

Problem 1: Data visualization: flights at ABIA

1.The total airplanes depar from Austin in 2008

Show the total airplanes of each airline:

Table 1

```
## # A tibble: 16 × 2
##   UniqueCarrier count
##   <chr>          <int>
## 1 WN            17438
## 2 AA             9997
## 3 CO             4614
## 4 YV             2496
## 5 B6             2400
## 6 XE             2307
## 7 OO             2007
## 8 OH             1491
## 9 MQ             1331
## 10 9E            1273
## 11 DL            1067
## 12 F9            1066
## 13 UA             933
## 14 US             729
## 15 EV             413
## 16 NW             61
```

Use the map to show top10 destination from Austin

```
##   Dest total                coordinate      lon
## 1   ATL  2252      33.6367, -84.428101      33.6367
## 2   AUS 49637  30.194499969482422, -97.6698989868164 30.194499969482422
## 3   DAL  5573      32.847099, -96.851799      32.847099
## 4   DEN  2673      39.861698150635, -104.672996521    39.861698150635
## 5   DFW  5506      32.896801, -97.038002      32.896801
## 6   HOU  2319      29.64539909, -95.27890015      29.64539909
## 7   IAH  3691  29.984399795532227, -95.34140014648438 29.984399795532227
## 8   LAX  1733      33.94250107, -118.4079971      33.94250107
## 9   ORD  2514      41.9786, -87.9048      41.9786
## 10  PHX  2783  33.43429946899414, -112.01200103759766 33.43429946899414
##
##           lat
## 1      -84.428101
## 2    -97.6698989868164
## 3      -96.851799
## 4    -104.672996521
## 5      -97.038002
```

```
## 6      -95.27890015
## 7     -95.34140014648438
## 8     -118.4079971
## 9      -87.9048
## 10   -112.01200103759766
```

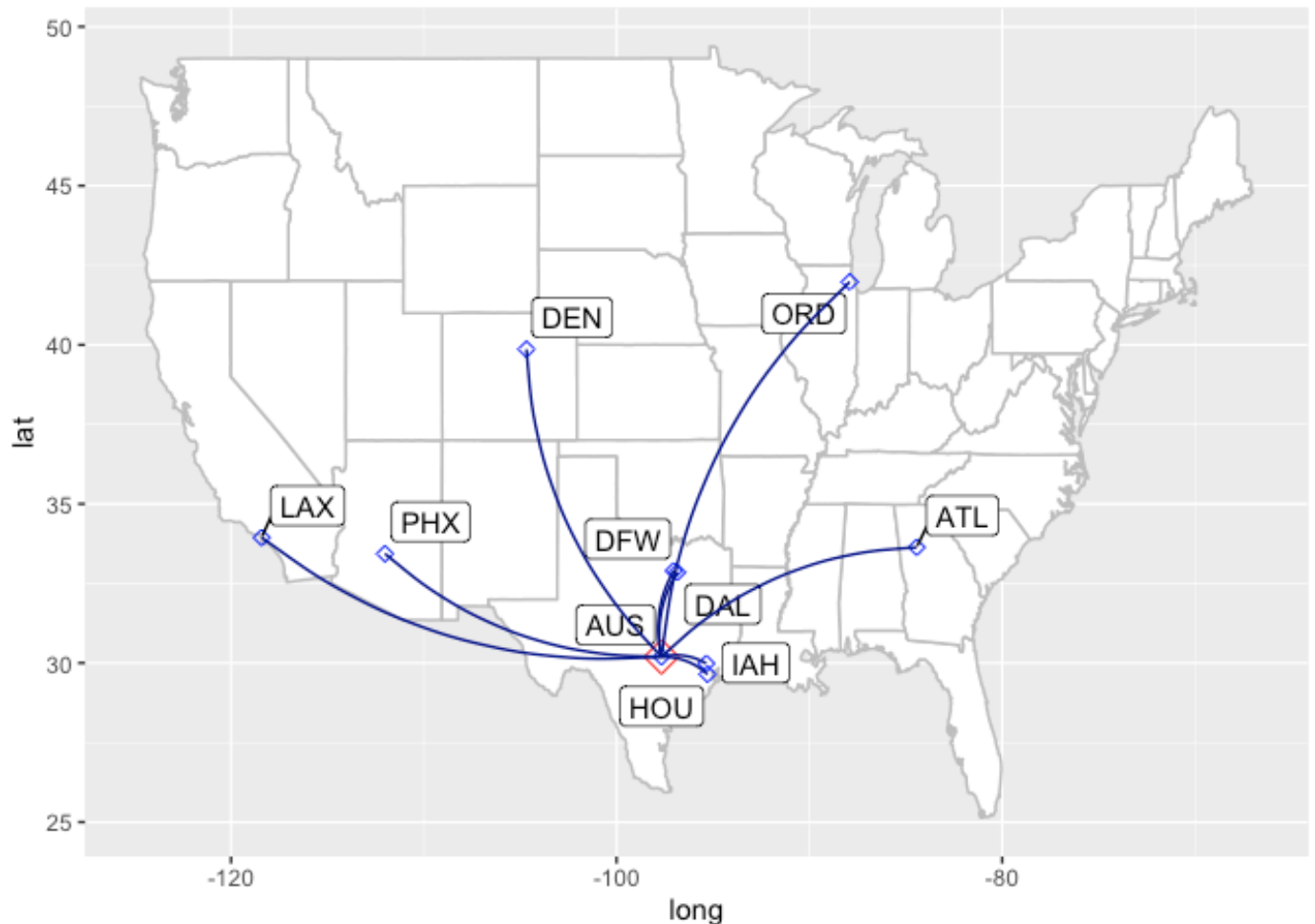


Figure 1: Top10 Destination from Austin

Summarize the total airplanes departing from Austin:

```
## # A tibble: 1 × 1
##   total
##   <int>
## 1 49623
```

In 2008, there were total 49623 airplanes departing from Austin. Among them, Southwest Airlines (WN) had the most departing flights.

2. Show the delay information of the departure of airplane for one week

Creat a new variable to find whether the departure of airplane is delay:

(1) Departure delay rate in the week

Take a look at which day has the highest delay rate in the week:

Table 2

```
## # A tibble: 7 × 4
##   DayOfWeek count num_delay rate_delay
##   <int> <int>     <dbl>     <dbl>
## 1      1    7299     2782     0.381
## 2      2    7265     2373     0.327
## 3      3    7294     2488     0.341
## 4      4    7274     2904     0.399
## 5      5    7270     3000     0.413
## 6      6    5618     1769     0.315
## 7      7    6871     2442     0.355
```

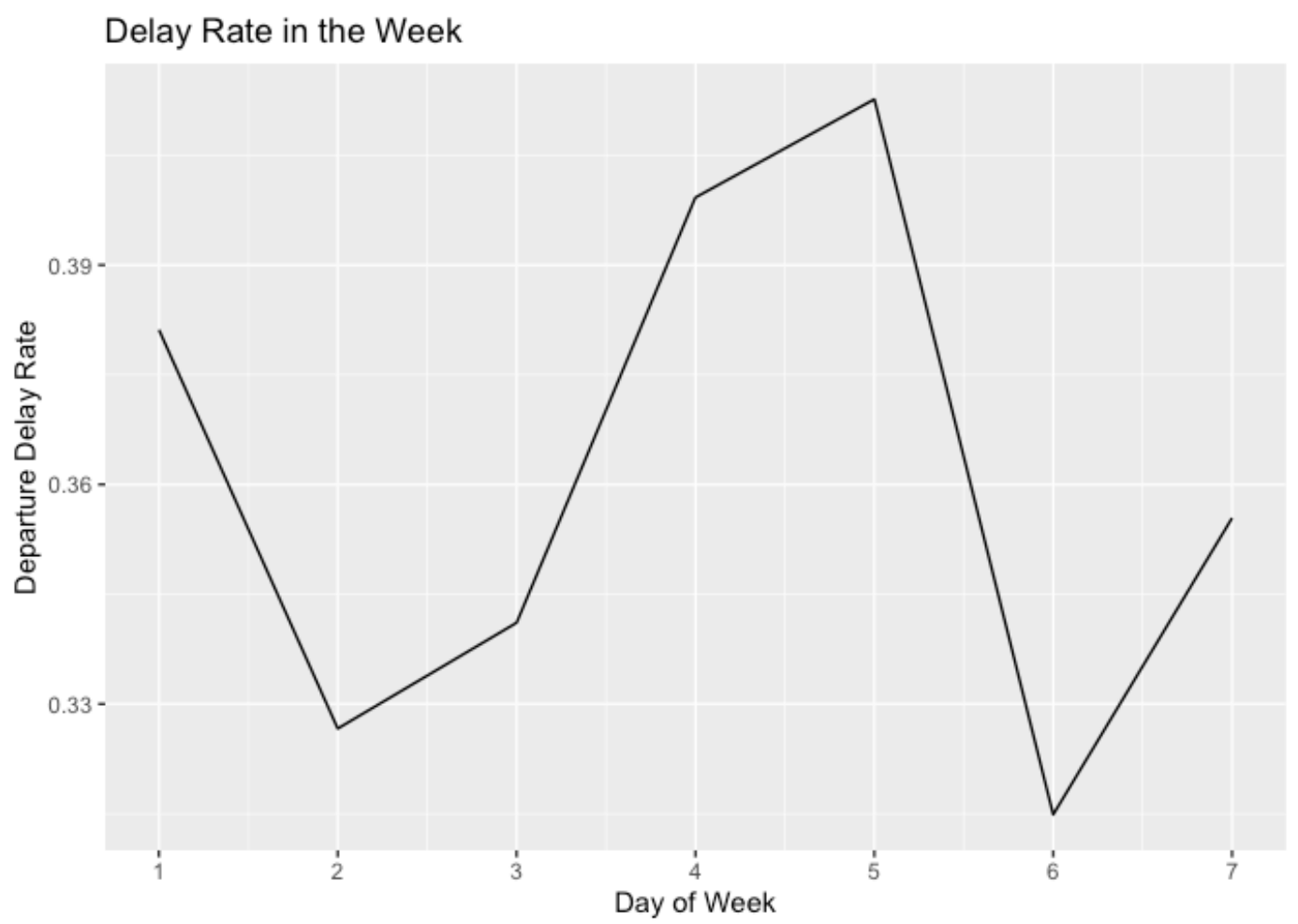


Figure 2: The Departure Delay Rate in the Week

From the line graphs, it shows that in one week AUS had the lowest departure delay rate on Tuesday and Saturday, and the highest departure delay rate on Friday.

