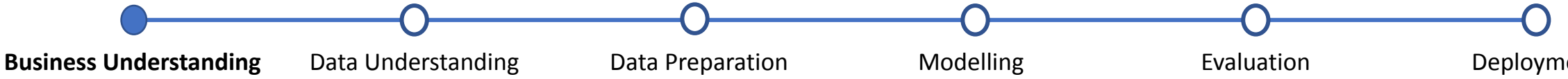
Leveraging Predictive Models on Real Estate Prices and Profitable Consumer Credit Card Data Segments

Group 5: Allison Tsang, Reshma Roy, Wangheng (Maggie) Hu, Rishik Adhikari



The start of solving any business problem begins with understanding the business and then translating it into an analytical problem that is solvable through data

Questions from Business Partners

How can we leverage this data? What should we do with this data? Is there anything valuable we can use this for?

Business Problem

- Inquiring about objectives, goals, plans, strategies specific to the business team or company

Who is the business/project owner?

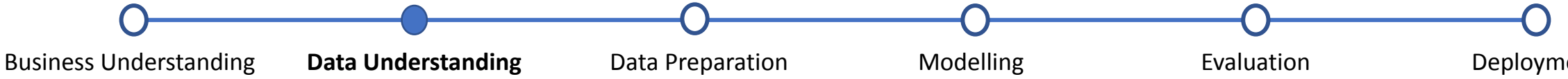
- Real Estate data can be used by Gov't/Non-profits, apps, marketplaces (Craigslist), etc
- Credit Card Data can by used by the companies themselves (marketing department), banks, market research firms, insurance companies,

Analytical Problem

- Translating the business problem into an analytical one
- Ensuing collected data can solve the problem

What is their goal?

- Predicting monthly rental pricing of properties based on property attributes (ie. # of bedrooms, sqft, etc.,) to help property listers
- Which subgroups/segments of customers would most likely respond to a credit card campaign.



Acquire Data

- Data acquired from Kaggle
- Typically, both sources would likely come internally

Discover & Access

Both data sets had many measures (attributes) within it. The team spent most of the time on data understanding through the step below:

1. Understanding of Attributes
2. Check Data Load
3. Check Data Types of each Column
4. Check for None values
5. **Correlation of Variables**
6. Skewness
7. **Outliers**
8. Data Normalization
9. **Exploratory Charts/Graphs to understand data**

Real Estate Data set:

- 384977 entries
- 22 columns
- dtypes: float64(3), int64(10), object(9)

