Homework 1 Writeup

Robert Petit

2/5/2022

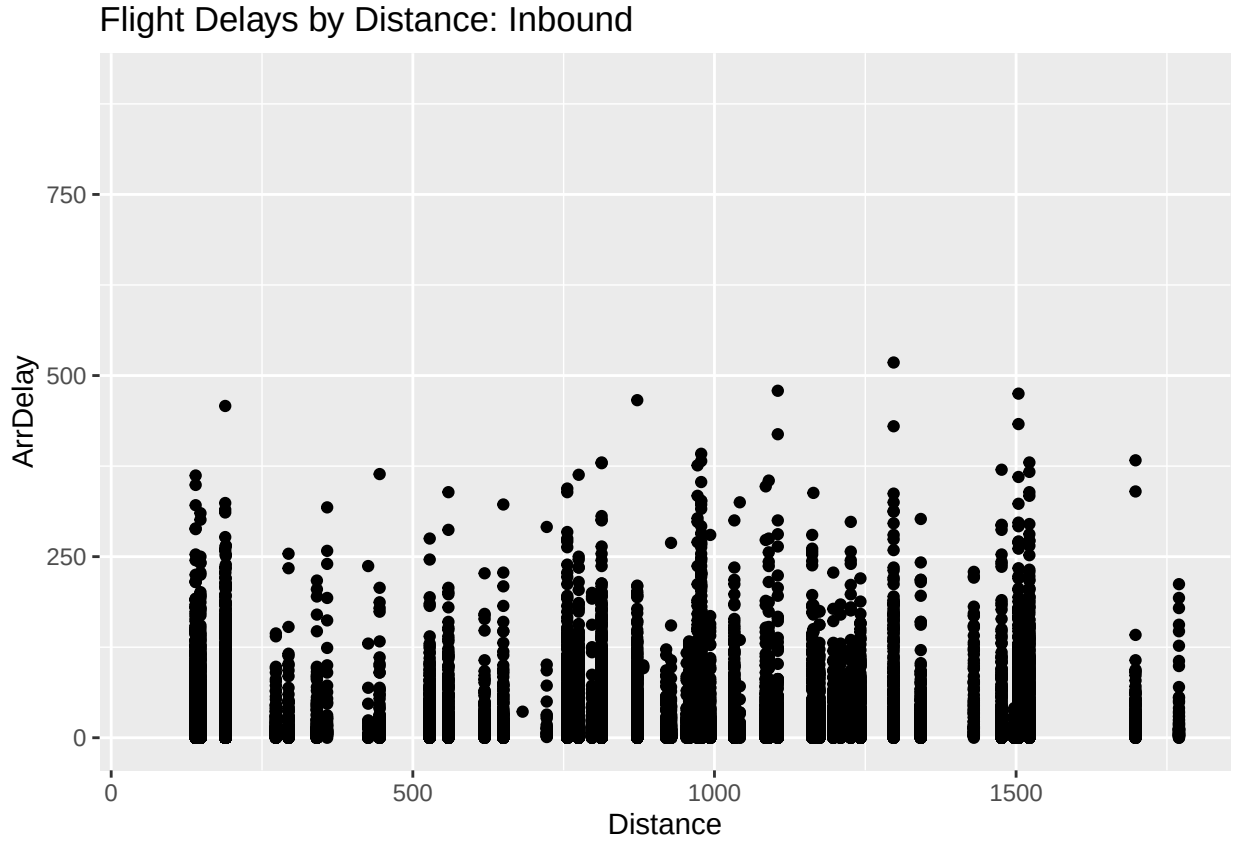
Question 1

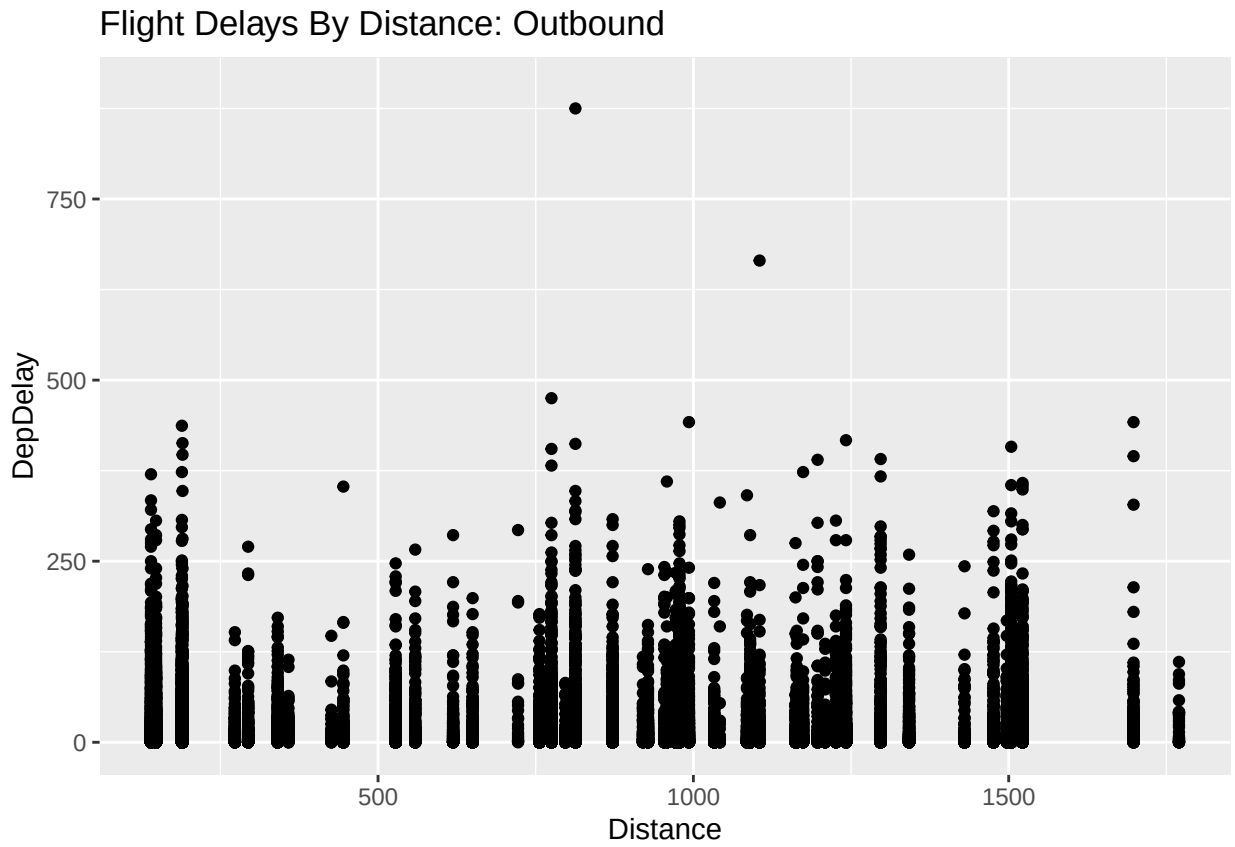
Key Question

The key question here is twofold: first, do flights across the US tend to be delayed by distance or on an airport-by-airport basis. Based on that result, we also want an idea of what the best day of the week is to fly.

