Homework #1

Rufus Petrie

Professor Magnolfi

Machine Learning

17 September, 2017

Note: Because the code to clean the data outputted a lot of useless information, I have silenced it in this file. The code is still visible in the script and Rmarkdown files included in this folder.

```
vois = c("mean_price","passenger_count","mean_distance")
kable(summary(carrier_data[vois]),format="markdown")
```

mean_price	passenger_count	mean_distance
Min. : 40.39	Min. : 1.0	Min. : 100.0
1st Qu.: 183.79	1st Qu.: 42.0	1st Qu.: 874.7
Median : 222.22	Median : 177.0	Median :1227.3
Mean : 236.50	Mean : 810.9	Mean :1380.1
3rd Qu.: 274.24	3rd Qu.: 1136.5	3rd Qu.:1825.5
Max. :2403.00	Max. :13612.0	Max. :7437.8

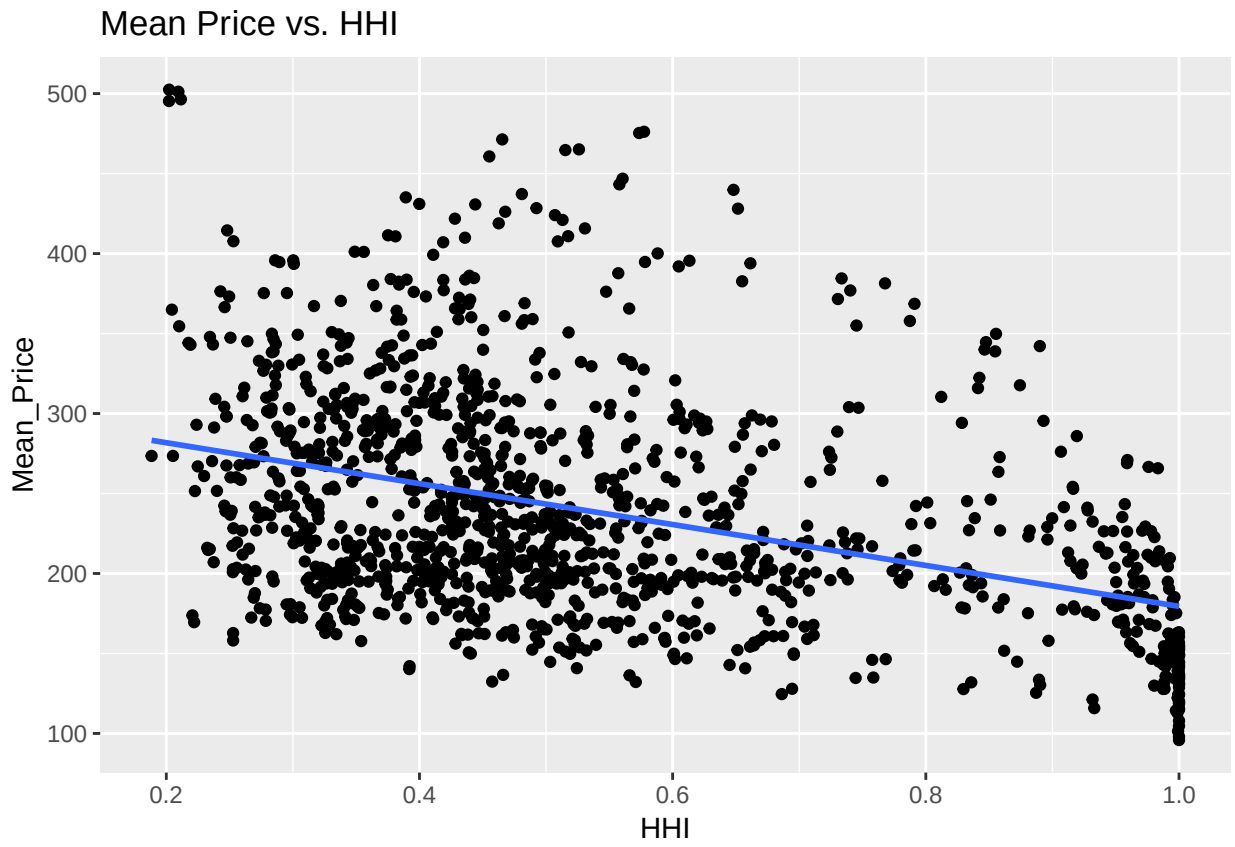
Above are summary statistics for the market-carrier-level data. All of the data looks fairly straightforward except for the passenger count data, where the median is about one seventh of the mean. This indicates that most passengers fly on a few routes, and there are many routes with smaller amounts of passengers.

```
VOIS=c("Mean_Price","Mean_Distance","Firms","HHI","market_size")
kable(summary(market_data[VOIS]),format="markdown")
```

