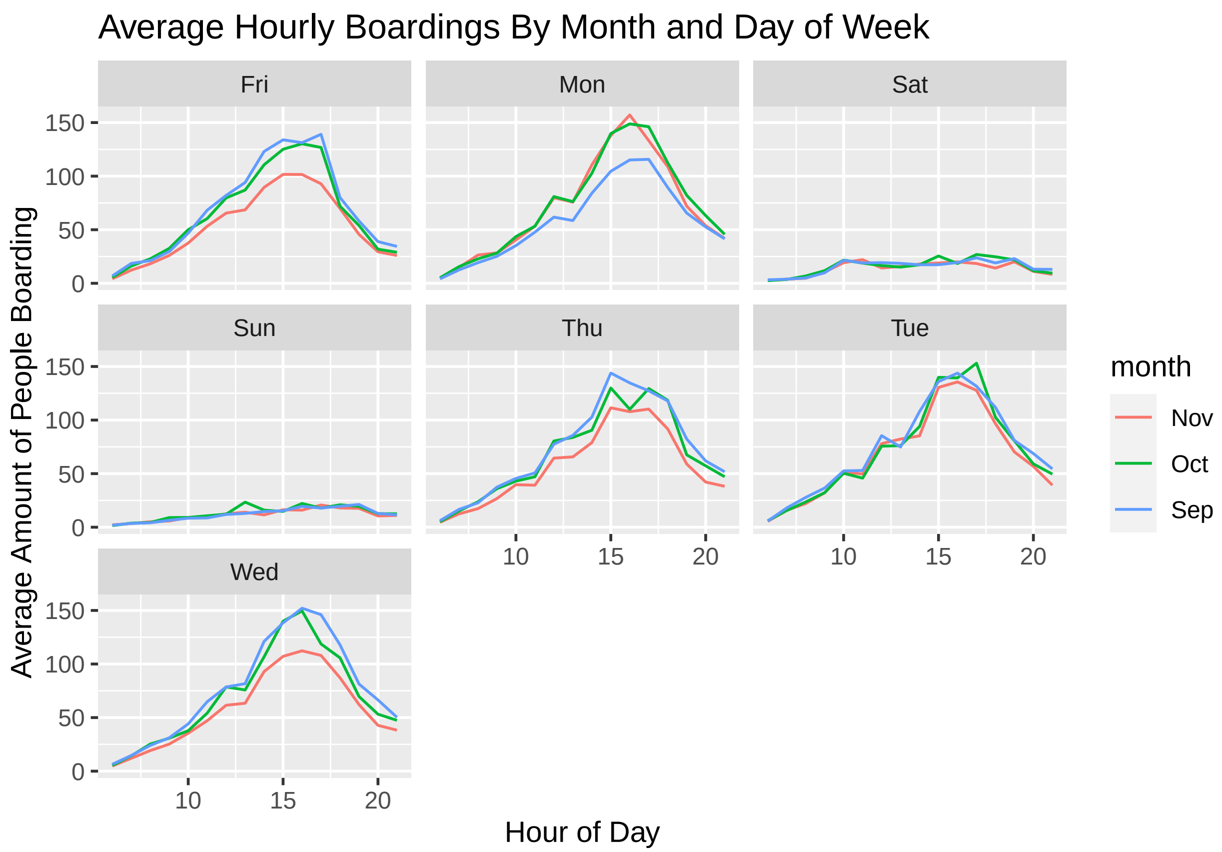
Assignment 2

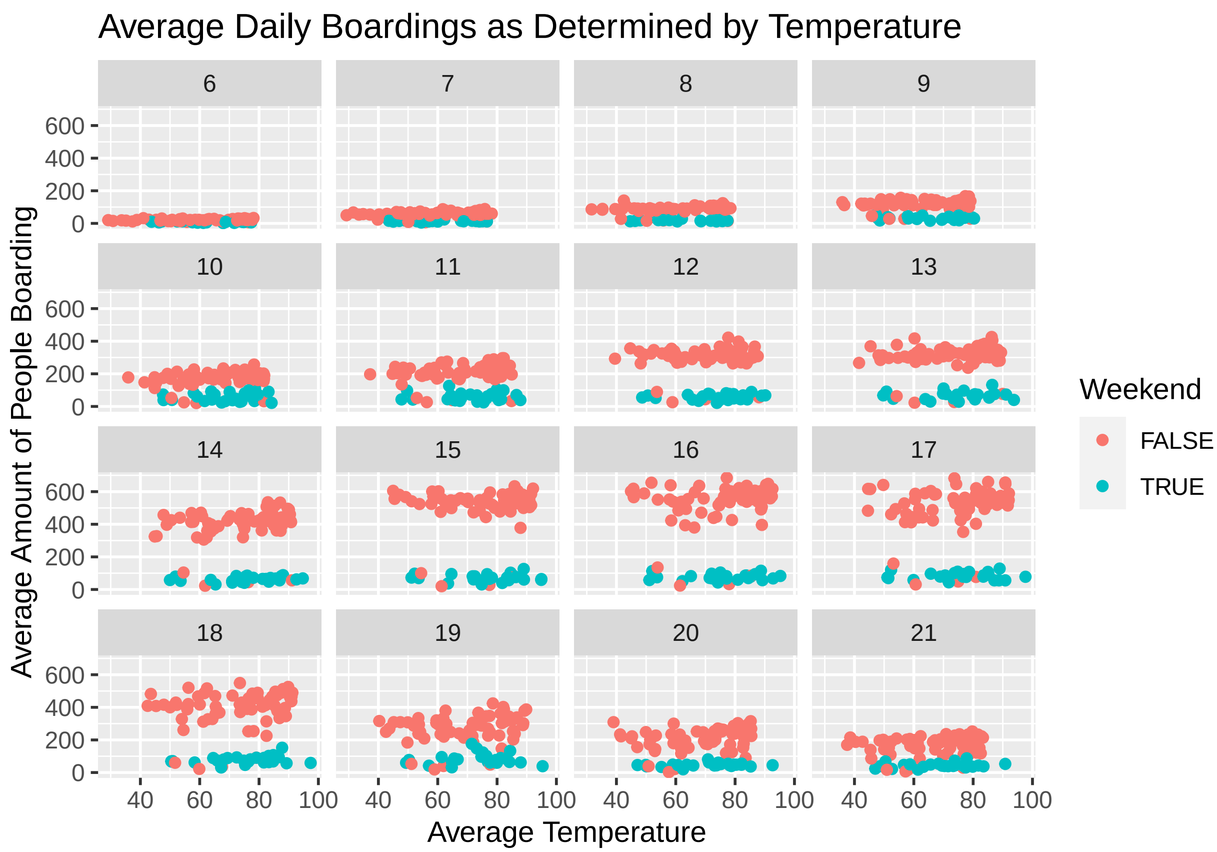
Elliot Spears

#Question 1

```
## # A tibble: 6 x 4
## # Groups:   month, hour_of_day [1]
##   month hour_of_day day_of_week avg_board
##   <chr>       <dbl> <chr>           <dbl>
## 1 Nov             6 Fri              4.2
## 2 Nov             6 Mon              4.38
## 3 Nov             6 Sat              2.62
## 4 Nov             6 Sun              2.19
## 5 Nov             6 Thu              4.65
## 6 Nov             6 Tue              5.44
```



It appears from this set of graphs that the peak boarding times are very similar across days. The peak seems to be between 4pm and 6pm, which is when many people get off of work. On the weekends this is not quite the case. On Sunday there is a highpoint around noon, which may be explained by people going to brunch or coming home from church. It may be the case that there are less boardings on Mondays in the month of September because labor day always falls on a Monday in September, which drastically reduces the demand for bus rides for that Monday alone, which depresses the mean value of boardings for Mondays in September. One possible explanation as to why boardings are lower on wed/thur/fri in November is because many UT students go home for thanksgiving break on Tuesday, which is the last school day before spring break. For the rest of that one week demand for buses around UT is substantially lower, which again depresses the average value for those days overall in the month of November.




```
When holding hour of day and weekend status constant, temperature doesn't seem to have a very significa
```

