HW1

Madeline Lin

2/10/2019

R Markdown

Question 1:

Question 1 Probelm Summary:

Environmentally conscious buildings have tangible and intangible benefits. The decision whether it is worthwhile to invest in eco-friendly buildings or not has under fierce discussion. In this problem, a data guru thinks it is a good financial move to build the green building upon his analysis. However, the reasons he gives for advocating building green buildings are not so convincing.

What we need to do: If we agree with the data guru, we need to elaborate more to support building eco-friendly buildings. If not, we should explain where and why the analysis goes wrong, and how it can be improved.

From my perspective, I agree with the idea that investing in a green building will be worth it from an economics perspective. Nevertheless, the reasons presented by the data guru base on some assumptions that are not very solid. Therefore, I would like to point out the unsolid assumptions and elaborate more evidence in support of building eco-friendly buildings.

I think the data guru's way to clean data makes sense, so I clean the data first and foremost.

```
greenbuildings_cleandata <- subset(greenbuildings, age<=116 & leasing_rate>=10)
```

Firstly, the data guru supposes that the rent is constant in green buildings(27.6$ per square foot per year) and non-green buildings(25$ per square foot per year). However, this is not accurate. If we extract the buildings into different categories, that is, dividing them into sub sets, we will discover that the rent does not remain constant.

For example, one variable called net is an indicator whether tenants pay their own utility costs or not. Intuitively, if rent is quoted on a ''net contract''basis, then the rents will be higher than''gross contract" basis rent. So, I divide greenbuildings and non-green building into two subsets separately.

```
green_net = subset(greenbuildings_cleandata, green_rating == 1 & net == 1)
dim(green_net)
```

```
## [1] 39 23
```

```
green_notnet = subset(greenbuildings_cleandata, green_rating == 1 & net == 0)
dim(green_notnet)
```

```
## [1] 645  23
```

```
notgreen_net = subset(greenbuildings_cleandata, green_rating == 0 & net == 1)
dim(notgreen_net)
```

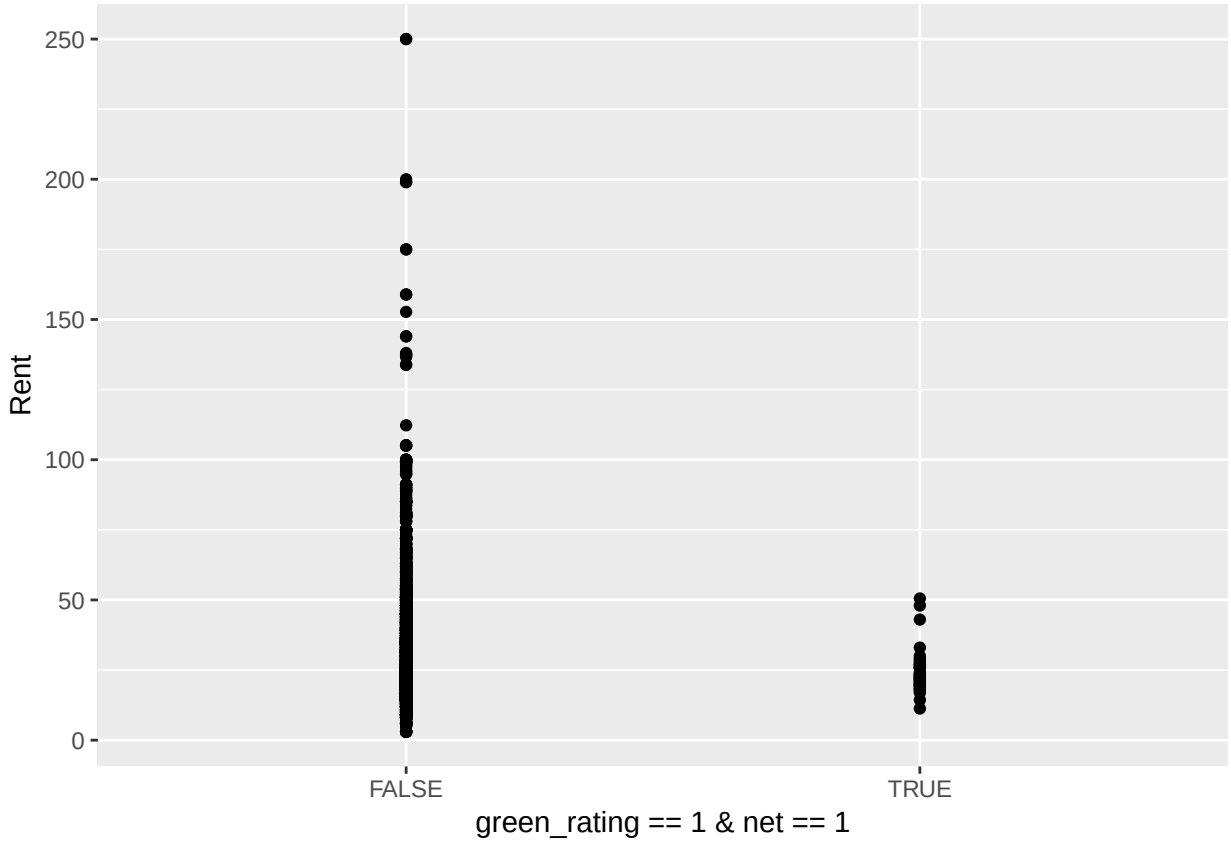
```
## [1] 234  23
```

```
notgreen_notnet = subset(greenbuildings_cleandata, green_rating == 0 & net == 0)
dim(notgreen_notnet)
```

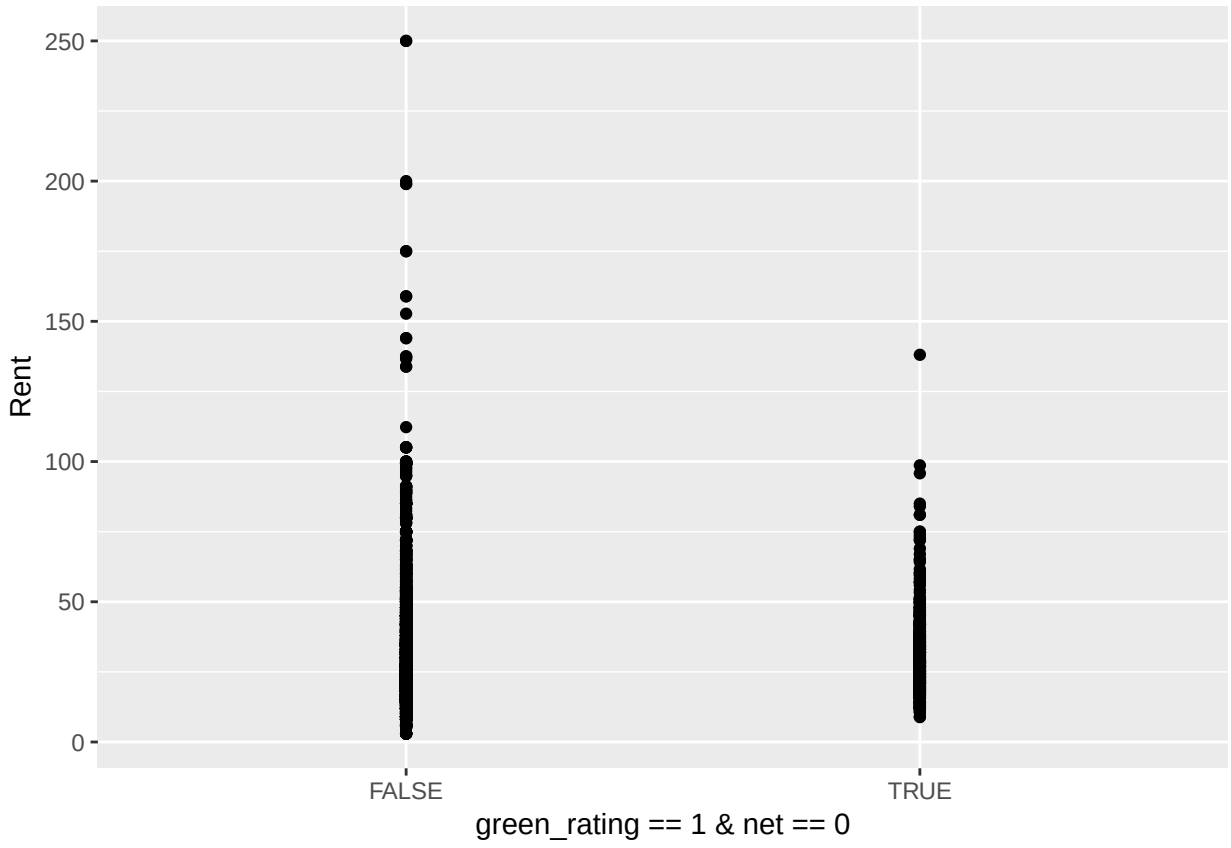
```
## [1] 6626   23
```

Then, we can look at the relationship between rent and green_net/green_notnet (between rent and notgreen_net/notgreen_notnet).

