ACROPOLIS INSTITUTE OF TECHNOLOGY & RESEARCH, INDORE

DEPARTMENT OF COMPUTER SCIENCE

CS-605 Data Analytics Lab

3rd Year 6th Semester

2023-2024

SUBMITTED BY –

Anushka Patel

(0827CS211030)

SUBMITTED TO -

Prof. ANURAG PUNDE

S.No.	Experiment	Remarks
1.	Data Analysis Questions: i. Data Analysis Principles ii. Statistical Analytics iii. Hypothesis Testing iv. Regression v. Correlation vi. ANOVA	
2.	Dashboards: i. Store Data Analysis ii. Sales Data Analysis iii. Comprehensive Analysis of Car Attributes: Insights from a Car Collection Dataset iv. Understanding Sales: Orders, Regions, and Segments v. Analysis of Cookie Sales Performance Across Countries vi. Analysis of Loan Applicants vii. Analysis of Sales Performance: Unveiling Insights from Sales Data	
3.	Reports: i. Store Data Analysis ii. Sales Data Analysis iii. Comprehensive Analysis of Car Attributes: Insights from a Car Collection Dataset iv. Understanding Sales: Orders, Regions, and Segments v. Analysis of Cookie Sales Performance Across Countries vi. Analysis of Loan Applicants vii. Analysis of Sales Performance: Unveiling Insights from Sales Data	
4.	Analysis of Forecasted Trends in Flipkart's Closing Stock Prices	

Comprehensive Guide to Data Analysis: Principles, Statistical Analytics, Hypothesis Testing, Regression, Correlation, and ANOVA

Data Analysis Principles

Introduction to Data Analysis

Data analysis is the systematic process of inspecting, cleaning, transforming, and modeling data to discover meaningful patterns, relationships, and insights. It involves using. statistical techniques, mathematical algorithms, and computational methods to extract valuable. information from raw data. The goal of data analysis is to uncover hidden patterns and make. predictions, and support decision-making based on evidence and empirical findings. It plays a crucial role in various fields, including scientific research, business intelligence, and marketing. research, and data-driven decision-making.

There are some principles of data analytics.

1. Accuracy: Ensuring that the data you're analyzing is reliable and free from errors.
2. Relevance: Focusing on the data that is most relevant to your analysis goals.
3. Objectivity: Approaching the analysis without bias or preconceived notions.
4. Interpretation: Analyzing the data in a meaningful way to draw insights and conclusions.
5. Communication: Presenting the analysis findings clearly and effectively.
The principles help guide the process of data analysis and ensure that it is accurate, meaningful, and useful.

Statistical Analytics Concepts

Statistical analytics in data analytics refers to the use of statistical methods and techniques to analyze and interpret data and draw meaningful conclusions. It's a fundamental concept in data analytics that helps in extracting valuable insights from data. It involves applying statistical models, hypothesis testing, regression analysis, and other statistical tools to uncover patterns, relationships, and insights within the data.

By using statistical analytics, analysts can make data-driven decisions, identify trends, and predict future outcomes based on the analysis of numerical data. It's a powerful tool in many fields, including business, finance, healthcare, and social sciences.

Some common statistical models used in data analytics include linear regression, logistic regression, decision trees, random forests, support vector machines (SVM), k-means clustering, and time series analysis. These models help in understanding relationships between variables, making predictions, identifying patterns, and grouping similar data points. Each model has its strengths and is used based on the specific analysis goals and characteristics of the data.

Hypothesis Testing

A hypothesis is a statement or assumption that is made about the relationship between variables or the characteristics of a population. It serves as a starting point for analysis and is tested using statistical methods. The hypothesis can be either null (no relationship or difference) or alternative (there is a relationship or difference).

By testing the hypothesis with data, analysts can determine if there is enough evidence to support or reject the hypothesis. Hypothesis testing is an important part of data analytics as it helps in making informed decisions and drawing conclusions from data.

There are two types of hypotheses: the null hypothesis and the alternative hypothesis.

1. **The null hypothesis**, denoted as H0, states that there is no significant relationship or difference between variables or populations. It assumes that any observed differences are due to chance or random variation.
2**. The alternative hypothesis**, denoted as Ha or H1, proposes that there is a significant relationship or difference between variables or populations. It suggests that the observed differences are not due to chance alone. Hypothesis testing involves collecting and analyzing data to determine whether there is enough evidence to support or reject the null hypothesis in favor of the alternative hypothesis.

Regression and its Types

Regression is a statistical technique used in data analytics to understand the relationship between a dependent variable and one or more independent variables. It helps in predicting the value of the dependent variable based on the values of the independent variables.

There are several types of regression, including:

1**. Linear Regression:** This is the most common type of regression, where the relationship between the dependent variable and independent variable(s) is assumed to be linear. It aims to find the best-fit line that minimizes the difference between the observed and predicted values.

Let's say we want to predict a student's exam score based on the number of hours they studied. Linear regression would help us find the line that best fits the data points, allowing us to predict the exam score for a given number of study hours.

2. Logistic Regression: Logistic regression is used when the dependent variable is categorical or binary. It predicts the probability of an event occurring based on the values of the independent variables.

Suppose we want to predict whether a customer will churn or not based on their demographic information. Logistic regression can help us calculate the probability of churn based on variables like age, gender, and income.

3. Polynomial Regression: Polynomial regression is an extension of linear regression where the relationship between the dependent variable and independent variable(s) is modeled using higher-degree polynomial equations.

Imagine we have data on the number of years of experience and corresponding salary for a group of employees. Polynomial regression can help us model a curve that fits the data points, allowing us to predict salary based on years of experience.

4. Ridge Regression: Ridge regression is a regularization technique used to handle multicollinearity (high correlation) among independent variables. It adds a penalty term to the regression equation to reduce the impact of multicollinearity.

Let's say we have a dataset with highly correlated independent variables, such as height and weight. Ridge regression can help us handle this multicollinearity issue by adding a penalty term, allowing us to make more accurate predictions.

5. Lasso Regression: Lasso regression is another regularization technique that not only handles multicollinearity but also performs feature selection by shrinking the coefficients of less important variables to zero.

Suppose we want to predict the price of a house based on various features like square footage, number of bedrooms, and location. Lasso regression can help us select the most important features and shrink the coefficients of less important variables, improving the model's predictive power.

Correlation

Correlation in data analytics is like a way to see if two things are connected or related. It helps us understand if there's a pattern between two or more things. It refers to the statistical relationship between two or more variables. It helps us understand how changes in one variable are associated with changes in another variable. Correlation is measured using a correlation coefficient, which ranges from -1 to 1.

A correlation coefficient of 0 suggests no linear relationship between the variables.

When we say there's a positive correlation, it means that when one thing goes up, the other thing also tends to go up. For example, if you study more, your exam scores might also go up. A positive correlation (between 0 and 1) indicates that as one variable increases, the other variable tends to increase as well. For example, there might be a positive correlation between the amount of studying done and exam scores.

A negative correlation means that when one thing goes up, the other thing tends to go down. For instance, if you spend more time watching TV, your physical fitness level might go down. A negative correlation (between -1 and 0) indicates that as one variable increases, the other variable tends to decrease. For instance, there might be a negative correlation between the number of hours spent watching TV and physical fitness level.

But correlation doesn't always mean that one thing causes the other. It just shows a relationship between them. We need more analysis to figure out if there's a cause-and-effect relationship.

Correlation is helpful in many areas, like figuring out trends in economics, understanding relationships in social sciences, or making predictions in marketing. It's a cool tool to see how things are connected!

ANOVA (Analysis of Variance)

ANOVA, or analysis of variance, is a statistical technique used to compare the means of three or more groups. It helps us determine if there are any significant differences between the groups.

Imagine you have different groups of people, and you want to know if there is a difference in their average heights. ANOVA can tell you if the differences you observe are statistically significant or just due to chance.
It does this by partitioning the total variation in the data into two components: the variation between the groups and the variation within the groups. The variation between the groups is compared to the variation within the groups to determine if there is a significant difference in the means.

By analyzing the variance between groups and within groups, ANOVA helps us understand if there is a significant difference in the means of the groups. It's commonly used in research, psychology, and other fields to compare multiple groups and draw conclusions.

ANOVA calculates an F-statistic, which is the ratio of the between-group variation to the within-group variation. If the F-statistic is large enough, it indicates that the group means are significantly different.
ANOVA is commonly used in experimental studies and research to compare the effects of different treatments or interventions on a dependent variable. It helps researchers determine if there is a significant effect of the independent variable(s) on the outcome variable.
The basic formula for calculating the F-statistic in ANOVA is:

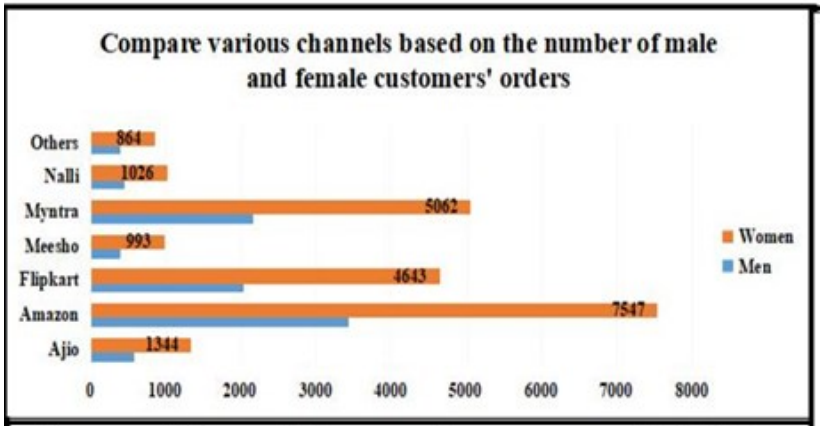
F = (Between-group variation / Within-group variation)

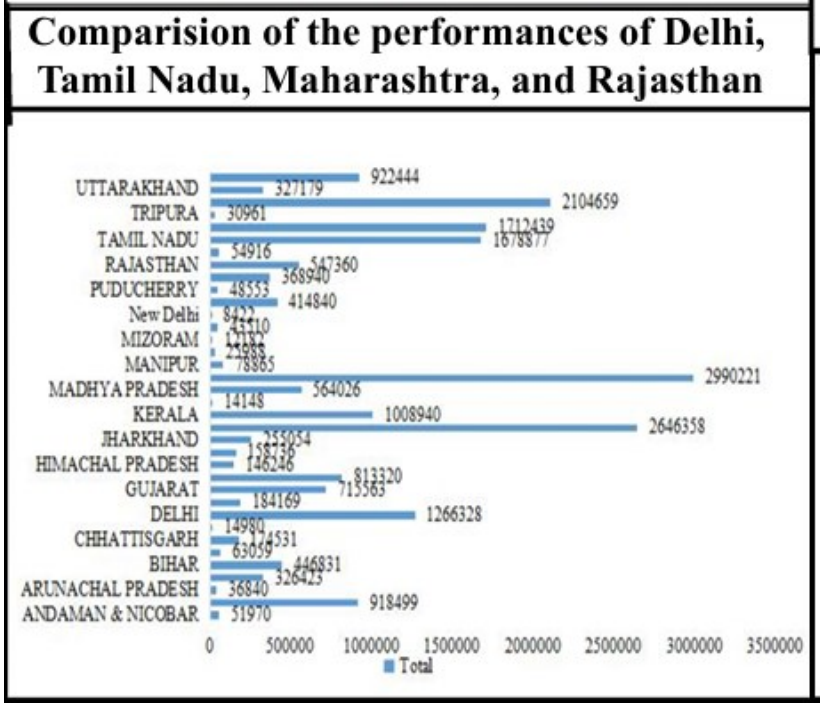
To calculate the between-group variation, the sum of squares between (SSB), which measures the differences between the group means. The within-group variation is calculated using the sum of squares within (SSW), which measures the variability within each group.

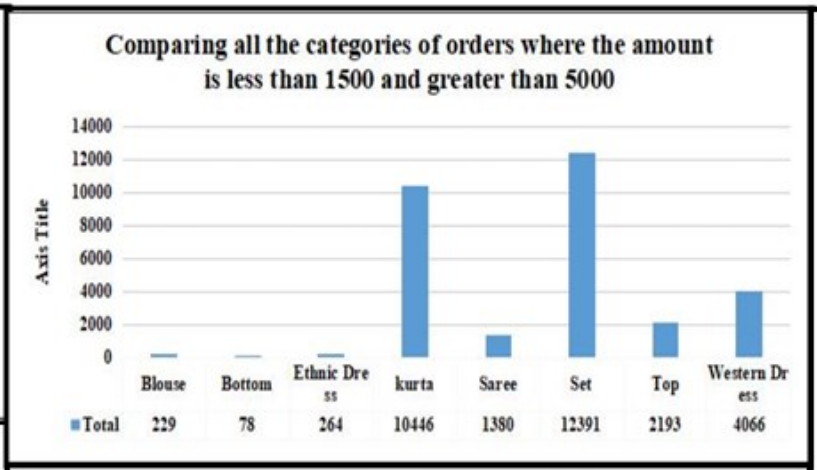
The sum of squares is obtained by summing the squared differences between each observation and the group mean. Then, these sums of squares are divided by their respective degrees of freedom (DFB and DFW) to calculate the mean squares.

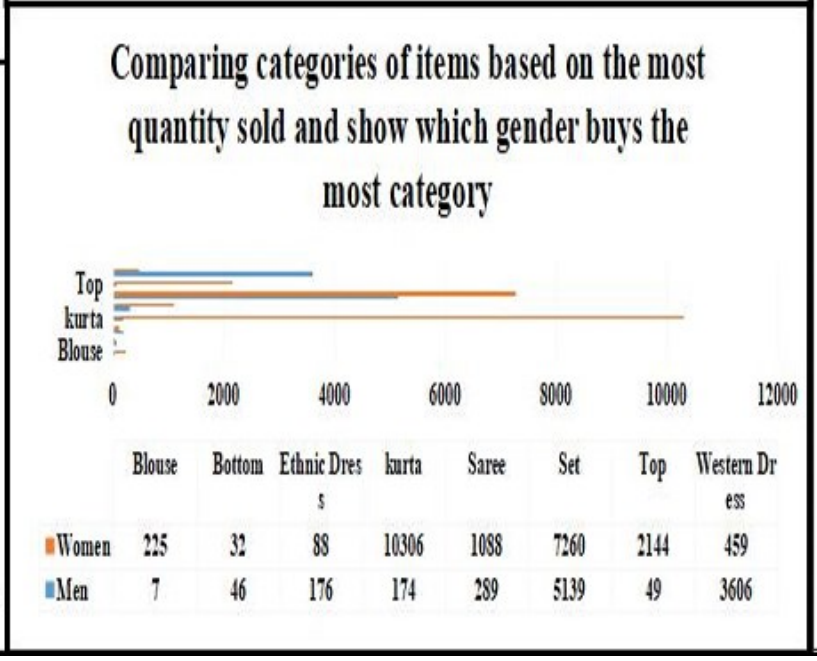
Finally, the F-statistic is calculated by dividing the mean square between by the mean square within.

Store Data Analysis

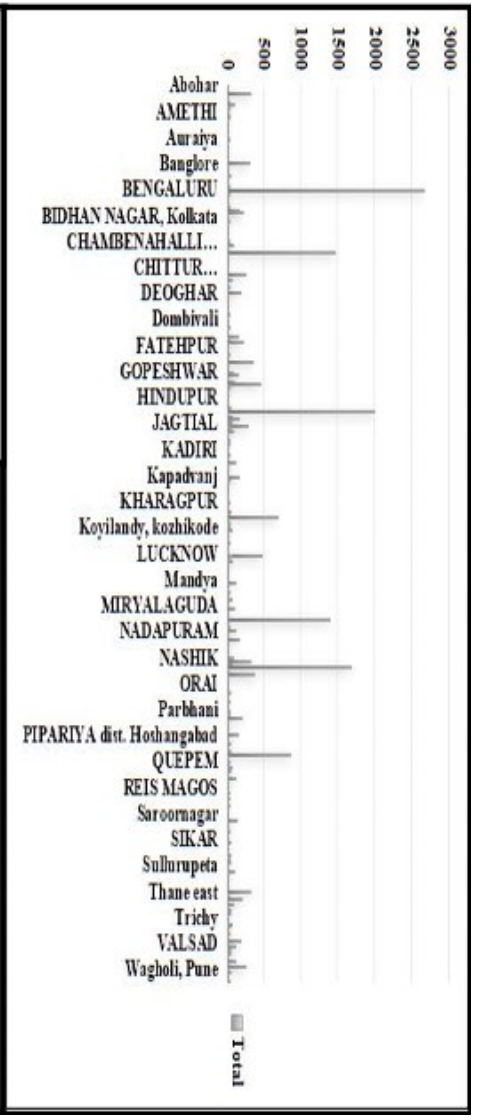




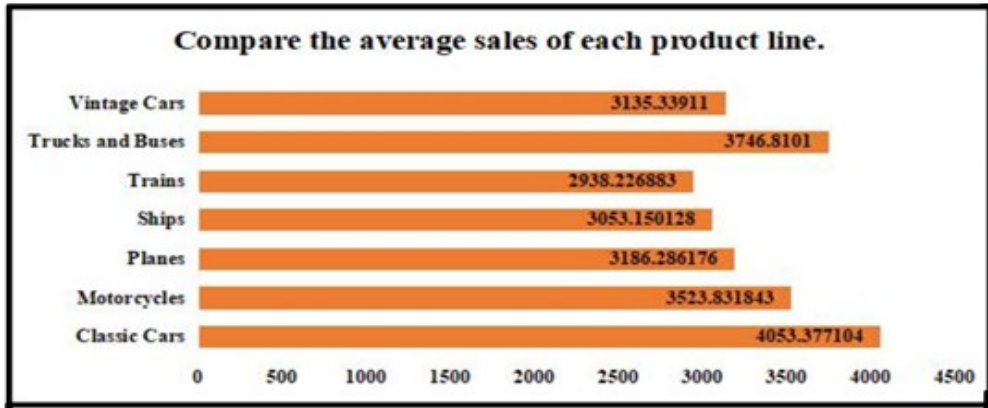


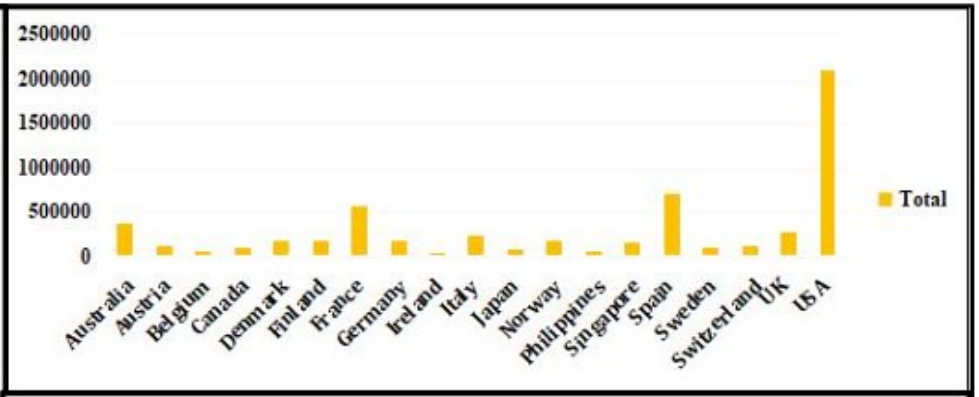


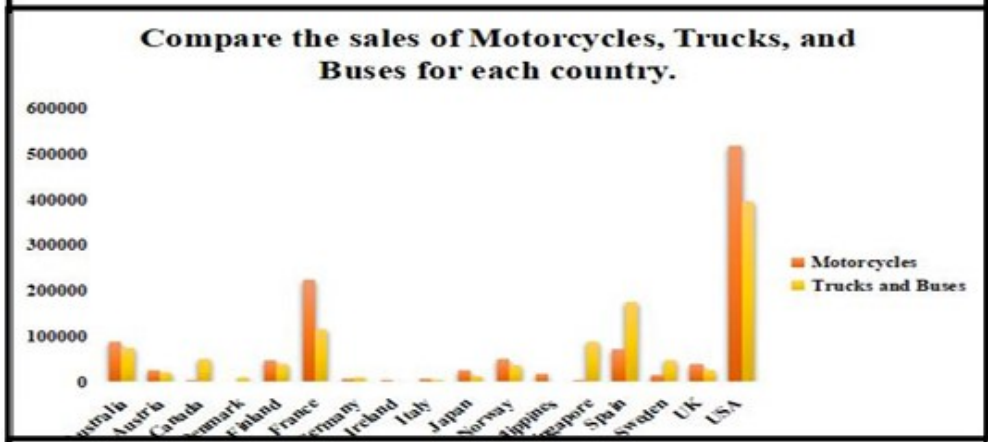
	Blouse	Bottom	Ethnic Dress	kurta	Saree	Set	Top	Western Dress
Women	225	32	88	10306	1088	7260	2144	459
Men	7	46	176	174	289	5139	49	3606

