TheAnalyticsTeam

Sprocket Central Pty Ltd

Data analytics approach

[Division Name] - [Engagement Manager], [Senior Consultant], [Junior Consultant]

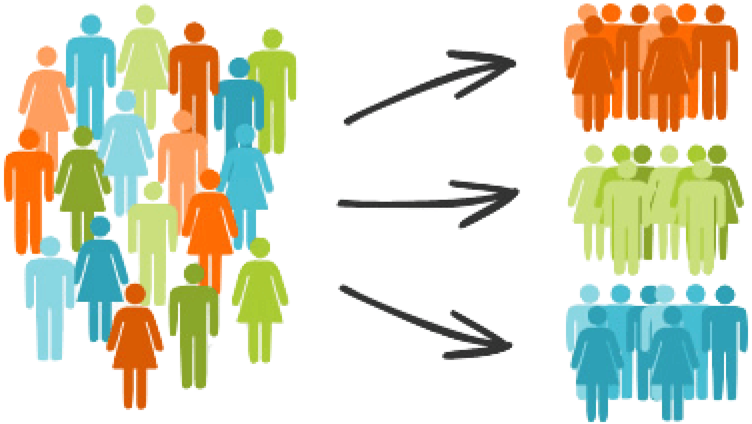
Agenda

1. Introduction
2. Data Exploration
3. Model Development
4. Model Evaluation
5. Interpretation

Introduction

We are here to identify top 1000 Customer to boost business by analysing their existing customer dataset to determine customer behaviour and trend.

To analyse customer behaviour and trend, we are using 3 dataset provided by **Sprocket Central company** that specializes in high-quality bikes and cycling accessories.

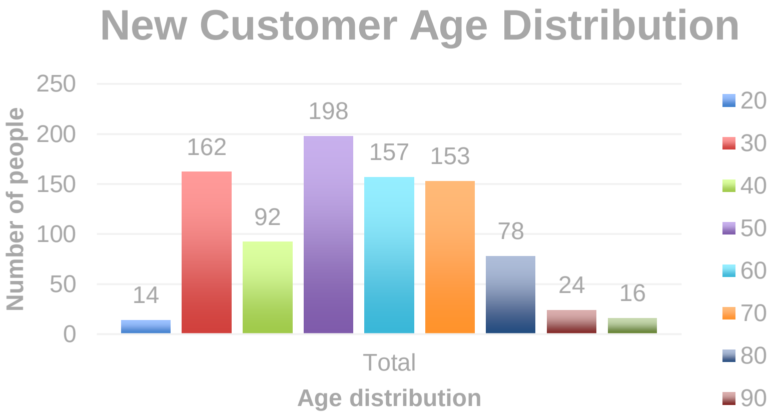


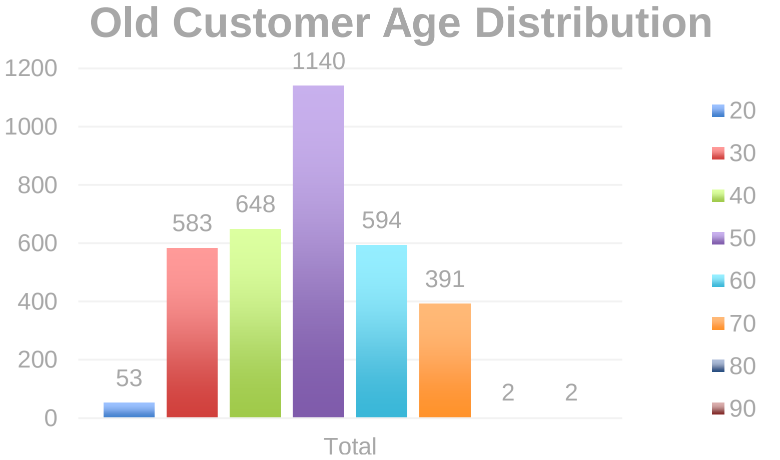
Customer Segmentation

Data Exploration

Customer Age Distributions

- In both graphs, most customers are aged between 30-50.
- In New customer list, there is slight increase in number of customer over 59 years old.
- The lowest age groups are under 20 and over 80 in New customer list and Old customer list respectively.
- The old customer list suggest 50-59 age group is most populated.

