Exercise 1 - Data Mining

Kylie Taylor, David Fraire, & Larisa Barreto

1/28/2019

Data Visualization 1: Green Buildings

The investor is intersted in the return of investment on a new "certified-green building" in Austin, TX. There are inherently many questions that must be answered in order to make a decision on whether to develop or not. In this analysis, we will not be preforming any predictive models through the use of regression. What we will be doing is conducting an analysis based on visualzations of the data at hand.

The data is sourced from a dataset titled "Green Buildings" constructed by real-estate economists, observing 1,360 green-certified buildings throughout the United States. Building characteristics were only available for 685 green buildings. For comparison purposes, the 685 green buildings were clustered with 12 non-green buildings within a quarter-mile radius of the certified green building, using data sourced from the CoStar database. The merged datasets consist of a total of 7,894 observations spanning the entire United States (685 clusters of approximately 12 buildings).

In order to become a certified-green building, a commercial property must fit within specific environmental criteria and is certified by an outside engineer. Some of the criteria a green building must satisfy are energy efficiency, carbon footprint, site selection, and sustainable building materials. Green buildings can be awarded either LEED or EnergyStar certifications.

There are 21 variables in the dataset. A summary of these variables follows:

1. *CS.PropertyID*: the building's unique identifier in the CoStar database.
2. *cluster*: an identifier for the building cluster
3. *size*: the total square footage of available rental space in the building.
4. *empl.gr*: the year-on-year growth rate in employment in the building's geographic region.
5. *Rent*: the rent charged to tenants in the building, in dollars per square foot per calendar year.
6. *leasing.rate*: a measure of occupancy; the fraction of the building's available space currently under lease.
7. *stories*: the height of the building in stories.
8. *age*: the age of the building in years.
9. *renovated*: whether the building has undergone substantial renovations during its lifetime.
10. *class.a, class.b*: These are relative classifications within a specific market. Class A buildings are the highest-quality properties. Class B buildings are a notch down. Class C buildings are the least desirable properties.
11. *green.rating*: an indicator for whether the building is either LEED- or EnergyStar-certified.
12. *LEED, Energystar*: indicators for the two specific kinds of green certifications.
13. *net*: an indicator as to whether the rent is quoted on a "net contract"" basis.
14. *amenities*: an indicator of whether at least one of the following amenities is available on-site: bank, convenience store, dry cleaner, restaurant, retail shops, fitness center.
15. *cd.total.07*: number of cooling degree days in the building's region in 2007.
16. *hd.total07*: number of heating degree days in the building's region in 2007.
17. *total.dd.07*: the total number of degree days (either heating or cooling) in the building's region in 2007.
18. *Precipitation*: annual precipitation in inches in the building's geographic region.
19. *Gas.Costs*: a measure of how much natural gas costs in the building's geographic region.
20. *Electricity.Costs*: a measure of how much electricity costs in the building's geographic region.
21. *cluster.rent*: a measure of average rent per square-foot per calendar year in the building's local market.

The first visualization we will make is a table of summary statisitcs for relevant variables

Table 1: Table continues below

GB.size	GB.empl_gr	GB.Rent	GB.leasing_rate
Min. : 1624	Min. :-24.950	Min. : 2.98	Min. : 0.00
1st Qu.: 50891	1st Qu.: 1.740	1st Qu.: 19.50	1st Qu.: 77.85
Median : 128838	Median : 1.970	Median : 25.16	Median : 89.53
Mean : 234638	Mean : 3.207	Mean : 28.42	Mean : 82.61
3rd Qu.: 294212	3rd Qu.: 2.380	3rd Qu.: 34.18	3rd Qu.: 96.44
Max. :3781045	Max. : 67.780	Max. :250.00	Max. :100.00
NA	NA's :74	NA	NA

Table 2: Table continues below

GB.stories	GB.age	GB.cd_total_07	GB.hd_total07
Min. : 1.00	Min. : 0.00	Min. : 39	Min. : 0
1st Qu.: 4.00	1st Qu.: 23.00	1st Qu.: 684	1st Qu.:1419
Median : 10.00	Median : 34.00	Median : 966	Median :2739
Mean : 13.58	Mean : 47.24	Mean :1229	Mean :3432
3rd Qu.: 19.00	3rd Qu.: 79.00	3rd Qu.:1620	3rd Qu.:4796
Max. :110.00	Max. :187.00	Max. :5240	Max. :7200
NA	NA	NA	NA

Table 3: Table continues below

GB.total_dd_07	GB.Precipitation	GB.Gas_Costs	GB.Electricity_Costs
Min. :2103	Min. :10.46	Min. :0.009487	Min. :0.01780
1st Qu.:2869	1st Qu.:22.71	1st Qu.:0.010296	1st Qu.:0.02330
Median :4979	Median :23.16	Median :0.010296	Median :0.03274
Mean :4661	Mean :31.08	Mean :0.011336	Mean :0.03096
3rd Qu.:6413	3rd Qu.:43.89	3rd Qu.:0.011816	3rd Qu.:0.03781
Max. :8244	Max. :58.02	Max. :0.028914	Max. :0.06280
NA	NA	NA	NA

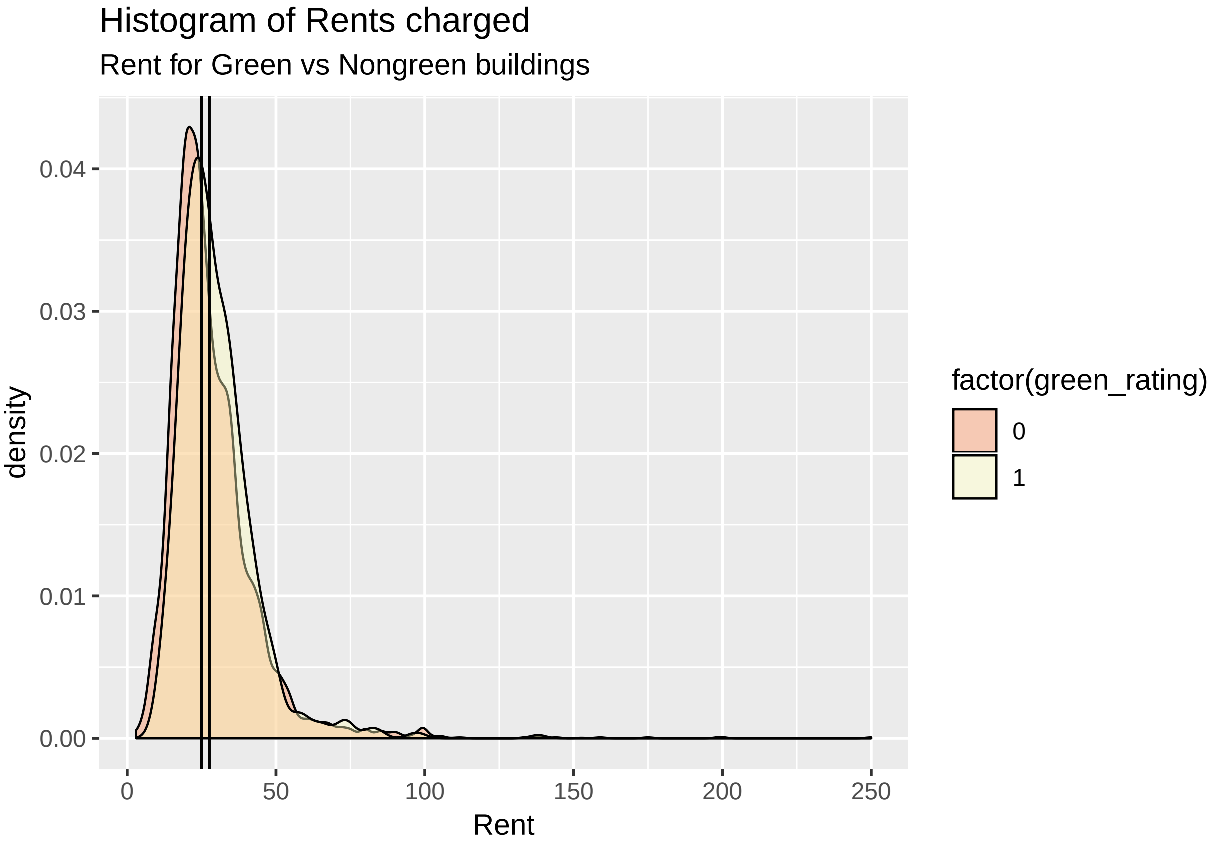
GB.cluster_rent
Min. : 9.00
1st Qu.:20.00
Median :25.14
Mean :27.50
3rd Qu.:34.00
Max. :71.44
NA