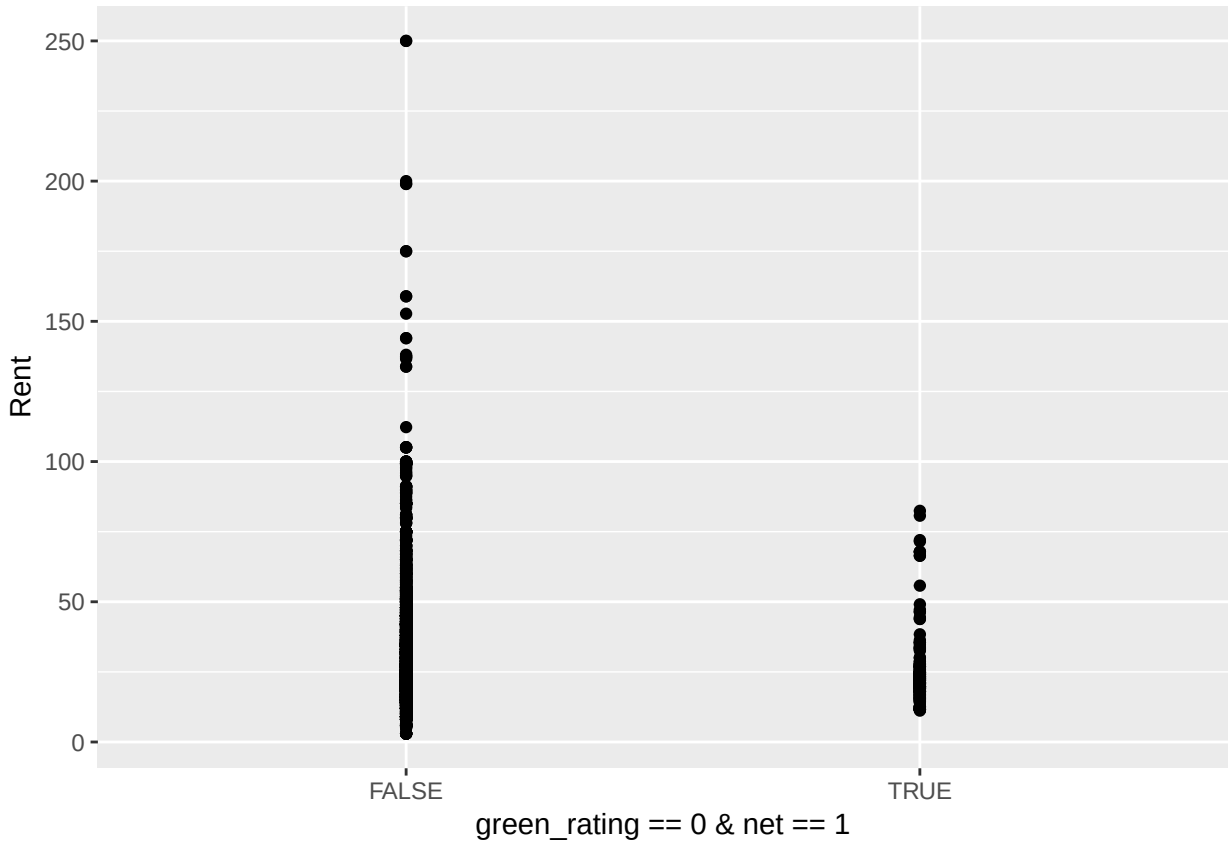
```
ggplot(data = greenbuildings_cleandata) +
  geom_point(mapping = aes(x = green_rating == 1 & net == 1, y = Rent))
```



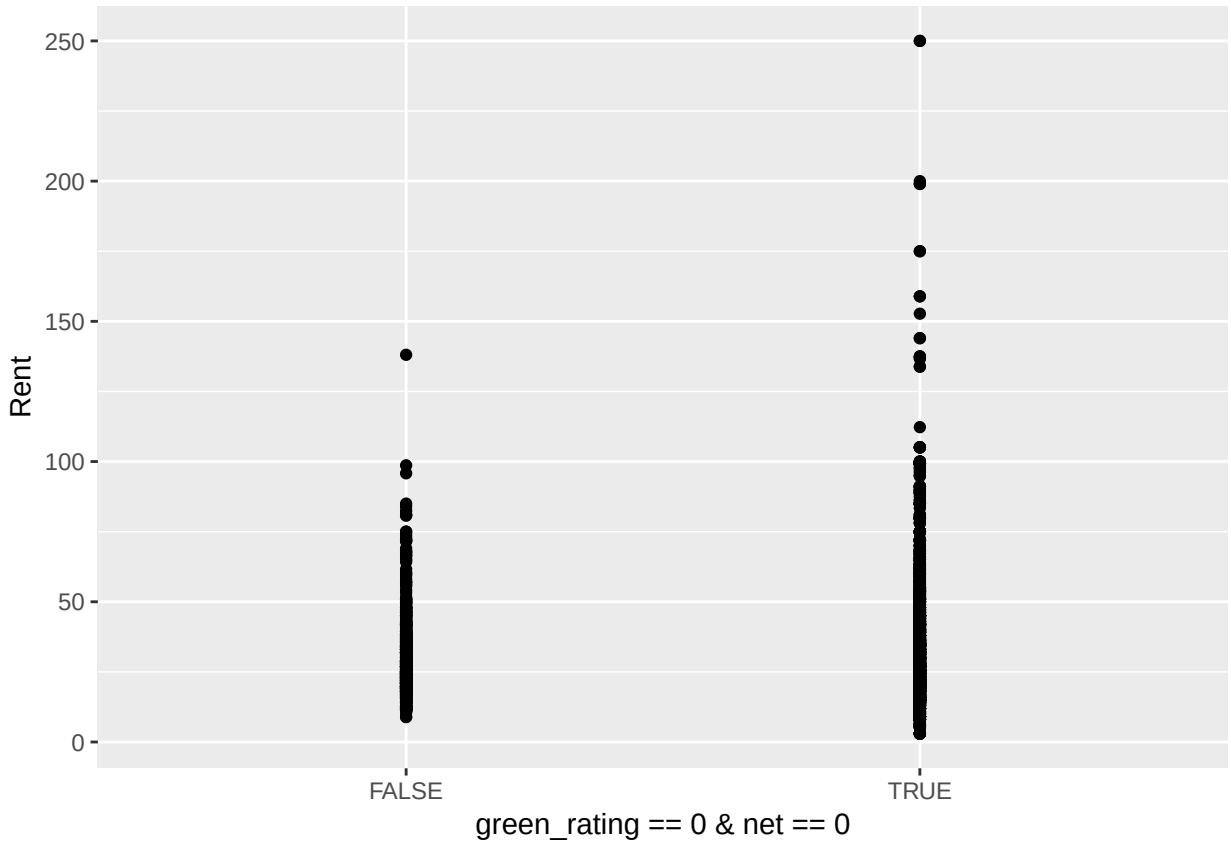

```
ggplot(data = greenbuildings_cleandata) +
  geom_point(mapping = aes(x = green_rating == 1 & net == 0, y = Rent))
```




```
ggplot(data = greenbuildings_cleandata) +
  geom_point(mapping = aes(x = green_rating == 0 & net == 1, y = Rent))
```




```
ggplot(data = greenbuildings_cleandata) +
  geom_point(mapping = aes(x = green_rating == 0 & net == 0, y = Rent))
```

