140SL Final Project Report

Andy Hierl (304427889), Sidney Lee (104566230),
Leon Luc (304443064), Carson Zhao (504448169)

12/10/2017

For this project, we analyzed the data on California STAR testing scores for 2010 and 2012 (cst2010 and cst2012) that were posted under Week 3 on CCLE. We decided to append both of those datasets together so that we could have more data for our analysis. Some of the variables in the dataset include County.Name, Zip.Code, and Mean.Scale.Score. After deciding on Mean.Scale.Score (average score per grade for each school for the STAR testing) as our primary response variable, we became interested in researching whether the various California counties had significantly different Mean.Scale.Scores from one another. We also wanted to find out what factors ultimately contribute to Mean.Scale.Scores.

Since we aimed to get more data about the counties, we found and merged two external datasets with the appended cst data. The first external dataset contained information about the county population, population density, per capita income, and median household income while the second one gave us the male percent ratio in each county. This data, however, was based off of 2017 as we were unable to find the same ideal data for 2010 and 2012. Once we looked into how median household income affected the Mean.Scale.Score, we were interested in determining why this variable could have affected the response variable. We became interested in seeing metrics such as, the amount of spending on education per student for each of the counties and also how many tutors per 1000 students a county had.

We continued to clean the data by removing the NA's for Mean.Scale.Score. We first checked to see if there was a pattern in the missing values (e.g. we checked to see if a certain county was missing significantly more Mean.Scale.Scores than any of the other counties). We determined that there was no strange pattern in the missing values and so we decided to delete them from the data. We then created additional categorical variables based off of the variables added by the first external dataset. These new categorical versions each had four levels where each level represented the quartile of the original numeric data value fell in (i.e. a county with median income in the 42nd percentile would be in the second quartile level of the categorical median income variable). This allowed us to produce boxplots to compare the different quartiles with the mean scale score.

The last data transformation that we did was geocoding the zip codes in order to obtain the longitude and latitude values needed to create maps. We chose to geocode zip codes instead of counties because there were over a thousand unique zip code values, which would allow for our map of California to be more filled out and thus show more of the variation in mean scale scores throughout the state and also within each county.

Having completed the initial data cleaning and transformation, we then began to explore the data in depth to answer our research question. The first thing that we looked into was the mean scale scores. Using a histogram to group the scores into bins, we found that the variable was pretty normally distributed. We then utilized our geocoding to map the scores for zip codes across California. The scores were colored similarly to how we categorized the external variables using the quartiles they fell under. From the map, we could see that the largely populated areas of California, such as the San Francisco Bay Area and Los Angeles County, contained most of the higher scores of blue and purple. On the other hand, as we shift towards the central and outer parts of the state such as the San Joaquin Valley and the mountain and desert areas, these rural and less populated regions exhibited lower scores of the red and green points. This made us interested in seeing whether various demographic variables such as population density and income could have an significant effect on the mean scale scores.

Next, we wanted to find out if the various counties had significantly different mean scale scores from each other. Based on the map previously mentioned, we hypothesized that the counties did differ due to the distribution of the scores. A boxplot of counties and mean scale scores shows that there is definitely some

variation, but it was difficult to determine whether the variation was significant based on the plot. Therefore, we conducted an ANOVA and with a p-value less than 0.001, we rejected the null hypothesis and concluded that the average mean scale score is not the same for all counties. This may be because some counties are wealthier on average than others, or have better school districts than others.

We then looked at population density, which is the number of people divided by the land area. For example, Los Angeles county would have a much higher population density than San Luis Obispo county. Through summary statistics, we found that mean scale score typically increases as population density increases. An ANOVA of the categorical variable version of population density and the response variable mean scale score yielded a p-value of less than 0.001. This means that the average mean scale scores is not the same among the four quartiles of population density.

We also looked at each county's median income. When we looked at the boxplot of all counties' mean scale scores color coded by the median income, we saw that the counties with high variation in scores were also the counties with relatively high median income. In order to dig in further as to why this is, we did some more analyses.

Mean scale score and median income had a positive correlation. This could be because counties with higher median income also receive more funding for education, different family backgrounds, and many other possible factors to consider. When we graphed mean scale score with median income and population density, we found that the higher median income counties were also counties with high population densities. So we decided to do some research about what those counties were. Those counties included Los Angeles County, Santa Clara County, and San Mateo County. These counties all cover big cities–LA, SF, and San Jose–and a large amount of suburban areas as well. We can expect that the demographics among neighborhoods within their own counties are very different, and also different from counties that only cover suburban areas, or areas without big cities. This explained the high variation of scores for counties with high median income.

Andy contributed to the project by helping explore the data, looking for the new county data that would be merged with the original dataset, creating the boxplots for the Mean.Scale.Score as a function of the counties and whether or not the county had a male majority ($> 50\%$ males living in the county), combining our code into one script, and writing the report. Carson contributed to the project by helping explore the data, descriptive statistics, cleaning the data, removing the NAs in the response variable, converting variables that should be categorical to factors, creating new categorical data, creating boxplots and other graphics, research question formulation, linear regression, ANOVA, and report editing. Leon contributed to the project by geocoding the zip codes to allow for mapping, brainstorming and confirming ideas, and organization and proofreading of the slides and report for better aesthetic viewing. Sidney contributed to the project by helping explore the data, creating scatter/boxplots, discussing which data to explore, question formulation, t-tests, report/presentation editing.

There's always more analysis that can be done, and in the future, we hope to explore this topic further. From our analyses, we were able to determine that county, median income, and population density all were correlated with mean scale score. Future analyses should try to investigate why these variables have an effect on mean scale score. Do higher income families pay for more tutors for their children? Does the California government pay more per student on education in densely populated counties? Do higher income families place more of an importance on education? Having data on the number of tutors per child, government spending per student (by county), and the importance of education to parents would give us more insight as to why county, median income, and population density would be correlated to mean scale score. Education is a vital part of the future of America, and although standardized testing may not be a perfect metric, it can indicate the relative proficiency level of students, and allow us to compare different counties. Below is the appendix, which contains the R code we used to analyze the California STAR Testing data set.

