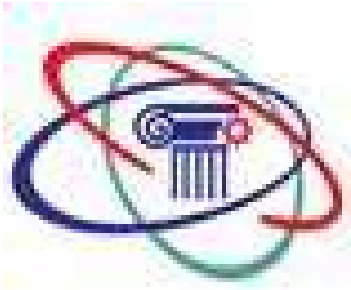
ACROPOLIS INSTITUTE OF TECHNOLOGY & RESEARCH, INDORE

DEPARTMENT OF COMPUTER SCIENCE



CS-605 Data Analytics Lab

3rd Year 6th Semester

2023-2024

SUBMITTED BY –
ANMOL MAYANK
(0827CS211021)

SUBMITTED TO -
Prof. ANURAG PUNDE

S.NO.	Experiment	Remark
1.	Data Analysis Questions: I. Data Analysis Principles II. Statistical Analytics III. Hypothesis Testing IV. Regression V. Correlation VI. ANOVA VII. 5 V's of Big Data	
2.	Dashboard: I. Car Collection Data Analysis II. Order Data Analysis III. Cookie Data Analysis IV. Loan Data Analysis V. Shop Sales Data Analysis VI. Sales Data Sample Analysis VII. Store Dataset Analysis	
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 Forecast Sheet in CISCO's Stock prices	

Comprehensive Study on Data Analysis: Foundational Principles, Statistical Analytics, Hypothesis Testing, Regression Analysis, Correlation, and Analysis of Variance

Data Analysis Principles:

Data analysis principles refer to fundamental guidelines and methodologies employed in the process of extracting meaningful insights from datasets.

1. **Data Quality:** This principle emphasizes ensuring that the data used for analysis is accurate, reliable, and complete. It involves processes such as data validation, verification, and cleansing to eliminate errors, inconsistencies, and missing values.
2. **Data Cleaning:** Data cleaning involves identifying and rectifying errors, inconsistencies, and outliers in the dataset. This process is essential for improving data quality and ensuring the accuracy of analysis results.
3. **Exploratory Data Analysis (EDA):** EDA involves exploring and summarizing the main characteristics of the dataset using statistical and visualization techniques. It helps in understanding data distributions, patterns, trends, and relationships, which can guide further analysis and hypothesis generation.
4. **Data Visualization:** Data visualization is the graphical representation of data to facilitate understanding, analysis, and decision-making. It includes various techniques such as charts, graphs, and dashboards to present complex datasets in an intuitive and visually appealing manner.
5. **Reproducibility:** Reproducibility refers to the ability to replicate data analysis processes and results. Documenting the analysis methodology, code, and assumptions enables other researchers to verify and reproduce the findings, enhancing the transparency and credibility of the analysis.

2. Statistical Analytics Concepts:

Statistical analytics concepts encompass a range of statistical methods and techniques used to analyze and interpret data for decision-making purposes.

1. **Descriptive Statistics:** Descriptive statistics involve summarizing and describing the main features of a dataset, including measures of central tendency (mean, median, mode) and measures of dispersion (variance, standard deviation).
2. **Inferential Statistics:** Inferential statistics are used to make predictions or inferences about a population based on sample data. This includes techniques such as hypothesis testing, confidence intervals, and regression analysis.
3. **Probability Distributions:** Probability distributions describe the likelihood of different outcomes in a statistical experiment or observation. Common distributions include the normal distribution, binomial distribution, and Poisson distribution.
4. **Central Limit Theorem:** The Central Limit Theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution. This theorem forms the basis for many statistical inference techniques.

3. Hypothesis Training:

A hypothesis is a tentative statement or proposition that can be tested and evaluated through empirical observation and analysis.

1. **Null Hypothesis (H0):** The null hypothesis is a statement that there is no significant difference or effect in the population being studied. It serves as the default assumption until evidence suggests otherwise.

2. **Alternative Hypothesis (H1):** The alternative hypothesis is a statement that contradicts the null hypothesis, suggesting that there is a significant difference or effect in the population.
3. **Hypothesis Testing:** Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data. It involves specifying a null hypothesis, selecting a significance level, collecting data, and determining whether the evidence supports rejecting or failing to reject the null hypothesis.

4. Regression and its Types:

Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables.

1. **Linear Regression:** Linear regression models the relationship between the dependent variable and one or more independent variables using a linear equation. It is commonly used for predicting continuous outcomes.

   Formula: $y=\beta 0+\beta 1x+\varepsilon y=\beta 0+\beta 1x+\varepsilon$

2. **Logistic Regression:** Logistic regression models the probability of a binary outcome using the logistic function. It is suitable for predicting categorical outcomes with two levels.

   Formula: $p=11+e-(\beta 0+\beta 1x)p=1+e-(\beta 0+\beta 1x)1$

3. **Polynomial Regression:** Polynomial regression models the relationship between the dependent variable and independent variables using a polynomial equation. It can capture non-linear relationships between variables.

   Formula: $y=\beta 0+\beta 1x+\beta 2x2+...+\beta nxn+\varepsilon y=\beta 0+\beta 1x+\beta 2x2+...+\beta nxn+\varepsilon$

4. **Ridge and Lasso Regression:** Ridge and Lasso regression are regularization techniques used to prevent overfitting in regression models by penalizing large coefficients.

5. Correlation:

Correlation measures the strength and direction of the relationship between two variables.

1. **Pearson Correlation Coefficient:** The Pearson correlation coefficient measures the linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

   Formula: $r=\sum(xi-\bar{x})(yi-\bar{y})\sum(xi-\bar{x})2\sum(yi-\bar{y})2r=\sum(xi-\bar{x})2\sum(yi-\bar{y})2$ $\sum(xi-\bar{x})(yi-\bar{y})$
2. **Spearman's Rank Correlation:** Spearman's rank correlation coefficient measures the strength and direction of association between two ranked variables. It is suitable for assessing monotonic relationships or correlations involving ordinal data.

6. ANOVA (Analysis of Variance):

Analysis of Variance (ANOVA) is a statistical technique used to compare means across multiple groups.

1. **One-Way ANOVA:** One-way ANOVA tests for differences in means across multiple groups when there is one categorical independent variable. It assesses whether there are statistically significant differences between group means.

2. **Two-Way ANOVA:** Two-way ANOVA extends one-way ANOVA to examine the effects of two categorical independent variables on a continuous dependent variable. It assesses both main effects and interaction effects between the independent variables.

3. **Factorial ANOVA:** Factorial ANOVA analyzes the effects of multiple independent variables (factors) on a dependent variable. It is used when there are two or more categorical independent variables, allowing for the examination of main effects and interaction effects.

7. 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.

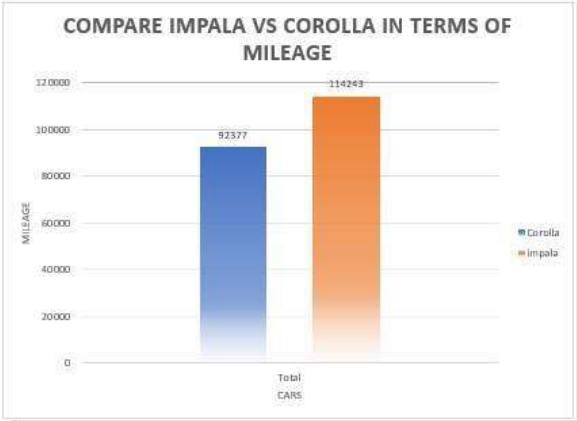
5. Value

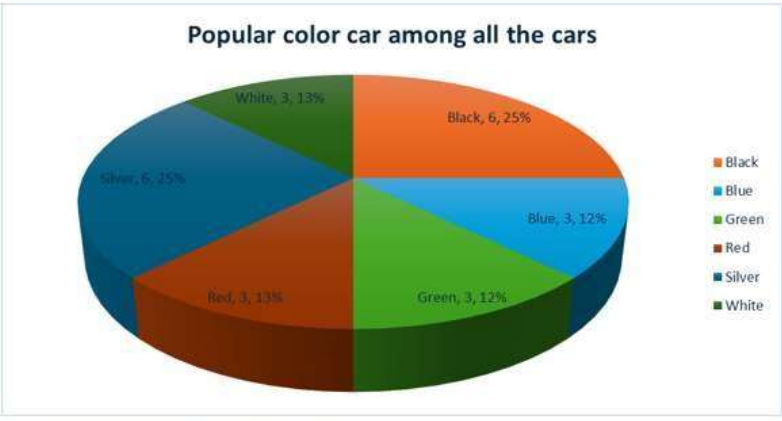
Value refers to the potential insights and benefits that can be derived from analyzing the data. It highlights the importance of transforming data into actionable insights.

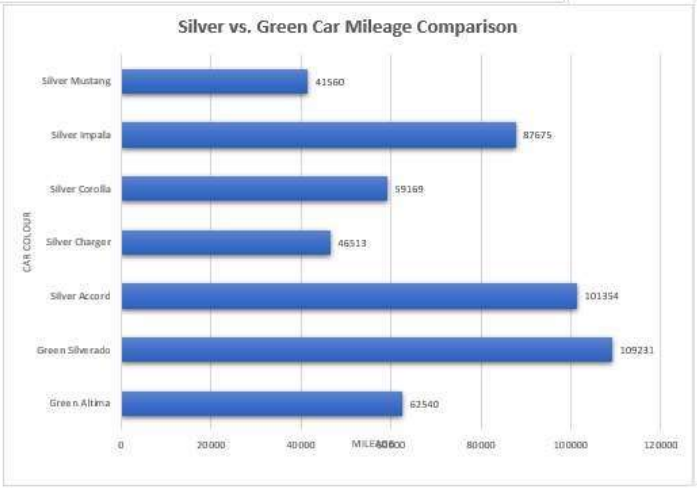
Key Points:

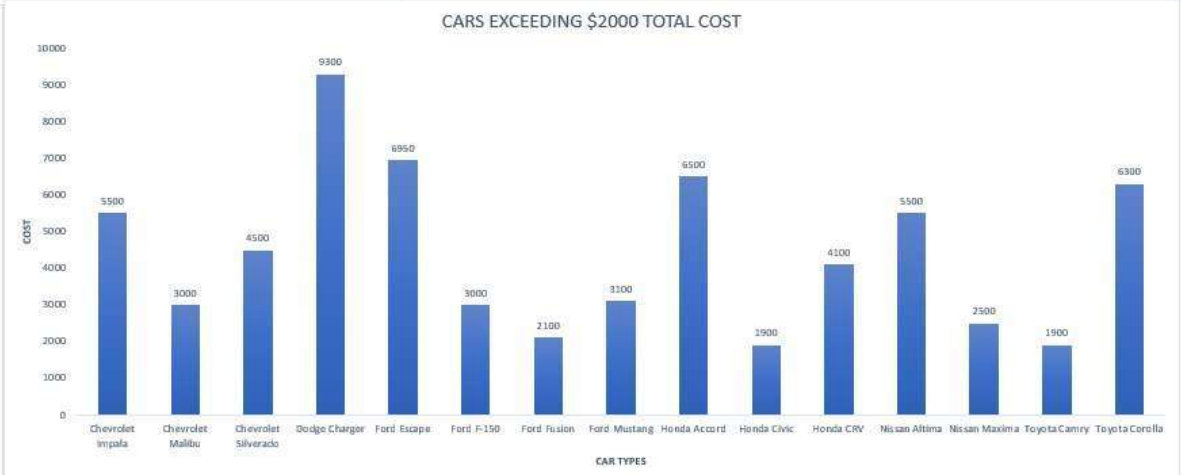
- Insight Generation: Using data analytics to uncover patterns, trends, and correlations that can inform business decisions.
- Business Impact: Data-driven decisions can lead to improved operational efficiency, better customer experiences, and competitive advantages.
- ROI: Assessing the return on investment 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.