Methods and Figures Pt. 1

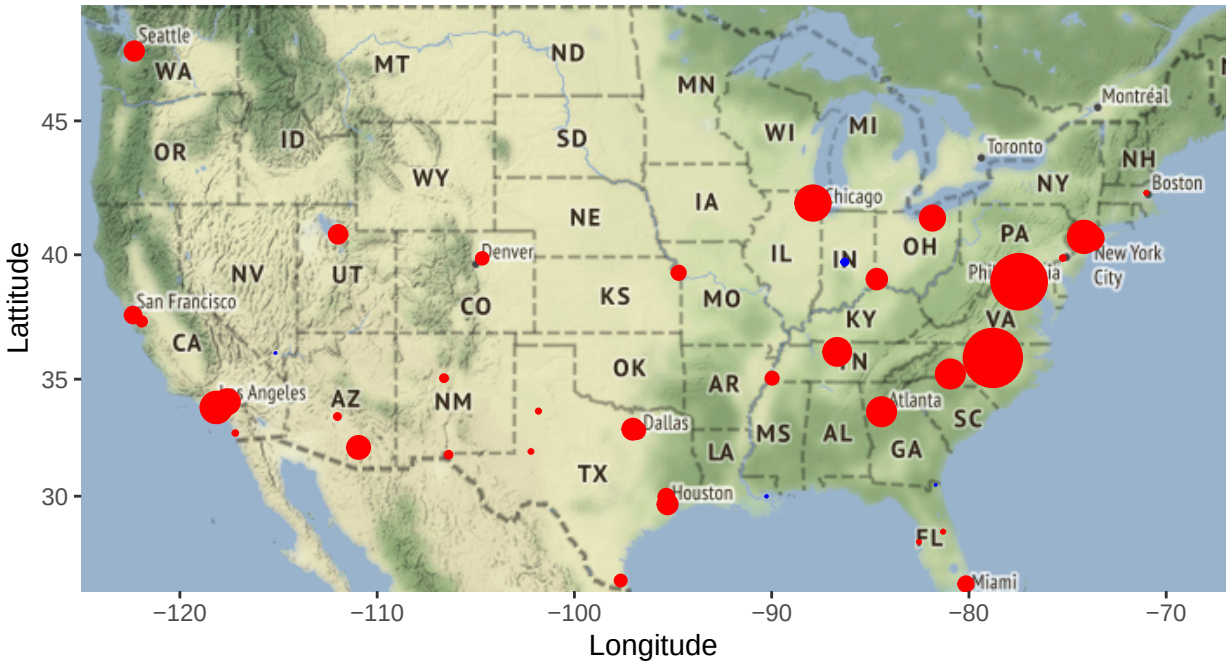
The answer to the first questions becomes fairly obvious when checking two different graphs. First, we can simply plot the distance and expected delays from incoming and outgoing flights



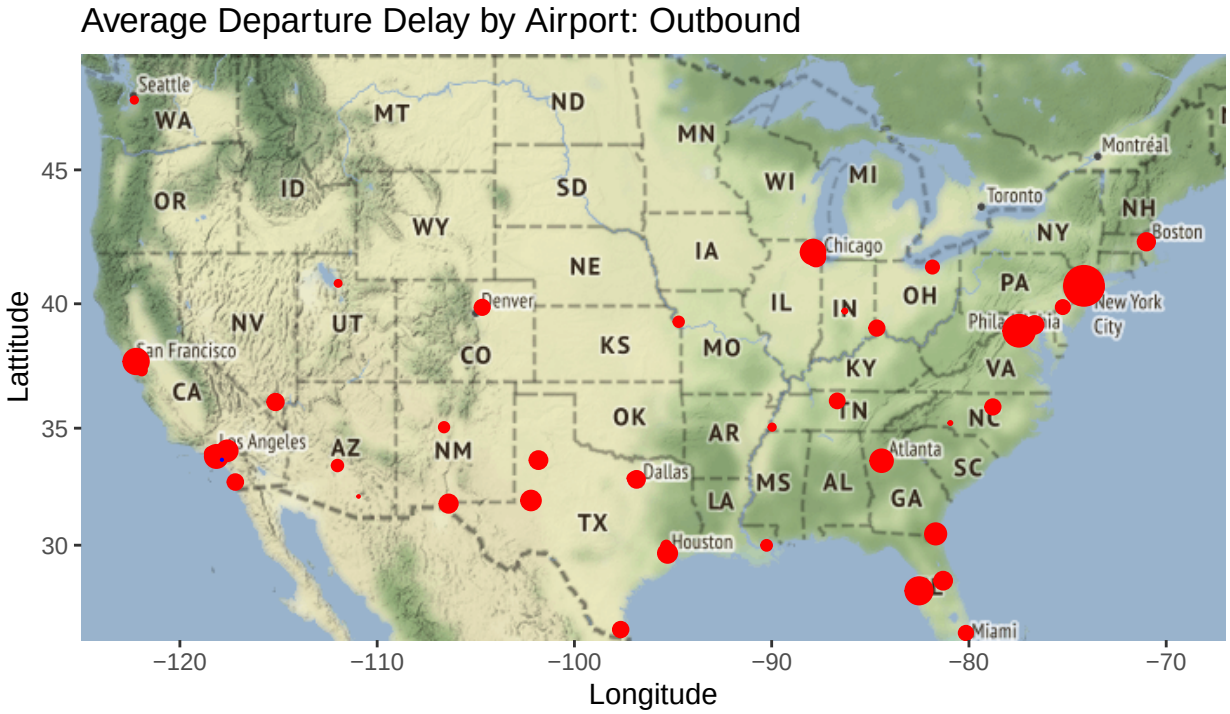


To do an eyeball test of the airport-dependency of delays, we can map the delays to the US

