ECO 395 Homework 1:

Soo Jee Choi, Annie Nguyen, and Tarini Sudhakar

2023-01-28

1) Data Visualization: Flights at ABIA

Introduction

In this section, we are looking at a 2008 dataset that contains information on every commercial flight coming in and out of the Austin-Bergstrom Internationial Airport. We are interested in seeing which cities in the dataset have the highest average departure delays.

Methods

To explore the data and create visualizations, we first need to import the `tidyverse` and `ggplot2` packages and read the `ABIA.csv` file which contains the data. Since we are also interested in visualizing the data on a geographical map, we will also import the `ggmap` package and read the `airport-codes` files, which contain the latitude and longitude reading of each airport.

First, the `airport_codes` dataframe is cleaned to exclude closed airports and drop rows with NA's or no available IATA-codes. Then, the `airport_codes` dataframe is joined with `abia` dataframe, so each US airport in our targeted dataframe has a respective value for longitude and latitude.

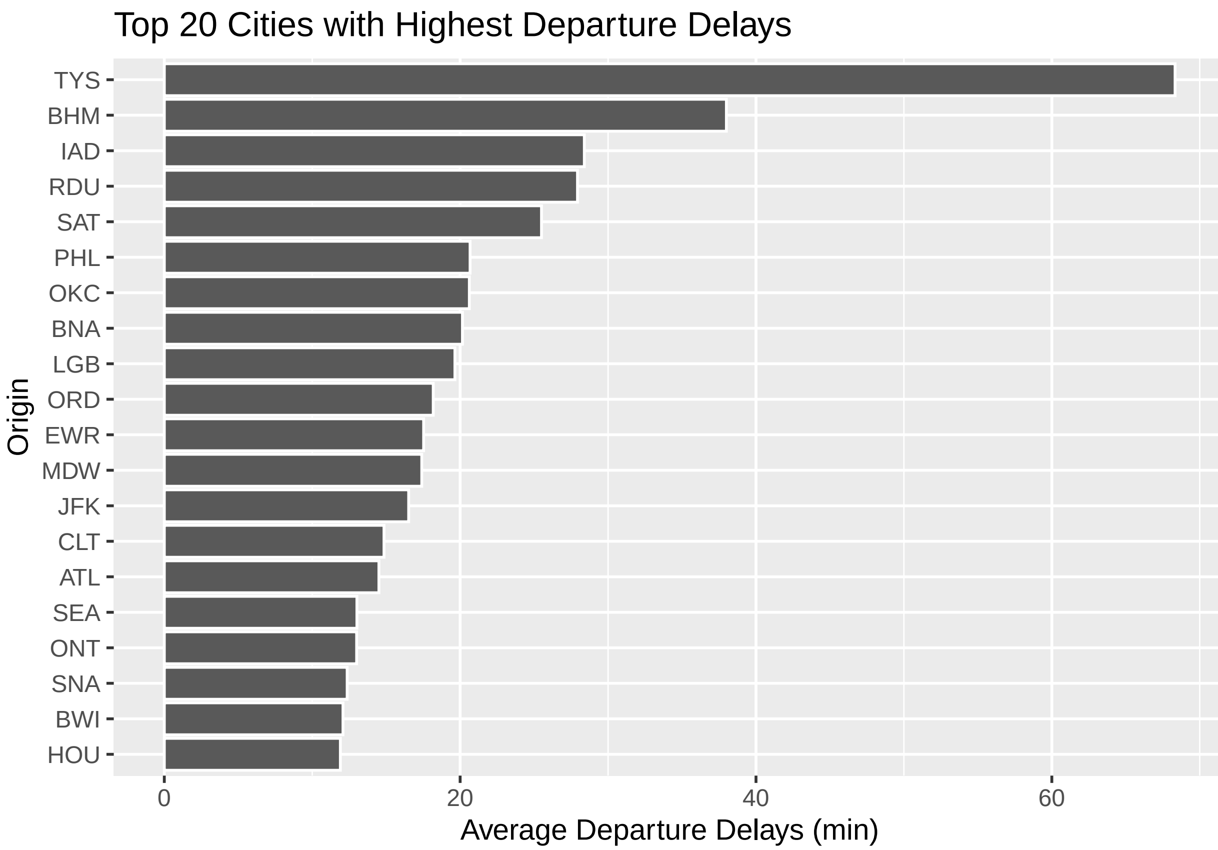
Then, we calculate the average departure delay from each Origin airport. Airports with the highest departure delays include: Knoxville, TN; Birmingham, AL; Washington, D.C.; Raleigh/Durham, NC; San Antonio, TX; Philadelphia, PA.

Table 1: Top 6 Cities with Highest Departure Delays

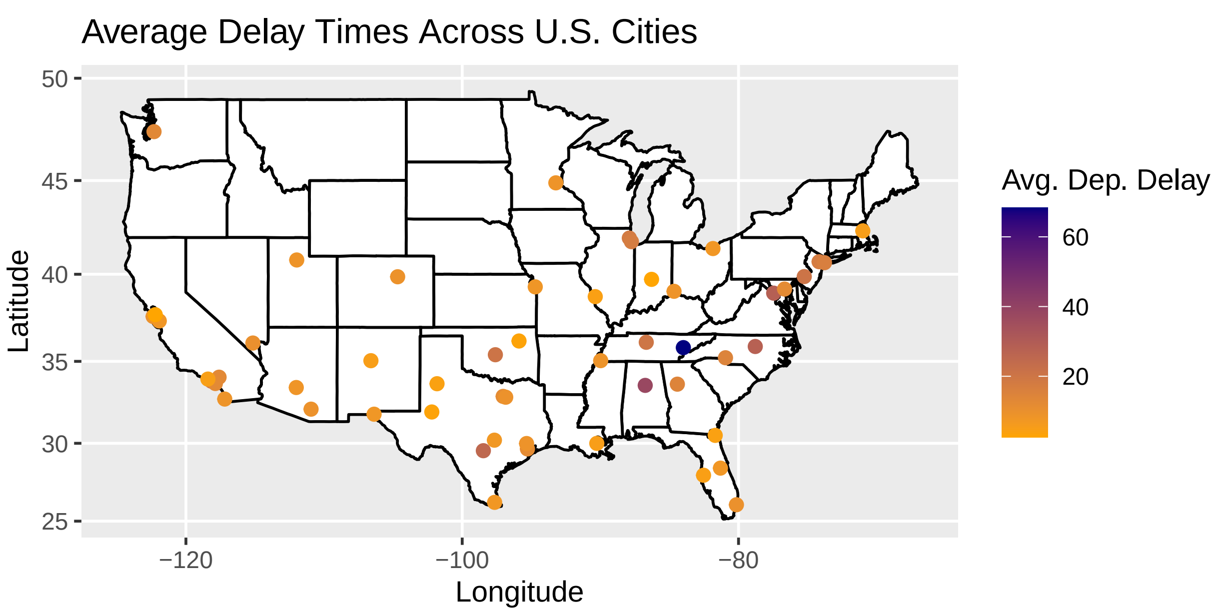
Origin	avg_delay_time
TYS	68.33333
BHM	38.00000
IAD	28.39550
RDU	27.93450
SAT	25.50000
PHL	20.67241

We will use a barplot to visualize a larger portion of the dataset, looking at the top 20 cities with the highest departure delays.

From the plot, we see that Knoxville has, on average, significantly higher delays than Birmingham (with departure delays close to 70 minutes for Knoxville and near 40 minutes for Birmingham). In comparison, the average delays for the remaining cities stay below 30 minutes.



We will visualize the data for all 53 cities included on the dataset by plotting the points on the US map. This will not only show us average airport delays but also give us a brief look at all of the US cities that flyers can travel to from Austin-Bergstrom Internationial Airport.



Results and Conclusion

Based on the results of the plots, we see that Knoxville and Birmingham experience significantly higher departure delays compared to other cities.

From the data visualization generated in this section, we were able to identify airports that experienced the highest average departure delays. This information can help future flyers know what to expect in terms of departure delays when traveling to Austin.

Of course, this brief data exploration and visualization report comes with its own set of limitations. Departure delays are influenced by holiday rushes and weather conditions, so outliers from unusual weather or from other potential problems may skew results. Furthermore, the dataset only includes observations from 2008 and delays may affect each airport differently on a yearly basis.

2) Wrangling the Olympics

A) What is the 95th percentile of heights for female competitors across all Athletics events (i.e., track and field)? Note that "sport" is the broad sport (e.g. Athletics) whereas "event" is the specific event (e.g. 100 meter sprint).

The 95th percentile of heights for female competitors across all Athletics events was obtained by first subsetting the values in the 'olympics_top20.csv' file to observations with the sex variable equal to "F" and the sport variable equal to "Athletic". Then, the quantile function was used to get the 95th quantile of the height of female track and field ("Athletic") Olympians, which was 183cm.

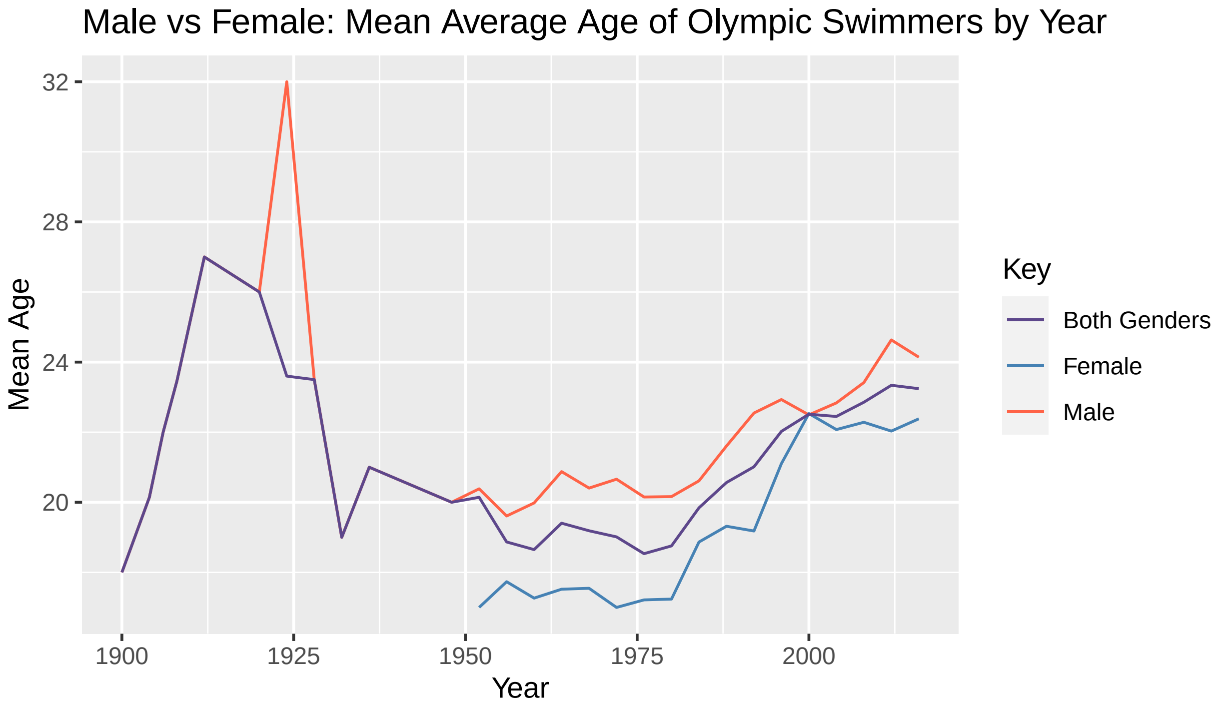
B) Which single women's "event" had the greatest variability in competitor's heights across the entire history of the Olympics, as measured by the standard deviation?

