Data Analysis using RFM Customer Segmentation Dashboard

A PROJECT REPORT

Submitted by

Priyansh Agarwal [Reg No: RA2111033010062]

Shashwat Tiwari [Reg No: RA2111029010026]

Under the Guidance of

DR. UMA MAHESHWARI

Associate Professor, Department of Data Science and Business Systems

In partial fulfilment of the requirements for the degree of **BACHELOR OF TECHNOLOGY**

in

COMPUTER SCIENCE AND ENGINEERING

with a specialization in Software Engineering

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

COLLEGE OF ENGINEERING AND TECHNOLOGY SRM INSTITUTE OF SCIENCE AND TECHNOLOGY KATTANKULATHUR –

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that this B.Tech project report titled " **Data Analysis using RFM Customer Segmentation Dashboard** " is the bonafide work of **Priyansh Agarwal [Reg. No.: RA2111033010062] , Shashwat Tiwari [Reg. No.:RA2111029010026]** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

DR. UMA MAHESHWARI
SUPERVISOR
Associate Professor
Department of Data Science and Business Systems

DR. M. LAKSHMI
HEAD OF THE DEPARTMENT
Department of Data Science and Business Systems

SIGNATURE OF INTERNAL EXAMINER

SIGNATURE OF EXTERNAL EXAMINER

Department of Data Science and Business Systems

SRM Institute of Science and Technology Own Work Declaration Form

Degree/ Course : B.Tech in Computer Science and Engineering with a specialization in Software Engineering

Student Name : Priyansh Agarwal, Shashwat Tiwari

Registration Number: RA2111033010062, RA2111029010026

Title of Work : Data Analysis using RFM Customer Segmentation Dashboard

We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism, as listed in the University Website, Regulations, and the Education Committee guidelines.

We confirm that all the work contained in this assessment is our own except where indicated, and that we have met the following conditions:

- Clearly references / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books,

web, etc.)

- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:
I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.
If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

ACKNOWLEDGEMENT

We express our humble gratitude to **Dr. C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, **Dr. T.V.Gopal**, for his invaluable support.

We wish to thank **Dr. Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We are incredibly grateful to our Head of the Department, **Dr. M. Lakshmi,** Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our Project Coordinator, **Dr. P. Rajasekar**, Assistant Professor, Panel Head, **Dr. A. Murugan**, Associate Professor and members, **Dr. R. Rajkumar,** Assistant Professor, **Dr. T. Karthick,** Assistant Professor and **Dr. S. Jeeva,** Assistant Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for their inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor, **Dr. K. Priyadarsini**, Assistant Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to our guide, **Dr. A. Murugan**, Associate Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for providing us with an opportunity to pursue our project under his mentorship. He provided us with the freedom and support to explore the research topics of our interest. His passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Data Science and Business Systems Department staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

Priyansh Agarwal [RA2111033010062]

TABLE OF CONTENTS

ABSTRACT

The Data Analysis using RFM Customer Segmentation Dashboard is a robust and user-friendly tool that revolutionizes business decision-making by delving into customer behavior data. Leveraging the RFM analysis framework, the dashboard categorizes customers based on recency, frequency, and monetary dimensions, providing a visually intuitive representation of distinct customer segments. Through dynamic segmentation criteria, businesses can flexibly explore and adapt strategies to align with evolving objectives. The dashboard goes beyond static visualizations, offering detailed customer profiles that highlight average transaction values, purchase frequency, and time since the last transaction for each segment. This actionable insight is complemented by suggested marketing strategies tailored to each customer group, facilitating personalized promotions and engagement campaigns. Real-time monitoring of performance metrics enables businesses to assess the impact of their strategies, while the user-friendly interface ensures accessibility for stakeholders with varying levels of data expertise. Integrated with existing data sources, the dashboard ensures that analyses are based on the latest information, empowering organizations to make informed decisions and enhance customer satisfaction, retention, and overall business growth.

CHAPTER 1

INTRODUCON

1.1. General

RFM (Recency, Frequency, Monetary) analysis is a data-driven marketing technique used to segment and understand customer behavior. This approach involves evaluating three key aspects of customer transactions: recency (how recently a customer made a purchase), frequency (how often a customer makes a purchase), and monetary value (how much money a customer spends). By analyzing these dimensions, businesses can categorize customers into different segments, allowing for targeted and personalized marketing strategies. The goal is to identify high-value customers, understand their preferences, and tailor marketing efforts to maximize engagement, retention, and revenue. RFM analysis is often complemented by data visualization tools and dashboards to provide a clear and actionable overview of customer segments, enabling businesses to make informed decisions and optimize their marketing efforts.

1.2. Purpose

RFM analysis serves a crucial purpose in modern marketing strategies by offering a nuanced understanding of customer behavior. This technique, which evaluates the Recency, Frequency, and Monetary value of customer transactions, enables businesses to categorize their customer base into distinct segments. The primary goal is to go beyond broad demographics and understand the intricacies of customer interactions. By doing so, businesses can tailor their marketing efforts more precisely. High-value customers, identified through RFM analysis, become focal points for targeted strategies aimed at maximizing retention and engagement. This approach not only optimizes resource allocation but also facilitates data-driven decision-making, allowing businesses to adapt and thrive in dynamic market environments. Ultimately, RFM analysis empowers companies to foster meaningful customer relationships, enhance customer satisfaction, and drive sustainable business growth.

1.3. Scope

For the project on hand, we have collected CCTV footage of highways with flowing traffic consisting of a wide range of vehicles with different characteristics. Consistently aggressive patterns include Cutting Lanes, passing on shoulders (exit ramps / entry ramps), Speeding and/or Unusual Acceleration, Lateral Displacement (Lane Switching), Sharp Turns, Unusual Deceleration (excessive braking). However, patterns and variables such as Congestion and Traffic and Factors such as Age, Gender, Presence/Absence of passengers, Honking and tailgating will not be included in the scope our project.

Our solution requires video that is recorded from a static point of view. This is to enable consistent results which wouldn’t be possible with large disparities in standards and formats. Videos for our project are required to have vehicles in question to be in frame for at least a few seconds to enable our model to understand the characteristics and trajectories of the vehicle. Videos which contain a vehicle being in frame for longer durations will ensure better accuracy in terms of understanding the behaviors of the particular vehicle caused by the driver.

CHAPTER 2

LITERATURE REVIEW

While I can't provide an actual literature review, I can give you an overview of how researchers and practitioners have approached and discussed RFM analysis in the literature.

RFM analysis has been a widely adopted methodology in marketing literature and practice for several decades. The concept was first introduced in the context of direct marketing, emphasizing the importance of recency, frequency, and monetary value in understanding and predicting customer behavior. Numerous studies have explored the effectiveness of RFM analysis across various industries, highlighting its versatility and applicability.

