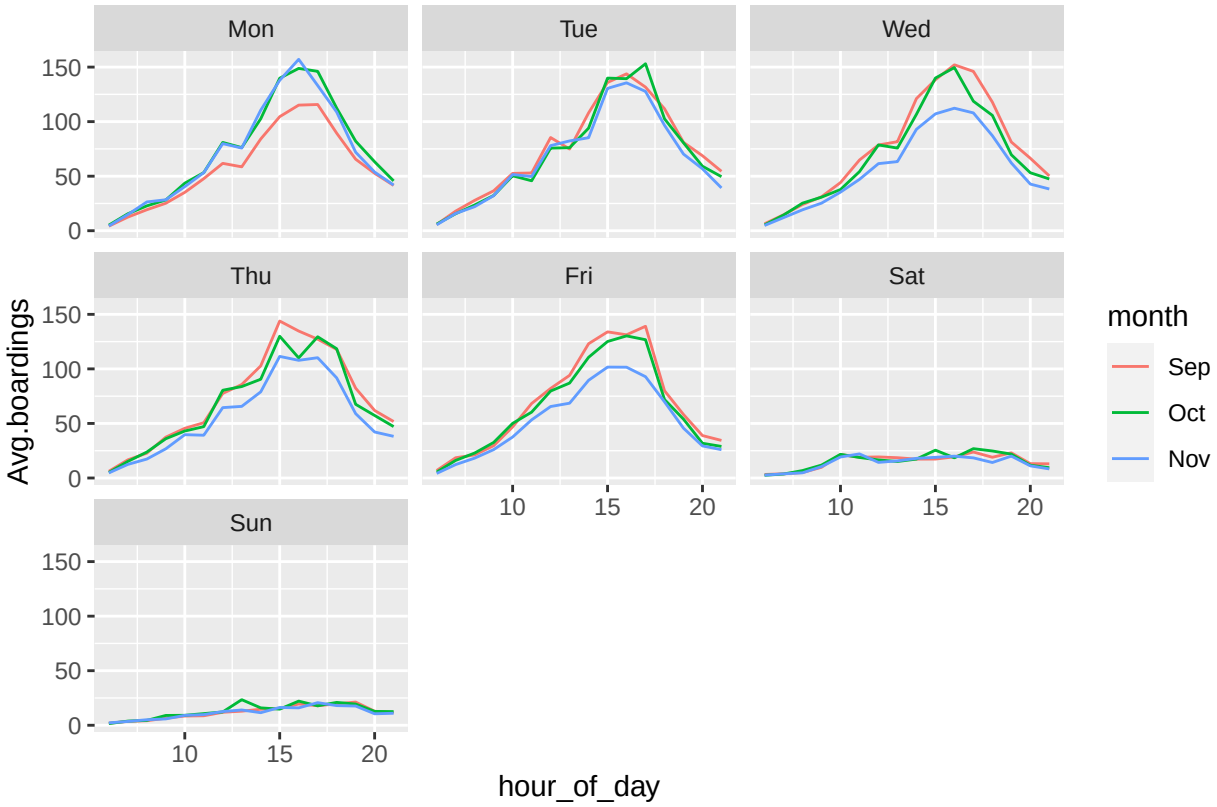
DataMining_HW2

Wen-Hsin Chang

2021/3/1

Problem 1: Visualization

Panel A



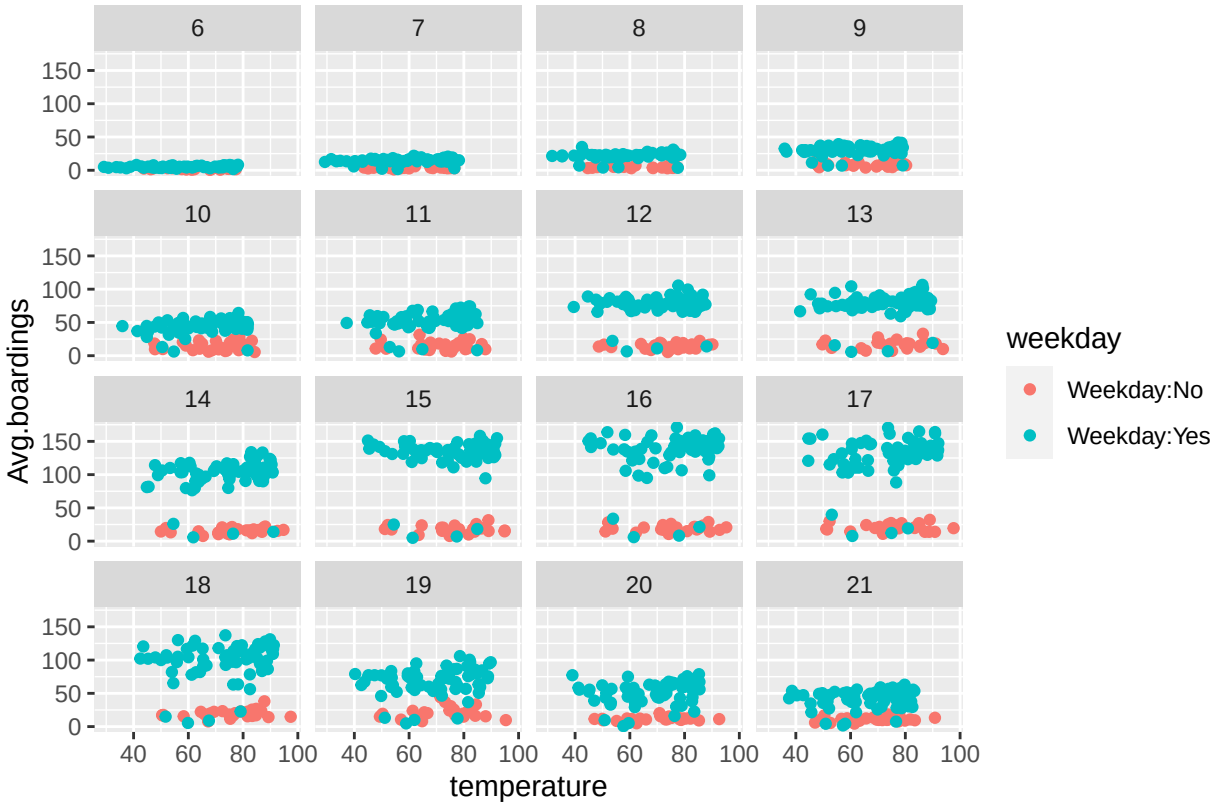
Plot A: Average bodardings by hour of day, colored by month, faceted according to day of a week

Does the hour of peak boardings changes from day to day? According to the graph above, the hour of peak boardings is around 17:30 during workday and the pattern is similar from day to day. However, the boardings remain low during weekend and there isn't a clear peak as indicated by the flat line.

Why do you think average boardings on Mondays in September look lower? We can clearly see that the average boardings in September on Monday are lower compared to other workdays and other months. One possible reason may be that when the fall semester first starts in September, students suffer more from "Monday blues", making them less likely to go to school.

Why do you think average boardings on Weds/Thurs/Fri in November look lower Based on the graph above, it does show lower average boardings on Weds/Thurs/Fri in November. My guess is that the Thanksgiving holiday is going on by the end of November, and thus students are very likely to be off-campus.

Panel B



Plot B: Average bodardings by temperature, colored by weekday,and faceted by hour

Based on Panel B, we can see that when we hold hour of day and weekend status constant, the temperature seems not to have a noticeable effect on the number of UT students riding the bus since there isn't a clear upward or downward trend in the scatter plots. One possible explanation may be that the weather rage in Austin during the sample period is not too spread out and therefore not too extreme to influence a student's willingness to commute.

