Psych 100A Spring 2019: Week 1 Slides

Amanda Montoya

March 31, 2019

Psych 100A: Psychological Statistics

- **Textbook/Homework**: Please enroll for the canvas site at: https://canvas.instructure.com/enroll/H67TL3
- **PollEverywhere**: You should have received an email invitation. You need to make an account. The email was sent to your official university account. If you don't know what account that is you can use https://moodle2.sscnet.ucla.edu/docs/How_do_I_update_my_official_email_addressto look it up.
- **AskForMe**: Ask for me is a way of submitting your questions anonymously. One of our brilliant TAs will manage the page and ask questions on your behalf. You can also just raise your hand if you want! http://bit.ly/AFMMontoya100A
- **Course Dashboard**: If you go to http://bit.ly/amanda100Aclass you will get a course dashboard with PollEverywhere, AskforMe, and R in DataCamp.

Welcome to Psych 100A: Psychological Statistics

Instructor: Amanda K. Montoya

Office Hours: M(10-11), W(12-1)

Figure 1: Amanda Montoya

Welcome to Psych 100A: Psychological Statistics

TA: Meredith Boyd

Office Hours: Th (11-1)

Figure 2: Meredith Boyd

Welcome to Psych 100A: Psychological Statistics

TA: Isabella Boyadjian

Office Hours:Tu (11-1)

Figure 3: Isabella Boyadjian

Today's Learning Outcomes

- Get connected to Canvas
- Get connected to PollEverywhere (and test it out)
- Familiarize yourself some of the big ideas of statistics
- Course Requirements and Grading
- A quick introduction to R

Connecting to Canvas

- **Textbook/Homework**: Please enroll for the canvas site at: https://canvas.instructure.com/enroll/H67TL3
- Canvas textbook used for both "Reading" and "Homework"
- Information is *interleaved* with questions and `R` exercises.
- Research suggests that *interleaving* helps with learning by testing your understanding as you develop it in a **low stakes** environment.
- Homework is graded on completion not correctness (but I will check for non-sense)

PollEverywhere

- **PollEverywhere**: You should have received an email invitation. You need to make an account. The email was sent to your official university account. If you don't know what account that is you can use This Resource to look it up.
- PollEverywhere is a free alternative to clickers, and provides more flexibility in what can be done. Let's give it a try.

PollEverywhere

[Insert Poll]

Statistics is about understanding variation

Statistics is an incredibly useful tool for understanding what's going on in the world.

The world is full of variation that's difficult to make sense of without tools like statistics.

Variation is when we see differences among the same types of things (observations).

Figure 4: Variation in Shell Coloring

Statistics is about understanding variation

We can see variation in this class here. Our **cases** are students in Psych 100A

You're all different on many **variables**:

- Hair color
- Experience with statistics
- What brand of computer you use
- Where you went to high school
- Height
- etc.

Variation doesn't mean that everyone has to be different (e.g., some of you may have gone to the same high school). Lack of variation is when everyone is the same on some variable.

When is there not variation?

In this class, there is not variation in your statistics instructor. So we can't look at variation in statistics instructor.

At UCLA, we can't study variation in undergraduate institution, because everyone is at the same institution.

If we looked at everyone on the men's football team, we likely wouldn't find much variation in gender.

Most of these examples are pretty contrived, because most of the time in reality, there is a lot of variability.

How can we use variability in statistics?

Most of the time when we observe variability in the world, we can then try to **measure** it and record it as **data**.

Then we take the data and **analyze** it using statistics.

Statistics helps us take our observations and apply them to the real world.

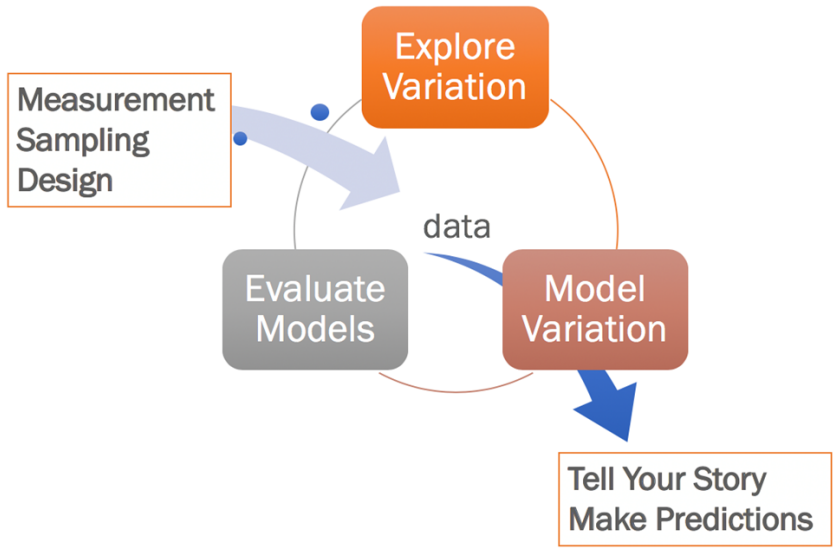


Figure 5: The Statistics Cycle

Elements of This Class

This class is broken up into three sections which are meant to align with the Statistics Cycle:

- Exploring Variation
- Modeling Variation
- Evaluating Models

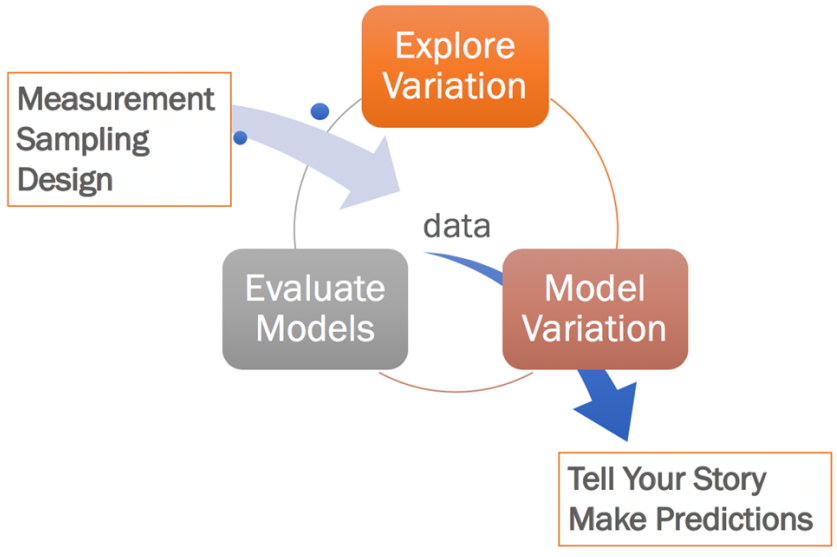


Figure 6: The Statistics Cycle

How this class fits with others

Up to this point in psychology, your education has been focused on learning facts or processes which are important findings in psychology.

Now you're transitioning to the training which asks you to "Think like a researcher." You begin to learn the answer to the question **How do we learn new facts about psychology?**

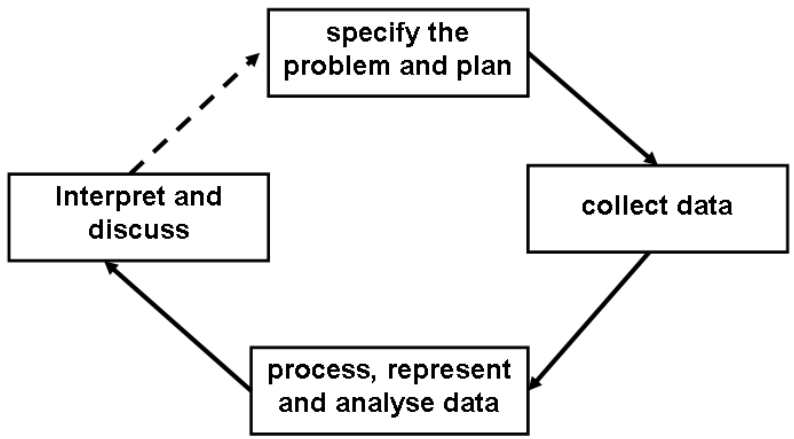


Figure 7: The Research Cycle

How this class fits with others

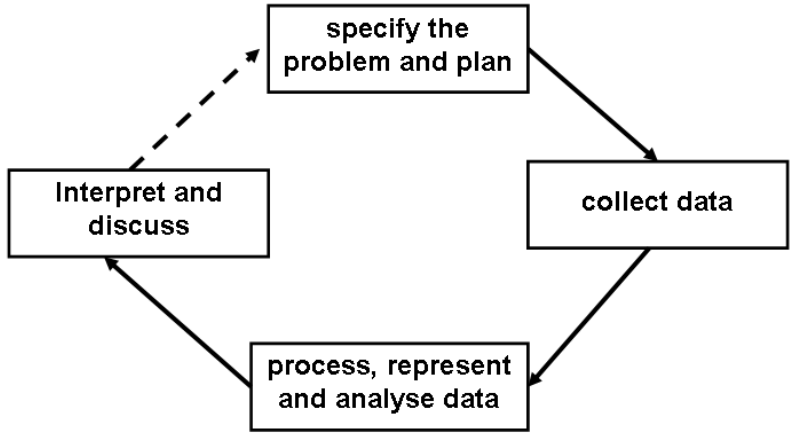


Figure 8: The Research Cycle

Psych 100B: Research Design will teach you to "Specify a problem and a plan" and how to "Collect Data."

We find it's helpful to know what you might do with the data first, so in 100A we teach you how to take data, and **process, represent, and analyze data**.

If you're interested in how we measure things in psychology, check out Psych M144

Starting with Data

For the most part in this class we start with data that's already collected for some specific purpose. Let's start by thinking about what data is and how it's collected.

Each piece of information has three important characteristics:

- The **case** it came from (person, school, company)
- The **variable** being measured
- The **value** of the observation

For example: Alejandro has taken 2 courses in statistics before.

- Case: Alejandro
- Variable: Courses in statistics
- Value: 2

Starting with Data

Think, Pair, Share:

Think about the two examples below, and try to identify the case, variable, and value. **Pair** up with someone(s) around you, and talk about what you think. If you don't agree, try to understand what you think is different. **Share** by raising your hand to indicate what your group thought.

Example 1: Of all the companies in the world, Apple makes the most money at \$45.7B/year

Example 2: When looking at all the regions of the brain, the cerebrum makes up 75% of brain volume.

Food for Thought: What other cases do you think might be included in a dataset with each example? What other variables do you think be included in each example?

What can we do with Data

Once we have data what do we do with it? That's largely the purpose of this class.

We will start with *exploring variation*. This involves creating **numeric summaries** of data and creating **visual depictions** of data.

Let's look at an example. This data comes from the National Longitudinal Study examining adolescents through development into adulthood.

In this data there are 4224 people selected to be a close representation of the US. We have data for 15 variables for them.

One of the variables we might look at is how many hours of sleep people got.

