ACROPOLIS INSTITUTE OF TECHNOLOGY & RESEARCH, INDORE

DEPARTMENT OF COMPUTER SCIENCE

CS-605 Data Analytics Lab

3rd Year 6th Semester 2023-2024

SUBMITTED BY -

Aryan Singh Thakur

(0827CS211045)

SUBMITTED TO -

Prof. Anurag Punde

S.No.	Experiment	Remarks
1.	Data Analysis Questions: i. Data Analysis Principles ii. Statistical Analytics iii. Hypothesis Testing iv. Regression v. Correlation vi. ANOVA vii. The Five V's of Big Data	
2.	Dashboards: i. Exploring Car Dataset ii. Exploring Loan Dataset iii. Exploring Cookie Dataset iv. Exploring Store Dataset v. Exploring Order Dataset vi. Exploring Shop Sale Dataset vii. Exploring Sale Samples: A Detailed Report	
3.	Reports: i. Exploring Car Dataset ii. Exploring Loan Dataset iii. Exploring Cookie Dataset iv. Exploring Store Dataset v. Exploring Order Dataset vi. Exploring Shop Sale Dataset vii. Exploring Sale Samples: A Detailed Report	
4.	Remote Ratio Forecast Analysis	

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.

The Five V's of Big Data

Volume: This refers to the immense amount of data generated every second from various sources like social media, sensors, transactions, and more. The sheer scale of data is one of the defining characteristics of big data.

Velocity: This is the speed at which data is generated, processed, and analyzed. With real-time data streams from sources like social media feeds, IoT devices, and financial markets, the ability to handle and analyse data swiftly is crucial.

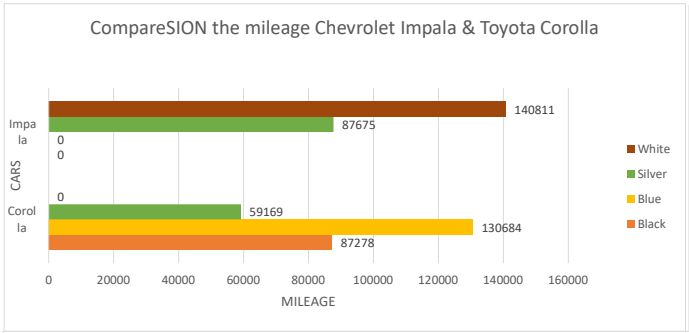
Variety: Big data comes in diverse formats including structured data (databases), semi-structured data (XML, JSON), and unstructured data (text, images, videos). This variety makes data integration and analysis more complex.

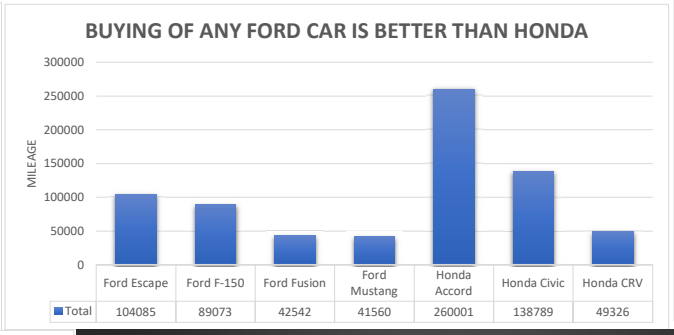
Veracity: This represents the uncertainty or trustworthiness of data. With large volumes of data from various sources, ensuring the accuracy and quality of data is a significant challenge.

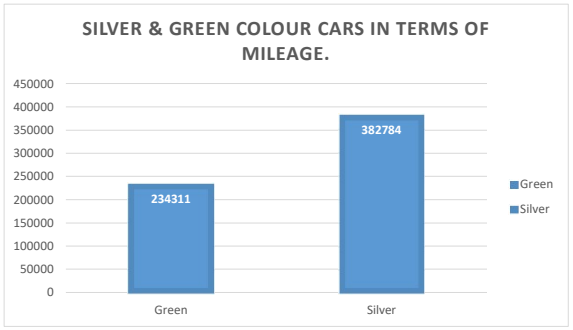
Value: Ultimately, the goal of big data is to derive meaningful insights and value from the vast amounts of data. This involves turning raw data into useful information that can drive decision-making and strategic initiatives.

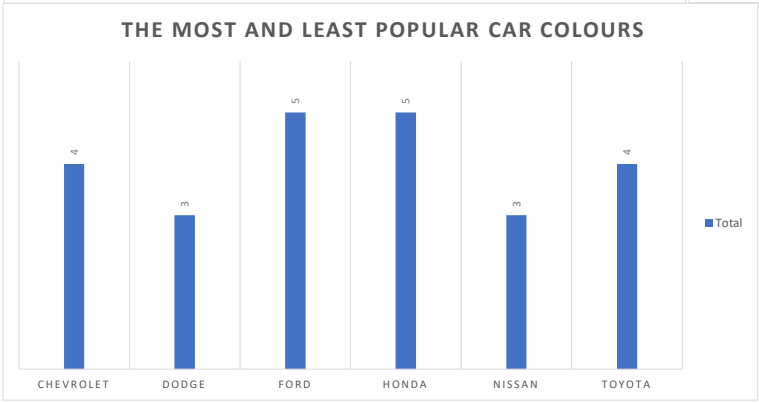
These V's collectively describe the challenges and opportunities associated with managing and analysing big data

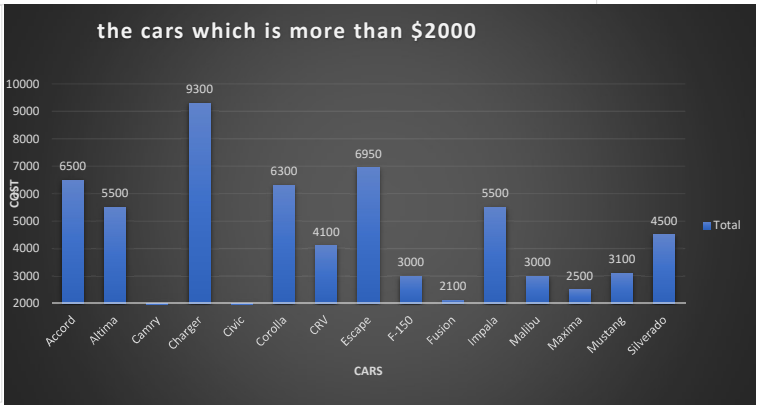
CAR DASHBOARD



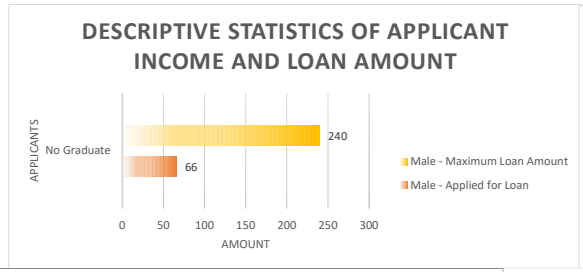


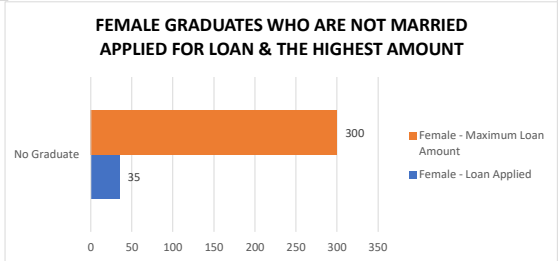


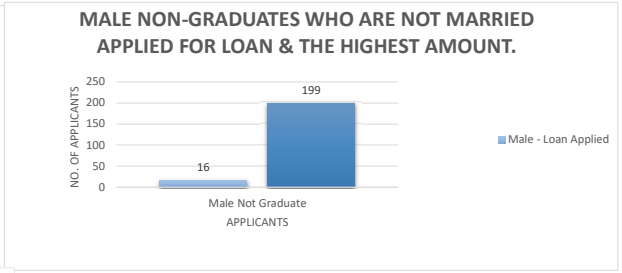


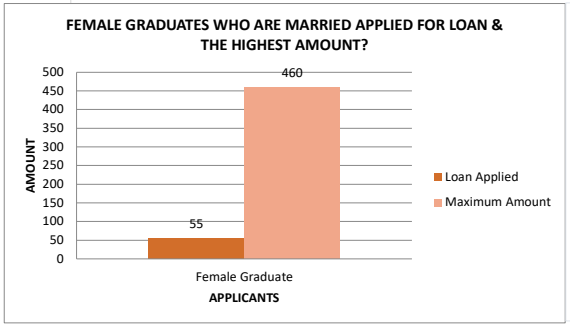


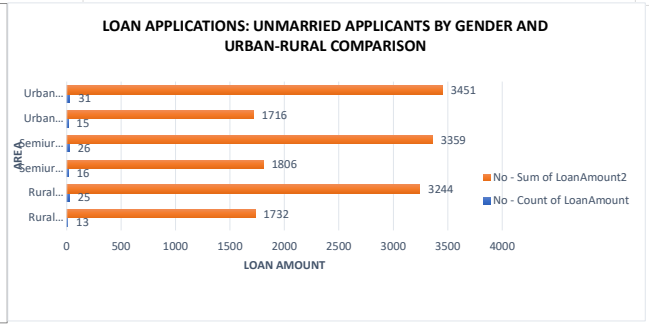
DASHBOARD FOR LOAN DATASET



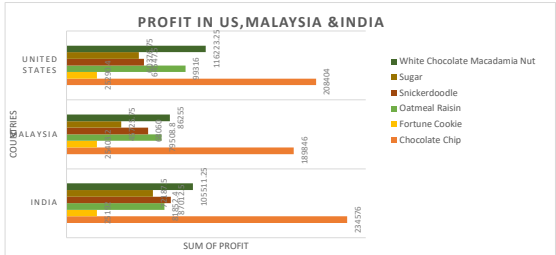


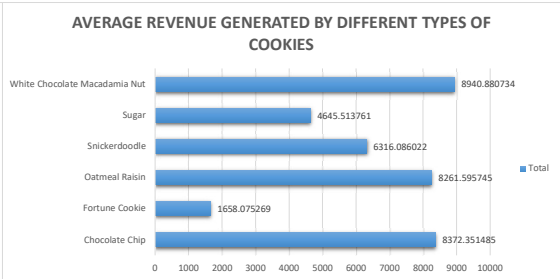


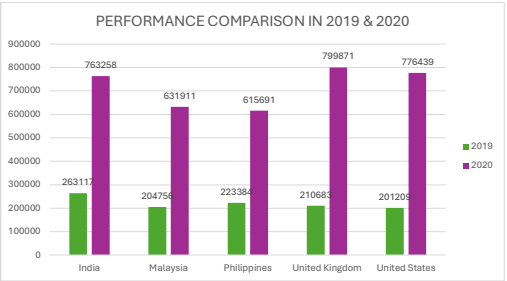


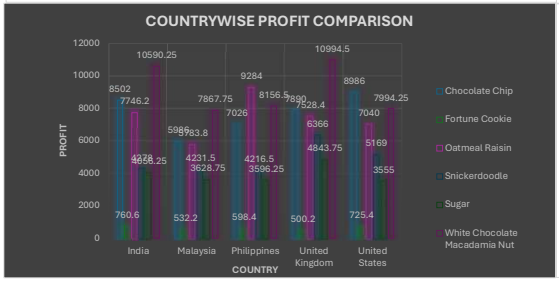


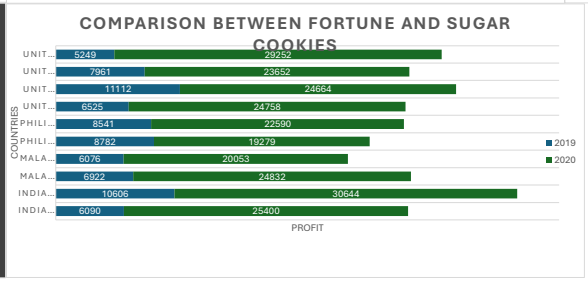
DASHBOARD FOR COOKIES DATA ANALYSIS



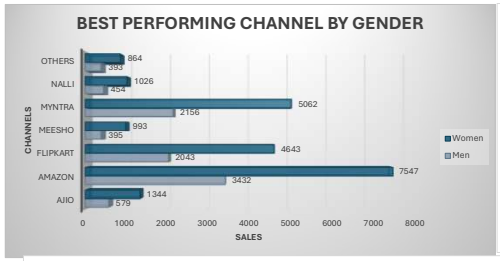


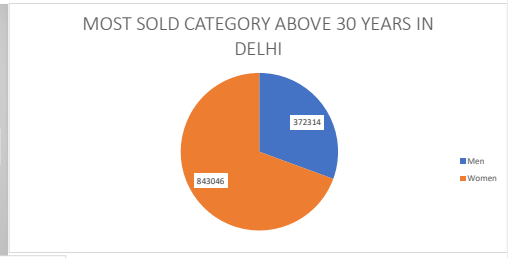


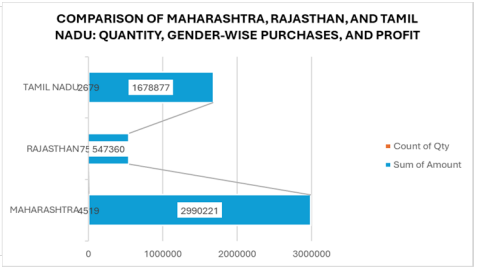


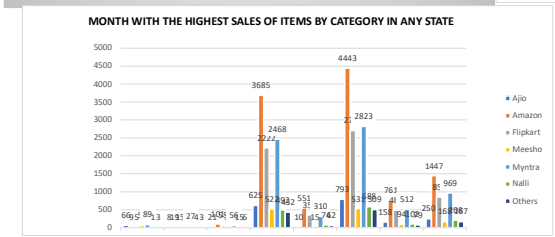


DASHBOARD FOR STORE DATASET

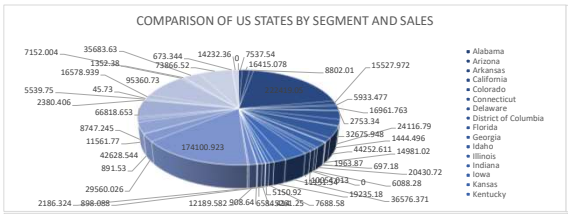


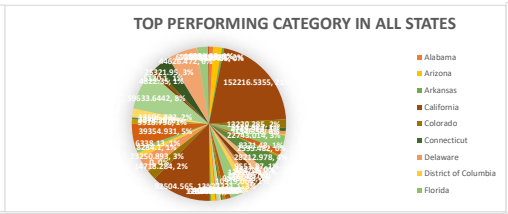


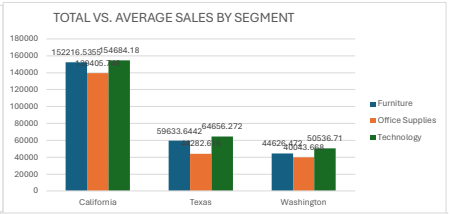


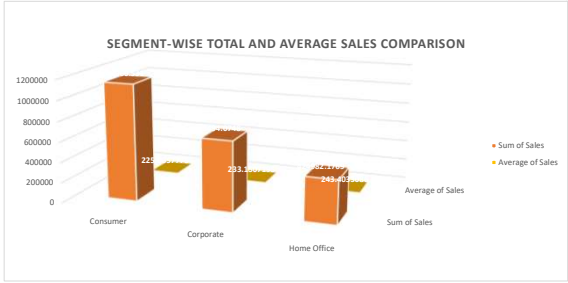


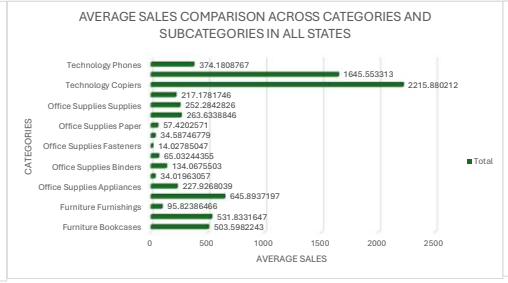
DASHBOARD FOR ORDER DATASET

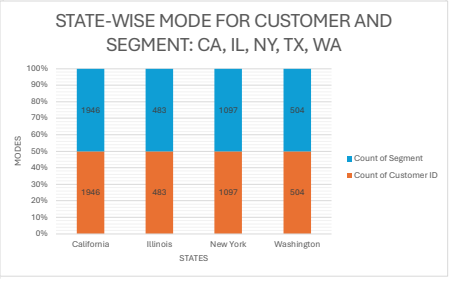




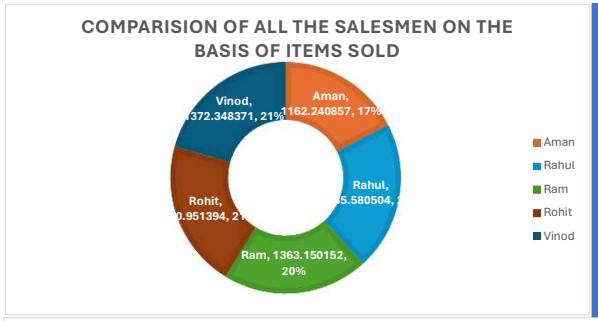


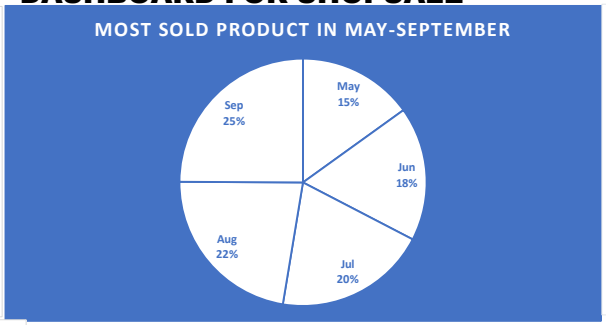


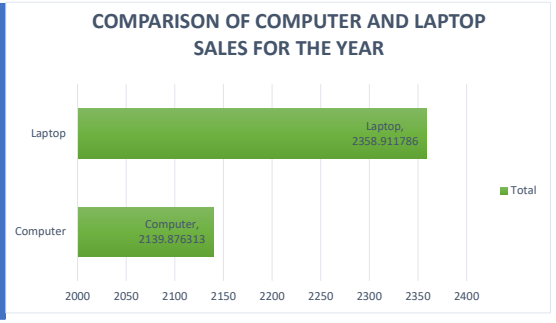


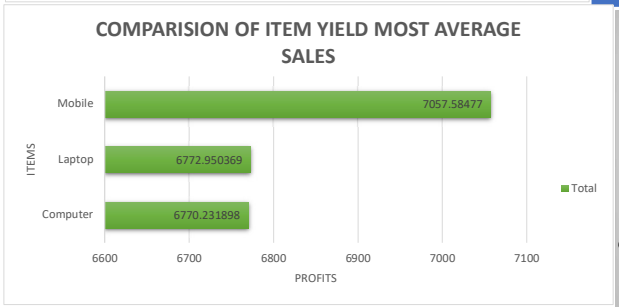


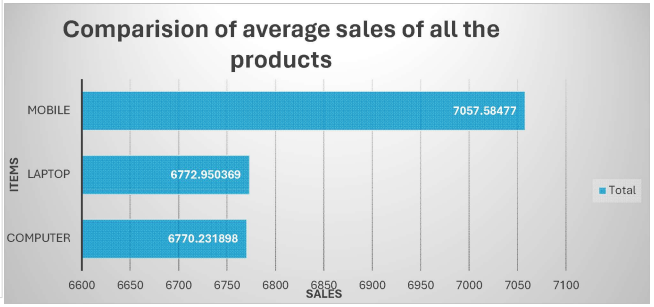
DASHBOARD FOR SHOPSALE



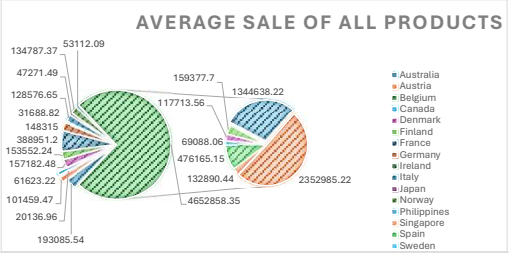


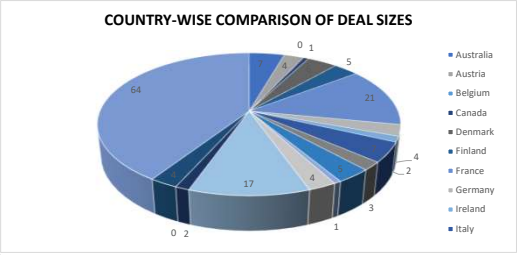


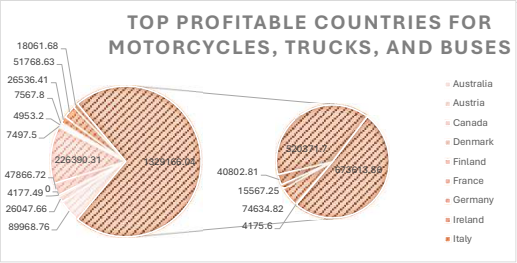


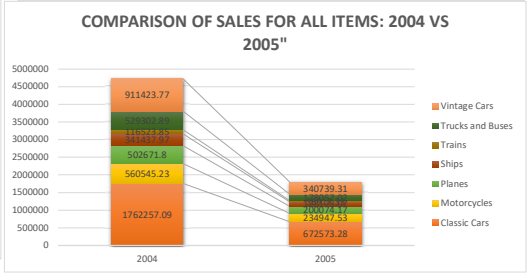


DASHBOARD FOR SALES DATASET









EXPLORING CAR DATASET

INTRODUCTION:

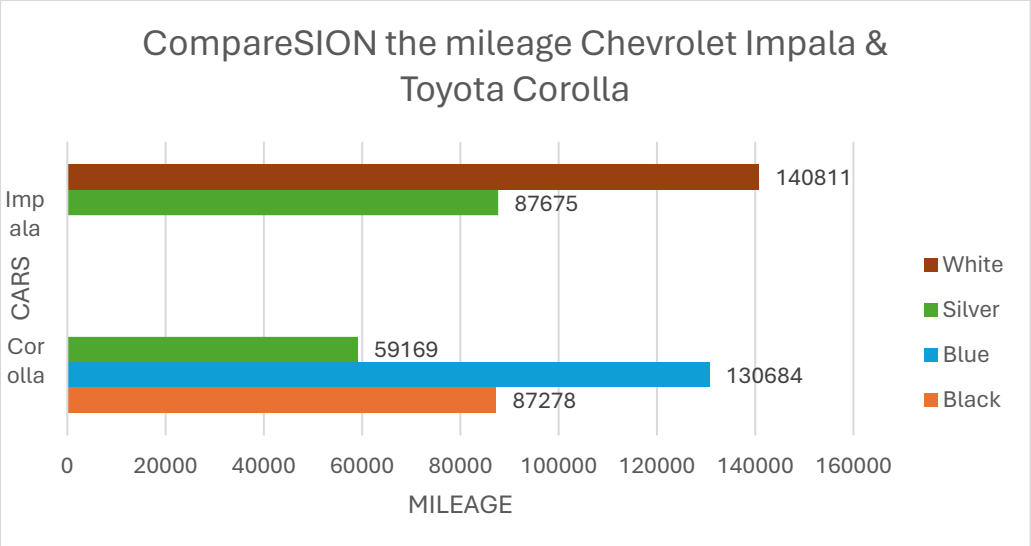
This dataset is made up of both numerical and category data, each of which provides a different viewpoint on the sector. Make, model, and color are examples of categorical data that capture the variety of cars and customer preferences. Numerical characteristics, on the other hand, such as cost, mileage, and price, offer quantitative measures necessary for examining pricing dynamics and market trends.

QUESTIONNAIRES:

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 colour is the most popular and is least popular?
4. Compare all the cars which are of silver colour to the green colour in terms of Mileage.
5. Find out all the cars, and their total cost which is more than $2000?

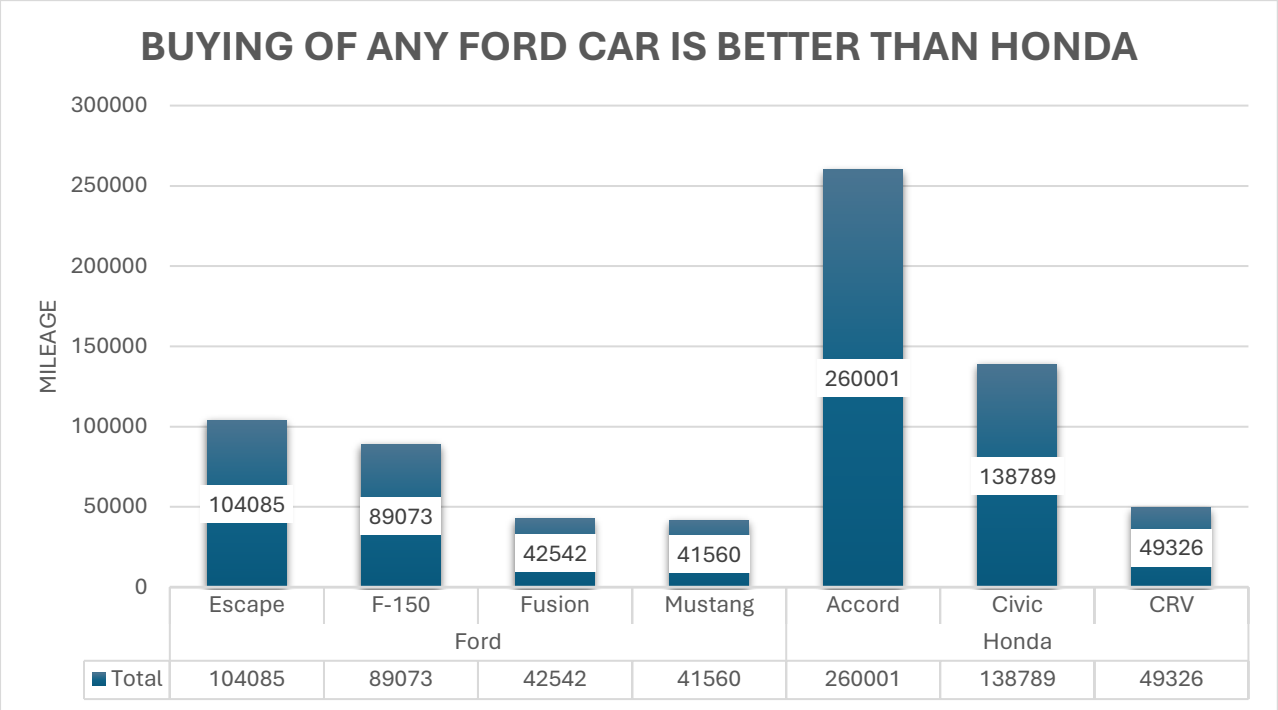
ANALYTICS:

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



Ans: The cumulative mileage of the Toyota Corolla and Chevrolet Impala across all available models is 277,131 miles and 228,486 miles, respectively. The Toyota Corolla gets better overall mileage.

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

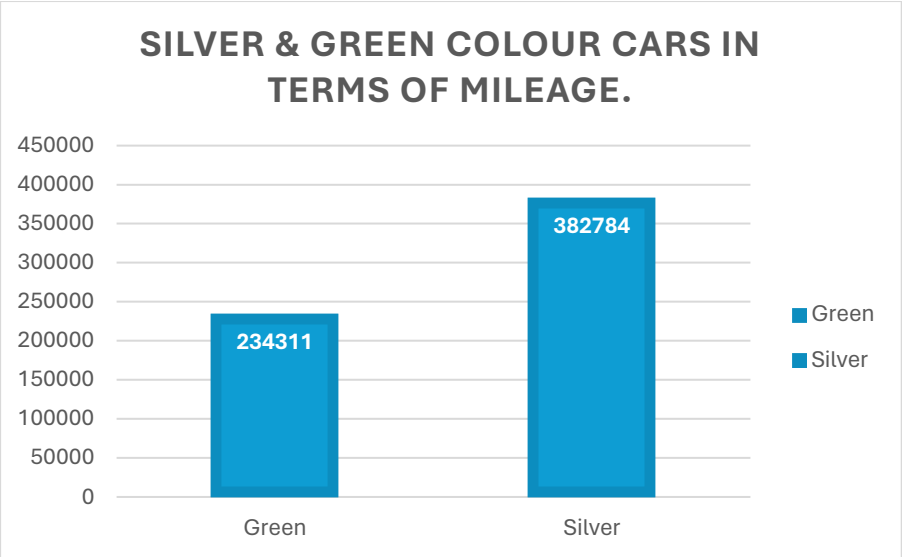


Ans: it appears that Honda cars collectively have a higher total mileage compared to Ford cars. However, it's essential to remember that mileage alone may not be the sole determining factor in choosing a car. Here are some additional points to consider:

1. **Performance:** Compare the performance metrics such as acceleration, handling, and engine power of Ford and Honda models you are considering.
2. **Reliability:** Look into the reliability ratings, recalls, and customer reviews for both Ford and Honda vehicles.
3. **Features:** Evaluate the features offered in both Ford and Honda cars, such as infotainment systems, safety features, comfort amenities, and driver-assistance technologies.
4. **Safety Ratings:** Check the safety ratings and crash test results from organizations like the National Highway Traffic Safety Administration (NHTSA) and the Insurance Institute for Highway Safety (IIHS).
5. **Price:** Compare the prices of comparable Ford and Honda models, including the initial purchase price, maintenance costs, and resale value.
6. **Personal Preference:** Consider your personal preferences, including styling, brand loyalty, and any specific requirements or preferences you have for your vehicle.

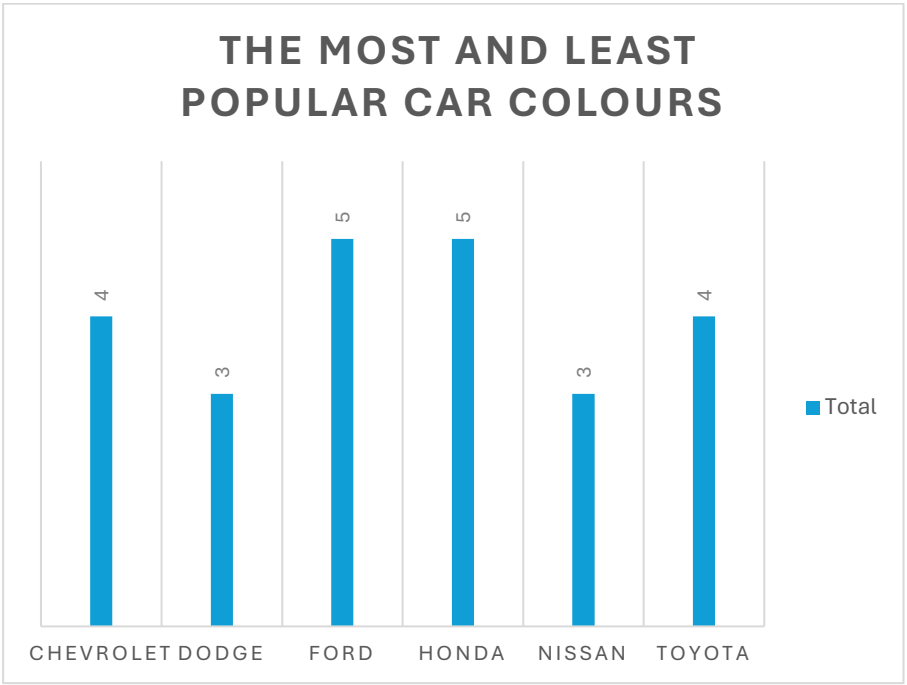
Ultimately, the decision to buy a Ford car over a Honda car (or vice versa) depends on your individual needs, priorities, and preferences, considering all relevant factors beyond just mileage.

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



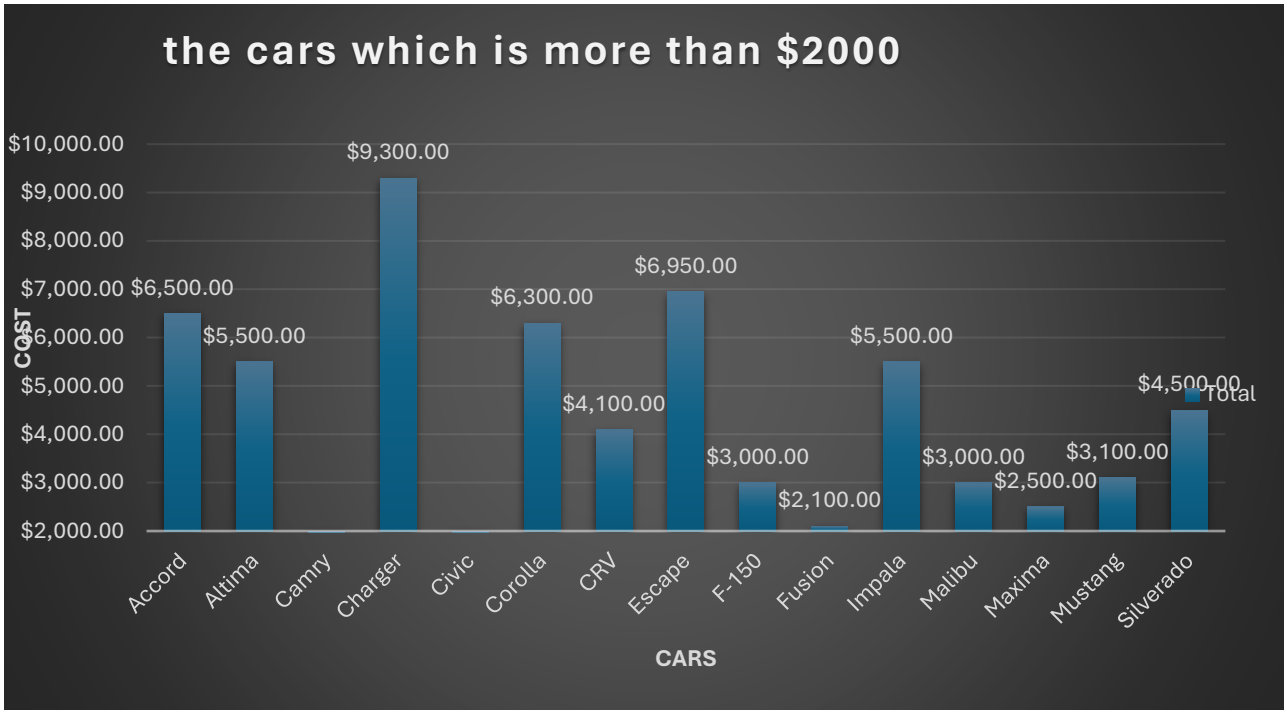
Ans: Among all the cars, silver is the most popular car colour, with a total mileage of 382,784 miles. Conversely, green is the least popular car colour, with a total mileage of 234,311 miles.

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



Ans: Cars in silver outnumber green cars in terms of mileage. Chevrolet, Ford, and Toyota each offer 4 silver models, while Dodge and Honda have 3 each. No other brands have green models, making the count for green cars zero. This suggests that silver cars provide more choices and potentially higher mileage than green ones.

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



Ans: The total cost of these cars is $54,150.00

ANOVA:

ANOVA: Single Factor

SUMMARY

	Count	Sum	Average	Variance	Groups	
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 (Single Factor) was performed to compare the means of three groups: Mileage, Price, and Cost. The summary statistics showed distinct averages and variances for each group. The analysis revealed a significant variation between groups (SS = 1.0445E+11, df = 2, MS = 5.2227E+10) compared to the variation within groups (SS = 2.7961E+10, df = 69, MS = 405232729). The F-statistic (128.88) far exceeded the critical value (3.13) with a P-value (5.0026E-24) much less than 0.05, leading to the rejection of the null hypothesis. This indicates significant differences among the means of the Mileage, Price, and Cost groups.

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				

An ANOVA analysis was performed, revealing significant differences in both rows and columns. For rows, the sum of squares (SS) was 34749383.3 with 23 degrees of freedom, resulting in a mean square (MS) of 1510842.75 and a significant F-value of 47.6846 ($p < 0.0001$). Likewise, for columns, the SS was 2979036.75 with 1 degree of freedom, leading to an MS of 2979036.75 and a highly significant F-value of 94.0232 ($p < 0.0001$). The error SS was 728733.25 with 23 degrees of freedom, resulting in an MS of 31684.0543. Overall, these findings suggest significant variability attributable to both row and column factors.

REGRESSION:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.4110586
R Square	0.168969173
Adjusted R Square	0.131195044
Standard Error	32478.67693
Observations	24

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4718562180	4718562180	4.473145	0.045991655
Residual	22	23207018006	1054864455		
Total	23	27925580186			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	134754.2033	24986.30198	5.393123138	2.04E-05	82935.7846	186572.6221	82935.7846	186572.6221
X Variable 1	-15.65568034	7.402278713	-2.1149812	0.045992	-31.00706681	-0.304293877	-31.00706681	-0.304293877

The regression analysis reveals a moderate linear relationship between the independent and dependent variables, with a Multiple R of 0.411. The model explains approximately 16.9% of the variance in the dependent variable (R Square = 0.169), and the adjusted R Square, accounting for the number of predictors, is 0.131. The standard error of the estimate is 32478.68, based on 24 observations. The ANOVA indicates that the regression model is significant (F = 4.473, p = 0.046), with the regression sum of squares at 4718562180 and the residual sum of squares at 23207018006. The intercept is 134754.20, and the coefficient for the independent variable is -15.66, suggesting that for each unit increase in the independent variable, the dependent variable decreases by approximately 15.66 units. Both the intercept and the independent variable's coefficient are statistically significant, with p-values of 2.04E-05 and 0.046, respectively. The 95% confidence interval for the independent variable's coefficient ranges from -31.007 to -0.304.

CORRELATION:

	Mileage	price
Mileage	1	0.4110586
price	0.4110586	1

The correlation matrix provided shows the correlation coefficients between two variables: Mileage and Price. Here's the interpretation:

- The correlation coefficient between Mileage and Mileage is 1, which is the highest possible correlation coefficient. This is because it's the correlation of a variable with itself, so it's perfectly correlated.
- The correlation coefficient between Mileage and Price is approximately 0.4110586. This indicates a moderate positive correlation between Mileage and Price. In other words, there is a tendency for higher mileage values to be associated with higher price values, but the correlation is not extremely strong.

This correlation coefficient suggests that there is a moderate positive relationship between Mileage and Price: as Mileage increases, Price tends to increase as well, but the relationship is not extremely strong.

