Final Project: Are First-borns Superior?

Author: Jenny Pan

Introduction

Are first-born siblings smarter, better, or cooler than their younger ones? In my family, this is a question that is often discussed around the dinner table. As a younger sibling, I passionately resent and disagree with this statement (and of course, my parents would agree!), but what does a more rigorous analysis suggest? On one hand, research suggests that first-born children enjoy more parental attention and investment. Anecdotes from parents corroborates this fact. Does enjoying more parental attention mean that first-born children are more successful than their younger ones? On the other hand, we cannot forget that first-born children are also the children to first-time parents. Younger siblings have the advantage of more experienced, mature, and possible more financially-stable parents. From these two perspectives, the jury's still out on whether the first-born is the better sibling. Is it better to have more doting parents, or more experienced parents? Does it not matter? However, through analyzing online personality surveys, we can get a better insight on these pressing questions.

The data is from the Firstborn Personality Test from the Open-Source Psychometrics Project. In the personality test, participants were asked a series of 25 questions relating to intellect (questions mostly focus on intellect), openness to experience, extroversion, emotional stability, agreeableness, and conscientiousness. The questions that were included in the study are listed in the link I provided. Each question was on a 5 point likert scale, and the greater the total score, the more likely you are to have a "firstborn personality".

As a preface, the survey is not making a qualitative judgement on which traits are associated "better". What personality traits are considered "better" is a matter of personal opinion. They are simply noting whether there is a difference in personality between first-borns and younger siblings. The source of my data did do a preliminary analysis and data visualization using the data, but not more in depth methods. I haven't seen somebody else do an analysis with this dataset in particular, but perhaps I wasn't looking hard enough. Despite this, this topic (whether birth order has an effect on personality) is of interest to a lot of researchers. There are dozens of journal articles online investigating the same topic, however, they did not use this dataset.

Results

Data wrangling: Mutating to get new columns for analysis

```
library(dplyr)
# making a dummy variable. 0 represents non first-born
# children, 1 represents first-born children
fbps <- read.csv("FBPS_ValidationData.csv")

fbps_1 <- fbps %>%
```

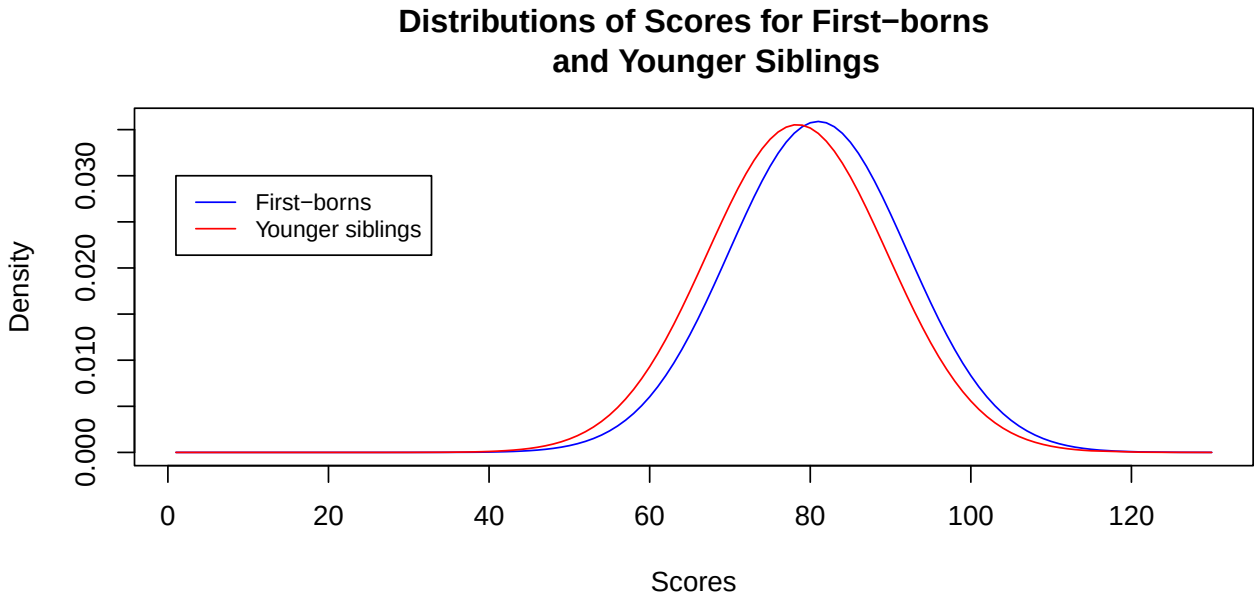
```
    mutate(first_born = ifelse(fbps$birthpos == 1, 1, 0), total_score = Q1 +
        Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9 + Q10 + Q11 + Q12 +
        Q13 + Q14 + Q15 + Q16 + Q17 + Q18 + Q19 + Q20 + Q21 +
        Q22 + Q23 + Q24 + Q25 + Q26) %>%
    filter(total_score != 0)

# make the first_born variable as a factor
fbps_1 <- fbps_1 %>%
    mutate(first_born = as.factor(fbps_1$first_born))
```

