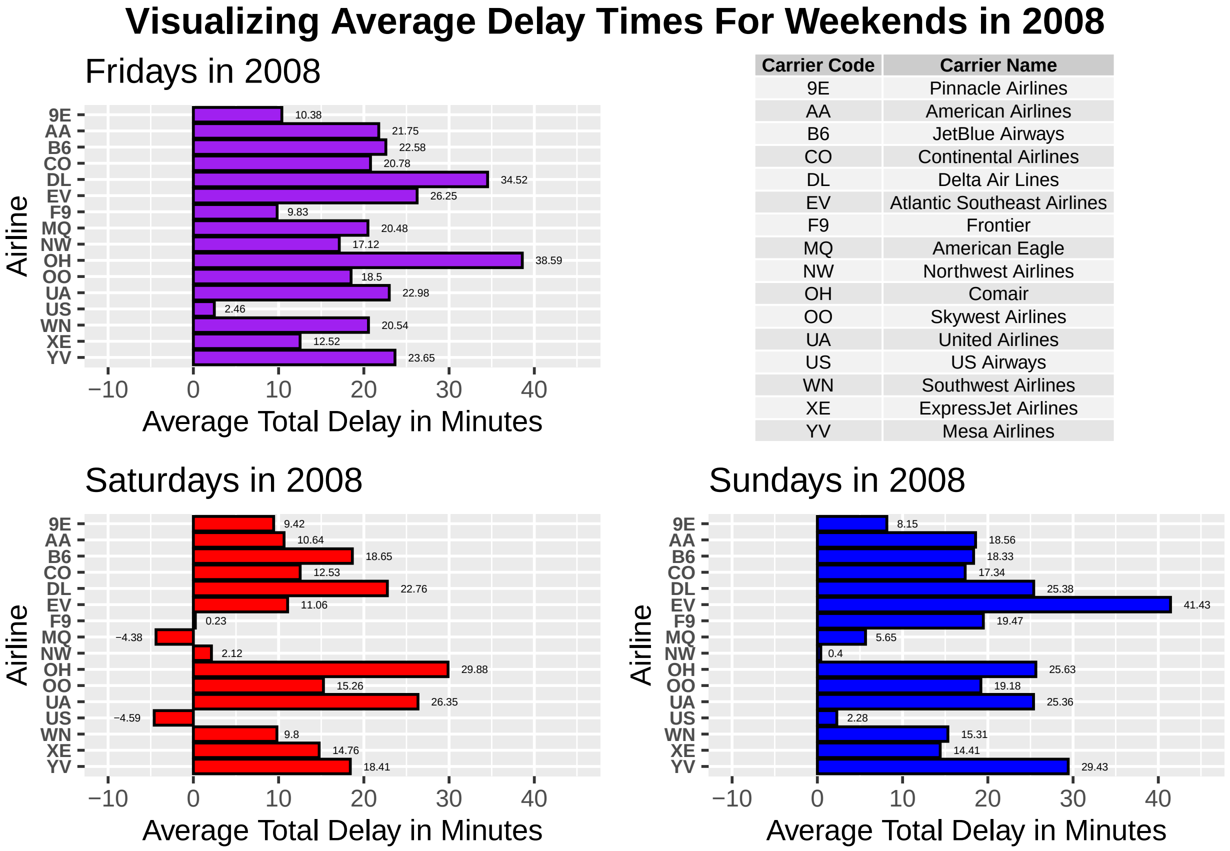
Exercises 1

Jordan Despain and Jack Freeman

01/30/2023

Problem 1



Carrier Code	Carrier Name
9E	Pinnacle Airlines
AA	American Airlines
B6	JetBlue Airways
CO	Continental Airlines
DL	Delta Air Lines
EV	Atlantic Southeast Airlines
F9	Frontier
MQ	American Eagle
NW	Northwest Airlines
OH	Comair
OO	Skywest Airlines
UA	United Airlines
US	US Airways
WN	Southwest Airlines
XE	ExpressJet Airlines
YV	Mesa Airlines

With the "ABIA.csv" data set, we wanted to see which airlines were the best/worst when it comes to delay times for days over the weekend. We also included a table to match carrier codes with the carriers name. We see that US Airways was easily the best in terms of delay times for each day, and Comair and Atlantic Southeast Airlines were among the worst.

