Sprocket Central Pty Ltd

Data analytics approach

The Analytics Team– Virtual Junior Consultant-KPMG]

Agenda

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

Introduction

Data Quality Analysis Overview

Dataset	General Overview	Issues Faced	Strategy to overcome
Transaction	13 columns, 20000 rows. Data types :int64, datetime64, float64 and object, no duplicate records.	**'product_first_sold_date'** column isn't in the appropriate date time format Missing values in columns = **'online order','product_line', 'product_class', 'product_size', 'standard_cost', 'product_first_sold_date'.**	The required column will be converted into appropriate datetime format. As the percentage of missing values are not alarmingly high, they will not be dropped but can be treated with mode/mean values.

Introduction

Data Quality Analysis Overview

Datasets	General Overview	Issues Faced	Strategy to overcome
Customer Demographic	There are 4000 rows (unique values of 'customer-id') with 13 columns. Data types are int64, datetime64, float64 and object. No duplicates.	Missing values in columns = **'last_name', 'DOB', 'job_title', 'job_industry_category', 'default' and 'tenure'.**Inconsistencies in the **'gender'** columnIrrelevant and incomprehensible data in **'default'** columnThere is an error in the year of the **'DOB'** column.More clarification needed in regards of how **'wealth-segment', 'owns_car' and 'tenure' columns** are related.	- For columns with missing values, we would employ replacement methods such as KNN imputer or through statistical techniques like mode or mean. Below are the columns which have more than 5 percent of missing values: **Job_title - 12.65** **Job_industry_category - 16.40** **Wealth_segment - 7.55** - To synchronize the values in the **'gender'** column, we will change 'female' to 'F', 'femal' to 'F', 'Male' to 'M' and 'U' to 'NA'.**'default'** column will be dropped as it is irrelevant to our analysis.Ages of the customers can be calculated from the **'DOB'** column. As there is an error of mistype error, that would be rectified too. From 1844 to 1944.

Introduction

Data Quality Analysis Overview

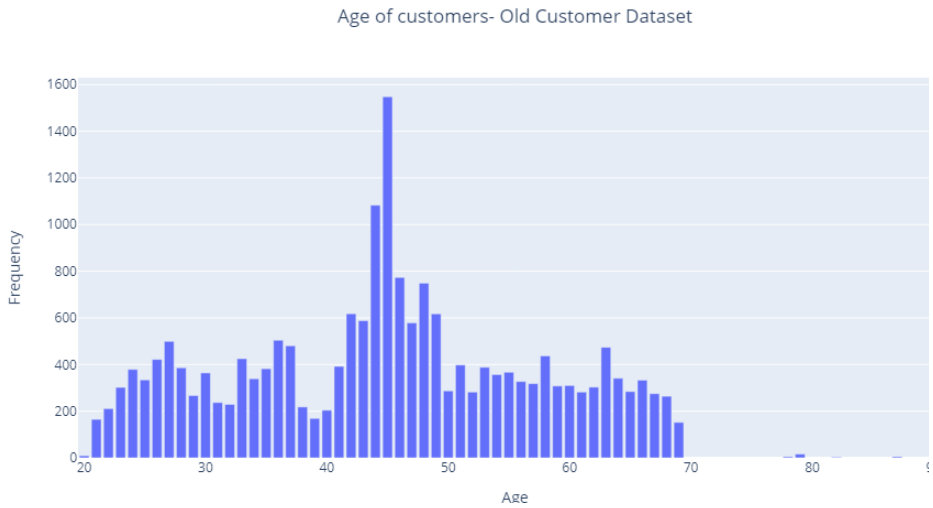
Datasets	General Overview	Issues Faced	Strategy to overcome
Customer Address	There are 3999 rows with 6 columns. Data types are int64 and object. No duplicates.	The ‘state’ column has names of states in full form and abbreviations mixed up.	The following will be changed: NSW to New South Wales, VIC to Victoria, QLD to Queensland