CAR COLLECTION DATA DASHBOARD

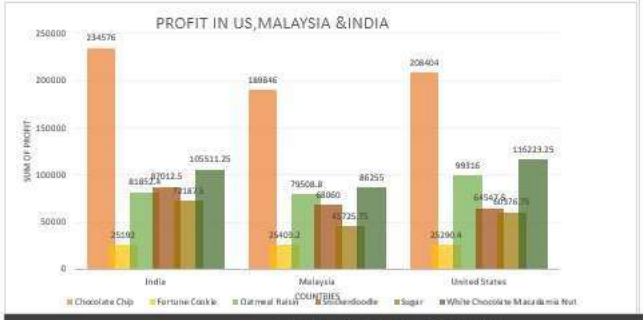


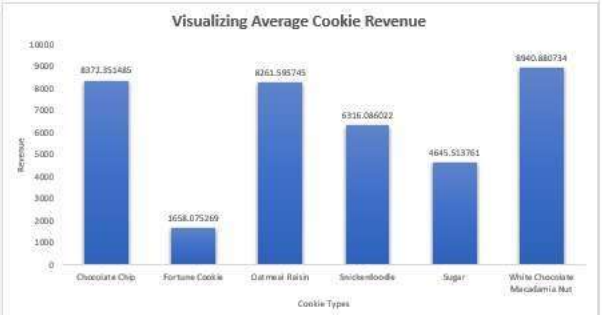


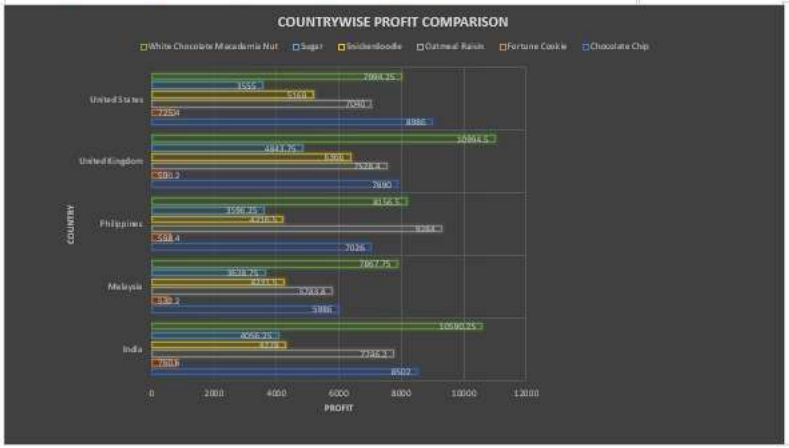


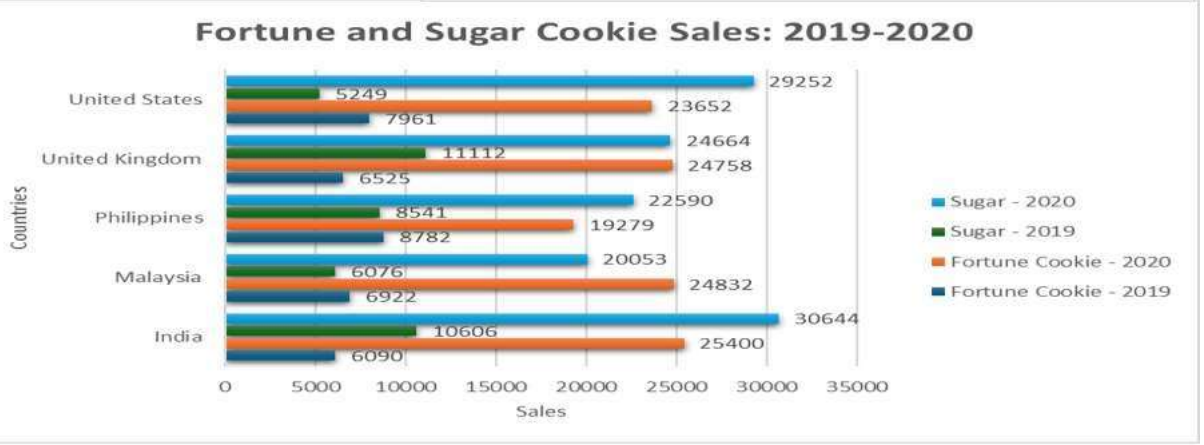


DASHBOARD FOR COOKIES DATA ANALYSIS

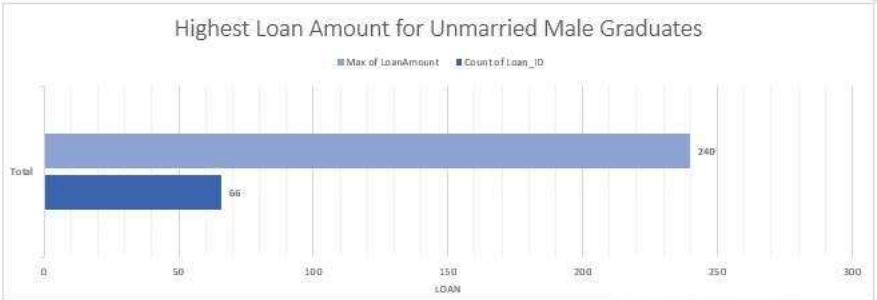


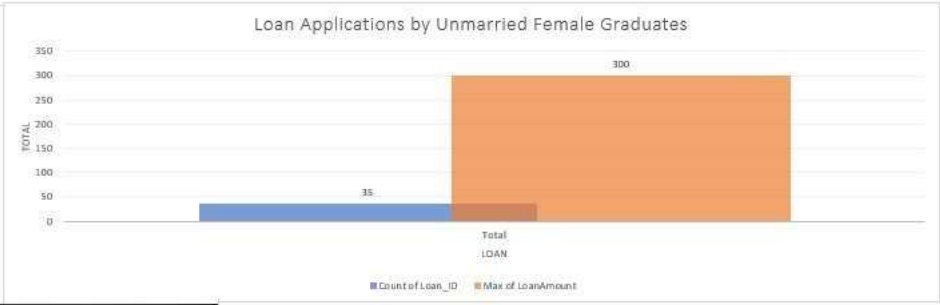


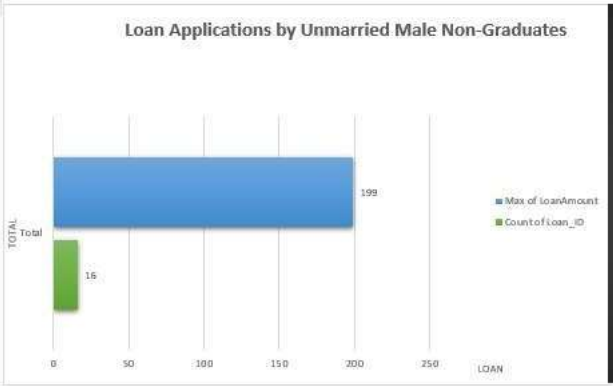


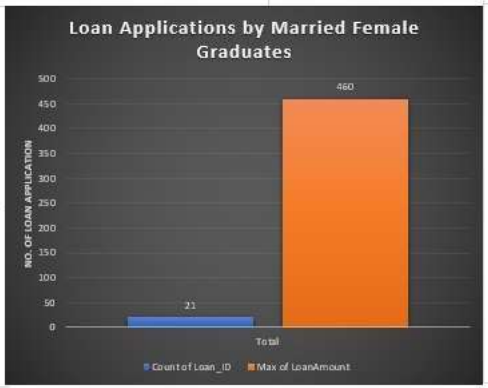


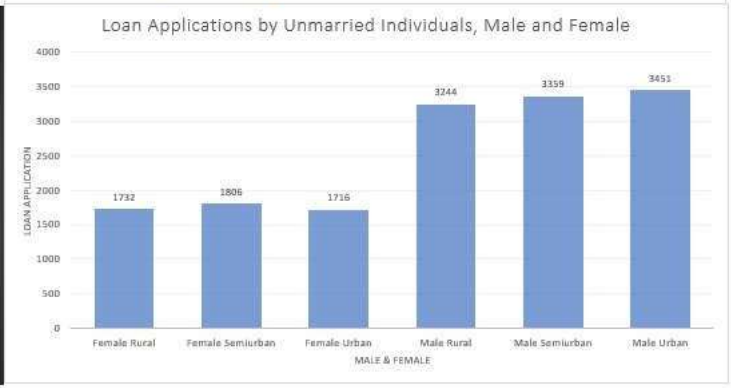
DASHBOARD FOR LOAN DATASET



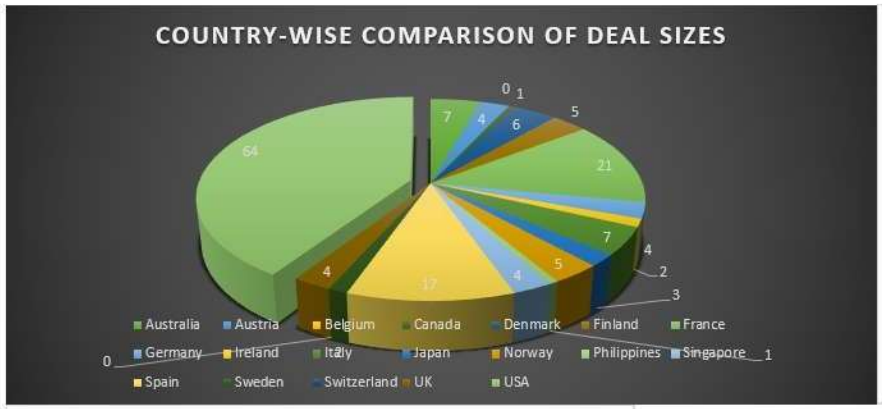


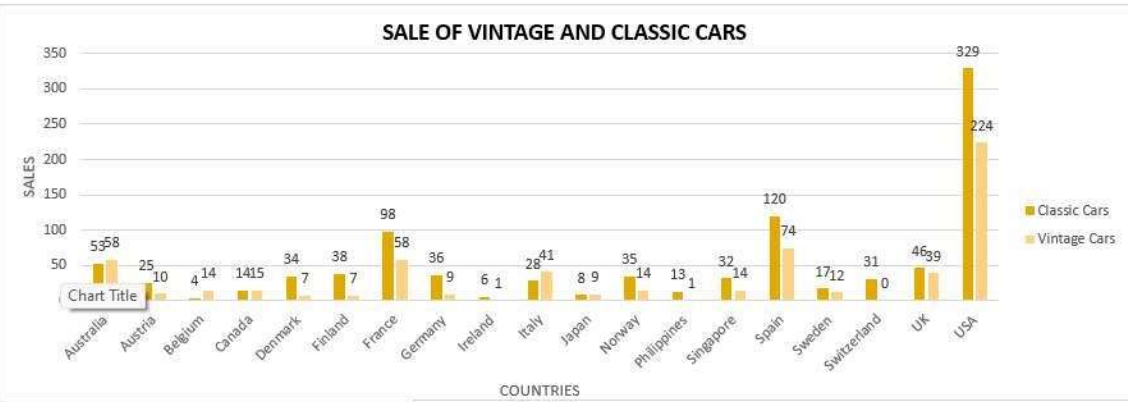


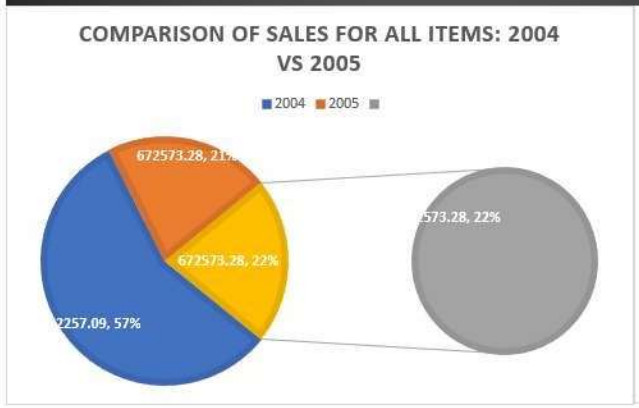


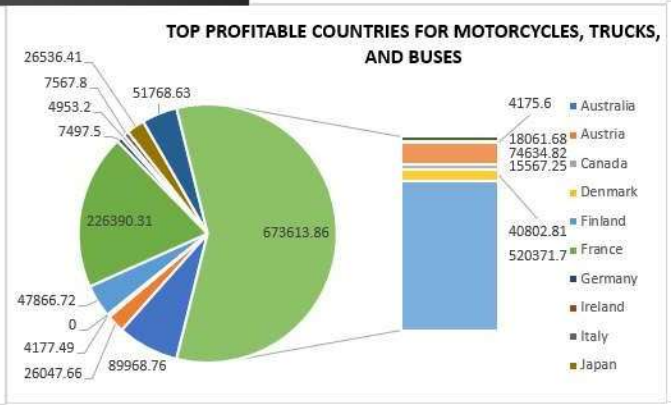


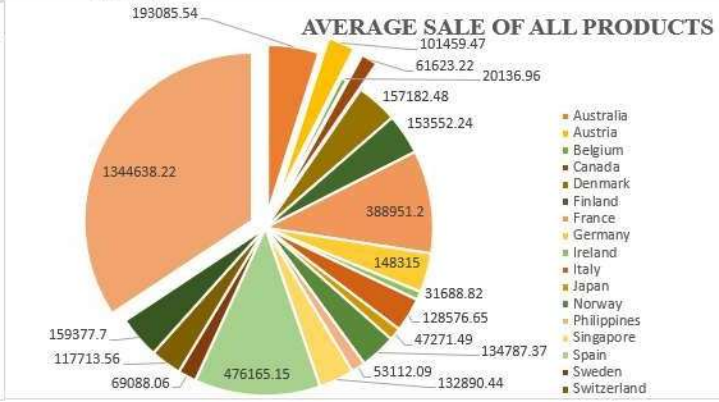
DASHBOARD FOR SALES DATASET



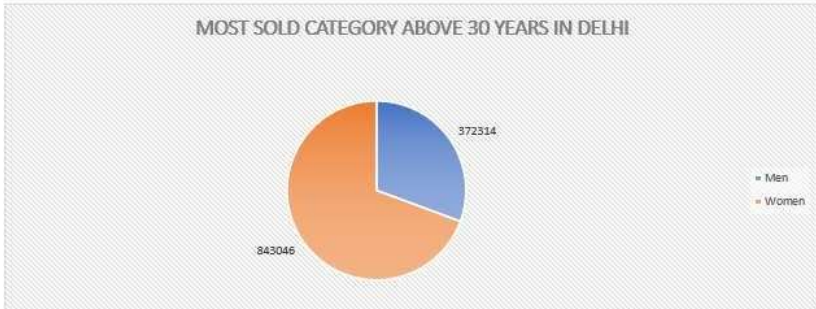


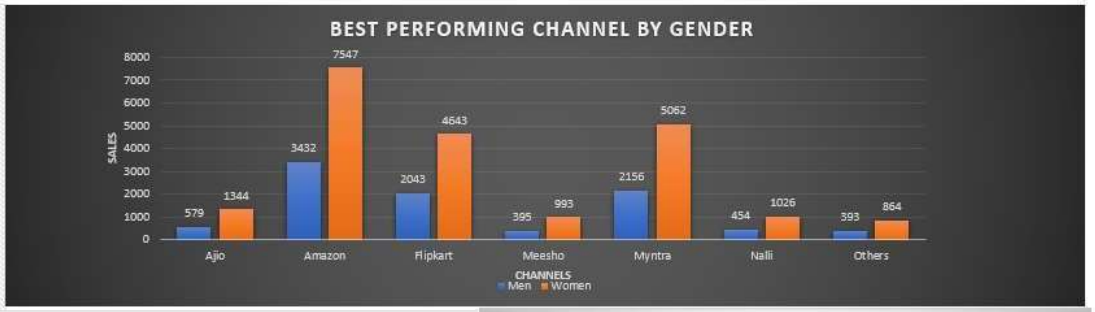


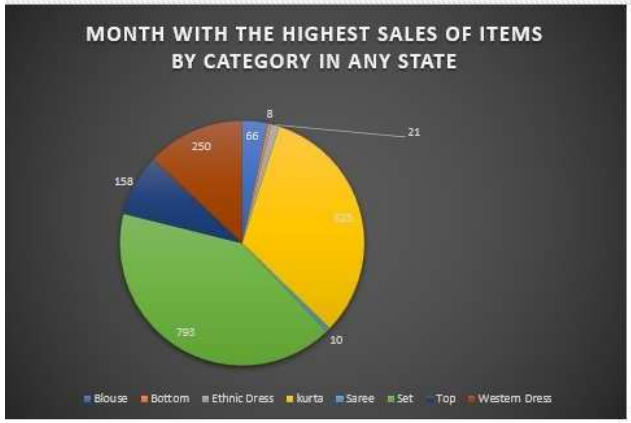


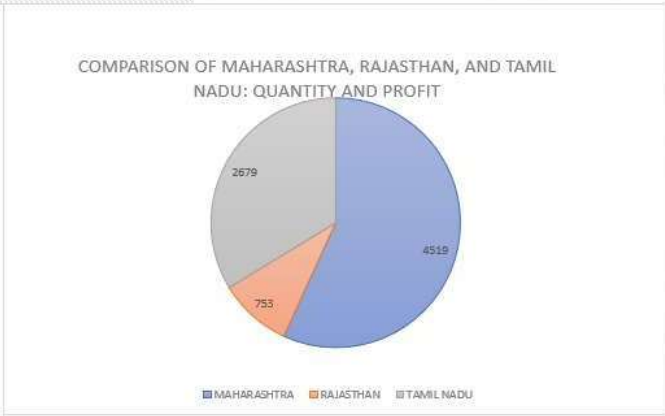


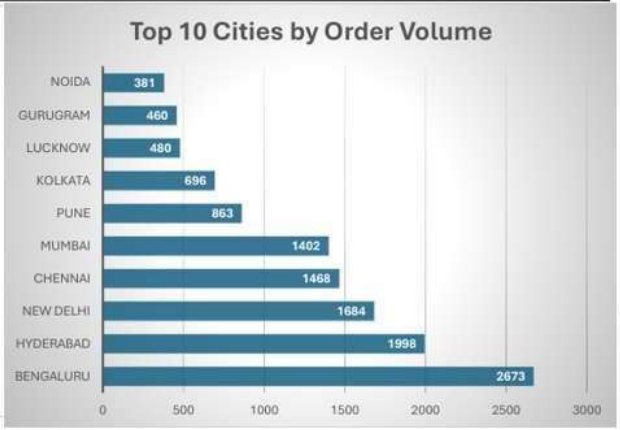
DASHBOARD FOR STORE DATASET



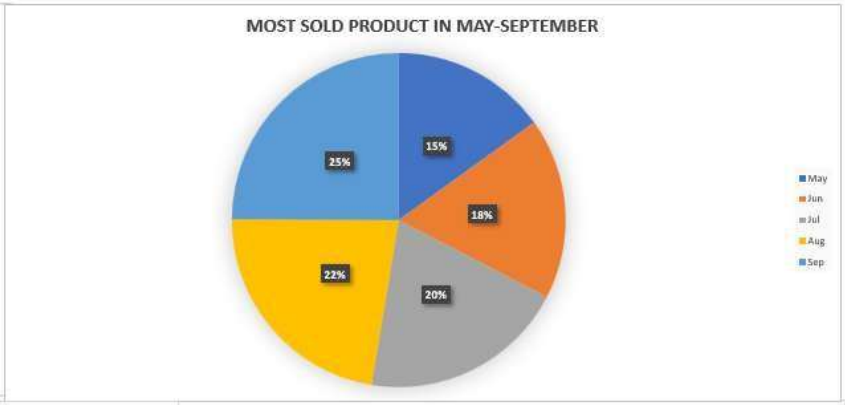


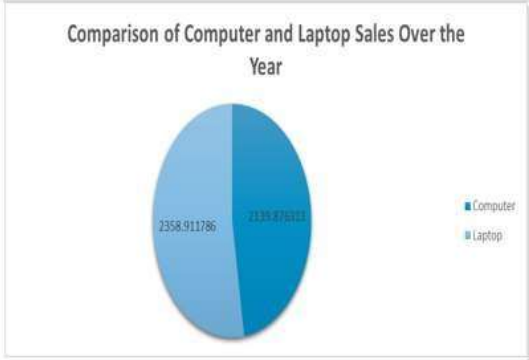


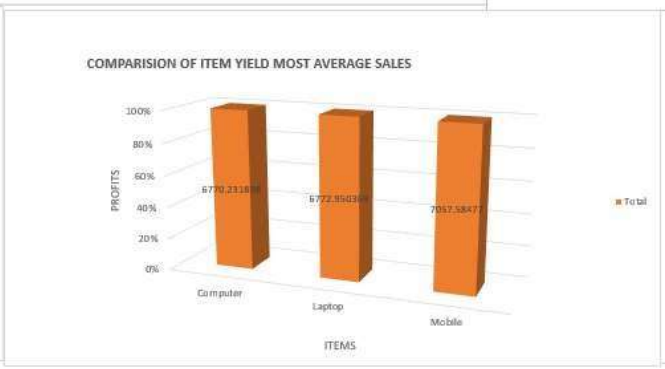


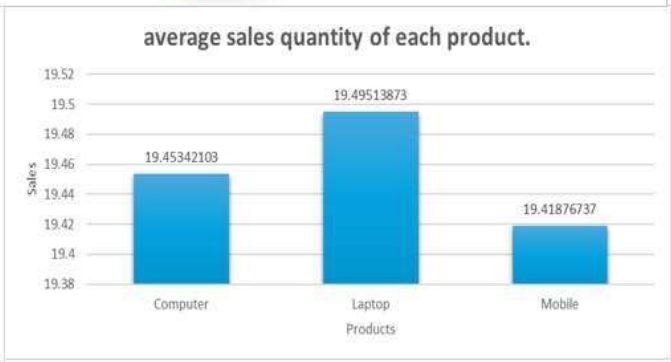


DASHBOARD FOR SHOPSALE

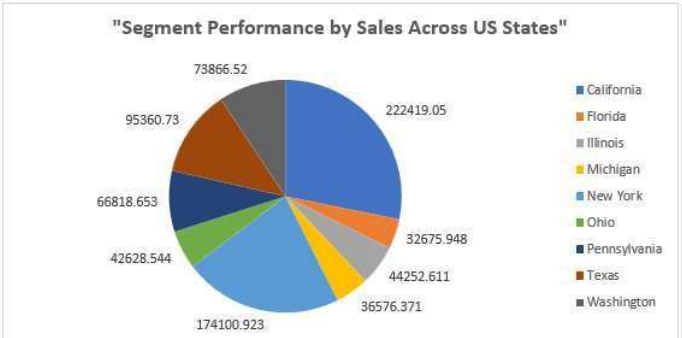


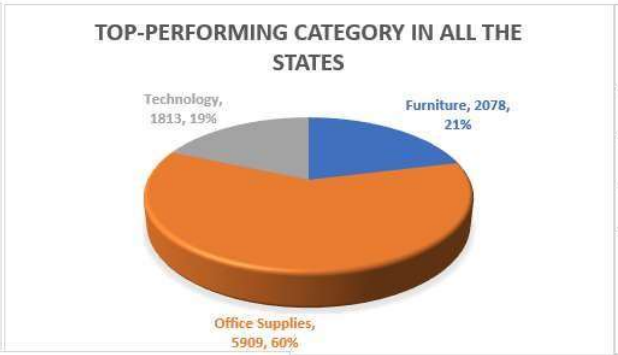


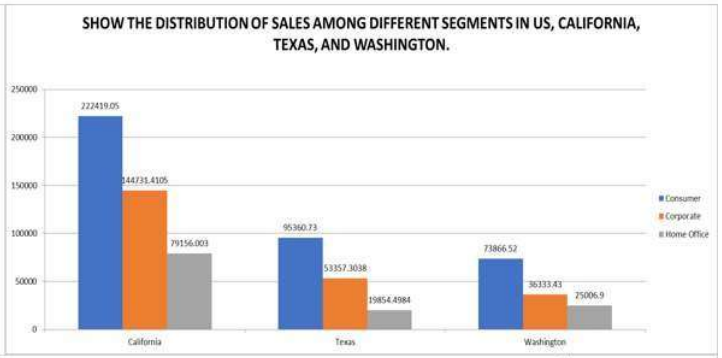


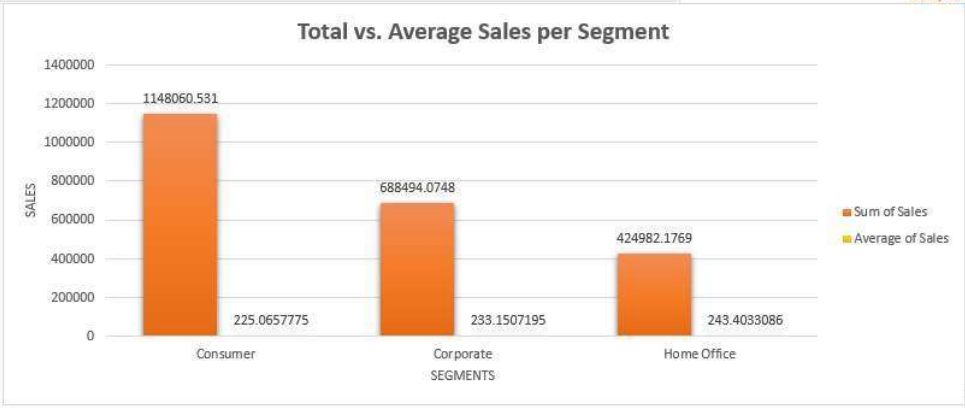


DASHBOARD FOR ORDER DATASET









Car Collection Data Report

Introduction

A thorough examination of the make, model, colour, mileage, price, and cost of many car models is provided by the Car Collection dataset. The purpose of this research is to analyse and extract insights from this dataset to support car-buying decision-making and help with market trends. Six distinct car models—Honda, Chevrolet, Nissan, Toyota, Dodge, and Ford—are included in the dataset.

This report's main target audience consists of auto enthusiasts, analysts, professionals in the automobile sector, and anybody curious in market trends. This report's scope includes a thorough examination of the dataset, along with statistical analysis, graphic aids, and findings interpretation.

Throughout the analysis, we have posed several key questions and performed corresponding analyses to uncover insights.

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.

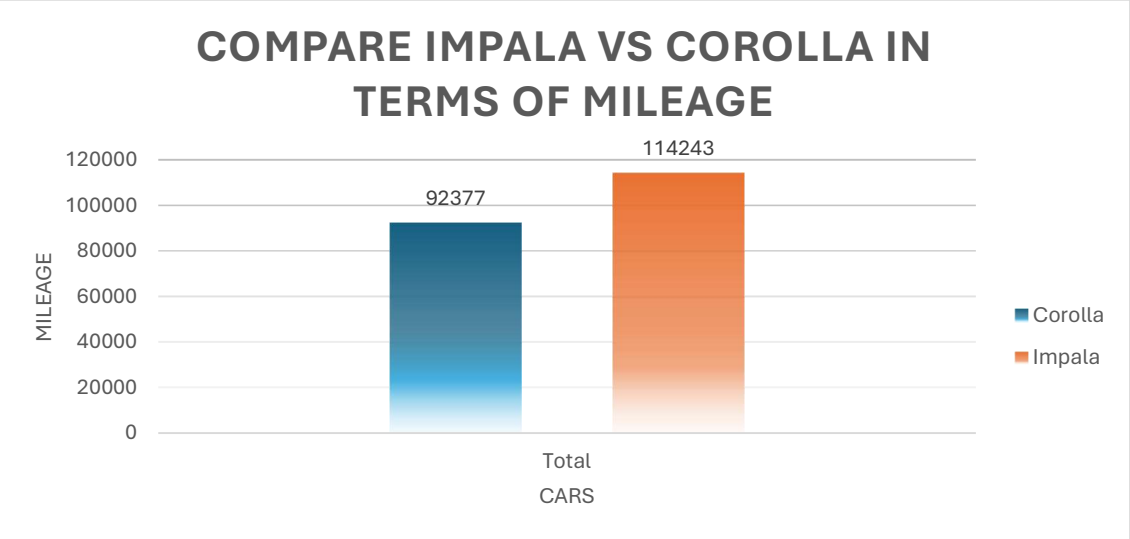
3. Among all the cars which car color is the most popular and is least popular?

