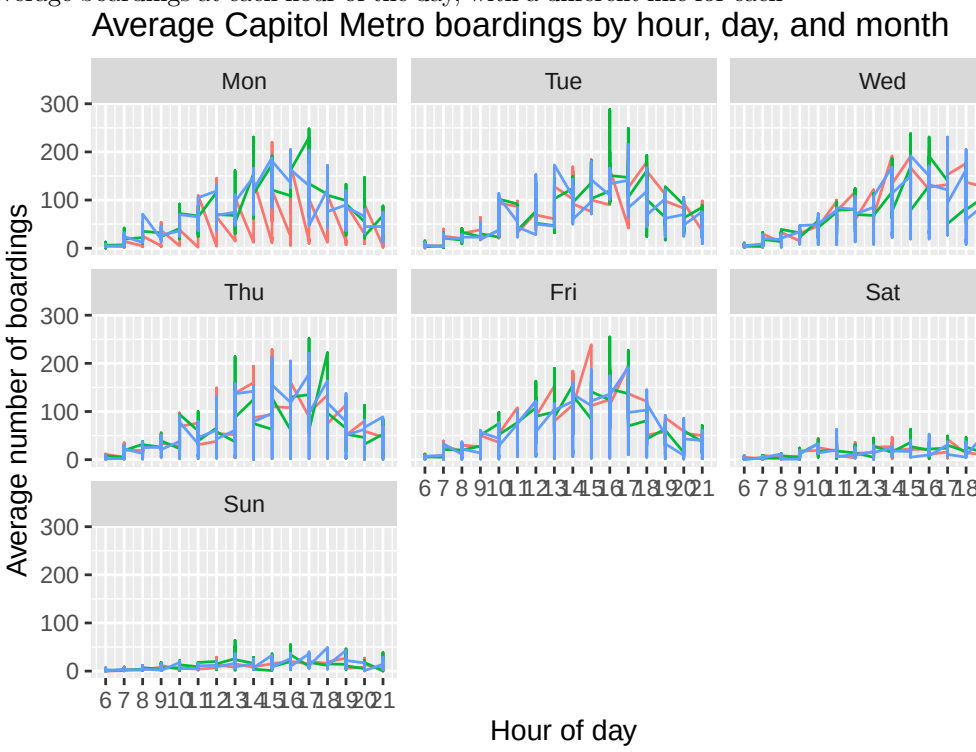
Assignment 2

Maria Gilbert

3/1/2021

Data Visualization: Capitol Metro

A) One panel of line graphs that plot average boardings at each hour of the day, with a different line for each month, faceted by day of the week.



B) One panel of scatter plots of boardings by temperature, faceted by hour of day and colored according to

weekday versus weekend.

Saratoga House Prices

Linear model

RMSE of the professor's simple model:

```
## [1] 72724.65
```

RMSE of the professor's moderate model:

```
## [1] 66019.17
```

RMSE of the professor's advanced model:

```
## [1] 68044.23
```

I found that a linear regression of price, log(lotSize+1), age, log(livingArea/rooms), landValue, livingArea, bedrooms, fireplaces, bathrooms, log(rooms+1), heating, centralAir, bedrooms times bathrooms, age times pctCollege, and newConstruction resulted in an RMSE of:

```
## [1] 62481.76
```

I tested its relative improvement over the professor's moderate model several times, finding an improvement between 10% and 20%.

```
## [1] 1.056615
```

KNN model

For the KNN model, I tested RMSE using different values of K, based on the scaled parameters of log(lotSize+1),age,log(livingArea/rooms), landValue, livingArea, bedrooms, fireplaces, bathrooms, log(rooms+1), heating, centralAir, bedrooms times bathrooms, age times pctCollege, and newConstruction.

