HW1 Dataming

Data Visualization 1: Green Buildings

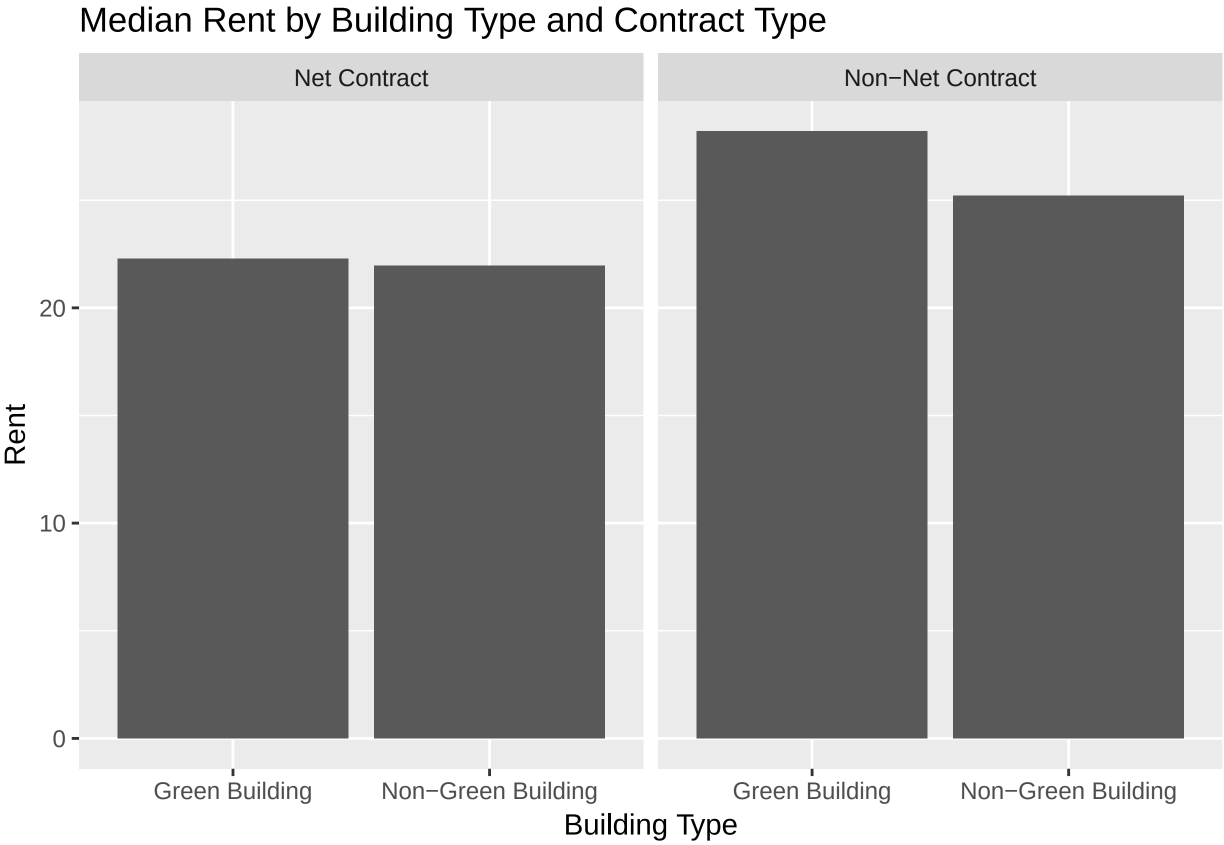
The “data-guru” did not consider other factors that could be causing the discrepancy in rents between green buildings and non-green buildings. First, the rents between net contract and non net contract rents should be analyzed because non-net contract rents include the price of utilities. For all of the below analysis buildings with lease rates 10 percent and below were excluded from the data.

The plot below shows the median rents for green buildings and non-green buildings for each contract type.

Plot

You can also embed plots, for example:

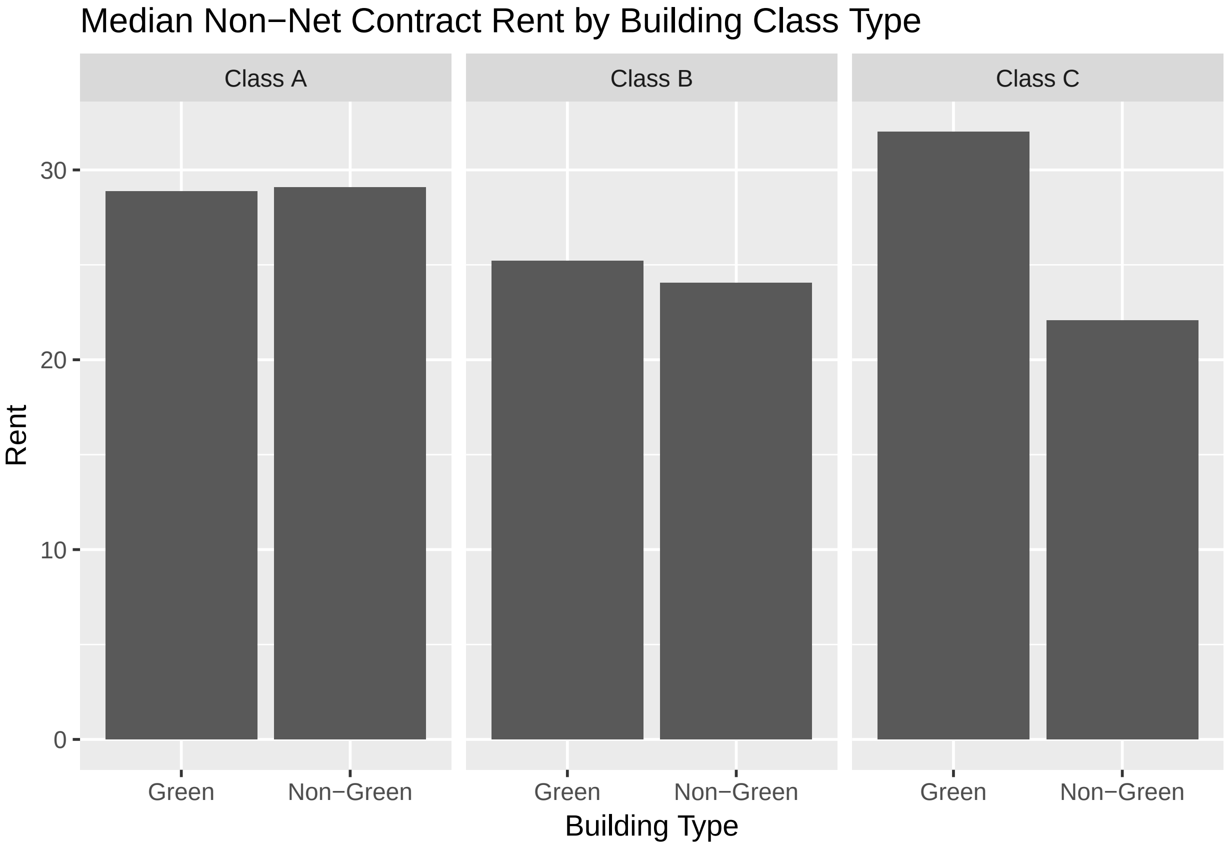
```
ggplot(df, aes(x=green, y=rent)) +
  geom_bar(stat='identity') +
  labs(
    title = "Median Rent by Building Type and Contract Type",
    x = "Building Type",
    y = "Rent") +
  facet_wrap(~ net)
```



As shown above, green buildings have higher rents compared to non-green buildings for both non-net contracts and net contracts.

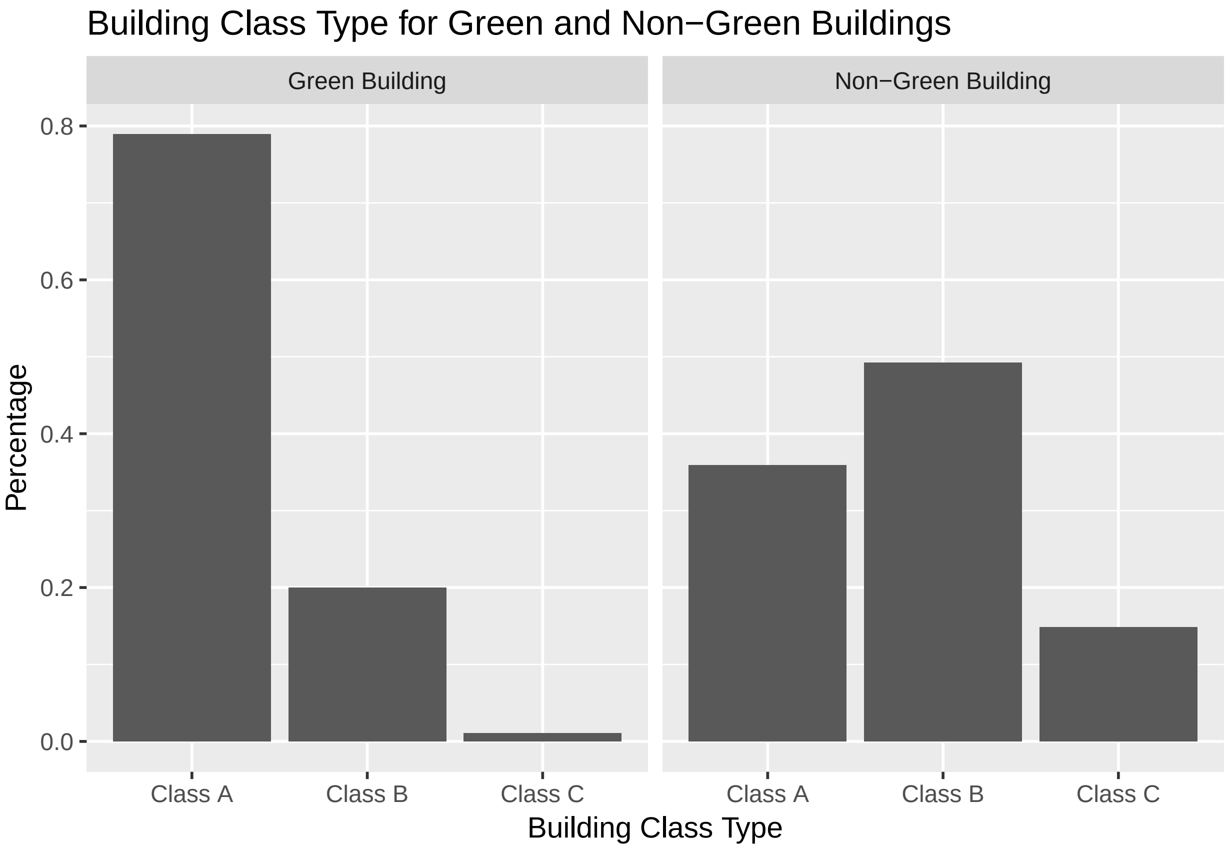
However, rents could be influenced by other factors such as building class type. The plots below show the median non-net contract rents by building class type for green buildings and non-green buildings.

```
ggplot(df_class, aes(x=greentype, y=rent)) +
  geom_bar(stat='identity') +
  labs(
    title = "Median Non-Net Contract Rent by Building Class Type",
    x = "Building Type",
    y = "Rent") +
  facet_wrap(~ typelabel)
```