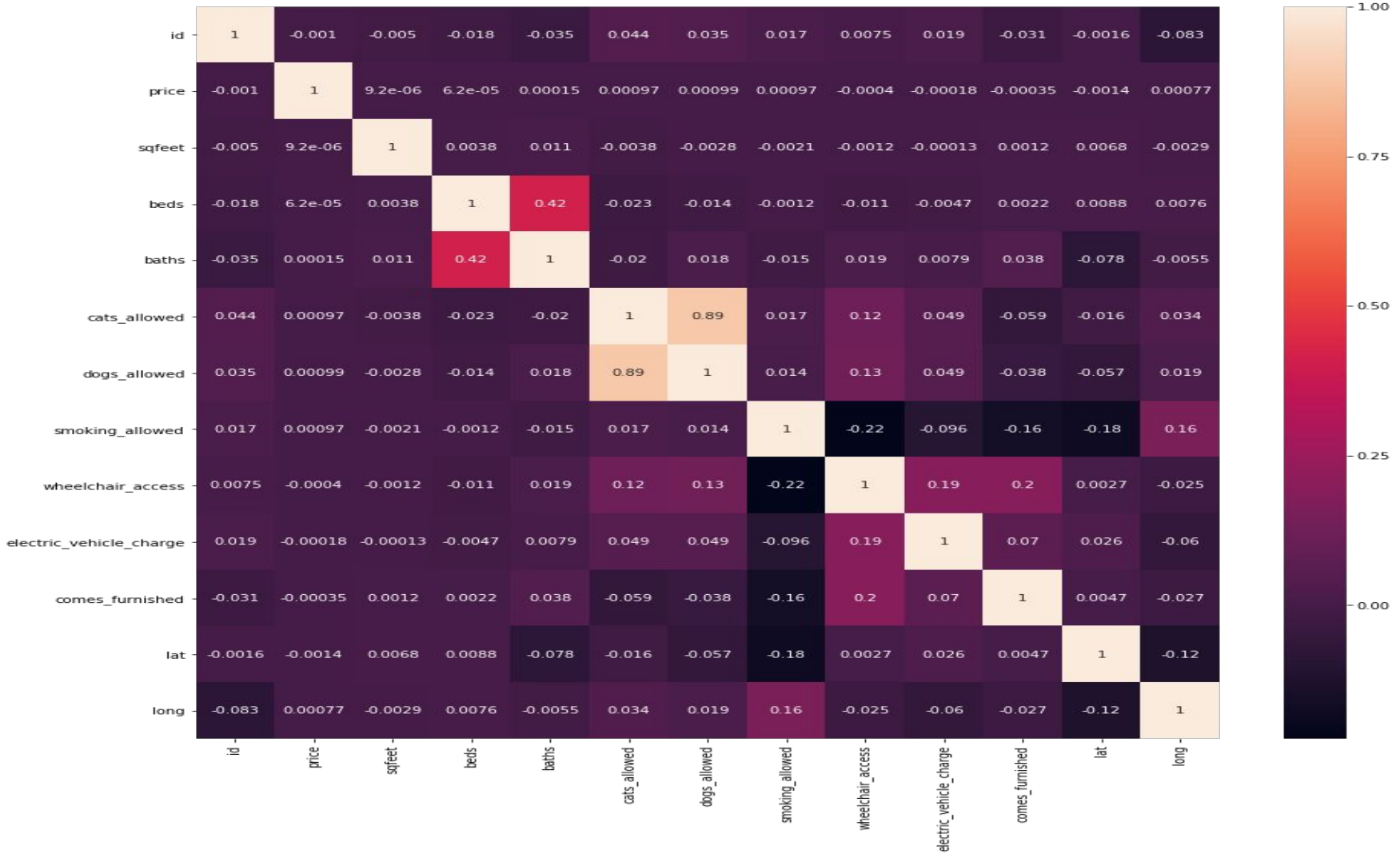
id	0
url	0
region	0
region_url	0
price	0
type	0
sqfeet	0
beds	0
baths	0
cats_allowed	0
dogs_allowed	0
smoking_allowed	0
wheelchair_access	0
electric_vehicle_charge	0
comes_furnished	0
laundry_options	79026
parking_options	140687
image_url	0
description	2
lat	1918
long	1918
state	0

Credit Card Dataset:

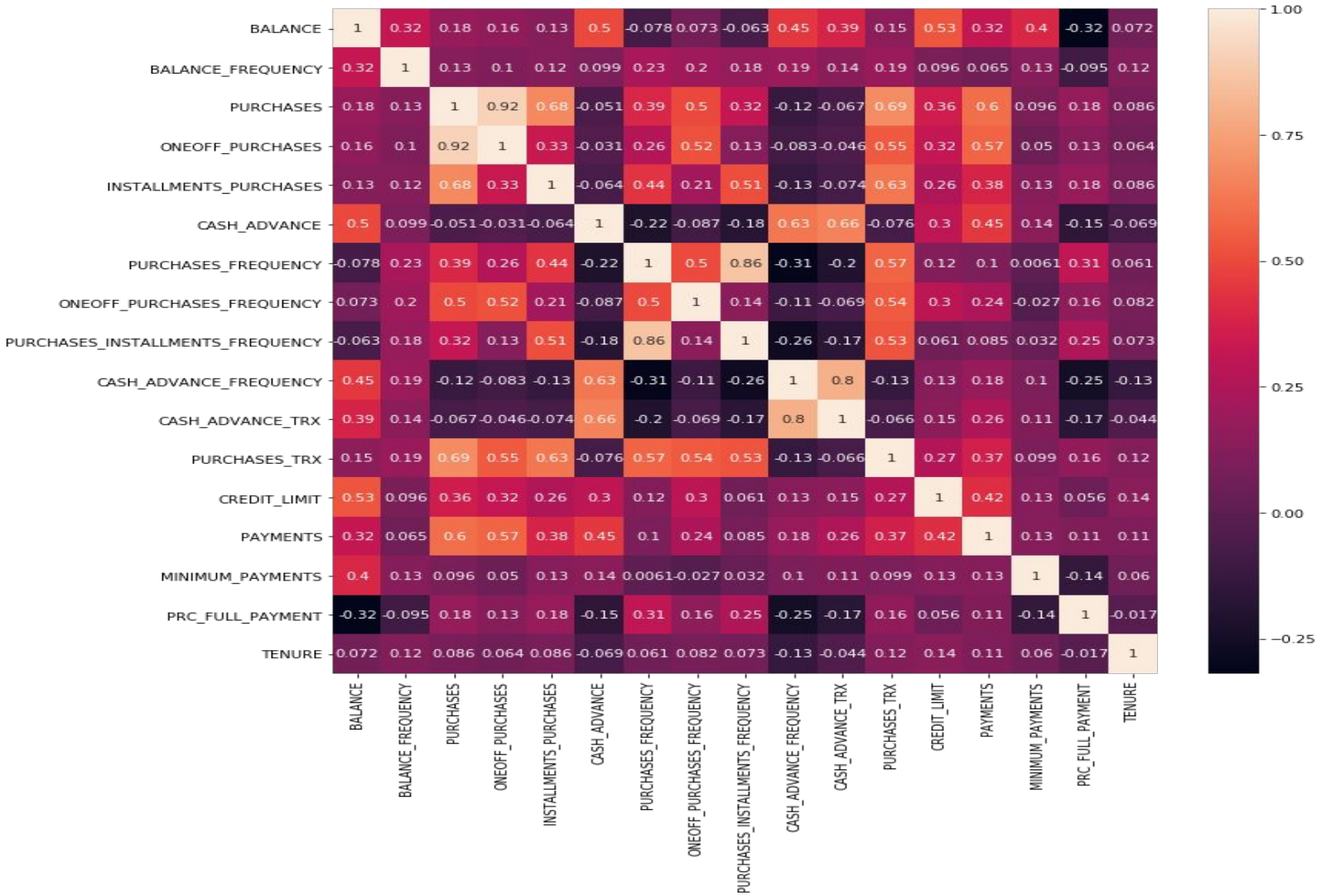
- 8950 entries
- 17 columns
- dtypes: float64(14), int64(3)

CUST_ID	0
BALANCE	0
BALANCE_FREQUENCY	0
PURCHASES	0
ONEOFF_PURCHASES	0
INSTALLMENTS_PURCHASES	0
CASH_ADVANCE	0
PURCHASES_FREQUENCY	0
ONEOFF_PURCHASES_FREQUENCY	0
PURCHASES_INSTALLMENTS_FREQUENCY	0
CASH_ADVANCE_FREQUENCY	0
CASH_ADVANCE_TRX	0
PURCHASES_TRX	0
CREDIT_LIMIT	1
PAYMENTS	0
MINIMUM_PAYMENTS	313
PRC_FULL_PAYMENT	0
TENURE	0