For K = 25, RMSE is:

```
## [1] 79365.46
```

For K = 26, RMSE is:

```
## [1] 79516.31
```

For K = 27, RMSE is:

```
## [1] 79271.76
```

For K = 28, RMSE is:

```
## [1] 79323.93
```

For K = 29, RMSE is:

```
## [1] 79698.69
```

For K = 30, RMSE is:

```
## [1] 80005.32
```

For K = 31, RMSE is:

```
## [1] 80177.63
```

For K = 32, RMSE is:

```
## [1] 80413.18
```

For K = 33, RMSE is:

```
## [1] 80591.99
```

For K = 34, RMSE is:

```
## [1] 80560.22
```

For K = 35, RMSE is:

```
## [1] 80801.28
```

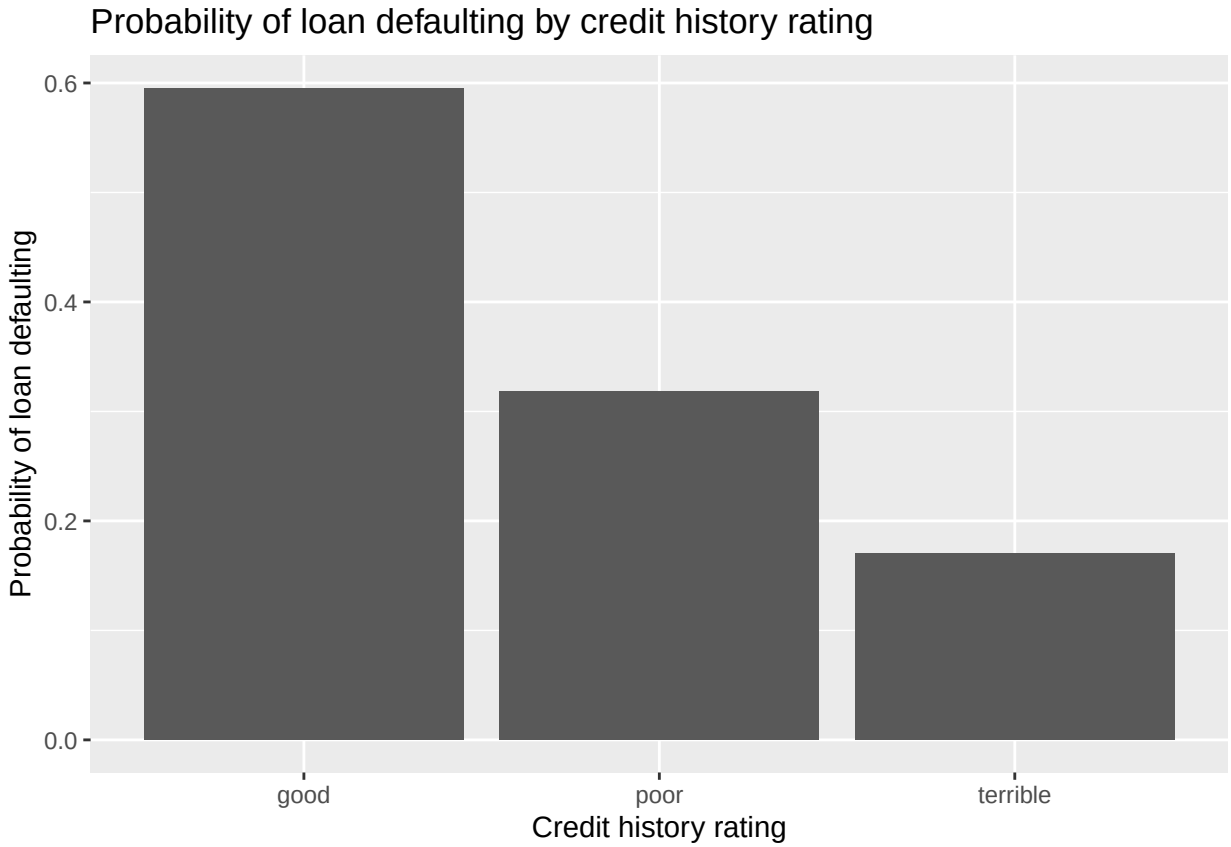
For K = 36, RMSE is:

```
## [1] 80823.84
```

Because the training set and testing set are randomly regenerating each time I knit my document, there is some variance in which K results in the lowest RMSE. The relative improvement of my KNN model over the professor's moderate model is:

```
## [1] 0.788197
```