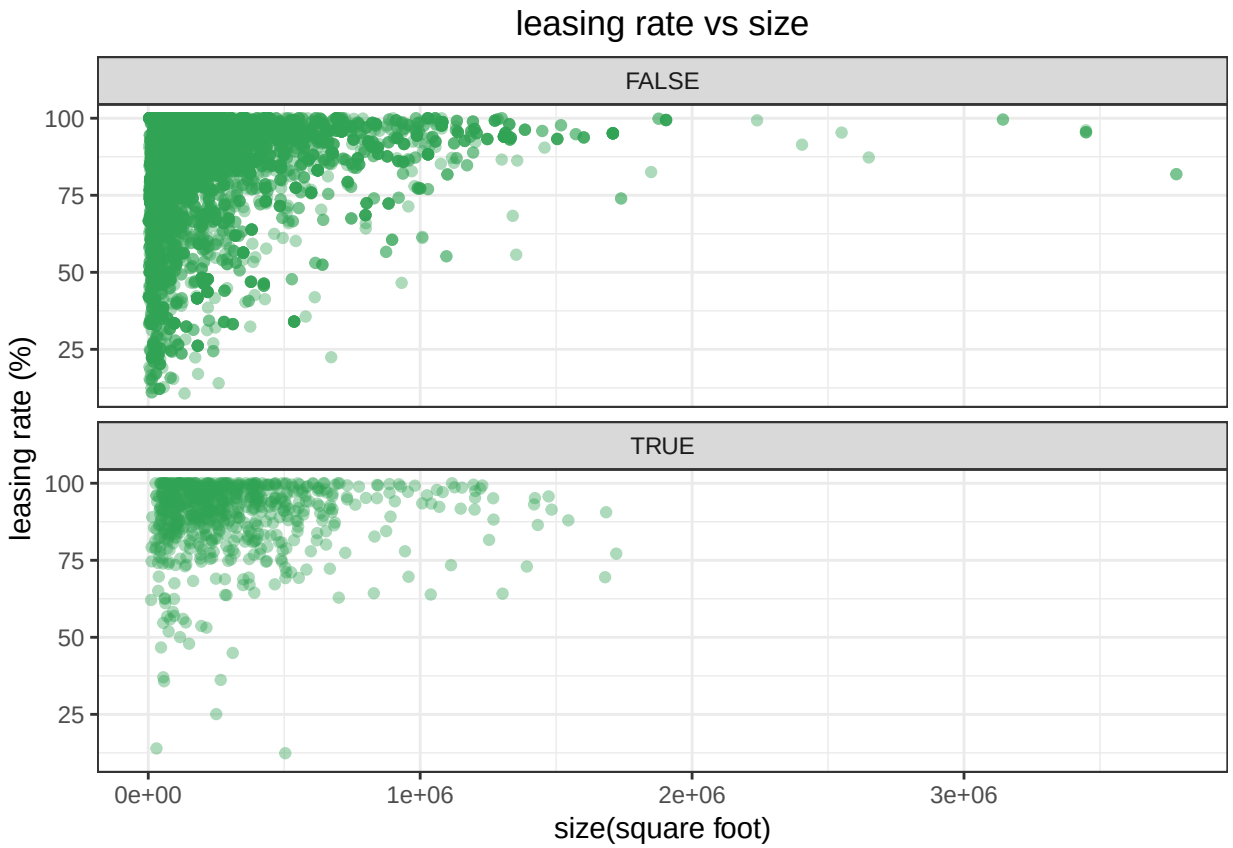


As we can see in the above four graphs, the rent of different subsets vary differently. So, we can not use two overall rates which are 27.6$ and 25$ to calculate the cost difference between reen buildings and non-green buildings.

Furthermore, since the data guru uses the 27.6$ per square foot per year for green buildings and 25$ per square foot per year, he supposes that all buildings have the similar leasing rate after he scrubs low occupancy buidlings from the data set. However, this again is not accurate. We need to analyze leasing rate according to different variables such as size, story and age.

For example, let's try the relationship between leasing_rate and size.

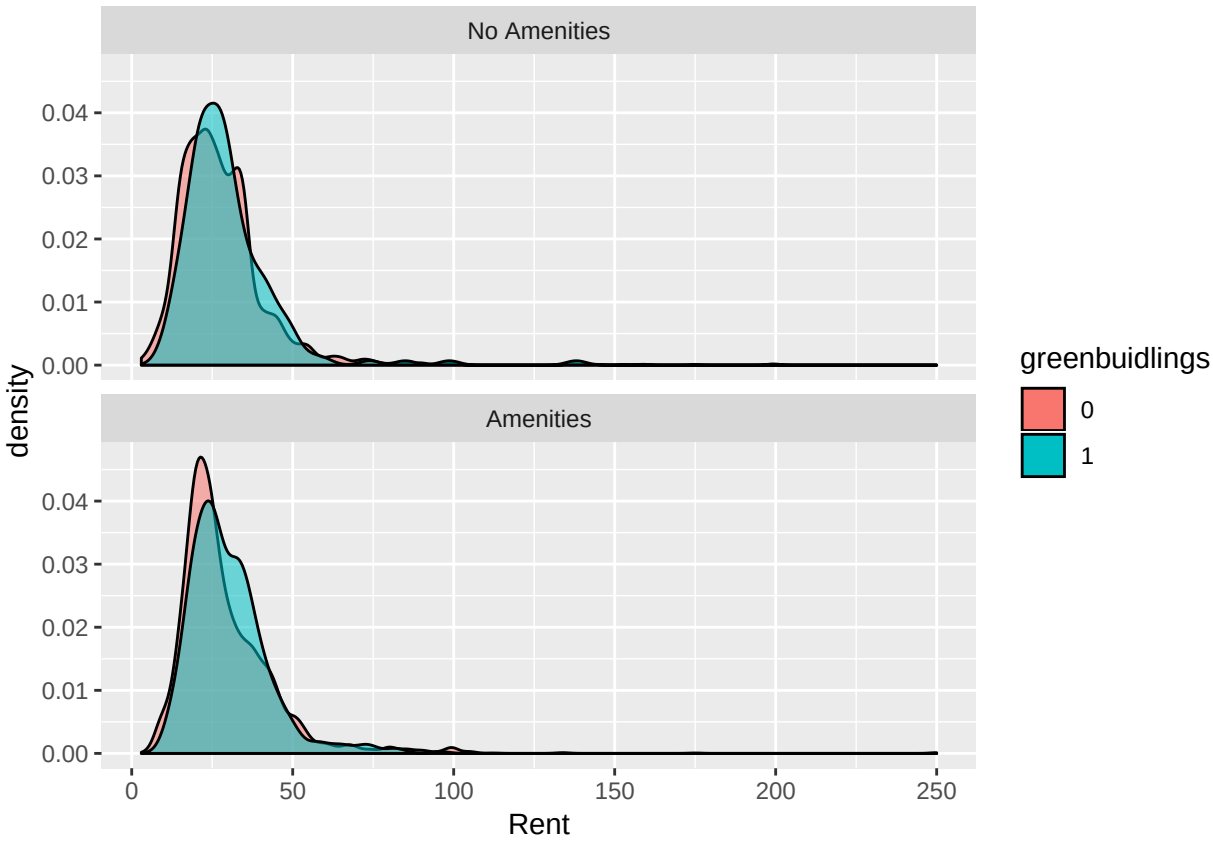
```
# facets
ggplot(data = greenbuildings_cleandata) +
  facet_wrap(~ green_rating == 1, nrow = 2)+
  geom_point(mapping = aes(x = size, y = leasing_rate), alpha = 0.4, col = brewer.pal(6, "Greens")[5]) +
  labs(title = "leasing rate vs size",
       y= "leasing rate (%)",
       x="size(square foot)")+
  theme_bw()+
  theme(plot.title=element_text(hjust = 0.5))
```



In the above two graphs, it is indicated that there are not much discrepancy of leasing rate as size goes up in both green buildings and none-greenbuildings, but, the leasing rate is not always 100%, some even are less than 25%. So, it is not sensible to say that all buildings have similar occupancy rate, which means that we cannot use the same median market rent since not all tenants will pay the rent.

Additionally, we can explore the costs including amentities, gas costs and electricity costs to determine whether it is worthwhile to invest on green buildings in the long run. Let's take amentities for example.

```
gg=ggplot(data=greenbuildings_cleandata, aes(x= Rent))
gg+geom_density(aes(fill = factor(green_rating), alpha = 0.4))+
  facet_wrap(~amenities, nrow = 2,labeller = labeller(amenities = c(`0` = "No Amenities", `1` = "Amenit
  labs( fill = "greenbuidlings")+
  scale_alpha(guide = 'none')
```



As it is shown in the two graphs, greenbuildings with amenities have larger differences in rent compared to greenbuildings with no amenities. Therefore, amenities paly a role in determining the rent of greenbuildings. So do gas costs and electricity costs as we can conjecture.

Conclusion: Since there are so many confounding variables to be taken consideration, we need to make analysis to produce better apples-to-apples comparison. It is not simply just assume the rent remains constant in terms of different sizes, stories, ages (and so on) of a building. We should support to build the green building for its shorter payback duration but not merely based on the data guru's reasons. I am sorry that I cannot provide a quantitive way to calculate the payback duration for investing greenbuildings because I don't know how to find an exact function. But indeed I provide lots data visualizations here to validate my opinion that investing in greenbuildings is a right decision.

Question 2

Question 2 Probelm Summary:

Flights at ABIA contains information on every commercial flight in 2008 that eitherdeparted from or landed at Austin-Bergstrom International Airport.