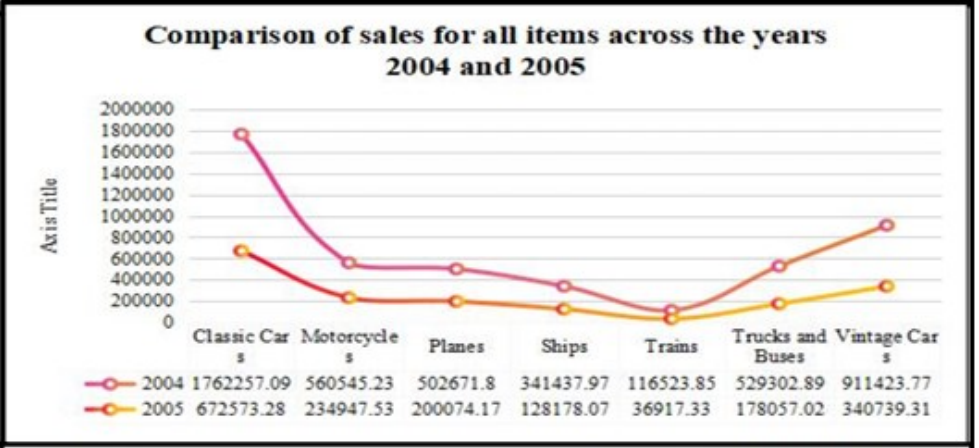


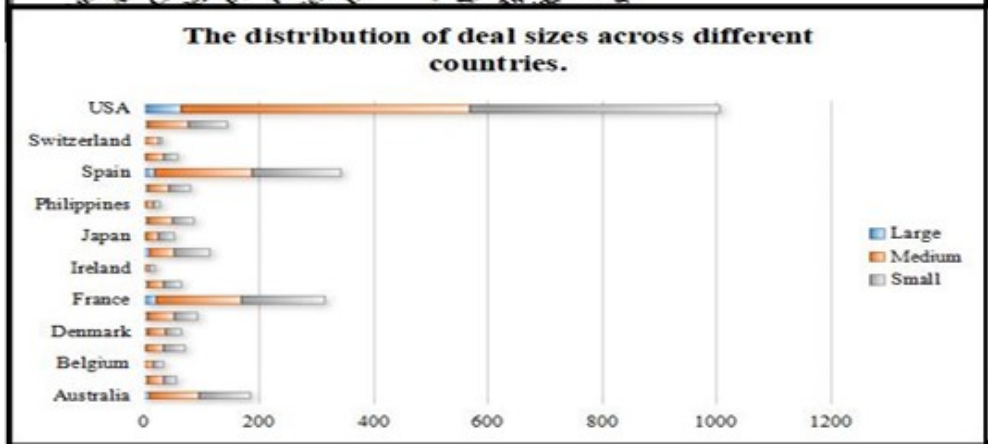
Sales Data Analysis





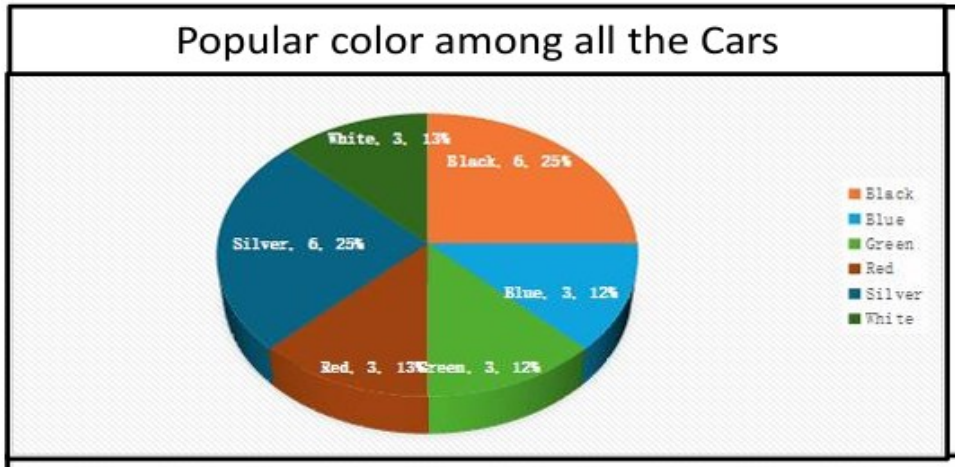


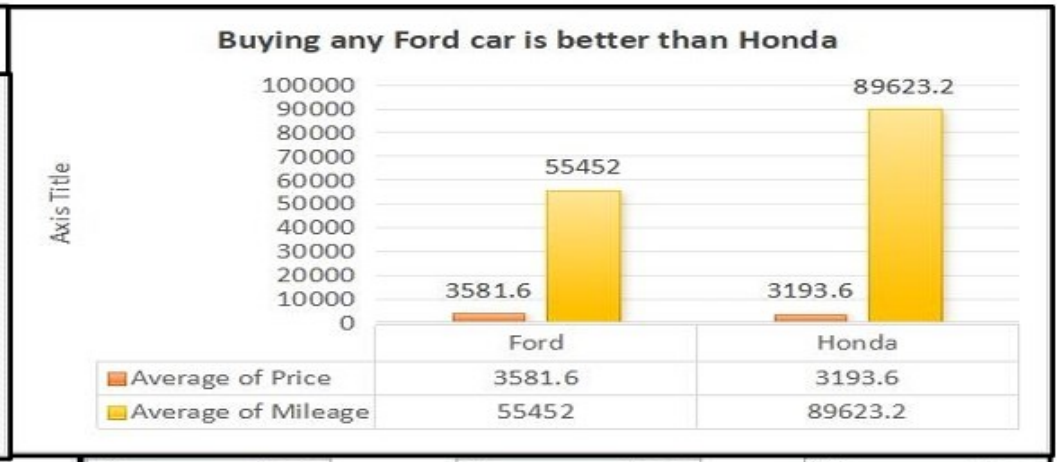


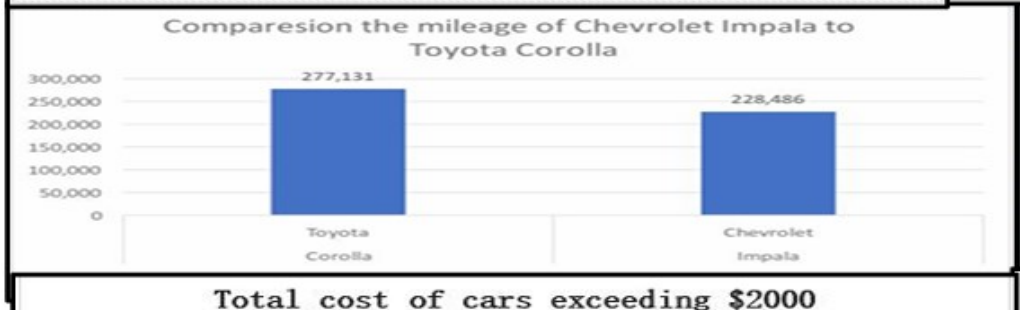


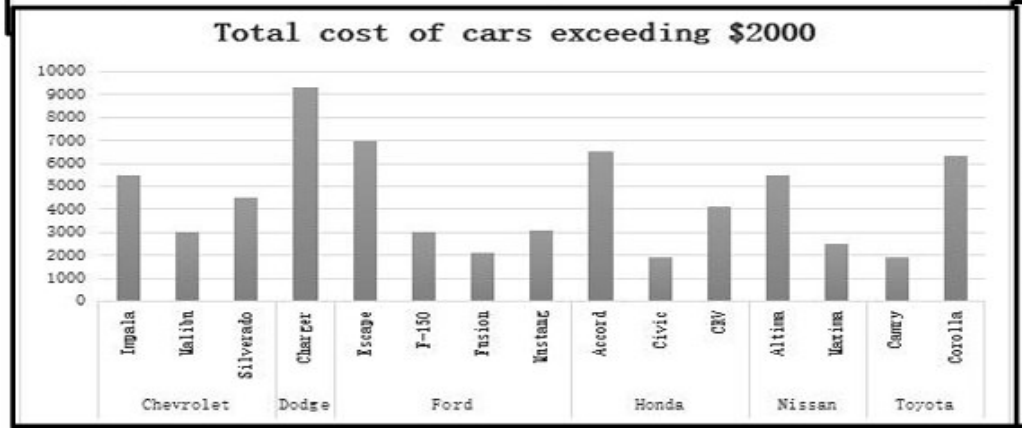
Mean	3553.889072
Standard Error	34.66589212
Median	3184.8
Mode	3003
Standard Deviation	1841.865106
Sample Variance	3392467.068
Kurtosis	1.792676469

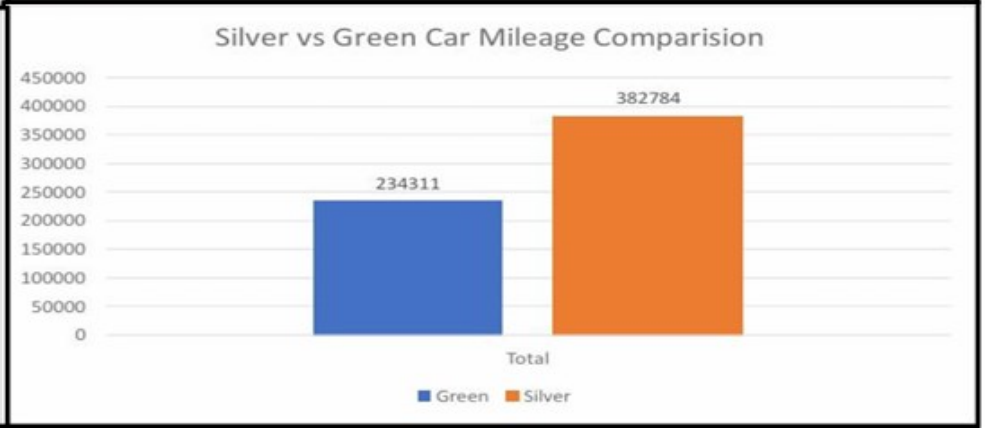
Comprehensive Analysis of Car Attributes: Insights from a Car Collection Dataset



