# Test Scores

1.) The observational unit in this study is **Section**

2.) The variables mentioned in the narrative paragraph are:
**Section – Categorical**
**Score – Quantitative**

3.) See code below

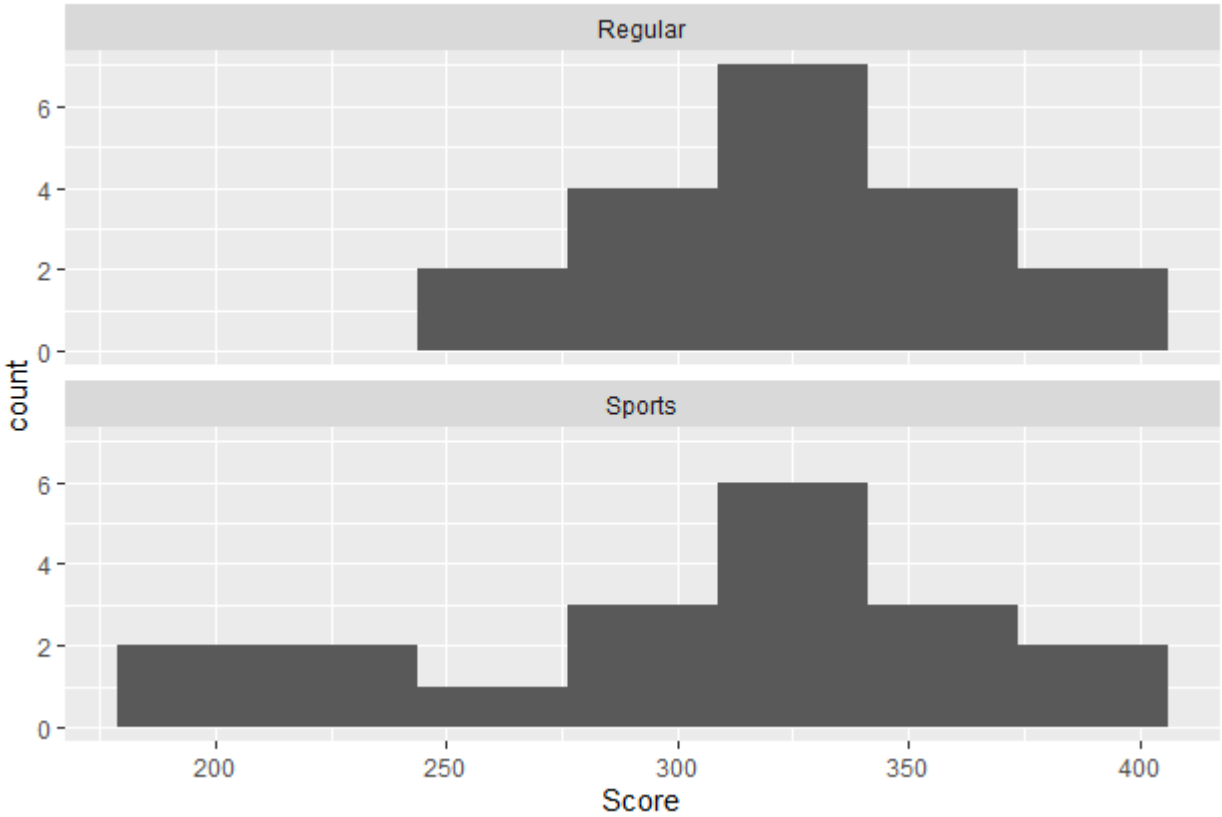
4.) See code below

a. The regular section had tended to score slightly higher scores than the sports section. Both the mean and median of the regular section were higher than the sports section.

b. No, one section did not have every, or even most, students scoring higher than the other section. In this case there were a few cases of students scoring much lower than the average in the sports section which seemed to bring down the average for that group.