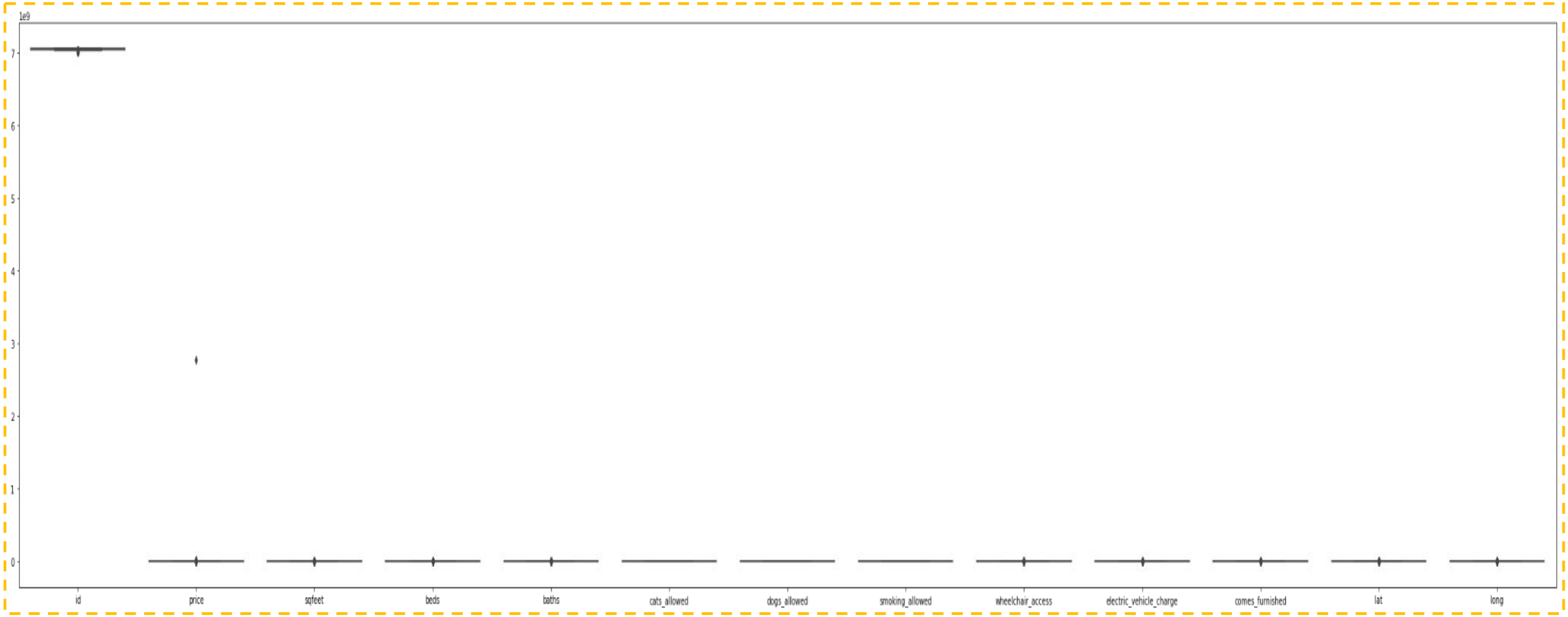
Real Estate : Correlation Matrix



Credit Card : Correlation Matrix



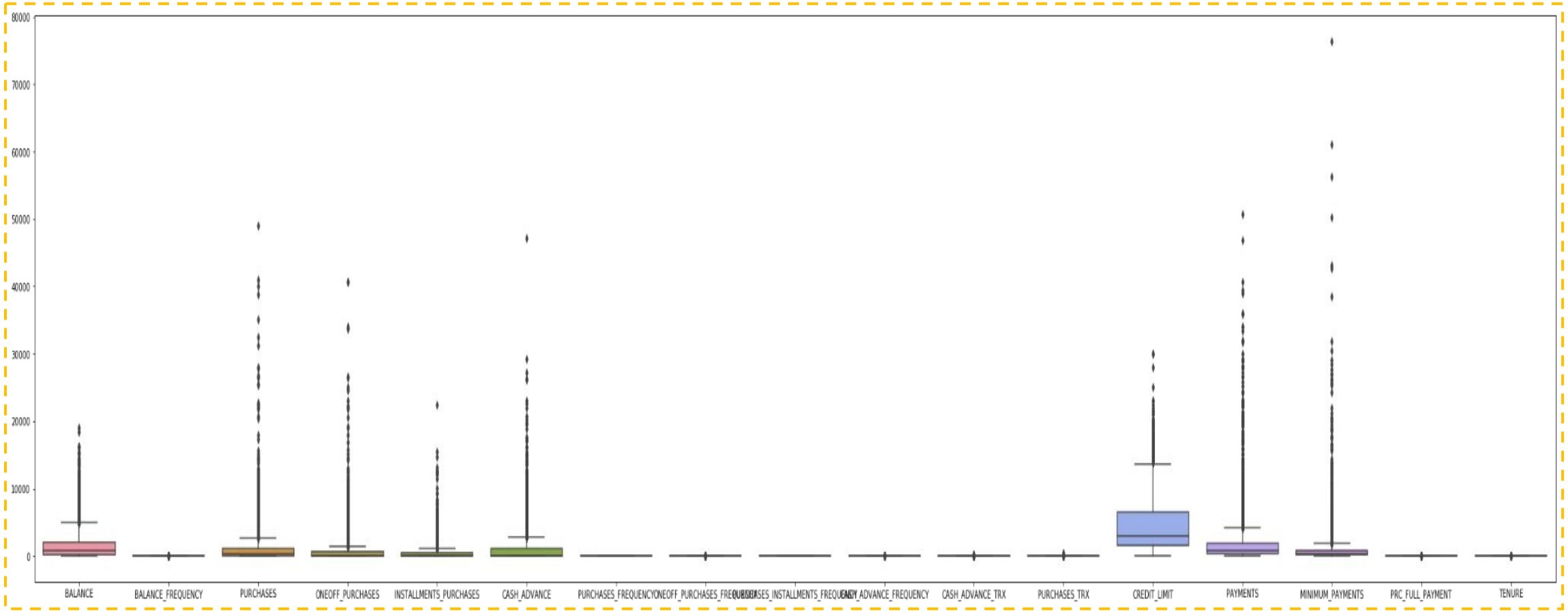
Real Estate: Box plot



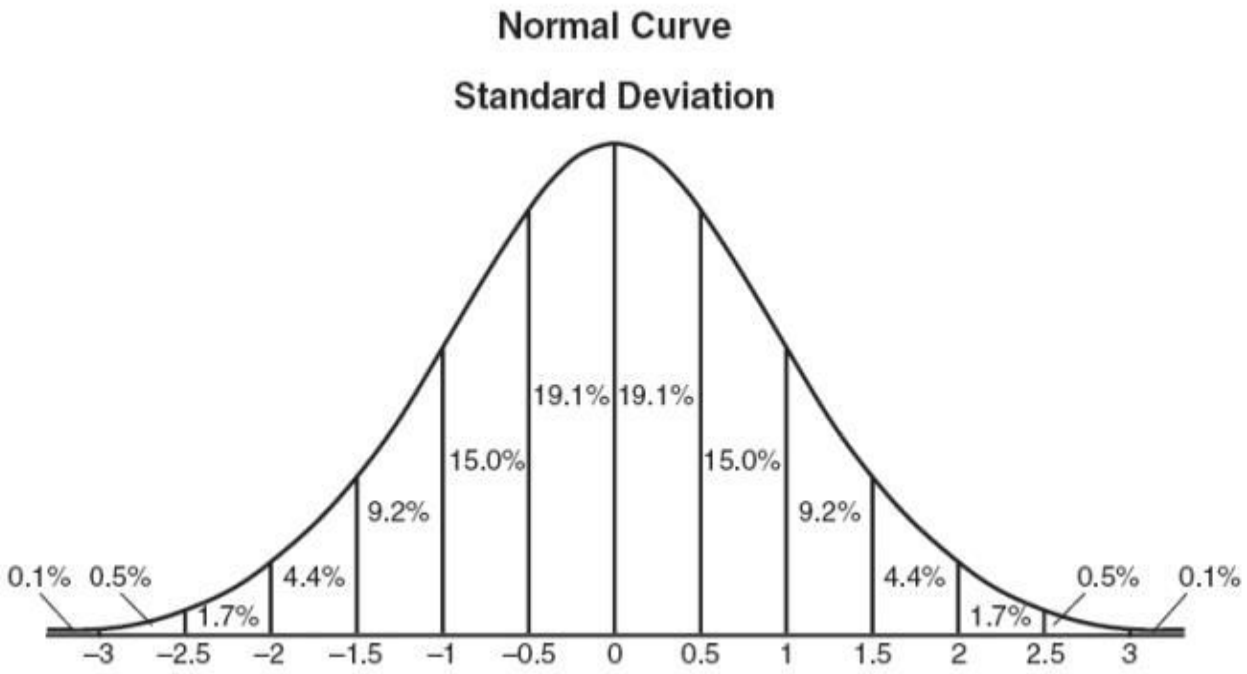
Real Estate: Summary

	count	mean	std	min	25%	50%	75%	max
id	384977.0	7.040982e+09	8.800376e+06	7.003808e+09	7.035979e+09	7.043320e+09	7.048426e+09	7.051292e+09
price	384977.0	8.825722e+03	4.462200e+06	0.000000e+00	8.050000e+02	1.036000e+03	1.395000e+03	2.768307e+09
sqfeet	384977.0	1.059900e+03	1.915076e+04	0.000000e+00	7.500000e+02	9.490000e+02	1.150000e+03	8.388607e+06
beds	384977.0	1.905345e+00	3.494572e+00	0.000000e+00	1.000000e+00	2.000000e+00	2.000000e+00	1.100000e+03
baths	384977.0	1.480718e+00	6.180605e-01	0.000000e+00	1.000000e+00	1.000000e+00	2.000000e+00	7.500000e+01
