Sports Analytics

Aaron Nielsen, Department of Statistics, Colorado State University

2022-05-26

Contents

About

This book serves as the course textbook for the following courses at Colorado State University:

- STAT 351 (Sports Statistics and Analytics 1)
- STAT 451 (Sports Statistics and Analytics 2)

CSU students contributed to the creation of this book. Many thanks to the following student collaborators:

- Levi Kipp
- Ellie Martinez
- Isaac Moorman

Current Tasks

Updated: “2022-05-26”

Team Tasks and Tips

1. Find datasets from various sports to use as examples for EDA and later chapters
2. Show how to get basic sumamry statistics from these datasets using dplyr, tidy
3. Describe and calculate useful team and individual (descriptive statistics). Example: Baseball: calculate AVG, OBP, OPS, WOBA
4. (High quality) Visualizations using ggplot
5. Look for relevant “sports” R packages
6. Include examples from CSU and Colorado sports teams when possible
7. Sports to be included: Baseball/Softball, Football, Basketball, Soccer, Hockey, Volleyball
8. Sports to be potentially included: Lacrosse, Cricket, Handball,

---

Aaron:

Sports:

Chapters: Currently working to add content to chapters 1-4

Ellie:

Sports: Soccer, Volleyball

Chapters: EDA, Probability

Levi:

Sports: Basketball, Hockey

Chapters: EDA, Probability

Isaac:

Sports: Baseball, Football, Tennis

Chapters: EDA, Scraping

Chapter 1

Exploratory Data Analysis

1.1 Getting Started With R

1.1.1 Installing R

For this class, you will be using R Studio to complete statistical analyses on your computer.

To begin using R Studio, you will need to install "R" first and then install "R Studio" on your computer.

Step 1: Download R
(a) Visit https://www.r-project.org/
(b) Click **CRAN** under **Download** (c) Select any of the mirrors
(d) Click the appropriate link for your type of system (Mac, Windows, Linux)
(e) Download R on this next page.
(For Windows, this will say **install R for the first time**. For Mac, this will be under **Latest release** and will be something like **R-4.1.0.pkg** – the numbers may differ depending on the most recent version)
(f) Install R on your computer

Step 2: Download R Studio
(a) Visit https://www.rstudio.com/products/rstudio/download/#download
(b) Click to download
(c) Install R Studio on your computer

Step 3: Verify R Studio is working
(a) Open R Studio

(b) Let's enter a small dataset and calculate the average to make sure everything is working correctly.
(c) In the console, type in the following dataset of Sammy Sosa's season home run totals from 1998–2002:

```
sosa.HR <- c(66,63,50,64,49)
```

(d) In the console, calculate the average season home run total for Sammy Sosa between 1998–2002:

```
mean(sosa.HR)
```

```
## [1] 58.4
```

(e) Did you find Slammin' Sammy's average home run total from 1998–2002 was 58.4? If so, you should be set up correctly!

1.1.2 Some R Basics

For the following examples, let's consider Peyton Manning's career with the Denver Broncos. In his four seasons with the Broncos, Manning's passing yard totals were: 4659, 5477, 4727, 2249. Let's enter this data into R. To enter a vector of data, use the **c()** function.

```
peyton <- c(4659, 5477, 4727, 2249)
```

To look at the data you just put in the variable *peyton*, type *peyton* into the console and press enter.

```
peyton
```

```
## [1] 4659 5477 4727 2249
```

Some basic function for calculating summary statistics include **summary**, **mean()**, **median()**, **var()**, and **sd()**.

```
summary(peyton)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2249    4056    4693    4278    4914    5477
```

```
mean(peyton)
```

```
## [1] 4278
```

```
sd(peyton)
```

```
## [1] 1402.522
```

R allows you to install additional packages (collections of functions) that aren't offered in the base version of R. To install a package, use **install.packages()** and to load a package, use **library()**.

One package that we will use frequently is **tidyverse**. This package includes several other packages and functions such as **ggplot** (plotting function), **dplyr** (data manipulation package), and **stringr** (string manipulation package).

```
install.packages("tidyverse")
library("tidyverse")
```

You will also need to know how to load datasets from files. For this class, we will typically provide data files is .csv format.

Here is how to load a file:

```
# load readr package and load example dataset
library(readr)
NFL_2021_Team_Passing <- read_csv("data/NFL_2021_Team_Passing.csv")

# we can look at the header (first few entries) using "head()"
head(NFL_2021_Team_Passing)
```

```
## # A tibble: 6 x 25
##      Rk Tm             G   Cmp   Att `Cmp%`   Yds    TD `TD%`   Int `Int%`   Lng
##   <dbl> <chr>      <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1     1 Tampa Bay~    17   492   731   67.3  5229    43   5.9    12    1.6    62
## 2     2 Los Angel~    17   443   674   65.7  4800    38   5.6    15    2.2    72
## 3     3 Dallas Co~    17   444   647   68.6  4800    40   6.2    11    1.7    73
## 4     4 Kansas Ci~    17   448   675   66.4  4791    37   5.5    13    1.9    75
## 5     5 Los Angel~    17   406   607   66.9  4642    41   6.8    18    3      79
## 6     6 Las Vegas~    17   429   628   68.3  4567    23   3.7    14    2.2    61
## # ... with 13 more variables: `Y/A` <dbl>, `AY/A` <dbl>, `Y/C` <dbl>,
## #   `Y/G` <dbl>, Rate <dbl>, Sk <dbl>, SKYds <dbl>, `Sk%` <dbl>, `NY/A` <dbl>,
## #   `ANY/A` <dbl>, `4QC` <dbl>, GWD <dbl>, EXP <dbl>
```

1.2 Descriptive Statistics

1.2.1 Definitions

Definition 1.1. A ***population*** is a well-defined complete collection of objects.