I downloaded the data as a csv and then imported it into R for analysis. The dataset was mostly already cleaned, but I needed to add a few things for the analysis.There were 89 variables and 41841 cases in the original dataset. The variables included how the participant responded to Q1-Q26, age, if English was their native language, gender, birth position, how many kids in the family, submit time, time spent. There were also answers to questions to a host survey (there variables start with EXT, EST, AG, CSN, OPN) which were irrelevant for my purposes.

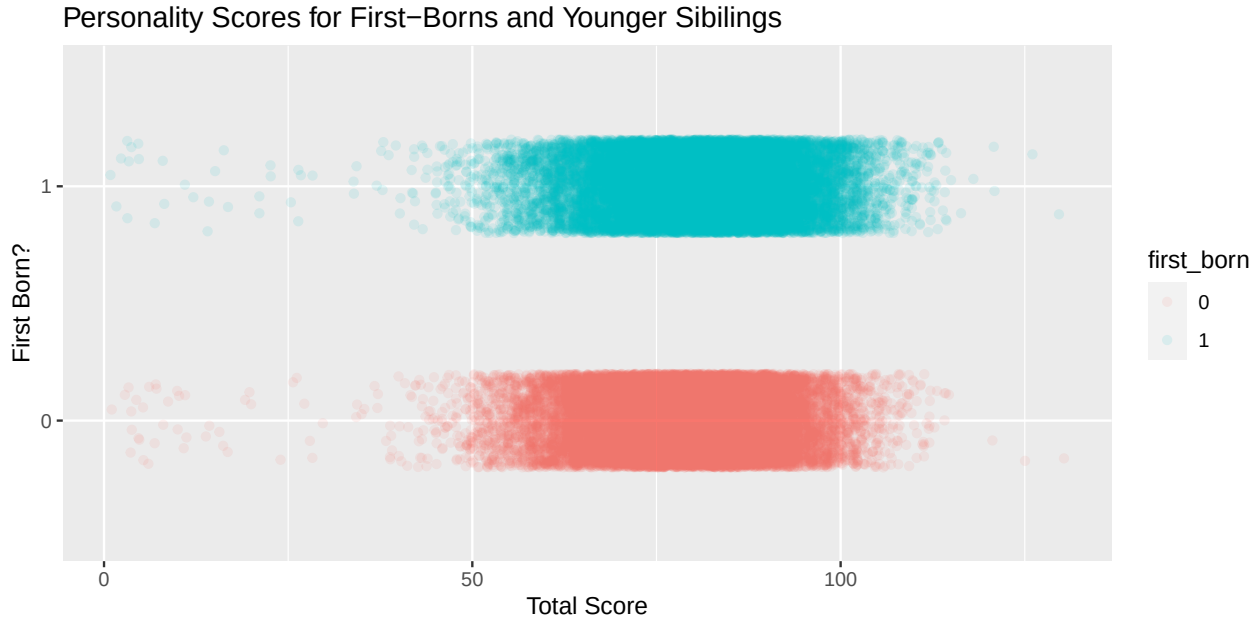
For outliers, I took out cases with a total score of 0 because that means that the particpant never answered any of the questions in the personality test. Their results would not be insightful in addressing questions of interest. In the original dataset, there was a column indicating the participant's birth order (i.e. 1 = first-born, 2 = second-born, etc). Since I want to compare first-borns with younger siblings, I needed to create a dummy variable (first_born) that codes first-borns as a 1, and younger siblings as a 0. I also needed to calculate the total score the particpant received on the test, as this was not already provided for us. Finally, I made my dummy variable (first_born) into a factor so R recognizes it as a categorical variable.