Average Arrival Delay by Airport: Inbound



Larger points correspond to larger delays/gains.
Red dots are delays, blue dots are early arrivals. Note the lack of blue dots

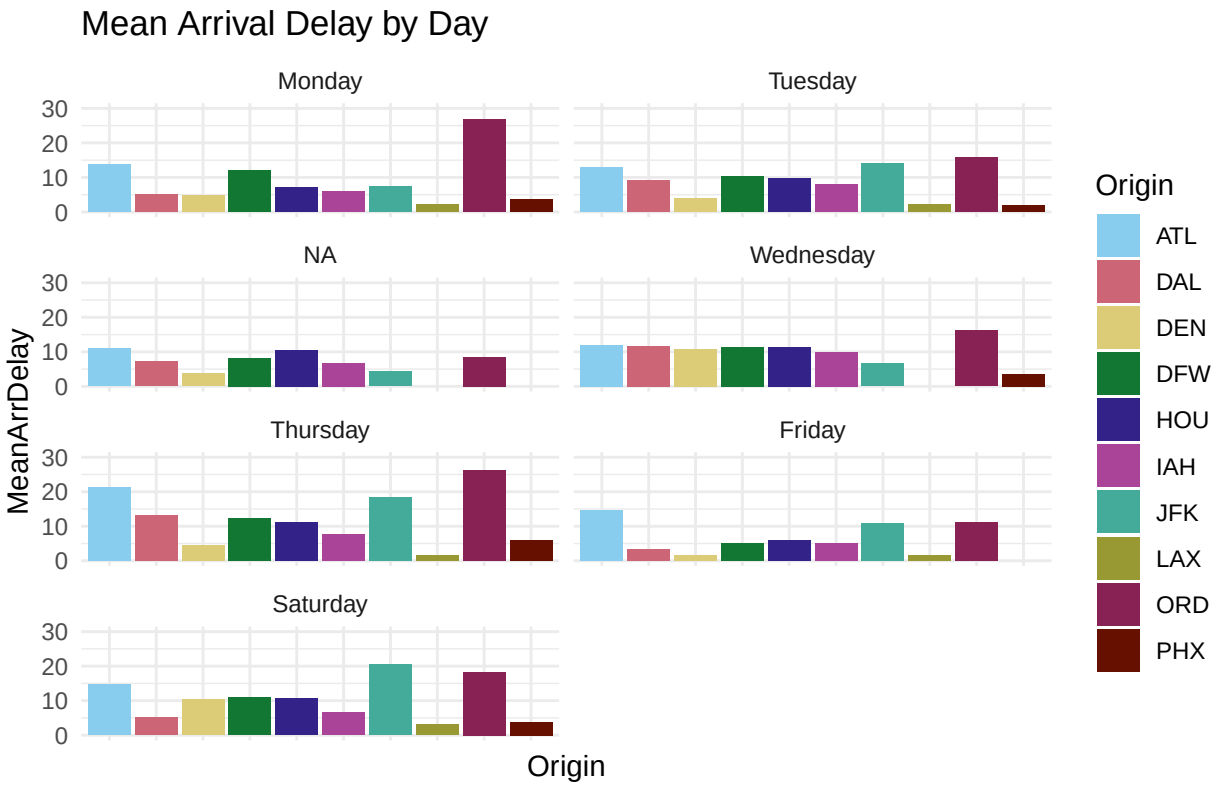


Larger points correspond to larger delays/gains.
Red dots are delays, blue dots are early departures Note the lack of blue dots

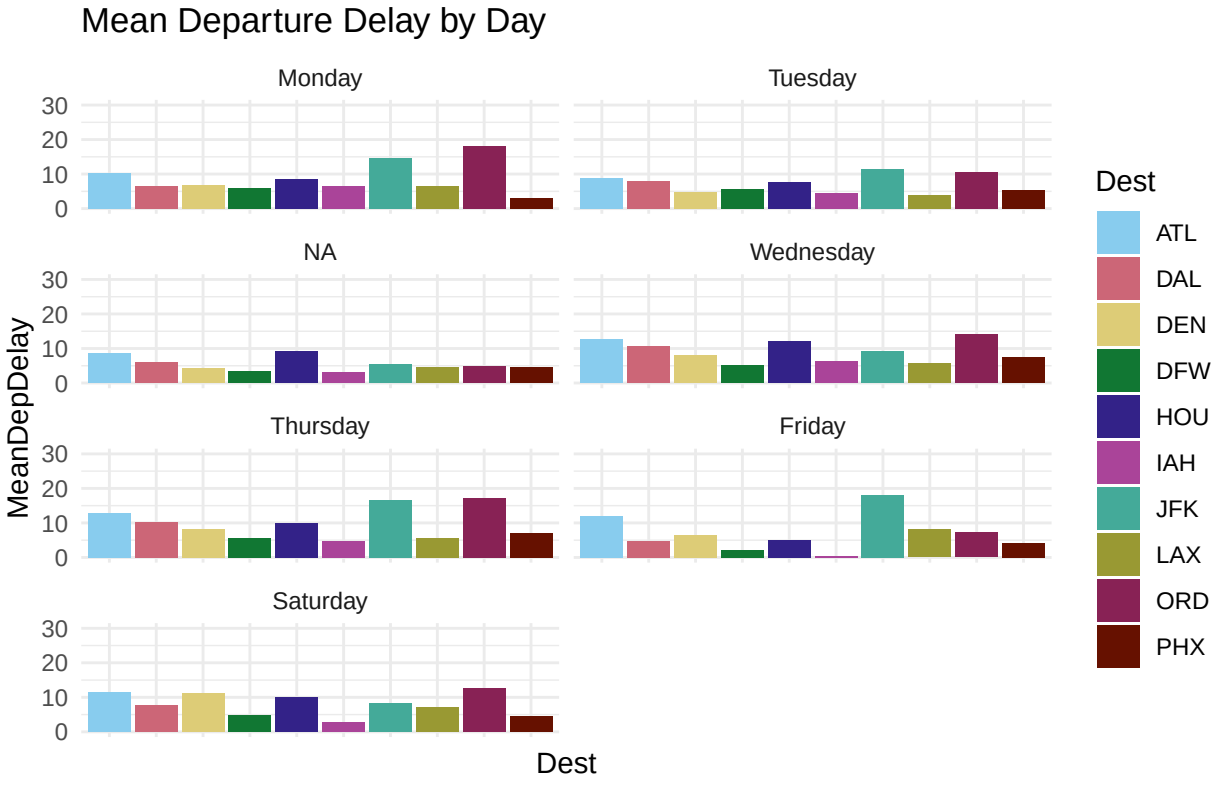
Results Pt. 1

This is a pretty clear indication that the delays based are on a per-airport basis and not a flight-distance basis. There are certainly more formal regressions we could run to examine their significance, but for our purposes here, an eyeball test is more than adequate, if only because this is extremely intuitive and the purpose of this question was more to graph things on a map. For more interesting analysis, we can jump to our second question: Which days are best for each of the largest airports.