(2) The departure delay rates of airlines in total

Take a look at which airline has the highest delay rate in total:

Table 3

```
## # A tibble: 16 × 4
```

##	UniqueCarrier	count	num_delay	rate_delay
##	<chr>	<int>	<dbl>	<dbl>
## 1	9E	1245	251	0.202
## 2	AA	9709	2805	0.289
## 3	B6	2367	757	0.320
## 4	CO	4554	1357	0.298
## 5	DL	1056	384	0.364
## 6	EV	407	176	0.432
## 7	F9	1064	279	0.262
## 8	MQ	1245	390	0.313
## 9	NW	61	19	0.311
## 10	OH	1463	482	0.329
## 11	OO	1976	570	0.288
## 12	UA	923	255	0.276
## 13	US	727	136	0.187
## 14	WN	17343	8621	0.497
## 15	XE	2296	762	0.332
## 16	YV	2455	514	0.209

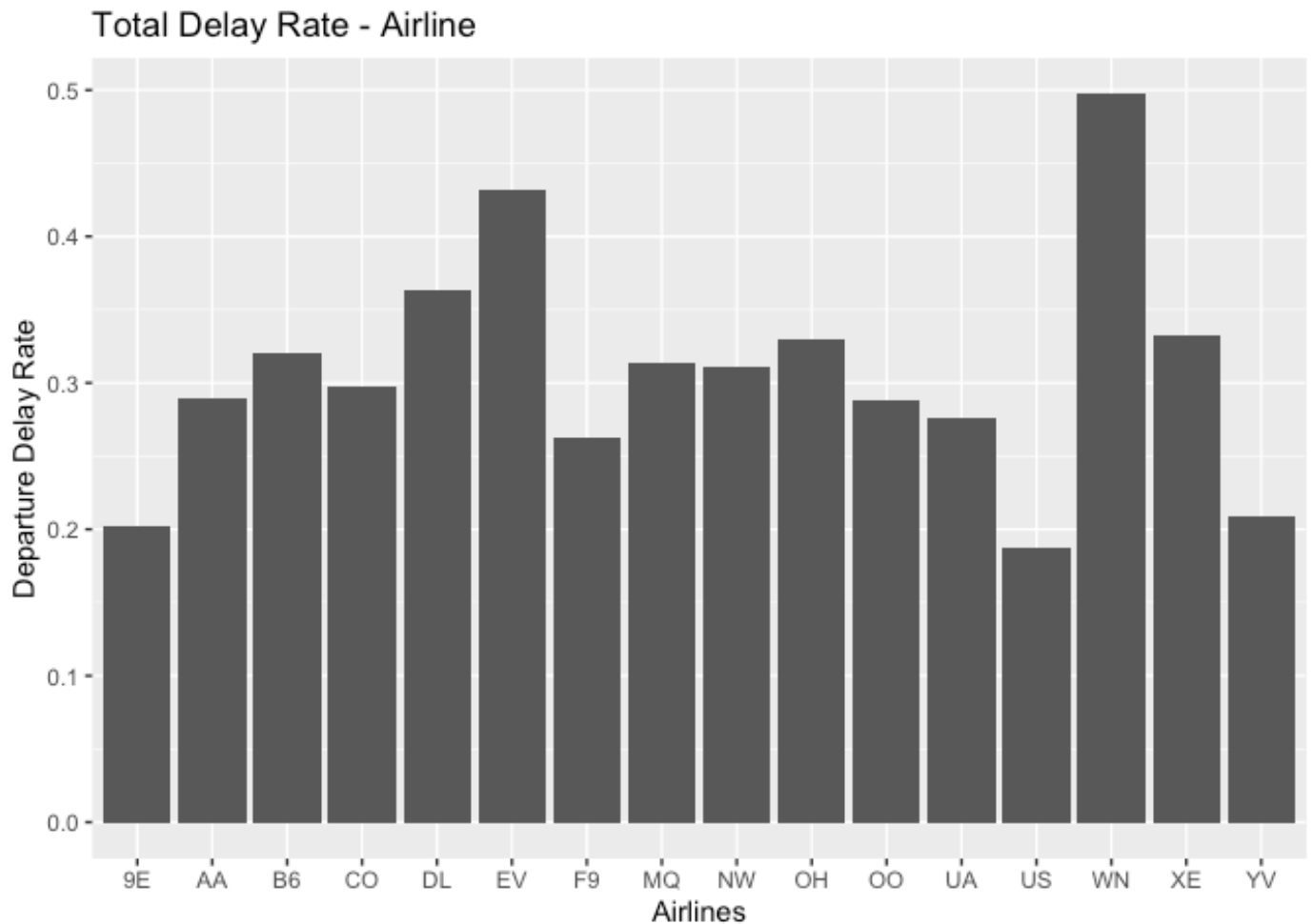


Figure 3: The Total Delay Rate of Airlines

From the barplot, we can see that Southwest Airlines (WN) and ExpressJet (EV) had the highest departure delay rate, US Airways (US) has the lowest departure delay rate.

(3) The departure delay rates of airlines in the week

Take a look at which day did the airline has the highest delay rate in the week:

Table 4

```
## # A tibble: 112 × 5
## # Groups:   DayOfWeek [7]
##   DayOfWeek UniqueCarrier count num_delay rate_delay
##   <int> <chr> <int> <dbl> <dbl>
## 1 1 9E 184 38 0.207
## 2 1 AA 1433 450 0.314
## 3 1 B6 343 122 0.356
## 4 1 CO 716 233 0.325
## 5 1 DL 164 60 0.366
## 6 1 EV 55 29 0.527
## 7 1 F9 156 44 0.282
## 8 1 MQ 192 67 0.349
## 9 1 NW 6 0 0
## 10 1 OH 217 83 0.382
```

... with 102 more rows

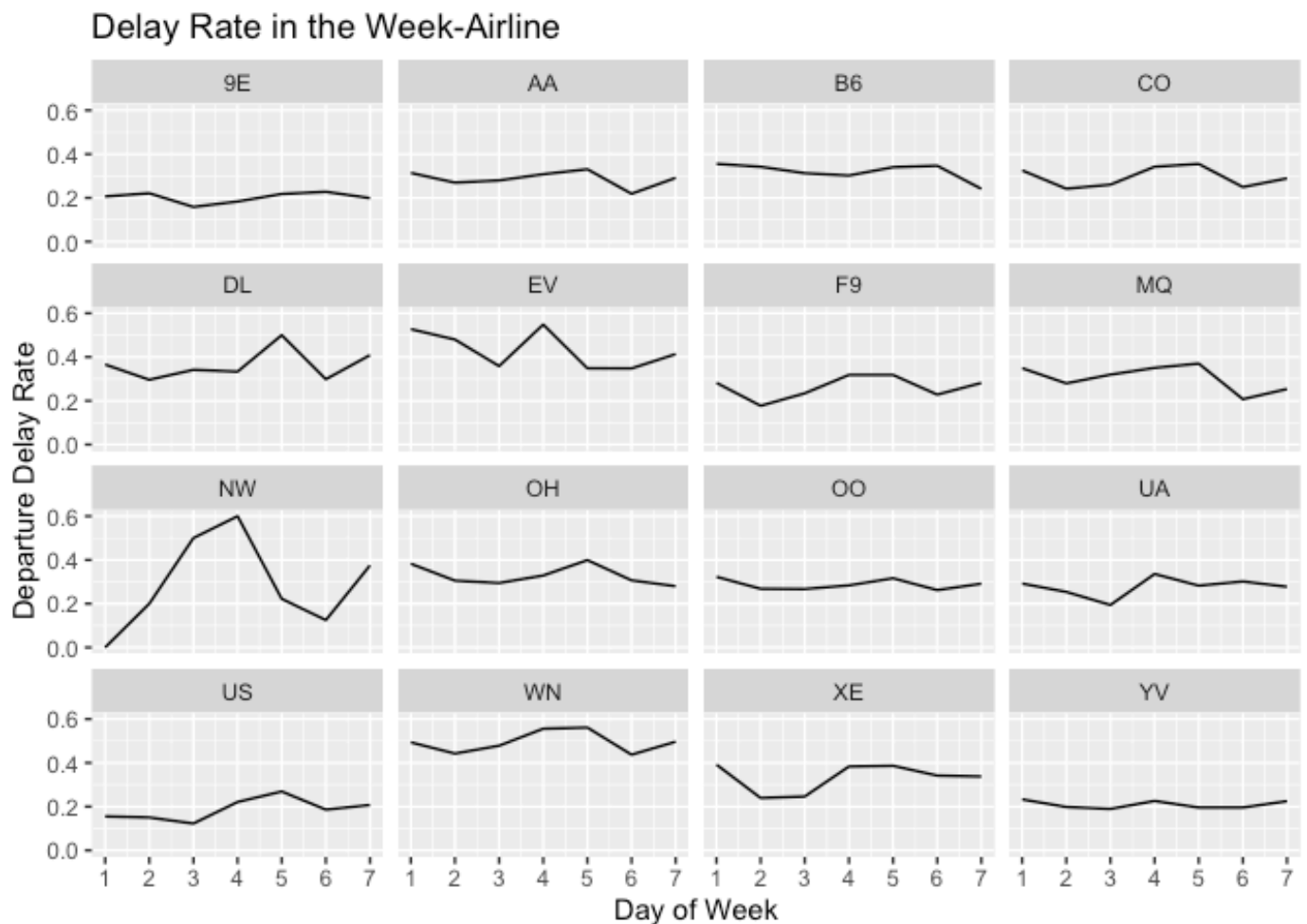


Figure 4: The Delay Rate of Airlines in the Week

From the line graph, it shows that in AUS, Southwest (WN), the airline with the highest number of flights, had the lowest departure delays on Tuesdays and Saturdays in 2008.

