Rudy Fasano

Data Glacier, Week 2

Cab Data Data Investigation

In [2]:
```
import pandas as pd
import matplotlib as mpl
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats
import seaborn as sns; sns.set_theme(color_codes=True)
%matplotlib inline
```

Reading in necessary files in order to get a snapshot at the collective data.

In [4]:
```
cab_data = pd.read_csv("D:/DataGlacier/Week2/DataSets-main/Cab_Data.csv")
cab_data.head()
```

Out[4]:

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
0	10000011	42377	Pink Cab	ATLANTA GA	30.45	370.95	313.635
1	10000012	42375	Pink Cab	ATLANTA GA	28.62	358.52	334.854
2	10000013	42371	Pink Cab	ATLANTA GA	9.04	125.20	97.632
3	10000014	42376	Pink Cab	ATLANTA GA	33.17	377.40	351.602
4	10000015	42372	Pink Cab	ATLANTA GA	8.73	114.62	97.776

In [5]:
```
city = pd.read_csv("D:/DataGlacier/Week2/DataSets-main/City.csv")
city.head()
```

Out[5]:

	City	Population	Users
0	NEW YORK NY	8,405,837	302,149
1	CHICAGO IL	1,955,130	164,468
2	LOS ANGELES CA	1,595,037	144,132
3	MIAMI FL	1,339,155	17,675
4	SILICON VALLEY	1,177,609	27,247

In [9]:
```
customer_id = pd.read_csv("D:/DataGlacier/Week2/DataSets-
main/Customer_ID.csv")
```

```
customer_id.head()
```

Out[9]:

	Customer ID	Gender	Age	Income (USD/Month)
0	29290	Male	28	10813
1	27703	Male	27	9237
2	28712	Male	53	11242
3	28020	Male	23	23327
4	27182	Male	33	8536

In [7]:

```
transaction_id = pd.read_csv("D:/DataGlacier/Week2/DataSets-
main/Transaction_ID.csv")
transaction_id.head()
```

Out[7]:

	Transaction ID	Customer ID	Payment_Mode
0	10000011	29290	Card
1	10000012	27703	Card
2	10000013	28712	Cash
3	10000014	28020	Cash
4	10000015	27182	Card

identifying number of observations and variables we have in the data sets.

In [6]:

```
print('cab data shape:', cab_data.shape,
      '\ncity shape:', city.shape,
      '\ncustomer ID shape,:', customer_id.shape,
      '\ntransaction_id shape:', transaction_id.shape)
```

```
cab data shape: (359392, 7)
city shape: (20, 3)
customer ID shape,: (49171, 4)
transaction_id shape: (440098, 3)
```

identifying any null values - none identified

In [7]:

```
print('cab data null values:', cab_data.isnull().sum(),
      '\ncity null values:', city.isnull().sum(),
      '\ncustomer ID null values,:', customer_id.isnull().sum(),
      '\ntransaction_id null values:', transaction_id.isnull().sum())
```

```
cab data null values: Transaction ID    0
Date of Travel    0
Company           0
City              0
KM Travelled      0
```

```
Price Charged     0
Cost of Trip      0
dtype: int64
city null values: City          0
Population    0
Users         0
dtype: int64
customer ID null values,: Customer ID           0
Gender                 0
Age                    0
Income (USD/Month)     0
dtype: int64
transaction_id null values: Transaction ID    0
Customer ID       0
Payment_Mode      0
dtype: int64
```

filtering cab data into two dataframes including pink and yellow cab company data.

In [8]:
```
p_cab = cab_data['Company'] == 'Pink Cab'
y_cab = cab_data['Company'] == 'Yellow Cab'
```

Pink Cab data exploration.