Methods and Figures Pt. 2



Mean arrival delay by day
Seperated by origin airport and day of the week



Mean departure delay
Seperated by destination and day of the week

And, in addition to the plots, a we can manually find the best day for each:

```
## # A tibble: 10 x 4
## # Groups:   Origin [10]
##    Origin DayOfWeek Count MeanArrDelay
##    <chr>      <int> <int>        <dbl>
##  1 DFW            6   728        5.12
##  2 IAH            6   408        5.18
##  3 DAL            6   404        3.42
##  4 DEN            6   371        1.61
##  5 ORD            3   363        8.60
##  6 ATL            3   327       11.0
##  7 PHX            6   307       -0.756
##  8 LAX            3   256       -0.727
##  9 HOU            6   212        5.93
## 10 JFK            3   194        4.33
```

```
## # A tibble: 10 x 4
## # Groups:   Dest [10]
##    Dest  DayOfWeek Count MeanDepDelay
##    <chr>     <int> <int>        <dbl>
##  1 DFW           6   721        2.07
##  2 PHX           1   414        2.94
##  3 DAL           6   399        4.76
##  4 IAH           6   395        0.506
##  5 DEN           3   394        4.23
##  6 ORD           3   355        4.74
##  7 ATL           3   329        8.55
##  8 LAX           2   258        4.04
##  9 HOU           6   213        5.11
## 10 JFK           3   195        5.47
```

Results Pt. 2

From the above, we can clearly see the best days to fly into or out of each of these airports. DFW, which is the most trafficked airport for Austin, is best flown from and to on a Saturday whereas Pheonix should be flown from on Saturday, but flown to on a Monday.

Question 2

Part A

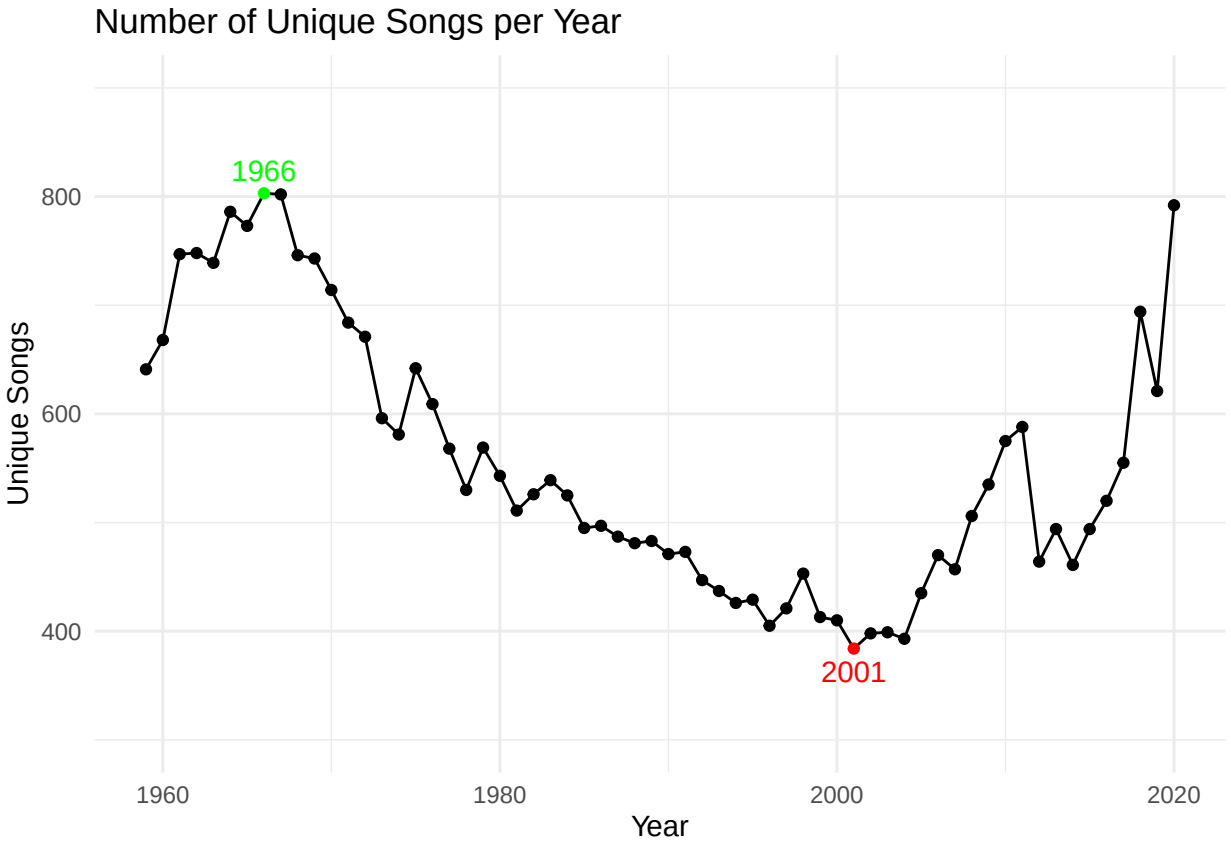
A table of the top 10 songs, measured by weeks in the top 100 chart

```
## # A tibble: 10 x 3
## # Groups:   song [10]
##    song                              performer                          Count
##    <chr>                             <chr>                              <int>
##  1 Radioactive                       Imagine Dragons                       87
##  2 Sail                              AWOLNATION                            79
##  3 Blinding Lights                   The Weeknd                            76
##  4 I'm Yours                         Jason Mraz                            76
##  5 How Do I Live                     LeAnn Rimes                           69
##  6 Counting Stars                    OneRepublic                           68
##  7 Party Rock Anthem                 LMFAO Featuring Lauren Bennett & G~    68
```

```
##  8 Foolish Games/You Were Meant For Me Jewel                          65
##  9 Rolling In The Deep                 Adele                          65
## 10 Before He Cheats                    Carrie Underwood               64
```

