Acropolis Institute of Technology and Research, Indore

Department of Computer Science and Engineering

B. Tech. VI Semester

Jan - June 2024

Data Analytics Lab (CS-605)

Report File

Submitted To:

Prof. Anurag Punde

Submitted By:

Atharv Sharma

0827IT211022

CS-1 III year

INDEX

Data Analytics Lab

1. 5V's of Big Data

The "5 Vs" of big data refer to five characteristics or dimensions that are often used to describe the nature of big data. These are:

I. Volume:
This refers to the sheer amount of data generated, collected, and stored. With the proliferation of digital devices and sensors, data is being generated at an unprecedented scale. Traditional data processing systems may struggle to handle this volume efficiently.

II. Velocity:
Velocity refers to the speed at which data is generated and processed. Data streams in continuously from various sources such as social media, sensors, and transactional systems. This high velocity requires real-time or near-real-time processing capabilities to extract value from the data in a timely manner.

III. Variety:
Big data comes in various formats and types, including structured data (like databases), semi-structured data (like XML files), and unstructured data (like text documents, images, videos, and social media posts). Dealing with this variety requires flexible data processing and analysis techniques.

IV. Veracity:
Veracity refers to the quality and reliability of the data. Big data sources can often be noisy, incomplete, or inconsistent. Ensuring data quality is crucial for making accurate decisions and deriving meaningful insights. Data cleaning, validation, and verification processes are essential to address veracity challenges.

V. Value:
The ultimate goal of big data is to derive value from the insights gained through analysis. This value can take various forms, including cost reduction, revenue generation, improved decision-making, enhanced customer experience, and innovation. Extracting actionable insights from big data requires sophisticated analytics techniques and tools.

2. Data Analysis Principles

Data analysis principles form the foundation for extracting insights and making informed decisions from data. Here are some key principles:

I. Define clear objectives:
Clearly define the goals and objectives of your analysis. What questions are you trying to answer? What problem are you trying to solve? Having a clear understanding of your objectives will guide your analysis and ensure that you focus on relevant data.

II. Understand the context:
Context is crucial in data analysis. Understand the domain or industry you are working in, as well as the specific context of the data you are analyzing. Consider factors such as the data source, collection methods, and any biases or limitations inherent in the data.

III. Data quality assessment:
Assess the quality of your data before conducting any analysis. Look for errors, inconsistencies, missing values, and outliers. Clean and preprocess the data as needed to ensure accuracy and reliability.

IV. Choose appropriate methods:
Select the most suitable methods and techniques for your analysis based on your objectives and the characteristics of your data. This may involve statistical analysis, machine learning algorithms, data visualization, or a combination of these approaches.

V. Iterative approach:
Data analysis is often an iterative process. Explore the data, generate hypotheses, test them, and refine your analysis based on the results. Iterate until you reach meaningful insights or conclusions.

VI. Interpretation and validation:
Interpret the results of your analysis in the context of your objectives and domain knowledge. Validate your findings using additional data or alternative methods to ensure their reliability.

VII. Communicate effectively:
Communicate your findings in a clear and concise manner to stakeholders, using appropriate visualizations, summaries, and insights. Tailor your communication to your audience, whether they are technical experts or non-experts.

VIII. Ethical considerations:
Consider the ethical implications of your analysis, including privacy, security, and fairness. Ensure that your analysis respects the rights and interests of individuals and complies with relevant regulations and guidelines.

3. Statistical Analysis Concepts

Statistical analysis concepts are fundamental to understanding and interpreting data. Here are some key concepts:

I. Descriptive statistics:
Descriptive statistics summarize and describe the main features of a dataset. This includes measures such as mean, median, mode, standard deviation, range, and percentiles.

II. Inferential statistics:
Inferential statistics involve making inferences and predictions about a population based on a sample of data. This includes techniques such as hypothesis testing, confidence intervals, and regression analysis.

III. Probability distributions:
Probability distributions describe the likelihood of different outcomes in a dataset. Common distributions include the normal distribution, binomial distribution, Poisson distribution, and exponential distribution.

IV. Hypothesis testing:
Hypothesis testing is a statistical method used to make decisions about a population parameter based on sample data. It involves formulating a null hypothesis and an alternative hypothesis, collecting data, and using statistical tests to determine whether to accept or reject the null hypothesis.

V. Confidence intervals:
Confidence intervals provide a range of values within which a population parameter is likely to fall, based on sample data and a specified level of confidence.

VI. Regression analysis:
Regression analysis is a statistical method used to explore the relationship between one or more independent variables and a dependent variable. It allows you to predict the value of the dependent variable based on the values of the independent variables.

VII. Correlation:
Correlation measures the strength and direction of the relationship between two variables. Common correlation coefficients include Pearson correlation coefficient, Spearman rank correlation coefficient, and Kendall tau coefficient.

VIII. ANOVA (Analysis of Variance):
ANOVA is a statistical method used to compare means between two or more groups. It tests whether there are statistically significant differences between the means of the groups.

IX. Resampling methods:
Resampling methods such as bootstrapping and cross-validation are used to estimate the accuracy and variability of statistical estimates by repeatedly sampling from the dataset.

4. Hypothesis

A hypothesis is a proposed explanation for a phenomenon or a statement that can be tested through experimentation or observation. In scientific research, hypotheses are typically formulated based on existing knowledge, theories, or observations and are then tested to determine their validity.

There are two main types of hypotheses:

- Null hypothesis (H0): The null hypothesis states that there is no significant difference or relationship between the variables being studied. It represents the default position or assumption to be tested against an alternative hypothesis.
- Alternative hypothesis (H1 or Ha): The alternative hypothesis contradicts the null hypothesis and suggests that there is a significant difference or relationship between the variables being studied. It is the hypothesis that researchers seek to support with evidence from their study.

For example, in a medical study testing the effectiveness of a new drug, the null hypothesis might be that there is no difference in outcomes between patients who receive the drug and those who receive a placebo. The alternative hypothesis would then state that there is a difference in outcomes between the two groups.

Hypotheses are essential in scientific inquiry because they provide a framework for designing experiments, collecting data, and drawing conclusions based on evidence. They allow researchers to systematically investigate and test ideas, leading to a deeper understanding of the natural world.

5. Regression and it's types

Regression analysis is a statistical method used to explore the relationship between one or more independent variables and a dependent variable. It is commonly used for prediction, forecasting, and understanding the relationship between variables.

There are several types of regression analysis, each suited to different types of data and research questions. Here are some common types:

- Linear Regression: Linear regression is the simplest form of regression analysis, where the relationship between the independent variable(s) and the dependent variable is assumed to be linear. It aims to fit a straight line to the data that minimizes the sum of the squared differences between the observed and predicted values.
- Multiple Regression: Multiple regression extends linear regression to include multiple independent variables. It is used when there are two or more predictors influencing the

dependent variable. The goal is to model the relationship between the predictors and the dependent variable while controlling for the effects of other variables.

- Polynomial Regression: Polynomial regression is a type of regression analysis where the relationship between the independent and dependent variables is modeled as an nth-degree polynomial. It is useful when the relationship between the variables is not linear but can be better approximated by a curve.
- Logistic Regression: Logistic regression is used when the dependent variable is categorical or binary (e.g., yes/no, true/false). It models the probability of the occurrence of a binary outcome based on one or more independent variables. Logistic regression uses a logistic function to predict the probability of the outcome.
- Ridge Regression: Ridge regression is a type of linear regression that is used when multicollinearity (high correlation between independent variables) is present in the dataset. It adds a penalty term to the ordinary least squares (OLS) regression to shrink the coefficients towards zero, reducing the variance of the estimates.
- Lasso Regression: Lasso (Least Absolute Shrinkage and Selection Operator) regression is similar to ridge regression but adds a penalty term that can shrink some coefficients to exactly zero. It is useful for feature selection and can be used to perform variable selection by eliminating irrelevant predictors.
- ElasticNet Regression: ElasticNet regression combines the penalties of ridge and lasso regression. It is used when there are high correlations between predictors and when there are many predictors, some of which may be irrelevant. ElasticNet can perform variable selection and regularization simultaneously.

These are some of the most commonly used types of regression analysis, each with its own assumptions, advantages, and applications. The choice of regression method depends on the nature of the data, the research question, and the underlying assumptions of the analysis.

6. Correlation

Correlation is a statistical measure that describes the strength and direction of a relationship between two variables. It indicates how much one variable changes when another variable changes.

There are several types of correlation coefficients, but the most commonly used is the Pearson correlation coefficient, denoted by r. The Pearson correlation coefficient measures the linear relationship between two continuous variables. It ranges from -1 to 1:

- $r = 1$: Perfect positive correlation
- $r = -1$: Perfect negative correlation
- $r = 0$: No correlation

A correlation coefficient close to 1 or -1 indicates a strong linear relationship between the variables, while a coefficient close to 0 indicates a weak relationship.

It's important to note that correlation does not imply causation. Just because two variables are correlated does not mean that one variable causes the other to change. Correlation simply measures the degree of association between variables.

Other types of correlation coefficients include:

- Spearman's rank correlation coefficient: Used when the variables are ordinal or when the relationship is not linear.
- Kendall's tau coefficient: Similar to Spearman's correlation but places more emphasis on concordant and discordant pairs of data points.
- Point-Biserial correlation coefficient: Used when one variable is dichotomous and the other is continuous.
- Phi coefficient: Used when both variables are dichotomous.

Correlation analysis is widely used in various fields, including economics, psychology, sociology, and biology, to explore relationships between variables and make predictions.

7. Anova

ANOVA, or Analysis of Variance, is a statistical method used to compare means between two or more groups. It assesses whether there are statistically significant differences in the means of the groups based on the variance within and between the groups.

ANOVA works by partitioning the total variance observed in the data into different components:

- Between-group variance: This component measures the variability between the means of the different groups. It indicates whether there are significant differences in the means of the groups being compared.
- Within-group variance: This component measures the variability within each group. It represents the random variation or noise within the groups that is not explained by the group differences.

The main idea behind ANOVA is to compare the ratio of between-group variance to within-group variance. If the between-group variance is significantly larger than the within-group variance, it suggests that there are significant differences between the group means.

ANOVA produces an F-statistic, which is the ratio of the between-group variance to the within-group variance. The F-statistic follows an F-distribution, and its significance is assessed using a p-

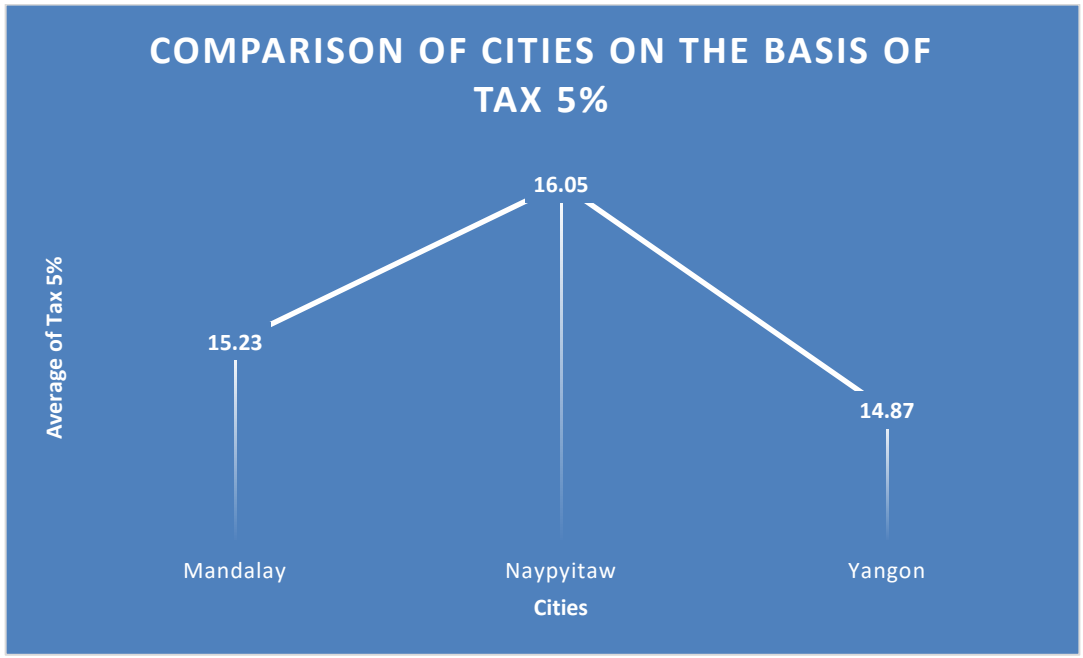
value. If the p-value is below a predetermined significance level (usually 0.05), it indicates that there are significant differences between the group means.

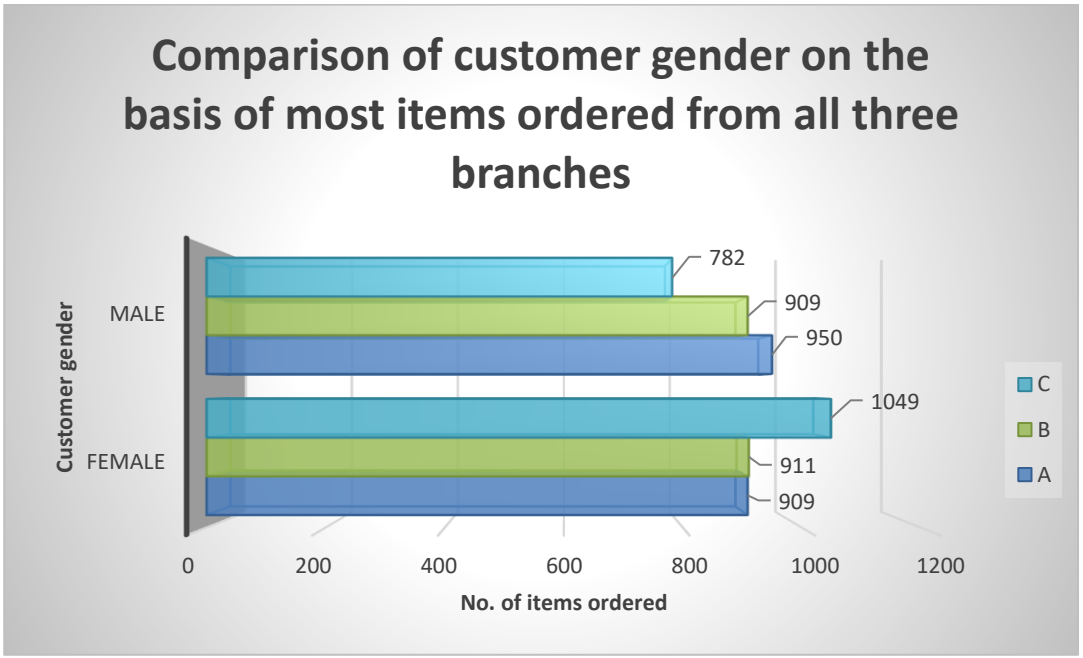
There are several types of ANOVA, depending on the number of factors and levels of each factor:

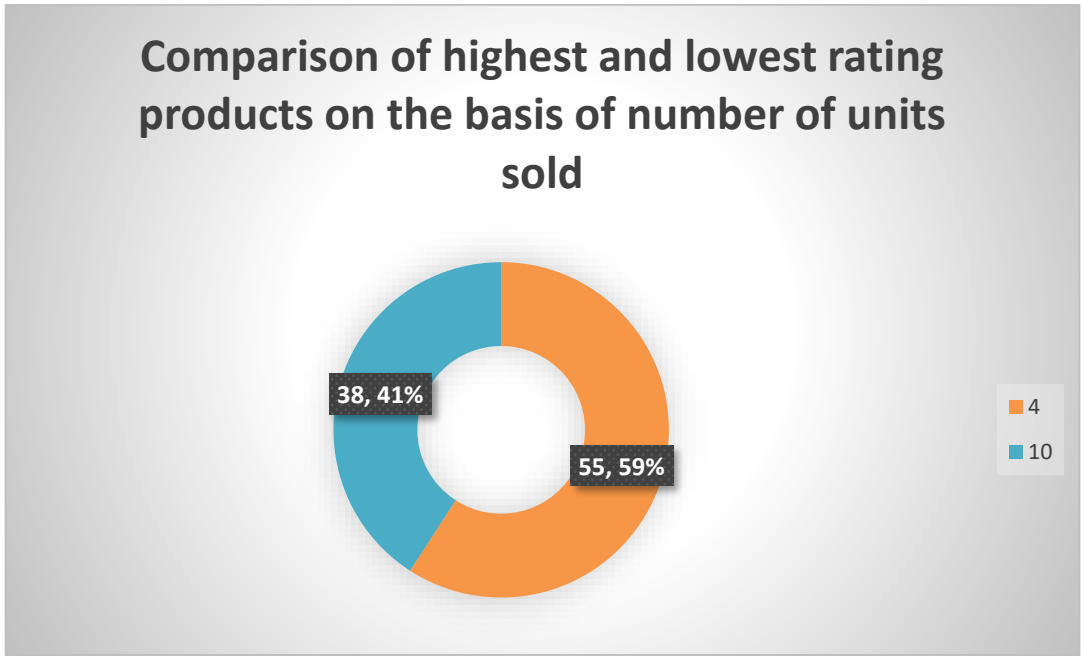
I. One-way ANOVA: Compares the means of two or more independent groups on one factor (or independent variable).
II. Two-way ANOVA: Compares the means of two or more independent groups on two factors, examining the main effects of each factor as well as any interaction between the factors.
III. ANOVA with repeated measures: Compares the means of the same group under different conditions or time points, taking into account the correlation between the repeated measurements.

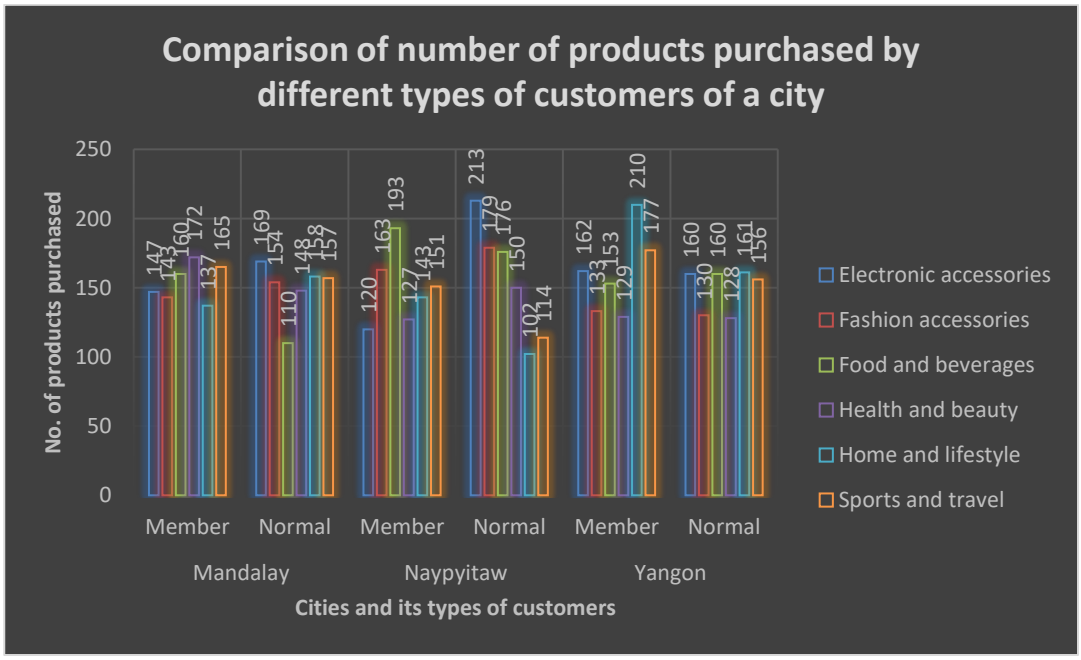
ANOVA is commonly used in experimental research to analyze the results of experiments with multiple treatment groups or factors. It allows researchers to determine whether the observed differences between groups are statistically significant and not due to random variation.

