Prediction Models for Wildfires

IS 467 - Fundamentals of Data Science

Spring 2018

David Abramov

Alicia Boyd

Erin Murphy

Abstract

Wildfires can result in devastating damages, including homes and lives. According to the National Interagency Fire Center, an average of 822,703 acres per year (between 2008 and 2018) have been touched by wildfires in the United States (Meyer, 2017). In the past decade, wildfires have caused $5.1 billion in damages (Short, 2017). Understanding wildfire trends and predictors can help in aiding professionals to prevent and contain wildfires, as well as reduce the amount of damage (both financially and psychologically). Focusing on 2015 data from the Geospatial database of U.S. wildfires, our initial questions are: (1) What are good predictors for determining duration, size, cause and frequency of wildfires in the US? and (2) How have wildfire trends changed over time?

We used two data mining techniques analyze this data set: clustering and decision trees. Hierarchical agglomerative clustering was used to conduct exploratory analysis in an attempt to create new categorical descriptions of the data. Using this technique, causes of fires could be grouped based on the cause of the fire (human, natural or industrial) as well as the region in which they occured. This dataset could however be missing some confounding variables, such as precipitation or temperature. The final clustering model used the variables Discovery Day of Year (normalized), Cause of Fire, Time of Day (binned), Fire Elapsed Hours (normalized), Owner (binned), Fire Size (binned), Latitude (normalized) and Longitude (normalized). Three clusters were chosen. Fire duration, cause and sizes were three target variables used to create various decision tree models. For each target variable, we looked at the performance of the final model on a test set with equal distributions of the classifier variables, and the performance on the initial population before it was undersampled. It was determined that the decision tree predicting fire size was the best model.

Problem Description

It seems that more and more, new reports are coming out about wildfires occurring in the United States, especially on the west coast. It could be that temperatures seem to be warmer, or conditions drier than normal (Meyer, 2017). Wildfires can result in devastating damages, including homes and lives. According to the National Interagency Fire Center, an average of 822,703 acres per year (between 2008 and 2018) have been touched by wildfires in the United States (Meyer, 2017). In the past decade, wildfires have caused $5.1 billion in damages (Short, 2017). Therefore, understanding wildfire trends and predictors can help in aiding professionals to prevent and contain wildfires, as well as reduce the amount of damage (both financially and psychologically).

According to Walters (n.d.), The “U.S. wildfires between 1992-2015” dataset was downloaded from Kaggle.com to create predictive models, specifically using hierarchical agglomerative clustering and decision tree techniques, to attempt to predict causes, sizes and durations of wildfires in the United States. Of all the variables in the dataset, specifically for the decision tree modeling, it was decided that three different models would be created, and would use the classifiers (for each respective model) for duration of fire, fire size, and cause of the fire. These classifier variables were chosen because they seemed like the most logical variables in which to classify the data. Understanding how long fires burn, their size, and what causes wildfires can potentially be useful information for governmental and private sectors to use to either prevent wildfires or contain them in a way so as to reduce their size, duration and resultantly, their devastating impacts.

The Dataset

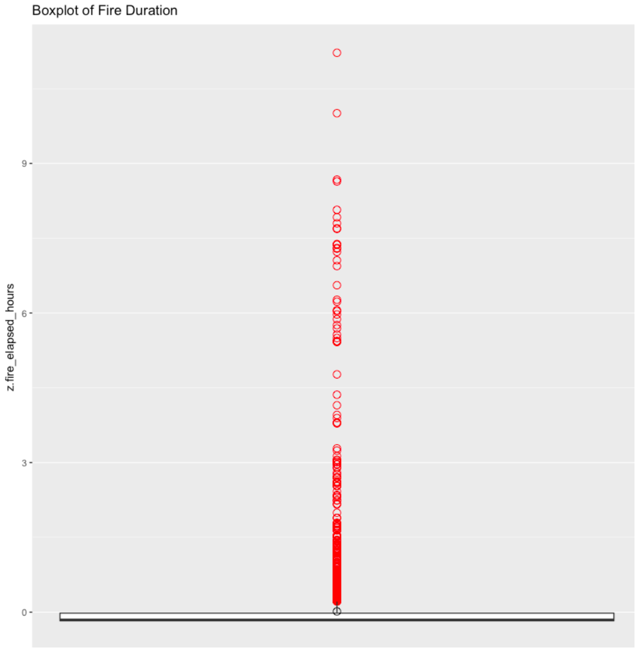
As mentioned, the wildfires dataset was downloaded from Kaggle.com as a SQLite database. The data contained latitude and longitude metrics, as well as 35 additional categorical and numerical attributes (see Appendix - Dataset Attributes for a list of all variables and their definitions). The records contained data from governmental and private entities, and contained more than 1.88 million records on wildfires that had occurred from 1992 - 2015.

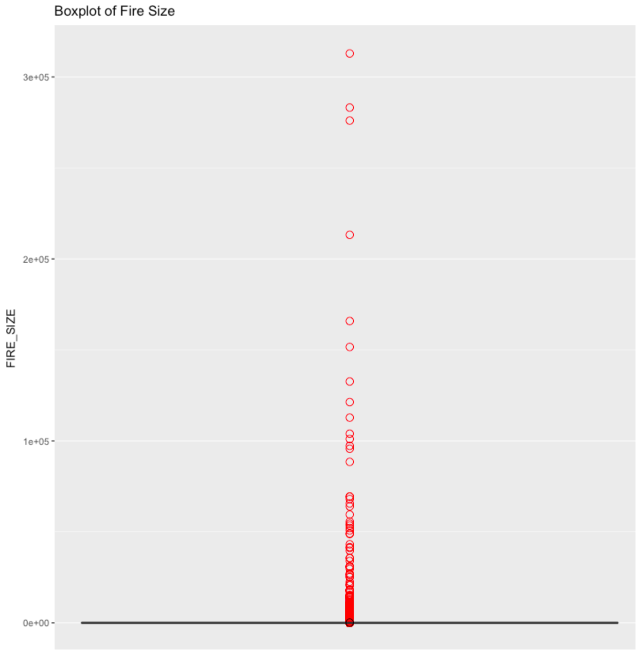
By performing an initial exploratory data analysis on all data in 2015, visualizations were created in Tableau to show “Causes of Wildfires", “Fire Duration” and “Fire Size”. It appears that more fires in Alaska and the Northwest are caused by lightening, whereas fires in the East and Southeast are predominantly caused by human factors. When looking at the “Fire Duration” and “Fire Size” visualizations, it also appears that fires that have a longer duration and larger size tend to occur in Alaska and the Northwest, whereas smaller, shorter fires occur in the South and Southeast. Reasons for this could potentially be due to less access and/or resources to contain wildfires in more rural areas. See Appendix - Data Visualizations - Exploratory Analysis for more information.

Preprocessing Steps

In SQLite, two new variables were created: fire_elapsed_hours (the number of hours from CONT_DATE + CONT_TIME and DISCOVERY_DATE + DISCOVER_TIME) and season (where fires in January, February and December were labeled as ‘Winter’; March, April and May as ‘Spring’; June, July and August as ‘Summer’; and September, October and November as ‘Fall’). See Appendix - SQL Code for the actual code used to create these variables. This data was then exported to a text file and read into R Studio, and the rest of the analysis was performed in R. In R, data was subsetted to only include data for wildfires that occurred in 2015. Of this subset, records with any missing data elements were removed from the sample.

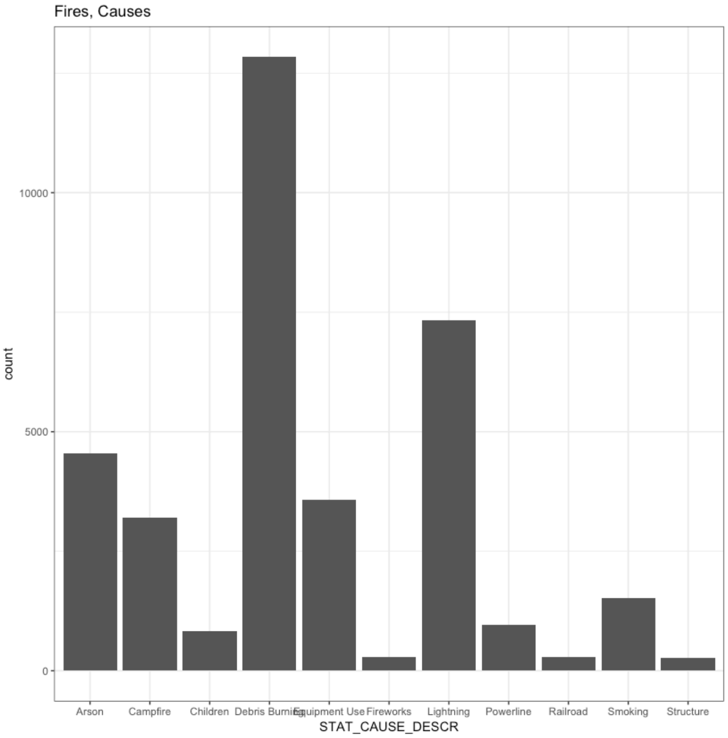
New variables were also created once the data was imported into R, which included binned variables. Based on the box plots of the classifier variables, fire duration and fire size were skewed to the right and needed to be binned for a more normal distribution.





Fire causes also showed an uneven distribution based on the histogram of fire causes (with “Miscellaneous” and “Missing/Undefined” removed).

As a result, fire_elapsed_hours (fire duration), FIRE_SIZE and STAT_CAUSE_DESCR (Cause of Fire) were all binned. The fire_elapsed_hours variable was binned so that the variable was converted from representing hours to days (< 0.025, 0.025-0.05, 0.05-0.125, 0.125-0.25, 0.25-0.5, 0.5-1, 1, 2 and 3+ days). FIRE_SIZE was categorized into 3 different bins (1, 2 and 3 - 7), and STAT_CAUSE_DESCR was binned into four bins: People, Industry, Debris Burning, and Lightning. This created more evenly distributed variables (see Appendix - Binned Histograms), yet there were still some classifier variables that had much less representation in the data than others. As a result, three final datasets were created--1 for each classifier, and within the 3 final datasets, the binned classifier variables were undersampled so that each bin within each final dataset had an equal number of records represented per bin.



Three more variables were created: Region (based on state and geographic region of the country) (US Census Bureau, 2017), owner.binned (describing if the land where the fire originated was privately owned or not), and time.of.day (6am - 12pm = morning, 12pm to 4pm = afternoon, 4pm - 9pm = evening, and the rest = night). The following variables were used to create new variables:

- FIRE_YEAR: Calendar year the fire was discovered or confirmed to exist
- DISCOVERY_DATE: Date the fire was discovered or confirmed to exist
- DISCOVERY_TIME: Time the fire was discovered or confirmed to exist
- DISCOVERY_DOY: Day of year the fire was discovered or confirmed to exist
- CONT_DATE: Date the fire was declared contained or otherwise controlled
- CONT_TIME: Time the fire was declared contained or otherwise controlled
- FIRE_SIZE: Acre estimate of the perimeter of the final fire size
- OWNER_DESC: Name of entity or owner managing the land where the fire originated
- STATE: State where the fire originated
- STAT_CAUSE_DESCR: Categorical description of the cause of the fire

Initial analysis of the data showed that nearly 30% of the data was made up of STAT_CAUSE_DESCR equal to "Miscellaneous" and "Missing/Undefined". It was therefore determined that these were considered unnecessary data points because they were not useful when used as a classifier (given that was one of the variables that was going to be predicted), and theses records were removed. This resulted in a sample dataset of 32,622 records.

After all of the data was cleaned and transformed, chi-square tests (of categorical variables) and correlation analyses were performed on the data to see if any variables exhibit multicollinearity (or dependence) with each other. With regards to the chi-square tests, and by observing the data visualizations that showed the fire duration, causes and size by region, the data showed that there is a slight correlation between Region and Owner, Cause and Duration. With regards to the correlation analyses, none of the continuous variables showed multicollinearity. As a result, it was decided to keep all of the variables in the dataset.

The final datasets that were used to build the three decision trees included the following variables: cause.binned, sizes.binned, owner.binned, region, Season, fire.duration. The final dataset for fire.duration contained 4500 records (the classifier had 9 bins, each with 500 samples); cause.binned contained 18,000 records (the classifier had 4 bins, and each bin contained 4500 records); and size.binned contained 12,000 records (the classifier had 3 bins, and each bin contained 4000 records).

The final dataset used for the hierarchical agglomerative cluster analysis used Discovery Day of Year (z-score scaled), cause.binned, time.of.day (binned), Fire Elapsed Hours (z-score scaled), Owner (binned), Fire Size (binned), Latitude (z-score scaled) and Longitude (z-score scaled)

Data Mining Techniques

Hierarchical Agglomerative Cluster Analysis

A hierarchical agglomerative clustering technique was used in exploratory analysis of this dataset. This is method is visualized as a dendrogram. It is a bottom-up algorithm, meaning each object is a leaf in the tree, and each leaf is considered a cluster. Pairs of clusters are compared and agglomerated (combined) until all of the clusters are combined into a one agglomerate containing all the original objects. This is an exploratory technique, used to identify potential patterns and relationships that are not immediately obvious from looking at the data -- perhaps there are classes that could be inferred from the data that was collected. This method was chosen over other “flat clustering” methods, like k-means or k-nearest neighbors, because it does not require a predetermined number of clusters in order to run the algorithm. Another difference is that the sizes of the clusters that result from agglomerative algorithms may be of various sizes, unlike k-means, which tends to produce groups of equal sizes. The main drawback to hierarchical clustering is that the complexity of the algorithm is at least quadratic with regards to the number of records analyzed. Because of this, a subset of only 1000 records from the original dataset were randomly selected in order to maximize efficiency.

There are several methods for comparing groups in agglomerative cluster analysis, namely single-linkage, complete-linkage, average-linkage, and ward-linkage. Each method was compared when running the first iteration of the cluster analysis. Single-linkage takes into account the similarities of only the most similar objects within each cluster (i.e. a single comparison); by far this method performed the worst. Complete-linkage on the other hand uses the distance between the most dissimilar objects within each cluster, resulting in clusters with relatively small diameters. This makes it particularly sensitive to outliers within the data. Average-linkage, as the name suggests, determines the distance between each cluster by taking averaging the distance between each object in one cluster with each object in the other. Ward's method is different in that it agglomerates clusters based on minimizing the error sum of squares in the resulting cluster. This provides a more holistic approach to combining clusters, as it doesn't place emphasis on any one particular object. Ward's method gave the largest agglomerative coefficient (0.998) when running the first iteration, so it was used for all subsequent clustering.