Exploring Variation: Example

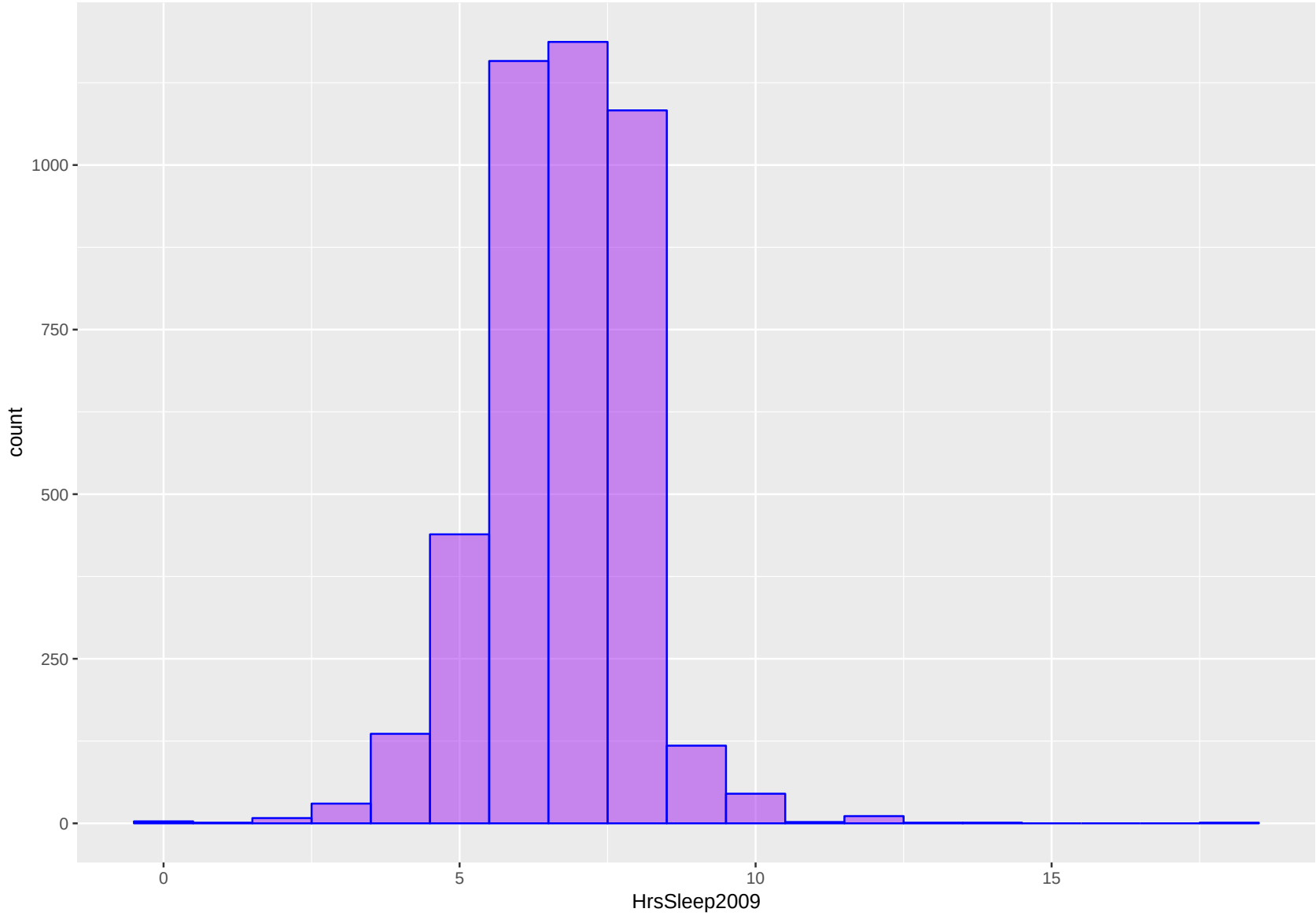
By the end of next week you'll be able to create *numeric summaries* and visualizations of data.

```
##
##    0    1    2    3    4    5    6    7    8    9
##    3    1    8   30  136  439 1158 1187 1083  118
```

```
##
## 10 11 12 13 14 18
## 45  2 11  1  1  1
```

Exploring Variation: Example

By the end of next week you'll be able to create visualizations and numeric summaries of data.



Modeling Data

In this class we'll use the *general linear model* (GLM) for creating and describing all models in this class. The general linear model is used to explain variation in a particular outcome, and can be expressed like this:

$$Outcome = Model + Error$$

Each different type of test that we do uses a different *Model*, however, the structure is always the same. So if you've taken a class before that describes a variety of tests, for example:

- z-test
- t-test
- regression
- ANOVA

These can all be expressed as a general linear model, so we use this approach to teach all of these different types of tests.

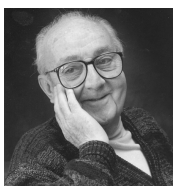
Why do we make models?

The term "model" is used a lot in statistics. But it isn't the only place that we use this term. And it's important to realize that it means the same thing inside and outside statistics.

Models are a representation of something used to approximate it in a certain way.

Think of a model car. That may be useful for understanding the proportions of a certain car, but not for understanding how cars move in traffic.

A globe is a physical model of the earth. It's good for understanding where certain places are in relation to each other, but perhaps not great for understanding the different levels of the earths core.

"All models are wrong but some are useful" ~ George Box 

Why do we need models?

Statistical models are very useful for understanding what's going on in the world. I can think of three ways that models can be used:

1. Approximating the **Data Generating Process** (DGP). We can use a statistical model as an **attempt** to approximate how we think the data comes about in the world.

Example: If I think social media use influences depression in kids, I may use social media use as a **predictor** of the **outcome** depression.

$$Model_{FromData} \approx Model_{IntheWorld}$$

Why do we need models?

Statistical models are very useful for understanding what's going on in the world. I can think of three ways that models can be used:

2. Improving complex systems. When a system is complex, many factors can influence the outcomes. If we can identify some of those factors and manipulate them, we can improve the efficiency of the system.

Example: Many factors may come together to result in depression for an individual, but if I can learn that developing strong relationships with peers can prevent depression, we may encourage kids who are "at-risk" to develop stronger friendships.

Figure 9: The Brain is a Complex System

Why do we need models?

Statistical models are very useful for understanding what's going on in the world. I can think of three ways that models can be used:

3. Predicting the future. Given information about specific predictors of an outcome, we may be able to guess the outcome fairly accurately.

Example: If we have been able to explain depression using a set of predictors in current data, then we can use our model to predict future outcomes. This means we may be able to identify kids who are most "at risk" and focus resources toward them.

Figure 10: Statistics: Becoming a Psychic

Checking our understanding

Let's imagine that we're interested in March Madness. March Madness is a series of basketball games where teams are paired up, then the winning team moves on in the "bracket." Lots of people make statistical models to figure out which team will win, which is very difficult because you have to guess the outcome of many games.

Use PollEverywhere to indicate which *purpose* you think these statistical models have: Approximating the DGP, Improving a complex system, or predicting future outcomes.

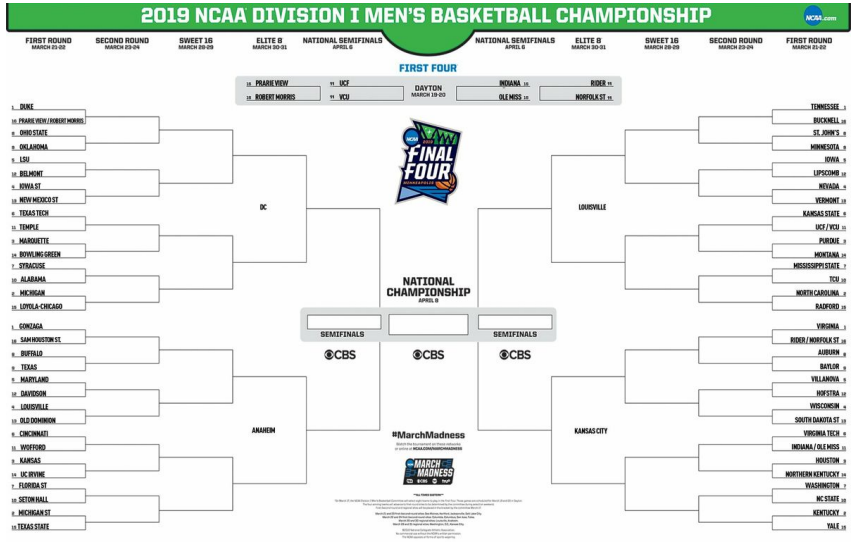


Figure 11: March Madness 2019 Bracket

Evaluating Models

It's one thing to make a model, but it's another to know if it does a good job.

Some models don't explain much error (underfitting), while other models explain too much and don't generalize to new cases (overfitting).

In this class we'll learn to take advantage of *sampling variability* to compare models, and choose the ones that fit just right.

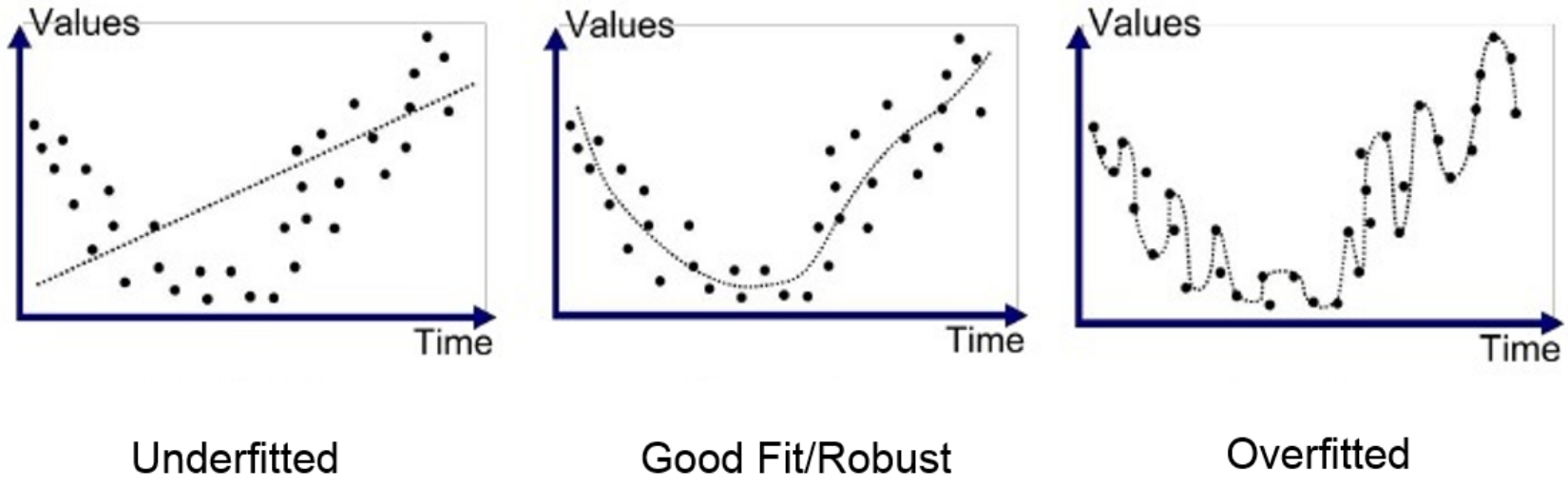


Figure 12: Over and underfitting

Using a Recipe vs. Cooking

Using statistics can look similar to cooking. There are some people who are very good at following the recipe, but the don't *understand* the process well enough to go off script.

Our goal is to get you to write your own recipes based on what you learn in the class, but ultimately feeling comfortable going **off script.**

"What if we did this another way? What would change, and what would stay the same?"