As shown above, for class A and B buildings there is little difference between the rents for green buildings and non-green buildings. However, class C buildings have much higher rents. As a result, the population mix of class types for green buildings and non-green buildings should be analyzed.

```
ggplot(df_class_perc, aes(x=typelabel, y=classpercent)) +
  geom_bar(stat='identity') +
  labs(
    title = "Building Class Type for Green and Non-Green Buildings",
    x = "Building Class Type",
    y = "Percentage") +
  facet_wrap(~ greentype)
```

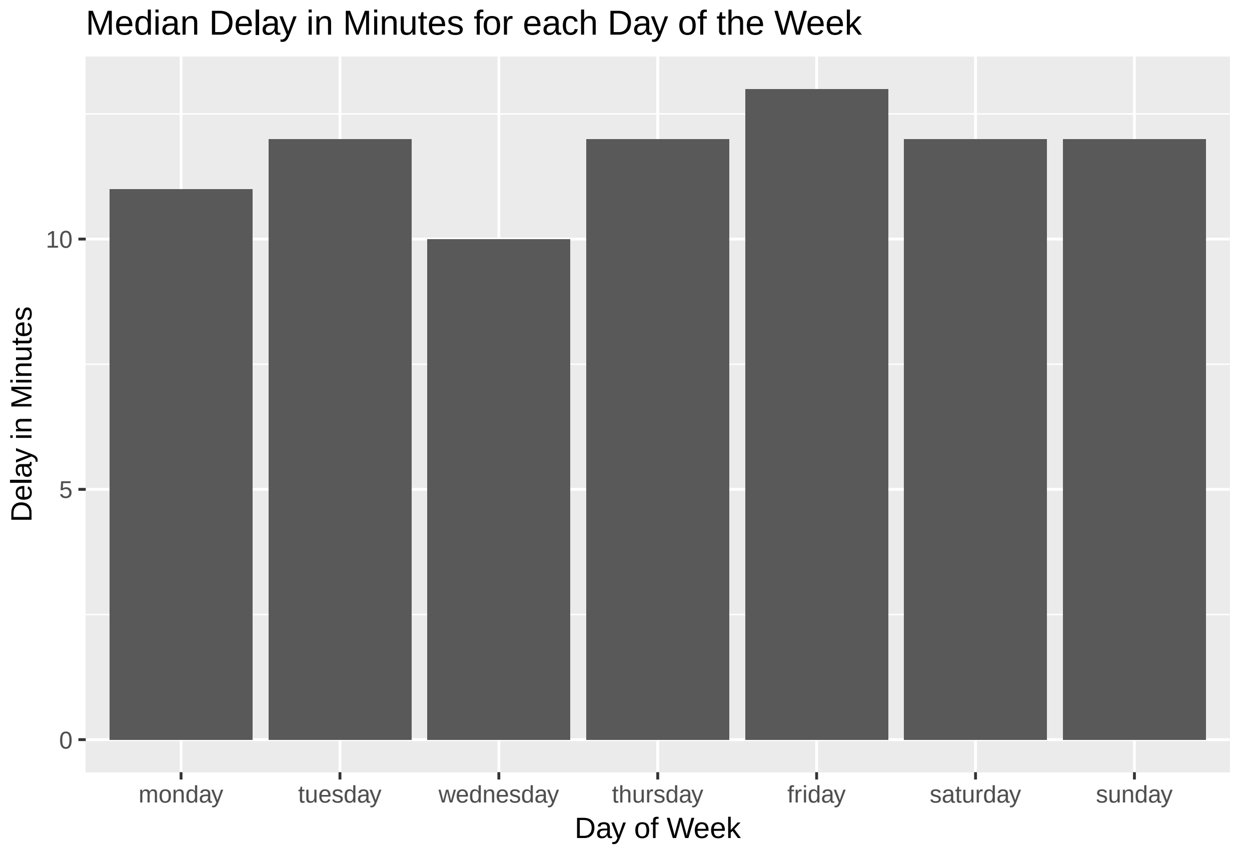


The above shows that a majority of green buildings are considered class A buildings. However, only around 36% of non-green buildings are class A. Class A buildings on average have higher rents which skews the median rents of green buildings higher compared to non-green buildings. As a result, if a class A green building is built it is not economically viable to build compared to a non-green building.