Linkage Method	**Agglomerative Coefficient (Iteration 1)**
Average	0.956873892460619
Single	0.841983439314069
Complete	0.975065373471492
Ward	0.997917539375865

For the first two iterations, data was selected from between 2012 to 2015, whereas in iterations 3 and 4 data was selected only from 2015. By isolating values to a single year, the Fire Year (normalized) variable was lost in the cluster analysis, resulting in a drop in the PCA result. This could be related to differences in precipitation in different regions and years, which would be a confounding variable not present in this dataset. In the first iteration, the State and Season for each object was included in the cluster analysis, however upon their removal the PCA result jumped nearly 7%. This is likely due to the strong relationships between State and Latitude/Longitude, as well as Season with Discovery Day of Year. Because of this, the Season and State features were excluded from all subsequent iterations. For the third iteration, most of the continuous numeric variables were binned (Time of Day, Fire Duration, Region, Owner (Private vs. Non-private land). A slightly greater PCA result was found in Iteration 4 by using the unbinned Fire Elapsed Hours as well as Latitude and Longitude in replace of the binned regions.

Results for Iteration 4 of the cluster analysis are shown in the table below. The total number of records sampled was 1,000 from 2015. A number of clusters *k* = 3 was chosen because it seemed to best highlight the main different causes for the fires and the time of year at which they occured. In the first cluster, there were 278 records, with the majority by far being lightning (72.3%) followed by smoking (14%), as well as a few caused by power lines, structures, railroads and fireworks (14.4% cumulatively). The top two regions were Mountains (82) and Pacific (68), suggesting that remote areas might be more prone to wildfires caused by lightning. The majority of these cases occurred during the summer. This category could be considered “Remote summer fires triggered by lightning and accidents.”

The second cluster contained nearly completely half of the records. The top cause was debris burning (68.7%), followed by equipment use (23.4%), with a few caused by children, campfires and fireworks (7.8%). The region containing the most of the fires was South Atlantic (31.7%), followed by more inland locations in the West North Central and East South Central regions. Interestingly, these occurred mostly in spring (47.5%). This cluster could be summarized as “Rural spring garbage and industrial fires.”

The third and final cluster contained about a quarter of the records. The leading cause of fires in this cluster was Arson (57%), followed by campfires (42.2%) and finally children (<1%). Most of these occurred in spring (40.2%) and about an equal distribution between winter, summer and fall. Arson and campfires are usually started on purpose, so this cluster can be called “Intentional fires gone wrong.”

Cluster Analysis Results.

	Iteration 1 (2012-2015)	**Iteration 2 (2012-2015)**	**Iteration 3 (2015)**	**Iteration 4 (2015)**
Variables chosen for Agglomerative Clustering	Fire Year (normalized)	Fire Year (normalized)	Discovery Day of Year (normalized)	Discovery Day of Year (normalized)
	Fire Elapsed Hours (normalized)	Fire Elapsed Hours (normalized)	Cause of Fire	Cause of Fire
	Discovery Day of Year (normalized)	Discovery Day of Year (normalized)	Time of Day (binned)	Time of Day (binned)
	Latitude (normalized)	Latitude (normalized)	Fire duration (binned)	Fire Elapsed Hours (normalized)
	Longitude (normalized)	Longitude (normalized)	Region (binned)	Owner (binned)

	Cause of Fire	Cause of Fire	Owner (binned)	Fire Size (binned)
	Fire size (binned)	Fire size (binned)	Fire size (binned)	Latitude (normalized)
	State			Longitude (normalized)
	Season			
Agglomerative Coefficient	1	0.99	0.99	0.99
PCA Result	39.99%	46.93%	38.22%	38.99%

Iteration 4	**Cause (Frequency)**	**Top 4 Regions (Frequency)**	**Season**	**New Cluster Names**
Cluster 1 (n = 278)	Lightning: 201 Smoking: 39 Powerline: 19 Structure: 10 Railroad: 6 Fireworks: 5	Mountain: 82 Pacific: 68 South Atlantic: 48 Middle Atlantic: 37	Winter: 7 Spring: 70 Summer: 166 Fall: 35	Remote Summer Fires Triggered by Lightning and Accidents
Cluster 2 (n = 499)	Debris Burning: 343 Equipment Use: 117 Children: 27 Campfire: 6 Fireworks: 6	South Atlantic: 158 West North Central: 98 East South Central: 61 Pacific: 55	Winter: 87 Spring: 237 Summer: 80 Fall: 95	Rural Spring Garbage and Industrial Fires
Cluster 3 (n = 223)	Arson: 127 Campfire: 94 Children: 2	South Atlantic: 35 East South Central: 34 Middle Atlantic: 34 Pacific: 31	Winter: 41 Spring: 92 Summer: 49 Fall: 41	Intentional Fires Gone Wrong

Decision Trees

Decision tree is one of many supervised learning techniques mainly used for classification. This method can be used on continuous and categorical variables. One of the numerous features of decision trees is the ability to split the sample into homogeneous subgroups based on the variables in question. Our reasons for implementing the decision tree technique on our data aligns with the advantages of using this method. First, our group agreed we needed a fairly easy way to interpret and intuitively understand the data using a graphic representation. Secondly, a decision tree helps us to explore the data and understand trends along the variables determining which is more significant. Lastly, we didn’t need to transform (or scale) the data. Although, there was some binning of the data based from our feedback in the in-class presentation.

For our project, we wanted to use the decision tree to predict explore different aspects how forest fires are started. Before seeking an answer to this question, we first had to clean the data and format it in before creating a decision tree. The variables used for our decision tree model are the following: Fire Duration (numeric), Cause (categorical), Sizes (categorical), Owner Description (categorical), Region (categorical), and Season (categorical). Note: In parentheses designates the type of variable used. Originally, we split the data into 80 percent for training and 20 percent for testing. Although we settled for 60 percent training and 40 percent testing for the size of the data set. We made this change due to the fact there were large number of records and we wanted an equal representation of the data.

Our methodology for growing our decision tree was to set the minimum number of observations before attempting to split the tree was five. Our complexity parameter was set at zero. We start with zero to get the biggest tree, then we will utilize a function in R called printcp() to determine the best complexity parameter, which is lowest xerror divided by penalty factor. We explored three different target variables to better predict forest fires in the United States. Before our in-class presentation, we did not bin of the variables in the data. After receiving the feedback to minimize the presence of outliers we removed the miscellaneous observations in the data. In addition, it was recommended by the professor to equally distribute our target variables and compare the results to the model as whole as it is without equal distributions.

Accuracy and Misclassification Rates for 3 Decision Tree Models.

Target Variable	Accuracy (equal distribution)	Misclassification Rate (equal distribution)	Accuracy	Misclassification Rate	Important Variables
Fire Duration	.2755556	.7244444	.3309423	.6690577	cause, owner, region, season and sizes
Cause	.5463889	.4536111	.5607565	.4392435	fire duration, owner, region, season and sizes
Sizes	.630625	.369375	.6374839	.3625161	Region and Fire duration

Our first target variable was fire duration where we converted and binned this variable. We converted the variable to correspond to the number of days. (Note: Please refer to earlier discussion about the conversion of variables for more details.) We wanted to see if we could predict forest fires based on fire duration. Running the code in R, we determined our complexity parameter was 0.005899705 based on function printcp(). The following two target variables--cause and sizes-- had a complexity parameter of numeric(0) and .00293495, respectively.

The first model for the equal distributions is a better model for sizes. The variables that were used in tree construction: Analyzing the three different target variables along with their equal distribution and whole models, all of the models which represented equal distributions had higher accuracy scores. Focusing on the equal distribution accuracy scores for each of the target variables, it appears the fire duration model has the highest accuracy score compared to Cause and Size models.

Conclusions

The cluster algorithm chosen to do exploratory analysis on this dataset was the hierarchical agglomerative clustering technique. This is a bottom-up approach where each object is considered a cluster, and is subsequently compared with all other clusters until there is one big cluster containing all the objects. It is visualized in a dendrogram, and does not require a predetermined number of clusters in order to be run. The best PCA resulting from this method on this dataset accounted for 46.93% of the variation. The PCA resulting from the final clustering iteration was 38.99%. The main features that stood out as related within each cluster were the cause of the fire, the region in which the fire occurred, and the time of year. This makes sense considering different regions may have different environmental regulations regarding fire safety, or some regions may be more remote than others. Spring and summer had the most fires, which makes sense considering it is warmer and potentially drier. The causes of fires could be further described as natural (lightning), and human related. The three clusters created from this analysis could be described as Remote Summer Fires Triggered by Lightning and Accidents, Rural Spring Garbage and Industrial Fires, and Intentional Fires Gone Wrong. Some ways to possibly improve the quality of the clustering would be to categorize the regions based on Koppen climate type rather than purely geographic regions, or to combine this dataset with historic precipitation data and use that instead of the time variables.

The best decision tree model was the tree which predicted fire size. This was based on a number of factors which included a high accuracy rate on both the equal distribution data set and the original data set, as well as a smaller, more pruned tree. Pruning the tree was based on a CP of 0.002724796 because after this value, there is little gain in terms of accuracy with adding more splits to, or growing the tree. As a result, it also has the lowest rate of misclassification error. The most important factors of the sizes decision tree are fire duration and region.

It would be recommended to test these models on previous years of the data, as well as data from 2016 and 2017. Especially with the perceived increase in wildfires (National Interagency Fire Center, 2018; www.iii.org, 2018) , it is important to be able to understand not only what potential causes for wildfires are, but also their size and duration. Seeing that wildfires in the Northwest and Alaska see to have the largest sizes, this model could, for one application, show if measures used to decrease wildfire sizes are indeed working.

Another recommendation would be to use this model on data of wildfires in other parts of the world and not just the United States. This would be an interesting follow-up, and if shown to be accurate, would increase the validity of the model.

References

"Facts + Statistics: Wildfires." Facts + Statistics: Wildfires | III, 2018, www.iii.org/fact-statistic/facts-statistics-wildfires.

Meyer, R. (2017, September 07). Has Climate Change Intensified 2017's Western Wildfires? Retrieved June 6, 2018, from https://www.theatlantic.com/science/archive/2017/09/why-is-2017-so-bad-for-wildfires-climate-change/539130/

National Interagency Fire Center, 2018, www.nifc.gov/fireInfo/nfn.htm.

Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

U.S. Census Bureau. *Census Regions and Divisions of the United States*. *Census Regions and Divisions Of the United States*, https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf.

Walters, Troy. "Wildfire Exploratory Analysis." *Kaggle*, www.kaggle.com/captcalculator/wildfire-exploratory-analysis/data.

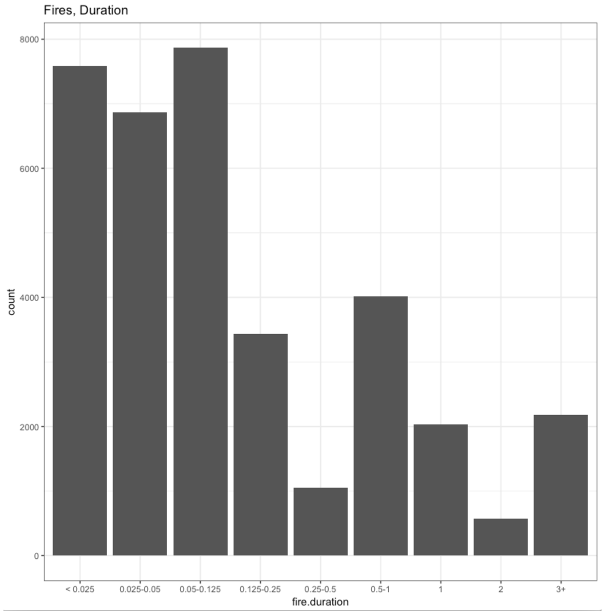
Appendices

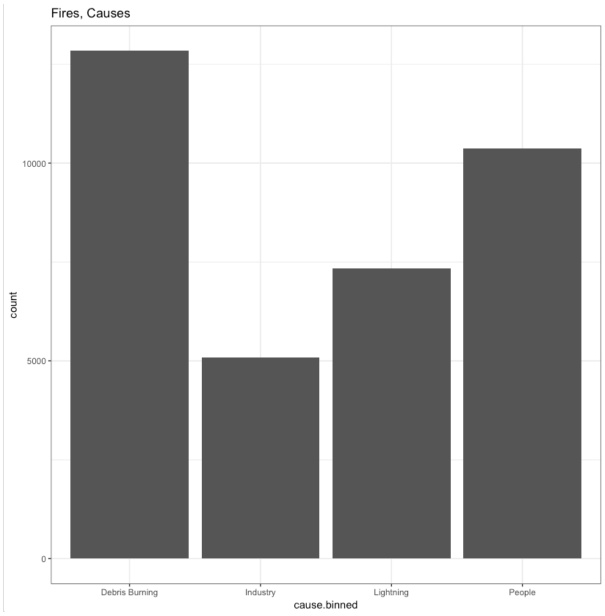
Dataset Attributes

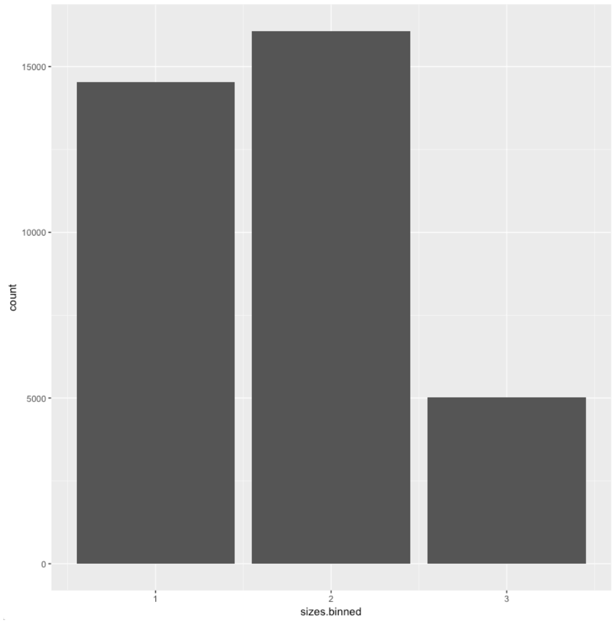
Fires: Table including wildfire data for the period of 1992-2015 compiled from US federal state and local reporting systems.	
# of Records = 1.88 Million	
Variable Name	**Variable Description**
FOD_ID	Global unique identifier.
FPA_ID	Unique identifier that contains information necessary to track back to the original record in the source dataset.
SOURCE_SYSTEM_TYPE	Type of source database or system that the record was drawn from
SOURCE_SYSTEM	Name of or other identifier for source database or system that the record was drawn from.
NWCG_REPORTING_AGENCY	Active National Wildlife Coordinating Group (NWCG) Unit Identifier for the agency preparing the fire report
NWCG_REPORTING_UNIT_ID	Active NWCG Unit Identifier for the unit preparing the fire report.
NWCG_REPORTING_UNIT_NAME	Active NWCG Unit Name for the unit preparing the fire report.
SOURCE_REPORTING_UNIT	Code for the agency unit preparing the fire report
SOURCE_REPORTING_UNIT_NAME	Name of reporting agency unit preparing the fire report
LOCAL_FIRE_REPORT_ID	Number or code that uniquely identifies an incident report for a particular reporting unit and a particular calendar year.
LOCAL_INCIDENT_ID	Number or code that uniquely identifies an incident for a particular local fire management organization within a particular calendar year.
FIRE_CODE	Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression (https://www.firecode.gov/).
FIRE_NAME	Name of the incident
ICS_209_INCIDENT_NUMBER	Incident (event) identifier
ICS_209_NAME	Name of the incident
MTBS_ID	Incident identifier
MTBS_FIRE_NAME	Name of the incident
COMPLEX_NAME	Name of the complex under which the fire was ultimately managed