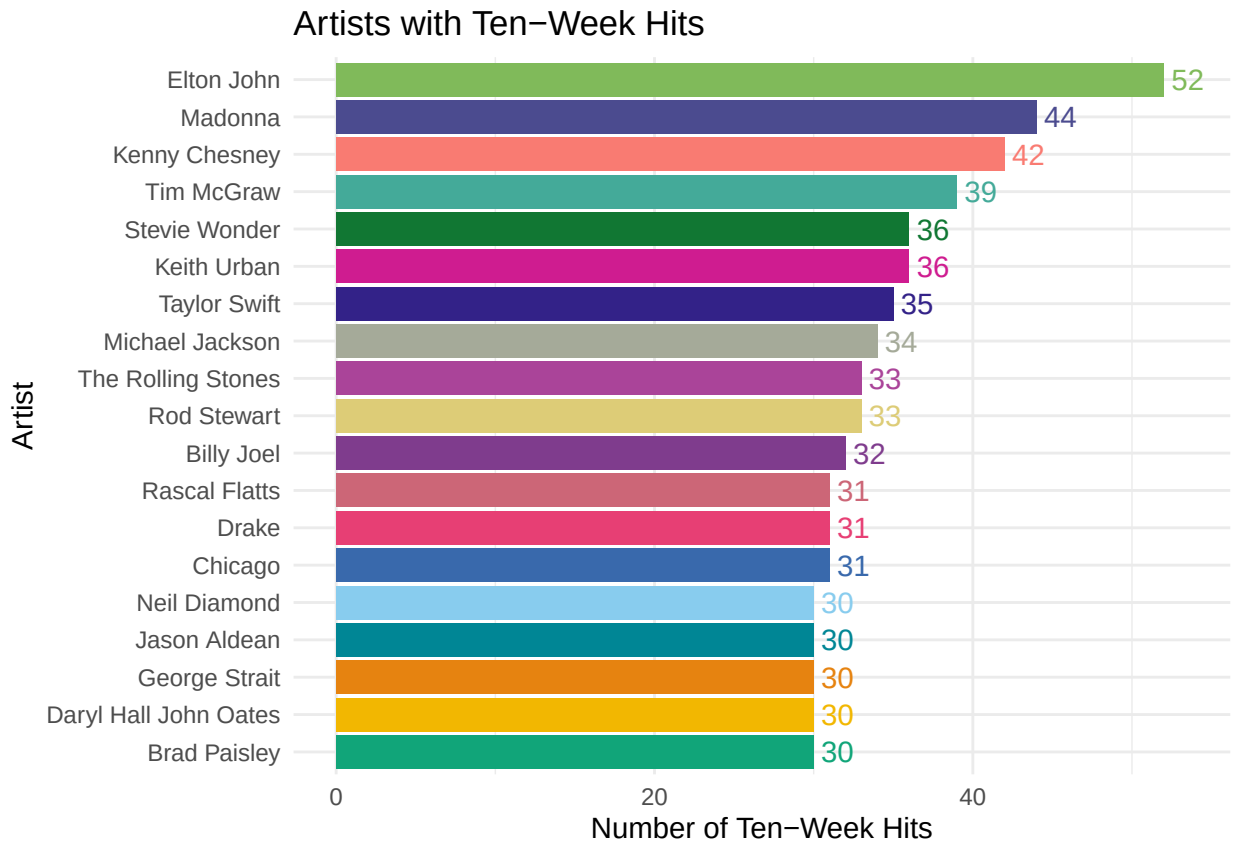
Part B

Diversity by year, measured by the number of unique songs on the in the top 100 each year



Part C

A list of artists with more than 30 songs that spent at least ten weeks in the top 100, listed by the number of 10-



week hits

Question 3

Part A

This question is written in a way that has me confused. I am unsure if we are supposed to get *each* events' 95th percentile or simply for Athletics' events as a whole. In the case of each event individually, it is given by the following table.

```
## # A tibble: 27 x 2
##    event                                    p0.95
##    <chr>                                    <int>
##  1 Athletics Women's 1,500 metres             172
##  2 Athletics Women's 10 kilometres Walk       170
##  3 Athletics Women's 10,000 metres            170
##  4 Athletics Women's 100 metres               182
##  5 Athletics Women's 100 metres Hurdles       178
##  6 Athletics Women's 20 kilometres Walk       173
##  7 Athletics Women's 200 metres               182
##  8 Athletics Women's 3,000 metres             170
##  9 Athletics Women's 3,000 metres Steeplechase 178
## 10 Athletics Women's 4 x 100 metres Relay     182
## # ... with 17 more rows
```

However, if we are looking simply for the 95th percentile of all Athletics medalists, it is produced in the

