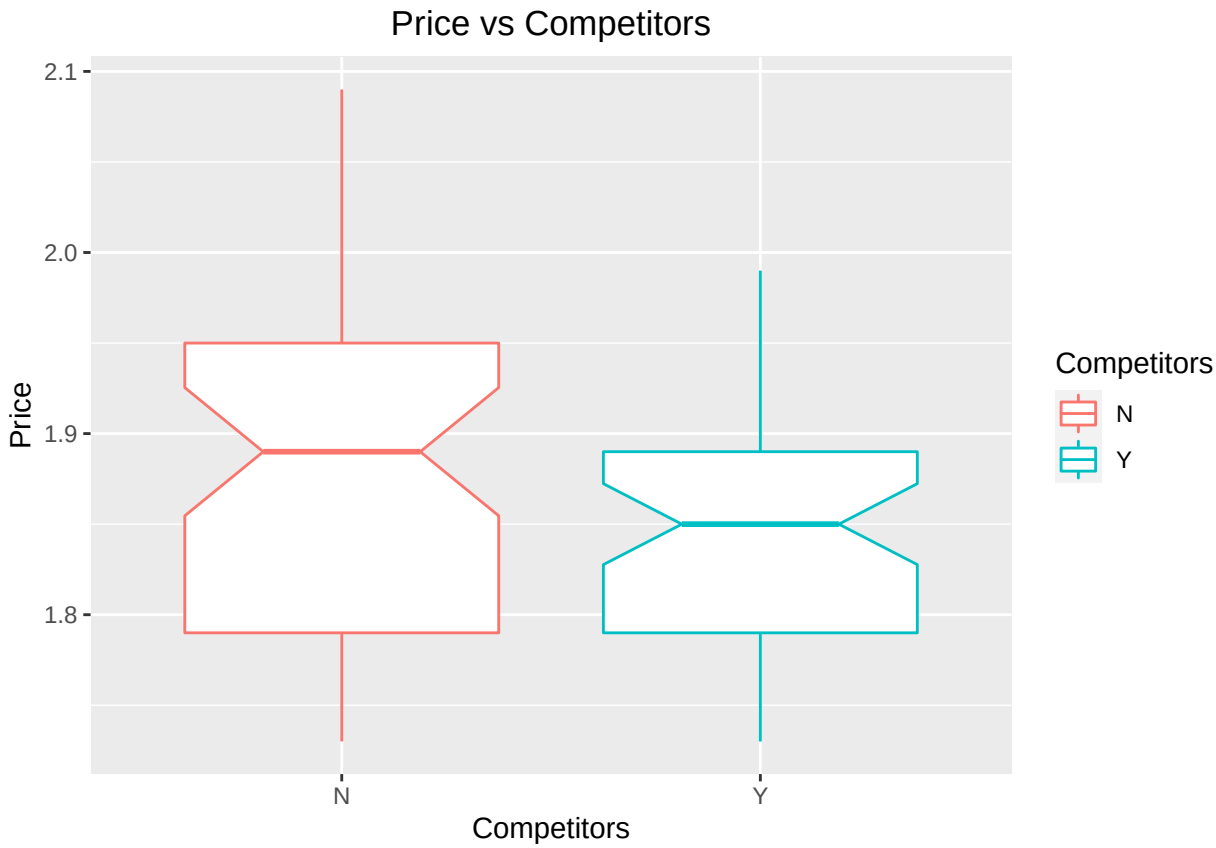
Data Mining Assignment 1

Patrick Chase

2/5/2021

1A)

```
price.comp <- ggplot(GasPrices, aes(x=Competitors, y=Price, color=Competitors)) +
  geom_boxplot(notch = TRUE) + ggtitle("Price vs Competitors")+
  theme(plot.title = element_text(hjust = 0.5))
price.comp
```



Given traditional economic theory, if a station has competitors we would expect a lower average price. "Price vs Competitors" provides evidence that this is true. Gas stations with competitors have both an average lower price, as well as a distribution that is lower than those without competitors.