Course Evaluation

1. Homework (20%)

- Graded on completeness not correctness

2. In-class Activities (15%)

- PollEverywhere, Group Activites, etc.

3. Quizzes (20%)

- Four Quizzes, Every other Friday
- Drop lowest quiz score (No Makeups)

4. Midterm (20%)
5. Final (25%)

Quizzes, Midterm, and Final all allow Rcheatsheet and single page of handwritten notes. Nothing is graded on a curve. **Everyone can get an A in this class!**

Course Schedule

Course Schedule

This is a tentative schedule and subject to change, with schedule adjustments posted on CCLE announcements.

Week	Monday (HW Due)	Tuesday (In-Class)	Thursday (In-Class)	Friday (Lab Section)
1 (4/1 – 4/5)		Chpt 1,2: Modeling Approach & Understanding Data	Chpt 3,4: Examining Distributions & Explaining Variability	Lab Activity: Intro to R
2 (4/8 – 4/12)	Chapters 1,2,3,4	Chpt 1-4: Review and Recap	Chpt 5 (Intro): The Simple Model	Quiz: Chpt 1 – 4
3 (4/15 – 4/19)	Chapter 5	Chpt 5 (Outro): The Simple Model	Chpt 6 (Intro): Quantifying Error	Lab Activity: Exploring Sums of Squares
4 (4/22 – 4/26)	Chapter 6	Chpt 6 (Outro): Quantifying Error	Chpt 7 (Intro): Adding an Explanatory Variable	Quiz: Chpt 5/6
5 (4/29 – 5/3)	Chapter 7	Chpt 7 (Outro): Adding an Explanatory Variable	Chpt 8 (Intro): Quantitative Explanatory Variables	Lab Activity: The Paired T-test
6 (5/6 – 5/10)	Chapter 8	Chpt 8 (Outro): Quantitative Explanatory Variables	Chpt 9 (Intro): Distributions of Estimates	Midterm (1 – 8)

Figure 14:

Checking out R in DataCamp

Before Next Time

- Try out Canvas Text (Chapter 1 and 2)
- Familiarize yourself with R in DataCamp

Thursday (4/4/2019)

Today's Learning Outcomes

- Describe how data can be represented in a dataframe
- Create numeric summaries of a single variable
- Create visual summaries of a single variable
- Identify when to use visual vs. numeric summaries.

Representing Data

Data can be represented in many ways, but in this class we will prefer the "tidy" approach to data.

This means three things:

- Cases are represented in rows of a dataframe
- Variables are represented in columns of a dataframe
- Data from different types of cases (e.g., people vs. companies) are stored in separate dataframes.

Representing Data

Let's look at out NLS data, remember this data comes from 4224 people in the National Longitudinal Study.

It might be too much to look at the whole dataset so let's check out just the first 6 cases and the first 5 variables.

```
head(NLSdata)[,1:5]
```

```
##   ID HrsSleep2009    Sex BdayMonth YearBorn
## 1  1             7 Female         9     1981
## 2  2             8   Male         7     1982
## 3  4             6 Female         2     1981
## 4 13             5   Male        11     1984
## 5 23             6 Female         1     1983
## 6 27             5   Male         5     1981
```

Examining the structure of data

We can use R to tell us information about what type of data the we're seeing, or what is the structure of the data.

```
str(NLSdata[,1:9])
```

```
## 'data.frame':    4224 obs. of  9 variables:
##  $ ID           : int  1 2 4 13 23 27 28 31 32 33 ...
##  $ HrsSleep2009: int  7 8 6 5 6 5 7 6 7 7 ...
##  $ Sex          : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 2 1 2
##  $ BdayMonth    : int  9 7 2 11 1 5 11 7 7 12 ...
##  $ YearBorn     : int  1981 1982 1981 1984 1983 1981 1983 1982 1981 1
##  $ Ethnicity    : Factor w/ 4 levels "Black","Hispanic",..: 4 2 2 2 2
##  $ ASVAB        : int  45070 58483 37012 67533 6423 9161 17058 62806
##  $ LifeSat2008 : int  8 9 10 7 5 8 6 8 8 8 ...
##  $ Income2008   : int  43500 55000 65000 20000 55000 3500 32000 77000
```

Examining the structure of data

We can use R to tell us information about what type of data the we're seeing, or what is the structure of the data.

```
str(NLSdata[,10:15])
```

```
## 'data.frame':    4224 obs. of  6 variables:
##  $ HeightFt2010: int  5 5 5 5 4 5 6 5 5 5 ...
##  $ HeightIn2010: int  7 7 1 5 11 6 0 9 8 4 ...
##  $ Weight2010  : int  152 189 160 136 115 165 210 136 150 138 ...
##  $ Computer2010: int  6 6 6 5 6 5 6 5 6 3 ...
##  $ HrsSleep2010: int  6 7 6 5 6 7 6 8 7 7 ...
##  $ Cohab2009   : logi  FALSE FALSE TRUE TRUE FALSE TRUE ...
```

Looking at some variables

Let's choose a couple variables to look at in depth.

Let's check out Hours of Sleep per night in 2009, this is a continuous/quantitative variable.

```
str(NLSdata$HrsSleep2009)
```

```
##  int [1:4224] 7 8 6 5 6 5 7 6 7 7 ...
```

```
summary(NLSdata$HrsSleep2009)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.000   7.000   6.742   8.000  18.000
```

Looking at some variables

Ethnicity is a "factor" according to R. It means that we use numbers to represent categories. But those numbers are arbitrary. It doesn't matter if White = 1 or White = 3. It's just a trick to make it easy on these computers. We call these categorical variables.

```
str(NLSdata$Ethnicity)
```

```
##  Factor w/ 4 levels "Black","Hispanic",..: 4 2 2 2 2 1 1 4 4 4 ...
```

```
summary(NLSdata$Ethnicity)
```

```
##                          Black                       Hispanic
##                            973                            816
## Mixed Race (Non-Hispanic)  Non-Black / Non-Hispanic
##                             40                           2395
```

Looking at some variables

Cohabitation is whether someone is living with their romantic partner. In R it can be either TRUE (cohabitating) or FALSE (not cohabitating). This is a special type of categorical data. When there is only two categories, we call it dichotomous. "Di" means "two" (the original word comes from greek and meant to cut in two).

```
str(NLSdata$Cohab2009)
```

```
##  logi [1:4224] FALSE FALSE TRUE TRUE FALSE TRUE ...
```

```
summary(NLSdata$Cohab2009)
```

```
##    Mode   FALSE    TRUE
## logical    1491    2733
```

Creating Variables

Data doesn't always come to you the way you want it to.

Let's look at how Height is given to us: We have two height variables `HeightFt2010` and `HeightIn2010`

```
table(NLSdata$HeightFt2010)
```

```
##
##    3    4    5    6    7
##    1   44 3270  904    5
```

```
table(NLSdata$HeightIn2010)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12
## 428 294 394 367 386 306 370 377 293 340 354 314   1
```

Creating New Variables (2)

If we want to think about Height, we may want to create *one variable* which includes information about both feet and inches.

There are two ways we could do this:

1. Create a new variable which is total number of inches

```
NLSdata$HeightTotalIn <- NLSdata$HeightFt2010*12 +
                         NLSdata$HeightIn2010
```

2. Create a new variable which is total number of feet

```
NLSdata$HeightTotalFt <- NLSdata$HeightFt2010 +
                         NLSdata$HeightIn2010/12
```

Creating New Variables

Let's take a look at our new variables to make sure everything makes sense.

We can look at one person's data, to see if their numbers added up correctly:

```
NLSdata[1,]
```

```
##   ID HrsSleep2009    Sex BdayMonth YearBorn                Ethnicity
## 1  1             7 Female          9     1981 Non-Black / Non-Hispanic
##   LifeSat2008 Income2008 HeightFt2010 HeightIn2010 Weight2010 Comput
## 1           8      43500             5            7        152
##   HrsSleep2010 Cohab2009 HeightTotalIn HeightTotalFt
## 1            6     FALSE            67      5.583333
```

There's a lot of extra information here, so let's use a new function to simplify.

```
select(NLSdata[1,], c(HeightFt2010, HeightIn2010,
                      HeightTotalIn, HeightTotalFt))
```

Creating New Variables (4)

```
select(NLSdata[1,], c(HeightFt2010, HeightIn2010,
                      HeightTotalIn, HeightTotalFt))
```

```
##   HeightFt2010 HeightIn2010 HeightTotalIn HeightTotalFt
## 1            5            7            67      5.583333
```

Let's make sure this makes sense to us. This person is 5 ft 7 in tall:

```
5*12 + 7
```

```
## [1] 67
```

```
5+7/12
```

```
## [1] 5.583333
```

Creating New Variables (5)

We probably don't want to check everyone but let's pick someone else randomly.

The `sample()` function will generate a random number from the range we specify:

```
randomPerson <- sample(1:4224, 1)
randomPerson
```

```
## [1] 3274
```

```
select(NLSdata[randomPerson,], c(HeightFt2010, HeightIn2010,
                                 HeightTotalIn, HeightTotalFt))
```

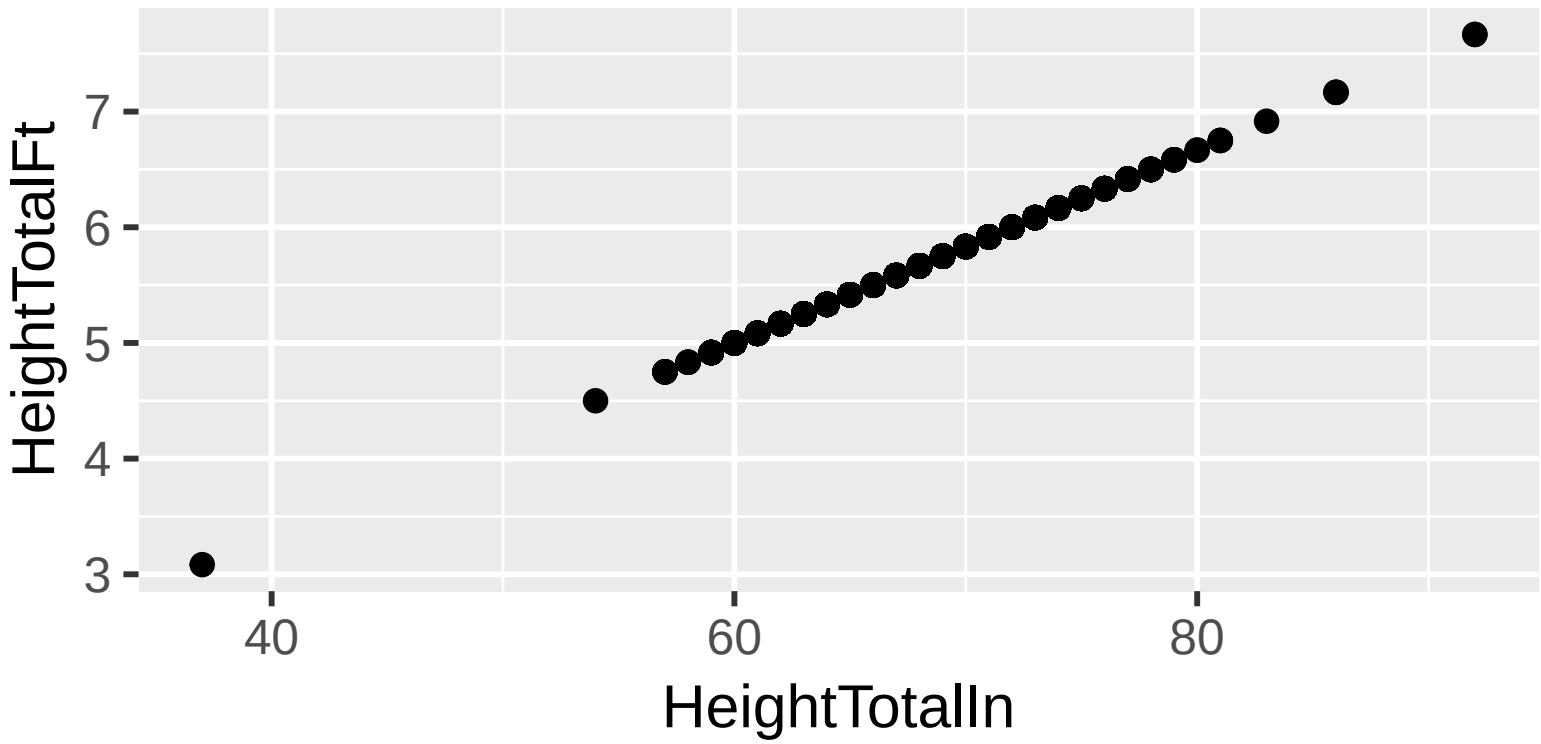
```
##      HeightFt2010 HeightIn2010 HeightTotalIn HeightTotalFt
## 3274            5            8            68      5.666667
```