In [9]:
```
### Pink cab data
f1 = cab_data[p_cab]
f1
```

Out[9]:

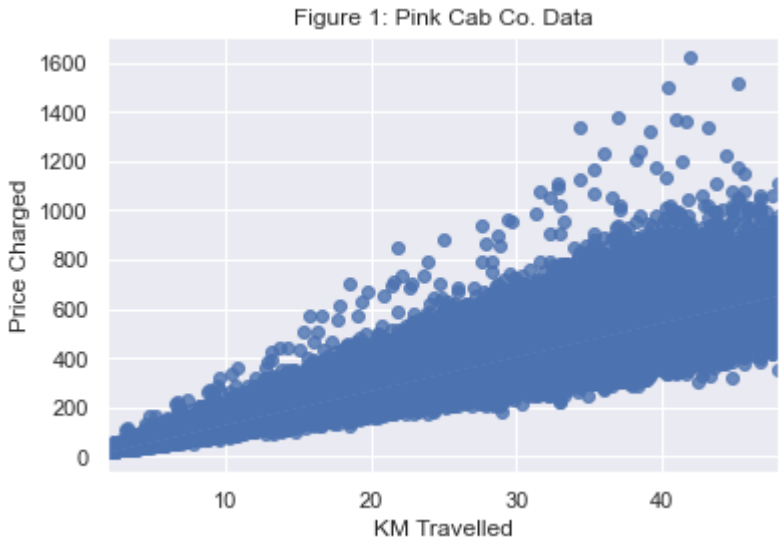
	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
0	10000011	42377	Pink Cab	ATLANTA GA	30.45	370.95	313.635
1	10000012	42375	Pink Cab	ATLANTA GA	28.62	358.52	334.854
2	10000013	42371	Pink Cab	ATLANTA GA	9.04	125.20	97.632
3	10000014	42376	Pink Cab	ATLANTA GA	33.17	377.40	351.602
4	10000015	42372	Pink Cab	ATLANTA GA	8.73	114.62	97.776
...	...	...	...	...	...	...	...
357449	10437610	43106	Pink Cab	WASHINGTON DC	13.56	184.19	135.600
357450	10437611	43465	Pink Cab	WASHINGTON DC	29.68	388.08	302.736
357451	10437612	43107	Pink Cab	WASHINGTON DC	28.50	369.04	310.650
357452	10437614	43102	Pink Cab	WASHINGTON DC	16.10	194.17	162.610
357453	10437615	43105	Pink Cab	WASHINGTON DC	22.20	287.46	244.200

84711 rows × 7 columns

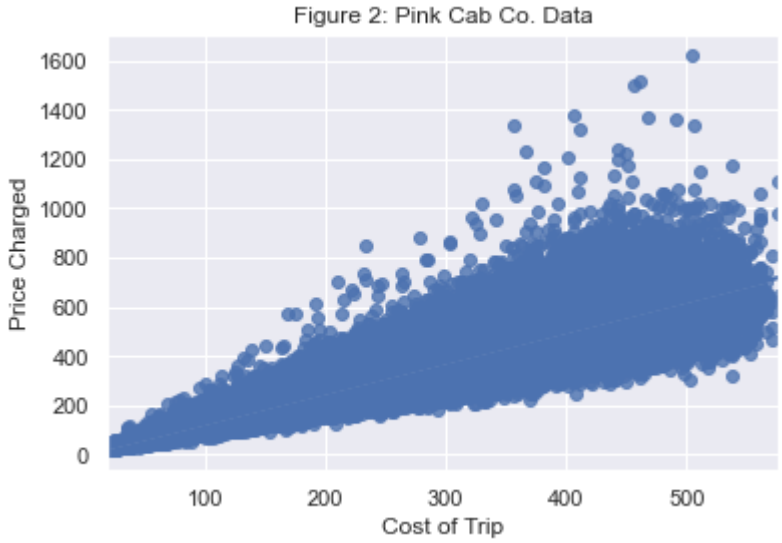
In [10]:
```
### finding simple statistics for the Pink cab company
f1.describe()
f1.sum(axis = 0, skipna = True)
```

