Data Analytics
CS301
The *Great* Review

Week 15
Spring 2023
Oliver BONHAM-CARTER

Course Summary

CMPSC-301 Data Analytics (4 Credits)

An introduction to computational methods of data analysis with an emphasis on understanding and reflecting on the social, cultural, and political issues surrounding data and its interrogation. Participating in hands-on activities that often require teamwork, students study, design, and implement analytics software and learn how to extract knowledge from, for instance, financial, political, and scientific sources of data. Students also investigate the biases, discriminatory views, and stereotypes that may be present during the collection and analysis of data, reflecting on the ethical implications of using the resulting computational techniques. During a weekly laboratory session, students use state-of-the-art statistical software to complete projects, reporting on their findings through both written documents and oral presentations. Students are invited to use their own departmentally approved laptop in this course; a limited number of laptops are available for use during class and lab sessions. Prerequisite: FS*102 or permission of the instructor. Distribution Requirements: QR, PD.

Requisites:
FS*102 or permission of the instructor - Must be completed prior to taking this course.
Locations:
Main Campus

Course Objectives

Course Objectives

Students successfully completing this class will have developed:

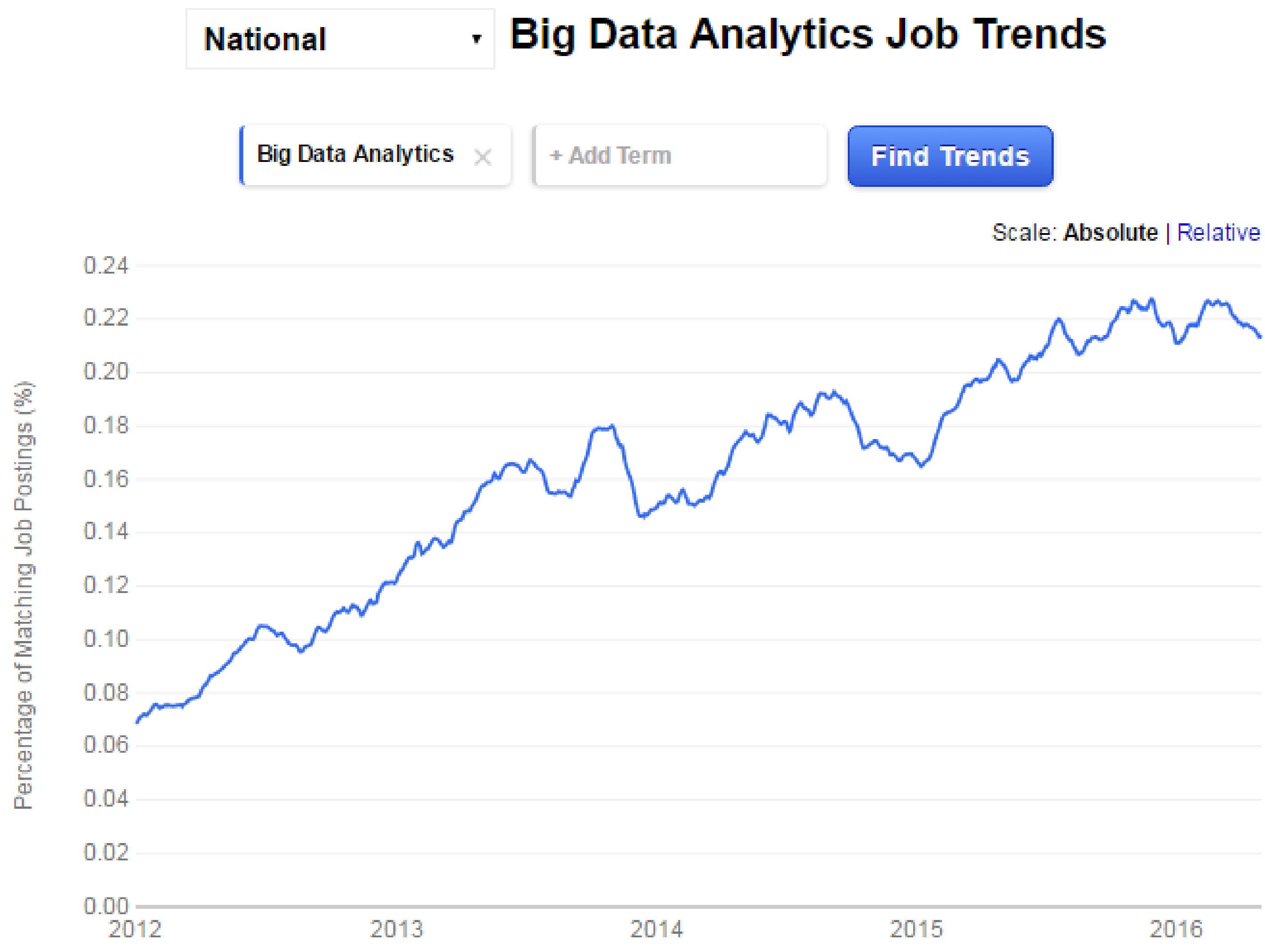
1. A “big-picture” view of data analytics.
2. An understanding of the objectives and limitations of data analytics.
3. An understanding of the main data analytics methods.
4. Practical skills using relevant software tools and programming techniques.
5. An understanding of the contemporary roles of power and difference as they relate to the knowledge derived from a data set.

How Did We Achieve These Objectives?

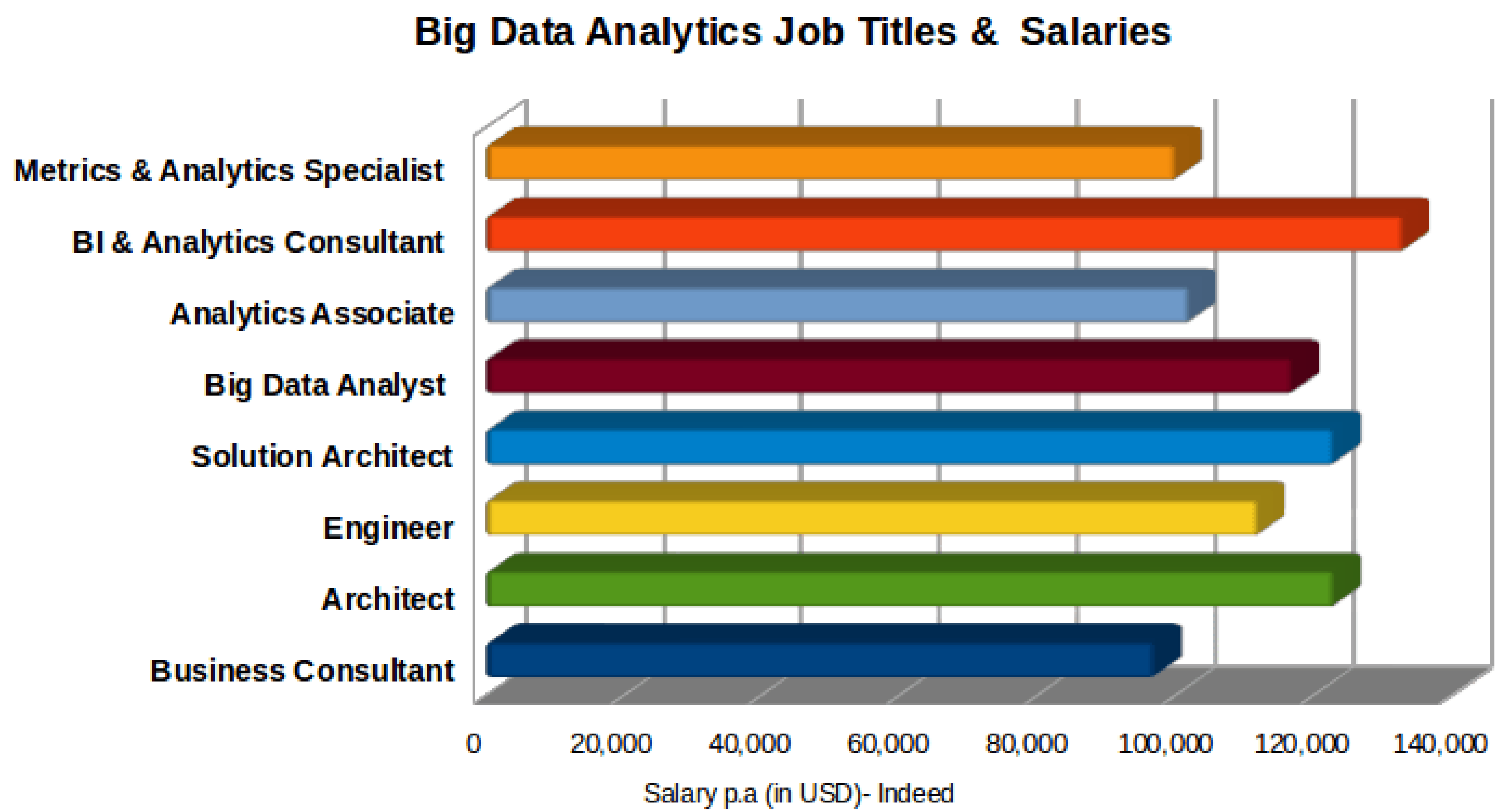
- Class lessons with activities using R to explore data sets
- Labs where students were given opportunity to apply classroom learning to data sets to uncover own researches
- Guest speakers who came to discuss how their own research involves DA
  - Ron Mattocks
  - Lydia Eckstein, PhD
  - Yee Mon Thu, PhD
  - Steven Onyeiwu, PhD
  - Chelsea Peebles, Jesse Sealand (Erie Insurance)

Jobs and Careers

- Many more job posting where analytics is absolutely required



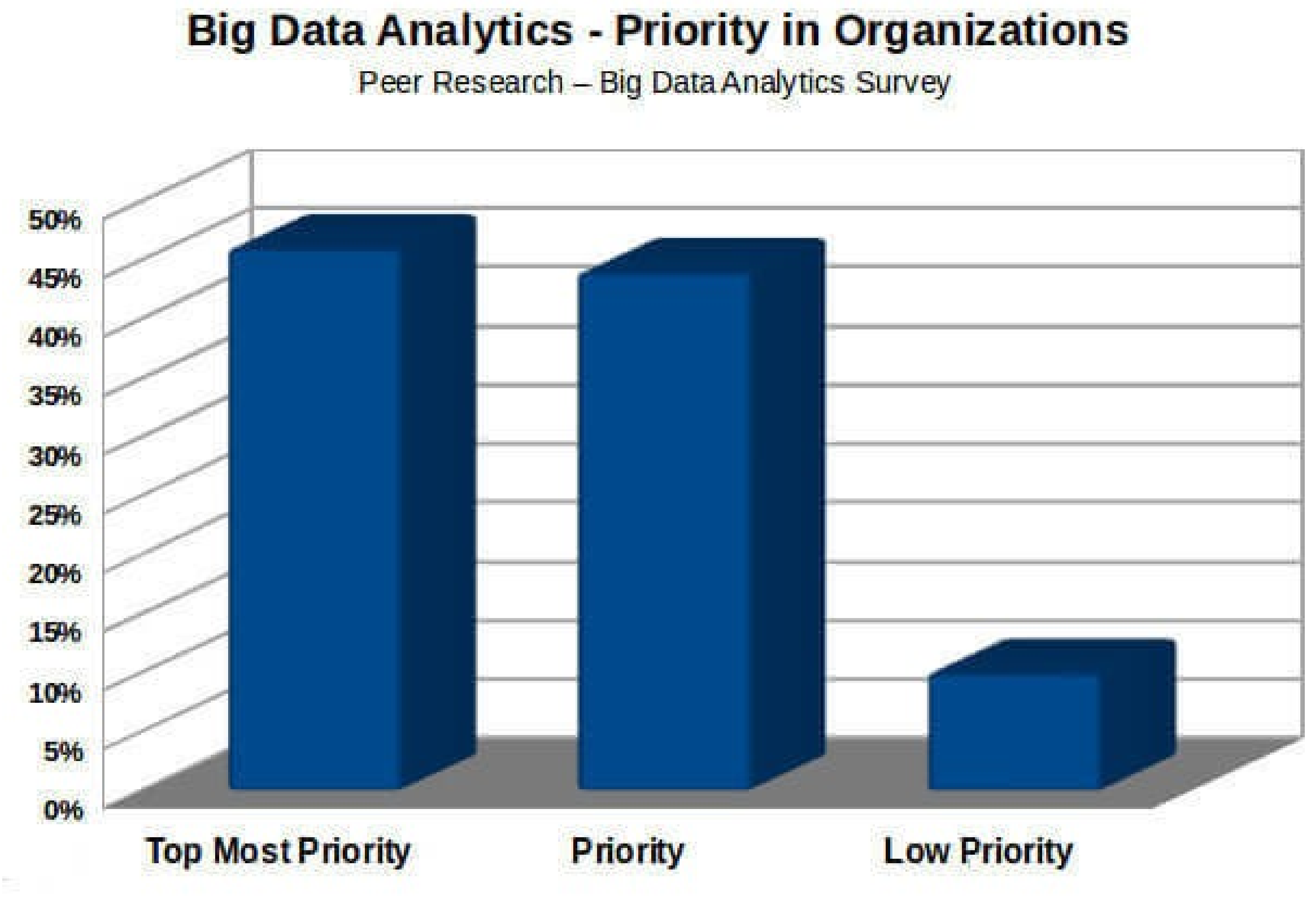
Money to Be Made



- High-paying salaries DA
- These careers have security due to the ever-presence of data in research, industry and everywhere else.