Classification and retrospective sampling




```
##
## Call:
## glm(formula = Default ~ duration + amount + installment + age +
##     history + purpose + foreign, family = binomial, data = german_credit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3464  -0.8050  -0.5751   1.0250   2.4767
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -7.075e-01  4.726e-01  -1.497  0.13435
## duration             2.526e-02  8.100e-03   3.118  0.00182 **
## amount               9.596e-05  3.650e-05   2.629  0.00856 **
## installment          2.216e-01  7.626e-02   2.906  0.00366 **
## age                 -2.018e-02  7.224e-03  -2.794  0.00521 **
## historypoor         -1.108e+00  2.473e-01  -4.479 7.51e-06 ***
## historyterrible     -1.885e+00  2.822e-01  -6.679 2.41e-11 ***
## purposeedu           7.248e-01  3.707e-01   1.955  0.05058 .
## purposegoods/repair  1.049e-01  2.573e-01   0.408  0.68346
## purposenewcar        8.545e-01  2.773e-01   3.081  0.00206 **
## purposeusedcar      -7.959e-01  3.598e-01  -2.212  0.02694 *
## foreigngerman       -1.265e+00  5.773e-01  -2.191  0.02849 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.7  on 999  degrees of freedom
## Residual deviance: 1070.0  on 988  degrees of freedom
## AIC: 1094
##
## Number of Fisher Scoring iterations: 4
```

It appears that having a poor or terrible credit history tends to lead to higher probability of a customer paying off their loan. I would guess that the reason for this could be that the bank only approves loans of small amounts, high interest rates, and/or some sort of collateral for those with poor and terrible credit scores, but is willing to offer riskier loans to those with good credit scores.

Given that over half of those with good credit scores end up defaulting on their loans, I think that the bank should mitigate some of the risk within that group by limiting the amounts of their loans, increasing their interest rates, or requiring collateral. If they mitigate some of the risk of the good credit score group using the methods that they are likely using for the lower credit score groups, that could help to decrease the amount of people that are defaulting on their loans.

Children and hotel reservations

Model Building

Baseline 1: a small model using only market segment, number of adults, type of customer, and whether or not the customers are repeat guests.

Confusion matrix:

```
##    yhat
## y      0
##   0 41365
##   1  3635
```

Out-of-sample performance, based on confusion matrix:

```
## [1] 0.9192222
```

This model has a 92% out-of-sample performance.

Baseline 2: a big model that uses all possible predictors except for arrival date.

Confusion matrix:

```
##    yhat
## y       0     1
##   0 40800   565
##   1  2357  1278
```

Out-of-sample performance, based on confusion matrix:

```
## [1] 0.9350667
```

I found that the model including every predictor except arrival date had a 93.5% out-of-sample performance.