Table 2: Top 6 Single Women's Olympic Events with Greatest Variability in Height

event	sd_height
Rowing Women's Coxed Fours	10.865490
Basketball Women's Basketball	9.700255
Rowing Women's Coxed Quadruple Sculls	9.246396
Rowing Women's Coxed Eights	8.741931
Swimming Women's 100 metres Butterfly	8.134398
Volleyball Women's Volleyball	8.101521

To find which single women's event had the greatest variability in competitor's heights (as measured by the standard deviation), we first subset the original data set to contain only females observations. Then the data was grouped by sporting event, and then the standard deviation of the heights for each sport was calculated. Arranging the results in descending order, we find that Rowing Women's Coxed Fours had the greatest variation in competitor's heights (as measured by the standard deviation) at 10.865490.

C) How has the average age of Olympic swimmers changed over time? Does the trend look different for male swimmers relative to female swimmers? Create a data frame that can allow you to visualize these trends over time, then plot the data with a line graph with separate lines for male and female competitors. Give the plot an informative caption answering the two questions just posed.



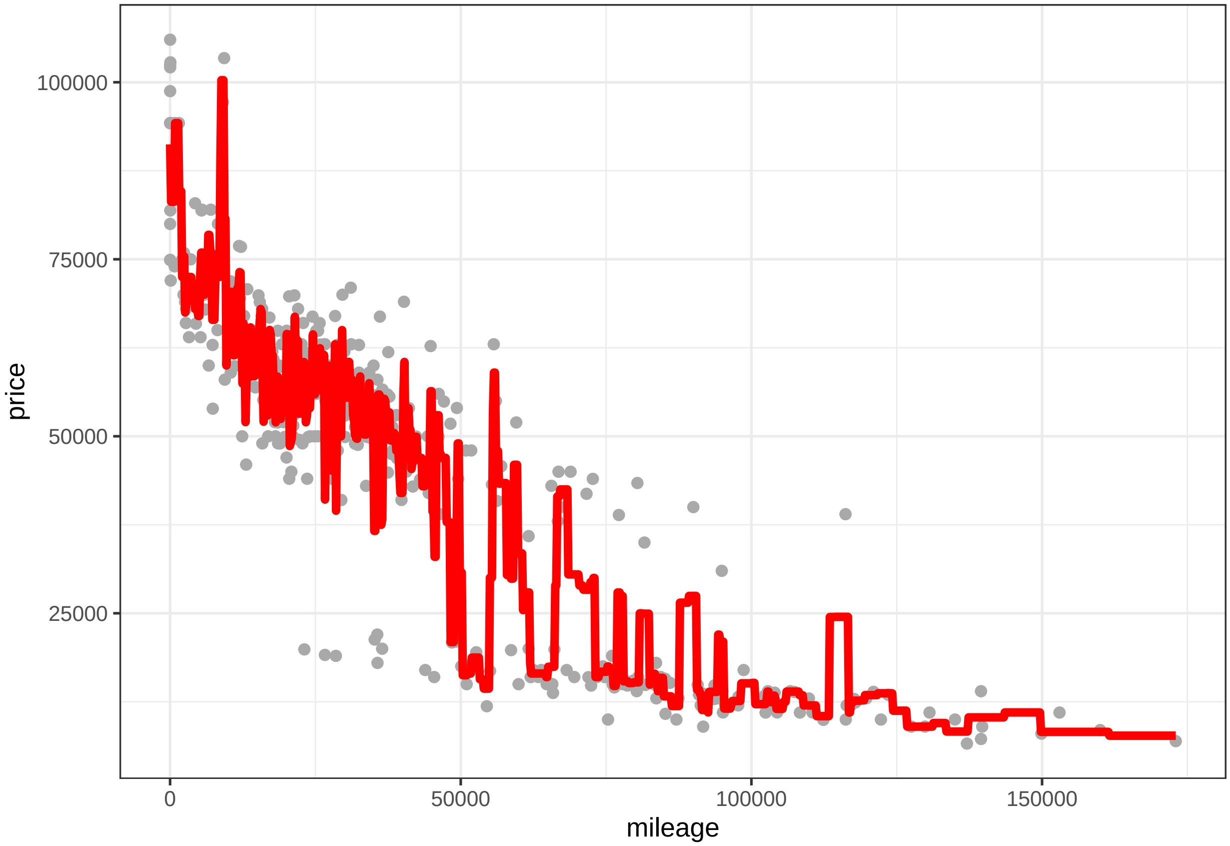
Average age of Olympic swimmers experiences a sharp rise and fall in the early 1900s, and follows an increasing trend thereafter. The average age of male Olympic swimmers generally trends higher than the average age of female Olympic swimmers. However, the average age of male and female Olympic Swimmers notably converge in 2000.

The dataset on average age of Olympic swimmers shows that women did not participate in Olympic swimming competitions from 1900 to 1948, with the exception of 1924. We see the average age of Olympic swimmers started at 18 years old in 1900 and reached a peak in 1912 at 27 years old. The data then follows a decreasing trend reaching the lowest average age of 18.53390 in 1976. Then the data follows an increasing trend in the years that follow, with 23.24211 years old in 2016 being the most recent average Olympic swimmers' age data value available.

We see, in the data available for both genders, that male Olympic swimmers, on average, have consistently been older than female Olympic swimmers. However, in 2000, female Olympic swimmer were, on average, 22.53191 years old while male Olympic swimmers were, on average, 22.49451 years old. Hence, we observe a convergence in the graph of the mean average age of Olympic swimmers in the year 2000. Additionally, the average age of male Olympic swimmers experienced fluctuations in the early 1900s (where data on female Olympic swimmers was not available), notably reaching a peak of 32 years old in 1924. Conversely, the data on female Olympic swimmer shows a general increase in average age over time.

3) K-Nearest Neighbors: Cars

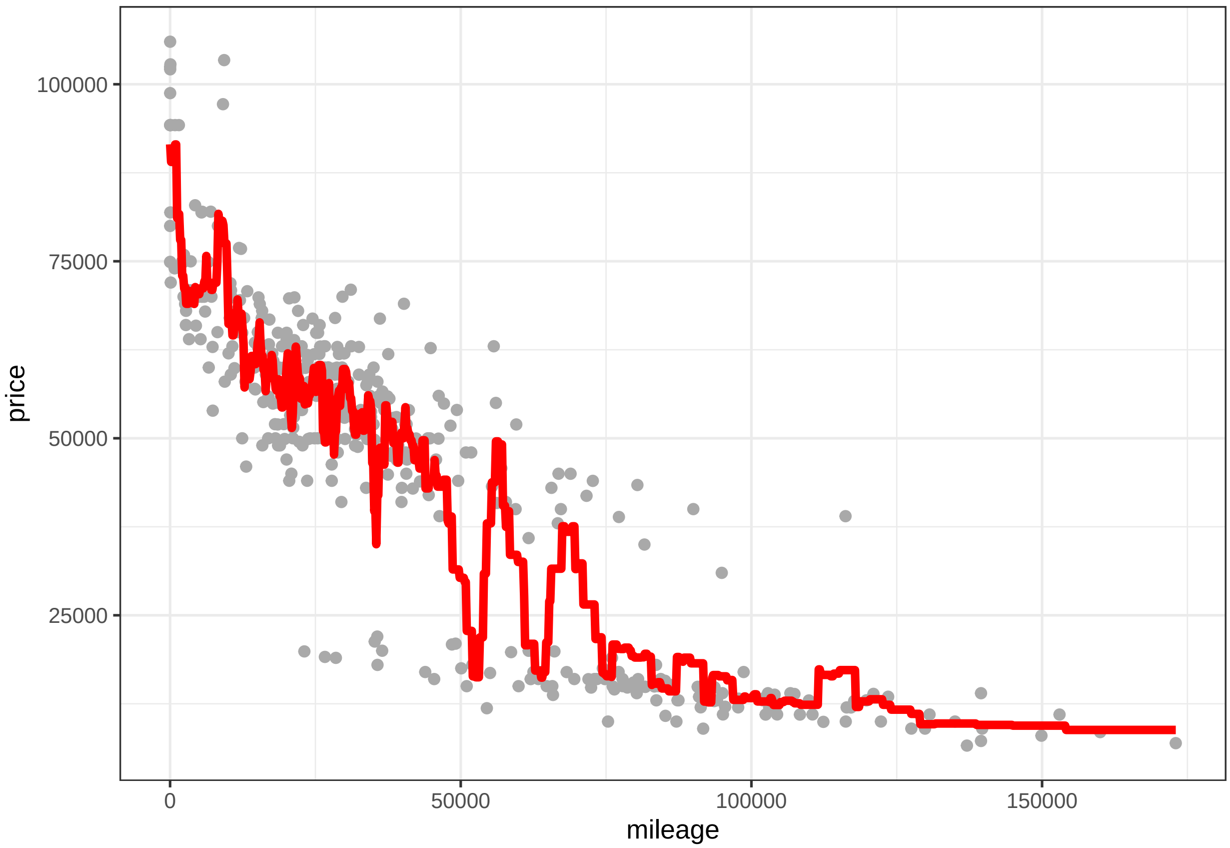
K=2



100000
75000
50000
25000
price
0
50000
100000
150000
mileage


```
## [1] 11397.94
```

