Welcome to DATA 151

I’m so glad you’re here!

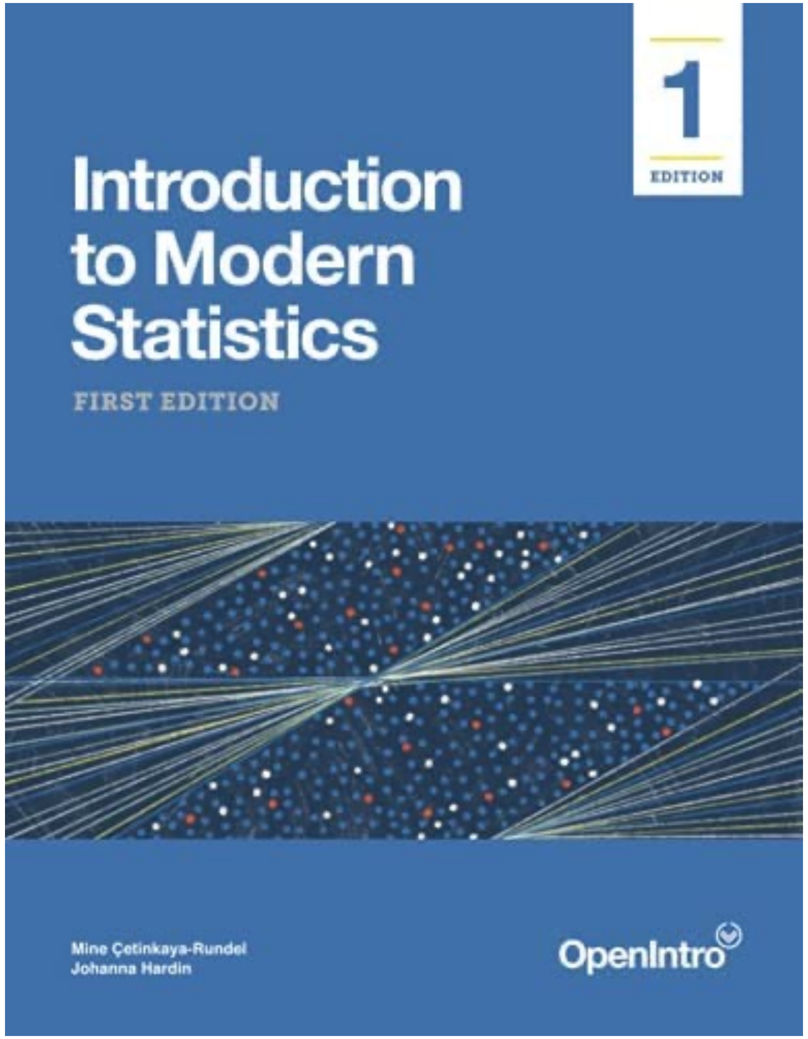
DATA 151: CLASS 12A
INTRODUCTION TO DATA SCIENCE (WITH R)

LINEAR REGRESSION WITH SUBGROUPS

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING



Introduction to Data Science:

- Tuesday
- Introduction to Modern Statistics
  - Ch 7: Relationships between two variables

HOMEWORK REMINDER

Due this week:

- ***DUE 11/17*** *Project Milestone #6*
  - Relationships between two numeric variables
  - Linear regression
- ***CANCELLED***
  - ~~***DUE 11/17*** *HW #10: DC Correlation and Regression*~~

EXPLORING SUB-GROUPS

INDICATORS AND INTERACTIONS

Example: **Shipping Books**

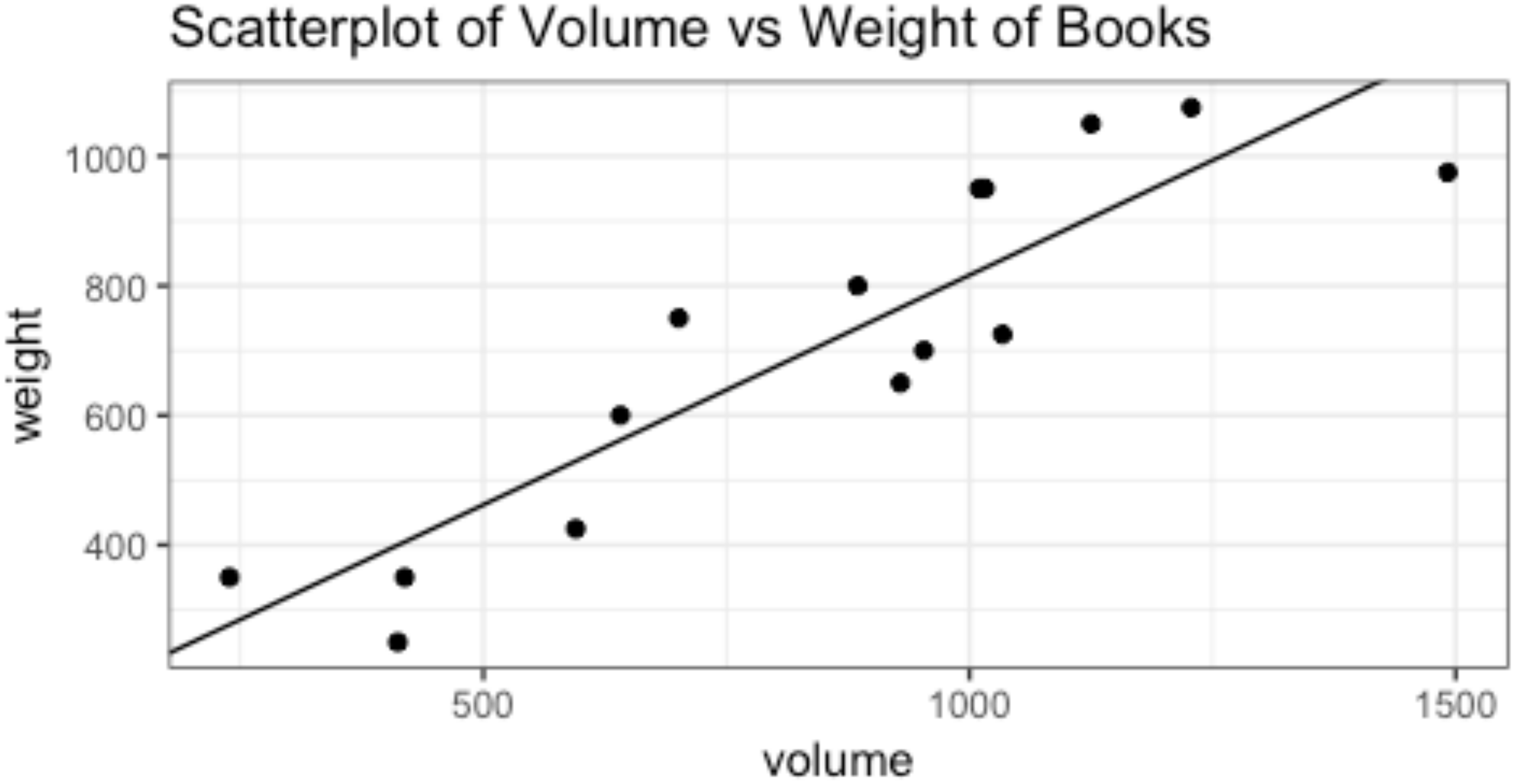
When you buy a book off Amazon, you get a quote for how much it costs to ship. This is based on the weight of the book. If you didn't know the weight of the book, what other characteristics of it could you measure to help predict the weight?

