This is an example exploratory analysis report intended to give you a sense of the scope of Assignment 2. Although this report targeted an earlier version of the assignment (with a slightly different set of instructions), it is nevertheless an example of ***exemplary work*** as it starts by identifying issues with the data quality, groups data into semantically-meaningful bins such as “New England” and the “Pacific” to better understand spatial distribution, explores a diverse range of visualization designs (including a particularly creative windmap!), and includes annotations of reference/trend lines and richly detailed captions. *Note:* “exemplary work” does not mean that this report is *perfect*—there are some mistakes made (for instance, with the design of some visualizations); nevertheless, it is exemplary from the point of view of the goals of the assignment which is to engage in systematic and rigorous exploratory data analysis and, thus, **earned a full 100% grade**.

Assignment 2 Example

Yao Zhao, MIT 6.894, Spring 2019

Dataset

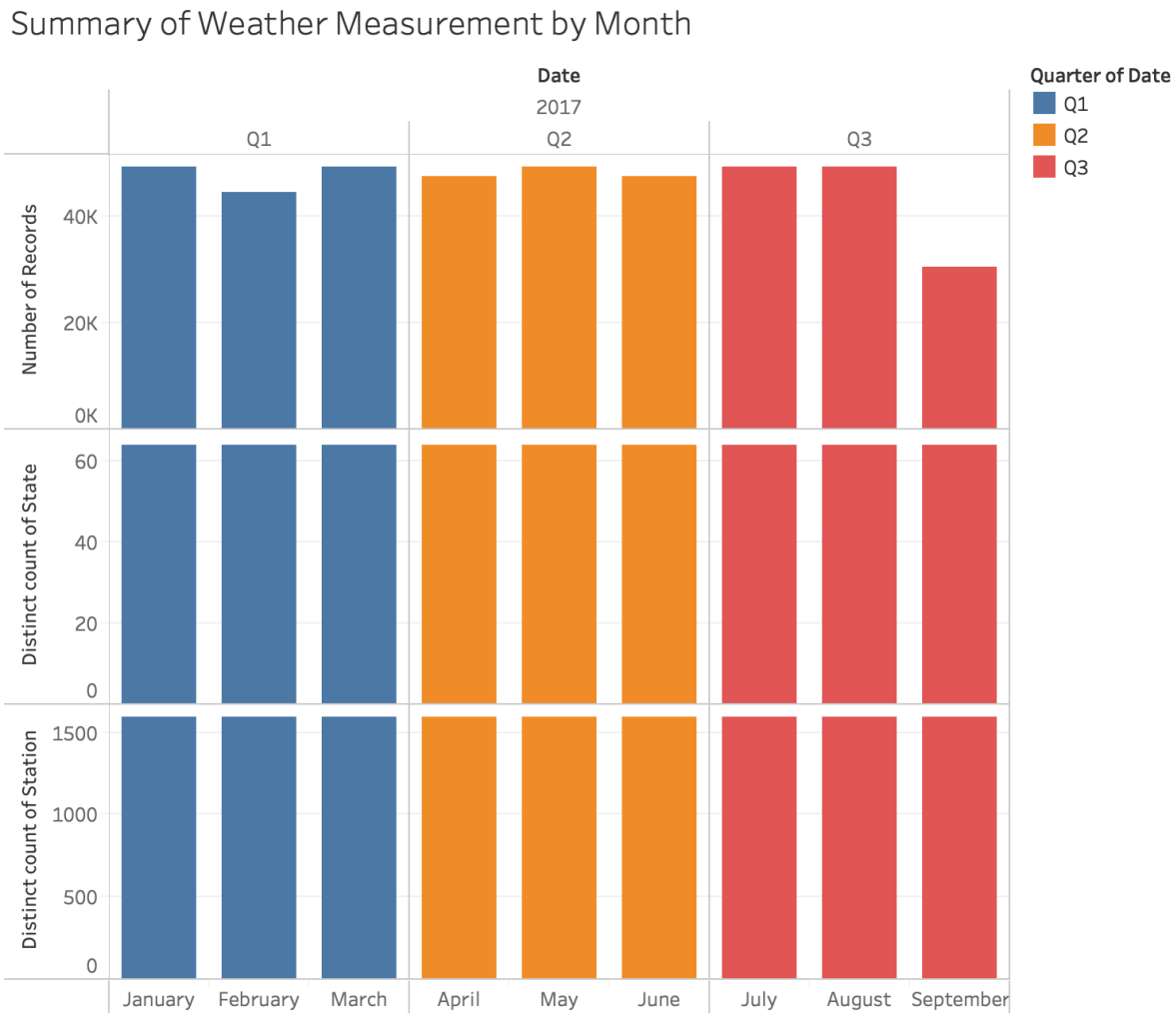
This dataset contains statistics for a sample of 416,937 U.S. daily weather measurements in 2017, which were recorded by different weather stations across the country. It basically contains two sets of information: (1) the geography of weather stations, such as its state, latitude, longitude and elevation; (2) measurements, including precipitation, snowfall, temperature, wind speed, etc (see here for more detailed description). This dataset is provided by NOAA Daily Global Historical Climatology Network. Some data transformation has been done: weather stations with only sparse measurements have been filtered out. I'm interested in this data because as a spatial-temporal dataset it not only has a large size but shows great resolution in terms of both time and space, which enables us to explore detailed patterns of the U.S. climates and make comparisons.