Creating New Variables (6)

Looks like things are going well. But let's check one more thing. The variables `HeightTotalIn` and `HeightTotalFt` should be perfectly related to each other. There is just an equation to move from one to the other.

Let's plot them both to make sure that is true.

```
gf_point(HeightTotalFt~HeightTotalIn, data = NLSdata)
```



Summary Statistics: Numeric Summaries

When we used the `summary()` function in `R` it gave us six numbers:

- Minimum: lowest value observed
- Maximum: highest value observed
- Mean: Arithmetic average (Add all observations up, divide by number of observations)
- Median: Middle of the data (50% of data is below this point and 50% of the data is above this point)
- 1st Quartile: Lowish point in the data (25% of the data is below this point and 75% is above)
- 3rd Quartile: Highish point in the data (75% of the data is below this point and 25% is above)

```
summary(NLSdata$HeightTotalFt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.083   5.333   5.667   5.650   5.917   7.667
```

Summary Statistics: Numeric Summaries

- 1st Quartile: Lowish point in the data (25% of the data is below this point and 75% is above)
- 3rd Quartile: Highish point in the data (75% of the data is below this point and 25% is above)

Where do you think we can find the mean and the median of the distribution?

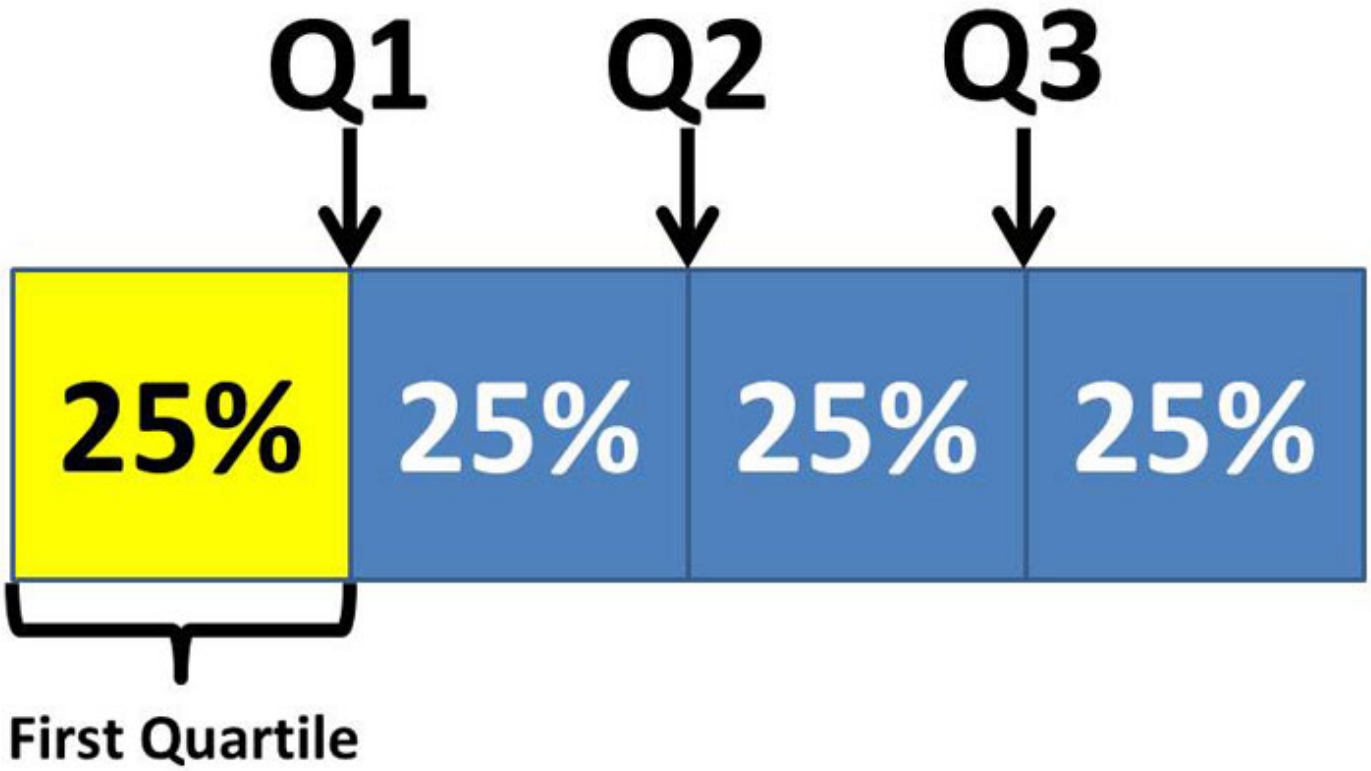


Figure 15: Quartiles

Summary Statistics: Numeric Summaries

Numeric summaries can give us some very specific information about the distribution. But often times it's easier to make a visualization.

Visualizations help us see the **big picture**.

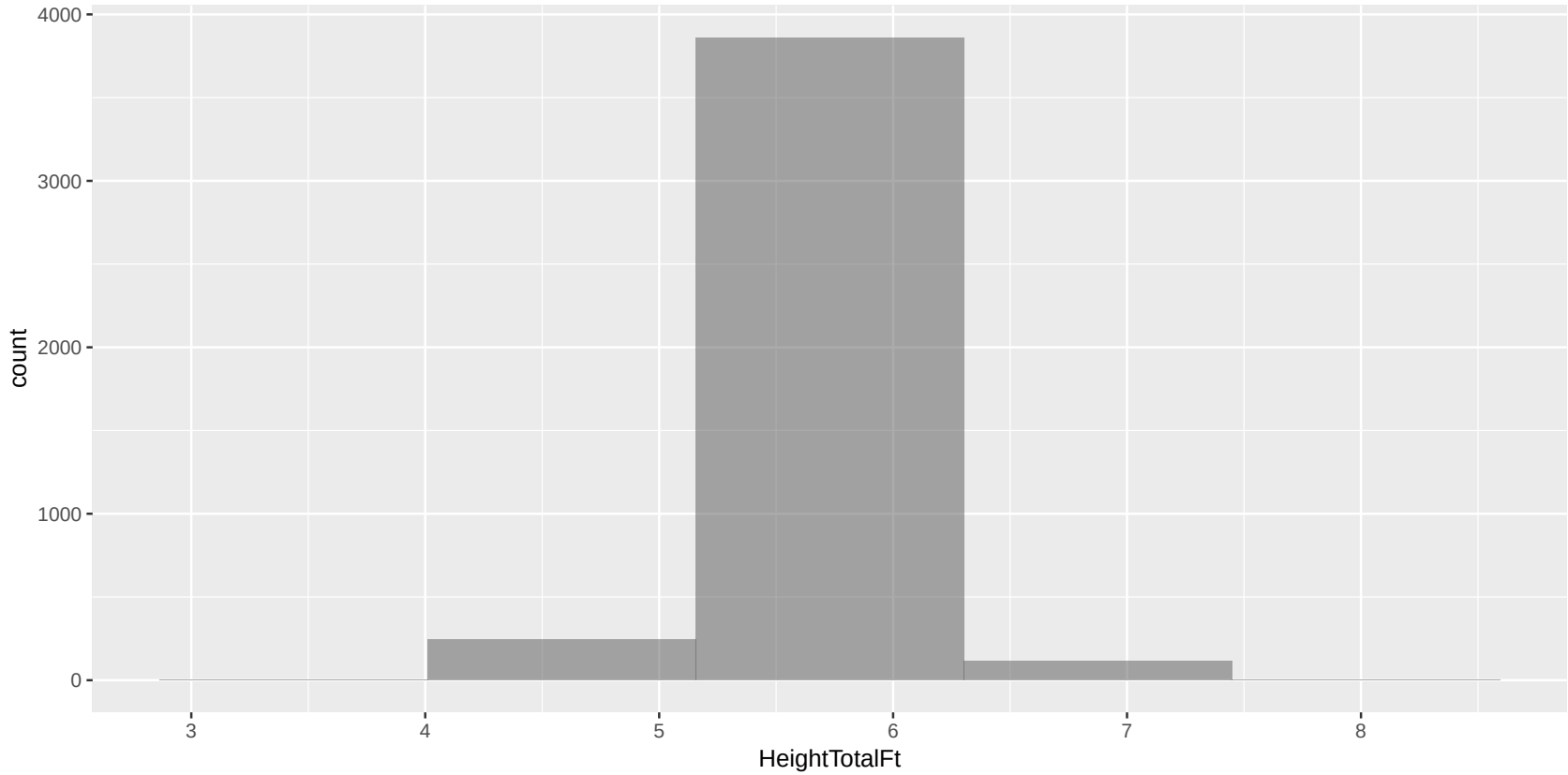
Numeric summaries are good for looking at the **finer detail**.

Most of the time we want both the big picture and finer detail. So we use both approaches.

Histograms

I've shown you a few histograms, but let's make sure we know how the R function works, and what the function is doing.