Definition 1.2. A ***sample*** is a subset of the population.

Example 1.1. Suppose we are interested in studying Peyton's Manning's season passing yards totals. How could you define the population and what is one possible sample?

Definition 1.3. ***Quantitative data*** is numeric data or numbers. It can be broken into two further categories: discrete and continuous data.

Definition 1.4. ***Discrete data*** is quantitative data with a finite or countably infinite number of values.

Definition 1.5. ***Continuous data*** is quantitative data with an uncountably infinite number of values or data taken from an interval.

Example 1.2. What are possible discrete and continuous data associated with Peyton Manning?

Definition 1.6. ***Qualitative data*** refers to names, categories, or descriptions. It can also be broken down into two further categories, nominal data and ordinal data.

Definition 1.7. ***Nominal data*** is qualitative data with no natural ordering.

Definition 1.8. ***Ordinal data*** is qualitative data with a natural ordering.

Example 1.3. What are possible nominal and ordinal data associated with Peyton Manning?

1.2.2 Descriptive Statistics

While we will learn about some descriptive statistics that are unique to specific sports, there are some descriptive statistics that are frequently used in many applications.

1.2.2.1 Descriptive Statistics for Quantitative Data

There are different descriptive statistics depending on the type of data you are analyzing. We will begin by looking at descriptive statistics for quantitative data.

To begin, let $x_1, x_2, \ldots, x_n$ represent a numerical dataset with a sample of size n, where x_i is the i^{th} value in the dataset.

Definition 1.9. The ***sum*** of the data values is given by: $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$

Definition 1.10. The ***sample mean*** (or sample average), $\bar{x}$, of the numerical dataset is given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Definition 1.11. The ***population mean*** (or population average), μ, is the mean value for the entire population.

The mean can be thought of as a measure of center or more generally, a measure of location.

Example 1.4. Recall that Peyton Manning's season passing yards total while with the Broncos were: 4659, 5477, 4727, 2249. Calculate the sample mean of these values.

```r
# Calculate the sample of Peyton Manning's passing yards season totals with Colts
peyton.broncos <- c(4659, 5477, 4727, 2249)
mean(peyton.broncos)
```

```
## [1] 4278
```

In sports statistics, we often have to choose between using a descriptive statistic that summarizes a quantity versus a descriptive statistic that summarizes a rate. For instance, in basketball, we can compare two players based on how many points they score in a game (total quantity) or we can compare two players based on how many points per minute played (rate statistic). Many applications in sports analytics focus more on rate statistics rather than quantity statistics. Why?

We can measure the spread or variability of a dataset using *variance* and *standard devatiation.*

Definition 1.12. The ***sample variance***, s^2, of the numerical dataset is a measure of spread and is given by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Definition 1.13. The ***sample standard deviation***, s, of the numerical dataset is a measure of spread and is given by $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Definition 1.14. The ***population variance***, σ^2, is the variance for an entire population.

Definition 1.15. The ***population standard deviation***, σ, is the standard deviation for an entire population.

We often prefer to work with standard deviations as a measure of spread as opposed to variance because standard deviations are given in our original units.

```
# Calculate the variance and standard deviation of Peyton Manning's passing yards seas
var(peyton.broncos) # units: yards^2
```

```
## [1] 1967068
```

```
sd(peyton.broncos) # units: yards
```

```
## [1] 1402.522
```

Definition 1.16. The **sample median**, $\tilde{x}$, of a numerical dataset is the middle value when the data are ordered from smallest to largest. In other words, let $x_1, x_2, \ldots, x_n$ be the (unordered) dataset and let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the same dataset but ordered from smallest to largest. If n is odd, then $\tilde{x} = x_{(n+1)/2}$ and if n is even, then $\tilde{x} = \frac{1}{2} \cdot \left[x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}\right]$.

Example 1.5. Calculate the sample median of Peyton Manning's season passing yards total while with the Colts (3739, 4135, 4413, 4131, 4200, 4267, 4557, 3747, 4397, 4040, 4002, 4500, 4700).

Like sample mean, sample median is a measure of center. It gives you an idea of where the "middle" of your dataset is.

We can calculate sample mean and sample median in R as follows:

```
# Calculate the median of Peyton Manning's passing yards season totals with Broncos an
peyton.colts <- c(3739, 4135, 4413, 4131, 4200, 4267, 4557, 3747, 4397, 4040, 4002, 450
median(peyton.broncos)
```

```
## [1] 4693
```

```
median(peyton.colts)
```

```
## [1] 4200
```

Definition 1.17. A ***percentile*** is a measure of relative standing. The p^{th} percentile is the number where at least p% of the data values are less than or equal to this number.

Definition 1.18. A ***quantile*** is a measure of relative standing and are the cut points for breaking a distribution of values into equal sized bins.

