TheAnalyticsTeam

Sprocket Central Pty Ltd

Data analytics approach

Manuel Martínez
Data Analytics Consulting Virtual Intern

Agenda

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

Introduction

Sprocket Central Pty Ltd is a long-standing ***KPMG*** client whom specializes in high-quality bikes and accessible cycling accessories to riders. Their marketing team is looking to boost business by analyzing their existing customer dataset to determine customer trends and behavior.

Using the existing 3 datasets:

- Customer demographic
- Customer address
- Transactions

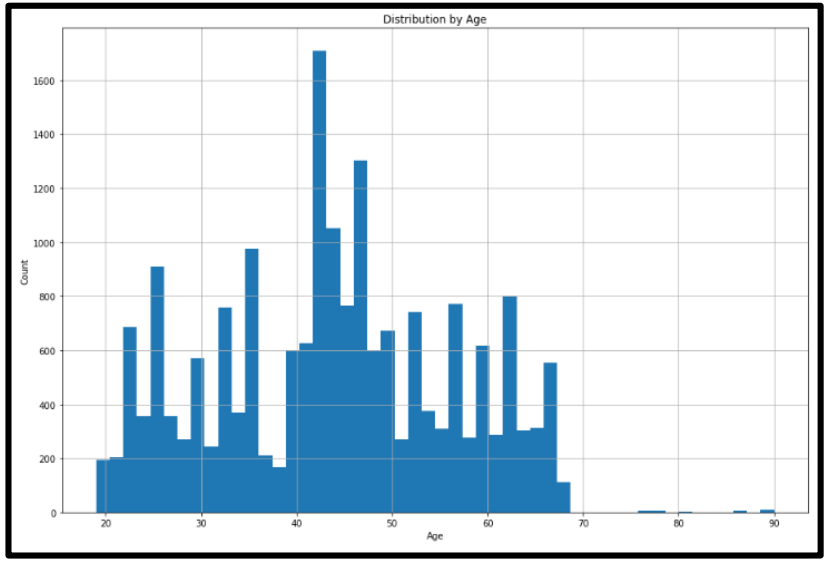
The objective is to recommend which of these 1000 new customers should be targeted to drive the most value for the organization.
The next steps are the approach that we will take:

Table name	No. of records	Unique Customer Ids
Customer demographic	4000	4000
Customer address	3999	3999
Transactions	20000	3494

Data Exploration

In this step, the dataset is explored in an unstructured way to uncover initial patterns, characteristics, and points of interest.

- Understand the features of given fields in the underlying data such as variable distribution, whether the dataset is skewed towards a certain demographic and the data validity of the fields. For example, a training dataset may be highly skewed towards the younger age bracket. If so, how will this impact your results when using it to predict over the remaining customer base.



Data Exploration

- There are some limitations in the given datasets like some values are missing and some data types are different according to their value.

	Customer demographic	Customer address	Transactions
Accuracy	✗ DOB: Inaccuracy	✗ Customer_id: Not in sync	✗ Customer_id: Not in sync
Completeness	✗ last_name, DOB, job_title, tenure, job_industry_category, default	✓ There are not missing values	✗ online_order , Brand, product_line, product_class, product_size, standard_cost, product_first_sold_date
Consistency	✗ Gender: Inconsistency	✗ State: Inconsistency	✗ product_first_sold_date: Fomat
Currency	✓ They are update	✓ They are update	✓ They are update
Relevancy	✗ Default : Exclude Feature	✓ They are relevant	✗ Order_status: Exclude Cancelled
Validity	✓ They are validated	✓ They are validated	✗ product_id: A lot zero values
Uniqueness	✓ There are not duplicated rows	✓ There are not duplicated rows	✓ There are not duplicated rows

Introduction

As part of the Data Exploration. We must clean the data

- Join the three datasets by Customer IDs.
- The features above must be impute with the mode:
  - job_title
  - job_industry_category
  - Tenure
  - Postcode
  - state
  - country
  - property_valuation
  - online_order
  - order_status
  - Brand
  - product_line
  - product_class
  - product_size
  - list_Price
  - product_size
  - standard_cost

- The rest of records with missing values must be were dropped.
- Values inaccuracy must be dropped.
- Features and records not relevant must be dropped.
- Features inconsistency must be fixed.

Table name	Combined table
No. of records	19354
Unique Customer Ids	3494

Data Exploration

- Use internal and external data to create new features that could be useful for modeling purposes. This may include bringing in ABS data at different geographic levels and creating additional features for the model.

  For example:
  - The geographic remoteness of different postcodes may be used as an indicator of proximity to consider to whether a customer is in need of a bike to ride to work.
  - The age of each customer to see what age is our main market.
  - When was the last purchase of each customer to analyze what clients are active.
  - What is the profit for each customer, that way we can analyze what premium clients we have,

Data Exploration

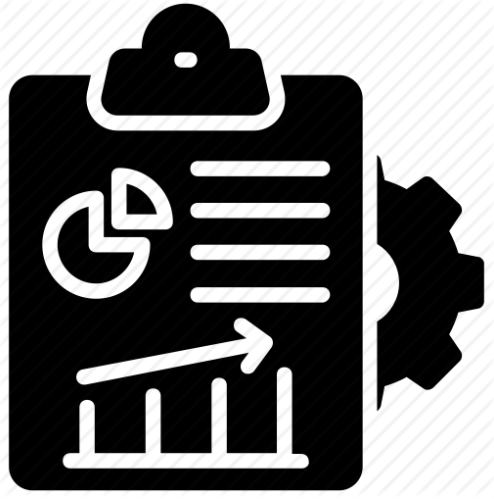
- Exploration of interactions between different variables through correlation analysis and look out for multicollinearity by creating interaction variables.

- Document assumptions, limitations and exclusions for the data. As well as how you would further improve in the next stage if there was additional time to address assumptions and remove limitations.

- Transform some features of the data in an appropriate format for analysis.

Model Development

Model development is an iterative process, in which many models are derived, tested and built upon until a model fitting the desired criteria is built.

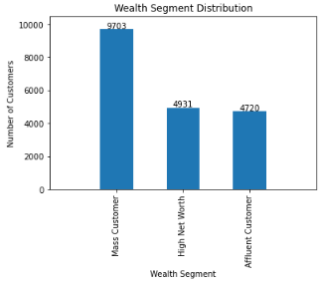


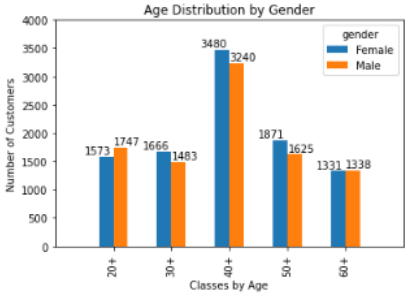
- Determine a hypothesis related to the business question that can be answered with the data. Perform statistical testing to determine if the hypothesis is valid or not.
- Create calculated fields based on existing data:
  - Age
  - Last purchase
  - Profit
- Test the performance of the model using criteria for the given model chosen. As residual deviance, AIC, ROC curves and R Squared.
- Appropriately document model performance, assumptions and limitations.