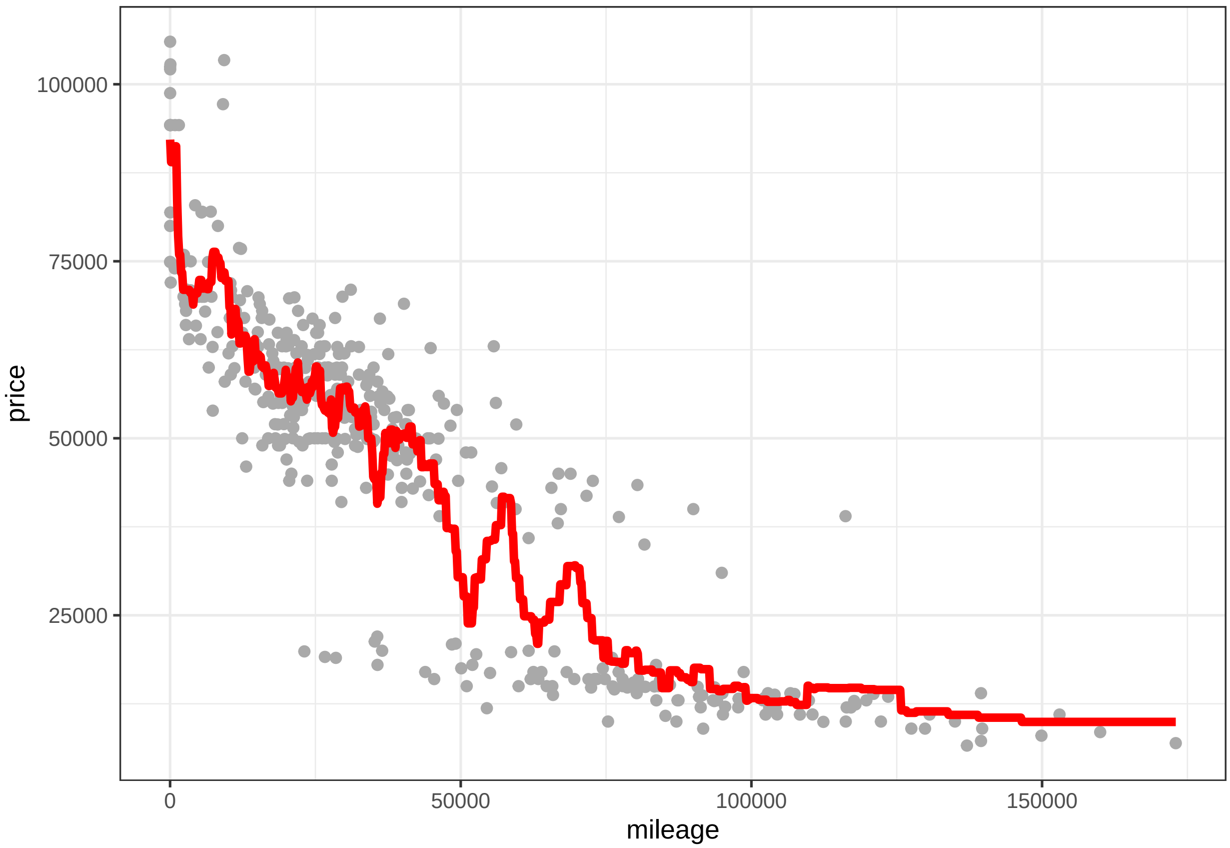
K=5



price
100000
75000
50000
25000
0
50000
100000
150000
mileage

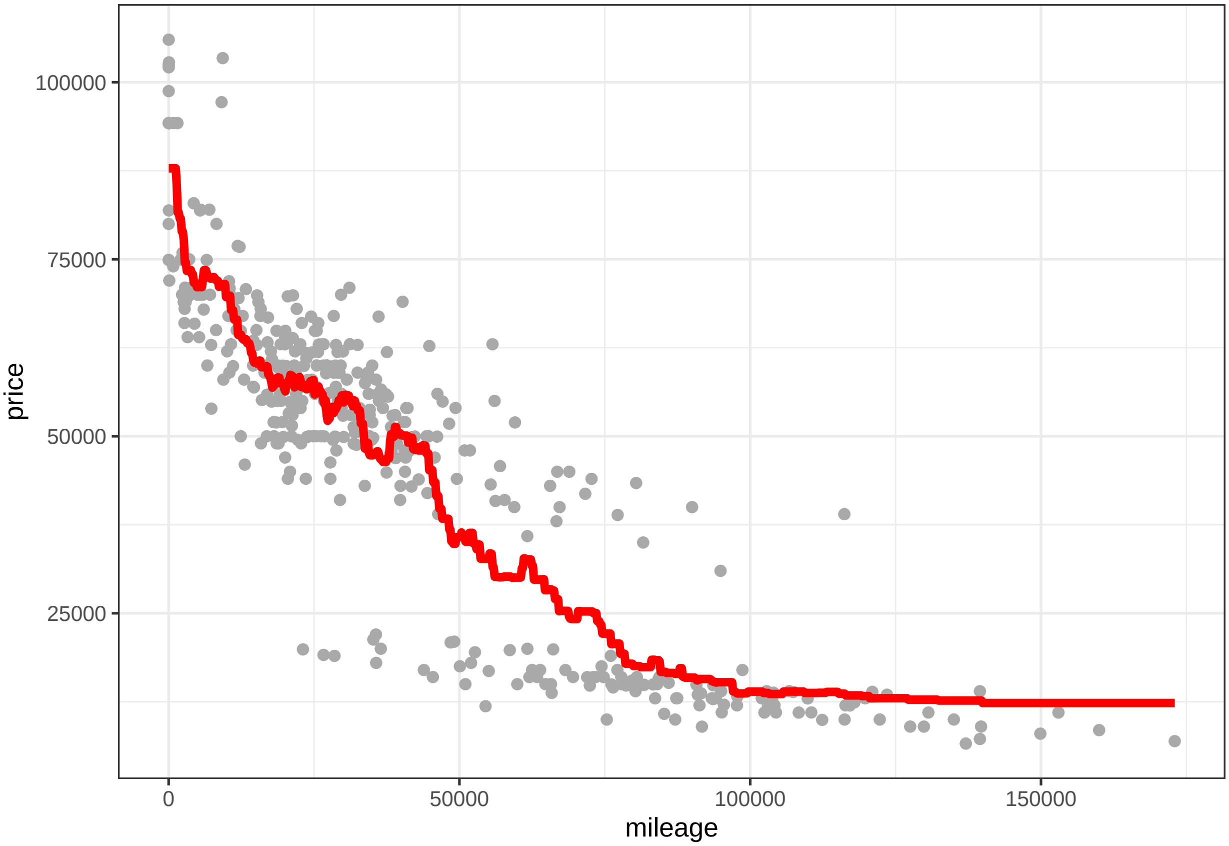

```
## [1] 9894.631
```

K=10



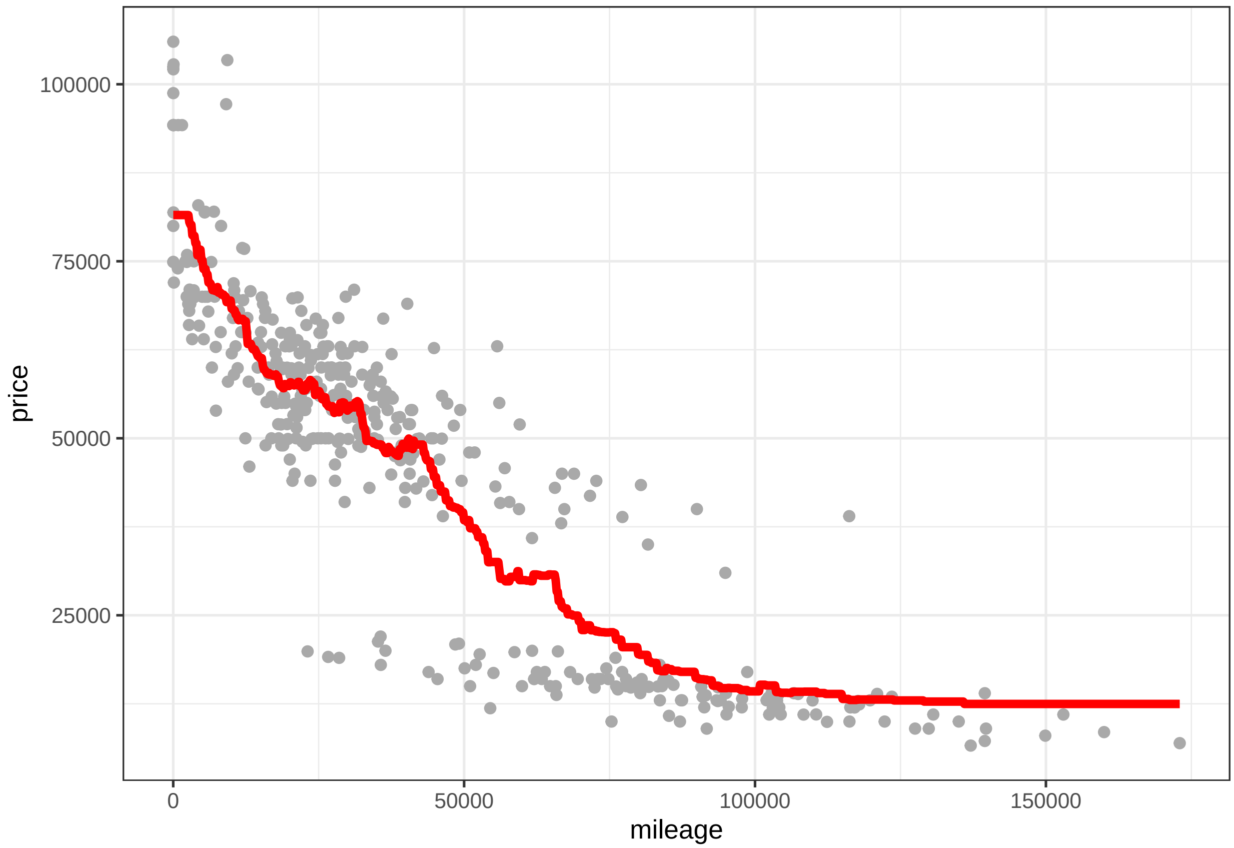

```
## [1] 9906.748
```

K=20



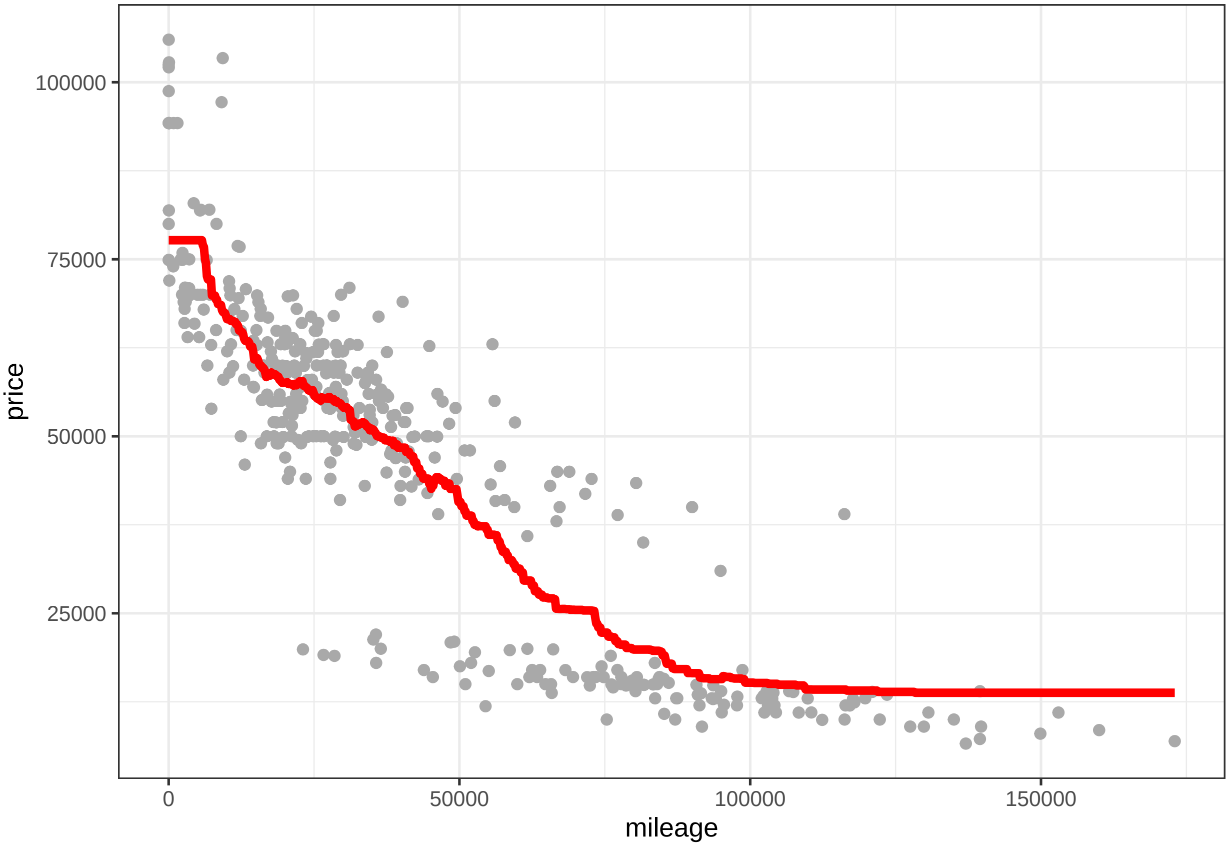

```
## [1] 9448.855
```

K=30



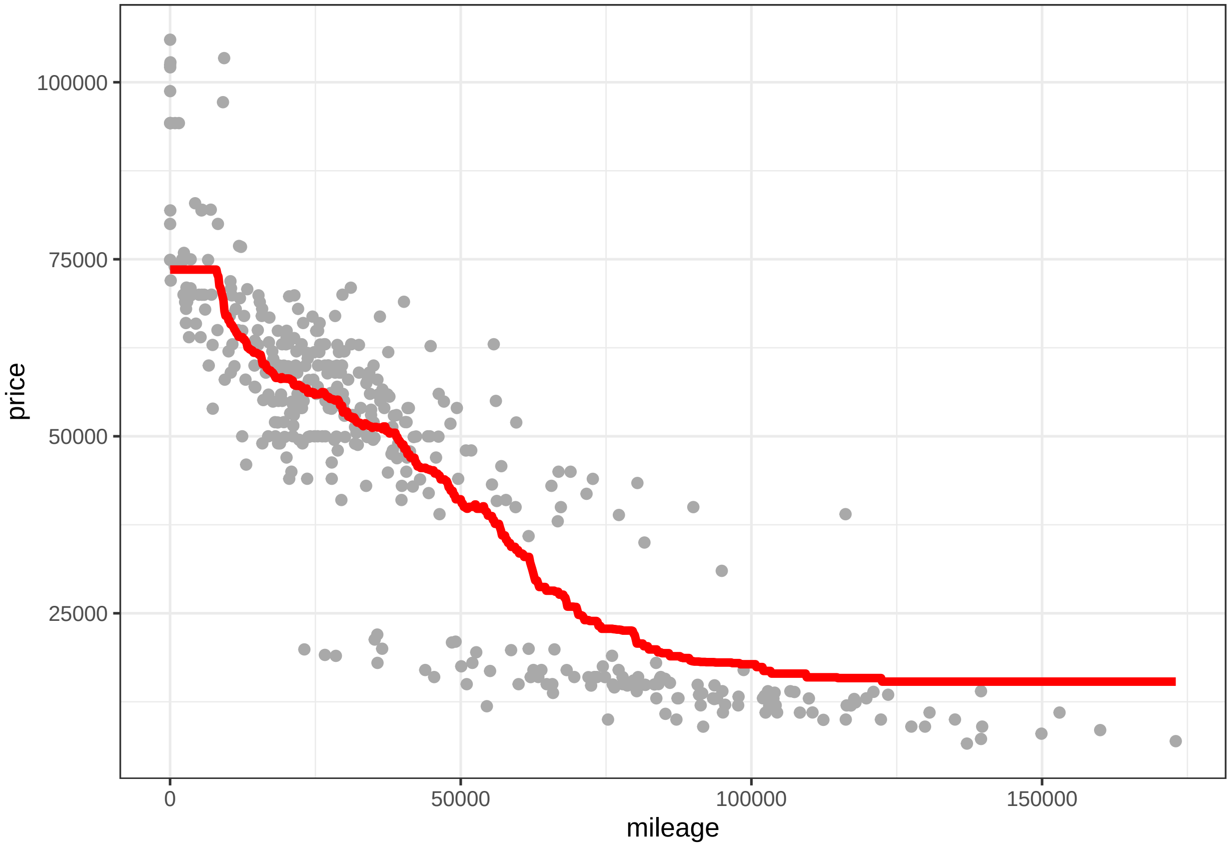

```
## [1] 9286.452
```

K=50



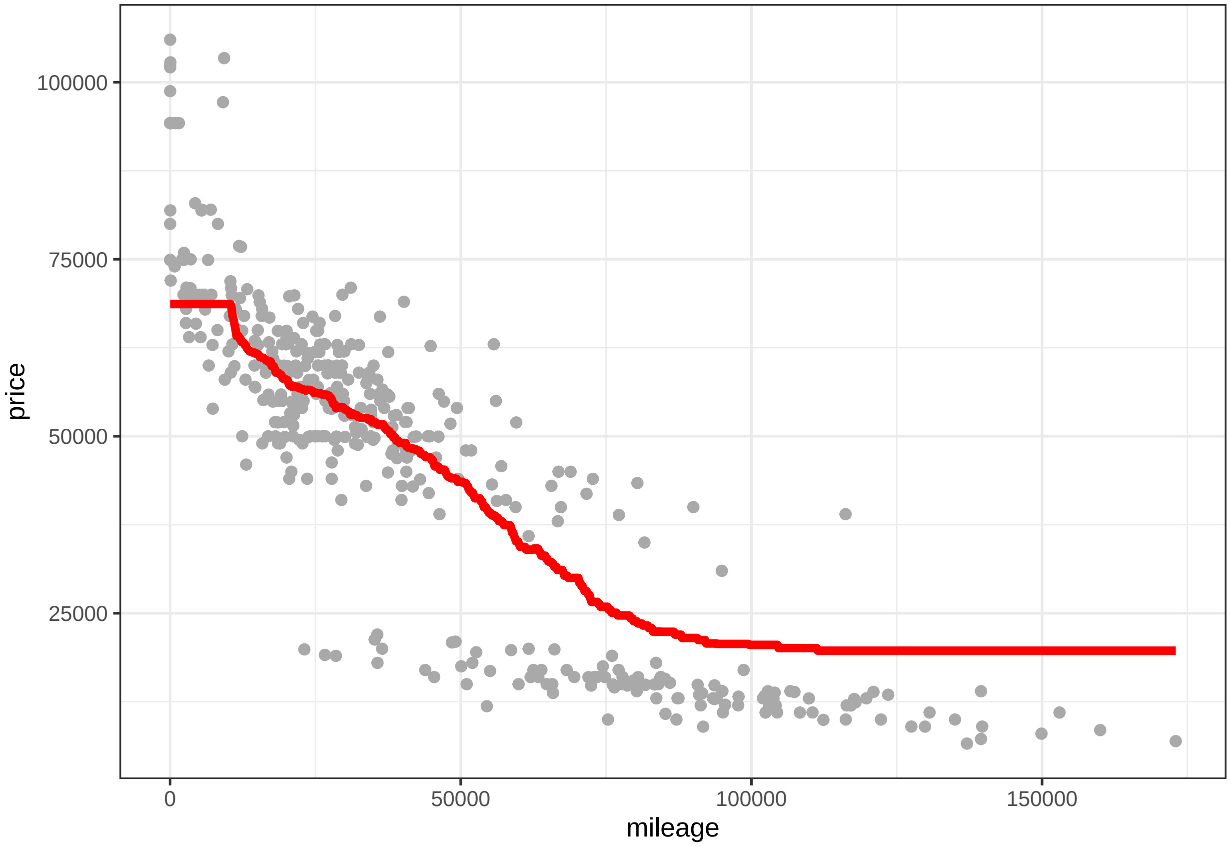

```
## [1] 8961.997
```

K=70



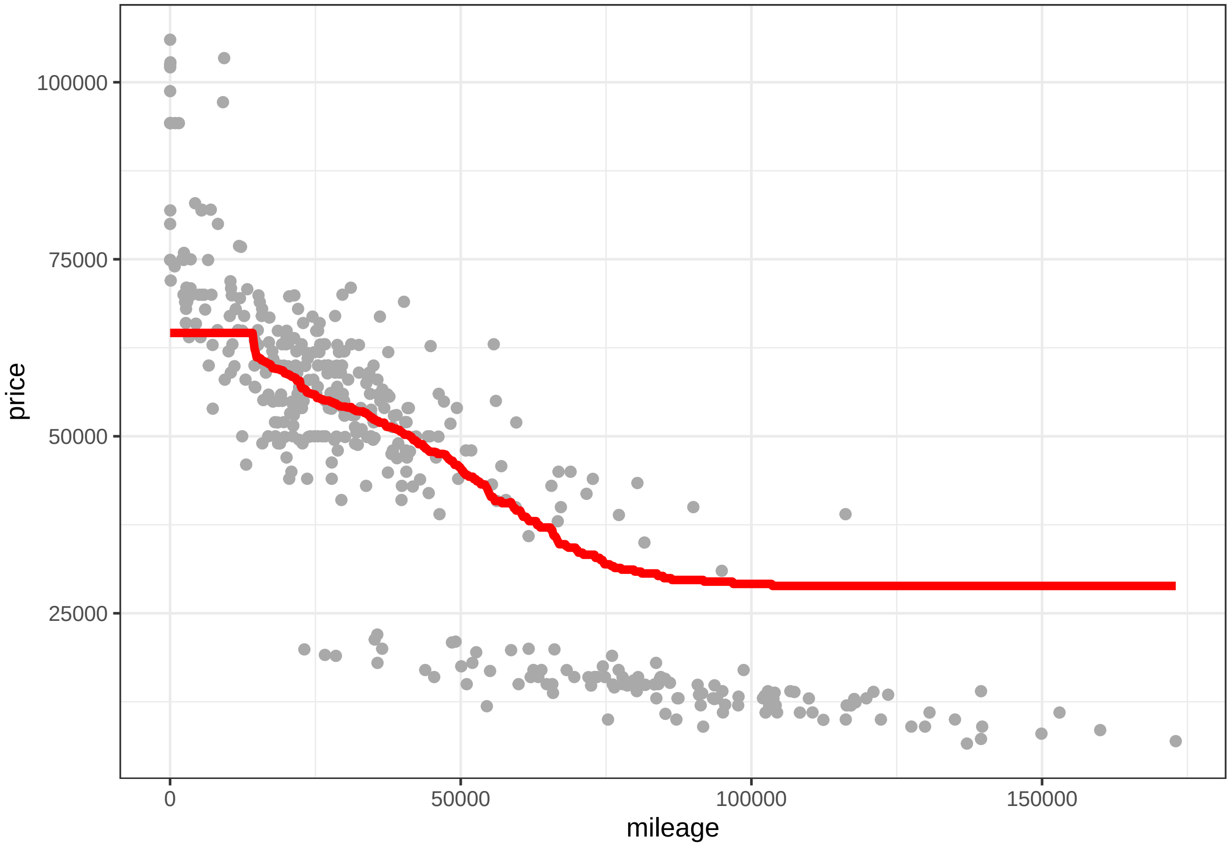

```
## [1] 9047.568
```

K=100



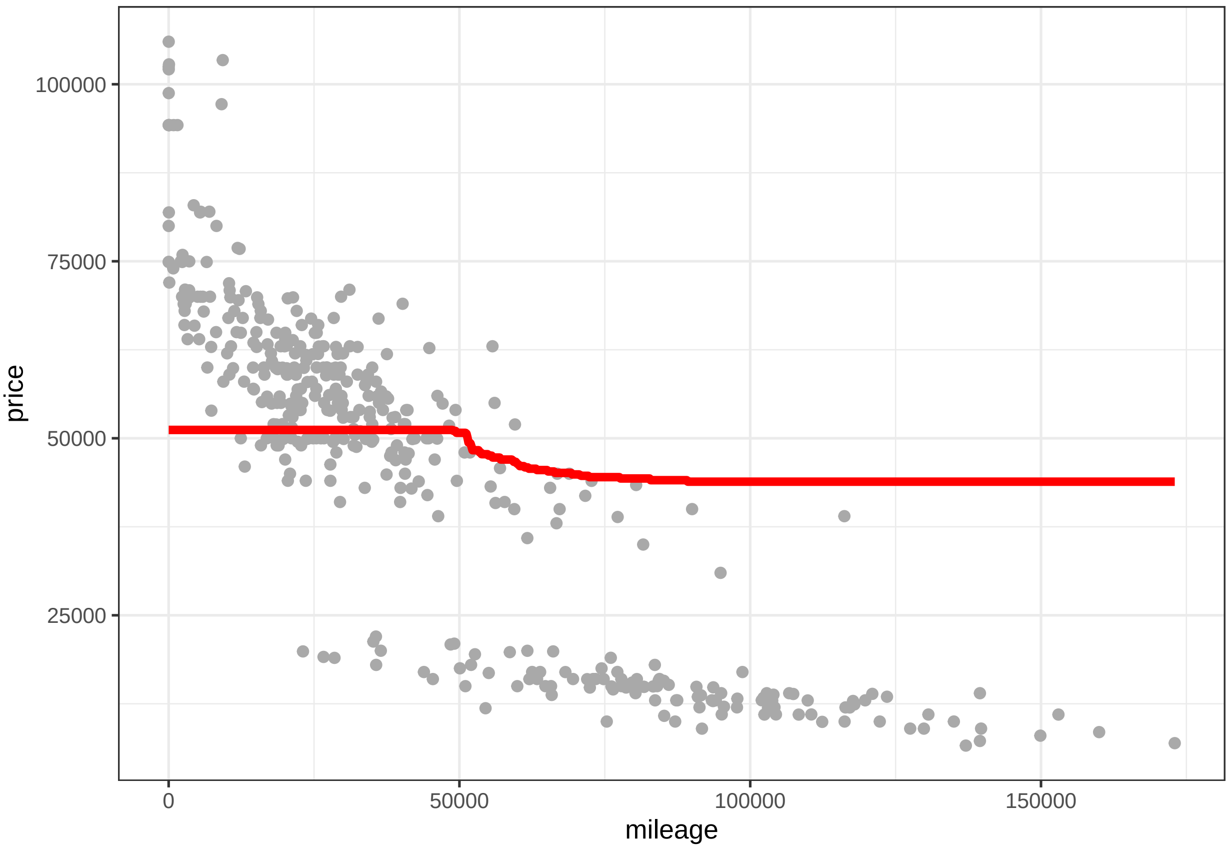

```
## [1] 9639.039
```

K=150



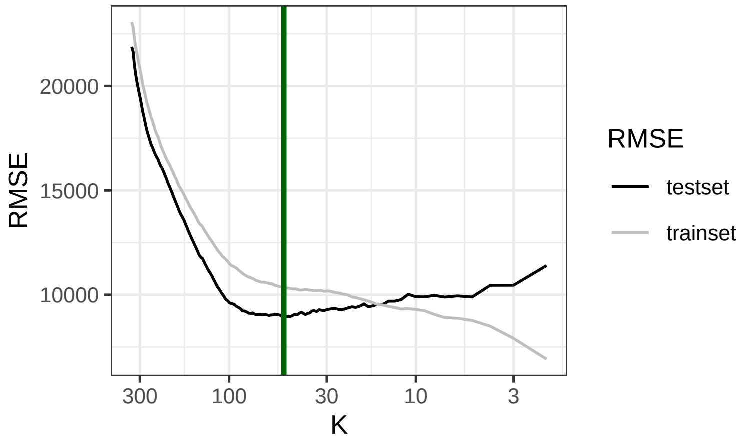

```
## [1] 12231.35
```

K=300

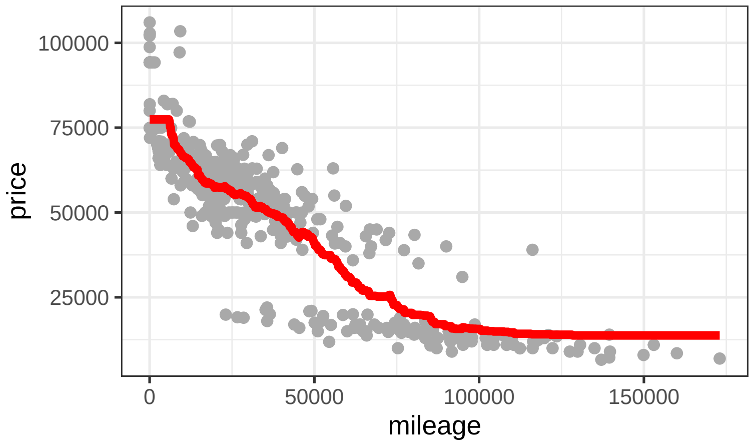



```
## [1] 19486.61
```

K-nearest neighbors: test

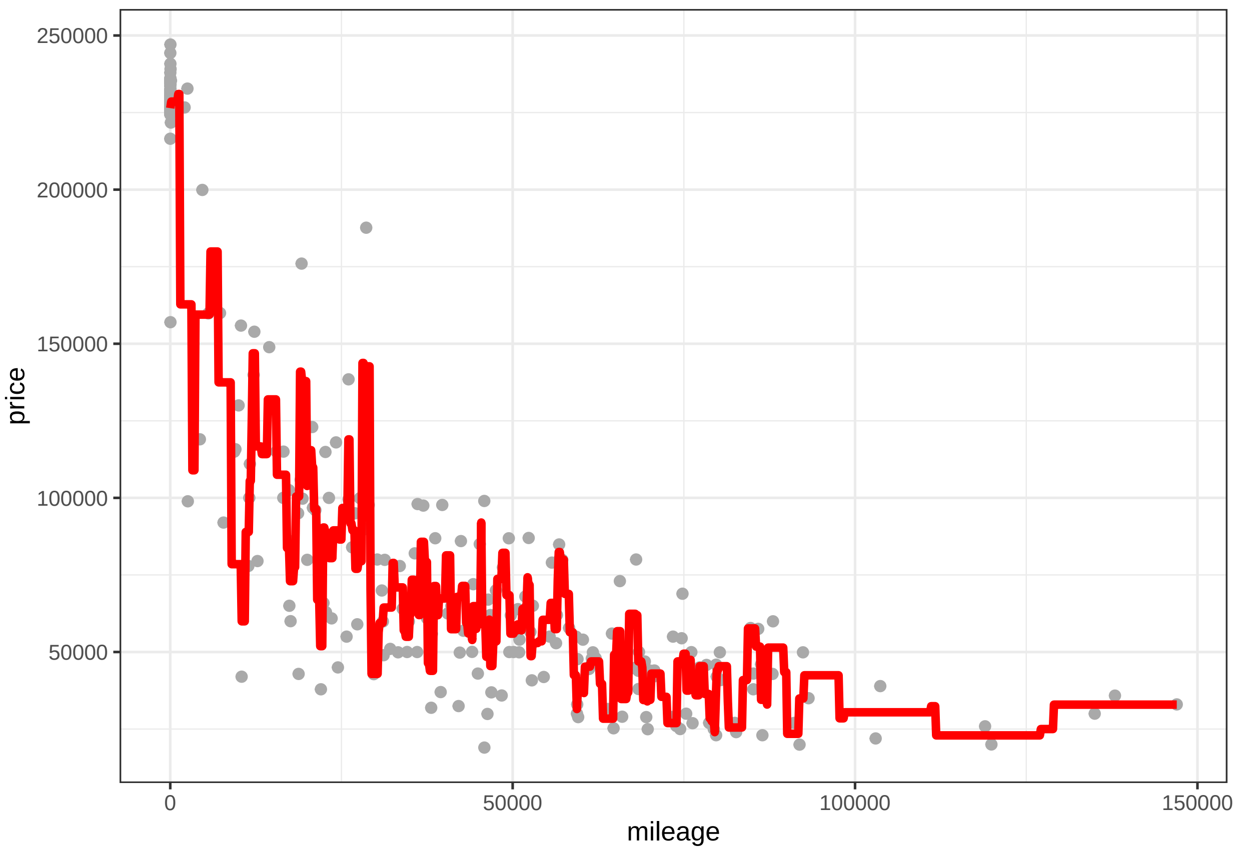


K-nearest neighbors: at the optimal k



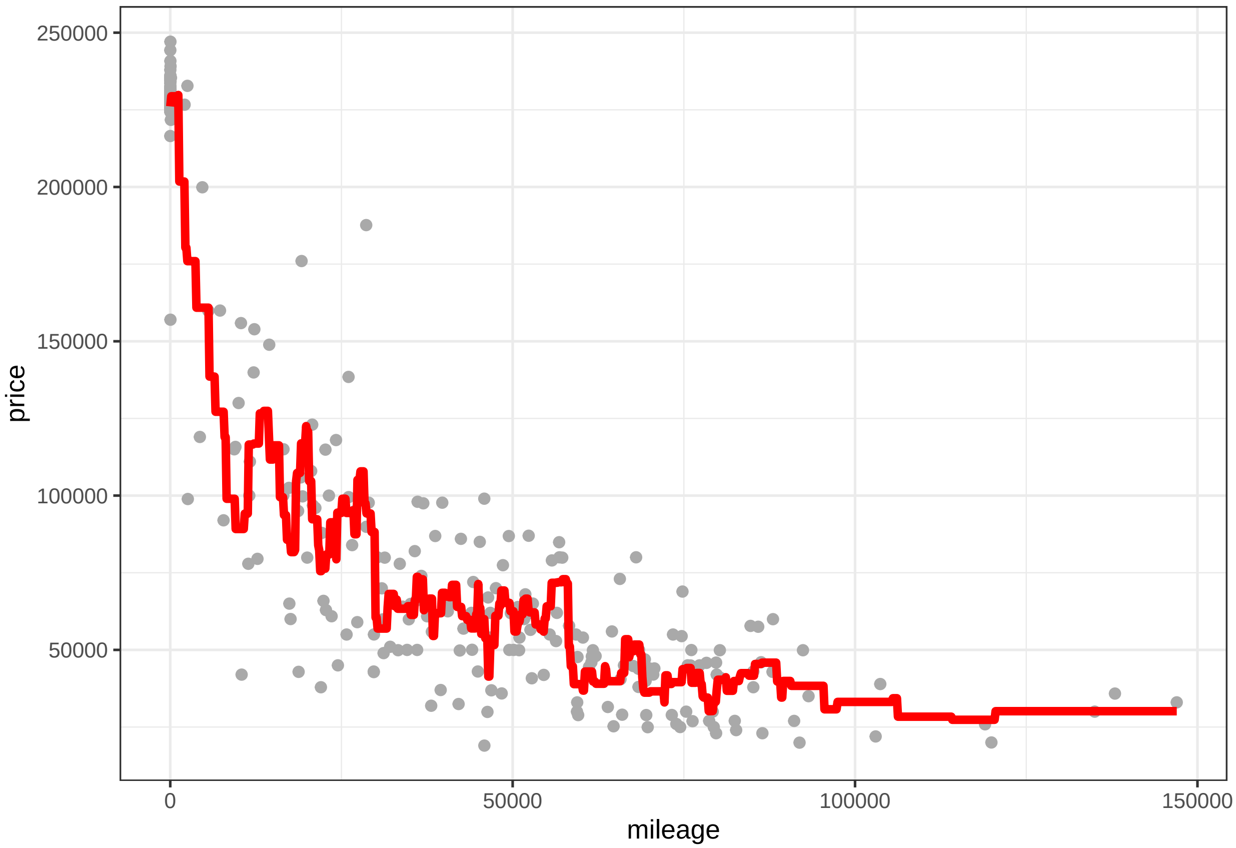

```
## [1] 8920.503
```

K=2




```
## [1] 26684.55
```