Introduction

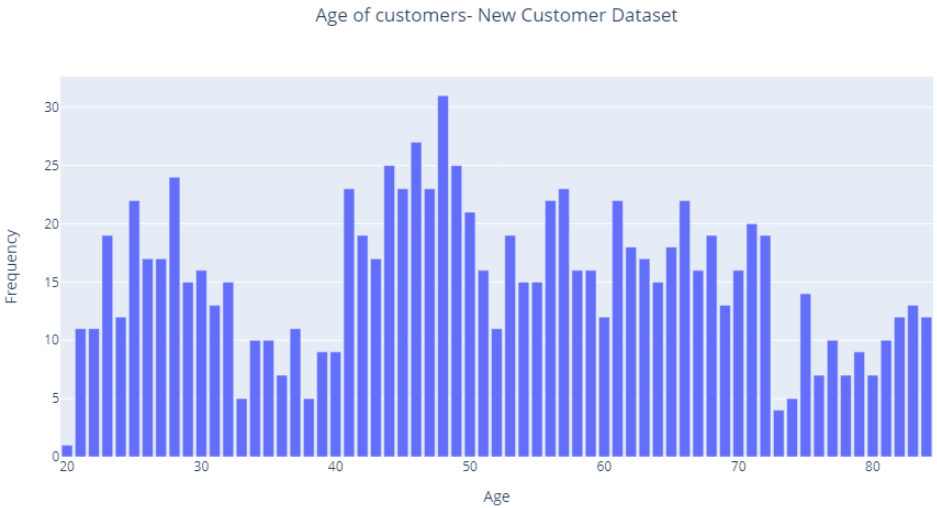
Data Quality Analysis Overview

Datasets	General Overview	Issues Faced	Strategy to overcome
Customer Address	There are 1000 rows with 22 columns. Data types are int64, datetime64, float64, int64 and object. No duplicates.	Missing values in columns = **‘last_name’, ‘DOB’, ‘job_title’, and ‘job_industry_category’** The values in **‘gender’** column to be replaced to match with the old customer details. There are some unnamed columns with data which are vague to other columns. There is no unique customer id assigned to any of the customers.	For columns with missing values, we would employ replacement methods such as KNN imputer or through statistical techniques like mode or mean. To synchronize the values in the **‘gender’** column, we will change 'female' to 'F', 'Male' to 'M' and ‘U’ to ‘NA’. The unnamed columns will be dropped. We can assign a new column ‘customer_id’ where the unique ids will be a continuation of the old customers’.

