"Data Analytics Report"

A Data Analytics Report Submitted to
Rajiv Gandhi Proudyogiki Vishwavidyalaya

Towards Partial Fulfillment for the Award of
Bachelor of Engineering in Computer Science Engineering

Submitted by:

Arjav Jain(0827CS211035)

Guided by:

Prof. Anurag Punde
Computer Science and Engineering

Acropolis Institute of Technology & Research, Indore
Jan-June 2024

Name: Arjav Jain

Enrollment Numer :0827CS211035

S.No.	Experiment	Remarks
1.	Data Analysis Questions: i. Data Analysis Principles ii. Statistical Analytics iii. Hypothesis Testing iv. Regression v. Correlation vi. ANOVA	
2.	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	
3.	Forcasting of TCS Shares	

Assignment-1: Key Concepts in Data Analysis

Data Analysis Principles

Data Analysis entails the systematic use of statistical and logical methods to understand, summarize, and evaluate data. Essential principles involve understanding the origin, context, and quality of the data; cleaning the data to remove inaccuracies; exploring the data with descriptive statistics and visualization techniques; modeling the data with statistical methods to make predictions or inferences; and interpreting the results to draw meaningful conclusions and make informed decisions.

Statistical Analysis

Statistical Analysis employs statistical techniques to gather, review, analyze, and derive conclusions from data. This encompasses descriptive statistics (mean, median, mode, range, variance, standard deviation) to summarize data features, inferential statistics (hypothesis testing, confidence intervals, regression analysis) to extend conclusions beyond the immediate data, predictive analytics to forecast future outcomes, and prescriptive analytics to recommend actions based on the data analysis.

Hypothesis Testing

Hypothesis Testing is a method used for decision-making based on data from experiments or studies. It involves a null hypothesis (H0) of no effect or difference and an alternative hypothesis (H1) indicating an effect or difference. The p-value represents the probability of observing the data if H0 is true, with small p-values providing strong evidence against H0. Type I errors (false positives) occur when H0 is incorrectly rejected, and Type II errors (false negatives) occur when H0 is incorrectly not rejected. The significance level (α), often set at 0.05, is the threshold for rejecting H0.

Regression Analysis

Regression Analysis helps understand the relationships between dependent and independent variables. Linear regression fits a linear equation to the data, multiple regression uses several independent variables, logistic regression predicts probabilities for categorical outcomes, and polynomial regression models relationships as nth degree polynomials.

Correlation

Correlation measures the strength and direction of the relationship between two variables using the correlation coefficient (r), which ranges from -1 to 1. A positive correlation means

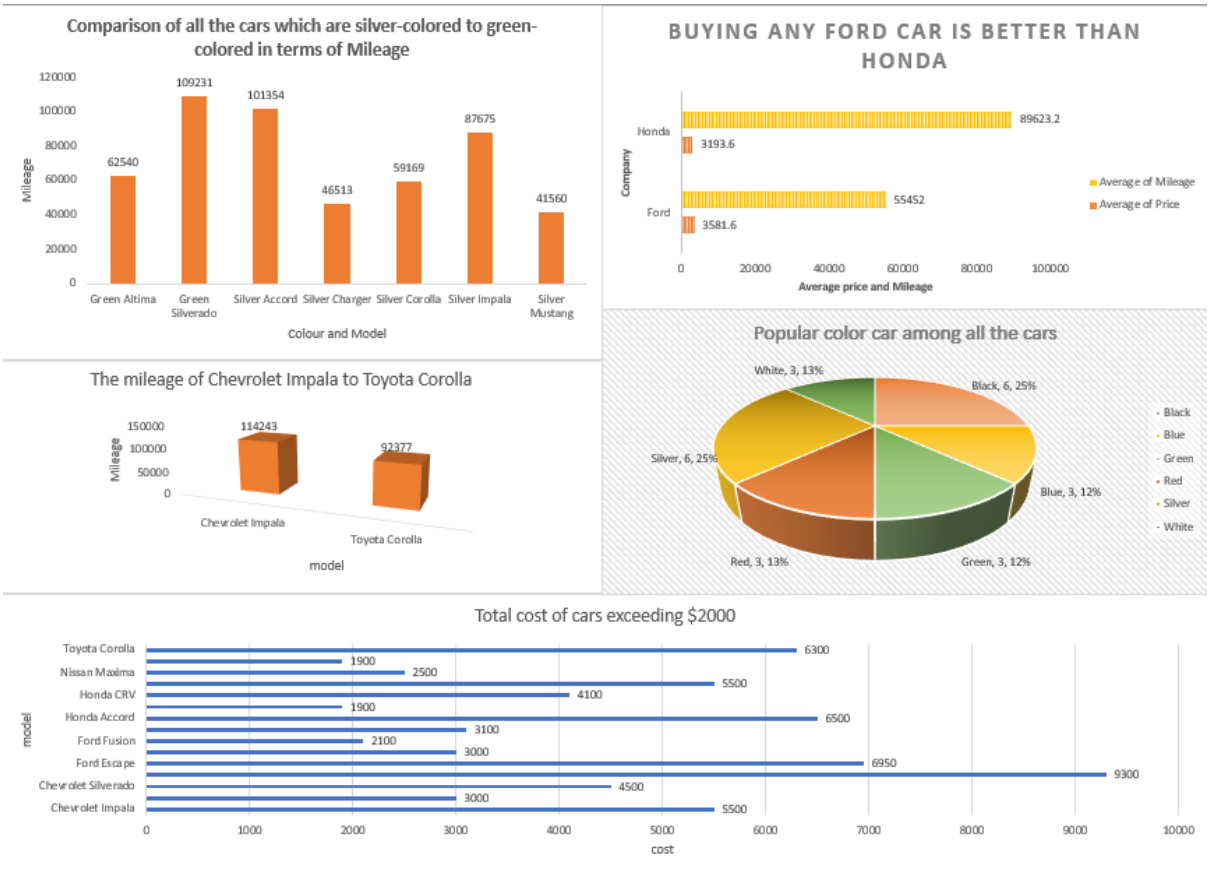
both variables move in the same direction, a negative correlation means one increases as the other decreases, and no correlation indicates no relationship. Importantly, correlation does not imply causation; it merely indicates a relationship between variables.

ANOVA (Analysis of Variance)

ANOVA is a method used to compare means across multiple groups to determine if at least one group mean significantly differs. One-way ANOVA compares means across one factor with multiple levels, while two-way ANOVA examines the influence of two categorical variables. ANOVA assumes normality, homogeneity of variances, and independence of observations. The F-statistic, the ratio of variance between group means to variance within groups, is used to determine the p-value for the test.

Car Collection Data Report

Dashboard



Introduction

A thorough examination of the make, model, colour, mileage, pricing, and cost of many car models is provided by the Car Collection dataset. The purpose of this research is to examine 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.

We have asked a number of important questions and carried out related analyses throughout the analysis 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.

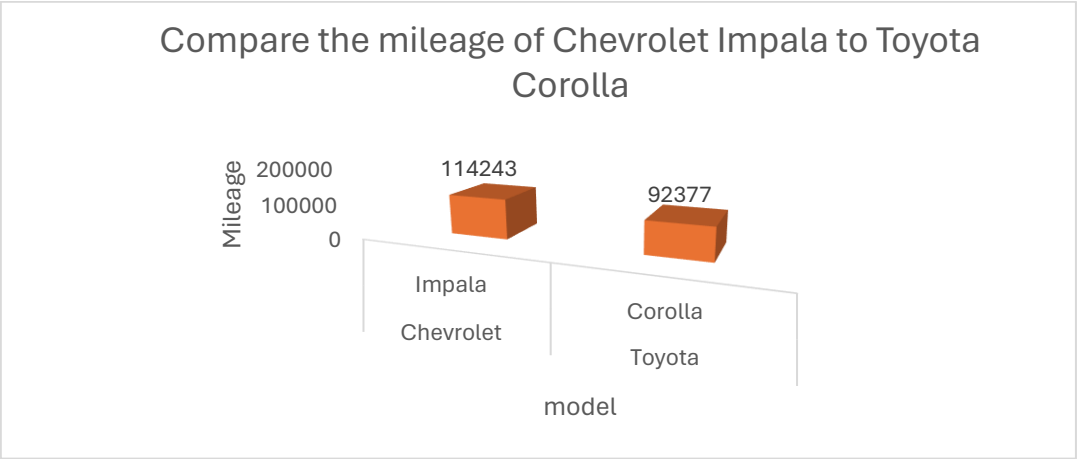
3. Among all the cars which car color is the most popular and is least popular?

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?

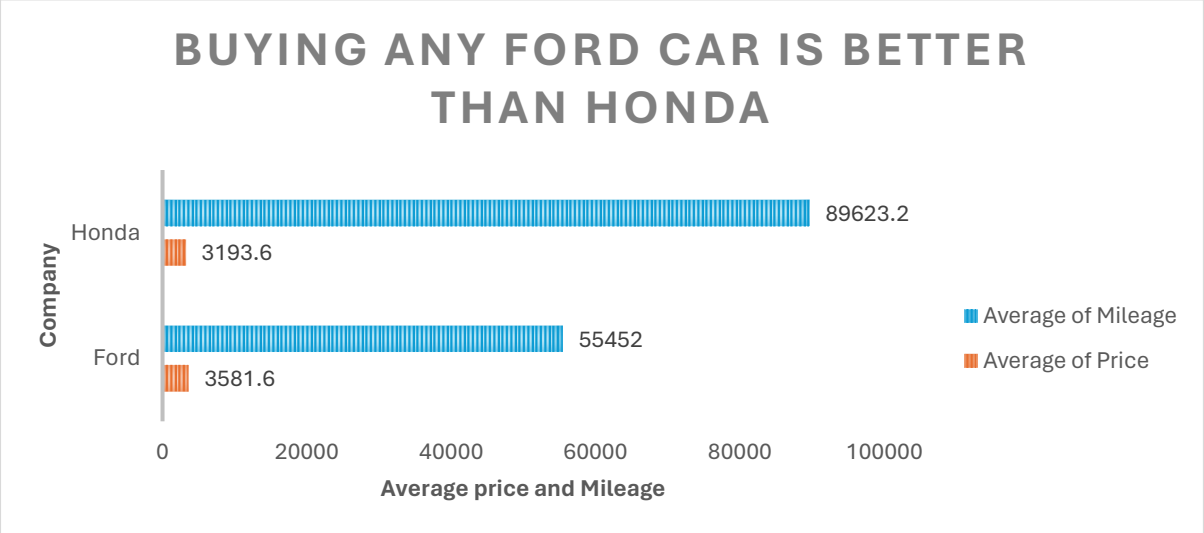
Analytics

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



In this analysis, the fuel efficiency (mileage) of two popular car models—the Chevrolet Impala and the Toyota Corolla—is compared. A column chart is created and the dataset is filtered to isolate data. The analysis's conclusion is that the Chevrolet Impala (114243) has a higher mileage than the Toyota Corolla (92377).

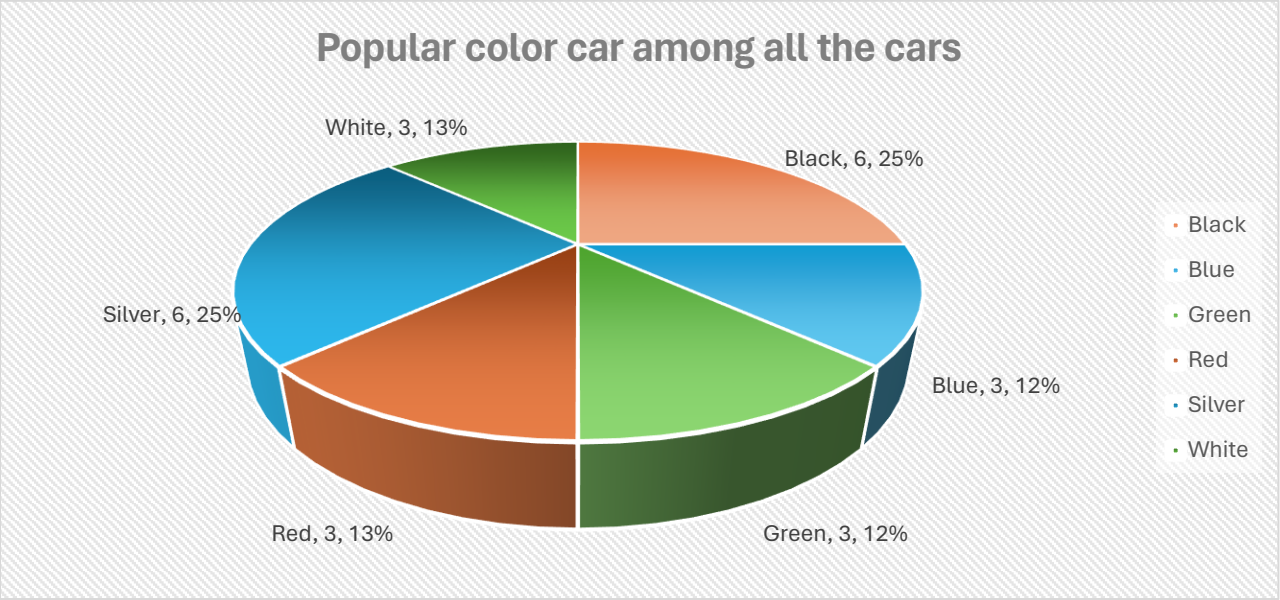
2. Justify, Buying of any Ford car is better than Honda.



By contrasting their relative qualities and putting a special emphasis on pricing, this analysis 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 (393.6) and average mileage (89623.3).

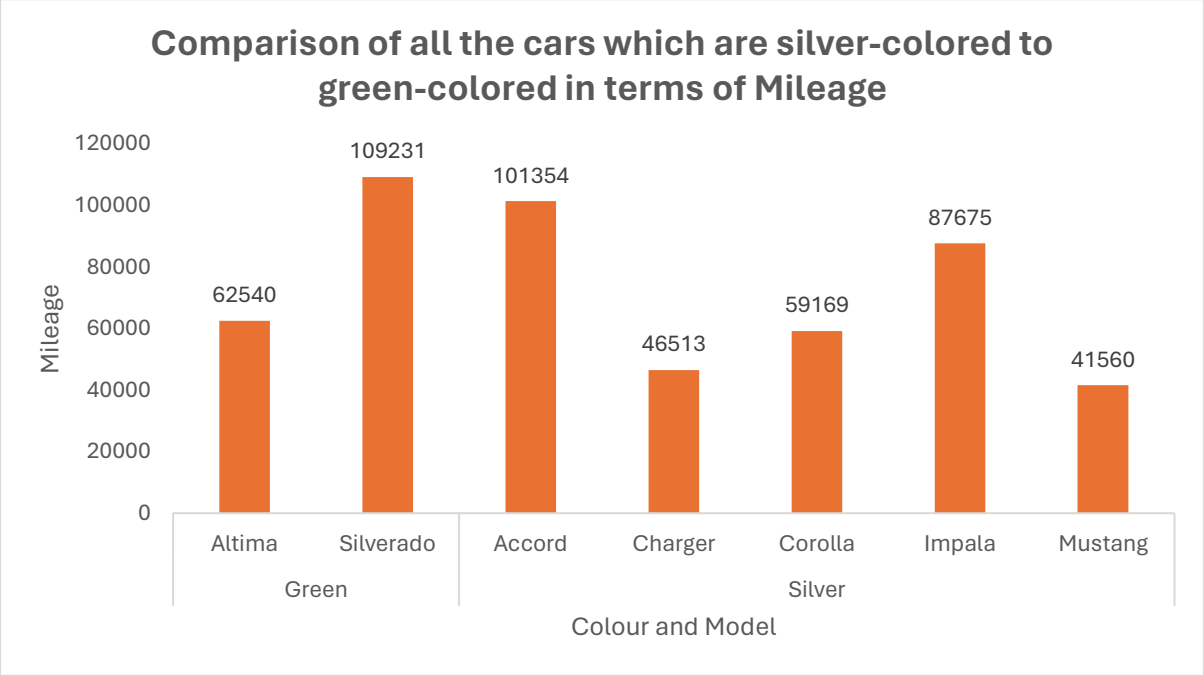
3. Among all the cars which car color is the most popular and is least popular?



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

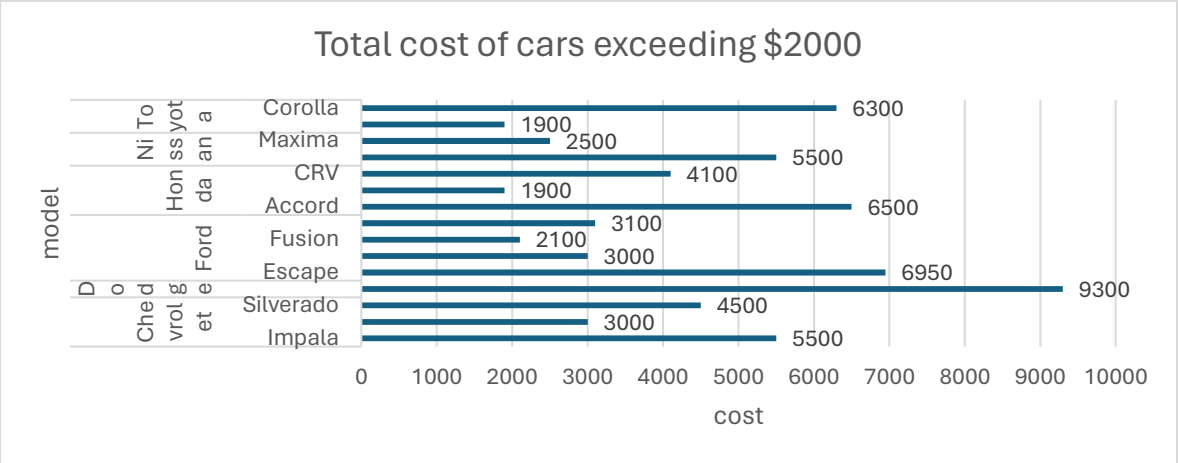
According to the data, the two most popular automobile colours are black and white, which account for 25% of the company's production, while green and blue cars account for 12% of the total.