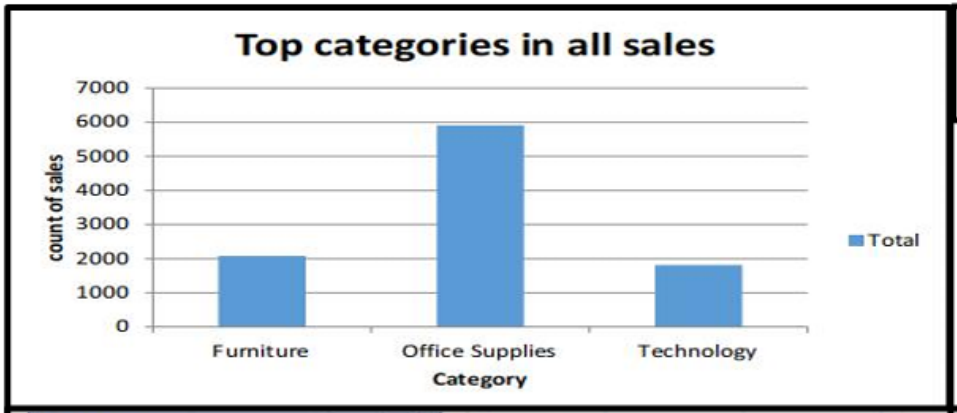


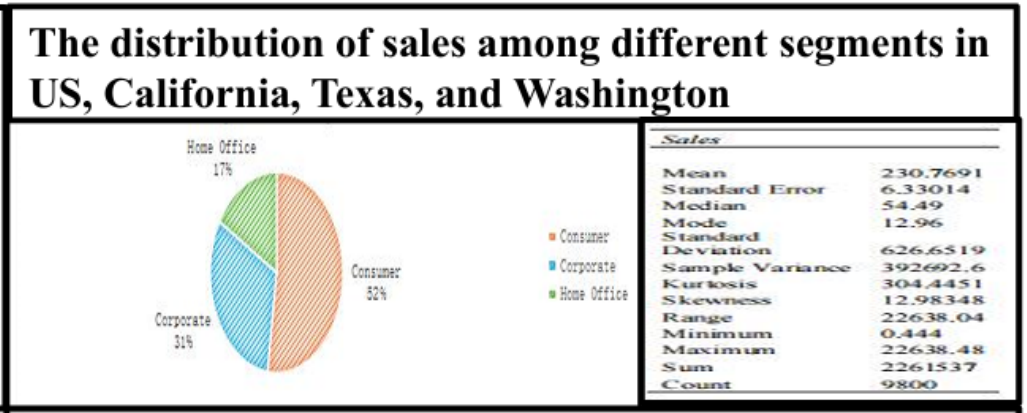


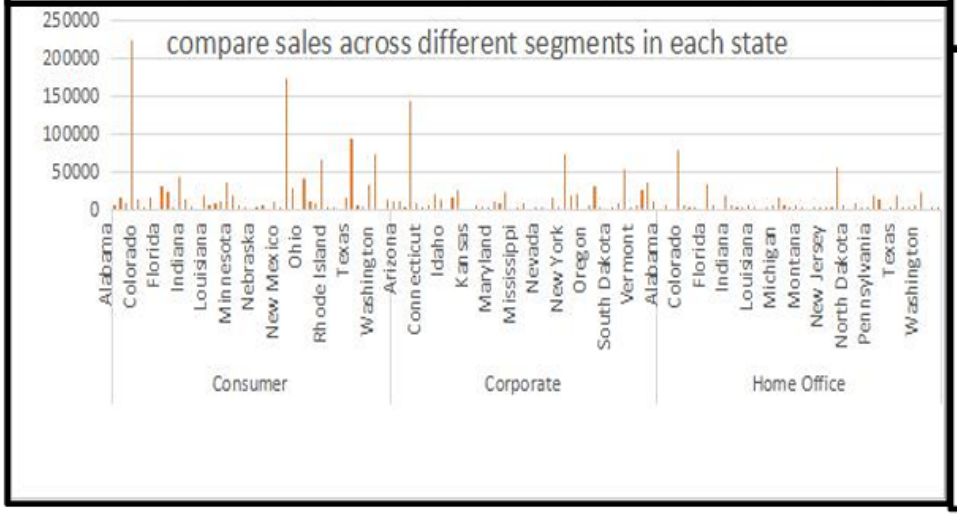


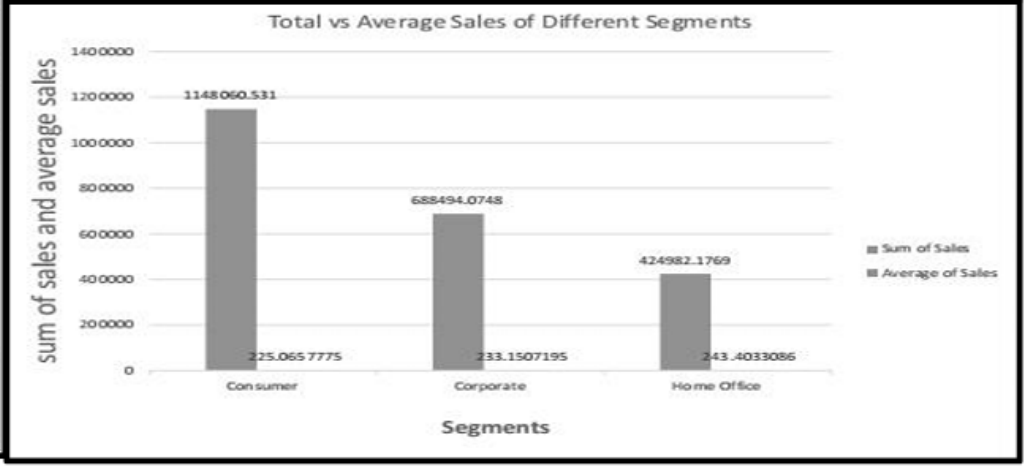


Order Data Analysis

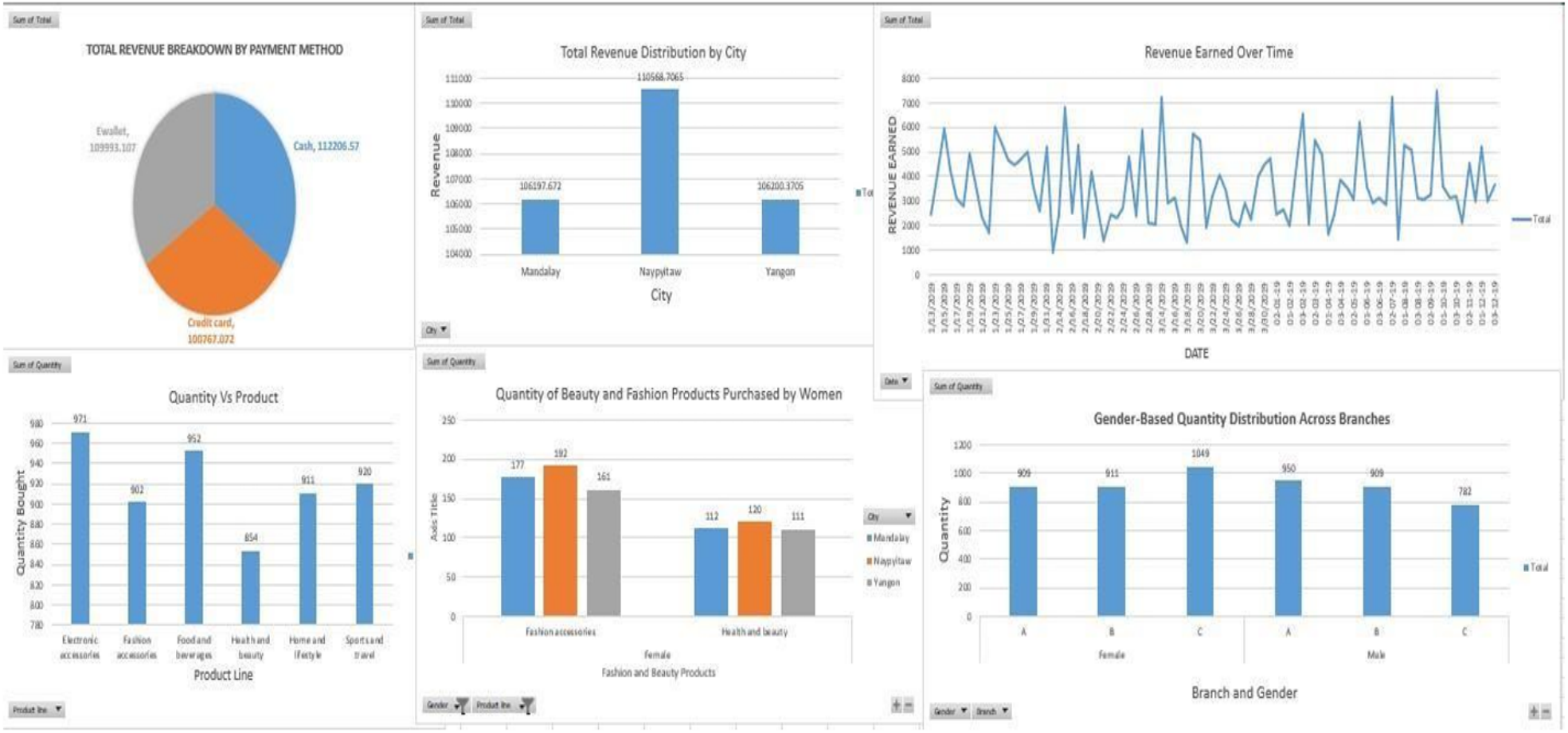




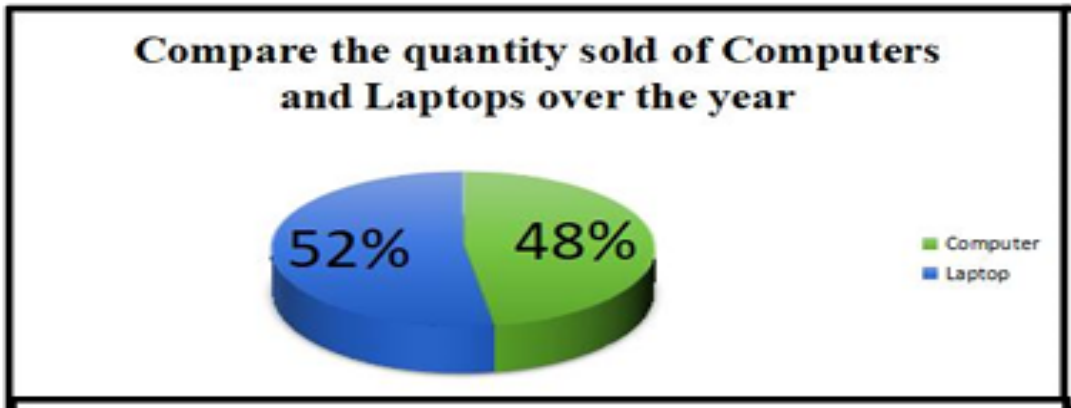


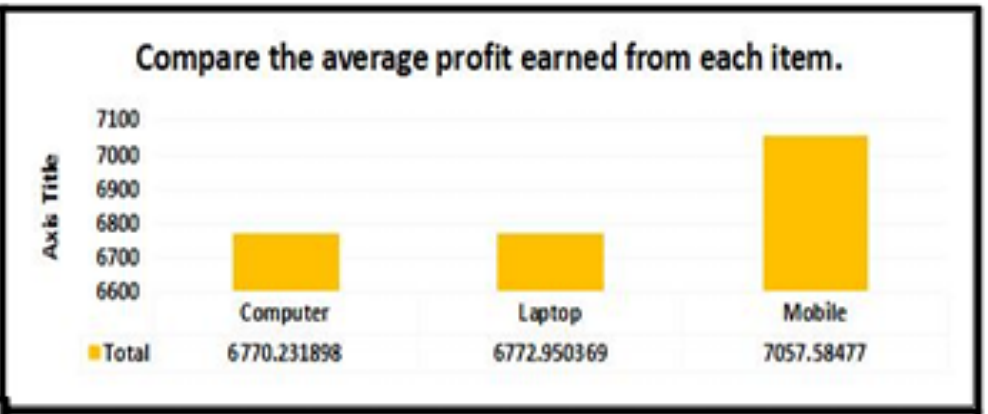


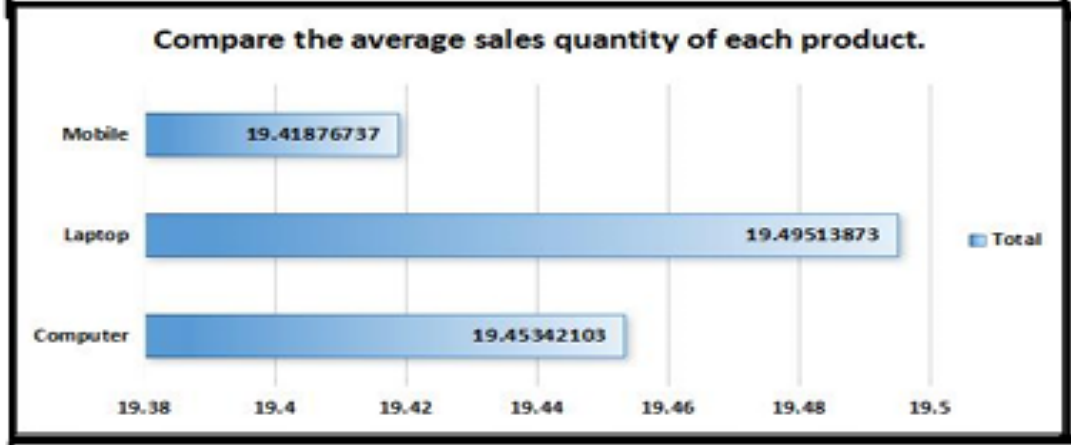
Super Market Data Analysis

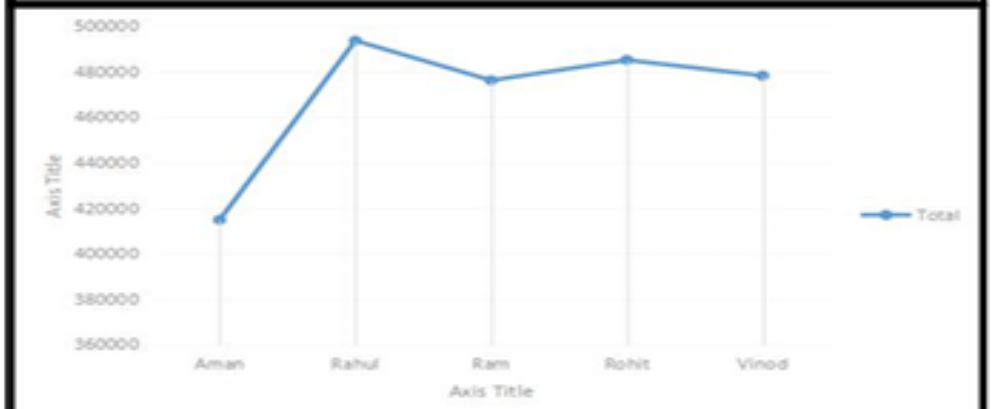


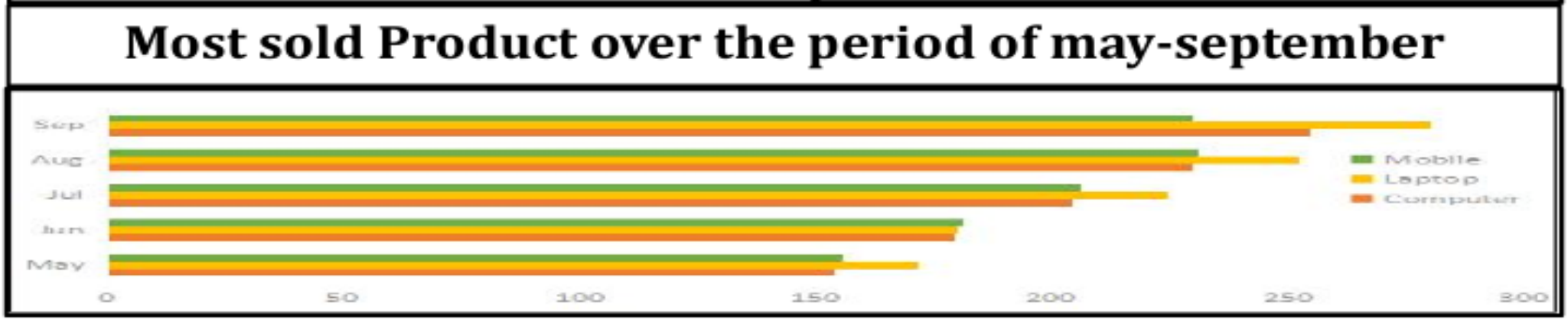
Shop Sales Data Analysis



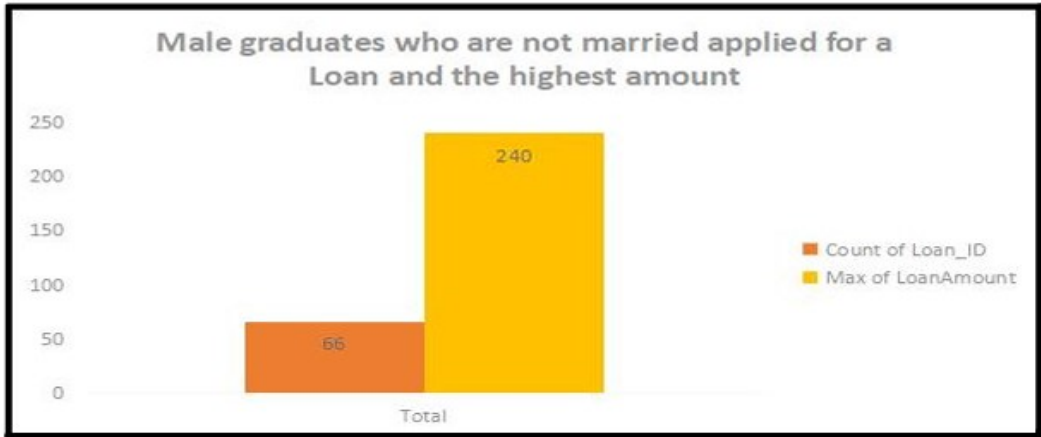


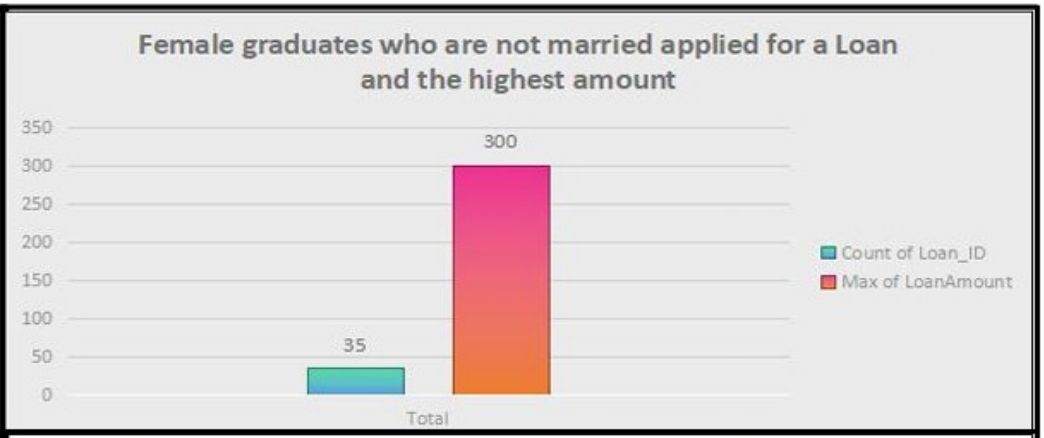


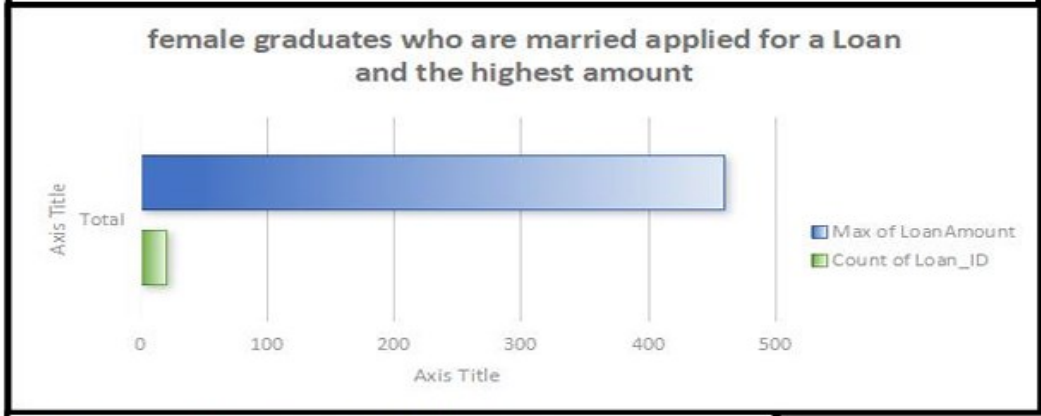


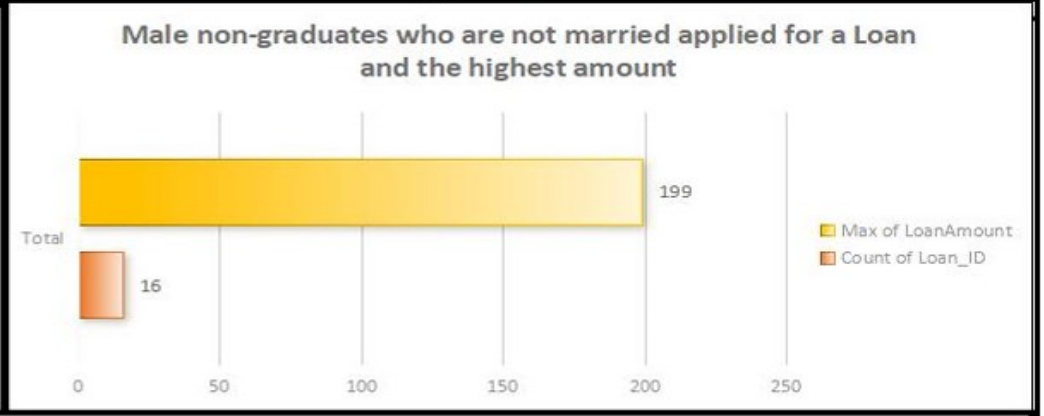


Loan Data Analysis

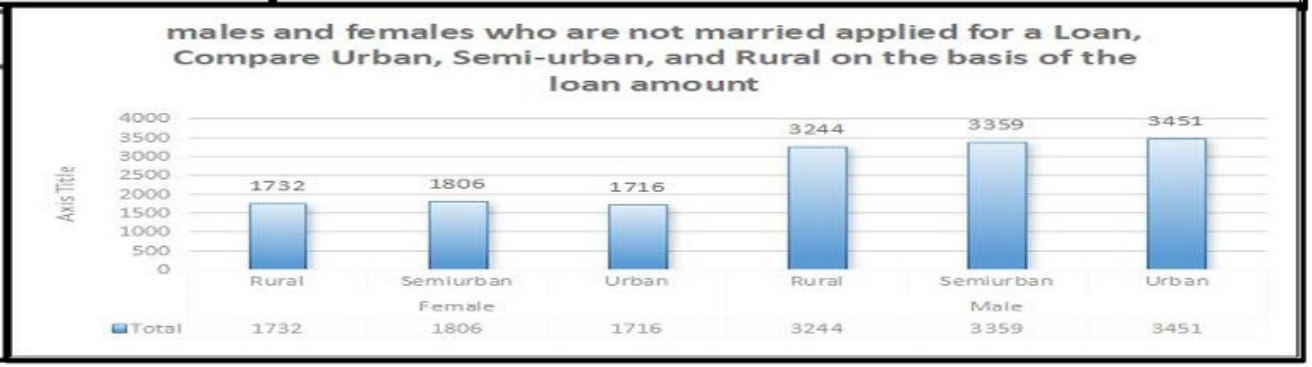




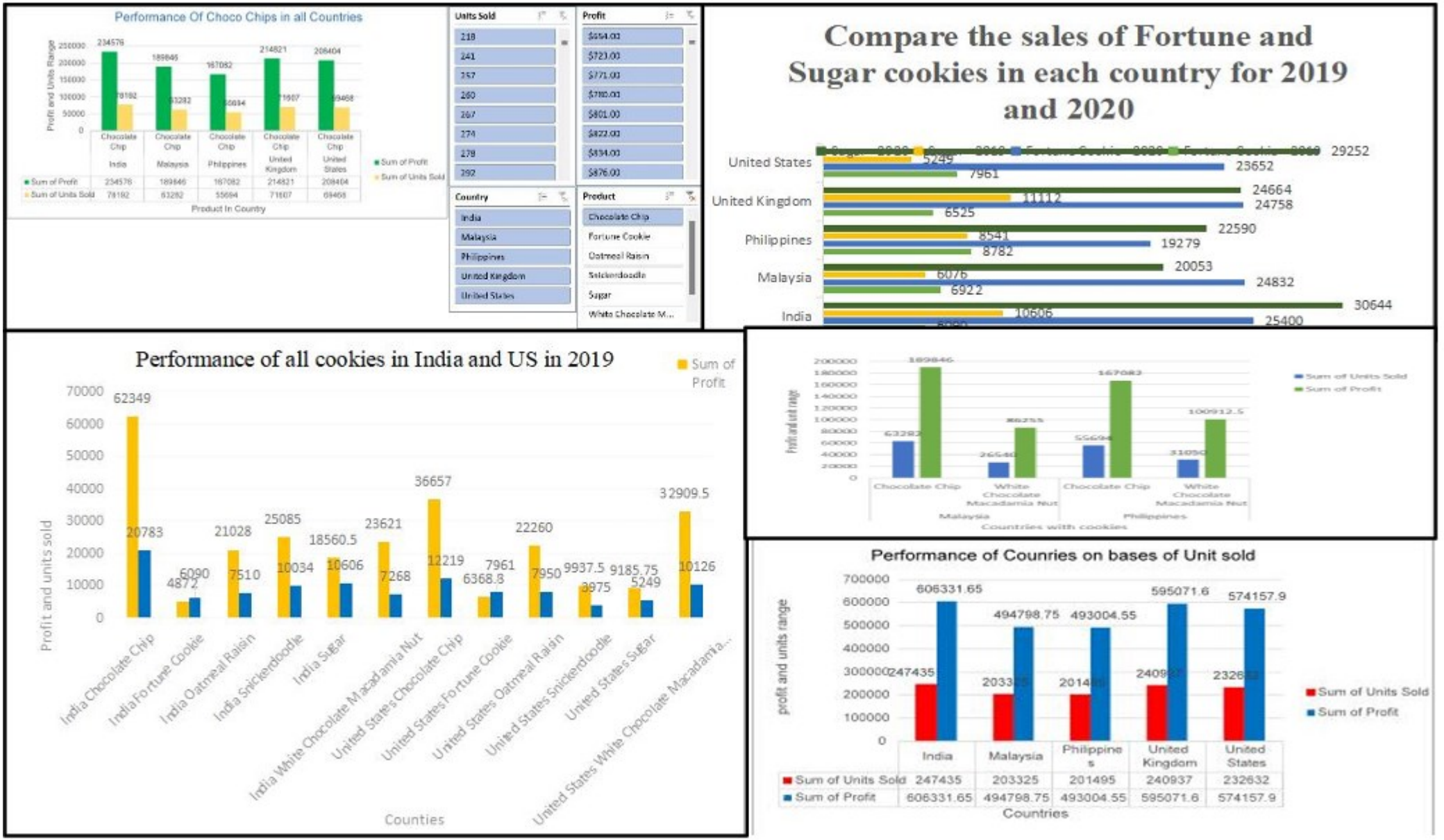




Regression Statistics	
Multiple R	0.45908096
R Square	0.21075532
Adjusted R Square	0.20858707
Standard Error	56.0766111
Observations	366



Cookie Data Analysis



Store Data Analysis

Introduction

This dataset contains retail sales data, featuring customer demographics (Gender, Age Group), transaction details (Order ID, Status), product specifics (Category, SKU), and shipping details. Our analysis focuses on understanding customer behavior and product trends to reveal patterns and preferences. These insights empower businesses to enhance marketing strategies, optimize inventory management, and improve customer satisfaction levels.

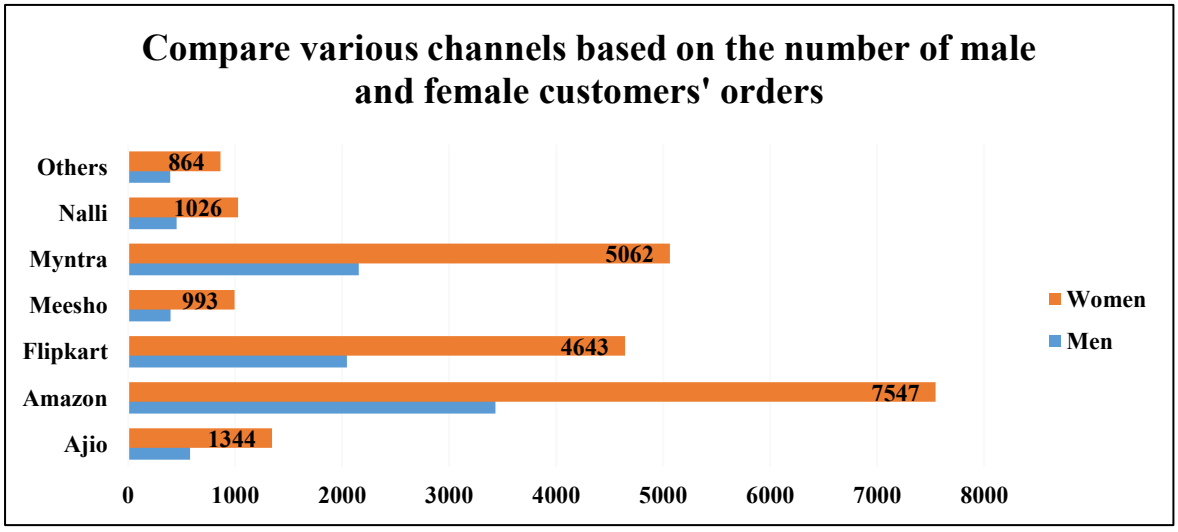
Questionnaire

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 the 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

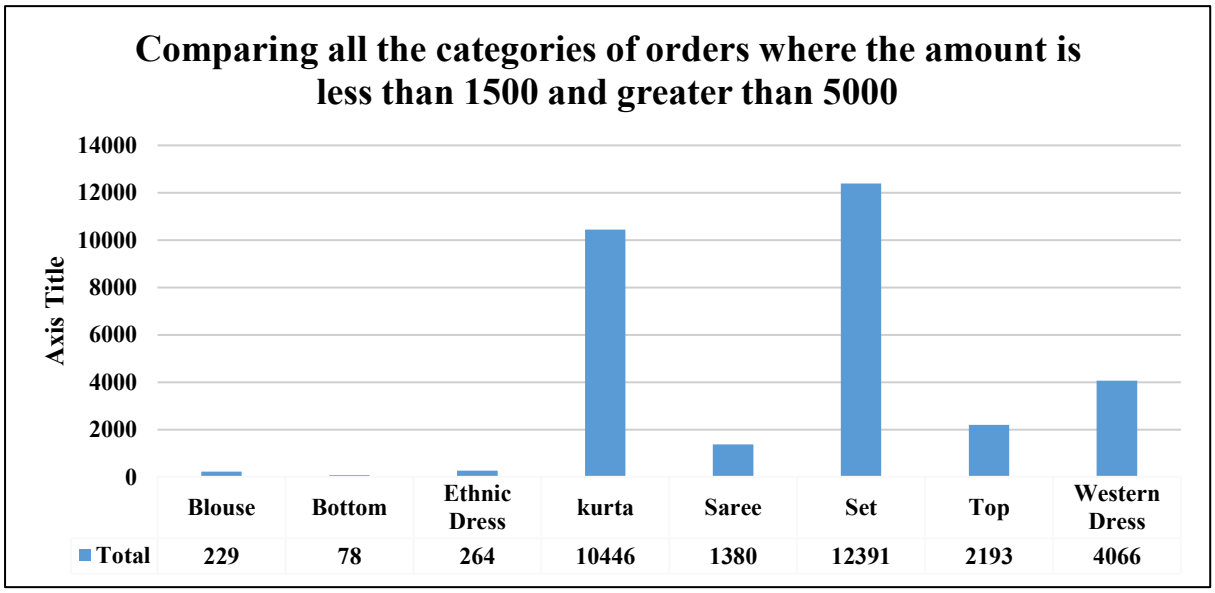
Q1. Compare various channels based on how many male customers order and female customer order?

ANS: Amazon leads in both men's and women's categories, outselling Myntra and Flipkart. Approximately 3,432 units were sold by Amazon in men's fashion, and nearly 7,547 units in women's fashion. Myntra sold around 2,156 units for men and 5,062 units for women.



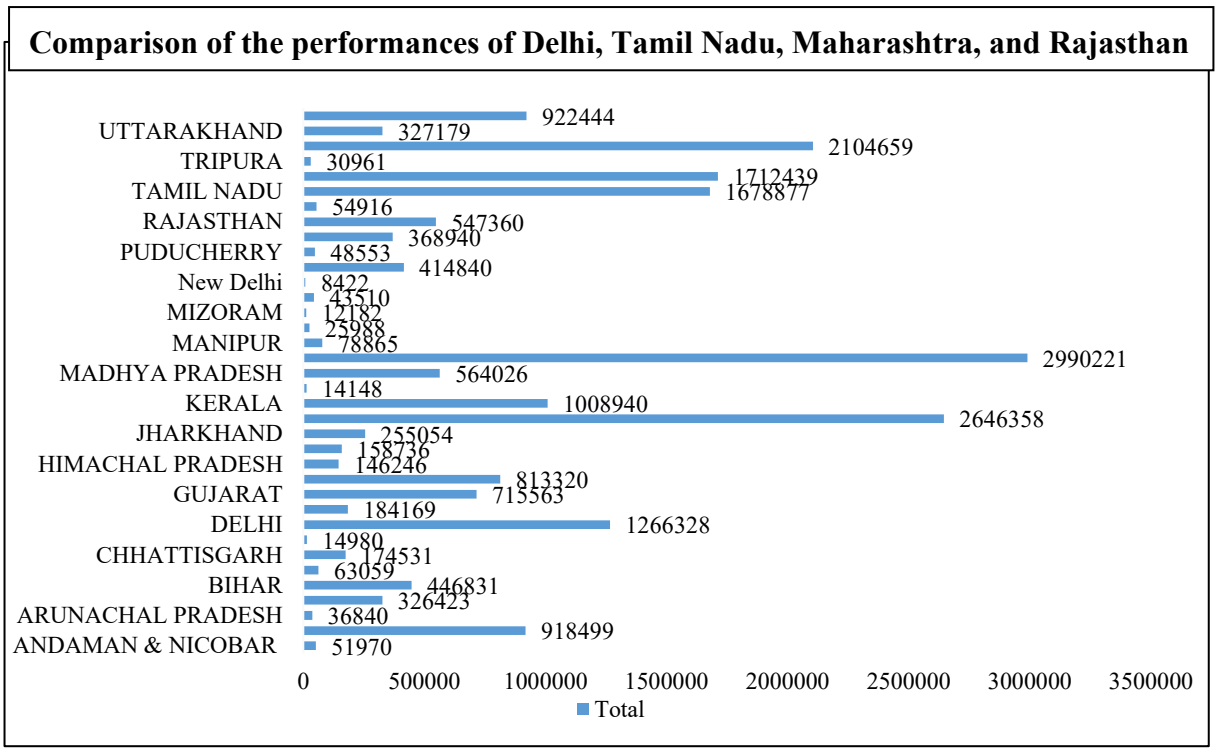
Q2. Compare all the categories of order where amount is less than 1500 and greater than 5000.

ANS: This analysis compares order categories with amounts less than 1500 and greater than 5000. Kurta and set have the highest order counts, followed by western dress, top, and saree, indicating consumer preferences across different price ranges.



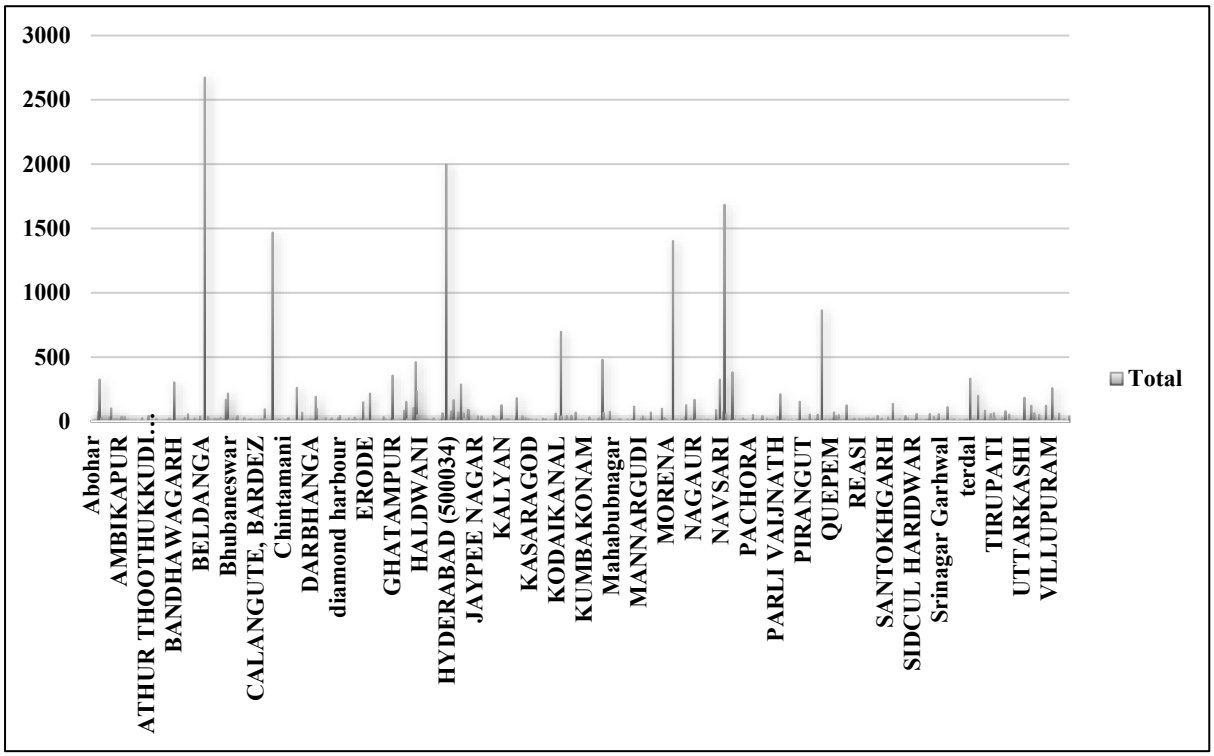
Q3. Which of the following state perform better than other, Delhi, Tamil Nadu, Maharashtra, Rajasthan.

ANS: This analysis identifies states outperforming those mentioned earlier, with Karnataka leading at 2,646,358 units sold, followed by Uttar Pradesh at 2,104,659 units. These states demonstrate superior sales performance. Understanding the top-performing states allows businesses to focus resources effectively and tailor strategies to capitalize on high-performing regions for enhanced sales and profitability.



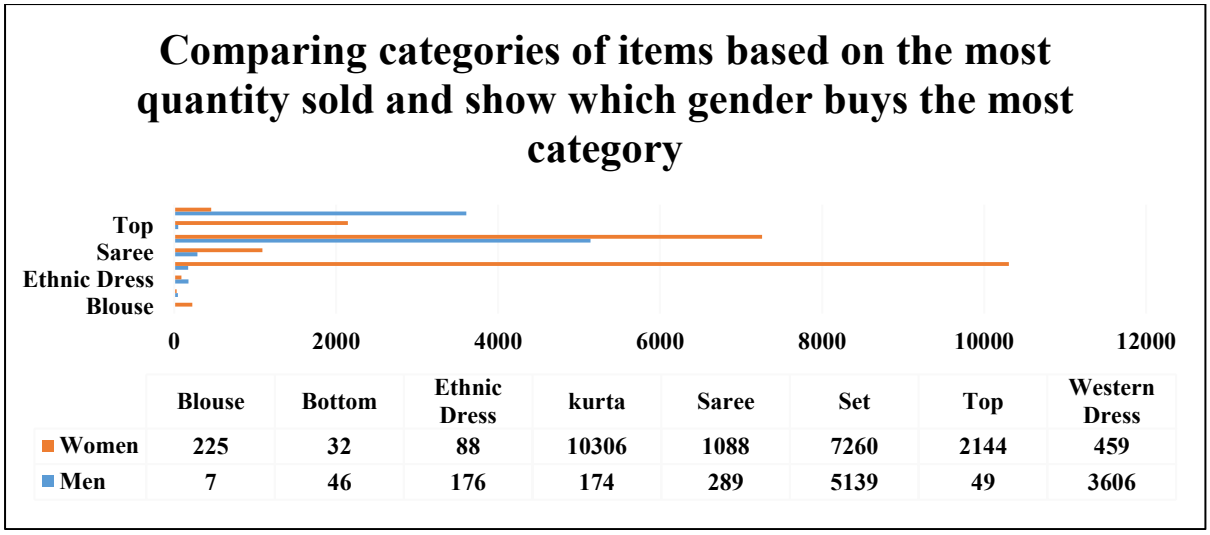
Q5. Which city performed better than all other cities based on highest order placed.

ANS: According to the recorded graph, Bangalore emerges as the top-performing city with the highest number of orders placed, totaling 2,673 orders, followed by Hyderabad with 1,998 orders. This indicates Bangalore's strong market presence and significant order volume.