1B)

```
price.inc <- ggplot(data = GasPrices) +
  geom_point(mapping = aes(x=Income, y=Price, color = Competitors)) +
  ggtitle("Price vs Income") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
price.inc
```

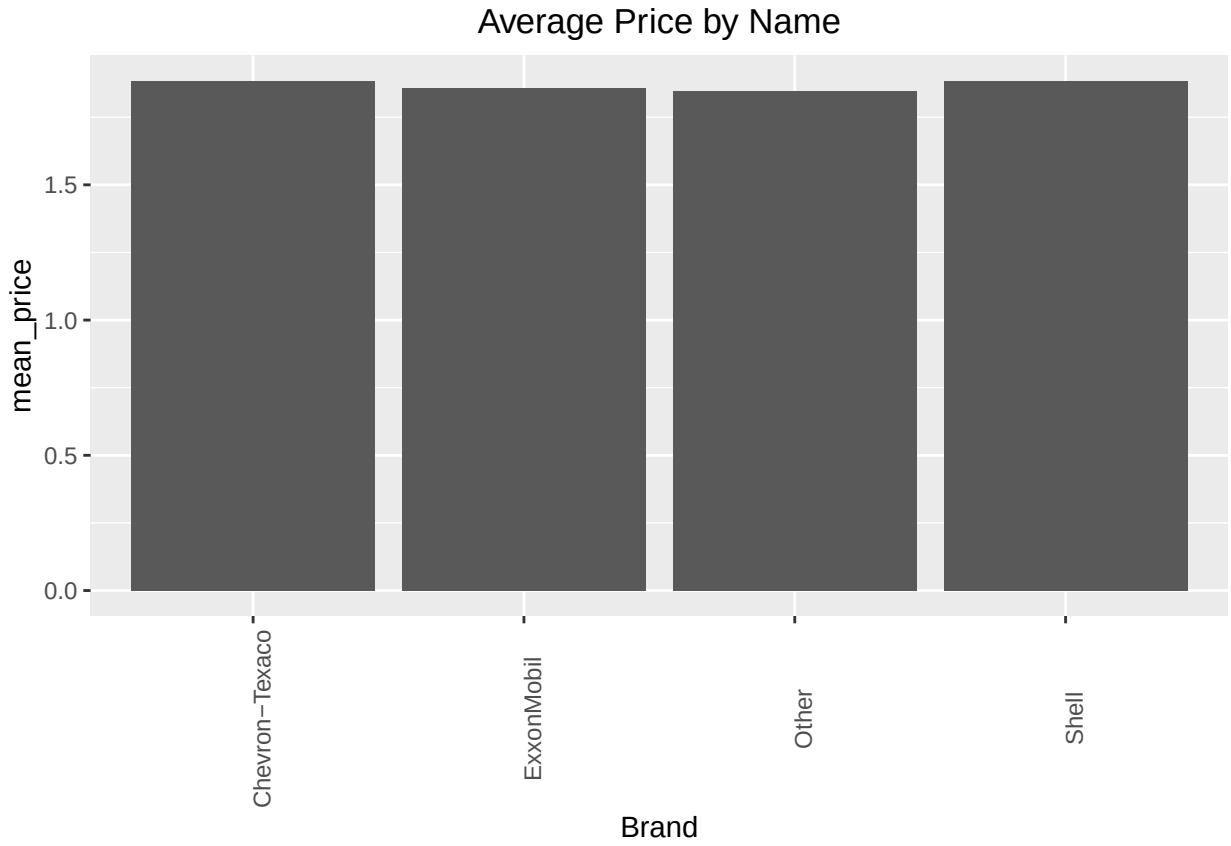
The claim that richer areas tend to have higher gas prices seems to be mildly supported by the available data. On my own, I chose to color each observation based on whether or not there were competitors near by, which shows an interesting relationship. There seems to be less competition at the extremes of income.

1C)

```
brand_price <- GasPrices %>%
  group_by(Brand) %>%
  summarize(mean_price = mean(Price))
brand_price
```

```
## # A tibble: 4 x 2
##   Brand          mean_price
## * <chr>               <dbl>
## 1 Chevron-Texaco       1.88
## 2 ExxonMobil           1.86
## 3 Other                1.85
## 4 Shell                1.88
```

```
ggplot(data = brand_price) +
  geom_col(mapping = aes(x=Brand, y=mean_price),
           position = 'dodge') +
  theme(axis.text.x = element_text(angle = 90)) +
  ggtitle("Average Price by Name")+
  theme(plot.title = element_text(hjust = 0.5))
```