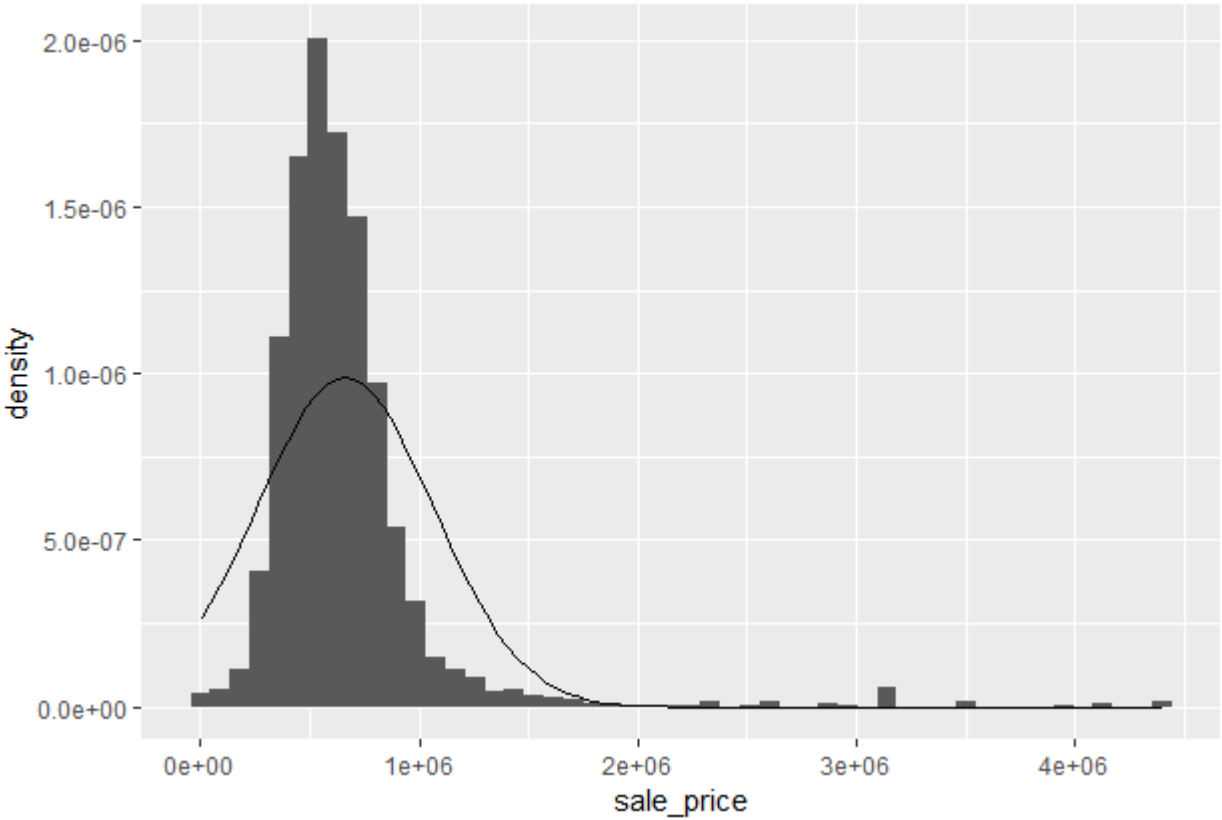
c. The variable that was not discussed in the narrative was the '**Count**'. At first glance this variable seemed to be the number of students in each section that attained the same score, however given that all the instances of count were multiples of ten with nothing above thirty this seemed unlikely. Upon further analysis of the data there were also two Counts of twenty in the sports section for the same score, 320, which further discounted the idea that this was the number of students achieving a single score.

```
scores <- read.csv("data\\scores.csv", header=TRUE, stringsAsFactors=TRUE)

sport_scores <- subset(scores, scores$Section == "Sports")
reg_scores <- subset(scores, scores$Section == "Regular")

sport_hist <- ggplot(sport_scores, aes(x=Score))
sport_hist + geom_histogram(bins = 7) +
  labs(x="Total Points", y="Number of Students") +
  ggtitle("Sports Section Scores")

reg_hist <- ggplot(reg_scores, aes(x=Score))
reg_hist + geom_histogram(bins = 7) +
  labs(x="Total Points", y="Number of Students") +
  ggtitle("Regular Section Scores")

ggplot(scores, aes(x=Score)) +
  geom_histogram(bins=7) +
  facet_wrap(~Section, ncol=1)

stat.desc(sport_scores$Score, basic=FALSE, norm=TRUE)
stat.desc(reg_scores$Score, basic=FALSE, norm=TRUE)
```


Regular
Sports
count
0
2
4
6
200
250
300
350
400
Score

# Housing Data

```
library("readxl")
library("pastecs")
library("ggplot2")
library("plyr")

housing_data <- read_xlsx("data\\week-7-housing.xlsx")
head(housing_data)

colnames(housing_data)[1:2] <- c("sale_date", "sale_price")

#Create at least two new variables
housing_data$row_num <- seq.int(nrow(housing_data))

sale_year <- format(housing_data$sale_date, format = "%Y")
sale_year <- matrix(sale_year)

sale_year <- apply(sale_year, 2, as.numeric)  # Use the apply function on a
variable in your dataset
sale_year <- sale_year[,1]
housing_data$sale_year <- sale_year


#Use the aggregate function on a variable in your dataset
aggregate(square_feet_total_living ~ year_built, housing_data, median)

#Use the plyr function on a variable in your dataset

sum_baths <- function(house_data){
  c(total_baths = house_data$bath_full_count +
                  house_data$bath_half_count * .5 +
                  house_data$bath_3qtr_count * .75)
}

total_baths <- ddply(housing_data, "row_num", sum_baths)
housing_data$total_baths <- total_baths$total_baths

#Check the distributions of the data
stat.desc(housing_data$sale_price[1:5000], basic=FALSE, norm=TRUE)
ggplot(housing_data, aes(sale_price)) + geom_histogram(bins=50,
aes(y=..density..)) +
  stat_function(fun=dnorm, args=list(mean=mean(housing_data$sale_price, na.rm
= TRUE),
                                     sd=sd(housing_data$sale_price,
na.rm=TRUE)))

#Indentify if there are any outliers
# There are some outliers in the sales price data.
# Specifically the prices exceeding two million dollars.
```


2.0e-06
1.5e-06
1.0e-06
5.0e-07
0.0e+00
density
0e+00
1e+06
2e+06
3e+06
4e+06
sale_price