Welcome to DATA 151

I'm so glad you're here!

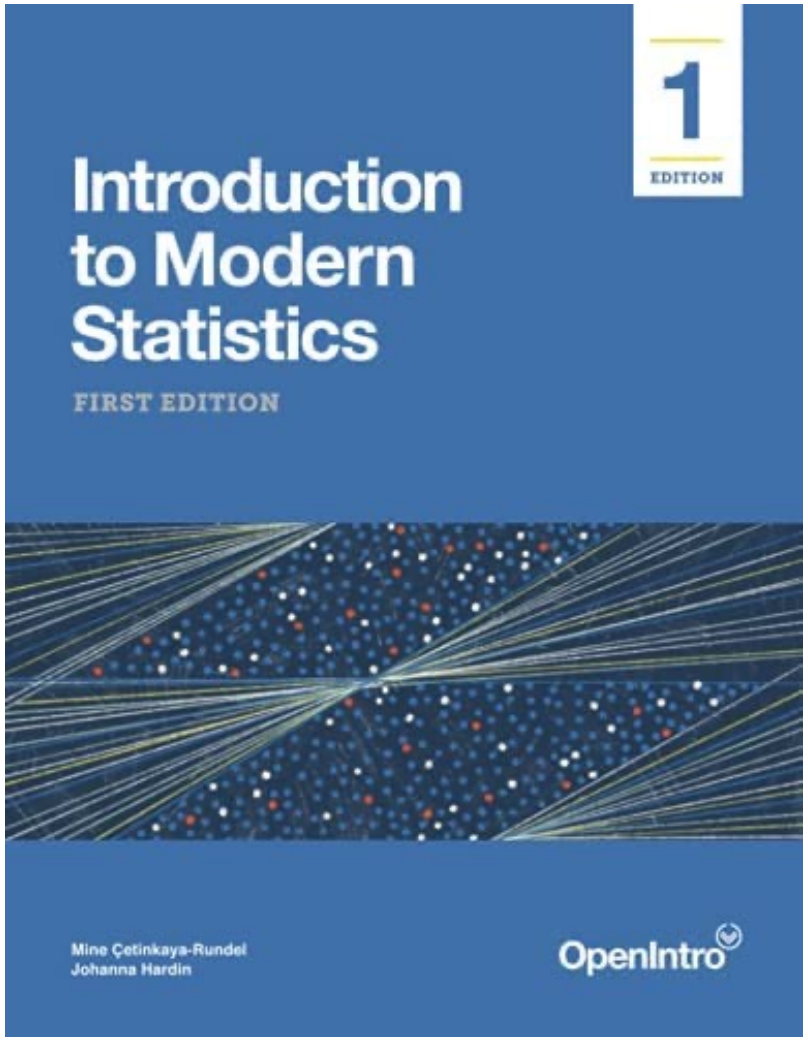
DATA 151: CLASS 8B
INTRODUCTION TO DATA SCIENCE (WITH R)

NUMERIC DATA ANALYSIS: SUMMARY VALUES AND GRAPHS

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING



Introduction to Data Science:

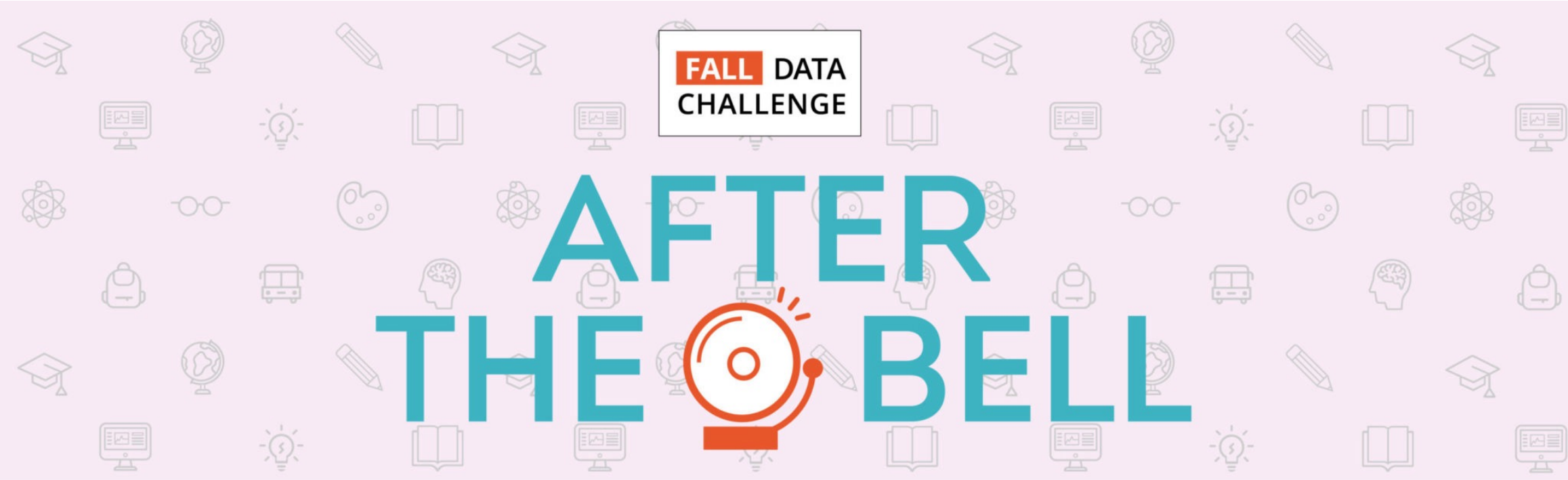
- Tuesday:
  - Introduction to Modern Statistics
  - Ch 4: Exploring Categorical Data
- Thursday:
  - Introduction to Modern Statistics
  - Ch 5: Exploring Numeric Data

HOMEWORK REMINDER

Due next 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***

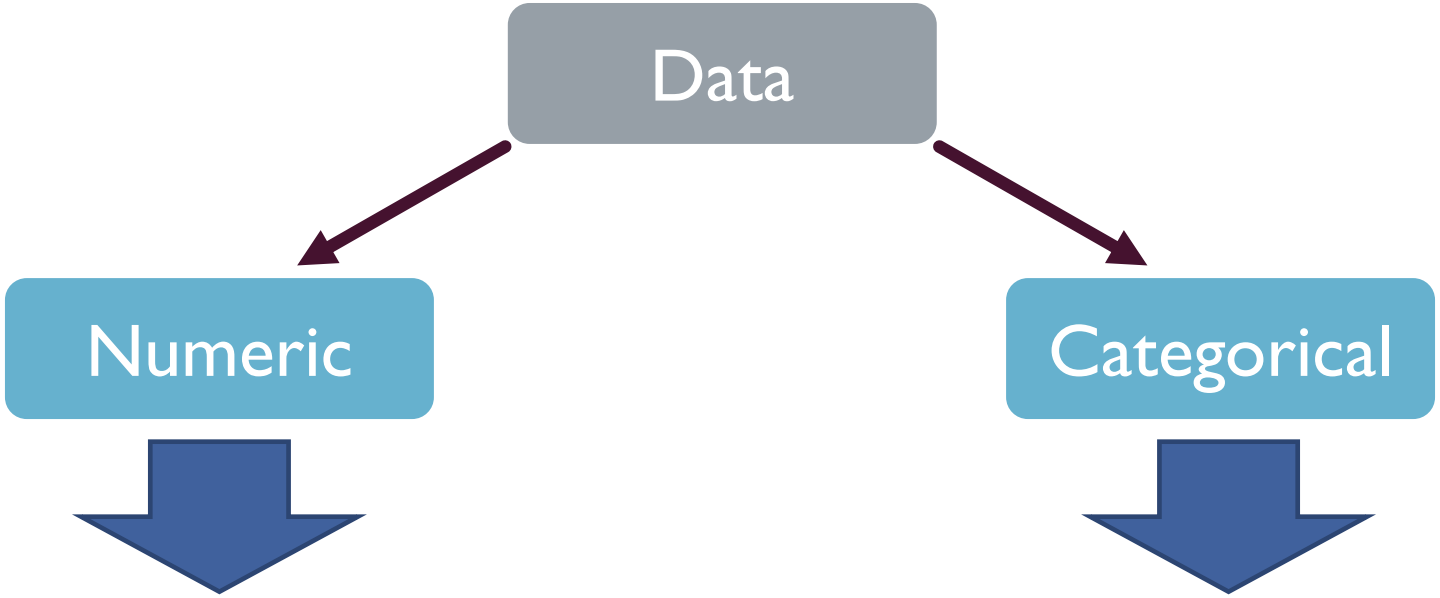
ANOTHER EXTRA CREDIT OPPORTUNITY



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.

UNIVARIATE GRAPHICAL TOOLS



- Stem-and-Leaf plots
- Histograms
- Density (approximation) plot
- Boxplot / Box-and-whisker plot

- Bar graphs
- Pie charts

WHAT TO DO WHEN YOU HAVE CATEGORICAL DATA

The distribution of a categorical variable lists the categories and provides the count or percent of individuals who fall into each category.

- **Tables**
  - Shows the counts/proportion of observations that fall within a category or categories
  - Ex: Joint, marginal, conditional distributions
- **Bar Graphs**
  - Represent each category as a bar whose heights show the category counts or percent.
- **Pie Charts**
  - Show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percent for the categories relative to the whole
  - ***NOTE: Statisticians avoid using pie charts because differences in angles are difficult for the view to perceive. We want to elucidate the story that the data is telling, not muddle it.***

EXPLORING YOUR NUMERIC DATA

WHAT TO DO WHEN YOU HAVE NUMERIC DATA

The distribution of a numeric variable tells us what values the variable takes on and how often it takes those values. If you have a numeric variable, you have a couple of options:

- **Stem-and-Leaf Plot**
  - A tabular visualization where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit)
  - Great for small data sets!
- **Histogram**
  - Diagram consisting of rectangles whose base is "binned" into intervals and whose height is relative to the frequency of observations in a given "bin"
  - Great for larger data sets!

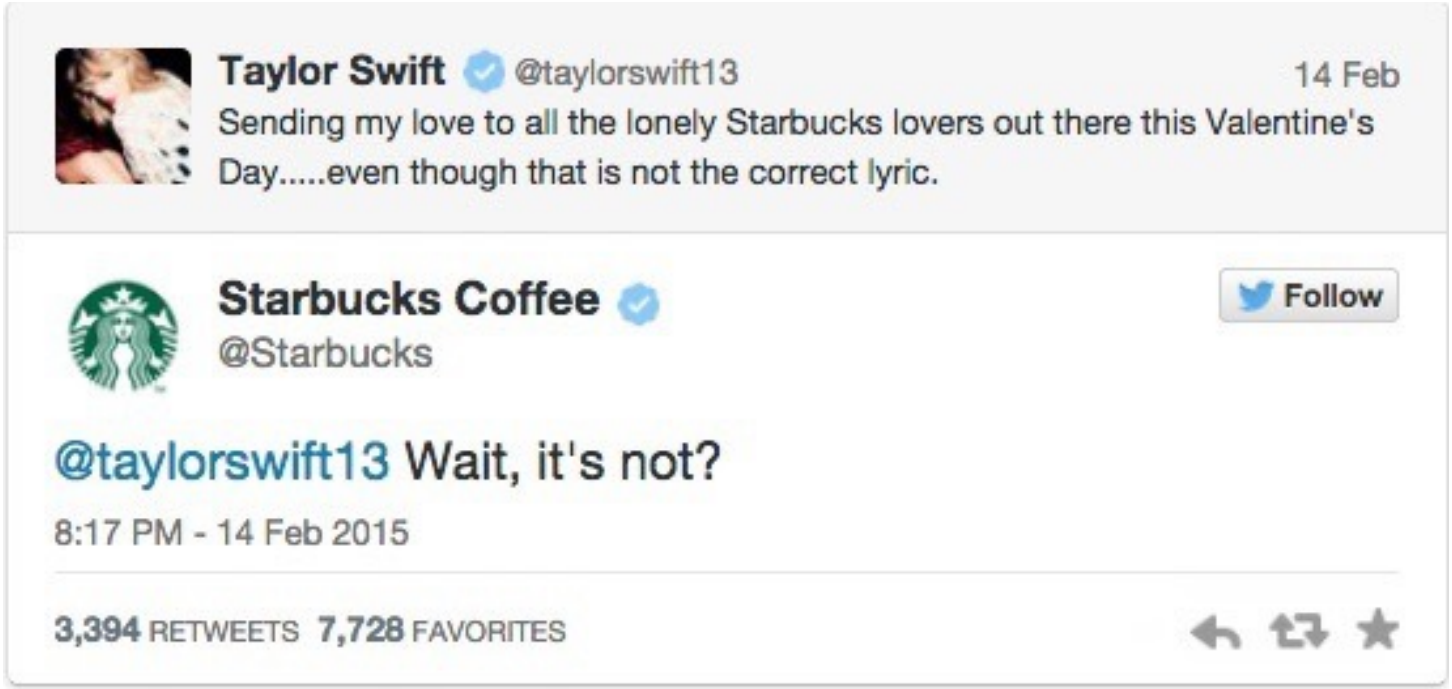
WHAT TO DO WHEN YOU HAVE NUMERIC DATA

More…

- **Density (approximation) Plot**
  - Smooths/approximates a continuous distribution for a numeric variable
- **Boxplot / Box-and-Whisker Plot**
  - Displays the five-number summary of a set of data; minimum, first quartile, median, third quartile, and maximum

