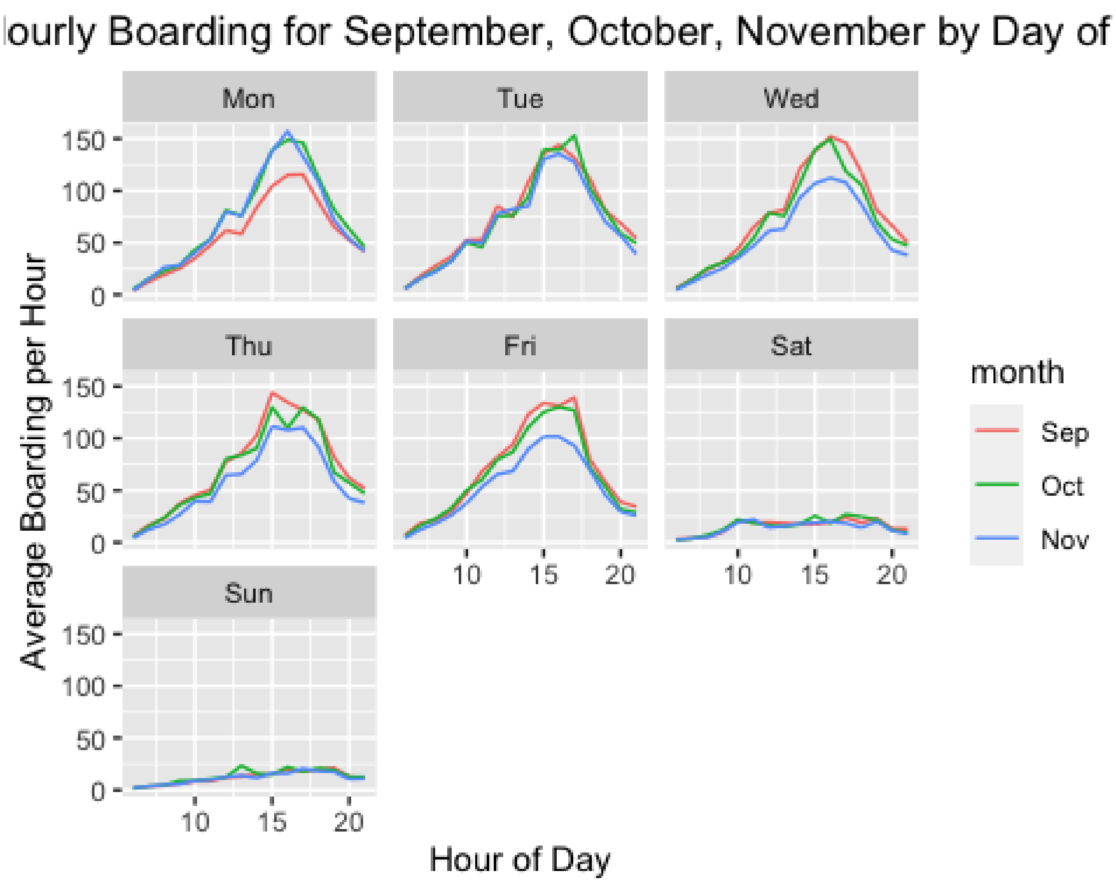
Data Mining Assignment 2

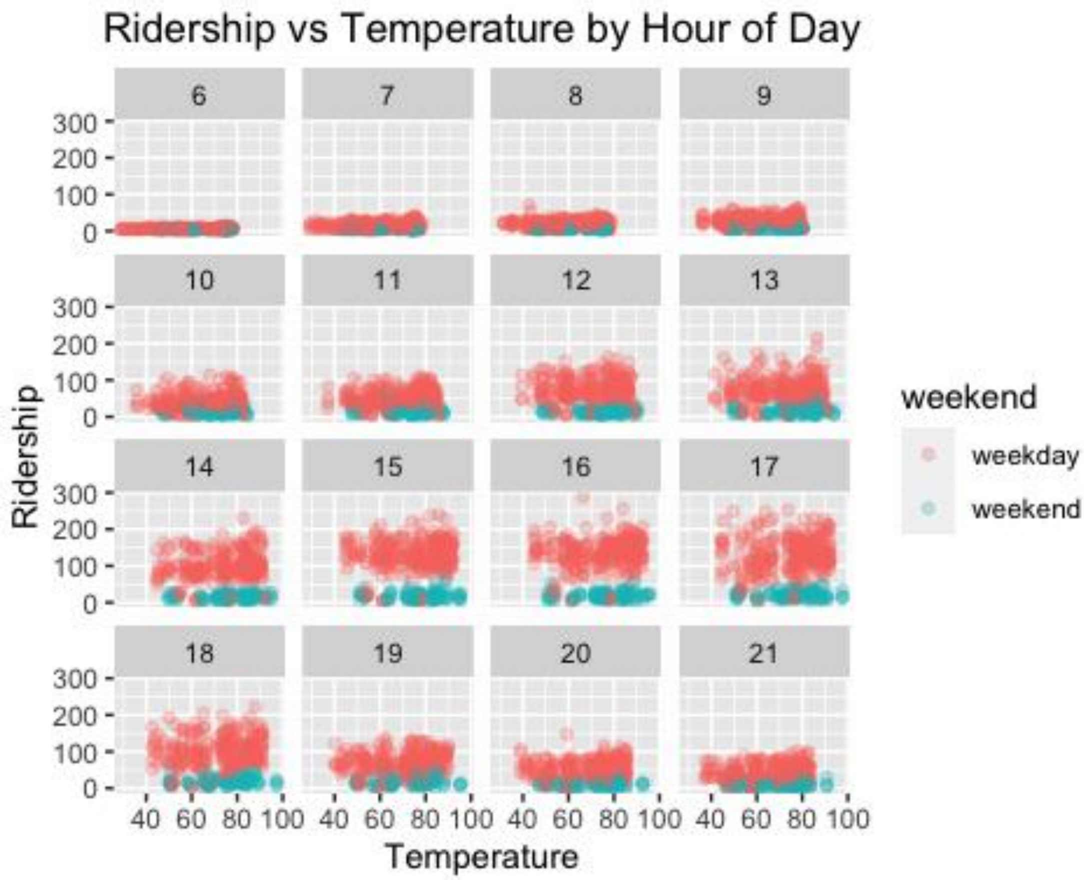
Patrick Chase

3/6/2021

1.



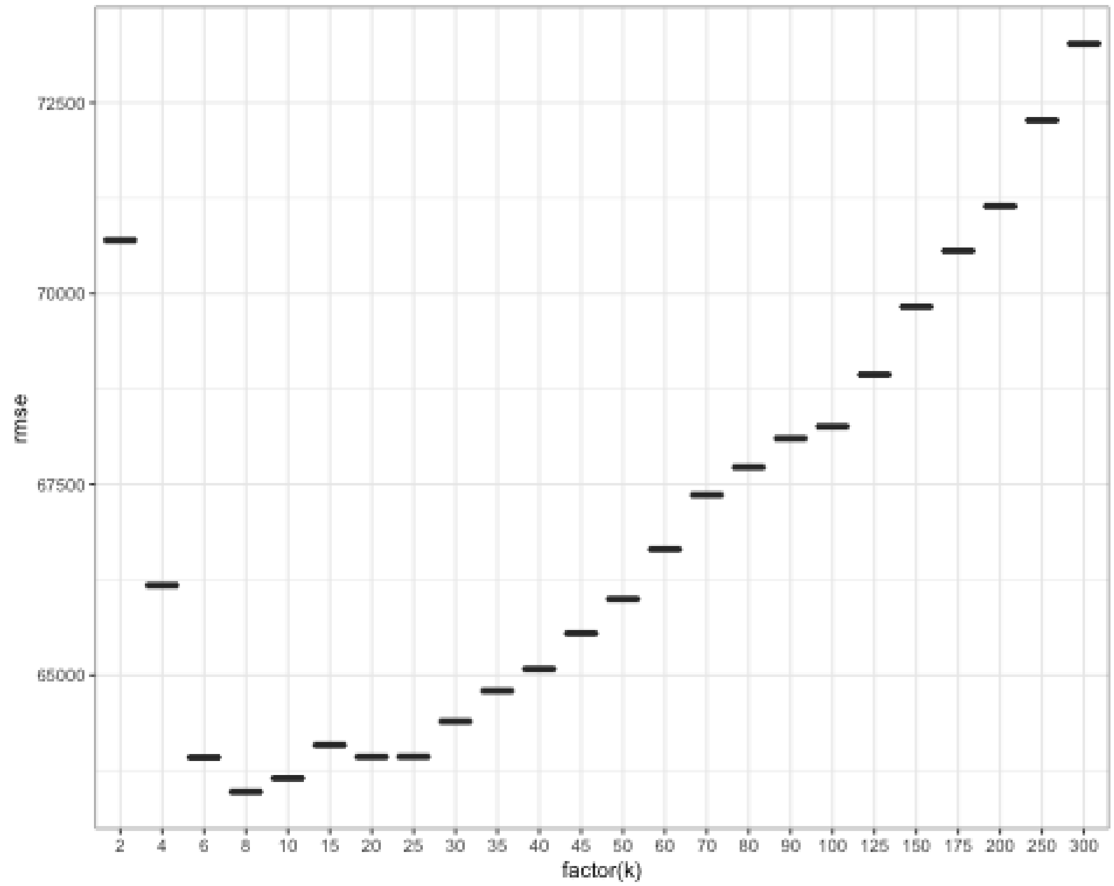
Plot 1a is showing the average hourly boarding for September, October, and November by day of the week. On weekdays, we see a clear peak around 4:00 pm every day however, on weekends ridership remains relatively flat. A possible reason we see lower ridership on Mondays for September is probably because of the Labor Day holiday that falls on Monday. This caused 1 Monday in September to have drastically lower boardings, thus causing a decrease in it's Monday's average for September. Similarly, for Wednesday through Friday in November, we can see a decrease likely because of the Thanksgiving holiday.