Visualize the data: Density Curves & Jitter Plot

```
library(ggplot2)

# making data frames with just first borns and with just
# younger sibs
first_born_df <- fbps_1 %>%
    filter(first_born == 1)
younger_sibs_df <- fbps_1 %>%
    filter(first_born == 0)

# making the density curve for the the first_born data
x1 <- sort(first_born_df$total_score)
y1 <- dnorm(x1, mean(x1), sd(x1))

# making the density curve for the younger_sibs data
x2 <- sort(younger_sibs_df$total_score)
y2 <- dnorm(x2, mean(x2), sd(x2))

# plotting the density curves on top of each other
plot(x1, y1, type = "l", col = "blue", main = "Distributions of Scores for First-borns
     and Younger Siblings",
    xlab = "Scores", ylab = "Density")
lines(x2, y2, type = "l", col = "red")
legend(1, 0.03, legend = c("First-borns", "Younger siblings"),
    col = c("blue", "red"), lty = 1:1, cex = 0.8)
```




```
# making a jitter plot visualization
ggplot(fbps_1, aes(x = total_score, y = first_born, col = first_born)) +
    geom_jitter(alpha = 0.1, position = position_jitter(height = 0.2)) +
    ggtitle("Personality Scores for First-Borns and Younger Sibilings") +
    xlab("Total Score") + ylab("First Born?")
```



Plotting the histograms of first_born data and younger_sibs data, I noticed that these histograms were approximately normal (see it in the appendix!). I then decided to plot the density curve of the normal distribution from the data. From the distributions of the density curves above, we see that the first-born distribution (in blue) is shifted to the right of the younger siblings distribution (in red). Although the effect is small, it suggests that first-borns may score a higher on the personality test than younger siblings.

From the jitter plot, it's even more clear that the distributions for first-borns and younger siblings are practically the same. The distribution of the first-born is shifted a little to the right compared to the distribution of the younger siblings, suggesting that first borns might be a little more likely to score higher

on this personality test than younger siblings, but the shift is marginal. Comparing the jitter plot to the histogram, I think the jitter plot does a better job at communicating the spread of the data. That's why I decided to include both.

Analyses: Hypothesis Test & Logistic Regression

Hypothesis Test: Is there a difference between the scores of first-borns and younger siblings?

```
# check conditions sample size?
(n1 <- nrow(first_born_df))
```

```
## [1] 20428
```

```
(n2 <- nrow(younger_sibs_df))
```

```
## [1] 21099
```

```
# equal sds?
(sd_1 <- sd(first_born_df$total_score))
```

```
## [1] 11.11647
```

```
(sd_2 <- sd(younger_sibs_df$total_score))
```

```
## [1] 11.2253
```

```
# running a t-test
t.test(total_score ~ first_born, alternative = "two.sided", data = fbps_1)
```