Researchers have delved into different aspects of RFM analysis, from the development of sophisticated algorithms for automated segmentation to the integration of machine learning techniques to enhance predictive modeling. Studies have demonstrated the utility of RFM analysis in customer segmentation for personalized marketing campaigns, emphasizing its role in improving customer retention and increasing the efficiency of marketing expenditures.

Furthermore, the literature highlights the evolving nature of RFM analysis in the digital age. With the proliferation of online transactions and the availability of vast amounts of customer data, scholars have explored the integration of RFM with advanced analytics and big data techniques. This integration aims to extract deeper insights into customer behavior, enabling businesses to adapt to changing market dynamics and consumer expectations.

Challenges and limitations of RFM analysis have also been a focus of scholarly discussions. Some researchers have addressed issues related to the static nature of traditional RFM models and the need for dynamic approaches that account for changing customer behaviors over time.

In summary, the literature on RFM analysis is rich and diverse, covering a spectrum of topics ranging from the foundational principles of recency, frequency, and monetary value to advanced applications in the era of big data and machine learning. The body of research collectively underscores the enduring relevance and adaptability of RFM analysis in contemporary marketing strategies.

CHAPTER 3

PROPOSED METHODOLOGY

Proposing a methodology for conducting RFM (Recency, Frequency, Monetary) analysis involves a systematic approach to gather, analyze, and interpret customer data. Below is a general outline for a proposed methodology:

Define Objectives:

Clearly articulate the objectives of the RFM analysis. Whether the focus is on customer segmentation, personalized marketing strategies, or improving retention, defining specific goals will guide the entire methodology.

Data Collection:

Gather relevant data from your database or data sources. This includes transactional data such as purchase dates, order frequencies, and monetary values. Ensure data quality by addressing issues like missing values or outliers.

Data Preprocessing:

Clean and preprocess the data to make it suitable for analysis. This involves handling missing values, outliers, and ensuring consistency in data formats. Normalize or scale variables if necessary.

RFM Calculation:

Calculate the RFM scores for each customer based on recency, frequency, and monetary metrics. Assign scores or percentiles to each customer based on their behavior in these dimensions.

Segmentation:

Utilize clustering techniques such as k-means clustering or hierarchical clustering to group customers based on their RFM scores. Experiment with different numbers of clusters to find the most meaningful segmentation for the business context.

Profile Analysis:

Conduct a thorough analysis of each segment's characteristics. Understand the average recency, frequency, and monetary values within each segment. Identify patterns and traits that distinguish one segment from another.

Visualization:

Develop visualizations to represent the segmented customer groups. Heatmaps, scatter plots, and bar charts can effectively communicate the distribution of customers across different RFM segments.

Interpretation and Strategy Formulation:

Interpret the findings in the context of the defined objectives. Formulate marketing and operational strategies tailored to each RFM segment. Consider factors such as customer preferences, potential value, and the business's overall goals.

Implementation:

Implement the formulated strategies, whether they involve targeted marketing campaigns, loyalty programs, or personalized communication. Monitor the impact of these strategies on customer behavior and key performance indicators.

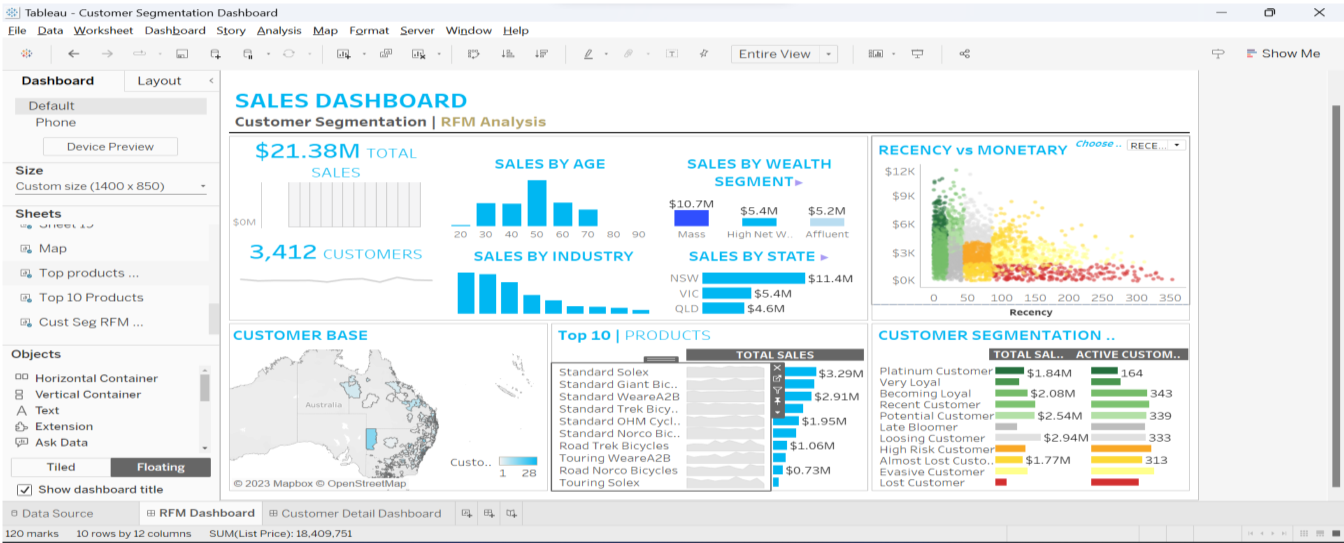
Evaluation and Iteration:

Continuously evaluate the effectiveness of the implemented strategies. Use feedback and performance metrics to iterate and refine the approach. This iterative process ensures the methodology remains adaptive to changing business conditions.

Documentation:

Document the entire methodology, including data sources, preprocessing steps, analysis techniques, and outcomes. This documentation serves as a reference for future analyses and allows for transparency in the decision-making process.

RESULT



RFM ANALYSIS DASHBOARD – It shows the Customer Segmentation of the automobile bicycle company using the RFM method for Data Analysis.

- **Recency, Frequency, and Monetary (RFM) Segmentation**: It shows the relationship using scatterplot of recency vs monetary and frequency vs monetary.

- **Total Sales Over Time:** Display a line or bar chart showing the total sales over a specific time period, allowing users to identify trends and seasonality.

- **Sales by Age:** Illustrate the distribution of customer ages in the dataset. This could be a histogram or a bar chart.