Initial Analysis Questions

1. What is the spatial distribution pattern of the U.S. weathers?
2. Which part of the U.S. has the biggest variation/seasonality in temperature over time?
3. What factors contribute to larger variation in temperature?
4. Where are the extreme weathers in the U.S.? What factors are associated with snowfall?

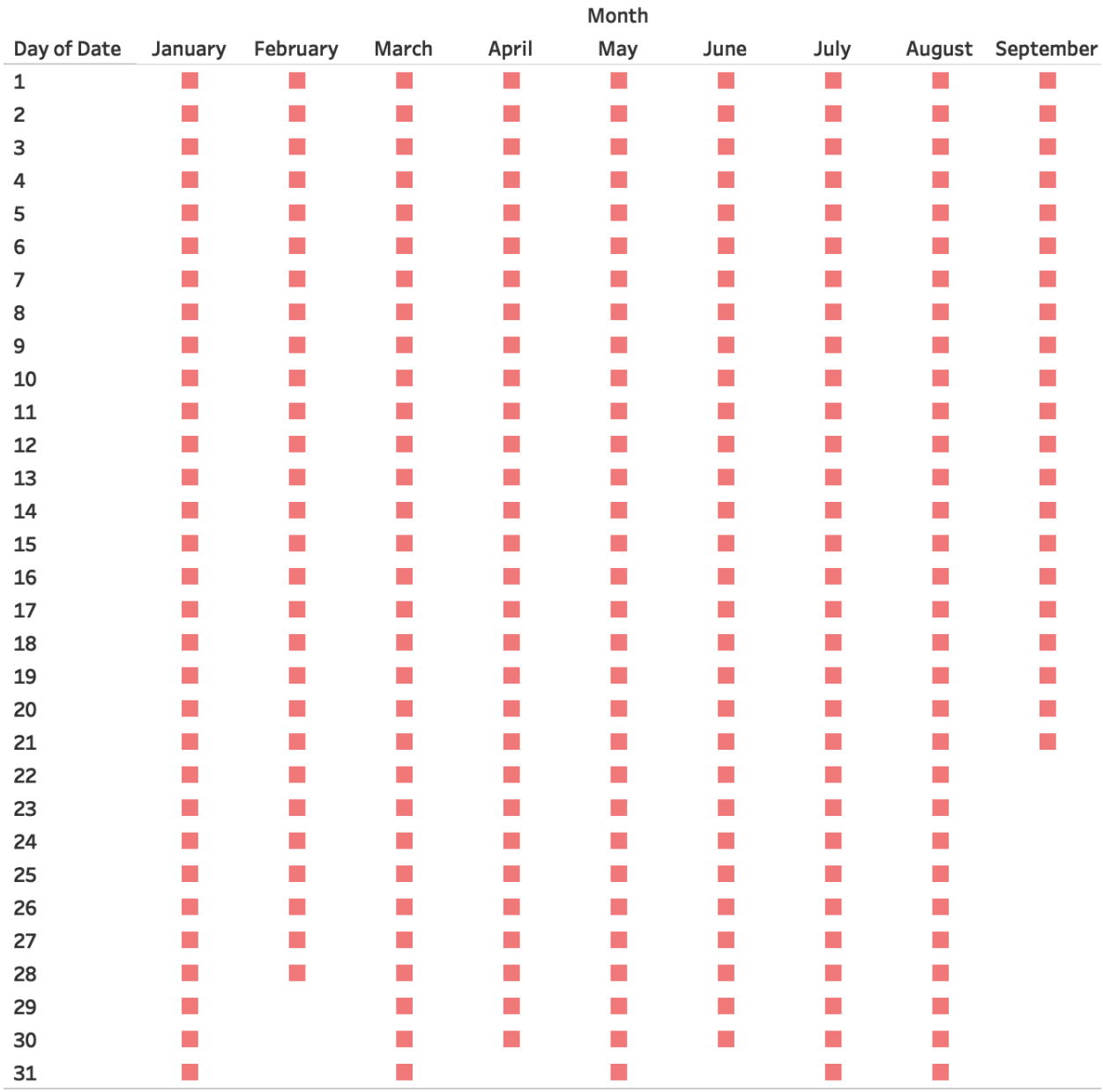
Discoveries & Insights

To assess the data quality and resolution, we first explore the spatial and temporal coverage of this dataset by plotting some basic statistics of weather measurements against time and space. Then we dive into the individual measurements and their correlations to answer our analysis questions.



This bar chart shows the number of weather records, the distinct count of states, and the distinct count of weather stations covered in each month. As we can see, this dataset covers only the first nine months of 2017, namely January to September, rather than the whole year as we expect. The weather records are almost evenly distributed over the first eight months, while September witnesses significantly fewer records. The second and third rows indicate that the spatial coverage is well balanced among all nine months, since they each cover almost the same number of states and weather stations. However, there're over 60 distinct states in this dataset, which are more than the total number of states in the U.S. Hence, two further questions regarding data distribution arise: (1) How come does Sptember have much fewer records? (2) What's the spatial coverage of this dataset?