By inspection of the summary statistics, the numerical variables appear to behave well and do not send any alarming signals that there is an error. We have made the assumption that all missing values and non-sensible observations were dealt with during the data cleaning process.

The first mistake the "data guru" made when considering his analysis was dropping observations from buildings that have an occupancy rate less than 10%. After calculating that only 215 buildings of the 7,894 buildings in the data set have low occupancy rates, we conclude that it is neccessary, or adivsed to not drop these buildings from the analysis for two reasons. Their existence in the analysis is likely to have very little effect on our outcomes, and there is likely valid reasons for low occupancy, like renovations, over-priced rent, or other specific factors.

```
## [1] 215
```

The next statistic the "excel-guru" looked at was the median market rent grouped by green and non-green buildings. The median rent we calculated was the same as the excel guru's calculation, $25 for non-green buildings and $27.60 for green buildings.



We also thought it would be useful to examine the mean and found that the mean rent for green buildings is $30.02, while the mean rent for nongreen buildings is $28.27. The difference in these means is $1.75.

In the next step of our analysis was to examine the variance of rent of the green and non-green buildings. It is crucial to look at the variance, since the variance will tell us more about how much potential revenue is "garaunteed". This enables us to see the differences between possible profitability for each type of building and estimate the ranges of possible rent prices.

The variance of rent for green buildings was $167.70, or a standard deviation of $12.95. The variance of rent for non-green buildings was $232.70, or a standard deviation of $15.25.

The 95% confidence interval of mean rent for a green building is (4.12, 55.92), while the 95% confidence interval of mean rent for a nongreen building is (-2.24, 58.78). This reveals that rent is highly skewed and that potential revenues can vary greatly. This weakens his analysis of median, or even mean, rent becuase the distributions of rents are hihgly skewed and have high variance. The validity of using this information to preform future calculations on, specifically potential revenues, is inherently flawed.

green_rating	Rent
0	25

green_rating	Rent
1	27.6

green_rating	Rent
0	28.27
1	30.02

green_rating	Rent
0	232.7
1	167.7

```
## [1] 15.25451
```

```
## [1] 12.9499
```

```
## [1] 1.75
```

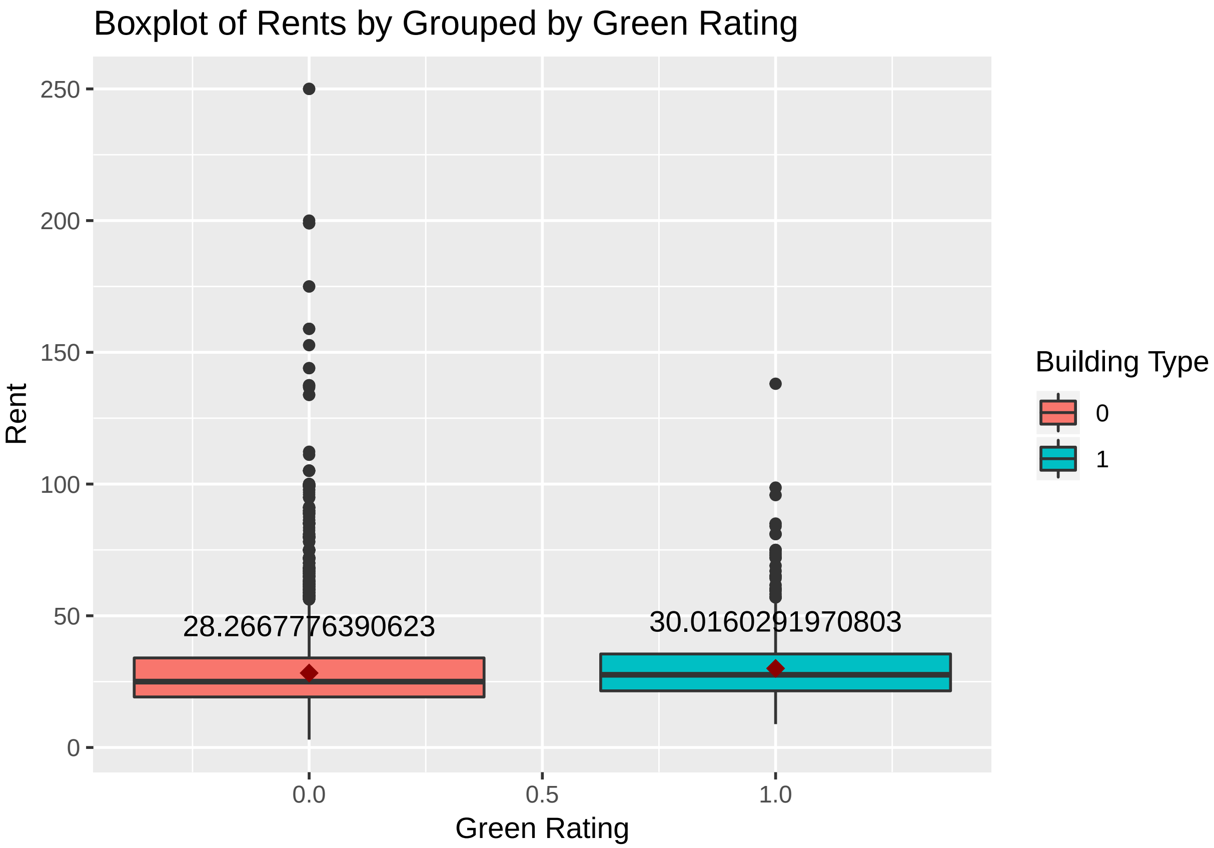
```
## [1] 4.120193
```

```
## [1] 55.91981
```

```
## [1] -2.239015
```

