Lab 2: What Makes a Product Successful? Fall 2021

w203: Statistics for Data Science

November 1, 2021

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
```

```
# movies <- read.csv("datasets/movies.csv")
#
# cast_gender <- read.csv("datasets/cast_gender.csv")
#
# movies_cast_gender <- inner_join(movies, cast_gender, by = c("star" = "name"))
#
# movies_cast_gender$genre_ <- as.numeric(factor(movies_cast_gender$genre))
# movies_cast_gender$rating_ <- as.numeric(factor(movies_cast_gender$rating))
# movies_cast_gender$director_ <- as.numeric(factor(movies_cast_gender$director))
# movies_cast_gender$writer_ <- as.numeric(factor(movies_cast_gender$writer))
# movies_cast_gender$star_ <- as.numeric(factor(movies_cast_gender$star))
# movies_cast_gender$country_ <- as.numeric(factor(movies_cast_gender$country))
# movies_cast_gender$company_ <- as.numeric(factor(movies_cast_gender$company))
#
# movies_cast_gender <- movies_cast_gender %>%
#   mutate(
#     net_profit = gross - budget
#   )
#
# movies_cast_gender <- movies_cast_gender %>%
#   mutate(male_indicator = case_when(
#     gender == 1 ~ 0,
#     gender == 2 ~ 1
#   )) %>%
#   mutate(female_indicator = case_when(
#     gender == 1 ~ 1,
#     gender == 2 ~ 0
#   ))
```

```
#
# write.csv(movies_cast_gender,"datasets/movies_cast_gender.csv", row.names = FALSE)

movies_cast_gender <- read.csv("datasets/movies_cast_gender.csv")

NROW(movies_cast_gender)
```

```
## [1] 7075
```

```
movies_cast_gender <- movies_cast_gender %>%
  mutate(gender_str = case_when(
    gender == 1 ~ "female",
    gender == 2 ~ "male"
  ))

movies_dataset <- movies_cast_gender[, c("score", "year","votes","budget","gross", "net_profit",
                                         "runtime", "gender", "genre_", "rating_", "director_",
                                         "writer_", "star_", "country_", "company_",
                                         "female_indicator")]


NROW(movies_dataset)
```

```
## [1] 7075
```

```
movies_dataset <- movies_dataset[!is.na(movies_dataset$score), ]
movies_dataset <- movies_dataset[!is.na(movies_dataset$year), ]
movies_dataset <- movies_dataset[!is.na(movies_dataset$votes), ]
movies_dataset <- movies_dataset[!is.na(movies_dataset$budget), ]
movies_dataset <- movies_dataset[!is.na(movies_dataset$gross), ]
movies_dataset <- movies_dataset[!is.na(movies_dataset$runtime), ]