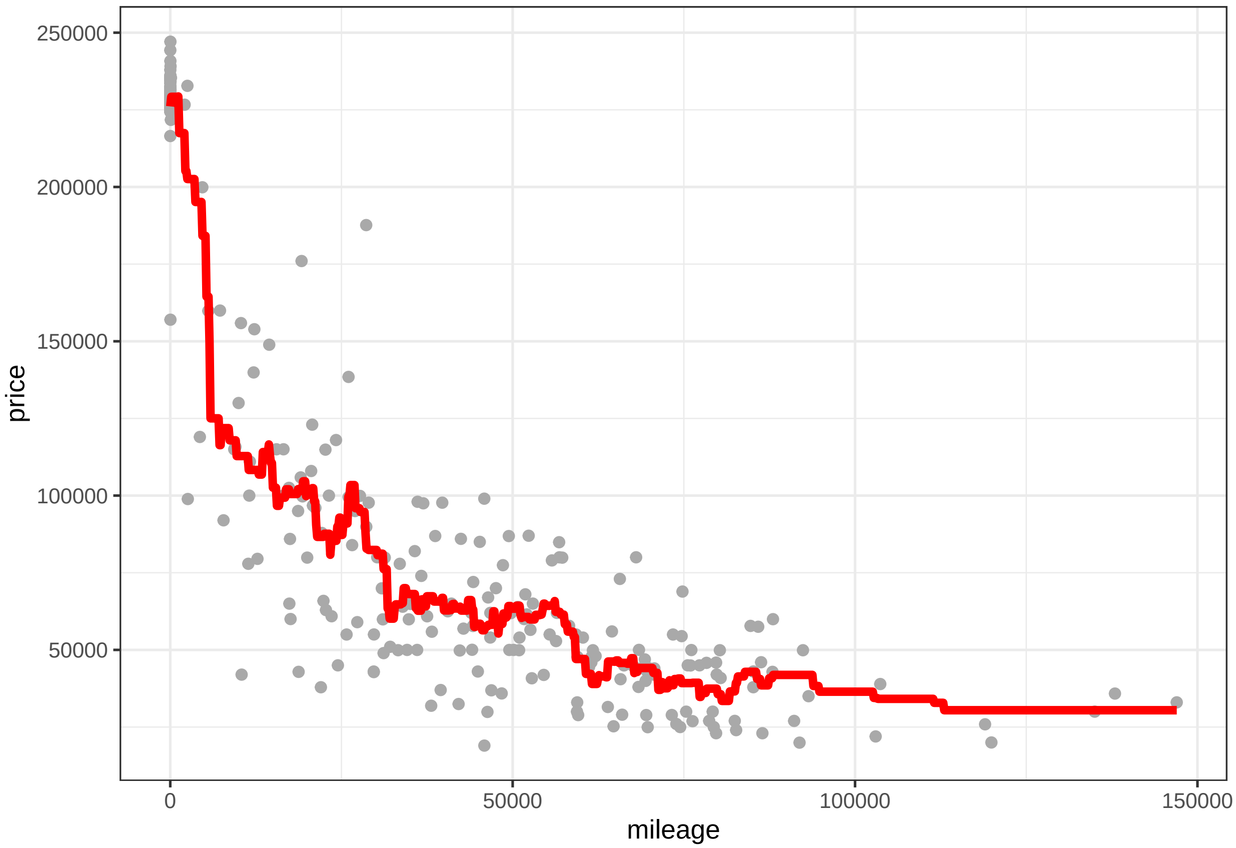
K=5



250000
200000
150000
100000
50000
price
0
50000
100000
150000
mileage

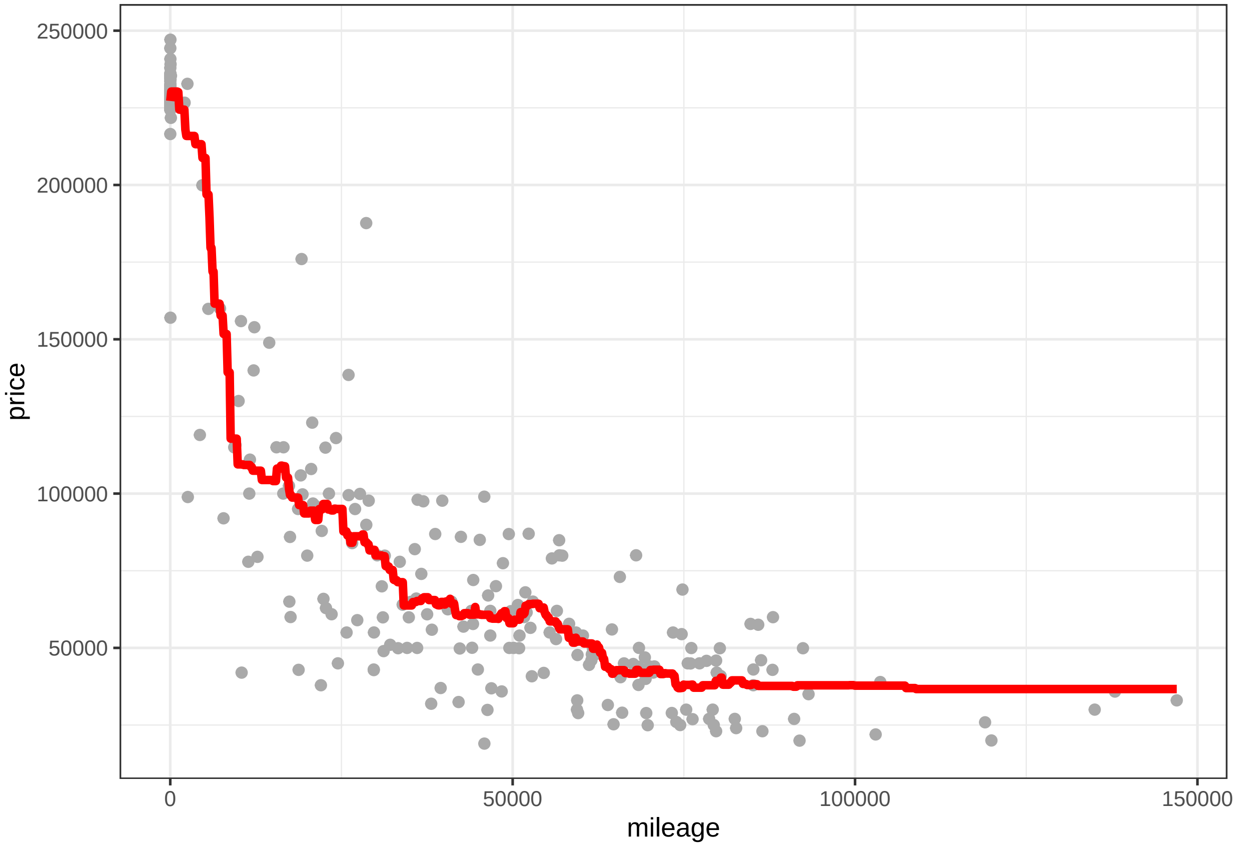

```
## [1] 23351.11
```

K=10


price
250000
200000
150000
100000
50000
0
50000
100000
150000
mileage

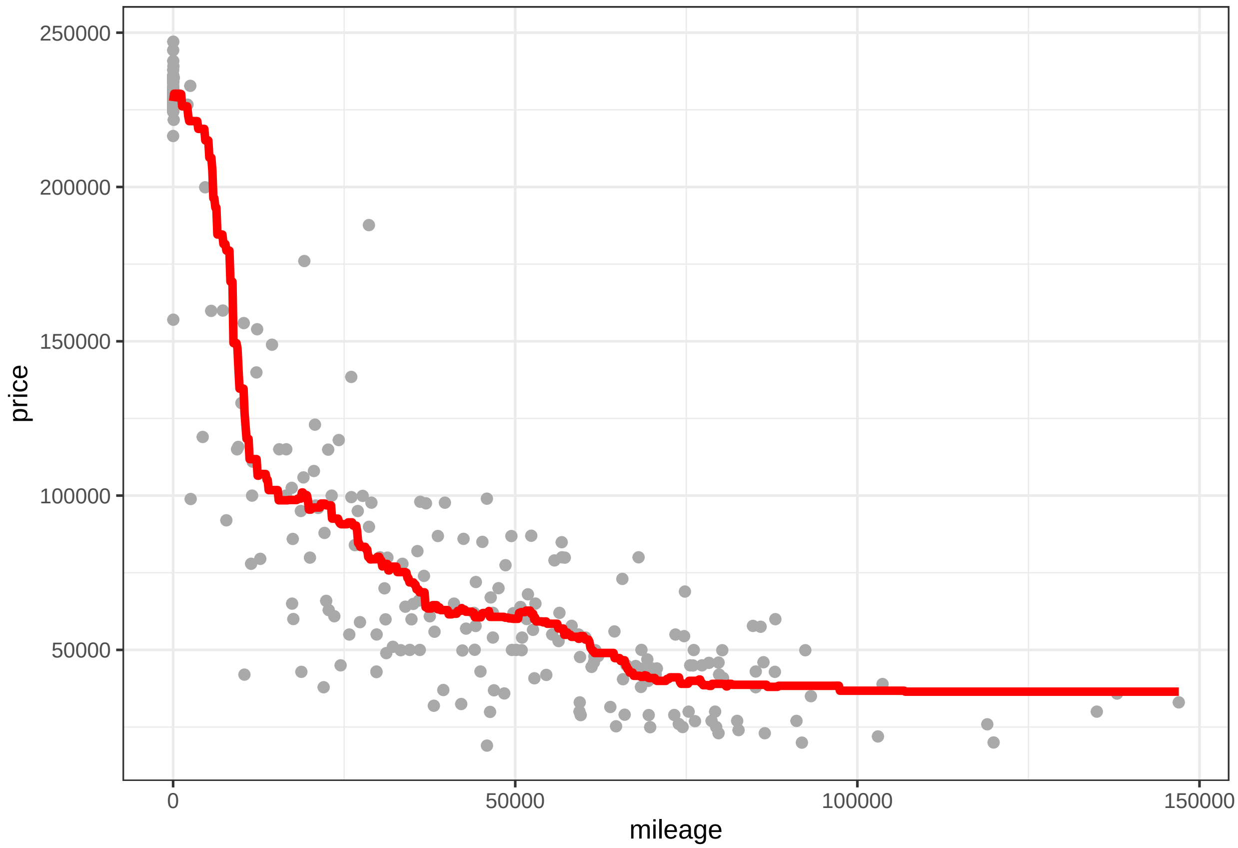
```
## [1] 19079.6
```

K=20



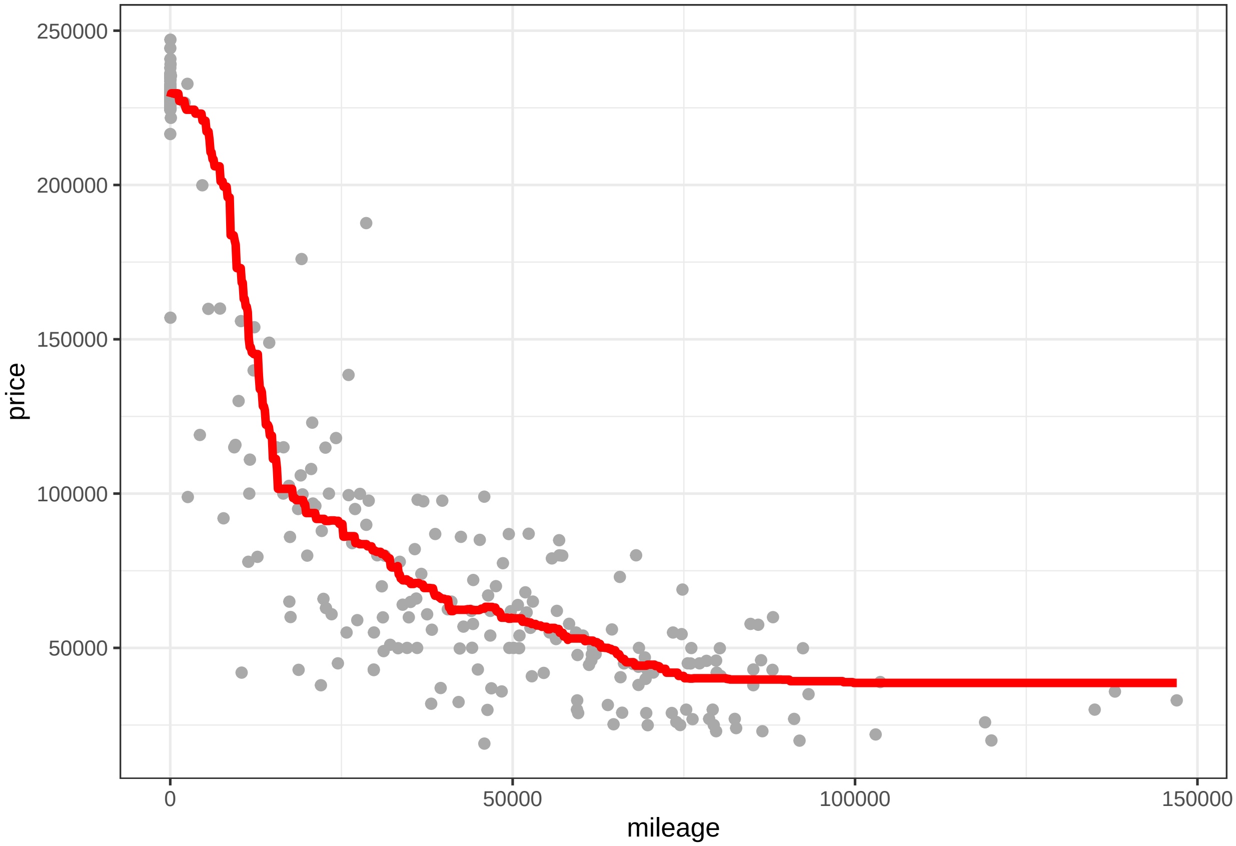

```
## [1] 20732.35
```

K=30



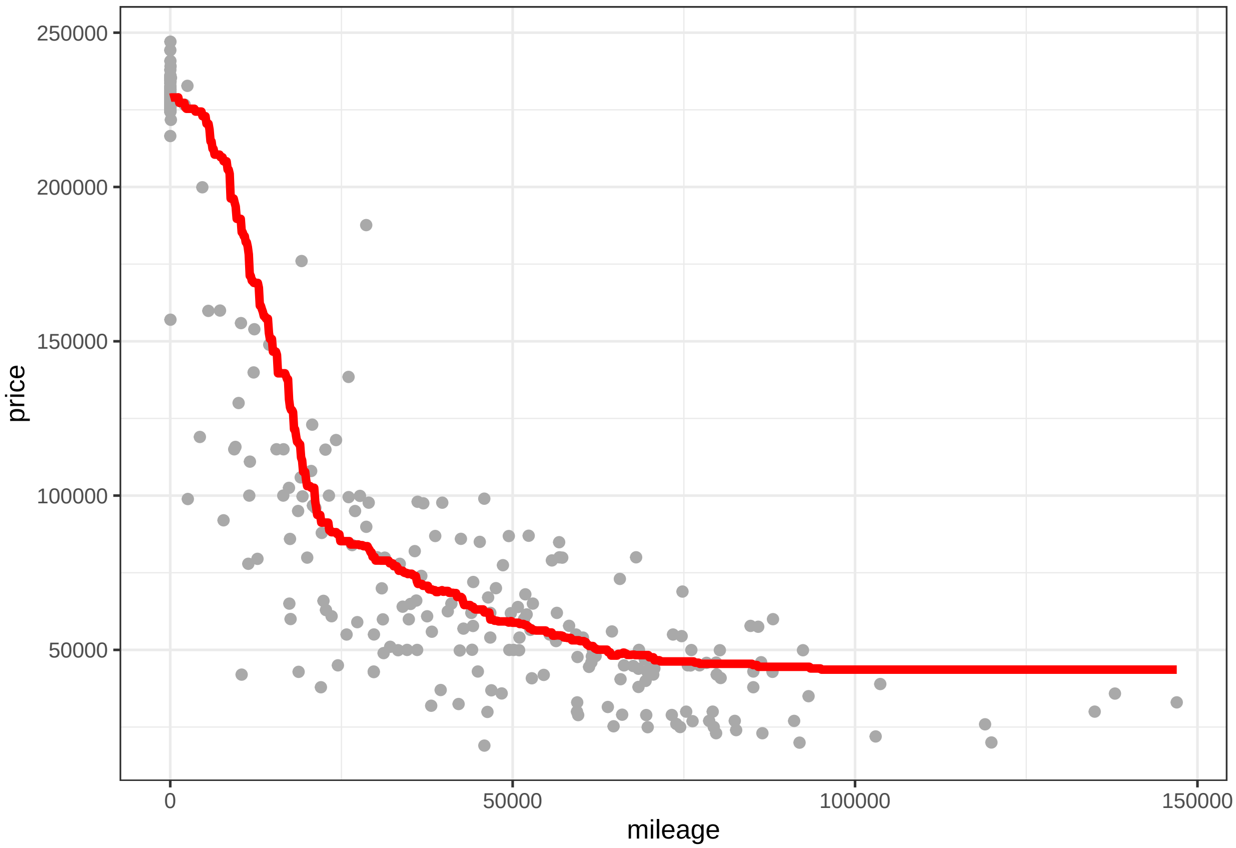

```
## [1] 22286.91
```

K=50



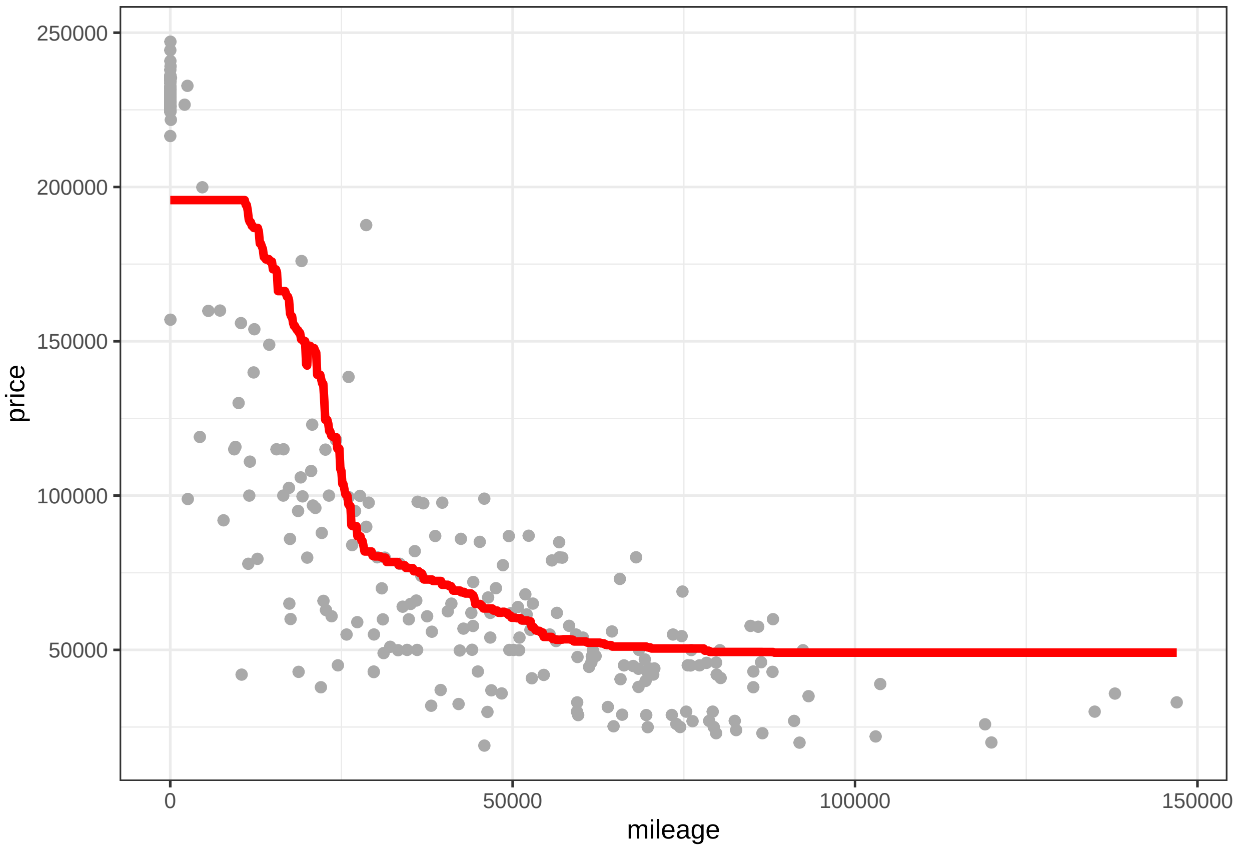

```
## [1] 25423.35
```

K=70



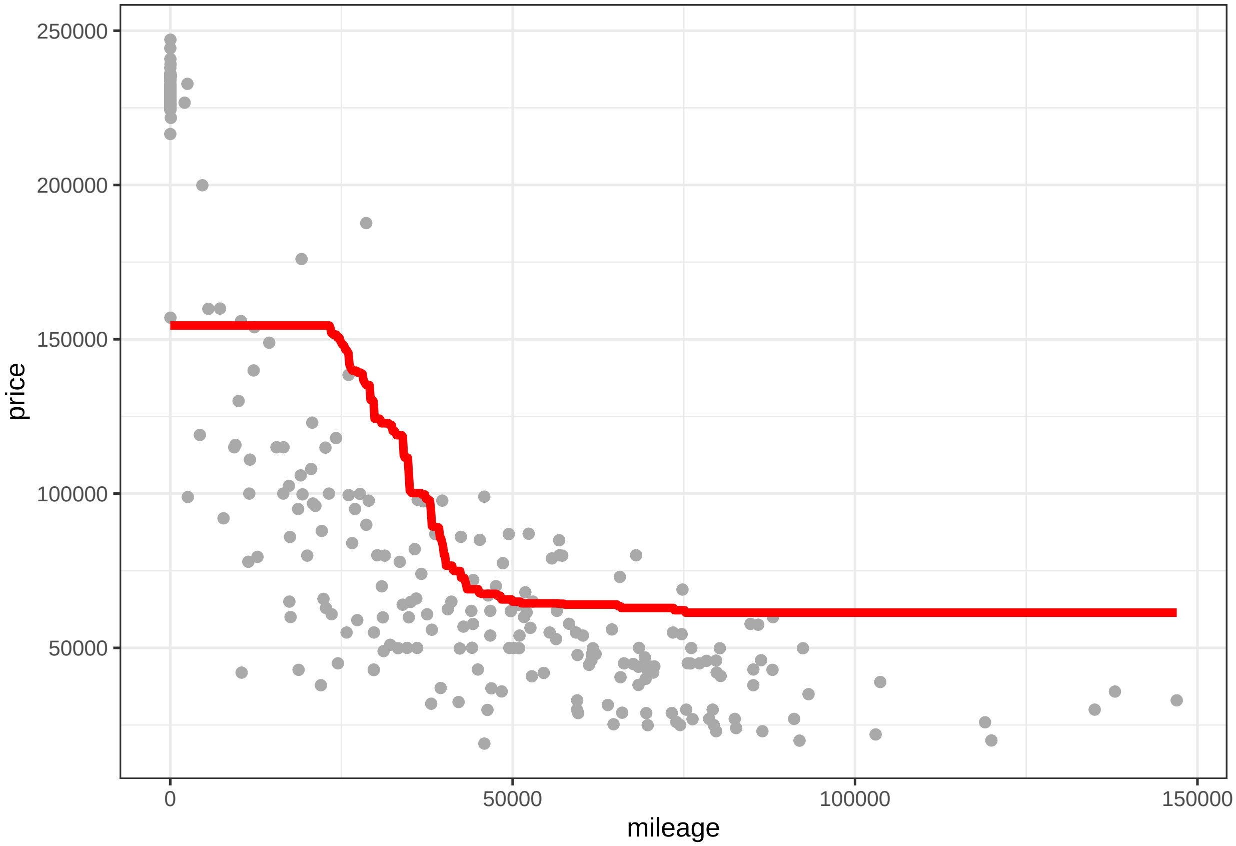

```
## [1] 28435.91
```

K=100



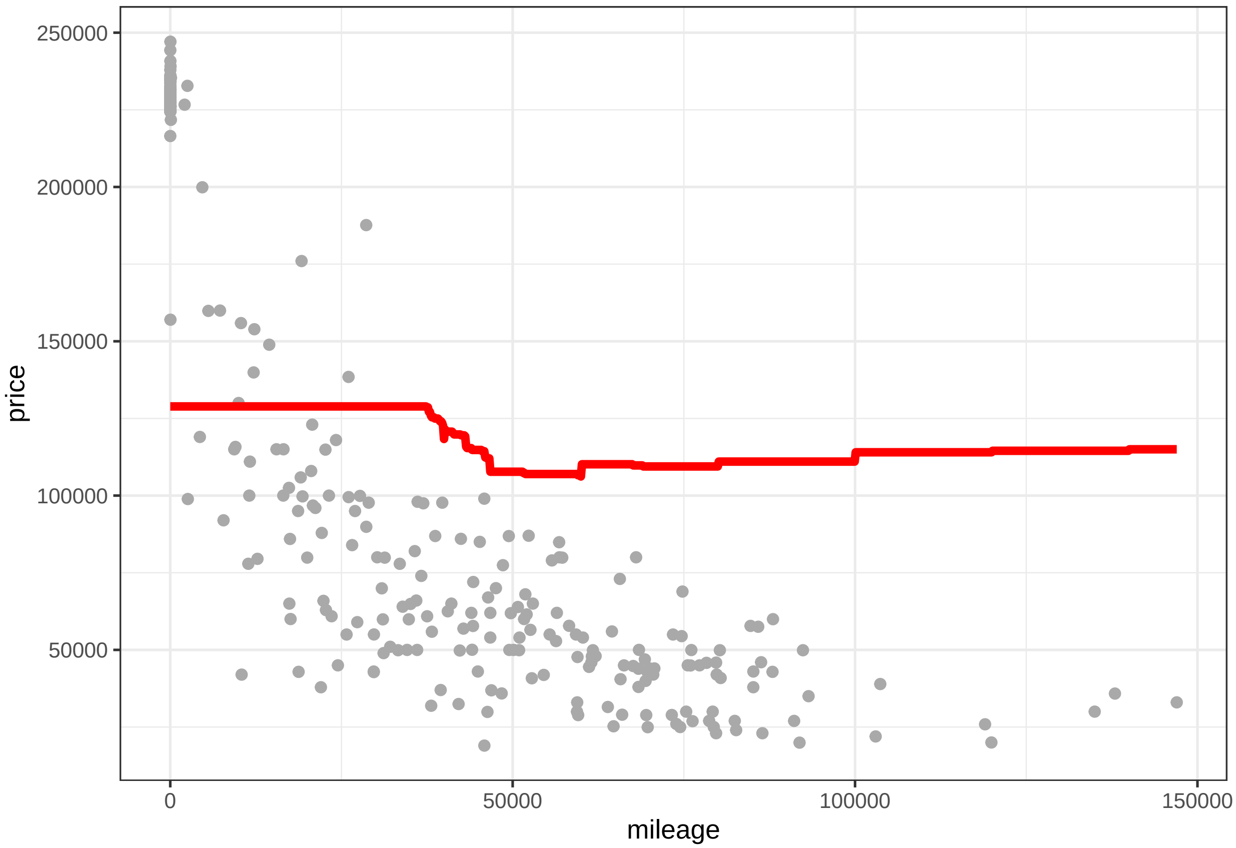

```
## [1] 36754.42
```

K=150



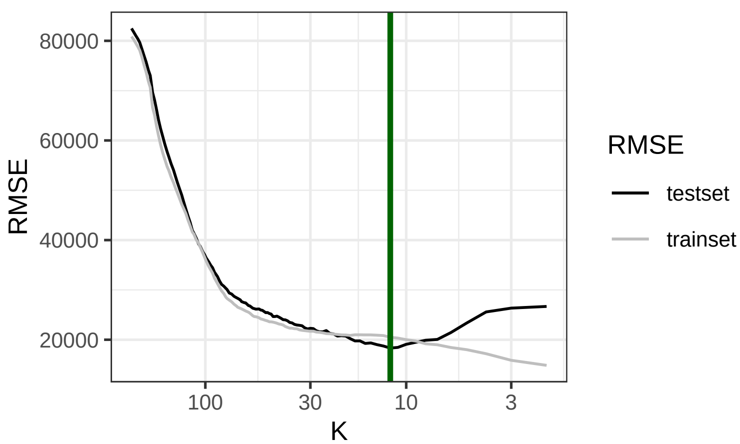

```
## [1] 56140.25
```

K=200

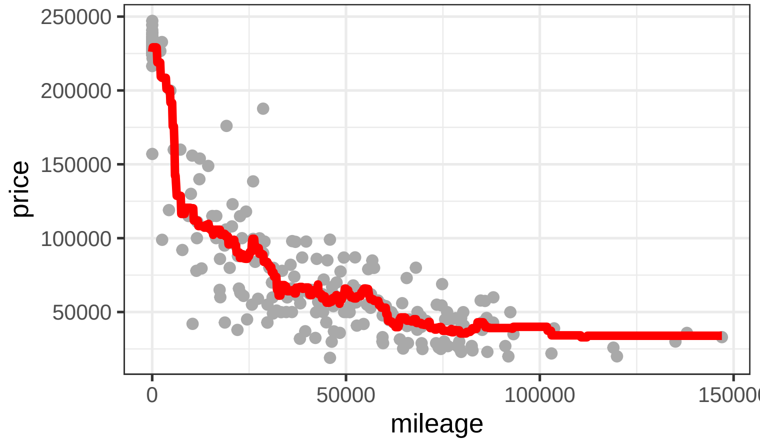



```
## [1] 76670.94
```

K-nearest neighbors: test



K-nearest neighbors: at the optimal k




```
## [1] 18319.42
```