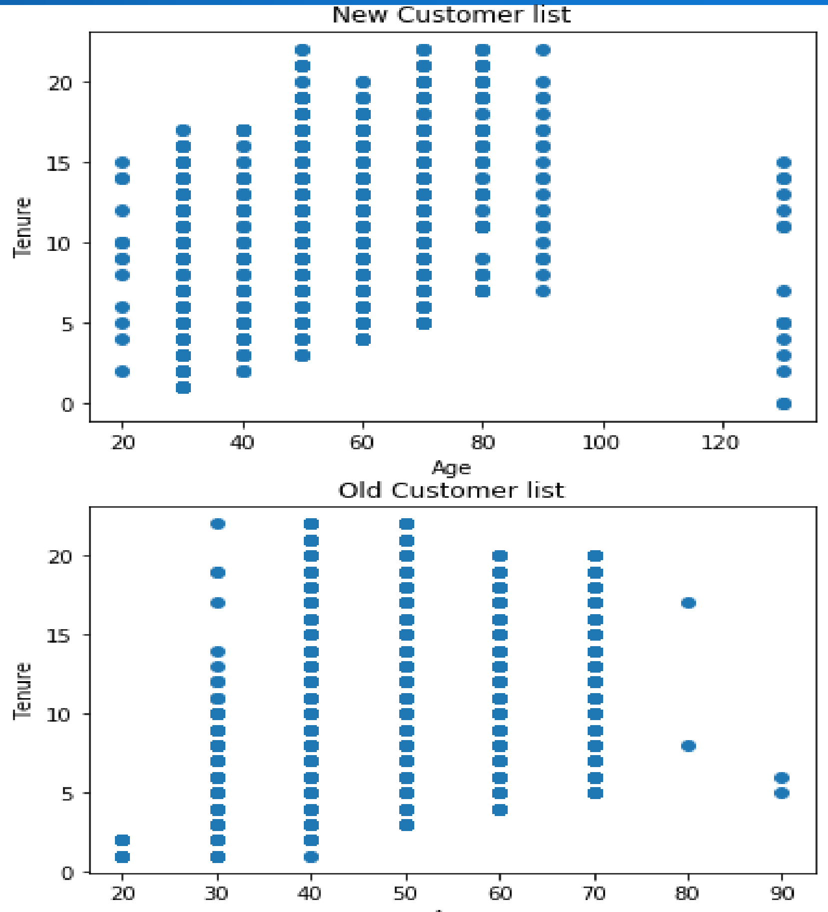




Data Exploration

Scatter plot between Age & tenure

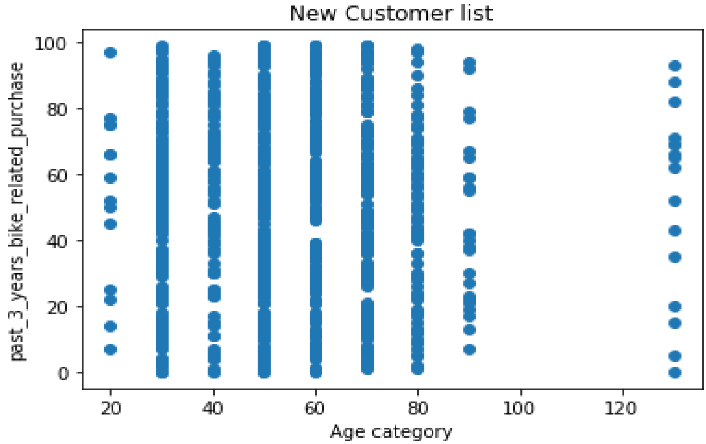
- There is Outliers in New customer list at 120 age group, which we have to remove.
- High tenure under 30-59 age group in Old customer list.
- High tenure under 50-99 age group in New customer list.

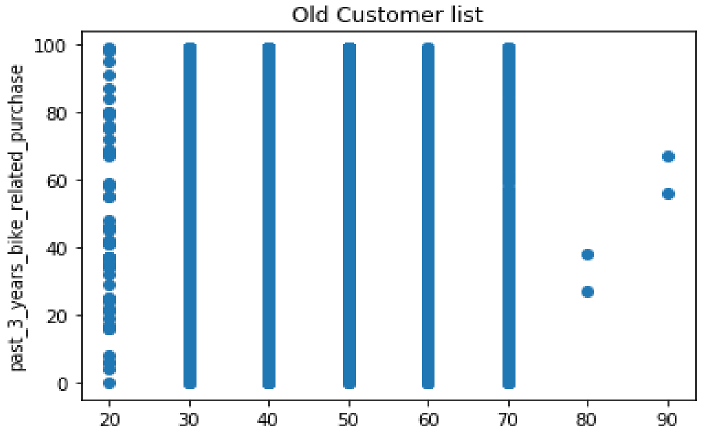


Data Exploration

Scatter plot between Age & Bike related purchase

- High amount of bike related purchases for almost each age group in Old customer list.
- Only few customer have low amount of purchases under 80-99 age group in Old customer list.
- In New Customer list, we have high amount of purchasing under 30-89 age group.