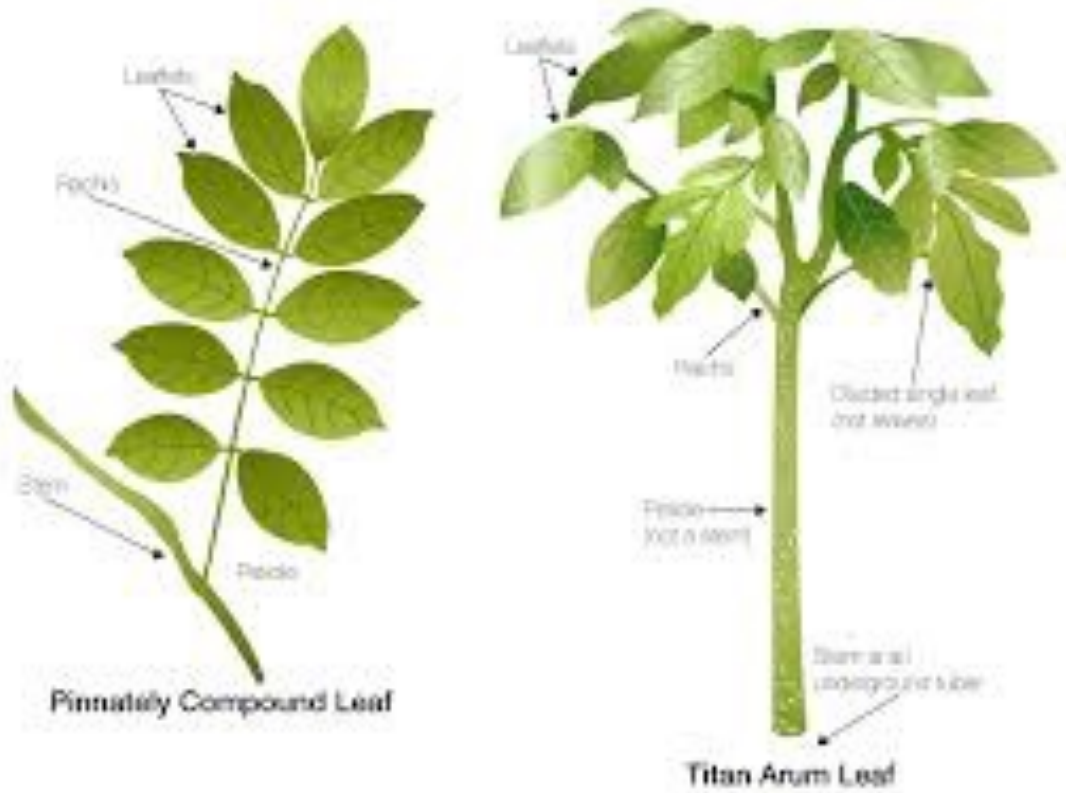
EXAMPLE #1: STARBUCKS LOVERS

Let's start with a small data set so we can learn the concepts.



EXAMPLE #1: STARBUCKS LOVERS

	product_name	milkName	size	calories	sugar_g
1	Chai tea Latte	2%	grande	240	42
2	Caramel Frappuccino Blended	2%	grande	270	60
3	London Fog Tea Latte	Nonfat	short	80	15
4	Iced Caffè Mocha	Soy	grande	210	24
5	Caffè Mocha	Soy	grande	330	28
6	Iced Coffee with milk	Nonfat	tall	80	17
7	Iced Black tea	None	venti	120	30
8	Chocolate Smoothie	Coconut	grande	290	32
9	Iced Coffee	None	tall	60	15
10	brewed coffee – True North Blend Blonde roast	None	tall	4	0
11	Iced White Chocolate Mocha	Coconut	tall	310	39
12	Caffè Latte	Whole	venti	290	22



STEM-AND-LEAF PLOTS

STEM-AND-LEAF PLOTS

- **Step 1:** Re-write values in numerical order

- **Step 2:** Separate each observation into a stem and a leaf
  - **Stem:** all but the right most digit (usually)
  - **Leaf:** the final digit of each observation (usually)

STEM-AND-LEAF PLOTS

- **Step 3:** Write the stems in a vertical column in numerical order beginning with the smallest (at the top)
  - Note: include *all* integer values between your smallest and largest stems, where or not they appear in the data set.

- **Step 4:** Draw a vertical line directly to the right of the column with the stems.

STEM-AND-LEAF PLOTS

- **Step 5:** Write each leaf in the row to the right of its stem, in increasing order.
  - Keep the leaves in neat columns
  - Write all leaves the same size (Recall: the area rule)

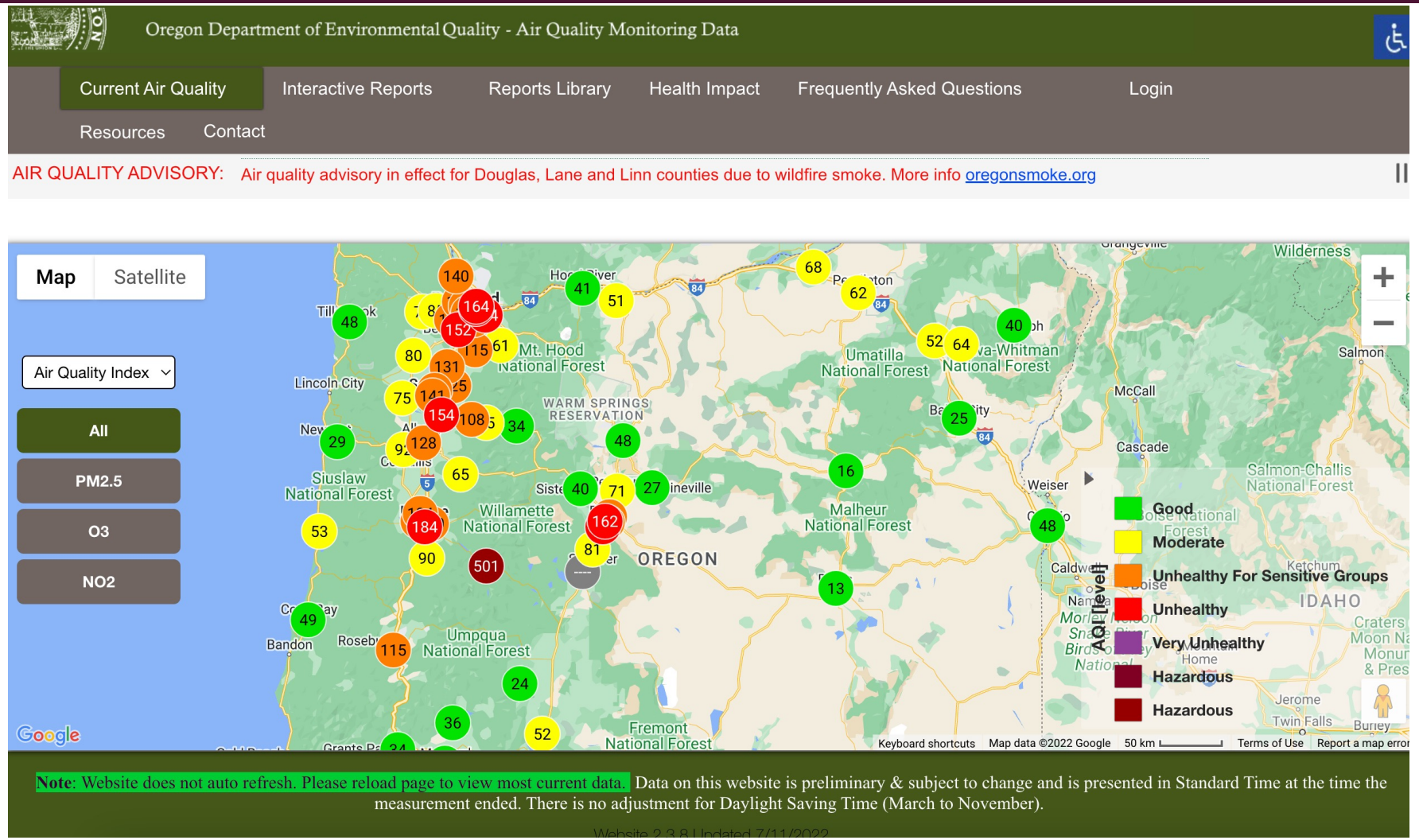
- **Step 6:** Give the graphical display a title and a legend (write the units)

TRY THE PROBLEM ON THE WORKSHEET

STEM-AND-LEAF PLOTS

Data	0	15	15	17	22	24	28	30	32	39	42	60
Stem	0	1	1	1	2	2	2	3	3	3	4	6
Leaf	0	5	5	7	2	4	8	0	2	9	2	0

EXAMPLE #2:OREGON AQI



Air Quality Index information

Inbox ×

Don Thomson, Associate Dean for Health & Well-Being dthomson@willamette.edu via e2ma.net 11:02 AM (46 minutes ago)

to me

Dear Willamette Community,

Smoke from area wildfires has created air quality concerns across the state, including in Portland and Salem. Willamette's campuses are far from the fires and are not in any danger, but we write to alert our community of deteriorating air quality that is expected to last through Friday.

Please take care to protect yourselves from the health hazards associated with smoky air. You can check the Air Quality Index in your area at Oregon's DEQ website. Willamette Emergency Medical Services (WEMS) also has an air quality index monitor for Portland and Salem on their web page.

Poor air quality most notably affects pregnant people, children, older adults & people with chronic health conditions. Sensitive groups should stay indoors and watch for coughing and shortness of breath.

- Additional recommendations for everyone include:
- Avoid strenuous outdoor activity.
- Keep windows and doors closed to reduce the smoke that enters your room or home.
- If you have an HVAC system with a fresh air intake, set the system to recirculate mode, or close the outdoor intake damper.

For Salem student athletes, Athletics training staff are actively monitoring the air quality and will communicate with teams regarding practices this week as appropriate.

