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 Amazon's Closing Stock Prices	

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

Ques 1. Data analysis principles
Ans 1: 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, 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, market 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.

Ques 2. Statistical analytics concept
Ans 2: 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 own strengths and is used based on the specific analysis goals and characteristics of the data.

Ques 3. hypothesis
Ans 3: 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.

Ques 4. Regression and its types
Ans 4: 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.

Ques 5. Correlation

Ans 5: 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!

Ques 6. ANOVA

Ans 6: 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:

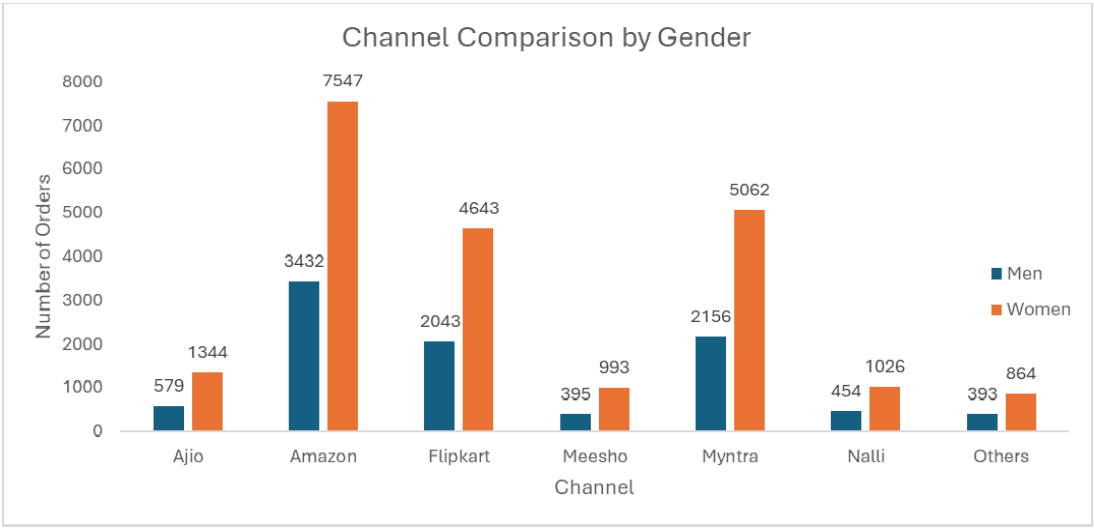
F = (Between-group variation / Within-group variation)

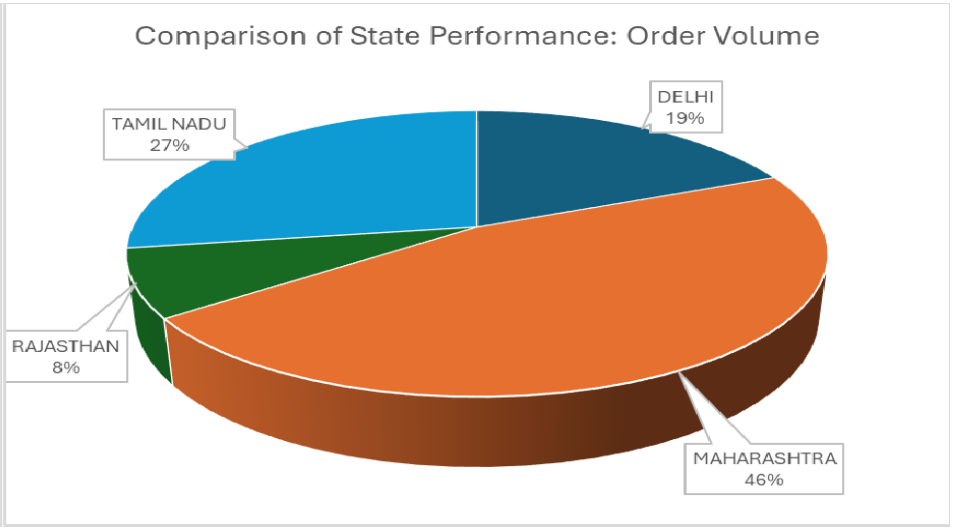
To calculate the between-group variation, sum of squares between (SSB), which measures the differences between the group means. The within-group variation is calculated using the sum of squares within (SSW), which measures the variability within each group.