Data Visualization 2: ABIA

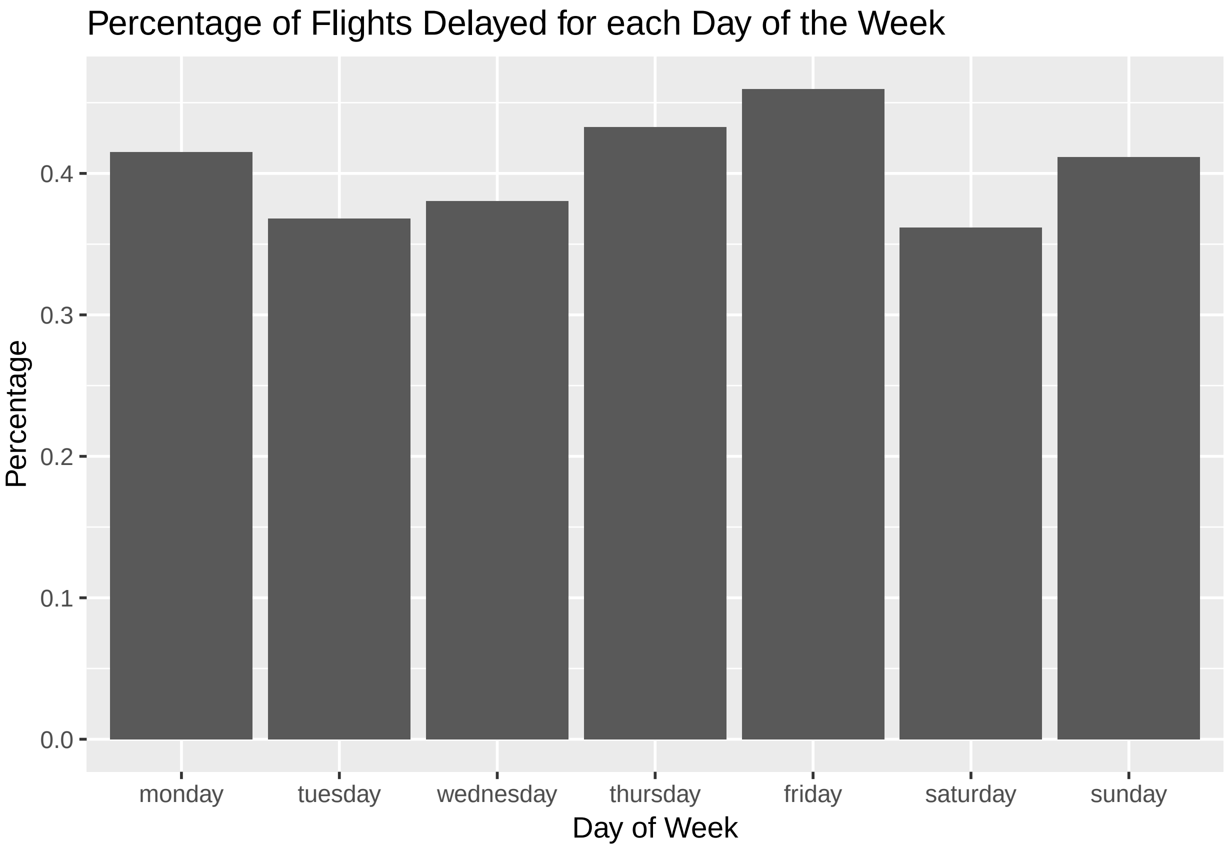
I analyzed which day of the week is the best day to minimize delays. The first plot below show the median delay for each flight on each day of the week.

```r
ggplot(df_day, aes(x=reorder(dayname, daynum), y=depdelay_med)) +
  geom_bar(stat='identity') +
  labs(
    title = "Median Delay in Minutes for each Day of the Week",
    x = "Day of Week",
    y = "Delay in Minutes")
```



Wednesday has the lowest median delay out of all days of the week. However, to minimize delays the percentage of flights delayed should also be considered.

```r
ggplot(df_day_perc, aes(x=reorder(dayname, daynum), y=depdelay_perc)) +
  geom_bar(stat='identity') +
  labs(
    title = "Percentage of Flights Delayed for each Day of the Week",
    x = "Day of Week",
    y = "Percentage")
```