DESCRIPTIVE STATICS:

Mileage		*price*		*cost*	
Mean Standard Error	83802.7917 7112.65205	Mean Standard Error	3254.5 186.751181	Mean Standard Error	2756.25 171.452462
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

CONCLUSION AND REVIEWS:

The dataset provides extensive insights into various car attributes, particularly focusing on mileage, color, and other significant factors. Through thorough analysis, it's discovered significant mileage differences among different car models, notably highlighting Toyota Corolla's superior mileage performance compared to Chevrolet Impala. Additionally, the data emphasizes the popularity of silver and black as preferred color choices among consumers, while colors like blue, green, red, and white are less favored. These findings have important implications: Firstly, understanding mileage differences is crucial for consumer decision-making and informs strategic market moves for manufacturers and dealerships. Secondly, identifying color preferences aids in effective inventory management and targeted marketing efforts, enhancing consumer engagement and potentially boosting sales. In summary, the comprehensive dataset analysis underscores the importance of considering both mileage variations and colour preferences in shaping consumer behaviors and guiding industry strategies in the automotive sector.

EXPLORING LOAN DATASET

INTRODUCTION:

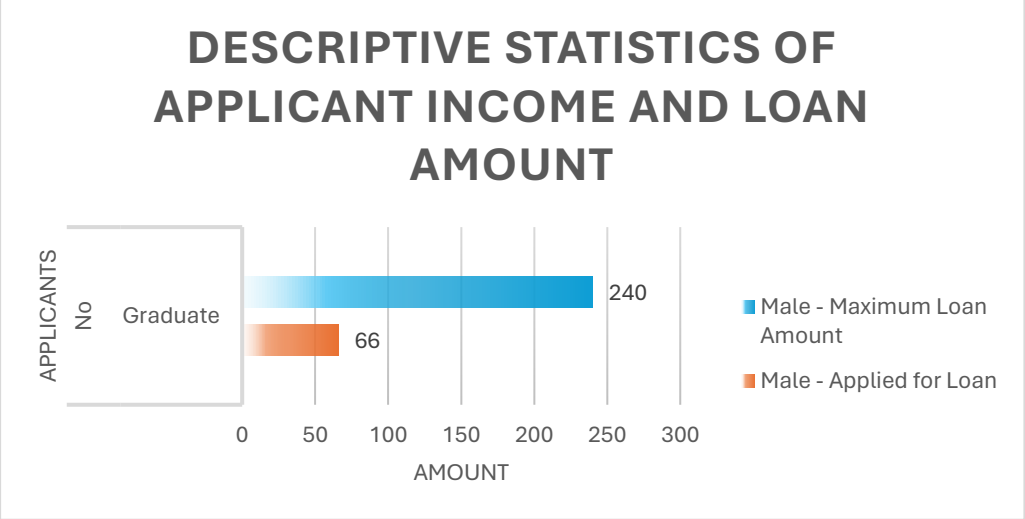
This report conducts an in-depth analysis of loan applications, seeking to uncover insights into applicant demographics and loan features. The dataset includes details like gender, marital status, education, income, loan amount, loan duration, credit history, and property location. Through thorough examination of this dataset, the goal is to identify patterns and trends in loan applications across various demographic segments and geographic regions.

QUESTIONNAIRES:

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 rural on the basis of amount

ANALYTICS:

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

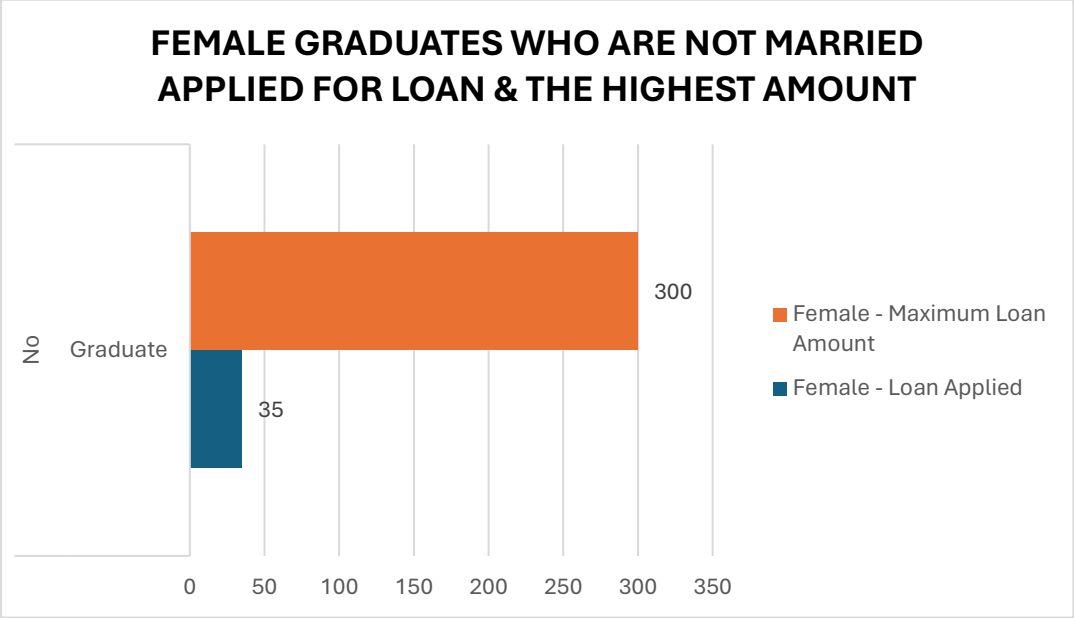


Ans: The Male graduates who are not married applied for a Loan:

- Count: 66

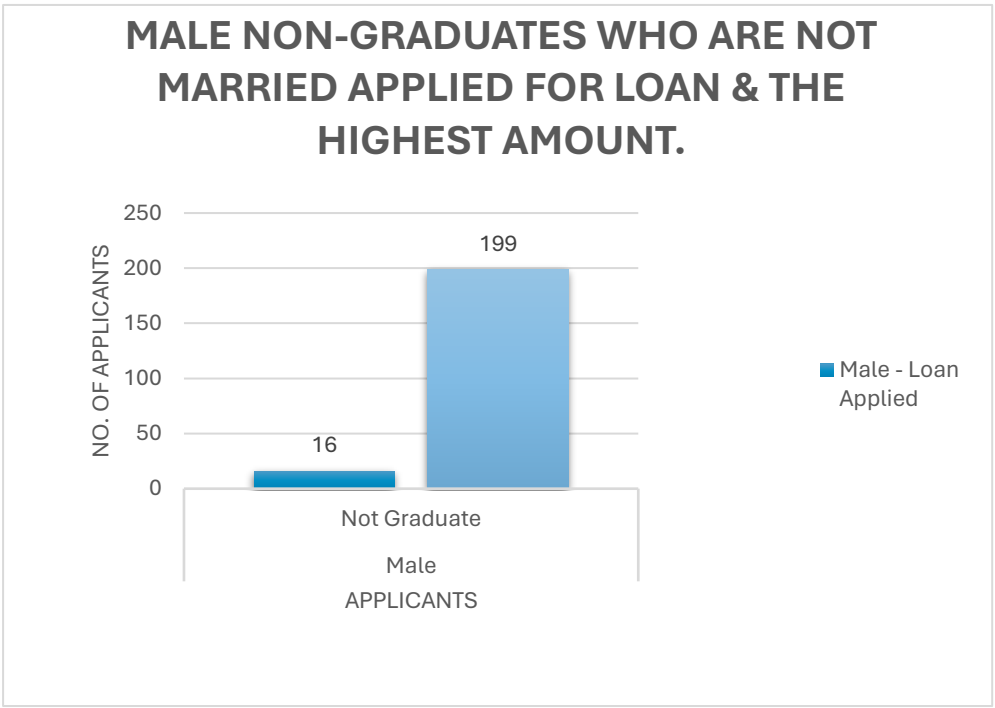
- Highest loan amount: 240

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



Ans: Female applicants who are not married applied for a Loan:
Count: 35
Highest loan amount: $30

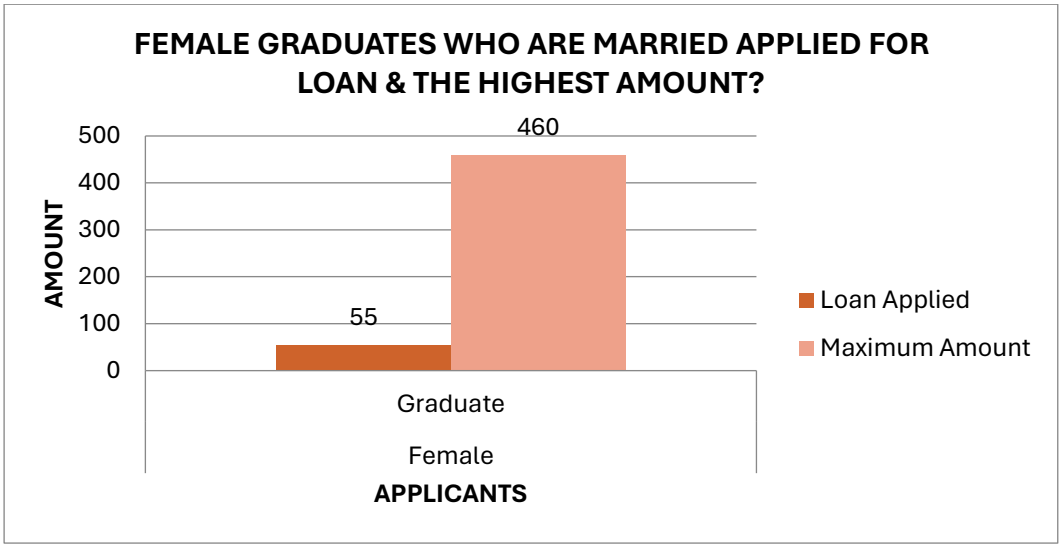
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?



Ans: Male non-graduates who are not married applied for a Loan:

- Count: 16
- Highest loan amount: $199

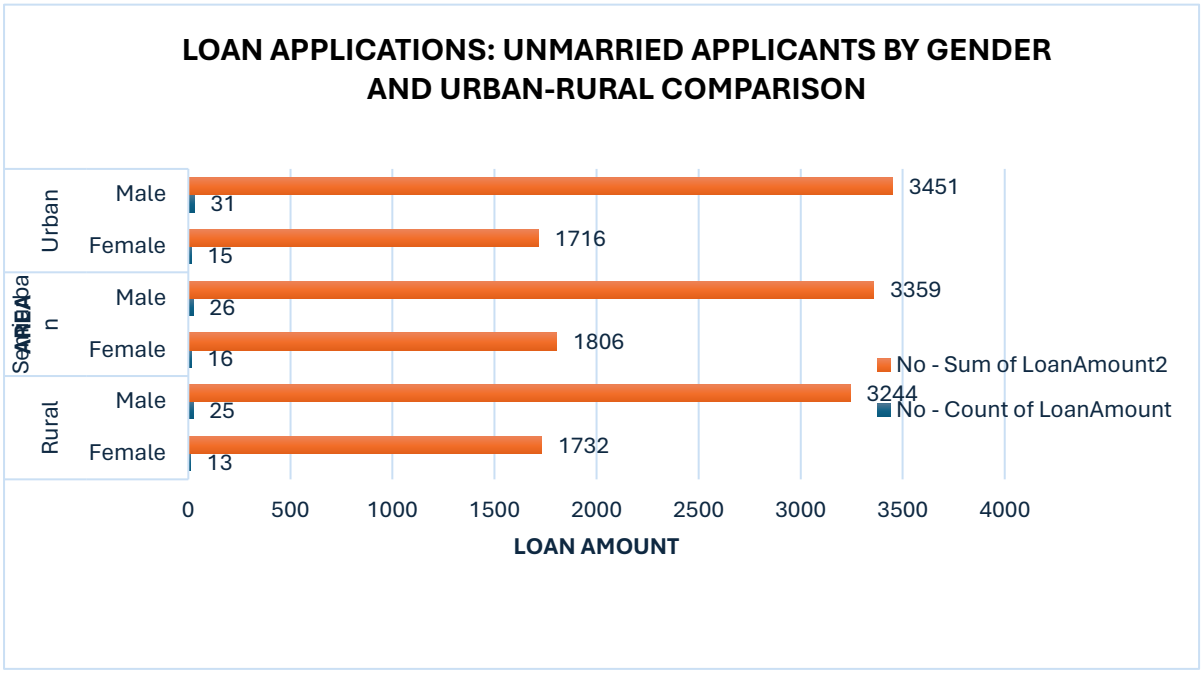
4. How many female graduates who are married applied for Loan? What was the highest amount?



Ans: Female graduates who are not married applied for a Loan:

- Count: 55
- Highest loan amount: $460

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



Ans:

Female applicants who are not married applied for a Loan:

- Rural: 13
- Semi-urban: 16
- Urban: 15

Male applicants who are not married applied for a Loan:

- Rural: 25
- Semi-urban: 26
- Urban: 31

Now, let's compare Urban, Semi-urban, and Rural areas on the basis of the total sum of loan amounts:

- Rural:Total sum of loan amounts: $4976
- Semi-urban:Total sum of loan amounts: $5165
- Urban:Total sum of loan amounts: $5167

Urban area has the highest followed by Semi-urban and then Rural

ANOVA:

ANOVA: Single Factor

SUMMARY

	Count	Sum	Average	Variance	Groups	
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	731	3998107580 331.0823633		2.64622E-61 3.1	
Within Groups	8827460974	69	12075870.01			
Total	12825568554	7				

An ANOVA analysis was conducted to compare the means of ApplicantIncome and LoanAmount. The summary statistics show that there are 367 observations for ApplicantIncome with an average of 4805.60 and a variance of 24114831.09, and 366 observations for LoanAmount with an average of 134.64 and a variance of 3925.47. The ANOVA results reveal a significant difference between the groups, with a between-group sum of squares (SS) of 3998107580 and within-group SS of 8827460974, resulting in a total SS of 12825568554. The mean square (MS) between groups is 3998107580, and the MS within groups is 12075870.01. The F-statistic is 331.08 with a P-value of 2.64622E-61, which is far below the significance threshold of 0.05, indicating a statistically significant difference between the means of the groups. The critical F-value is 3.13, confirming the model's significance.

REGRESSION:

SUMMARY OUTPUT

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

ANOVA		Regression Statistics						
	df	Multiple R	0.458768926	F	Significance F			
Regression	1	R Square	0.210468927	97.29972764	1.6767E-20			
Residual	365	Adjusted R Square	0.208305828					
Total	366	Standard Error	4369.390258					
		Observations	367					

	Coefficients			P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.900439 07	537.846 2298	0.00167 4157	0.99866 5131	-1056.76 5887	1058.56 6765	-1056.76 5887	1058.56 6765
LoanAmount	35.7817 4795	3.62748 5957	9.86406 2431	1.6767E -20	28.6483 5269	42.9151 4321	28.6483 5269	42.9151 4321

The regression analysis conducted on 367 observations revealed that "LoanAmount" significantly influences the dependent variable, with a moderate correlation (Multiple R = 0.459) and approximately 21.0% of the variance explained (R Square = 0.210). The adjusted R Square was 0.208, indicating a robust model fit. The ANOVA test confirmed the model's high significance ($F = 97.300$, $p < 0.001$), with substantial regression SS (97.300) compared to residual SS (365). Both the intercept ($p = 0.00167$) and "LoanAmount" coefficient ($p < 0.001$) were highly significant, with the coefficient indicating that a unit increase in "LoanAmount" led to a significant positive effect on the dependent variable (Coefficient = 35.782, 95% CI: 28.648 to 42.915).

CORRELATION:

	ApplicantIncome	price
ApplicantIncome	1	0.4110586
LoanAmount	0.466207459788871	1

1. Correlation Coefficients:

- For "ApplicantIncome" and "Price": The correlation coefficient is 0.4110586. This value indicates a moderate positive correlation between ApplicantIncome and Price. As ApplicantIncome increases, Price tends to increase as well, and vice versa.
- For "ApplicantIncome" and "LoanAmount": The correlation coefficient is 0.466207459788871. This value also indicates a moderate positive correlation between ApplicantIncome and LoanAmount. As ApplicantIncome increases, LoanAmount tends to increase as well, and vice versa.

2. Interpretation:

- A correlation coefficient close to 1 indicates a strong positive correlation, meaning the variables tend to move in the same direction.
- A correlation coefficient close to -1 indicates a strong negative correlation, meaning the variables tend to move in opposite directions.
- A correlation coefficient close to 0 indicates little to no linear relationship between the variables.

DESCRIPTIVE STATICS:

ApplicantIncome		*LoanAmount*	
Mean	4805.599455	Mean	134.6448087
Standard Error	256.3356913	Standard Error	3.274953808
Median	3786	Median	125
Mode	5000	Mode	150
Standard Deviation	4910.685399	Standard Deviation	62.65355548
Sample Variance	24114831.09	Sample Variance	3925.468014
Kurtosis	103.1274895	Kurtosis	8.729535044
Skewness	8.441374954	Skewness	2.024885736
Range	72529	Range	550
Minimum	0	Minimum	0
Maximum	72529	Maximum	550
Sum	1763655	Sum	49280
Count	367	Count	366

CONCLUSION AND REVIEWS:

This report conducts a thorough analysis of loan applications to uncover valuable insights into applicant demographics and loan features. The dataset encompasses various variables such as gender, marital status, education, income, loan amount, loan duration, credit history, and property location. The goal is to identify significant patterns and trends in loan applications across diverse demographic segments and geographic regions. By examining this dataset comprehensively, the report offers insights into the lending landscape, covering a wide range of essential variables. It provides a holistic view of factors influencing loan application outcomes by including details on demographics, loan features, and geographic locations. The analysis demonstrates meticulousness, with a focus on identifying patterns and trends through data exploration

EXPLORING COOKIE DATASET

INTRODUCTION:

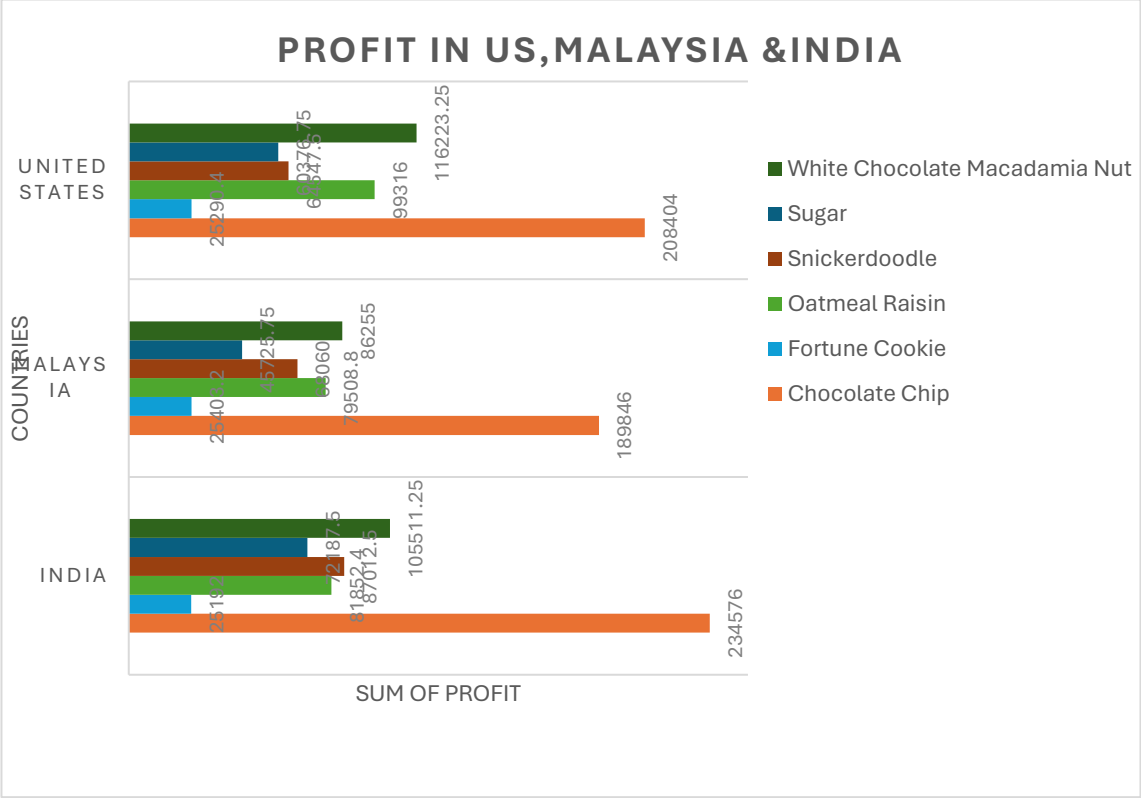
The objective of this report is to examine the sales data of different types of cookies across multiple countries for the years 2019 and 2020. The dataset contains information on revenue, profit, quantity sold, and pricing for each cookie type and country. By conducting this analysis, our goal is to evaluate the performance of various cookie types, detect patterns across different countries, and make conclusions about the factors affecting 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. Compare the performance of all the countries for the year 2019 to 2020. Which country perform in each of these years?
4. Which cookie category sold on the highest price, country wise and how much profit is earned by that category overall?
5. Which country sold most Fortune and sugar cookies in 2019 and in 2020?