GROUP CODING

START WITH SLR

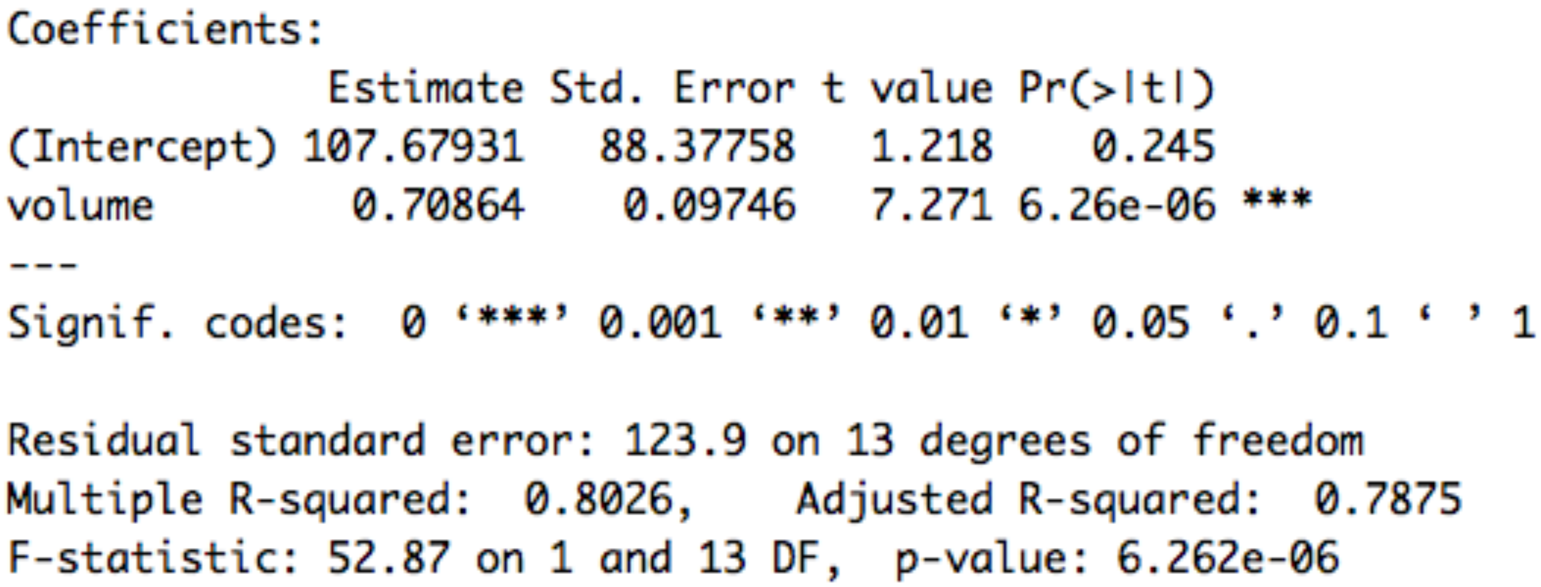
Example: Shipping Books

```
m3<-lm(weight~volume, data=books)
summary(m3)
```



START WITH SLR

Example: Shipping Books (Model Output)



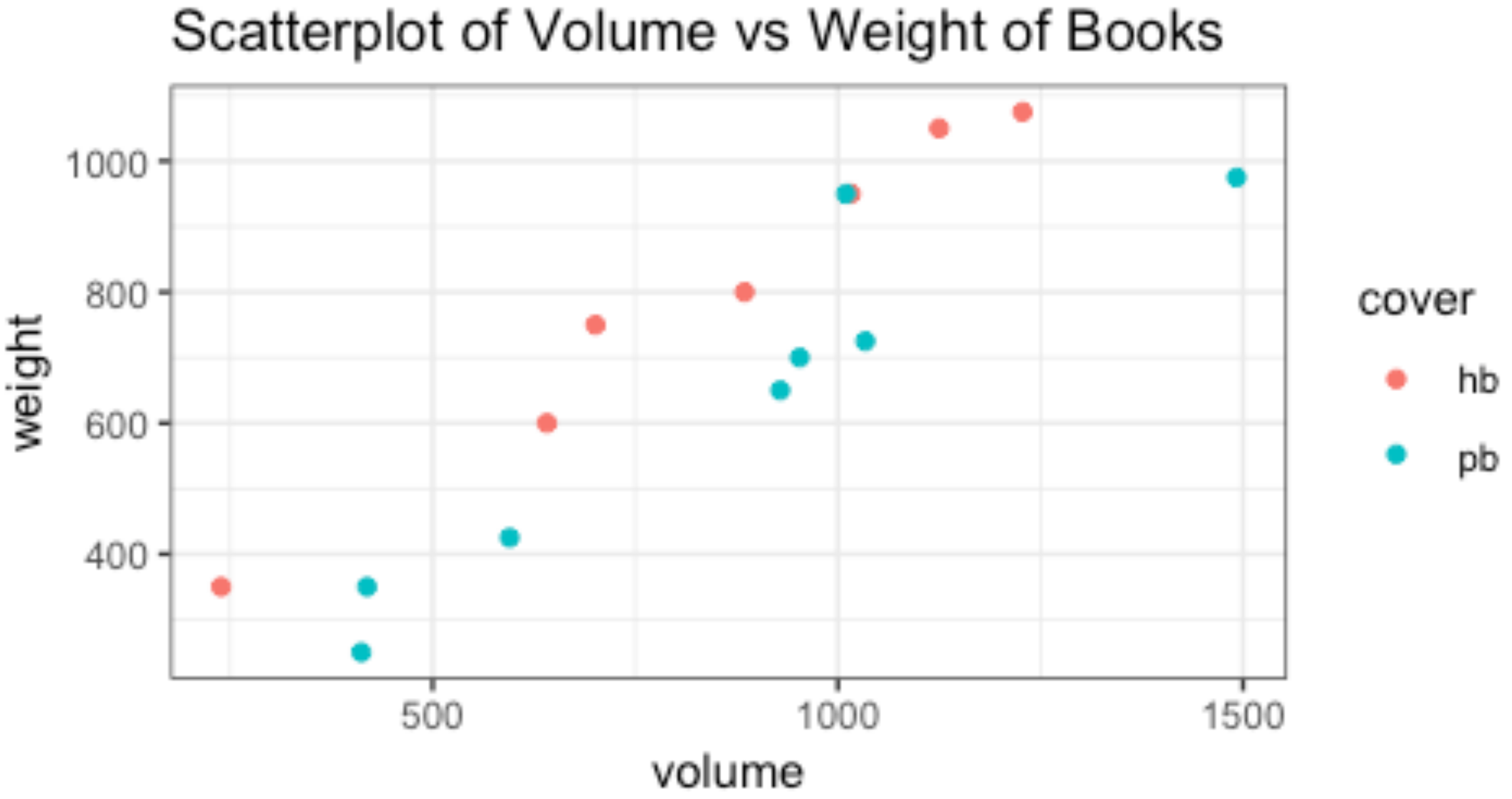
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 107.67931   88.37758   1.218    0.245    
volume        0.70864    0.09746   7.271 6.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared:  0.8026,	Adjusted R-squared:  0.7875 
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