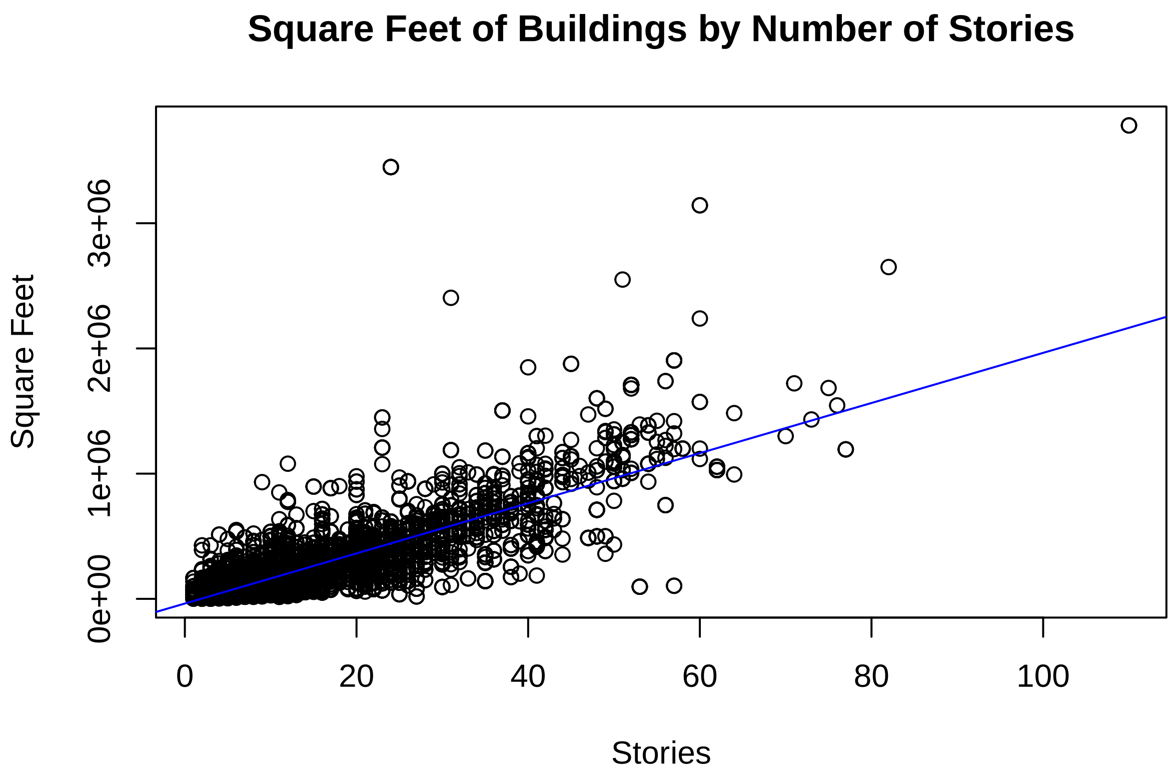
```
## [1] 58.77902
```

```
## Warning: `show_guide` has been deprecated. Please use `show.legend`
## instead.
```



The “data guru” goes on to calculate the extra revenue expected each year for a 250,000 square foot building. He took the difference in median rents, $2.60 and multiplied that by 250,000 sq ft to render extra revenues of $650,000 per year. The first, out of many flaws, of this calculation is that it is unknown as to why he is calculating expected extra revenue for a 250,000 sq ft building, when we only know that the building is 15 stories high. In reality, the expected square feet of a 15 story building is 262,977, refer to the plot below. If using the same difference in median rent to calculate extra revenues, this would render $683,741, which is $33,741 more dollars per year than his original estimate.

```
## (Intercept)     stories
##   -37335.16    20020.83
```



```
## [1] 262977.3
```

```
## [1] 683741
```

The next very obvious flaw of his calculations is that he does not take into account that less than 100% of the 250,000 (actually 262,977) square feet of the building will even be leasable, meaning the owner will not be collecting rent on every square foot of the building, unless the whole building is leased out. The final calculation for excess revenue per year has the possibility to vary greatly since, as we found above, the average rent of both green and non-green buildings vary an awful lot.

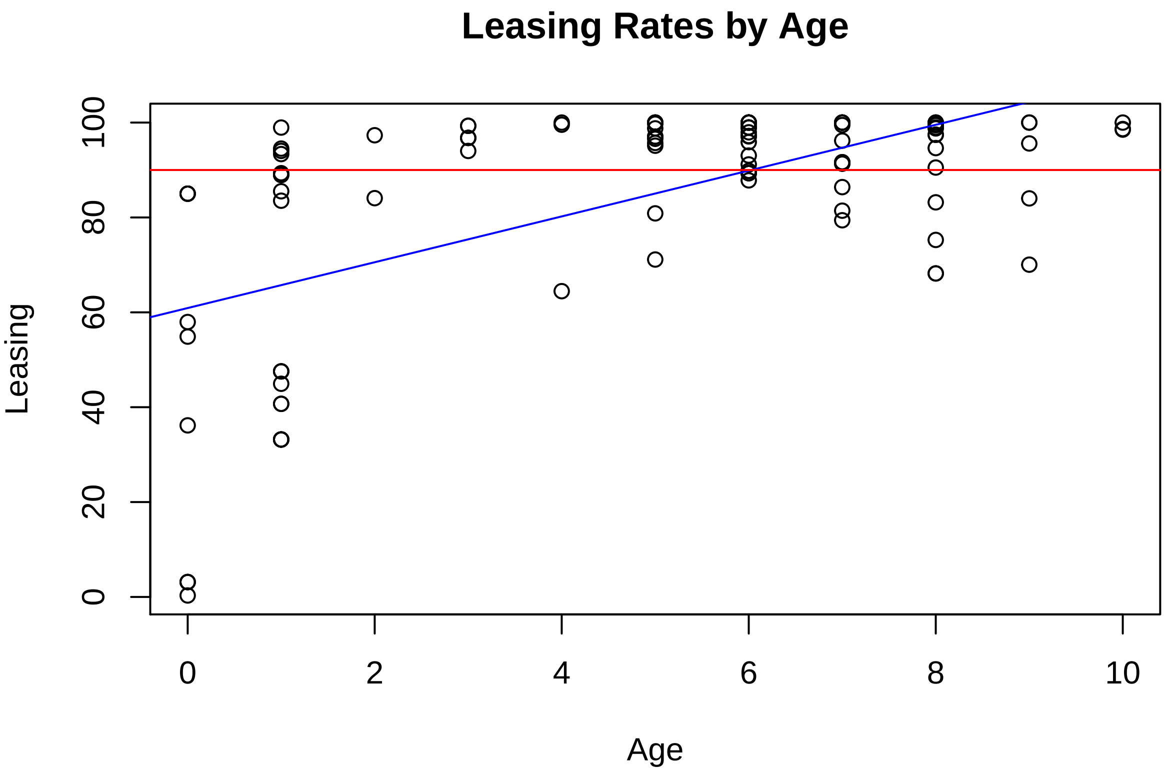
For buildings with rents one standard deviation lower than the average market rent, green buildings can expect an average rent of $17.07 per square foot, while non-green buildings can expect an average rent of $13.02 per square foot. This means that for "lower" end/rent buildings, green buildings still preform better than non-greeen buildings by $4.05 per square foot better to be exact.