Out[10]:
```
Transaction ID                                          866080030579
Date of Travel                                           3640470764
Company           Pink CabPink CabPink CabPink CabPink CabPink C...
City              ATLANTA GAATLANTA GAATLANTA GAATLANTA GAATLANT...
KM Travelled                                          1.91107e+06
Price Charged                                         2.63283e+07
Cost of Trip                                          2.10209e+07
dtype: object
```

In [10]:
```
### plotting for visualizations
ax = sns.regplot(x='KM Travelled', y='Price Charged',
data=f1).set_title('Figure 1: Pink Cab Co. Data')
```



In [11]:
```
## visualizations cont.
ax1 = sns.regplot(x='Cost of Trip', y='Price Charged',
data=f1).set_title('Figure 2: Pink Cab Co. Data')
```



Pink Cab Co. Hypotheses: As KM travelled increases, as does price charged. As cost of trip increases as does the price being charged proportionately increases.

In [13]:
```
### finding correlation and P-value of pink cab prices charges by cost of trip
cor, p = scipy.stats.pearsonr(x=f1['KM Travelled'], y=f1['Price Charged'])
print('Pink Cab Co. Data: The correlation coefficient between KM Travelled and Price Charged is:', cor, 'which indicates \nhigh positive correlation.'
      'The P-value between the same variables is', p, 'indicating statistical significance. This provides \nsufficient evidence to reject the null hypothesis in favor ot the alternative hypothesis.')
```

```
Pink Cab Co. Data: The correlation coefficient between KM Travelled and Price Charged i
s: 0.9277652782594881 which indicates
high positive correlation.The P-value between the same variables is 0.0 indicating stati
stical significance. This provides
sufficient evidence to reject the null hypothesis in favor ot the alternative hypothesi
s.
```

In [14]:
```
cor1, p1 = scipy.stats.pearsonr(x=f1['Cost of Trip'], y=f1['Price Charged'])
print('Pink Cab Co. Data: The correlation coefficient between Cost of Trip and Price Charged is:', cor1, 'which indicates \nhigh positive correlation.'
      'The P-value between the same variables is', p1, 'indicating statistical significance. This provides \nsufficient evidence to reject the null hypothesis in favor ot the alternative hypothesis.')
```

```
Pink Cab Co. Data: The correlation coefficient between Cost of Trip and Price Charged i
s: 0.9218956544941218 which indicates
high positive correlation.The P-value between the same variables is 0.0 indicating stati
stical significance. This provides
sufficient evidence to reject the null hypothesis in favor ot the alternative hypothesi
s.
```

Yellow Cab data exploration.

In [13]:
```
### Yellow cab data
f2 = cab_data[y_cab]
f2
```

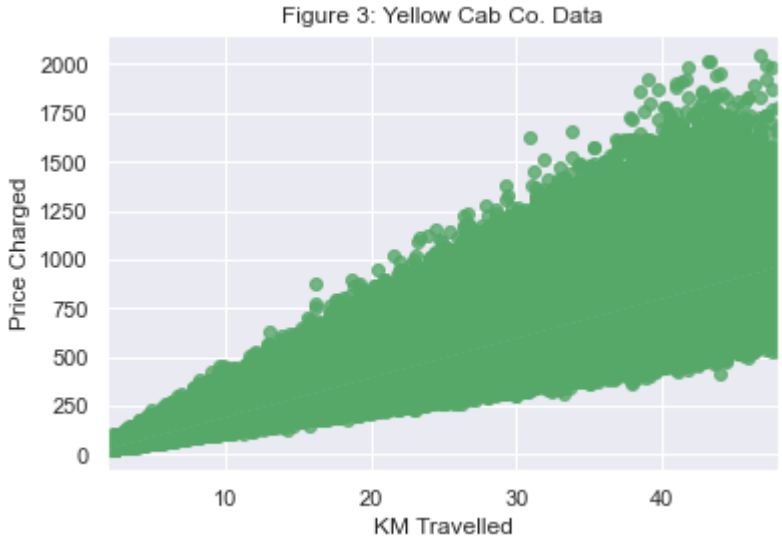
Out[13]:

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
233	10000384	42371	Yellow Cab	ATLANTA GA	33.93	1341.17	464.1624
234	10000385	42378	Yellow Cab	ATLANTA GA	42.18	1412.06	516.2832
235	10000386	42372	Yellow Cab	ATLANTA GA	10.60	364.62	132.2880
236	10000387	42375	Yellow Cab	ATLANTA GA	26.75	838.00	333.8400
237	10000388	42376	Yellow Cab	ATLANTA GA	46.02	1540.61	596.4192
...	...	...	...	...	...	...	...
359387	10440101	43108	Yellow Cab	WASHINGTON DC	4.80	69.24	63.3600
359388	10440104	43104	Yellow Cab	WASHINGTON DC	8.40	113.75	106.8480
359389	10440105	43105	Yellow Cab	WASHINGTON DC	27.75	437.07	349.6500
359390	10440106	43105	Yellow Cab	WASHINGTON DC	8.80	146.19	114.0480
359391	10440107	43102	Yellow Cab	WASHINGTON DC	12.76	191.58	177.6192

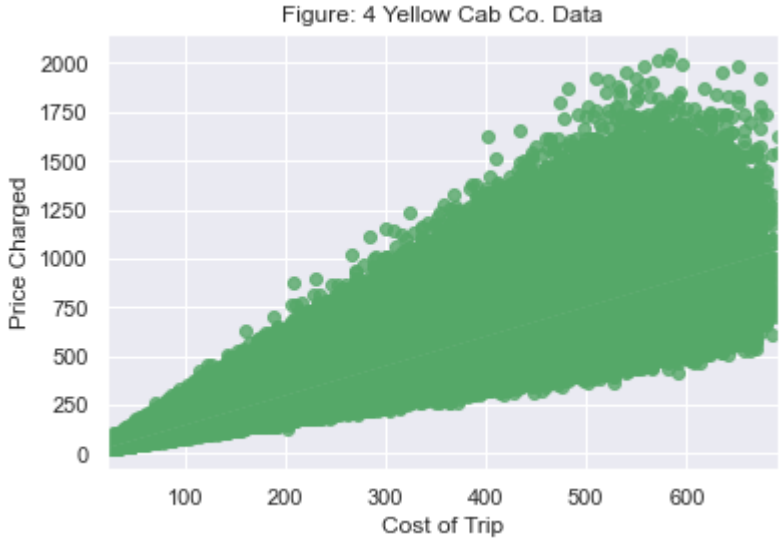
274681 rows × 7 columns

In [ ]:
```
### discovering simple statistics for the yellow cab company
f2.describe()
f2.sum(axis = 0, skipna = True)
```

In [14]:
```
#plotting for visualizations
ax2 = sns.regplot(x='KM Travelled', y='Price Charged', color='g',
data=f2).set_title('Figure 3: Yellow Cab Co. Data')
```



In [15]:
```
## visualizations cont.
ax3 = sns.regplot(x='Cost of Trip', y='Price Charged', color='g',
data=f2).set_title('Figure: 4 Yellow Cab Co. Data')
```



Yellow Cab Co. Alt. Hypotheses: As KM travelled increases, as does price charged. As cost of trip increases as does the price being charged proportionately increases.