We can see that the trim level of 65 AMG has an optimal k of 12. On the other hand, the 350 trim has an optimal k value of 51. When we have a higher k-value, we will tend to have a higher bias but lower variance. Prices for cars with the 350 trim level vary less than that of cars with the 65 AMG level. Cars with the 350 trim are mostly clustered between 0 to 50000 miles for a price range between $50,000 and $75,000 or below $25,000 when offering mileage between 50,000 and 1,50,000 miles. Here, a higher k-value would be reasonable since we have lower variance within the dataset to be concerned about.

With a lower k-value, we risk having a higher variance but lower bias. Cars with a 65 AMG level trim are more scattered, and therefore, we will need a lower k-value to make more accurate price predictions.

The S65 AMG has features of a sports car and thus, will call a lower price on the second-hand market since its working condition will be more critically evaluated than that of a standard sedan. In this case, that is the S350. S65 AMGs with 0 mileage cost around $25,000 while S350 start off with $50,0000. It may also explain why we have a smaller set of observations for the S65 since there would be a smaller market for a sports car than that for a sedan. The dataset for cars with 350 trim level has 416 observations while the dataset of cars with 65 AMG trim has 292 observations.

Checking RMSE through K-cross validation

###For car 350

```
##    id trim subTrim condition isOneOwner mileage year  color displacement   fuel
## 1 282  350    unsp       CPO          f   21929 2012  Black        3.0 L Diesel
## 2 284  350    unsp       CPO          f   17770 2012 Silver        3.0 L Diesel
## 3 285  350    unsp      Used          f   29108 2012  Black        3.0 L Diesel
## 4 288  350    unsp       CPO          f   35004 2013  White        3.0 L Diesel
## 5 289  350    unsp      Used          t   66689 2012  Black        3.0 L Diesel
## 6 290  350    unsp       CPO          f   19567 2012  Black        3.0 L