Welcome to DATA 151

I'm so glad you're here!

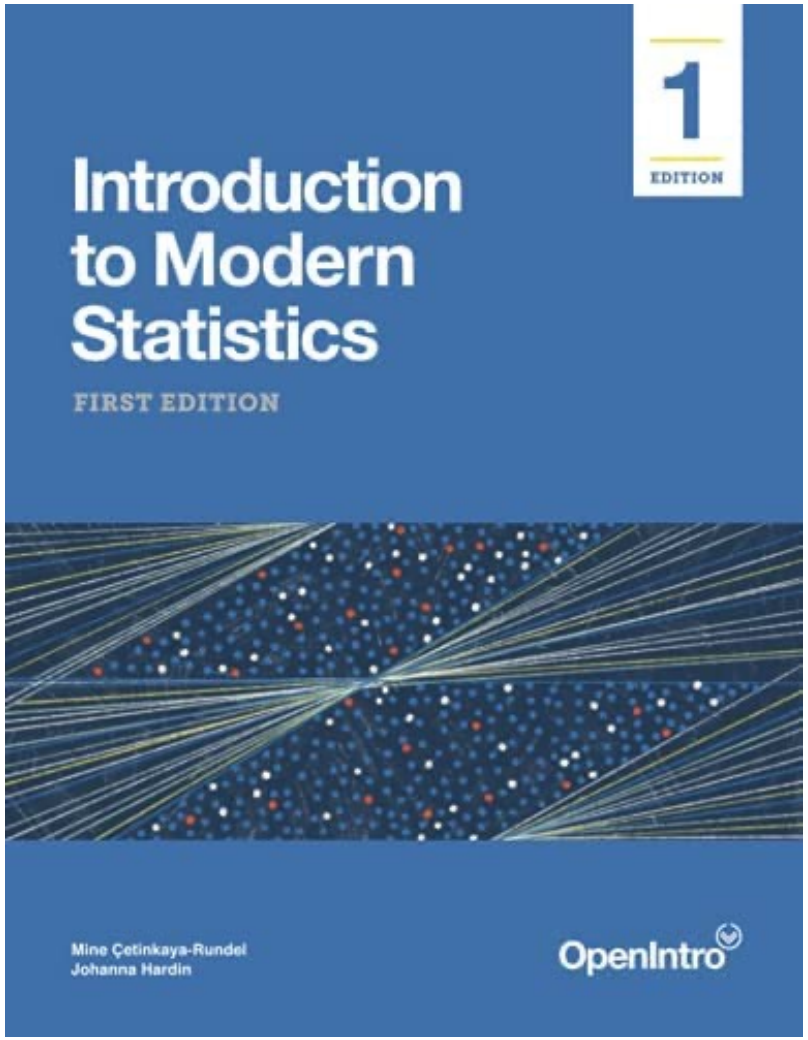
DATA 151: CLASS 9A
INTRODUCTION TO DATA SCIENCE (WITH R)

COMPARING DISTRIBUTIONS

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING



Introduction to Data Science:

- Tuesday:
  - Introduction to Modern Statistics
  - Ch 5: Exploring Numeric Data

HOMEWORK REMINDER

Due this week:

- ***DUE 10/25*** *Project Milestone #4: EDA Step 2*
  - Create Tables and Bar Graphs
- ***DUE 10/27*** *HW #8: DC Exploratory Data Analysis with Numeric Data*
  - *Just one chapter*
  - ***No submission on WISE necessary, do on DataCamp***

HOMEWORK REMINDER

Due next week:

- ***DUE 11/1*** *Project Milestone #5: EDA Step 3*
  - Numeric Distributions and Summary Statistics
- ***DUE 11/3*** *HW #9: DC Exploratory Data Analysis with Numerical Summaries*
  - *Just one chapter*
  - ***No submission on WISE necessary, do on DataCamp***

FRIENDLY REMINDER

Midterm #2 is Next Thursday

(content from weeks 5-9)