Age Distribution

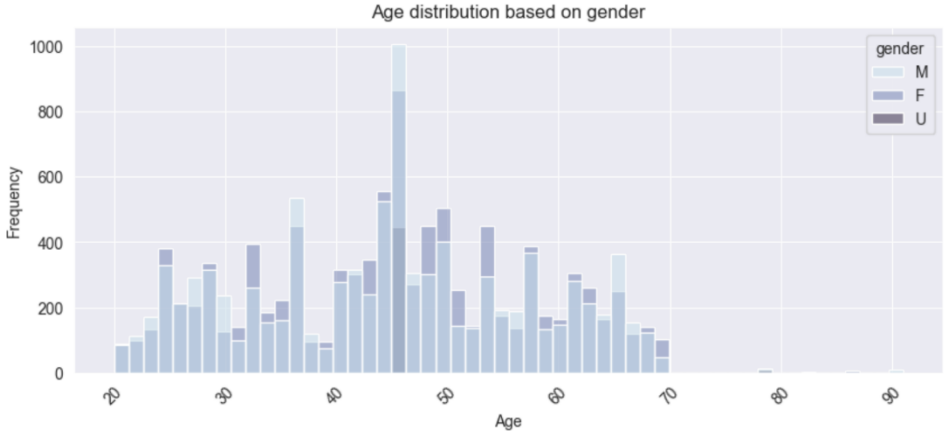


Min age : 20
Max age : 91
Avg age : 44.9



Min age : 20
Max age : 84
Avg age : 51.3

Age Distribution

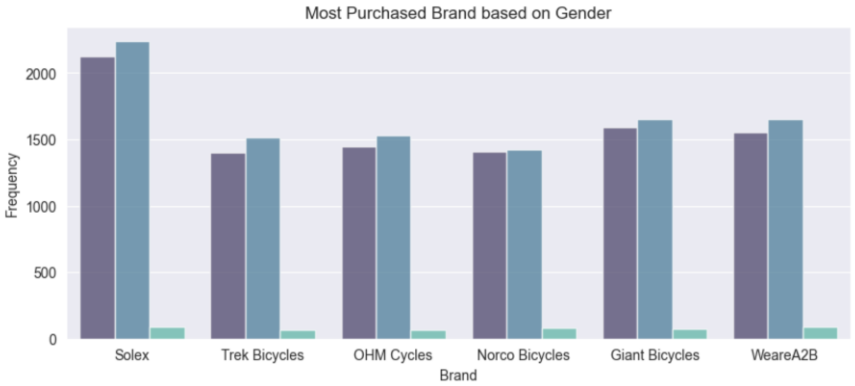


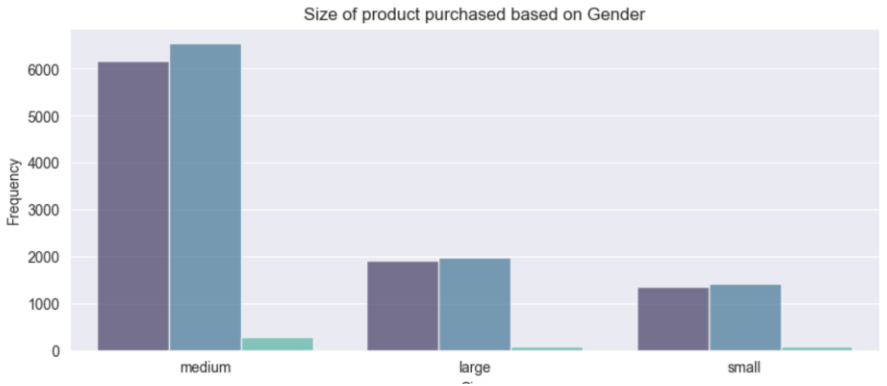
U : Unspecified ,all the values of U were 'nan', we replaced with mode values

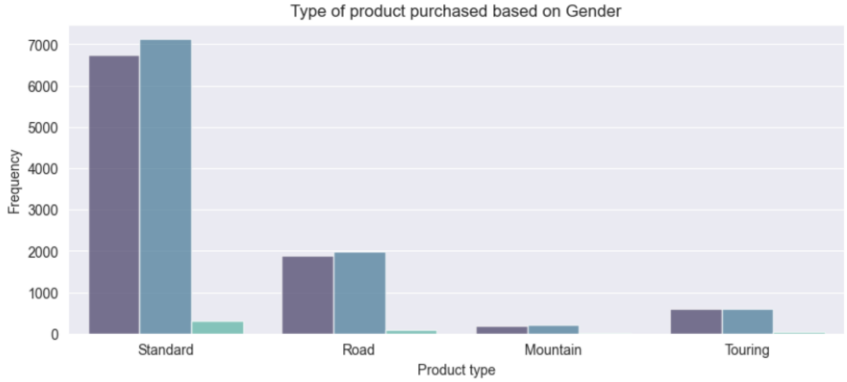
gender	count	mean	std	min	25%	50%	75%	max
F	10011.0	45.059235	12.426117	20.0	35.0	45.0	54.0	87.0