Problem 2: Saratoga house prices

First, to get a sense of the significance of all the factors in linear form, my way is to look at the regression result of all covariates.

```
##
## Call:
## lm(formula = price ~ ., data = SaratogaHouses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -228655  -35225   -4929   27480  457325
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.098e+05  1.968e+04   5.581 2.77e-08 ***
## lotSize                  7.599e+03  2.241e+03   3.391 0.000713 ***
## age                     -1.304e+02  5.839e+01  -2.234 0.025600 *
## landValue                9.219e-01  4.757e-02  19.379  < 2e-16 ***
## livingArea               6.996e+01  4.615e+00  15.159  < 2e-16 ***
## pctCollege              -1.102e+02  1.515e+02  -0.727 0.467139
## bedrooms                -7.835e+03  2.567e+03  -3.052 0.002309 **
## fireplaces               1.037e+03  2.986e+03   0.347 0.728504
## bathrooms                2.311e+04  3.370e+03   6.859 9.66e-12 ***
## rooms                    3.020e+03  9.619e+02   3.139 0.001722 **
## heatinghot water/steam -1.045e+04  4.190e+03  -2.495 0.012679 *
## heatingelectric        -8.245e+01  1.232e+04  -0.007 0.994661
## fuelelectric           -1.093e+04  1.213e+04  -0.901 0.367799
## fueloil                -4.381e+03  5.015e+03  -0.874 0.382466
## sewerpublic/commercial -1.524e+03  3.667e+03  -0.416 0.677775
## sewernone              -4.845e+03  1.712e+04  -0.283 0.777239
## waterfrontNo           -1.202e+05  1.554e+04  -7.734 1.77e-14 ***
## newConstructionNo       4.544e+04  7.308e+03   6.219 6.29e-10 ***
## centralAirNo           -9.953e+03  3.478e+03  -2.862 0.004266 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58260 on 1709 degrees of freedom
## Multiple R-squared:  0.6534, Adjusted R-squared:  0.6498
## F-statistic:   179 on 18 and 1709 DF,  p-value: < 2.2e-16
```

From the regression result of the medium sample, we can tell that lotSize, landvalue, living area, bedrooms, bathrooms, waterfront, and NewConstruction are very important factors in determining price (significant at 1%). In terms of economic significance, I also decide to include the interaction term between landvalue and lotSize because it makes intuitive sense.

