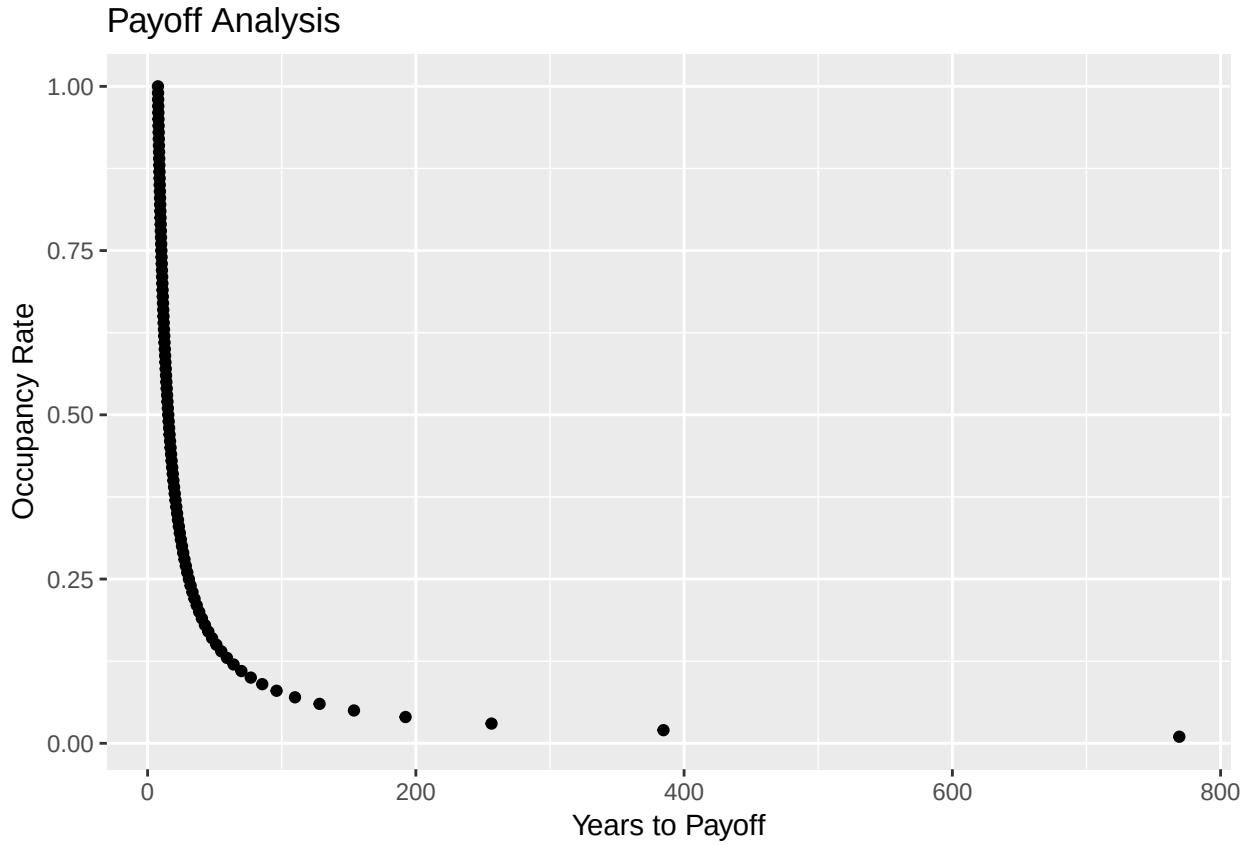
Data Mining and Statistical Learning: Exercise 1