In conclusion, in AUS, to avoid departure delays, you should try to avoid traveling on Friday and choosing Southwest Airlines. ExpressJet or choosing traveling on Tuesdays or Saturdays are good choices.

3.Show the delay information of the departure of airplane in a day

Delay Rate and Average_Delay_Time:

Make the chart of average departure delay:

Average departure delay

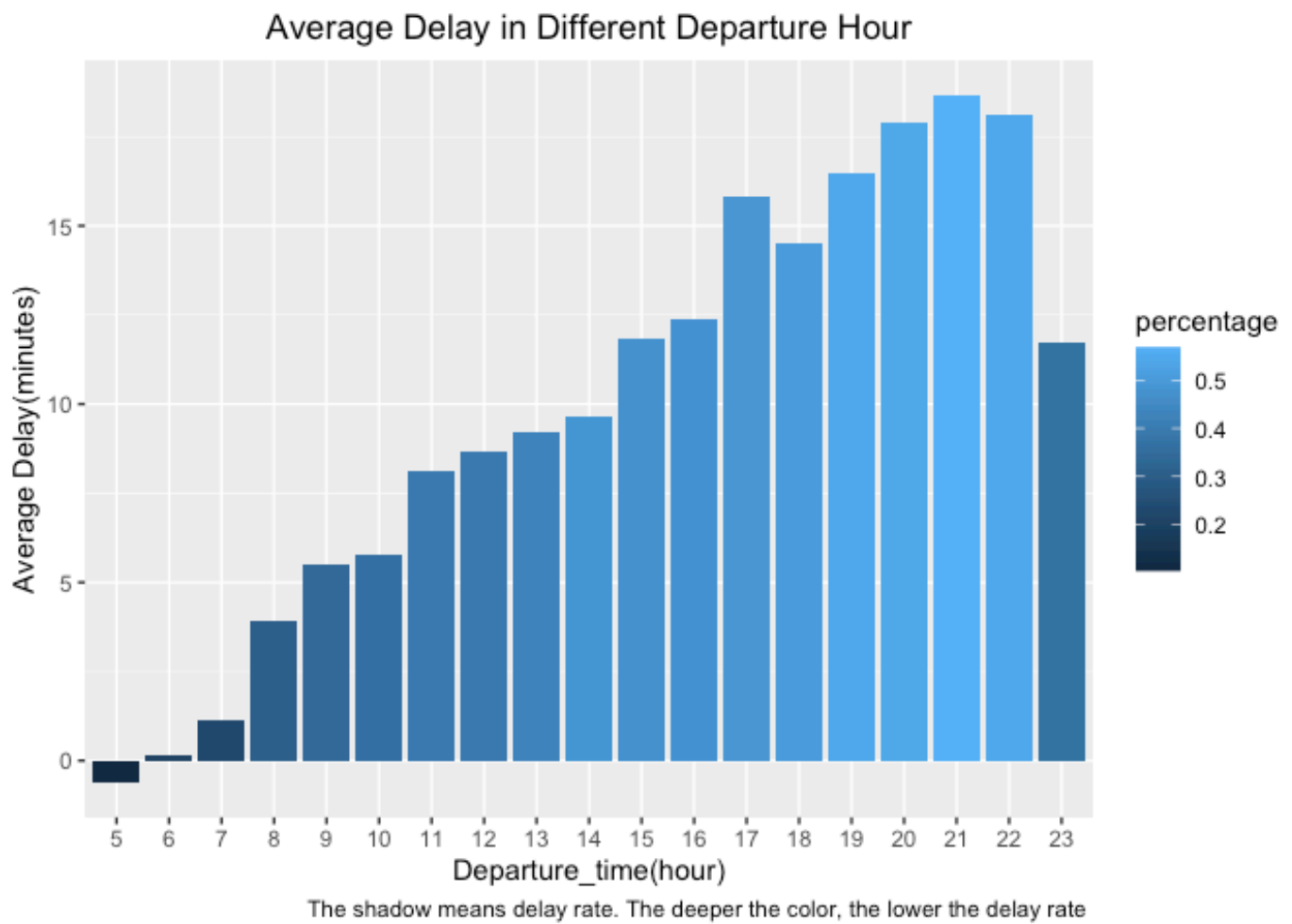
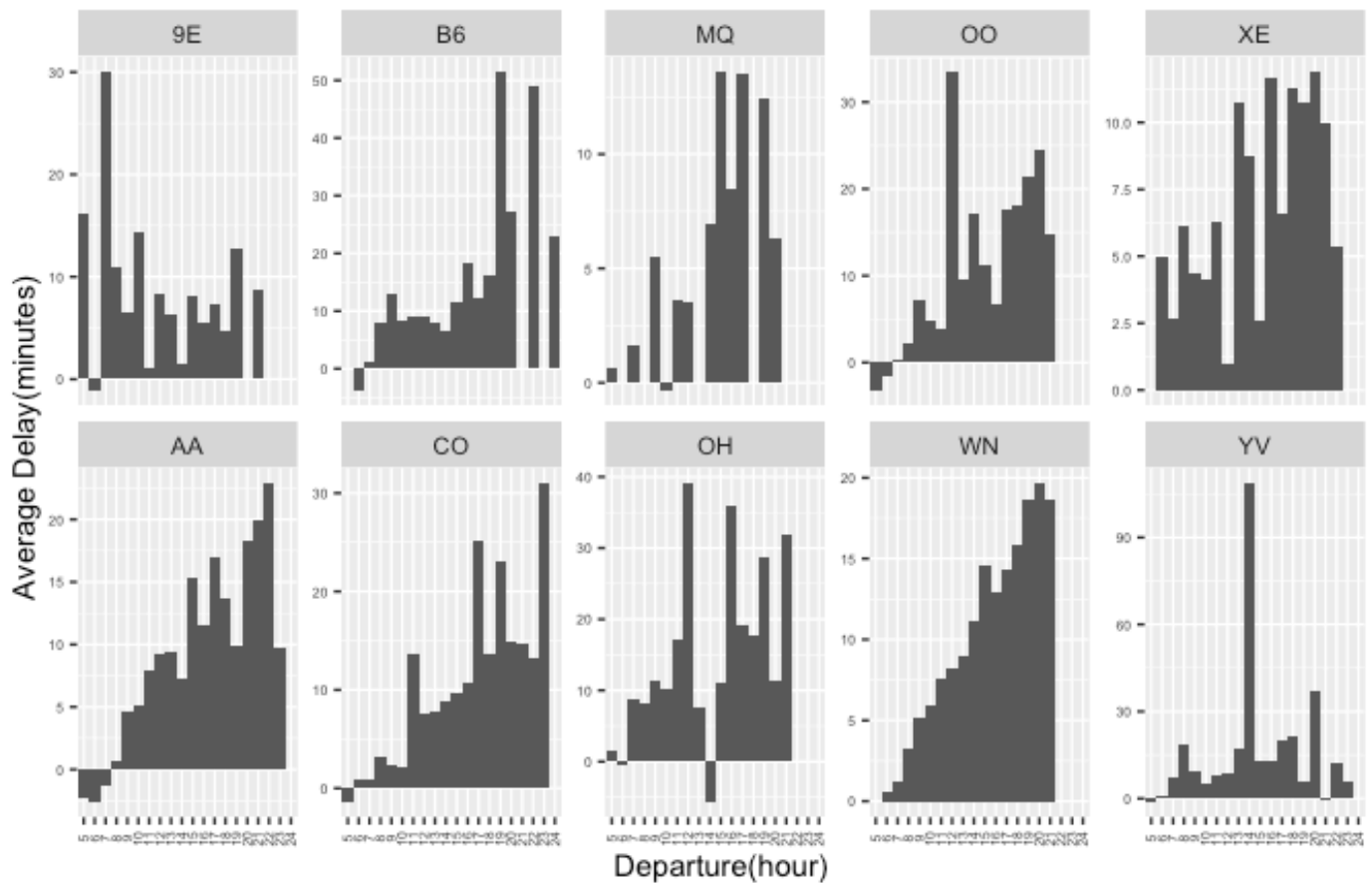


Figure 5

From the bar plot, the departure delay rate and the average delay time is lowest in the morning. Evening had the highest delay rate and delay time.

Take a look at the departure delay hour in one day around the 10 airlines that have the most airplanes.

Average Delay in Different Departure Hour



The data is top 10 airline

Figure 6

From the bar plot, Southwest Airlines (WN), the airline with the most airplanes, had the most departure delay rate in the evening.

In conclusion, the best departure time to optimize your flight plan from Austin-Bergstrom International Airport would be from 5:00 to 10:00. In this time slot, the delay rate is lower than 30%, much lower than that in other time slot. Also, the amount of flight is sufficient. Besides, as to top-10 airlines, especially top 3 airlines (WN, AA and OO), the best time to get rid of departure delay is from 5:00 to 10:00 within 5 mins delay in average.