M	9531.0	44.791837	12.805794	20.0	35.0	45.0	55.0	91.0
U	455.0	45.672527	4.739516	45.0	45.0	45.0	45.0	79.0

gender	count	mean	std	min	25%	50%	75%	max
F	513.0	51.814815	16.866514	20.0	41.0	51.0	65.0	84.0
M	470.0	50.687234	17.291988	21.0	37.0	50.0	65.0	84.0
U	17.0	48.000000	0.000000	48.0	48.0	48.0	48.0	48.0

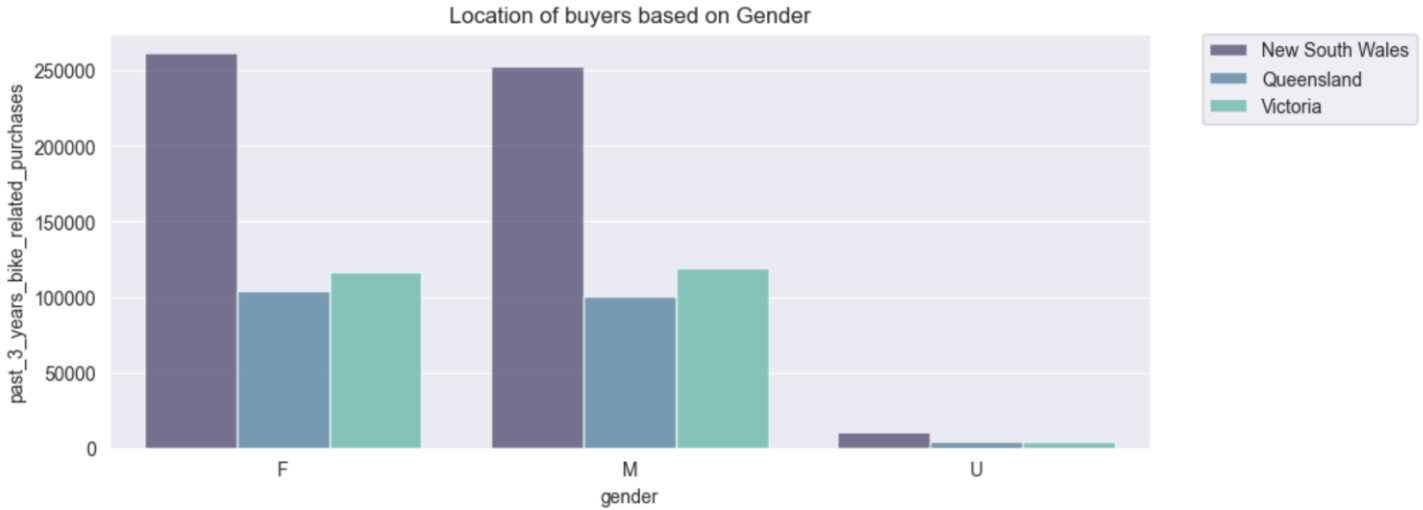
Most purchased brand, type/size/class of product based of gender