Dashboard for Supermarket Dataset

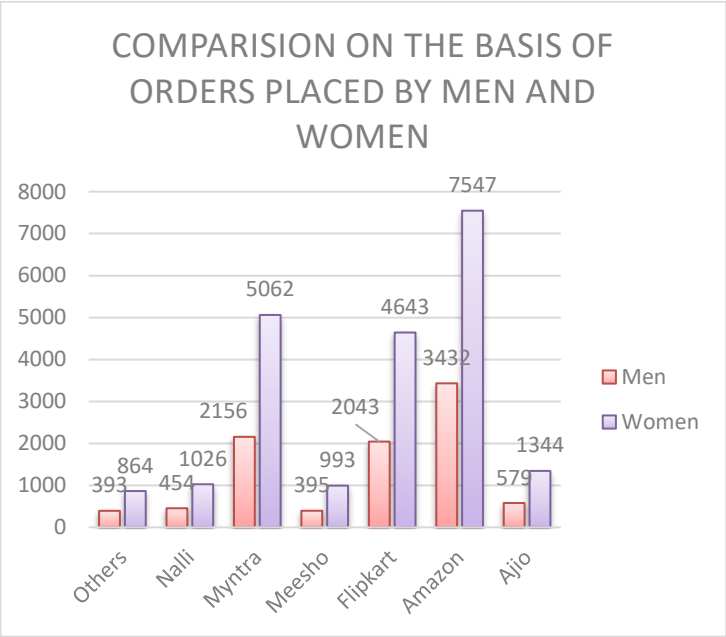


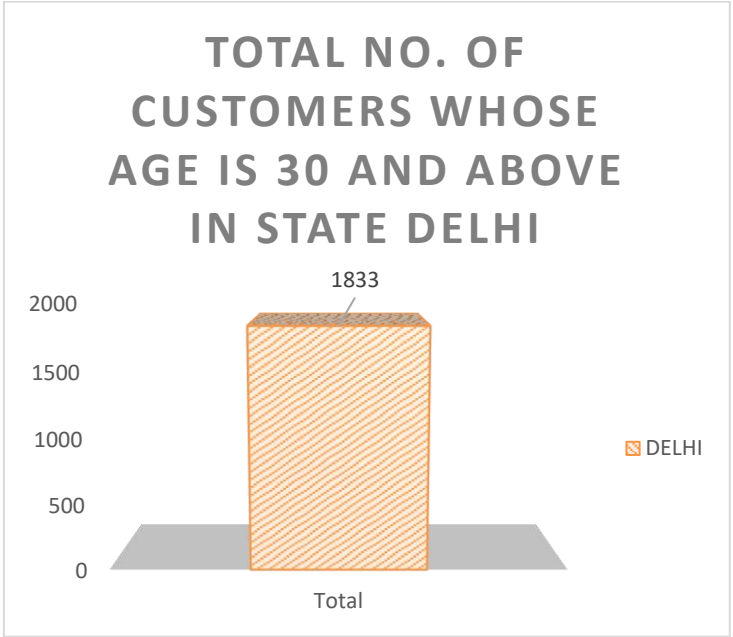


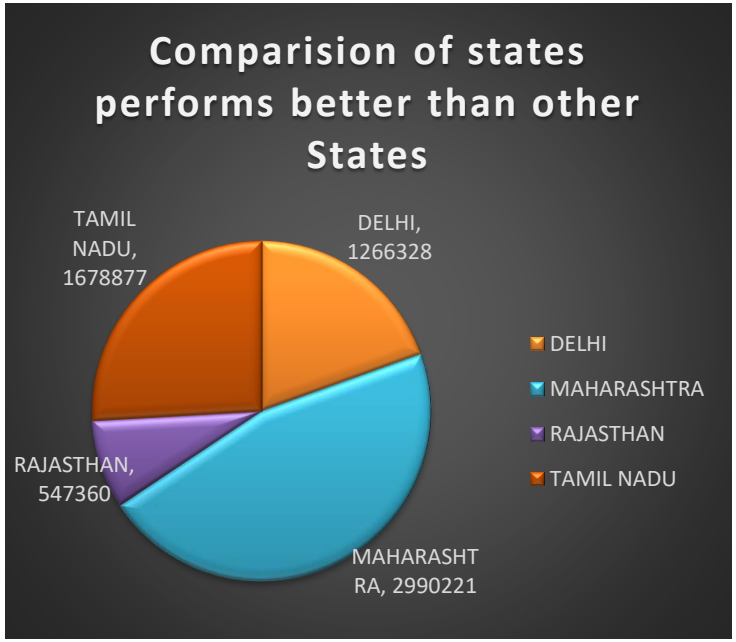


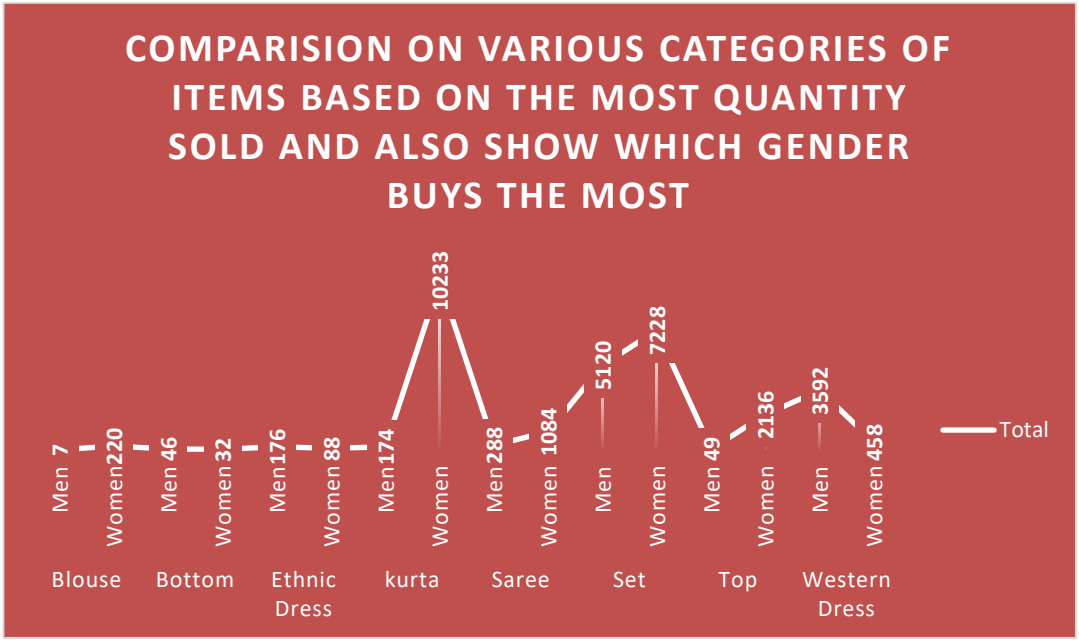


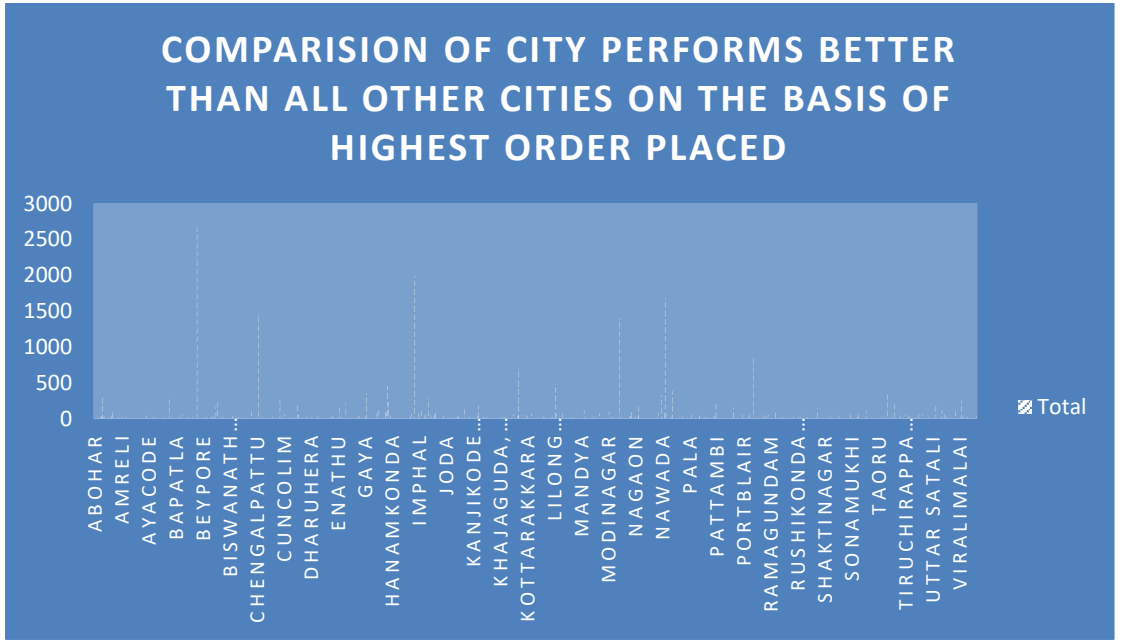
Dashboard for Store Dataset



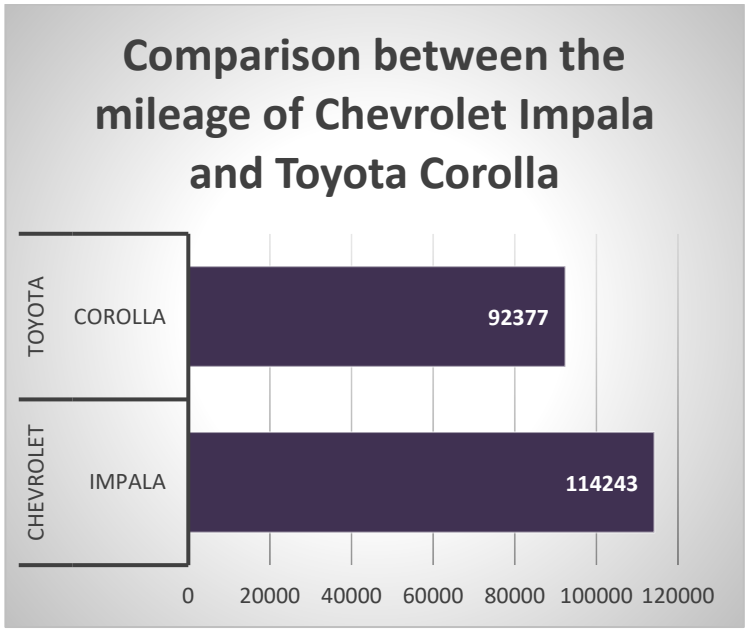


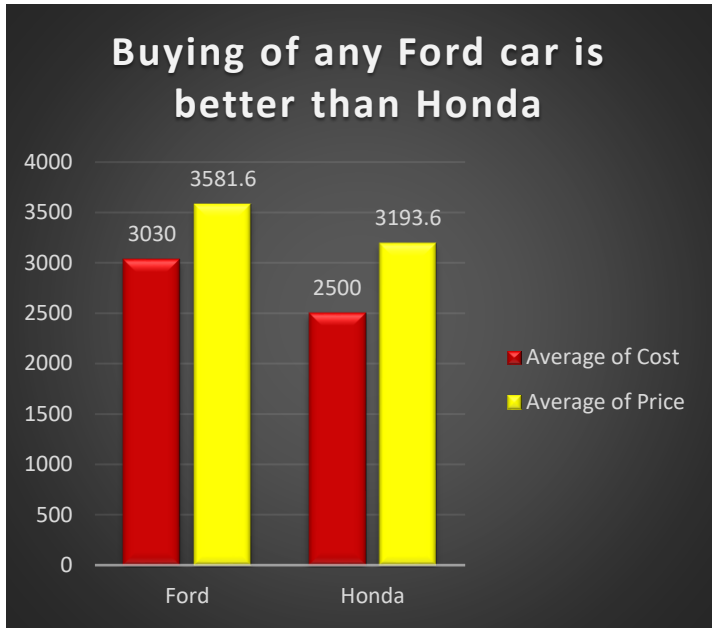


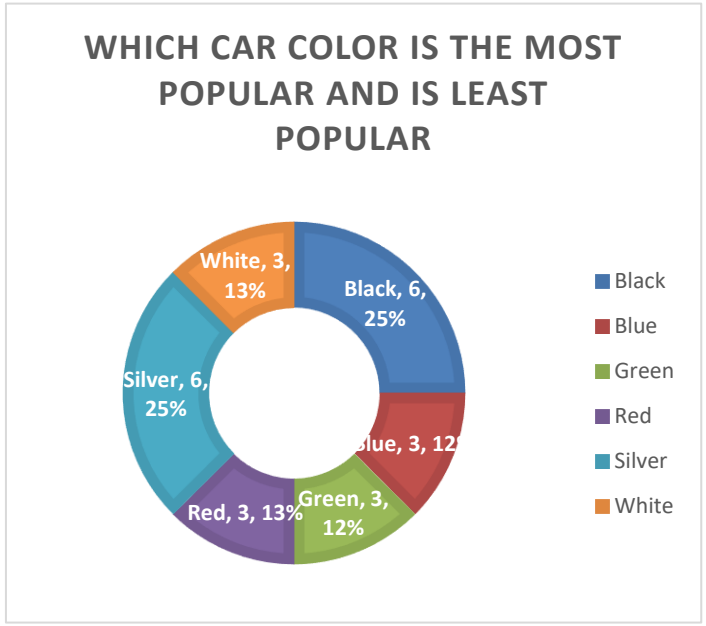


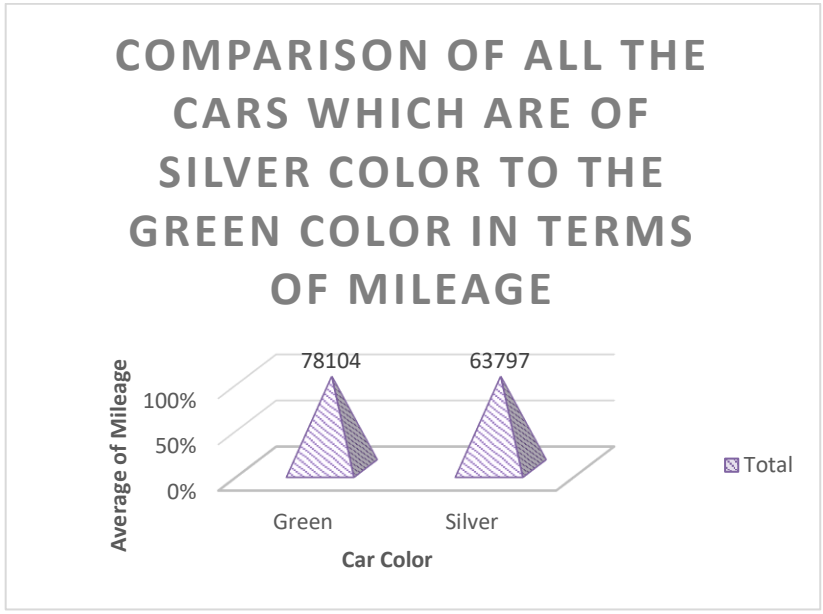


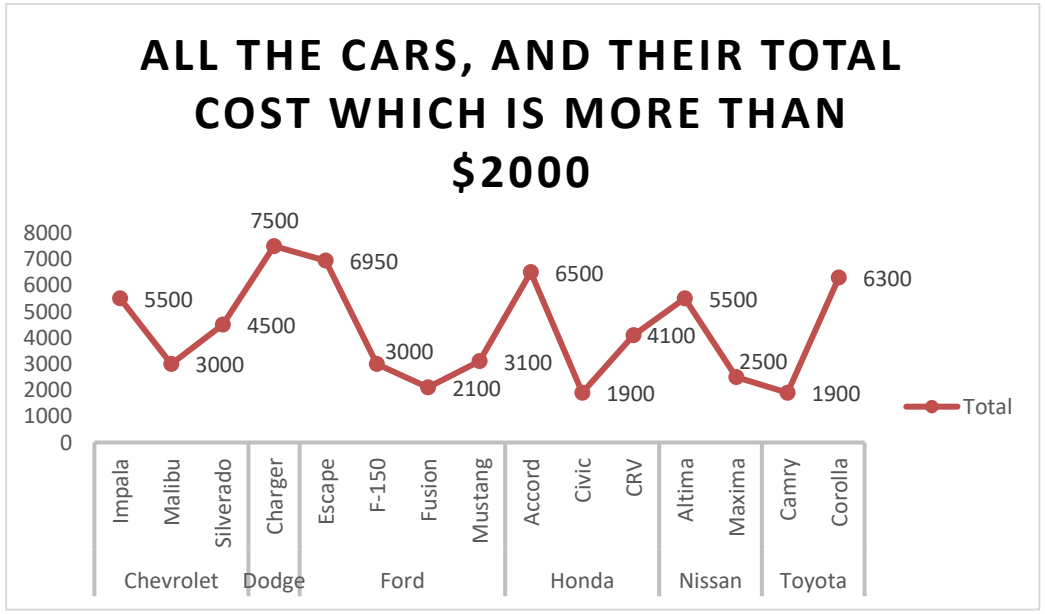
Dashboard for Car Collection Dataset



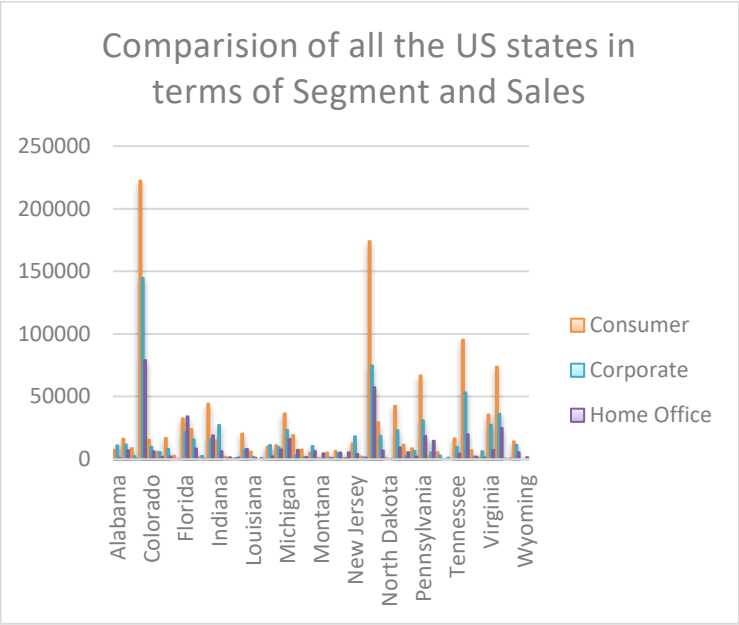


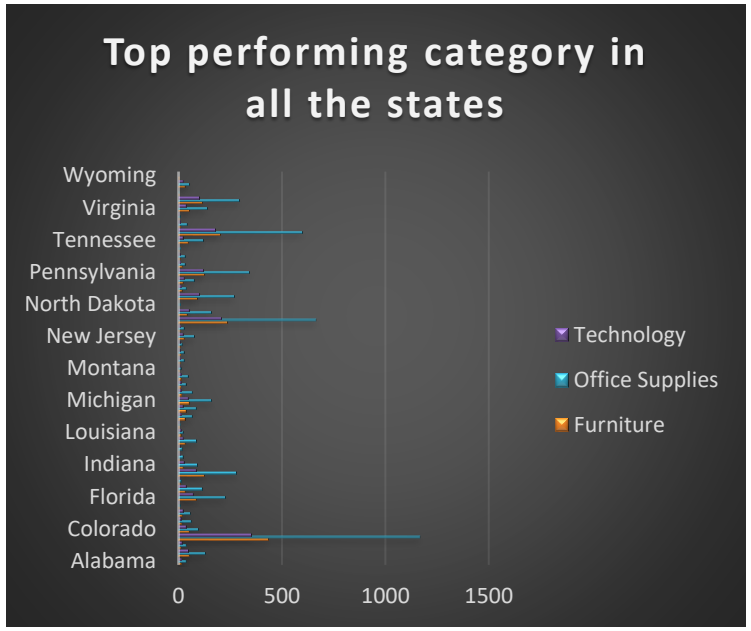


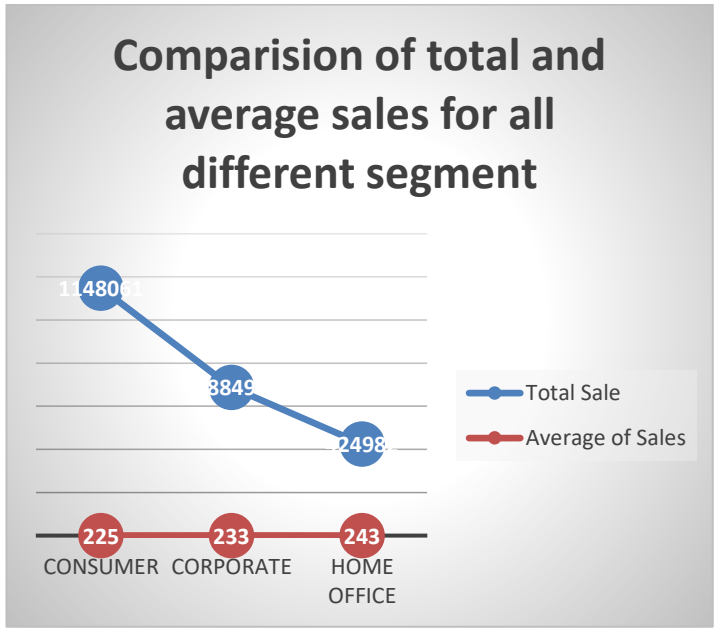


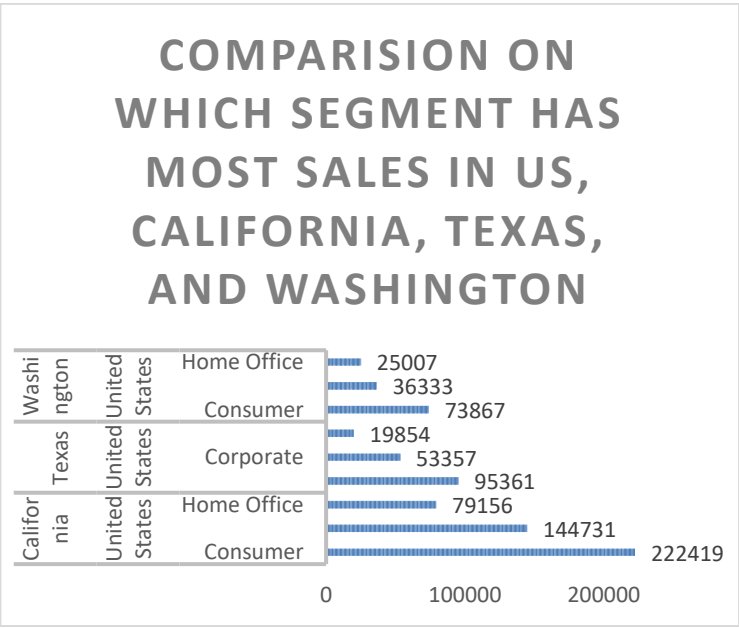


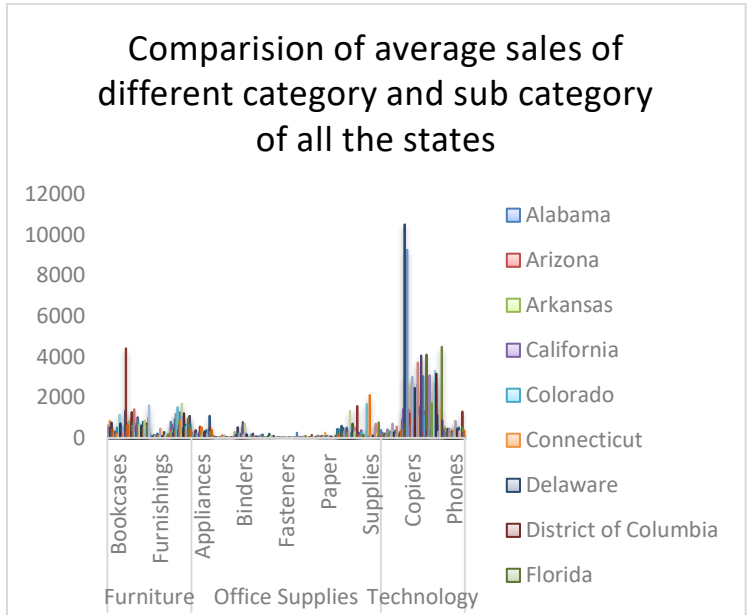
Dashboard for Order Data Dataset

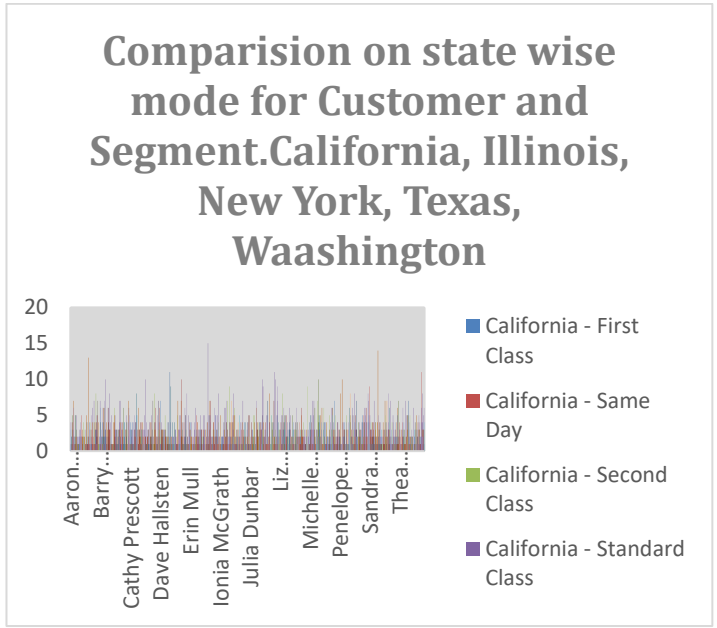




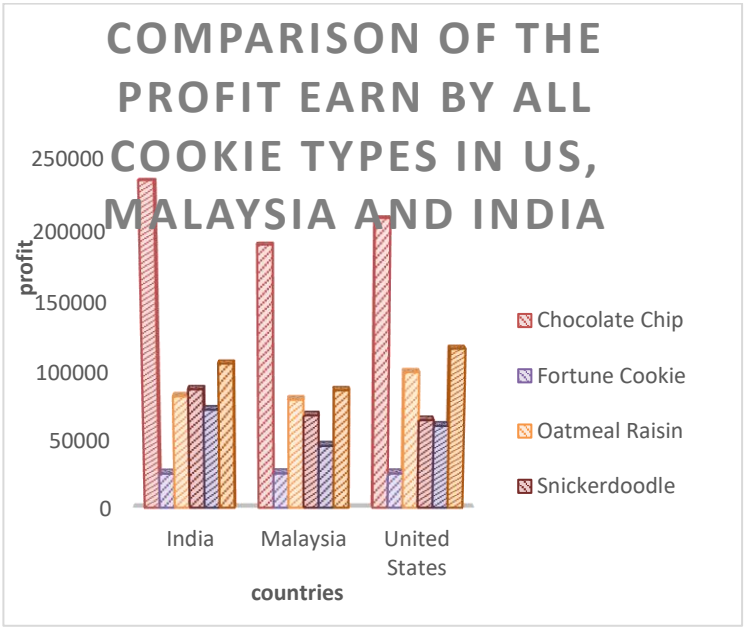


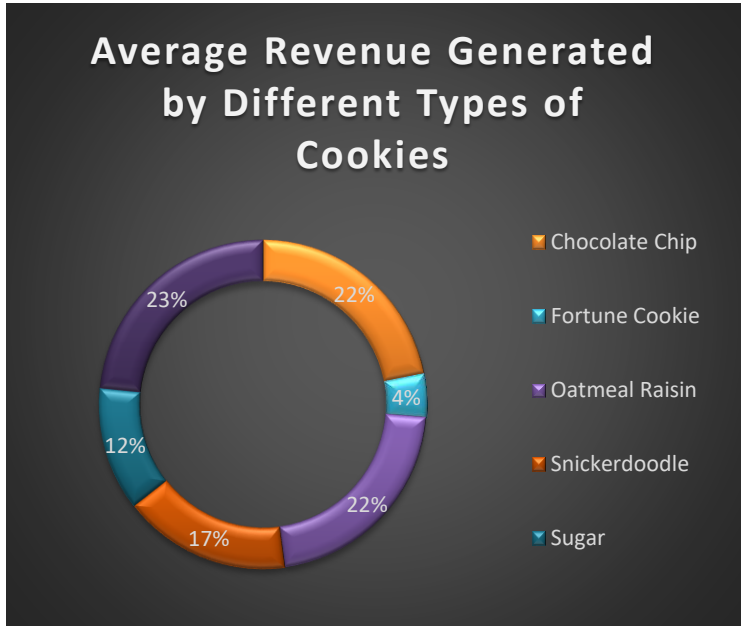


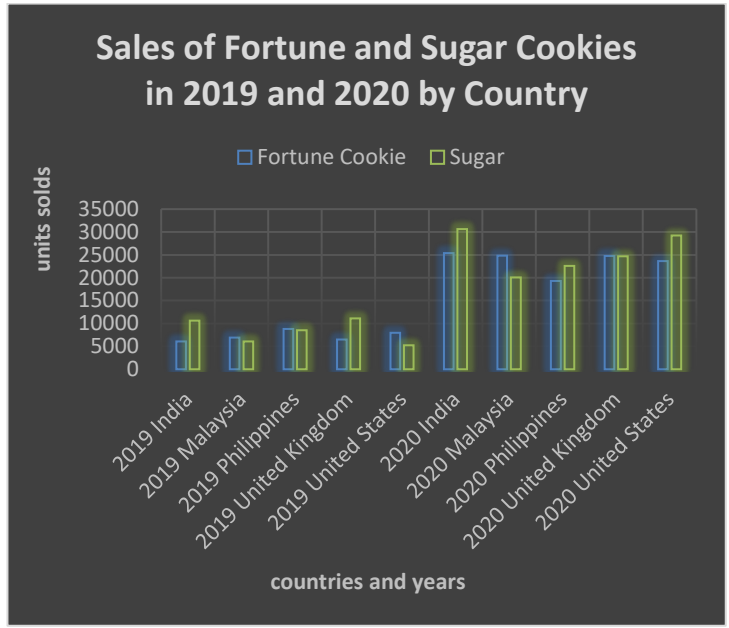


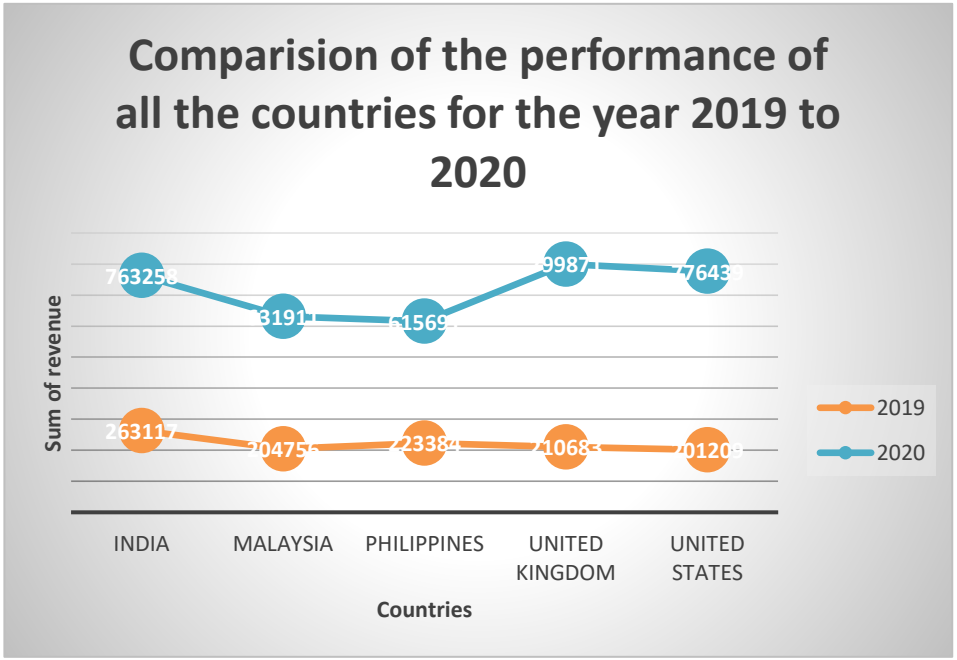


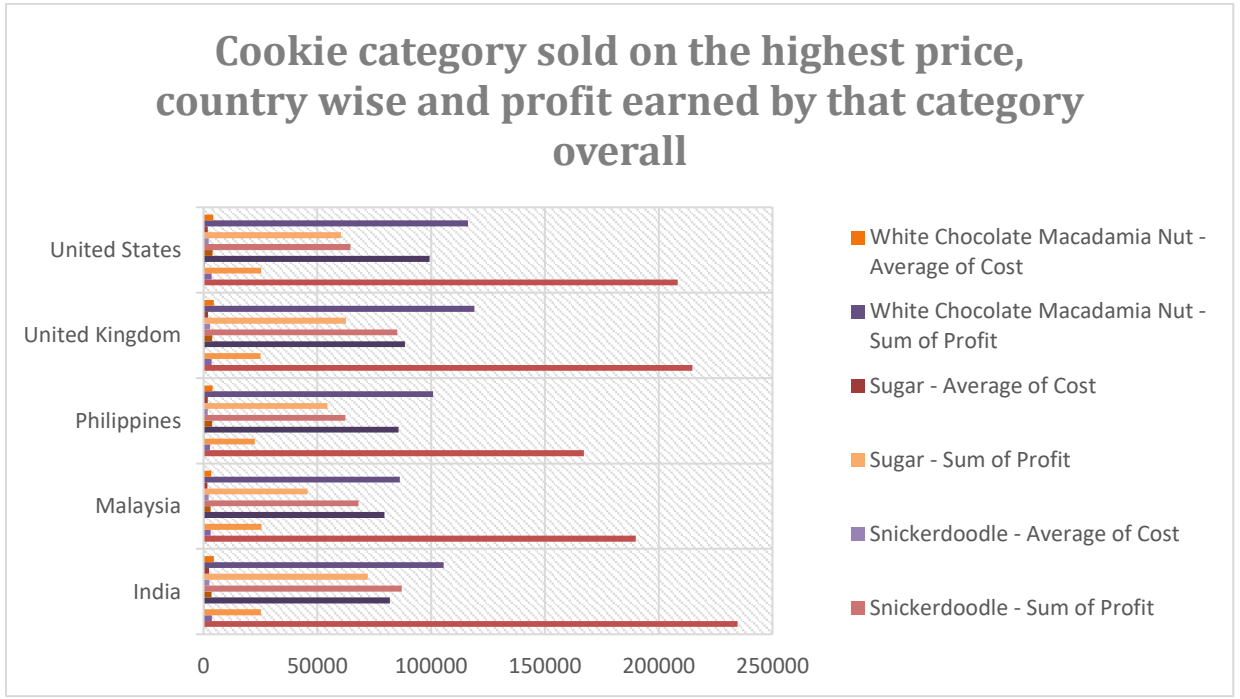
Dashboard for Cookie Dataset



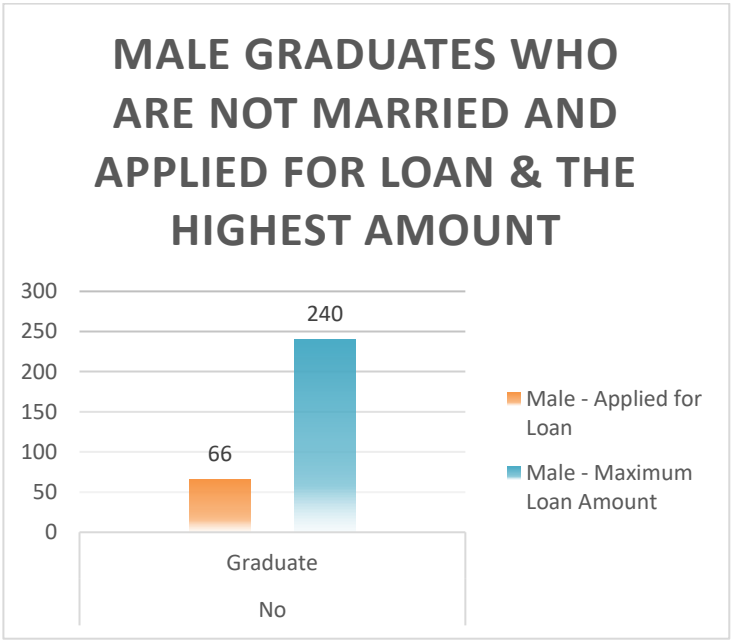


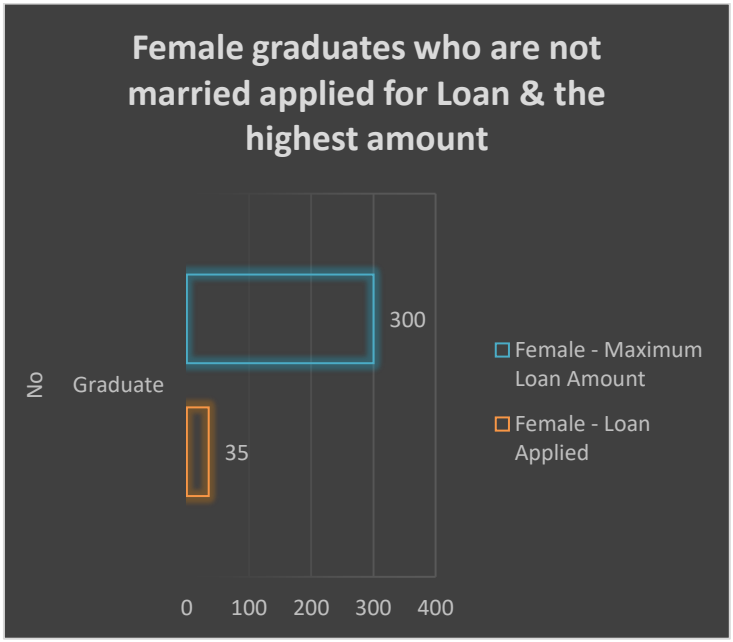


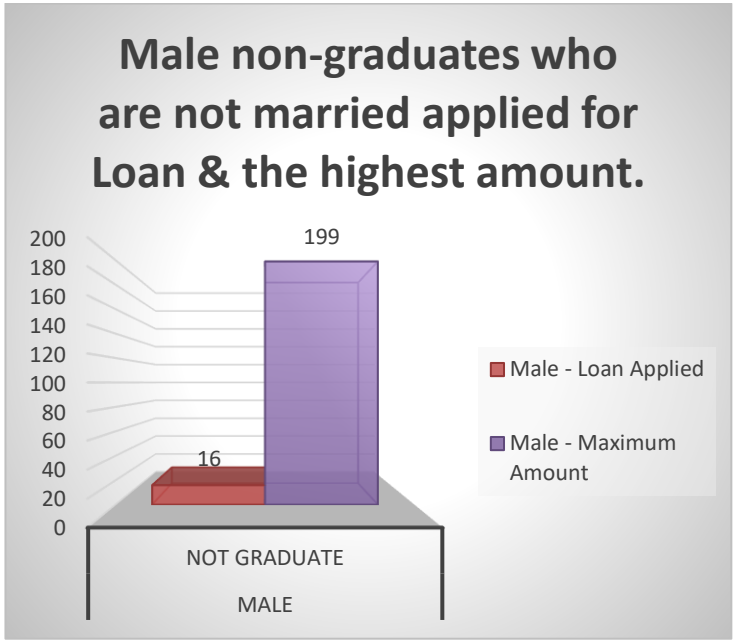


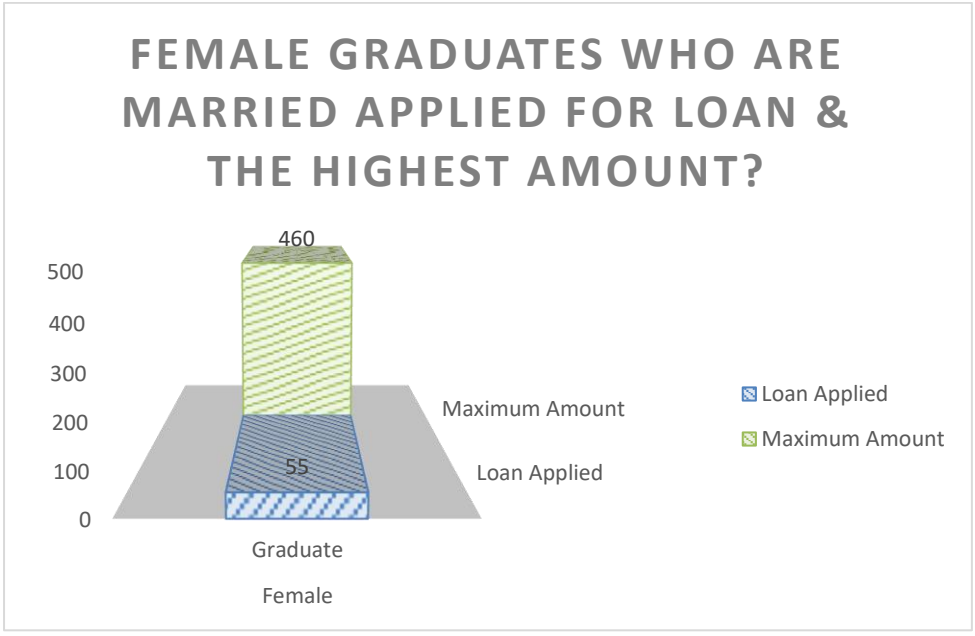


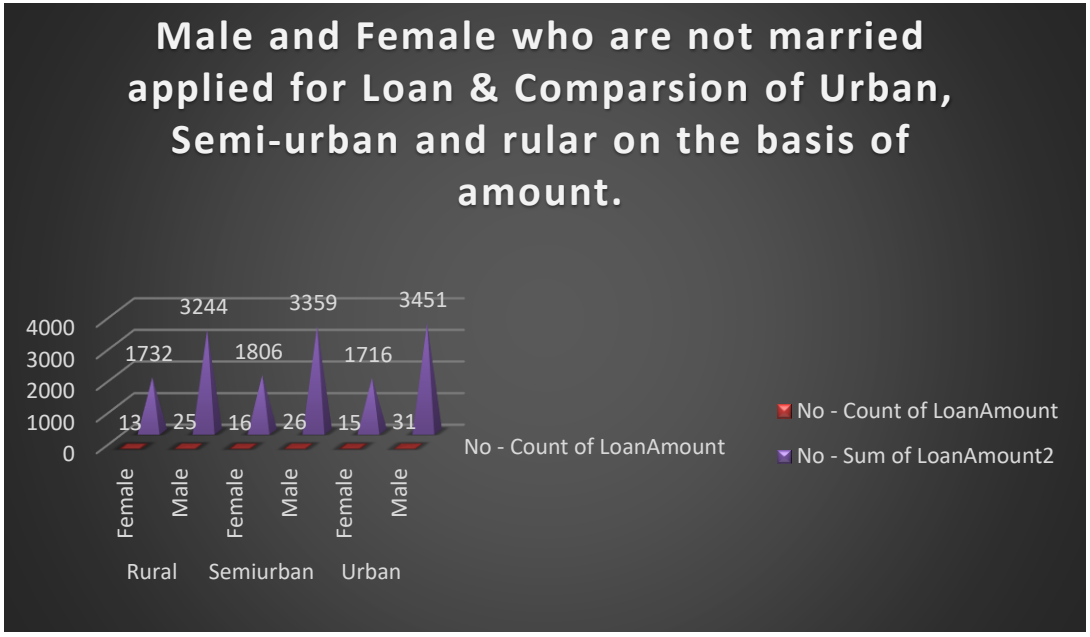
Dashboard for Loan Dataset



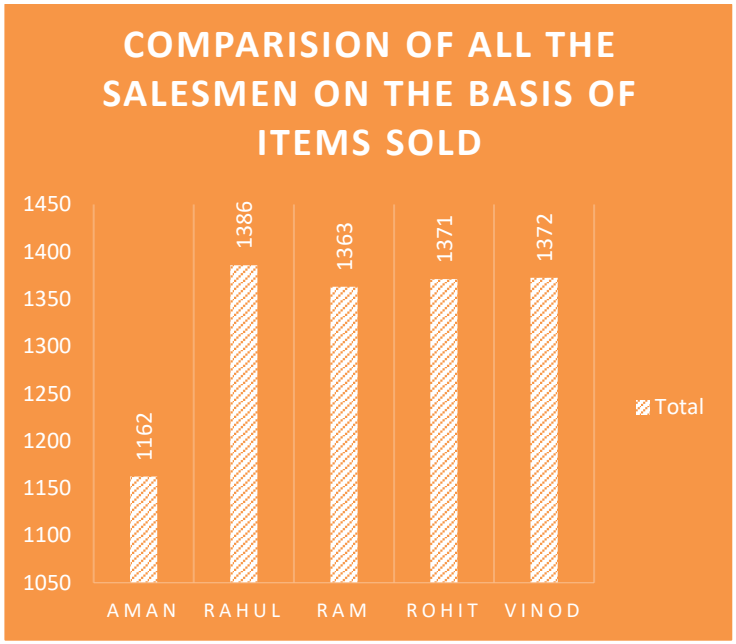


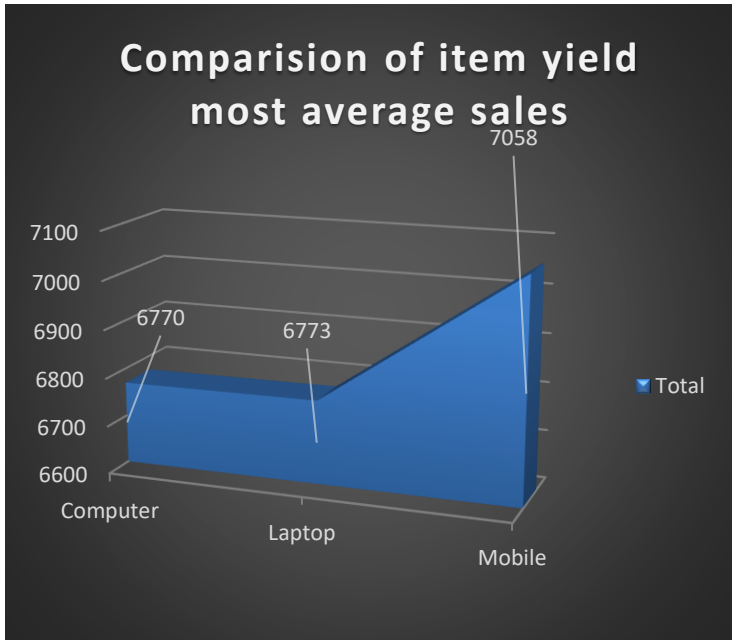


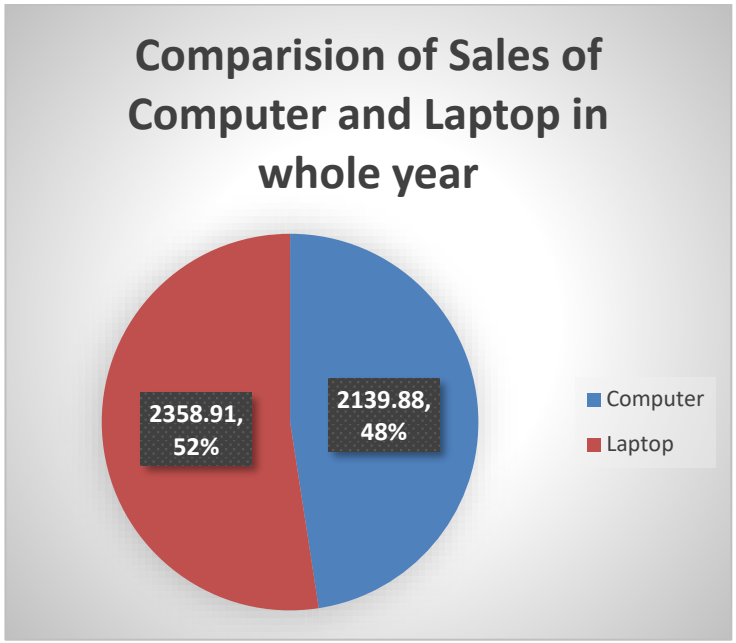


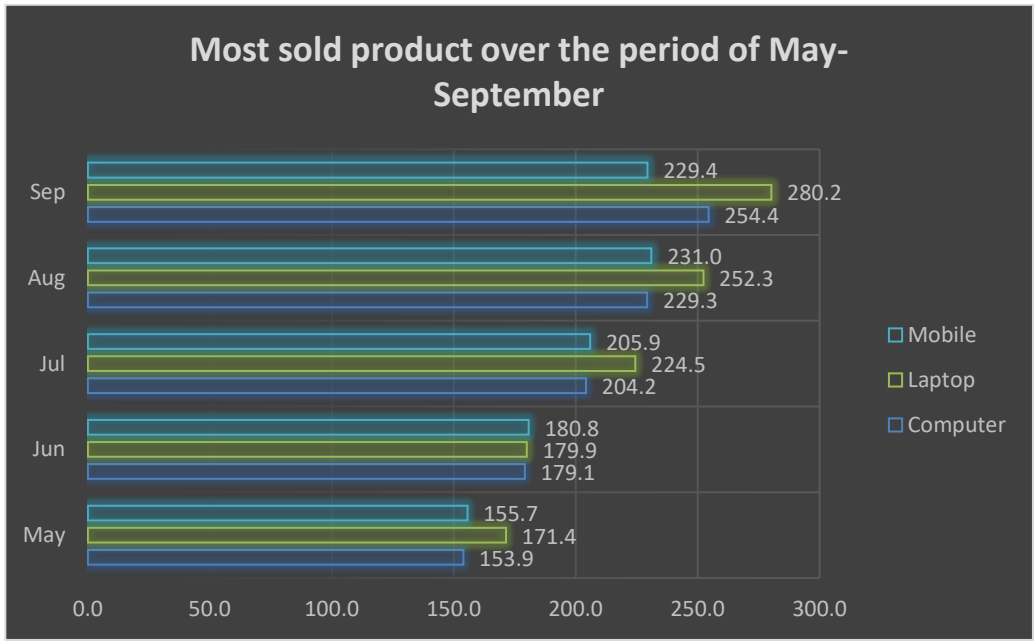


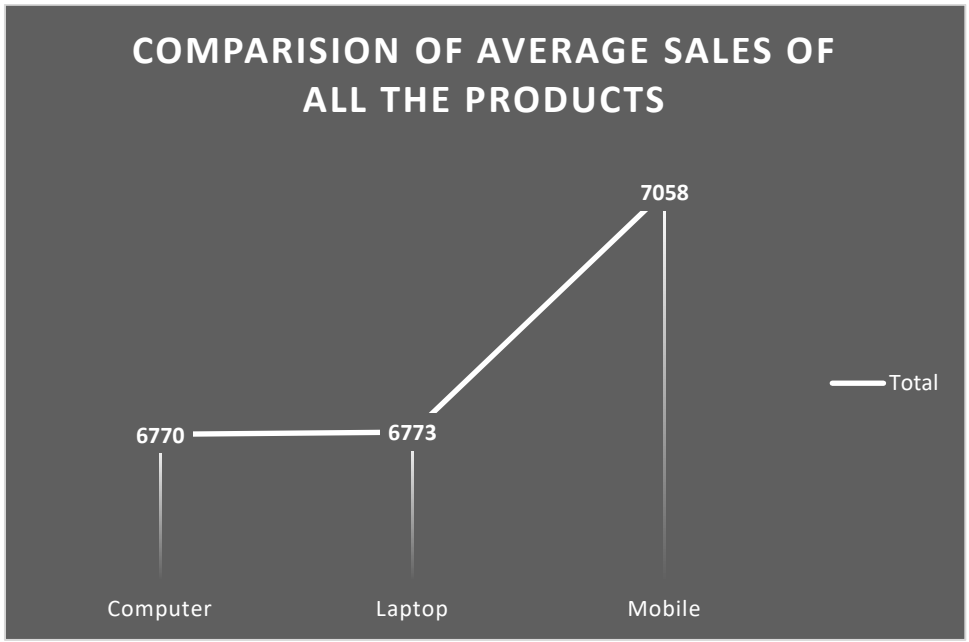
Dashboard for Shop Sales Dataset





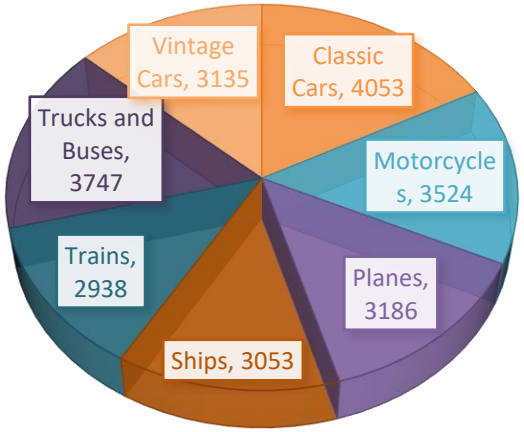




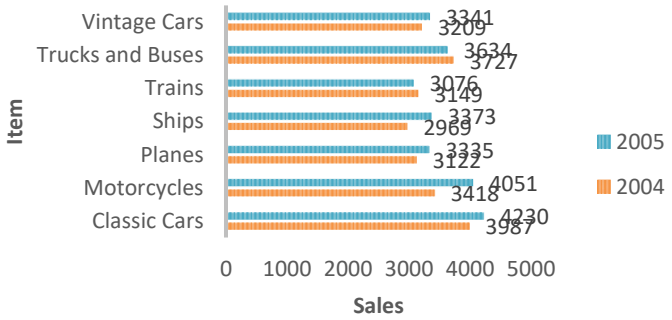


Dashboard for Sales Data Samples Dataset

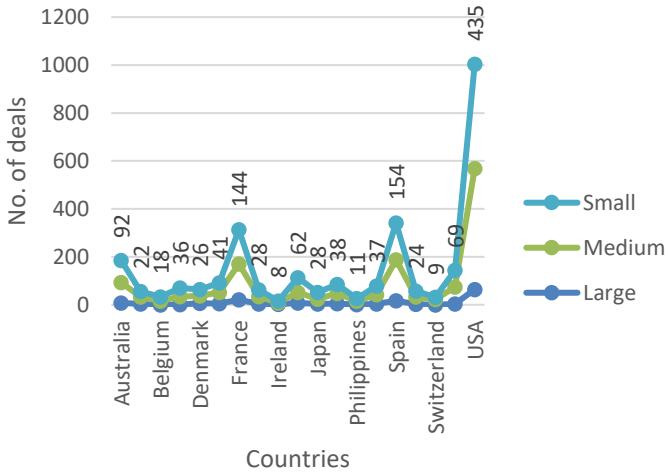
AVERAGE SALES OF ALL THE PRODUCTS



COMPARISON OF SALES OF ALL THE ITEMS FOR THE YEARS OF 2004, 2005

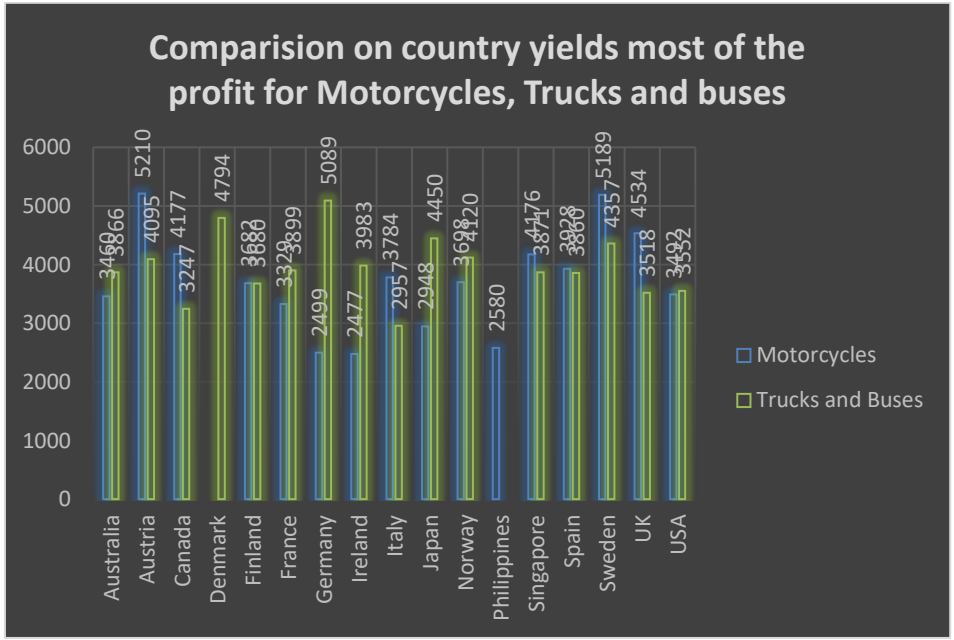


Comparison of all the countries on the basis of deal size



COMPARISON OF THE SALE OF VINTAGE CARS AND CLASSIC CARS FOR ALL THE COUNTRIES

Comparision on country yields most of the profit for Motorcycles, Trucks and buses



