Data Science Career Track

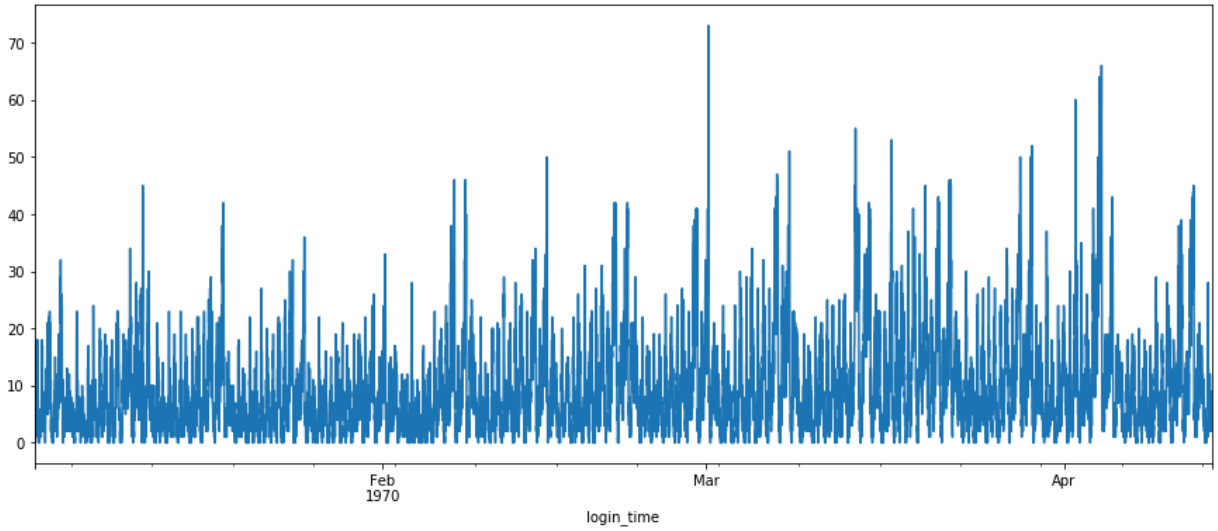
TAKE HOME CHALLENGE -

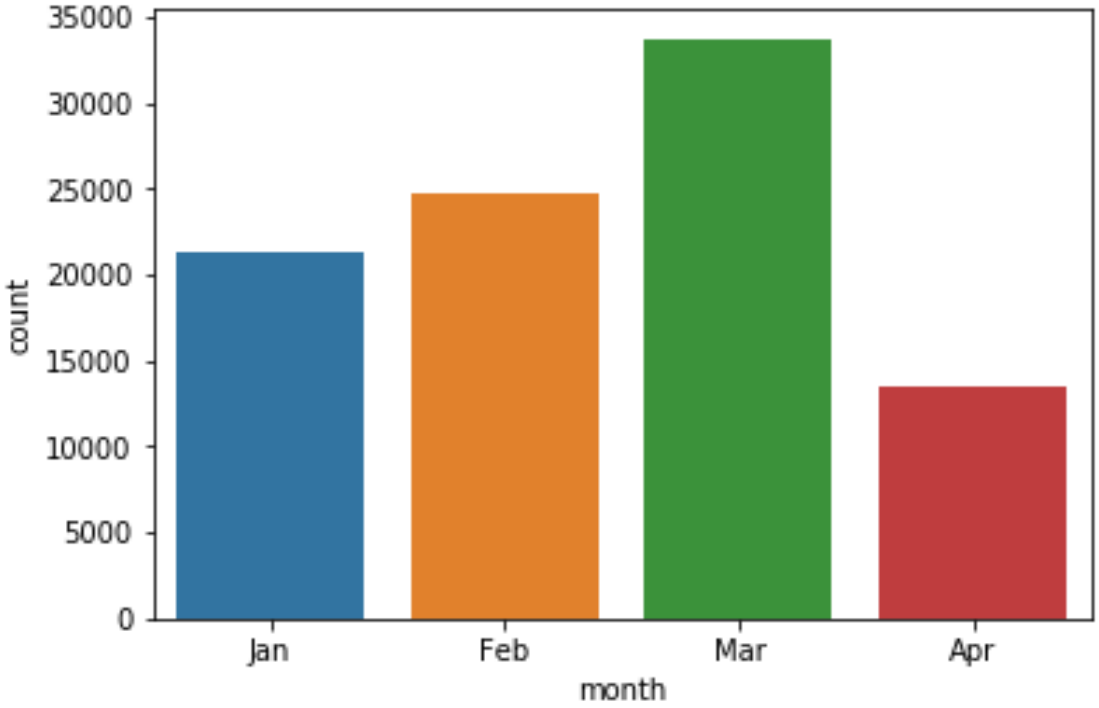
Ultimate Report

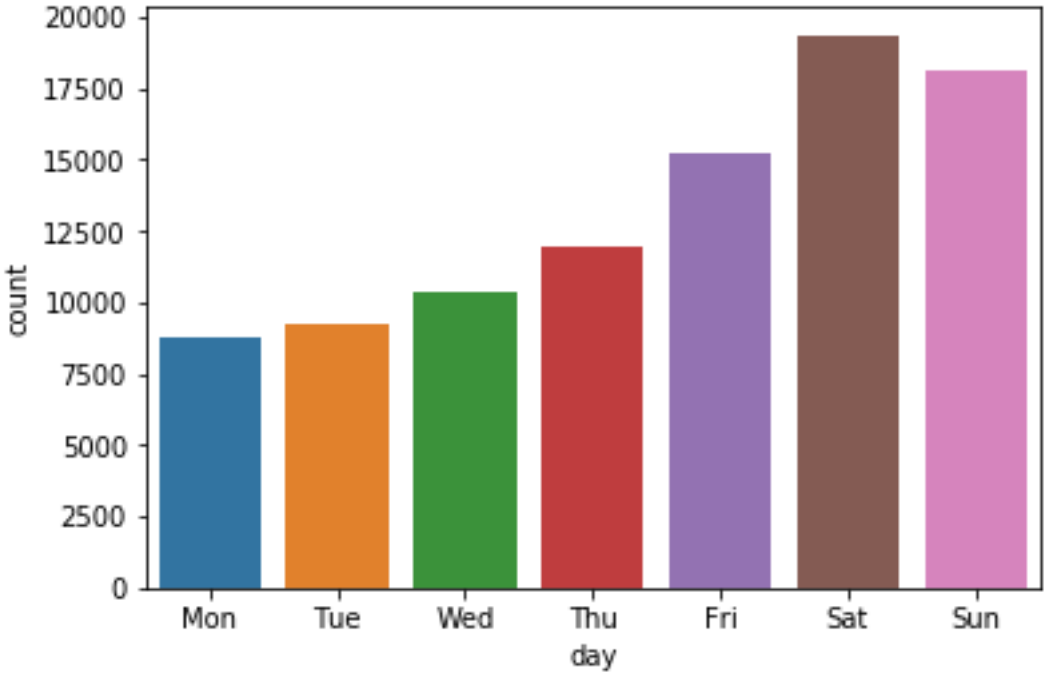
by Edward Franke

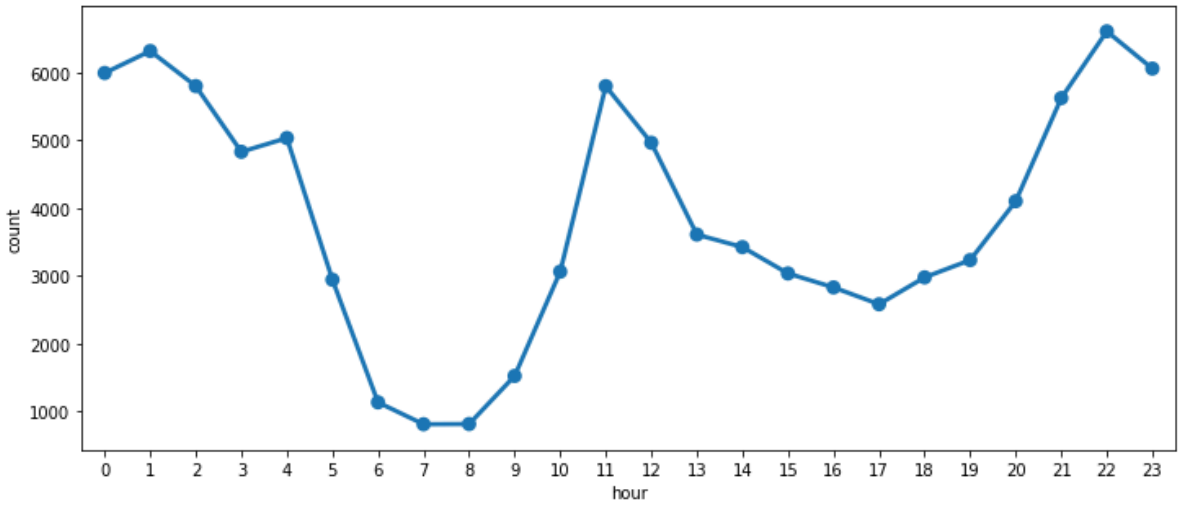
06/02/2019

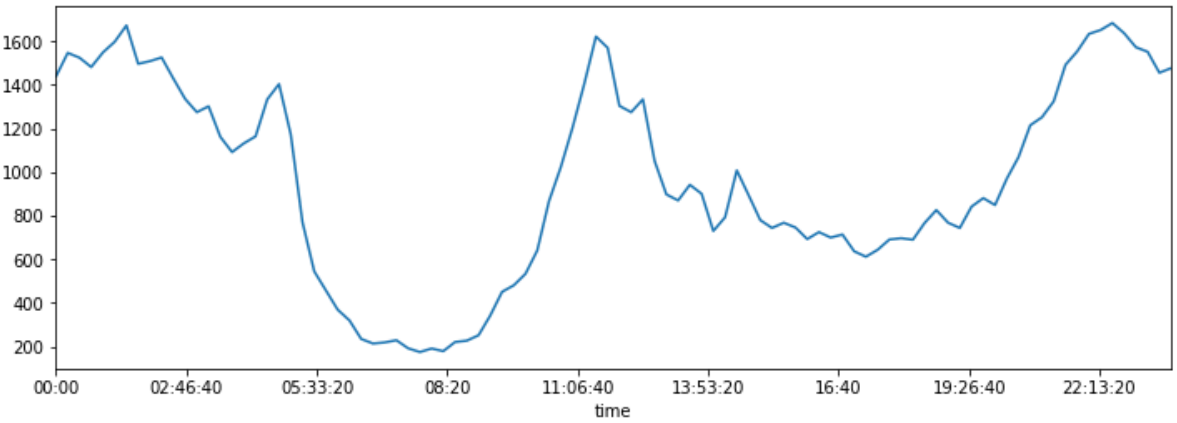
Part 1 - Exploratory data analysis


70
60
50
40
30
20
10
0
Feb
1970
Mar
Apr
login_time


35000
30000
25000
20000
15000
10000
5000
0
count
Jan
Feb
Mar
Apr
month


20000
17500
15000
12500
10000
7500
5000
2500
0
count
Mon
Tue
Wed
Thu
Fri
Sat
Sun
day





