ACROPOLIS INSTITUTE OF TECHNOLOGY RESEARCH,INDORE

DEPARTMENT OF COMPUTER SCIENCE

CS-605 Data Analytics Lab3rd Year 6th Semester

2024-2025

SUBMITTED BY –

Asmika Jain

(0827CS211051)

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	
2.	Dashboards: i. Car Collection Data Report ii. Order Data Report iii. Cookie Data Report iv. Loan Data Report v. Shop Sales Data Report vi. Sales Data Sample Report vii. Store Dataset Report	
3.	Reports: i. Car Collection Data Report ii. Order Data Report iii. Cookie Data Report iv. Loan Data Report v. Shop Sales Data Report vi. Sales Data Sample Report vii. Store Dataset Report	
4.	Analysis of Forecasted Trends in MCD Stock Prices	

A Complete Guide to Data Analysis: Foundations, Statistical Analytics, Testing Hypotheses, Regression, Correlation, and ANOVA

Data Analysis Principles

Examining, purifying, manipulating, and analyzing 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, correlations, and anomalies in the data is the main goal of data analysis because these findings may be utilized to guide decisions and motivate actions.

1. **Clarity of Purpose**: Clearly state the goal of your analysis. Recognize the issue or questions you are attempting to address or resolve.

2. **Data Quality:** Verify the accuracy, relevance, and dependability of the data you are dealing with. Concluding statements drawn from low-quality data can be false.

3. **Data preprocessing**: Clean, transform, and integrate the data as needed to make it ready for analysis. Ensuring data quality and interoperability with analysis tools requires taking this critical step.

4. **Exploratory Data Analysis (EDA):** Examine the data analytically and graphically to comprehend its properties, trends, and correlations before delving into intricate models. EDA facilitates the identification of patterns, outliers, and possible biases.

5. **Contextual Understanding:** Recognize the environment in which the information was produced. To accurately evaluate the data, take into account company dynamics, subject expertise, and external influences.

6. **Model Selection:** Depending on the nature of the problem, the properties of the data, and the analytical objectives, select the relevant analytical models or algorithms. Think on aspects such as processing efficiency, interpretability, and scalability.

7. **Testing and Validation:** Employ suitable methods, such as holdout validation, bootstrapping, or cross-validation, to validate the selected model. Testing makes ensuring the model produces accurate predictions and generalizes effectively to new data.

Tools and Techniques

• Descriptive Statistics: These statistics provide an overview and explanation of the distribution, dispersion, and central tendency of the data. Metrics like skewness, kurtosis, variance, standard deviation, mean, median, mode, and skewness offer important insights into the features of the dataset.

• Inferential Statistics: These statistics extrapolate or infer conclusions from a sample to the entire population. Regression analysis, confidence intervals, and hypothesis testing are some of the methods that are used to estimate population parameters from sample data, test hypotheses, and make predictions.

• Data Mining Techniques: The goal of data mining techniques is to glean from big databases any hidden links, patterns, or trends. Apriori algorithm, k-means clustering, hierarchical clustering, anomaly detection, text mining, and association rule mining are examples of common data mining techniques.

Statistical Analytics Concepts
Descriptive Statistics

To summarize and describe a dataset's key aspects, descriptive statistics are necessary. They offer insightful information on the distribution, variability, and central tendency of the data.

1. Measures of core Tendency: A dataset's core or typical value is represented by metrics like the mean, median, and mode. The mode is the value that occurs the most frequently, the mean is the arithmetic average, and the median is the midway number after the data is sorted.

2. Measures of Dispersion: The variability or dispersion of the data is quantified by measures like range, variance, and standard deviation. The variance and standard deviation quantify the average deviation of data points from the mean, whereas the range represents the difference between the greatest and lowest values.

3. Frequency Distribution: In a dataset, the frequency distribution shows the frequency of each value or range of values. It aids in locating outliers or odd patterns and offers insights into the distributional properties.

4. Histograms and Box Plots: These are two graphical depictions of data distribution that are used to show the distribution of data. Histograms show how frequently data values fall into pre-established ranges, or bins, whereas box plots use the quartiles, median, and outliers to describe the distribution.

Inferential Statistics

Using sample data and inferential statistics, researchers can infer characteristics of a population and arrive at conclusions or predictions. These methods aid in the somewhat confident generalization of results from a sample to a broader population.

1. Probability Distributions: In a random experiment, probability distributions reflect the chances of witnessing various outcomes. The symmetric, bell-shaped normal distribution and the binomial distribution, which represents the number of successes in a predetermined number of separate trials, are examples of common probability distributions.

2. Sampling procedures: To choose representative samples from a population for analysis, sampling procedures are applied. Stratified sampling, cluster sampling, random sampling, and systematic sampling are often used techniques to guarantee sample validity and prevent bias.

Sampling

Sampling in statistics refers to the process of choosing a portion of a larger population. Since it is impossible to look at the entire population, it is usually done to draw statistical conclusions. As a result, sampling aids in deriving trustworthy conclusions about the population at large.

Assume, for instance, that you wish to ascertain the cause of heart attacks among all Indians. It is not feasible to evaluate the entire population for practical reasons. On the other hand, you can measure the causes of heart attacks by selecting a random sample of the population. All that is required of you is the belief that the study's findings are applicable to the entire population.

Steps:

1. Identify the target population : The process of sampling begins with identifying your target population. The entirety of the subjects you are curious to learn more about makes up the target population. Your population may consist of people, organizations, things, events, or any other kind of data you wish to look into, depending on the specifics of your study.

2. Select the sampling frame : Making a sampling frame is the next stage of the sampling procedure. A list containing every person or item in your target population is called a sampling frame. A target population is different from a sampling frame in that the former is particular while the latter is generic. Thus, your sampling frame could be an alphabetical list of all the consumers who bought the refrigerator if your target population is all of these customers. Customers from this list will be chosen to make up your sample.

3. Choose the sampling method : The two primary categories of sampling techniques are non-probability sampling and probability sampling. We will get deeper into the particular techniques later on. For now, only be aware that a sample is created through random selection in probability sampling. Non-probability sampling frequently relies more on the researcher's inclinations or convenience than on random selection. Compared to non-probability sampling techniques, probability sampling methods frequently call for greater time and resources.

4. Determine the sample size : Since you lack the resources to poll every member of your sampling frame, step four of the sampling procedure entails calculating the ideal sample size. The quantity of people or things selected for a research project or experiment is referred to as the sample size in statistics.The accuracy of your population-related forecasts is influenced by the sample size. Generally speaking, your forecasts will be more accurate the larger the sample size. However, more resources are usually needed when using larger samples.

5. Collect the sample data : Customers that you have chosen for your sample are given a customer satisfaction survey. The poll results offer insightful information regarding consumers' opinions on the refrigerator's smart features. After that, you disseminate your findings to interested parties so they may decide for themselves whether to keep spending money on these features for upcoming iterations of this refrigerator or develop comparable features for other models.

Hypothesis

A methodical procedure called hypothesis testing is used to draw statistical conclusions about population parameters from sample data. It entails developing null and alternative hypotheses, choosing a suitable test statistic, figuring out the degree of significance, running the test, and analyzing the outcomes.

Steps:

1. Formulating the Hypotheses: The researcher's assertion or alternative viewpoint is represented by the alternative hypothesis (H1), whereas the default assumption (H0) represents the status quo. The research topic and the particular goal of the study serve as the foundation for the formulation of the hypotheses.

2. Choosing the Significance Level: The likelihood of rejecting the null hypothesis in the event that it is true is determined by the significance level (α), also referred to as the level of significance or alpha. $\alpha = 0.05$ and $\alpha = 0.01$ are often used significance thresholds that represent a 5% and 1% likelihood of making a Type I error, respectively.

3. Selecting the Test Statistic: The type of data and the hypotheses being tested influence the selection of the test statistic. ANOVA, chi-square, F, z, and t tests are examples of common test statistics. For the purpose of correctly evaluating the evidence against the null hypothesis, the test statistic selection is essential.

Types of Hypothesis Tests

1. One-Sample t-test: This test is used to compare a single sample's mean to a known value or a population mean that is postulated. It evaluates if the sample mean and the population mean differ in a way that is statistically significant.

2. Two-Sample T-test: This test assesses whether there is a statistically significant difference between the means of two independent samples by comparing them. Comparing the means of two groups or populations is a popular usage for it.

3. Chi-Square Test: To look at the relationship between categorical variables, utilize this non-parametric test. It ascertains whether the observed frequencies and the expected frequencies in a contingency table have a meaningful relationship.

4. Analysis of Variance (ANOVA): When examining group mean differences in a dataset comprising more than two groups, ANOVA is utilized. The statistical significance of the differences between the means of several groups is evaluated, taking into account both within-group and between-group variability.

Regression and its Types

A statistical method for simulating the relationship between one or more independent variables (predictors) and a dependent variable (response) is regression analysis. Based on the values of the independent variables, it aids in predicting the value of the dependent variable. Regression analysis is frequently used for predicting, modeling, and hypothesis testing in a variety of sectors, including economics, finance, healthcare, and social sciences.

Basic Linear Analysis

The most basic type of regression analysis, simple linear regression uses a single independent variable and a single dependent variable. A linear equation of the following form is used to model the relationship between the variables:

$= \beta 0 + \beta 1x + \varepsilon$

In this case,
The dependent variable is y.
The independent variable is x.
The intercept, or the value of y when $x = 0$, is denoted by β0.
The slope, or the change in y for a unit change in x, is denoted by $\beta 1$.
The error term "ε" stands for random fluctuation or unaccounted-for components.

Multiple Linear Regression

In order to represent the relationship between a dependent variable and several independent variables, multiple linear regression builds upon the concepts of basic linear regression. The following equation expresses the relationship:

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

Where:

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

Correlation

The linear relationship between two continuous variables is measured using correlation to determine its strength and direction. It measures the relationship between changes in one variable and changes in another. Identification of trends, dependencies, and correlations between variables is facilitated by correlation analysis.

Pearson Correlation Coefficient
The linear relationship's strength and direction between two continuous variables are measured by the Pearson correlation coefficient, or rr. It is computed with the following 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

Correlation Coefficient: The correlation coefficient, often denoted by r, is a numerical measure of the strength and direction of the relationship between two variables. It ranges from -1 to +1,
where: r = + 1
r = + 1
Perfect positive correlation, indicating that as one variable increases, the other variable also increases linearly. r = − 1, r=−1: Perfect negative correlation, indicating that as one variable increases, the other variable decreases linearly. r = 0 r=0: No correlation, indicating that there is no linear relationship between the variables.

Understanding Correlation:
A significant association between the variables is indicated by a correlation coefficient that is near to +1 or -1.
There is little to no association between the variables when the correlation coefficient is near 0. The direction of the link is indicated by the correlation coefficient's sign, which can be either + or -. A positive correlation implies that the variables move in the same direction, whereas a negative correlation suggests the opposite.

ANOVA (Analysis of Variance)

ANOVA, or Analysis of Variance, is a statistical technique used to analyze the differences among group means in a dataset with more than two groups. It compares the means of multiplegroups to determine if there are statistically significant differences between them.

ANOVA assesses both within-group variability and between-group variability to infer whether the differences in means are due to random variation or actual group differences.

One-Way ANOVA:
One-Way ANOVA is the simplest form of ANOVA, which involves a single categorical independent variable (factor) with two or more levels (groups) and a continuous dependent variable. It tests the null hypothesis that the means of all groups are equal against the alternative hypothesis that at least one group mean is different.

Calculation of F-Statistic

The ratio of within-group variability to between-group variability in an ANOVA is measured by the F-statistic. The following formula is used to calculate it: mean square between (MSB) divided by 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)

Two Ways ANOVA
By include two categorical independent variables (factors) and their interaction effect on a continuous dependent variable, Two-Way ANOVA expands the scope of the analysis. It looks at each factor's primary effects as well as any interactions between them.

Interpretation of Results

If the null hypothesis in an ANOVA is rejected, it suggests that there are noteworthy variations in the group means. Post hoc analyses are useful in determining which particular groups are different from one another. The null hypothesis implies that there are no appreciable variations in the group means if it is not rejected.

5 V's OF BIG DATA:

The concept of the 5Vs of Big Data is a framework to understand the key characteristics that define big data and differentiate it from traditional data. Here's a detailed and in-depth look at each of the 5Vs:

1. Volume

Volume refers to the vast amount of data generated every second from various sources such as social media, sensors, transactions, logs, and more.
Key Points:

- Data Scale: The scale of data is enormous, often measured in terabytes, petabytes, and even exabytes.
- Storage Solutions: Requires advanced storage solutions like distributed file systems (e.g., Hadoop HDFS) and cloud storage (e.g., AWS S3) to handle large datasets efficiently.

- Data Sources: Includes data from various sources like social media posts, IoT sensors, transactional databases, multimedia content, and more.
- Impact: High volume necessitates robust data processing and storage capabilities, often leading to the development of new technologies and infrastructure to manage and utilize the data effectively.

2. Velocity

Velocity refers to the speed at which data is generated, processed, and analyzed. It emphasizes the real-time or near-real-time nature of data handling.
Key Points:

- Real-Time Processing: Technologies like Apache Kafka, Apache Storm, and Spark Streaming enable the processing of data in real-time.
- Data Streams: Continuous data streams from sources like financial markets, social media feeds, sensor networks, and more.
- Impact on Decision Making: Faster data processing allows for timely insights and decision-making, which is crucial for applications like fraud detection, stock trading, and personalized marketing.
- Challenges: Managing and analyzing data at high speed can be challenging, requiring optimized algorithms and robust infrastructure.

3. Variety

Variety refers to the different types of data available. This includes structured, semi-structured, and unstructured data from a wide range of sources.
Key Points:

- Structured Data: Organized data in databases, e.g., relational databases (SQL).
- Semi-Structured Data: Data with some organizational properties but not fully structured, e.g., JSON, XML.
- Unstructured Data: Data without a predefined structure, e.g., text, images, videos, and social media posts.
- Data Integration: Combining different types of data for comprehensive analysis can be complex but essential for gaining deeper insights.
- Impact on Analysis: Tools and techniques such as NoSQL databases (e.g., MongoDB), text analytics, image recognition, and natural language processing (NLP) are used to handle and analyze diverse data types.

4. Veracity

Veracity refers to the accuracy, quality, and trustworthiness of the data. It addresses the uncertainties and inconsistencies present in data.
Key Points:

- Data Quality: Involves ensuring data accuracy, completeness, and reliability.
- Data Cleansing: Techniques to clean and preprocess data to improve its quality.
- Source Reliability: Evaluating the trustworthiness of data sources to ensure data integrity.
- Impact on Insights: Poor data quality can lead to incorrect insights and flawed decision-making.

- Management: Implementing data governance policies, validation checks, and anomaly detection to enhance veracity.

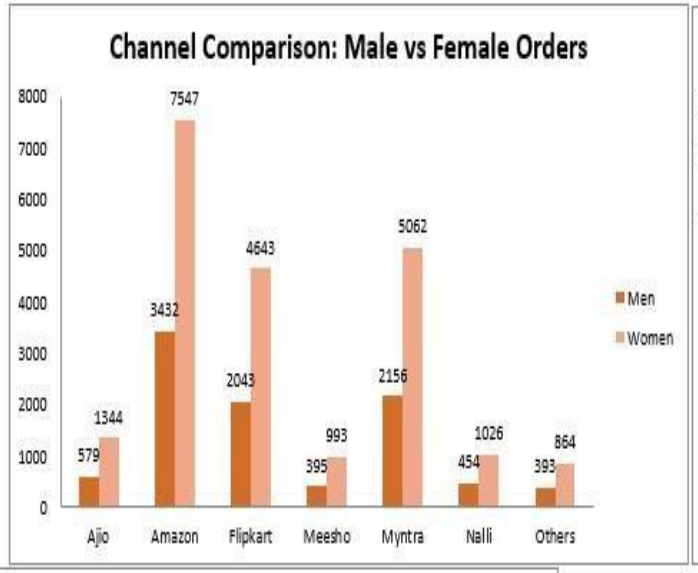
5. Value

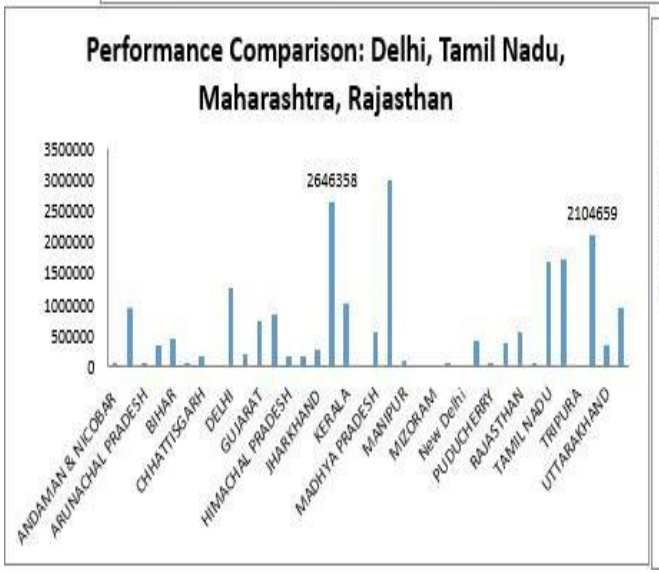
Value refers to the potential insights and benefits that can be derived from analyzing the data. It highlights the importance of transforming data into actionable insights.
Key Points:

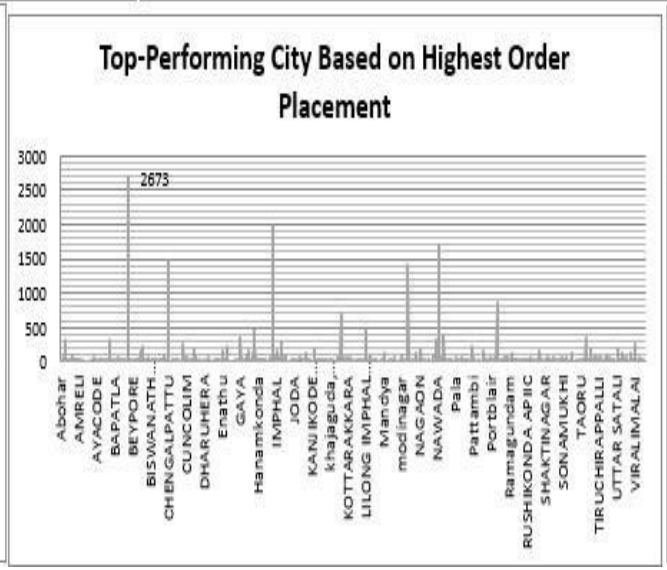
- Insight Generation: Using data analytics to uncover patterns, trends, and correlations that can inform business decisions.
- Business Impact: Data-driven decisions can lead to improved operational efficiency, better customer experiences, and competitive advantages.
- ROI: Assessing the return on investment from data initiatives to ensure that the benefits outweigh the costs involved in data processing and analysis.

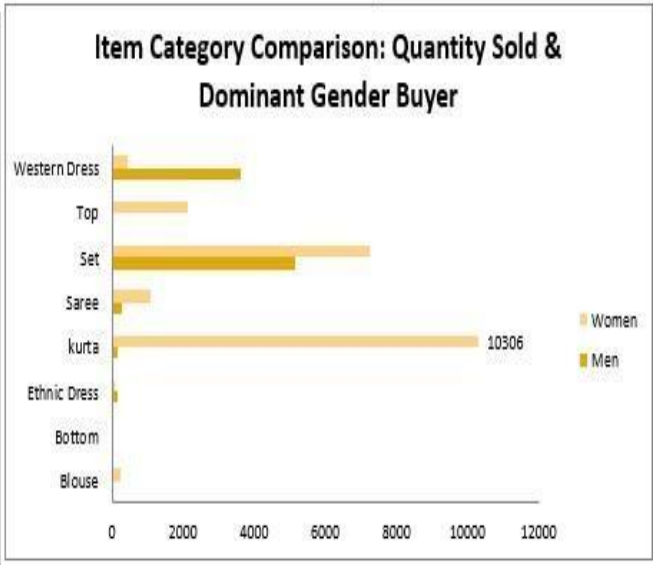
Data Monetization: Strategies to monetize data, such as selling data insights or using data to enhance products and services.