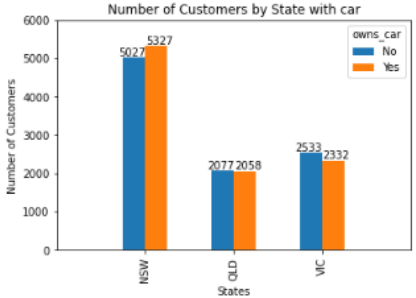
Interpretation

The final phase in any experiment is to interpret and report the results. Finding the answer to a challenging question is the goal of any study

- Visualization and presentation of findings. This may involve interpreting the significant variables and co-efficient from a business perspective.
- Demonstrate the accuracy of the model develop.
- Insights are presented



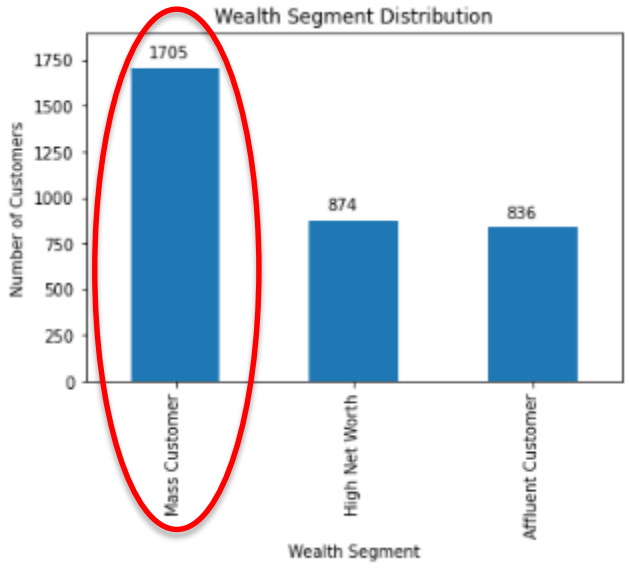




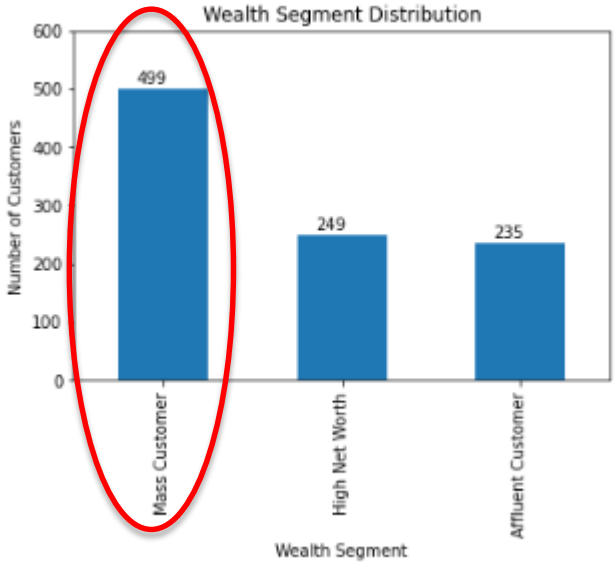
Analysis

Data Exploration

Wealth Segment by customers



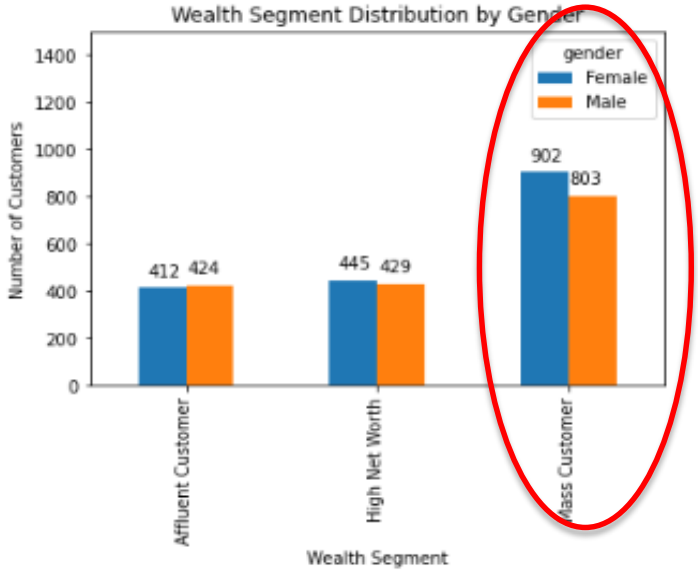
Old Customers



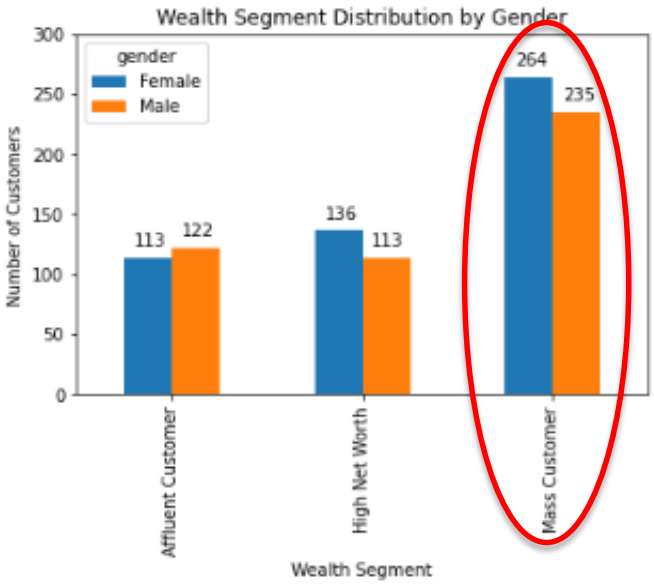
New Customers

Data Exploration