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 utilising 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 study comparing the Toyota Corolla and Chevrolet Impala's mileage showed that the Impala has superior 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.

common Car Colours: Based on the analysis, the most common car colours 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 colours, making up a mere 12% of all cars produced.Comparing

Silver vs. Green Vehicles: Of the silver-colored vehicles, the Accord had the highest average mileage, while the Silverado had the highest mileage among green-colored cars.

Cars Costing more than $2000: The analysis determined that the total cost of cars exceeding $2000 amounted 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 investigation was comprehensive, and the results were presented well using the right visualizations—column charts and bar graphs, for example. 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

1. **Regression Statistics:**
    - Multiple R: This is the correlation coefficient, representing the strength and direction of the linear relationship between the predictor variables and the response variable. In this case, it's quite high at 0.962639, indicating a strong positive correlation.
    - R Square: Also known as the coefficient of determination, it represents the proportion of the variance in the dependent variable that is predictable from the independent variables. In your case, it's 0.926673, indicating that 92.67% of the variance in the response variable is predictable from the predictor variables.
    - Adjusted R Square: This is a modified version of R Square that adjusts for the number of predictor variables in the model. It penalizes the addition of irrelevant variables. Here, it's slightly lower than R Square at 0.91969.
    - Standard Error: This represents the average deviation of the observed values from the regression line. It's 259.2716 in your analysis.
    - Observations: The number of data points in your dataset is 24.
2. **ANOVA (Analysis of Variance):**
    - ANOVA table breaks down the variance in the dependent variable into components attributed to different sources. Here, it shows the sum of squares (SS), degrees of freedom (df), mean squares (MS), F statistic, and significance level (p-value) for both regression and residual (error) terms.
    - The regression is statistically significant with a very low p-value (1.22E-12), indicating that the overall regression model fits the data well.

3. **Coefficients:**
    - This table provides the coefficients for each predictor variable.
    - The "Intercept" is the value of the dependent variable when all independent variables are zero.
    - "X Variable 1" and "X Variable 2" are the coefficients for the respective predictor variables.
    - Each coefficient has a standard error, t-statistic, and p-value associated with it.
    - "Lower 95%" and "Upper 95%" represent the lower and upper bounds of the 95% confidence interval for each coefficient.

Anova: one factor

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				

1. **Summary:**
    - You have three columns/groups, each with 24 observations.
    - For each group, you've provided the sum, average (mean), and variance.
2. **ANOVA (Analysis of Variance):**
    - ANOVA decomposes the total variance in the data into variance between groups and variance within groups.
    - "Between Groups" represents the variation between the group means.
    - "Within Groups" represents the residual variation within each group after accounting for the group means.
    - "Total" is the sum of the variance between and within groups.
3. **ANOVA Table:**

- "Source of Variation" indicates whether the variation is between groups or within groups.
- "SS" stands for sum of squares, representing the squared deviations from the mean.
- "df" is the degrees of freedom associated with each source of variation.
- "MS" is the mean square, calculated as SS divided by its degrees of freedom.
- "F" is the F-statistic, which is the ratio of mean square between groups to mean square within groups.
- "P-value" indicates the significance level of the F-statistic, testing whether the group means are significantly different.
- "F crit" is the critical F-value, which is the threshold for determining statistical significance.

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				

1. **Summary:**
    - The data is presented in a table format with 24 rows and 3 columns.
    - Each row represents a category, with 3 observations per category.
    - For each row, you've provided the count, sum, average, and variance.
2. **ANOVA (Analysis of Variance):**
    - ANOVA decomposes the total variance into variance between rows, variance between columns, and residual (error) variance.
    - "Rows" represent the variation between different categories (rows).
    - "Columns" represent the variation between different columns.
    - "Error" represents the residual variance within the cells after accounting for row and column effects.
    - "Total" is the sum of all sources of variation.
3. **ANOVA Table:**
    - "Source of Variation" indicates whether the variation is between rows, columns, or due to error.
    - "SS" stands for sum of squares, representing the squared deviations from the mean.
    - "df" is the degrees of freedom associated with each source of variation.
    - "MS" is the mean square, calculated as SS divided by its degrees of freedom.
    - "F" is the F-statistic, which is the ratio of mean square for each source of variation.
    - "P-value" indicates the significance level of the F-statistic, testing whether the variation is significant.
    - "F crit" is the critical F-value, which is the threshold for determining statistical significance.

Based on the provided ANOVA results, it seems that there is no statistically significant difference between the rows (categories) since the p-value (0.549982) is higher than the typical significance level (e.g., 0.05). However, there is a statistically significant difference between the columns (categories), as the p-value (2.05E-19) is much lower than 0.05.

Descriptive Statistics

mileage		***price***		***Cost***	
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

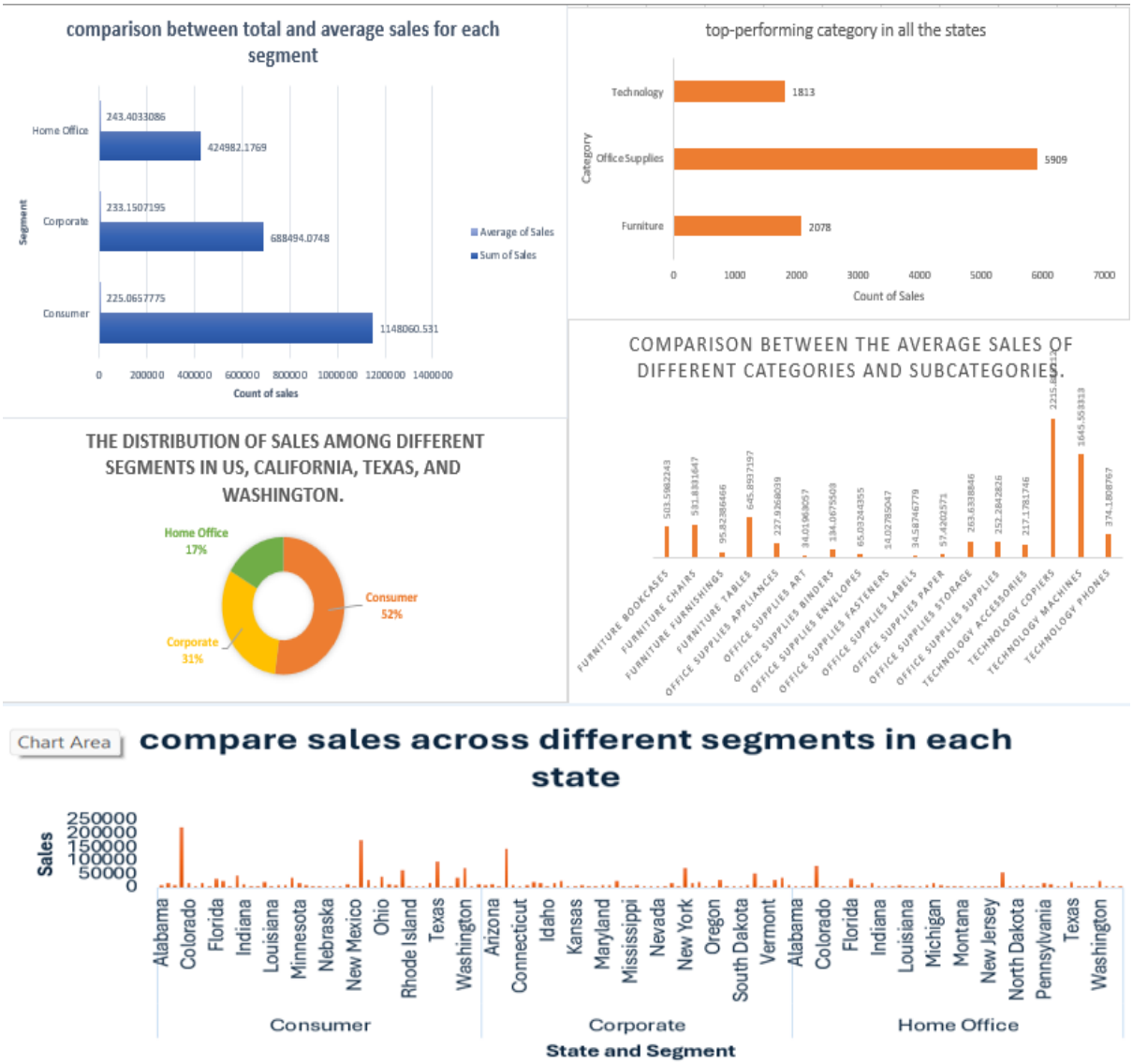
Correlation

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

	mileage	***price***
mileage	1	
price	-0.41106	1

Order Data Report

Dashboard



Introduction

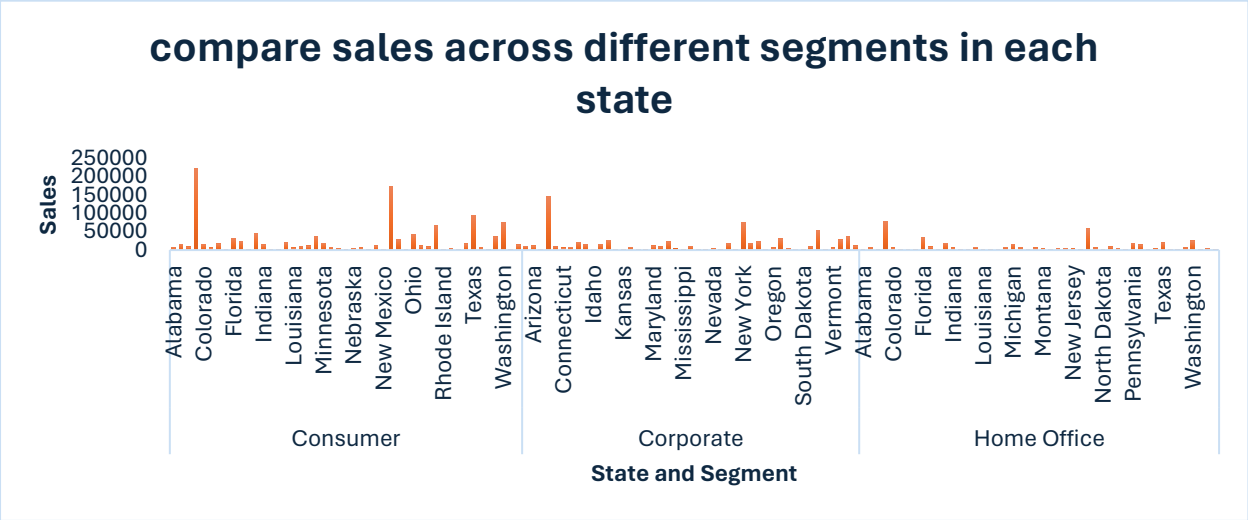
This study explores a vast dataset that records sales transactions in the automotive sector. It includes a variety of variables, including Order ID, Order Date, Ship Date, Customer Information, Product Details, and Sales Figures. Finding practical insights to guide decision-making and promote corporate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

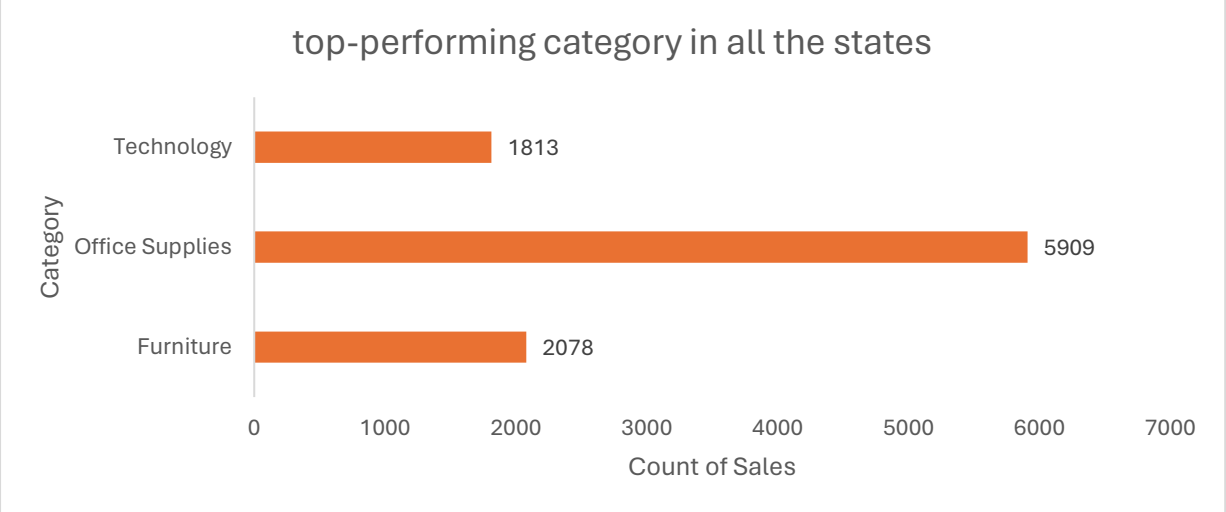
Analytics

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



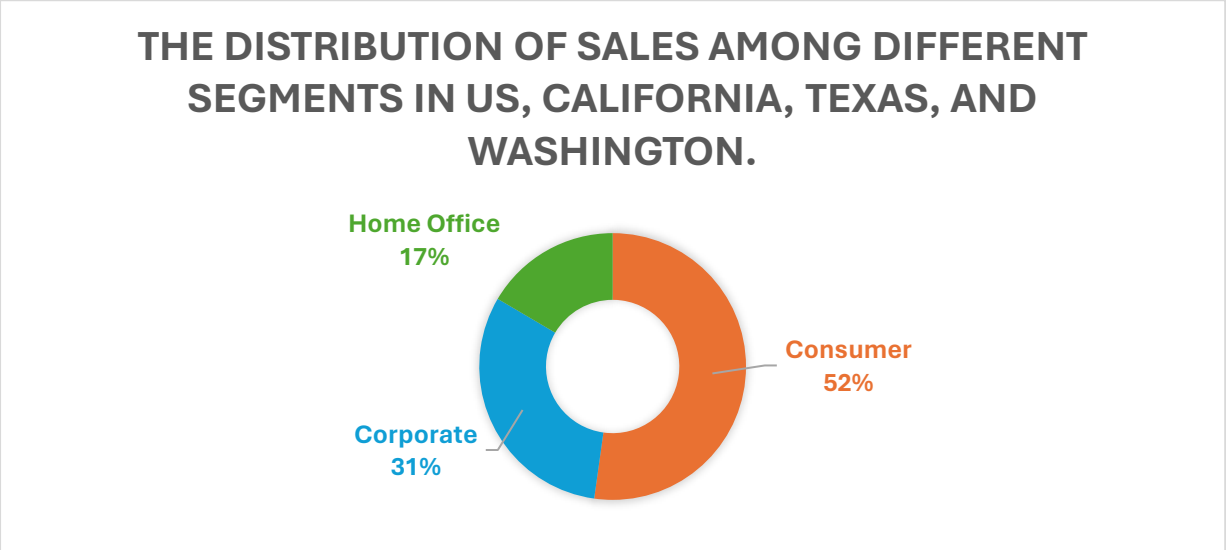
After comparing all the states in terms of segment and sales , California(222419.05) emerged as the state with the highest number of sales. Consumer(1148060.531) segment performed well in all the states.