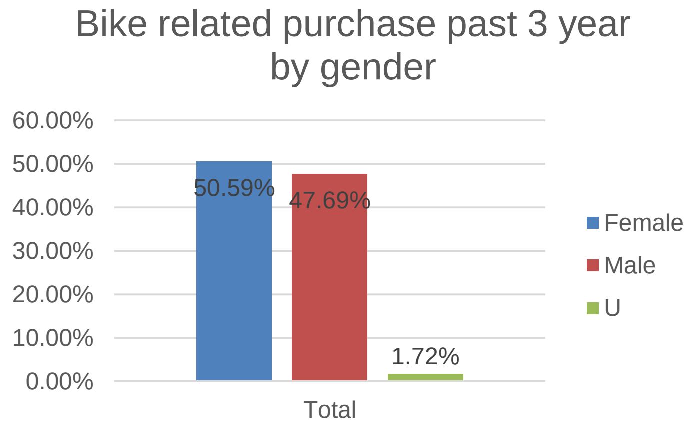


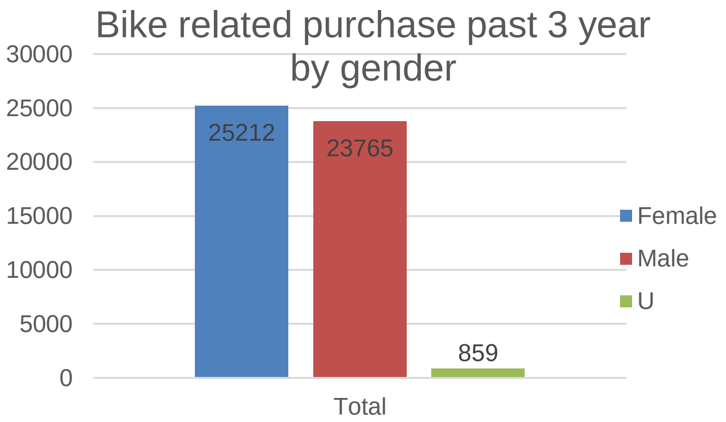


Data Exploration

Bike related purchases over last 3 years by gender

- Over the last 3 years about 50% of bike related purchases were made by females to 47% of purchases made by males. Approximately 2% were made by unknown gender.
- Numerically female purchases 25212 and male purchases 23765.
- So we can focus a little more on female customer than male customers.

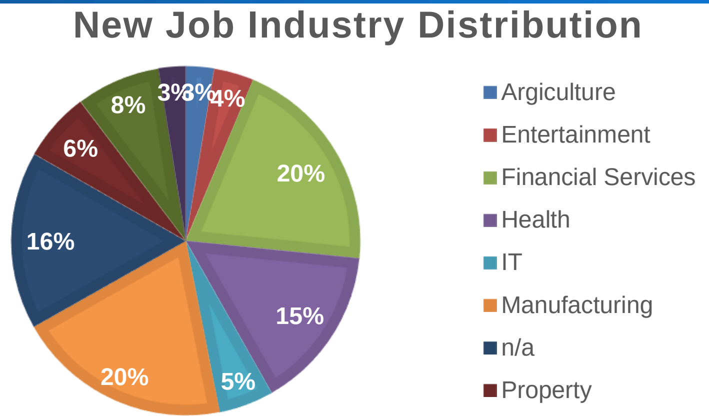


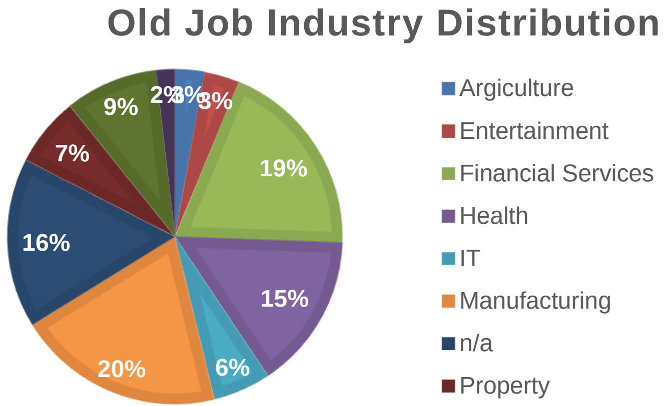


Data Exploration

Job Industry Distribution

- Most of New customers are from Manufacturing and financial services.
- Small number of customer are in Agriculture, Telecommunication and Entertainment.
- Most of old customers are also in Manufacturing and Financial Services.

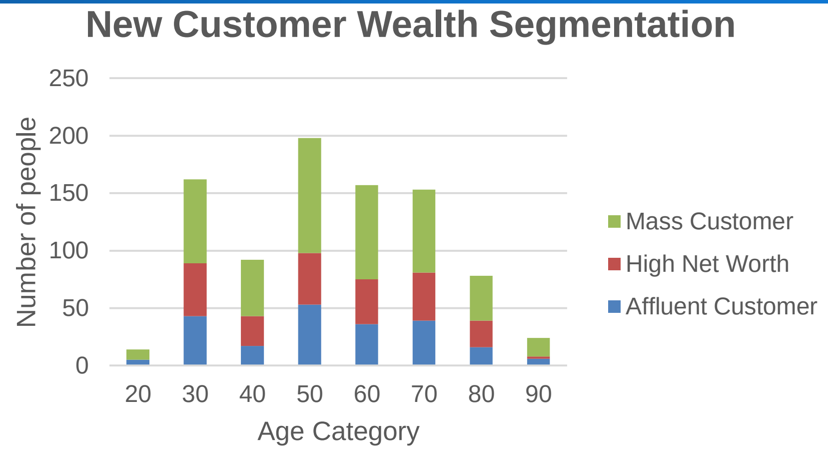


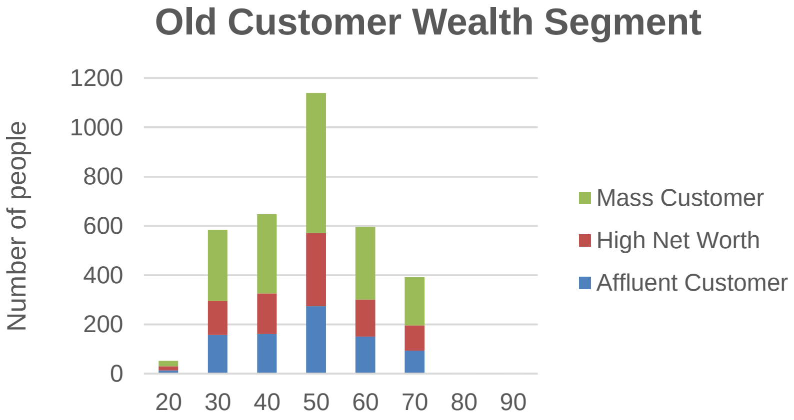


Data Exploration

Wealth Segmentation by age

- In all age categories, number of Mass customers is high, so we should focus on this class.
- The second highest class is High Net worth.
- But Affluent customer can outperforms the High Net worth customer in 50-59 age group.