cats_allowed	384977.0	7.268902e-01	4.455574e-01	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
dogs_allowed	384977.0	7.079176e-01	4.547206e-01	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
smoking_allowed	384977.0	7.317710e-01	4.430381e-01	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
wheelchair_access	384977.0	8.211140e-02	2.745347e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
electric_vehicle_charge	384977.0	1.287090e-02	1.127177e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
comes_furnished	384977.0	4.812755e-02	2.140360e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
lat	383059.0	3.723349e+01	5.546171e+00	-4.353330e+01	3.345470e+01	3.764780e+01	4.113830e+01	1.020360e+02
long	383059.0	-9.270063e+01	1.653198e+01	-1.638940e+02	-1.007750e+02	-8.774510e+01	-8.117960e+01	1.726330e+02

Credit card: Box plot



Augment & Integration



- Examples of dropped columns:
  - Real Estate: 'id', 'url', 'region', 'region_url', 'laundry_options', 'parking_options', 'image_url', 'description','lat',long', 'state'
  - Credit Card: 'CUST_ID', 'CREDIT_LIMIT'
- Blank data cells filled

- To clean both data sets, any data outside 3 standard deviations of the dataset was removed

A multiple regression would be best suited for US real estate price.

```
In [50]: #Converting the coefficient values to a dataframe
coeffcients = pd.DataFrame([x_train.columns,lm.coef_]).T
coeffcients = coeffcients.rename(columns={0: 'Attribute', 1: 'Coefficients'})
coeffcients
```

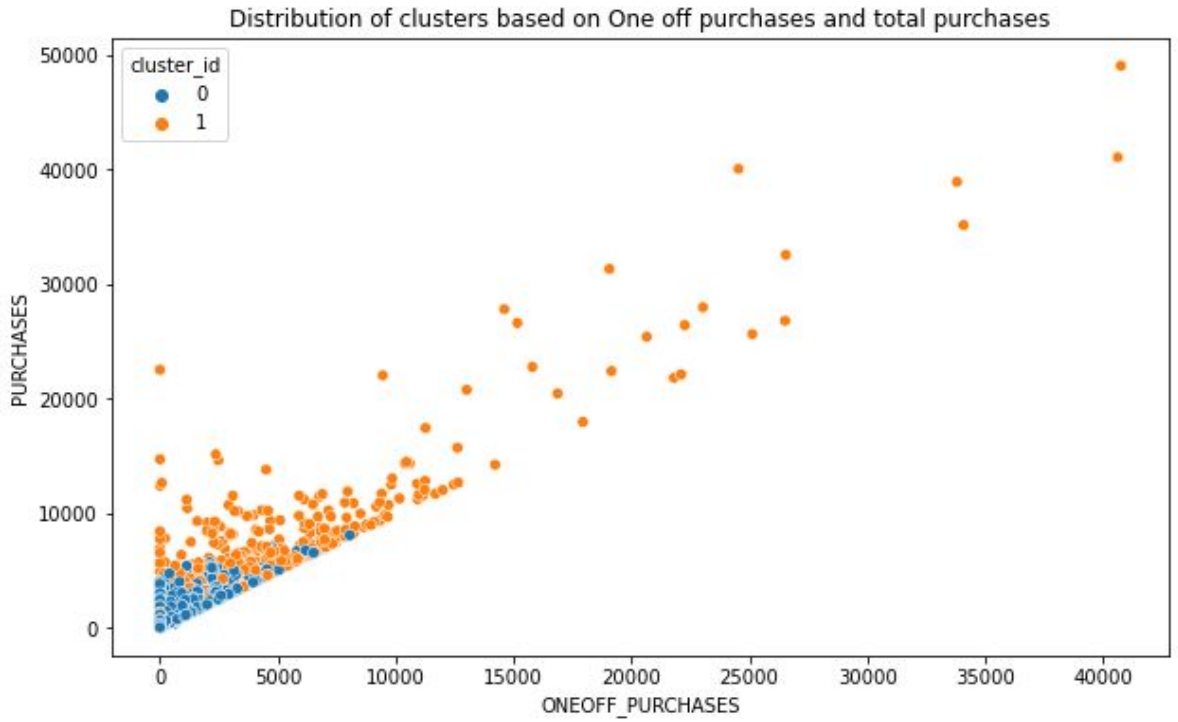
Out[50]:

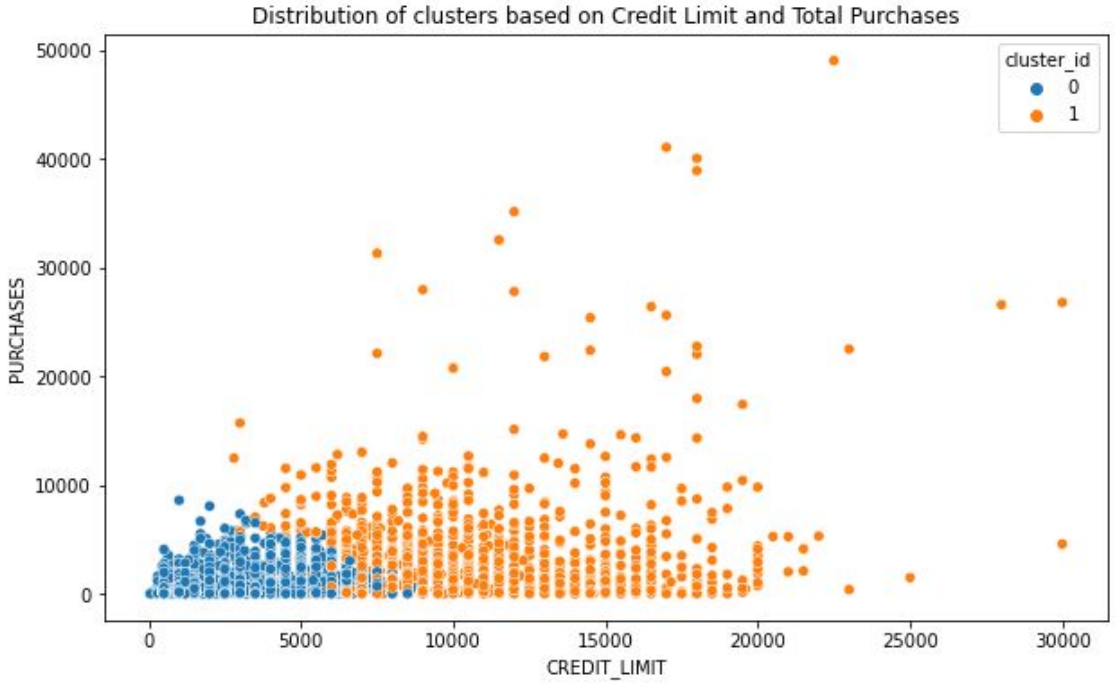
	Attribute	Coefficients
0	sqfeet	0.00266053
1	beds	9.86725
2	baths	1528.81
3	cats_allowed	6692.08
4	dogs_allowed	8564.05
5	smoking_allowed	12467.4
6	wheelchair_access	-7746.14
7	electric_vehicle_charge	-4575.59
8	comes_furnished	-2859.42

Multiple Linear Regression Model:

price = 0.0027 sqfeet + 9.87 beds + 1528.81 baths + 6692.08 cats_allowed + 8564.05 dogs allowed + 12467.4 smoking_allowed - 7746.14 wheelchair_access - 4575.59 electric_vehicle change +-2859.42 comes_furnished

- **KMeans clustering algorithm works best for credit card example.**
- **KMeans clustering is an unsupervised clustering algorithm which groups together similar data in same cluster to form k clusters.**





Our model has clustered customers with low usage of credit card in one cluster and model with higher usage of clusters in other.