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



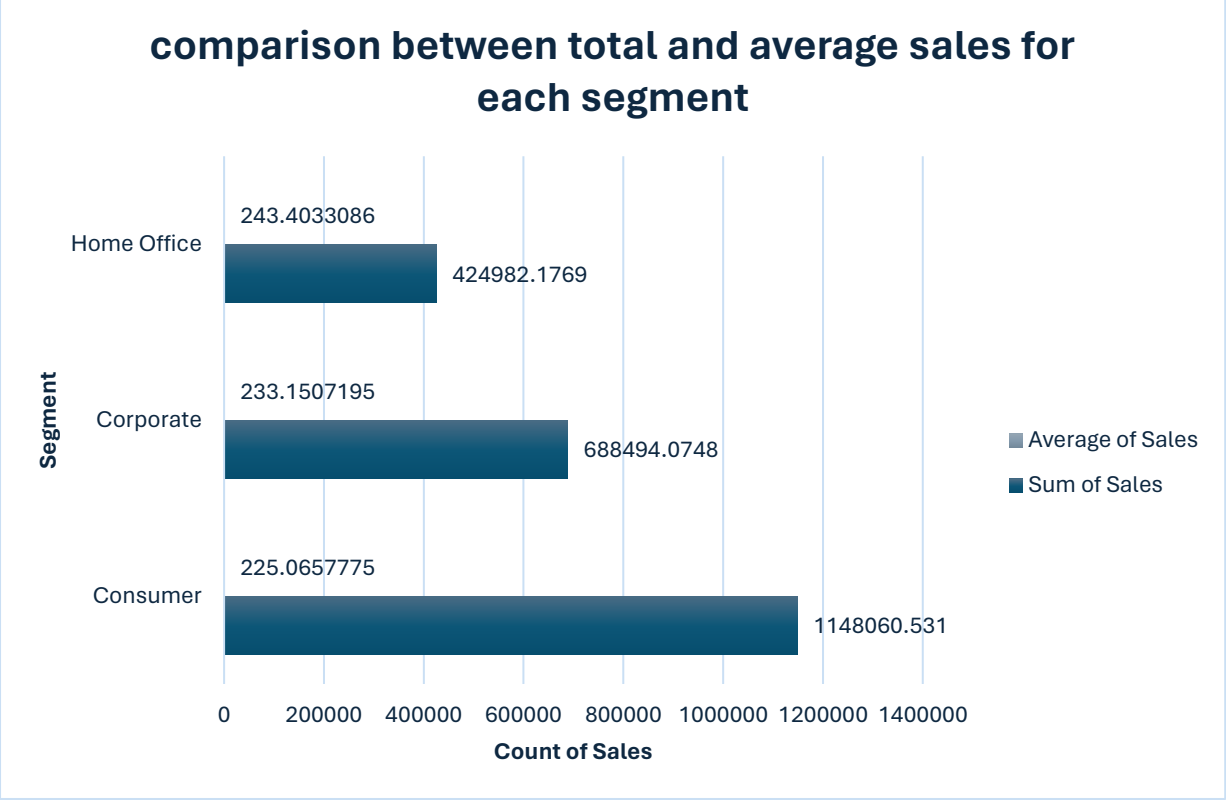
Office Supplies is the top performing category in all the states with total count of sales of 5909 followed by furniture(2078) and technology(1813).

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



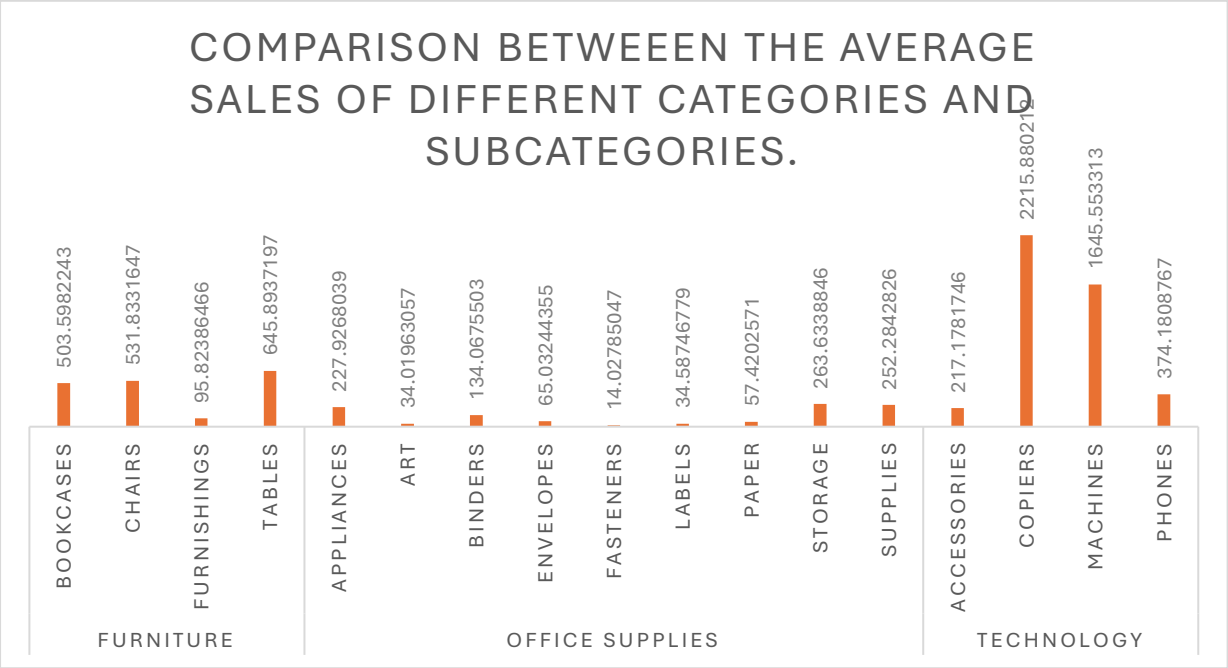
Filtering the states for the total sales count and showing the percentage of distribution through pie chart. The consumer segment has the most sales in US, California, Texas, and Washington.

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 analysis of data on sales in the automotive sector yields several 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 consistently led by the consumer market, especially in California, Texas, and Washington.

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

Regression

Regression of this data contains the dependent variable input as sales and independent variables as Id, the R Square value is 1.88E-07 and the Error value is 625.334, with the total df value as 9788 and total SS value of 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

The regression analysis aims to understand the relationship between an independent variable (X Variable 1) and a dependent variable. Here's a detailed explanation of the key points from the regression output:

Regression Statistics:

- Multiple R (0.000434): This value is the correlation coefficient, which indicates the strength and direction of the linear relationship between the independent and dependent variables. A value close to 0 suggests a very weak relationship.

- R Square (1.88E-07): This value represents the proportion of variance in the dependent variable that can be explained by the independent variable. An R Square value of 0.000000188 indicates that the model explains virtually none of the variance in the dependent variable.

- Adjusted R Square (-0.0001): This value adjusts the R Square for the number of predictors in the model. A negative value here implies that the model does worse than simply using the mean of the dependent variable as the predictor.

-Standard Error (625.334): This value measures the average distance that the observed values fall from the regression line. Higher values indicate a less precise model.

- Observations (9789): This is the number of data points used in the analysis.

ANOVA (Analysis of Variance):

Regression :

- df (degrees of freedom) (1) : This indicates the number of predictors in the model.
- SS (Sum of Squares) (721.1637) : This represents the variability explained by the model.
- MS (Mean Square) (721.1637) : This is the average variability explained by each predictor.
- F (0.001844) : The F-statistic assesses the overall significance of the model. A very low F value indicates that the model is not statistically significant.
- Significance F (0.965747) : This is the p-value for the F-test. A value much greater than 0.05 indicates that the model is not significant.

Residual :

- df (9787) : This is the number of observations minus the number of predictors and the intercept.
- SS (3.83E+09) : This represents the variability not explained by the model.
- MS (391042.6) : This is the average variability not explained by the model.

Total :

- df (9788) : This is the total number of observations minus one.
- SS (3.83E+09) : This is the total variability in the dependent variable.

Coefficients:

Intercept :

- Value (230.5863) : This is the expected value of the dependent variable when the independent variable is zero.
- Standard Error (12.63999) : This measures the variability of the intercept estimate.
- t Stat (18.24261) : This tests whether the intercept is significantly different from zero.
- P-value (3.83E-73) : This very low value indicates that the intercept is highly significant.
- 95% Confidence Interval (205.8093 to 255.3633) : This range suggests that we are 95% confident that the true intercept lies within this interval.

X Variable 1 :

- Value (-9.6E-05) : This is the estimated change in the dependent variable for a one-unit change in the independent variable.
- Standard Error (0.002235) : This measures the variability of the coefficient estimate.
- t Stat (-0.04294) : This tests whether the coefficient is significantly different from zero.

- P-value (0.965747) : This high value indicates that the coefficient is not significant.

- 95% Confidence Interval (-0.00448 to 0.004286) : This range includes zero, indicating that the true effect of the independent variable could be zero.

Interpretation:

- The very low Multiple R and R Square values suggest a negligible relationship between the independent variable and the dependent variable.

- The negative Adjusted R Square indicates that adding the independent variable does not improve the model.

- The high p-values for the F-test and the coefficient of the independent variable indicate that neither the model as a whole nor the independent variable is statistically significant.

- The intercept is significant, suggesting that while the baseline level of the dependent variable is well-estimated, the independent variable does not add explanatory power.

In summary, the analysis shows that the independent variable does not significantly predict the dependent variable, implying that other factors may be more important in explaining the variance in the dependent variable.

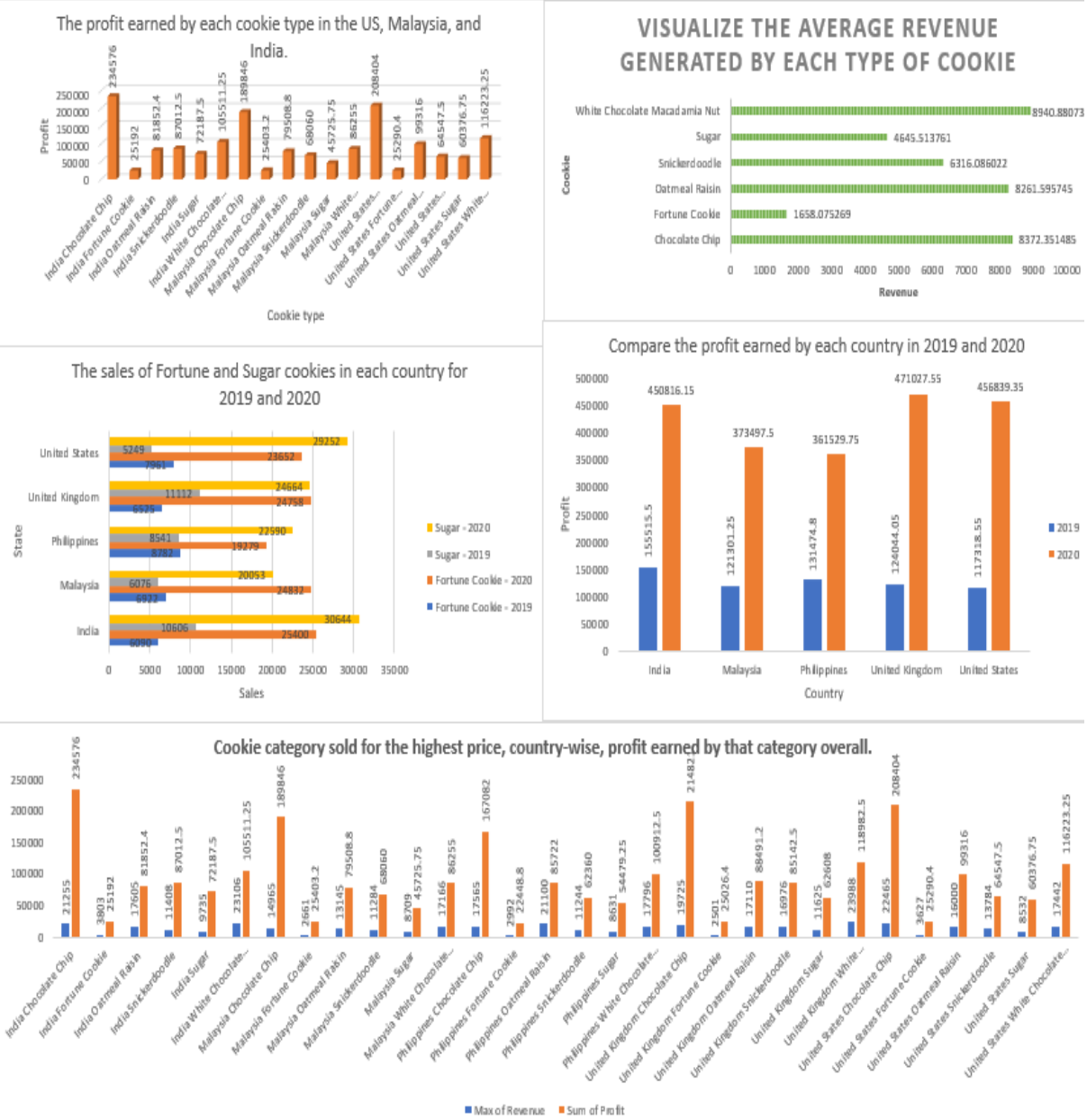
Descriptive Statistics

Sales	
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

For the sales column, descriptive statistics display the mean value of 230.1162, the median value of 54.384, the mode value of 12.96, the standard deviation of 625.3021, the variance of 391002.7, and numerous other variables like as kutosis, skewness, range, max, and min.

Cookie Data Report

Dashboard



Introduction

Six distinct types of cookies are included in our cookie data set: chocolate chip, fortune cookie, sugar, oatmeal raisin, Snickerdoodle and white chocolate macadamia nut. We possess an abundance of information regarding 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 various 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. Prepare to learn some amazing facts about the cookie industry and how it affects companies like yours.

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?

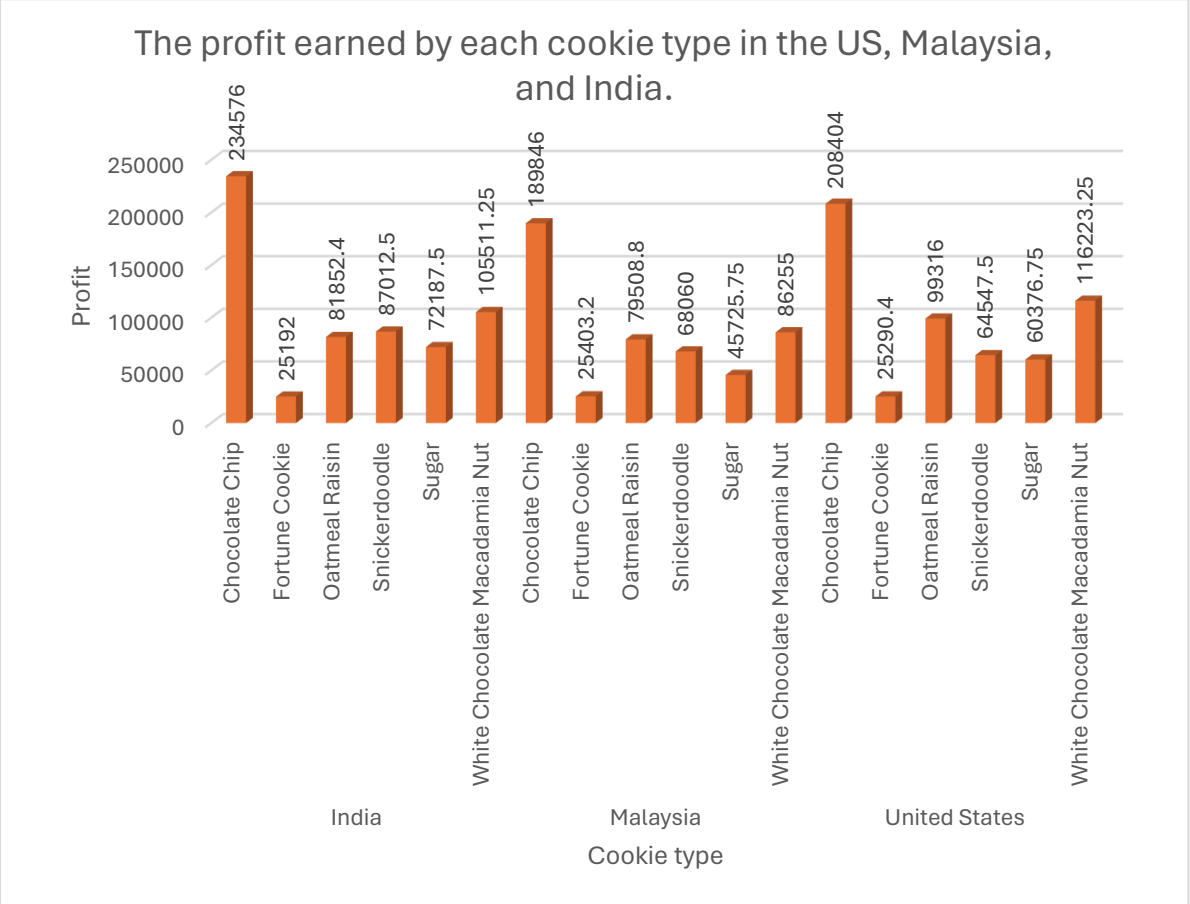
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?

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?