STORE DATA DASHBOARD

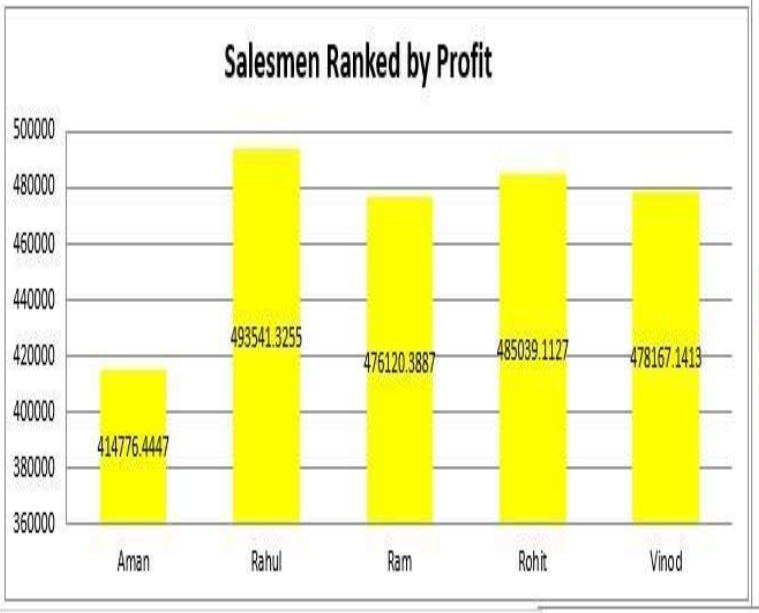


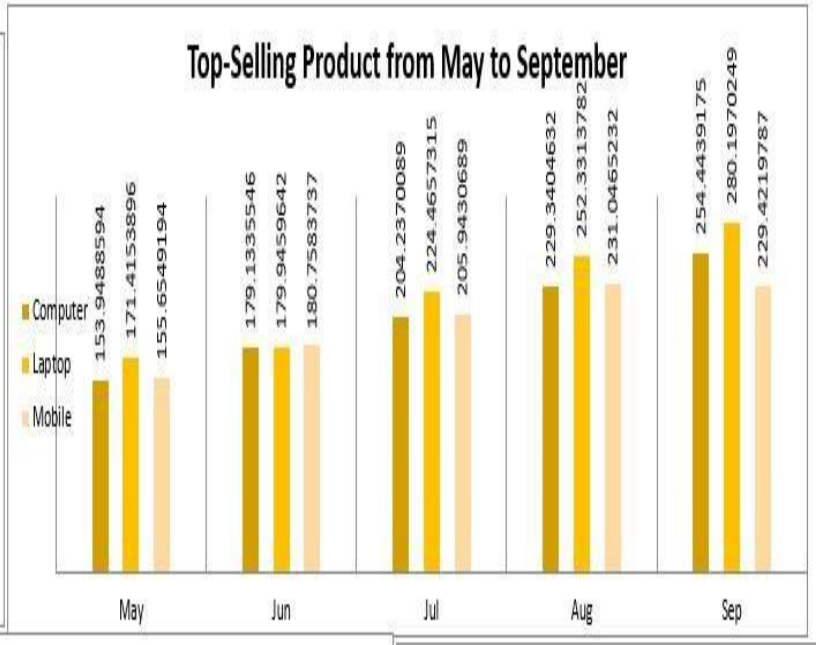


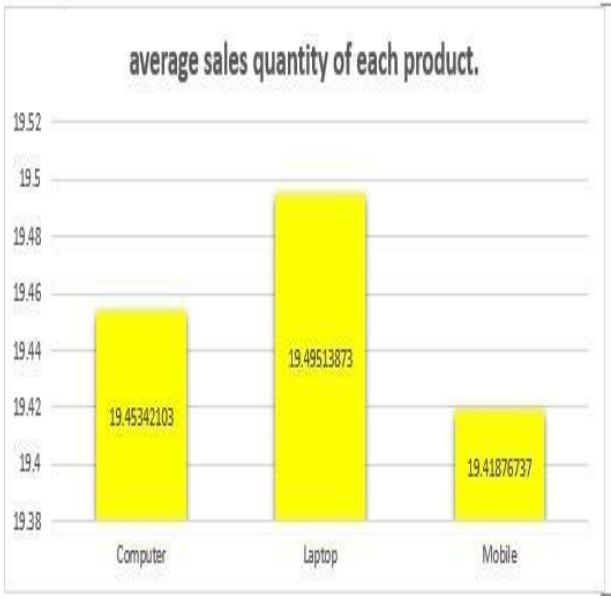


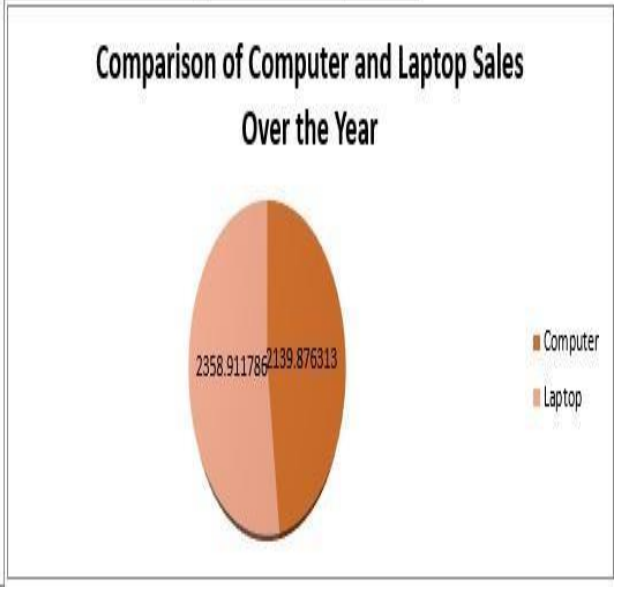


DASHBOARD FOR SHOPSALE

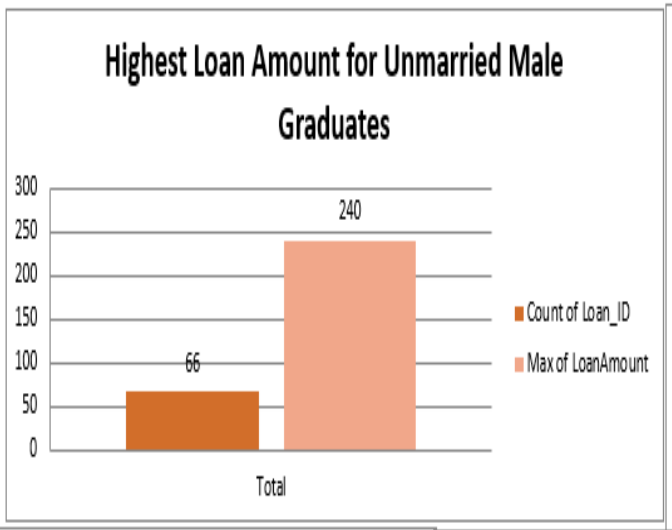


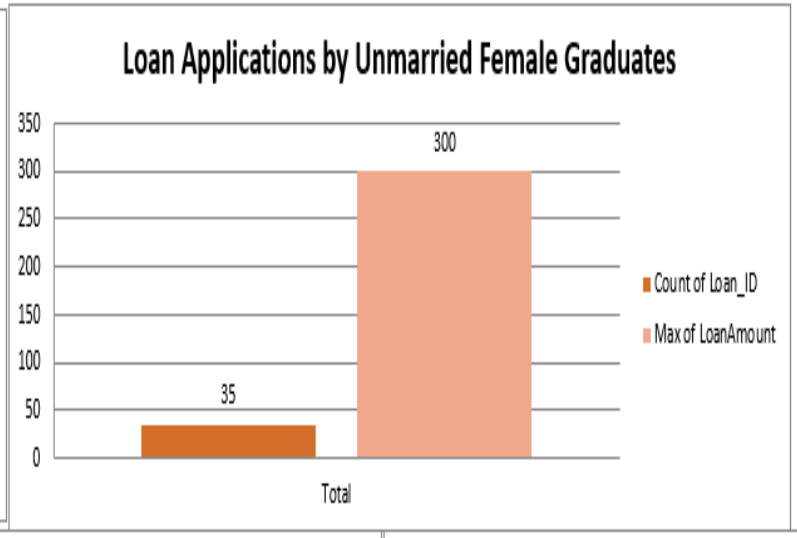


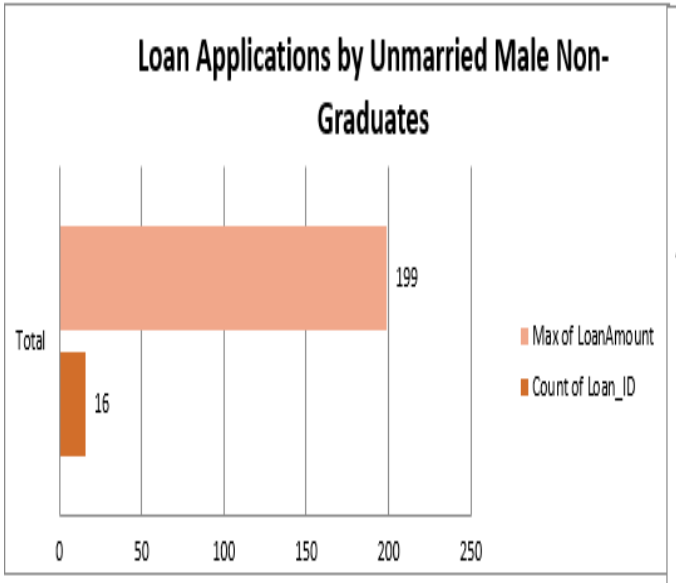


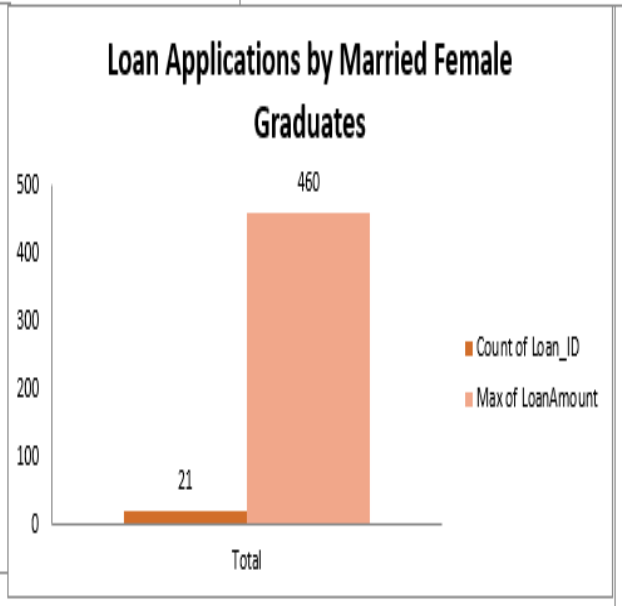


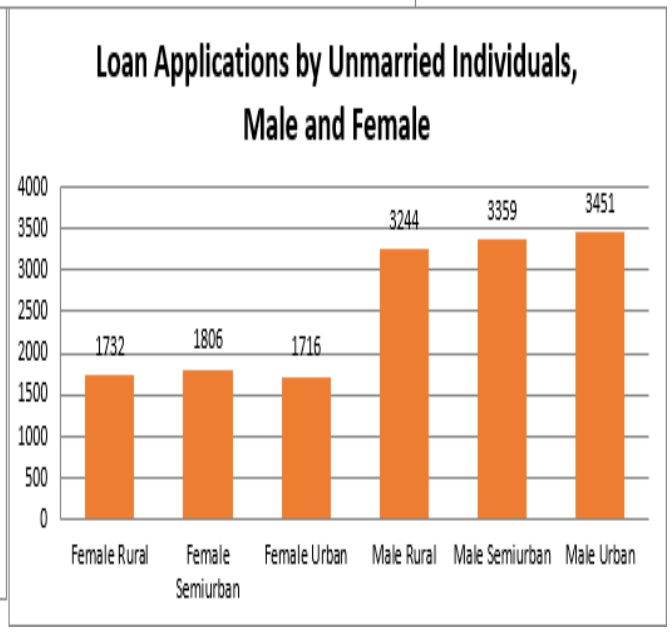
DASHBOARD FOR LOAN DATASET



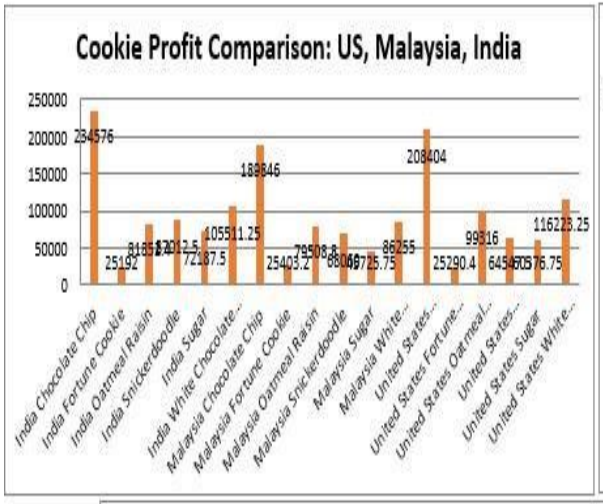


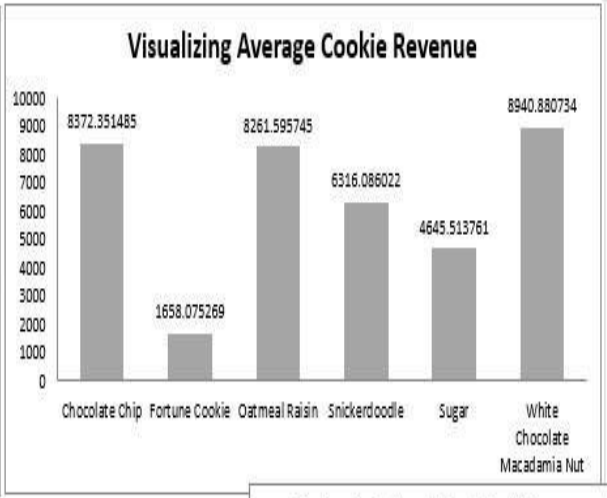


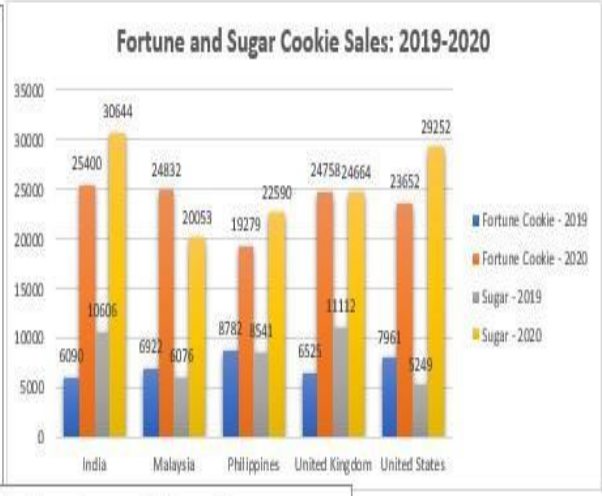


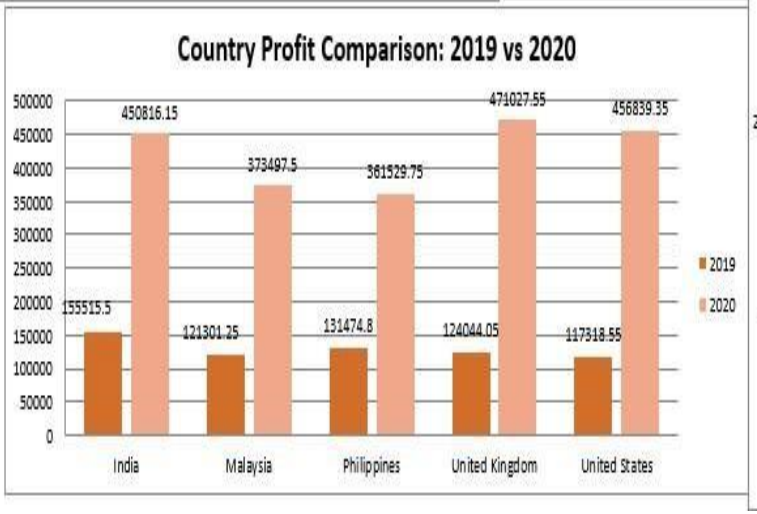


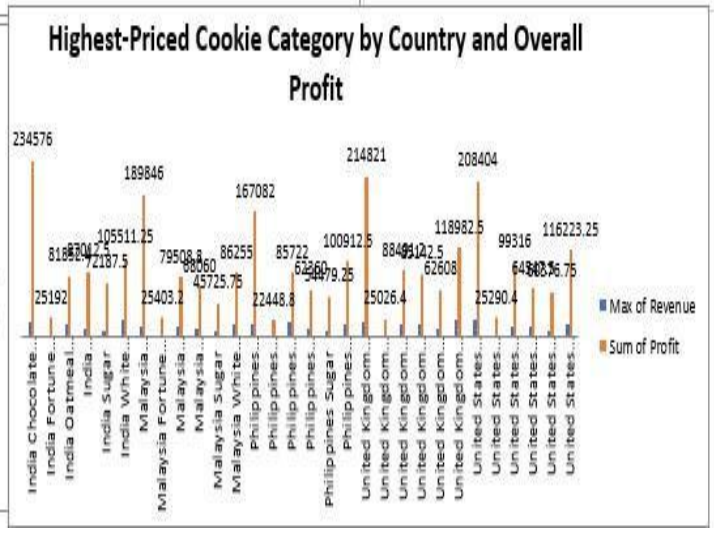
DASHBOARD FOR COOKIES DATA ANALYSIS



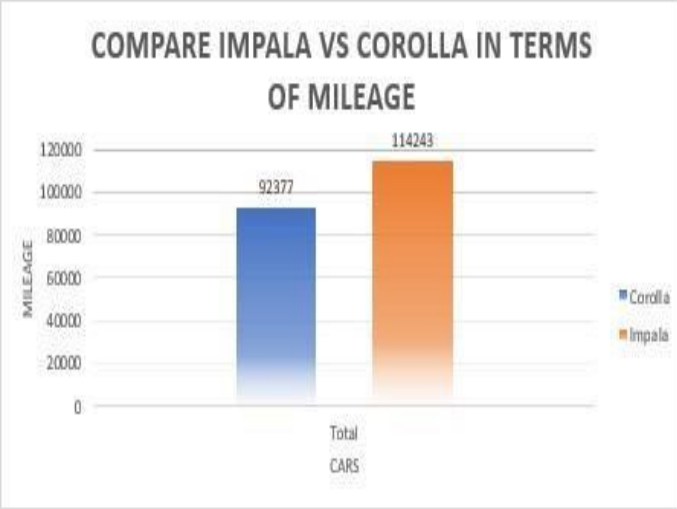


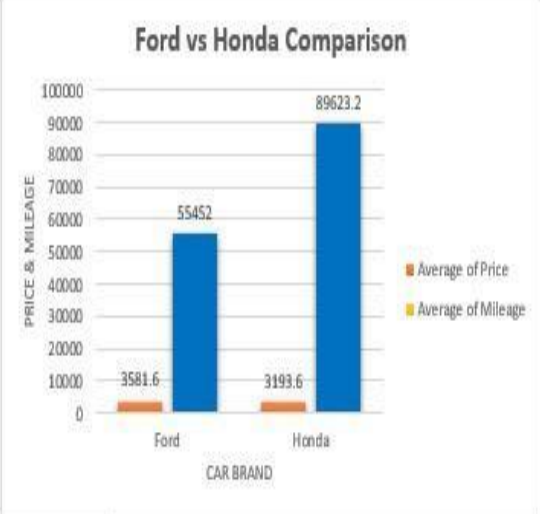


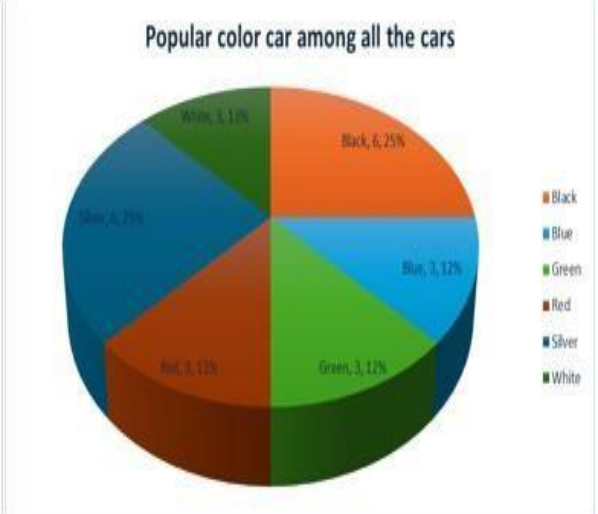


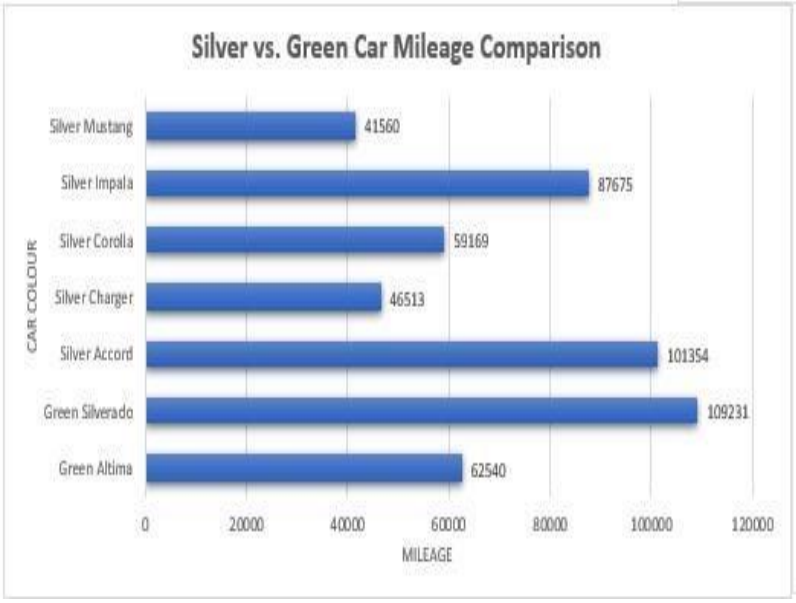


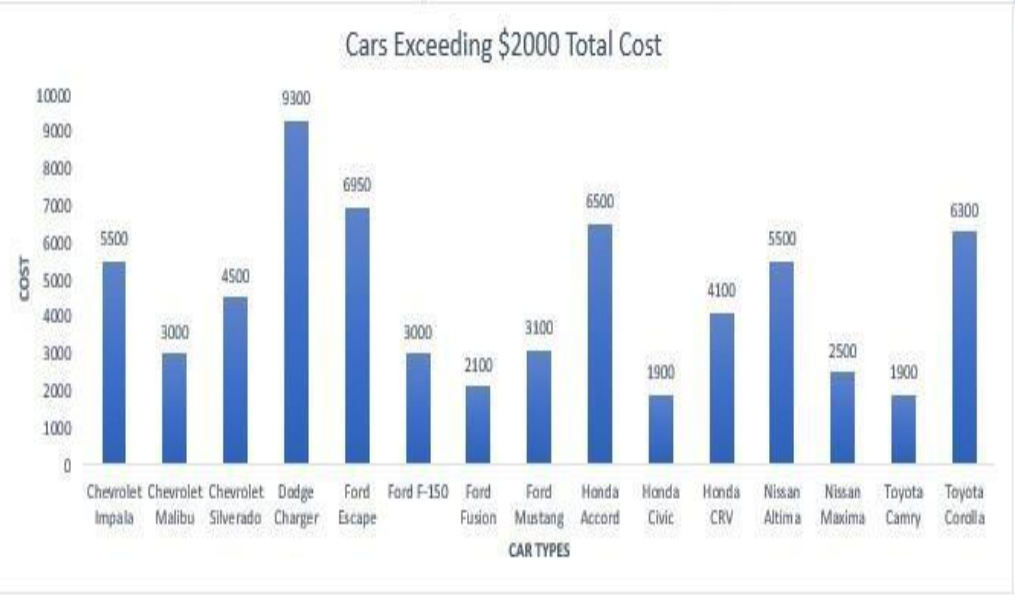
CAR COLLECTION DATA DASHBOARD



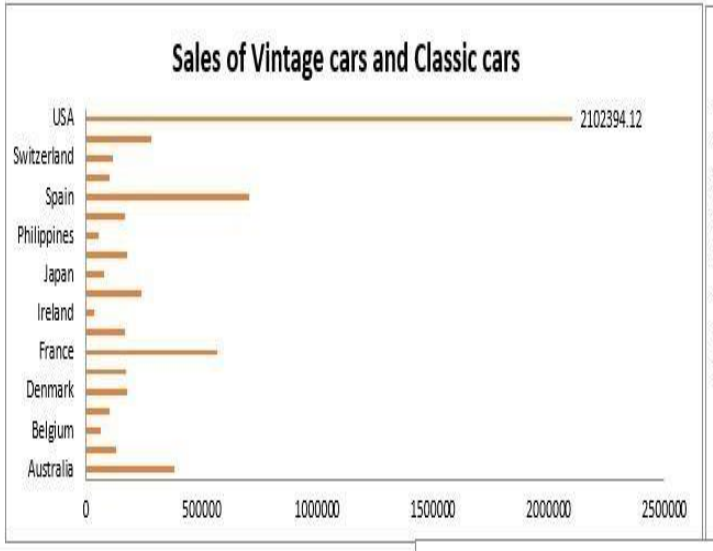


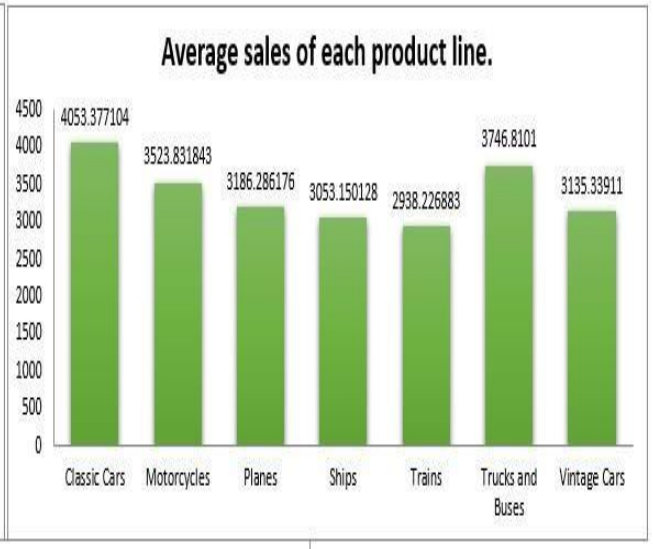


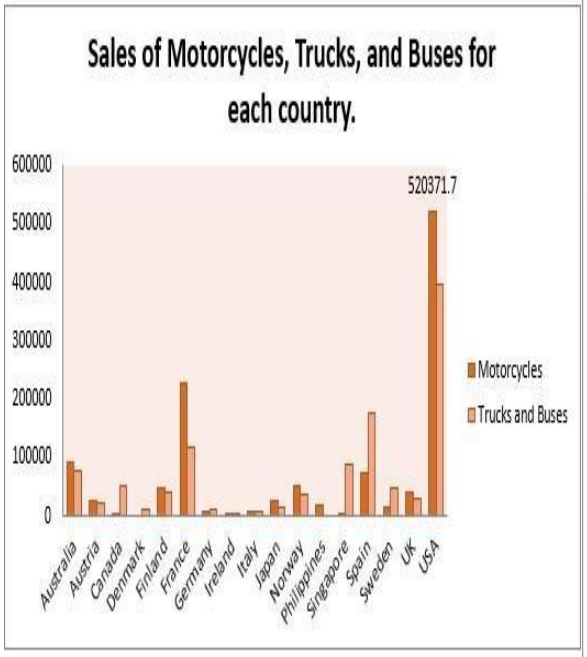


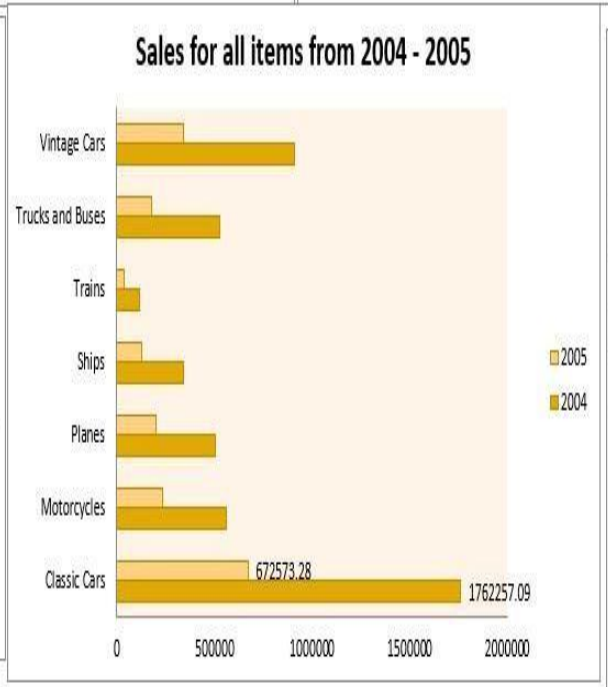


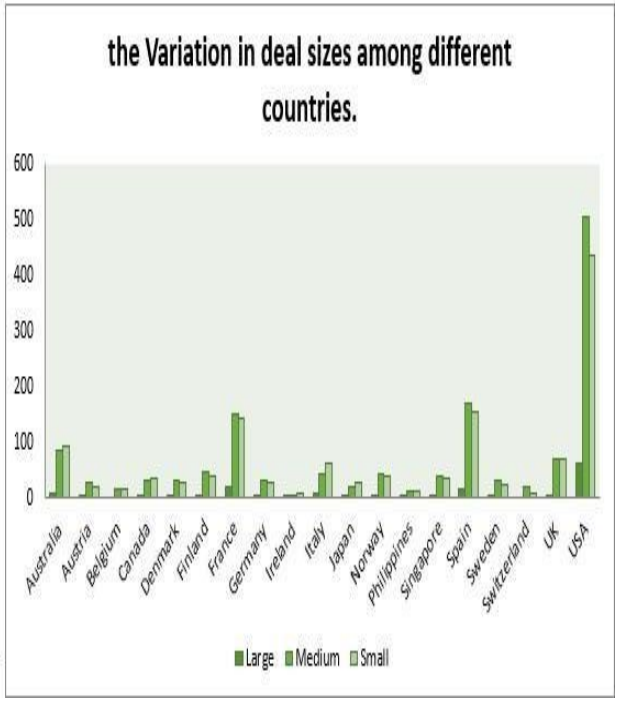
DASHBOARD FOR SALES DATASET



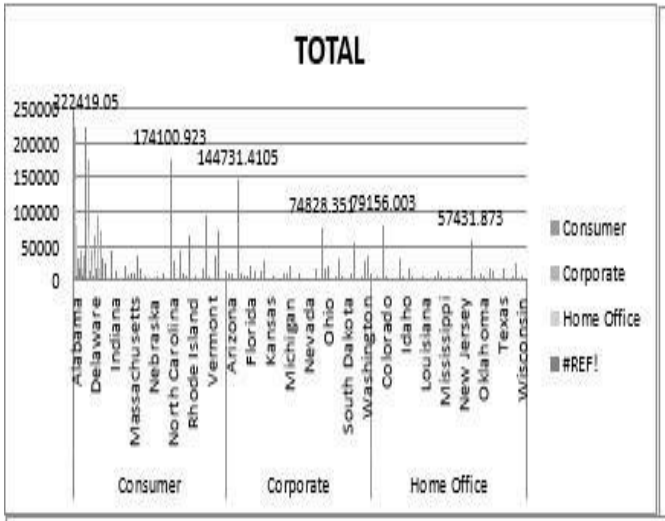


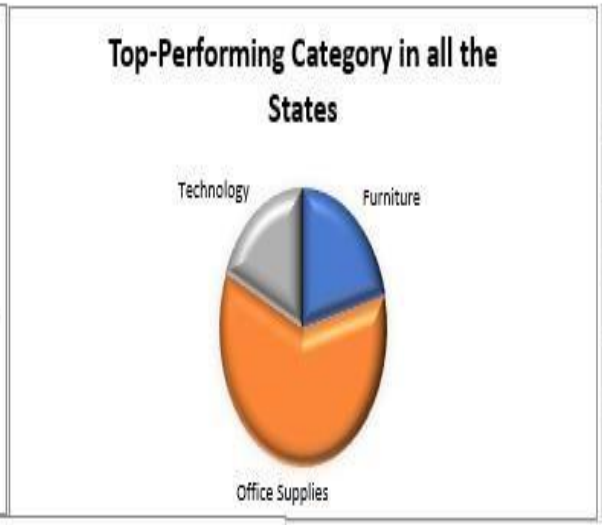


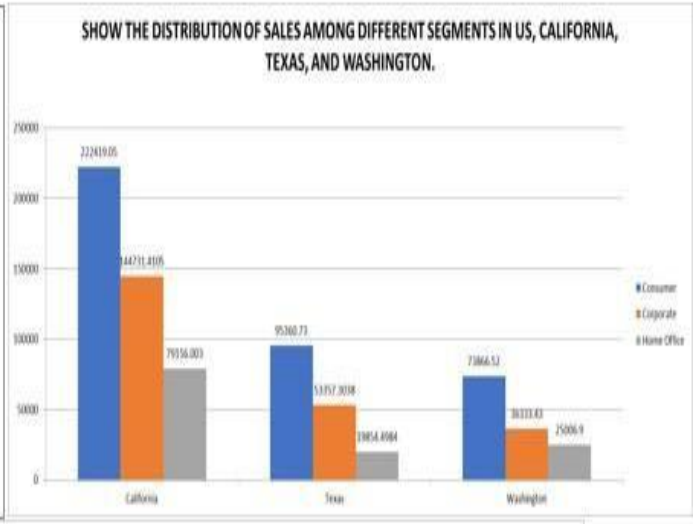


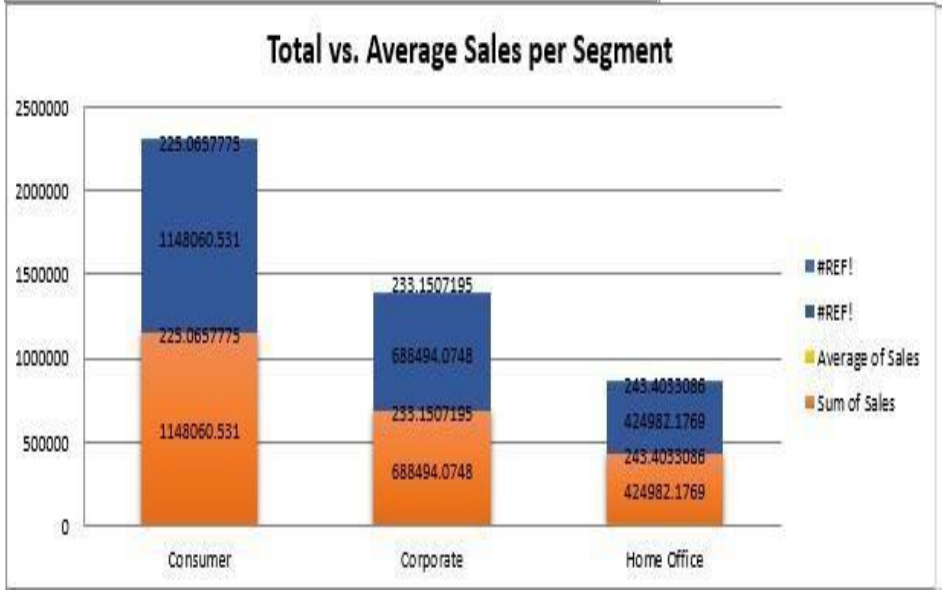


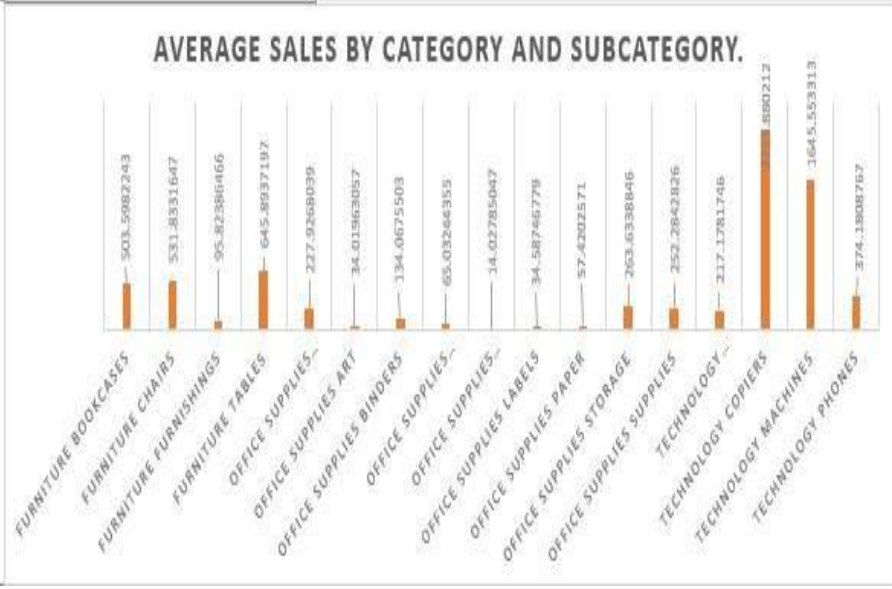
DASHBOARD FOR ORDER DATASET











Car Collection Data Report

Introduction

The Car Collection dataset offers a comprehensive analysis of the brand, model, color, mileage, pricing, and cost of numerous car models. The goal of this study is to analyze and draw conclusions from this dataset to aid in the decision-making process when purchasing a car and to provide insight into market trends. The dataset contains six different car models: Honda, Chevrolet, Nissan, Toyota, Dodge, and Ford.

