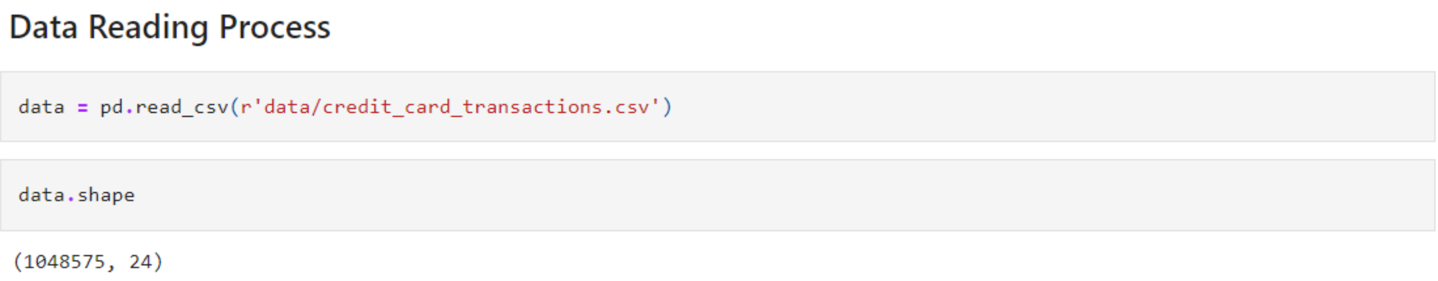
Exploratory Data Analysis for Credit Card Transactions

Proponent: Engr. Matthew David M. Loquinerio, REE

Questions to Answer:

1. What is the correlation of spending based on gender?
2. Which days and months have the most transaction records?
3. Which age group has the highest spending scores?
4. What merchant companies are the most spent on by the users?
5. Who are the individuals with above mean spending in their current cities?

For this EDA Project I decided to conduct data analyst practice for credit card transactions for the timeline starting in 2019 up to the first quarter of 2020. The CSV file is acquired from Kaggle an open data science forum website.



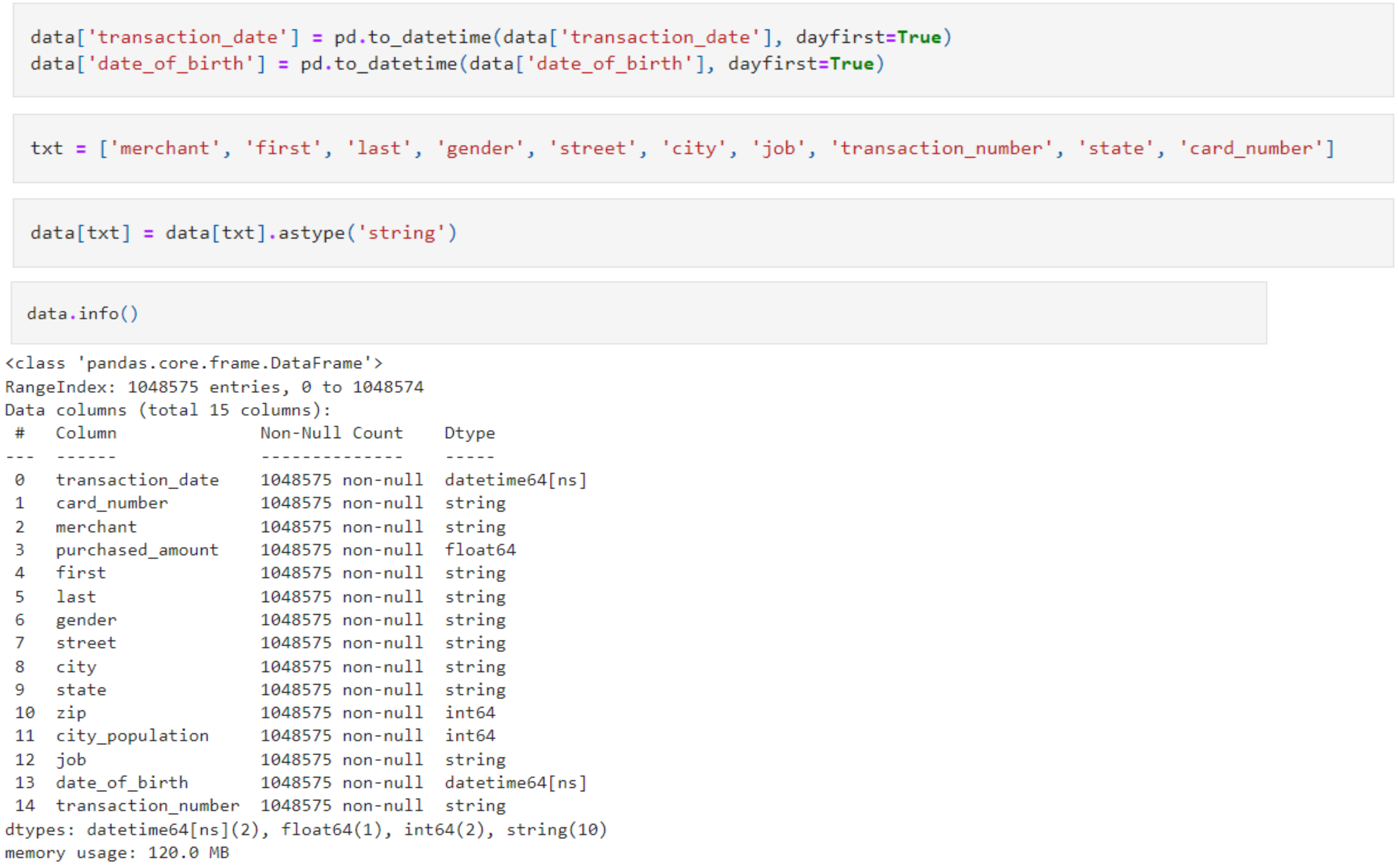
Data Reading Process

```
data = pd.read_csv(r'data/credit_card_transactions.csv')
```

```
data.shape
```

```
(1048575, 24)
```