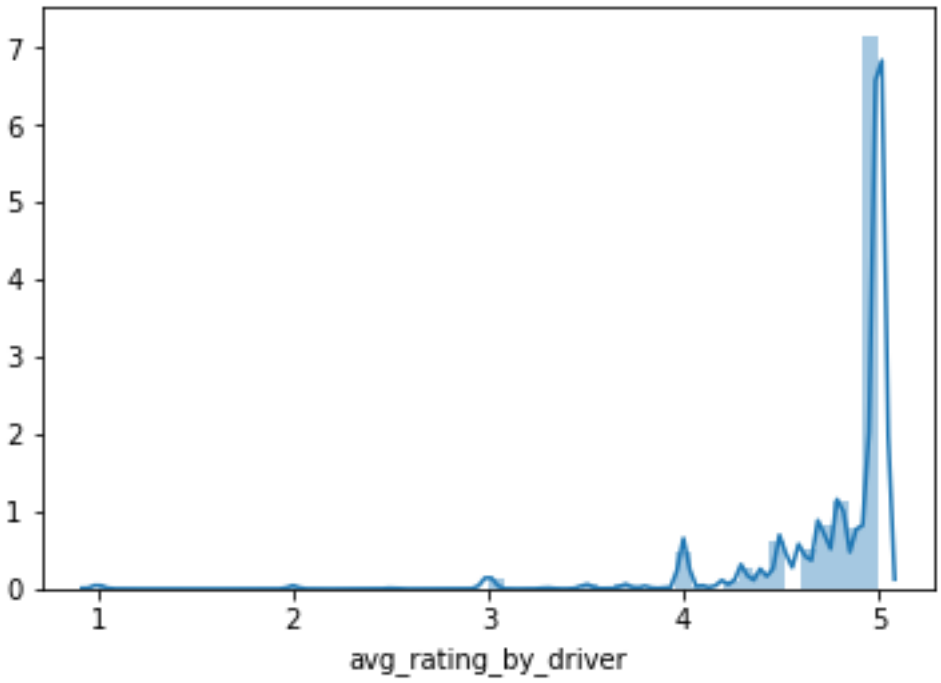
Part 2 - Experiment and metrics design

	avg_dist	avg_rating_by_driver	avg_rating_of_driver	avg_surge	city	last_trip_date	phone	signup_date	surge_pct	trips_in_first_30_days	ultimate_black_user	weekday_pct
0	3.67	5.0	4.7	1.10	King's Landing	2014-06-17	iPhone	2014-01-25	15.4	4	True	46.2
1	8.26	5.0	5.0	1.00	Astapor	2014-05-05	Android	2014-01-29	0.0	0	False	50.0
2	0.77	5.0	4.3	1.00	Astapor	2014-01-07	iPhone	2014-01-06	0.0	3	False	100.0
3	2.36	4.9	4.6	1.14	King's Landing	2014-06-29	iPhone	2014-01-10	20.0	9	True	80.0