The best linear model I can build

```
##
## Call:
## lm(formula = children ~ hotel + lead_time + meal + stays_in_weekend_nights +
##     stays_in_week_nights + adults + market_segment + distribution_channel +
##     is_repeated_guest + log(1 + previous_bookings_not_canceled) +
##     reserved_room_type + assigned_room_type + deposit_type +
##     days_in_waiting_list + average_daily_rate + customer_type +
##     total_of_special_requests + adults_squared + weekend_squared +
##     week_squared, data = hotels_dev, family = binomial)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91656 -0.08396 -0.03744  0.00949  1.12451
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -1.751e-01  2.453e-02  -7.138 9.60e-13
## hotelResort_Hotel                   -3.725e-02  2.827e-03 -13.180  < 2e-16
## lead_time                            3.992e-05  1.431e-05   2.789 0.005297
## mealFB                               5.016e-02  1.710e-02   2.932 0.003366
## mealHB                              -5.533e-04  3.676e-03  -0.151 0.880369
```

```
## mealSC                                      -5.475e-02  4.241e-03 -12.909  < 2e-16
## mealUndefined                                2.256e-02  1.079e-02   2.090 0.036630
## stays_in_weekend_nights                      5.217e-03  2.623e-03   1.989 0.046737
## stays_in_week_nights                         1.207e-03  1.257e-03   0.960 0.336963
## adults                                       1.064e-01  9.772e-03  10.884  < 2e-16
## market_segmentComplementary                  6.625e-02  2.675e-02   2.477 0.013255
## market_segmentCorporate                      5.522e-02  2.269e-02   2.433 0.014966
## market_segmentDirect                         5.057e-02  2.444e-02   2.069 0.038571
## market_segmentGroups                         6.540e-02  2.375e-02   2.753 0.005899
## market_segmentOffline_TA/TO                  7.325e-02  2.380e-02   3.077 0.002090
## market_segmentOnline_TA                      6.809e-02  2.375e-02   2.867 0.004147
## distribution_channelDirect                   1.286e-02  1.022e-02   1.258 0.208330
## distribution_channelGDS                     -7.918e-02  2.495e-02  -3.173 0.001509
## distribution_channelTA/TO                   -9.283e-03  8.504e-03  -1.092 0.275023
## is_repeated_guest                           -9.947e-03  7.853e-03  -1.267 0.205278
## log(1 + previous_bookings_not_canceled) -2.429e-02  5.195e-03  -4.676 2.93e-06
## reserved_room_typeB                          2.273e-01  1.351e-02  16.823  < 2e-16
## reserved_room_typeC                          5.343e-01  1.438e-02  37.165  < 2e-16
## reserved_room_typeD                         -6.612e-02  4.308e-03 -15.349  < 2e-16
## reserved_room_typeE                         -3.001e-02  7.745e-03  -3.875 0.000107
## reserved_room_typeF                          3.032e-01  1.155e-02  26.261  < 2e-16
## reserved_room_typeG                          4.289e-01  1.559e-02  27.510  < 2e-16
## reserved_room_typeH                          6.381e-01  2.911e-02  21.920  < 2e-16
## reserved_room_typeL                         -8.798e-02  1.648e-01  -0.534 0.593343
## assigned_room_typeB                          1.473e-02  9.083e-03   1.621 0.104946
## assigned_room_typeC                          1.022e-01  8.323e-03  12.276  < 2e-16
## assigned_room_typeD                          6.528e-02  3.738e-03  17.465  < 2e-16
## assigned_room_typeE                          5.762e-02  6.913e-03   8.335  < 2e-16
## assigned_room_typeF                          6.751e-02  9.904e-03   6.817 9.40e-12
## assigned_room_typeG                          1.005e-01  1.365e-02   7.366 1.78e-13
## assigned_room_typeH                          8.300e-02  2.497e-02   3.324 0.000888
## assigned_room_typeI                          1.083e-01  1.648e-02   6.572 5.01e-11
## assigned_room_typeK                          7.911e-02  1.865e-02   4.241 2.23e-05
## deposit_typeNon_Refund                       3.049e-02  3.004e-02   1.015 0.310201
## deposit_typeRefundable                       2.518e-02  2.582e-02   0.975 0.329518
## days_in_waiting_list                        -5.526e-05  7.708e-05  -0.717 0.473432
## average_daily_rate                           9.262e-04  2.962e-05  31.270  < 2e-16
## customer_typeGroup                           4.051e-03  1.423e-02   0.285 0.775834
## customer_typeTransient                       1.858e-02  6.146e-03   3.023 0.002501
## customer_typeTransient-Party                -1.953e-02  6.636e-03  -2.944 0.003246
## total_of_special_requests                    3.331e-02  1.499e-03  22.224  < 2e-16
## adults_squared                              -4.269e-02  2.729e-03 -15.644  < 2e-16
## weekend_squared                             -4.307e-04  1.128e-03  -0.382 0.702565
## week_squared                                -8.144e-05  1.817e-04  -0.448 0.653984
##
## (Intercept)                                 ***
## hotelResort_Hotel                           ***
## lead_time                                   **
## mealFB                                      **
## mealHB
## mealSC                                      ***
## mealUndefined                               *
## stays_in_weekend_nights                     *
## stays_in_week_nights
```

```
## adults                                    ***
## market_segmentComplementary               *
## market_segmentCorporate                   *
## market_segmentDirect                      *
## market_segmentGroups                      **
## market_segmentOffline_TA/TO               **
## market_segmentOnline_TA                   **
## distribution_channelDirect
## distribution_channelGDS                   **
## distribution_channelTA/TO
## is_repeated_guest
## log(1 + previous_bookings_not_canceled) ***
## reserved_room_typeB                       ***
## reserved_room_typeC                       ***
## reserved_room_typeD                       ***
## reserved_room_typeE                       ***
## reserved_room_typeF                       ***
## reserved_room_typeG                       ***
## reserved_room_typeH                       ***
## reserved_room_typeL
## assigned_room_typeB
## assigned_room_typeC                       ***
## assigned_room_typeD                       ***
## assigned_room_typeE                       ***
## assigned_room_typeF                       ***
## assigned_room_typeG                       ***
## assigned_room_typeH                       ***
## assigned_room_typeI                       ***
## assigned_room_typeK                       ***
## deposit_typeNon_Refund
## deposit_typeRefundable
## days_in_waiting_list
## average_daily_rate                        ***
## customer_typeGroup
## customer_typeTransient                    **
## customer_typeTransient-Party              **
## total_of_special_requests                 ***
## adults_squared                            ***
## weekend_squared
## week_squared
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2328 on 44951 degrees of freedom
## Multiple R-squared:  0.2707, Adjusted R-squared:  0.2699
## F-statistic: 347.6 on 48 and 44951 DF,  p-value: < 2.2e-16
```