The file contains 1,048,575 rows and 24 columns of data. Understanding the dataset some fields are not in the proper data types for transformation. It is substantial for a data analyst to clean the data set it includes: proper usage of data types, replacing unwanted string values, choosing columns that are relevant for analysis, removing duplicated rows, and dropping null subsets.



```
data['transaction_date'] = pd.to_datetime(data['transaction_date'], dayfirst=True)
data['date_of_birth'] = pd.to_datetime(data['date_of_birth'], dayfirst=True)
```

```
txt = ['merchant', 'first', 'last', 'gender', 'street', 'city', 'job', 'transaction_number', 'state', 'card_number']
```

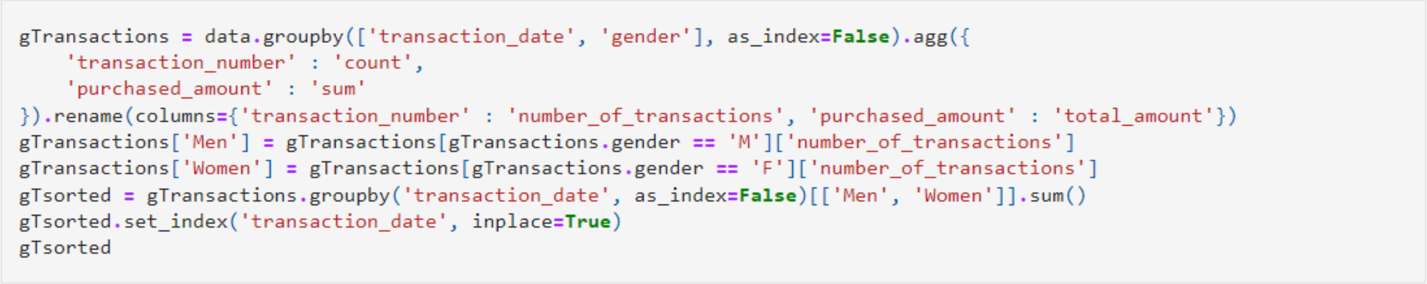
```
data[txt] = data[txt].astype('string')
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 15 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   transaction_date    1048575 non-null  datetime64[ns]
 1   card_number         1048575 non-null  string
 2   merchant            1048575 non-null  string
 3   purchased_amount    1048575 non-null  float64
 4   first               1048575 non-null  string
 5   last                1048575 non-null  string
 6   gender              1048575 non-null  string
 7   street              1048575 non-null  string
 8   city                1048575 non-null  string
 9   state               1048575 non-null  string
 10  zip                 1048575 non-null  int64
 11  city_population     1048575 non-null  int64
 12  job                 1048575 non-null  string
 13  date_of_birth       1048575 non-null  datetime64[ns]
 14  transaction_number  1048575 non-null  string
dtypes: datetime64[ns](2), float64(1), int64(2), string(10)
memory usage: 120.0 MB
```

