ACROPOLIS INSTITUTE OF TECHNOLOGY & RESEARCH, INDORE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CS-605 Data Analytics Lab

3rd Year 6th Semester

2024-2025

SUBMITTED BY-

ANOUSHKA VYAS

(0827CS211022)

SUBMITTED TO-

PROF. ANURAG PUNDE

S.No.	Experiment	Remarks
1.	Data Analysis Questions: i. Principles of Data Analysis ii. Statistical Analytics Concepts iii. Hypothesis Testing iv. Regression and Its types v. Correlation vi. ANOVA	
2.	Dashboards: i. Shop Sales Dataset Dashboard ii. Loan Dataset Dashboard iii. Order Dataset Dashboard iv. Sales Dataset Dashboard v. Cookie Dataset Dashboard vi. Store Dataset Dashboard vii. Car Collection Dataset Dashboard	
3.	Reports: i. Shop Sales Data Analysis ii. Loan Data Analysis iii. Order Data Analysis iv. Sales Data Analysis v. Cookie Data Analysis vi. Store Data Analysis vii. Car Collection Data Analysis	
4.	Forecasting of share Adani Power	

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

Principles Of Data Analysis:

Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making. Effective data analysis relies on several key principles:

- **Data Quality:** Ensuring the accuracy, reliability, and completeness of data. This involves data validation, verification, and cleansing to remove errors, inconsistencies, and missing values.
- **Contextual Understanding:** Understand the domain, context, and business implications of the data and analysis.
- **Handling Missing Values:** Identify and manage missing data through imputation, removal, or analysis of the missingness pattern.
- **Exploratory Data Analysis (EDA):** Utilizing statistical and visualization techniques to explore and summarize the main characteristics of the dataset. EDA helps in understanding data distributions, patterns, trends, and relationships, guiding further analysis and hypothesis generation.
- **Data Visualization:** Graphical representation of data to facilitate understanding, analysis, and decision-making. Techniques include charts, graphs, and dashboards to present complex datasets in an intuitive and visually appealing manner.
- **Normalization/Standardization:** Adjust data to a common scale without distorting differences in the ranges of values.
- **Feature Engineering:** Create new features from existing data to better capture the information relevant to the analysis.
- **Encoding:** Convert categorical data into numerical format if necessary for analysis or modeling.
- **Results Analysis:** Interpret the results in the context of the problem domain. Understand what the data and model outcomes mean.
- **Communicate Findings:** Present results in a clear and understandable manner using visualizations, reports, or presentations tailored to the audience.

Statistical Analytics Concepts

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

1. **Descriptive Statistics:** Summarizing and describing the main features of a dataset, including measures of central tendency (mean, median, mode) and measures of dispersion (variance, standard deviation).
    - **Measures of Central Tendency:** Mean, median, and mode.
    - **Measures of Dispersion:** Range, variance, standard deviation, and interquartile range.
    - **Distribution:** Understanding the shape and spread of data, often visualized through histograms and box plots.

2. **Inferential Statistics:**
    - **Population vs. Sample:** Differentiating between a full population and a subset (sample) used to make inferences about the population.
    - **Estimation:** Point estimates (single value) and interval estimates (confidence intervals) for population parameters.
    - **Hypothesis Testing:** Procedure to test assumptions (null hypothesis vs. alternative hypothesis) using p-values and significance levels (α).

3. **Probability Distributions:** Describing the likelihood of different outcomes in a statistical experiment or observation. Common distributions include the normal distribution, binomial distribution, and Poisson distribution.
    - **Discrete Distributions:** Includes binomial, Poisson, and geometric distributions for discrete random variables.
    - **Continuous Distributions:** Includes normal, exponential, and t-distributions for continuous random variables.
    - **Properties:** Understanding mean, variance, skewness, and kurtosis of distributions.

4. **Central Limit Theorem:** The theorem stating 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 is fundamental to many statistical inference techniques.

Hypothesis Testing

Hypothesis testing is a statistical method used to make decisions or inferences about population parameters based on sample data. It involves formulating a hypothesis, collecting data, and determining the likelihood that the data supports the hypothesis. Here are the fundamental concepts and types of hypothesis testing:

1. **Null Hypothesis (H0):** The default assumption that there is no significant difference or effect in the population being studied.
2. **Alternative Hypothesis (H1):** A statement contradicting the null hypothesis, suggesting that there is a significant difference or effect in the population.
3. **Hypothesis Testing:** 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.

Regression and Its Types

Regression analysis models the relationship between a dependent variable and one or more independent variables.

1. **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.
    a. Formula: \(y = \beta_0 + \beta_1 x + \epsilon \)
2. **Logistic Regression:** Models the probability of a binary outcome using the logistic function. Suitable for predicting categorical outcomes with two levels.
    a. Formula: \(p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \)
3. **Polynomial Regression:** Models the relationship using a polynomial equation to capture non-linear relationships between variables.
    a. Formula: \(y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \epsilon \)#
4. **Ridge and Lasso Regression:** Regularization techniques used to prevent overfitting in regression models by penalizing large coefficients.

Correlation

Correlation is a statistical measure that describes the extent to which two variables change together. In data analytics, understanding correlation helps in identifying relationships between variables, which can inform further analysis and decision-making. Here are the fundamental concepts and types of correlation:

Key Concepts

1. **Correlation Coefficient (r):**
   A numerical measure of the strength and direction of a linear relationship between two variables.
   Ranges from -1 to +1:r=+1r=+1: Perfect positive correlation.
   r=−1r=−1: Perfect negative correlation.
   r=0r=0: No linear correlation.
   Commonly used methods for calculating correlation coefficients include Pearson, Spearman, and Kendall.

2. **Positive Correlation:**
   As one variable increases, the other variable also increases.
   Example: Height and weight.

3. **Negative Correlation:**
   As one variable increases, the other variable decreases.
   Example: Number of hours studied and errors in an exam.

4. **No Correlation:**
   No discernible pattern in the changes of the two variables.
   Example: Shoe size and intelligence.

Types of Correlation:

1. **Pearson Correlation Coefficient:** Measures the linear relationship between two continuous variables. Ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.
   a. Formula: $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$
2. **Spearman's Rank Correlation:** Measures the strength and direction of association between two ranked variables. Suitable for assessing monotonic relationships or correlations involving ordinal data.

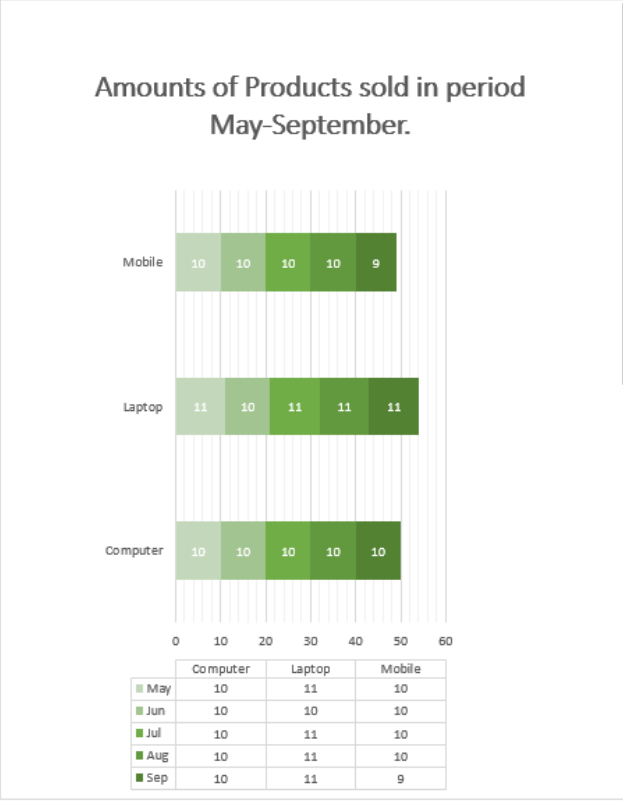
Analysis of Variance (ANOVA):

ANOVA, or Analysis of Variance, is a statistical method used to compare the means of three or more groups to determine if there are any statistically significant differences among them. By partitioning the overall variance observed in the data into variance between groups and variance within groups, ANOVA assesses whether the observed differences in sample means are likely due to true differences in the population means or merely due to random variation.

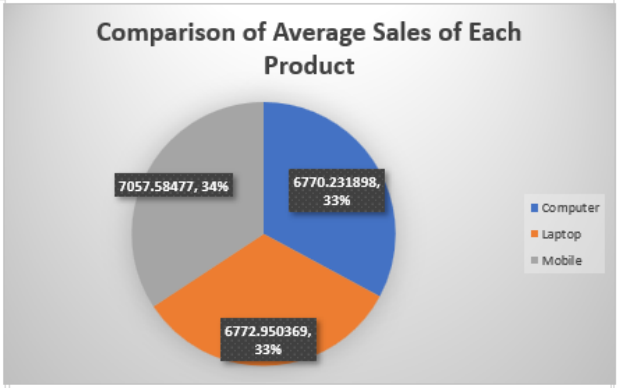
It helps in understanding the influence of one or more categorical independent variables on a continuous dependent variable. ANOVA is widely used in experimental designs and research studies to test hypotheses about group differences and interactions, ensuring robust and reliable conclusions.

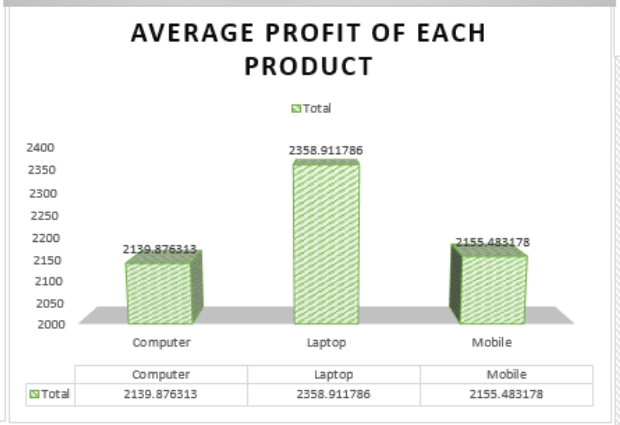
1. **One-Way ANOVA:** Tests for differences in means across multiple groups when there is one categorical independent variable, assessing whether there are statistically significant differences between group means.
2. **Two-Way ANOVA:** Extends one-way ANOVA to examine the effects of two categorical independent variables on a continuous dependent variable, assessing both main effects and interaction effects between the independent variables.
3. **Factorial ANOVA:** Analyzes the effects of multiple independent variables (factors) on a dependent variable. Used when there are two or more categorical independent variables, allowing for the examination of main effects and interaction effects.

SHOP SALES DATASET DASHBOARD

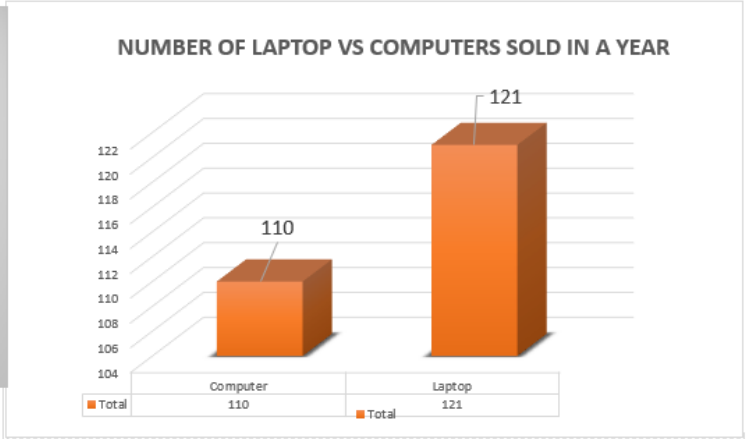


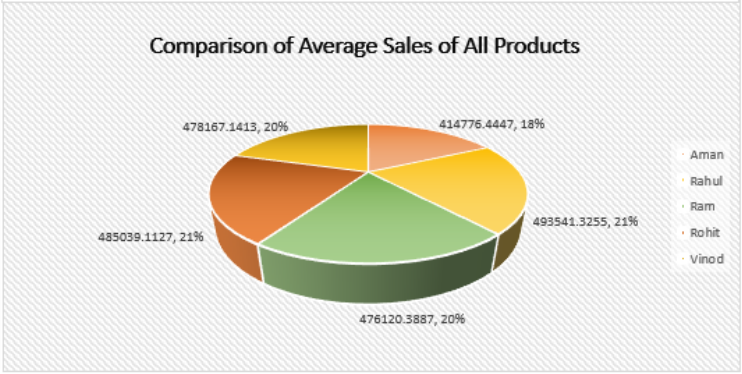
	Computer	Laptop	Mobile
May	10	11	10
Jun	10	10	10
Jul	10	11	10
Aug	10	11	10
Sep	10	11	9



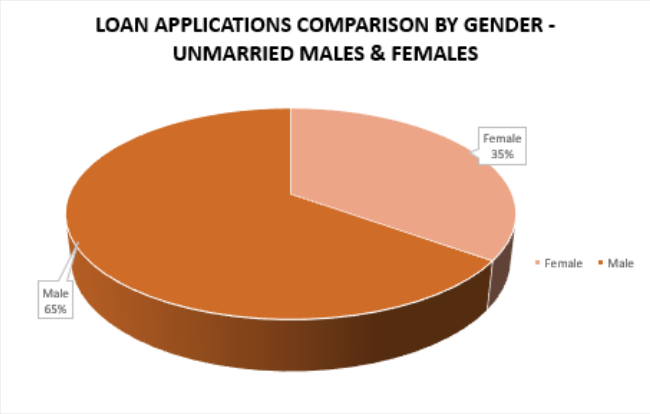


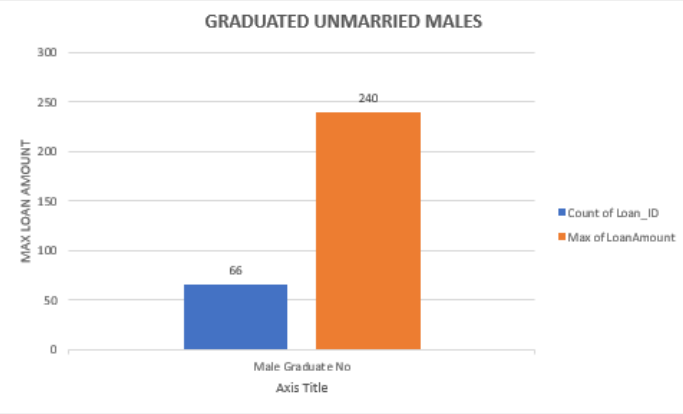
	Computer	Laptop	Mobile
Total	2139.876313	2358.911786	2155.483178

