databricks Rating Prediction

1.EDA Analysis

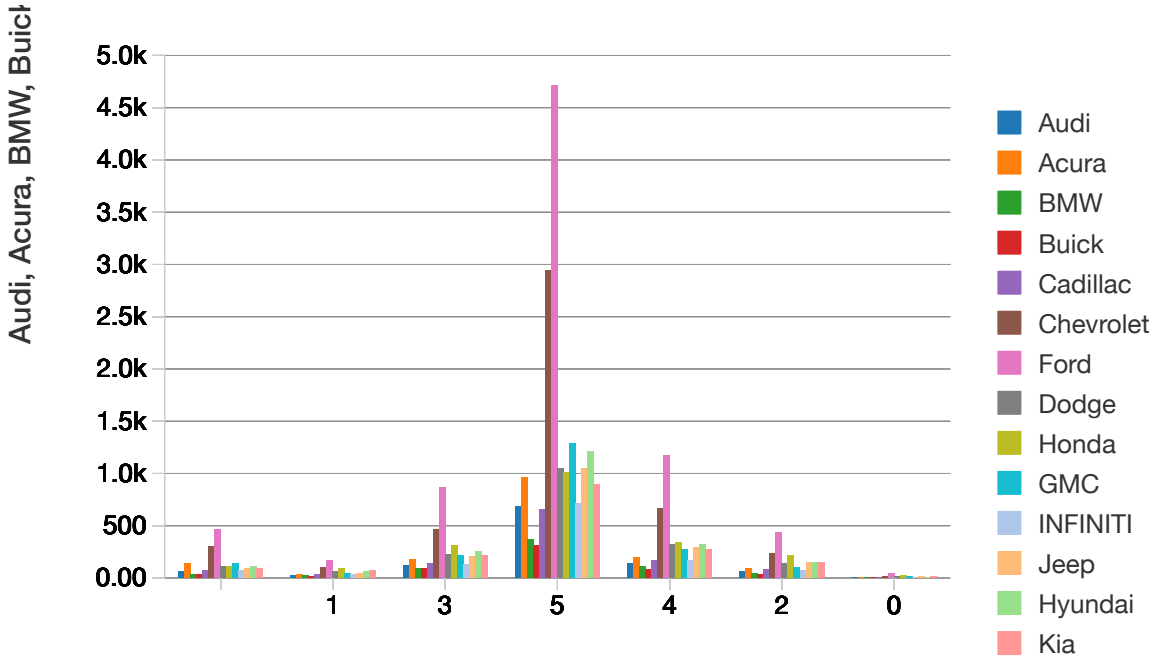
1. Inspecting inactive drivers who has not provided a drive

```
rides = spark.read.parquet("/mnt/cis442f-data/duocar/clean/rides/")
drivers = spark.read.parquet("/mnt/cis442f-data/duocar/clean/drivers/")
drivers.select('id').subtract(rides.select('driver_id')).count()
# There are 103 inactive drivers.
```