Mean_Price	Mean_Distance	Firms	HHI	market_size
Min. : 95.88	Min. : 100.0	Min. : 1.000	Min. :0.1884	Min. : 199497
1st Qu.:192.61	1st Qu.: 632.5	1st Qu.: 4.000	1st Qu.:0.3731	1st Qu.: 2534375
Median :225.63	Median :1023.4	Median : 6.000	Median :0.4725	Median : 3844297
Mean :239.05	Mean :1125.4	Mean : 5.427	Mean :0.5340	Mean : 4422687
3rd Qu.:281.47	3rd Qu.:1479.9	3rd Qu.: 7.000	3rd Qu.:0.6427	3rd Qu.: 5577746
Max. :502.45	Max. :2959.7	Max. :11.000	Max. :1.0000	Max. :16338394
NA	NA	NA	NA	NA's :21

Above are summary statistics for the market-level data. There is nothing too alarming about this data except for some missing values.

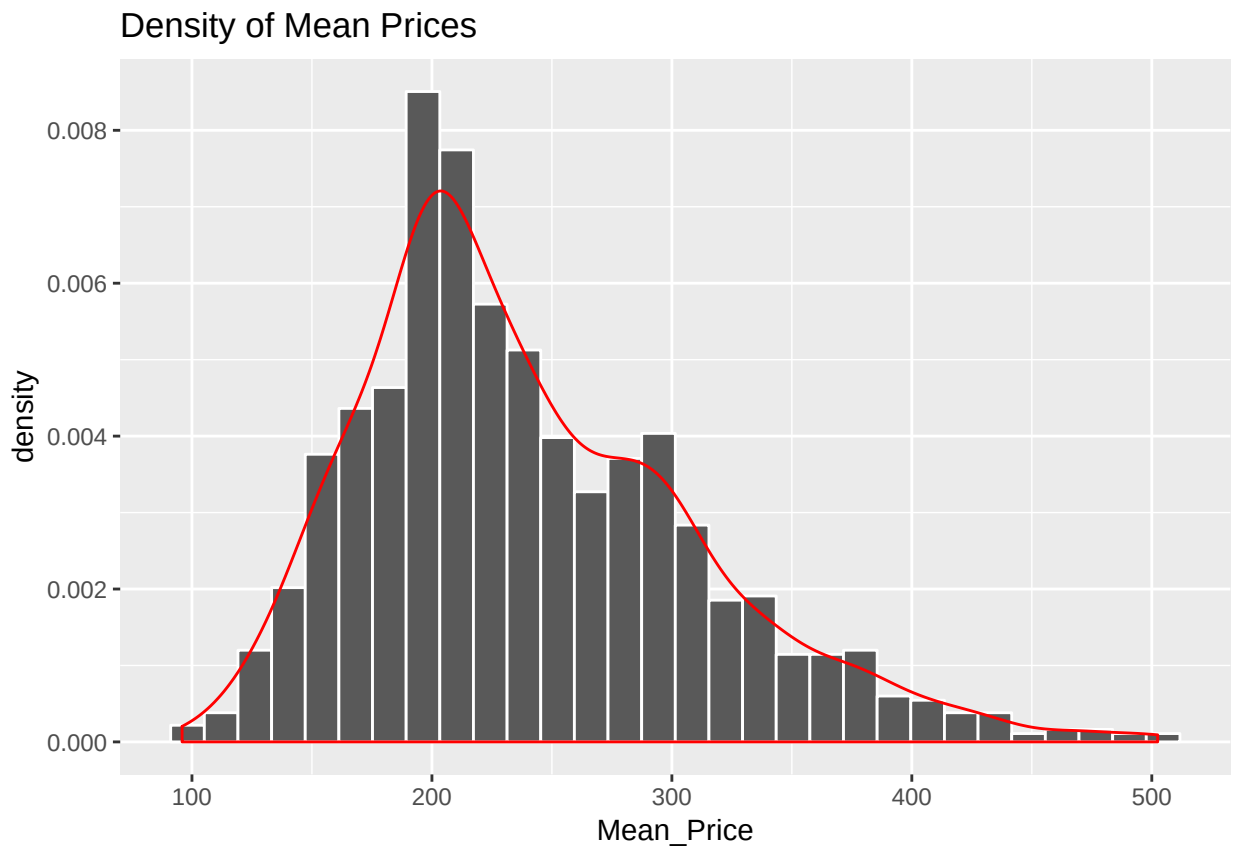
```r
ggplot(market_data, aes(x=HHI, y=Mean_Price)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE) +
  ggtitle("Mean Price vs. HHI")
```



