Chicago Taxi Trips

Group 8:
Andrew Shimshock, Claire Zyfers, Robby Konrath, Mina Tawfik and Travis Vickers

Dataset

This Dataset include taxi trips for 2016 in the City of Chicago. Each row represents one taxi ride. Columns include:

taxi_id: identification number of taxi

trip_start_timestamp: time taxi driver begins trip for passenger

trip_end_timestamp: time taxi driver ends trip for passenger

trip_seconds: total time of trip counted in seconds for passenger

trip_miles: total miles traveled during trip for passenger

pickup_community_area: specialized community area where passenger picked up

dropoff_community_area: specialized community area where passenger dropped off up

https://www.kaggle.com/chicago/chicago-taxi-rides-2016

fare: amount to go from point a to b

tips: amount passenger tipped taxi driver

tolls: extra toll payments

extras: extra money spent by passenger for taxi ride

trip_total: total amount passenger spent on taxi ride

payment_type: payment method of customer

company: company taxi driver works for

pickup_lattiude: pickup destination latitude point

pickup_longitude: pickup destination longitude point

dropoff_lattiude: dropoff destination latitude point

dropoff_longitude: dropoff destination longitude point

Data Cleanup Part 1

The dataset as a whole contains 20 million rows so we took a ten percent sample in order to work with the data. Most columns had some NaN.

```
Out[6]: taxi_id                      283
        trip_start_timestamp           0
        trip_end_timestamp           250
        trip_seconds                 329
        trip_miles                    22
        pickup_census_tract      1986616
        dropoff_census_tract      772817
        pickup_community_area     275671
        dropoff_community_area    308533
        fare                          30
        tips                          30
        tolls                         30
        extras                        30
        trip_total                    30
        payment_type                   0
        company                   763833
        pickup_latitude           275629
        pickup_longitude          275629
        dropoff_latitude          304747
        dropoff_longitude         304747
        dtype: int64
```

- Any numerical value->replaced with mean of that value
- Census columns both dropped
- For missing trip_end_timestamp->added on the avg seconds
- taxi_id->replaced with ‘Unknown
- Community areas filled with ‘Unknown’, but not used anyway

Data Cleanup Part 2

- Unfortunately the creator of the dataset had encoded the meaning of some of the columns including longitude, latitude, and company in a json file. We had to fix this to have the correct values

Before

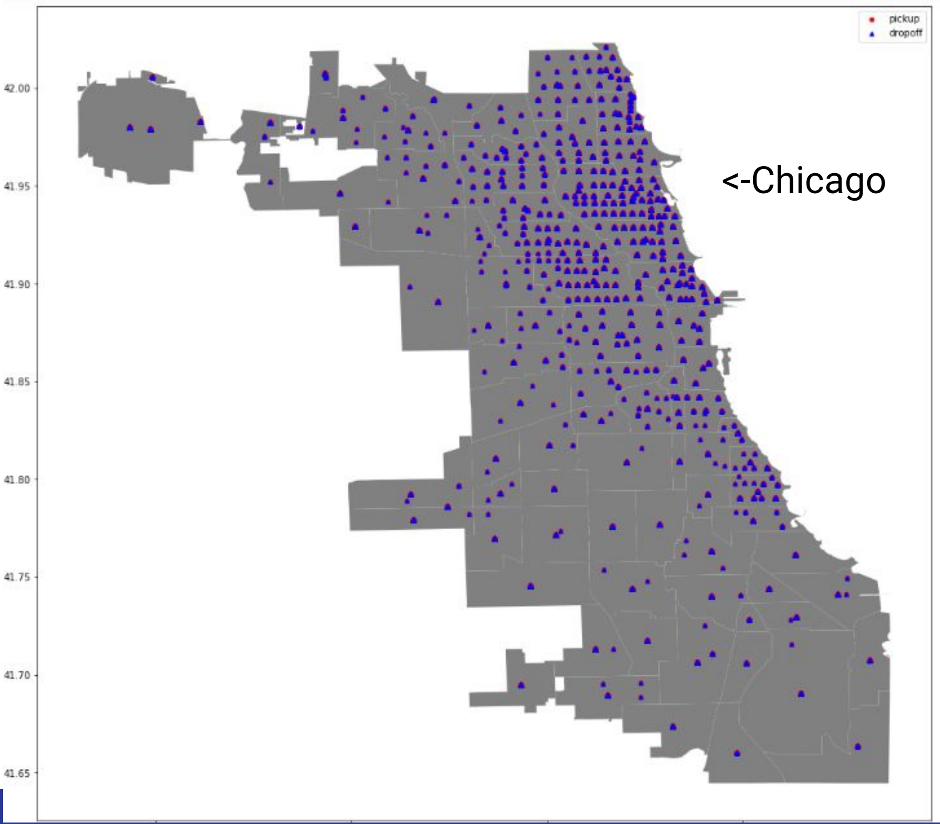
company	pickup_latitude	pickup_longitude	dropoff_latitude	dropoff_longitude
Unknown	210	470	411	545
Unknown	255	300	753	551
Unknown	411	545	433	757
Taxi Affiliation Services	686	500	660	120
Chicago Medallion Leasing INC	Unknown	Unknown	Unknown	Unknown