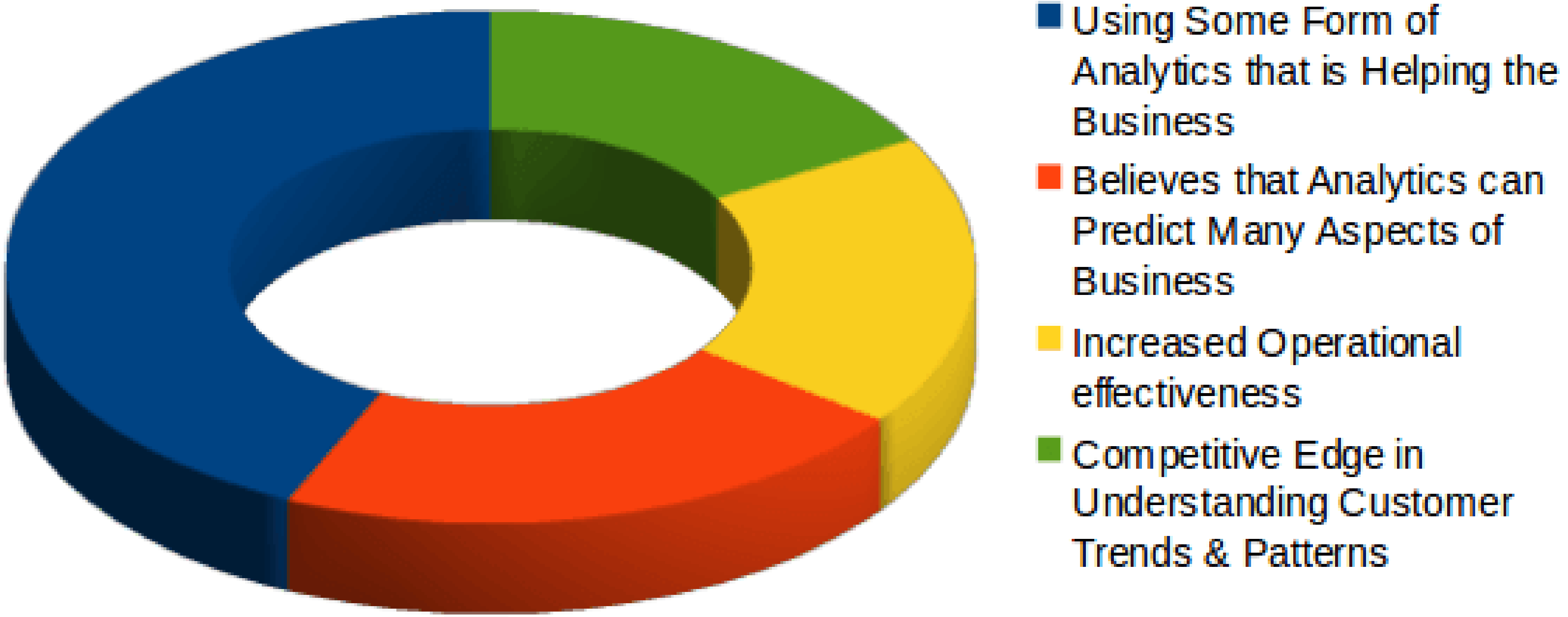
https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move

More Important Each Day



- Organizations are looking for people to help them process, understand what is in their data to make decisions.

Helping Business Development



- Organizations realize that data drives their business.
- When will your career in DA begin?!

Forbes

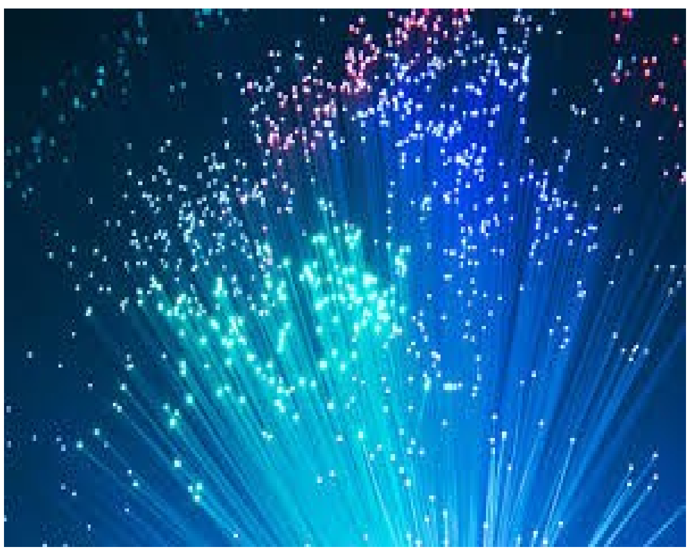
ALLEGHENY COLLEGE

- ***“75% of firms are prioritizing big data and analytics expertise in their hiring decisions****, stating that having these skills is critical for any candidate to be considered an IoT (Internet of Things) expert.”*

https://www.forbes.com/sites/louiscolumbus/2017/08/21/big-data-analytics-is-the-most-wanted-expertise-by-75-of-iot-providers/#52082a4e5188

Computers and Information

- In this class, you learned how to use machines to harness the power of data.

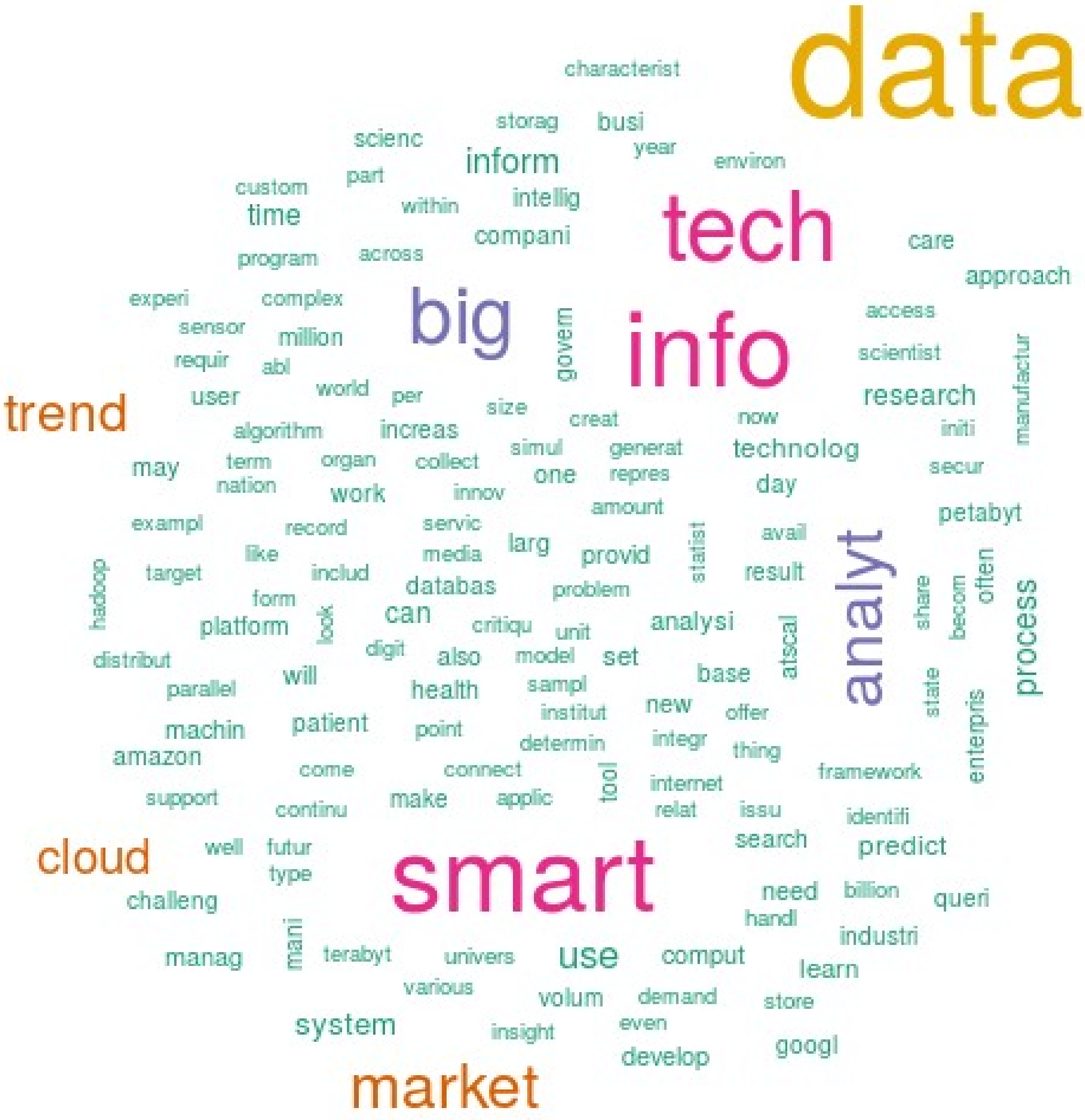


Raw Data

Meaningful Information

How to Analyze Data?

- How do we harness this power of data analytics?



Topics Covered

- **Google Analytics**
  - Web traffic Information: terms and plots
- **Visualizations: types and meanings**
- **R Statistics**
  - Basic syntax and methods
- **Library features**
  - Tidyverse, nycflights13, lubridate
- **Concepts**
  - Exploratory data analysis
  - Tidy data manipulation
  - Managing date and time
  - Others from recent lessons

We Have
Ethical Backbone

- We discussed ethical concepts.
- For example: *Twelve Million Phones, One Dataset, Zero Privacy*, A New York Times opinion piece
- Link: https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html