Frank Chou

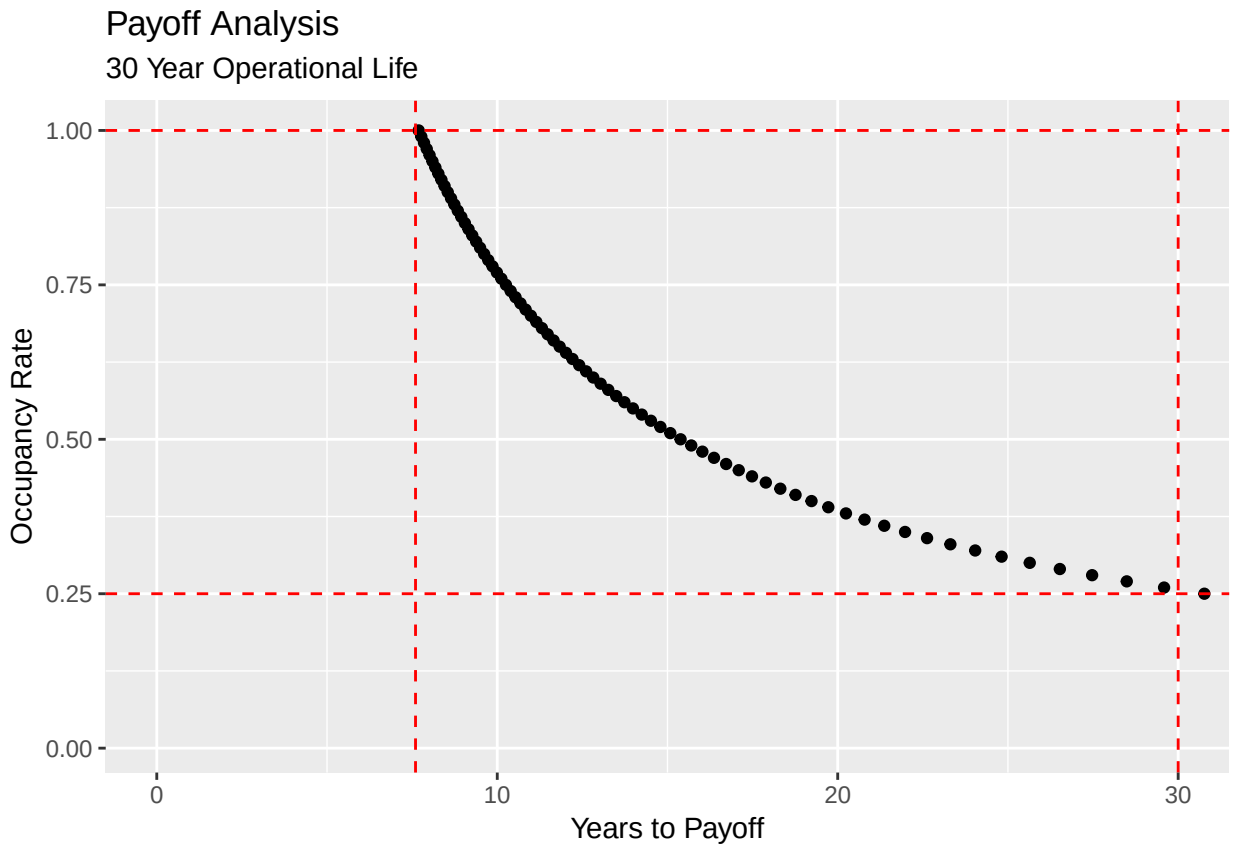
February 9, 2019

Question 1

The Analyst's Analysis



A rudimentary analysis of Analyst's parameters gives us a basic understanding of how long it would take to pay off the initial investment of building a green building: from a low of 7.6 years at 100% occupancy rate to over 800 years if we have a 1% occupancy rate. However, this is unrealistic, becuase if we were analyze the entire data set, we find that the rate of occupancy for green to non-green buildings are significantly different. But before we continue, let us focus on the actual building operational lifecycle of 30 years.



Here we see a more accurate estimate of the payoff timetable for the building if we built it green. If we have 100% occupancy, we expect to pay it off within 7.6 years, however if we see 25% occupacy rates, we would break-even at the end of the building's estimated operational lifecycle, and any lower than 25% we would never break-even.
Regardless of the analysis there are a number of points of fault within the Analyst's analysis.

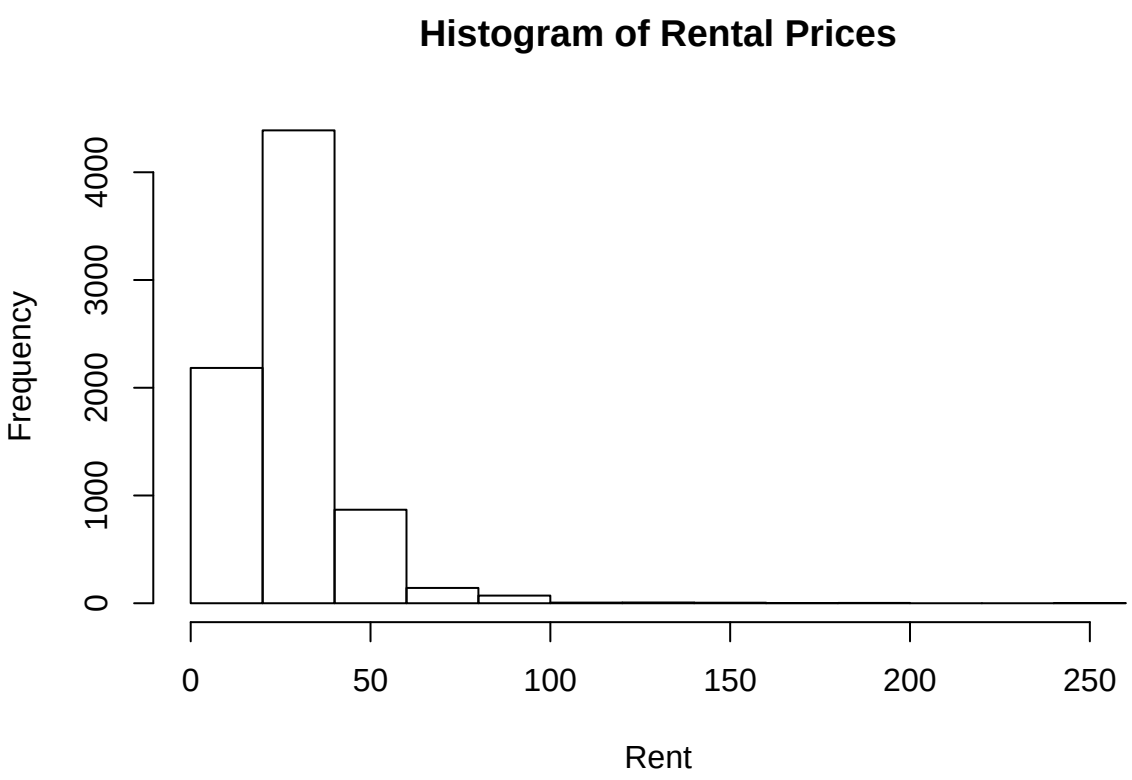
- One: by eliminating buildings where its occupancy rates are below 10%, we exclude important data that would give us insight as to how the market is doing, namely the fact that we don't expect occupancy rates to remain constant nor at max-capacity at the onsent of the building's commission. Becuase the data does not give us longitudial information regarding a given building's occupancy rate over the time of its lifetime, we do not have an idea of how a building going from non-green to green affects its rental value.