Appendix: R Code

```
# Load Packages
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(readxl)
library(readr)
library(ggmap)
```

```
## Loading required package: ggplot2
```

```
library(maps)
library(reshape2)
```

```
# Cleaning the Data
load("C:/cst2012.Rdata")
load("C:/cst2010.Rdata")

# Merge 2010 and 2012 data sets
cst <- rbind(cst2010, cst2012)

# Merge socio-economic county data
county <- read_excel("/Users/carsonzhao/Downloads/county.xlsx")
county$County.Name <- county$County
county$County <- NULL

cst <- inner_join(cst, county, by = "County.Name")

# Add Gender ratios per county
MalePercCounty <- read_csv("/Users/carsonzhao/Downloads/MalePercCounty.csv")

cst <- cst %>%
  inner_join(MalePercCounty, by = "County.Name")

# does a county have a male majority?
cst$maleMajority <- cst$`Male % Ratio` > 0.5 %>% as.numeric()

# remove NAs in response variable
cst2 <- cst[complete.cases(cst$Mean.Scale.Score), ]
names(cst2)[27:30] <- c("Pop.Density", "Per.Capita.Income", "Median.Household.Income",
                        "Median.Family.Income")
cst3 <- cst2

# Convert a few variables to categorical variables
cst3$County.Name <- as.factor(cst3$County.Name)
```

```r
cst3$District.Name <- as.factor(cst3$District.Name)
cst3$School.Name <- as.factor(cst3$School.Name)

cst <- cst3


# Change Charter.Number to a factor with 2 levels
cst2 <- cst
cst2$Charter.Number[cst2$Charter.Number != 0] <- 1
table(cst2$Charter.Number)
cst2$Charter.Number <- as.factor(cst2$Charter.Number)
cst <- cst2


# New categorical variables and exploratory analysis on them
cst4 <- mutate(cst,
               Med_Income = cut(Median.Household.Income,
                                breaks = c(0, 54100, 55870, 68507, 1000000),
                                labels = c("first_quartile", "second_quartile",
                                           "third_quartile", "fourth_quartile")),
               Pop_Density = cut(Pop.Density,
                                 breaks = c(0, 159.2, 756.7, 2457.9, 1000000),
                                 labels = c("first_quartile", "second_quartile",
                                            "third_quartile", "fourth_quartile")),
               Pop = cut(Population,
                         breaks = c(0, 701050, 1841569, 3183143, 99740000),
                         labels = c("first_quartile", "second_quartile",
                                    "third_quartile", "fourth_quartile"))
)
cst <- cst4
rm(cst4)
rm(cst3)
rm(cst2)

summary(cst)

write.csv(cst, file = "cst.csv")
```

```r
cst <- read_csv("/Users/carsonzhao/Desktop/Stats 140SL/140SL Assignment 8/cst.csv")
```

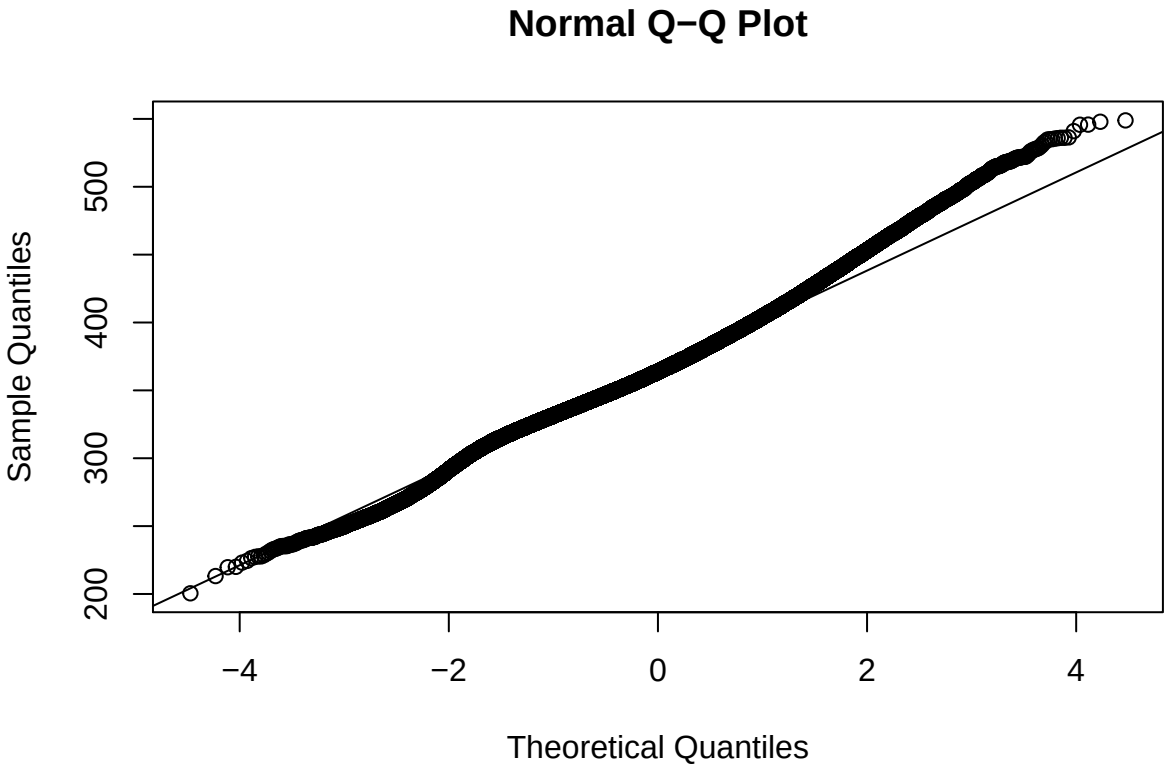
```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Percent.Tested = col_double(),
##   Mean.Scale.Score = col_double(),
##   County.Name = col_character(),
##   District.Name = col_character(),
##   School.Name = col_character(),
##   Pop.Density = col_double(),
##   `Male % Ratio` = col_double(),
##   maleMajority = col_logical(),
##   Med_Income = col_character(),
```

