DATA SOCIETY™

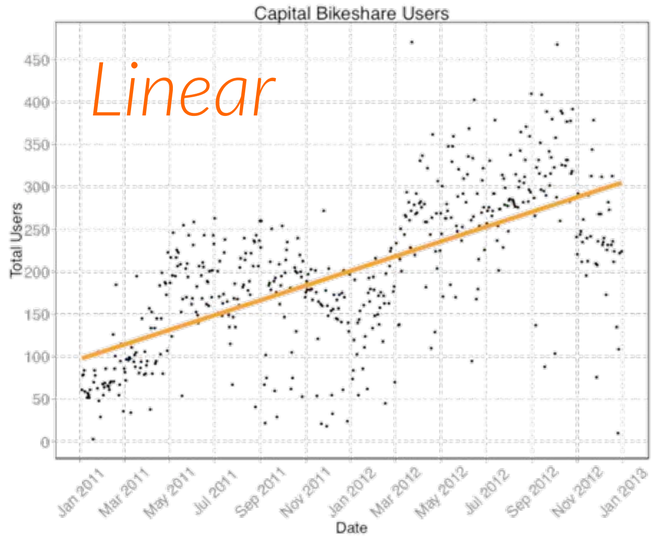
"If you can't explain it simply, you don't understand it well enough."

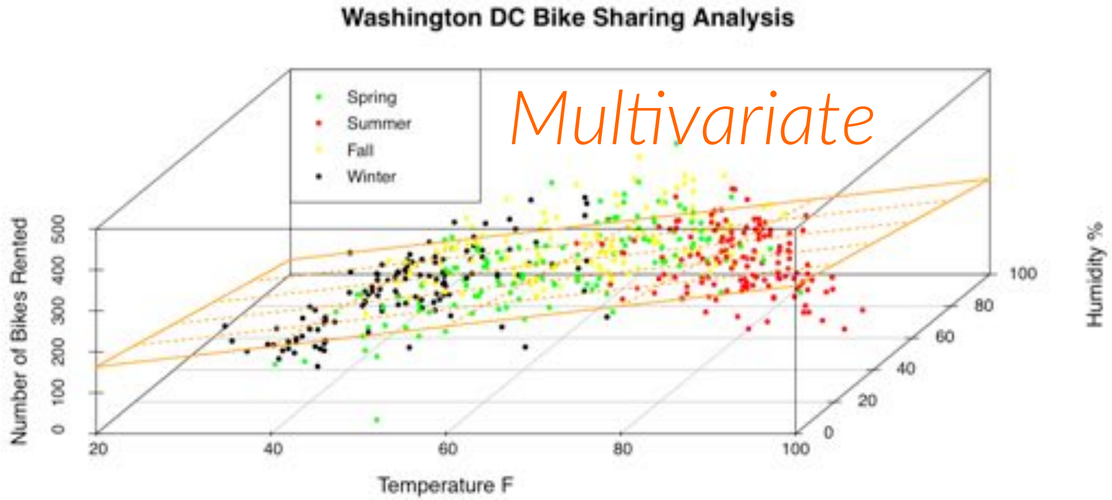
- Albert Einstein

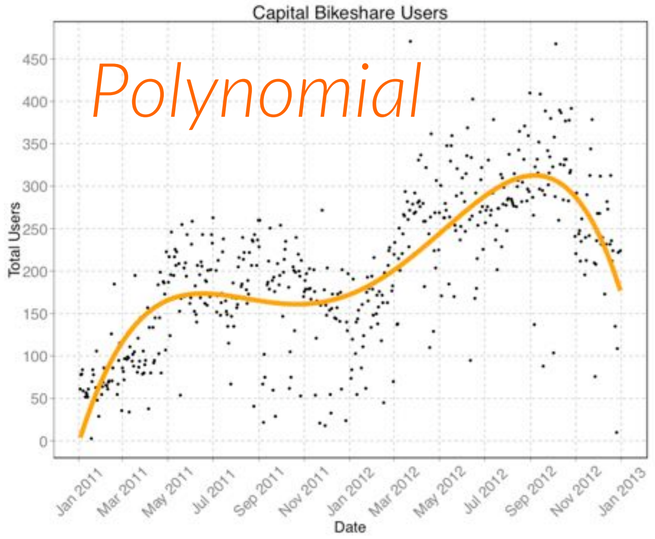
What will you learn?

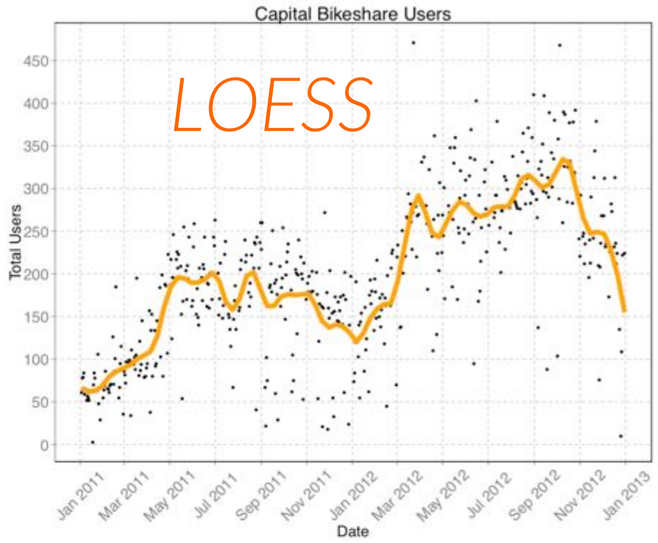
1. How do many variables combine in complex relationships to predict how many [bike rentals, boat rentals, concert tickets, etc.] will people buy?
2. How do seeming non-quantifiable variables (whether it rains, the color of the car, etc.) affect a numerical outcome (number of donations, number of visitors, etc.)?
3. What is the pattern of behavior across time?
   - Identify how latent cultural patterns, nature's cycles and other patterns affect outcomes
4. How do relationships change over time?
   - As time goes by, various factors affect stock prices in different ways – we'll show you how to quantify how these and other relationships change!

