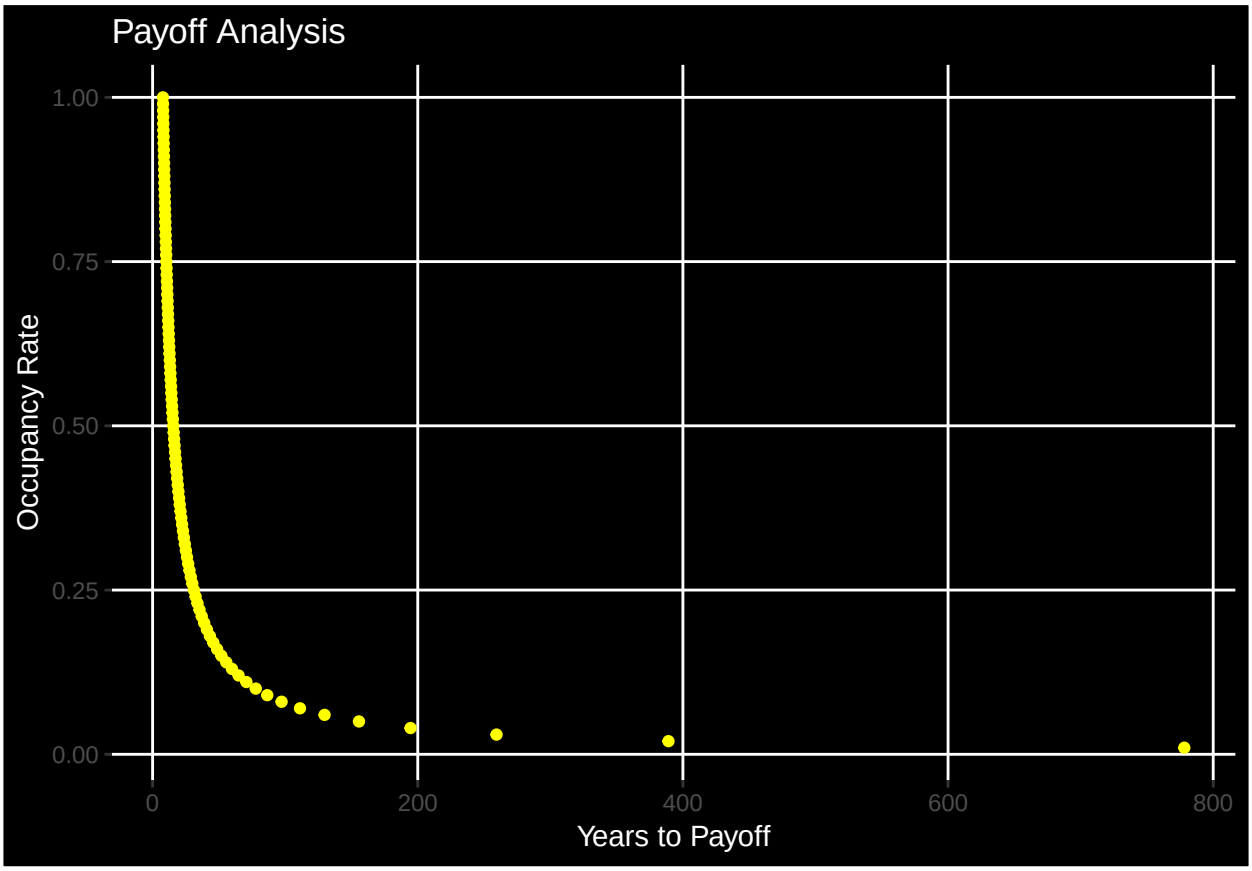
Data Mining and Statistical Learning: Exercise 1

Tejaswi Pukkalla

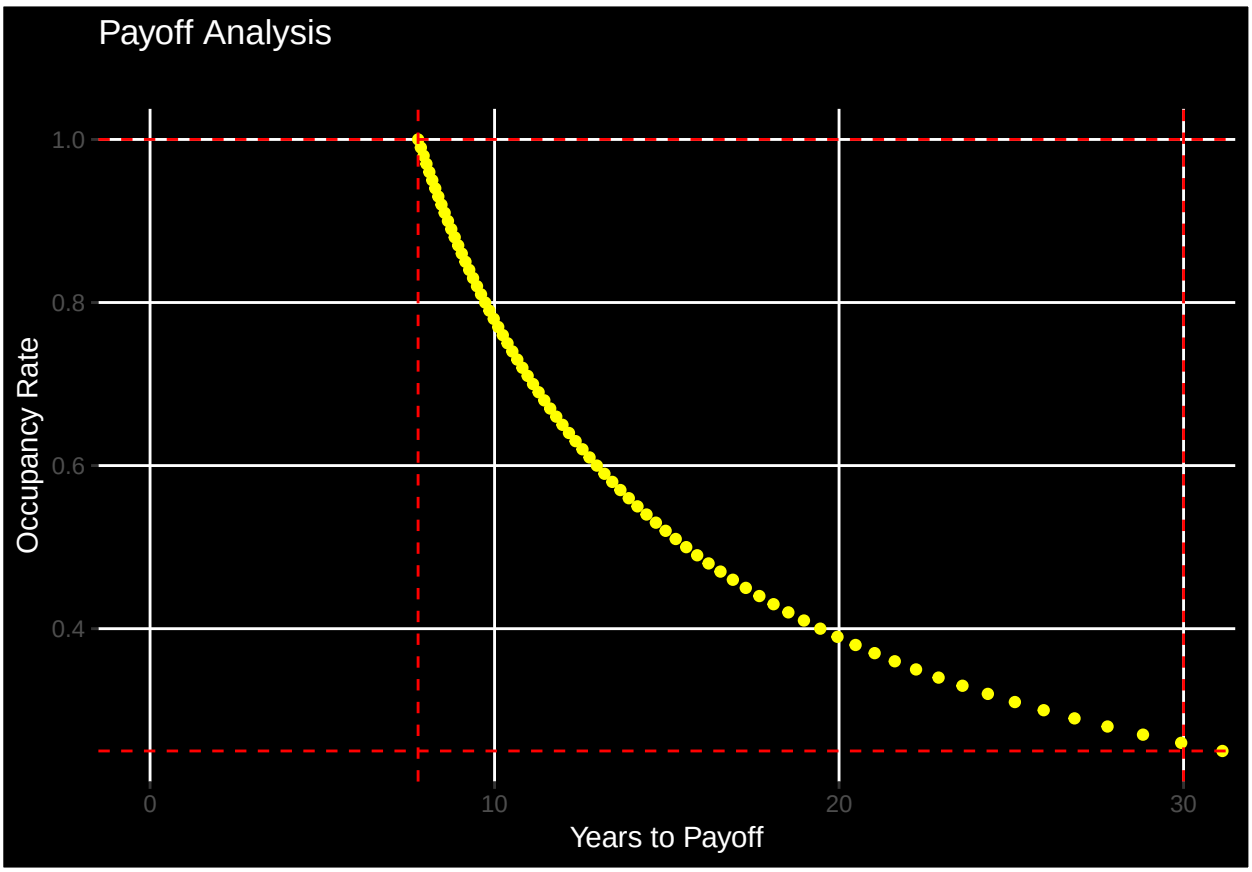
February 11, 2019

Question 1

The Analyst's Analysis



When we replicate the Analyst's data analysis steps, we are able to visualize the amount of time taken to break even the costs of building a green building. At an optimistic 100% occupancy rate, the company would break even in 7.78 years. However, the extreme negative case is when occupancy rate is 1%, it would take about 800 years to break even. We understand this is more of a calculation technicality than a real life scenario.

