Erin Werner Shift Data Challenge - Data Analyst

```
porscheListings <- read.csv("~/Downloads/porscheListings.csv")
```

1.) What are the different Porsche "models" (eg. Porsche Cayenne, Porsche Panamera, etc) contained within the dataset? Which is the most common?

```
different_models <- sort(unique(porscheListings$model))
different_models
```

```
## [1] 718 Boxster 718 Cayman  911         Boxster     Cayenne     Cayman
## [7] Macan       Panamera
## 8 Levels: 718 Boxster 718 Cayman 911 Boxster Cayenne Cayman ... Panamera
```

```
most_common <- sort(table(porscheListings$model),decreasing=TRUE)[1]
most_common
```

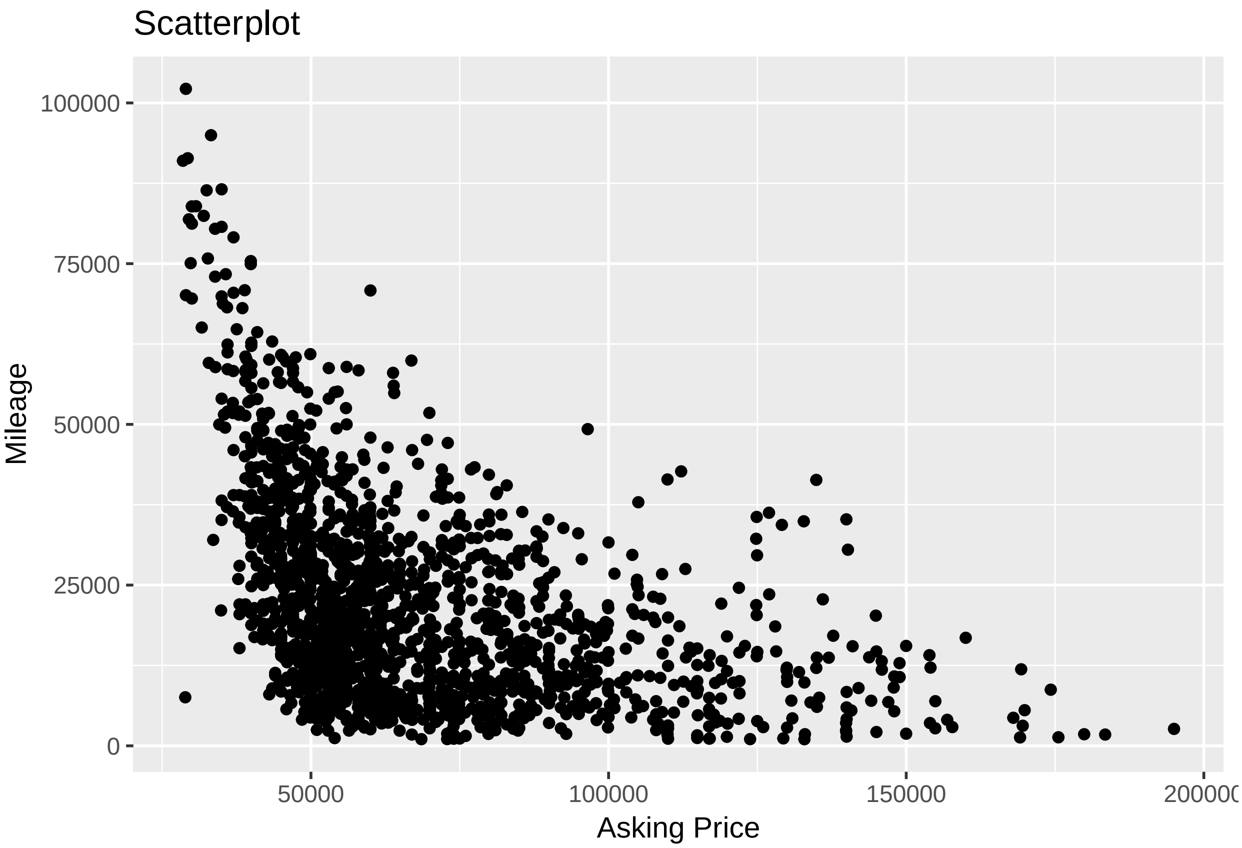
```
## Cayenne
##     630
```

2.) What can you say quantitatively about the relationship between asking price and mileage?

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

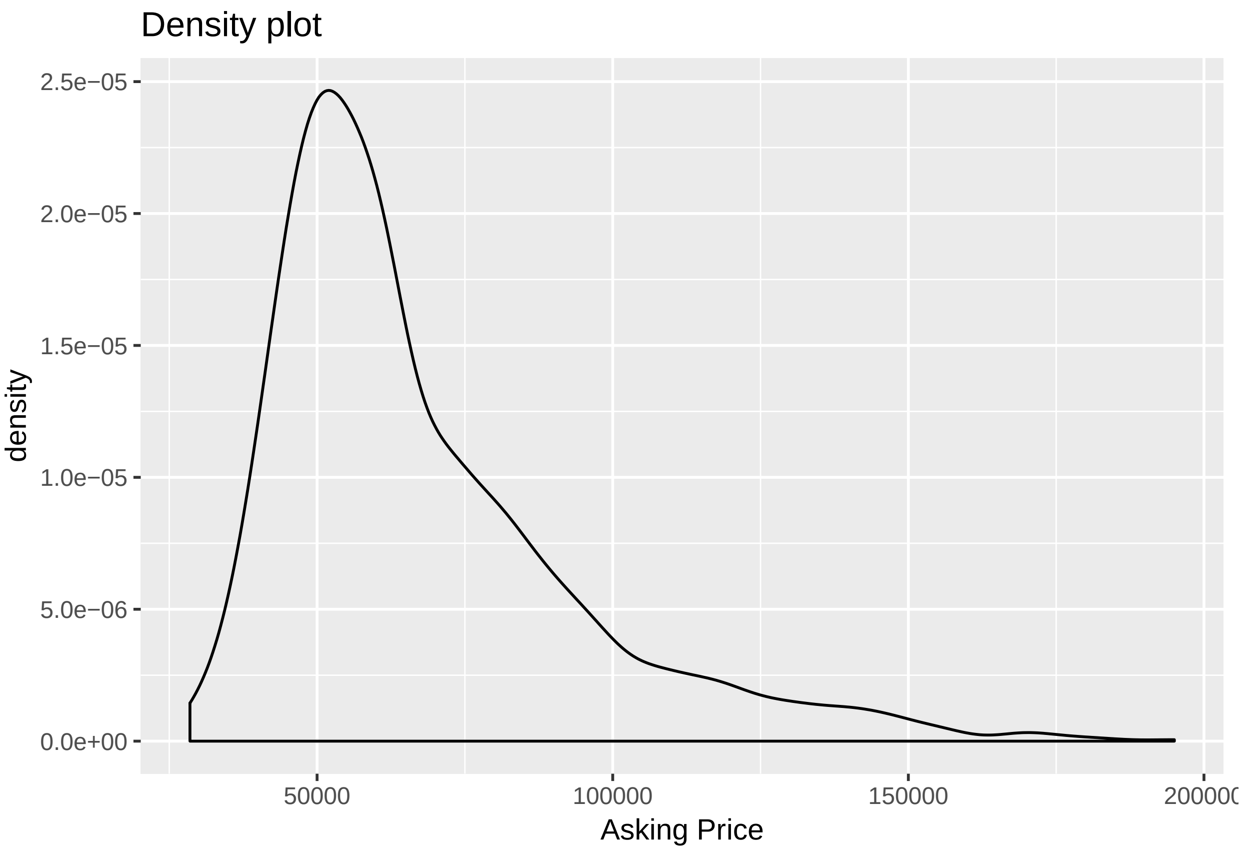
```
gg <- ggplot(porscheListings, aes(x=asking_price, y=mileage)) +
  geom_point() +
  labs(y="Mileage",
       x="Asking Price",
       title="Scatterplot")
gg
```



The scatter plot shows that there is a strong negative linear relationship between the two features. The asking price and the mileage thus have a negative correlation. So, the higher the mileage, the lower the asking price will be, and vice versa.