What you'll do: regression analyses

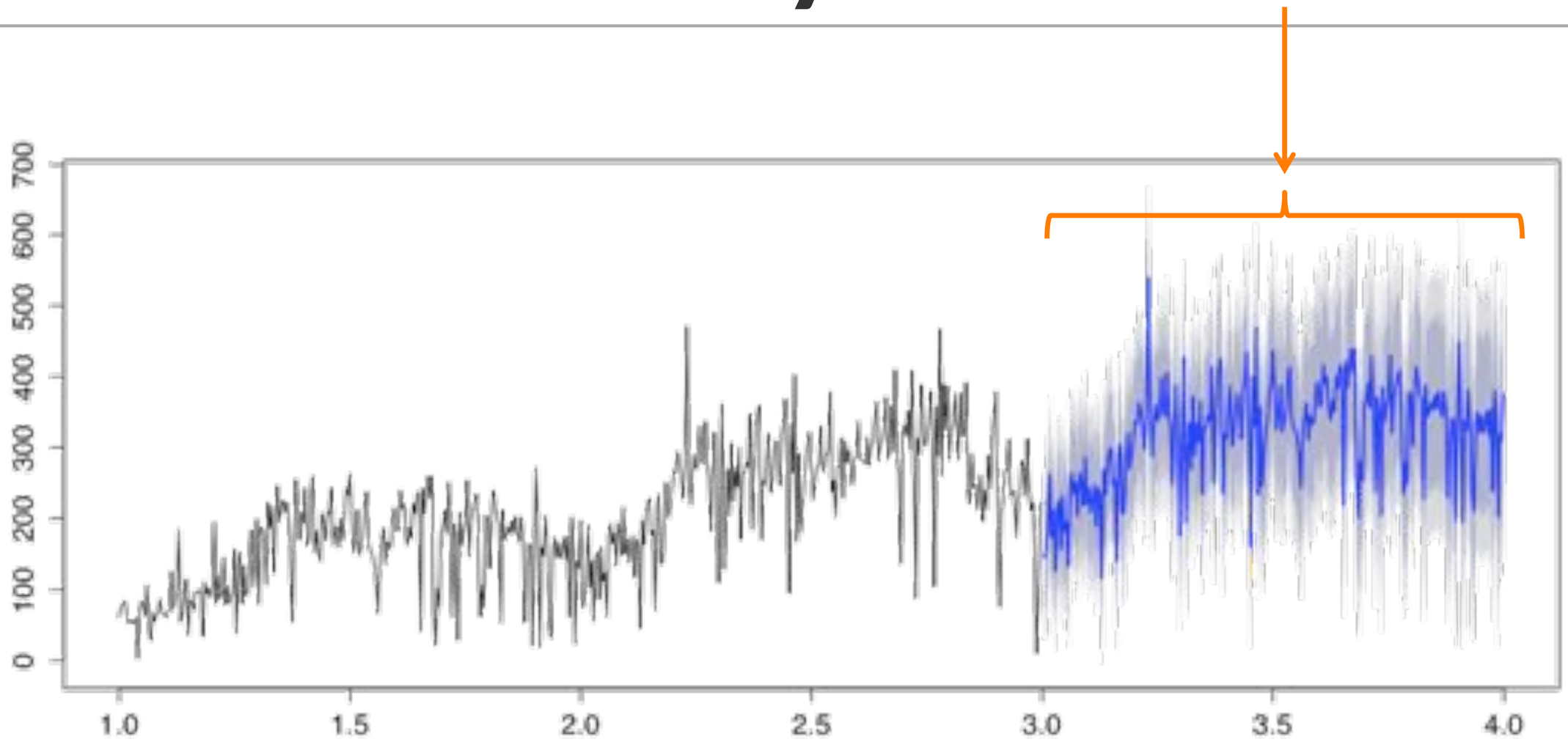








Time-series analyses & forecasts



Setting expectations

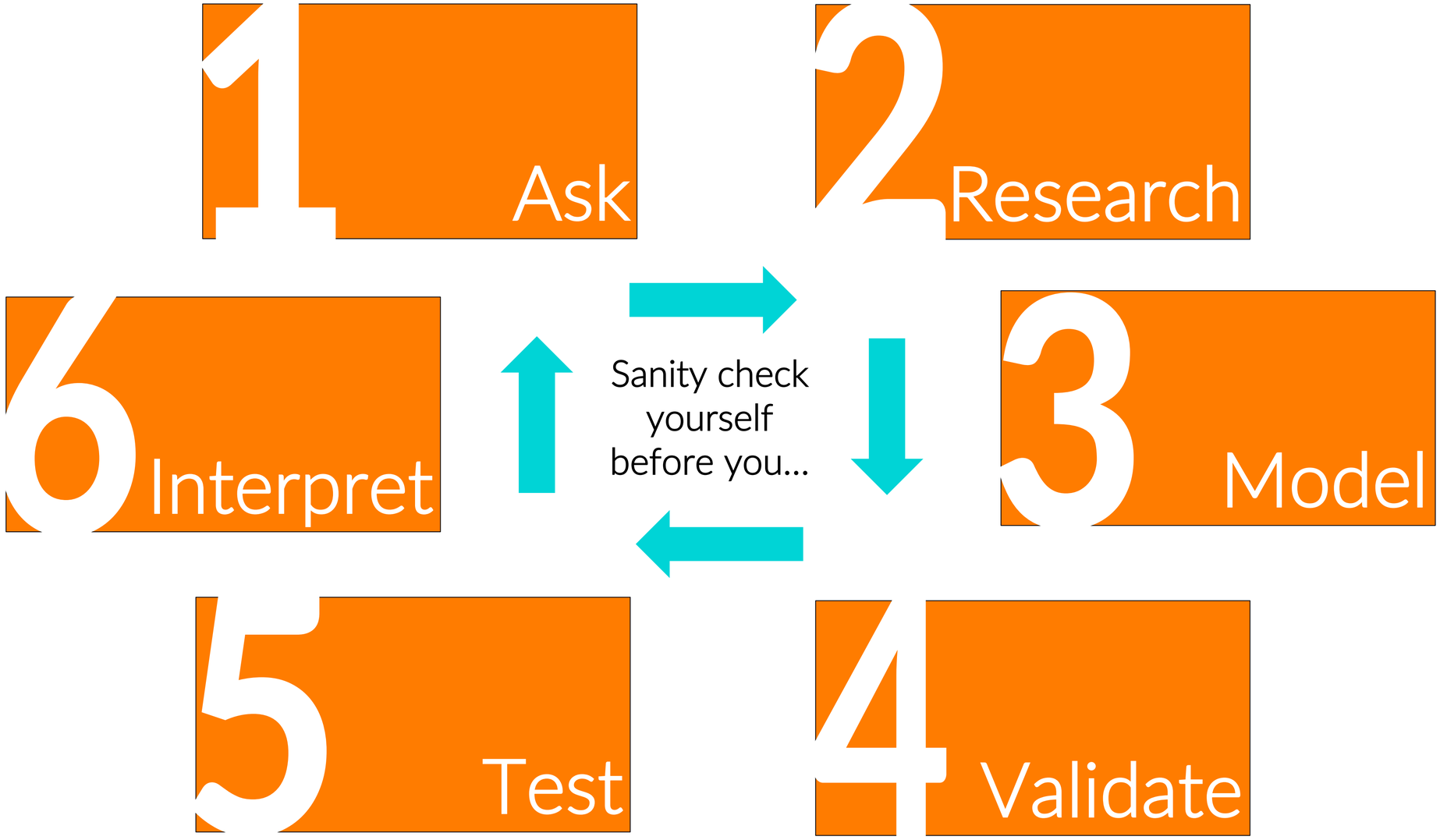
Data science takes dedication! You will need to:

1. Take this course ☺
2. Practice coding
3. Review class material on your own
4. Practice coding
5. Complete concept reviews and exercises outside of class
6. Practice coding
7. Share and read latest news

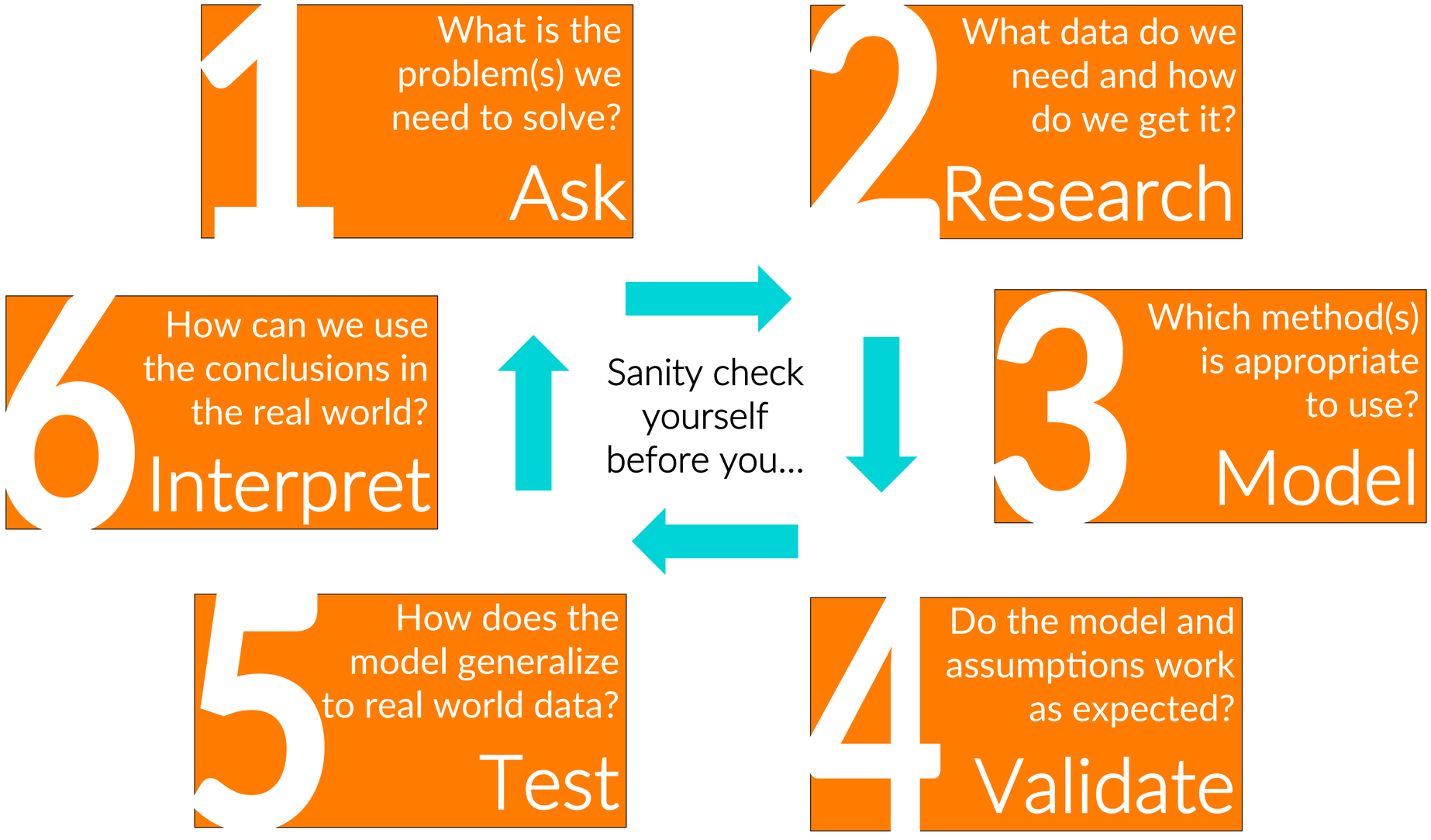
Functions of a business

Functions of a business

Data science control cycle



Data science control cycle



Forecasting: definition

To predict something after looking at the available information

-Merriam-Webster dictionary

Regression vs. classification

Regression

- Regression can:
  - Predict the numerical value of a variable based on the value of another variable(s)
- Regression can't:
  - Predict probabilities (except for logistic regression)
  - Confirm causation (since A happened B will occur, only controlled experiments can confirm causation)

Regression vs. classification

Regression	Classification
Predict an amount	Predict group membership
Continuous/ordered output variable	Categorical/ unordered output variable
Performance Measure: Difference between the predicted value and the actual value	Performance Measure: Misclassification rate

Forecasting: the truth

The only thing true about a forecast is that it's wrong!

A good forecaster isn't smarter than anyone else;
they just have their ignorance better organized

-Anonymous

Forecasting: use cases

- GlaxoSmithKline predicts the supply of participants for clinical trials of new drugs in order to allocate drug trial resources
- Amazon uses forecasts of consumer demand to manage inventory levels
- Ford uses forecasts of seasonally adjusted annual sales rates to manufacture cars
- New South Wales, Australia, predicts travel time in Sydney based on events in the city and weather forecasts
- The Heritage Provider Network predicts the number of days a patient will spend in the hospital over the next year

Source: Predictive Analytics by Eric Siegel

Outline

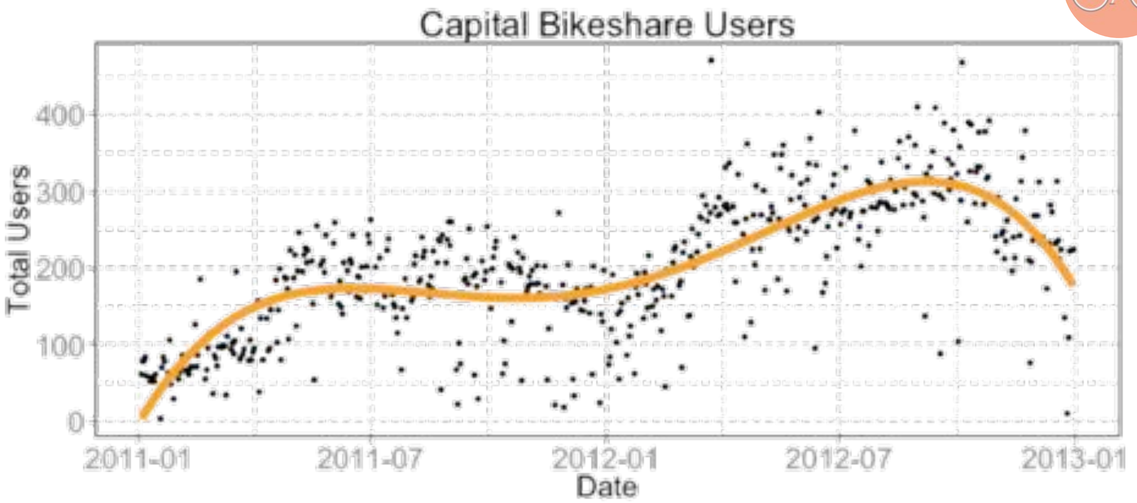
Application	Question to answer	Technique
How do various aspects of the weather and the time of year affect demand for bike rentals?	How do many variables combine in complex relationships to predict how many times an event will occur?	Linear regression Polynomial regression Multivariate regression
How do different types of weather (rain, hail, etc.) affect demand for bike rentals?	How do categorical variables affect a numerical outcome?	Datafication
How do different seasons of the year, and days of the week affect demand for bike rentals?	What is the pattern of behavior across time?	Seasonality analysis
How can you predict demand for bike rentals in the short term in the absence of rich, contextual data?	How do relationships change over time?	Local regression