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?

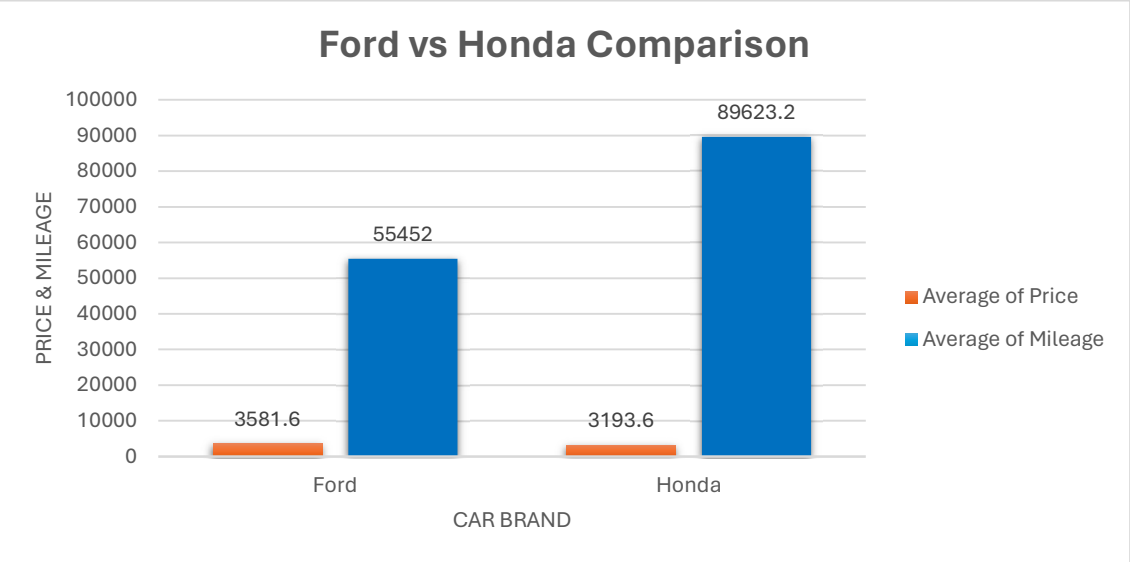
Analytics

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



The fuel economy (mileage) of the Chevrolet Impala and Toyota Corolla, two well-known automobile models, is compared in this comparison. In order to do this, the dataset was filtered to remove irrelevant information, and a column chart was made. The study revealed that the Chevrolet Impala (114243) gets higher gas mileage than the Toyota Corolla (92377).

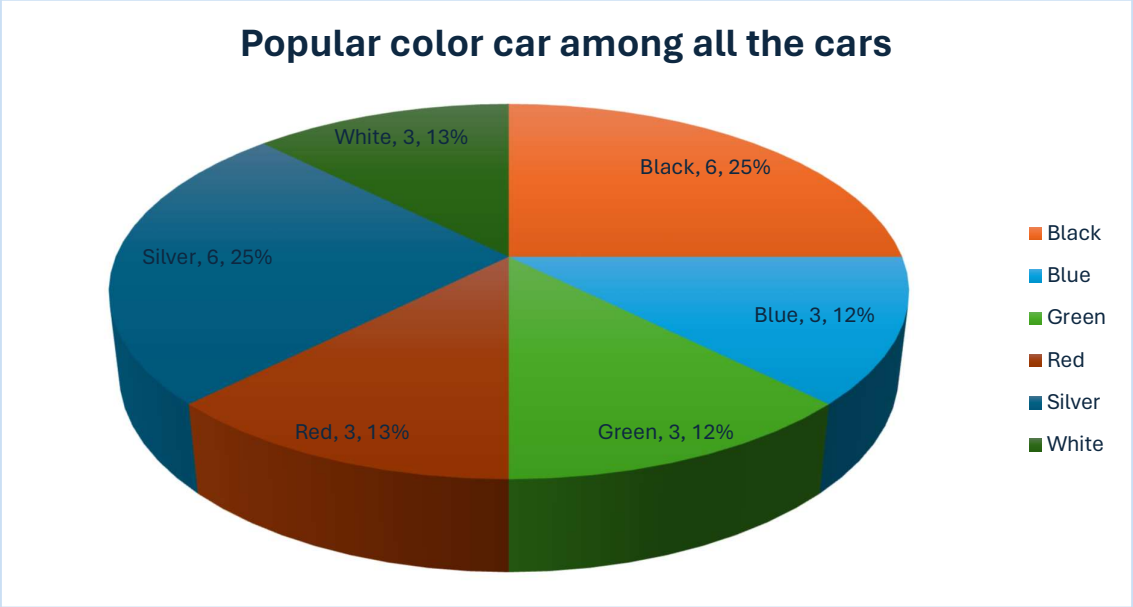
2. Justify, buying of any Ford car is better than Honda.



By contrasting their relative qualities and putting a special emphasis on pricing, this research seeks to justify buying any Ford vehicle over a Honda.

However, the dataset analysis that was done did not support the claim; rather, Honda vehicles outperform Ford vehicles in terms of average price and average mileage.

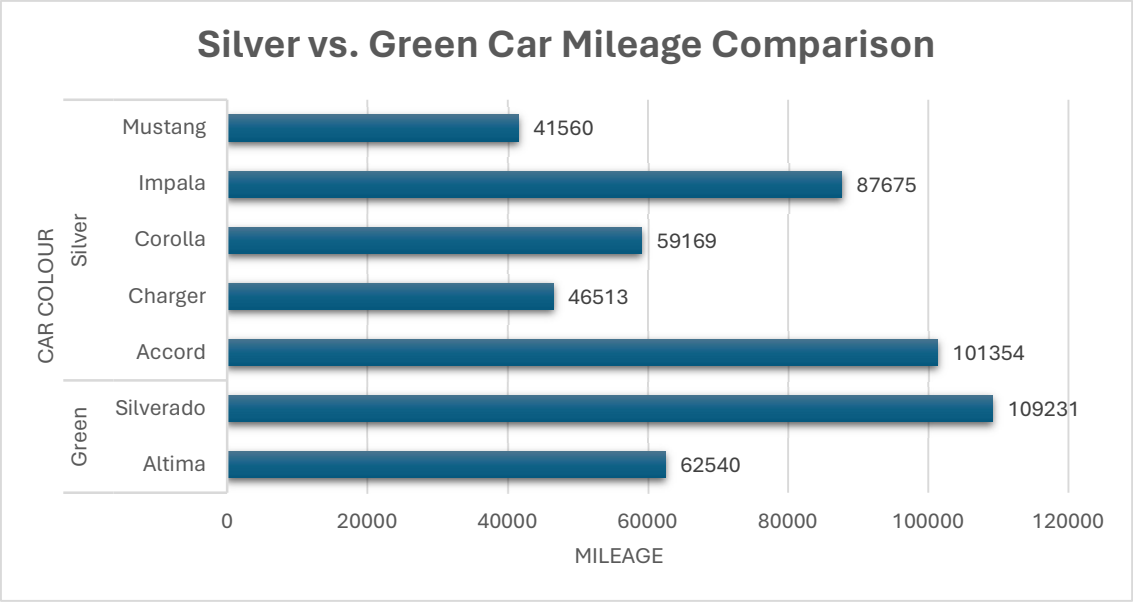
3. Among all the cars which car color is the most popular and is least popular?



This analysis aims to identify the most popular and least popular car colors among all the cars in the dataset based on the count of the make.

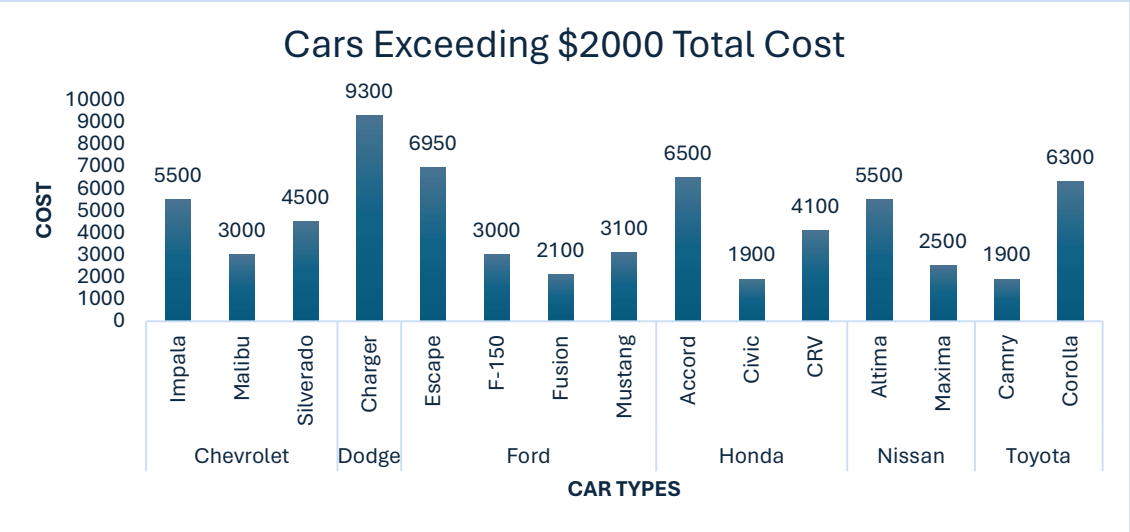
According to the data, the two most popular automobile colors are black and silver, which account for 25% of the company's production, while green and blue cars account for 12% of the total.

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



The objective of this analysis is to determine which automobiles, in terms of mileage, are silver to green. The results show that there are five silver cars: the Charger, Accord, Mustang, Impala, and Corolla. Of them, the Accord has the greatest average mileage (101354). And there were two green cars: an Altima and a Silverado, with the Silverado having the greatest miles (109231).

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



The goal of this analysis is to determine how much the car costs over $2,000. It also displays the intended outcome by utilizing a bar graph and calculating value as the total cost. All cars over $2000 have a grand total cost of $66150.

Conclusion and Review

Comparison: The analysis comparing the mileage of Chevrolet Impala and Toyota Corolla revealed that Chevrolet Impala provides better fuel efficiency.

Ford vs. Honda Comparison: The investigation refuted the basic assumption that Ford vehicles are more cost-effective and had higher mileage than Honda vehicles. When comparing average mileage and pricing to Ford vehicles, Honda vehicles performed better.

Proper Car Colors: Based on the data, the most common car colors are black and white, which account for 25% of all car production. Green and blue, on the other hand, were discovered to be the least common colors, making up a mere 12% of all cars produced. .

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: Based on the data, the total amount spent on cars over $2000 came to $66150.

The research offered insightful information about a number of dataset components, such as mileage comparisons, the popularity of different automobile colours, and financial considerations. But there were differences between the first hypotheses and the results, especially when comparing Ford and Honda vehicles. The analysis was comprehensive and used suitable visualizations to properly display the results, like bar graphs and column charts.

All things considered, the study provides insightful information to consumers, business professionals, and scholars who wish to comprehend market developments. It's crucial to be aware of the analysis's limitations, too, including the dataset's completeness and the need for more research into other variables impacting auto 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

This regression analysis examines the relationship between two predictors, Price and Cost, and the Total Cost of Cars using multiple linear regression. The analysis reveals a moderate linear relationship, indicated by the Multiple R value of 0.414. However, the coefficient of determination (R Square) is relatively low at 0.171, suggesting that only a small proportion of the variance in the Total Cost of Cars can be explained by Price and Cost. Adjusted R Square further adjusts this value for the number of predictors in the model, yielding a value of 0.092. The Standard Error of the estimate is 33202.50, indicating the average deviation between observed and predicted values. The ANOVA table evaluates the overall statistical significance of the regression model, with a p-value of 0.140 suggesting that the model may not be statistically significant at conventional levels.

In the Coefficients table, the intercept is estimated at 133934.06, representing the Total Cost when both predictors are zero. The coefficients for Price and Cost are -9.58 and -6.87, respectively, suggesting a minimal change in Total Cost for a one-unit increase in each predictor.

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				

This ANOVA compares the means of two groups, Price and Cost, regarding their influence on Total Cost of Cars. The Price group averages $3254.5 with a total sum of $78,108, while the Cost group averages $2756.25 with a total sum of $66,150. The analysis shows a slight difference in means between the groups, but it's not statistically significant at the conventional significance level ($p = 0.0554$). Further investigation with a larger sample size may be needed for conclusive results.

This p-value indicates a slight tendency toward a difference in means between the two groups but is not significant at the conventional $\alpha = 0.05$ level. The "Within Groups" SS is 35478117 with 46 df, yielding an MS of 771263.4. The total SS is 38457153 with 47 df.

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 involves a summary of data across 24 rows and 3 columns, detailing counts, sums, averages, and variances. Each row represents a distinct category, while columns denote different attributes. For instance, the first column holds 24 observations with a total sum of $2,011,267, averaging $83,802.79, and a variance of $1.21 billion. The ANOVA table reveals the sources of variation: rows, columns, and error. Notably, the rows' sum of squares (SS) is $8.95 billion with 23 degrees of freedom (df) and a mean square (MS) of $389 million, resulting in an F-value of 0.941 and a non-significant p-value of 0.55. Conversely, the columns' SS is $104 billion with 2 df, yielding an F-value of 126.36 and an extremely low p-value of 2.05E-19, indicating significant differences between columns. The error SS is $19 billion with 46 df. The total SS is $132 billion.

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

This dataset provides summaries for three columns: Column1, Column2, and Column3, each representing distinct attributes. Column1 displays larger monetary values with a mean of $83,802.79 and considerable variability, evident in its standard deviation of $34,844.74 and a wide range from $34,853 to $140,811. Column2 portrays smaller values, such as $2,756.25 as the mean, with less variability compared to Column1, as reflected in its standard deviation of $839.94 and a narrower range from $1,500 to $4,500. Column3 exhibits values similar in magnitude to Column2 but with slightly higher variability, illustrated by its mean of $3,254.5, standard deviation of $914.89, and a range from $2,000 to $4,959. Each column's statistics, including mean, median, mode, standard deviation, skewness, kurtosis, and range, offer insights into the distribution and characteristics of the respective attributes they represent across 24 observations.

Correlation

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

The correlation data provided suggests a relationship between Column 1 and Column 2. A correlation coefficient of 1 for Column 1 with itself indicates a perfect positive correlation, which is expected as it represents the correlation of a variable with itself. Meanwhile, the correlation coefficient of -0.41106 between Column 1 and Column 2 suggests a moderate negative correlation. This negative correlation indicates that as values in Column 1 increase, values in Column 2 tend to decrease, and vice versa. Although the correlation is not very strong, it still implies a discernible pattern in the relationship between the two variables. This insight can be valuable for understanding how changes in one variable may impact the other, potentially informing decision-making or further analysis depending on the context of the data.

Order Data Report

Introduction

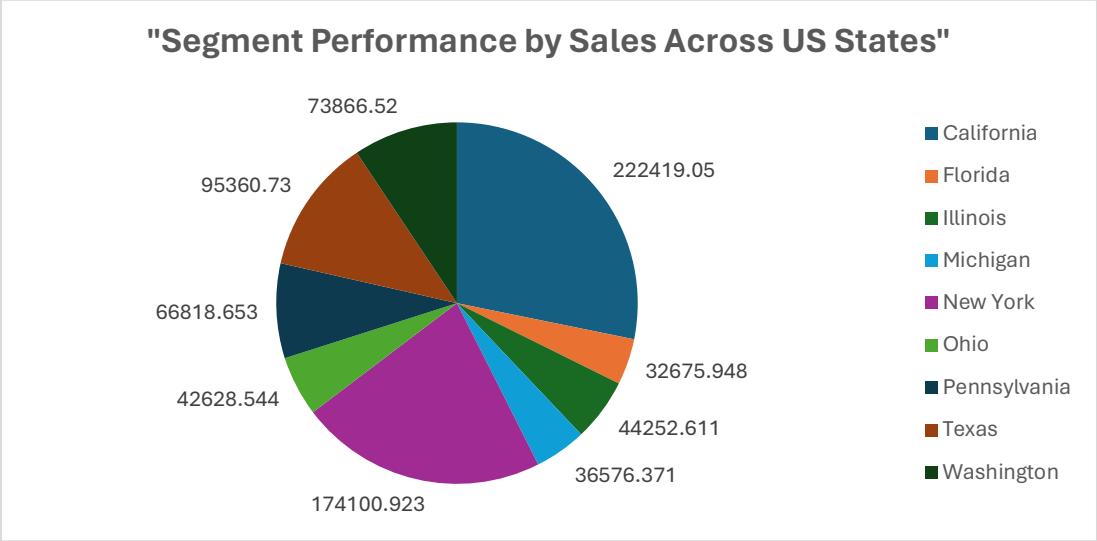
This report explores a vast dataset that records sales transactions in the automotive sector. It includes a variety of variables, including Order ID, Order Date, Ship Date, Customer Information, Product Details, and Sales Figures. Finding practical insights to guide decision-making and promote corporate expansion in the automobile industry is the main goal of this investigation. This analysis looks at sales data from several US states, sectors, categories, and subcategories in order to pinpoint important trends, high-performing segments, and possible growth prospects. The insights obtained from this study will be extremely beneficial to stakeholders in the automobile sector, such as executives, marketers, and sales managers, who are looking to maximize income, improve customer happiness, and optimize sales methods.

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.