4. Show the delay information of the arrival of airline in a day

Find the best time of the day through arrival delay rate

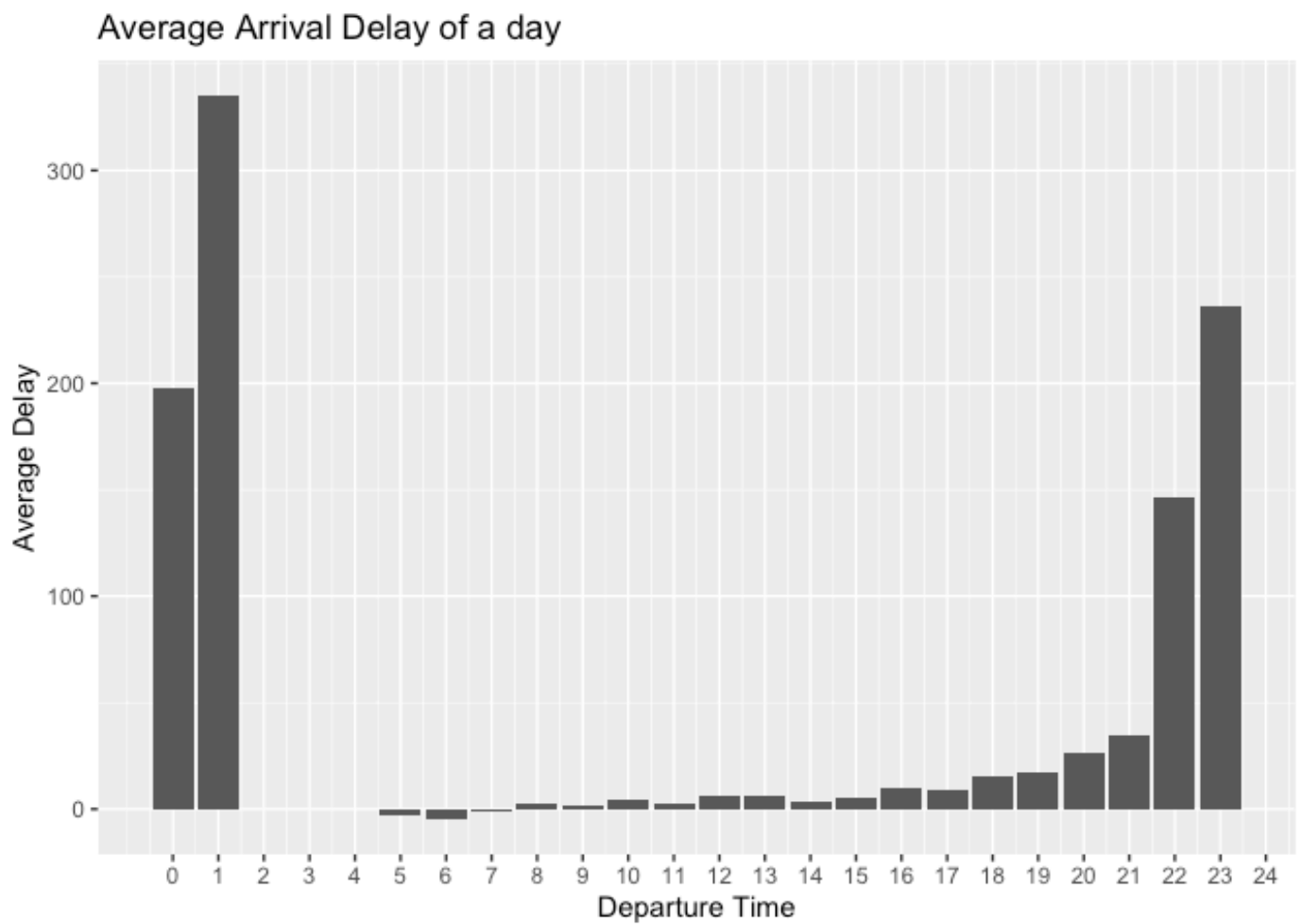


Figure 7

From the bar plot, we can see that the arrival delay rate in the morning (5am-11pm) is relatively less, and even has the chance to take off early.

Average Arrival Delay of a day across Airlines

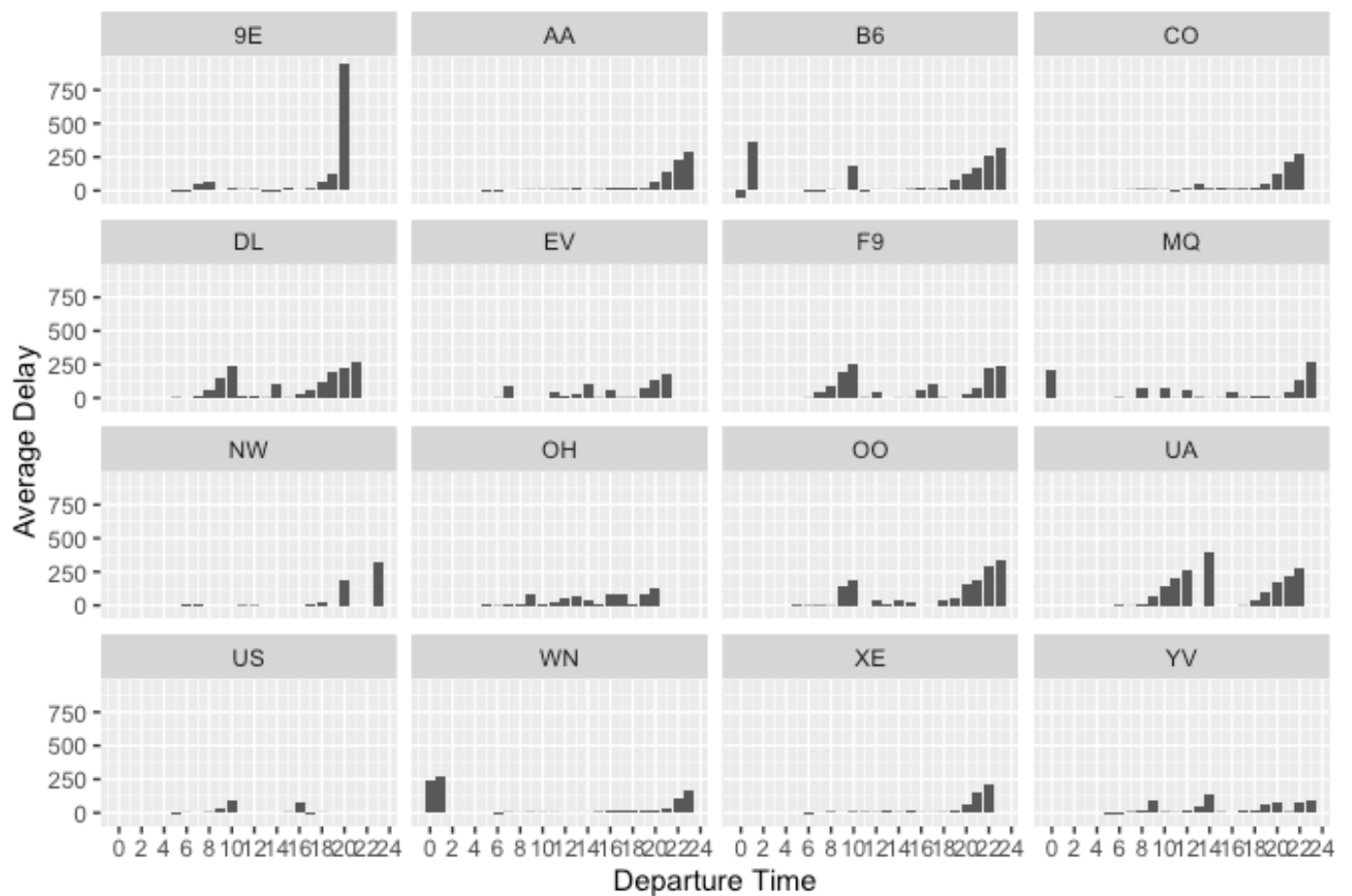


Figure 8

From the bar plot, WN, with the most departing flight, the arrival delay rate at 6am-14pm is relatively less. This time period is a good choice.

Take a look at the arrival delay rate of the popular airport in total:

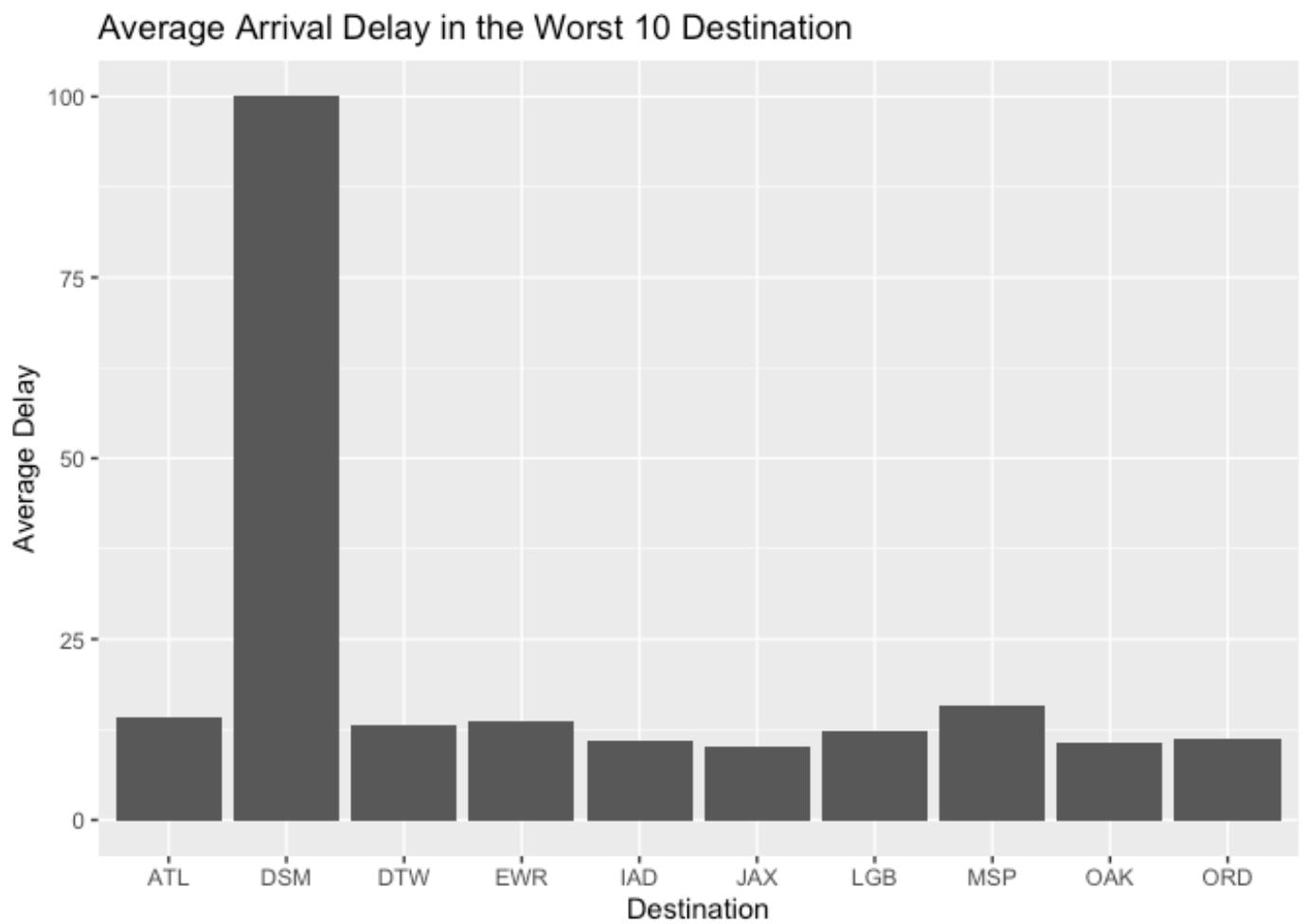


Figure 9

From the bar plot, we can see that Des Moines Intl Airport (DSM) had the highest delay rate as the arrival airport in the whole year.