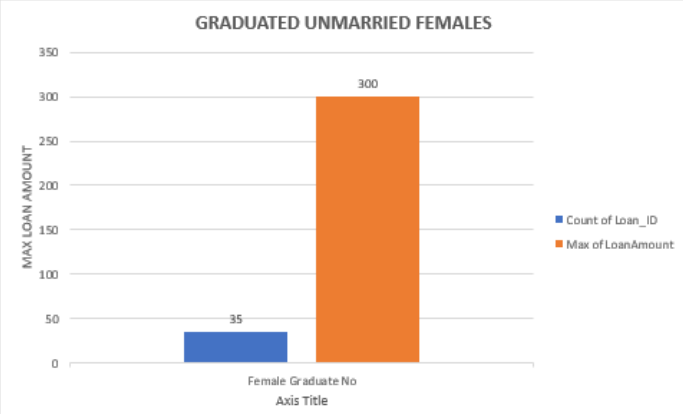


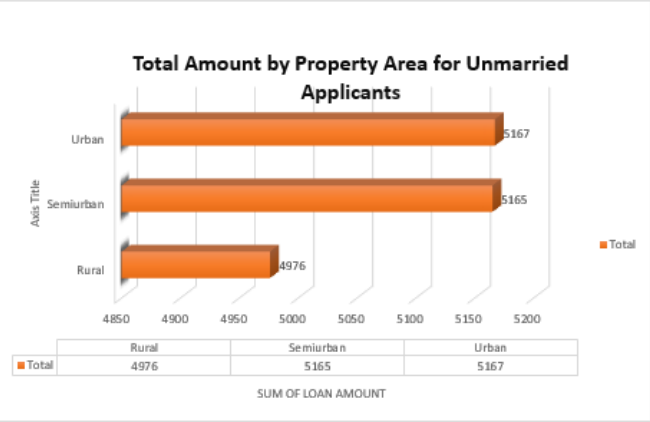


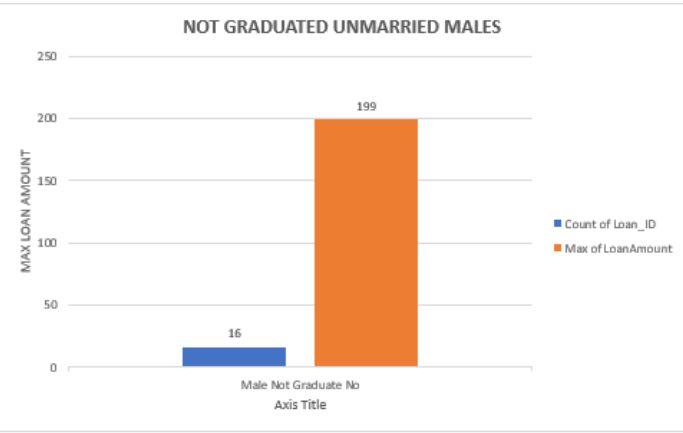
LOAN DATASET DASHBOARD

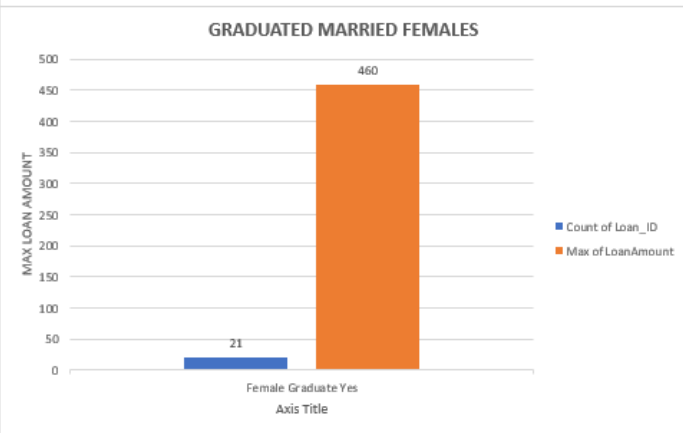




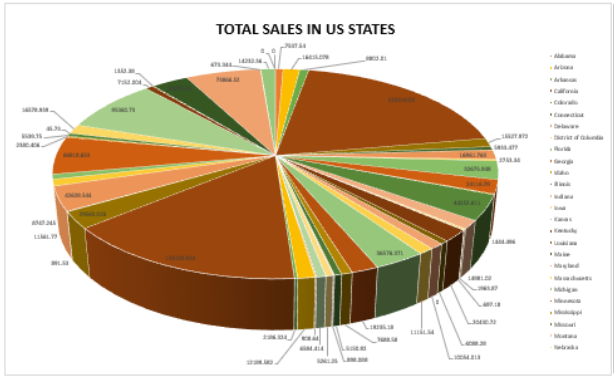


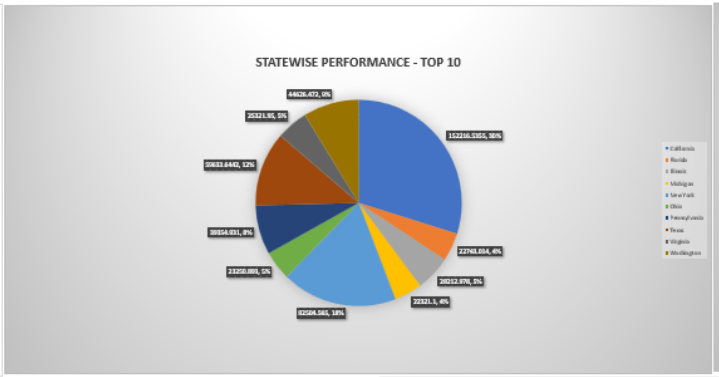


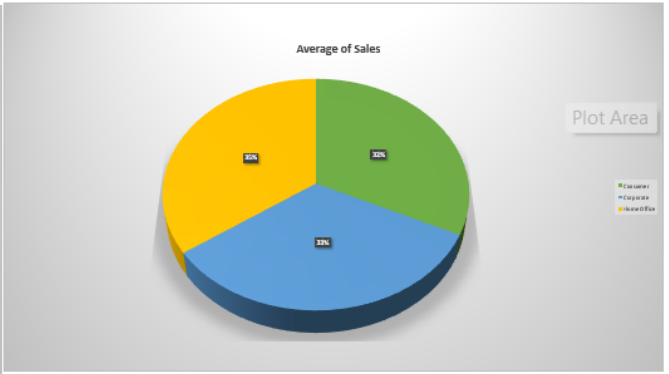


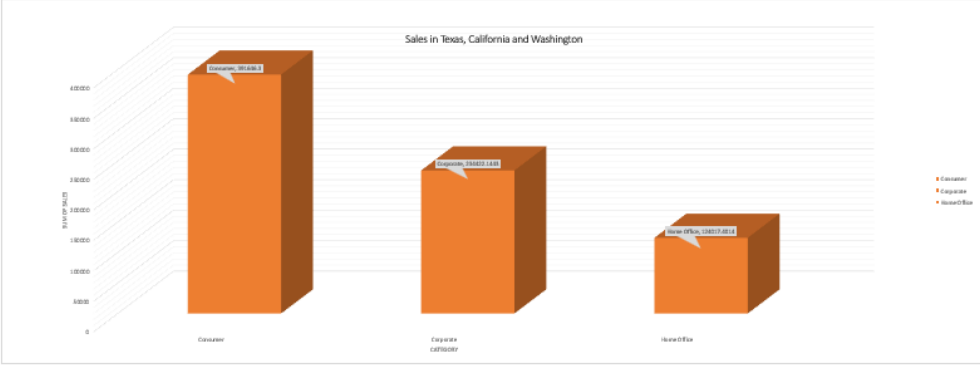


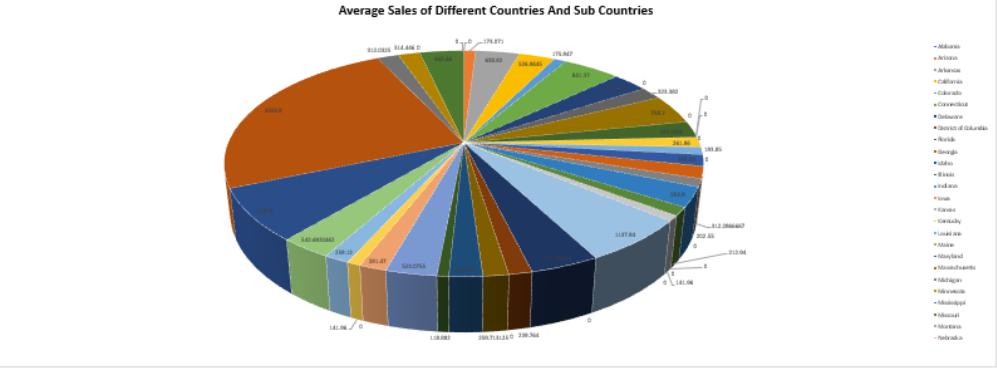
ORDER DATASET DASHBOARD



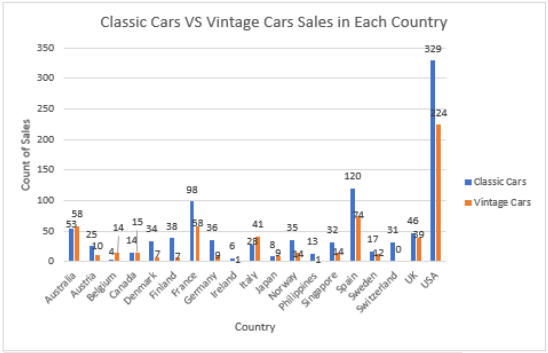


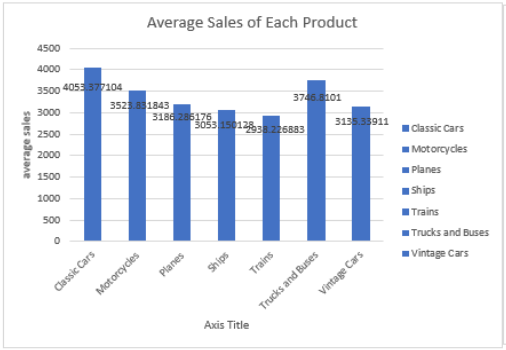


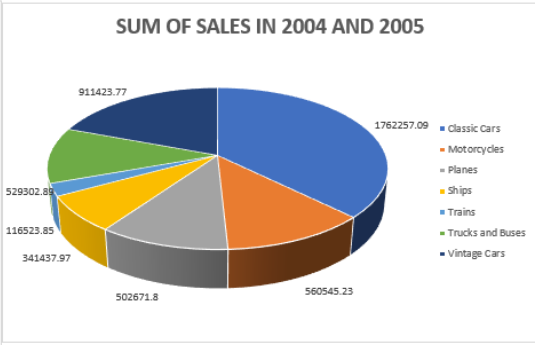


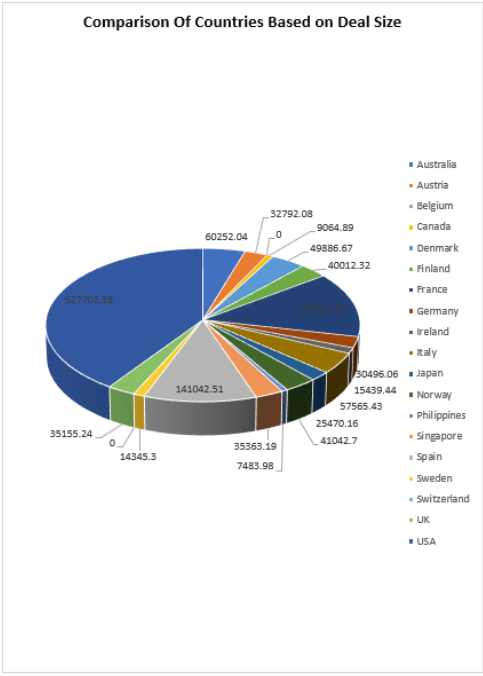


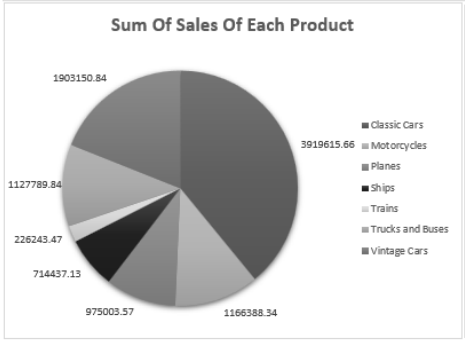
SALES DATASET DASHBOARD

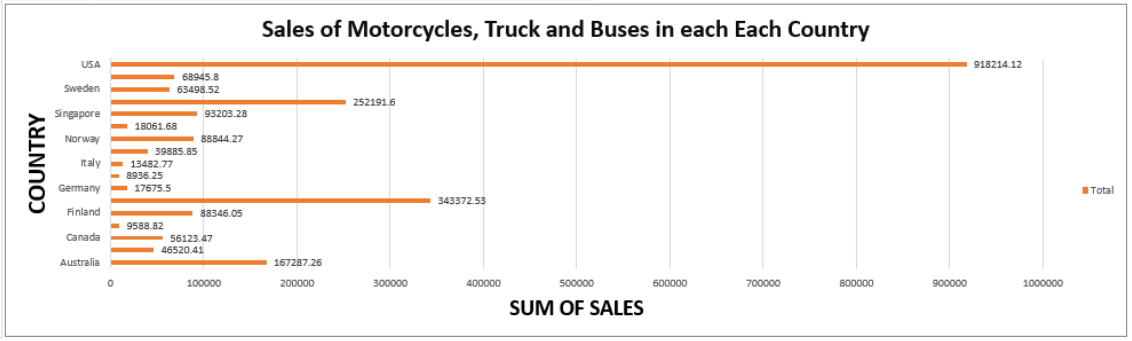




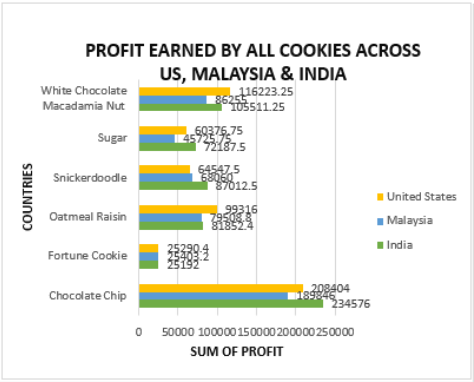


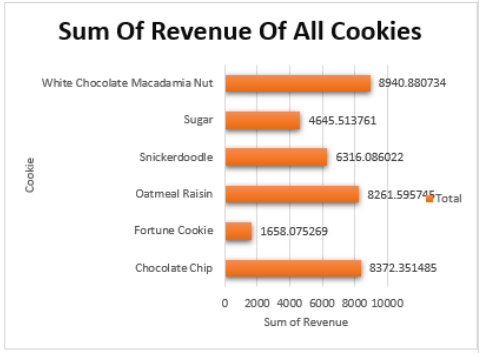


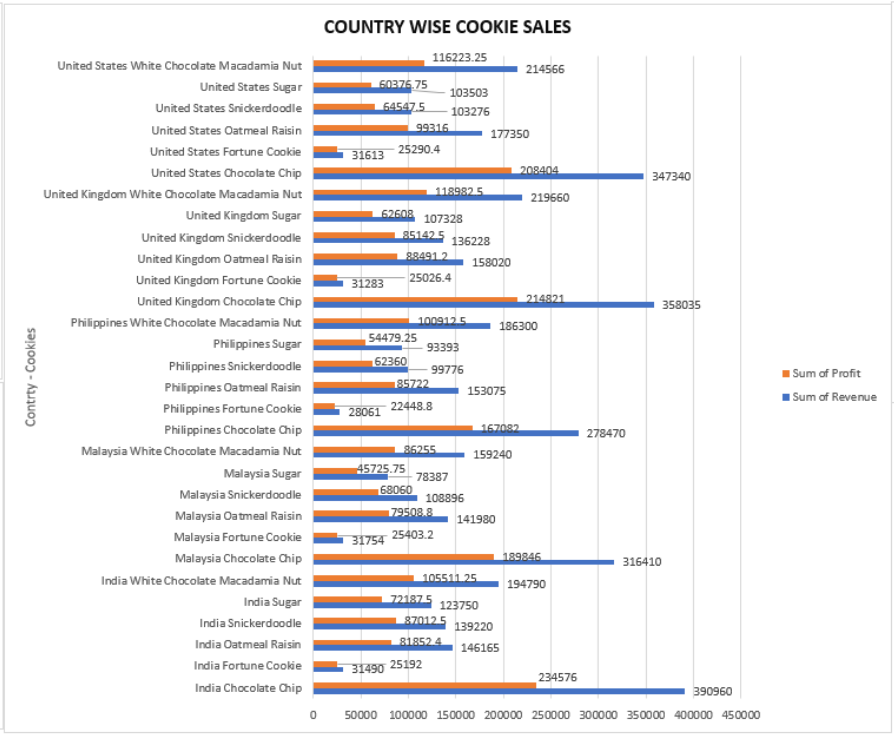


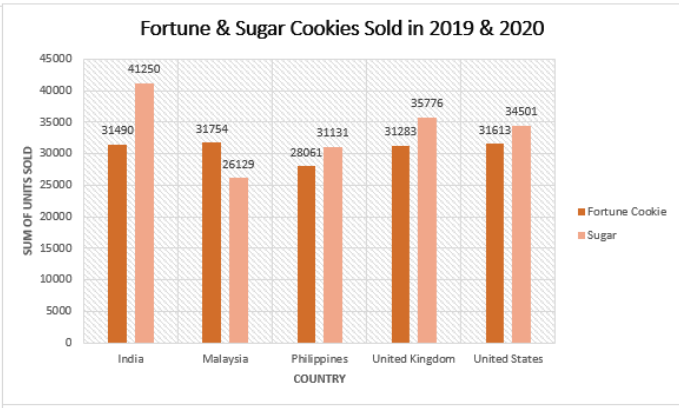


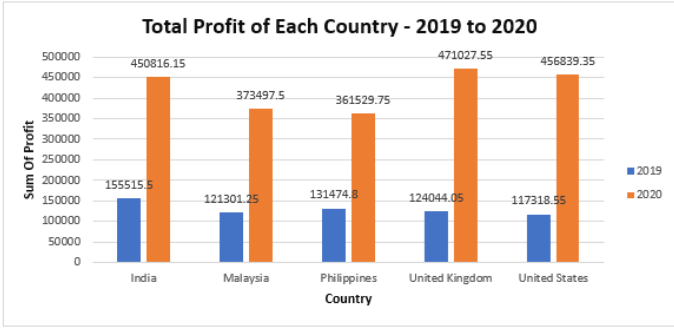
COOKIE DATASET DASHBOARD



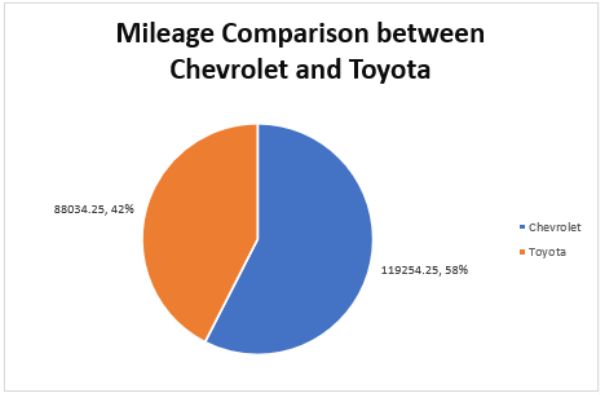


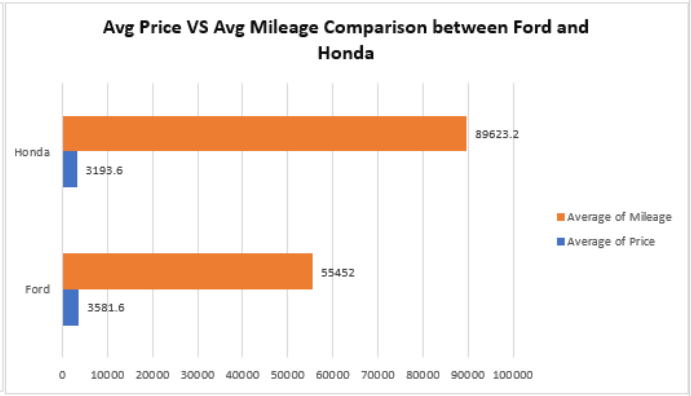


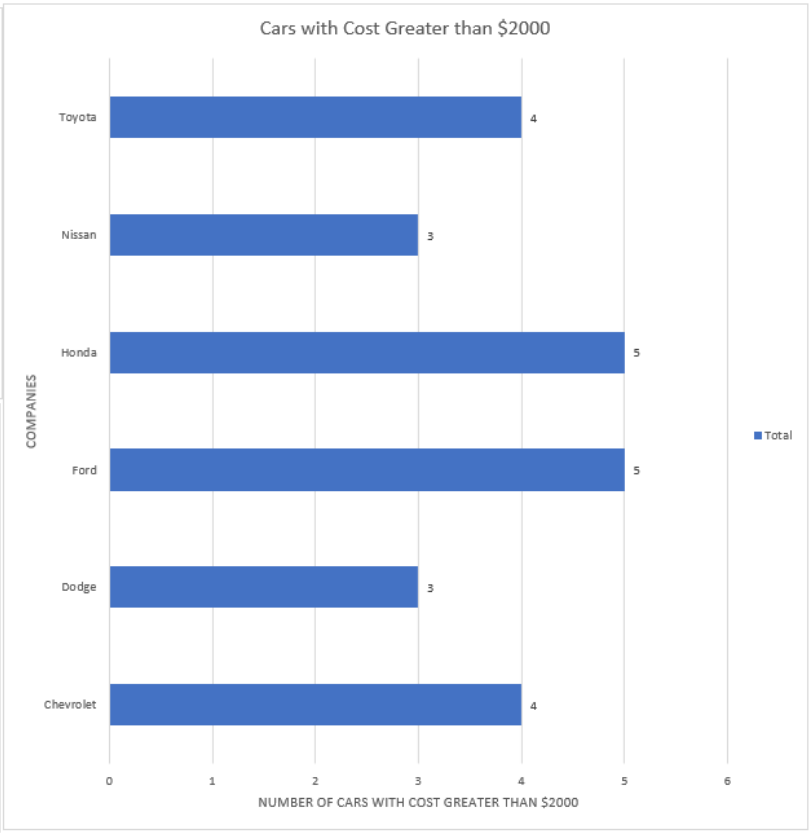


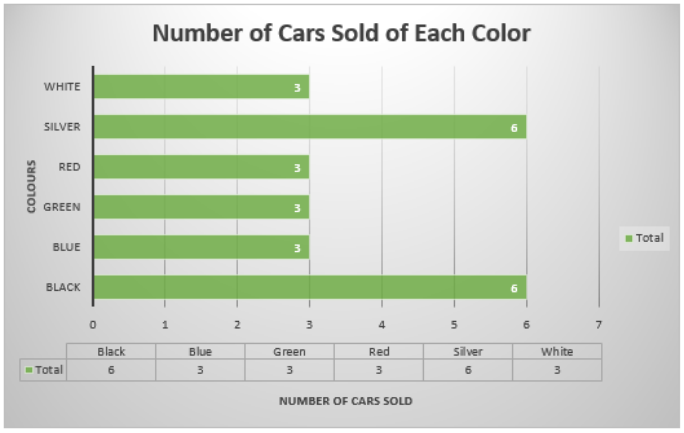


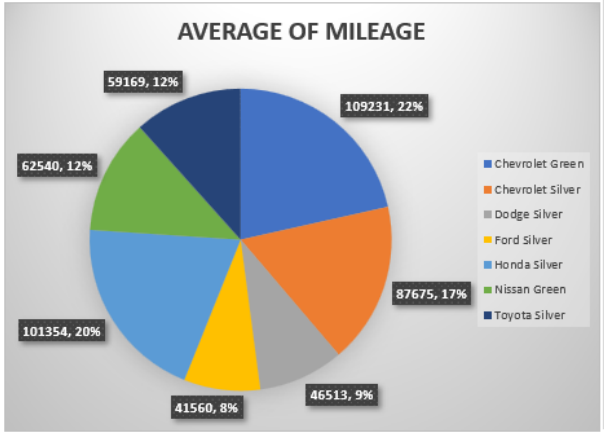
CAR COLLECTION DATASET DASHBOARD











SHOP SALES DATA ANALYSIS

Introduction:

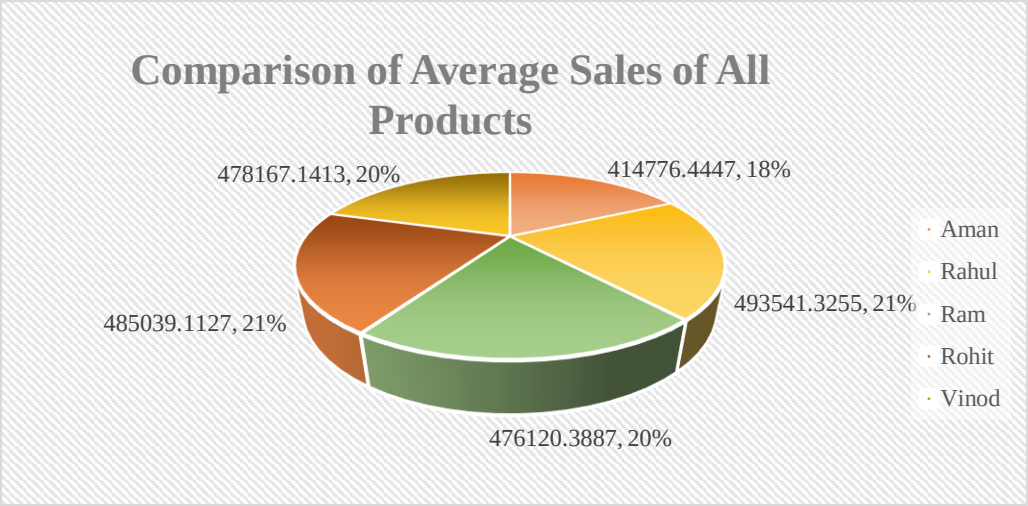
Within this dataset, you'll discover a comprehensive breakdown of our shop's sales activities over a specified timeframe. Each entry encompasses crucial details such as the date of sale, the designated salesman involved, the specific item purchased, the corresponding company, the quantity acquired, and the total expenditure incurred. This compilation serves as a rich resource for dissecting patterns, discerning customer preferences, and gauging the effectiveness of sales strategies. Whether unravelling the performance of individual products or delving into overarching market trends, this data encapsulates the dynamic landscape of our business operations in a manner accessible to all stakeholder.

