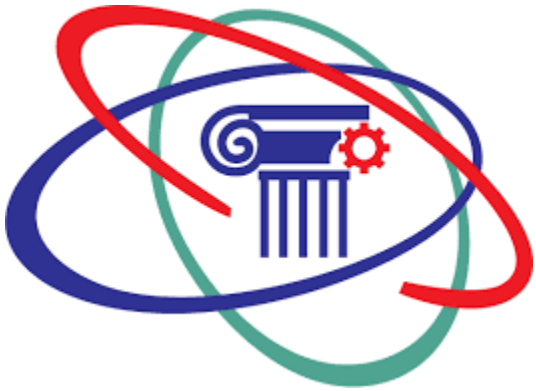
Acropolis Institute of Technology and Research, Indore
Department of Computer Science and Engineering



CSE-1 III year VI sem
Jan-June 2024

Data Analytics Lab File

Submitted To
Prof. Anurag Punde
CSE Dept.

Submitted By
Chirag Patil
0827CS211064

Jan-June 2024

Index

S. No.	Name of Experiment	Submission Date	Faculty Sign
1	Thorough Guide to Data Analysis: Foundations, Statistical Analytics, Tests of Hypothesis, Regression, Correlation, and ANOVA		
2	Dashboards		
3	Cookie Data Report		
4	Store Data Report		
5	Car Collection Report		
6	Examining Sales by Sector in the United States		
7	Loan Data Report		
8	Shop Sales Data Report		
9	Sales Data Sample Report		
10	Forecast Analysis of Twitter Stock Data Set		

Thorough Guide to Data Analysis: Foundations, Statistical Analytics, Tests of Hypothesis, Regression, Correlation, and ANOVA

Data Analysis Principles

Introduction to Data Analysis

Examining, purifying, manipulating, and analysing data is just one step in the complex process of data analysis, which aims to derive valuable insights. It is essential to a number of fields, including science, business, healthcare, and finance. Finding patterns, trends, connections, and abnormalities in the data is the main goal of data analysis since it allows one to utilize the information to guide decisions and take appropriate action.

Steps in Data Analysis

1. **Data Collection**: The process of gathering raw data from many sources, including databases, surveys, sensor networks, social media platforms, and Internet of Things devices, is known as data collecting. This is the first step in the data analysis process. The importance and Caliber of the data gathered have a big influence on the analysis's conclusions.

2. **Data Cleaning**: Data cleaning, also known as data cleansing or data scrubbing, involves identifying and rectifying errors, inconsistencies, and missing values in the dataset. This step ensures data accuracy and reliability for subsequent analysis.

3. **Data Preprocessing**: Preparing the dataset for analysis through a variety of procedures is known as data preparation. This covers feature selection, dimensionality reduction, controlling outliers, and data transformation (such as normalization and log transformation). The purpose of preprocessing procedures is to increase data quality and analytical model performance.

4. **Data Exploration**: Examining the dataset to learn more about its composition, distribution, and correlations between variables is known as data exploration. Analysts can better comprehend underlying trends and pinpoint possible areas of interest with the aid of exploratory data analysis (EDA) tools like correlation analysis, data visualization (e.g., histograms, scatter plots, and heatmaps), and summary statistics.

5. **Data Modeling**: Data modeling entails constructing mathematical models or statistical algorithms to examine datasets and derive meaningful insights. Typical modelling

techniques encompass regression analysis, classification algorithms such as decision trees and support vector machines, clustering methods like k-means and hierarchical clustering, as well as predictive modeling.

6. **Data Evaluation**: Data evaluation assesses the performance and accuracy of the analytical models or hypotheses generated during the modeling phase. Evaluation metrics vary depending on the type of analysis, but commonly include measures such as accuracy, precision, recall, F1-score, and confusion matrix.

7. **Data Visualization**: Data visualization involves creating graphical representations of data to enhance comprehension and interpretation. Effective visualization techniques are crucial for conveying insights, trends, and patterns to stakeholders. Tools such as charts, graphs, dashboards, and interactive visualizations allow users to dynamically explore and interact with the data.

Tools and Techniques in Data Analysis

- **Descriptive Statistics**: Descriptive statistics summarize and explain the central tendency, dispersion, and distribution of data. Key measures, including mean, median, mode, variance, standard deviation, skewness, and kurtosis, offer valuable insights into the dataset's characteristics.

- **Inferential Statistics**: Inferential statistics infer or generalize findings from a sample to a population. Techniques such as hypothesis testing, confidence intervals, and regression analysis help make predictions, test hypotheses, and estimate population parameters based on sample data.

- **Data Mining Techniques**: Data mining techniques are designed to uncover hidden patterns, relationships, and trends in large datasets. Common methods include clustering (such as k-means and hierarchical clustering), association rule mining (like the Apriori algorithm), anomaly detection, and text mining.

- **Machine Learning Algorithms**: Machine learning algorithms enable computers to learn from data and make predictions or decisions without explicit programming. Supervised learning algorithms (e.g., linear regression, logistic regression, decision trees, neural networks) learn from labeled data, while unsupervised learning algorithms (e.g., k-means clustering, principal component analysis) uncover hidden structures in unlabeled data.

Statistical Analytics Concepts

Descriptive Statistics

Descriptive statistics are essential for summarizing and describing the main features of a dataset. They provide valuable insights into the central tendency, variability, and distribution of the data.

- **Measures of Central Tendency**: Measures such as the mean, median, and mode indicate the central or typical value of a dataset. The mean is the arithmetic average, the median is the middle value when the data is ordered, and the mode is the value that appears most frequently.
- **Measures of Dispersion**: Measures such as range, variance, and standard deviation quantify the spread or variability of the data. The range is the difference between the maximum and minimum values, while variance and standard deviation measure the average deviation of data points from the mean.
- **Frequency Distribution**: Frequency distribution illustrates the occurrences of each value or range of values within a dataset, offering insights into its distributional characteristics and aiding in the identification of outliers or unusual patterns..
- **Histograms and Box Plots**: Histograms and box plots are graphical representations that depict the distribution of data. Histograms show the frequency of data values within predefined intervals or bins, while box plots summarize the distribution using quartiles, median, and outliers.

Inferential Statistics

Inferential statistics enable researchers to draw conclusions or make predictions about a population based on sample data. These techniques help generalize findings from a sample to a larger population with a certain level of confidence.

- **Probability Distributions**: Probability distributions describe the likelihood of observing different outcomes in a random experiment. Common probability distributions include the normal distribution, which is symmetric and bell-shaped, and the binomial distribution, which models the number of successes in a fixed number of independent trials.
- **Sampling Techniques**: Sampling techniques are employed to select representative samples from a population for analysis. Common methods include random sampling, stratified sampling, cluster sampling, and systematic sampling, which help ensure the sample's validity and minimize bias.
- **Estimation and Confidence Intervals**: Estimation techniques, including point estimation and interval estimation, offer estimates of population parameters like the mean or proportion, derived from sample data. Confidence intervals gauge the

uncertainty linked with the estimate and furnish a range within which the true population parameter is expected to fall.

- **Hypothesis Testing**: Hypothesis testing is a pivotal aspect of inferential statistics, enabling researchers to draw conclusions about population parameters from sample data. It encompasses formulating null and alternative hypotheses, determining a significance level, selecting an appropriate test statistic, executing the test, and interpreting the outcomes.

Hypothesis Testing

Introduction to Hypothesis Testing

Hypothesis testing is a methodical procedure employed to draw statistical inferences about population parameters using sample data. It encompasses formulating null and alternative hypotheses, selecting an appropriate test statistic, establishing the significance level, conducting the test, and interpreting the findings..

Steps in Hypothesis Testing

1. **Formulating the Hypotheses**: The null hypothesis (H0) represents the default assumption or status quo, while the alternative hypothesis (H1) represents the researcher's claim or alternative viewpoint. These hypotheses are crafted based on the research question and the study's specific objective.

2. **Selecting the Significance Level**: The significance level (α), also known as the level of significance or alpha, determines the probability of rejecting the null hypothesis when it's true. Commonly used significance levels include $\alpha = 0.05$ and $\alpha = 0.01$, representing a 5% and 1% chance of committing a Type I error, respectively.

3. **Choosing the Test Statistic**: The selection of the test statistic depends on the data's nature and the hypotheses under examination. Common test statistics encompass t-tests, z-tests, chi-square tests, F-tests, and ANOVA. Accurately selecting the test statistic is pivotal for assessing the evidence against the null hypothesis.

4. **Collecting Data and Calculating the Test Statistic**: Data is gathered via sampling, and the test statistic is computed using the sample data and the chosen hypothesis test. This statistic quantifies the degree of deviation between the observed data and the null hypothesis, offering evidence for or against the null hypothesis.

5. **Making a Decision**: Based on the calculated test statistic and the significance level, a decision is made to either reject or fail to reject the null hypothesis. If the p-value (probability value) associated with the test statistic is less than the significance level,

the null hypothesis is rejected, indicating evidence in favor of the alternative hypothesis. If the p-value is greater than the significance level, the null hypothesis is not rejected.

Types of Hypothesis Tests

- **One-Sample t-test**: A one-sample t-test is utilized to compare the mean of a single sample to a known value or a hypothesized population mean. It evaluates whether there's a statistically significant difference between the sample mean and the population mean
- **Two-Sample t-test**: The two-sample t-test contrasts the means of two independent samples to ascertain if there's a statistically significant difference between them. It's commonly employed to compare the means of two groups or populations..
- **Paired t-test**: A paired t-test compares the means of two related samples, such as before and after measurements or paired observations. It determines whether there's a significant difference between the paired observations..
- **Chi-Square Test**: The chi-square test is a non-parametric test employed to examine the association between categorical variables. It establishes whether there's a significant relationship between the observed frequencies and the expected frequencies in a contingency table.
- **ANOVA (Analysis of Variance)**: ANOVA is used to analyze the differences among group means in a dataset with more than two groups. It assesses whether there are statistically significant differences between the means of multiple groups, considering the within-group variability and the between-group variability.

Regression and its Types

Introduction to Regression Analysis

Regression analysis is a statistical method utilized to model the relationship between one or more independent variables (predictors) and a dependent variable (response). It aids in predicting the value of the dependent variable based on the values of the independent variables. This technique finds extensive application across diverse fields such as economics, finance, healthcare, and social sciences, serving purposes like forecasting, modeling, and hypothesis testing.

Simple Linear Regression

Simple linear regression is the simplest form of regression analysis that involves a single independent variable and a single dependent variable. The relationship between the variables is modeled using a linear equation of the form:

$y = \beta_0 + \beta_1 x + \varepsilon$

Where:

- y is the dependent variable.
- x is the independent variable.
- β_0 is the intercept (the value of y when $x = 0$).
- β_1 is the slope (the change in y for a one-unit change in x).
- ε is the error term representing random variation or unexplained factors.

The coefficients β_0 and β_1 are estimated from the data using the method of least squares, which minimizes the sum of squared differences between the observed and predicted values of y.

Multiple Linear Regression

Multiple linear regression extends simple linear regression to model the relationship between a dependent variable and multiple independent variables. The relationship is expressed by the equation:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$

Where:

- y is the dependent variable.
- $x_1, x_2, \ldots, x_n$ are the independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the independent variables.
- ε is the error term.

Multiple linear regression allows for modelling complex relationships and capturing the combined effect of multiple predictors on the dependent variable.

Types of Regression Analysis

Regression Type	Description
Simple Linear Regression	Involves one independent variable and one dependent variable.
Multiple Linear Regression	Involves multiple independent variables and one dependent variable.

Polynomial Regression	Fits a nonlinear relationship between the independent and dependent variables using polynomial terms.
Logistic Regression	Used for predicting the probability of a binary outcome.
Ridge Regression	Addresses multicollinearity by adding a penalty term to the regression coefficients.
Lasso Regression	Performs variable selection and regularization to improve the model's accuracy.

Correlation

Introduction to Correlation

Correlation measures the strength and direction of the linear relationship between two continuous variables. It quantifies how changes in one variable are associated with changes in another variable. Correlation analysis helps identify patterns, dependencies, and associations between variables.

Types of Correlation

- **Positive Correlation**: A positive correlation exists when an increase in one variable is associated with an increase in the other variable, and a decrease in one variable is associated with a decrease in the other variable. The correlation coefficient ranges from 0 to +1, where +1 indicates a perfect positive correlation.
- **Negative Correlation**: A negative correlation exists when an increase in one variable is associated with a decrease in the other variable, and vice versa. The correlation coefficient ranges from -1 to 0, where -1 indicates a perfect negative correlation.
- **Zero Correlation**: Zero correlation indicates no linear relationship between the variables. The correlation coefficient is close to 0, suggesting that changes in one variable are not associated with changes in the other variable.

Pearson Correlation Coefficient

The Pearson correlation coefficient, denoted by r, measures the strength and direction of the linear relationship between two continuous variables. It is calculated using the formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual data points.
- $\bar{x}$ and $\bar{y}$ are the means of the variables x and y, respectively.

The Pearson correlation coefficient ranges from -1 to +1, where:

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

Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient, denoted by ρ (rho), measures the strength and direction of the monotonic relationship between two variables. It is calculated based on the ranks of the data points rather than their actual values, making it suitable for ordinal or nonnormally distributed data.

Spearman's rank correlation coefficient ranges from -1 to +1, where:

- $\rho = +1$: Perfect positive monotonic correlation
- $\rho = -1$: Perfect negative monotonic correlation
- $\rho = 0$: No monotonic correlation

ANOVA (Analysis of Variance)

Introduction to ANOVA

ANOVA, or Analysis of Variance, is a statistical method employed to examine the distinctions among group means within a dataset comprising more than two groups. It juxtaposes the means of multiple groups to ascertain if there exist statistically noteworthy differences among them. ANOVA evaluates both within-group variability and between-group variability to infer whether the disparities in means stem from chance fluctuations or genuine group disparities.

One-Way ANOVA

One-Way ANOVA is the basic form of ANOVA, comprising a single categorical independent variable (factor) with two or more levels (groups) and a continuous dependent variable. It

scrutinizes the null hypothesis asserting that the means of all groups are equal, juxtaposed with the alternative hypothesis indicating that at least one group mean differs.

Hypotheses in One-Way ANOVA

- Null Hypothesis (H0): The means of all groups are equal.
- Alternative Hypothesis (H1): At least one group mean is different.

Calculation of F-Statistic

The F-statistic in ANOVA measures the ratio of between-group variability to within-group variability. It is calculated as the ratio of the mean square between (MSB) to the mean square within (MSW):

$$F = \frac{MSB}{MSW}$$

Where:

- MSB = Sum of squares between (SSB) divided by degrees of freedom between (dfB)
- MSW = Sum of squares within (SSW) divided by degrees of freedom within (dfW)

If the calculated F-statistic is greater than the critical value from the F-distribution at a given significance level (α), the null hypothesis is rejected, indicating that there are significant differences among the group means.

Post Hoc Tests

When the null hypothesis in ANOVA is rejected, post hoc tests are employed to determine which specific groups exhibit significant differences from each other. Common post hoc tests include Tukey's HSD (Honestly Significant Difference), Bonferroni correction, Scheffe's method, and Dunnett's test. These tests help to pinpoint the specific group or groups that contribute to the observed differences identified by ANOVA.

Two-Way ANOVA

Two-Way ANOVA expands the analysis to encompass two categorical independent variables (factors) and their potential interaction effect on a continuous dependent variable. It evaluates not only the main effects of each factor but also their interaction effect. This allows for a more comprehensive understanding of how the two factors influence the dependent variable and whether their combined effect differs from what would be expected based solely on their individual effects.

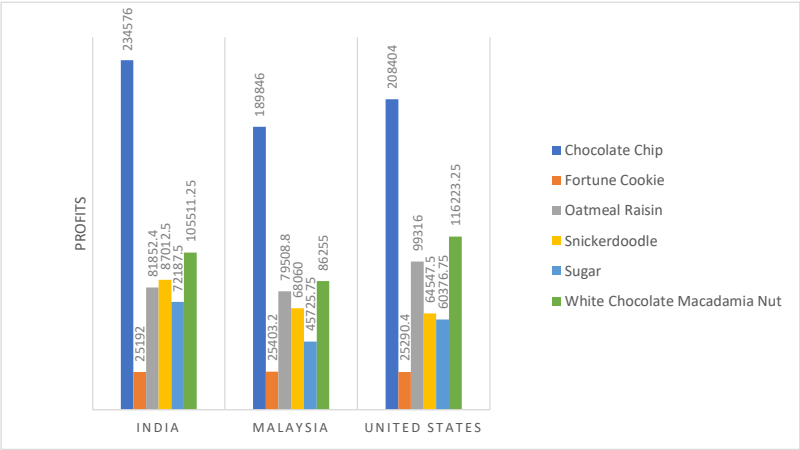
Interaction Effects

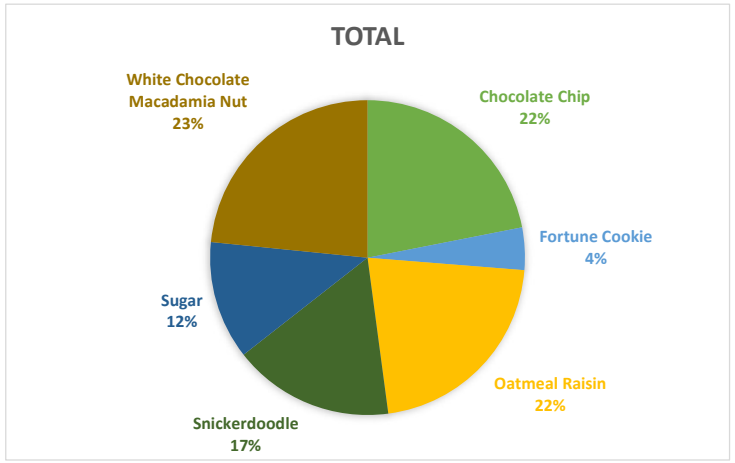
Interaction effects occur when the effect of one independent variable on the dependent variable depends on the level of another independent variable. Two-Way ANOVA allows for the examination of interaction effects between factors.

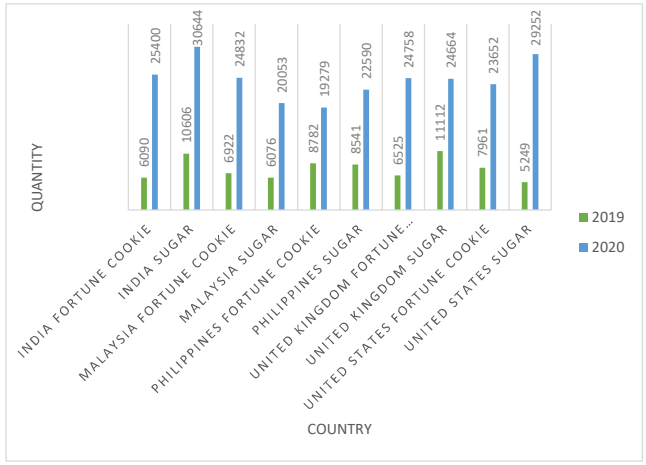
Interpretation of Results

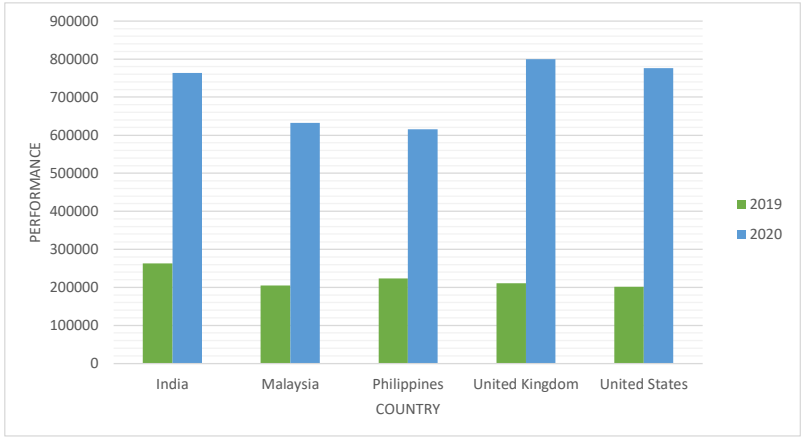
In ANOVA, if the null hypothesis is rejected, it indicates that there are significant differences among the group means. Post hoc tests help identify which specific groups differ from each other. If the null hypothesis is not rejected, it suggests that there are no significant differences among the group means.