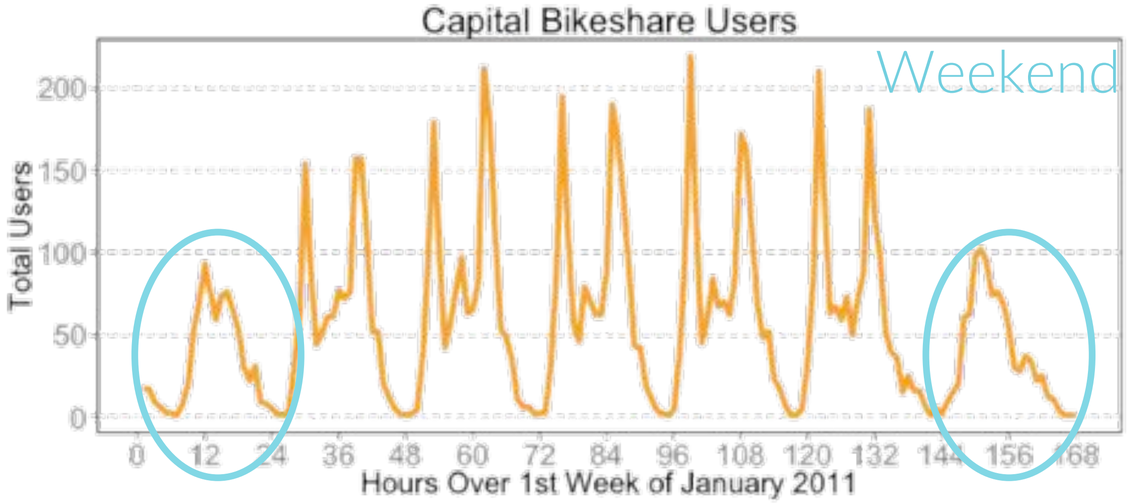
Washington DC bike sharing program

- Bike sharing systems are new generation rental models where the whole process from membership, rental and return is automatic
- There are over 500 bike sharing programs around the world with over 500,000 bikes
- The automated systems track numerous data points providing a treasure trove of data about the mobility of residents
- Since 2010 Washington DC's Capital Bikeshare program has provided over 3,000 bicycles at over 350 stations
- Bikes are available 24 hours/day, 365 days/year
- Objectives:
  1. Forecast the number of bikes required
  2. Adjust pricing to reflect demand

Capital Bikeshare data: considerations

1. Demand for bicycles is seasonal
   - More people ride bicycles during the summer when it's warm rather than the winter
2. Demand for bicycles is growing, yet there are peaks and troughs
3. Demand for bicycles during the day is greater than demand at night
4. Demand for bicycles during the week is greater than during the weekends or holidays





Capital Bikeshare data

Date	Season	Hour	Holiday	Day of the Week	Working Day	Weather Type	Temperature °F	Temperature Feels °F	Humidity %	Wind Speed (MPH)	Casual Users	Registered Users	Total Users
1/1/11	4	0	0	6	0	1	36.6	37.4	81	0.0	3	13	16
1/1/11	4	1	0	6	0	1	34.9	35.6	80	0.0	8	32	40
1/1/11	4	2	0	6	0	1	34.9	35.6	80	0.0	5	27	32
1/1/11	4	3	0	6	0	1	36.6	37.4	75	0.0	3	10	13

- Season: 1 = spring, 2 = summer, 3 = fall, 4 = winter
- Holiday: 0 = no, 1 = yes
- Working day: if day is neither weekend nor holiday is 1, otherwise is 0
- Weather type:
  1. Clear, Few clouds, Partly cloudy, Cloudy
  2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  4. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- Casual users: users who don't have a subscription
- Registered users: subscribers to Capital Bikeshare

Source: University of California at Irvine, https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

What questions can we answer?

A How does any single given factor (air temperature, humidity, wind speed) affect demand for bikes?

B How do several variables (air temperature, humidity, wind speed, day of the week, holidays, hour of the day) affect demand for bikes?

C How can you factor in seasonality and cyclicality when forecasting demand?

Why do we want to know this?

1. To anticipate demand and ensure a sufficient supply of bicycles

2. To adjust pricing according to demand to maximize revenues

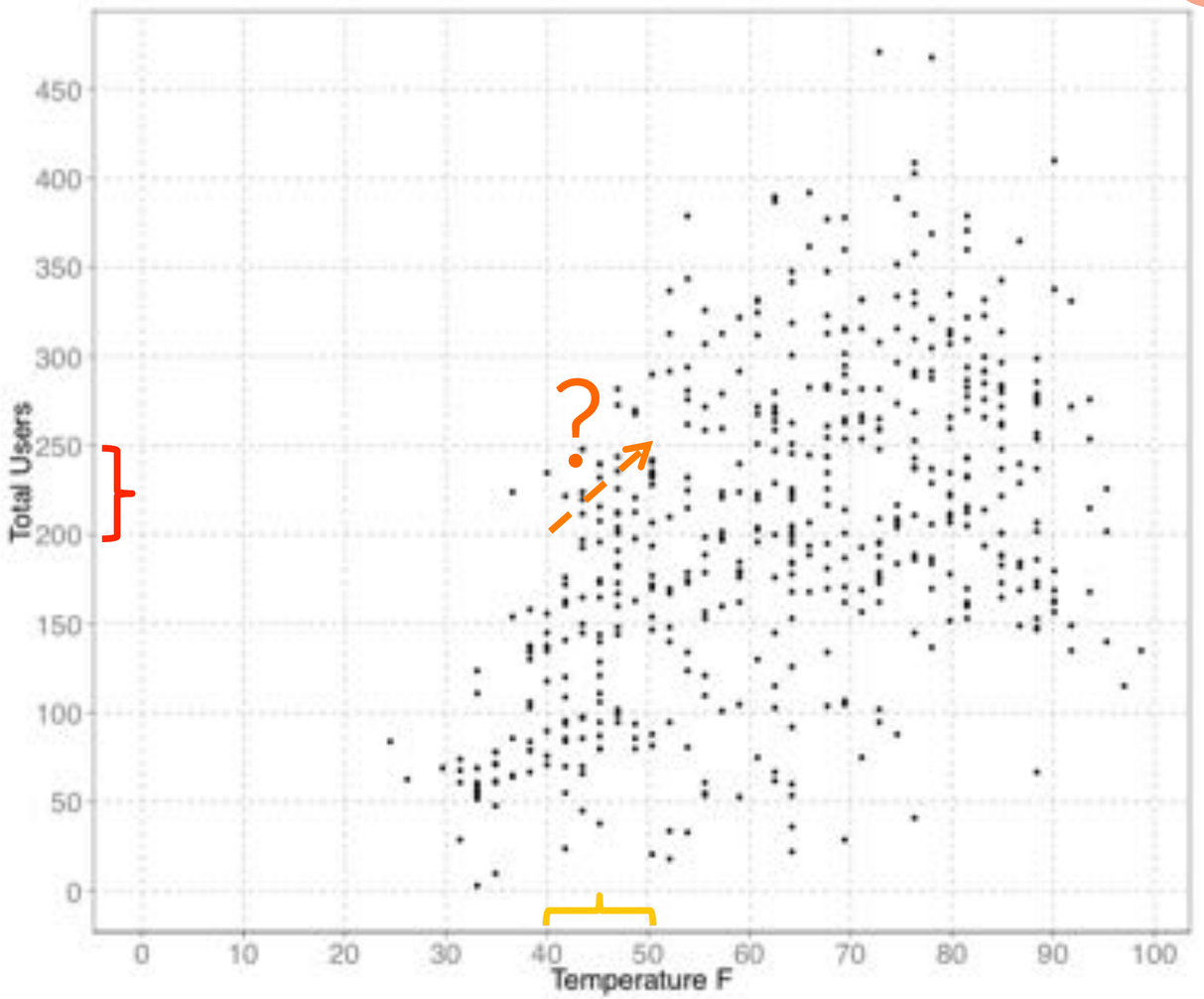
3. To anticipate maintenance expenditures and decrease costs by planning ahead

Regression: measuring relationships

Relation between the number of bikers and temperature

- On average, for every 10 degree increase in temperature, how many more users rent a bike?

- Note: we're assuming that a change in temperature drives a change in the number of users



A

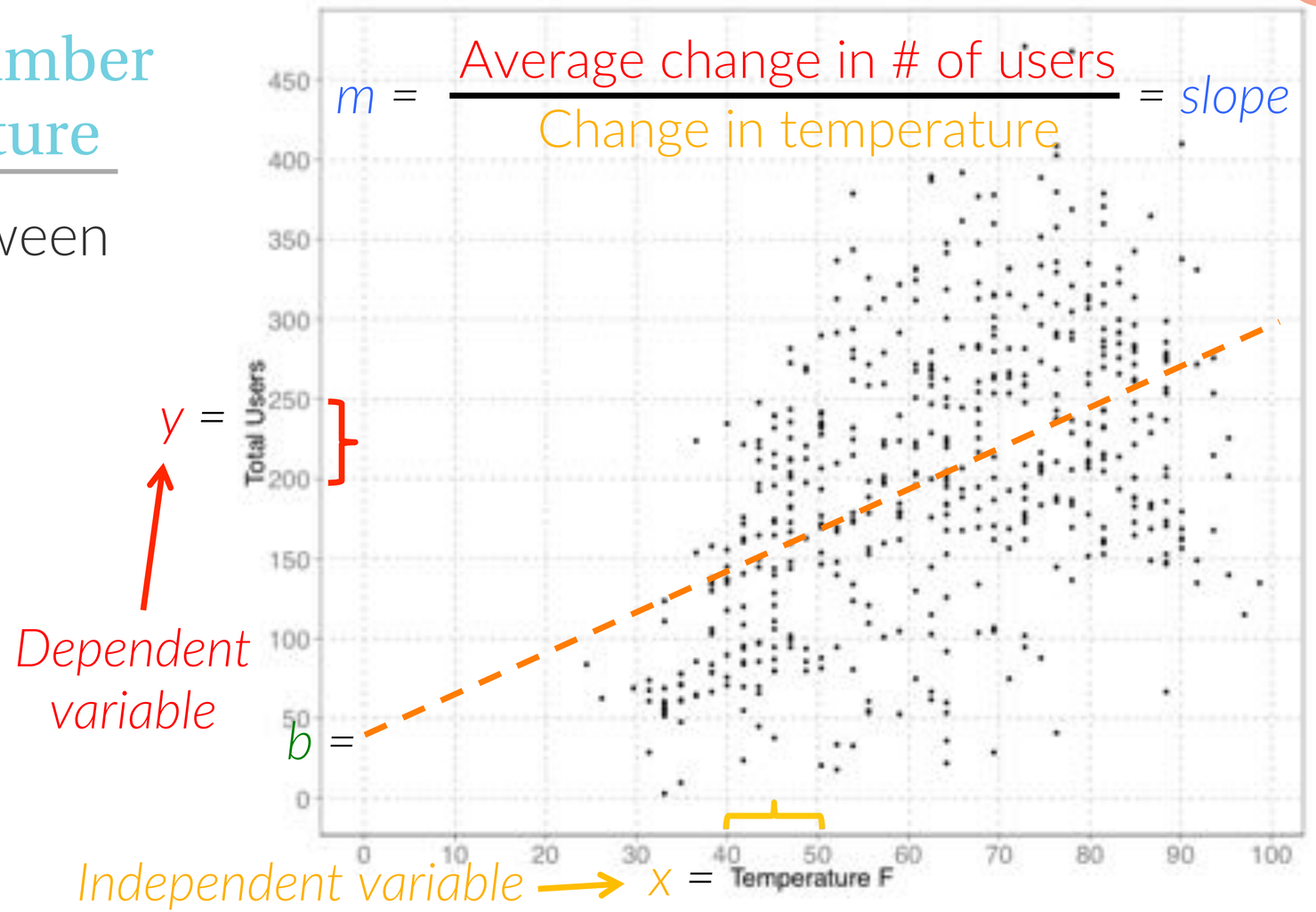
Regression: measuring relationships

Relation between the number of bikers and temperature

- A linear relationship between 2 variables is:

$$y = mx + b$$

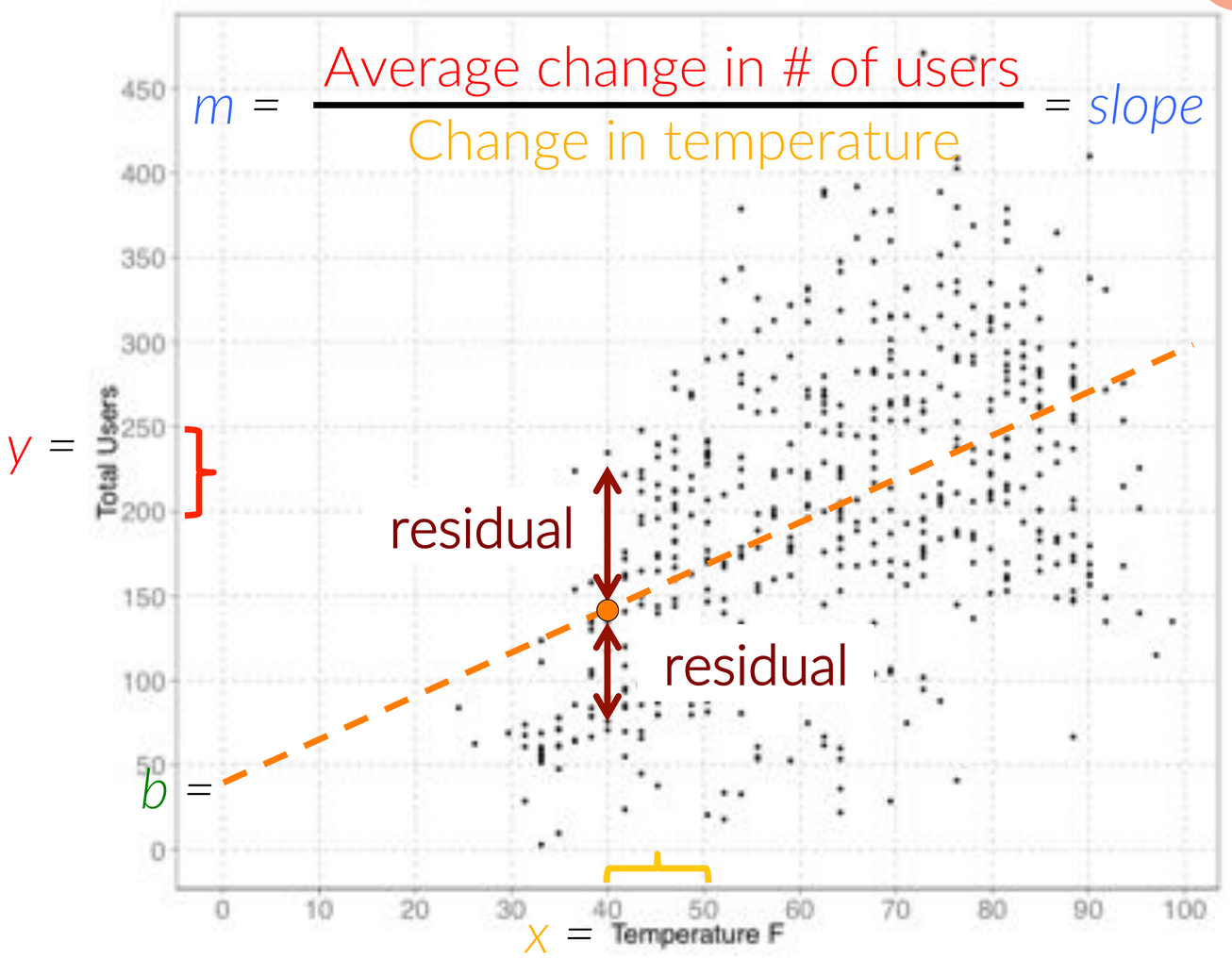
- y: Total Users
- x: Temperature F
- m: rate of change (slope)
- b: value of y when x = 0



What is the rate of change (slope)?

$y = mx + b$

- Slope = average change in y
- The average expected value of y is the based on m (average rate of change)
- Residual = distance from the actual data points (values of y) to the average expected value of y for every value of x

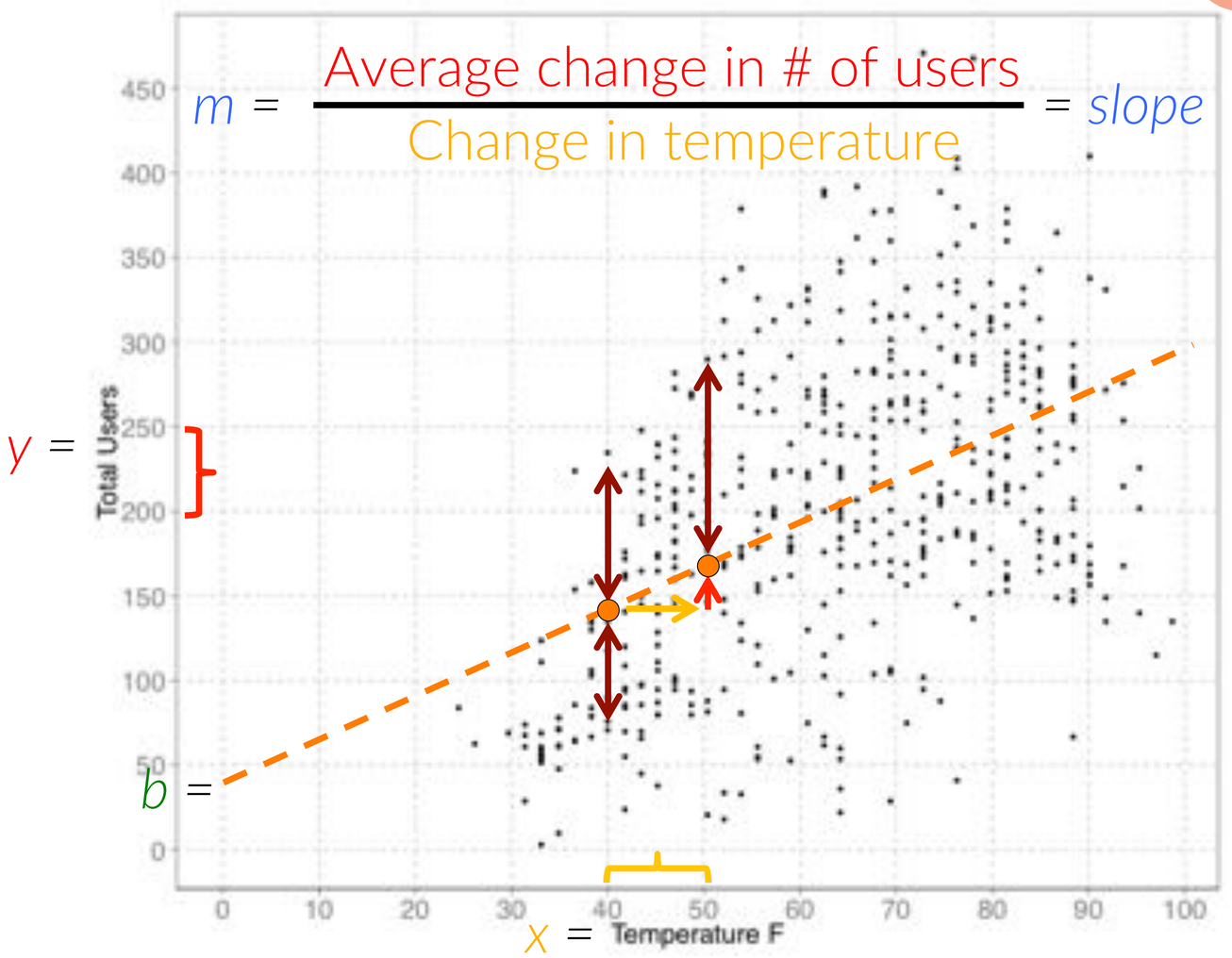


What is the rate of change (slope)?