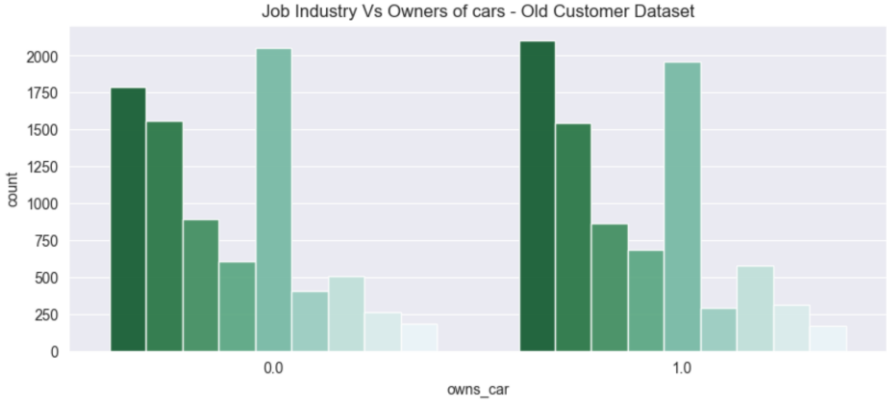




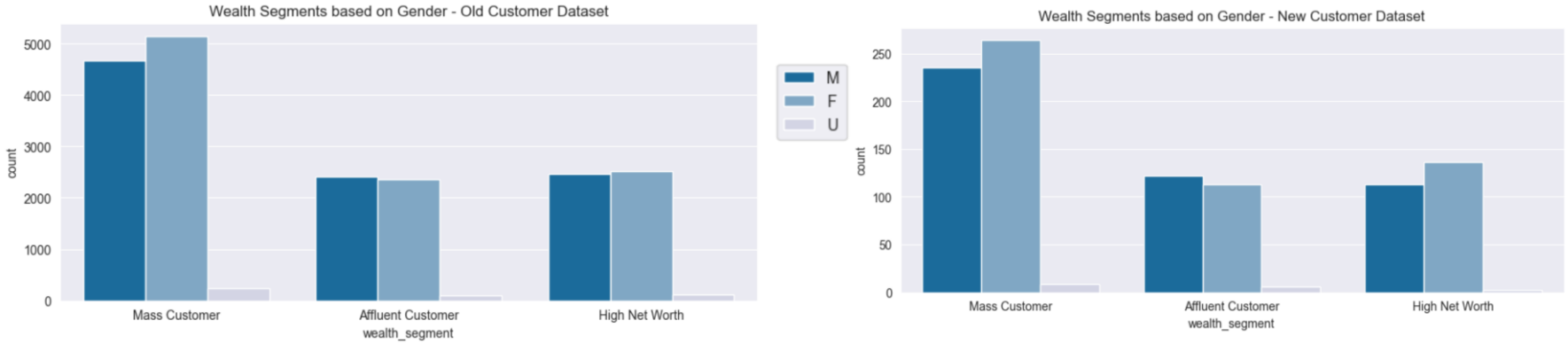
Location of buyer based on Gender



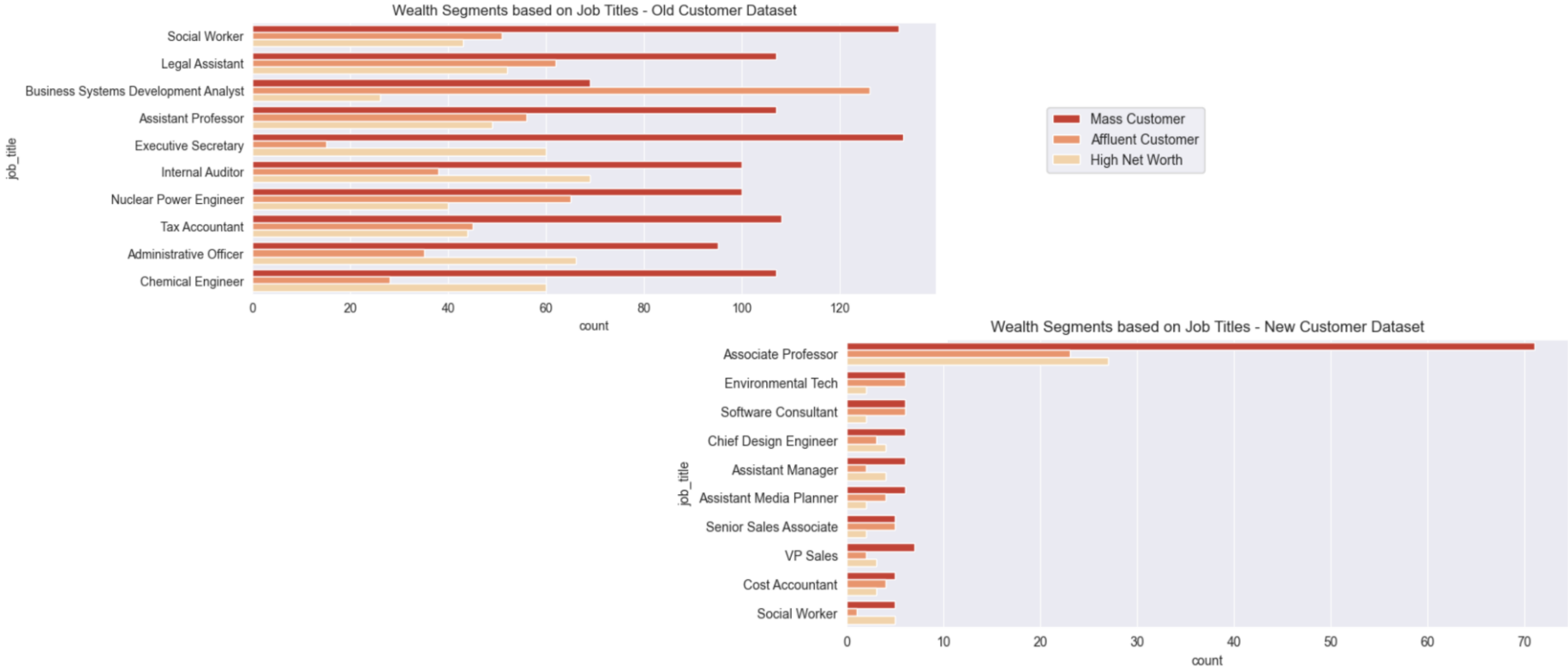
Job Industry Vs Car Owners



Wealth Segment based on Gender



Wealth Segment based on Gender



Observation

- There are more female customers in both the given old and new customers' dataset.
- The age range for old customer dataset is from 20 to 91 and 20 to 84 in the new customer dataset.
- The gender was plotted against the location. Most of the female customers are from New South Wales , followed by Victoria and the least is from Queensland. The reason could be that NSW has higher number of residents compared to Queensland. The similar pattern is also observed among the male customers.
- As for the case of Job Industry Vs Car Owners, highest number of vehicle ownership is recorded from customers working in the Financial Services (old customer dataset) while for the new customer dataset, it is from the Retail industry.
- Since most of the customers are females, there are large amount of females in all the wealth segments (mass/affluent/high net worth).
- The top most job title is Social Worker (old customer dataset) with distribution of mass customer being the highest followed by affluent and as for the new customer dataset it is Associate Professor with Mass Customers being the highest but followed by high net worth customers. There isn't any direct correlations from this aspect. Most probably the job industry which the job titles belong too is doing well compared to the others.

RFM Analysis

RFM uses sales data to segment a pool of customers based on their purchasing behavior. The resulting customer segments are neatly ordered from most valuable to least valuable. This makes it straightforward to identify best customers.

Recency stands for how long it takes for the customer to come back to you to purchase. Naturally, the longer it takes, the more chance it is for the customer to drop out or lose interest.