Wealth Segment Distribution by Gender



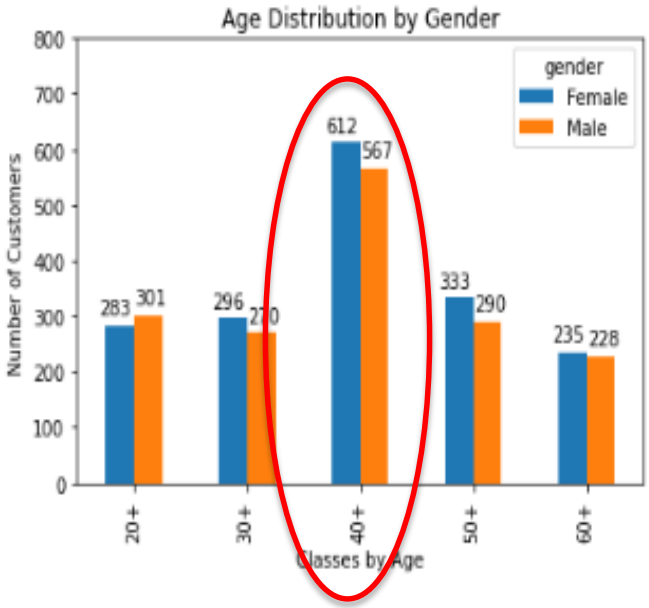
Old Customers



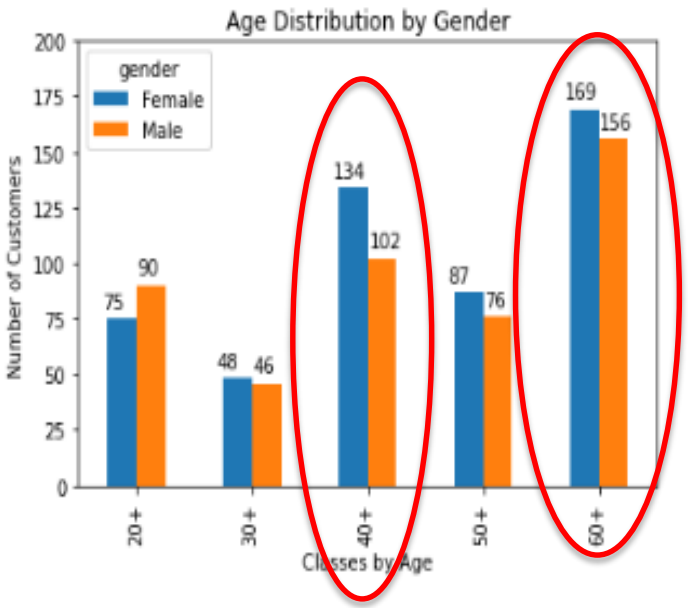
New Customers

Data Exploration

Age Distribution by Gender



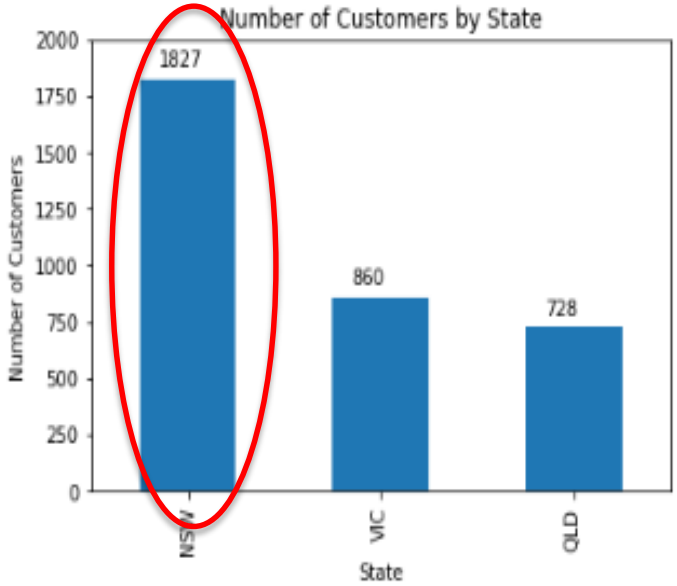
Old Customers



New Customers

Data Exploration

Number of Customers by State

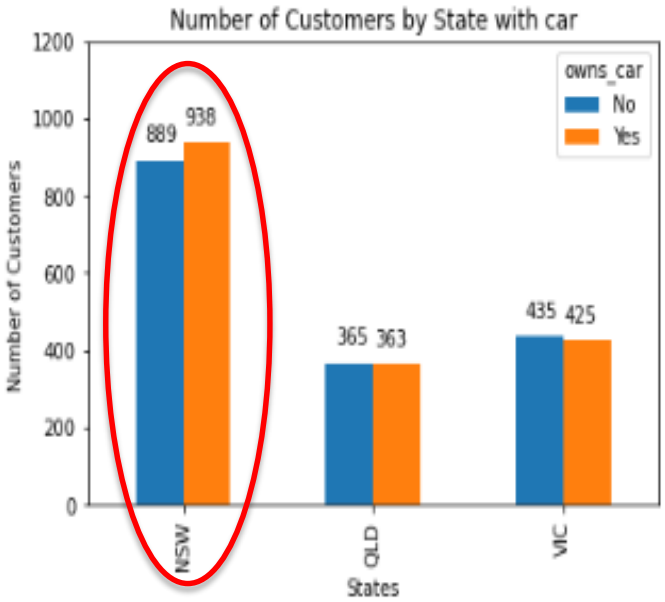


Old Customers

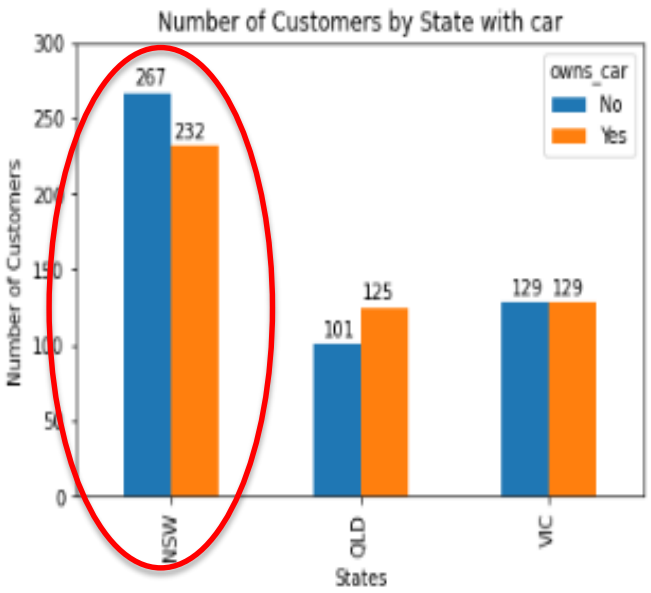
New Customers

Data Exploration

Number of Customers by State with car



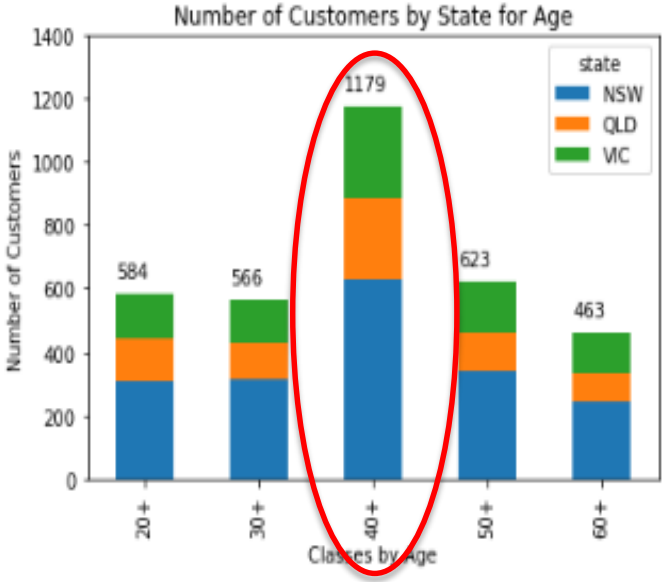
Old Customers



New Customers

Data Exploration

Number of Customers by State for Age

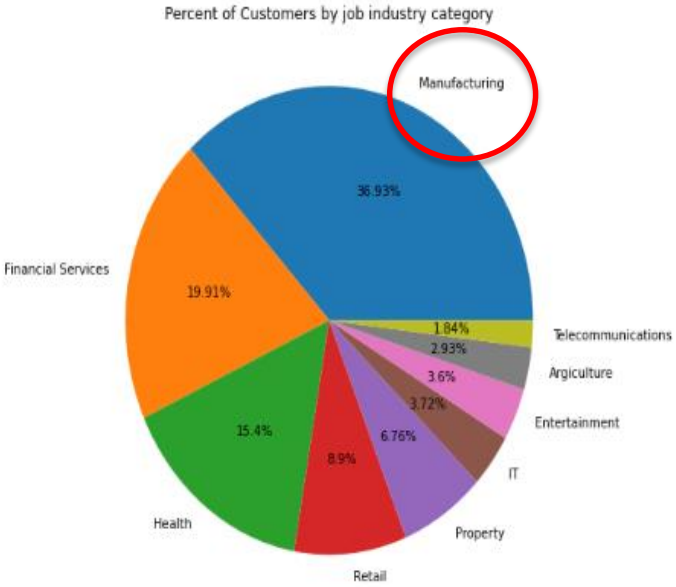


Old Customers