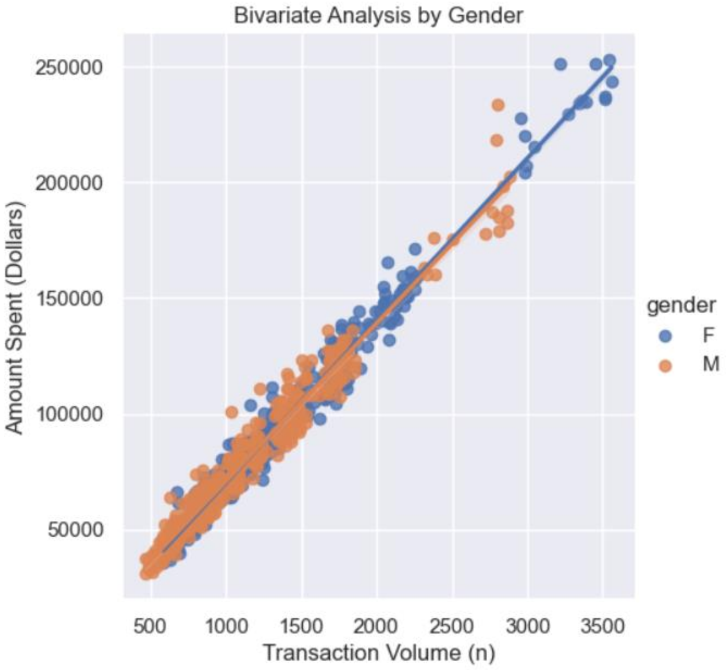
Gender Based Spending

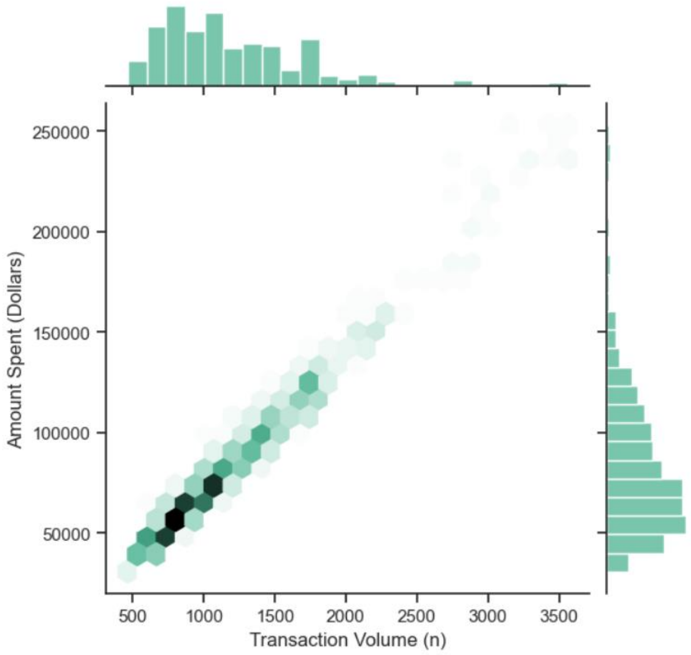
In this analysis we want to track down the mean difference on spending correlated from gender. By these we can observe which gender has the most potential in utilizing their credit cards, volume of transactions they usually spent, and amount per transaction.



```
gTransactions = data.groupby(['transaction_date', 'gender'], as_index=False).agg({
    'transaction_number' : 'count',
    'purchased_amount' : 'sum'
}).rename(columns={'transaction_number' : 'number_of_transactions', 'purchased_amount' : 'total_amount'})
gTransactions['Men'] = gTransactions[gTransactions.gender == 'M']['number_of_transactions']
gTransactions['Women'] = gTransactions[gTransactions.gender == 'F']['number_of_transactions']
gTsorted = gTransactions.groupby('transaction_date', as_index=False)[['Men', 'Women']].sum()
gTsorted.set_index('transaction_date', inplace=True)
gTsorted
```

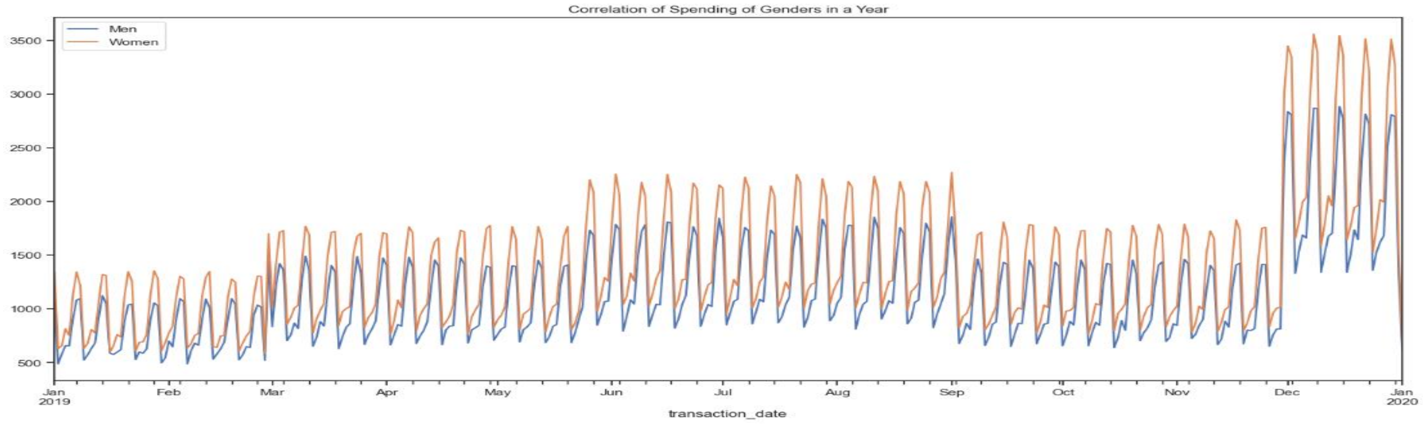
By plotting the data aggregates, we can visually comprehend the results.