NROW(movies_dataset)
```

```
## [1] 5175
```

```
movies_cor <- cor(movies_dataset)
movies_cor
```

```
##                       score         year        votes       budget
## score             1.000000000  0.052701379  0.482267487  0.074868106
## year              0.052701379  1.000000000  0.207067465  0.333253526
## votes             0.482267487  0.207067465  1.000000000  0.436971393
## budget            0.074868106  0.333253526  0.436971393  1.000000000
## gross             0.225250156  0.276494187  0.612559165  0.741543303
## net_profit        0.245795963  0.238827958  0.607626449  0.612845180
## runtime           0.417633819  0.067490269  0.359090442  0.321644146
## gender            0.097190194 -0.117028439  0.101485520  0.115517729
## genre_            0.036634757 -0.068859335 -0.135197297 -0.369615029
## rating_           0.070048006  0.020881717  0.010826353 -0.194808655
## director_         0.004779081 -0.040417066 -0.011458382 -0.009605375
## writer_           0.017041159 -0.027503391 -0.004414592 -0.044568979
## star_             0.007520196 -0.034741091 -0.018504572 -0.020655946
## country_         -0.035529424 -0.062108227  0.039861503  0.051335191
## company_          0.022694348 -0.008283255  0.116238972  0.167735111
## female_indicator -0.097190194  0.117028439 -0.101485520 -0.115517729
```

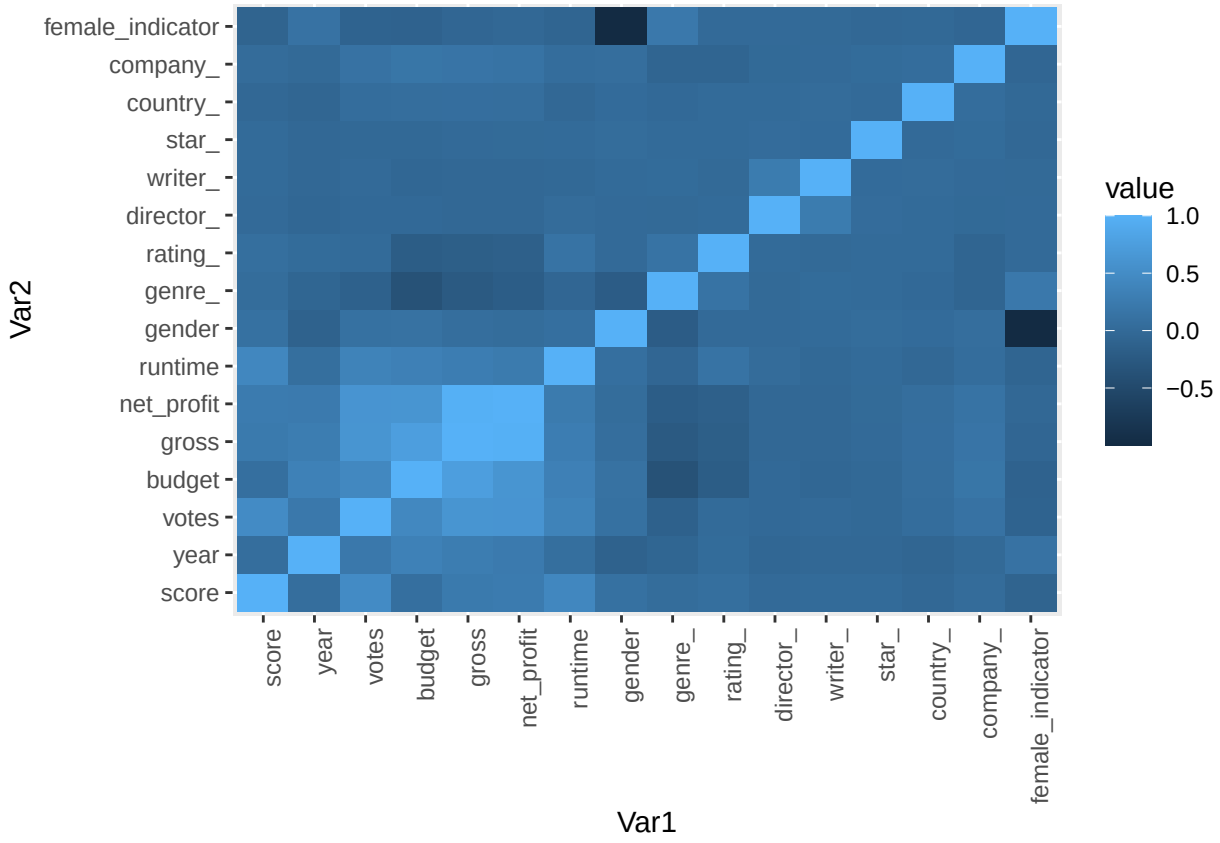
```
##                          gross   net_profit     runtime        gender
## score             0.2252501561  0.245795963  0.41763382  9.719019e-02
## year              0.2764941873  0.238827958  0.06749027 -1.170284e-01
## votes             0.6125591652  0.607626449  0.35909044  1.014855e-01
## budget            0.7415433031  0.612845180  0.32164415  1.155177e-01
## gross             1.0000000000  0.984602391  0.27893311  5.671136e-02
## net_profit        0.9846023908  1.000000000  0.24472547  3.669665e-02
## runtime           0.2789331150  0.244725470  1.00000000  7.716889e-02
## gender            0.0567113607  0.036696651  0.07716889  1.000000e+00
## genre_           -0.2444665223 -0.191631009 -0.05696529 -2.050282e-01
## rating_          -0.1733237408 -0.153384851  0.13563350  9.875482e-05
## director_        -0.0306488913 -0.033596025  0.02391512  3.384442e-03
## writer_          -0.0377319829 -0.032828604 -0.01678494  8.018437e-03
## star_             0.0005835714  0.006069387  0.01355148  3.760889e-02
## country_          0.0620267976  0.059680455 -0.03082369  1.730995e-02
## company_          0.1494471369  0.132316787  0.04727637  5.623046e-02
## female_indicator -0.0567113607 -0.036696651 -0.07716889 -1.000000e+00
##                         genre_       rating_    director_      writer_
## score             0.036634757  7.004801e-02  0.004779081  0.017041159
## year             -0.068859335  2.088172e-02 -0.040417066 -0.027503391
## votes            -0.135197297  1.082635e-02 -0.011458382 -0.004414592
## budget           -0.369615029 -1.948087e-01 -0.009605375 -0.044568979
## gross            -0.244466522 -1.733237e-01 -0.030648891 -0.037731983
## net_profit       -0.191631009 -1.533849e-01 -0.033596025 -0.032828604
## runtime          -0.056965288  1.356335e-01  0.023915118 -0.016784937
## gender           -0.205028215  9.875482e-05  0.003384442  0.008018437
## genre_            1.000000000  1.353537e-01 -0.007119397  0.020357760
## rating_           0.135353664  1.000000e+00  0.007751905 -0.007340871
## director_        -0.007119397  7.751905e-03  1.000000000  0.254616989
## writer_           0.020357760 -7.340871e-03  0.254616989  1.000000000
## star_             0.007873704  1.014168e-02  0.032584061  0.018594915
## country_         -0.012036691  9.780340e-03  0.011473200  0.023483570
## company_         -0.074327662 -8.182124e-02 -0.009006623 -0.001559621
## female_indicator  0.205028215 -9.875482e-05 -0.003384442 -0.008018437
##                          star_      country_     company_ female_indicator
## score             0.0075201963 -0.0355294243  0.022694348    -9.719019e-02
## year             -0.0347410911 -0.0621082269 -0.008283255     1.170284e-01
## votes            -0.0185045724  0.0398615032  0.116238972    -1.014855e-01
## budget           -0.0206559460  0.0513351908  0.167735111    -1.155177e-01
## gross             0.0005835714  0.0620267976  0.149447137    -5.671136e-02
## net_profit        0.0060693866  0.0596804553  0.132316787    -3.669665e-02
## runtime           0.0135514811 -0.0308236946  0.047276367    -7.716889e-02
## gender            0.0376088903  0.0173099477  0.056230462    -1.000000e+00
## genre_            0.0078737042 -0.0120366908 -0.074327662     2.050282e-01
## rating_           0.0101416778  0.0097803402 -0.081821236    -9.875482e-05
## director_         0.0325840613  0.0114732005 -0.009006623    -3.384442e-03
## writer_           0.0185949145  0.0234835697 -0.001559621    -8.018437e-03
## star_             1.0000000000 -0.0004477874  0.020811176    -3.760889e-02
## country_         -0.0004477874  1.0000000000  0.044770141    -1.730995e-02
## company_          0.0208111762  0.0447701414  1.000000000    -5.623046e-02
## female_indicator -0.0376088903 -0.0173099477 -0.056230462     1.000000e+00
```

