

FILTERING

FILTERING GROUPS AND SUBGROUPS

- **filtering data** - selecting cases (or subgroups) in your data >> to focus on them - or compare them against each other

INSTALL AND LOAD PACKAGES

```
pacman::p_load(pacman, rio, tidyverse)
```

LOAD AND PREPARE DATA

- import file BP created - **StateData**
- 49 continental US states / state code / region, governor (republican / democrat) /
 - psychRegion (from psych study of their profile (levels = friendly and conventional / relaxed and creative/ temperamental and uninhibited) at state level
 - 5 personality factors (extraversion, agreeableness, conscientiousness, neuroticism, openness)
 - results through google correlate (about relative popularity of search terms on state by state basis)


Select specific columns for analysis:

Change PsychRegions from **character** to **factor**

```
df <- import("data/StateData.xlsx") %>%
  as.tibble() %>%
  select(state_code,
         region,
         psychRegions,
         instagram:modernDance) %>%
  mutate(psychRegions = as.factor(psychRegions)) %>%
  # rename(y = psychRegions) %>%
  print()
```

- **results:** tibble 48 x 15

```
# A tibble: 48 x 15
  state_code region    psychRegions    instagram facebook retweet entrepreneur
  <chr>      <chr>      <fct>          <dbl>      <dbl>      <dbl>          <dbl>
1 AL        South    Friendly and C...  0.64      1.65      0.35          0.257
2 AZ        West     Relaxed and Cr...  0.183     -0.259   -0.566         0.562
3 AR        South    Friendly and C...  0.456     1.10     -0.598         0.245
4 CA        West     Relaxed and Cr...  1.47     -0.422    0.481         0.502
5 CO        West     Friendly and C... -1.03     -1.06    -0.902         0.023
6 CT        Northeast Temperamental ...  0.374     -0.982    1.14          0.069
7 DE        South    Temperamental ...  1.48     -1.12     1.19          2.55
8 FL        South    Friendly and C...  0.85      0.38     -0.23          0.783
9 GA        South    Friendly and C...  0.807     0.526    0.035          1.95
10 ID       West     Relaxed and Cr... -0.736    -0.269   -1.80          0.296
# i 38 more rows
# i 8 more variables: gdpr <dbl>, privacy <dbl>, university <dbl>,
# mortgage <dbl>, volunteering <dbl>, museum <dbl>, scrapbook <dbl>,
# modernDance <dbl>
# i Use `print(n = ...)` to see more rows
```

Data	
 df	48 obs. of 15 variables

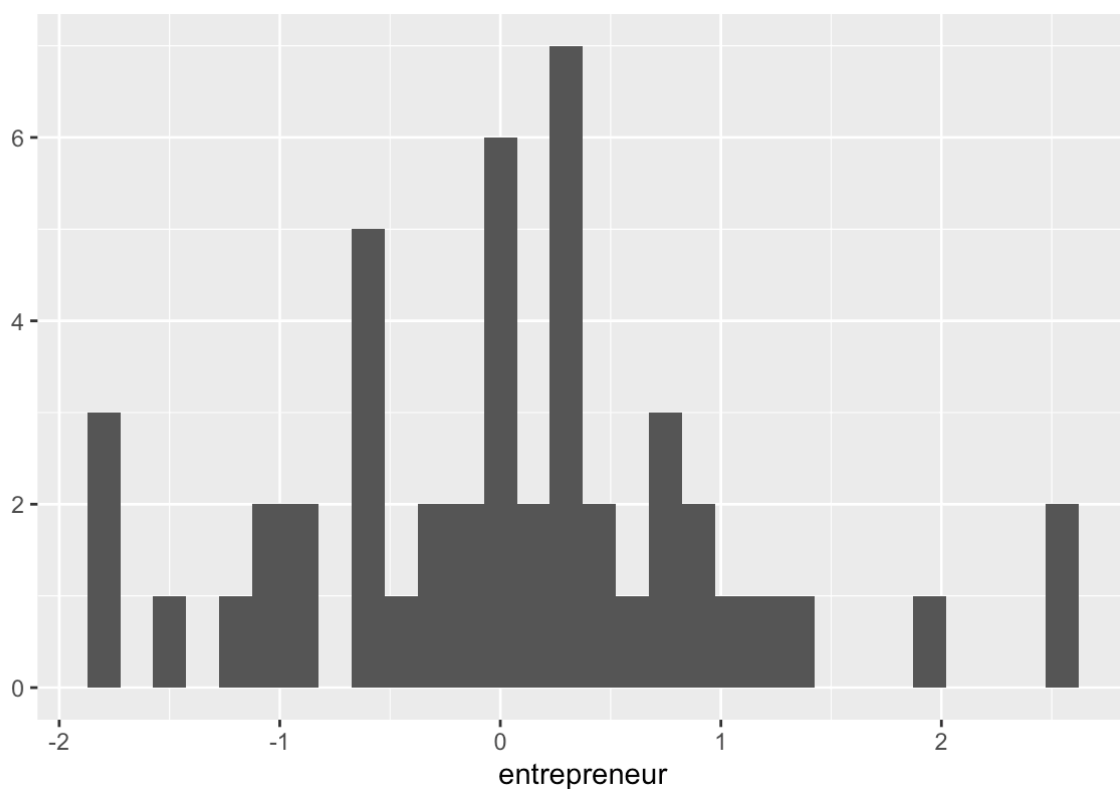
FILTER BY ONE VARIABLE**howto - QUANTITATIVE VARIABLE**