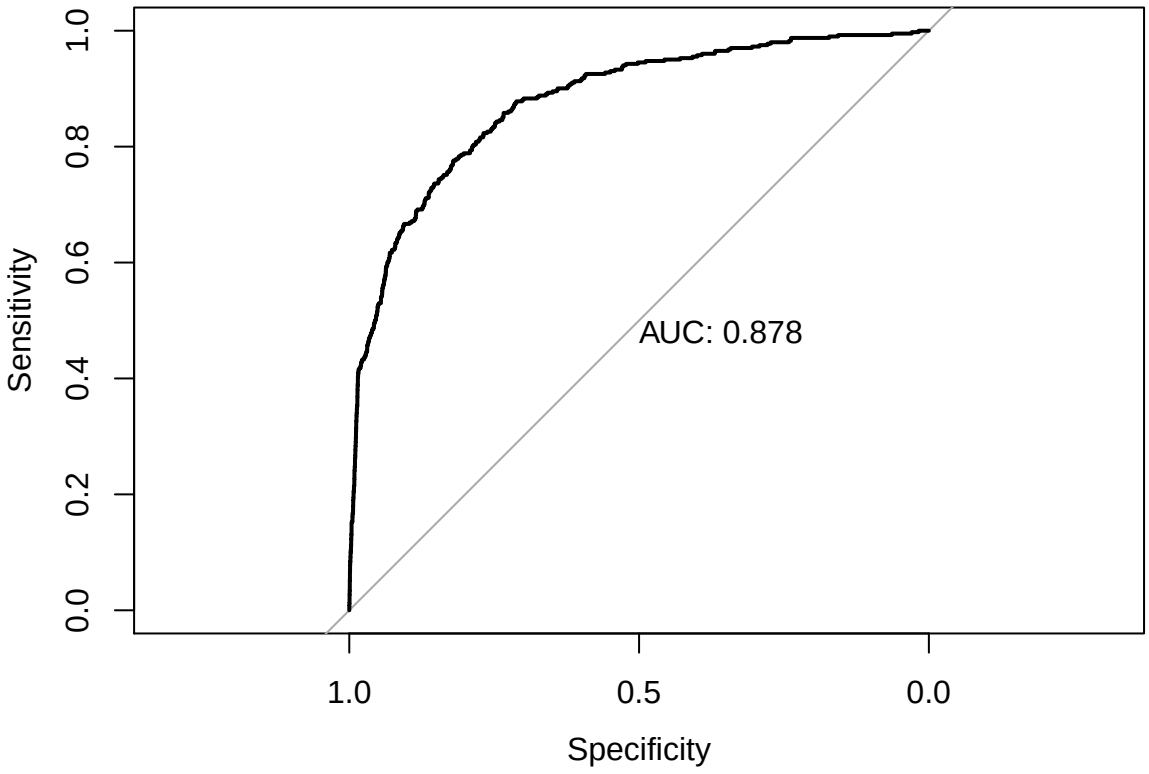
I found it very difficult to come up with a model that was any better than the second baseline model. My model has only slightly higher out-of-sample performance.

```
##    yhat
## y       0     1
```

```
##   0 40836   529
##   1  2337  1298
```