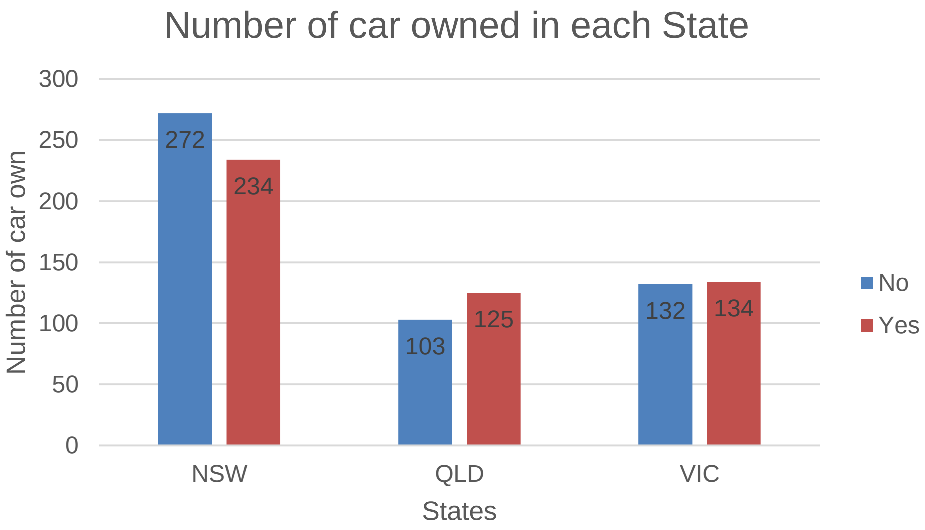




Data Exploration

Number of cars owned distribution by States.

- NSW has the largest amount of people that do not own a car. Also NSW is high populated.
- Victoria is quite even.
- QLD has relatively high number of customers that own a car.



Model Development

RFM Analysis and Customer Classification

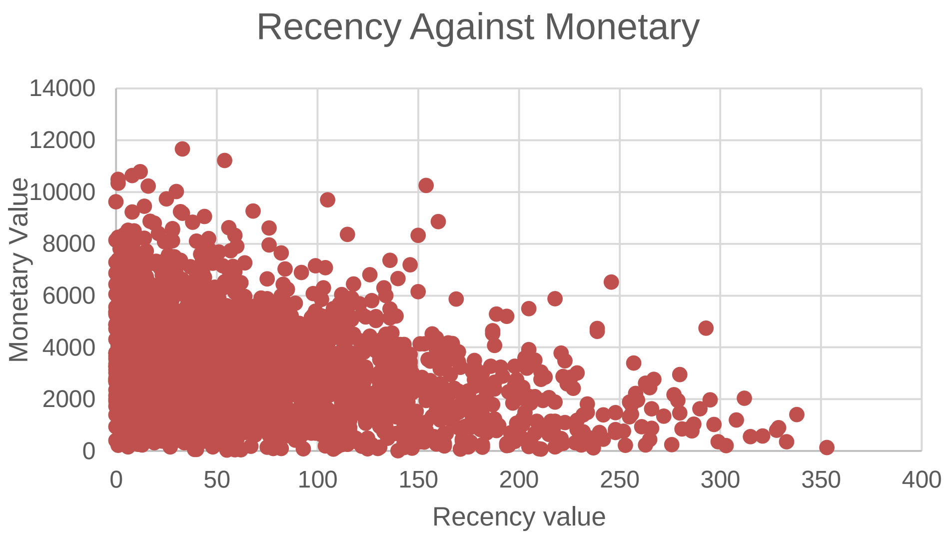
- We are doing RFM analysis to predict the valuable customers. It is used to increase business revenue and value.

- This method includes Recency, Frequency and Monetary values that shows number of customers that have displayed high level of engagement with the business.

Model Development

Scatter-Plot based on RFM Analysis

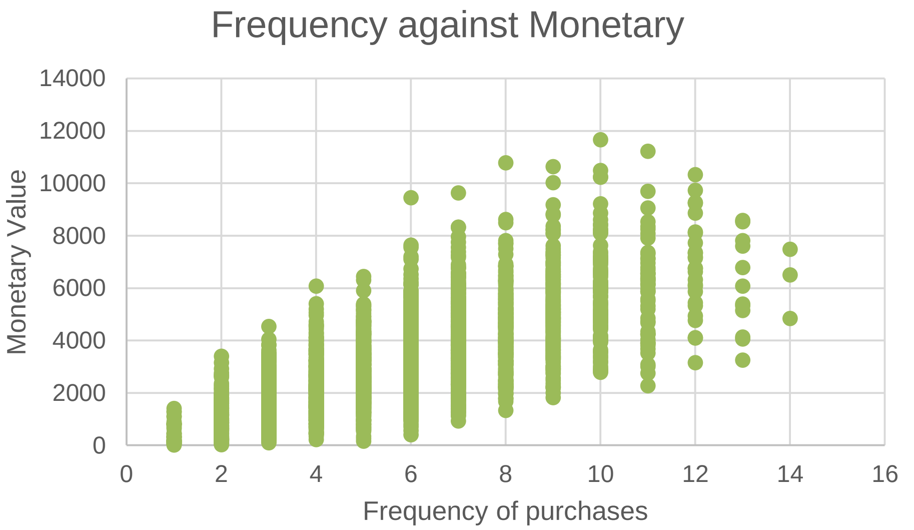
- This scatter plot shows the correlation between Recency and Monetary.
- It is representing customers who purchased more recently have generate more revenue, than customer who visited a while ago.
- Customers from recent past (0-100) showing to generate large amount of revenue.