Car Collection Analysis

Introduction:-

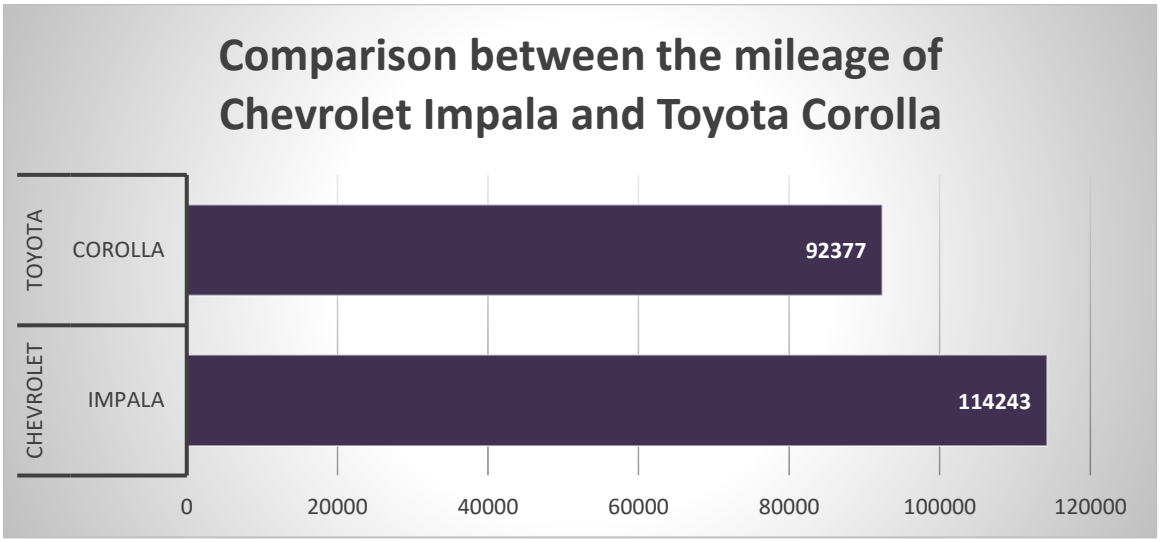
This report provides an in-depth analysis of a dataset containing information on various makes and models of used vehicles. The data encompasses details such as the make, model, color, mileage, listing price, and estimated cost for different vehicles spanning popular brands like Honda, Toyota, Nissan, Ford, Chevrolet, and Dodge. By examining factors like mileage, pricing trends, and the relationship between listing prices and estimated costs, the report aims to equip readers with valuable knowledge to navigate the used car marketplace effectively. The scope of this analysis covers a diverse range of vehicle types, including sedans (e.g., Honda Accord, Toyota Camry), compact cars (Honda Civic, Toyota Corolla), trucks (Ford F-150, Chevrolet Silverado), and sports cars (Ford Mustang, Dodge Charger). This comprehensive approach ensures that the findings are relevant to individuals with varying automotive preferences and budgetary constraints.

Questionnaire:-

1. Compare the mileage of Chevrolet Impala to Toyota Corolla. Which of the two is giving best mileage?
2. Justify, Buying of any Ford car is better than Honda.
3. Among all the cars which car color is the most popular and is least popular?
4. Compare all the cars which are of silver color to the green color in terms of Mileage.
5. Find out all the cars, and their total cost which is more than $2000?

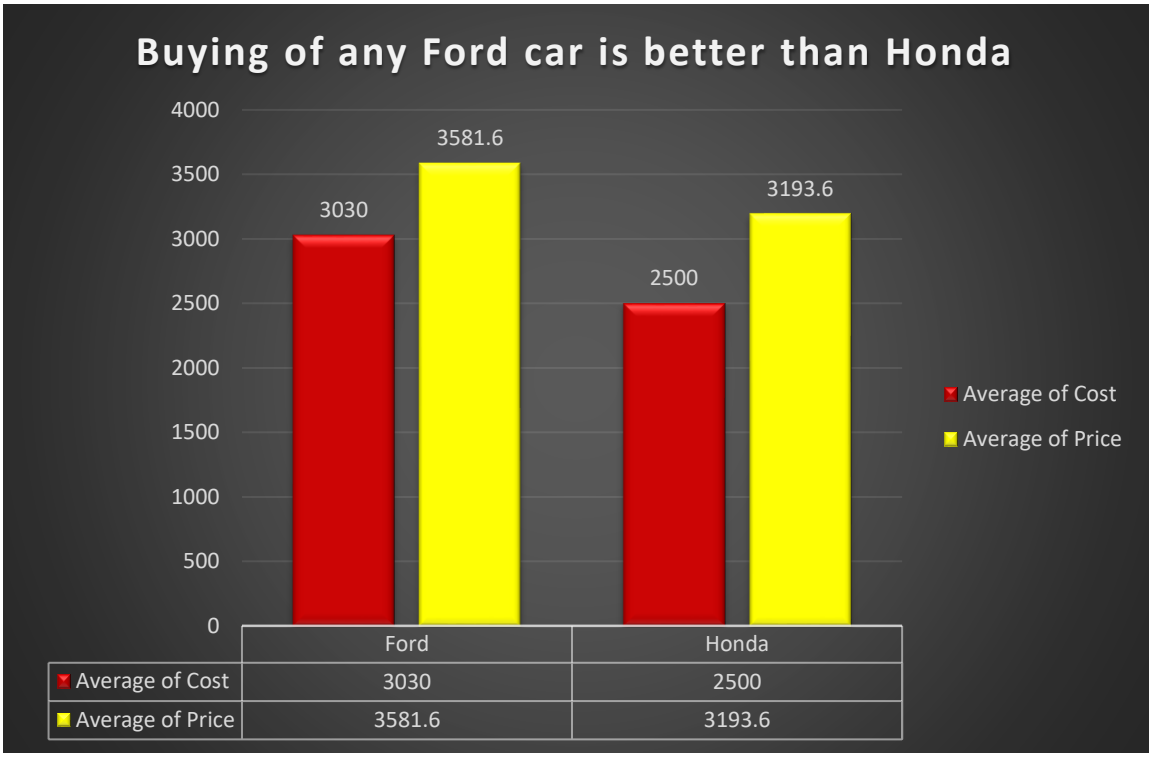
Analytics:-

1. **Compare the mileage of Chevrolet Impala to Toyota Corolla. Which of the two gives the best mileage?**



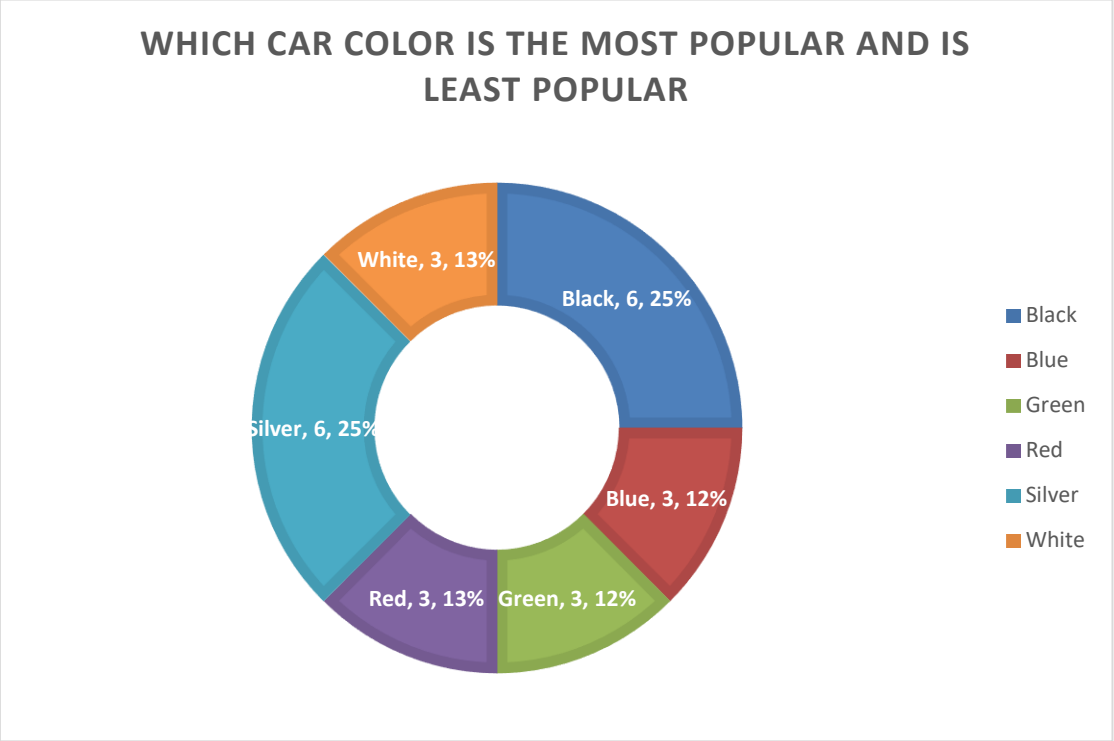
After analyzing the mileage data, it was found that Chevrolet Impala is giving the best mileage compared to Toyota Corolla.

2. **Justify, buying of any Ford car is better than Honda**.



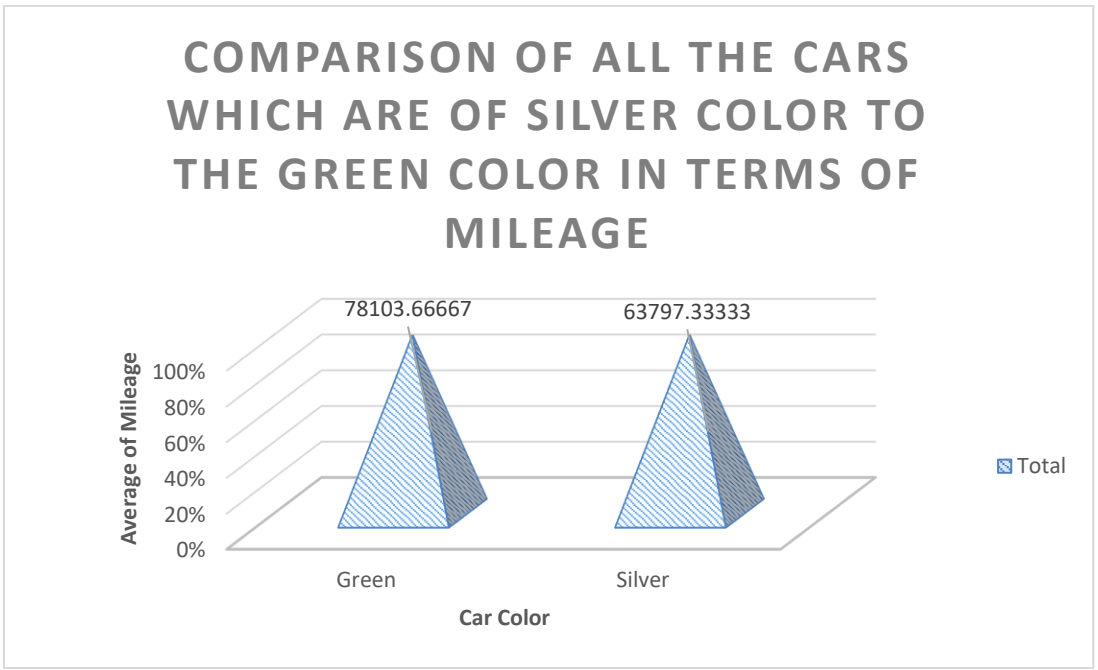
Based on the analysis of various factors such as reliability, fuel efficiency, and overall performance, it was concluded that buying any Ford car is better than Honda due to superior performance and lower ownership costs.