Time Coverage

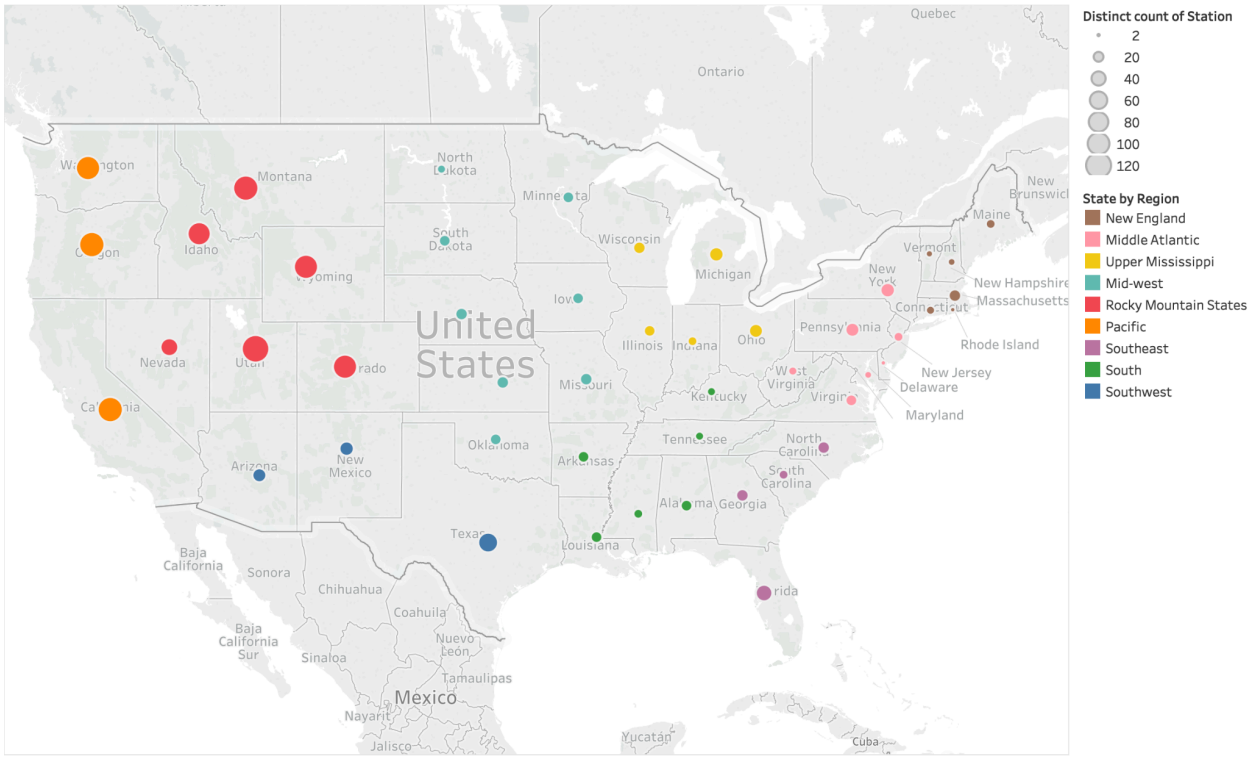


We explore the puzzle of September's fewer records with the help of a Gantt chart, where we mark a red square (i,j) if the i-th day of month j is recorded in our dataset. Clearly, the dataset fails to cover the whole September. Only the first 21 days are recorded. This implies the September weather data may not be representative for the whole month. But we don't want to drop Semptember, since we will not specifically conduct monthly comparisons in this analysis. However, for the quarterly statistics shown below, Q3 always end at Seps 21. Now we have a clearer mind about the temporal coverage of this dataset: it covers each day from Jan 01, 2017 to Sep 21, 2017, and the spatial coverage of each month is quite similar.