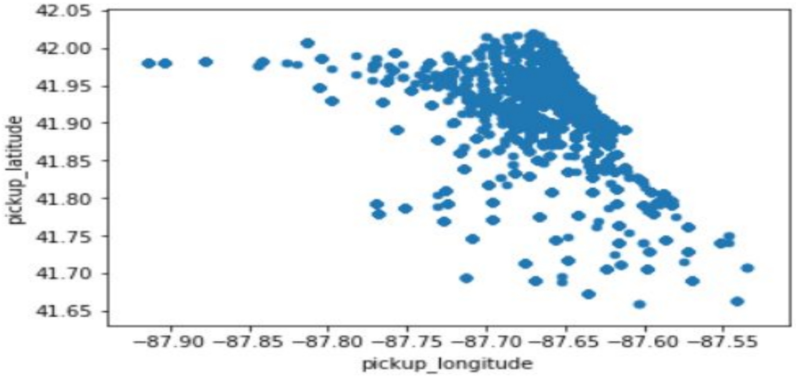
After

company	pickup_latitude	pickup_longitude	dropoff_latitude	dropoff_longitude
Unknown	41.892508	-87.626215	41.879255	-87.642649
Unknown	41.908379	-87.670945	41.906026	-87.675312
Unknown	41.879255	-87.642649	41.785999	-87.750934
Taxi Affiliation Services	41.944227	-87.655998	41.965812	-87.655879
Chicago Medallion easing INC	41.900879	-87.659032	41.900886	-87.653736
Chicago Elite Cab Corp. (Chicago Carriag	41.900879	-87.659032	41.900886	-87.653736