3.) What does the distribution of "asking_price" look like? How does the asking price vary across different models?

```r
gp <- ggplot(porscheListings, aes(asking_price)) +
  geom_density(alpha=0.8) +
  labs(title="Density plot",
       x="Asking Price")
gp
```



```
mean(porscheListings$asking_price)
```

```
## [1] 67426.77
```

```
median(porscheListings$asking_price)
```

```
## [1] 59990
```

```
quantile(porscheListings$asking_price)
```

```
##       0%      25%      50%      75%     100%
##  28500.0  49900.0  59990.0  78998.5 194993.0
```

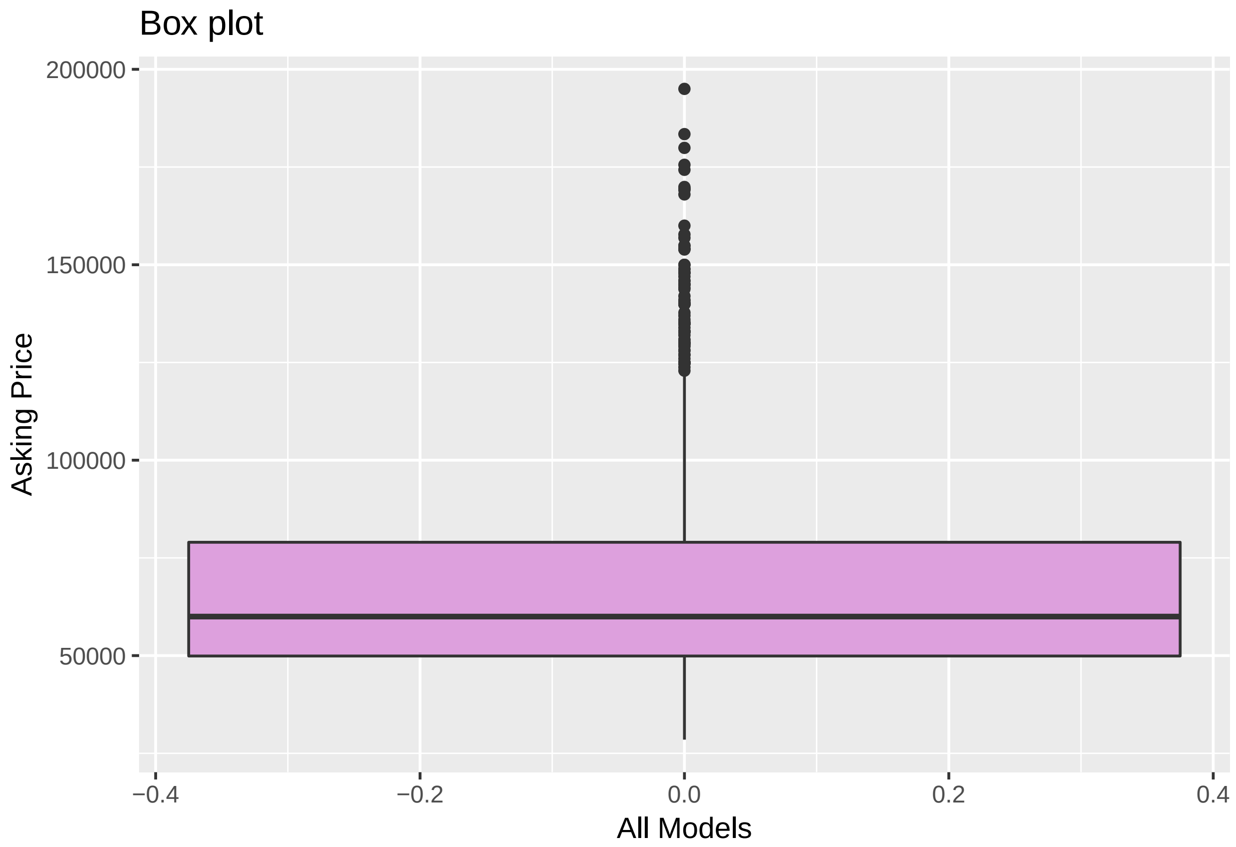
```
sd(porscheListings$asking_price)
```

```
## [1] 25697.99
```

```
asking_price_mode <- sort(table(porscheListings$asking_price),decreasing=TRUE)[1]
asking_price_mode
```

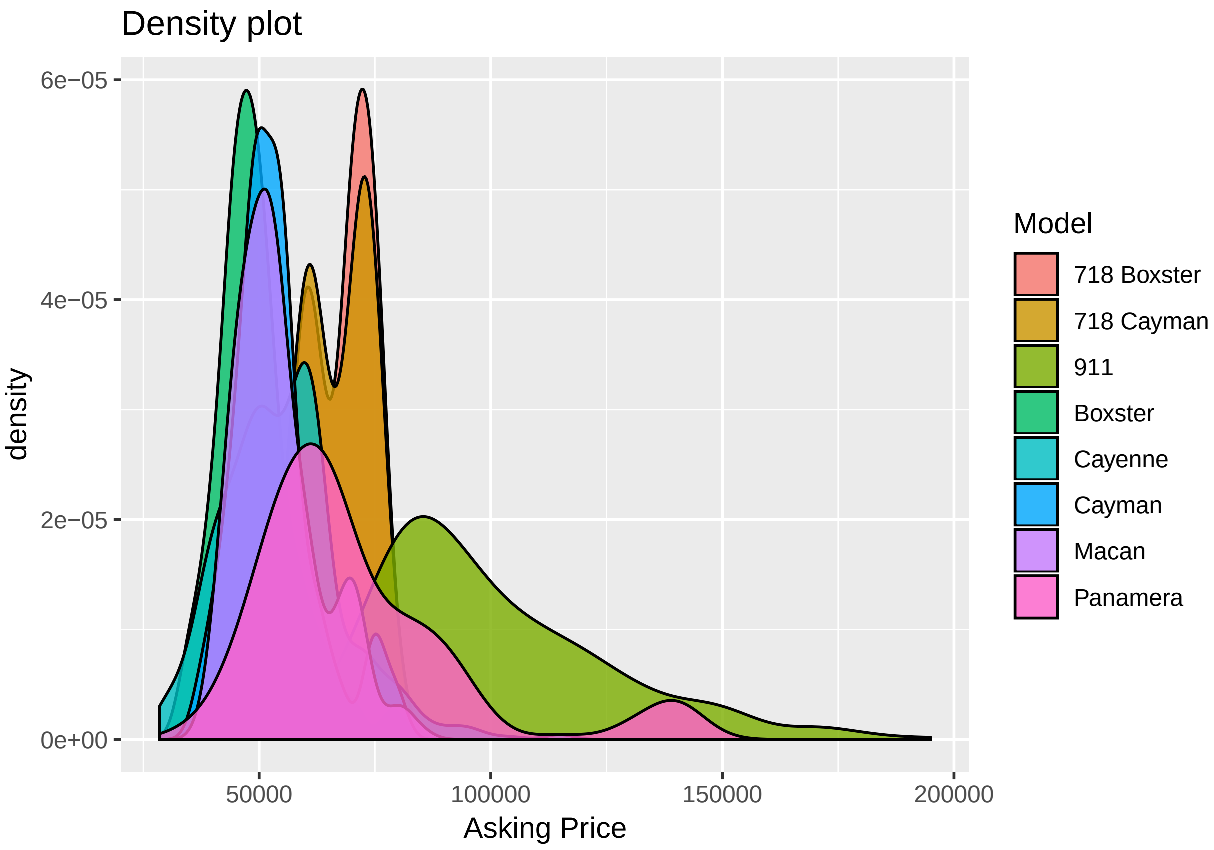
```
## 52995
##    15
```

```
gbox1 <- ggplot(porscheListings, aes(y = asking_price)) +
  geom_boxplot(varwidth=T, fill="plum") +
  labs(title="Box plot",
       x = "All Models",
       y="Asking Price")
gbox1
```