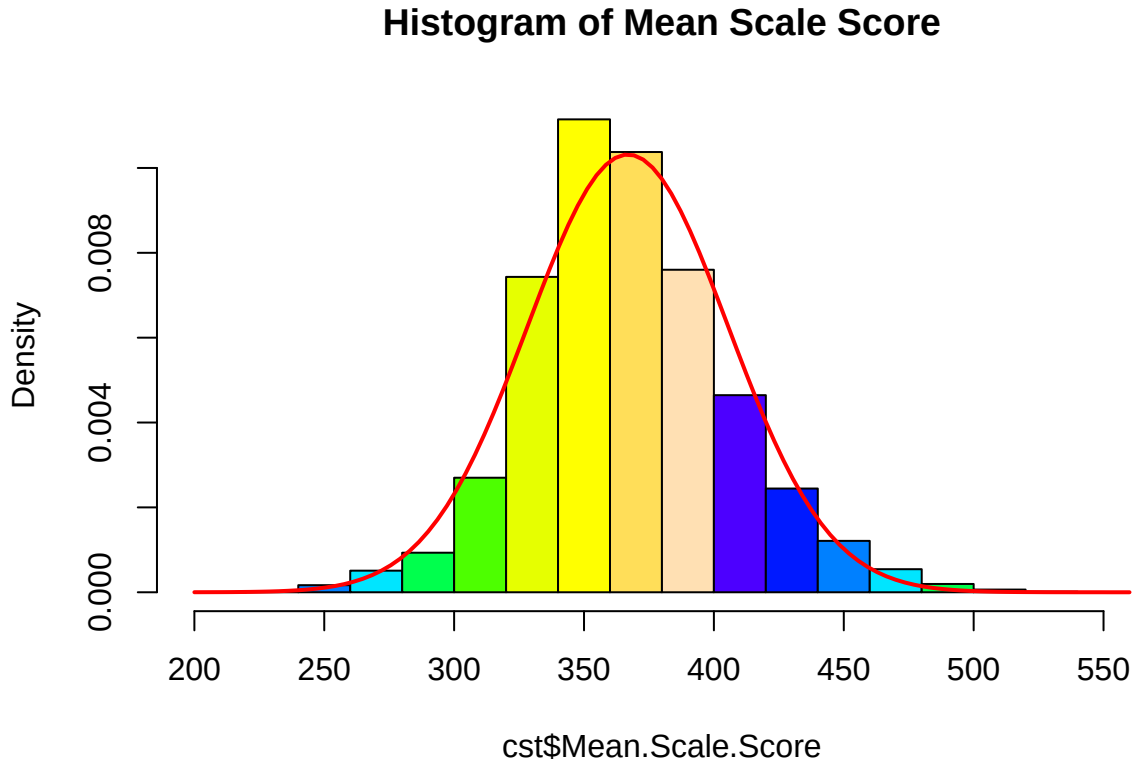
```
##   Pop_Density = col_character(),
##   Pop = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

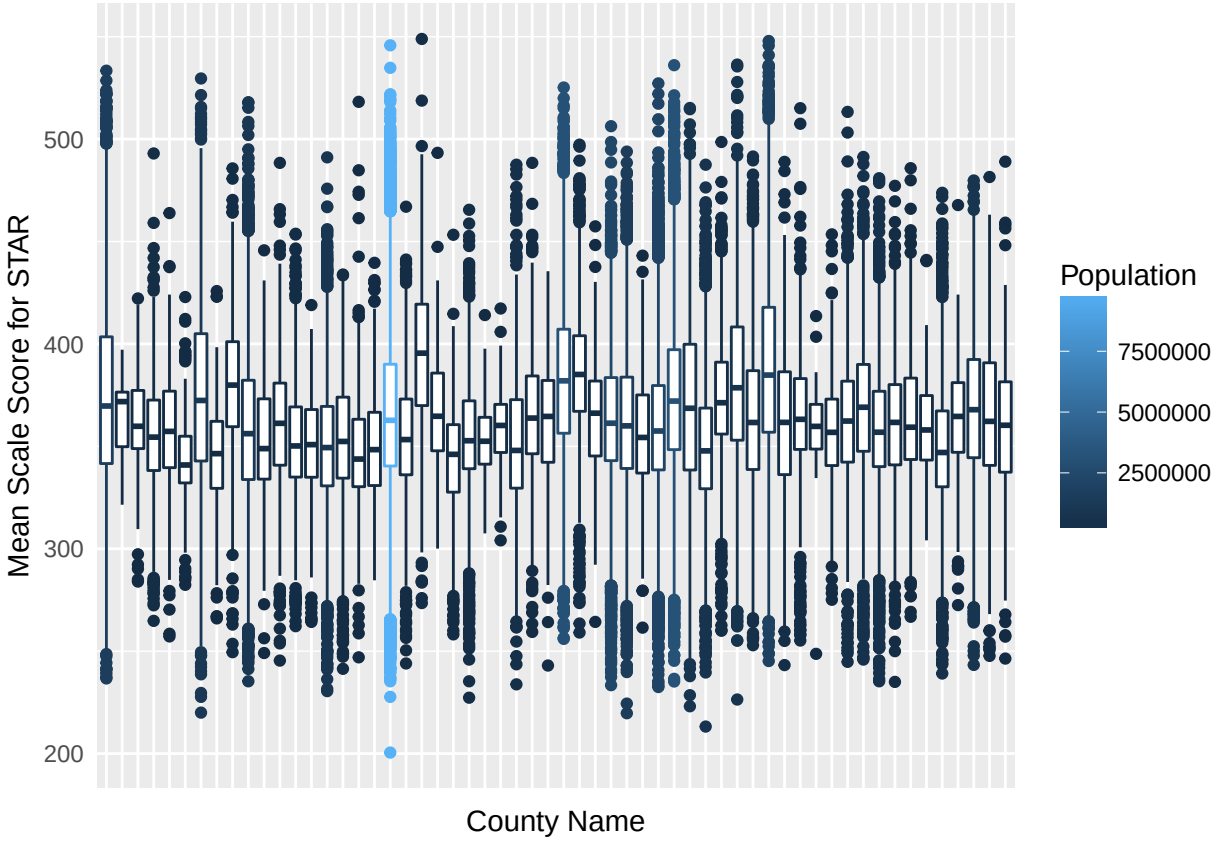
```
# Mean Scale Score
qqnorm(cst$Mean.Scale.Score)
qqline(cst$Mean.Scale.Score)
```