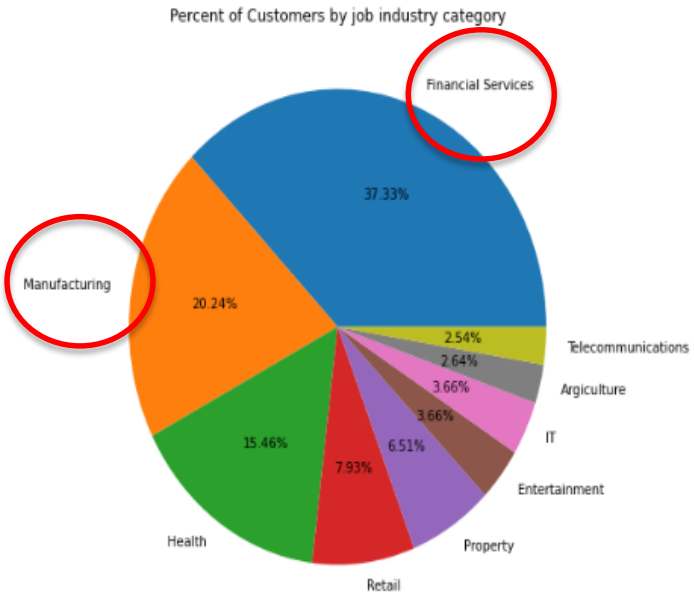
New Customers

Data Exploration

Percent of Customers by job industry category



Old Customers

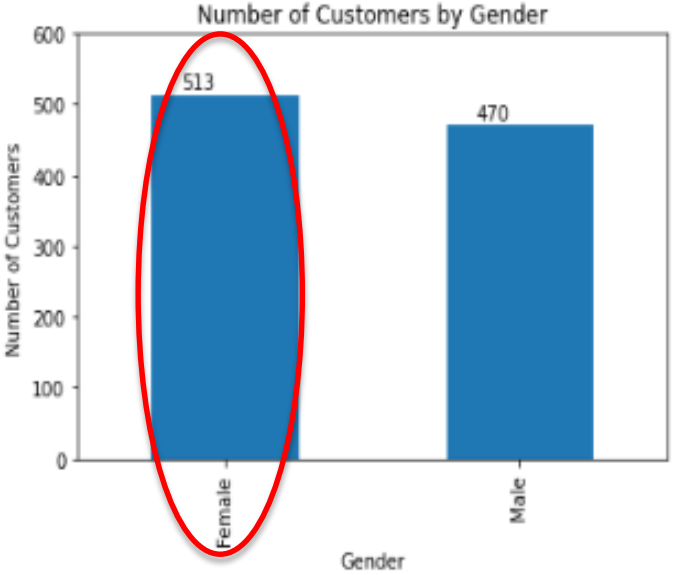


New Customers

Data Exploration

Number of Customers by Gender

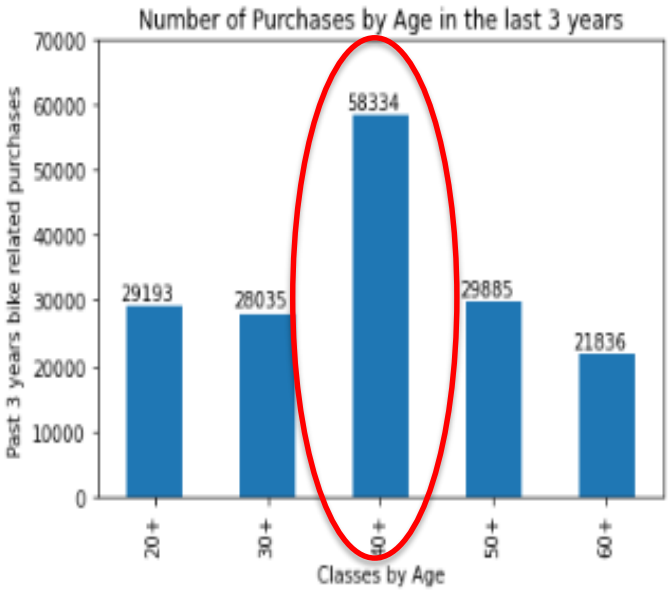
Old Customers



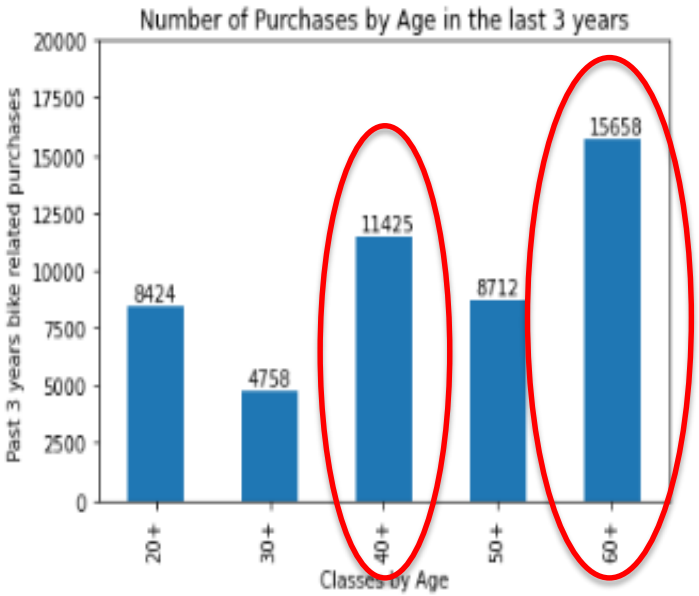
New Customers

Data Exploration

Number of Purchases by Gender in the last 3 years



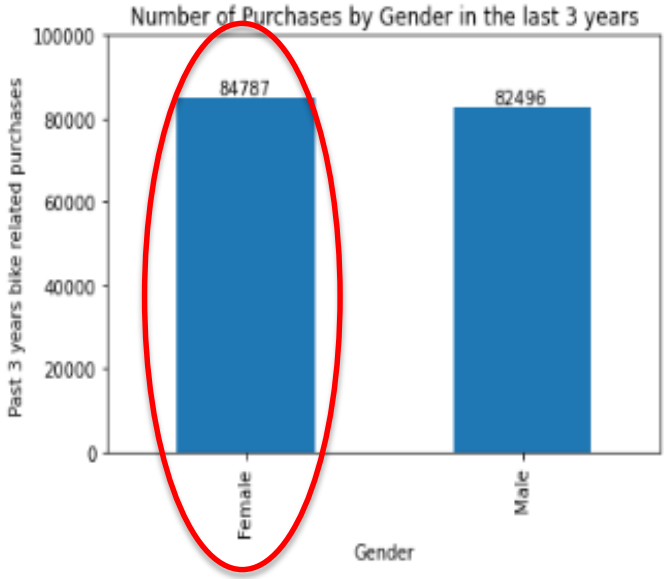
Old Customers



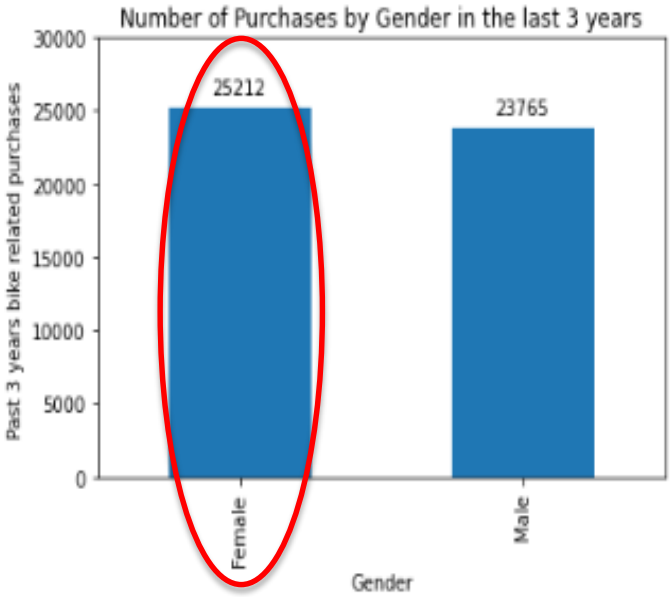
New Customers

Data Exploration

Number of Purchases by Gender in the last 3 years



Old Customers

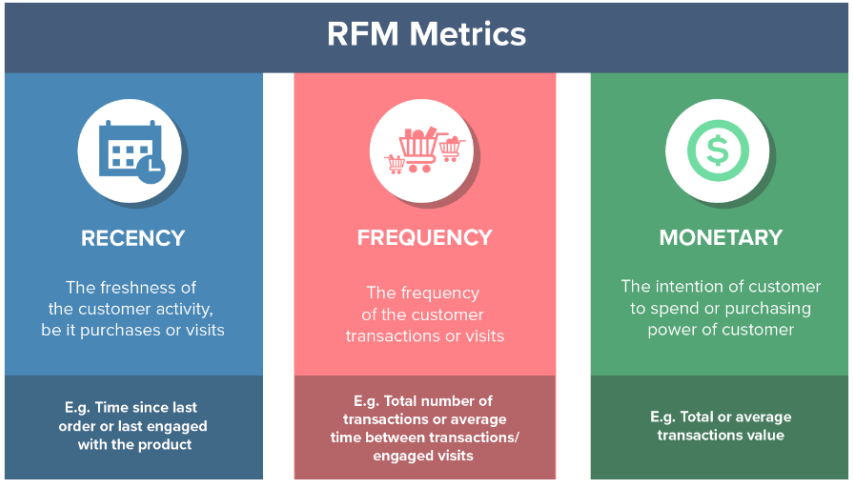


New Customers

Model Development