What we need to do:

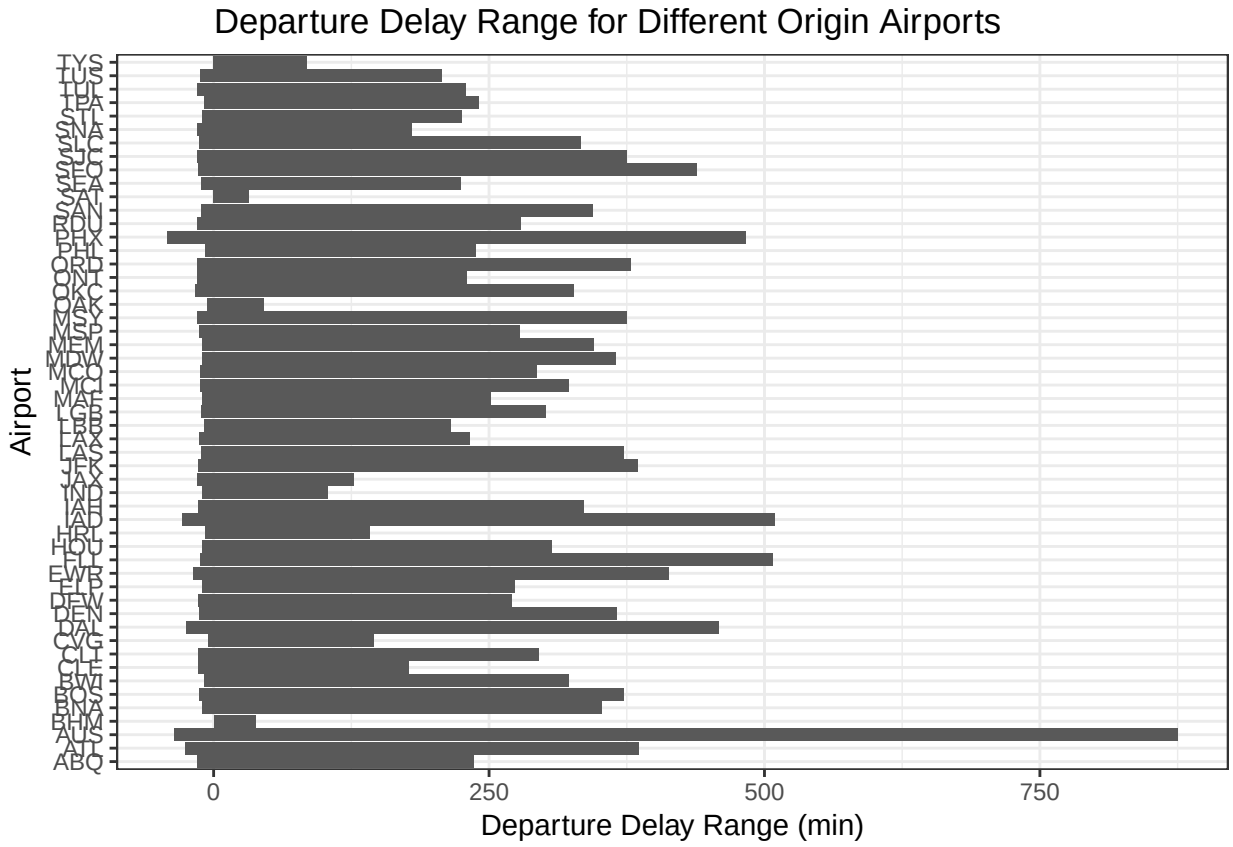
Create a figure, or set of related figures, that tell an interesting story about flights into and out of Austin. As for me, I would like to explore "What are the bad airports to fly to?" My analysis is as follows:

From our perspective, bad airports are airports which ususally have long departure delay range and long arrival delay range. Our logic here is to use a flipped table to show the departure and arrival delay range for each origin and destination airport during 2008, and, to see which airport has the longest or longer delay range compared to other airports.

Firstly, we should clean the data by not considering cancelled and diverted subsets since they are not included in the departure delays or arrival delays samples.

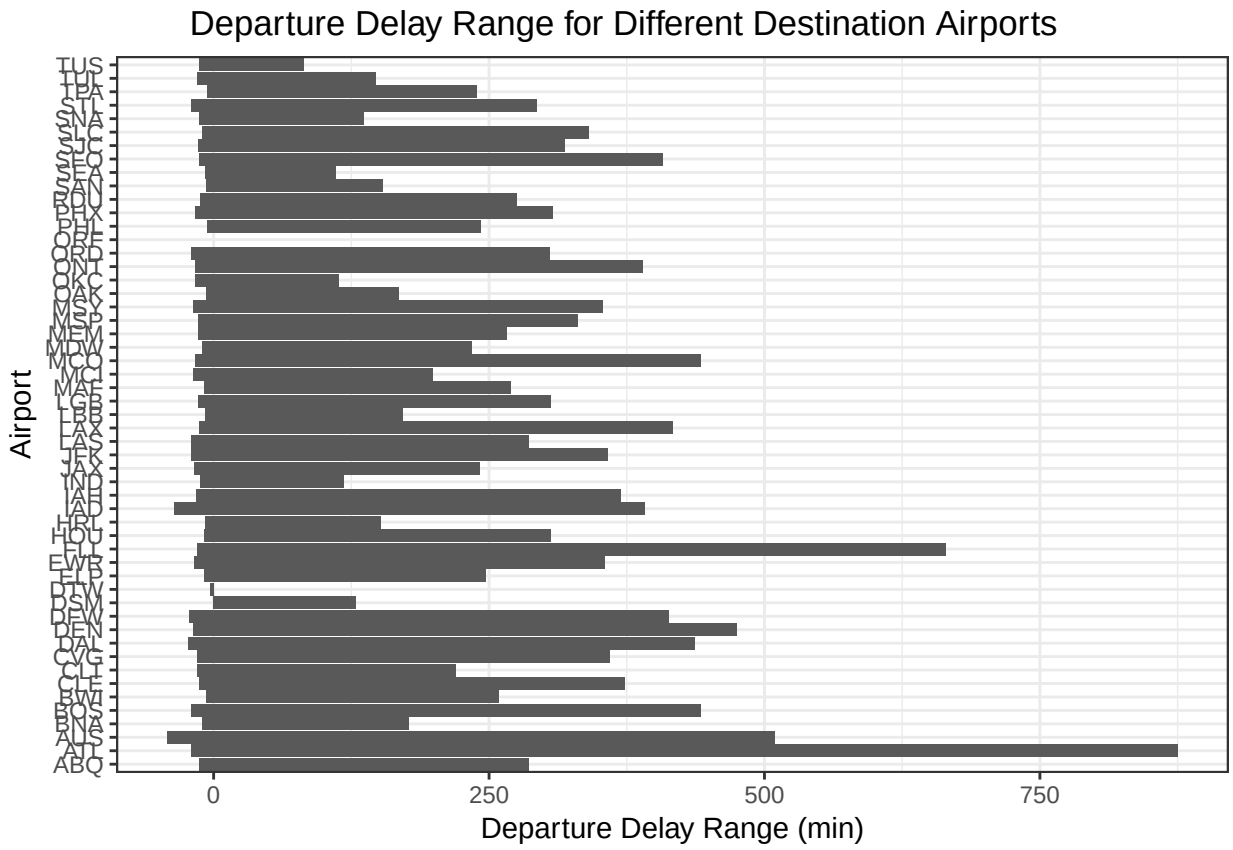
```
ggplot(data = ABIA)+
  geom_bar(aes(x = Origin, y = DepDelay),stat='identity',position='dodge')+
  coord_flip()+
  labs(title = "Departure Delay Range for Different Origin Airports", y = "Departure Delay Range (min)"
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.3))
```

```
## Warning: Removed 1413 rows containing missing values (geom_bar).
```