Problem 2

A)

With the “olympics_top20.csv” data set, we filtered by sex and sport to get all the female competitors who participated in an Athletics event. We then created a new set with just their names and height and sorted descending in height. Here is a preview of the set,

```
# A tibble: 10 x 2
# Groups:   name [10]
   name                                       height
   <chr>                                       <int>
 1 Valerie Kasanita Adams-Vili (-Price)          193
 2 Stephanie Karenmonica Brown-Trafton           193
 3 Blanka Vlai                                   193
 4 Anastasia Kelesidou                           192
 5 Yulimar del Valle Rojas Rodrguez              192
 6 Ruth Beita Vila                               191
 7 Nadine Kleinert-Schmitt                       190
 8 Tatyana Sergeyevna Chernova                   189
 9 Aleksandra Georgiyevna Chudina                188
10 Nataliya Venediktovna Lisovskaya (-Sedykh)    188
```

With this new data set, we now want to find the 95th percentile of heights for these competitors. We use the quantile function to get,

```
95%
183
```

B)

With the “olympics_top20.csv” data set, we filtered by sex to grab the data for female competitors and the women’s events. We then grouped by events and took the standard deviation of the heights of the women who competed in each event. Here is a glimpse of what these two steps return,

```
# A tibble: 10 x 3
# Groups:   name [10]
   name                                    height event
   <chr>                                    <int> <chr>
 1 Ann Kristin Aarnes                         182 Football Women's Football
 2 Mariya Vasilyevna Abakumova (-Tarabina)    179 Athletics Women's Javelin Thr~
 3 Tamila Rashidovna Abasova                  163 Cycling Women's Sprint
 4 Reema Abdo                                 173 Swimming Women's 4 x 100 metr~
 5 Irene Abel                                 160 Gymnastics Women's Team All-A~
 6 Elvan Abeylegesse                          159 Athletics Women's 5,000 metres
 7 Nelli Mikhaylovna Abramova                 171 Volleyball Women's Volleyball
 8 Svetlana Olegovna Abrosimova               188 Basketball Women's Basketball
 9 Ginko Abukawa-Chiba                        148 Gymnastics Women's Team All-A~
10 Andreea Roxana Acatrinei                   150 Gymnastics Women's Team All-A~
```

```
# A tibble: 10 x 2
   event                                       std_dev_height
   <chr>                                                <dbl>
 1 Athletics Women's 1,500 metres                        5.03
 2 Athletics Women's 10 kilometres Walk                  4.31
 3 Athletics Women's 10,000 metres                       5.41
 4 Athletics Women's 100 metres                          6.29
 5 Athletics Women's 100 metres Hurdles                  4.68
 6 Athletics Women's 20 kilometres Walk                  5.28
 7 Athletics Women's 200 metres                          5.14
 8 Athletics Women's 3,000 metres                        5.48
 9 Athletics Women's 3,000 metres Steeplechase           6.06
10 Athletics Women's 4 x 100 metres Relay                5.57
```

We then used the max function to find the event with the greatest variability. Here is the result,