$$\widehat{\text{weight}} = 107.68 + 0.71 \times volume$$

PARALLEL LINES

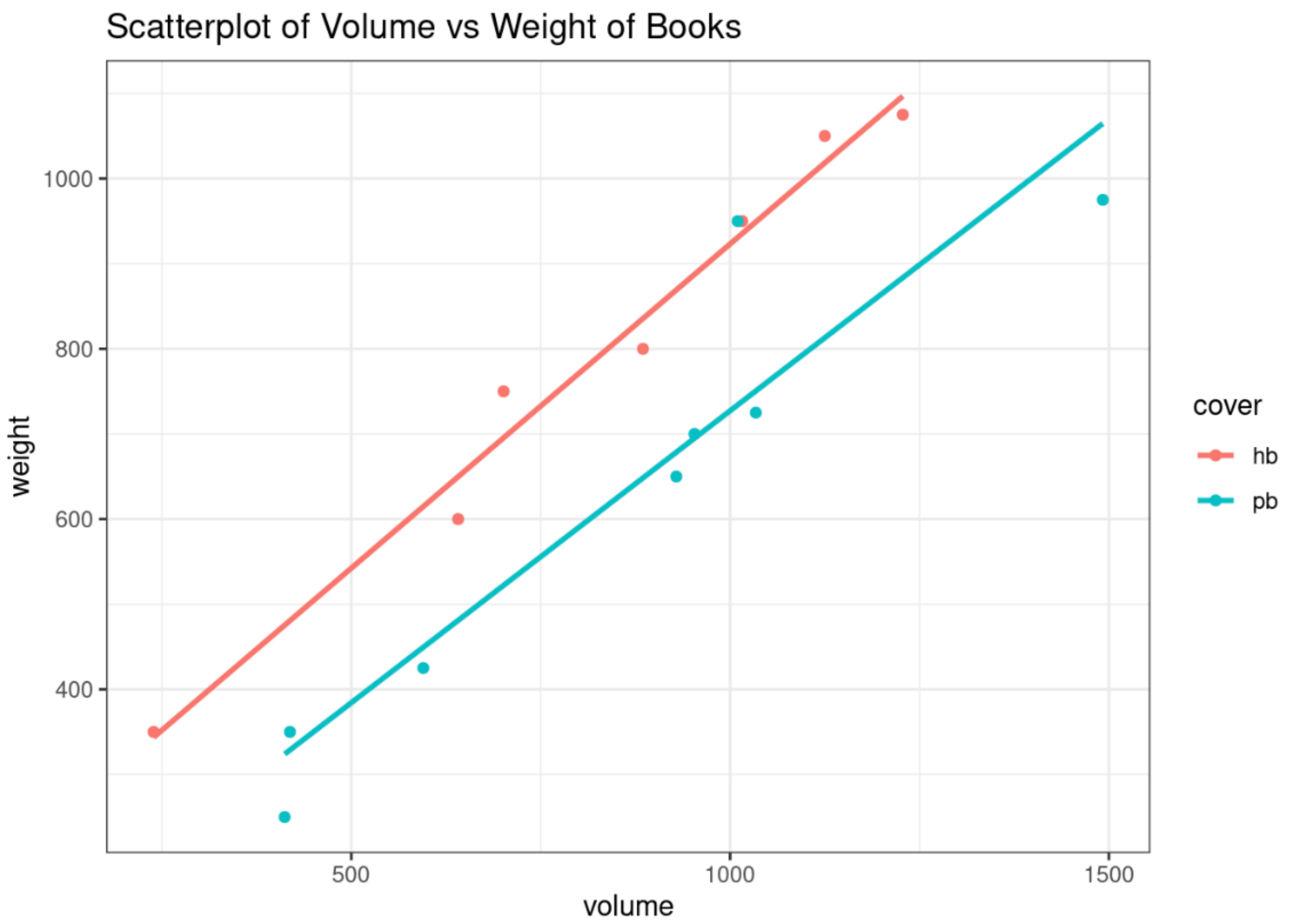
Would including cover type help out model explain more variation?



RStudio®

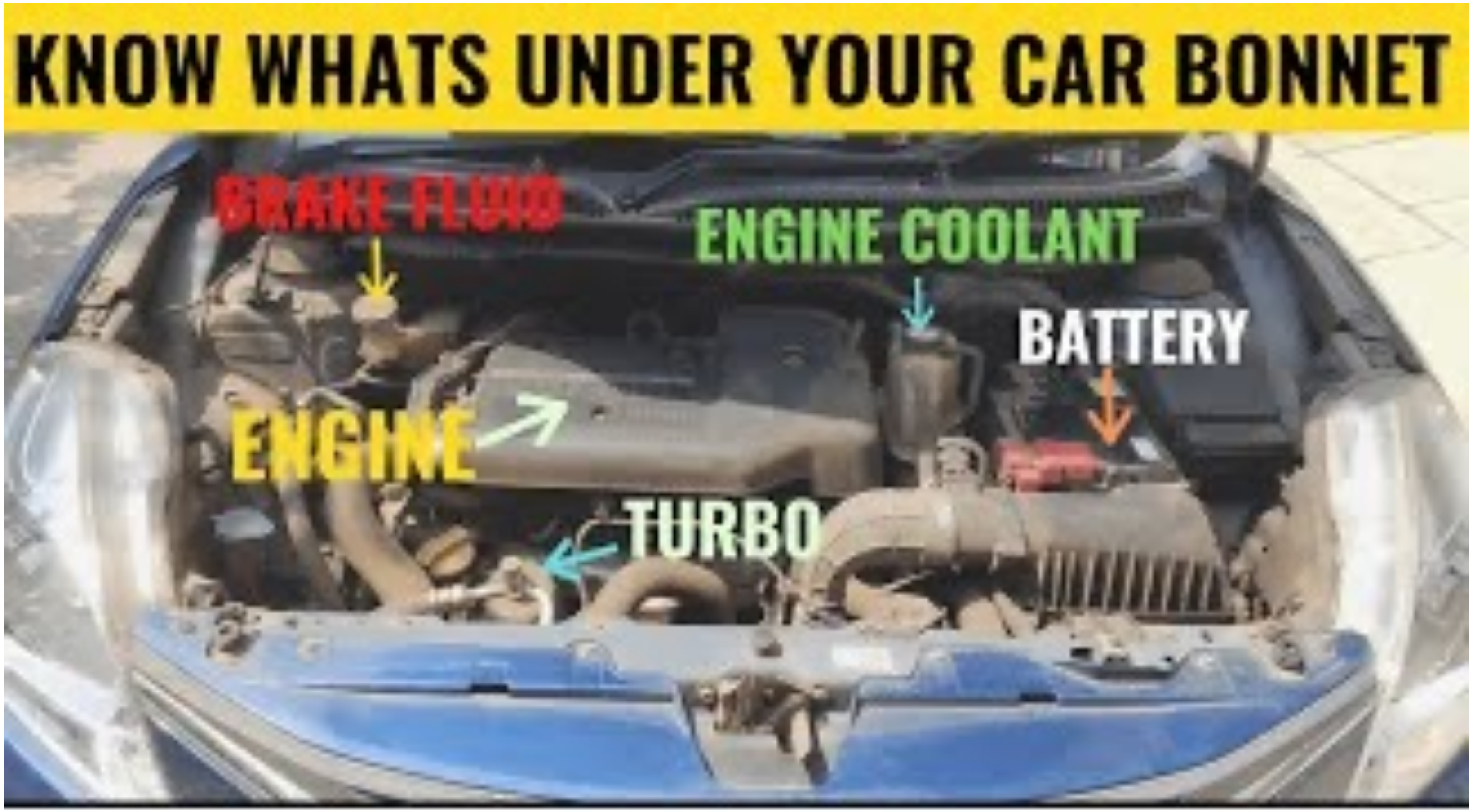
GROUP CODING

INTERACTIONS



HOW DOES THIS WORK? WHAT ARE THESE LINES?