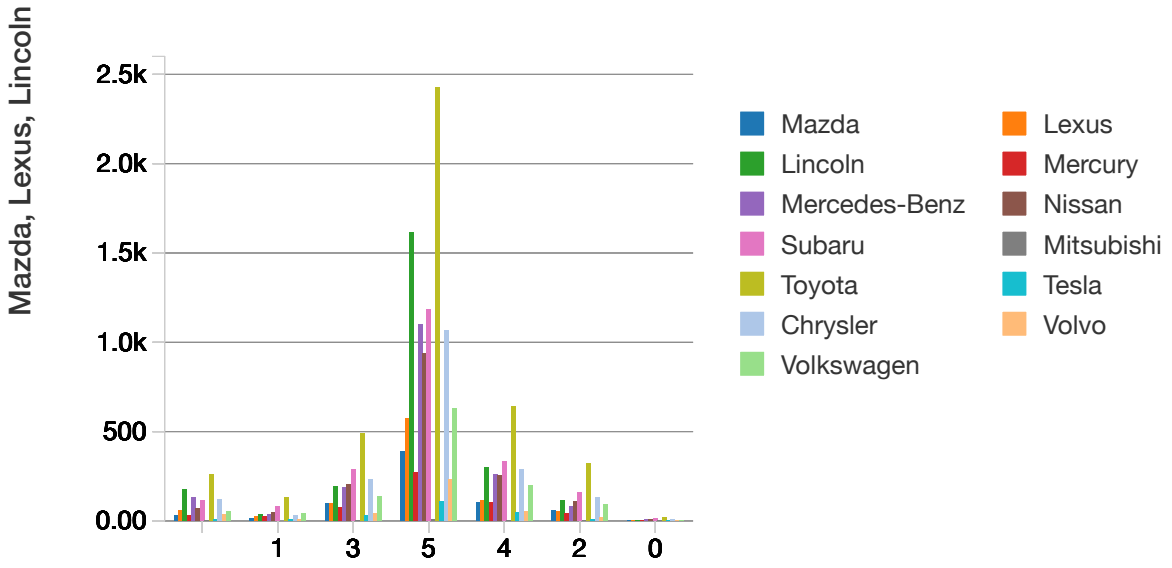
```
Out[3]: 103
```

2. Plot the ride rating in terms of vehicles

```
reviews = spark.read.parquet("/mnt/cis442f-data/duocar/clean/ride_reviews/")
joined =
rides.select('driver_id','star_rating').join(drivers.select('id','vehicle_make'
),rides.driver_id == drivers.id)
eda = joined.groupby('star_rating').pivot('vehicle_make').count()
display(eda)
```



```
display(eda)
```



From the barplot and line chart, star rating is not dependent on vehicle make. The bar plot and line chart both indicates that 5 stars is the most frequently given rating. And the distribution of rating is similar regardless of vehicle make. Therefore, we can conclude that star rating is not dependent on vehicle make.

3. Investigate student users

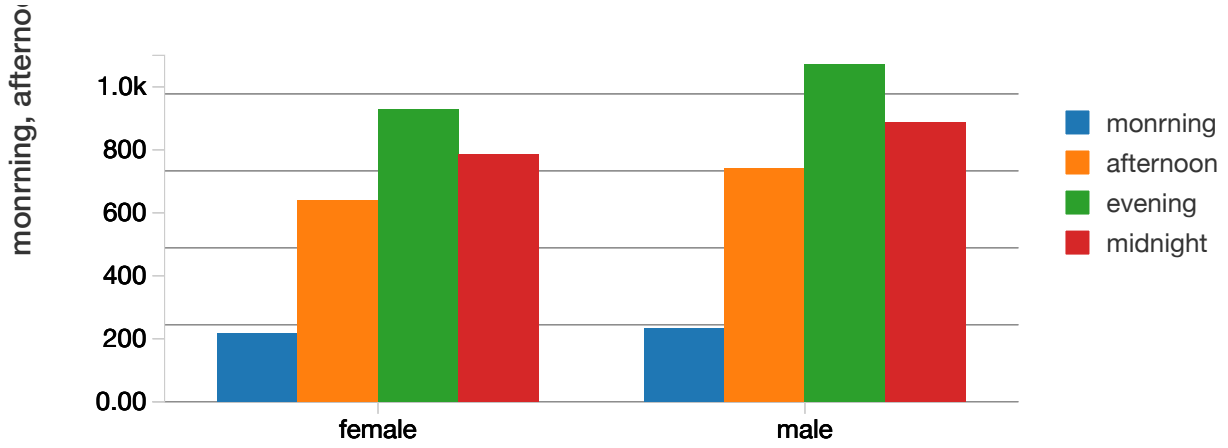
```
# create new columns of the time information (date_time in a day, day of a
week, month)
riders = spark.read.parquet("/mnt/cis442f-data/duocar/clean/riders/")
joined2 = rides.join(riders,rides.rider_id ==
riders.id).select('rider_id','student','sex','date_time')

from pyspark.sql.functions import dayofweek,month,hour,when
joined2 = joined2.select('rider_id','student','sex',\

'date_time',dayofweek('date_time').alias('dayofweek'),month('date_time').alias(
'month'))\
          .withColumn('date_time', when((hour(joined2.date_time) > 5) &
(hour(joined2.date_time) < 13),'monrning')\
                   .when((hour(joined2.date_time) > 12) &
(hour(joined2.date_time) < 19),'afternoon').\
                    when((hour(joined2.date_time) > 18) &
(hour(joined2.date_time) < 24),'evening').otherwise('midnight'))
joined2.show(5)
```