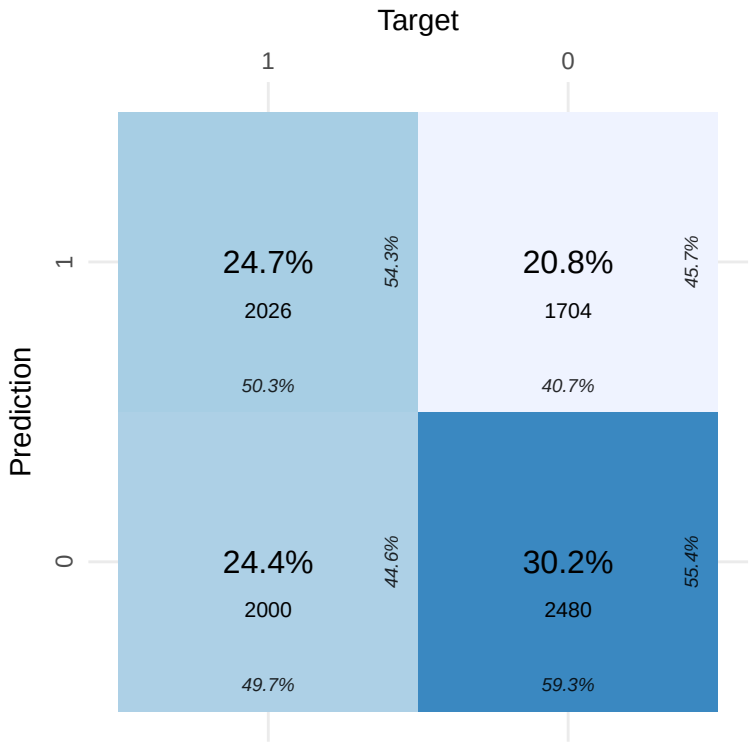
```
##
##  Welch Two Sample t-test
##
## data:  total_score by first_born
## t = -23.792, df = 41504, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.823517 -2.393713
## sample estimates:
## mean in group 0 mean in group 1
##        78.38864        80.99726
```

The assumptions for the t-test is that the data comes from a simple random sample. This might not be true because not everyone has an equal probability of taking the test (people who take it are perhaps interested in the topic, interested in proving a point, referred by a friend), but the each case is certainly independent of one another. Next, the data are normal. Looking at the plot made in the data visualization part, we see that when graphed, the data are normal. Next, we have a large sample size, as each group has well over 20,000 cases. Although Welch's t-test does not assume equal variances, our two groups have approximately equal variances as given by the values of standard deviation above.Even though our data satisfies most of the assumptions, as we learned in class, t-tests are fairly robust to violations in assumptions, so I feel very comfortable using this test.

$H_0 : \mu_{firstborn} - \mu_{youngersibs} = 0$
$H_A : \mu_{firstborn} - \mu_{youngersibs} \neq 0$
The null hypothesis is that there is no difference between the personality scores of first-born children and younger siblings. The alternative hypothesis is that there is a difference between the personality scores of first-born children and younger siblings. Running a two sided t-test, with a p-value of close to 0, I conclude that there is evidence to suggest that there is a difference between the personality scores of first-born children and younger children, with first-born children scores being slightly higher ($\mu_{firstborn} = 80.43420$, $\mu_{youngersiblings} = 77.75844$).

Logistic Regression: Can I predict who's a first-born and who's a younger sibling given their scores for specific questions?

```
library(cvms)
# train on 80% of data and test on 20%
indexSet <- sample(2, nrow(fbps_1), replace = TRUE, prob = c(0.8,
    0.2))
# create training set
train <- fbps_1[indexSet == 1, ]
# create test set
test <- fbps_1[indexSet == 2, ]

# Fit a model and train the model
lr_fit <- glm(first_born ~ total_score, data = train, family = "binomial")
# summary(lr_fit) too long, put in appendix

# Test the model
p <- predict(lr_fit, test, type = "response")

pred <- ifelse(p > 0.5, 1, 0)
pred_table <- table(Predicted = pred, Actual = test$first_born)

cfm <- as_tibble(pred_table)

plot_confusion_matrix(cfm, target_col = "Actual", prediction_col = "Predicted",
    counts_col = "n")
```




```
pred_vals <- as.numeric(pred_table)

# Find the accuracy (# of correct/total #)
(accuracy <- (pred_vals[1] + pred_vals[4])/(nrow(test)))
```

```
## [1] 0.5488429
```

My data meets most of the assumptions of logistic regression. First, the response variable (whether you are first_born or not) only takes on two possible outcomes. Next, the observations/cases are reasonably independent since each user is taking the test themselves. Next, since there is only one predictor/explantory variable, there is no multicolinearlity. Furthermore, there are no extreme outliers since we filtered out total scores on 0. Finally, our sample size is sufficiently large.