Q6. Compare various categories of items based on the quantity sold and also show which gender buys the most category.

ANS: The analysis underscores the popularity of kurta and set purchases among women, with kurta being the most bought item, followed by set. For men, western dress emerges as the top choice, while top is favored by both men and women, indicating its universal appeal across genders.



	Blouse	Bottom	Ethnic Dress	kurta	Saree	Set	Top	Western Dress
Women	225	32	88	10306	1088	7260	2144	459
Men	7	46	176	174	289	5139	49	3606

Conclusion and Review

The analysis underscores Amazon's dominance in men's and women's categories, with Myntra and Flipkart trailing closely. Leading sales items include kurta and set, with Karnataka and Bangalore demonstrating exceptional performance. These insights offer valuable guidance for retailers, although delving deeper into other sales-influencing factors could enrich the analysis. Overall, the findings provide crucial insights for refining sales strategies in competitive market landscapes.

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	523.3088	460.7472	523.308

Anova (One factor)

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

Sales Data Analysis

Introduction

The "Sales Dataset" comprises transactional records from a store, encompassing various attributes such as order number, quantity ordered, price each, order line number, sales revenue, order date, status, product line, MSRP, product code, customer name, phone, address details, city, state, postal code, country, territory, contact names, and deal size. This dataset holds substantial value in uncovering insights related to sales trends, product performance, customer behaviour, regional distribution, and market dynamics.

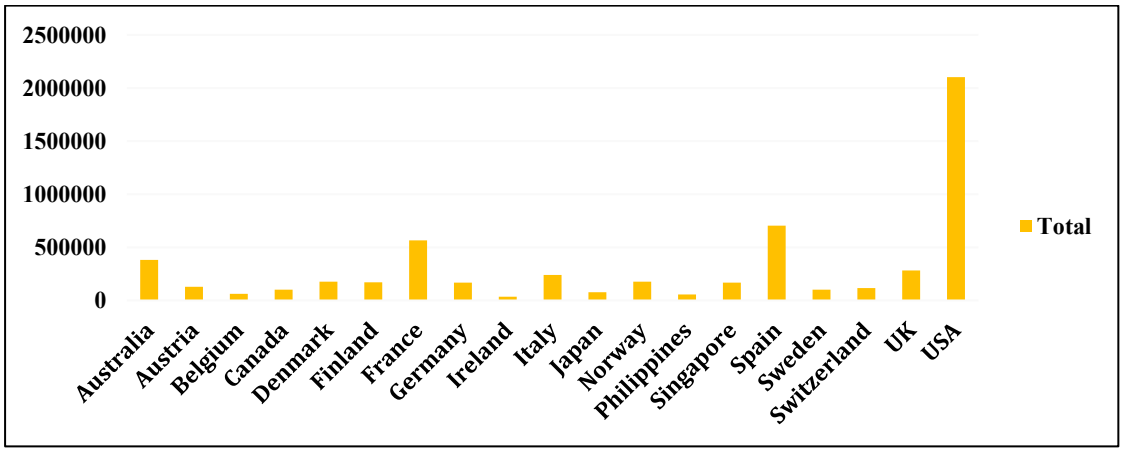
Questionnaire

1. Compare the sale of Vintage cars and Classic cars for all the countries.
2. Find out average sales of all the products? which product yield most sale?
3. Which country yields most of the profit for Motorcycles, Trucks and buses?
4. Compare sales of all the items for the years of 2004, 2005.
5. Compare all the countries based on deal size.

Analytics

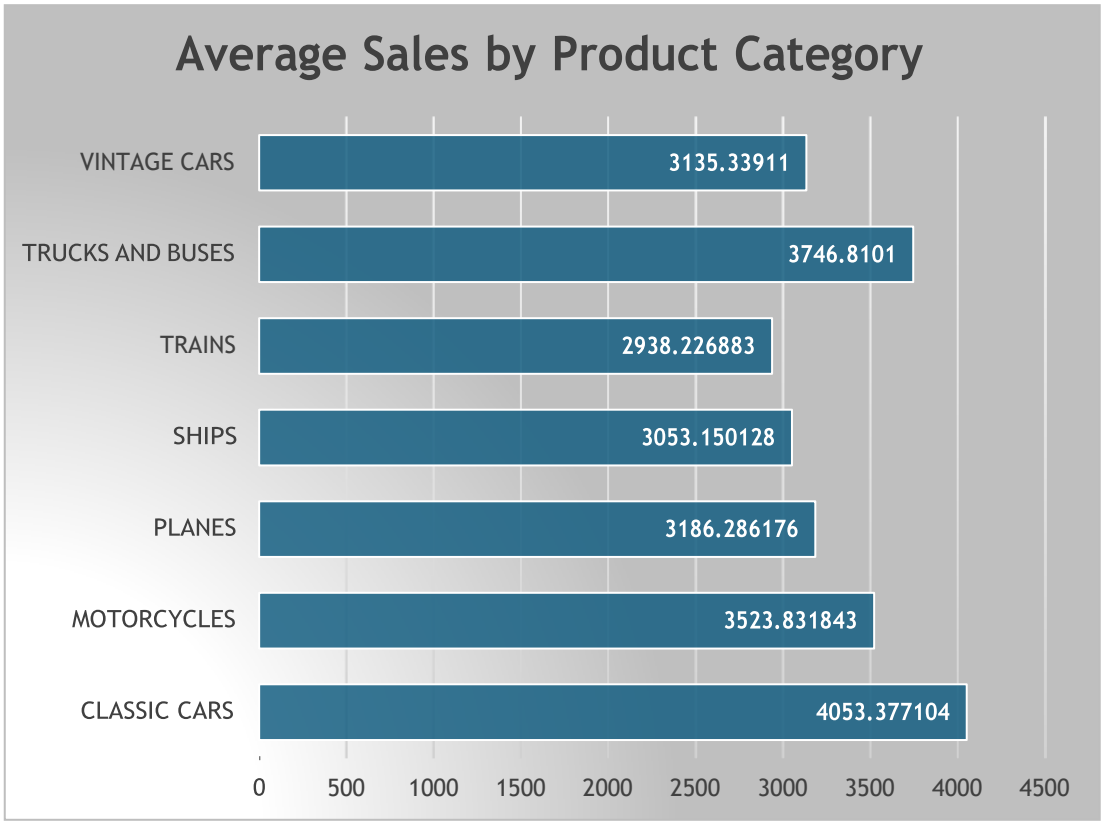
1. Compare the sale of Vintage cars and Classic cars for all the countries:

ANS: This analysis compares sales of Vintage cars and Classic cars across countries.
USA leads with sales of $2,102,394.02, followed by Spain, France, and Australia.



2. **Find out average sales of all the products? which product yield most sale?**
   2.1. **Methodology:** Calculate the average sales for each product and then identify the product with the highest average sales.

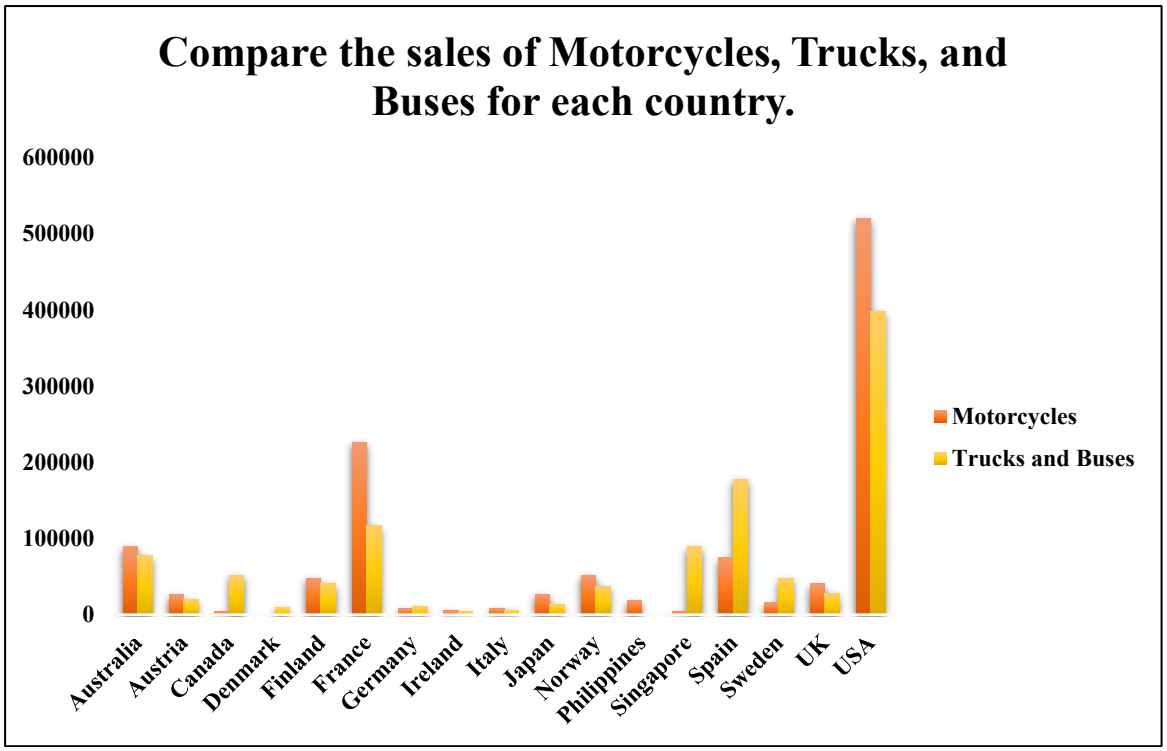
   2.2. Findings:



The analysis of average sales across product categories reveals that Classic Cars lead the pack with an average sales figure of $4,053.38 per transaction. Trucks and Buses

follow closely behind with an average sales value of $3,746.81. This data provides valuable insights into the relative sales performance of different product categories, enabling targeted resource allocation and strategic decision-making.

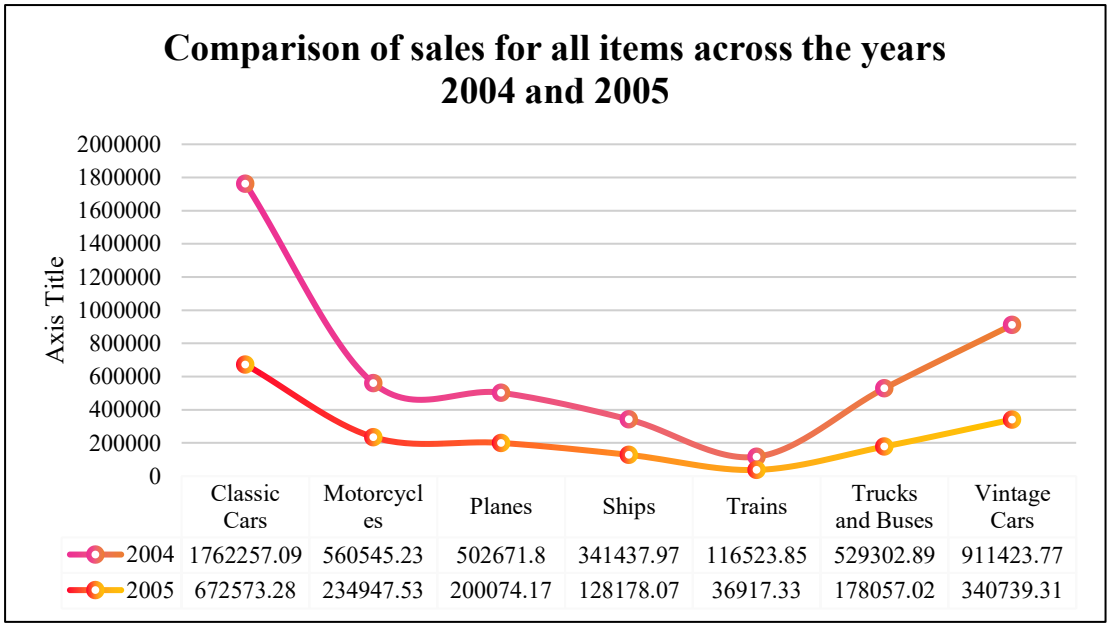
3. **Which country yields most of the profit for Motorcycles, Trucks, and buses?**



The analysis identifies the top-profitable country for Motorcycles, Trucks, and Buses. The USA leads in sales, followed by France and Spain.

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

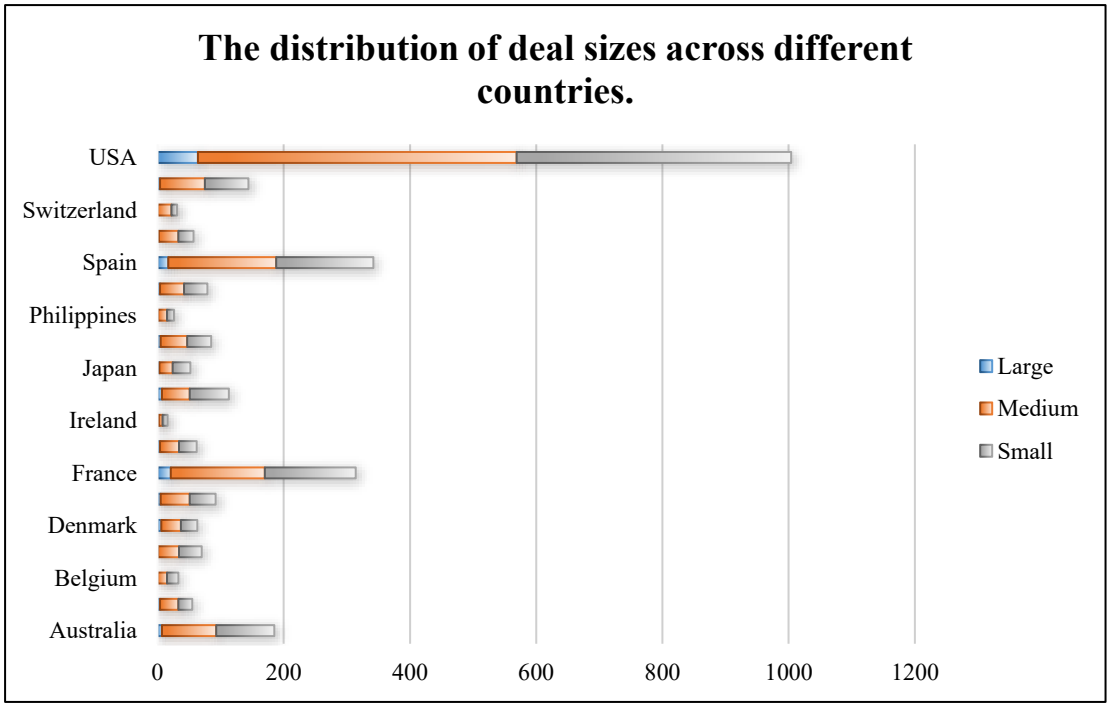
Ans: For the comparison of sales between the years 2004 and 2005, The line chart illustrates dynamic shifts in sales over the years, with Classic cars consistently leading in both 2004 and 2005, totaling $1,762,257.09 and $672,573.28, respectively. The chart visually depicts sales fluctuations across categories, highlighting Classic cars' consistent dominance throughout the analyzed period.



	Classic Cars	Motorcycles	Planes	Ships	Trains	Trucks and Buses	Vintage Cars
2004	1762257.09	560545.23	502671.8	341437.97	116523.85	529302.89	911423.77
2005	672573.28	234947.53	200074.17	128178.07	36917.33	178057.02	340739.31

5. Compare all the countries based on deal size.

Ans: To compare all countries based on deal size, we'll group the data by country to consolidate transactions from each location. Then, we'll count the occurrences of each deal size category (small, medium, large) for each country. This approach will provide insights into the distribution of deal sizes across different regions, enabling us to compare them effectively. By analyzing the frequency of deal size categories, we aim to identify any significant variations or patterns between countries. This structured approach will support informed decision-making and strategic planning efforts.



The analysis investigates deal size distribution across countries. The bar chart reveals the USA's dominance, with significantly larger deal sizes compared to other countries:64 large,505 medium, and 435 small deals

Conclusion and Review

In this report, we conducted a thorough analysis of the sales dataset to gain insights into customer behavior, product performance, and market dynamics. Through the examination of various key metrics and patterns, we aimed to derive actionable insights to support strategic decision-making.

The analysis revealed several noteworthy findings:

- Product Category Performance: We observed variations in sales performance across different product categories. Notably, "Classic Cars" and "Vintage Cars" emerged as the top-performing categories in terms of sales revenue, suggesting a strong demand for vintage and classic automobile models.
- Yearly Sales Comparison: Comparing sales between the years 2004 and 2005 revealed fluctuations in sales performance over time. Further exploration into the factors driving these variations could provide insights into market dynamics and consumerpreferences.
- Profitability by Country: Analysis of sales data across different countries highlighted variations in profitability. The USA emerged as the top-performing country in terms of sales revenue, followed by France and Spain. Understanding the factors contributing to the success of these markets could inform expansion strategies and resource allocation.
- Deal Size Distribution: Examination of deal sizes across countries revealed insights into transactional patterns. The USA exhibited the highest frequency of large and medium-sized deals, indicating potential opportunities for high-value transactions in the market.
- Customer Demographics: Demographic analysis uncovered insights into customer segmentation and preferences. Further exploration into customer demographics could enable targeted marketing strategies and personalized offerings to enhance customer satisfaction and loyalty.

The analysis provided valuable insights into customer behavior, product performance, and market trends. To capitalize on these insights, we recommend:

- Further exploration of customer segmentation to tailor marketing strategies.
- Continuous monitoring of sales trends to adapt strategies accordingly.

- Targeted efforts to leverage high-performing product categories and markets for revenue growth.

Overall, the analysis sets the foundation for data-driven decision-making and strategic planning, enabling the organization to navigate market complexities and drive sustainable growth in the sales domain.

Regression

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.877178
R Square	0.769441
Adjusted R Square	0.766629
Standard Error	896.6688
Observations	250

ANOVA

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

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

Anova (One factor)

Anova: Single Factor						
SUMMARY						
Groups	*Count*					
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 factors)

Anova: Two-Factor Without Replication				
SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	3	4097.66	1365.887	5069957
Row 2	3	2451.12	817.04	1725170
Row 3	3	1566	522	648687
Row 4	3	5095.24	1698.413	7507173
Row 5	3	5140.39	1713.463	7650609
Row 248	3	4386.35	1462.117	5944534
Row 249	3	2261.6	753.8667	1546167
Row 250	3	4176.72	1392.24	5420980
Column 1	250	903280.9	3613.123	3445221
Column 2	250	25534	102.136	1664.552
Column 3	250	8659	34.636	89.69428
ANOVA				

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

Descriptive Statistics

	SALES
Mean	3553.889072
Standard Error	34.66589212
Median	3184.8
Mode	3003
Standard Deviation	1841.865106
Sample Variance	3392467.068
Kurtosis	1.792676469
Skewness	1.161076001
Range	13600.67
Minimum	482.13
Maximum	14082.8
Sum	10032628.85
Count	2823

Correlation

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

Car Collection Data Report

Introduction

The Car Collection dataset provides detailed information on various car models, including their make, model, color, mileage, price, and cost. This report analyzes the dataset to derive insights for better decision-making in car purchases and understanding market trends. The dataset includes six different car models: Honda, Chevrolet, Nissan, Toyota, Dodge, and Ford.

This report targets car enthusiasts, industry professionals, analysts, and those interested in car market trends. It includes detailed dataset analysis, statistical evaluations, visualizations, and interpretation of results.

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

Questionnaire

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

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

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

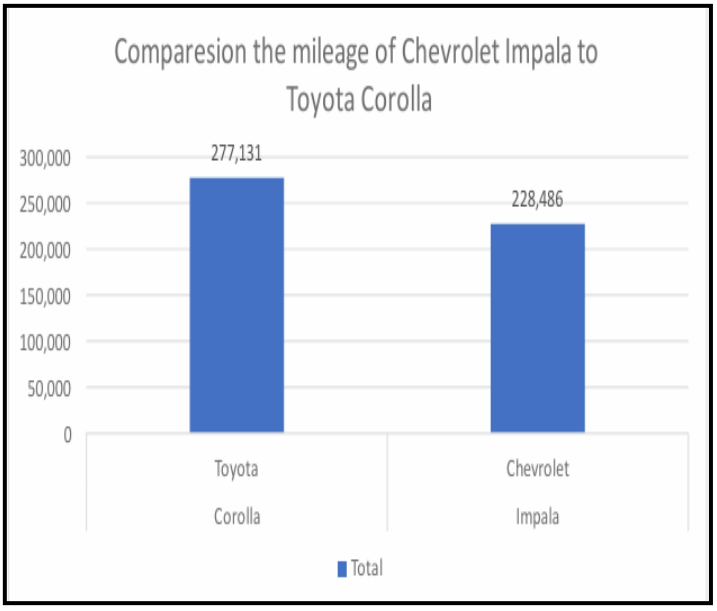
4. Compare all the cars which are of silver color to the green color in terms of Mileage.

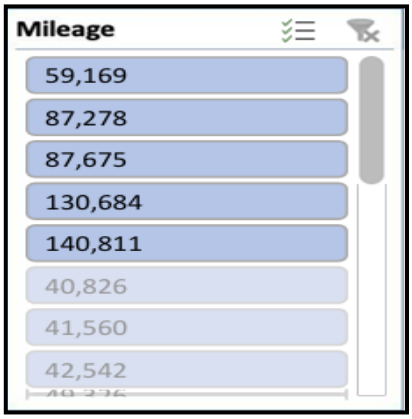
5. Find out all the cars, and their total cost which is more than $2000?

Analytics

Q1. Compare the mileage of Chevrolet Impala to Toyota Corolla. Which of the two is giving best mileage?

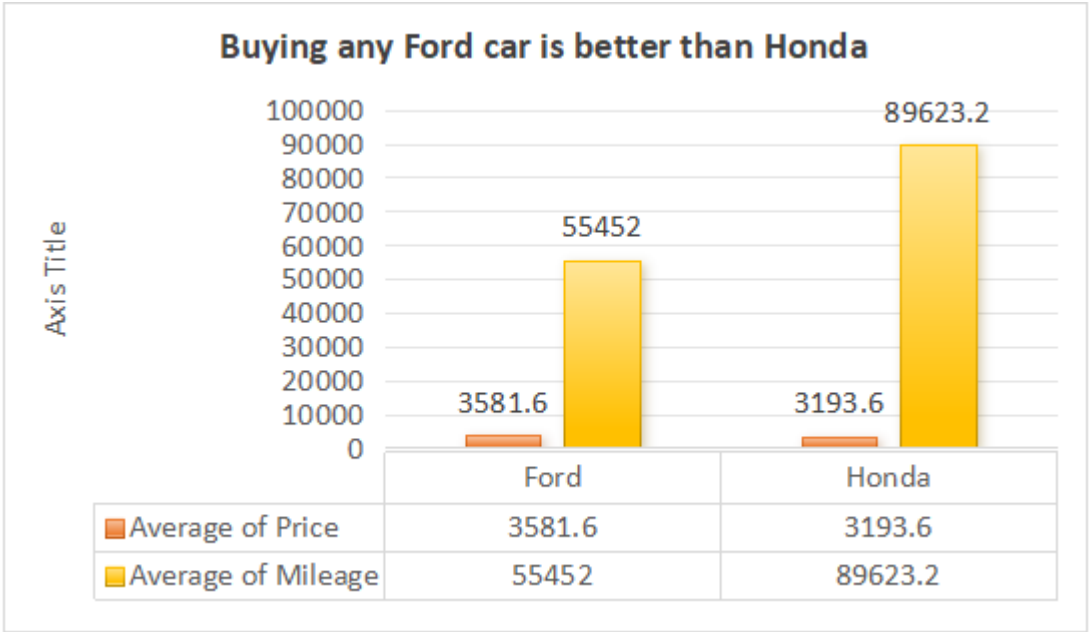
ANS: This analysis examines the fuel efficiency (mileage) of two popular car models, the Chevrolet Impala and the Toyota Corolla. To perform this comparison, the dataset was filtered to isolate relevant data, and a column chart was created to visualize the findings. Based on this analysis, it was concluded that the Chevrolet Impala, with a mileage of 114,243 miles, offers better fuel efficiency compared to the Toyota Corolla, which has a mileage of 92,377 miles. This insight helps potential buyers and industry analysts make informed decisions regarding these two car models' fuel performances.