#Question 2

```
##                         (Intercept)                             lotSize
##                               16608                               37741
##                          livingArea                                 age
##                                  70                                  63
##                          fireplaces                            bedrooms
##                              -14423                               -4677
##                     poly(lotSize^2)                           bathrooms
##                             -407169                               28253
##                           landValue                  poly(fireplaces^2)
##                                   1                              416834
##               livingArea:fireplaces                   lotSize:landValue
##                                   0                                   0
##                  bedrooms:bathrooms                         lotSize:age
##                               -1301                                -241
##           centralAirNo:heatingelectric       centralAirYes:heatingelectric
```

```
##                                    -1992                                    -9988
##            centralAirNo:heatinghot air          centralAirYes:heatinghot air
##                                     -216                                    12555
##   centralAirNo:heatinghot water/steam centralAirYes:heatinghot water/steam
##                                    -7751                                       NA
```

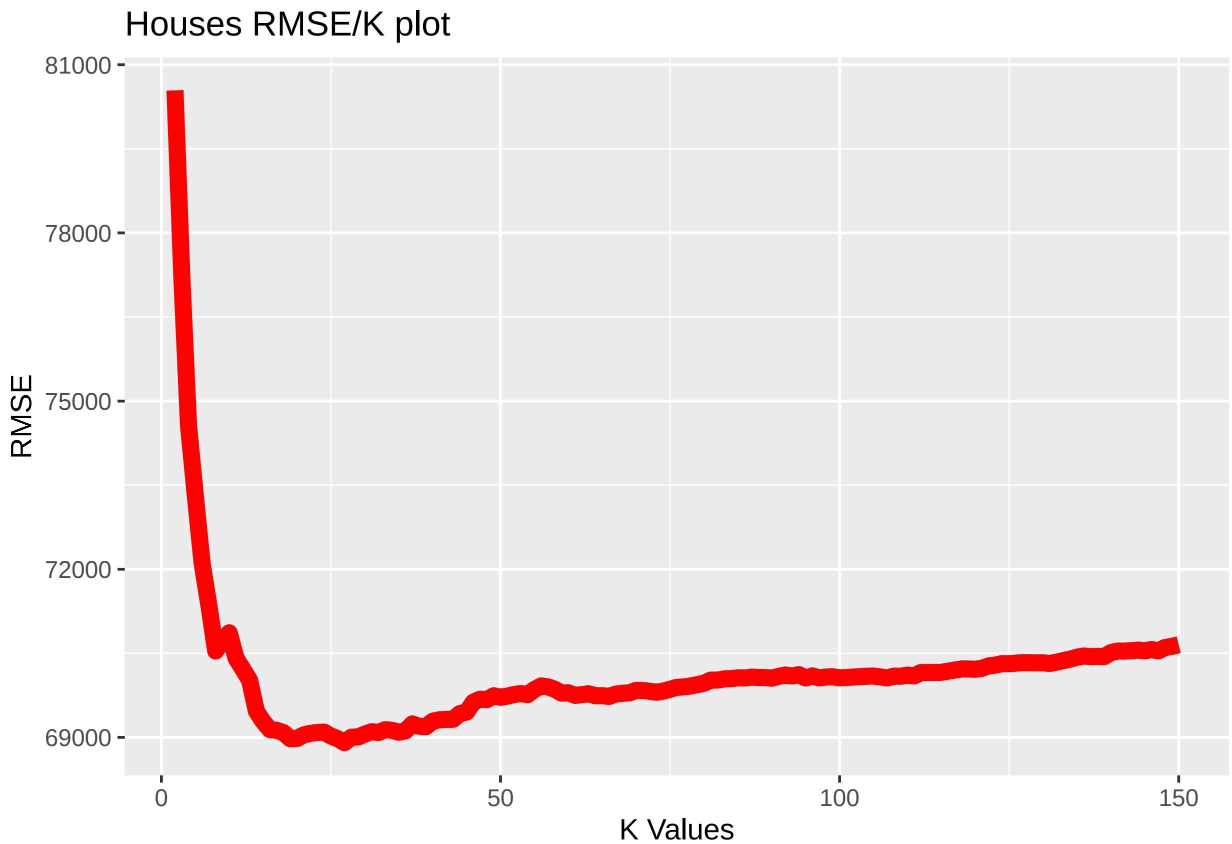
```
## [1] 59396.8
```

```
I was able to get the rmse down to $61280.12 after a few quick adjustments. I didn’t want to spend too r
```