MORE DETAILS IN DATA 152 AND DATA 252


KNOW WHATS UNDER YOUR CAR BONNET
BRAKE FLUID
ENGINE COOLANT
BATTERY
ENGINE
TURBO

INTERACTIONS

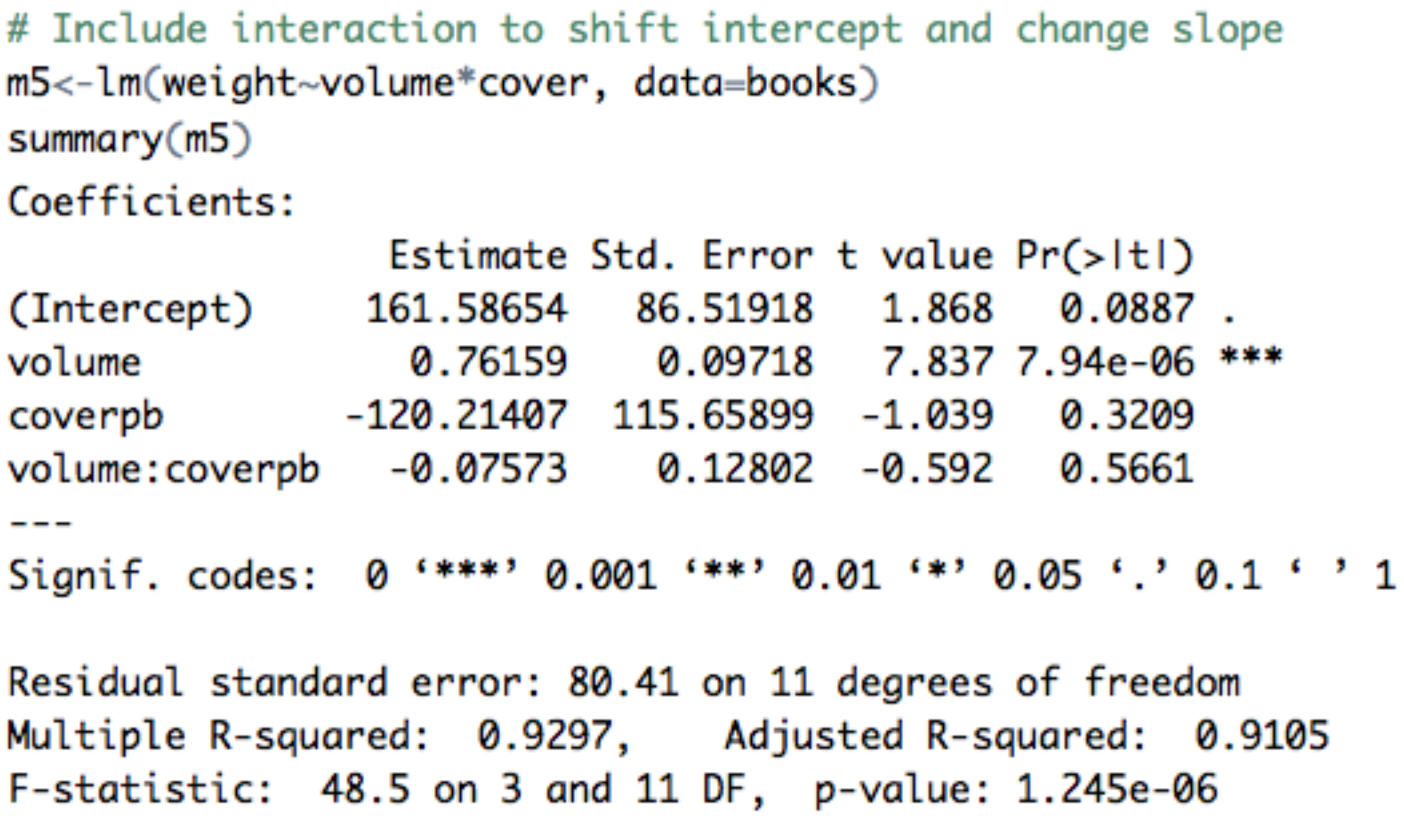
- In R
  - “*” All possible subsets of interactions (and main effects)
  - “:“ Only the specified interaction

- Test significance of interaction

- **Hierarchical principle**: If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant

INTERACTIONS

A shift in the intercept was significant, maybe we should also allow for different slopes.



```
# Include interaction to shift intercept and change slope
m5<-lm(weight~volume*cover, data=books)
summary(m5)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     161.58654   86.51918   1.868   0.0887 .
volume            0.76159    0.09718   7.837 7.94e-06 ***
coverpb        -120.21407  115.65899  -1.039   0.3209
volume:coverpb   -0.07573    0.12802  -0.592   0.5661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.41 on 11 degrees of freedom
Multiple R-squared:  0.9297,    Adjusted R-squared:  0.9105
F-statistic:  48.5 on 3 and 11 DF,  p-value: 1.245e-06
```

INTERACTIONS

`volume:cover` is an interaction term.

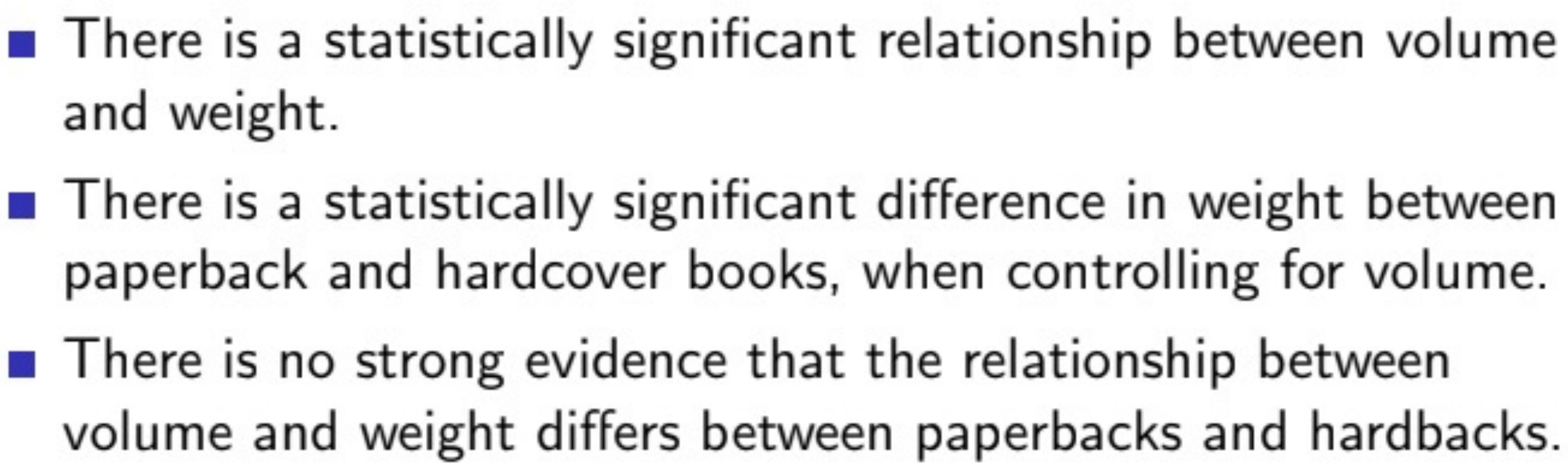
- It describes how the relationship between volume and weight may be different for the two cover type groups.