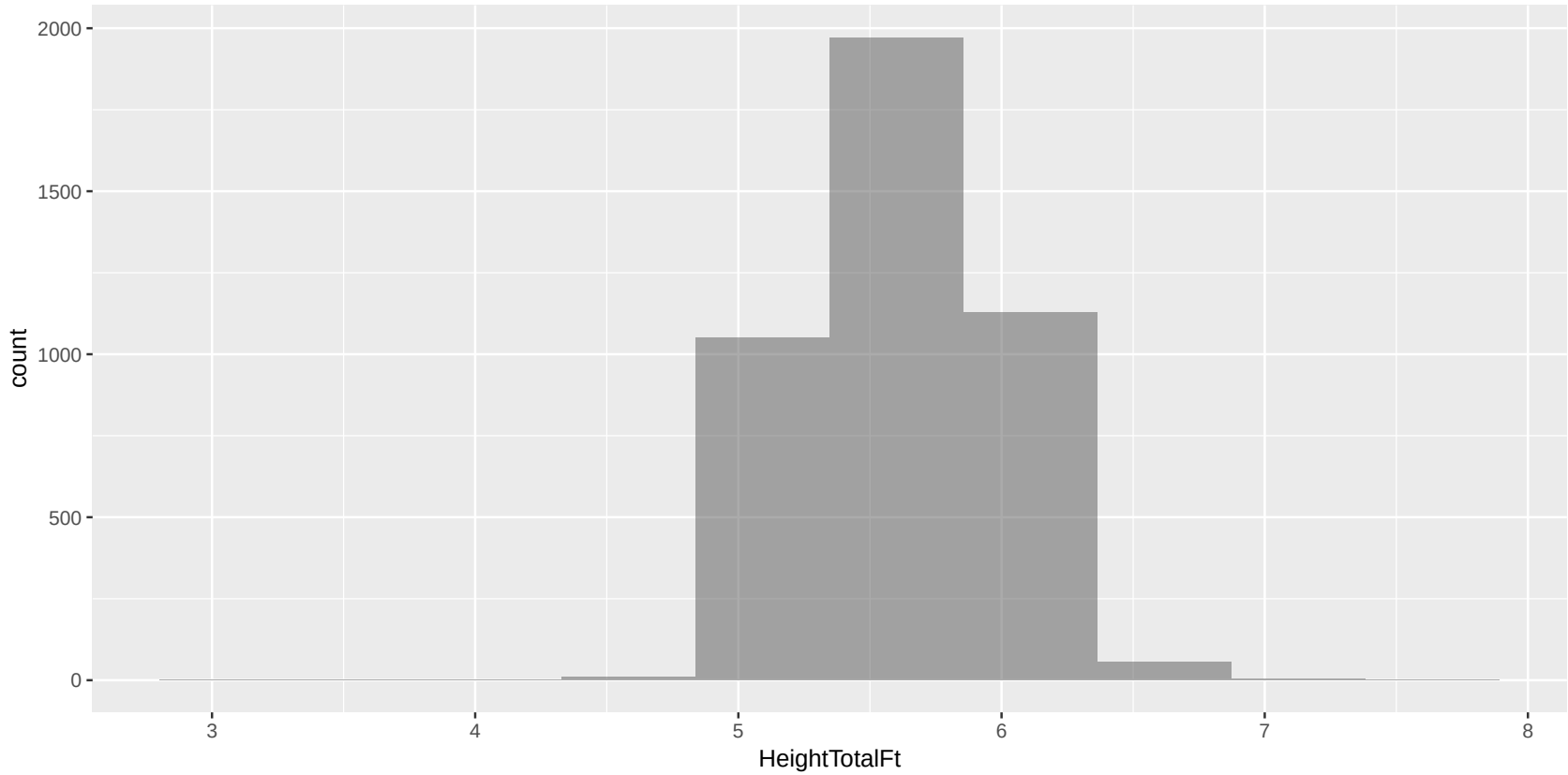
```
gf_histogram(~HeightTotalFt, data = NLSdata, bins = 5)
```



Histograms

I've shown you a few histograms, but let's make sure we know how the R function works, and what the function is doing.

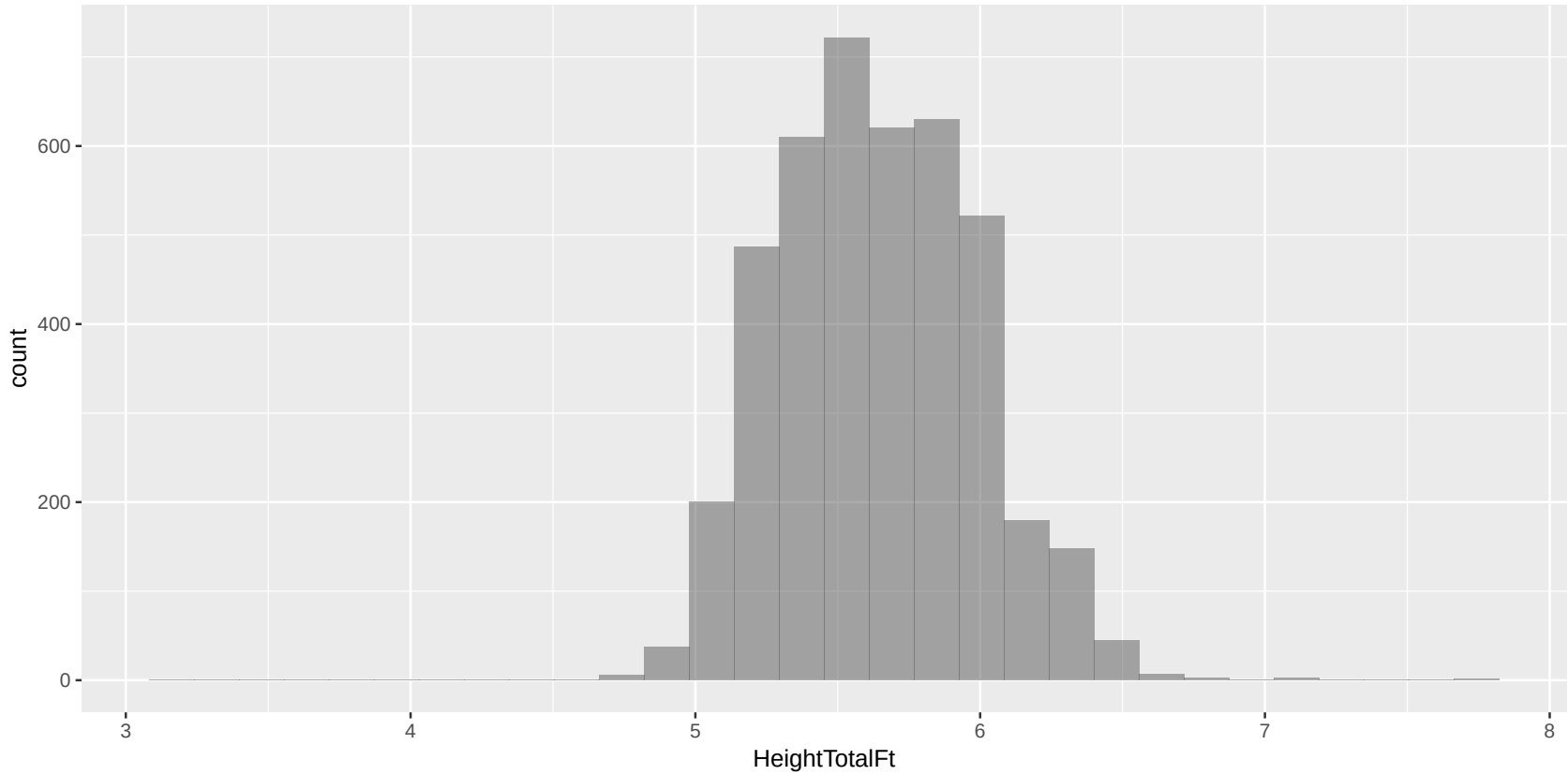
```
gf_histogram(~HeightTotalFt, data = NLSdata, bins = 10)
```



Histograms

I've shown you a few histograms, but let's make sure we know how the R function works, and what the function is doing.

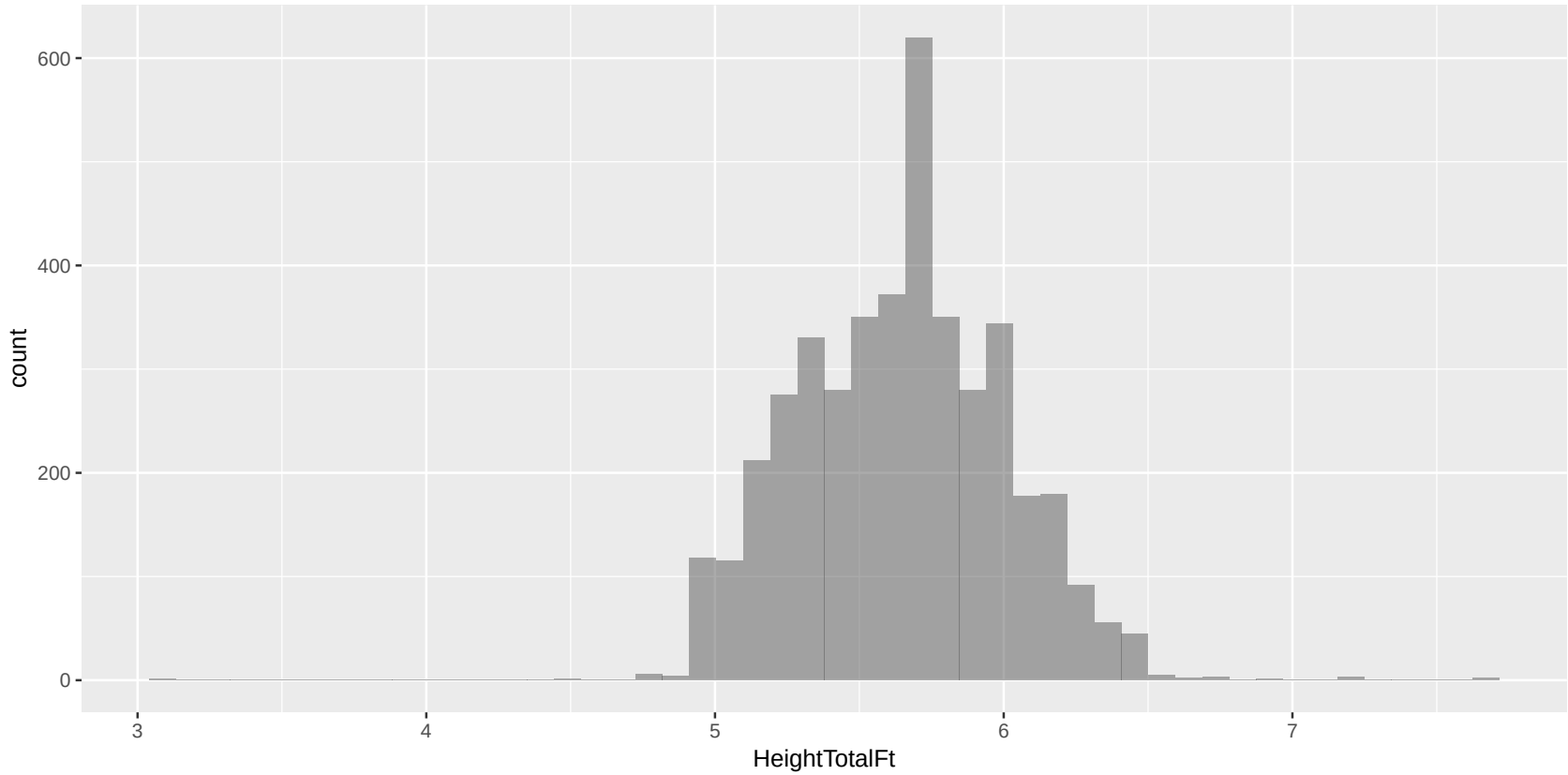
```
gf_histogram(~HeightTotalFt, data = NLSdata, bins = 30)
```