```
library(reshape2)
```

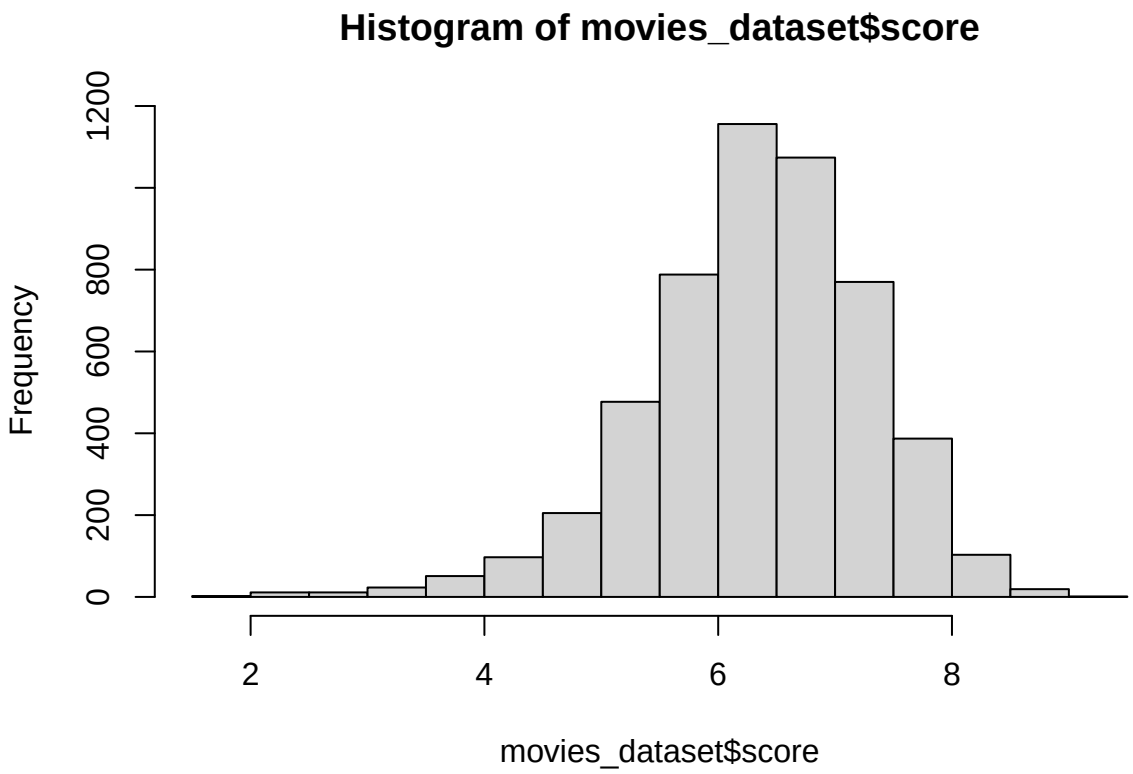
```
##
```

```
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

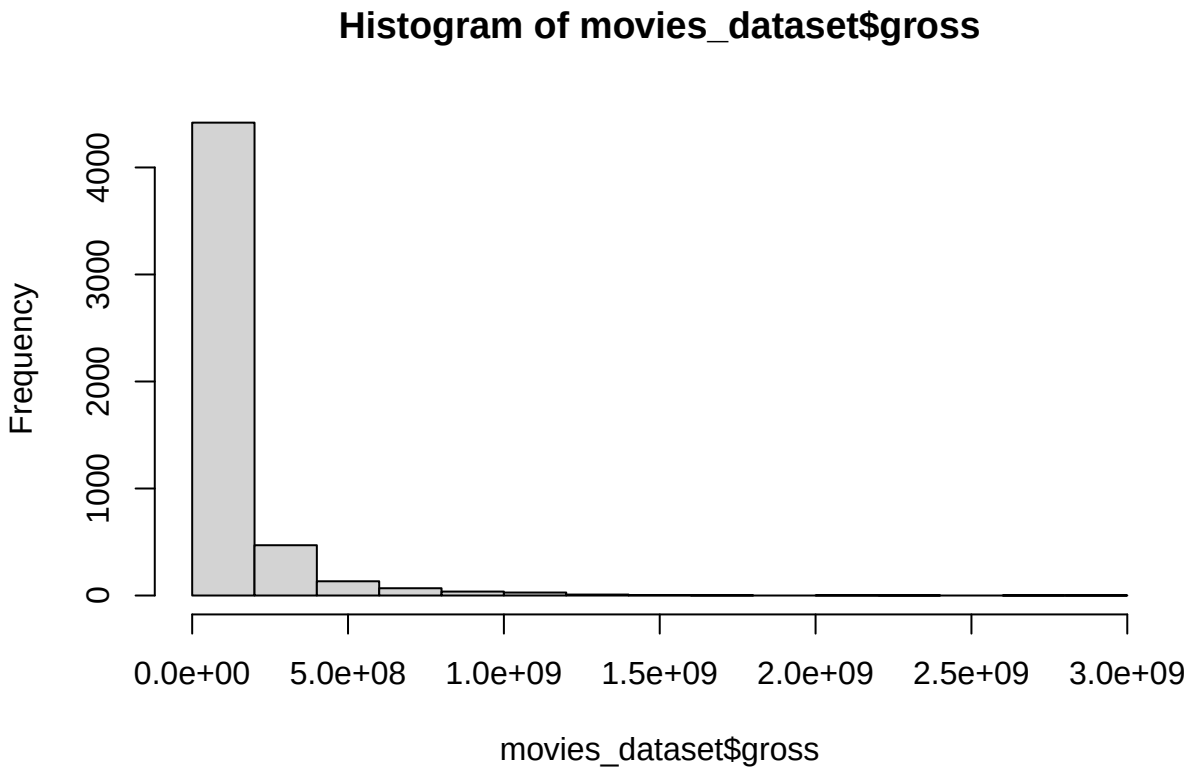
```
melted_movies_cor <- melt(movies_cor)

library(ggplot2)
ggplot(data = melted_movies_cor, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() + theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1))
```