Opinion | THE PRIVACY PROJECT

Twelve Million Phones, One Dataset, Zero Privacy

By Stuart A. Thompson and Charlie Warzel

DEC. 19, 2019

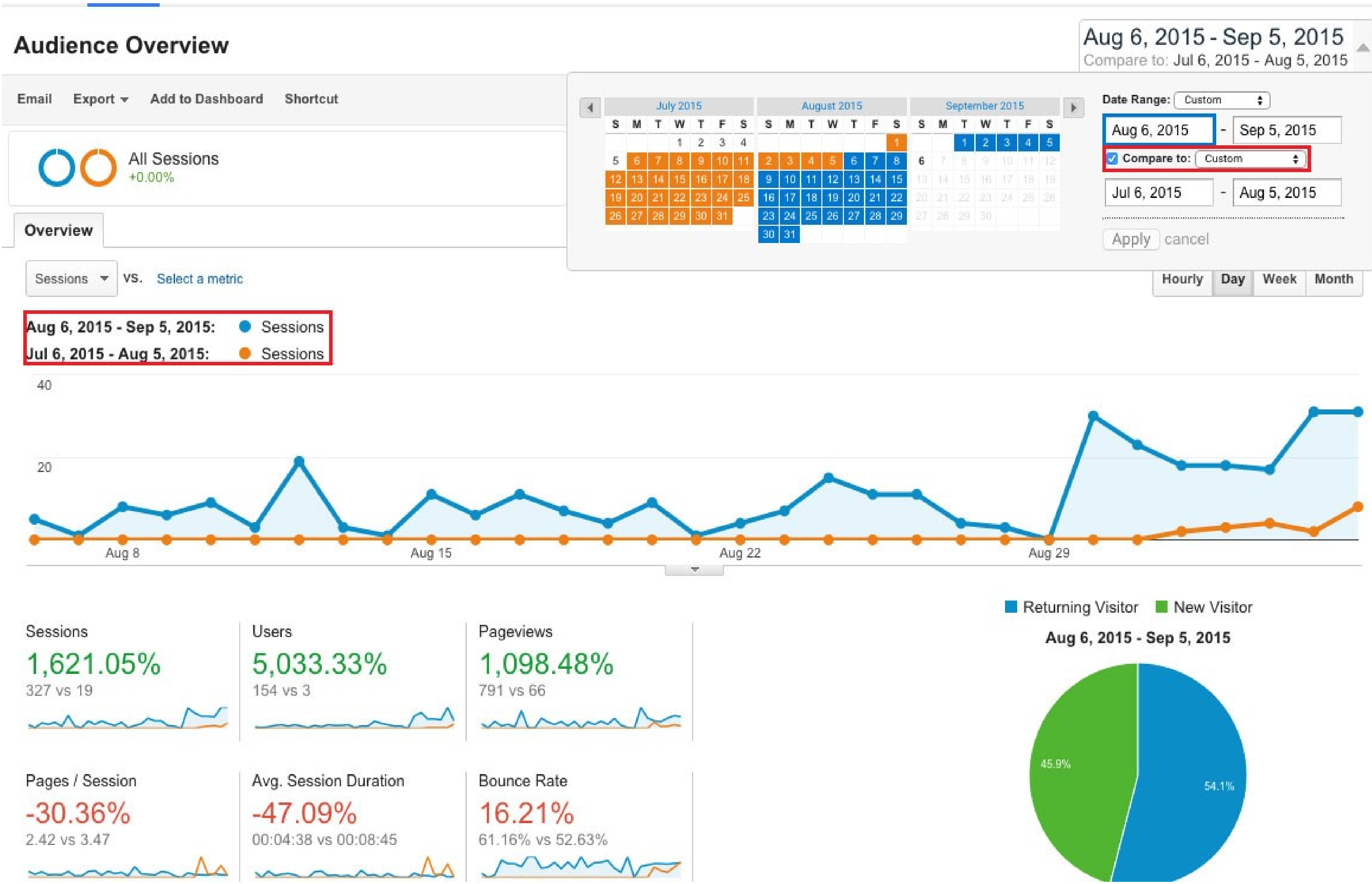
THINK

We Learned Analysis: Google Analytics

- An introduction to computational and analytical methods for finding patterns in large data sets.
- Google Analytics and web page data analysis
  - Page views?
  - How many users clicked on purchase buttons?
  - How many user downloaded (read, viewed) your hand-out newsletter?
  - How long to land on “check-out” page? Time to decide to buy?

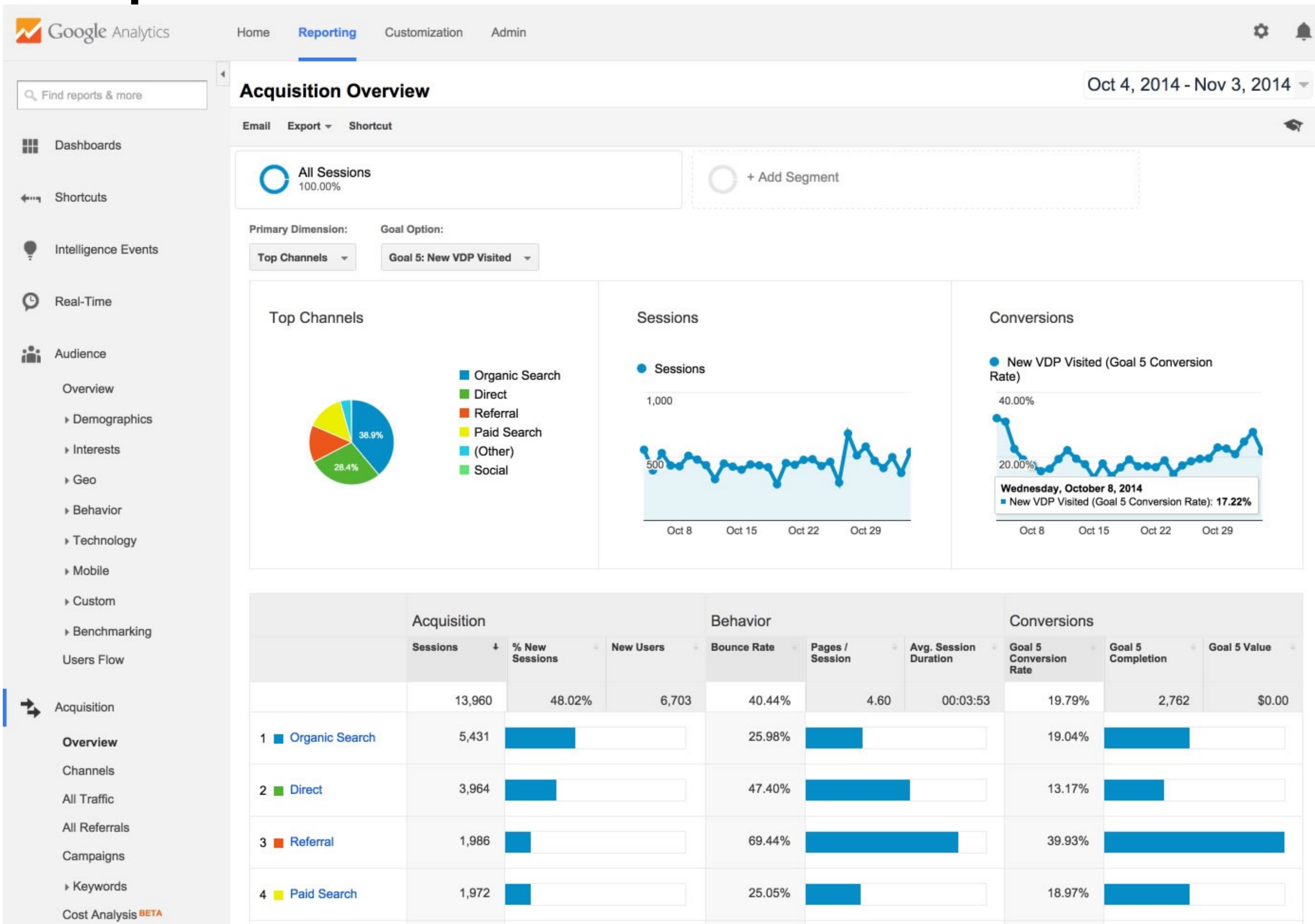
Audience

- Who are your users?
- When was that?



Acquisition

- How do these users get to your site?



Where To Now?

- Google Analytics is a tool allowing for convenient analysis of web sites
- The code was written by developers for this purpose.
- What if you need tools and there are no current developers to create them?

Develop Your Own Tools!!

Questions to Ask?

- Types of general questions for making discoveries
  - What type of variation occurs within my variables?
  - What type of covariation occurs between my variables?
  - What can I do to see more?!

The R Programming Language

- https://www.r-project.org/
- What is the R language?
  - An open source, well-developed programming platform for work in statistics, mathematics and data analytics
  - Built-in libraries to simplify programming
  - Language includes conditionals, loops, user-defined recursive functions and input and output facilities.
- Community Blogs:
  - https://www.r-bloggers.com/
  - https://twitter.com/rstudiotips

Code for a Simple GGPlot

- **library(tidyverse) or if not present,**
- **install.packages(“tidyverse”)**
- **ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy))**

- Establish the *canvas* (where the plot is shown)
- **Ggplot()**

- Link to the data (set is called, ‘mpg’)
  - **ggplot(data = mpg)**

- Compute the geometery of point placement on canvas
  - **geom_point(mapping = …)**

- Compute the aesthetics of the plot (titles, color, point type, etc)
  - **aes(x = displ, y = hwy)**

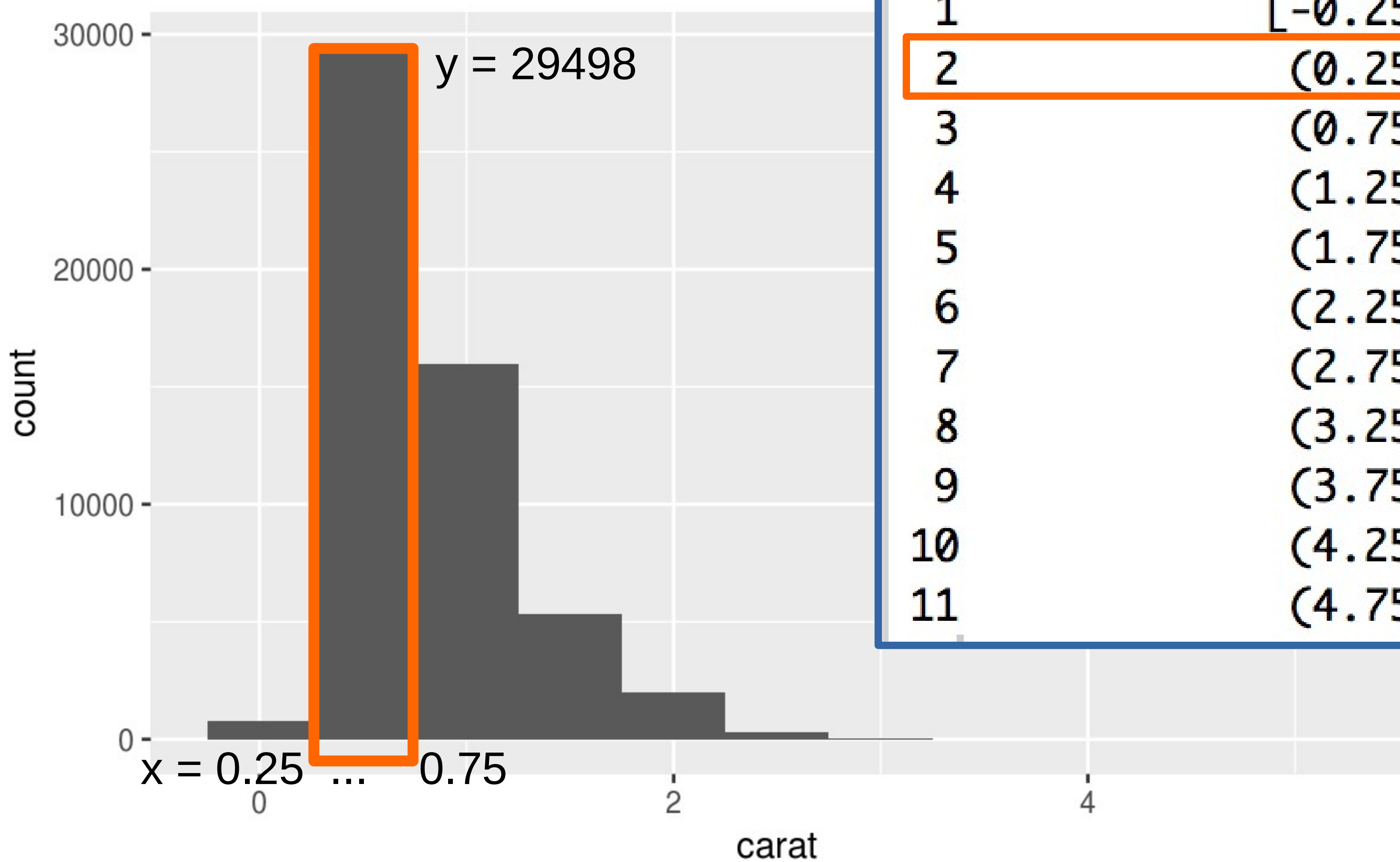
dplyr Basics

- Five key dplyr functions
  - Pick observations by their values (**filter()**).
  - Reorder the rows (**arrange()**).
  - Pick variables by their names (**select()**).
  - Create new variables with functions of existing variables (**mutate()**).
  - Collapse many values down to a single summary (**summarise()**).
- Find help for each: ?keyword