Plot 1b is showing boardings by temprature controlling for hour of the day and whether it is a weekday or a weekend. When holding hour of the day and weekend status constant, there doesn't seem to be a clear relationship between ridership and temperature. The fluctuations shown in these scatter plots could just as easily be explained by the normal commuting patterns of students.

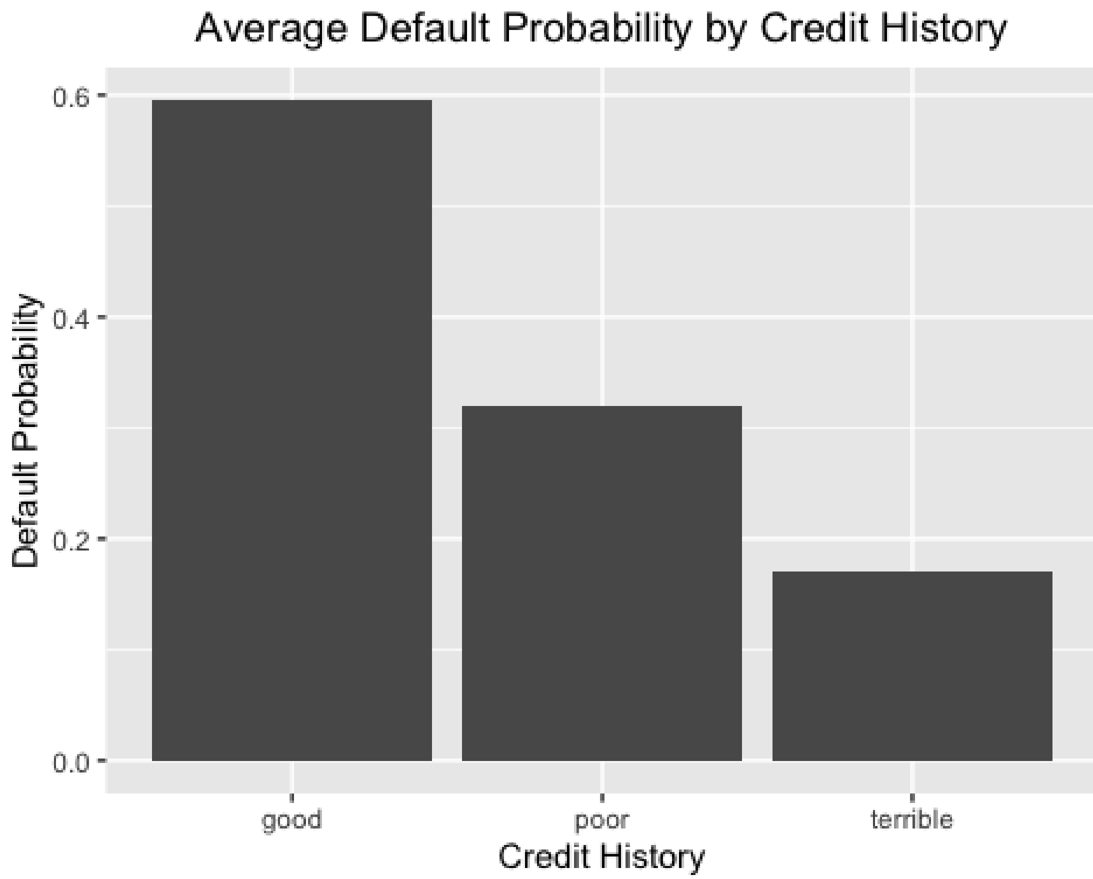
2.

```
##        1        1
## 66624.01 63106.94
```