4	3.13	4.9	4.4	1.19	Winterfell	2014-03-15	Android	2014-01-27	11.8	14	False	82.4

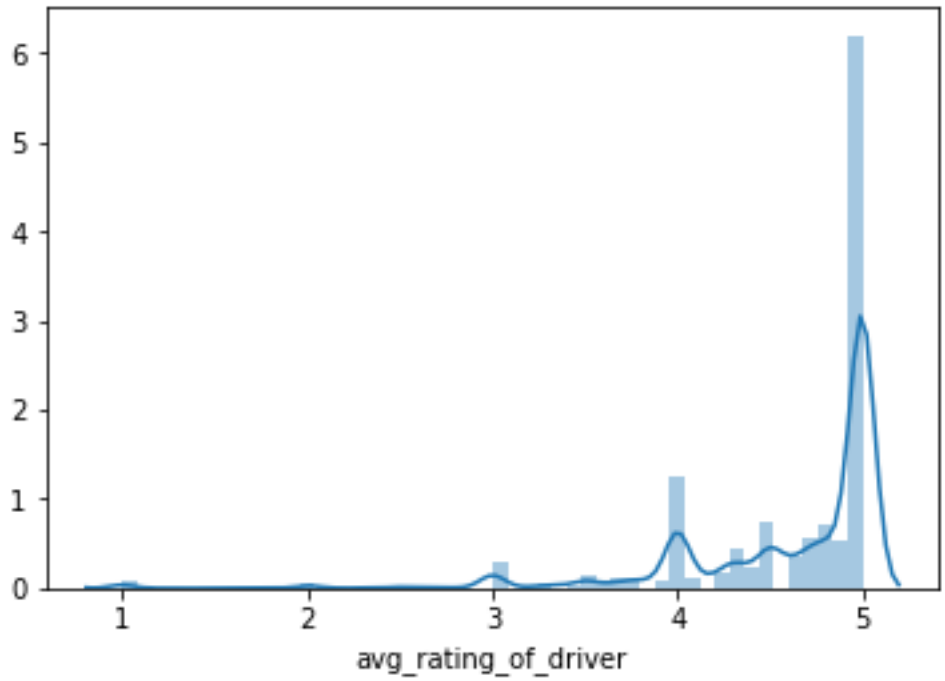
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 12 columns):
avg_dist                  50000 non-null float64
avg_rating_by_driver      49799 non-null float64
avg_rating_of_driver      41878 non-null float64
avg_surge                 50000 non-null float64
city                      50000 non-null object
last_trip_date            50000 non-null object
phone                     49604 non-null object
signup_date               50000 non-null object
surge_pct                 50000 non-null float64