Using their total score as a predictor for whether a person is first-born or not gives approximately 55% accuracy. 55% is slightly higher than chance, reinforcing the results from the t-test that showed that there is a slight difference between the scores of first-born and younger siblings (with first-borns scoring higher). Having only a 55% accuracy with our model is not because the questions do a bad job, as the test questions were carefully researched and chosen by the creator of the test. Rather, it's because there's not much of a difference between the answers of first-borns and younger siblings to begin with, so it makes predictions hard because it could go either way. Since it was 55% accurate (better than chance), we could say knowing how the participants did on the test, we could make better predictions about whether they were first-born or not than if we knew nothing at all.

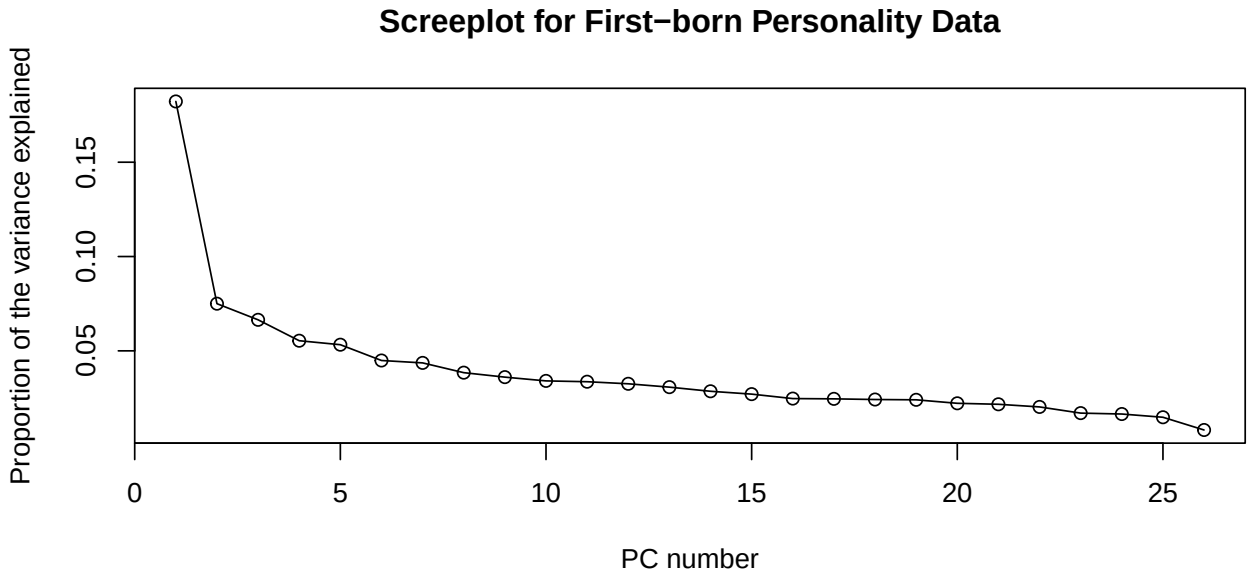
Principal component analysis: Are all these questions necessary to differentiate first-born and younger siblings?

```
# preparing our data for pca
pca_data <- fbps_1 %>%
    select(Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12,
        Q13, Q14, Q15, Q16, Q17, Q18, Q19, Q20, Q21, Q22, Q23,
```

```
        Q24, Q25, Q26)

# pca
pca.fbps <- prcomp(pca_data)

summary(pca.fbps)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.8135 1.80514 1.69901 1.55099 1.52113 1.39634 1.37649
## Proportion of Variance 0.1822 0.07501 0.06645 0.05537 0.05326 0.04488 0.04362
## Cumulative Proportion  0.1822 0.25722 0.32367 0.37904 0.43231 0.47719 0.52080
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.29231 1.25188 1.21672 1.20892 1.18890 1.15599 1.11416
## Proportion of Variance 0.03844 0.03608 0.03408 0.03364 0.03254 0.03076 0.02857
## Cumulative Proportion  0.55925 0.59532 0.62940 0.66304 0.69558 0.72634 0.75491
##                           PC15   PC16    PC17    PC18    PC19   PC20    PC21
## Standard deviation     1.08427 1.0359 1.03276 1.02550 1.02207 0.9821 0.97034
## Proportion of Variance 0.02706 0.0247 0.02455 0.02421 0.02405 0.0222 0.02167
## Cumulative Proportion  0.78198 0.8067 0.83123 0.85544 0.87949 0.9017 0.92336
##                           PC22   PC23    PC24   PC25    PC26
## Standard deviation     0.93858 0.8594 0.84700 0.8018 0.59122
## Proportion of Variance 0.02028 0.0170 0.01651 0.0148 0.00805
## Cumulative Proportion  0.94364 0.9606 0.97716 0.9919 1.00000
```

```
# pca.fbps$rotation (way too long! In the appendix)

# screeplot
plot(pca.fbps$sdev^2/sum(pca.fbps$sdev^2), type = "o", ylab = "Proportion of the variance explained",
    xlab = "PC number", main = "Screeplot for First-born Personality Data")
```