For building with rents one standard deviation higher than the average market rent, green buildings can expect an average rent of $42.97 per square foot, while non-green buildings can expect an average rent of $43.52 per square foot. This means that for "higher" end/rent buildings, non-green buildings preform better than greeen buildings by $0.65 per square foot better to be exact.

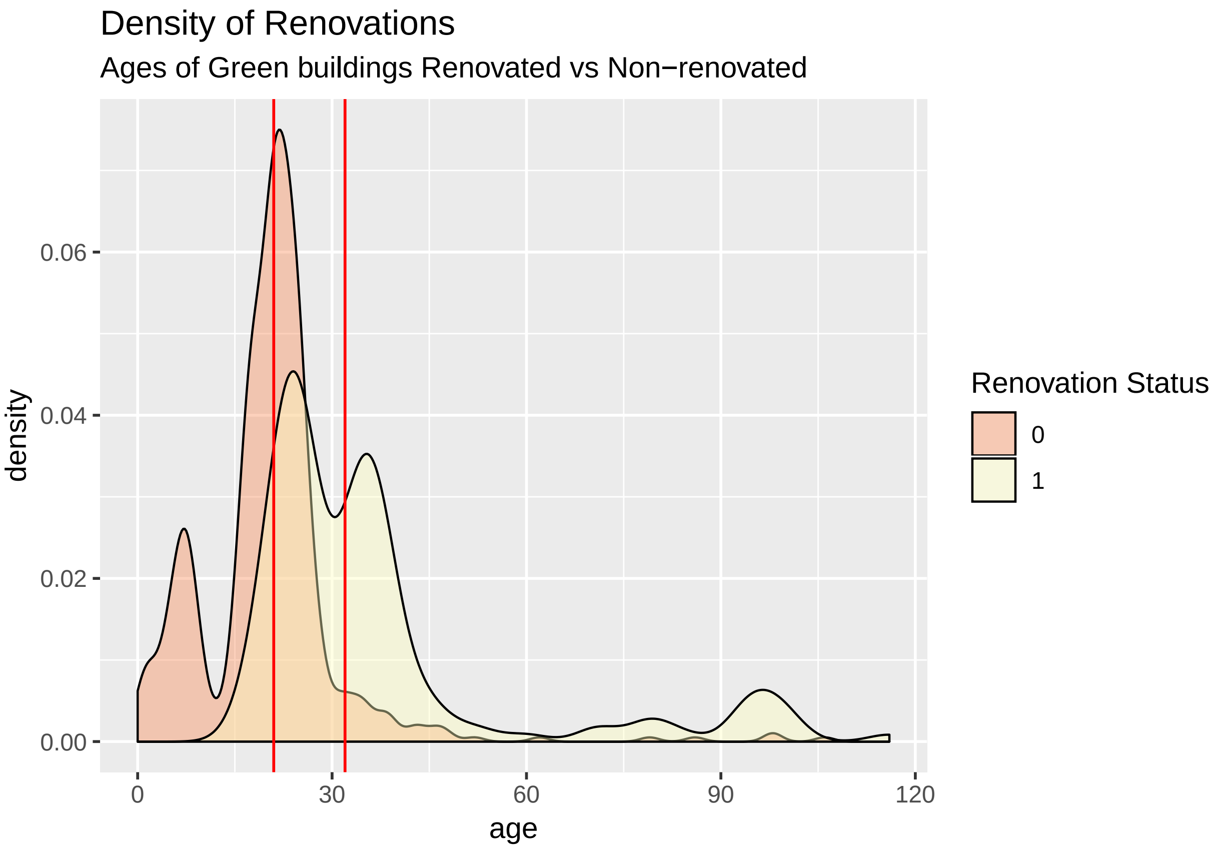
This is an interesting finding, because it clarifies which market finds greater value in energy-cost reducing measures. We found that a green certification appears to have a higher impact on lower rent buildings. We can derive that these potential tenants, while concerned with associating themselves as "eco-friendly", will more likely still place a higher value on their personal savings by preferring to save money on electricity costs. This is an important insight that the "data guru" did not include in his analysis by failing to take the variance of rent into account.

The next part of the analysis that we found to be misleading was the cost recouperation from the costs associated with obtaining green certification. He estimated that total costs for building a green building will be $105 million, which we cannot verify to be true, but will accept as true. Using the values we have obtained and his methodology, the extra revenue that we expect from obtaining a green rating, of $683,741 would be recouperated in 7.3 years.

This calculation still does not make much sense, because it implies 100% occupancy and does not account for variation in rent prices. As we can see from the plot below, new buildings normally do not have 100% occupancy. From what we can see from the market of buildings similar to ours, less than 10 years old and betwee 10 and 20 stories, average leasing rate does not surpass the 90% occupany until approximately 6 years old. This means that we would not recover green rating certification costs at the rate that he proposed, instead it would be significantly slower.