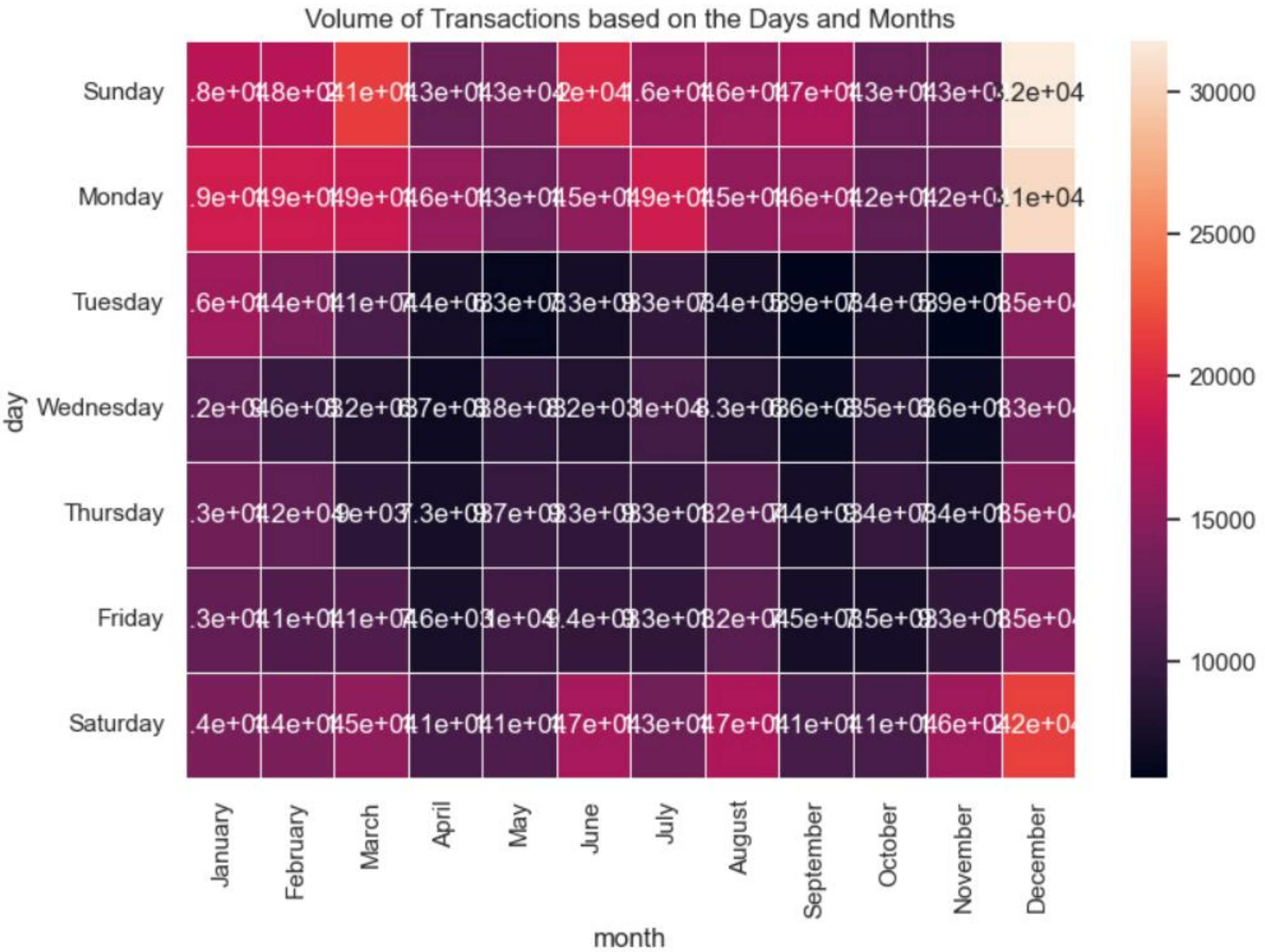


By bivariate analysis having amount spending and volume of transactions as its axis's. There is not much spread in both variables making it a low variance model. With low variance we can totally predict the trends by each transactions and spending. We can also observe that females tend to have more transactions and spend more, garnering both the maximum values in both variables.

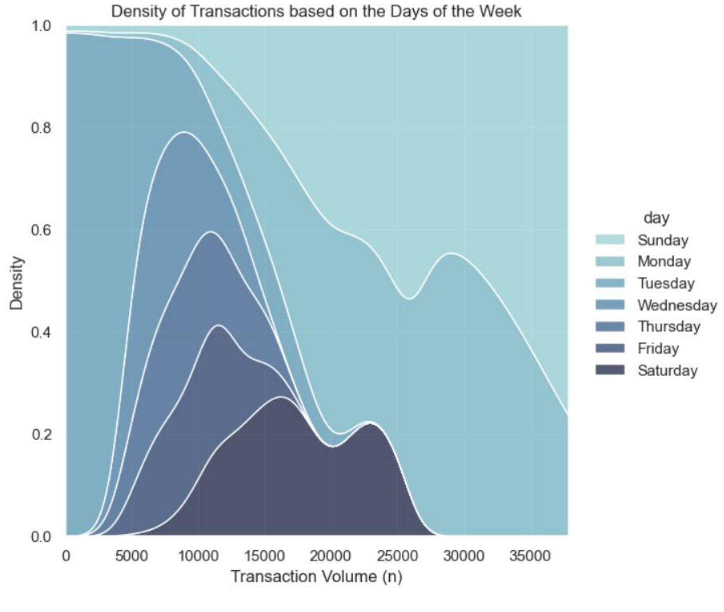
Also using joint plot, we can figure out the values with more instances of occurrence. With concentration of transaction volume between 1000 <= 1,500 and amount spending between 40,000 <=100,000.