Histograms

I've shown you a few histograms, but let's make sure we know how the R function works, and what the function is doing.

```
gf_histogram(~HeightTotalFt, data = NLSdata, bins = 50)
```



Histograms

- Histograms are used for Continuous/Quantitative variables
- Plot the number of cases which fall into different bins
- Bins are always equal in size
- Bins are made by chopping up the range of the variable

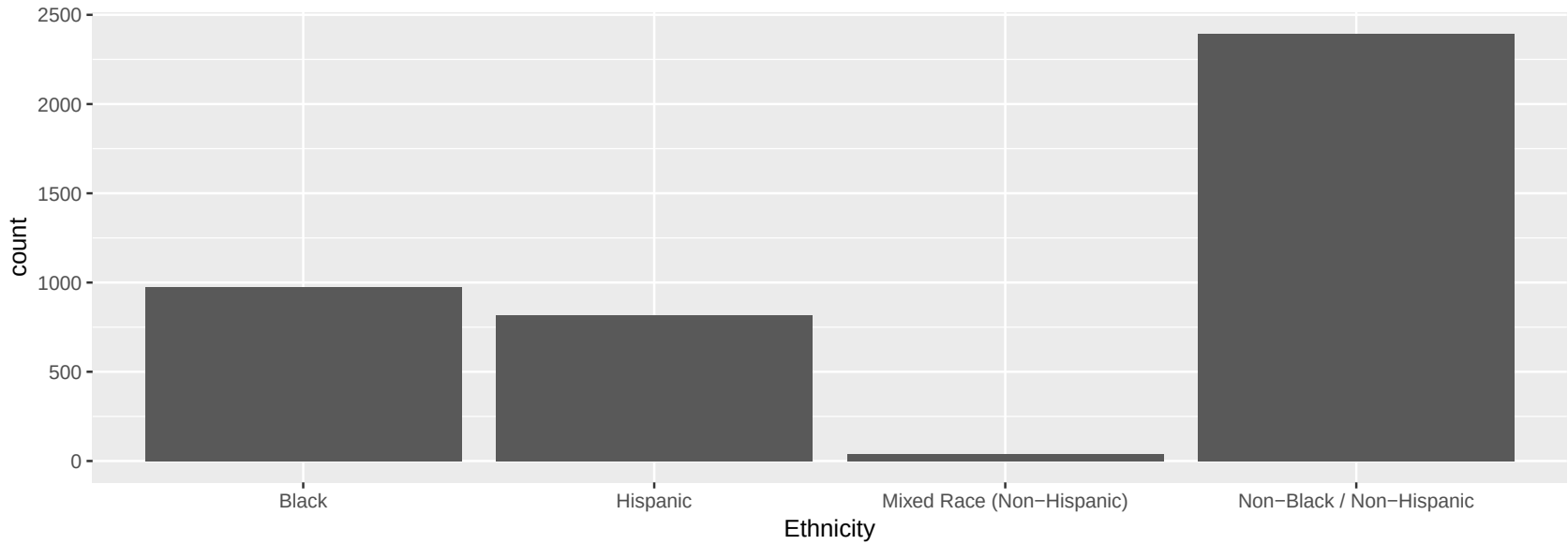
Frequency Plots

Frequency plots/bar plots are like histograms in the we count the number of cases in a "bin"

But frequency plots are used for categorical variables, so the bins come premade. We can't divide up the bins in different ways, and the order of the bins is arbitrary.

Let's examine a frequency plot for Ethnicity. Notice that we don't use the histogram function.

```
gf_bar(~Ethnicity, data = NLSdata)
```

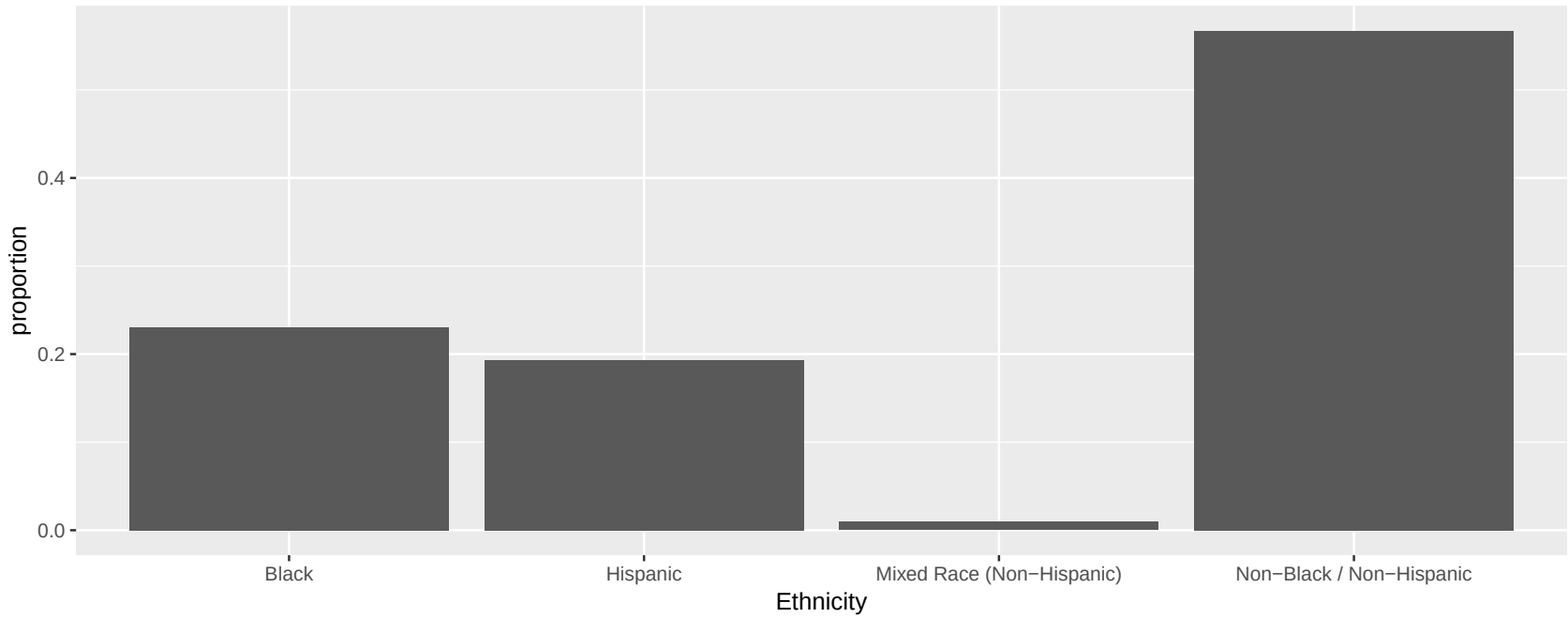


Relative Density Plot

When we made a histogram or a barplot the Y-axis tells us how many cases are in each bin. But sometimes it's easier to think in proportions. So we can use densities (proportions). Densities for bar plots take the number of cases and divide by the total.

For example we know that there were 973 Black respondents out of 4224. That means the proportion Black respondents will be 973/4224 =0.2303504.

```
gf_props(~Ethnicity, data = NLSdata)
```



Relative Density Plot

Density/relative frequency can also be used for continuous variables. The way these plots work is the the entire "area" of the bars add up to 1. Below each bar represents an inch (or 1/12 of a foot). If we want to know what proportion of the sample is 5'6" we can take the height of the bar (1) times the width of the interval (1/12). So `1*1/12 = 1/12`. Can we be sure this is true? Let's check!

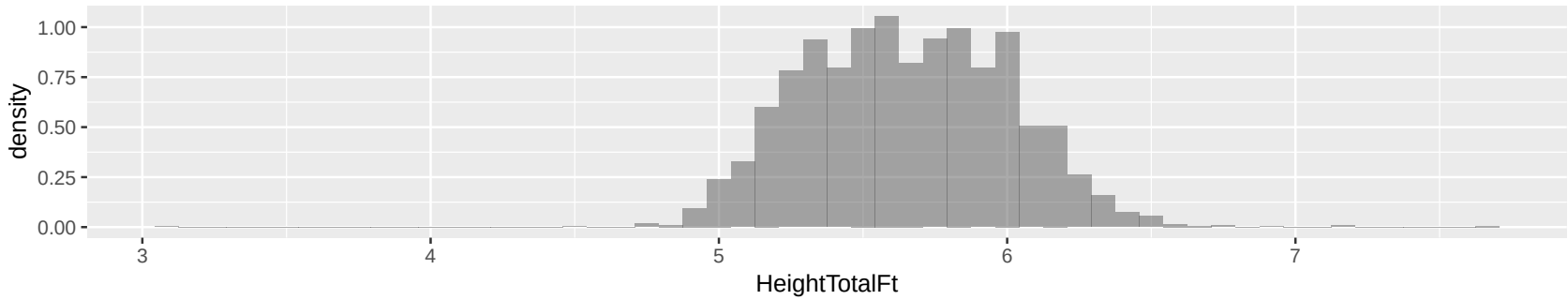
```
sum(NLSdata$HeightTotalFt == 5.5)/4224
```

```
## [1] 0.08285985
```

```
1/12
```