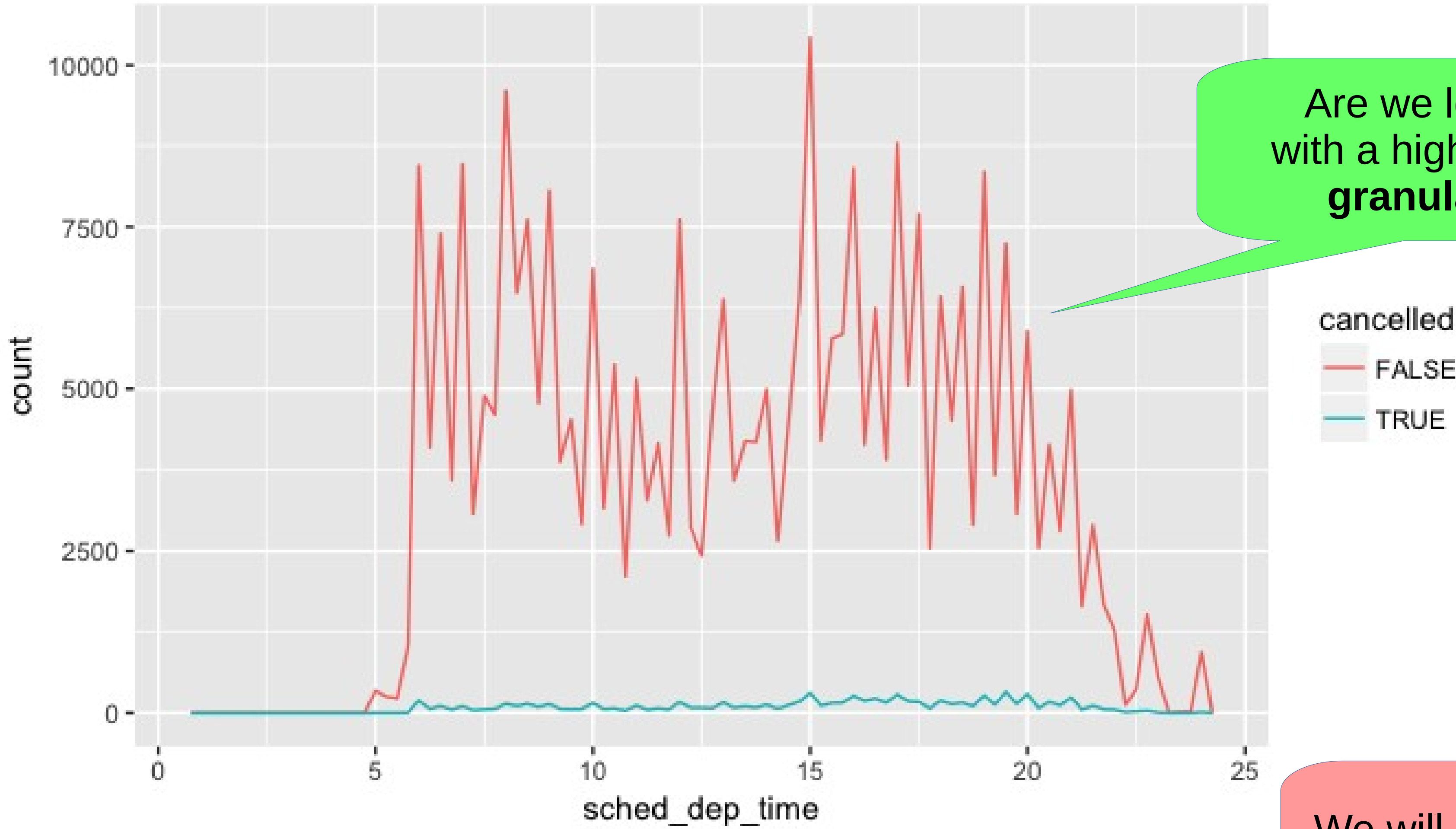
Bins
Dividing Continuous Data

- The cut_width() gives a textual representation of the histogram.

```
> diamonds %>%
+     count(cut_width(carat, 0.5))
# A tibble: 11 x 2
   `cut_width(carat, 0.5)`     n
                    <fctr> <int>
 1           [-0.25,0.25]   785
 2            (0.25,0.75] 29498
 3            (0.75,1.25] 15977
 4            (1.25,1.75]  5313
 5            (1.75,2.25]  2002
 6            (2.25,2.75]   322
 7            (2.75,3.25]    32
 8            (3.25,3.75]     5
 9            (3.75,4.25]     4
10            (4.25,4.75]     1
11            (4.75,5.25]     1
```



Potential Pitfalls in Theory



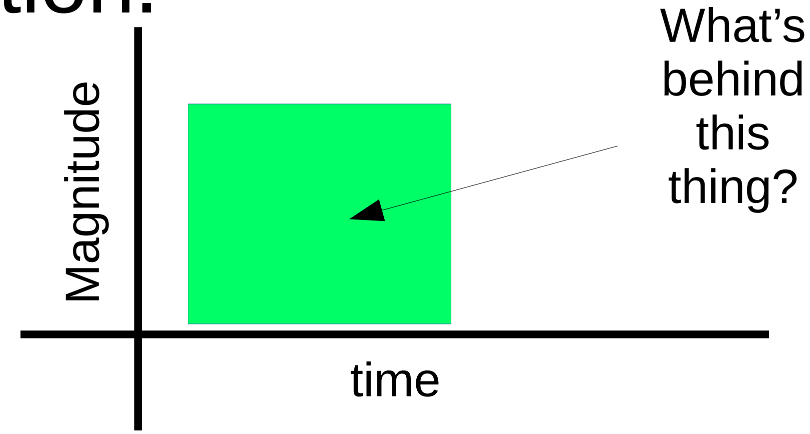
Are we looking with a high-enough **granularity**?

We will return to visual comparisons

- We get an slight idea of cancellations
- But many more non-cancelled flights than cancelled flights

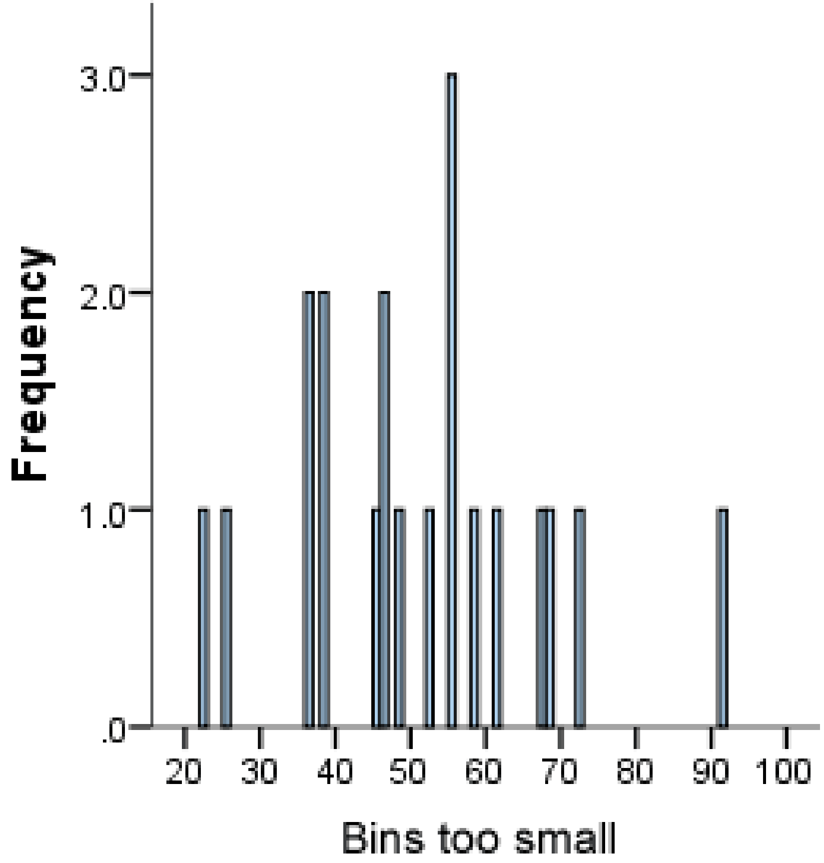
This Plot May Make It Hard To See The Phenomenon

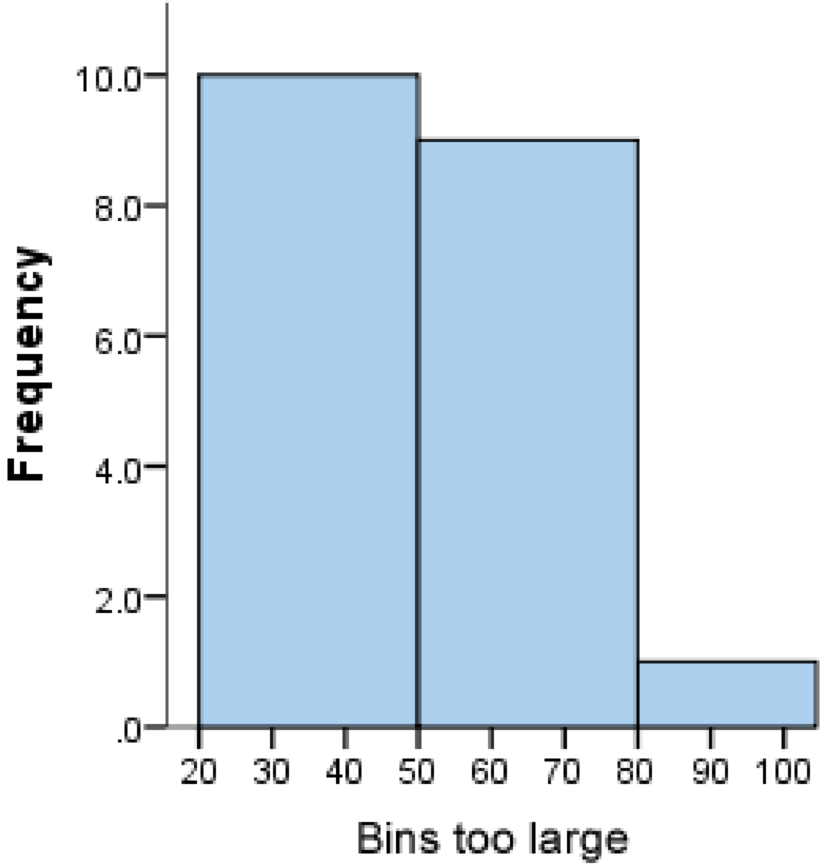
- The counts variable seems to have values from all over the range.
- This is noise in our plot
- If one group is much smaller than the others, then it is hard to see the differences in its distribution.