trips_in_first_30_days    50000 non-null int64
ultimate_black_user       50000 non-null bool
weekday_pct               50000 non-null float64
dtypes: bool(1), float64(6), int64(1), object(4)
memory usage: 4.2+ MB
```

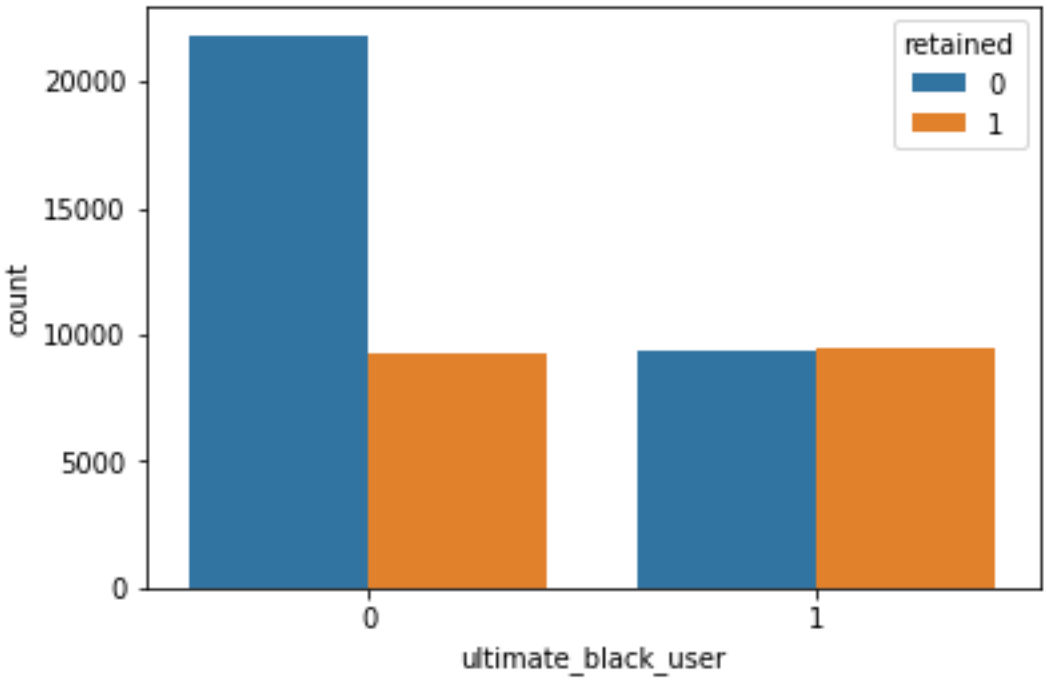
```
iPhone      34582
Android     15022
Name: phone, dtype: int64
```

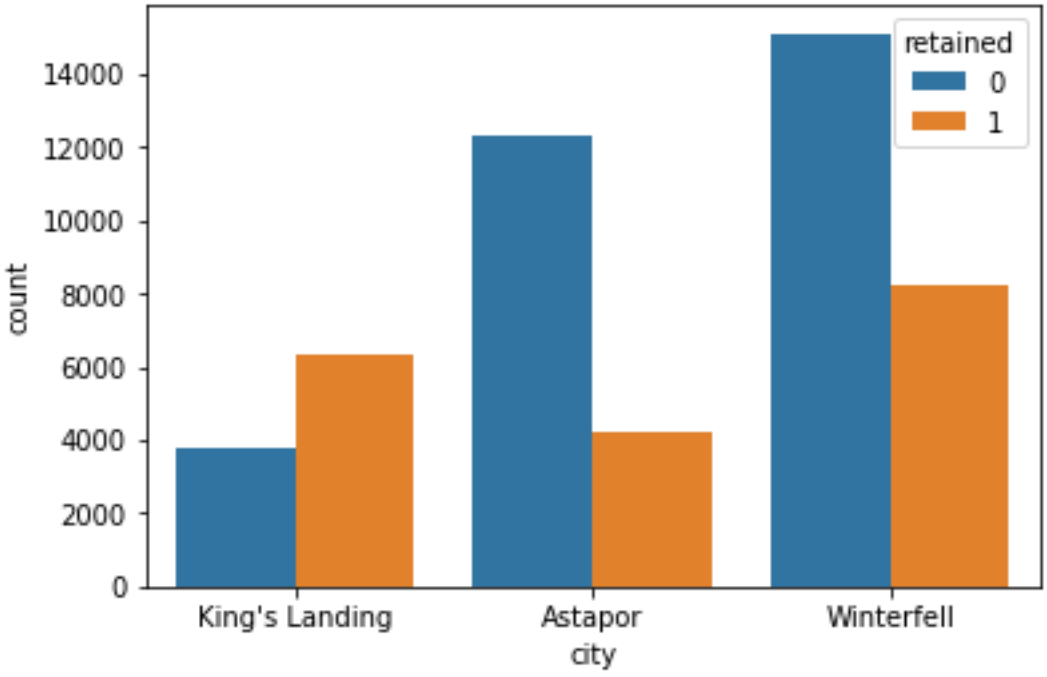
```
count    49799.000000
mean         4.778158
std          0.446652
min          1.000000
25%          4.700000
50%          5.000000
75%          5.000000
max          5.000000