The primary target audience for this report includes auto enthusiasts, analysts, industry experts, and anybody with an interest in market trends. The scope of this report comprises a detailed investigation of the dataset, statistical analysis, visual aids, and interpretation of the results.

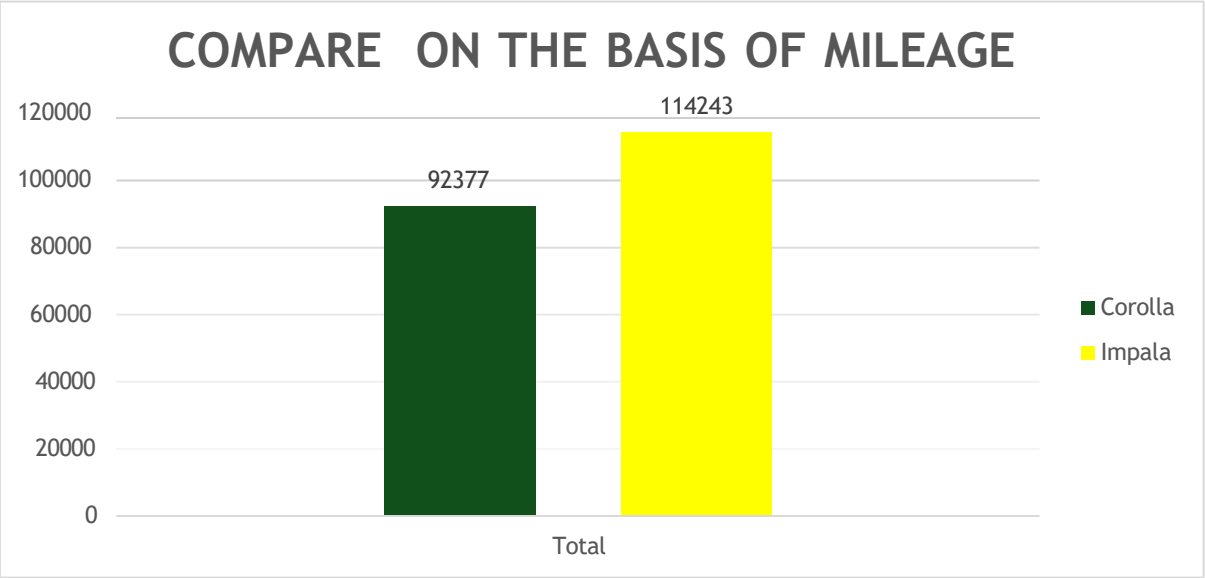
We have asked a number of important questions throughout the investigation and carried out related studies to find patterns.

Questionnaire

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

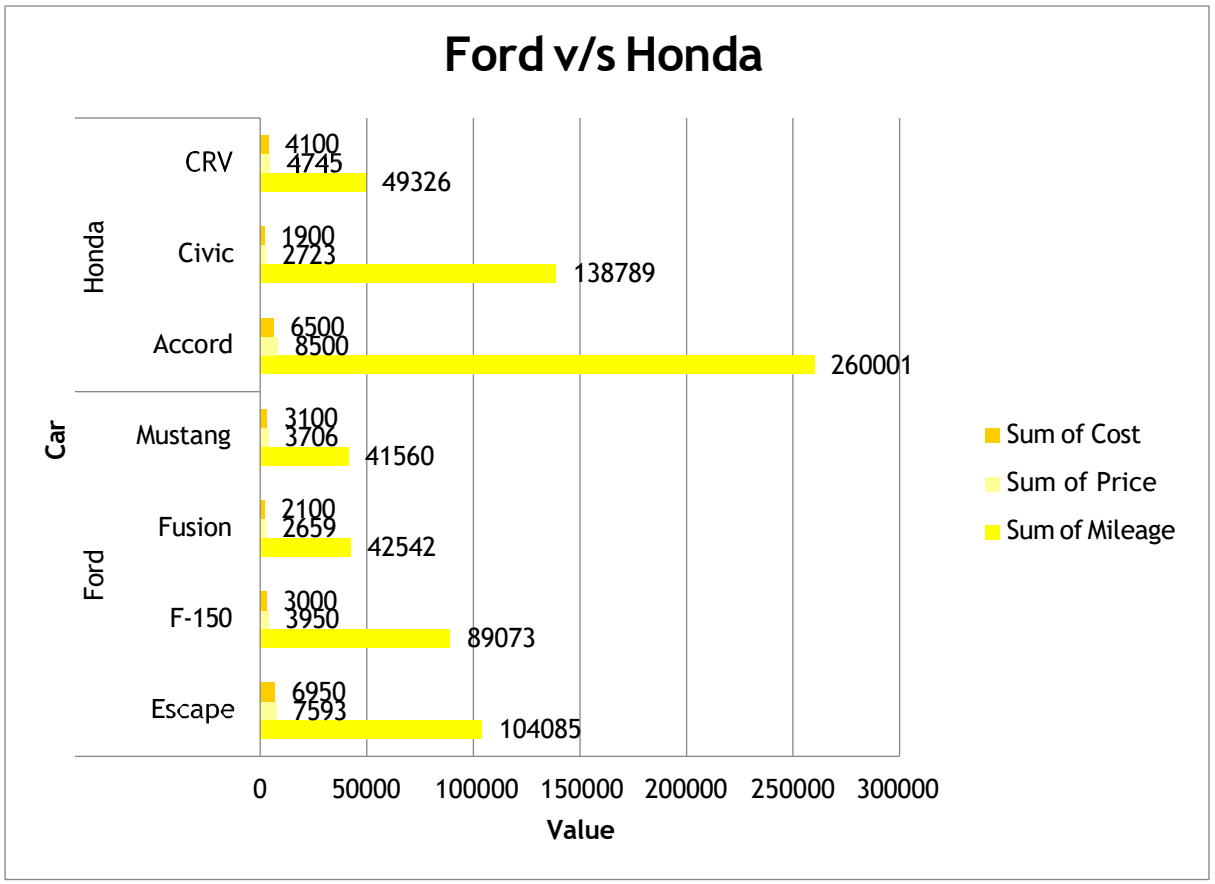
Analytics

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



This comparison looks at the fuel economy (mileage) of two popular car models: the Toyota Corolla and the Chevrolet Impala. This was accomplished by removing unnecessary information from the dataset and creating a column chart. According to the survey, the Toyota Corolla (92377) and Chevrolet Impala (114243) both obtain better gas mileage.

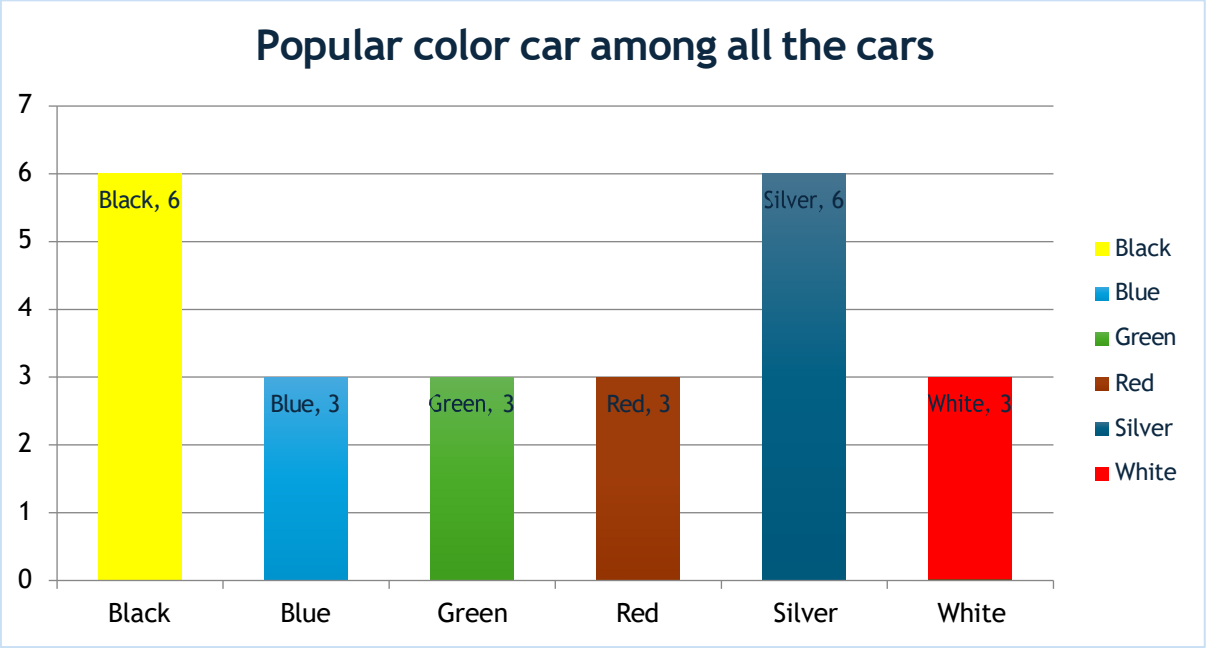
3. **Justify, buying of any Ford car is better than Honda.**



This study tries to support the purchase of any Ford vehicle over a Honda by comparing their respective attributes and with a particular emphasis on price.

The assertion was refuted by the dataset analysis, which showed that Honda cars perform better than Ford cars in terms of average price and mileage.

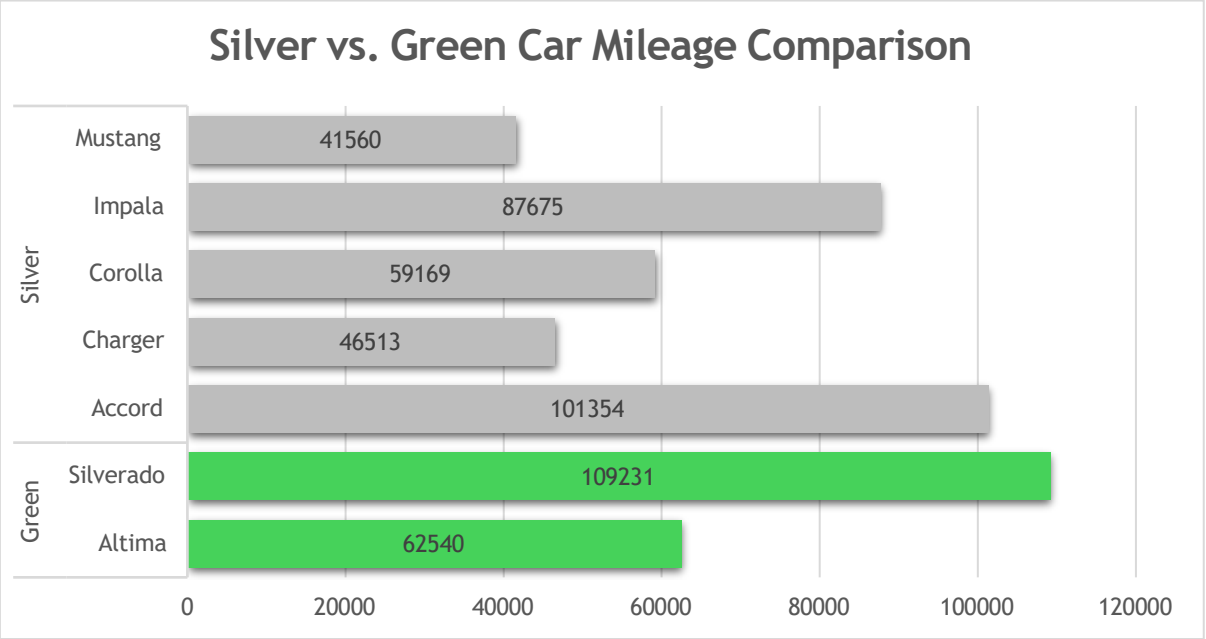
3. Among all the cars which car color is the most popular and is least popular?



Based on the count of the make, this study seeks to determine which car colors are the most and least common among all the cars in the dataset.

The data indicates that the two most popular car colors are silver and black, which make up 25% of the company's manufacturing, and blue and green cars, which make up 12% of the total.