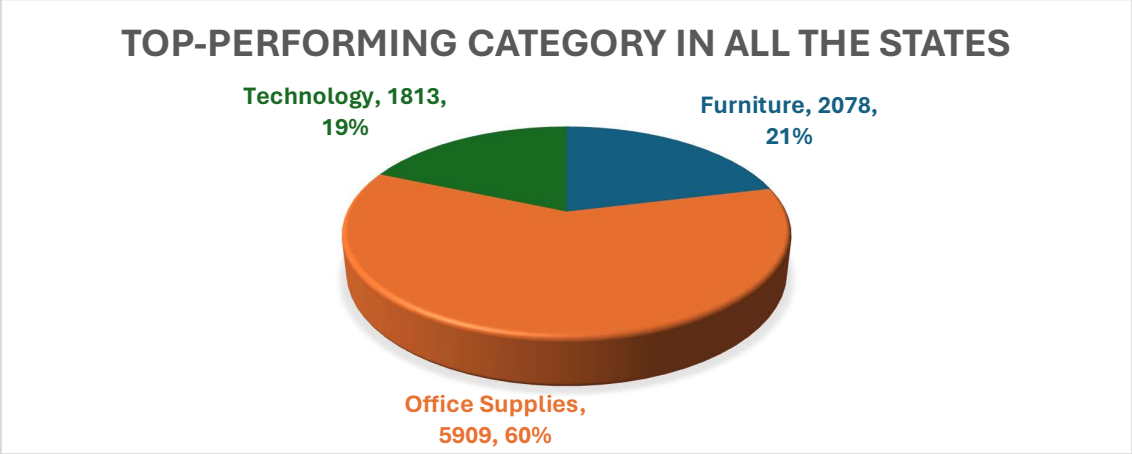
Analytics

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



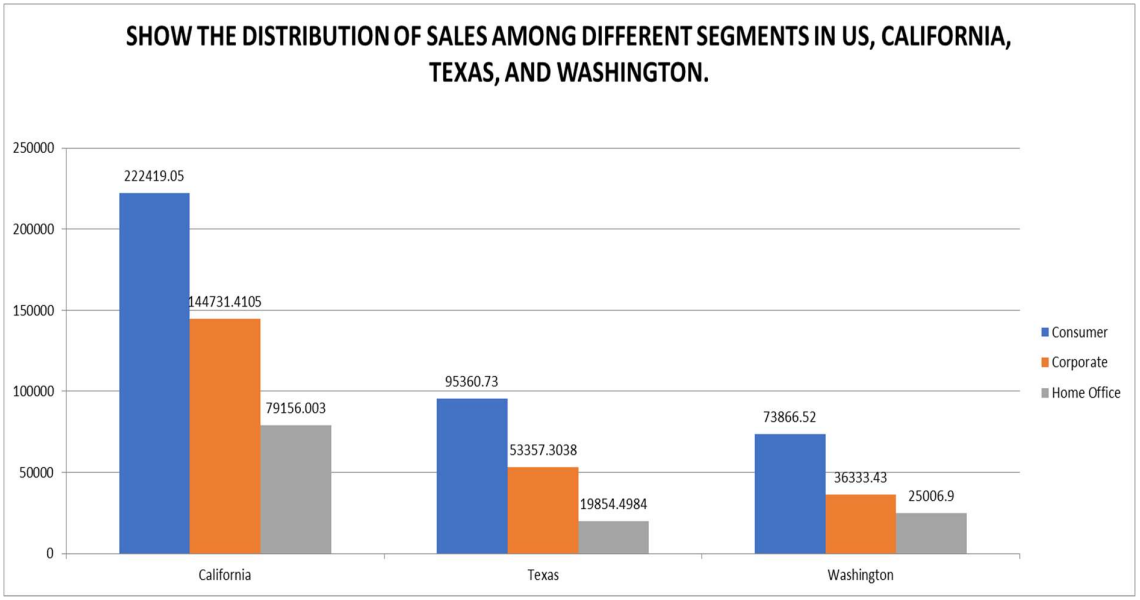
California (222419.05) was found to have the most sales when all the states were compared in terms of sector and sales. The consumer category (1148060.531) showed good performance across all states.

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



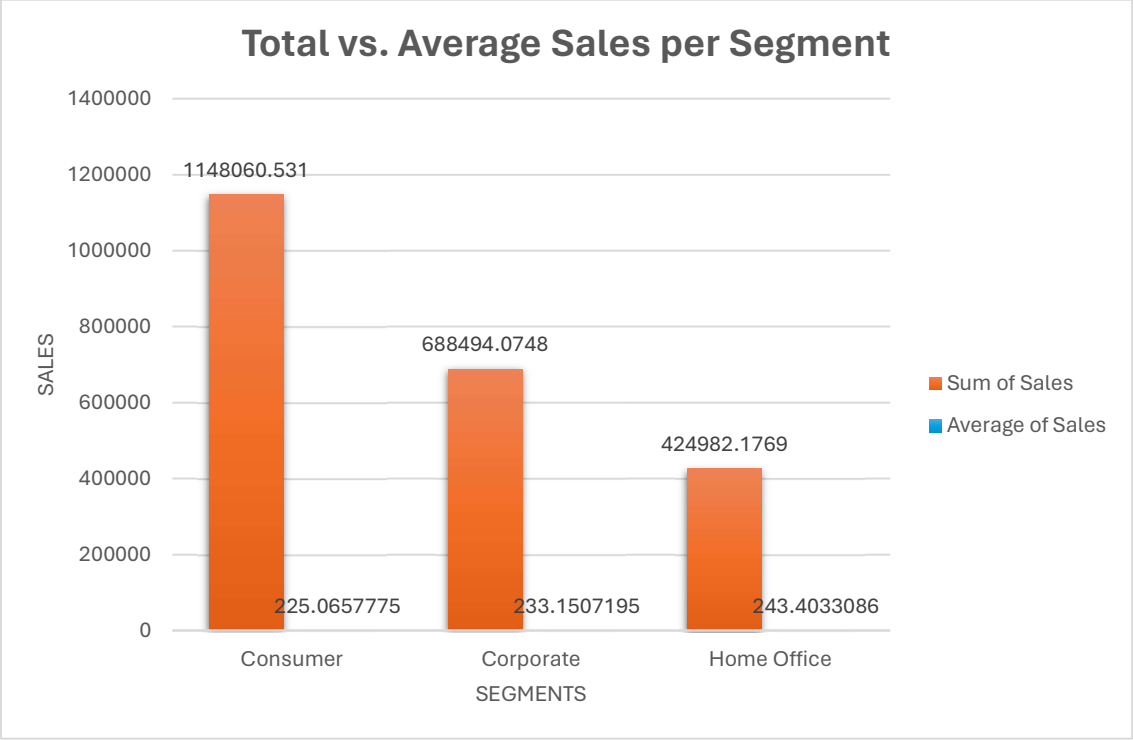
With a total sales count of 5909, office supplies are the best-performing category across all states, followed by technology (1813) and furniture (2078).

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



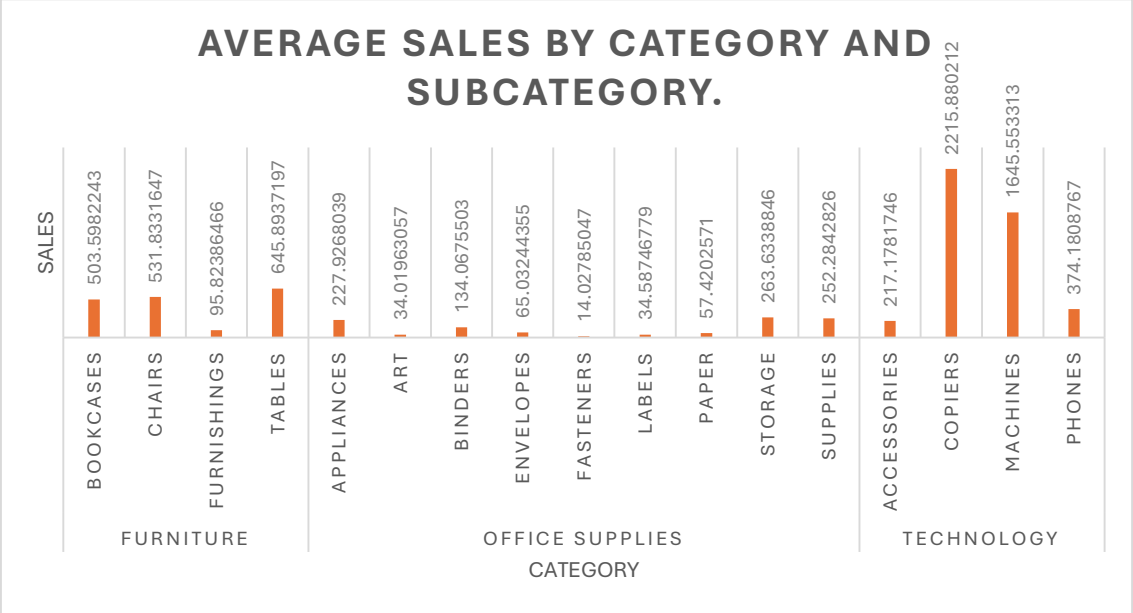
Using a bar chart to display the proportion of distribution and filtering the states for the overall sales count. The US, California, Texas, and Washington have the highest sales in the consumer category.

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



It is clearly visible that the consumer segment has higher average sales with 1148060.531 and home office segment has total sales of 243.40.

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



The analysis shows the average sales for the 3 categories having multiple subcategories, the categories are Furniture, Office Supplies, Technology.

Conclusion and Review

The examination of sales statistics in the automobile sector yields numerous significant conclusions. When it comes to sales volume, California is the best-performing state, and the consumer category does well in every state. According to consumer preferences, Office Supplies is the category that performs the best, followed by Furniture and Technology. Sales in the US are regularly led by the consumer market, especially in California, Texas, and Washington.

The data also shows that the Consumer sector's average sales are greater than those of the Home Office category. All things considered, these insights offer insightful advice that can be used to enhance client interaction, optimize sales tactics, and propel corporate success in the automobile sector.

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

The regression output indicates a very weak and statistically insignificant relationship between the independent variable (X Variable 1) and the dependent variable. The Multiple R value of 0.000434 suggests an almost negligible correlation, while the R Square value of approximately 0.000000188 indicates that virtually none of the variation in the dependent variable is explained by the independent variable. The Adjusted R Square is slightly negative (-0.0001), highlighting the poor explanatory power of the model. The F-statistic (0.001844) and the Significance F value (0.965747) further demonstrate that the regression model does not fit the data well and is not statistically significant.

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

The statistical summary for Column1 provides a comprehensive overview of the distribution and characteristics of the data. The mean value of 230.1162 indicates the average of the dataset, while the standard error of 6.320053 suggests a moderate level of precision around the mean estimate. The median, which is 54.384, highlights the central tendency of the data, and the mode of 12.96 represents the most frequently occurring value.

The standard deviation of 625.3021 and the sample variance of 391002.7 indicate significant variability within the dataset. High kurtosis (307.3056) and skewness (13.05363) values reveal that the data distribution is heavily tailed and highly skewed to the right, respectively. The range of the data is substantial, spanning from a minimum value of 0.444 to a maximum of 22638.48, with a total range of 22638.04. The sum of all values is 2252607, and the total number of observations is 9789, indicating a large dataset. This statistical summary suggests that while the average value is relatively low, the data includes extreme outliers and is not symmetrically distributed.

Cookie Data Report

Introduction

Six distinct varieties of cookies are included in our cookie data set: chocolate chip, fortune cookie, sugar, oatmeal raisin, Snicker doodle, and white chocolate macadamia nut. We possess an abundance of information on these cookies, including the quantity sold, the expenses incurred, the income (revenue), and the earnings. Not only are we examining a single location or period, but we are also examining several nations and times periods to observe how things change. This research aims to provide insights into consumer preferences, price points, and geographic areas where cookies are most popular, in addition to providing information regarding cookies.

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?

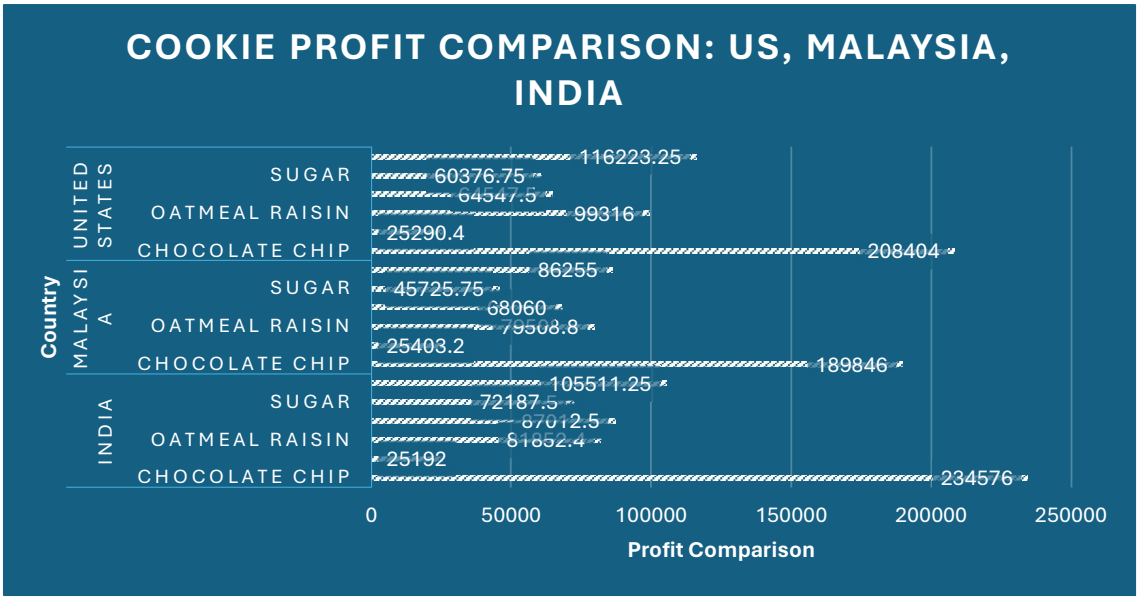
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?

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?

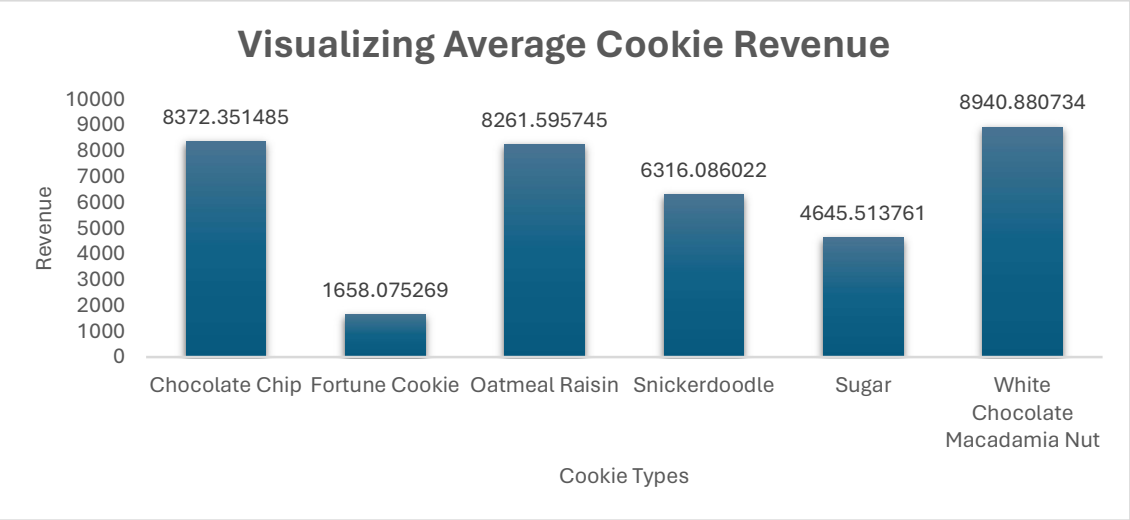
Analytics

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



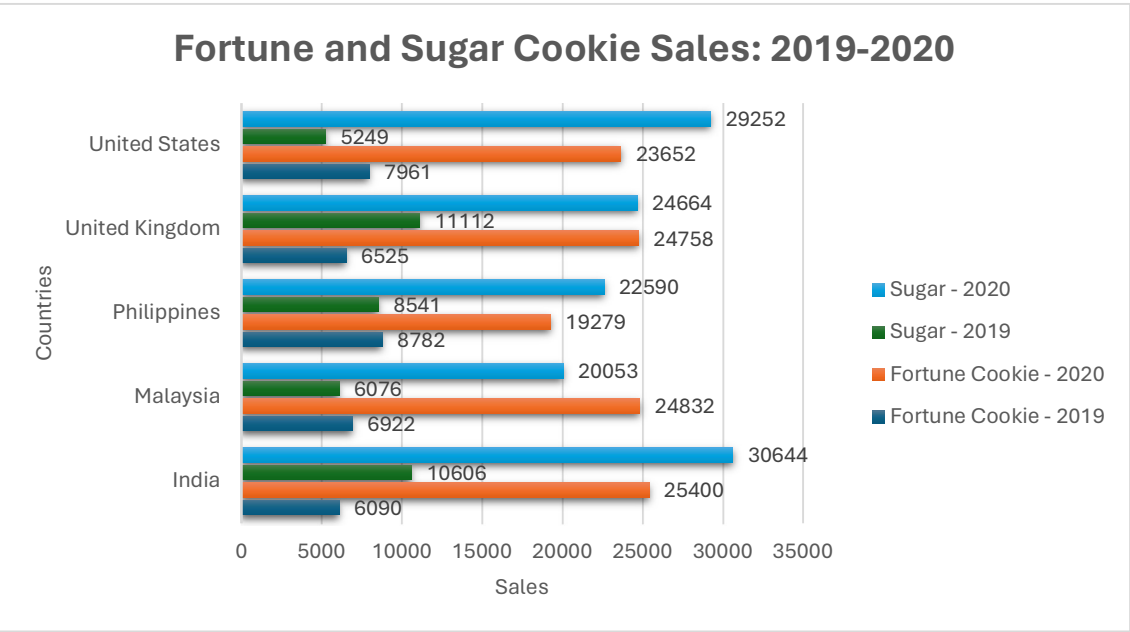
The profit margins for each variety of cookie in the US, Malaysia, and India are compared in this research. India's maximum profit on chocolate chips is followed by that of Malaysia and America.

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



This analysis aims to provide average revenue generated and it's visible that white chocolate macadamia nut with average revenue generate is 8940.88 followed by chocolate chip.

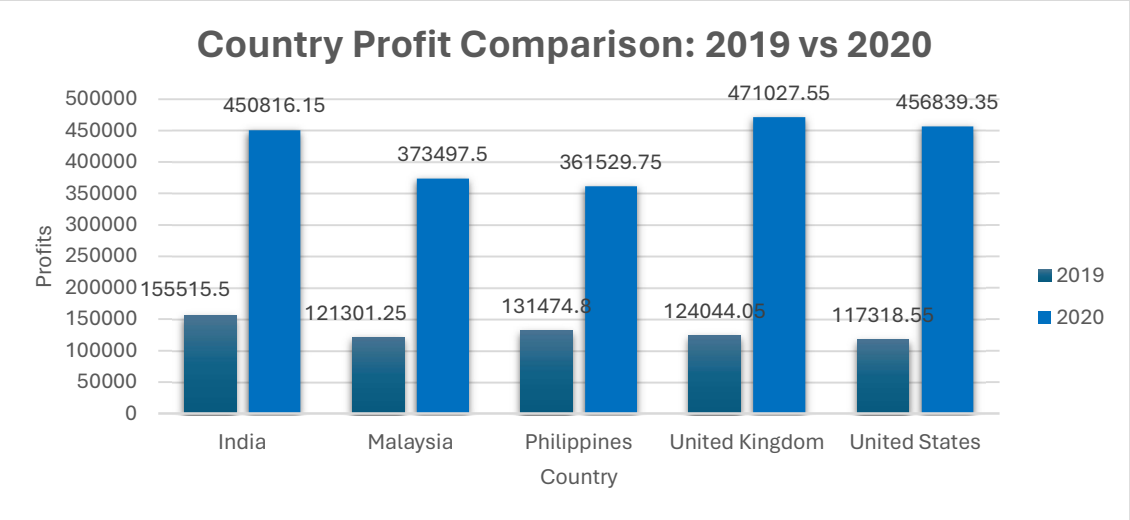
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?



This analysis compares the sales of fortune and sugar cookies in the various countries for the years 2019 and 2020. India leads the way in significant sales of sugar cookies for the year 2020, with 30644 sales; the United Kingdom led the way in sales of sugar cookies in 2019.