$$y = mx + b$$

- The average expected values of y minimize the sum of all the residuals

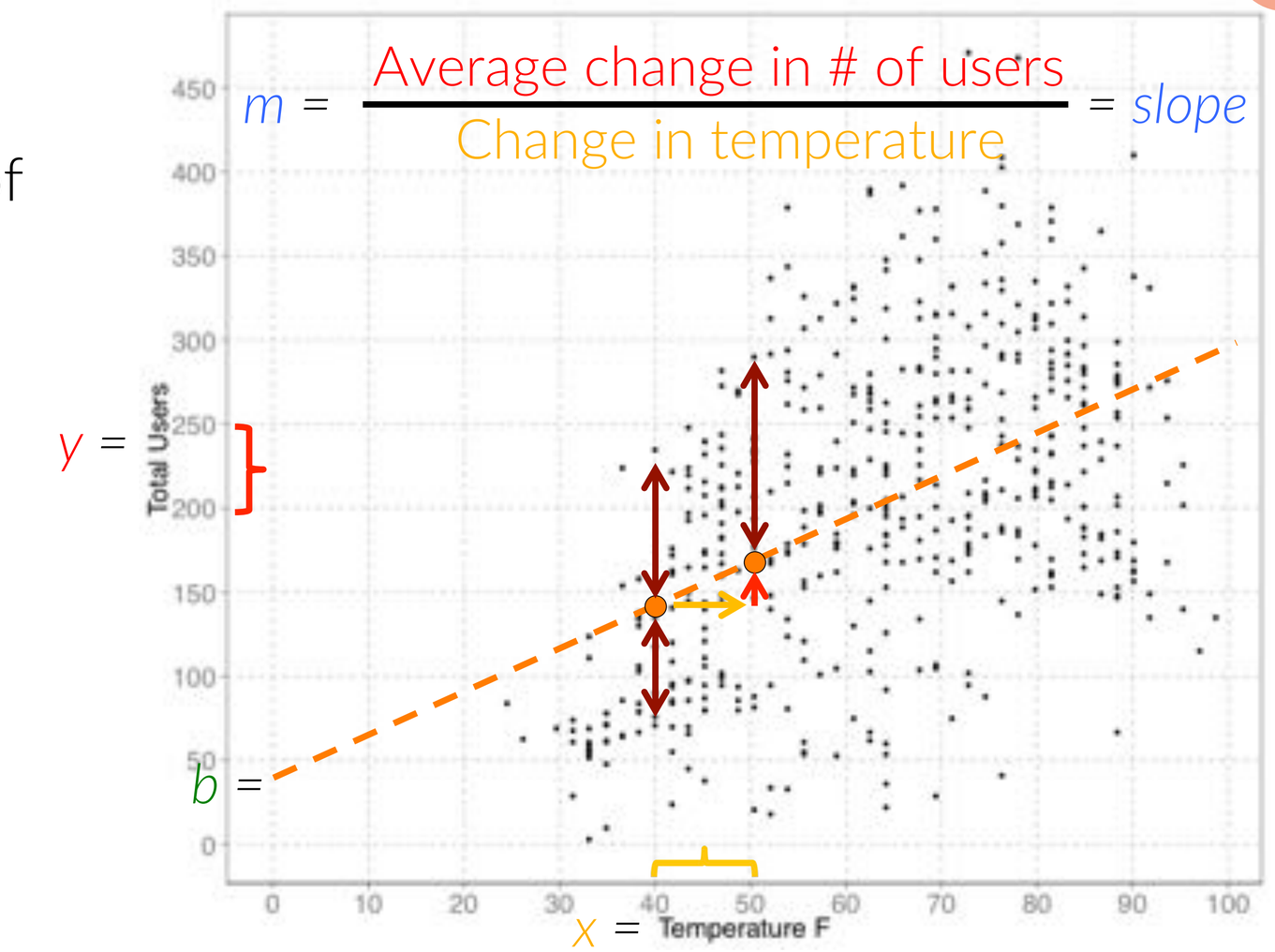
- Slope = change in the average expected value of y for every change in x



What is the rate of change (slope)?

$y = mx + b$

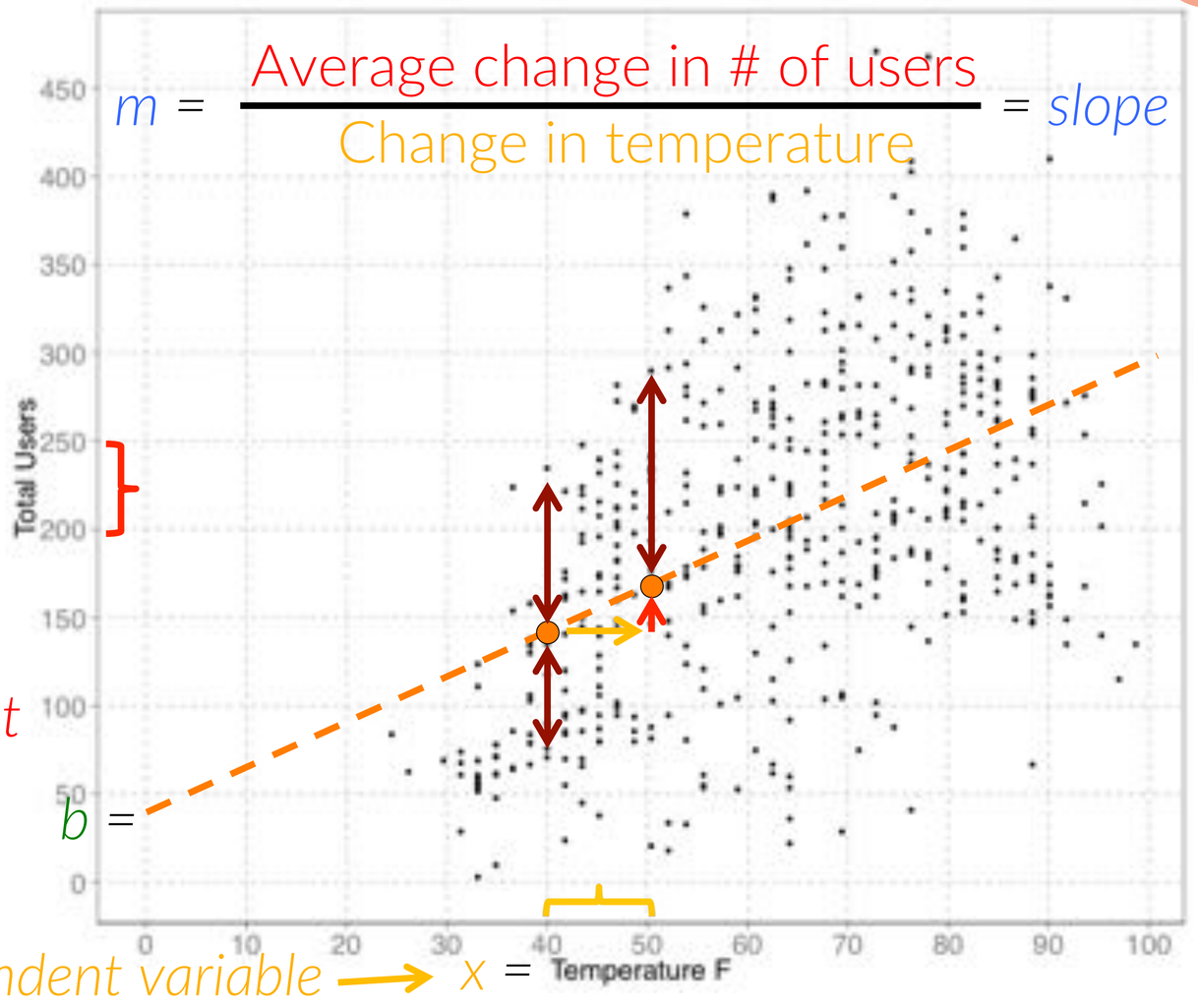
- The average expected values of y minimize the sum of all the residuals
- Slope = 2.6
- For every additional 1°F there are 2.6 additional bike users



y-intercept

$y = mx + b$

- The average expected values of y minimize the sum of all the residuals
- For every additional 1°F there are 2.6 additional bike users
- b = *y-intercept* = 37.6 (value of y when $x = 0$)



A

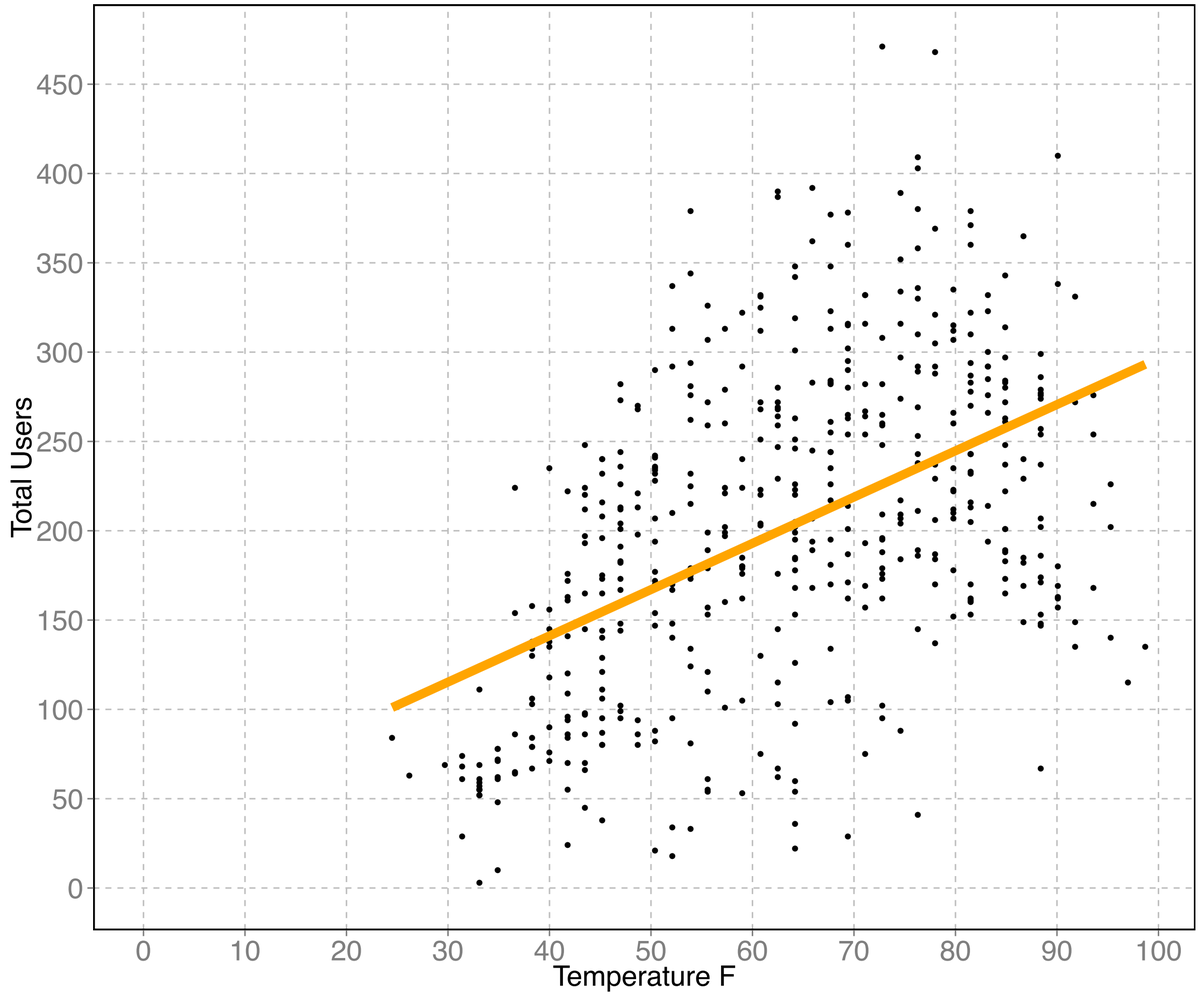
Temperature vs. bike users

$y = mx + b$

Number of bike users

=

2.6 * (Temperature °F) + 37.6



A

Load the data

Script

```
# Start by setting your working directory.
setwd("~/Desktop/Forecasting")
bike = read.csv("bike data.csv", check.names = FALSE)

# What does our data look like?
View(bike)

# We first want to isolate data to exclude variability based on hour of the day and
# weekends and holidays. Let's subset data only for weekdays at noon.
bike_weekday_noon = subset(bike,
                           `Working Day` == 1 & Hour == 12)

View(bike_weekday_noon)
```