- Two: by utilizing the median rental value of the subsets: green and non-green, we skew the data because we have more green versus non-green buildings:

```
##
##    0    1
## 7209  685
```

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

- Three: As for the decision to utilize the median instead of the mean value of rent, a cursory assessment of different values provide different results of a green vs. non-green building.



A histogram of the Analyst's data set provides a view of how the rental prices are positively skewed. There are more buildings with rental prices in the 0-50 range than >50. Using the median value prices would have a result less than the mean value.

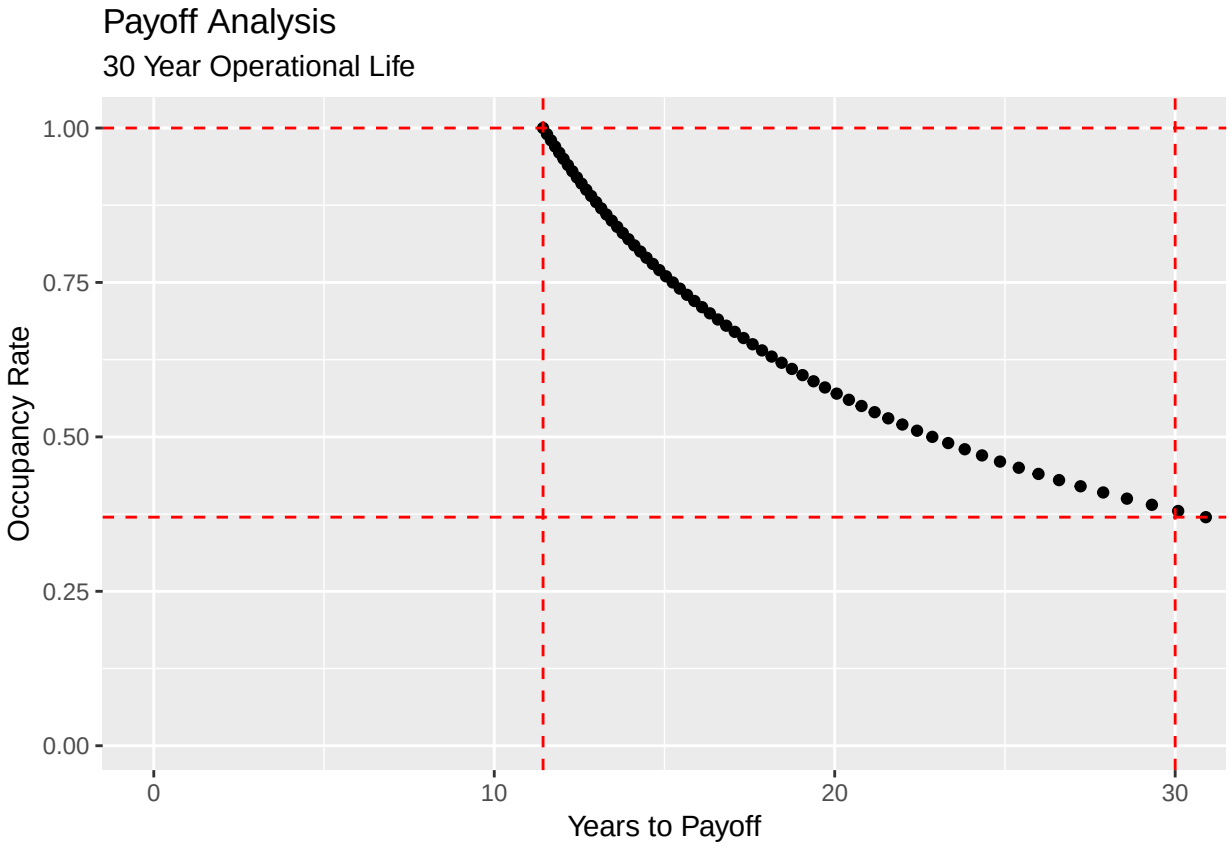
```
##                  Median     Mean
## Whole Dataset 25.29000 28.58585
```

However, if we were to compare the price differences within the green versus non-green building subsets, we see that the effect of green buildings are less than anticipated. Intead of a 2.6 difference using median-values, it will only be 1.7. This will greatly extend our anticipated break-even point.

```
##                  Median     Mean
## Whole Dataset 25.29000 28.58585
## Green         27.60000 30.01603
## Non-Green     25.00000 28.26678
```

A Better Approach

If we were to conduct the analysis with the full set of data and the mean-value, we see a different picture.