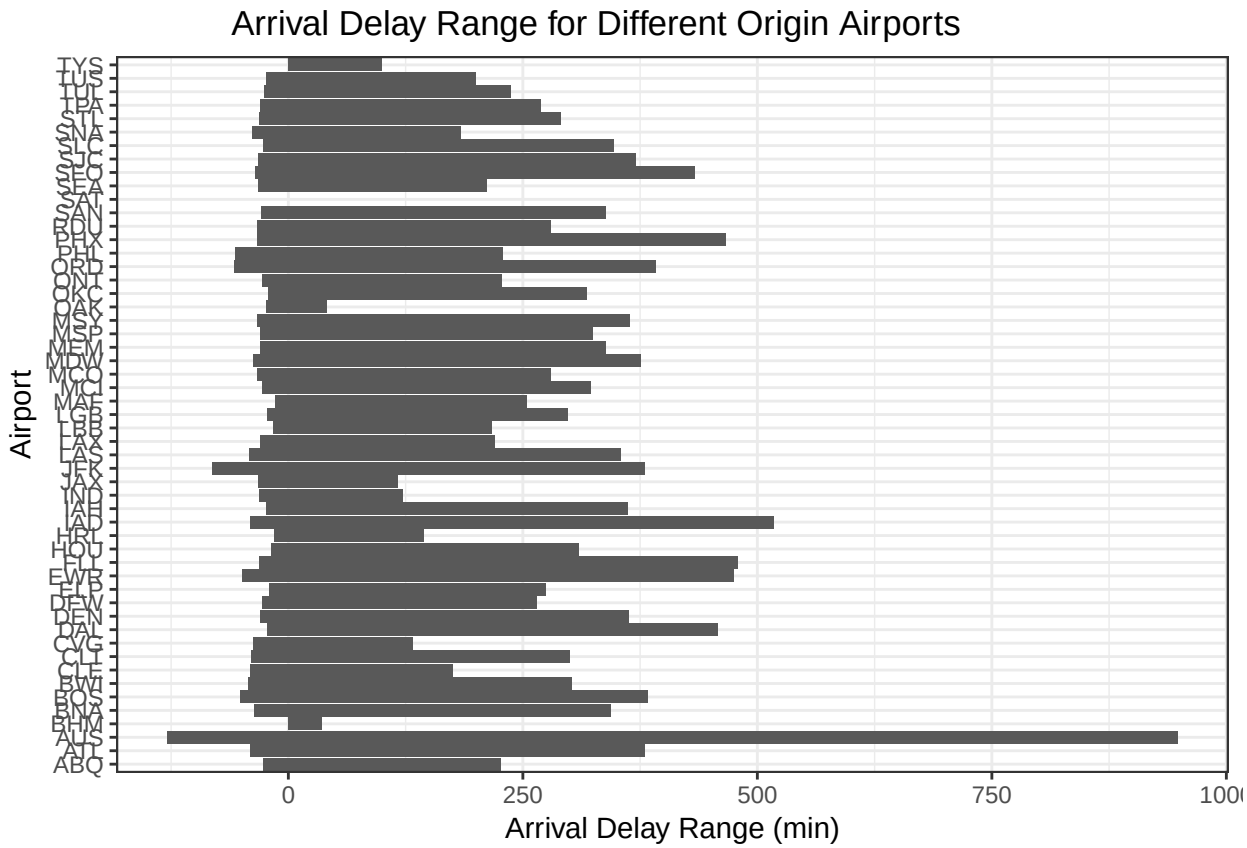

```
ggplot(data = ABIA)+
  geom_bar(aes(x =Dest, y = DepDelay),stat='identity',position='dodge')+
  coord_flip()+
  labs(title = "Departure Delay Range for Different Destination Airports", y = "Departure Delay Range (
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.3))
```

```
## Warning: Removed 1413 rows containing missing values (geom_bar).
```



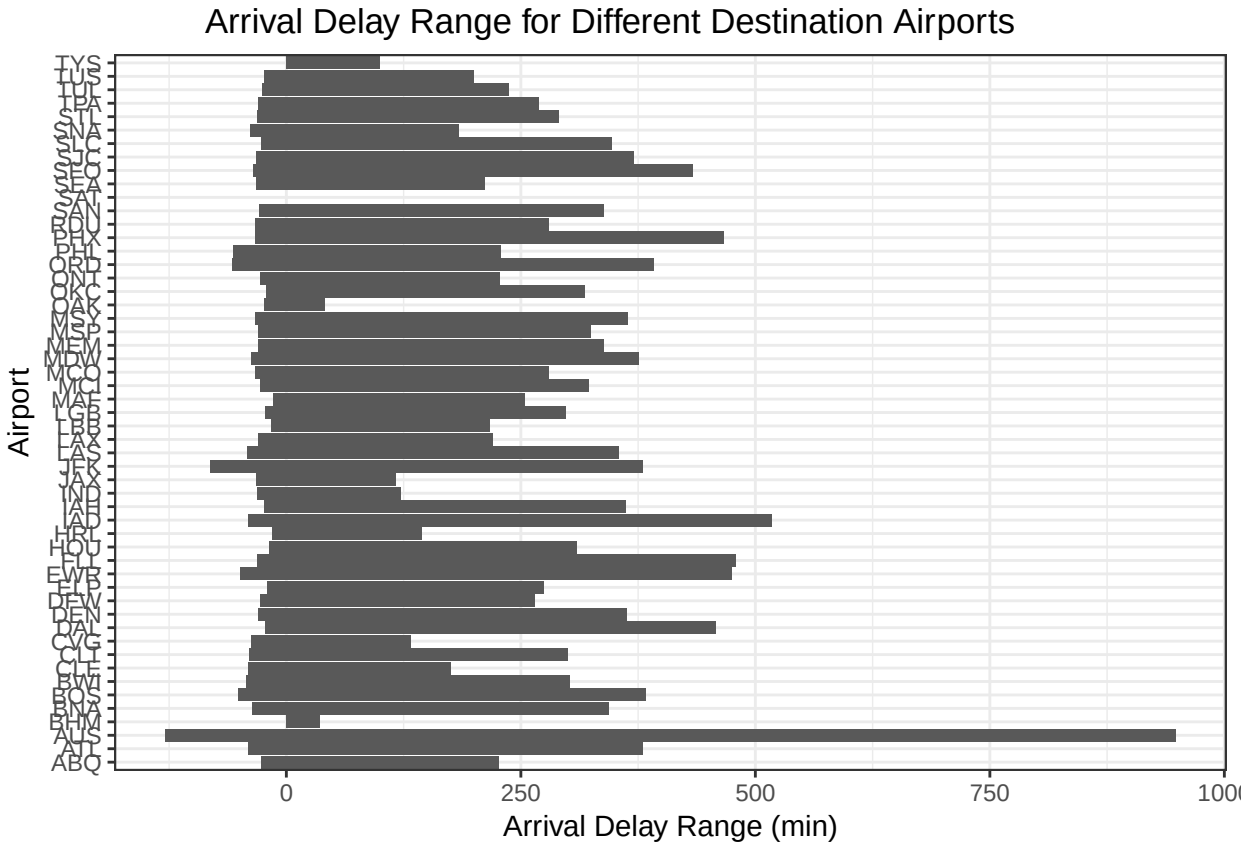

```
ggplot(data = ABIA)+
  geom_bar(aes(x = Origin, y = ArrDelay),stat='identity',position='dodge')+
  coord_flip()+
  labs(title = "Arrival Delay Range for Different Origin Airports", y = "Arrival Delay Range (min)", x =
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.3))
```

```
## Warning: Removed 1601 rows containing missing values (geom_bar).
```




```
ggplot(data = ABIA)+
  geom_bar(aes(x = Origin, y = ArrDelay),stat='identity',position='dodge')+
  coord_flip()+
  labs(title = "Arrival Delay Range for Different Destination Airports", y = "Arrival Delay Range (min)
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.3))
```

```
## Warning: Removed 1601 rows containing missing values (geom_bar).
```



In Conclusion, from these four graphs, we can see that AUS has the longest departure and arrival delays regardless of it is the original airport or destination airport, which is pretty surprising since I don't have much experience of delay when I depart from Austin or arrive at Austin. If you zoom in the graph(sorry in the pdf the graph is not very clear, a little bit crowded), you can also discover that IAD has a pretty long departure delay as for origin airport; and, FLL has a long departure delay as for destination airport.

Actually, I suppose my analysis will be more accurate if I can calculate the average departure and arrival delay for different origin and destination airport. However, I am not sure how to write the code about it.

All in all, these four graphs can roughly show what the bad airports fly to and I hope it will help.

Question 3

Question 3 Problem Summary:

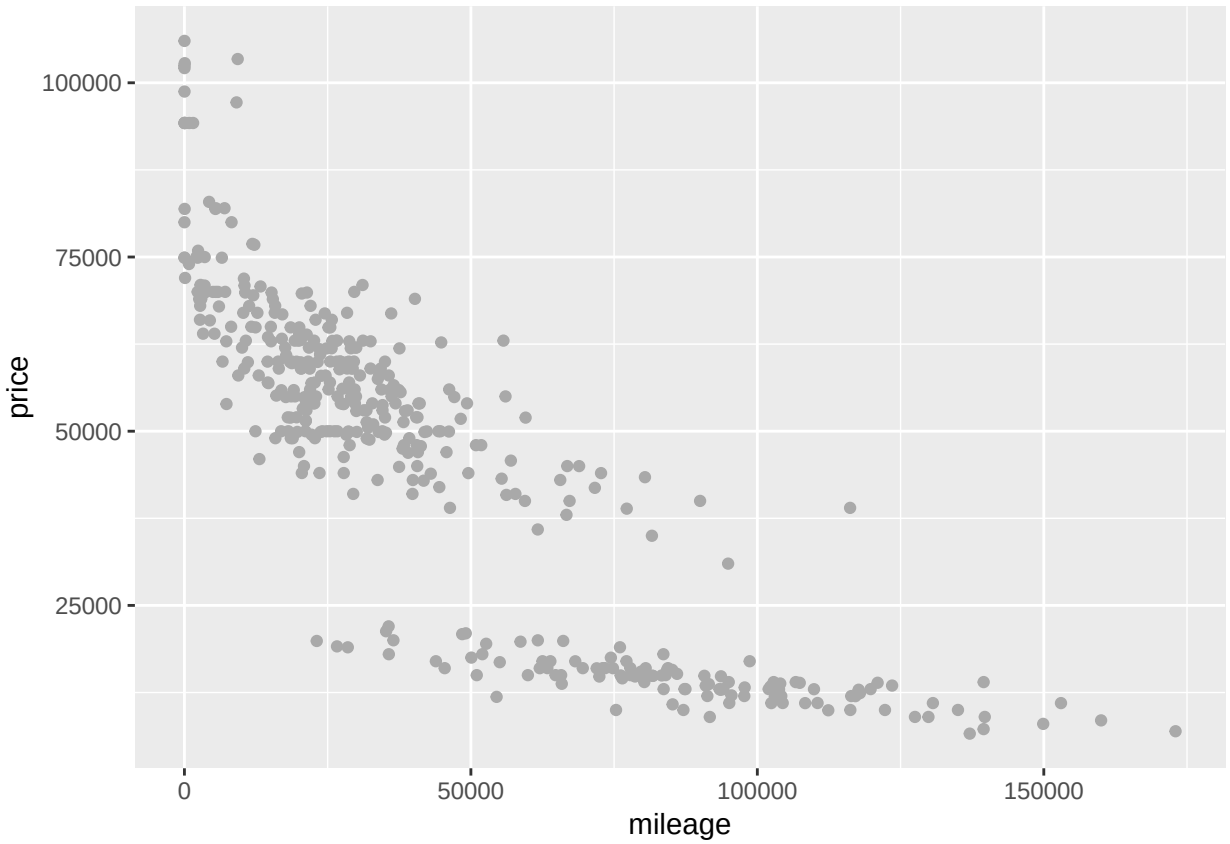
In this question, we only treat two trims: 350 and 65AMG as two separate data sets. We use KNN method to build a fitted model for price, given mileage.

What we need to do:

Under two trims, by using KNN method, we should find the optimal value of K which has the minimum out-of-sample RMSE. Then, for the optimal value of K, show the fitted model.

Firstly, let's focus on sclass350 trim.

```
# plot the data of sclass350
ggplot(data = sclass350) +
  geom_point(mapping = aes(x = mileage, y = price), color='darkgrey')
```



Then, split the data into a training and a testing set.

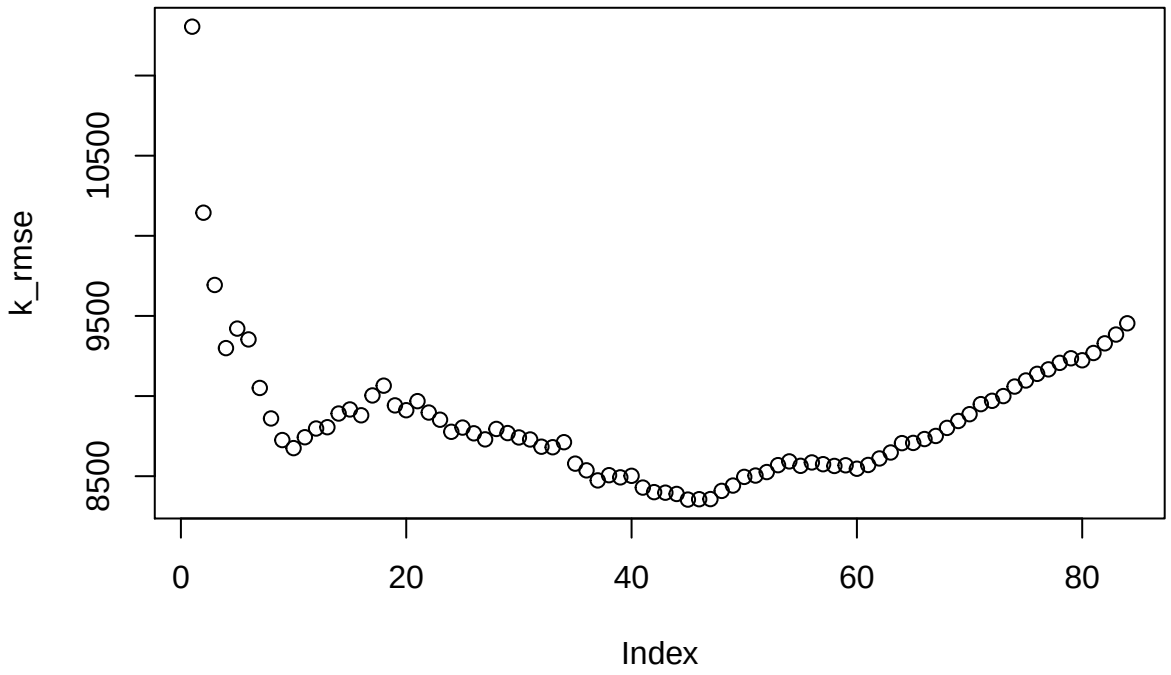
```
# Make a train-test split of sclass350
N = nrow(sclass350)
N_train = floor(0.8*N)
N_test = N - N_train
```

Now, let's run KNN for many different values of K, starting k=2 and going as high as we need to.

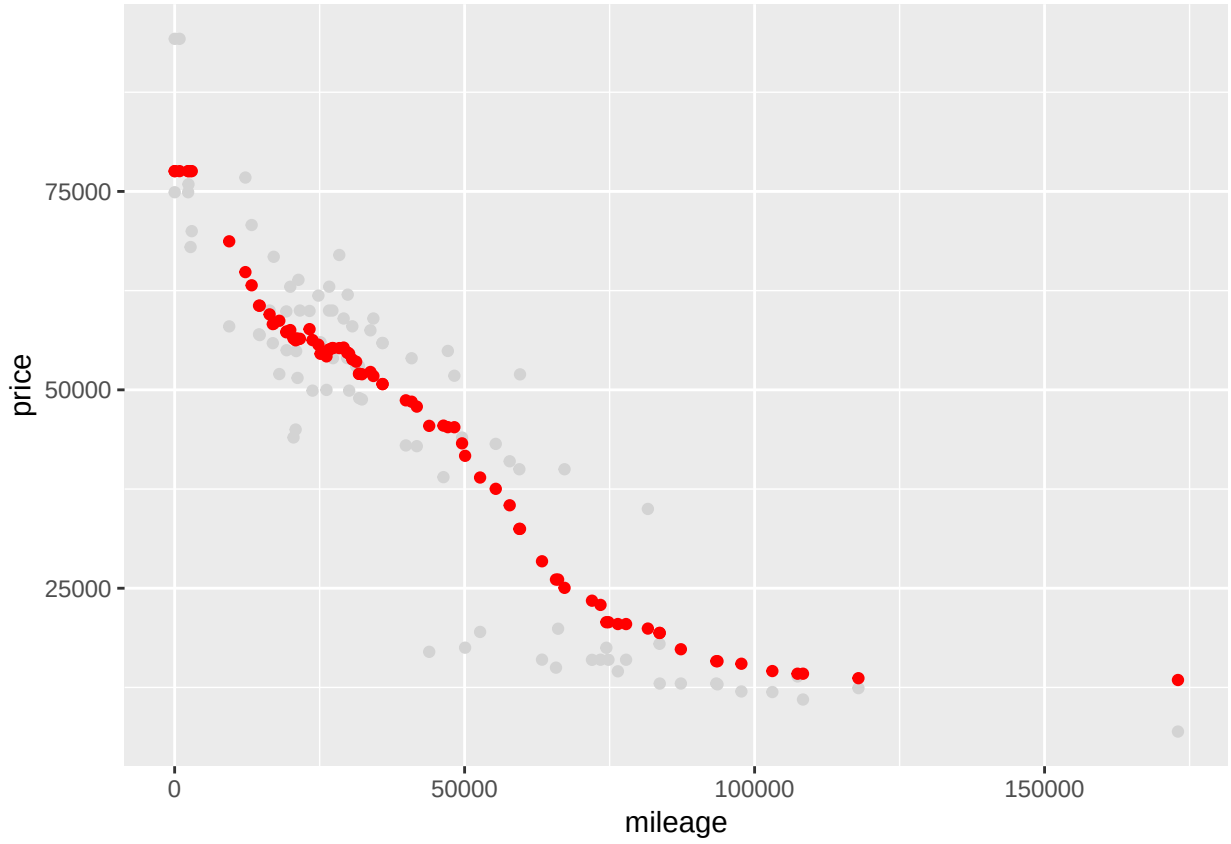
It seems that K=2 cannot run. Let's try K=3.