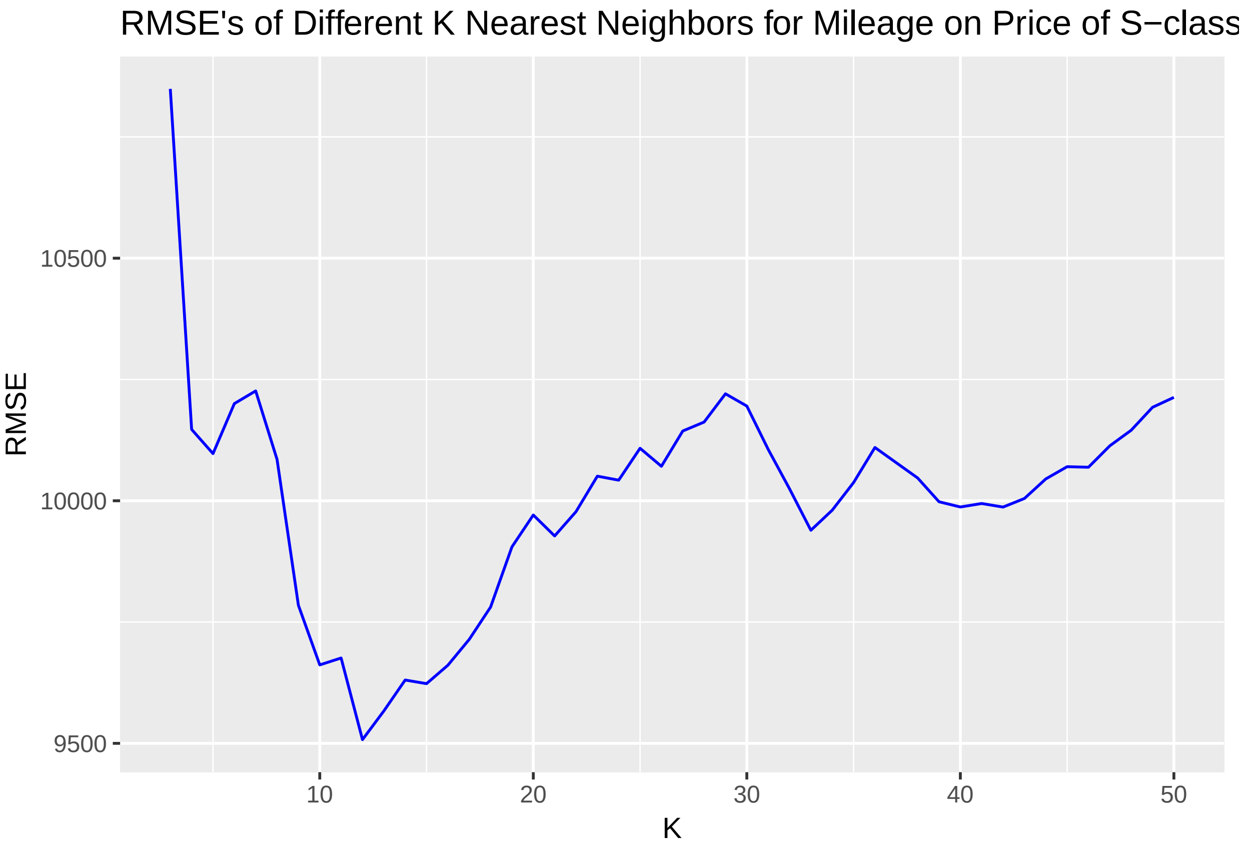
Saturdays have the lowest percentage of flights delayed at 36%. Flights on Wednesdays are delayed slightly more at 38%. As a result, Wednesdays are generally the best day of the week to fly in order to minimize delays.

Regression vs. KNN

For each trim type k values from 3 to 50 were tested in order to minimize out of sample RMSE. The data subsets for each trim type was split for training and testing at 70% and 30%, respectively.

The plot below shows the out of sample RMSE's of K Nearest Neighbors from 3 to 50 for predicting the price of an S-class model with a 350 trim based on mileage.

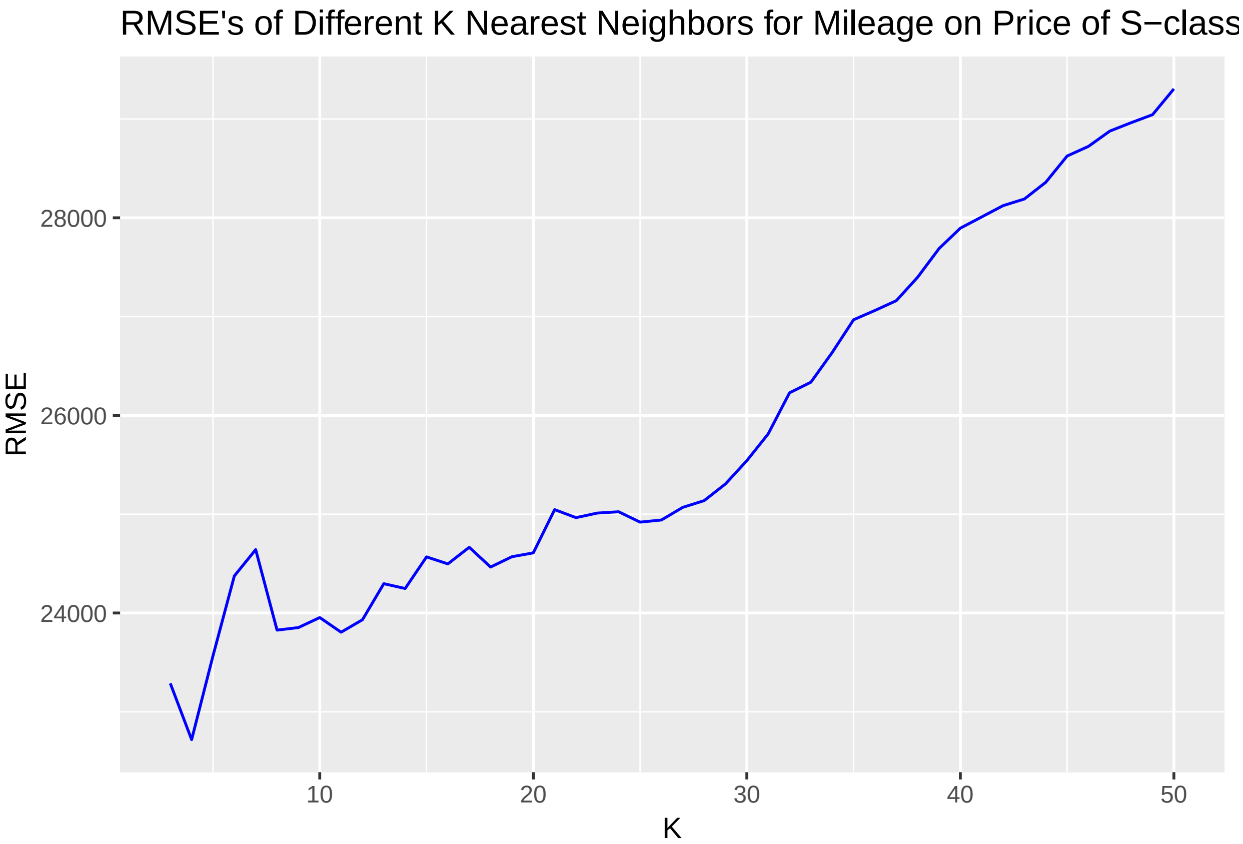
```
ggplot(data = df_350) +
  geom_path(aes(x = x_350, y = y_350), color='blue') +
  labs(
  title = "RMSE's of Different K Nearest Neighbors for Mileage on Price of S-class Mercedes with a 350
  x = "K",
  y = "RMSE"
)
```