Name: avg_rating_by_driver, dtype: float64
```

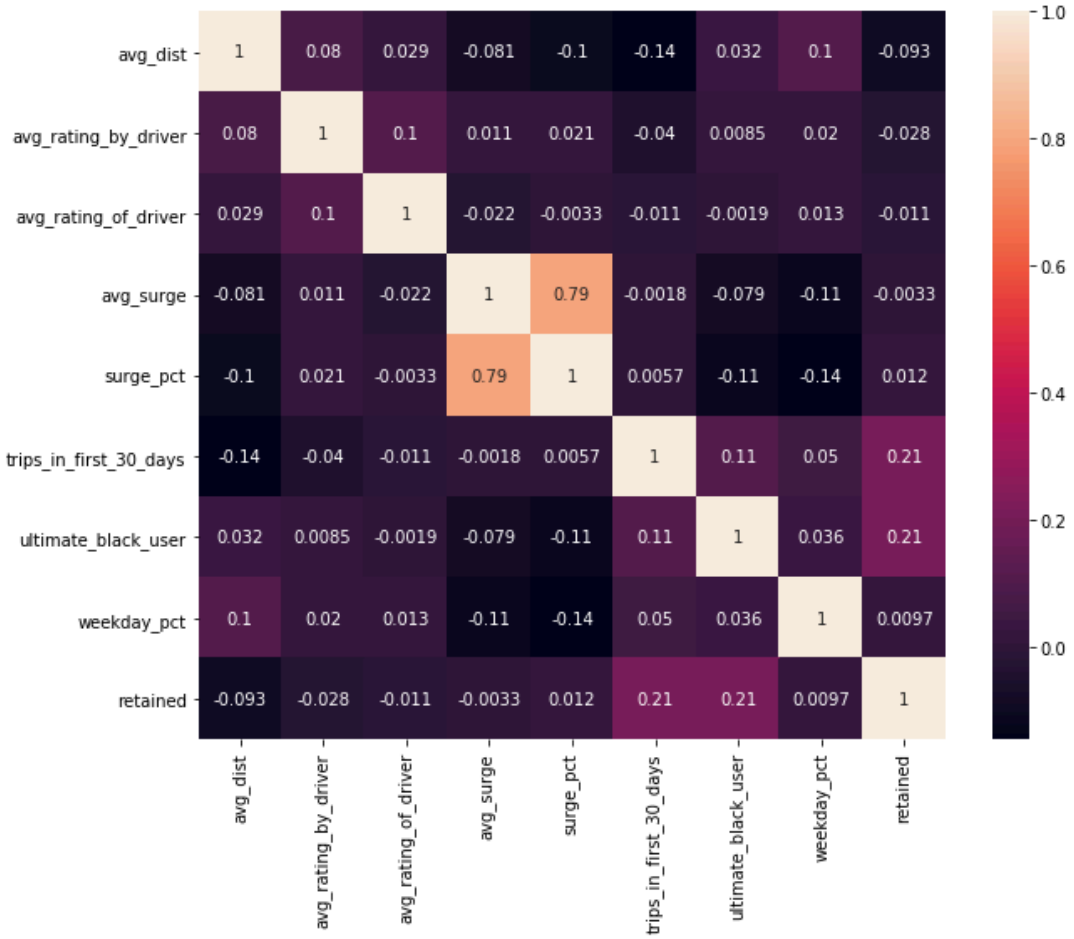


```
count    41878.000000
mean         4.601559
std          0.617338
min          1.000000
25%          4.300000
50%          4.900000
75%          5.000000
max          5.000000
Name: avg_rating_of_driver, dtype: float64
```




retained
0
1
20000
15000
10000
5000
0
count
0
1
ultimate_black_user

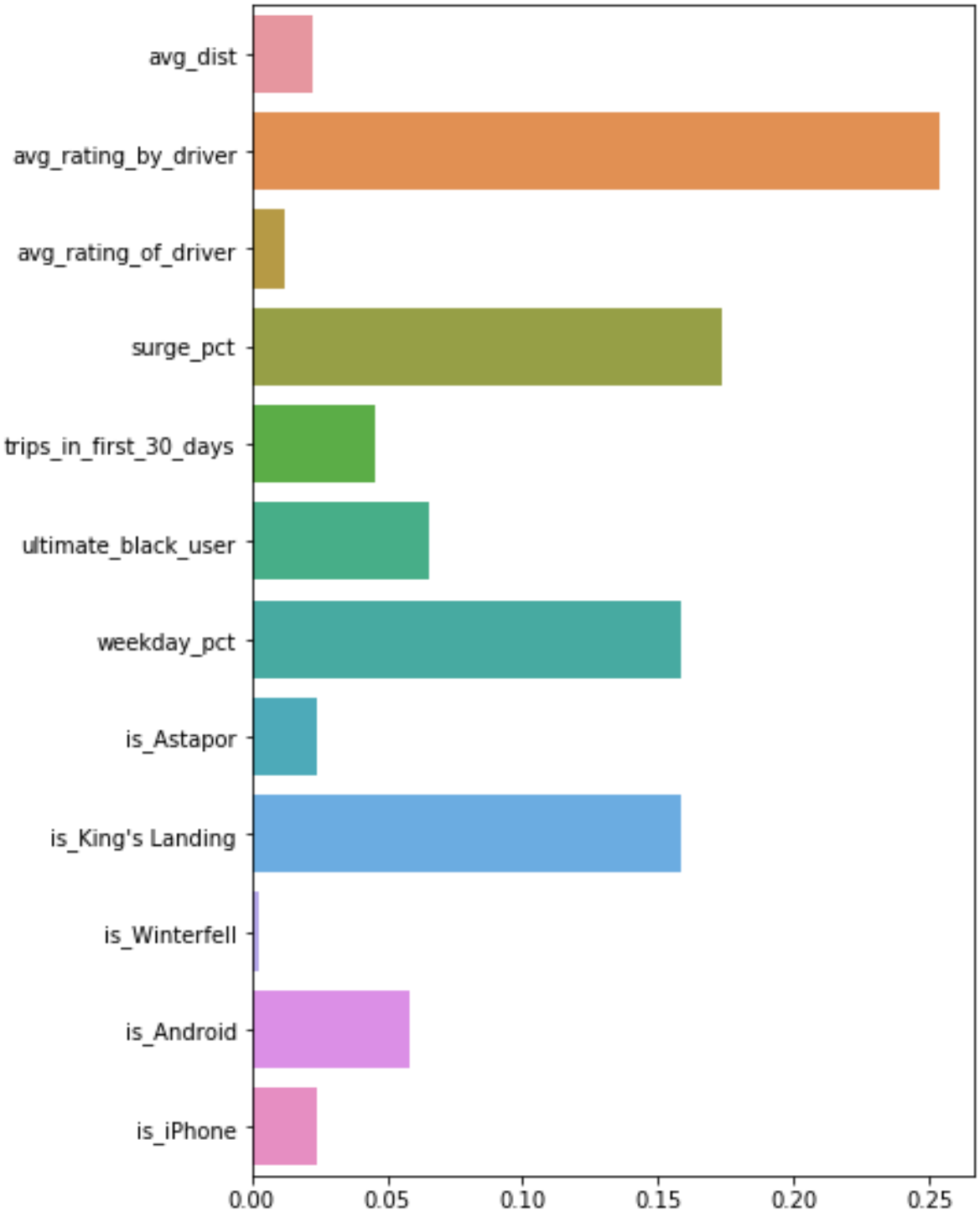

retained
0
1
14000
12000
10000
8000
6000
4000
2000
0
count
King's Landing
Astapor
Winterfell
city



Part 3 - Predictive modeling