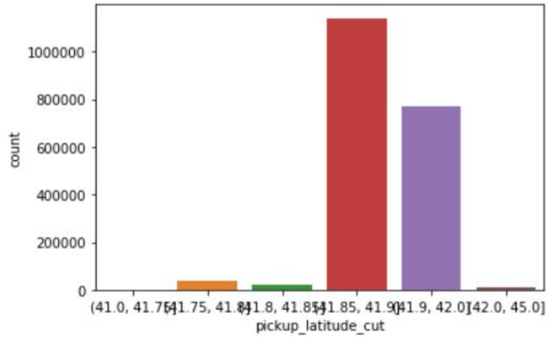
Finding #1

Best location-most consumers located in Chicago



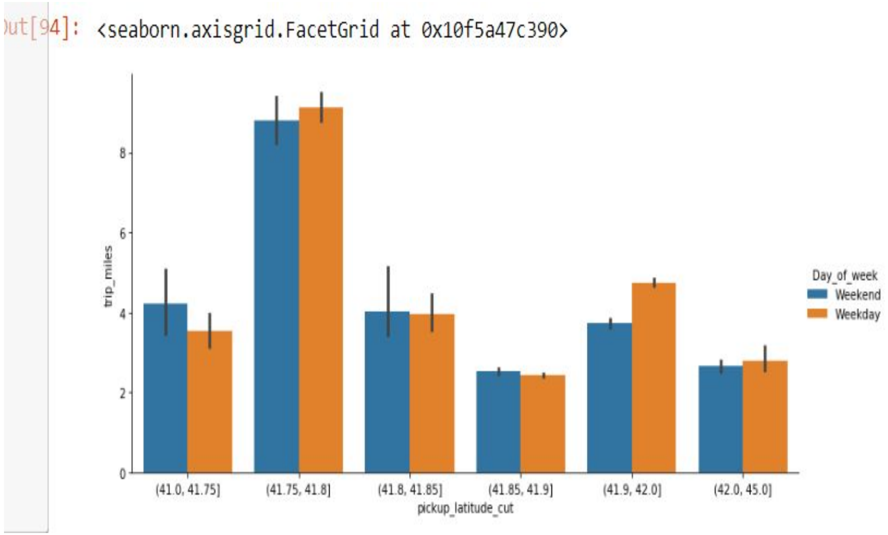


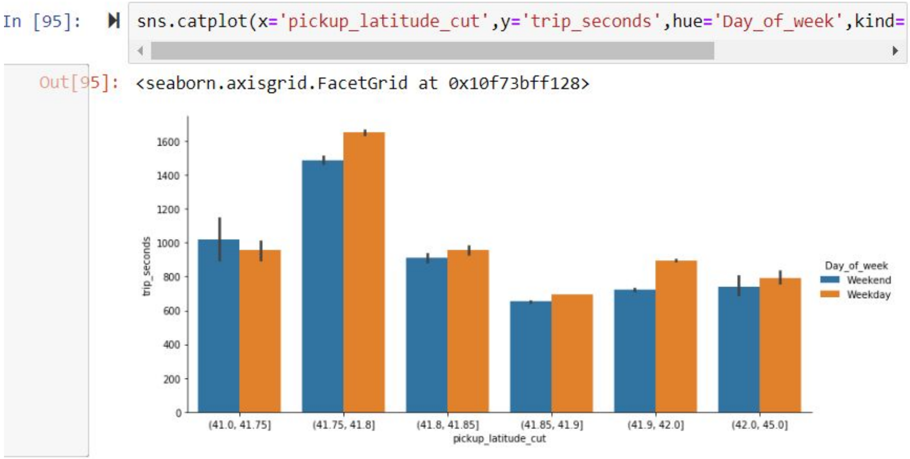

```
]: sns.countplot(x='pickup_latitude_cut',data=df)
[216]: <matplotlib.axes._subplots.AxesSubplot at 0x10e08a8aeb8>
```



Trip Analysis

- Trips south of Chicago are generally longer in miles. Chicago is located at 41.85, 41.9.
- Trips south of Chicago take more time.

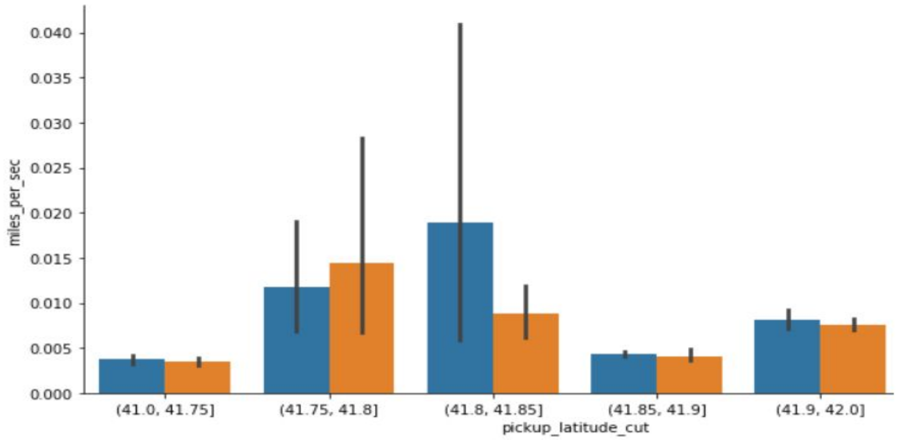




Trips Analysis Continued

- Trips south of Chicago are 3 times faster.
- Based on Chicago Taxi fare information we can see how important it is to go many miles and as fast as possible.

t[221]: <seaborn.axisgrid.FacetGrid at 0x10e08a33208>



<u>CITY OF CHICAGO TAXICAB FARE RATES and INFORMATION</u>

FARE RATES as of January 1, 2016

Flag Pull (Base Fare)	$ 3.25
Each additional mile	$ 2.25
Every 36 seconds of time elapsed	$ 0.20
First additional passenger	$ 1.00
Each additional passenger after first passenger*	$ 0.50
Vomit Clean-up Fee	$50.00
Illinois Airport Departure Tax**	$ 4.00

Summary

- Most consumers are located in Chicago
- After analysis, however, a driver can make much more per ride south of Chicago. This factors in multiple rides in Chicago as they are shorter usually.

```
Most one could make in one trip on average south of Chicago is 34.115
Most one could make in Chicago on average compared to one trip south of Chicago is 17.41
```

Managerial Insight

Large companies should have most of their taxis in Chicago. However, individual drivers may want to start South of Chicago to maximize profit

Finding #2

Efficiency analysis with Clustering

Calculated efficiency as (trip_total/trip_seconds) to find the revenue generated per second of the trip. Also calculated miles per second as we know fast trips are best