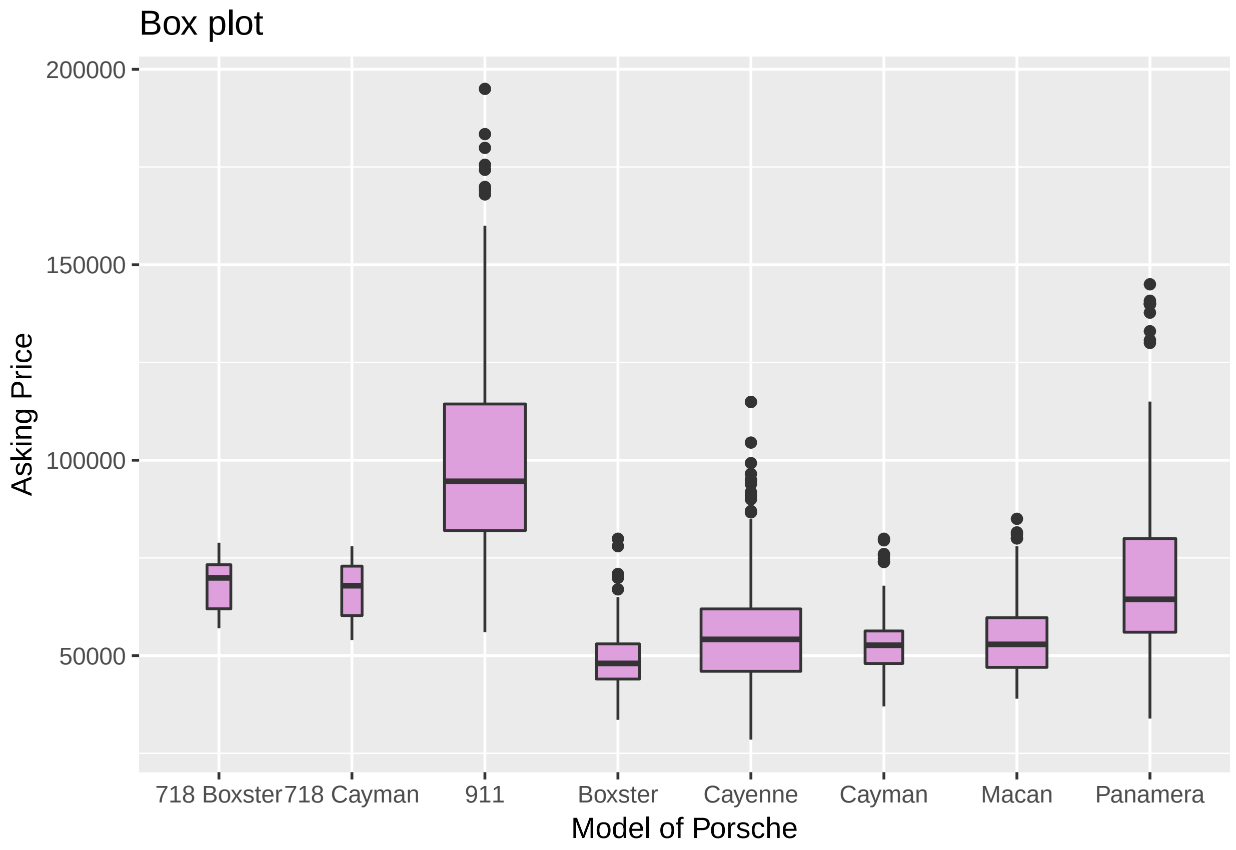


The distribution of asking prices is skewed to the right with a long tail. The distribution is unimodal, centered around the mean, which is $67426.77. The mean however, is slightly higher than both the median, $59990, and the mode, $52995. This confirms the positive right skewness of the distribution. The distribution has a standard deviation of $25697.99. The asking prices vary from about $28500 to $194993, with a number of outlying values with higher asking prices. This is because there are values beyond the 3rd quantile ($78998.50), which is more clear in the boxplot.

```
g <- ggplot(porscheListings, aes(asking_price)) +
  geom_density(aes(fill=factor(model)), alpha=0.8) +
  labs(title="Density plot",
       x="Asking Price",
       fill="Model")
g
```



```
gbox2 <- ggplot(porscheListings, aes(model, asking_price)) +
  geom_boxplot(varwidth=T, fill="plum") +
  labs(title="Box plot",
       x="Model of Porsche",
       y="Asking Price")
gbox2
```



The asking price distributions vary across the different models. Most of the plots retain a positive right skew. Some distributions stay unimodal, but models, like the 718 Boxster and the 718 Cayman, become bimodal. Most of the distribtions are still centered around the mean, but the 911 has an individual mean that is higher and has a longer tail. So, it is more common for the 911 model to have a higher asking price compared to the other models.

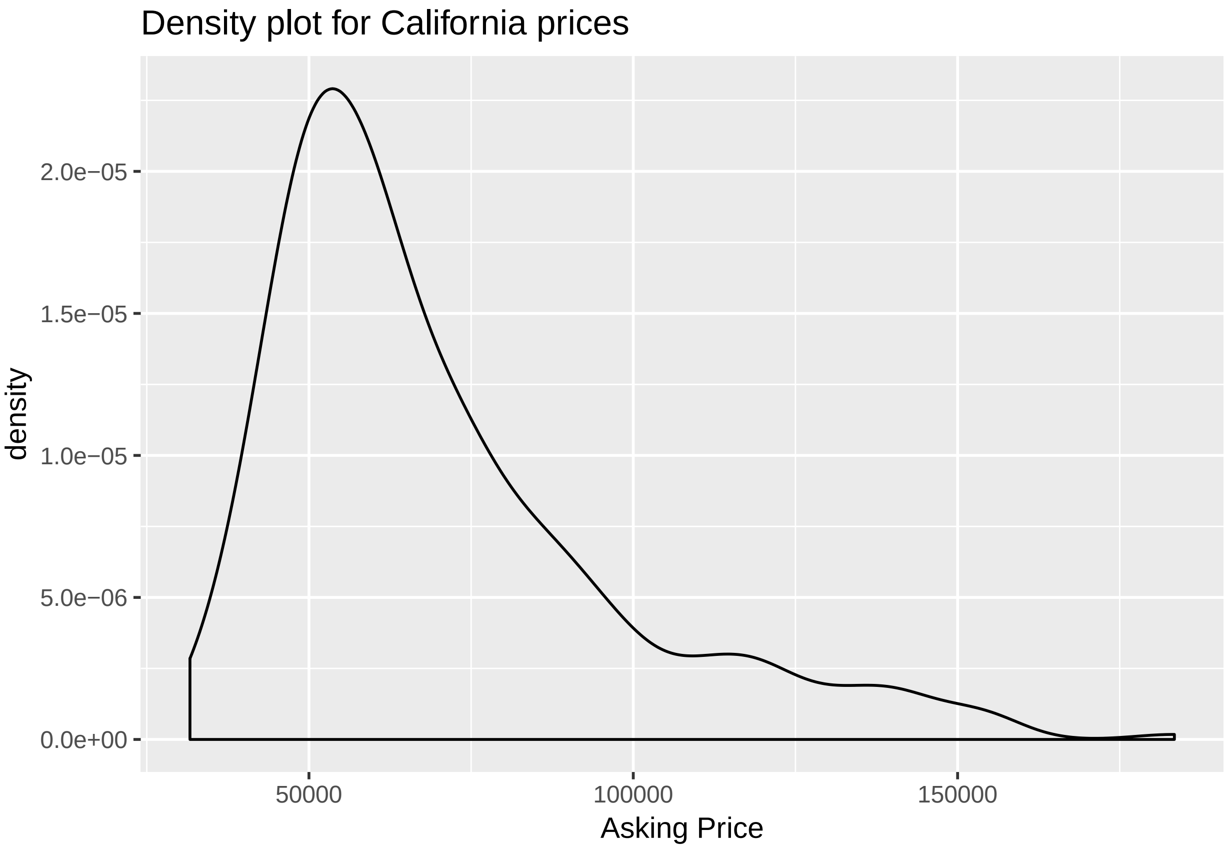
4.) How do the prices in California compare with the national prices?

```
sort(unique(porscheListings$state))
```

```
##  [1] AL         AZ         CA         CALIFORNIA CO         CT
##  [7] FL         GA         IA         ID         IL         IN
## [13] KS         LA         MA         MD         MI         MN
## [19] MO         MS         NC         NE         NH         NJ
## [25] NM         NV         NY         OH         OK         OR
## [31] PA         RI         SC         TN         TX         UT
## [37] VA         VT         WA         WI
## 40 Levels: AL AZ CA CALIFORNIA CO CT FL GA IA ID IL IN KS LA MA MD ... WI
```

```
ca <- porscheListings[which(porscheListings$state == "CA"
                            | porscheListings$state == "CALIFORNIA"),]
```

```
g_ca <- ggplot(ca, aes(asking_price)) +
  geom_density(alpha=0.8) +
  labs(title="Density plot for California prices",
       x="Asking Price")
g_ca
```



```
mean(ca$asking_price)
```

```
## [1] 69827.13
```

```
median(ca$asking_price)
```

```
## [1] 61898
```

```
quantile(ca$asking_price)
```

```
##        0%       25%       50%       75%      100%
##  31657.00  51104.50  61898.00  80624.75 183425.00
```