FIRE_YEAR	Calendar year in which the fire was discovered or confirmed to exist.
DISCOVERY_DATE	Date on which the fire was discovered or confirmed to exist.
DISCOVERY_DOY	Day of year on which the fire was discovered or confirmed to exist.
DISCOVERY_TIME	Time of day that the fire was discovered or confirmed to exist.
STAT_CAUSE_CODE	Code for the (statistical) cause of the fire.
STAT_CAUSE_DESCR	Description of the (statistical) cause of the fire.
CONT_DATE	Date on which the fire was declared contained or otherwise controlled
CONT_DOY	Day of year on which the fire was declared contained or otherwise controlled.
CONT_TIME	Time of day that the fire was declared contained or otherwise controlled
FIRE_SIZE	Estimate of acres within the final perimeter of the fire.
FIRE_SIZE_CLASS	Code for fire size based on the number of acres within the final fire perimeter expenditures
LATITUDE	Latitude (NAD83) for point location of the fire (decimal degrees).
LONGITUDE	Longitude (NAD83) for point location of the fire (decimal degrees).
OWNER_CODE	Code for primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
OWNER_DESCR	Name of primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
STATE	Two-letter alphabetic code for the state in which the fire burned (or originated)
COUNTY	County in which the fire burned (or originated)
FIPS_CODE	Three-digit code from the Federal Information Process Standards (FIPS) publication 6-4 for representation of counties and equivalent entities.
FIPS_NAME	County name from the FIPS publication 6-4 for representation of counties and equivalent entities.

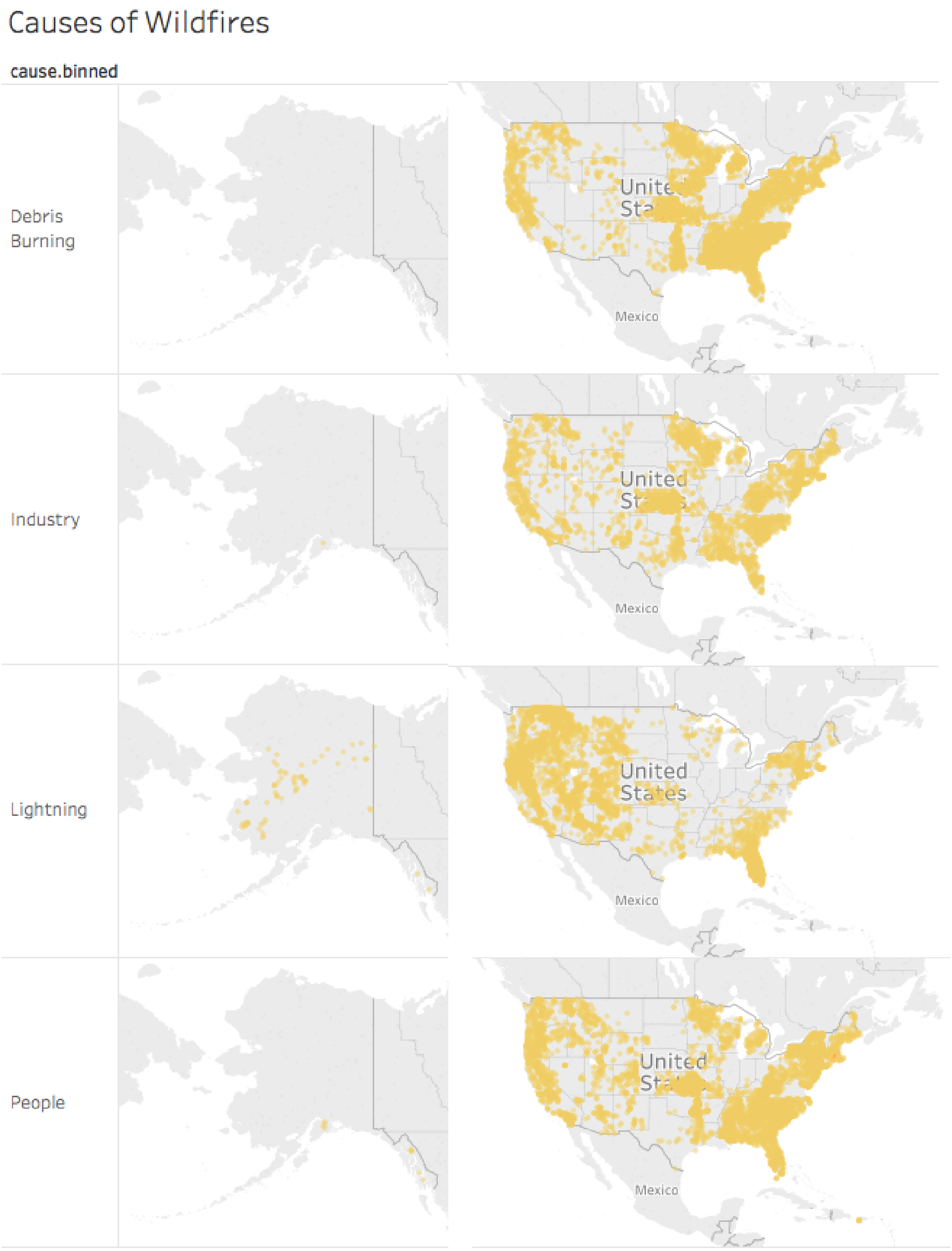
Binned Histograms



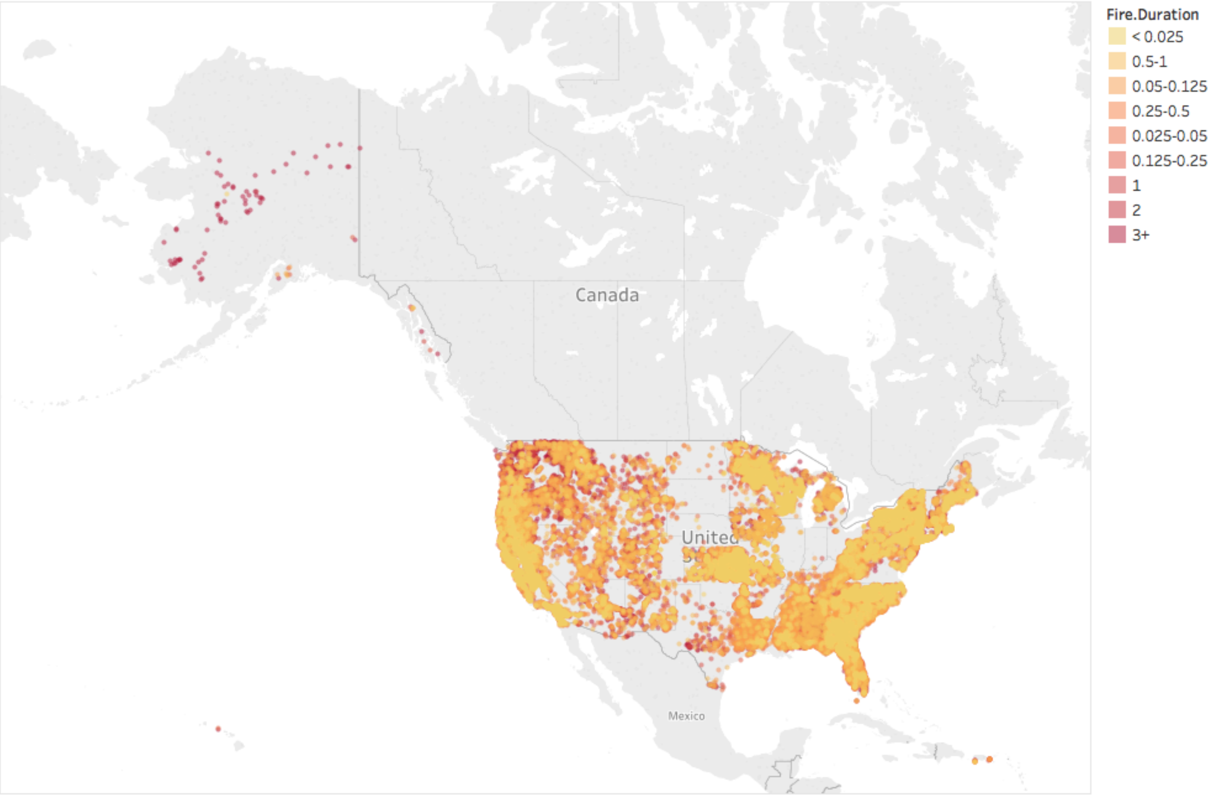




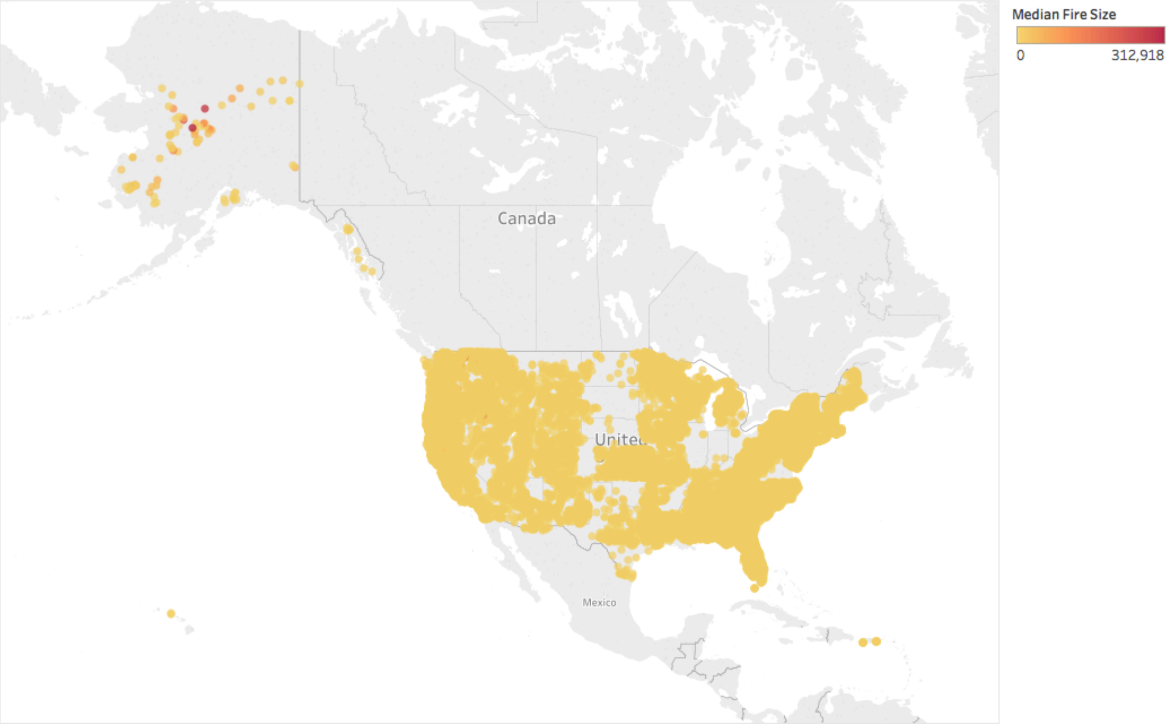
Data Visualizations - Exploratory Analysis