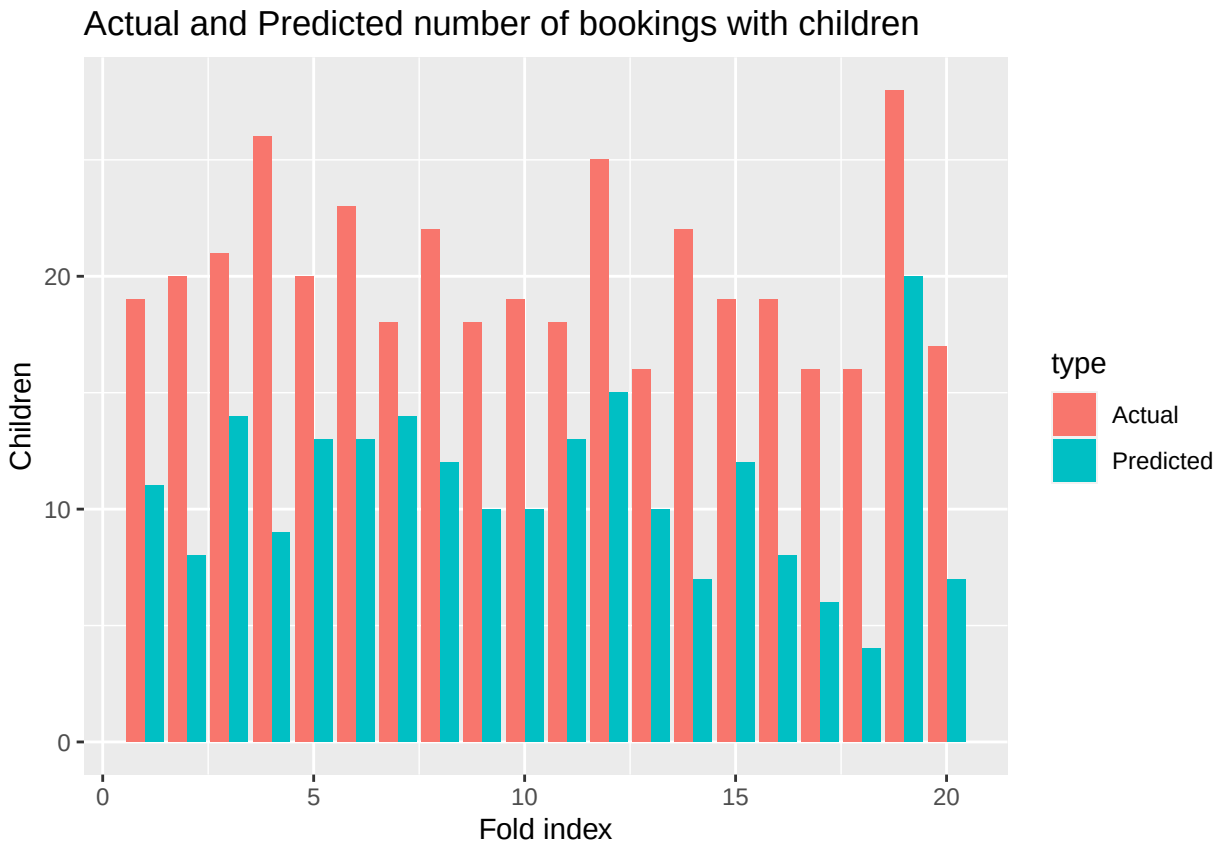
```
## [1] 0.9363111
```

Model Validation: Step 1



Model Validation: Step 2

For the next part, I created 20 random folds of data (all with 250 bookings each, except for one that only has 249 bookings). Then, I used my linear model to estimate whether or not we would expect to see children on each booking sample. I compared the sum of modeled number of bookings with children, with the actual number of bookings with children in each fold, and created a bar graph that shows both expected and actual number of bookings with children in each fold.



It looks like my model consistently underestimates the number of bookings with children within each fold by about an average of about 5.