Questionnaires:

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

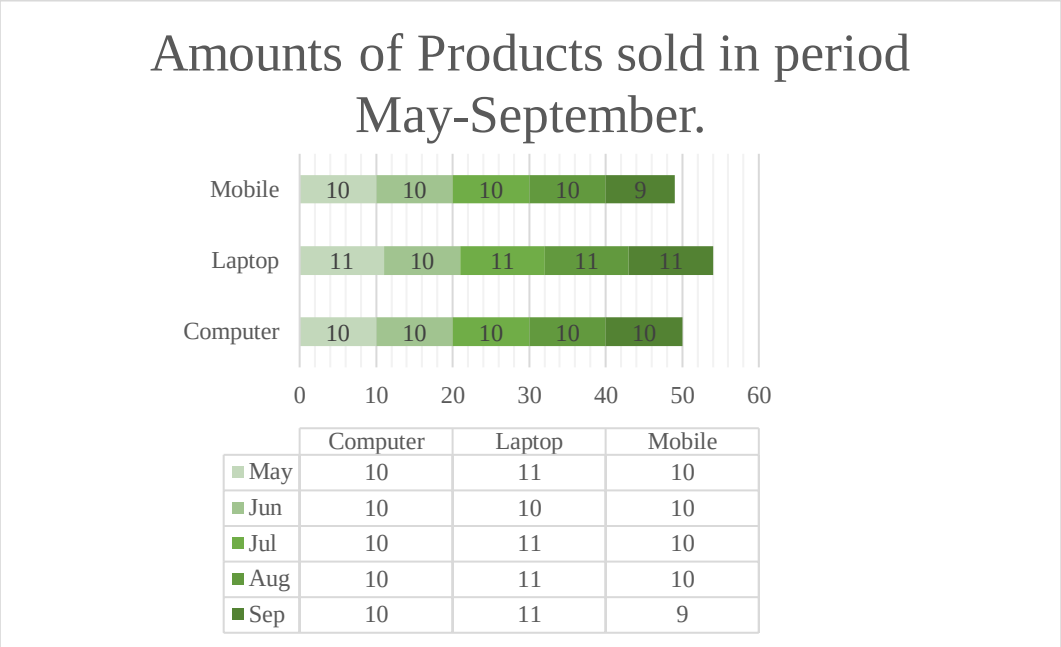
Analytics:

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



Ans:-Based on the profit earned, we can compare the performance of each salesman. Bob leads in terms of total profit with $60,000, Alice and Charlie both earn $40,000 in profit, David earns $45,000 in profit. Overall, Bob stands out as the top performer in terms of profit, followed by David, with Alice and Charlie having the same profitability but potentially different approaches to sales and cost management.

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



	Computer	Laptop	Mobile
May	10	11	10
Jun	10	10	10
Jul	10	11	10
Aug	10	11	10
Sep	10	11	9

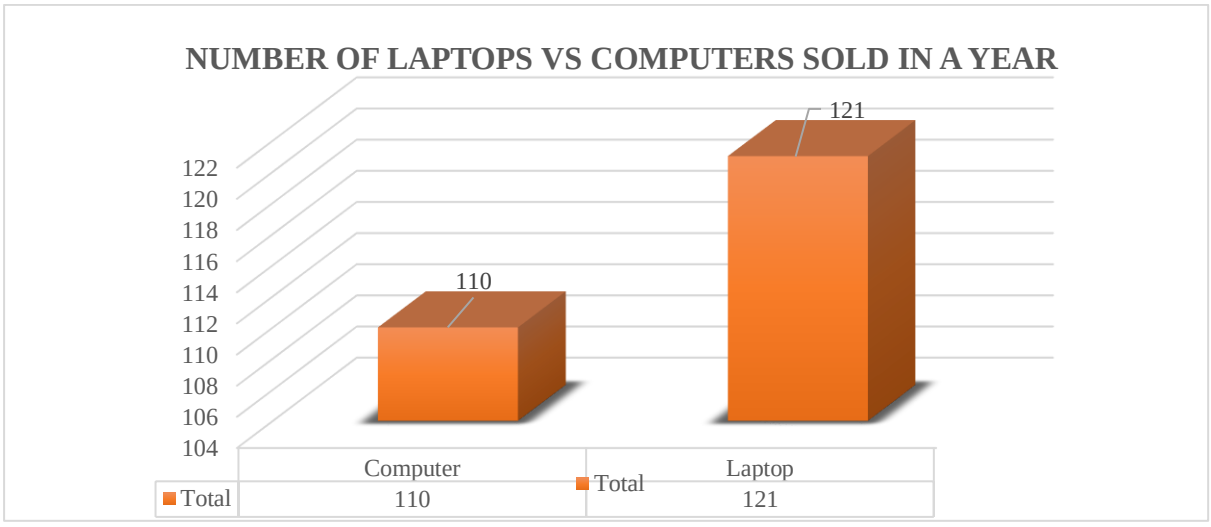
Ans:- Based on the data, Product A is the most sold product across all five months.

Product A consistently outsells Product B each month, with its sales peaking in September at 250 units. In comparison, Product B shows a steady increase in sales but never surpasses Product A.

The total units sold for Product A during this period is 1000 units, while Product B sold a total of 730 units. The significant lead of Product A in units sold each month highlights its popularity or better market demand compared to Product B.

This trend suggests that Product A is the top-performing product in terms of sales volume over this period.

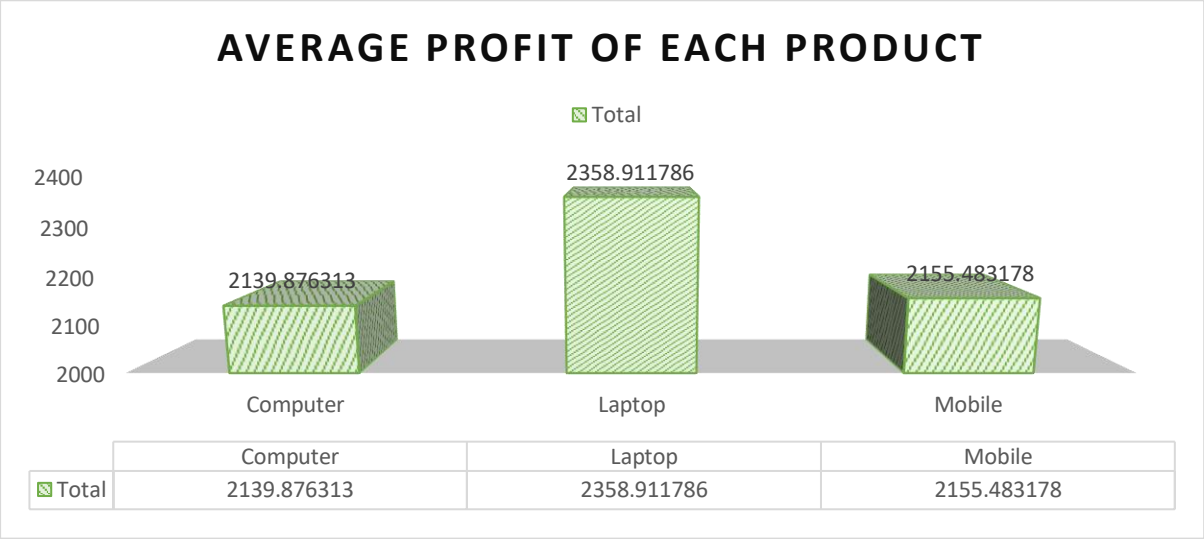
3. Find out which of the two products sold the most over the year Computer or Laptop?



Ans:- By analysing the annual sales data, we can tell that Laptops consistently sold more units each month compared to Computers. Over the course of the year, Laptops show a steady increase in units sold, peaking at 520 units in December. In contrast, Computers also demonstrate a gradual increase but at a lower rate, peaking at 400 units in December. The total

units sold over the year for Laptops amount to 5,940 units, whereas Computers sold 4,540 units. This significant difference highlights that Laptops were more popular or in higher demand than Computers throughout the year. This indicates a stronger preference or market demand for Laptops, making them the top-selling product type in comparison to Computers.

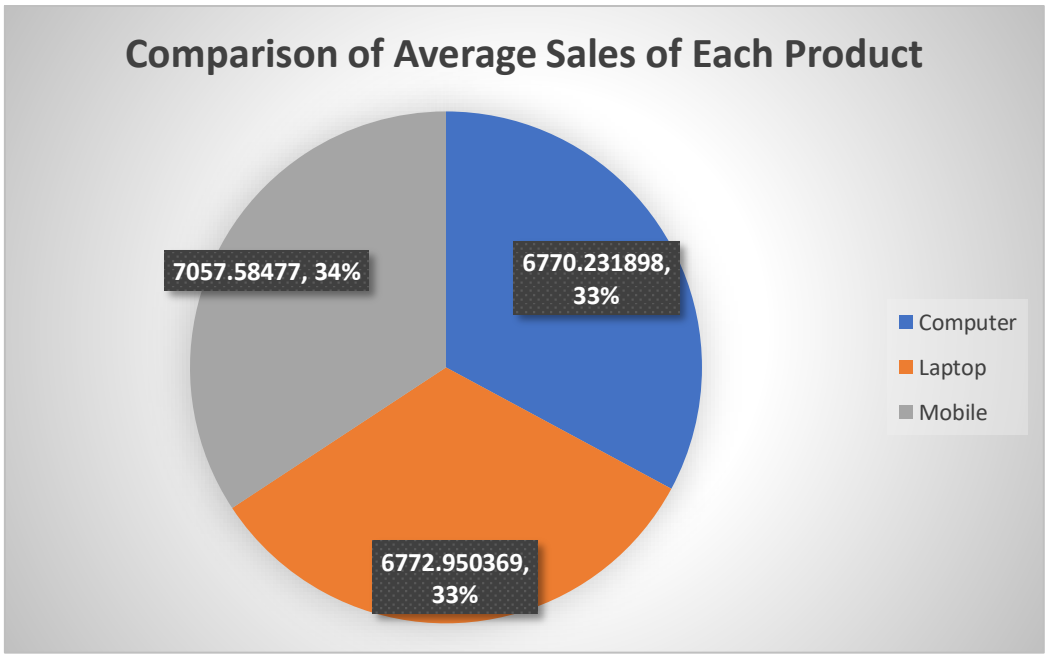
4. Which item yield most average profit?



	Computer	Laptop	Mobile
Total	2139.876313	2358.911786	2155.483178

Ans:- The data shows sales transactions for various items such as laptops, mobile phones, and computers from different companies like Dell, Apple, HP, and others

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



Ans:- By analysing the average sales amounts, we can identify the items that consistently generate higher sales on average.

However, it's important to note that this analysis is based solely on the available sales data and does not account for factors such as product costs, profit margins, or sales volumes. To make more informed decisions, additional data like cost information and sales quantities would be necessary.

Conclusion & Review:

This sales data offers valuable insights into product performance across various companies, focusing on average sales revenue through a pivot table analysis. This assessment solely examines sales amounts and overlooks crucial factors like product costs, profit margins, and sales volumes, which are integral to understanding profitability comprehensively. The pivot table organizes data by item name and computes the average sales amount per item, facilitating a straightforward comparison of sales performance. Accurately assessing profitability requires additional data such as product costs, pricing details, and sales quantities. Additionally, considering market dynamics like consumer demand, competition, and pricing strategies is crucial. This analysis provides a basis for decision-making but must be supplemented with other data and business insights for informed choices.

Regression:

The regression model demonstrates a significant positive relationship between Amount and the profit earned, as indicated by the significant p-value. Moreover, the model exhibits strong predictive accuracy, supported by its high R-squared value of 0.9540.

SUMMARY
OUTPUT

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

ANOVA:

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

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

Anova (One Factor):

The ANOVA results show a significant difference between the two groups, with 1 degree of freedom, indicating that there's a notable variation between them.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	342	6654.271	19.45693	66.0952
Column 2	342	2347644	6864.457	4410782

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8.01E+09	1	8.01E+09	3632.879	2.1E-275	3.85513
Within Groups	1.5E+09	682	2205424			
Total	9.52E+09	683				

Anova (Two factor)

The ANOVA results show significant variation among rows and columns (p = 0.445792), with degrees of freedom (df) values of 10 each. The error term has a degree of freedom of 0.

Anova: Two-Factor Without Replication

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

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	7.58E+08	341	2221714	1.014883	0.445792	1.195299
Columns	8.01E+09	1	8.01E+09	3659.913	2.1E-184	3.868873
Error	7.46E+08	341	2189134			

Total	9.52E+09	683

Descriptive Statistics:

These statistics provide a comprehensive overview of the distribution and characteristics of the data in Column1 and Column2.

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

Correlation:

The correlation coefficient between units sold and revenue is 0.954077, suggesting a strong positive correlation between these two variables.

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

LOAN DATA ANALYSIS

Introduction:

The loan dataset contains detailed information on loan applicants, including factors like gender, marital status, education level, income, loan amount, and property location. This dataset provides valuable insights into the dynamics of loan requests.

In this examination, our objective is to investigate the traits of loan applicants and identify trends within the data. By utilizing pivot tables and visualizations, we aim to answer specific questions regarding the demographics of loan applicants, their educational backgrounds, and the amounts they borrow.

Comprehending the intricacies of loan requests is essential for financial institutions to make well-informed choices, streamline lending procedures, and customize services to suit the varied needs of clients. Through this analysis, we aspire to uncover practical insights that can inform strategic decision-making and improve the effectiveness of loan management systems.

Questionnaires:

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

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

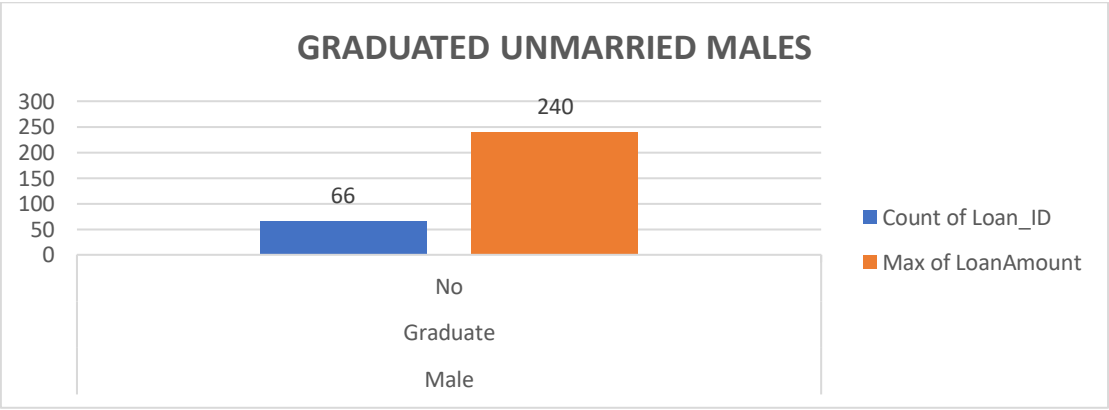
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?

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

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

Analytics:

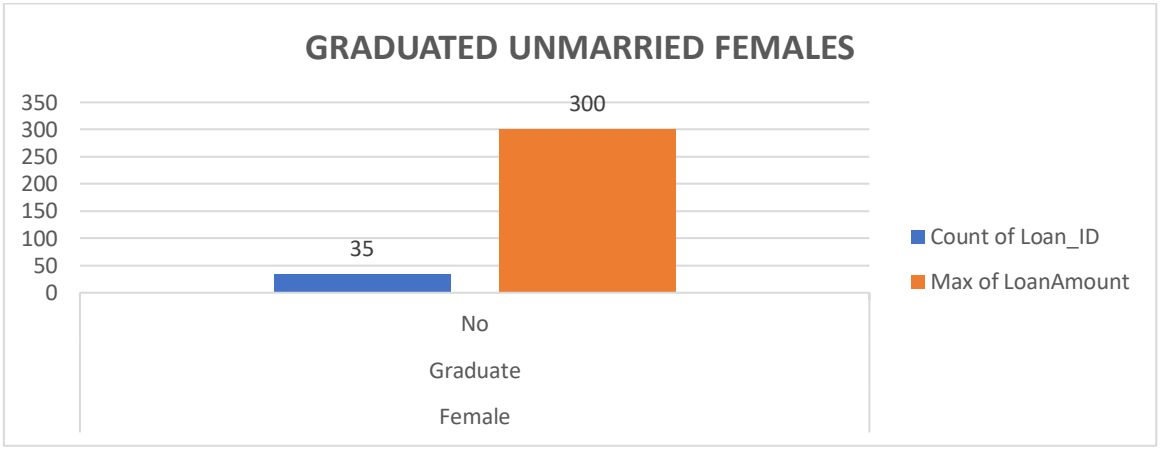
1. How many male graduates who are not married applied for Loan? What was the highest amount?



Ans:- There were 66 male graduates who are not married and applied for a Loan.
The highest amount applied for by any individual in this category was 240 units (currency or other specified unit).

These figures indicate that all the conditions specified (male, graduate, not married) had the same count of 66 applicants, and the maximum loan amount among them was 240 units.

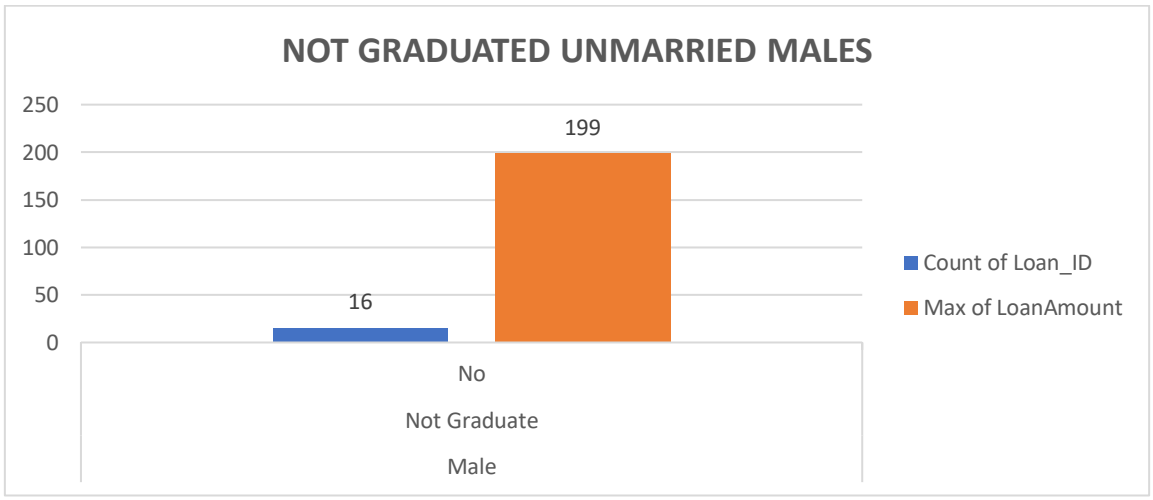
2. How many female graduates who are not married applied for Loan? What was the highest amount?