In [16]:
```
### finding correlation and P-value of pink cab prices charges by cost of
trip
cor2, p2 = scipy.stats.pearsonr(x=f2['KM Travelled'], y=f2['Price
Charged'])
print('Yellow Cab Co. Data: The correlation coefficient between KM
Travelled and Price Charged is:', cor2, 'which indicates \nhigh positive
correlation.'
      'The P-value between the same variables is', p2, 'indicating
```

```
statistical significance. This provides \nsufficient evidence to reject the
null hypothesis in favor of the alternative hypothesis.')
```

```
Yellow Cab Co. Data: The correlation coefficient between KM Travelled and Price Charged
is: 0.8597086294478448 which indicates
high positive correlation.The P-value between the same variables is 0.0 indicating stati
stical significance. This provides
sufficient evidence to reject the null hypothesis in favor of the alternative hypothesi
s.
```

In [17]:

```
### finding correlation and P-value of pink cab prices charges by cost of
trip
cor3, p3 = scipy.stats.pearsonr(x=f2['Cost of Trip'], y=f2['Price
Charged'])
print('Yellow Cab Co. Data: The correlation coefficient between Cost of
Trip and Price Charged is:', cor3, 'which indicates \nhigh positive
correlation.'
      'The P-value between the same variables is', p3, 'indicating
statistical significance. This provides \nsufficient evidence to reject the
null hypothesis in favor ot the alternative hypothesis.')
```

```
Yellow Cab Co. Data: The correlation coefficient between Cost of Trip and Price Charged
is: 0.8539958911683975 which indicates
high positive correlation.The P-value between the same variables is 0.0 indicating stati
stical significance. This provides
sufficient evidence to reject the null hypothesis in favor ot the alternative hypothesi
s.
```

merging of datasets for further investigation

In [12]:

```
######## merging datasets to find relationships
df = pd.merge(customer_id, transaction_id)
df


df1 = pd.merge(df, cab_data)
df1
```

Out[12]:

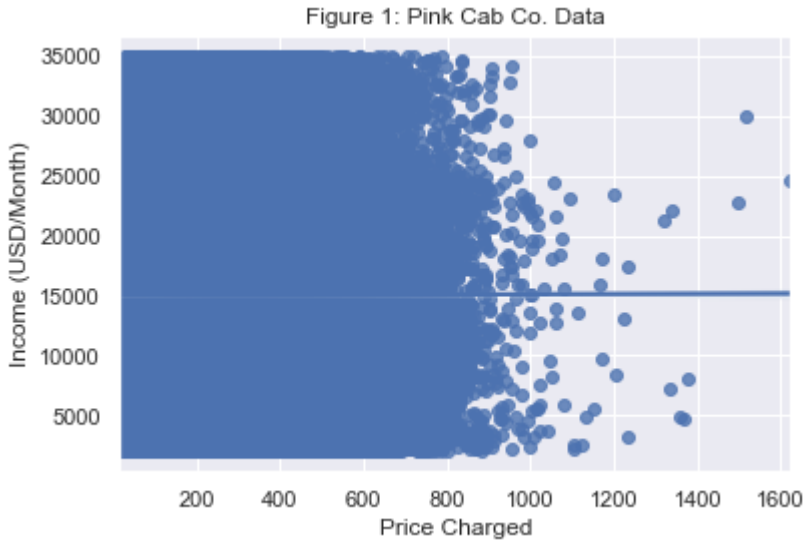
	Customer ID	Gender	Age	Income (USD/Month)	Transaction ID	Payment_Mode	Date of Travel	Company	Cit
0	29290	Male	28	10813	10000011	Card	42377	Pink Cab	ATLANT G
1	29290	Male	28	10813	10351127	Cash	43302	Yellow Cab	ATLANT G
2	29290	Male	28	10813	10412921	Card	43427	Yellow Cab	ATLANT G
3	27703	Male	27	9237	10000012	Card	42375	Pink Cab	ATLANT G