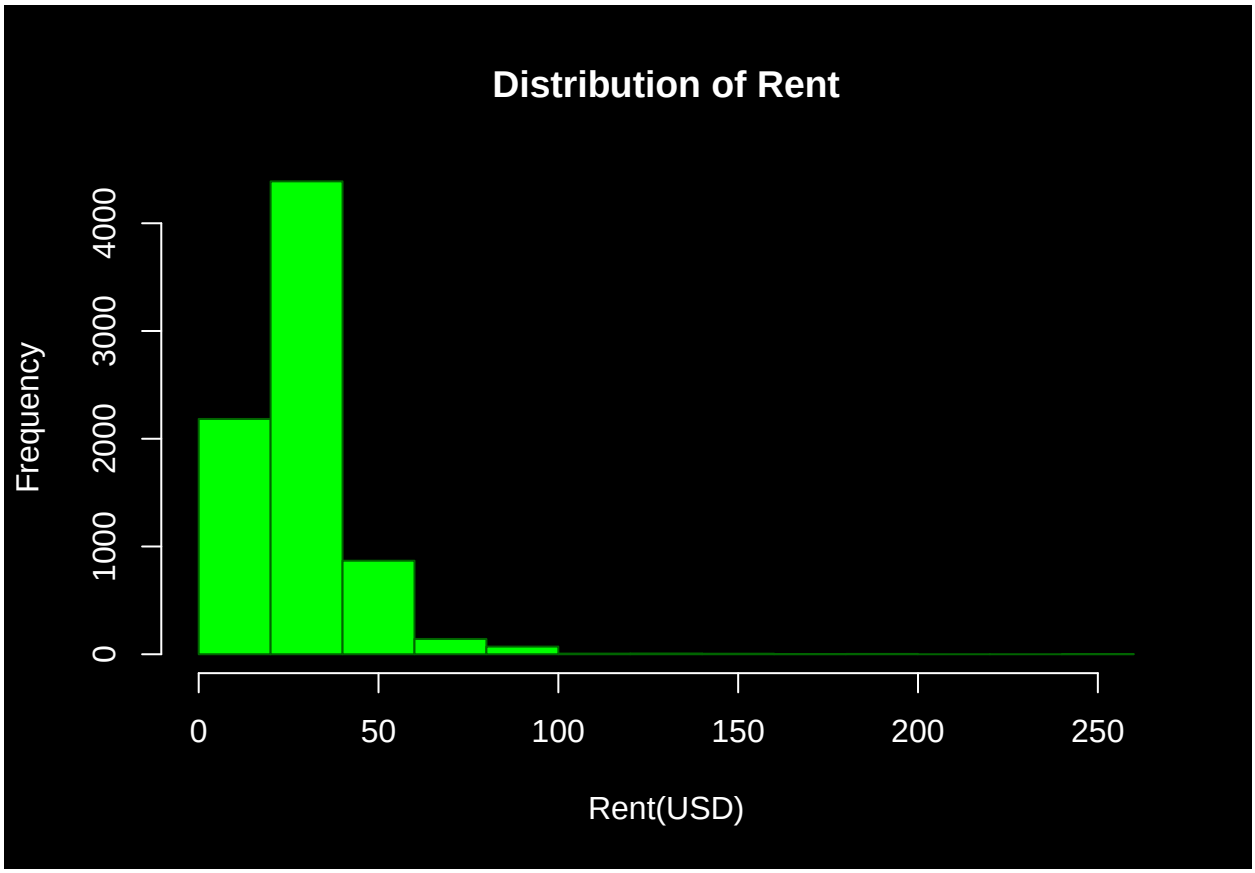


Since the operational lifetime of the building is 30 years, it makes more sense to look at this timeline in a more compact manner. Even if the occupancy rates go as low as 25%, the builders would break even at the end of 30 years.

Although the analyst's desire for simplicity in terms of cleaning and keeping the data is understandable, it is however not prudent to follow. If there are enough green buildings that have occupancy rate of less than 10%, that might be a necessary trend to be picked up during analysis. Using the median instead of the mean would skew the data as the number of non green buildings is massive compared to green buildings.

```
##
##    0    1
## 6995  684
```

Where "0" represents the number of non-green buildings and "1" represents the number of green buildings



The above histogram shows how skewed the data is to the right. Since most values lie between 0 and 50, this would mean the median would be significantly different from the mean.

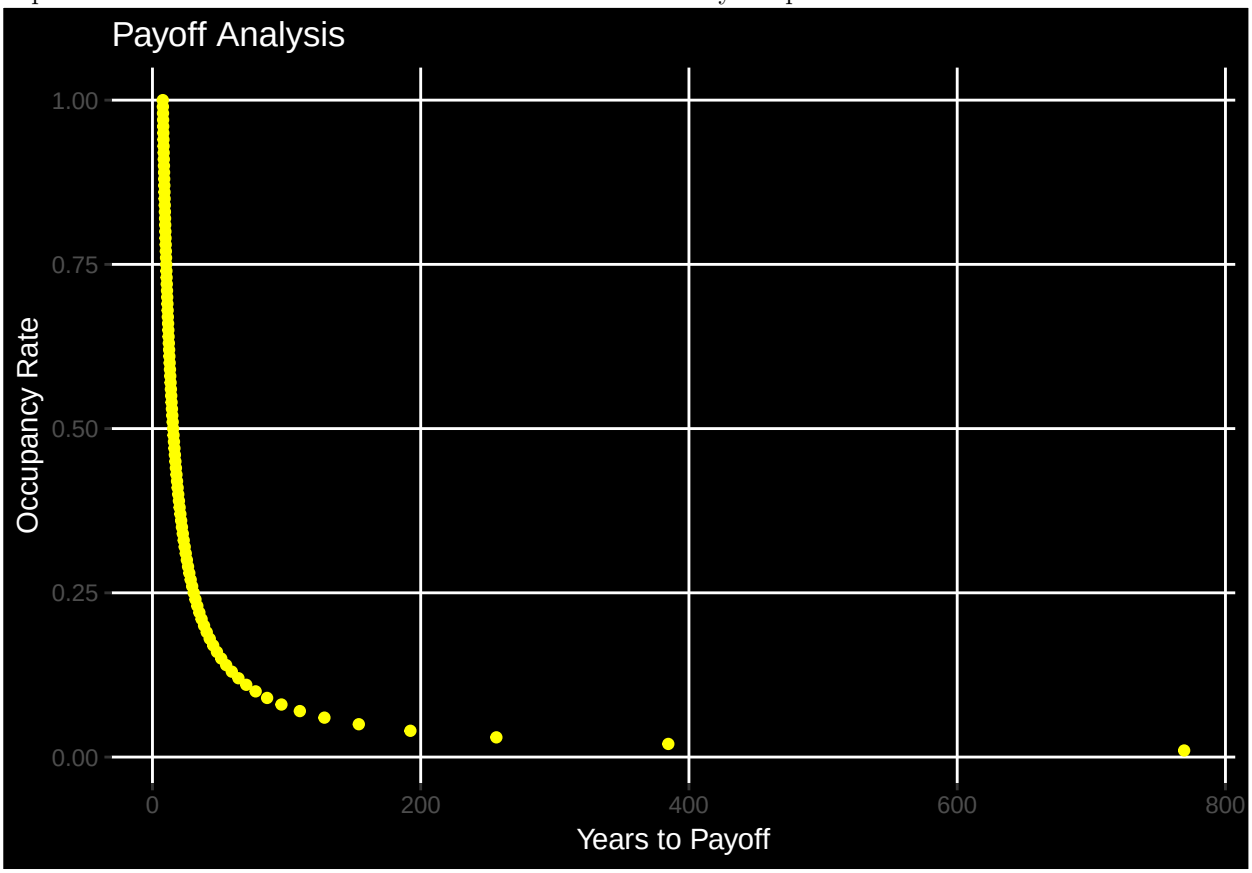
```
##                  Median     Mean
## Whole Dataset 25.29000 28.58585
```

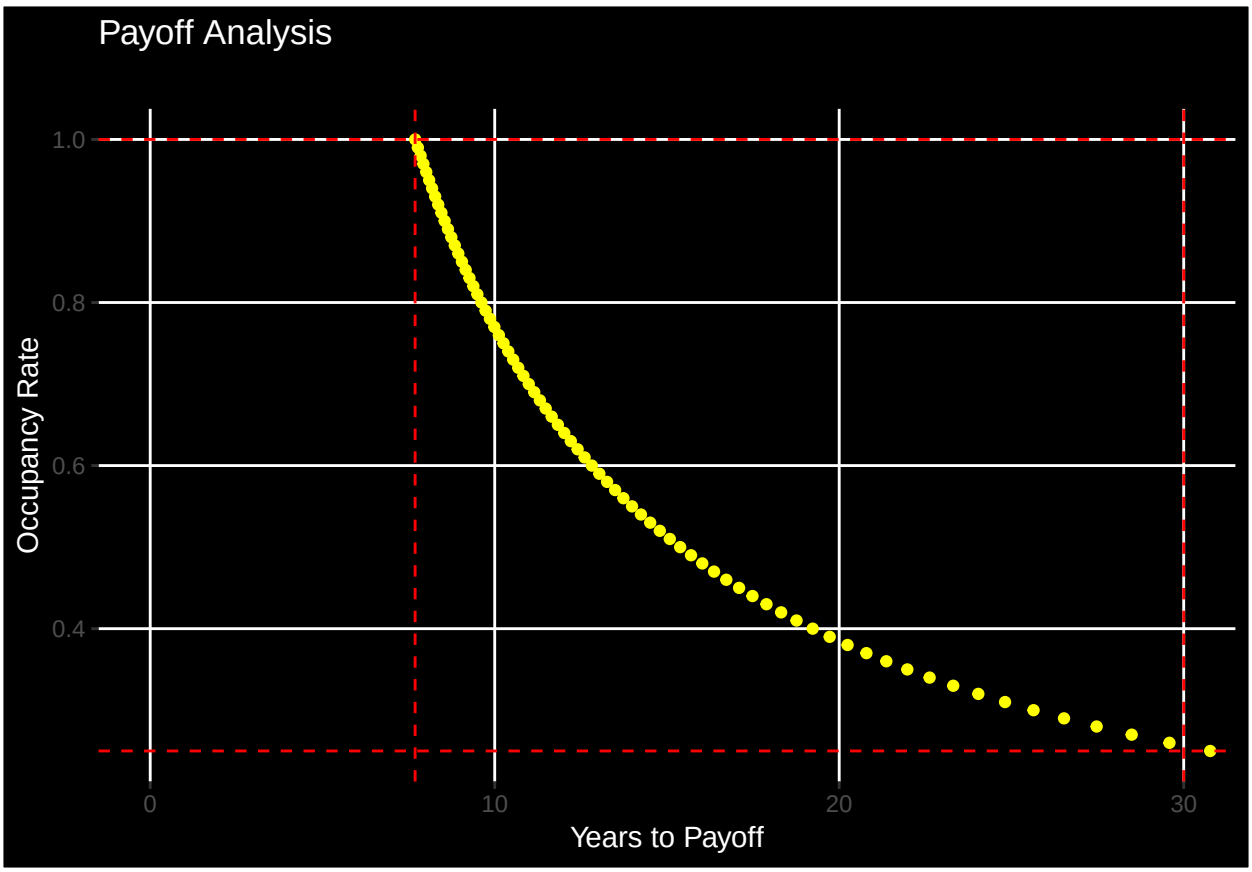
If we take mean into consideration however, we can see that the rent difference of 2.6 doesn't hold true and the difference comes down to as low as 1.7. This greatly impacts how we would want to calculate the time taken to break even and the profitability of takinng up this investment.

```
##                  Median     Mean
## Whole Dataset 25.29000 28.58585
## Green         27.60000 30.02848
## Non-Green     25.03000 28.44478
```