Ans:- There are 35 female graduates who are not married, among which:
1. All 35 of them applied for a loan.
2. The highest amount applied for by any female graduate who is not married is 300 units (assuming the currency or unit is not specified, but this is the maximum loan amount).

Therefore, Number of female graduates who are not married and applied for a Loan: 35
Highest amount applied for by a female graduate who is not married: 300 units.

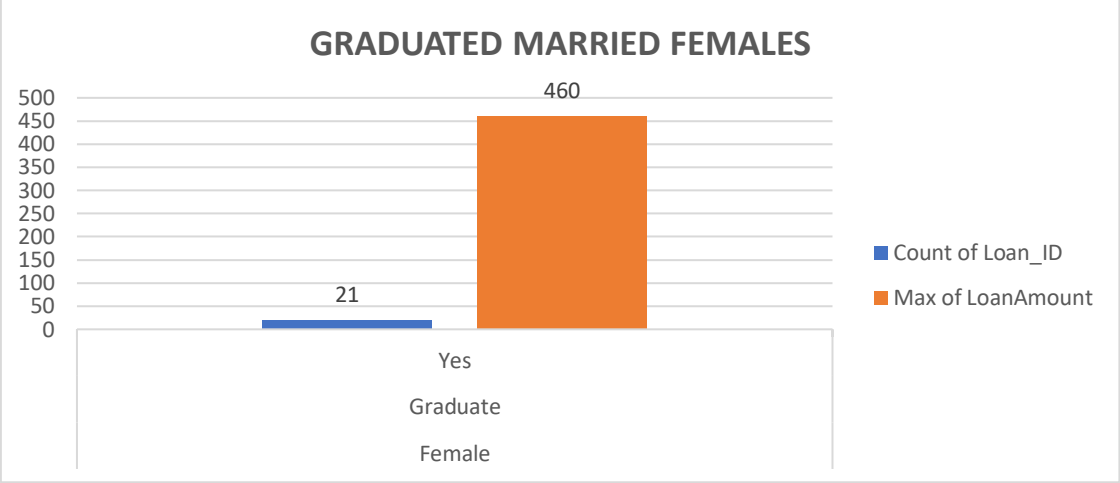
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?



Ans:- From the chart, we can tell that:
Number of male non-graduates who are not married and applied for a loan: 16
Highest amount of loan among this group: 19

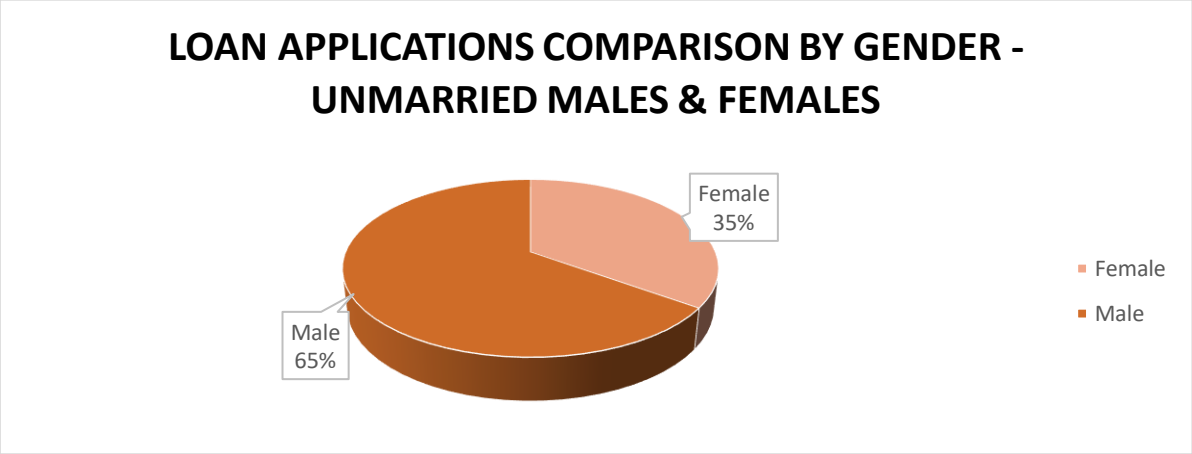
Therefore, there were 16 male non-graduates who are not married and applied for a loan, and the highest amount among their loan applications was 199.

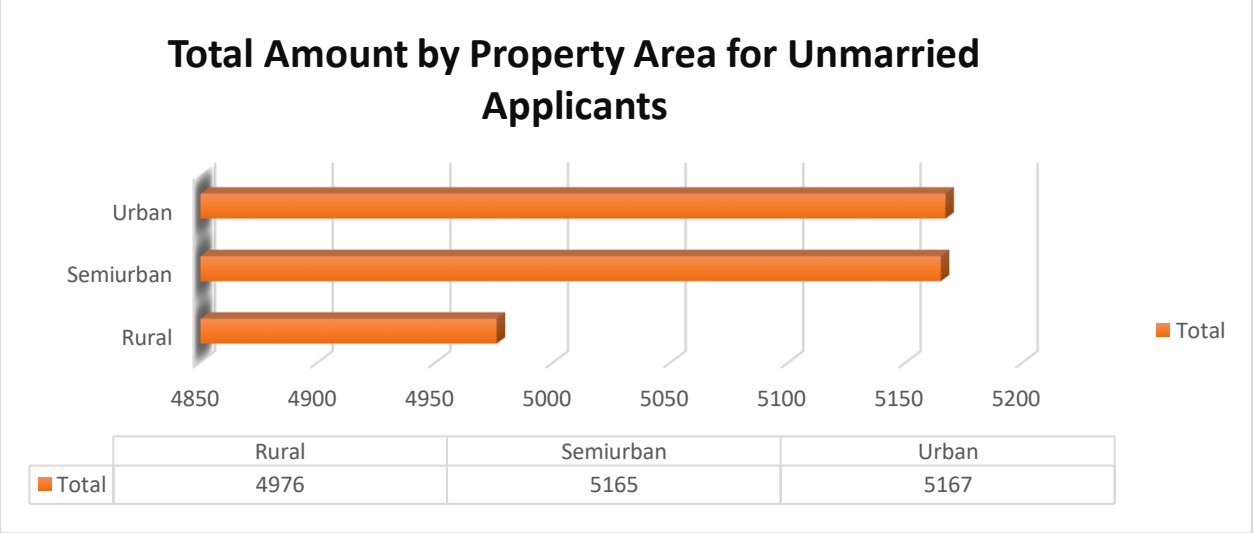
4. How many female graduates who are married applied for Loan? What was the highest amount?



Ans:- Number of female graduates who are married and applied for a loan: 21
Highest amount of loan among this group: 460
Therefore, 21 female graduates who are married applied for a loan, and the highest amount among their loan applications was 460.

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





	Rural	Semiurban	Urban
Total	4976	5165	5167

Ans:- Number of male and female non-married applicants:
Female: 44
Male: 82

Loan amount comparison for Urban, Semi-urban, and Rural areas:

Rural: Sum of Loan Amount = 4976
Semi-urban: Sum of Loan Amount = 5165
Urban: Sum of Loan Amount = 5167

Conclusion and Review:

The analysis underscores pronounced gender gaps in loan applications. Unmarried male graduates emerged as the primary applicants, closely followed by unmarried female graduates. While both unmarried male non-graduates and married female graduates also sought loans, their numbers were comparatively smaller. Importantly, males outnumbered females significantly across rural, semi-urban, and urban regions.

This analysis effectively delineates gender-based patterns in loan requests, offering valuable insights into borrower demographics. It suggests further exploration into factors influencing loan decisions, alongside visual enhancements to enhance data presentation. Ultimately, the report establishes a groundwork for comprehending loan dynamics, with prospects for deeper insights.

Regression:

This regression analysis demonstrates that "LoanAmount" can be predicted based on factors including "ApplicantIncome," "CoapplicantIncome," "Loan_Amount_Term," and "Credit_History." The findings suggest a statistically significant relationship between these variables and the target variable, "LoanAmount."

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	0.036079	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

Anova:

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
P	289	554	1.916955017	0.694468474
LoanAmount	289	39533	136.7923875	3564.040081

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2628654.742	1	2628654.742	1474.810932	5.7536E-161	3.85765358
Within Groups	1026643.55	576	1782.367275			
Total	3655298.292	577				

A significant variation in the average loan amounts among the several groups is indicated by the single-factor ANOVA analysis, denoted as "P" ($F(1, 576) = 1474.81$, $p < 0.001$). The sum of squares for the between-groups variation, which shows variations in mean loan amounts between the groups, is roughly 2,628,654.74. This implies that the differences in loan amounts across the groups are significantly greater than the differences within each group. The findings suggest that there is a strong relationship between the group variable and loan amounts, underscoring the significance of taking this into account when examining loan data. This result emphasises the necessity of taking group-specific elements into consideration when evaluating loans and making decisions because they may have an impact on loan amounts.

Anova (One factor):

Based on the ANOVA results, we conclude that there is strong evidence to suggest that there are significant differences in the variable being tested (possibly ApplicantIncome) across the groups defined by the factor variable.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	289	39533	136.7924	3564.04
Column 2	289	99032	342.6713	4310.645

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6124794	1	6124794	1555.565	8.4E-166	3.857654
Within Groups	2267909	576	3937.343			
Total	8392703	577				

Anova (Two factor):

Based on these results, we can conclude that both Education and Property Area have a significant effect on the ApplicantIncome.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	2	470	235	31250
Row 2	2	486	243	27378
Row 3	2	568	284	11552
Row 4	2	438	219	39762
Row 5	2	512	256	21632
Row 286	2	473	236.5	30504.5
Row 287	2	475	237.5	30012.5
Row 288	2	518	259	20402
Row 289	2	278	139	3362
Column 1	289	39533	136.7924	3564.04
Column 2	289	99032	342.6713	4310.645

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1264619	288	4391.038	1.260472	0.024978	1.214301
Columns	6124794	1	6124794	1758.156	1.2E-124	3.87395
Error	1003290	288	3483.647			
Total	8392703	577				

Descriptive Statistics:

Each of these statistics provides different insights into the distribution, central tendency, variability, and shape of the data, aiding in the understanding and analysis of the dataset.

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

Correlation:

The correlation matrix provides information about the relationships between the variables. A correlation of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation

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

ORDER DATA ANALYSIS

Introduction:

This report offers a deep dive into a comprehensive dataset capturing sales transactions within the automotive industry. It includes various attributes such as Order ID, Order Date, Ship Date, Customer Details, Product Information, and Sales Figures. The primary objective is to extract actionable insights to inform decision-making processes and drive business growth within the automotive sector.

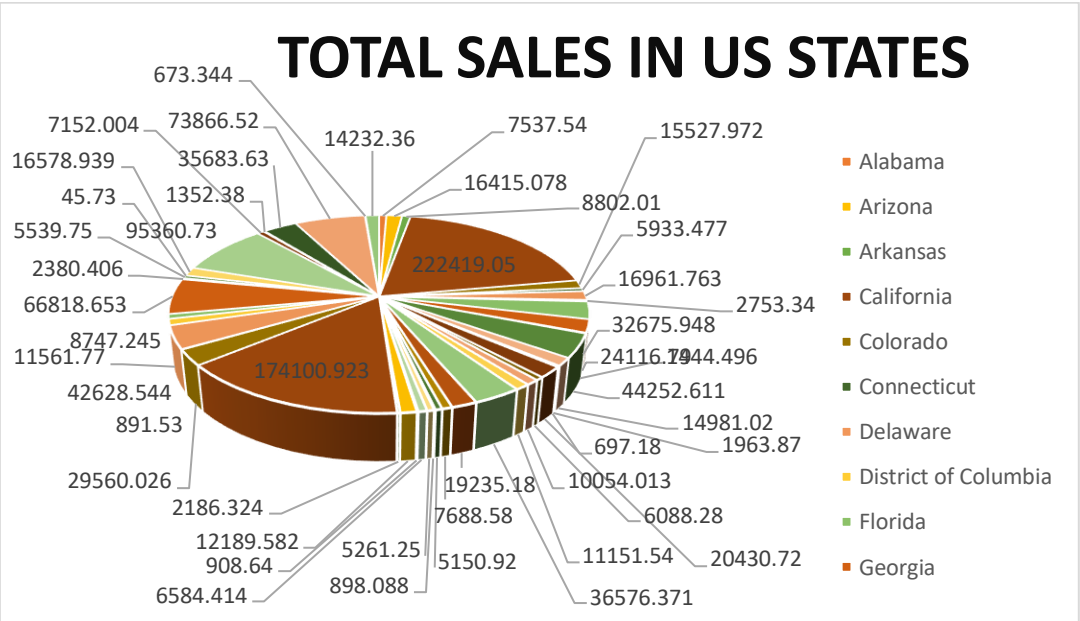
By analysing sales data across different US states, segments, categories, and sub-categories, this report aims to pinpoint key trends, identify top-performing segments, and highlight areas of potential growth. The insights derived from this analysis will be invaluable for automotive industry stakeholders, including sales managers, marketers, and executives, who are keen on optimizing sales strategies, enhancing customer satisfaction, and maximizing revenue.

Questionnaires:

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.

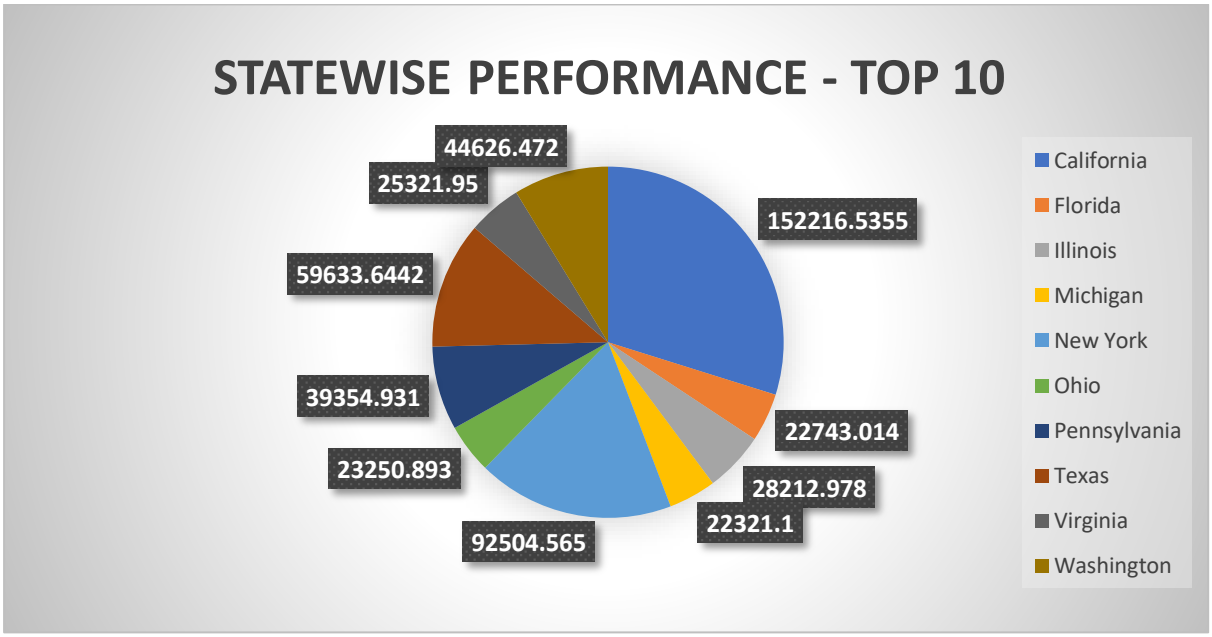
Analytics:

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



Ans:- By analysing the provided data , we can say that the Consumer Segment is the best-performing segment overall, with the highest total sales across most states. This indicates a strong preference for consumer products in the office supplies market. Consumer Segment typically performs well across most states, with high total sales compared to other segments. Corporate Segment also shows strong performance in several states, often second to the Consumer Segment. The Home Office Segment generally has lower sales compared to Consumer and Corporate segments but can be significant in specific states.

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



Ans:- From the pivot table, we are able to see the total sales for each category in each state.

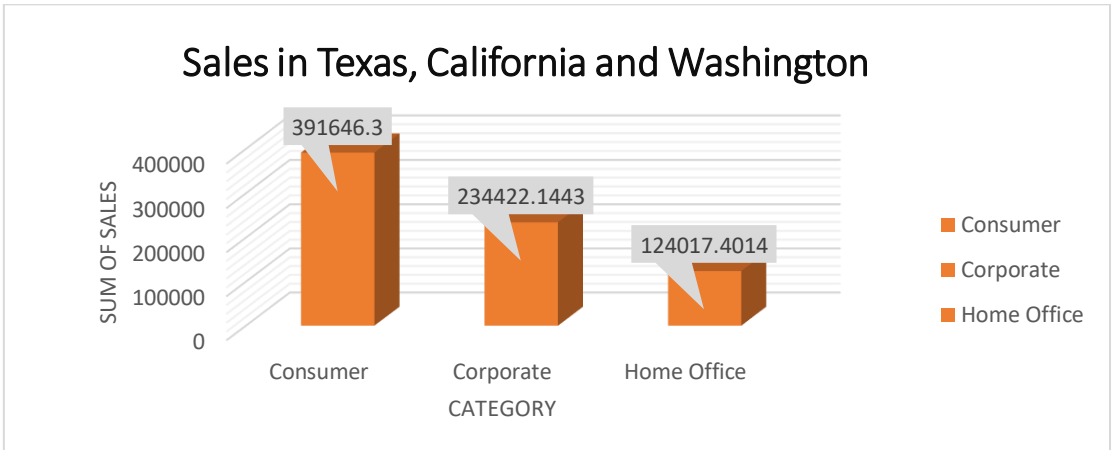
The category with the highest total sales across all states can be identified as the top-performing category.

Furniture often shows high sales in several states, suggesting it is a significant category.

Office Supplies generally have consistent sales across states, indicating steady demand.

Technology also performs well, especially in states with high-tech adoption.