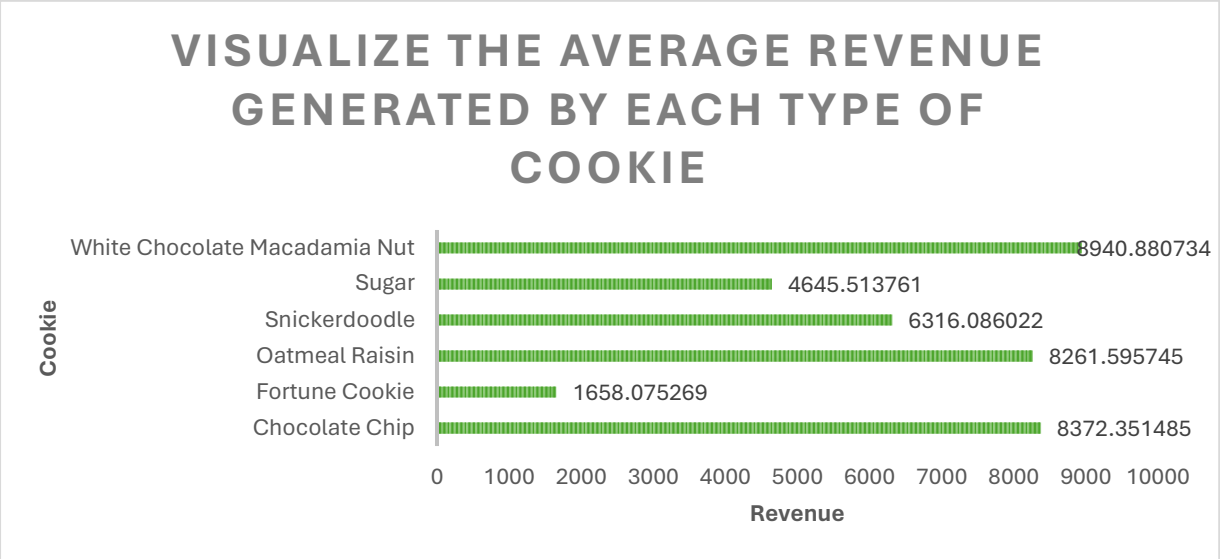
Analytics

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



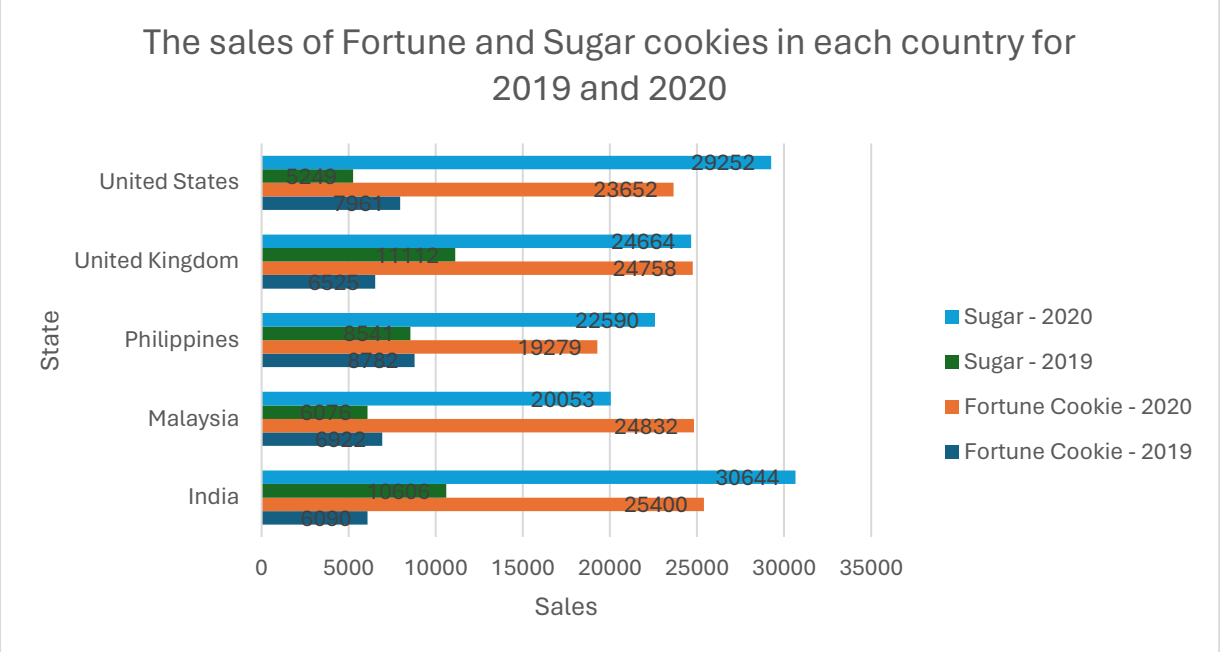
The profit margins for each type of cookie in the US, Malaysia, and India are compared in this analysis. India's maximum profit on chocolate chips is followed by that of Malaysia and the United States.

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



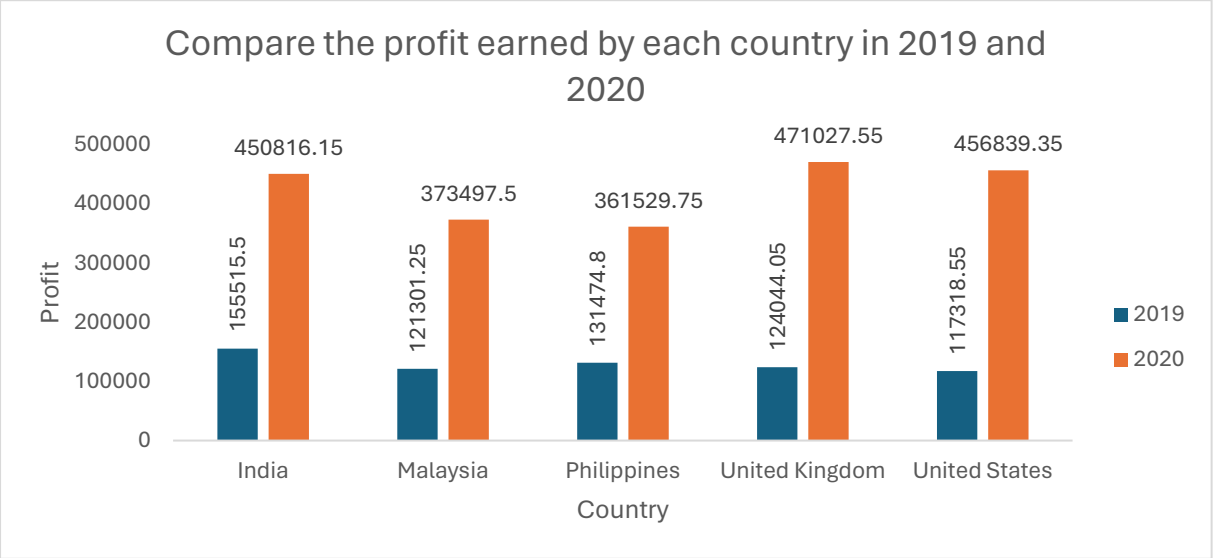
The goal of this analysis is to present the average revenue generated. It is evident that the product with the highest average revenue generate, white chocolate macadamia nut, is chocolate chip, at 8940.88.

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



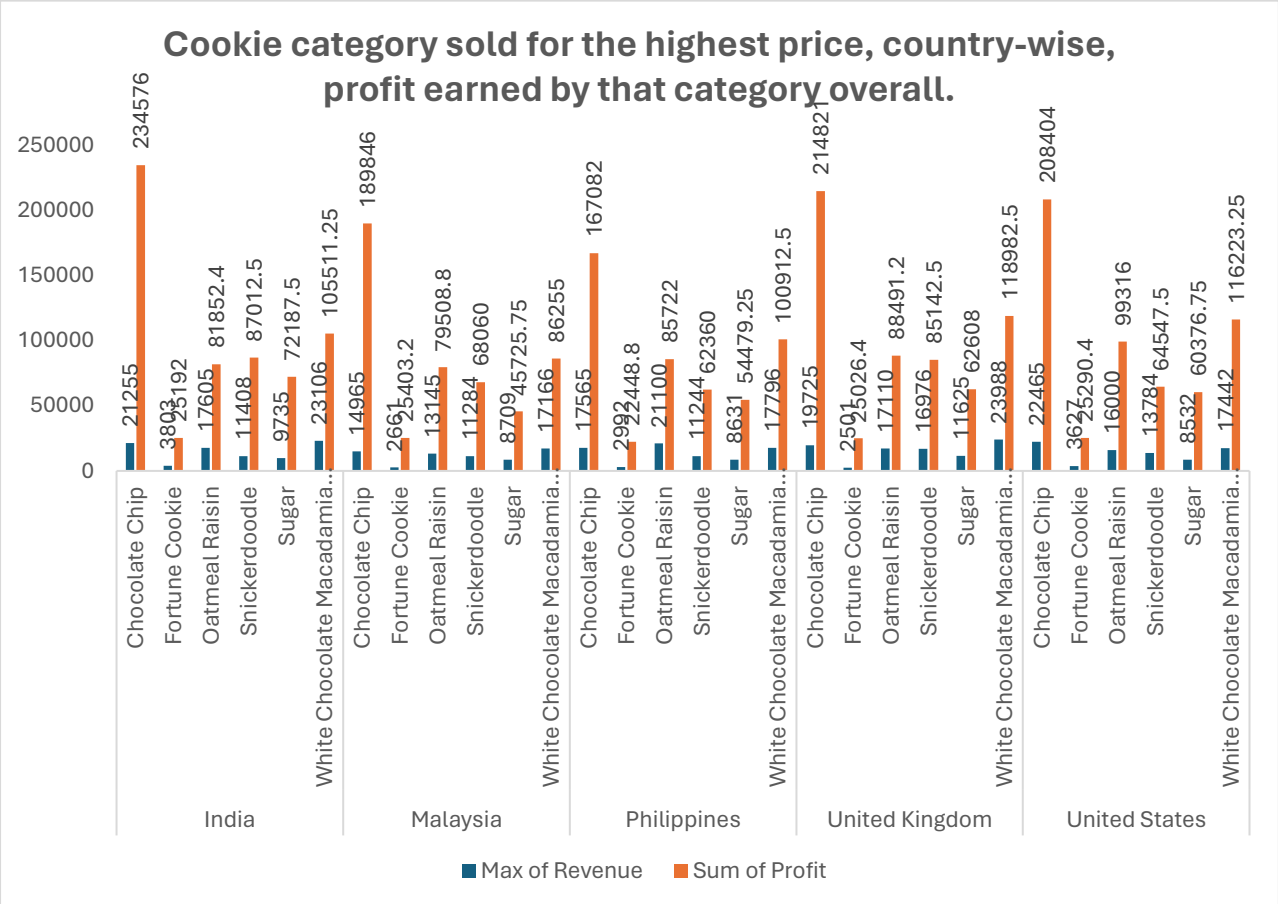
This analysis compares the sales of fortune and sugar cookies in the various countries for the years 2019 and 2020. India leads the way in significant sales of sugar cookies for the year 2020, with 30644 sales; the United Kingdom led the way in sales of sugar cookies in 2019. India again leads in sales of fortune cookies, with 25400, followed by Malaysia; the Philippines leads 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?



The objective of this investigation is to determine which cookie category sold for the most money, per country, and the profit earned by that category. The maximum revenue for chocolate chips (23988) and the total profit for sugar (2763364.45) are recorded for India, followed by the 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 highest 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 highest profits in 2020. In 2019, India had the highest profit, followed by the Philippines.

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

The report helped stakeholders 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 communicate the findings. It's crucial to recognise the need for more research into other variables affecting sales and profitability, though. Getting accurate and comprehensive data is essential to getting trustworthy insights.

Regression

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

Regression shows R square value of 1 and Error of 9.16E-12 with 700 observations having the Input dependent variable as Profit and the input dependent variable as Units Sold, Price, Revenue.

Anova: one factor

Anova single factor shows the variance in the characters of the data effecting the independent variables profit.

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

Anova: two factor

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

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

Descriptive Statistics

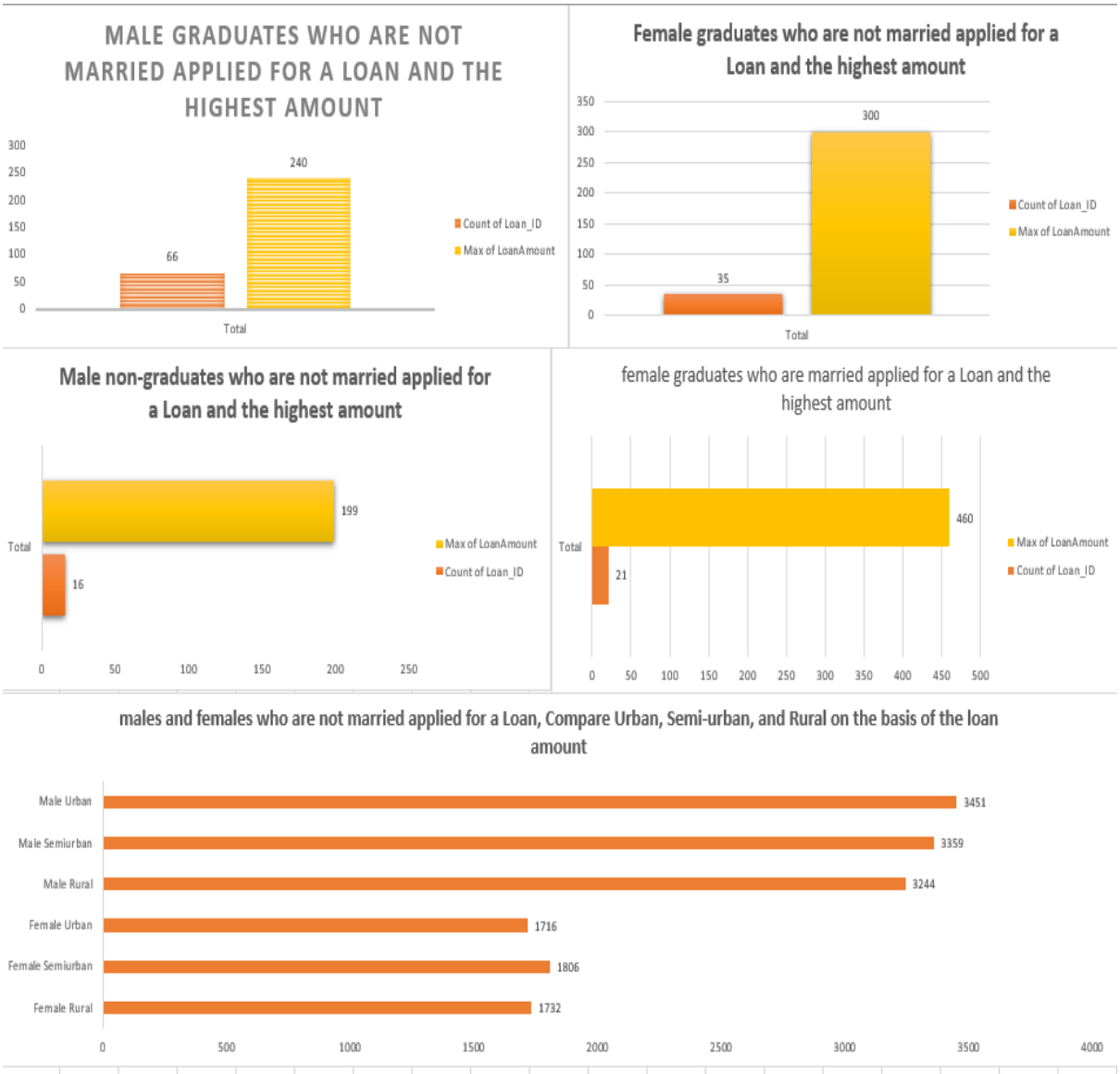
Column1		***Column2***		***Column3***		***Column4***	
Mean	1608.32	Mean	6700.456	Mean	2752.792	Mean	3947.664
Standard Error	32.78652	Standard Error	174.767	Standard Error	76.99166	Standard Error	98.86874
Median	1542.5	Median	5871.5	Median	2423.6	Median	3424.5
Mode	727	Mode	8715	Mode	3450	Mode	5229
Standard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

Correlation

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

Loan Data Report

Dashboard



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 about 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 particular 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 choices, streamline the lending process, and customise services to satisfy the wide range of client needs. Our goal in doing this study is to find practical insights that might inform strategic choices and improve the effectiveness of loan management systems.

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?

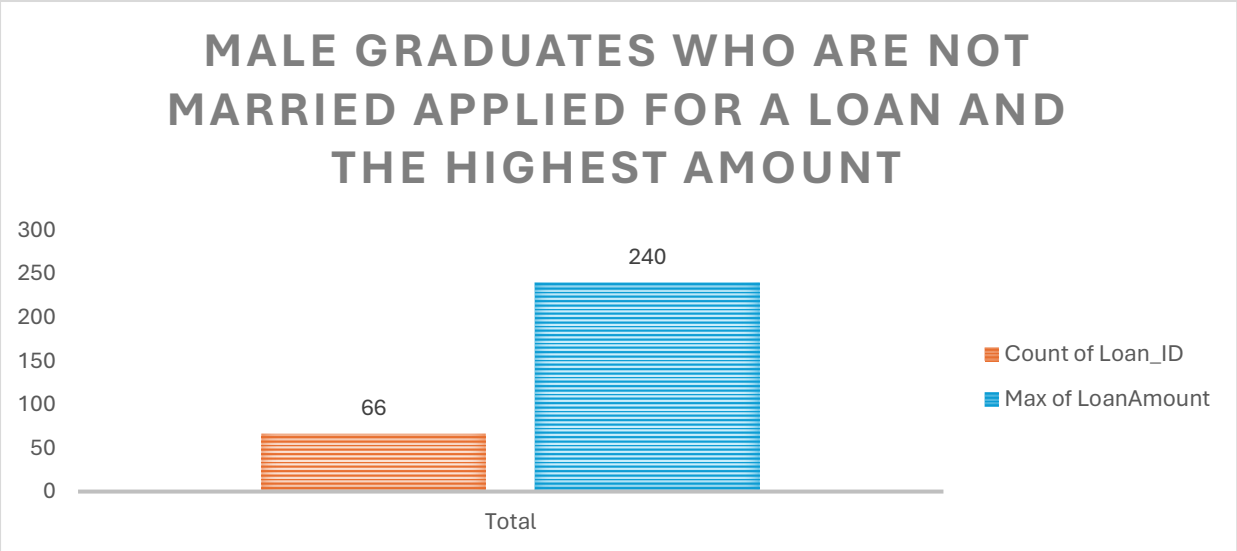
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?