Definition 1.19. A ***quartile*** is a measure of relative standing and are the cut points for breaking a distribution of values into four equal parts.

```
# Calculate the 10th and 90th percentile of Peyton Manning's passing yards season totals with Col
quantile(peyton.colts,0.10)
```

```
##  10%
## 3798
```

```
quantile(peyton.colts,0.90)
```

```
##    90%
## 4545.6
```

```
quantile(peyton.colts,c(0.1,0.9))
```

```
##    10%    90%
## 3798.0 4545.6
```

Special percentiles:
1. 25th percentile = 1st quartile = Q_1
2. 50th percentile = 2nd quartile = $Q_2 = \tilde{x}$
3. 75th percentile = 3rd quartile = Q_3

Definition 1.20. ***Range*** is a measure of spread, measures the full width of a dataset, and is given by: $Range = Max - Min$.

Definition 1.21. ***Interquartile range*** is a measure of spread, measures the width of the middle 50% of a dataset, and is given by: $IQR = Q_3 - Q_1$.

Definition 1.22. A ***five number summary*** describes the center, spread, and edges of a dataset and is given by: $(Min, Q_1, Q_2, Q_3, max)$.

```
summary(peyton.colts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3739    4040    4200    4218    4413    4700
```

```
quantile(peyton.colts,c(0,0.25,0.5,0.75,1))
```

```
##   0%  25%  50%  75% 100%
## 3739 4040 4200 4413 4700
```

1.2.2.2 Descriptive Statistics for Qualitative Data

In sports statistics, we also encounter qualitative (categorical) data which is names or labels which has its own descriptive statistics.

To begin, let $x_1, x_2, \ldots, x_n$ represent a categorical dataset with a sample of size n, where x_i is the i^{th} value in the dataset.

Definition 1.23. The ***proportion*** of sampled data that fall into a category is given by: $p = \frac{\text{\# in category}}{\text{\# total}}$

’’Proportion” and “Probability” are often used interchangeably. Both have a minimum value of 0 and a maximum value of 1.

Definition 1.24. The ***percentage*** of sampled data that fall into a category is given by: $P\% = 100 \cdot p = 100 \cdot \frac{\text{\# in category}}{\text{\# total}}$

Percentages in this context can have a minimum value of 0% and a maximum value of 100%.

Example 1.6. In 2014, Peyton Manning started as quarterback for the Denver Broncos. The result of the Broncos’ 16-game season was:
Win, Win, Loss, Win, Win, Win, Win, Loss, Win, Loss, Win, Win, Win, Win, Loss, Win

Calculate the proportion and percentage of Broncos’ winning games in 2014.

```
broncos2014 <- c("Win", "Win", "Loss", "Win", "Win", "Win", "Win", "Loss", "Win", "Loss
broncos.prop <- sum(broncos2014 == "Win")/length(broncos2014); broncos.prop
```

```
## [1] 0.75
```

```
broncos.perc <- 100*broncos.prop; broncos.perc
```

```
## [1] 75
```

We can also build a frequency table that summarizes the categories and their occurrences using **table()** in R. Note that **table()** works for quantitative and qualitative data.

```
table(broncos2014)
```

```
## broncos2014
## Loss  Win
##    4   12
```

1.3 Visualizations

Conveying information visually is also an important part in providing a description of a dataset.

R provides some basic plotting functions such as **plot**, **hist**, and **barplot**. These plotting functions are simple and not always very clean looking.

In this class, we will use analogous plotting functions in **ggplot2** that are much improved plotting functions.

If you have already installed the **tidyverse** package, it should have also installed the **ggplot2** package.

```
# You have likely already installed the tidyverse package but if not, use the following command u
# install.packages("tidyverse")

# Load the tidyverse package (which includes ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.7     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

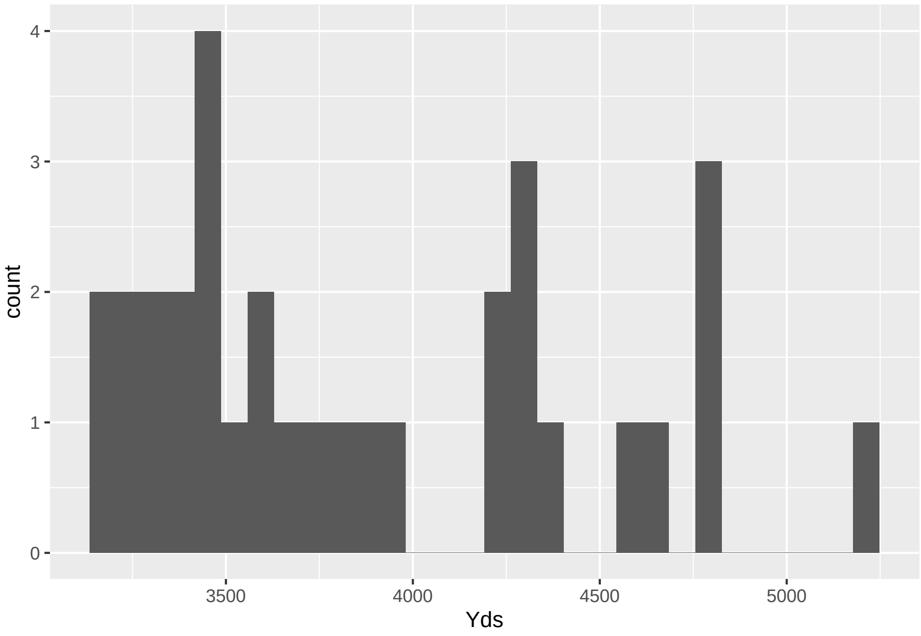
Let's load the file "NFL_2021_Team_Passing.csv" which contains NFL Team Passing Statistics, 2021

```
library(readr)
NFL_2021_Team_Passing <- read_csv("data/NFL_2021_Team_Passing.csv")
```

Histograms are one of the most common and basic ways to visualize a dataset's distribution of values. To make a histogram, you will use **ggplot** and **geom_histogram**.

Example 1.7. Create a histogram of the NFL Team Passing Yards in 2021.