Iter	Train Loss	Remaining Time
1	1.2792	12.28s
2	1.2427	9.00s
3	1.2110	7.92s
4	1.1834	7.32s
5	1.1612	6.99s
6	1.1409	6.77s
7	1.1229	6.64s
8	1.1074	6.48s
9	1.0884	6.36s
10	1.0761	6.23s
20	0.9913	5.40s
30	0.9582	4.95s
40	0.9400	4.66s
50	0.9291	4.33s
60	0.9218	3.95s
70	0.9169	3.60s
80	0.9131	3.27s
90	0.9094	2.95s
100	0.9065	2.65s
200	0.8876	0.00s

0.7837333333333333


avg_dist
avg_rating_by_driver
avg_rating_of_driver
surge_pct
trips_in_first_30_days
ultimate_black_user
weekday_pct
is_Astapor
is_King's Landing
is_Winterfell
is_Android
is_iPhone
0.00
0.05
0.10
0.15
0.20
0.25

OBSERVATIONS

- Saturday is the most popular day for users followed by Sunday and Friday respectively. Monday shows the least activity.
- The login counts have been recorded from January 1, 1970 to April 13, 1970.
- the number of logins has progressively increased over the months. This indicates an increasing user base.
- Maximum activity between 10 PM and 1 AM. There is also a sharp rise in user activity at around noon. The least activity is recorded during early morning between 6 AM and 8 AM.

RECOMMENDATIONS

- Increase operations in King's Landing as they tend to have greater probability of conversion. Alternatively, discover what is unique about King's Landing drivers and passengers and check if it can be implemented in the other cities, especially Astapor.
- If the user has taken a Ultimate Black, it indicates that s/he is more likely to stay.
- Provide additional perks to these people.
- People who use cabs on the weekdays are more likely to be retained. This is the most important feature as discovered by the Gradient Boosting Classifier.
- Provide more offers in the weekdays to encourage people to take cabs to work.