RFM Analysis

RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement.



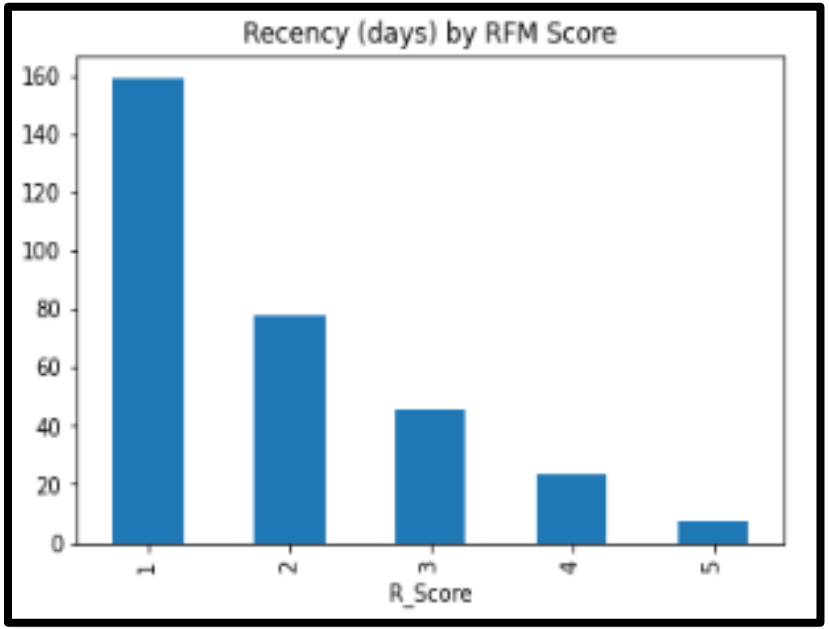
Model Development

RFM Analysis

- **Recency**
  - ❖ Number of days that have passed since the customer last purchased - How recently did the customer purchase?
  - ❖ Customers were divided into 5 quartiles and given a R_Score
- **Frequency**
  - ❖ Number of purchases in a specific period (for example, last 12 months) - How often do they purchase.
  - ❖ Customers were divided into 5 quartiles and given a F_Score
- **Monetary**
  - ❖ Sum of all purchases in the specific period – How much do they spend.
  - ❖ Customers were divided into 5 quartiles and given a M_Score