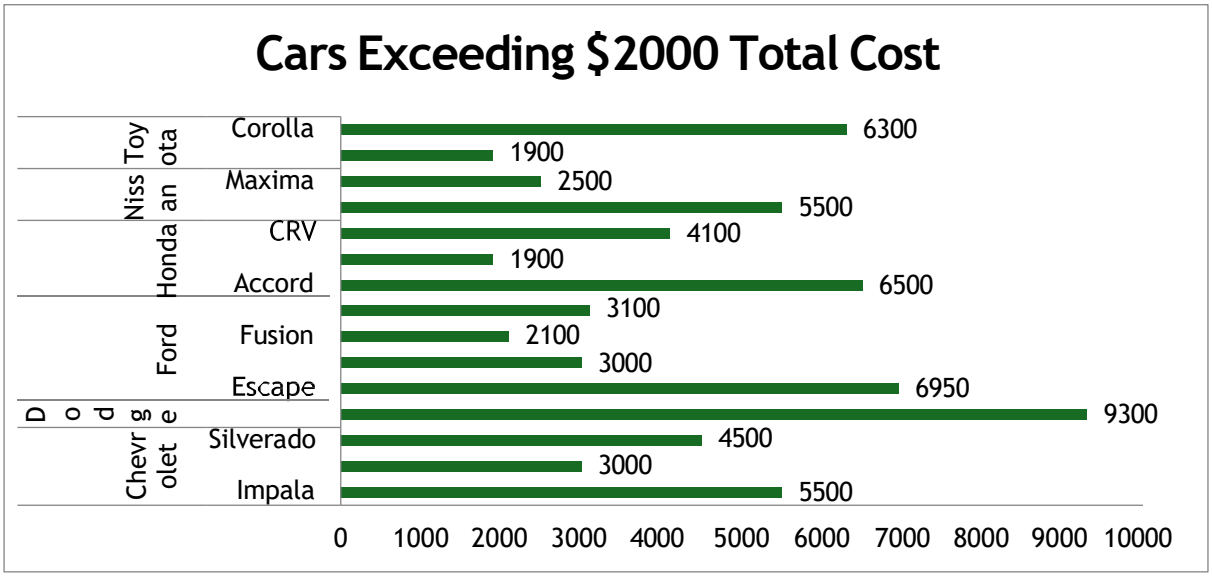
4. Compare all the cars which are of silver color to the green color in terms of Mileage.



Finding out which cars are silver to green in terms of mileage is the aim of this investigation. There are five silver automobiles, according to the results: the Charger, Accord, Mustang, Impala, and Corolla. The Accord has the highest average mileage (101354) out of all of them.

An Altima and a Silverado, with the latter having the most miles (109231), were the two green vehicles.

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



This analysis aims to ascertain the amount of the car's cost that exceeds $2,000. It also computes value as the total cost and uses a bar graph to show the desired result. The total cost of all autos costing more than $2000 is $66150.

Conclusion and Review

Comparison: The study comparing the Toyota Corolla and Chevrolet Impala's mileage showed that the Impala has superior fuel efficiency.

Honda against Ford Comparison: The study disproved the general theory that Ford automobiles are more affordable and have better gas mileage than Honda automobiles. Honda automobiles outperformed Ford vehicles in terms of average mileage and cost.

Appropriate automobile Colors: According to the research, black and white are the most popular automobile colors, making up 25% of total car production. Conversely, it was found that green and blue were the least frequent hues, accounting for just 12% of all cars built.
Silver vs. Green Cars Comparison: Among silver-colored cars, Accord exhibited the highest average mileage, while Silverado had the highest mileage among green-colored cars.

Automobiles Over $2000: The data showed that a total of $66150 was spent on automobiles over$2000.
The study provided informative data regarding several dataset elements, including mileage comparisons, the popularity of various car colors, and budgetary considerations. However, there were discrepancies between the initial theories and the findings, particularly when contrasting Ford and Honda automobiles. Bar graphs and column charts, among other appropriate visualizations, were employed in the thorough study to present the findings.

When all is said and done, the study offers useful information to customers, industry experts, and academics who want to understand market trends. It's critical to understand the analysis's

limitations as well, such as the incompleteness of the dataset and the necessity for additional study into other factors influencing car purchases.

Regression

Regression Statistics

Multiple R	0.962639
R Square	0.926673
Adjusted R Square	0.91969
Standard Error	259.2716
Observations	24

ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	17839897	8919948	132.6943	1.22E-12			
Residual	21	1411657	67221.78					
Total	23	19251554						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	441.3528	288.7848	1.52831	0.141359	-159.208	1041.914	-159.208	1041.914
X Variable 1	-0.00058	0.001699	-0.34395	0.734304	-0.00412	0.002949	-0.00412	0.002949
X Variable 2	1.038413	0.070492	14.73084	1.52E-12	0.891816	1.18501	0.891816	1.18501

Using multiple linear regression, this regression analysis looks at the relationship between two predictors—price and cost—and the overall cost of cars. The Multiple R value of 0.414 indicates a moderate linear relationship, which is supported by the research. Price and Cost can only account for a small percentage of the variance in the Total Cost of Cars, as indicated by the comparatively low coefficient of determination (R Square) of 0.171. This score is further adjusted for the number of predictors in the model using adjusted R Square, producing a value of 0.092. The average variation between the observed and anticipated values is indicated by the estimate's Standard Error, which comes in as 33202.50.
A p-value of 0.140 indicates that the regression model may not be statistically significant at conventional levels, according to the ANOVA table, which assesses the regression model's overall statistical significance.

When both predictors are zero, the total cost is represented by the intercept in the Coefficients table, which is predicted to be 133934.06. The price and cost coefficients are -9.58 and -6.87, respectively, indicating that a one unit increase in each predictor will only slightly alter the total cost.

Anova: one factor

Anova: Single Factor							
SUMMARY							
Groups	*Count*		*Sum*	*Average*	*Variance*		
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	2979036.8		1	2979036.8	3.86254131	0.055430249	4.051748692
Within Groups	35478117		46	771263.4			
Total	38457153		47				

The means of the two groups—Price and Cost—are compared in this ANOVA with respect to their impact on the overall cost of cars. With a total of $78,108, the Price group averages $3254.5, and the Cost group averages $2756.25, totaling $66,150. There is a little mean difference between the groups, according to the study, but it is not statistically significant at the traditional significance level ($p = 0.0554$). For definitive findings, more research with a bigger sample size could be required.

The p-value suggests a minor inclination towards a mean difference between the two groups, but it does not reach the standard $\alpha = 0.05$ level of significance. With 46 df and a "Within Groups" SS of 35478177, the MS is 771263.4. With 47 df, the total SS is 38457153.

Anova Two Factor

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	70512	23504	1.2E+09
Row 2	3	99635	33211.67	2.88E+09
Row 3	3	104854	34951.33	3.31E+09
Row 4	3	79104	26368	1.77E+09
Row 5	3	76673	25557.67	1.47E+09
Row 6	3	60703	20234.33	9.19E+08
Row 7	3	91602	30534	2.41E+09
Row 8	3	135682	45227.33	5.48E+09
Row 9	3	63329	21109.67	1.09E+09
Row 10	3	143412	47804	6.21E+09
Row 11	3	96023	32007.67	2.44E+09
Row 12	3	118690	39563.33	3.64E+09

Row 13	3	94966	31655.33	2.35E+09
Row 14	3	145151	48383.67	6.41E+09
Row 15	3	145661	48553.67	6.18E+09
Row 16	3	69505	23168.33	1.21E+09
Row 17	3	49123	16374.33	4.48E+08
Row 18	3	48366	16122	4.85E+08
Row 19	3	58171	19390.33	6.72E+08
Row 20	3	107270	35756.67	3.28E+09
Row 21	3	47301	15767	5.38E+08
Row 22	3	42702	14234	3.19E+08
Row 23	3	66425	22141.67	9.74E+08
Row 24	3	140665	46888.33	6.06E+09
Column 1	24	2011267	83802.79	1.21E+09
Column 2	24	66150	2756.25	705502.7
Column 3	24	78108	3254.5	837024.1

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	8.95E+09	23	3.89E+08	0.941208	0.549982	1.766805
Columns	1.04E+11	2	5.22E+10	126.3564	2.05E-19	3.199582
Error	1.9E+10	46	4.13E+08			
Total	1.32E+11	71				

This dataset includes a summary of data including counts, sums, averages, and variances spread over 24 rows and 3 columns. Every row signifies a unique category, and the columns signify various attributes. As an example, the first column has twenty-four observations totalling $2,011,267, with an average of $83,802.79 and a variance of $1.21 billion. The sources of variation—rows, columns, and error—are displayed in the ANOVA table. Notably, the rows have a mean square (MS) of $389 million and a sum of squares (SS) of $8.95 billion with 23 degrees of freedom (df). This results in an F-value of 0.941 and a non-significant p-value of 0.55.

The columns' SS, on the other hand, is $104 billion with 2 df, producing an F-value of 126.36 and an incredibly low p-value of 2.05E-19, which indicates that the columns differ significantly from one another. With 46 df, the error SS is $19 billion. SS is $132 billion in total.

Descriptive Statistics

Column1		*Column2*		*Column3*	

Mean	83802.79	Mean	2756.25	Mean	3254.5
Standard Error	7112.652	Standard Error	171.4525	Standard Error	186.7512
Median	81142	Median	2750	Median	3083
Mode	#N/A	Mode	3000	Mode	#N/A
Standard Deviation	34844.74	Standard Deviation	839.9421	Standard Deviation	914.8902
Sample Variance	1.21E+09	Sample Variance	705502.7	Sample Variance	837024.1
Kurtosis	-1.09718	Kurtosis	-0.81266	Kurtosis	-1.20291
Skewness	0.386522	Skewness	0.473392	Skewness	0.272019
Range	105958	Range	3000	Range	2959
Minimum	34853	Minimum	1500	Minimum	2000
Maximum	140811	Maximum	4500	Maximum	4959
Sum	2011267	Sum	66150	Sum	78108
Count	24	Count	24	Count	24

Three columns—Column1, Column2, and Column 3—each of which represents a different attribute—have summaries available in this dataset. Column 1 presents a wider range of monetary values, ranging from $34,853 to $140,811, with a mean of $83,802.79 and significant variability as indicated by its standard deviation of $34,844.74. In contrast to Column1, Column2 displays lower values, including the mean of $2,756.25, and a tighter range from $1,500 to $4,500. Its standard deviation is $839.94. Column 3 displays values that are comparable to those in Column 2, but with a little bit more variability, as seen by its mean of $3,254.5.
$914.89 is the standard deviation, and the range is $2,000 to $4,959. The statistics for each column—mean, median, mode, standard deviation, skewness, kurtosis, and range—provide information about the distribution and features of the corresponding qualities over a sample of twenty-four observations.

Correlation

	Column 1		*Column 2*
Column 1	1		
Column 2	-0.41106		1

Based on the presented correlation data, it appears that Column 1 and Column 2 are related. Given that it shows the correlation of a variable with itself, a correlation coefficient of 1 for Column 1 with itself denotes a perfect positive correlation, as would be expected. A moderately negative connection is shown by the correlation coefficient of -0.41106 between Column 1 and Column 2. According to this negative connection, values in Column 2 tend to decrease when values in Column 1 increase and vice versa. Despite its low strength, the correlation suggests a recognizable pattern in the relationship between the two variables. Understanding how changes in one variable may affect the other can be a helpful insight that, depending on the data, may help with decision-making or additional study.

Order Data Report

Introduction

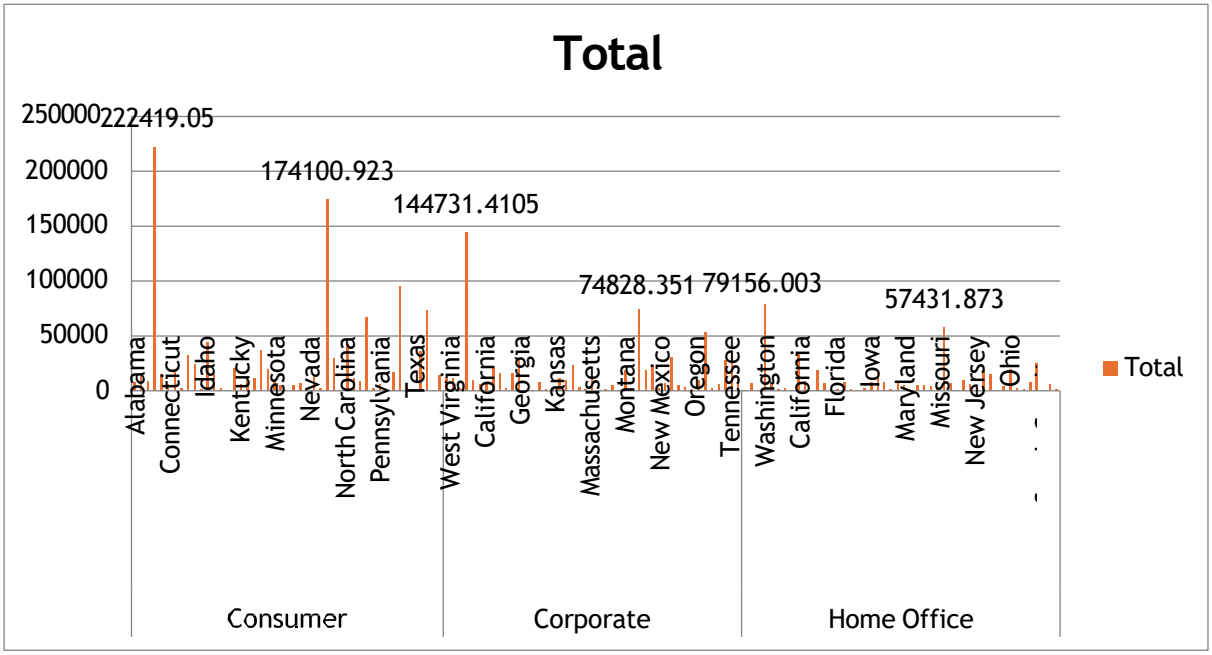
This study examines a sizable dataset that documents car sales transactions. Order ID, Order Date, Ship Date, Customer Information, Product Specifications, and Sales Figures are just a few of the many variables that are included. The primary objective of this study is to obtain useful information to support company growth and decision-making in the automotive sector. The goal of this analysis is to identify significant trends, high-performing segments, and potential areas for growth by examining sales data from multiple US states, industries, categories, and subcategories. The study's findings will be very helpful to those involved in the automotive industry, including executives, marketers, and sales managers, who want to increase revenue, enhance customer satisfaction, and optimize sales techniques.

Questionnaire

1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?
2. Find out top performing category in all the states?
3. Which segment has the most sales in the US, California, Texas, and Washington?
4. Compare total and average sales for all different segments?
5. Compare the average sales of different categories and subcategory of all the states.

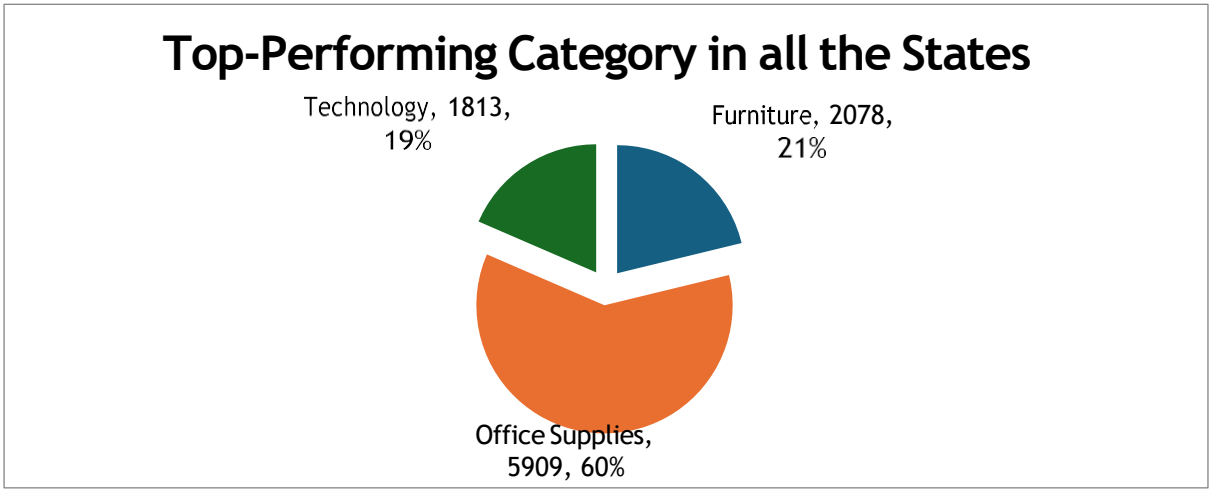
Analytics

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



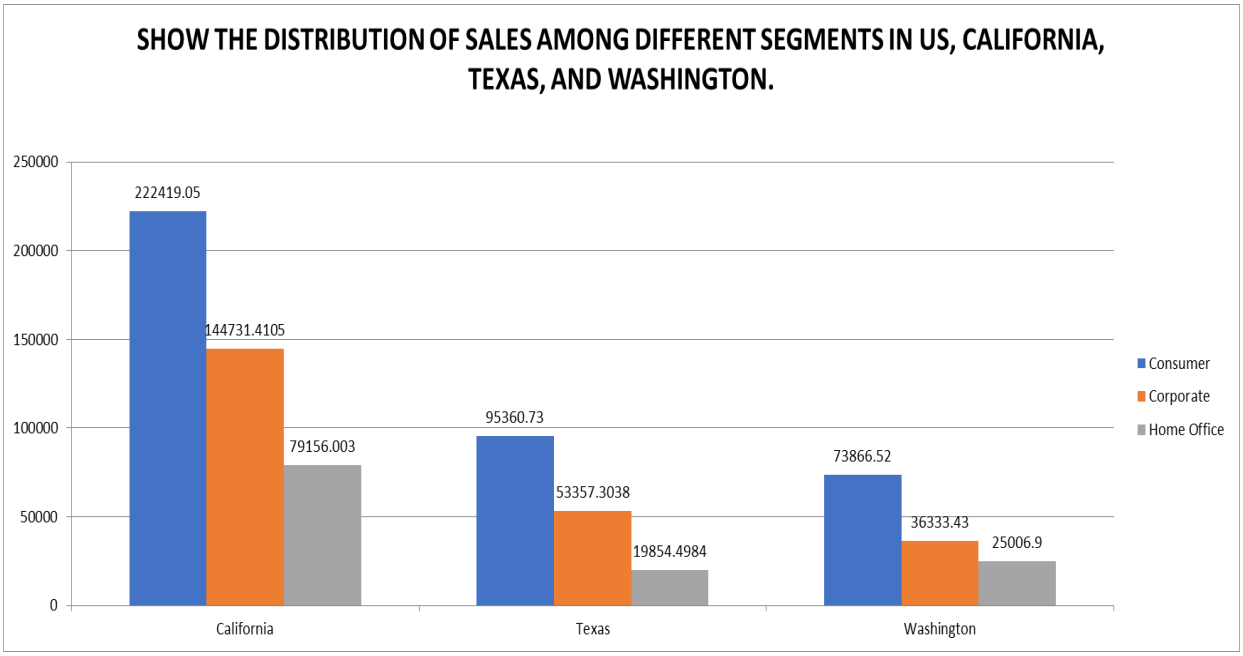
When the states were evaluated based on sales and sector, California came out on top with 222419.05. In every state, the consumer category (1148060.531) performed well.
When the states were evaluated based on sales and sector, California came out on top with 222419.05. In every state, the consumer category (1148060.531) performed well.

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



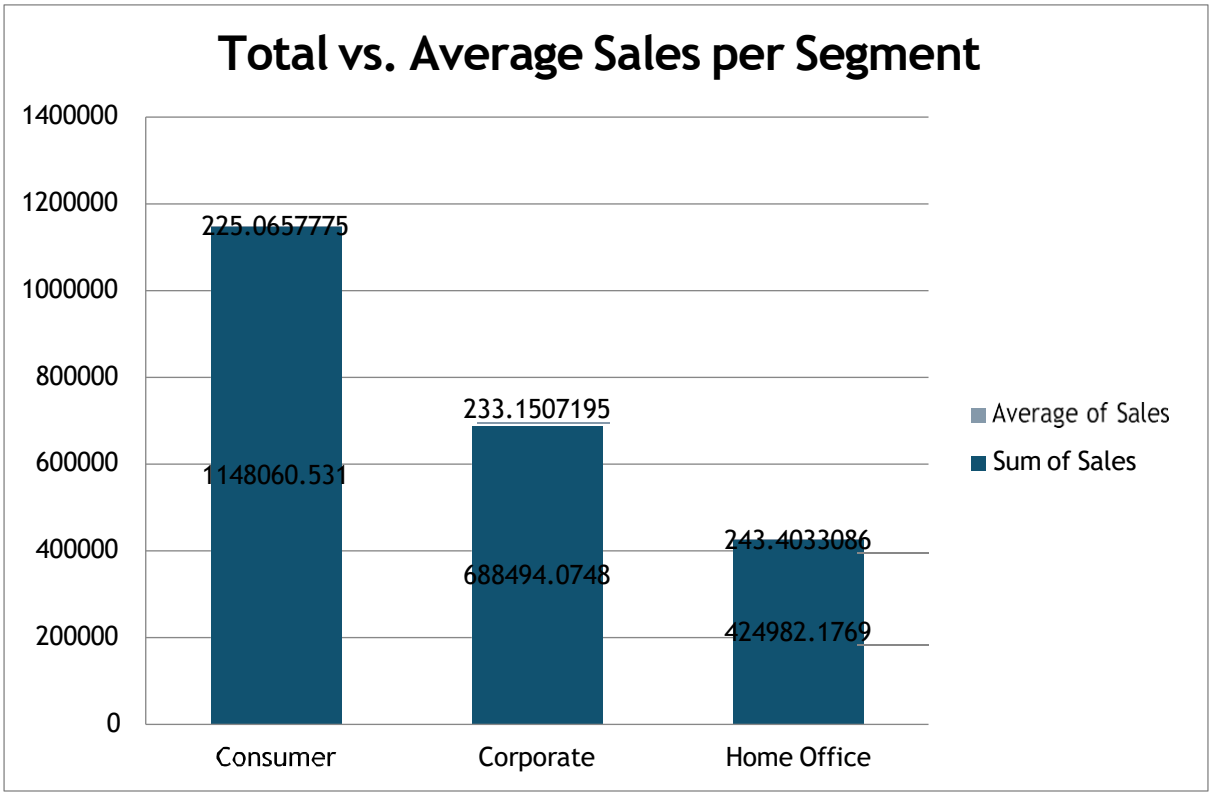
Office supplies, with 5909 total sales, are the top-performing category in all states, followed by furniture (2078) and technology (1813).

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



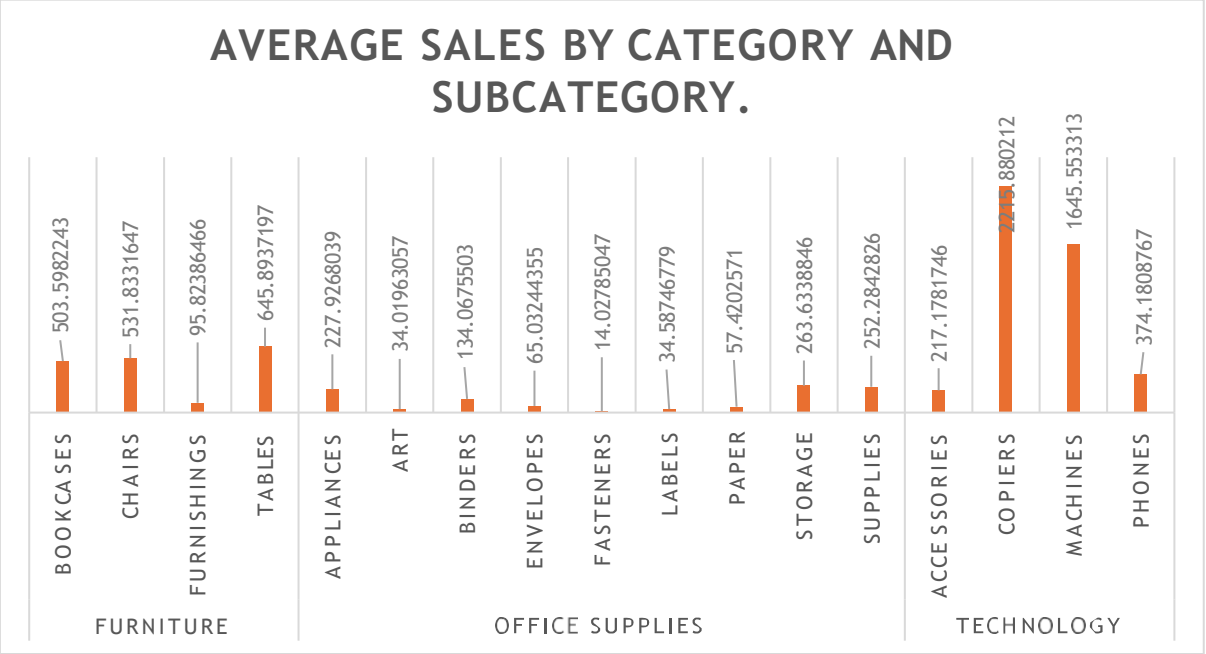
displaying the distribution's percentage using a bar chart and selecting the states to determine the total number of sales. In the consumer sector, the US, California, Texas, and Washington have the largest sales.