Our Analysis

Instead of removing data that isn't convenient to tell a particular side of the story, we include it so as to have a complete dataset. This affects our calculations but not in a very steep manner.



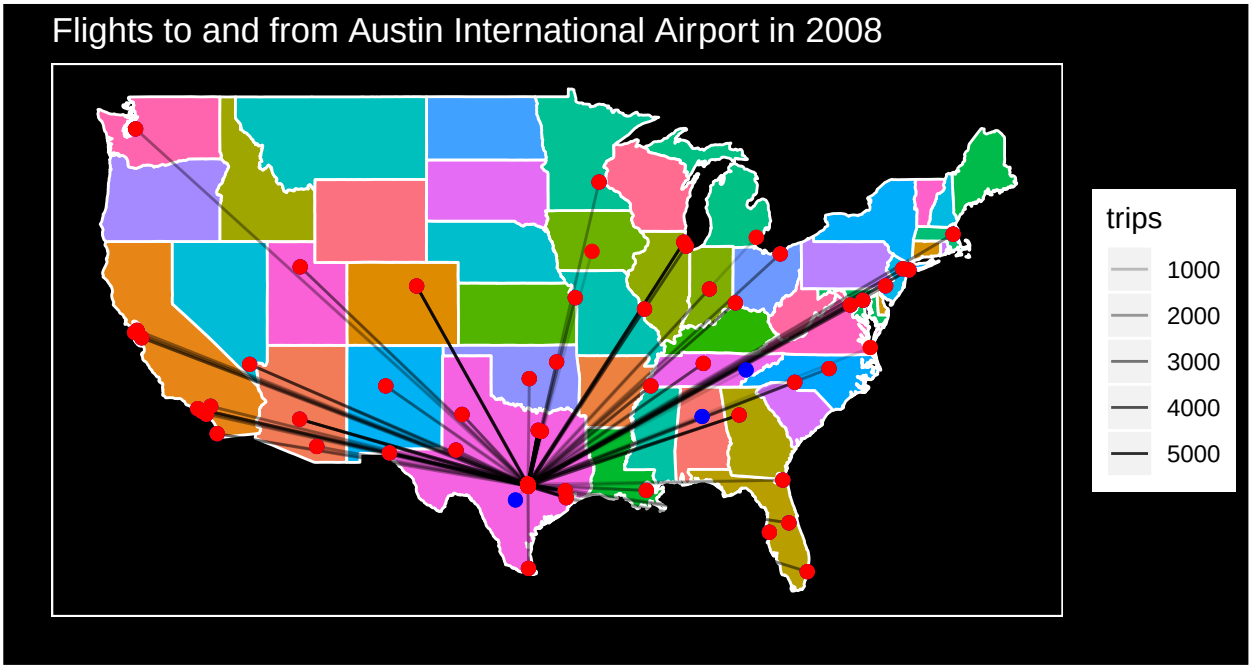


The time taken to break even varies a bit as compared to analyst's numbers. Ours is a more conservative value that states it would take about 12 years to break even if the occupancy rate is a complete 100% and the entire lifetime of the building if it goes down to occupancy rate of 37%. To summarize, it is more prudent to include the whole dataset and be a bit more conservative when making investment decisions rather than to simplify or massage data till it becomes a convenient story to tell.