```
hist(cst$Mean.Scale.Score, col=topo.colors(10),
     main="Histogram of Mean Scale Score", prob = T)
curve(dnorm(x, mean=mean(cst$Mean.Scale.Score), sd=sd(cst$Mean.Scale.Score)),
      col="red",
      lwd=2, add=TRUE, yaxt="n")   # lwd changes line size
```



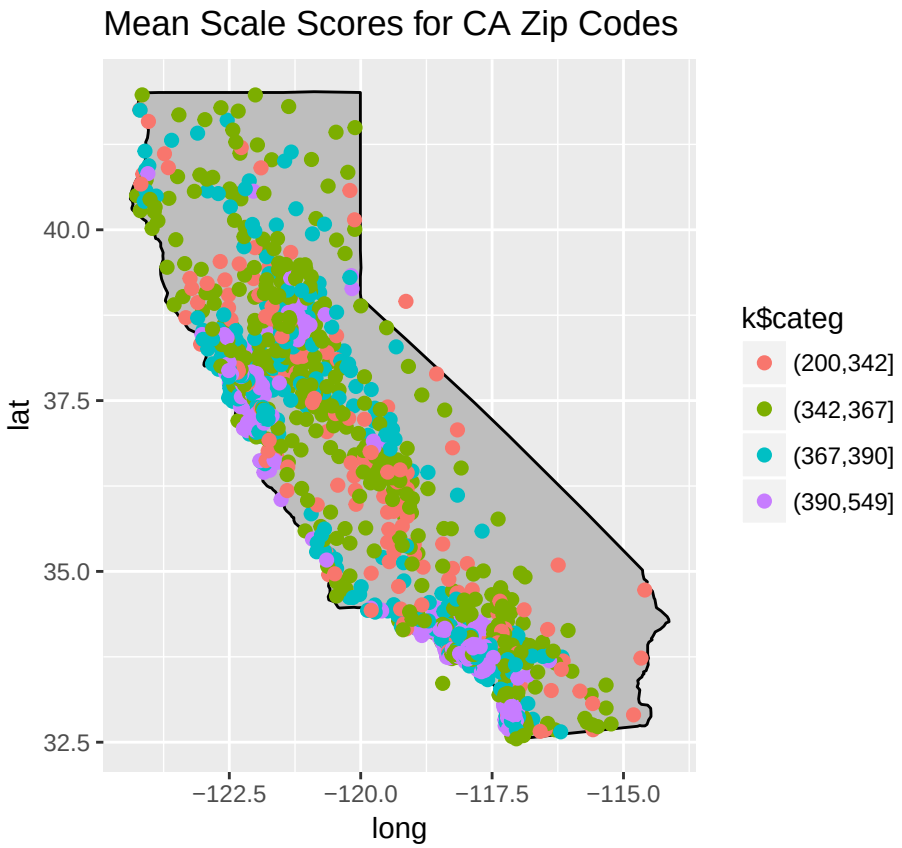

```
# Boxplots of different Mean Scale Scores for the different counties
ggplot(cst, aes(x = County.Name, y = Mean.Scale.Score, color = Population)) +
  geom_boxplot() + xlab("County Name") + ylab("Mean Scale Score for STAR") +
  theme(axis.text.x = element_blank(), axis.ticks = element_blank())
```




```
# code of CA map for different zip codes of the schools
k <- cst %>% group_by(Zip.Code) %>% summarise(MeanScaleScore=mean(Mean.Scale.Score))
# locg <- geocode(k$Zip.Code, output="latlona", messaging=FALSE, source="google")

load("/Users/carsonzhao/Downloads/locg.Rda")

k$lon <-locg$lon
k$lat <-locg$lat
#save(locg, file="locg.Rda")
k<- k %>% filter(lon <0) #not a part of California
summary(k$MeanScaleScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   253.9   350.0   363.4   367.0   383.4   465.3
```

```
k$categ <- cut(k$MeanScaleScore, c(summary(cst$Mean.Scale.Score)[1],summary(cst$Mean.Scale.Score)[2],su

states <- map_data("state")
cali <- subset(states, region %in% c("california"))

ggplot(data = cali) + ggtitle("Mean Scale Scores for CA Zip Codes") +
  coord_fixed(1.3) +  geom_polygon(aes(x = long, y = lat, group=group), fill = "gray", color = "black")
  geom_point(data=k, aes(x=lon, y=lat, color=k$categ) ,size=2)
```




```
# Mean Scale Score vs Income, Population Density, and Population

cst %>% group_by(Med_Income) %>%
  summarise(Avg_Score = mean(Mean.Scale.Score),
            Number_Of = n(), Std_Dev = sd(Mean.Scale.Score))
```