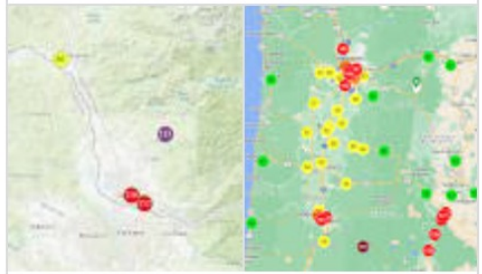
More information, including recommendations for people with health conditions can be found in the DEQ's "A Guide to Air Quality and Your Health".

Sincerely,
Don Thomson
Associate Dean for Health and Well-Being

https://twitter.com/search/oregon+air+quality

oregon air quality on Twitter

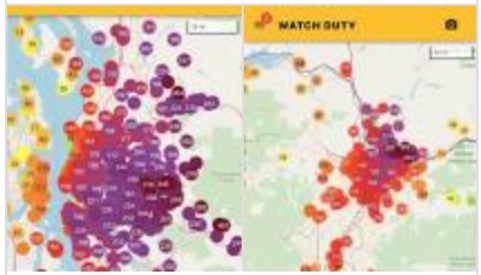
NWS Portland
@NWSPortland



Per @EcologyWA and @OregonDEQ sensors, air quality has deteriorated into the unhealthy for everyone category across much of the Portland/Vancouver metro & the Eugene/Springfield metro. Their recommendations are to close windows and limit outdoor activities. #pdxtst #orwx #wawx

Twitter · 23 hours ago

Michael Stein…
@MichaelWX18



Air quality in the Pacific Northwest is downright awful this morning. Way into the hazardous zone for the #Seattle metro, all the way down to #Portland and the Willamette Valley in Oregon. Pattern change is on the horizon however, and will clear out the air by tomorrow.

Twitter · 4 hours ago

Oregon Smok…
@ORSmokeInfo

Today, @OregonDEQ issued an #AirQuality advisory for Columbia, Clackamas, Marion, Multnomah & Washington Counties due to smoke from the Nakia Creek & Siouxon fires in Washington, and the Cedar Creek fire near Oakridge.
Details:
www.oregonsmoke.org
Photo: Nakia Creek Fire 1/2

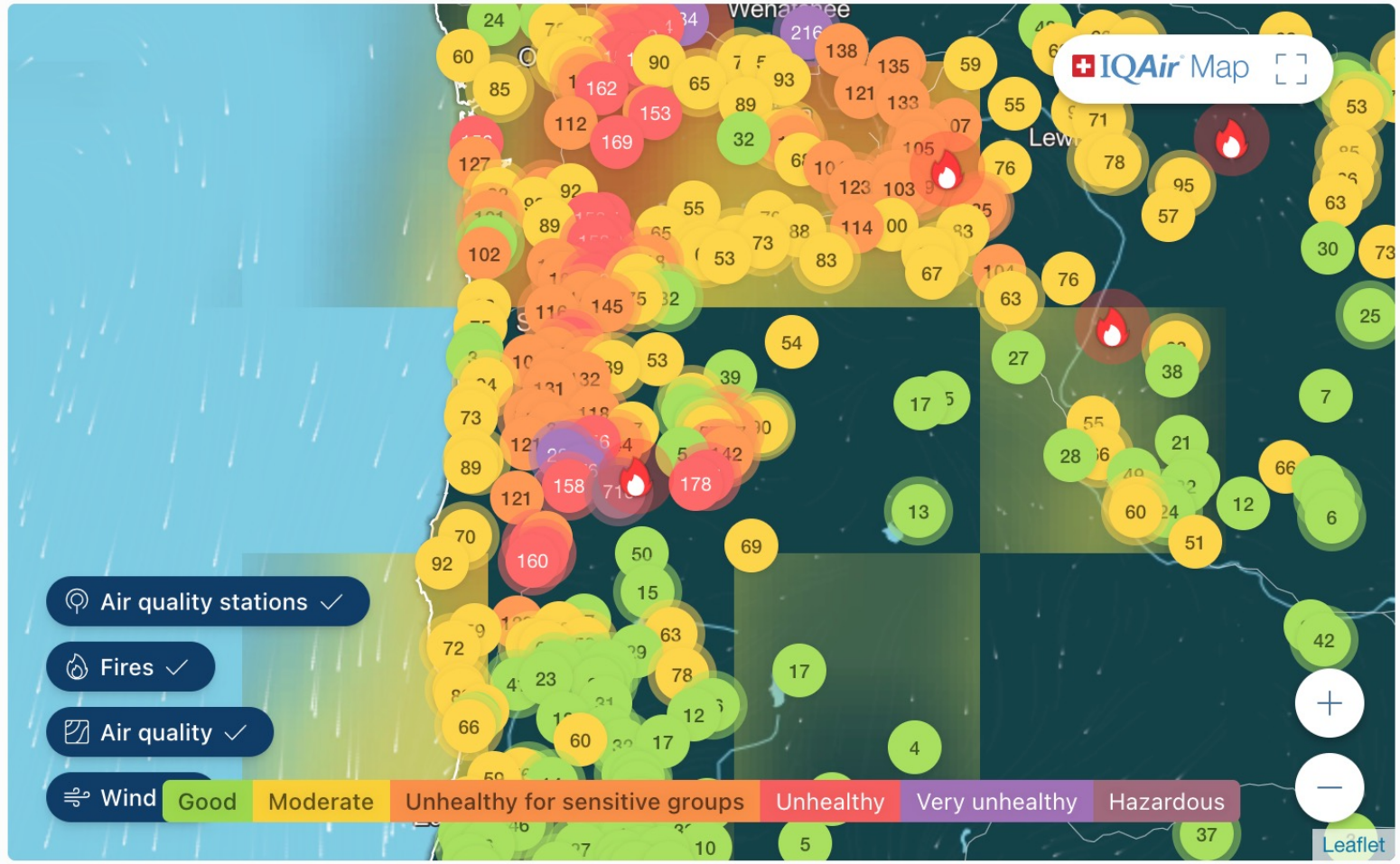
Twitter · 58 mins ago

LIVE AQI CITY RANKING

Real-time Oregon Most polluted city ranking

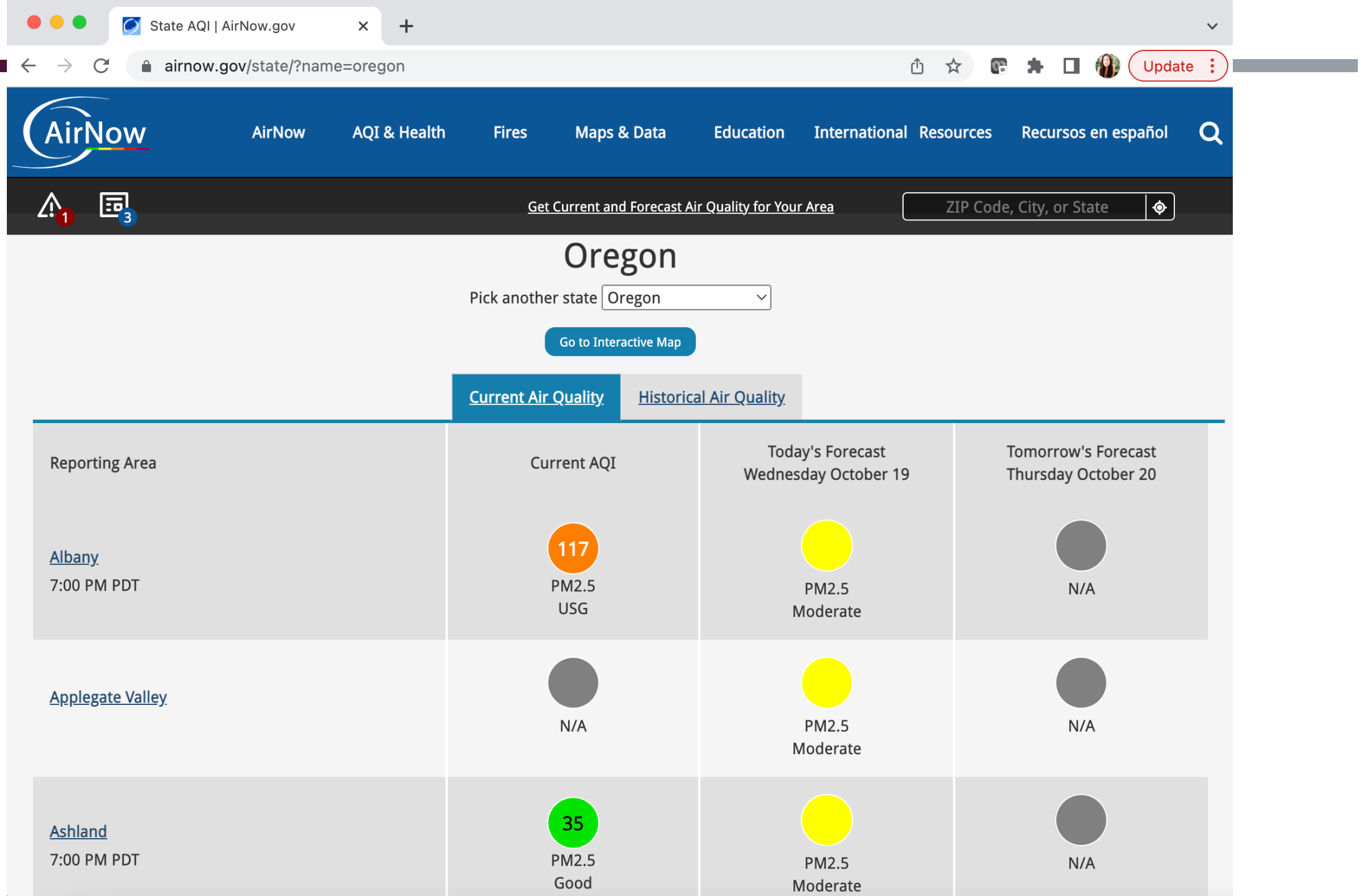
#	CITY	US AQI
1	Oakridge	721
2	Lowell	583
3	Eugene	193
4	Springfield	175
5	Oak Hills	174
6	Creswell	169
7	Deschutes River Woods	169
8	Coburg	168
9	Troutdale	168
10	Marion County	164

11:51, Oct 19 (local time)



How to best protect from air pollution?