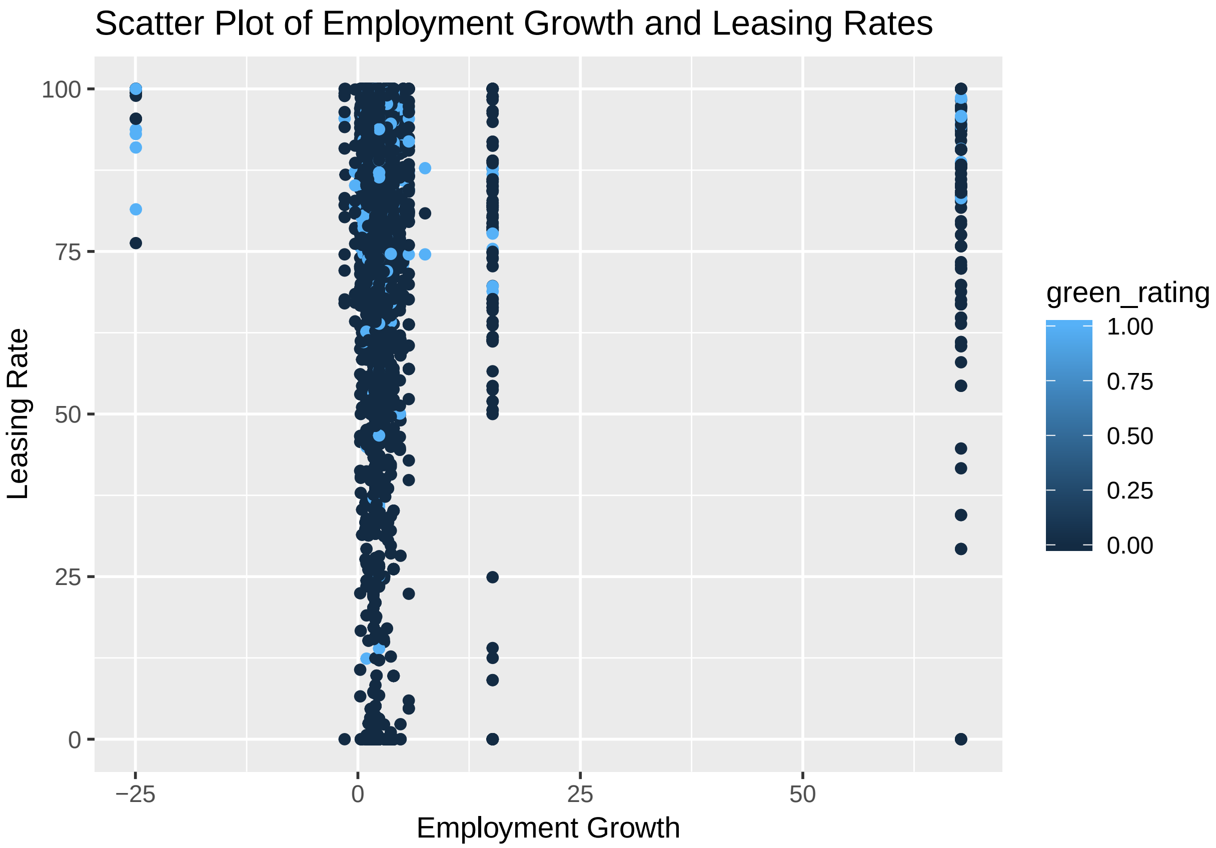


He then states that past the 9 year mark, the owner would be making the $650,000 in green certification revenue as profit for the next thirty years. Clearly the owner will be recovering from certification costs much later than 9 years after completion, and the claim that she will collect profit for the next 30 years is also outlandish. The "data guru" clearly did not take renovations or other unexpected costs into consideration when making that statement. The density plot below reveals that green buildings of 32 years have had at least one renovation in the past. This disputes his claim that we will earn profits for 30 years, since some of the profits will be allocated to renovations within those 30 years.



As an extension to the cost benefit analysis done above, we think it would be important to include how leasing rates are expected to change with employment growth rate. We are not garaunteed a 90% occupancy rate, so we must consider other confounding variables, such as employment growth rates in the region. We are intersted in how the leasing rate will change as the employment growth fluctuates. We notice that for negative employment growth rates, the leasing rate is very high for all buildings. This is interesting because it is not intuitive, this may be because there are such few observations for that employment growth rate. One thing that does stand out is that the leasing rates for green buildings is normally relatively high, it drops below 20 only a few times. This is indicitive that the owners ability to fill spots in the building as employment growth rates fluctuate should not be a major concern to her.

```
## Warning: Removed 74 rows containing missing values (geom_point).
```

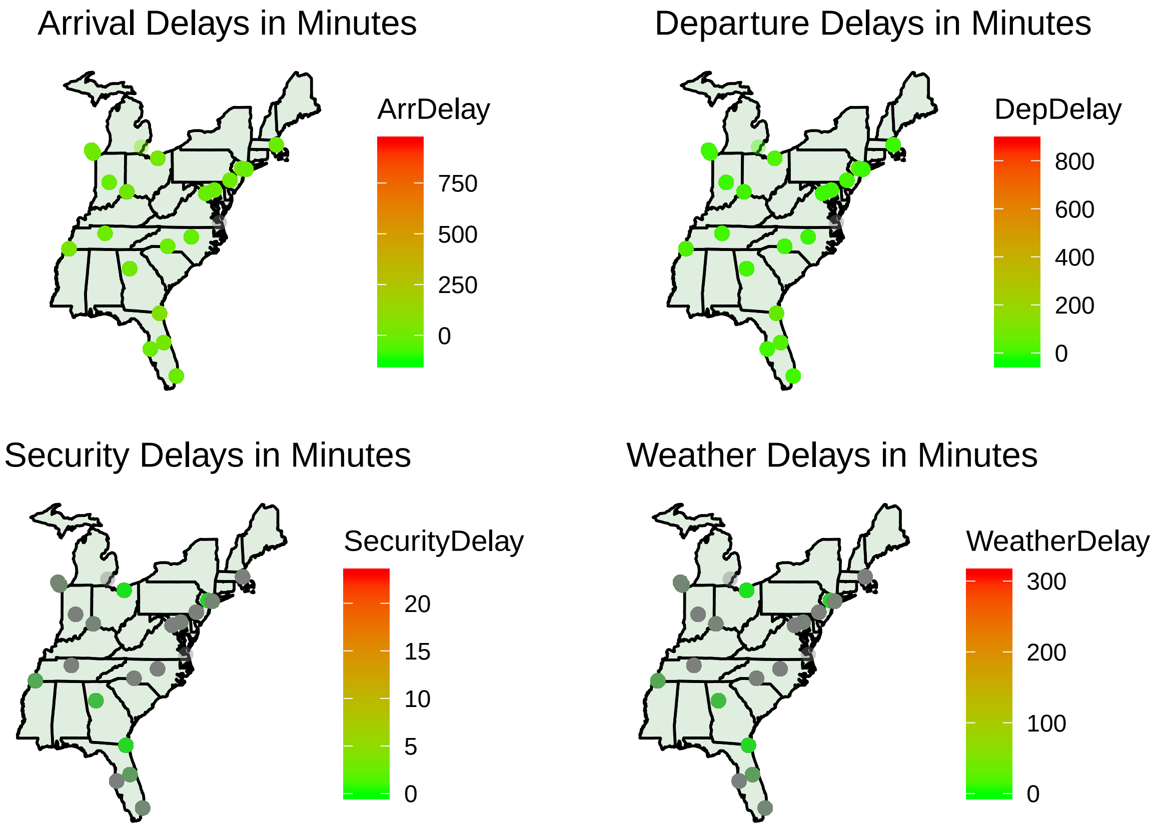


Data Visualization 2: Flights at ABIA

The following visualization shows four different types of delays in minutes for airports on the East coast of the United States. The graphs show us that duration of nearly all delays are low in minutes for flights departing or arriving from Austin International Bergstorm Airport. We also found carrier delays are also very low for airports on the East coast where data was available. The grey points on the graphs signify incomplete observations for those airports, while the colored points are a scale of how long the delay is in minutes. The color green reveals a delay close to zero minutes and the color red is the maximum minutes of delay, which was never reached for this region of the United States.