```
## # A tibble: 4 x 4
##        Med_Income Avg_Score Number_Of  Std_Dev
##             <chr>     <dbl>     <int>    <dbl>
## 1  first_quartile  355.7451     35134 33.70604
## 2 fourth_quartile  379.8130     32171 42.19650
## 3 second_quartile  365.4924     34743 38.05870
## 4  third_quartile  367.7032     26773 36.29592
```

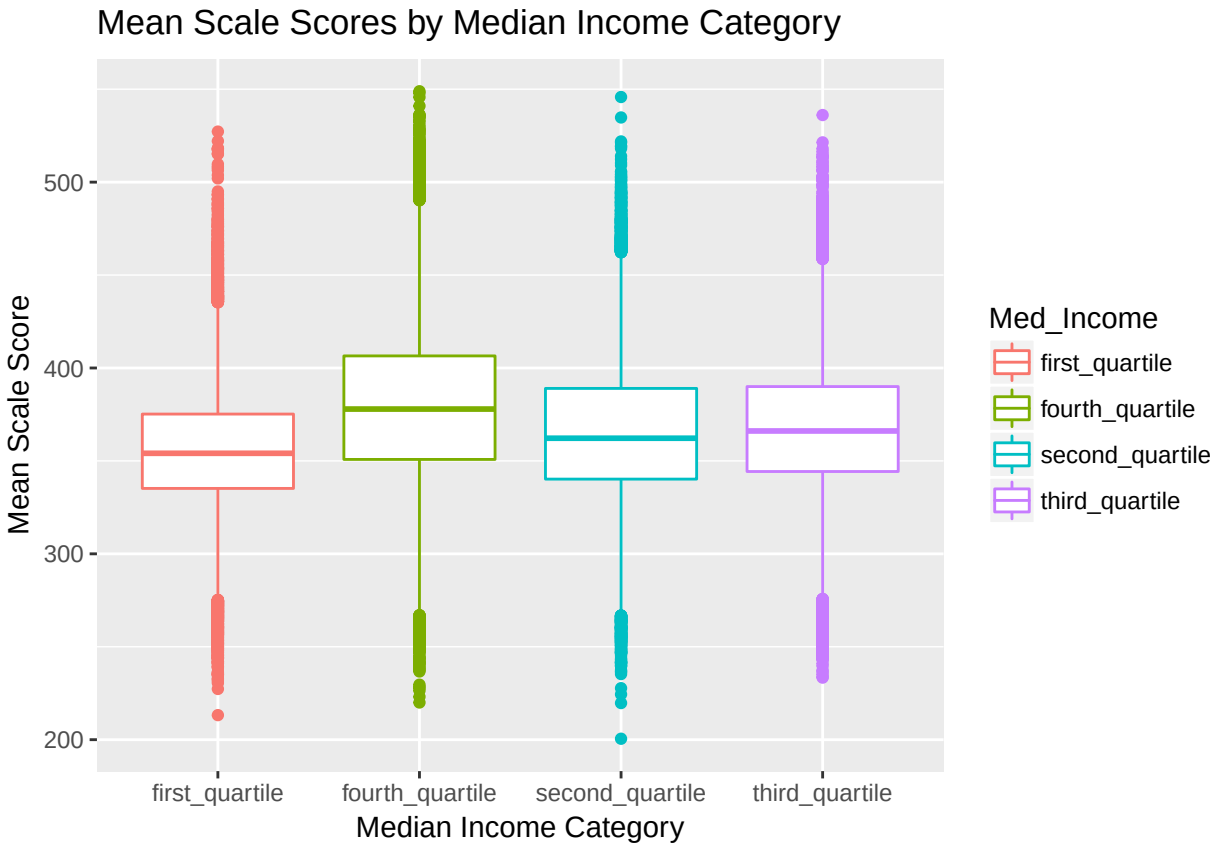
```
cst %>% group_by(Pop_Density) %>%
  summarise(Avg_Score = mean(Mean.Scale.Score),
            Number_Of = n(), Std_Dev = sd(Mean.Scale.Score))
```

```
## # A tibble: 4 x 4
##       Pop_Density Avg_Score Number_Of  Std_Dev
##             <chr>     <dbl>     <int>    <dbl>
## 1  first_quartile  357.6312     35214 34.43885
## 2 fourth_quartile  381.1067     10220 39.54197
## 3 second_quartile  367.2224     33622 36.36127
## 4  third_quartile  370.2450     49765 41.27196
```

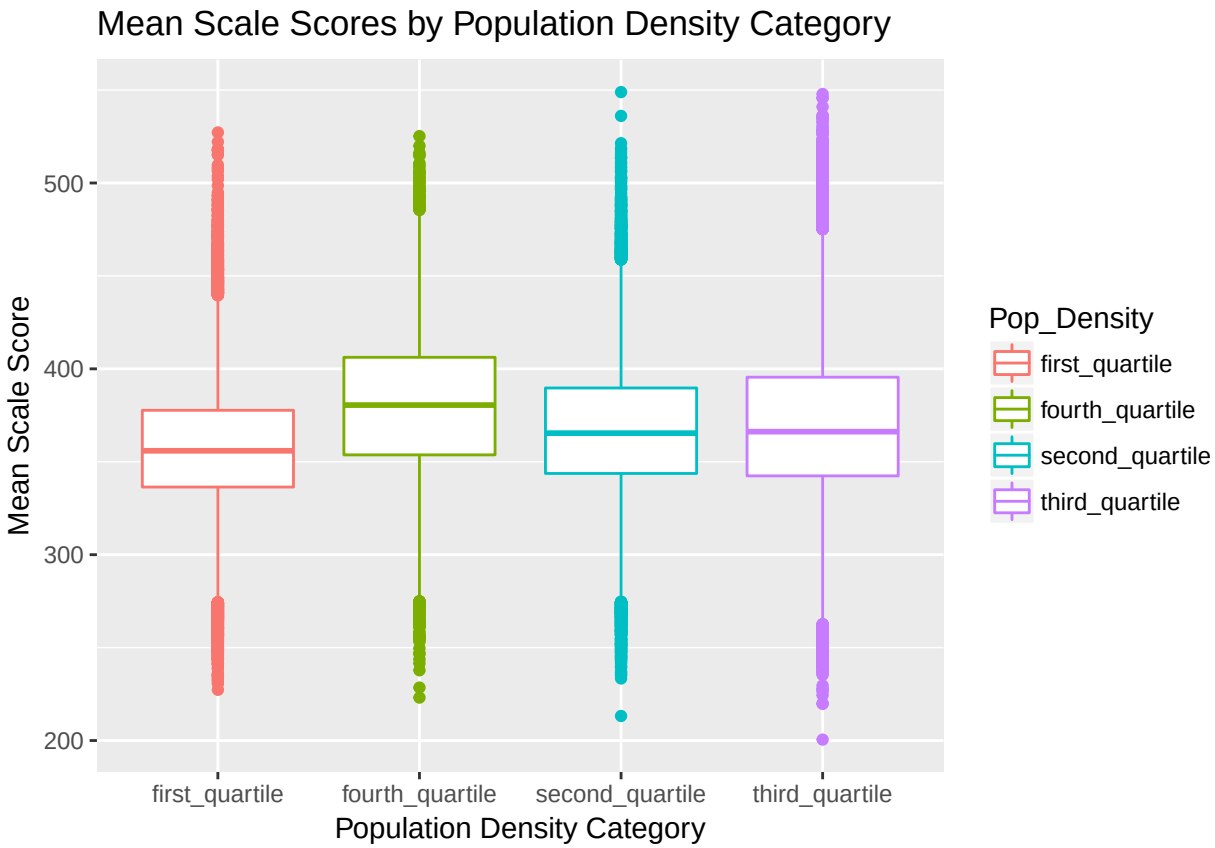
```
cst %>% group_by(Pop) %>%
  summarise(Avg_Score = mean(Mean.Scale.Score),
            Number_Of = n(), Std_Dev = sd(Mean.Scale.Score))
```