```
NFL_2021_Team_Passing %>% ggplot(aes(x=Yds)) + geom_histogram()
```

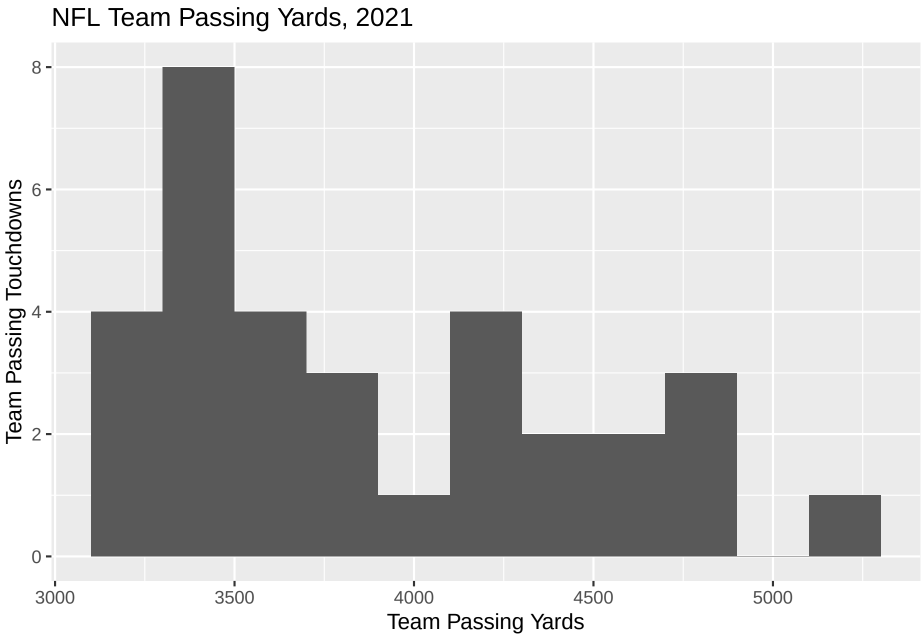


Notice how %>% is used to **pipe** the dataset into ggplot. This is using the pipe function from the **dplyr** package.

By default, **geom_histogram** uses 30 bins but this is customizable. Let's make the bins have a width of 200.

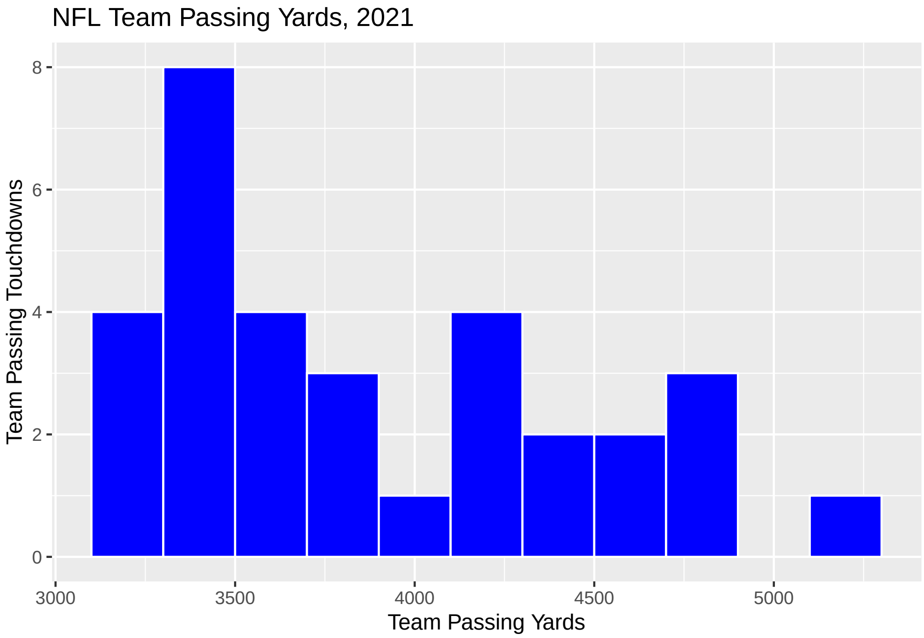
All good visualizations have good labels. Let's improve the axis labels and give the figure a title.