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.

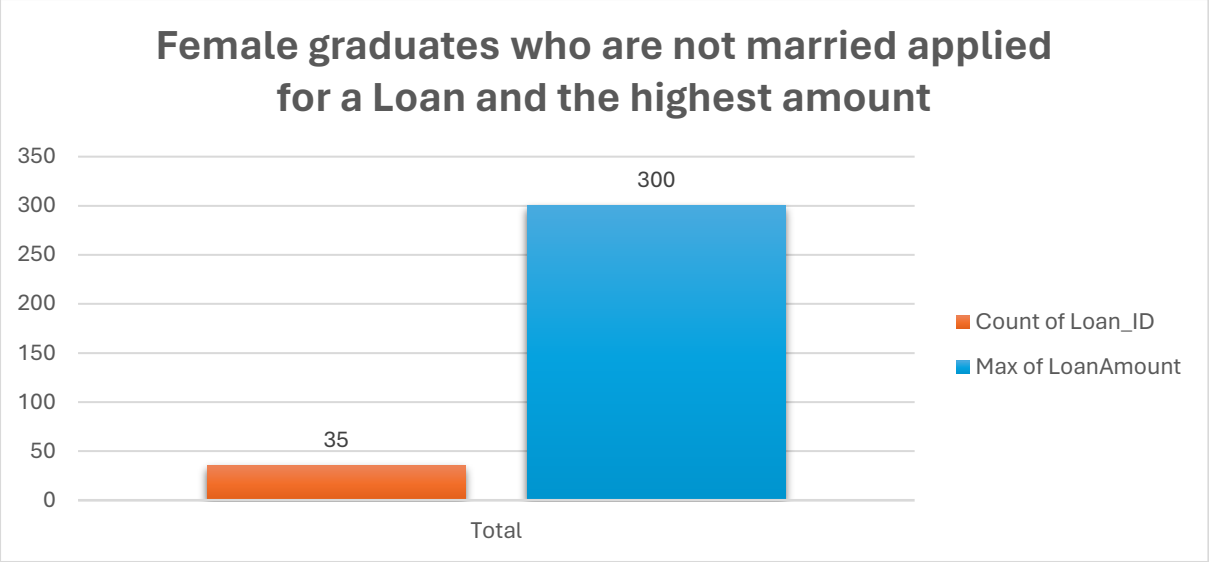
Analytics

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



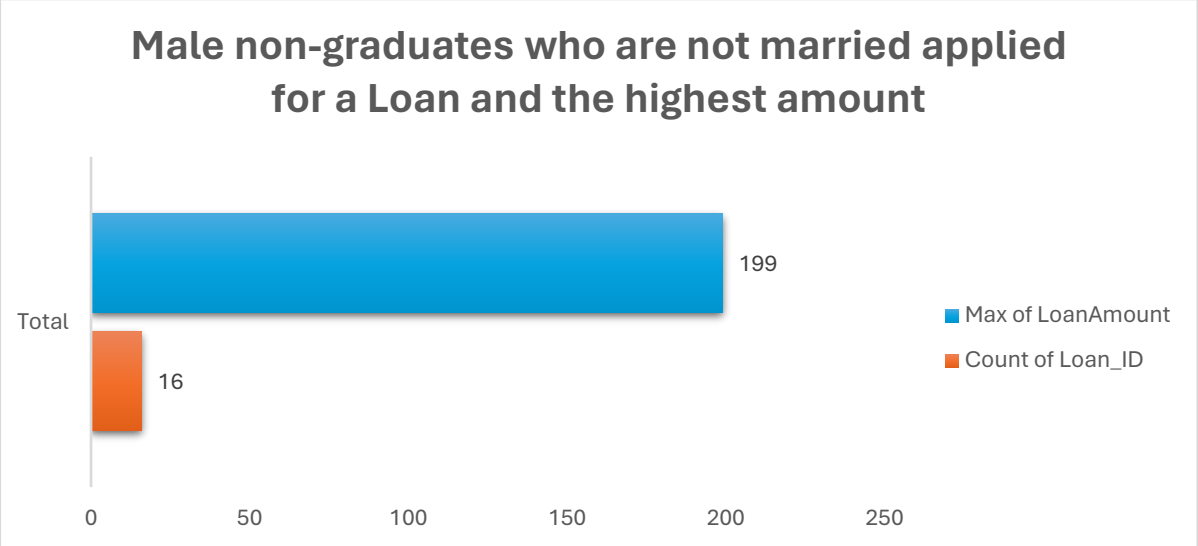
This analysis shows the no. of male graduates applied for the loan and are not married with the highest amount. As of analysed the total no. of loan applied is 66 and max loan amount is 240.

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



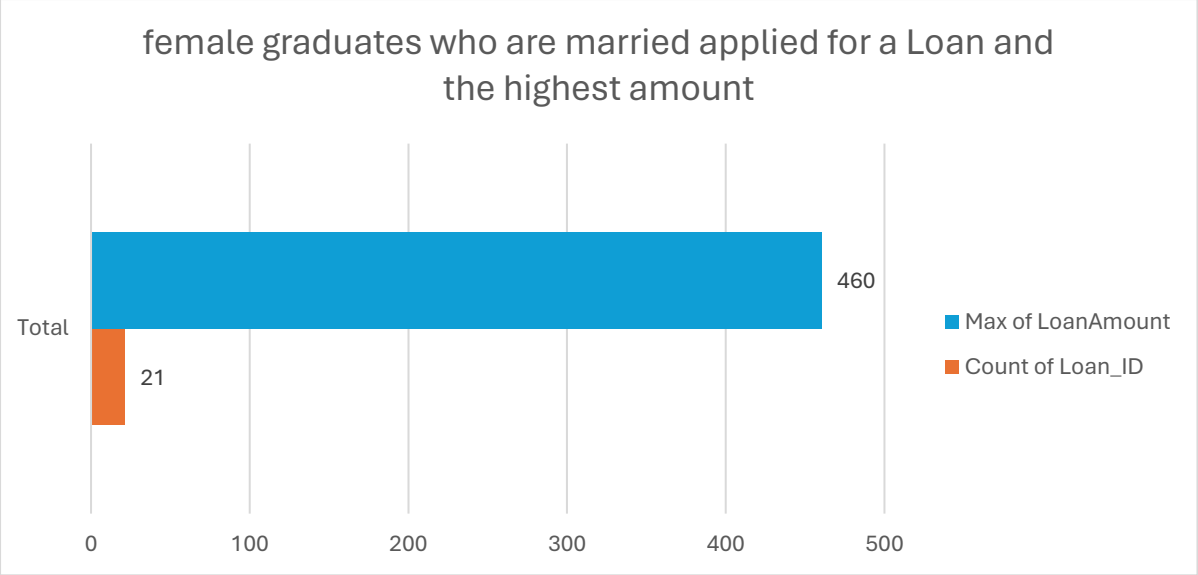
This analysis shows the no. of female graduates applied for the loan and are not married with the highest amount. As of analysed the total no. of loan applied is 35 and max loan amount is 300.

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



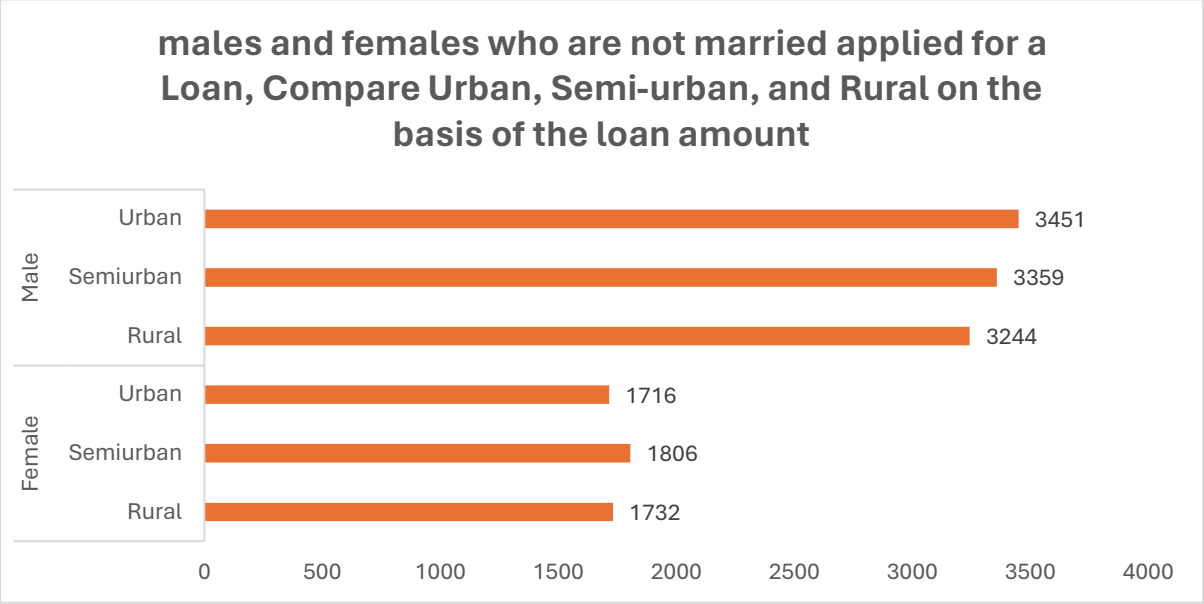
This analysis shows the no. of male non-graduates applied for the loan and are not married with the highest amount. As of analysed the total no. of loan applied is 16 and max loan amount is 199.

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



This analysis shows the no. of female graduates applied for the loan and are not married with the highest amount. As of analysed the total no. of loan applied is 21 and max loan amount is 460.

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



This analysis aims to compare the rural, semi urban, urban female and male who are not married and applied for the loan, where the no. is less in females but much higher in males

Females loan count in rural(1732),semiurban(1806), urban(1716) and males loan count in rural(3244),semiurban(3359),urban(3451).

Conclusion and Review

The analysis indicates clear gender disparities in loan applications. Male graduates not married dominated the applicant pool, followed by female graduates not married. Both male non-

graduates not married and married female graduates also applied for loans, albeit in smaller numbers. Notably, males significantly outnumbered females across rural, semi-urban, and urban areas.

The analysis effectively illustrates gender-based trends in loan applications and provides valuable insights into borrower demographics. Further exploration into factors influencing loan decisions is recommended, along with visual enhancements to improve data presentation. Overall, the report lays a foundation for understanding loan dynamics, with potential for deeper insights.

Regression

Regression shows the stats

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.531078663							
R Square	0.282044546							
Adjusted R Square	0.274487121							
Standard Error	50.85033905							
Observations	289							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	289502.8035	96500.93	37.32019	2.25609E-20			
Residual	285	736940.7397	2585.757					
Total	288	1026443.543						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	66.690952	16.26833015	4.099434	5.41E-05	34.66963005	98.71227396	34.66963	98.71227
X Variable 1	0.095771273	0.045649816	2.097955	0.03679	0.005917708	0.185624838	0.005918	0.185625
X Variable 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: one factor

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

Anova: two factor

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

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

Descriptive Statistics

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

Correlation

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

Shop Sales Data Report

Dashboard

Introduction

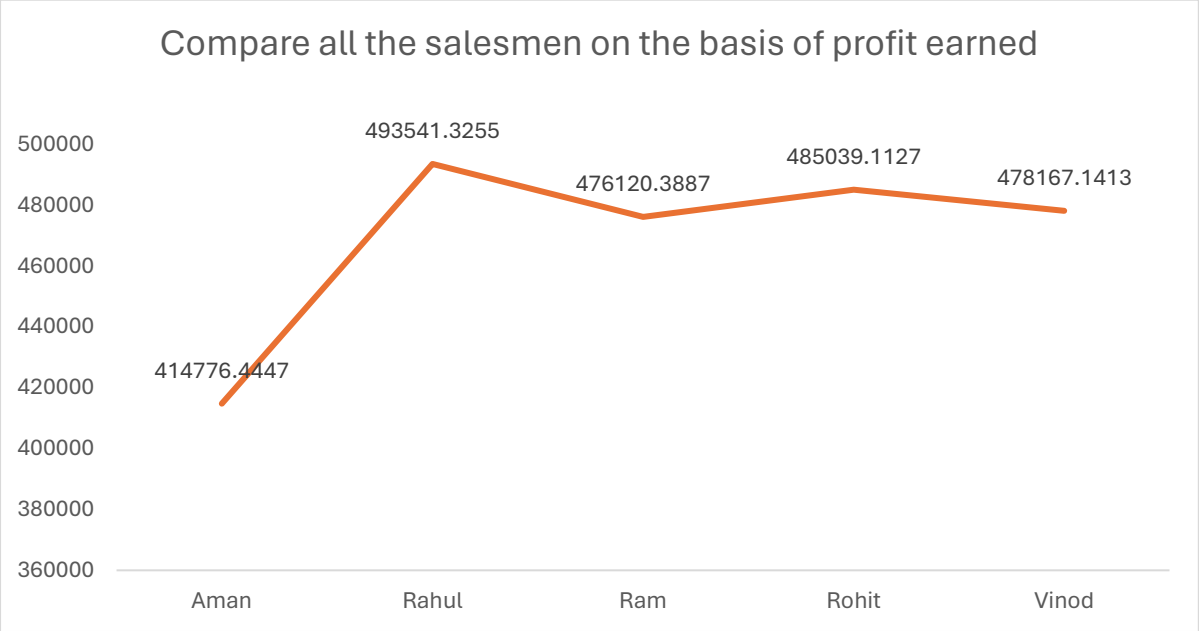
This paper examines a large sales dataset with an emphasis on sales performance analysis and product trends among sales representatives. The collection includes attributes including product specifications, sales volumes, earnings, and salesman details. Finding information that can improve corporate performance and guide the creation of sales strategies is the main goal of this investigation. The report's objectives are to identify top-performing salespeople, assess product popularity, and comprehend sales patterns by looking at sales data over a given 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 conducting this study is to offer practical insights that will help inform decisions and advance the success of the organisation as a whole.

Questionaries

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.

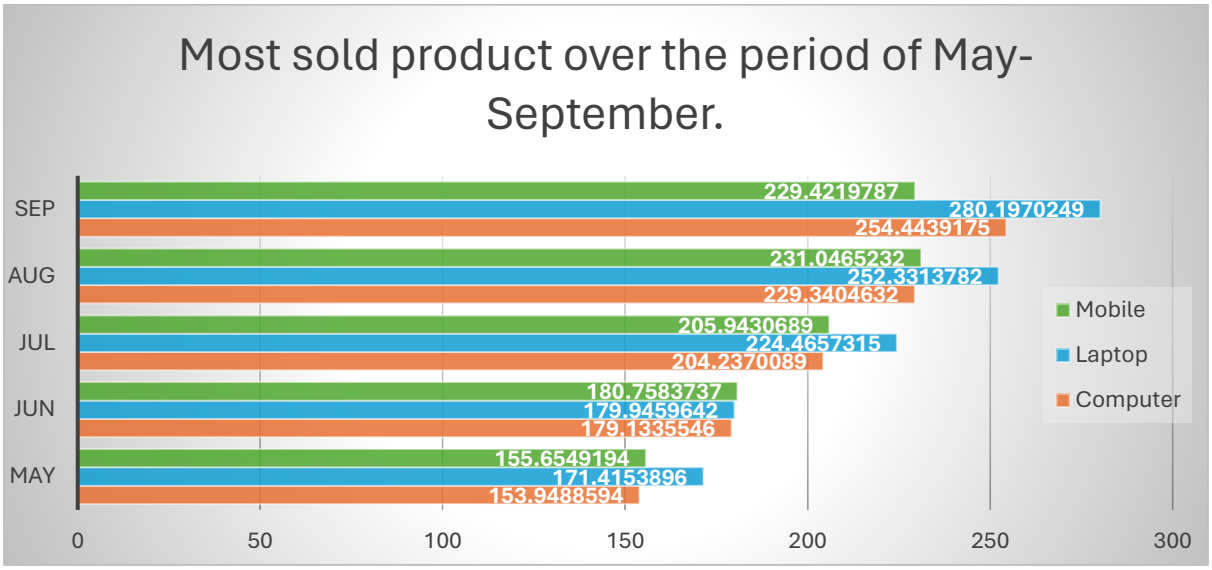
Analytics

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



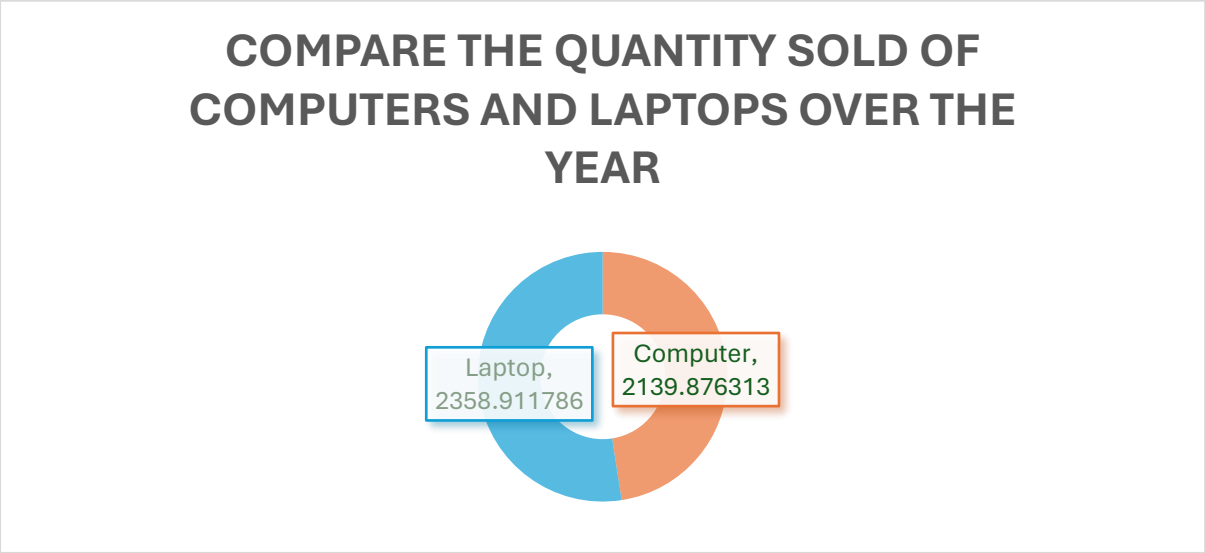
The comparison of all the salesmen on the basis of profit earned and the line chart shows that the rahul has the highest profit earned with value 493541.3255, compared to all the salesmen.