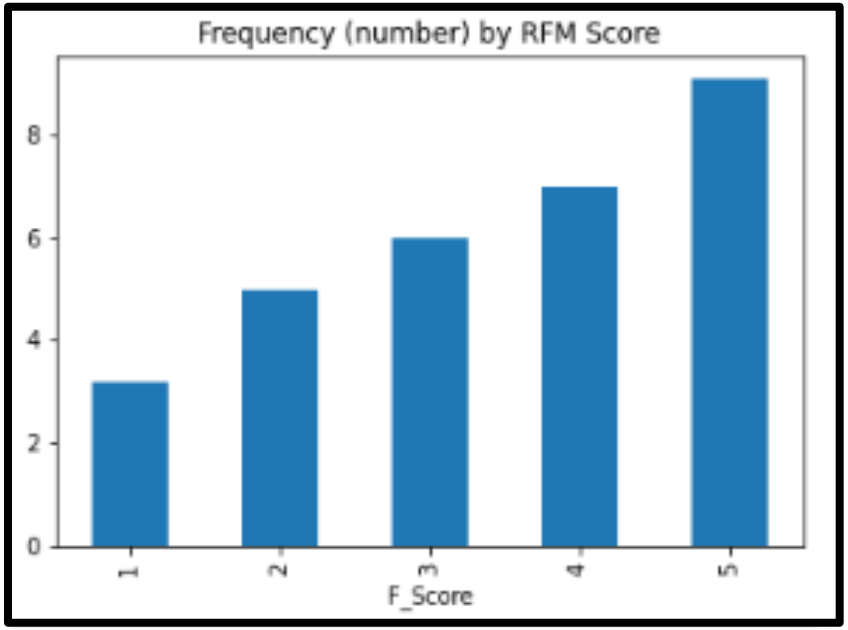
Model Development

RFM Analysis



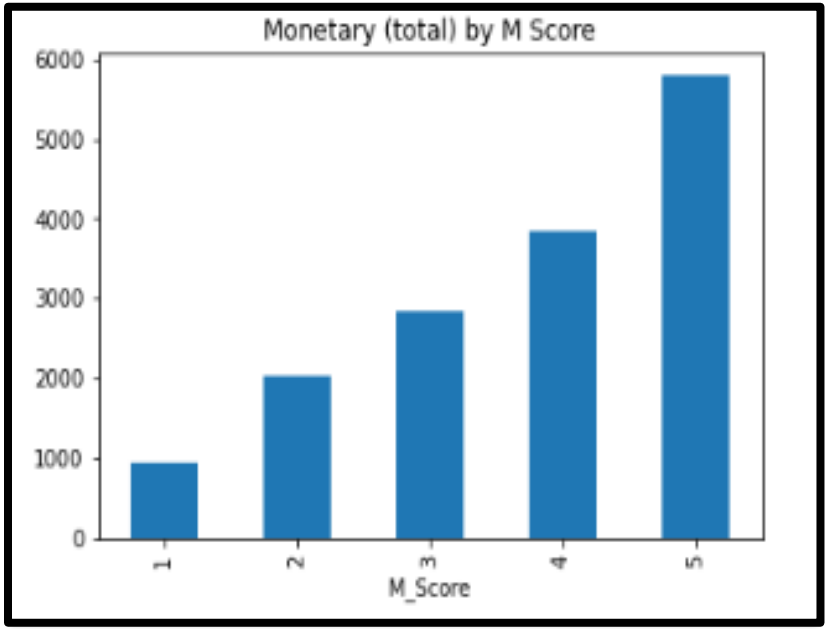
Model Development

RFM Analysis



Model Development

RFM Analysis



Model Development

RFM Analysis

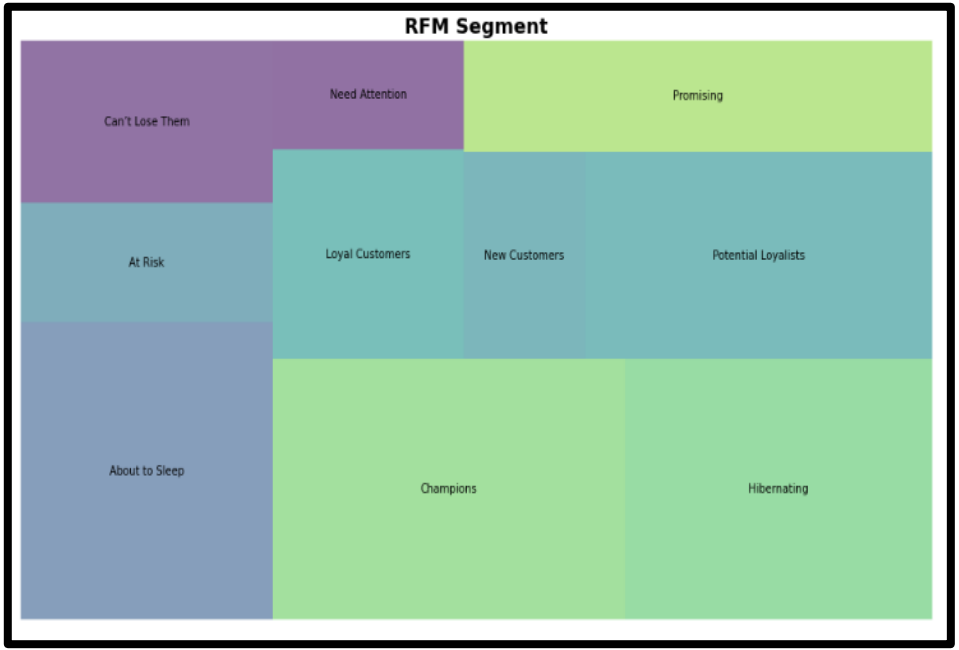
Segmenting customers by the RFM Score in 5 quartiles (considering that the sum of the 3 has the same weight) we obtain the following segmentation:

	recency (days)	frequency (number)	monetary (total)	
	mean	mean	mean	count
RFM_Level				
Bronze	130.28	3.14	1330.03	709
Gold	44.67	5.85	3237.07	638
Platinum	37.62	7.20	4125.47	577
Silver	62.87	4.54	2328.67	947
VIP	18.10	8.76	5454.28	544

- **VIP:** These customers have recently made a purchase, are frequent and are most profitable.
- **Platinum**
- **Gold**
- **Silver**
- **Bronze:** These customers have not recently made a purchase, are not frequent and do not spend a lot.