Then used clustering analysis to split dataset into three clusters

In [90]:
```
df3.groupby('cluster')[['trip_seconds','trip_miles', \
                        'fare','tips','tolls','extras','trip_total','time_0-6','time_6-12',\
                        'time_12-18','time_18-24','tipper','revenue_per_sec']].mean()
```

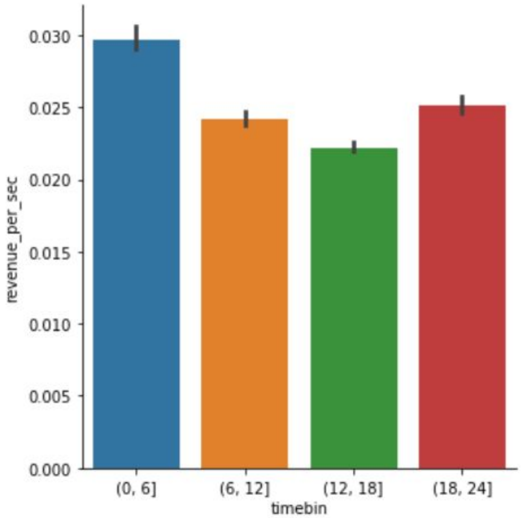
Out[90]:

cluster	trip_seconds	trip_miles	fare	tips	tolls	extras	trip_total	time_0-6	time_6-12	time_12-18	time_18-24	tipper	revenue_per_sec
0	594.985004	2.208847	10.033706	1.082239	0.001612	0.706507	11.873733	0.143915	0.089942	0.118803	0.276742	0.401733	0.024912
1	52102.507194	13.187302	72.511583	1.132698	0.865108	0.678777	75.241223	0.104317	0.169065	0.075540	0.118705	0.223022	0.001663
2	2421.573478	13.712285	39.445912	4.991370	0.010737	3.458138	47.990896	0.057451	0.096227	0.167396	0.212348	0.593480	0.020996

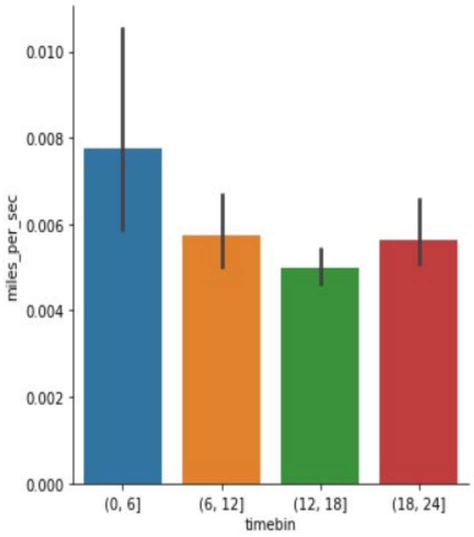
Summary

Clusters 0 and 2 had best efficiency ratings and also had the highest level of their trips occurring between the midnight to 6am time slot and the 6pm to midnight time slot.

```
In [80]:  sns.catplot(y='revenue_per_sec',x='timebin', kind='bar', data=df3)
Out[80]:  <seaborn.axisgrid.FacetGrid at 0x23e17a700b8>
```




```
n [270]:  sns.catplot(y='miles_per_sec',x='timebin', kind='bar', data=dft)
Out[270]: <seaborn.axisgrid.FacetGrid at 0x1104a03ba58>
```



Managerial Insight

To maximize efficiency, drivers should do more trips between midnight and 6am or 6pm and midnight, rather than normal business hours

Finding #3

Tipping and Payment type

```
df = df[df.trip_seconds!=0]
df = df[df.trip_total !=0]
```

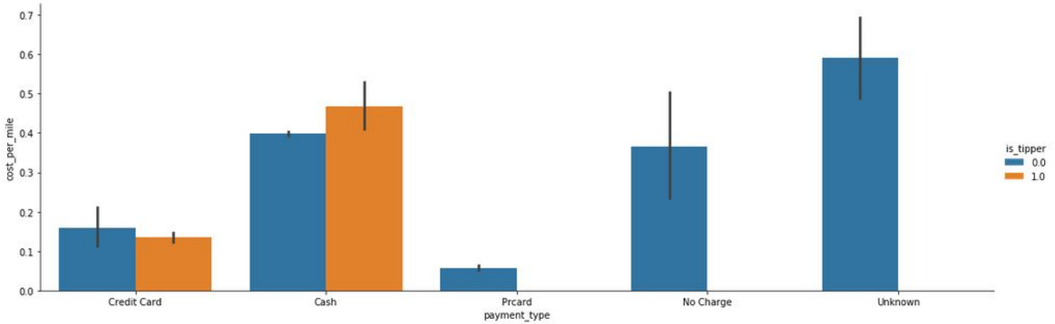
```
df['cost_per_mile']= df.trip_total/df.trip_miles
```

```
dfCostPerMile = df[(df.cost_per_mile > .01)&(df.cost_per_mile <.7)]
```

```
dfCostPerMile['is_tipper'] = dfCostPerMile.tips.apply(lambda x: 1.0 if x>0 else 0.0)
```

```
In [62]: sns.catplot(x='payment_type', y='cost_per_mile', data=dfCostPerMile, hue='is_tipper', kind='bar',aspect=3)
```

```
Out[62]: <seaborn.axisgrid.FacetGrid at 0x284c7859648>
```