following (much smaller) table

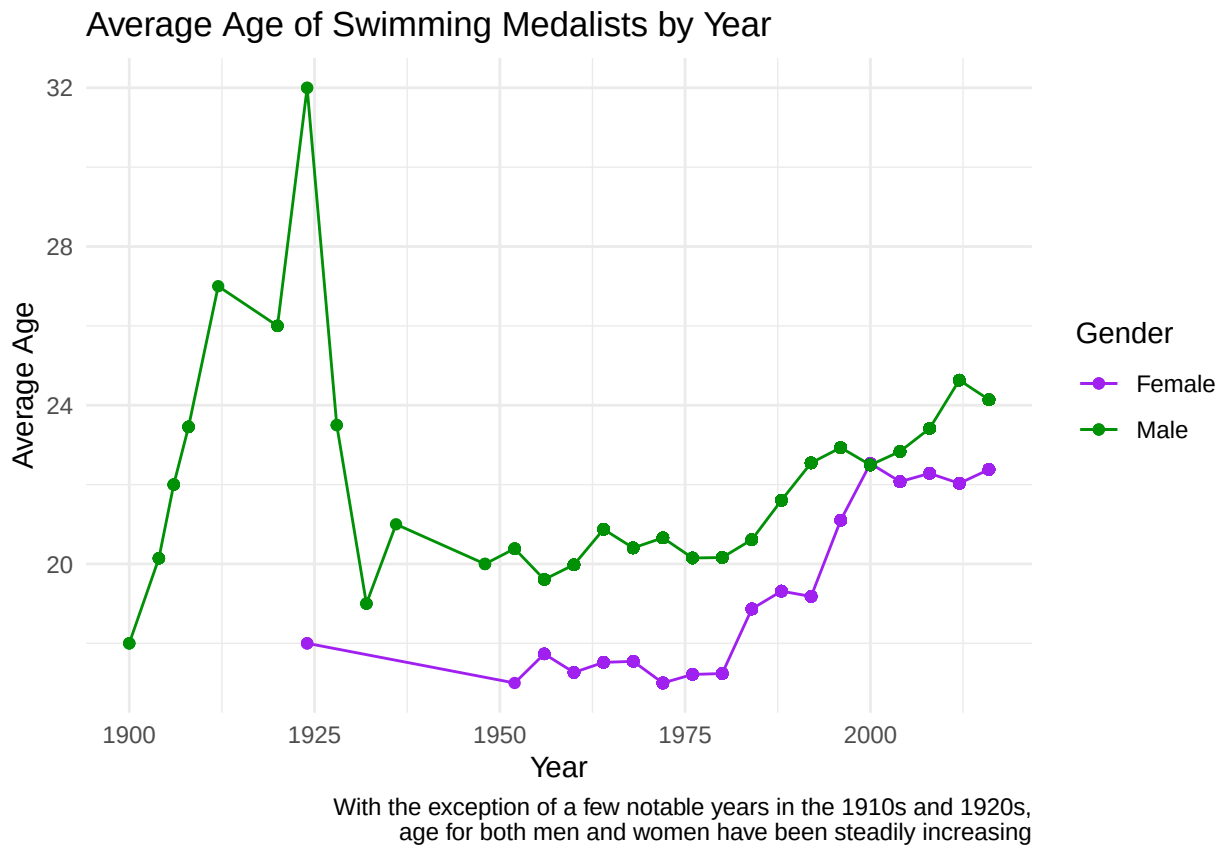
```
## 95%
## 183
```

Part B

The top variation among all events in females competitors heights is given by

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

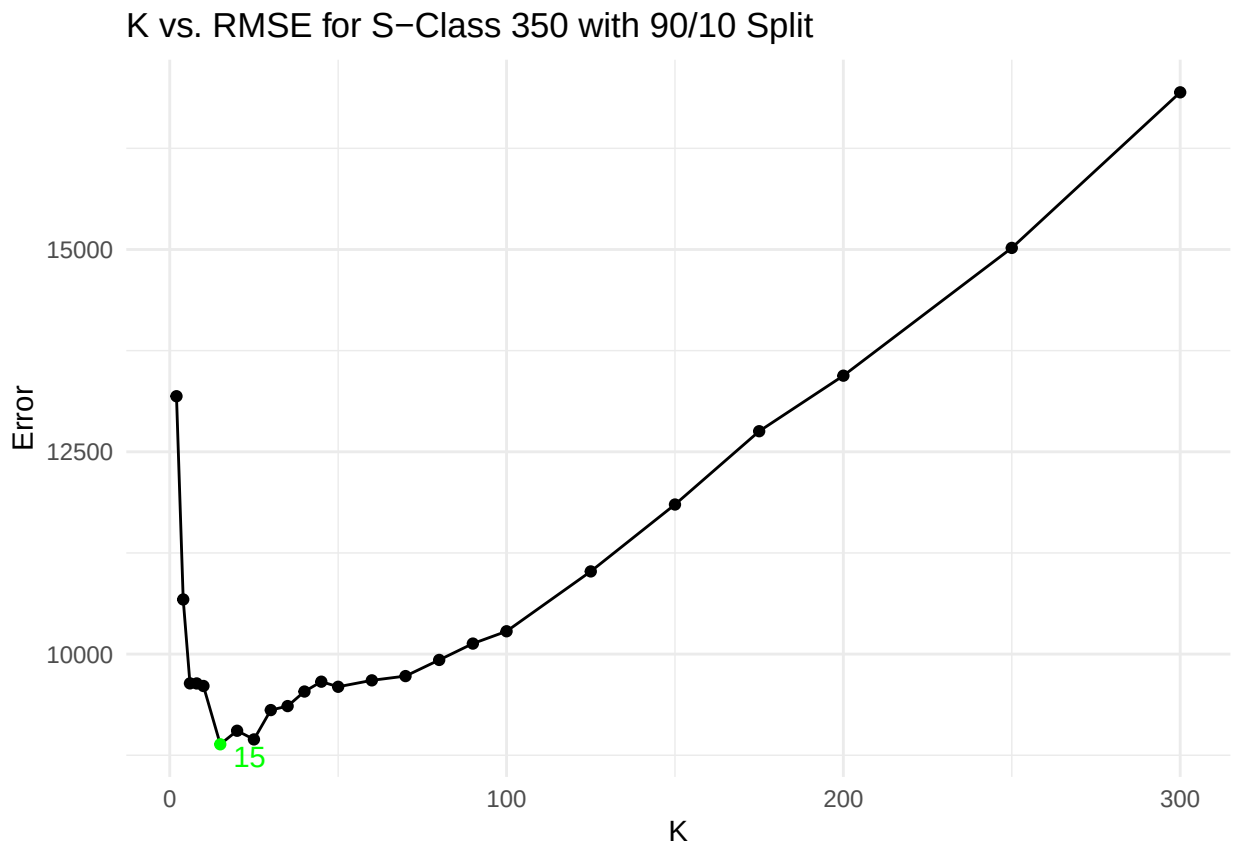
Part C



Question 4

RMSE values for a collection of K values with a 90/10 test-train split. Based on data for the S-Class 350

```
## $Kplot
```