The out of sample RMSE's bottom when k is around 20.

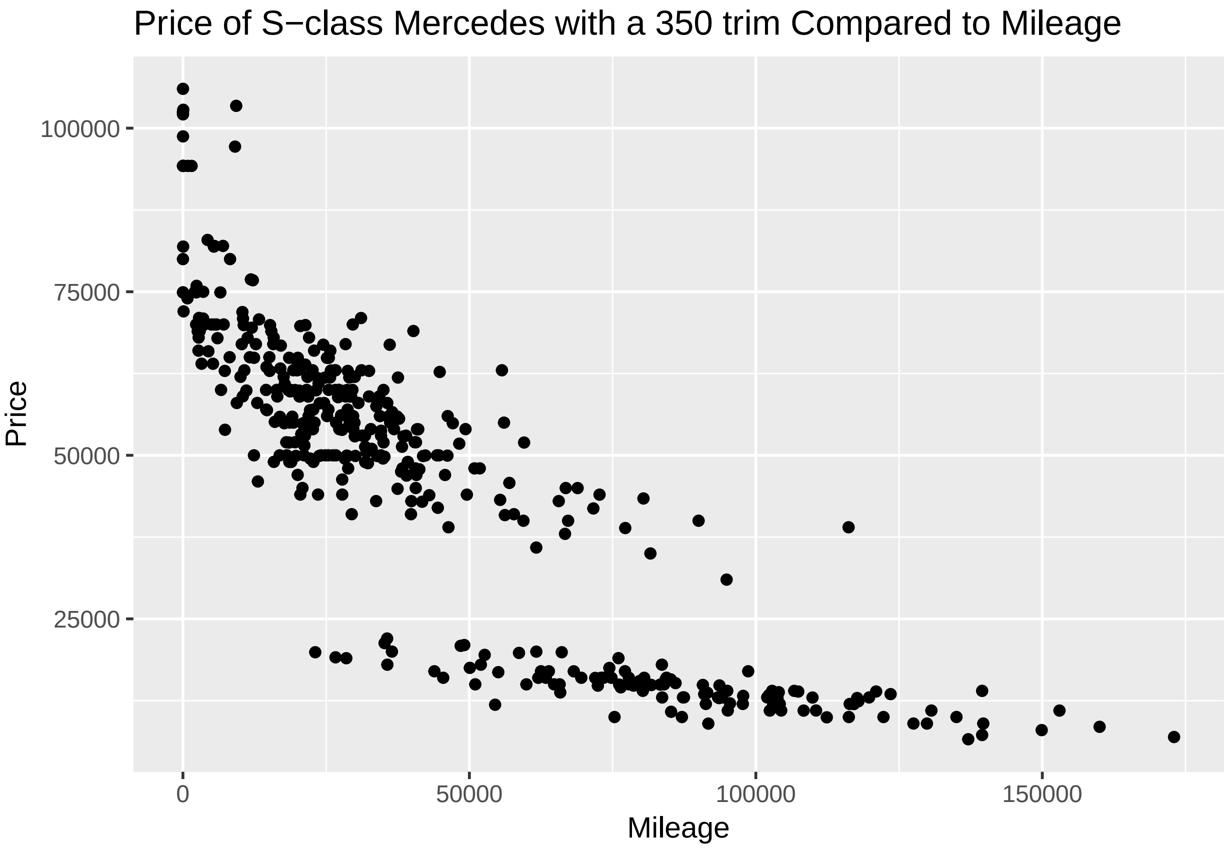
The plot below shows the out of sample RMSE's of K Nearest Neighbors from 3 to 50 for predicting the price of an S-class model with a 65 AMG trim based on mileage.

```
ggplot(data = df_65amg) +
  geom_path(aes(x = x_65amg, y = y_65amg), color='blue') +
  labs(
    title = "RMSE's of Different K Nearest Neighbors for Mileage on Price of S-class Mercedes with a 65
    x = "K",
    y = "RMSE"
  )
```