Reduce your air pollution exposure in Oregon


State AQI | AirNow.gov
airnow.gov/state/?name=oregon
Update
AirNow
AirNow
AQI & Health
Fires
Maps & Data
Education
International
Resources
Recursos en español
Get Current and Forecast Air Quality for Your Area
ZIP Code, City, or State
Oregon
Pick another state
Oregon
Go to Interactive Map
Current Air Quality
Historical Air Quality
Reporting Area
Current AQI
Today's Forecast
Wednesday October 19
Tomorrow's Forecast
Thursday October 20
Albany
7:00 PM PDT
117
PM2.5
USG
PM2.5
Moderate
N/A
Applegate Valley
N/A
PM2.5
Moderate
N/A
Ashland
7:00 PM PDT
35
PM2.5
Good
PM2.5
Moderate
N/A

Example : Oregon Air Quality Index (AQI)

These data were reported on AirNow on October 19, 2022 for the states of Oregon, Washington, and Colorado.

https://www.airnow.gov/state/?name=oregon

Step 0: Library Tidyverse

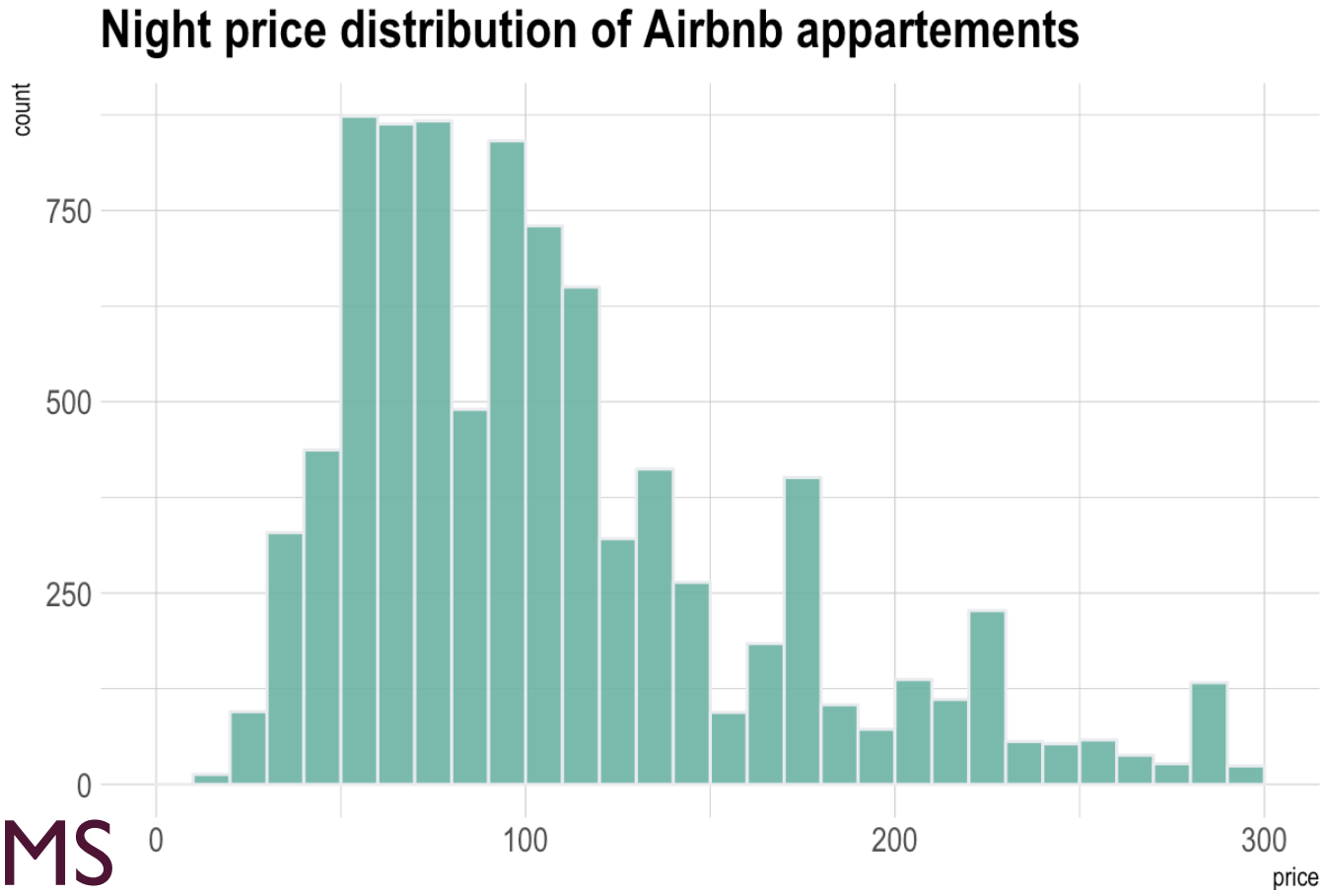
```
library(tidyverse)
```

Step 1: Load the Data

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

orAQI<-aqi%>%
  filter(State=="Oregon")
```

HISTOGRAMS


Night price distribution of Airbnb appartements
count
750
500
250
0
0
100
200
300
price

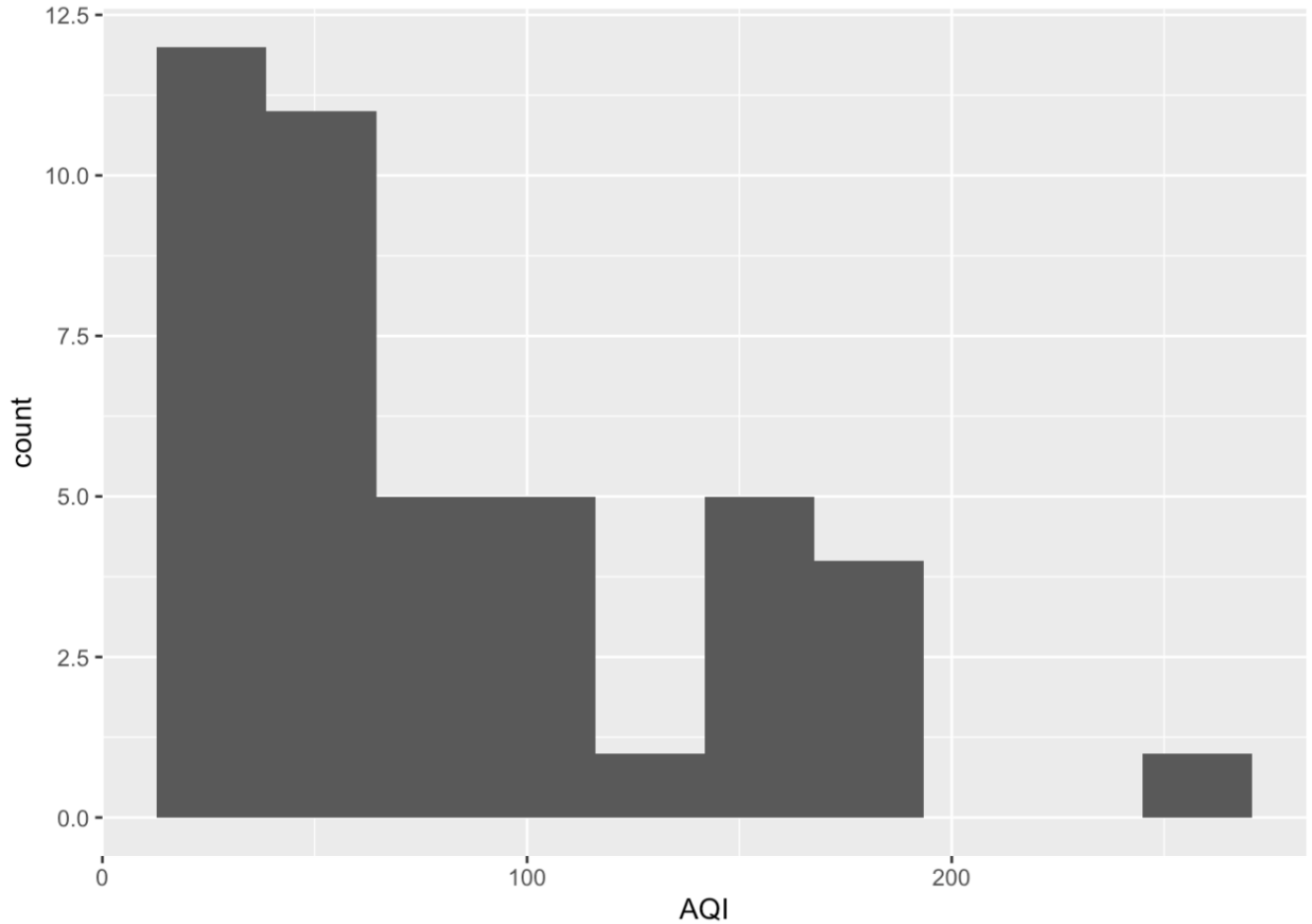
HOW TO CONSTRUCT A HISTOGRAM

- Since quantitative variables often take many values, we must first divide the possible values into classes of equal widths.
- Count how many observations fall into each class.
- Draw a picture representing the distribution
  - Bar heights are equivalent to the number of observations in each intervals

Step 2: Histogram

```
ggplot(orAQI, aes(x=AQI))+
  geom_histogram(bins=10)
```

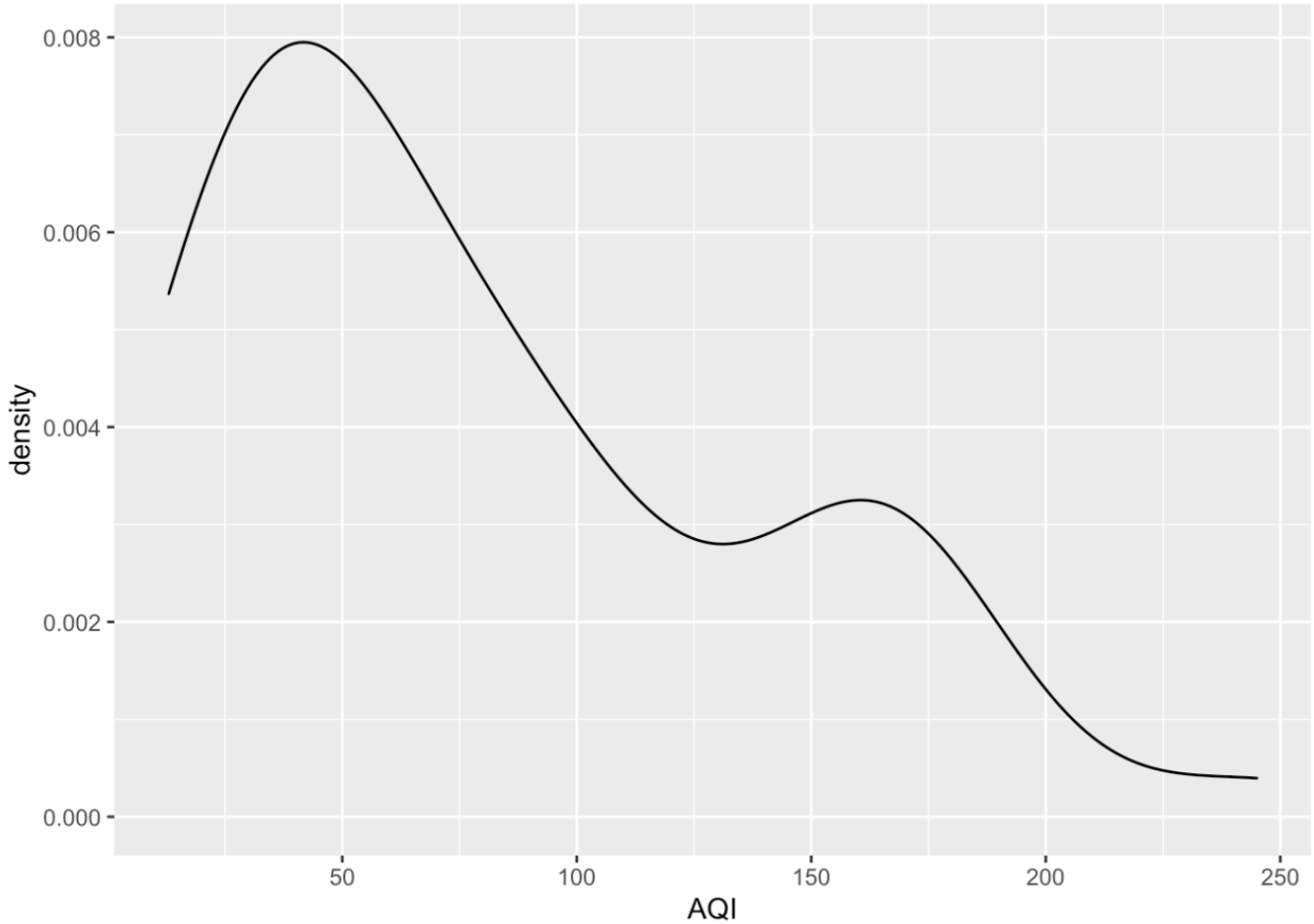
```
## Warning: Removed 3 rows containing non-finite values (stat_bi
n).
```



Step 3: Density Plot