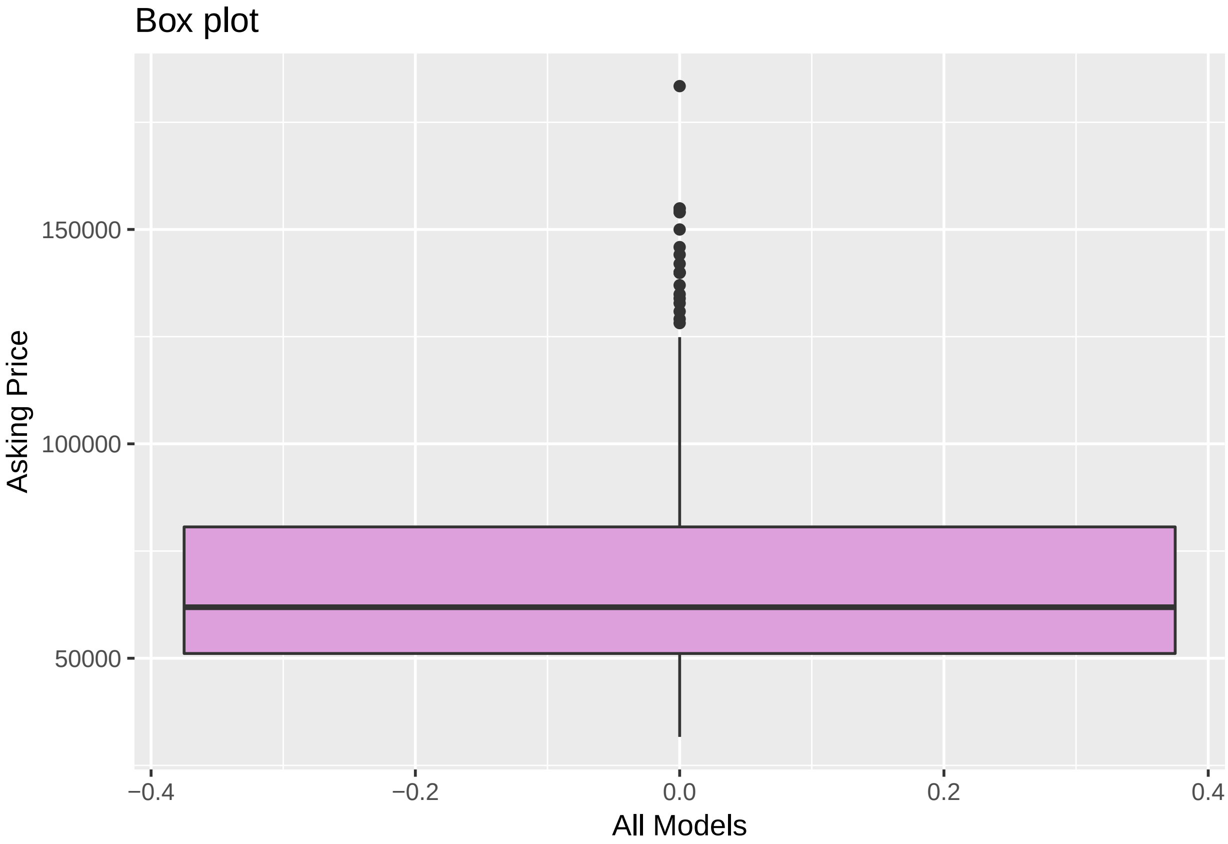
```
sd(ca$asking_price)
```

```
## [1] 26669.59
```

```
asking_price_mode_ca <- sort(table(ca$asking_price),decreasing=TRUE)[1]
asking_price_mode_ca
```

```
## 41998
##     4
```

```
gboxca <- ggplot(ca, aes(y = asking_price)) +
  geom_boxplot(varwidth=T, fill="plum") +
  labs(title="Box plot",
       x = "All Models",
       y="Asking Price")
gboxca
```



The distribution of asking prices is skewed to the right with a long tail, just like the national distribution. The distribution is also unimodal and centered around the mean, $669827, which is just slightly higher than the national mean. The CA mean however, is still higher than both the median ($61898) and the mode ($41998). This confirms the positive right skewness of the distribution. The distribution has a standard deviation of $26669.59, which is slightly higher than the national value. The asking prices vary from about $31657 to $183425, so the minimum price of a car is higher in California and the max price is only a little bit lower. There are still a number of outlying values with higher asking prices. This is because there are values beyond the 3rd quantile ($80624.75), which is more apparent in the boxplot. The value for the 3rd quantile is also higher compared to the national distribution.

In general, the asking price of a car in California will be higher compared to national prices.

5.) We ran an A/B test on our website in an attempt to target and acquire more Porsches. Given that A was the control and B was the experiment, what do the results say about the proposed change?

```
table(porscheListings$experiment_group)
```