Looking at the PCA results and the screeplot, I notice that it's relatively difficult to reduce the dimension of this data. It seems that most of the variables (questions on the personality test) contribute substantially to the variability in the data. Using a threshold value of capturing 90% of the variability, we would choose 20 principal components. Using 20 principal components would reduce our dimensions by only 6. Since it's

"hard" to reduce the dimensions of our data (or we wouldn't reduce it by a lot of dimensions), it suggests that most of these questions "belong" or are necessary in the personality test. This means that the creator of the test did a good job picking questions that serve a purpose in explaining the variability in the data. Furthermore, taking a look at PC1 (the principal component that explains the most variability out of all the other PC's) and its correlation to the questions, the highest correlations are questions related to intellect and challenging oneself. PC2 is most correlated to questions related to extroversion and "coolness" (the opposite of nerdiness). For the analysis of PC1 and PC2, I looked at the rotation of the pc.fbps object (it's too large to print out).

Conclusion

Are first-borns superior that younger siblings? Yes, marginally, but only if you consider traits like higher intellect, greater openness to experience and extroversion to be superior. From the t-test for difference of means, we observed a small, but statistically significant result: that the mean personality scores of first-borns were higher than that of younger siblings. The logistic regression also confirmed that given the personality scores of an individual, we could predict, at greater than chance accuracy, whether that individual is a first-born or not. However, we could only predict at a 55% accuracy rate, which is only 5% greater than chance. One might argue that this is because the questions on the personality test were not truly capturing traits that would be different between first-borns and younger siblings, but the author of the test made sure to emphasize that these questions were chosen from a much bigger pool and carefully researched. Also through the principal component analysis, we learned that the questions on the personality test were indeed well-chosen because it was hard to reduce the dimensions of the questions. Most of the questions explained a lot of the variability in the data and were "necessary" or belonged in the test. The overarching conclusion that I drew from all these analyses is that first-borns do see marginally higher scores on this personality test, suggesting that they are perhaps smarter, more open to experience, and more extroverted than their younger siblings. However, in practice, this difference is likely meaningless/not obvious. If you meet a new friend and they never explicitly tell you if they are the oldest in the family, you probably wouldn't be able to say with a lot of confidence.

The main limitation with drawing conclusions from this analysis is how the data was collected. Total scores were collected online with self-reported likert scale ratings on a series of questions. The largest studies on birth order effects on personality have found it to be very small and limited in scope. For example, Rohrer, Egloff, and Schmukle (2015) found only a 0.1 SD between 1st born and 2nd born children on intellect and openness to experience and no differences in other traits. However, in groups like professors, first-borns are overrepresented and occur more frequently than chance. This difference might be attributed to self-reporting and self-report bias. This data was all self-reported and so was the Rohrer, Egloff, and Schmukle research. My analyses using this data corroborates the results of the Rohrer study. However, looking upon populations like professors and noticing who's a first-born or not is not self-reported. Because of the limitation of how this data was collected, we aren't able to sift the effects of self-reporting.

For further studies, I would be interested/curious in a study/analysis that addresses the limitation I pointed out about self-report bias. People aren't the most accurate/honest judges of themselves. Perhaps, a study that asks similar personality questions as this test used, but instead of asking the person directly, asking parents, friends, teachers, etc. This would cirucmvent the issue of self-report bias, but perhaps this method would introduce other kinds of bias.