Summary of Weather Measurements by Space

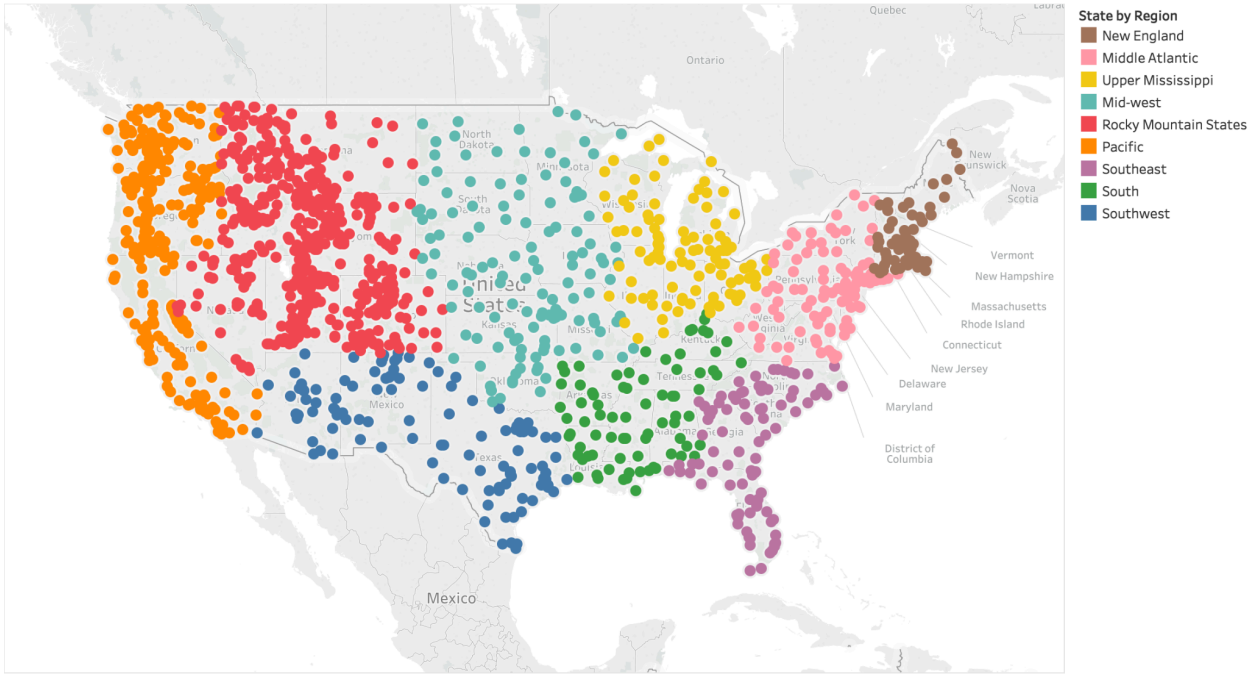
This map explores the puzzle of state numbers. As we can clearly see, this dataset includes not only the states of the U.S., but also those of Canada. Given that we intend to analyze the U.S. climate pattern, we select only the 48 states in the U.S. mainland, as marked in red. The size of each circle represents the number of distinct weather stations in the corresponding state. At a first glance, it seems that the states in the west have more weather stations. But that doesn't necessarily mean the weather stations are not evenly distributed in space, considering that those states in the west each occupy larger area as well --- we will save this question to answer later.

Summary of Weather Stations by Grouped States (Regions)



Let's take a closer view of the U.S. mainland, which is our analysis focus. Given that analyzing the weather by state could sometimes make us trapped in trivial details, we decide to group (and color) the states according to their geographic regions. As the legend shows, we have altogether 9 regional groups, including New England, Middle Atlantic, Upper Mississippi, Mid-west, Rocky Mountain States, Pacific, Southeast, South and Southwest. By doing so we maintain the spatial disparities between groups to the largest extent, and we don't need to worry about the within-group differences. Besides, these 9 groups are more comparable in size than 48 states. In the following analysis we'll choose the scale (region/state) according to different visualization goals.

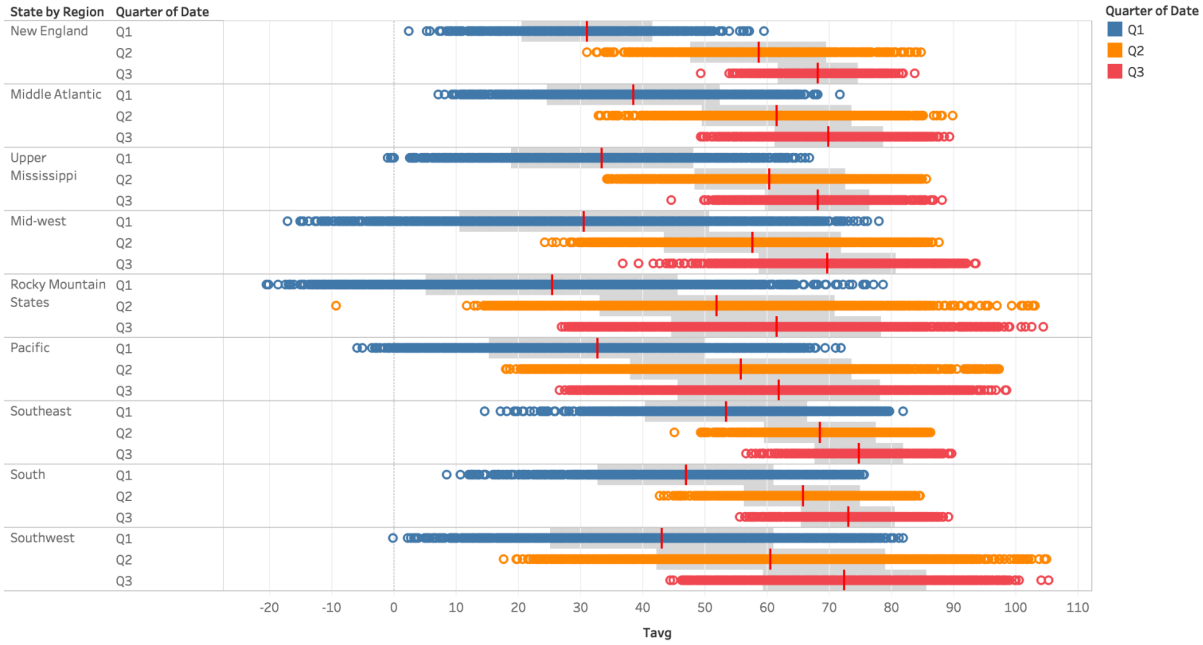
Where are the Weather Stations? - The Spatial Distribution



Now we get back to our previous question: are weather stations evenly distributed across space? In this map we visualize the location of each weather station as a dot, and color it by its corresponding regional group. Although the stations seem to be a bit sparser in the Mid-west and Southwest regions than others, we can still safely conclude that the the stations are evenly distributed. This is critical to our following analysis, since only with evenly distributed stations can we derive statistics of all stations in each regional group and assume those statistics are representative.