Model Development

RFM Analysis

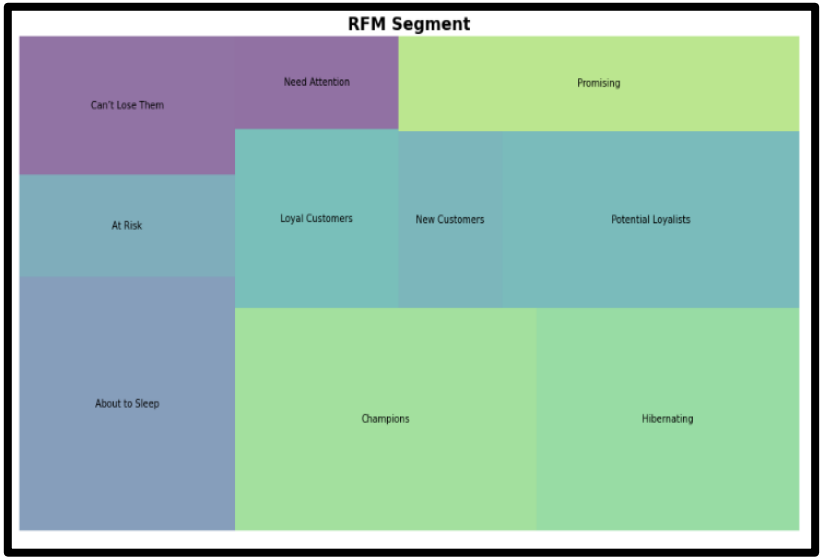


Segmenting customers only by the R and F Score. We can segment clients as follows:

	recency (days)	frequency (number)	monetary (total)	
	mean	mean	mean	count
RFM_type				
About to Sleep	71.12	4.12	2228.71	484
At Risk	104.13	6.00	3344.73	196
Can't Lose Them	91.08	7.88	4257.91	265
Champions	14.64	8.45	4665.54	593
Hibernating	166.38	3.37	1836.81	518
Loyal Customers	45.28	8.16	4533.61	258
Need Attention	46.73	6.00	3434.93	135
New Customers	7.53	3.48	2065.45	163
Potential Loyalists	15.09	5.50	3003.46	465
Promising	35.17	3.32	1815.39	338

Model Development

RFM Analysis



Segmenting customers only by the R and F Score. We can segment clients as follows:

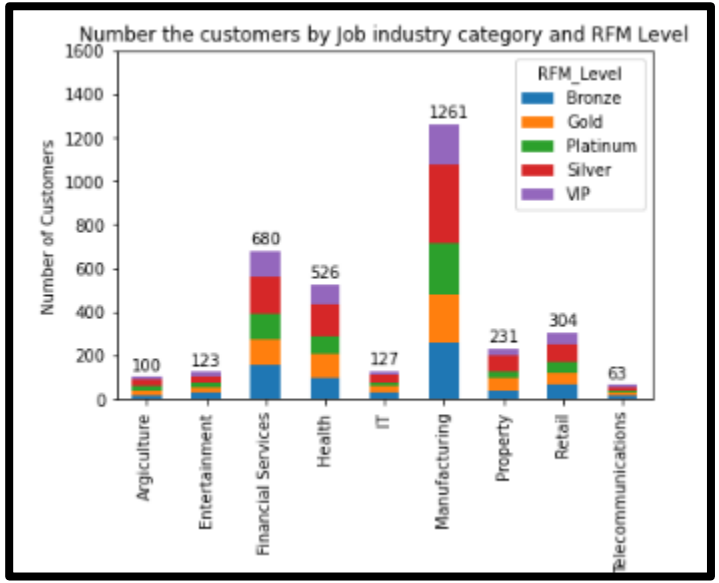
- **Champions:** Bought recently, buy often and spend the most.
- **Loyal Customers:** Spend good money often and responsive to promotions.
- **Potential Loyalists:** Recent customers, but spent a good amount and bought more than once

- **New Customers:** Bought most recently, but not often.
- **Promising:** Recent shoppers, but haven't spent much.
- **Need Attention:** Above average recency, frequency and monetary values. May not have bought very recently though.

- **About To Sleep:** Below average recency, frequency and monetary values. Will lose them if not reactivated.
- **Cannot Lose Them:** Made biggest purchases, and often. But haven't returned for a long time.
- **At Risk:** Engaged with your app and purchased often, but not for awhile. Time to bring them back.
- **Hibernating:** Last purchase was long ago. Low spenders and low number of orders

Model Development

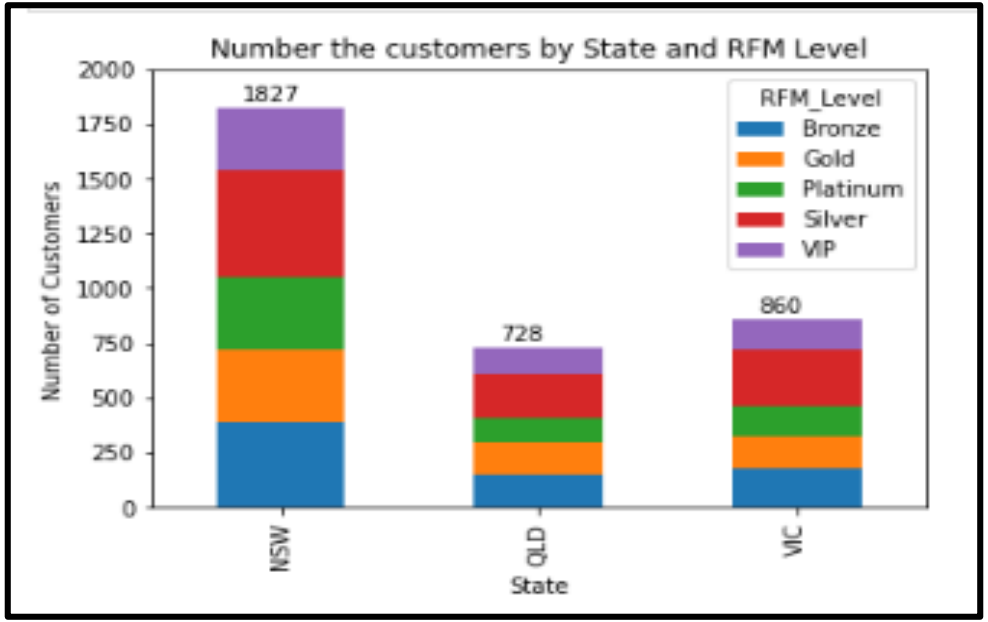
RFM Analysis



job_industry_category RFM_type	Argiculture	Entertainment	Financial Services	Health	IT	Manufacturing	Property	Retail	Telecommunications
About to Sleep	13	16	98	68	23	191	31	38	6
At Risk	8	10	31	29	11	75	9	21	2
Can't Lose Them	13	9	49	35	8	98	22	24	7
Champions	17	19	130	96	14	220	36	51	10
Hibernating	13	22	114	71	19	188	27	50	14
Loyal Customers	3	10	41	47	10	98	13	31	5
Need Attention	3	6	19	22	8	52	13	10	2
New Customers	7	6	27	32	3	54	12	18	4
Potential Loyalists	13	12	103	73	19	162	34	38	11
Promising	10	13	68	53	12	123	34	23	2