```
NFL_2021_Team_Passing %>% ggplot(aes(x=Yds)) +
  geom_histogram(binwidth = 200) +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards
```



We also have numerous options to change the appearance of plots when using **ggplot**. Let's change the bins color to *blue* and change the bin borders to *white*.

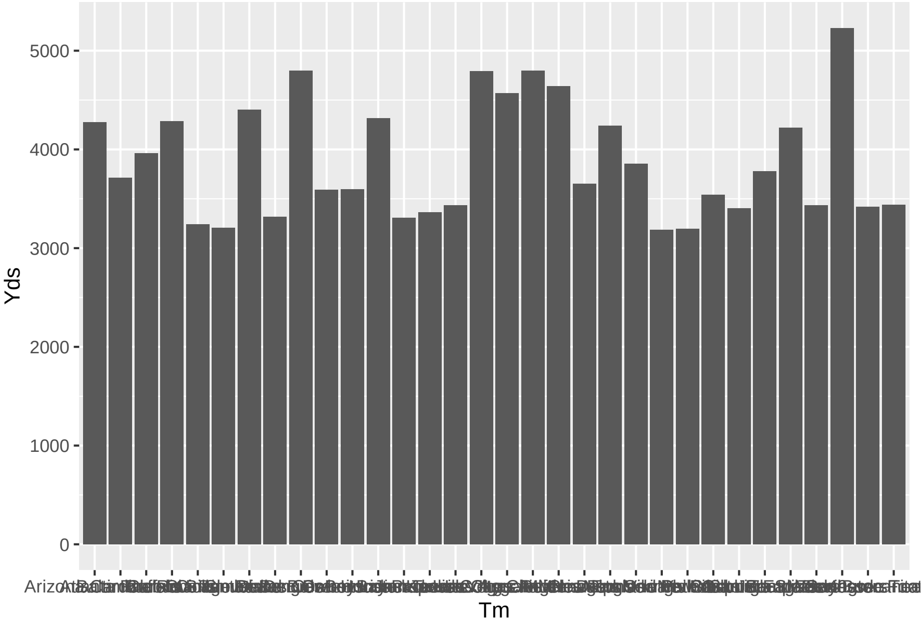
```
NFL_2021_Team_Passing %>% ggplot(aes(x=Yds)) +
  geom_histogram(color = "white", fill = "blue", binwidth = 200) +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards, 2021")
```



We can also create bar plots using ggplot using the **geom__bar** function.

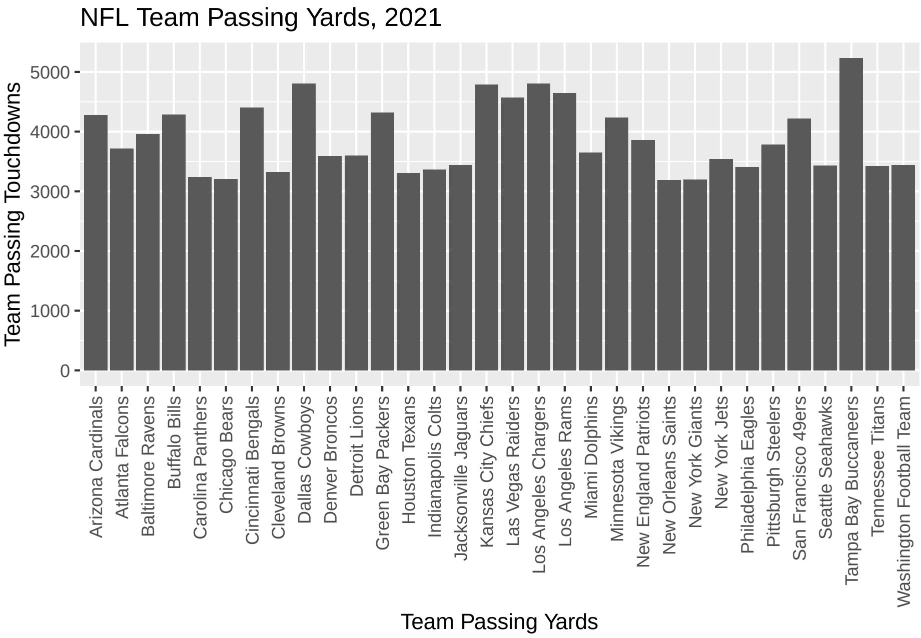
Example 1.8. Create a bar plot with teams on the horizontal axis and passing touchdowns on the vertical axis.

```
NFL_2021_Team_Passing %>% ggplot(aes(x=Tm,y=Yds)) +
  geom_bar(stat="identity")
```



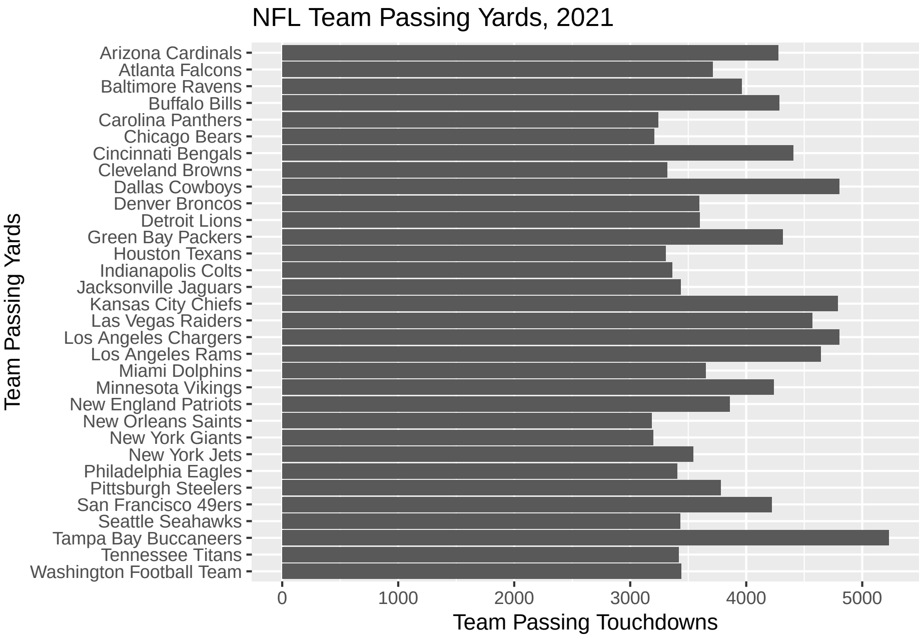
The team labels are a complete mess. Let's fix this and make some adjustments to the axis labels and figure title.