- **R square provides a measure of how well observed outcomes are replicated by the model. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values.**
- **The mean squared error (MSE) tells you how close a regression line is to a set of points. The smaller the MSE, the closer you are to finding the line of fit.**

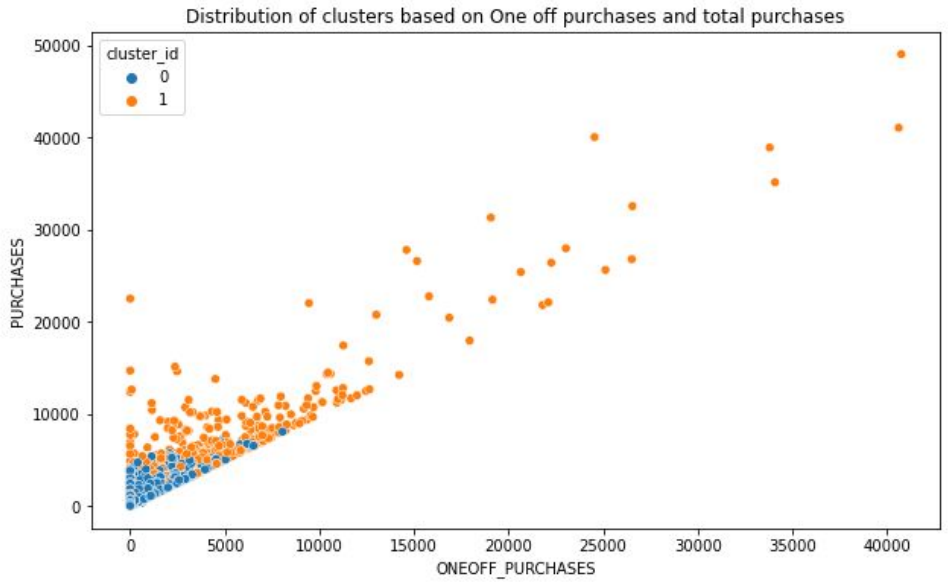
```
R^2: -0.09316107328237089
Adjusted R^2: -0.09319758324685545
MAE: 12732.988301940906
MSE: 2356332008.284829
RMSE: 48542.06431832941
```

- **A negative R square in US real estate price model indicates that the regression line we are currently using is worse than the mean value.**
- **A relatively large mean squared error (MSE) in this example also shows that this multiple linear regression model is not as good as we expect.**

- **One way to measure the model is to plot the graph of n versus innertia and as we plot the graph we can spot an elbow after which the innertia at a much lower rate.**
- **Silhouette score is a metric used to calculate the goodness of a clustering technique. The silhouette score of 1 means that the clusters are very dense and nicely separated. The score of 0 means that clusters are overlapping.**

```
from sklearn.metrics import silhouette_score
kmeans = KMeans(n_clusters=2, random_state=23)
kmeans.fit(X_red)
print('Silhoutte score of our model is' + str(silhouette_score(X_red,kmeans.labels_)))
```

```
Silhoutte score of our model is0.5208467678425495
```



- This seems like better clustering as it does segment out top half of customers having more than usual amount of usage of credit card and the customers which have very low usage. This seems to be more practical result if we want to direct our marketing strategies according to usage of credit card.

Deployment & Limitation

Deployment

- Assuming the data was processed internally, both companies with this type of data likely already have some sort of data infrastructure
- Data can be fed directly into Jupyter Notebook and code can be re-run.
- Model can be saved and implemented on various platforms such as softwares and websites.
- In this case, the regression model used in Real Estate can be deployed in websites for home buyers to know an estimated price and can be used by real estate agents to do a fair prediction of home prices.
- And the clustering model can be deployed as a software for Banks to launch a credit card campaign by targeting specific group of people who will most likely engage in that campaign.

Limitations

- Rental data may become obsolete if prices go up in an area, historical data may bring down the accuracy.
- Customer data / segment data should be refreshed to ensure that customers are still with the target segment of a campaign.

Q&A