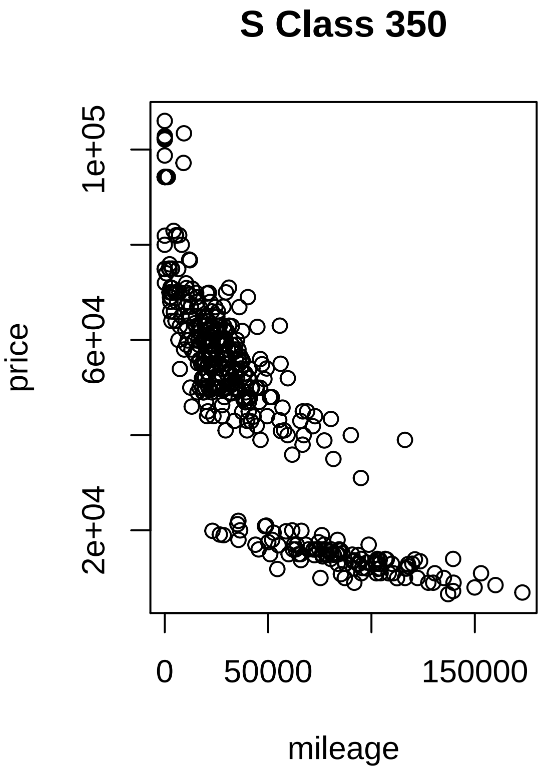
Overall, the data shows us that, on average, delays on the East coast coming to and from Austin did not comsume many minutes of air-travelers' time and might reveal some information about the efficiency of Austin Bergstorm International Airport.

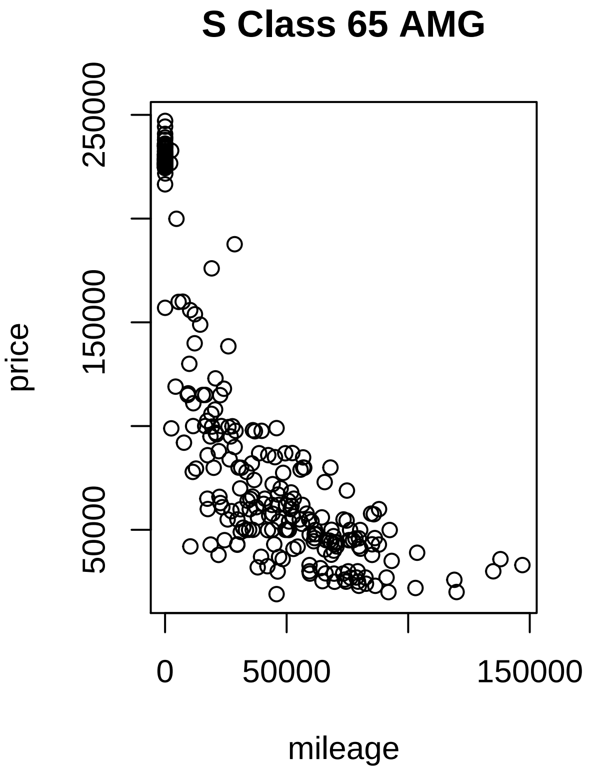
```
##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map

## Source : https://maps.googleapis.com/maps/api/staticmap?center=World&zoom=4&size=640x640&scale=2&map

## Source : https://maps.googleapis.com/maps/api/geocode/json?address=World&key=xxx-8XB9V3HPyEo
```


Arrival Delays in Minutes
ArrDelay
750
500
250
0
Departure Delays in Minutes
DepDelay
800
600
400
200
0
Security Delays in Minutes
SecurityDelay
20
15
10
5
0
Weather Delays in Minutes
WeatherDelay
300
200
100
0

Regression vs. KNN

First, we visualize what our data looks like. From the two scatter plots below, we can deduce that as the mileage of the car increases, the sale price of the car decreases. This is a good finding, because it follows inuition.





The following tables represent the RMSEs of the the simple linear models, linear model with a polynomial, and the 7 KNN models with varying number of neighbors Mercedes Benz S Class 350. We can see that in this case, the KNN with around 60 neighbors does the best job at predicting price of the S Class 350, with the smallest RMSE of 10721. The plots show the predictions of KNN estimates with varying neighbors and an orange line which is the fit of the linear model with mileage squared as the explanatory variables. We can see that as the number of neighbors increases, the error of the predictions increases. In this case, the optimal value of K is 60.

Table 8: Table continues below

rmse.ytest..ypred_lm1.	rmse.ytest..ypred_lm2.	rmse.ytest..ypred_knn4.
10975	10161	10427

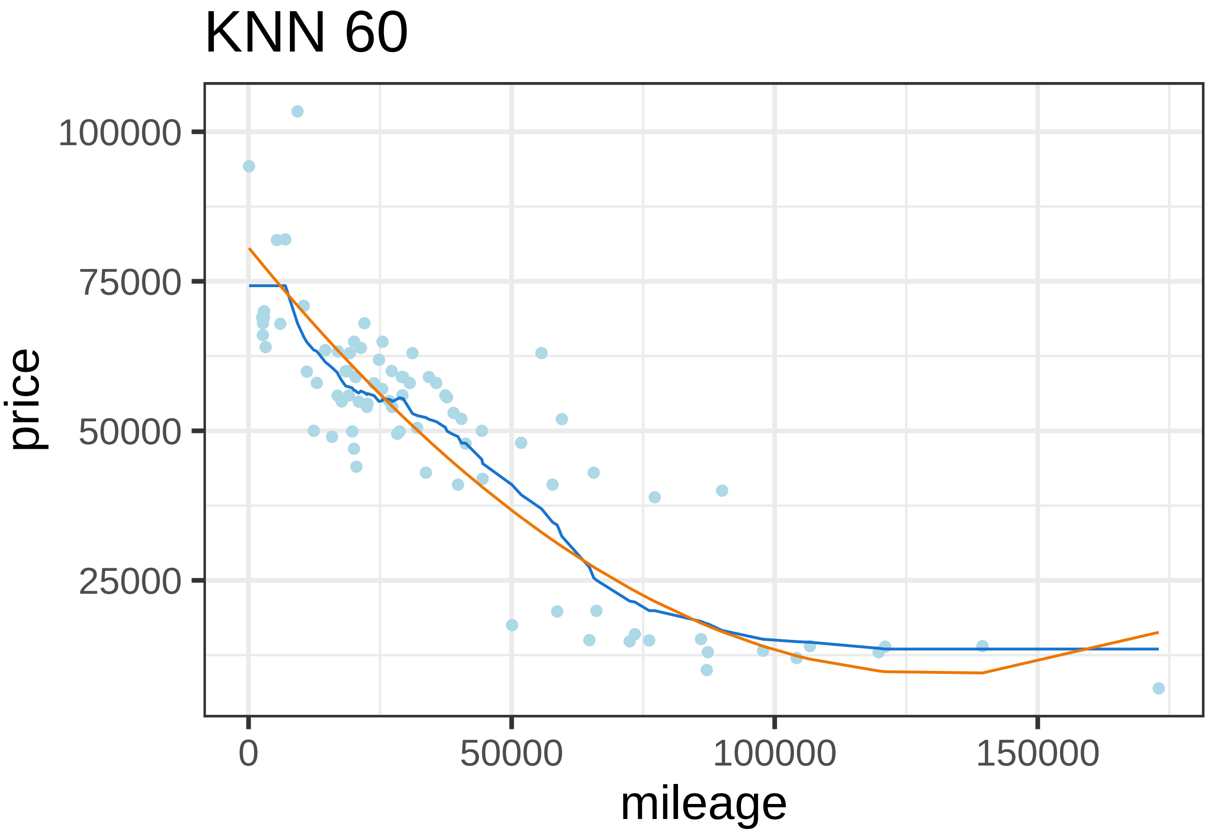
Table 9: Table continues below

rmse.ytest..ypred_knn10.	rmse.ytest..ypred_knn30.	rmse.ytest..ypred_knn60.
9964	10174	9439