```
NFL_2021_Team_Passing %>% ggplot(aes(x=Tm,y=Yds)) +
  geom_bar(stat="identity") +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards, 2021") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



We can flip this graph if we like as well. Note that when we flip the graph, our labels get in reverse ordering, so this can be fixed using **fct_rev()** which is part of the **forcats** package.

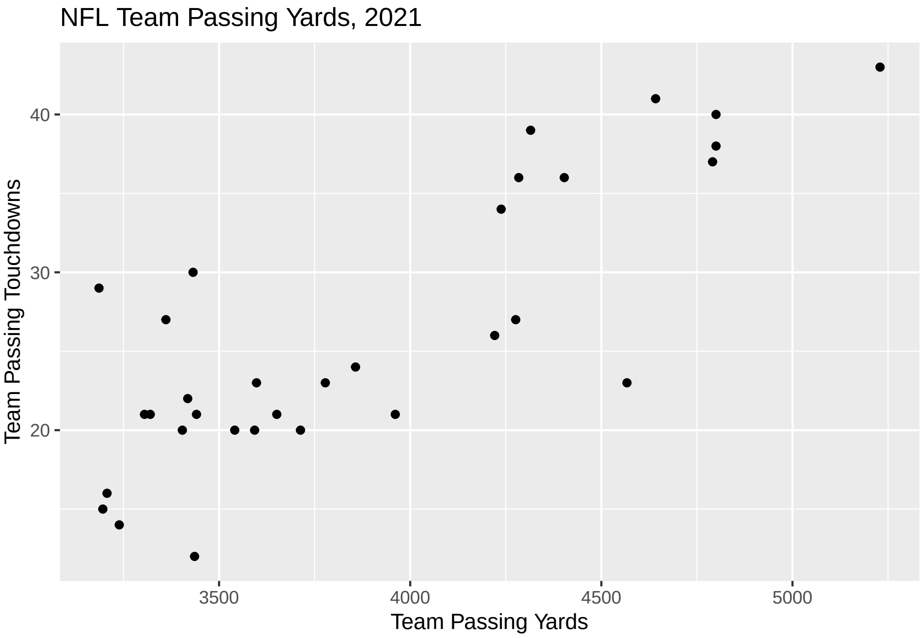
```
NFL_2021_Team_Passing %>%
  ggplot(aes(x=fct_rev(Tm),y=Yds)) +
  geom_bar(stat="identity") +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards
  coord_flip()
```



Another common and useful visualization is a scatterplot which shows the relationship between two numeric variable. In ggplot, you use **geom_point()**.

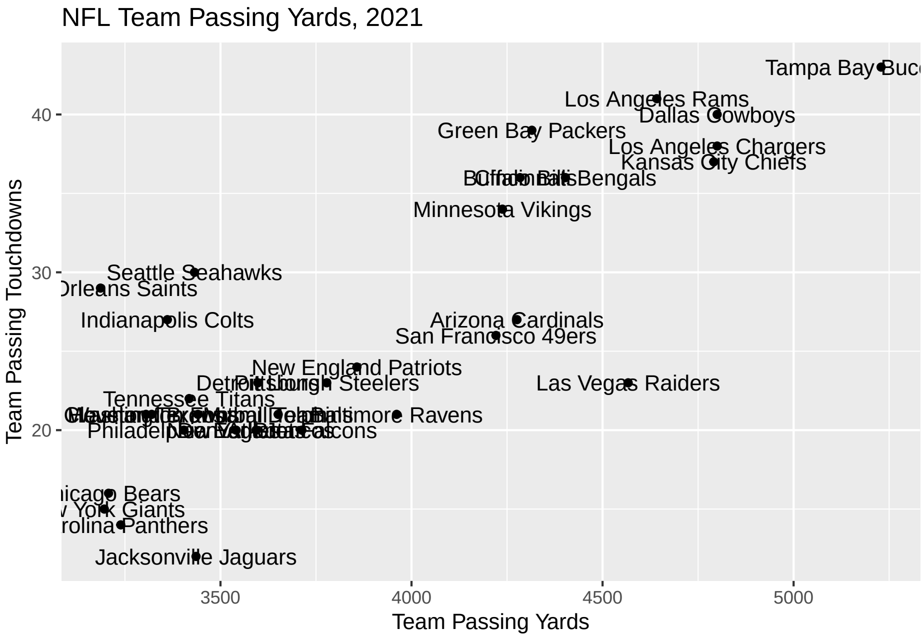
Example 1.9. Create a scatterplot of Team Passing Yards and Team Passing Touchdowns from the NFL 2021 dataset.

```
NFL_2021_Team_Passing %>%
  ggplot(aes(x=Yds,y=TD,label=Tm)) +
  geom_point() +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards, 2021")
```



We may want to include team labels on this plot, however, it can get messy very quickly with a lot of points.