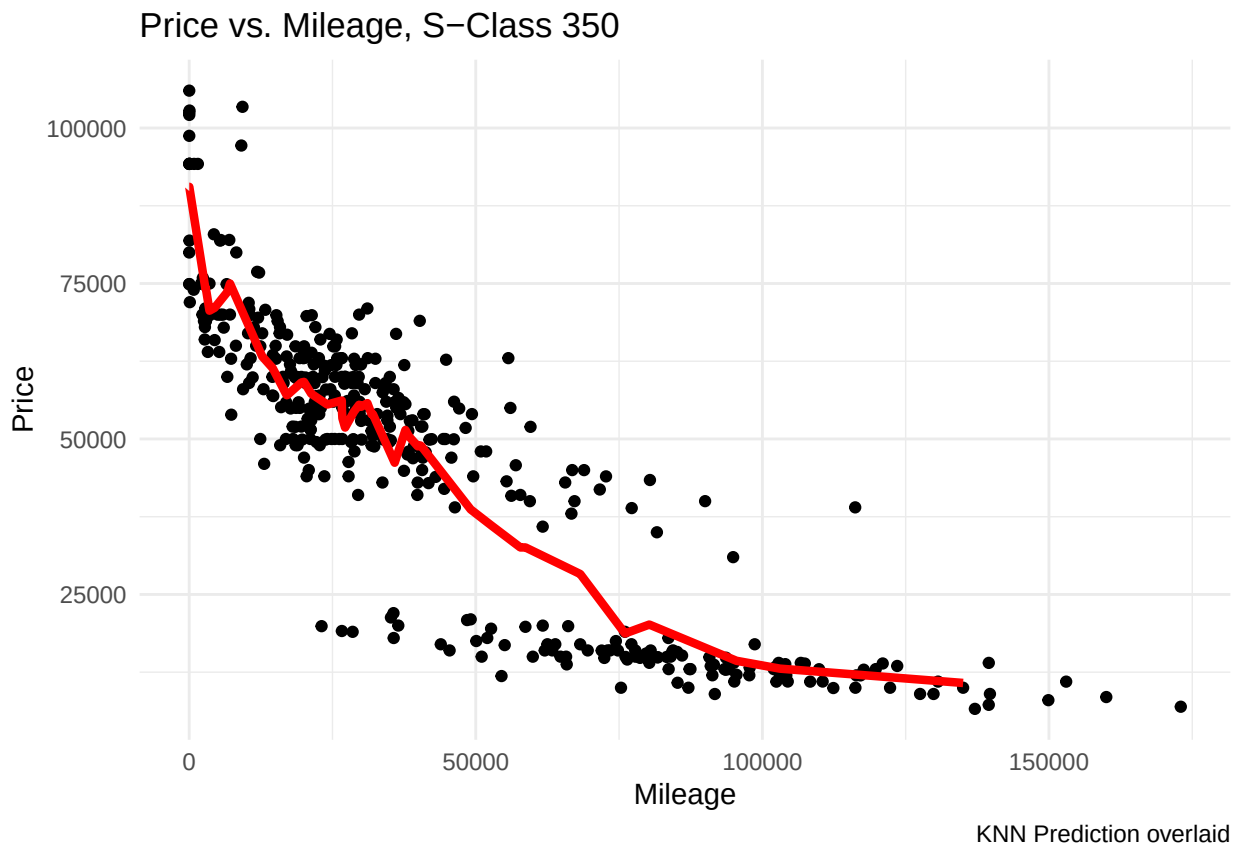


The best-case K, in this case `Sclass_350_Plots["KVal"]`, overlaid on the observed data

```
## [1] "For K of "
```

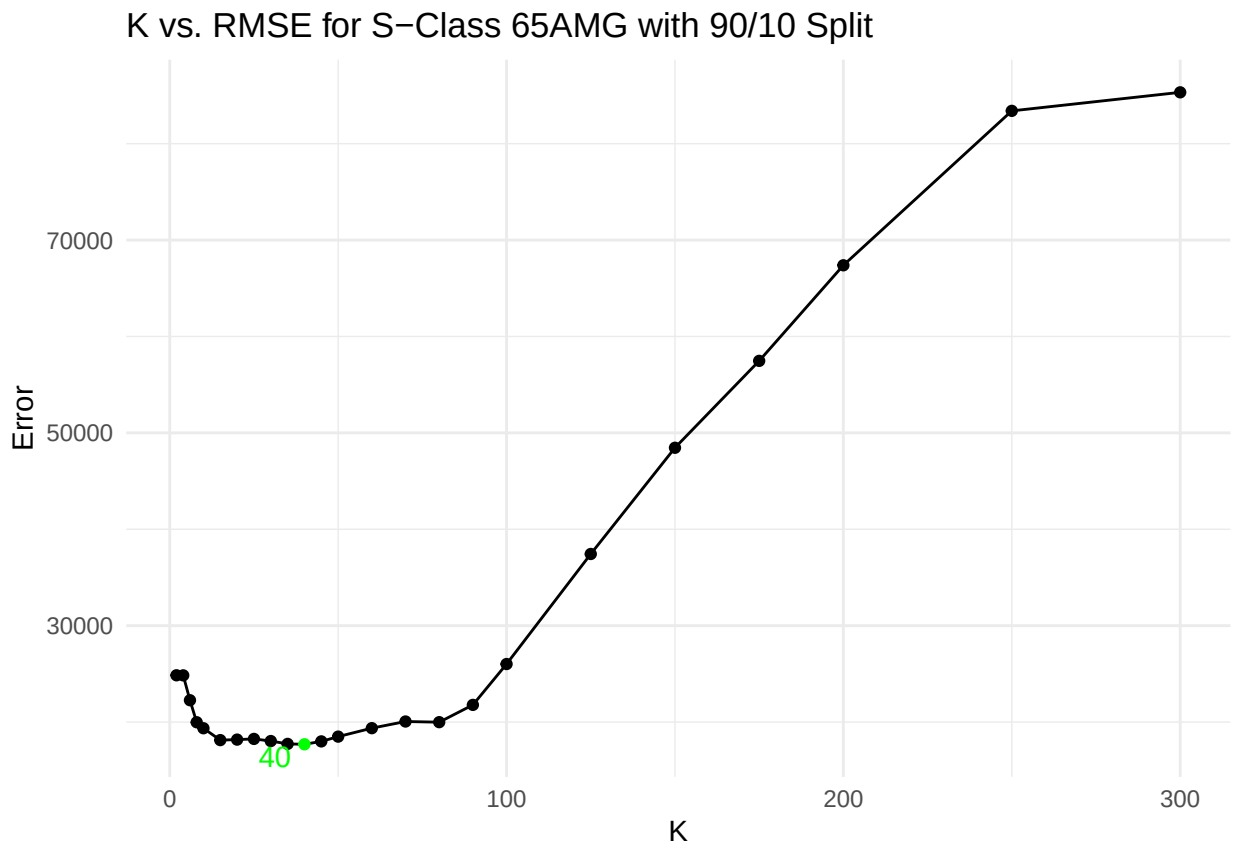
```
## $KVal
## [1] 15
```

```
## $PredictionPlot
```



For the S-Class 65AMG, we can see the errors of each of the same K values as above.

```
## $Kplot
```