```
## # A tibble: 4 x 4
##                Pop Avg_Score Number_Of  Std_Dev
##              <chr>     <dbl>     <int>    <dbl>
## 1  first_quartile  360.3154     34102 34.60695
## 2 fourth_quartile  366.2798     28145 38.71007
## 3 second_quartile  369.9589     34013 42.49025
## 4  third_quartile  371.0175     32561 37.51714
```

```
ggplot(cst, aes(Med_Income, Mean.Scale.Score, colour=Med_Income)) +
  geom_boxplot() +
  labs(title="Mean Scale Scores by Median Income Category", x = "Median Income Category",
       y = "Mean Scale Score")
```



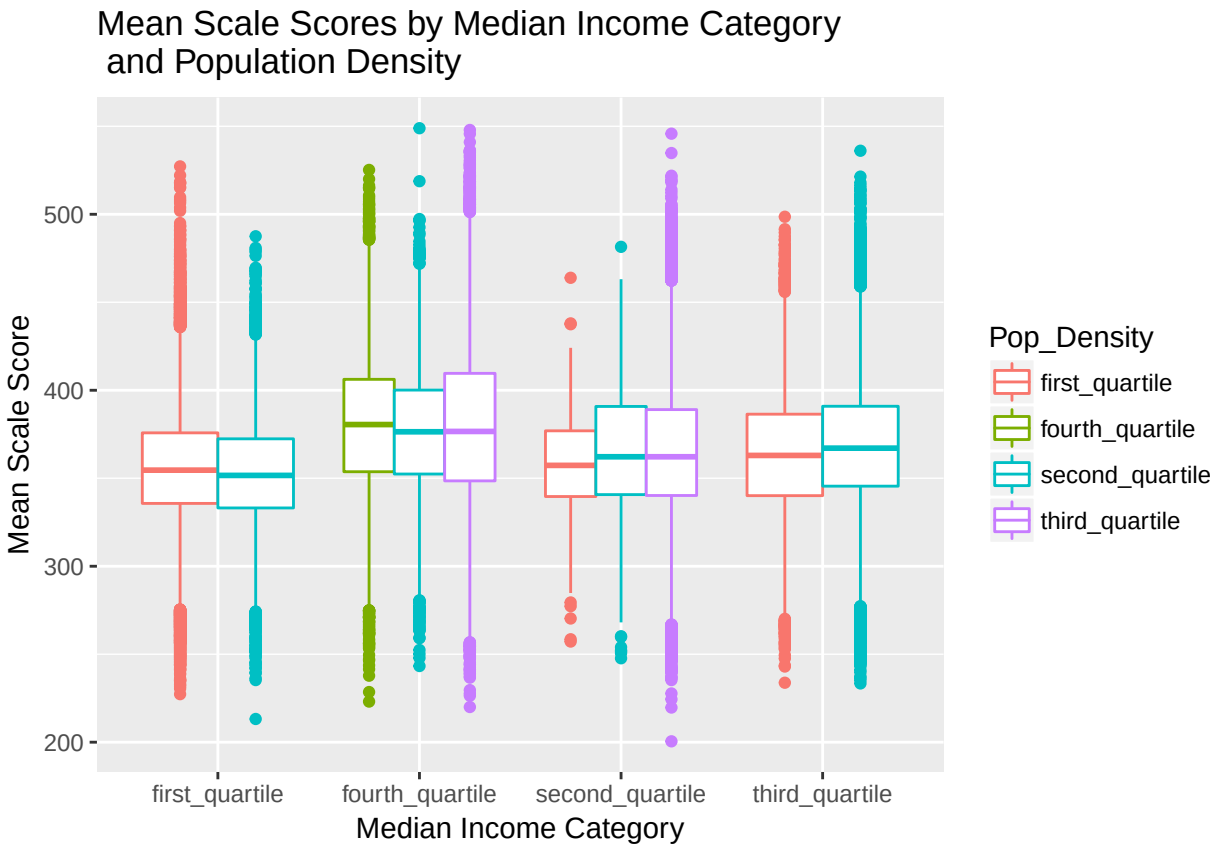