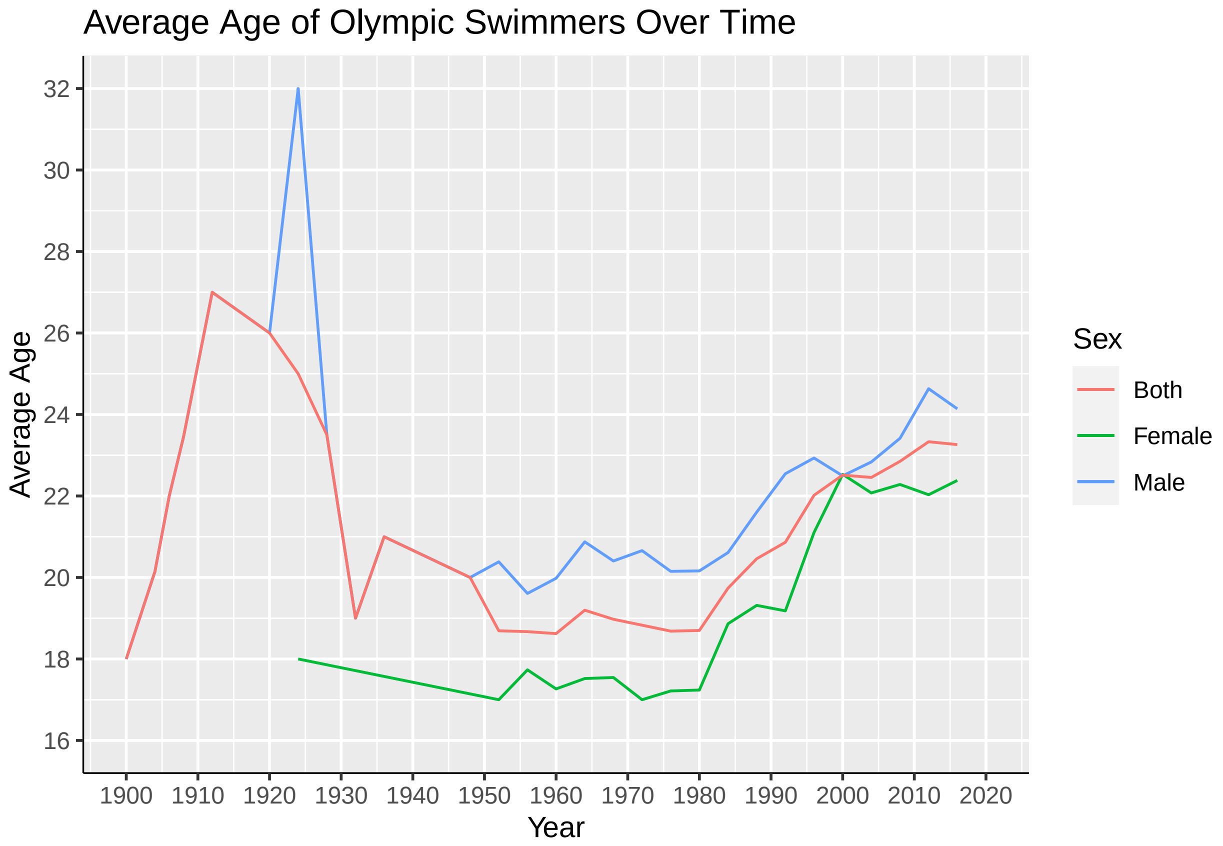
```
# A tibble: 1 x 2
  event                      std_dev_height
  <chr>                               <dbl>
1 Rowing Women's Coxed Fours           10.9
```

C)

With the “olympics_top20.csv” data set, we filtered the data for swimmers only. We then grouped by sex and year and took the average age of the swimmers. Here is a glimpse of what this data looks like,

```
# A tibble: 7 x 3
# Groups:   year [7]
   year sex   average_age
  <int> <chr>       <dbl>
1  1900 M            18
2  1904 M            20.1
3  1906 M            22
4  1908 M            23.5
5  1912 M            27
6  1920 M            26
7  1924 F            18
```

We now want to plot the data to show the trends in average age for females, males, and both combined. Here are the results,