Through several visualizations above we can draw conclusions about the spatial coverage of this dataset: it includes all states in the U.S. and Canada, but we'll focus on the U.S. mainland only. And the weather stations covered in this dataset are evenly distributed across space.

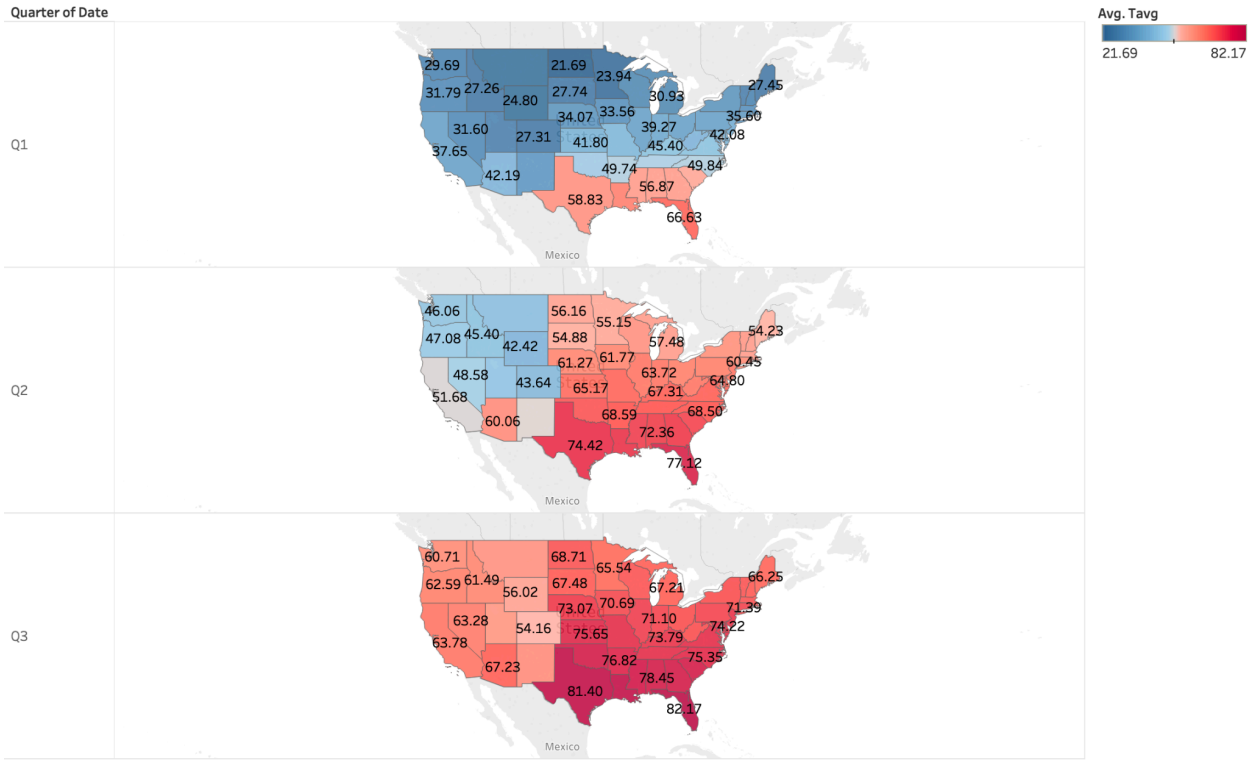
The Distribution of Average Temperatures by Region and Quarter



By quarter and by region, this graph depicts the mean of Tavg (the average daily temperature) recorded by each weather station. Hence, every circle stands for a weather station. In each row we visualize the average Tavg of each station in the corresponding region during the corresponding quarter, along with reference lines for the interquartile range (middle 50% of the data, in grey) and median values (in red). Several conclusions can be reached according to this graph: (1) For all regions, Q3 is generally the hottest quarter while Q1 is the coldest; (2) On average, the regions in the south (e.g. Southeast, South and Southwest) have higher temperature in Q1, but not necessarily in Q2 and Q3. The temperature differences among regions are most significant in Q1; (3) we can also see some outliers on both ends of the distribution , which may imply extreme weathers near some weather stations.