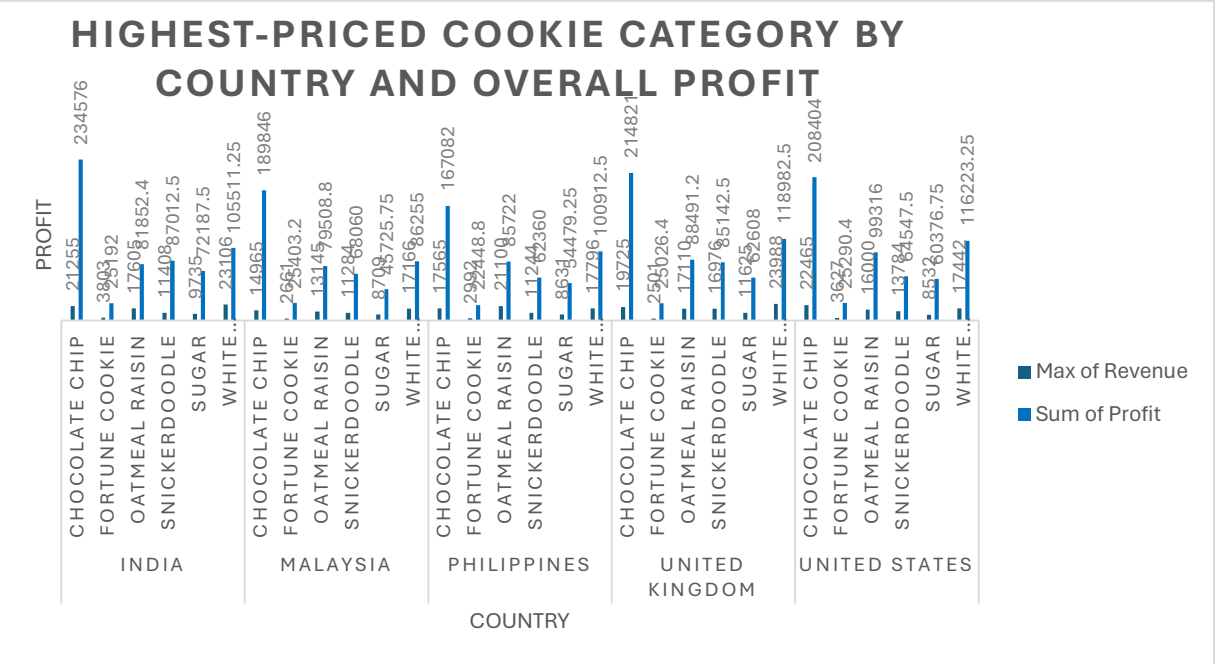
India again leads in sales of fortune cookies, with 25400, followed by Malaysia; the Philippines lead in sales of fortune cookies, with 8782, followed by the United States.

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



This analysis compares the profits made by the various countries in the fiscal years 2019 and 2020. The graph indicates that the United Kingdom made the most profit in 2020 with sales of 471027.55, followed by the United States with 456839.35, and that India made the most profit in 2019 with sales of 155515.5, followed by the Philippines with 131474.8.

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



This analysis aims to find the cookie category sold for the highest price, country-wise, profit earned by that category, max of revenue is recorded by chocolate chip (23988) and sum of profit is recorded by sugar (2763364.45) for the country India followed by United Kingdom.

Conclusion and Review

The study shed light on the profits made by several cookie varieties in the US, Malaysia, and India. The country that made the most money from chocolate chip cookies was India, followed by Malaysia and the US.

The cookies with the greatest average revenue were white chocolate macadamia nut cookies, closely followed by chocolate chip cookies.

In terms of sales, the United Kingdom led the world in sugar cookie sales in 2019, with India showing notable sales in 2020. Sales of fortune cookies were increasing in both years in Malaysia and India, with significant sales also coming from the US and the Philippines.

In terms of comparing profits by nation for 2019 and 2020, the United States and the United Kingdom both had the greatest profits in 2020. India and the Philippines had the biggest profits in 2019.

In terms of income, chocolate chip cookies brought in the most money, but altogether, sugar cookies made the most profit.

The report helped players understand market dynamics and make wise decisions by providing insightful information on the cookie sector. Visuals that were acceptable and easy to understand were used to successfully explain the findings. It's crucial to recognize the need for more research into other variables affecting sales and profitability, though. For trustworthy insights, data completeness and correctness must be guaranteed.

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

The regression statistics show a Multiple R of 1 and an R Square of 1, meaning that the independent variables explain 100% of the variance in the dependent variable. The Adjusted R Square also remains at 1, confirming that this perfect fit holds even when accounting for the number of predictors. The standard error is nearly zero (9.16E-12), indicating almost no deviation from the regression line.

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				

The ANOVA single-factor analysis compares two groups, Column 1 and Column 2, each with 700 observations. Column 1 has an average value of 2752.792 and a variance of 4149401, while Column 2 has a higher average of 3947.664 and a variance of 6842519. The ANOVA results show a significant difference between the groups, with an F-value of 90.92153, which is much higher than the critical value (F crit) of 3.848119. The very low p-value (6.36E-21) indicates that the differences in means are statistically significant, implying that the variations in the groups' means are not due to random chance.

Anova: two factors:

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				

The ANOVA two-factor without replication analysis compares the means of rows and columns for a dataset with three observations per row. The summary statistics show variability in averages and variances across both rows and columns. The ANOVA results indicate significant differences in both rows and columns. For rows, the F-value is 14.75112, much higher than the critical value (F crit) of 1.112595, and a p-value of 0, showing significant variation between row means. For columns, the F-value is 1484.458, also far exceeding the critical value of 3.002161, with a p-value of 0, indicating significant differences between column means. The large F-values and zero p-values suggest that the differences in both row and column means are highly significant, not due to random chance.

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

The descriptive statistics for four columns of data summarize the central tendency and variability within each column. Column 1 has a mean of 1608.32, with a standard deviation of 867.45, indicating moderate variability. Column 2 has a much higher mean of 6700.456 and a larger standard deviation of 4623.901, showing greater spread. Column 3's mean is 2752.792, with a standard deviation of 2037.008, and Column 4 has a mean of 3947.664 with a standard deviation of 2615.821, both showing significant variability. All columns exhibit positive skewness and kurtosis values near zero, indicating slightly skewed distributions with moderate tail heaviness. The ranges and sums further highlight the differences in data spread and total values across the columns, with Column 2 having the highest variability and total sum. Each column has 700 observations, ensuring a consistent sample size for comparison.

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

The correlation matrix shows the strength and direction of linear relationships between four columns of data. Column 1 is moderately to strongly correlated with the other columns, with correlation coefficients of 0.796 with Column 2, 0.743 with Column 3, and 0.829 with Column 4. Columns 2, 3, and 4 exhibit very strong positive correlations among themselves, with coefficients of 0.992 between Columns 2 and 3, 0.995 between Columns 2 and 4, and 0.975 between Columns 3 and 4. These high values indicate that as one of these columns increases, the others tend to increase as well, suggesting a strong linear relationship between these sets of data.

Loan Data Report

Introduction

The loan dataset includes a wealth of information about loan applicants, including details about their income, property area, gender, marital status, education level, and loan amount. This dataset provides a wealth of information on loan application behaviour.

Our goal in this research is to examine the traits of loan candidates and look for trends in the data. We use pivot tables and charts to try to answer certain questions about the educational backgrounds, loan amounts, and demographics of loan applicants.

Financial institutions must comprehend the subtleties of loan applications in order to make well-informed judgments, streamline the lending process, and customize services to satisfy the wide range of client demands. Our goal in doing this research is to find practical insights that can inform strategic choices and improve the effectiveness of loan management programs.

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?

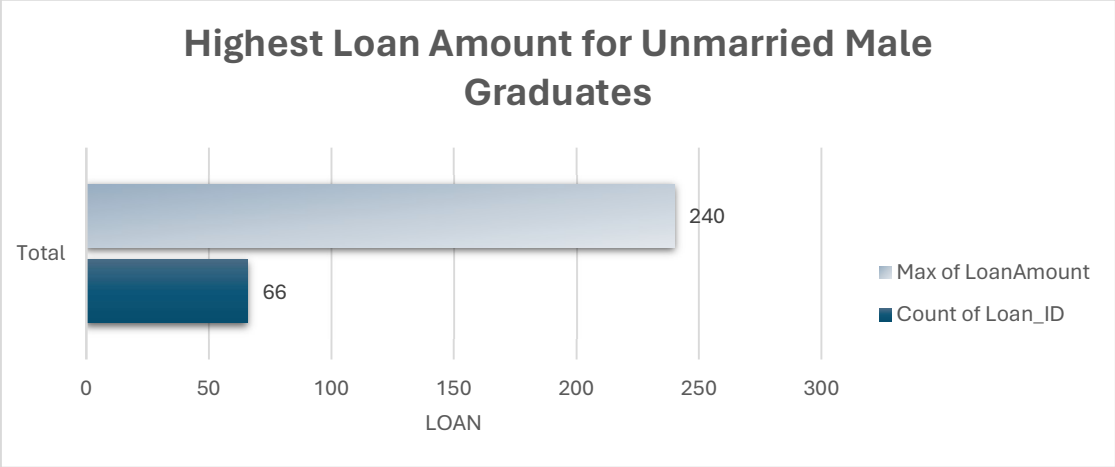
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?

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.

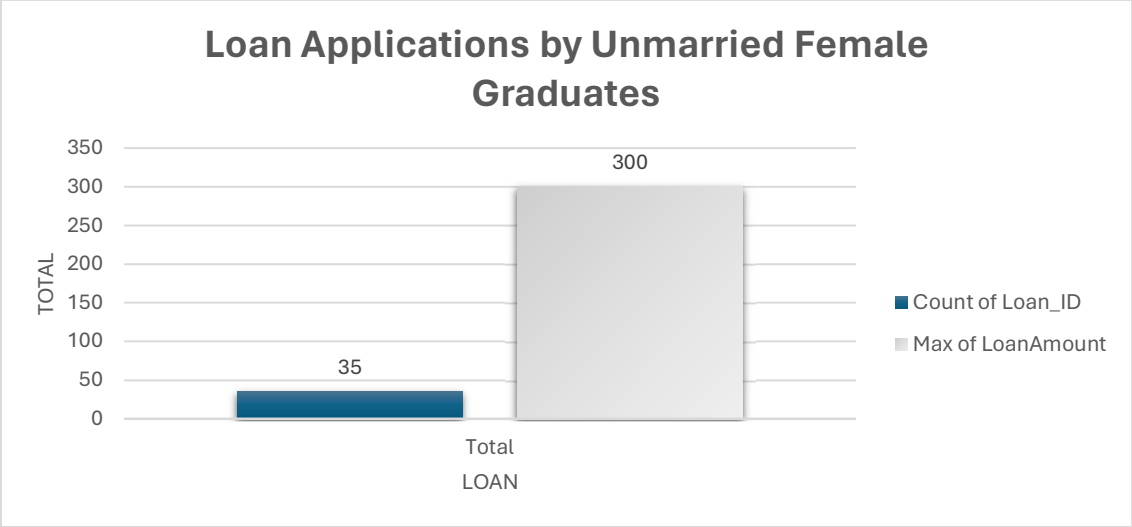
Analytics

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



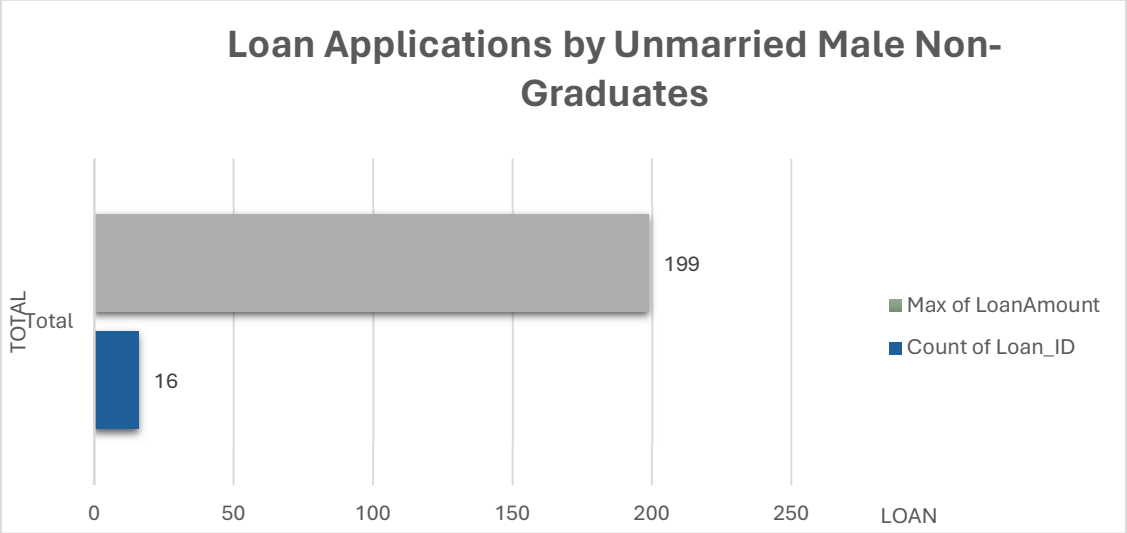
This analysis shows the no. of male graduates applied for the loan and are not married with the highest amount. As of analysed the total no. of loan applied is 66 and max loan amount is 240.

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



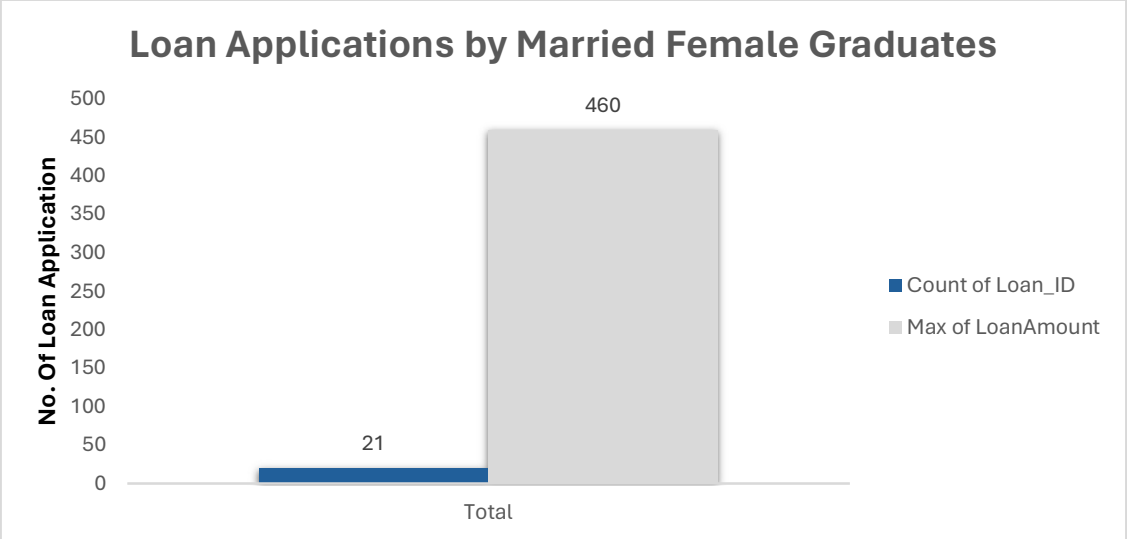
According to this data, the greatest number of female graduates who are single sought for loans. As of now, there have been 35 total loan applications, with a maximum loan amount of $300.

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



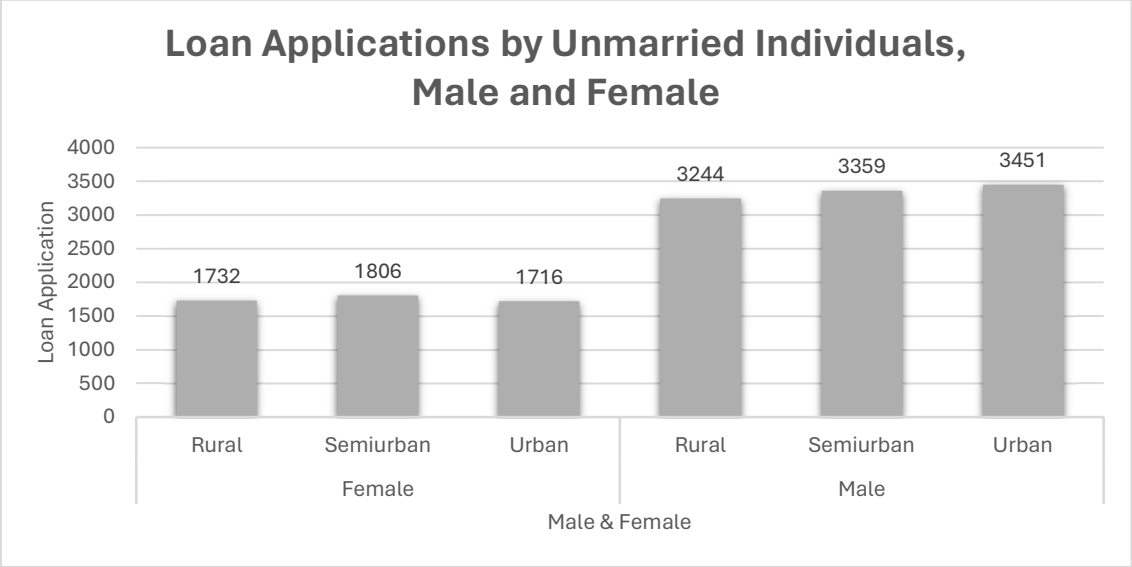
This research reveals the number of unmarried male non-graduates who asked for loans and the greatest amount they were denied. As of now, there have been 16 total loan applications, with a maximum loan amount of 199.

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



According to this data, the greatest number of female graduates who are single sought for loans. As of now, there have been 21 total loan applications, with a maximum loan amount of $460.

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



This research compares unmarried male and female applicants for loans in rural, semi-urban, and metropolitan areas; the number of applications for loans is much larger in males than in females.

Loan counts for women are as follows: women's (1732), semi urban (1806), and urban (1716); men's (3244), semi urban (3359), and urban (3451).

Conclusion and Review

The data shows glaring differences in loan applications based on gender. The application pool was dominated by single male grads, then single female graduates. Though in lower percentages, married female grads and unmarried male graduates also asked for loans. Interestingly, in rural, semi-urban, and urban regions, the number of men was far more than that of girls.

The research offers insightful information on borrower demographics and successfully depicts patterns in loan applications depending on gender. It is advised to carry out more research on the variables impacting loan choices and to improve the data presentation through visual improvements. In general, the paper provides a basis for comprehending loan dynamics, with need for further analysis.

Regression

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 2	0.005807787	0.000627861	9.250122	5.49E-18	0.004571955	0.007043619	0.004572	0.007044
X Variable 3	0.006772797	0.001264765	5.354983	1.76E-07	0.004283331	0.009262263	0.004283	0.009262

The provided summary output corresponds to a regression analysis conducted on a dataset comprising 289 observations. The model's overall performance is moderate, as indicated by a multiple R of approximately 0.531. The coefficient of determination (R-squared) of 0.282 suggests that around 28% of the variability in the dependent variable can be explained by the independent variables. The ANOVA table shows that the regression model is significant, with an F-value of 37.32 and a very low p-value, indicating that the model's explanatory power is significant. The coefficients table displays the intercept and coefficients for three predictor variables (X Variable 1, X Variable 2, and X Variable 3). All three predictor variables exhibit statistically significant relationships with the dependent variable, as their p-values are very low. Specifically, for every unit increase in X Variable 1, the dependent variable increases by approximately 0.096 units. Similarly, a unit increase in X Variable 2 and X Variable 3 results in approximately 0.006 and 0.007 units increase in the dependent variable, respectively. Overall, these findings suggest that the predictors significantly contribute to explaining the variability in the dependent variable.