My model seems to outperform lm_medium by about 5%-6%. That said the KNN model with k = 8 outperforms both by 15%-20%. Depending on the geographic market you're in, I'd say we need to rely heavily on the location because houses in the same neighborhood are generally going to relatively similar. This is especially true if you're looking at suburban neighborhoods that were all developed at the same time by the same development company. This strategy would likely have to change if we were in a rural area but we should look at the going rate of similar houses, particularly those that are physical close to the proprety in question, and use that to determine our valuation.

3.



With this data set, it would seem that the credit history rating is predicting in the exact opposite direction it was originally intended. Those with "terrible" credit history are the least likely to default. This is likely due to sampling bias. If very few individuals with "terrible" credit are being granted lines of credit then in all likelihood the bank has a good reason to give them credit that trumps their credit history. If the bank wanted to improve it's sampling method, it should be collecting attributes of those it denies credit to as well. This will expand the sample and probably increase the quality of those individuals

4.

4.1

```
##       AICs     Model
## 1 13978.44 step_mod2
## 2 18975.02 small_mod
## 3 13983.52   big_mod
```

4.2

```
##        3337        936        704       4782        866          4
612
##  -5.28115960 -5.08613624 -5.03432747 -2.74234470 -2.80011739 -1.15347
958
##        4180       1034       2114       2428       3648          3
835
##  -0.80081695 -2.18203396 -5.20321329 -5.26942791 -3.71315511 -3.12964
587
##          14       3235        136       1655        799
241
##  -3.16877785 -4.30755421 -6.34465539 -2.66312610 -2.18998536 -24.17670
335
##         222       2061       1777         12       2992          2
348
##  -2.96035289 -6.25508342 -1.86978976 -14.76738416 -0.92943926 -3.67285
363
##        1565       4443        963       1051       3323          2
112
##  -2.54560352 -2.56323619 -3.71722159 -3.18171531 -4.40545120 -2.63769
532
##        4891       1990       3331        200       3399          1
725
##  -2.65659964  0.68156618 -0.73393181 -3.75868771 -4.44427222 -2.39397
132
##        2541       3219       1340       2181        519          3
111
##  -3.92300175 -3.71730984  0.36018597 -2.67200314 -3.76062329 -2.87620
395
##        1228       3950       2303       4366       2044          4
461
##  -4.19166915 -3.38205329 -3.38887208 -6.05235710 -0.84858200 -2.89718
849
##        2446       2031        788       3029       2499          2
553
##  -4.38132031 -3.49988682 -4.64560216 -3.31769821  1.24335508 -6.98450
689
##        1728       1318       4688       1253       3196
76
##   0.25996901 -3.57361895 -1.64436466 -3.59039515 -4.23715111 -3.64174
404
##        4969       4008       1890        240       2487
648
##  -3.86782756 -2.66438261 -3.01759843 -3.39820319 -3.61997631 -3.63995
138
##        2875       3156        113       4608       3576          1
968
##  -5.14623838 -2.78296651 -6.50995381 -4.68850009 -2.98879526 -2.90930
394
```

```
##         101         3508          267          498          952            3
976
##  -4.90350459  -3.64489702  -3.82803216  -1.68783441  -3.44688232   2.17658
607
##        2645         4920         2299         4241          878            3
972
##  -4.36612691  -3.30782895  -2.82675919  -2.44213019  -5.34693798  -5.76003
501
##        4339         1125           83         2616         4875            4
505
##  -2.81966881  -2.08929425  -1.01554250  -4.63741826  -2.78123522  -2.75783
535
##         618         3758         3225         3611          204            1
823
##  -1.45952285  -3.20193162  -1.79507042  -3.46331681  -2.01746970  -3.59447
688
##        2189         2898          345         2351          841            1
092
##  -3.97746394  -3.19676414  -2.13205725  -3.28929781  -4.30576596  -2.45239
960
##        3528         3604         1734         3588          161
99
##  -2.66975458  -2.19527564 -16.09466470  -2.74094725  -2.91495647  -2.82911
977
##        3901         1801         2641         1613         3341            2
389