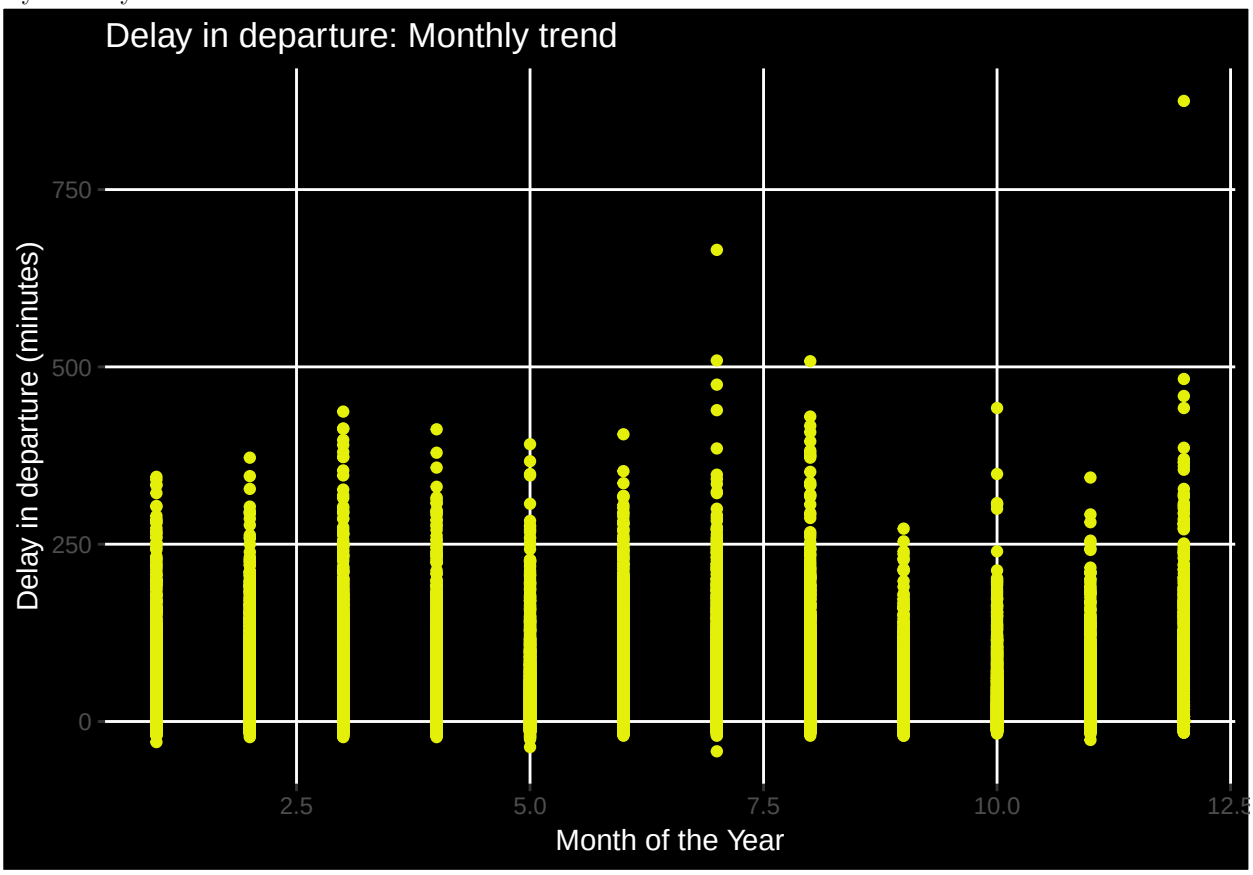
Question 2

Flights travelling to/from Austin

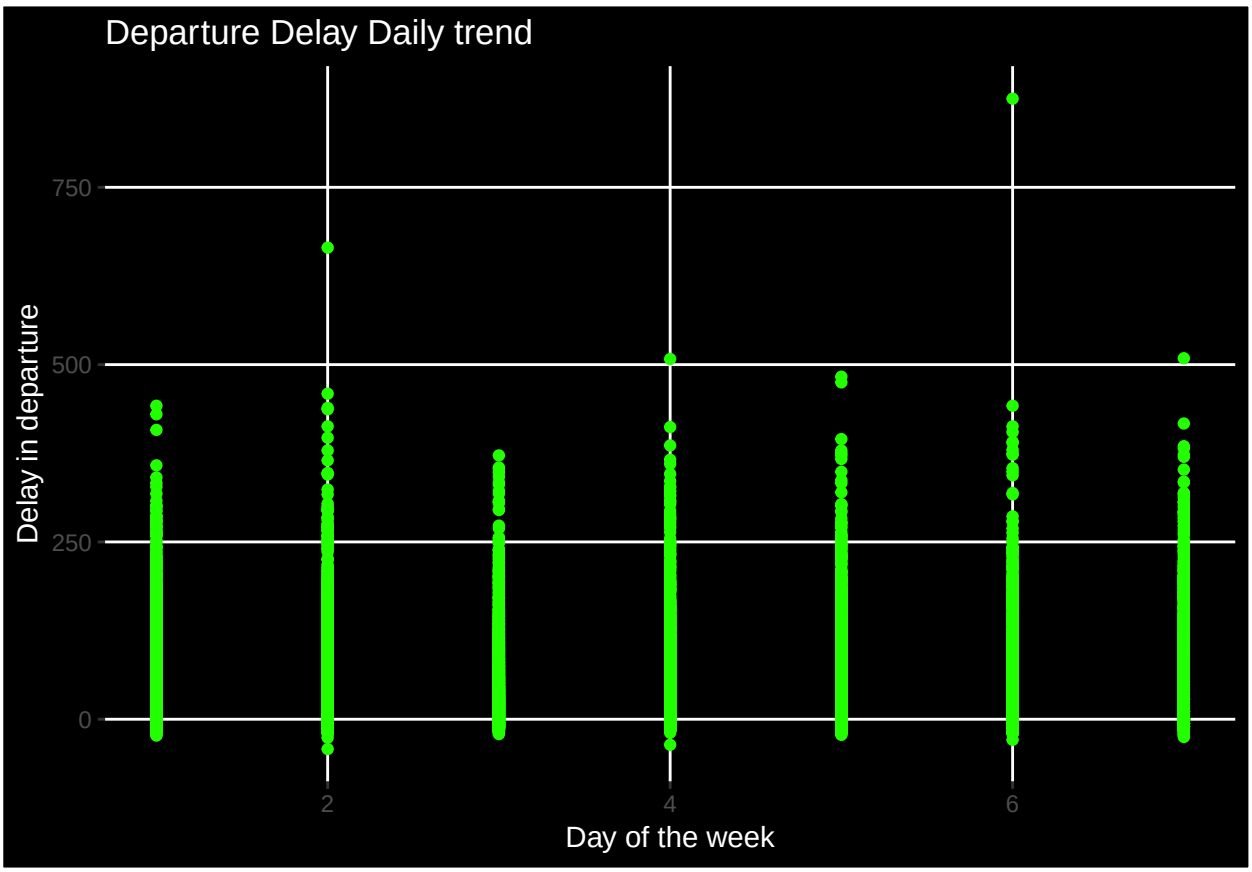
We try to show the map of the United States, depicting the travel routes of all flights flying to/from Austin. The path density is determined by how many trips occur between the origin and destination cities.



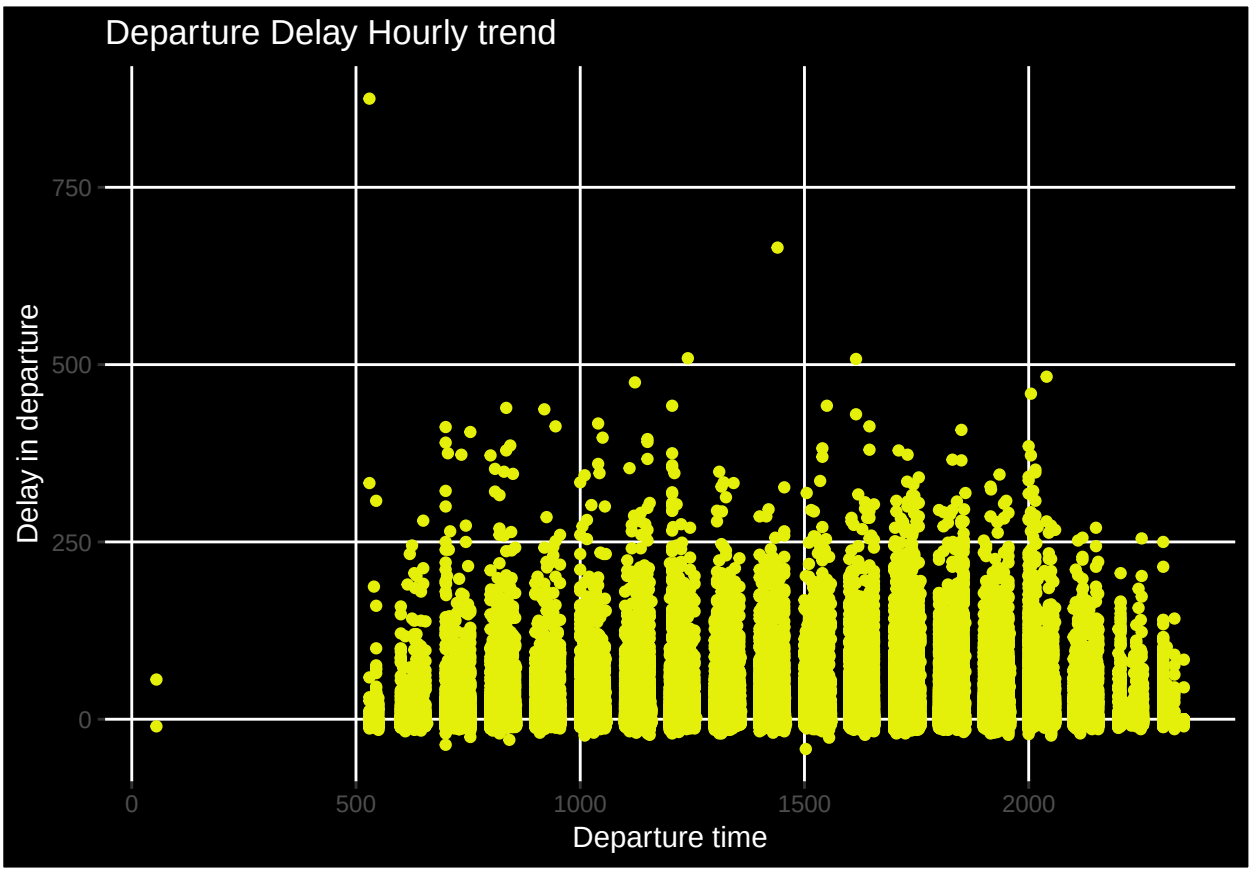
We try to look at a scatterplots in departure delays to see if there are any trends with respect to the time or day of the year.