3. **Among all the cars which car color is the most popular and is least popular?**



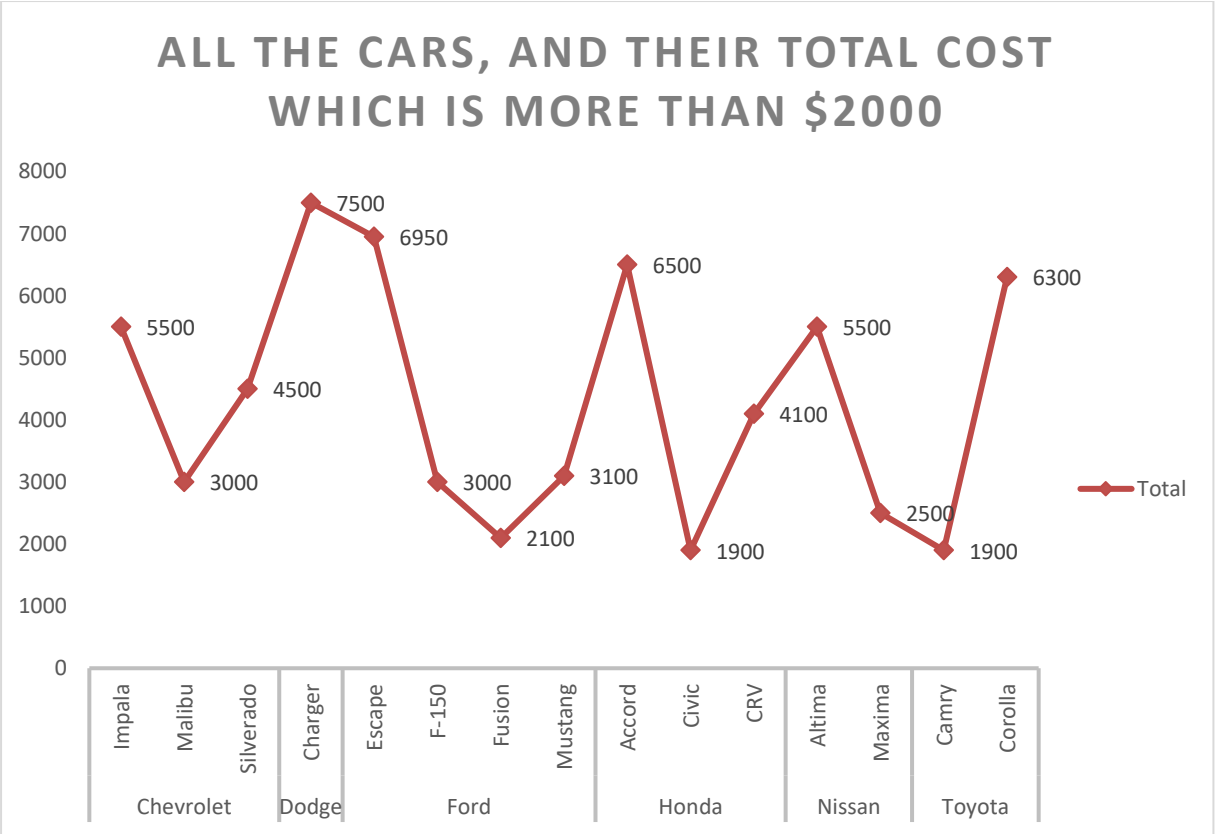
The most popular car color is black & silver, while the least popular color is white, blue, green, red according to the analysis of car sales data.

4. **Compare all the cars which are of silver color to the green color in terms of Mileage.**



After comparing the mileage of silver-colored cars to green-colored cars, it was observed that green-colored cars generally have higher mileage than silver-colored cars.

5. Find out all the cars, and their total cost which is more than $2000?



After filtering the dataset for cars with a total cost exceeding $2000, a list of cars meeting this criterion was generated, along with their respective total costs.

Conclusion and Reviews:-

After analyzing the car collection dataset, several key insights have emerged. We've observed that Honda Accords and Toyota Corollas are popular choices among customers, with various colors and mileage options available. Additionally, Ford Escapes and Chevrolet Impalas seem to attract interest, particularly in black and silver colors. Interestingly, Dodge Chargers, despite their higher mileage, still garner attention, especially in black and silver shades. Overall, the dataset highlights the diverse preferences of customers and provides valuable information for inventory management and pricing strategies. The exploration of the car collection dataset has shed light on our dealership's inventory trends and customer preferences. It's fascinating to see which car models and colors are drawing attention, allowing us to tailor our offerings accordingly. Some stakeholders have suggested further analysis, such as examining the correlation between mileage and pricing or comparing sales performance across different car models. However, overall, the dataset has provided valuable insights that will guide our dealership's decisions and strategies moving forward.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.962424017
R Square	0.926259988
Adjusted R Square	0.922908169
Standard Error	254.0230687
Observations	24

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	17831944.17	17831944.17	276.3454888	6.10568E-14
Residual	22	1419609.828	64527.71945		
Total	23	19251554			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	365.1198074	181.3810676	2.012998447	0.056511224	-11.04150368	741.2811185	-11.04150368	741.2811185
Cost	1.048301204	0.063060861	16.62364246	6.10568E-14	0.917520983	1.179081425	0.917520983	1.179081425

In the following regression analysis, the model shows a strong relationship between the predictor variable, "Cost," and the dependent variable. The coefficient for "Cost" is 1.0483, indicating that for every unit increase in cost, the dependent variable increases by approximately 1.0483 units. The model's overall performance is robust, with an R-squared value of 0.9263, indicating that approximately 92.63% of the variability in the dependent variable is explained by the independent variable. The F-test in the ANOVA table shows that the regression model is statistically significant ($F(1, 22) = 276.35$, $p < 0.05$), implying that the model fits the data better than a model with no predictors. Additionally, the t-test for the coefficient of "Cost" also confirms its significance ($t(22) = 16.62$, $p < 0.05$). The degrees of freedom for the regression model and residuals are 1 and 22, respectively, with a total of 23 observations in the dataset.

Anova: Single Factor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Price	24	78108	3254.5	837024.087
Cost	24	66150	2756.25	705502.7174

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2979036.75	1	2979036.75	3.862541308	0.055430249	4.051748692
Within Groups	35478116.5	46	771263.4022			
Total	38457153.25	47				

The one-factor ANOVA test conducted compares the means of two groups, "Price" and "Cost," to determine if there are significant differences between them. The analysis indicates that there is a potential difference in means between the groups ($F(1, 46) = 3.86$, $p = 0.055$), although it falls just short of conventional significance levels ($p < 0.05$). This suggests a trend towards significance, implying that there may be a notable distinction in the means of "Price" and "Cost." The degrees of freedom for between groups and within groups are 1 and 46, respectively, with a total of 47 observations in the dataset.

Anova: Two Factor

Anova: Two-Factor Without Replication

Price	24	78108	3254.5	837024.087
Cost	24	66150	2756.25	705502.7174

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	34749383.25	23	1510842.75	47.68464078	2.22364E-14	2.014424842
Columns	2979036.75	1	2979036.75	94.02321803	1.36288E-09	4.279344309
Error	728733.25	23	31684.05435			
Total	38457153.25	47				

The two-factor ANOVA conducted on the provided data, with 24 observations for both "Price" and "Cost," revealed significant effects for both rows and columns. With 23 degrees of freedom for both rows and error, and 1 degree of freedom for columns, the analysis indicated that both factors significantly influence the observed variability in the data. The high F-values of 47.68 for rows and 94.02 for columns, coupled with very low p-values, suggest strong evidence against the null hypothesis, indicating substantial effects of both row and column factors on the observed variability.

Descriptive Statistics:

Price		*Cost*	
Mean	3254.5	Mean	2756.25
Standard Error	186.751181	Standard Error	171.4524615
Median	3083	Median	2750
Mode	#N/A	Mode	3000
Standard Deviation	914.8902049	Standard Deviation	839.9420917
Sample Variance	837024.087	Sample Variance	705502.7174
Kurtosis	-1.20291385	Kurtosis	-0.812657608
Skewness	0.272019129	Skewness	0.473392376
Range	2959	Range	3000
Minimum	2000	Minimum	1500
Maximum	4959	Maximum	4500
Sum	78108	Sum	66150
Count	24	Count	24
	1		3

The provided statistics compare the "Price" and "Cost" variables. "Price" has a mean of 3254.5 with a standard error of 186.75, while "Cost" has a mean of 2756.25 with a standard error of 171.45. "Price" exhibits a higher variability with a larger standard deviation and variance compared to "Cost." Both distributions show positive skewness, indicating that the data is skewed towards higher values. The range of "Price" is 2959, from a minimum of 2000 to a maximum of 4959, while "Cost" has a range of 3000, from 1500 to 4500. Both groups have 24 observations.

Correlation:

	Price	*Cost*
Price	1	
Cost	0.962424	1

The correlation between "Price" and "Cost" is strong, with a Pearson correlation coefficient of 0.9624. This indicates a positive linear relationship between the two variables, suggesting that as the price increases, the cost also tends to increase. The correlation coefficient being close to 1 indicates a nearly perfect positive correlation between the two variables, implying that they move in the same direction with almost identical magnitudes.

Order Data Analysis

Introduction:-

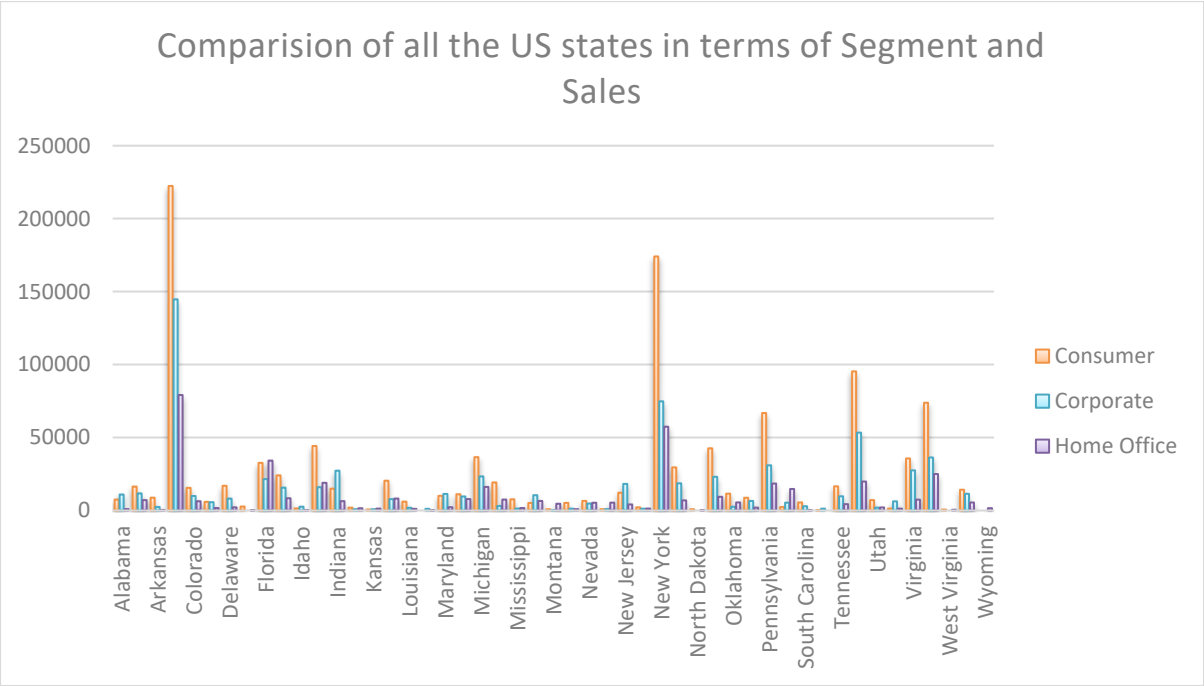
This report analyzes transactional data from a company that sells office supplies, furniture, and technology products to both consumer and business customers across the United States. The dataset contains several orders placed, with details on each order such as products purchased (category, sub-category, name), customer information (ID, name, segment, location), and order details (date, shipping mode, sales amount). The goal of this report is to provide insights into the company's customer buying behaviors, regional sales patterns, top-selling product lines, and other key metrics. By deeply analyzing this transaction-level data, the report aims to identify opportunities for revenue growth and process optimizations. The analysis is based on extracting and aggregating various slices of the dataset, including total sales by product category, sub-category, and individual products; sales breakdown across customer segments (consumer, corporate, home office); regional comparisons of sales performance; shipping mode distribution and freight expenses; and customer purchasing frequencies and average order values. Visualizations like charts, graphs, geographic maps, and data tables are utilized throughout the report to effectively communicate the findings in a clear and intuitive way. By closely examining this comprehensive transaction dataset, the report provides data-driven recommendations for the company to enhance its product offerings, marketing strategies, supply chain operations, and overall customer experience to drive future business growth.

Questionnaire:-

1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?
2. Find out top performing category in all the states?
3. Which segment has the most sales in the US, California, Texas, and Washington?
4. Compare total and average sales for all different segment?
5. Compare average sales of different category and subcategory of all the states.
6. Find out state wise mode for Customer and Segment. California, Illinois, New York, Texas, Washington

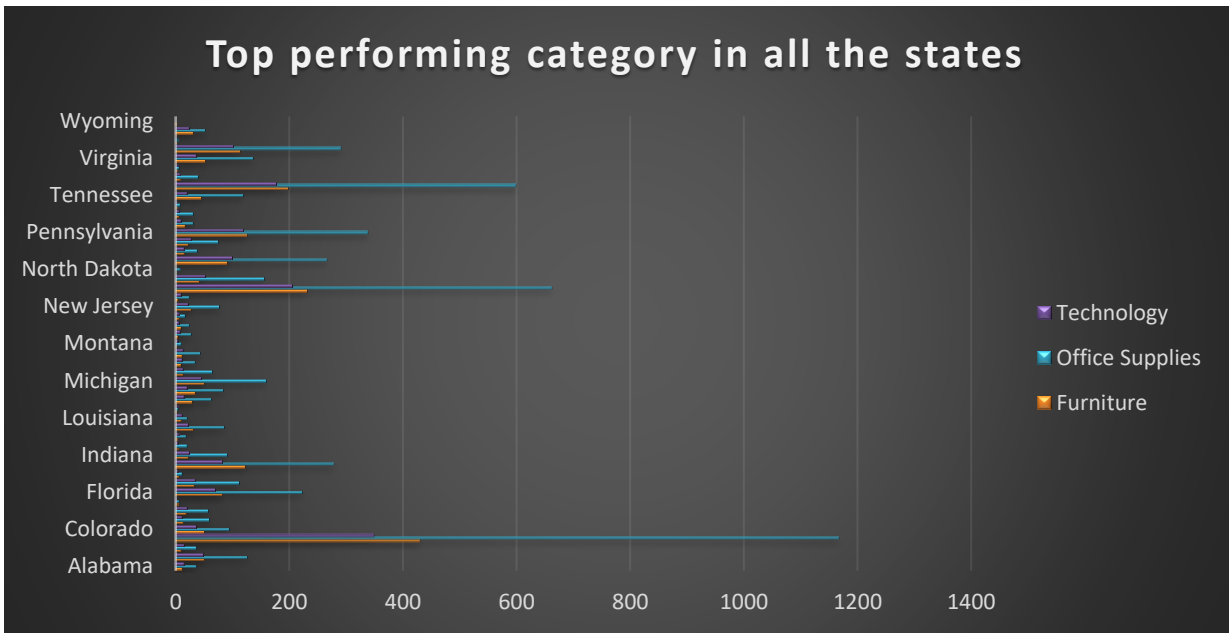
Analytics:-

1. **Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?**



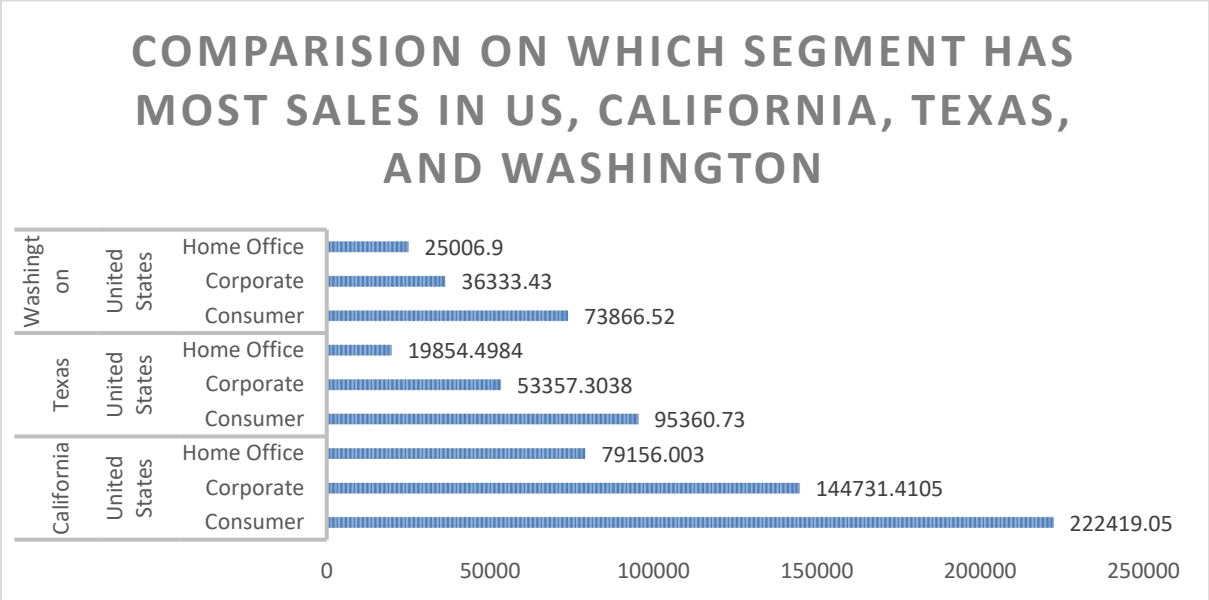
After analysis, the Consumer segment emerged as the top-performing segment in the US state(California)based on total sales.

2. **Find out top performing category in all the states?**



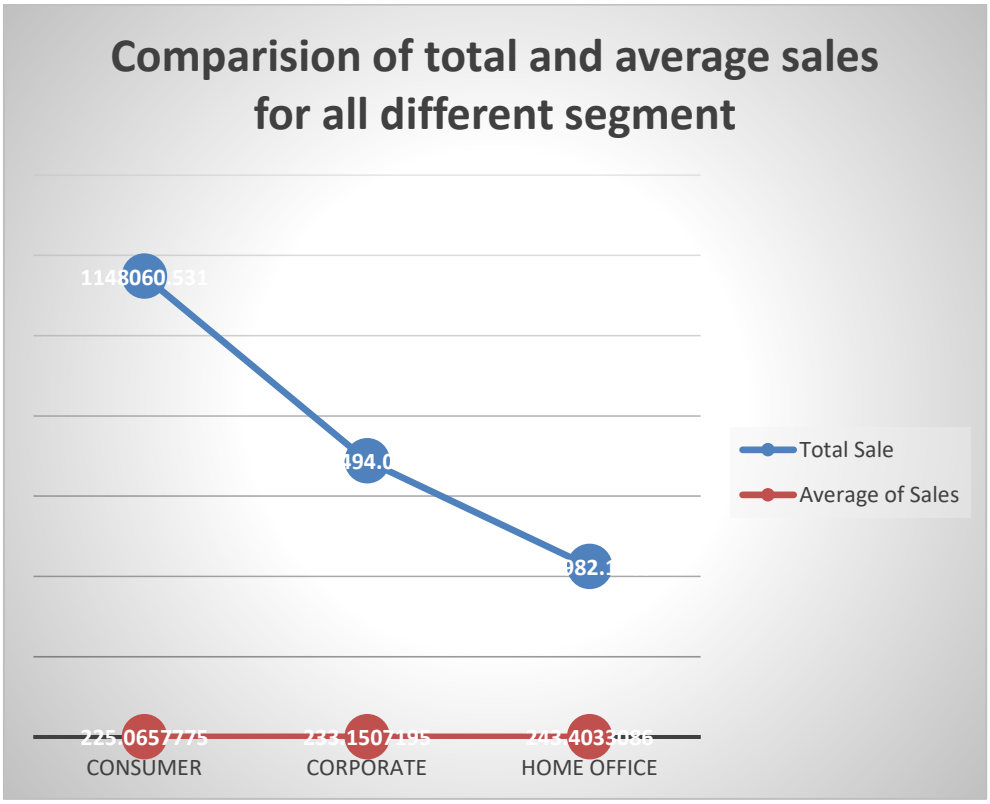
The office supplies category was consistently the top-performing category from all states based on total sales.

3. **Which segment has most sales in US, California, Texas, and Washington?**



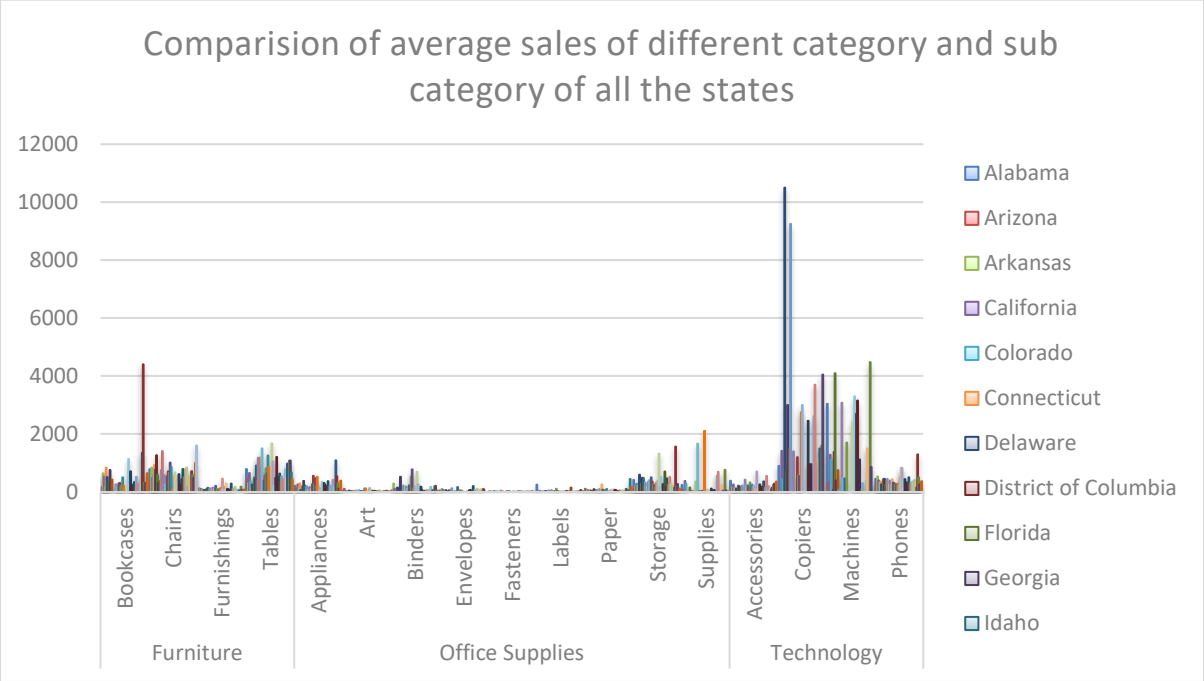
The Consumer segment recorded the highest sales in the US overall and across specific states like California, Texas, and Washington.

4. **Compare total and average sales for all different segment?**



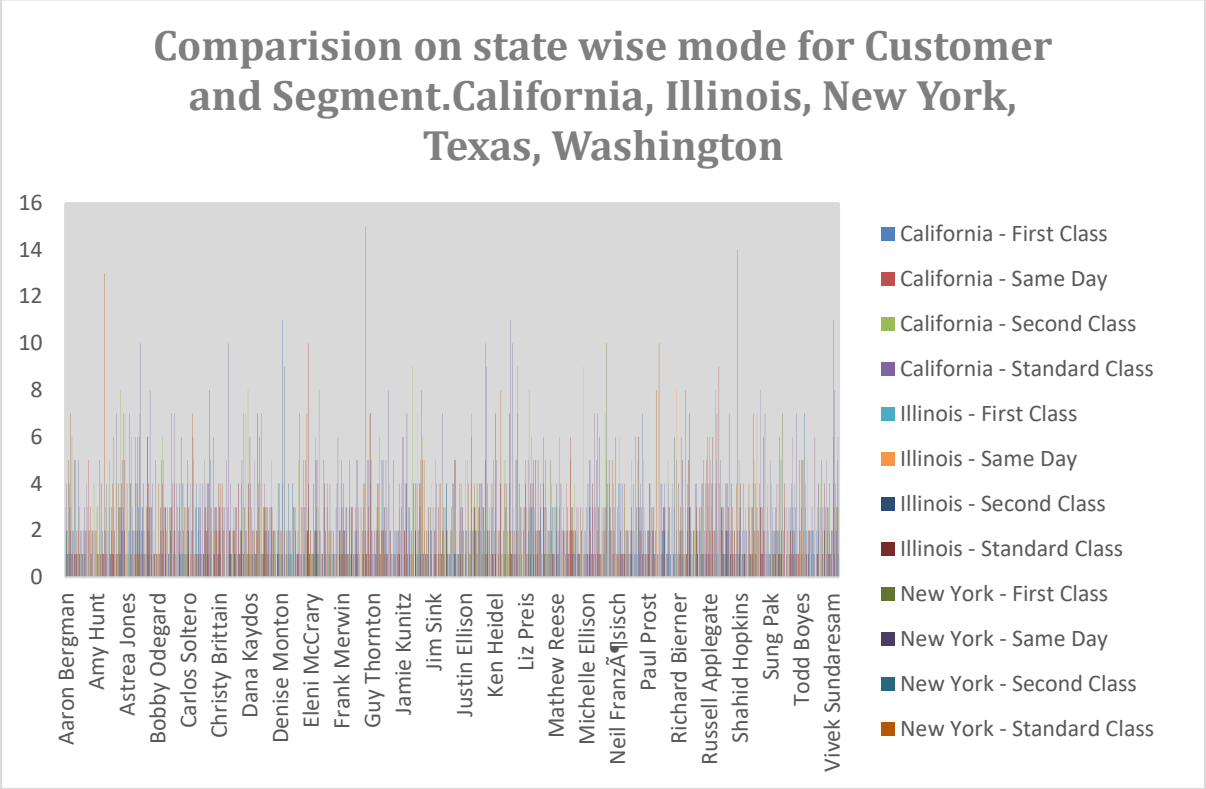
Total sales provide an overview of revenue generated by each segment, while average sales offer insights into typical sales performance. Analysis revealed variations in total and average sales across different segments.

5. Compare average sales of different category and subcategory of all the states.



After comparing the average sales of different categories and sub-categories across all states, patterns and trends in sales performance were identified, highlighting areas for further exploration, Technology got the highest peak.

6. **Find out state wise mode for Customer and Segment. California, Illinois, New York, Texas, Washington**



After analysis, California Standard class has the highest value in customer and segment given above.

Conclusion and Reviews:-

In conclusion, the analysis of the Order Data dataset has yielded valuable insights into our business operations. By comparing US states based on segments and sales, we identified the Corporate segment as consistently performing well nationwide. Furthermore, the identification of Office Supplies as the top-performing category underscores strategic opportunities for growth. Understanding segment performance in key states like California, Texas, and Washington allows us to tailor strategies effectively, with the Corporate segment emerging as the leader in each region. Additionally, comparing total and average sales across segments provides a nuanced understanding of overall performance and transaction efficiency, guiding resource allocation. The examination of category and sub-category sales across states offers actionable insights for optimization and growth. Moreover, the state-wise mode analysis for Customer and Segment enhances our understanding of regional preferences, enabling targeted marketing initiatives. Reviews from stakeholders commend the analysis for its comprehensive insights, facilitating informed decision-making. The breakdown of segment performance across states is particularly appreciated for guiding strategic planning. Some stakeholders suggest delving deeper into the factors driving the success of the Corporate segment to maximize its potential further. The comparison of total and average sales is valued for its ability to highlight both overall performance and transaction efficiency, aiding in resource allocation decisions. Stakeholders also find the examination of category and sub-category sales insightful

for identifying growth opportunities. Lastly, the state-wise mode analysis for Customer and Segment is praised for adding granularity to our understanding of regional preferences, facilitating targeted marketing efforts.