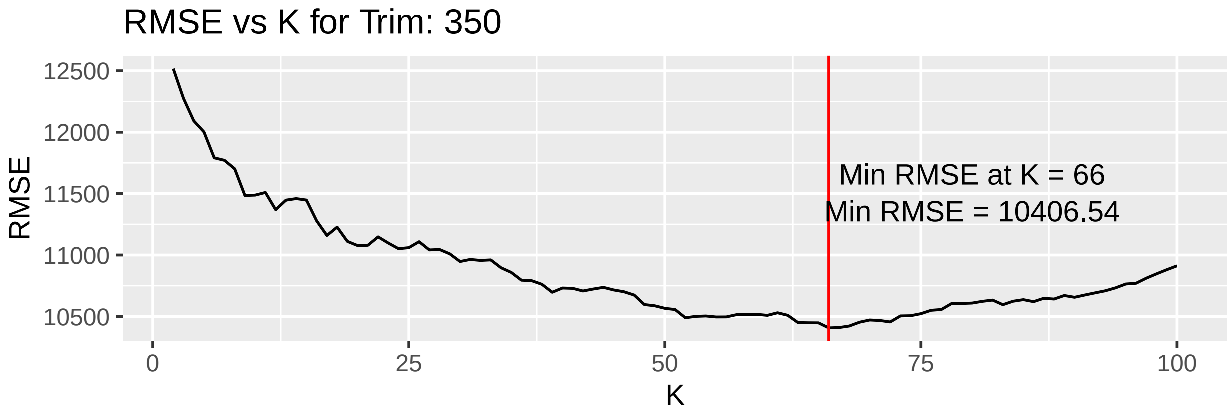


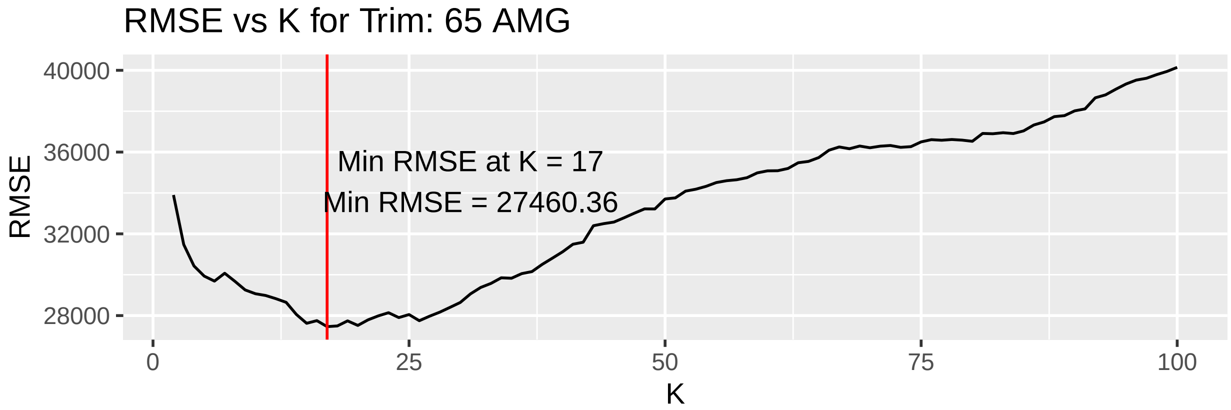
We see the average age of Olympic swimmers started out very volatile in the early 1900s with a sharp increase then a similar decline. Around the 1950s we see a bottoming out and the average age starts to increase through the most recent year in the data. There were not female swimmers in the data in the earliest years, but once they began to regularly show up in the data, the trend (an increasing in average age) is very similar for both female and male swimmers.

Problem 3

With the “sclass.csv” data set, we filtered the data into two smaller subsets, one for trim = 350 and the other for trim = 65 AMG. We then split these two subsets of data into a training and testing set for each individual trim. We then ran K-nearest neighbors on each subset starting at K=2 and increasing K by 1 up to K=100. For each value K, we fit the model to the training sets and made predictions on our test sets. We then calculated the RMSE for each value of K for both subsets.