Average Arrival Delay in the Worst 10 Destination over Month

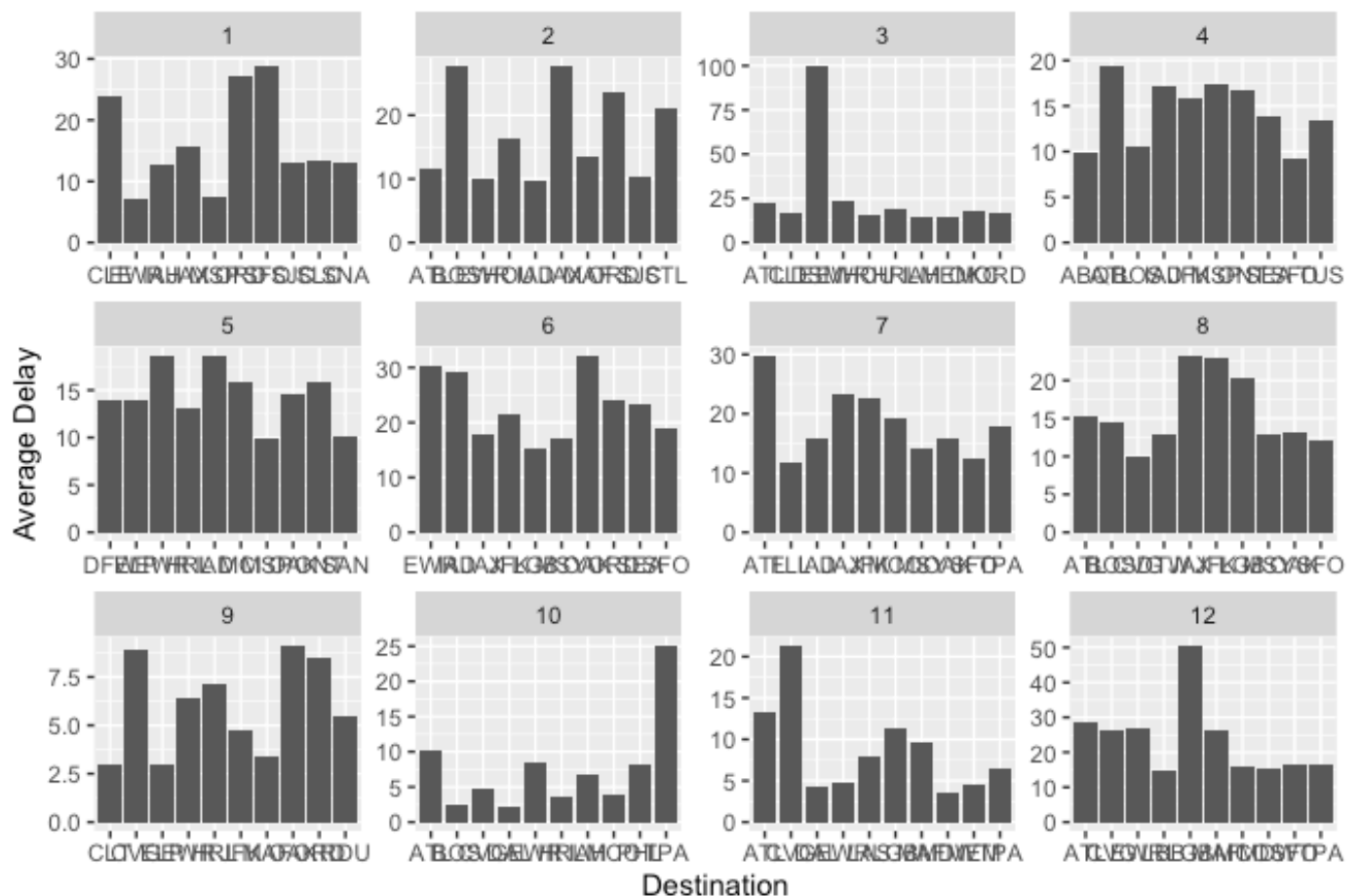


Figure 10

From the bar plot, we can see that September has the lowest arrival delay rate around the whole year. Arrival delays are relatively similar across popular airports in every month except March and October.

In conclusion, the best time of day to fly to minimize delays is the morning, and this basically doesn't change by airline. For customers who want to take off early, they may choose departure at 5 or 6 am. For the bad airports, the results do change over months, but the gap between top 10 bad airports each month is not significant, excluding March and October.

Problem 2: Wrangling the Billboard Top 100

Part A

Show the top 10 most popular songs since 1958:

Table 1

```
## # A tibble: 10 × 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

Show the number of unique songs that appeared in the Billboard Top 100 on given year:

Table 2

```
## # A tibble: 62 × 2
##   year count
##   <int> <int>
## 1  1959   663
## 2  1960   700
## 3  1961   779
## 4  1962   768
## 5  1963   754
## 6  1964   811
## 7  1965   800
## 8  1966   832
## 9  1967   827
## 10 1968   772
## # ... with 52 more rows
```

Use the line graph of unique songs to show the “musical diversity”

Musical Diversity over Year

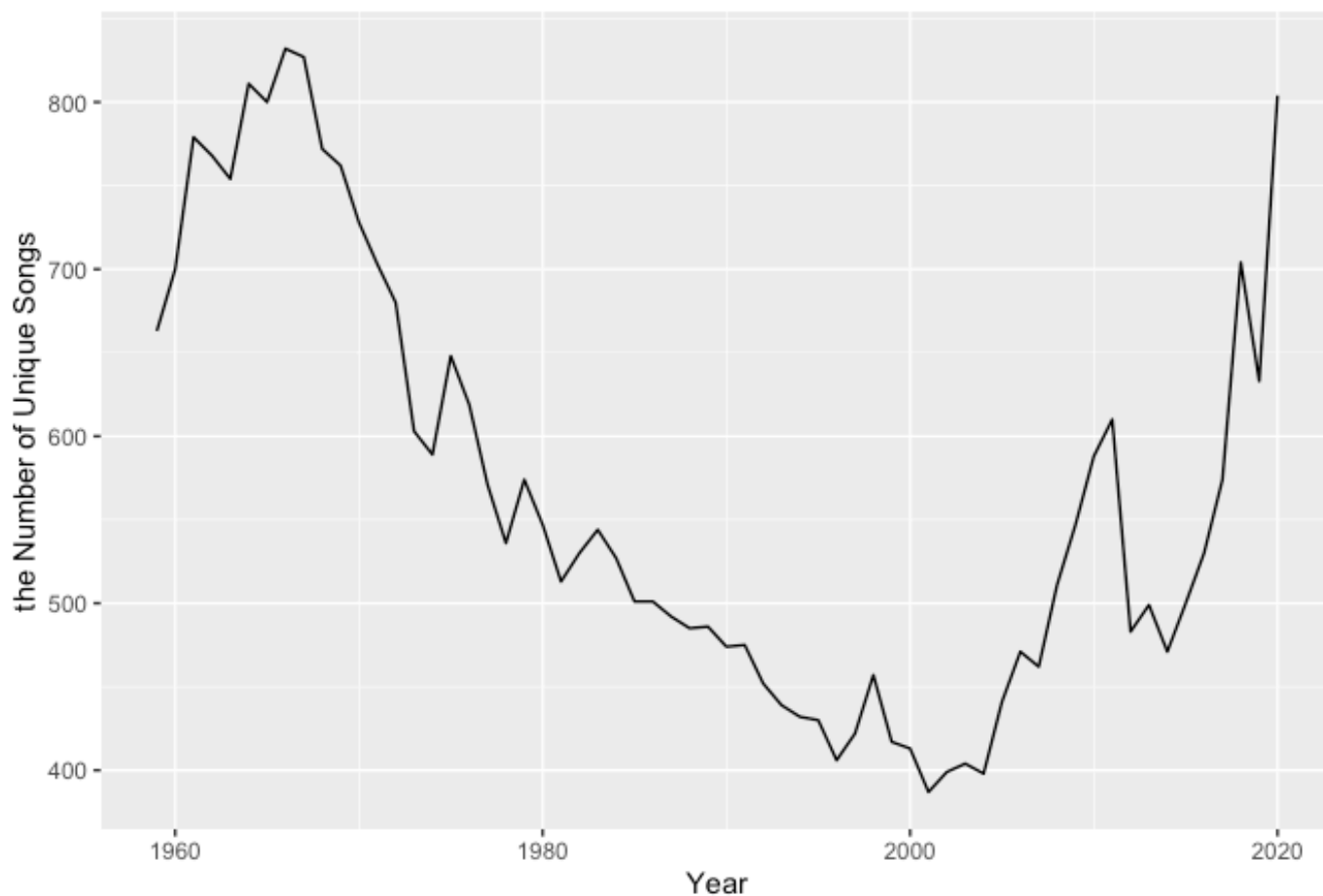


Figure 1

Part C

Show the songs that appeared on the Billboard Top 100 for at least ten weeks:

Table 3