Ensures R doesn't modify column names

Criteria for subsetting

Forecasting.R | bike_weekday_noon

Filter

	Date	Season	Hour	Holiday	Day of the Week	Working Day	Weather Type	Temperature F	Temperature Feels F	Humidity	Wind Speed	Casual Users	Registered Users	Total Users
58	1/3/2011	4	12	0	1	1	1	34.9	28.4	35	20	13	48	61
81	1/4/2011	4	12	0	2	1	1	34.9	30.2	51	11	12	66	78

Showing 1 to 2 of 497 entries

A

Plot the data

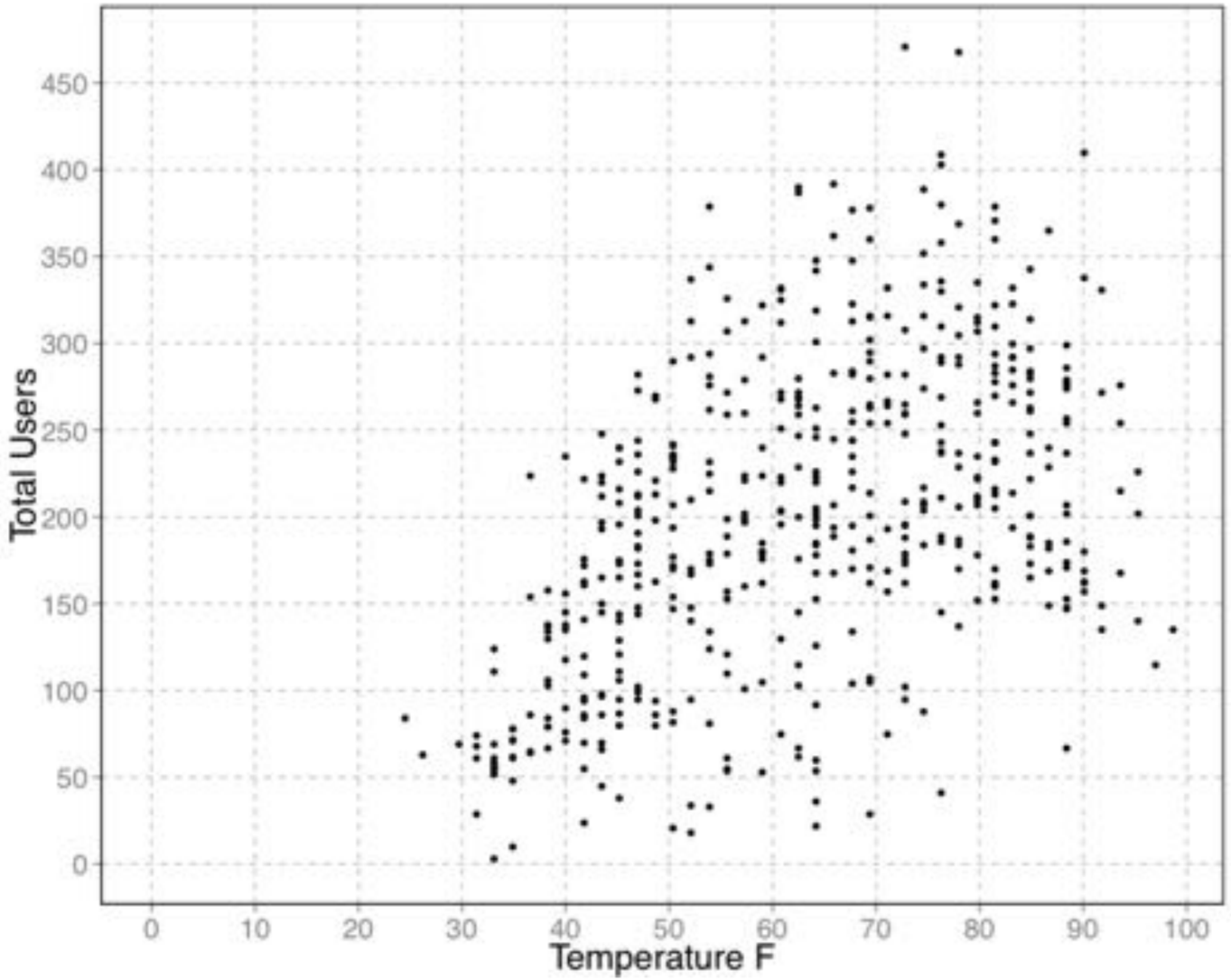
Script

```
# Let's plot temperature vs. the total count of bike rentals.
install.packages("ggplot2")
library(ggplot2)

ggplot(bike_weekday_noon,
       aes(x = `Temperature F`,
           y = `Total Users`)) +
  geom_point() +
  expand_limits(x = 0, y = 0) +
  scale_y_continuous(breaks = seq(0, 500, by = 50)) +
  scale_x_continuous(breaks = seq(0, 100, by = 10)) +
  theme(text = element_text(size = 20),
        panel.border = element_rect(color = "black",
                                    fill = NA,
                                    size = 1),
        panel.background = element_rect(fill = "white"),
        panel.grid.minor = element_line(color = NA),
        panel.grid.major = element_line(color = "grey",
                                        linetype = "dashed"))
```

1. Specify the data set
2. Specify data to plot on the x-axis
3. Specify data to plot on the y-axis
4. Plot the data as points
5. Sets minimum values for x and y axes
6. Sets y-axis limits
7. Sets x-axis limits
8. Sets font size for the whole graph
9. Border color is black
10. Tell R that there is no fill
11. Set border thickness
12. Set graph background
13. Remove minor grid marks
14. Make major grid marks grey
15. Make major grid mark lines dashed

A

Plot the data



A

Add linear regression line

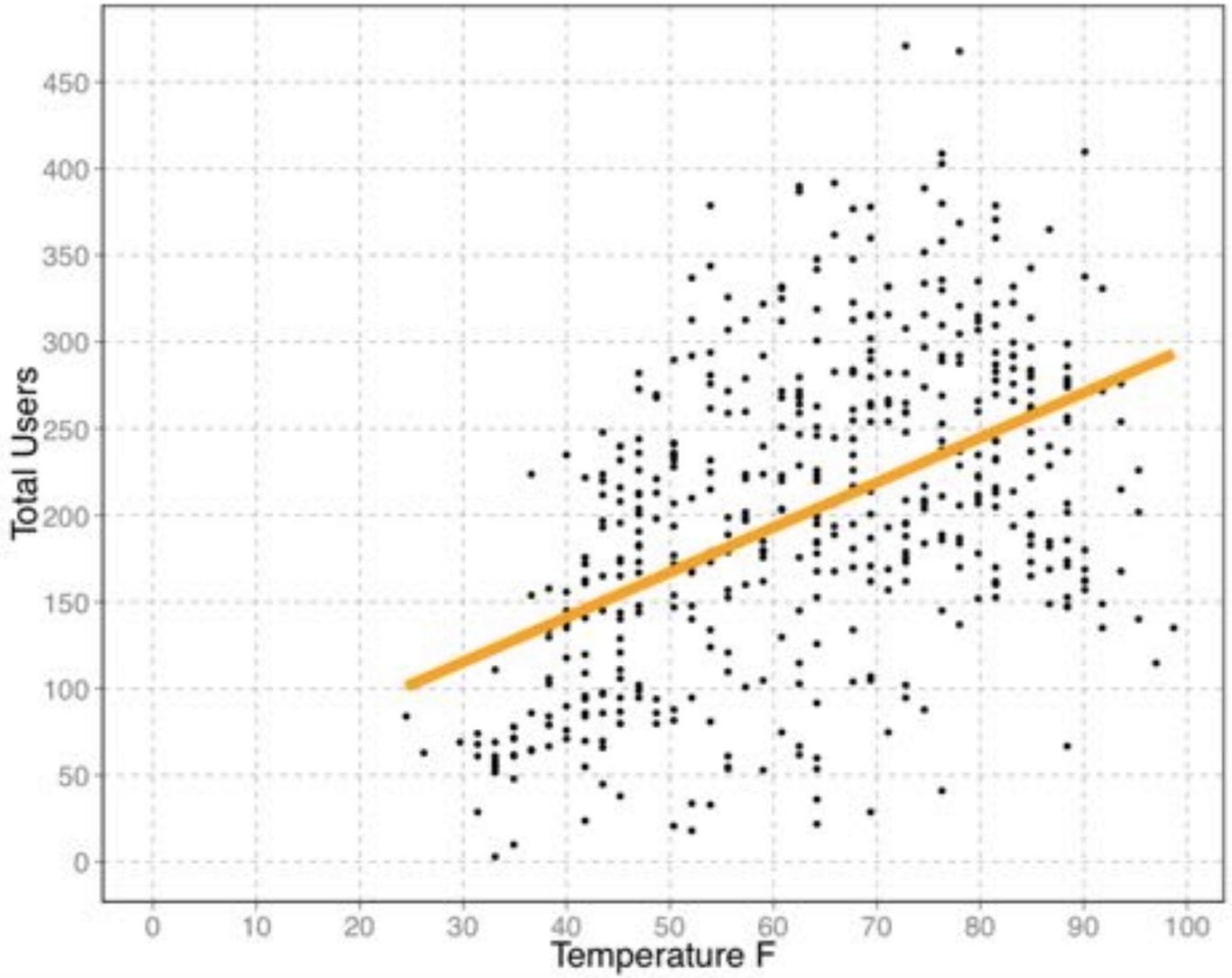
Script

```
# Let's add a best fit line.
ggplot(bike_weekday_noon,
       aes(x = `Temperature F`,
           y = `Total Users`)) +
  geom_point() +
  stat_smooth(method = "lm",
              se = FALSE,
              color = "orange",
              size = 3) +
  expand_limits(x = 0, y = 0) +
  scale_y_continuous(breaks = seq(0, 500, by = 50)) +
  scale_x_continuous(breaks = seq(0, 100, by = 10)) +
  theme(text = element_text(size = 20),
        panel.border = element_rect(color = "black",
                                    fill = NA,
                                    size = 1),
        panel.background = element_rect(fill = "white"),
        panel.grid.minor = element_line(color = NA),
        panel.grid.major = element_line(color = "grey",
                                        linetype = "dashed"))
```

1. Use linear regression
2. Don't plot the confidence interval, we'll discuss this later
3. Make the best fit line orange
4. Set the thickness of the best-fit line

A

Add linear regression line



A

Run linear regression

Script

```
# Let's run a linear regression model of temperature vs. number of bike rentals
# so you can see the results and use them in your analysis.
lm_weekday_noon_temp = lm(formula = `Total Users` ~ `Temperature F`,
                          data = bike_weekday_noon)

# Arguments used in lm()
# formula = is the model equation
# On the left of the '~' is the dependent variable
# On the right of the '~' is/are the independent variable(s)
# data = is the data from being used to fit the model on
# Other commonly used arguments:
# subset = a vector to specify a subset of observations to be
#          used in the model fitting process
# weights = a vector used to weight observations differently
# na.action = how to handle "na" values in the data
```