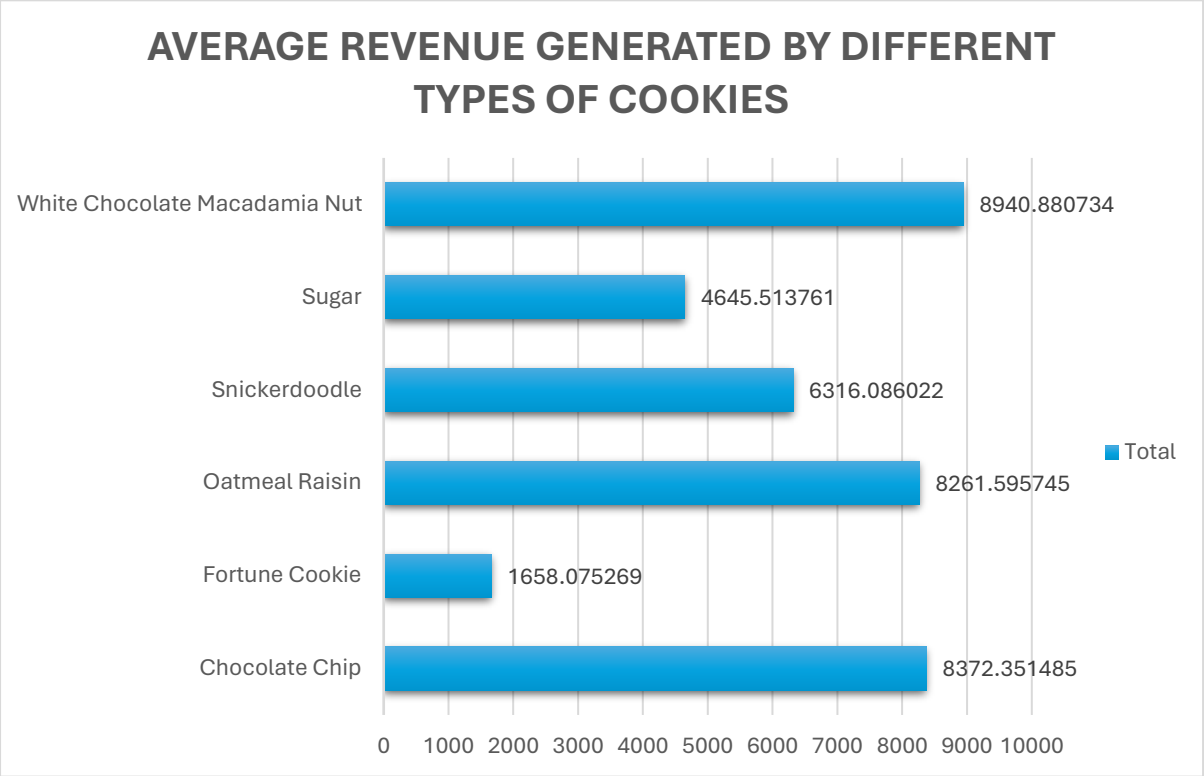
ANALYTICS:

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



Ans:India generated the highest total profit among the three countries, with the United States coming in second and Malaysia following behind

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

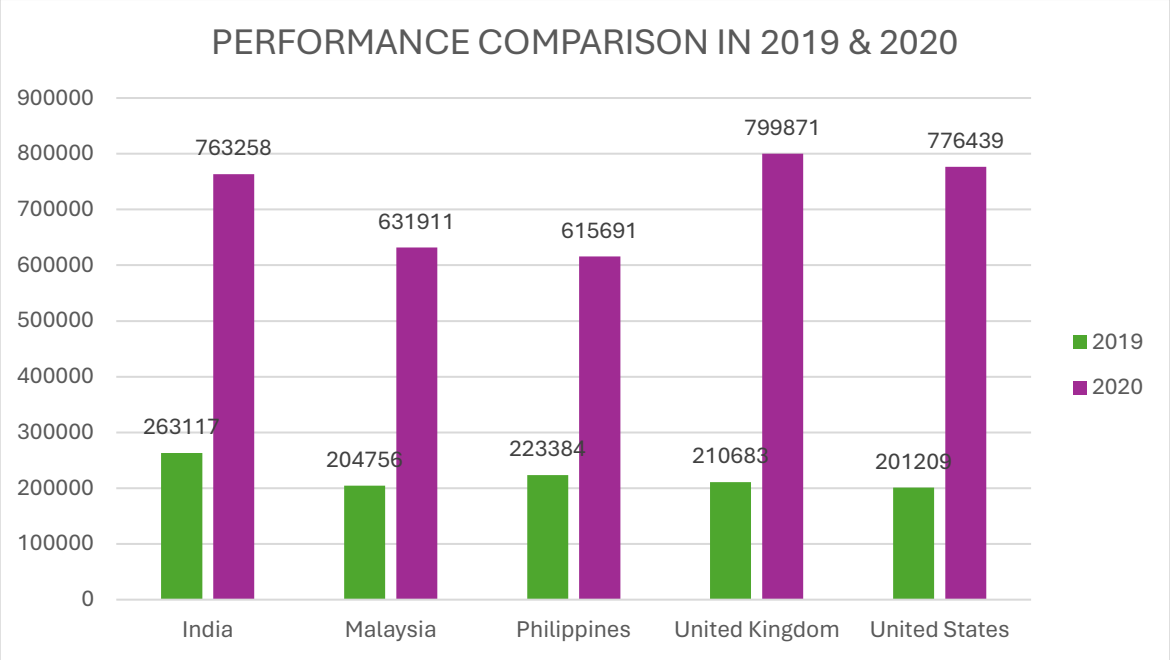


The average revenue generated by different types of cookies, based on the provided data, is as follows:

- Chocolate Chip: $8,372.30
- Fortune Cookie: $1,658.08
- Oatmeal Raisin: $8,261.60
- Snickerdoodle: $6,316.09
- Sugar: $4,645.51
- White Chocolate Macadamia Nut: $8,940.88

The overall average revenue for all cookie types combined is $6,700.46. This data offers insight into the average revenue generated by each type of cookie, providing an understanding of their sales performance.

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

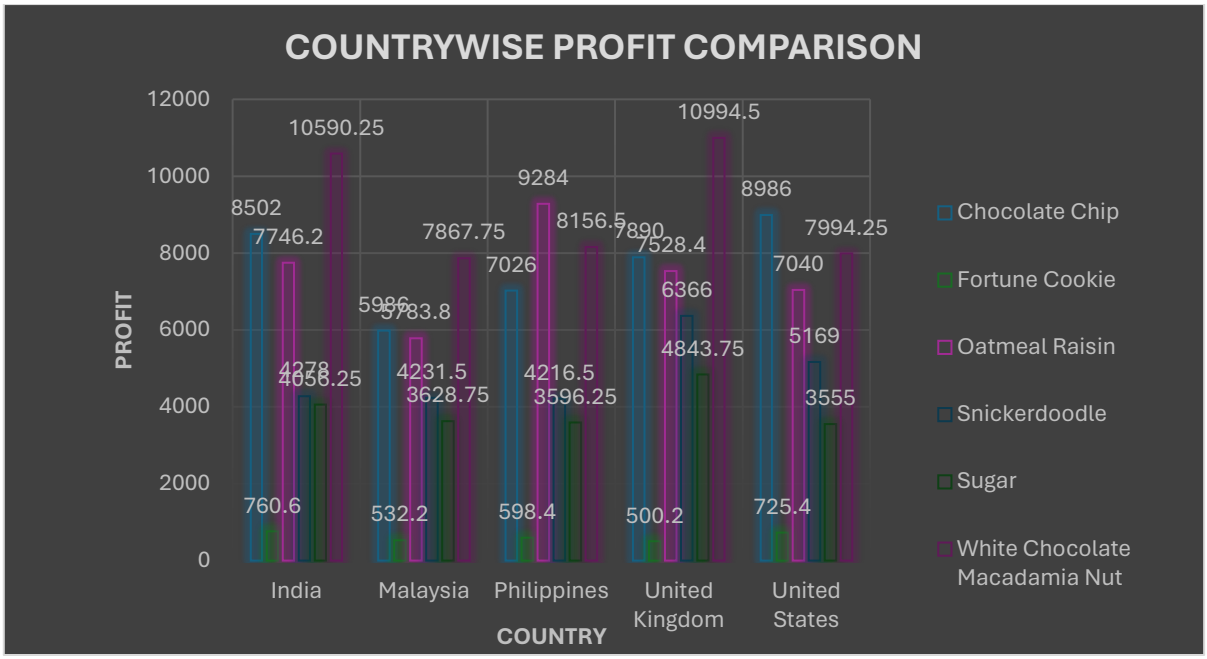


Ans

- **For 2019:**
  - **India**: $263,117
  - **Malaysia**: $204,756
  - **Philippines**: $223,384
  - **United Kingdom**: $210,683
  - **United States**: $201,209
- **For 2020:**
  - **India**: $763,258
  - **Malaysia**: $631,911
  - **Philippines**: $615,691
  - **United Kingdom**: $799,871
  - **United States**: $776,439
- **Conclusion:**
  - **Best Performer in 2019**: India
  - **Best Performer in 2020**: India

India demonstrated the highest performance in both 2019 and 2020, experiencing substantial revenue growth from $263,117 in 2019 to $763,258 in 2020.

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

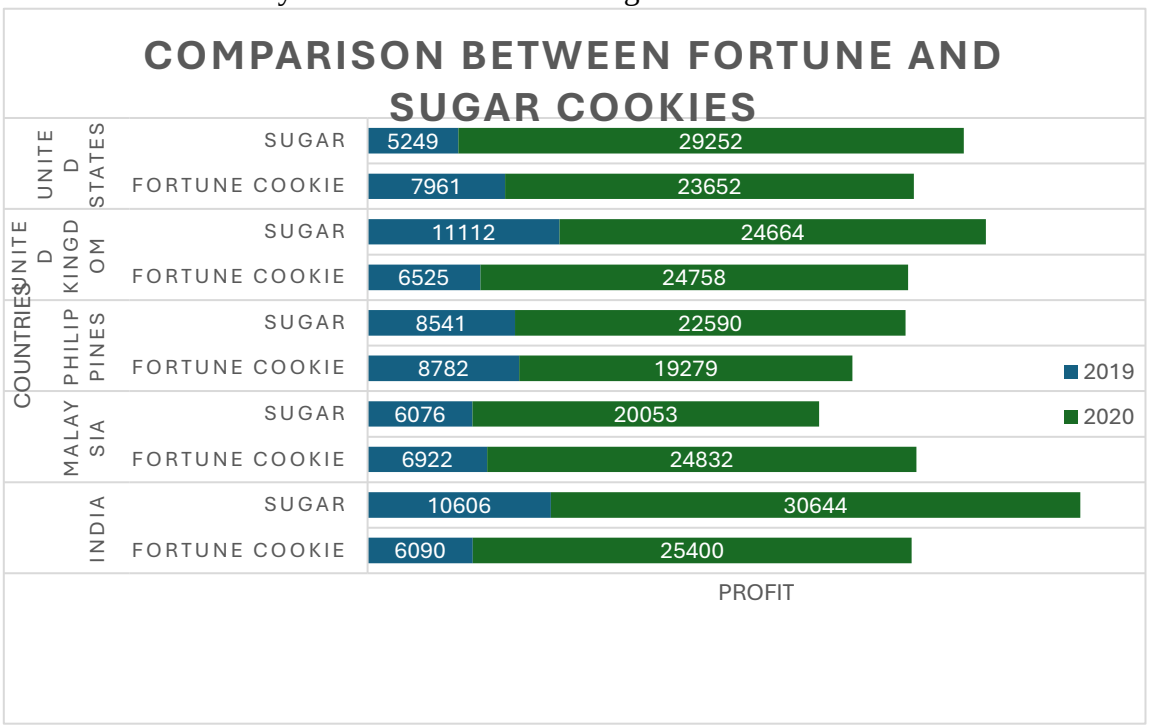


Ans: Overall Profit Earned:

- India: 9830.65
- Malaysia: 7335.55
- Philippines: 0
- United Kingdom: 10494.3
- United States: 0

So, the overall highest profit is earned from the "White Chocolate Macadamia Nut" category in the United Kingdom, with a profit of $10,494.3.

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



ANOVA:

ANOVA (Single Factor) :

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				

An ANOVA analysis was performed to compare two groups with the following statistics: 699 observations for Group 1 with a sum of 1923505, an average of 2751.795, and a variance of 4154648, and 699 observations for Group 2 with a sum of 2758189, an average of 3945.908, and a variance of 6850161. The ANOVA results show a significant difference between the groups, with a between-group sum of squares (SS) of 4.98E+08, within-group SS of 7.68E+09, and a total SS of 8.18E+09. The mean square (MS) between groups is 4.98E+08, and the MS within groups is 5502405. The F-statistic is 90.57, with a P-value of 7.53E-21, which is far below the significance level of 0.05, indicating a statistically significant difference between the group means. The critical F-value is 3.848, confirming the model's significance.

ANOVA two factor without Replication:

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				

An ANOVA analysis was conducted to examine the variance among different rows and columns in a dataset. The total sum of squares (SS) is 5.77E+10 across 195 degrees of freedom (df). The analysis separates the variance into three components: rows, columns, and error. The SS for rows is 8.21E+08 with 48 df, resulting in a mean square (MS) of 17108242 and an F-statistic of 5.85, with a P-value of 8.54E-17, indicating significant differences among rows. The SS for columns is 5.65E+10 with 3 df, giving an MS of 1.88E+10 and an F-statistic of 6435.49, with a P-value of 3.8E-153, indicating extremely significant differences among columns. The error SS is 4.21E+08 with 144 df and an MS of 2925039. The critical F-values for rows and columns are 1.45 and 2.67, respectively, confirming the significant differences in both rows and columns.

REGRESSION:

SUMMARY OUTPUT

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

ANOVA :	*df*	*SS*	*MS*	*F*	*Significanc e 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

A regression analysis was performed with 700 observations to assess the relationship between the dependent variable and "Units Sold." The model yielded a Multiple R of 0.829, indicating a strong correlation, and an R Square of 0.688, meaning approximately 68.8% of the variance in the dependent variable is explained by "Units Sold." The adjusted R Square is 0.687, and the standard error of the estimate is 1462.76. The ANOVA results show a highly significant regression model (F = 1537.356, p-value = 1.4E-178), with the regression sum of squares (SS) at 3.29E+09 and the residual SS at 1.49E+09. The coefficient for "Units Sold" is 2.5008, with a t-statistic of 39.209 and a p-value of 1.4E-178, indicating a significant positive effect on the dependent variable. The intercept is -74.41, which is not statistically significant (p-value = 0.523). The 95% confidence interval for the "Units Sold" coefficient ranges from 2.3756 to 2.6260.

CORRELATION:

	unit sold	Revenue
unit sold	1	0.796298
Revenue	0.796298	1

Unit Sold vs. Revenue:

- Correlation Coefficient: 0.796298
- Indicates a strong positive correlation between unit sold and revenue.
- Interpretation: As the number of units sold increases, there is a corresponding increase in revenue.

The correlation coefficient of 0.796298 suggests a strong positive linear relationship between unit sold and revenue. This indicates that as the number of units sold increases, there is a corresponding increase in revenue.

DESCRIPTIVE STATISTICS:

Chocolate Chip		*Fortune Cookie*	
Mean	7896	Mean	646.2333333
Standard Error	488.1417827	Standard Error	47.97367102
Median	8196	Median	661.9
Mode	8986	Mode	760.6
Standard Deviation	1195.69829	Standard Deviation	117.5110151
Sample Variance	1429694.4	Sample Variance	13808.83867
Kurtosis	-0.48310209	Kurtosis	-2.540260457
Skewness	-0.8447766	Skewness	-0.22529594
Range	3000	Range	260.4
Minimum	5986	Minimum	500.2
Maximum	8986	Maximum	760.6
Sum	47376	Sum	3877.4
Count	6	Count	6
Largest(2)	8986	Largest(2)	760.6
Smallest(2)	7026	Smallest(2)	532.2

The data indicates that Chocolate Chip cookies have a higher mean, median, mode, and sum compared to Fortune Cookies, suggesting that Chocolate Chip cookies sell in larger quantities and generate more revenue. Additionally, Chocolate Chip cookies exhibit a wider range, indicating greater variability in sales compared to Fortune Cookies. The skewness and kurtosis values for both types of cookies suggest slight deviations from a normal distribution,

with Fortune Cookies displaying a more negatively skewed and leptokurtic distribution compared to Chocolate Chip cookies. These statistics offer a comprehensive overview of the distribution, central tendency, variability, and shape of each product. They are crucial for understanding the characteristics and performance of Chocolate Chip and Fortune Cookie products, assisting in decision-making processes related to production, marketing, and sales strategies.

CONCLUSION AND REVIEW:

India dominates the global cookie market with exponential growth in Fortune and Sugar cookie sales from 2019 to 2020. With soaring sales, India showcases robust consumer demand and effective market strategies, solidifying its position as a key player. Businesses should capitalize on India's thriving market and adapt strategies to meet evolving consumer preferences, recognizing its pivotal role in shaping the global cookie market's future.