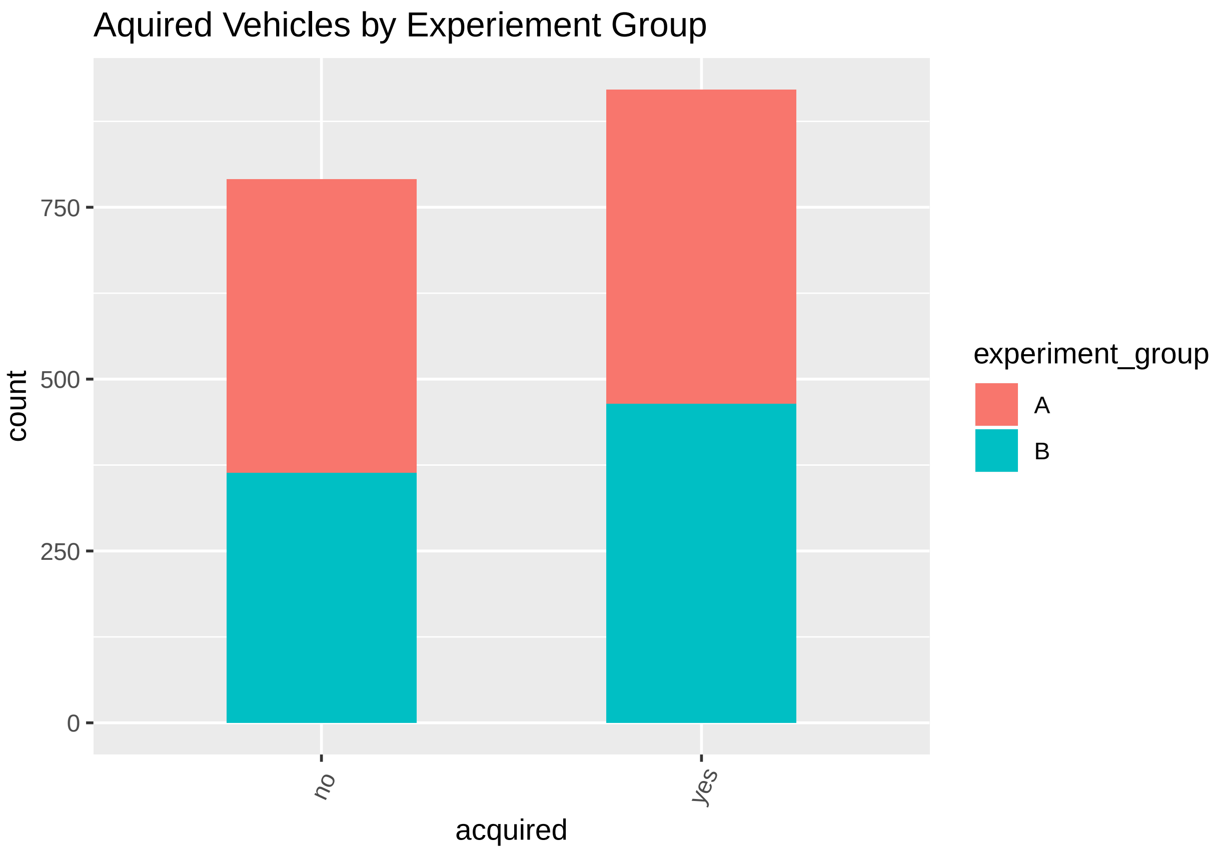
```
##
```

```
##   A   B
## 884 828
```

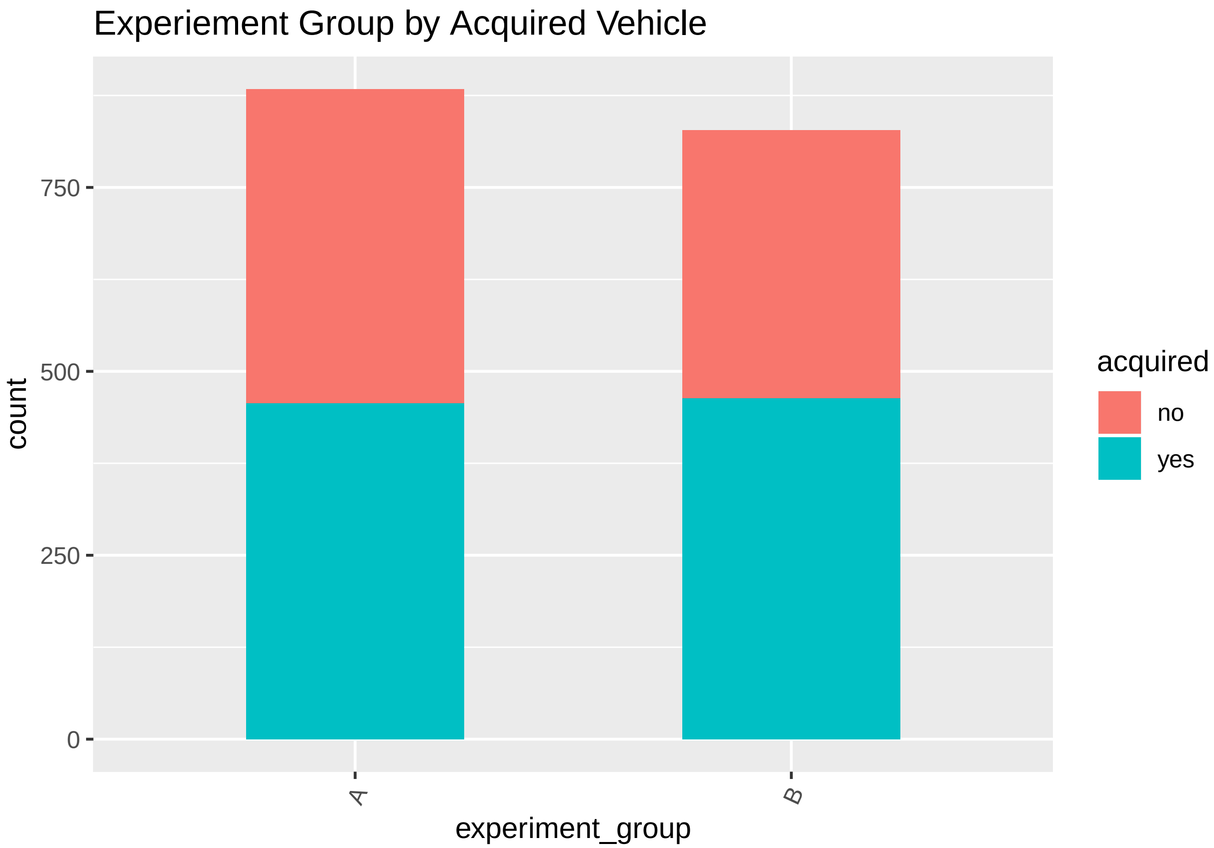
```
table(porscheListings$acquired)
```

```
##
##  no yes
## 791 921
```

```
ab <- ggplot(porscheListings, aes(acquired)) +
  geom_bar(aes(fill=experiment_group), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Aquired Vehicles by Experiement Group")
ab
```



```
ab_p <- ggplot(porscheListings, aes(experiment_group)) +
  geom_bar(aes(fill=acquired), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Experiement Group by Acquired Vehicle")
ab_p
```



```
acquired <- porscheListings[which(porscheListings$acquired=="yes"),]
table(acquired$experiment_group)
```

```
##
##   A   B
## 457 464
```

```
a_acq <- 457/nrow(acquired)
a_acq
```

```
## [1] 0.4961998
```

```
b_acq <- 464/nrow(acquired)
b_acq
```

```
## [1] 0.5038002
```

Rougly the same amount of acquired cars came from both A and B groups.

```
nonacquired <- porscheListings[which(porscheListings$acquired=="no"),]
table(nonacquired$experiment_group)
```

```
##
##   A   B
## 427 364
```

```
a_nacq <- 427/nrow(nonacquired)
a_nacq
```

```
## [1] 0.539823
```

```
b_nacq <- 364/nrow(nonacquired)
b_nacq
```

```
## [1] 0.460177
```

There are slightly more nonacquired cars from the A group comapred to the B group.

```
A <- porscheListings[which(porscheListings$experiment_group=="A"),]
table(A$acquired)
```

```
##
##  no yes
## 427 457
```

```
Ayes <- 457/nrow(A)
Ayes
```

```
## [1] 0.5169683
```

```
Ano <- 427/nrow(A)
Ano
```

```
## [1] 0.4830317
```

```
B <- porscheListings[which(porscheListings$experiment_group=="B"),]
table(B$acquired)
```

```
##
##  no yes
## 364 464
```

```
Byes <- 464/nrow(B)
Byes
```

```
## [1] 0.5603865
```

```
Bno <- 364/nrow(B)
Bno
```

```
## [1] 0.4396135
```

Depending on how significant the reults need to be, group B (the experiment group) does slightly better than group A. Of the acquired cars, 50% came from each group. Of the nonacquired cars, 54% came from A, indicating that control group A was not able to acquire more Porches than group B. Within group A, 52% of the cars were aquired. But in group B, 56% of the cars were acuired. As a result, experiment group B performed slightly better than control group A.