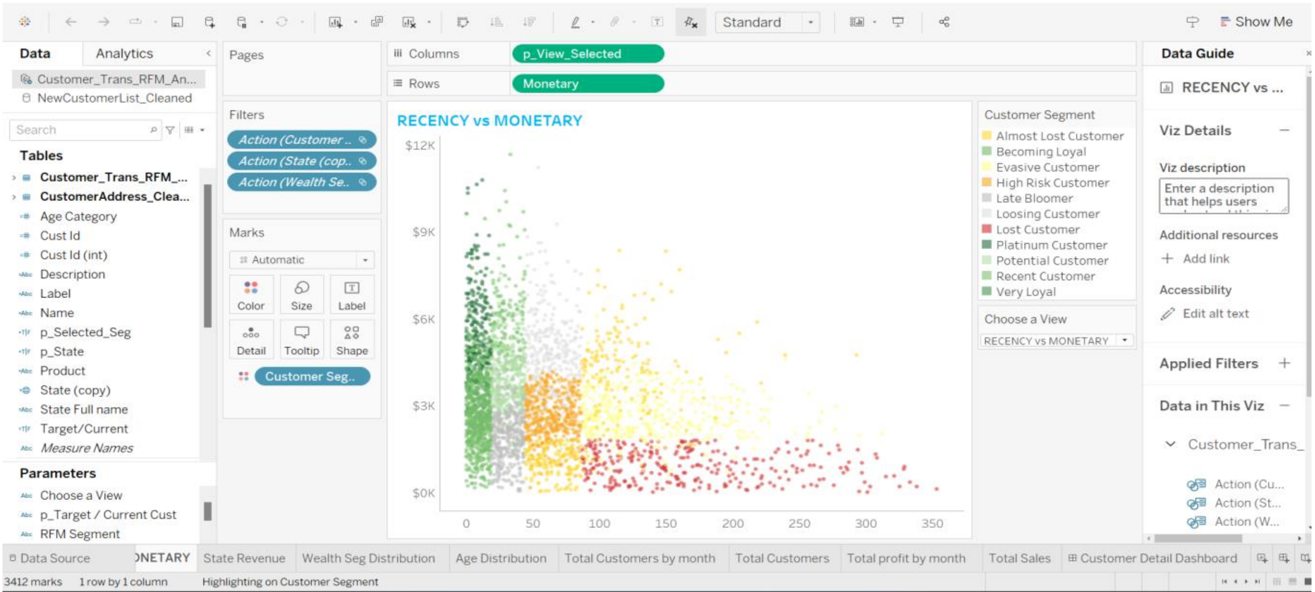
- **Sales by Industry:** Show a breakdown of sales by industry or sector through a pie chart or stacked bar chart.

- **Top Products:** Display a table or chart showcasing the top-selling products based on revenue or quantity sold.

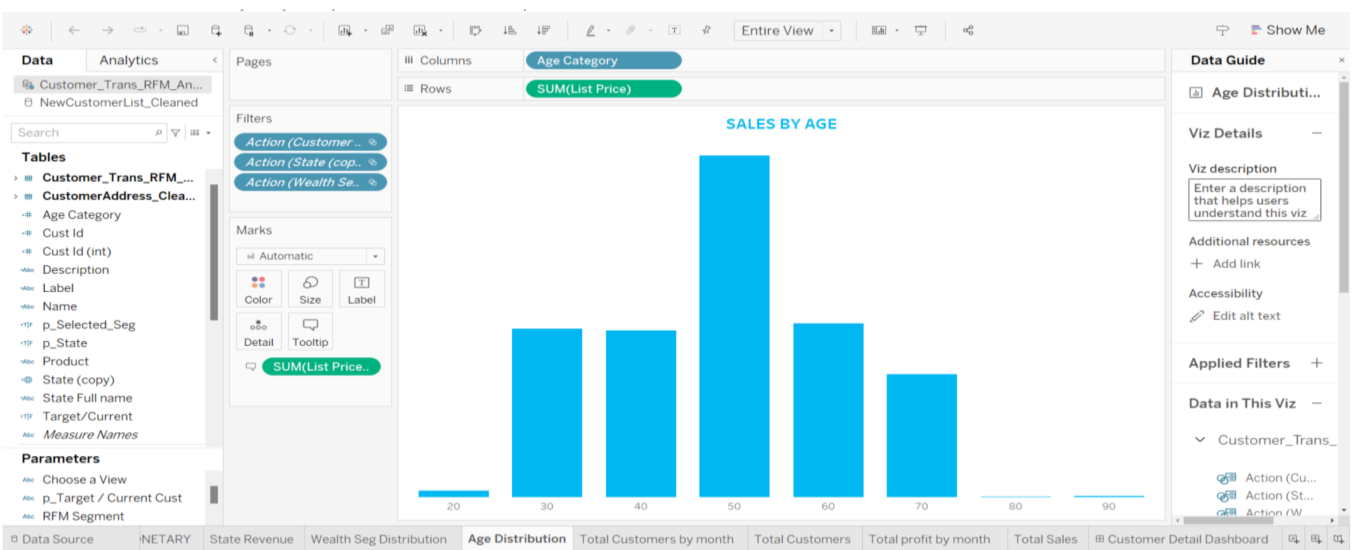
- **Maps of Customer Locations:** (Geographic Distribution) Use a map to visualize the locations of customers. This can be done using heat maps or markers on a geographical map.

- **Customer Demographics**: Include visualizations that provide an overview of customer demographics, such as age, gender, and location.

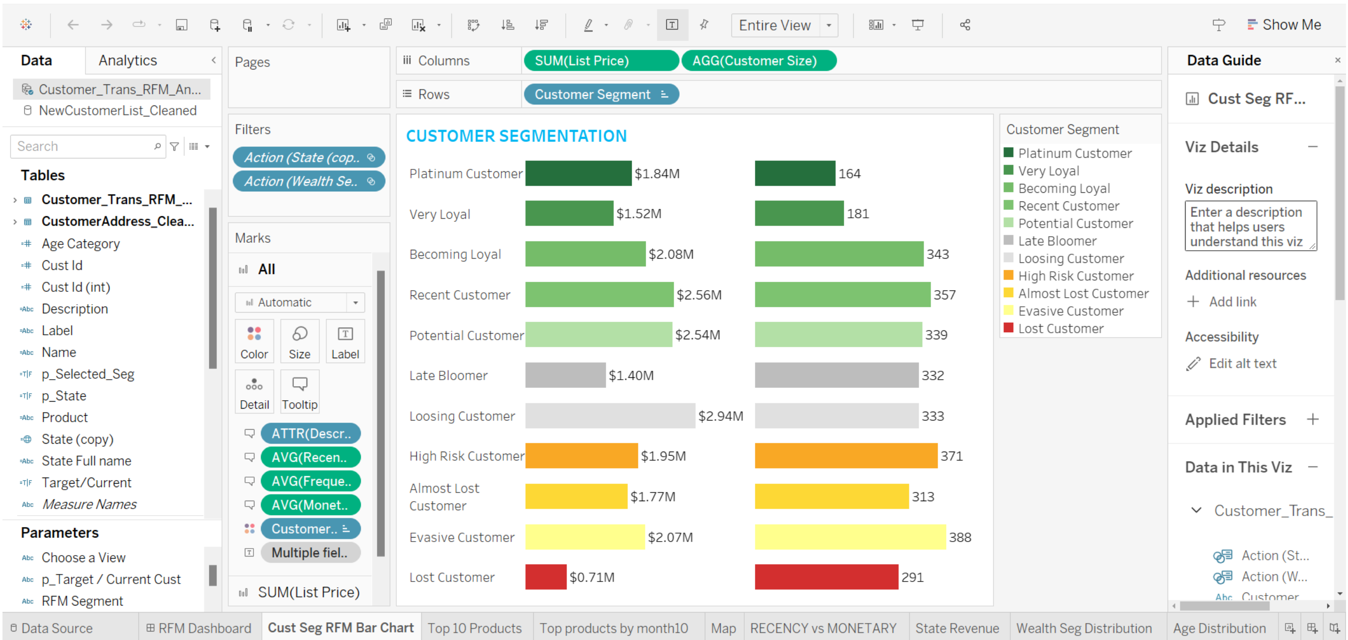
- **Customer segmentation Trends:** Display the trends in customer lifetime value over time.