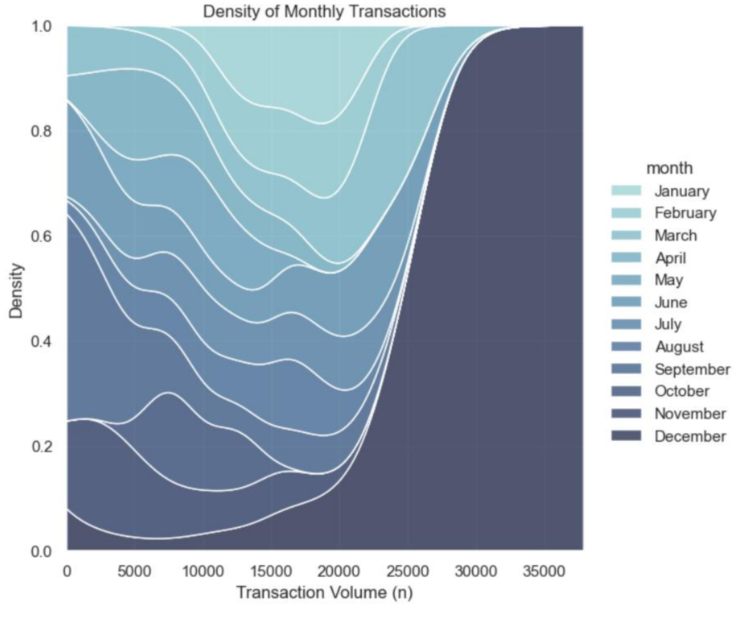


Timelines with Most Transactions and Amount Spending



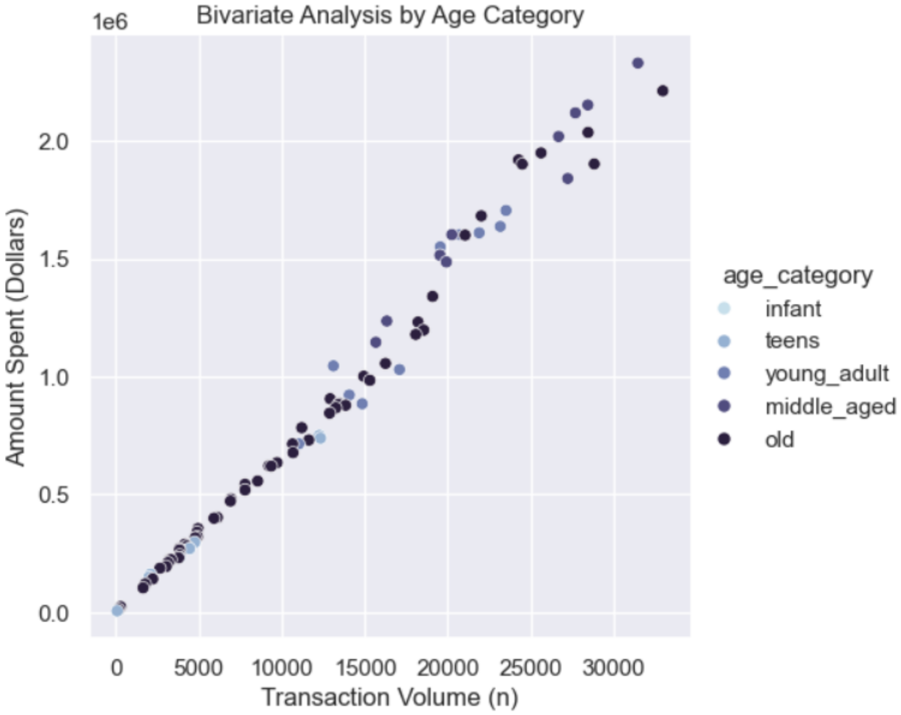
Heatmap visual is useful for this kind of analysis. Distinguishing values of occurrence by gradient assimilation. In here we can say that at the month of December and preferably at days of Sunday and Monday has the most numbers of transactions. We can deduce that most customers tend to spend more on weekends and at the holiday seasons than the rest of the year.



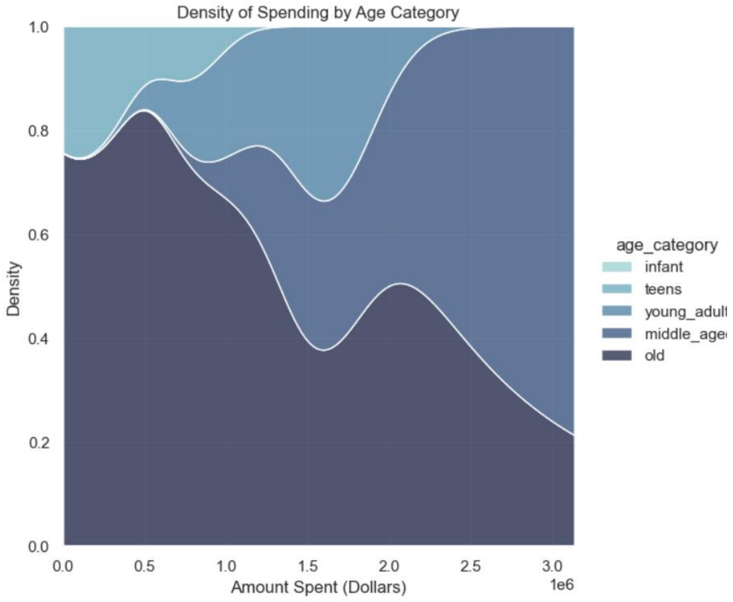


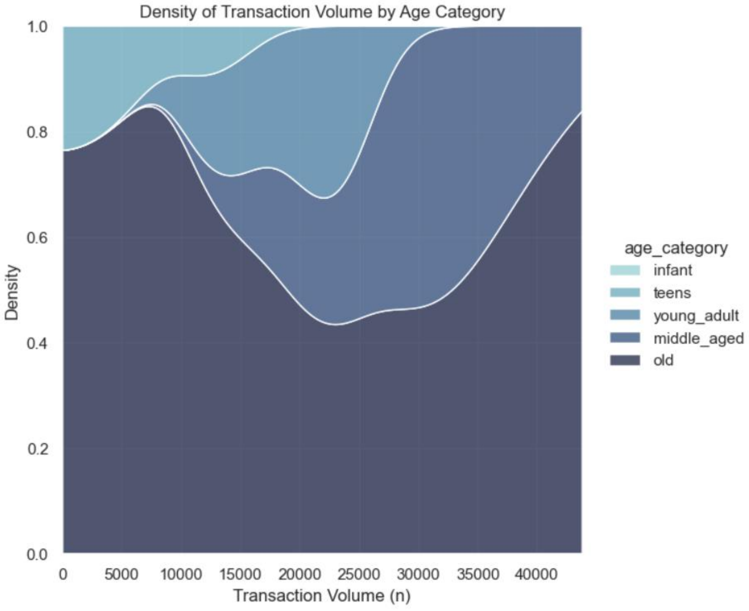
Age Group with Highest Spending Scores

This subset tends to discover which age groups has the highest trend in market transactions and amount spending. Knowing this we can track which customers at a specific age group has the highest potency in dealing with the market.