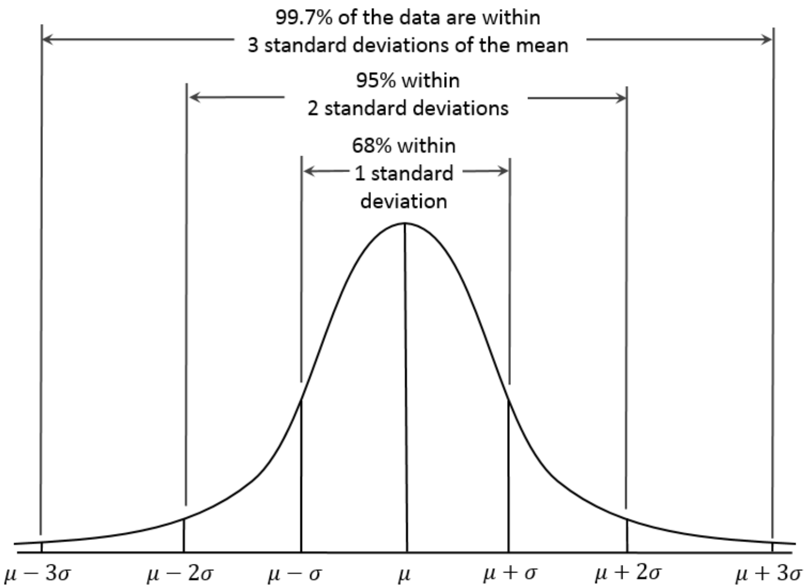
Different Bin Widths

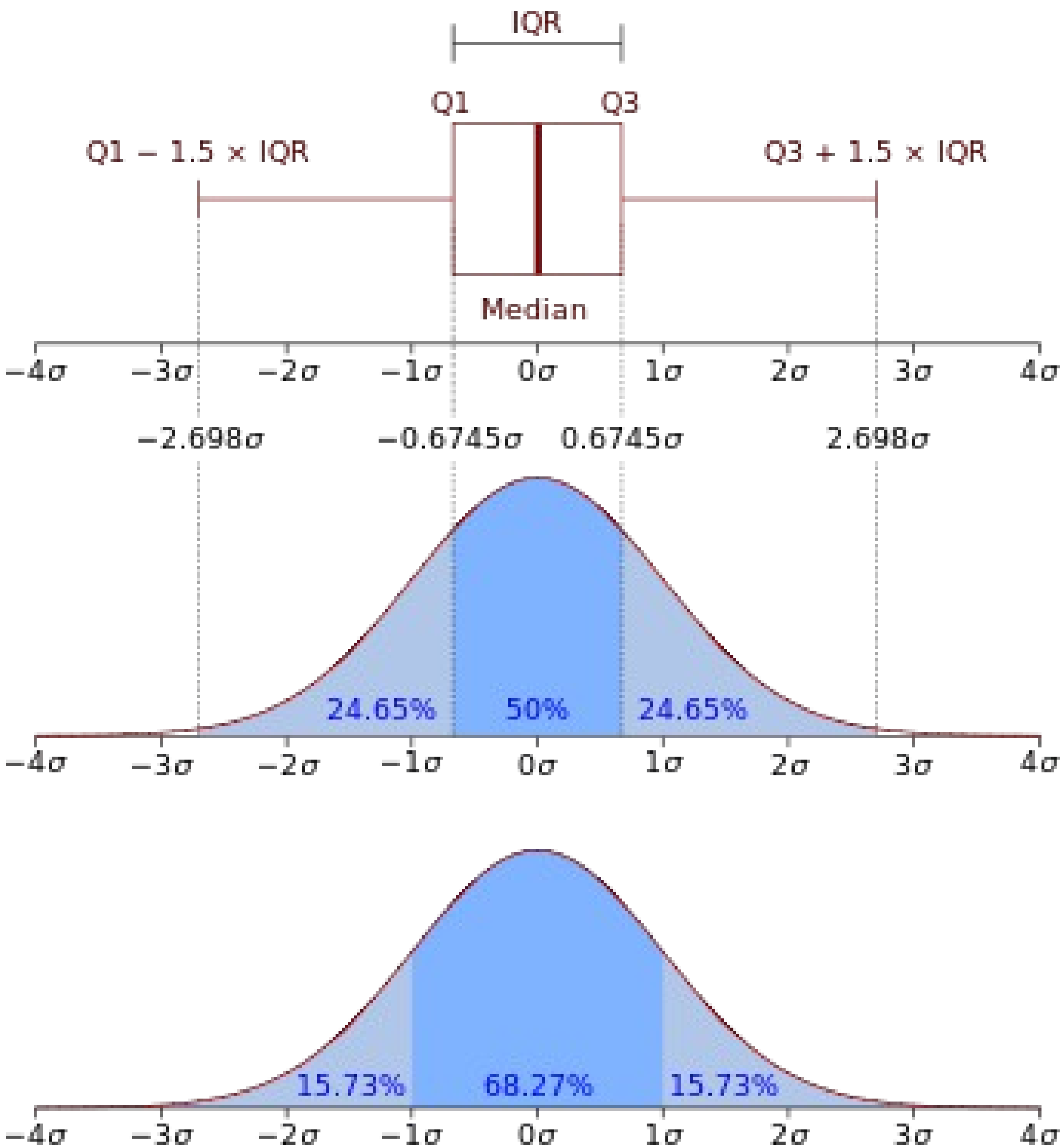
- Set the width of the intervals in a histogram with the binwidth argument, which is measured in the units of the *x* variable.
- Left histogram: bins are too small, too much individual data and hides underlying pattern (frequency distribution).
- Right histogram: bins are too large, hard to spot trends in the data.





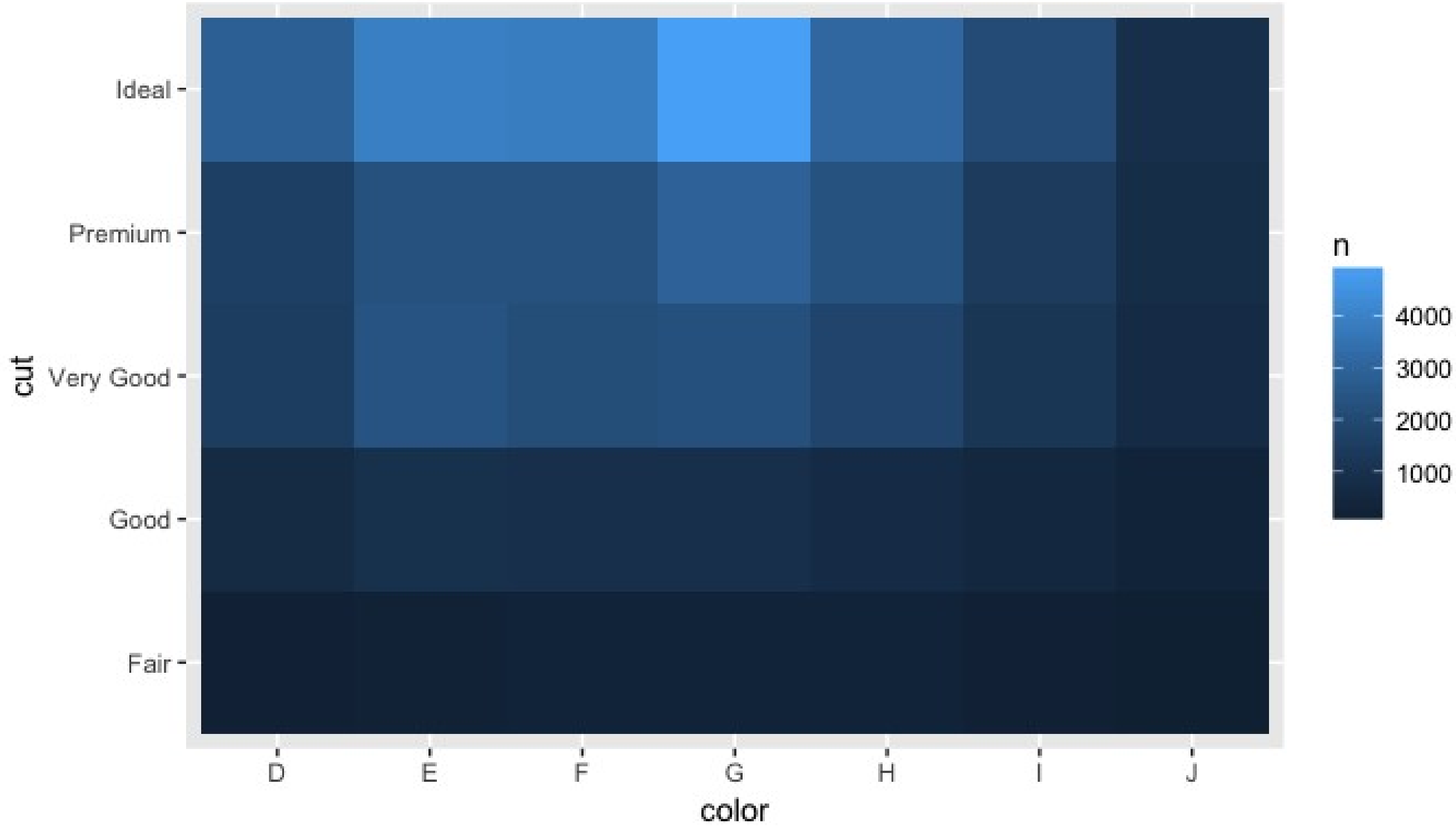
Box Plots





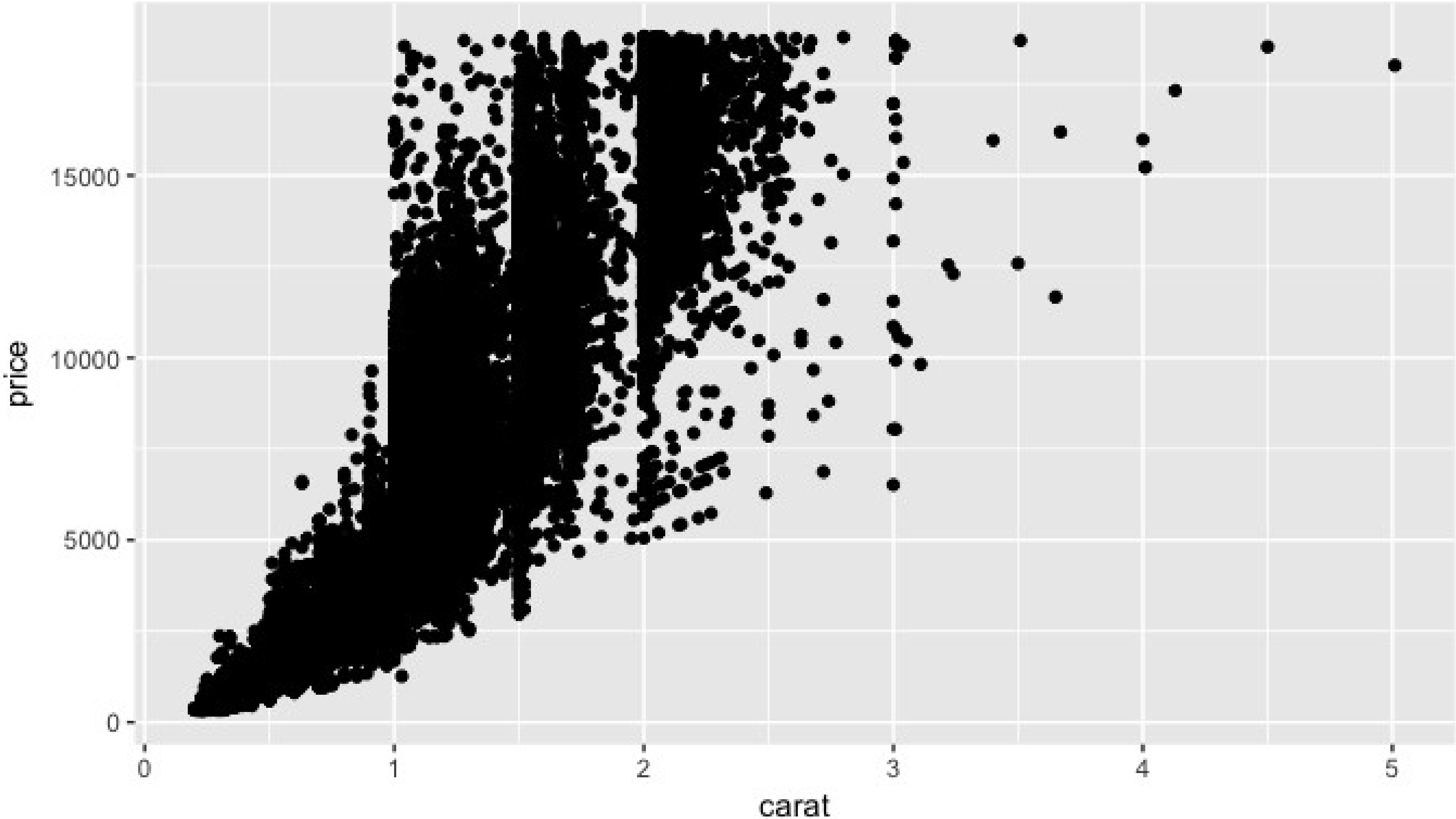
- For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