Frequency is how many transactions are there between the customer and you. Higher frequency means your products are of interest to them and they are valuable customers to the company.

Monetary helps us to understand how much a customer spends on average or in total which is the final measure of his or her value.

RFM Analysis

Steps taken to calculate Recency, Frequency and Monetary values.

- The last purchase data was identified and then that was used to calculate other values.

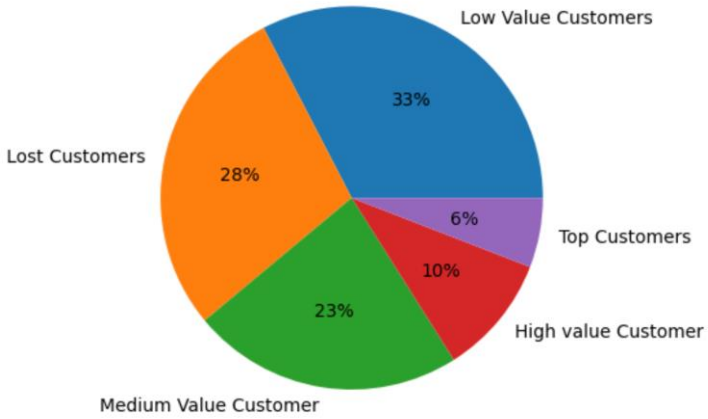
- The values were then normalized so that they would be at same scale.

- Finally the RFM score was calculated based on the formula below:
  0.15*Recency score + 0.28*Frequency score + 0.57 *Monetary score

- ***Score mapping was done as below:***
  rfm score =>5 : ***Top Customer***, rfm score => 4 : ***High Value Customer***, rfm score => 3 : ***Medium Value Customer***, rfm score => 2 : ***Low Value Customer***, rfm score => 1:***Lost Customer***

	CustomerID	Recency	Frequency	Monetary	R_rank_norm	F_rank_norm	M_rank_norm	RFM_Score	Customer_segment
0	1	7	11	3018.09	89.99	97.84	97.81	4.83	Top Customers
1	2	128	3	2226.26	12.92	12.37	12.36	0.62	Lost Customers
2	3	102	8	3362.81	19.01	83.40	83.37	3.69	Medium Value Customer
3	4	195	2	220.57	3.82	4.31	4.31	0.21	Lost Customers
4	5	16	6	2394.94	77.30	57.17	57.16	3.01	Medium Value Customer