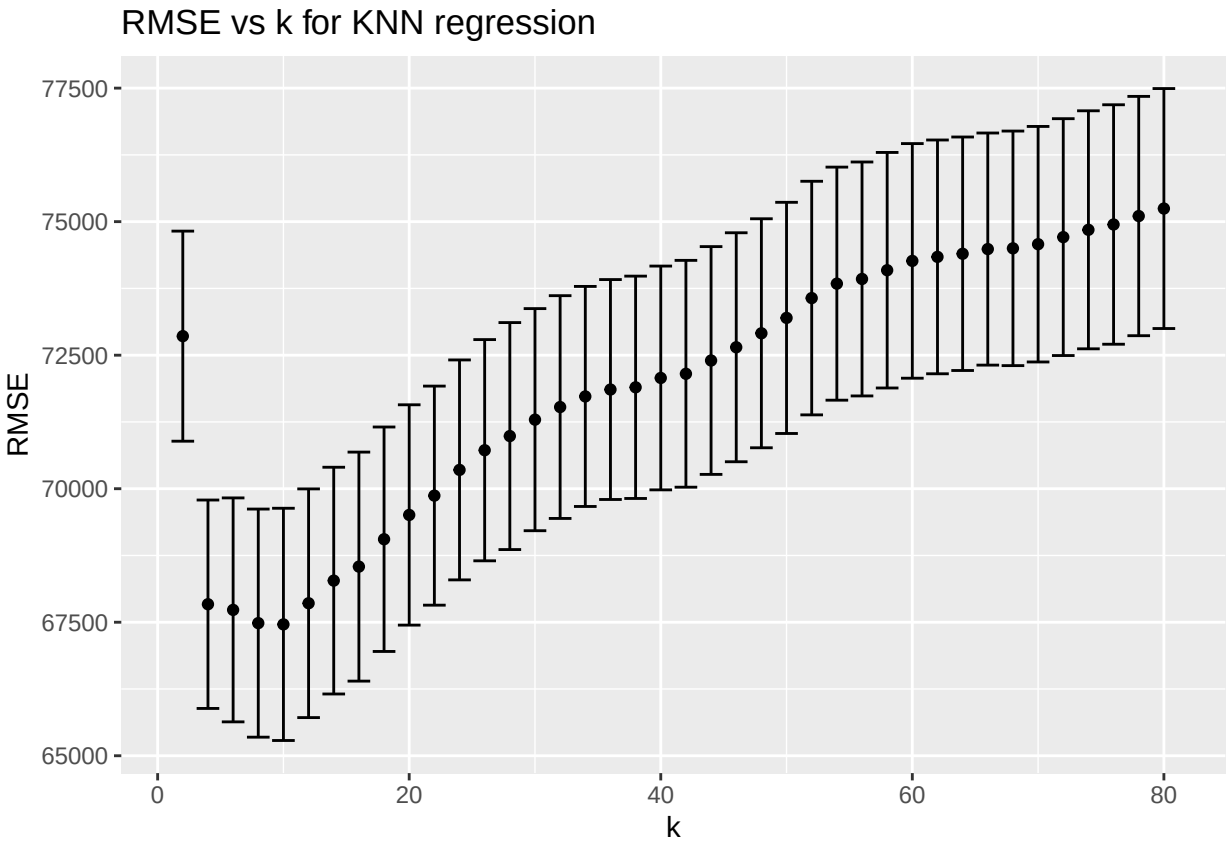
performance of the"medium" linear model versus the new linear model

```
## [1] "RMSE of medium model: 66945"
```

```
## [1] "RMSE of new linear model: 59670"
```

The result above shows that the RMSE of my new linear model is smaller than the RMSE of the medium model. I think the main reason may be that the medium model includes too many variables that are not highly related to the house price, making the prediction less accurate outside the training sample.

My best KNN model
I include all variables that are significant in the full regression above.



From the graph above, it seems that my KNN model attains the least mean square error around k=5. Therefore, I will use k=5 in the analysis afterward.