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



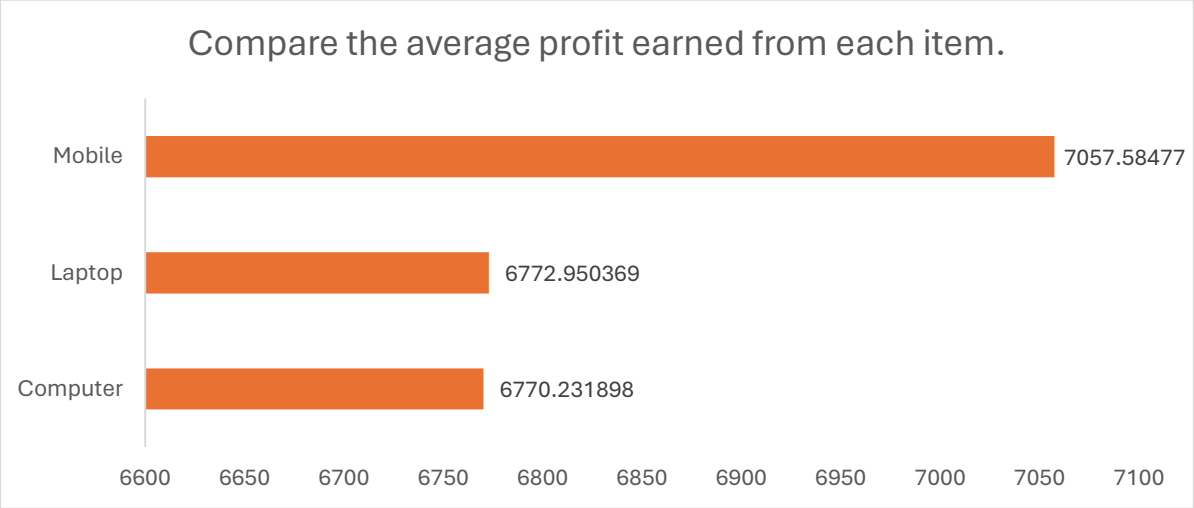
To identify the most sold product over the period of May-September, we would need to analyze the sales data within the timeframe. By aggregating the quantity sold for each product across all transactions during this period, and the most sold product over the period of May-September is Laptop with most sales in the September month with the value of 280.1970249.

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



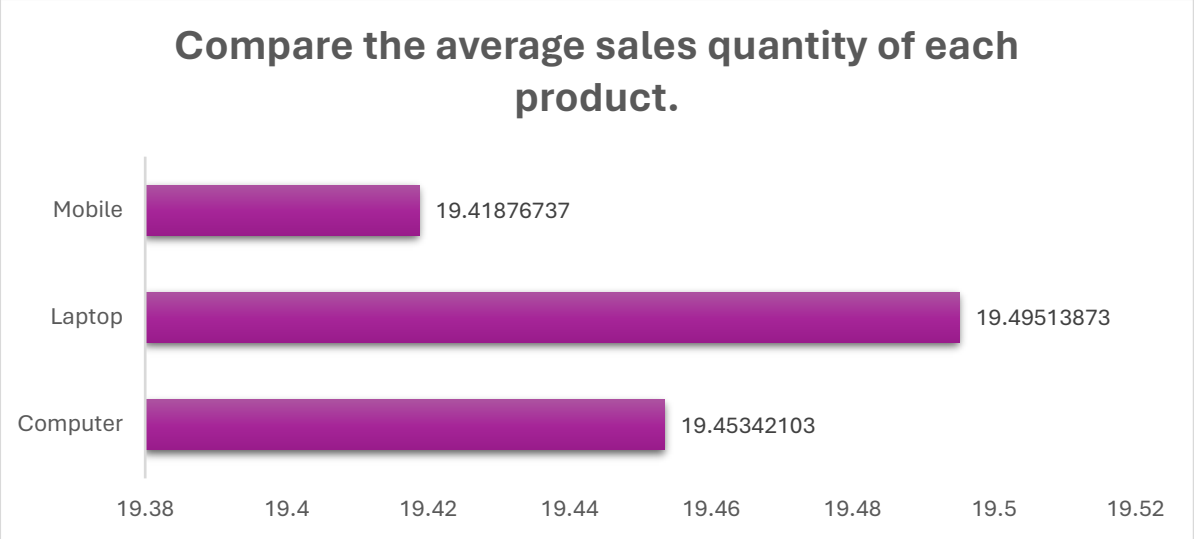
The two product sold the most over the year between computer or laptop where Computer has the sold quantity of 2139.876313 and laptop has 2358.911786 units sold quantity.

4 . Which item yield most average profit?



This analysis shows that the Mobile has the most Average profit earned among Mobile, Laptop, and Computer where Mobile has the average profit earned of 7057.58477.

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



The analysis shows that the average sales quantity of Laptop(19.49513873) is higher than the other products e.g. Mobile(19.41876737) and Computer(19.45342103).

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 perform better than computers. 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. Visualisations help in comprehending popular products and long-term patterns. Deeper understanding of the variables affecting product preferences and sales variations, however, might improve the analysis. All things considered, the research provides useful information for enhancing sales tactics and increasing profits.

Regression

The regression model, with a significant p-value indicates a strong positive relationship between Amount and the profit earned and the outcome variable. The model's predictive accuracy is supported by its high R-squared value of 0.9540.

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%*	*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

Correlation

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

	Unit sold	***revenue***
Column 1	1	
Column 2	0.954077	1

Anova (Single Factor)

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

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

Anova two factor

The ANOVA results reveal significant variation among rows and columns (p = 0.445792), with degrees of freedom (df) values of 10 respectively. 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				

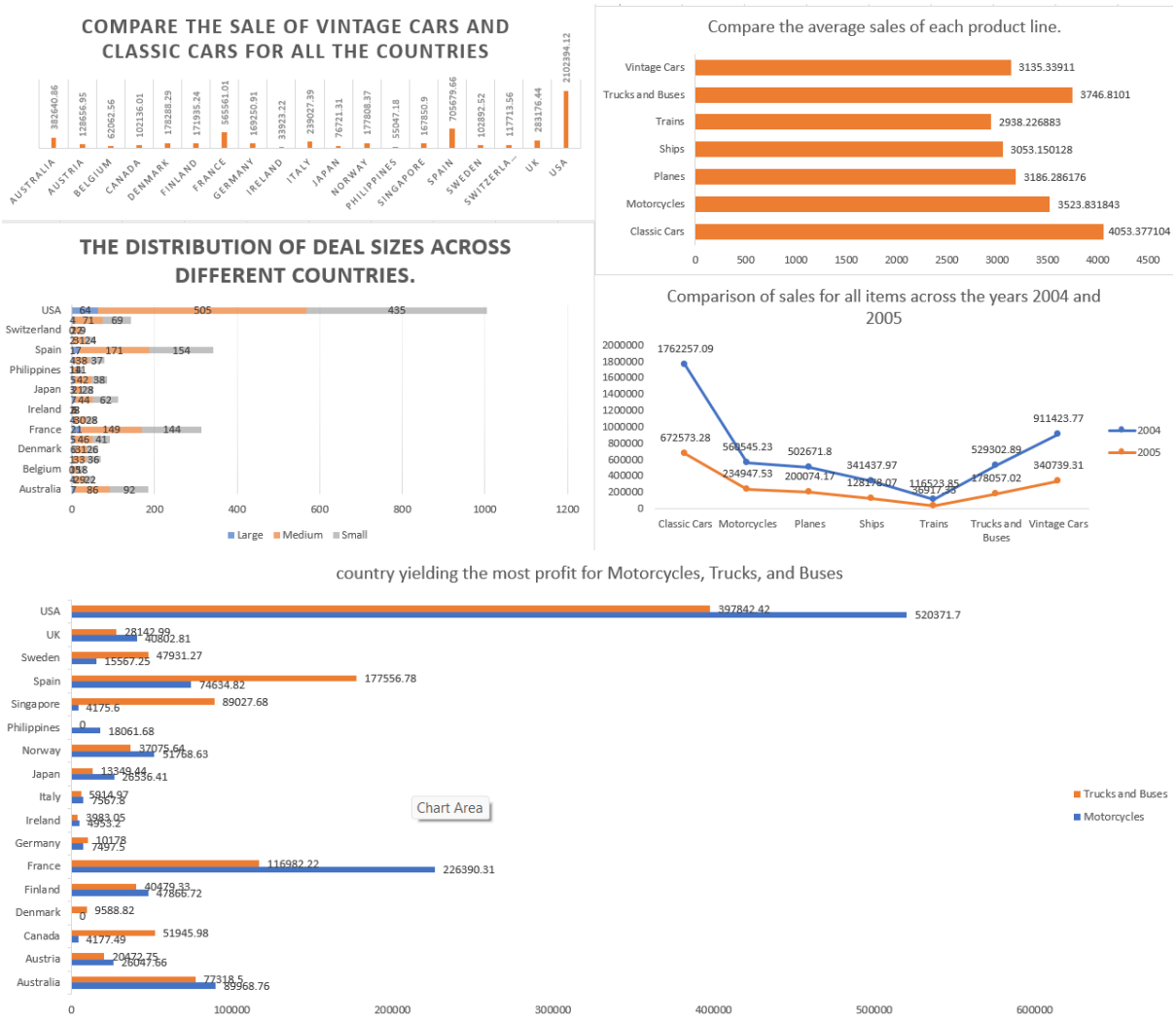
Descriptive Statistics:

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

Maximum	33.30852	Maximum	10279.85
Sum	6654.271	Sum	2347644
Count	342	Count	342

Sales Data Sample Report

Dashboard



Introduction

A large sales dataset with attributes like ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and SALES is analysed in this report. It seeks to draw conclusions that will direct sales tactics and improve corporate performance. Sales managers, marketers, and executives looking to maximise income and optimise sales operations are among the intended audience members. Important analyses include comparing the sales of classic and vintage cars, figuring out average sales, figuring out what items are best-selling, analysing the profit margin by country for particular product lines, comparing sales over time, and analysing countries according to the amount of deals. The report's goal is to offer practical insights that may be used to boost sales growth and enhance overall business outcomes through these assessments.

The project's scope includes assessing 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 looking for insights into 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.

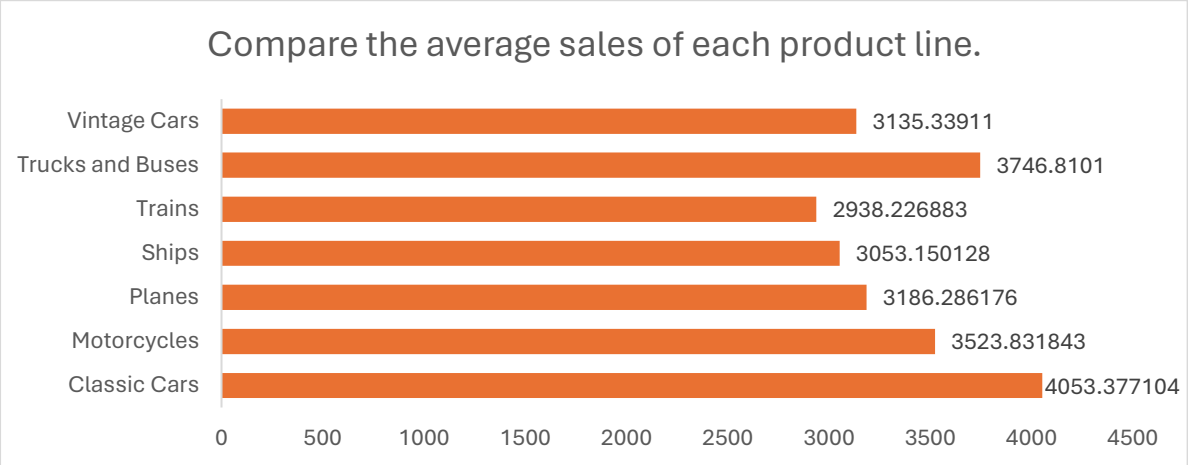
5. Comparative analysis of all countries based on deal size.

Analytics

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

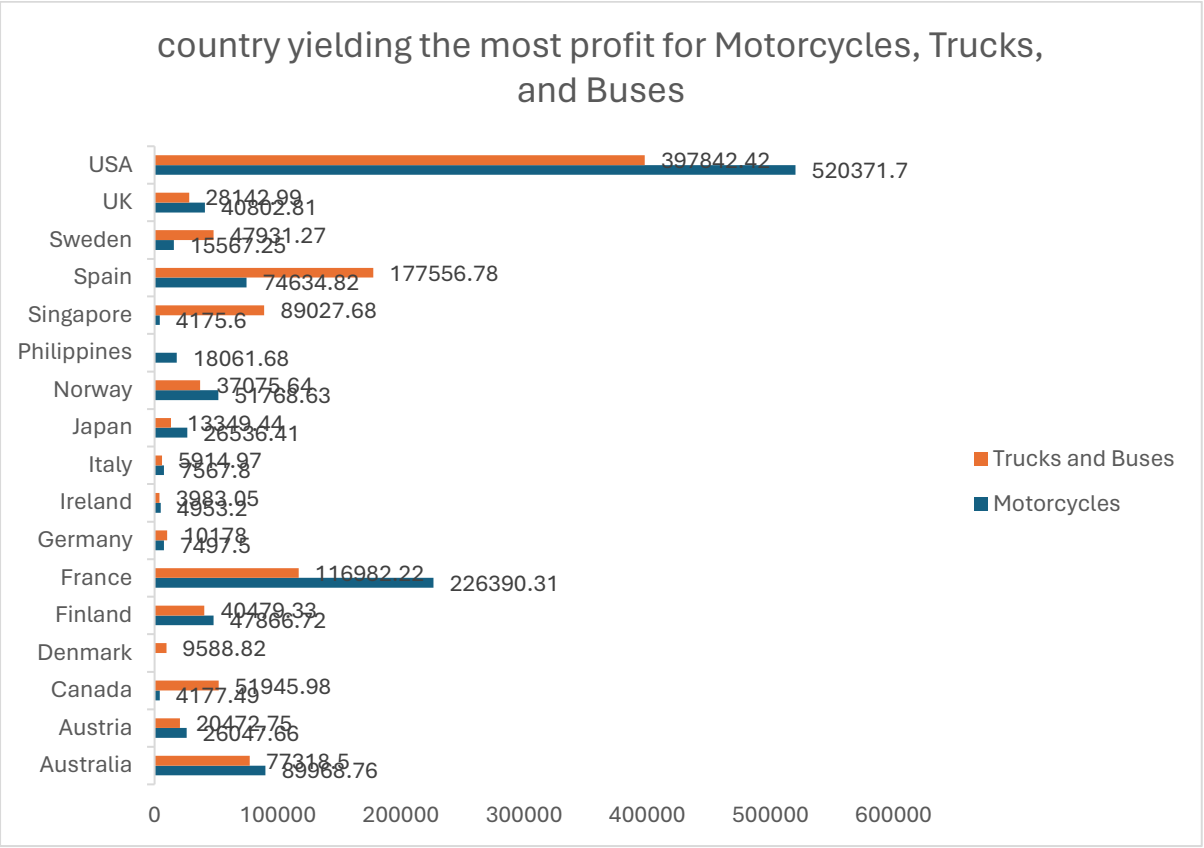
This analysis Compare the sale of Vintage cars and Classic cars for all the countries. Where USA(2102394.02) has the highest sales followed by Spain, France, and Australia.