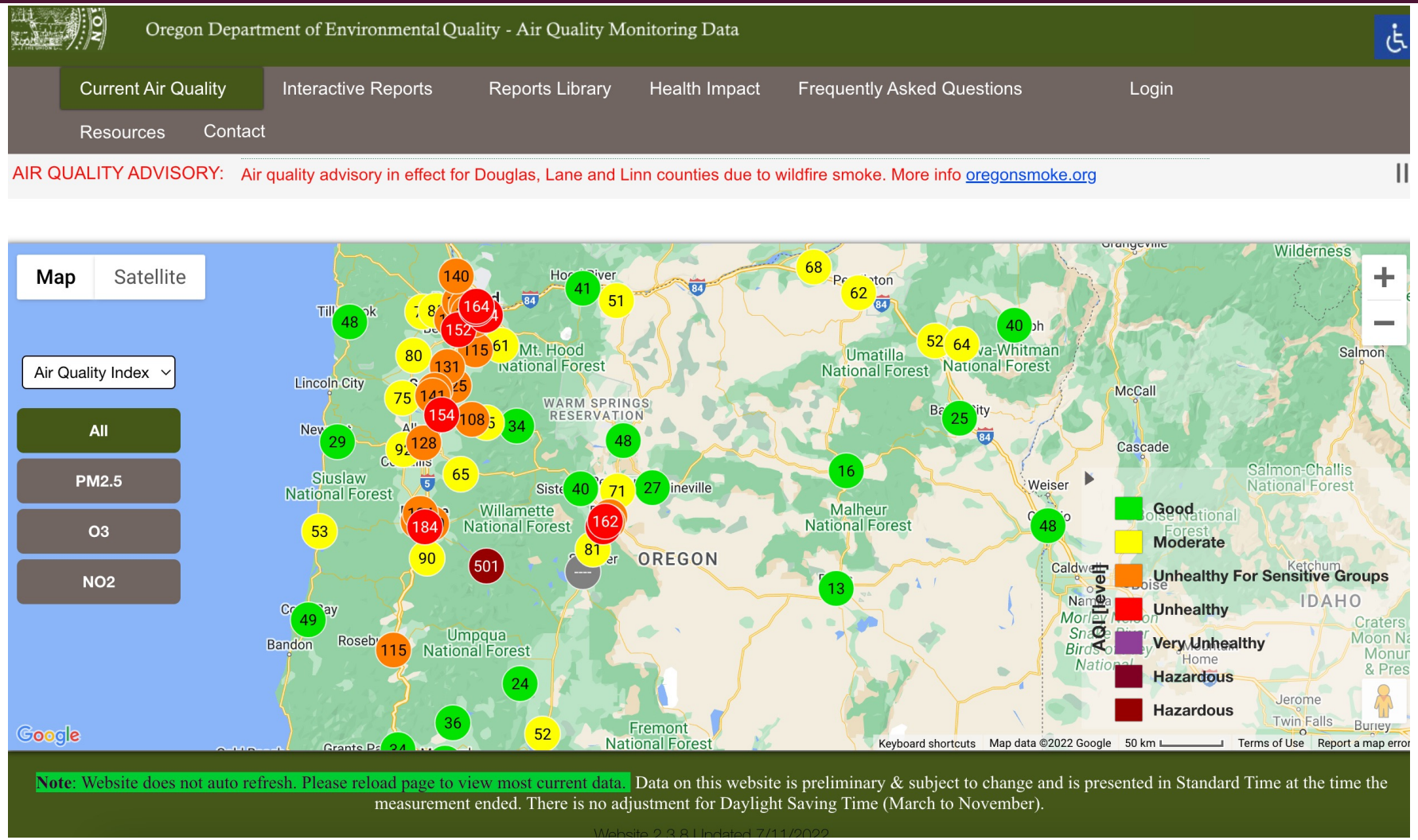
ANOTHER EXTRA CREDIT OPPORTUNITY

Meeting TODAY in FORD 224 at 4pm

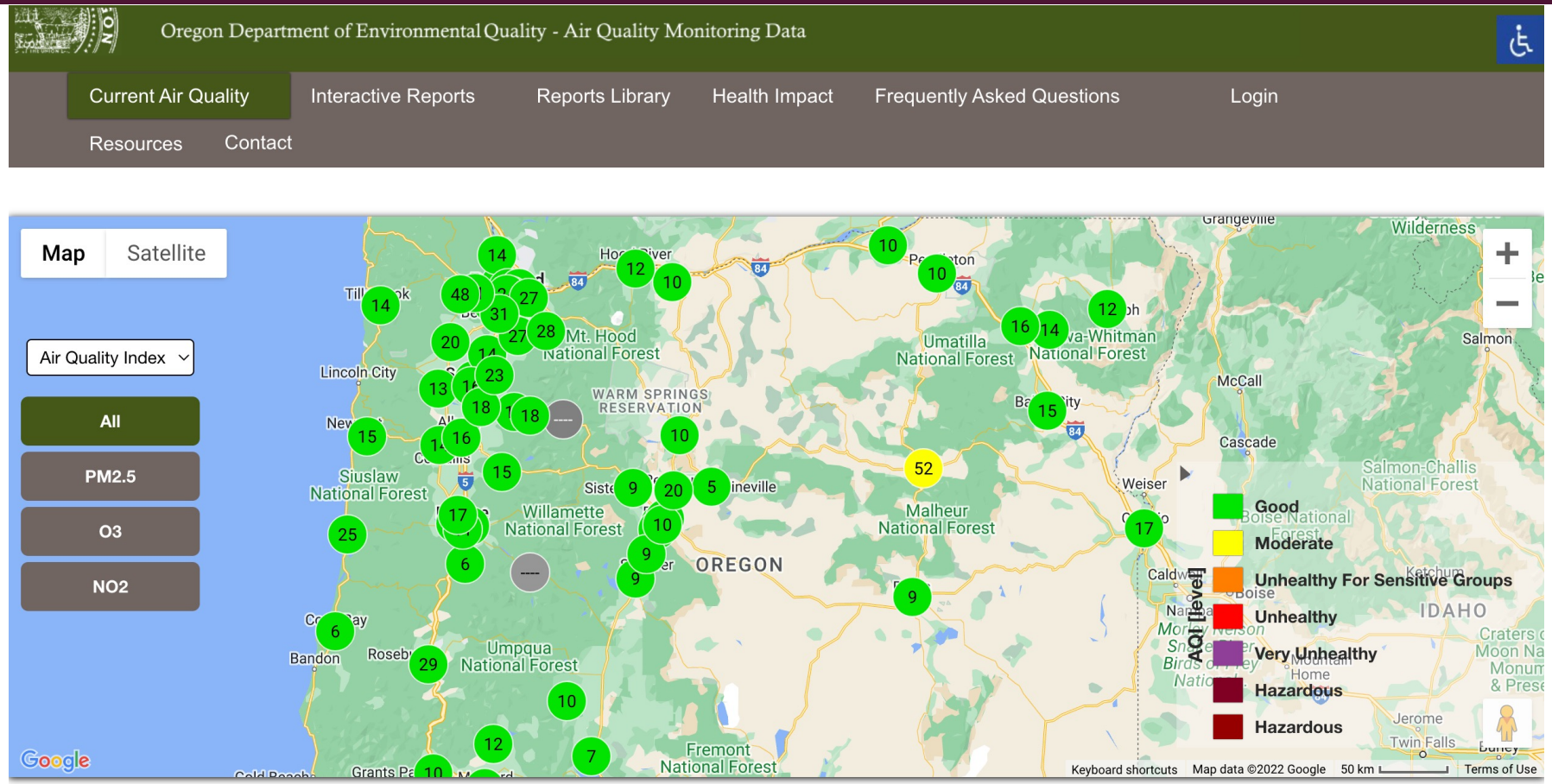
Students will work in teams as they analyze data on the K-12 educational experience "After the Bell." This year's theme will require student teams to dive into data on the impacts of school choice and family engagement in school activities and homework. Teams will provide recommendations on factors that best optimize family involvement and support of K-12 students' academic excellence.

Submissions will be accepted from October 17 to November 6, at 11:59pm EST.

OREGON AQI LAST WEEK



GOOD NEWS! OREGON AQI THIS WEEK



Note: Website does not auto refresh. Please reload page to view most current data. Data on this website is preliminary & subject to change and is presented in Standard Time at the time the measurement ended. There is no adjustment for Daylight Saving Time (March to November).

Website 2.3.8 Updated 7/11/2022

DATA151: Comparing Distributions

Kitada Smalley

Learning Objectives

In this lesson students will compare distributions from multiple populations of interest using:

- `dplyr`: `group_by` and `summarise`
- side-by-side `boxplots`
- side-by-side `violin` plots
- `beeswarm` plots
- faceting

Step 0: Library Tidyverse

```
library(tidyverse)
```

Step 1: Load the Data

```
aqi<-read.csv("https://raw.githubusercontent.com/kitadasmalley/DATA151/main/Data/fireAQI_OrCoWa_10192022.csv",
                      header=TRUE)
```

DESCRIBING DISTRIBUTIONS

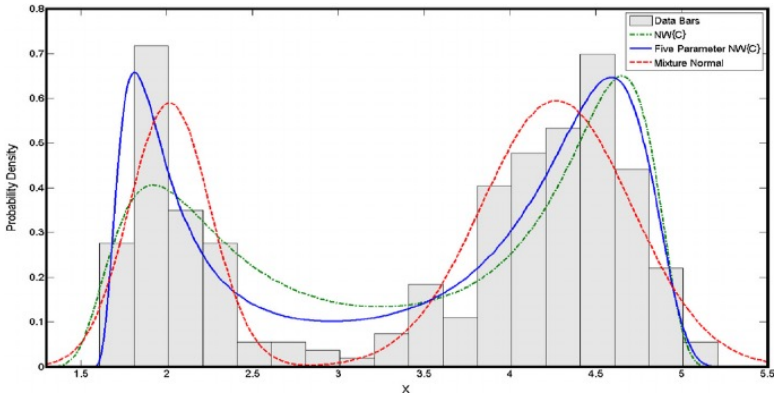
DESCRIBING DATA CHARACTERISTICS OF NUMERIC DATA

- **Shape :** Look for an overall pattern
  - Is the data symmetric?
  - Is it skewed? Right (positive) or left (negative) skewed?
- **Center :** Represents a typical value in a data set
  - If symmetric, use mean (or median)
  - If skewed, use median
- **Spread :** How spread out the values are
  - If symmetric, use standard deviation
  - If skewed, use interquartile range (IQR)
- **Unusual observations / Outliers**
  - Observations that are "far" from the others

We will rigorously define these terms when we talk about numerical summaries

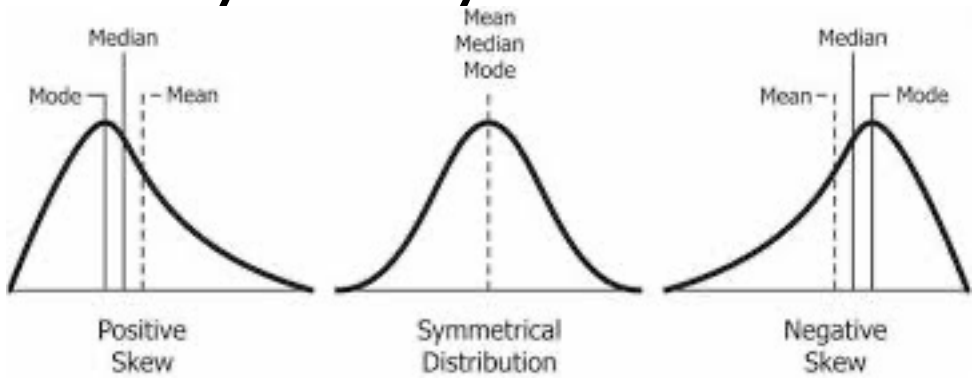
DESCRIBING SHAPE

- **Modality :** Number of peaks of mass in a numeric distribution
  - Unimodal (one peak)
  - Bimodal (two peaks)
  - Multimodal (many peaks)



Old Faithful Eruptions

- **Skewness** : Measure of asymmetry



NUMERICAL SUMMARIES

STATISTICAL NOTATION

- n = sample size
- $\bar{x}$ = sample mean. *This is the mean of some quantitative variable for a sample of size n.*
- s = sample standard deviation. *This is the standard deviation of some quantitative variable for a sample of size n.*
- s^2 = sample variance
- M = median
- Q_1 = The first quantile
- Q_3 = The third quantile
- IQR = Interquartile range

MEASURING CENTER

- Measures of center estimates… a variable's "typical" value
- The two kinds of measures of center will be the … mean and the median
- Depending on the shape of the data, we will talk about which measure of center is best to use when describing the data set.

THE MEAN

- The most common measure of center is the arithmetic average, or **mean**

- To find the sample mean, $\bar{x}$ (pronounced "x-bar")

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$

- The mean is NOT considered a resistant measurement of center.
  - It is influenced by outliers.

MEASURING SPREAD

- Reporting the center alone does not tell the whole story, you need to discuss the spread.
- Statistics is really the study of variability
- Its important to understand how data vary, so that we can eventually make accurate conclusions about a larger population.

SPREAD: STANDARD DEVIATION

Sample standard deviation: the square root of the sample variance.