```
## # A tibble: 6,126 × 2
##   performer      count
##   <chr>          <int>
## 1 Elton John      52
## 2 Madonna         44
## 3 Kenny Chesney   42
## 4 Tim McGraw      39
## 5 Keith Urban     36
## 6 Stevie Wonder   36
## 7 Taylor Swift    35
## 8 Michael Jackson 34
## 9 Rod Stewart     33
## 10 The Rolling Stones 33
## # ... with 6,116 more rows
```

There are 19 artists in U.S. musical history since 1958 who have had at least 30 songs that were “ten-week hits.”

Table 4

Use the bar plot to show the 19 artists:

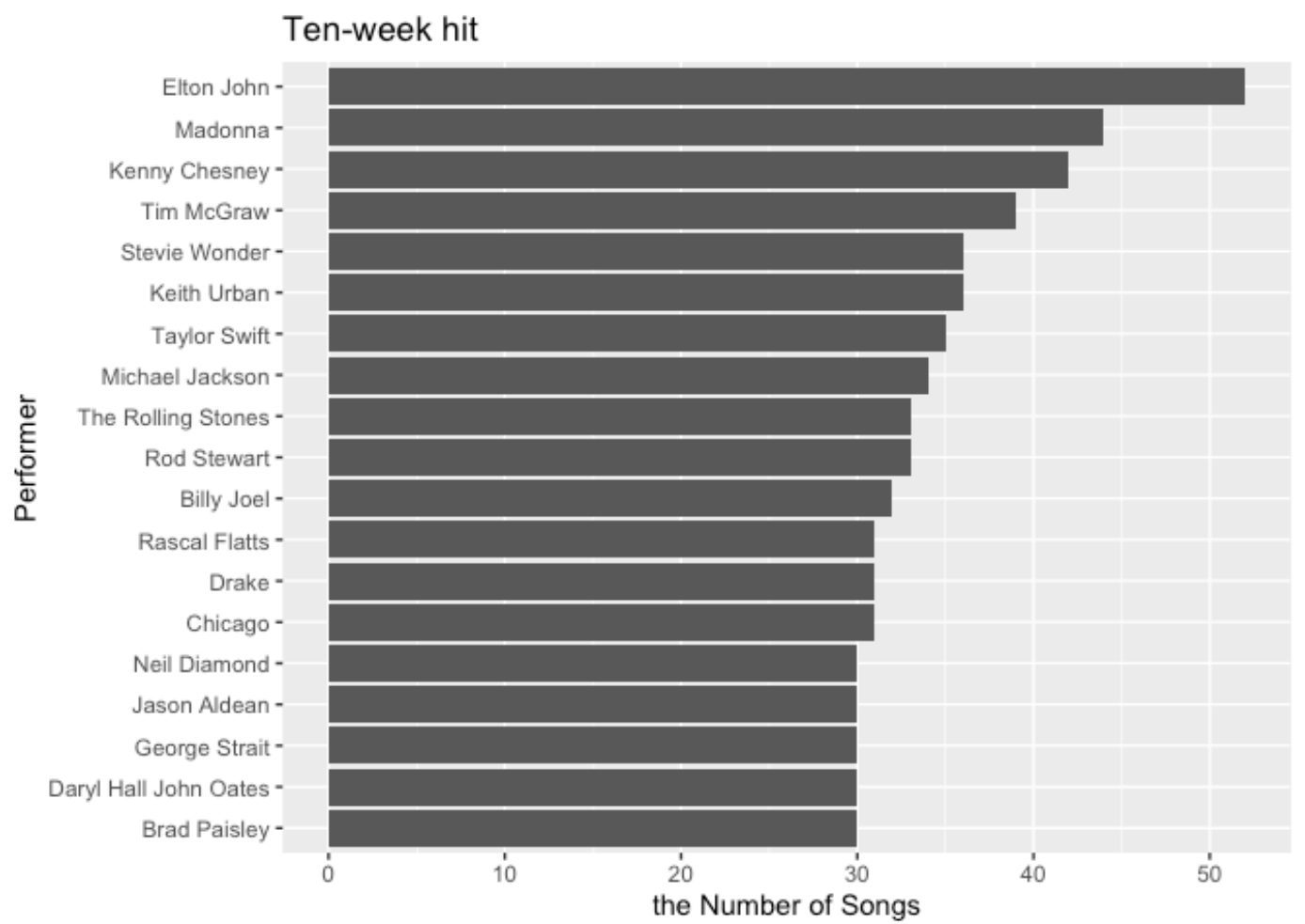


Figure 2

Problem 3: Wrangling the Olympics

Part A

Show the 95th percentile of heights for female competitors across all Athletics events:

```
##      q95_height
## 1           183
```

Part B

Show the top 10 variability in competitor’s heights of women’s event:

Table 1

```
## # A tibble: 10 × 2
##   event                                sd_height
##   <chr>                                <dbl>
## 1 Rowing Women's Coxed Fours          10.9
## 2 Basketball Women's Basketball       9.70
## 3 Rowing Women's Coxed Quadruple Sculls 9.25
## 4 Rowing Women's Coxed Eights         8.74
## 5 Swimming Women's 100 metres Butterfly 8.13
## 6 Volleyball Women's Volleyball       8.10
## 7 Gymnastics Women's Uneven Bars       8.02
## 8 Shooting Women's Double Trap         7.83
## 9 Cycling Women's Keirin              7.76
## 10 Swimming Women's 400 metres Freestyle 7.62
```

Part C

1. Show the trend of the average age of Olympic swimmers changed over time:

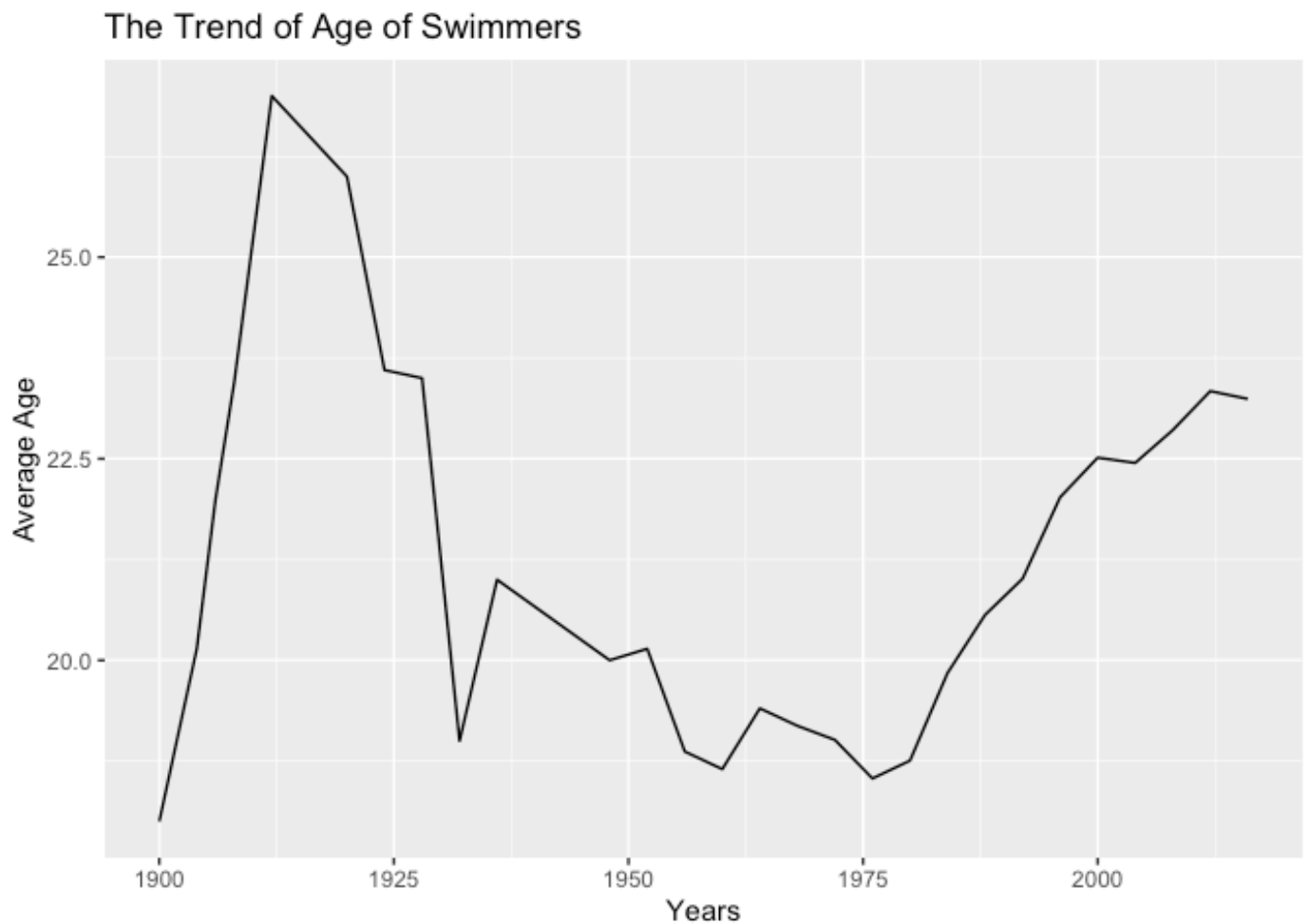


Figure 1: Trend for Total Swimmers

2. Show the trend of the average age of male swimmers and female swimmers changed over time:

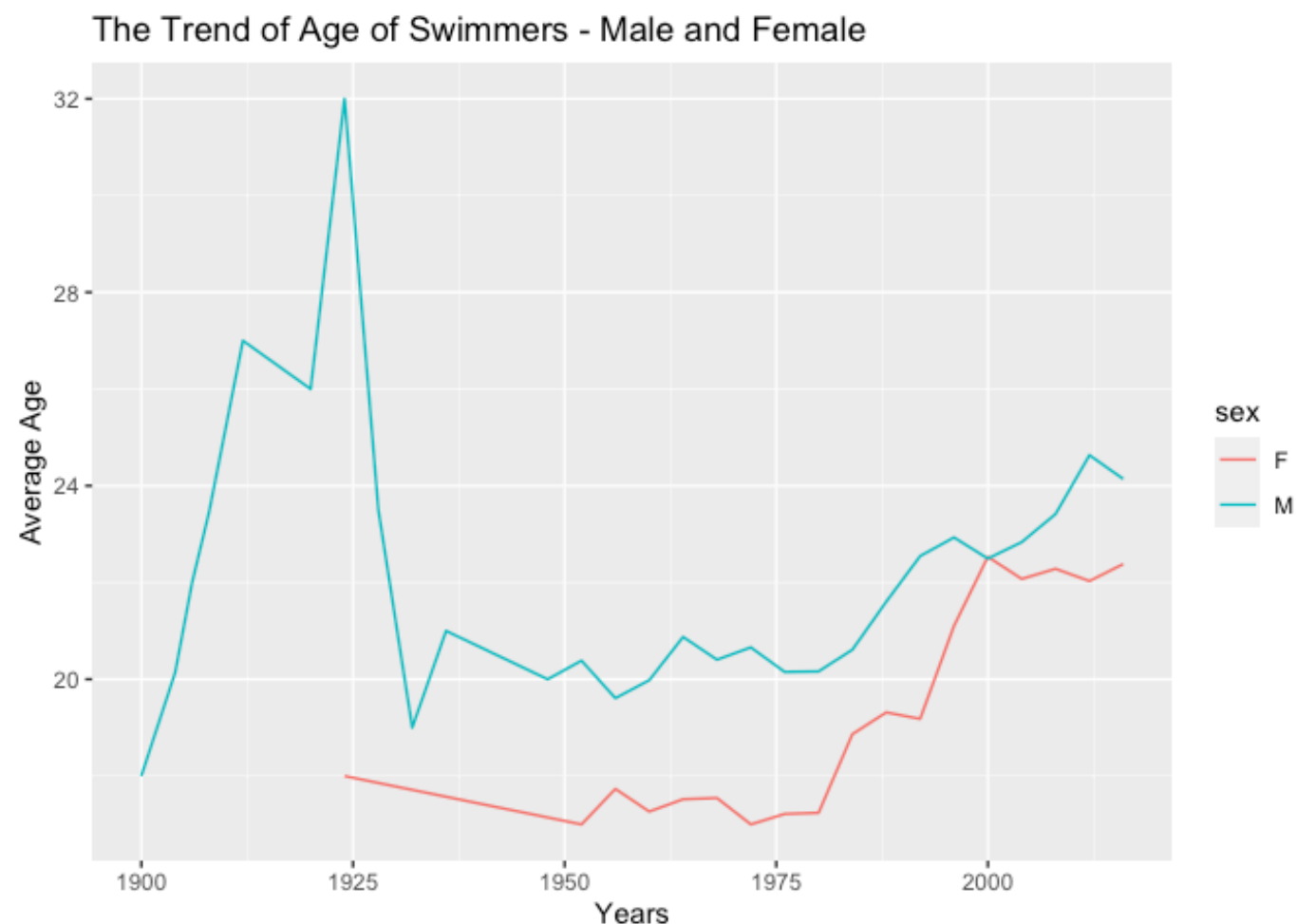


Figure 2: Trend for Male and Female Swimmers

Problem 4: Wrangling the Billboard Top 100

1. Take a look at 350 trim level

Creat a range of k:

Table 1