Reflection

There were a lot of up and down moments with this project. I immediately had a sense of what I project topic I wanted to do. I devoted a lot of time into researching and getting data for it, only to change my

project idea at the last minute because the data I had for my original idea wasn't comprehensive enough for an in-depth analysis. I found it frustrating to track down data in usable formats. At the last minute, I switched to this topic idea about first-borns and whether they have different personalities from younger siblings. I really enjoyed choosing what type of analyses would help me answer questions of interest. In the homeworks, choosing analyses/models for to answer questions was always spoonfed to us, but now we had to make decisions on what analyses would help us answer interesting questions. I struggled a lot with interpretation, especially with the primary component analysis because we didn't spend a lot of time going over this concept. It was hard to understand what I should be taking away from the pca object. I also struggled with thinking of good ways to visualize the data that was interesting and intuitive. I wanted to make a visualization that was creative, but I ended up going with distributions that I believe were helpful and easy to interpret. For analyses I did, but didn't end up including was a correlation test/matrix of all the questions. There are just too many variables for this to be meaningful, and correlation only tests relationships between two variables. I instead resorted to doing a primary component analysis instead.

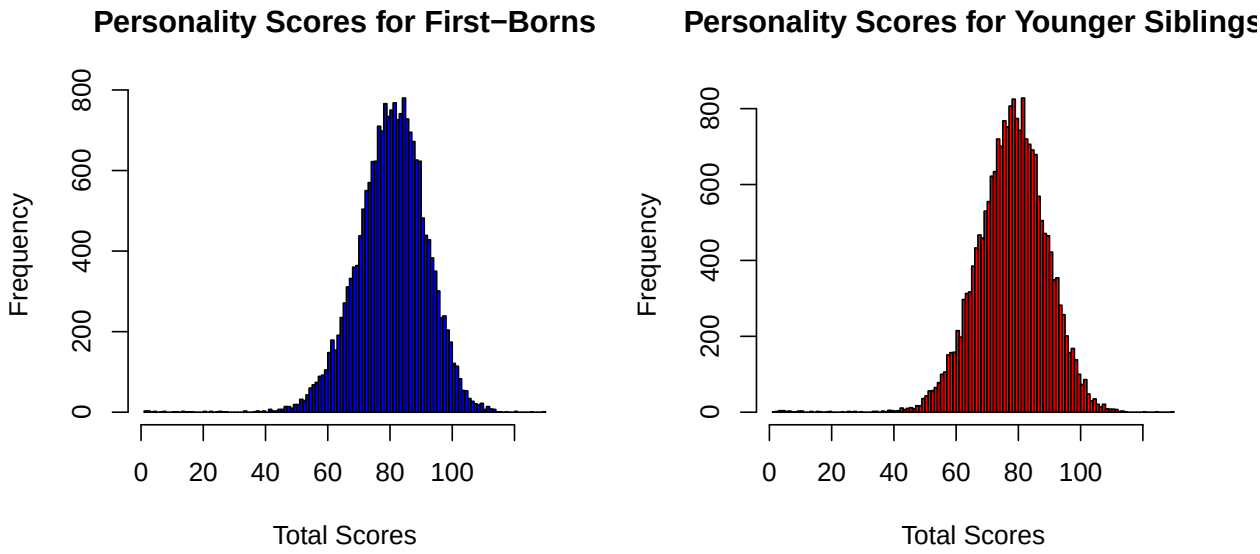
Appendix

Sources // 1. Data from https://openpsychometrics.org/tests/birthorder/development/

2. Victor Lavrenko, YouTube Video on how many dimensions to choose when doing PCA. Retrieved from: https://www.youtube.com/watch?v=KE_fxBCBS3w
3. Dr. Bharatendra Rai, YouTube Video on logistic regression. Retrieved from: https://www.youtube.com/watch?v=AVx7Wc1CQ7Y
4. UC Buisiness Analytics R Programming Guide, Principal Component Analysis. Retrieved from: https://uc-r.github.io/pca
5. Ludvig Renbo Olsen, Creating a confusion matrix with cvms. Retrieved from: https://cran.r-project.org/web/packages/cvms/vignettes/Creating_a_confusion_matrix.html#manually-creating-a-two-class-confusion-matrix