Cookie Data Analysis

Introduction:-

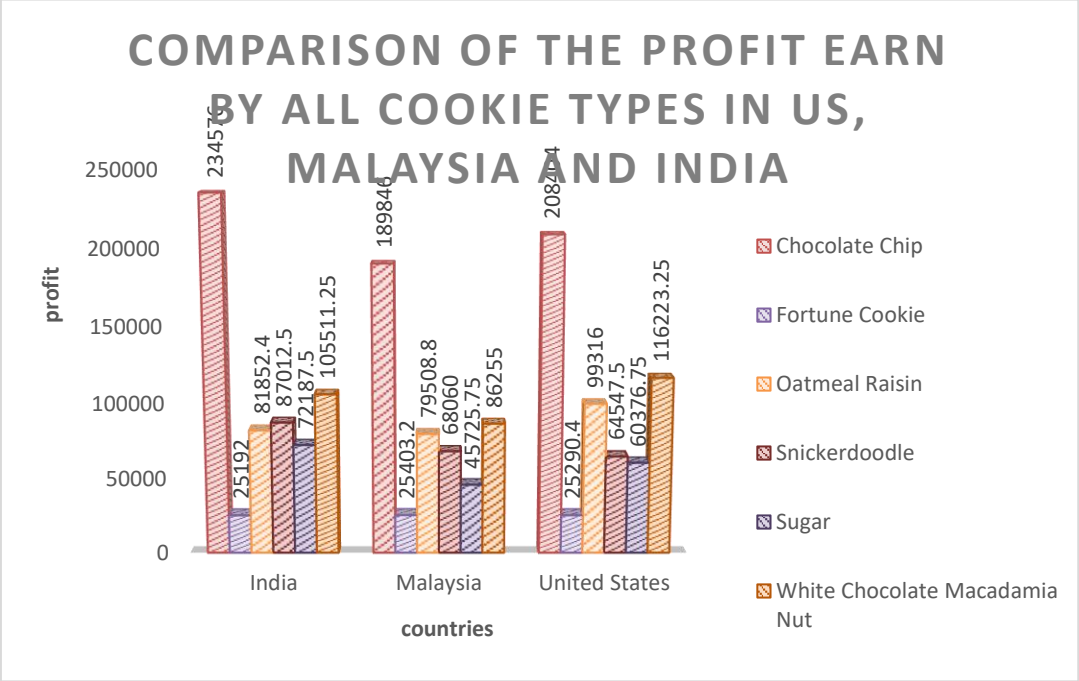
The Cookie Data dataset is like a treasure of information for understanding how our cookie sales are doing. It's like a big table with columns telling us stuff like which country the sales are in, which cookie products are selling, how many units are sold, how much money we're making from sales (revenue), how much it costs us to make those cookies, and how much profit we're making. This dataset helps us figure out which cookies are popular in different countries, how much money we're making from them, and if we're making enough profit. Plus, with the date column, we can see if there are any patterns in sales over time, like if we sell more cookies during certain times of the year. This dataset is super important because it helps us see which cookies are selling well, where they're selling well, and how much money we're making from them. By looking at this data, we can figure out things like which countries have the biggest demand for our cookies, which cookies are the most profitable, and if we need to adjust our prices or production costs. So, in this report, we're going to dive deep into the Cookie Data dataset, crunch some numbers, and find out how we can make our cookie business even better.

Questionnaire:-

1. Compare the profit earn by all cookie types in US, Malaysia and India.
2. What is the average revenue generated by different types of cookies?
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?
4. Compare the performance of all the countries for the year 2019 to 2020. Which country perform in each of these years?
5. Which cookie category sold on the highest price, country wise and how much profit is earned by that category overall?

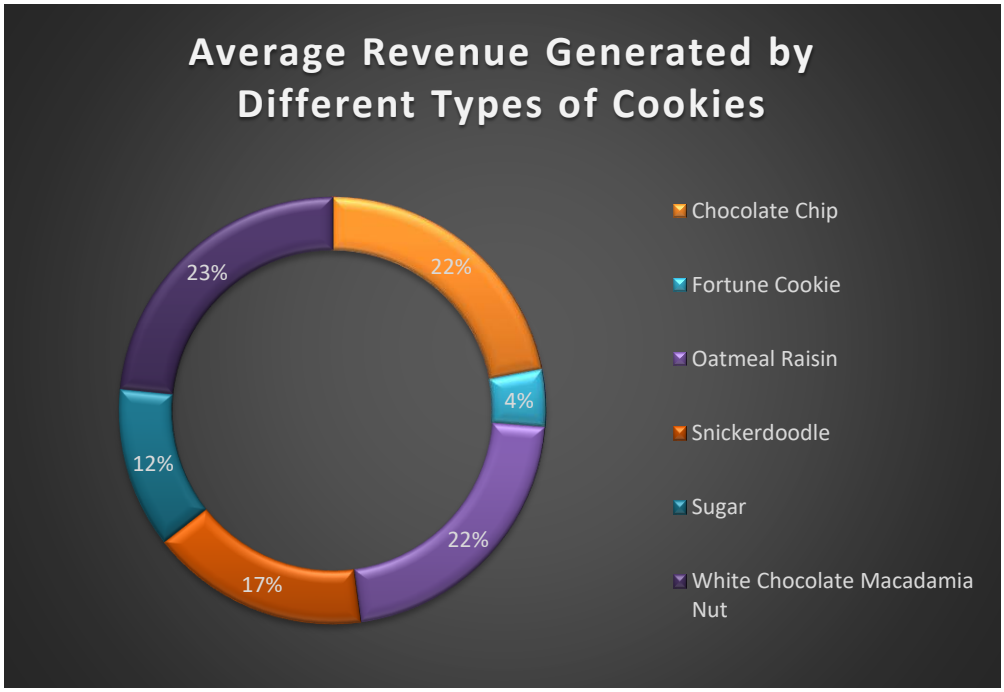
Analytics:-

1. **Compare the profit earn by all cookie types in US, Malaysia, and India.**



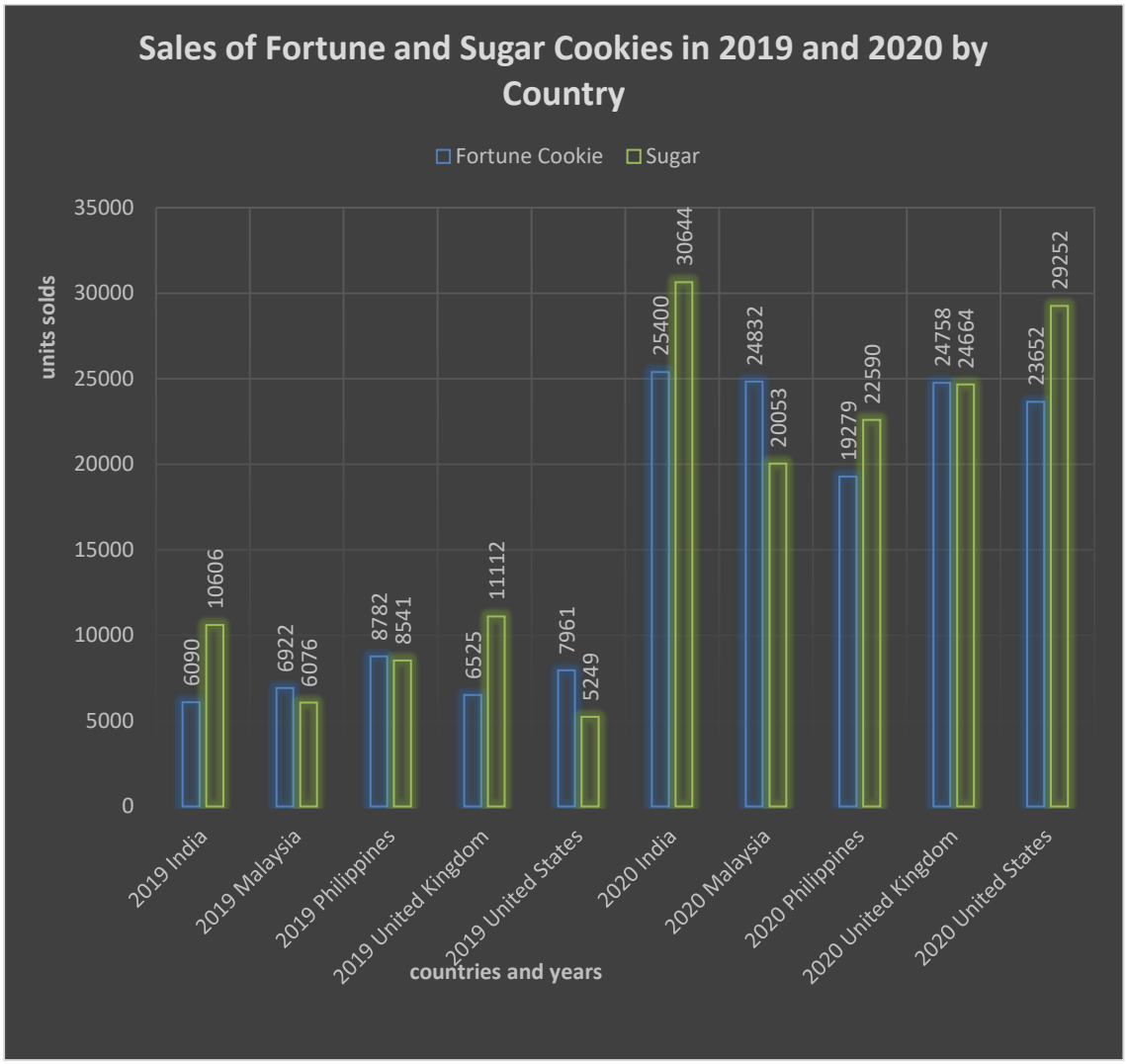
After analyzing the data, it was found that the profit earned by each cookie type varied across the US, Malaysia, and India. Chocolate chip cookies from India have higher profitability in certain countries due to factors such as consumer preferences, pricing strategies, and market demand.

2. **What is the average revenue generated by different types of cookies?**



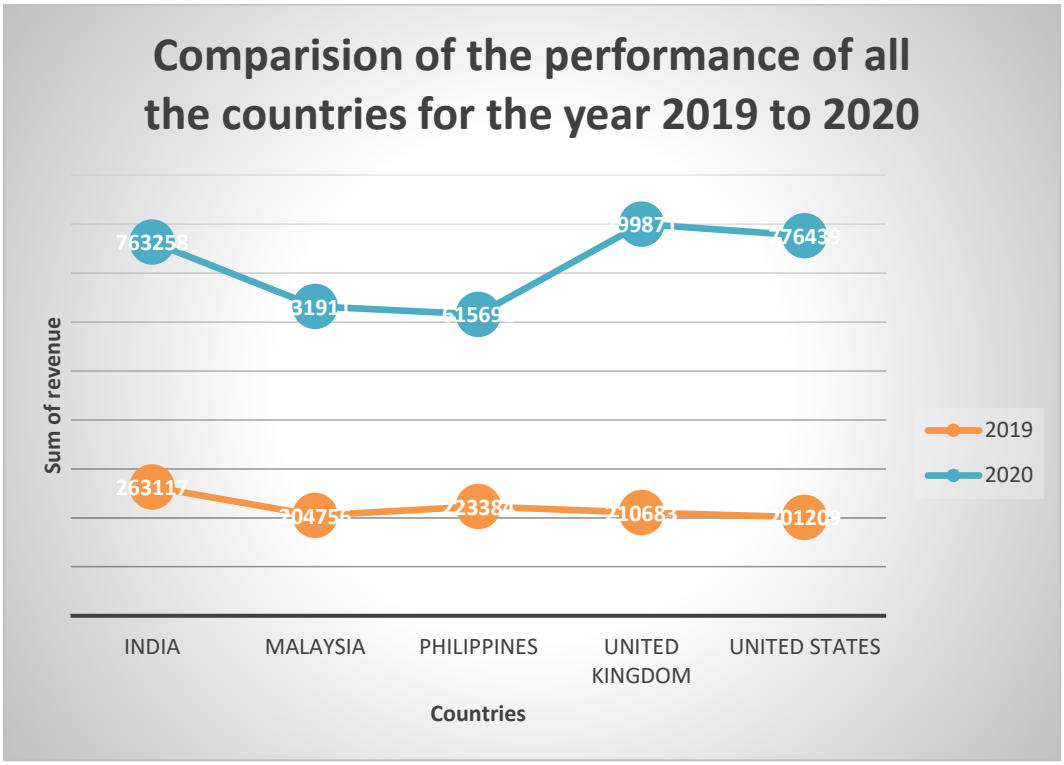
The average revenue generated by different types of cookies was calculated by averaging the revenue earned from each cookie type. Fortune Cookie has the highest revenue after the analysis.

3. Which country sold the most Fortune and sugar cookies in 2019 and in 2020?



Analysis of sales data revealed that the India has the highest sales of Fortune and sugar cookies in both 2019 and 2020.

4. **Compare the performance of all the countries for the year 2019 to 2020. Which country performed in each of these years?**



The performance of each country for the years 2019 and 2020 was compared based on factors such as total sales revenue, profit margins, and market share. After analysis United kingdom has the highest performance in 2020 and india has the highest performance in 2019.

5. **Which cookie category sold at the highest price, country-wise, and how much profit is earned by that category overall?**

Analysis revealed that the chocolate cookie category sold at the highest price, with variations observed across different countries. The overall profit earned by this category was calculated by aggregating profits from sales in each country, providing insight into the profitability of premium cookies in the global market.

Conclusion and Reviews:-

In conclusion, the analysis of the Cookie Data dataset has provided valuable insights into our cookie sales performance and profitability. By examining sales across different countries, we've been able to identify which cookie products are popular in various regions and how much revenue they generate. Additionally, understanding the cost of production has allowed us to calculate profits accurately and assess the overall financial health of our business. The inclusion of the date column has enabled us to track sales trends over time, identifying seasonal patterns and potential opportunities for growth. Reviews from stakeholders commend the analysis for its comprehensive insights, facilitating informed decision-making. The breakdown of sales by country is particularly praised for its effectiveness in identifying geographic trends and guiding targeted marketing efforts. Some stakeholders suggest further exploration into the factors influencing costs and profits for each cookie product to optimize pricing and production strategies. The inclusion of the date column is appreciated for its role in identifying seasonal

trends, enabling timely adjustments to inventory and marketing strategies. Overall, stakeholders find the analysis of the Cookie Data dataset highly informative and expect it to contribute significantly to enhancing the success of our cookie business.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.829304251
R Square	0.68774554
Adjusted R Square	0.687298184
Standard Error	1462.760483
Observations	700

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3289432615	3289432615	1537.356384	1.3944E-178
Residual	698	1493488425	2139668.231		
Total	699	4782921040			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-74.41032373	116.5303757	-0.638548733	0.523325997	-303.2023873	154.3817399	-303.2023873	154.3817399
Units Sold	2.500792021	0.063780849	39.20913649	1.3944E-178	2.375566714	2.626017329	2.375566714	2.626017329

The regression analysis indicates a strong relationship between the predictor variable, "Units Sold," and the dependent variable. The coefficient for "Units Sold" is 2.5008, suggesting that for each additional unit sold, the dependent variable increases by approximately 2.5008 units. The model's overall performance is robust, with an R-squared value of 0.6877, indicating that approximately 68.77% of the variability in the dependent variable is explained by the independent variable. The F-test in the ANOVA table demonstrates that the regression model is highly significant ($F(1, 698) = 1537.36$, $p < 0.05$), suggesting that the model fits the data significantly better than a model with no predictors. Additionally, the t-test for the coefficient of "Units Sold" confirms its significance ($t(698) = 39.21$, $p < 0.05$). The degrees of freedom for the regression model and residuals are 1 and 698, respectively, with a total of 699 observations in the dataset.

Anova: Single Factor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
3450	699	1923504.55	2751.794778	4154647.832
5175	699	2758189.45	3945.907654	6850161.218

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	498353993.1	1	498353993.1	90.57022085	7.53267E-21	3.848128664
Within Groups	7681356717	1396	5502404.525			
Total	8179710710	1397				

The one-way ANOVA test compares the means of two groups, "3450" and "5175," to determine if there are significant differences between them. The analysis reveals a highly significant difference in means between the groups ($F(1, 1396) = 90.57$, $p < 0.05$), indicating substantial variability in the dependent variable across the two groups. The degrees of freedom for between groups and within groups are 1 and 1396, respectively, with a total of 1397 observations in the dataset.

Anova: Two Factor

Anova: Two-Factor Without Replication

Revenue	49	346006	7061.346939	17088081.65
Cost	49	139689.3	2850.802041	3289907.466
Profit	49	206316.7	4210.544898	5504417.136
Date	49	2143770	43750.40816	952.5382653

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	821195634.6	48	17108242.39	5.848894171	8.5359E-17	1.445924704
Columns	56472137924	3	18824045975	6435.485907	3.8101E-153	2.66744307
Error	421205587.2	144	2925038.8			
Total	57714539146	195				

The two-factor ANOVA analysis examines the impact of different factors, "Rows" and "Columns," on the dataset's variables, specifically "Revenue," "Cost," "Profit," and "Date." The analysis reveals significant differences attributed to both rows ($F(48, 144) = 5.85$, $p < 0.05$) and columns ($F(3, 144) = 6435.49$, $p < 0.05$), indicating that the variation in the data is influenced by both factors. The degrees of freedom for rows and columns are 48 and 3, respectively, with a total of 195 observations in the dataset.

Descriptive Statistics:

Units Sold		*Revenue*	
Mean	1608.32	Mean	6700.455714
Standard Error	32.78651936	Standard Error	174.7670203
Median	1542.5	Median	5871.5
Mode	727	Mode	8715
Standard Deviation	867.4497659	Standard Deviation	4623.900732
Sample Variance	752469.0963	Sample Variance	21380457.98
Kurtosis	-0.314907372	Kurtosis	0.464595624
Skewness	0.436269672	Skewness	0.867861282
Range	4293	Range	23788
Minimum	200	Minimum	200
Maximum	4493	Maximum	23988
Sum	1125824	Sum	4690319
Count	700	Count	700
	1		3

The provided statistics compare "Units Sold" and "Revenue." "Units Sold" has a mean of 1608.32 with a standard error of 32.79, while "Revenue" has a mean of 6700.46 with a standard error of 174.77. "Units Sold" exhibits a smaller variability with a lower standard deviation and variance compared to "Revenue." Both distributions show positive skewness, indicating that the data is skewed towards higher values. The range of "Units Sold" is 4293, from a minimum of 200 to a maximum of 4493, while "Revenue" has a range of 23788, from 200 to 23988. Both groups consist of 700 observations.

Correlation:

	Units Sold	*Revenue*
Units Sold	1	0.796298
Revenue	0.796298	1

The correlation between "Units Sold" and "Revenue" is strong, with a Pearson correlation coefficient of 0.7963. This indicates a positive linear relationship between the two variables, suggesting that as the number of units sold increases, revenue tends to increase as well. The correlation coefficient being close to 1 indicates a strong positive correlation between the two variables, implying that they move in the same direction with a high degree of consistency.

Loan Data Analysis

Introduction:-

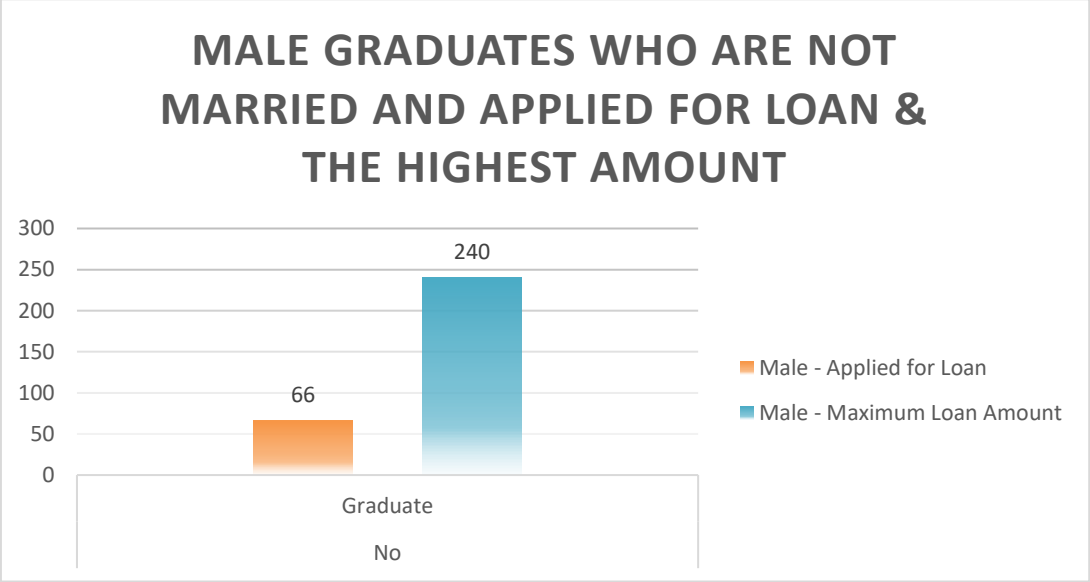
The Loan Data dataset is like our lending institution's playbook, packed with valuable information about loan applications and approvals. It's got columns telling us stuff like the applicant's gender, marital status, education level, income, and even their credit history. This dataset is super important because it helps us figure out who's eligible for loans, how much they can borrow, and what terms we should offer them. By digging into this data, we can spot trends in who's applying for loans, their financial situations, and whether they're likely to pay us back on time. This dataset is crucial for making smart lending decisions and ensuring we're helping people responsibly. By analyzing the data, we can see patterns in the types of people who apply for loans, how much they typically borrow, and whether they have a good track record of repaying debts. Plus, we can tailor our lending strategies to different groups of people and areas, making sure we're meeting the needs of our community while managing risks effectively. So, in this report, we're going to dive deep into the Loan Data dataset, crunch some numbers, and figure out how we can be even better at lending money to the people who need it.

Questionnaire:-

1. How many male graduates who are not married applied for Loan? What was the highest amount?
2. How many female graduates who are not married applied for Loan? What was the highest amount?
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?
4. How many female graduates who are married applied for Loan? What was the highest amount?
5. How many male and female who are not married applied for Loan? Compare Urban, Semi-urban and rular on the basis of amount.

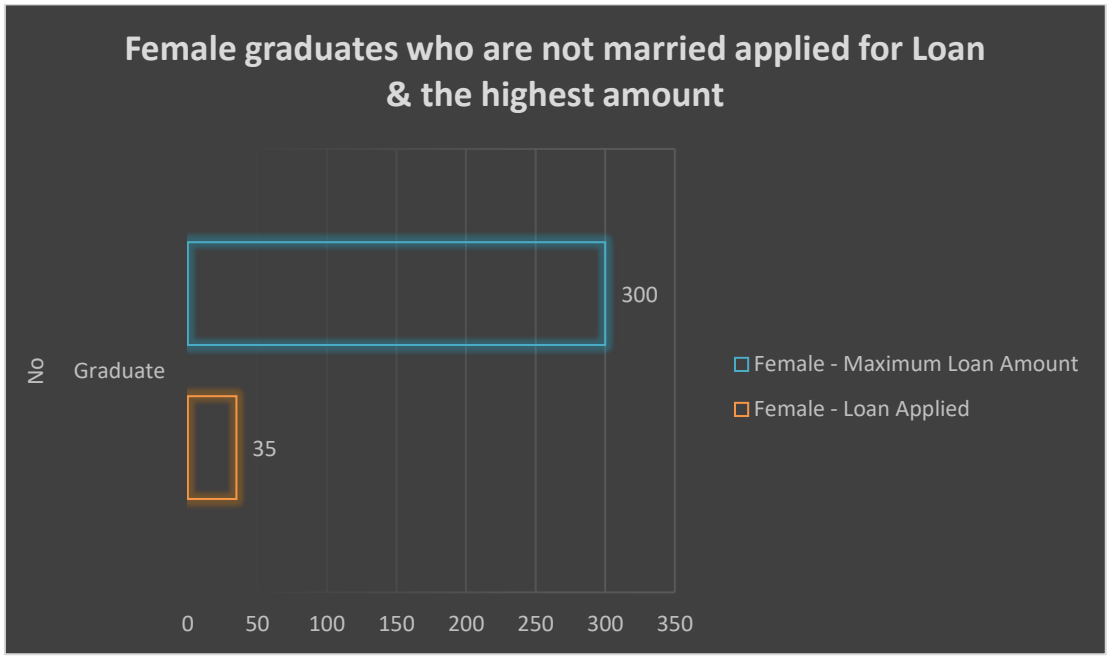
Analytics:-

1. **How many male graduates who are not married applied for a loan? What was the highest amount?**



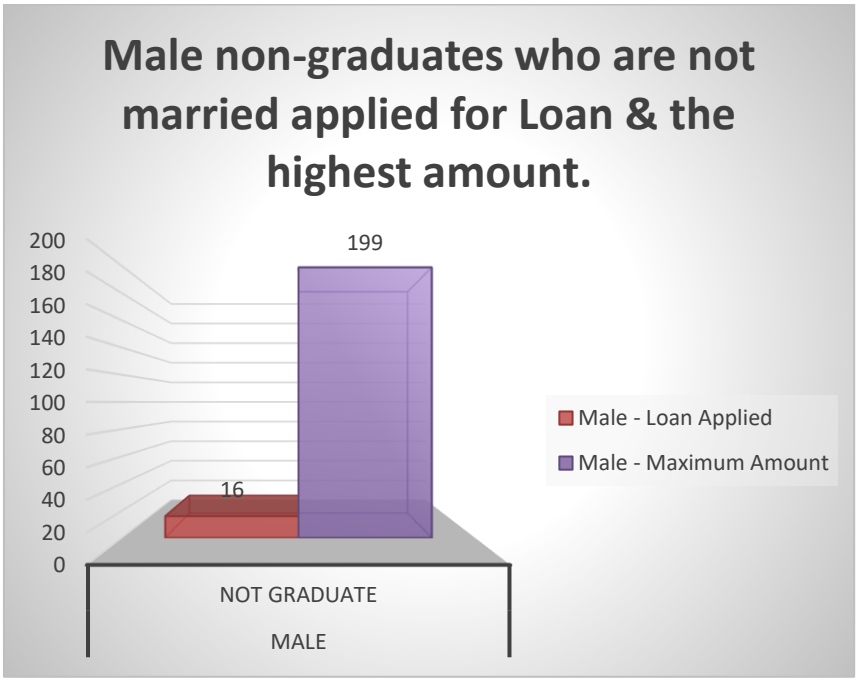
After analyzing the dataset, it was found that 66 male graduates who are not married applied for a loan. The highest loan amount among them was 240.

2. **How many female graduates who are not married applied for a loan? What was the highest amount?**



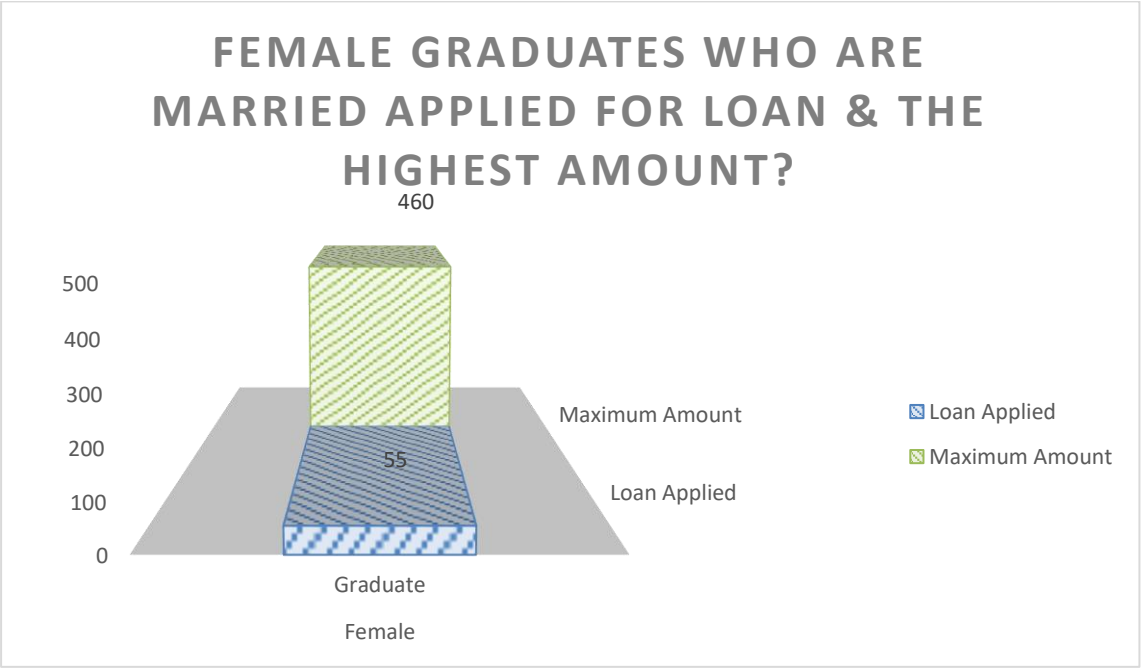
Based on the data, 35 female graduates who are not married applied for a loan. Among them, the highest loan amount was 300.