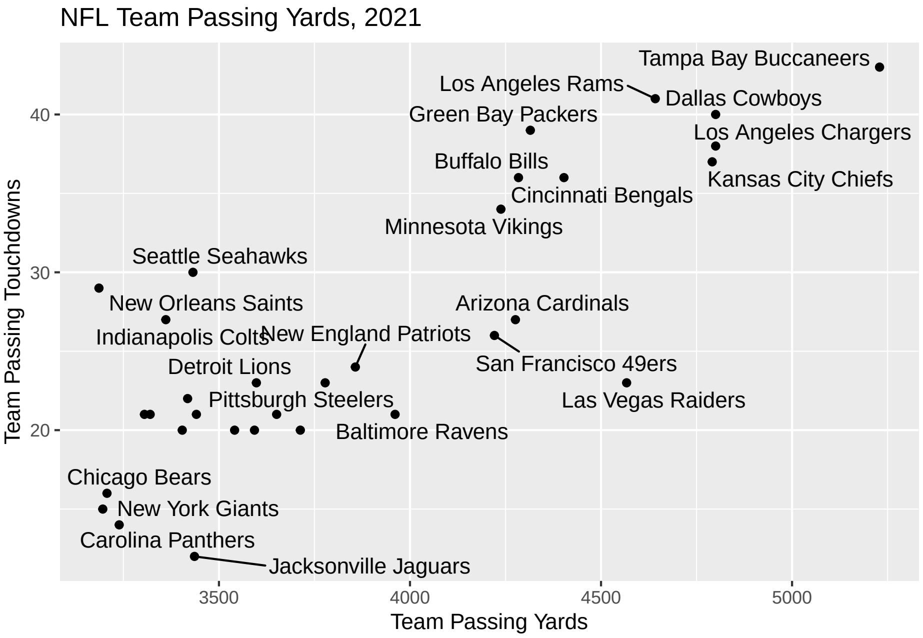
```
NFL_2021_Team_Passing %>%
  ggplot(aes(x=Yds,y=TD,label=Tm)) +
  geom_point() +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards
  geom_text()
```



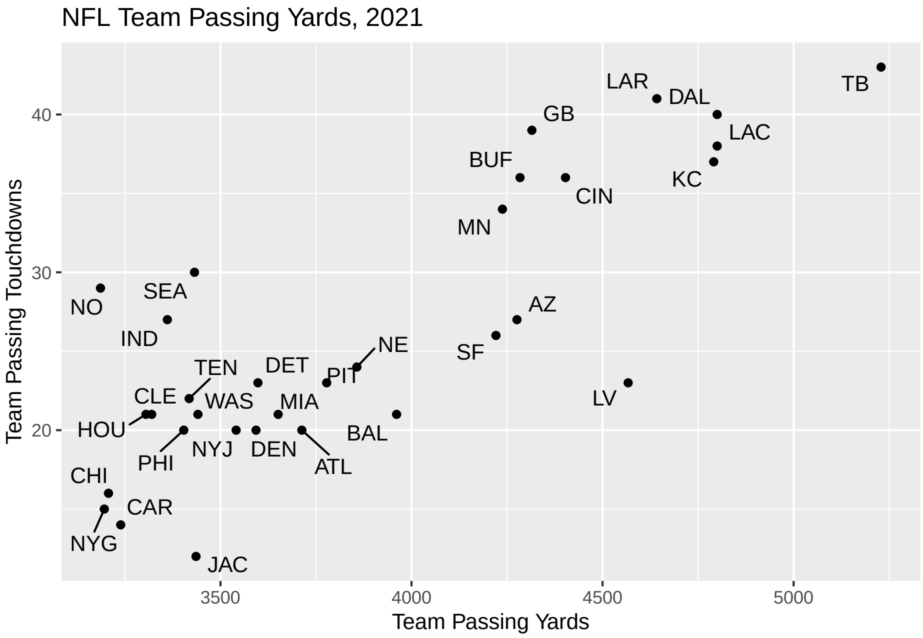
Many sports leagues have around 30 teams, so a clean scatterplot with labels can be tricky to make. Here are some options below.

```
# install ggrepel package
library(ggrepel)
NFL_2021_Team_Passing %>%
  ggplot(aes(x=Yds,y=TD,label=Tm)) +
  geom_point() +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards, 2021") +
  geom_text_repel()
```

```
## Warning: ggrepel: 9 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
NFL_2021_Team_Passing$Abbr <- c("TB","LAC","DAL","KC","LAR","LV","CIN","GB","BUF","AZ"
                                "BAL","NE","PIT","ATL","MIA","DET","DEN","NYJ","WAS","
                                "TEN","PHI","IND","CLE","HOU","CAR","CHI","NYG","NO")
NFL_2021_Team_Passing %>%
  ggplot(aes(x=Yds,y=TD,label=Abbr)) +
  geom_point() +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yard
  geom_text_repel(box.padding = 0.3)
```