The plot above shows the relationship between mean prices and the HHI in the data. Notice that as HHI increases, prices start to decrease. This could be due to a variety of reasons, including endogeneity, niche markets, or competitive pricing.

```
ggplot(market_data) +
  geom_histogram(aes(x=Mean_Price,y=..density..), position="identity",color="white") +
  geom_density(aes(x=Mean_Price,y=..density..),color="red") +
  ggtitle("Density of Mean Prices")
```

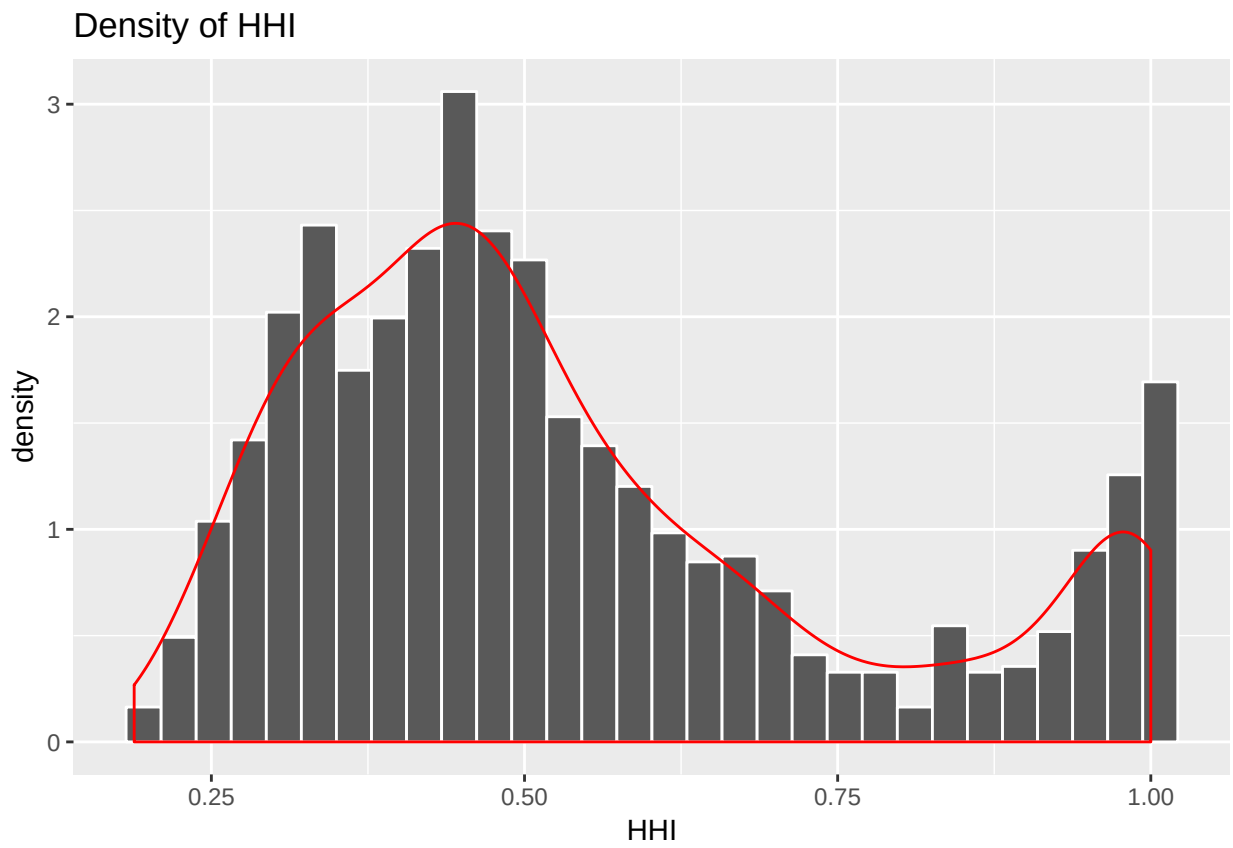
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot above shows the density of mean prices in the market-level data. Notice that the density looks skewed and has a fat tail containing a bunch of high-priced tickets.

```
ggplot(market_data) +
  geom_histogram(aes(x=HHI,y=..density..), position="identity",color="white") +
  geom_density(aes(x=HHI,y=..density..),color="red") +
  ggtitle("Density of HHI")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot above has a histogram of HHI values for the market level data and a geometric density line. Notice that the density increases until about 0.5, decreases until about 0.8, and then increases again. Once again, this could be due to niche markets not being able to support multiple firms or some sort of competitive pricing.