A

Run linear regression

Script

```
# Check the output of the linear model.
lm_weekday_noon_temp
summary(lm_weekday_noon_temp)
```

Number of bike users

=

2.6*(Temperature °F) + 37.6

We will review what these terms mean in the subsequent slides

```
Console ~/Desktop/Forecasting/
> lm_weekday_noon_temp

Call:
lm(formula = `Total Users` ~ `Temperature F`, data = bike_weekday_noon)

Coefficients:
    (Intercept)  `Temperature F`
         37.596           2.589

> summary(lm_weekday_noon_temp)

Call:
lm(formula = `Total Users` ~ `Temperature F`, data = bike_weekday_noon)

Residuals:
     Min       1Q   Median       3Q      Max
-199.499  -57.298   -2.569   54.835  244.896

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      37.5956    13.4588   2.793  0.00542 **
`Temperature F`   2.5894     0.2062  12.557  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 77.82 on 495 degrees of freedom
Multiple R-squared:  0.2416,    Adjusted R-squared:  0.2401
F-statistic: 157.7 on 1 and 495 DF,  p-value: < 2.2e-16
```

A

Run linear regression

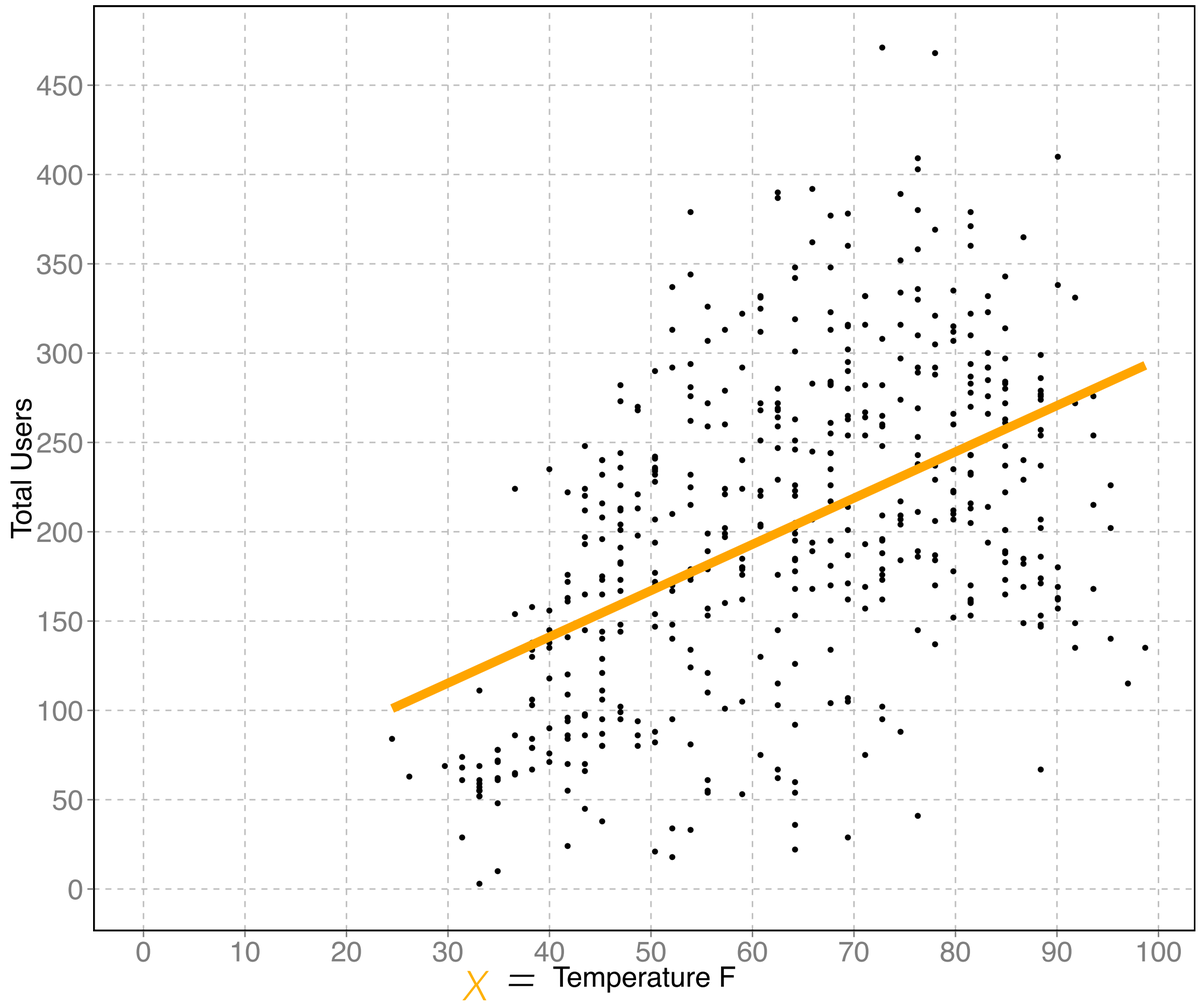
Script

```
# The linear regression model produces several objects. Use ls() to see
# a list of contents.
ls(lm_weekday_noon_temp)

# Let's go through what the outputs mean:
lm_weekday_noon_temp$assign

lm_weekday_noon_temp$call
lm_weekday_noon_temp$coefficients
lm_weekday_noon_temp$df.residual
lm_weekday_noon_temp$effects
lm_weekday_noon_temp$fitted.values
lm_weekday_noon_temp$model
lm_weekday_noon_temp$qr
lm_weekday_noon_temp$rank

lm_weekday_noon_temp$residuals
lm_weekday_noon_temp$xlevels
```

1. List of elements, 0 for y-intercept, 1 for the 1st independent variable, 2 for the 2nd independent variable, etc.
2. The formula used for the equation
3. A named vector of coefficients
4. The degrees of freedom of the residual
5. This is outside of the scope of this course, see the R script for details
6. The fitted mean y values used to plot the best fit line
7. The table of the x and y values used to plot the best fit line
8. Values for QR decomposition, not in the scope of this course
9. Number of linearly independent columns in the model, we will discuss this later
10. The values of the residuals in the model
11. Comes into play when categorical variables are used in regression, we will look at this later

How good is our prediction?

- Looks like the data points are widely dispersed (the residuals are large)
- How can we measure accuracy of our linear model?



Measuring errors: variance

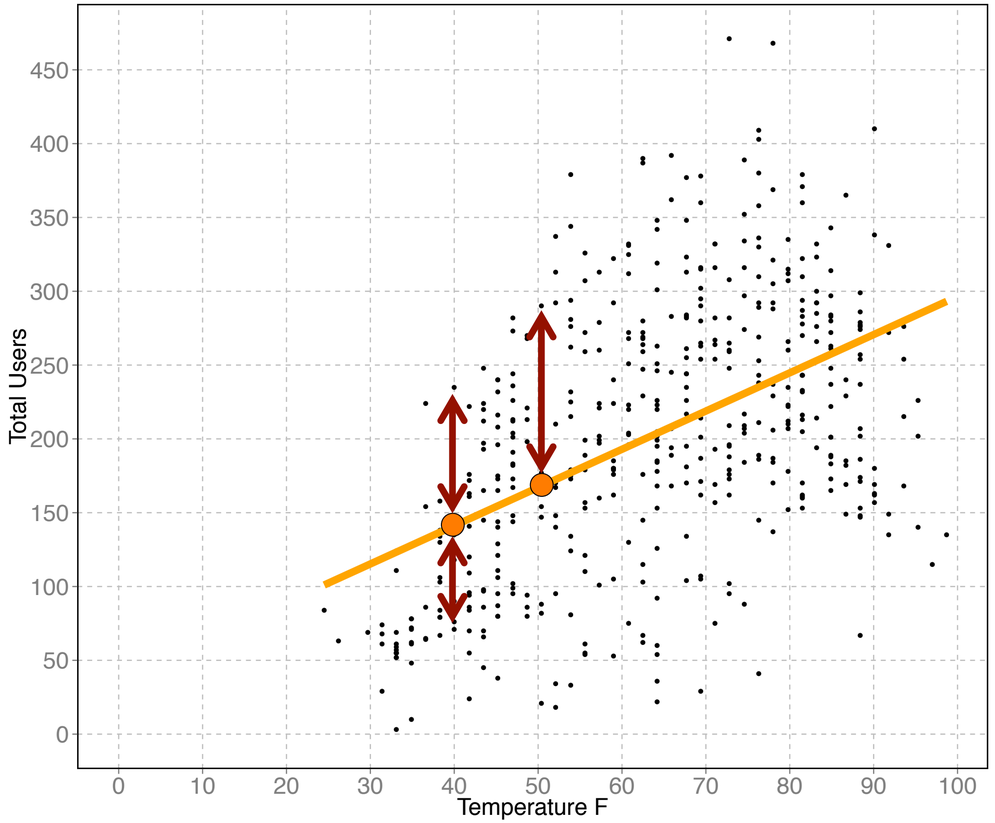
- Variance = $\frac{(\text{actual data point} - \text{expected data point})^2}{\text{Number of data points}}$ = average squared deviation from the mean

This can also be the average

- Variance is denoted by σ^2 "sigma" squared

- Variance = on average, how widely actual data is dispersed around [the predicted values, the mean, etc.]