Model Development

Scatter-Plot based on RFM Analysis

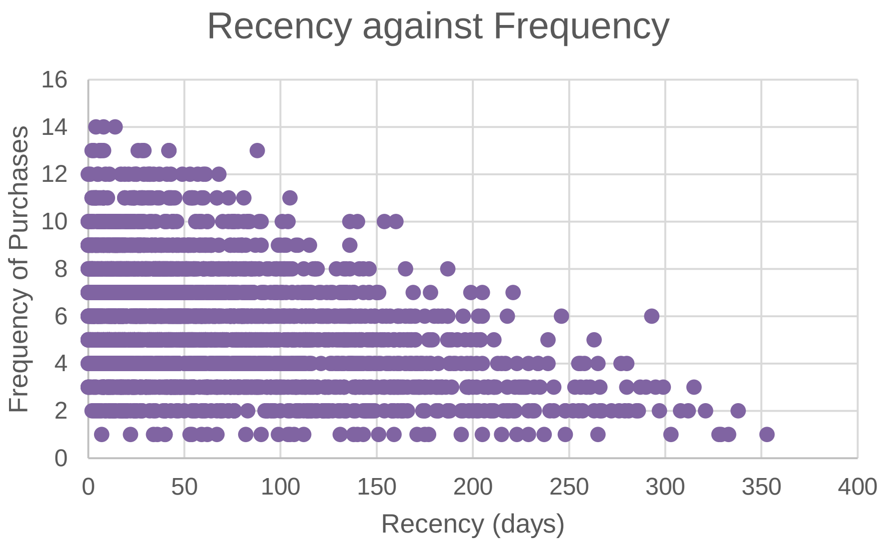
- This scatter plot shows the correlation between Frequency and Monetary.
- There is a positive relationship between frequency and monetary as with the increase of frequency, revenue is also increasing.
- High revenue customers are between 8 to 12 frequency.



Model Development

Scatter-Plot based on RFM Analysis

- ❑ This scatter plot shows the correlation between Recency and Frequency.
- ❑ Very low frequency of 0-2 correlated with high recency values.
- ❑ Customers that have visited more recently(0-50) have a higher chance of visiting more frequently.