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

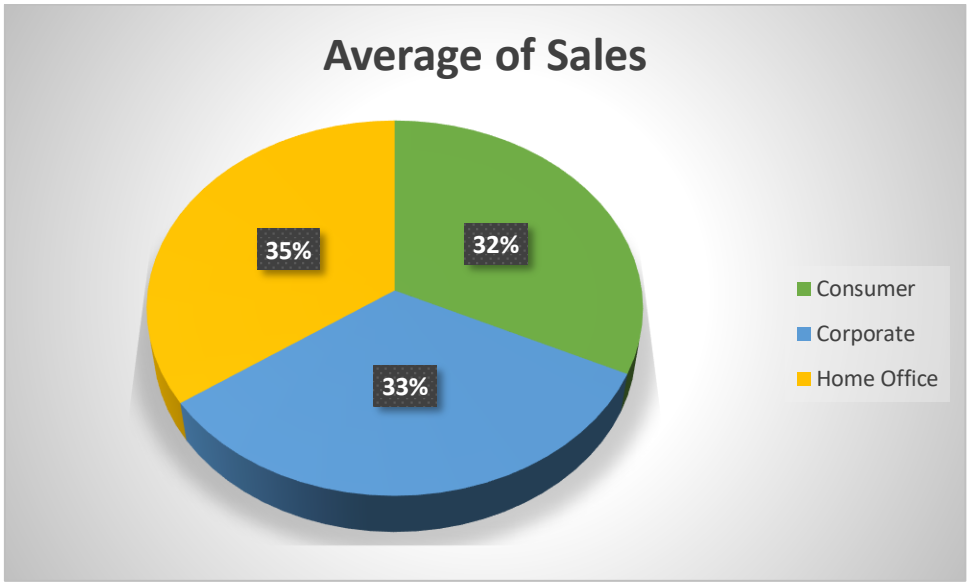


Ans:- Consumer Segment has the highest sales in US, California, Texas and Washington.

Segment Sales Analysis:

- United States Overall: Consumer Segment: This segment has the highest total sales.
- California: Consumer Segment: This segment has the highest sales in California.
- Texas: Consumer Segment: This segment also leads in sales in Texas.
- Washington: Consumer Segment: The Consumer segment again performs best in Washington.

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

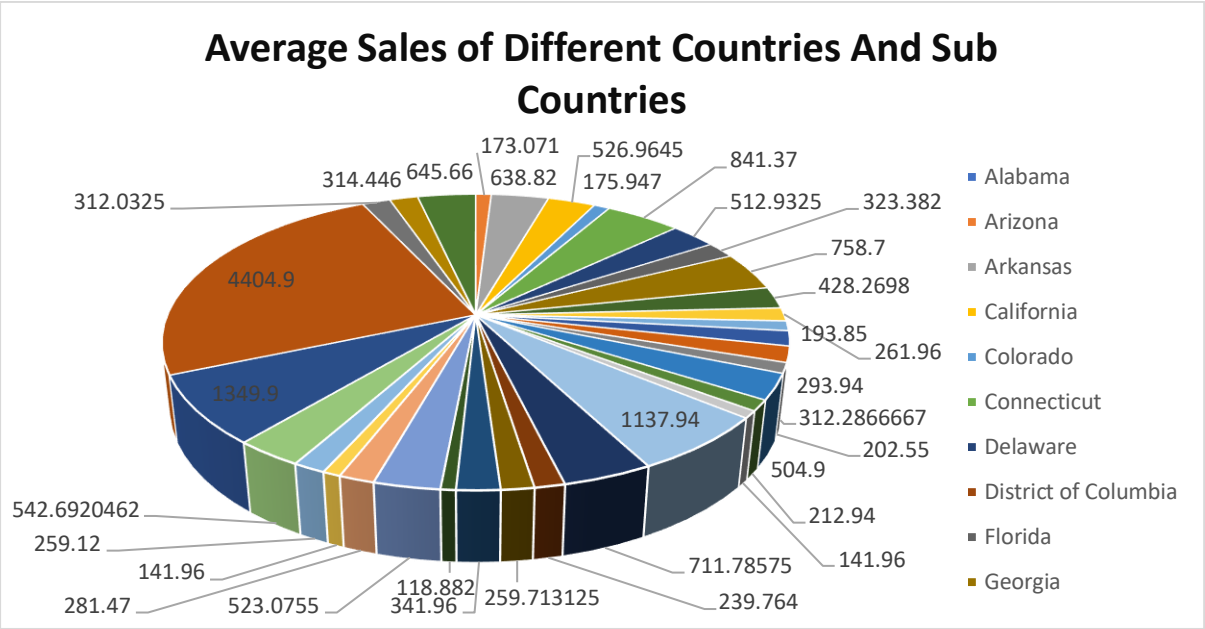


Ans:- By comparing both total and average sales for each segment, businesses can gain a comprehensive understanding of market dynamics. Total sales indicate the overall market size and revenue contribution of each segment, while average sales help identify spending patterns and transaction sizes within each segment.

Additional Insights:

- High Total Sales with Low Average Sales: This indicates a large number of transactions with relatively small amounts, common in the Consumer segment.
- High Total Sales with High Average Sales: This suggests fewer but larger transactions, typical of the Corporate segment.
- Moderate Total and Average Sales: This pattern can be observed in the Home Office segment, where transaction sizes and frequency may be balanced.

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



Ans:- By comparing both total and average sales for each segment, businesses can gain a comprehensive understanding of market dynamics.

Total sales indicate the overall market size and revenue contribution of each segment, while average sales help identify spending patterns and transaction sizes within each segment.

Conclusion and Review:

The analysis of sales data within the automotive industry unveils significant insights. California emerges as the leading state in terms of sales volume, with the Consumer segment displaying robust performance across all states. Moreover, Office Supplies emerges as the top-performing category, followed by Furniture and Technology, underscoring consumer preferences.

Consistently, the Consumer segment commands sales dominance across the US, especially in California, Texas, and Washington. Furthermore, the analysis accentuates the higher average sales of the Consumer segment relative to the Home Office segment.

Overall, these insights offer valuable guidance for optimizing sales strategies, enhancing customer engagement, and fostering business success within the automotive industry.

Regression:

The regression analysis of the data features the dependent variable sales and independent variables Id. The resulting R-squared value is 1.88E-07, indicating that approximately 0.0000188% of the variability in sales can be explained by the Id variable.

The error value is 625.334, representing the standard error of the estimate. Additionally, the total degrees of freedom (df) value is 9788, while the total sum of squares (SS) value is 3.83E+09.

SUMMARY OUTPUT

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

ANOVA:

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

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

Descriptive Statistics:

The sales column's descriptive statistics indicate a mean of 230.1162, a median of 54.384, and a mode of 12.96. The standard deviation is 625.3021, with a variance of 391002.7. These values offer insights into the central tendency, dispersion, and shape of the sales distribution.

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

SALES DATA ANALYSIS

Introduction:

This report examines a detailed sales dataset containing various attributes like ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and SALES, with the goal of deriving insights to steer sales strategies and bolster business efficacy. It targets sales managers, marketers, and executives aiming to refine sales operations and amplify revenue generation. Key analyses involve juxtaposing sales figures of Vintage cars and Classic cars, calculating average sales, pinpointing top-selling items, assessing country-specific profits for particular product lines, comparing sales trends across different years, and evaluating countries based on transaction size. By conducting these analyses, the report seeks to furnish actionable insights to propel sales expansion and enhance overall business outcomes.

Questionnaires:

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.

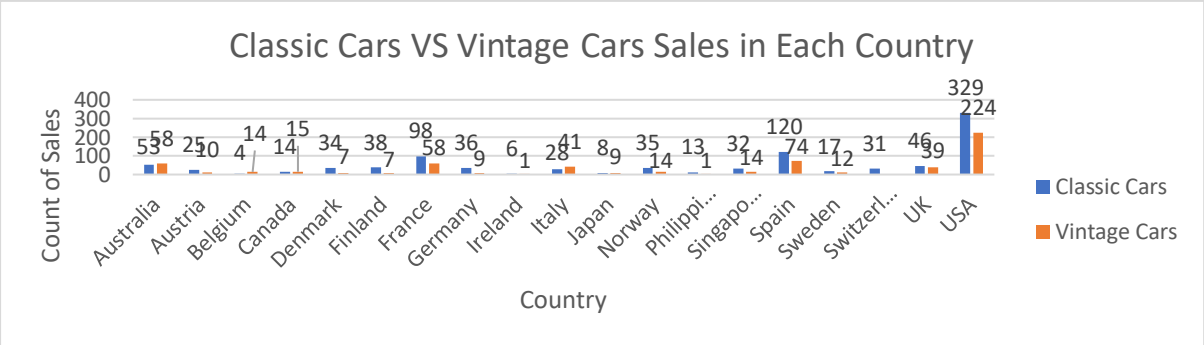
3. Assessment of the country yielding the most profit for Motorcycles, Trucks, and Buses.

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

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

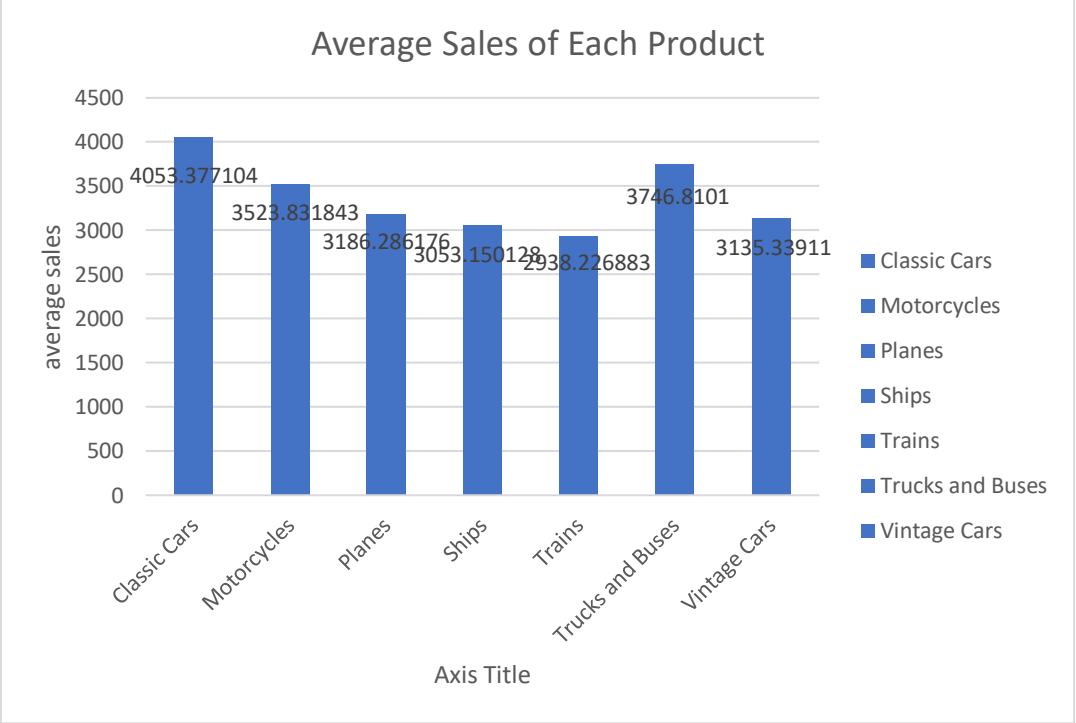
Analytics:

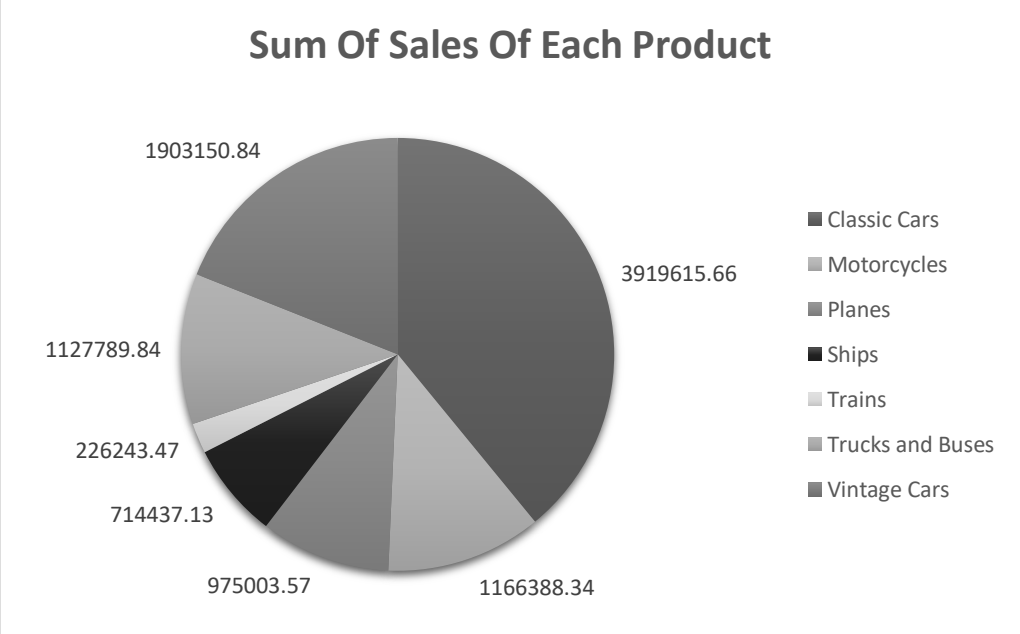
1. Comparison of sales between Vintage cars and Classic cars across all countries.



Ans:- 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. This is represented by using line graph.

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

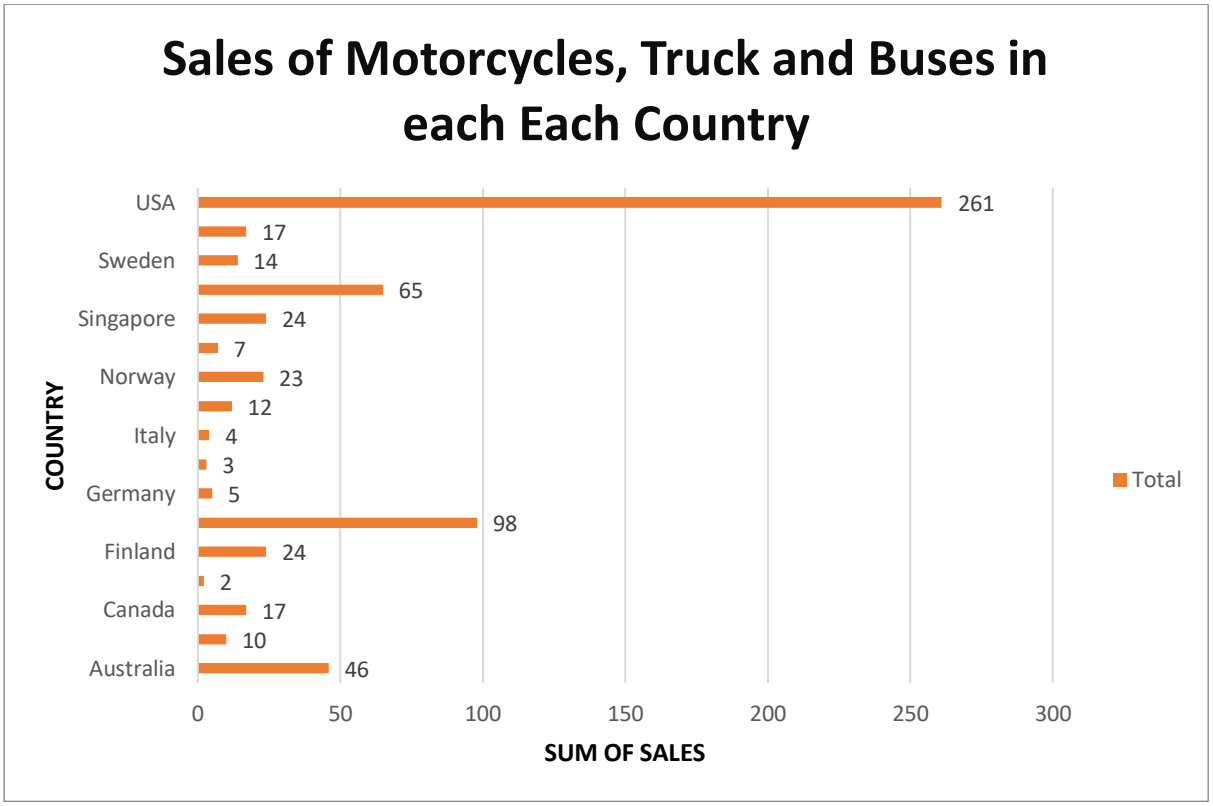




Ans:- By analysing the data, we can see that classic cars have the highest average sales.

From the following charts – sum of sales & average sales, extracted through the given database, we can clearly conclude that Classic cars has the most sales among all.

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



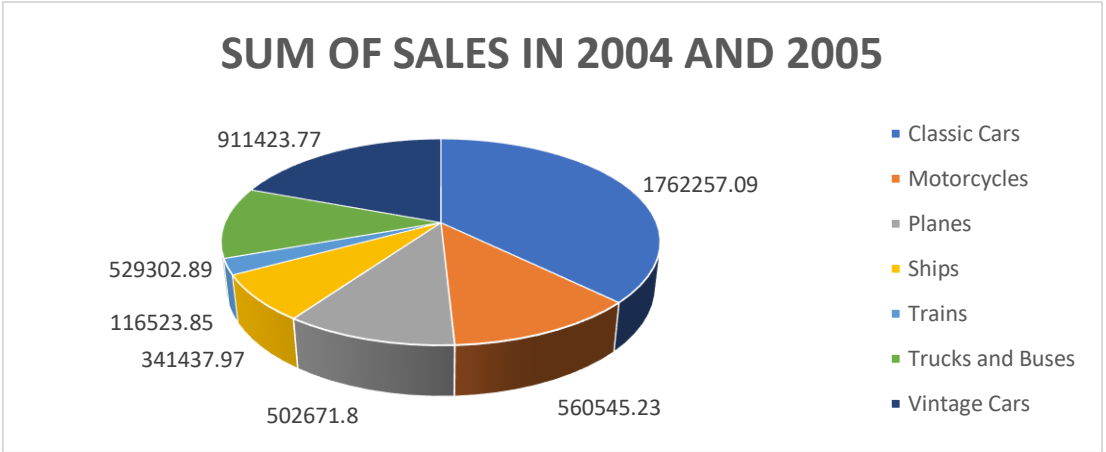
Ans: Based on the analysing provided data, the country that yields the most profit for Motorcycles, Trucks, and Buses is the USA, with total sales of $918,214.12.

The USA has the highest sum of sales compared to all other countries, contributing significantly to the overall total.

The total sales from the USA ($918,214.12) are substantially higher than the next highest, France, which has sales of $343,372.53.

This dominant figure underscores the USA as the leading market in terms of profitability for Motorcycles, Trucks, and Buses, making it the most profitable country in this segment according to the data provided.

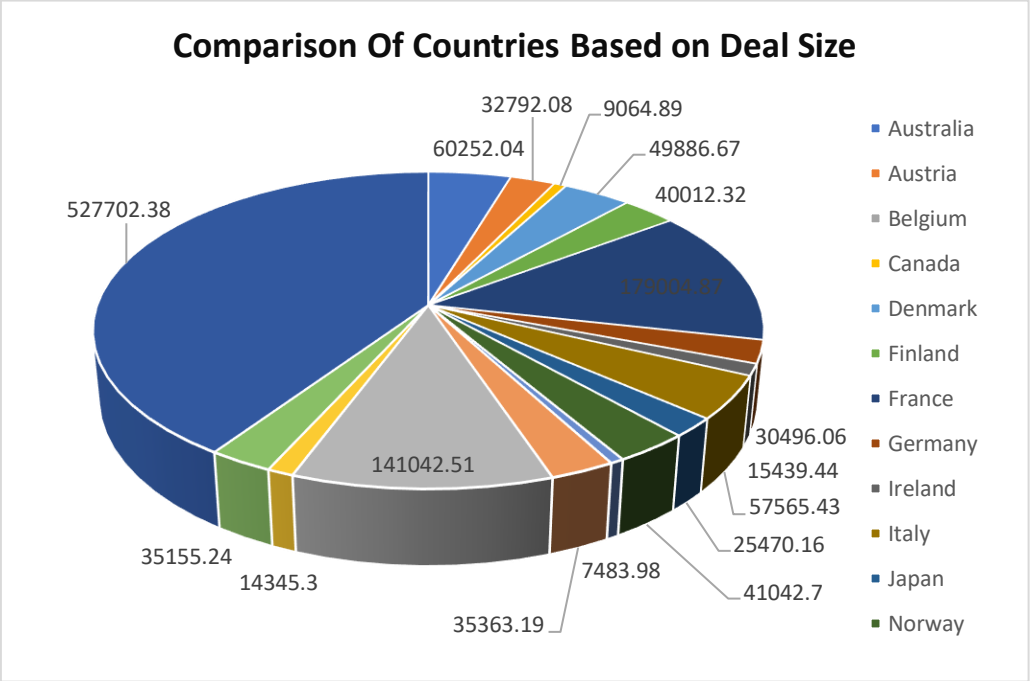
4. Comparison of sales for all items across the years 2004 and 2005.