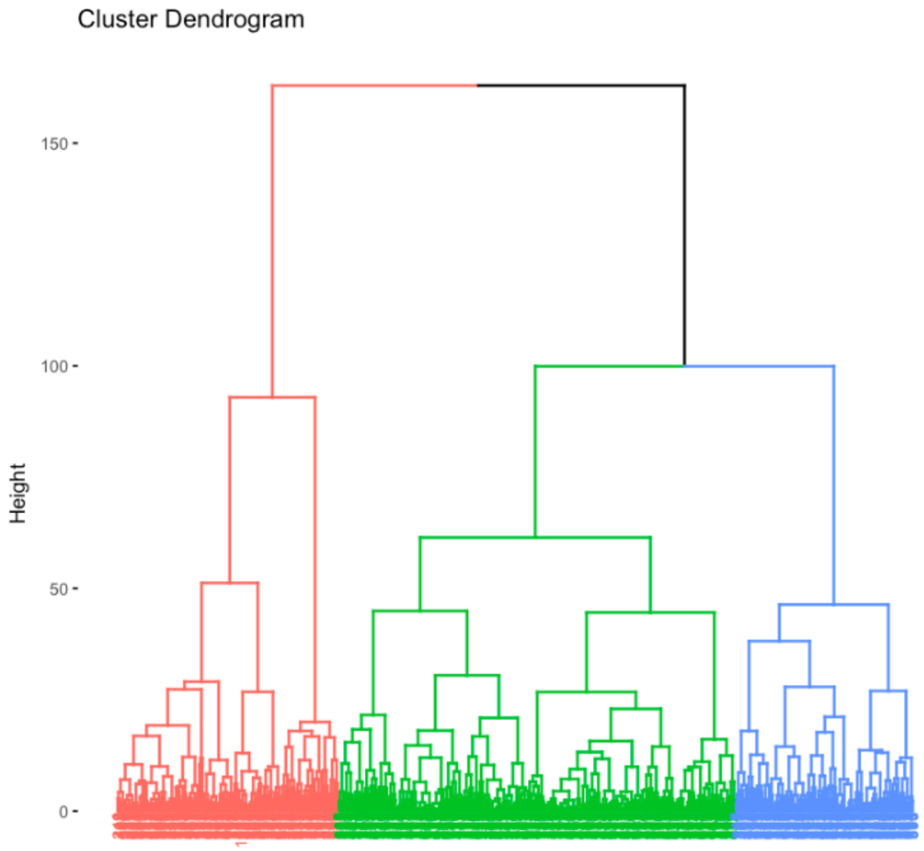
Fire Duration, 2015

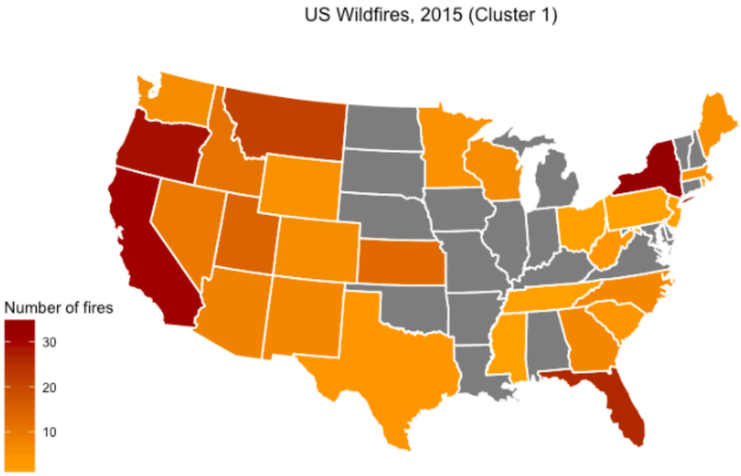


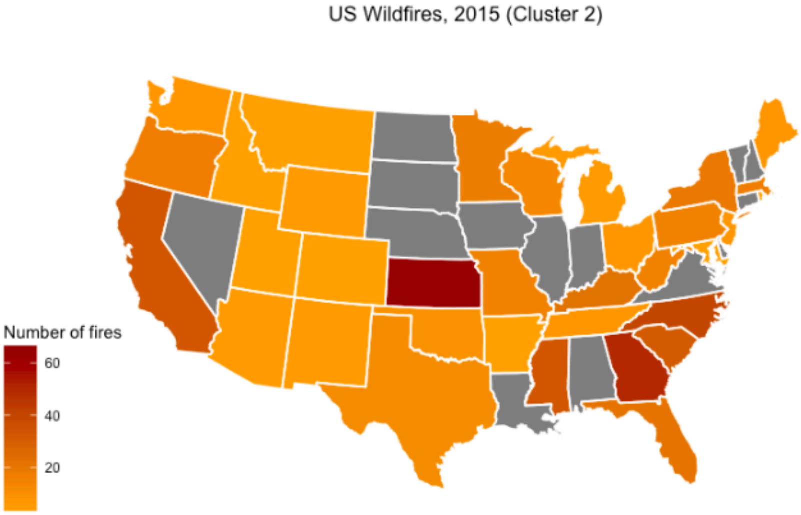
Median Fire Size

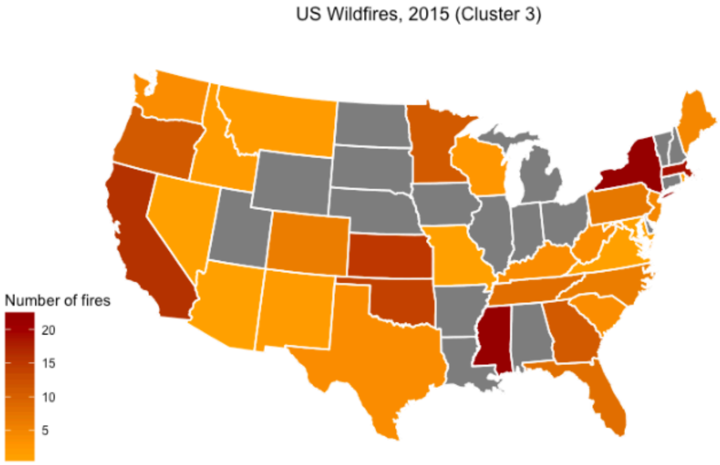


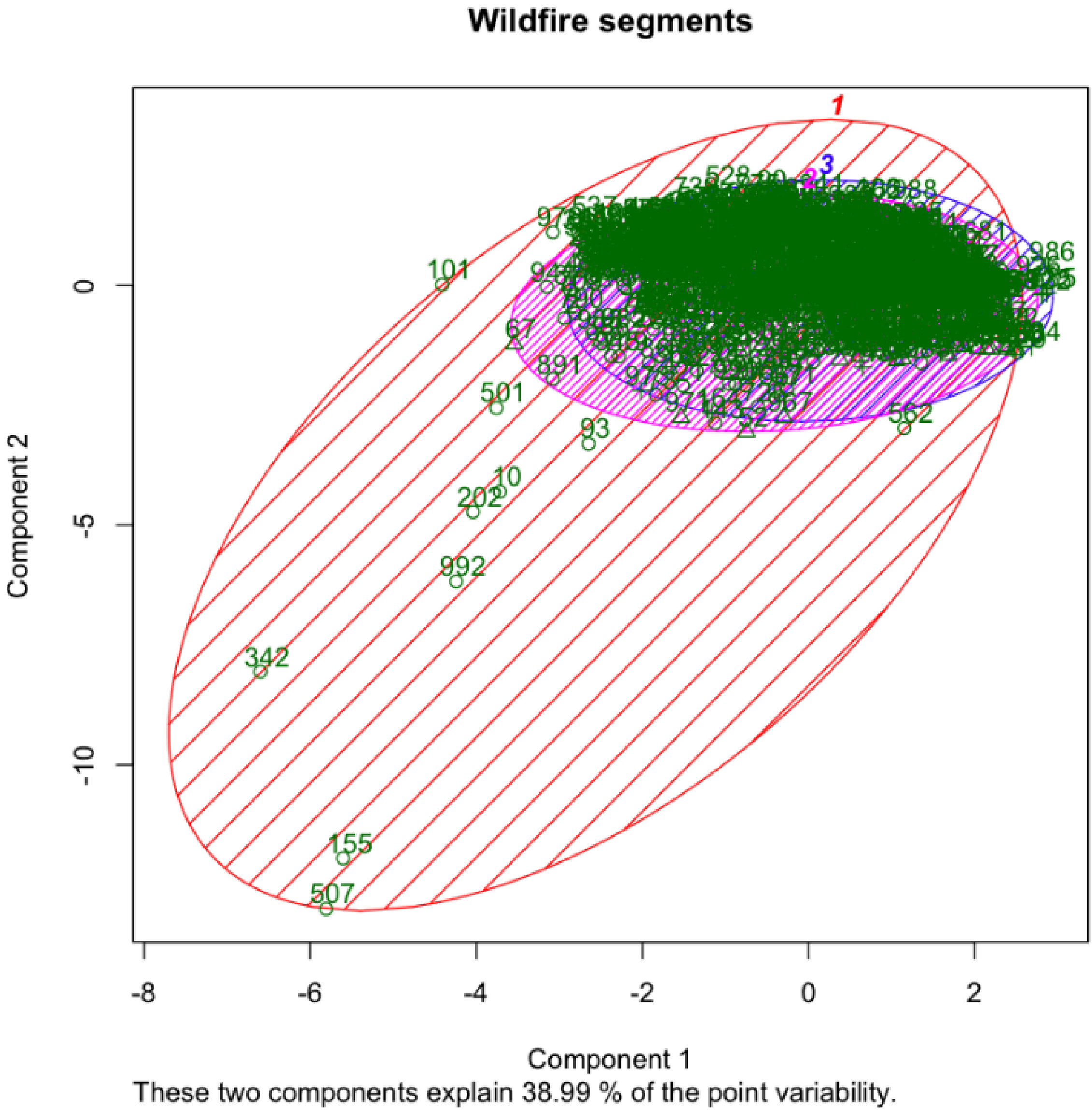
Data Visualizations - Hierarchical Agglomerative Cluster Analysis and PCA