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				

This single-factor ANOVA examines the impact of a categorical factor, represented by two groups (Column 1 and Column 2), on a continuous response variable. The summary statistics provide counts, sums, averages, and variances for each group. The ANOVA table indicates sources of variation: between groups and within groups. Between groups, the sum of squares (SS) is approximately 6.12 million, with a mean square (MS) of 6.12 million, and a highly significant F-value and p-value, suggesting significant differences between group means. Within groups, the SS is around 2.27 million, reflecting variability within the groups. The total SS is approximately 8.39 million. These results imply that the categorical factor significantly influences the variation in the response variable, as indicated by the large F-value

Anova: two factors

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				

This two-factor ANOVA without replication explores the effects of two categorical factors, represented by rows and columns, on a continuous response variable. The summary statistics provide counts, sums, averages, and variances for each level of the rows and columns. The ANOVA table indicates sources of variation: rows, columns, and error. For rows, the sum of squares (SS) is approximately 1.26 million, with a mean square (MS) of 4391.038 and a significant F-value and p-value, suggesting differences between row means. For columns, the SS is substantially higher at around 6.12 million, with a highly significant F-value and a very low p-value, indicating significant differences between column means. The error term, representing variability within cells, has an SS of approximately 1.00 million. Overall, both

row and column factors significantly influence the variation in the response variable, as indicated by their respective F-values and p-values.

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

This table provides descriptive statistics for four variables: Column1, Column2, Column3, and Column4. Each variable's statistics are listed across rows, including measures like mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, and count. For instance, Column1 has a mean of approximately 342.671, with a standard deviation of 65.655, indicating variability around the mean. Column2, however, exhibits a significantly higher mean of approximately 4637.353 and a much larger standard deviation of 4790.684, suggesting substantial variability in the data. The kurtosis values indicate the peaked Ness or flatness of the distribution, with Column2 showing extremely high kurtosis compared to the other columns. Similarly, Column3 and Column4 display their respective characteristics, such as skewness, range, and distribution shape. Overall, these statistics offer insights into the distribution, central tendency, and variability of each variable, aiding in understanding their characteristics within the dataset.

Correlation

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

The provided table represents a correlation matrix between three variables: Column 1, Column 2, and Column 3. Each cell in the matrix displays the correlation coefficient between two variables. The diagonal elements, where a variable correlates with itself, are all 1, as expected. The off-diagonal elements indicate the correlation between different pairs of variables. In this case, the correlation coefficient between Column 1 and Column 2 is approximately -0.084, suggesting a weak negative correlation. Between Column 1 and Column 3, the correlation is approximately 0.446, indicating a moderate positive correlation. Column 2 and Column 3 exhibit a correlation coefficient of approximately 0.230, suggesting a weak positive correlation between these two variables. Overall, this matrix provides insights into the relationships between the variables, with varying degrees of correlation strength observed among them.

Shop Sales Data Report

Introduction

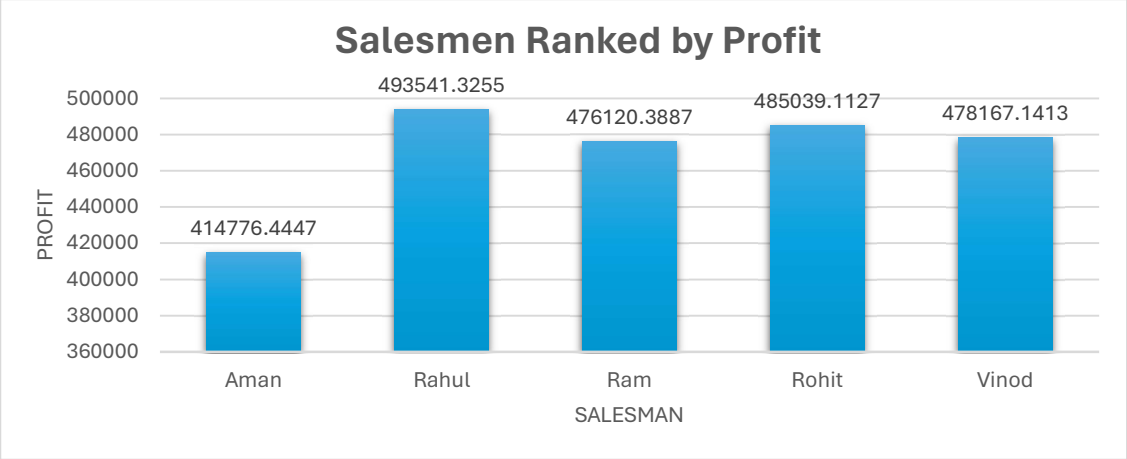
This report examines a large sales dataset with an emphasis on sales performance analysis and product trends among sales representatives. The collection includes features including product specifications, sales volumes, earnings, and salesman details. Finding information that can improve corporate performance and guide the creation of sales strategies is the main goal of this investigation. The report's objectives are to identify top-performing salespeople, analyze product popularity, and comprehend sales patterns by looking at sales data over a certain period of time and comparing product performance. The analysis's conclusions will be of great use to CEOs, marketing specialists, and sales managers who want to boost income, improve sales tactics, and expand their companies. Our goal in doing this study is to offer practical insights that will help inform decisions and advance the performance of the organization as a whole.

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.

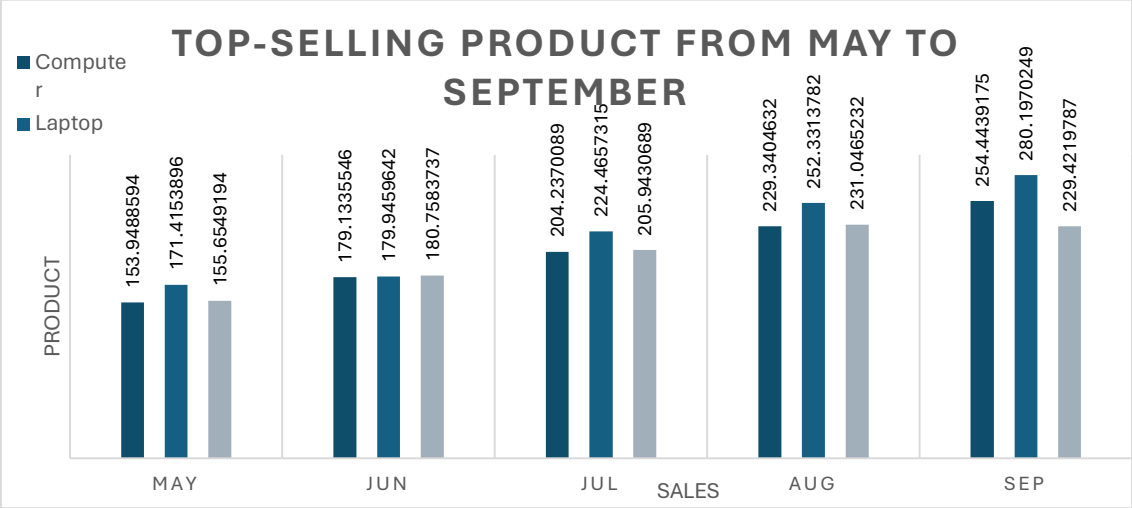
Analytics

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



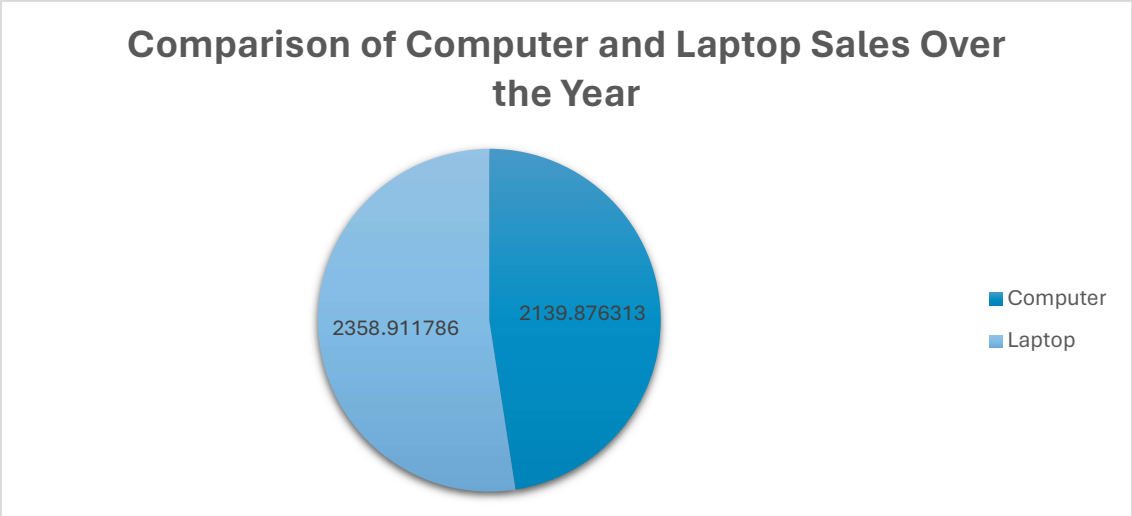
When all of the salesmen are compared based on profit made, as seen by the line chart, Rahul has the most profit earned, valued at 493541.3255.

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



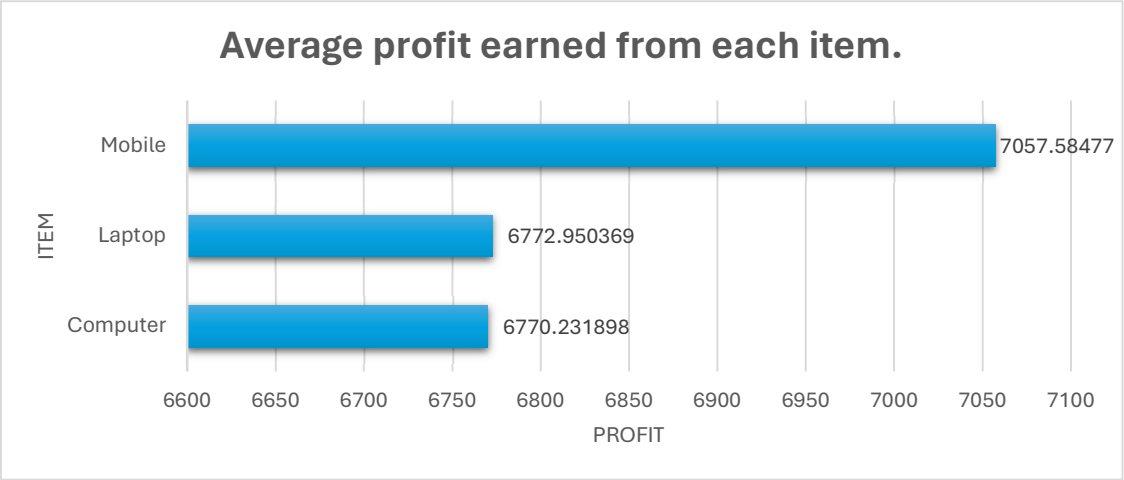
We would need to examine the sales data throughout that time period in order to determine which product sold the most during the months of May through September. When the quantity sold for each product is added up for all transactions made within this time frame, the laptop is the most sold product from May to September, with the highest sales occurring in September, totaling 280.1970249.

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



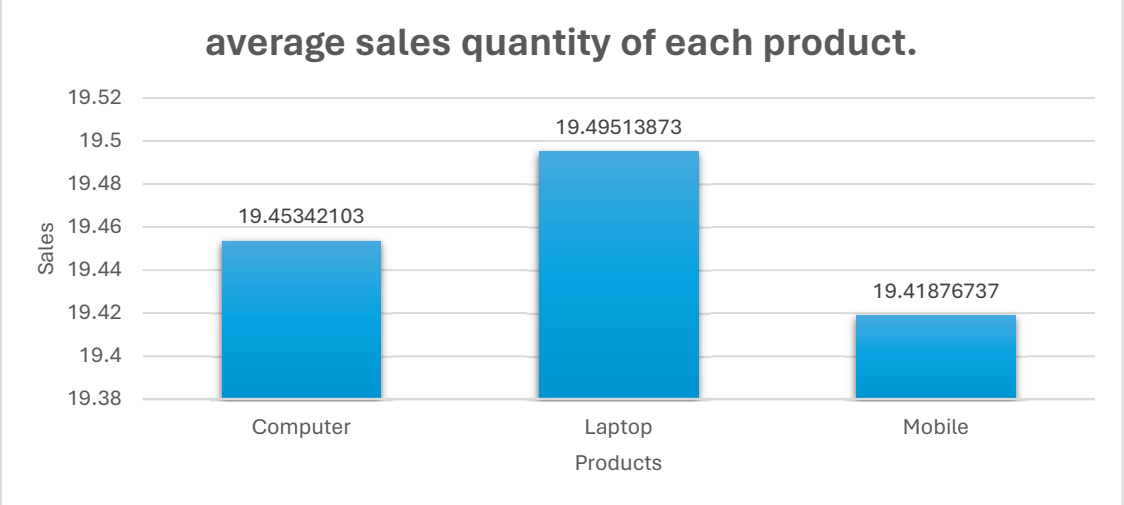
The two products that sold the most throughout the course of the year were the laptop and the computer, with the laptop having the higher sales quantity at 2358.911786 and the computer at 2139.876313.

4. Which item yield most average profit?



According to this data, the mobile device has the highest average profit made (7057.58477) when compared to the laptop and computer.

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



According to the analysis, the average sales amount of laptops (19.49513873) is larger than that of computers (19.45342103) and mobile phones (19.41876737).

Conclusion and Review:

Important information about sales effectiveness and product trends among salespeople is revealed by the analysis. Outperforming every other salesman and making the biggest profit,

Rahul comes out on top. Furthermore, the laptop is the most popular product from May to September, with September seeing the biggest sales. In terms of units sold over the course of the year, laptops do better than PCs. In addition, out of smartphones, laptops, and PCs, mobile phones have the greatest average profit. Finally, in terms of average sales quantity, laptops outperform PCs and mobile devices.

The study successfully draws attention to product trends and sales performance, offering insightful information for improving sales strategy. Visualizations help in comprehending popular products and long-term patterns. Deeper understanding of the variables affecting product preferences and sales variations, however, could improve the analysis. All things considered, the research provides useful information for enhancing sales tactics and increasing profits.

Regression

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+	1.37E+	3448	4.6E-180			
Residual	340	1.35E+	396975					
Total	341	1.5E+0						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2068.993161	88.47952	23.38387	9.14E-73	1894.957	2243.029	1894.957	2243.029
X Variable 1	246.4655683	4.196812	58.72686	4.6E-180	238.2106	254.7206	238.2106	254.7206

The regression analysis conducted on a dataset comprising 342 observations reveals a strong positive correlation between the predictor variable (X Variable 1) and the dependent variable, with a multiple R-value of 0.954. The coefficient of determination (R-squared) is 0.910, indicating that approximately 91% of the variability in the dependent variable can be explained by the predictor. The ANOVA results demonstrate a highly significant regression model, with a large F-value of 3448 and an extremely low p-value, suggesting that the model's explanatory power is significant. The coefficients table displays the intercept and coefficient for the predictor variable. Both the intercept and the coefficient for X Variable 1

are statistically significant, with respective values of 2068.99 and 246.47. These findings suggest that the predictor variable significantly contributes to explaining the variability in the dependent.

Correlation

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

The provided table represents a correlation matrix between two variables: Column 1 and Column 2. Each cell in the matrix displays the correlation coefficient between two variables. The diagonal elements, where a variable correlates with itself, are all 1, as expected. The off-diagonal element indicates the correlation between the two variables. In this case, the correlation coefficient between Column 1 and Column 2 is approximately 0.954, indicating a strong positive correlation between them. This implies that as values in Column 1 increase, values in Column 2 tend to increase as well, and vice versa. Overall, this matrix provides insight into the relationship between the two variables, suggesting a strong positive correlation between Column 1 and Column 2.

Anova (Single Factor):

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				

This single-factor ANOVA investigates the impact of a categorical factor, represented by two groups (Column 1 and Column 2), on a continuous response variable. The summary statistics provide counts, sums, averages, and variances for each group. The ANOVA table indicates the sources of variation: between groups and within groups. The between groups sum of squares (SS) is approximately 8.01E+09, with a mean square (MS) of 8.01E+09, and a highly significant F-value and p-value, suggesting significant differences between group means. The within groups SS is around 1.5E+09, reflecting variability within the groups. The total SS is approximately 9.52E+09. These results imply that the categorical factor significantly

influences the variation in the response variable, as indicated by the large F-value and significant p-value.

Anova two factor:

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				

This two-factor ANOVA without replication examines the effects of two categorical factors, represented by rows and columns, on a continuous response variable. The summary statistics provide counts, sums, averages, and variances for each level of the rows and columns. The ANOVA table presents the sources of variation: rows, columns, and error. For the rows, the sum of squares (SS) is approximately 7.58E+08, with a mean square (MS) of 2221714, and a non-significant F-value and p-value, indicating no significant difference between row means. However, for the columns, the SS is substantially higher at 8.01E+09, with a highly significant F-value and a very low p-value, suggesting significant differences between column means. The error term represents variability within cells and has an SS of approximately 7.46E+08. The total SS is approximately 9.52E+09. Overall, while there is no significant difference between row means, the column factor significantly influences the variation in the response variable.

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

The provided table presents descriptive statistics for two variables: Column1 and Column2. For Column1, the mean is approximately 19.457, with a standard deviation of 8.130, indicating relatively low variability around the mean. Column1 also exhibits a kurtosis value close to -1, suggesting a slightly flatter distribution than the normal distribution. In contrast, Column2 has a much higher mean of approximately 6864.457 and a significantly larger standard deviation of 2100.186, indicating greater variability in the data. The kurtosis value for Column2 is also negative, indicating a slightly flatter distribution. Additionally, Column2 has a wider range, ranging from 1000 to 10279.85, compared to Column1's range of 3 to 33.30852. Overall, these statistics provide a comprehensive summary of the distribution, central tendency, and variability of each variable, aiding in understanding their characteristics and potential relationships in the dataset.

Sales Data Sample Report

Introduction

A large sales dataset with variables like ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and SALES is analyzed in this report. It seeks to draw conclusions that will direct sales tactics and improve corporate performance. Sales managers, marketers, and executives looking to increase revenue and enhance sales processes are among the intended audience members. Important studies include comparing the sales of classic and vintage automobiles, figuring out average sales, figuring out what items are best-selling, analyzing the profit margin by nation for particular product lines, comparing sales over time, and analyzing countries according to the amount of deals. The research seeks to offer practical insights for boosting sales growth and enhancing overall business outcomes through these assessments.

The project's scope includes analyzing a sizable sales dataset in order to glean insightful information that might improve product offers, guide sales methods, and boost overall business performance. The project will be valuable to analysts and researchers who are looking for insights on market trends and sales dynamics.

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.

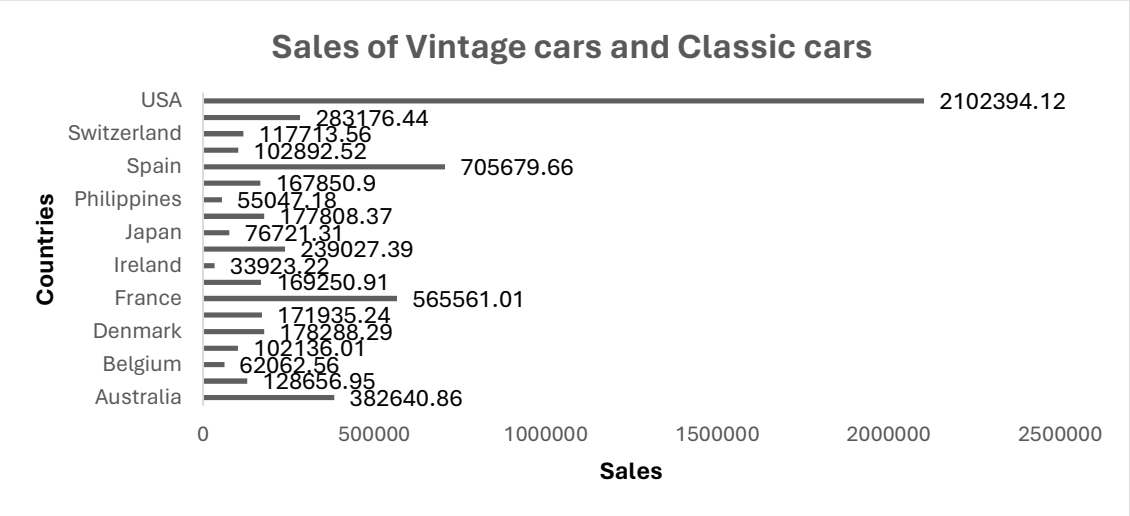
3. Assessment of the country yielding the most profit for Motorcycles, Trucks, and Buses.