- The "average" distance of the observations from their mean

Used because units are the same as the data

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

USING DPLYR TO COMPARE GROUPS

Step 2: Numeric Summaries

How does the air quality compare in Oregon, Washington, and Colorado?

```
## LET'S START WITH MEAN and STANDARD DEV
aqi%>%
  group_by(State)%>%
  summarise(n=n(),
            avgAQI=mean(AQI, na.rm = TRUE),
            sdAQI=sd(AQI, na.rm = TRUE),
            NAs=sum(is.na(AQI)))
```

```
## # A tibble: 3 × 5
##   State          n avgAQI sdAQI   NAs
##   <fct>      <int>  <dbl> <dbl> <int>
## 1 Colorado      11   35.6  7.07     0
## 2 Oregon        47   81.0 57.9      3
## 3 Washington    52  117.  72.0      1
```

FIVE NUMBER SUMMARY

- The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation
  - Min
  - Q_1
  - Med (Q_2)
  - Q_3
  - Max

- The five-number summary is used to … get a quick summary of both center and spread, combine all five numbers

THE MEDIAN

- The **median *M*** is the midpoint of a distribution … the number such that half of the observations are smaller and the other half are larger

- To find the median of a distribution
  1. Arrange all observations from smallest to largest
  2. If the number of observations *n* is odd, the median *M* is the center observation in the ordered list.
  3. If the number of observations *n* is even, the median *M* is the average of the TWO CENTER observations in the ordered list.

Now let's add more metrics! Let's use the quartiles. Cutting our data into quartiles means that we have split the data into four even parts.

- Q_1: The first quartile is the 25th percentile
- Q_2 (Median): The second quartile is also known as the median and is the 50th percentile
- Q_3: The third quartile is the 75th percentile

We can use the `quantile()` function to get any desired quantile from our data set. Quantile is synonymous with percentile. When we use the quantile function we specify the fraction of data below a desired cut off point.

```
# The first quartile
quantile(aqi$AQI, prob=.25, na.rm = TRUE)
```

```
## 25%
##  39
```

We can do this for each state:

```
aqi%>%
  group_by(State)%>%
  summarise(n=n(),
            min=min(AQI, na.rm = TRUE),
            Q1=quantile(AQI, prob=.25, na.rm = TRUE),
            med=median(AQI, na.rm = TRUE),
            Q3=quantile(AQI, prob=.75, na.rm = TRUE),
            max=max(AQI, na.rm = TRUE))
```

```
## # A tibble: 3 × 7
##   State          n   min    Q1   med    Q3   max
##   <fct>      <int> <int> <dbl> <dbl> <dbl> <int>
## 1 Colorado      11    19  33.5  36     38     49
## 2 Oregon        47    13  35.8  62.5  116.   245
## 3 Washington    52    15  67    84    170    359
```

What do you observe?

BOX PLOTS

- The **five-number summary** is illustrated in a boxplot

- How to make a boxplot:
  1. Draw and label a number line that includes the range of the distribution
  2. Draw a central box from Q_1 to Q_3
  3. Note the median inside the box
  4. Extend lines (whiskers) from the box out to the minimum and maximum values, *that are not outliers*

QUARTILES AND THE INTERQUARTILE RANGE

- How to calculate the quartiles and the interquartile range:
- To calculate the **quartiles**:
  1. Arrange the observations in increasing order to locate the overall median *M*
  2. The **first quartile, Q_1,** is the median of the observations located to the left of the overall median.
  3. The **third quartile, Q_3,** is the median of the observations located to the right of the overall median.

- The interquartile range (IQR) is defined as: **$IQR = Q_3 - Q_1$**

OUTLIERS

- The 1.5 x IQR Rule for Outliers (the book calls these "fences")

- A data point is an "outlier" if
  - $< Q_1 - 1.5\ (IQR)$
  - $> Q_3 + 1.5\ (IQR)$

- Note: Fences are for construction, not for displays

TRY IT OUT ON YOUR WORKSHEET ☺

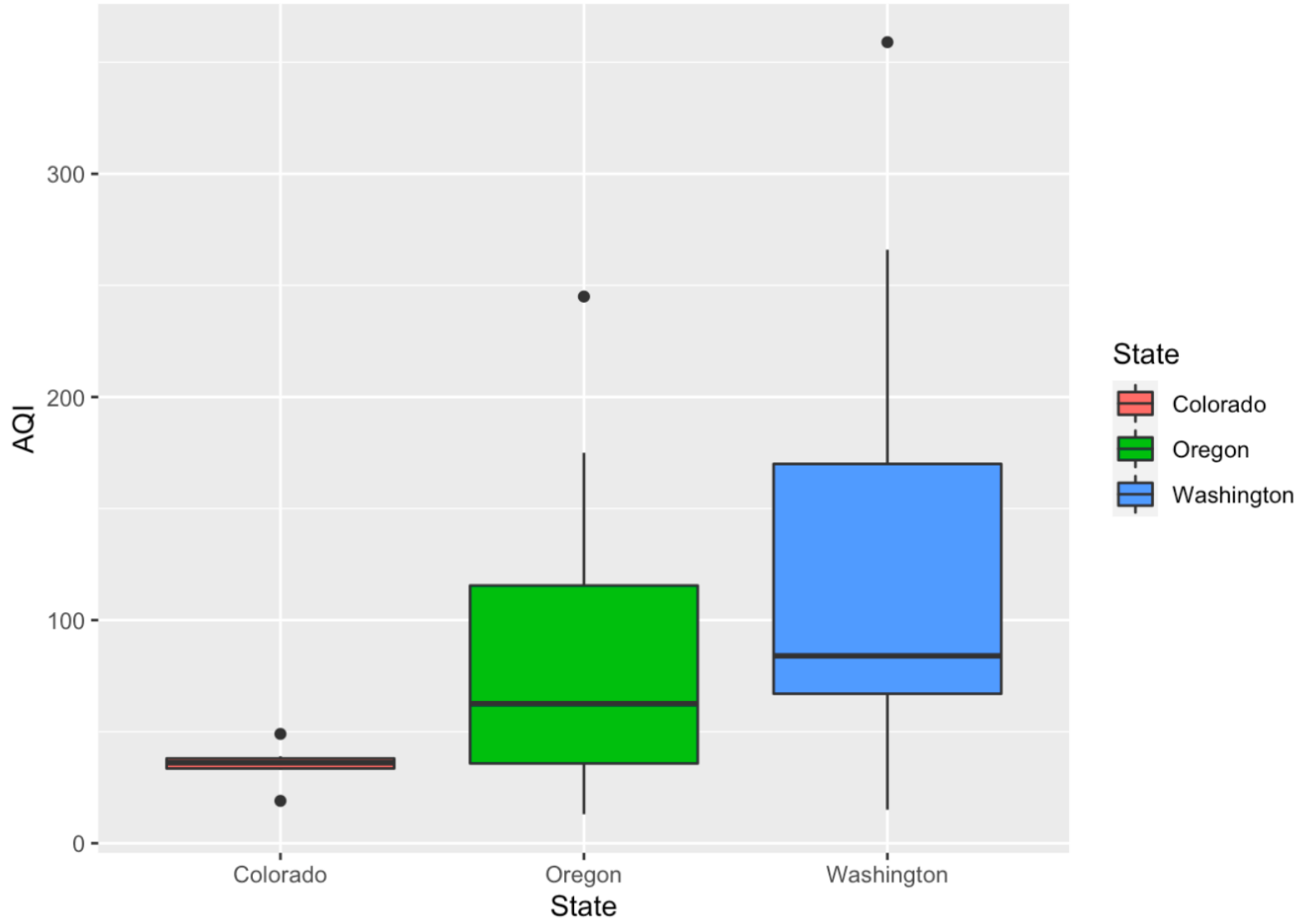
SIDE-BY-SIDE BOXPLOTS

But, my favorite is a side-by-side boxplot

```
# BOXPLOT


ggplot(aqi, aes(x=State, y=AQI, fill=State))+
  geom_boxplot()
```

But, my favorite is a side-by-side boxplot

```
# BOXPLOT

ggplot(aqi, aes(x=State, y=AQI, fill=State))+
  geom_boxplot()
```



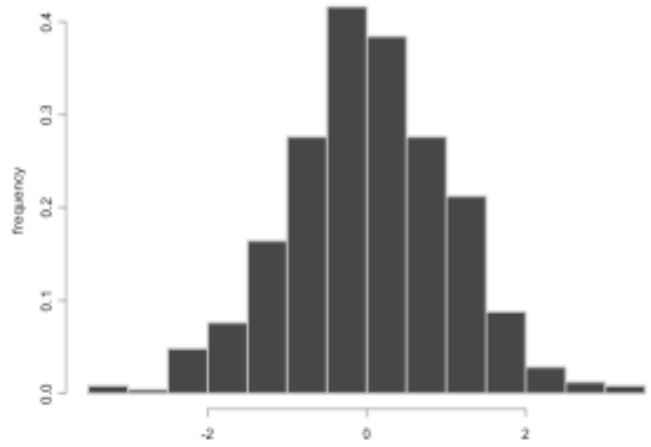
BOXPLOTS VS HISTOGRAMS

Histograms	Boxplots
• Most useful for larger data sets (usually more than 30 or so) • Provides more detail about the shape of the data • Caution: May be harder to see outliers	• Also most useful for large data sets • Contains less information about the shape of the data • You can see all the quartiles and median clearly. • It is easier to see outliers