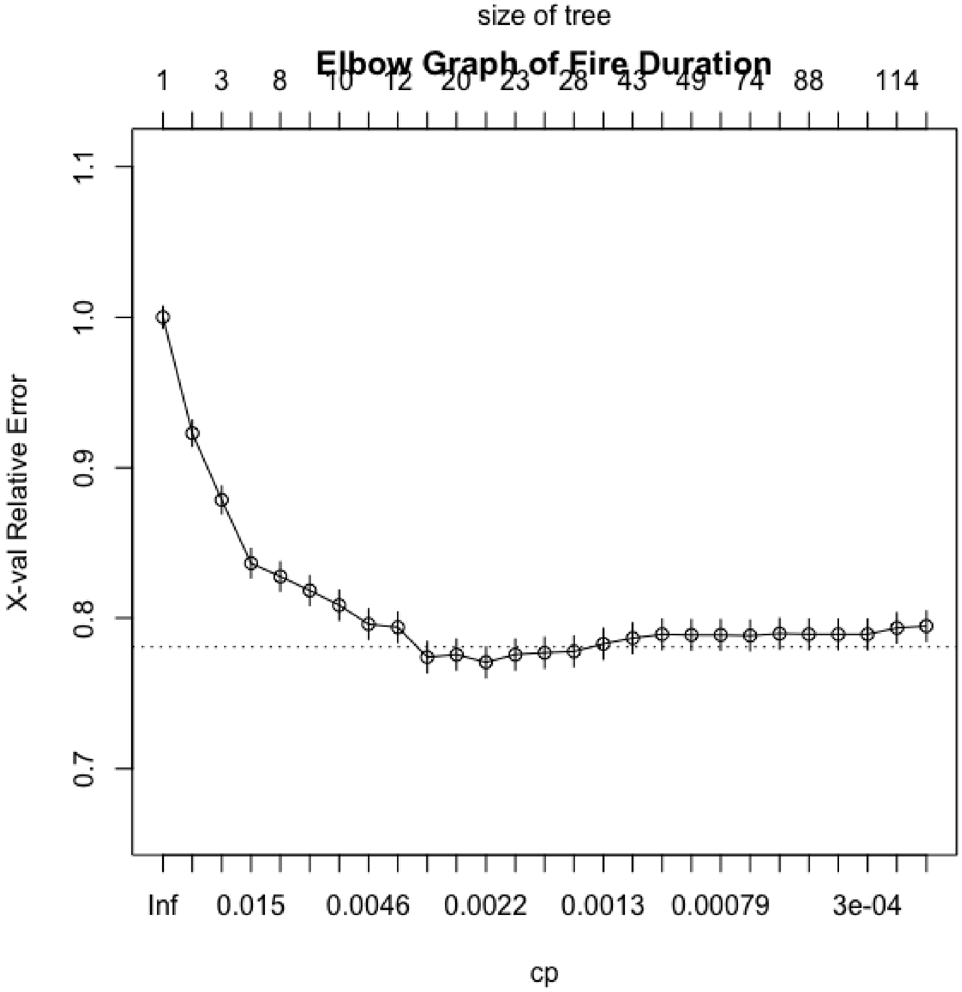

US Wildfires, 2015 (Cluster 1)
Number of fires
30
20
10


US Wildfires, 2015 (Cluster 2)
Number of fires
60
40
20

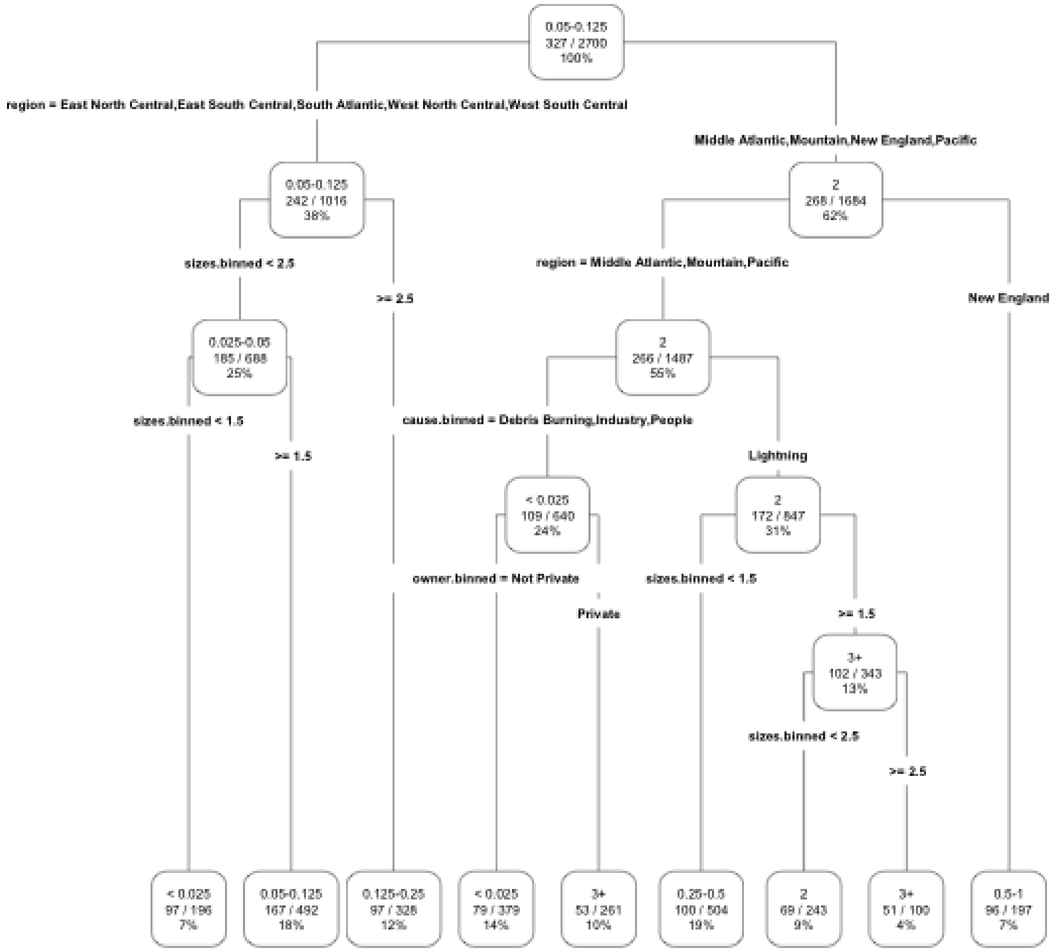

US Wildfires, 2015 (Cluster 3)
Number of fires
20
15
10
5

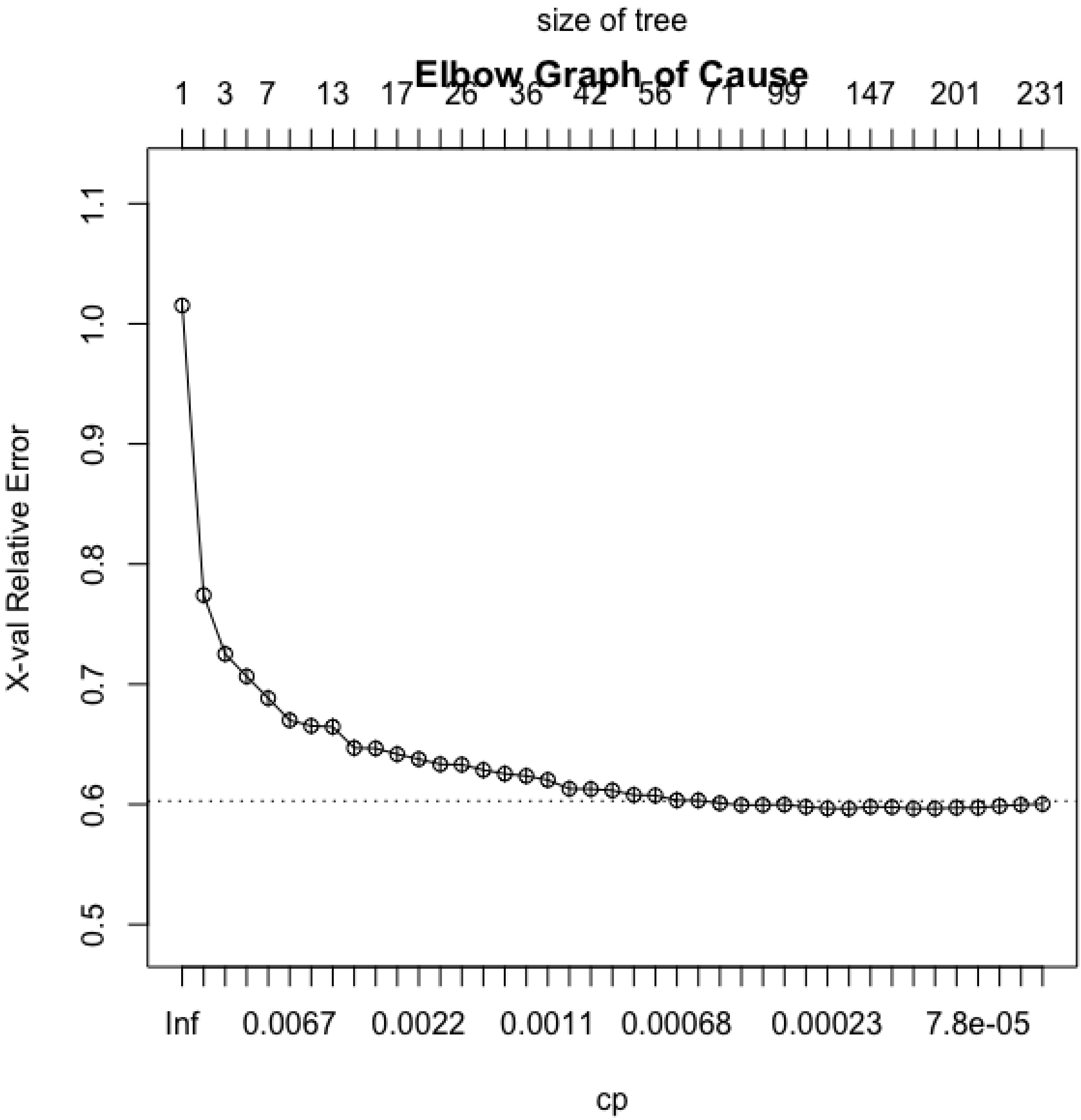

Wildfire segments
Component 1
Component 2
These two components explain 38.99 % of the point variability.
-8
-6
-4
-2
0
2
0
-5
-10
1
2
3
101
67
891
501
93
10
202
992
342
155
507
562
986
528
537
581
94
97

Data Visualizations - Decision Tree



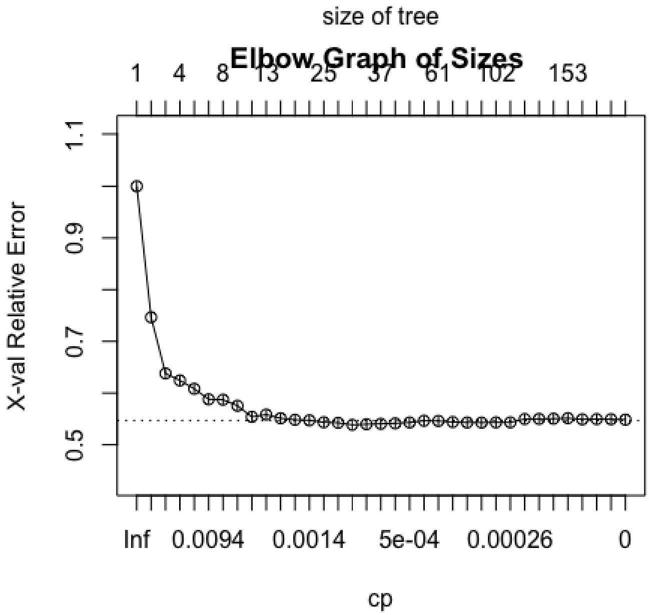
Pruned Classification Tree for Fire Duration

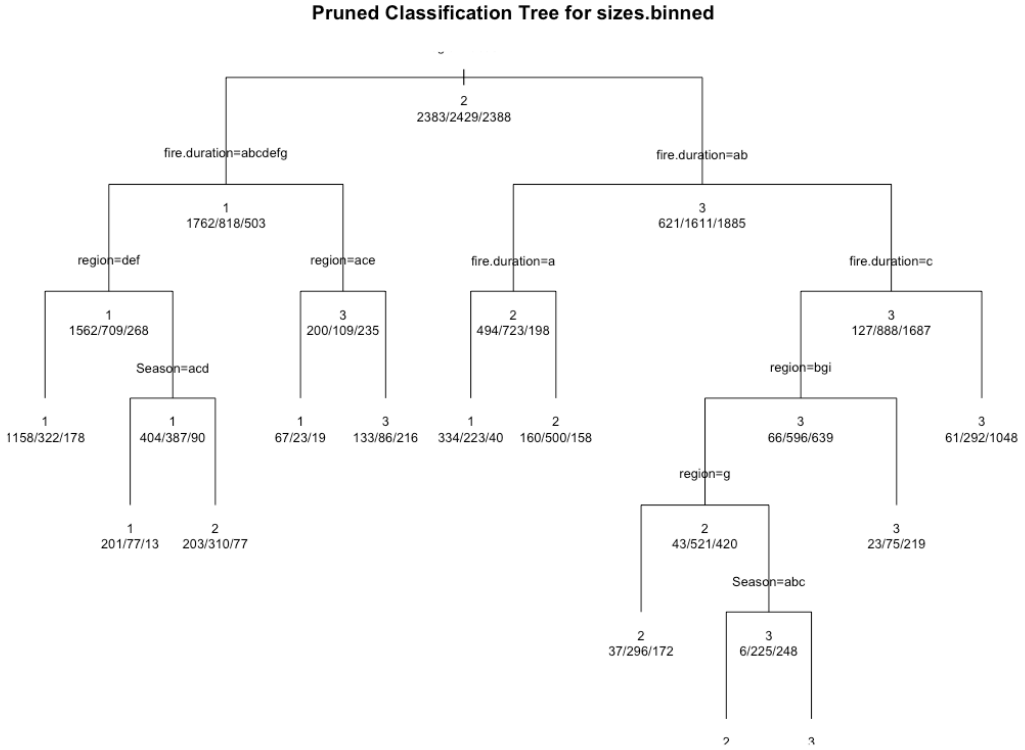



size of tree
Elbow Graph of Cause
1
3
7
13
17
26
36
42
56
71
99
147
201
231
X-val Relative Error
1.1
1.0
0.9
0.8
0.7
0.6
0.5
Inf
0.0067
0.0022
0.0011
0.00068
0.00023
7.8e-05
cp

Pruned Classification Tree for Cause

Debris Burning
Industry
Lightning
People
Debris Burning
Season = Fall,Spring,Winter
Summer
Lightning
Middle Atlantic,Mountain,Pacific
People


size of tree
Elbow Graph of Sizes
1
4
8
13
25
37
61
102
153
X-val Relative Error
1.1
0.9
0.7
0.5
Inf
0.0094
0.0014
5e-04
0.00026
0
cp


Pruned Classification Tree for sizes.binned
2
2383/2429/2388
fire.duration=abcdefg
fire.duration=ab
1
1762/818/503
3
621/1611/1885
region=def
region=ace
fire.duration=a
fire.duration=c
1
1562/709/268
3
200/109/235
2
494/723/198
3
127/888/1687
Season=acd
region=bgi
1
1158/322/178
1
404/387/90
1
67/23/19
3
133/86/216
1
334/223/40
2
160/500/158
3
66/596/639
3
61/292/1048
region=g
1
201/77/13
2
203/310/77
2
43/521/420
3
23/75/219
Season=abc
2
37/296/172
3
6/225/248
2
3

Pre-processing, Transformation and Reduction Code (SQL)

```
select
  x.*
  , case  when (discovery_datetime is null or containment_datetime is null) then mean_fire_time
      else cast((strftime('%s',containment_datetime)-strftime('%s',discovery_datetime)) AS real)/60/60
    end AS fire_elapsed_hours
from
(
    select f.*
    , case  when strftime('%m',DISCOVERY_DATE) in ('01','02','12') THEN 'Winter'
        when strftime('%m',DISCOVERY_DATE) in ('03','04','05') THEN 'Spring'
        when strftime('%m',DISCOVERY_DATE) in ('06','07','08') THEN 'Summer'
        when strftime('%m',DISCOVERY_DATE) in ('09','10','11') THEN 'Fall'
      end as Season
    , datetime(DISCOVERY_DATE, time(cast(substr(DISCOVERY_TIME, 1, 2) || ":" ||
substr(DISCOVERY_TIME, 3, 2) as char(5)))) as discovery_datetime
    , datetime(CONT_DATE, time(cast(substr(CONT_TIME, 1, 2) || ":" || substr(CONT_TIME, 3, 2) as
char(5)))) as containment_datetime
    from Fires as f
    where FIRE_YEAR > 2011
) as x
left join
(
    select
      Season
      , STAT_CAUSE_DESCR
      , avg(elapsed) as mean_fire_time
    from
    (
    select
      x.*
      , cast((strftime('%s',containment_datetime)-strftime('%s',discovery_datetime)) AS real)/60/60 AS
elapsed
    from
    (
        select f.*
        , case  when strftime('%m',DISCOVERY_DATE) in ('01','02','12') THEN 'Winter'
            when strftime('%m',DISCOVERY_DATE) in ('03','04','05') THEN 'Spring'
            when strftime('%m',DISCOVERY_DATE) in ('06','07','08') THEN 'Summer'
            when strftime('%m',DISCOVERY_DATE) in ('09','10','11') THEN 'Fall'
          end as Season
        , datetime(DISCOVERY_DATE, time(cast(substr(DISCOVERY_TIME, 1, 2) || ":" ||
substr(DISCOVERY_TIME, 3, 2) as char(5)))) as discovery_datetime
        , datetime(CONT_DATE, time(cast(substr(CONT_TIME, 1, 2) || ":" || substr(CONT_TIME, 3,
2) as char(5)))) as containment_datetime
```

```
                from Fires as f
                where FIRE_YEAR > 2011
        ) as x
        ) as x1
group by 1, 2
) as a
on x.Season = a.Season
and x.STAT_CAUSE_DESCR = a.STAT_CAUSE_DESCR
;
```