This means that the proposed changes helped to target and acuire more Porches.

6.) Let's say that this information was contained in a SQL table called porsche_listings , and we wanted to see the unique horsepower values for each year and model. The table that has this information is called engines and the first 5 rows are shown below. Write a SQL query that would get this information.

- porsche_listings has year, make, style_id, model, etc.

- engines has style_id, horsepower, make

SELECT DISTINCT engines.horsepower, porsche_listings.year, porsche_listings.model

FROM engines

JOIN porsche_listings ON porsche_listings.style_id = engines.style_id

7.) Shift currently only buys and sells cars in California, but we're looking to expand into new markets. Given this dataset of Porsche listings, what state might you recommend Shift expands to next, and why? What other data would you like to have in order to make this decision?

```
table(porscheListings$state)
```

```
##
##         AL         AZ         CA CALIFORNIA         CO         CT
##         11         13        334         36         19         56
##         FL         GA         IA         ID         IL         IN
##        263         79          2          6        111          6
##         KS         LA         MA         MD         MI         MN
##          2          6         56         18         16         14
##         MO         MS         NC         NE         NH         NJ
##         29          1         43          2          3         83
##         NM         NV         NY         OH         OK         OR
##          1         14         87         99         14          2
##         PA         RI         SC         TN         TX         UT
##         96         10         42         48         27          1
##         VA         VT         WA         WI
##         21          1         30         10
```

```
sort(table(porscheListings$state),decreasing=TRUE)[1:4]
```

```
##
##  CA  FL  IL  OH
## 334 263 111  99
```

So, behind CA, FL, IL, and OH are the next biggest markets to possibly pursue.

```
fl <- porscheListings[which(porscheListings$state=="FL"),]
table(fl$acquired)
```

```
##
##  no yes
## 121 142
```

```
max(fl$asking_price)
```

```
## [1] 179900
```

```
min(fl$asking_price)
```

```
## [1] 29800
```

```
mean(fl$asking_price)
```

```
## [1] 68293.15
```

```
mean(fl$mileage)
```

```
## [1] 19836.14
```

```
il <- porscheListings[which(porscheListings$state=="IL"),]
table(il$acquired)
```

```
##
##   no yes
##   44  67
```

```
max(il$asking_price)
```

```
## [1] 160000
```

```
min(il$asking_price)
```

```
## [1] 29000
```

```
mean(il$asking_price)
```

```
## [1] 61683.98
```

```
mean(il$mileage)
```

```
## [1] 24524.7
```

```
oh <- porscheListings[which(porscheListings$state=="OH"),]
table(oh$acquired)
```

```
##
##   no yes
##   52  47
```

```
max(oh$asking_price)
```

```
## [1] 194993
```

```
min(oh$asking_price)
```

```
## [1] 28991
```

```
mean(oh$asking_price)
```

```
## [1] 68651.32
```

```
mean(oh$mileage)
```

```
## [1] 24206.82
```

FL has the most cars. 54% of those cars were acquired. The average asking price for cars in FL is $68,293.15. FL also has the lowest average mileage of the three at 19836.14.

IL has the next most cars. 60% of those cars were acquired. The average asking price for cars in IL is $61,683.98. IL has the highest average mileage at 24524.7.

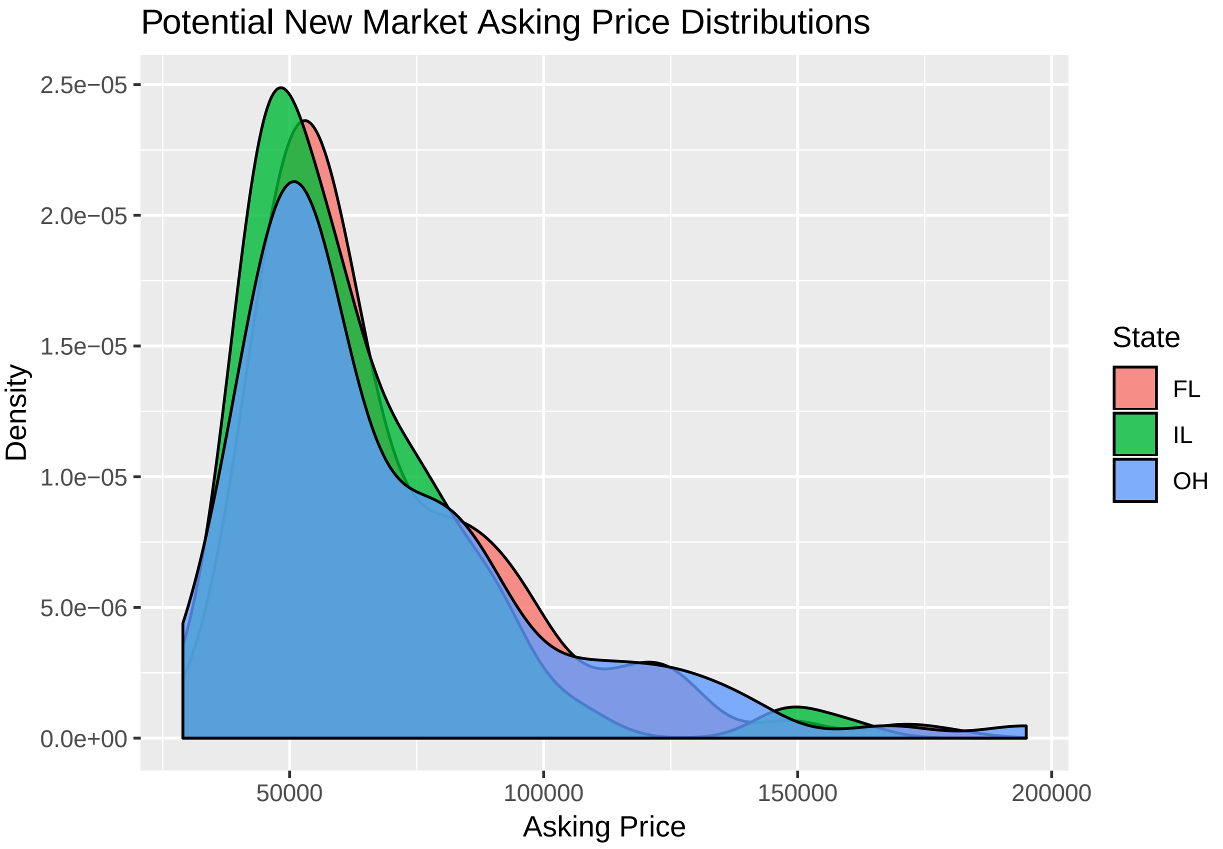
OH has slightly less cars. 47% of those cars were acquired. The average asking price for cars in OH is $68,651.32. OH has an average mileage of 24206.82

Overall, FL looks like the most profitable option of the three states. Although IL acquires more of its cars, FL has a higher average asking price. Even though OH has an asking price that is $358.17 greater than FL, OH acquires much less of its cars. FL has the largest number of cars of these three states and it has the second most cars out of the entire dataset. FL also has the best ratio of average mileage to average asking price, with a high average asking price and a low average mileage.

Plots can help to reaffirm this notion.