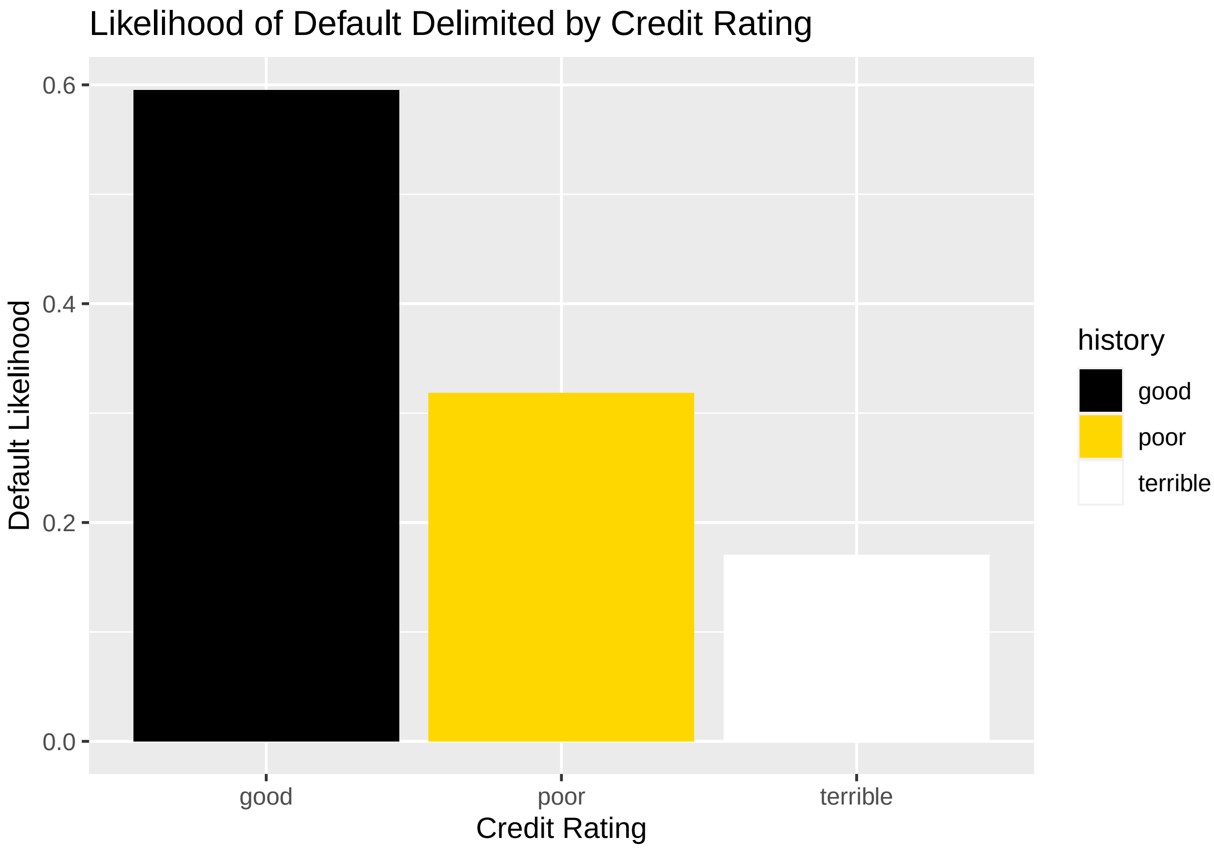
```
##  [1] 74004.58 69431.98 82911.19 70357.37 68727.25 72897.20 75866.28 66598.59
##  [9] 77599.80 70566.32
```

```
## [1] 72896.06
```

```
I was able to get down to a mean rmse of $71157.28. I started off by incorporating all of the same vari
```



As can be seen in the graph above, the optimal k value for this regression is around 25-30, but closer to 28.