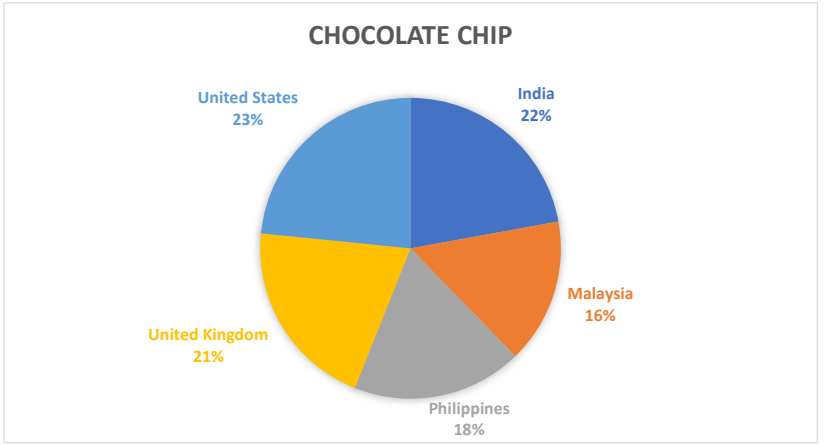
Cookie Data Report



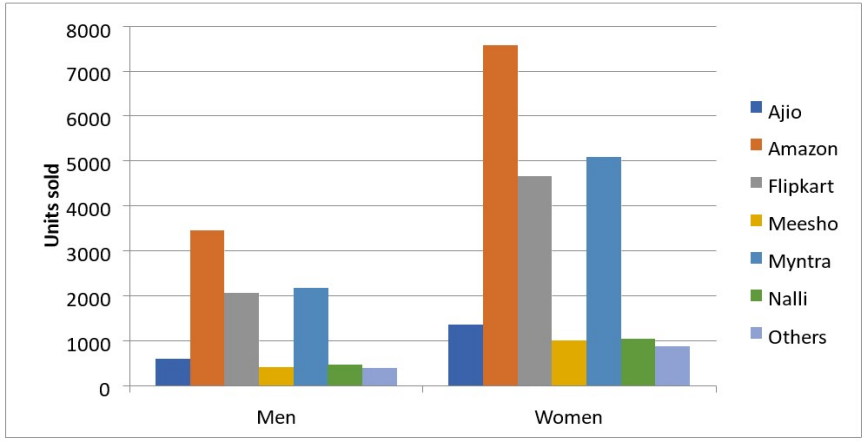


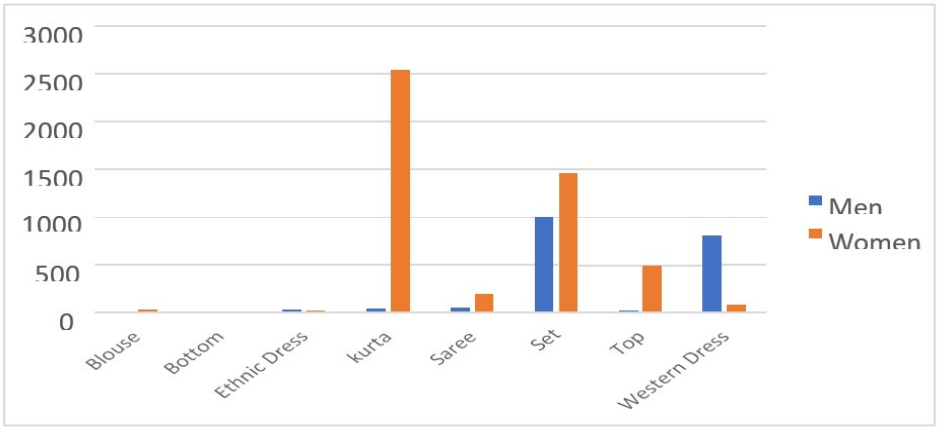


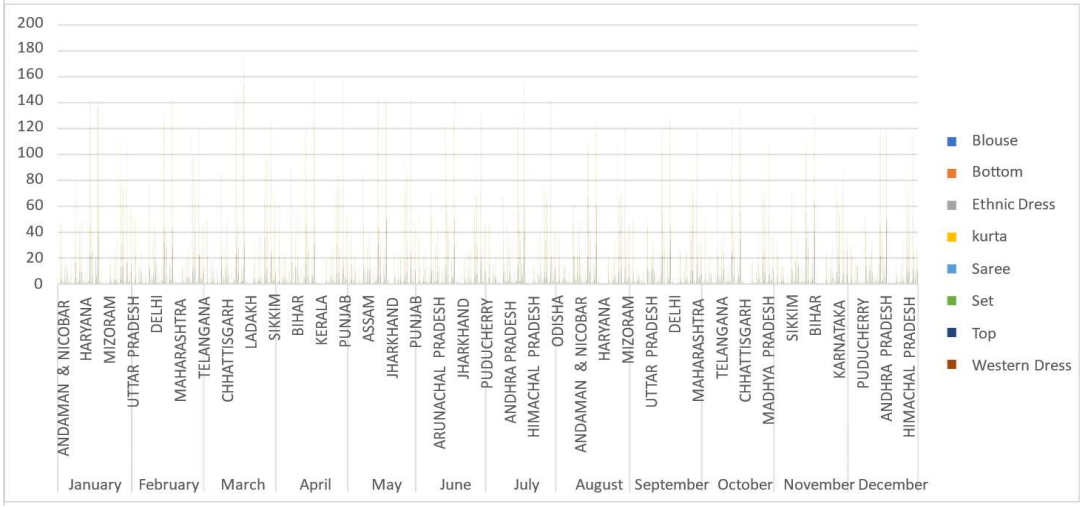


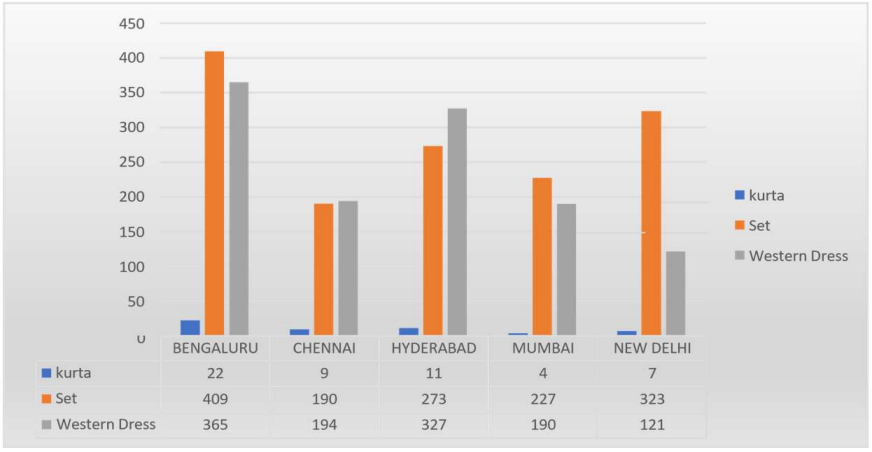


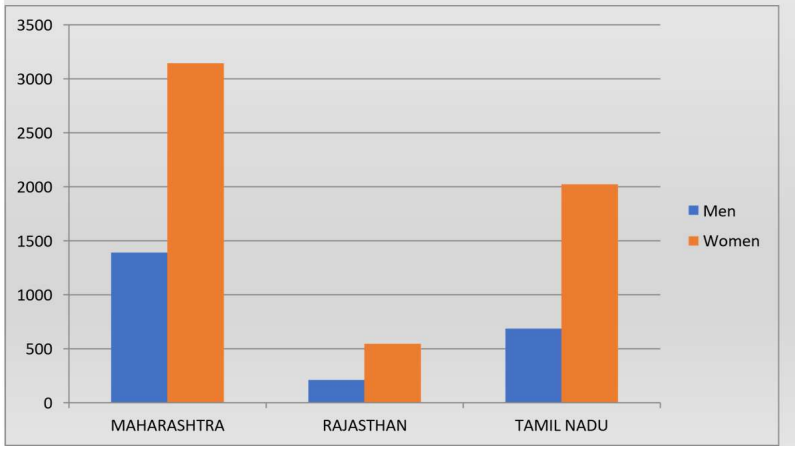
Store Data Report



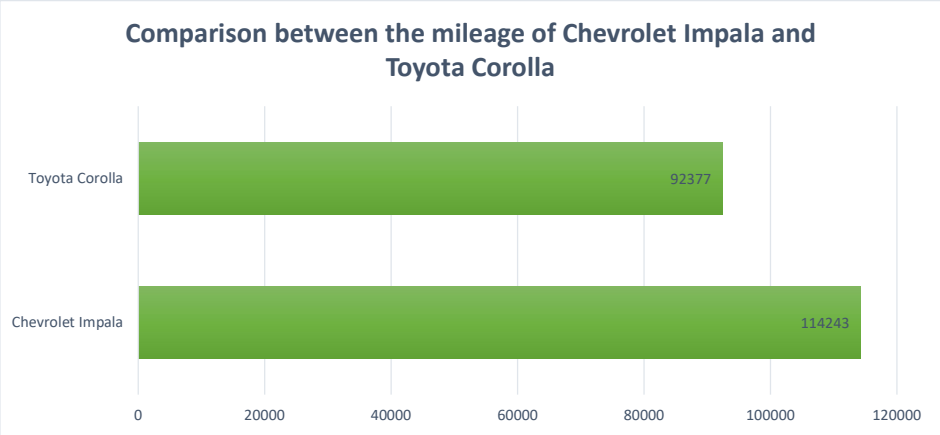


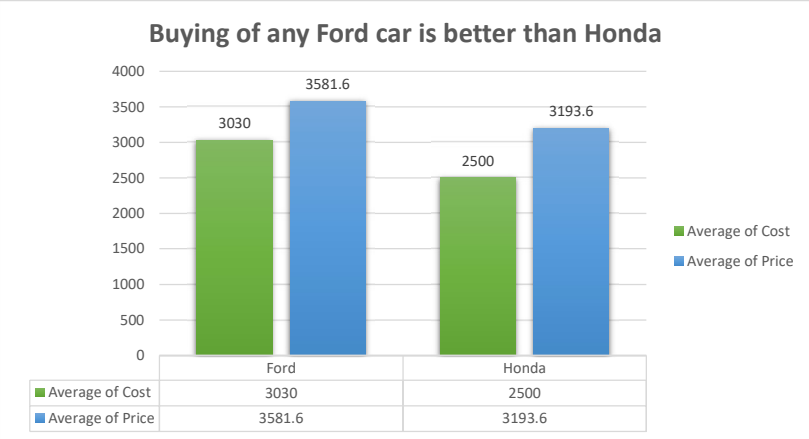


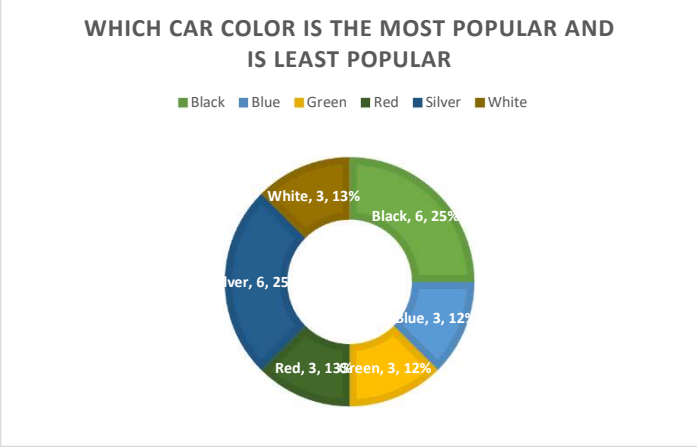


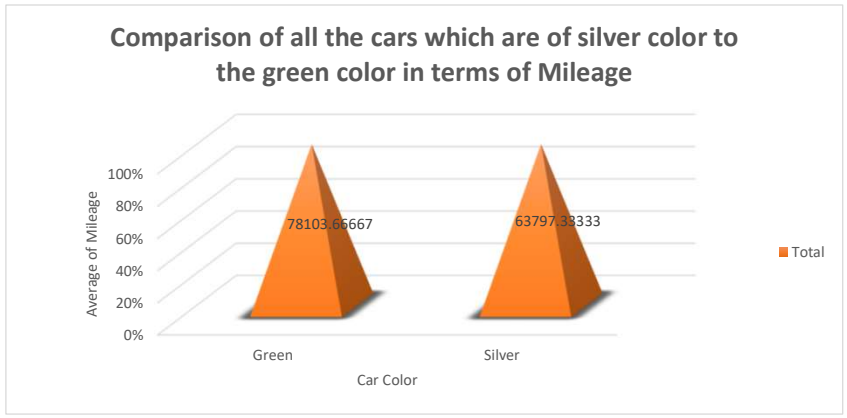


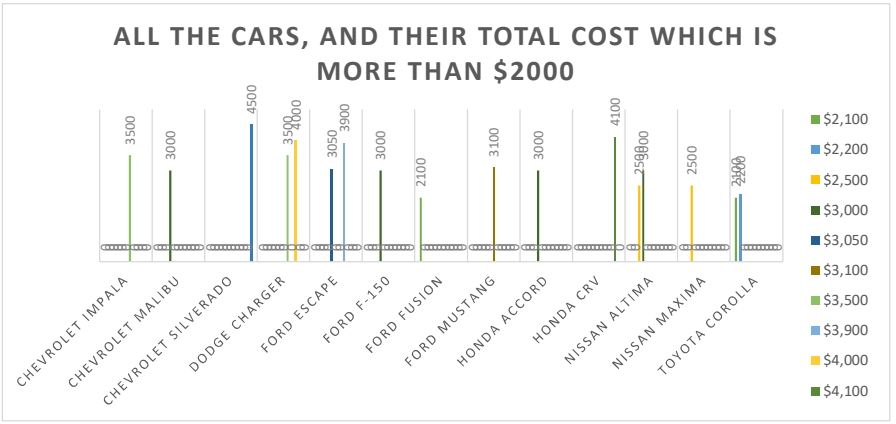
Car Collection Report



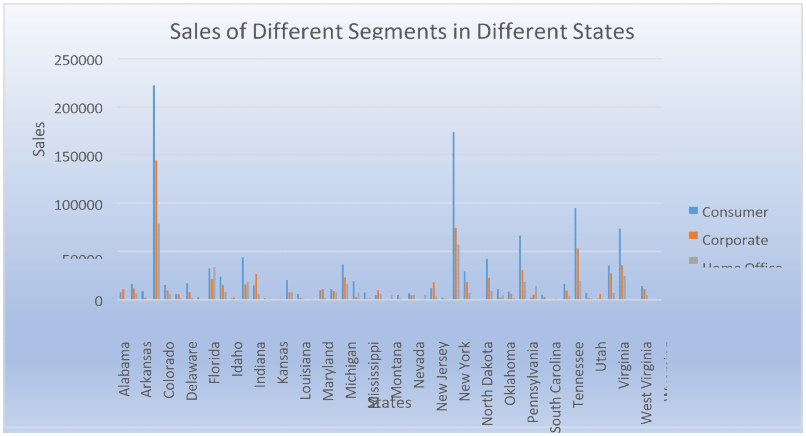


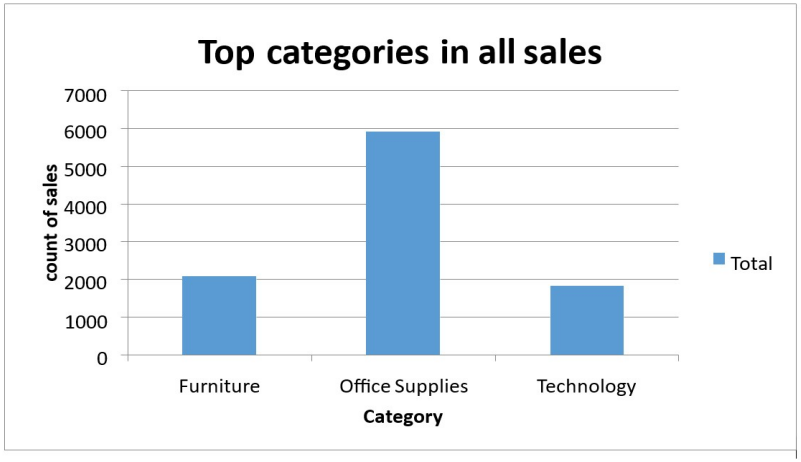


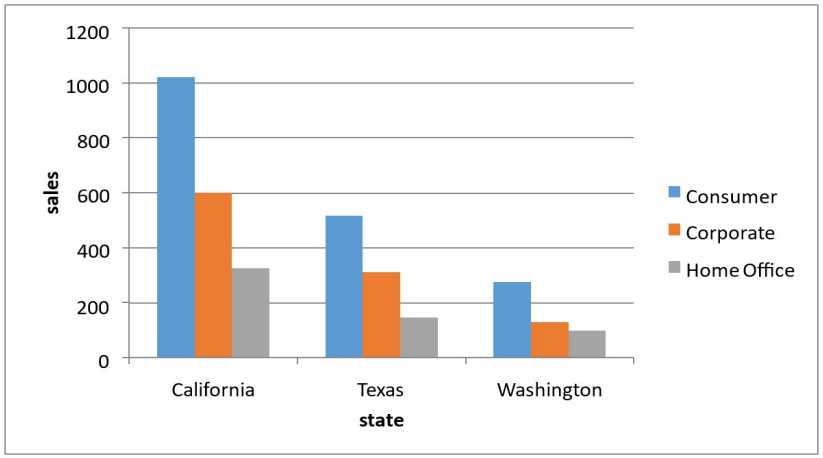


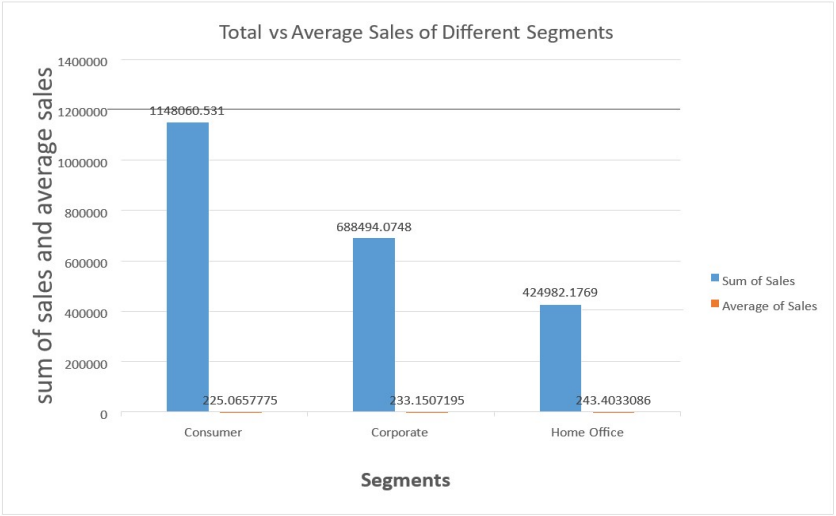


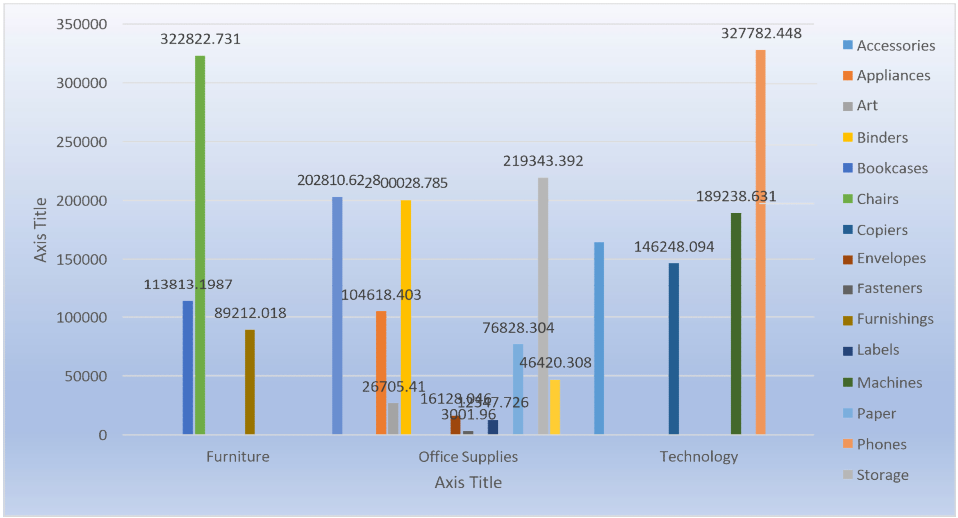
Examining Sales by Sector in the United States