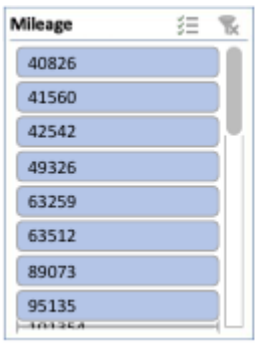


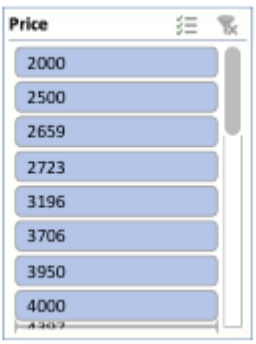


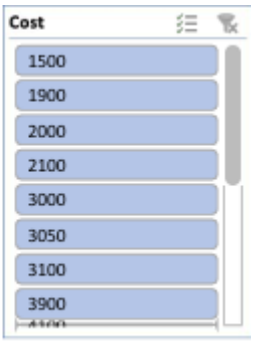
Q2. Justify, Buying of any Ford car is better than Honda.

ANS: This analysis compares Ford and Honda cars, focusing on price. Contrary to the initial statement, Honda cars have better average mileage (89,623.3) and a lower average price ($3,193.6) than Ford cars. Honda offers higher mileage at a lower cost. However, if low mileage and cost are prioritized, Ford is the better choice. Thus, the decision hinges on whether the buyer values higher mileage or lower cost.



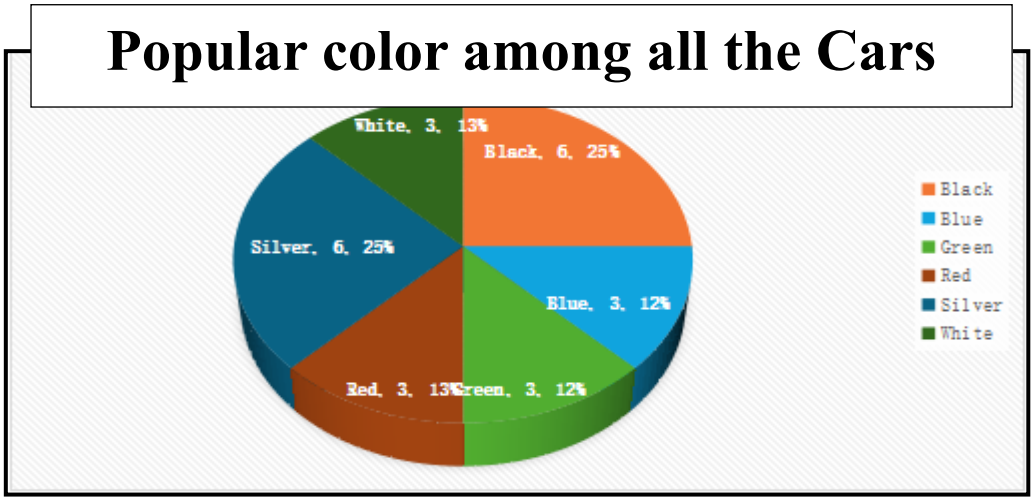






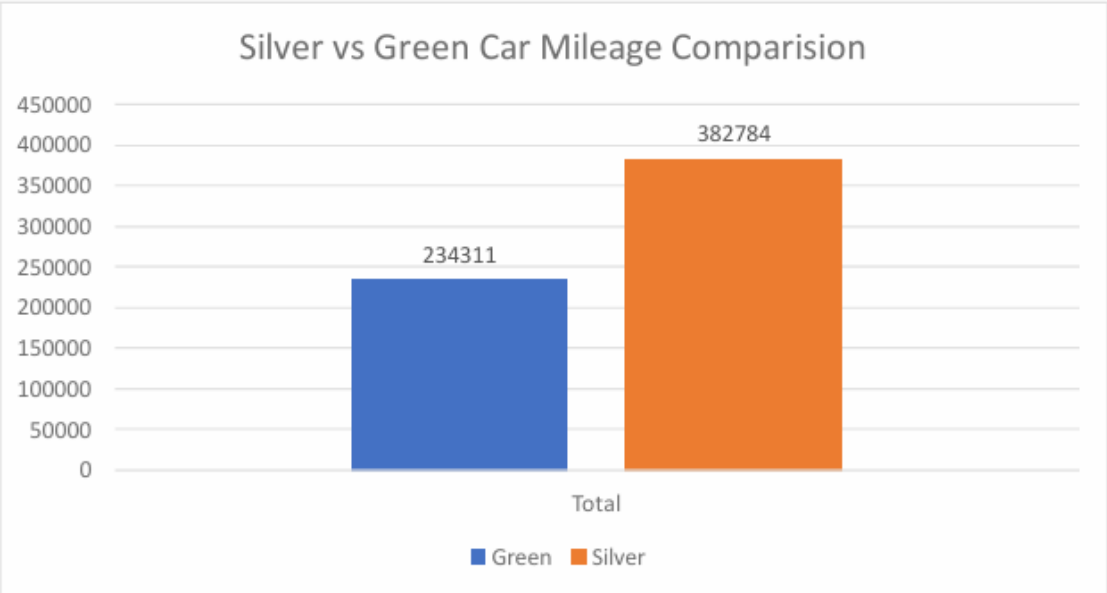
3. Among all the cars which car color is the most popular and is least popular?

ANS: This analysis identifies the most and least popular car colors in the dataset based on make counts. The results show that Black and White are the most popular colors, each accounting for 25% of the cars made by the company. Green and Blue are the least popular, each with 12% of the make. Additionally, Silver and Black are the top colors, each appearing six times, while Blue, Green, Red, and White are the least common, each appearing three times.



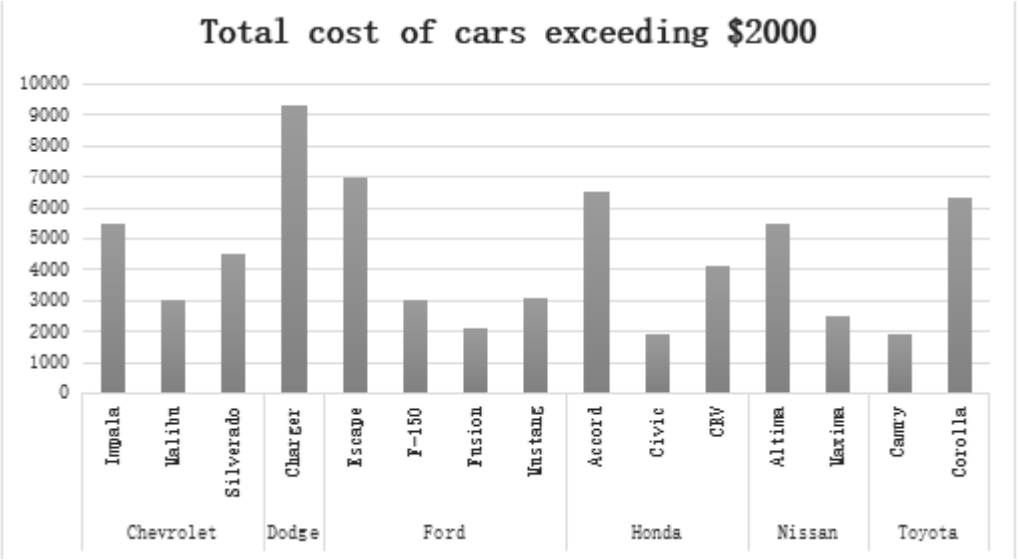
Q4. Compare all the cars which are of silver color to the green color in terms of Mileage.

ANS: This analysis compares silver and green cars by mileage. There are five silver cars (Mustang, Impala, Corolla, Charger, Accord) with Accord having the highest mileage at 101,354. Among the two green cars (Silverado, Altima), Silverado has the highest mileage at 109,23



Q5. Find out all the cars, and their total cost which is more than $2000?

ANS: This analysis aims to find out the car's costs more than $2000. And by using bar graph and taking value as sum of cost it shows the desired result. The total grand cost of all cars exceeding $2000 is $66150.



Conclusion and Review

Mileage Comparison: The analysis revealed that the Chevrolet Impala offers better fuel efficiency compared to the Toyota Corolla.

Ford vs. Honda: Contrary to the initial assumption, Honda cars were found to have better average mileage and price compared to Ford cars, disproving the claim that Ford cars are superior.

Popular Car Colors: Black and White were identified as the most popular car colors, each comprising 25% of the production, while Green and Blue were the least popular, each at 12%.

Silver vs. Green Cars: Among silver cars, the Accord had the highest average mileage. The Silverado had the highest mileage among green cars.

Cost Analysis: The total cost of cars priced over $2000 was $66,150.

The analysis provided valuable insights into mileage, car color popularity, and costs. It highlighted discrepancies, particularly in the Ford vs. Honda comparison. The use of visualizations like column charts and bar graphs effectively presented findings. This report is valuable for car buyers, industry professionals, and researchers, but further exploration into other factors influencing car purchases is recommended.

Regression

Regression analysis examined the dataset, using mileage as the dependent variable and cost and price as independent variables, to determine the statistical relationship between mileage, cost, and price.
Regression Analysis Findings

The regression analysis shows a moderate positive relationship between the predictor and response variables, with a correlation coefficient of 0.40. The model explains 16% of the variance ($R^2 = 0.16$). Each unit increase in the predictor results in a 16.66 decrease in the response variable, with a marginally significant p-value of 0.056.

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

Correlation

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

Anova means the Analysis of variance.

The Anova one factor shows the summary of columns having count, sum, average, variance. And the source of variance with ss and df. For total of three columns mileage, price and cost the count for column1, column2, and column3 is shown below.

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				

Descriptive Statistics

The descriptive statistics provide insights into three variables: Mileage, Price, and Cost. The Mileage variable ranges from 34,853 to 140,811 miles, with an average of approximately 83,803 miles. Price ranges from $2,000 to $4,959, and Cost ranges from $1,500 to $4,500. The means and standard deviations offer insights into the central tendencies and variability within each variable. These statistics offer a comprehensive overview of the dataset, enhancing our understanding of the data distribution and characteristics.

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

Understanding Sales: Orders, Regions, and Segments

Introduction

The dataset comprises crucial information regarding our orders, encompassing details such as Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, and Sales. With a focus on understanding regional variations and segment-specific trends, this dataset serves as a valuable resource for strategic decision-making and market optimization.

Questionnaire

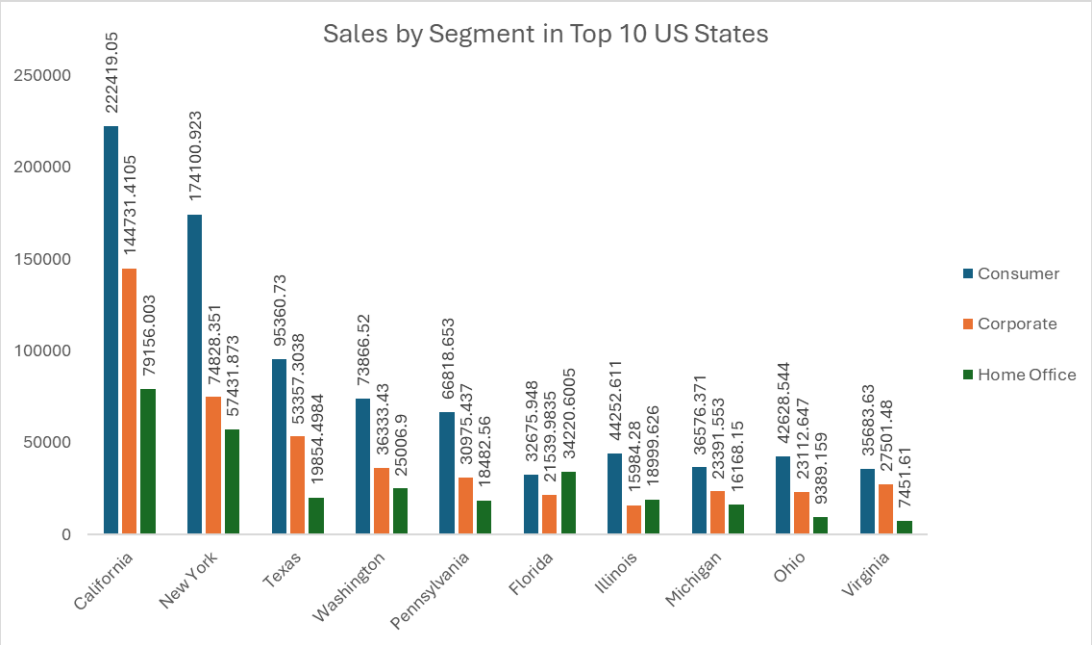
1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?
2. Find out top performing category in all the states?
3. Which segment has most sales in US, California, Texas, and Washington?
4. Compare total and average sales for all different segments?
5.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

Analytics

1. Compare all the US states in terms of Segment and Sales. Which Segment performedwell in all the states:

1.1. **Methodology:** To address this question, we first filter the dataset to include only records from the United States. Then, we calculate the total sales for each segment in each state. Next, we identify the top 10 states with the highest total sales and determine the segment with the highest sales in each of these states. By focusing on the top-performing states, we can discern trends in segment performance across different regions of the country.

1.2. Findings:

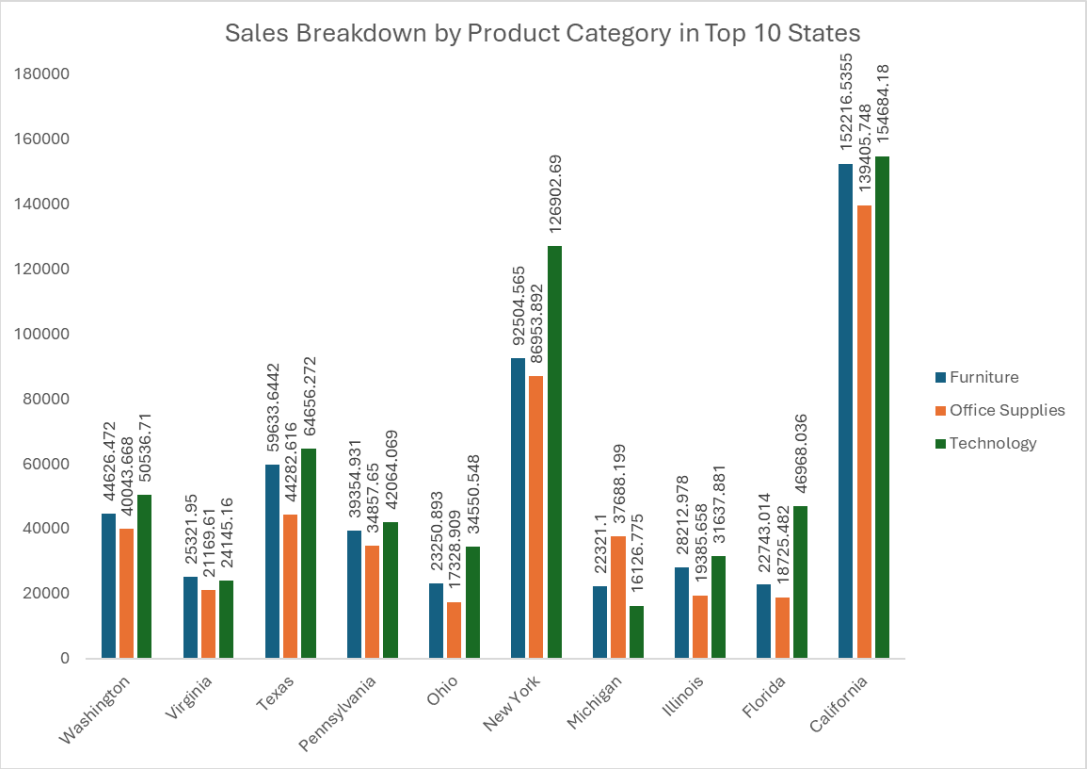


Upon analyzing the sales data for the top 10 US states, it is evident that the Consumer segment has the highest sales across all states. California leads in total sales across all segments, with Consumer segment contributing the most to its sales revenue. Similarly, in New York and Texas, the Consumer segment also dominates in terms of sales. This trend is consistent across most of the top states, indicating a strong preference for consumer goods in these regions. Corporate and Home Office segments also contribute to the overall sales revenue but are generally lower compared to the Consumer segment. These insights underscore the importance of understanding segment-specific preferences and tailoring marketing strategies to effectively target different customer segments for enhanced sales performance.

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

2.1. **Methodology:** Similar to the previous question, we filter the dataset to include only records from the United States and calculate the total sales for each category in each state. We then identify the top 10 states with the highest total sales and determine the category with the highest sales in each of these states. This analysis provides insights into the most popular categories of products across different states, aiding in strategic decision-making and marketing efforts.

2.2. Findings:

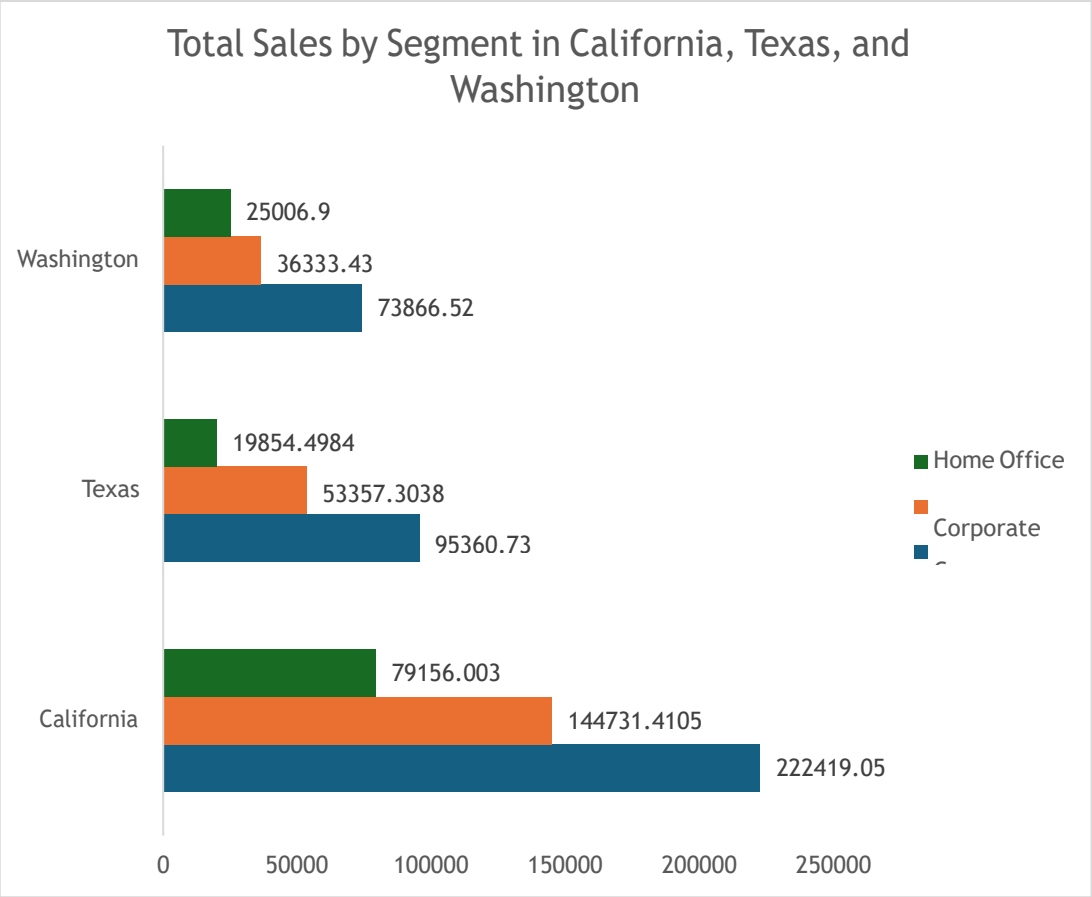


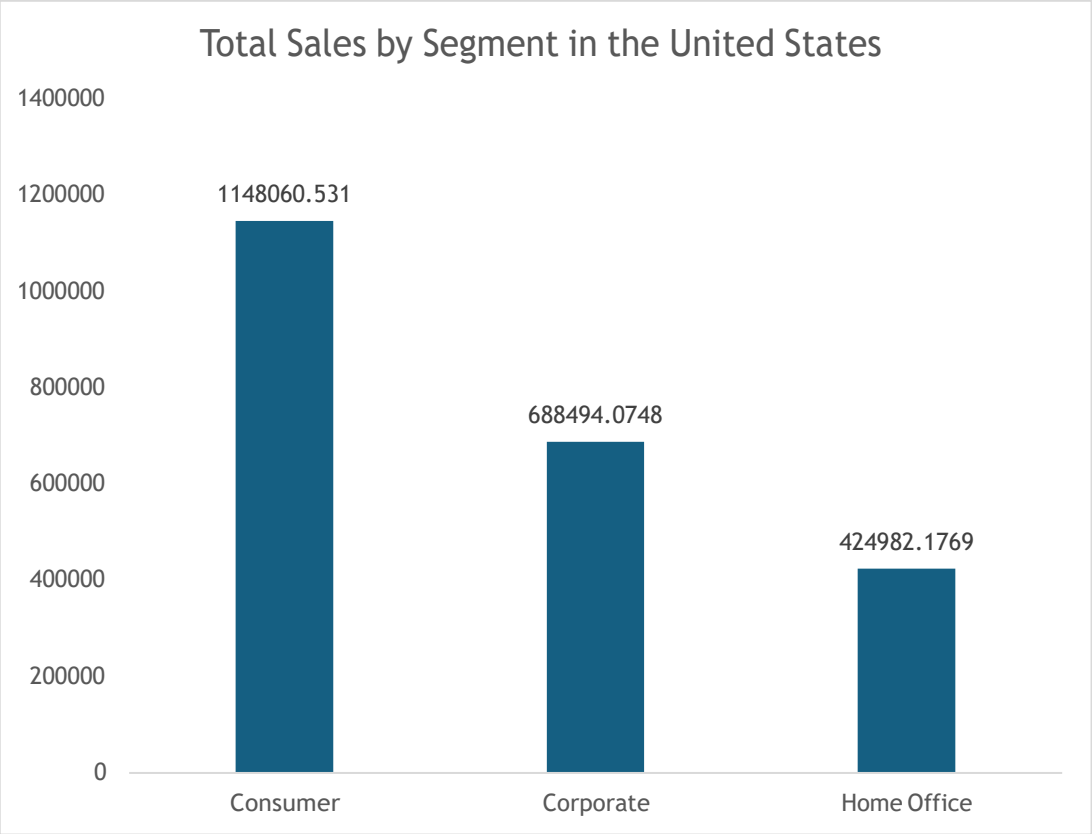
Upon analysing the sales breakdown by product category in the top 10 states, several notable insights emerge. In Washington, California, and New York, the Technology category stands out as the highest-selling category, contributing significantly to the overall sales figures. Meanwhile, in states like Virginia, Pennsylvania, and Ohio, the Furniture category appears to have the highest sales, indicating variations in consumer preferences across different regions. In Texas, Florida, and Illinois, there's a more balanced distribution of sales across all three categories, with Technology often leading but not by a significant margin. Overall, the Technology category emerges as a consistent performer across most states, contributing substantially to the total sales figures. These insights can inform strategic decision-making regarding product assortment, marketing strategies, and inventory management tailored to the preferences of customers in each state.

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

3.1. **Methodology:** For this question, we focus on specific states - California, Texas, and Washington - along with the overall United States. We filter the dataset accordingly and calculate the total sales for each segment in each of these states. By comparing the sales figures across segments, we can identify which segment dominates in terms of sales volume in each state and across the entire US.