So we really have two different lines with different intercepts and slopes,

- Hardcover: $weight = 161.59 + 0.76 \times volume + (-120.21) \times 0 + (-0.08) \times volume \times 0$
  $\rightarrow weight = 161.59 + 0.76 \times volume$
- Paperback: $weight = 161.59 + 0.76 \times volume + (-120.21) \times 1 + (-0.08) \times volume \times 1$
  $\rightarrow weight = 41.38 + 0.68 \times volume$

INDICATORS AND INTERACTIONS

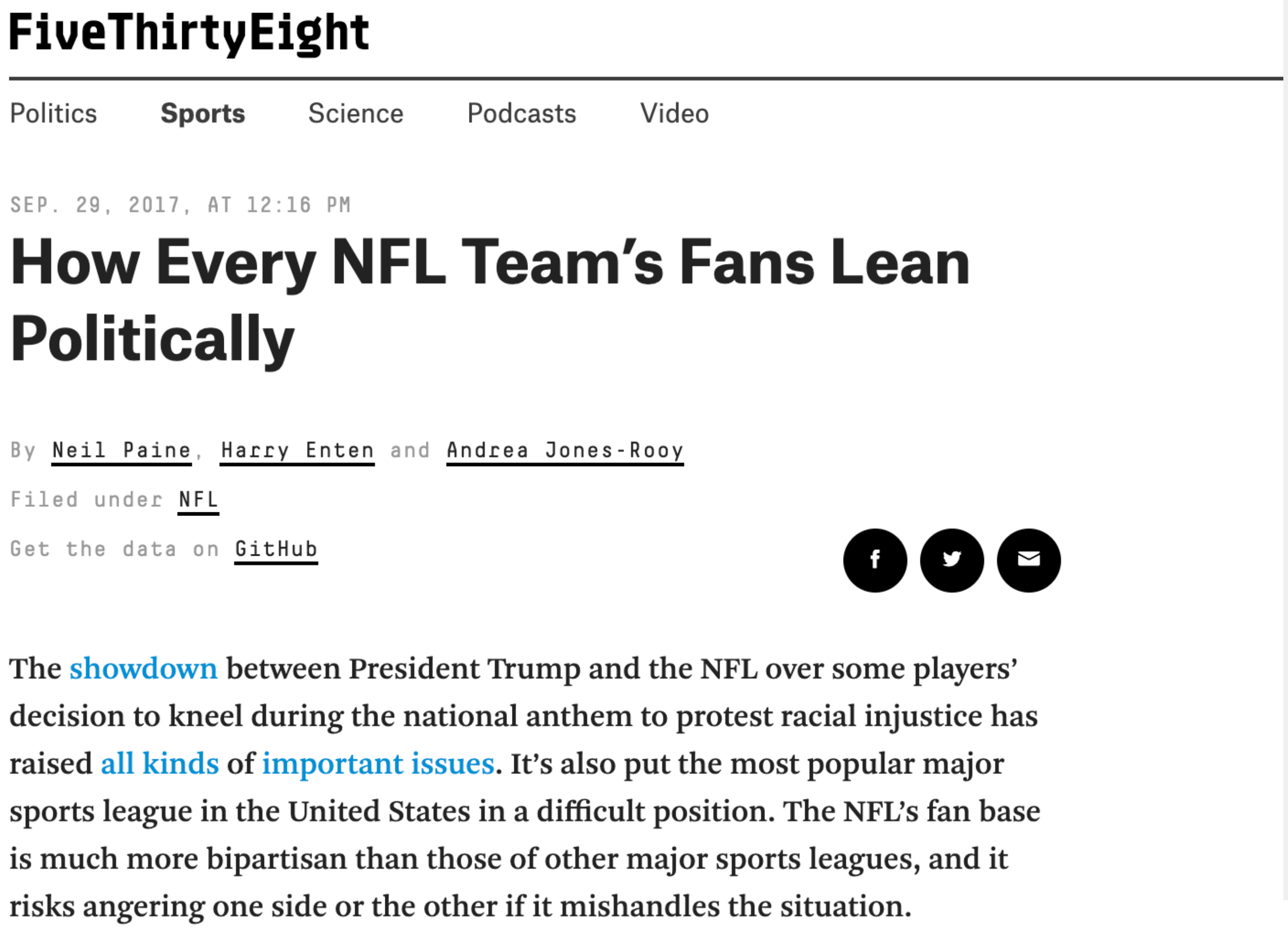
Take home messages:



- There is a statistically significant relationship between volume and weight.
- There is a statistically significant difference in weight between paperback and hardcover books, when controlling for volume.
- There is no strong evidence that the relationship between volume and weight differs between paperbacks and hardbacks.

FIVETHIRTYEIGHT ACTIVITY

READ THE ARTICLE



FiveThirtyEight

Politics Sports Science Podcasts Video

SEP. 29, 2017, AT 12:16 PM

How Every NFL Team’s Fans Lean Politically

By Neil Paine, Harry Enten and Andrea Jones-Rooy

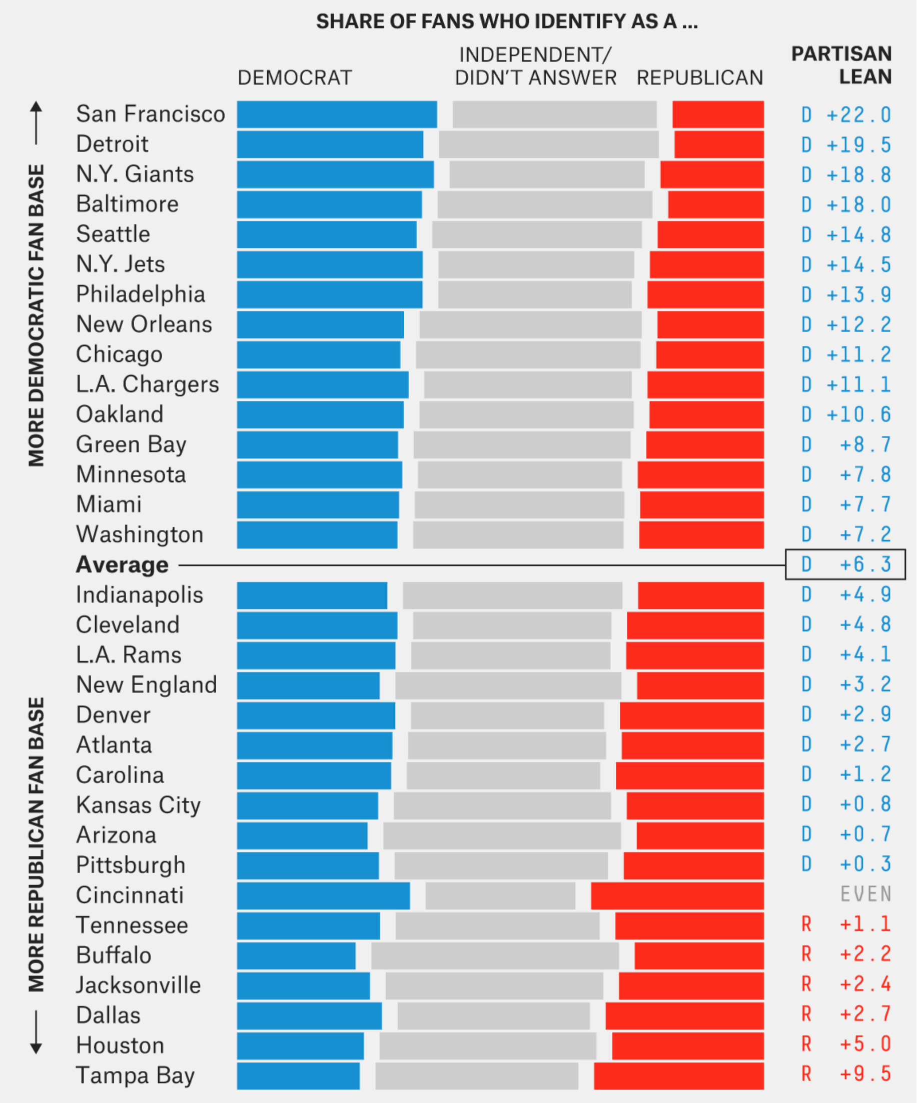
Filed under NFL

Get the data on GitHub

The showdown between President Trump and the NFL over some players’ decision to kneel during the national anthem to protest racial injustice has raised all kinds of important issues. It’s also put the most popular major sports league in the United States in a difficult position. The NFL’s fan base is much more bipartisan than those of other major sports leagues, and it risks angering one side or the other if it mishandles the situation.