** My linear model versus KNN model (after standardizing)**

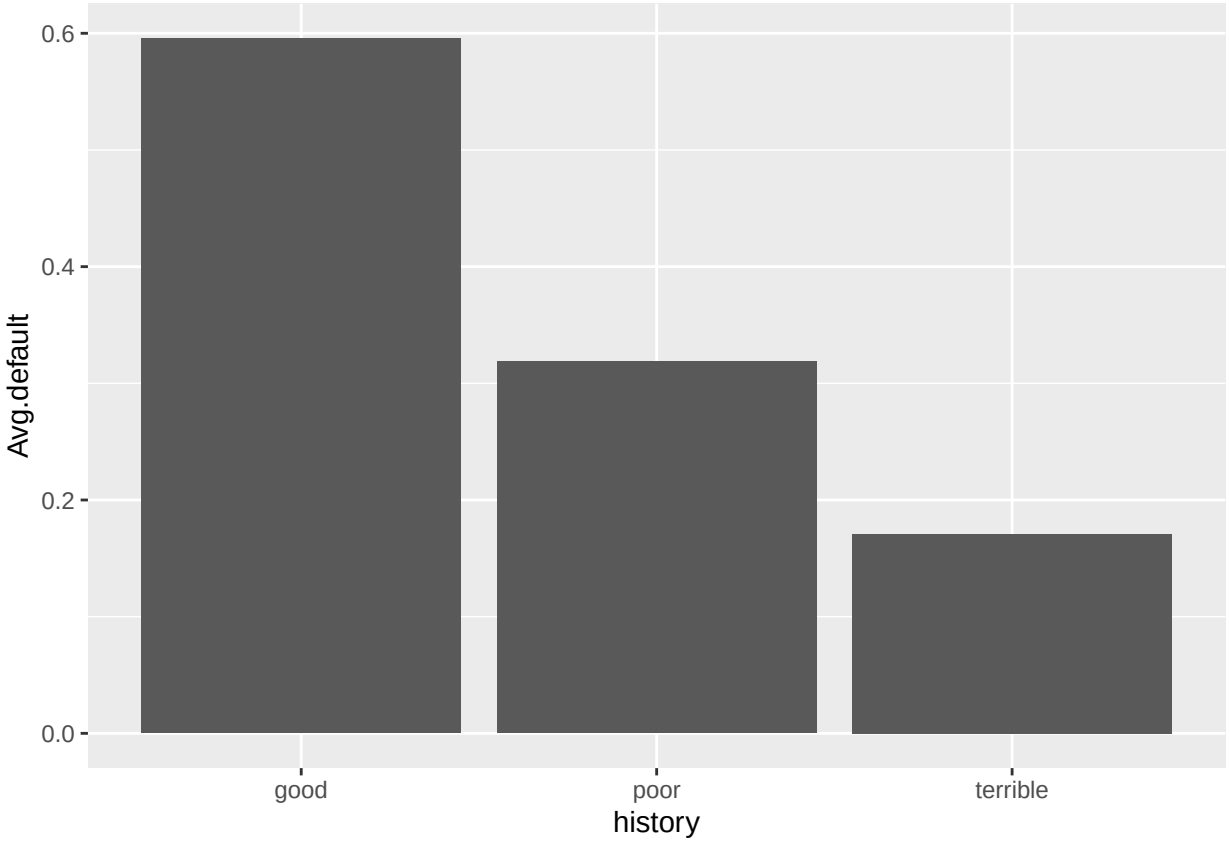
```
## [1] "RMSE of linear model: 67113"
```

```
## [1] "RMSE of KNN model: 72250"
```

According to the average estimate of out-of-sample RMSE over many train/test split, my linear model seems to perform better than the KNN model. Therefore, I recommend building a linear model using variable lotSize, landvalue, living area, bedrooms, bathrooms, waterfront,and NewConstruction and adding an interatciton term between landvalue and living area. However, the model is always subject to improvement. If we can obtain a more diverse data category, then maybe it will work better overall or under the KNN method.

Problem 3: Classification and retrospective sampling

First, build a bar plot of **default probability by credit history**. It seems that the credit history of "good" is more likely to trigger defaults.



Next, build a logistic regression model.

```
##
## 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
```

The history variables show that relative to good history firms, poor history and terrible history firms are less likely to default, which is extremely counter-intuitive. I think this data set is not a good input for building a predictive model of defaults because the bank should not over-sample the defaults. For high credit ranking firms with default, it is less likely to find a matched pair. I recommend using a random sample from the overall portfolio and adding more variables to control for other factors.