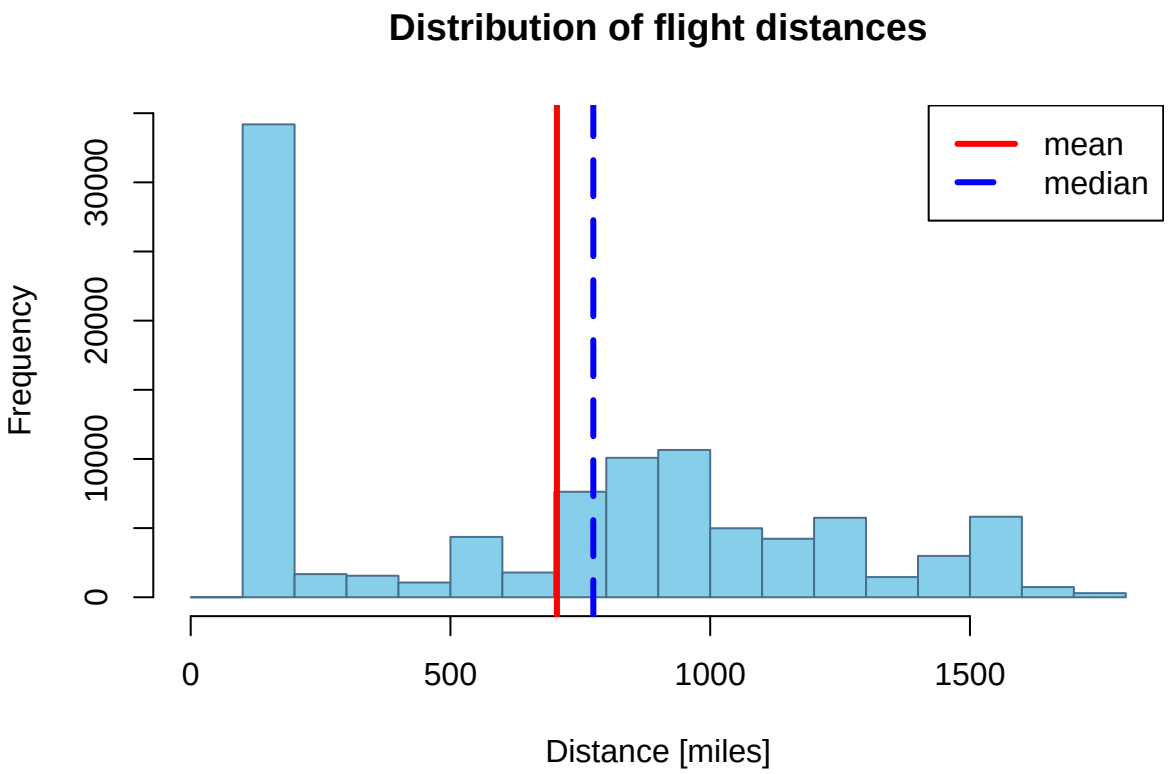


Here we see a more accurate estimate of the payoff timetable for the building if we built it green using mean-values. If we have 100% occupancy, we expect to pay it off within 11.43 years, however if we see 37% occupacy rates, we would break-even at the end of the building's estimated operational lifecycle, and any lower than 37% we would never break-even. In conclusion, we see that a more accurate assessment of the viability of building green is represented by using the full data set with mean-values instead of a truncated data set with median-values.