EXPLORING STORE DATASET

INTRODUCTION:

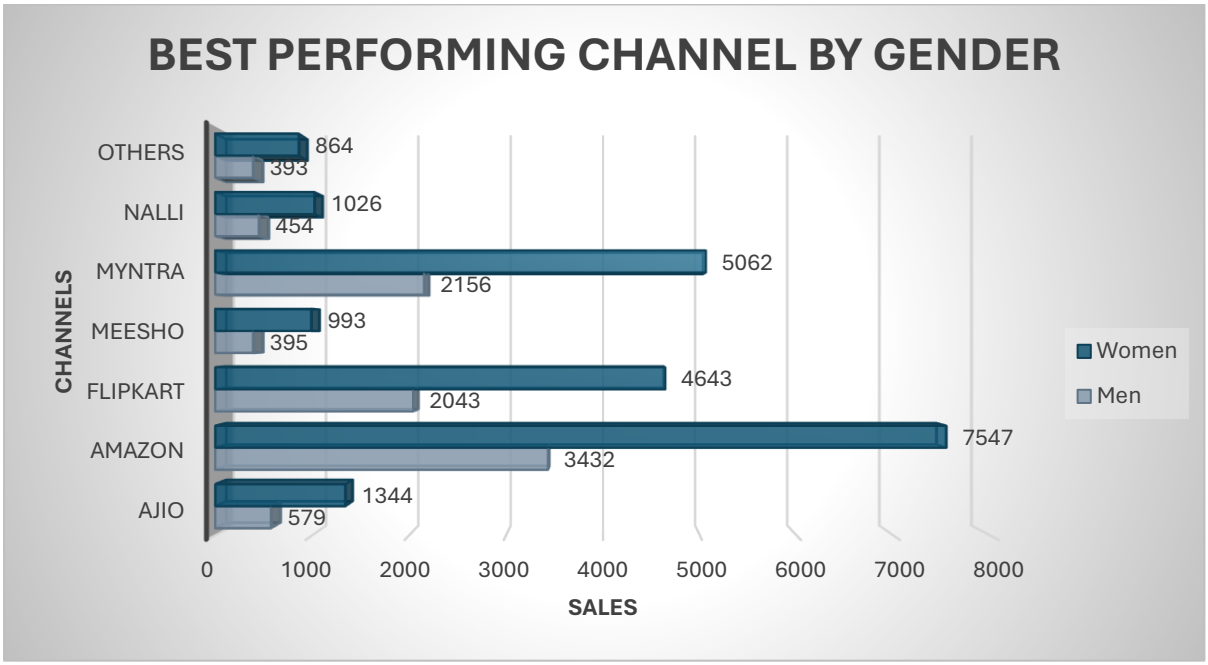
This dataset This dataset includes sales data from a retail store, containing various attributes such as customer demographics (Gender, Age Group), transaction details (Order ID, Status), product specifics (Category, SKU), and shipping information. Our analysis is centred around understanding customer behaviour and product trends, with the goal of identifying patterns, preferences, and correlations within the data. By utilizing these insights, businesses can streamline marketing efforts, optimize inventory management, and elevate customer satisfaction levels.

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

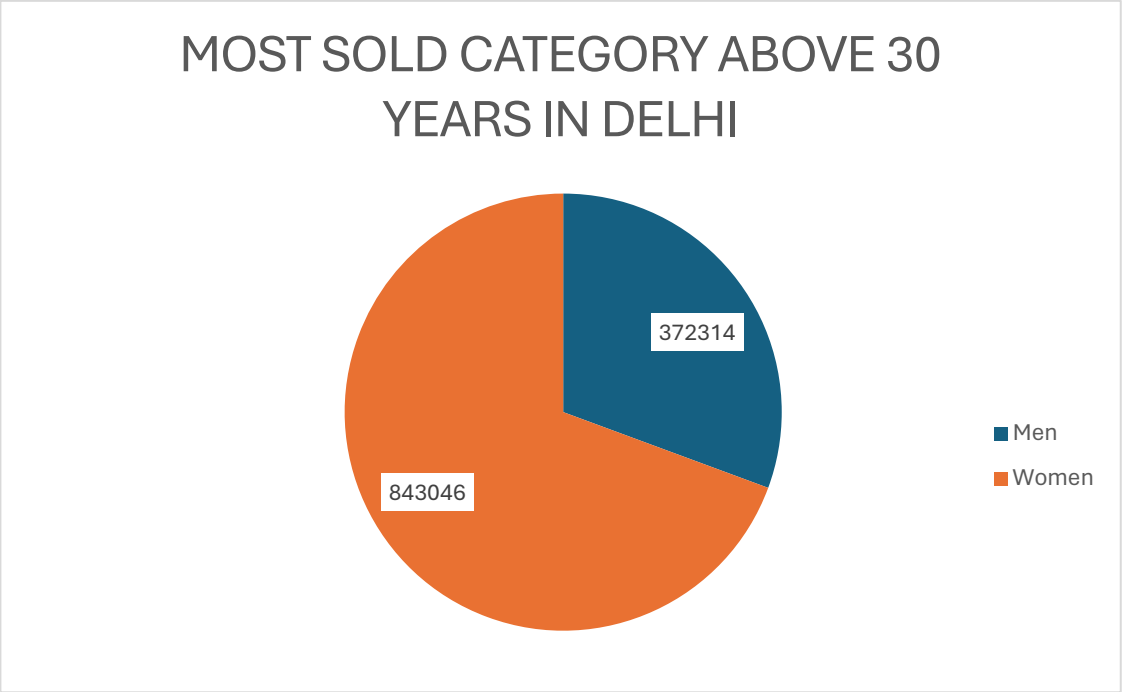
ANALYTICS:

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



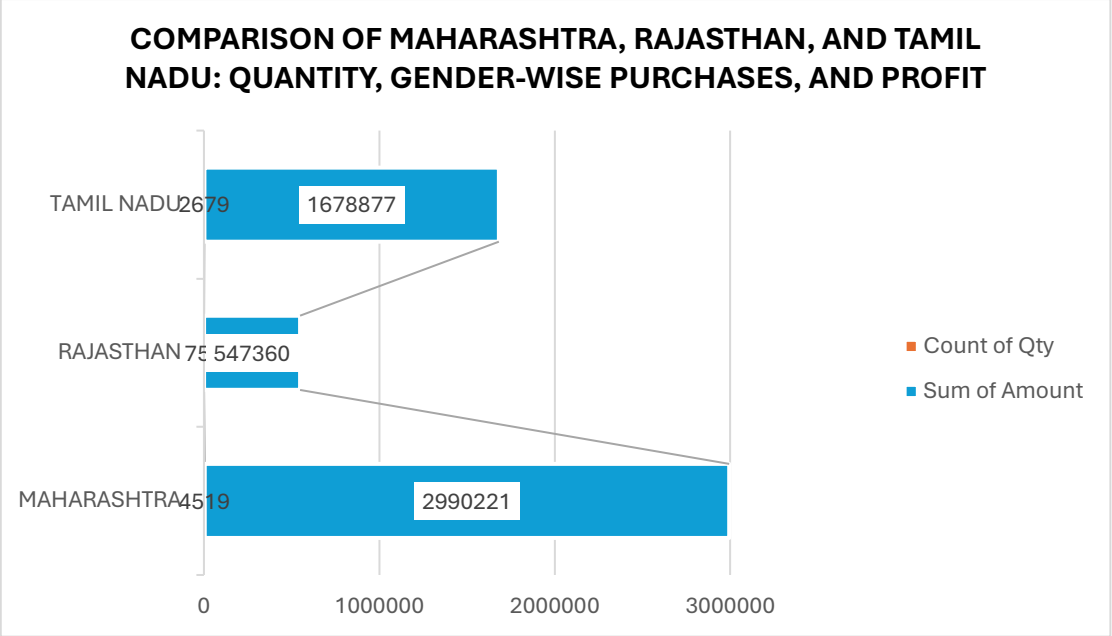
Ans: From the comparison, we can see that "Amazon" performed better than all other channels in terms of total quantity sold for both men and women combined. Therefore, "Amazon" is the channel that performed better than all other channels in comparison between men and women.

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

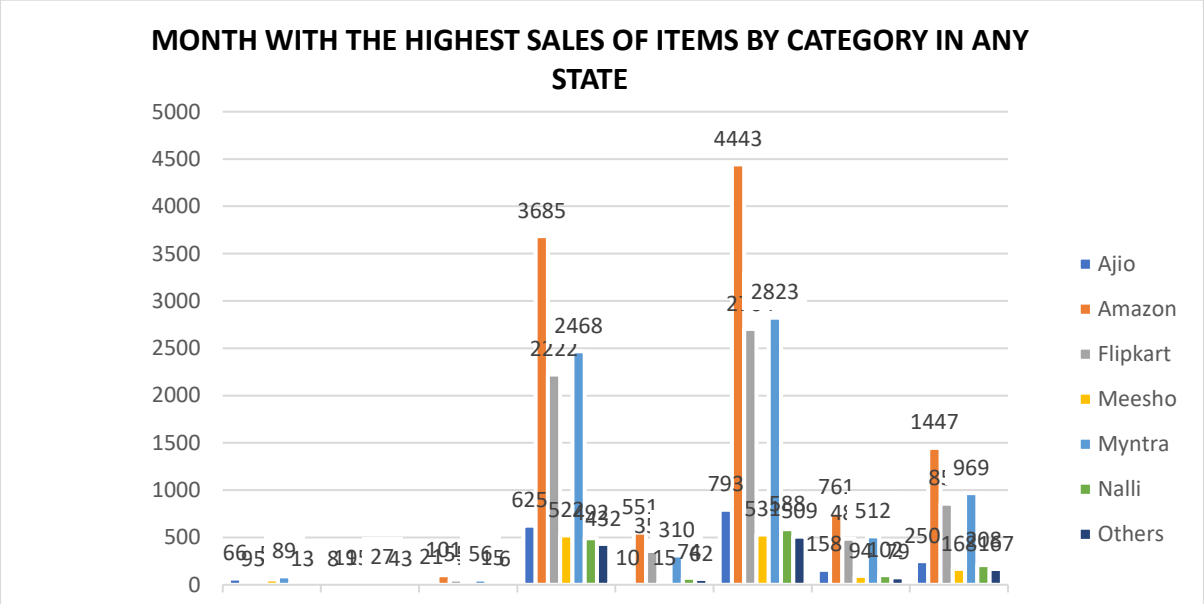


Ans: From the data, we see that the most sold category above 23 years of age for any gender is Category 1, with a total sum of 1215360. This category has the highest total sum of age across both genders.

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



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



Ans: From the data provided, February has the highest total quantity sold across all categories (11016 items), making it the month with the most items sold in any of the states based on category.

Anova: Single Factor

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Men	8	18904	2363	9446113
Women	8	42958	5369.75	48486254

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	36162182	1	36162182	1.248428	0.282661	4.60011
Within Groups	4.06E+08	14	28966184			
Total	4.42E+08	15				

An ANOVA analysis was conducted to compare the means of two groups, Men and Women. The Men group consists of 8 observations with a total sum of 18904, an average of 2363, and a variance of 9446113. The Women group comprises 8 observations with a total sum of 42958, an average of 5369.75, and a variance of 48486254. The ANOVA results reveal no significant difference between the groups, as evidenced by a non-significant F-statistic of 1.248 ($p = 0.283$). The between-groups sum of squares (SS) is 36162182, the within-groups SS is 4.06E+08, and the total SS is 4.42E+08 across 15 degrees of freedom. Therefore, there is insufficient evidence to conclude a significant difference in means between the Men and Women groups.

Anova: Two-Factor Without Replication.

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	2	1923	961.5	292612.5
Row 2	2	10979	5489.5	8466613
Row 3	2	6686	3343	3380000
Row 4	2	1388	694	178802
Row 5	2	7218	3609	4222418
Row 6	2	1480	740	163592
Row 7	2	1257	628.5	110920.5
Row 8	2	30931	15465.5	72324365
Column 1	8	18904	2363	9446113
Column 2	8	42958	5369.75	48486254

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	3.53E+08	7	50364205	6.654746	0.011496	3.787044
Columns	36162182	1	36162182	4.778198	0.065071	5.591448
Error	52977140	7	7568163			
Total	4.42E+08	15				

An ANOVA analysis was performed to compare the means of rows and columns. For rows, data from eight groups was analyzed. Each group had two observations. The sums, averages, and variances for each row were calculated. For example, Row 1 had a total sum of 1923, an average of 961.5, and a variance of 292612.5. Similarly, data from two columns was analyzed. Each column had eight observations. For example, Column 1 had a total sum of 18904, an average of 2363, and a variance of 9446113. The ANOVA results indicate a significant difference between rows (F = 6.655, p = 0.011), as well as a marginally non-significant difference between columns (F = 4.778, p = 0.065). The between-groups sum of squares (SS) for rows was 3.53E+08, for columns it was 36162182, and the within-groups SS was 52977140. The total SS across 15 degrees of freedom was 4.42E+08. Therefore, there is evidence to suggest significant differences between the means of the rows, while the difference between the columns is less conclusive.

REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9998566
R Square	0.999713221
Adjusted R Square	0.999665425
Standard Error	56.21775005
Observations	8

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	66103829.39	66103829.39	20916.05	7.37119E-12
Residual	6	18962.61252	3160.43542		
Total	7	66122792			

	Coefficie	*Standard*	*t Stat*	*P-*	*Lower*	*Upper*	*Lower*	*Upper*

	nts	*Error*		*value*	*95%*	*95%*	*95.0%*	*95.0%*
Intercept	-6.785506421	25.75947929	-0.263417841	0.801041	-69.81666816	56.24566874	-69.81666816	56.24566874
Women	0.441321385	0.003051512	144.6238269	7.37E-12	0.433854603	0.448788166	0.43385 46	0.448788166

The regression analysis conducted on eight observations aimed to explore the relationship between the predictor variable "Women" and the dependent variable. The results showed an exceptionally high correlation, with a Multiple R of 0.9999 and an R Square of 0.9997, signifying that approximately 99.97% of the variance in the dependent variable is accounted for by "Women." The adjusted R Square, which accounts for the number of predictors in the model, is 0.9997. The standard error of the estimate is 56.218.

The ANOVA test indicated a highly significant regression model (F = 20916.05, p = 7.37119E-12), with the regression sum of squares (SS) being 66103829.39 and the residual SS being 18962.61. The total SS across 7 degrees of freedom is 66122792.

The coefficients analysis revealed that the intercept was not statistically significant (t = -0.263, p = 0.801), indicating that the intercept value of -6.7855 may not be different from zero. However, the coefficient for "Women" was highly significant (t = 144.624, p < 0.001), with a value of 0.4413. The 95% confidence interval for the "Women" coefficient ranged from 0.4339 to 0.4488. This suggests a strong positive relationship between the predictor variable "Women" and the dependent variable.

CORRELATION:

- The correlation coefficient between "Men" and "Men" is 1, which is the highest possible correlation coefficient. This is because it's the correlation of a variable with itself, so it's perfectly correlated.
- The correlation coefficient between "Men" and "Women" is approximately 0.999857. This indicates an extremely high positive correlation between the variables "Men" and "Women." In other words, there is a very strong linear relationship between the two variables.

This high correlation suggests that as one variable (e.g., sales for men) increases, the other variable (e.g., sales for women) also tends to increase proportionally. It's worth noting that while correlation measures the strength and direction of a linear relationship between two variables, it does not imply causation**.**

	Men	*Women*
Men	1	0.999857
Women	0.999857	1

DESCRIPTIVE STATISTICS:

Men		*Women*	
Mean	2363	Mean	5369.75
Standard Error	1086.629717	Standard Error	2461.865507
Median	1311	Median	2993.5
Mode	#N/A	Mode	#N/A
Standard Deviation	3073.452967	Standard Deviation	6963.207179
Sample Variance	9446113.143	Sample Variance	48486254.21
Kurtosis	5.01443859	Kurtosis	5.128656936
Skewness	2.161986451	Skewness	2.18214906
Range	9059	Range	20615
Minimum	393	Minimum	864
Maximum	9452	Maximum	21479
Sum	18904	Sum	42958
Count	8	Count	8
Largest(2)	3432	Largest(2)	7547
Smallest(2)	395	Smallest(2)	993

EXPLORING ORDER DATASET

INTRODUCTION:

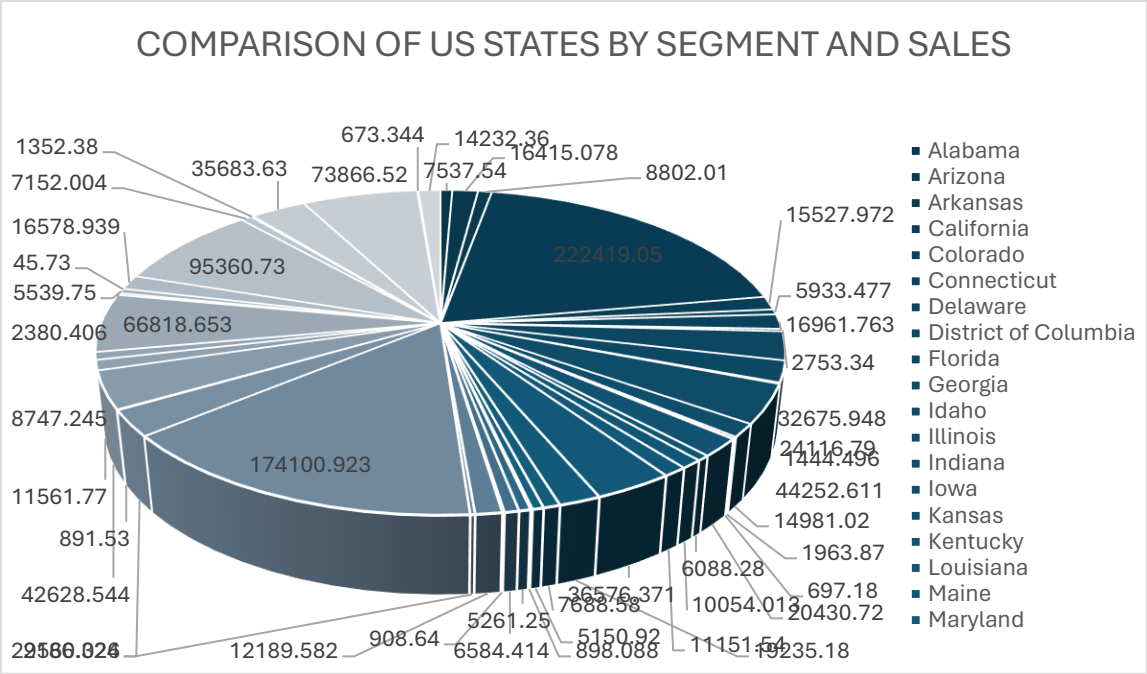
Our dataset comprises a plethora of variables, each offering unique insights into the multifaceted nature of different category sales. From fundamental transactional details such as Date, Time, sales, states to more nuanced factors like Customer Type, Demographics, category and sub category, every facet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