	Customer ID	Gender	Age	Income (USD/Month)	Transaction ID	Payment_Mode	Date of Travel	Company	Cit
4	27703	Male	27	9237	10320494	Card	43211	Yellow Cab	ATLANT G
...	...	...	...	...	...	...	...	...	
359387	38520	Female	42	19417	10439790	Card	43107	Yellow Cab	SEATTL W
359388	12490	Male	33	18713	10439799	Cash	43103	Yellow Cab	SILICO VALLE
359389	41414	Male	38	3960	10439838	Card	43104	Yellow Cab	TUCSO A
359390	41677	Male	23	19454	10439840	Cash	43106	Yellow Cab	TUCSO A
359391	39761	Female	32	10128	10439846	Card	43104	Yellow Cab	TUCSO A

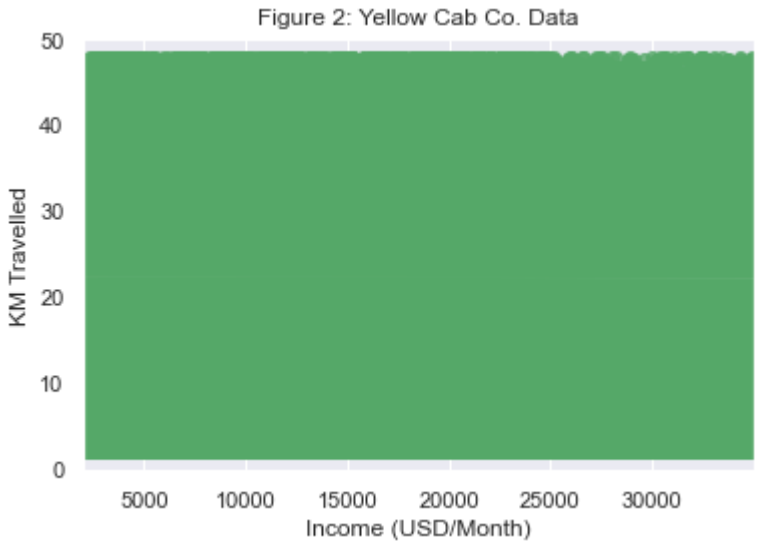
359392 rows × 12 columns

In [14]:
```
## separating into two dataframes to pull information from both cab companies
p_cab1 = df1['Company'] == 'Pink Cab'
y_cab1 = df1['Company'] == 'Yellow Cab'
f3 = df1[p_cab1]
f4 = df1[y_cab1]
```

In [16]:
```
### plotting for visualizations
ax = sns.regplot(x='Price Charged', y='Income (USD/Month)',
data=f3).set_title('Figure 1: Pink Cab Co. Data')
```



In [17]:
```
ax = sns.regplot(x='Income (USD/Month)', y='KM Travelled', data=f4,
color='g').set_title('Figure 2: Yellow Cab Co. Data')
```



Merged dataset Alternative Hypotheses: As income increases, so does KM travelled for both cab companies.

In [ ]:
```
cor, p = scipy.stats.pearsonr(x=f1['Income (USD/Month)'], y=f1['KM
Travelled'])
print('Pink Cab Co. Data: The correlation coefficient between Income in USD
and KM Travelled is:', cor, 'which indicates \n low and slightly positive
correlation.'
      'The P-value between the same variables is', p, 'indicating no
statistical \nsignificance. This provides sufficient evidence to not reject
the null hypothesis.')
```

In [ ]:
```
cor1, p1 = scipy.stats.pearsonr(x=f2['Income (USD/Month)'], y=f2['KM
Travelled'])
print('Yellow Cab Co. Data: The correlation coefficient between Income in
USD and KM Travelled is:', cor1, 'which indicates \n low and slightly
negative correlation.'
      'The P-value between the same variables is', p1, 'indicating no
statistical \nsignificance. This provides sufficient evidence to not reject
the null hypothesis.')
```

In [18]:
```
f3.describe()
```

Out[18]:

	Customer ID	Age	Income (USD/Month)	Transaction ID	Date of Travel	KM Travelled	Pri Charg
count	84711.000000	84711.000000	84711.000000	8.471100e+04	84711.000000	84711.000000	84711.0000