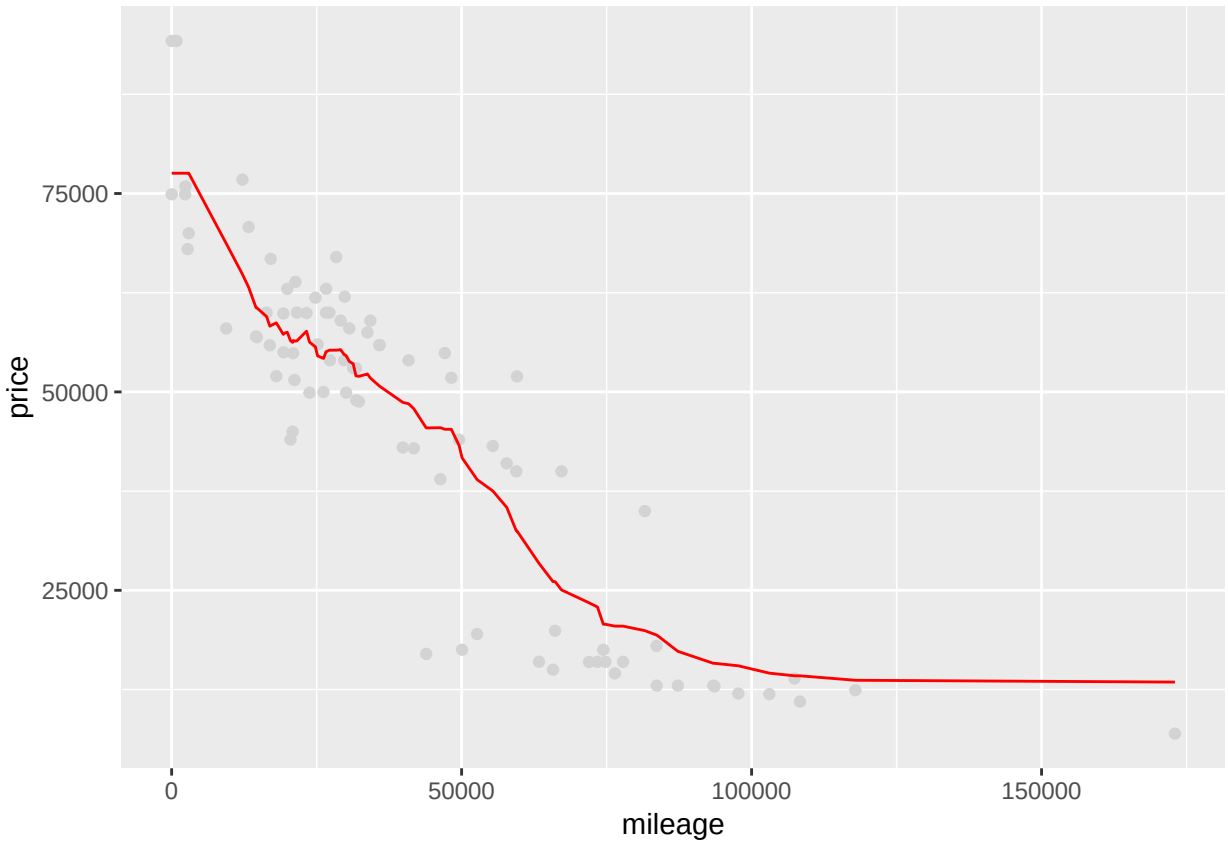
Then, try K = 4.

Now we should run a for loop with K=3,4,5... and find the optimal K.



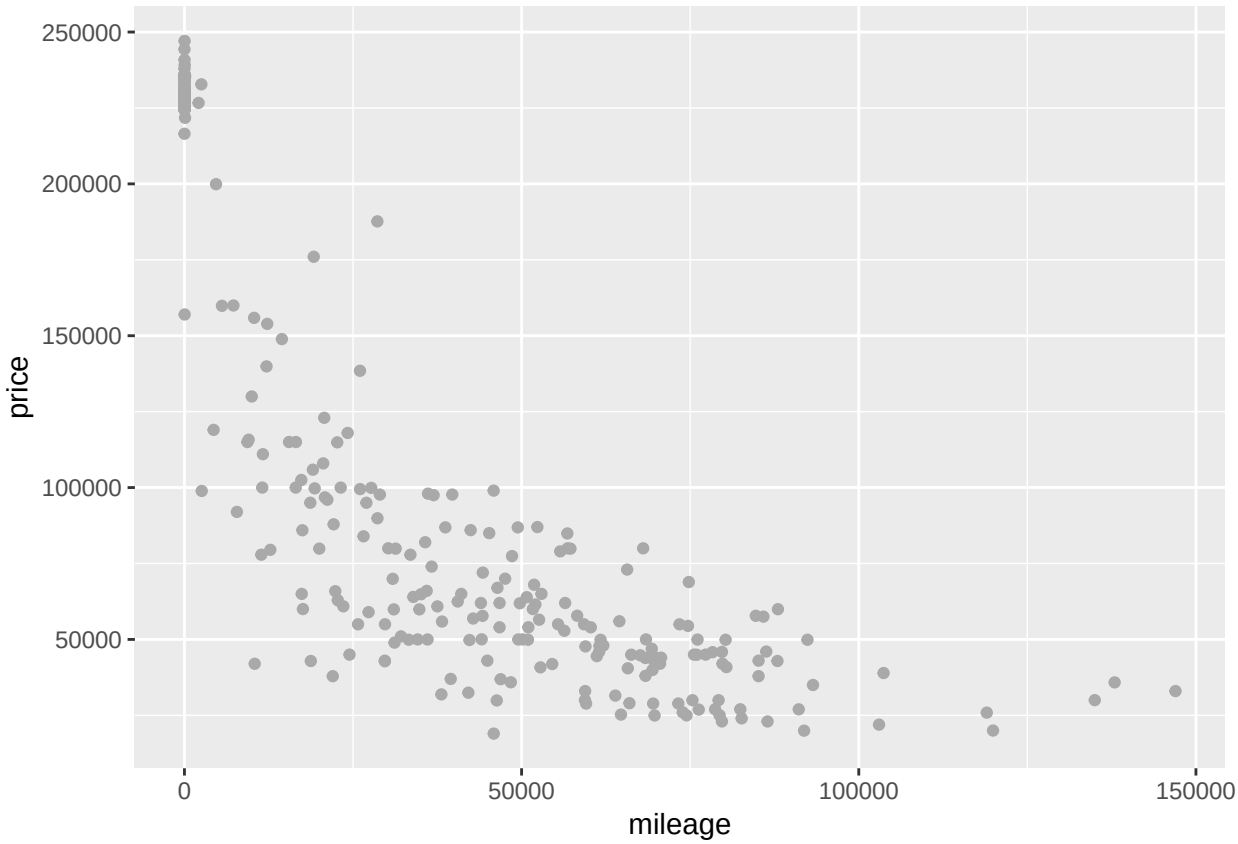
Using the optimal value of K, plot the fitted model as follows.





Secondly, let's focus on 65AMG trim.

```
# plot the data of sclass65AMG
ggplot(data = sclass65AMG) +
  geom_point(mapping = aes(x = mileage, y = price), color='darkgrey')
```



Then, split the data into a training and a testing set.

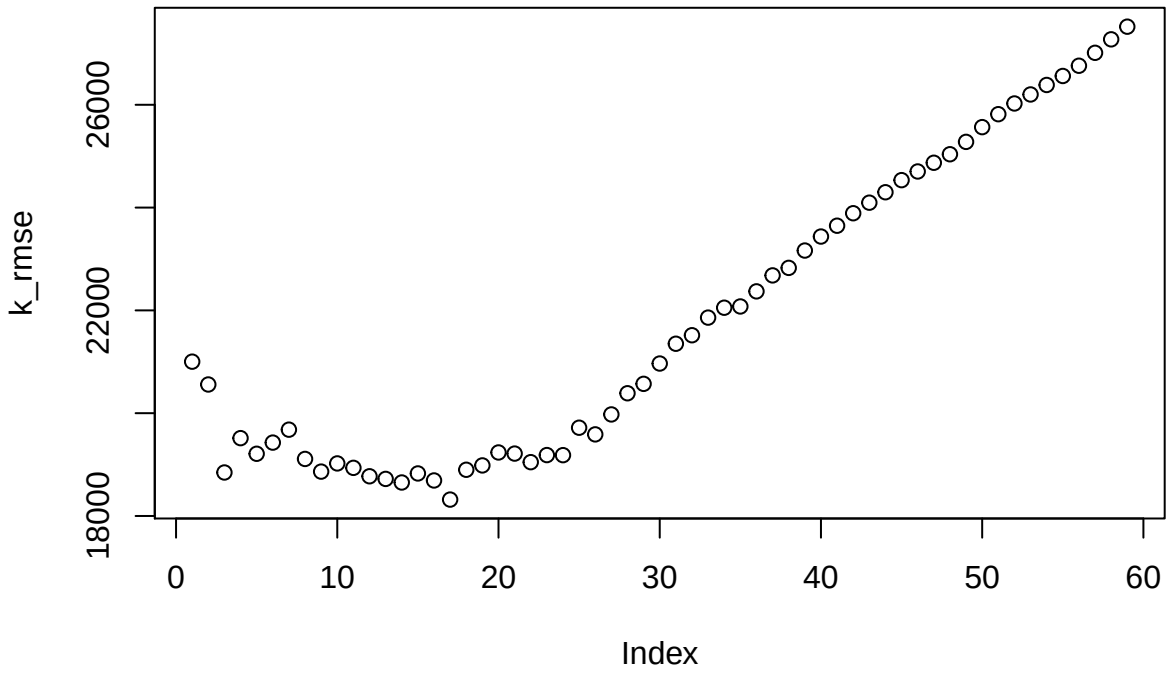
```
# Make a train-test split of sclass65AMG
N = nrow(sclass65AMG)
N_train = floor(0.8*N)
N_test = N - N_train
```