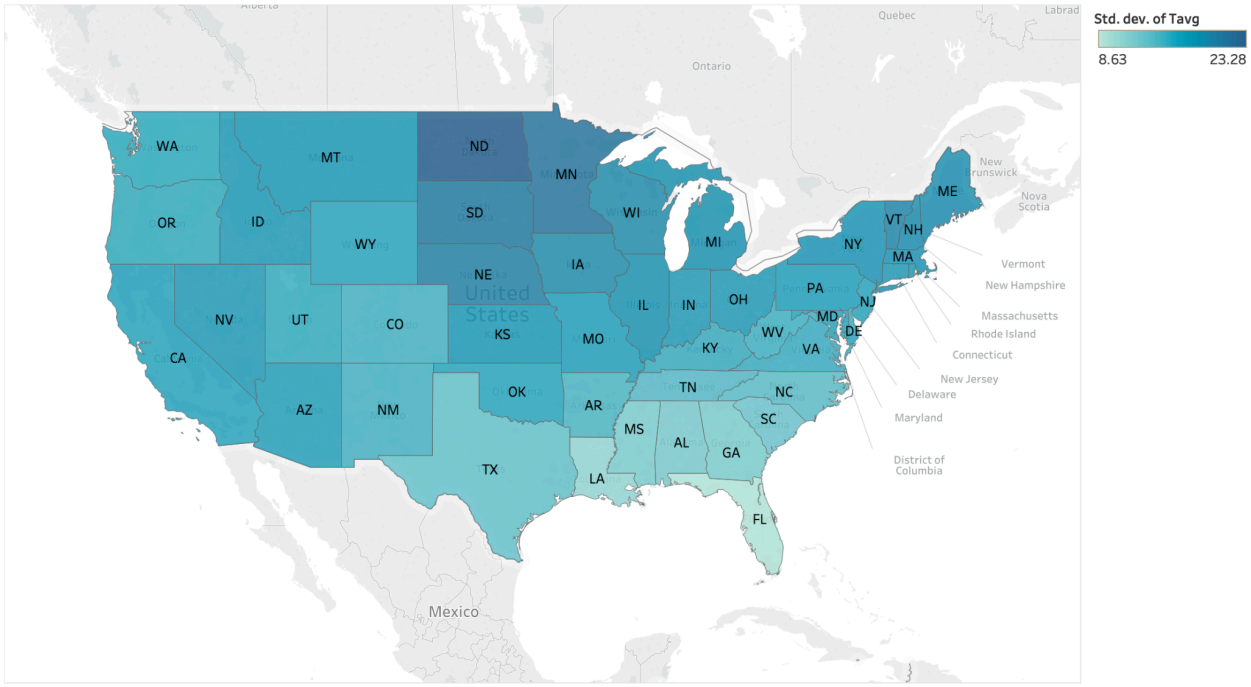
The pattern shown in this graph also raises our curiosity: which region/state has the biggest variation in temperature over time? What factors may be correlated with the temperature variation?

How Do Average Temperatures Vary across Space?