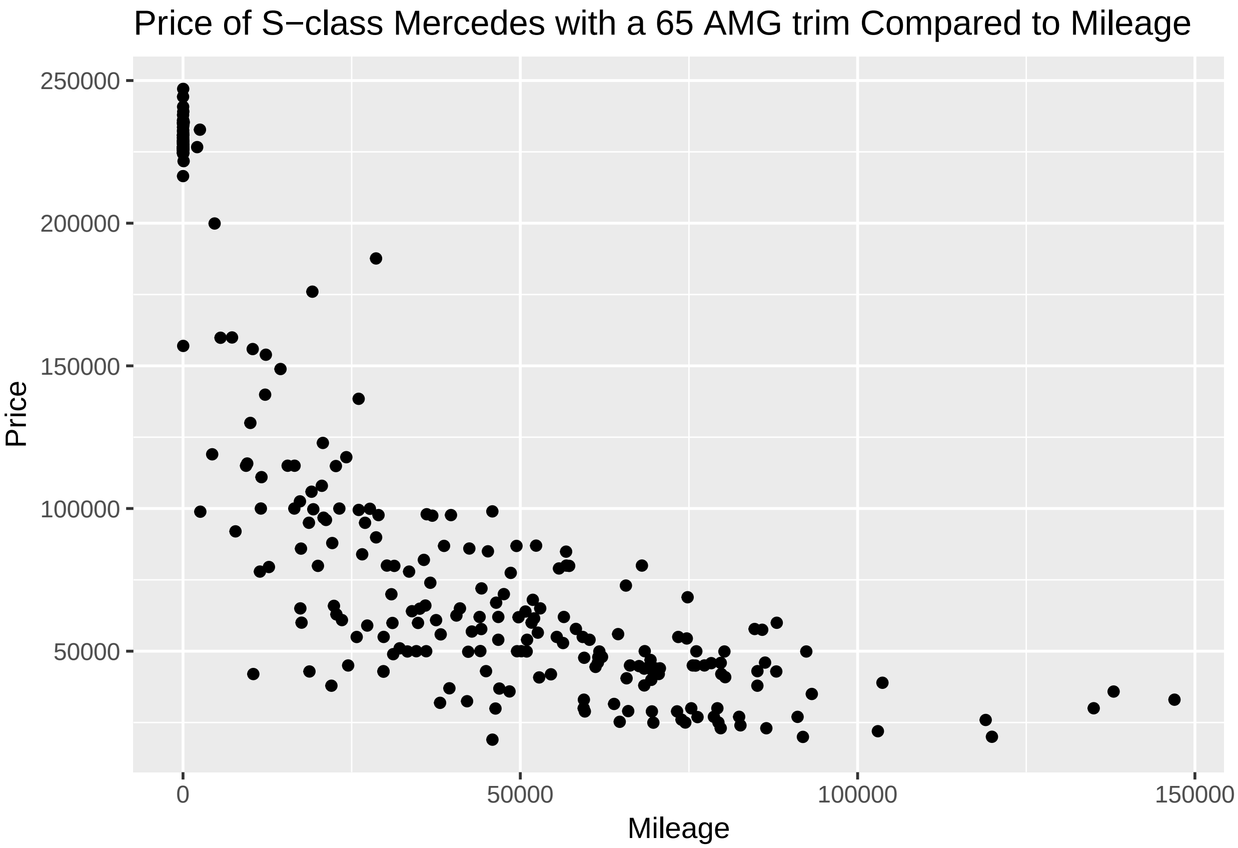


The optimal K value is generally larger for the 65 AMG trim compared to the 350 trim. This is because there is greater variation in price in the 65 AMG trim compared to the 350 trim as shown below.

```
ggplot(data = sclass_350) +
  geom_point(mapping = aes(x = mileage, y = price)) +
labs(
    title = "Price of S-class Mercedes with a 350 trim Compared to Mileage",
    x = "Mileage",
    y = "Price")
```



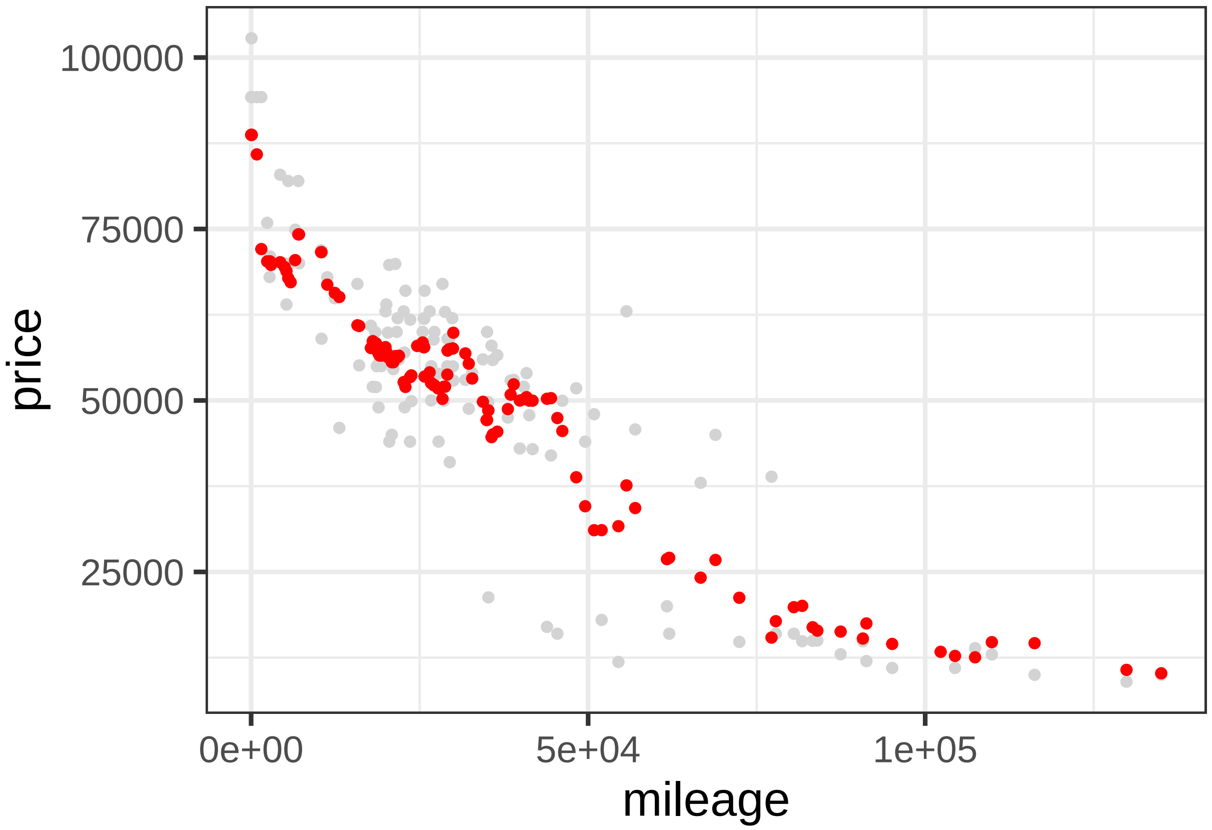
```r
ggplot(data = sclass_65amg) +
  geom_point(mapping = aes(x = mileage, y = price)) +
labs(
    title = "Price of S-class Mercedes with a 65 AMG trim Compared to Mileage",
    x = "Mileage",
    y = "Price")
```