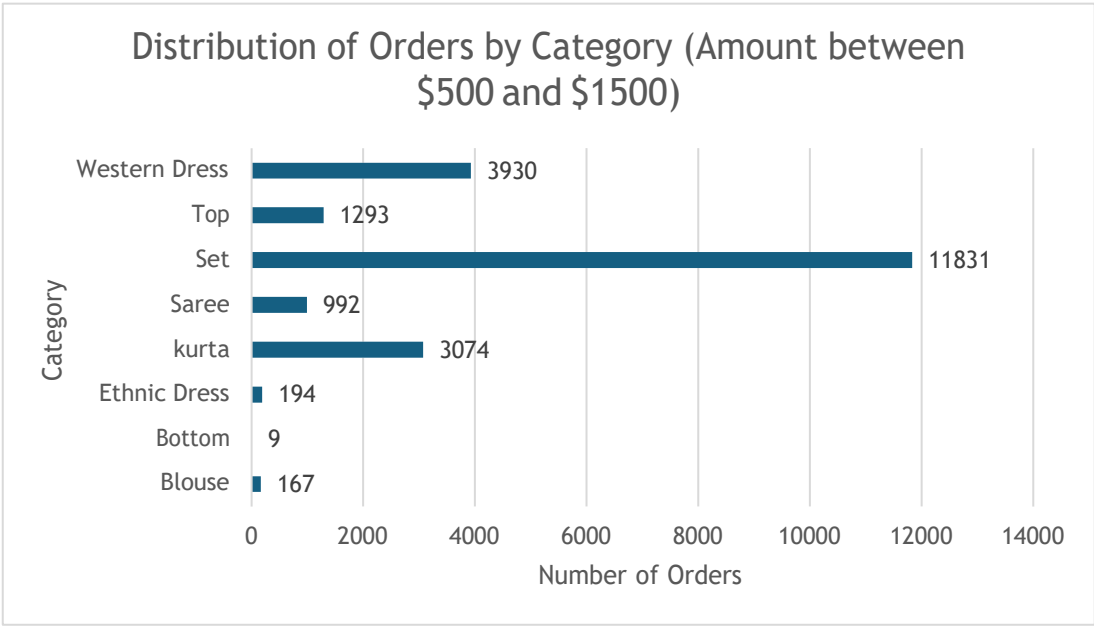
The sum of squares is obtained by summing the squared differences between each observation and the group mean. Then, these sums of squares are divided by their respective degrees of freedom (DFB and DFW) to calculate the mean squares.

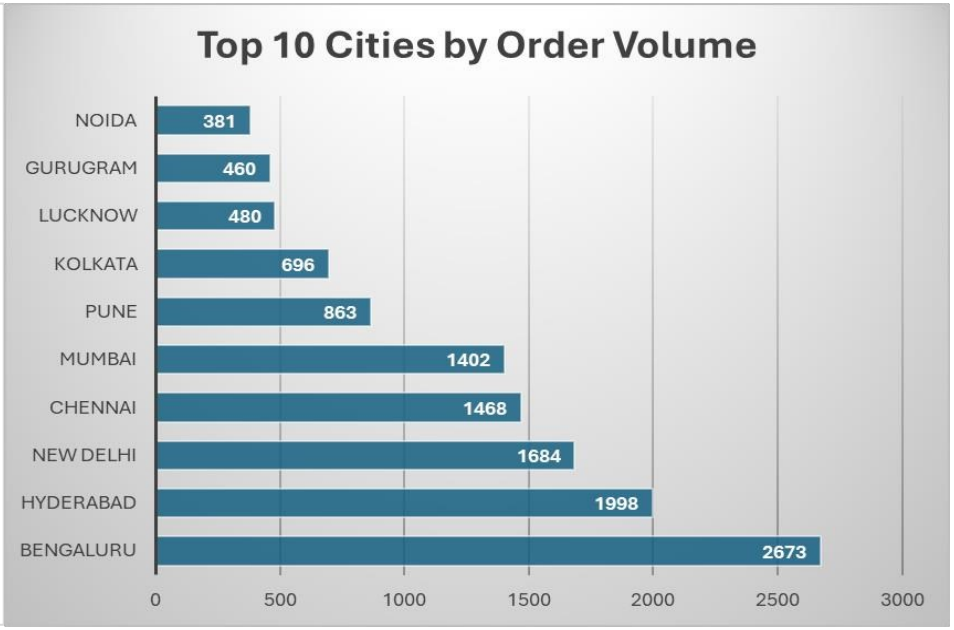
Finally, the F-statistic is calculated by dividing the mean square between by the mean square within.