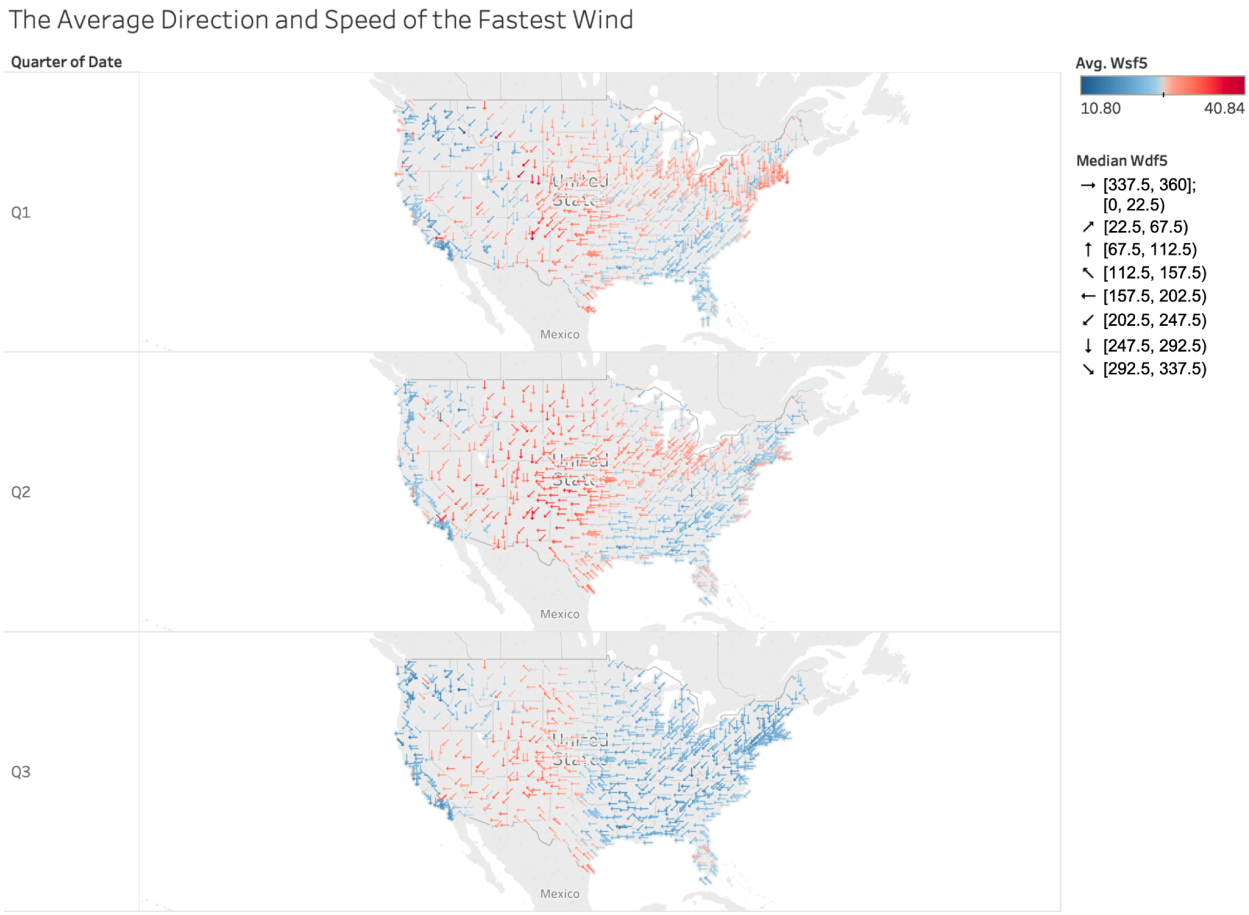


Before answering the question we first explore the average Tavg (the average daily temperature) by state and by quarter. We visually encode the value by a color palette, where red represents higher temperature while blue means coldness. Generally, the more southern a state is, the higher temperature it will have. Again, we find that the temperature disparities are the smallest in Q3, when all states get hot. As expected, most states will suffer from cold in Q1. Surprisingly, we find that two coasts show very different weather patterns in Q2: the winter in regions near west coast (including Pacific and Rocky Mountain States) seems to last longer and extend to Q2, which makes the average temperature there 10-15 degrees lower than the east coast.

Climate: Seasonality vs Constant?



In addition to the average temperature in different states within each quarter, we also care about to what extent the temperature vary over time in different states, which is a indicator of seasonality. This graph tells the temperature variation story by showing the standard deviation of monthly average Tavg in each state. We conduct the monthly average to partial out the variation arising from space. We use darker color to represent larger standard deviation of temperature. As we can see, ND (North Dakota) is the most seasonal state in the U.S, while the states in the Southeast are the least seasonal ones, because the weathers there are hot throughout the year.

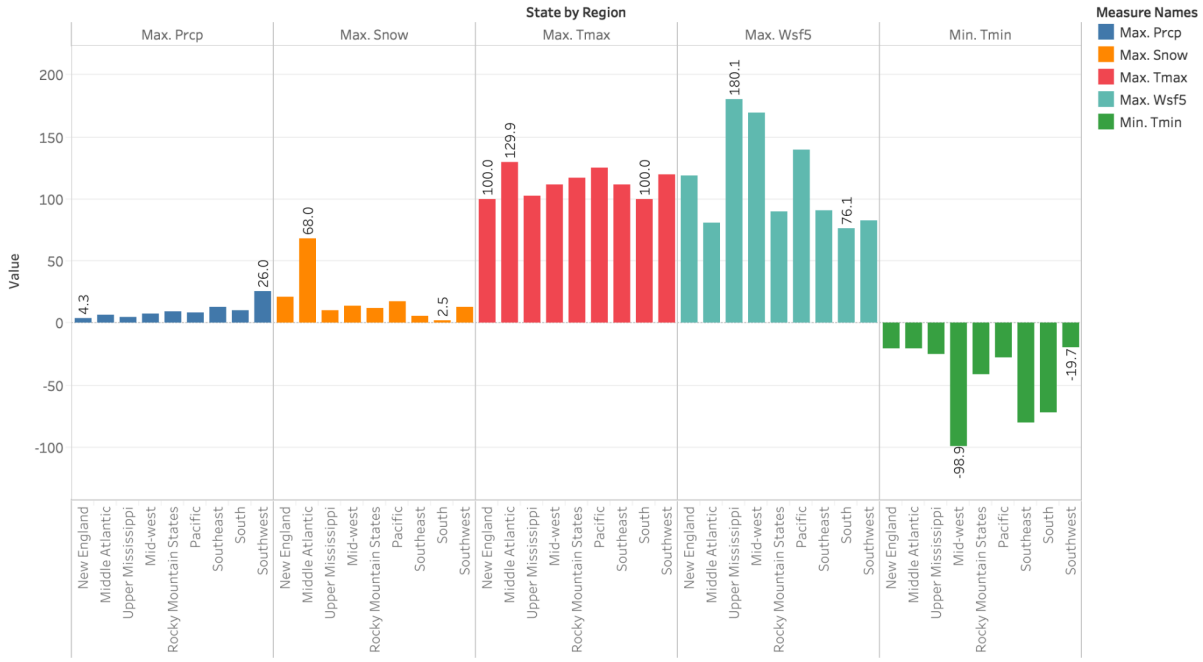


We guess the strong wind might be an important factor to cause temperature variation. To explore their relationship, we first visualize the fastest wind in a map. For each station and each quarter, we take the median direction and average speed of the fastest wind it records as its 'typical' fastest wind in that quarter. For the direction of wind, we group the continuous quantitative values into 8 groups, where each has a range of 45 degrees. We represent all degrees in one group by the same arrow (see legend). The wind speed is encoded as blue (lighter) and red (stronger) palette.