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



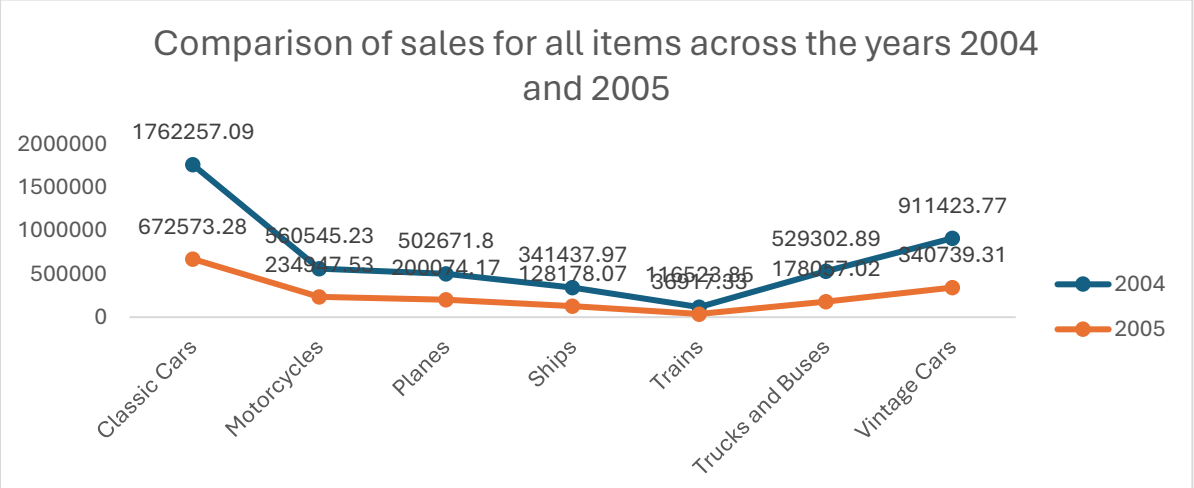
This analysis aims to provide average sales of all products and identification of the highest-selling product. And through the graph we can see that Classic Cars have the highest sales with 4053.377104 average sales followed by Trucks and Buses and Motorcycles.

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



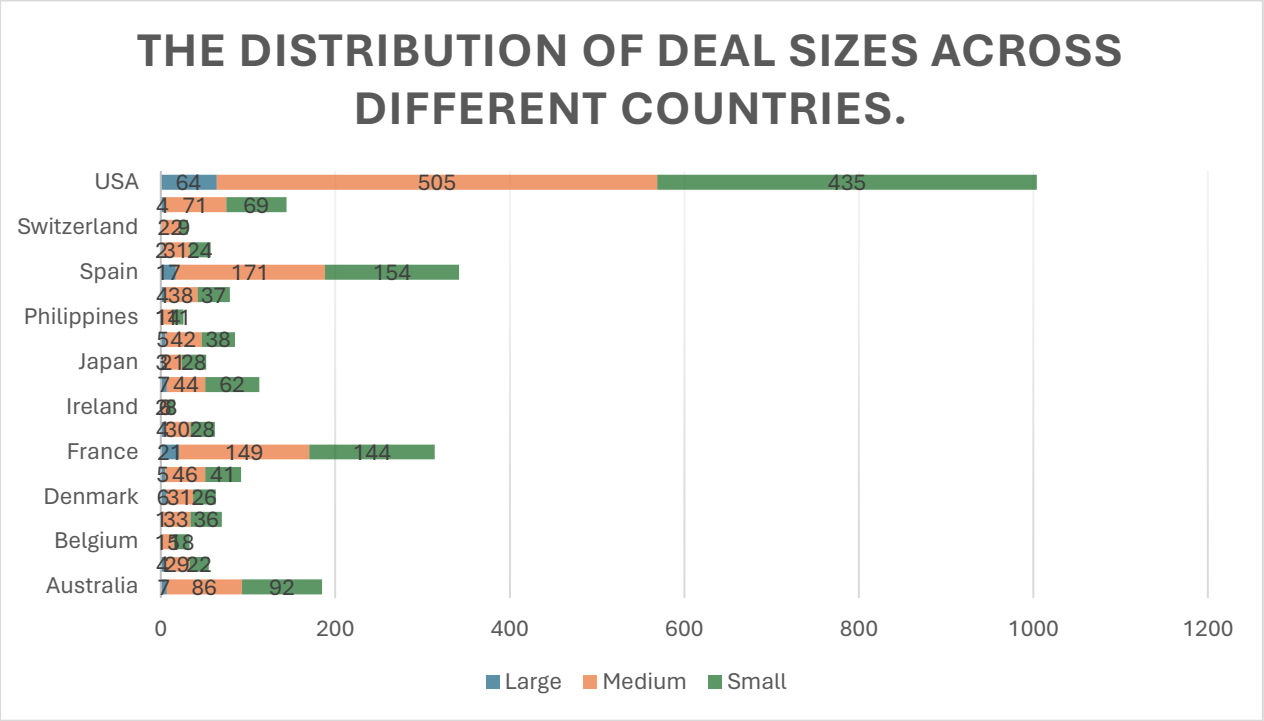
This analysis aims to identify the country yielding the most profit for Motorcycles, Trucks, and Buses. And bar chart shows that the USA has the highest sales with 397842.42 sum of sales for Trucks and Buses while 520371.7 sum of sales for Motorcycles followed by France and Spain.

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



This analysis aims to compare the sales for all the items across the years 2004 and 2005, with the line chart we can see that the sales for all the items across the years are shifting at very rate where the sales for Classic cars are highest among all the categories in both the years with 1762257.09 sales in 2004 and 672573.28 sales in 2005.

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



This analysis aims to find out the distribution of deal sizes across the different countries. And the bar chart shows that the deal size in the USA with large deal size of 64, medium deal size of 505, and small deal size of 435 is way higher than all the other countries.

Conclusion and Review

The analysis uncovers significant insights into sales dynamics and profitability across categories and countries. Notably, the USA emerges as a key market leader, exhibiting strong sales performance in Vintage and Classic cars, Trucks, Buses, and Motorcycles. Classic Cars stand out as the highest-selling product, contributing significantly to overall sales revenue. Moreover, the USA demonstrates exceptional profitability, particularly in the Trucks, Buses, and Motorcycles categories. Sales for Classic cars remain consistently robust throughout the years 2004 and 2005, indicating sustained demand for this category. Additionally, the USA showcases markedly larger deal sizes compared to other countries, underscoring its dominance in sales volume.

While the analysis effectively presents key findings through visualizations, further exploration into factors influencing sales fluctuations and deal size disparities could provide deeper insights. Overall, the report offers valuable insights for optimizing sales strategies and driving business growth.

Regression

Regression shows...

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

Anova: one factor

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

Anova: two factor

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

Descriptive Statistics

Column1		***Column2***		***Column3***		***Column4***	
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

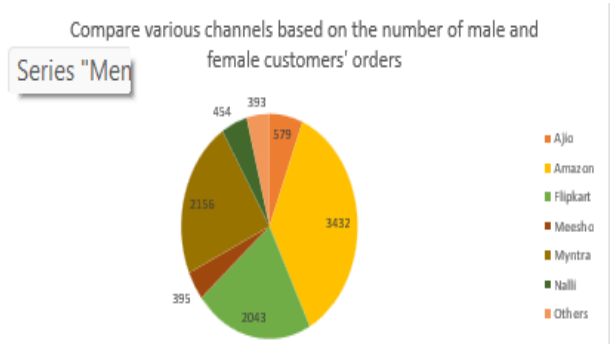
Standard Deviation	9.470706	Standard Deviation	1856.131	Standard Deviation	40.79892	Standard Deviation	20.22993
Sample Variance	89.69428	Sample Variance	3445221	Sample Variance	1664.552	Sample Variance	409.2499
Kurtosis	-0.64676	Kurtosis	1.127057	Kurtosis	-0.19836	Kurtosis	-0.40344
Skewness	0.256745	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

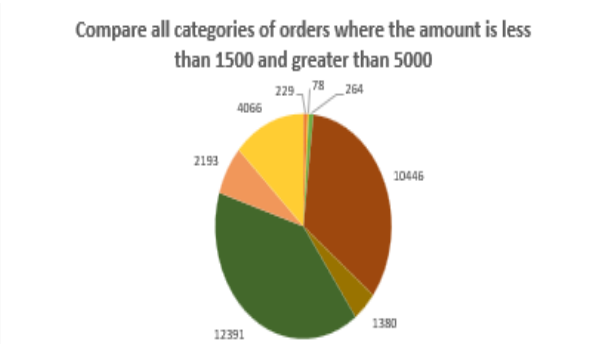
Correlation

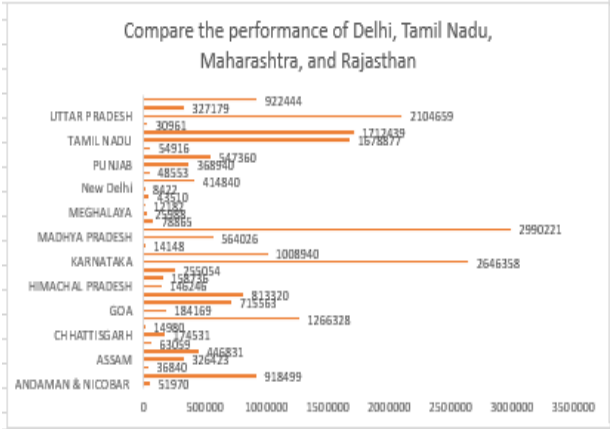
	Column 1	***Column 2***	***Column 3***
Column 1	1		
Column 2	0.513951	1	
Column 3	-0.01254	0.663973	1

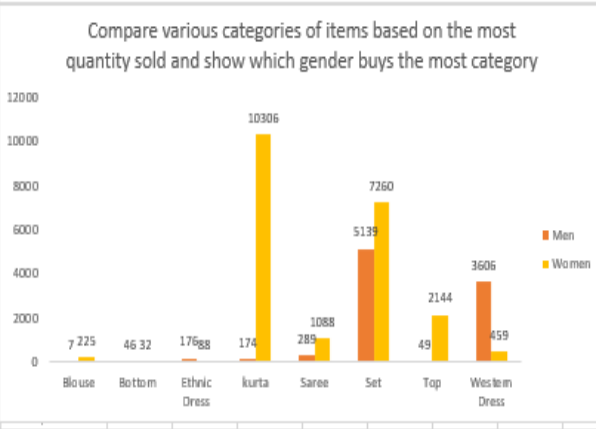
Store Dataset Report

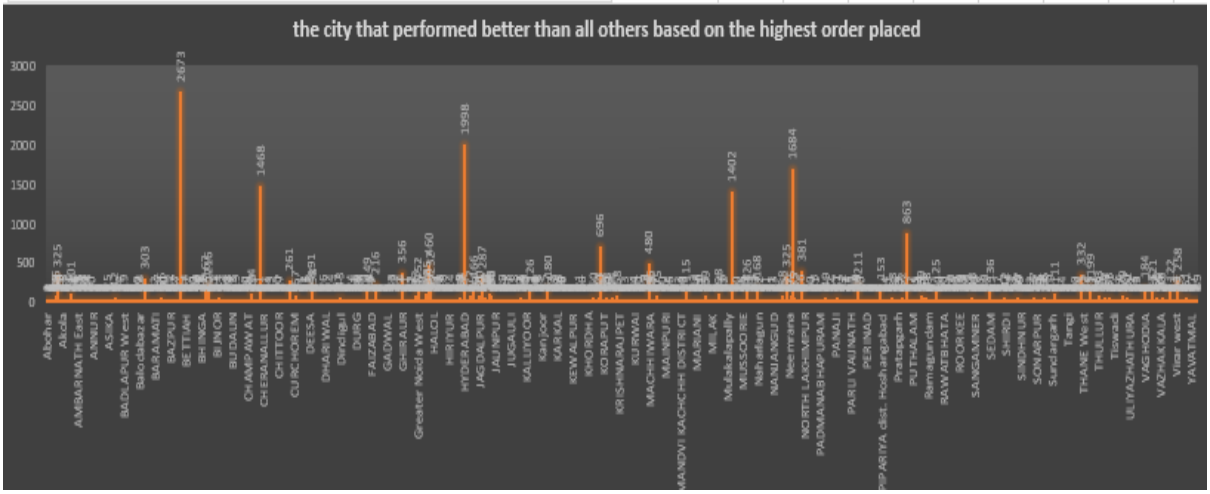
Dashboard











Introduction

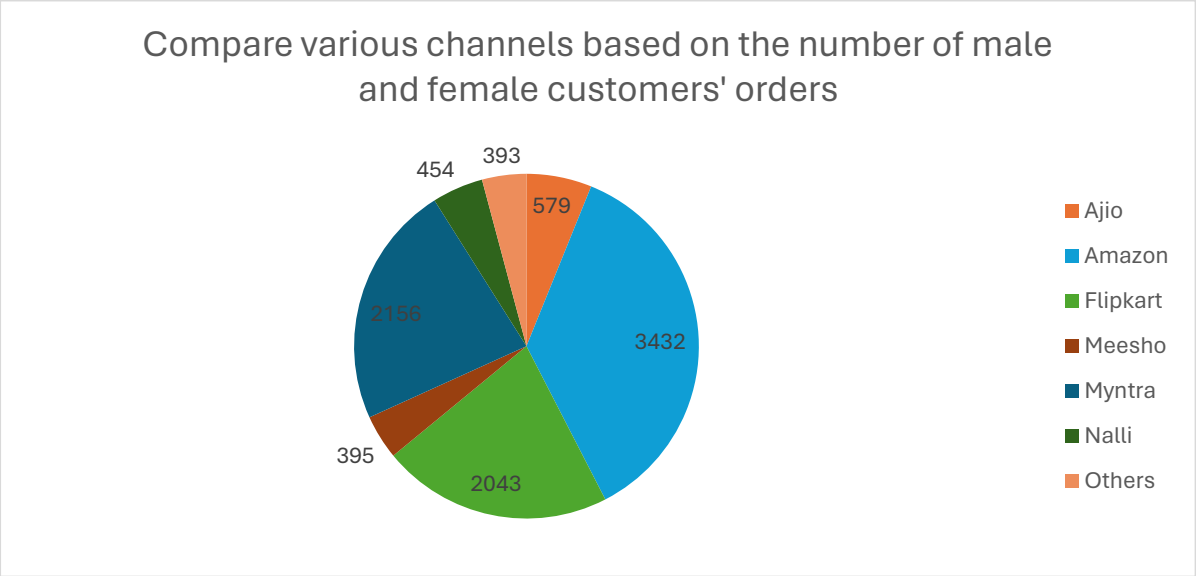
This dataset includes sales information from a retail location that includes a variety of features, including product details (category, SKU), transaction details (order ID, status), and consumer demographics (gender, age group). Finding patterns, preferences, and connections in the data is the aim of our analysis, which aims to clarify consumer behaviour and product trends. Businesses can improve overall consumer satisfaction, expedite inventory management, and improve marketing methods by utilising these insights.