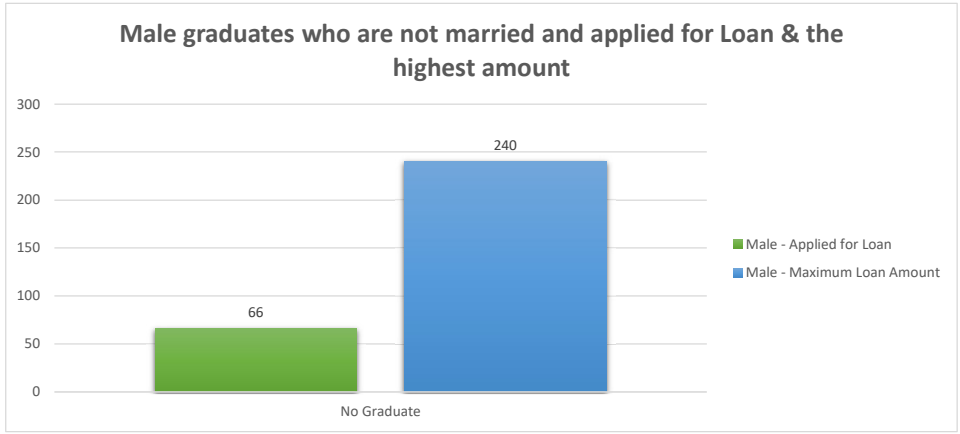


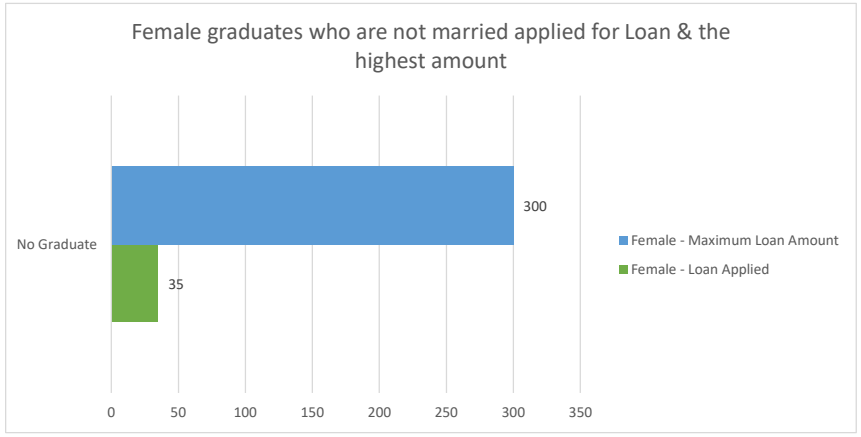


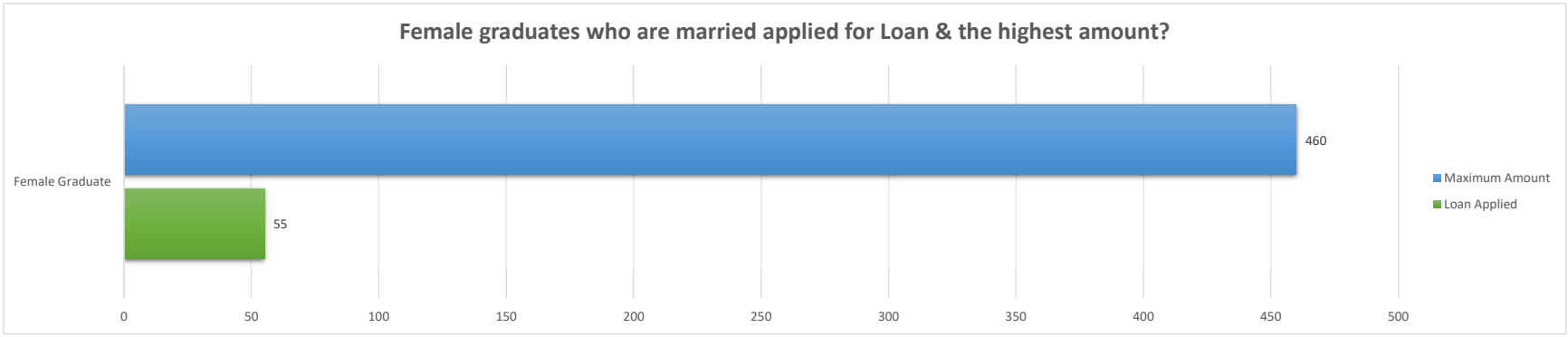


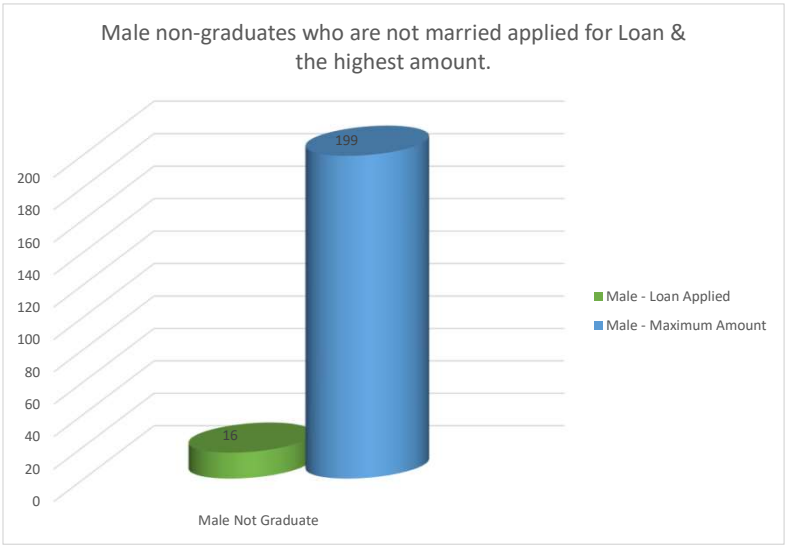


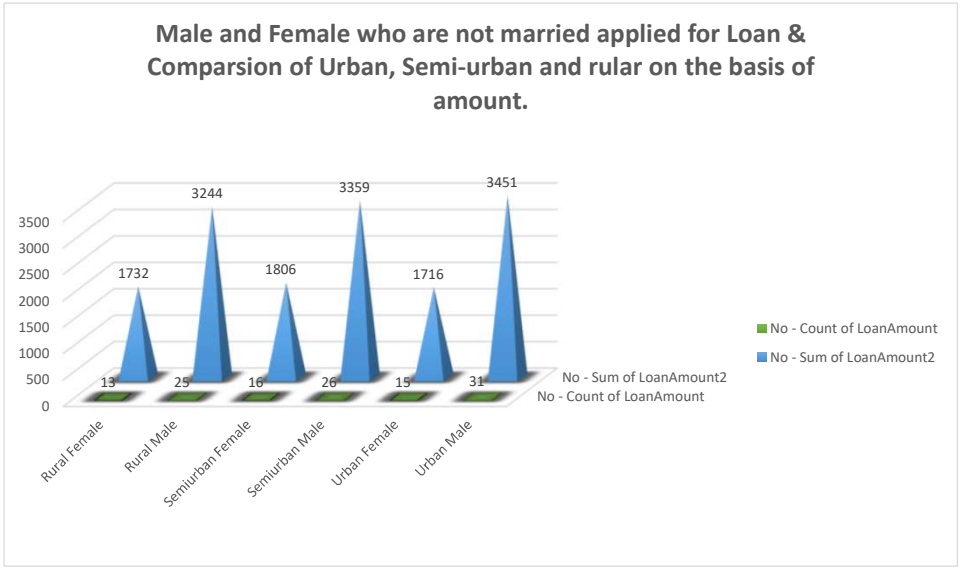
Loan Data Report



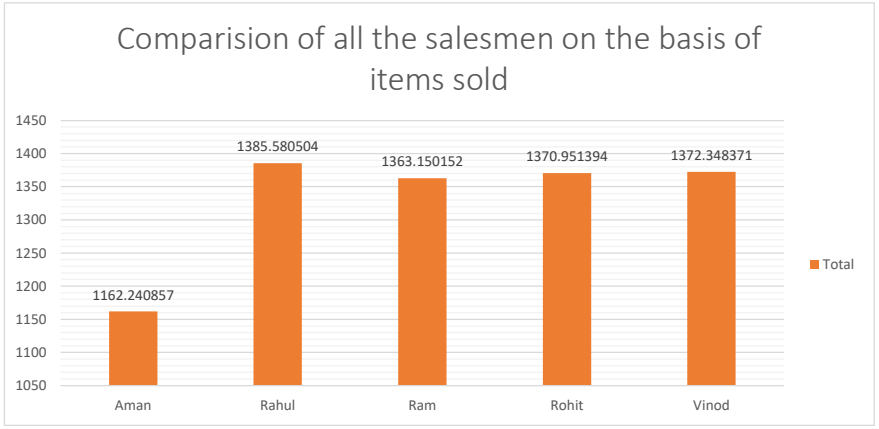


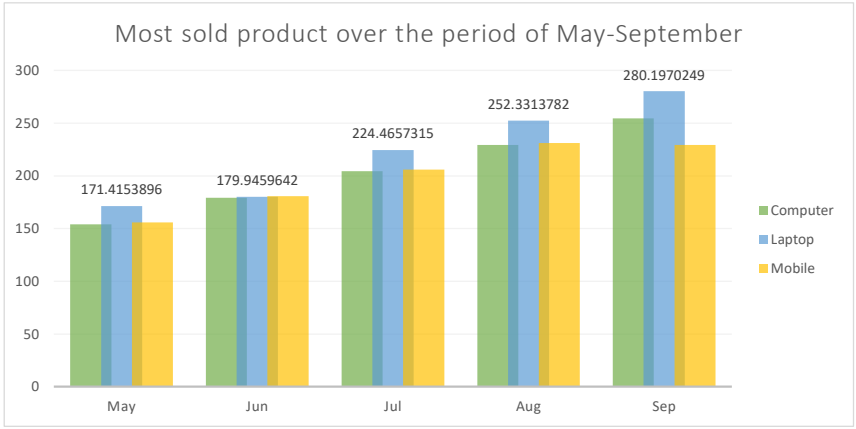


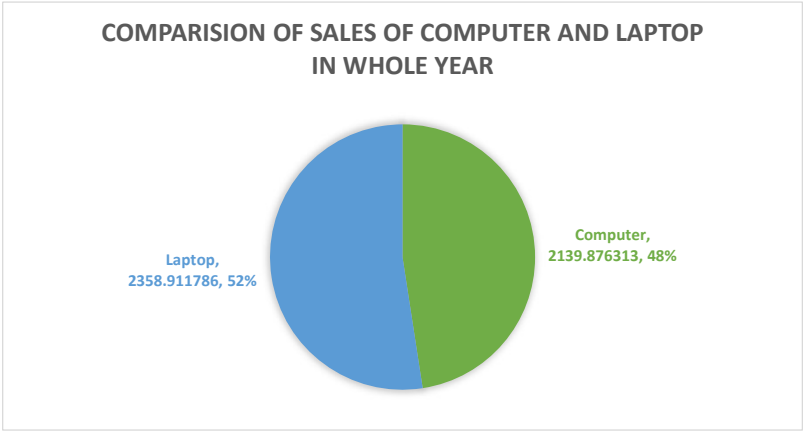


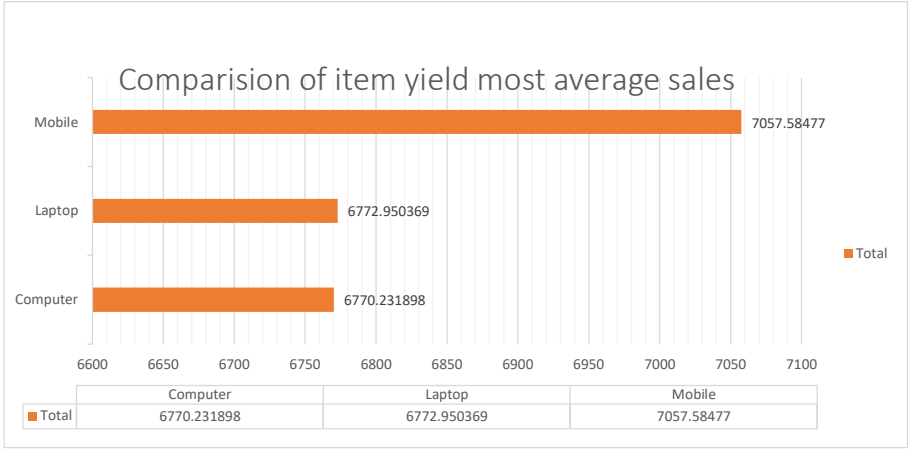


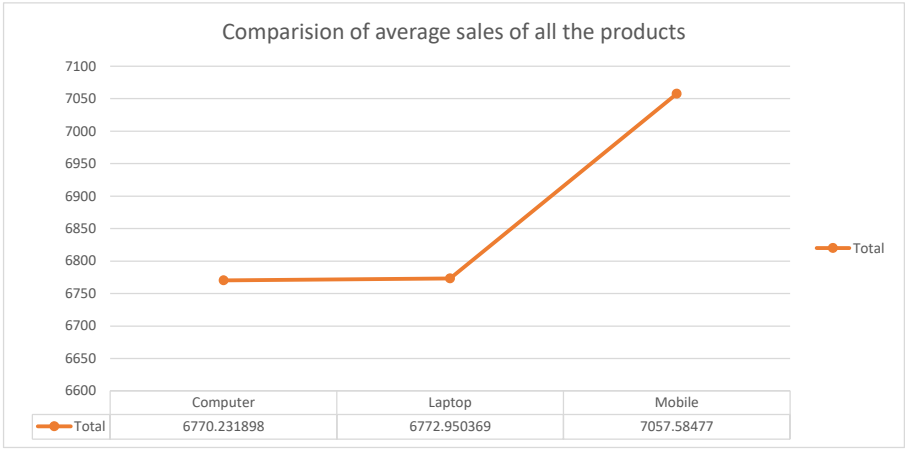
Shop Sales Data Report



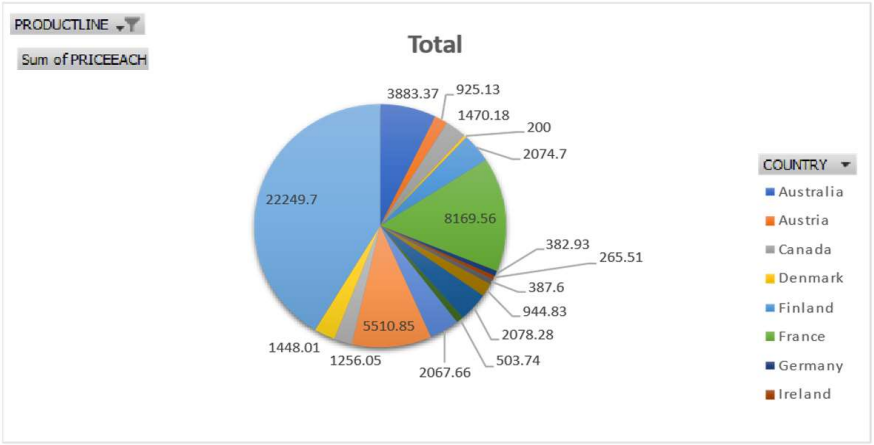


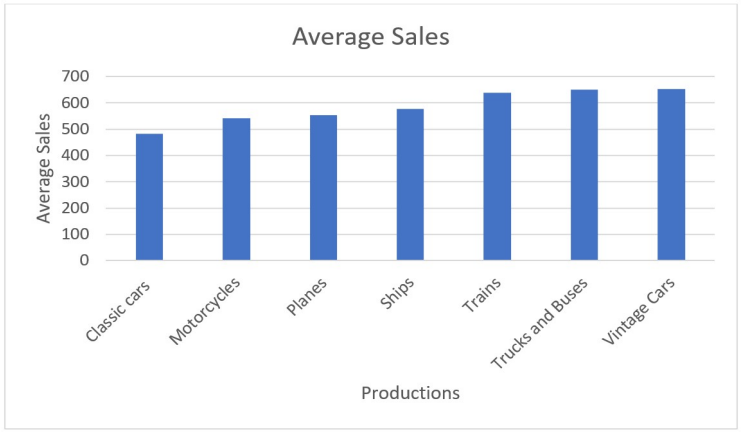


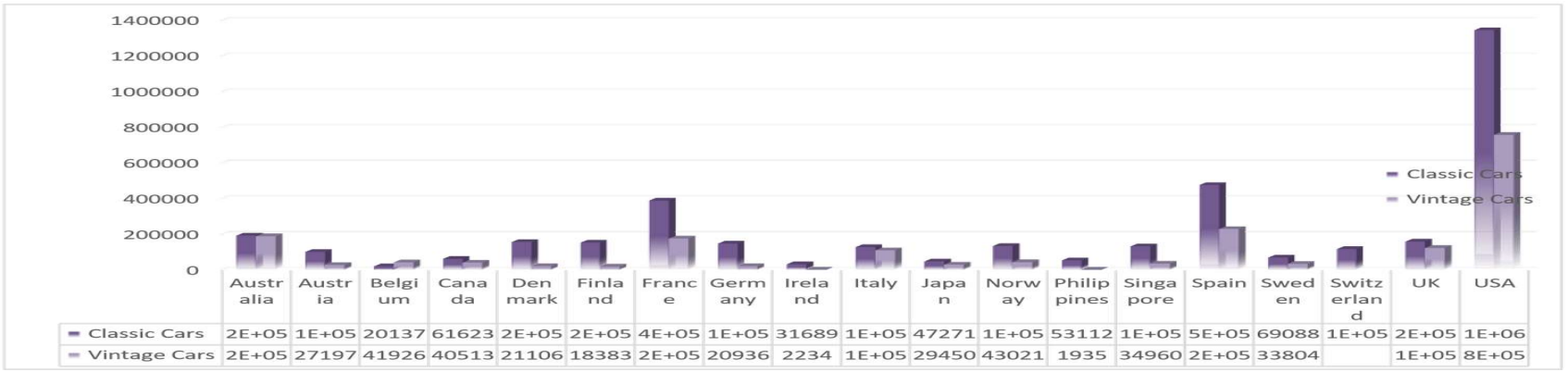


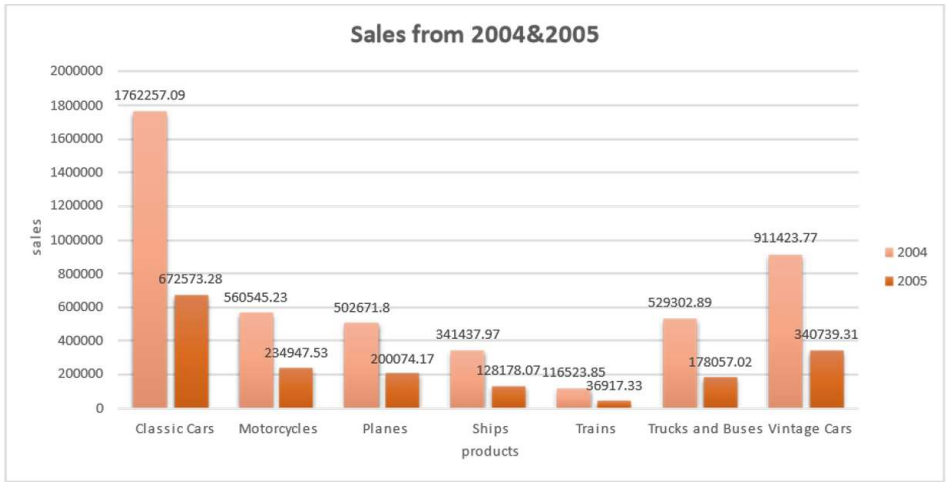


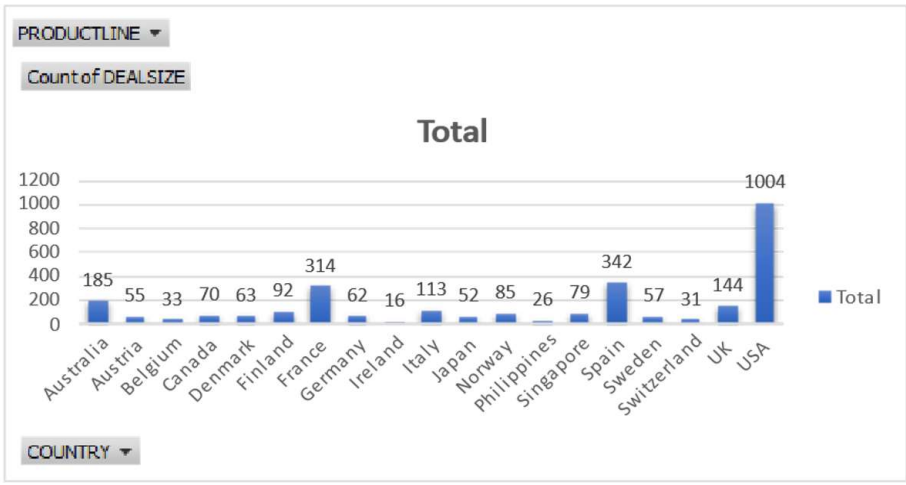
Sales Data Samples Report











Cookie Data Report

Introduction:-

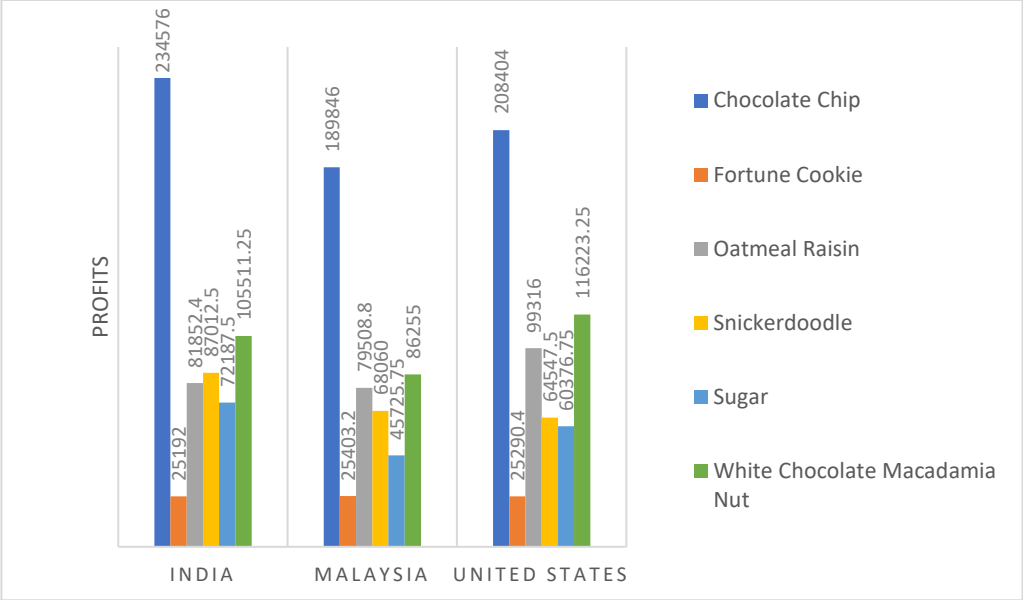
The purpose of this report is to analyse the sales data of various cookie types across different countries for the years 2019 and 2020. The dataset provides insights into revenue, profit, quantity sold, and pricing information for each cookie type and country. Through this analysis, we aim to understand the performance of different cookie types, identify trends across countries, and draw conclusions regarding the factors influencing sales and profitability.

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:-

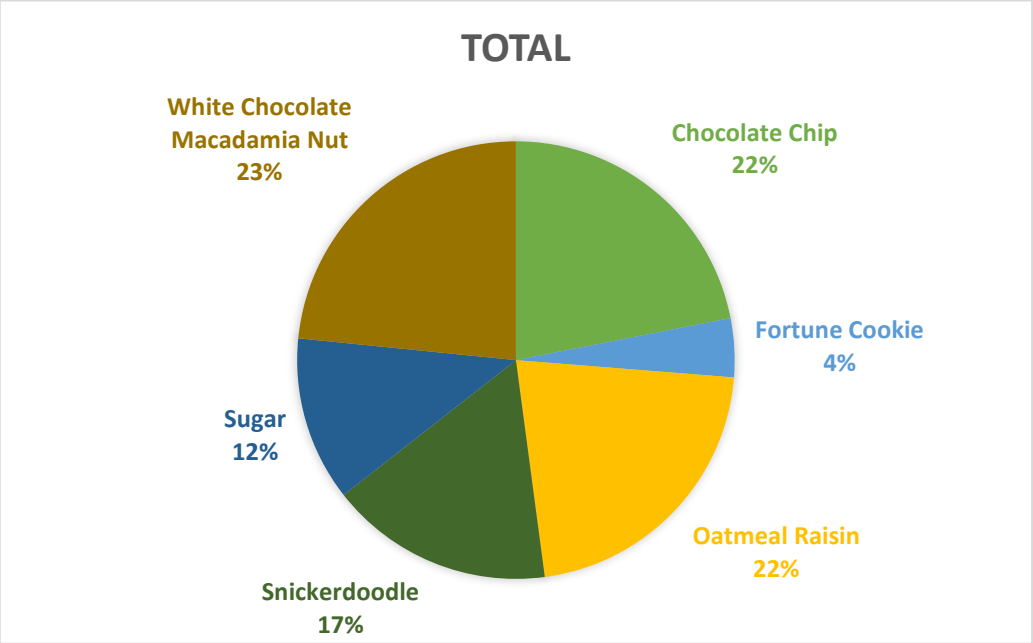
1. Compare the profit earn by all cookie types in US, Malaysia and India.



Countries

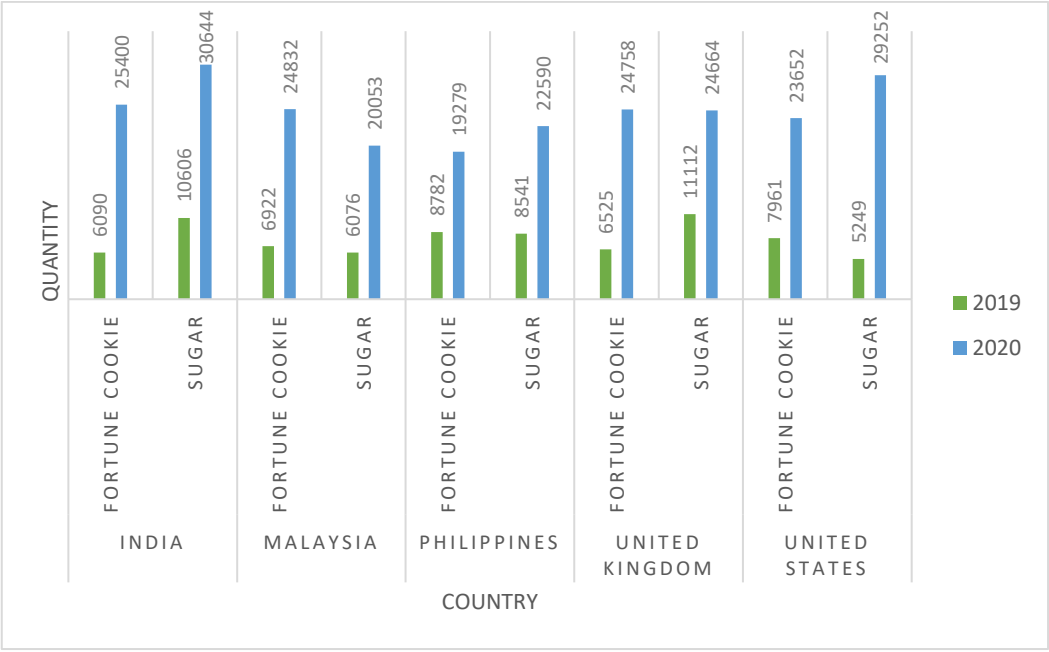
Ans: When compared to other treats like fortune cookies, oatmeal raisin cookies, and white chocolate macadamia, chocolate chips yield larger profits in these countries.