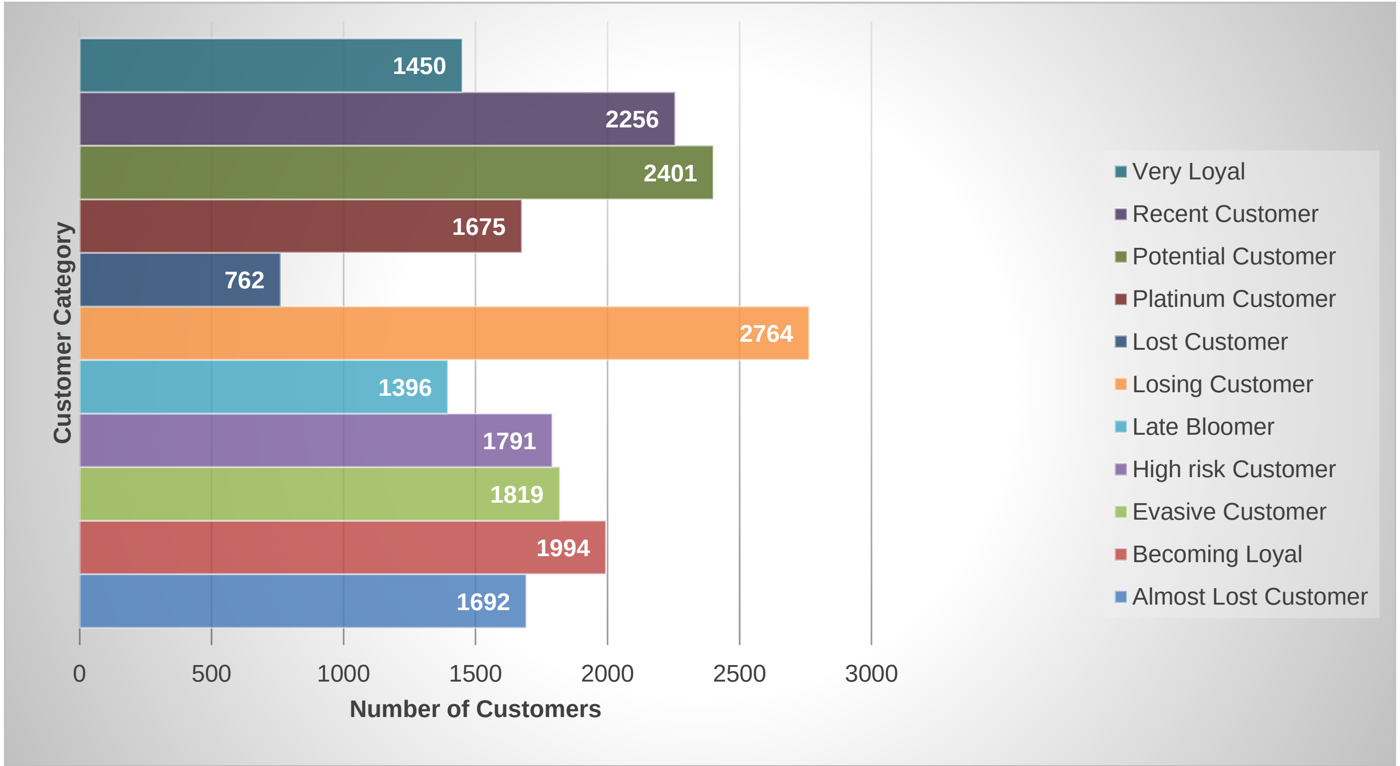
Model Development

Customer Categories Description with RFM values

Customer Categories	Description	RMF Value
Lost Customer	Very Low RFM	111
Evasive Customer	Very Low recency & frequency, small amount spent	112
Almost Lost Customer	Very Low recency & frequency, high amount spent	124
High Risk Customer	Purchase was long time ago, frequency & amount spent is high	212
Losing Customer	Purchase was a while ago, below average RFM value	224
Late Bloomer	No recent Purchase, RFM value is high	311
Potential Customer	Bought recently, never bought before, spent small amount	323
Recent Customer	Bought recently, not very often, average money spent	344
Becoming loyal	Recent customer, spends large amount of money	421
Very Loyal	Most recent, buys often, spends large amount of money	433
Platinum Customer	Most recent, buys often, most spent	444

Model Development

Customer Distribution in Combined dataset



Model Development

Data cleaning

```
df.columns
```

```
Index(['transaction_id', 'product_id', 'customer_id', 'transaction_date',
       'recency', 'online_order', 'order_status', 'brand', 'product_line',
       'product_class', 'product_size', 'list_price', 'standard_cost',
       'product_first_sold_date', 'profit', 'gender',
       'past_3_years_bike_related_purchases', 'DOB', 'Age', 'Age category',
       'job_title', 'job_industry_category', 'wealth_segment',
       'deceased_indicator', 'owns_car', 'tenure', 'address', 'postcode',
       'state', 'country', 'property_valuation', 'Customer Title',
       'RFM value'],
      dtype='object')
```

Removing unnecessary variables

```
#feature selection
data=data.drop(['address','country','DOB','job_title','online_order','order_status','brand',
               'product_class','product_size','product_line','product_first_sold_date','Customer Title','Age','profit','postcode'],a
```

Model Development

Missing values treatment

Missing values in training set with percentage:

We will remove missing values from Gender , Age and Bike related purchases.
In variable job industry category we have 3232 missing records
Which is approximately 16% of the data.

we decide to fill these missing records with a new class named 'Unknown'.

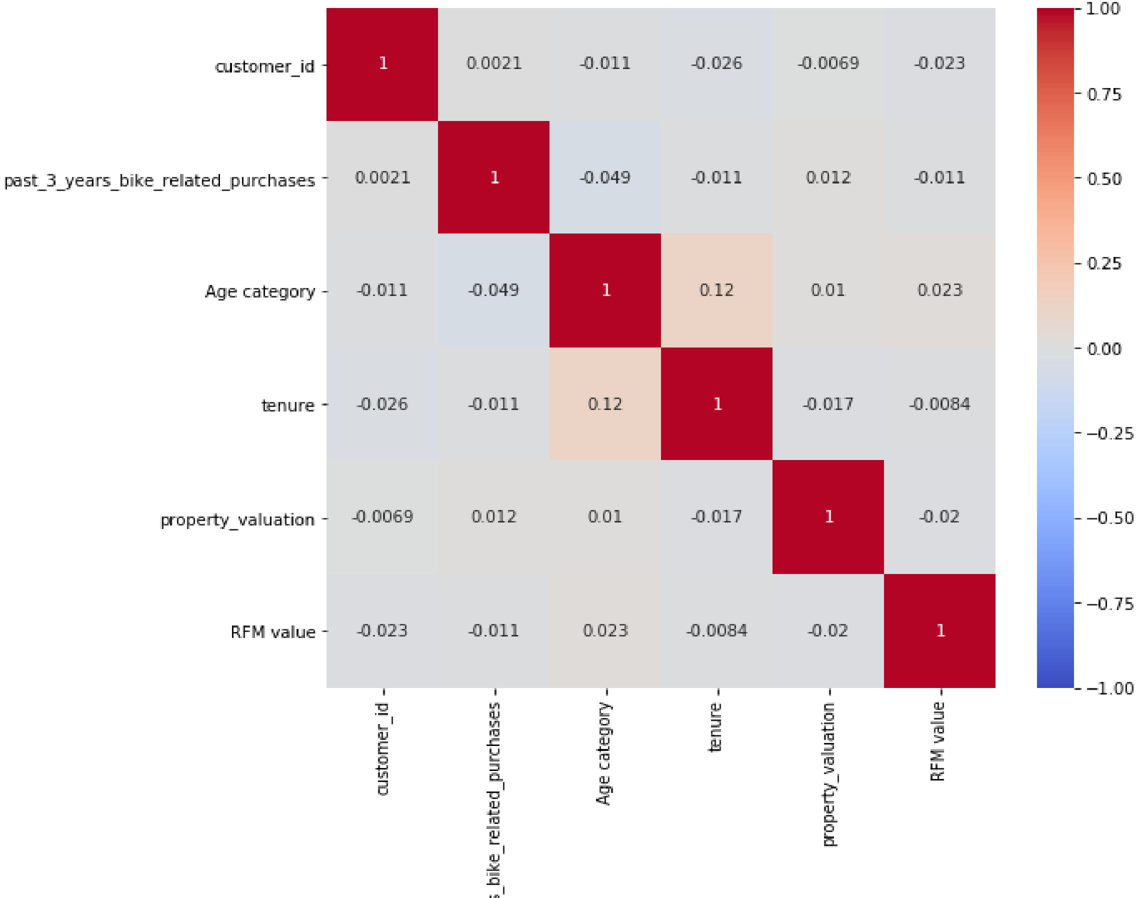
```
#imputing value
data.job_industry_category.fillna('unknown',inplace= True)
```