##  -2.90645009  -5.12301106  -3.86995528  -3.32249740  -0.67920325  -2.93644
132
##         391          340         4786         3599         4150            3
842
##  -6.29673453  -4.78263698  -3.78844801  -4.28788428  -3.89489497  -2.32584
064
##        3145         1628         4834         3517         3150            1
722
##  -2.63143691  -2.55736266  -4.88228042  -3.64938878  -1.62745339  -4.13044
672
##        1192          326          149         2611         1066            2
505
##  -3.28311855  -6.82586491  -2.03939428  -3.75344177  -2.43312120  -2.62601
990
##        1414         4693         1834          163         2689            2
520
##  -5.47737117  -4.73377269   0.05180186  -2.37929263  -3.05332741  -5.87688
541
##        2804          820         3801         2109         4689            2
720
##  -1.95426611  -3.41916507  -2.77118974  -3.35164339  -4.06009049  -2.81364
258
##        1822         4501         3075         3283         1690            4
592
```

```
##  -2.85533801  -4.22353697  -3.07915969  -2.39631778  -4.84509407  -1.50405
039
##          503         1380         1342         1789         4039
90
##  -7.12024694 -16.20309101  -3.07287219  -3.20796237  -5.21599303  -0.29519
650
##         1116         1129         3043         1038         3360           1
385
##  -5.41473736  -5.94319992  -2.37709830  -2.11086870  -1.98903131  -2.66420
784
##         2184         1358         4261         3059         1277
817
##  -2.78960190  -2.49612027  -2.52332212  -1.14513194  -3.36129087  -3.31715
395
##         4074          729         2429         4775         1341           3
626
##  -3.15857414  -7.49857178  -5.22137610  -2.65673235  -3.36880271  -4.20603
176
##         3979         4251         3401           88         3571           1
093
##  -3.04194989  -4.64777263  -2.80659256  -0.76755304  -4.12821147 -12.48285
753
##          722         4238         1383         3954          753           4
176
##  -3.34853972  -2.42261931  -2.64034820  -4.01011596  -2.09224032  -3.55106
444
##         2599         2277          540         4718          951           3
458
##  -2.86598766  -0.39773140  -4.11626680  -4.51179158  -1.23517657  -3.71632
746
##         3660         1971         1119         2425         4151           3
695
##  -4.17176301  -3.51334249  -0.66180310  -3.04686086  -1.70222643  -2.48708
516
##         1933         4799         2219         1958         3100           1
496
##  -3.05042637  -2.81442789  -1.94367613  -3.55287384  -2.65659964  -3.34185
431
##         2531         1367         1820         3383         2827           4
701
##  -4.80069279  -2.55369961  -3.99846494  -4.04243838  -2.46354076  -3.91343
587
##         3128         3144         4056         1833         4122           1
028
##  -4.63216630  -2.46490783  -1.99666492  -2.71151642  -3.43645845  -4.03890
165
##         3918         2677         3467         3444         4952           2
730
##  -2.01365718  -2.03371444  -2.04068956  -2.98114001  -5.90840855  -1.85395
206
```

```
##         3306         2831          311         3404         2260           3
688
##  -1.08086423  -5.77784834  -7.71198345  -4.91236911  -2.27641837   2.06938
484
##         1356         4953         4935         1484          218           2
417
## -17.40129408  -3.32449961  -3.16781342  -3.21545906  -3.73156677  -3.06030
888
##          257         2462         1961          564         1156           1
641
##   0.19696009  -3.36520532  -2.73803651  -6.18330467  -1.59139706  -2.81985
036
##         4616         1939         2232          317          447           4
307
##  -3.10306749  -5.30352767  -4.19606525  -2.51658396  -2.80843043  -5.09550
679
##         4146          941         2066
##  -3.30148157  -4.76782507  -2.32852178

##      3337       936       704      4782       866      4612
## -5.281160 -5.086136 -5.034327 -2.742345 -2.800117 -1.153480
```

So I've tried this a bunch of different ways and I haven't been able to figure out how to arrive at the asked for outputs. After consulting with Rui, I'm pretty sure I've done the 20-fold validation correctly but I'm not sure how to move from that into a neat table to get the predicted values, summed probabilities, and the actual bookings into a neat table.