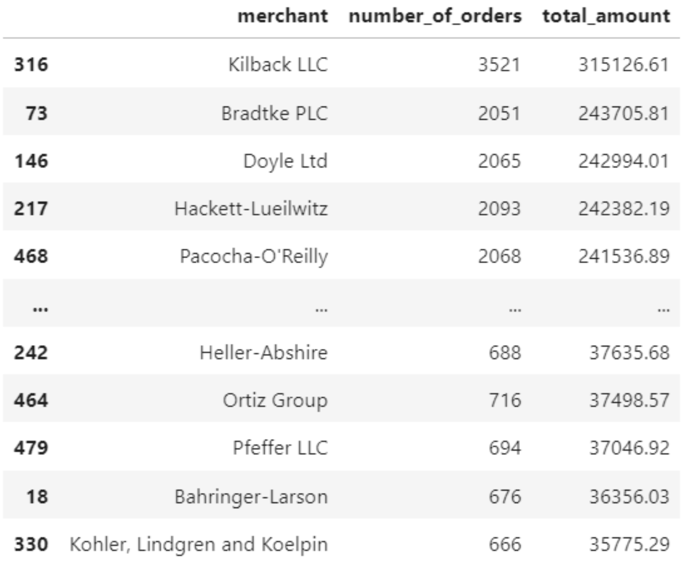
Using bivariate method, we can clearly see that there are more instances of acquisition in the old age group. This age group is binned with age value more than 50 years old. It is also understandable that there are no occurrences beyond infant age group with 10 years old less values. With spending alike with gender variance there is not much spread in this dataset thus the trend will persist.



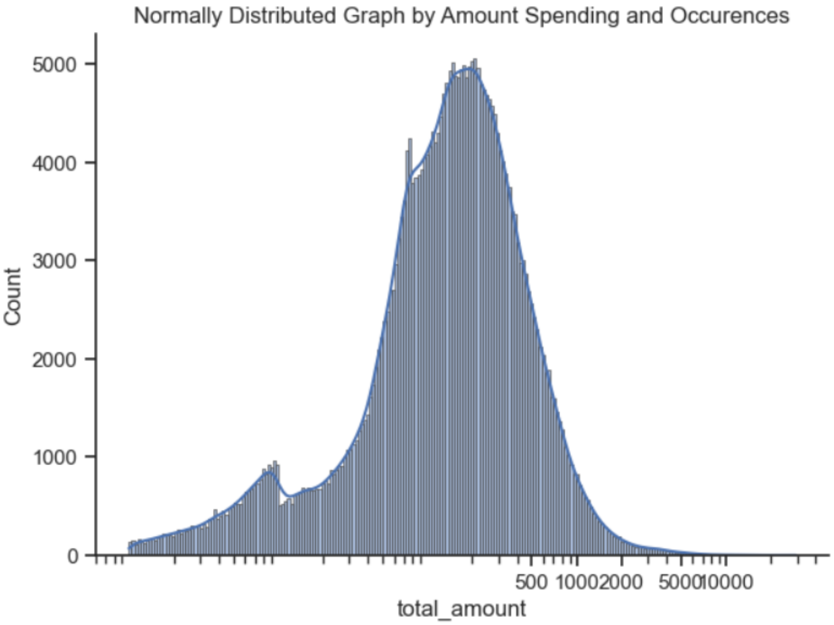


Merchant Companies with Most Value

This case study will help monitor which merchants does our customers have more likely to engage with. Being able to know which merchants have the most up value will let us predict trends of purchases the customers will persist on buying in the future.



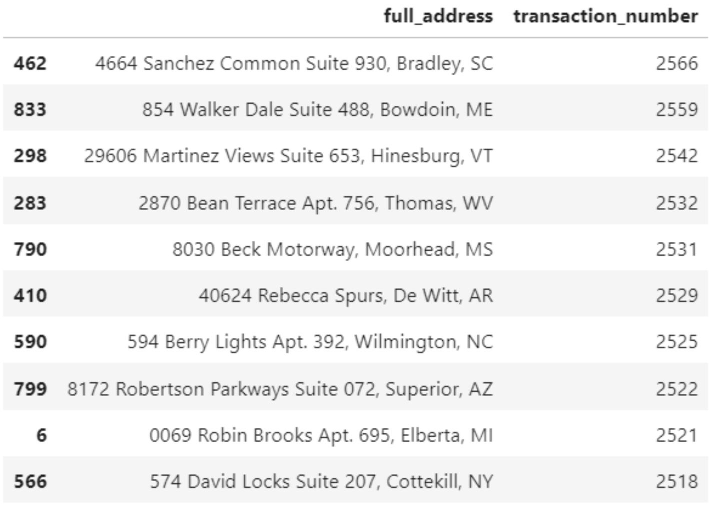
	merchant	number_of_orders	total_amount
316	Kilback LLC	3521	315126.61
73	Bradtke PLC	2051	243705.81
146	Doyle Ltd	2065	242994.01
217	Hackett-Lueilwitz	2093	242382.19
468	Pacocha-O'Reilly	2068	241536.89
...	...	...	...
242	Heller-Abshire	688	37635.68
464	Ortiz Group	716	37498.57
479	Pfeffer LLC	694	37046.92
18	Bahringer-Larson	676	36356.03
330	Kohler, Lindgren and Koelpin	666	35775.29