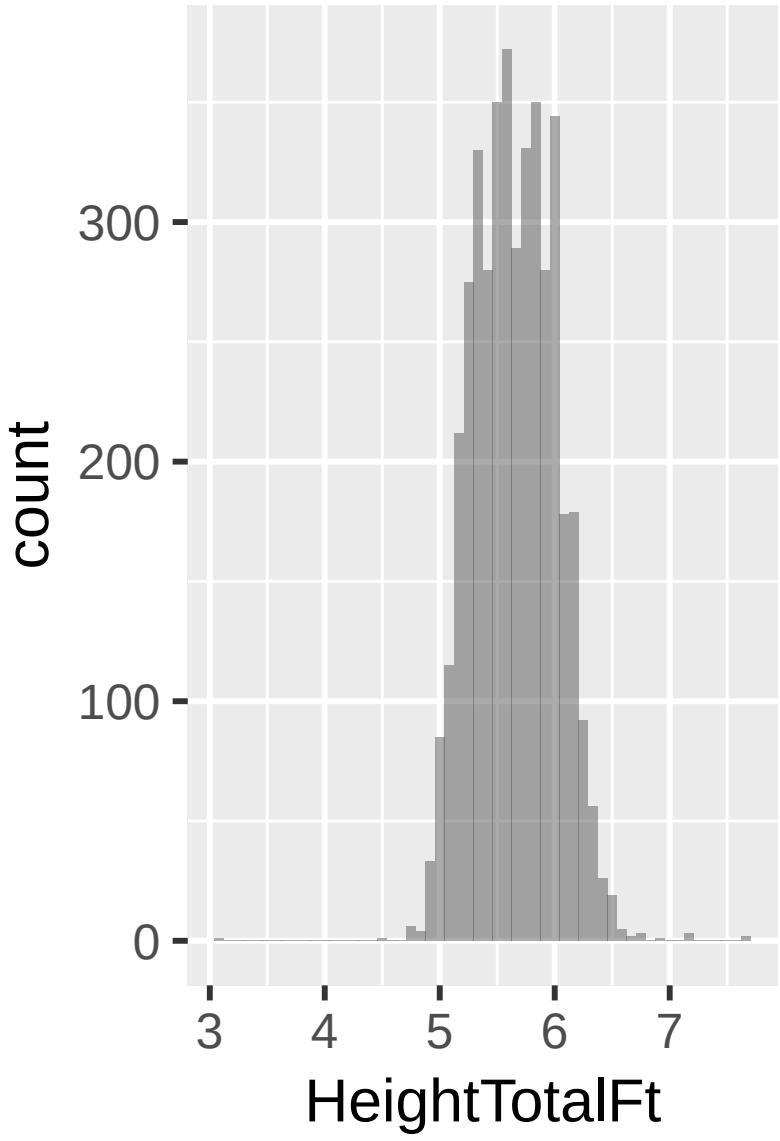
```
## [1] 0.08333333
```

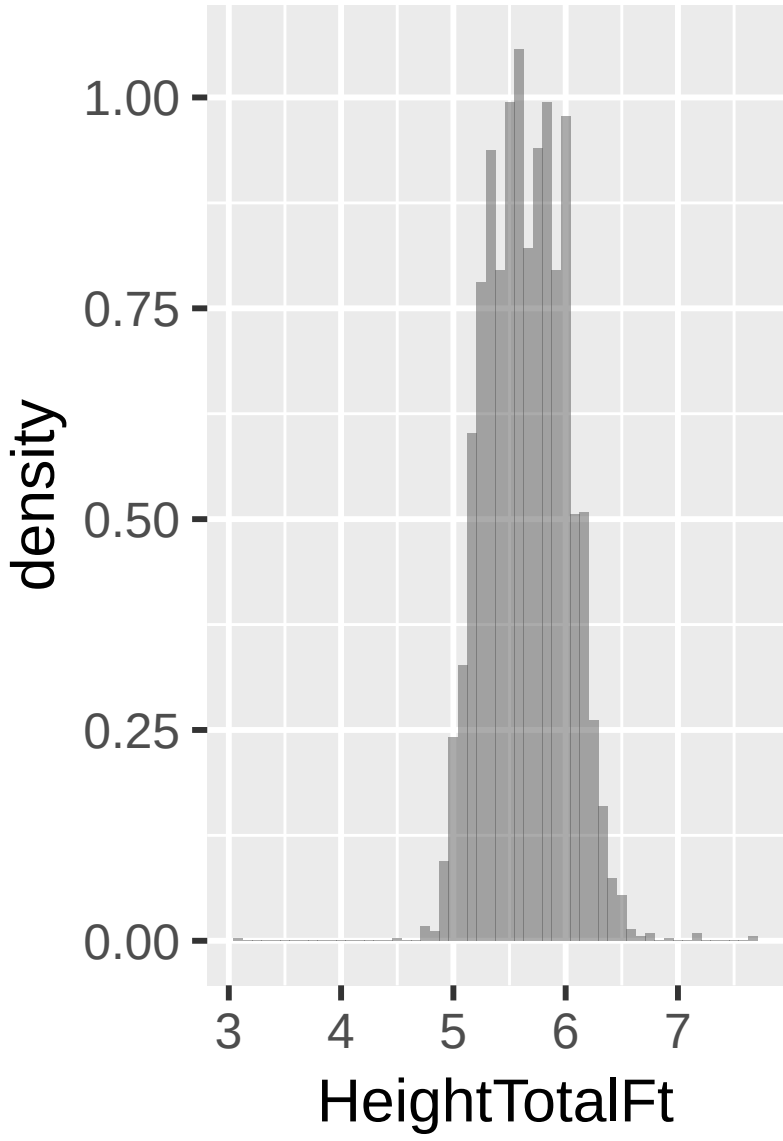
```
gf_dhistogram(~HeightTotalFt, data = NLSdata, binwidth=1/12)
```



Relative Density Plot

It's important to notice the the **shape** of the distribution doesn't change whether you use a density plot or a histogram. What is changing though?





Describing Shape: How many peaks?

We can describe a distribion based on how many peaks/modes it has.

Unimodal: One peak

Bimodal: Two peaks

Uniform: No peaks (all data is equally probable)

Can you come up with a word that might mean there are more than two peaks?

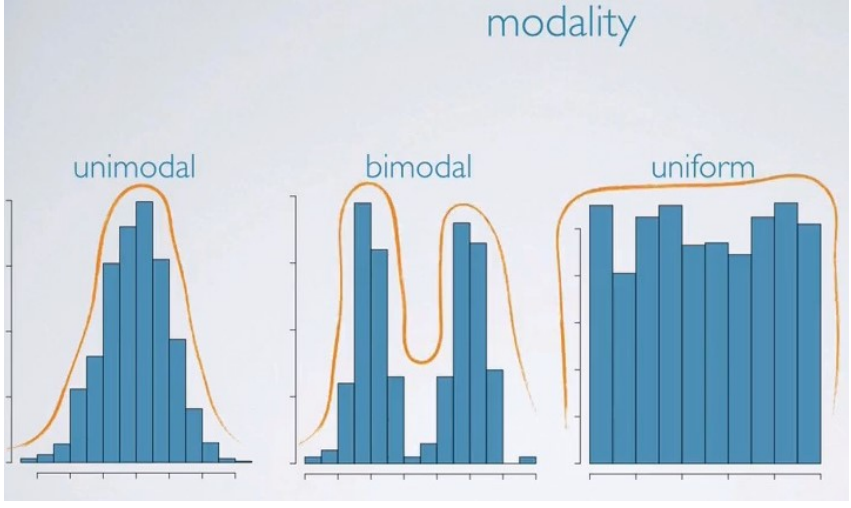


Figure 16: How many peaks/modes

Describing Shape: Skew

Skew of a distribution means that the left side and right side don't look similar.

The opposite of skewed is "symmetrical".

Skew can go in two directions "right" and "left". This corresponds to the direction of the "tail" (the part of the distribution which stretches out)

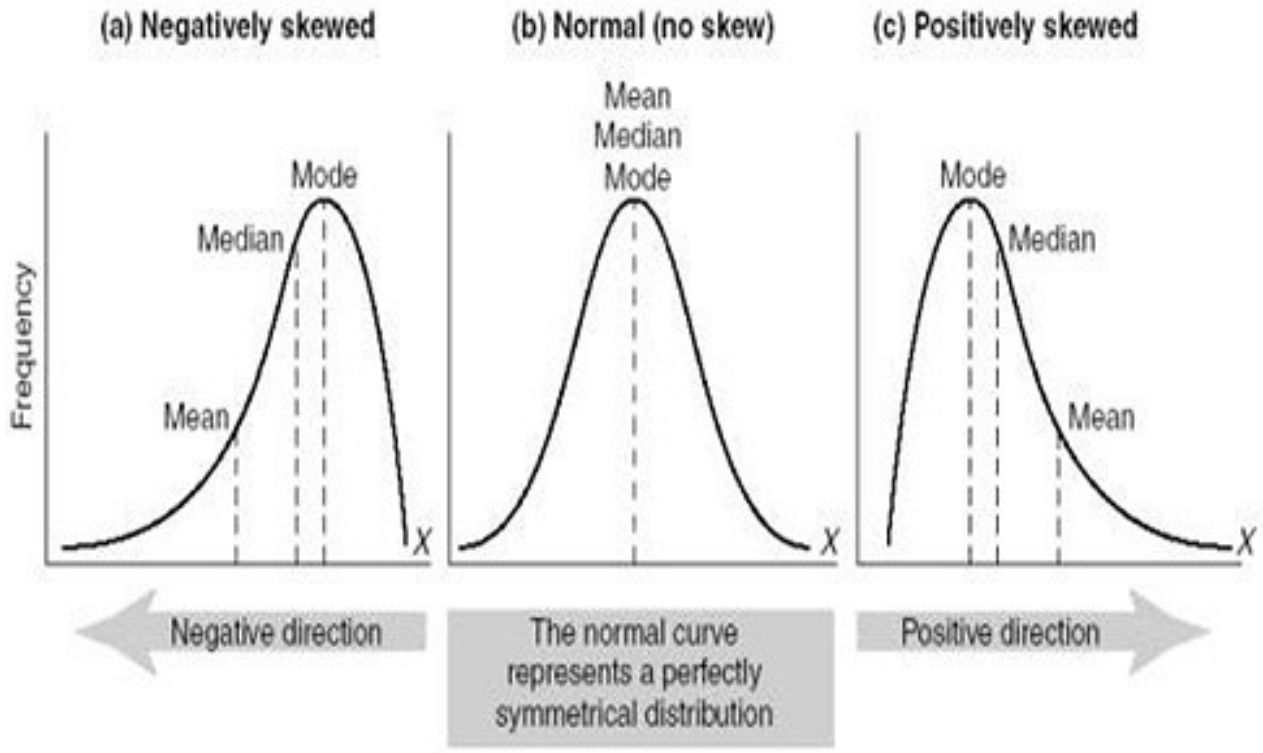


Figure 17: Distributions with Different Skew

Normal Distribution

The normal distribution can go by many names: bell-shaped, Gaussian distribution

It's very popular in statistics because it turns out that many things have a normal distribution.

In particular when we take averages of things (like heights, hours of work time, salary, etc) the averages turn out to have a normal distribution. This was proven by some very smart statisticians, and it's called the *Central Limit Theorem*. (We'll talk about this more later)

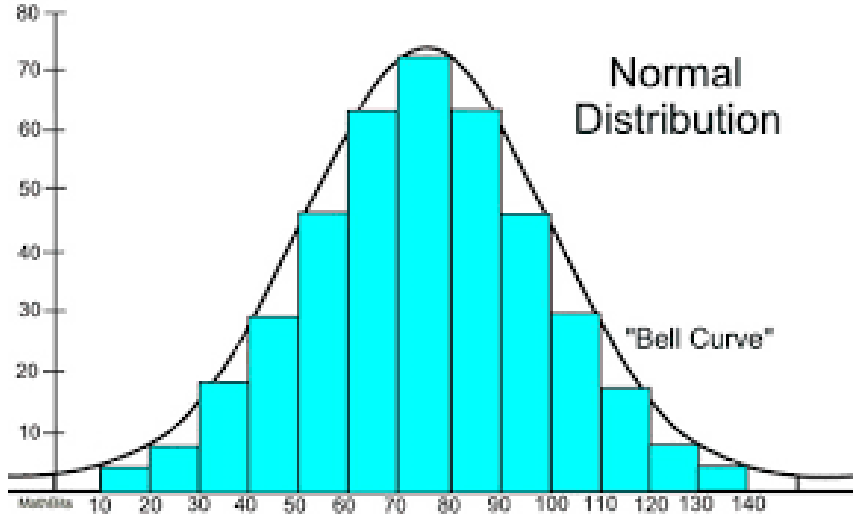


Figure 18: Normal Distribution

Center and Spread

Two characteristics that can tell us a lot about a distribution are:

Center: Where are more of the cases? Where is the middle of the distribution?

Spread: How spread out are the cases? How similar or dissimilar should we expect two cases to be?

We've already talked about a couple measures of both of these.

Center: Mean and Median, we can also add Mode (most frequent outcome)

Spread: Quartiles.

Measuring Spread

We learned that quartiles are useful for describing the data, but how do we use them to measure spread?

The **interquartile range** is a useful statistics to describe spread.

We take the 1st Quartile and subtract it from the 3rd Quartile.