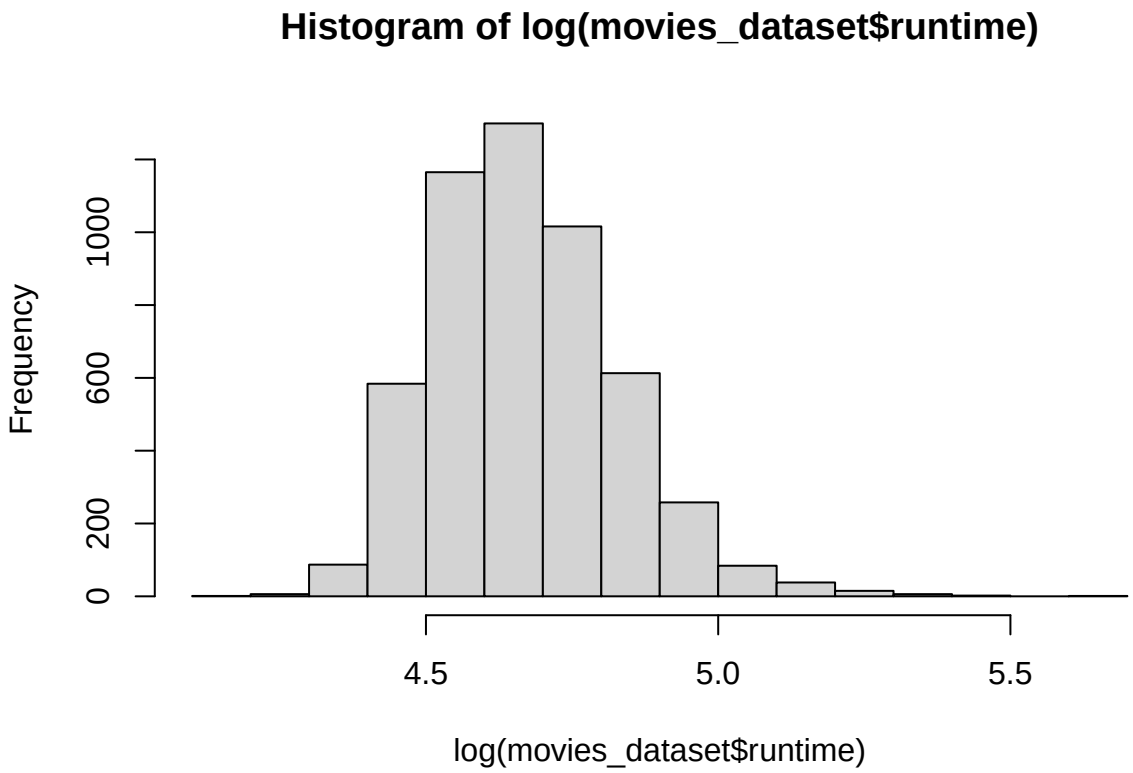

```
hist(movies_dataset$score)
```



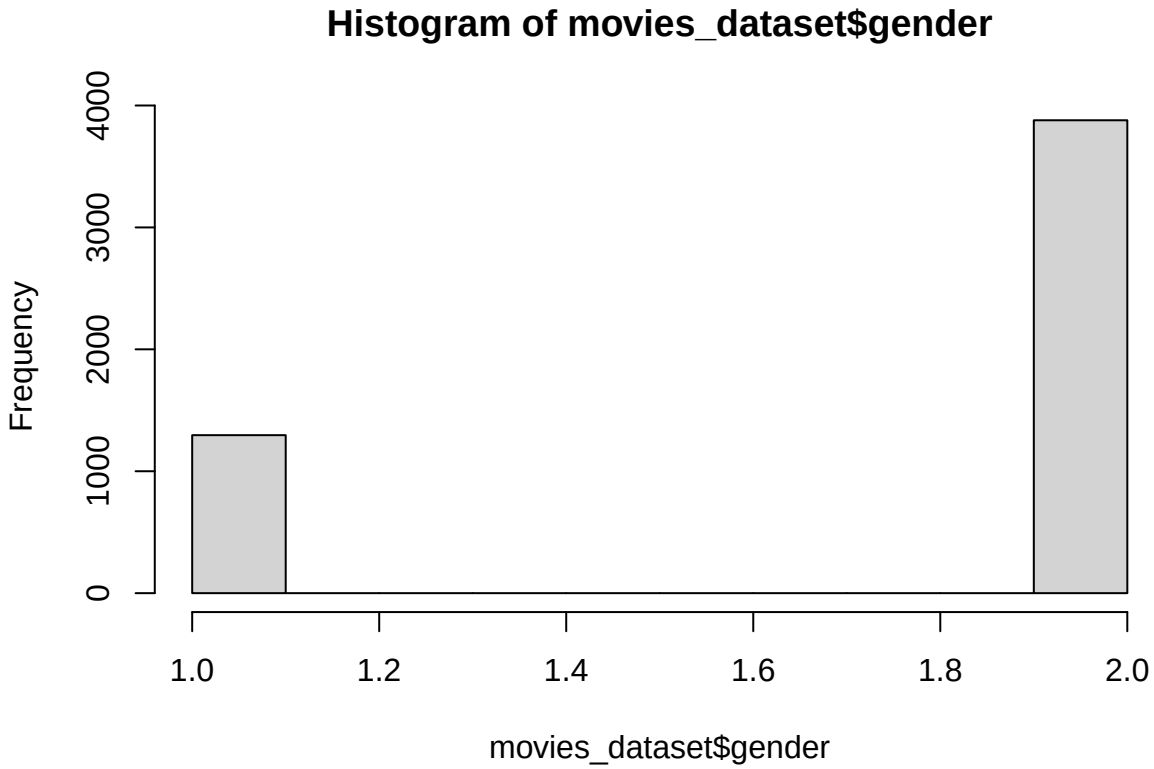

```
# hist(movies_dataset$votes)
# hist(movies_dataset$budget)
hist(movies_dataset$gross)
```



```
# hist(movies_dataset$net_profit)
hist(log(movies_dataset$runtime))
```




```
hist(movies_dataset$gender)
```



```r
# hist(movies_dataset$country_)

movies_cast_gender %>% group_by(gender, gender_str) %>% summarise(count_genders = n()) %>% arrange(desc
```

```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 2 x 3
## # Groups:   gender [2]
##   gender gender_str count_genders
##    <int> <chr>              <int>
## 1      2 male                5177
## 2      1 female              1898
```

```r
movies_cast_gender %>% group_by(genre, genre_) %>% summarise(count_genres = n()) %>% arrange(desc(count_
```

```
## `summarise()` has grouped output by 'genre'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 15 x 3
## # Groups:   genre [15]
##    genre     genre_ count_genres
##    <chr>      <int>        <int>
##  1 Comedy         5         2077
##  2 Action         1         1618
##  3 Drama          7         1364
##  4 Crime          6          509
##  5 Adventure      2          411
##  6 Biography      4          409
##  7 Animation      3          294
##  8 Horror        10          286
##  9 Fantasy        9           43
## 10 Mystery       11           20
## 11 Thriller      14           16
## 12 Romance       12            9
## 13 Sci-Fi        13            9
## 14 Family         8            7
## 15 Western       15            3
```

```r
movies_cast_gender %>% group_by(rating, rating_) %>% summarise(count_rating = n()) %>% arrange(desc(cou
```

```
## `summarise()` has grouped output by 'rating'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 13 x 3
## # Groups:   rating [13]
##    rating      rating_ count_rating
##    <chr>         <int>        <int>
##  1 "R"               8         3460
##  2 "PG-13"           7         2000
##  3 "PG"              6         1170
##  4 "Not Rated"       5          199
##  5 "G"               3          136
##  6 "Unrated"        12           39
##  7 ""                1           38
##  8 "NC-17"           4           20
##  9 "TV-MA"          10            6
```

```
## 10 "X"              13            3
## 11 "TV-PG"          11            2
## 12 "Approved"        2            1
## 13 "TV-14"           9            1
```

```
movies_cast_gender %>% group_by(director, director_) %>% summarise(count_directors = n()) %>% arrange(d
```

```
## `summarise()` has grouped output by 'director'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 2,710 x 3
## # Groups:   director [2,710]
##    director          director_ count_directors
##    <chr>                 <int>           <int>
##  1 Woody Allen            2684              33
##  2 Clint Eastwood          471              29
##  3 Steven Spielberg       2452              26
##  4 Ron Howard             2247              24
##  5 Directors               677              23
##  6 Ridley Scott           2148              23
##  7 Steven Soderbergh      2451              23
##  8 Joel Schumacher        1254              22
##  9 Barry Levinson          195              20
## 10 Tim Burton             2532              19
## # ... with 2,700 more rows
```

```
movies_cast_gender %>% group_by(writer) %>% summarise(count_writers = n()) %>% arrange(desc(count_write
```

```
## # A tibble: 4,188 x 2
##    writer              count_writers
##    <chr>                       <int>
##  1 Woody Allen                    32
##  2 Stephen King                   30
##  3 Luc Besson                     25
##  4 John Hughes                    22
##  5 David Mamet                    15
##  6 William Shakespeare            15
##  7 Joel Coen                      12
##  8 Pedro Almodóvar                12
##  9 Wes Craven                     12
## 10 Leigh Whannell                 11
## # ... with 4,178 more rows
```

```
movies_cast_gender %>% group_by(star) %>% summarise(count_stars = n()) %>% arrange(desc(count_stars))
```

```
## # A tibble: 2,334 x 2
##    star               count_stars
##    <chr>                    <int>
##  1 Nicolas Cage                43
##  2 Robert De Niro              41
##  3 Tom Hanks                   41
##  4 Denzel Washington           37
##  5 Bruce Willis                34
##  6 Tom Cruise                  34
##  7 Johnny Depp                 33
##  8 Sylvester Stallone          32
##  9 John Travolta               31
```

```
## 10 Kevin Costner               29
## # ... with 2,324 more rows
```

```
movies_cast_gender %>% group_by(country) %>% summarise(count_country = n()) %>% arrange(desc(count_coun
```