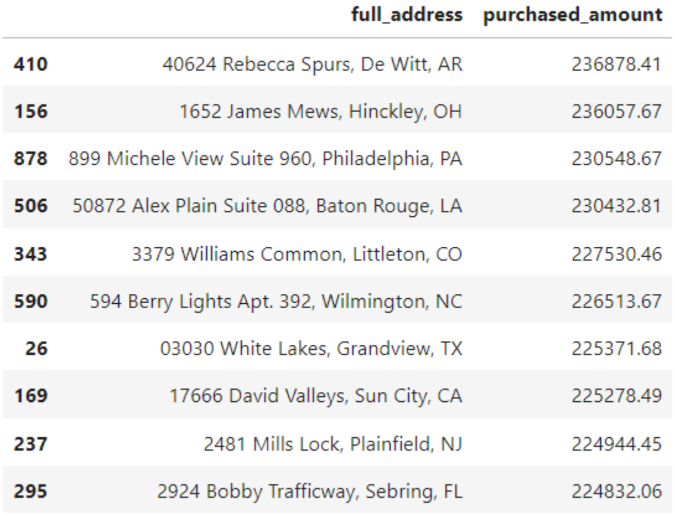
In here we showcase the top 5 merchants and the bottom 5 merchants by counting the number of transactions and summing the total amounts of purchasing. As the dataset creates a normally distributed graph, we can infer that the highest totals rest beyond the middle. Thus, values with more occurrences adheres to near the mean value.

Funneled Individuals with Highest Transaction Record and Amount Spending by Location

Funneling values helps the merchant develop a keen insight on which customers has given the most contribution by engaging in transactions and rate of spending. Detecting a specific persona by qualifying to addressed mechanics/variables of detection

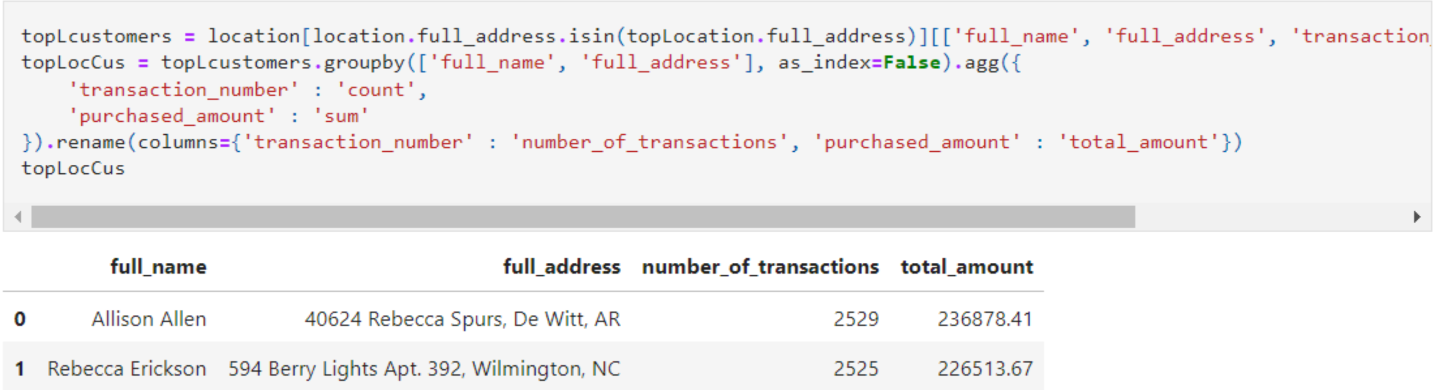


	full_address	transaction_number
462	4664 Sanchez Common Suite 930, Bradley, SC	2566
833	854 Walker Dale Suite 488, Bowdoin, ME	2559
298	29606 Martinez Views Suite 653, Hinesburg, VT	2542
283	2870 Bean Terrace Apt. 756, Thomas, WV	2532
790	8030 Beck Motorway, Moorhead, MS	2531
410	40624 Rebecca Spurs, De Witt, AR	2529
590	594 Berry Lights Apt. 392, Wilmington, NC	2525
799	8172 Robertson Parkways Suite 072, Superior, AZ	2522
6	0069 Robin Brooks Apt. 695, Elberta, MI	2521
566	574 David Locks Suite 207, Cottekill, NY	2518



	full_address	purchased_amount
410	40624 Rebecca Spurs, De Witt, AR	236878.41
156	1652 James Mews, Hinckley, OH	236057.67
878	899 Michele View Suite 960, Philadelphia, PA	230548.67
506	50872 Alex Plain Suite 088, Baton Rouge, LA	230432.81
343	3379 Williams Common, Littleton, CO	227530.46
590	594 Berry Lights Apt. 392, Wilmington, NC	226513.67
26	03030 White Lakes, Grandview, TX	225371.68
169	17666 David Valleys, Sun City, CA	225278.49
237	2481 Mills Lock, Plainfield, NJ	224944.45
295	2924 Bobby Trafficway, Sebring, FL	224832.06

Here we lessen the sample size by enforcing methods determining the locations with both highest transaction volume and highest rate of spending. Acquiring both the top 10 on each method, merging both data frames with inner values we are left with locations both present on each data frames. The left data is now mapped to the main data to track customers that has the congruency with the top locations.