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



It is evident that the home office segment has total sales of 243.40 while the consumer segment has higher average sales of 1148060.531.

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



The data displays the average sales for the three categories—office supplies, technology, and furniture—each of which has several subcategories.

Conclusion and Review

Analyzing sales data in the automotive industry produces a number of noteworthy findings. California is the state with the highest sales volume, while the consumer sector performs well across the board. Office Supplies is the category that performs the best in terms of customer preferences, followed by Furniture and Technology. The consumer sector consistently leads sales in the US, particularly in California, Texas, and Washington.

Additionally, the data indicates that the average sales of the Consumer sector are higher than those of the Home Office category. All things considered, these observations provide perceptive guidance that may be applied to improve customer service, maximize sales strategies, and advance business success in the automotive industry.

Regression

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.000434
R Square	1.88E-07
Adjusted R Square	-0.0001
Standard Error	625.334
Observations	9789

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	721.1637	721.1637	0.001844	0.965747
Residual	9787	3.83E+09	391042.6		
Total	9788	3.83E+09			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	230.5863	12.63999	18.24261	3.83E-73	205.8093	255.3633	205.8093	255.3633
X Variable 1	-9.6E-05	0.002235	-0.04294	0.965747	-0.00448	0.004286	-0.00448	0.004286

Descriptive Statistics

Column1	
Mean	230.1162
Standard Error	6.320053
Median	54.384
Mode	12.96

Standard Deviation	625.3021
Sample Variance	391002.7
Kurtosis	307.3056
Skewness	13.05363
Range	22638.04
Minimum	0.444
Maximum	22638.48
Sum	2252607
Count	9789

Cookie Data Report

Introduction

Our cookie data collection consists of the following six types of cookies: sugar, chocolate chip, fortune cookie, oatmeal raisin, Snicker doodle, and white chocolate macadamia nut. We have a great deal of data on these cookies, such as the number of units sold, the costs incurred, the revenue, and the profits. To see how things change, we are not only looking at one place or historical period, but also looking at multiple countries and eras. In addition to offering information about cookies, this study attempts to shed light on customer preferences, price points, and the regions where cookies are most popular.

Questionnaire

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

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

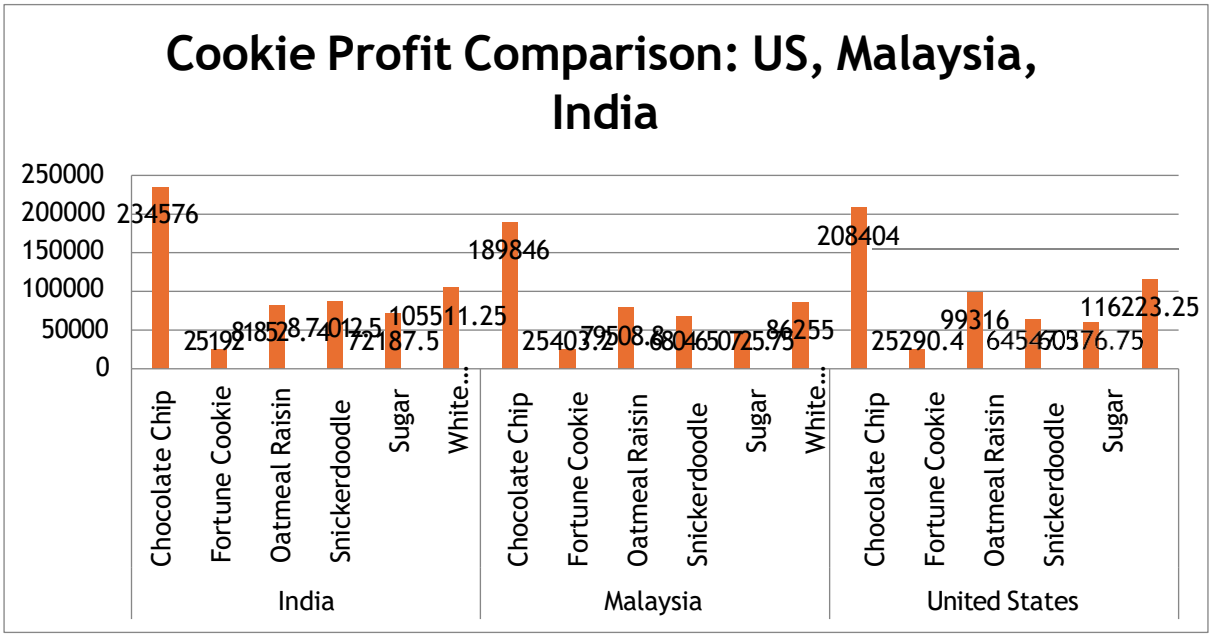
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?

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

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

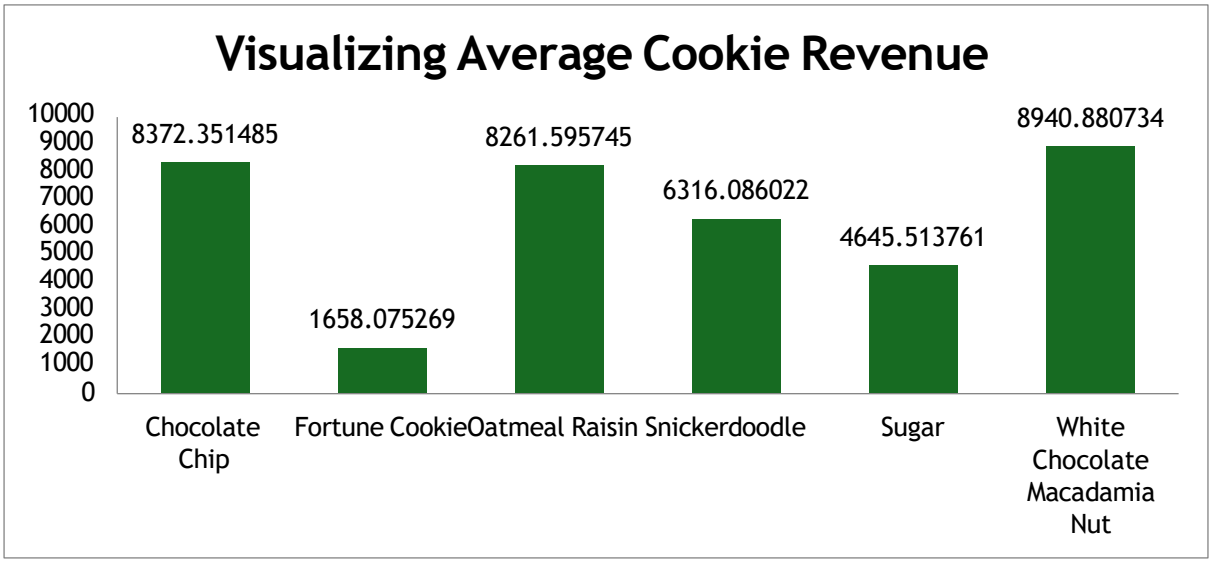
Analytics

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



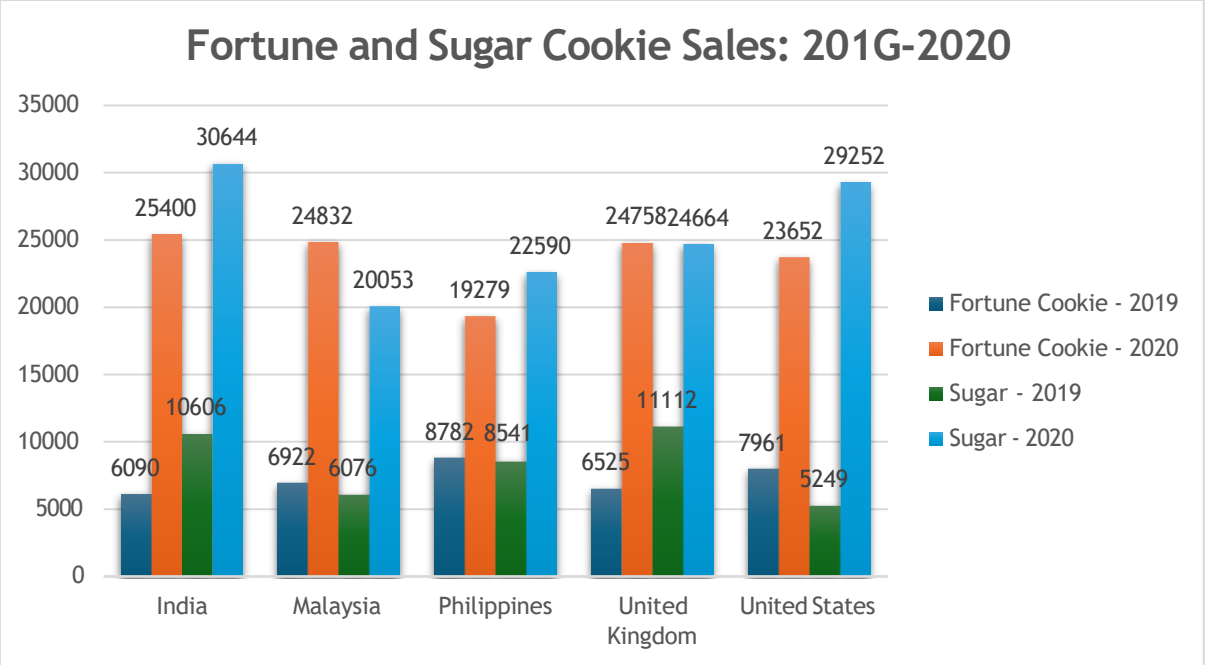
This study compares the profit margins for each type of cookie in the US, Malaysia, and India. The countries with the highest chocolate chip profits are Malaysia, the United States, and India.

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



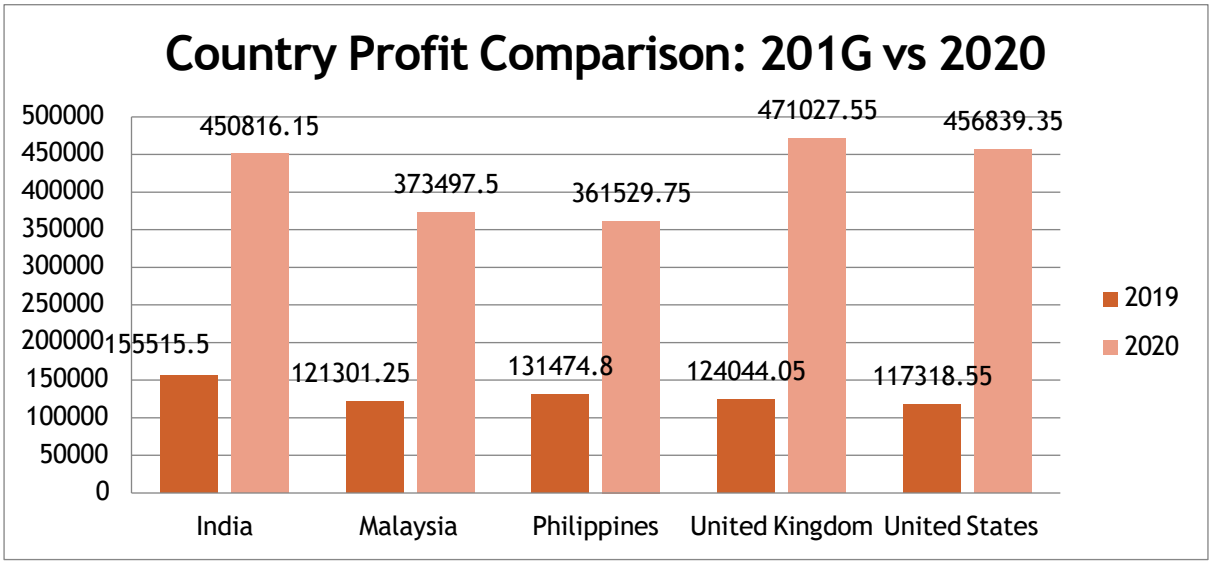
The average income generated is the goal of this investigation, and it is evident that chocolate chip comes in second place with an average revenue generate of 8940.88, followed by white chocolate macadamia nuts.

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



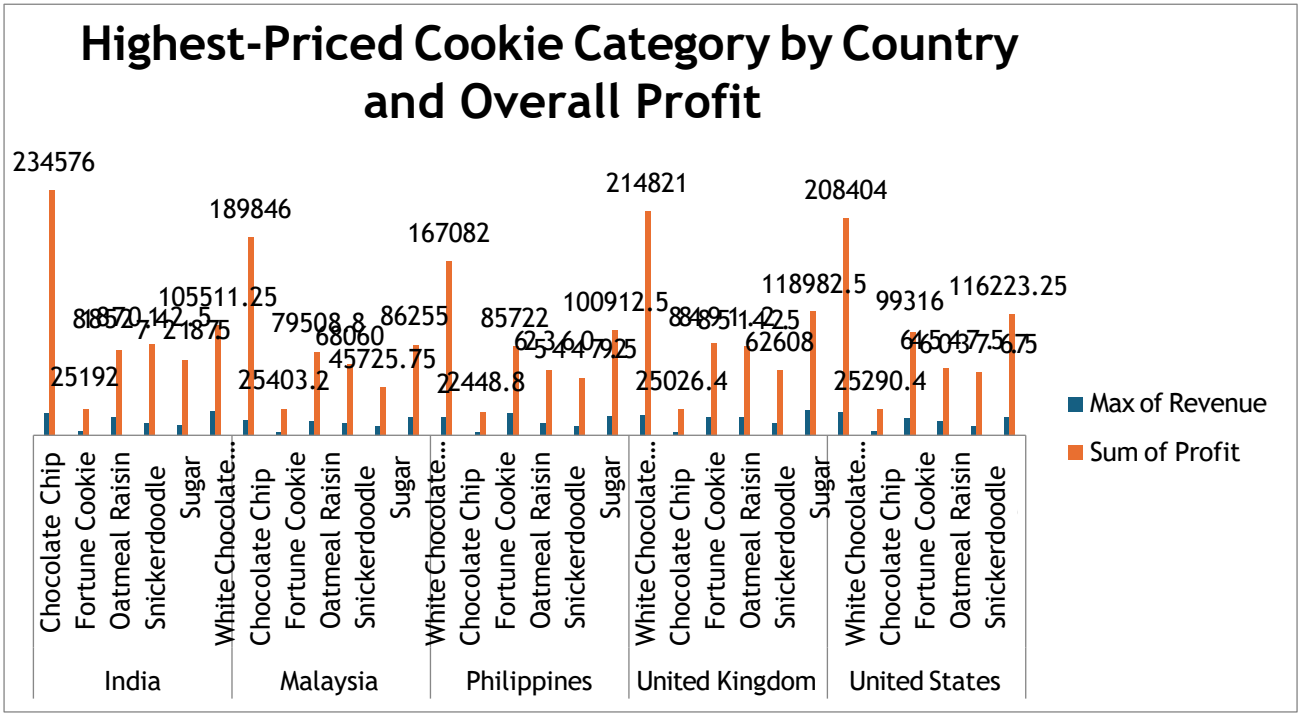
The fortune and sugar cookie sales for the years 2019 and 2020 are compared in this analysis across the different countries. With 30644 sales, India leads the world in significant sugar cookie sales for the year 2020; in 2019, the UK led the world in sugar cookie sales. Once more, India tops the fortune cookie sales charts with 25,400, followed by Malaysia; the Philippines tops the charts with 8782, followed by the US.

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



The earnings generated by the different nations in the 2019 and 2020 fiscal years are compared in this research. According to the graph, India made the most profit in 2019 with sales of 155515.5, followed by the Philippines with 131474.8, and the United Kingdom made the most profit in 2020 with sales of 471027.55, followed by the United States with 456839.35.

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



The goal of this analysis is to determine which cookie category sold for the most money, per country, and the profit made by that category. The maximum revenue for chocolate chips

(23988) and the total profit for sugar (2763364.45) are reported for India and the United Kingdom, respectively.

Conclusion and Review

The study clarified how much money certain cookie variants brought in from the US, Malaysia, and India. India was the nation with the highest revenue from chocolate chip cookies, followed by the US and Malaysia.
White chocolate macadamia nut cookies had the highest average income among the cookies, with chocolate chip cookies coming in second.

In terms of sales, the UK topped the globe in 2019 for sugar cookie sales, while India had significant sales in 2020. Fortune cookie sales were rising in Malaysia and India in both years, while the US and the Philippines also contributed significantly to these sales.

By country, the United States and the United Kingdom had the highest profits in 2020 when comparing their earnings in 2019 and 2020. In 2019, India and the Philippines yielded the highest revenues.

Although sugar cookies made the most money overall, chocolate chip cookies were the most profitable in terms of revenue.

The report's analytical information on the cookie industry assisted participants in understanding market dynamics and making informed decisions. The results were effectively explained by the use of visually appealing and easily comprehensible graphics. Nonetheless, it's critical to acknowledge the need for additional study into other factors influencing sales and profitability. Data accuracy and completeness must be ensured for reliable insights.

Regression

Regression shows.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	1
R Square	1
Adjusted R Square	1
Standard Error	9.16E-12
Observations	700

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	4.78E+09	1.59E+09	1.9E+31	0
Residual	696	5.84E-20	8.39E-23		
Total	699	4.78E+09			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.3E-11	7.3E-13	-18.0657	4.09E-60	-1.5E-11	-1.2E-11	-1.5E-11	-1.2E-11
X Variable 1	6.56E-17	8.42E-16	0.077892	0.937936	-1.6E-15	1.72E-15	-1.6E-15	1.72E-15
X Variable 2	1	8.38E-16	1.19E+15	0	1	1	1	1
X Variable 3	-1	1.72E-15	-5.8E+14	0	-1	-1	-1	-1

Anova: one factor

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	700	1926955	2752.792	4149401		
Column 2	700	2763364	3947.664	6842519		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	5E+08	1	5E+08	90.92153	6.36E-21	3.848119
Within Groups	7.68E+09	1398	5495960			
Total	8.18E+09	1399				

Anova: two factor

Anova: Two-Factor Without Replication				
SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	17250	5750	6943125
Row 2	3	21520	7173.333	10805909
Row 3	3	23490	7830	12874869
Row 4	3	12280	4093.333	3518629
Row 5	3	13890	4630	4501749
Column 1	700	4690319	6700.456	21380458
Column 2	700	1926955	2752.792	4149401
Column 3	700	2763364	3947.664	6842519

ANOVA						

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1.99E+10	699	28507277	14.75112	0	1.112595
Columns	5.74E+09	2	2.87E+09	1484.458	0	3.002161
Error	2.7E+09	1398	1932550			
Total	2.84E+10	2099				

Descriptive Statistics

Column1		*Column2*		*Column3*		*Column4*	
Mean	1608.32	Mean	6700.456	Mean	2752.792	Mean	3947.664
Standard Error	32.78652	Standard Error	174.767	Standard Error	76.99166	Standard Error	98.86874
Median	1542.5	Median	5871.5	Median	2423.6	Median	3424.5
Mode	727	Mode	8715	Mode	3450	Mode	5229
Standard Deviation	867.4498	Standard Deviation	4623.901	Standard Deviation	2037.008	Standard Deviation	2615.821
Sample Variance	752469.1	Sample Variance	21380458	Sample Variance	4149401	Sample Variance	6842519
Kurtosis	-0.31491	Kurtosis	0.464596	Kurtosis	0.810043	Kurtosis	0.338621
Skewness	0.43627	Skewness	0.867861	Skewness	0.930442	Skewness	0.840484
Range	4293	Range	23788	Range	10954.5	Range	13319
Minimum	200	Minimum	200	Minimum	40	Minimum	160
Maximum	4493	Maximum	23988	Maximum	10994.5	Maximum	13479
Sum	1125824	Sum	4690319	Sum	1926955	Sum	2763364
Count	700	Count	700	Count	700	Count	700

Correlation

	Column 1	*Column 2*	*Column 3*	*Column 4*
Column 1	1			
Column 2	0.796298	1		
Column 3	0.742604	0.992011	1	
Column 4	0.829304	0.995163	0.974818	1

Loan Data Report

Introduction

A plethora of information about loan applicants is contained in the loan dataset, including information about their income, property area, gender, marital status, degree of education, and loan amount. There is a lot of information on loan application behavior in this dataset. Examining the characteristics of loan candidates and searching for patterns in the data are the objectives of this study. To try to provide answers to some of the inquiries regarding the loan amounts, applicant demographics, and educational backgrounds, we employ pivot tables and charts.
To make educated decisions, expedite the lending process, and tailor services to meet the diverse needs of their clientele, financial institutions need to be aware of the nuances of loan applications. Finding useful information that can guide strategic decisions and increase the efficacy of loan management initiatives is the aim of this research.

Questionnaire

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

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

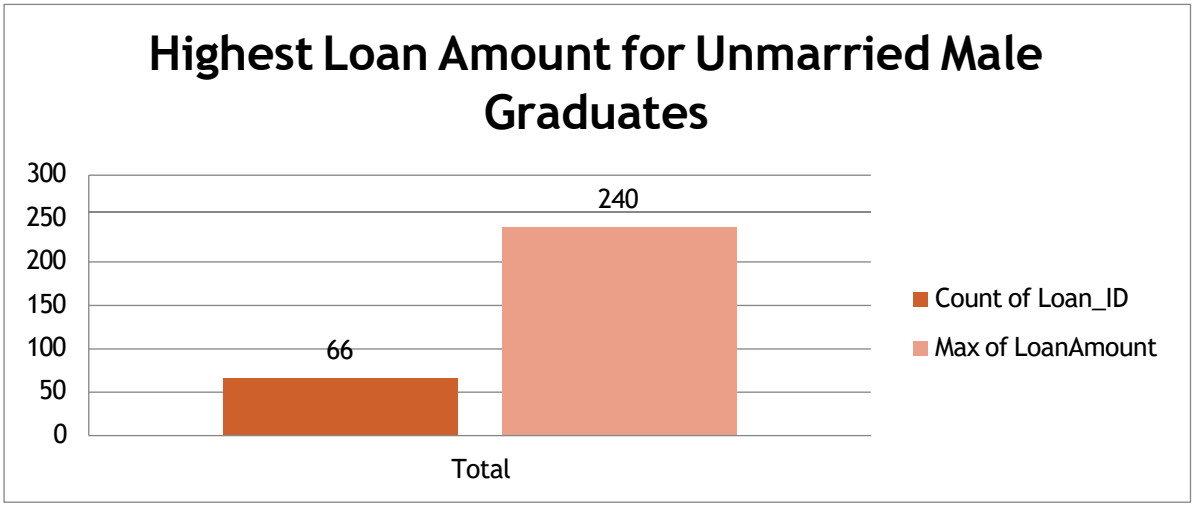
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?