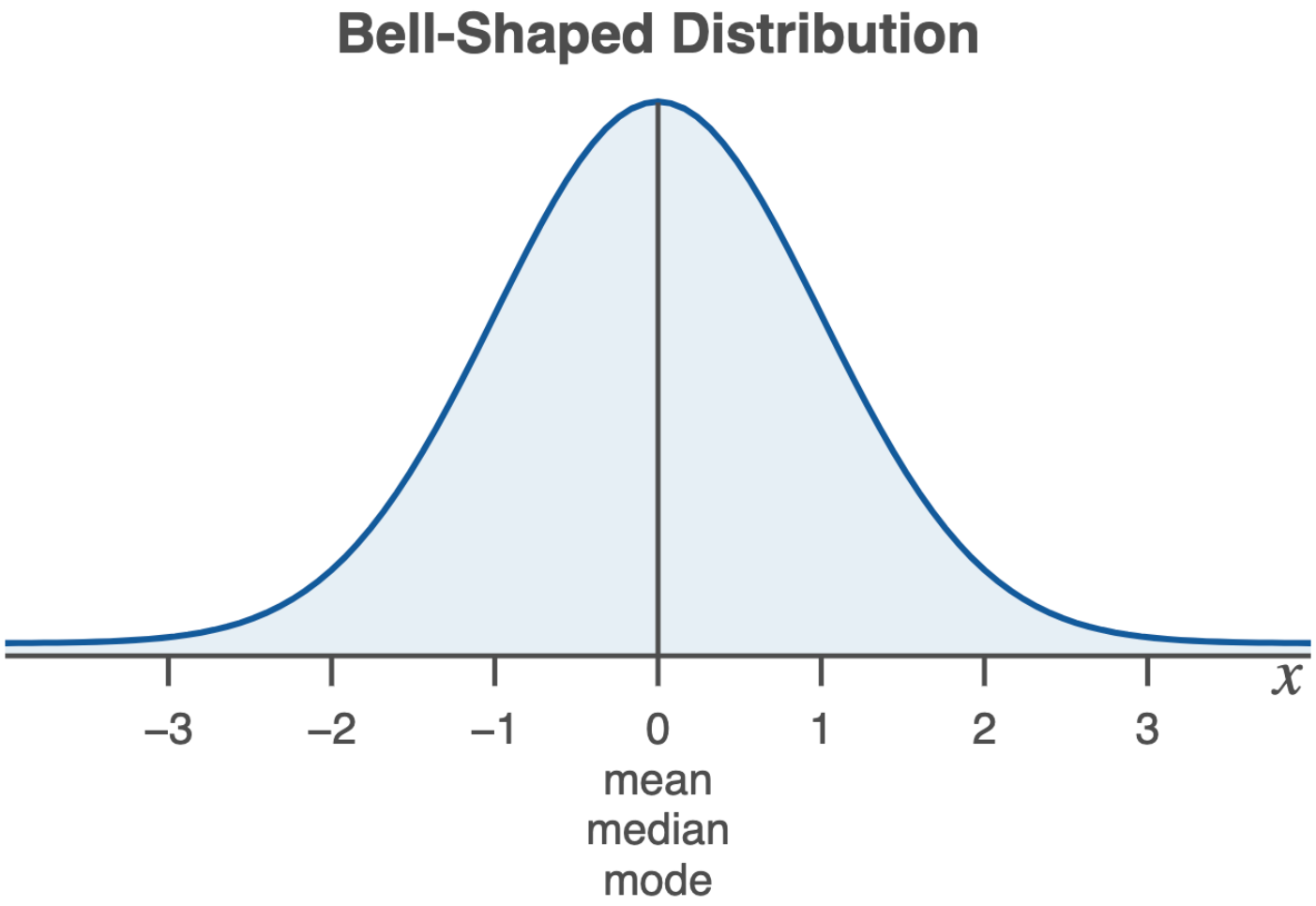
CRITICAL THINKING

CRITICAL THINKING ABOUT THE MEAN AND MEDIAN

- There is a relationship between mean, the median, and the shape of the data
- If the data are **symmetric** (or approximately symmetric):
  - In perfectly symmetric data, the mean and the median are equal.
  - In data that is approximately symmetric, the mean and the median are close to the same value.

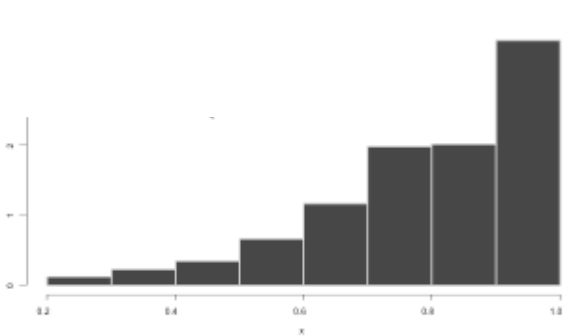


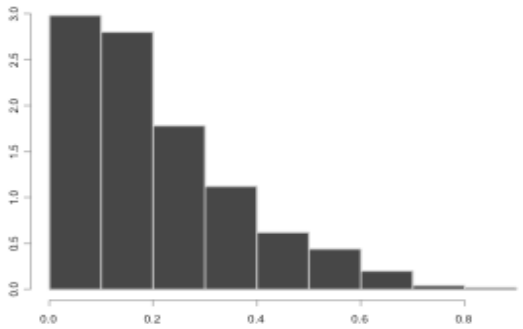
CRITICAL THINKING ABOUT THE MEAN AND MEDIAN



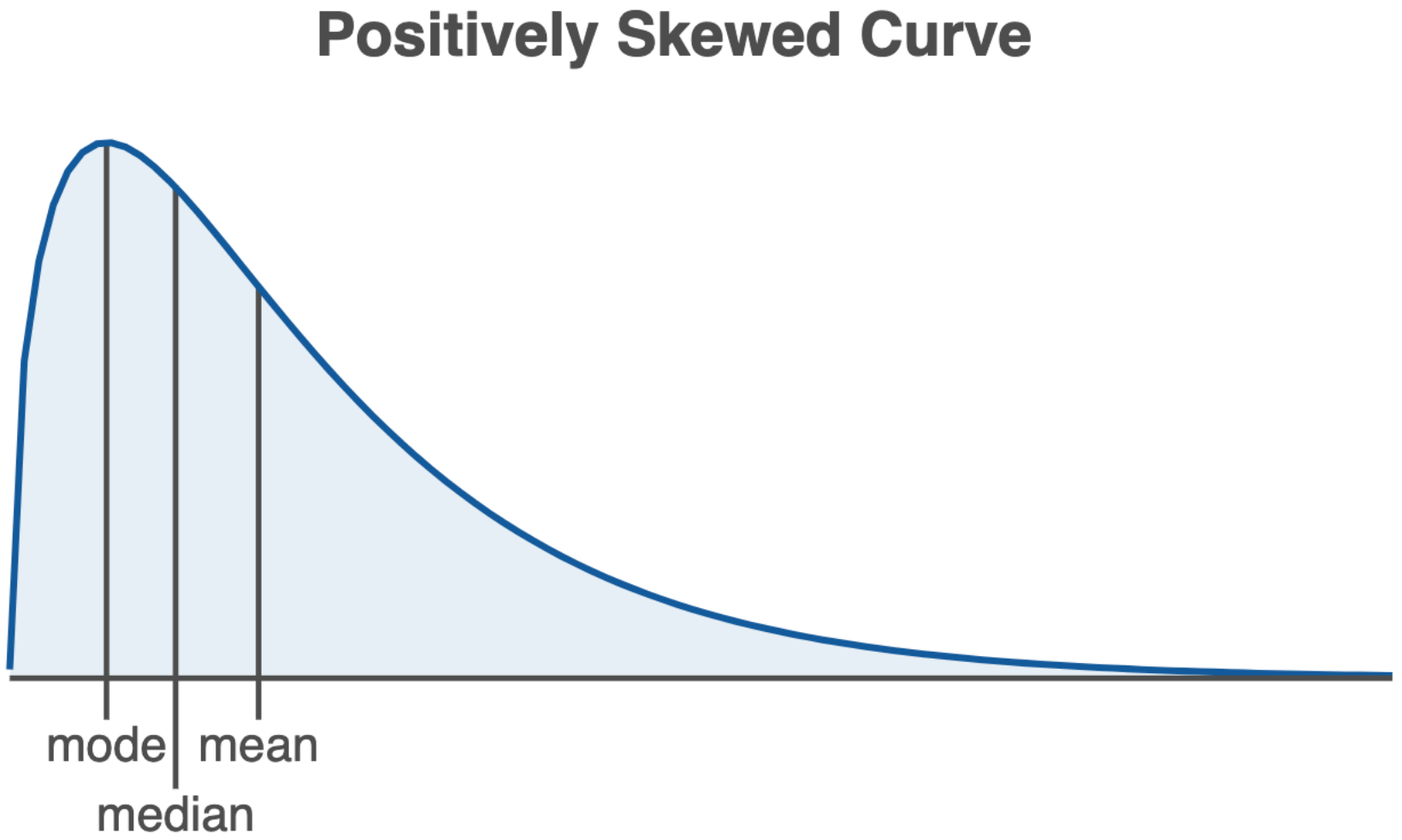
CRITICAL THINKING ABOUT THE MEAN AND MEDIAN

- There is a relationship between mean, the median, and the shape of the data
- If the data are **skewed:**
  - Because the mean is heavily influence by very large (or small) values in the data set relative to the rest of the data, it is usually more appropriate to use the median when describing the center of skewed data.