```
ggplot(orAQI, aes(x=AQI))+
  geom_density()
```

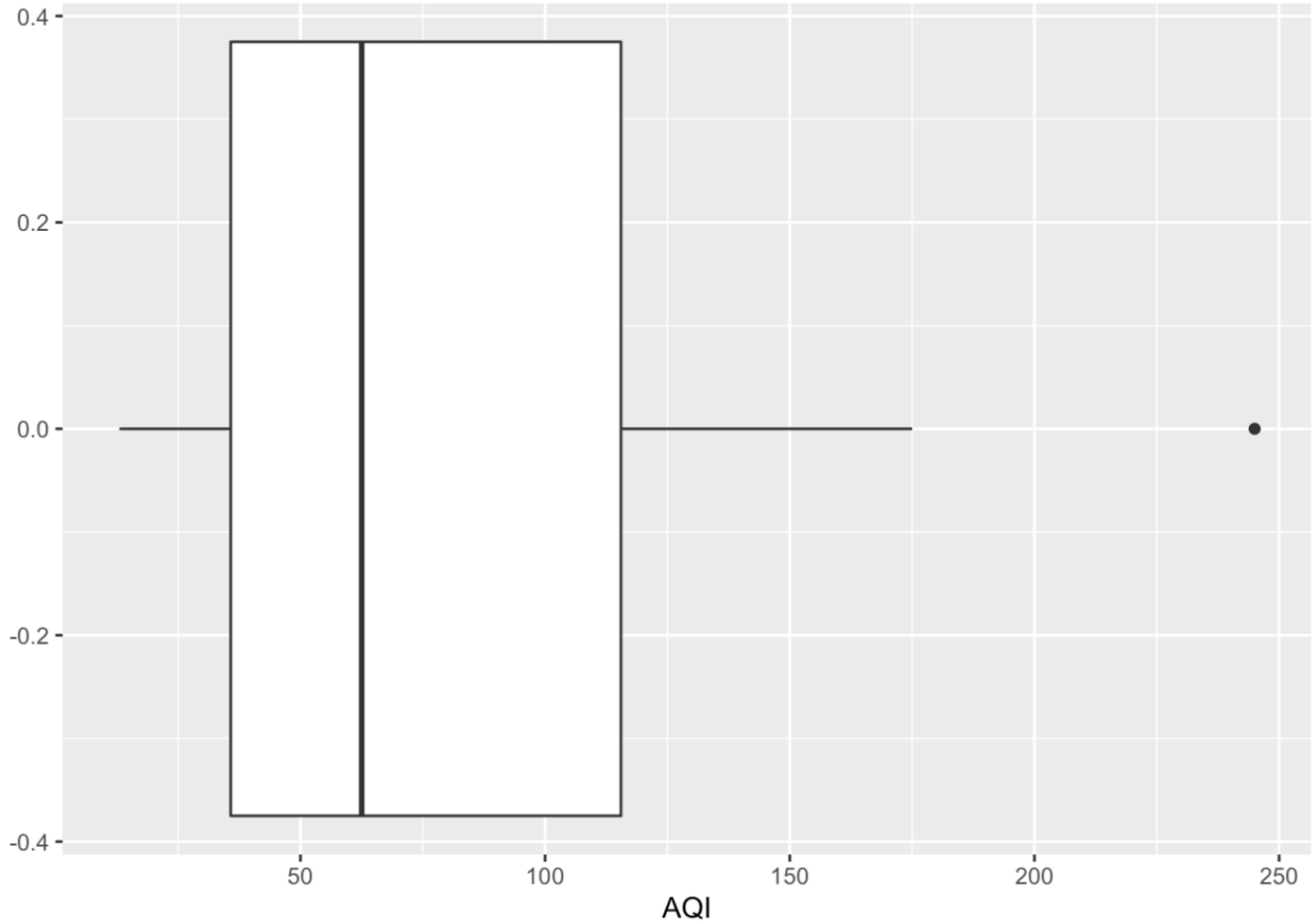
```
## Warning: Removed 3 rows containing non-finite values (stat_den
sity).
```



Step 4: Box Plot

```
ggplot(orAQI, aes(x=AQI))+
  geom_boxplot()
```

```
## Warning: Removed 3 rows containing non-finite values (stat_box
plot).
```



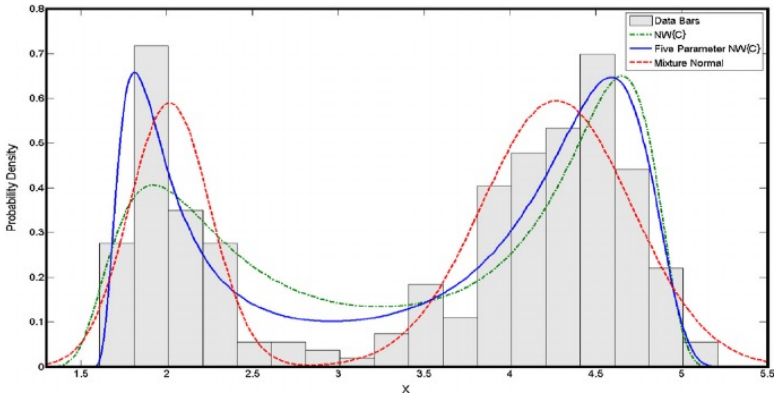
DESCRIBING DATA CHARACTERISTICS OF NUMERIC DATA

- **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

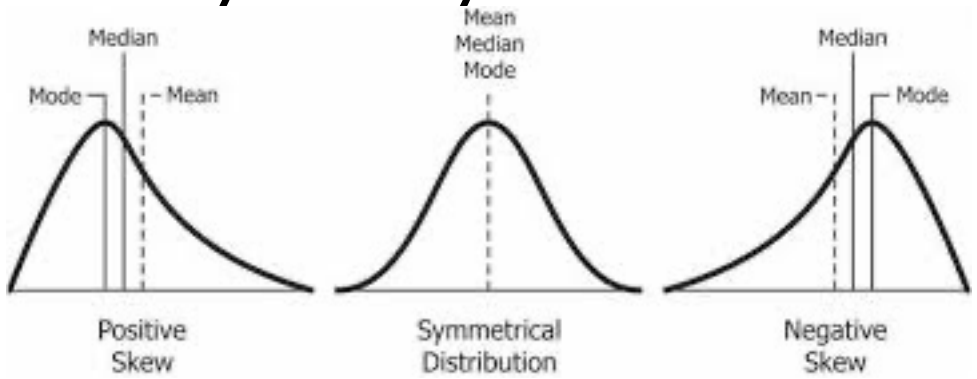
DESCRIBING SHAPE

- **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



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

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.

Step 5: Quantiles

Quantiles split up a data set into four even parts given a relative ordering.

```
## This will give the five number summary
summary(orAQI$AQI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   13.00   35.75   62.50   81.02  115.50  245.00       3
```

```
## If we only want a given quantiles use
quantile(orAQI$AQI, 0.25, na.rm = TRUE)
```

```
##   25%
## 35.75
```

BOX PLOTS

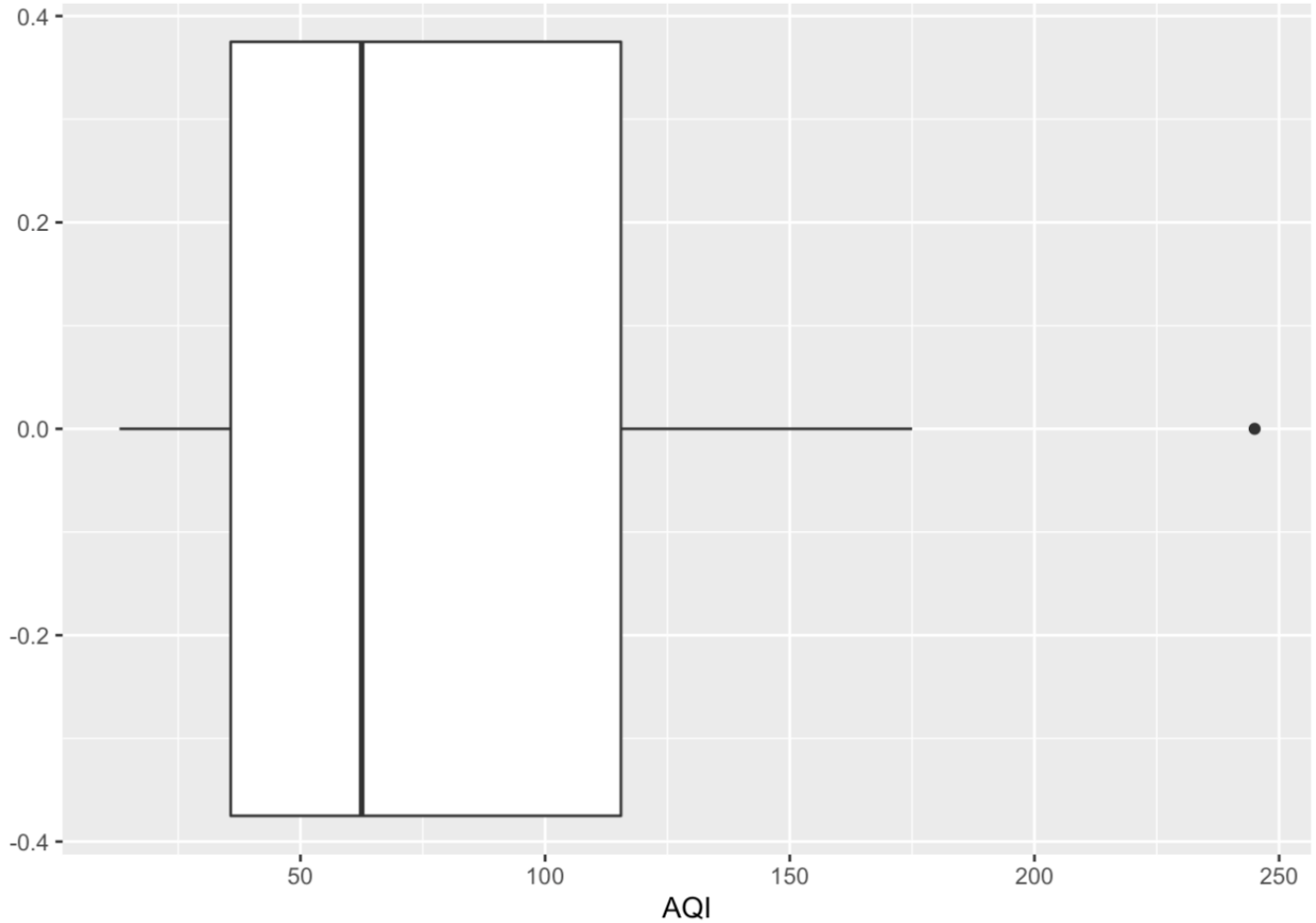
- The **five-number summary** is illustrated in a boxplot

- 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*

Step 4: Box Plot

```
ggplot(orAQI, aes(x=AQI))+
  geom_boxplot()
```

```
## Warning: Removed 3 rows containing non-finite values (stat_box
plot).
```



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: **$IQR = Q_3 - Q_1$**

IQR (Interquartile range)

```
q1<-35.75
q3<-115.50
iqr<-q3-q1
iqr
```

```
## [1] 79.75
```

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

Defining Outliers

We create “fences” to highlight possible outliers in our data.

A data point is highlighted as an outlier if

- it is greater than $Q_3 + 1.5 \times IQR$
- it is less than $Q_1 - 1.5 \times IQR$

```
## upper fence
upper<-q3+1.5*iqr
upper
```

```
## [1] 235.125
```

```
## lower fence
lower<-q1-1.5*iqr
lower
```

```
## [1] -83.875
```

Can we find any outliers?