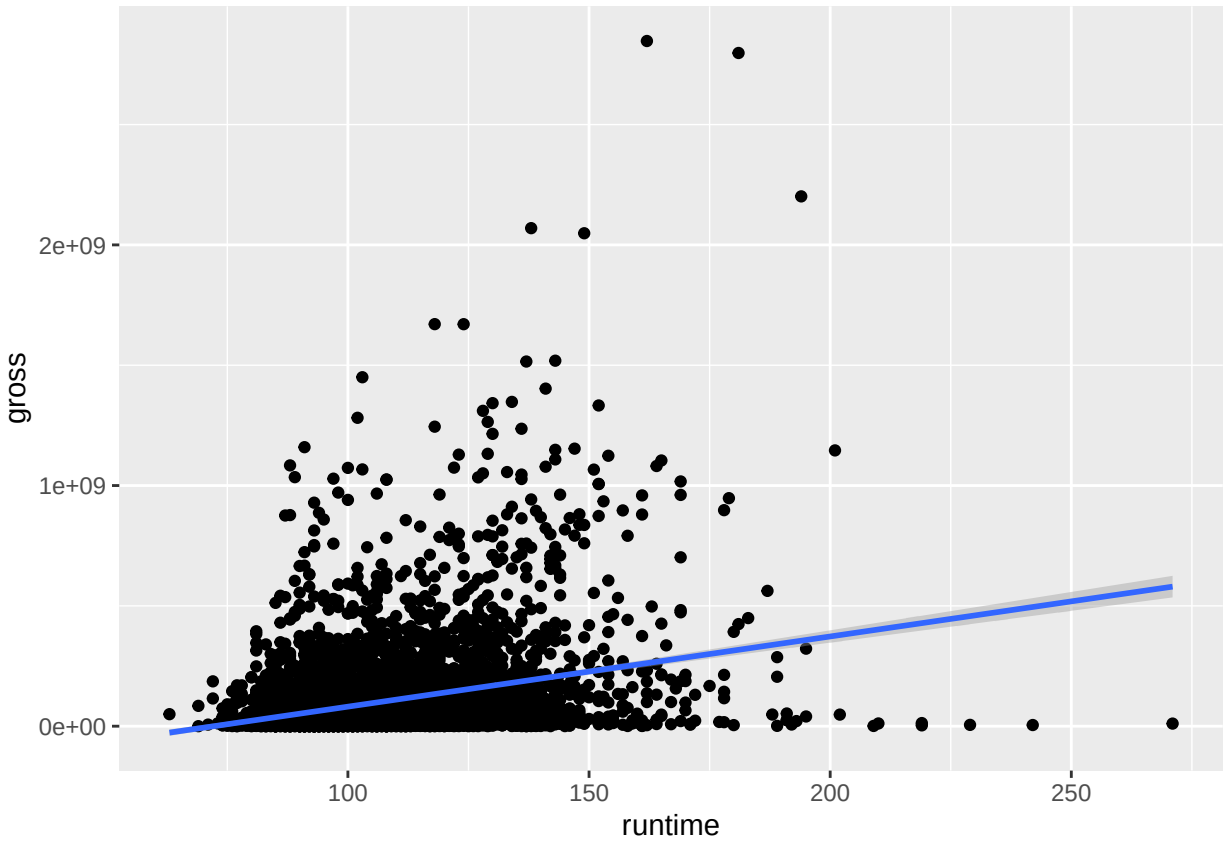
```
## # A tibble: 56 x 2
##    country        count_country
##    <chr>                  <int>
##  1 United States           5181
##  2 United Kingdom           763
##  3 France                   244
##  4 Canada                   171
##  5 Germany                  112
##  6 Australia                 84
##  7 Italy                     55
##  8 India                     54
##  9 Japan                     48
## 10 Spain                     44
## # ... with 46 more rows
```

```
movies_cast_gender %>% group_by(company) %>% summarise(count_company = n()) %>% arrange(desc(count_compa
```

```
## # A tibble: 2,134 x 2
##    company                   count_company
##    <chr>                             <int>
##  1 Universal Pictures                  368
##  2 Warner Bros.                        327
##  3 Columbia Pictures                   318
##  4 Paramount Pictures                  306
##  5 Twentieth Century Fox               235
##  6 New Line Cinema                     163
##  7 Touchstone Pictures                 128
##  8 Walt Disney Pictures                119
##  9 Metro-Goldwyn-Mayer (MGM)           114
## 10 TriStar Pictures                     91
## # ... with 2,124 more rows
```