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

5. How many male and female who are not married applied for Loan? Compare Urban, Semi-urban and rural based on amount.

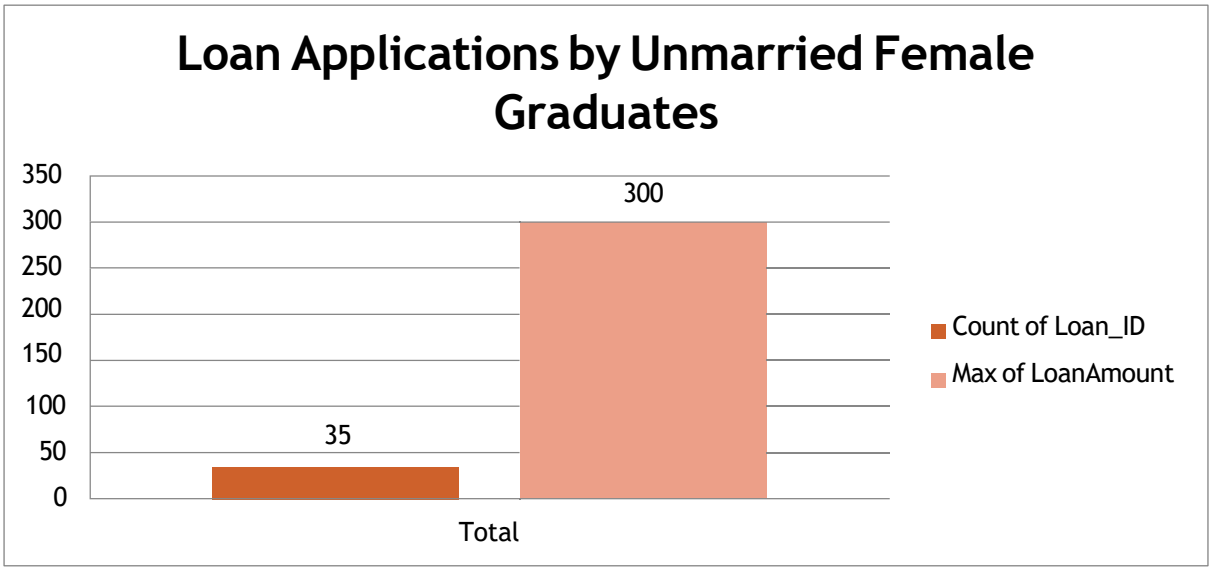
Analytics

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



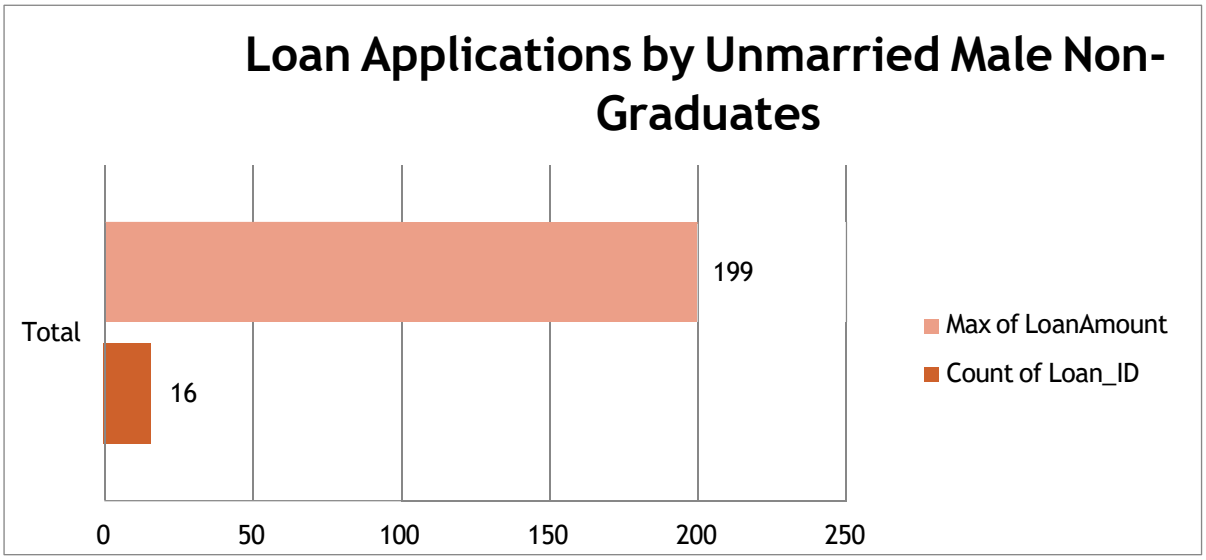
According to this data, the greatest number of male graduates who are single asked for loans. As of now, there have been 66 total loan applications, with a maximum loan amount of 240.

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



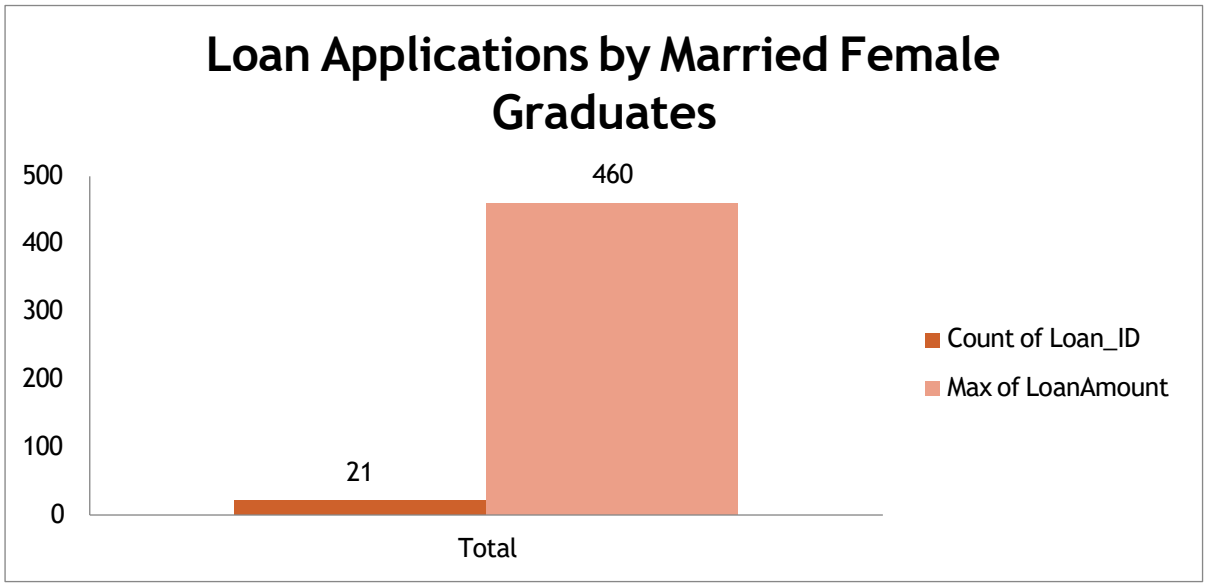
Based on the available data, the majority of unmarried female graduates applied for loans. There have been 35 loan applications submitted thus far, with a $300 maximum loan amount.

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



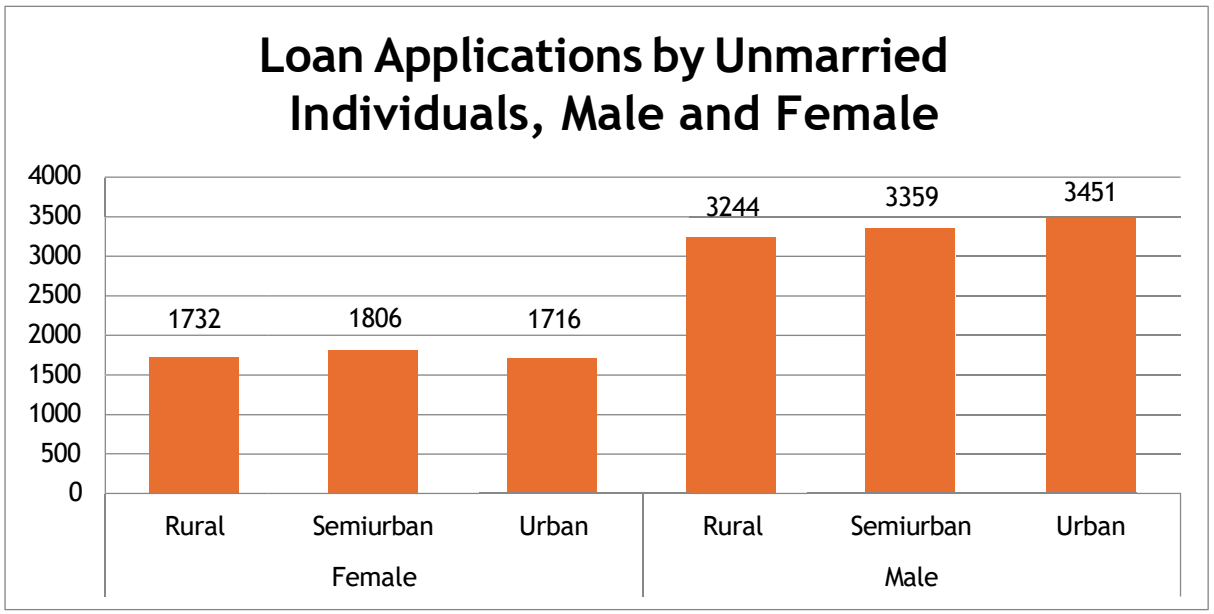
This study shows how many single male undergrads applied for loans and how much of them were turned down. There have been 16 loan applications submitted thus far, with a maximum loan amount of 199.

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



Based on the available data, the majority of unmarried female graduates applied for loans. There have been 21 loan applications submitted thus far, with a $460 maximum loan amount.

5. How many males and female who are not married applied for Loan? Compare Urban, Semi-urban and rural based on amount.



This study examines loan applications from single men and women in rural, semi-urban, and urban areas; the number of applications from men is significantly higher than that from women. Men's (1744), semi-urban (3359), and urban (3451) loan counts are as follows: women's (1732), semi-urban (1806), and urban (1716).

Conclusion and Review

The data demonstrates stark variations in loan applications according to gender. Male graduates alone made up the majority of the application pool, followed by female graduates alone. Married female graduates and single male graduates also requested loans, though in smaller proportions. Remarkably, there were much more men than women in rural, semi- urban, and metropolitan areas.

The study provides useful data on borrower demographics and effectively illustrates gender-based trends in loan applications. It is suggested that additional research be done on the factors influencing loan decisions and that the data be presented more visually. Overall, the work offers a foundation for understanding loan dynamics; however, more research is necessary.

Regression

Regression shows the stats

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.531078663
R Square	0.282044546
Adjusted R Square	0.274487121
Standard Error	50.85033905
Observations	289

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	289502.8035	96500.93	37.32019	2.25609E-20
Residual	285	736940.7397	2585.757		
Total	288	1026443.543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	66.690952	16.26833015	4.099434	5.41E-05	34.66963005	98.71227396	34.66963	98.71227
X Variable 1	0.095771273	0.045649816	2.097955	0.03679	0.005917708	0.185624838	0.005918	0.185625
X Variable	0.005807	0.000627	9.2501	5.49E-	0.004571	0.007043	0.0045	0.0070

2	787	861	22	18	955	619	72	44
X Variable 3	0.006772	0.001264	5.3549	1.76E-07	0.004283	0.009262	0.0042	0.0092
	797	765	83		331	263	83	62

Anova: one factor

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	289	39533	136.7924	3564.04		
Column 2	289	99032	342.6713	4310.645		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6124794	1	6124794	1555.565	8.4E-166	3.857654
Within Groups	2267909	576	3937.343			
Total	8392703	577				

Anova: two factor

Anova: Two-Factor Without Replication				
SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	2	470	235	31250
Row 2	2	486	243	27378
Row 3	2	568	284	11552
Row 4	2	438	219	39762
Row 5	2	512	256	21632
Row 286	2	473	236.5	30504.5
Row 287	2	475	237.5	30012.5
Row 288	2	518	259	20402
Row 289	2	278	139	3362
Column 1	289	39533	136.7924	3564.04
Column 2	289	99032	342.6713	4310.645

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1264619	288	4391.038	1.260472	0.024978	1.214301
Columns	6124794	1	6124794	1758.156	1.2E-124	3.87395

Error	1003290	288	3483.647			
Total	8392703	577				

Descriptive Statistics

Column1		*Column2*		*Column3*		*Column4*	
Mean	342.6713	Mean	4637.353	Mean	1528.263	Mean	136.7924
Standard Error	3.862088	Standard Error	281.8049	Standard Error	139.8588	Standard Error	3.51174
Median	360	Median	3833	Median	879	Median	126
Mode	360	Mode	5000	Mode	0	Mode	150
Standard Deviation	65.6555	Standard Deviation	4790.684	Standard Deviation	2377.599	Standard Deviation	59.69958
Sample Variance	4310.645	Sample Variance	22950653	Sample Variance	5652978	Sample Variance	3564.04
Kurtosis	8.62994	Kurtosis	141.612	Kurtosis	32.96701	Kurtosis	5.739804
Skewness	-2.64147	Skewness	10.41123	Skewness	4.510775	Skewness	1.780616
Range	474	Range	72529	Range	24000	Range	432
Minimum	6	Minimum	0	Minimum	0	Minimum	28
Maximum	480	Maximum	72529	Maximum	24000	Maximum	460
Sum	99032	Sum	1340195	Sum	441668	Sum	39533
Count	289	Count	289	Count	289	Count	289

Correlation

	Column 1	*Column 2*	*Column 3*
Column 1	1		
Column 2	-0.08435	1	
Column 3	0.445695	0.230355	1

Shop Sales Data Report

Introduction

With a focus on product trends among sales reps and sales performance analysis, this paper looks at a sizable sales dataset. Features in the collection include product specs, sales numbers, earnings, and information about salespeople. The primary objective of this research is to gather data that will help guide the development of sales strategies and enhance business performance. By examining sales data over a specific time period and comparing product performance, the study aims to identify top-performing salespeople, examine product popularity, and understand sales patterns. The analysis's findings will be very helpful to CEOs, marketing experts, and sales managers who wish to increase revenue, enhance sales strategies, and grow their businesses. The purpose of this study is to provide useful information that will support decision-making and improve the overall performance of the company.

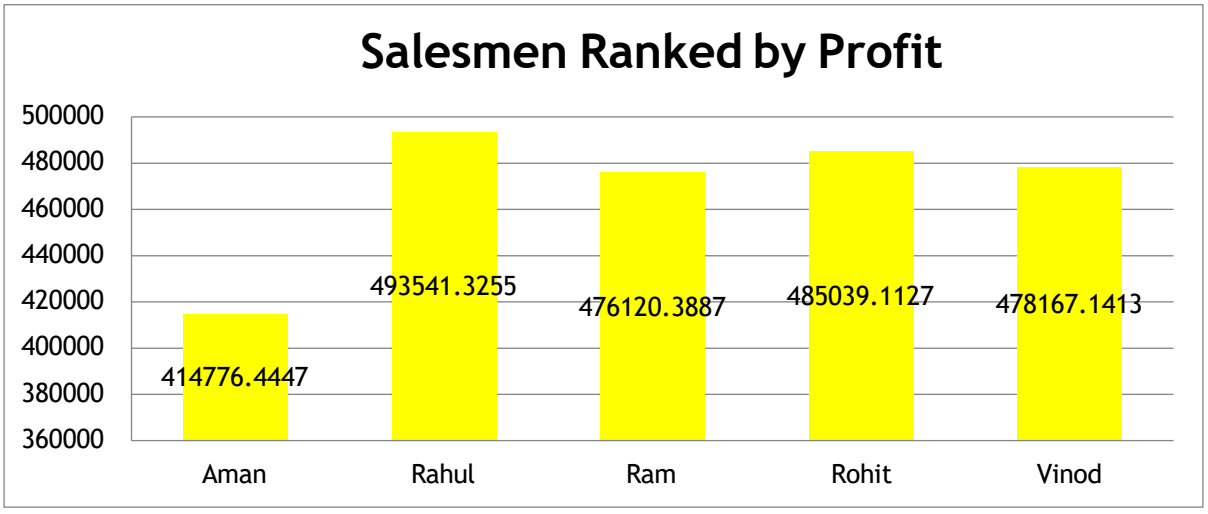
Questionnaires

1. Compare all the salesmen based on 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

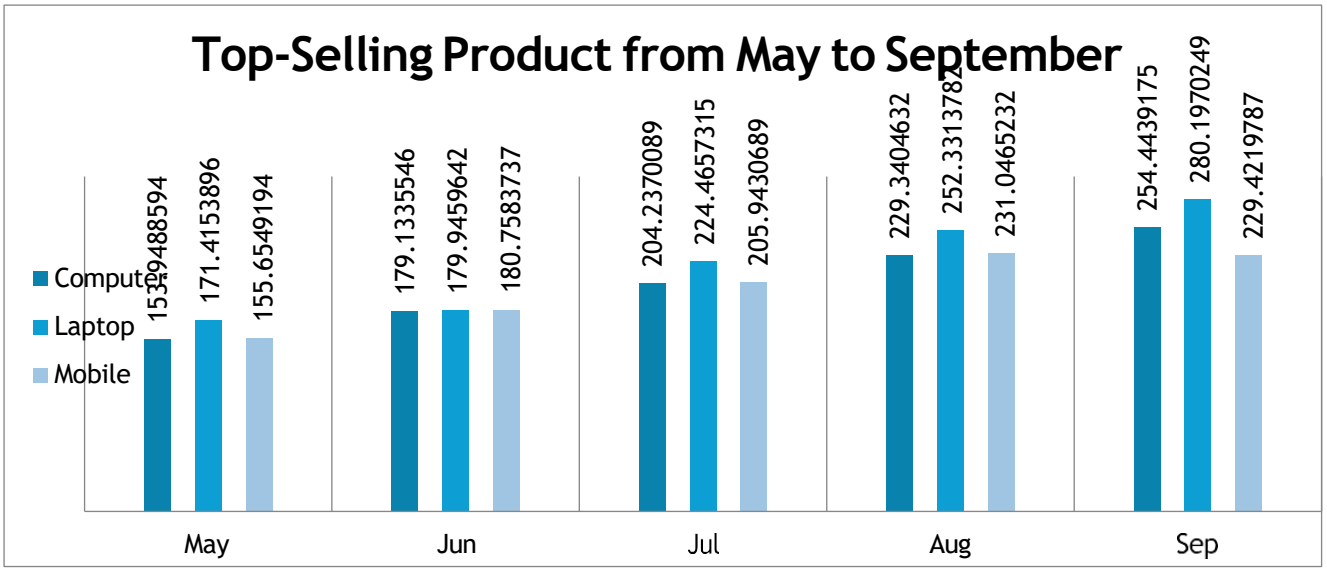
1. Compare all the salesmen on the basis of profit earn.

Rahul had the most profit earned, valued at 493541.3255, when all of the salesmen are compared based on profit made, as shown by the line chart.



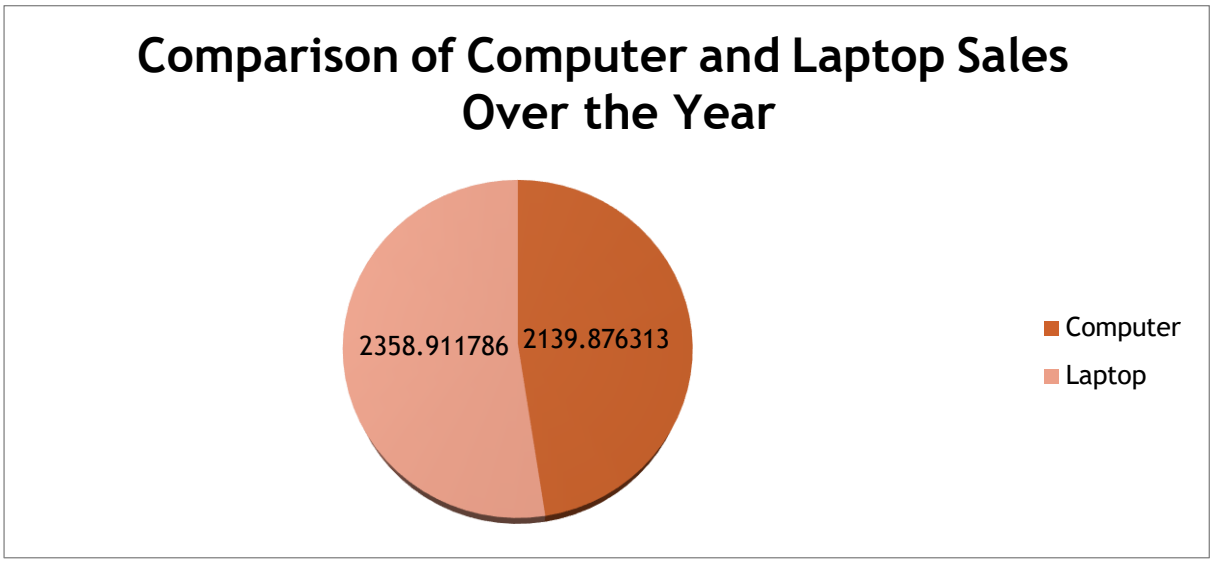
2. Find out most sold product over the period of May-September.

To find out which product sold the most from May through September, we would have to look at the sales data for that entire period. Compiling all transactions within this time period and adding the quantity sold for each product, the laptop is the most sold product from May to September, with September having the most sales, reaching 280.1970249.



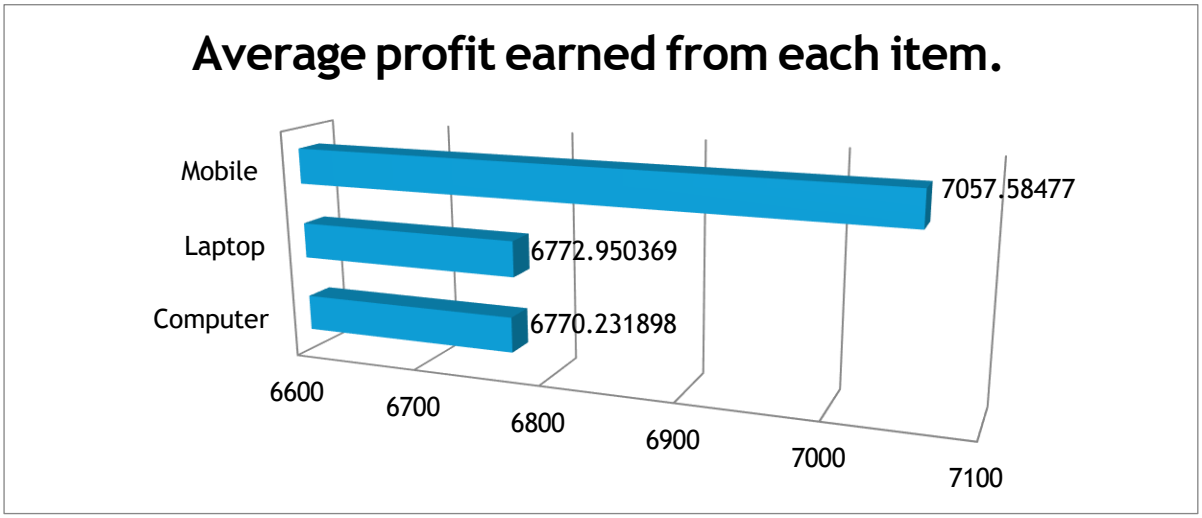
3. Find out which of the two product sold the most over the year Computer or Laptop?

Over the course of the year, the laptop and the computer were the two most popular items, with the laptop having a greater sales quantity at 2358.911786 and the computer at 2139.876313.