```
orAQI%>%
  filter(AQI < lower | AQI > upper)
```

```
##    State      City AQI           Level
## 1 Oregon Oakridge 245 Very Unhealthy
```

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

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.

Step 6: Popular Numeric Summaries

Sample Mean

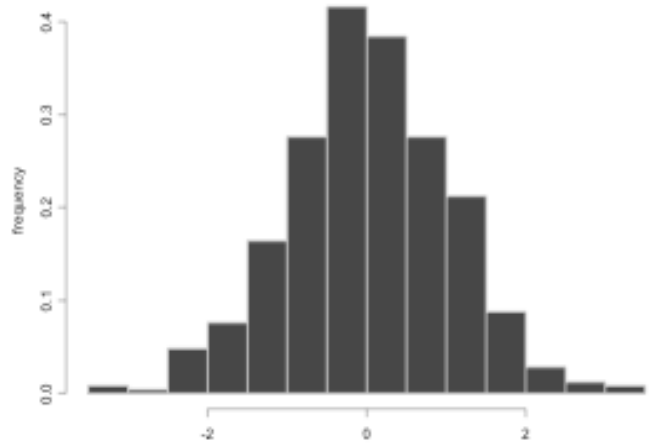
The average (or mean) is the most commonly used metric for center.

```
mean(orAQI$AQI, na.rm=TRUE)
```

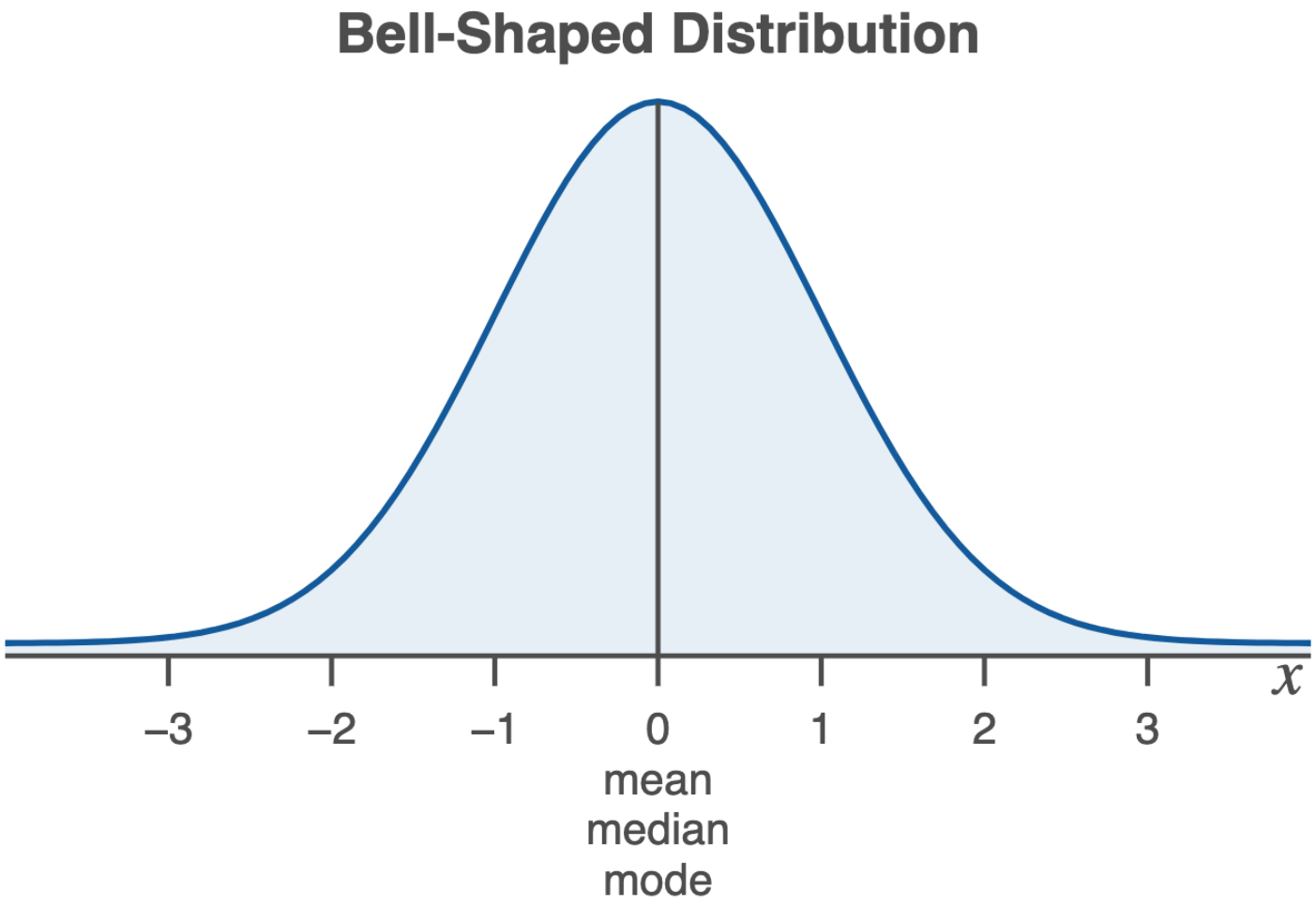
```
## [1] 81.02273
```

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.

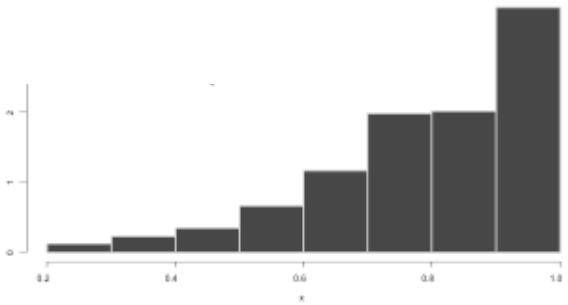


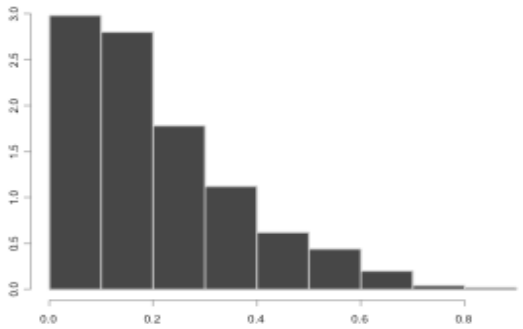
CRITICAL THINKING ABOUT THE MEAN AND MEDIAN



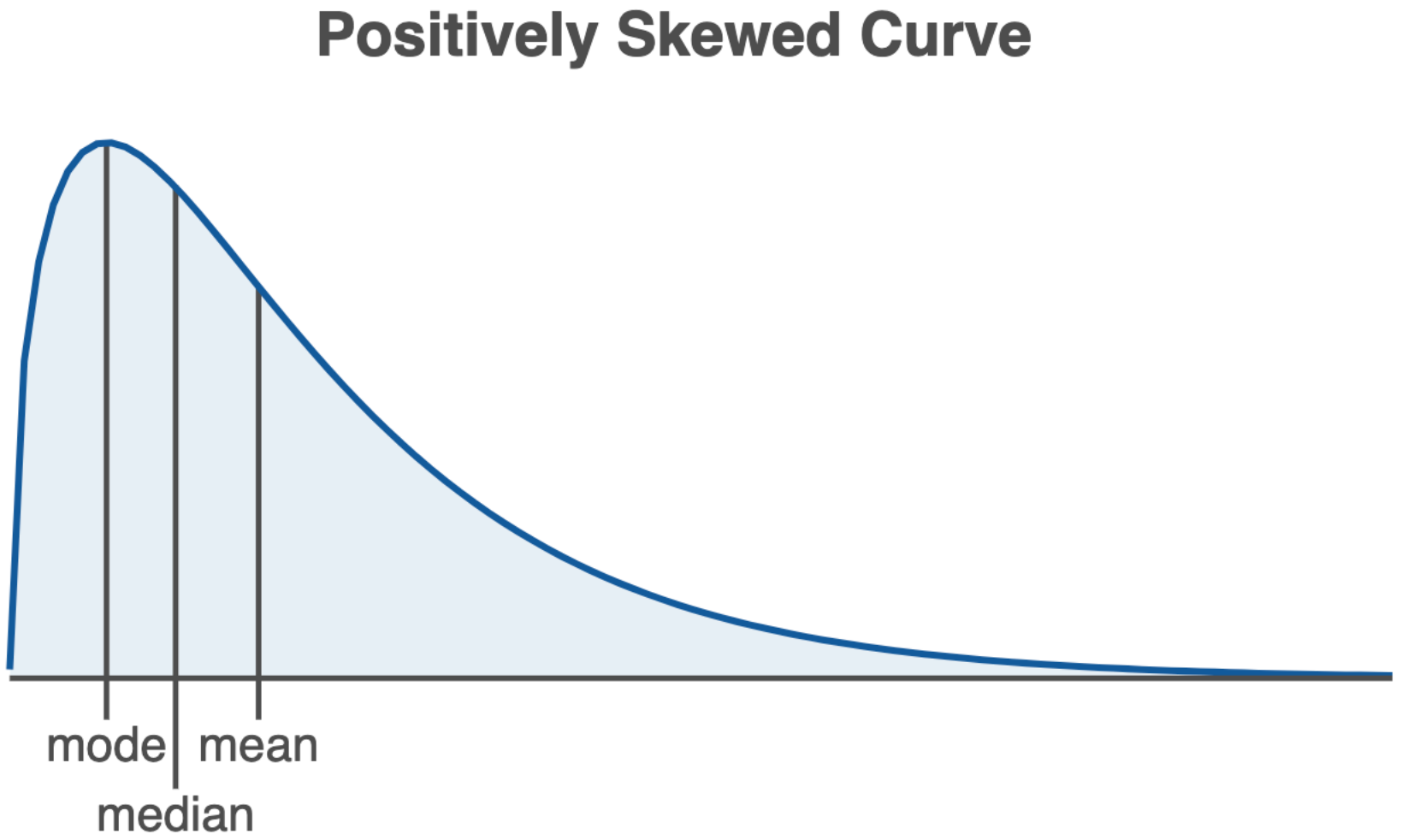
CRITICAL THINKING ABOUT THE MEAN AND MEDIAN

- 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.

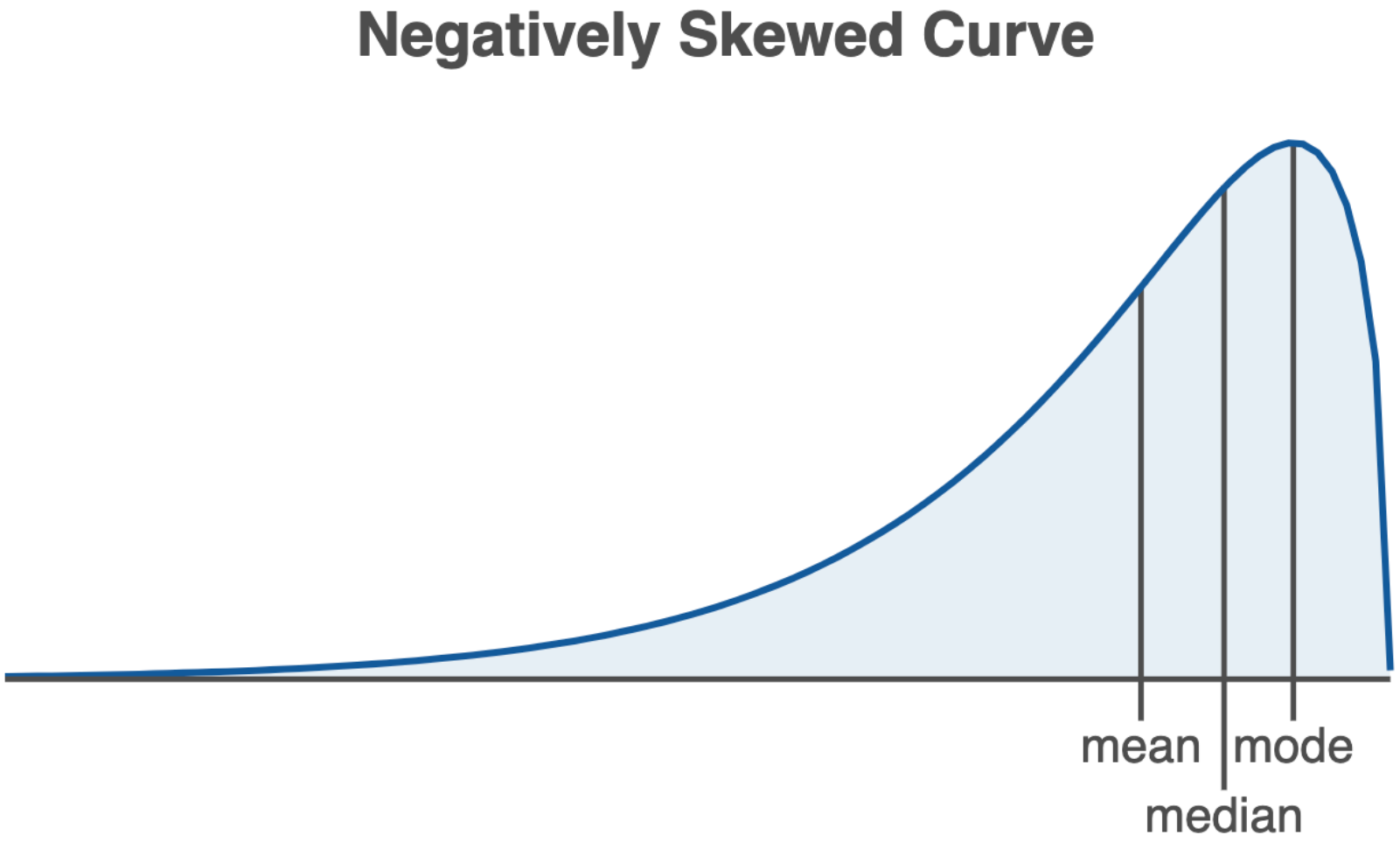




CRITICAL THINKING ABOUT THE MEAN AND MEDIAN



CRITICAL THINKING ABOUT THE MEAN AND MEDIAN



CRITICAL THINKING ABOUT THE MEAN AND MEDIAN

Table 4.1.4 — Sensitivity to Outliers

	Not Sensitive	Very Sensitive
Mean		✔
Median	✔	
Mode	✔	
Trimmed Mean	✔	

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

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

EXAMPLE: STANDARD DEVIATION

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

Steps:

1. Calculate the mean
2. Calculate each deviation. A deviation is the difference between the observation and the mean.
3. Square each devaition
4. Add
5. Divide by *n-1* ← STOP HERE FOR VARIANCE
6. Take the squareroot ← STOP HERE FOR STANDARD DEVIATION

Sample Standard Deviation

The standard deviation is the most common metric for spread. It is in the same units that the data are in gives a rough sense of how far data points are from the sample mean.