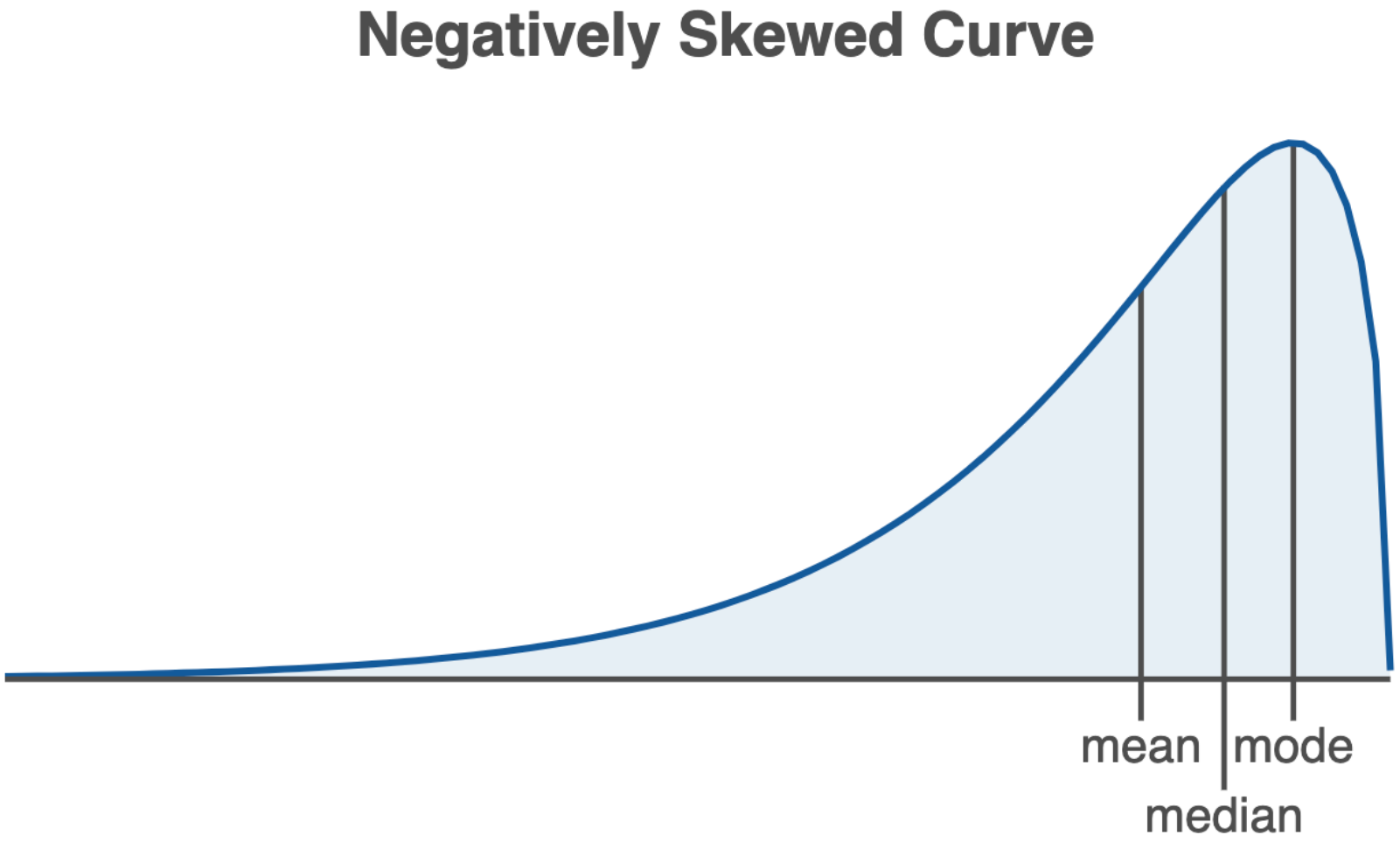




CRITICAL THINKING ABOUT THE MEAN AND MEDIAN



CRITICAL THINKING ABOUT THE MEAN AND MEDIAN



KNOWLEDGE CHECK

COMPREHENSION QUESTION: SPREAD

Which measure(s) of spread would be sensitive to the presence of outliers?

1. Variance
2. Standard deviation
3. IQR
4. Range

COMPREHENSION QUESTION: CENTER

Which measure(s) of center would be sensitive to the presence of outliers?

1. Mean
2. Median
3. Mode

COMPREHENSION QUESTION: STANDARD DEVIATION

A standard deviation can be negative.

- TRUE
- FALSE

COMPREHENSION QUESTION: STANDARD DEVIATION

A standard deviation can be negative.

- TRUE
- FALSE

FALSE, when calculating we square the deviations and the result will always be positive.

COMPREHENSION QUESTION: STANDARD DEVIATION

A standard deviation can be 0.

- TRUE
- FALSE

TRUE, when all values are exactly the same (5, 5, 5, 5) the data set will have zero spread

CONCLUSION

Choosing measures of center and spread:

- Skewed distortions or distributions with extreme outliers
  - Use median and quartiles

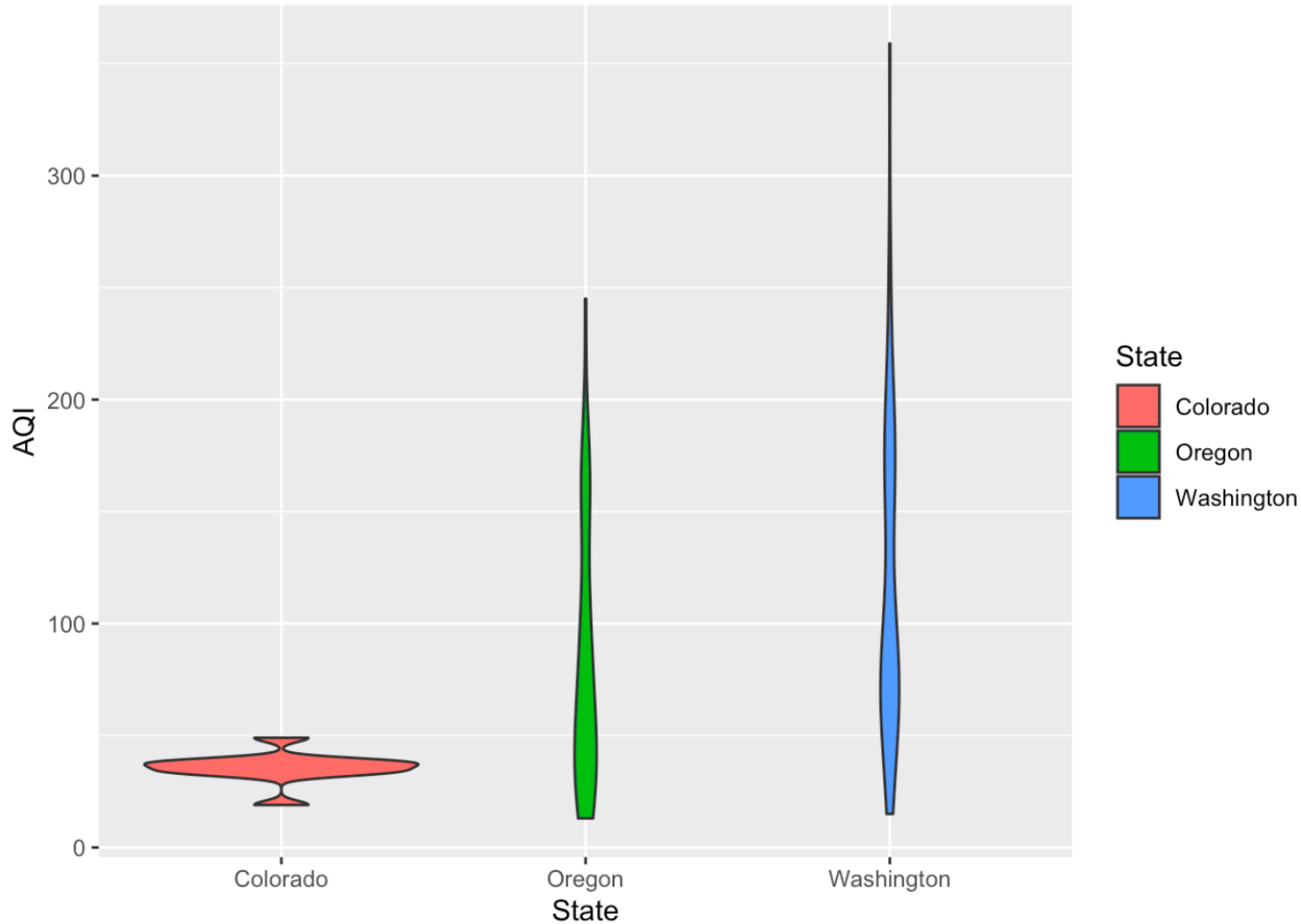
- Approximately symmetric distribution (with no outliers)
  - Use mean and standard deviation

MORE FUN PLOTS!