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



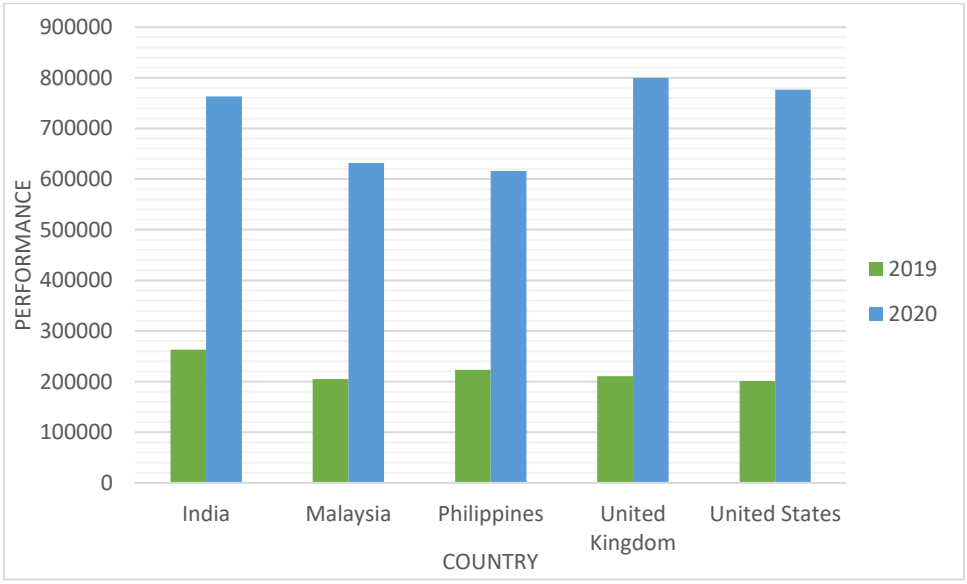
Ans: Based on the above figure, oatmeal is the cookie with the second-highest revenue generation, followed by white chocolate macadamia nut cookies, which earn the highest average revenue of all the cookies.

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



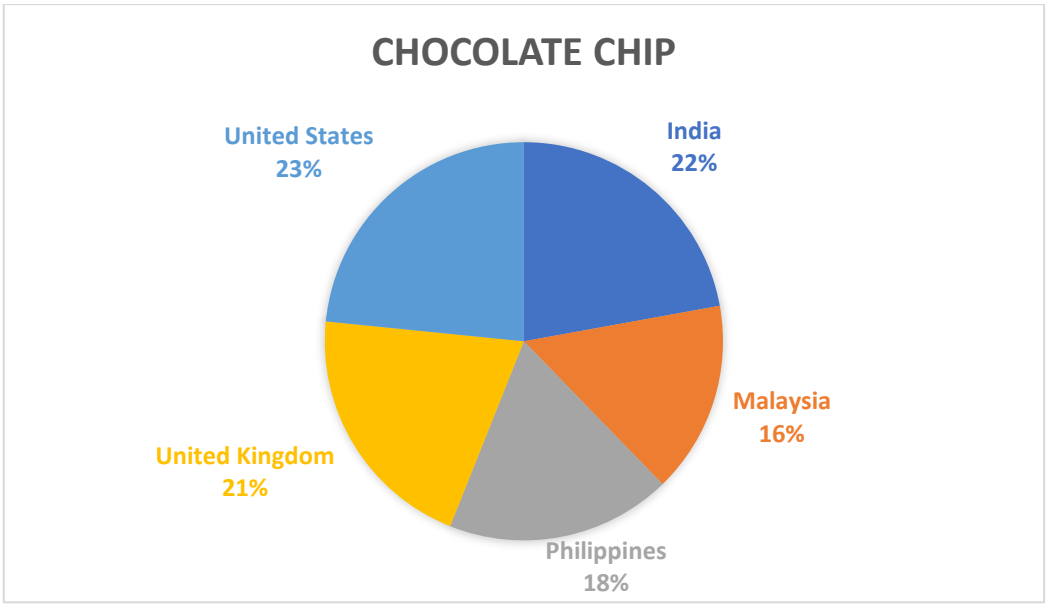
Ans: According to the graph above, the United States ranked second in terms of sales of sugar and fortune cookies in 2020, followed by India in the United Kingdom and the Philippines in 2019.

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



Ans: United Kingdom performed the best out of all the nations in 2020, with India taking the lead in 2019.

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



Ans: India was leading at position 22%, while chocolate chip cookies sold for the highest price of 23% total in the United States for the biggest profit made.

Conclusion and Reviews: -

In summary, analyzing cookie sales data has provided invaluable insights into consumer preferences, industry trends, and profitability across many countries and cookie types. We were able to gain a complete understanding of the factors affecting sales success by analyzing data on revenue, profit, quantity sold, and price. Thanks to this analysis, we have been able to increase profitability and better satisfy customer requests by finding growth potential, refining product offers, and adjusting marketing strategies. Future study and adaptation based on these insights will be necessary to keep a competitive edge in the rapidly changing cookie market. Taking everything into account, the meticulous examination of sales data has proven essential in directing strategic decisions and ensuring the long-term sustainability of our cookie business.

Regression:

The regression model, with a significant p-value ($p < 0.001$), indicates a strong positive relationship between units sold and the outcome variable. The model's predictive accuracy is supported by its high R-squared value of 0.688, suggesting that approximately 68.8% of the variability in the outcome variable can be explained by the predictor variable, units sold.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.829304
R Square	0.687746
Adjusted R Square	0.687298
Standard Error	1462.76
Observations	700

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.29E+09	3.29E+09	1537.356	1.4E-178
Residual	698	1.49E+09	2139668		
Total	699	4.78E+09			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-74.4103	116.5304	-0.63855	0.523326	-303.202	154.3817	-303.202	154.3817
Units Sold	2.500792	0.063781	39.20914	1.4E-178	2.375567	2.626017	2.375567	2.626017

Co-relation:

The correlation coefficient between units sold and revenue is 0.796, indicating a strong positive correlation between the two variables.

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

Anova (Single Factor) :

The AN VA results indicate a significant difference between the two groups ($p < 0.001$), with 1 degree of freedom. The within-group error is 7681356717, and the total R-squared value is 0.06, suggesting that the model explains 6% of the variability in the data.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
3450	699	1923505	2751.795	4154648
5175	699	2758189	3945.908	6850161

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4.98E+08	1	4.98E+08	90.57022	7.53E-21	3.848129
Within Groups	7.68E+09	1396	5502405			
Total	8.18E+09	1397				

Anova two factor without Replication:

The AN VA results reveal significant variation among rows and columns ($p < 0.001$), with degrees of freedom (df) values of 48 and 3, respectively. The error term has a degree of freedom of 144.

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Rows	8.21E+08	48	17108242	5.848894	8.54E-17	1.445925

Columns	5.65E+10	3	1.88E+10	6435.486	3.8E-153	2.667443
Error	4.21E+08	144	2925039			
Total	5.77E+10	195				

Anova two factor with Replication:

The AN VA results show that there is a significant difference among the samples, columns, and their interaction, with p-values less than 0.001. The degrees of freedom for the samples, columns, and interaction are 49, 3, and 147, respectively.

Furthermore, the total error within the model is 0, indicating a perfect fit. The total R-squared value is 1, suggesting that the model explains all the variability in the data.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	8.55E+08	49	17443674	65535	#NUM!	#NUM!
Columns	5.78E+10	3	1.93E+10	65535	#NUM!	#NUM!
Interaction	4.39E+08	147	2983765	65535	#NUM!	#NUM!
Within	0	0	65535			
Total	5.91E+10	199				

Descriptive Statistics:

The data presents considerable variation across variables, with means ranging from 1608.15 to 43949.81. Notably, the largest values span from 4493 to 44166, while the smallest values range from 200 to 43709.

1725		*8625*		*3450*		*5175*
Mean	1608.153	Mean	6697.702	Mean	2751.795	Mean
Standard Error	32.83303	Standard Error	174.9955	Standard Error	77.09541	Standard Error
Median	1540	Median	5868	Median	2422.2	Median
Mode	727	Mode	8715	Mode	3486	Mode
Standard Deviation	868.0597	Standard Deviation	4626.638	Standard Deviation	2038.295	Standard Deviation
Sample Variance	753527.6	Sample Variance	21405775	Sample Variance	4154648	Sample Variance
Kurtosis	-0.31828	Kurtosis	0.463405	Kurtosis	0.807696	Kurtosis
Skewness	0.436551	Skewness	0.869254	Skewness	0.931429	Skewness
Range	4293	Range	23788	Range	10954.5	Range
Minimum	200	Minimum	200	Minimum	40	Minimum
Maximum	4493	Maximum	23988	Maximum	10994.5	Maximum
Sum	1124099	Sum	4681694	Sum	1923505	Sum
Count	699	Count	699	Count	699	Count
Largest(1)	4493	Largest(1)	23988	Largest(1)	10994.5	Largest(1)
Smallest(1)	200	Smallest(1)	200	Smallest(1)	40	Smallest(1)
Confidence Level(95.0%)	64.46334	Confidence Level(95.0%)	343.5807	Confidence Level(95.0%)	151.3667	Confidence Level(95.0%)

Store Data Report

Introduction:

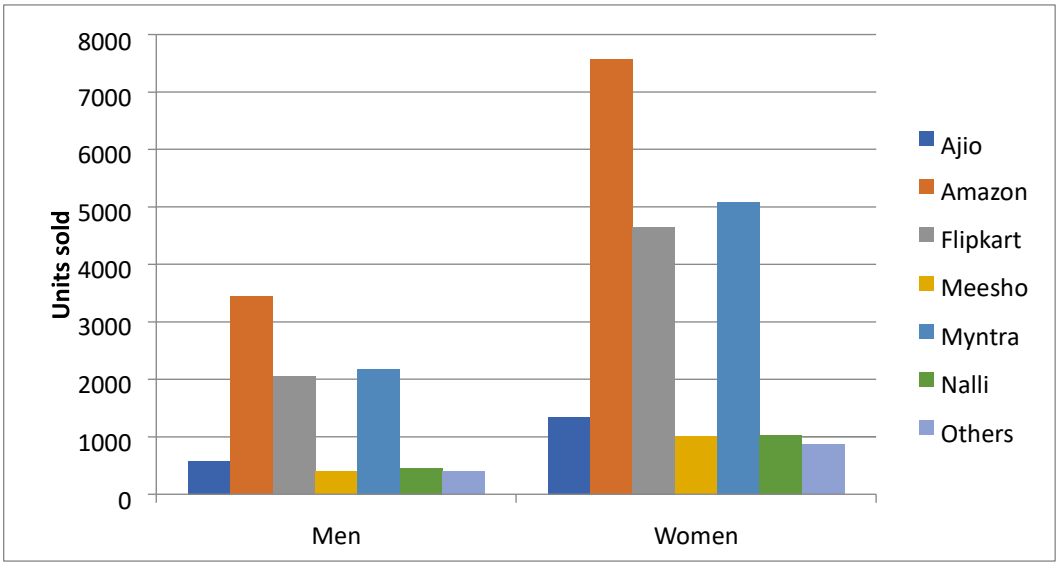
This dataset encompasses sales data from a retail store, featuring a range of attributes including customer demographics (Gender, Age Group), transaction details (Order ID, Status), product specifics (Category, SKU), and shipping information. With a focus on understanding customer behaviour and product trends, our analysis aims to uncover patterns, preferences, and correlations within the data. By leveraging these insights, businesses can optimize marketing efforts, enhance inventory management, and improve customer satisfaction.

Questionnaire:

1. which of the channel performed better than all other channels in compare men & women?
2. Compare category. Find out most sold category above 23 years of age for any gender.
3. Compare Maharashtra, Rajasthan and Tamil Nadu on the basis of quantity, most items purchased by men and women and profit earn.
4. Which city sold most of following categories:
    a. Kurta
    b. Set
    c. Western wears
5. In which month most items sold in any of the state on the basis of category.

Analytics:

1. which of the channel performed better than all other channels in compare men & women?



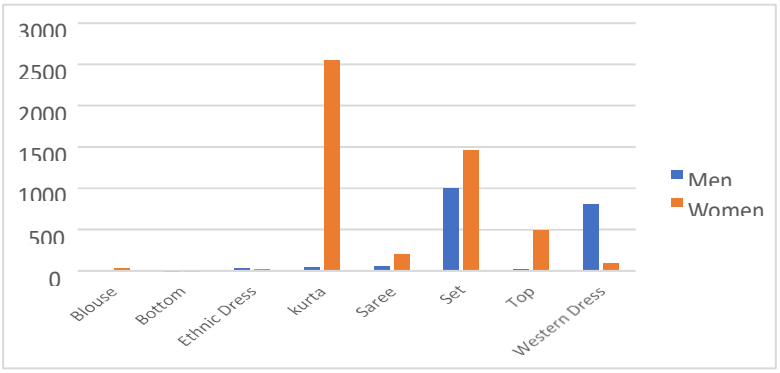
Ans: Amazon leads in sales for both genders, with Myntra and Flipkart following closely behind. In the women's category, Amazon sold almost 7,500 copies, and in the men's category, almost 3500 units. 2000 pieces were sold in Myntra's men's section.

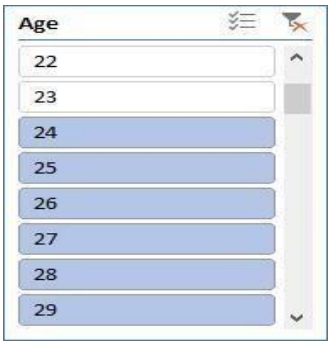
2. Compare category. Find out most sold category above 23 years of age for any gender.

The table of items sold is given below:

Item	Men	Women	Grand Total
Blouse	6	190	196
Bottom	40	28	68
Ethnic Dress	150	77	227
kurta	156	8820	8976
Saree	261	941	1202
Set	4365	6204	10569
Top	45	1825	1870
Western Dress	3078	380	3458
Grand Total	**8101**	**18465**	**26566**

The graph is as follows:

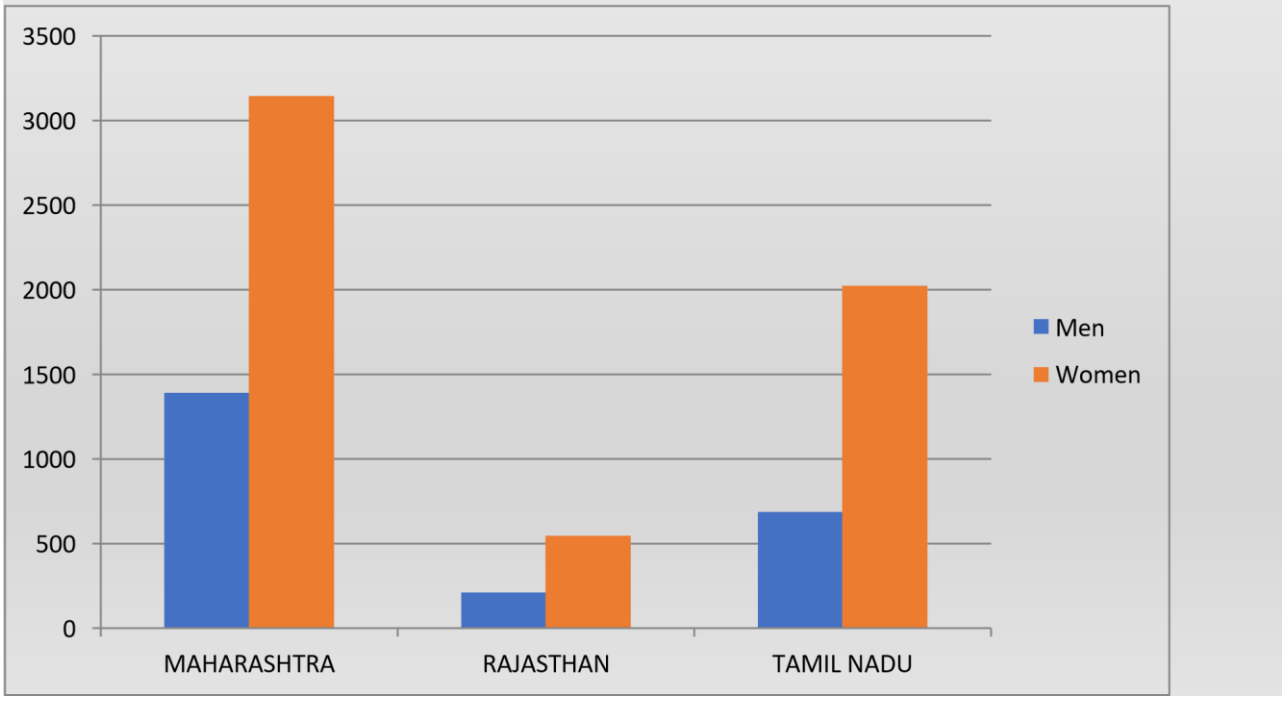




Ans: In the women's segment, kurtas are the most popular category, with 8820 items sold to customers over 23. Sets are the most popular category in the men's area and the second most popular category in the women's area, with 4365 units sold.

3. Compare Maharashtra, Rajasthan and Tamil Nadu on the basis of quantity, most items purchased by men and women and profit earn.

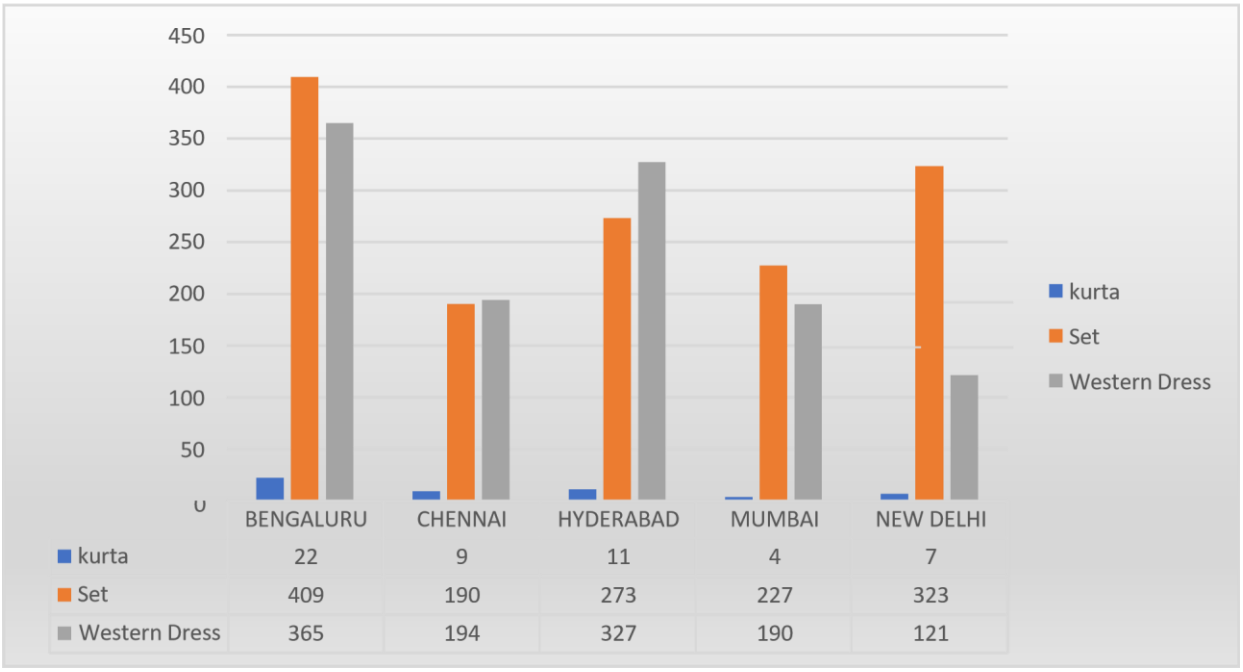
State	Men	Women	Grand Total
MAHARASHTRA	1390	3144	4534
RAJASTHAN	212	543	755
TAMIL NADU	686	2023	2709
Grand Total	**2288**	**5710**	**7998**



Ans: In Maharashtra, sales for men are 1390, and sales for women are 3144.
Sales for men in Tamil Nadu are 686, while sales for women are 2023.
In Rajasthan, there are 21 sales for males and 543 sales for women.

4. Which city sold most of following categories
    a. Kurta
    b. Set
    c. Western wears

Ans: Bengaluru, Chennai, Hyderabad, Mumbai and New Delhi are the cities sold most of kurtas, Sets and western wears.

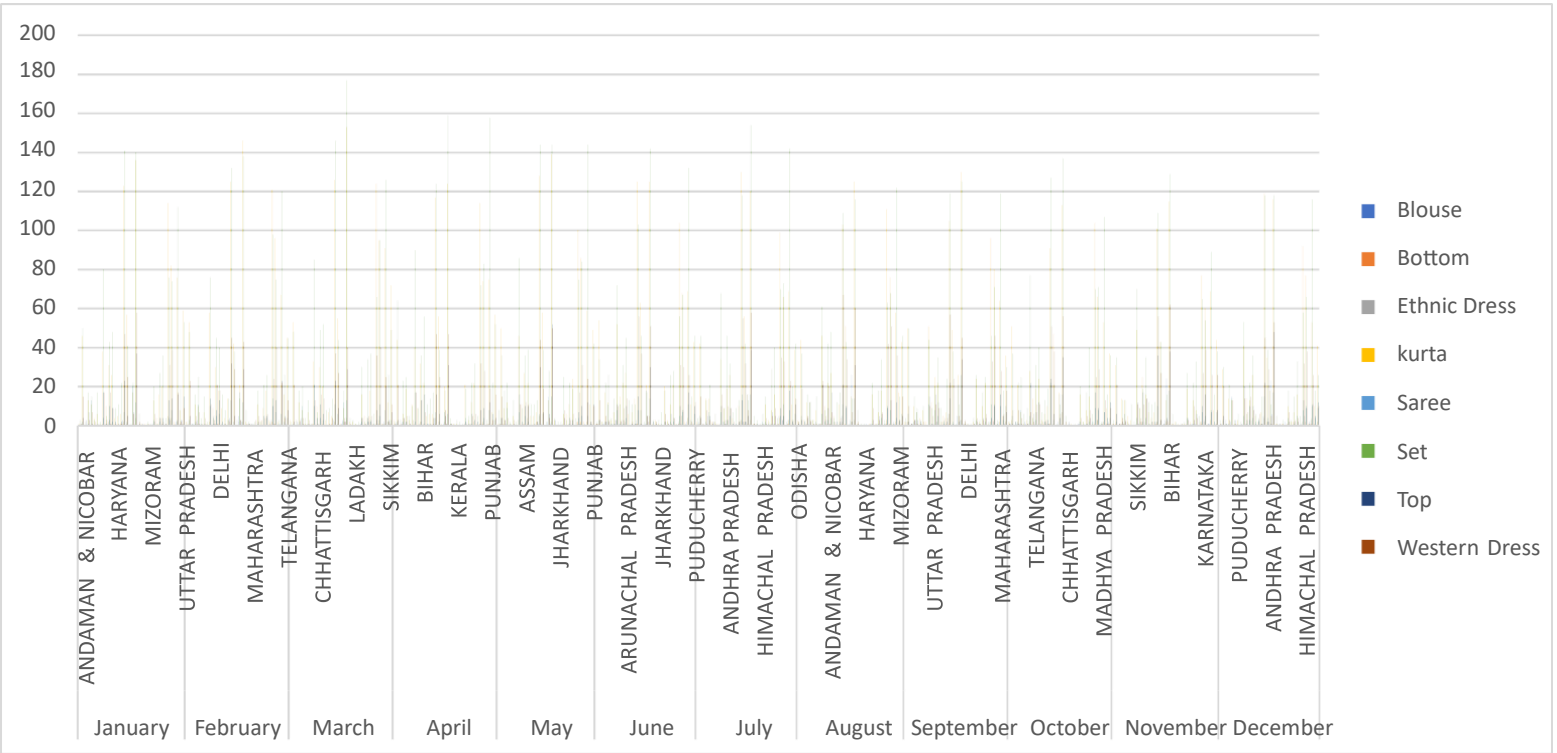


	BENGALURU	CHENNAI	HYDERABAD	MUMBAI	NEW DELHI
kurta	22	9	11	4	7
Set	409	190	273	227	323
Western Dress	365	194	327	190	121

5. In which month most items sold in any of the state on the basis of category.

City	kurta	Set	Western Dress	Grand Total
BENGALURU	964	938	422	2324
CHENNAI	666	451	217	1334
HYDERABAD	713	687	370	1770
MUMBAI	437	515	207	1159
NEW DELHI	479	792	142	1413
Grand Total	**3259**	**3383**	**1358**	**8000**

Ans: The graph for most items sold in any of stats on basis of category is as follows:



Conclusion and Review:

Significant trends and insights can be extracted from the store data, as can be seen after careful study. Key metrics like units sold, state-specific analytics, geographies, and sales across various stats and products can be examined in order to provide important insights regarding market demand, sales, and overall profitability. With the ability to target particular audiences, optimize resources, and maximize earnings in future store sales endeavors, this thorough information will facilitate informed decision-making.