Remember that the 1st Quartile is the point where 25% of the data is below this point and 75% is above. The 3rd Quartile is the point where 75% of the data is below and 25% is above.

How much of the distribution do you think will be between Q1 and Q3 (i.e., inside the interquartile range?)

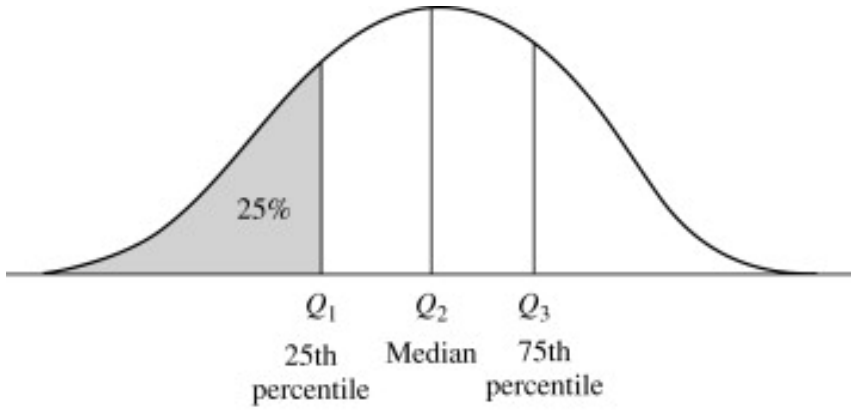


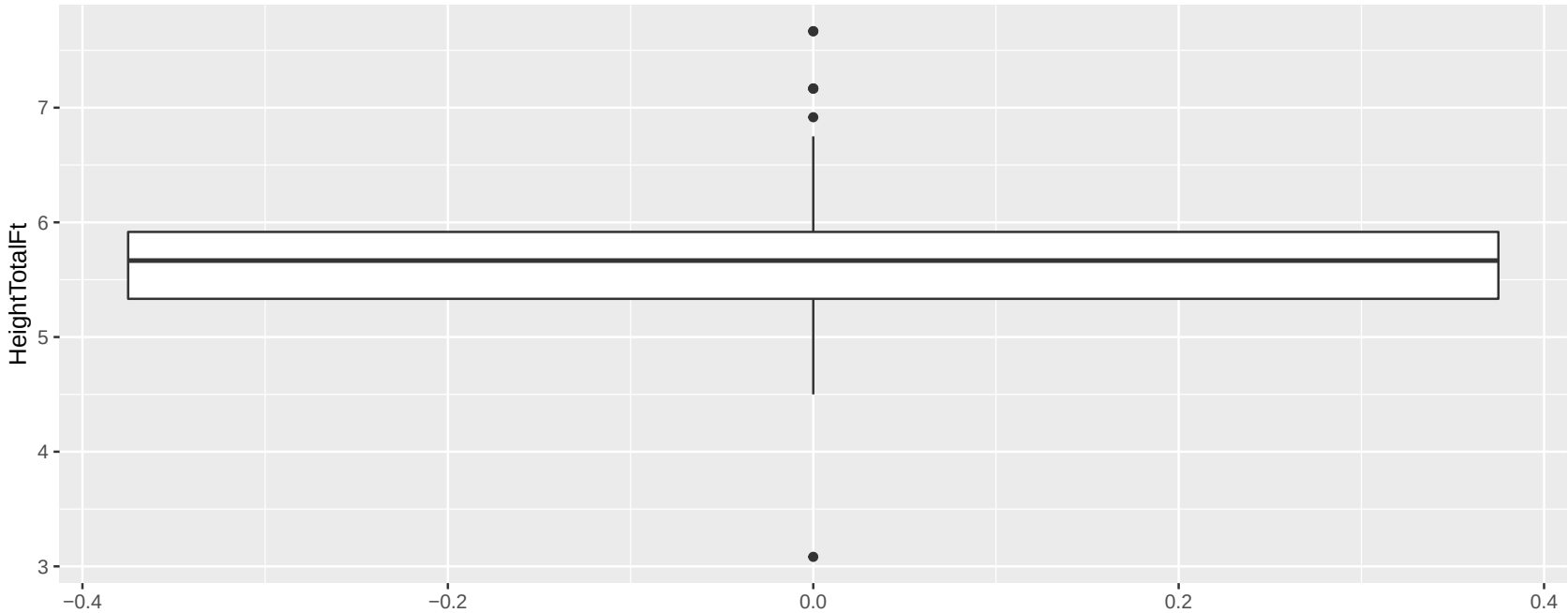
Figure 19: Interquartile Range

Box and Whisker Plots

Box and Whisker plots are a way to take the five number summary (min, Q1, Median, Q3, and max) and create a visualization.

These plots are particularly ideal for spotting unusual cases (i.e., outliers)

```
gf_boxplot(~HeightTotalFt, data = NLSdata)
```




```
summary(NLSdata$HeightTotalFt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.083   5.333   5.667   5.650   5.917   7.667
```

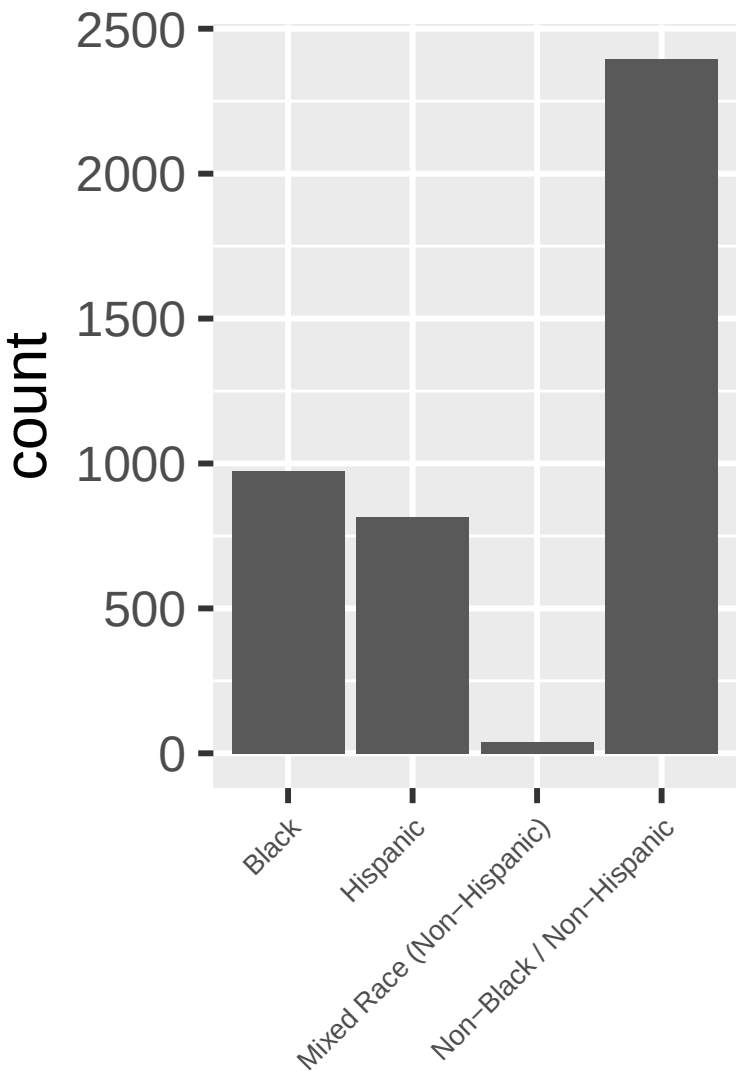
Thinking about Categorical Data

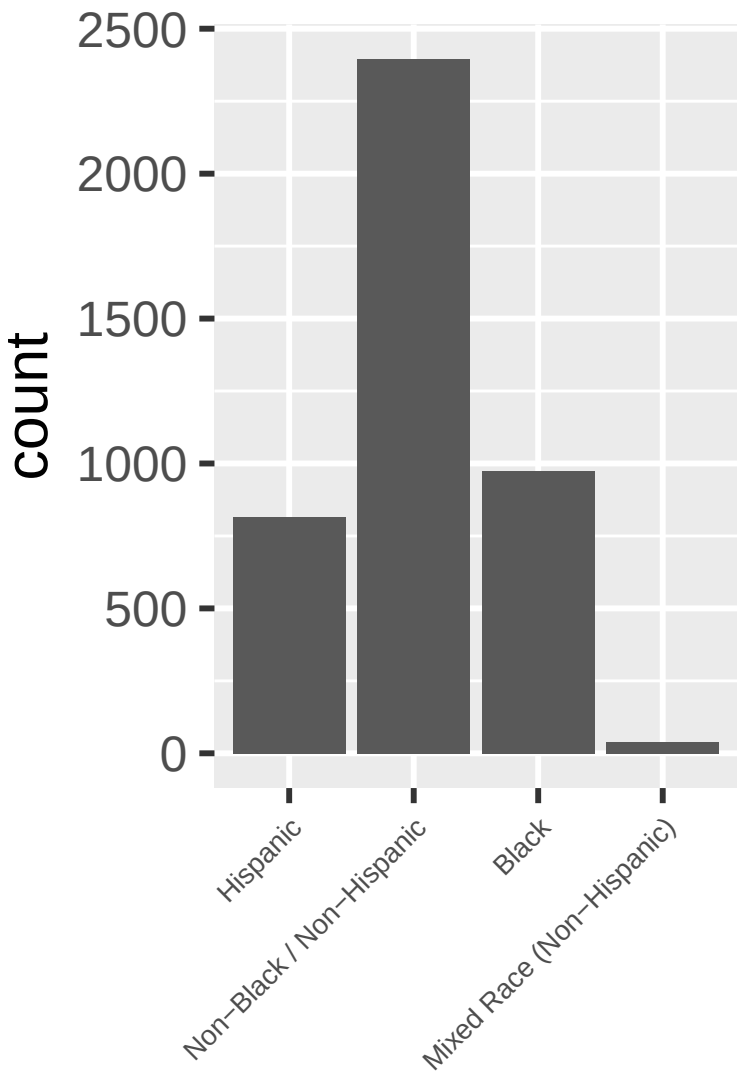
We've just learned how to describe the shape of a distribution of a quantitative variable, but what if the variable is categorical/qualitative. Does it make sense to talk about the shape and skew of a distribution of qualitative observations?

Thinking about Categorical Data

Does it make sense to talk about the shape and skew of a distribution of qualitative observations?

The ordering of the categories is arbitrary, so we could reorganize the distribution and it would look totally different!





Big Picture vs. Finer Detail

Let's try to break down some of the ways we've just learned to summarize data.

Imagine you work for an car insurance company as a data-specialist. Below are four requests from your boss and four approaches to representing data. Choose the approach that you think is most appropriate for each request:

1. We think there are generally two groups of people: people who drive a lot and people who drive very little. Can you look at our "hours of driving" data to tell us if that's the case?
2. When do most people get their driver's license, is there a range that's typical for people to get their first drivers license? Can we use the data about what age people were when they got their license?
3. What is the average amount of money we typically spend when someone files a claim? Can you pull the data and give me your best estimate?
4. Laura in the Wisconsin office is curious what proportion of car crashes happen on the freeway as compared to arterials and other roadways. Could you give her some information about where crashes happen?

Options:

1. Interquartile Range
2. Mean
3. Histogram
4. Bar Graph with Densities