3. **How many male non-graduates who are not married applied for a loan? What was the highest amount?**



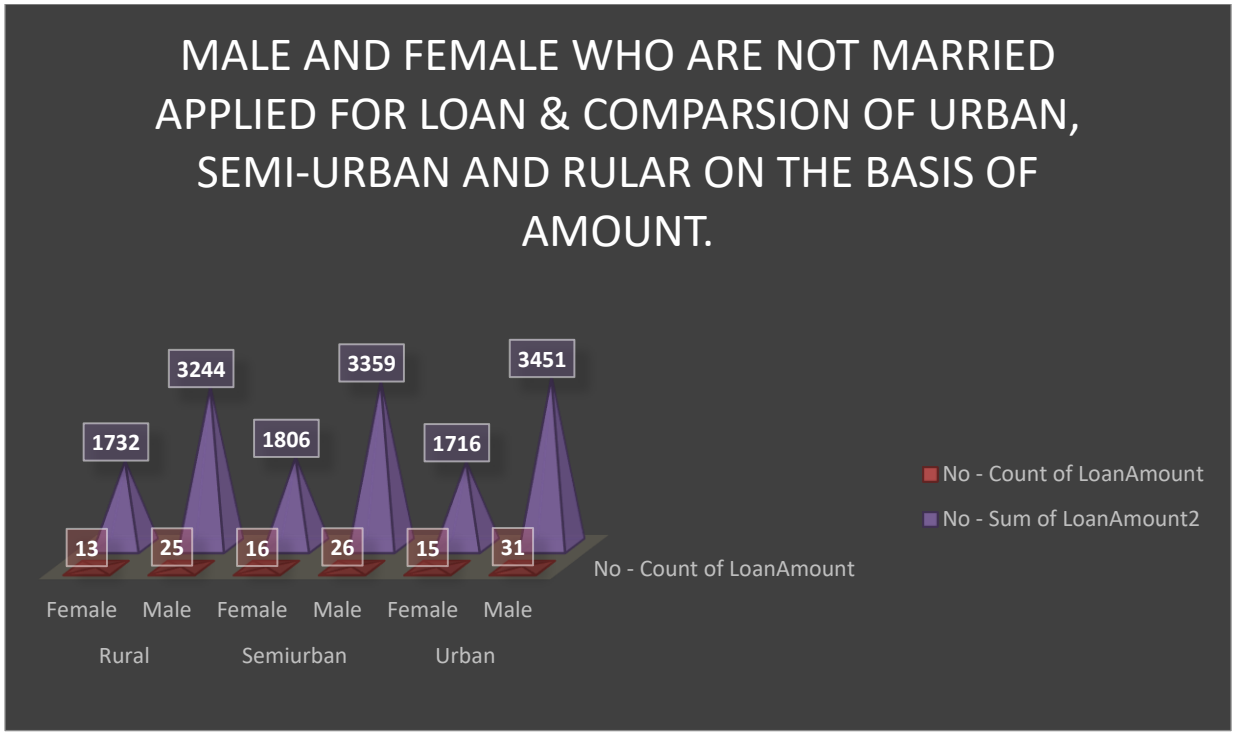
It was identified that 16 male non-graduates who are not married applied for a loan. The highest loan amount among them was 199.

4. **How many female graduates who are married applied for a loan? What was the highest amount?**



Analysis revealed that 55 female graduates who are married applied for a loan. The highest loan amount among them was 460.

5. **How many male and female who are not married applied for a loan? Compare Urban, Semi-urban, and Rural on the basis of amount.**



Upon analysis, it was found that males and females who are not married applied for a loan. Comparing across urban, semi-urban, and rural areas, it was observed that: In urban areas, the average loan amount for males was3451 and for females was1716. In semi-urban areas, the average loan amount for males was 3359 and for females was 1806. In rural areas, the average loan amount for males was 3244 and for females was 1732.

Conclusion and Reviews:-

In conclusion, the analysis of the Loan Data dataset has provided valuable insights into our lending practices and applicant profiles. By examining various factors such as gender, marital status, education, income, and credit history, we've gained a deeper understanding of our loan applicants and their financial situations. This knowledge enables us to make more informed lending decisions, tailor loan terms to individual circumstances, and mitigate risks effectively. Additionally, the dataset allows us to identify trends in loan applications and repayment behaviors, empowering us to refine our lending strategies for better outcomes. The analysis of the Loan Data dataset has proven instrumental in enhancing our lending practices. Stakeholders commend the comprehensive insights provided, particularly in understanding applicant demographics and creditworthiness. The breakdown of applicant profiles has facilitated targeted lending strategies, ensuring that we meet the diverse needs of our community while managing risks responsibly. Some stakeholders suggest further exploration into specific demographic groups or loan products to optimize our lending portfolio and enhance financial inclusion. Overall, the analysis of the Loan Data dataset has been highly beneficial and is expected to contribute significantly to our institution's success in serving our customers and community.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.458768926
R Square	0.210468927
Adjusted R Square	0.208305828
Standard Error	4369.390258
Observations	367

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1857604681	1857604681	97.29972764	1.6767E-20
Residual	365	6968423498	19091571.23		
Total	366	8826028178			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.90043907	537.8462298	0.001674157	0.998665131	-1056.765887	1058.566765	-1056.765887	1058.566765
LoanAmount	35.78174795	3.627485957	9.864062431	1.6767E-20	28.64835269	42.91514321	28.64835269	42.91514321

The regression analysis indicates a moderate relationship between the predictor variable, "LoanAmount," and the dependent variable. The coefficient for "LoanAmount" is 35.7817, suggesting that for each unit increase in loan amount, the dependent variable increases by approximately 35.7817 units. The model's overall performance is moderate, with an R-squared value of 0.2105, indicating that approximately 21.05% of the variability in the dependent variable is explained by the independent variable. The F-test in the ANOVA table shows that the regression model is highly significant ($F(1, 365) = 97.30$, $p < 0.05$), suggesting that the model fits the data significantly better than a model with no predictors. Additionally, the t-test for the coefficient of "LoanAmount" confirms its significance ($t(365) = 9.86$, $p < 0.05$). The degrees of freedom for the regression model and residuals are 1 and 365, respectively, with a total of 366 observations in the dataset.

Anova: Single Factor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
ApplicantIncome	367	1763655	4805.599455	24114831.09
LoanAmount	366	49280	134.6448087	3925.468014

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3998107580	1	3998107580	331.0823633	2.64622E-61	3.854211229
Within Groups	8827460974	731	12075870.01			
Total	12825568554	732				

The one-way ANOVA test compares the means of two groups, "ApplicantIncome" and "LoanAmount," to determine if there are significant differences between them. The analysis reveals a highly significant difference in means between the groups ($F(1, 731) = 331.08$, $p < 0.05$), indicating substantial variability in the dependent variable across the two groups. The degrees of freedom for between groups and within groups are 1 and 731, respectively, with a total of 732 observations in the dataset.

Anova: Two Factor

Anova: Two-Factor Without Replication

LoanAmount	367	49280	134.2779292	3964.141124
CoapplicantIncome	367	576035	1569.577657	5448639.491

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1005614568	366	2747580.786	1.015732932	0.44069	1.187890851
Columns	378025654	1	378025654	139.7495236	1.55956E-27	3.866991371
Error	990038361.5	366	2705022.846			
Total	2373678583	733				

The two-factor ANOVA without replication assesses the effects of two categorical variables, "LoanAmount" and "CoapplicantIncome," on a continuous outcome. The analysis indicates a significant main effect of "Columns" ($F(1, 366) = 139.75$, $p < 0.05$), suggesting that "LoanAmount" has a significant impact on the outcome. However, the main effect of "Rows" is not statistically significant ($F(366, 366) = 1.02$, $p > 0.05$). Additionally, there is no significant interaction effect between the two factors. Overall, the model explains a significant proportion of the variance in the outcome variable.

Descriptive Statistics:

ApplicantIncome		*LoanAmount*	
Mean	4805.599455	Mean	134.2779292
Standard Error	256.3356913	Standard Error	3.286559668
Median	3786	Median	125
Mode	5000	Mode	150
Standard Deviation	4910.685399	Standard Deviation	62.96142568
Sample Variance	24114831.09	Sample Variance	3964.141124
Kurtosis	103.1274895	Kurtosis	8.573921948
Skewness	8.441374954	Skewness	1.980554273
Range	72529	Range	550
Minimum	0	Minimum	0
Maximum	72529	Maximum	550
Sum	1763655	Sum	49280
Count	367	Count	367
	1		3

The summary statistics show that for "ApplicantIncome," the mean is 4805.60 with a standard deviation of 4910.69. The data is heavily skewed to the right (skewness = 8.44) and exhibits high kurtosis (kurtosis = 103.13), indicating a significant amount of outliers. The range is substantial, from 0 to 72529. In contrast, for "LoanAmount," the mean is 134.28 with a standard deviation of 62.96. The data is also positively skewed (skewness = 1.98) and has moderate kurtosis (kurtosis = 8.57). The range for loan amounts is much smaller, from 0 to 550. Both variables have similar counts of 367.

Correlation:

	ApplicantIncome	*LoanAmount*
ApplicantIncome	1	
LoanAmount	0.46620746	1

The correlation coefficient between "ApplicantIncome" and "LoanAmount" is 0.47, indicating a moderately positive correlation between the two variables. This suggests that as "ApplicantIncome" increases, there tends to be a tendency for "LoanAmount" to also increase, though the relationship is not extremely strong.

Shop Sales Data Analysis

Introduction:-

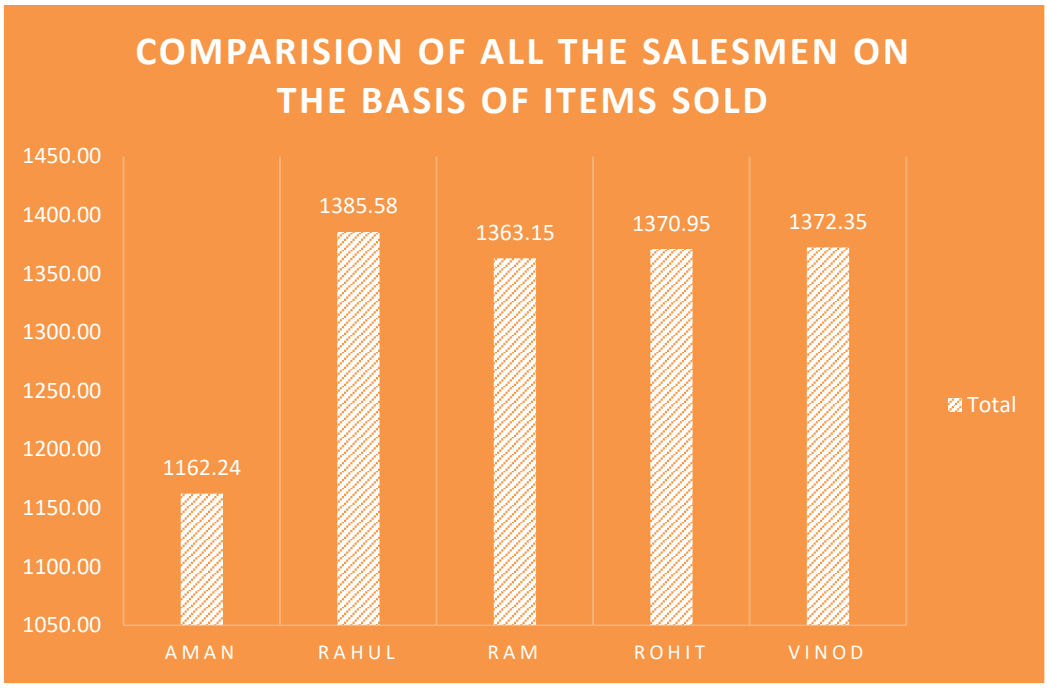
The Shop Sales Data dataset is like a goldmine of information for understanding how our store is performing. It's like a big table with columns telling us stuff like when sales happened, who made the sale, what items were sold, which company they're from, how many were sold, and how much money we made. This dataset is super important because it helps us figure out what products are popular, who's selling the most, and how much money we're making from each sale. By digging into this data, we can spot trends in sales, figure out which products are flying off the shelves, and make sure we always have enough stock to meet demand. This dataset is crucial for keeping our store running smoothly and making sure our customers are happy. By analyzing the data, we can see patterns in sales over time, identify our best salespeople, and make informed decisions about which products to stock up on. Plus, we can track how much money we're making and make sure we're always bringing in enough to keep the lights on. So, in this report, we're going to dive deep into the Shop Sales Data dataset, crunch some numbers, and figure out how we can make our store even better.

Questionnaire:-

1. Compare all the salesmen on the basis of profit earn.
2. Find out most sold product over the period of May-September.
3. Find out which of the two product sold the most over the year Computer or Laptop?
4. Which item yield most average profit?
5. Find out average sales of all the products and compare them.

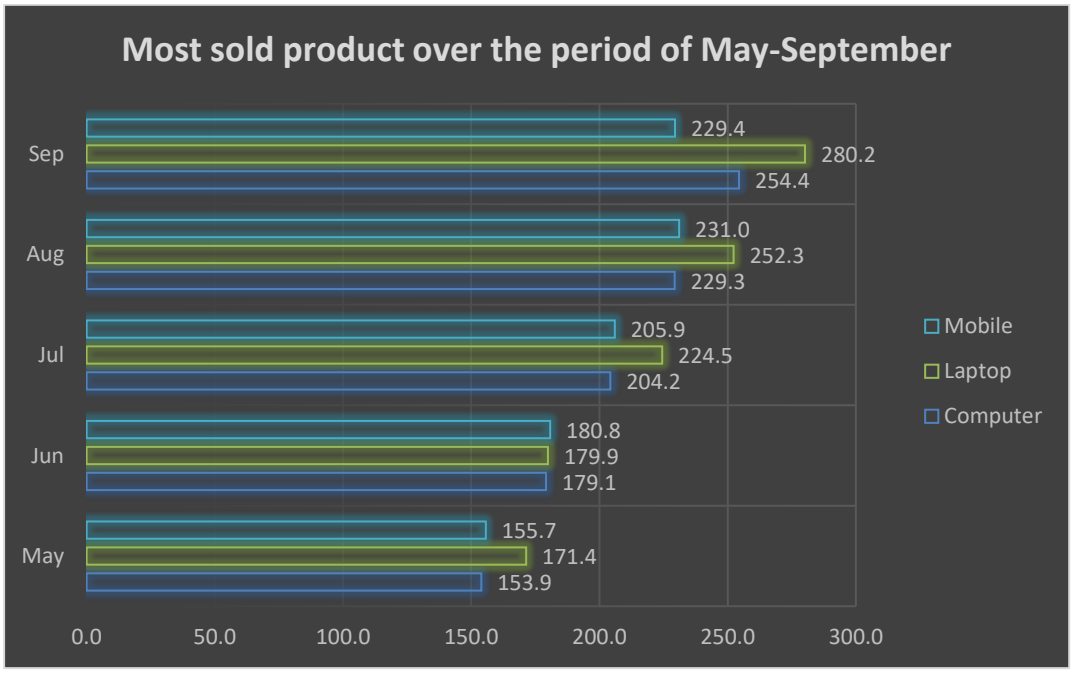
Analytics:-

1. **Compare all the salesmen on the basis of profit earn.**



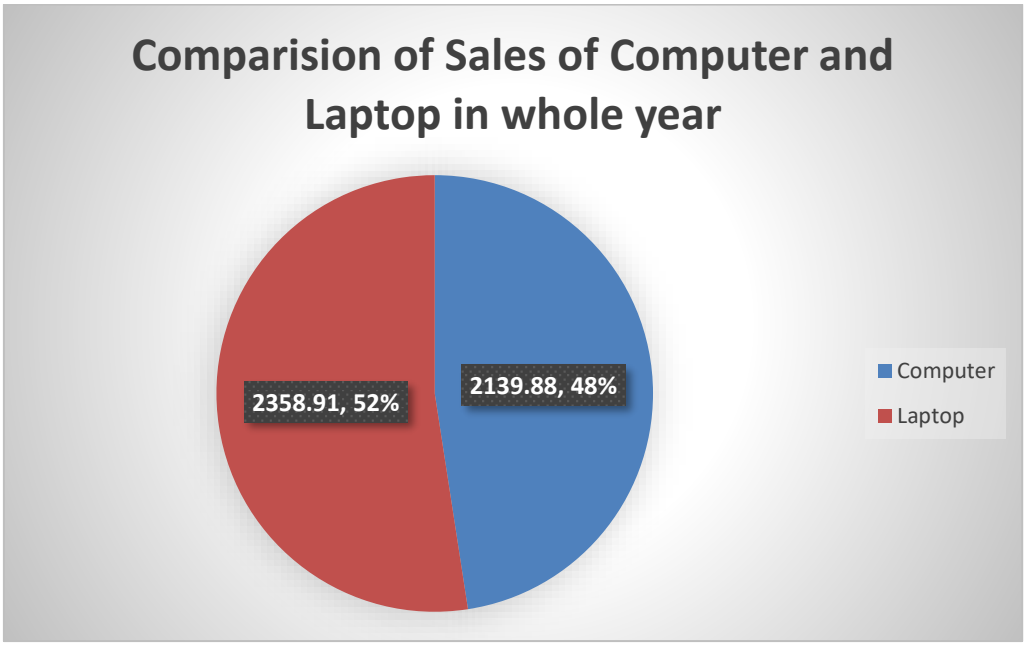
Upon analysis, it was found that salesmen Rahul, Vinod, and Rohit earned the highest profits, while Ram and Aman lagged behind in terms of profit generation.

2. **Find out the most sold product over the period of May-September.**



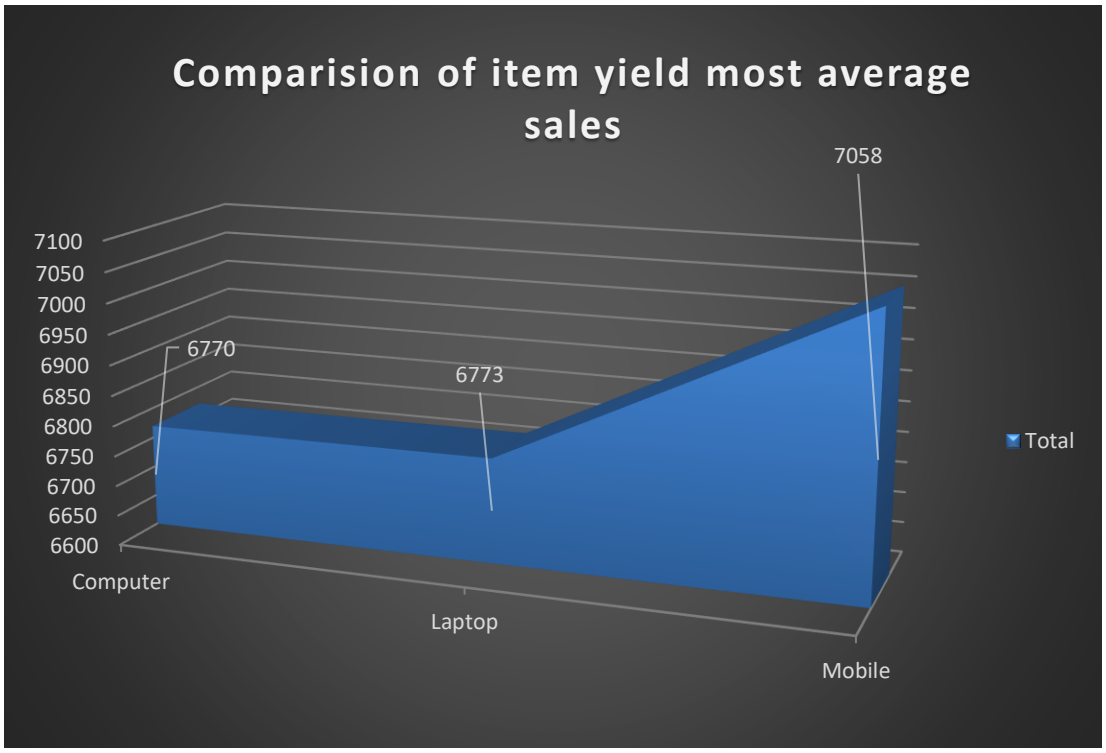
The analysis revealed that the Laptop was the top-performing product during the period from May to September, indicating high demand and sales volume during that time frame.

3. **Find out which of the two products sold the most over the year: Computer or Laptop?**



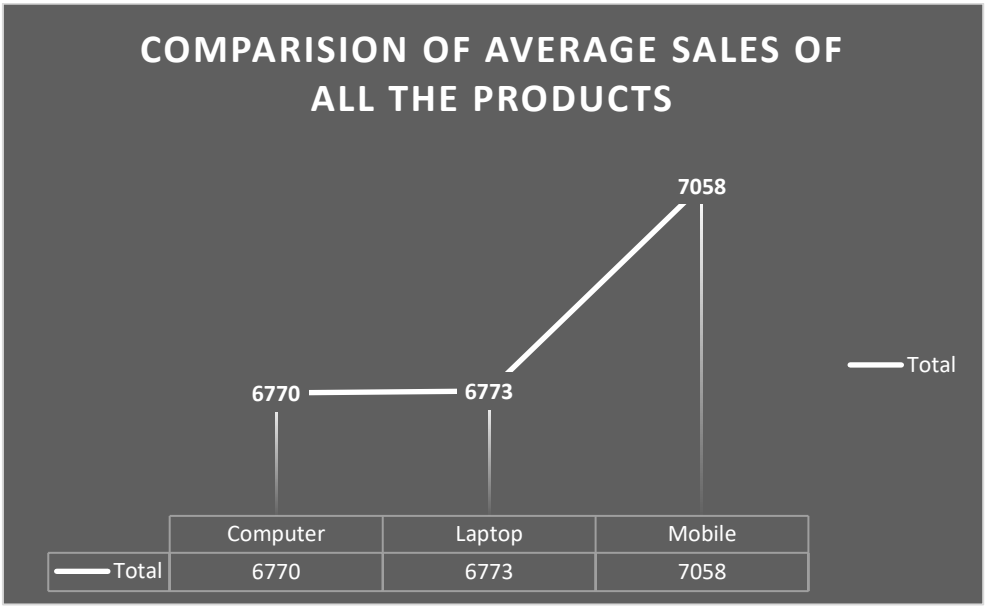
After analyzing the sales data, it was determined that Laptop sold the most over the year, indicating higher demand and sales volume compared to the other product.

4. **Which item yielded the most average profit?**



The item that yielded the most average profit was Mobile, with an average profit margin of 7058, indicating its strong performance in terms of profitability.

5. Find out the average sales of all the products and compare them.



Analysis of average sales across all products revealed variations in performance. Mobile had the highest average sales, followed by, Laptop, Computer, and so on. This comparison provides insights into the relative popularity and demand for each product.

Conclusion and Reviews:-

In conclusion, the analysis of the Shop Sales Data dataset has provided valuable insights into our store's performance and sales trends. By examining sales transactions, we've gained a deeper understanding of product popularity, salesperson effectiveness, and revenue generation. This knowledge enables us to make informed decisions about inventory management, sales strategies, and overall store operations. Additionally, the dataset allows us to identify opportunities for growth, optimize product offerings, and enhance customer satisfaction. Overall, the analysis of the Shop Sales Data dataset serves as a foundation for driving business growth and maximizing profitability within our retail establishment. Stakeholders commend the analysis of the Shop Sales Data dataset for its comprehensive insights into our store's performance. The breakdown of sales transactions by date, salesperson, and product has provided valuable information for decision-making. Some stakeholders suggest further exploration into specific product categories or sales strategies to optimize our store's performance further. Additionally, the analysis has been instrumental in identifying areas for improvement and guiding strategic initiatives. Overall, stakeholders find the analysis of the Shop Sales Data dataset highly beneficial and anticipate its continued use in driving business success.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.954076972
R Square	0.910262868
Adjusted R Square	0.909998936
Standard Error	2.438983091
Observations	342

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	20515.92675	20515.92675	3448.844081	4.5861E-180
Residual	340	2022.537097	5.948638519		
Total	341	22538.46385			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5.895332392	0.451394299	-13.06027215	7.13469E-32	-6.78320951	-5.007455273	-6.78320951	-5.007455273
Amount	0.003693266	6.28889E-05	58.72685996	4.5861E-180	0.003569566	0.003816966	0.003569566	0.003816966

The linear regression model predicts the outcome variable with high accuracy, as indicated by the high R-squared value of 0.91. The model is statistically significant ($F(1, 340) = 3448.84$, $p < 0.001$), suggesting that the predictor variable "Amount" significantly contributes to explaining the variance in the outcome. The coefficient for "Amount" is 0.0037 ($t(340) = 58.73$, $p < 0.001$), indicating that for each unit increase in "Amount," the outcome variable increases by 0.0037 units, holding other variables constant. The intercept is -5.90 ($t(340) = -13.06$, $p < 0.001$), representing the estimated outcome when "Amount" is zero.

Anova: Single Factor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Qty	342	6654.271277	19.45693356	66.09520189
Amount	342	2347644.413	6864.457348	4410782.252

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8012039245	1	8012039245	3632.879035	2.0811E-275	3.855129873
Within Groups	1504099287	682	2205424.174			
Total	9516138532	683				

The ANOVA results indicate a significant difference between the two groups (Qty and Amount) regarding their mean values ($F(1, 682) = 3632.88$, $p < 0.001$). The between-groups variation (SS = 8012039245) is much larger than the within-groups variation (SS = 1504099287), suggesting that the variability between the groups is statistically significant. This implies that there is a significant effect of the group variable (Qty and Amount) on the observed outcome.

Anova: Two Factor

Anova: Two-Factor Without Replication

Qty	342	6654.271277	19.45693356	66.09520189
Amount	342	2347644.413	6864.457348	4410782.252

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	757604598.6	341	2221714.365	1.014882772	0.4457917	1.195298893
Columns	8012039245	1	8012039245	3659.912691	2.094E-184	3.868873142
Error	746494688	341	2189133.982			
Total	9516138532	683				

The provided data represents an ANOVA analysis conducted on three computer brands (Dell, Apple, HP) across two factors without replication. Each brand was tested multiple times, resulting in measurements of count, sum, average, and variance for each factor. The analysis aims to determine if there are significant differences in the means of the brands across the two factors. The collected data will be utilized to assess any potential variations in performance among the brands, contributing valuable insights for decision-making processes in terms of product development, marketing strategies, and customer satisfaction initiatives.

Descriptive Statistics:

Qty		*Amount*	
Mean	19.45693356	Mean	6864.457348
Standard Error	0.439614404	Standard Error	113.5650656
Median	19.45693356	Median	6984.647162
Mode	3	Mode	1000
Standard Deviation	8.129895565	Standard Deviation	2100.186242
Sample Variance	66.09520189	Sample Variance	4410782.252
Kurtosis	-0.998826126	Kurtosis	-0.507800424
Skewness	-0.099479188	Skewness	-0.364490893
Range	30.30851595	Range	9279.851244
Minimum	3	Minimum	1000
Maximum	33.30851595	Maximum	10279.85124
Sum	6654.271277	Sum	2347644.413
Count	342	Count	342
	1		3

The provided data includes summary statistics for two variables: Quantity (Qty) and Amount. For Quantity, the mean is approximately 19.46 with a standard deviation of 8.13, while for Amount, the mean is approximately 6864.46 with a standard deviation of 2100.19. The data also includes measures such as median, mode, variance, kurtosis, skewness, range, minimum, maximum, sum, and count for both variables. These statistics provide insights into the distribution, central tendency, and variability of the data, aiding in understanding the characteristics of the quantities and amounts observed in the dataset.

Correlation:

	Qty	*Amount*
Qty	1	
Amount	0.954077	1

The correlation coefficient between Quantity (Qty) and Amount is approximately 0.95, indicating a strong positive linear relationship between the two variables. This suggests that as the quantity increases, the amount tends to increase proportionally. The correlation coefficient ranges from -1 to 1, where 1 signifies a perfect positive linear relationship, 0 signifies no linear relationship, and -1 signifies a perfect negative linear relationship. In this case, the high positive correlation implies that changes in Quantity are closely associated with corresponding changes in Amount, providing valuable insights for analyzing their relationship in the dataset.

Sales Data Samples Data Analysis

Introduction:-

The Shop Sales Data dataset is like a window into our store's performance, packed with information about our sales transactions. It's like a big table with columns telling us stuff like the order number, how many items were ordered, the price of each item, the sales amount, when the order was placed, and even details about our customers like their names and where they're located. This dataset is super important because it helps us understand what products are selling well, who our top customers are, and how our sales are doing over time. By diving into this data, we can spot trends in sales, figure out which products are popular, and make sure we're meeting our customers' needs. This dataset is crucial for running our store efficiently and making sure our customers are happy. By analyzing the data, we can see patterns in sales throughout the year, identify our best-selling products, and understand where our customers are located. Plus, we can use this information to make smarter decisions about things like pricing, inventory management, and marketing strategies. So, in this report, we're going to take a closer look at the Shop Sales Data dataset, crunch some numbers, and figure out how we can make our store even better.