#Question 3

Here I have created three bar graphs that show default percentage as broken down by credit history. Obviously, there is a problem here. The people that fall into the category of "good" credit have the highest default rate according to the data, which ended up being near 60%. The folks with "terrible" credit have a default rate of about 20%, which is 40% lower than the "good" credit folks. This result doesn't make sense as we would expect the opposite result.

```
##
## Call:  glm(formula = Default ~ duration + amount + installment + age +
##     history + purpose + foreign, family = binomial, data = credit)
##
## Coefficients:
##         (Intercept)             duration               amount
##          -7.075e-01            2.526e-02            9.596e-05
##         installment                  age          historypoor
##           2.216e-01           -2.018e-02           -1.108e+00
##     historyterrible           purposeedu  purposegoods/repair
##          -1.885e+00            7.248e-01            1.049e-01
##       purposenewcar       purposeusedcar        foreigngerman
##           8.545e-01           -7.959e-01           -1.265e+00
##
## Degrees of Freedom: 999 Total (i.e. Null);  988 Residual
## Null Deviance:       1222
## Residual Deviance: 1070  AIC: 1094
```

```
Clearly, as the credit history worsens, the partial effect of the history variables goes down. Although
We need to find out what the default rate is for each category overall before we manipulate the dataset
```

#Question 4

```
##                   (Intercept)  market_segmentComplementary
##                           -15                           12
##      market_segmentCorporate         market_segmentDirect
##                            10                           12
##         market_segmentGroups  market_segmentOffline_TA/TO
##                            10                           11
##      market_segmentOnline_TA                       adults
##                            12                            0
##           customer_typeGroup      customer_typeTransient
##                             0                            0
## customer_typeTransient-Party             is_repeated_guest
##                             0                           -1
```

```
## [1] 3.141624
```

Above we have the RMSE that was yielded by the first baseline model.

```
##                    (Intercept)              hotelResort_Hotel
##                            -17                             -1
##                      lead_time        stays_in_weekend_nights
##                              0                              0
##            stays_in_week_nights                         adults
##                              0                             -1
##                         mealFB                         mealHB
##                              1                              0
##                         mealSC                  mealUndefined
##                             -1                              0
##    market_segmentComplementary        market_segmentCorporate
##                             13                             12
##           market_segmentDirect           market_segmentGroups
##                             13                             12
##    market_segmentOffline_TA/TO        market_segmentOnline_TA
##                             13                             13
##      distribution_channelDirect        distribution_channelGDS
##                              1                            -14
##       distribution_channelTA/TO              is_repeated_guest
##                              0                             -1
##         previous_cancellations previous_bookings_not_canceled
##                             -1                              0
##            reserved_room_typeB             reserved_room_typeC
##                              2                              3
##            reserved_room_typeD             reserved_room_typeE
##                             -1                              0
##            reserved_room_typeF             reserved_room_typeG
##                              1                              2
##            reserved_room_typeH             reserved_room_typeL
##                              3                            -14
##            assigned_room_typeB             assigned_room_typeC
##                              0                              2
##            assigned_room_typeD             assigned_room_typeE
##                              1                              1
```

```
##                assigned_room_typeF                 assigned_room_typeG
##                                  1                                   1
##                assigned_room_typeH                 assigned_room_typeI
##                                  2                                   2
##                assigned_room_typeK                     booking_changes
##                                  0                                   0
##             deposit_typeNon_Refund             deposit_typeRefundable
##                                  0                                 -11
##               days_in_waiting_list                 customer_typeGroup
##                                  0                                   0
##             customer_typeTransient       customer_typeTransient-Party
##                                  0                                   0
##                 average_daily_rate required_car_parking_spacesparking
##                                  0                                   0
##          total_of_special_requests
##                                  0
```

```
## [1] 4.05003
```

The RMSE calculated above is the RMSE of the big model, which includes all possible predictors, with the exception of "arrival_date." The RMSE calculated from the larger model is greater than what we obtained from the smaller baseline model. Next, I will try and build the best possible model that I can.