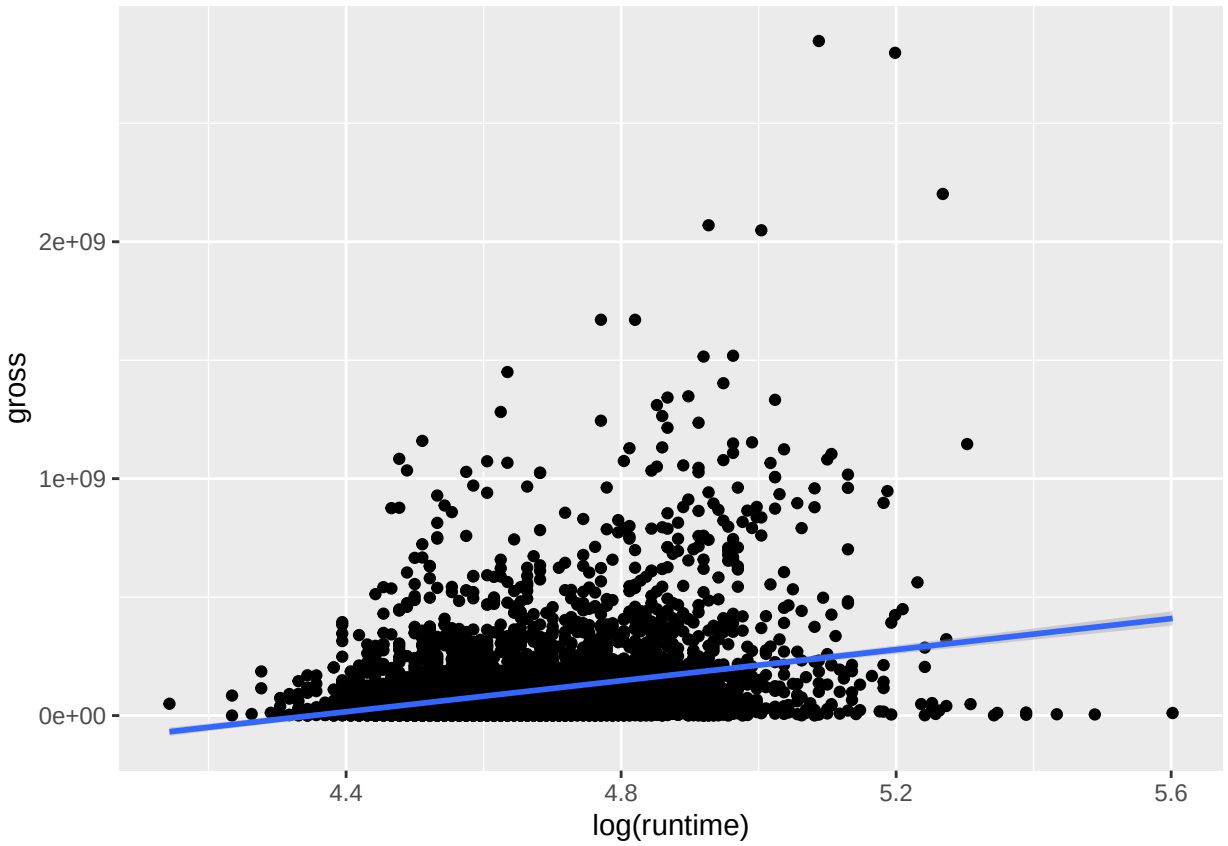
```
ggplot(movies_dataset, aes(x = runtime, y = gross)) +
  geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# movies_dataset_runtime <- movies_dataset[movies_dataset$runtime < 150,]
# ggplot(movies_dataset_runtime, aes(x = runtime, y = gross)) +
#   geom_point() + geom_smooth(method = "lm")

ggplot(movies_dataset, aes(x = log(runtime), y = gross)) +
  geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#https://learn.datascience.berkeley.edu/ap/courses/556/sections/0c4128e8-08ad-402c-a8c3-e8697aba5feb/co
#indicator: female -> gross
#
# lm = lm(gross ~ runtime, movies_dataset)
# summary(lm)

lm = lm(log(gross) ~ runtime, movies_dataset)
summary(lm)
```

```
##
## Call:
## lm(formula = log(gross) ~ runtime, data = movies_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9139  -0.9777   0.1962   1.2827   4.0792
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.459147   0.152404   94.87   <2e-16 ***
## runtime      0.025742   0.001388   18.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.81 on 5173 degrees of freedom
## Multiple R-squared:  0.06231,    Adjusted R-squared:  0.06213
## F-statistic: 343.8 on 1 and 5173 DF,  p-value: < 2.2e-16
```

```r
lm = lm(log(gross) ~ runtime + female_indicator, movies_dataset)
summary(lm)
```

```
##
## Call:
## lm(formula = log(gross) ~ runtime + female_indicator, data = movies_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9488  -0.9776   0.2092   1.2861   4.1548
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.512617   0.154597   93.87   <2e-16 ***
## runtime           0.025523   0.001392   18.33   <2e-16 ***
## female_indicator -0.118790   0.058244   -2.04   0.0414 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.81 on 5172 degrees of freedom
## Multiple R-squared:  0.06307,    Adjusted R-squared:  0.0627
## F-statistic: 174.1 on 2 and 5172 DF,  p-value: < 2.2e-16
```

```r
# ggplot(data=movies_dataset, aes(x=gender, y=score)) +
#     geom_point()+
#     geom_smooth(method='lm', formula=y~x)

lm = lm(score ~ runtime, movies_dataset)
summary(lm)
```

```
##
## Call:
## lm(formula = score ~ runtime, data = movies_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3873 -0.4964  0.0596  0.5763  2.5551
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.0117453  0.0729425   55.00   <2e-16 ***
## runtime     0.0219680  0.0006645   33.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8665 on 5173 degrees of freedom
## Multiple R-squared:  0.1744, Adjusted R-squared:  0.1743
## F-statistic:  1093 on 1 and 5173 DF,  p-value: < 2.2e-16
```

```r
lm = lm(score ~ runtime + female_indicator, movies_dataset)
summary(lm)
```

```
##
## Call:
## lm(formula = score ~ runtime + female_indicator, data = movies_dataset)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4149 -0.4940  0.0575  0.5796  2.5137
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.0764810  0.0738311   55.21  < 2e-16 ***
## runtime            0.0217028  0.0006649   32.64  < 2e-16 ***
## female_indicator -0.1438183  0.0278157   -5.17 2.42e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8644 on 5172 degrees of freedom
## Multiple R-squared:  0.1787, Adjusted R-squared:  0.1783
## F-statistic: 562.5 on 2 and 5172 DF,  p-value: < 2.2e-16
```

Introduction

Imagine that you are part of a team of product data scientists at Acme, Inc. Your manager, Mx. Coy Ote, has given you the freedom to choose your own product to investigate, and evaluate a way to make it more successful.

Your task is to select and develop a research question, find appropriate data, then conduct a regression study.