Histograms for Normality in t-test Normality of data is an assumption for a t-test. Let's use histograms to visually check this!

```
par(mfrow = c(1, 2))

hist(first_born_df$total_score, breaks = 100, main = "Personality Scores for First-Borns",
    xlab = "Total Scores", col = "blue")

hist(younger_sibs_df$total_score, breaks = 100, main = "Personality Scores for Younger Siblings",
    xlab = "Total Scores", col = "red")
```



These histograms look approximately normal because they are unimodal and symmetric.

Summary for lr_fit in Logistic Regression

```
summary(lr_fit)
```

```
##
## Call:
## glm(formula = first_born ~ total_score, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5924  -1.1583  -0.9352   1.1694   1.9378
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.732798   0.081042  -21.38   <2e-16 ***
## total_score  0.021361   0.001007   21.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46179  on 33316  degrees of freedom
## Residual deviance: 45715  on 33315  degrees of freedom
## AIC: 45719
##
## Number of Fisher Scoring iterations: 4
```

The logistic regression shows that for an increase in total_score, the odds of being a first_born increases because the sign on the coefficient is positive (albiet small).

Principal factor analysis: PC analysis

```r
a <- pca.fbps$rotation

a[1:26, 1:3]
```

```
##             PC1         PC2         PC3
## Q1   0.28106566 -0.18394384  0.24945030
## Q2   0.22035284  0.48390073  0.03985486
## Q3   0.31059512 -0.05722398  0.08419164
## Q4   0.24439909  0.17434665 -0.12859742
## Q5  -0.15315402 -0.25368741 -0.06584813
## Q6   0.21664375 -0.29417589  0.33744645
## Q7   0.29011513  0.00715567 -0.01351893
## Q8   0.28374797  0.03204751  0.02903495
## Q9   0.25145536 -0.02701909  0.09696824
## Q10  0.20142985 -0.19784119  0.16656253
## Q11  0.25797729 -0.02265463 -0.04829994
## Q12  0.18223688 -0.22499473 -0.19267547
## Q13  0.05351536  0.04677168 -0.27684178
## Q14  0.20674595 -0.13006845  0.12392113
## Q15  0.08785913 -0.14754335 -0.22589482
## Q16  0.23241143 -0.20560574 -0.02899318
## Q17  0.02150350  0.05711792 -0.05172001
## Q18 -0.10001272 -0.31050660 -0.19539117
## Q19  0.09727228  0.03328279 -0.39795009
## Q20  0.03578779 -0.02882119  0.05734158
## Q21 -0.19816559 -0.19013727 -0.07354607
## Q22 -0.08555479 -0.26054703  0.20975390
## Q23  0.15387791  0.05595167 -0.02878344
## Q24  0.20699095  0.17862542 -0.21917311
## Q25 -0.05461543 -0.27661840 -0.13303681
## Q26  0.18156087 -0.23451164 -0.50436530
```

Sorry if it's really long! For PC1, I noticed the highest correlation questions were Q3, 1, 7, and 8. The largest negative correlation questions were Q21 and Q5. The questions are as follows:

Q1 I have read an absurd number of books.(+)

Q3 I love to read challenging material.(+)

Q7 I use difficult words.(+)

Q8 I have a rich vocabulary.(+)

Q21 I find too much thinking exhausting. (-)

Q5 I like simple work. (-)

These questions seem to be related to an individual's intellect. Reading books, using difficult words, liking a challenge, and finding thinking not exhausting are charactersitics I associate with somebody's intellect.

For PC2, I noticed the highest correlation questions was Q2. The highest negative correlations were Q6, Q18, Q12. These questions are as follows:

Q2 I have traveled alone in a foreign country. (+)

Q6 I would rather read a book than go to a party. (-)

Q18 I miss my childhood. (-)

Q12 I like science fiction.(-)

These questions seem to be related to a person's openness to trying new things and extroversion.