1.4 Baseball

1.5 Football

1.6 Basketball

1.7 Soccer

1.8 Volleyball

1.9 Hockey

Chapter 2

Probability

2.1 Definitions and Axioms

2.2 Theorems and Laws

2.3 Random Variables

Chapter 3

Simulation

Chapter 4

Statistical Inference

4.1 One Sample and Two Sample t-tests and confidence intervals

Chapter 5

Correlation

Chapter 6

Linear Regression

Chapter 7

Data Scraping

Chapter 8

Principal Component Analysis

Chapter 9

Clustering

Chapter 10

Classification

Chapter 11

Decision Trees

11.1 Random Forests

11.2 Gradient Boosting

Chapter 12

Non-parametric Statistics

Chapter 13

Baseball

Chapter 14

Football

Chapter 15

Basketball

Chapter 16

Soccer

Chapter 17

Hockey

Chapter 18

Volleyball

18.1 Resources

Women's Volleyball D1 Statistics

Chapter 19

Other Sports

Chapter 20

Ellie's stuff

Chapter 21

Levi's stuff

Chapter 22

Isaac's stuff

Chapter 23

Aaron's stuff

23.1 Notes for Chapter 2 (Probability)

Axioms of Probability:

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \ldots, A_n$ are disjoint events, then $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$

Theorem 23.1 (Bayes theorem)**.** *Let A and B be events in Ω such that $P(B) > 0$. Then we have the following:*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

23.2 Suggested Readings

23.2.1 Moneyball

Moneyball, Chapter 2, How to Find a Ballplayer (Lewis, 2004)

Near the end of the chapter (page 40), Michael Lewis give a list of players the Oakland Athletics hoped to draft. How did these players turn out? Find the WAR for each of the players in their pre-free agency years and compare it against the Rockies draft picks in the same rounds from the same draft.

23.2.2 Future Value

Future Value, Chapter 7, How to Scout (Longenhagen and McDaniel, 2020)

If a player receives a running grade of 40, approximately what proportion of MLB players have a lower have a lower running grade?

For a given tool, about 95% of all player grades fall between what two bounds? (Consider the middle 95% of the distribution of grades.)

23.3 Notes for Chapter 4 (Simulation)

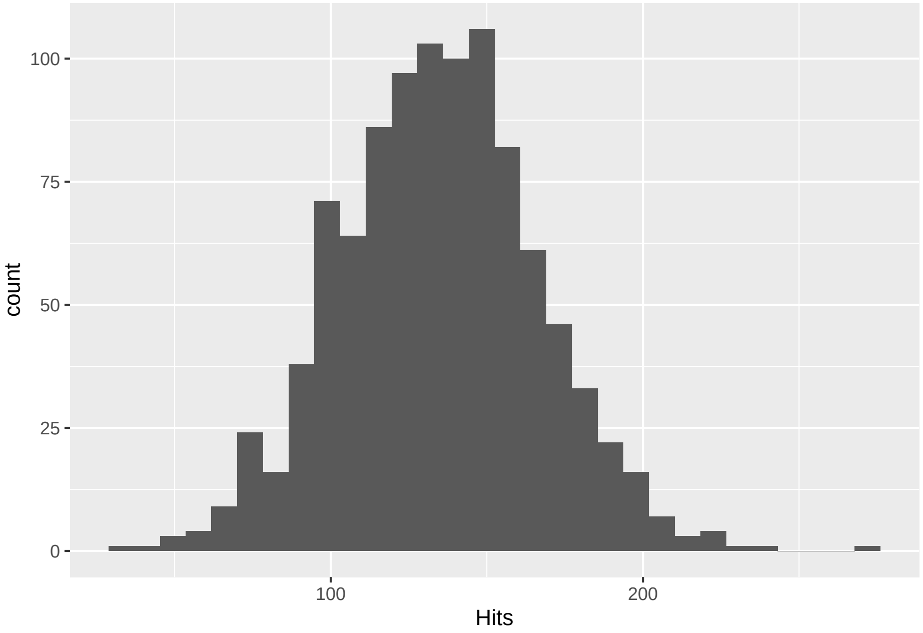
23.3.1 Baseball Simulation Example

```
library(tidyverse)
```

This is a baseball example for chapter 4.

```
set.seed(2022)
n.sims <- 1000
hits <- rep(0,n.sims)
avg <- 0.300
atbats.mean <- 450
atbats.sd <- 100
sim.atbats <- round(rnorm(n.sims,atbats.mean,atbats.sd))

for(i in 1:n.sims){
  sim.hits <- rbinom(1,sim.atbats[i],avg)
  hits[i] = sim.hits
}
hits.df <- data.frame(Hits=hits)
hits.df %>% ggplot(aes(x=Hits)) + geom_histogram()
```



Reference: Blocks

23.4 Equations

Here is an equation.

$$f\left(k\right) = \binom{n}{k} p^k\left(1-p\right)^{n-k} \tag{23.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (23.1).

23.5 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 23.2.

Theorem 23.2. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

23.6 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Reference: Footnotes and citations

23.7 Footnotes

Footnotes are put inside the square brackets after a caret `^[]`. Like this one [1].

23.8 Citations

Reference items in your bibliography file(s) using `@key`.

For example, we are using the **bookdown** package (Xie, 2016) (check out the last code chunk in index.Rmd to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015) (this citation was added manually in an external file book.bib). Note that the `.bib` files need to be listed in the index.Rmd with the YAML `bibliography` key.

The RStudio Visual Markdown Editor can also make it easier to insert citations: https://rstudio.github.io/visual-markdown-editing/#/citations

[1] This is a footnote.

Chapter 24

References

Bibliography

Lewis, M. (2004). *Moneyball: The art of winning an unfair game.* WW Norton & Company.

Longenhagen, E. and McDaniel, K. (2020). *Future Value: The battle for baseball's soul and how teams will find the next superstar.* Triumph Books.

Xie, Y. (2015). *Dynamic Documents with R and knitr.* Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown.* R package version 0.3.9.