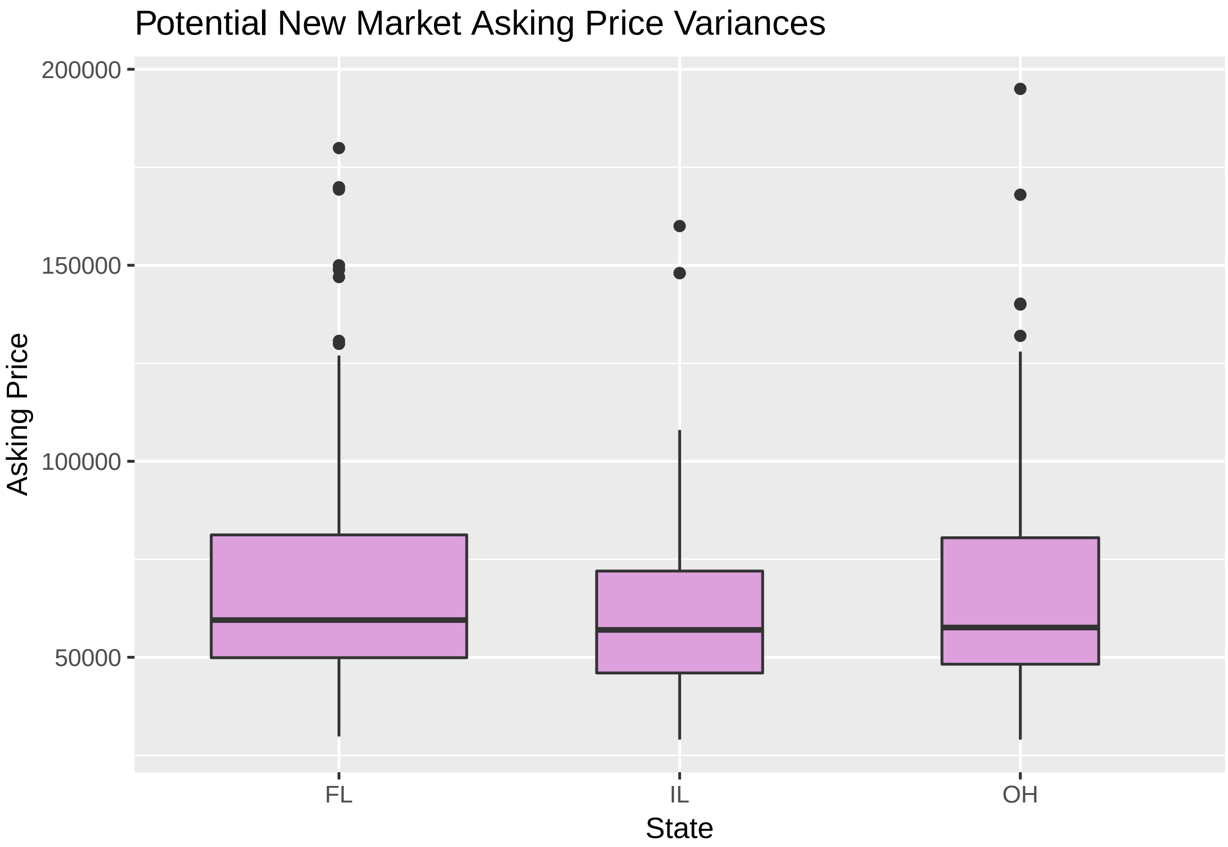
```
pot_new_market <- porscheListings[which(porscheListings$state=="FL" |
                                         porscheListings$state=="IL" |
                                         porscheListings$state=="OH"),]
```

```
ggplot(pot_new_market, aes(asking_price)) +
  geom_density(aes(fill=state), alpha=0.8) +
  labs(title="Potential New Market Asking Price Distributions",
       y = "Density",
       x="Asking Price",
       fill="State")
```



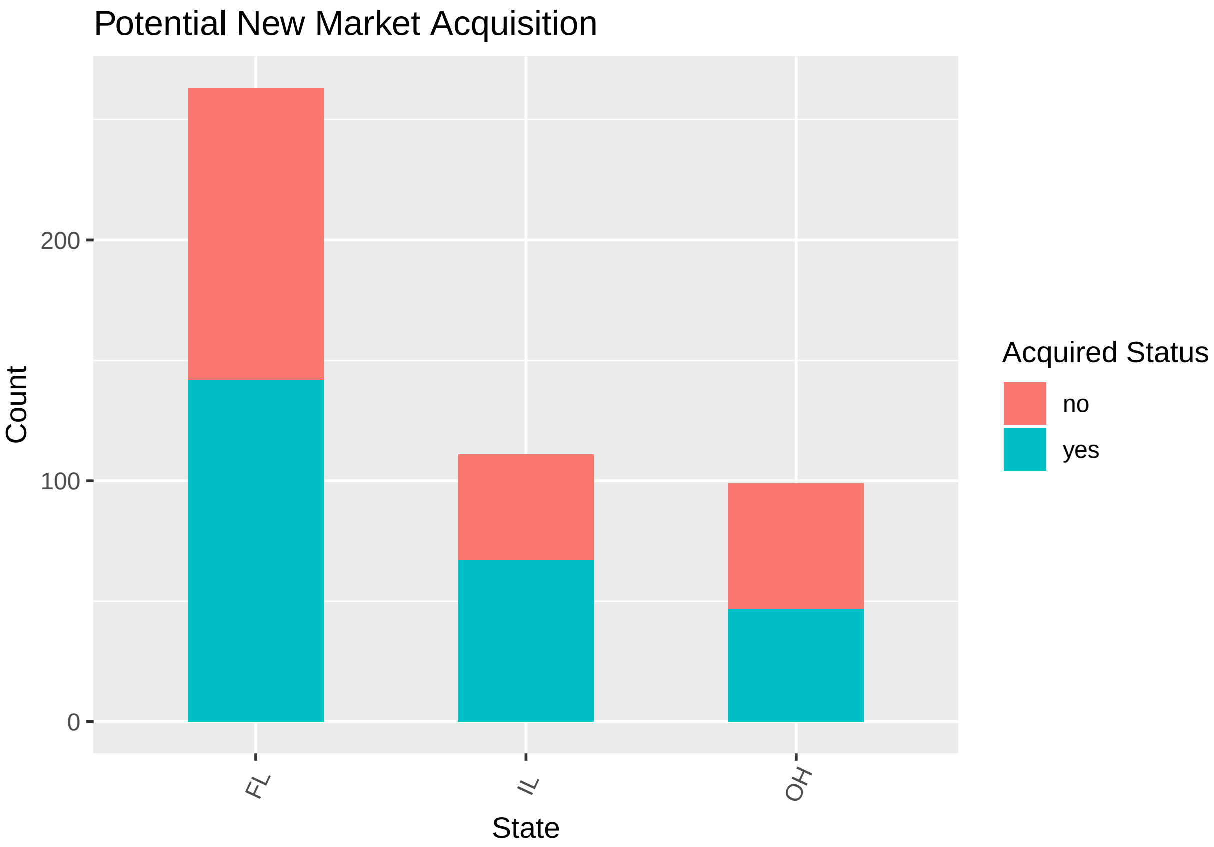
All three states have roughly the same distribution of asking prices. They are all skewed to the right and have long tails. IL has the tallest peak, but it is centered around a mean that is slightly lower than that of FL.

```
ggplot(pot_new_market, aes(state, asking_price)) +
  geom_boxplot(varwidth=T, fill="plum") +
  labs(title="Potential New Market Asking Price Variances",
       x="State",
       y="Asking Price")
```