Visualization - Mini Distributions: Cut vs Color



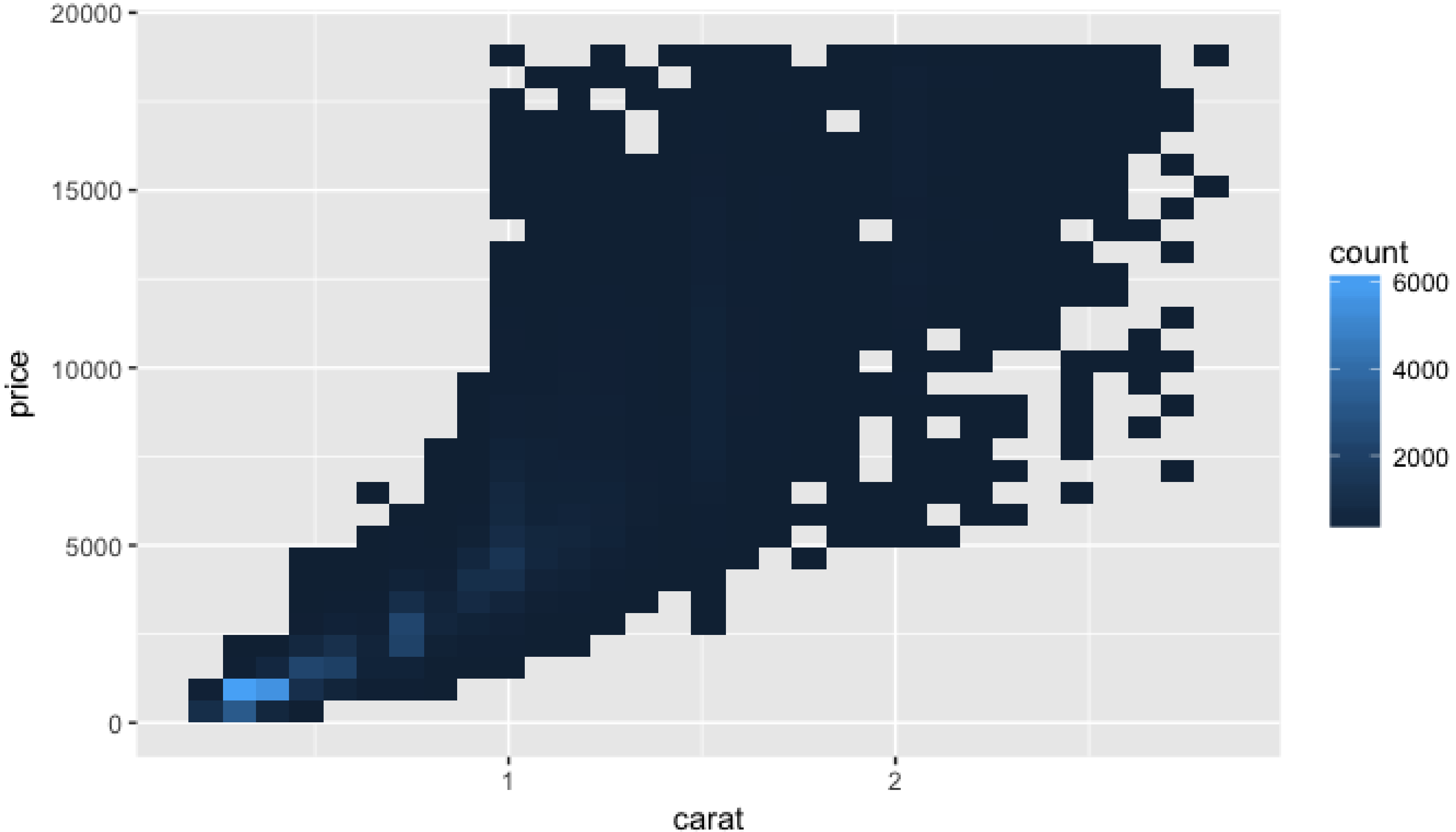

```
diamonds %>%
  count(color, cut) %>%
  ggplot(mapping = aes(x = color, y = cut)) +
    geom_tile(mapping = aes(fill = n))
```

Visualization - Mini Distributions: Carat vs Price




```
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price))
```

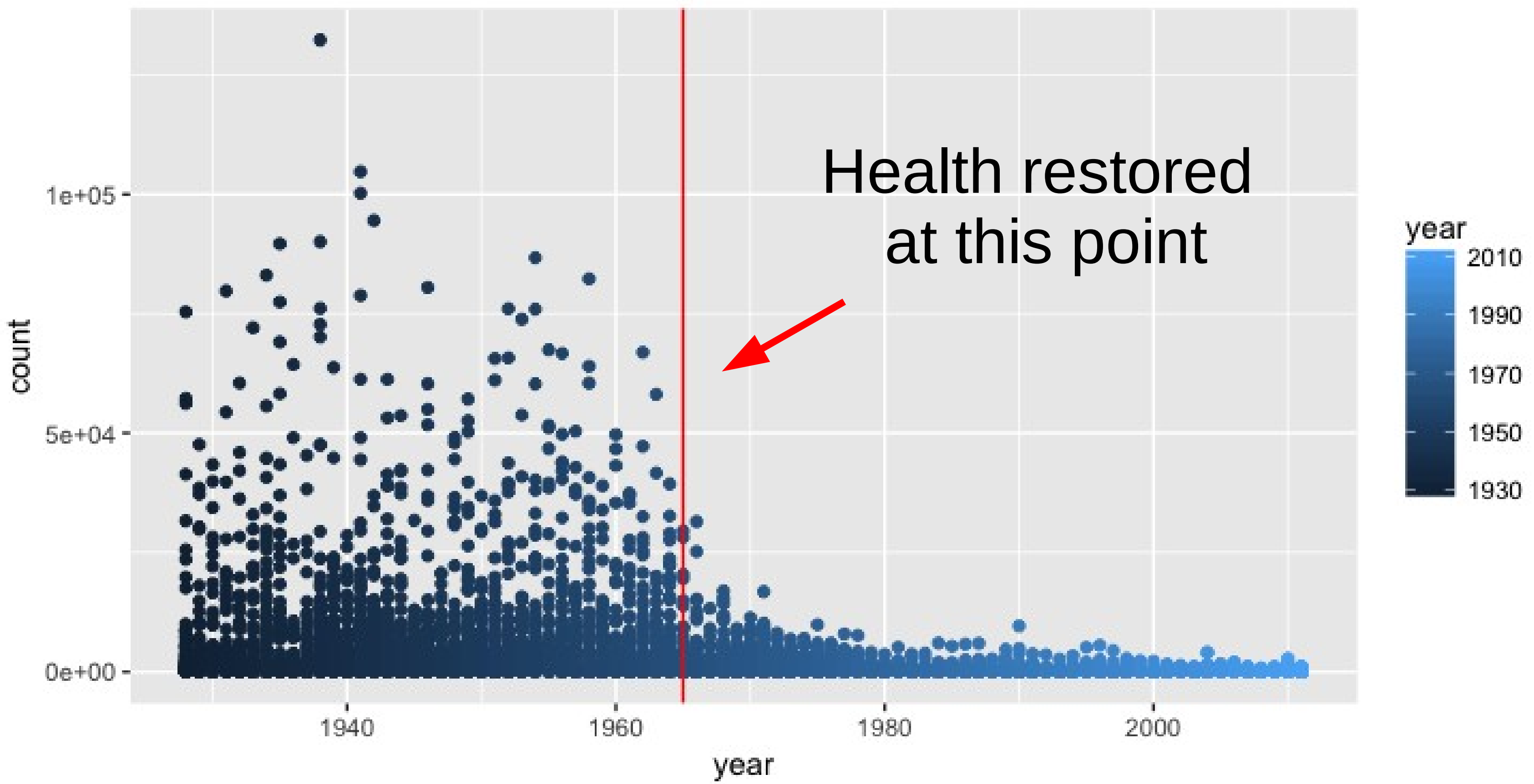
Visualization - Mini Distributions: Carat vs Price




```
ggplot(data = smaller) +
  geom_bin2d(mapping = aes(x = carat, y = price))
```

Vaccine Lab 3: What Does Our Data Say About (All) Vaccines of Data?

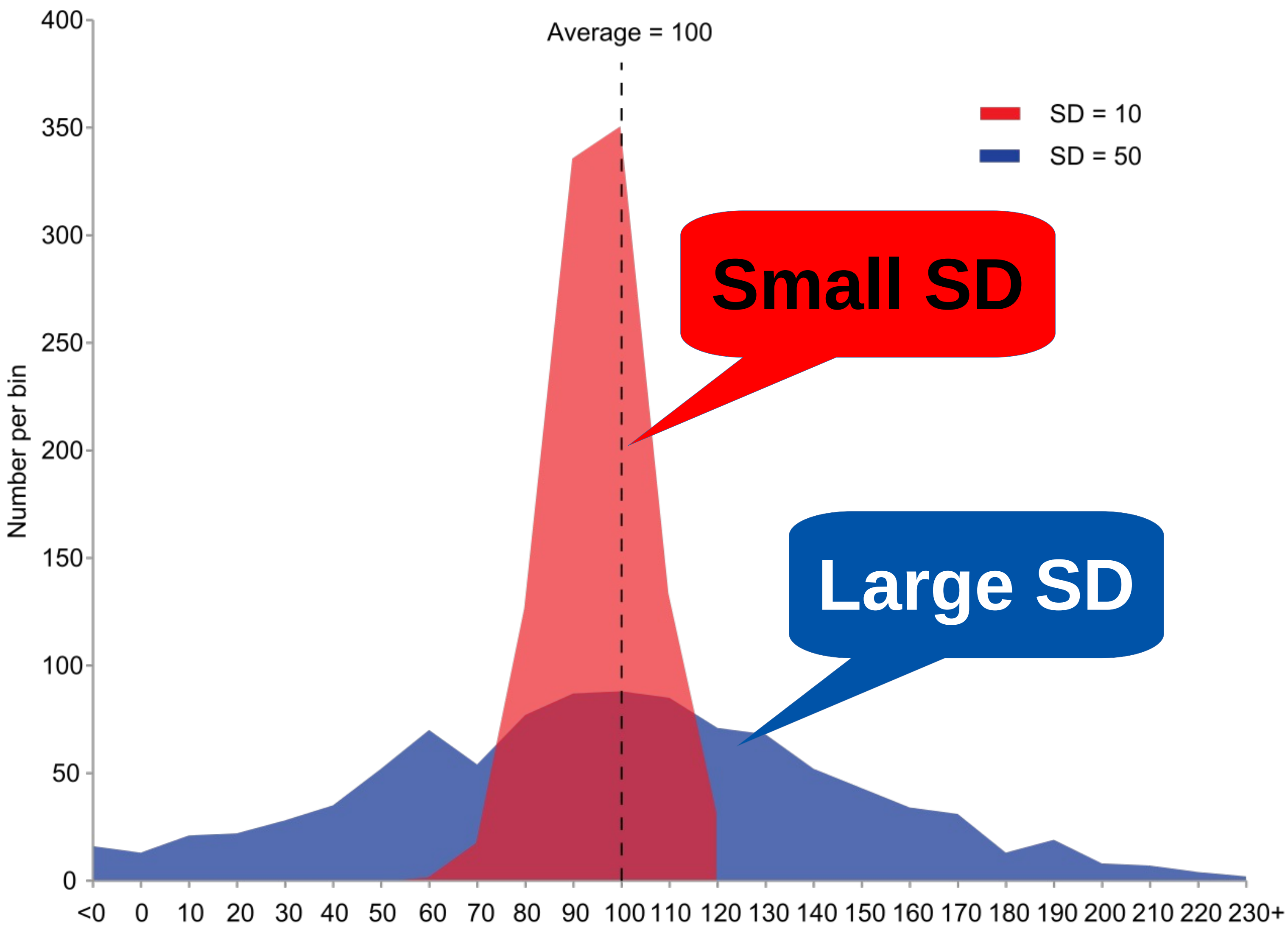
```
library(tidyverse)

library(dslabs)

library(dplyr)

ggplot(data = us_contagious_diseases) +  geom_point(mapping = aes(x = year, y = count, color = year)) + geom_vline(xintercept = 1965, color = "red")
```