Research Question

High grossing movies still feature mostly men. https://abcnews.go.com/Business/high-grossing-movies-feature-men-study-finds/story?id=32920292

Our research question: Does high male cast ratio in a movie/TV improve rotten-tomatoes rating?
###

Your research question must be specific, it should clearly state an X and a Y, defined at a conceptual level. Your X should be a design property or characteristic of a product that could be modified in the production process, and your Y should be a metric of success.

In selecting your research question, you will have to use the skills you developed in RDADA to work on a question that can be addressed using a modeling approach from this course. It is not appropriate to ask "What product features increase success?" or "How does product design affect sales?". These types of questions are not amenable to a modeling based approach and your study would likely become a fishing expedition. Instead, your team will have to use your background knowledge to identify a relationship you want to measure between a specific design feature and a specific metric of success.

If your data set is large enough, you can begin your process by splitting the data into an exploration set and a testing set. As a rough guideline, you might put 30% of your data into the exploration set, but make sure that both sets have a minimum of 200 rows of data. Use the exploration set to build your intuition, explore how the data is distributed, and identify your X and Y variables. Then use the testing set to fit your models.

Because your manager is interested in *changes* to a product, they are fundamentally asking you to perform an explanatory study. As we have noted in the class, given observational data, an OLS regression is usually not a credible way to measure a causal effect. We have purposefully selected a domain in which the one-equation structural model is at least partially defensible. The most prominent causal pathways will go in

one direction, from product design characteristics to success. While not a perfect reflection of reality, we expect your model to be plausible enough to make your results interesting. At the same time, you will need to analyze potential violations of the one-equation structural model and what effect any violations may have on your results.

Data

For this lab, you and your team will be responsible for gathering the data that you use. The data should be publicly available, and should be relevant to your research question. To increase the diversity of products investigated, we are asking students to avoid working on data that is sourced from Yelp and Airbnb. There are very, very good data resources available, for example:

- New York Times
- Tidy Tuesday
- ICPSR for social and political data
- Data.world
- Dataverse for published research data
- UC Irvine Machine Learning Data Repository
- Google Dataset Search
- Amazon Open Data Registry
- Azure Open Data Registry

Requirements for your data:

- Data should be cross-sectional (i.e. not have multiple measurements for a single unit of observation). You may, for example, take a single cross section from a larger panel.
- We recommend a minimum of 100 or 200 observations. A team could choose to use an interesting dataset that is smaller than this, however, this will then require the team to assess and satisfy the more stringent CLM assumptions.
- The outcome (or outcomes) that you use should be plausibly metric (i.e. number of sales of a product; number of views of a video). For this lab however, to make it easier to find data, teams may use an ordinal outcome variable if necessary. If using an ordinal outcome such as a 1-7 Likert scale, the team should clearly discuss the consequences of failing to satisfy the assumptions of the OLS regression model.
- For any omitted variable that would call your results into question, the data should include the best possible variable that operationalizes this concept. At a minimum, the data should have a variable that serves as an imperfect measure - or *proxy* - for the omitted variable.
- Your models must include a mixture of numeric and categorical inputs (this requirement is for learning purposes). If it is appropriate, you may bin a metric variable into a categorical variable.

You may draw different variables from different data sources. You may use data sources not on the above list. You must document any data source you use in your report.

Example of a Research Question

Suppose that your team is interested in learning how the length of lanyard attached to a catapult affects customers' satisfaction with the catapult. (A classic question from Roadrunner cartoons.)

You work to develop a primary outcome: proportion of boulders that land on their target.

On Acme's servers, you find data on lanyard length, maximum-rated weight for the catapult and sales region. However, when you are reasoning about the product, you also note that length of the catapult arm and size of the catapult wheels are also likely to affect customer satisfaction and are correlated with lanyard length. Because any model that does not include these confounding variables would yield estimates that conflate the importance of wheels and arms with the lanyard, you determine that the off-the-shelf data is not complete and that you need to encode the data yourself.

In the modeling phase of your project, your team proposes to build three models. One model estimates the relationship between targeting accuracy and lanyard length by itself. A second model is similar, but adds a set of covariates including length of catapult arm and size of catapult wheels. Finally, a third model includes an interaction term between lanyard length and customer type (first time or repeat), allowing you to investigate whether the effect of lanyard length is heterogeneous depending on the person operating the catapult.

A Group Assignment

This is a group assignment. Your live session instructor will coordinate the formation of groups. We would like to encourage teams to focus on using the lab as a way to learn how to work as a team of collaborating data scientists on shared code; how to clean and organize data; and, how to present work in a compelling way. As a result, we encourage teams to allow individuals to take risks and be supportive in the face of successes and failures. Create an opportunity for people who want to improve a particular skill to do so – this might be project coordination, management of code through git, plotting, or any of the many aspects that you'll work on. *We hope that you can support and learn from one another through this team-based project.*

Deliverables

Deliverable Name	Week Due	Grade Weight
Research Proposal	Week 12	10%
Within-Team Review	Week 12	5%
Final Presentation	Week 14	10%
Final Report	Week 14	75%

Final Project Components

Research Proposal

After a week of work, the project team will produce a one-page research proposal that defines the teams' research question, data sources and plan of action.

The research question should be informed by an understanding of the data and information that is available. This means that the team will need to pursue at least some preliminary exploratory data analysis. A motivated team might form their research question, and begin to build a functioning data pipeline as an investment in ongoing project success.

The research proposal is intended to provide a structure for the team to have an early conversation with their instructor. It will be graded credit/no credit for completeness (i.e. a reasonable effort by the team will receive full marks). Your instructor will read these proposals and will contact the team with any necessary course corrections, suggestions, or feedback.

This proposal is due in week 12, in Gradescope, with one submission for the whole team.

Within-Team Review

Being an effective, supportive team member is a crucial part of data science work. Your performance in this lab includes the role you play in supporting your teammates. This includes being responsive, creating an environment in which all members feel included, and above all treating each other with respect. In line with this perspective, we will ask each team member to write two paragraphs to their instructor about the progress they have made individually, and the team has made as a whole toward completing their report.

This self-assessment should:

- Reflect on the strengths and weaknesses of the team and the team's process to this point in the project.
  - Where your collaboration has worked well, how will you work to ensure that these successful practices continue to be employed?
  - If there are places where collaboration has been challenging, what can the team do jointly to improve?
- If there are any individual performances that deserve special recognition, please let your instructor know in this evaluation.
- If there are any individual performances that require special attention, please also let your instructor know in this evaluation.

Instructors will treat these reviews as confidential and will not take any action without first consulting you.