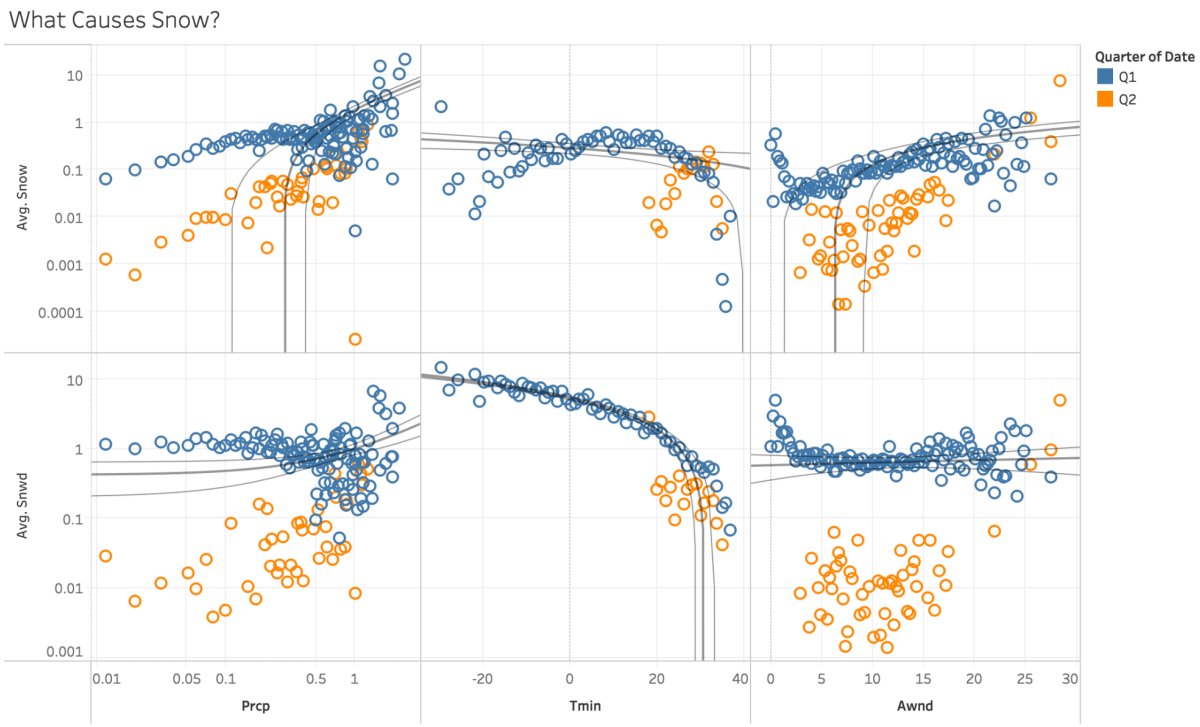
As we can see from the graph, the fastest wind pattern doesn't vary too much over time. It might be because we take the median direction. For all three quarters, it seems that the middle part (especially Mid-west) has the strongest wind. In comparison, coastal parts have much lighter winds.

To better reveal the relationship between temprature's standard deviation and the fastest wind, we use a scatter plot to show the two dimensions for all 48 states in the U.S. mainland. A linear model is employed to fit their relationship. In the graph above, the straight line is the linear fitting line, which is accompanied by two curves showing the confidence bands. The positive correlation between the temprature's standard deviation and the speed of the fastest wind is pretty robust.

Where are the Extreme Weathers in U.S.?



Our next step is to investigate the extreme weathers in the U.S. In this graph, we use bar charts to visualize the maximum Prcp (precipitation), the maximum Snow (snowfall), the maximum Tmax (maximum daily temperature), the maximum WSF5 (fastest 5-second wind speed), and the minium Tmin (minimum daily temperature). We label the maximum and minimum values in each column, in order to know which region have the most extreme or the most normal weather. For example, in terms of snowfall (our next interest), Middle Atlantic region shows the biggest maximum snowfall while the South region shows the smallest maximum. This means it's most likely to witness a snowstorm in Middle Atlantic among all regions, while a snowstorm will be extremely rare in South. Similar conclusions can be drawn with respect to other measures in this chart.



Our last question is what factors are related to the amount of snow. We propose three potential factors, including Prcp (percipitation), Tmin (minimum daily temperature), as well as Awnd (average daily wind speed). Since there're two measures of snow: Snow (snowfall) and Snwd (snow depth), we make one scatter plot for each pair from the two snow meaures and the three potential factors --- hence, we have six plots in total. For the two snow measures and the percipitation, we choose to visualize them using a logarithmic scale, given their distributions (too many observations cluster around low values). For the other two factors, namely minimum temperature and average wind, we keep the original scale. We conduct a linear fitting for each of the six cells/pairs, and the confidence bands are also presented. The linear fitting may not look linear due to the change of axis scale. Each data point in this graph stands for a weather station. We color the data by quarter but the linear fitting is for all data points rather than by quarter. Q3 is left out because there's no snow in the third quarter across the U.S.

Clearly, all three factors seem to play a role: they're linearly correlated with snow. To be more specific, more snow is associated with higher percipitation, lower minimum daily temperature, and faster average wind speed.

Summary

In this exploratory data analysis, we explore the spatial and temporal distribution of the U.S. weather measurements. We first check the data quality and resolution, and find that (1) this dataset covers each day from Jan 01, 2017 to Sep 21, 2017, and the spatial coverage of each month is quite similar; (2) it includes all states in the U.S. and Canada, but we focus on the U.S.

mainland only; and (3) the weather stations covered in this dataset are evenly distributed across space.

We then dive deeper into the weather measurements and try to address our interested questions. In addition to the spatial distribution patterns of temperature (average temperature and its standard deviation), we find a robust positive correlation between the temperature's standard deviation (an indicator of seasonality) and the fastest wind speed. We also explore the extreme weathers in the U.S., and reach the conclusion that more snow is associated with higher percipitation, lower minimum temperature, and faster wind speed.