Store Data Analysis

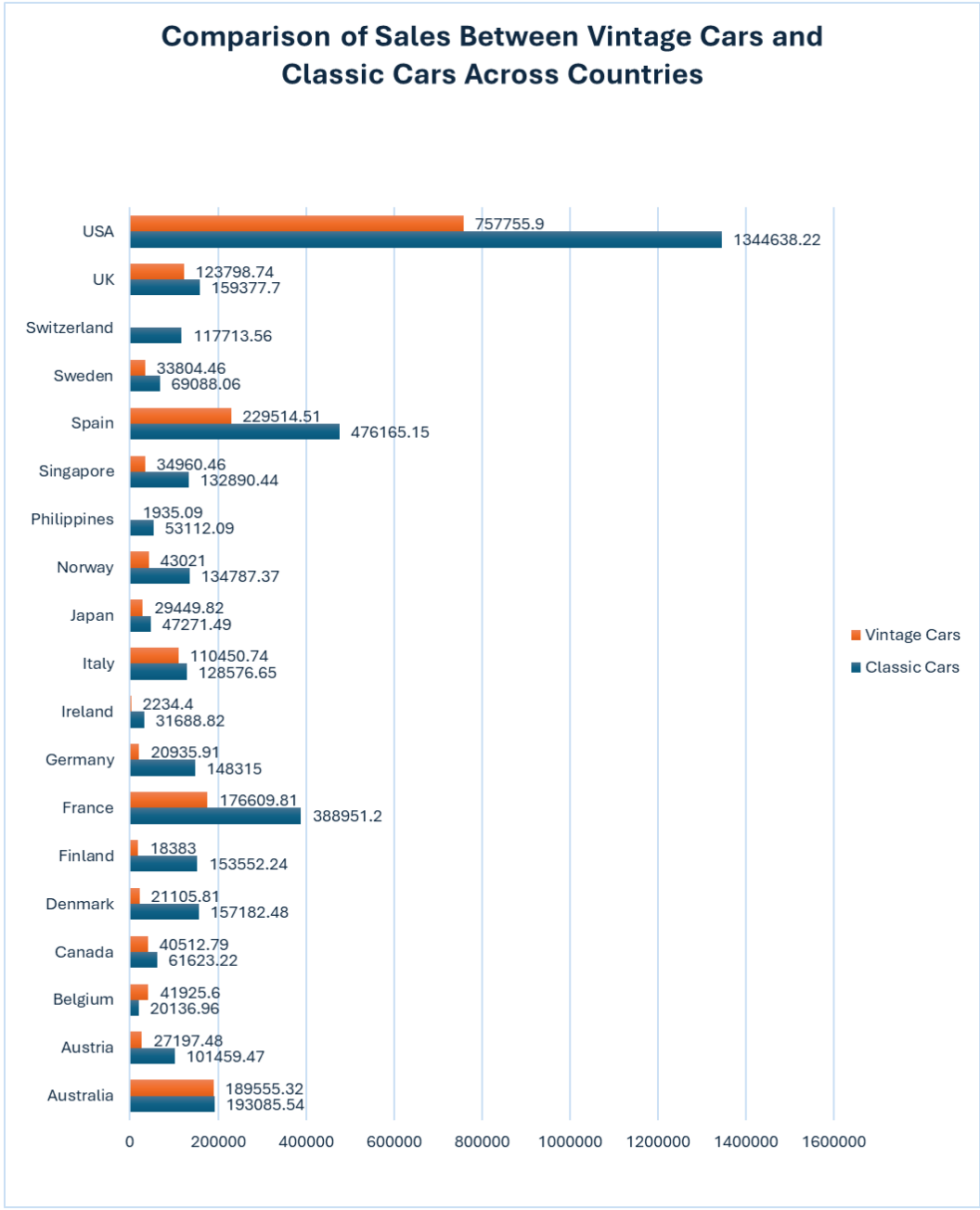


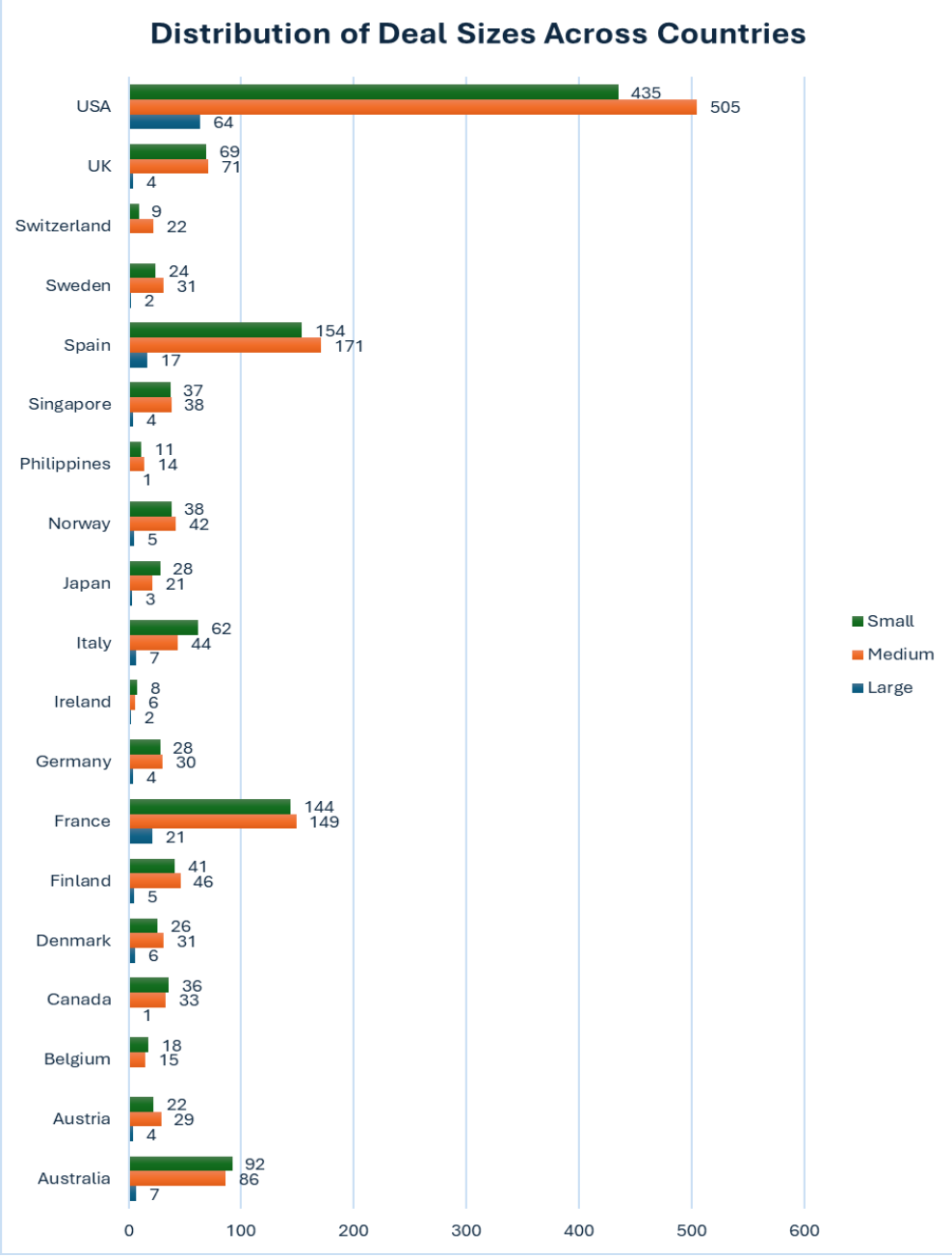