	0	1	2
0	customer_id	0.000000	0
1	gender	2.705997	555
2	past_3_years_bike_related_purchases	2.705997	555
3	Age category	0.039005	8
4	job_industry_category	15.758167	3232
5	wealth_segment	0.000000	0
6	deceased_indicator	0.000000	0
7	owns_car	0.000000	0
8	tenure	0.000000	0
9	state	0.000000	0
10	property_valuation	0.000000	0
11	RFM value	0.000000	0

Model Development

Data Correlation

Tenure is slightly correlated with Age. Other than that everything is normal.



Model Development

Dummy variable

For categorical variables we created dummy variables.
Categorical variables are:

- Gender
- Job industry category
- Wealth segment
- Deceased indicator
- Owns car
- State

After creating dummy variables and missing value treatment we are left with 20 variables.

```
df2.shape
```

```
(19437, 20)
```

Model Development

We created RFM value as a Target variable to make predictions.

```
X1= df2.drop(['RFM value'],axis =1)
y2= df2['RFM value']
```

```
from sklearn.preprocessing import StandardScaler
X1 = StandardScaler().fit_transform(X1)
```

```
from sklearn.model_selection import train_test_split
X_tr, X_te, y_tr, y_te= train_test_split(X1, y2, test_size=0.2,random_state=0)
print("shape of X_train,Y_train:",X_tr.shape,y_tr.shape)
print("shape of X_test,Y_test:",X_te.shape,y_te.shape)
```

```
shape of X_train,Y_train: (15549, 19) (15549,)
shape of X_test,Y_test: (3888, 19) (3888,)
```

Model Development

Training the Dataset with Decision Tree Algorithm

```
from sklearn.tree import DecisionTreeRegressor

# create a regressor object
regressor = DecisionTreeRegressor(max_depth=110,random_state = 0)

# fit the regressor with X and Y data
regressor.fit(X_tr, y_tr)
```

```
DecisionTreeRegressor(ccp_alpha=0.0, criterion='mse', max_depth=110,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort='deprecated',
                      random_state=0, splitter='best')
```

Model Evaluation

Model Evaluation with Residual Sum of Square & R2-Score

```
y_pred2 = regressor.predict(X_te)
from sklearn.metrics import r2_score
print("Residual sum of squares: %.2f"
      % np.mean((y_pred2 - y_te) ** 2))
print("R2-score: %.2f" % r2_score(y_pred2 , y_te) )
```

```
Residual sum of squares: 261.99
R2-score: 0.98
```

Testing set with Accuracy 98%

```
y_hat3 = regressor.predict(X_tr)
from sklearn.metrics import r2_score
print("Residual sum of squares: %.2f"
      % np.mean((y_hat3 - y_tr) ** 2))
print("R2-score: %.2f" % r2_score(y_hat3 , y_tr) )
```