Step 4: Side-by-side Violin Plots

```
ggplot(aqi, aes(x=State, y=AQI, fill=State))+
  geom_violin()
```

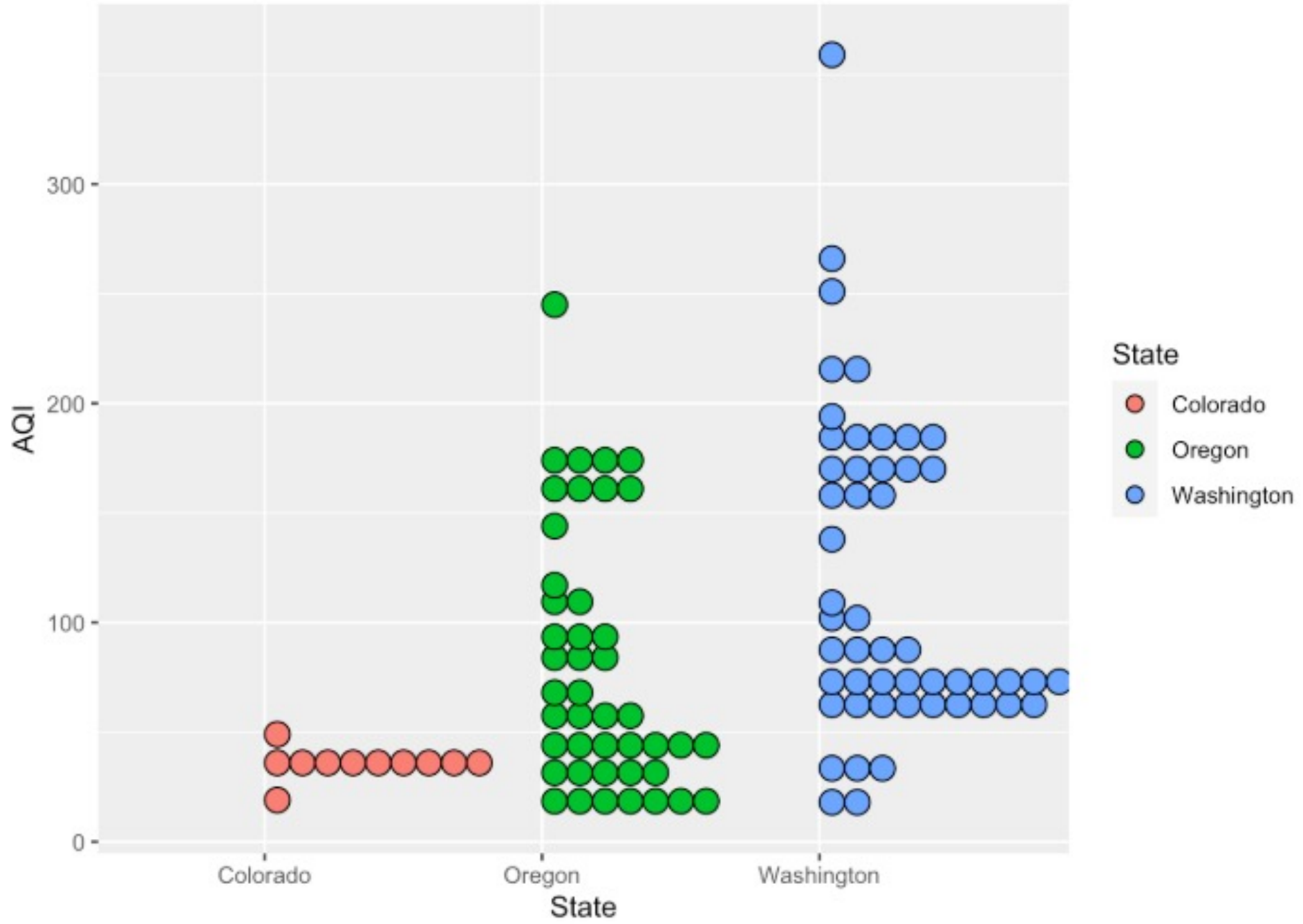
```
ggplot(aqi, aes(x=State, y=AQI, fill=State))+
  geom_violin()
```



Step 5: Dot Plots

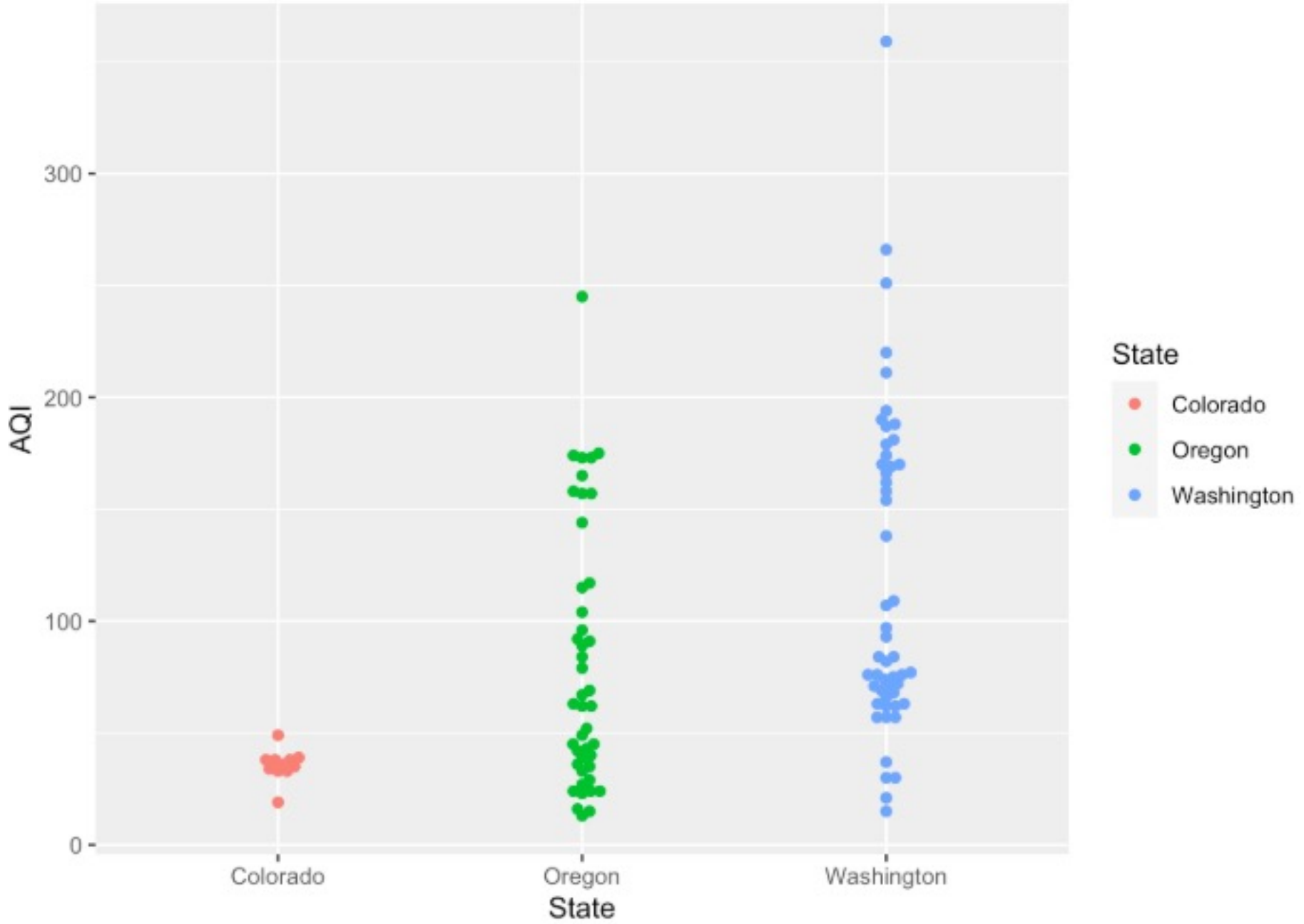
You could make a dot plot!

```
# Dot plot


ggplot(aqi, aes(x=State, y=AQI, fill=State))+
  geom_dotplot(binaxis='y')
```