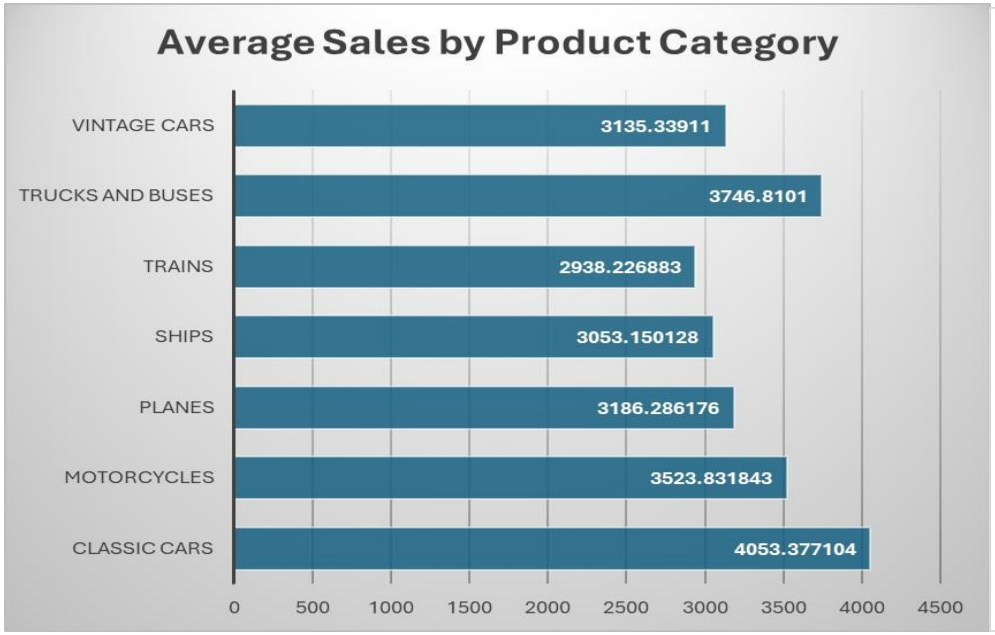


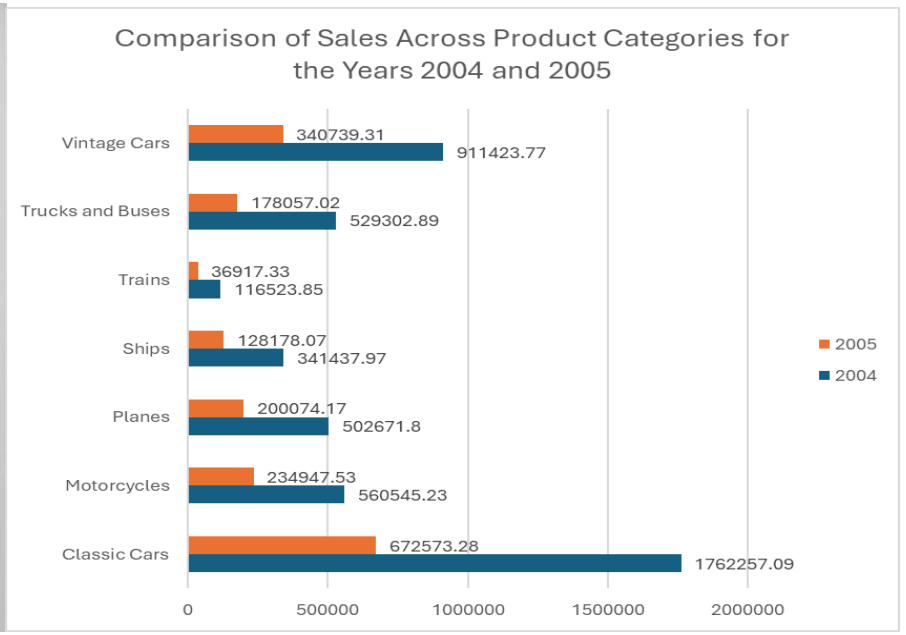


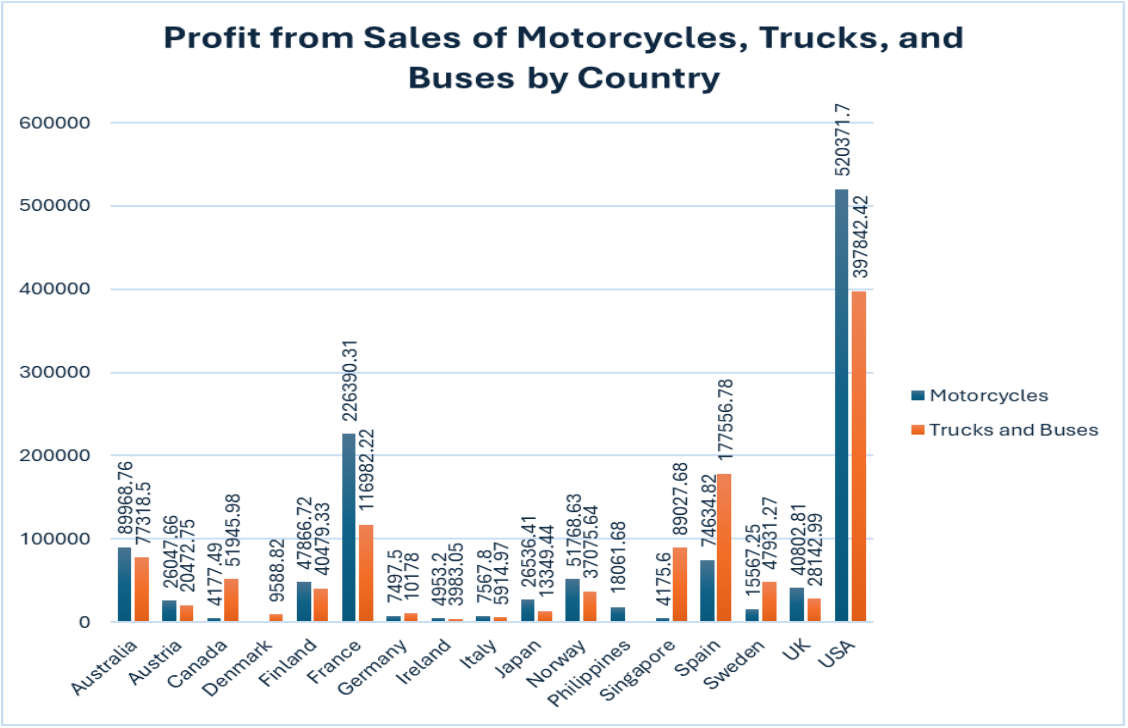