3.2. Findings:


Total Sales by Segment in California, Texas, and Washington
Washington
25006.9
36333.43
73866.52
Texas
19854.4984
53357.3038
95360.73
California
79156.003
144731.4105
222419.05
Home Office
Corporate
0
50000
100000
150000
200000
250000

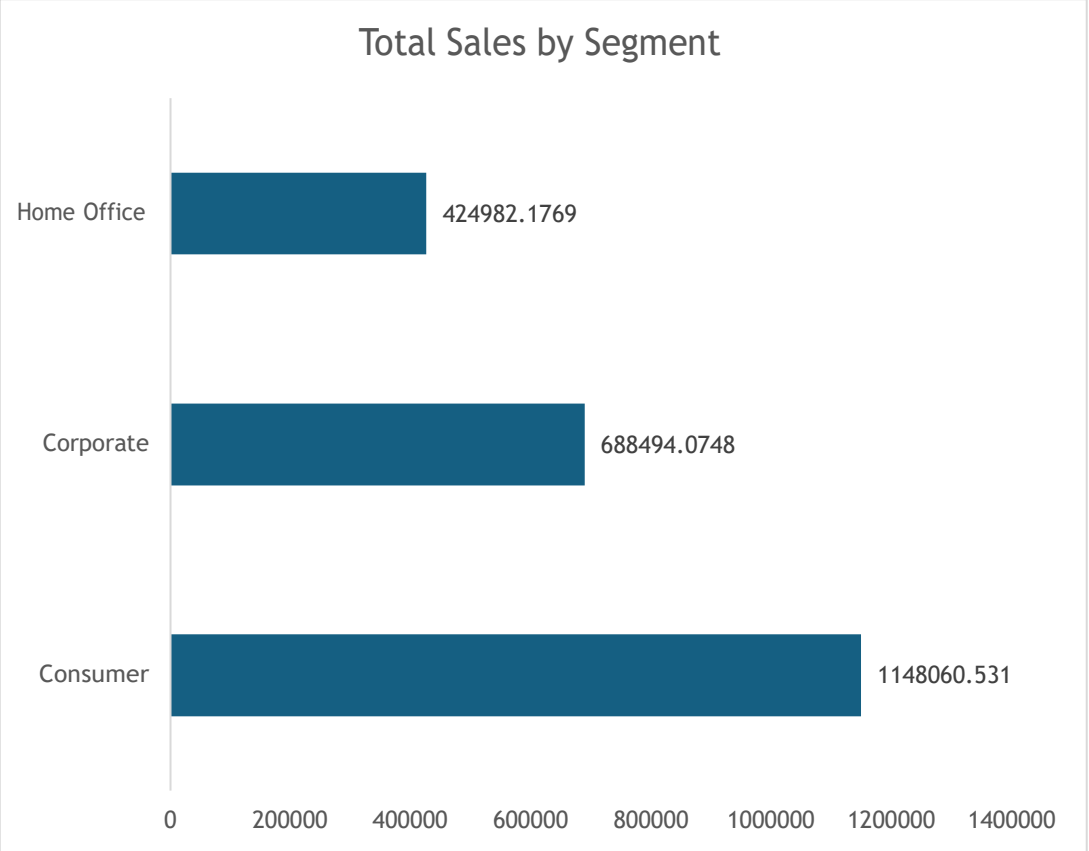

Total Sales by Segment in the United States
1400000
1200000
1000000
800000
600000
400000
200000
0
1148060.531
688494.0748
424982.1769
Consumer
Corporate
Home Office

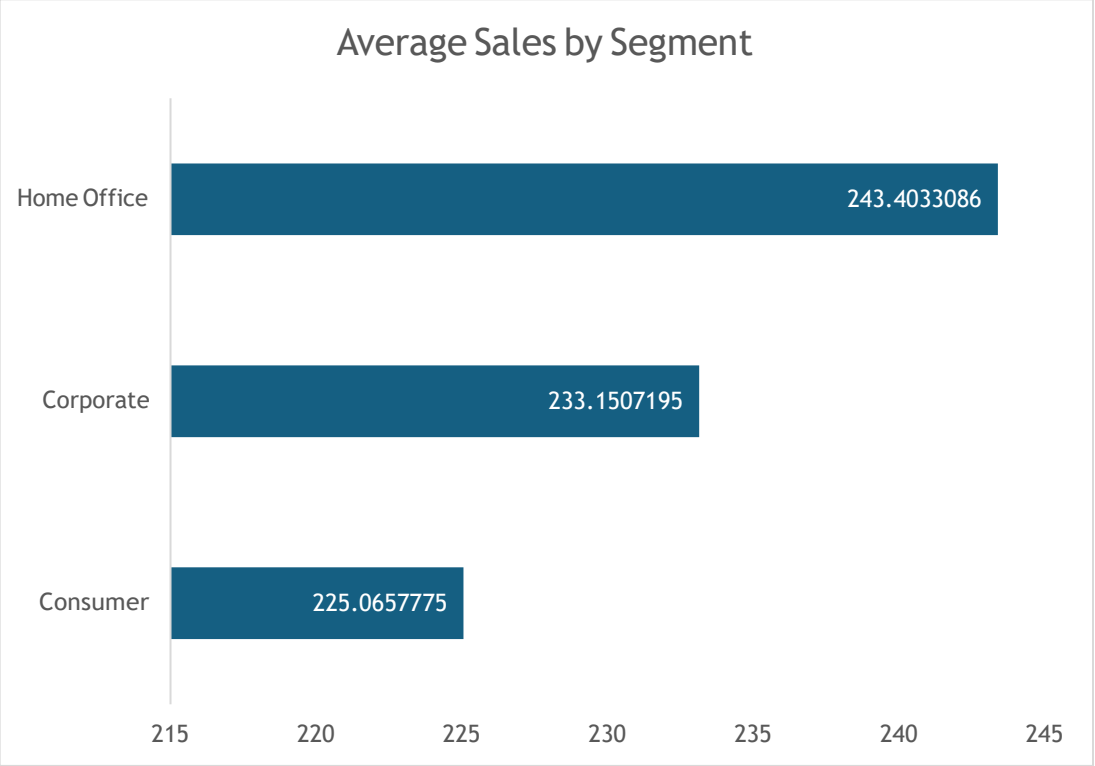
The analysis of total sales by segment in California, Texas, and Washington illustrates consistent trends across the three states. In all three states, the consumer segment dominates sales, with California leading in total sales followed by Texas and Washington. Specifically, the consumer segment exhibits the highest sales figures, followed by corporate and home office segments in each state. This trend is mirrored at the national level, where the consumer segment also emerges as the top performer, contributing the highest total sales across the United States. These findings underscore the significance of the consumer segment in driving sales, both regionally and nationally, highlighting its crucial role in shaping market dynamics and informing strategic decision-making processes.

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

4.1. **Methodology:** The comparison of total and average sales for different segments involves calculating the total sales by summing up the sales values across all segments and determining which segment has the highest overall sales. Additionally, we calculate the average sales by dividing the total sales by the number of segments to identify the segment with the highest average sales. Analyzing these figures provides insights into the relative performance of each segment, highlighting any significant differences or trends. This comparison aids in strategic decision-making and market optimization by identifying areas of strength and potential areas for improvement within each segment.

4.2. Findings:

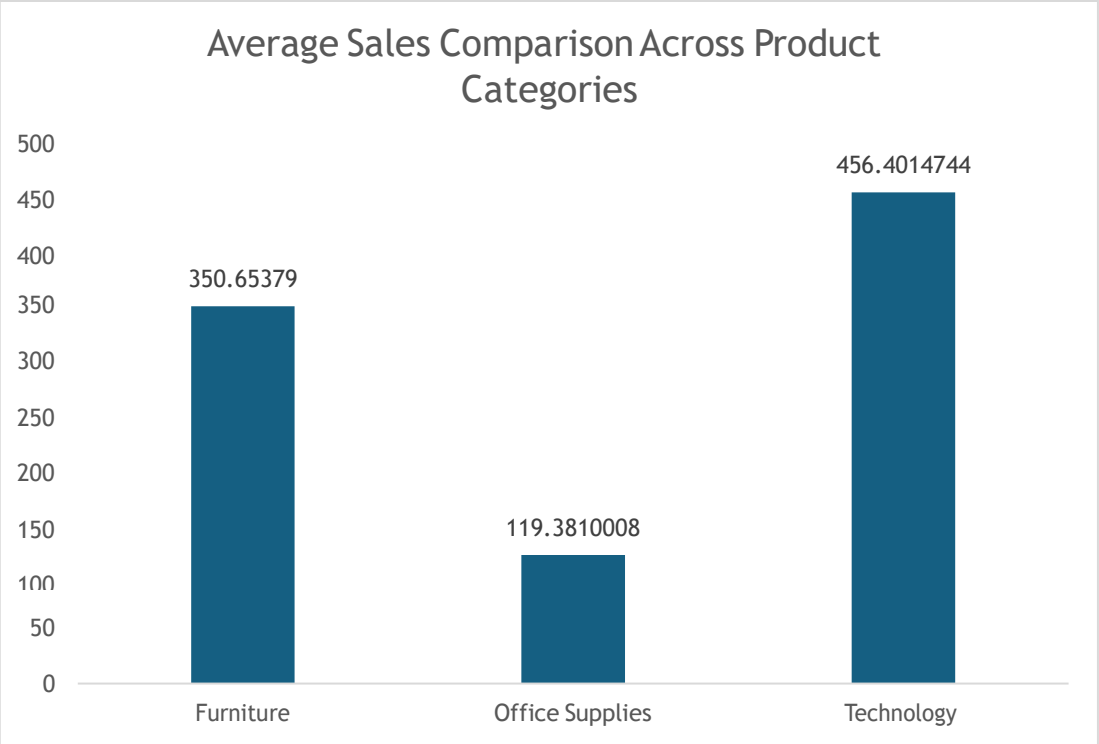


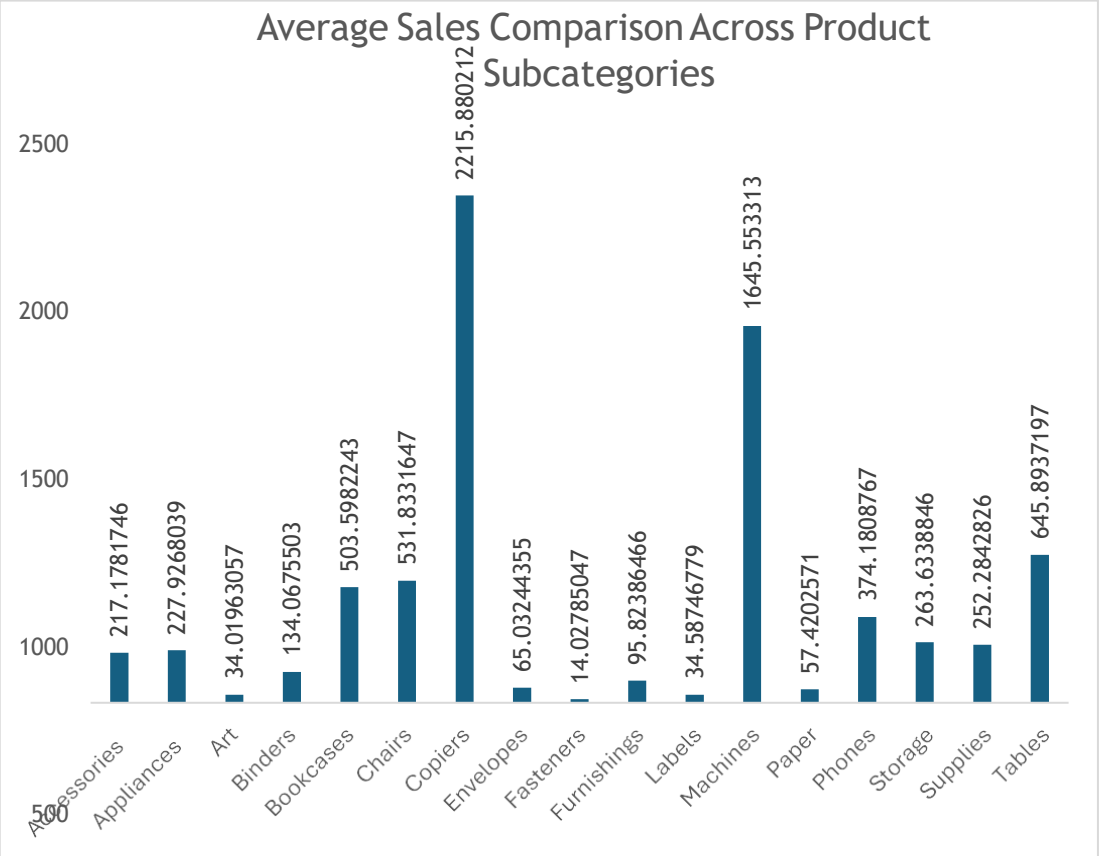


The analysis of total sales by segment reveals that the consumer segment leads with a total sales figure of $1,148,060.53, followed by corporate ($688,494.07) and home office ($424,982.18) segments. On the other hand, when considering average sales per segment, the home office segment emerges with the highest average sales of $243.40, followed by corporate ($233.15) and consumer ($225.07) segments. These findings indicate that while the consumer segment generates the highest total sales, the home office segment achieves the highest average sales per transaction. Such insights can inform strategic decisions aimed at optimizing sales strategies and resource allocation to maximize profitability across different segments.

5. **Compare the average sales of different categories and subcategories in all states.**
   5.1. **Methodology:** In this analysis, we calculate the average sales for each category and subcategory across all states. By comparing the average sales figures, we can identify which categories and subcategories tend to perform better on average. This information is valuable for understanding consumer preferences and market trends, enabling targeted marketing strategies and product development efforts.

5.2. Findings:





Upon examining the average sales across different product categories, we observe significant variations in performance. The Technology category emerges as the top

performer, with an average sales figure of $456.40, indicating strong demand for technological products among customers. Following closely behind is the Furniture category, with an average sale of $350.65, suggesting a steady market for home and office furniture. Meanwhile, the Office Supplies category lags with an averagesale of $119.38, indicating comparatively lower demand for office supplies products. Analysing the average sales across various product subcategories reveals significantdisparities in performance. The highest average sales are observed in categories suchas Copiers ($2215.88), Machines ($1645.55), and Tables ($645.89), indicating strongdemand for these high-value products. In contrast, categories like Fasteners ($14.03)and Art ($34.02) exhibit much lower average sales, suggesting relatively lowerconsumer interest or niche market segments. Understanding these differences inaverage sales can inform product pricing strategies, inventory management, andmarketing efforts to optimize revenue generation and meet customer demandseffectively.

6. **Find out the state-wise mode for Customer and Segment in California, Illinois, New York, Texas, and Washington.**
   6.1. **Methodology:** Focusing on the specified states - California, Illinois, New York, Texas, and Washington - we determine the mode (most frequent value) for both Customer and Segment. This analysis provides insights into the predominant customer and segment types in each state, facilitating personalized marketing strategies and customer segmentation efforts tailored to the preferences of each region.

   6.2. **Findings**: Upon analyzing the dataset for state-wise mode in both Customer and Segment categories, several noteworthy findings emerged. In California, the most common customer is William Brown, with the prevailing segment being Consumer. Similarly, in Illinois, the dominant customer is identified as Rob Lucas, and the primary segment remains Consumer. Moving on to New York, Seth Vernon emerges as the predominant customer, aligning with the prevalent segment of Consumer. In Texas, Matt Collister is the most frequent customer, accompanied by the dominant segment of Consumer. Lastly, in Washington, Dennis Kane is identified as the mode customer, with Consumer being the prevailing segment. These insights offer valuable guidance for tailoring marketing strategies and customer engagement approaches according to the predominant customer profiles and segments in each state.

Conclusion and Review

In conclusion, our analysis revealed key insights into regional sales performance, segment dynamics, and product category preferences. The dominance of the Consumer segment nationwide and the strong demand for Technology products emerged as prominent trends. By understanding these patterns, businesses can refine their strategies to target specific market segments effectively. Our findings offer actionable guidance for optimizing sales strategies and driving growth in the retail industry. Through transparent methodologies and clear presentation, our analysis provides valuable insights for informed decision-making and sustained success in a competitive marketplace.

Anova

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Regionn	9800	25201	2.571530612	1.350531385
Sales	9800	2261536.783	230.7690595	392692.5722

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	255163149.6	1	255163149.6	1299.552322	1.3384E-275	3.841933358
Within Groups	3848007749	19598	196346.9614			
Total	4103170899	19599				

The single-factor ANOVA analysis indicates a significant difference in mean sales across different regions ($F(1, 19598) = 1299.55$, $p < 0.001$). The between-groups variation, which represents differences in mean sales among regions, is substantial, with a sum of squares of approximately 255,163,149.6. This suggests that the variation in sales among regions is much larger than the variation within each region. The results imply that the region significantly influences sales, highlighting the importance of considering regional factors when analyzing sales data. This finding underscores the potential impact of geographical location on sales performance and the need to tailor marketing strategies and business operations accordingly.

Descriptive Statistics

Sales	
Mean	230.7690595
Standard Error	6.330139859
Median	54.49
Mode	12.96
Standard Deviation	626.6518748
Sample Variance	392692.5722
Kurtosis	304.4450883
Skewness	12.98348287
Range	22638.036
Minimum	0.444
Maximum	22638.48
Sum	2261536.783
Count	9800

The descriptive statistics for the sales data reveal a mean sales value of approximately $230.77, with a standard error of $6.33. The distribution exhibits high kurtosis (304.45) and skewness (12.98), indicating significant deviation from a normal distribution and a pronounced right

skew. The median sales value is $54.49, and the mode is $12.96, suggesting that lower sales values are more common. The standard deviation is relatively large at $626.65, indicating considerable variability in sales data. The range spans from a minimum sales value of $0.44 to a maximum of $22638.48, illustrating the wide dispersion of sales amounts. Overall, these statistics provide insights into the central tendency, variability, and distributional characteristics of the sales data, which can inform decision-making and strategic planning processes.

Analysis of Cookie Sales Performance Across Countries

Introduction

The dataset under consideration contains detailed information on cookie sales, including the country of sale, cookie type, units sold, revenue, cost, profit, and date of sale. This analysis aims to provide insights into the performance of cookie sales across different countries, with a focus on profitability, revenue generation, and product trends. By examining key metrics and trends, this report aims to offer actionable recommendations for optimizing sales strategies and maximizing profits in the cookie market.

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 performs 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?

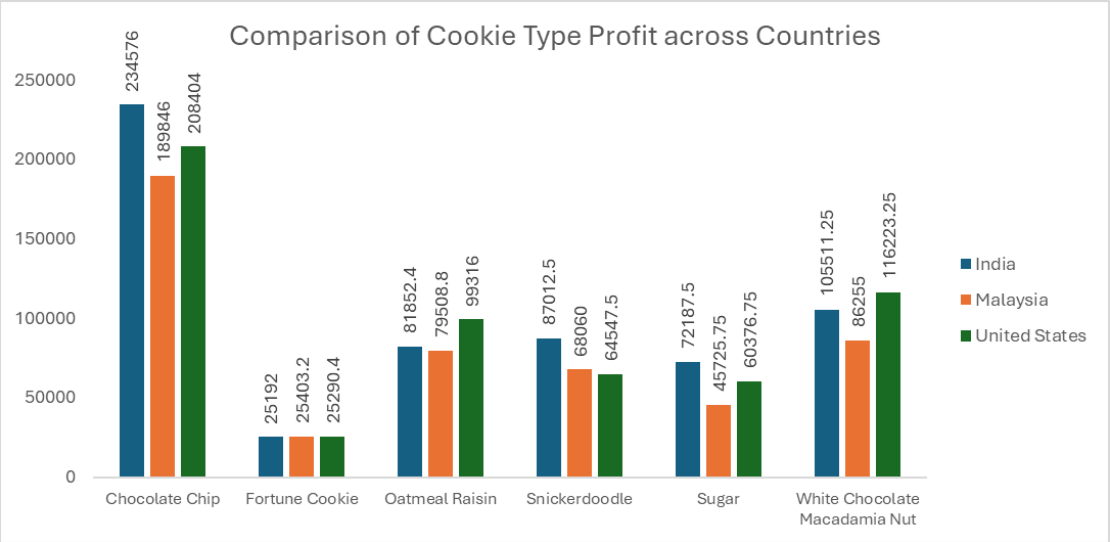
Analytics

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

1.1. Methodology:

To compare cookie profitability across the US, Malaysia, and India, we filter sales data by country and compute total profits for each cookie type. This method allows us to analyze and compare profitability trends across different cookie types in diverse market contexts, aiding strategic decision-making.

1.2. Findings:

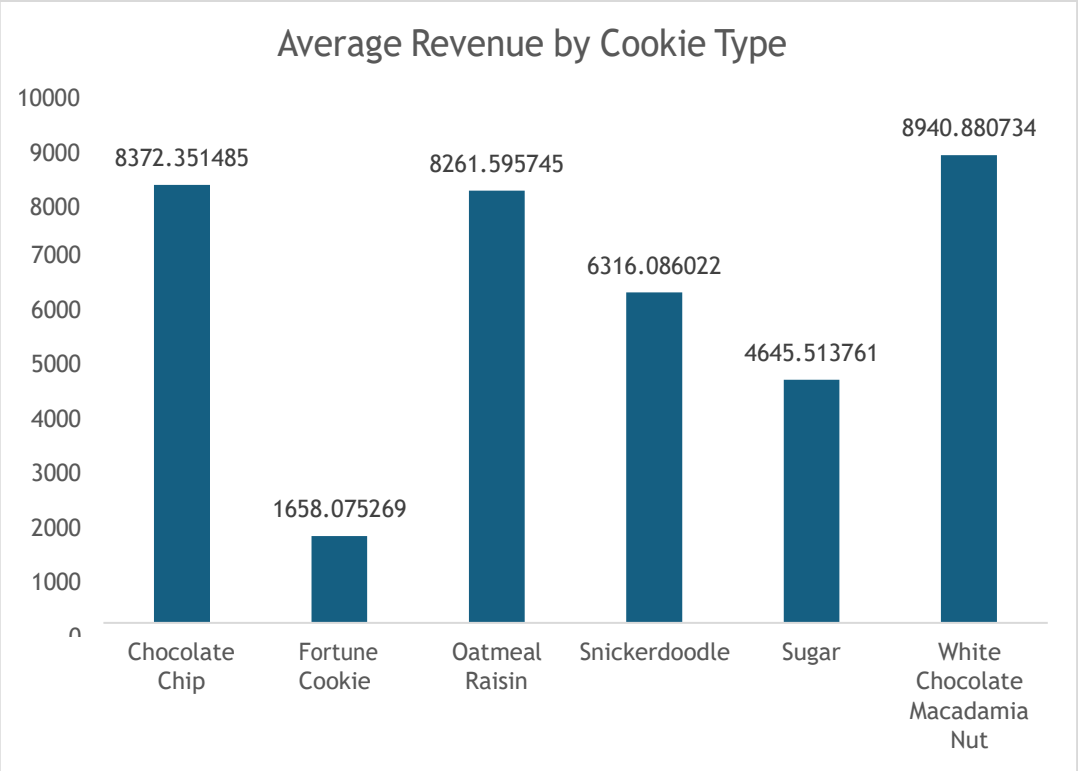


The analysis of cookie profitability across India, Malaysia, and the United States reveals interesting insights into sales performance. Among the various cookie types, Chocolate Chip cookies generate the highest total profits, with India contributing $234,576, Malaysia $189,846, and the United States $208,404, resulting in a grand total of $632,826. Other notable performers include Oatmeal Raisin and White Chocolate Macadamia Nut cookies, contributing $260,677.20 and $307,989.50 to the overall profits, respectively. While Snickerdoodle cookies also perform well in terms of total profits, earning $219,620 overall, Fortune and Sugar cookies lag behind in profitability. These findings provide valuable insights into the sales dynamics of different cookie types across the analyzed countries, aiding in strategic decision-making for market expansion and product optimization.

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

2.1. **Methodology:** To determine the average revenue generated by different types of cookies, we employ a straightforward methodology. First, we aggregate the revenue for each cookie type across all sales transactions. Then, we divide the total revenue for each cookie type by the number of units sold to compute the average revenue per unit. This analysis provides insights into the average pricing and sales performance of each cookie type, enabling us to identify which types yield higher average revenues. Additionally, comparing the average revenues across different cookie types facilitates understanding of consumer preferences and market demand for specific products.