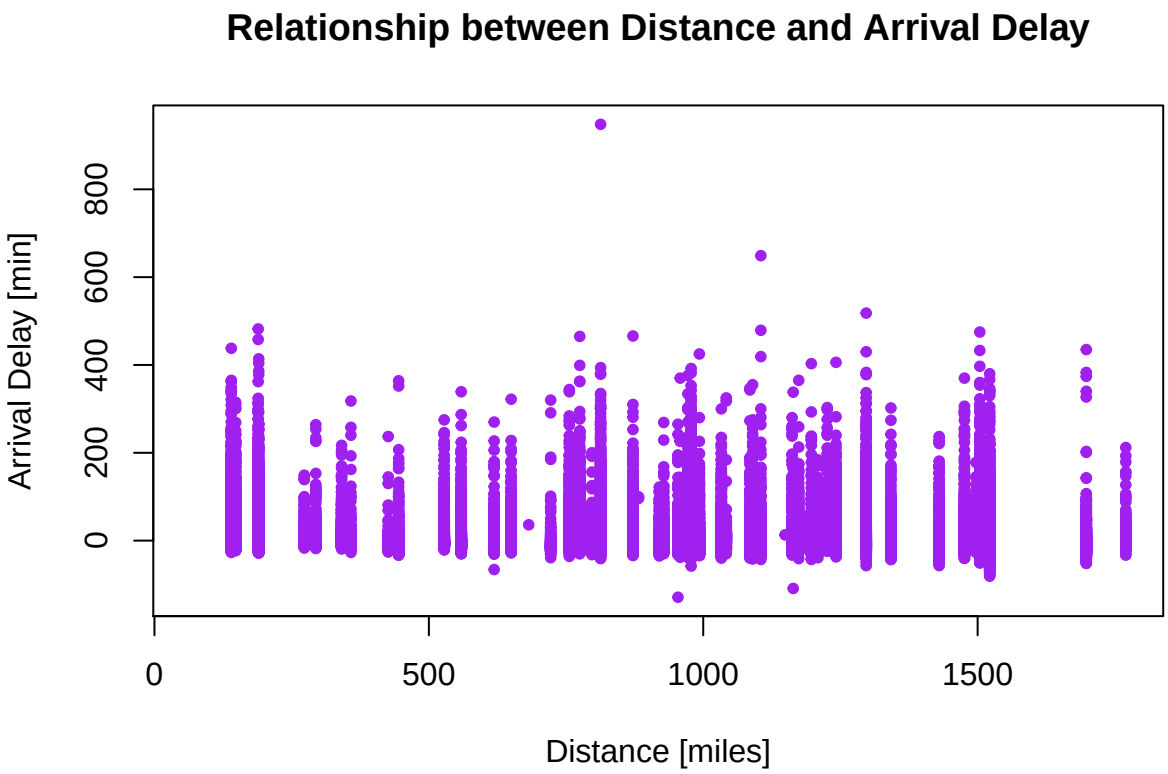
Question 2

A flight to Austin

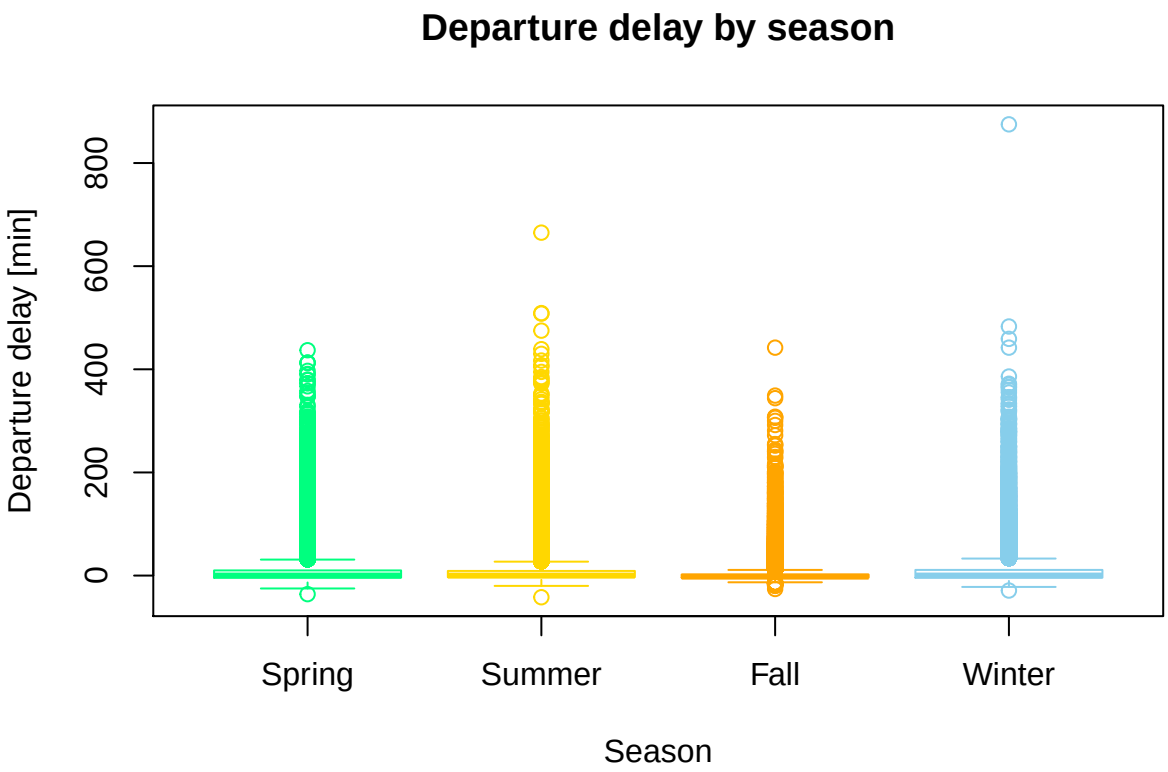
We can get a better understanding of which cities are connected to Austin by air. By plotting all of the flights to and from Austin to other cities in the United States, we can see which cities fly more to and from Austin.