Problem 4: Children and hotel reservations

Model Building Firstly, I would like to take a look at the linear probability model coefficient.

```
##                        (Intercept)                  hotelResort_Hotel
##                      -5.775695e-02                       -3.307893e-02
##                          lead_time             stays_in_weekend_nights
##                       3.979883e-05                        3.577001e-03
##               stays_in_week_nights                              adults
##                      -8.994214e-04                       -3.800720e-02
##                             mealFB                              mealHB
##                       4.684703e-02                        2.647172e-05
##                             mealSC                       mealUndefined
##                      -5.304903e-02                        2.223866e-02
##         market_segmentComplementary             market_segmentCorporate
##                       5.263609e-02                        4.656078e-02
##               market_segmentDirect                market_segmentGroups
##                       4.783874e-02                        5.892673e-02
##         market_segmentOffline_TA/TO             market_segmentOnline_TA
##                       7.127149e-02                        6.539160e-02
##          distribution_channelDirect             distribution_channelGDS
##                       1.895338e-02                       -7.584360e-02
##           distribution_channelTA/TO                   is_repeated_guest
##                       1.233949e-03                       -3.084634e-02
##              previous_cancellations     previous_bookings_not_canceled
##                       6.904447e-04                       -2.246336e-03
##                 reserved_room_typeB                 reserved_room_typeC
##                       2.080364e-01                        5.367538e-01
##                 reserved_room_typeD                 reserved_room_typeE
##                      -6.699691e-02                       -2.556823e-02
##                 reserved_room_typeF                 reserved_room_typeG
##                       3.079356e-01                        4.306056e-01
##                 reserved_room_typeH                 reserved_room_typeL
##                       6.009265e-01                       -8.548933e-02
##                 assigned_room_typeB                 assigned_room_typeC
##                       1.226022e-02                        9.064864e-02
##                 assigned_room_typeD                 assigned_room_typeE
##                       5.929701e-02                        5.020963e-02
##                 assigned_room_typeF                 assigned_room_typeG
##                       6.288261e-02                        9.305579e-02
##                 assigned_room_typeH                 assigned_room_typeI
##                       7.277336e-02                        8.871559e-02
##                 assigned_room_typeK                    booking_changes
##                       2.755465e-02                        1.989851e-02
##              deposit_typeNon_Refund              deposit_typeRefundable
##                       2.773819e-02                        2.127366e-02
##                days_in_waiting_list                  customer_typeGroup
##                      -2.471735e-05                       -4.031377e-03
##              customer_typeTransient       customer_typeTransient-Party
##                       1.256707e-02                       -2.954751e-02
##                  average_daily_rate required_car_parking_spacesparking
##                       8.754251e-04                       -3.406172e-04
##           total_of_special_requests
##                       3.268004e-02
```

For my best linear model, I expand baseline 2 and add an interaction term between adults and total_of_special_requests. The main reason is that holding the number of adults fixed, when a room requests more special requests, it may be an indicator that there are "hidden children" in the room.

Out-of-sample performance for baseline1, baseline2, and my best linear model My best linear model seems to perform the best among the three.