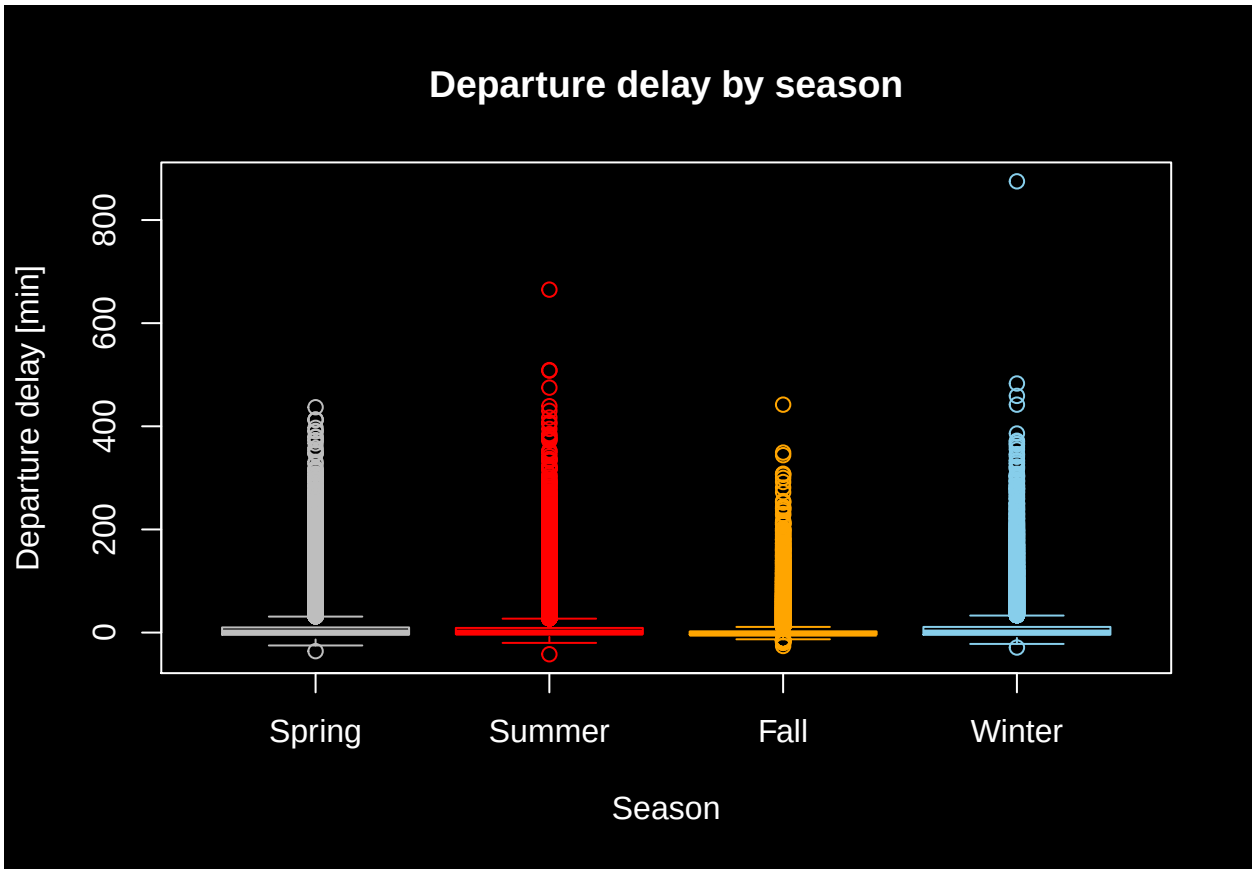
The delays seem to be concentrated more around the months of August and September which is the holiday season, so it makes sense that more number of flights lead to more delay.



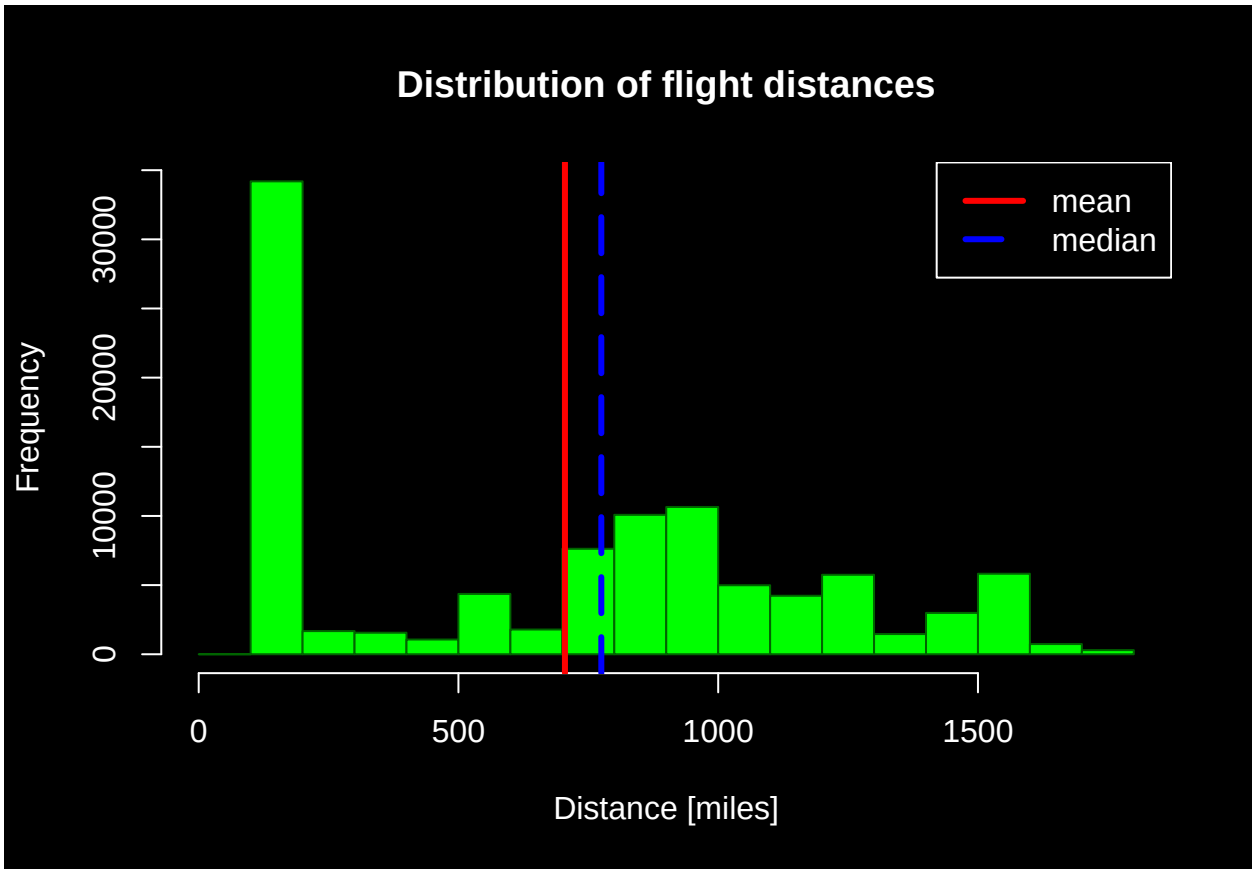
There doesnt seem to be a strong difference between delays on different days of the week, although weekends are a bit of a rush.



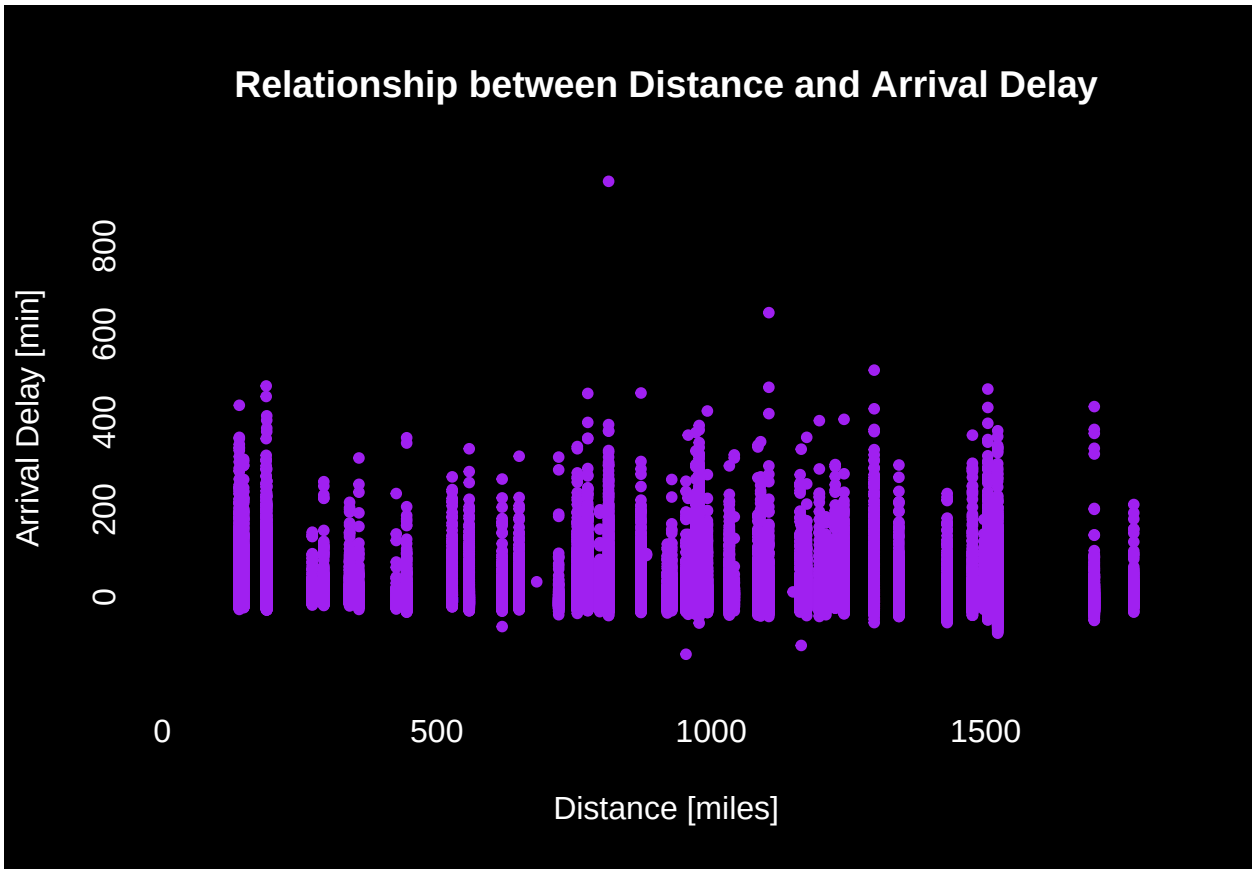
There are hardly any delays from midnight to 5a.m. This is to be expected as there aren't many flights during that time of the day. The delays seem to increase as the day progresses peaking in the evening and then again going down as the night passes.