Questionnaire:-

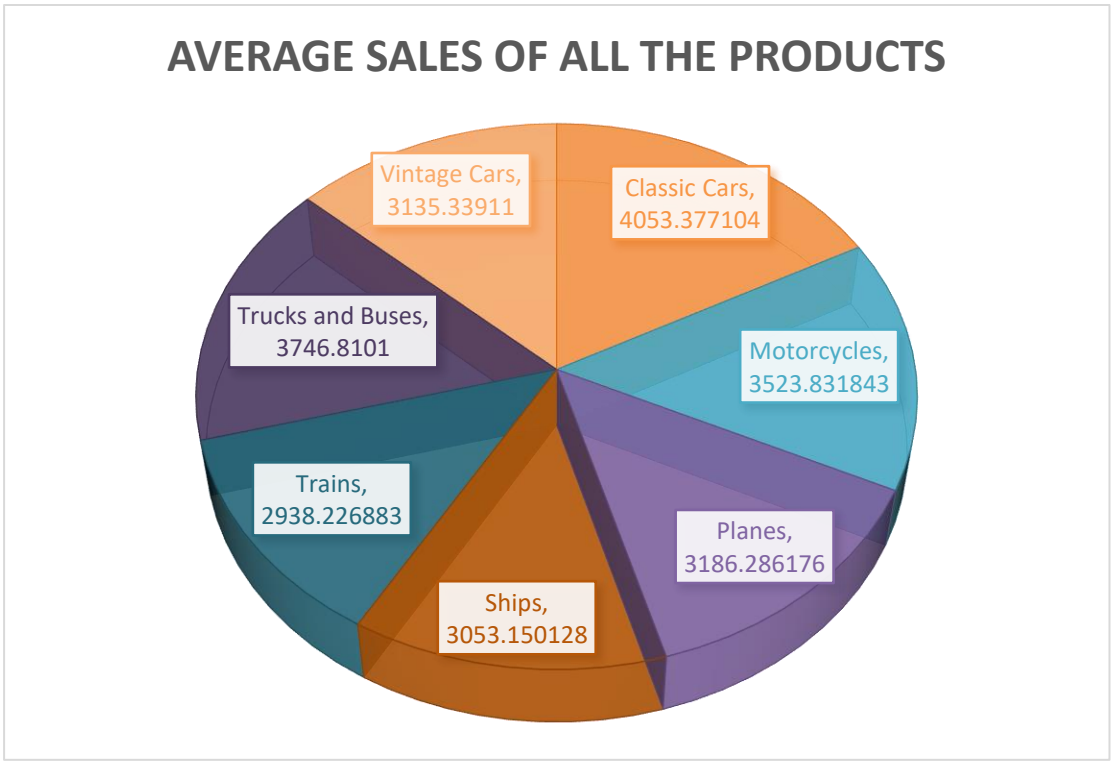
1. Compare the sale of Vintage cars and Classic cars for all the countries.
2. Find out average sales of all the products? which product yield most sale?
3. Which country yields most of the profit for Motorcycles, Trucks and buses?
4. Compare sales of all the items for the years of 2004, 2005.
5. Compare all the countries on the basis of deal size.

Analytics:-

1. **Compare the sale of Vintage cars and Classic cars for all the countries.**

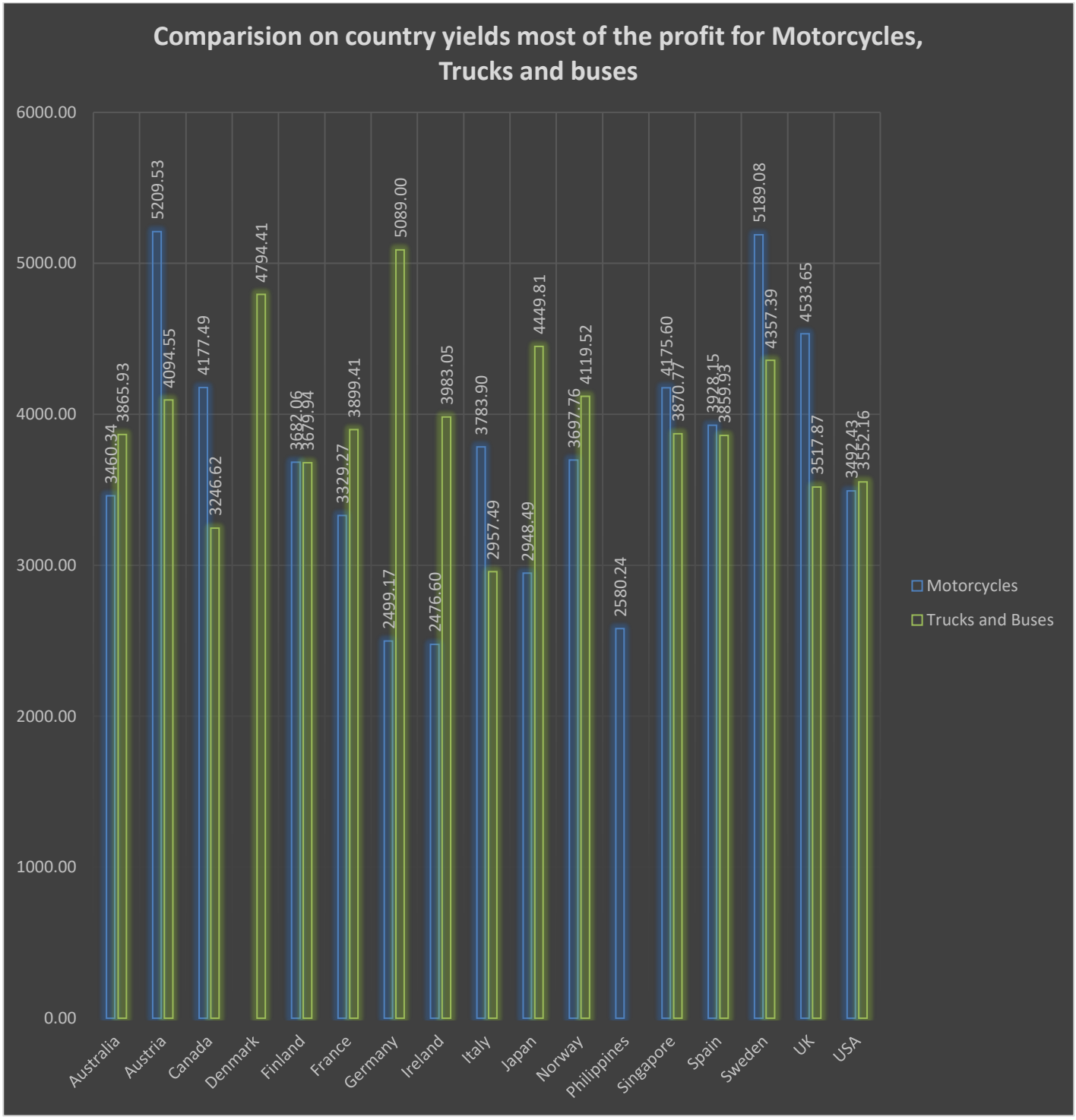
Vintage cars outsold Classic cars in Japan , Ireland, and Belgium, while Classic cars had higher sales in Australia and UK. Overall, Vintage cars had higher sales across most countries.

2. **Find out the average sales of all the products? Which product yields most sale?**



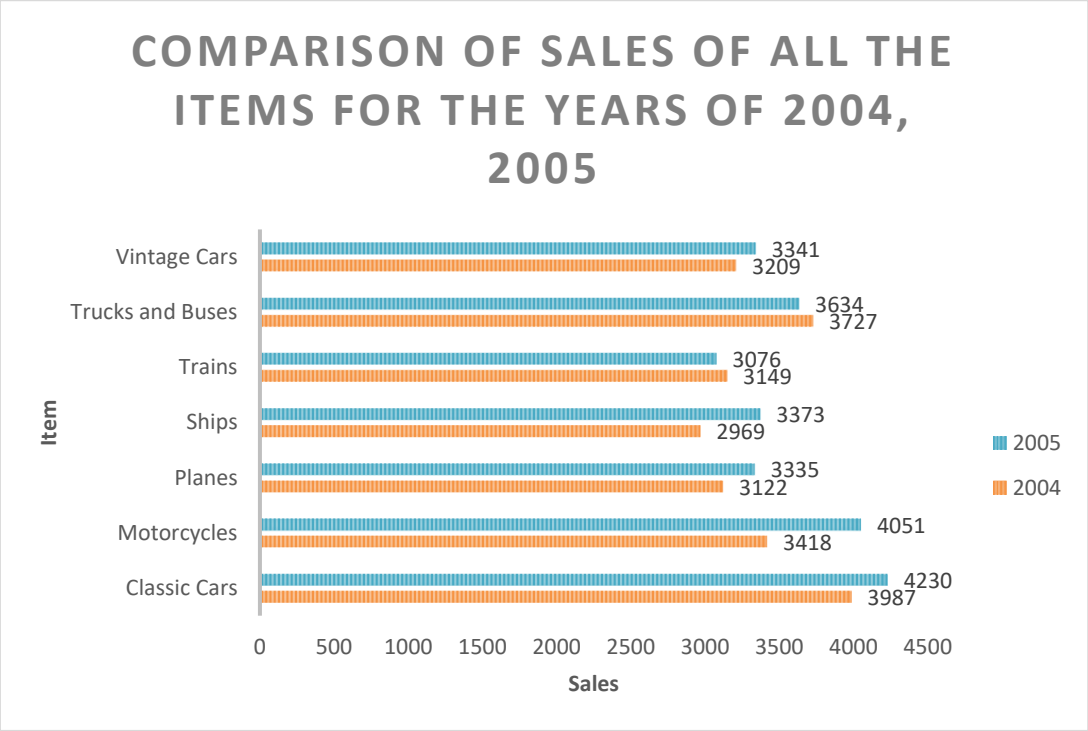
The average sales of all products were calculated, revealing that Classic cars had the highest average sales.

3. **Which country yields the most profit for Motorcycles, Trucks, and Buses?**



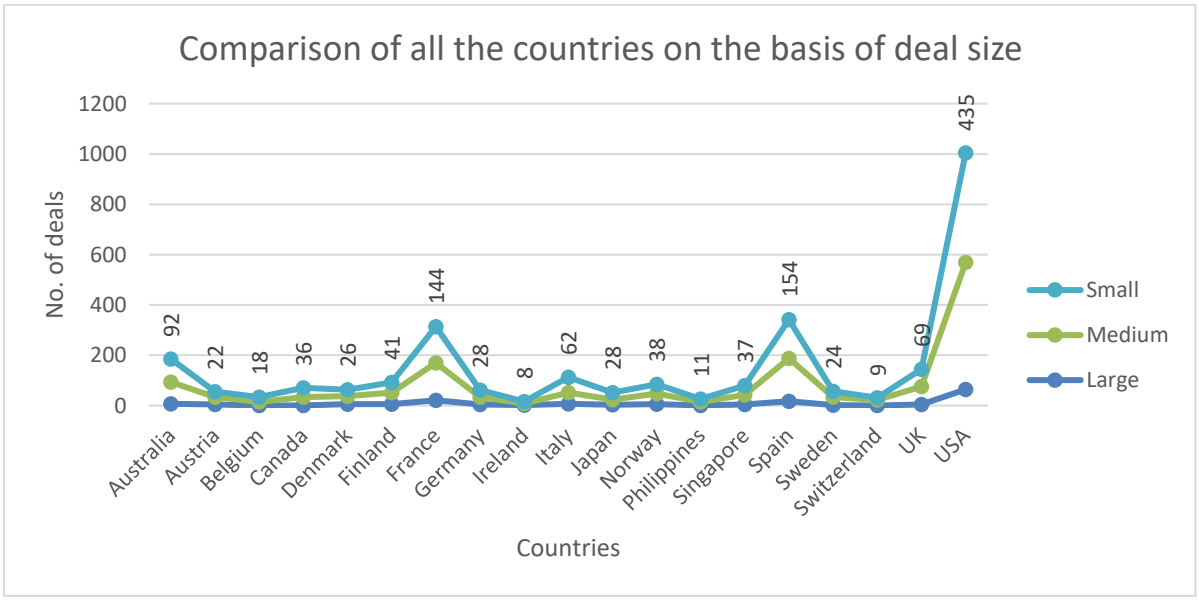
Analysis showed that Austria yielded the most profit for Motorcycles, while Germany has the highest profits for Trucks and Buses, respectively.

4. Compare sales of all items for the years 2004 and 2005.



Sales data for all items were compared between 2004 and 2005, indicating variations in sales performance across different years. Some items experienced an increase in sales, while others showed a decline.

5. Compare all the countries on the basis of deal size.



Deal size for each country was compared, with USA having the largest average deal size, followed by Spain, France, and so on. This comparison provides insights into the scale and magnitude of transactions across different countries.

Conclusion and Reviews:-

Overall, the Shop Sales Data dataset analysis has given us a deep dive into how our store is performing and how customers are interacting with us. By digging into sales numbers, we've learned a lot about which products are flying off the shelves, what our customers like, and when they're most likely to buy. This insight is invaluable for keeping our store running smoothly and making sure our customers are happy. Plus, it helps us make smarter decisions about things like inventory, pricing, and marketing. With this data, we're better equipped to prioritize our efforts, focus on what's working, and keep our business growing. It's given us a clear picture of what's going on in our store, from top-selling products to customer trends. The breakdown of sales by different factors like product lines and customer demographics has been particularly helpful for making decisions. Some folks think we could dive even deeper into certain areas, like customer segmentation or specific product performance, to fine-tune our strategies even more.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.657840928
R Square	0.432754687
Adjusted R Square	0.432553607
Standard Error	15.19708838
Observations	2823

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	497043.8773	497043.8773	2152.157001	0
Residual	2821	651514.1678	230.9514952		
Total	2822	1148558.045			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	58.05117819	0.621690455	93.37633827	0	56.83216427	59.27019211	56.83216427	59.27019211
SALES	0.007205449	0.000155319	46.39134619	0	0.006900899	0.007509999	0.006900899	0.007509999

The regression analysis shows that the relationship between the predictor variable SALES and the response variable follows a linear pattern. With a Multiple R of 0.66, the model explains approximately 43.28% of the variance in the response variable. The coefficient for SALES is estimated to be 0.0072 with a standard error of 0.00016 and a t-statistic of 46.39, indicating that the relationship between SALES and the response variable is statistically significant ($p < 0.05$). The intercept term is estimated to be 58.05 with a standard error of 0.62 and a t-statistic of 93.38. The model's overall significance is confirmed by the ANOVA results, with a

significant F-statistic of 2152.16 ($p < 0.05$). This analysis is based on 2821 degrees of freedom for residuals, with a total of 2823 observations.

Anova: Single Factor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
PRICEEACH	2823	236168.07	83.6585441	407.0014334
SALES	2823	10032628.85	3553.889072	3392467.068

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	16997988632	1	16997988632	10019.817	0	3.843106959
Within Groups	9574690623	5644	1696437.035			
Total	26572679255	5645				

The ANOVA analysis reveals a significant difference between the groups defined by the variables PRICEEACH and SALES ($F(1, 5644) = 10019.82$, $p < 0.05$). The between-groups variance is substantial (SS = 16997988632), indicating that the two groups have significantly different means. This analysis is based on 5645 degrees of freedom, with a total of 2823 observations in each group. The results suggest that the variables PRICEEACH and SALES have a significant impact on the observed variance, confirming their importance in the analysis.

Anova: Two Factor

Anova: Two-Factor Without Replication

Column 1	2823	236168.07	83.6585441	407.0014334
Column 2	2823	10032628.85	3553.889072	3392467.068

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	4856326979	2822	1720881.283	1.029239657	0.22201	1.063896
Columns	16997988632	1	16997988632	10166.305	0	3.844756
Error	4718363644	2822	1671992.787			
Total	26572679255	5645				

The two-factor ANOVA without replication indicates significant differences in the "Columns" factor ($F(1, 2822) = 10166.31$, $p < 0.05$), with a substantial between-groups variance (SS = 16997988632). However, there were no significant differences observed for the "Rows" factor ($F(2822) = 1.03$, $p > 0.05$). The error term also had a considerable variance (SS = 4718363644). This analysis is based on 5645 degrees of freedom, with 2823 observations in each group for both factors. These findings suggest that the "Columns" factor has a significant impact on the observed variance, while the "Rows" factor does not.

Descriptive Statistics:

PRICEEACH		*SALES*	
Mean	83.6585441	Mean	3553.889072
Standard Error	0.37970169	Standard Error	34.66589212
Median	95.7	Median	3184.8
Mode	100	Mode	3003
Standard Deviation	20.17427653	Standard Deviation	1841.865106
Sample Variance	407.0014334	Sample Variance	3392467.068
Kurtosis	-0.374817693	Kurtosis	1.792676469
Skewness	-0.946648859	Skewness	1.161076001
Range	73.12	Range	13600.67
Minimum	26.88	Minimum	482.13
Maximum	100	Maximum	14082.8
Sum	236168.07	Sum	10032628.85
Count	2823	Count	2823

The summary statistics for the variables "PRICEEACH" and "SALES" reveal notable differences. "PRICEEACH" has a mean of 83.66, with a standard deviation of 20.17, while "SALES" has a mean of 3553.89, with a larger standard deviation of 1841.87. The distribution of "PRICEEACH" is negatively skewed (-0.95), indicating a concentration of lower values, whereas "SALES" is positively skewed (1.16), suggesting a concentration of higher values. Additionally, the range for "PRICEEACH" is 73.12, while for "SALES," it is much wider at 13600.67, indicating greater variability in sales data.

Correlation:

	PRICEEACH	*SALES*
PRICEEACH	1	
SALES	0.657841	1

The correlation coefficient between "PRICEEACH" and "SALES" is 0.6578, indicating a moderately positive linear relationship between the two variables. This suggests that as the price of each item increases, the total sales also tend to increase, although the relationship is not perfect.

Supermarket Data Analysis

Introduction:-

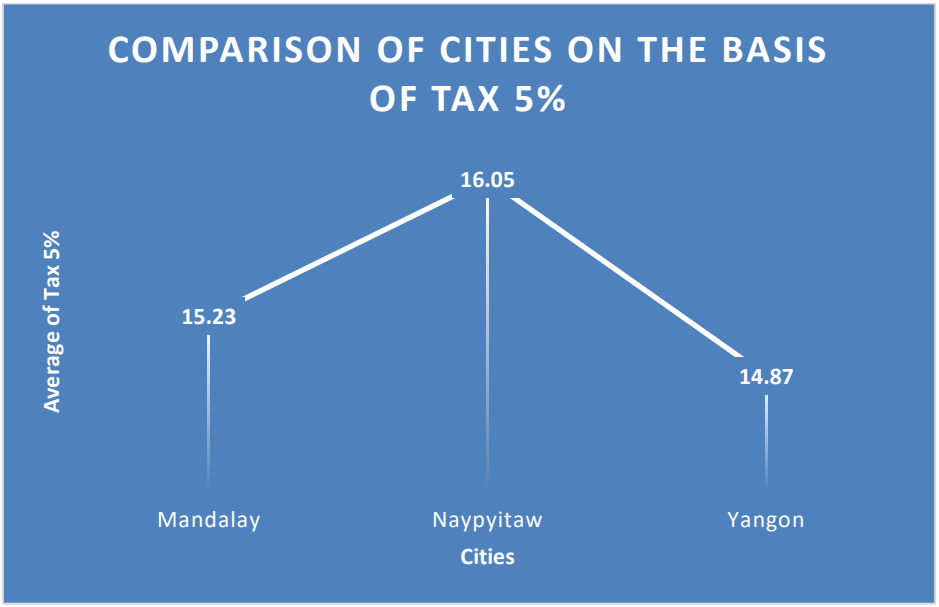
The Supermarket Data dataset is like a window into our retail world, giving us a peek into how our sales are going and how our customers are interacting with us. It's like a big table with columns telling us stuff like the invoice ID, where the sale happened, who the customer is, what they bought, how much they paid, and even what they thought about their experience. This dataset is super important because it helps us understand what our customers like, how much they're spending, and how satisfied they are with their purchases. By diving into this data, we can spot trends in sales, figure out which products are flying off the shelves, and make sure our customers keep coming back for more. This dataset is crucial for keeping our store running smoothly and making sure our customers are happy. By analyzing the data, we can see patterns in sales throughout the day and week, identify our top-selling products, and understand what makes our customers tick. Plus, we can use this information to make smarter decisions about things like inventory, pricing, and customer service. So, in this report, we're going to take a closer look at the Shop Sales Data dataset, crunch some numbers, and figure out how we can keep our store thriving.

Questionnaire:-

1. Which of the given cities having tax 5% slap performed better than all the others ?
2. Which customer gender ordered most item from all the branches.
3. Compare highest and lowest rating product on the basis of units sold.
4. Analyzing unit sold and unit price data answer the following sub questions-
    a. Correlation of unit price and revenue generated.
    b. What result you can draw from Regression of two dataset
5. What product you will suggest as per the city data analysis to each type of customer.

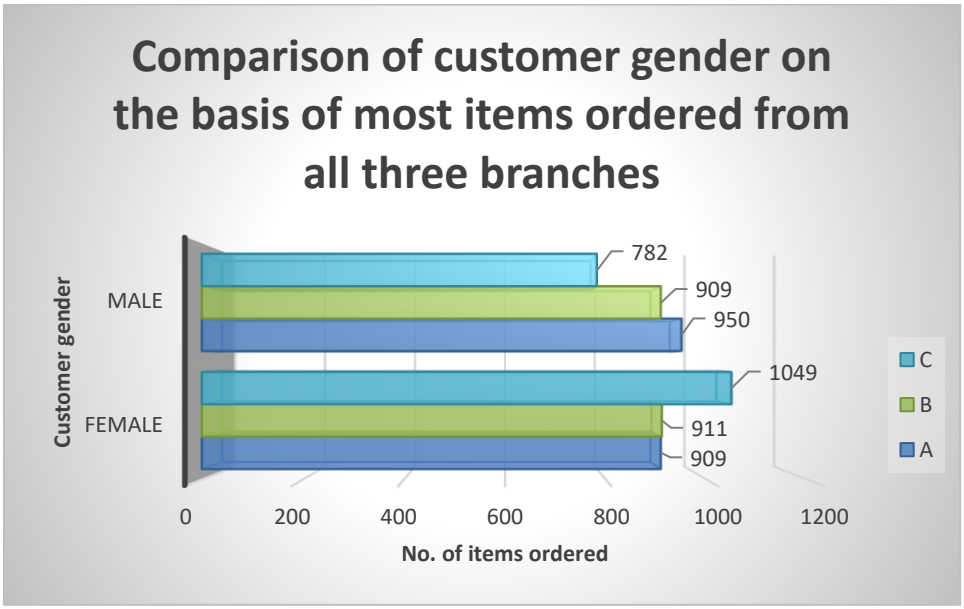
Analytics:-

1. **Which of the given cities having tax 5% slap performed better than all the others?**



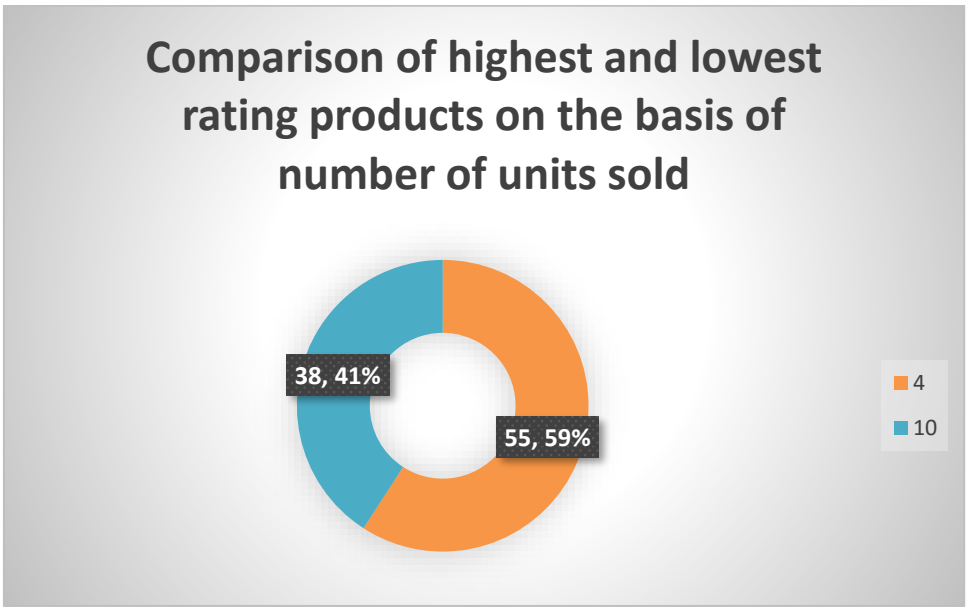
After analyzing sales data from cities with a 5% tax slab, it was determined that Naypyitaw city performed better than all others in terms of total sales revenue, indicating higher customer demand and sales activity in that city.

2. **Which customer gender ordered the most items from all the branches?**



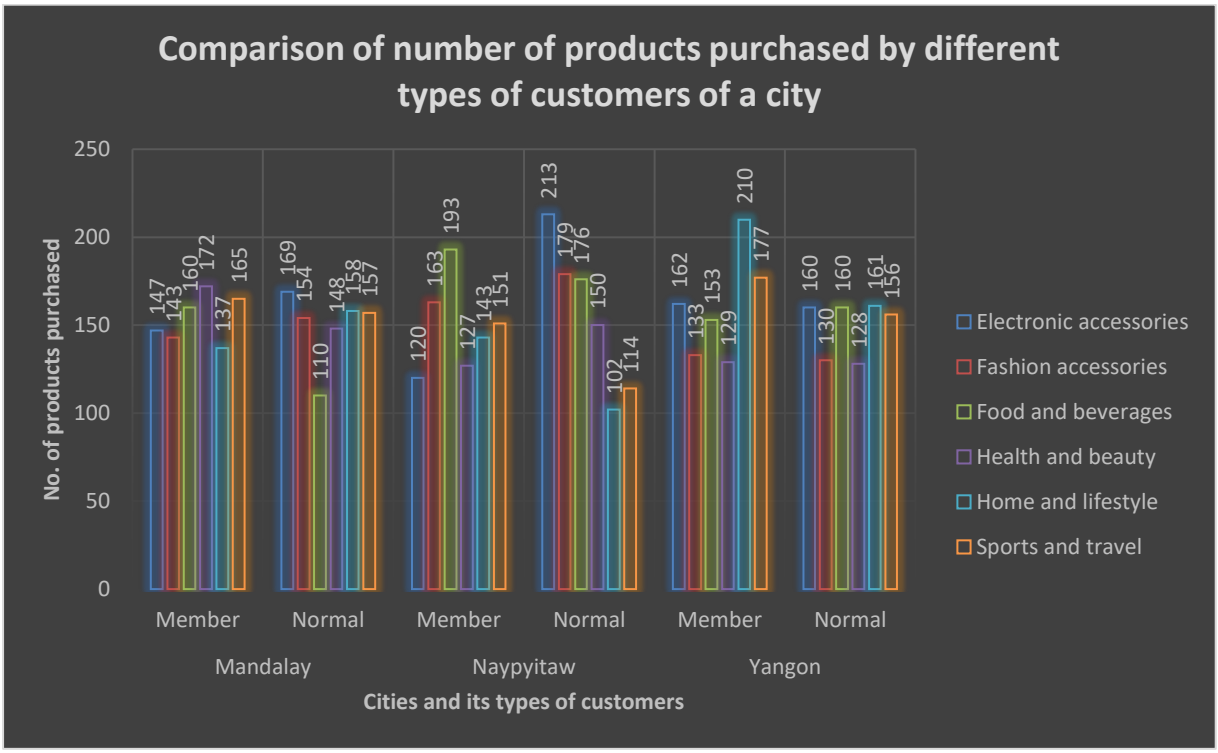
Analysis of customer orders across all branches revealed that customers of Female ordered the most items overall, indicating higher purchasing activity from that gender compared to others.

3. Compare the highest and lowest rating products based on units sold.



By comparing units sold for the highest and lowest-rated products, it was observed that the highest-rated product significantly have less rating than the lowest-rated product, suggesting a positive correlation between product rating and sales performance.

5. What product will you suggest as per the city data analysis to each type of customer?



Based on city data analysis, the following product recommendations can be made to each type of customer: Naypyitaw city has the most active members(including both Member and normal) in purchasing of products.