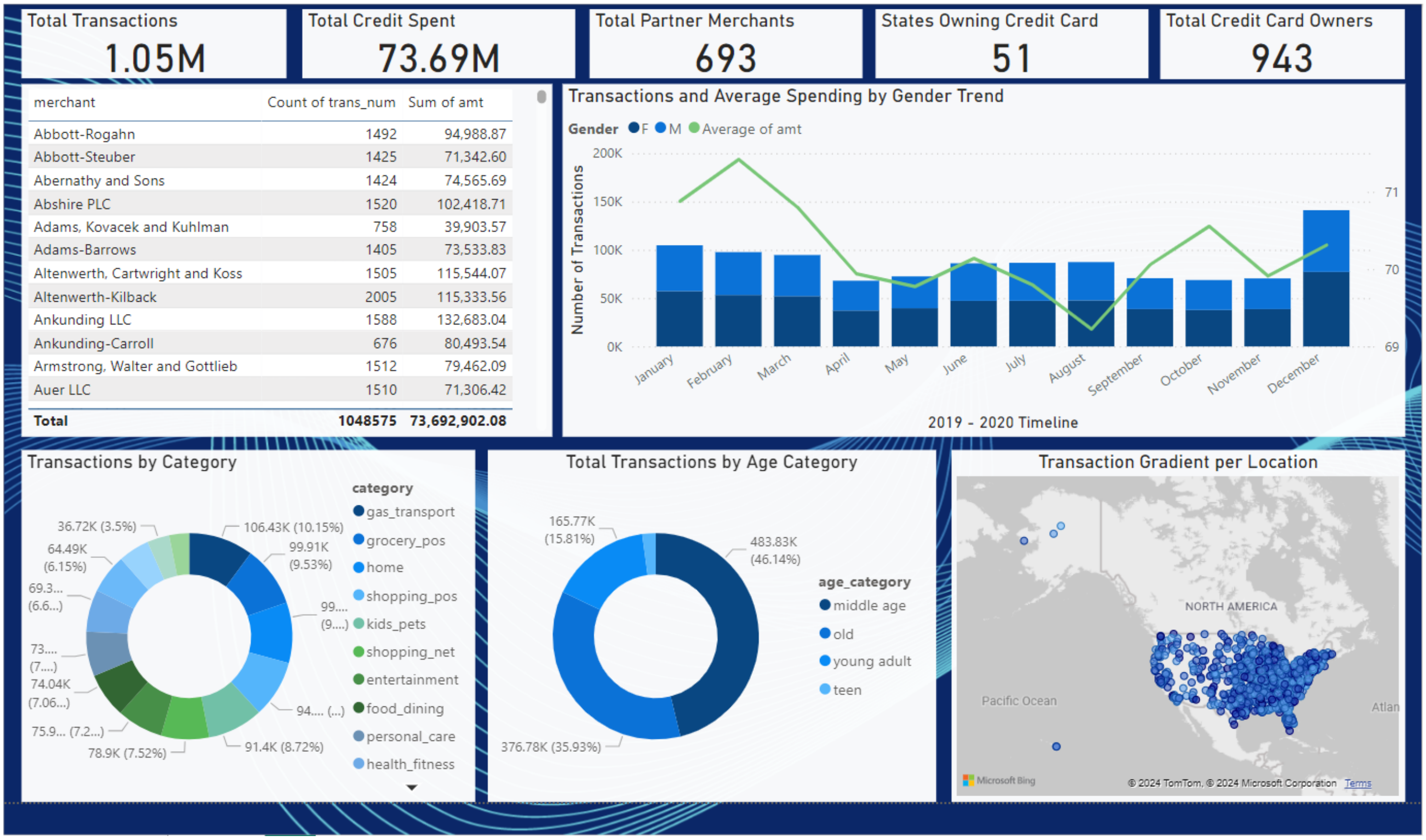


```
topLcustomers = location[location.full_address.isin(topLocation.full_address)][['full_name', 'full_address', 'transaction
topLocCus = topLcustomers.groupby(['full_name', 'full_address'], as_index=False).agg({
    'transaction_number' : 'count',
    'purchased_amount' : 'sum'
}).rename(columns={'transaction_number' : 'number_of_transactions', 'purchased_amount' : 'total_amount'})
topLocCus
```

	full_name	full_address	number_of_transactions	total_amount
0	Allison Allen	40624 Rebecca Spurs, De Witt, AR	2529	236878.41
1	Rebecca Erickson	594 Berry Lights Apt. 392, Wilmington, NC	2525	226513.67

Finally passing the argument the result has yielded the names of Allison Allen and Rebecca Erickson. The findings do not reflect on the maximums of both variables but, with the congruence of both variables by location.

Sample Dashboard Created in Power BI



Total Transactions 1.05M

Total Credit Spent 73.69M

Total Partner Merchants 693

States Owning Credit Card 51

Total Credit Card Owners 943

merchant	Count of trans_num	Sum of amt
Abbott-Rogahn	1492	94,988.87
Abbott-Steuber	1425	71,342.60
Abernathy and Sons	1424	74,565.69
Abshire PLC	1520	102,418.71
Adams, Kovacek and Kuhlman	758	39,903.57
Adams-Barrows	1405	73,533.83
Altenwerth, Cartwright and Koss	1505	115,544.07
Altenwerth-Kilback	2005	115,333.56
Ankunding LLC	1588	132,683.04
Ankunding-Carroll	676	80,493.54
Armstrong, Walter and Gottlieb	1512	79,462.09
Auer LLC	1510	71,306.42
Total	**1048575**	**73,692,902.08**