```
sd(orAQI$AQI, na.rm=TRUE)
```

```
## [1] 57.85969
```

COMPARING SUBGROUPS

Step 8: Comparing Groups

Numeric Summaries

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

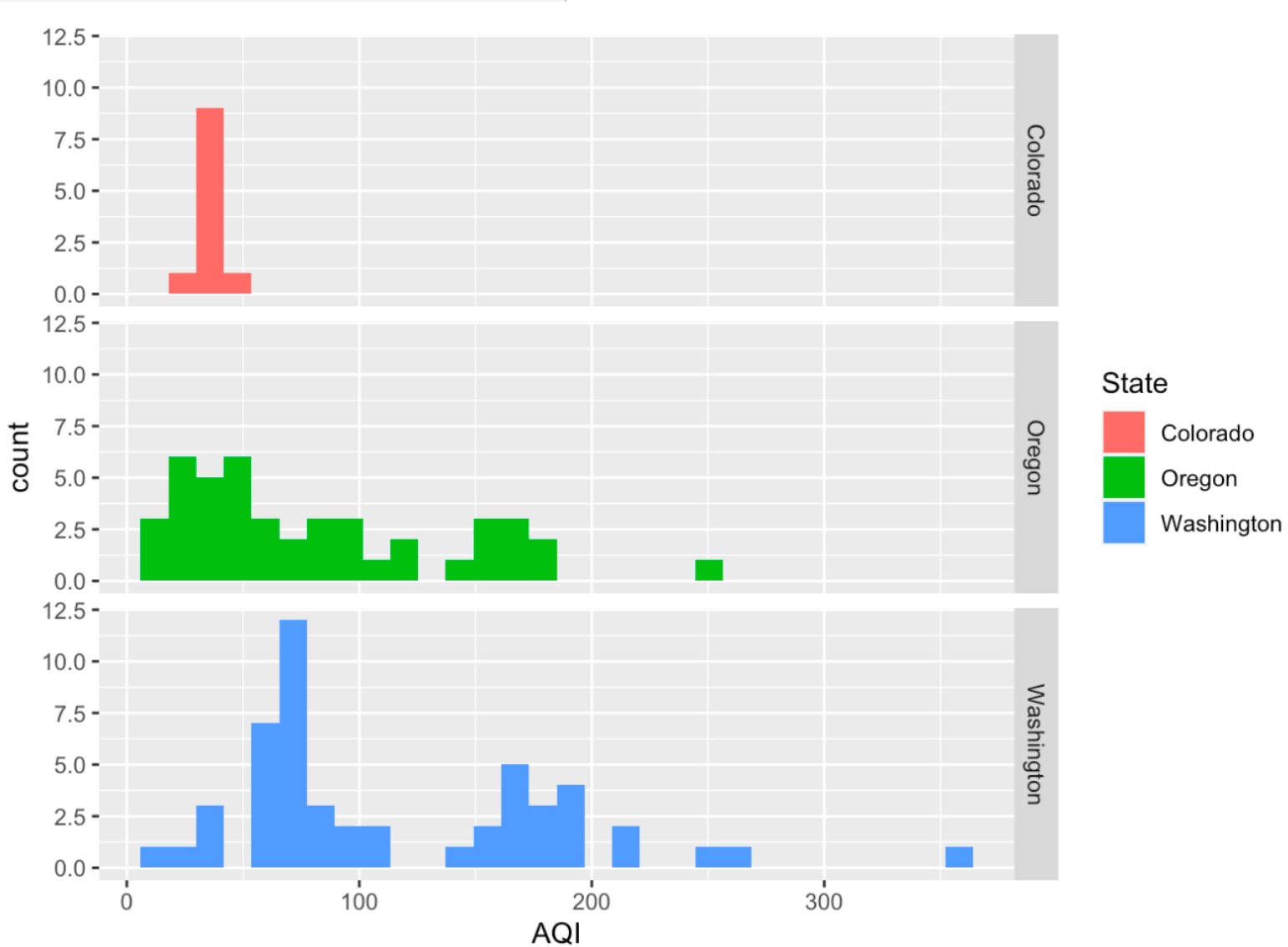
```
aqi%>%
  group_by(State)%>%
  summarise(n=n(),
            medAQI=median(AQI, na.rm = TRUE),
            avgAQI=mean(AQI, na.rm = TRUE),
            sdAQI=sd(AQI, na.rm = TRUE))
```