The political leanings of every NFL team’s fans

Based on a national survey of 2,290 American NFL fans conducted from Sept. 1 to Sept. 7



FiveThirtyEight

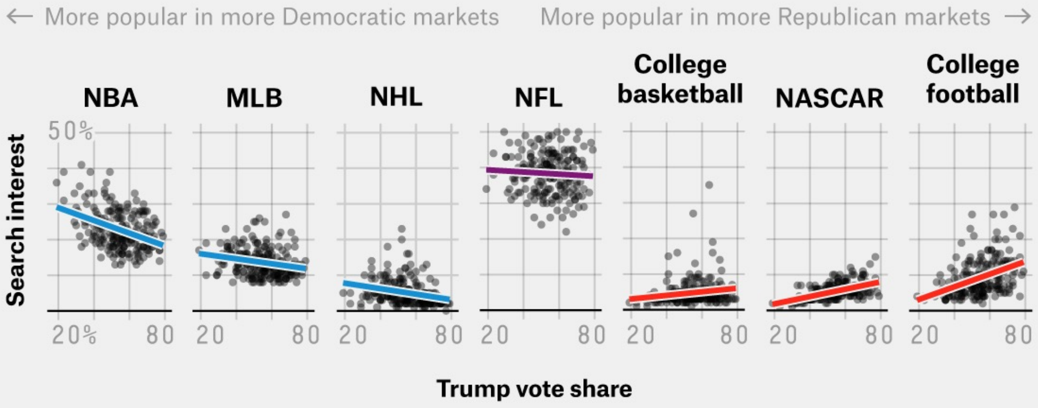
SOURCE: SURVEYMONKEY AUDIENCE

DISCUSS IN SMALL GROUPS

1. How are graphics used to tell the author’s story?
2. What geometries are used?

The NFL has appeal everywhere

Donald Trump’s 2016 vote share compared with search interest for seven major sports, by media market



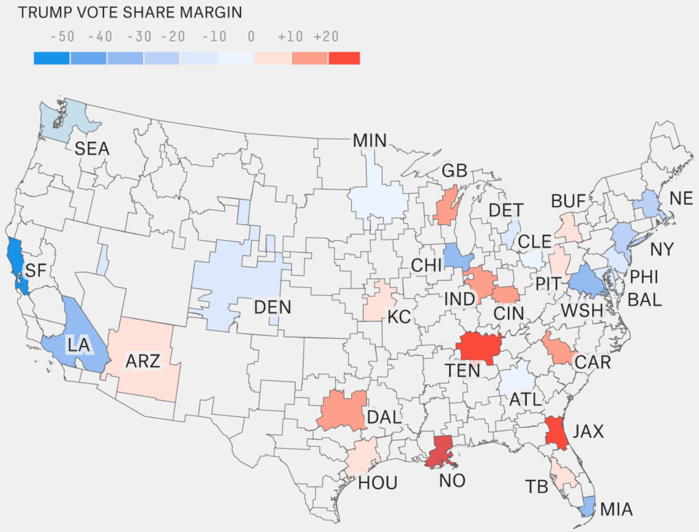
Search interest based on Google Trends data from 2012 to 2017

FiveThirtyEight

SOURCES: GOOLE TRENDS, ECHELON INSIGHTS

Partisanship and the NFL

Team media markets by Donald Trump’s 2016 presidential vote share margin over Hillary Clinton

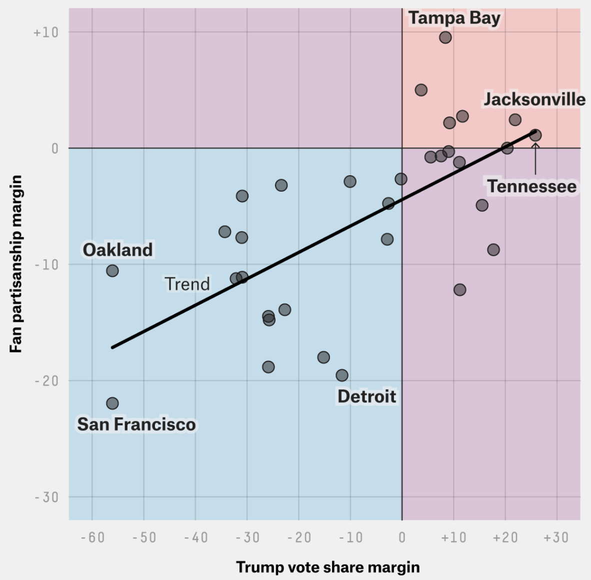


FiveThirtyEight

SOURCES: ECHELON INSIGHTS, NIELSEN

NFL fan partisanship follows regional voting patterns

Difference in the share of self-identified Republican and Democratic fans (according to a SurveyMonkey Audience poll) vs. Donald Trump’s 2016 vote share margin over Hillary Clinton, by NFL media market

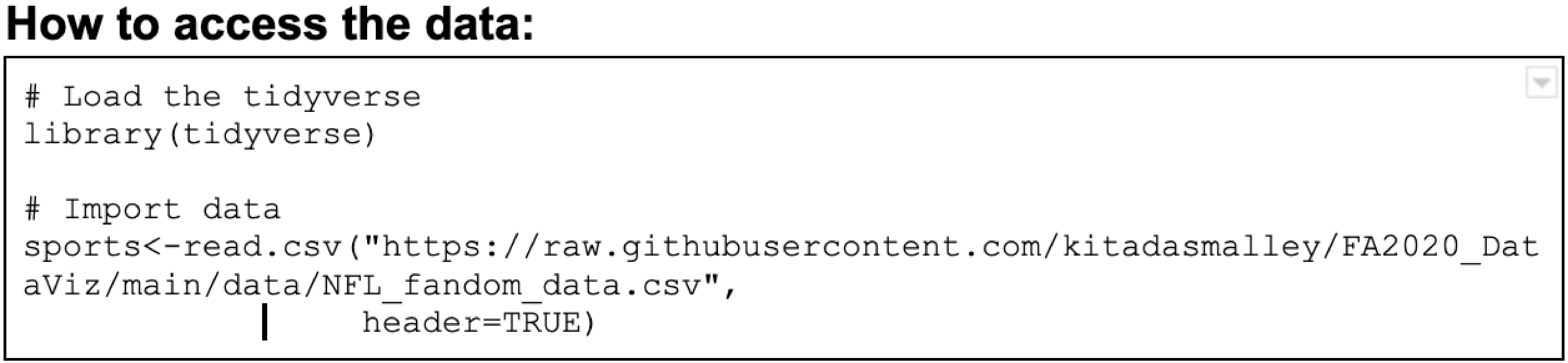


Fan partisanship is based on self-reported party affiliation in a national survey of 2,290 American NFL fans that was conducted Sept. 1-7, 2017. To be affiliated with a team, a respondent had to rank that team among his or her three favorites.

FiveThirtyEight

SOURCES: SURVEYMONKEY AUDIENCE, ECHELON INSIGHTS

WHAT DOES THE RAW DATA LOOK LIKE?