```
## [1] 2 4 6 8 10 12 14 16 18 20 25 30 35 40 45 50 55 60 65
## [20] 70 75 80 85 90 95 100
```

K-fold cross validation

Model across the train/test splits

Table 2

```
##          k      err  std_err
## result.1  2 12173.54 581.7625
## result.2  4 10854.16 586.5408
## result.3  6 10189.53 558.9616
## result.4  8 10123.46 594.6117
## result.5 10 10068.49 559.0305
## result.6 12 10053.43 584.1031
```

Plot means and standard errors across k

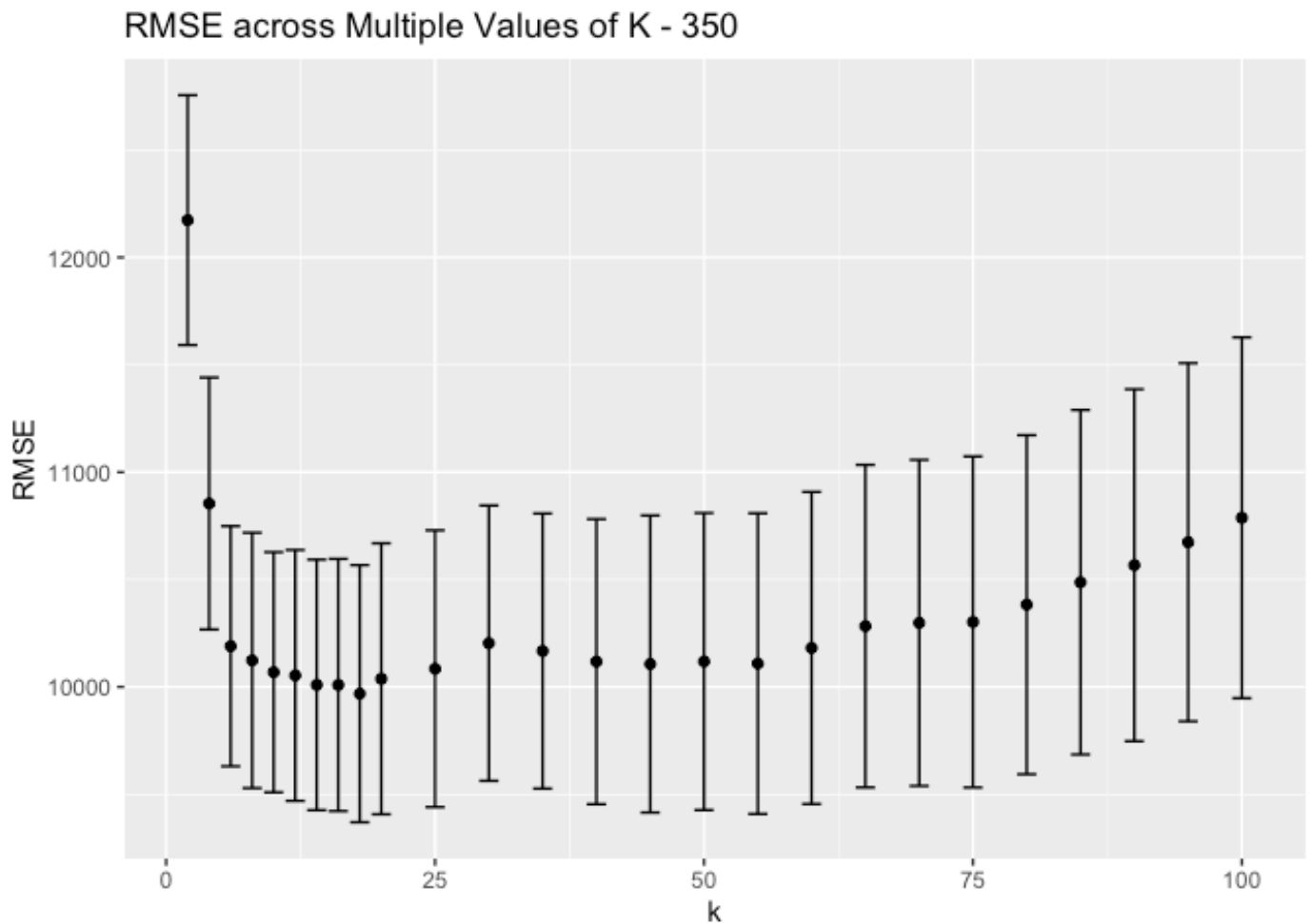


Figure 1

Find the optimal value of k

KNN with optimal k

Attach the predictions to the data and add the predictions

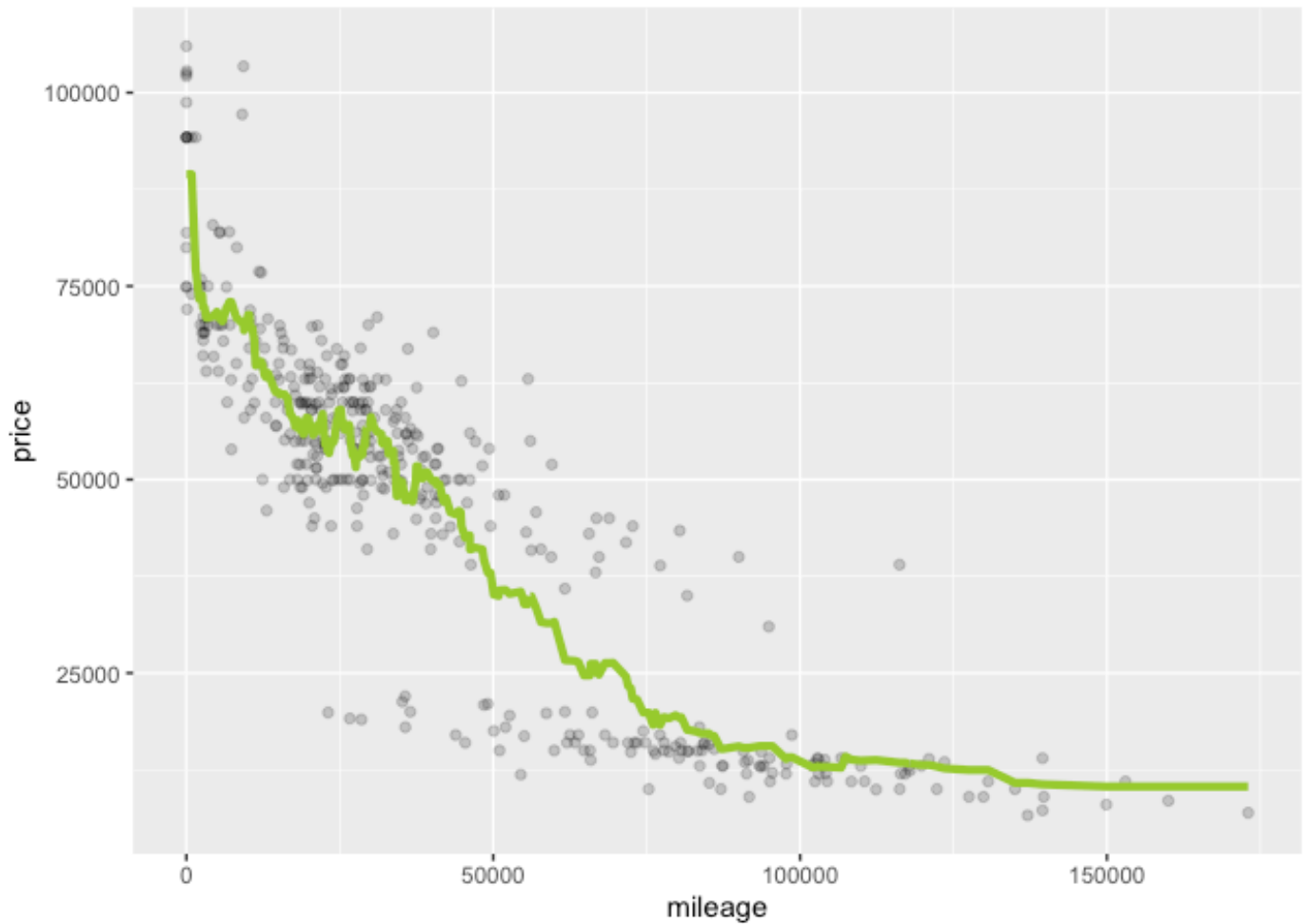