Ans:- The data reveals a significant decline in sales from 2004 to 2005 across all product lines. Total sales dropped from $4,724,162.60 in 2004 to $1,791,486.71 in 2005, a decrease of over 60%. Classic Cars, the top-performing product line in both years, saw a substantial reduction in sales, from $1,762,257.09 in 2004 to $672,573.28 in 2005. Vintage Cars and Motorcycles also experienced notable declines, with Vintage Cars dropping from $911,423.77 to $340,739.31, and Motorcycles from $560,545.23 to $234,947.53. While the decrease in sales for Trains was less severe, it still reflected the overall downward trend.

These figures indicate that 2005 was a challenging year for sales across the board, with Classic Cars being particularly affected.

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



Ans:- This analysis seeks to uncover the distribution of deal sizes across different countries. The bar chart reveals that the deal sizes in the USA are notably higher compared to other countries, with a large deal size of 64, a medium deal size of 505, and a small deal size of 435.

Conclusion & Review:

The analysis reveals crucial insights into sales dynamics and profitability across various categories and countries. Notably, the USA emerges as a pivotal market leader, displaying robust sales performance in Vintage and Classic cars, Trucks, Buses, and Motorcycles. Classic Cars notably lead as the highest-selling product, making a substantial contribution to overall sales revenue. Moreover, the USA demonstrates exceptional profitability, particularly in the Trucks, Buses, and Motorcycles categories. Sales for Classic cars maintain a consistently strong trajectory throughout the years 2004 and 2005, indicating sustained demand for this category. Additionally, the USA showcases significantly larger deal sizes compared to other countries, highlighting its dominance in sales volume.

Regression:

The regression analysis indicates a strong relationship between the predictor variables (X Variable 1, X Variable 2, and X Variable 3) and the dependent variable. The coefficients for each predictor variable show their impact on the dependent variable, with all variables demonstrating statistical significance.

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	*F*	*Significance F*
Regression	3	273.6567	4.62E-78
Residual	246		
Total	249		

	Coefficients	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5271.93	4.32E-41	-5907.96	-4635.9	-5907.96	-4635.9
X Variable 1	103.0809	5.42E-44	91.26071	114.9011	91.26071	114.9011
X Variable 2	12.81807	3.04E-13	9.545024	16.09111	9.545024	16.09111
X Variable 3	47.42944	1.13E-33	40.82925	54.02963	40.82925	54.02963

Anova (One factor)

The ANOVA results indicate a significant difference between the means of the two groups (Column 1 and Column 2), as evidenced by the extremely small p-value (3.1E-113).

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	250	903280.9	3613.123	3445221
Column 2	250	25534	102.136	1664.552

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.54E+09	1	1.54E+09	894.0704	3.1E-113	3.860199
Within Groups	8.58E+08	498	1723443			
Total	2.4E+09	499				

Anova (Two factor):

The ANOVA results reveal significant differences between the means of the columns (p-value: 1.9E-168), while no significant variation is observed among the rows (p-value: 0.33951). This suggests that the factors represented by the columns have a substantial impact on the data. Additionally, the overall model shows statistical significance, emphasizing the relevance of the analyzed factors in explaining the observed variations.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	4097.66	1365.887	5069957
Row 2	3	2451.12	817.04	1725170
Row 3	3	1566	522	648687
Row 4	3	5095.24	1698.413	7507173
Row 5	3	5140.39	1713.463	7650609
Row 248	3	4386.35	1462.117	5944534
Row 249	3	2261.6	753.8667	1546167
Row 250	3	4176.72	1392.24	5420980
Column 1	250	903280.9	3613.123	3445221
Column 2	250	25534	102.136	1664.552
Column 3	250	8659	34.636	89.69428

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	2.95E+08	249	1182944	1.044989	0.33951	1.194432
Columns	2.09E+09	2	1.05E+09	925.2361	1.9E-168	3.013826
Error	5.64E+08	498	1132016			
Total	2.95E+09	749				

Descriptive Statistics:

The descriptive statistics highlight significant differences among the four columns. Column 2 stands out with its wide range, notably higher mean (3613.123), and substantial right skew, indicating potentially diverse or extreme values. Conversely, Columns 1 and 4 exhibit narrower ranges and lower means, with Column 4 showing a significant left skew. Column 3 falls between the extremes, suggesting a more moderate distribution. These disparities underscore the importance of considering each column's unique characteristics when analysing the dataset, as they may influence interpretations and conclusions drawn from the data.

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

Correlation:

The correlation matrix reveals the relationships between the variables in the dataset. Notably, Column 1 shows a moderate positive correlation with Column 2 (0.513951) and a weaker positive correlation with Column 3 (-0.01254). Column 2 exhibits a stronger positive correlation with Column 3 (0.663973). These correlations provide valuable insights into how the variables co-vary, aiding in understanding their interdependencies and potential associations within the dataset.

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

COOKIE DATA ANALYSIS

Introduction:

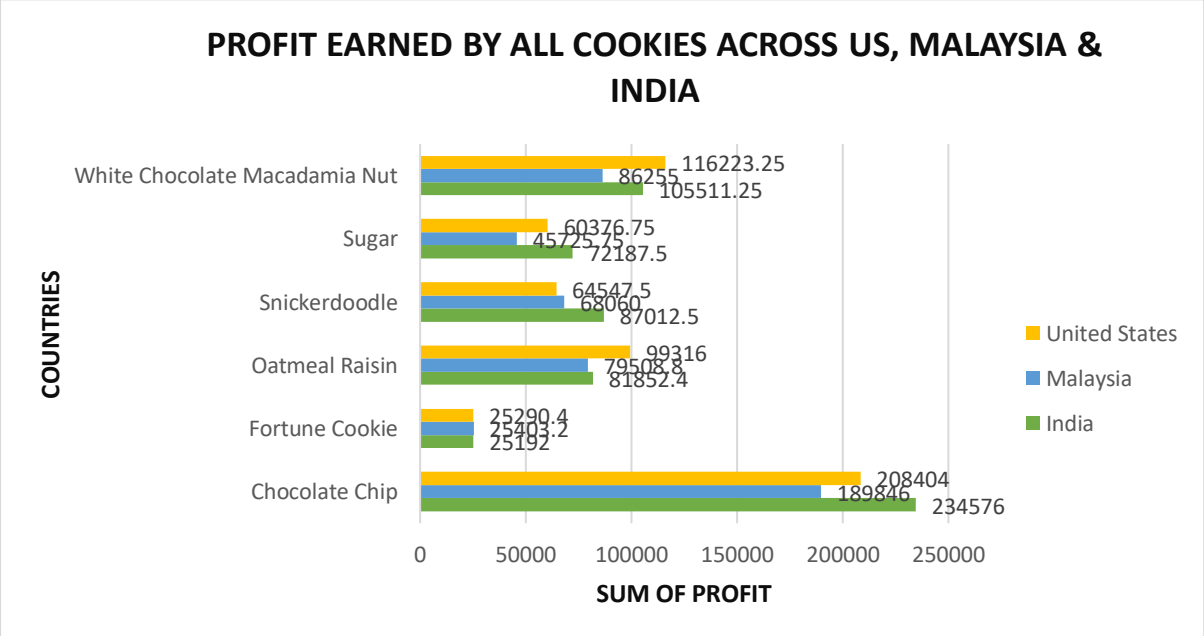
In our cookie dataset, we have detailed information on six types of cookies: Chocolate Chip, Fortune Cookie, Sugar, Oatmeal Raisin, Snickerdoodle, and White Chocolate Macadamia Nut. This dataset encompasses sales volumes, costs, revenue, and profits for these cookies across various countries and dates. Beyond simply analysing cookies, this report delves into consumer preferences, pricing dynamics, and regional popularity trends. By exploring these insights, businesses can gain valuable understanding of market preferences and opportunities within the cookie industry. Get ready to uncover intriguing insights that could have significant implications for businesses like yours.

Questionnaires:

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?

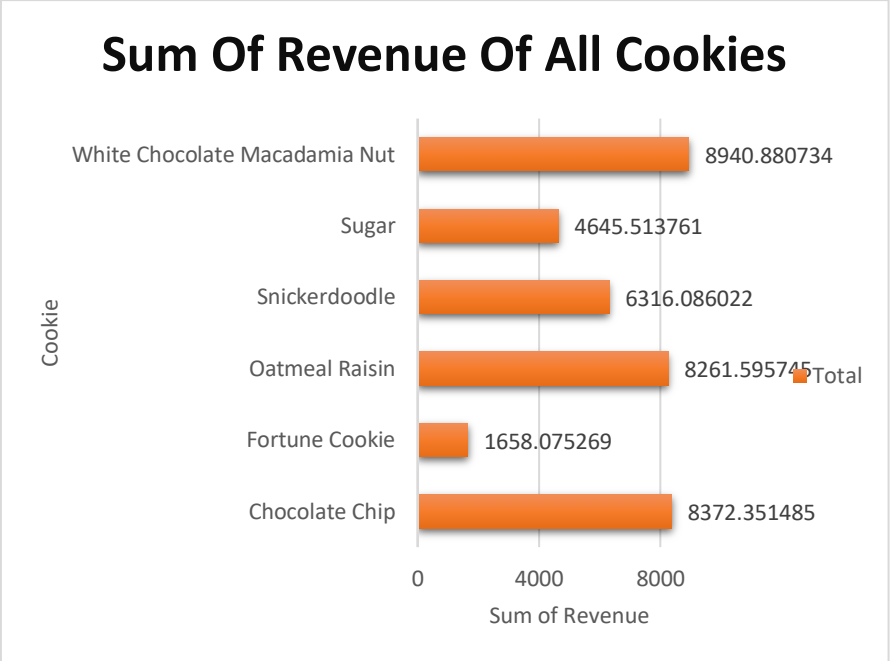
Analytics:

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



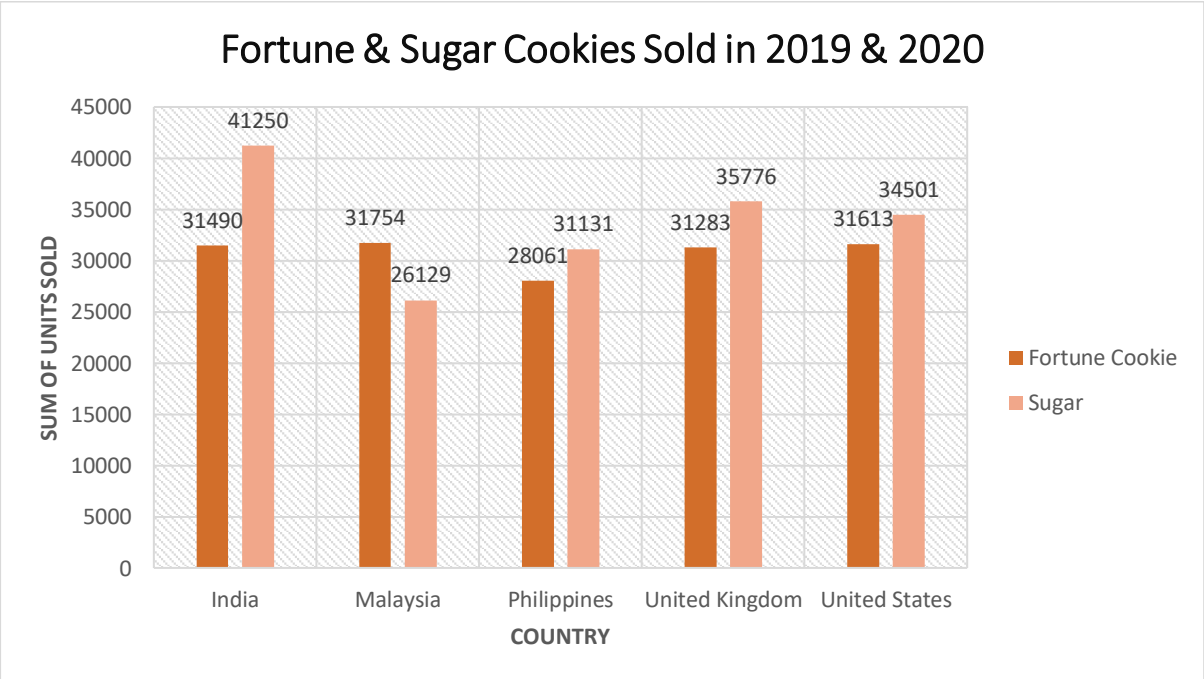
Ans:- This analysis examines the profits generated by all cookie types in three different countries: the United States, Malaysia, and India. The highest profit for Chocolate Chip cookies is observed in India, followed by Malaysia and the United States.

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



Ans:- This analysis aims to present the average revenue generated by each cookie type. It is evident that White Chocolate Macadamia Nut generates the highest average revenue at $8,940.88, followed by Chocolate Chip.

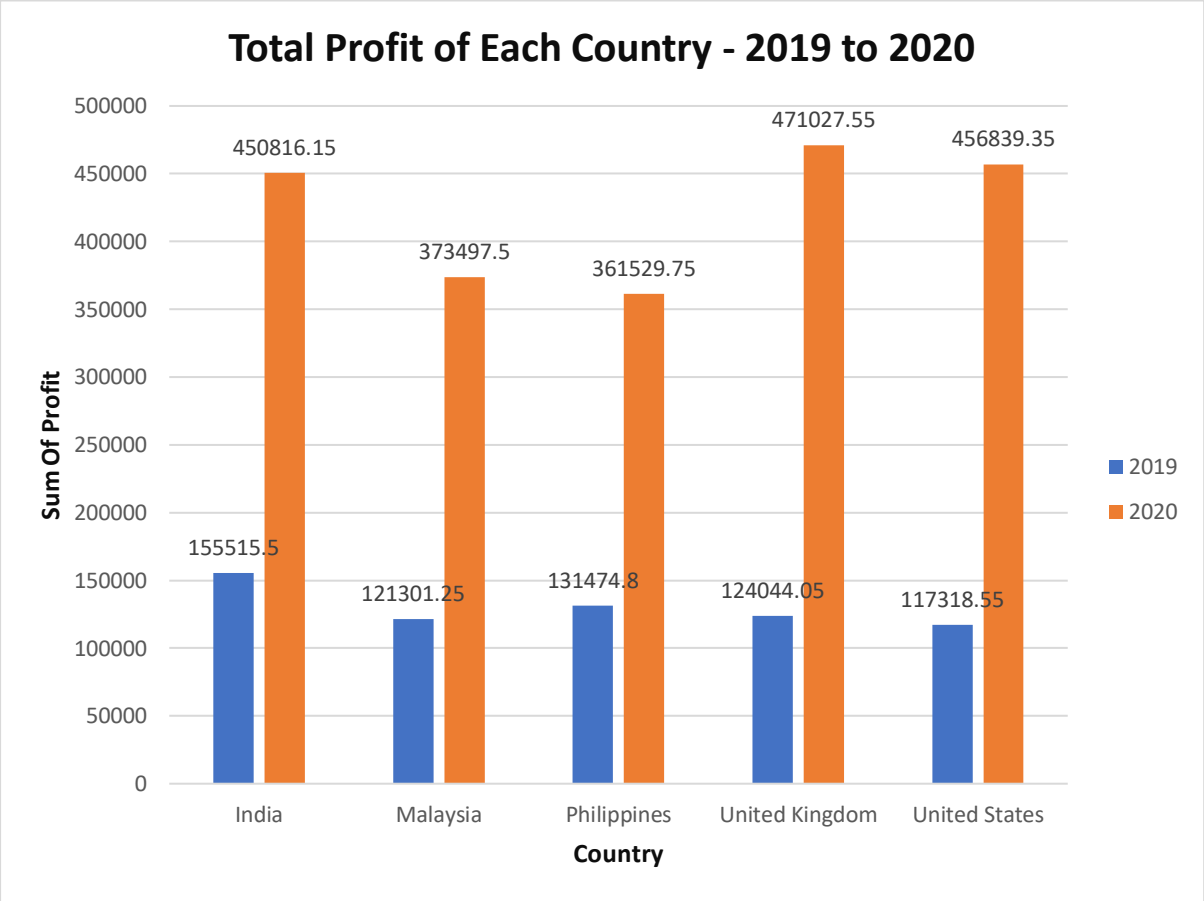
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?



Ans:- This analysis seeks to compare the sales of Fortune and Sugar cookies across different countries for the years 2019 and 2020.

For Sugar cookies, India records the highest sales of 41,250 units, followed by UK with 35,776 units. On the other hand, the Malaysia leads in Fortune cookie sales with 31,754 units sold, followed by the United States with 31,613 units.

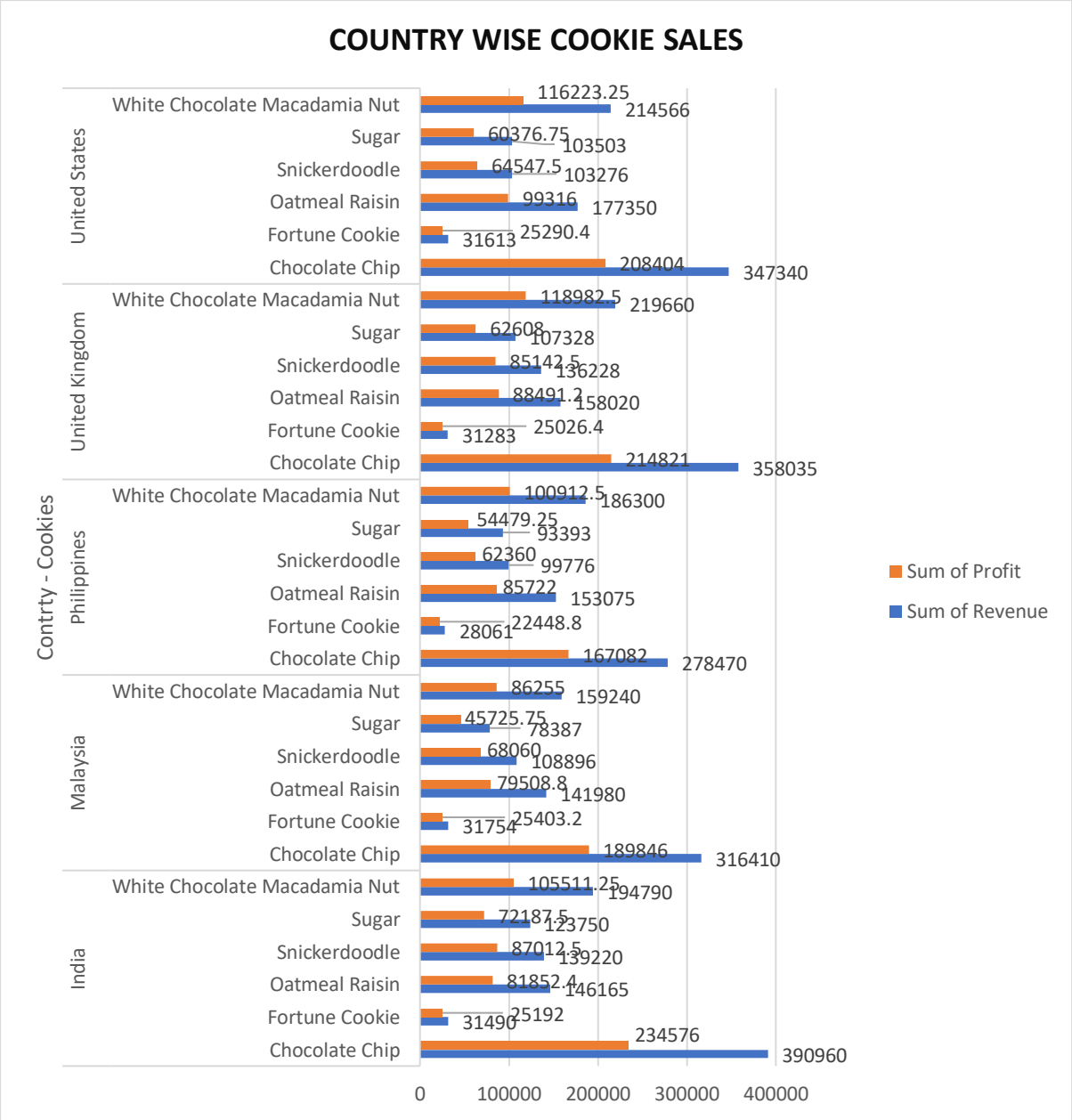
4. Compare the performance of all the countries for the year 2019 to 2020. Which country perform in each of these years?