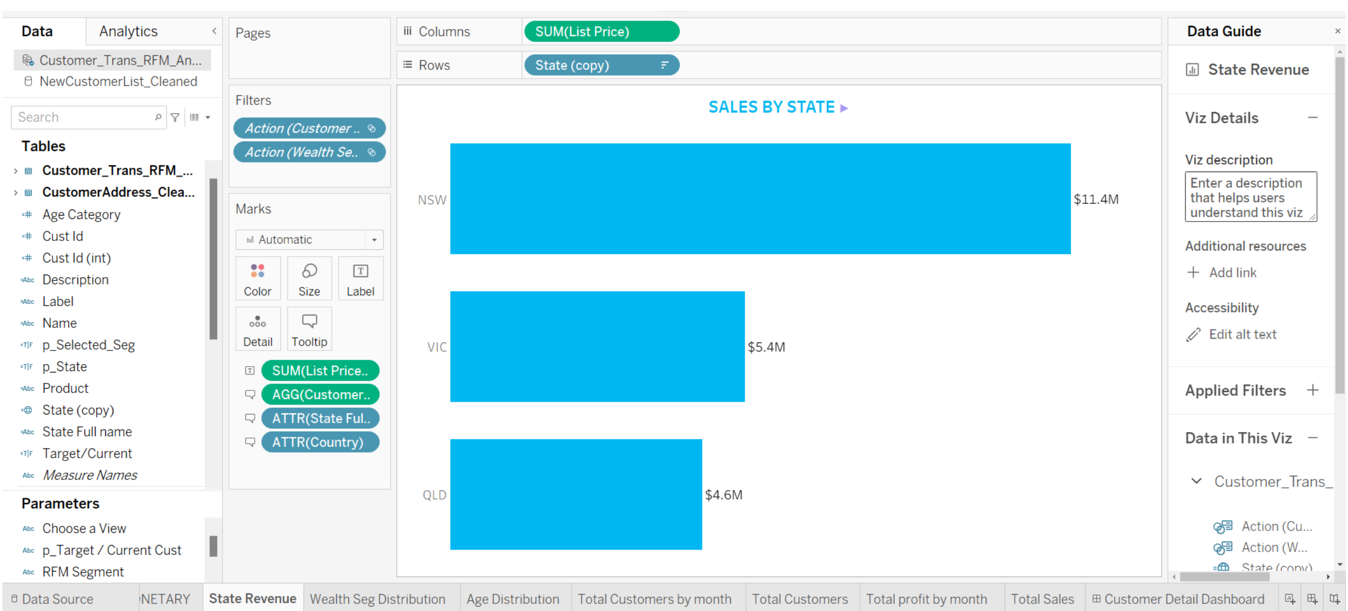
A scatter plot of Recency versus Monetary spending for a car or bike company shows us a quick picture of customer behavior. On one side, you've got folks who recently bought and spent a lot—those are your VIPs. On the other side, there are customers who used to splurge but haven't shopped lately—they might need a nudge. New customers are in a different corner, worth keeping an eye on. The plot helps spot trends, outliers, and understand the typical customer journey. It's like a snapshot guiding where to focus marketing efforts and how to keep customers happy.



It shows the Age Distribution of Customers that bought the Bicycles . The Age of Customers who have bought the most bicyles is shown in Bar Graph.



Using the RFM method for Analysis. We have showed the different segements of Customers that are there and those who have bought the bicycles over the years. The customers who buy a lot of bicyles are the Platinum Customers are showed in Bar Graph.



This Bar Graph shows the states with the most number of sales and the customers who give the most sales belong to the state with high revenue.

CHAPTER 5

CONCLUSION

In conclusion, RFM (Recency, Frequency, Monetary) analysis stands as a powerful methodology in the realm of customer segmentation and marketing strategy. The proposed methodology outlined above provides a systematic and comprehensive approach to leveraging RFM principles for actionable insights. By meticulously collecting and preprocessing customer data, calculating RFM scores, and employing clustering techniques for segmentation, businesses can gain a nuanced understanding of their customer base. The subsequent analysis and visualization of segments enable the formulation of tailored marketing and operational strategies, fostering personalized engagement and driving business growth. The iterative nature of the methodology ensures adaptability to changing market dynamics, while continuous evaluation and documentation contribute to the refinement and enhancement of future analyses. Ultimately, RFM analysis empowers organizations to move beyond one-size-fits-all approaches, optimizing resource allocation, and fostering meaningful customer relationships in the dynamic landscape of modern business

APPENDIX 1

This section contains details on the language, software and packages used in our project.

This project is developed in **Python** and the Data Analysis is implemented on **Tableau**. **Python** which is a general-purpose interpreted, interactive, object oriented and high-level programming language. It offers concise and readable code. Despite being highly complex with versatile workflows, the AI and ML algorithms, when written in Python, can help the developers create robust and reliable machine intelligent systems.

Tableau simplifies the creation of interactive and visually compelling scatter plots, offering ease of use, dynamic data connections, and customization for effective analysis of customer recency versus spending patterns in the automobile or bicycle industry.

The list of Python packages used in our project are:

- **Numpy:** It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation.