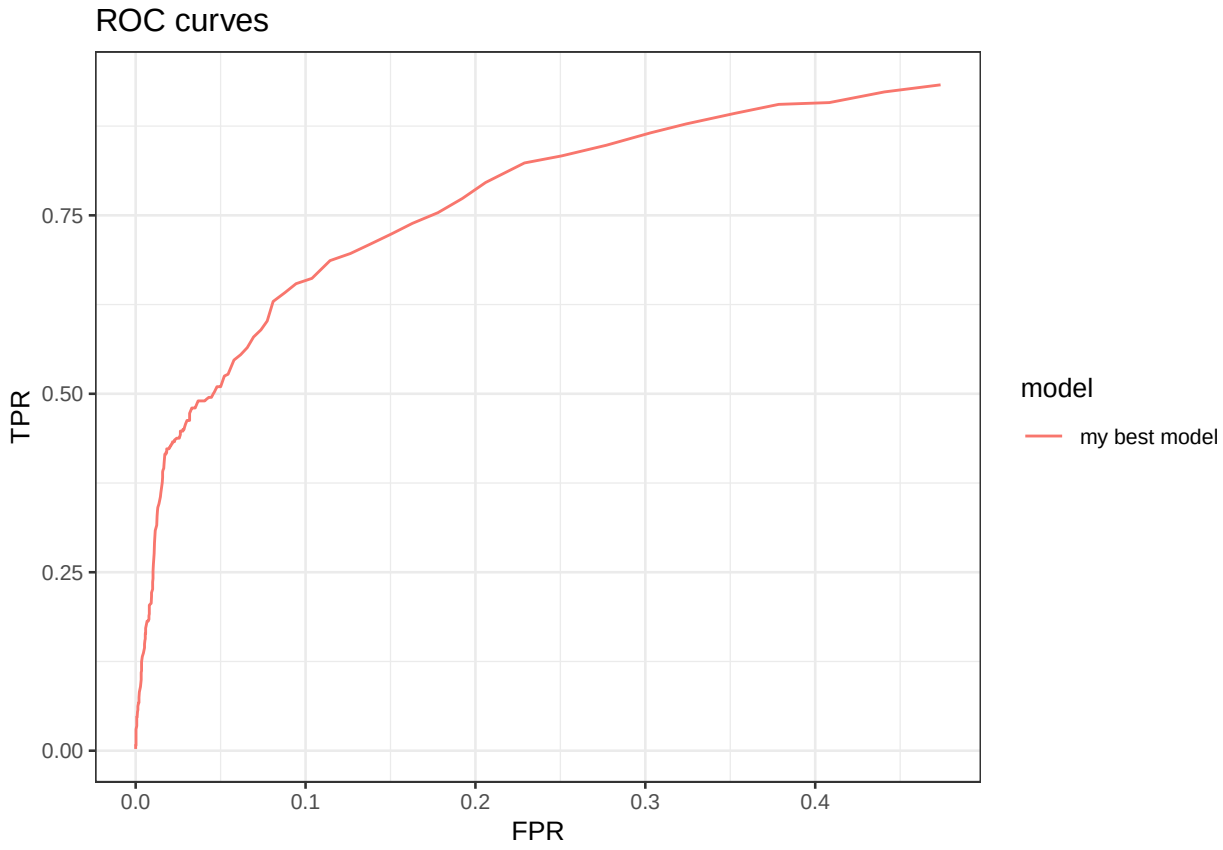
```
## [1] "out-of-sample performance- Baseline1: 0.919222"
```

```
## [1] "out-of-sample performance- Baseline2: 0.935067"
```

```
## [1] "out-of-sample performance- My best model: 0.935267"
```