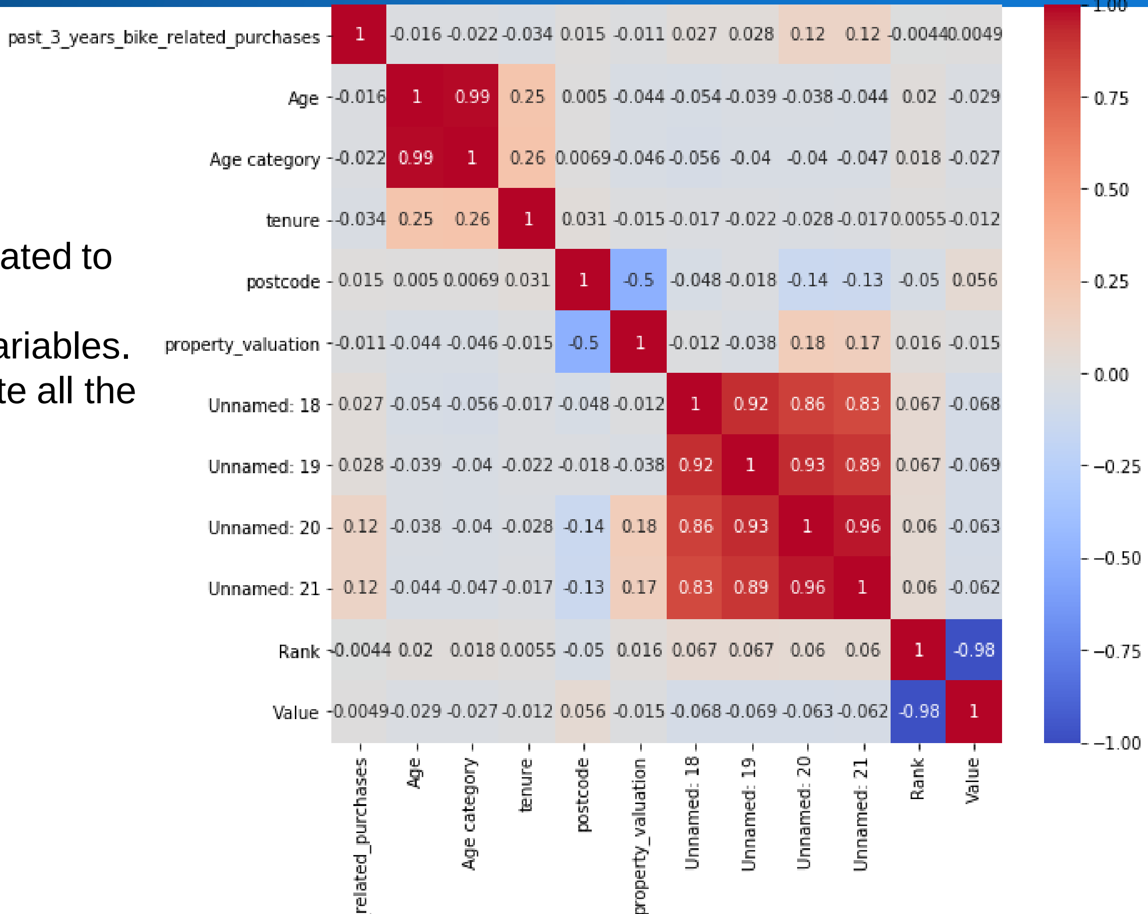
```
Residual sum of squares: 6.61
R2-score: 1.00
```

Training set with Accuracy 100%

Interpretation

New Customer list Data

There are so many variables which are correlated to each other.
Such as Rank with Value and all Unnamed variables.
Therefore, we decide to keep Value and delete all the other variables.



Interpretation

New Customer list Data

Data Cleaning

Removing unnecessary columns:

```
nc.columns
```

```
Index(['first_name', 'last_name', 'gender',
       'past_3_years_bike_related_purchases', 'DOB', 'Age', 'Age category',
       'job_title', 'job_industry_category', 'wealth_segment',
       'deceased_indicator', 'owns_car', 'tenure', 'address', 'postcode',
       'state', 'country', 'property_valuation', 'Unnamed: 18', 'Unnamed: 19',
       'Unnamed: 20', 'Unnamed: 21', 'Rank', 'Value'],
      dtype='object')
```

```
df=nc.drop(['first_name', 'last_name','DOB', 'Age','job_title','address', 'postcode','country', 'Unnamed: 18', 'Unnamed: 19',
       'Unnamed: 20', 'Unnamed: 21', 'Rank'],axis=1)
```

Interpretation

New Customer list Data

Missing values Treatment

We only have missing values in the column Job Industry Category which is 165 records.

we decide to fill these empty records as:

```
df.job_industry_category.fillna('unknown',inplace=True)
df.isnull().sum()
```

```
df.replace(r'^\s*$', np.nan, regex=True)
df.isnull().sum()
```

```
gender                                  0
past_3_years_bike_related_purchases     0
Age category                            0
job_industry_category                 165
wealth_segment                          0
deceased_indicator                      0
owns_car                                0
tenure                                  0
state                                   0
property_valuation                      0
Value                                   0
dtype: int64
```

Model Development

New Customer List Data

Dummy Variables

For categorical variables we created dummy variables.
Categorical variables are:

- Gender
- Job industry category
- Wealth segment
- Deceased indicator
- Owns car
- State

After creating dummy variables and missing value treatment we are left with 20 variables.

```
df.shape
```

```
(1000, 20)
```

Interpretation

New Customer list Data after prediction

Now, we can easily find our best customers with these predicted values.

Customer Categories	Description	RMF Value
Lost Customer	Very Low RFM	111
Evasive Customer	Very Low recency & frequency, small amount spent	112
Almost Lost Customer	Very Low recency & frequency, high amount spent	124
High Risk Customer	Purchase was long time ago, frequency & amount spent is high	212
Losing Customer	Purchase was a while ago, below average RFM value	224
Late Bloomer	No recent Purchase, RFM value is high	311
Potential Customer	Bought recently, never bought before, spent small amount	323
Recent Customer	Bought recently, not very often, average money spent	344
Becoming loyal	Recent customer, spends large amount of money	421
Very Loyal	Most recent, buys often, spends large amount of money	433
Platinum Customer	Most recent, buys often, most spent	444

Predicted_value	First_name_x
312.0	Chickie
143.0	Morly
112.0	Ardelis
111.0	Lucine
422.0	Melinda
424.0	Druci
421.0	Rutledge
422.0	Nancie
111.0	Duff
211.0	Barthel
313.0	Rockwell
332.0	Wheeler
224.0	Olag
422.0	Melba
242.0	Mandie