Model Development

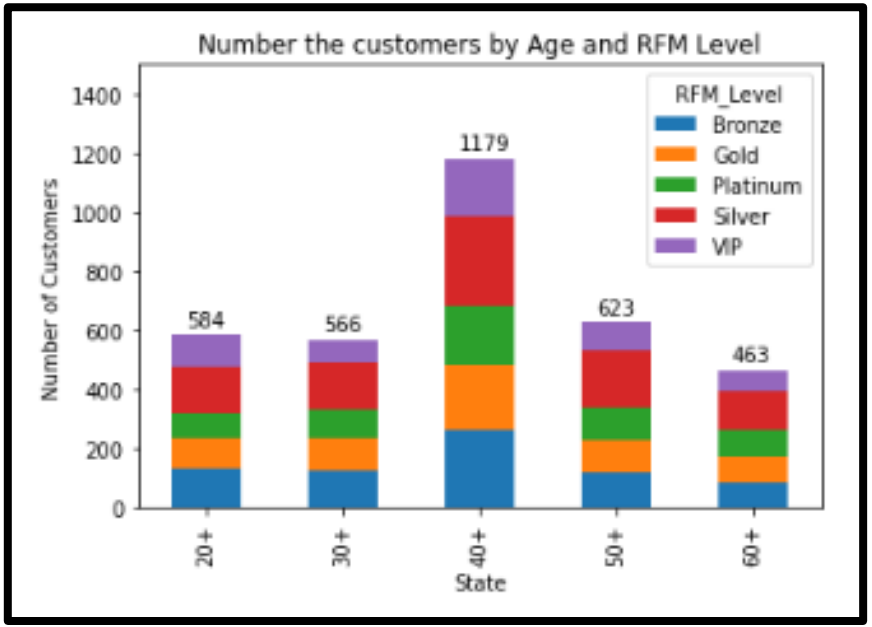
RFM Analysis



state	NSW	QLD	VIC
RFM_type			
About to Sleep	253	109	122
At Risk	99	44	53
Can't Lose Them	147	64	54
Champions	324	113	156
Hibernating	270	101	147
Loyal Customers	135	63	60
Need Attention	74	25	36
New Customers	87	37	39
Potential Loyalists	254	104	107
Promising	184	68	86

Model Development

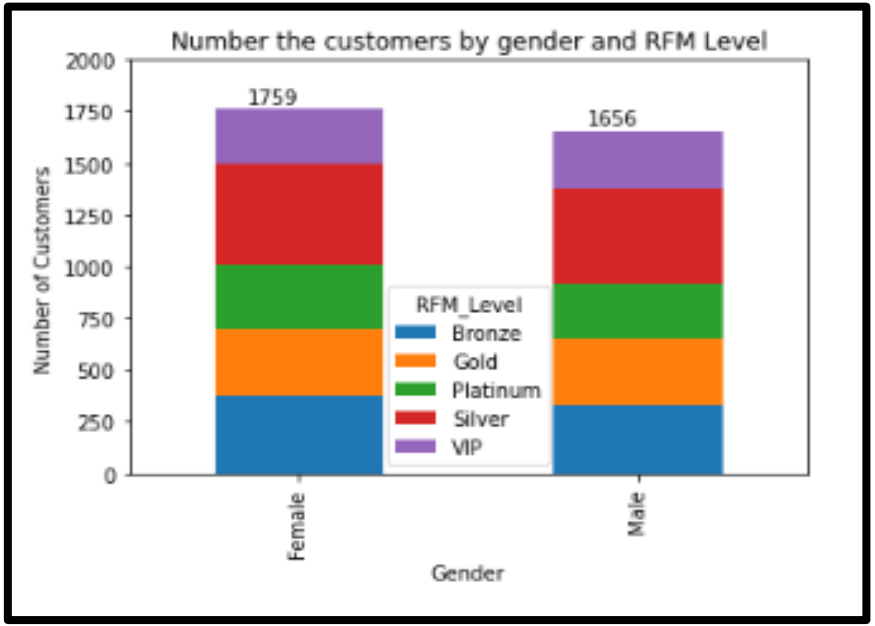
RFM Analysis



Age_C	20+	30+	40+	50+	60+
RFM_type					
About to Sleep	84	93	149	93	65
At Risk	29	34	65	36	32
Can't Lose Them	41	40	104	49	31
Champions	114	90	205	103	81
Hibernating	96	79	190	94	59
Loyal Customers	43	38	88	48	41
Need Attention	15	28	48	22	22
New Customers	30	33	45	31	24
Potential Loyalists	75	72	166	94	58
Promising	57	59	119	53	50

Model Development

RFM Analysis



gender	Female	Male
RFM_type		
About to Sleep	251	233
At Risk	102	94
Can´t Lose Them	135	130
Champions	297	296
Hibernating	265	253
Loyal Customers	134	124
Need Attention	68	67
New Customers	82	81
Potential Loyalists	236	229
Promising	189	149

Interpretation

Insights:

- Our largest wealth segment is the mass customer for the old and new customers.
- The largest number of clients we currently have are in their 40s, but in this new list of clients the largest number is in their 60s.
- Most of our clients live in the state of NSW and most of the new list also reside in that state.
- The proportion between customers who own a car and not, is almost the same, there is very little difference.
- Most of our current clients work in the Manufacturing sector and in the new list of potential clients the greatest number is in financial services.
- We have a greater number of female clients but not by much. In the new client list there is a greater number of female clients as well.

Conclusion:

To attract these new potential clients, we must focus on the fact that the majority are over 60 years of age who mainly belong to financial services and are from the NSW state. So our promotions and commercials must be focused on this group.

Our best clients are mainly from the Manufacturing and financial services sector, they are from the NSW state, are in the 40s and are equally male and female

Appendix

Appendix

This is an optional slide where you may place any supporting items.