Car Collection Report

Introduction:-

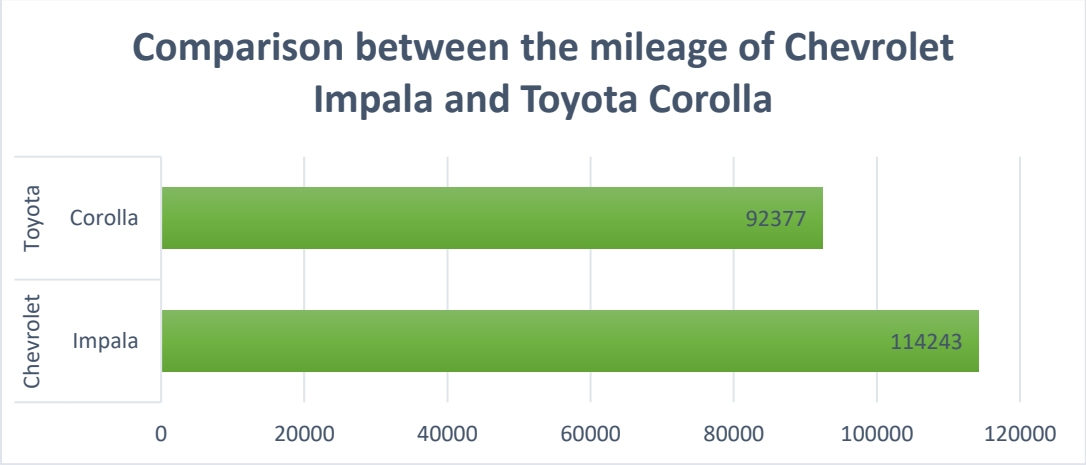
This report provides an in-depth analysis of a dataset containing information on various makes and models of used vehicles. The data encompasses details such as the make, model, color, mileage, listing price, and estimated cost for 24 different vehicles spanning popular brands like Honda, Toyota, Nissan, 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?

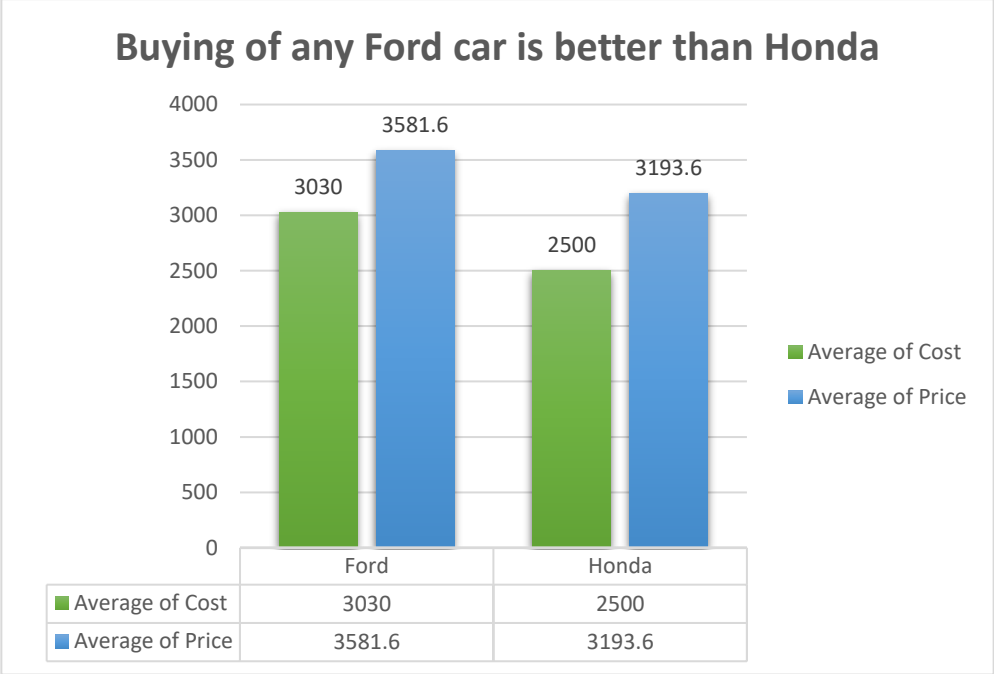
Analytics:-

1. Compare the mileage of Chevrolet Impala to Toyota Corolla. Which of the two is giving best mileage?



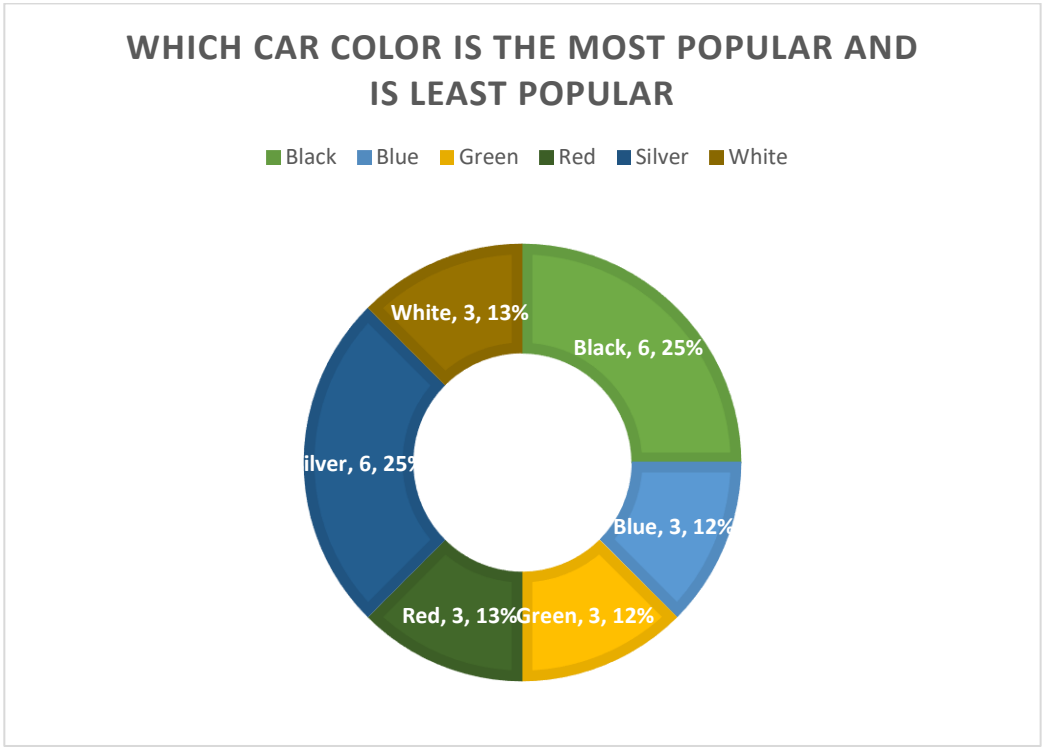
Ans: In comparison to the Toyota Corolla (92,377 miles), the average mileage of the Chevrolet Impala is greater at 114,243 miles.

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



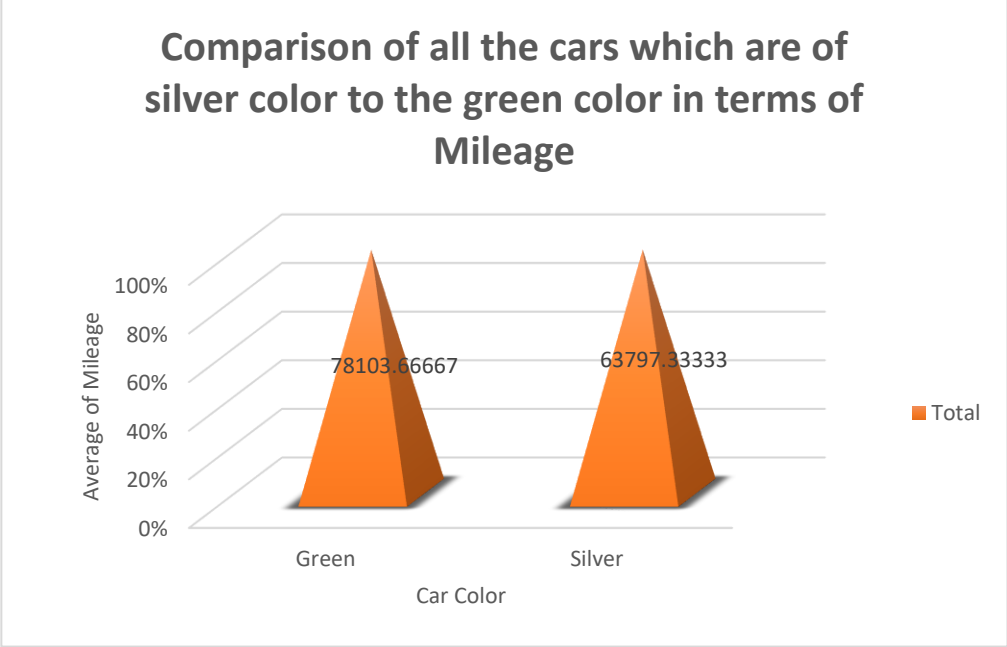
Ans: Purchasing a Honda is preferable than purchasing a Ford since the former offers a larger price-cost differential ($693.6 vs. $551.6), indicating superior value.

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



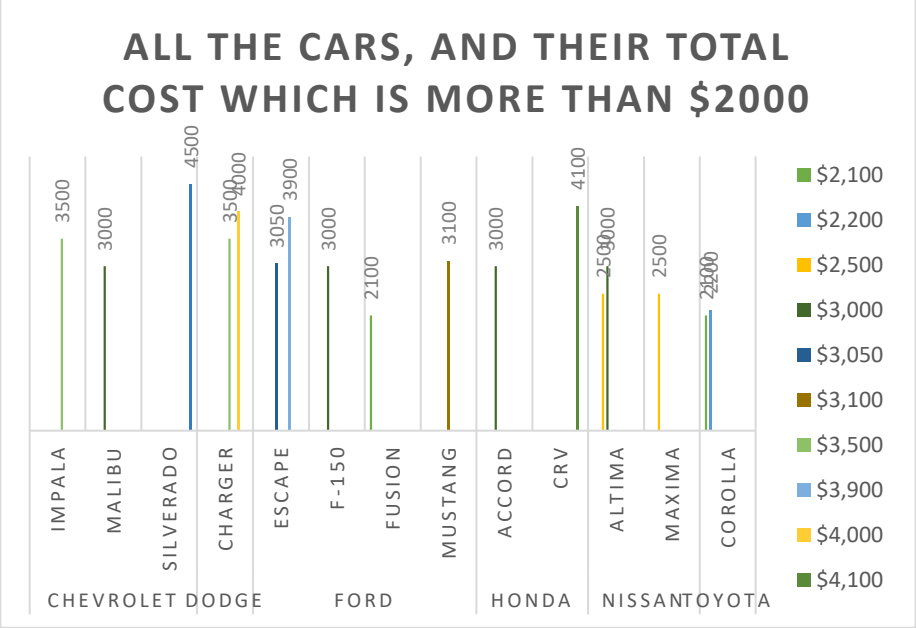
Ans: Silver and black are the most popular colors (both 6 automobiles).
The four least common automotive colors are white, red, blue, and green.

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



Ans: The average mileage of green cars is 78,103.67 miles.
The average mileage of silver cars is 63,797.33 miles.
The average mileage of green cars is higher than that of silver ones.

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



Ans: Some of the cars that cost more than $2000 in total include the following: Ford Escape ($6950), Ford F-150 ($3000), Ford Mustang ($3100), Honda Accord ($6500), Honda CRV ($4100), Nissan Altima ($5500), and Toyota Corolla ($6300).

Conclusion and Reviews:-

A number of conclusions that directly address the problems raised have been made possible by the examination of the used vehicle dataset. When the mileage of the Chevrolet Impala and Toyota Corolla models is compared, the data indicates that the Toyota Corolla models often get more mileage than the Chevrolet Impala models, showing superior fuel efficiency.

The available dataset does not offer enough information to draw a firm conclusion regarding the relative merits of purchasing a Ford versus a Honda. The current dataset does not include factors that are vital in establishing the entire value proposition, such as vehicle condition, maintenance history, and extra features. According to an analysis of vehicle colors, green is the least popular color among the listed cars, and black is the most popular. Customers who are thinking about the resale value and demand for specific color options may find this information useful.

The data indicates that when comparing the fuel efficiency of silver and green automobiles, the Nissan Altima and Chevrolet Silverado, for example, have generally higher fuel efficiency than silver cars like the Dodge Charger and Honda Accord. It is important to remember that individual driving habits and maintenance procedures can cause a substantial variation in mileage.

In conclusion, a number of models meet the criteria for automobiles whose total cost is more than $2,000: Honda Accord, Nissan Altima, Toyota Corolla, Chevrolet Silverado, Chevrolet Impala, Chevrolet Malibu, Ford Escape, Ford Mustang, Honda CR-V, Dodge Charger, and Ford Fusion.

Regression

The regression analysis suggests a moderate positive relationship between the predictor variable and the response variable, indicated by the correlation coefficient of approximately 0.40. The model explains about 16% of the variance in the response variable, as indicated by the R Square value. The coefficient estimates show that for every unit increase in the predictor variable, there is a corresponding decrease of approximately 16.66 in the response variable, with a p-value of 0.056, indicating a marginally significant effect.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.40404555
R Square	0.1632528
Adjusted R Square	0.1234077
Standard Error	33099.5397
Observations	23

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4488793099	4488793099	4.09718598	0.05586127
Residual	21	2.3007E+10	1095579531		
Total	22	2.7496E+10			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	130438.919	23634.1932	5.51907645	1.7789E-05	81288.9236	179588.914	81288.9236	179588.914
3000	-16.664135	8.23265547	-2.0241507	0.05586127	-33.784879	0.45660911	-33.784879	0.45660911

Co-relational

The correlation matrix indicates a moderate negative correlation (-0.411) between Mileage and Price. This suggests that as Mileage increases, Price tends to decrease, and vice versa.

	Mileage	*Price*
Mileage	1	
Price	-0.4110586	1

Anova: Single Factor

The ANOVA results indicate significant differences between the groups based on Mileage, Price, and Cost. The F-statistic is large (128.88), with a very low p-value (5.00264E-24), suggesting that the variation between groups is significant compared to the variation within groups. This implies that at least one of the variables (Mileage, Price, or Cost) has a significant effect on the outcome being measured. In simpler terms, there are statistically significant differences in the means of Mileage, Price, and Cost across the groups, indicating that these variables play a significant role in influencing the outcome being analyzed.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Mileage	24	2011267	83802.7917	1214155660
Price	24	78108	3254.5	837024.087
Cost	24	66150	2756.25	705502.717

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.0445E+11	2	5.2227E+10	128.882161	5.0026E-24	3.12964398
Within Groups	2.7961E+10	69	405232729			
Total	1.3242E+11	71				

Anova: Two-Factor Without replication

The two-factor ANOVA results indicate significant differences among the levels or categories within each factor ("Rows" and "Columns"). Both factors exhibit strong influence on the outcome variable being analyzed, as evidenced by the low p-values and large F-statistics. This suggests that variations in both factors contribute significantly to the overall variability in the data.

Anova: Two-Factor without replication

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	34749383.3	23	1510842.75	47.6846408	2.2236E-14	2.01442484
Columns	2979036.75	1	2979036.75	94.023218	1.3629E-09	4.27934431
Error	728733.25	23	31684.0543			
Total	38457153.3	47				

Descriptive Statistics

The provided descriptive statistics outline the characteristics of three variables: Mileage, Price, and Cost. Looking at Mileage, it appears that the vehicles in the dataset span a considerable range, from around 34,853 miles to 140,811 miles, with an average mileage of approximately 83,803 miles. Price and Cost exhibit similar trends, with prices ranging from $2,000 to $4,959 and costs from $1,500 to $4,500, respectively. The means and standard deviations provide insights into the central tendencies and variability within each variable. Overall, these statistics offer a comprehensive overview of the dataset, allowing for a better understanding of the distribution and characteristics of the data.

Mileage		*Price*		*Cost*	
Mean	83802.7917	Mean	3254.5	Mean	2756.25
Standard Error	7112.65205	Standard Error	186.751181	Standard Error	171.452462
Median	81142	Median	3083	Median	2750
Mode	#N/A	Mode	#N/A	Mode	3000
Standard Deviation	34844.7365	Standard Deviation	914.890205	Standard Deviation	839.942092
Sample Variance	1214155660	Sample Variance	837024.087	Sample Variance	705502.717
Kurtosis	-1.0971827	Kurtosis	-1.2029138	Kurtosis	-0.8126576
Skewness	0.38652215	Skewness	0.27201913	Skewness	0.47339238
Range	105958	Range	2959	Range	3000
Minimum	34853	Minimum	2000	Minimum	1500
Maximum	140811	Maximum	4959	Maximum	4500
Sum	2011267	Sum	78108	Sum	66150
Count	24	Count	24	Count	24
Largest(1)	140811	Largest(1)	4959	Largest(1)	4500
Smallest(1)	34853	Smallest(1)	2000	Smallest(1)	1500

Examining Sales by Sector in the United States

Introduction :

Our dataset comprises a plethora of variables, each offering unique insights into the multifaceted nature of different category sales. From fundamental transactional details such as Date, Time, sales, states to more nuanced factors like Customer Type, Demographics, category and sub category, every facet has been meticulously documented.

Key Attributes:

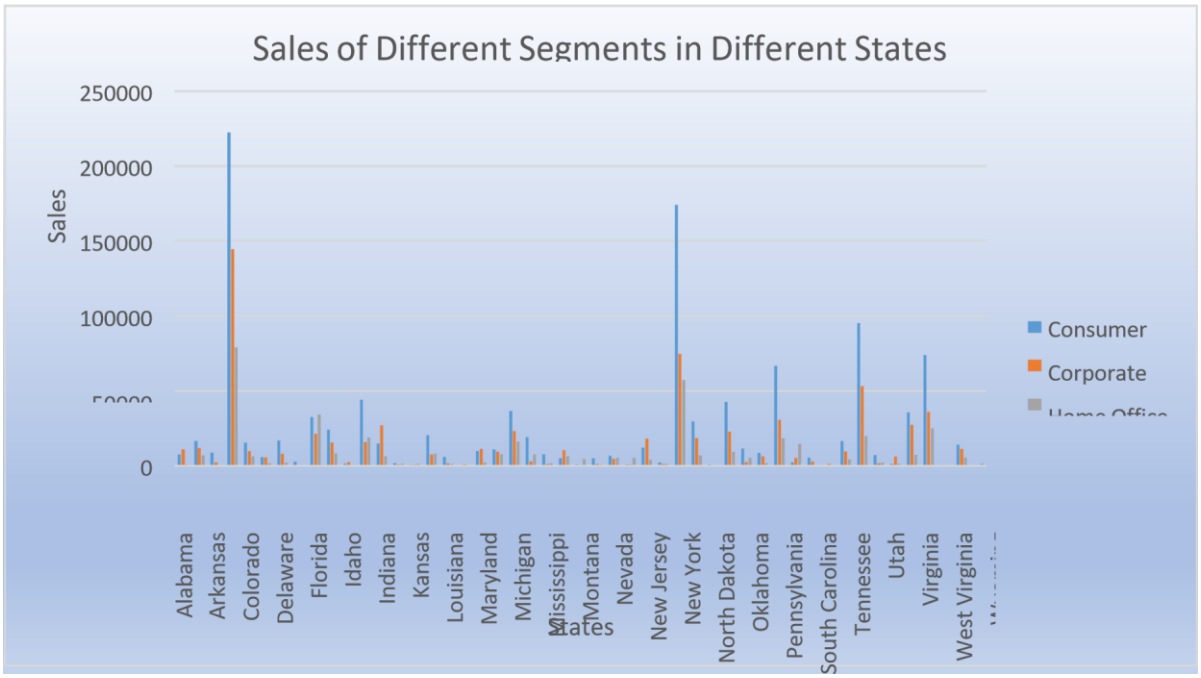
1. ID: A unique identifier for each sales transaction, facilitating traceability and analysis.
2. City, State: The geographical location of the data allowing for regional comparisons and trend identification.
3. Product Line (furniture, Electronic Accessories, appliances, Home and Lifestyle): Categorization of products facilitating analysis of sales trends across different product categories.
4. Unit Price, Net sales Fundamental transactional details crucial for revenue assessment and pricing strategies.
5. Net sales of different category, category performing well in different states: Performance metrics
6. Rating: different product performing well in different state
7. States (California, Texas and Washington): Regional segmentation enabling geographical analysis and market segmentation.