RFM table



Distribution of customer segmentation

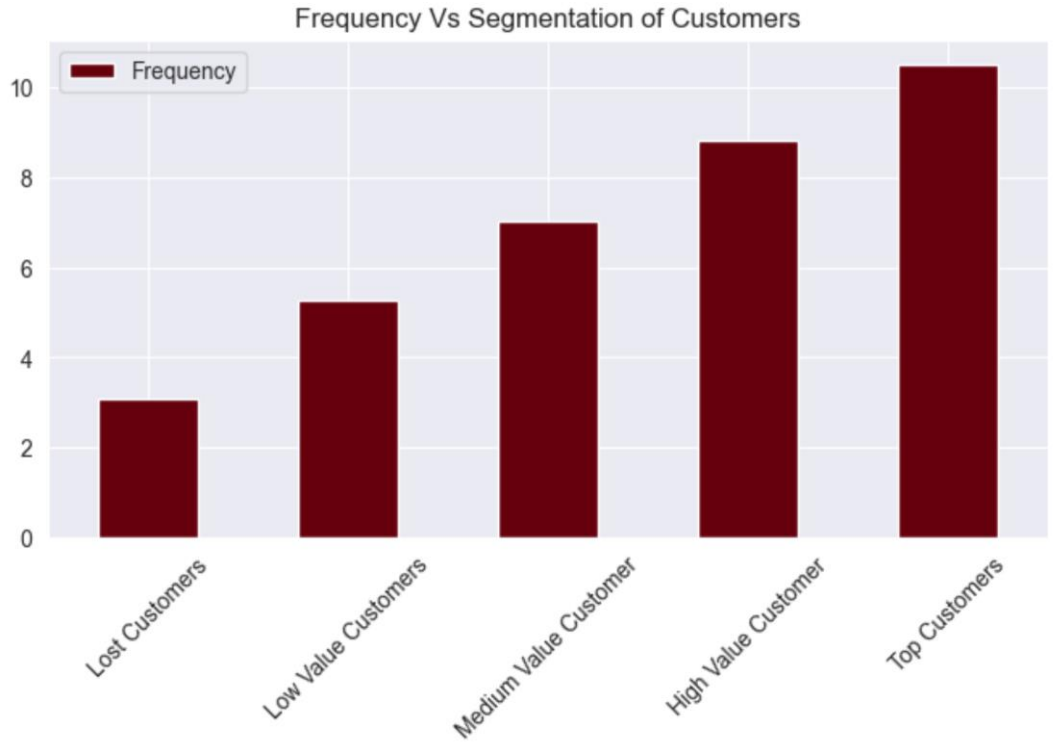
RFM Analysis- Model Development

	Monetary
count	5.000000
mean	3843.201055
std	1619.869877
min	1675.432216
25%	2966.724447
50%	3866.492247
75%	4871.570366
max	5835.786000

RFM Analysis

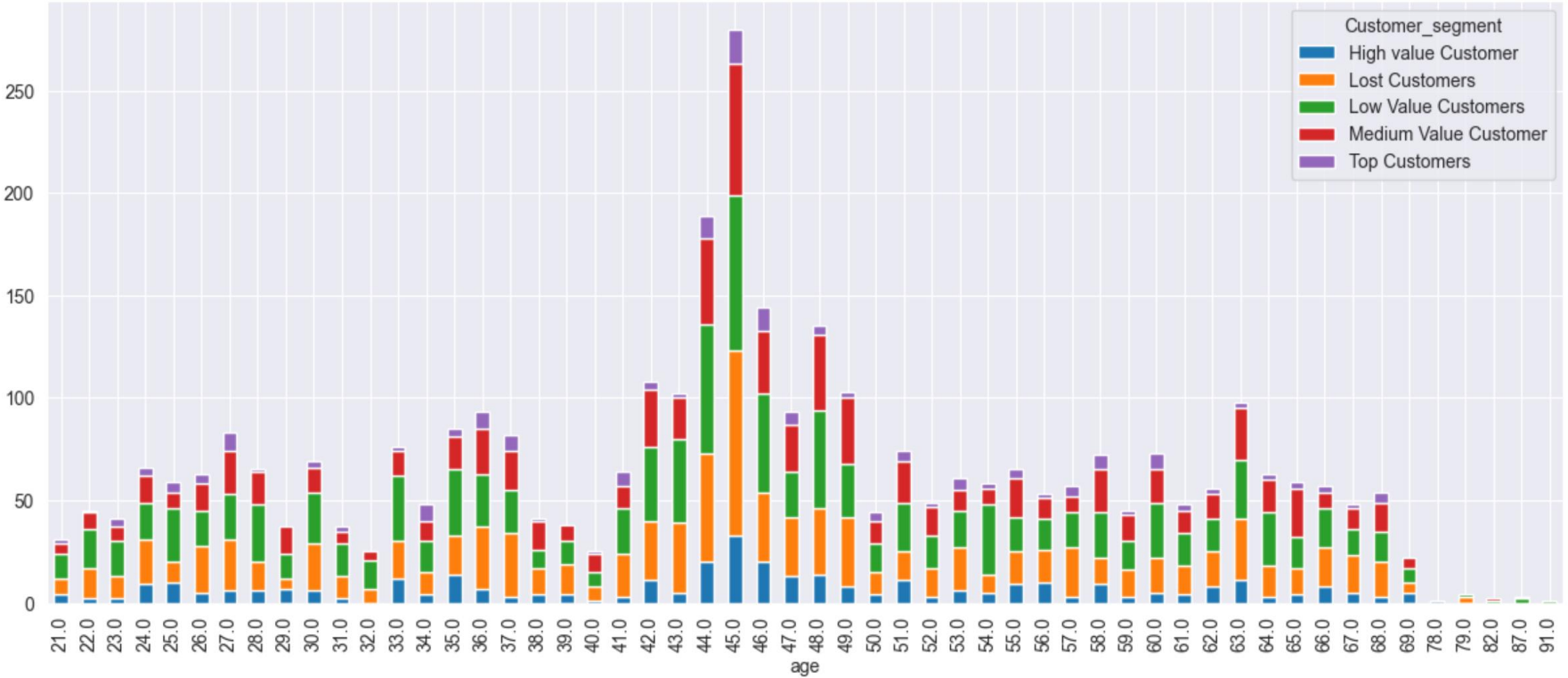
	Recency
count	5.000000
mean	49.311865
std	31.139008
min	14.682927
25%	32.861972
50%	43.584270
75%	58.189474
max	97.240685