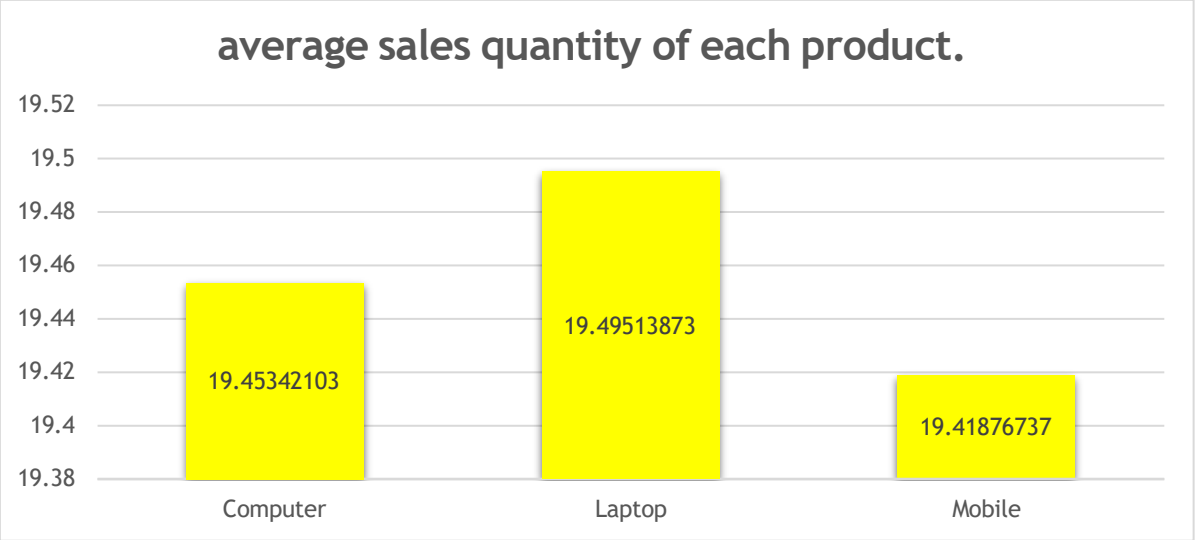
4. Which item yield most average profit?

When compared to a laptop and computer, the mobile device has the largest average profit produced (7057.58477).



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

Based on the analysis, the average sales quantity of laptops (19.49513873) is higher than that of mobile phones (19.41876737) and computers (19.45342103).



Conclusion and Review:

The analysis provides valuable insights into sales effectiveness and product trends among sales representatives. Rahul wins by outperforming all other salespeople and generating the largest profit. In addition, the laptop is the best-selling item from May to September, with the highest sales occurring in September. Over the course of the year, laptops outsell PCs in terms of units sold. Additionally, mobile phones have the highest average profit among PCs,

laptops, and smartphones. Finally, laptops do better than PCs and mobile devices in terms of average sales quantity.
The study effectively highlights sales performance and product trends while providing useful data for enhancing sales strategy. Understanding enduring patterns and popular products is aided by visualizations. However, a deeper comprehension of the factors influencing product preferences and sales fluctuations could enhance the analysis. When all is said and done, the research offers helpful knowledge for improving sales strategies and raising revenue.

Regression

The outcome variable and the amount earned have a strong positive association, as indicated by the regression model with a significant p-value. The model's strong R-squared score of 0.910 attests to its predictive accuracy.

SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.954076972
R Square	0.910262868
Adjusted R Square	0.909998936
Standard Error	630.0595983
Observations	342

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.37E+09	1.37E+09	3448.844	4.6E-180
Residual	340	1.35E+08	396975.1		
Total	341	1.5E+09			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	
Intercept	2068.993161	88.47952	23.38387	9.14E-73	1894.957	2243.029	1894.957	2
X Variable 1	246.4655683	4.196812	58.72686	4.6E-180	238.2106	254.7206	238.2106	2

Correlation

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

	Column 1	*Column 2*
Column 1	1	
Column 2	0.954077	1

Anova (Single Factor)

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

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	342	6654.271	19.45693	66.0952		
Column 2	342	2347644	6864.457	4410782		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8.01E+09	1	8.01E+09	3632.879	2.1E-275	3.85513
Within Groups	1.5E+09	682	2205424			
Total	9.52E+09	683				

Anova two factor

With degrees of freedom (df) values of 10 for rows and columns, respectively, the ANOVA results show considerable variation between them ($p < 0.001$). There is zero degree of freedom for the error term.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	2	1003	501.5	497004.5
Row 2	2	7804	3902	30388808
Row 3	2	3005	1502.5	4485013
Row 4	2	2304	1152	2635808
Row 5	2	7003	3501.5	24479005
Row 339	2	10252.82	5126.411	51884342
Row 340	2	10272.93	5136.467	52087770
Row 341	2	10293.05	5146.523	52291595
Row 342	2	10313.16	5156.58	52495819
Column 1	342	6654.271	19.45693	66.0952
Column 2	342	2347644	6864.457	4410782
ANOVA				

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*

Rows	7.58E+08	341	2221714	1.014883	0.445792	1.195299
Columns	8.01E+09	1	8.01E+09	3659.913	2.1E-184	3.868873
Error	7.46E+08	341	2189134			
Total	9.52E+09	683				

Descriptive Statistics:

Column1		*Column2*	
Mean	19.45693	Mean	6864.457
Standard Error	0.439614	Standard Error	113.5651
Median	19.45693	Median	6984.647
Mode	3	Mode	1000
Standard Deviation	8.129896	Standard Deviation	2100.186
Sample Variance	66.0952	Sample Variance	4410782
Kurtosis	-0.99883	Kurtosis	-0.5078
Skewness	-0.09948	Skewness	-0.36449
Range	30.30852	Range	9279.851
Minimum	3	Minimum	1000
Maximum	33.30852	Maximum	10279.85
Sum	6654.271	Sum	2347644
Count	342	Count	342

Sales Data Sample Report

Introduction

This report analyzes a sizable sales dataset that includes variables like ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and SALES. It aims to make inferences that will guide sales strategies and enhance organizational effectiveness. The target audience includes executives, marketers, and sales managers who want to improve sales procedures and boost revenue. Comparing the sales of classic and vintage cars, calculating average sales, identifying the best-selling items, examining the profit margin by country for specific product lines, comparing sales over time, and examining countries based on the volume of deals are all examples of significant research. With these evaluations, the research hopes to provide useful guidance for increasing sales growth and improving overall business outcomes.

The project's scope involves examining a substantial sales dataset to extract valuable insights that could enhance product offerings, direct sales strategies, and enhance overall business performance. Researchers and analysts searching for information on market trends and sales dynamics will find the project useful.

Questionnaire

1. Comparison of sales between Vintage cars and Classic cars across all countries.

2. Determination of the average sales of all products and identification of the highest-selling product.

3. Assessment of the country yielding the most profit for Motorcycles, Trucks, and Buses.

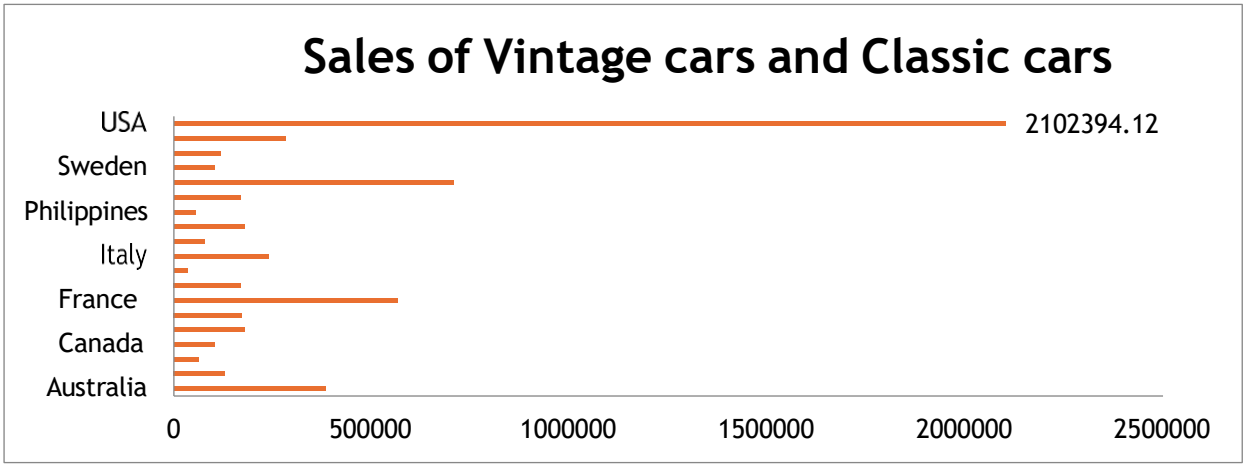
4. Comparison of sales for all items across the years 2004 and 2005.

5. Comparative analysis of all countries based on deal size.

Analytics

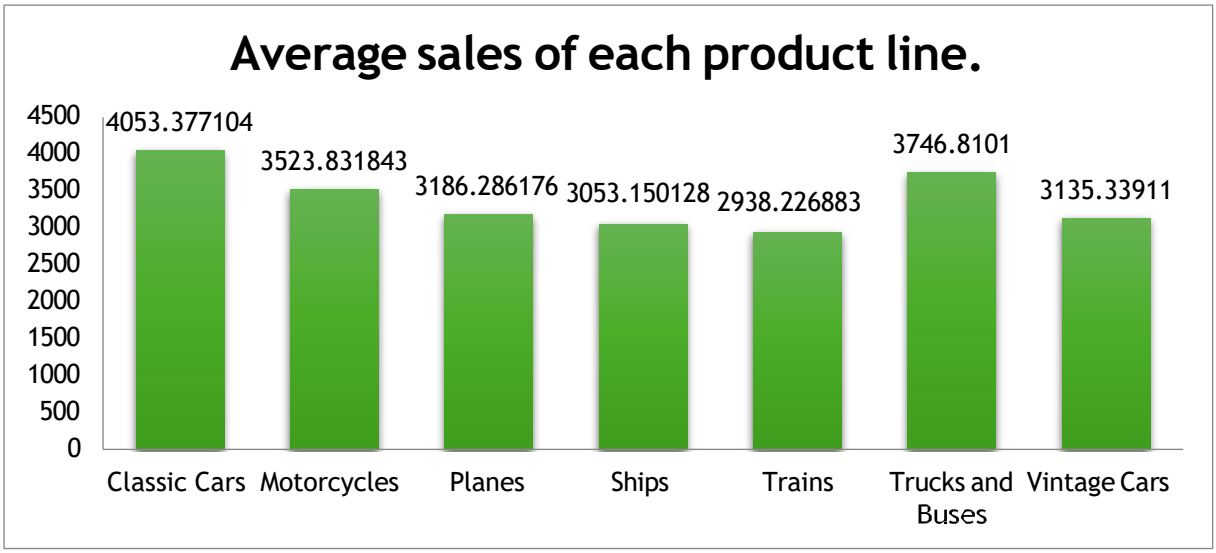
1. Comparison of sales between Vintage cars and Classic cars across all countries.

This examination Compare the sales of classic and vintage automobiles across all nations. The USA (2102394.02) has the largest sales, with Australia, Spain, and France following.

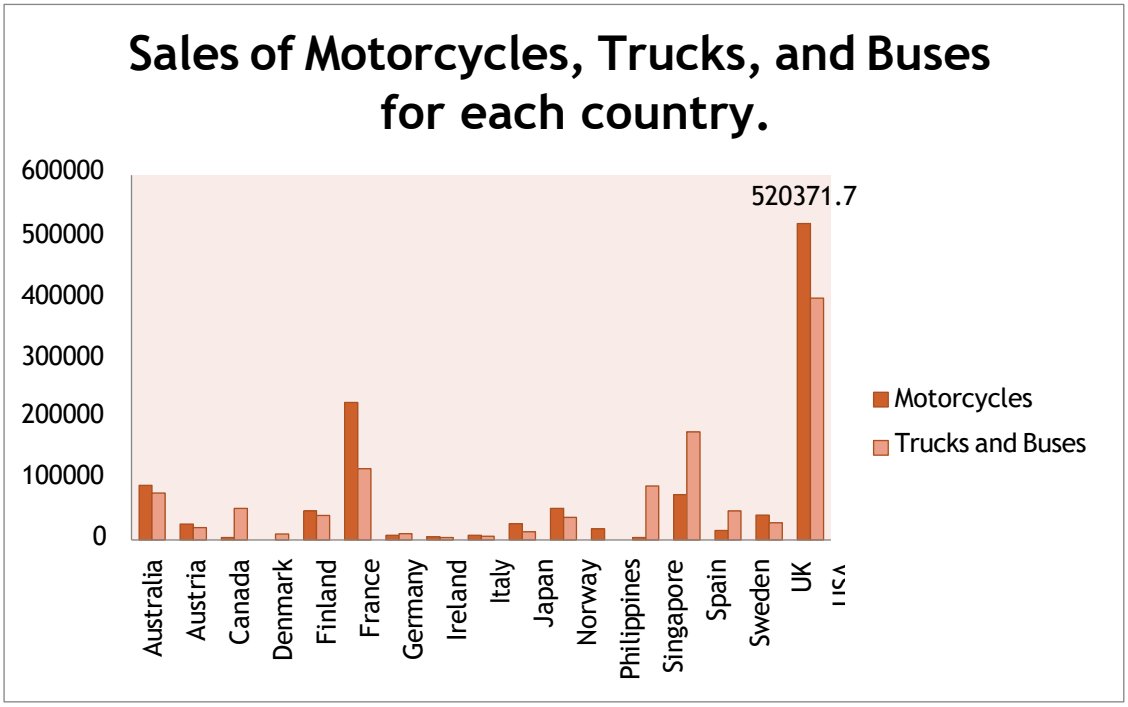


2. Determination of the average sales of all products and identification of the highest-selling product.

The two objectives of this inquiry are to determine the top-selling product and the average sales of each product. Furthermore, the graph indicates that Classic Cars have the highest average sales, at 405.377104, followed by Trucks & Buses and Motorcycles.

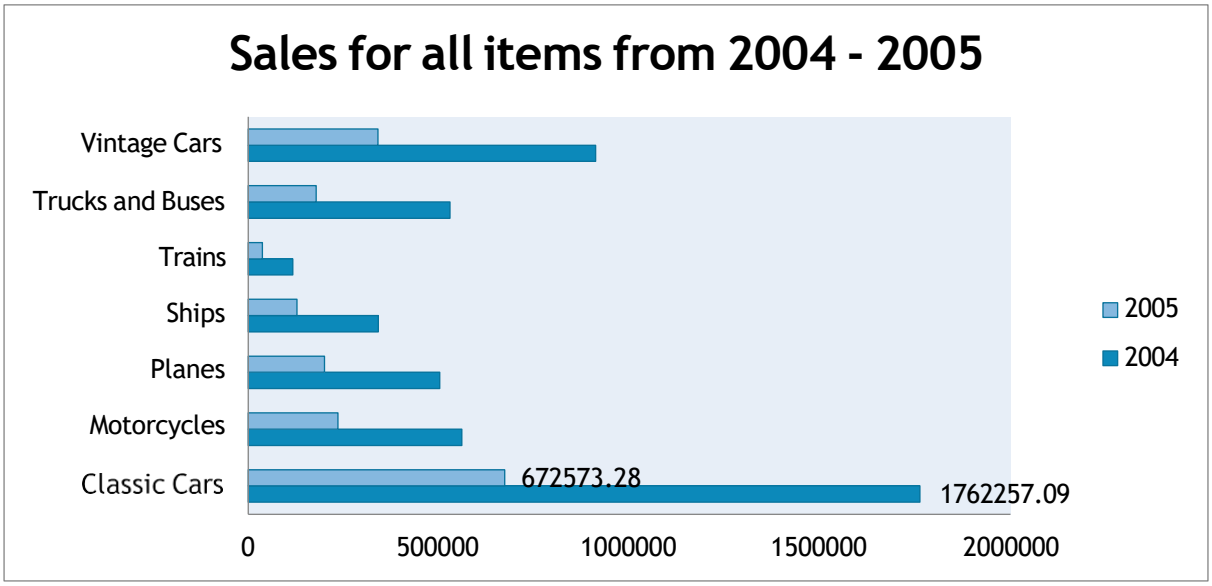


Assessment of the country yielding the most profit for Motorcycles, Trucks, and Buses.Finding the country that generates the most revenue from trucks, buses, and motorbikes is the aim of this investigation. A bar graph shows that the USA leads the globe in motorcycle sales (520371.7), followed by truck and bus sales (397842.42) and motorcycle sales (France and Spain).



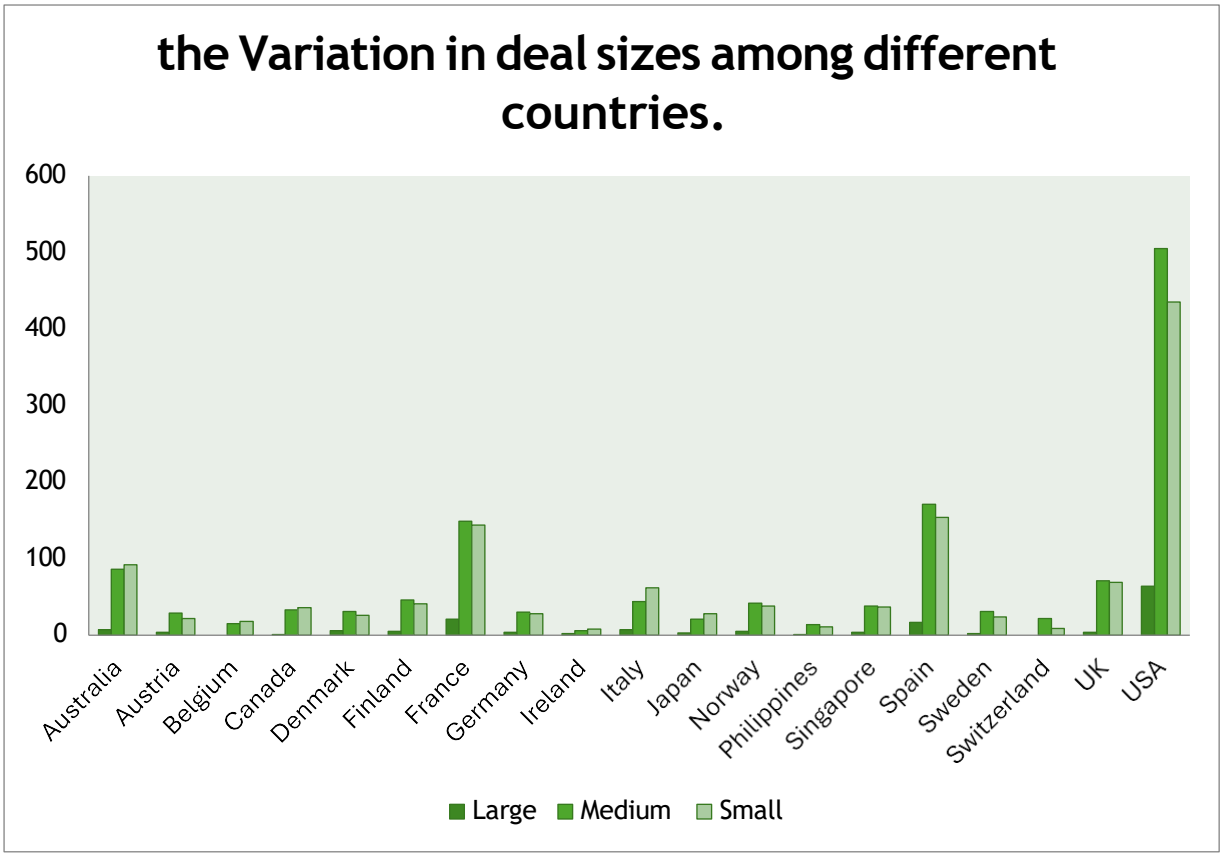
4. Comparison of sales for all items across the years 2004 and 2005.

Comparing the sales of each item in 2004 and 2005 is the aim of this investigation. With the exception of historic cars, which had the highest sales of any category in both years—1762257.09 in 2004 and 672573.28 in 2005—the line chart demonstrates how quickly the sales of every item are changing.



5. Comparative analysis of all countries based on deal size.

The aim of this study is to ascertain the distribution of deal sizes across the different countries. The bar chart, which values massive agreements at 64, medium deals at 505, and small deals at 435, also shows how much larger deals are made in the USA than in any other nation.



Conclusion and Review

The research offers insightful information on profitability and sales trends by country and category. In terms of trucks, buses, motorcycles, and vintage and classic cars, the USA leads the world market. The best-selling item, which contributes significantly to overall sales income, is classic cars. Furthermore, the USA is incredibly profitable, particularly in the truck, bus, and motorcycle sectors. Throughout 2004 and 2005, sales of classic vehicles were robust, indicating a sustained demand for this product category. Furthermore, compared to other nations, the USA exhibits far bigger transaction sizes, indicating its superiority in terms of sales volume.
Even if the analysis presents important insights in a visually appealing manner, a deeper examination of the factors that influence sales volatility and deal size variations could produce more informative outcomes. Overall, the study offers insightful information that may be used to improve sales tactics and quicken company expansion.

Regression

Regression shows...

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.877178
R Square	0.769441
Adjusted R	0.766629

Square	
Standard Error	896.6688
Observations	250

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	6.6E+08	2.2E+08	273.6567	4.62E-78
Residual	246	1.98E+08	804014.9		
Total	249	8.58E+08			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5271.93	322.9166	-16.326	4.32E-41	-5907.96	-4635.9	-5907.96	-4635.9
X Variable 1	103.0809	6.001152	17.17685	5.42E-44	91.26071	114.9011	91.26071	114.9011
X Variable 2	12.81807	1.661734	7.713668	3.04E-13	9.545024	16.09111	9.545024	16.09111
X Variable 3	47.42944	3.350938	14.15408	1.13E-33	40.82925	54.02963	40.82925	54.02963

Anova: one factor

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	250	903280.9	3613.123	3445221		
Column 2	250	25534	102.136	1664.552		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.54E+09	1	1.54E+09	894.0704	3.1E-113	3.860199
Within Groups	8.58E+08	498	1723443			
Total	2.4E+09	499				

Anova: two factor

Anova: Two-Factor Without Replication				
SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	4097.66	1365.887	5069957
Row 2	3	2451.12	817.04	1725170
Row 3	3	1566	522	648687
Row 4	3	5095.24	1698.413	7507173
Row 5	3	5140.39	1713.463	7650609
Row 248	3	4386.35	1462.117	5944534
Row 249	3	2261.6	753.8667	1546167
Row 250	3	4176.72	1392.24	5420980
Column 1	250	903280.9	3613.123	3445221
Column 2	250	25534	102.136	1664.552
Column 3	250	8659	34.636	89.69428
ANOVA				
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	2.95E+08	249	1182944	1.044989	0.33951	1.194432
Columns	2.09E+09	2	1.05E+09	925.2361	1.9E-168	3.013826
Error	5.64E+08	498	1132016			
Total	2.95E+09	749				

Descriptive Statistics

Column1		*Column2*		*Column3*		*Column4*	
			3613.123				84.45296
Mean	34.636	Mean		Mean	102.136	Mean	
Standard Error	0.59898	Standard Error	117.392	Standard Error	2.58035	Standard Error	1.27945
Median	34	Median	3263.96	Median	99	Median	100
Mode	29	Mode	#N/A	Mode	118	Mode	100

Standard Deviation	9.470706	Standard Deviation	1856.131	Standard Deviation	40.79892	Standard Deviation	20.22993
Sample Variance	89.69428	Sample Variance	3445221	Sample Variance	1664.552	Sample Variance	409.2499
Kurtosis	-0.64676	Kurtosis	-1.12705	Kurtosis	-0.19836	Kurtosis	-0.40344
Skewness	0.256745	Skewness	1.013489	Skewness	0.517104	Skewness	-0.96781
Range	51	Range	10626.85	Range	181	Range	73.12
Minimum	15	Minimum	652.35	Minimum	33	Minimum	26.88
Maximum	66	Maximum	11279.2	Maximum	214	Maximum	100
Sum	8659	Sum	903280.9	Sum	25534	Sum	21113.24
Count	250	Count	250	Count	250	Count	250

Correlation

	Column 1	*Column 2*	*Column 3*
Column 1	1		
Column 2	0.513951	1	
Column 3	-0.01254	0.663973	1

Store Dataset Report

Introduction

Retail store sales data is included in this collection. specifications like gender, age, product specifications (category, SKU), transaction details (order ID, status), and shipping information are all included. Our mission is to assist you in comprehending the ways in which your clients engage with your offerings. We search your data for correlations, patterns, and preferences. You may raise consumer satisfaction, manage inventory, and enhance marketing with these information.