Pre-processing, Transformation and Reduction Code (R)

```
library(rpart)
library(rpart.plot)
library(dplyr)
library(e1071)
library(caret)
library(MASS)

tbl1 <- read.csv(file="/Users/erinmurphy/Documents/school/IS 467/Final Project/final_fires_data.csv",
header=FALSE, sep=",")
names(tbl1) <- c("OBJECTID", "FOD_ID", "FPA_ID", "SOURCE_SYSTEM_TYPE",
"SOURCE_SYSTEM", "NWCG_REPORTING_AGENCY", "NWCG_REPORTING_UNIT_ID",
"NWCG_REPORTING_UNIT_NAME", "SOURCE_REPORTING_UNIT",
"SOURCE_REPORTING_UNIT_NAME", "LOCAL_FIRE_REPORT_ID", "LOCAL_INCIDENT_ID",
"FIRE_CODE", "FIRE_NAME", "ICS_209_INCIDENT_NUMBER", "ICS_209_NAME", "MTBS_ID",
"MTBS_FIRE_NAME", "COMPLEX_NAME", "FIRE_YEAR", "DISCOVERY_DATE",
"DISCOVERY_DOY", "DISCOVERY_TIME", "STAT_CAUSE_CODE", "STAT_CAUSE_DESCR",
"CONT_DATE", "CONT_DOY", "CONT_TIME", "FIRE_SIZE", "FIRE_SIZE_CLASS", "LATITUDE",
"LONGITUDE", "OWNER_COD", "OWNER_DESCR", "STATE", "COUNTY", "FIPS_CODE",
"FIPS_NAME", "Shape", "Season", "discovery_datetime", "containment_datetime","fire_elapsed_hours")

##################################################
# Pre-Processing, Cleansing and Transformation
##################################################
tbl1$z.FIRE_YEAR <- scale(x = tbl1$FIRE_YEAR, center = TRUE, scale = TRUE)
tbl1$z.FIRE_SIZE <- scale(x = tbl1$FIRE_SIZE, center = TRUE, scale = TRUE)
tbl1$z.fire_elapsed_hours <- scale(x = tbl1$fire_elapsed_hours, center = TRUE, scale = TRUE)
tbl1$z.DISCOVERY_DOY <- scale(x = tbl1$DISCOVERY_DOY, center = TRUE, scale = TRUE)
tbl1$z.DISCOVERY_TIME <- scale(x = tbl1$DISCOVERY_TIME, center = TRUE, scale = TRUE)
tbl1$z.LATITUDE <- scale(x = tbl1$LATITUDE, center = TRUE, scale = TRUE)
tbl1$z.LONGITUDE <- scale(x = tbl1$LONGITUDE, center = TRUE, scale = TRUE)

# remove NA values and causes = "Miscellaneous", "Missing/Undefined"
tbl2 <- subset(tbl1, FIRE_YEAR == 2015 & is.na(FIRE_SIZE) == FALSE &
is.na(DISCOVERY_DATE) == FALSE & is.na(CONT_DATE) == FALSE & is.na(OWNER_DESCR)
== FALSE & is.na(FIRE_SIZE) == FALSE & is.na(FIRE_SIZE_CLASS) == FALSE & is.na(STATE) ==
FALSE & is.na(FIPS_CODE) == FALSE & STAT_CAUSE_DESCR %in% c("Arson", "Campfire",
"Children", "Debris Burning", "Equipment Use", "Fireworks", "Lightning", "Powerline", "Railroad",
"Smoking", "Structure"))
tbl2$time.of.day <- ifelse(tbl2$DISCOVERY_TIME < '0600' | tbl2$DISCOVERY_TIME >= '2100',
'Night'
          , ifelse(tbl2$DISCOVERY_TIME >= '0600' & tbl2$DISCOVERY_TIME < '1200', 'Morning'
                , ifelse(tbl2$DISCOVERY_TIME >= '1200' & tbl2$DISCOVERY_TIME < '1600',
'Afternoon'
                      , 'Evening')))


##############################
# Binning the Duration attribute
##############################
table(tbl2$fire_elapsed_hours)
tbl2$fire.duration <- ifelse(tbl2$fire_elapsed_hours < 0.6, '< 0.025'
                    , ifelse(tbl2$fire_elapsed_hours >= .6 & tbl2$fire_elapsed_hours < 1.2, '0.025-0.05'
                          , ifelse(tbl2$fire_elapsed_hours >= 1.2 & tbl2$fire_elapsed_hours < 3,
'0.05-0.125'
                                , ifelse(tbl2$fire_elapsed_hours >= 3 & tbl2$fire_elapsed_hours < 6,
'0.125-0.25'
                                      , ifelse(tbl2$fire_elapsed_hours >= 6 & tbl2$fire_elapsed_hours < 12,
'0.25-0.5'
                                            , ifelse(tbl2$fire_elapsed_hours >= 12 & tbl2$fire_elapsed_hours
< 24, '0.5-1'
                                                  , ifelse(tbl2$fire_elapsed_hours >= 24 &
tbl2$fire_elapsed_hours < 48, '1'
                                                        , ifelse(tbl2$fire_elapsed_hours >= 48 &
tbl2$fire_elapsed_hours < 72, '2'
                                                              , '3+'))))))))
table(tbl2$fire.duration)
##############################
# Binning the CAUSE attribute
##############################
table(tbl2$STAT_CAUSE_DESCR)
tbl2$cause.binned <- ifelse(tbl2$STAT_CAUSE_DESCR %in% c('Campfire', 'Children', 'Fireworks',
'Smoking', 'Arson'), 'People'
                   , ifelse(tbl2$STAT_CAUSE_DESCR %in% c('Powerline', 'Railroad', 'Structure',
'Equipment Use'), 'Industry'
                         , ifelse(tbl2$STAT_CAUSE_DESCR %in% c('Debris Burning'), 'Debris Burning'
                               , ifelse(tbl2$STAT_CAUSE_DESCR %in% c('Lightning'), 'Lightning'
                                     , 'Misc'))))
table(tbl2$cause.binned)


##############################
# Binning the Sizes attribute
##############################
```

```
table(tbl2$FIRE_SIZE_CLASS)
tbl2$sizes <- factor(tbl2$FIRE_SIZE_CLASS, levels = c("A", "B", "C", "D", "E", "F", "G"))

# % of each fire size in the data
x <- table(tbl2$sizes)
x
(x[1] / sum(x)) * 100
(x[2] / sum(x)) * 100
(x[3] / sum(x)) * 100
(x[4] / sum(x)) * 100
(x[5] / sum(x)) * 100
(x[6] / sum(x)) * 100
(x[7] / sum(x)) * 100

# bin the data to make it more even
((x[1]) / sum(x)) * 100
(x[2] / sum(x)) * 100
((x[3] + x[4] + x[5] + x[6] + x[7]) / sum(x)) * 100

tbl2$sizes.binned <- ifelse(tbl2$sizes == 'A', 1
                     , ifelse(tbl2$sizes == 'B', 2
                            , 3))
table(tbl2$sizes.binned)


tbl2$region <- ifelse(tbl2$STATE %in% c('AK', 'HI', 'WA', 'OR', 'CA'), 'Pacific'
                 , ifelse(tbl2$STATE %in% c('MT', 'ID', 'NV', 'WY', 'UT', 'CO', 'AZ', 'NM'), 'Mountain'
                        , ifelse(tbl2$STATE %in% c('ND','SD','MN','IA','NE', 'KS', 'MO'), 'West North Central'
                               , ifelse(tbl2$STATE %in% c('OK', 'AR', 'TX', 'LA'), 'West South Central'
                                      , ifelse(tbl2$STATE %in% c('WI', 'MI', 'IL', 'IN', 'OH'), 'East North Central'
                                             , ifelse(tbl2$STATE %in% c('KY', 'TN', 'MS', 'AL'), 'East South
Central'
                                                    , ifelse(tbl2$STATE %in% c('NY', 'PA', 'NJ'), 'Middle Atlantic'
                                                           , ifelse(tbl2$STATE %in% c('FL', 'GA', 'SC', 'NC', 'VA',
'WV', 'DC', 'MD', 'DE'), 'South Atlantic'
                                                                  , 'New England'))))))))

tbl2$fire_elapsed_hours <- round(tbl2$fire_elapsed_hours, digits = 2)
head(tbl2$fire_elapsed_hours)


##############################
# Binning the OWNER_DESCR attribute
##############################
table(tbl2$OWNER_DESCR)

tbl2$owner.binned <- ifelse(tbl2$OWNER_DESCR %in% c('PRIVATE'), 'Private'
                     , 'Not Private')
```

```
table(tbl2$owner.binned)


# output data in a text file to be read into Tableau
output.tbl <- tbl2
write.table(output.tbl, "/Users/erinmurphy/Documents/school/IS 467/Final Project/tbl2_fires_data.txt",
sep="\t")

# 3 dependents variables: Fire duration, Cause, Size
# Over sample so that each of these dependent variables are equally distributed

tbl3 <- na.omit(tbl2[, c("cause.binned", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration", "time.of.day", "z.FIRE_YEAR", "z.fire_elapsed_hours", "z.DISCOVERY_DOY",
"z.LATITUDE", "z.LONGITUDE", "STAT_CAUSE_DESCR")])

# final cleaned and processed dataset with categorical variables
final.tbl <- tbl3
final <- tbl3[, c("STAT_CAUSE_DESCR", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration")]

# final cleaned and processed dataset with continuous variables
final.tbl.corr <- na.omit(tbl2[, c("z.FIRE_YEAR", "z.fire_elapsed_hours", "z.DISCOVERY_DOY",
"z.LATITUDE", "z.LONGITUDE")])

# fire duration
table(tbl3$fire.duration)
final.tbl.duration0 <- tbl3[ sample( which( tbl3$fire.duration == "< 0.025" ) , 500 ) , ]
final.tbl.duration0.025 <- tbl3[ sample( which( tbl3$fire.duration == "0.025-0.05" ) , 500 ) , ]
final.tbl.duration0.05 <- tbl3[ sample( which( tbl3$fire.duration == "0.05-0.125" ) , 500 ) , ]
final.tbl.duration0.125 <- tbl3[ sample( which( tbl3$fire.duration == "0.125-0.25" ) , 500 ) , ]
final.tbl.duration0.25 <- tbl3[ sample( which( tbl3$fire.duration == "0.25-0.5" ) , 500 ) , ]
final.tbl.duration0.5 <- tbl3[ sample( which( tbl3$fire.duration == "0.5-1" ) , 500 ) , ]
final.tbl.duration1 <- tbl3[ sample( which( tbl3$fire.duration == "1" ) , 500 ) , ]
final.tbl.duration2 <- tbl3[ sample( which( tbl3$fire.duration == "2" ) , 500 ) , ]
final.tbl.duration3 <- tbl3[ sample( which( tbl3$fire.duration == "3+" ) , 500 ) , ]

final.tbl.duration <- rbind(final.tbl.duration0, final.tbl.duration0.025, final.tbl.duration0.05,
final.tbl.duration0.125, final.tbl.duration0.25, final.tbl.duration0.5, final.tbl.duration1, final.tbl.duration2,
final.tbl.duration3)
remove(final.tbl.duration0, final.tbl.duration0.025, final.tbl.duration0.05, final.tbl.duration0.125,
final.tbl.duration0.25, final.tbl.duration0.5, final.tbl.duration1, final.tbl.duration2, final.tbl.duration3)

# cause
table(tbl3$cause.binned)
final.tbl.cause.debris <- tbl3[ sample( which( tbl3$cause.binned == "Debris Burning" ) , 4500 ) , ]
final.tbl.cause.industry <- tbl3[ sample( which( tbl3$cause.binned == "Industry" ) , 4500 ) , ]
final.tbl.cause.lightening <- tbl3[ sample( which( tbl3$cause.binned == "Lightning" ) , 4500 ) , ]
final.tbl.cause.people <- tbl3[ sample( which( tbl3$cause.binned == "People" ) , 4500 ) , ]

final.tbl.cause <- rbind(final.tbl.cause.debris, final.tbl.cause.industry, final.tbl.cause.lightening,
final.tbl.cause.people)
remove(final.tbl.cause.debris, final.tbl.cause.industry, final.tbl.cause.lightening, final.tbl.cause.people)

# table size
table(tbl3$sizes.binned)
final.tbl.cause.1 <- tbl2[ sample( which( tbl2$sizes.binned == "1" ) , 4000 ) , ]
final.tbl.cause.2 <- tbl2[ sample( which( tbl2$sizes.binned == "2" ) , 4000 ) , ]
final.tbl.cause.3 <- tbl2[ sample( which( tbl2$sizes.binned == "3" ) , 4000 ) , ]

final.tbl.size <- rbind(final.tbl.cause.1, final.tbl.cause.2, final.tbl.cause.3)
remove(final.tbl.cause.1, final.tbl.cause.2, final.tbl.cause.3)

####################################
# Data / Stats Descriptive Analysis
####################################
# Chi square tests for independence
fires <- sample_n(final.tbl.duration, 100)
chisq.test(fires$fire.duration, fires$sizes.binned)
chisq.test(fires$fire.duration, fires$owner.binned)
chisq.test(fires$fire.duration, fires$region)
chisq.test(fires$fire.duration, fires$Season)
chisq.test(fires$fire.duration, fires$cause.binned)

fires <- sample_n(final.tbl.cause, 100)
chisq.test(fires$cause.binned, fires$sizes.binned)
chisq.test(fires$cause.binned, fires$owner.binned)
chisq.test(fires$cause.binned, fires$region)
chisq.test(fires$cause.binned, fires$Season)
chisq.test(fires$cause.binned, fires$fire.duration)

fires <- sample_n(final.tbl.size, 100)
chisq.test(fires$sizes.binned, fires$cause.binned)
chisq.test(fires$sizes.binned, fires$owner.binned)
chisq.test(fires$sizes.binned, fires$region)
chisq.test(fires$sizes.binned, fires$Season)
chisq.test(fires$sizes.binned, fires$fire.duration)


# Scatter plots
fires.corr <- sample_n(final.tbl.corr, 100)
ggplot( data=fires.corr, aes(x=z.DISCOVERY_DOY, y=z.fire_elapsed_hours) ) + geom_point() +
ggtitle("Scatter Plot of Petal Length vs Petal Width") + xlab("Petal Length (cm)") + ylab("Petal Width
(cm)") + theme_bw()
```

```
ggplot( data=fires.corr, aes(x=z.LATITUDE, y=z.fire_elapsed_hours) ) + geom_point() + ggtitle("Scatter
Plot of Petal Length vs Petal Width") + xlab("Petal Length (cm)") + ylab("Petal Width (cm)") +
theme_bw()
ggplot( data=fires.corr, aes(x=z.LONGITUDE, y=z.fire_elapsed_hours) ) + geom_point() +
ggtitle("Scatter Plot of Petal Length vs Petal Width") + xlab("Petal Length (cm)") + ylab("Petal Width
(cm)") + theme_bw()
ggplot( data=fires.corr, aes(x=z.DISCOVERY_DOY, y=z.LATITUDE) ) + geom_point() +
ggtitle("Scatter Plot of Petal Length vs Petal Width") + xlab("Petal Length (cm)") + ylab("Petal Width
(cm)") + theme_bw()
ggplot( data=fires.corr, aes(x=z.DISCOVERY_DOY, y=z.LONGITUDE) ) + geom_point() +
ggtitle("Scatter Plot of Petal Length vs Petal Width") + xlab("Petal Length (cm)") + ylab("Petal Width
(cm)") + theme_bw()
ggplot( data=fires.corr, aes(x=z.LATITUDE, y=z.LONGITUDE) ) + geom_point() + ggtitle("Scatter Plot
of Petal Length vs Petal Width") + xlab("Petal Length (cm)") + ylab("Petal Width (cm)") + theme_bw()



#   Fires by duration
#tbl2$fire_elapsed_hours.ln <- log10(tbl2$fire_elapsed_hours)
ggplot(final.tbl.duration, aes(x = 1, y = z.fire_elapsed_hours)) + geom_boxplot(outlier.size = 3,
outlier.color = "Red", outlier.shape = 1) +
  stat_summary(fun.y = "mean", geom = "point", shape = 21, size = 3) +
  scale_x_continuous(breaks = NULL) +
  theme(axis.title.x = element_blank()) + ggtitle("Boxplot of Fire Duration")

#   Fires by cause
#tbl2$fire_elapsed_hours.ln <- log10(tbl2$fire_elapsed_hours)
ggplot(final.tbl.cause, aes(x = 1, y = STAT_CAUSE_DESCR)) + geom_boxplot(outlier.size = 3,
outlier.color = "Red", outlier.shape = 1) +
  stat_summary(fun.y = "mean", geom = "point", shape = 21, size = 3) +
  scale_x_continuous(breaks = NULL) +
  theme(axis.title.x = element_blank()) + ggtitle("Boxplot of Fire Cause")

# Boxplots
#   Fires by size
#tbl2$FIRE_SIZE.ln <- log10(tbl2$FIRE_SIZE)
ggplot(tbl2, aes(x = 1, y = FIRE_SIZE)) + geom_boxplot(outlier.size = 3, outlier.color = "Red",
outlier.shape = 1) +
  stat_summary(fun.y = "mean", geom = "point", shape = 21, size = 3) +
  scale_x_continuous(breaks = NULL) +
  theme(axis.title.x = element_blank()) + ggtitle("Boxplot of Fire Size")


ggplot(tbl2, aes(x = 1, y = z.DISCOVERY_DOY)) + geom_boxplot(outlier.size = 3, outlier.color = "Red",
outlier.shape = 1) +
  stat_summary(fun.y = "mean", geom = "point", shape = 21, size = 3) +
  scale_x_continuous(breaks = NULL) +
  theme(axis.title.x = element_blank()) + ggtitle("Boxplot of Discovery, Day of Year")
```

```
ggplot(tbl2, aes(x = 1, y = DISCOVERY_TIME)) + geom_boxplot(outlier.size = 3, outlier.color = "Red",
outlier.shape = 1) +
  stat_summary(fun.y = "mean", geom = "point", shape = 21, size = 3) +
  scale_x_continuous(breaks = NULL) +
  theme(axis.title.x = element_blank()) + ggtitle("Boxplot of Discovery, Time of Day")


# Histograms
#   Fires by duration
ggplot(tbl2, aes(x = fire_elapsed_hours)) + geom_histogram(binwidth = 75, position = "identity", alpha =
1) + ggtitle("Fires, Duration") + theme_bw() + xlim(0, 7000) + ylim(0, 300)
ggplot(tbl2, aes(x = fire.duration)) + geom_bar() + ggtitle("Fires, Duration") + theme_bw()
# Oversampled
table(final.tbl.duration$fire.duration)

#   Fires by causes
ggplot(tbl2, aes(x = STAT_CAUSE_DESCR)) + geom_bar() + ggtitle("Fires, Causes") + theme_bw()
ggplot(tbl2, aes(x = cause.binned)) + geom_bar() + ggtitle("Fires, Causes") + theme_bw()
# Oversampled
table(final.tbl.cause$cause.binned)

#   Fires by size
ggplot(tbl2, aes(x = FIRE_SIZE)) + geom_histogram(binwidth = 5, position = "identity", alpha = 0.5) +
ggtitle("Fires, Size") + theme_bw() + xlim(0, 110) + ylim(0, 1000)
ggplot(data.frame(tbl2), aes(x=sizes)) + geom_bar()
ggplot(data.frame(tbl2), aes(x=sizes.binned)) + geom_bar()
# Oversampled
table(final.tbl.size$sizes.binned)

#   # of fires in 2015 (heatmap?)
ggplot(tbl2, aes(x = DISCOVERY_DOY)) + geom_histogram(binwidth = 25, position = "identity", alpha
= 0.5) + ggtitle("Fires, DOY") + theme_bw()
ggplot(tbl2, aes(x = DISCOVERY_TIME)) + geom_histogram(binwidth = 1, position = "identity", alpha
= 1) + ggtitle("Fires, Time of Day") + theme_bw()
```