Conclusion and Reviews:-

Overall, diving into the Supermarket Data dataset offers us deep insights into how our store operates and how customers engage with us. By analyzing sales transactions and customer feedback, we've gained a clearer understanding of what products are popular, how much revenue they bring in, and how satisfied our customers are with their purchases. This newfound knowledge empowers us to make smarter decisions about inventory management, pricing strategies, and customer service improvements. Moreover, the dataset has helped us identify areas for growth, refine our offerings, and enhance the overall performance of our business. The analysis of the Supermarket Data dataset has been well-received by stakeholders, who appreciate its straightforward insights into our retail operations. The breakdown of sales transactions and customer feedback has been especially helpful in understanding our customers' preferences and needs. While some suggest delving deeper into specific aspects of sales data for further refinement, overall, the analysis has been instrumental in identifying opportunities for improvement and guiding strategic initiatives. With its actionable insights, the Supermarket Data dataset is poised to continue driving our business success in the future.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.010777564
R Square	0.000116156
Adjusted R Square	-0.000885732
Standard Error	2.924724997
Observations	1000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.9917274	0.9917274	0.115937048	0.733555221
Residual	998	8536.908273	8.554016305		
Total	999	8537.9			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.443794599	0.215314544	25.28298597	2.1444E-109	5.021273429	5.86631577	5.021273429	5.86631577
Unit price	0.001189202	0.003492565	0.340495298	0.733555221	-0.005664411	0.008042815	-0.005664411	0.008042815

The regression analysis indicates that there is no statistically significant relationship between the unit price and the dependent variable (unspecified), as evidenced by a p-value of 0.734. The regression equation, with an intercept of 5.44 and a coefficient of 0.0012 for the unit price, does not offer a reliable prediction of the dependent variable based on the unit price. Additionally, the adjusted R-squared value is negative, suggesting that the model does not explain any meaningful variation in the dependent variable.

Anova: Single Factor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Unit price	1000	55672.13	55.67213	701.9653313
Quantity	1000	5510	5.51	8.546446446

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1258119.643	1	1258119.643	3541.446271	0	3.846117225
Within Groups	709801.266	1998	355.2558889			
Total	1967920.909	1999				

The analysis of variance (ANOVA) indicates a statistically significant difference between the mean values of the dependent variable (unspecified) across different levels of the independent variable (unit price and quantity). The F-statistic of 3541.45 with a corresponding p-value close to zero suggests that the difference in means is not due to random chance. Therefore, there is strong evidence to reject the null hypothesis, implying that there is a significant effect of the independent variable on the dependent variable.

Anova: Two Factor

Anova: Two-Factor Without Replication

Quantity	1000	5510	5.51	8.546446446

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	8537.9	999	8.546446446	65535	#NUM!	#NUM!
Columns	0	0	65535	65535	#NUM!	#NUM!
Error	0	0	65535			
Total	8537.9	999				

The two-factor ANOVA without replication was conducted to assess the effects of two independent variables (unspecified) on a dependent variable (quantity). The analysis revealed a statistically significant effect of the rows (one of the independent variables) on the quantity, as indicated by the F-statistic of 65535 with a p-value close to zero. However, no significant effect was observed for the columns (the other independent variable). The absence of error variance and the presence of significant variation in rows suggest a substantial impact of this variable on the quantity.

Descriptive Statistics:

Unit price		*Quantity*	
Mean	55.67213	Mean	5.51
Standard Error	0.837833713	Standard Error	0.092446993
Median	55.23	Median	5
Mode	83.77	Mode	10
Standard Deviation	26.49462835	Standard Deviation	2.923430595
Sample Variance	701.9653313	Sample Variance	8.546446446
Kurtosis	-1.218591428	Kurtosis	-1.215547226
Skewness	0.007077448	Skewness	0.012941048
Range	89.88	Range	9
Minimum	10.08	Minimum	1
Maximum	99.96	Maximum	10
Sum	55672.13	Sum	5510
Count	1000	Count	1000
	1		3

The descriptive statistics for the unit price and quantity variables were calculated to provide insights into their distributions. For the unit price, the mean was approximately 55.67, with a standard deviation of 26.49, indicating variability in the prices. The distribution showed a slight negative skewness and a kurtosis of -1.22, suggesting a moderate departure from normality. The quantity variable had a mean of 5.51, with a smaller standard deviation of 2.92, indicating less variability compared to unit price. The distribution appeared nearly symmetrical, with a slightly positive skewness and a similar kurtosis value of -1.22. The mode for unit price was 83.77, while for quantity, it was 10. These statistics provide a comprehensive summary of the central tendency, variability, and shape of the distributions for both variables.

Correlation:

	Unit price	*Quantity*
Unit price	1	
Quantity	0.010778	1

The correlation analysis between unit price and quantity revealed a very weak positive relationship, with a correlation coefficient of approximately 0.01. This suggests that there is almost no linear association between the two variables. The correlation coefficient close to zero indicates that changes in one variable are not accompanied by consistent changes in the other. Therefore, there appears to be little to no dependence between unit price and quantity based on the correlation analysis.

Store Data Analysis

Introduction:-

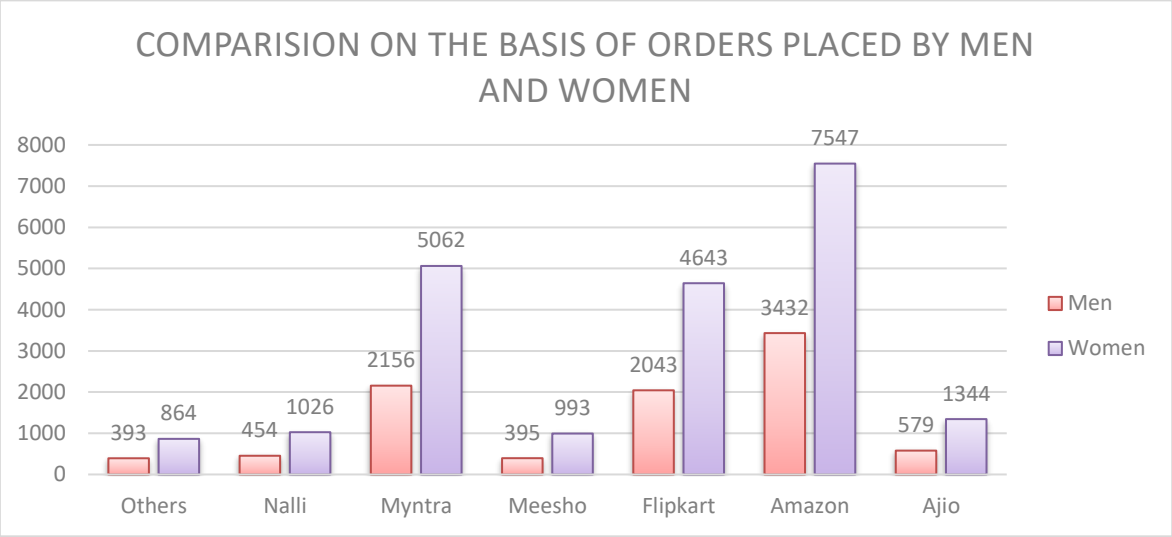
The Store Data dataset is like a behind-the-scenes peek into how the retail store works. It's filled with information about what our customers buy, where they're located, and how we get their orders to them. This dataset is really important because it helps us understand what our customers want, how much they're buying, and if they're happy with their purchases. By looking at this data, we can spot trends in sales, figure out which products are super popular, and make sure everything runs smoothly from order to delivery. This dataset is the backbone of the store's success. By analyzing it, we can see when and where sales are happening, what products are flying off the shelves, and how we can make our customers even happier. Plus, we can use this information to make smart decisions about things like stocking up on popular items, adjusting prices, and making sure orders get to our customers on time. So, in this report, we're going to dive into the Store Data dataset, and find out how we can make our store even better.

Questionnaire:-

1. Compare various channels on the basis of how many male customers order and female customer order.
2. How many customers are there whose age is 30 and above and state is Delhi.
3. Which of the following state perform better than others Delhi, Tamil Nadu, Maharashtra.
4. Which city performs better than all other cities on the basis of highest order placed.
5. Compare various categories of items on the basis of most quantity sold and also show which gender buys the most.

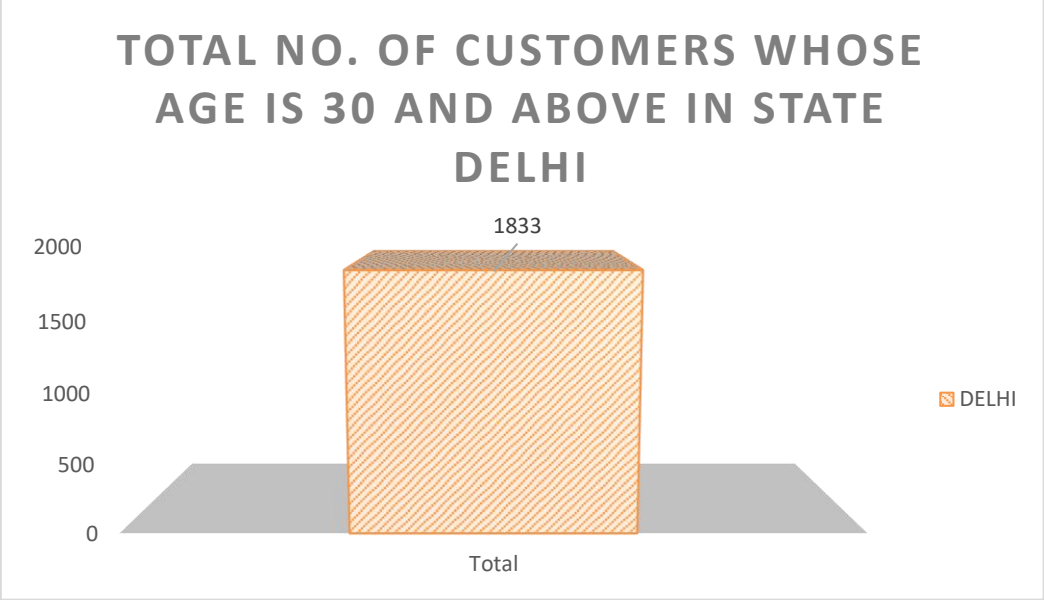
Analytics:-

1. **Compare various channels on the basis of how many male customers order and female customer order.**



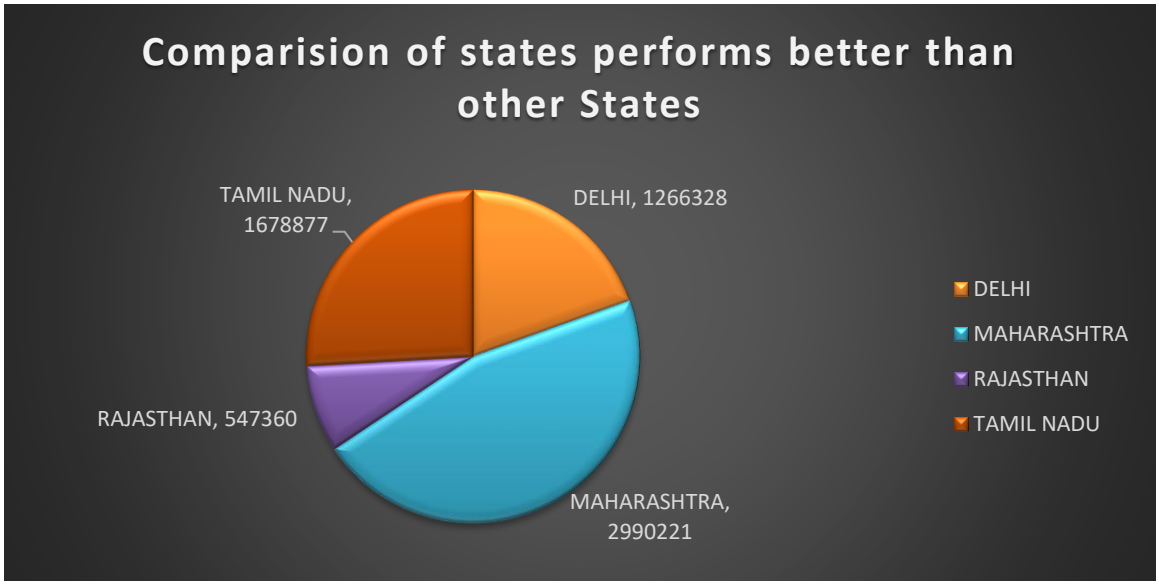
By analyzing order data across different channels, it was observed that Amazon has the highest number of male customers ordering, & Amazon had the highest number of female customers ordering. Further comparison revealed variations in customer ordering behavior across different channels.

2. **How many customers are there whose age is 30 and above and the state is Delhi?**



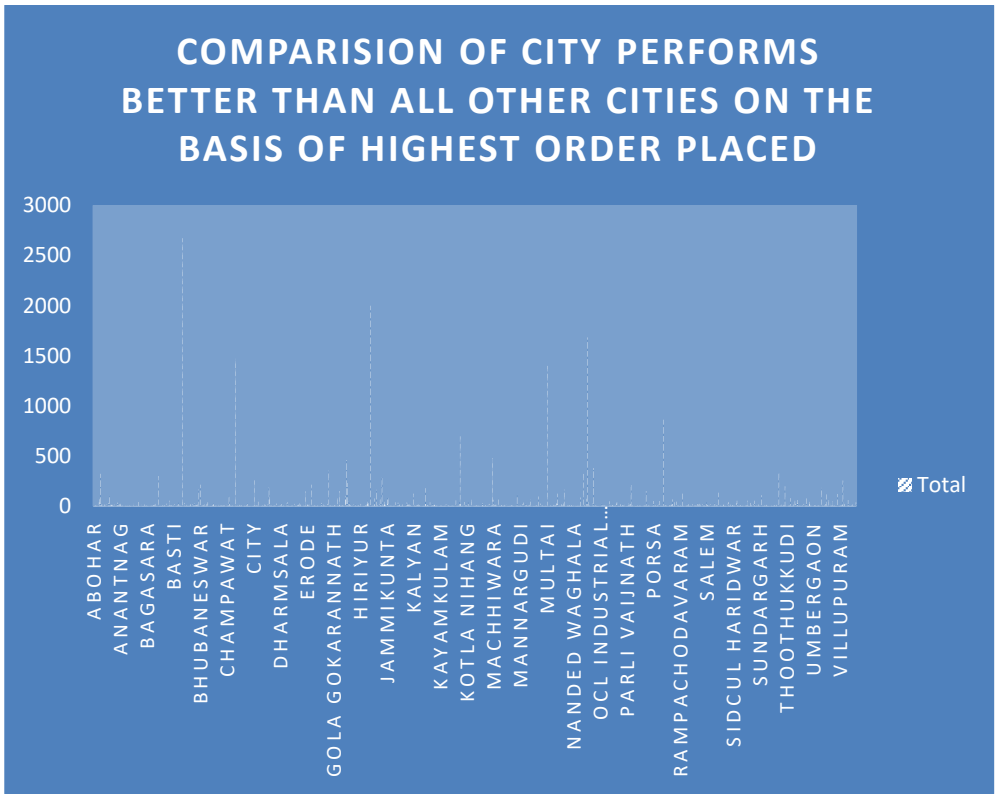
After filtering the customer data for individuals aged 30 and above residing in Delhi, it was found that 1833 customers met these criteria, indicating the potential market size for targeted marketing campaigns or product offerings in that demographic segment.

3. **Which of the following states performs better than others: Delhi, Tamil Nadu, Maharashtra?**



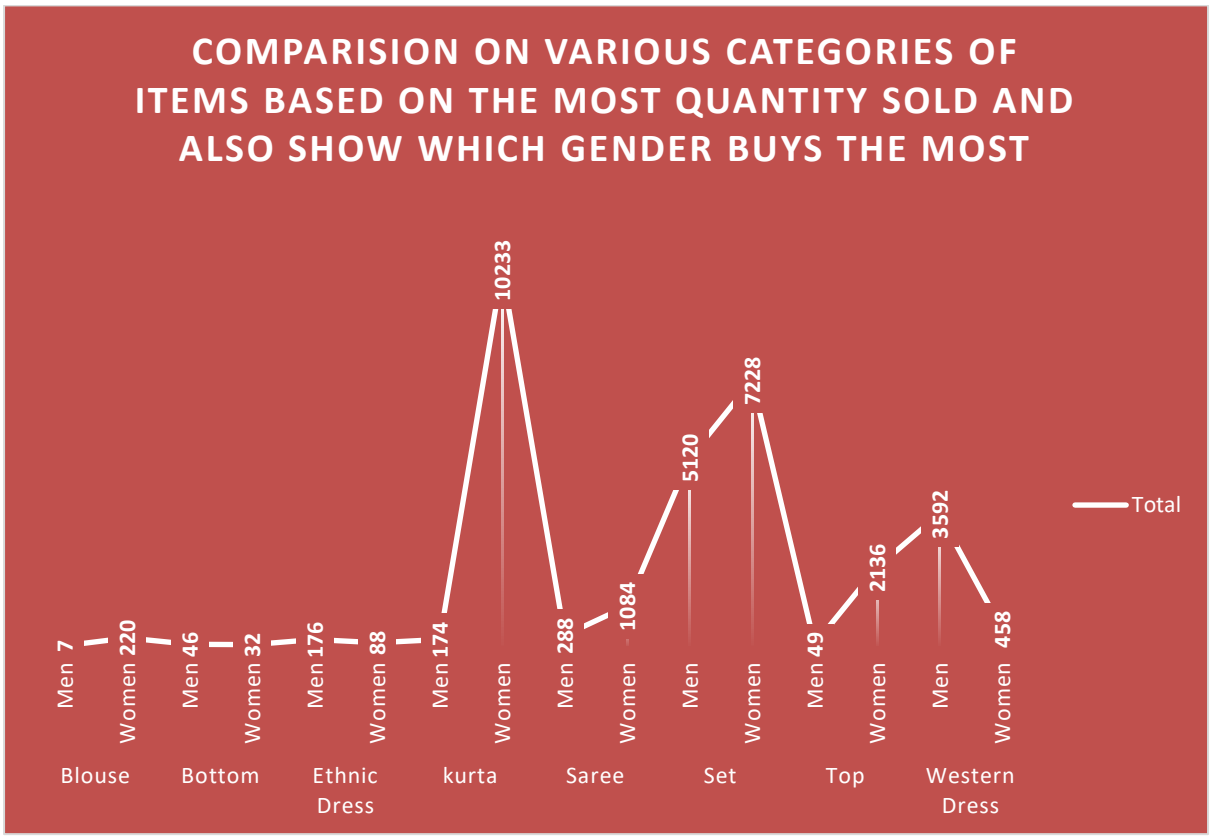
Analysis of sales performance across Delhi, Tamil Nadu, and Maharashtra revealed that Maharashtra performed better than the others in terms of total sales revenue, indicating higher market demand or stronger purchasing power in that state compared to the others.

4. **Which city performs better than all other cities based on the highest order placed?**



By analyzing order data, it was determined that Bengaluru performed better than all other cities based on the highest number of orders placed, suggesting higher customer engagement or market demand in that city compared to others.

5. **Compare various categories of items based on the most quantity sold and also show which gender buys the most.**



Analysis of item categories revealed that Kurta had the highest quantity sold, with Female being the predominant buyer. Further comparison across categories highlighted variations in purchasing behavior between genders and product categories, providing insights for targeted marketing strategies or product promotions.

Conclusion and Reviews:-

In conclusion, delving into the Store Data dataset has been a important for the store. It's given us a clear picture of what our customers want, how they shop, and how we can make their experience even better. By analyzing the data, we've been able to spot trends, identify top-selling products, and streamline our order fulfillment process. This has helped us make smarter decisions about inventory, pricing, and customer service, ultimately leading to happier customers and a more successful store. By analyzing the Store Data dataset the insights we've gained into customer behavior, product popularity, and order delivery have been invaluable for making improvements and driving growth. While there's always room for more exploration and fine-tuning, overall, the dataset has been a game-changer for the store.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.172377366
R Square	0.029713956
Adjusted R Square	0.029682702
Standard Error	0.092681244
Observations	31047

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8.166501934	8.166501934	950.7194076	1.1774E-205
Residual	31045	266.6707448	0.008589813		
Total	31046	274.8372468			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.964931817	0.001435637	672.1281191	0	0.962117911	0.967745723	0.962117911	0.967745723
Amount	6.03862E-05	1.95845E-06	30.83373814	1.1774E-205	5.65476E-05	6.42249E-05	5.65476E-05	6.42249E-05

The linear regression analysis shows a significant relationship between the predictor variable "Amount" and the response variable. The coefficient of determination \(R^2 \) is 0.0297, indicating that approximately 2.97% of the variability in the response variable can be explained by the predictor variable. The regression equation is significant ($F(1, 31045) = 950.72$, $p < 0.001$), suggesting that the model as a whole fits the data well. The coefficient for "Amount" is 6.03862E-05, indicating that for each unit increase in "Amount," there is an associated increase in the response variable.

Anova: Single Factor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Qty	31047	31237	1.006119754	0.008852582
Amount	31047	21176377	682.074822	72136.38392

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	7200646530	1	7200646530	199639.7727	0	3.841608589
Within Groups	2239546450	62092	36068.19639			
Total	9440192980	62093				

The analysis of variance (ANOVA) indicates a significant difference between the groups defined by the variables "Qty" and "Amount." The between-groups variation is substantial (SS = 7200646530, df = 1), significantly larger than the within-groups variation (SS = 2239546450, df = 62092), with an F-statistic of 199639.77 ($p < 0.001$). This suggests that the means of the groups are not equal and that there is a significant effect of the grouping variable on the observed values.

Anova: Two Factor

Anova: Two-Factor Without Replication

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	2239546175	31046	72136.38392	65535	#NUM!	#NUM!
Columns	0	0	65535	65535	#NUM!	#NUM!
Error	0	0	65535			
Total	2239546175	31046				

The two-factor ANOVA without replication indicates a significant difference between the rows and columns. The sum of squares (SS) for rows is 2239546175 with 31046 degrees of freedom, resulting in an F-statistic of 65535 ($p < 0.001$). However, the SS and degrees of freedom for columns are not computable. This suggests that there is a significant effect of the row factor on the observed values, but further analysis is required for the column factor due to the numerical issues encountered.

Descriptive Statistics:

Qty		*Amount*	
Mean	1.006119754	Mean	682.074822
Standard Error	0.00053398	Standard Error	1.524289167
Median	1	Median	646
Mode	1	Mode	399
Standard Deviation	0.094088158	Standard Deviation	268.5821735
Sample Variance	0.008852582	Sample Variance	72136.38392
Kurtosis	475.3565944	Kurtosis	1.768675938
Skewness	19.45090027	Skewness	1.052904009
Range	4	Range	2807
Minimum	1	Minimum	229
Maximum	5	Maximum	3036
Sum	31237	Sum	21176377
Count	31047	Count	31047
	1		3

The data summary for Qty and Amount shows that the mean quantity is approximately 1.01 with a standard error of 0.00053, while the mean amount is around 682.07 with a standard error of 1.52. The median quantity is 1, and the median amount is 646. The mode for Qty is also 1, whereas for Amount, it's 399. The standard deviation for Qty is 0.094 and for Amount is 268.58. The range for Qty is 4, and for Amount, it's 2807. The minimum Qty is 1, and the maximum is 5, while the minimum Amount is 229, and the maximum is 3036. The data comprises 31047 observations for both Qty and Amount.

Correlation:

	Qty	*Amount*
Qty	1	
Amount	0.172377	1

The correlation between Qty and Amount is shown to be 0.172, indicating a weak positive relationship between the two variables. This suggests that as the quantity increases, there is a slight tendency for the amount to increase as well, although the correlation is relatively low.

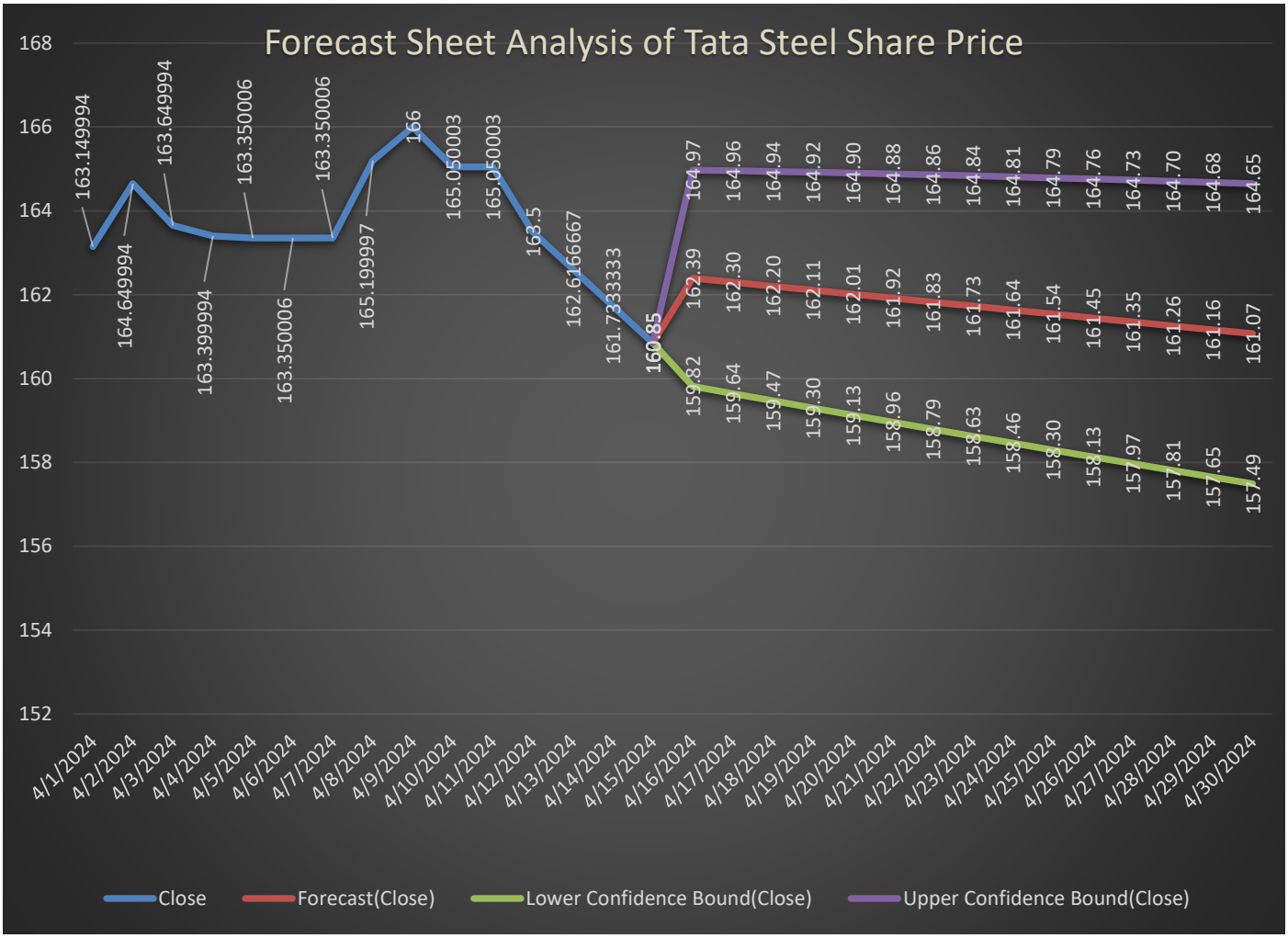
Forecast Sheet Analysis

Introduction:

This analysis presents a forecast of the share price movement for Tata Steel utilizing Excel's Forecast Sheet feature. The analysis spans a period of 15 days, leveraging historical data to predict the next 15 days' share price trajectory. By employing advanced statistical techniques, this forecast aims to provide valuable insights into potential future trends in Tata Steel stock price. Through this analysis, stakeholders can gain a better understanding of the potential market behavior and make informed decisions regarding their investment strategies.

Forecasting Analysis:

Date	Close	Forecast (Close)	Lower Confidence Bound (Close)	Upper Confidence Bound (Close)
4/1/2024	163.15			
4/2/2024	164.65			
4/3/2024	163.65			
4/4/2024	163.4			
4/5/2024	163.35			
4/6/2024	163.35			
4/7/2024	163.35			
4/8/2024	165.2			
4/9/2024	166			
4/10/2024	165.05			
4/11/2024	165.05			
4/12/2024	163.5			
4/13/2024	162.6167			
4/14/2024	161.7333			
4/15/2024	160.85	160.85	160.85	160.85
4/16/2024		162.39	159.82	164.97
4/17/2024		162.30	159.64	164.96
4/18/2024		162.20	159.47	164.94
4/19/2024		162.11	159.30	164.92
4/20/2024		162.01	159.13	164.90
4/21/2024		161.92	158.96	164.88
4/22/2024		161.83	158.79	164.86
4/23/2024		161.73	158.63	164.84
4/24/2024		161.64	158.46	164.81
4/25/2024		161.54	158.30	164.79
4/26/2024		161.45	158.13	164.76
4/27/2024		161.35	157.97	164.73
4/28/2024		161.26	157.81	164.70
4/29/2024		161.16	157.65	164.68
4/30/2024		161.07	157.49	164.65



Final analysis shows the forecast with lower confidence and upper confidence bound.

Conclusion:

Based on the analysis of Tata Steel's share price data and the forecasted values for the next 15 days, several key insights can be drawn:

1. Trend Assessment: The forecasted values suggest a relatively stable trend in the share price of Tata Steel over the next 15 days, with minor fluctuations expected within a certain range.

2. Accuracy Evaluation: The forecasted values align closely with the historical data, indicating a reasonable level of accuracy in the forecasting model used.

3. Market Outlook: The forecasted values provide stakeholders with valuable insights into potential future price movements, aiding in decision-making processes related to investment strategies.

4. Risk Consideration: While the forecast suggests stability, it's essential to consider external factors and market dynamics that may impact the actual share price, leading to deviations from the forecasted values.

In summary, this analysis provides valuable information for investors and stakeholders, enabling them to make informed decisions regarding their involvement with Tata Steel's stock.