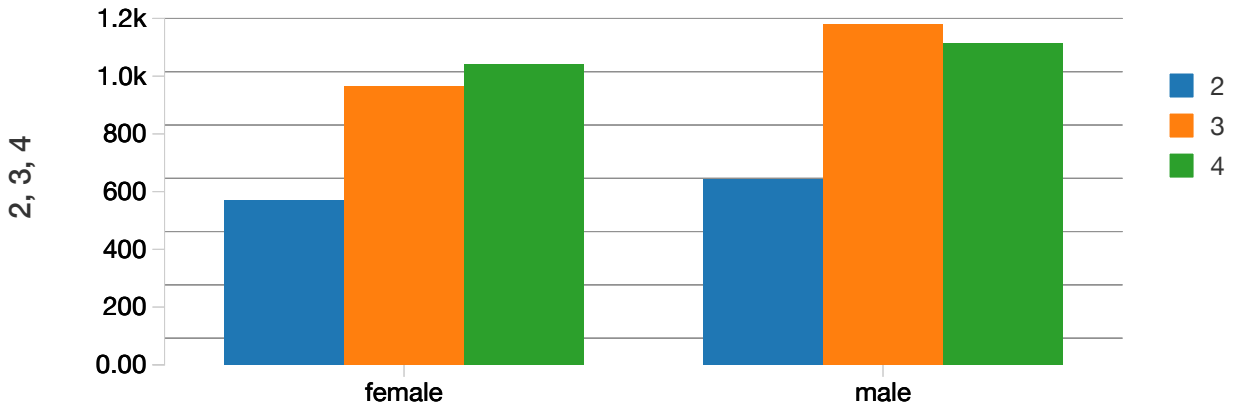
```
+------------+-------+------+---------+---------+-----+
|    rider_id|student|   sex|date_time|dayofweek|month|
+------------+-------+------+---------+---------+-----+
|220200000084|  false|female| monrning|        4|    2|
|220200000462|  false|  male| monrning|        4|    2|
|220200000489|  false|  male| monrning|        4|    2|
|220200000057|  false|  male| monrning|        4|    2|
|220200000012|   true|  null| monrning|        4|    2|
+------------+-------+------+---------+---------+-----+
only showing top 5 rows
```

```
# 1. time period during a day
datetime = joined2.dropna(subset = ['sex']).filter(joined2.student ==
True).groupby('sex').pivot('date_time').count()
display(datetime)
```



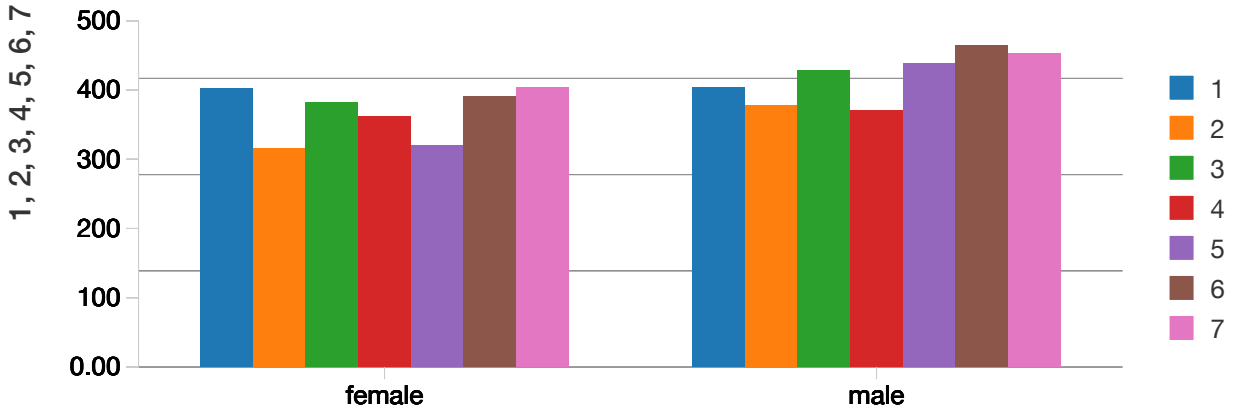
The bar chart indicates that student prefer taking a ride during evening and midnight (7pm to 5am the next day). Student seldom take a ride in the morning. Also, the distribution has no sex difference.

```
# 2.Month
month = joined2.dropna(subset = ['sex']).filter(joined2.student == True).groupby('sex').pivot('month').count()
display(month)
```



The bar chart indicates that females prefer taking a ride in April to February and March, while males prefer March. Both genders tend to take less rides in February.

```
# 3. day of week
dayofweek = joined2.dropna(subset = ['sex']).filter(joined2.student == True).groupby('sex').pivot('dayofweek').count()
display(dayofweek)
```



The bar chart shows that males tend to take more rides than females. Males prefer to take a ride during weekends and Fridays. And females take the most rides in Mondays and Sundays.

4. Explore the distance data

```
rides.where(rides.distance.isNull()).select('cancelled','distance').distinct().show()
```

```
+---------+--------+
|cancelled|distance|
+---------+--------+
|     true|    null|
+---------+--------+
```

All rides records that has nulls in distance is due to cancelled trips.