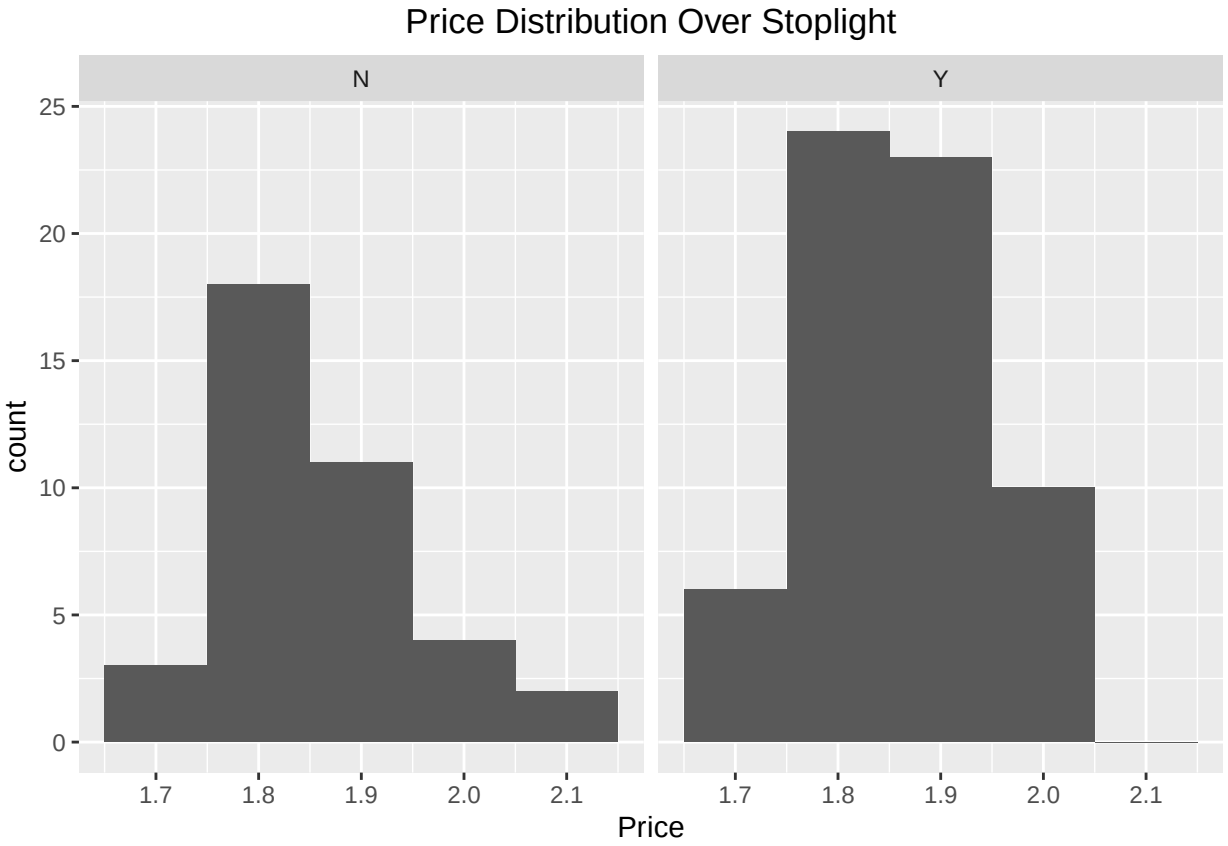
The claim that Shell gas stations charge more than others does not seem to be supported by this data. Visually, it appears that shell is charging about the same average price as all the other stations.

1D)

```
stoplight <- GasPrices %>%
  group_by(Stoplight) %>%
  summarize(mean_price = mean(Price))
stoplight
```

```
## # A tibble: 2 x 2
##   Stoplight mean_price
## * <chr>          <dbl>
## 1 N               1.87
## 2 Y               1.86
```