Cluster Analysis Code

```
set.seed(444)
fires <- final
fires <- na.omit(fires)
require(data.table)
fires <- data.table(fires)
#fires <- fires[ , .SD[sample(1:.N,min(100,.N))], by=STAT_CAUSE_DESCR]
fires <- sample_n(fires, 1000)
```

```
head(fires)
str(fires)
fires$sizes.binned <- as.factor(fires$sizes.binned)
fires$owner.binned <- as.factor(fires$owner.binned)
fires$region.bin <- as.factor(fires$region.bin)
fires$fire.duration <- as.factor(fires$fire.duration)
fires$cause.binned <- as.factor(fires$cause.binned)
fires$time.of.day <- as.factor(fires$time.of.day)
fires_nostate <- fires[, 1:8]
#Initialize subset of data
#determine best method to use
#m <- c( "average", "single", "complete", "ward")
#names(m) <- c( "average", "single", "complete", "ward")
#ac <- function(x) {
#  agnes(fires, method = x)$ac
#}
#map_dbl(m, ac)
#Ward gave the highest ac
agn1 <- agnes(fires_nostate, metric = "manhattan", diss=FALSE,stand = FALSE, method = "ward")
#Cut tree into clusters
groups <- cutree(agn1, k=3)
plot(agn1)
rect.hclust(agn1, k=3, border="red")
clusplot(fires_nostate, groups, main='Wildfire segments',color=TRUE, shade=TRUE,labels=2, lines=3)
groups <- cutree(agn1, k=3)
#plot(agn1)
#rect.hclust(agn1, k=3, border="red")
head(fires)
#install.packages('PerformanceAnalytics')
#library(PerformanceAnalytics)
fires.clust <- transform(fires, cluster=groups)

c <- c("sizes.binned", "owner.binned", "z.fire_elapsed_hours", "time.of.day","STAT_CAUSE_DESCR",
"z.DISCOVERY_DOY", "z.LATITUDE", "z.LONGITUDE", "STATE", "cause.binned",
"Season","region.bin")

print("Cluster 1")
fires.clust1 <- subset(fires.clust, groups == 1,c)
summary(fires.clust1)

print("Cluster 2")
fires.clust2 <- subset(fires.clust, groups == 2,c)
summary(fires.clust2)

print("Cluster 3")
fires.clust3 <- subset(fires.clust, groups == 3,c)
summary(fires.clust3)
```

```
#Cause graphs
ggplot(fires.clust1, aes(STAT_CAUSE_DESCR)) +
  geom_bar(fill = "#0073C2FF") + xlab("Cause") + ylab("Count") + ggtitle("Cluster 1")

ggplot(fires.clust2, aes(STAT_CAUSE_DESCR)) +
  geom_bar(fill = "#0073C2FF") + xlab("Cause") + ylab("Count") + ggtitle("Cluster 2")

ggplot(fires.clust3, aes(STAT_CAUSE_DESCR)) +
  geom_bar(fill = "#0073C2FF") + xlab("Cause") + ylab("Count") + ggtitle("Cluster 3")

#Region graphs
ggplot(fires.clust1, aes(region.bin)) +
  geom_bar(fill = "#0073C2FF") + xlab("Region") + ylab("Count") + ggtitle("Cluster 1") + coord_flip()

ggplot(fires.clust2, aes(region.bin)) +
  geom_bar(fill = "#0073C2FF") + xlab("Region") + ylab("Count") + ggtitle("Cluster 2") + coord_flip()

ggplot(fires.clust3, aes(region.bin)) +
  geom_bar(fill = "#0073C2FF") + xlab("Region") + ylab("Count") + ggtitle("Cluster 3") + coord_flip()

state.abb <- append(state.abb, c("DC", "PR"))
state.name <- append(state.name, c("District of Columbia", "Puerto Rico"))
fires$region <- map_chr(fires$STATE, function(x) { tolower(state.name[grep(x, state.abb)]) })
fires.clust1$region <- map_chr(fires.clust1$STATE, function(x) { tolower(state.name[grep(x, state.abb)])
})
fires.clust2$region <- map_chr(fires.clust2$STATE, function(x) { tolower(state.name[grep(x, state.abb)])
})
fires.clust3$region <- map_chr(fires.clust3$STATE, function(x) { tolower(state.name[grep(x, state.abb)])
})
state_map <- map_data('state')
head(fires)

#fires$region
fires %>%
    select(region) %>%
    group_by(region) %>%
    summarize(n = n()) %>%
    right_join(state_map, by = 'region') %>%
    ggplot(aes(x = long, y = lat, group = group, fill = n)) +
    geom_polygon() +
    geom_path(color = 'white') +
    scale_fill_continuous(low = "orange",
                          high = "darkred",
                          name = 'Number of fires') +
    theme_map() +
    coord_map('albers', lat0=30, lat1=40) +
    ggtitle("US Wildfires, 2015 (n = 4000)") +
    theme(plot.title = element_text(hjust = 0.5))
```

```
fires.clust1 %>%
    select(region) %>%
    group_by(region) %>%
    summarize(n = n()) %>%
    right_join(state_map, by = 'region') %>%
    ggplot(aes(x = long, y = lat, group = group, fill = n)) +
    geom_polygon() +
    geom_path(color = 'white') +
    scale_fill_continuous(low = "orange",
                            high = "darkred",
                            name = 'Number of fires') +
    theme_map() +
    coord_map('albers', lat0=30, lat1=40) +
    ggtitle("US Wildfires, 2015 (Cluster 1)") +
    theme(plot.title = element_text(hjust = 0.5))

fires.clust2 %>%
    select(region) %>%
    group_by(region) %>%
    summarize(n = n()) %>%
    right_join(state_map, by = 'region') %>%
    ggplot(aes(x = long, y = lat, group = group, fill = n)) +
    geom_polygon() +
    geom_path(color = 'white') +
    scale_fill_continuous(low = "orange",
                            high = "darkred",
                            name = 'Number of fires') +
    theme_map() +
    coord_map('albers', lat0=30, lat1=40) +
    ggtitle("US Wildfires, 2015 (Cluster 2)") +
    theme(plot.title = element_text(hjust = 0.5))

fires.clust3 %>%
    select(region) %>%
    group_by(region) %>%
    summarize(n = n()) %>%
    right_join(state_map, by = 'region') %>%
    ggplot(aes(x = long, y = lat, group = group, fill = n)) +
    geom_polygon() +
    geom_path(color = 'white') +
    scale_fill_continuous(low = "orange",
                            high = "darkred",
                            name = 'Number of fires') +
    theme_map() +
    coord_map('albers', lat0=30, lat1=40) +
    ggtitle("US Wildfires, 2015 (Cluster 3)") +
    theme(plot.title = element_text(hjust = 0.5))
```