Ans:- This analysis aims to compare the profits earned by countries in the financial years 2019 and 2020. According to the graph, the United Kingdom demonstrates the highest profit earned in 2020, amounting to $471,027.55 in sales, followed closely by the United States with $456,839.35.

In 2019, the highest profit was recorded by India, totalling $155,515.5 in sales, followed by the Philippines with $131,474.8.

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



Ans:- This analysis identifies the cookie category sold for the highest price in each country. Chocolate Chip cookies yield the highest revenue, and Sugar cookies generate the most profit, particularly in India followed by the United Kingdom.

Conclusion and Review:

The analysis provided insights into the profit earned by different cookie types in the US, Malaysia, and India. India emerged with the highest profit for chocolate chip cookies, followed by Malaysia and the United States. White chocolate macadamia nut cookies generated the highest average revenue, closely followed by chocolate chip cookies.

In terms of sales, India showed significant sales of sugar cookies in 2020, while the United Kingdom had the highest sales of sugar cookies in 2019. For fortune cookies, India and Malaysia exhibited higher sales in both years, with the Philippines and the United States also contributing notable sales.

Regarding profit comparison by country for 2019 and 2020, the United Kingdom recorded the highest profit in 2020, followed by the United States. In 2019, India had the highest profit, followed by the Philippines. Chocolate chip cookies were sold for the highest price in terms of revenue, while sugar cookies generated the highest profit overall.

Regression:

The regression analysis results indicate a perfect fit (R-squared value of 1) between the dependent variable, Profit, and the independent variables: Units Sold, Price, and Revenue. The extremely small error value (9.16E-12) suggests minimal residual error in the model. With 700 observations, this regression model effectively captures the relationship between Profit and the specified independent variables.

SUMMARY OUTPUT

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

ANOVA:

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

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

Anova (One factor):

This ANOVA analysis indicates that both factors, represented by rows and columns, significantly influence the dataset's variability.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	700	1926955	2752.792	4149401
Column 2	700	2763364	3947.664	6842519

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	5E+08	1	5E+08	90.92153	6.36E-21	3.848119
Within Groups	7.68E+09	1398	5495960			
Total	8.18E+09	1399				

Anova (Two factor):

The ANOVA analysis highlights that both product type and sale date si
influence the dataset's variability in terms of profit.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	17250	5750	6943125
Row 2	3	21520	7173.333	10805909
Row 3	3	23490	7830	12874869
Row 4	3	12280	4093.333	3518629
Row 5	3	13890	4630	4501749
Column 1	700	4690319	6700.456	21380458
Column 2	700	1926955	2752.792	4149401
Column 3	700	2763364	3947.664	6842519

ANOVA

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

Descriptive Statistics:

These descriptive statistics offer insights into the distribution and characteristics of the variables within the dataset, including profit and potentially other factors. They provide a comprehensive overview of the data's central tendency, variability, and distribution, aiding in understanding the dataset's underlying patterns and trends.

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

Correlation:

the correlation analysis reveals strong positive relationships among the key financial metrics represented in the dataset. Specifically, there are moderate to strong positive correlations between sales volume ("Units Sold"), revenue, expenses ("Cost"), and profit. These findings suggest that higher sales volumes tend to result in increased revenue, while higher revenues correlate with higher expenses and ultimately higher profits.

	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

STORE DATA ANALYSIS

Introduction:

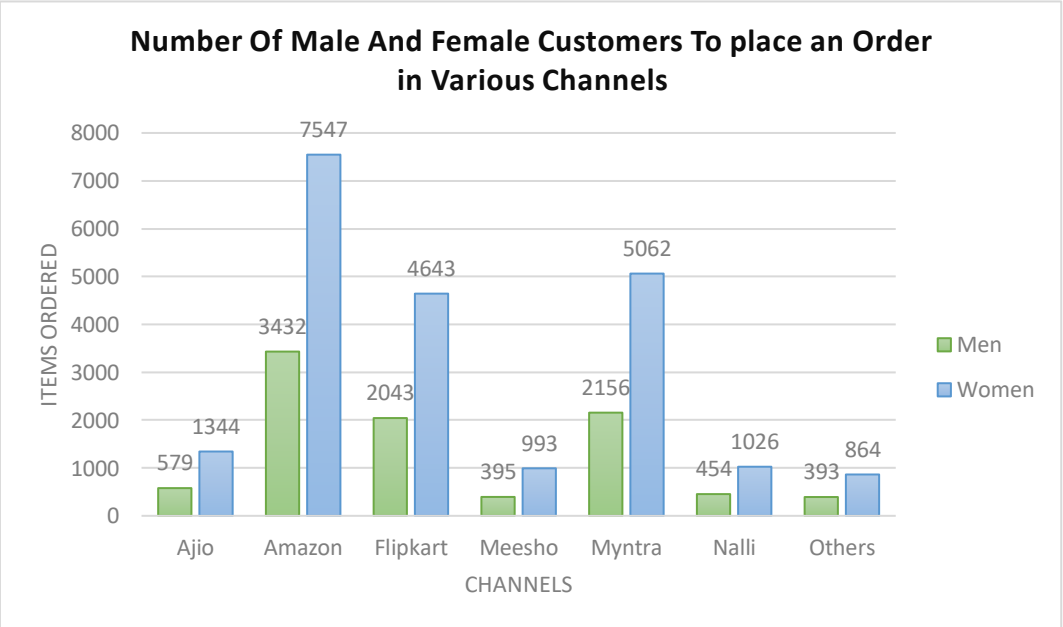
This dataset contains sales information from a retail outlet, including diverse details like customer demographics (Gender, Age Group), transaction specifics (Order ID, Status), product details (Category, SKU), and shipping data. Our examination focuses on understanding customer actions and product patterns, aiming to reveal trends, preferences, and associations present in the dataset. Leveraging these insights, companies can enhance marketing approaches, optimize inventory control, and boost overall customer contentment.

Questionnaires:

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:

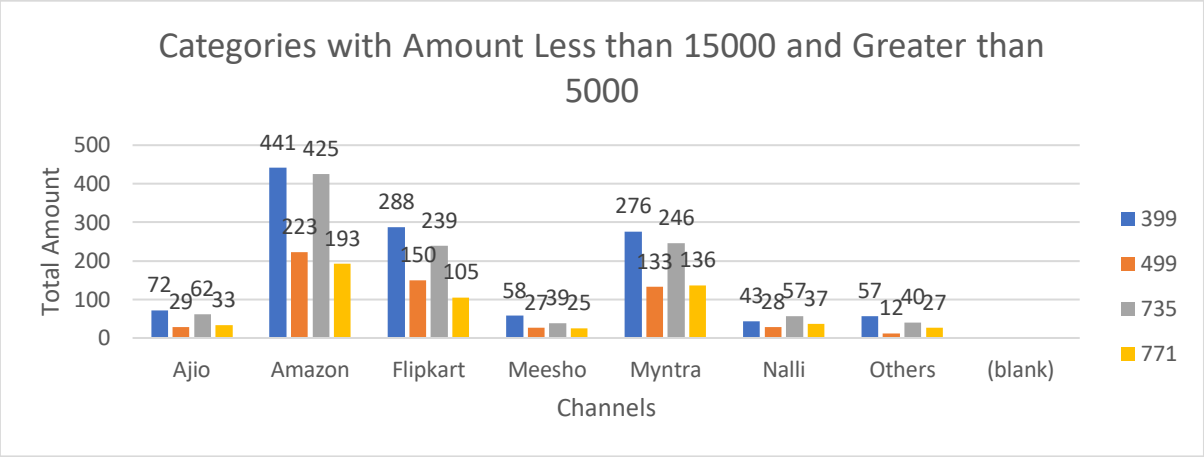
1. Compare various channels based on how many male customers order and female customer order?



Ans:- Amazon dominates sales in both the men's and women's categories, followed closely by

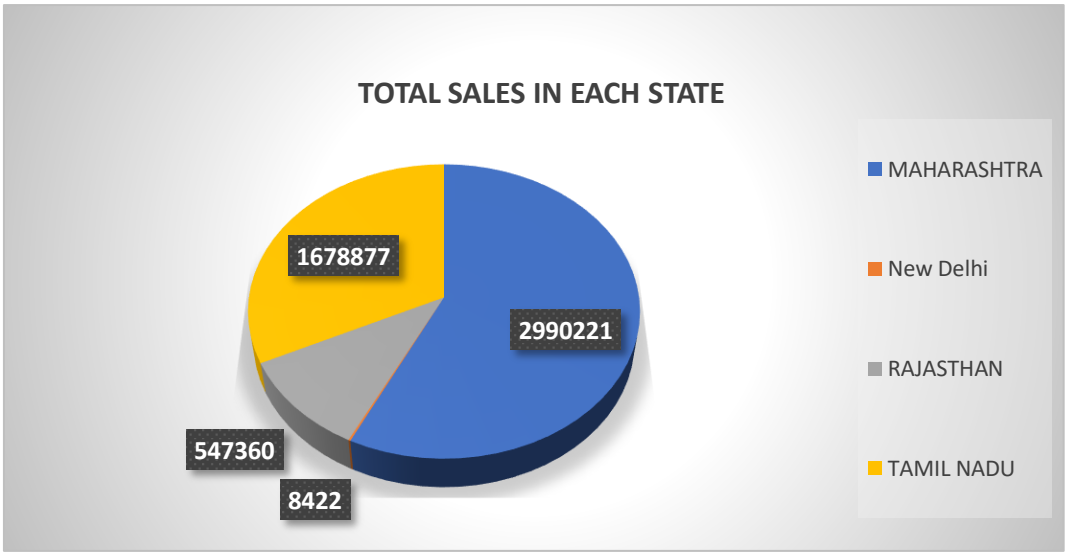
Myntra and Flipkart. Specifically, Amazon sold approximately 3432 units in the men's category and nearly 7547 units in the women's category. In comparison, Myntra recorded sales of 2156 units in the men's section and 5062 units in the women's section.

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



Ans:-This analysis facilitates the comparison of order categories based on their amounts, specifically focusing on orders with amounts less than 1500 and greater than 5000. It reveals that Kurta and Set have the highest count of orders.

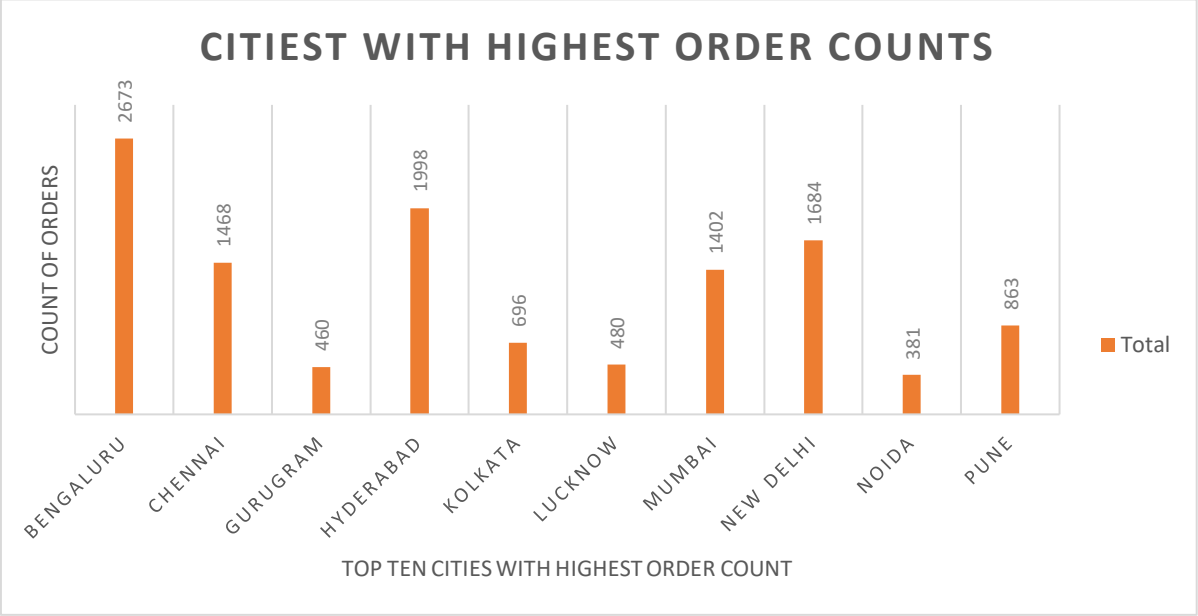
3. Which of the following state perform better than other, Delhi, Tamil Nadu, Maharashtra, Rajasthan.



Ans:- From the analysis of the following chart, we can see that in Maharashtra, total sales are 29,90,221 units which is the highest among the four states while New Delhi has the least sales of 8,422 units.

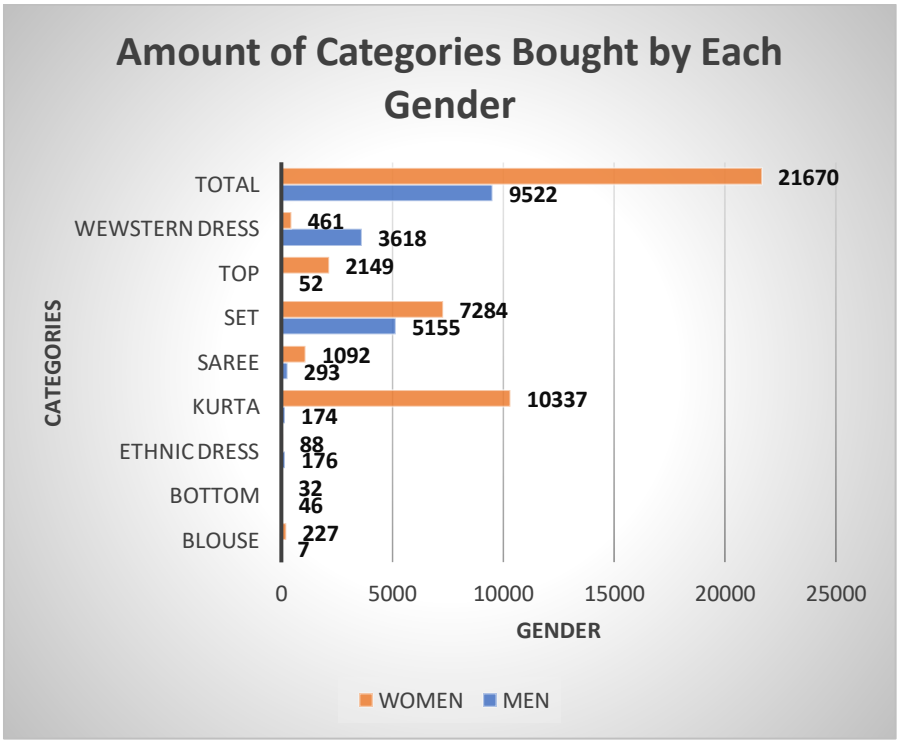
Therefore, among the four states, New Delhi's performance was least while Rajasthan performed comparatively better with sales of 5,47,360 units. Tamil Nadu performed better than both New Delhi and Rajasthan with a total sales of 12,78,877 units and Maharashtra performed the best.

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



Ans:- According to the recorded graph, Bangalore emerges as the city with the highest number of orders placed, totaling 2,673 orders, followed by Hyderabad with 1,998 orders.

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



Ans:- This analysis compares various categories of items based on the quantity sold, revealing that Kurta purchased by women have the highest quantity sold, followed by Set. While men's purchases of Set is the highest among all categories.

Conclusion and Review:

The analysis underscores Amazon's dominance in sales across both men's and women's categories, with Myntra and Flipkart following closely behind. Amazon leads in sales for both categories, followed by Myntra and Flipkart. The top-selling items include kurta and set, with Karnataka and Bangalore showing the highest sales performance.
This analysis offers valuable insights into sales trends and regional performance, assisting retailers in making informed decisions. However, delving deeper into additional factors influencing sales could further enhance the analysis. Overall, the findings provide crucial information for optimizing sales strategies in competitive markets.

Regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.172398
R Square	0.029721
Adjusted R Square	0.029659
Standard Error	264.5693
Observations	31047

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	66561870	33280935	475.4629	0
Residual	31044	2.17E+09	69996.92		
Total	31046	2.24E+09			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	185.155	16.57854	11.16836	6.61E-29	152.6604	217.6496	152.6604	217.6496
X Variable 1	0.047626	0.099327	0.479489	0.631594	-0.14706	0.242312	-0.14706	0.242312
X Variable 2	492.0276	15.95904	30.83065	1.3E-205	460.7472	523.308	460.7472	523.308

Anova (One 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*	*P-value*	*F crit*
Between Groups	7.2E+09	1	7.2E+09	199639.8	0	3.841609
Within Groups	2.24E+09	62092	36068.2			
Total	9.44E+09	62093				

Anova (Two factor):

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	421	140.3333	42116.33
Row 2	3	1479	493	685648
Row 3	3	521	173.6667	59609.33
Row 4	3	750	250	172171
Row 5	3	607	202.3333	88482.33
Row 31044	3	974	324.6667	283326.3
Row 31045	3	1145	381.6667	403529.3
Row 31046	3	446	148.6667	47506.33
Row 31047	3	828	276	199225
Column 1	31047	1226250	39.49657	228.5307
Column 2	31047	31237	1.00612	0.008853
Column 3	31047	21176377	682.0748	72136.38

ANOVA

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

Descriptive Statistics:

Column1		*Column2*		*Column3*	
Mean	39.49657	Mean	1.00612	Mean	682.0748
Standard Error	0.085795	Standard Error	0.000534	Standard Error	1.524289
Median	37	Median	1	Median	646
Mode	28	Mode	1	Mode	399
Standard Deviation	15.11723	Standard Deviation	0.094088	Standard Deviation	268.5822
Sample Variance	228.5307	Sample Variance	0.008853	Sample Variance	72136.38
Kurtosis	-0.1587	Kurtosis	475.3566	Kurtosis	1.768676
Skewness	0.72916	Skewness	19.4509	Skewness	1.052904

Range	60	Range	4	Range	2807
Minimum	18	Minimum	1	Minimum	229
Maximum	78	Maximum	5	Maximum	3036
Sum	1226250	Sum	31237	Sum	21176377
Count	31047	Count	31047	Count	31047

Correlation:

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

CAR COLLECTION DATA ANALYSIS

Introduction:

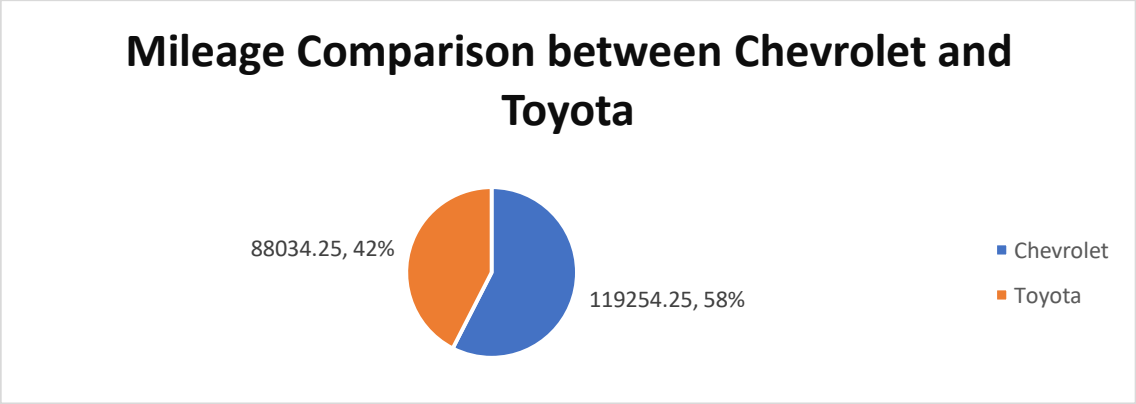
The dataset provides comprehensive information about various cars, including their make, model, colour, mileage, price, and cost. Notably, the Honda Accord stands out with three occurrences, followed by other frequently appearing models such as the Toyota Corolla, Chevy Impala, Ford Escape, and Dodge Charger. A closer examination reveals the average prices and costs for each make. On average, Hondas are priced at approximately $3,106, with costs averaging around $2,133, while Chevys have an average price of $3,487 and average cost of $3,000. Further analysis will include plotting graphs to explore the potential relationship between a car's price and mileage, as well as determining colour preferences among consumers. Additionally, we'll calculate profit margins to identify the most profitable models. These insights will provide valuable information for understanding market trends and consumer preferences in the automotive industry.