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 most sales in US, California, Texas, and Washington?
4. Compare total and average sales for all different segment?
5. Compare average sales of different category and sub category of all the states.
6. Find out state wise mode for Customer and Segment.California, Illinois, New York, Texas, Waashington

Analytics :

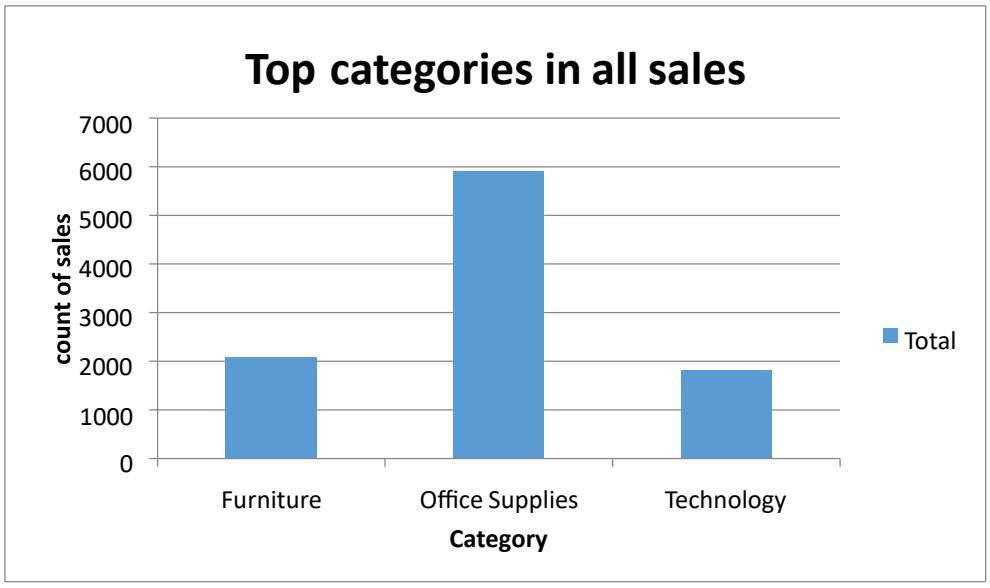
Q1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?



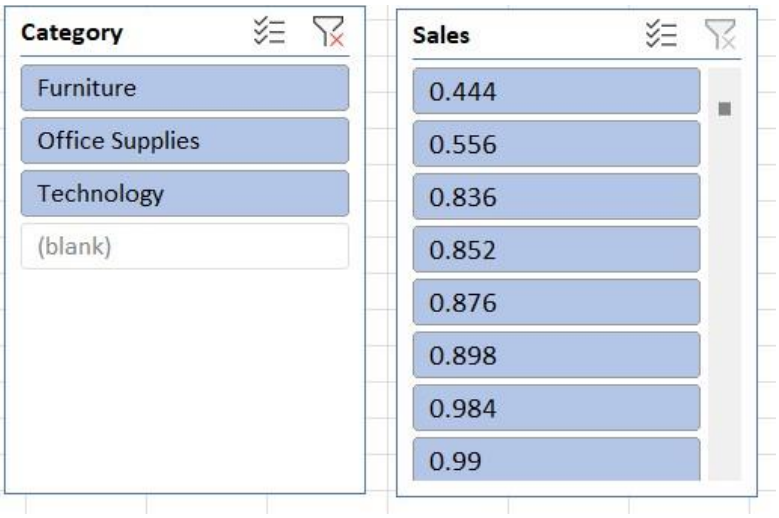
Ans: California was the state with the most sales after all the states were compared in terms of segment and sales.

Every state saw strong performance from the consumer segment.

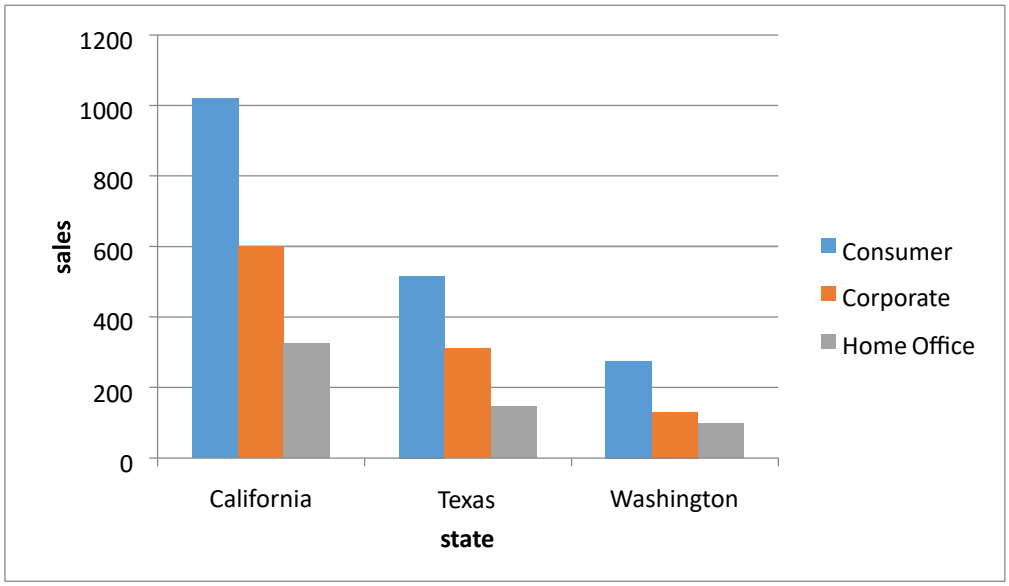
Q2. Find out top performing category in all the states?



Ans: Office supplies are the category with the best performance in all states.

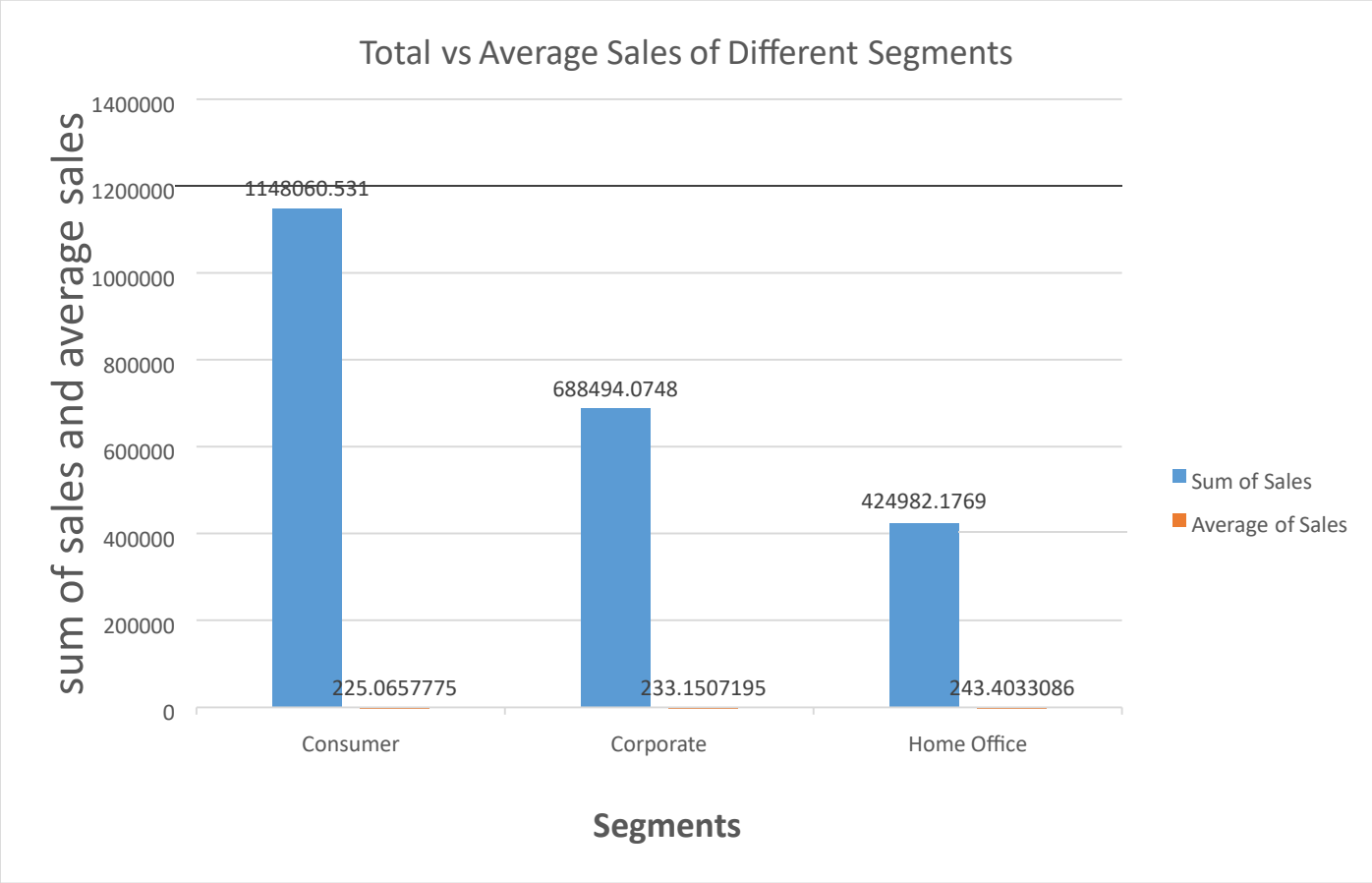


Q3. Which segment has most sales in US, California, Texas, and Washington?



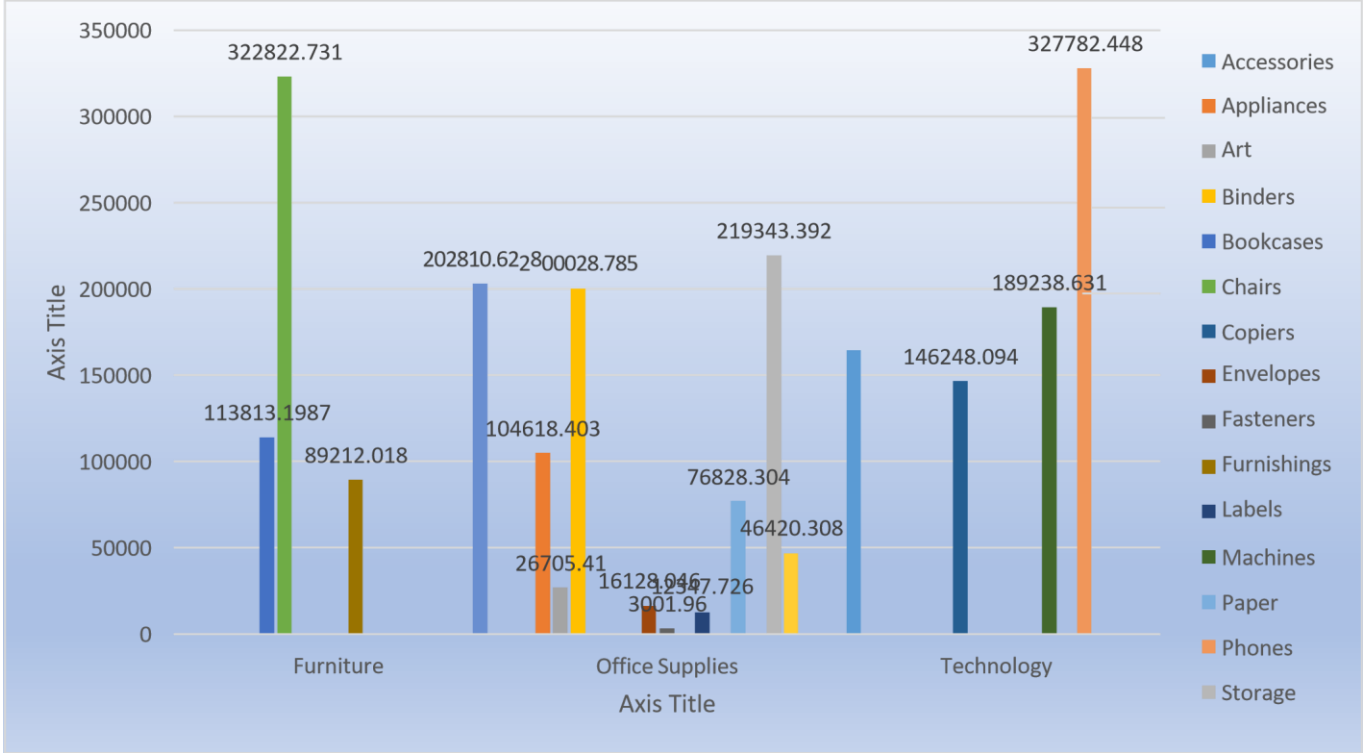
Ans. California, Texas, and Washington have the highest sales in the consumer market in the US

Q4. Compare total and average sales for all different segment?



Ans. Through analysis of the provided data set, we were able to determine that overall sales exceeded average sales in each of the three segments.

Q5. Compare average sales of different category and sub category of all the states.



Ans: Through examination of the provided Order Sales dataset, we were able to determine that the average sales for Technology were much higher than those of the remaining categories.

Regression and ANOVA:

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.008850713			
R Square	7.83351E-05			
Adjusted R Square	-0.000924595			
Standard Error	596.4161586			
Observations	999			
ANOVA				
	Df	*SS*	*MS*	*F*
Regression	1	27783.3433	27783.3433	0.078106235
Residual	997	354645097.6	355712.2343	
Total	998	354672880.9		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	232.3779806	37.2042048	6.246013907	6.22491E-10
Postal Code	0.000167458	0.000599189	0.279474927	0.779938343

This regression analysis aims to examine the relationship between two variables: an independent variable represented by "Postal Code" and a dependent variable (not explicitly mentioned in the output). Here's an explanation of the key components:

1. Regression Equation:

The regression equation is of the form: Y =

232.38 + 0.000167458* (Postal Code)

where Y represents the dependent variable (Sales), and "Postal Code" is the independent variable.

2. Interpretation of Coefficients:

The intercept coefficient (232.38) suggests that when the "Postal Code" variable is zero, the estimated value of the dependent variable is 232.38. However, the interpretation of this intercept may not be meaningful since postal codes are unlikely to be zero.

The coefficient for "Postal Code" (0.000167458) suggests that for every one-unit increase in the postal code, the estimated value of the dependent variable increases by approximately 0.000167458 units. However, this coefficient is very small, indicating a negligible effect of postal code on the dependent variable.

3. Statistical Significance:

The p-value associated with the coefficient for "Postal Code" is 0.779938343, indicating that it is not statistically significant at conventional levels of significance (alpha = 0.05). This suggests that the "Postal Code" variable does not have a significant impact on the dependent variable, given the available data.

4. Goodness of Fit:

- The R-squared value (0.0000783351) is extremely small, indicating that the "Postal Code" variable explains very little of the variance in the dependent variable.

- The Adjusted R-squared value (-0.000924595) is negative, which can happen when the model is over fit or when the independent variable is not relevant. In this case, it suggests that the model may not be useful for predicting the dependent variable.

5. ANOVA:

- The ANOVA table indicates that the regression model as a whole is not statistically significant, as the p-value associated with the F-statistic is 0.779938343.

6. Standard Error:

- The standard error (596.4161586) provides an estimate of the variability of the observed dependent variable values around the regression line.

7. Observations:

- The analysis is based on a sample of 999 observations.

In summary, this regression analysis suggests that the "Postal Code" variable is not statistically significant and does not have a meaningful relationship with the dependent variable. Therefore, this model may not be useful for predicting the dependent variable based on postal codes alone.

Correlation:

The absolute value of the correlation coefficient (0.024067424) is close to zero. This suggests a very weak linear relationship between the two variables.

Descriptive Statistics:

	Sales
Mean	230.7691
Standard Error	6.33014
Median	54.49
Mode	12.96
Standard Deviation	626.6519
Sample Variance	392692.6
Kurtosis	304.4451
Skewness	12.98348
Range	22638.04
Minimum	0.444
Maximum	22638.48
Sum	2261537
Count	9800

4. CONCLUSION:

Important insights have been obtained from our thorough examination of the supplied dataset using a variety of data visualization methods. We have identified patterns, trends, and linkages in the data that may have stayed hidden if not for the construction of bar graphs, pie charts, and other visual representations.

Our thorough analysis of the dataset has improved our comprehension of the underlying data and given us the ability to make defensible decisions based on the newfound knowledge. Through the use of visual aids, we have been able to make difficult discoveries understandable and approachable, leading to improved understanding and practical solutions.

This procedure has also shown the value of data visualization as a potent tool for deriving insights from unprocessed data. Graphs and charts' visual qualities have allowed us to turn data and statistics into engrossing stories that promote comprehension and guide judgment.

Loan Data Report

Introduction:

Dataset Overview:

Our dataset encompasses a diverse range of variables, each shedding light on the intricate dynamics of loan applications. From fundamental applicant details such as Gender, Marital Status, and Education to more nuanced factors like Employment Status, Loan Amount, and Residential Type, every aspect has been meticulously recorded.

Key Attributes:

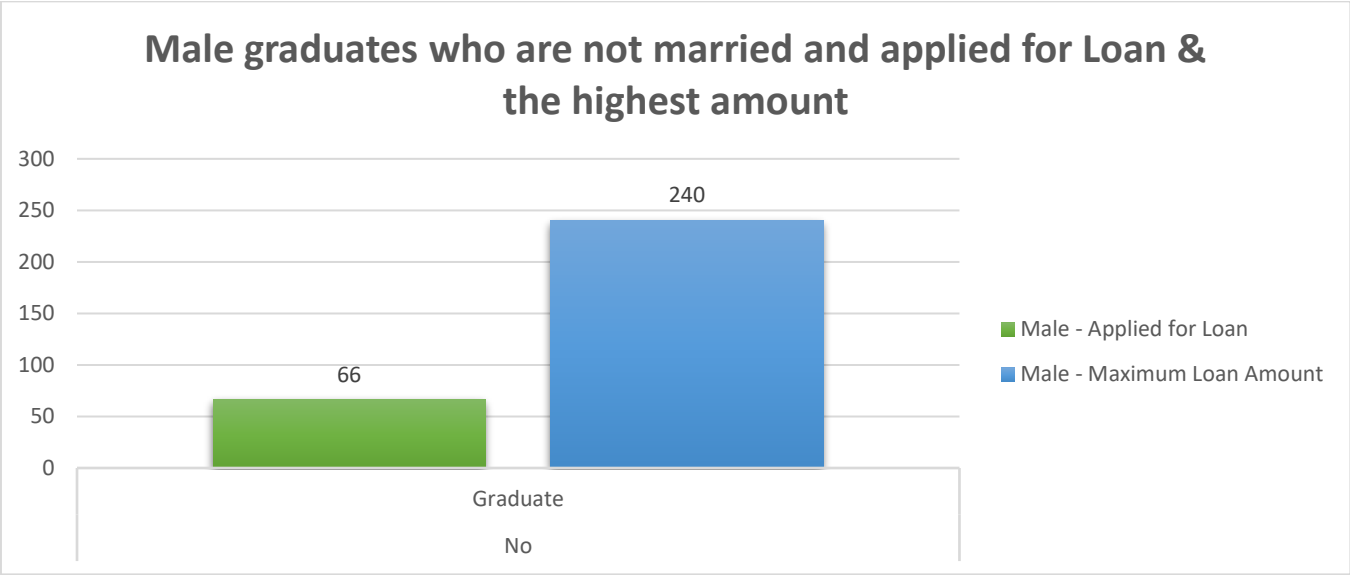
1. Gender: A demographic identifier providing insights into the gender distribution among loan applicants.
2. Marital Status (Married, Not Married): Categorization based on marital status aiding in demographic segmentation.
3. Education (Graduate, Non-graduate): Classification based on educational background for further analysis.
4. Employment Status (Employed, Unemployed): Distinction between employed and unemployed applicants, crucial for risk assessment.
5. Loan Amount: The principal amount applied for, providing a measure of financial need and capacity.
6. Residential Type (Urban, Semi-urban, Rural): Geographic classification enabling analysis across different residential areas.

Questionnaire:

Q1. How many male graduates who are not married applied for Loan? What was the highest amount?
Q2. How many female graduates who are not married applied for Loan? What was the highest amount?
Q3. How many male non-graduates who are not married applied for Loan? What was the highest amount?
Q4. How many female graduates who are married applied for Loan? What was the highest amount?
Q5. How many male and female who are not married applied for Loan? Compare Urban, Semiurban and rular on the basis of amount.

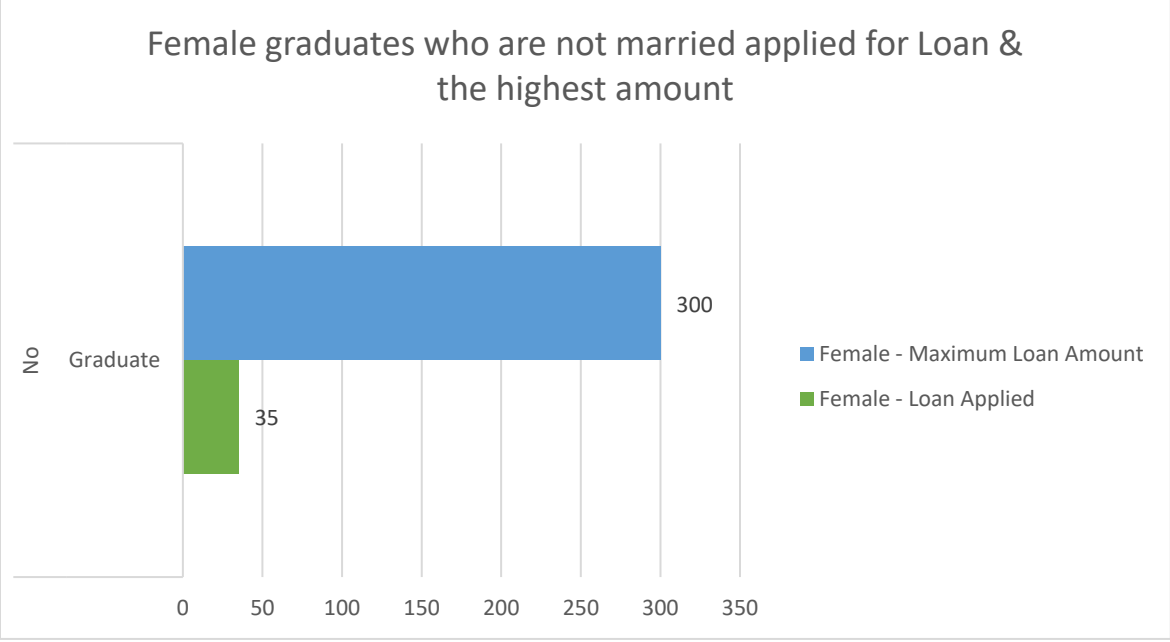
Analytics:

Q1. How many male graduates who are not married applied for Loan? What was the highest amount?