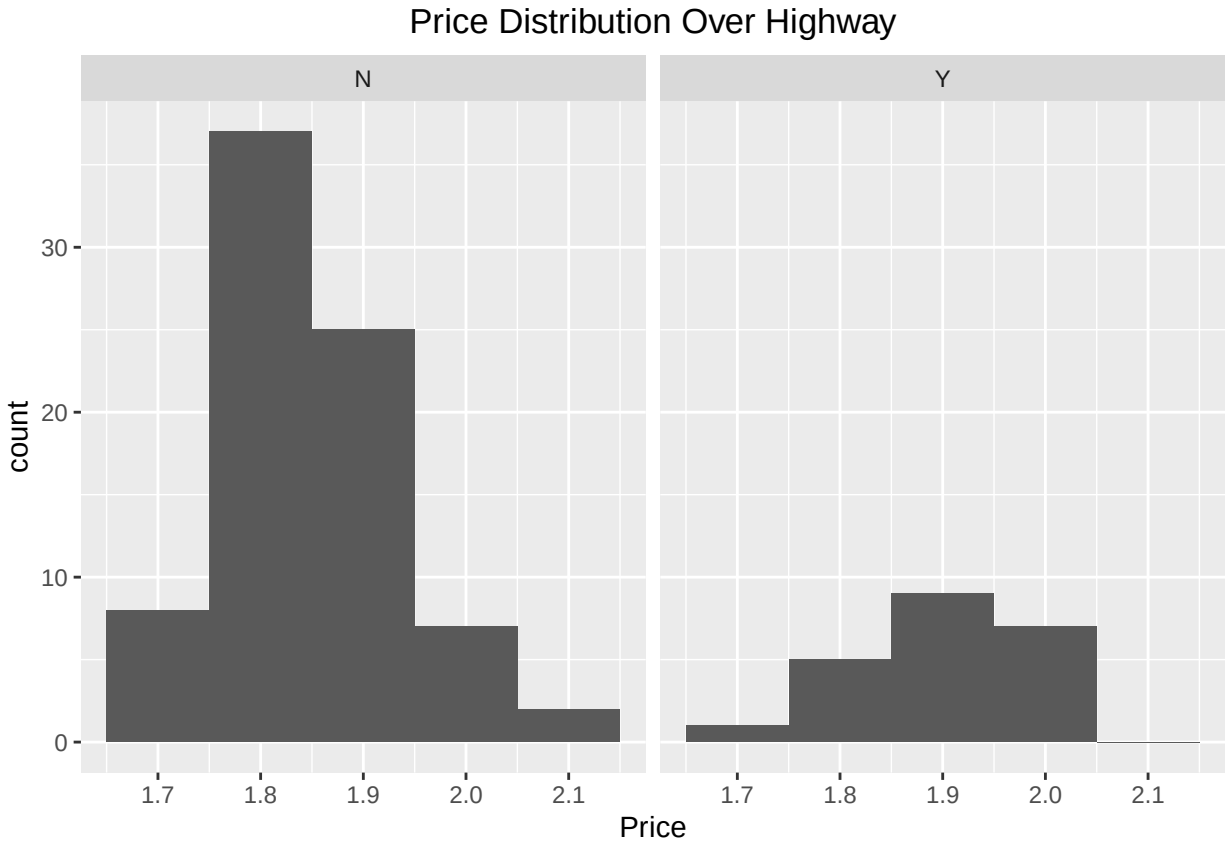
```
ggplot(data = GasPrices) +
  geom_histogram(aes(x=Price), binwidth = .1) +
  facet_wrap(~Stoplight) +
  ggtitle("Price Distribution Over Stoplight")+
  theme(plot.title = element_text(hjust = 0.5))
```



I don't think this visualization supports the claim that gas stations near stoplights charge more for gas. The average price near stoplights is probably higher, however prices that aren't near a stop light have a wider range and a higher max price.

1E)

```
ggplot(data = GasPrices) +
  geom_histogram(aes(x=Price), binwidth = .1) +
  facet_wrap(~Highway) +
  ggtitle("Price Distribution Over Highway")+
  theme(plot.title = element_text(hjust = 0.5))
```



I chose to generate a faceted histogram in order to show the difference in between prices given distance from a highway. Preliminarily, I'd say that there is some evidence that suggests that prices are higher when one is close to a highway. That said, the differing counts between the two indicate that we may have some selection bias. Our sample of stations near the highway may not be representative and as such should be taken with a grain of salt.

2)

```
bikeshare <- read.csv("https://raw.githubusercontent.com/jgscott/ECO395M/master/data/bikeshare.csv")
head(bikeshare)
```

```
##   instant     dteday season yr mnth hr holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1  0       0       6          0          1
## 2       2 2011-01-01      1  0    1  1       0       6          0          1
## 3       3 2011-01-01      1  0    1  2       0       6          0          1
## 4       4 2011-01-01      1  0    1  3       0       6          0          1
## 5       5 2011-01-01      1  0    1  4       0       6          0          1
## 6       6 2011-01-01      1  0    1  5       0       6          0          2
##   temp total
## 1 0.24    16
## 2 0.22    40
## 3 0.22    32
## 4 0.24    13
## 5 0.24     1
## 6 0.24     1
```

Plot A

```
df1 <- bikeshare %>%
  group_by(hr) %>%
```

```
  summarize(totalhr = sum(total),
            avgperhr = totalhr/730)
df1
```

```
## # A tibble: 24 x 3
##       hr totalhr avgperhr
##  * <int>   <int>    <dbl>
##  1     0   39130    53.6
##  2     1   24164    33.1
##  3     2   16352    22.4
##  4     3    8174    11.2
##  5     4    4428     6.07
##  6     5   14261    19.5
##  7     6   55132    75.5
##  8     7  154171   211.
##  9     8  261001   358.
## 10     9  159438   218.
## # ... with 14 more rows
```

```
ggplot(data = df1, aes(x = hr, y = avgperhr)) +
  geom_line(color="blue") +
  ggtitle("Average Total per Hour")+
  theme(plot.title = element_text(hjust = 0.5))
```