Sales Data Analysis



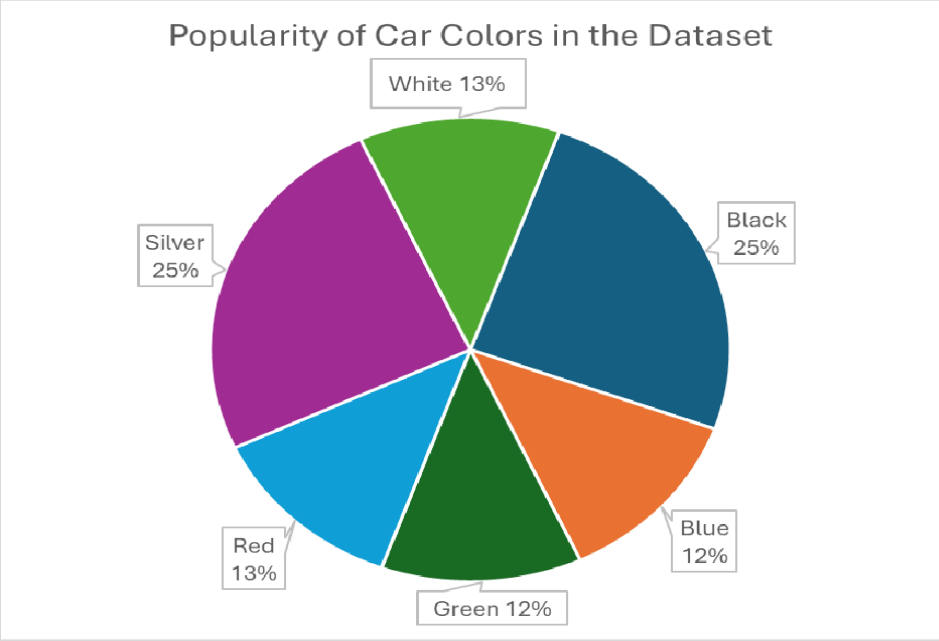


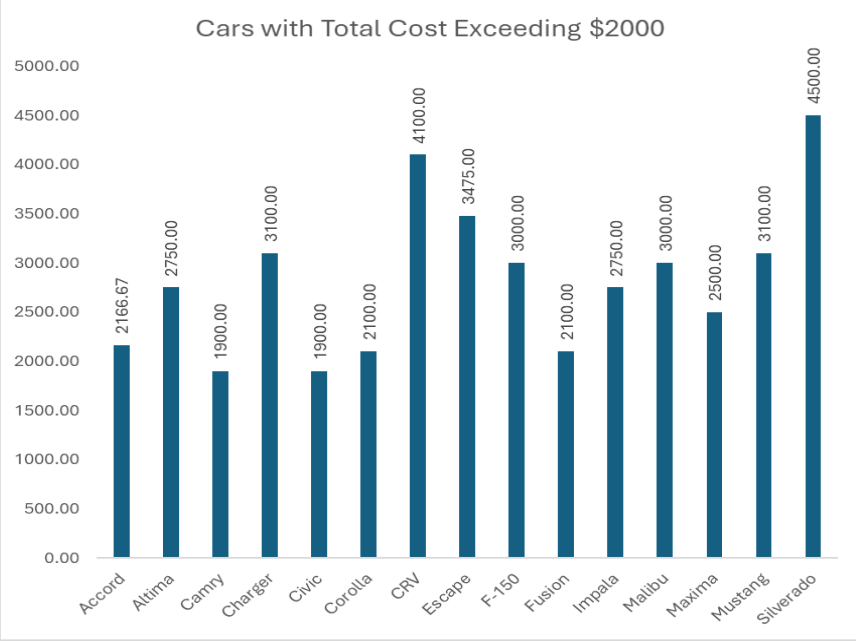

Average Sales by Product Category
VINTAGE CARS
TRUCKS AND BUSES
TRAINS
SHIPS
PLANES
MOTORCYCLES
CLASSIC CARS
3135.33911
3746.8101
2938.226883
3053.150128
3186.286176
3523.831843
4053.377104
0
500
1000
1500
2000
2500
3000
3500
4000
4500