Figure 2

2. Take a look at 65 AMG trim level

K-fold cross validation

Model across the same train/test splits

Table 3

##	k	err	std_err
## result.1	2	24456.34	726.4251
## result.2	4	22337.66	812.3805
## result.3	6	22131.99	1021.7378
## result.4	8	21753.93	1079.4855
## result.5	10	21304.36	1236.1598
## result.6	12	21139.84	1321.0863

Plot means and standard errors versus k

RMSE across Multiple Values of K - 65AMG

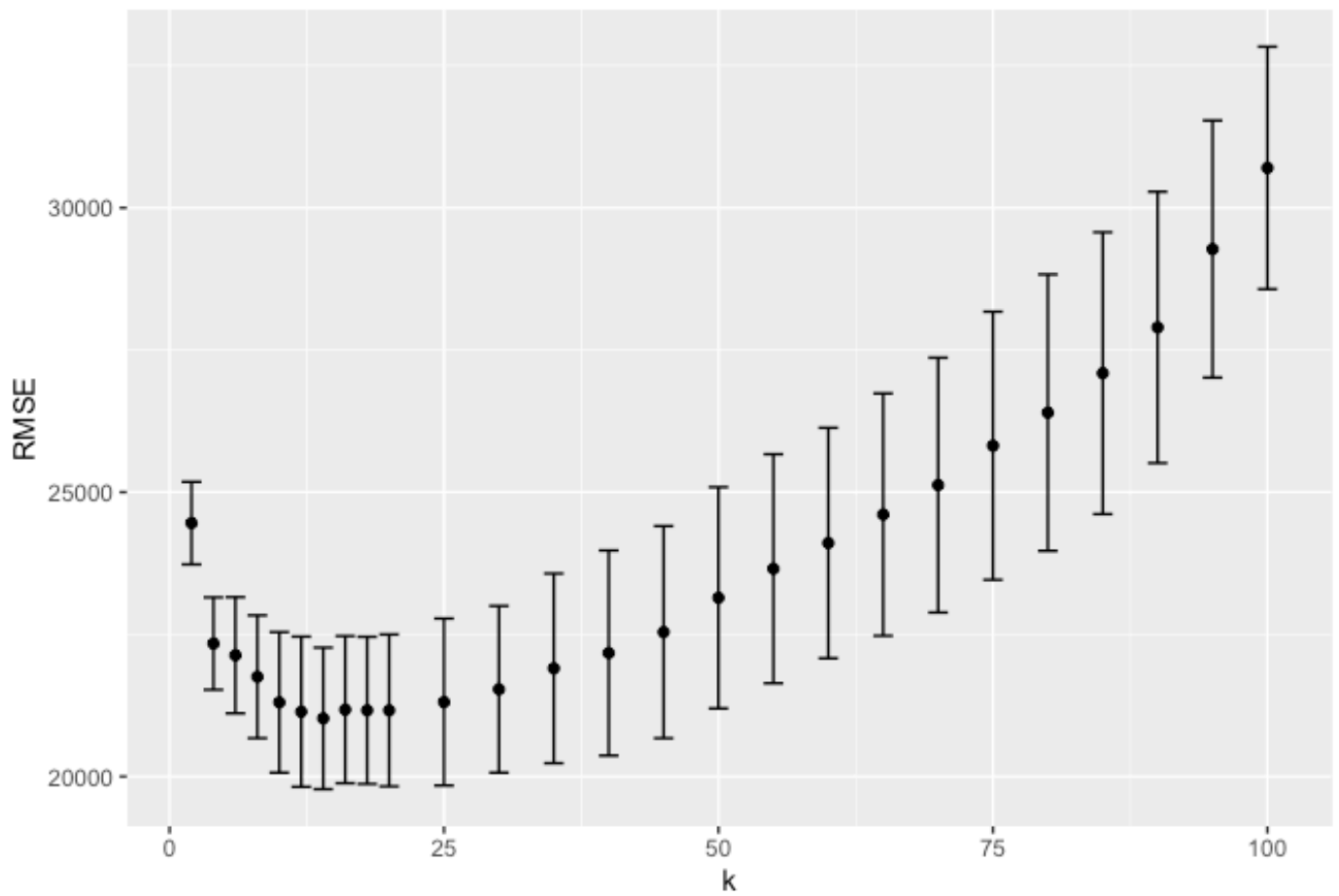


Figure 3

Find the optimal value of k

KNN with optimal k

Attach the predictions to the data and add the predictions

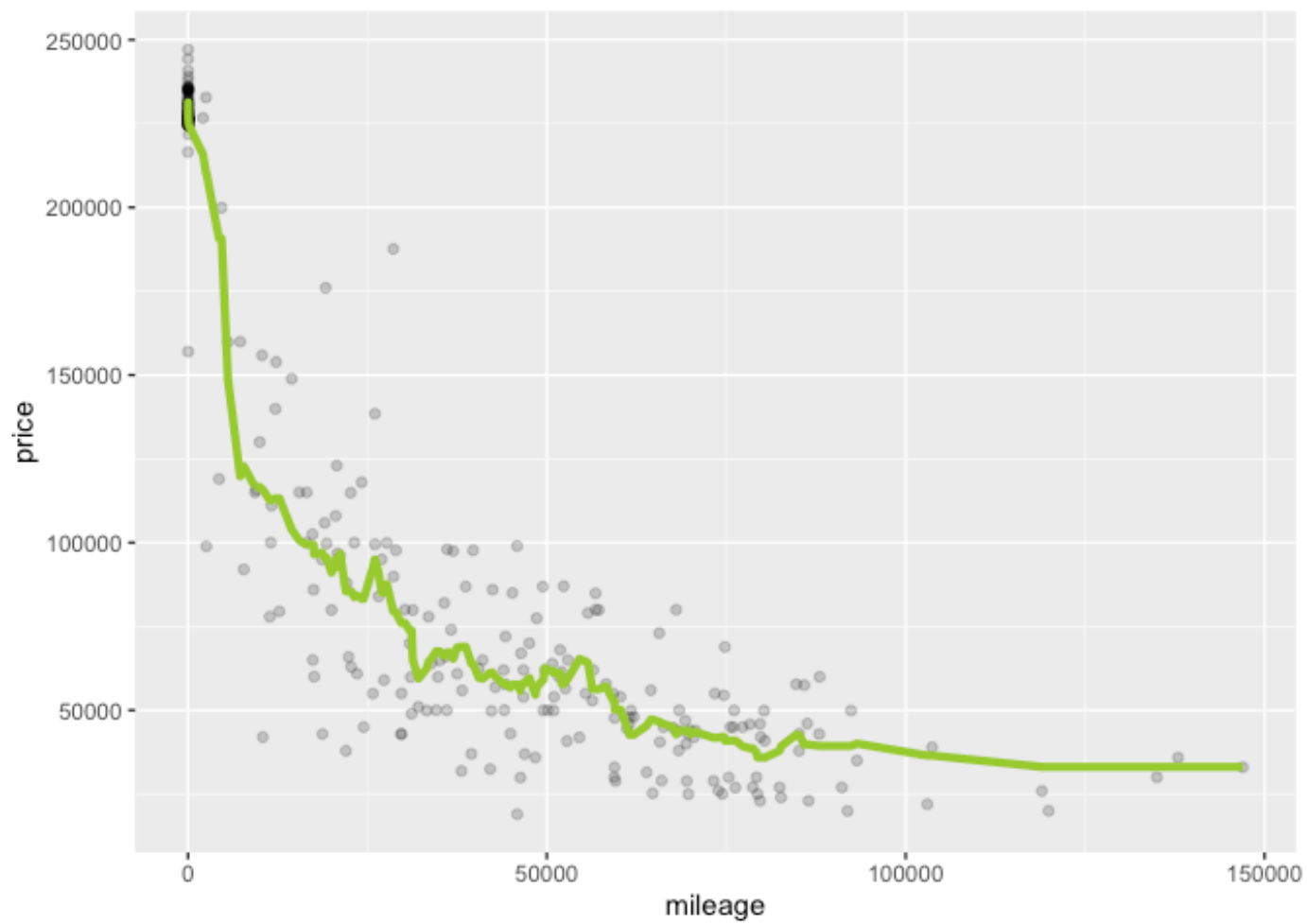


Figure 4