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

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

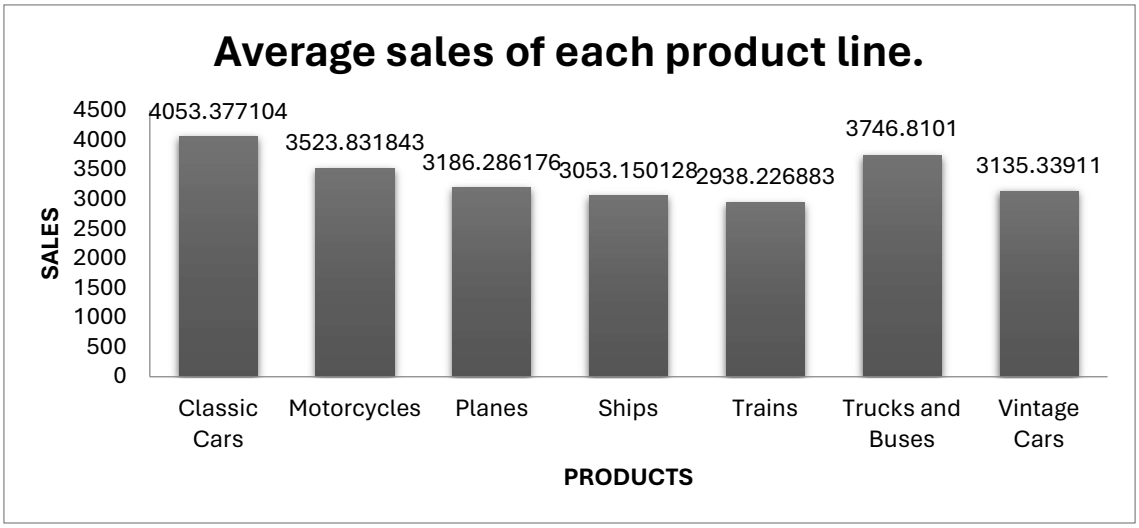
Analytics

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



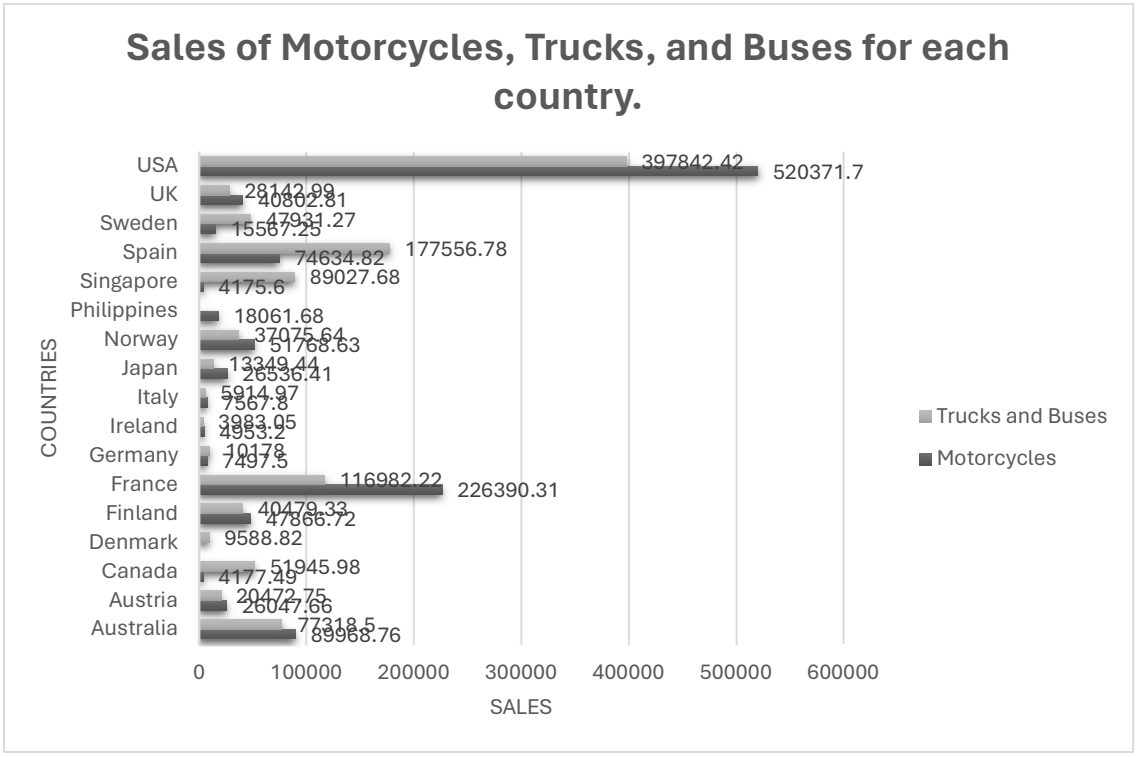
This analysis Compare the sale of Vintage cars and Classic cars for all the countries. Where USA (2102394.02) has the highest sales followed by Spain, France, and Australia.

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



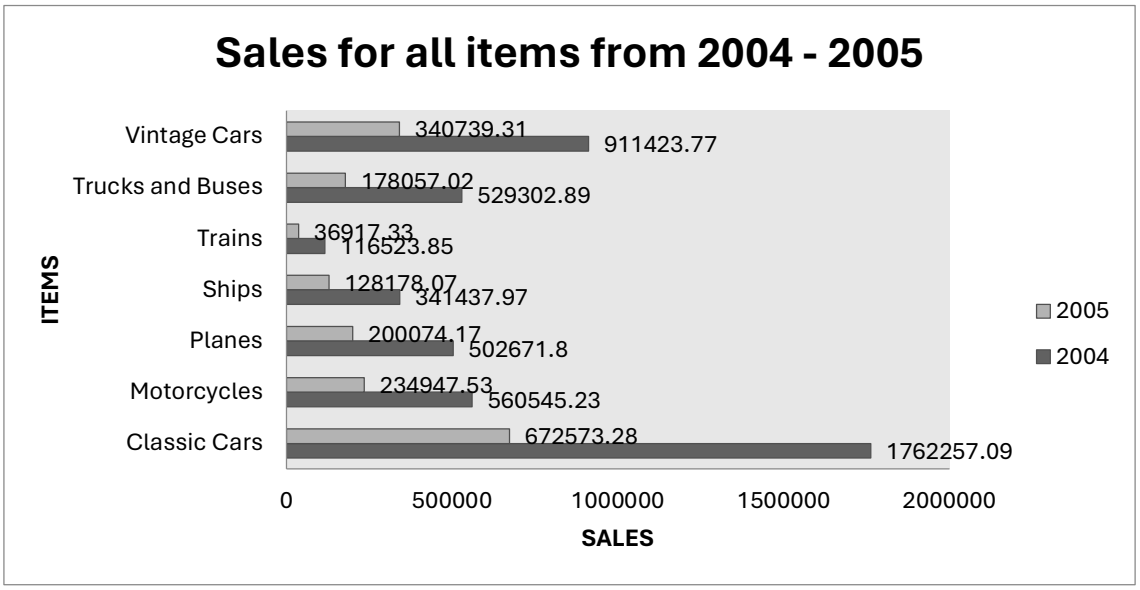
The average sales of every product and the top-selling product are the two goals of this investigation. Additionally, the graph shows that, with an average sale of 405.377104, Classic Cars have the greatest sales, followed by Trucks and Buses and Motorcycles.

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



The goal of this analysis is to determine which nation makes the most money from trucks, buses, and motorcycles. According to a bar graph, the USA leads the world in motorcycle sales with 520371.7, followed by France and Spain, and the world in truck and bus sales with 397842.42.

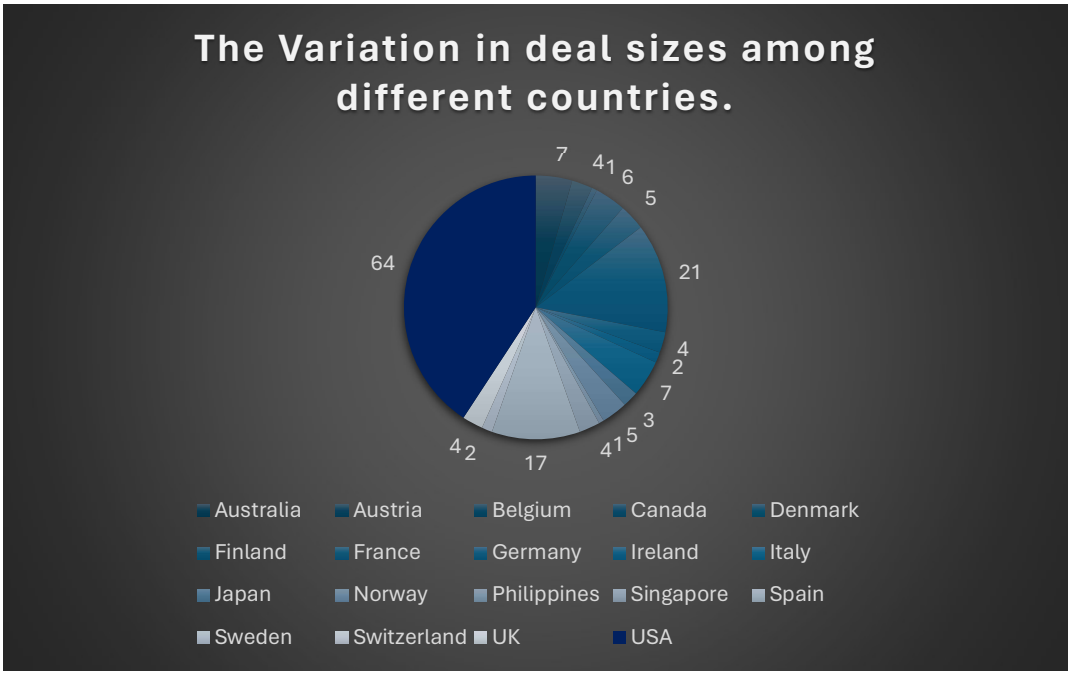
4. Comparison of sales for all items across the years 2004 and 2005.



The goal of this analysis is to compare the sales of every item in the years 2004 and 2005. The line chart shows that sales of every item are changing at a very rapid rate, with the

exception of classic cars, which had the highest sales of any category in both years, with 1762257.09 in 2004 and 672573.28 in 20

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



The purpose of this research is to determine how deal sizes are distributed among the various nations. Additionally, the bar chart demonstrates how much larger deals are made in the USA than in every other country, with huge deals valued at 64, medium deals at 505, and small deals at 435.

Conclusion and Review

The analysis provides valuable insights into sales trends and profitability by category and by country. The USA comes out on top as a market leader in Vintage & Classic cars, in Trucks, in Buses, and in Motorcycles. Classic Cars are the top-selling product, accounting for a significant portion of total sales revenue. In addition, the USA shows exceptional profitability, especially in the Trucks & Buses & Motorcycles categories. Sales for Classic cars remain strong throughout 2004 and 2005, showing that there is a continuing demand for this product category. Also, the USA shows significantly larger deal sizes than other countries, demonstrating its dominance in terms of sales volume.

While the analysis provides visualizations of key findings, more in-depth analysis into the drivers of sales volatility and deal size differences could yield more insightful results. All in all, the report provides valuable insights to optimize sales strategies and accelerate business growth.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.877178
R Square	0.769441
Adjusted R Square	0.766629
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

The regression analysis conducted on a dataset consisting of 250 observations indicates a multiple R-value of 0.877, suggesting a strong positive correlation between the predictors and the dependent variable. The R-squared value of 0.769 implies that approximately 77% of the variability in the dependent variable can be explained by the predictors. The ANOVA results reveal a highly significant regression model, with a large F-value of 273.66 and a very low p-value, indicating that the model's explanatory power is significant. The coefficients table displays the intercept and coefficients for three predictor variables. All predictor variables, X Variable 1, X Variable 2, and X Variable 3, exhibit statistically significant relationships with the dependent variable, with respective coefficients of 103.08, 12.82, and 47.43. These findings suggest that the predictors contribute significantly to explaining the variability in the dependent variable.

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				

This single-factor ANOVA investigates the impact of a categorical factor, represented by two columns, on a continuous response variable. The summary statistics provide counts, sums, averages, and variances for each level of the factor. The ANOVA table partitions the variation into between groups and within groups. Between groups, the sum of squares (SS) is approximately 1.54E+09, with a mean square (MS) of 1.54E+09 and a highly significant F-value and p-value, indicating significant differences between group means. Within groups, the SS is around 8.58E+08, reflecting variability within the groups. The total SS is approximately 2.4E+09. These results suggest that the categorical factor significantly influences the variation in the response variable, as indicated by the large F-value and significant p-value.

Anova: two factors:

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				

This two-factor ANOVA without replication examines the effects of two categorical factors, rows and columns, on a continuous response variable. The summary statistics present counts, sums, averages, and variances for each level of the rows and columns. The ANOVA table delineates the sources of variation: rows, columns, and error. For the rows, the sum of squares (SS) is approximately 2.95E+08, with a mean square (MS) of 1182944 and a non-significant F-value and p-value, indicating no significant difference between row means. However, for the columns, the SS is substantially higher at 2.09E+09, with a highly significant F-value and a very low p-value, suggesting significant differences between column means. The error term represents variability within cells and has an SS of approximately 5.64E+08

Descriptive Statistics:

Column1		*Column2*		*Column3*		*Column4*	
Mean	34.636	Mean	3613.123	Mean	102.136	Mean	84.45296
Standard Error	0.59898	Standard Error	117.392	Standard Error	2.58035	Standard Error	1.279453
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	1.127057	Kurtosis	-0.19836	Kurtosis	-0.40344
Skewness	0.256745	Skewness	1.013489	Skewness	0.517104	Skewness	-0.9678
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

This table presents descriptive statistics for four variables: Column1, Column2, Column3, and Column4. Each variable's statistics are listed across rows, including measures like mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, and count. For instance, Column1 has a mean of approximately 34.636 and a standard deviation of 9.471, indicating variability around the mean. Column2, with a mean of approximately 3613.123 and a standard deviation of 1856.131, shows much higher variability compared to the other columns. Column3 and Column4 display means of around 102.136 and 84.453, respectively, with moderate standard deviations. The skewness and kurtosis values provide insights into the distribution's shape and tail behavior for each variable. Overall, these statistics offer a comprehensive summary of the distribution, central tendency, and variability of each variable, aiding in understanding their characteristics and potential relationships in the dataset.

Correlation:

	Column 1	*Column 2*	*Column 3*
Column	1		

1			
Column 2	0.513951	1	
Column 3	-0.01254	0.663973	1

The provided table represents a correlation matrix between three variables: Column 1, Column 2, and Column 3. Each cell in the matrix displays the correlation coefficient between two variables. The diagonal elements, where a variable correlates with itself, are all 1, as expected. The off-diagonal elements indicate the correlation between different pairs of variables. For instance, the correlation between Column 1 and Column 2 is approximately 0.514, suggesting a moderate positive correlation. Similarly, the correlation between Column 1 and Column 3 is close to zero at -0.013, indicating a weak negative correlation. Column 2 and Column 3 exhibit a correlation of approximately 0.664, indicating a moderate positive correlation between these two variables. Overall, this matrix provides insight into the relationships between the variables, with varying degrees of correlation strength observed among them.

Store Dataset Report

Introduction

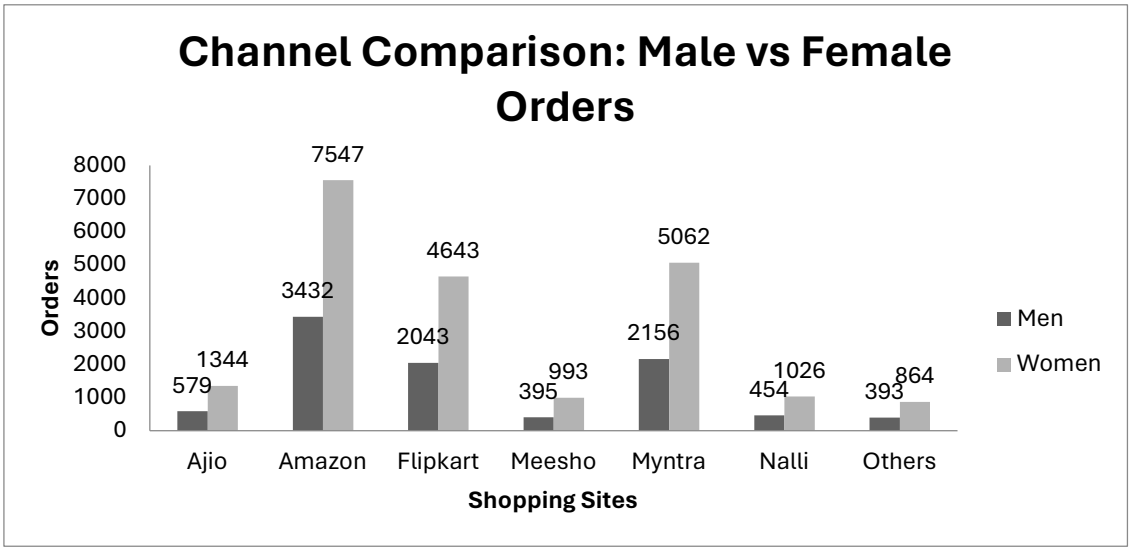
This dataset contains sales data from a retail store. It includes things like gender, age, transaction details (order ID, status), product details (category, SKU) and shipping details. Our goal is to help you understand how your customers interact with your products and how they interact with your products. We look for patterns, preferences and correlations within your data. With these insights, you can improve your marketing, manage your inventory and increase your customer satisfaction.

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.