Above is a boxplot of delays grouped by season. This shows that thee are more delays in summer and winter as compared to spring and fall. This again resonates with our hypothesis of more delays during holiday season.



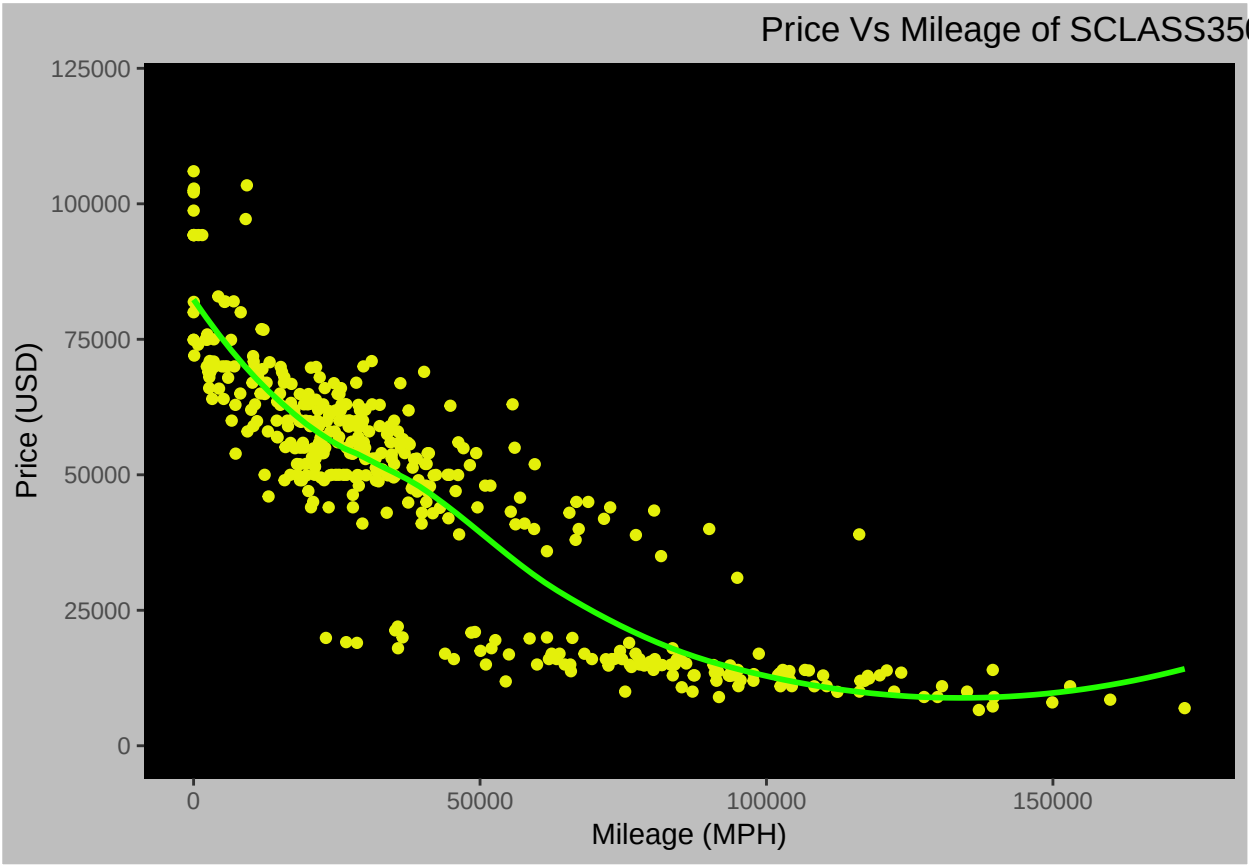
When we look at the relationship between number of flights and distance travelled, we see that most flights are for short distances only and very few flights travel across the country.



We also notice a slight increase in average delay with an increase in distance. This sounds reasonable as the long distance flights would require additional effort and time to get ready for the journey ahead.