Here we have a histogram of all of the flights originating and destination from AUS, most distances are within 500 miles.



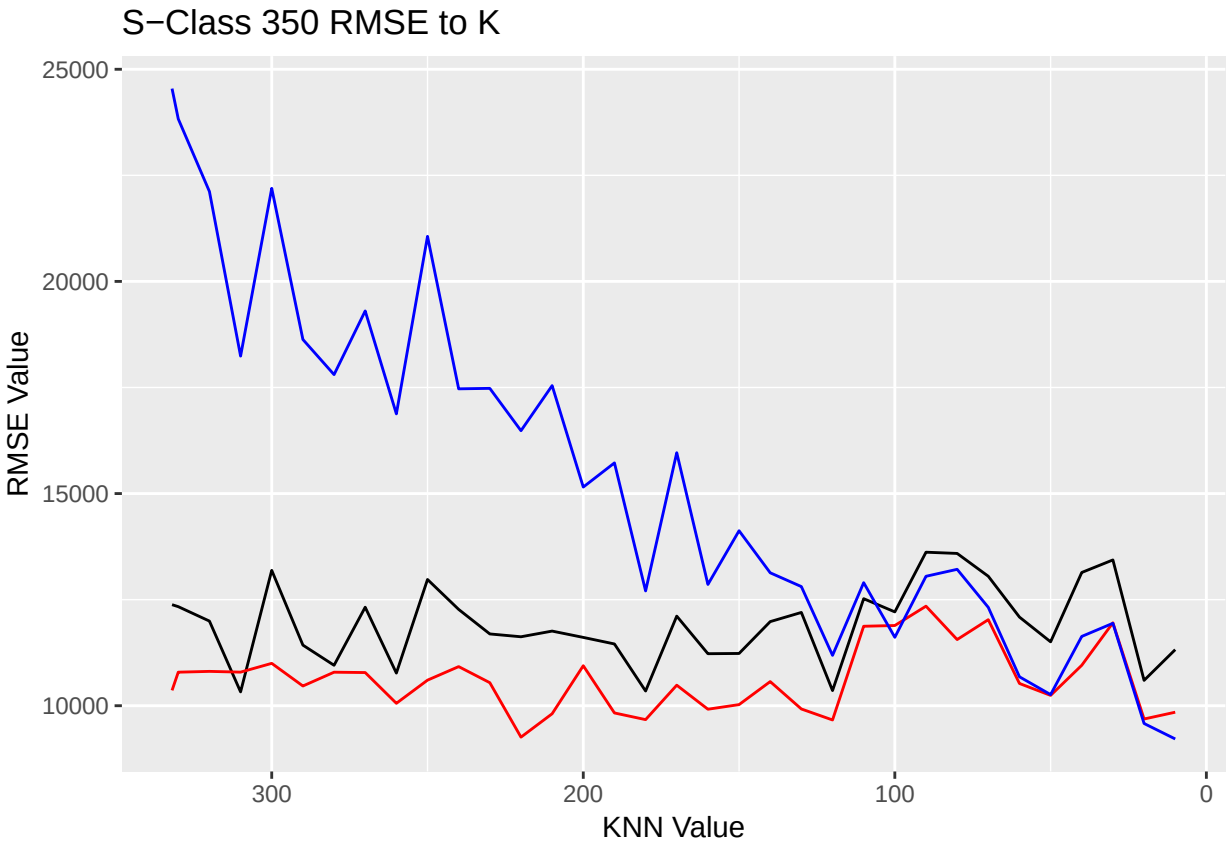
Although difficult to tell with the naked eye, there is a marked increase in delay time when the flight distance increases based on the density of the plot near the 1000 mile range.



Looking strictly a the outlier data points, we find that there are outliers of delayed flights during the Summer and Winter Season.