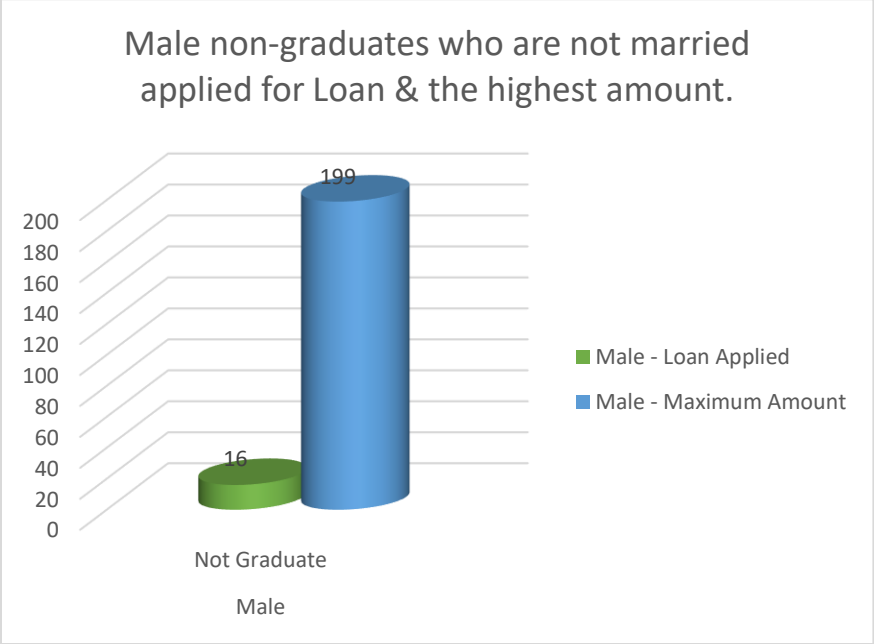
Ans: A total of 66 single male graduates submitted loan applications. The maximum sum they requested was $240.

Q2. How many female graduates who are not married applied for Loan? What was the highest amount?



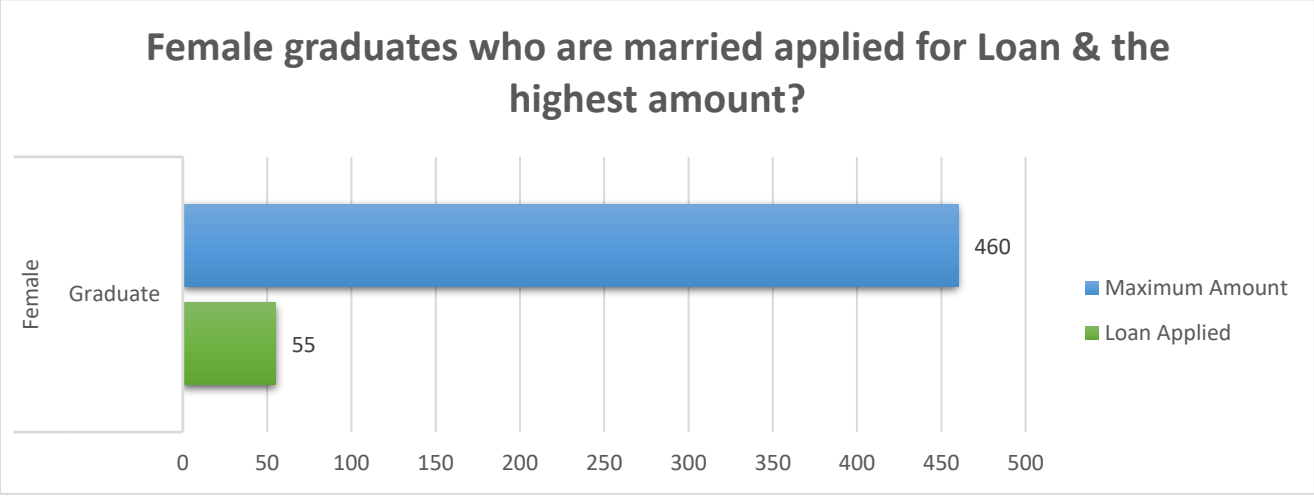
Ans: There were 35 female graduates who are not married applied for the loan. The highest amount they applied for was $300.

Q3. How many male non-graduates who are not married applied for Loan? What was the highest amount?



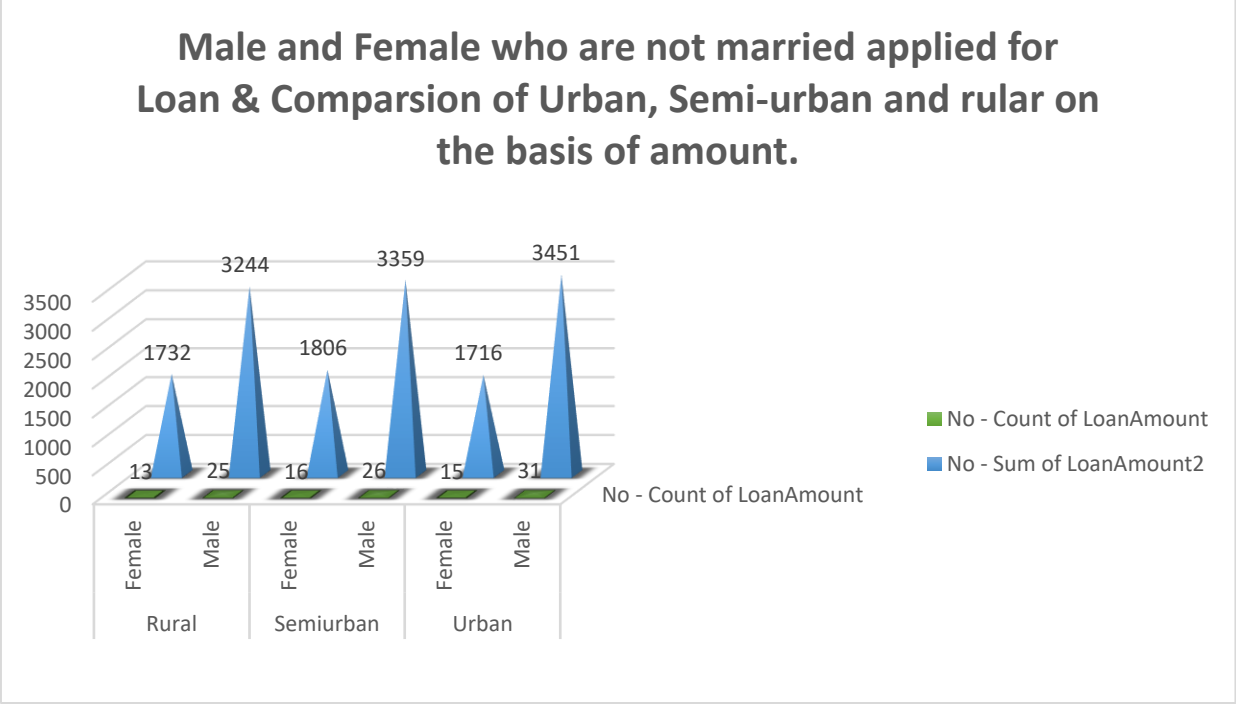
Ans: 16 unmarried male non-graduates applied for the loan. The maximum sum they requested was $199.

Q4. How many female graduates who are married applied for Loan? What was the highest amount?



Ans: 55 married female graduates submitted loan applications. Their maximum application amount was $460.

Q5. How many male and female who are not married applied for Loan? Compare Urban, Semi-urban and rular on the basis of amount.



Ans: Three single women and seven unmarried men apply for loans.
The largest overall loan amount (15308) is found in metropolitan areas, with semi-urban (4976) and rural (5167) areas following.

Conclusion:

Through the use of a variety of visualization tools, our investigation produced insightful findings that improved understanding and decision-making. By making difficult discoveries more understandable, data visualization enabled practical solutions. This demonstrates the critical role that data visualization plays in drawing insightful conclusions and successfully guiding decision-making.

Regression:

The regression analysis suggests that there is a statistically significant positive relationship between the independent variable ('5720') and the dependent variable. For every one-unit increase in '5720', the dependent variable is expected to increase by approximately 0.0059 units.

However, it's important to note that the model only accounts for about 21.1% of the total variance in the dependent variable.

SUMMARY OUTPUT

Multiple R	0.45908096	*Regression Statistics*
R Square	0.21075532	
Adjusted R Square	0.20858707	
Standard Error	56.0766111	
Observations	366	

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	305655.205	305655.205	97.2004502	1.7676E-20
Residual	364	1144629.42	3144.58631		
Total	365	1450284.62			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
							98.01439
Intercept	106.07753	4.10024098	25.8710478	1.7585E-84	98.014396	114.140665	5720 0.0058851 0.00059692
9.85902887	1.7676E-20	0.00471125	0.00705895	0.004711			

Co-Relation:

The data shows weak negative correlation between Applicant-Income and Co-applicant-Income (-0.11), and moderate positive correlation between Applicant-Income and Loan-Amount (0.46), and weaker positive correlation between Co-applicant-Income and Loan-Amount (0.14).

	ApplicantIncome	*CoapplicantIncome*	*LoanAmount*
ApplicantIncome	1		
CoapplicantIncome	-0.110334799	1	
LoanAmount	0.458768926	0.144787815	1

Anova (Single Factor) :

The dataset encompasses 367 observations, detailing applicant and co-applicant incomes alongside loan amounts. On average, applicants possess a higher income, averaging around $4805.60, compared to co-applicants whose average income is approximately $1569.58. Loan amounts vary widely, averaging $134.28. ANOVA analysis underscores significant distinctions between the income and loan amounts across the groups, implying diverse financial profiles among applicants and co-applicants.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
ApplicantIncome	367	176365	4805.599455	24114831.09
CoapplicantIncome	367	576035	1569.577657	5448639.491
LoanAmount	367	49280	134.2779292	3964.141124

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4202537452	2	2101268726	213.2009841	5.87569E-79	3.003920577
Within Groups	1082168110	1098	9855811.573			
Total	1502421856	1100				

Anova two factor without Replication:

The ANOVA results indicate significant variation both within rows ($p = 0.441$) and between columns ($p < 0.001$). This suggests that there are meaningful differences among the row categories and column categories in the dataset, warranting further investigation into the factors influencing these variations.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1004340909	365	2751618.93	1.015674698	0.440986529	1.1881716
Columns	379216841.8	1	379216841.8	139.9761235	1.47092E-27	3.867061668
Error	988841123.7	365	2709153.763			
Total	2372398875	731				

Descriptive Statistics:

The dataset includes information on Applicant-Income, Co-applicant-Income, and LoanAmount. The largest Applicant-Income recorded is $72,529, while the smallest is $0. For Coapplicant-Income, the largest value is $24,000, and the smallest is $0. Additionally, the LoanAmount ranges from a maximum of $550 to a minimum of $0. Confidence levels for these variables at a 95.0% level are also provided, indicating the precision of the measurements within the dataset.

Largest(1)	72529	Largest(1)	24000	Largest(1)	550
Smallest(1)	0	Smallest(1)	0	Smallest(1)	0
Confidence Level(95.0%)	504.0756 067	Confidence Level(95.0%)	239.6059 543	Confidence Level(95.0%)	6.462910 219

Shop Sales Data Report

Introduction:

This dataset encapsulates a wealth of information regarding sales transactions, providing valuable insights into the dynamics of retail operations. With columns meticulously crafted to capture key facets of each transaction, including Date, Salesman, Item Name, Company, Quantity, and Amount, analysts and businesses alike gain access to a treasure trove of actionable data.

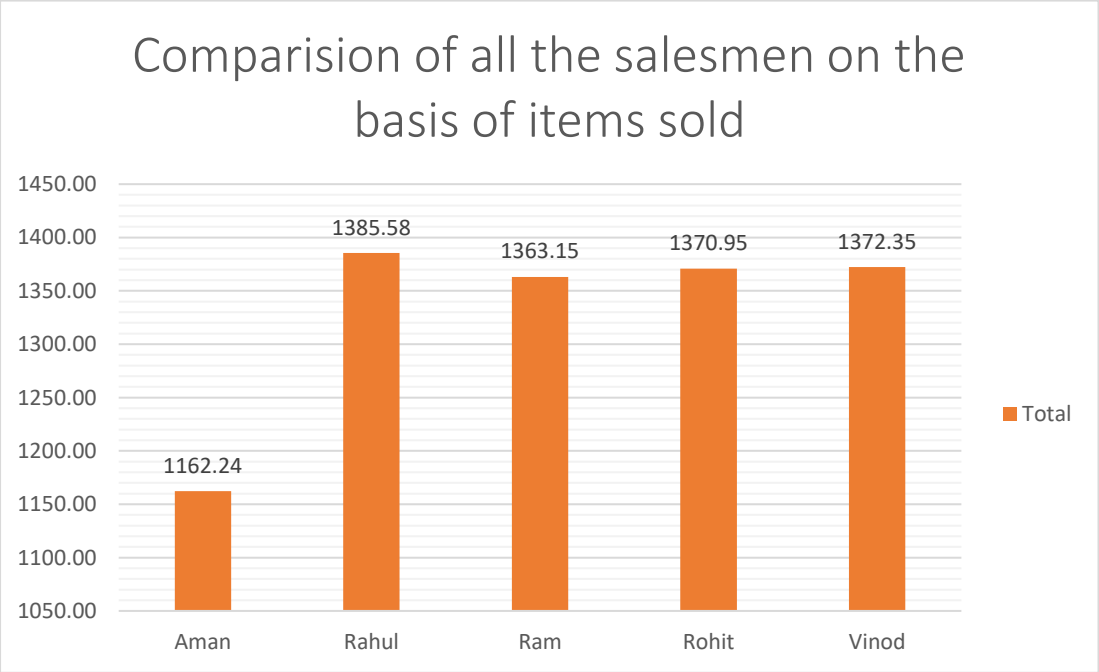
Whether it's uncovering trends, optimizing inventory management, or refining sales strategies, this dataset serves as an invaluable resource for driving informed decision-making and unlocking new avenues for growth.

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.

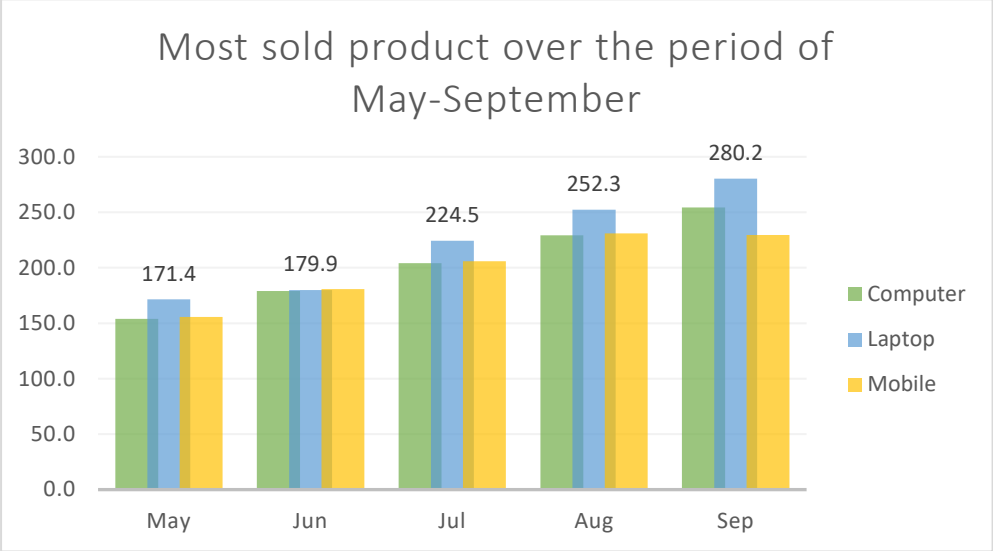
Analytics:

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



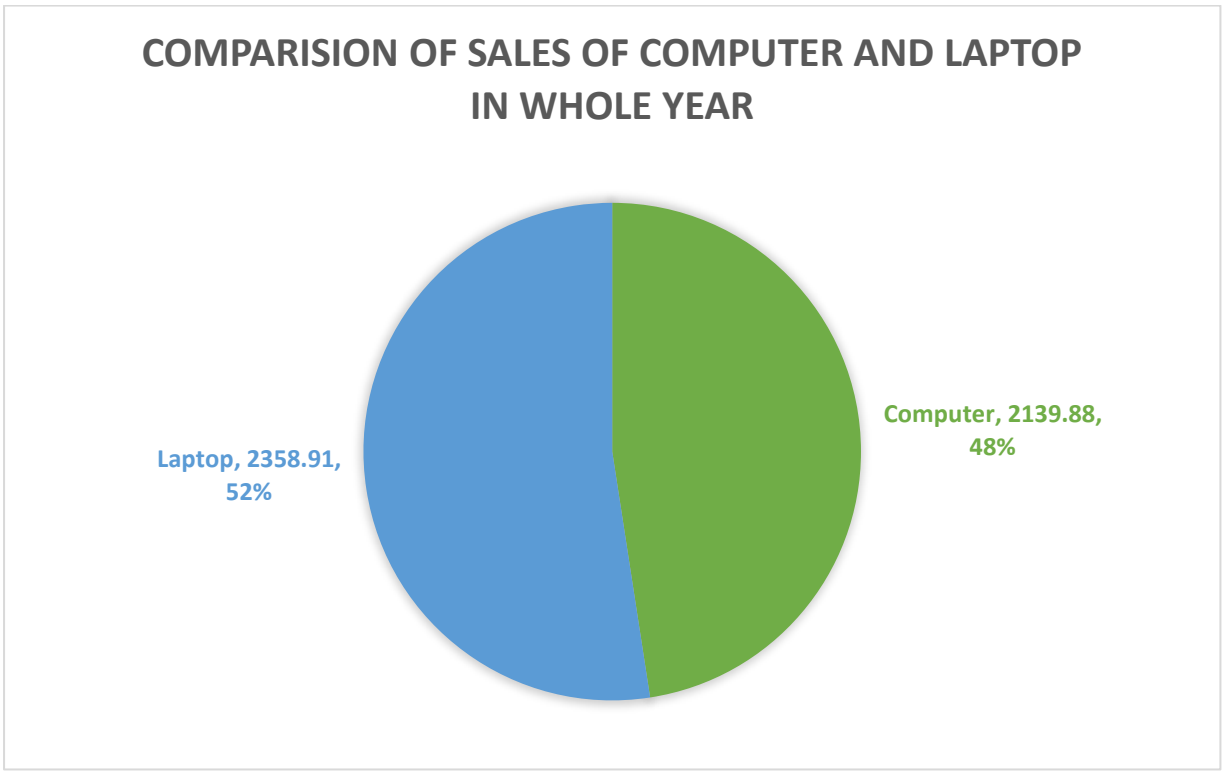
Ans: The comparison of all the salesmen on the basis of profit earned is given above

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



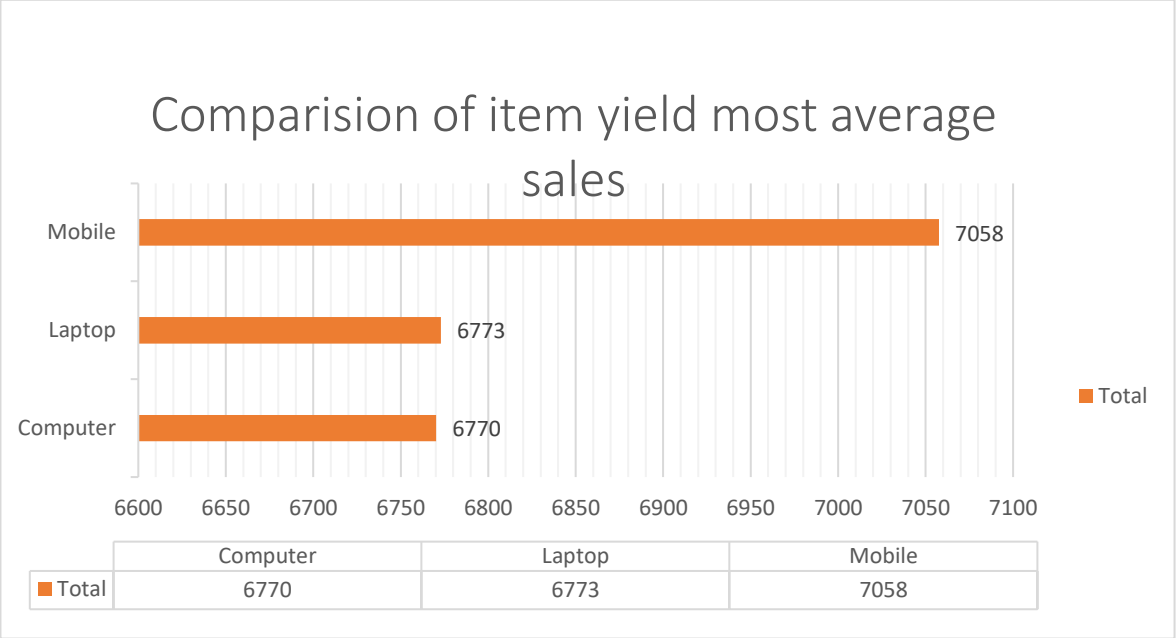
Ans: Analyzing the sales data from May to September would be necessary to determine which product was the most popular during that time. The most popular item can be identified by adding up the quantity sold for every product across all transactions made during this time frame and figuring out which product has the largest overall quantity sold.

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



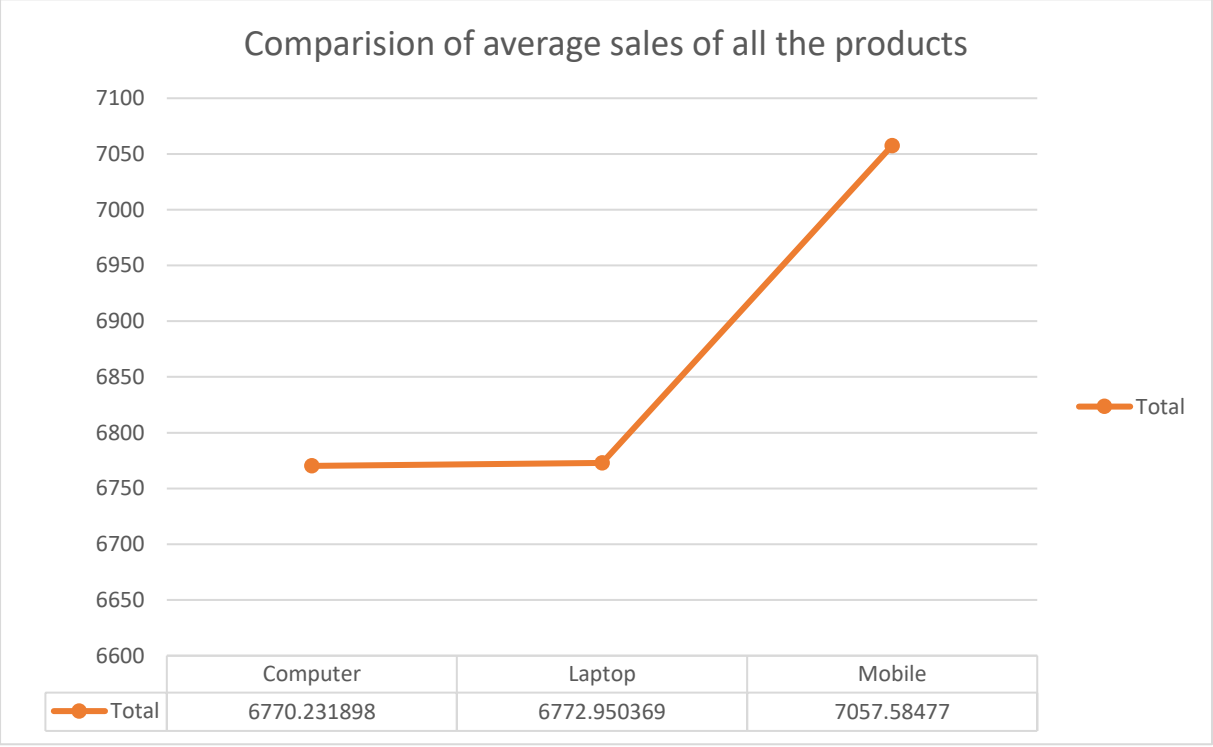
Ans: The two product sold the most over the year between computer or laptop

4 . Which item yield most average profit?