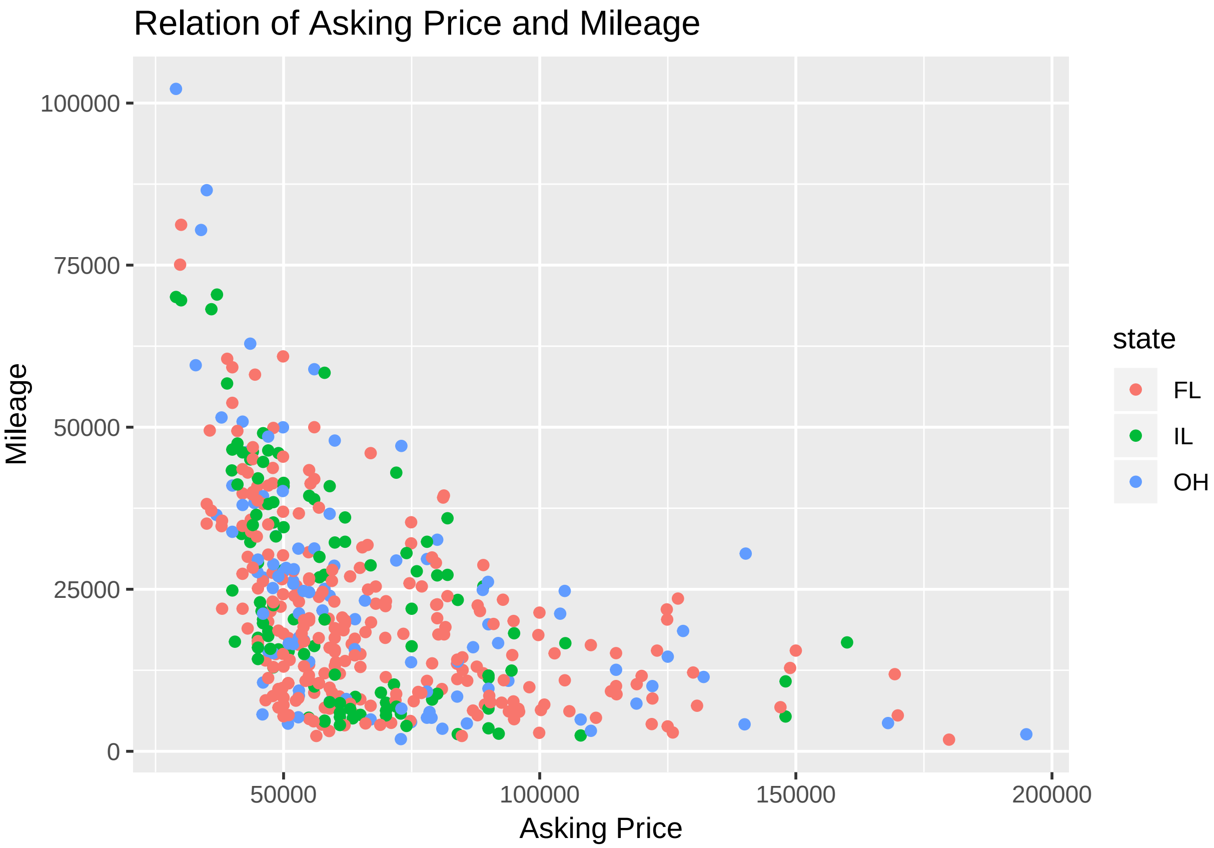
All three states have a similar spread. IL, however, has a slightly smaller spread which indicates that the asking prices tend to stay within a smaller range. Each state has a few outliers, but OH's outliers are a bit more extreme and spread out.

```r
ggplot(pot_new_market, aes(state)) +
  geom_bar(aes(fill=acquired), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Potential New Market Acquisition",
       x = "State",
       y = "Count",
       fill = "Acquired Status")
```



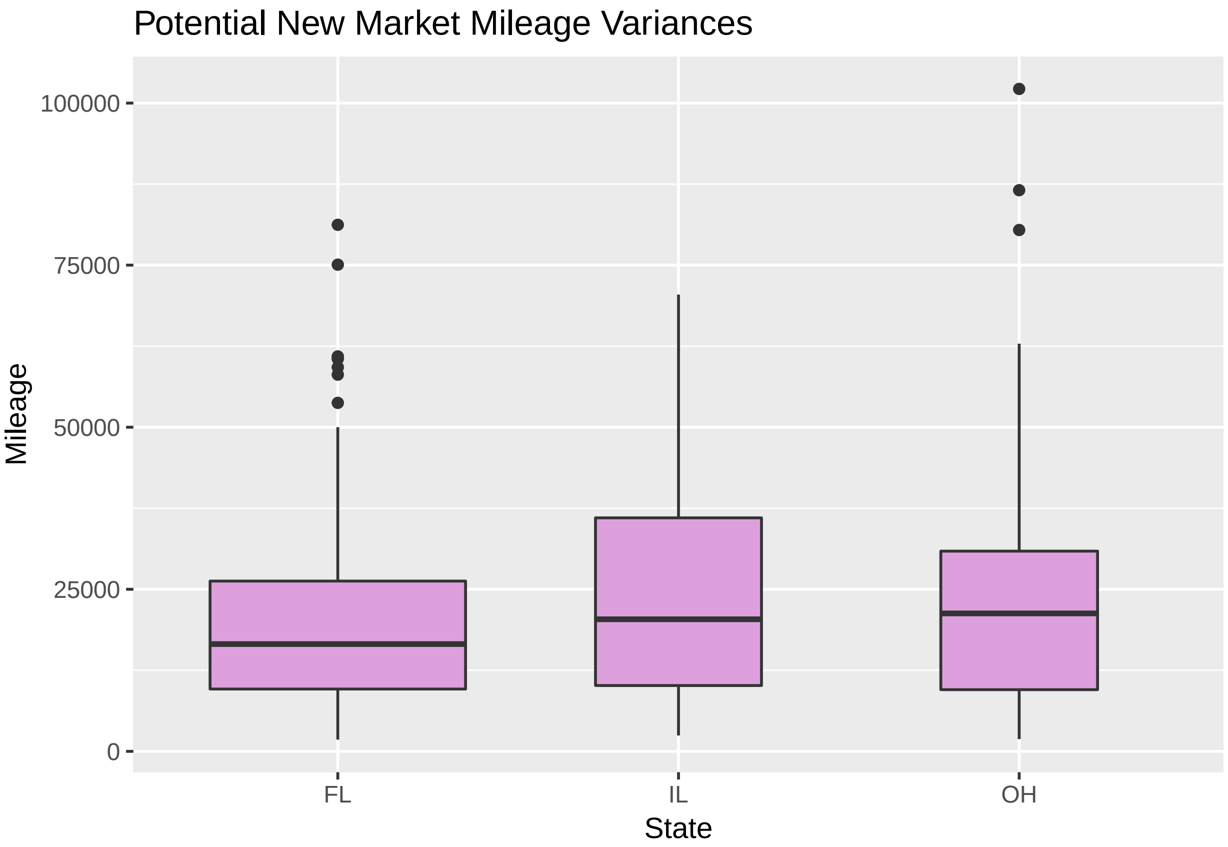
All states have a roughly similar proportion of acquired cars. Yet, OH has a bit less acquired than non aquired.

```r
ggplot(pot_new_market, aes(x=asking_price, y=mileage)) +
  geom_point(aes(col = state)) +
  labs(y="Mileage",
       x="Asking Price",
       title="Relation of Asking Price and Mileage")
```



There are no major clusters in the scatterplot. But, we can see that IL has more cars with a lower asking price and higher milage.

```
ggplot(pot_new_market, aes(state, mileage)) +
  geom_boxplot(varwidth=T, fill="plum") +
  labs(title="Potential New Market Mileage Variances",
       x="State",
       y="Mileage")
```



All three states have similar distributions of mileages. Yet, IL has the largest spread and no outliers. OH has a high average mileage and a few extreme outliers. FL has the lowest average mileage and the smallest spread.

Given this dataset and all the evidence above, I would recommend Shift to expand to Florida next.

In order to make an even more well informed decision, there are some other factors that would be useful. For instance, other vehicle makes would be helpful as it could be quite possible that porsche's are particularly popular in Florida. There could be more cars and makes in other states, making them more profitable but the data doesn't reflect that. It is also important to consider what the selling price is in addition to the asking price. Some states could sell much lower than asking on average, making them a less profitable possible next expansion. Another feature that would also be informative is fuel efficieny of the vehicle. That is a very important aspect that people consider in buying/selling cars. Some states could have a greater demand for fuel efficient cars, making those states more viable as the next market to expand to.