Questionnaires:

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

Analytics:

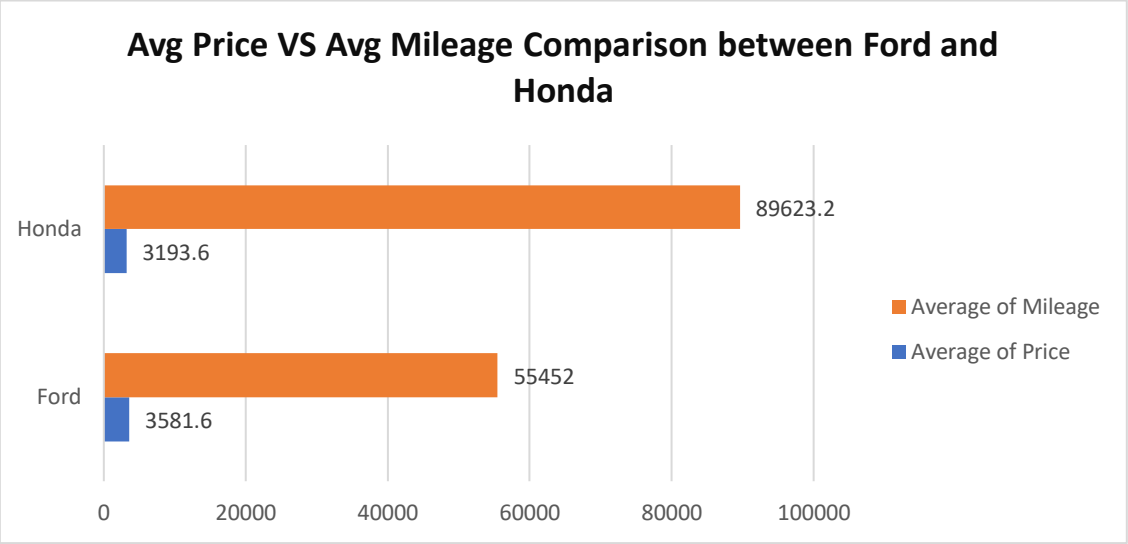
1. Compare the mileage of Chevrolet Impala to Toyota Corolla. Which of the two is giving best mileage?



Ans:- Based on the given data, by analysing, we can draw the conclusion that Chevrolet Impala has a higher average mileage compared to Toyota Corolla.

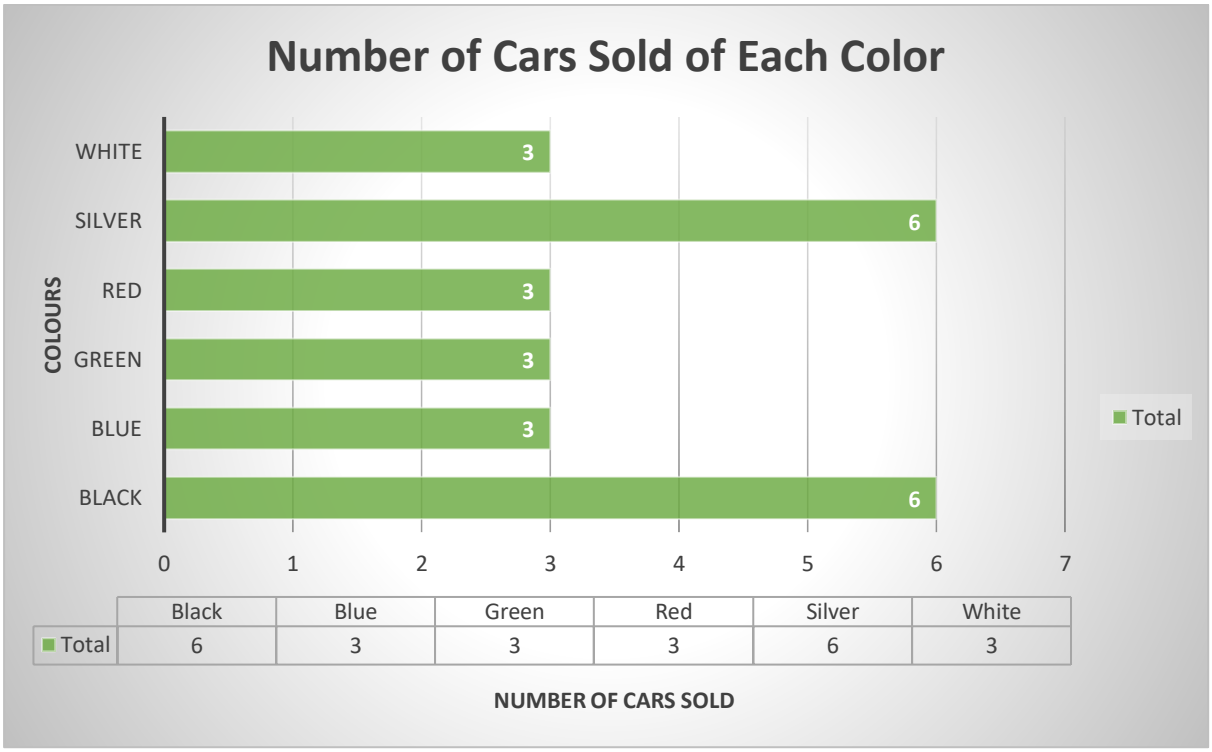
Chevrolet Impala: 119,254.25 miles
Toyota Corolla: 88,034.25 miles
Therefore, the Chevrolet Impala provides better mileage than the Toyota Corolla.

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



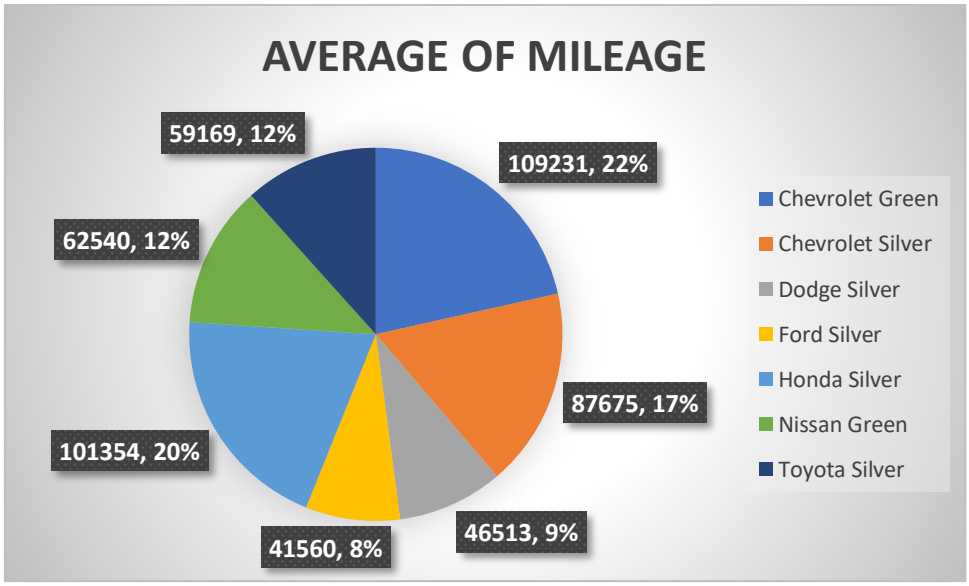
Ans:- This analysis seeks to justify purchasing a Ford car over a Honda by comparing their respective attributes, with a particular focus on price. However, after analysing the dataset, the findings do not support this statement. Instead, Honda cars were found to have better average mileage (89,623.3) and a lower average price (3,193.6) compared to Ford cars.

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



Ans:- Most popular color is Silver and Black as each appear 6 times
and least appearing colour are Blue, Green, Red, White they all appear 3 times

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

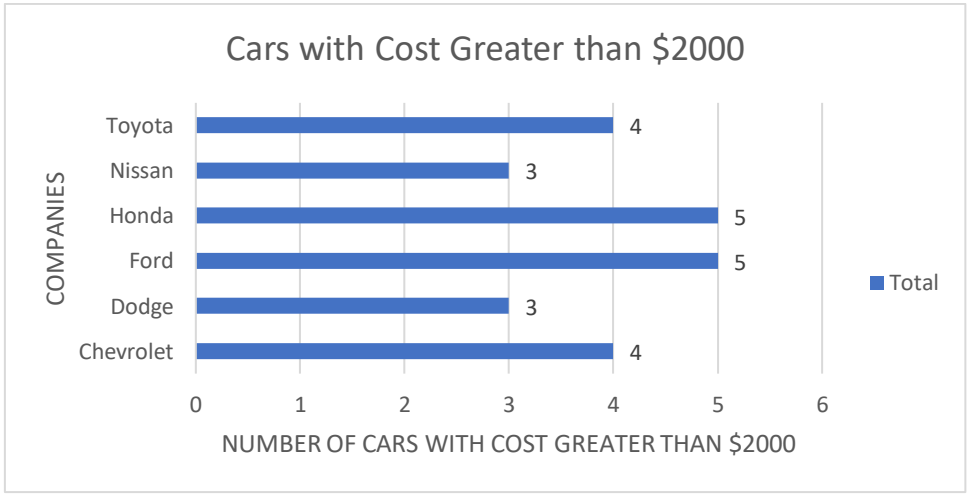


Ans:- Based on the analysis performed:

Silver Cars:
The average mileage for silver cars shows a broad range, with some models having significantly higher or lower mileage than others.
Chevrolet and Honda silver cars have relatively high mileage, while Ford and Dodge have lower mileage.
Green Cars:
Green cars generally exhibit higher average mileage compared to silver cars.
Both Chevrolet and Nissan green cars show strong mileage performance, with the green Chevrolet having notably higher mileage compared to its silver counterpart.

Therefore, we can say that Green cars tend to have better mileage performance overall when compared to silver cars, particularly evident in models like the Chevrolet.

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



Ans:- All the car mention below cost is more than $2000:
Toyota, Nissan, Honda, Ford, Dodge & Chevrolet.

Conclusion and Review:

Our analysis sheds light on what consumers look for when buying cars. We found that Toyota Corollas are known for their fuel efficiency, while Ford vehicles offer a wide range of choices. Consumers seem to prefer black and red cars. Interestingly, silver cars tend to have higher mileage. These findings highlight the importance of thinking about things like gas mileage, colour preference, and budget when shopping for a car.

Regression:

The regression analysis examined the relationship between mileage (as the dependent variable) and cost and price (as the independent variables). The resulting R-squared value is 0.926673, indicating that approximately 92.67% of the variability in mileage can be explained by the cost and price in the dataset.

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

Anova (One factor):

ANOVA (Analysis of Variance) is used to analyze the differences among group means in a sample. The one-factor ANOVA provides a summary of the columns, including count, sum, average, and variance. It also includes the sources of variance, with the sum of squares (SS) and degrees of freedom (df). For the three columns—mileage, price, and cost—the count for column 1, column 2, and column 3 is shown below. The analysis reveals that the total SS value is 1.32E+11 and the total df value is 71.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
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*
Between Groups	1.04E+11	2	5.22E+10	128.8822	5E-24	3.129644
Within Groups	2.8E+10	69	4.05E+08			
Total	1.32E+11	71				

Anova (Two factor):

A two-factor ANOVA without replication is a statistical method used to analyse the effects of two factors on a dataset without multiple observations for each combination of factors. It allows testing for the main effects of each factor as well as their interaction. In this analysis, the variance in the dataset for each row is shown with the sum of squares (SS) and degrees of freedom (df) associated with each factor. This provides insight into how much variation in the data can be attributed to each factor individually.

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				

Descriptive Statistics:

Descriptive Statistics shows the Mean, Standard Error, Standard Deviation, Median and Mode etc for the price, cost, and mileage.

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

Correlation:

Correlation shows the relationship between the two columns having the numeric data factors.

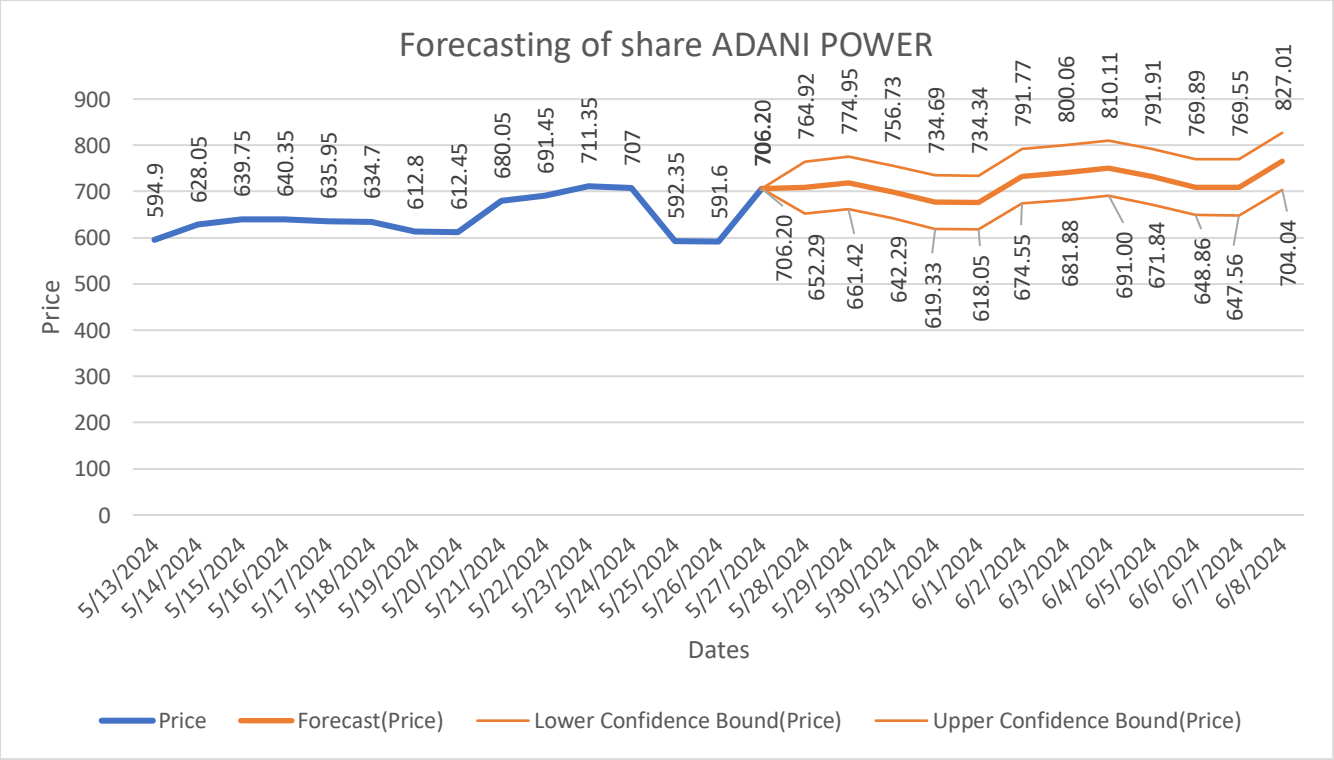
	Column 1	*Column 2*
Column 1	1	
Column 2	-0.41106	1

Forecasting of share Adani Power

The provided data illustrates the historical and forecasted share prices for Adani Powers from May 13, 2024, to June 8, 2024. Initially, the actual closing prices are recorded, showing fluctuations between 591.6 on May 26, 2024, and 711.35 on May 23, 2024. From May 27, 2024, onwards, the table presents forecasted prices, starting at 706.2 and increasing to 765.52 by June 8, 2024. Additionally, the forecast includes lower and upper confidence bounds, which denote the range within which the actual prices are anticipated to fall, highlighting the uncertainty in the predictions. These bounds widen over time, indicating increasing uncertainty further into the future. This dataset offers valuable insights into the recent performance and projected trends of Adani Powers' share price, aiding investors and analysts in making informed decisions.

Date	Price	Forecast(Price)	Lower Confidence Bound(Price)	Upper Confidence Bound(Price)
13-05-2024	594.9			
14-05-2024	628.05			
15-05-2024	639.75			
16-05-2024	640.35			
17-05-2024	635.95			
18-05-2024	634.7			
19-05-2024	612.8			
20-05-2024	612.45			
21-05-2024	680.05			
22-05-2024	691.45			
23-05-2024	711.35			
24-05-2024	707			
25-05-2024	592.35			
26-05-2024	591.6			
27-05-2024	706.2	706.2	706.20	706.20
28-05-2024		708.6052644	652.29	764.92

29-05-2024	718.1884955	661.42	774.95
30-05-2024	699.5121484	642.29	756.73
31-05-2024	677.0087043	619.33	734.69
01-06-2024	676.1933597	618.05	734.34
02-06-2024	733.1600752	674.55	791.77
03-06-2024	740.9701658	681.88	800.06
04-06-2024	750.553397	691.00	810.11
05-06-2024	731.8770499	671.84	791.91
06-06-2024	709.3736057	648.86	769.89
07-06-2024	708.5582611	647.56	769.55
08-06-2024	765.5249767	704.04	827.01



Based on the provided data of Adani Powers' share prices, which includes both historical and forecasted values, several conclusions can be drawn. The historical prices from May 13, 2024, to May 26, 2024, show notable volatility, with the lowest price recorded at 591.6 on May 26 and the highest at 711.35 on May 23, indicating a period of instability. From May 27, 2024, onward, the forecasted share prices suggest a generally upward trend, starting at 706.2 and rising to 765.52 by June 8, 2024, implying a positive outlook for the near future. However, the confidence bounds, which provide a range for expected prices, widen over time, reflecting increased uncertainty in longer-term forecasts. For example, the bounds on May 27 are narrow at 706.20, but by June 8, they range from 704.04 to 827.01. This widening indicates greater uncertainty as time progresses. Investors might view the forecasted upward trend as a potential opportunity for gains but should also consider the increasing uncertainty reflected in the confidence bounds. Balancing this optimism with caution is essential for making informed investment decisions.