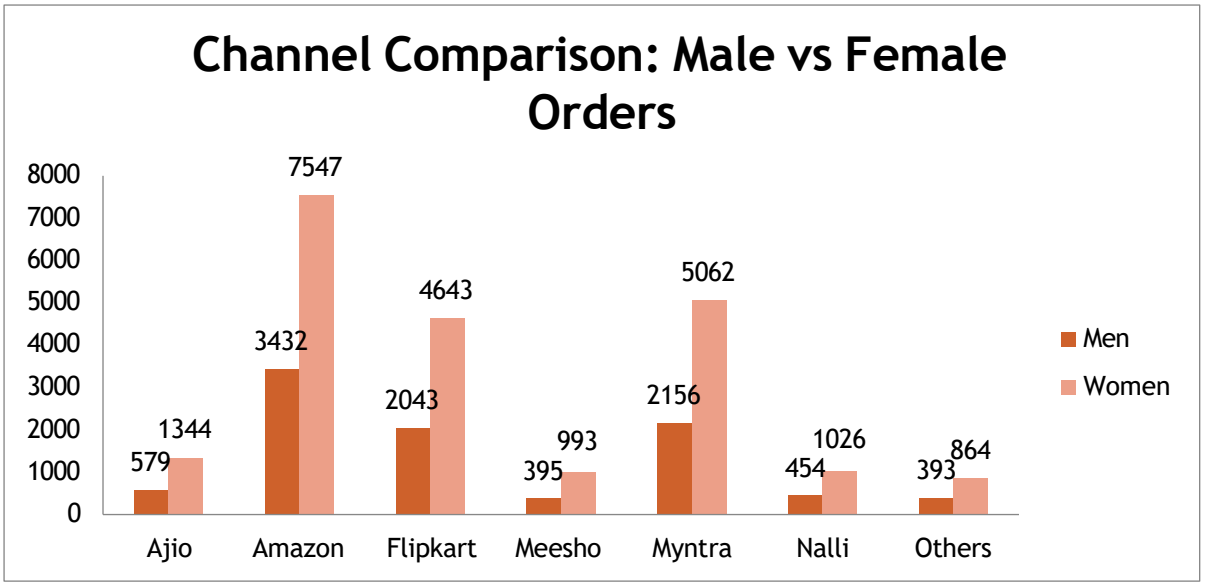
Questionnaire

1. Compare various channels based on how many male customers order and female customer order.
2. Compare all the categories of order where amount is less than 1500 and greater than 5000.
3. How many Customers are there whose age is 30 and above and state is Delhi.
4. Which of the following state perform better than other, Delhi, Tamil Nadu, Maharashtra, Rajasthan.
5. Which city performed better than all other cities based on highest order placed.
6. Compare various categories of items based on most quantity sold and show which gender buys the most category.

Analytics

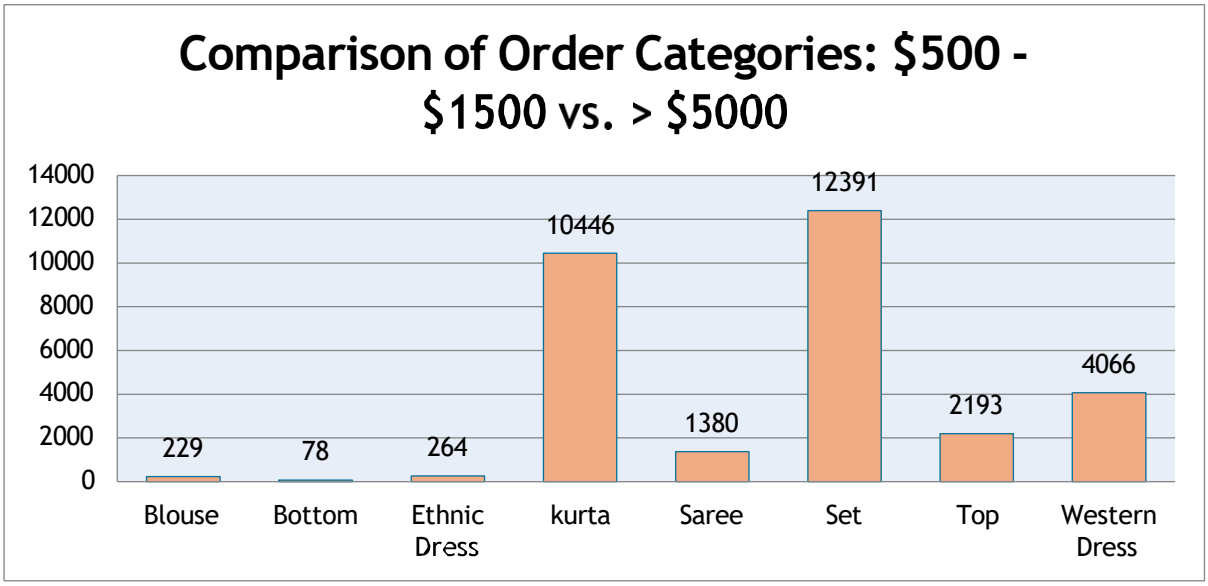
1. Compare various channels based on how many male customers order and female customer order?

Amazon leads the market in terms of sales for both men and women, followed by Myntra and Flipkart. Amazon sold approximately 3432 units in the men's category and approximately 7547 units in the women's category. On Myntra, there were 2156 units sold in the men's section and 5062 units in the women's.



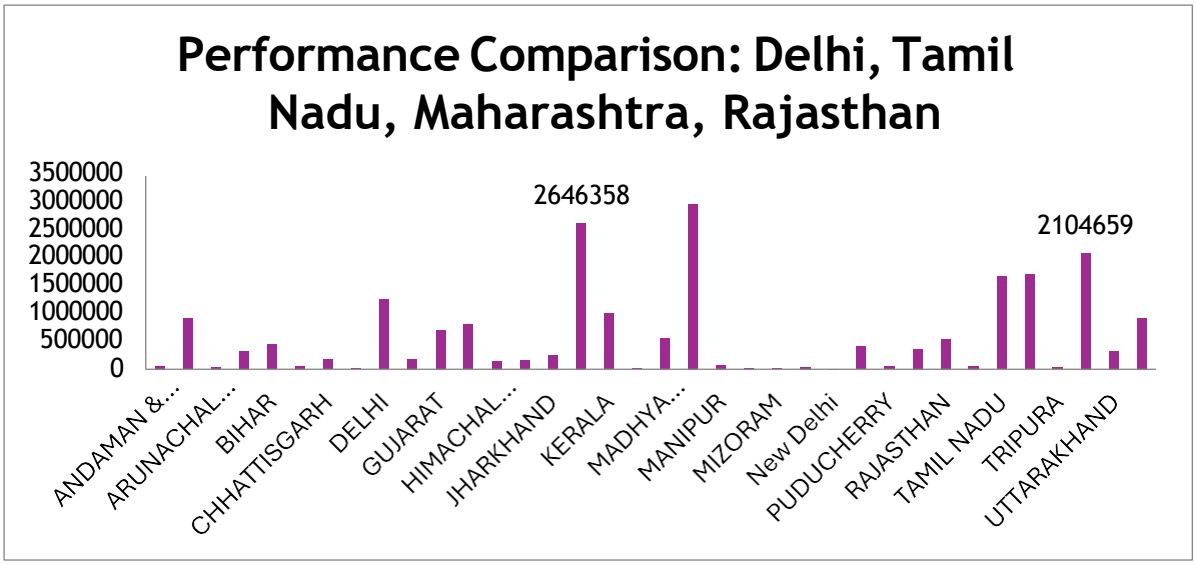
2. Compare all the categories of order where amount is less than 1500 and greater than 5000.

This analysis facilitates the comparison of order categories where the quantity is between 1500 and 5000. putting the kurta (10446) and set (12391) in order of highest order count, then the saree, top, and western clothing.



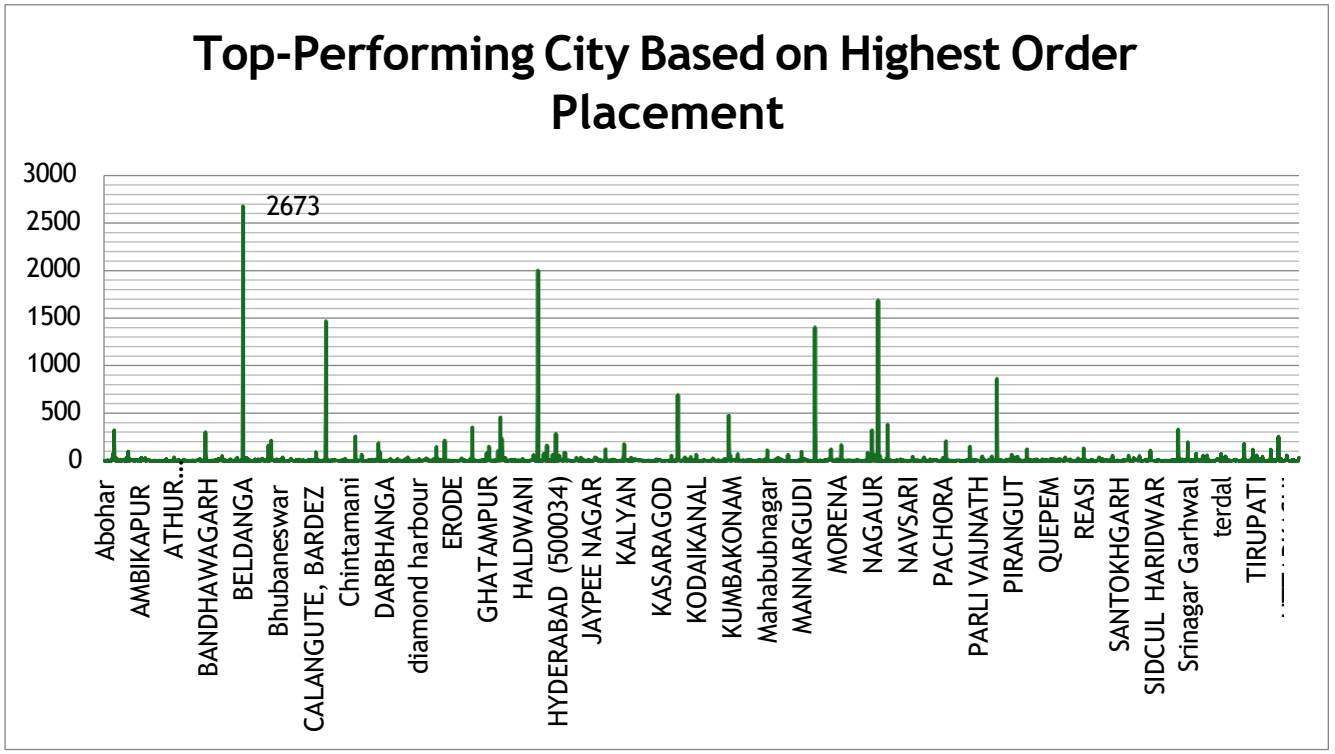
4. Which of the following state perform better than other, Delhi, Tamil Nadu, Maharashtra, Rajasthan.

Karnataka (2646358) and Uttar Pradesh (2104659) were the two states with the best results. Which states performed better than the states listed above is revealed by this research.



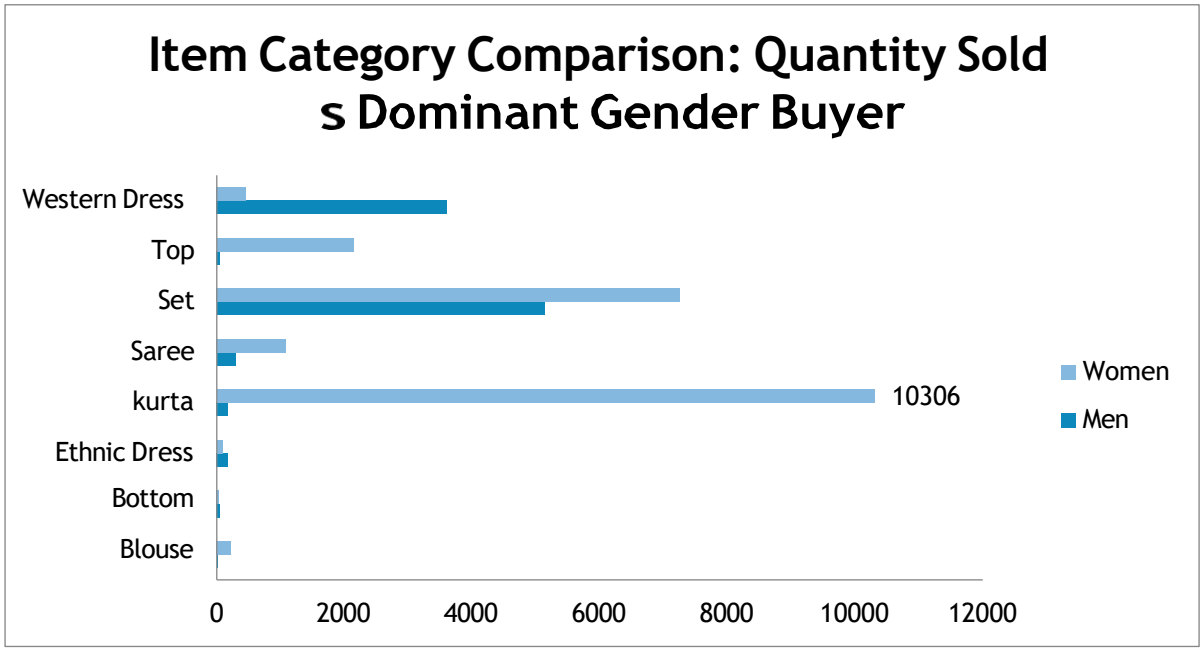
5. Which city performed better than all other cities based on highest order placed.

With 2673 orders, Bengaluru had the greatest order put, followed by Hyderabad (1998). We can clearly see which city performed better than the others based on the largest order put on the graph that was recorded.



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

Western apparel is the most popular item for both men and women, with women purchasing kurtas being the most common category of goods, followed by men's purchases. Based on sales volume, these several product categories are compared in this research.



Conclusion and Review

The study indicates that Amazon is the market leader in terms of sales for all genders, with Myntra and Flipkart following closely behind. Amazon leads in sales for both men's and women's categories, with Myntra and Flipkart following closely after. Kurtas and sets are some of the best-selling items, with the highest sales numbers seen in Bangalore and Karnataka.

The study, which provides useful information regarding regional performance and sales patterns, may help retailers make better judgments. Nevertheless, examining additional factors that impact sales could enhance the analysis. When all is said and done, the findings offer valuable information for optimizing sales strategies in competitive markets.

Regression

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.172398							
R Square	0.029721							
Adjusted R Square	0.029659							
Standard Error	264.5693							
Observations	31047							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	66561870	33280935	475.4629	0			
Residual	31044	2.17E+09	69996.92					
Total	31046	2.24E+09						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	185.155	16.57854	11.16836	6.61E-29	152.6604	217.6496	152.6604	217.6496
X Variable 1	0.047626	0.099327	0.479489	0.631594	-0.14706	0.242312	-0.14706	0.242312
X Variable 2	492.0276	15.95904	30.83065	1.3E-205	460.7472	523.308	460.7472	523.308

Anova-1 factor

Anova: Single Factor				
SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	31047	31237	1.00612	0.008853
Column 2	31047	21176377	682.0748	72136.38

ANOVA				
Source of Variation	*SS*	*df*	*MS*	*F*
Between Groups	7.2E+09	1	7.2E+09	199639.8
Within Groups	2.24E+09	62092	36068.2	
Total	9.44E+09	62093		

Anova- 2 factor

Anova: Two-Factor Without Replication				
SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	421	140.3333	42116.33
Row 2	3	1479	493	685648
Row 3	3	521	173.6667	59609.33
Row 4	3	750	250	172171
Row 5	3	607	202.3333	88482.33
Row 31044	3	974	324.6667	283326.3
Row 31045	3	1145	381.6667	403529.3
Row 31046	3	446	148.6667	47506.33
Row 31047	3	828	276	199225
Column 1	31047	1226250	39.49657	228.5307
Column 2	31047	31237	1.00612	0.008853
Column 3	31047	21176377	682.0748	72136.38

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	7.49E+08	31046	24134.08	1.000774	0.468198	1.016275
Columns	9.09E+09	2	4.54E+09	188446.6	0	2.995877
Error	1.5E+09	62092	24115.42			
Total	1.13E+10	93140				

Descriptive Statistics

Column1		*Column2*		*Column3*	
Mean	39.49657	Mean	1.00612	Mean	682.0748
Standard Error	0.085795	Standard Error	0.000534	Standard Error	1.524289
Median	37	Median	1	Median	646
Mode	28	Mode	1	Mode	399
Standard Deviation	15.11723	Standard Deviation	0.094088	Standard Deviation	268.5822
Sample Variance	228.5307	Sample Variance	0.008853	Sample Variance	72136.38
Kurtosis	-0.1587	Kurtosis	475.3566	Kurtosis	1.768676
Skewness	0.72916	Skewness	19.4509	Skewness	1.052904
Range	60	Range	4	Range	2807
Minimum	18	Minimum	1	Minimum	229
Maximum	78	Maximum	5	Maximum	3036
Sum	1226250	Sum	31237	Sum	21176377
Count	31047	Count	31047	Count	31047

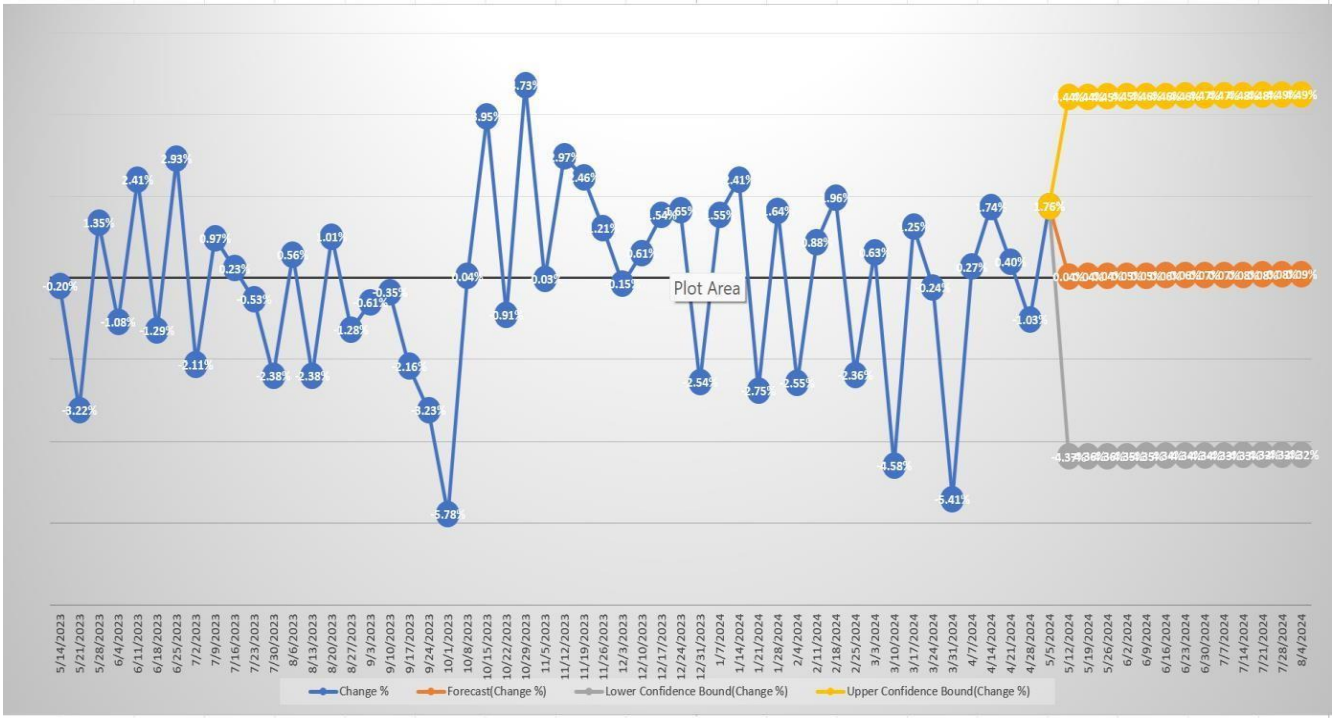
Correlation

	Column 1	*Column 2*	*Column 3*
Column 1	1		
Column 2	0.004884	1	
Column 3	0.003522	0.172377	1

Analysis of Forecasted Trends in MCD Stock Prices

Timeline	Values	Forecast	Lower Confidence Bound	Upper Confidence Bound
5/14/2023	-0.20%			
5/21/2023	-3.22%			
5/28/2023	1.35%			
6/4/2023	-1.08%			
6/11/2023	2.41%			
6/18/2023	-1.29%			
6/25/2023	2.93%			
7/2/2023	-2.11%			
7/9/2023	0.97%			
7/16/2023	0.23%			
7/23/2023	-0.53%			
7/30/2023	-2.38%			
8/6/2023	0.56%			
8/13/2023	-2.38%			
8/20/2023	1.01%			
8/27/2023	-1.28%			
9/3/2023	-0.61%			
9/10/2023	-0.35%			
9/17/2023	-2.16%			
9/24/2023	-3.23%			
10/1/2023	-5.78%			
10/8/2023	0.04%			
10/15/2023	3.95%			
10/22/2023	-0.91%			
10/29/2023	4.73%			
11/5/2023	-0.03%			
11/12/2023	2.97%			
11/19/2023	2.46%			
11/26/2023	1.21%			
12/3/2023	-0.15%			
12/10/2023	0.61%			
12/17/2023	1.54%			
12/24/2023	1.65%			
12/31/2023	-2.54%			
1/7/2024	1.55%			
1/14/2024	2.41%			
1/21/2024	-2.75%			
1/28/2024	1.64%			
2/4/2024	-2.55%			
2/11/2024	0.88%			
2/18/2024	1.96%			
2/25/2024	-2.36%			
3/3/2024	0.63%			
3/10/2024	-4.58%			
3/17/2024	1.25%			
3/24/2024	-0.24%			
3/31/2024	-5.41%			
4/7/2024	0.27%			
4/14/2024	1.74%			
4/21/2024	0.40%			

4/28/2024	-1.03%			
5/5/2024	1.76%	1.76%	1.76%	1.76%
5/12/2024		0.04%	-4.37%	4.44%
5/19/2024		0.04%	-4.36%	4.44%
5/26/2024		0.04%	-4.36%	4.45%
6/2/2024		0.05%	-4.35%	4.45%
6/9/2024		0.05%	-4.35%	4.46%
6/16/2024		0.06%	-4.34%	4.46%
6/23/2024		0.06%	-4.34%	4.46%
6/30/2024		0.07%	-4.34%	4.47%
7/7/2024		0.07%	-4.33%	4.47%
7/14/2024		0.08%	-4.33%	4.48%
7/21/2024		0.08%	-4.32%	4.48%
7/28/2024		0.08%	-4.32%	4.49%
8/4/2024		0.09%	-4.32%	4.49%



The predicted trajectory of MC Donald's closing stock prices is shown in the line graph forecast. Beyond just past data, this projection provides insights into possible future price fluctuations.

The forecast accounts for the inherent uncertainty in stock price prediction by providing a range of possible outcomes, accompanied with lower and upper confidence ranges. By defining the expected variability in the projected values, these boundaries give stakeholders an idea of the possible risk involved in the projection.

The synopsis emphasizes the amount of analysis used to predict future patterns in the price of McDonald's shares. Stakeholders are given important information by this predictive research to help them make strategic decisions in the financial markets.