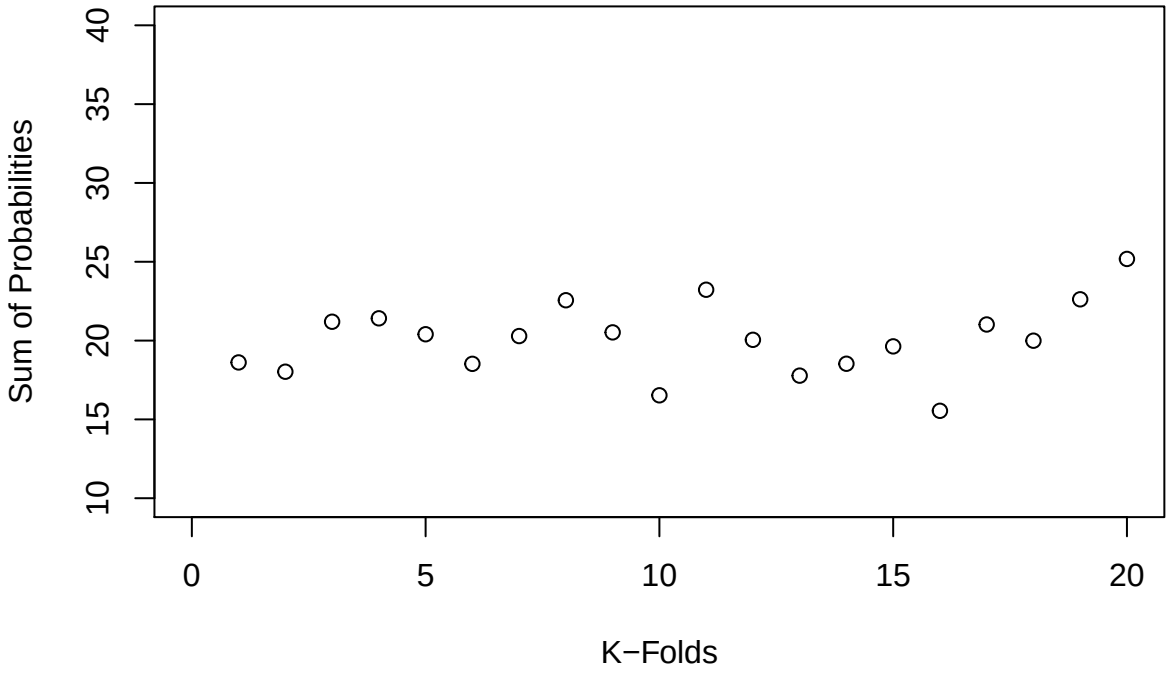
Model validation:step 1

Below is the ROC curve for my best model. The shape seems pretty standard.

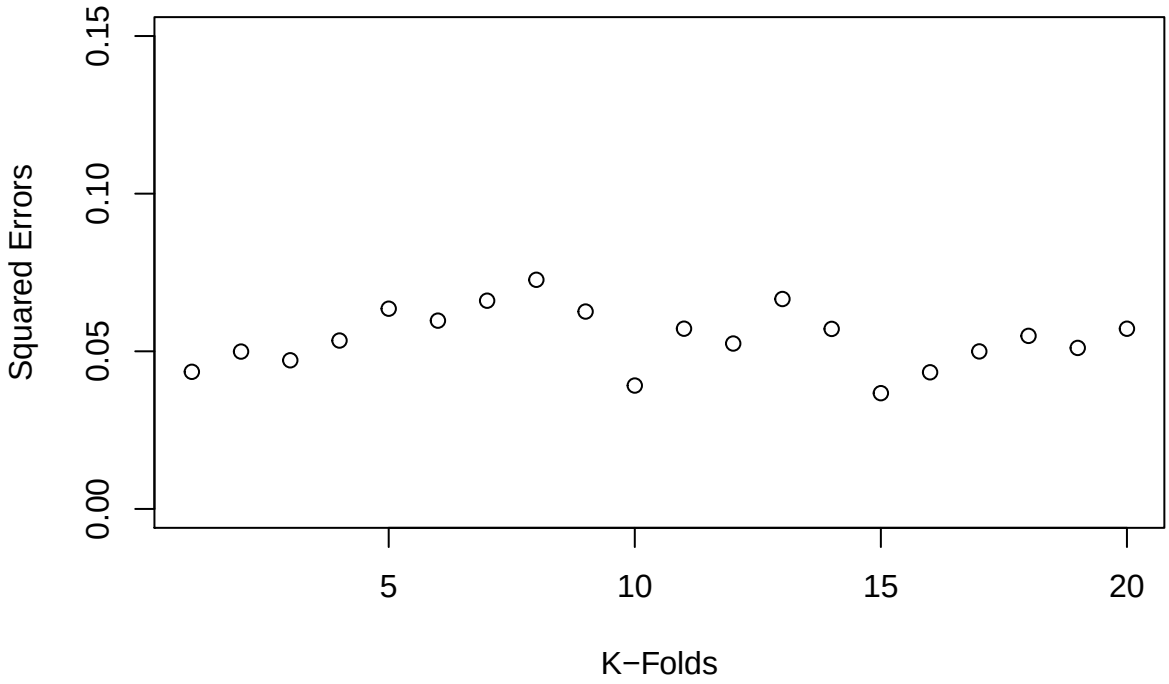


Model validation:step 2

I create 20 folds within the testing data. Within each fold, I predict whether each booking will have children on it. Moreover, I sum up the predicted probabilities for all the booking in the fold. Below is the graph of the **expected number of bookings across 20 folds**.



Below is the **squared error of expected number of bookings across 20 folds** (versus actual number of bookings with children in that fold.)



My model does relatively well at predicting the total number of bookings. The sum of square errors is quite small. Overall, the K-folds methods give me more confidence to validate my model.