- **Pandas:** It is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

- **Matplotlib:** It is a comprehensive library for creating static, animated, and interactive visualizations in Python. It can create publication quality plots, make interactive figures that can zoom, pan, update, customize visual style and layout, export to many file formats and can be embedded in JupyterLab and Graphical User Interfaces.
- **Seaborn:** Seaborn is a data visualization library in Python built on top of Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. Seaborn comes with several built-in themes and color palettes to make it easy to create aesthetically pleasing visualizations

APPENDIX 2

This section tells about the coding and the Datasets used.
We have used **4 Datasets** –

- Customer_Trans_RFM_Analysis.csv
- NewCustomerList_Cleaned.csv
- Transactions_Cleaned.csv
- CustomerDemographic_Cleaned.csv

A1 | fx transaction_id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	transactio	product_i	customer_	transaction_d	online_ord	order_stat	brand	product_li	product_c	product_s	list_price	standard_	product_fi	Profit	first_name	last_name	gender	past_3_ye	DOB
2	1	2	2950	25-02-2017	0	Approved	Solex	Standard	medium	medium	71.49	53.62	41245	17.87	Kristos	Anthony	Male	19	11-01-
3	11065	1	2950	16-10-2017	0	Approved	Giant Bicy	Standard	medium	medium	1403.5	954.82	37659	448.68	Kristos	Anthony	Male	19	11-01-
4	18923	62	2950	26-04-2017	0	Approved	Solex	Standard	medium	medium	478.16	298.72	40487	179.44	Kristos	Anthony	Male	19	11-01-
5	2	3	3120	21-05-2017	1	Approved	Trek Bicyc	Standard	medium	large	2091.47	388.92	41701	1702.55	Lauree	O'Donnell	Female	89	04-02-
6	6862	4	3120	05-10-2017	0	Approved	Giant Bicy	Standard	high	medium	1129.13	677.48	40649	451.65	Lauree	O'Donnell	Female	89	04-02-
7	9155	91	3120	30-10-2017	1	Approved	Solex	Standard	medium	medium	100.35	75.26	38002	25.09	Lauree	O'Donnell	Female	89	04-02-
8	11409	95	3120	30-01-2017	0	Approved	Giant Bicy	Standard	medium	large	569.56	528.43	34556	41.13	Lauree	O'Donnell	Female	89	04-02-
9	13066	38	3120	20-01-2017	1	Approved	Solex	Standard	medium	medium	1577.53	826.51	34071	751.02	Lauree	O'Donnell	Female	89	04-02-
10	15829	41	3120	10-12-2017	0	Approved	Solex	Road	medium	medium	416.98	312.74	41848	104.24	Lauree	O'Donnell	Female	89	04-02-
11	18612	10	3120	24-10-2017	0	Approved	WeareA2B	Touring	medium	medium	1466.68	363.25	38216	1103.43	Lauree	O'Donnell	Female	89	04-02-
12	3	37	402	16-10-2017	0	Approved	OHM Cycl	Standard	low	medium	1793.43	248.82	36361	1544.61	Berne	Donegan	Male	9	03-06-
13	1674	47	402	04-11-2017	1	Approved	Trek Bicyc	Road	low	small	1720.7	1531.42	38991	189.28	Berne	Donegan	Male	9	03-06-
14	13820	12	402	30-09-2017	1	Approved	WeareA2B	Standard	medium	medium	1231.15	161.6	38216	1069.55	Berne	Donegan	Male	9	03-06-
15	17127	3	402	29-01-2017	0	Approved	Trek Bicyc	Standard	medium	large	2091.47	388.92	41167	1702.55	Berne	Donegan	Male	9	03-06-
16	17824	61	402	05-05-2017	1	Approved	OHM Cycl	Standard	low	medium	71.16	56.93	42172	14.23	Berne	Donegan	Male	9	03-06-
17	18312	77	402	03-10-2017	0	Approved	Norco Bicy	Road	medium	large	1240.31	795.1	40553	445.21	Berne	Donegan	Male	9	03-06-
18	4	88	3135	31-08-2017	0	Approved	Norco Bicy	Standard	medium	medium	1198.46	381.1	36145	817.36	Titus	Worsall	Male	83	14-01-
19	1134	59	3135	27-02-2017	1	Approved	Solex	Standard	medium	large	1061.56	733.58	34170	327.98	Titus	Worsall	Male	83	14-01-
20	7000	98	3135	08-02-2017	0	Approved	Trek Bicyc	Standard	high	medium	358.39	215.03	38002	143.36	Titus	Worsall	Male	83	14-01-
21	8683	59	3135	17-05-2017	0	Approved	Solex	Standard	medium	large	1061.56	733.58	34170	327.98	Titus	Worsall	Male	83	14-01-

Customer_Trans_RFM_Analysis

	A	B	C	D	E	F	G	H	I	J	K	L
1	customer_	first_nam	last_name	gender	past_3_ye	DOB	job_title	job_indus	wealth_se	deceased_	owns_car	tenure
2	1	Laraine	Medendo	Female	93	########	Executive	Health	Mass Cust	N	Yes	11
3	2	Eli	Bockman	Male	81	########	Administr	Financial S	Mass Cust	N	Yes	16
4	3	Arlin	Dearle	Male	61	########	Recruiting	Property	Mass Cust	N	Yes	15
5	4	Talbot	None	Male	33	########	Missing	IT	Mass Cust	N	No	7
6	5	Sheila-katl	Calton	Female	56	########	Senior Edi	Missing	Affluent C	N	Yes	8
7	6	Curr	Duckhous	Male	35	########	Missing	Retail	High Net V	N	Yes	13
8	7	Fina	Merali	Female	6	########	Missing	Financial S	Affluent C	N	Yes	11
9	8	Rod	Inder	Male	31	########	Media Ma	Missing	Mass Cust	N	No	7
10	9	Mala	Lind	Female	97	########	Business S	Argicultur	Affluent C	N	Yes	8
11	10	Fiorenze	Birdall	Female	49	########	Senior Qu	Financial S	Mass Cust	N	Yes	20
12	11	Uriah	Bisatt	Male	99	########	Missing	Property	Mass Cust	N	No	9
13	12	Sawyere	Flattman	Male	58	########	Nuclear Pc	Manufact	Mass Cust	N	No	8
14	13	Gabriele	Norcross	Male	38	########	Developer	Financial S	High Net V	N	Yes	8
15	14	Rayshell	Kitteman	Female	85	########	Account E	Financial S	Affluent C	N	No	6
16	15	Erroll	Radage	Male	91	########	Junior Exe	Manufact	Mass Cust	N	No	1
17	16	Harlin	Parr	Male	38	########	Media Ma	Missing	Mass Cust	N	Yes	18
18	17	Heath	Faraday	Male	57	########	Sales Asso	Missing	Affluent C	N	Yes	15
19	18	Marjie	Neasham	Female	79	########	Professor	Missing	Affluent C	N	No	11
20	19	Sorcha	Keyson	Female	76	########	Geological	Manufact	High Net V	N	No	1
21	20	Basile	Firth	Male	72	########	Project Ma	Manufact	Mass Cust	N	No	11