Decision Tree Code

```
library(rpart)
library(rpart.plot)
library(dplyr)
library(e1071)
library(caret)
library(MASS)

tbl1 <- read.csv(file="final_fires_data.csv", header=FALSE, sep=",")
names(tbl1) <- c("OBJECTID", "FOD_ID", "FPA_ID", "SOURCE_SYSTEM_TYPE",
"SOURCE_SYSTEM", "NWCG_REPORTING_AGENCY", "NWCG_REPORTING_UNIT_ID",
"NWCG_REPORTING_UNIT_NAME", "SOURCE_REPORTING_UNIT",
"SOURCE_REPORTING_UNIT_NAME", "LOCAL_FIRE_REPORT_ID", "LOCAL_INCIDENT_ID",
"FIRE_CODE", "FIRE_NAME", "ICS_209_INCIDENT_NUMBER", "ICS_209_NAME", "MTBS_ID",
"MTBS_FIRE_NAME", "COMPLEX_NAME", "FIRE_YEAR", "DISCOVERY_DATE",
"DISCOVERY_DOY", "DISCOVERY_TIME", "STAT_CAUSE_CODE", "STAT_CAUSE_DESCR",
"CONT_DATE", "CONT_DOY", "CONT_TIME", "FIRE_SIZE", "FIRE_SIZE_CLASS", "LATITUDE",
"LONGITUDE", "OWNER_COD", "OWNER_DESCR", "STATE", "COUNTY", "FIPS_CODE",
"FIPS_NAME", "Shape", "Season", "discovery_datetime", "containment_datetime","fire_elapsed_hours")

#################################################
# Pre-Processing, Cleansing and Transformation
#################################################
tbl1$z.FIRE_YEAR <- scale(x = tbl1$FIRE_YEAR, center = TRUE, scale = TRUE)
tbl1$z.FIRE_SIZE <- scale(x = tbl1$FIRE_SIZE, center = TRUE, scale = TRUE)
tbl1$z.fire_elapsed_hours <- scale(x = tbl1$fire_elapsed_hours, center = TRUE, scale = TRUE)
tbl1$z.DISCOVERY_DOY <- scale(x = tbl1$DISCOVERY_DOY, center = TRUE, scale = TRUE)
tbl1$z.DISCOVERY_TIME <- scale(x = tbl1$DISCOVERY_TIME, center = TRUE, scale = TRUE)
tbl1$z.LATITUDE <- scale(x = tbl1$LATITUDE, center = TRUE, scale = TRUE)
tbl1$z.LONGITUDE <- scale(x = tbl1$LONGITUDE, center = TRUE, scale = TRUE)

# remove NA values and causes = "Miscellaneous", "Missing/Undefined"
tbl2 <- subset(tbl1, FIRE_YEAR == 2015 & is.na(FIRE_SIZE) == FALSE &
is.na(DISCOVERY_DATE) == FALSE & is.na(CONT_DATE) == FALSE & is.na(OWNER_DESCR)
== FALSE & is.na(FIRE_SIZE) == FALSE & is.na(FIRE_SIZE_CLASS) == FALSE & is.na(STATE) ==
FALSE & is.na(FIPS_CODE) == FALSE & STAT_CAUSE_DESCR %in% c("Arson", "Campfire",
"Children", "Debris Burning", "Equipment Use", "Fireworks", "Lightning", "Powerline", "Railroad",
"Smoking", "Structure"))
tbl2$time.of.day <- ifelse(tbl2$DISCOVERY_TIME < '0600' | tbl2$DISCOVERY_TIME >= '2100',
'Night'
            , ifelse(tbl2$DISCOVERY_TIME >= '0600' & tbl2$DISCOVERY_TIME < '1200', 'Morning'
                  , ifelse(tbl2$DISCOVERY_TIME >= '1200' & tbl2$DISCOVERY_TIME < '1600',
'Afternoon'
                         , 'Evening')))
```

```
##############################
# Binning the Duration attribute
##############################
table(tbl2$fire_elapsed_hours)
tbl2$fire.duration <- ifelse(tbl2$fire_elapsed_hours < 0.6, '< 0.025'
                    , ifelse(tbl2$fire_elapsed_hours >= .6 & tbl2$fire_elapsed_hours < 1.2, '0.025-0.05'
                          , ifelse(tbl2$fire_elapsed_hours >= 1.2 & tbl2$fire_elapsed_hours < 3,
'0.05-0.125'
                                , ifelse(tbl2$fire_elapsed_hours >= 3 & tbl2$fire_elapsed_hours < 6,
'0.125-0.25'
                                      , ifelse(tbl2$fire_elapsed_hours >= 6 & tbl2$fire_elapsed_hours < 12,
'0.25-0.5'
                                            , ifelse(tbl2$fire_elapsed_hours >= 12 & tbl2$fire_elapsed_hours
< 24, '0.5-1'
                                                  , ifelse(tbl2$fire_elapsed_hours >= 24 &
tbl2$fire_elapsed_hours < 48, '1'
                                                        , ifelse(tbl2$fire_elapsed_hours >= 48 &
tbl2$fire_elapsed_hours < 72, '2'
                                                              , '3+'))))))))
table(tbl2$fire.duration)
##############################
# Binning the CAUSE attribute
##############################
table(tbl2$STAT_CAUSE_DESCR)
tbl2$cause.binned <- ifelse(tbl2$STAT_CAUSE_DESCR %in% c('Campfire', 'Children', 'Fireworks',
'Smoking', 'Arson'), 'People'
                   , ifelse(tbl2$STAT_CAUSE_DESCR %in% c('Powerline', 'Railroad', 'Structure',
'Equipment Use'), 'Industry'
                         , ifelse(tbl2$STAT_CAUSE_DESCR %in% c('Debris Burning'), 'Debris Burning'
                               , ifelse(tbl2$STAT_CAUSE_DESCR %in% c('Lightning'), 'Lightning'
                                     , 'Misc'))))
table(tbl2$cause.binned)


##############################
# Binning the Sizes attribute
##############################
table(tbl2$FIRE_SIZE_CLASS)
tbl2$sizes <- factor(tbl2$FIRE_SIZE_CLASS, levels = c("A", "B", "C", "D", "E", "F", "G"))

# % of each fire size in the data
x <- table(tbl2$sizes)
x
(x[1] / sum(x)) * 100
(x[2] / sum(x)) * 100
(x[3] / sum(x)) * 100
(x[4] / sum(x)) * 100
```

```
(x[5] / sum(x)) * 100
(x[6] / sum(x)) * 100
(x[7] / sum(x)) * 100

# bin the data to make it more even
((x[1]) / sum(x)) * 100
(x[2] / sum(x)) * 100
((x[3] + x[4] + x[5] + x[6] + x[7]) / sum(x)) * 100

tbl2$sizes.binned <- ifelse(tbl2$sizes == 'A', 1
               , ifelse(tbl2$sizes == 'B', 2
                    , 3))
table(tbl2$sizes.binned)


tbl2$region <- ifelse(tbl2$STATE %in% c('AK', 'HI', 'WA', 'OR', 'CA'), 'Pacific'
            , ifelse(tbl2$STATE %in% c('MT', 'ID', 'NV', 'WY', 'UT', 'CO', 'AZ', 'NM'), 'Mountain'
                 , ifelse(tbl2$STATE %in% c('ND','SD','MN','IA','NE', 'KS', 'MO'), 'West North Central'
                      , ifelse(tbl2$STATE %in% c('OK', 'AR', 'TX', 'LA'), 'West South Central'
                           , ifelse(tbl2$STATE %in% c('WI', 'MI', 'IL', 'IN', 'OH'), 'East North Central'
                                , ifelse(tbl2$STATE %in% c('KY', 'TN', 'MS', 'AL'), 'East South
Central'
                                     , ifelse(tbl2$STATE %in% c('NY', 'PA', 'NJ'), 'Middle Atlantic'
                                          , ifelse(tbl2$STATE %in% c('FL', 'GA', 'SC', 'NC', 'VA',
'WV', 'DC', 'MD', 'DE'), 'South Atlantic'
                                               , 'New England'))))))))

tbl2$fire_elapsed_hours <- round(tbl2$fire_elapsed_hours, digits = 2)
head(tbl2$fire_elapsed_hours)


##############################
# Binning the OWNER_DESCR attribute
##############################
table(tbl2$OWNER_DESCR)

tbl2$owner.binned <- ifelse(tbl2$OWNER_DESCR %in% c('PRIVATE'), 'Private'
               , 'Not Private')

table(tbl2$owner.binned)


# output data in a text file to be read into Tableau
output.tbl <- tbl2
write.table(output.tbl, "tbl2_fires_data.txt", sep="\t")

# 3 dependents variables: Fire duration, Cause, Size
# Over sample so that each of these dependent variables are equally distributed
```

```
tbl3 <- na.omit(tbl2[, c("cause.binned", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration", "time.of.day", "z.FIRE_YEAR", "z.fire_elapsed_hours", "z.DISCOVERY_DOY",
"z.LATITUDE", "z.LONGITUDE", "STAT_CAUSE_DESCR")])

# final cleaned and processed dataset with categorical variables
final.tbl <- tbl3
final <- tbl3[, c("STAT_CAUSE_DESCR", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration")]

# final cleaned and processed dataset with continuous variables
final.tbl.corr <- na.omit(tbl2[, c("z.FIRE_YEAR", "z.fire_elapsed_hours", "z.DISCOVERY_DOY",
"z.LATITUDE", "z.LONGITUDE")])

# fire duration (undersampling the population so you get an equal distribution in the bins)
table(tbl3$fire.duration)
final.tbl.duration0 <- tbl3[ sample( which( tbl3$fire.duration == "< 0.025" ) , 500 ) , ]
final.tbl.duration0.025 <- tbl3[ sample( which( tbl3$fire.duration == "0.025-0.05" ) , 500 ) , ]
final.tbl.duration0.05 <- tbl3[ sample( which( tbl3$fire.duration == "0.05-0.125" ) , 500 ) , ]
final.tbl.duration0.125 <- tbl3[ sample( which( tbl3$fire.duration == "0.125-0.25" ) , 500 ) , ]
final.tbl.duration0.25 <- tbl3[ sample( which( tbl3$fire.duration == "0.25-0.5" ) , 500 ) , ]
final.tbl.duration0.5 <- tbl3[ sample( which( tbl3$fire.duration == "0.5-1" ) , 500 ) , ]
final.tbl.duration1 <- tbl3[ sample( which( tbl3$fire.duration == "1" ) , 500 ) , ]
final.tbl.duration2 <- tbl3[ sample( which( tbl3$fire.duration == "2" ) , 500 ) , ]
final.tbl.duration3 <- tbl3[ sample( which( tbl3$fire.duration == "3+" ) , 500 ) , ]

final.tbl.duration <- rbind(final.tbl.duration0, final.tbl.duration0.025, final.tbl.duration0.05,
final.tbl.duration0.125, final.tbl.duration0.25, final.tbl.duration0.5, final.tbl.duration1, final.tbl.duration2,
final.tbl.duration3)
remove(final.tbl.duration0, final.tbl.duration0.025, final.tbl.duration0.05, final.tbl.duration0.125,
final.tbl.duration0.25, final.tbl.duration0.5, final.tbl.duration1, final.tbl.duration2, final.tbl.duration3)

# cause
table(tbl3$cause.binned)
final.tbl.cause.debris <- tbl3[ sample( which( tbl3$cause.binned == "Debris Burning" ) , 4500 ) , ]
final.tbl.cause.industry <- tbl3[ sample( which( tbl3$cause.binned == "Industry" ) , 4500 ) , ]
final.tbl.cause.lightening <- tbl3[ sample( which( tbl3$cause.binned == "Lightning" ) , 4500 ) , ]
final.tbl.cause.people <- tbl3[ sample( which( tbl3$cause.binned == "People" ) , 4500 ) , ]

final.tbl.cause <- rbind(final.tbl.cause.debris, final.tbl.cause.industry, final.tbl.cause.lightening,
final.tbl.cause.people)
remove(final.tbl.cause.debris, final.tbl.cause.industry, final.tbl.cause.lightening, final.tbl.cause.people)

# table size
table(tbl3$sizes.binned)
final.tbl.cause.1 <- tbl2[ sample( which( tbl2$sizes.binned == "1" ) , 4000 ) , ]
final.tbl.cause.2 <- tbl2[ sample( which( tbl2$sizes.binned == "2" ) , 4000 ) , ]
final.tbl.cause.3 <- tbl2[ sample( which( tbl2$sizes.binned == "3" ) , 4000 ) , ]
final.tbl.size <- rbind(final.tbl.cause.1, final.tbl.cause.2, final.tbl.cause.3)
remove(final.tbl.cause.1, final.tbl.cause.2, final.tbl.cause.3)

############################
# Decision Tree - Duration
############################
final.tbl <- final.tbl.duration[, c("cause.binned", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration")]


# Random sample (60% of data in train, 40% in test)
smp_size <- floor(0.6 * nrow(final.tbl))

## set the seed to make your partition reproducible
set.seed(200)
train_ind1 <- sample(seq_len(nrow(final.tbl)), size = smp_size)

train1 <- final.tbl[train_ind1, ]
test1 <- final.tbl[-train_ind1, ]

# ensure each bin is present
table(train1$fire.duration)


# grow tree

term.obs = 5 #minimum number of observations before attempting to split.
cp = 0 #complexity parameter. Any split that does not decrease the overall lack of fit by a factor of cp is
not attempted. Start with 0 to get the biggest tree, then use printcp() to determine the best cp (lowest
xerror / penalty factor).


fit1 <- rpart(fire.duration ~ ., data = train1, method="class", control = rpart.control(minbucket = term.obs,
cp = cp ))
plot(fit1, uniform=TRUE, main="Classification Tree for Fire Duration")
text(fit1, use.n=TRUE, all=TRUE, cex=.8)


# The knee / elbow occurrs at 8
plotcp(fit1, main="Elbow Graph of Fire Duration")
printcp(fit1)
newCP <- fit1$cptable[ which(fit1$cptable[,"nsplit"]==8) ]
newCP

# Pruning the tree
# cp = smallest cross-validated error, choose the cp with the smallest xerror from printcp()
prune.fit1 <- prune(fit1, cp = newCP)
```

```
plot(prune.fit1, uniform = TRUE, main = "Pruned Classification Tree for Fire Duration")
text(prune.fit1, use.n = TRUE, all = TRUE, cex = .8)
rpart.plot(prune.fit1, type=4, extra=102, main= "Pruned Classification Tree for Fire Duration")

# create a confusion matrix from the training data
predict.train1 <- predict(prune.fit1, type = "class")
table(predict.train1)

#confusionMatrix(actual, predicted)
confusionMatrix(factor(train1$fire.duration), predict.train1)

# create a confusion matrix from the test data
predict.test1 <- predict(object = prune.fit1, test1, type="class")
table(predict.test1)

confusionMatrix(factor(test1$fire.duration), predict.test1)
x <- confusionMatrix(factor(test1$fire.duration), predict.test1)
x$overall[1]
1 - x$overall[1]

# create a confusion matrix from the initial data
final.tbl.to.predict <- tbl3[, c("cause.binned", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration")]
predict.all <- predict(object = prune.fit1, final.tbl.to.predict, type="class")
table(predict.all)

confusionMatrix(factor(tbl3$fire.duration), predict.all)
x <- confusionMatrix(factor(tbl3$fire.duration), predict.all)
x$overall[1]
1 - x$overall[1]


###########################
# Decision Tree - Causes
###########################
final.tbl <- final.tbl.cause[, c("cause.binned", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration")]



# Random sample (60% of data in train, 40% in test)
smp_size <- floor(0.6 * nrow(final.tbl))

## set the seed to make your partition reproducible
set.seed(200)
train_ind1 <- sample(seq_len(nrow(final.tbl)), size = smp_size)
```

```
train1 <- final.tbl[train_ind1, ]
test1 <- final.tbl[-train_ind1, ]


table(train1$cause.binned)

# grow tree

term.obs = 5 #minimum number of observations before attempting to split.
cp = 0 #complexity parameter. Any split that does not decrease the overall lack of fit by a factor of cp is not attempted. Start with 0 to get the biggest tree, then use printcp() to determine the best cp (lowest xerror / penalty factor).


fit1 <- rpart(cause.binned ~ ., data = train1, method="class", control = rpart.control(minbucket = term.obs, cp = cp ))

plot(fit1, uniform=TRUE, main="Classification Tree for Cause")
text(fit1, use.n=TRUE, all=TRUE, cex=.8)

# The knee / elbow occurrs at 91 (created a more accurate model)
plotcp(fit1, main="Elbow Graph of Cause")
printcp(fit1 )
newCP <- fit1$cptable[ which(fit1$cptable[,"nsplit"]==91) ]
newCP

# Pruning the tree
# cp = smallest cross-validated error, choose the cp with the smallest xerror from printcp()
prune.fit1 <- prune(fit1, cp = newCP)

plot(prune.fit1, uniform = TRUE, main = "Pruned Classification Tree for Cause")
text(prune.fit1, use.n = TRUE, all = TRUE, cex = .8)
rpart.plot(prune.fit1, type=4, extra=102, main = "Pruned Classification Tree for Cause")

# create a confusion matrix from the training data
predict.train1 <- predict(prune.fit1, type = "class")
table(predict.train1)

#confusionMatrix(actual, predicted)
confusionMatrix(factor(train1$cause.binned), predict.train1)

# create a confusion matrix from the test data
predict.test1 <- predict(object = prune.fit1, test1, type="class")
table(predict.test1)

confusionMatrix(factor(test1$cause.binned), predict.test1)
x <- confusionMatrix(factor(test1$cause.binned), predict.test1)
x$overall[1]
```

```
1 - x$overall[1]

# create a confusion matrix from the initial data
final.tbl.to.predict <- tbl3[, c("cause.binned", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration")]
predict.all <- predict(object = prune.fit1, final.tbl.to.predict, type="class")
table(predict.all)

confusionMatrix(factor(tbl3$cause.binned), predict.all)
x <- confusionMatrix(factor(tbl3$cause.binned), predict.all)
x$overall[1]
1 - x$overall[1]


###########################
# Decision Tree - Sizes
###########################
final.tbl <- final.tbl.size[, c("cause.binned", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration")]


# Random sample (60% of data in train, 40% in test)
smp_size <- floor(0.6 * nrow(final.tbl))

## set the seed to make your partition reproducible
set.seed(200)
train_ind1 <- sample(seq_len(nrow(final.tbl)), size = smp_size)

train1 <- final.tbl[train_ind1, ]
test1 <- final.tbl[-train_ind1, ]

# ensure each bin is present
table(train1$sizes.binned)


# grow tree
term.obs = 5 #minimum number of observations before attempting to split.
cp = 0 #complexity parameter. Any split that does not decrease the overall lack of fit by a factor of cp is
not attempted. Start with 0 to get the biggest tree, then use printcp() to determine the best cp (lowest
xerror / penalty factor).


fit1 <- rpart(sizes.binned ~ ., data = train1, method="class", control = rpart.control(minbucket = term.obs,
cp = cp ))
#, cp = cp
# , maxdepth = max.depth
plot(fit1, uniform=TRUE, main="Classification Tree for Sizes")
text(fit1, use.n=TRUE, all=TRUE, cex=.8)
```

```
# The knee / elbow occurrs at 11
plotcp(fit1, main =  "Elbow Graph of Sizes ")
printcp(fit1)
newCP <- fit1$cptable[ which(fit1$cptable[,"nsplit"]==11) ]
newCP

# Pruning the tree

prune.fit1 <- prune(fit1, cp = newCP)

plot(prune.fit1, uniform = TRUE, main = "Pruned Classification Tree for Sizes")
text(prune.fit1, use.n = TRUE, all = TRUE, cex = .8)
rpart.plot(prune.fit1, type=4, extra=102,  main = "Pruned Classification Tree for Sizes")

# create a confusion matrix from the training data
predict.train1 <- predict(prune.fit1, type = "class")
table(predict.train1)

#confusionMatrix(actual, predicted)
confusionMatrix(factor(train1$sizes.binned), predict.train1)

# create a confusion matrix from the test data
predict.test1 <- predict(object = prune.fit1, test1, type="class")
table(predict.test1)

confusionMatrix(factor(test1$sizes.binned), predict.test1)
x <- confusionMatrix(factor(test1$sizes.binned), predict.test1)
x$overall[1]
1 - x$overall[1]

# create a confusion matrix from the initial data
final.tbl.to.predict <- tbl3[, c("cause.binned", "sizes.binned", "owner.binned", "region", "Season",
"fire.duration")]
predict.all <- predict(object = prune.fit1, final.tbl.to.predict, type="class")
table(predict.all)

confusionMatrix(factor(tbl3$sizes.binned), predict.all)
x <- confusionMatrix(factor(tbl3$sizes.binned), predict.all)
x$overall[1]
1 - x$overall[1]
```