Question 3

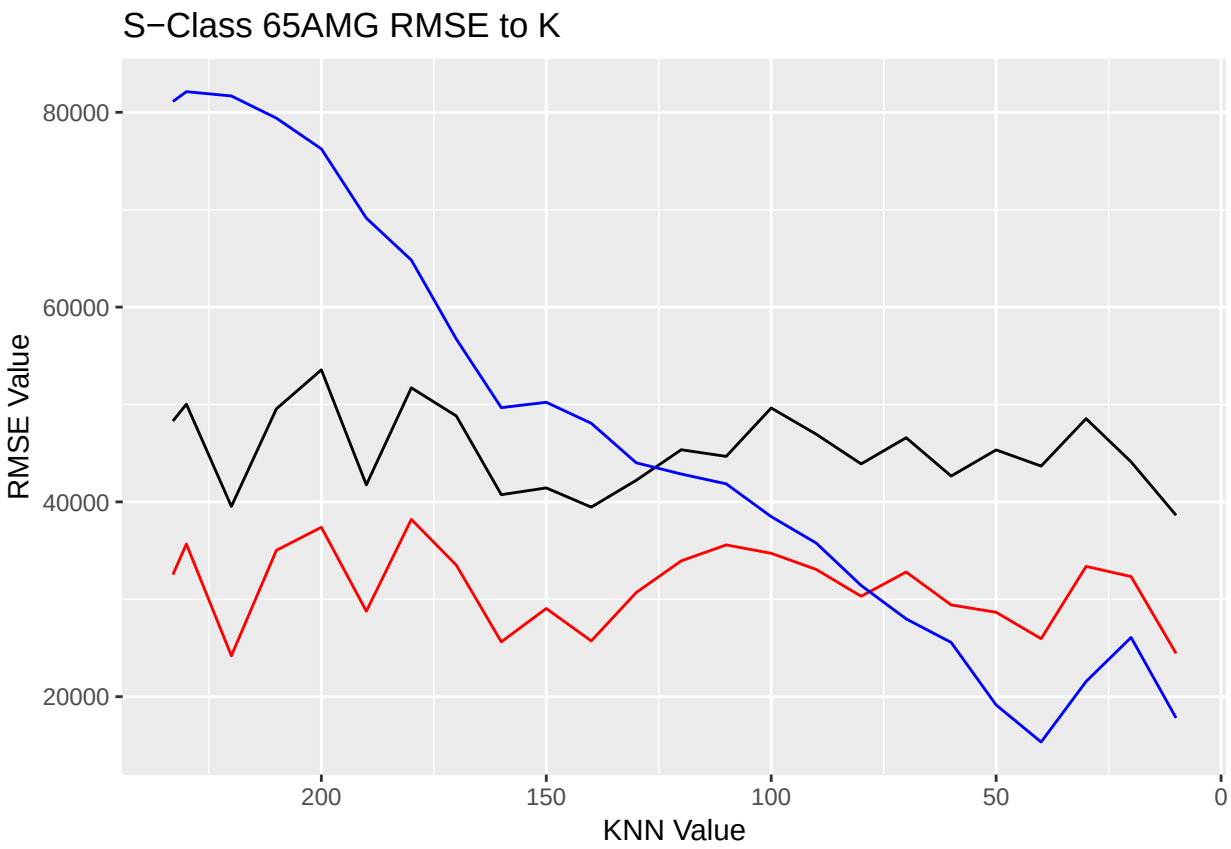
Comparing RMSE to K



Legend:

- Linear RMSE in Black
- Polynomial RMSE in Red
- KMM RMSE in Blue

Based on the blue line for KMM RMSE values, we find that RMSE value start to bottom out when K<150.