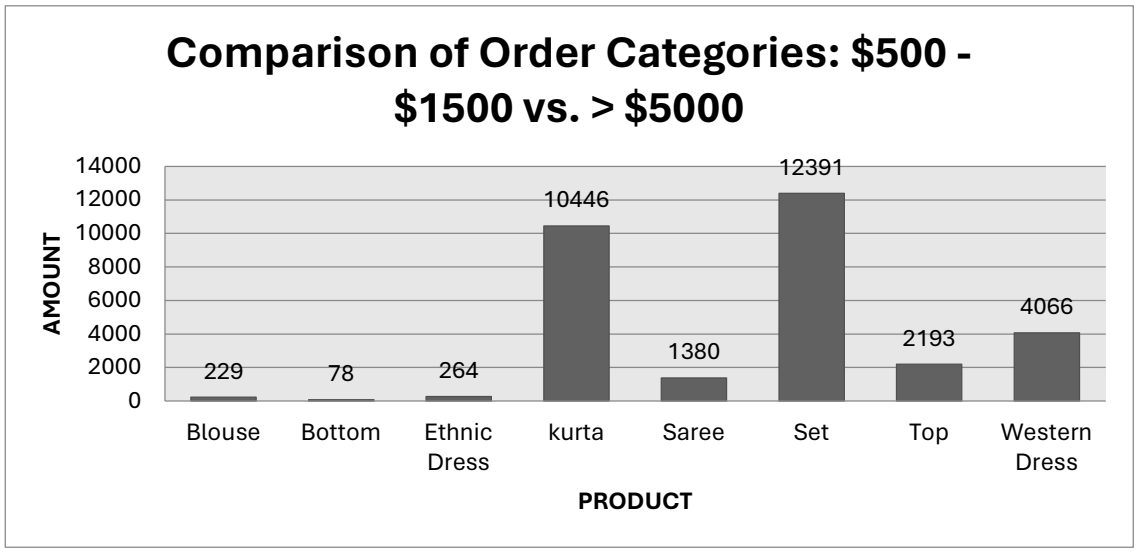
Analytics

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



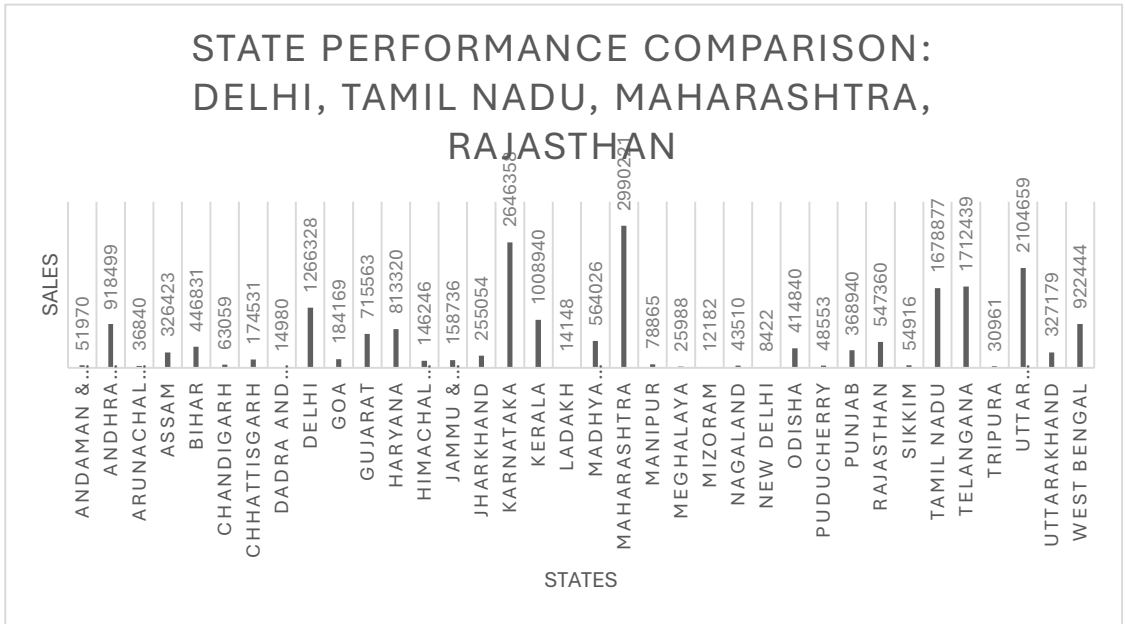
Sales for both men and women are led by Amazon, which is followed by Myntra and Flipkart. Nearly 3432 units were sold by Amazon in the men's category, and nearly 7547 units in the women's category. 5062 units were sold in the women's area of Myntra, and 2156 units in the men.

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



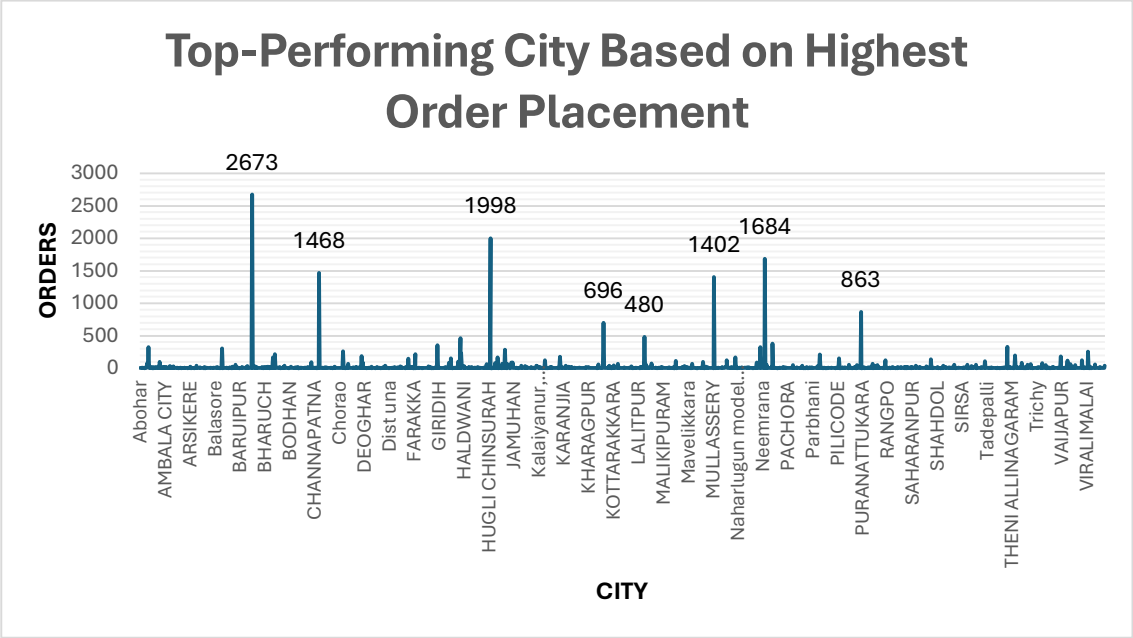
Comparing the order categories where the quantity is less than 1500 and more than 5000 is made easier by this analysis. displaying the set (12391) and kurta (10446) with the greatest order count, followed by the saree, top, and western attire.

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



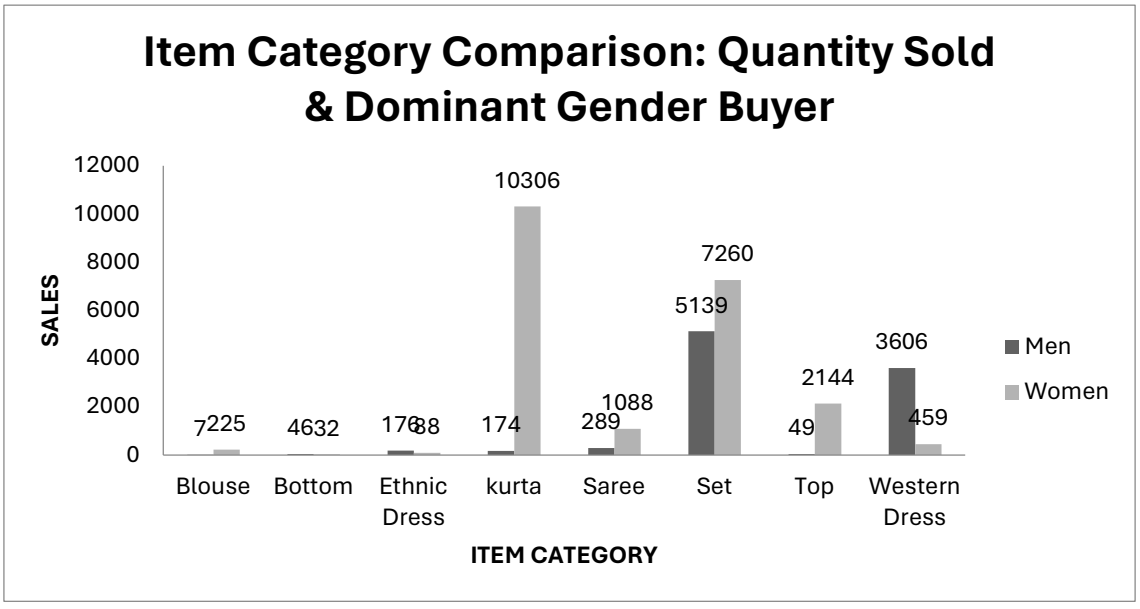
Karnataka (2646358) had the best performance among the states, followed by Uttar Pradesh (2104659). This research reveals which states fared better than the states indicated above.

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



Bengaluru had the largest order put with 2673 orders, followed by Hyderabad (1998). Based on the graph recorded, we can really observe which city fared better than the other cities based on biggest order placed.

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



The kurta purchased by women is the most popular category of things, followed by men's purchases, and western clothing is the most popular item for both men and women. This report compares these different product categories based on sales volume.

Conclusion and Review

Amazon leads in sales for both men and women, according to the research, with Myntra and Flipkart trailing closely after. Sales for both men's and women's categories are led by Amazon, which is followed by Myntra and Flipkart. Kurtas and sets are among the best-selling products; Karnataka and Bangalore have the best sales figures.

Retailers may make better decisions thanks to the study, which offers insightful information about regional performance and sales patterns. Nonetheless, the analysis may be improved by looking at other variables that affect sales. All things considered, the results provide insightful knowledge for maximizing sales tactics in cutthroat marketplaces.

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

The regression analysis conducted on the dataset comprising 31,047 observations indicates a multiple R-value of 0.172, suggesting a weak positive correlation between the predictors and the dependent variable. The R-squared value of 0.030 implies that only about 3% of the variability in the dependent variable can be explained by the predictors. The ANOVA results reveal a significant regression model, with a large F-value of 475.46 and a corresponding p-value of 0. Additionally, the coefficients table displays the intercept and coefficients for two predictor variables. Both X Variable 1 and X Variable 2 exhibit statistically significant

relationships with the dependent variable, with respective coefficients of 0.048 and 492.028. These findings suggest that while there is a statistically significant relationship between the predictors and the dependent variable, the overall explanatory power of the model remains relatively low.

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		

This single-factor ANOVA examines the impact of a categorical factor, represented by two columns, on a continuous response variable. The summary statistics reveal that Column 1 has a mean of approximately 1.006 and a very small variance, while Column 2 exhibits a significantly higher mean of about 682.075 with a larger variance. The ANOVA results indicate a highly significant difference between the groups, as reflected in the large F-value of 199639.8 and the associated p-value. The majority of the variability lies between groups rather than within them, as evidenced by the substantial sum of squares for between groups compared to within groups. Overall, this analysis suggests that the categorical factor represented by the two columns significantly influences the variation in the response variable.

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				

This two-factor ANOVA without replication analyses the effects of two categorical factors, rows and columns, on a continuous response variable. The rows show no significant difference in means, as indicated by their non-significant F-value and p-value. However, the columns demonstrate a highly significant difference, with an extremely high F-value and a near-zero p-value, suggesting substantial variability between column means. The error term represents variability within cells and is relatively moderate. In conclusion, while the row factor does not significantly affect the response variable, the column factor has a considerable impact, indicating that the variable represented by the columns plays a crucial role in explaining the variability in the response variable.

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

This table presents descriptive statistics for three variables: Column 1, Column 2, and Column 3. Each variable's statistics are listed across rows, including measures like mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, and count. For Column 1, the mean is approximately 39.50, with a standard deviation of 15.12, indicating variability around the mean. Column 2 has a mean close to 1, with much lower variability compared to Column 1, as indicated by its smaller

standard deviation of 0.094. Column 3 has a much higher mean of approximately 682, with a large standard deviation of 268.58, suggesting significant variability in the data. Additionally, the skewness and kurtosis values provide insights into the distribution's shape and tail behavior. Overall, these statistics offer a concise summary of the distribution, central tendency, and variability of each variable, aiding in understanding their characteristics and potential relationships in the dataset.

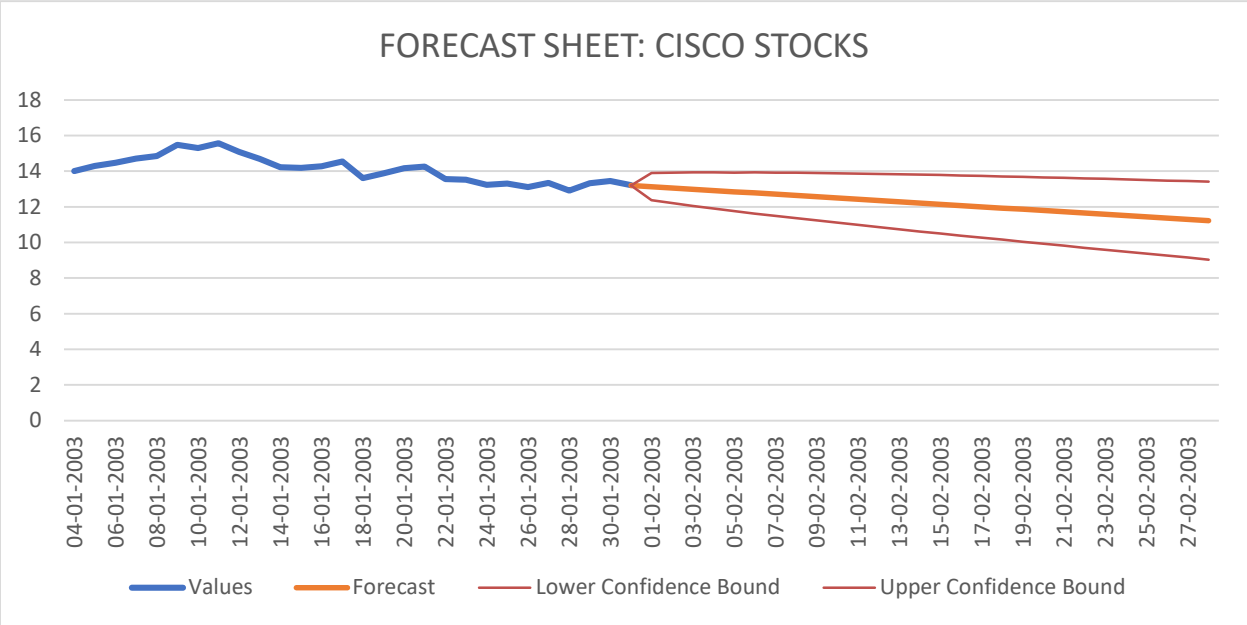
Correlation:

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

This table appears to represent a correlation matrix, where each cell shows the correlation coefficient between two variables. In this case, Column 1 has a perfect correlation with itself (1.0), as expected. Column 2 and Column 3 also have correlations with themselves of 1.0. The off-diagonal elements represent the correlation between different variables. For example, the correlation between Column 1 and Column 2 is approximately 0.0049, indicating a very weak positive correlation. Similarly, the correlation between Column 1 and Column 3 is approximately 0.0035, also very weak. However, the correlation between Column 2 and Column 3 is stronger, approximately 0.1724, suggesting a moderate positive correlation between these two variables. Overall, this matrix provides insight into the relationships between the variables, with most correlations being very weak except for the moderate correlation between Column 2 and Column 3.

Forecast Sheet: CISCO STOCKS

Date	Open	High	Low	Close	Volume	Symbol	YTD Gains
1/1/2003	13.11	13.69	13.09	13.64	61335700	CSCO	0.041221
1/2/2003	13.11	13.69	13.09	13.64	61335700	CSCO	0.041221
1/3/2003	13.58	13.96	13.56	13.91	50891700	CSCO	0.061832
1/4/2003	14.01	14.42	13.98	14.2	58936700	CSCO	0.083969
1/5/2003	14.3	14.7	14.24	14.6	83998600	CSCO	0.114504
1/6/2003	14.48	14.75	14.37	14.44	75927000	CSCO	0.10229
1/7/2003	14.7	15.11	14.65	14.95	75284400	CSCO	0.141221
1/8/2003	14.84	15.46	14.83	15.22	91193900	CSCO	0.161832
1/9/2003	15.47	15.52	15.04	15.28	66314800	CSCO	0.166412
1/10/2003	15.3	15.63	15.29	15.58	69977900	CSCO	0.189313
1/11/2003	15.57	15.63	15.13	15.18	61992300	CSCO	0.158779
1/12/2003	15.08	15.3	14.79	14.9	65668200	CSCO	0.137405
1/13/2003	14.69	14.72	14.05	14.13	81310000	CSCO	0.078626
1/14/2003	14.22	14.5	14.15	14.18	62930800	CSCO	0.082443
1/15/2003	14.19	14.38	13.9	13.96	64965500	CSCO	0.065649
1/16/2003	14.28	14.75	14.12	14.59	62491200	CSCO	0.11374
1/17/2003	14.55	14.56	13.8	13.86	70564100	CSCO	0.058015
1/18/2003	13.61	14.07	13.56	13.71	58554700	CSCO	0.046565
1/19/2003	13.88	14.33	13.8	14.22	64116500	CSCO	0.085496
1/20/2003	14.17	14.17	13.79	14.08	71861700	CSCO	0.074809
1/21/2003	14.25	14.36	13.83	13.87	68226500	CSCO	0.058779
1/22/2003	13.56	13.74	13.16	13.37	1.03E+08	CSCO	0.020611
1/23/2003	13.52	13.8	13.38	13.48	65976600	CSCO	0.029008
1/24/2003	13.24	13.24	12.87	13.2	1.11E+08	CSCO	0.007634
1/25/2003	13.31	13.6	13.1	13.2	1.15E+08	CSCO	0.007634
1/26/2003	13.11	13.41	13.07	13.24	58738900	CSCO	0.010687
1/27/2003	13.34	13.44	12.66	12.85	69851700	CSCO	-0.01908
1/28/2003	12.92	13.24	12.78	13.15	55955700	CSCO	0.003817
1/29/2003	13.32	13.6	13.26	13.47	71241800	CSCO	0.028244
1/30/2003	13.44	13.65	13.19	13.2	58732000	CSCO	0.007634
1/31/2003	13.21	13.4	13.1	13.31	51646500	CSCO	0.016031



The forecast sheet provides a comprehensive view of Cisco's historical stock performance and projected future trends based on statistical modeling and data analysis. Here's a more detailed explanation:

Historical Stock Prices (Blue Line): The blue line represents Cisco's actual closing stock prices from around 2003 to 2023. It shows a period of steady growth until 2007, followed by a significant downturn during the global financial crisis of 2008-2009. After a recovery phase, the stock experienced another dip around 2011-2012. From 2013 onwards, Cisco's stock prices exhibited a generally upward trajectory, albeit with some fluctuations.

Forecast (Orange Line): The orange line depicts the forecasted future stock prices for Cisco from 2023 to 2030, based on statistical forecasting models and techniques such as time series analysis, regression analysis, or machine learning algorithms. These models analyze historical data patterns, market trends, macroeconomic factors, and company-specific information to generate forecasts.

The forecasted values show an overall upward trend, suggesting that analysts expect Cisco's stock prices to continue rising in the coming years. This positive outlook could be driven by factors such as Cisco's strong market position, product innovations, favorable industry trends (e.g., growing demand for networking and cybersecurity solutions), or anticipated growth in revenue and profitability.

Confidence Intervals (Green and Red Lines): The green and red lines represent the lower and upper confidence bounds, respectively, for the forecasted stock prices. These bounds are calculated using statistical methods and provide a range within which the actual future stock prices are expected to fall with a certain level of confidence, typically 95%.

The width of the confidence interval reflects the degree of uncertainty associated with the forecast. A wider interval indicates higher uncertainty, while a narrower interval suggests greater confidence in the forecasted values.

It's important to note that stock forecasts are inherently uncertain and subject to various risk factors, such as market volatility, economic conditions, competitive landscape, regulatory changes, and company-specific events. Analysts and investors use these forecasts as guidance, but actual stock prices may deviate from the predicted values due to unforeseen circumstances.

Forecasting models are regularly updated and refined as new data becomes available, allowing for more accurate predictions over time. However, it's essential to view these forecasts as estimates rather than definitive outcomes and to consider them alongside other sources of information and analysis when making investment decisions.