CustomerDemographic_Cleaned

A1 first_name

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	first_nam	last_name	gender	past_3_ye	DOB	job_title	job_indus	wealth_se	deceased_	owns_car	tenure	address	postcode	state	country	property_	Rank	Value	Age	Ag
2	Chickie	Brister	Male	86	########	General M	Manufactu	Mass Cust	N	Yes	14	45 Shopko	4500	QLD	Australia	6	1	1.71875	63	
3	Morly	Genery	Male	69	########	Structural	Property	Mass Cust	N	No	16	14 Mccorr	2113	NSW	Australia	11	1	1.71875	51	
4	Ardelis	Forrester	Female	10	########	Senior Cos	Financial S	Affluent C	N	No	10	5 Colorado	3505	VIC	Australia	5	1	1.71875	46	
5	Lucine	Stutt	Female	64	########	Account R	Manufactu	Affluent C	N	Yes	5	207 Annar	4814	QLD	Australia	1	4	1.703125	42	
6	Melinda	Hadlee	Female	34	########	Financial A	Financial S	Affluent C	N	No	19	115 Monta	2093	NSW	Australia	9	4	1.703125	55	
7	Druci	Brandli	Female	39	########	Assistant I	Entertainn	High Net V	N	Yes	22	89105 Pea	4075	QLD	Australia	7	6	1.671875	70	
8	Rutledge	Hallt	Male	23	########	Compensa	Financial S	Mass Cust	N	No	8	7 Nevada	2620	NSW	Australia	7	6	1.671875	44	
9	Nancie	Vian	Female	74	########	Human Re	Retail	Mass Cust	N	Yes	10	85 Carioca	4814	QLD	Australia	5	8	1.65625	48	
10	Duff	Karlowicz	Male	50	########	Speech Pa	Manufactu	Mass Cust	N	Yes	5	717 West	2200	NSW	Australia	10	8	1.65625	49	
11	Barthel	Docket	Male	72	########	Accountin	IT	Mass Cust	N	Yes	17	80 Scofield	4151	QLD	Australia	5	10	1.640625	35	
12	Rockwell	Matson	Male	94	########	Programm	Retail	High Net V	N	No	3	3682 Crow	4573	QLD	Australia	6	10	1.640625	26	
13	Wheeler	Winward	Male	48	########	Environme	Manufactu	Mass Cust	N	No	10	3 Golden L	3216	VIC	Australia	8	12	1.625	21	
14	Olag	None	Male	60	########	Human Re	Telecomm	Mass Cust	N	No	9	0484 Nort	2032	NSW	Australia	11	13	1.609375	31	
15	Melba	Spellacy	Female	38	########	VP Market	Health	Mass Cust	N	No	4	0591 Anzi	2232	NSW	Australia	10	14	1.59375	44	
16	Mandie	Feares	Female	32	########	Clinical Sp	Health	Mass Cust	N	No	10	39 Kedzie	4053	QLD	Australia	8	14	1.59375	57	
17	Dukie	Swire	Male	88	########	Missing	Manufactu	Affluent C	N	Yes	5	64 Granby	2500	NSW	Australia	8	16	1.5625	67	
18	Marcelia	Monkleigh	Female	61	########	Associate	Manufactu	Mass Cust	N	Yes	4	610 Swallo	4051	QLD	Australia	6	17	1.546875	27	
19	Winnifred	Beswether	Female	83	########	Actuary	Financial S	Mass Cust	N	No	14	61 4th Stre	3040	VIC	Australia	10	17	1.546875	44	
20	Odilia	Quick	Female	65	########	General M	Manufactu	Affluent C	N	Yes	11	1550 Russ	2222	NSW	Australia	11	19	1.53125	82	
21	Karly	Willavize	Female	2	########	Internal A	Manufactu	High Net V	N	No	12	193 North	2190	NSW	Australia	10	19	1.53125	66	