```
## # A tibble: 3 × 5
##   State           n medAQI avgAQI sdAQI
##   <fct>       <int>  <dbl>  <dbl> <dbl>
## 1 Colorado       11   36     35.6  7.07
## 2 Oregon         47   62.5   81.0 57.9
## 3 Washington     52   84    117.  72.0
```

What do you observe?

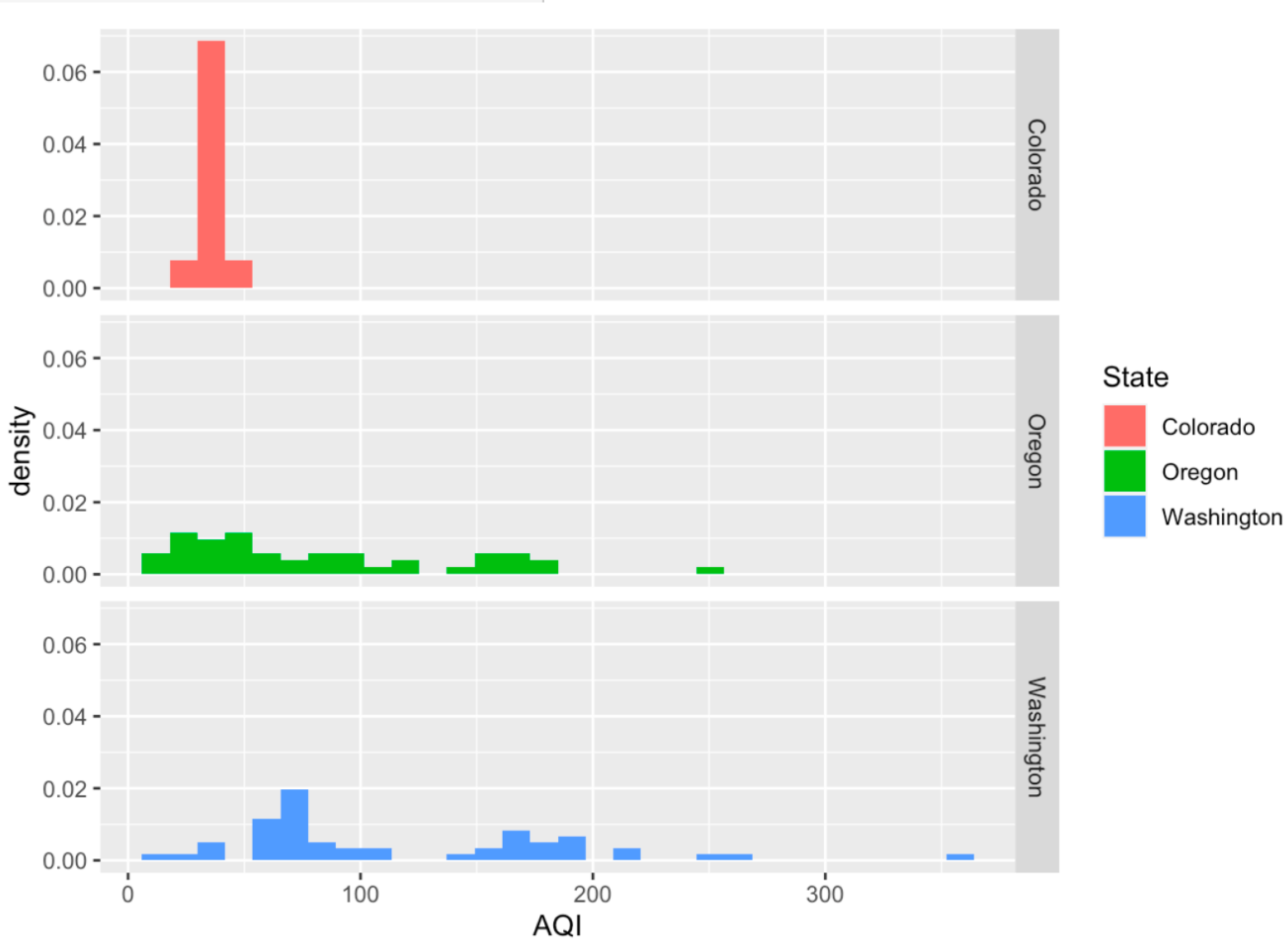
Here is a histogram where the y-axis is count.

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

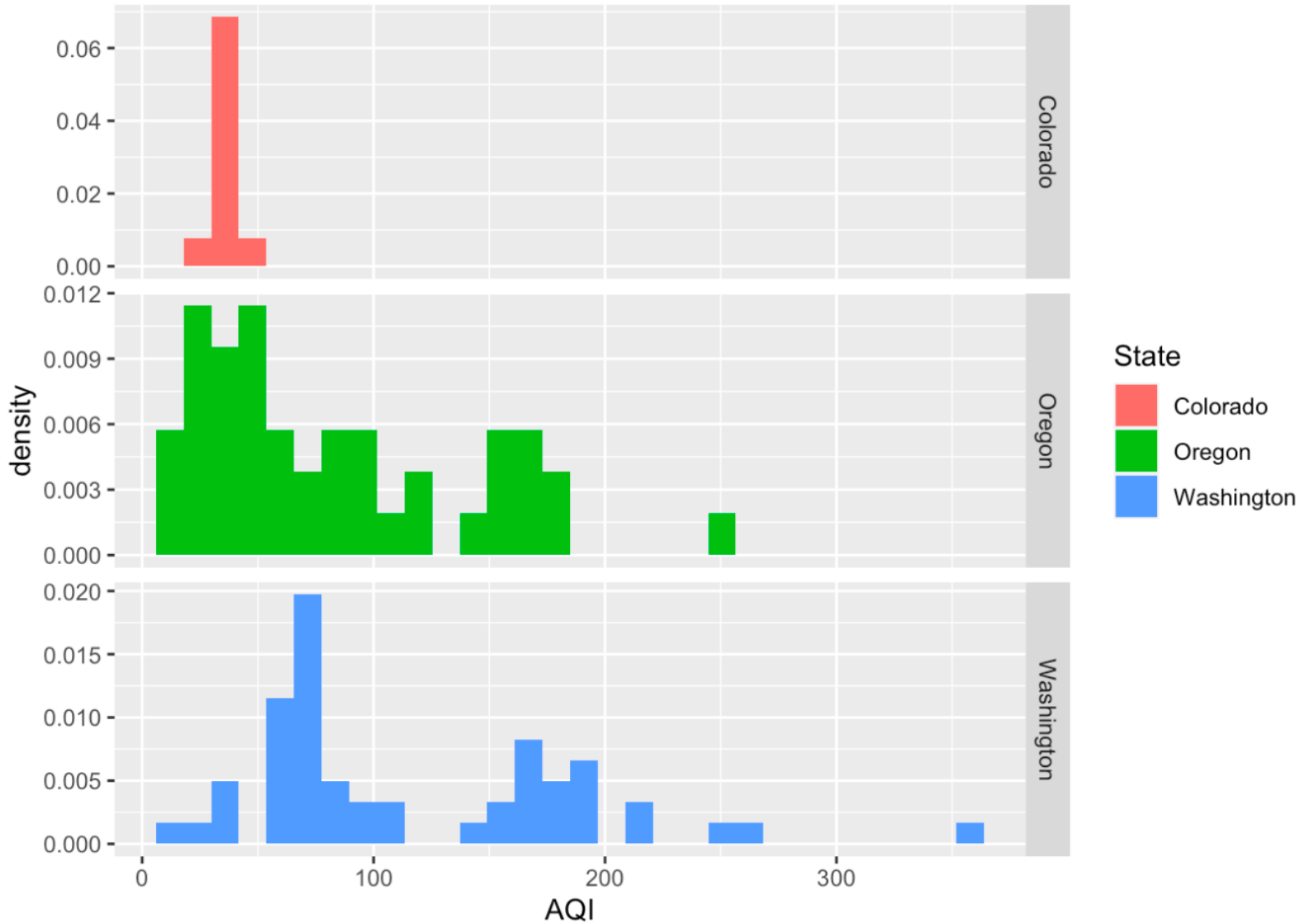


We can change the y-axis to density (proportions).

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

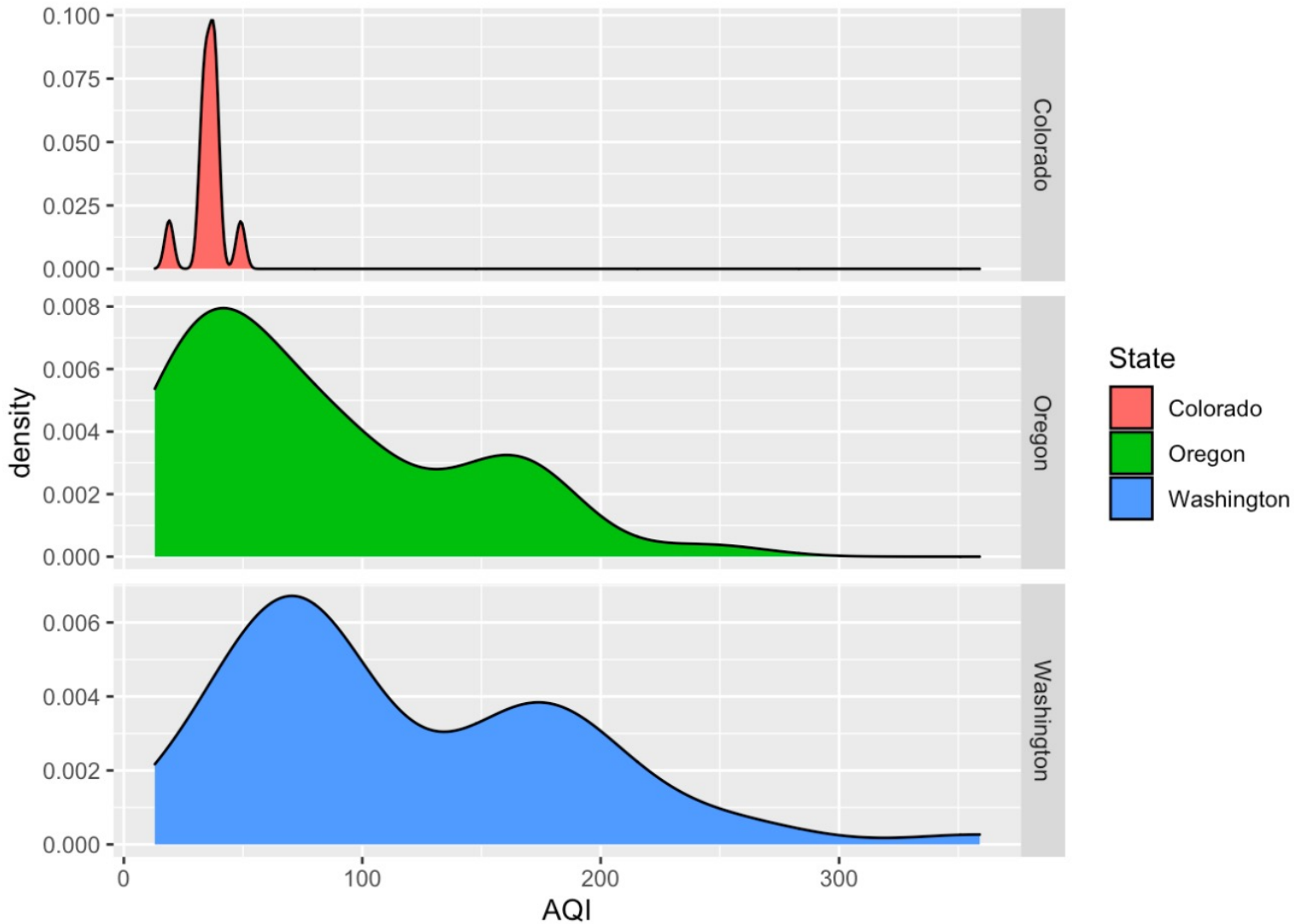


```r
### Histogram (Density - Free_y)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_histogram(aes(y=..density..))+
  facet_grid(State~., scales = "free_y")
```



We can also make a density plot!

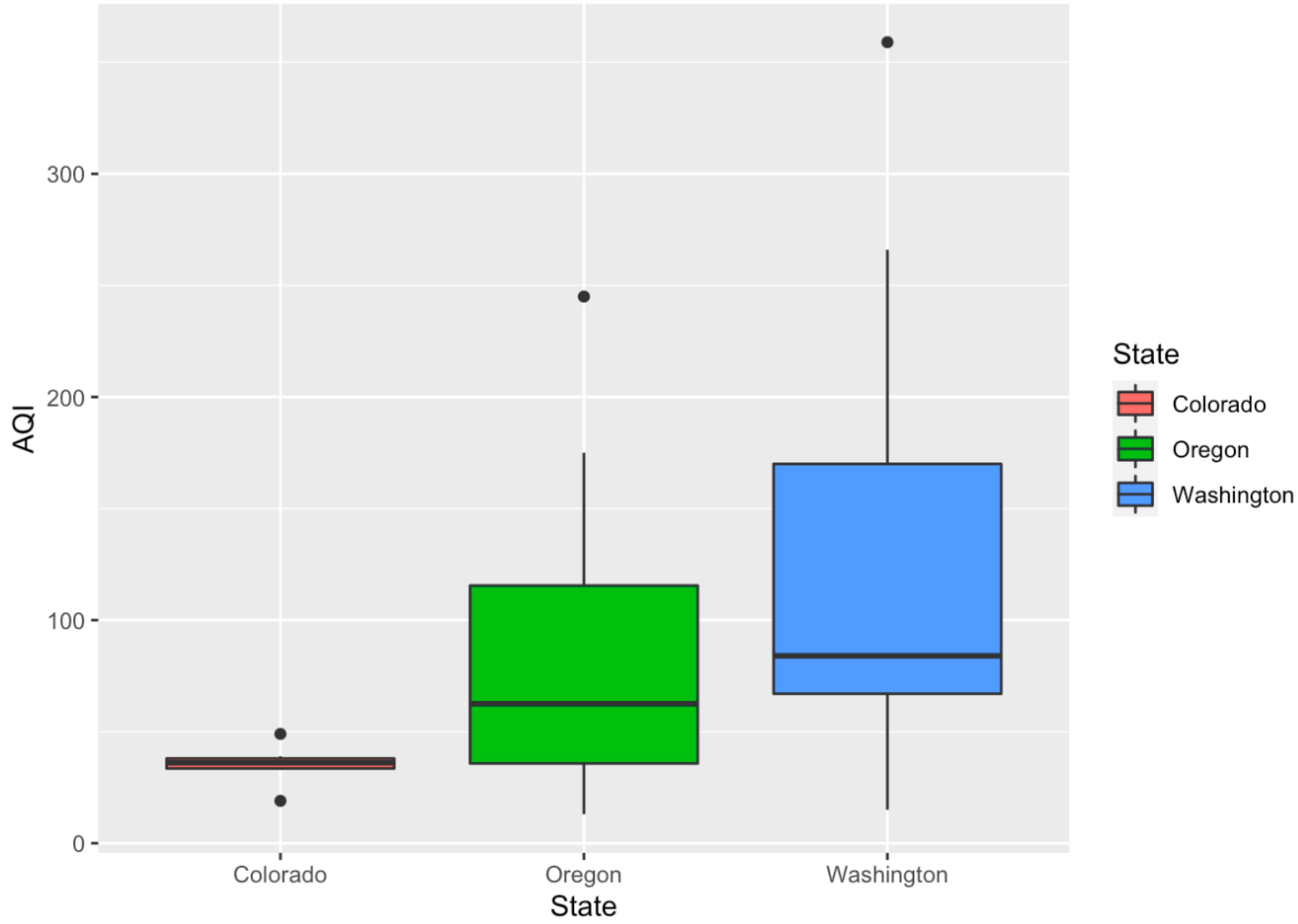
```
### Density Plot (free_y)
ggplot(aqi, aes(x=AQI, fill=State))+
  geom_density()+
  facet_grid(State~., scales = "free_y")
```



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

```
# BOXPLOT

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



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

HISTOGRAM (VS BAR CHART) WARNINGS

- Histograms and bar charts may look similar, but they are fundamentally different!
  - Don't make a histogram of categorical data
  - Don't look for shape, center, and spread of a bar chart
- Don't use bars in every display (i.e. only use them for "piles")
- Be careful to choose an appropriate bin width