RFM Analysis

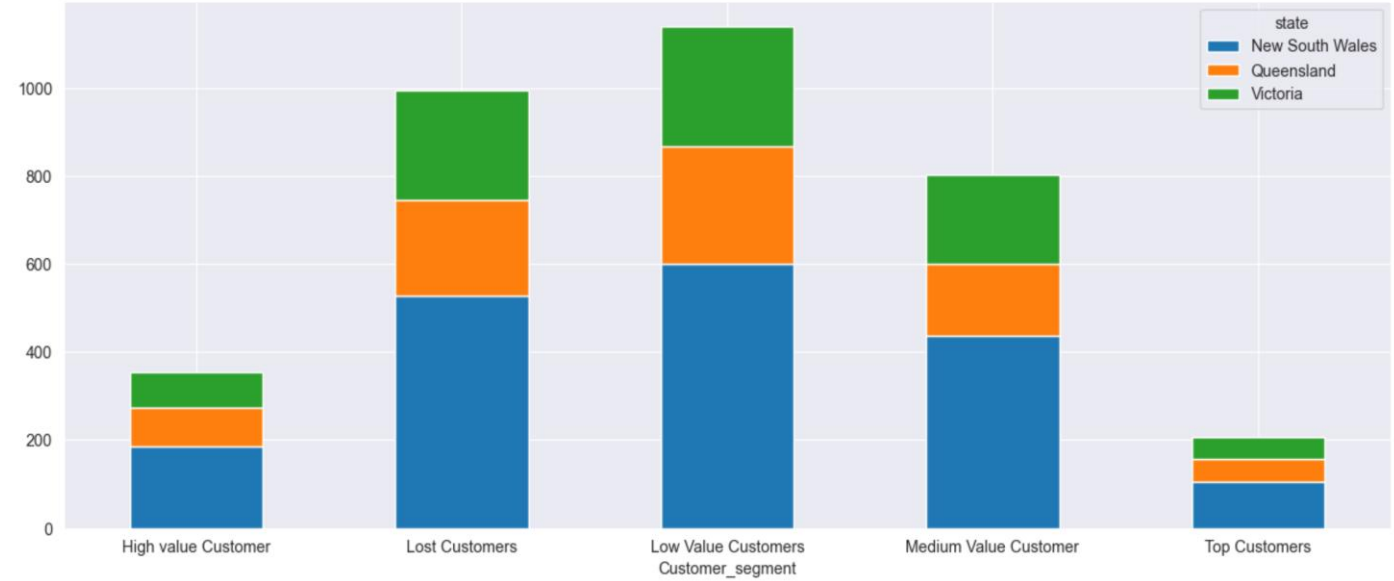


	Frequency
count	5.000000
mean	6.951000
std	2.923057
min	3.081571
25%	5.265789
50%	7.044944
75%	8.830986
max	10.531707

RFM Analysis



RFM Analysis



Customer_segment	state	
High value Customer	New South Wales	186
	Queensland	87
	Victoria	82
Lost Customers	New South Wales	528
	Victoria	248
	Queensland	217
Low Value Customers	New South Wales	601
	Victoria	273
	Queensland	266
Medium Value Customer	New South Wales	438
	Victoria	200
	Queensland	163
Top Customers	New South Wales	104
	Queensland	52
	Victoria	49

RFM Insights

- The most important aspects of RFM is Recency because you want the customers to keep coming back and purchase products from your company.
- We can see from the graph, the top customers' recency values are shorter in comparison to the lost customers.
- The top customers' monetary range is $5835.79, while the customers who have lost interest or dropped off is capped at $ 1675.43. This is also related to the number of purchases made by the different segments of customers (as is visible in the plotted graph)

Strategies

- The customers can be easily segmented based on RFM score. RFM analysis will help us to identify the kind of approach that should be taken for each segment of customers.

- For the lost customers, more discounts can be allocated while for the loyal ones, point system could be implemented. The more they buy, the more points can be accumulated which at later stage can be used to redeemed in other purchases.

- The brand/product which is being sold the most can be continued to be promoted to the group of customers who bought them, perhaps, more variety of accessories or upgrades relevant to the brand.

- Most bikes are used for standard usage and is identified bought the most by female customers. Awareness can created pertaining to health and benefits of cycling can be further spread. Besides endorsing or giving importance to the health factor, mini competitions can be organized by the company to further promote the brands/company.

The End