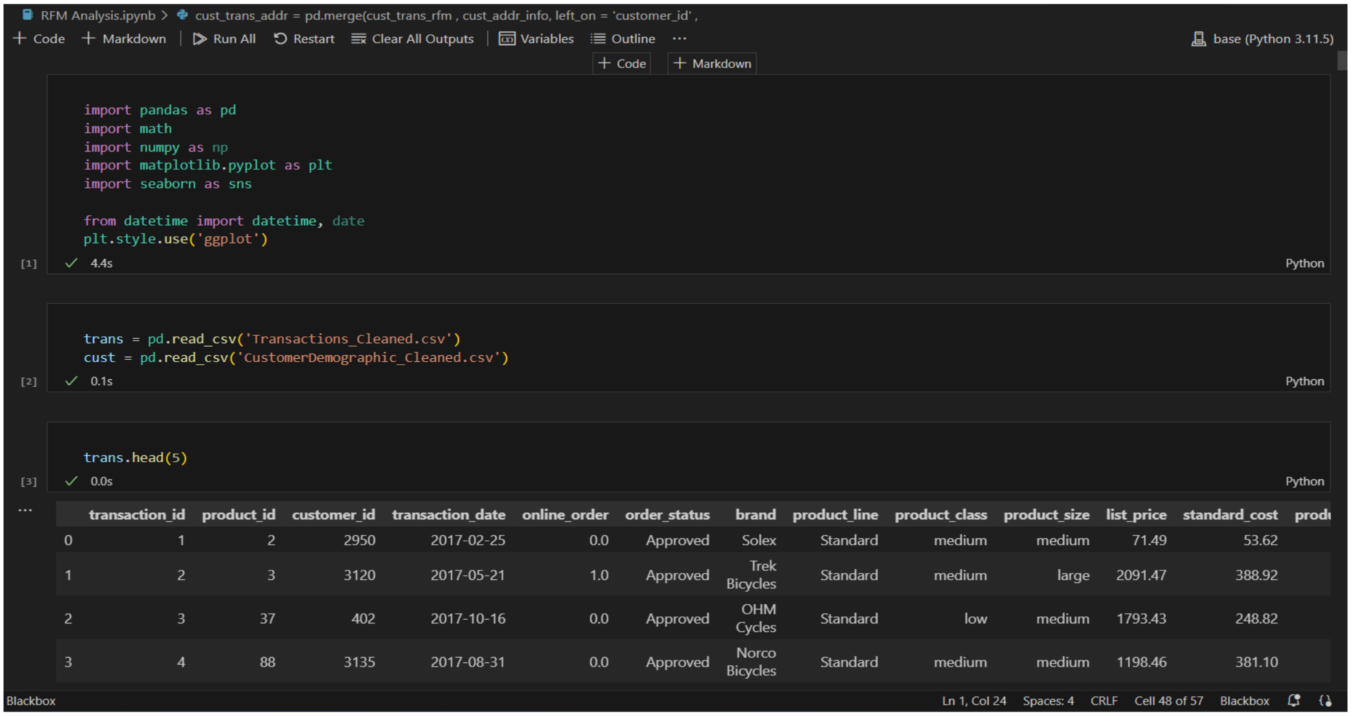
NewCustomerList_Cleaned

A1 transaction_id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	transactio	product_i	customer_	transaction_d	online_or	order_sta	brand	product_li	product_c	product_s	list_price	standard_	product_f	Profit
2	1	2	2950	25-02-2017	0	Approved	Solex	Standard	medium	medium	71.49	53.62	41245	17.87
3	2	3	3120	21-05-2017	1	Approved	Trek Bicyc	Standard	medium	large	2091.47	388.92	41701	1702.55
4	3	37	402	16-10-2017	0	Approved	OHM Cycl	Standard	low	medium	1793.43	248.82	36361	1544.61
5	4	88	3135	31-08-2017	0	Approved	Norco Bicy	Standard	medium	medium	1198.46	381.1	36145	817.36
6	5	78	787	01-10-2017	1	Approved	Giant Bicy	Standard	medium	large	1765.3	709.48	42226	1055.82
7	6	25	2339	08-03-2017	1	Approved	Giant Bicy	Road	medium	medium	1538.99	829.65	39031	709.34
8	7	22	1542	21-04-2017	1	Approved	WeareA2B	Standard	medium	medium	60.34	45.26	34165	15.08
9	8	15	2459	15-07-2017	0	Approved	WeareA2B	Standard	medium	medium	1292.84	13.44	39915	1279.4
10	9	67	1305	10-08-2017	0	Approved	Solex	Standard	medium	large	1071.23	380.74	33455	690.49
11	10	12	3262	30-08-2017	1	Approved	WeareA2B	Standard	medium	medium	1231.15	161.6	38216	1069.55
12	11	5	1986	17-01-2017	0	Approved	Trek Bicyc	Mountain	low	medium	574.64	459.71	40784	114.93
13	12	61	2783	05-01-2017	1	Approved	OHM Cycl	Standard	low	medium	71.16	56.93	42172	14.23
14	13	35	1243	26-02-2017	1	Approved	Trek Bicyc	Standard	low	medium	1057.51	154.4	34527	903.11
15	14	16	2717	10-09-2017	0	Approved	Norco Bicy	Standard	high	small	1661.92	1479.11	34586	182.81
16	15	12	247	11-06-2017	0	Approved	Giant Bicy	Standard	medium	large	1765.3	709.48	38193	1055.82
17	16	3	2961	10-10-2017	0	Approved	Trek Bicyc	Standard	medium	large	2091.47	388.92	37873	1702.55
18	17	79	2426	03-04-2017	0	Approved	Norco Bicy	Standard	medium	medium	1555.58	818.01	38206	737.57
19	18	33	1842	02-06-2017	0	Approved	Giant Bicy	Standard	medium	small	1311.44	1167.18	33888	144.26
20	19	54	2268	06-04-2017	1	Approved	WeareA2B	Standard	medium	medium	1292.84	13.44	39915	1279.4
21	20	25	3002	28-01-2017	1	Approved	Giant Bicy	Road	medium	medium	1538.99	829.65	37337	709.34

Transactions_Cleaned

RFM ANALYSIS:



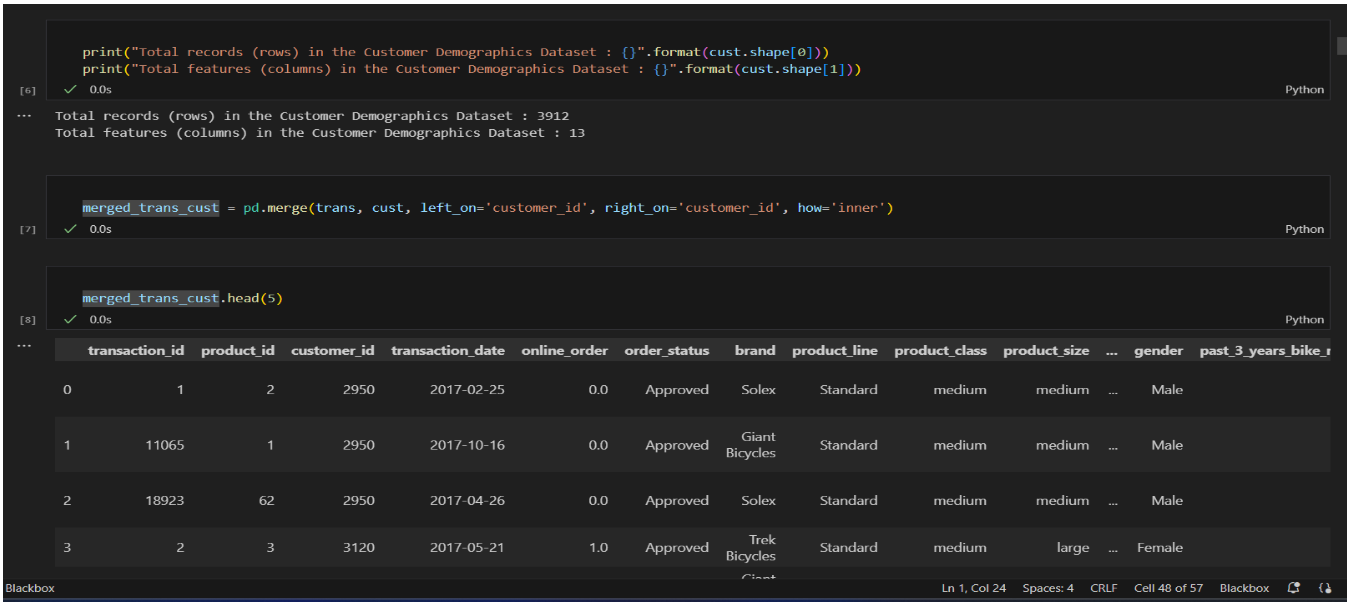
```python
import pandas as pd
import math
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from datetime import datetime, date
plt.style.use('ggplot')
```

```python
trans = pd.read_csv('Transactions_Cleaned.csv')
cust = pd.read_csv('CustomerDemographic_Cleaned.csv')
```

```python
trans.head(5)
```

	transaction_id	product_id	customer_id	transaction_date	online_order	order_status	brand	product_line	product_class	product_size	list_price	standard_cost	prodi
0	1	2	2950	2017-02-25	0.0	Approved	Solex	Standard	medium	medium	71.49	53.62	
1	2	3	3120	2017-05-21	1.0	Approved	Trek Bicycles	Standard	medium	large	2091.47	388.92	
2	3	37	402	2017-10-16	0.0	Approved	OHM Cycles	Standard	low	medium	1793.43	248.82	
3	4	88	3135	2017-08-31	0.0	Approved	Norco Bicycles	Standard	medium	medium	1198.46	381.10	