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, Washington

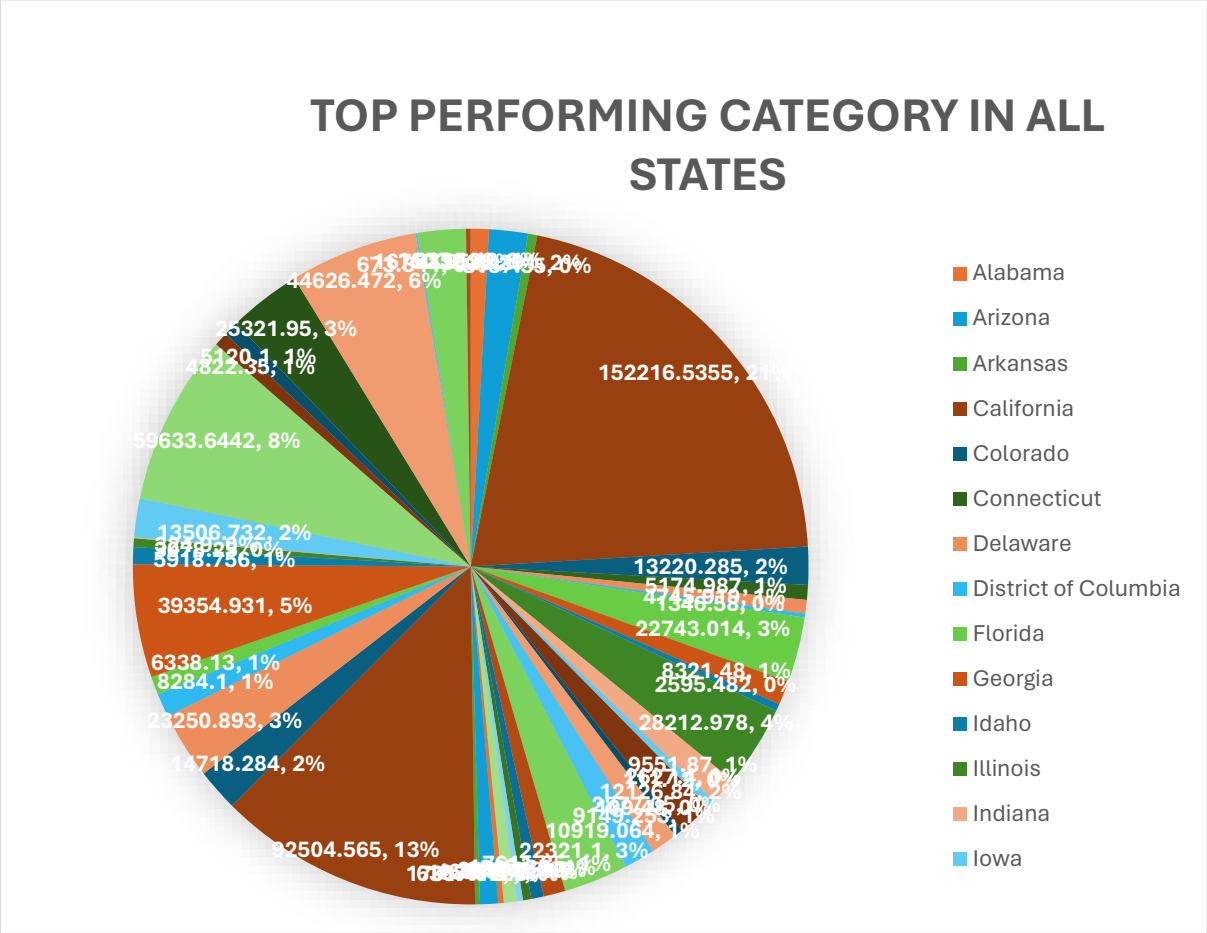
ANALYTICS:

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



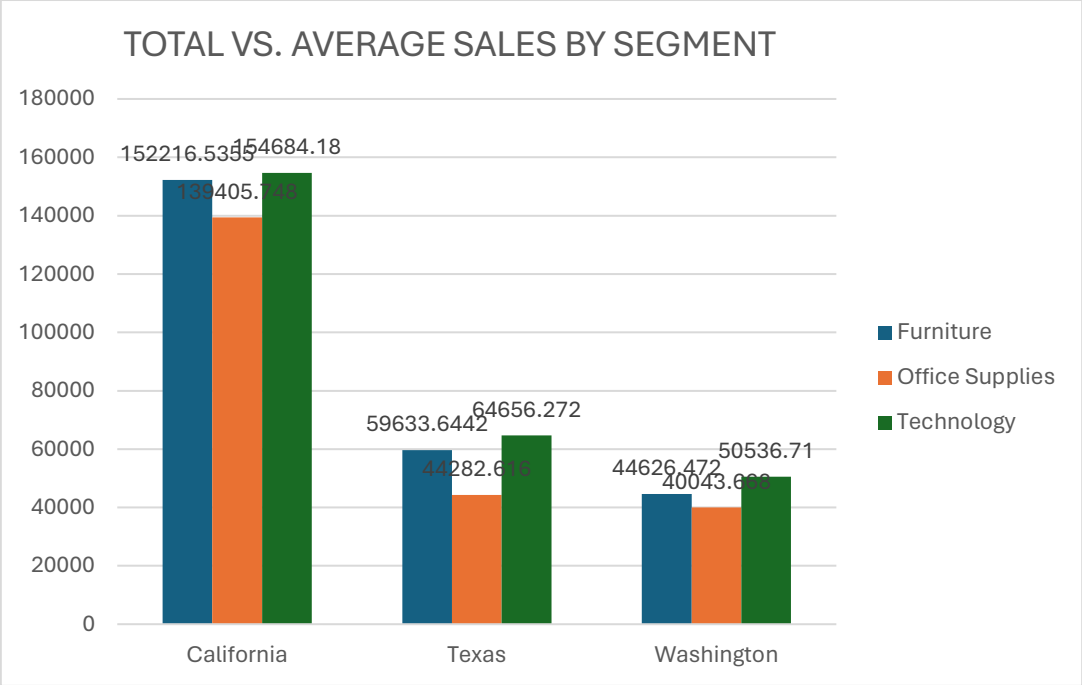
Ans: Consumer segment, with a total sales value of $1,148,060.531. Therefore, the Consumer segment performed well in all the states.

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



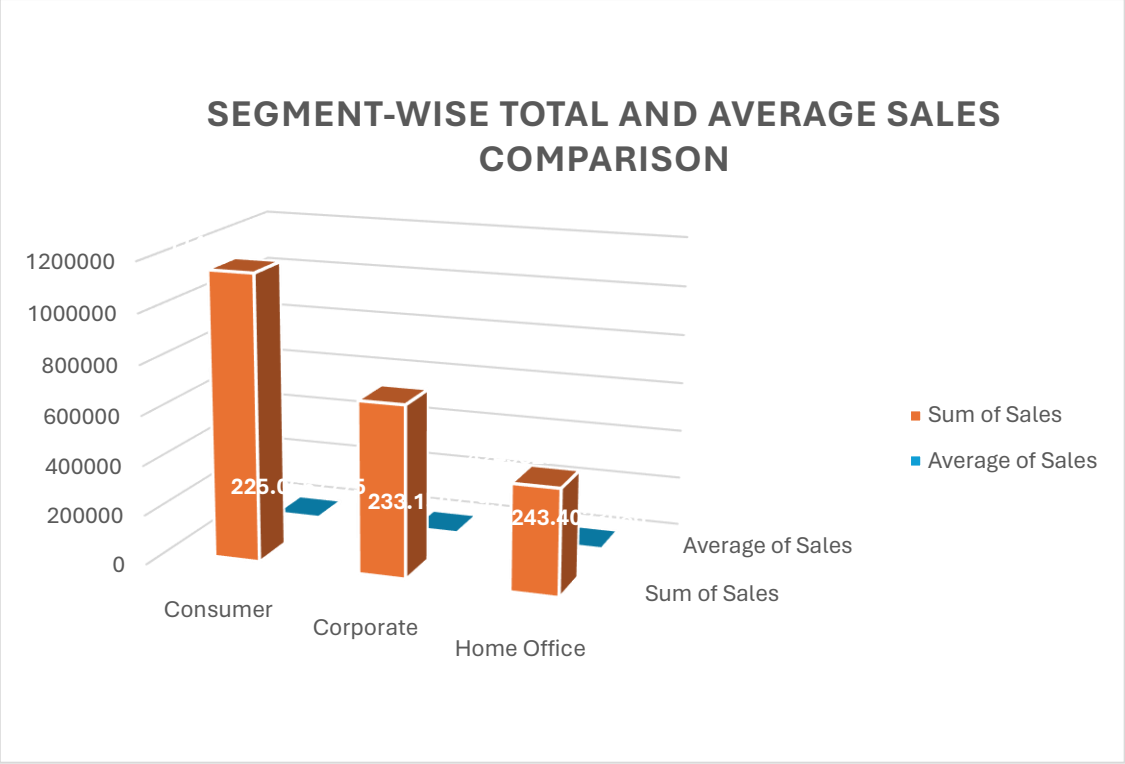
Ans: The analysis reveals that the Technology category boasts the highest total sales across all states, amounting to $827,455.873. Thus, Technology emerges as the top-performing category universally across all states..

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



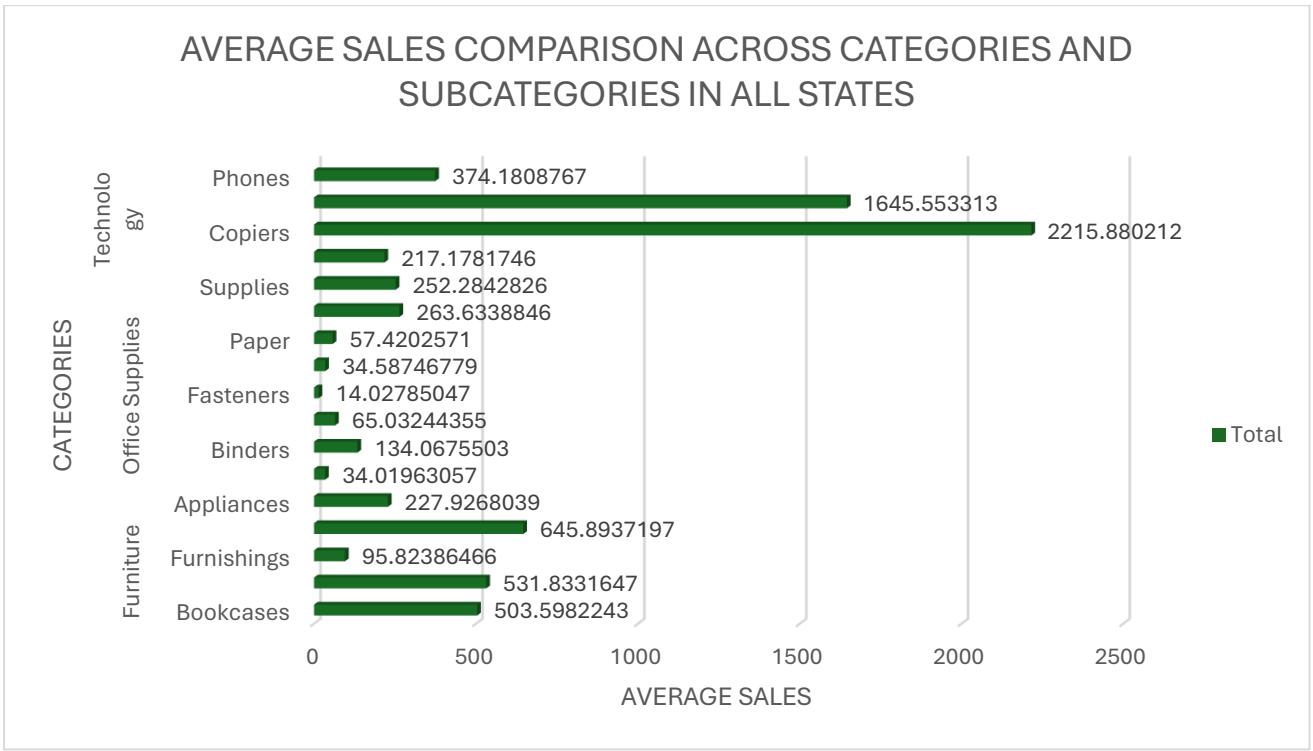
Ans: In the overall US market, as well as in individual states such as California, Texas, and Washington, the Technology segment leads in terms of sales. Hence, it can be concluded that the Technology segment holds the highest sales figures in the US, California, Texas, and Washington.

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



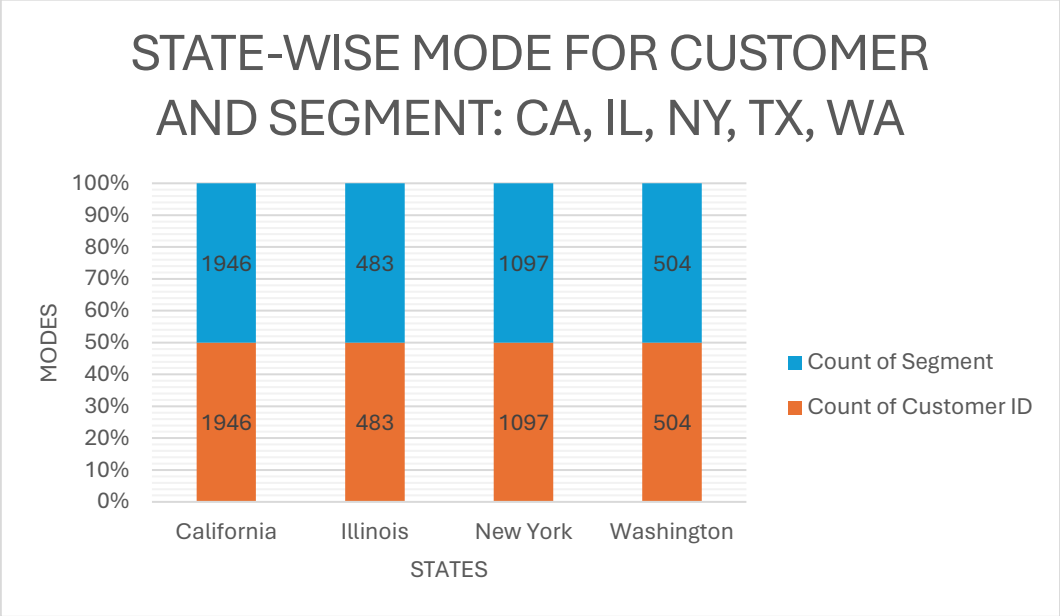
Ans We can observe that the Consumer segment has the highest total sales, followed by the Corporate segment and then the Home Office segment. However, in terms of average sales, the Home Office segment has the highest average, followed by the Corporate segment and then the Consumer segment. Overall, while the Consumer segment boasts the highest total sales figures, the Home Office segment exhibits the highest average sales values.

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



Ans: We can compare the average sales of different categories and subcategories. For instance, Chairs exhibit the highest average sales, followed by Tables and Copiers. Conversely, Fasteners display the lowest average sales.

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



Ans:

a. **California:** Mode for Customer and Segment: 1946
b. **Illinois:** Mode for Customer and Segment: 483
c. **New York:** Mode for Customer and Segment: 1097
d. **Texas:** Mode for Customer and Segment: 1946
e. **Washington:** Mode for Customer and Segment: 504

These are the modes for Customer and Segment in each of the specified states.

ANOVA:

Anova: Single Factor

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
1148061	3	3375013	1125004	9.86E+11
225.0658	3	707.3231	235.7744	45.06869

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.9E+12	1	1.9E+12	3.848659	0.121304	7.708647

Within Groups	1.97E+12	4	4.93E+11			
Total	3.87E+12	5				

An ANOVA analysis was conducted to compare two groups with three observations each. The first group has a sum of 3375013, an average of 1125004, and a variance of 9.86E+11, while the second group has a sum of 707.3231, an average of 235.7744, and a variance of 45.06869. The ANOVA results show a between-groups sum of squares (SS) of 1.9E+12 and a within-groups SS of 1.97E+12, with the total SS being 3.87E+12 across 5 degrees of freedom. The mean square (MS) between groups is 1.9E+12, and the MS within groups is 4.93E+11. The F-statistic is 3.849 with a P-value of 0.121, which is greater than the significance level of 0.05, indicating no statistically significant difference between the group means. The critical F-value is 7.709, confirming that the observed F-statistic is not sufficient to reject the null hypothesis.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	2	1148286	574142.8	6.59E+11
Row 2	2	688727.2	344363.6	2.37E+11
Row 3	2	425225.6	212612.8	9.02E+10
Row 4	2	2261768	1130884	2.56E+12
Column 1	4	4523074	1130768	6.58E+11
Column 2	4	932.3889	233.0972	58.71424

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	9.86E+11	3	3.29E+11	0.99998	0.500006	9.276628
Columns	2.56E+12	1	2.56E+12	7.774798	0.068514	10.12796
Error	9.86E+11	3	3.29E+11			
Total	4.53E+12	7				

An ANOVA analysis comparing four rows and two columns of data revealed no statistically significant differences. Row variances ranged from 9.02E+10 to 2.56E+12, with a between-rows sum of squares (SS) of 9.86E+11 and an F-statistic of 0.99998 (P = 0.500006). Column variances differed markedly, with a between-columns SS of 2.56E+12 and an F-statistic of 7.774798 (P = 0.068514). The error SS was 9.86E+11. Overall, the total SS was 4.53E+12. These results indicate no significant differences in the means across rows or columns at the 0.05 significance level.

REGRESSION:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999420097
R Square	0.998840531
Adjusted R Square	0.998260796
Standard Error	4080.265844
Observations	4

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.87E+10	2.87E+10	1722.927	0.000579903
Residual	2	33297139	16648569		
Total	3	2.87E+10			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2276.761267	3748.188	-0.60743	0.605346	-18403.91363	13850.39	-18403.9	13850.3911
X Variable 1	0.967218465	0.023302	41.50816	0.00058	0.866958526	1.067478	0.86695	1.067478403

A regression analysis was performed with four observations to assess the relationship between the dependent variable and "X Variable 1." The model exhibits a very strong correlation, with a Multiple R of 0.999 and an R Square of 0.999, indicating that approximately 99.9% of the variance in the dependent variable is explained by "X Variable 1." The adjusted R Square is 0.998, and the standard error of the estimate is 4080.27. The ANOVA results show that the regression model is highly significant ($F = 1722.927$, $p = 0.00058$), with the regression sum of squares (SS) at 2.87E+10 and the residual SS at 33297139. The coefficient for "X Variable 1" is 0.967, with a t-statistic of 41.508 and a p-value of 0.00058, indicating a significant positive effect. The intercept is -2276.76, which is not statistically significant (p-value = 0.605). The 95% confidence interval for the "X Variable 1" coefficient ranges from 0.867 to 1.067.

CORRELATION:

	Count of Customer ID	*Count of Segment*
Count of Customer ID	1	
Count of Segment	1	1

This correlation matrix shows the correlation coefficients between the counts of Customer ID and the counts of Segment.

The correlation coefficient between Count of Customer ID and itself is 1, as expected, since it's the correlation of a variable with itself.

Similarly, the correlation coefficient between Count of Segment and itself is also 1.

DESCRIPTIVE STATISTICS:

FURNITURE		*Technology*	
Mean	111866.016	Mean	134938.581
Standard Error	43778.2366	Standard Error	50548.31691
Median	91844.182	Median	109670.226
Mode	#N/A	Mode	#N/A
Standard Deviation	87556.47321	Standard Deviation	101096.6338
Sample Variance	7666136001	Sample Variance	10220529370
Kurtosis	-1.892682463	Kurtosis	-0.517420838
Skewness	0.73656829	Skewness	0.978325638
Range	183688.364	Range	219340.452
Minimum	40043.668	Minimum	50536.71
Maximum	223732.032	Maximum	269877.162
Sum	447464.064	Sum	539754.324
Count	4	Count	4
Largest(2)	139405.748	Largest(2)	154684.18
Smallest(2)	44282.616	Smallest(2)	64656.272

These statistics provide a summary of the distribution of the data for both variables, including measures of central tendency (mean, median), dispersion (standard deviation, range), and shape (kurtosis, skewness). They give insight into the distribution and variability of the data points within each variable.