- How much does a state search for term “entrepreneur” on google (quantitative variable) - as compared to other states - as percentage

```
qplot(entrepreneur, geom = "histogram", data = df)
```

- results:

z-scores - 0 indicates that on national average



- Filter- to see which states unusually high for searches on “entrepreneurs”
 - (standard deviation > 1)

```
df %>%
  filter(entrepreneur > 1) %>%
  print()
```

- results:

console: tibble 5 X 15

-tf 5 states above std dev of 1 for searches on entrepreneurship (DE GA MD NC UT)) unusually high on relative popularity of entrepreneurship as a search term

```
# A tibble: 5 x 15
  state_code region psychRegions  instagram facebook retweet entrepreneur  gdpr
  <chr>      <chr>    <fct>          <dbl>    <dbl>    <dbl>      <dbl> <dbl>
1 DE        South  Temperamenta...  1.48    -1.12    1.19      2.55  1.21
2 GA        South  Friendly and...  0.807    0.526    0.035     1.95  0.403
3 MD        South  Temperamenta...  0.895    -1.47    1.56      1.17  0.603
4 NC        South  Relaxed and ...  0.357    0.443    0.454     1.27  0.065
5 UT        West   Relaxed and ... -0.089   -1.57   -0.779     2.54  0.158
# i 7 more variables: privacy <dbl>, university <dbl>, mortgage <dbl>,
#   volunteering <dbl>, museum <dbl>, scrapbook <dbl>, modernDance <dbl>
```

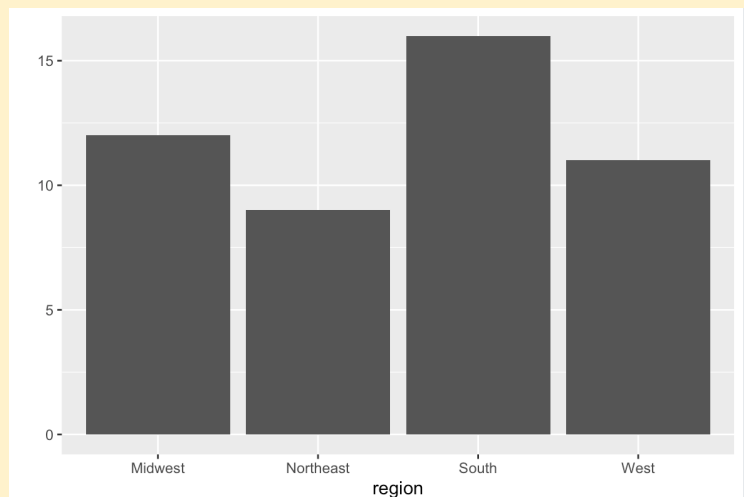
howto - CHARACTER VARIABLE

- What are the regions (character variable) - used to indicate a category membership

```
qplot(region, data = df)
```

- results:

creates **bar chart for Region** (midwest / northeast / south / west)
 -more state in south - relatively few in north-east



- Look at just the states listed in the South in this dataset:

```
df %>%
  filter(region == "south") %>%
  print()
```

- results:

console : lists all the southern states (tibble 16 obs x 15 variables)

```
# A tibble: 16 x 15
  state_code region psychRegions    instagram facebook retweet entrepreneur
  <chr>      <chr>    <fct>          <dbl>    <dbl>    <dbl>        <dbl>
1 AL        South    Friendly and Conv...  0.64     1.65     0.35         0.257
2 AR        South    Friendly and Conv...  0.456    1.10    -0.598        0.245
3 DE        South    Temperamental and...  1.48    -1.12     1.19         2.55
4 FL        South    Friendly and Conv...  0.85     0.38    -0.23         0.783
5 GA        South    Friendly and Conv...  0.807    0.526   0.035         1.95
6 KY        South    Friendly and Conv... -0.221    1.12    0.325        -0.647
7 LA        South    Friendly and Conv...  1.52     0.114  -0.087         0.044
8 MD        South    Temperamental and...  0.895    -1.47     1.56         1.17
9 MS        South    Friendly and Conv...  1.29     1.86    -0.508        0.404
10 NC       South    Relaxed and Creat...  0.357    0.443   0.454         1.27
11 OK       South    Friendly and Conv... -0.162   -0.421  -0.472        -0.942
12 SC       South    Friendly and Conv...  0.684    1.42    0.033         0.693
13 TN       South    Friendly and Conv...  0.288    0.812   0.326         0.818
14 TX       South    Temperamental and...  0.438    0.827  -0.382         0.178
15 VA       South    Relaxed and Creat... -0.719    1.04    0.006        -0.538
16 WV       South    Temperamental and... -0.114    2.25    1.25        -1.77
# i 8 more variables: gdp <dbl>, privacy <dbl>, university <dbl>,
# mortgage <dbl>, volunteering <dbl>, museum <dbl>, scrapbook <dbl>,
# modernDance <dbl>
```

summary: to specify subgroup 1) use == 2) put value in quotes

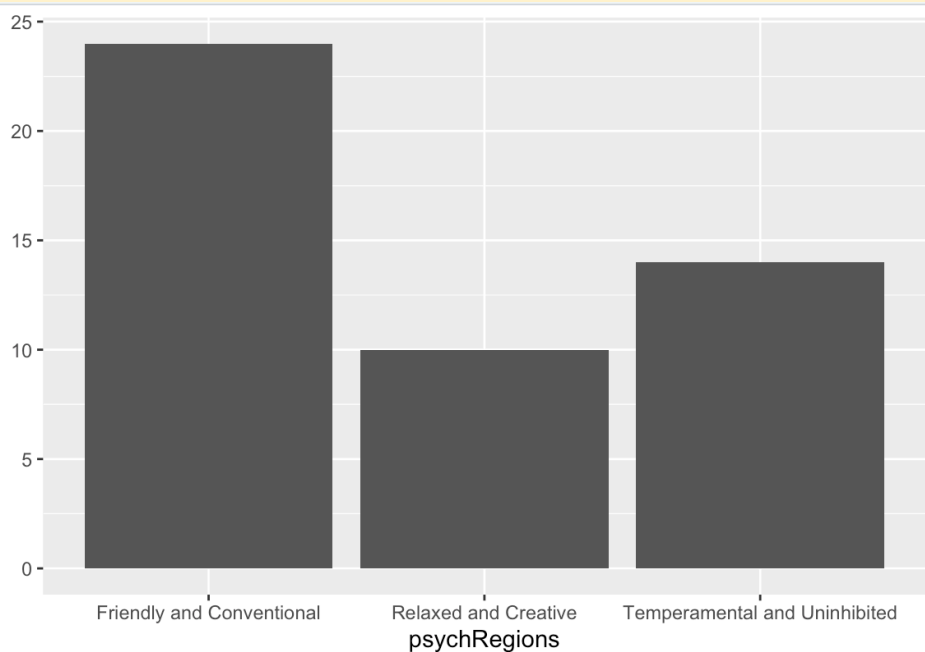
howto - FACTOR VARIABLE

```
qplot(psychRegions, data = df)
```

- **results:**

bar chart for psych regions

-most state are 'friendly and conventional' vs 'relaxed and creative' is small number



```
df %>%
  filter(psychRegions == "Relaxed and Creative") %>%
  print()
```

- **results:**

list of 10 states that are relaxed and creative

```
# A tibble: 10 x 15
  state_code region psychRegions    instagram facebook retweet entrepreneur
  <chr>      <chr>    <fct>          <dbl>      <dbl>      <dbl>          <dbl>
1 AZ        West    Relaxed and Creat...  0.183    -0.259    -0.566         0.562
2 CA        West    Relaxed and Creat...  1.47     -0.422     0.481         0.502
3 ID        West    Relaxed and Creat... -0.736    -0.269    -1.80          0.296
4 NV        West    Relaxed and Creat...  1.62     -0.244     0.276         0.275
5 NM        West    Relaxed and Creat... -0.17     0.857    -1.39         -0.957
6 NC        South   Relaxed and Creat...  0.357     0.443     0.454         1.27
7 OR        West    Relaxed and Creat... -0.001     1.46    -0.735        -1.80
8 UT        West    Relaxed and Creat... -0.089    -1.57    -0.779         2.54
9 VA        South   Relaxed and Creat... -0.719     1.04     0.006        -0.538
10 WA       West    Relaxed and Creat... -0.654    -1.11    -0.285        -1.09
# i 8 more variables: gdpr <dbl>, privacy <dbl>, university <dbl>,
# mortgage <dbl>, volunteering <dbl>, museum <dbl>, scrapbook <dbl>,
# modernDance <dbl>
```