This shows the imported libraries that we are going to use for performing analysis on our data and to clean the data.



```python
print("Total records (rows) in the Customer Demographics Dataset : {}".format(cust.shape[0]))
print("Total features (columns) in the Customer Demographics Dataset : {}".format(cust.shape[1]))
```

```
Total records (rows) in the Customer Demographics Dataset : 3912
Total features (columns) in the Customer Demographics Dataset : 13
```

```python
merged_trans_cust = pd.merge(trans, cust, left_on='customer_id', right_on='customer_id', how='inner')
```

```python
merged_trans_cust.head(5)
```

	transaction_id	product_id	customer_id	transaction_date	online_order	order_status	brand	product_line	product_class	product_size	...	gender	past_3_years_bike_r
0	1	2	2950	2017-02-25	0.0	Approved	Solex	Standard	medium	medium	...	Male	
1	11065	1	2950	2017-10-16	0.0	Approved	Giant Bicycles	Standard	medium	medium	...	Male	
2	18923	62	2950	2017-04-26	0.0	Approved	Solex	Standard	medium	medium	...	Male	
3	2	3	3120	2017-05-21	1.0	Approved	Trek Bicycles	Standard	medium	large	...	Female	

This shows the file which we have read through pandas in the python and we start performing analysis on it.

```
merged_trans_cust.info()
```
[10] ✓ 0.0s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19354 entries, 0 to 19353
Data columns (total 26 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   transaction_id                    19354 non-null  int64
 1   product_id                        19354 non-null  int64
 2   customer_id                       19354 non-null  int64
 3   transaction_date                  19354 non-null  object
 4   online_order                      19354 non-null  float64
 5   order_status                      19354 non-null  object
 6   brand                             19354 non-null  object
 7   product_line                      19354 non-null  object
 8   product_class                     19354 non-null  object
 9   product_size                      19354 non-null  object
 10  list_price                        19354 non-null  float64
 11  standard_cost                     19354 non-null  float64
 12  product_first_sold_date           19354 non-null  float64
 13  Profit                            19354 non-null  float64
 14  first_name                        19354 non-null  object
 15  last_name                         18732 non-null  object
 16  gender                            19354 non-null  object
 17  past_3_years_bike_related_purchases 19354 non-null  int64
 18  DOB                               19354 non-null  object
 19  job_title                         19354 non-null  object
...
 24  tenure                            19354 non-null  float64
 25  Age                               19354 non-null  int64
dtypes: float64(6), int64(5), object(15)
memory usage: 3.8+ MB
```
Blackbox — Ln 1, Col 18 Spaces: 4 CRLF Cell 48 of 57 Blackbox

This shows the info of the file which we are going to use for further classifying our data.

RFM Analysis.ipynb > cust_trans_addr = pd.merge(cust_trans_rfm , cust_addr_info, left_on = 'customer_id' ,

+ Code + Markdown | Run All | Restart | Clear All Outputs | Variables | Outline ... base (Python 3.11.5)

```
rfm_table = merged_trans_cust.groupby(['customer_id']).agg({'transaction_date': lambda date : (comparison_date - date.max()).days,
                                                            'product_id' : lambda prod_id : len(prod_id),
                                                            'Profit' : lambda p : sum(p)})
```
[14] ✓ 0.8s Python

```
rfm_table.columns
```
[15] ✓ 0.0s Python

```
Index(['transaction_date', 'product_id', 'Profit'], dtype='object')
```

```
rfm_table.rename(columns={'transaction_date' : 'recency',
                          'product_id' : 'frequency',
                          'Profit' : 'monetary'} , inplace=True)
```
[16] ✓ 0.0s Python

```
rfm_table['r_quartile'] = pd.qcut(rfm_table['recency'], 4, ['4','3','2','1'])
rfm_table['f_quartile'] = pd.qcut(rfm_table['frequency'], 4, ['1','2','3','4'])
rfm_table['m_quartile'] = pd.qcut(rfm_table['monetary'], 4, ['1','2','3','4'])
```
[17] ✓ 0.0s Python

```
rfm_table
```
[18] ✓ 0.0s Python

Blackbox — Ln 1, Col 18 Spaces: 4 CRLF Cell 48 of 57 Blackbox

This shows the Combing and cleaning of data to group it in more easier way to visualize.

RFM Analysis.ipynb > cust_trans_addr = pd.merge(cust_trans_rfm , cust_addr_info, left_on = 'customer_id' ,

+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ··· base (Python 3.11.5)

[18] ✓ 0.0s Python

	recency	frequency	monetary	r_quartile	f_quartile	m_quartile
customer_id						
1	7	11	3018.09	4	4	3
2	128	3	2226.26	1	1	2
3	102	8	3362.81	1	4	3
4	195	2	220.57	1	1	1
5	16	6	2394.94	4	2	2
...	...	...	...	...	...	...
3496	256	4	2045.84	1	1	2
3497	52	3	1648.32	2	1	1
3498	127	6	3147.33	1	2	3
3499	51	7	4955.25	2	3	4
3500	144	6	1785.86	1	2	1

3416 rows × 6 columns

```python
rfm_table['rfm_score'] = 100*rfm_table['r_quartile'].astype(int)+10*rfm_table['f_quartile'].astype(int)+rfm_table['m_quartile'].astype(int)
```

[19] ✓ 0.0s Python

This is the RFM table which we will be going to use for our classification of data.

RFM Analysis.ipynb > cust_trans_addr = pd.merge(cust_trans_rfm , cust_addr_info, left_on = 'customer_id' ,

+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ··· base (Python 3.11.5)

[24] ✓ 0.0s Python

Comment Code |

```python
def cust_score_title_lkup(cols):

    rfm_score = cols[0]

    if rfm_score >= 444:
        return 'Platinum Customer'
    elif rfm_score >=433 and rfm_score < 444:
        return 'Very Loyal'
    elif rfm_score >=421 and rfm_score < 433:
        return 'Becoming Loyal'
    elif rfm_score >=344 and rfm_score < 421:
        return 'Recent Customer'
    elif rfm_score >=323 and rfm_score < 344:
        return 'Potential Customer'
    elif rfm_score >=311 and rfm_score < 323:
        return 'Late Bloomer'
    elif rfm_score >=224 and rfm_score < 311:
        return 'Loosing Customer'
    elif rfm_score >=212 and rfm_score < 224:
        return 'High Risk Customer'
    elif rfm_score >=124 and rfm_score < 212:
        return 'Almost Lost Customer'
    elif rfm_score >=112 and rfm_score < 124:
        return 'Evasive Customer'
    else :
        return 'Lost Customer'
```

[25] ✓ 0.0s Python

Blackbox Ln 1, Col 18 Spaces: 4 CRLF Cell 48 of 57 Blackbox

This is the RFM method which we are going to use for customer segmentation.