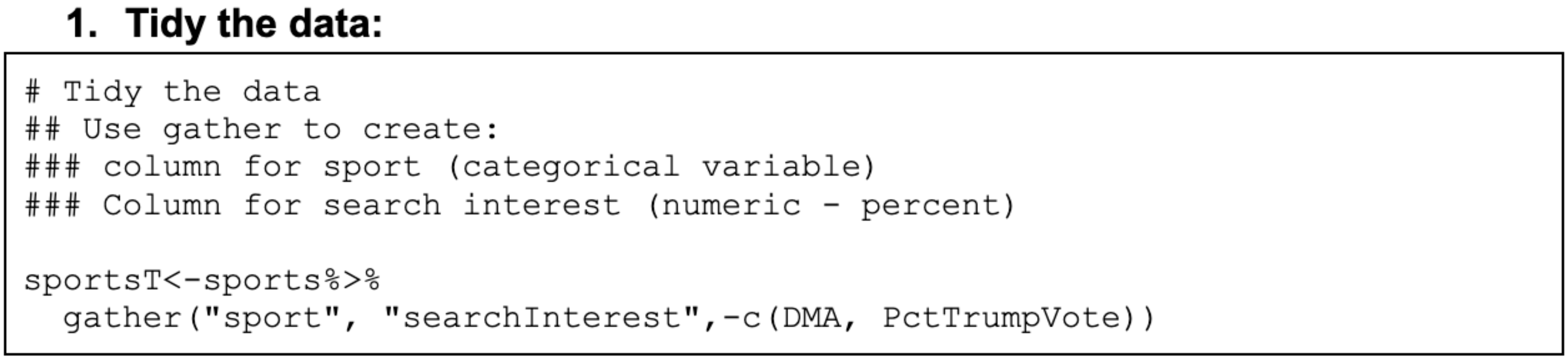


How to access the data:

```
# Load the tidyverse
library(tidyverse)

# Import data
sports<-read.csv("https://raw.githubusercontent.com/kitadasmalley/FA2020_Dat
aViz/main/data/NFL_fandom_data.csv",
                 header=TRUE)
```

ARE WE GOING TO NEED TO TIDY THE DATA?

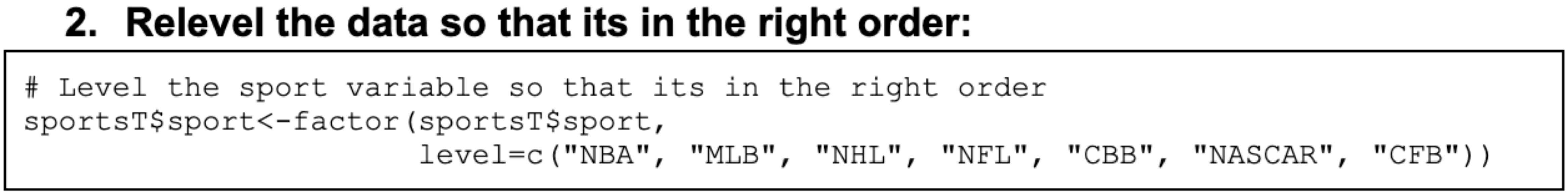


1. **Tidy the data:**

```
# Tidy the data
## Use gather to create:
### column for sport (categorical variable)
### Column for search interest (numeric - percent)

sportsT<-sports%>%
  gather("sport", "searchInterest",-c(DMA, PctTrumpVote))
```

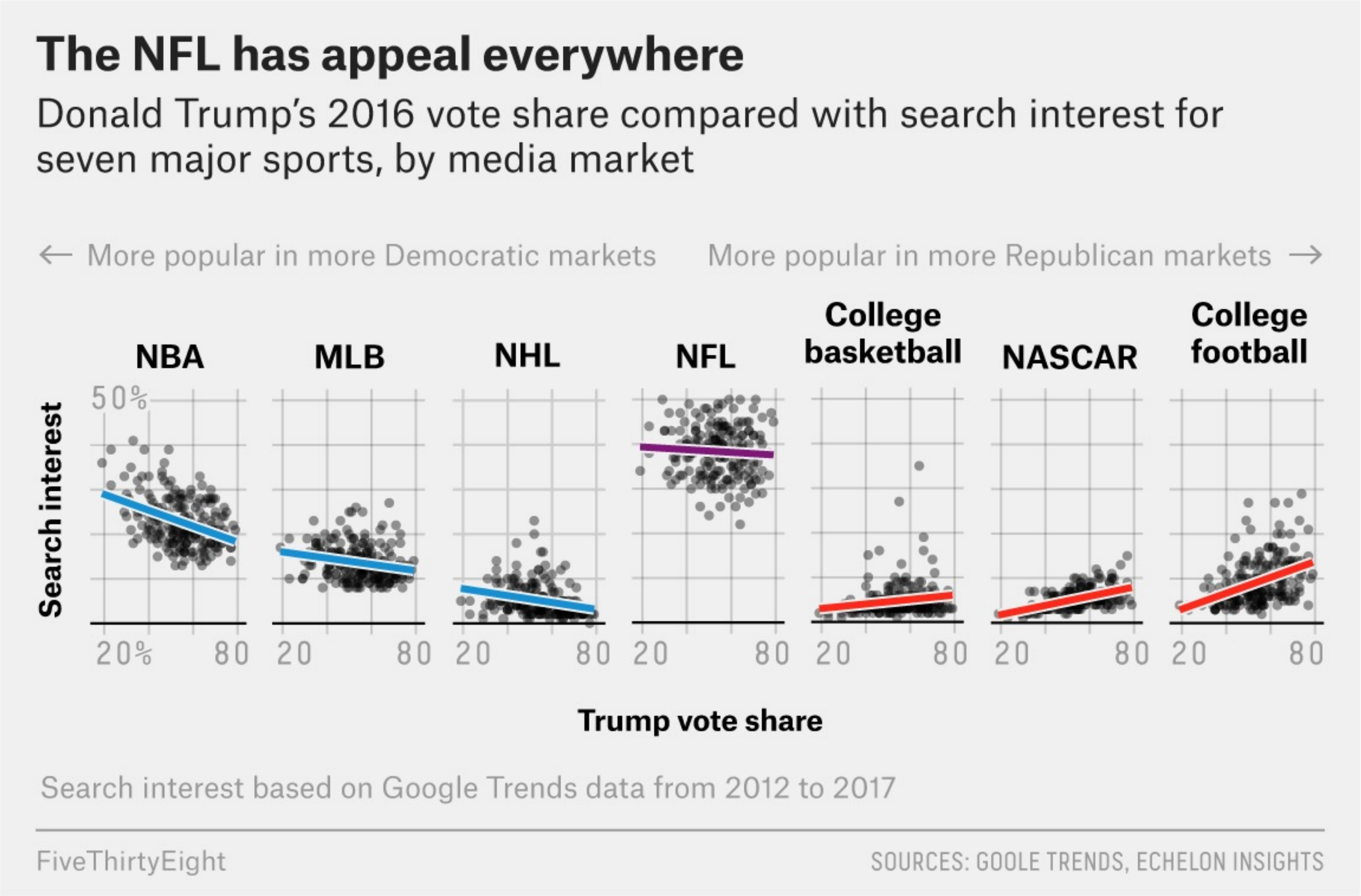
WE MIGHT WANT TO RELEVEL THE SPORTS



2. **Relevel the data so that its in the right order:**

```
# Level the sport variable so that its in the right order
sportsT$sport<-factor(sportsT$sport,
                      level=c("NBA", "MLB", "NHL", "NFL", "CBB", "NASCAR", "CFB"))
```

RECREATE THIS GRAPH IN SMALL GROUPS



The NFL has appeal everywhere

Donald Trump's 2016 vote share compared with search interest for seven major sports, by media market