AQI
300
200
100
0
Colorado
Oregon
Washington
State
State
Colorado
Oregon
Washington

Step 6: Bee swarm Plots

```
#install.packages('ggbeeswarm')
library(ggbeeswarm)

ggplot(aqi, aes(x=State, y=AQI, color=State))+
  geom_beeswarm()
```


AQI
300
200
100
0
Colorado
Oregon
Washington
State
State
Colorado
Oregon
Washington

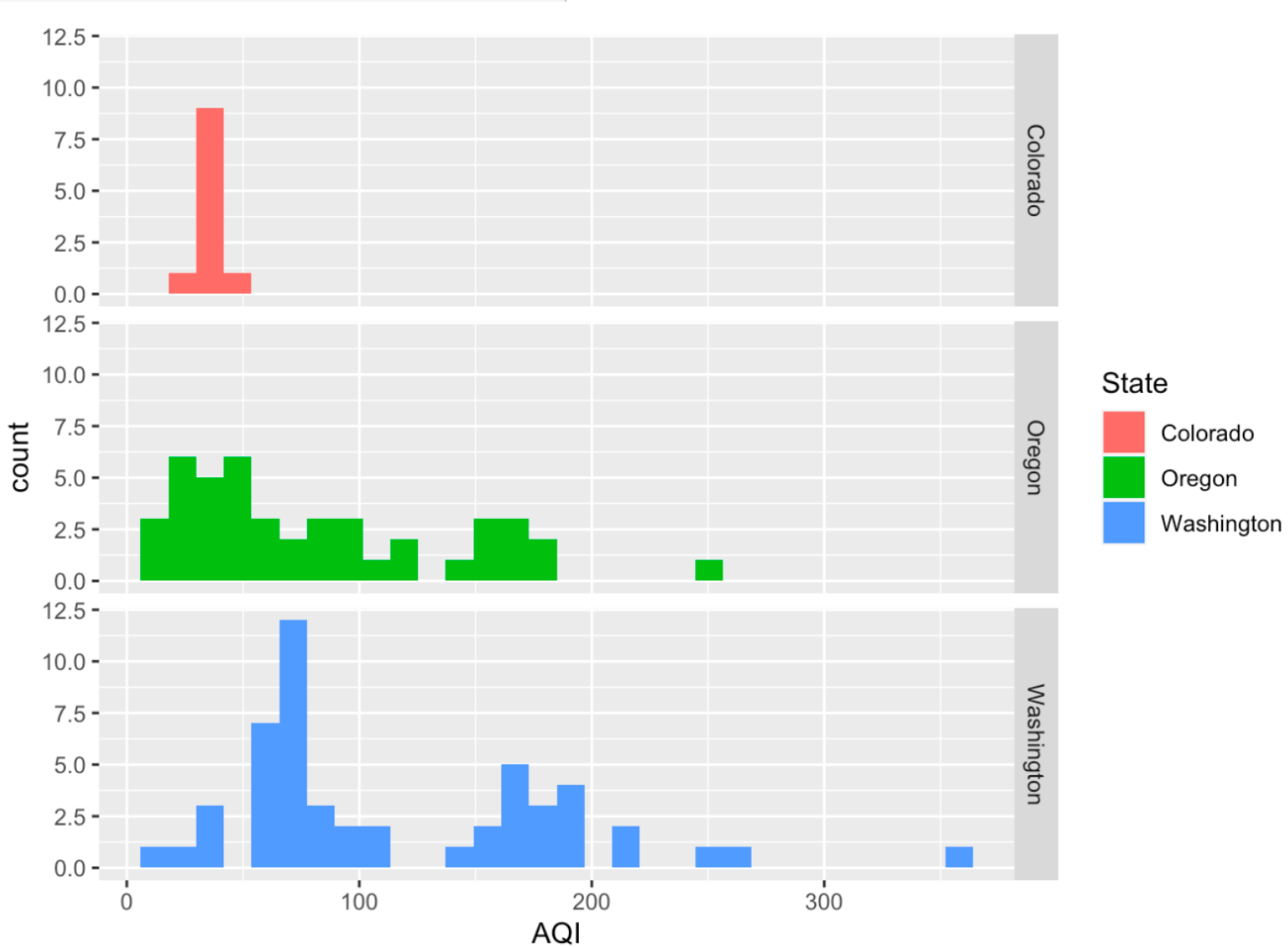
FACETS

Here is a histogram where the y-axis is count.

```
### Histogram (Counts)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_histogram()+
  facet_grid(State~.)
```

Here is a histogram where the y-axis is count.

```
### Histogram (Counts)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_histogram()+
  facet_grid(State~.)
```

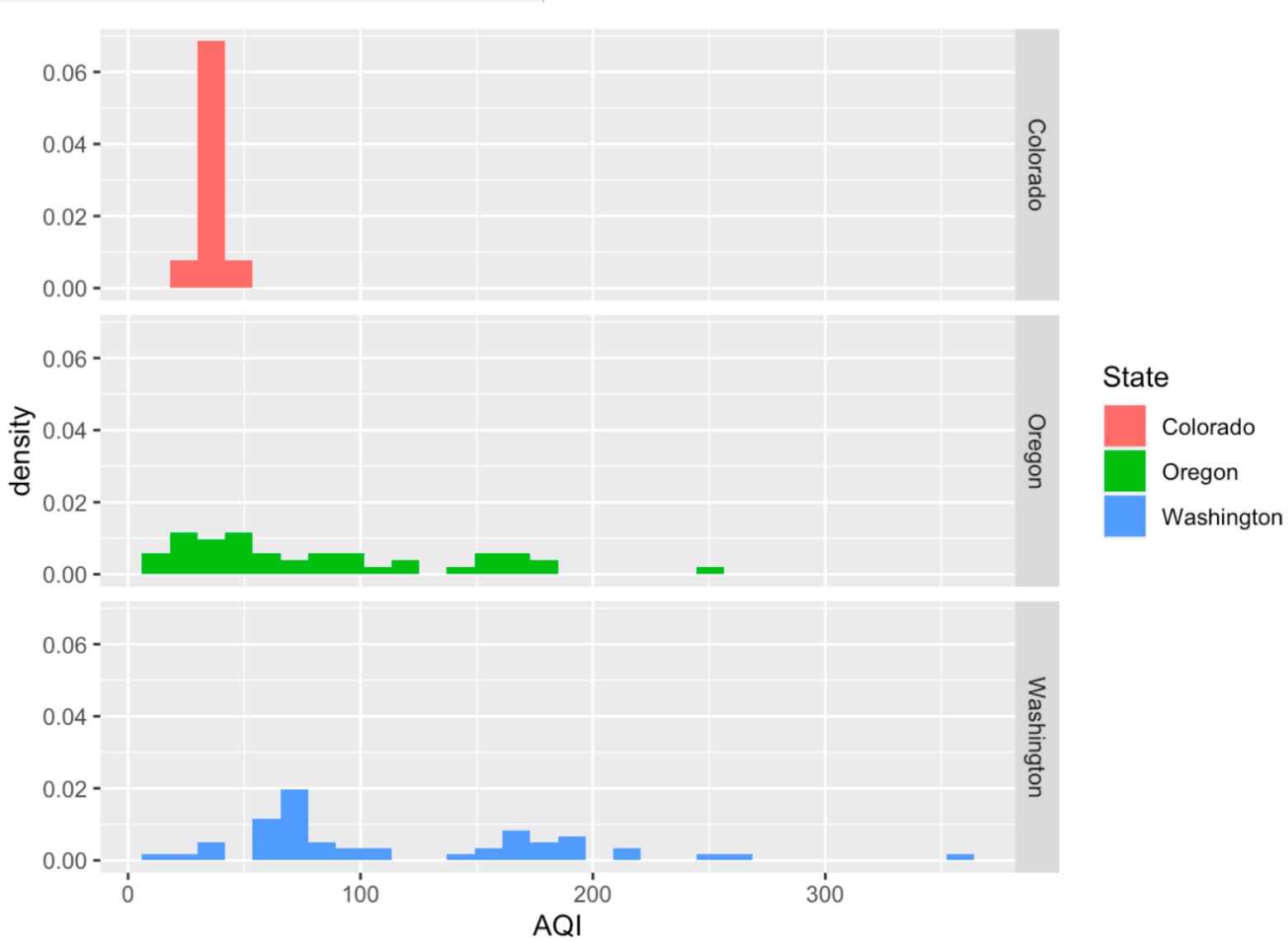


We can change the y-axis to density (proportions).

```
### Histogram (Density)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_histogram(aes(y=..density..))+
  facet_grid(State~.)
```