Questionnaire

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

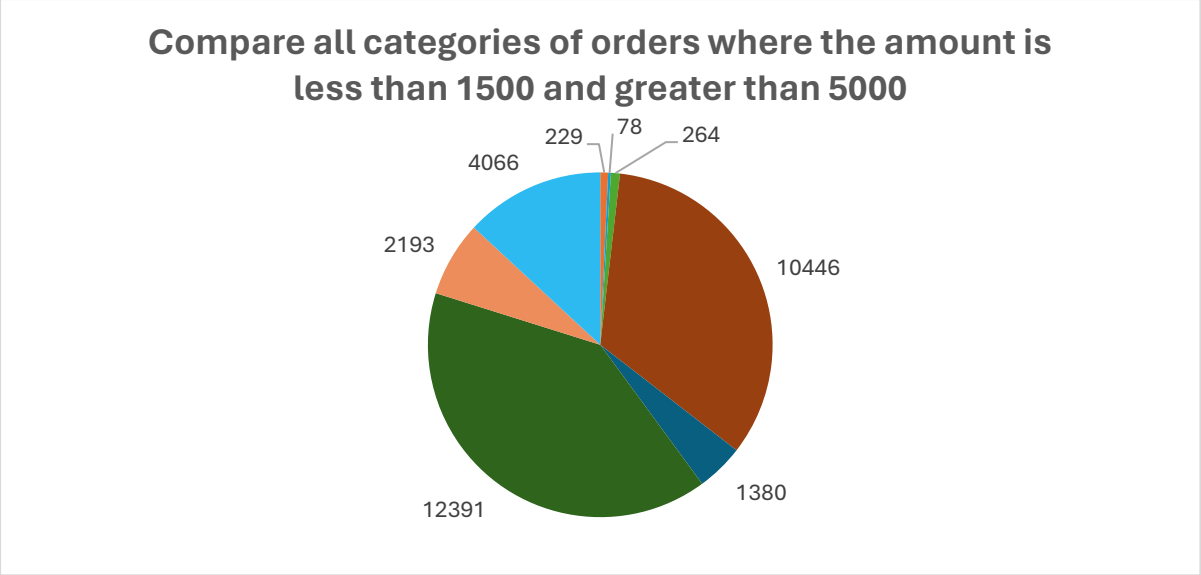
Analytics

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



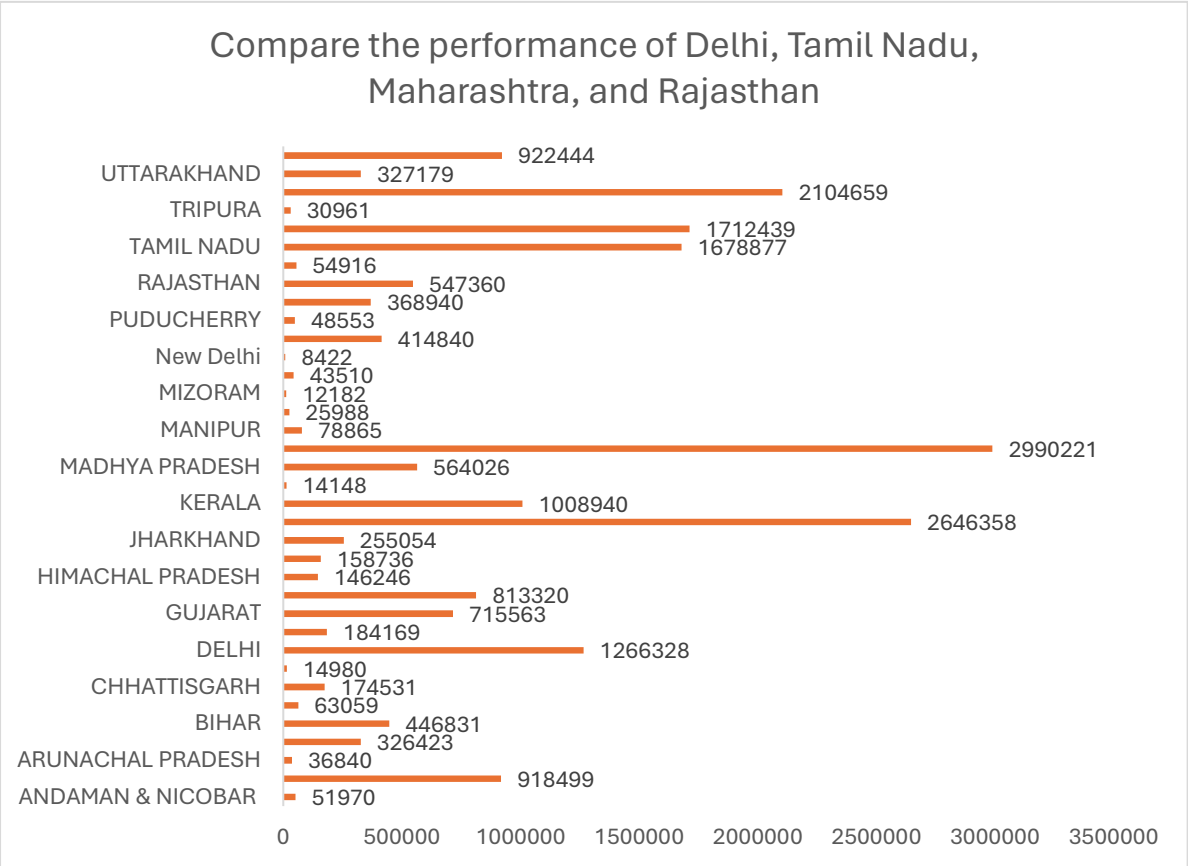
Amazon leads in the sales in both men and women category followed by Myntra and Flipkart. Amazon sold almost 3432 units in men category and almost 7547 units in women category. Myntra sold 2156 units in men section and 5062 units in women section.

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



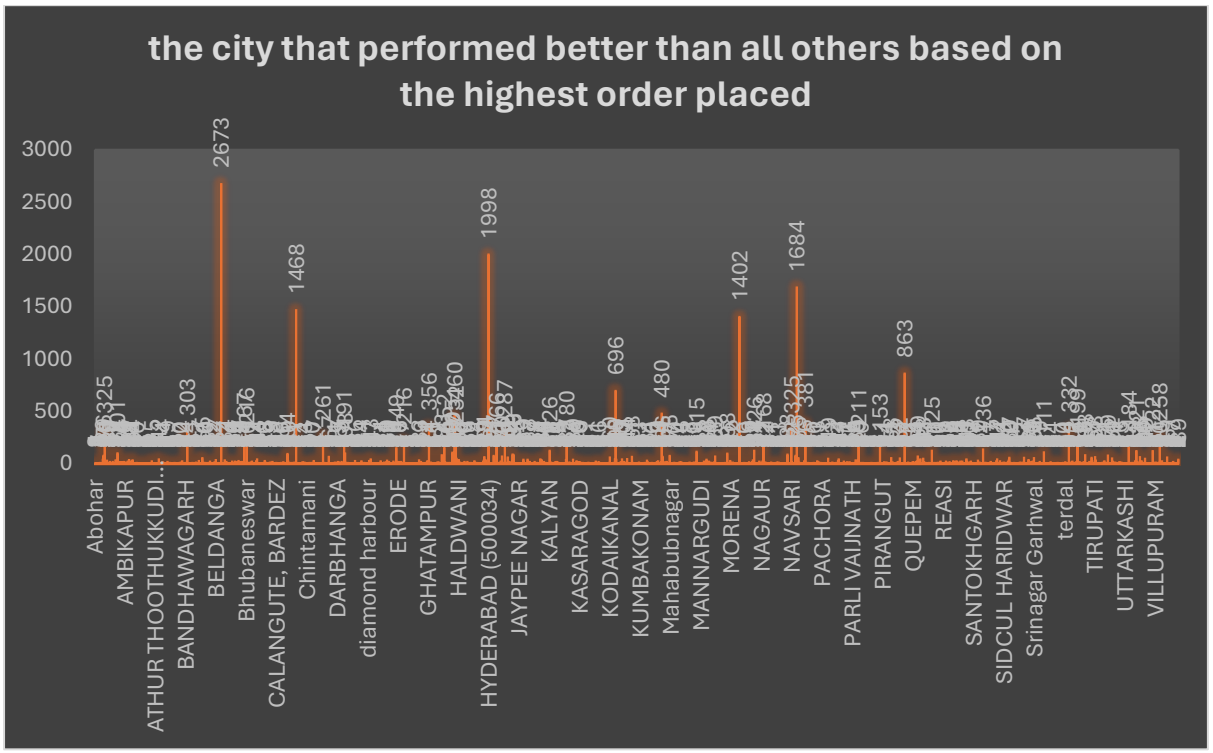
This analysis helps in comparing the categories of order where amount is less than 1500 and greater than 5000. Showing the kurta(12391) and set(10446) with highest count of the orders followed by western dress, top and saree.

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



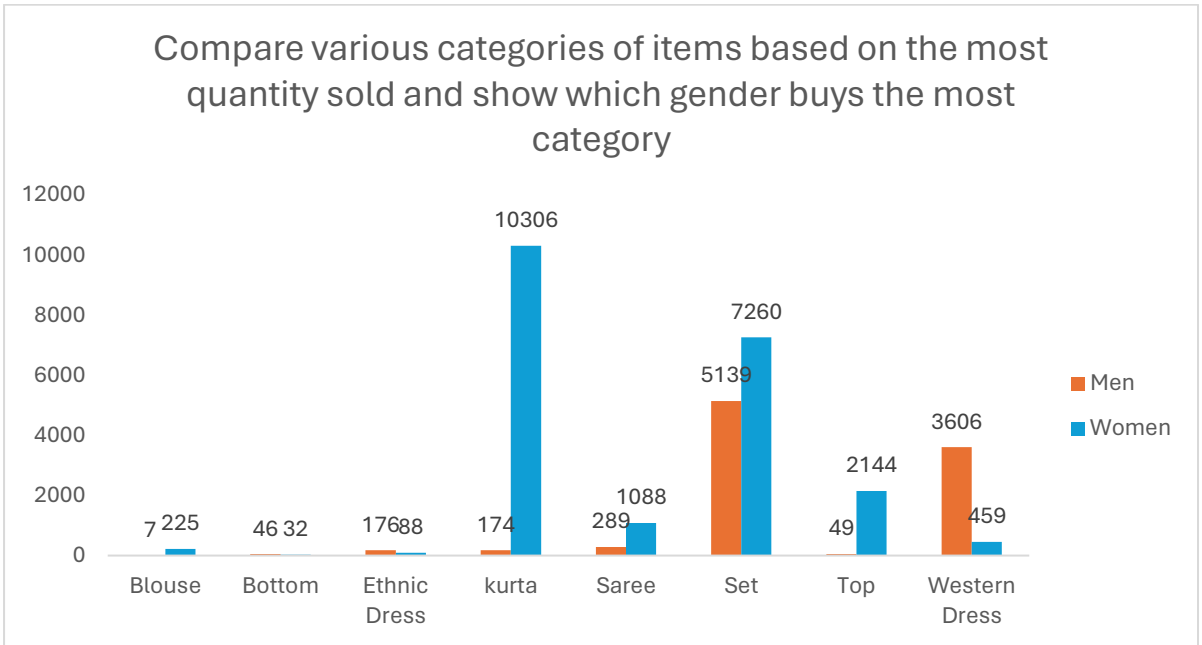
This analysis shows which states performed better than the states mentioned above and Karnataka (2646358) has the highest performance than the other states followed by Uttar Pradesh(2104659).

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



Based on the graph recorded we can actually see which city performed better than all other cities based on highest order placed, so according to graph Bangluru has the highest order placed with 2673 orders followed by Hyderabad(1998).

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



This analysis shows the comparison of various categories of items based on most quantity sold which is kurta bought by women set bought by women followed by men and western dress

followed by top for both men and women.

Conclusion and Review

Amazon leads in sales for both men and women, according to the research, with Myntra and Flipkart trailing closely behind. 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 might be improved by looking into more variables that affect sales. All things considered, the results provide insightful knowledge for maximising 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*			
Regression	2	66561870	33280935	475.4629			
Residual	31044	2.17E+09	69996.92				
Total	31046	2.24E+09					
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	185.155	16.57854	11.16836	6.61E-29	217.6496	152.6604	217.6496
X Variable 1	0.047626	0.099327	0.479489	0.631594	0.242312	-0.14706	0.242312
X Variable 2	492.0276	15.95904	30.83065	1.3E-205	523.308	460.7472	523.308

Anova-1 factor

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	31047	31237	1.00612	0.008853		
Column 2	31047	21176377	682.0748	72136.38		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*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- 2 factor

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

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

Descriptive Statistics

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

Correlation

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

Car Data Forecast

The share datset is showing the share price of TCS share from 1st janary 2024 to 30 january 2024.

Date	Open	Forecast(Open)	Lower Confidence Bound(Open)	Upper Confidence Bound(Open)
01-01-2024	122.8			
02-01-2024	121.2375			
03-01-2024	123.3125			
04-01-2024	123.75			
05-01-2024	123.7375			
06-01-2024	125.75			
07-01-2024	129.9875			
08-01-2024	129.375			
09-01-2024	124.5			
10-01-2024	124.625			
11-01-2024	123.75			
12-01-2024	123.875			
13-01-2024	125.625			
14-01-2024	127.25			
15-01-2024	125.875			
16-01-2024	126.5			
17-01-2024	129.05			
18-01-2024	128.1075			
19-01-2024	131.05			
20-01-2024	131			

21-01-2024	128.6375			
22-01-2024	128.4375			
23-01-2024	127			
24-01-2024	127.4075			
25-01-2024	129.45			
26-01-2024	128.625			
27-01-2024	132.125			
28-01-2024	135.0375	135.037506	135.04	135.04
29-01-2024		135.0476403	131.20	138.90
30-01-2024		135.350153	130.17	140.53
31-01-2024		135.6526657	129.42	141.89
01-02-2024		135.9551785	128.82	143.09
02-02-2024		136.2576912	128.32	144.20
03-02-2024		136.560204	127.89	145.23
04-02-2024		136.8627167	127.52	146.21

The data table presents a detailed account of historical open prices from 01-01-2024 to 27-01-2024 and forecasted open prices from 28-01-2024 to 04-02-2024, alongside their respective confidence bounds. The historical open prices display a range from 121.2375 on 02-01-2024 to a high of 132.125 on 27-01-2024, showing some fluctuations but generally an upward trend over the period. The data indicates a gradual increase, highlighting potential stability and growth in the market during the observed days.

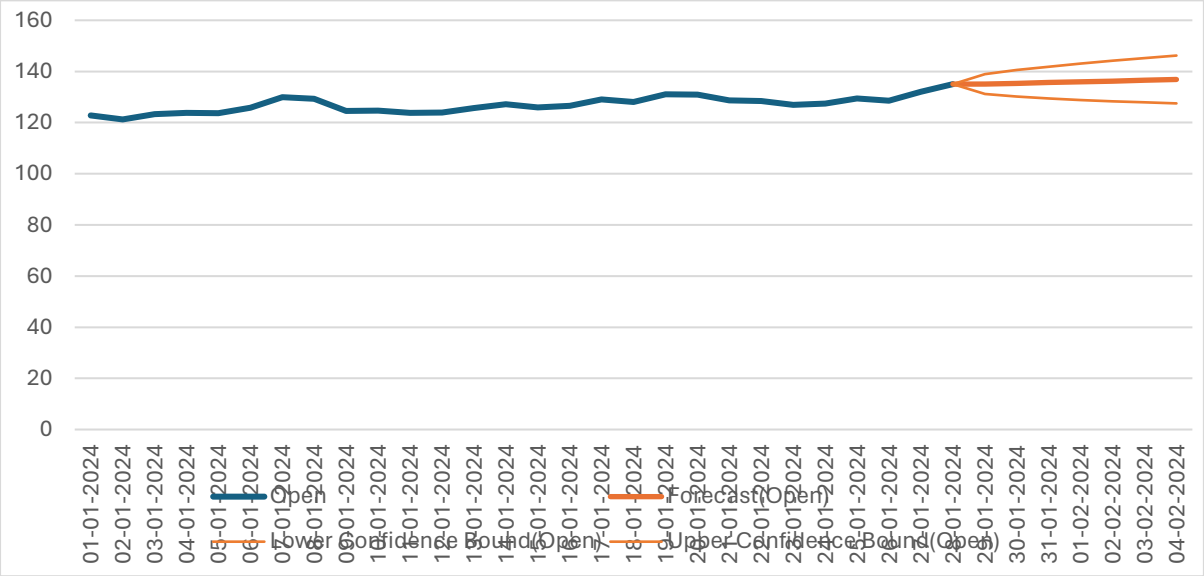
The forecasted open prices begin at 135.037506 on 28-01-2024 and rise steadily to 136.8627167 by 04-02-2024. Each forecasted value is accompanied by a lower and upper confidence bound, providing a range within which the actual open prices are expected to lie. For example, the forecasted open price on 29-01-2024 is 135.0476403, with a lower bound of 131.20 and an upper bound of 138.90. These confidence intervals are relatively narrow, suggesting a high level of certainty and reliability in the forecasts.

The consistent increase in both historical and forecasted open prices suggests a positive market outlook. Investors can interpret this trend as a signal of potential growth, making it a useful

indicator for making informed investment decisions. The confidence bounds further reinforce the reliability of these forecasts, giving investors a clearer picture of the expected price range and helping to manage risks.

Overall, the data provides valuable insights into market trends, showing a steady increase in open prices and offering reliable forecasts with narrow confidence intervals. This information can aid investors, analysts, and market watchers in understanding the market's direction, making strategic decisions, and anticipating future price movements. The upward trend and high forecast certainty underscore the positive market sentiment, making this analysis a critical tool for financial planning and investment strategies.

The predicted graph will be as follow:-



The chart illustrates the historical open prices of a stock from 01-01-2024 to 27-01-2024, followed by forecasted open prices and their corresponding confidence bounds from 28-01-2024 to 04-02-2024.

Key Observations:

1. Historical Data :

- The blue line represents the actual open prices from 01-01-2024 to 27-01-2024.

- The open prices show slight fluctuations but generally maintain an upward trend, starting around 122.8 on 01-01-2024 and reaching approximately 132.125 by 27-01-2024.

2. Forecasted Data :

- The orange line depicts the forecasted open prices starting from 135.037506 on 28-01-2024, continuing to rise to 136.8627167 by 04-02-2024.

- The shaded area around the forecasted line, bounded by the upper and lower confidence bounds (orange lines), indicates the range within which the actual open prices are expected to fall with a high degree of confidence.

- The confidence bounds widen as the forecast moves further into the future, reflecting increased uncertainty.

Implications:

- Upward Trend : Both the historical data and forecasts indicate a general upward trend in the open prices, suggesting positive market performance.

- Forecast Reliability : The relatively narrow confidence bounds close to the forecasted dates (28-01-2024) imply a high level of confidence in the near-term predictions. However, the bounds widen over time, indicating growing uncertainty as the forecast horizon extends.

- Investment Insights : Investors can use this information to anticipate potential price movements. The upward trend and reliable short-term forecasts may encourage buying decisions, while the widening confidence bounds highlight the need for caution in the longer term.

Conclusion:

The chart provides a comprehensive visual representation of the stock's performance, combining historical data with forward-looking forecasts. The overall positive trend and reliable short-term predictions offer valuable insights for investors and analysts monitoring market trends and making informed decisions.