Diesel
##   state region   soundSystem wheelType wheelSize featureCount price fold_id
## 1    MA    New          unsp      unsp      unsp           82 55994       1
## 2    IL    ENC       Premium     Alloy      unsp           72 60900       5
## 3    VA    SoA          unsp      unsp      unsp            5 54995       1
## 4    NH    New Harman Kardon      unsp      unsp           83 59988       2
## 5    NJ    Mid Harman Kardon     Alloy      unsp           79 37995       5
## 6    LA    WSC       Premium     Alloy      unsp           76 59977       1
```

```
## [1] 11364.93 12374.61 10675.73 11076.96 11866.60
```

```
## [1] 11471.77
```

```
## [1] 297.6562
```

###For car 65 AMG

```
##     id   trim subTrim condition isOneOwner mileage year  color displacement
## 1 1060 65 AMG    unsp       New          f     106 2015  Black        6.0 L
## 2 1062 65 AMG    unsp       New          f      11 2015  Black        6.0 L
## 3 1387 65 AMG    unsp      Used          f   74461 2006 Silver        6.0 L
## 4 2068 65 AMG    unsp      Used          f   73415 2007   Gray        6.0 L
## 5 2141 65 AMG    unsp       CPO          f   17335 2011  Black        6.0 L
## 6 2310 65 AMG    unsp       New          f       7 2015  White        6.0 L
##       fuel state region soundSystem wheelType wheelSize featureCount  price
## 1 Gasoline    NJ    Mid     Premium     Alloy      unsp           73 235375
## 2 Gasoline    CA    Pac     Premium      unsp        20           83 226465
## 3 Gasoline    IL    ENC        unsp     Alloy      unsp           50  24995
## 4 Gasoline    CA    Pac     Premium      unsp      unsp           17  54981
## 5 Gasoline    OH    ENC        unsp      unsp      unsp           92 102500
## 6 Gasoline    CA    Pac        unsp      unsp      unsp            1 230860
##   fold_id
## 1       4
## 2       1
## 3       4
## 4       5
## 5       5
## 6       5

## [1] 43950.73 36972.11 28851.85 39004.58 30680.45

## [1] 35891.95

## [1] 2761.538
```

"‘