Table 10: Table continues below

rmse.ytest..ypred_knn80.	rmse.ytest..ypred_knn100.
9353	9714

rmse.ytest..ypred_knn300.
18254



The following tables represent the RMSEs of the the simple linear models, linear model with a polynomial, and the 7 KNN models with varying number of neighbors for Mercedes Benz S Class 65 AMG. We can see that in this case, the KNN with 10 neighbors does the best job at predicting price of the S Class 65 AMG, with the smallest RMSE of 23719. The plots show the predictions of KNN estimates with varying neighbors and an orange line which is the fit of the linear model with mileage squared as the explanatory variables. We can see that as the number of neighbors increases, the error of the predictions dramatically increases. In this case, the optimal value of K is approximately 10.

Table 12: Table continues below

rmse.ytest..ypred_lm1.	rmse.ytest..ypred_lm2.	rmse.ytest..ypred_knn4.
43077	29757	25089

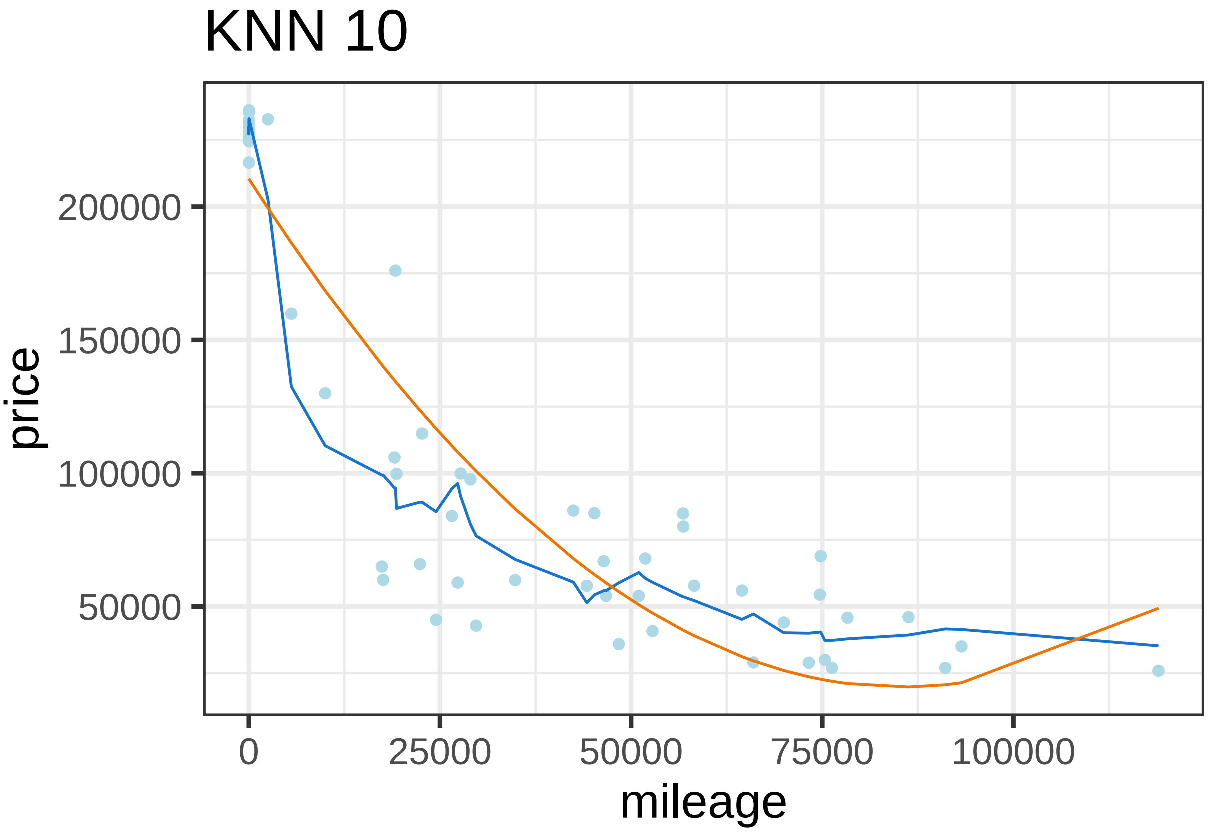
Table 13: Table continues below

rmse.ytest..ypred_knn10.	rmse.ytest..ypred_knn30.	rmse.ytest..ypred_knn60.
20469	19563	22335