```
ggplot(cst, aes(Pop_Density, Mean.Scale.Score, colour=Pop_Density)) +
  geom_boxplot() +
  labs(title="Mean Scale Scores by Population Density Category",
       x = "Population Density Category",
       y = "Mean Scale Score")
```



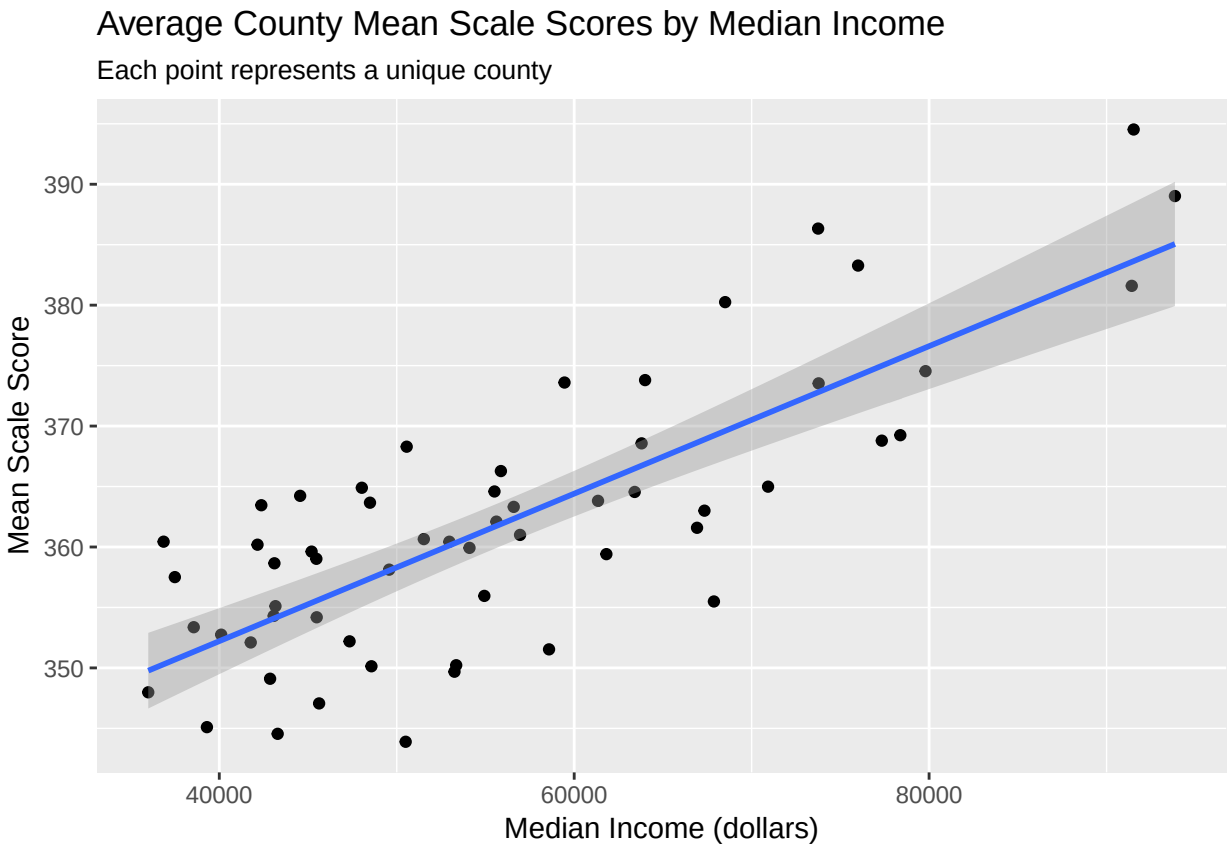

```
cst %>% group_by(Med_Income, Pop_Density) %>%
  summarise(Avg_Score = mean(Mean.Scale.Score),
            Number_Of = n(), Std_Dev = sd(Mean.Scale.Score))
```

```
## # A tibble: 10 x 5
## # Groups:   Med_Income [?]
##         Med_Income     Pop_Density Avg_Score Number_Of  Std_Dev
##              <chr>           <chr>     <dbl>     <int>    <dbl>
##  1  first_quartile  first_quartile  356.3353     28928 34.03163
##  2  first_quartile second_quartile  352.9936      6206 32.00415
##  3 fourth_quartile fourth_quartile  381.1067     10220 39.54197
##  4 fourth_quartile second_quartile  376.9147      5932 36.26032
##  5 fourth_quartile  third_quartile  380.0608     16019 45.68807
##  6 second_quartile  first_quartile  355.9550       260 30.67079
##  7 second_quartile second_quartile  364.5927       737 36.85458
##  8 second_quartile  third_quartile  365.5855     33746 38.12726
##  9  third_quartile  first_quartile  363.9242      6026 35.81969
## 10  third_quartile second_quartile  368.8009     20747 36.36039
```

```
ggplot(cst, aes(Med_Income, Mean.Scale.Score, colour=Pop_Density)) +
  geom_boxplot() +
  labs(title="Mean Scale Scores by Median Income Category \n and Population Density",
       x = "Median Income Category",
       y = "Mean Scale Score")
```




```
# Code for income analysis

# unique county scatterplot
cst_order <- cst[order(cst$County.Name), ]
uniq_inc <- unique(cst_order$Median.Household.Income)
uniq_score <- cst %>% group_by(County.Name) %>%
  summarise(mean = mean(Mean.Scale.Score))
df1 <- as.data.frame(cbind(uniq_score, uniq_inc))

ggplot(df1, aes(uniq_inc, mean)) + geom_point() +
  geom_smooth(se = TRUE, method = "lm") +
  labs(title="Average County Mean Scale Scores by Median Income",
       x = "Median Income (dollars)",
       y = "Mean Scale Score",
       subtitle = "Each point represents a unique county")
```