Cases of Illness

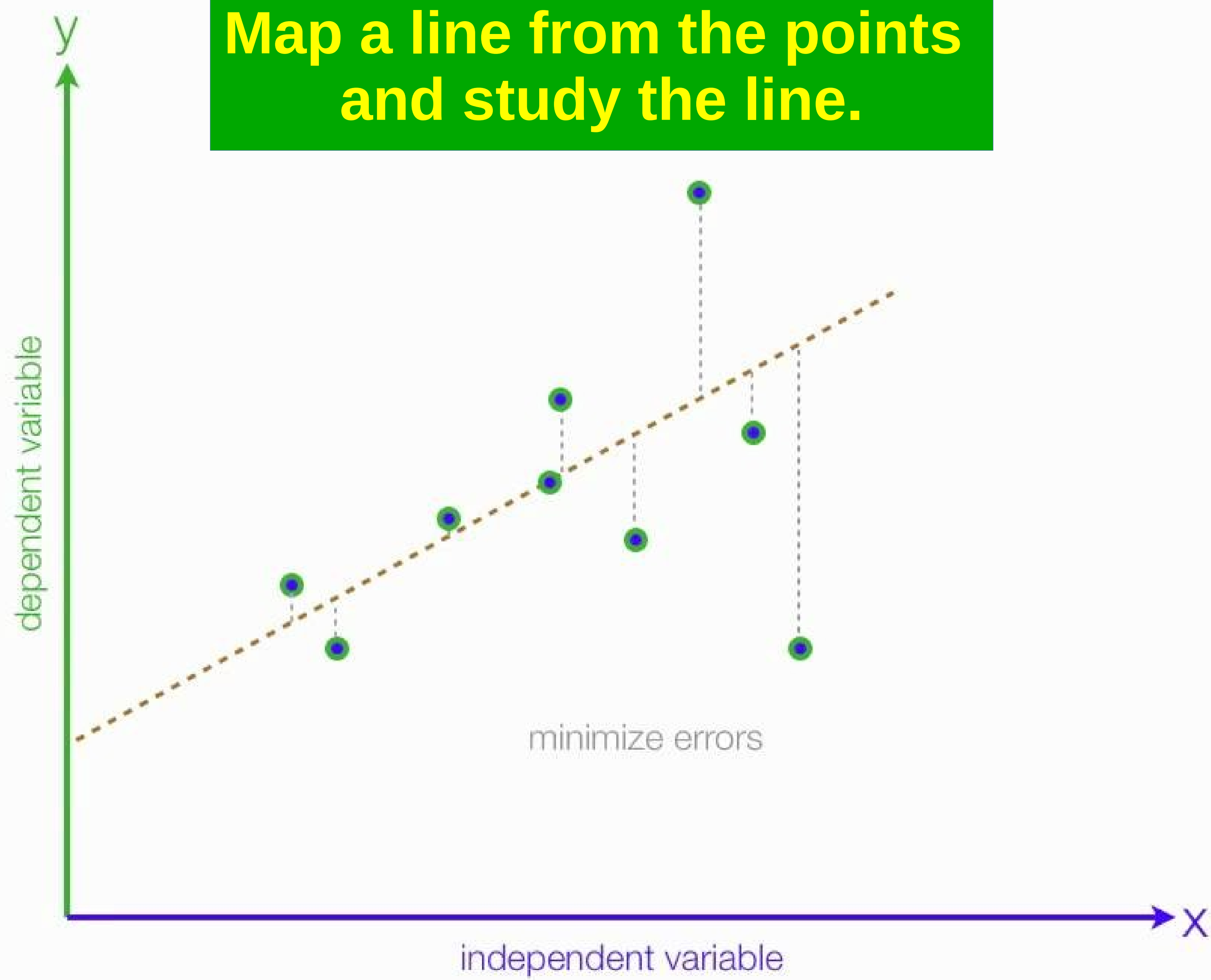
Basic Stats
Standard Deviation

- A quantity calculated to indicate the extent of deviation for a group as a whole.

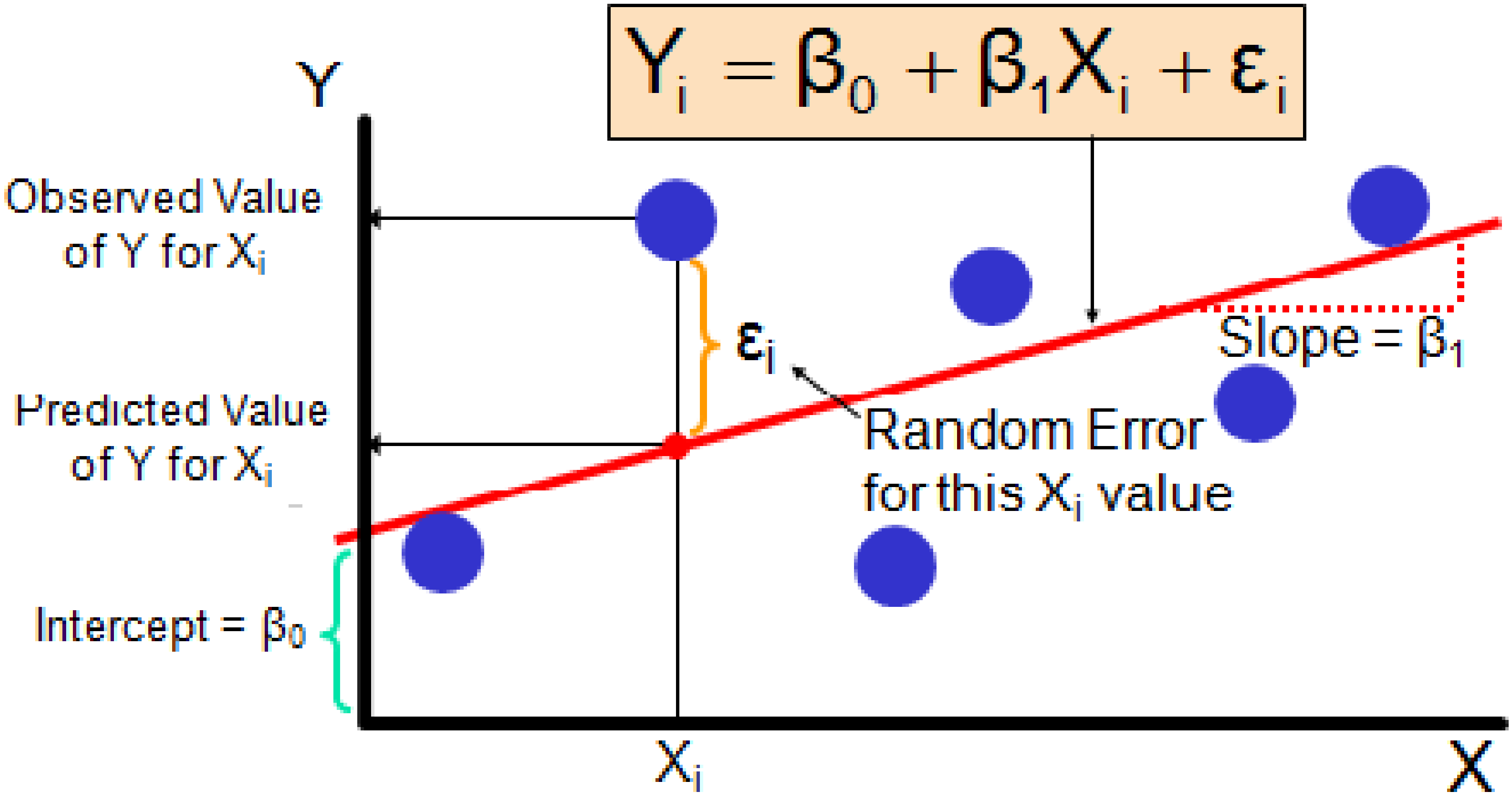


Linear Regression

Map a line from the points and study the line.



Let's Talk Linear Models



Linear Regression

- Is one thing able to influence another thing?
- A linear approach for modeling the relationship between a scalar dependent variable ***y*** and one or more explanatory variables, or independent variables, denoted by ***x***.
- *Simple linear regression*: Single explanatory variable; **models x and y**
- *Multiple linear regression*: More than one explanatory variable (***y's***); **models x and y1, y2**

Let’s Hit the Code

- Linear model syntax

lm

Model formula:
response ~ predictor(s)

data

```
mod <- lm(tc2009 ~ low, data = crime)
```

Linear Regression: Code

```
ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
group <- gl(2, 10, 20, labels = c("Ctl","Trt"))
weight <- c(ctl, trt)
lm.D9 <- lm(weight ~ group)
lm.D90 <- lm(weight ~ group - 1) # omitting intercept
summary(lm.D9)
```

- **H0: there is no relationship between vars, m =0**
- **Ha: There is a relationship between vars, m!=0**

  # Check the p-value:
  - **If p-val =< alpha = 0.05: reject H0.**
  - **If p-val > alpha = 0.05: do not reject H0.**

So, Are My Models Made From Sampling Full Data Set Any Good?

- Use *Parametric statistics* to check your model before you use it!

```
> summary(mod)

Call:
lm(formula = earn ~ height, data = pWages)

Residuals:
   Min     1Q Median     3Q    Max 
-49392 -17589  -4448  10236 108209 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -138901.1    50897.3  -2.729 0.007530 ** 
height         2607.4      760.6   3.428 0.000891 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29100 on 98 degrees of freedom
Multiple R-squared:  0.1071,	Adjusted R-squared:  0.09795 
F-statistic: 11.75 on 1 and 98 DF,  p-value: 0.0008909
```

mod = lm(earn ~ height, data = pWages)

Basic Stats: T-Tests

- Suppose: We are the producers of two kinds o drinks: green and purple. Each drink comes in a bottle and we would like to know whether the green and the purple drink are filled to the same levels.
- We randomly select 9 bottles from our entire set of 100000 bottles

Basic Stats: T-Tests

- By inspection,
  - **Purple bottles seem a little under-filled**
  - **Green bottles seem a little over-filled**
- Can we use a statistical test to conclude whether the whole batch is under- or over-filled?

T-Test: Hypotheses

- We want to know: **Is there a statistically significant difference between the two groups in terms of the average extent to which the bottles are filled?**
  - **Null hypothesis (Ho)**: The bottles are filled the same
  - **Alternative hypothesis (Ha)**: There is a difference between the filling of bottles.
- Remember: we have a sample of only nine bottles from the super set of 100000 bottles. Statistics is used to extrapolate from the small set to the larger set.

Use *p*-Values

- The p-Value says that we are sure that our sample size that we randomly selected if a very good representation of our larger super set.
- 95 confidence interval range: Our selected bottled fit within 95 per cent of the entire set → a good representation of the whole set of 100000 bottles.
- **Reject the Null Hypothesis when $p < 0.05$** (when *p* is close to zero.)

Basic Stats: T-Tests

```
data_drinks <- data_drinks %>%
    select(Colour, percentFull)
#Run the t-test: a comparison of means.
t.test(data = data_drinks, percentFull ~ Colour)
```

 # Check the p-value:

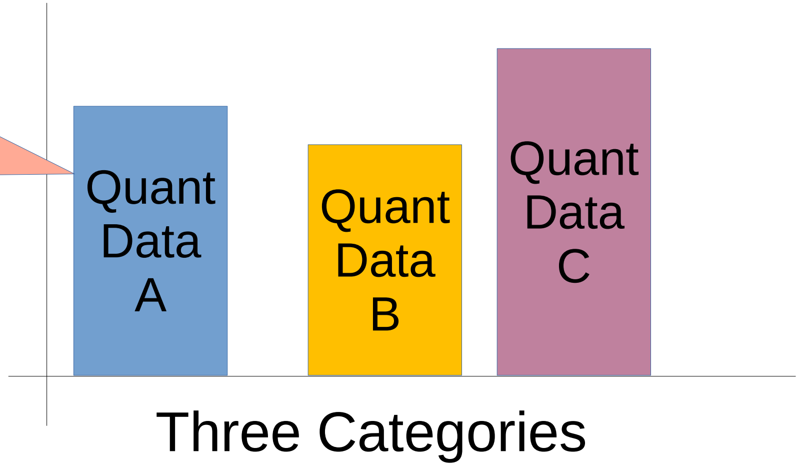
- **If p-val =< alpha = 0.05: reject H0.**
- **If p-val > alpha = 0.05: do not reject H0.**

- **What do we conclude about our *data_drinks*?**

When to Use ANOVA?