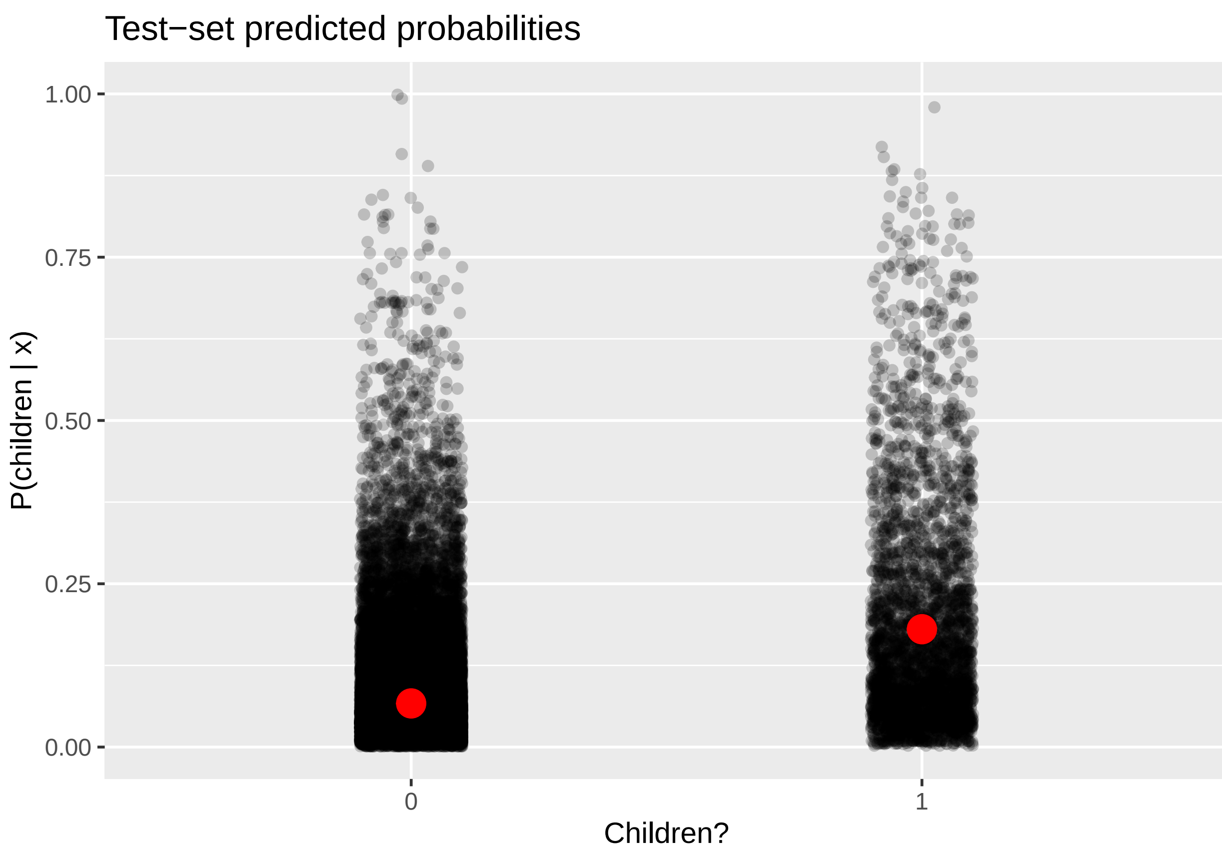
```
##            (Intercept)      stays_in_week_nights                     adults
##                     -6                         0                          1
##                 mealFB                    mealHB                     mealSC
##                      0                         0                         -1
##          mealUndefined        average_daily_rate adults:average_daily_rate
##                     -1                         0                          0
```

```
## [1] 3.05709
```