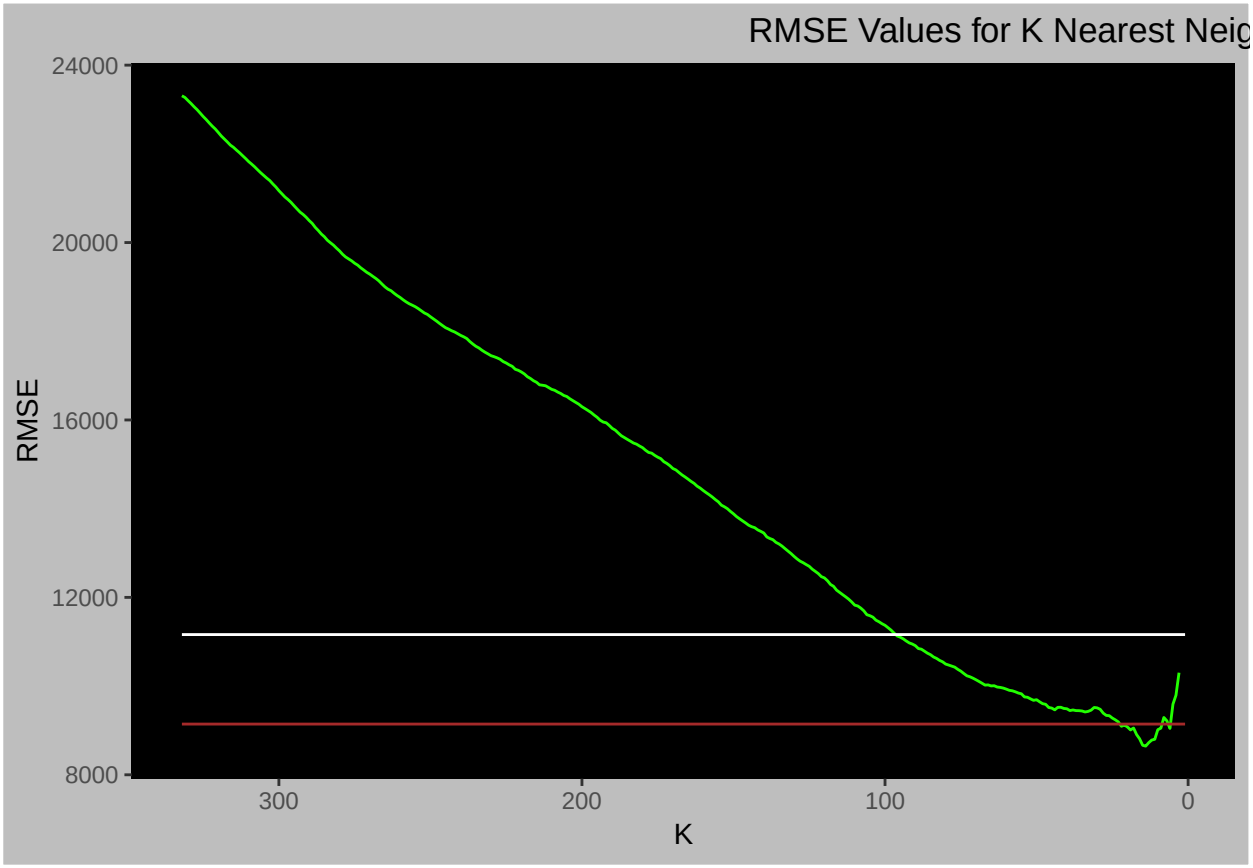
Question 3

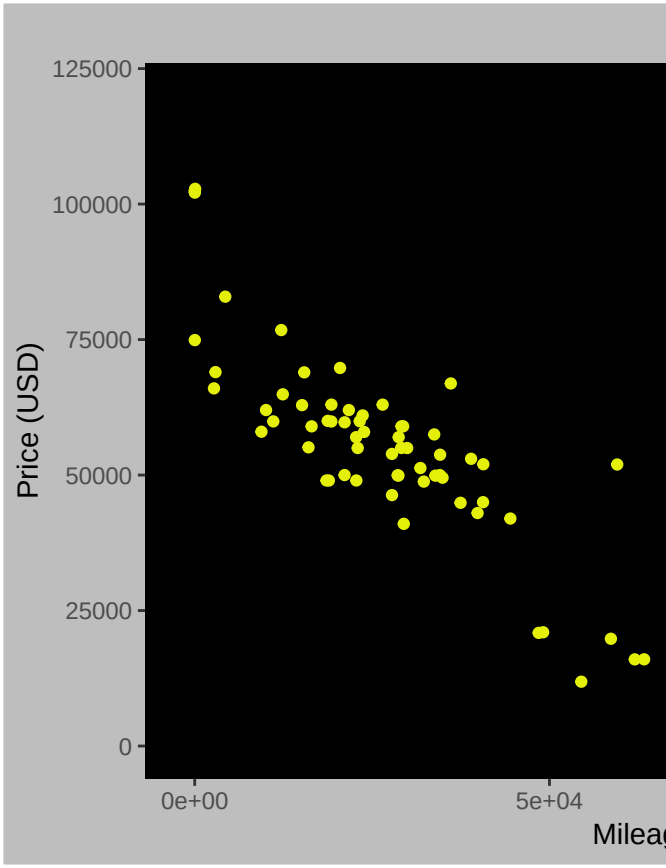
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The above graph shows thedistribution of prices of SCLASS350 with respect to mileage offered. We have added a smoothing filter so see how the trend line should look like.

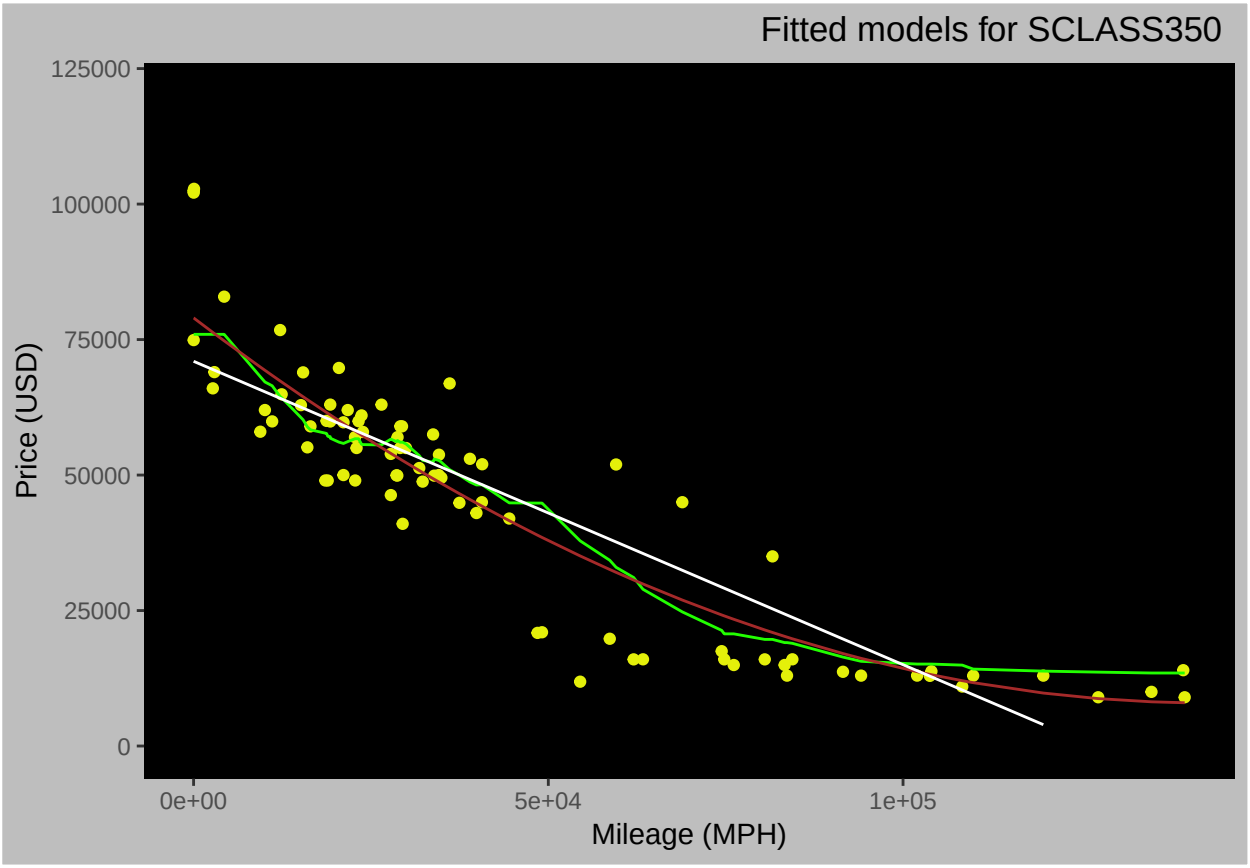
Below is a graph showing the RMSE values of linear models of order 1 and 2 and KNN ranging from K=3 TO 332.