Search interest based on Google Trends data from 2012 to 2017

FiveThirtyEight

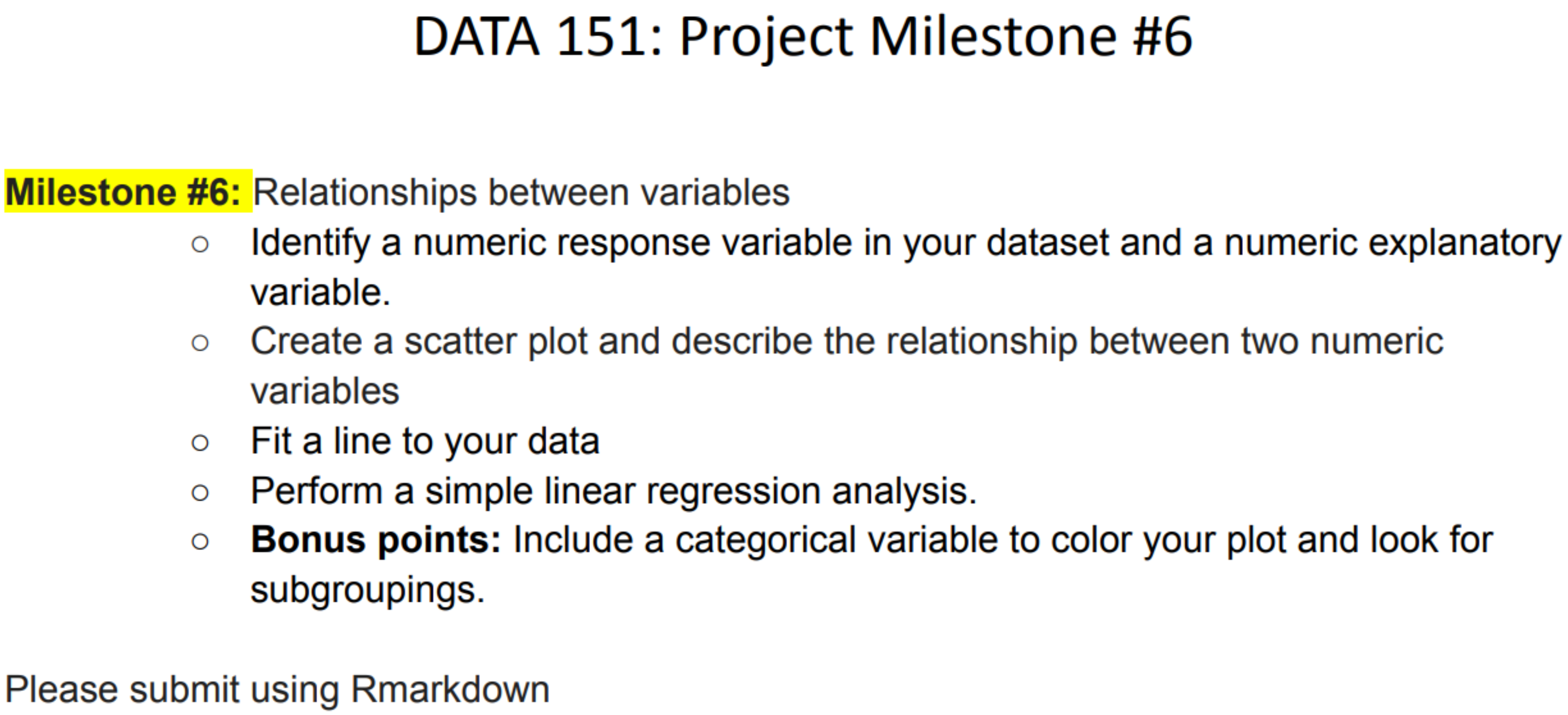
SOURCES: GOOLE TRENDS, ECHELON INSIGHTS

Task: Using the tools we have covered so far, recreate this graph.

Bonus Challenge: *Change the color of the lines.*

TIME FOR GROUP WORK

MILESTONE #6



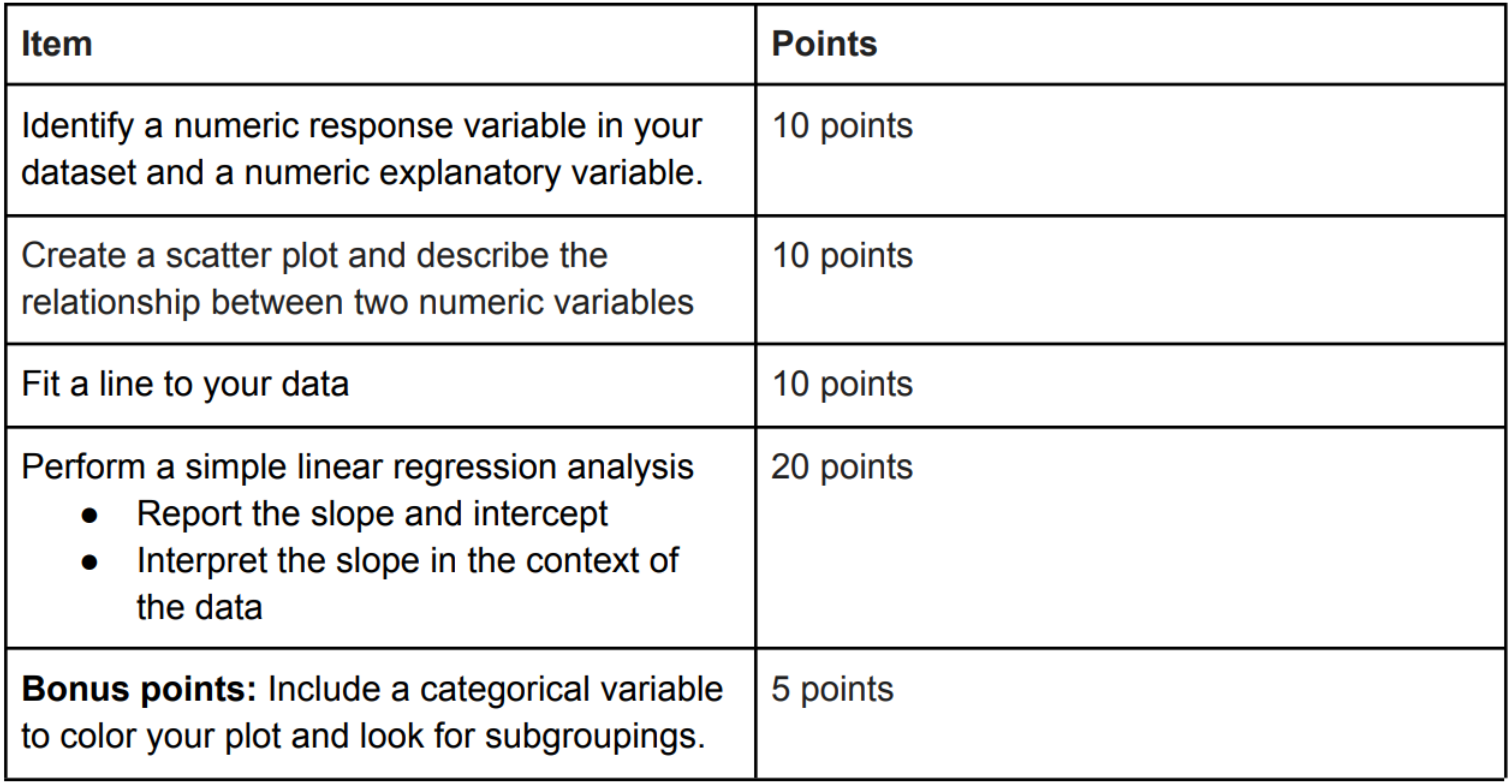
DATA 151: Project Milestone #6

Milestone #6: Relationships between variables

- Identify a numeric response variable in your dataset and a numeric explanatory variable.
- Create a scatter plot and describe the relationship between two numeric variables
- Fit a line to your data
- Perform a simple linear regression analysis.
- **Bonus points:** Include a categorical variable to color your plot and look for subgroupings.

Please submit using Rmarkdown

MILESTONE #6



Item	**Points**
Identify a numeric response variable in your dataset and a numeric explanatory variable.	10 points
Create a scatter plot and describe the relationship between two numeric variables	10 points
Fit a line to your data	10 points
Perform a simple linear regression analysis ● Report the slope and intercept ● Interpret the slope in the context of the data	20 points
Bonus points: Include a categorical variable to color your plot and look for subgroupings.	5 points