Table 14: Table continues below

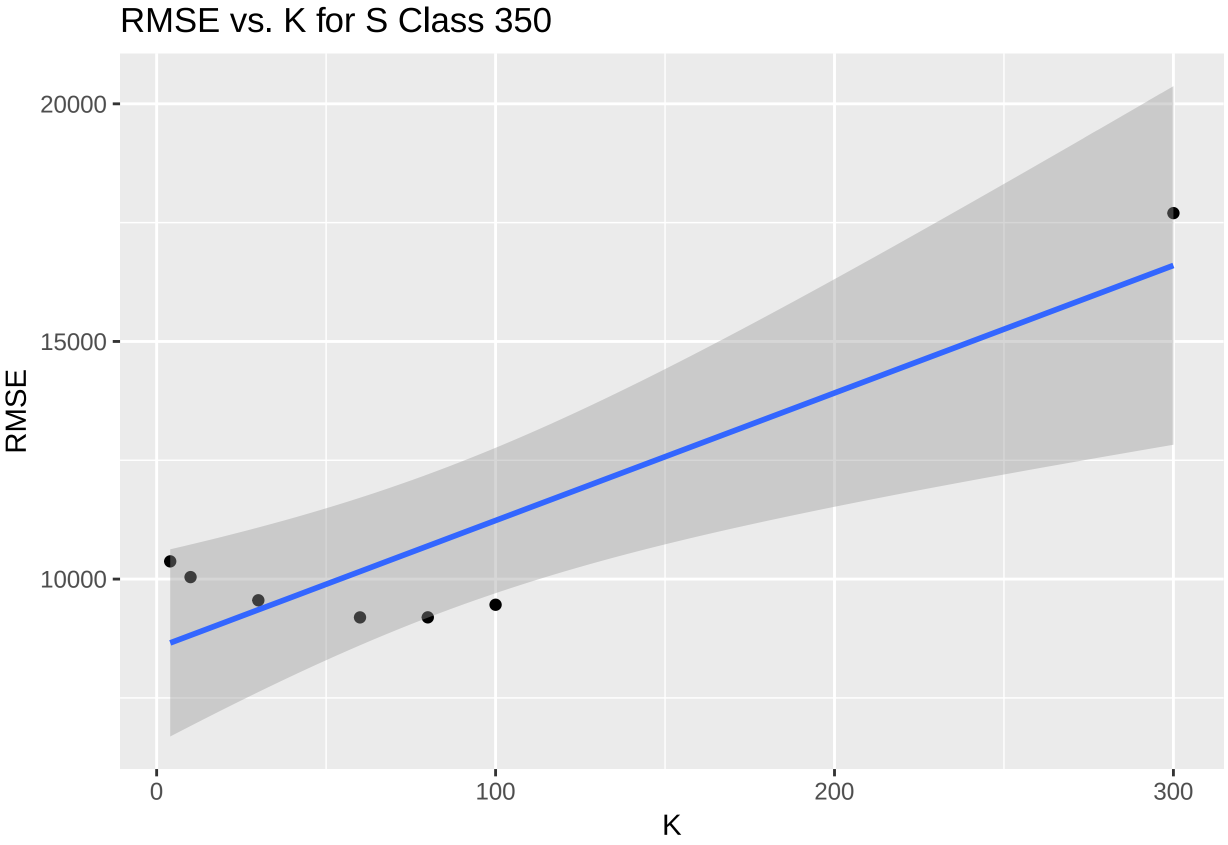
rmse.ytest..ypred__knn80.	rmse.ytest..ypred__knn100.
24938	32692

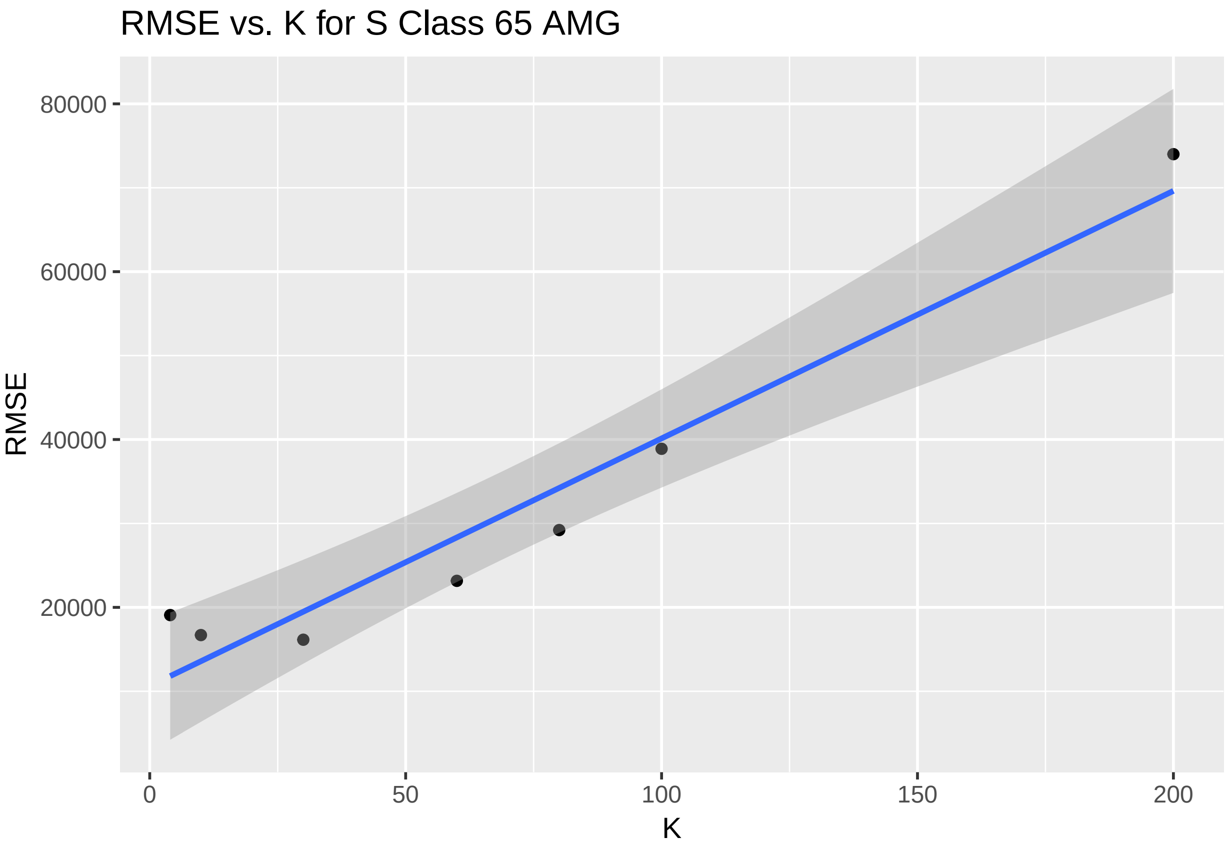
rmse.ytest..ypred__knn300.
70240



The two following graphs show the movement of the RMSE as the number of neighbors increases for both types of S Class Mercedes Benz. They appear to move in very similar directions, having RMSE increase with K. Looking at the plot of RMSE against K for the S Class 350, we see that as K approaches 300 and further, the RMSE becomes much larger. Since we are looking for the smallest RMSE, we verify that the optimal K for this class of car is closer to 60.
Looking at the plot of RMSE against K for the S Class 65 AMG's, we see that as the number of neighbors we use to predict price increases, the RMSE also increases. One thing to make note of is the magnitude of the RMSE for this subset of cars. The RMSE is much higher in magnitude than its 350 counterpart. We are able to verify that the optimal K for this class is close to 10. The higher K reveals that we are estimating f(x) using many points, possibly far away (this increases bias), and the lower K reveals that we are using not very many points that are likely close by (this reduces bias).

Each class of car is telling us different stories about what the optimal K is. The S Class 350 is revealing that a slightly larger K is optimal, whereas the S Class 65 AMG is telling us that a small K is optimal. The reason that this may be is becuase of the size, quality and behavior of the data sets. Data sets with large numbers of observations are inherently better and normally behave better than data sets with low numbers of observations. The KNN estimations for the S Class 65 AMG price have quite a few less observations than the S Class 350. Although there is more data for the S Class 350, it appears to contain more clusters than the S Class 65 AMG data. Since the data appears to have clusters, the higher optimal K value for the S Class 350 reveals that the we are likely having to use points slightly further away to estimate f(x), compared to the S Class 65 AMG.




RMSE vs. K for S Class 65 AMG
RMSE
80000
60000
40000
20000
K
0
50
100
150
200