- The independent variable should have at least three levels (i.e. at least **three different groups** or categories)
- ANOVA tells you if the *dependent* variable changes according to the level of the *independent* variable

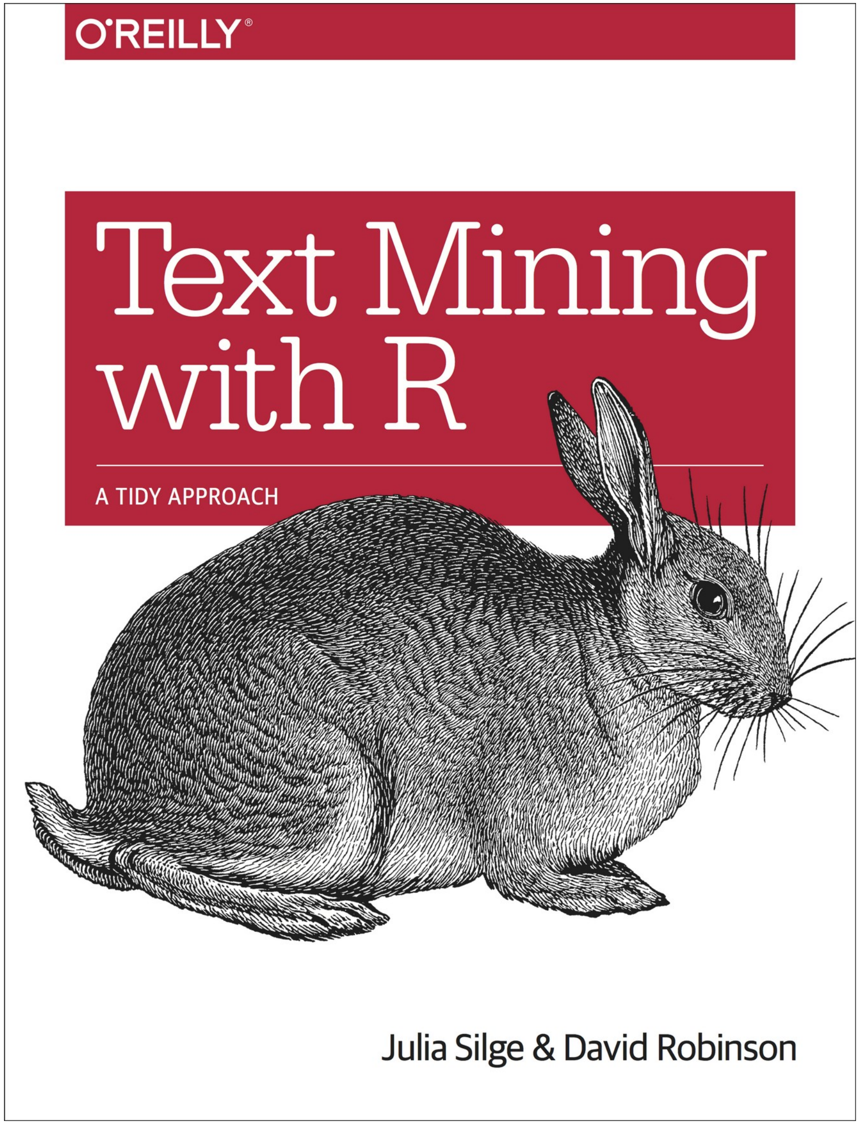
Quantitative data is data that can be counted or measured in numerical values. The two main types of quantitative data are discrete data and continuous data.



Testing Hypotheses in Anova

- Null hypothesis (H0) of the ANOVA: *there is no difference between means*
  - Written as, **H0: μ1 == μ2 == μ3**
- Alternative hypothesis (Ha): *the means are different from one another*
  - Written as, **Ha: μ1 != μ2 != μ3**
- For,
  - **H0** = the null hypothesis,
  - **Ha** = the alternative hypothesis,
  - **μ1** = the mean of population 1
  - **μ2** = the mean of population 2
  - **μ3** = the mean of population 3

In The Textbook



This slide material below has been taken from Silge *et al*.

Chapter: 2
Sentiment analysis with tidy data

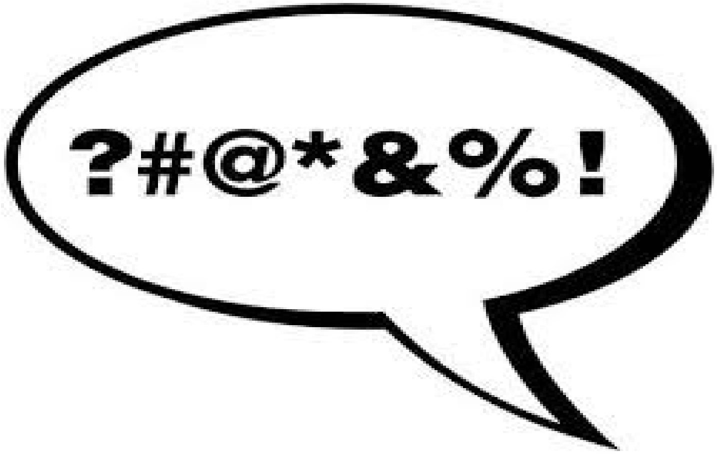
https://www.tidytextmining.com/sentiment.html

And Now. Back to the Research Question!

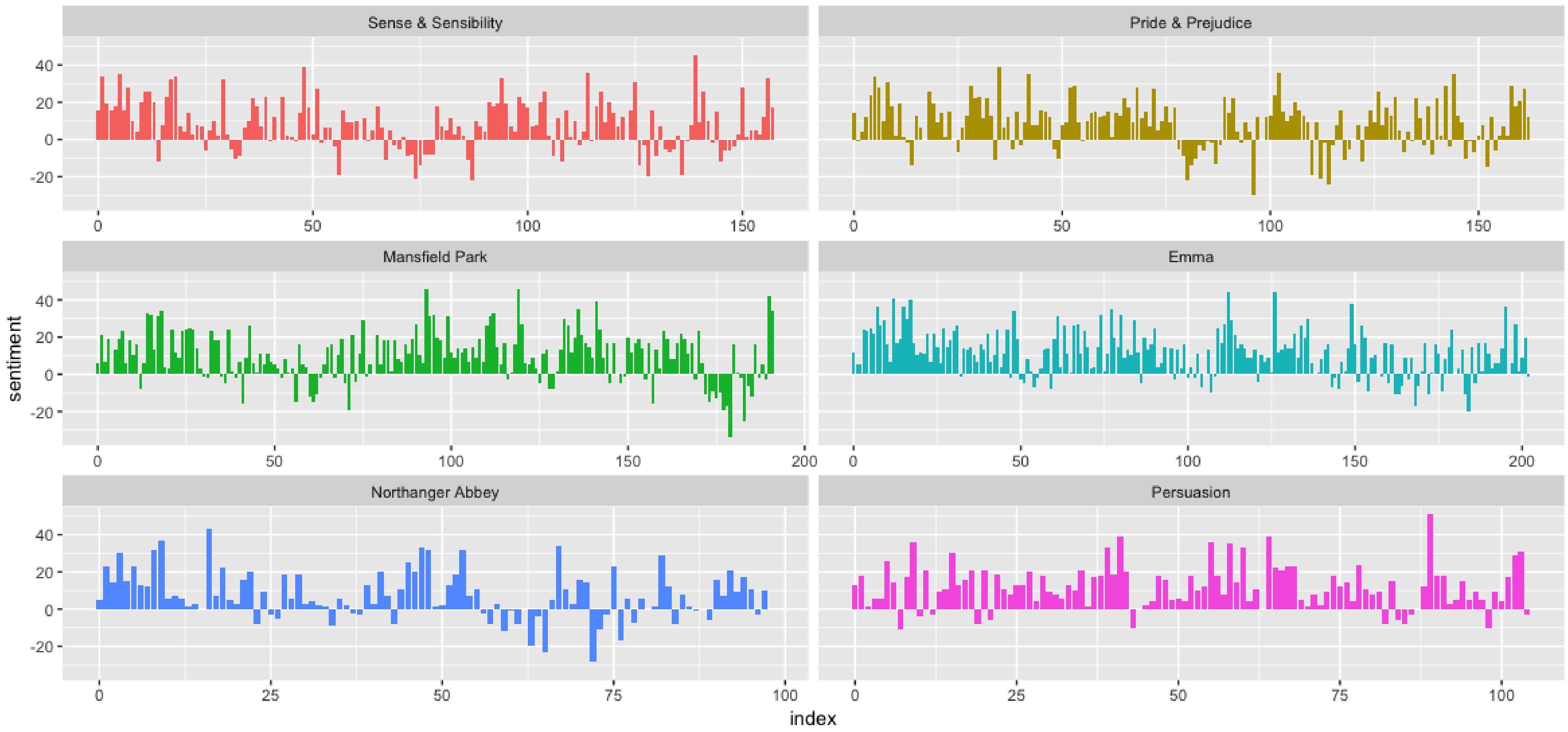
- Jane Austen's written work:

How many *Bad (pessimistic)* words did she use?

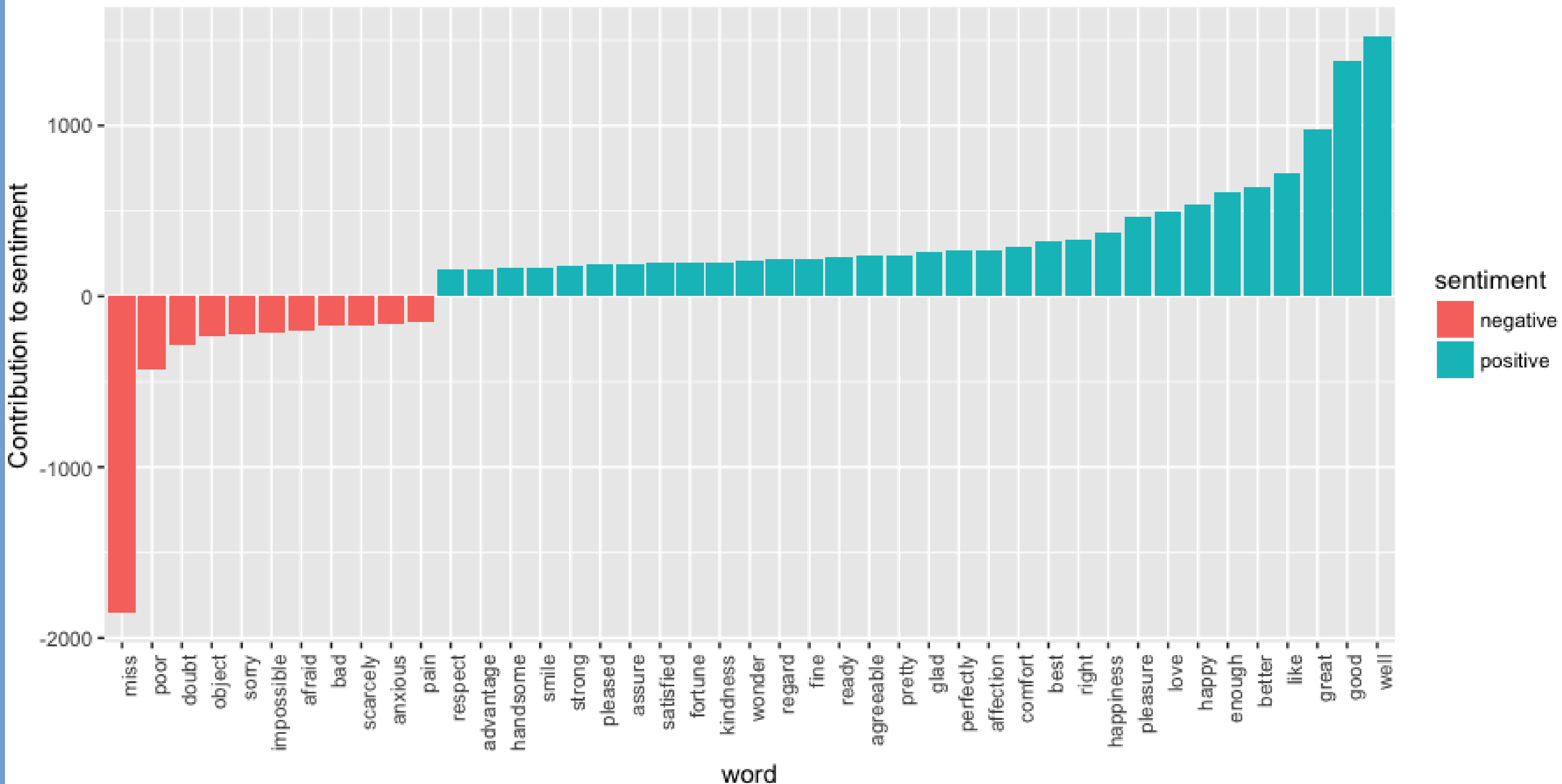
How many *Good* (optimistic) words did she use?



Plot the Good and Bad Words Across Each of Jane Austen's Books



Plot of Positive and Negative Sentiment Words



What Else Did We Cover?

SO MUCH

MORE

It’s Your Turn!

Now, Go Decorate Your Resume!!

Machine Learning, Anyone?

- An Introduction to Machine Learning with R
  - https://lgatto.github.io/IntroMachineLearningWithR/
- Machine Learning in R for beginners
  - https://www.datacamp.com/community/tutorials/machine-learning-in-r#five
- Your First Machine Learning Project in R Step-By-Step
  - https://machinelearningmastery.com/machine-learning-in-r-step-by-step/
- Intro to Machine Learning with R & caret
  - https://www.youtube.com/watch?v=z8PRU46I3NY
- Machine learning with the "diabetes" data set in R
  - https://towardsdatascience.com/machine-learning-with-the-diabetes-data-set-in-r-11fa7ae944d0