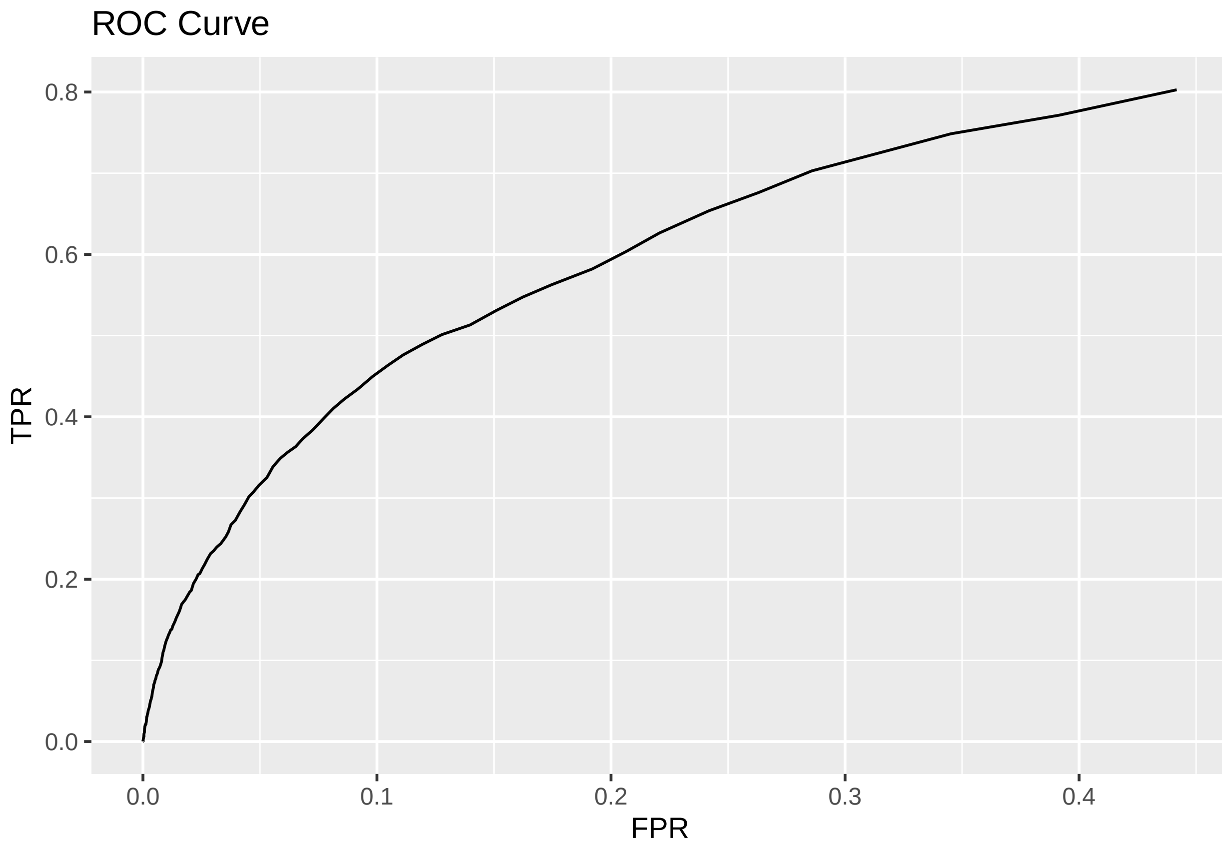
This was the lowest RMSE I could muster. It beats the first baseline model by .02 RMSE points. It includes an interaction between adults and average daily rate as I figured that those variables together may help predict the number of children present.

```
##            (Intercept)      stays_in_week_nights                     adults
##                     -5                         0                          0
##                 mealFB                    mealHB                     mealSC
##                      1                         0                         -1
##          mealUndefined        average_daily_rate adults:average_daily_rate
##                     -1                         0                          0
```

```
## [1] 3.170306
```




```
##          yhat
## children     0     1
##        0 41179   186
##        1  3392   243
```



The hotel_val data has validated my previous model as the RMSE it turned out is virtually the same and differs from the previous model by only .002 RMSE points.However, the ROC curve is anything but impressive in terms of our true positive rate that we obtained. This means that my model that I constructed is not a very reliable predictor of whether a guests will be bringing children.

```
## Training
##   |                                                                          |
```

```
## Fold: 11/20
##   |                                                                      |
```

Round	Accuracy	Better_Than_Original
1	95.6	Yes
2	94.8	Yes
3	93.6	Yes
4	92.4	Yes
5	92.4	Yes
6	92.8	Yes
7	93.6	Yes
8	87.6	
9	95.2	Yes
10	93.6	Yes
11	89.6	
12	92.8	Yes
13	92.8	Yes
14	93.6	Yes
15	93.6	Yes
16	92.4	Yes
17	94.4	Yes
18	94.0	Yes
19	91.6	Yes
20	94.8	Yes

For the last part of the assignment, I tested the model over 20 folds. I've provided a graph to list off the occurrences where the model indicated was better than the baseline.