FILTER BY MULTIPLE VARIABLES**howto - OR**

- by combining search terms
- “or” is vertical pipe |

```
df %>%
  filter(region == "South" |
         psychRegions == "Relaxed and Creative") %>%
  print()
```

- **results:**

24 states (fall into either one of the listed categories)

```
# A tibble: 24 x 15
  state_code region psychRegions      instagram facebook retweet entrepreneur
  <chr>      <chr>   <fct>                <dbl>      <dbl>    <dbl>          <dbl>
1 AL        South   Friendly and Conv...    0.64       1.65     0.35           0.257
2 AZ        West    Relaxed and Creat...    0.183     -0.259   -0.566         0.562
3 AR        South   Friendly and Conv...    0.456     1.10    -0.598         0.245
4 CA        West    Relaxed and Creat...    1.47     -0.422    0.481         0.502
5 DE        South   Temperamental and...    1.48     -1.12     1.19          2.55
6 FL        South   Friendly and Conv...    0.85      0.38    -0.23          0.783
7 GA        South   Friendly and Conv...    0.807     0.526    0.035          1.95
8 ID        West    Relaxed and Creat...   -0.736    -0.269   -1.80          0.296
9 KY        South   Friendly and Conv...   -0.221     1.12     0.325        -0.647
10 LA       South   Friendly and Conv...    1.52      0.114   -0.087         0.044
# i 14 more rows
# i 8 more variables: gdpr <dbl>, privacy <dbl>, university <dbl>,
# mortgage <dbl>, volunteering <dbl>, museum <dbl>, scrapbook <dbl>,
# modernDance <dbl>
# i Use `print(n = ...)` to see more rows
```

howto - AND

- “and” is ampersand &

```
df %>%
  filter(region == "South" &
         psychRegions == "Relaxed and Creative") %>%
  print()
```

- **results:**

combined, joint set - which of southern states have also been classified as relaxed and creative > 2 states

```
# A tibble: 2 x 15
  state_code region psychRegions      instagram facebook retweet entrepreneur gdpr
  <chr>      <chr>   <fct>                <dbl>      <dbl>    <dbl>          <dbl> <dbl>
1 NC        South   Relaxed and ...    0.357     0.443    0.454          1.27 0.065
2 VA        South   Relaxed and ...   -0.719     1.04     0.006         -0.538 0.168
# i 7 more variables: privacy <dbl>, university <dbl>, mortgage <dbl>,
# volunteering <dbl>, museum <dbl>, scrapbook <dbl>, modernDance <dbl>
```

howto - NOT

- “not” is an exclamation point !

```
df %>%
  filter(region == "South" &
    !psychRegions == "Relaxed and Creative") %>%
  print()
```

- result:

filter all southern states that are NOT classified as relaxed and creative (joint search) = 14 states

```
# A tibble: 14 x 15
  state_code region psychRegions    instagram facebook retweet entrepreneur
  <chr>      <chr>    <fct>          <dbl>      <dbl>    <dbl>          <dbl>
1 AL        South    Friendly and Conv...  0.64       1.65     0.35           0.257
2 AR        South    Friendly and Conv...  0.456      1.10    -0.598         0.245
3 DE        South    Temperamental and...  1.48      -1.12     1.19           2.55
4 FL        South    Friendly and Conv...  0.85       0.38    -0.23           0.783
5 GA        South    Friendly and Conv...  0.807      0.526    0.035           1.95
6 KY        South    Friendly and Conv... -0.221      1.12     0.325          -0.647
7 LA        South    Friendly and Conv...  1.52       0.114   -0.087           0.044
8 MD        South    Temperamental and...  0.895     -1.47     1.56           1.17
9 MS        South    Friendly and Conv...  1.29       1.86    -0.508           0.404
10 OK       South    Friendly and Conv... -0.162     -0.421   -0.472          -0.942
11 SC       South    Friendly and Conv...  0.684      1.42     0.033           0.693
12 TN       South    Friendly and Conv...  0.288      0.812    0.326           0.818
13 TX       South    Temperamental and...  0.438      0.827   -0.382           0.178
14 WV       South    Temperamental and... -0.114      2.25     1.25           -1.77

# i 8 more variables: gdpr <dbl>, privacy <dbl>, university <dbl>,
# mortgage <dbl>, volunteering <dbl>, museum <dbl>, scrapbook <dbl>,
# modernDance <dbl>
```