This reflection is due in week 12, in Gradescope and requires one submission per person. You will submit this through Gradescope, and like all parts of your educational record, this will be treated confidentially by the instructional team.

Final Presentation

During the Unit 14 live session, each team will give a presentation of their work to their classmates, who will be seated with you as collaborating data scientists. As collaborating data scientists, your classmates will need to be informed of the specific product and research question that you are addressing.

Presentation Guidelines

- **The presentation should be structured as 10 minutes of presentation and 5 minutes of questions from our classmates.** Please note that this is an *incredibly* limited amount of time to present.
- There should be no more than two slides that set-up your research question and these slides should take no more than two minutes to present. On this slide, it is quite alright to state bluntly: "**Research Question**: Do shorter lanyards increase the accuracy of catapult launches?" (2 minutes)
- You should ground the audience in an understanding of the data that you are using in your models. Take a short amount of time to describe key attributes of the variables that you are including in your model. This should minimally include a description of the outcome and key explanatory feature, but should also include any other variables or context that is necessary to reason about your model and results. (2-3 minutes)
- Do not present R code, discuss data wrangling, or normality - details like this are best left to the full analysis. It is tempting to want to share these process based stories with your peers, but save that time for after the presentation.
- There should then be several slides that present what you've learned from your models. It is a good practice to show your final regression table on a slide by itself. If you show a regression table, you need to provide your audience with enough time to read and engage with it. As a general rule, any model table you show will take at least two minutes to discuss. For any table (or plot) that you show, you should minimally interpret the variables (or axes) and the key point that you are making with that piece of evidence.

Finally, a few more general thoughts:

- Practice your talk with a timer!
- If you divide your talk with your teammates, practice your section with a timer so that you do not spill over into your teammates' time.
- There is no need to say, "Now I am going to hand it off to Becca." And for Becca to say, "Thank you Adam." Whoever's turn it is to talk can simply move the presentation forward.

Final Report

Your final deliverable is a written statistical analysis documenting your findings. **Please limit your submission to 6,000 words, excluding code cells and R output.**

The exact format of your report is flexible, but it should include the following elements.

1. An Introduction

Your introduction should present a research question and explain the concept that you're attempting to measure and how it will be operationalized. This section should pave the way for the body of the report, preparing the reader to understand why the models are constructed the way that they are. It is not enough to simply say, "We are looking for product features that enhance product success." Your introduction must do work for you, focusing the reader on a specific measurement goal, making them care about it, and propelling the narrative forward. This is also a good time to put your work into context, discuss cross-cutting issues, and assess the overall appropriateness of the data.

2. A description of the Data and Research Design

After you have presented the introduction and the concepts that are under investigation, what data are you going to use to answer the questions? What type of research design are you using? What type of models are you going to estimate, and what goals do you have for these models?

2a. A Model Building Process

You will next build a set of models to investigate your research question, documenting your decisions. Here are some things to keep in mind during your model building process:

1. *What do you want to measure*? Make sure you identify one, or a few, variables that will allow you to derive conclusions relevant to your research question, and include those variables in all model specifications. How are the variables that you will be modeling distributed? Provide enough context and information about your data for your audience to understand whatever model results you will eventually present.
2. What covariates help you achieve your modeling goals? Are there problematic covariates? either due to *collinearity*, or because they will absorb some of a causal effect you want to measure?
3. What *transformations*, if any, should you apply to each variable? These transformations might reveal linearities in the data, make our results relevant, or help us meet model assumptions.
4. Are your choices supported by exploratory data analysis (*EDA*)? You will likely start with some general EDA to *detect anomalies* (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to *guide* your decisions. You can also leverage statistical *tests* to help assess whether variables, or groups of variables, are improving model fit.

At the same time, it is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust (or sensitive) your results are to modeling choices, and to show that you're not just cherry-picking the specification that leads to the largest effects.

At a minimum, you need to estimate at least three model specifications:

The first model you include should include *only the key variables* you want to measure. These variables might be transformed, as determined by your EDA, but the model should include the absolute minimum number of covariates (usually zero or one covariate that is so crucial it would be unreasonable to omit it).

Additional models should each be defensible, and should continue to tell the story of how product features contribute to product success. This might mean including additional right-hand side features to remove omitted variable bias identified by your casual theory; or, instead, it might mean estimating a model that examines a related concept of success, or a model that investigates a heterogeneous effect. These models, and your modeling process should be defensible, incremental, and clearly explained at all points.

Your goal is to choose models that encircle the space of reasonable modeling choices, and to give an overall understanding of how these choices impact results.

4. A Results Section

You should display all of your model specifications in a regression table, using a package like `stargazer` to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table.

In your text, comment on both *statistical significance and practical significance*. You may want to include statistical tests besides the standard t-tests for regression coefficients. Here, it is important that you make clear to your audience the practical significance of any model results. How should the product change as a result of what you have discovered? Are there limits to how much change you are proposing? What are the most important results that you have discovered, and what are the least important?

5. Limitations of your Model

5a. Statistical limitations of your model As a team, evaluate all of the large sample model assumptions. However, you do not necessarily want to discuss every assumption in your report. Instead, highlight any assumption that might pose significant problems for your analysis. For any violations that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies.

Note that you may need to change your model specifications in response to violations of the large sample model.

5b. Structural limitations of your model What are the most important *omitted variables* that you were not able to measure and include in your analysis? For each variable you name, you should *reason about the direction of bias* caused by omitting this variable and whether the omission of this variable calls into question the core results you are reporting. What data could you collect that would resolve any omitted variables bias?

7. Conclusion

Make sure that you end your report with a discussion that distills key insights from your estimates and addresses your research question.

Encouragement for the Project

This project touches on many of the skills that you have developed in the course.

- When you are reasoning about the world and the way that it works, you are implicitly reasoning about *random variables*. Although you might not reason with specific functions (e.g. $f_x(x) = x^2$) to describe these random variables, you are very likely to be reasoning about conditional expectations.
- This class is not a class in pure theory! And so, theories you have about the world need to be informed by samples of data. These samples might be iid, or they might not be. The team will have to assess how this, and other possible violations of model assumptions shape what they learn.
- Given a set of input variables, OLS regression produces an estimate of the BLP. But, how good of a predictor is this predictor? And, does the team have enough data to rely on large-sample theory, or does the team need to engage with the requirements of the smaller-sample?
- Throughout, you will have to communicate both to a technical and non-technical audience.

Finally, have fun with this project. You have worked hard this semester to build a foundation for reasoning about the world through statistical models. This project is a chance for you and a team of peers to work to apply this reasoning.