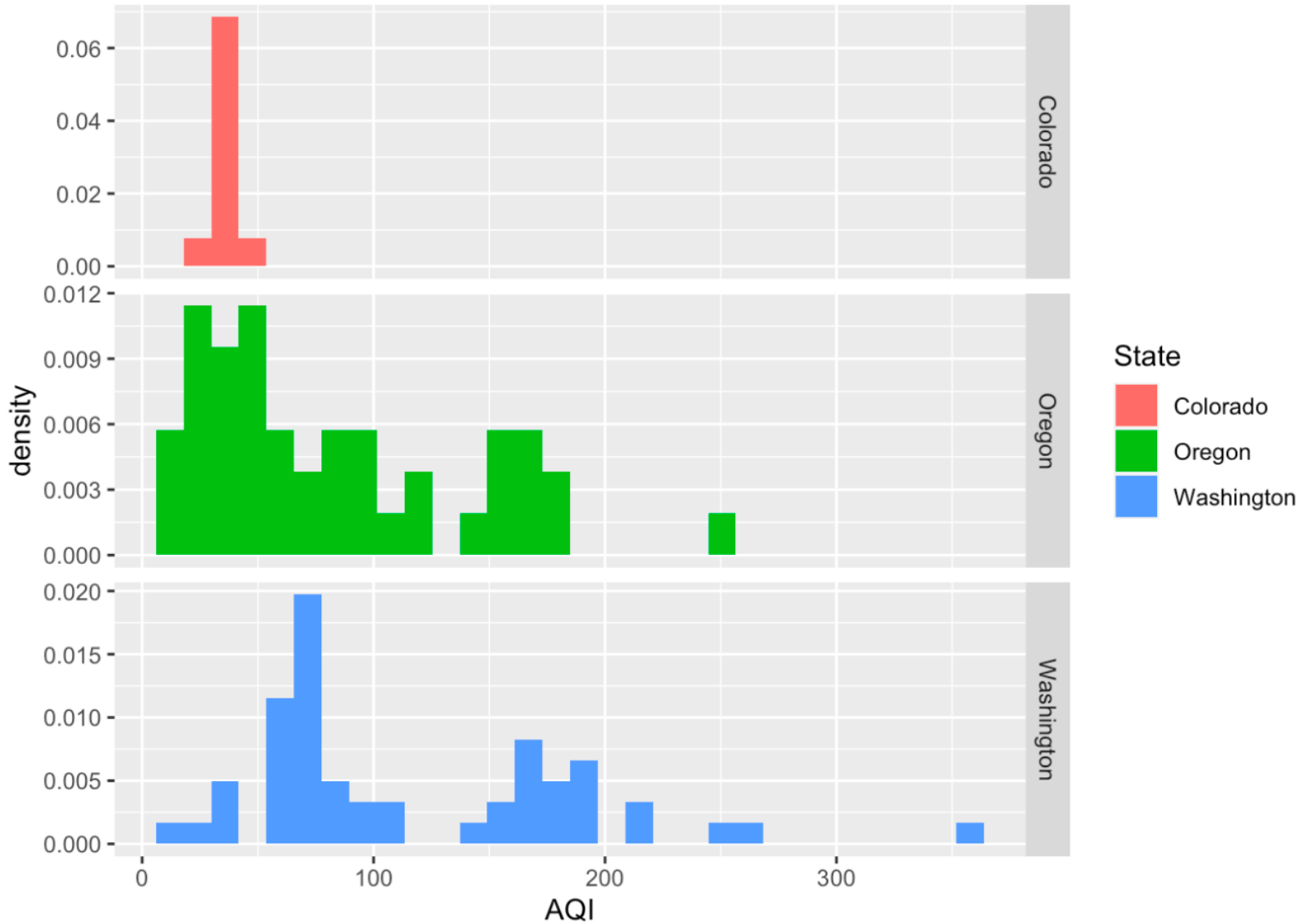
We can change the y-axis to density (proportions).

```
### Histogram (Density)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_histogram(aes(y=..density..))+
  facet_grid(State~.)
```



```
### Histogram (Density - Free_y)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_histogram(aes(y=..density..))+
  facet_grid(State~., scales = "free_y")
```

```r
### Histogram (Density - Free_y)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_histogram(aes(y=..density..))+
  facet_grid(State~., scales = "free_y")
```

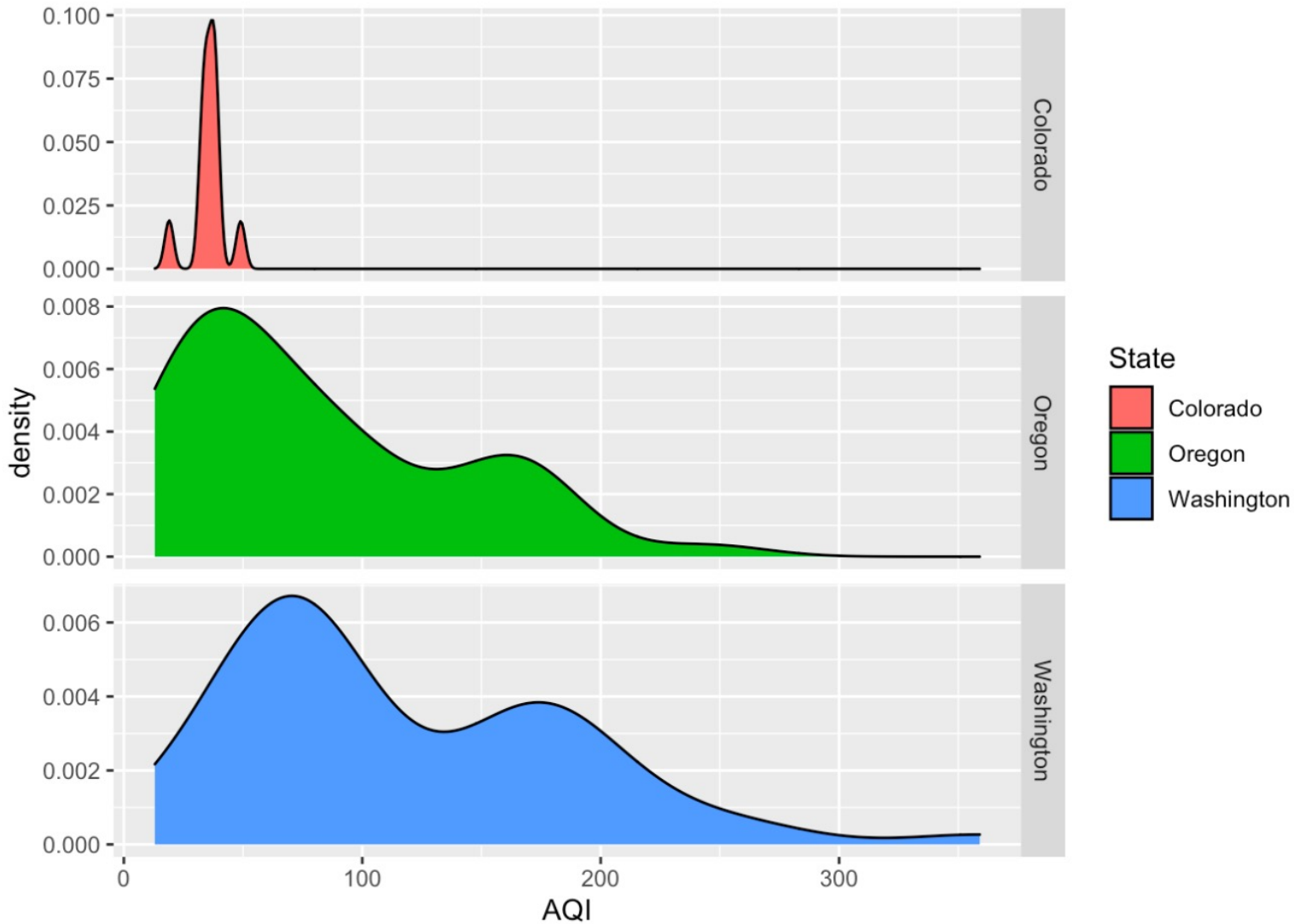


We can also make a density plot!

```
### Density Plot (free_y)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_density()+
  facet_grid(State~., scales = "free_y")
```

We can also make a density plot!

```
### Density Plot (free_y)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_density()+
  facet_grid(State~., scales = "free_y")
```



TIME FOR GROUP WORK

MILESTONE #5

DATA 151: Project Milestone #5

Due 11 - Milestone #5: Exploratory Data Analysis Step #3
Distributions, Summary Statistics, and Comparing Subgroups

Goal: Work to answer at least one of your questions of interest for numeric variables of interest.

- Describe the shape of the distribution for a numeric variable of your choice. Convey the appropriate summary statistics
- Explore possible subgroups

Please submit using Rmarkdown