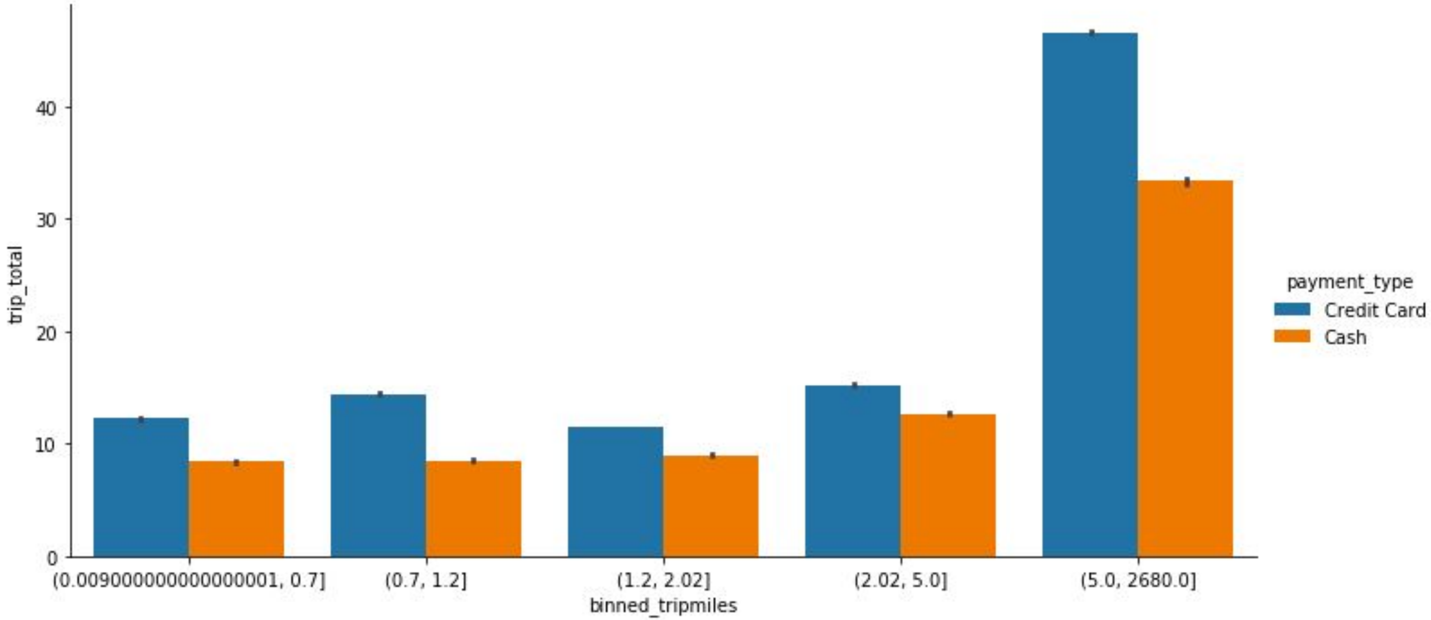
```
: dfm = df[(df.payment_type == 'Cash') | (df.payment_type == 'Credit Card')]

: dfm2 = dfm[['payment_type', 'trip_total', 'trip_miles']]

: dfm2 = dfm2[dfm2.trip_miles != 0]

: dfm2['binned_tripmiles']=pd.qcut(dfm2.trip_miles, 5)

: sns.catplot(y='trip_total', data=dfm2, x='binned_tripmiles', aspect=2, kind='bar', hue = 'payment_type')
```



Managerial Insight

Credit card users tend to both pay more for every binned distance traveled and on average have a longer trip distance (indicated by lower cost/mile), this leads us to say that credit card users are more likely to tip so all taxi companies should incentivize credit card payments with some sort of discount program.

Thank you!