```
# Linear Regression
mod3 <- lm(mean ~ uniq_inc, data = df1)
summary(mod3)    # R-squared = 0.623
```

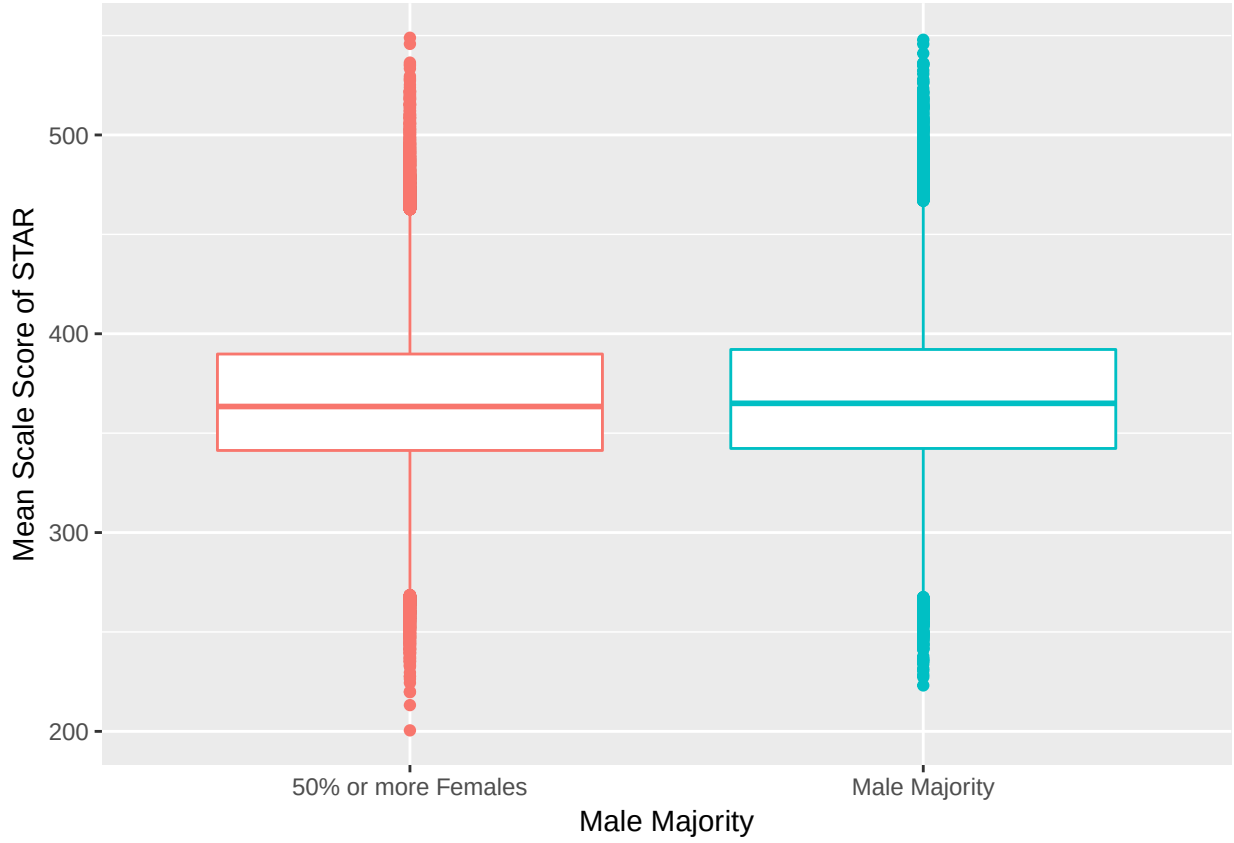
```
##
## Call:
## lm(formula = mean ~ uniq_inc, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7279  -5.7561   0.2666   4.5057  13.5369
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.278e+02  3.669e+00  89.340  < 2e-16 ***
## uniq_inc    6.099e-04  6.345e-05   9.613 1.87e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.912 on 56 degrees of freedom
## Multiple R-squared:  0.6227, Adjusted R-squared:  0.6159
## F-statistic: 92.41 on 1 and 56 DF,  p-value: 1.875e-13
```

```
# Income significant for possibly many reasons, would want further data to dig into why income
# is significant in determining STAR scores.

gender <- cst %>%
  group_by(County.Name) %>%
```

```
  summarize(maleMajority = mean(maleMajority))

table(gender$maleMajority)
```

```
##
##  0  1
## 34 24
```

```
# there are 24 counties with a male majority

ggplot(cst, aes(x = maleMajority, y = Mean.Scale.Score, color = maleMajority)) +
  geom_boxplot() +
  scale_x_discrete(labels = c("50% or more Females", "Male Majority")) +
  xlab("Male Majority") + ylab("Mean Scale Score of STAR") +
  theme(legend.position = "none")
```




```
t.test(cst$Mean.Scale.Score[cst$maleMajority == 1],
       cst$Mean.Scale.Score[cst$maleMajority == 0])
```

```
##
##  Welch Two Sample t-test
##
## data:  cst$Mean.Scale.Score[cst$maleMajority == 1] and cst$Mean.Scale.Score[cst$maleMajority == 0]
## t = 8.1765, df = 43694, p-value = 3e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.664005 2.713308
## sample estimates:
```

```
## mean of x mean of y
##  368.5779  366.3893
```

```
# significant difference
# but is it different by 10 points?
# Not according to the confidence interval


# Anova
mod1 <- lm(cst$Mean.Scale.Score ~ cst$County.Name)
anova(mod1)  # Mean Scale Score and County
```

```
## Analysis of Variance Table
##
## Response: cst$Mean.Scale.Score
##                     Df    Sum Sq Mean Sq F value    Pr(>F)
## cst$County.Name     57  13835408  242726  174.83 < 2.2e-16 ***
## Residuals       128763 178773208    1388
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(cst$Mean.Scale.Score ~ cst$Med_Income)
anova(mod2)  # Mean Scale Score and Median Income
```

```
## Analysis of Variance Table
##
## Response: cst$Mean.Scale.Score
##                    Df    Sum Sq Mean Sq F value    Pr(>F)
## cst$Med_Income      3   9822174 3274058  2307.4 < 2.2e-16 ***
## Residuals      128817 182786441    1419
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod3 <- lm(Mean.Scale.Score ~ Pop_Density, data = cst)
anova(mod3)  # Mean Scale Score and Population Density
```

```
## Analysis of Variance Table
##
## Response: Mean.Scale.Score
##                   Df    Sum Sq Mean Sq F value    Pr(>F)
## Pop_Density        3   5648205 1882735  1297.2 < 2.2e-16 ***
## Residuals     128817 186960411    1451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```