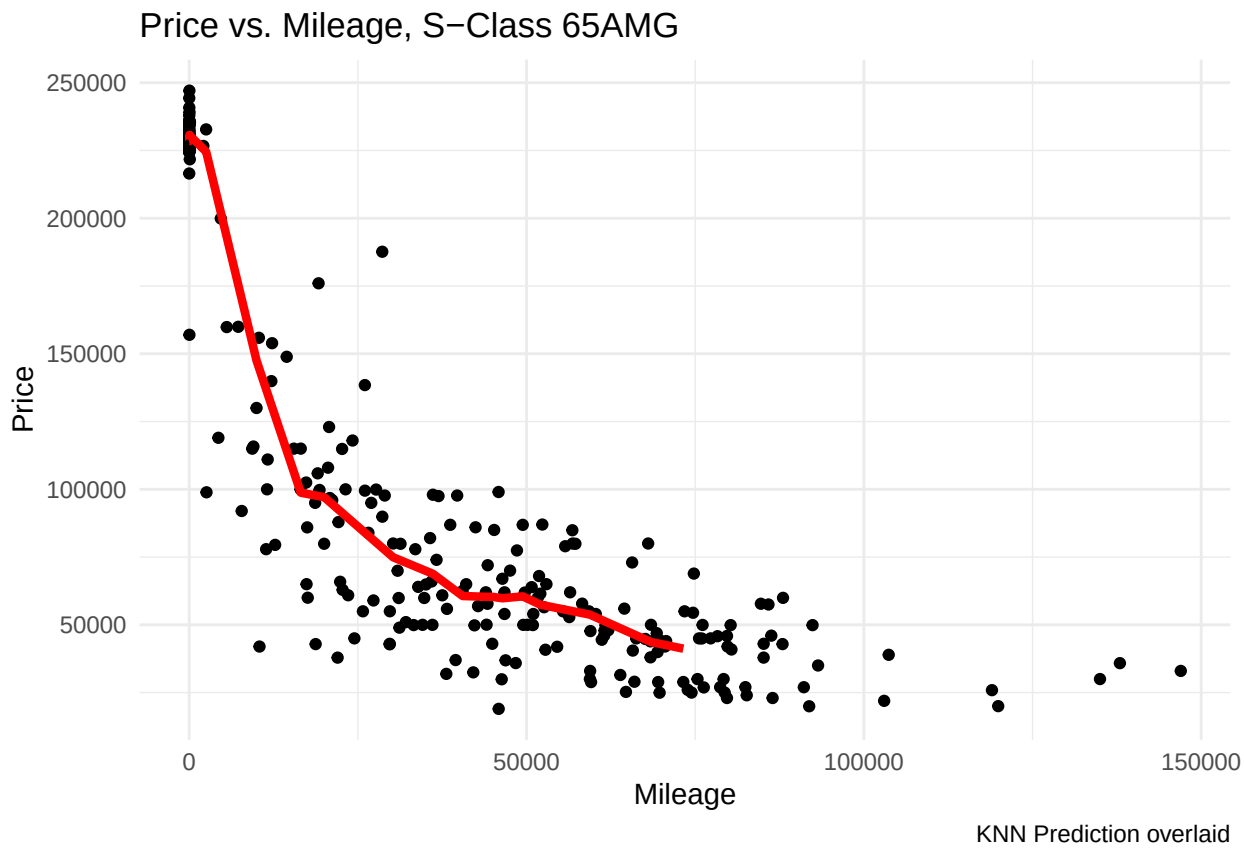


And again, with a best-case K, we can see how it fits the data

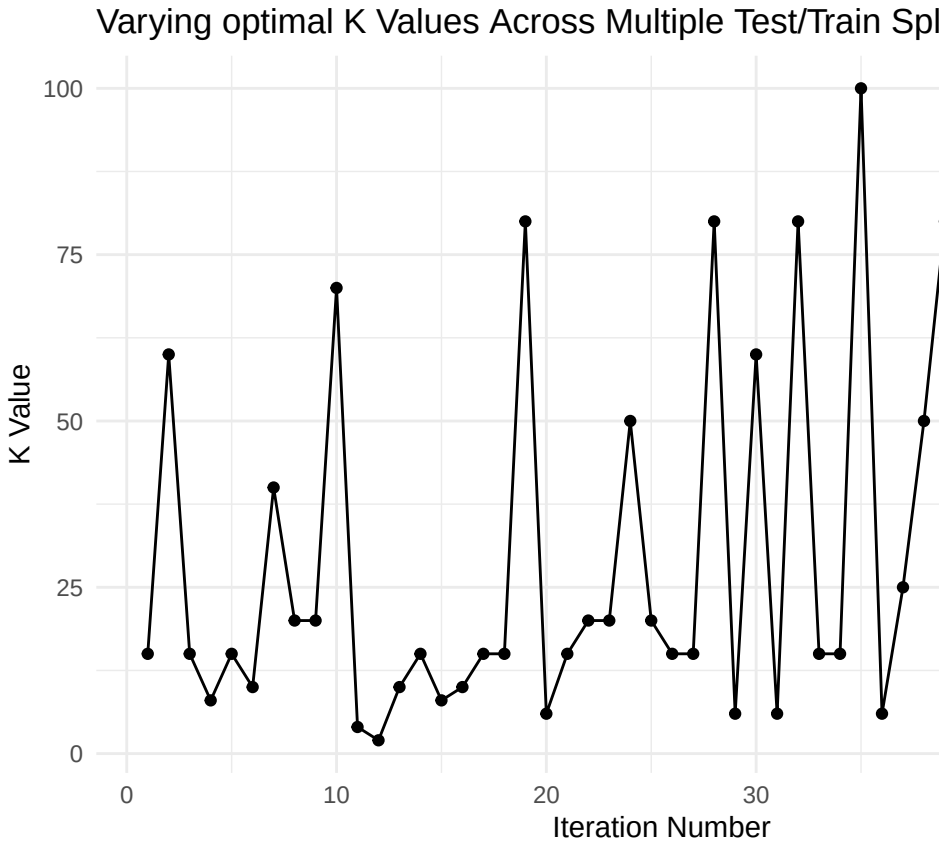
```
## [1] "For K of "
```

```
## $KVal
## [1] 40
```

```
## $PredictionPlot
```



It is worth noting that, in this case, the above values are not at all stable. Below, we can see the variation across



a collection of tests on the S-Class 350 data.