Now, let's run KNN for many different values of K, starting k=2 and going as high as we need to.

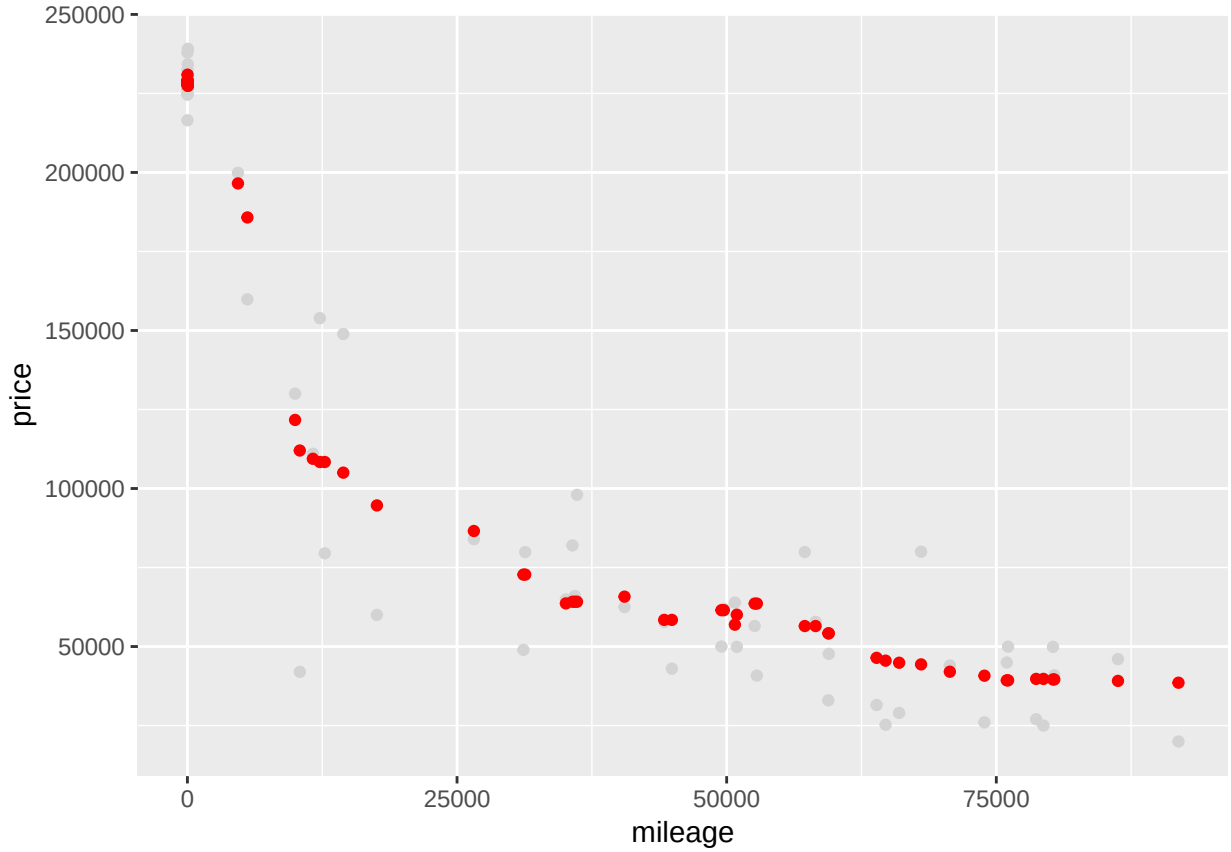
It seems that K=2 cannot run. Let's try K=3.

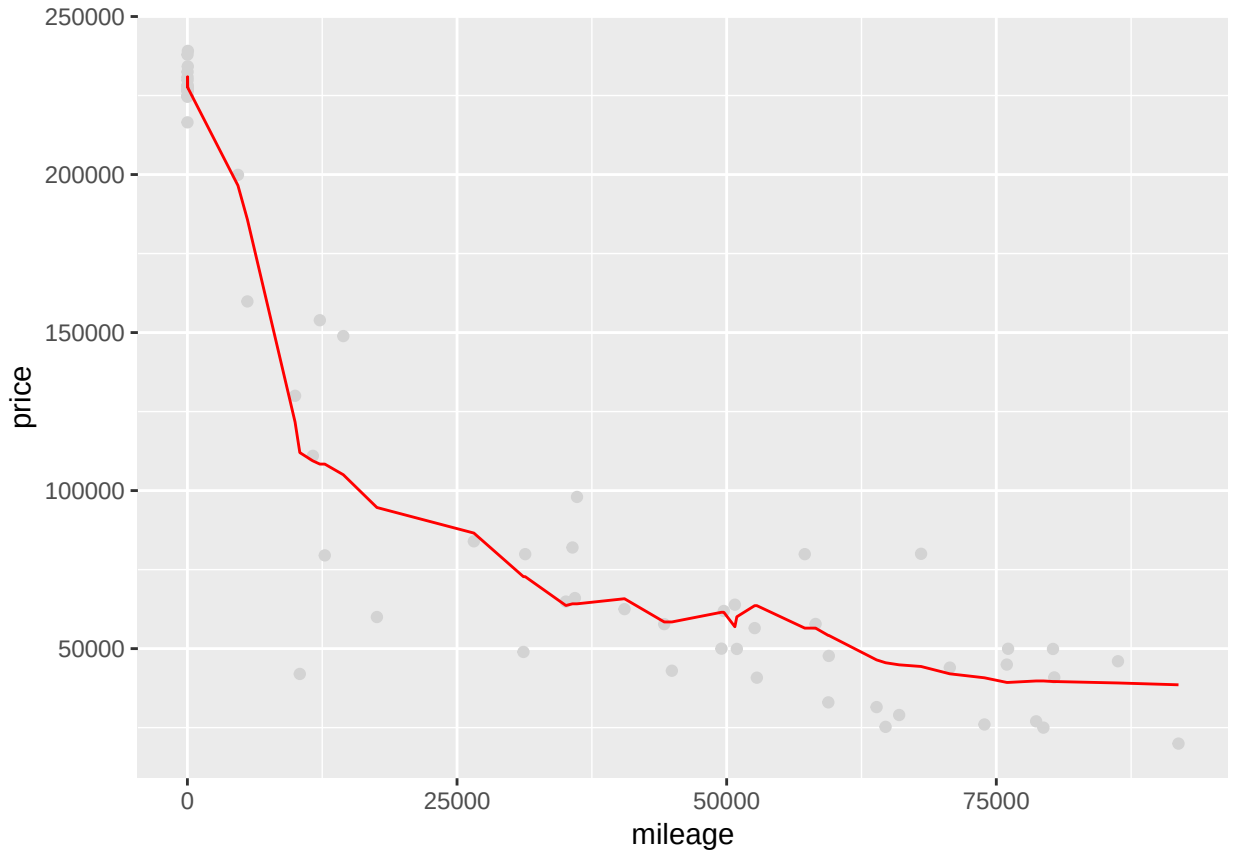
Then, try K = 4.

Now we should run a for loop with K=3,4,5... and find the optimal K.



Using the optimal value of K, plot the fitted model as follows.





Conclusion:

Since we randomly choose the train and test data sets, optimal K will not be the same for each time. However, no matter how many times we generate for each optimal K, optimal K of Sclass350 is always larger than optimal K of Sclass65AMG. I suppose this is because Sclass350 has larger sample size, which leads to a larger optimal K.