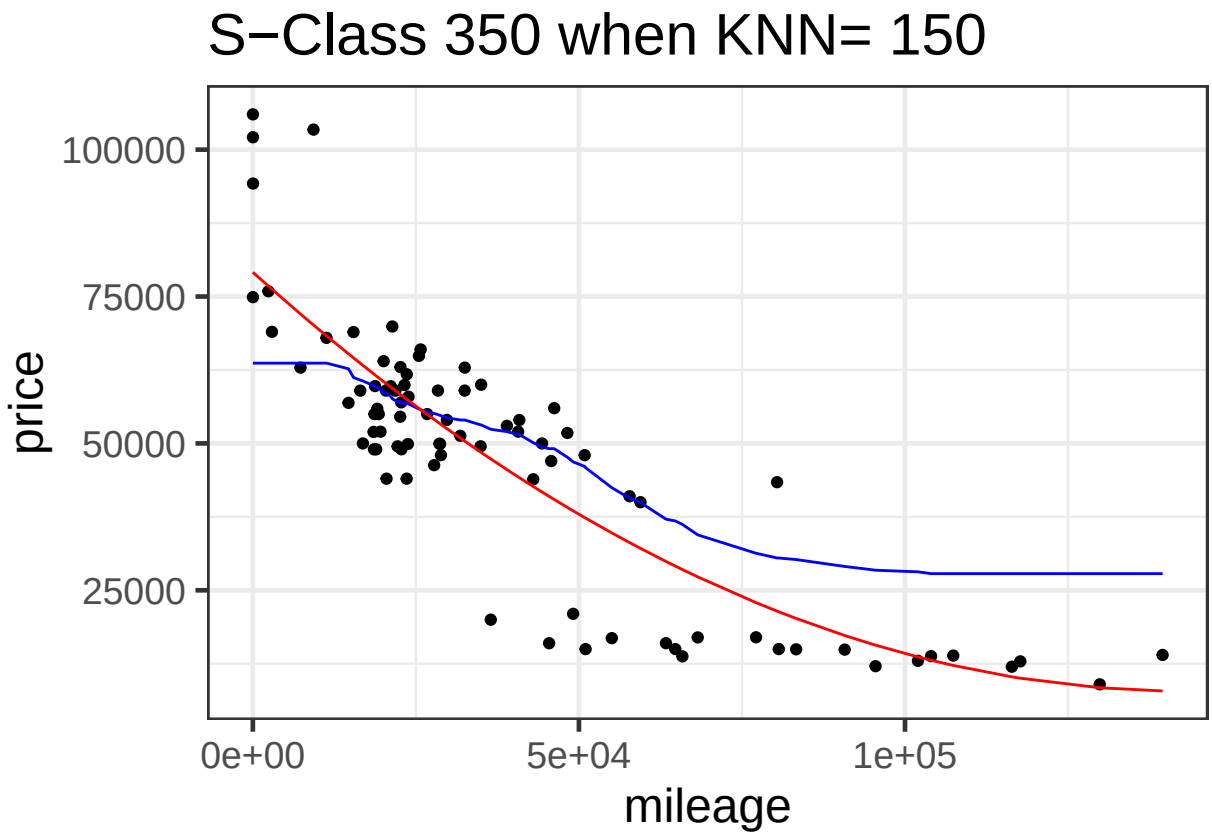


Legend:

- Linear RMSE in Black
- Polynomial RMSE in Red
- KMM RMSE in Blue

Based on the blue line for KMM RMSE values, we find that RMSE value start to bottom out when K<60.

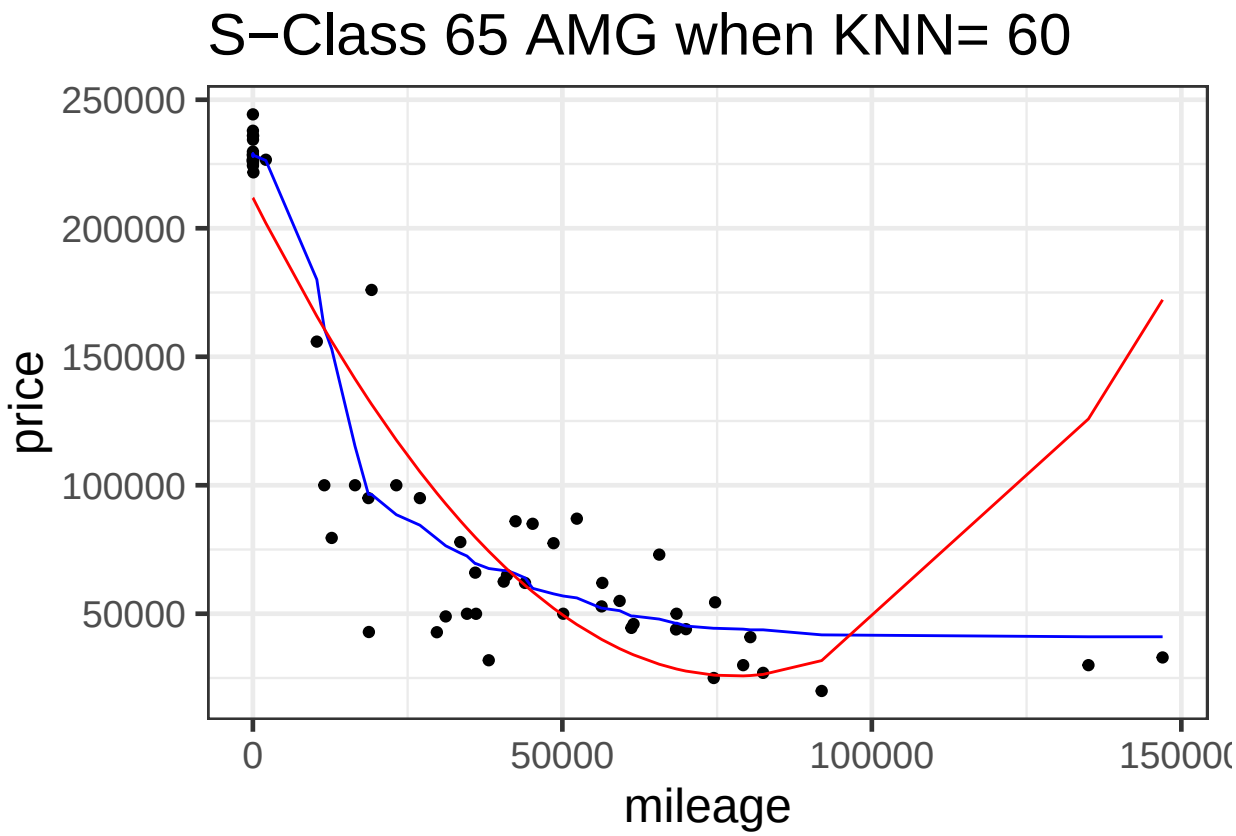
S-Class 350



Legend:

- Polynomial RMSE in Red
- KMM RMSE in Blue

S-Class 65 AMG



Legend:

- Polynomial RMSE in Red
- KMM RMSE in Blue

In conclusion, we find that the S-Class 350 trim offers a more optimal value of K becuase of the larger dataset it provides for the training data, reducing the tail-end bias effects.