- *The higher the variance of the residuals the less accurate the model*

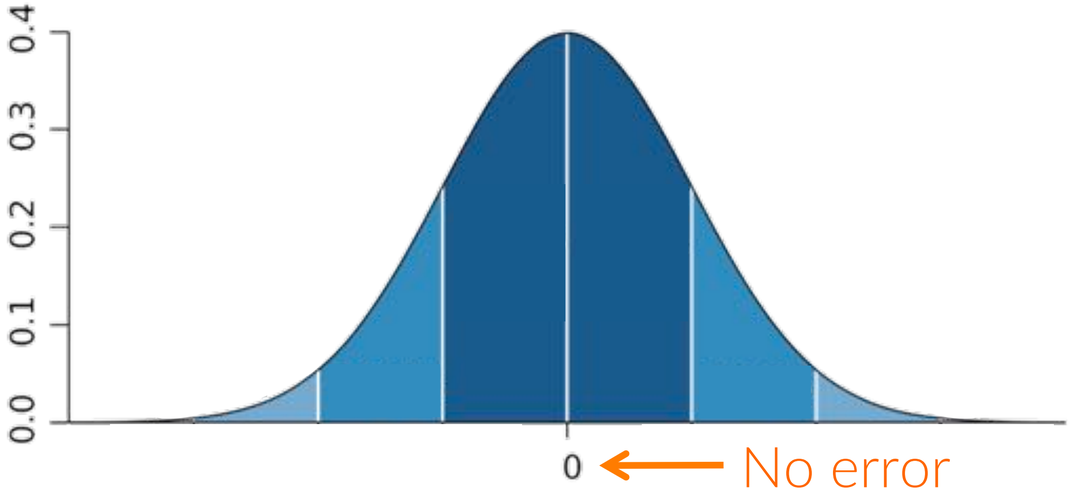


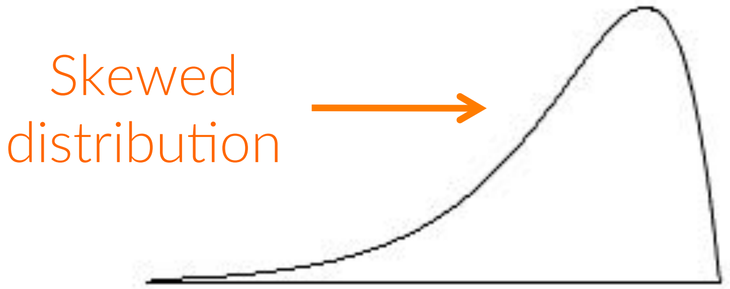
A

Measuring errors: randomness

- In a non-biased model errors will be random
- If errors are not random it indicates that there is a "bias" in the model
- If errors are not random it means you're not taking something into account
- Random errors follow a bell curve
  - The bell curve is called a "normal" distribution

"Normal distribution" of errors



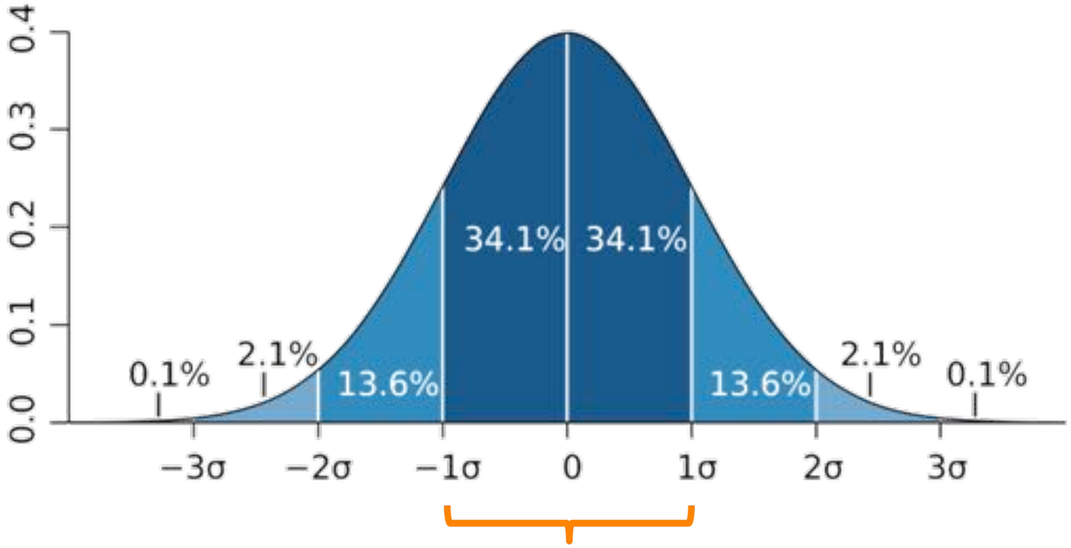


A

Measuring errors: standard deviation

- Standard deviation = $\sqrt{\text{variance}} = \sigma$
- Standard deviation is a standardized measure of how dispersed data points are around the average or the expected value
- Standard deviation tells you what proportion of data points falls within a given range
- *The smaller the standard deviation of the residuals the more accurate the model*

"Normal distribution" of errors



- 68.2% of errors are within 1σ away from the average or best fit line
- 95.4% of errors are within 2σ
- 99.6% of errors are within 3σ

A

Measuring errors: certainty

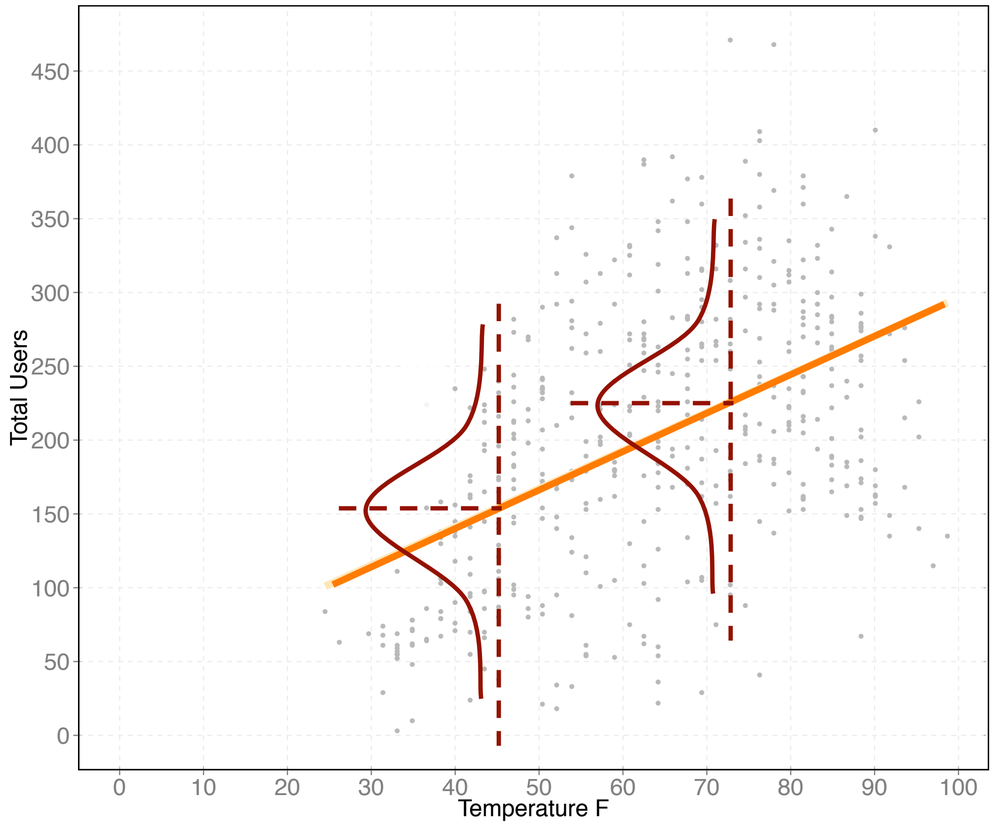
- Understanding standard deviation tells you
  the likelihood that a value will be within a given range!

- A 95% confidence interval is within 2σ away from the average
  (or expected value)

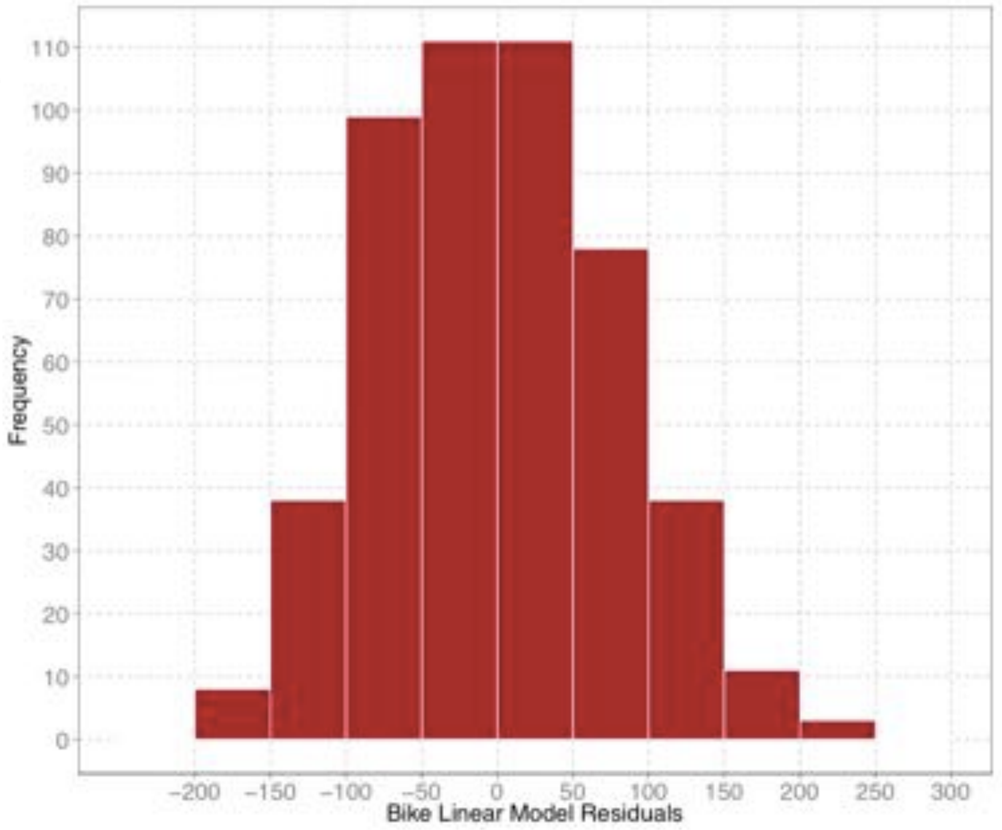
- The objective of a good model: decrease the standard deviation of the residuals
  - Put another way, to explain as much of the variance as possible

Understanding residuals

- We can check that residuals are normally distributed by plotting them as a histogram: no bias in our model!



Pooling the residuals together



A

Plot the residuals

Script

```
# Create a new variable to represent residuals. Cast is as a data frame
# so that you can tell ggplot which column to plot on the x-axis.
lm_weekday_noon_temp_resid = as.data.frame(lm_weekday_noon_temp$residuals)

# Plot the distribution of the residuals.
ggplot(lm_weekday_noon_temp_resid,
       aes(x = lm_weekday_noon_temp_resid[, 1])) +
  geom_bar(color = "white",
           fill = "brown",
           binwidth = 50) +
  scale_y_continuous(breaks = seq(0, 500, by = 50)) +
  scale_x_continuous(breaks = seq(0, 100, by = 10)) +
  theme(text = element_text(size = 20),
        panel.border = element_rect(color = "black", fill = NA, size = 1),
        panel.background = element_rect(fill = "white"),
        panel.grid.minor = element_line(color = NA),
        panel.grid.major = element_line(color = "grey",
                                        linetype = "dashed")) +
  labs(y = "Frequency",
       x = "Residuals")
```

1. Make the outline of the bars white
2. Color the bars brown
3. Each bar should encompass 50 units
4. Y-axis label
5. X-axis label

A

Plot the residuals