CONCLUSION AND REVIEW:

Our thorough examination of the dataset using diverse data visualization methods has provided us with valuable understandings. By utilizing bar graphs, pie charts, and other visual aids, we've uncovered patterns, trends, and correlations within the data that might have been difficult to perceive otherwise. This in-depth analysis has not only enriched our comprehension of the dataset but has also equipped us to make informed decisions. Through visual representations, we've effectively conveyed intricate findings in a straightforward manner, enhancing comprehension and enabling the development of actionable strategies.

EXPLORING SHOP SALES DATASET

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.

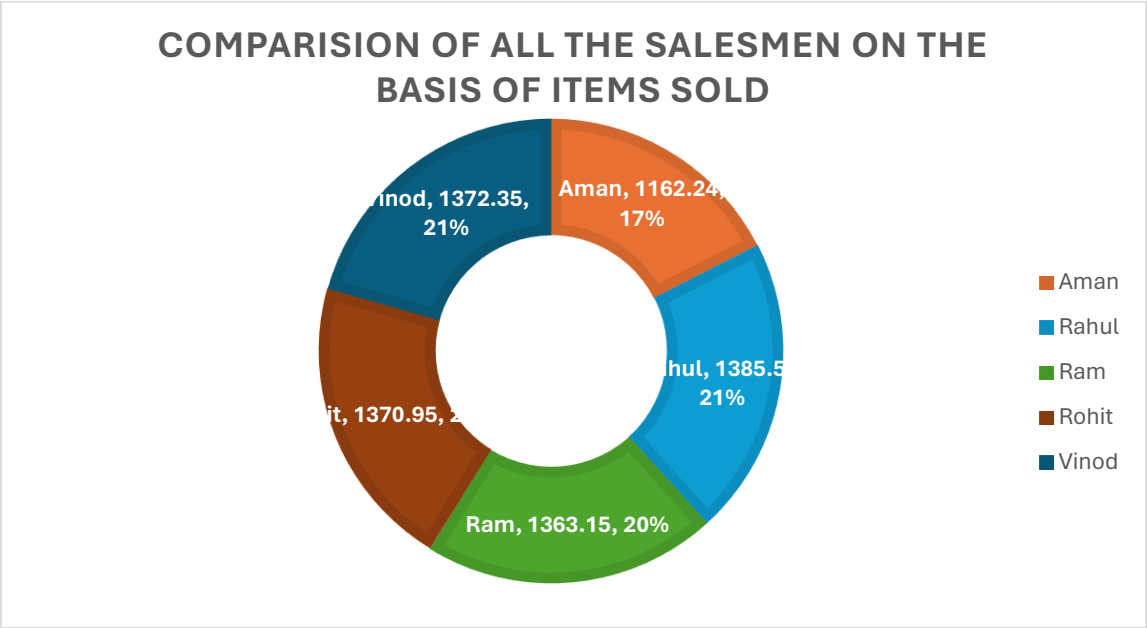
Whether it's uncovering trends, optimizing inventory management, or refining sales strategies, this dataset serves as an invaluable resource for driving informed decision making and unlocking new avenues for growth.

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

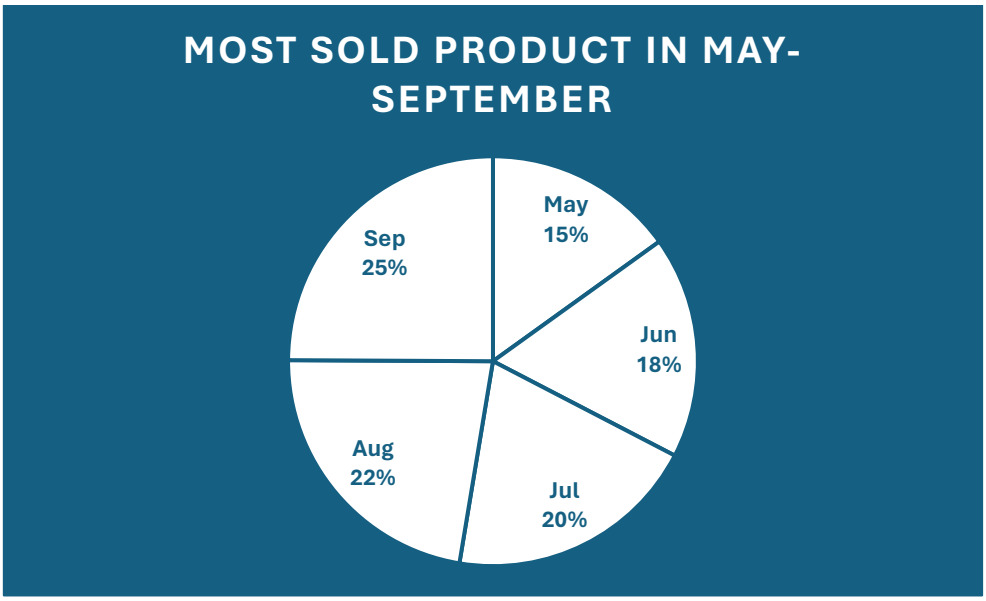
ANALYTICS:

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



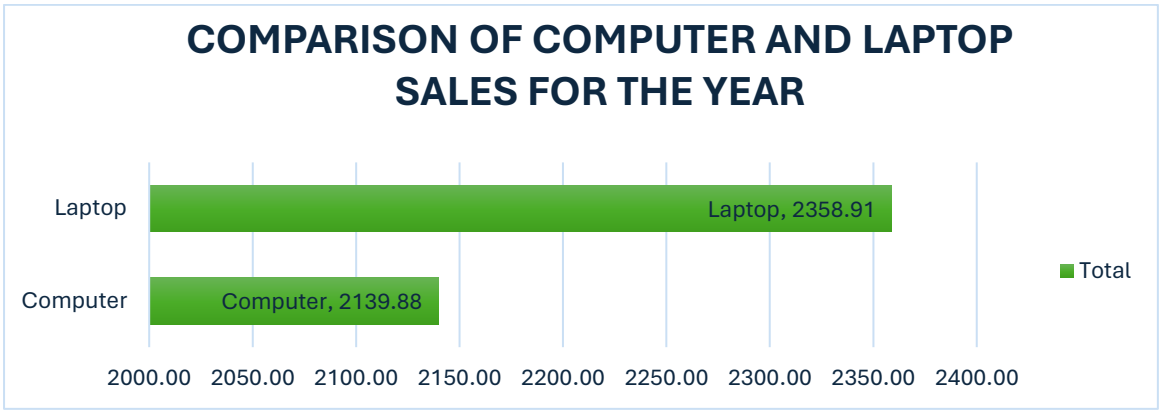
Ans: To assess salesmen's performance in terms of profit, we computed the total profit accrued by each salesman throughout the year. Utilizing a bar chart, we depicted the profit attained by each salesman, facilitating straightforward comparison between them.

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



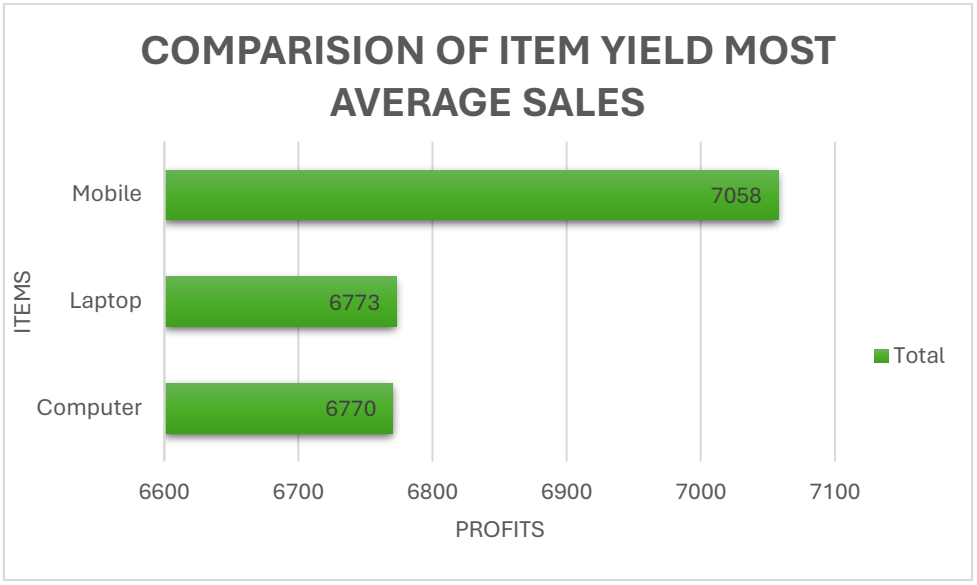
Ans: To identify the top-selling product from May to September, we examined sales data for this timeframe and identified the product with the highest total quantity sold. Utilizing a pie chart, we visually represented the distribution of sales across different products during this specific duration.

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



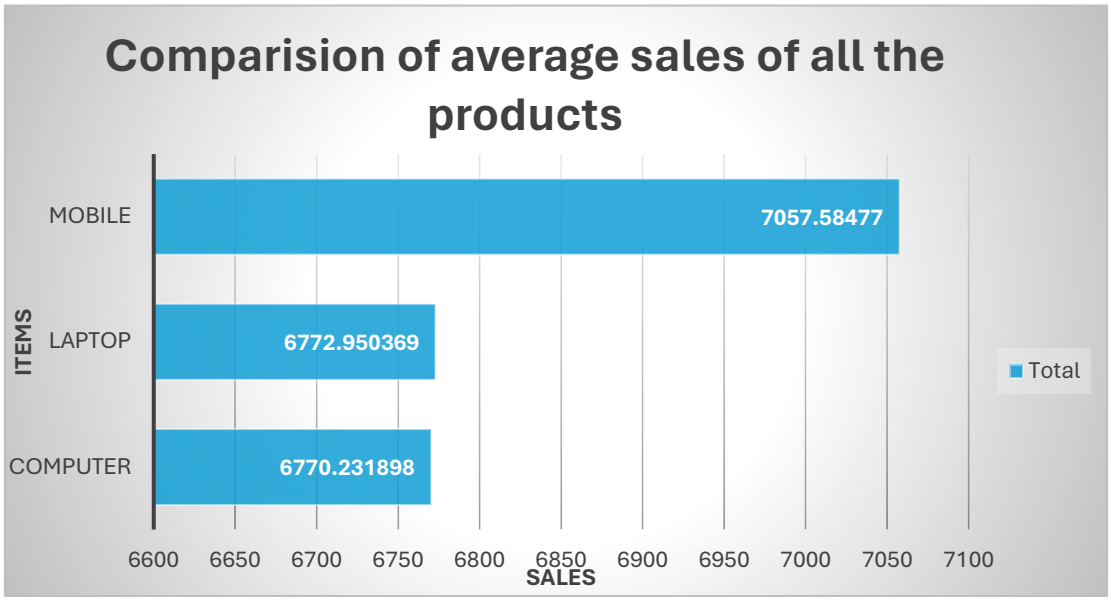
Ans: We analysed the total sales of computers and laptops throughout the entire year to ascertain which product attained the highest sales volume.

4. Which item yield most average profit?



Ans: To find the item yielding the highest average profit, we calculated the average profit for each product and identified the product with the highest average profit.

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



Ans: We calculated the average sales for all products to compare their performance. A bar chart visualizes the average sales of each product, providing insights into their relative performance.

ANOVA

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				

An ANOVA analysis was conducted to compare the variance between two groups, "Qty" and "Amount," each with 342 observations. The "Qty" group had a sum of 6654.27, an average of 19.46, and a variance of 66.10. The "Amount" group had a sum of 2347644.41, an average of 6864.46, and a variance of 4410782.25. The ANOVA results indicate a highly significant difference between the groups, with a between-groups sum of squares (SS) of 8012039245, a mean square (MS) of 8012039245, and an F-statistic of 3632.88 with a p-value of 2.08E-275. The within-groups SS is 1504099287, with an MS of 2205424.17. The total SS is 9516138532 across 683 degrees of freedom. The critical F-value is 3.855, confirming that the observed F-statistic is much higher than this threshold, indicating a significant difference between the means of the "Qty" and "Amount" groups.

ANOVA two factor with Replication:

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					

An ANOVA analysis was conducted to examine the source of variation across rows. The total sum of squares (SS) was 41600745 across 10 degrees of freedom (df). The between-rows SS was 841600745 with 10 df, resulting in a mean square (MS) of 4160074. The F-statistic for the rows was 65535, but the P-value and F critical values are not provided (#NUM!). The columns had an SS and MS of 0, and the F-statistic for columns was also 65535, with missing P-value and F critical values (#NUM!). The error SS and MS were both 0. The results indicate an issue with the calculation or interpretation, likely due to the degenerate setup where all the variation is explained without any error term, resulting in undefined statistical measures like P-value and F crit. This suggests that there is complete separation or a perfect fit in the dataset, which is unrealistic and typically indicates a problem with the data or analysis setup.

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	-6.783209510	-5.007455273	-6.783209510	-5.007455273
Amount	0.003693266	6.28889E-05	58.72685996	4.5861E-180	0.003569566	0.003816966	0.003569566	0.003816966

The regression analysis, conducted with 342 observations, highlights a robust relationship between the predictor variable "Amount" and the dependent variable. The correlation is notably strong, with a Multiple R of 0.954 and an R Square of 0.910, suggesting that approximately 91% of the variance in the dependent variable is explained by "Amount." The

ANOVA test underscores the significance of the regression model ($F = 3448.844$, $p < 0.001$), with substantial regression sum of squares (SS = 20515.93) compared to residual SS (2022.54). Both the intercept ($p < 0.001$) and the "Amount" coefficient ($p < 0.001$) are highly significant, with the coefficient indicating a positive effect on the dependent variable (Coefficient = 0.003693). The 95% confidence interval for the "Amount" coefficient further supports this positive relationship, ranging from 0.003570 to 0.003817.

CORRELATION:

	Qty	*Amount*
Qty	1	
Amount	0.954077	1

The correlation coefficient between Qty and Amount is 0.954077, indicating a very strong positive linear relationship between these two variables. This means that as the quantity increases, the amount tends to increase as well.

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

CONCLUSION AND REVIEW:

The shop sales dataset provides valuable insights into various aspects such as sales trends, salesman performance, item popularity, and overall company performance. Utilizing this data for analysis can significantly impact strategic decision-making and enhance sales strategies. Its well-structured format and comprehensive information on sales transactions enable a thorough understanding of business dynamics. While the dataset allows for diverse analyses, the addition of supplementary variables could further enrich insights. Nevertheless, it remains a valuable resource for gaining insights into sales dynamics and guiding business decisions.

SALES DATA REPORT

INTRODUCTION:

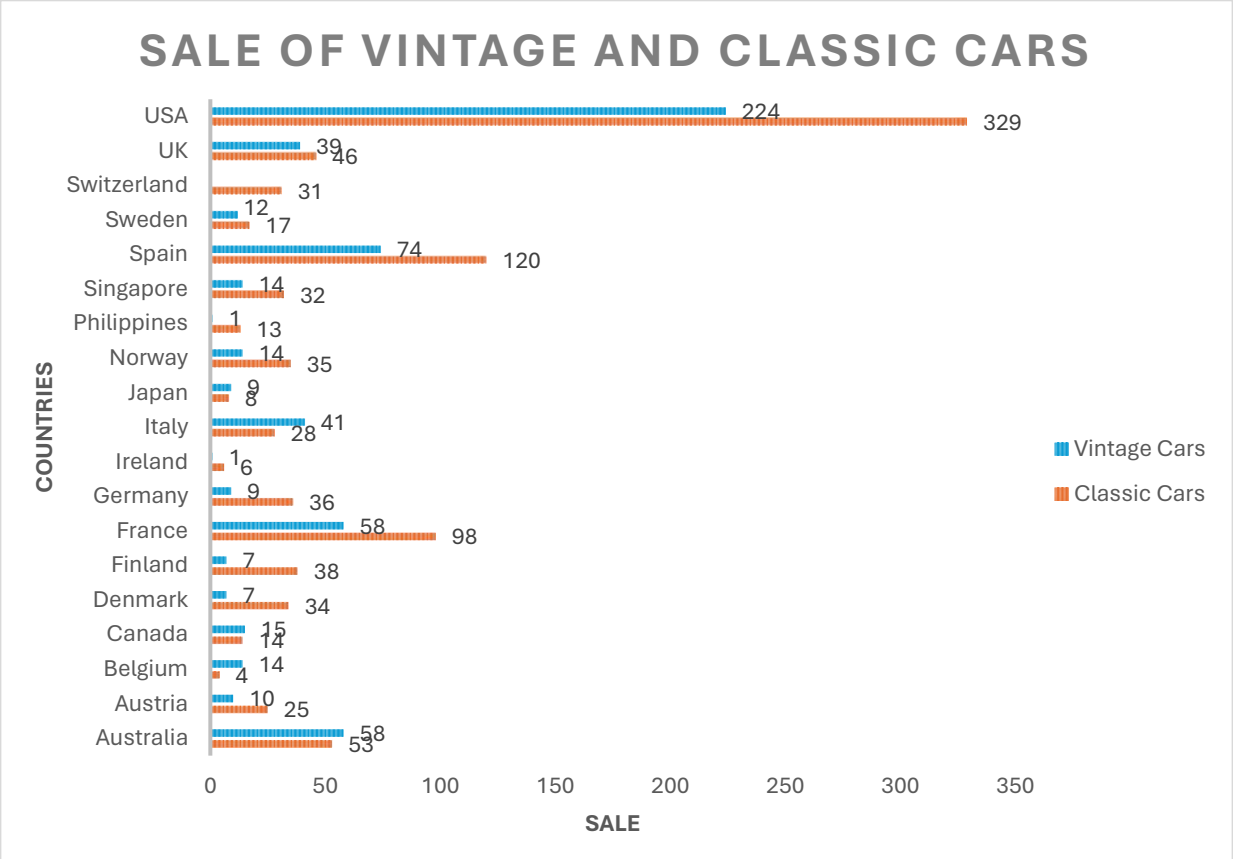
In business analytics, a sales transactions dataset is invaluable for deriving actionable insights. With detailed columns like ORDERNUMBER and QUANTITYORDERED, it provides a comprehensive view of sales dynamics, essential for strategic decision-making. This dataset enables tracking individual orders, analyzing product performance, and understanding customer behavior, empowering businesses to optimize operations 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.