2.2. Findings:



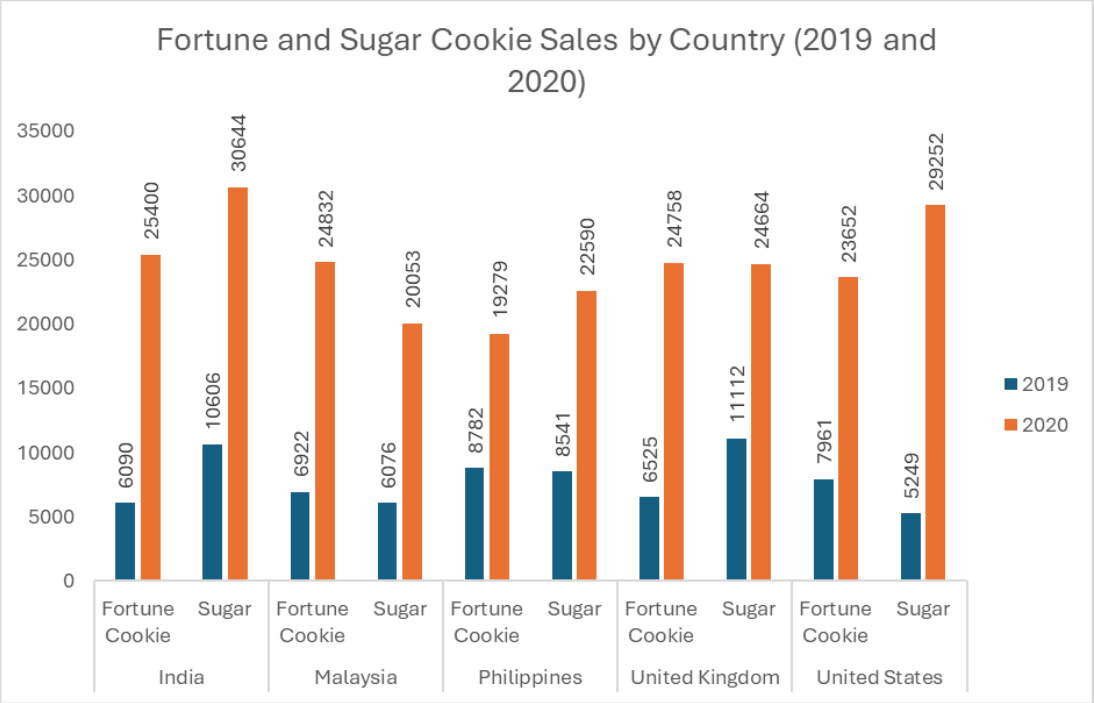
The analysis of average revenue generated by different types of cookies reveals interesting insights into their respective sales performance. Among the various cookie types, the highest average revenue is observed for the White Chocolate Macadamia Nut cookies, with an average revenue of $8,940.88. This indicates strong consumer demand and willingness to pay premium prices for this cookie variant. Following closely behind are Chocolate Chip and Oatmeal Raisin cookies, with average revenues of $8,372.35 and $8,261.60, respectively. These findings suggest that these traditional favorites also command significant market value and contribute positively to overall sales revenue. On the other hand, Fortune Cookies exhibit comparatively lower average revenue of $1,658.08, indicating potentially lower consumer interest or lower pricing for this type of cookie. Similarly, Snickerdoodle and Sugar cookies also show moderate average revenues of $6,316.09 and $4,645.51, respectively. These insights into average revenue by cookie type provide valuable guidance for pricing strategies, product development, and marketing initiatives aimed at maximizing revenue and profitability in the cookie market.

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

3.1. **Methodology:** To determine which country sold the most Fortune and sugar cookies in 2019 and 2020, we need to filter the dataset based on the cookie types (Fortune and Sugar), and then further segment the data based on the respective years (2019 and 2020). Next, we calculate the total units sold for each country within these specific categories and years. By analyzing the total units sold for Fortune and Sugar cookies in each country for both years, we can identify the country with the highest sales volume for each cookie type in each year. This analysis allows us to understand the

market trends and preferences for Fortune and Sugar cookies across different countries over the two years, providing valuable insights for strategic decision-making and targeted marketing efforts.

3.2. Findings:

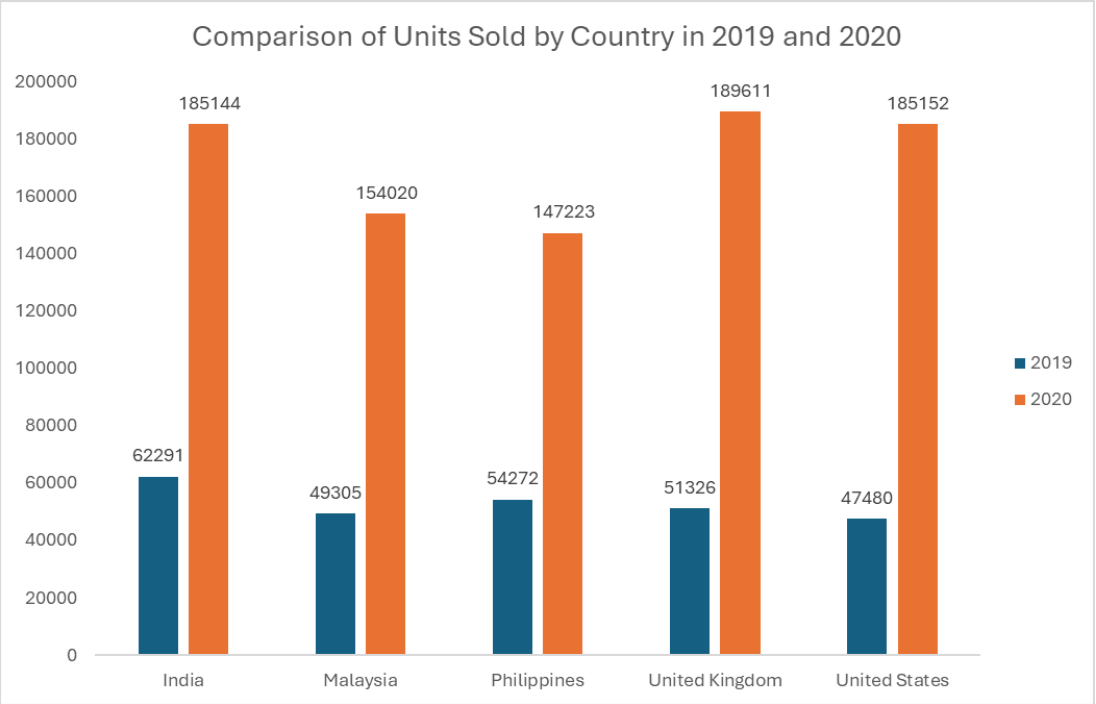


In 2019, the Philippines emerged as the top seller of Fortune cookies, while the United Kingdom led in Sugar cookie sales. This indicates distinct consumer preferences in different regions, with the Philippines showing a preference for Fortune cookies and the UK favoring Sugar cookies. However, in 2020, India surpassed all other countries, becoming the highest seller of both Fortune and Sugar cookies. This significant shift suggests evolving market dynamics and potentially changing consumer tastes, highlighting India's growing prominence in the cookie market. Overall, these findings underscore the importance of monitoring market trends and adapting strategies to capitalize on emerging opportunities in the confectionery industry.

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

4.1. **Methodology:** To compare the performance of all countries for the years 2019 and 2020, we first filter the dataset to include records from these specific years. Then, we calculate the total units sold for each country in both 2019 and 2020. Next, we analyze the data to identify any fluctuations or trends in sales performance across the two years for each country. This comparison allows us to assess the relative performance of each country over time and identify any notable changes or patterns in sales dynamics. By examining year-on-year variations in sales, we can gain insights into factors influencing market demand and tailor strategic decisions accordingly to optimize sales performance.

4.2. Findings:



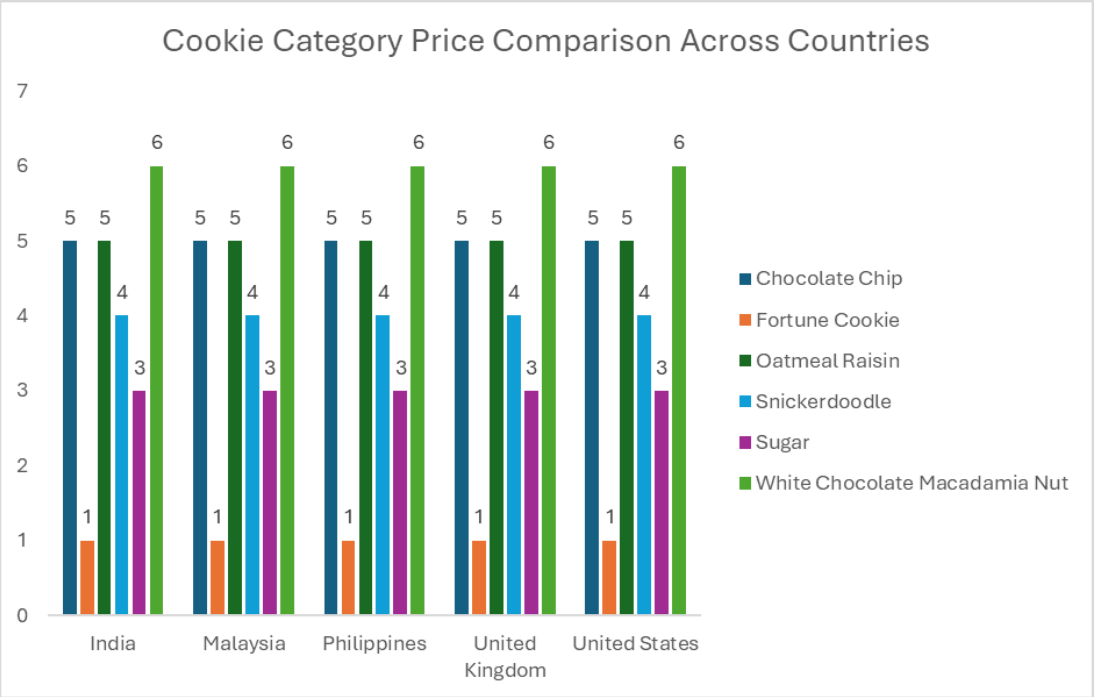
In examining the units sold by each country in 2019 and 2020, notable trends emerge. India demonstrated a substantial increase in units sold from 2019 to 2020, with a rise from 62,291 units to 185,144 units, marking a significant surge in sales. Similarly, Malaysia experienced a considerable uptick in units sold, climbing from 49,305 units in 2019 to 154,020 units in 2020, indicating robust growth in market demand. The Philippines and the United Kingdom also demonstrate growth in units sold, though not as pronounced as India and Malaysia. Conversely, the United States shows a relatively modest increase in units sold from 2019 to 2020. Overall, the data underscores the dynamic nature of sales performance across different countries, with India and Malaysia notably leading in terms of growth in units sold over the two years.

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

5.1. Methodology:

To find the cookie category with the highest price in each country and the overall profit earned by that category, we calculate the average price per unit sold for each category in every country. Then, we determine the category with the highest average price in each country and compute the total profit generated by that category globally. This approach enables us to understand pricing variations across countries and evaluate the overall profitability of each cookie category.

5.2. Findings:



After analyzing the data, it is evident that the White Chocolate Macadamia Nut cookie category is sold at the highest price across all countries, priced at $6 per unit. Despite its higher price point, it has generated the highest overall profit compared to other cookie categories. The total profit earned from selling the White Chocolate Macadamia Nut cookies amounts to $527,884.50. This indicates not only the popularity of this cookie category but also its significant contribution to the overall profitability of the cookie business.

Conclusion and Review

In conclusion, the analysis of cookie sales performance across various countries provides valuable insights into market trends, consumer preferences, and profitability. Chocolate Chip cookies emerged as the top performers in terms of profitability across India, Malaysia, and the United States, while White Chocolate Macadamia Nut cookies commanded the highest average price and generated the highest overall profit globally. The comparison of Fortune and Sugar cookie sales revealed shifting consumer preferences, with India emerging as the top seller in 2020. Furthermore, the examination of sales performance from 2019 to 2020 highlighted significant growth in units sold in India and Malaysia, indicating expanding market demand in these regions. These findings underscore the importance of adapting strategies to evolving market dynamics and tailoring product offerings to meet consumer preferences effectively.

The analysis conducted in this report provides comprehensive insights into the performance of cookie sales across different countries, offering actionable recommendations for businesses operating in the confectionery industry. By examining key metrics such as profitability, revenue generation, and sales performance over time, this report enables stakeholders to make informed decisions regarding product development, pricing strategies, and market expansion efforts.

However, it is important to note that the findings presented in this report are based on the available dataset and may be subject to limitations such as data quality and sample size. Therefore, further research and analysis may be warranted to validate the findings and explore additional factors influencing cookie sales performance. Overall, this report serves as a valuable resource for businesses seeking to optimize their operations and maximize profitability in the competitive cookie market.

Regression

Regression Statistics	
Multiple R	0.829304251
R Square	0.68774554
Adjusted R Square	0.687298184
Standard Error	485.0757185
Observations	700

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	361737578.4	361737578.4	1537.356384	1.3944E-178
Residual	698	164238319.9	235298.4526		
Total	699	525975898.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	522.6687569	33.20853031	15.73899092	5.42127E-48	457.468176	587.8693377	457.468176	587.8693377
Profit	0.27501109	0.007013954	39.20913649	1.3944E-178	0.261240114	0.288782067	0.261240114	0.288782067

The regression analysis indicates a strong positive relationship between the predictor variable "Profit" and the dependent variable, with a coefficient of 0.275 ($p < 0.001$). This suggests that for every unit increase in profit, there is an expected increase of $0.275 in the dependent variable. The regression model explains approximately 68.8% of the variance in the dependent variable, as indicated by the R-squared value of 0.688. Both the multiple R and adjusted R-squared values are high, indicating a good fit of the regression model to the data. The ANOVA results are highly significant ($p < 0.001$), suggesting that the regression model is a significant predictor of the dependent variable. Overall, the regression analysis suggests that profit is a strong predictor of the dependent variable, with higher profits associated with higher values of the dependent variable.

Anova

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Country	700	2100	3	2.00286123
Revenue	700	4690319	6700.455714	21380457.98

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	15699569566	1	15699569566	1468.59044	3.0547E-220	3.84811911
Within Groups	14944941526	1398	10690229.99			
Total	30644511091	1399				

The single-factor ANOVA analysis indicates a significant difference in mean revenues across different countries ($F(1, 1398) = 1468.59$, $p < 0.001$). The between-groups variation, which represents differences in mean revenues among countries, is substantial, with a sum of squares of approximately 15,699,569,566. This suggests that the variation in revenues among countries is much larger than the variation within each country. The results imply that the country significantly influences revenue generation, highlighting the importance of considering geographical factors when analyzing revenue data. This finding underscores the potential impact of country-specific factors on revenue performance and the need to tailor business strategies to different market environments.

Descriptive Statistics

Units Sold		*Revenue*		*Cost*		*Profit*	
Mean	1608.32	Mean	6700.455714	Mean	2752.792214	Mean	3947.6635
Standard Error	32.78651936	Standard Error	174.7670203	Standard Error	76.99165581	Standard Error	98.86873961
Median	1542.5	Median	5871.5	Median	2423.6	Median	3424.5
Mode	727	Mode	8715	Mode	3450	Mode	5229
Standard Deviation	867.4497659	Standard Deviation	4623.900732	Standard Deviation	2037.007743	Standard Deviation	2615.820975
Sample Variance	752469.0963	Sample Variance	21380457.98	Sample Variance	4149400.545	Sample Variance	6842519.371
Kurtosis	-0.314907372	Kurtosis	0.464595624	Kurtosis	0.81004281	Kurtosis	0.338621291
Skewness	0.436269672	Skewness	0.867861282	Skewness	0.930442063	Skewness	0.840484415
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	1926954.55	Sum	2763364.45
Count	700	Count	700	Count	700	Count	700

The descriptive statistics provide insights into the characteristics of the variables Units Sold, Revenue, Cost, and Profit. The mean number of units sold is approximately 1608.32, with a standard deviation of 867.45. Revenue has a mean of $6700.46, with a wide range from $200 to $23988, indicating significant variability in sales amounts. The mean cost is $2752.79, while the mean profit is $3947.66, suggesting that on average, profits exceed costs. Both revenue and profit exhibit positive skewness, indicating a tendency towards higher values. The median values provide a sense of central tendency, with units sold at 1542.5 and revenue at $5871.5. Overall, these statistics provide a comprehensive overview of the sales data, including measures of central tendency, variability, and distributional characteristics, which can inform decision-making and strategic planning processes.

Correlation

	Units Sold	*Revenue*	*Cost*	*Profit*
Units Sold	1			
Revenue	0.796297786	1		
Cost	0.74260418	0.992010548	1	
Profit	0.829304251	0.995162738	0.974818454	1

The correlation analysis reveals strong positive correlations among the variables Units Sold, Revenue, Cost, and Profit. Units Sold and Revenue exhibit a correlation coefficient of approximately 0.796, indicating a significant positive relationship, while Units Sold and Profit have a correlation coefficient of approximately 0.829, suggesting a similarly strong positive association. Additionally, Revenue and Profit demonstrate a correlation coefficient of approximately 0.995, indicating a nearly perfect positive relationship. Cost is highly correlated with both Revenue and Profit, with correlation coefficients of approximately 0.992 and 0.975, respectively. These findings suggest that as Units Sold increase, there is a corresponding increase in Revenue, Cost, and Profit, highlighting the interconnectedness of these variables in the sales process.

Loan Data Report

Introduction

The loan dataset offers comprehensive insights into applicants' details like gender, marital status, education, income, loan amount, and property area. This analysis aims to uncover patterns and demographics using pivot tables and charts. Understanding these nuances is crucial for financial institutions to optimize lending processes and cater to diverse customer needs effectively. The insights derived will inform strategic decisions and enhance loan management systems' efficiency.

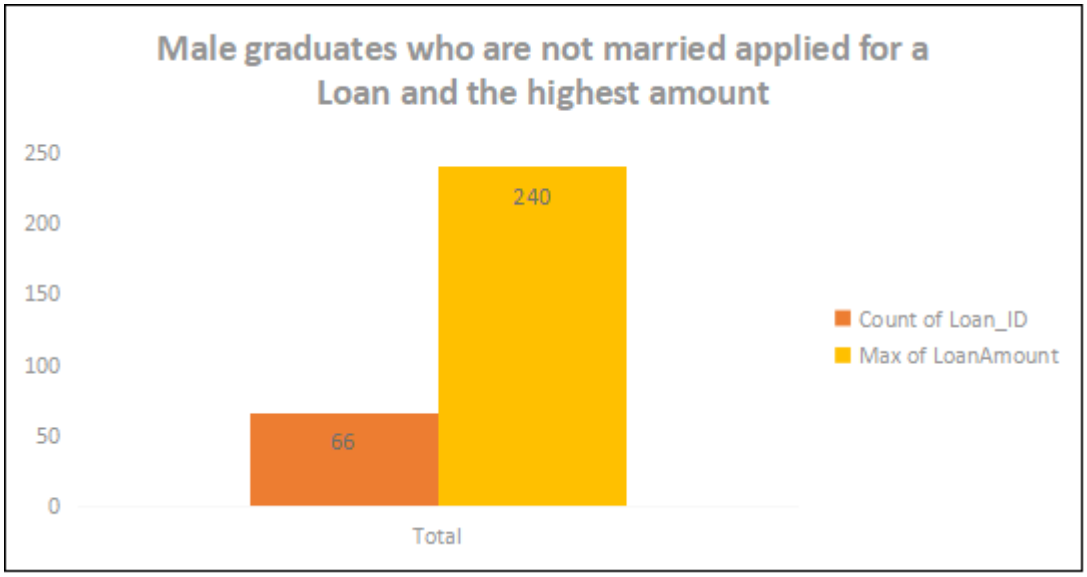
Questionnaire

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

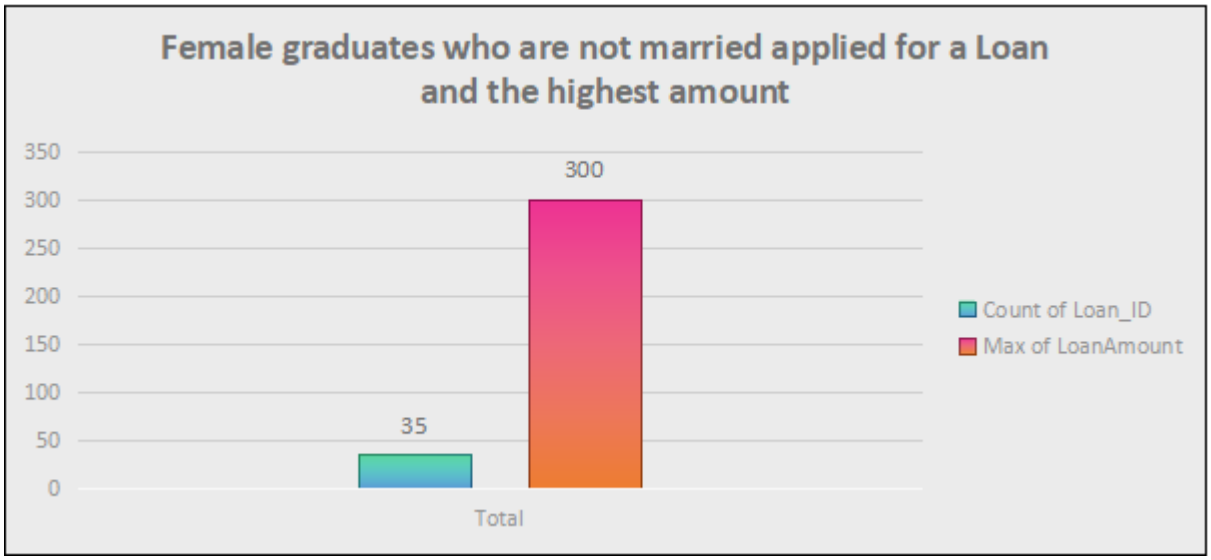
Q1. How many male graduates who are not married applied for Loan? What was the highest amount?

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



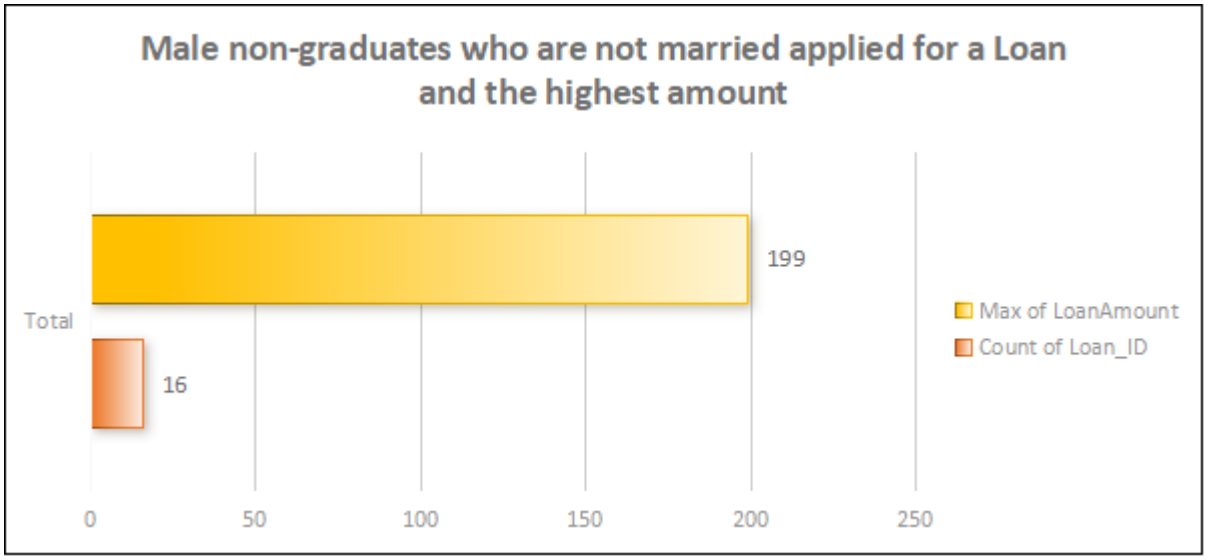
Q2. How many female graduates who are not married applied for Loan? What was the highest amount?