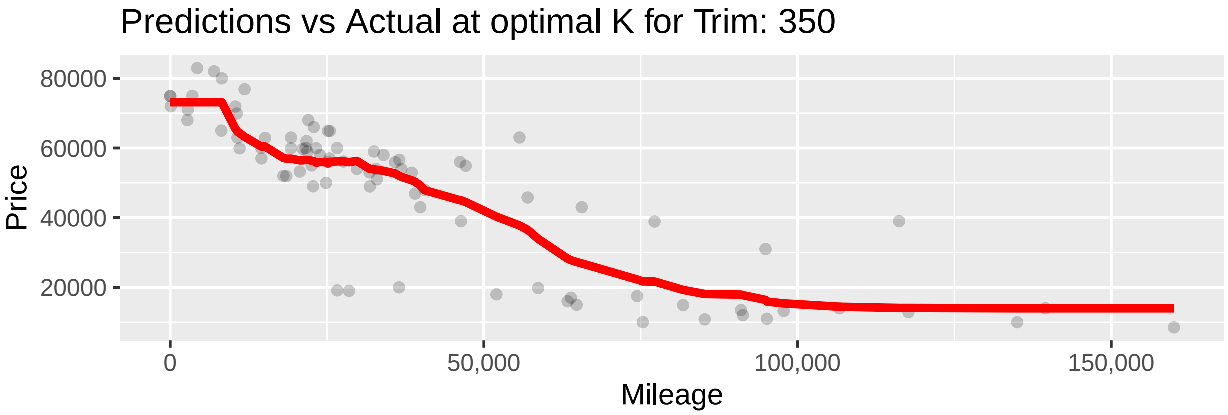
We now want to plot RMSE versus K for each trim to see where it bottoms out. Here are the plots for both trims,

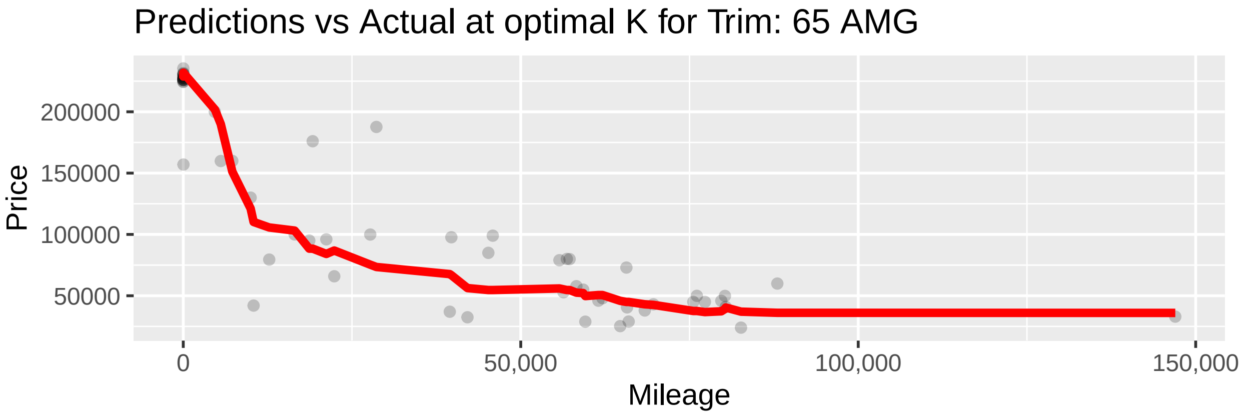




We see the RMSE bottoms out at K=66 for the 350 trim and K=17 for the 65 AMG trim. These are our optimal value of K for each trim.

With our optimal K values for each trim, we now want to plot the fitted models. Here are the models for both trims,





It looks like we have pretty good predictions for both trims.

The 350 trim yields a higher optimal K than the 65 AMG trim most of the time it seems. We believe this is because the 350 trim data set is larger than the 65 AMG data set, and a larger K may help it capture more information about the data. We ran our code many times though and it seemed to often alternate which one is lower, so it is hard to say. We think it has a lot to do with how the data is split and what values happen to go into the training set and which go into the test set.