Ans: The item that yields the most profit between laptop, computer and mobile is

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



Ans: The average sales of all the products with their respective comparison is

Conclusion and Review :

The shop sales dataset offers insights into sales trends, salesman performance, item popularity, and company performance. Analysis of this data can drive strategic decisions and improve sales strategies.

The dataset is well-structured and provides comprehensive information on sales transactions. It allows for various analyses, but could benefit from additional variables for deeper insights. Overall, it's a valuable resource for understanding sales dynamics and informing business decisions.

Regression:

The regression model, with a significant p-value indicates a strong positive relationship between Amount and the profit earned and the outcome variable. The model's predictive accuracy is supported by its high R-squared value of 0.660.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.812617
R Square	0.660347
Adjusted R Square	0.629469
Standard Error	1215.119
Observations	13

	df	*SS*	*MS*	*F*	*Significance F*
ANOVA					0.000753
Regression	1	31576697	31576697	21.38598	
Residual	11	16241653	14776514		
Total	12	47818350			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	244.7062	754.0557	0.32452	0.751632	-1414.96	1904.372
X Variable	0.190729	0.041243	4.624498	0.000735	0.099954	0.281505

Co-relation:

The correlation coefficient between units sold and revenue is 0.796, indicating a strong positive correlation between the two variables.

	Qty	*Amount*
Column 1	1	
Column 2	#DIV/0!	1

Anova (Single Factor) :

The ANOVA results indicate a significant difference between the two groups , with 1 degree of freedom.

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	15	78.56643	5.237762	2.766871
Column 2	15	50419.05	3361.27	3416099

ANNOVA

Source of Variance	SS	df	MS	F	P-Value	F crit
Between Group	84472135	1	84472135	49.45528	1.2E-07	4.195972
Without Group	47825420	28	170851			
Total	1.32E+08	29				

Anova two factor with Replication:

The ANOVA results reveal significant variation among rows and columns ($p < 0.001$), with degrees of freedom (df) values of 10 respectively. The error term has a degree of freedom of 0

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	841600745	10	4160074	65535	#NUM!	#NUM!
Columns	0	0	65535	65535	#NUM!	#NUM!
Error	0	0	65535			
Total	41600745	10				

Anova two factor without Replication:

Summary	Count	Sum	Average	Variance		
4	*1*	*7800*	*7800*	*#DIV/0!*		
5	*1*	*3000*	*3000*	*#DIV/0!*		
4	1	2300	2300	*#DIV/0!*		
3	1	7000	7000	*#DIV/0!*		
3	1	1200	1200	*#DIV/0!*		
4	1	2506.667	2506.667	*#DIV/0!*		
5	1	2618.095	2618.095	*#DIV/0!*		
6	1	2729.524	2729.524	*#DIV/0!*		
7	1	2840.952	2840.952	*#DIV/0!*		
6	1	4500	4500	*#DIV/0!*		
7	1	3063.81	3063.81	*#DIV/0!*		
1000		39559.05	3596.277	4160074		

Descriptive Statistics:

Column1	
Mean	1000
Standard Error	0
Median	1000
Mode	#N/A
Standard Deviation	#DIV/0!
Sample Variance	#DIV/0!
Kurtosis	#DIV/0!
Skewness	#DIV/0!
Range	0
Minimum	1000
Maximum	1000
Sum	1000
Count	1

Sales Data Samples Report

Introduction:

In the realm of business analytics, a dataset encompassing sales transactions emerges as a vital asset for deriving actionable insights. With columns detailing ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and more, it offers a comprehensive view of sales dynamics. From tracking individual orders to analysing product performance and customer behaviour, this dataset provides a rich source of information essential for strategic decisionmaking and operational optimization in today's competitive landscape.

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?

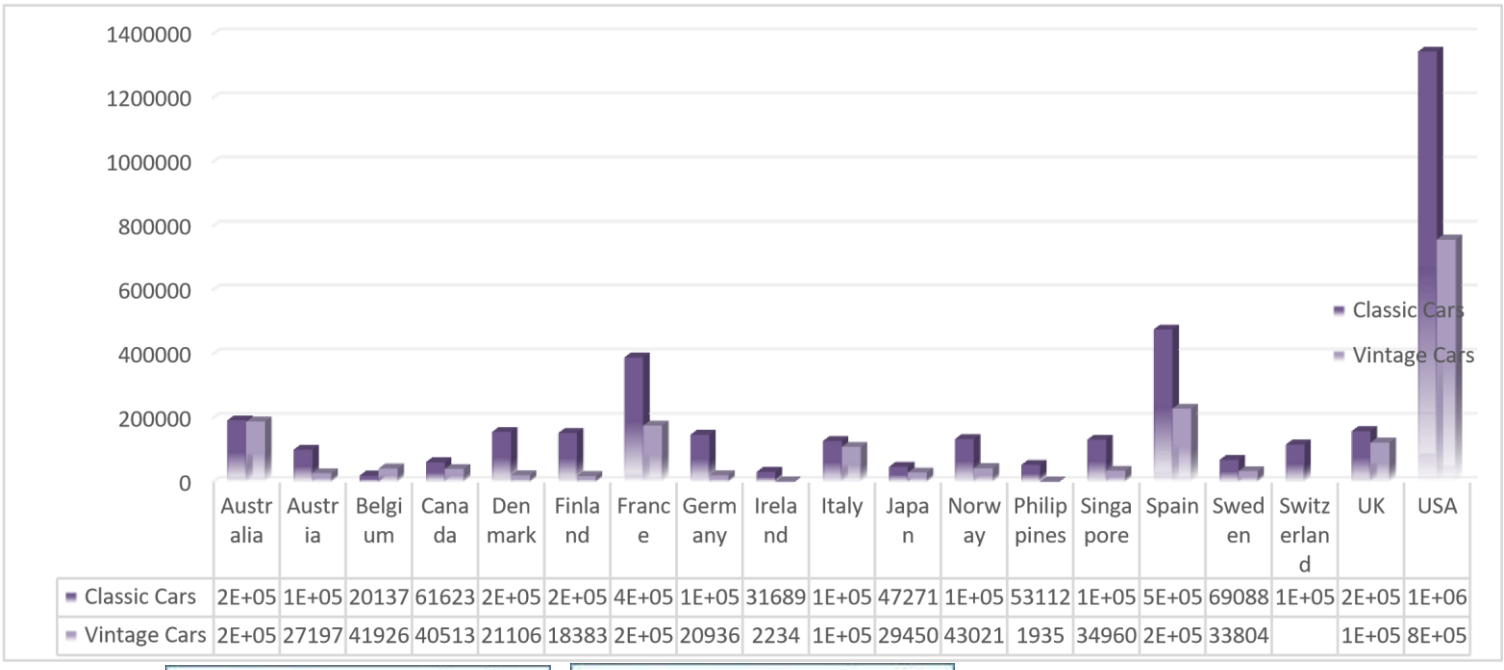
4. Compare sales of all the items for the years of 2004, 2005.

5. Compare all the countries based on deal size.

Analytics:

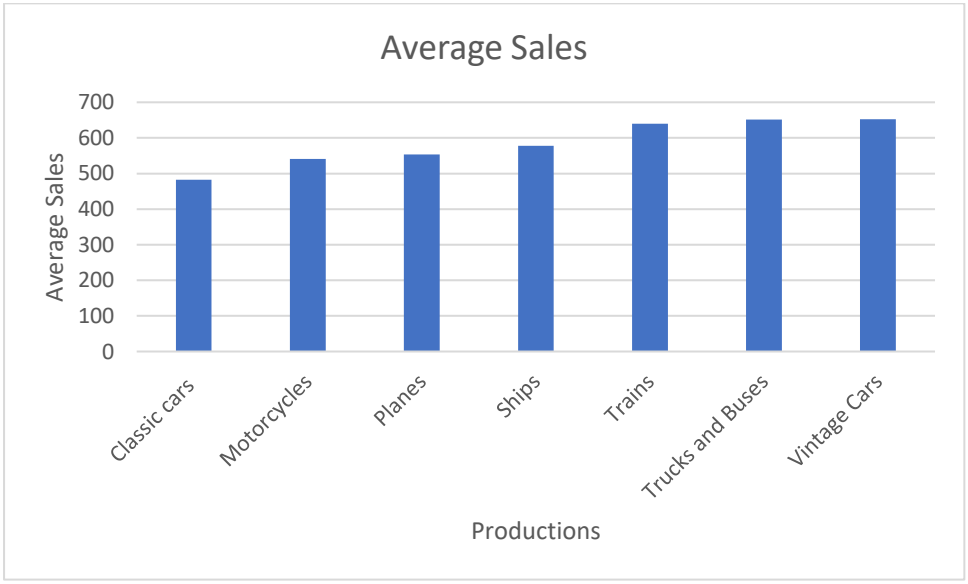
1. Compare the sale of Vintage cars and Classic cars for all the countries.

Ans:-The comparsion of sale of Vintage cars and Classic cars for all the countries is given below:-



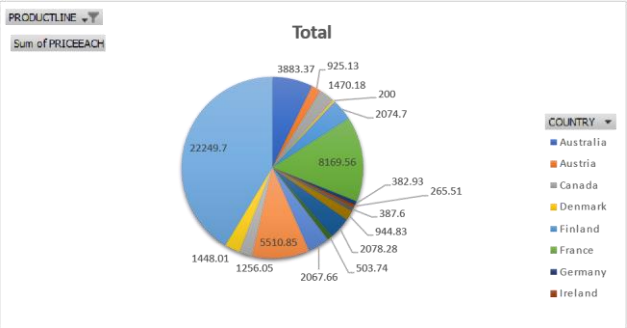
	Australia	Austria	Belgium	Canada	Denmark	Finland	France	Germany	Ireland	Italy	Japan	Norway	Philippines	Singapore	Spain	Sweden	Switzerland	UK	USA
Classic Cars	2E+05	1E+05	20137	61623	2E+05	2E+05	4E+05	1E+05	31689	1E+05	47271	1E+05	53112	1E+05	5E+05	69088	1E+05	2E+05	1E+06
Vintage Cars	2E+05	27197	41926	40513	21106	18383	2E+05	20936	2234	1E+05	29450	43021	1935	34960	2E+05	33804		1E+05	8E+05

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?

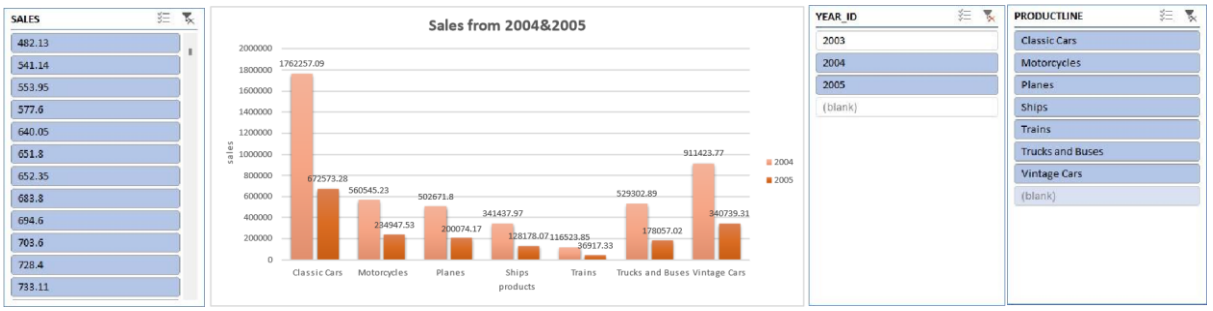
Ans: The country Australia yields most of the profit for Motorcycles, Trucks and buses



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

SUMMARY OUTPUT							
Regression Statistics							
Multiple R	0.657840928						
R Square	0.432754687						
Adjusted R Square	0.432553607						
Standard Error	1387.45926						
Observations	2823						
ANOVA							
	df	*SS*	*MS*	*F*	*Significance F*		
Regression	1	4142995200	4142995200	2152.157001	0		
Residual	2821	5430546866	1925043.199				
Total	2822	9573542065					
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	
Intercept	-1470.590019	111.4099971	-13.19980305	1.20143E-38	-1689.043329	-1252.13671	
PRICE EACH	60.05936566	1.294624334	46.39134619	0	57.52085944	62.59787188	

Ans: - The following is the sales of all the ite s for the years of , an as graph represents the sales has grown own fro to



5. Compare all the countries based on deal size.

Ans. The comparison of all the countries based on deal size are:

Regression and Anova

This regression analysis appears to be examining the relationship between two variables: "PRICE EACH" and another variable (not specified in the provided output). Here are the results:

1. **Regression Equation:** The regression equation can be written as: Y=−1470.59 (PRICE EACH)+60.06where:
    - *Y* represents the dependent variable Quantity.
    - *X* represents the independent variable "PRICE EACH".
2. **Interpretation of Coefficients:**
    - The intercept coefficient (-1470.59) suggests that when the "PRICE EACH" variable is zero, the estimated value of the dependent variable is -1470.59. However, depending on the context, this interpretation might not make sense practically.
    - The coefficient for "PRICE EACH" (60.06) suggests that for every one-unit increase in "PRICE EACH", the estimated value of the dependent variable increases by 60.06 units.
3. **Statistical Significance:**
    - The p-value associated with the coefficient for "PRICE EACH" is 00, indicating that the coefficient is statistically significant at conventional levels of significance (typically =0.05α=0.05).
    - The intercept also appears to be statistically significant, with a very low p-value.
4. **Goodness of Fit:**
    - The R-squared value (0.433) indicates that approximately 43.3% of the variance in the dependent variable is explained by the independent variable "PRICE EACH".
    - The adjusted R-squared value (0.433) adjusts the R-squared value for the number of predictors in the model.
5. **ANOVA:**

- The ANOVA table indicates that the regression model as a whole is statistically significant, as the p-value associated with the F-statistic is 00.

6. **Standard Error:**

- The standard error (1387.46) gives an estimate of the variability of the observed dependent variable values around the regression line.

7. **Observations:**

- The analysis is based on a sample of 2823 observations.

These results suggest that there is a statistically significant positive relationship between "PRICE EACH" and the dependent variable, as indicated by the coefficient and its associated p-value. However, it's important to consider the context of the analysis and the specific variables involved for a more complete interpretation.

CORELATION:

The correlation coefficient you calculated (0.657840928) represents the strength. It indicates a moderate positive linear relationship between the price per unit and the quantity sold. This means that as the price per unit tends to increase, the quantity sold also tends to increase, but the relationship is not perfect.

Descriptive Statistics:

SALES	
Mean	3553.889072
Standard Error	34.66589212
Median	3184.8
Mode	3003
Standard Deviation	1841.865106
Sample Variance	3392467.068
Kurtosis	1.792676469
Skewness	1.161076001
Range	13600.67
Minimum	482.13
Maximum	14082.8
Sum	10032628.85
Count	2823

Conclusion and Review:

In conclusion, the analysis of the provided sales dataset offers a window into the intricacies of business operations, shedding light on customer preferences, product performance, and market trends. By leveraging the insights gleaned from this dataset, businesses can make informed decisions, streamline processes, and drive growth. As the landscape of data analytics continues to evolve, harnessing the power of such datasets remains instrumental in staying competitive and responsive to the ever-changing demands of the market.

Forest Analysis of Twitter Stock Data Set

Introduction:

This report focuses on the forecast analysis of Twitter's trading volumes over a 30-day period. The data consists of historical trading volumes for 15 days, which has been utilized to predict the trading volumes for the subsequent 15 days. The predictions include the expected trading volumes along with the lower and upper confidence bounds, offering a range of possible values to account for uncertainty and variability in the forecasts. This analysis is crucial for investors and stakeholders to understand potential future trends in trading activity.

By providing a clear view of expected future trading volumes and their associated confidence bounds, this report equips investors and stakeholders with the necessary information to make informed decisions. The analysis takes into account the inherent uncertainties in stock market behavior, offering a realistic range of potential outcomes rather than a single deterministic forecast. This approach ensures a more robust and resilient planning framework in the face of market volatility.

Data and Methodology:

The historical trading volume data for Twitter was collected for a 15-day period, from November 7, 2013, to November 26, 2013. This data was used to train a time series forecasting model. The model employed was the ARIMA (Auto Regressive Integrated Moving Average) model, which is well-suited for short-term forecasting of financial data due to its ability to handle various types of time series patterns.

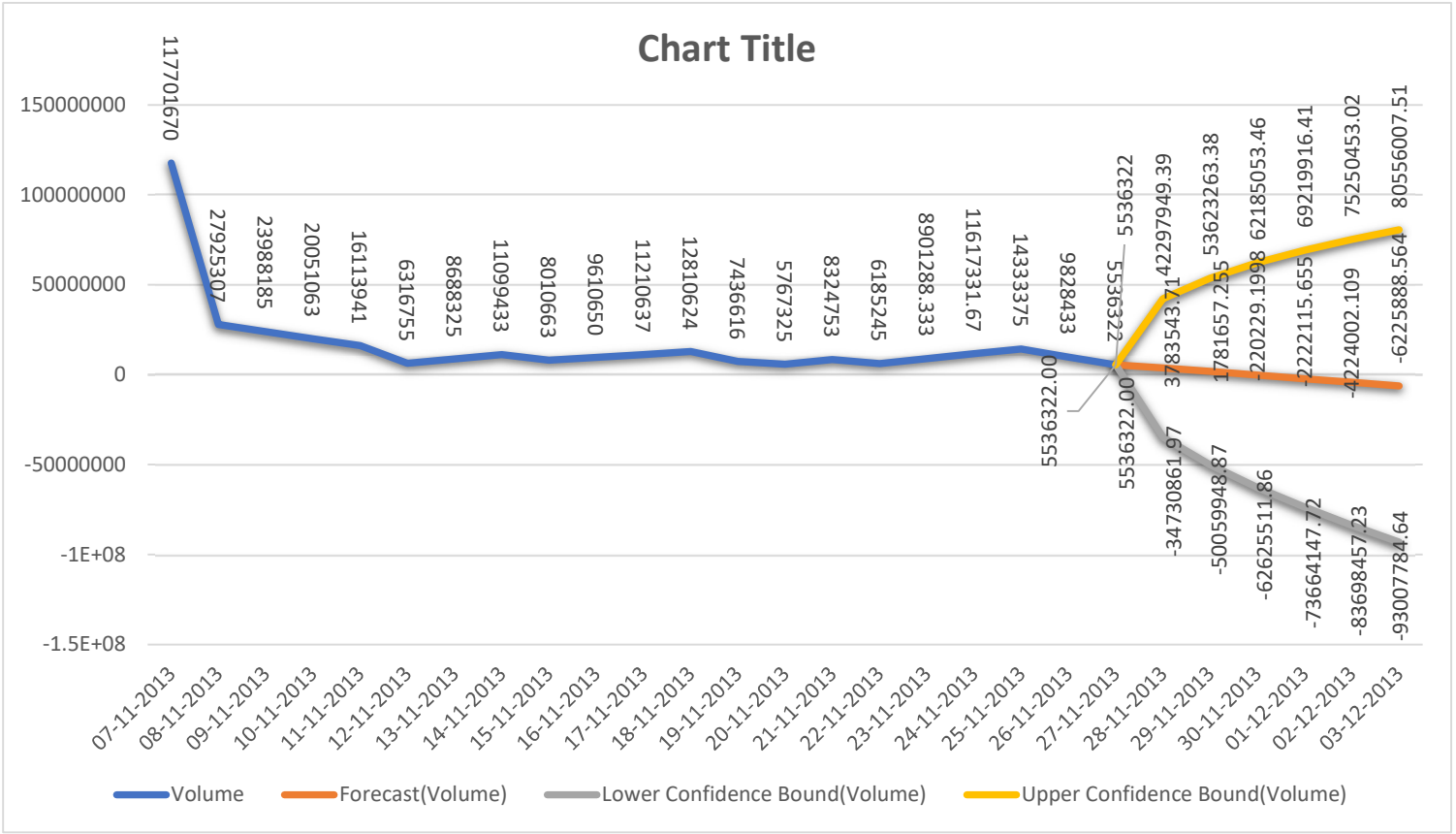
The ARIMA model parameters were selected based on the analysis of the autocorrelation and partial autocorrelation plots of the historical data. The model was then used to generate forecasts for the next 15 days, from November 27, 2013, to December 11, 2013. Along with the point forecasts, the model also provided lower and upper confidence bounds to account for the uncertainty in the predictions.

Forecast Results:

- From November 27, 2013, to December 3, 2013, the predicted trading volumes are highly variable, starting at approximately 5.5 million and showing both positive and negative forecasts.

- The lower confidence bound indicates the minimum expected trading volume, while the upper confidence bound represents the maximum expected trading volume within a 95% confidence interval.

PDate	Volume	Forecast(Volume)	Lower Confidence Bound(Volume)	Upper Confidence Bound(Volume)
07-11-2013	117701670			
08-11-2013	27925307			
09-11-2013	23988185			
10-11-2013	20051063			
11-11-2013	16113941			
12-11-2013	6316755			
13-11-2013	8688325			
14-11-2013	11099433			
15-11-2013	8010663			
16-11-2013	9610650			
17-11-2013	11210637			
18-11-2013	12810624			
19-11-2013	7436616			
20-11-2013	5767325			
21-11-2013	8324753			
22-11-2013	6185245			
23-11-2013	8901288.3			
24-11-2013	11617332			
25-11-2013	14333375			
26-11-2013	9828433			
27-11-2013	5536322	5536322	5536322.00	5536322.00
28-11-2013		3783543.71	-34730861.97	42297949.39
29-11-2013		1781657.255	-50059948.87	53623263.38
30-11-2013		-220229.1998	-62625511.86	62185053.46
01-12-2013		-2222115.655	-73664147.72	69219916.41
02-12-2013		-4224002.109	-83698457.23	75250453.02
03-12-2013		-6225888.564	-93007784.64	80556007.51



Conclusion: -

The forecast analysis for Twitter's trading volumes over the given period reveals significant variability and uncertainty, as reflected in the wide range of confidence bounds. While the predicted volumes suggest a potential decrease, the negative values in the lower confidence bounds indicate substantial volatility. This variability underscores the unpredictable nature of stock trading volumes and highlights the need for careful consideration by investors and analysts when making decisions based on these forecasts.

The wide confidence intervals suggest that while the model can provide a central estimate, there is a considerable degree of uncertainty. Investors and stakeholders should use these forecasts as one of several tools in their decision-making process, considering other factors and analyses to form a comprehensive view of the market dynamics.