	Customer ID	Age	Income (USD/Month)	Transaction ID	Date of Travel	KM Travelled	Pri Charg
mean	18422.581577	35.322414	15059.047137	1.022394e+07	42975.183435	22.559917	310.8008
std	18084.830799	12.644780	7991.077762	1.261782e+05	305.502235	12.231092	181.9956
min	1.000000	18.000000	2000.000000	1.000001e+07	42371.000000	1.900000	15.6000
25%	5317.500000	25.000000	8371.000000	1.011014e+07	42700.000000	12.000000	159.9700
50%	8876.000000	33.000000	14713.000000	1.022590e+07	43000.000000	22.440000	298.0600
75%	27190.000000	42.000000	21055.000000	1.033642e+07	43252.000000	32.960000	441.5050
max	60000.000000	65.000000	35000.000000	1.043762e+07	43465.000000	48.000000	1623.4800

In [19]:
```
f4.describe()
```

Out[19]:

	Customer ID	Age	Income (USD/Month)	Transaction ID	Date of Travel	KM Travelled	Price
count	274681.000000	274681.000000	274681.000000	2.746810e+05	274681.000000	274681.000000	27468
mean	19428.831732	35.341112	15045.669817	1.021978e+07	42960.640022	22.569517	45
std	21830.791423	12.578625	7962.727062	1.269829e+05	307.990287	12.234298	28
min	1.000000	18.000000	2000.000000	1.000038e+07	42371.000000	1.900000	2
25%	2403.000000	25.000000	8439.000000	1.011084e+07	42695.000000	11.990000	22
50%	6445.000000	33.000000	14676.000000	1.021987e+07	42984.000000	22.440000	42
75%	38916.000000	42.000000	21023.000000	1.032939e+07	43225.000000	32.960000	63
max	60000.000000	65.000000	34996.000000	1.044011e+07	43465.000000	48.000000	204

In [22]:
```
print('The Pink Cab Co. had', len(f3), 'transactions during from 2016-2018.
\nThe Yellow Cab Co. had', len(f4), 'transactions in turn')
```

```
The Pink Cab Co. had 84711 transactions during from 2016-2018.
The Yellow Cab Co. had 274681 transactions in turn
```

In [23]:
```
print('The grand total of charged trips for the Pink Cab Co. was',
      f3['Price Charged'].sum(),'from 2016-2018, \nwhere as the Yellow Cab
Co. accrued', f4['Price Charged'].sum(), 'during this same time period.')
```

```
The grand total of charged trips for the Pink Cab Co. was 26328251.33 from 2016-2018,
where as the Yellow Cab Co. accrued 125853887.19 during this same time period.
```

In [26]:
```
#simple statistics
sil = 274681 / 84711
print('The Yellow Cab Co. is', sil, 'the size of the Pink Cab Co.
\nMathematically, if this rate of sales continue up to the point of the
```

```
Yellow Cab Co., \nthe Pink Cab Co. would still experience less earnings,
resulting in', final)
```

```
The Yellow Cab Co. is 3.2425659005323983 the size of the Pink Cab Co.
Mathematically, if this rate of sales continue up to the point of the Yellow Cab Co.,
the Pink Cab Co. would still experience less earnings, resulting in 85303533.24000001
```

In [25]:

```
#simple statistics cont.
final = 26328251 * 3.24
print('The Yellow Cab Co. is', sil, 'the size of the Pink Cab Co.
\nMathematically, if this rate of sales continue up to the point of the
Yellow Cab Co., \nthe Pink Cab Co. would still experience less earnings,
resulting in', final)
```

```
The Yellow Cab Co. is 3.2425659005323983 the size of the Pink Cab Co.
Mathematically, if this rate of sales continue up to the point of the Yellow Cab Co.,
the Pink Cab Co. would still experience less earnings, resulting in 85303533.24000001
```