The optimal K value for 350 trim is 10

The fitted plot for the 350 trim with K=10 is below.

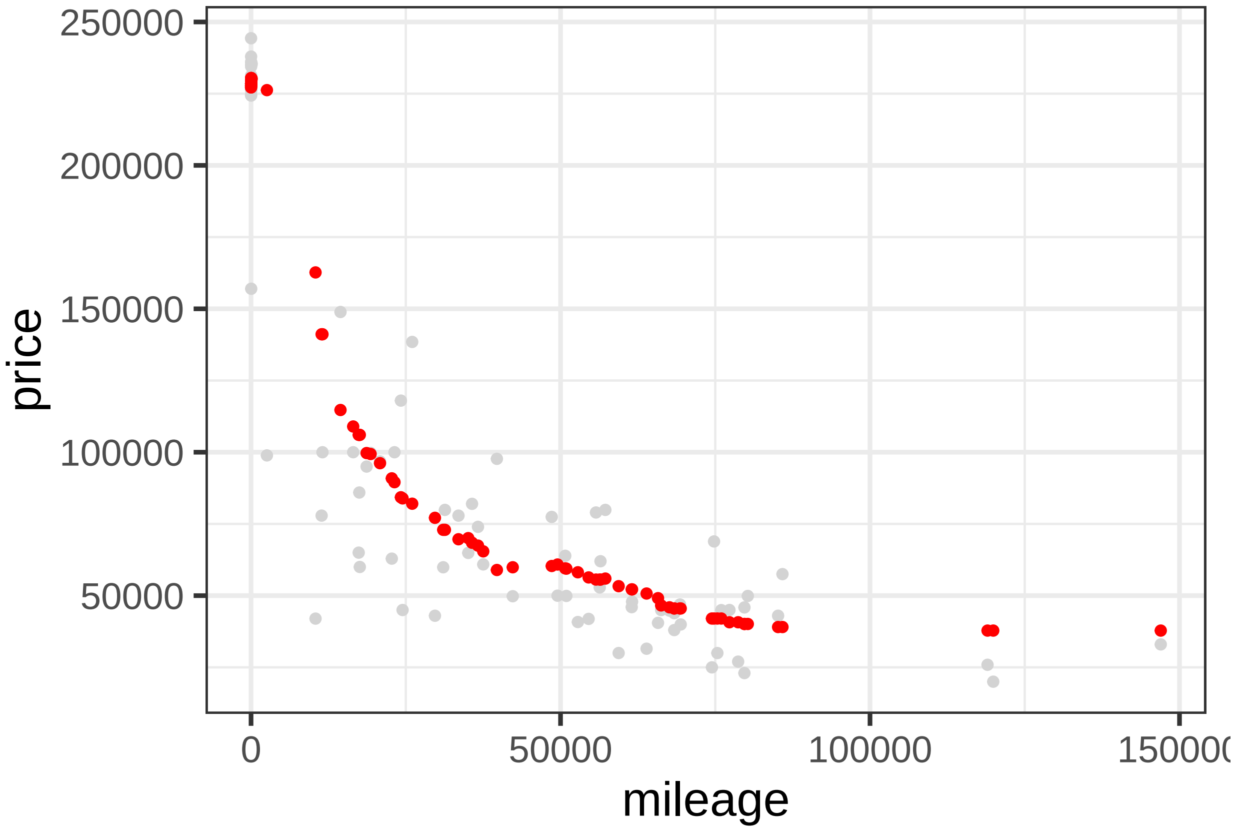
```
p_test + geom_point(aes(x = mileage, y = ypred_350), color='red')
```



The optimal K value for 65 AMG trim is 35

The fitted plot for the 65 AMG trim with K=35 is below.

```
p_test + geom_point(aes(x = mileage, y = ypred_65amg), color='red')
```



price
250000
200000
150000
100000
50000
0
50000
100000
150000
mileage