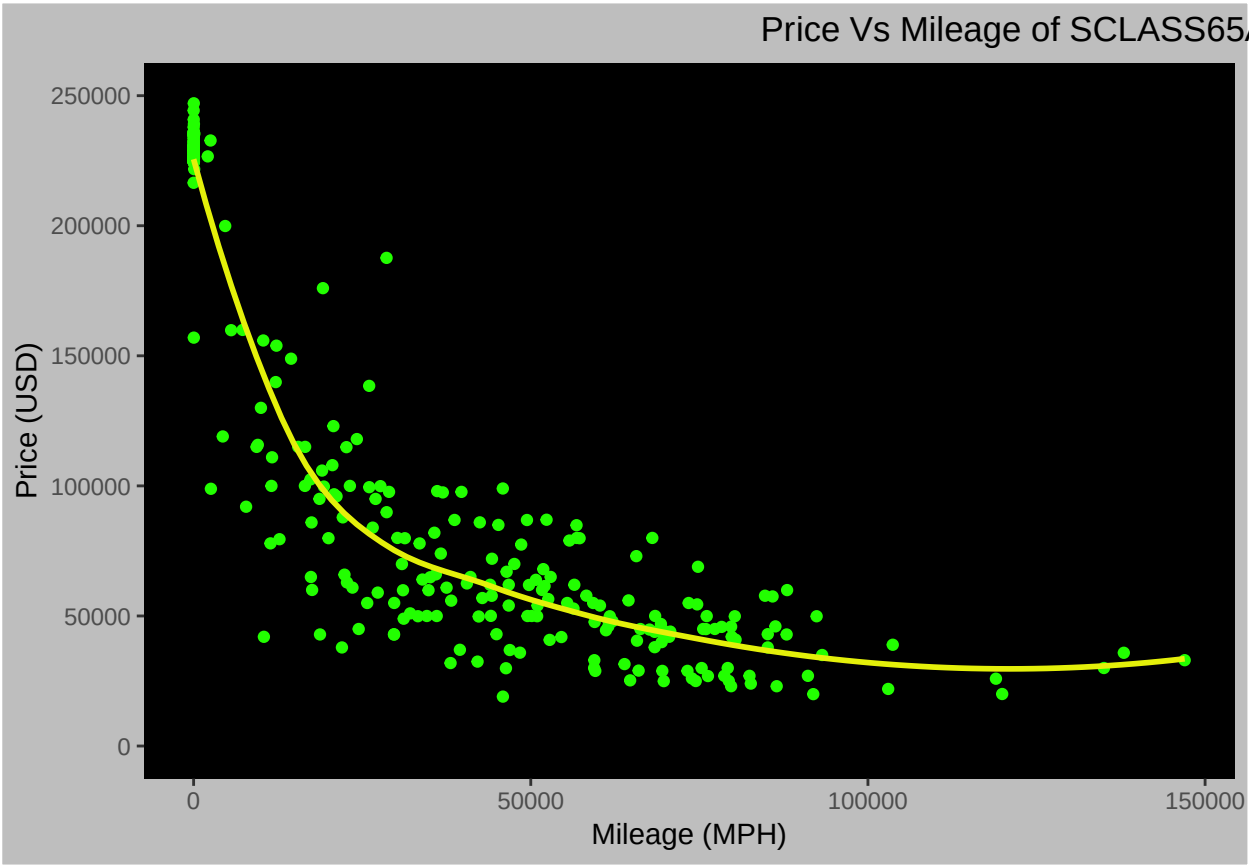


Comparing the scatterplot with the fitted models (K=50), we get,

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```

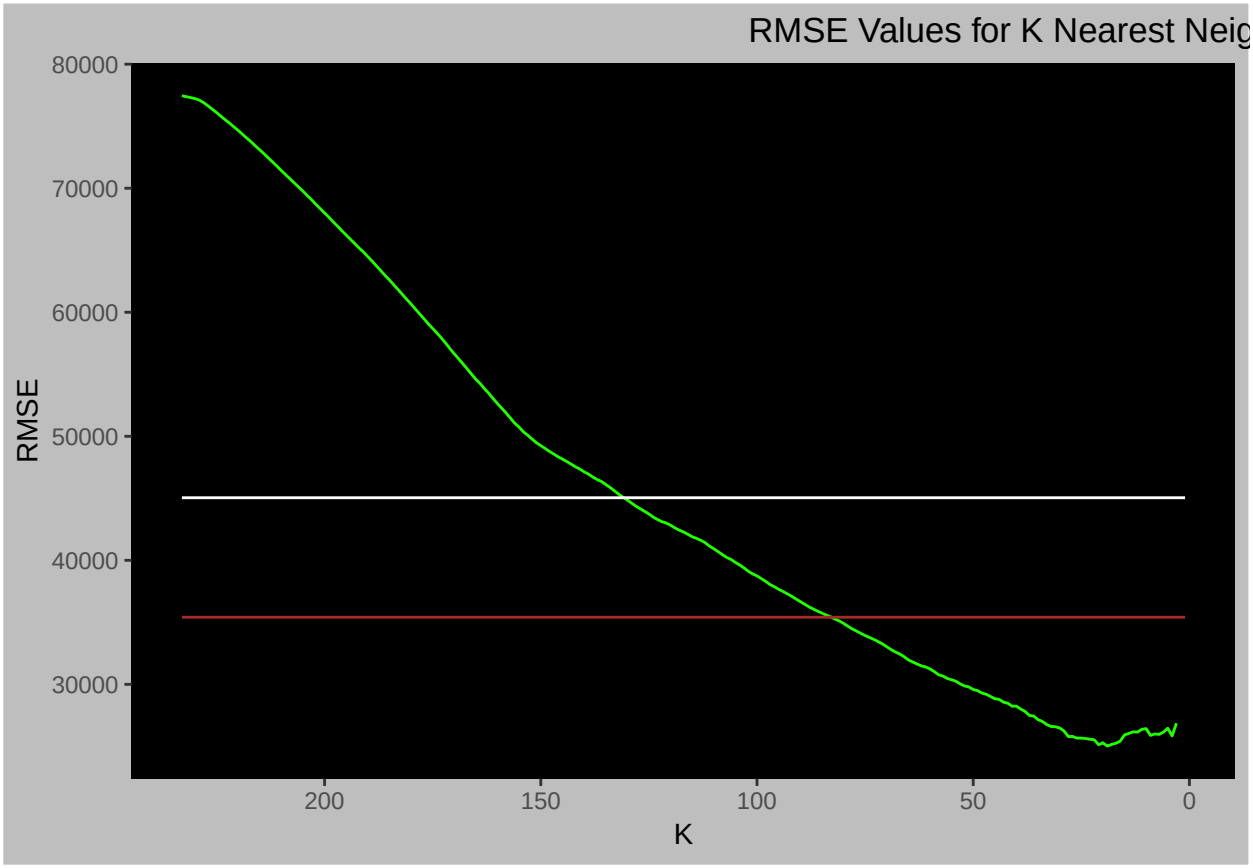


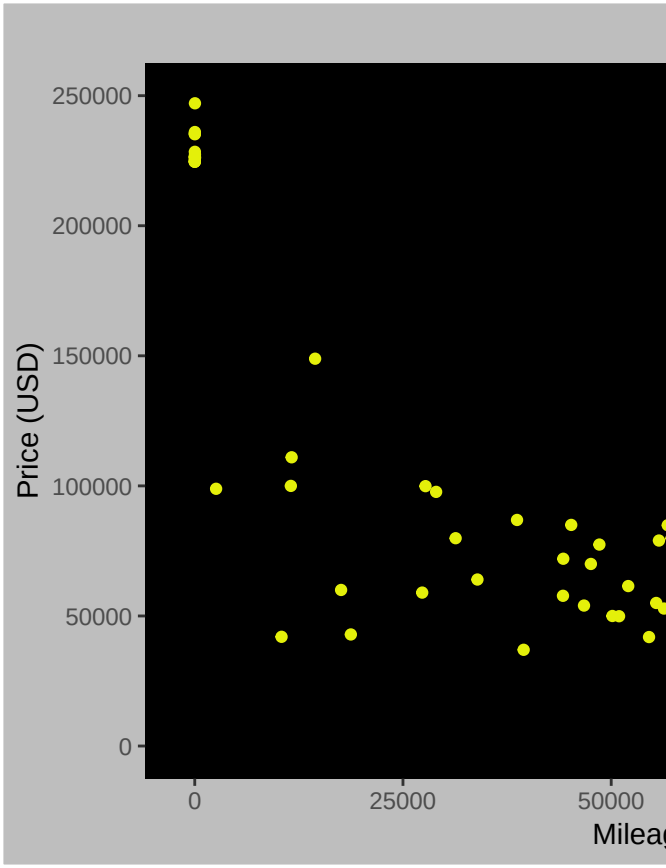

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The above graph shows thedistribution of prices of SCLASS65AMG with respect to mileage offered. We have added a smoothing filter so see how the trend line should look like.

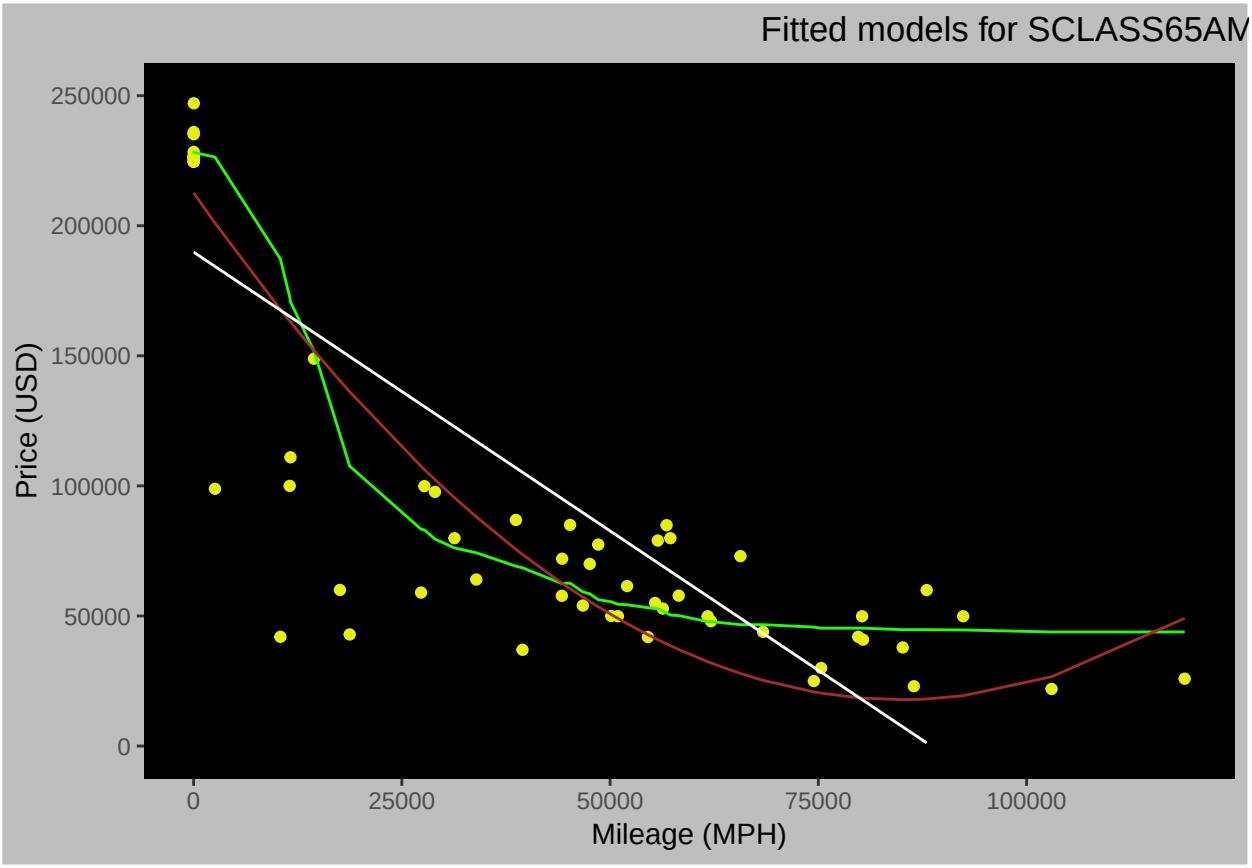
Below is a graph showing the RMSE values of linear models of order 1 and 2 and KNN ranging from K=3 TO 332.





Comparing the scatterplot with the fitted models (K=70), we get,

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```



The K value is higher for SCLASS65AMG as it has more outliers than SCLASS350