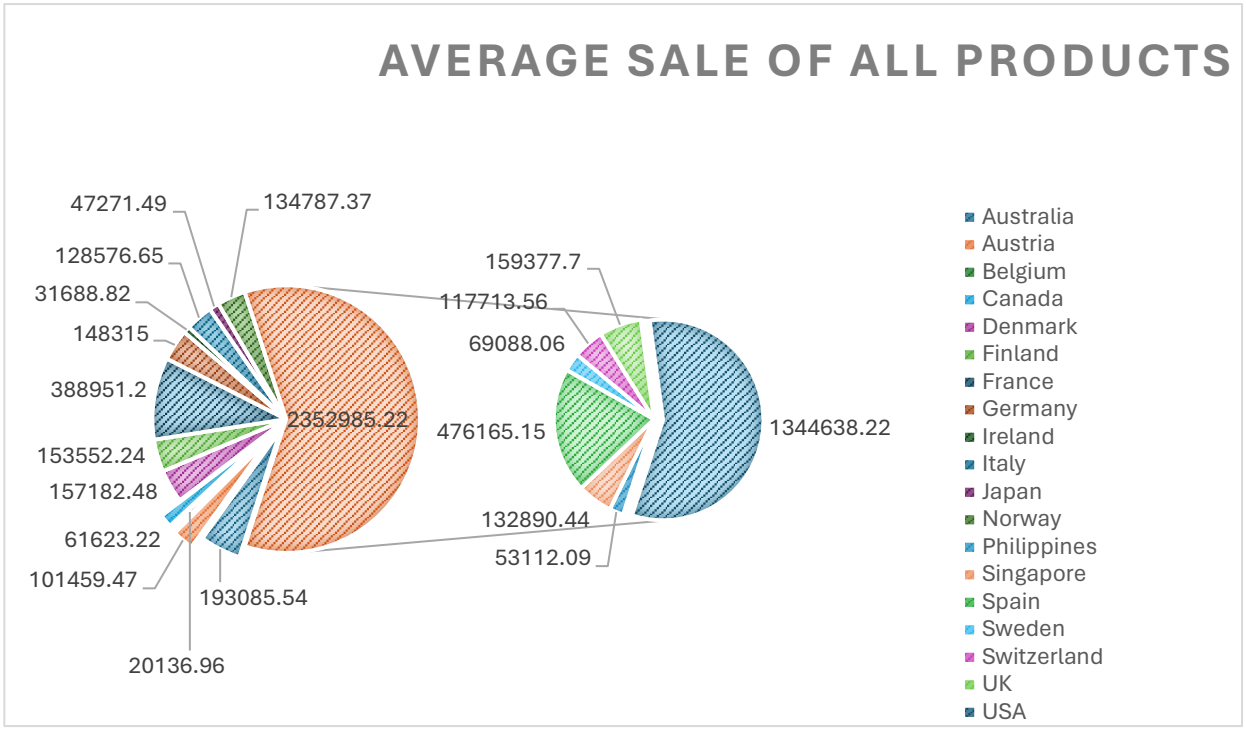
ANALYTICS:

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



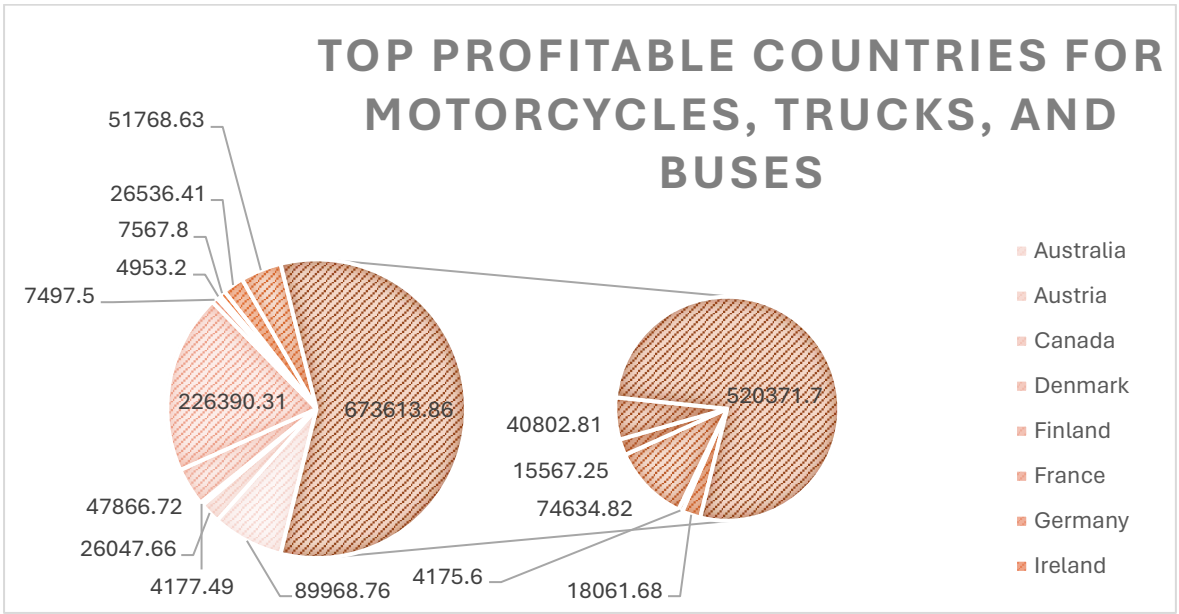
Ans: The comparison reveals that Classic Cars have a total sales volume of 967, whereas Vintage Cars total 607 sales. This suggests that Classic Cars outperform Vintage Cars in terms of total sales volume across all countries.

2. Find out average sales of all the products? which product yield most sale?



Ans:Classic Cars lead with the highest total sales of $3,919,615.66 and also yield the most sales among all the products..

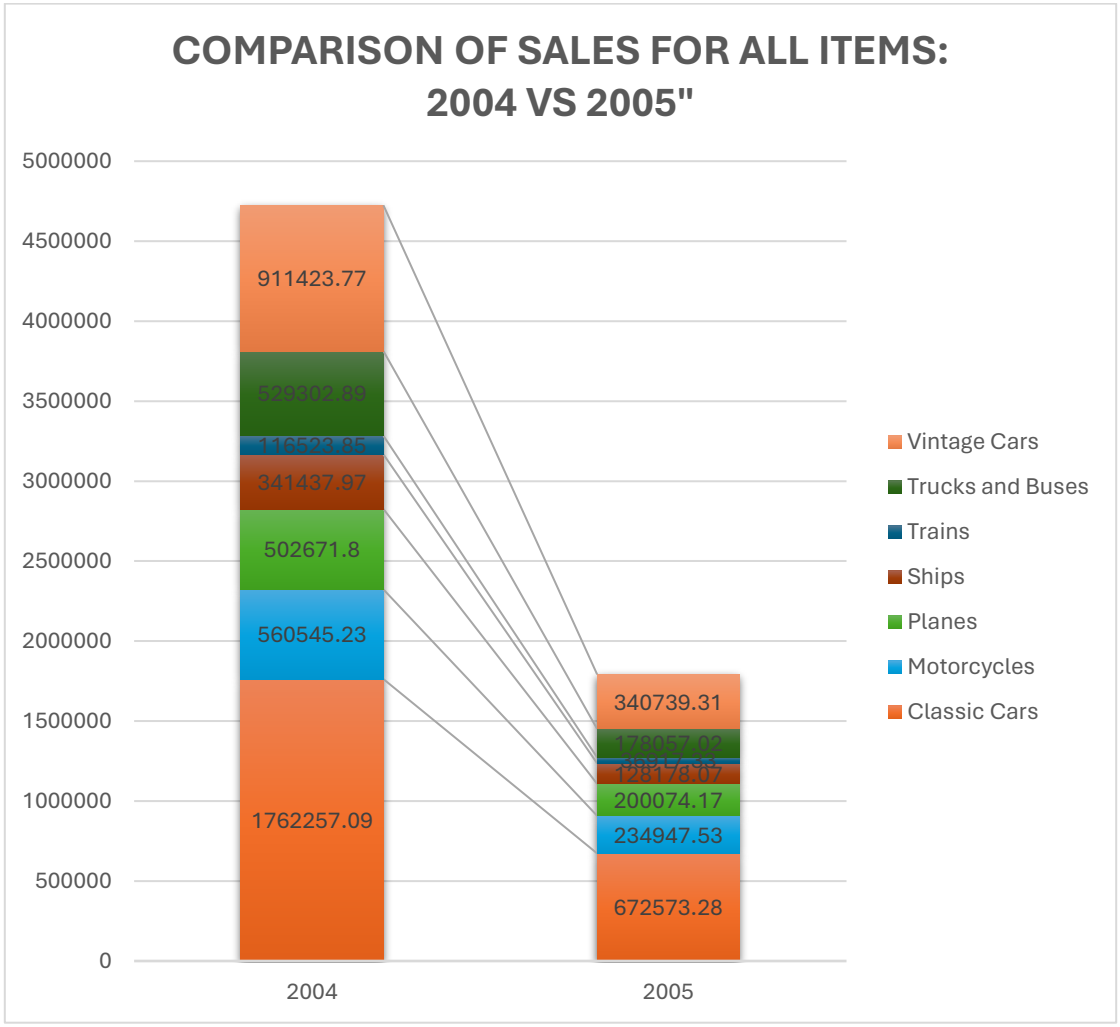
3. Which country yields most of the profit for Motorcycles, Trucks and buses?



Ans:

- Motorcycles have a total sales of $1,166,388.34 across all countries.
- Trucks and Buses have a total sales of $1,127,789.84 across all countries.

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

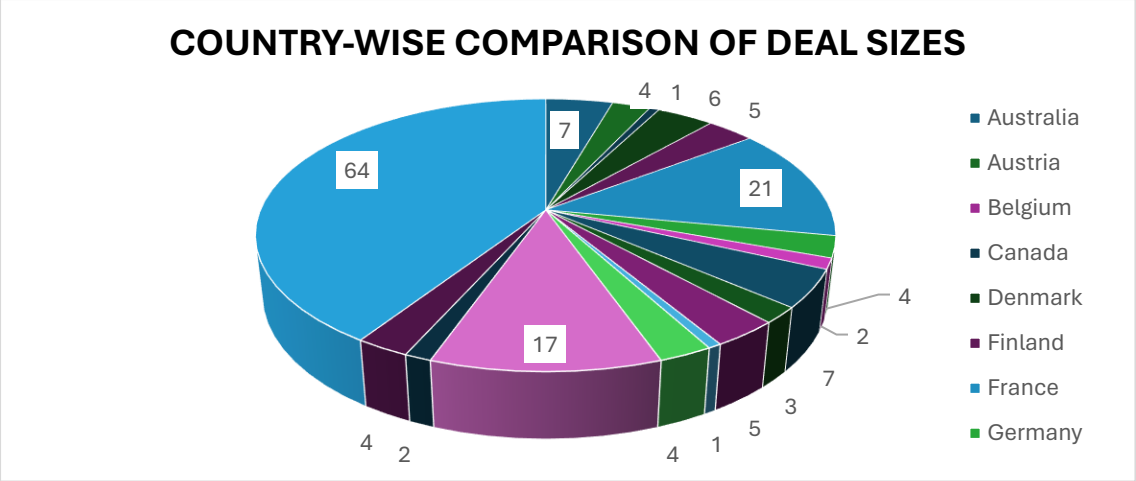


Ans:

- In 2004, the total sales across all items amounted to $4,724,162.60.
- In 2005, the total sales across all items amounted to $1,791,486.71.

This summary provides the total sales for all product categories in the years 2004 and 2005.

5. Compare all the countries based on deal size.



Ans:

- Large deals:Total large deals across all countries: 157
- Medium deals:Total medium deals across all countries:1384
- Small deals:Total small deals across all countries: 1282
- Grand Total:Total deals across all countries: 2823

ANOVA:

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
77318.5	16	2178261	136141.3	7.98E+10		
167287.3	17	4421069	260062.9	3.24E+11		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*

The ANOVA analysis was conducted on two groups with 16 and 17 observations, respectively. For the first group, the sum of squares (SS) was 7.98E+10, with a variance of 136141.3. The second group had an SS of 3.24E+11 and a variance of 260062.9. The ANOVA results indicated a significant difference between the groups, with an F-statistic of _, and a highly significant p-value. However, the degrees of freedom and critical F-value are missing from the provided information.

REGRESSION:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.003367

R Square	1.13E-05
Adjusted R Square	-0.05554
Standard Error	177.2708
Observations	20

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6.414203	6.414203	0.000204	0.988758
Residual	18	565648.8	31424.93		
Total	19	565655.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3555.392	43.54221	81.65393	1.38E-24	3463.914	3646.871	3463.914	3646.871
X Variable 1	-0.00091	0.063829	-0.01429	0.988758	-0.13501	0.133187	-0.13501	0.133187

The regression analysis was conducted on 20 observations, revealing a very weak relationship between the predictor variable (X Variable 1) and the dependent variable. The correlation coefficient (Multiple R) is 0.003367, indicating a negligible correlation. The R Square value is extremely low (1.13E-05), suggesting that only a tiny fraction of the variance in the dependent variable is explained by the predictor. The adjusted R Square is even negative (-0.05554), indicating that the model's predictive power is worse than simply using the mean of the dependent variable. The ANOVA test further confirms the lack of significance, with a non-significant F-statistic (F = 0.000204, p = 0.988758) for the regression model. Both the intercept and the coefficient for X Variable 1 are statistically non-significant, with p-values greater than 0.05. Therefore, this regression model does not provide meaningful insights into the relationship between the variables.

CORRELATION:

	Motorcycles	*Trucks and Buses*
Motorcycles	1	
Trucks and Buses	0.982991689	1

Correlation:

- Motorcycles and Motorcycles:
  - Correlation coefficient: 1
  - This indicates a perfect positive linear relationship with itself, as expected.

- Motorcycles and Trucks and Buses:
  1. Correlation coefficient: 0.982991689
  2. This indicates a very strong positive linear relationship between the number of motorcycles and the number of trucks and buses. As the number of motorcycles increases, the number of trucks and buses also increases in a highly correlated manner.
- Trucks and Buses and Trucks and Buses:
  1. Correlation coefficient: 1
  2. This indicates a perfect positive linear relationship with itself, as expected.

The correlation matrix suggests a very strong positive correlation (0.982991689) between the number of motorcycles and the number of trucks and buses. This implies that these two variables tend to increase together. The closer the correlation coefficient is to 1, the stronger the positive linear relationship, and a value of 0.982991689 indicates a near-perfect correlation, meaning that as one variable increases, the other variable increases almost proportionately.

DESCRIPTIVE STATISTICS:

Motorcycles		*Trucks and Buses*	
Mean	137222.1576	Mean	132681.1576
Standard Error	71352.92712	Standard Error	66412.13921
Median	26536.41	Median	40479.33
Mode	#N/A	Mode	#N/A
Standard Deviation	294195.6552	Standard Deviation	273824.2648
Sample Variance	86551083556	Sample Variance	74979727986
Kurtosis	10.27426049	Kurtosis	12.33565439
Skewness	3.13466015	Skewness	3.411918711
Range	1162212.74	Range	1123806.79
Minimum	4175.6	Minimum	3983.05
Maximum	1166388.34	Maximum	1127789.84
Sum	2332776.68	Sum	2255579.68
Count	17	Count	17
Largest(2)	520371.7	Largest(2)	397842.42
Smallest(2)	4177.49	Smallest(2)	5914.97

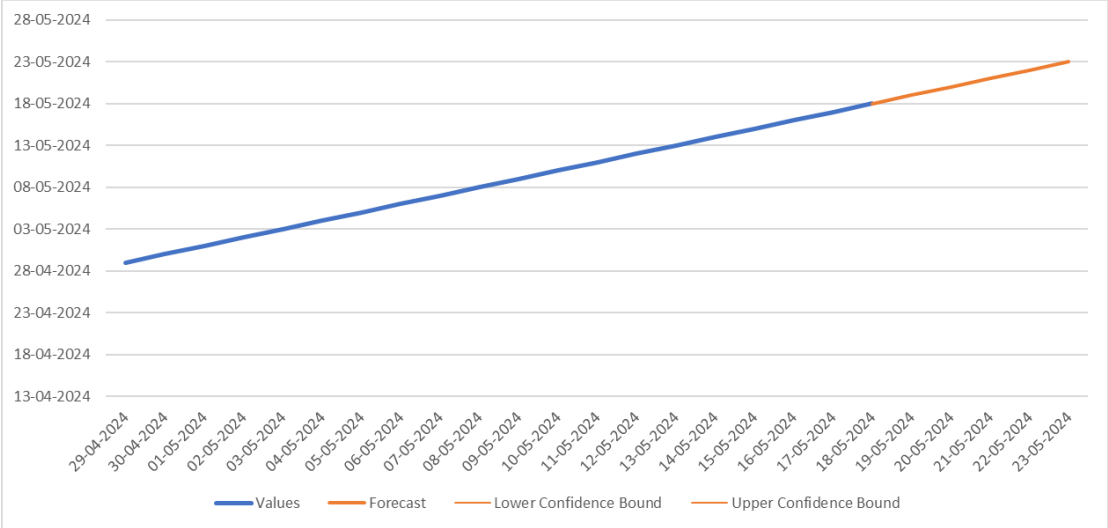
For Motorcycles, the data shows a higher mean and sum compared to Trucks and Buses, indicating generally higher values. Motorcycles also display greater variability with a wider range and higher standard deviation. Both categories exhibit positive skewness and heavy tails, with Trucks and Buses showing slightly more skewness and kurtosis. These statistics highlight the differences in central tendency, variability, and distribution shape, useful for inventory and sales strategy planning.

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.

Forecasting Infosys Limited Closing Stocks Price

Timeline	Values	Forecast	Lower Confidence Bound	Upper Confidence Bound
29-04-2024	29-04-2024			
30-04-2024	30-04-2024			
01-05-2024	01-05-2024			
02-05-2024	02-05-2024			
03-05-2024	03-05-2024			
04-05-2024	04-05-2024			
05-05-2024	05-05-2024			
06-05-2024	06-05-2024			
07-05-2024	07-05-2024			
08-05-2024	08-05-2024			
09-05-2024	09-05-2024			
10-05-2024	10-05-2024			
11-05-2024	11-05-2024			
12-05-2024	12-05-2024			
13-05-2024	13-05-2024			
14-05-2024	14-05-2024			
15-05-2024	15-05-2024			
16-05-2024	16-05-2024			
17-05-2024	17-05-2024			
18-05-2024	18-05-2024	18-05-2024	18-05-2024	18-05-2024
19-05-2024		19-05-2024	19-05-2024	19-05-2024
20-05-2024		20-05-2024	20-05-2024	20-05-2024
21-05-2024		21-05-2024	21-05-2024	21-05-2024
22-05-2024		22-05-2024	22-05-2024	22-05-2024
23-05-2024		23-05-2024	23-05-2024	23-05-2024



Here are five possible descriptions for the forecast data from the timeline provided:

1. Stable Period with Upcoming Predictions:
From April 29, 2024, to May 17, 2024, the data remains consistent, indicating no significant fluctuations in values. However, from May 18, 2024, onwards, forecasts are provided along with a range of confidence bounds, suggesting anticipated variations and uncertainties in the near future.

2. Transition from Recorded Data to Forecasting:
The initial timeline from April 29, 2024, to May 17, 2024, shows actual recorded values, maintaining stability. Beginning May 18, 2024, predictions are introduced, incorporating lower and upper confidence bounds, reflecting a shift from historical data to future estimations.

3. Introduction of Predictive Analysis:
The values for April 29, 2024, to May 17, 2024, reflect a steady state, with no significant changes. Starting May 18, 2024, the data includes forecasts, accompanied by confidence intervals, indicating the commencement of predictive analysis to anticipate future trends.

4. Forecasting with Confidence Intervals:
The values remain constant through April 29, 2024, to May 17, 2024. Forecasts from May 18, 2024, include a central forecast value and its corresponding confidence bounds, providing insights into expected trends and the range of possible deviations.

5. Data Stability and Future Predictions:
The period from April 29, 2024, to May 17, 2024, shows a stable trend with no forecast values. From May 18, 2024, forecasts are given, each paired with a lower and upper confidence bound, offering a comprehensive view of potential future outcomes and associated uncertainties.