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



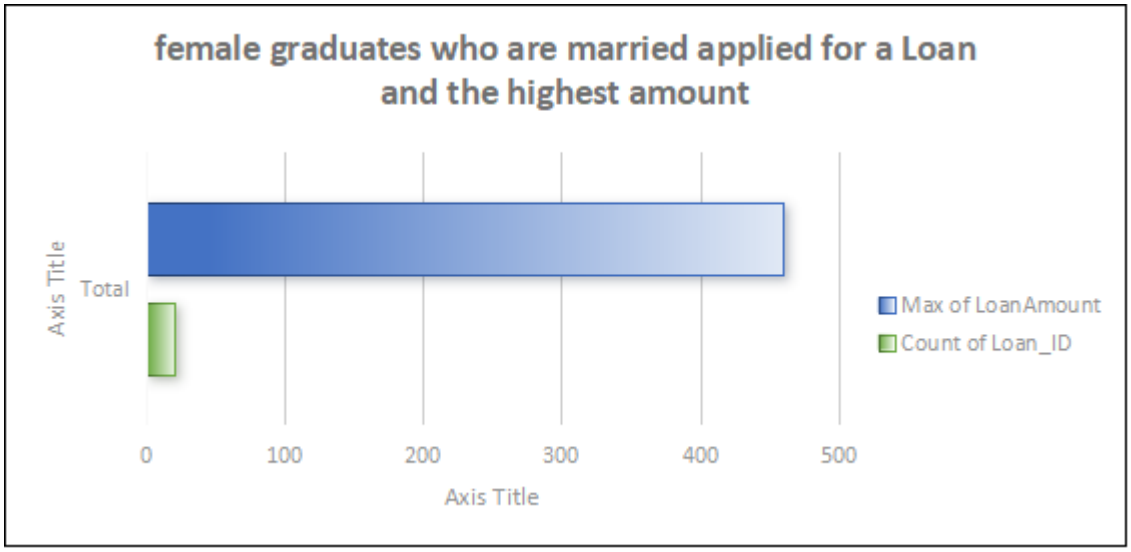
Q3. How many male non-graduates who are not married applied for Loan? What was the highest amount?

ANS: The analysis reveals that among loan applicants, the highest number consists of unmarried male non-graduates, with a total of 16 applications. The maximum loan amount requested is $199.



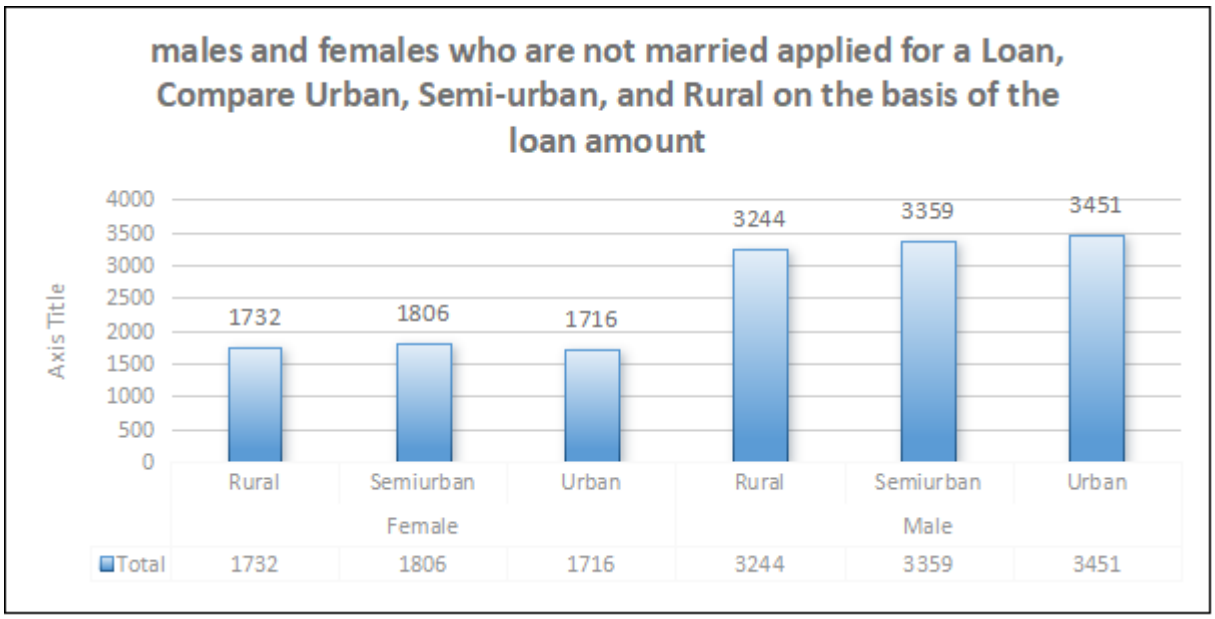
Q4. How many female graduates who are married applied for Loan? What was the highest amount?

ANS: The analysis indicates that the highest number of loan applicants are unmarried female graduates, totalling 21 applications. The maximum loan amount requested among them is $460.



Q5. How many male and female who are not married applied for Loan? Compare Urban, Semi-urban, and rural based on amount.

ANS: This analysis compares loan applications from unmarried females and males across rural, semi-urban, and urban areas. Females have lower application counts: rural (1,732), semi-urban (1,806), and urban (1,716). In contrast, males show significantly higher counts: rural (3,244), semi-urban (3,359), and urban (3,451), indicating a gender disparity in loan applications.



Conclusion and Review

The analysis reveals stark gender disparities in loan applications, with unmarried male graduates leading, followed by unmarried female graduates. Smaller numbers of unmarried male non-graduates and married female graduates also applied. Males outnumbered females across rural, semi-urban, and urban areas. The report effectively illustrates gender-based trends, offering valuable insights into borrower demographics, while suggesting further exploration and visual enhancements for deeper insights.

Regression

Regression Statistics	
Multiple R	0.45908096
R Square	0.21075532
Adjusted R Square	0.20858707
Standard Error	56.07661 11
Observations	366

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

Descriptive Statistics

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

Correlation

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

Shop Sales Data Report

Introduction

This report delves into a comprehensive sales dataset, analysing sales performance and product trends among salesmen. It aims to uncover insights for sales strategy formulation and business enhancement. By examining sales data and comparing product performance, it identifies top salesmen, analyses product popularity, and understands sales trends. These insights are invaluable for optimizing strategies and driving business growth.

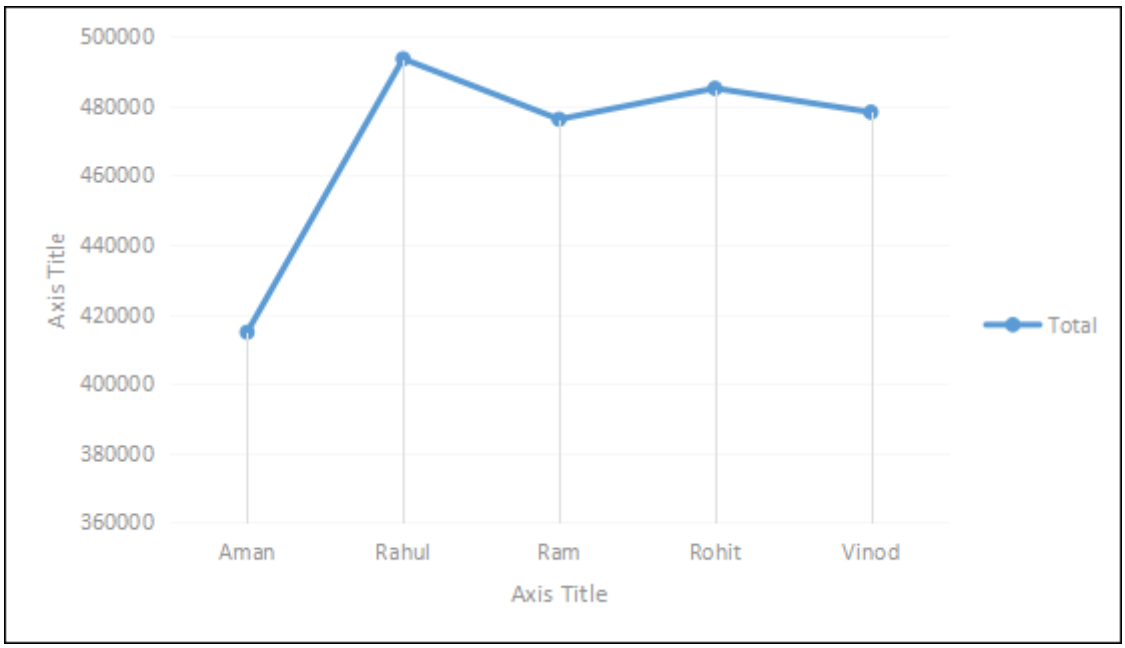
Questionaries

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

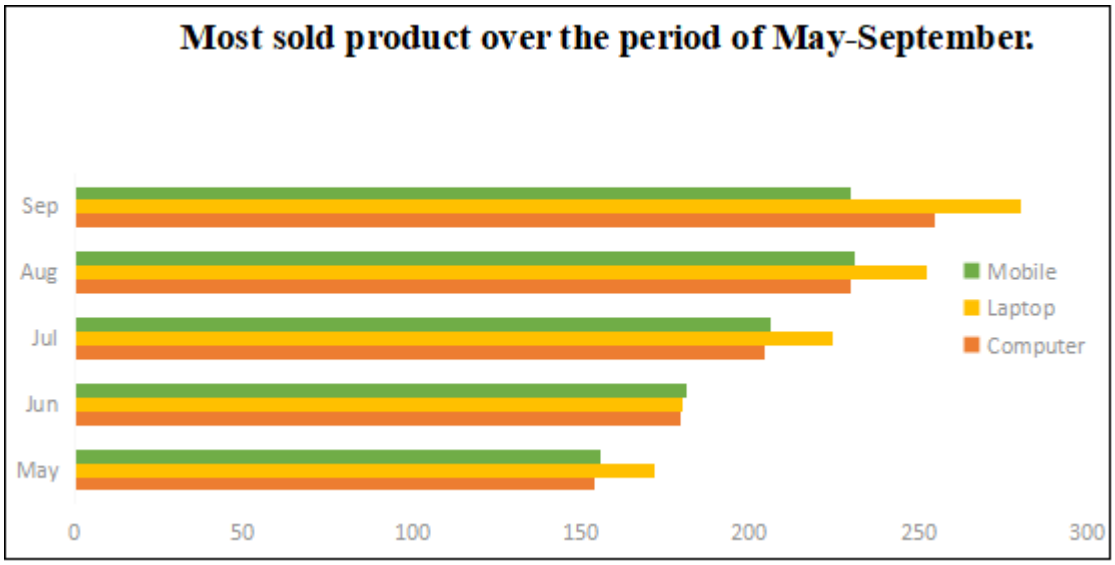
Q 1. Compare all the salesmen based on profit earned.

ANS: The comparison of all the salesmen based on profit earned and the line chart shows that Rahul has the highest profit earned with the value of 493541.3255, compared to all the salesmen.



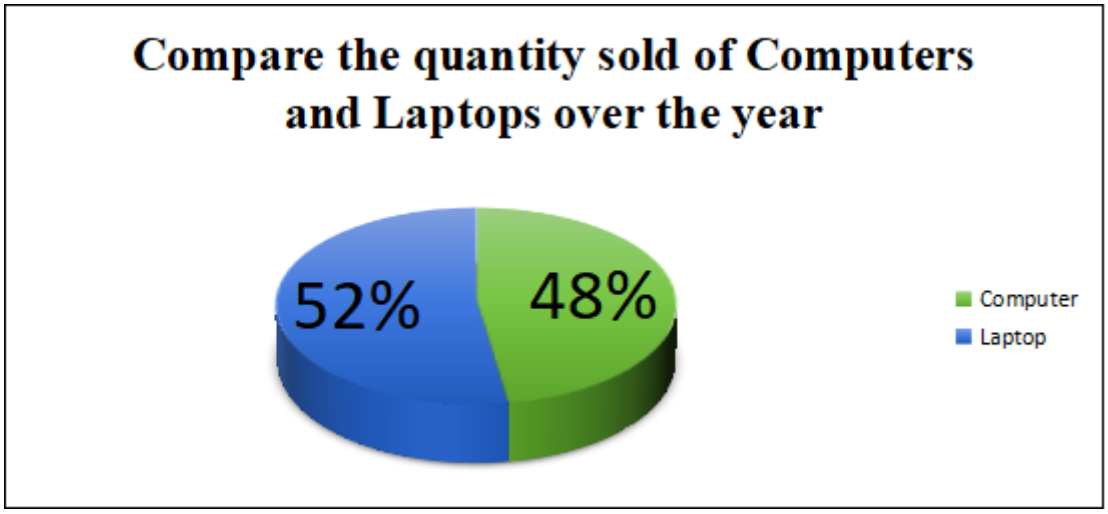
Q2. Find out most sold product over the period of May-September.

ANS: To pinpoint the most sold product from May to September, we analyze sales data within this time frame. Aggregating product quantities across all transactions reveals that the laptop was the best-selling item, particularly in September, with sales totaling 280.1970249 units.



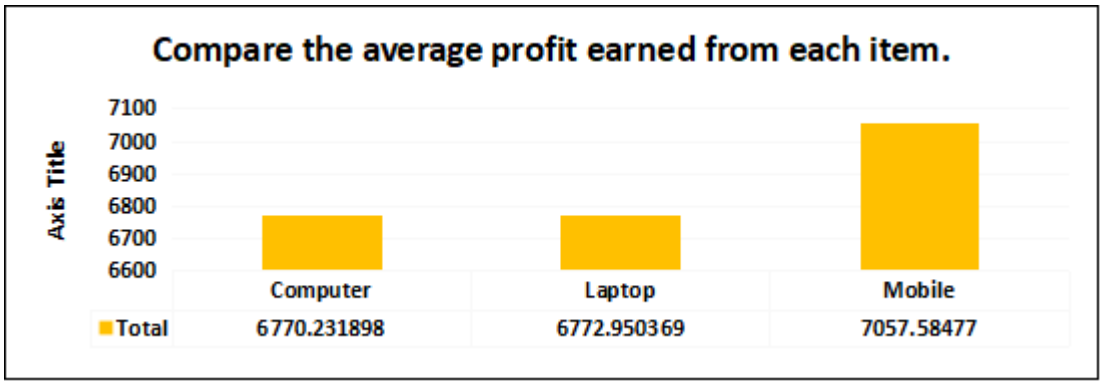
Q3. Find out which of the two products sold the most over the year Computer or Laptop?

ANS: Between computers and laptops, laptops were the best-selling product with 2,358.911786 units sold, compared to computers with 2,139.876313 sold over a year.



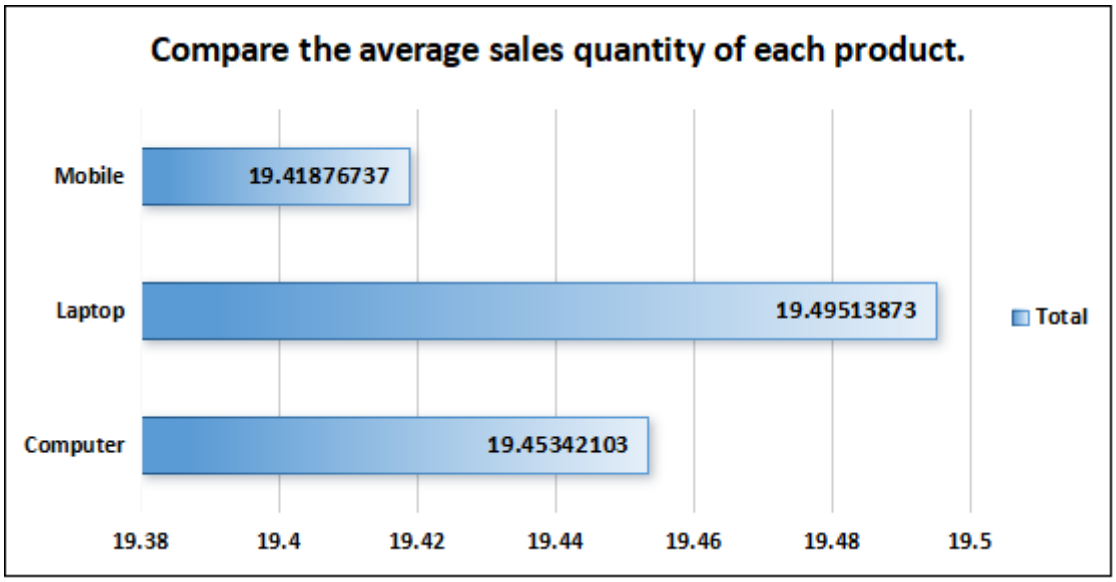
Q4. Which item yields the most average profit?

ANS: This analysis shows that Mobile has the highest Average profit earned among Mobile, Laptop, and Computer whereas Mobile has the average profit earned of 7057.58477.



Q5. Find out the average sales of all the products and compare them.

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



Conclusion and Review

The analysis uncovers crucial insights into sales performance and product trends among salesmen. Rahul emerges as the top performer, achieving the highest profit. The Laptop emerges as the best-selling product from May to September, with peak sales in September. Laptops outperform computers in units sold throughout the year. Mobile phones yield the highest average profit among devices, while laptops demonstrate the highest average sales quantity. Though providing valuable insights, deeper exploration into sales fluctuations and product preferences could enhance understanding. Overall, the report offers actionable insights for optimizing sales strategies and maximizing revenue, supported by visualizations aiding trend comprehension and product popularity assessment.

Correlation

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

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

Regression

The regression model, with a significant p-value, indicates a strong positive relationship between the Amount and the profit earned and the outcome variable. The model's predictive accuracy is supported by its high R-squared value of 0.910.

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

	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

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 factors)

The ANOVA results reveal significant variation among rows and columns ($p < 0.001$), 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				

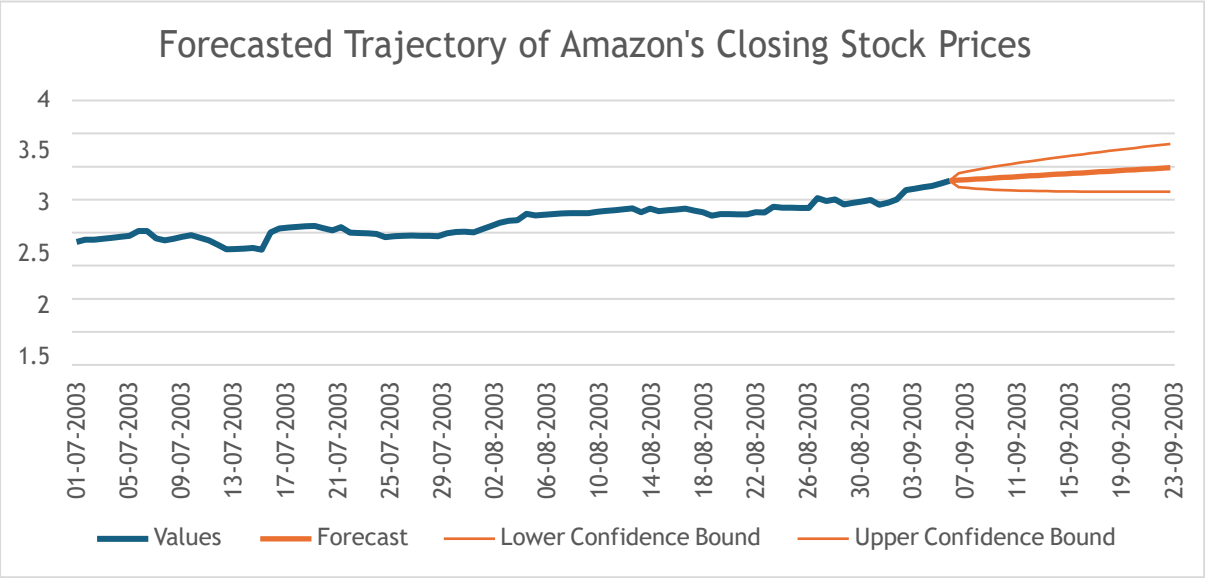
Descriptive Statistics

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

Analysis of Forecasted Trends in Amazon's Closing Stock Prices

Timeline	Values	Forecast	Lower Confidence Bound	Upper Confidence Bound
01-07-2003	1.8625			
02-07-2003	1.8925			
03-07-2003	1.896			
04-07-2003	1.910125			
05-07-2003	1.92425			
06-07-2003	1.938375			
07-07-2003	1.9525			
08-07-2003	2.0255			
09-07-2003	2.025			
10-07-2003	1.9125			
11-07-2003	1.8825			
12-07-2003	1.9095			
13-07-2003	1.9365			
14-07-2003	1.9635			
15-07-2003	1.9215			
16-07-2003	1.884			
17-07-2003	1.816			
18-07-2003	1.749			
19-07-2003	1.754833			
20-07-2003	1.760667			
21-07-2003	1.7665			
22-07-2003	1.7435			
23-07-2003	2.0055			
24-07-2003	2.0655			
25-07-2003	2.08			
26-07-2003	2.088			
27-07-2003	2.096			
28-07-2003	2.104			
29-07-2003	2.0695			
30-07-2003	2.033			
31-07-2003	2.082			
01-08-2003	2.0015			
02-08-2003	1.995167			
03-08-2003	1.988833			
04-08-2003	1.9825			
05-08-2003	1.9335			
06-08-2003	1.947			
07-08-2003	1.9505			
08-08-2003	1.9575			
09-08-2003	1.953833			
10-08-2003	1.950167			
11-08-2003	1.9465			
12-08-2003	1.9925			
13-08-2003	2.0105			
14-08-2003	2.015			
15-08-2003	2.005			
16-08-2003	2.053667			
17-08-2003	2.102333			
18-08-2003	2.151			
19-08-2003	2.1785			
20-08-2003	2.188			
21-08-2003	2.285			
22-08-2003	2.261			
23-08-2003	2.2715			
24-08-2003	2.282			
25-08-2003	2.2925			
26-08-2003	2.2965			
27-08-2003	2.294			
28-08-2003	2.297			
29-08-2003	2.316			

30-08-2003	2.328625			
31-08-2003	2.34125			
01-09-2003	2.353875			
02-09-2003	2.3665			
03-09-2003	2.3095			
04-09-2003	2.3645			
05-09-2003	2.326			
06-09-2003	2.338667			
07-09-2003	2.351333			
08-09-2003	2.364			
09-09-2003	2.334			
10-09-2003	2.311			
11-09-2003	2.2595			
12-09-2003	2.284			
13-09-2003	2.281167			
14-09-2003	2.278333			
15-09-2003	2.2755			
16-09-2003	2.312			
17-09-2003	2.308			
18-09-2003	2.3945			
19-09-2003	2.379			
20-09-2003	2.377167			
21-09-2003	2.375333			
22-09-2003	2.3735			
23-09-2003	2.522			
24-09-2003	2.4805			
25-09-2003	2.5025			
26-09-2003	2.428			
27-09-2003	2.449667			
28-09-2003	2.471333			
29-09-2003	2.493			
30-09-2003	2.4215			
01-10-2003	2.456			
02-10-2003	2.5045			
03-10-2003	2.6445			
04-10-2003	2.6655			
05-10-2003	2.6865			
06-10-2003	2.7075			
07-10-2003	2.7455			
08-10-2003	2.785	2.7850001	2.79	2.79
09-10-2003		2.7928846	2.69	2.90
10-10-2003		2.8007691	2.68	2.93
11-10-2003		2.8086536	2.67	2.95
12-10-2003		2.8165382	2.66	2.97
13-10-2003		2.8244227	2.65	3.00
14-10-2003		2.8323072	2.65	3.02
15-10-2003		2.8401917	2.64	3.04
16-10-2003		2.8480762	2.64	3.06
17-10-2003		2.8559607	2.63	3.08
18-10-2003		2.8638453	2.63	3.10
19-10-2003		2.8717298	2.63	3.11
20-10-2003		2.8796143	2.63	3.13
21-10-2003		2.8874988	2.62	3.15
22-10-2003		2.8953833	2.62	3.17
23-10-2003		2.9032678	2.62	3.18
24-10-2003		2.9111524	2.62	3.20
25-10-2003		2.9190369	2.62	3.22
26-10-2003		2.9269214	2.62	3.23
27-10-2003		2.9348059	2.62	3.25
28-10-2003		2.9426904	2.62	3.27
29-10-2003		2.9505749	2.62	3.28
30-10-2003		2.9584595	2.62	3.30
31-10-2003		2.966344	2.62	3.31
01-11-2003		2.9742285	2.62	3.33
02-11-2003		2.982113	2.62	3.34



The forecast depicted in the line graph illustrates the projected trajectory of Amazon's closing stock prices from October 8, 2003, onwards. This forecast extends beyond the historical data, offering insights into potential future price movements.

Accompanied by lower and upper confidence bounds, the forecast provides a range of possible outcomes, accounting for the inherent uncertainty in predicting stock prices. These bounds delineate the expected variability in the forecasted values, offering stakeholders a perspective on the potential risk associated with the forecast.

The summary highlights the analytical depth achieved in anticipating future trends in Amazon's stock prices. This predictive analysis equips stakeholders with valuable insights for strategic decision-making in financial markets.