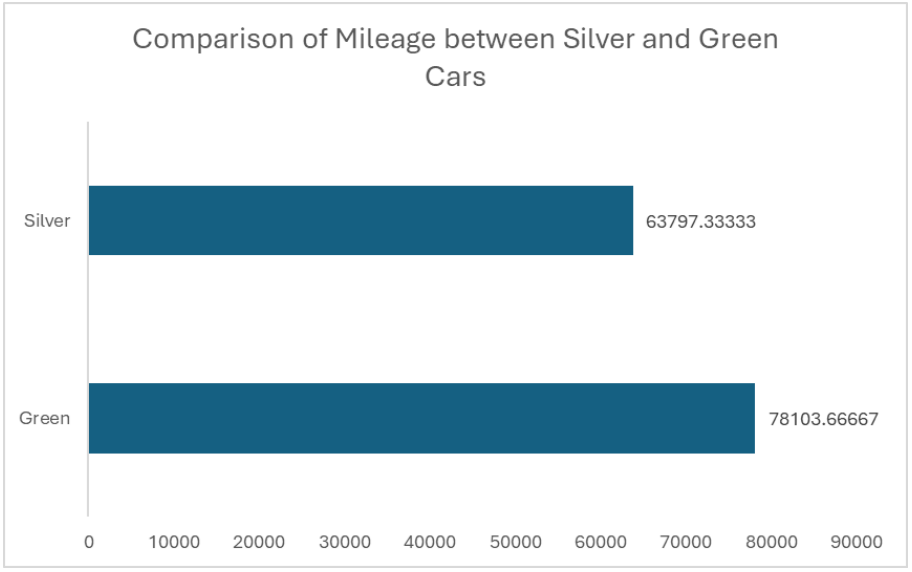

Comparison of Sales Across Product Categories for the Years 2004 and 2005
Vintage Cars
Trucks and Buses
Trains
Ships
Planes
Motorcycles
Classic Cars
340739.31
911423.77
178057.02
529302.89
36917.33
116523.85
128178.07
341437.97
200074.17
502671.8
234947.53
560545.23
672573.28
1762257.09
2005
2004
0
500000
1000000
1500000
2000000

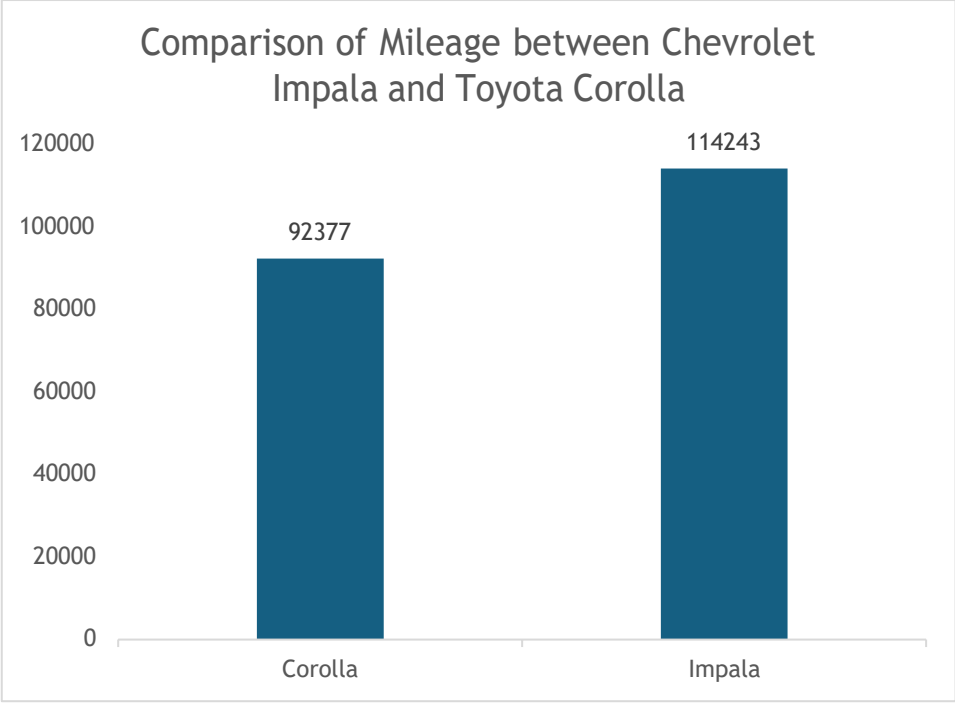

Profit from Sales of Motorcycles, Trucks, and Buses by Country
600000
500000
400000
300000
200000
100000
0
Australia 89968.76 77318.5
Austria 26047.66 20472.75
Canada 4177.49 51945.98
Denmark 9588.82
Finland 47866.72 40479.33
France 226390.31 116982.22
Germany 7497.5 10178
Ireland 4953.2 3983.05
Italy 7567.8 5914.97
Japan 26536.41 13349.44
Norway 51768.63 37075.64
Philippines 18061.68
Singapore 4175.6 89027.68
Spain 74634.82 177556.78
Sweden 15567.25 47931.27
UK 40802.81 28142.99
USA 520371.7 397842.42
Motorcycles
Trucks and Buses

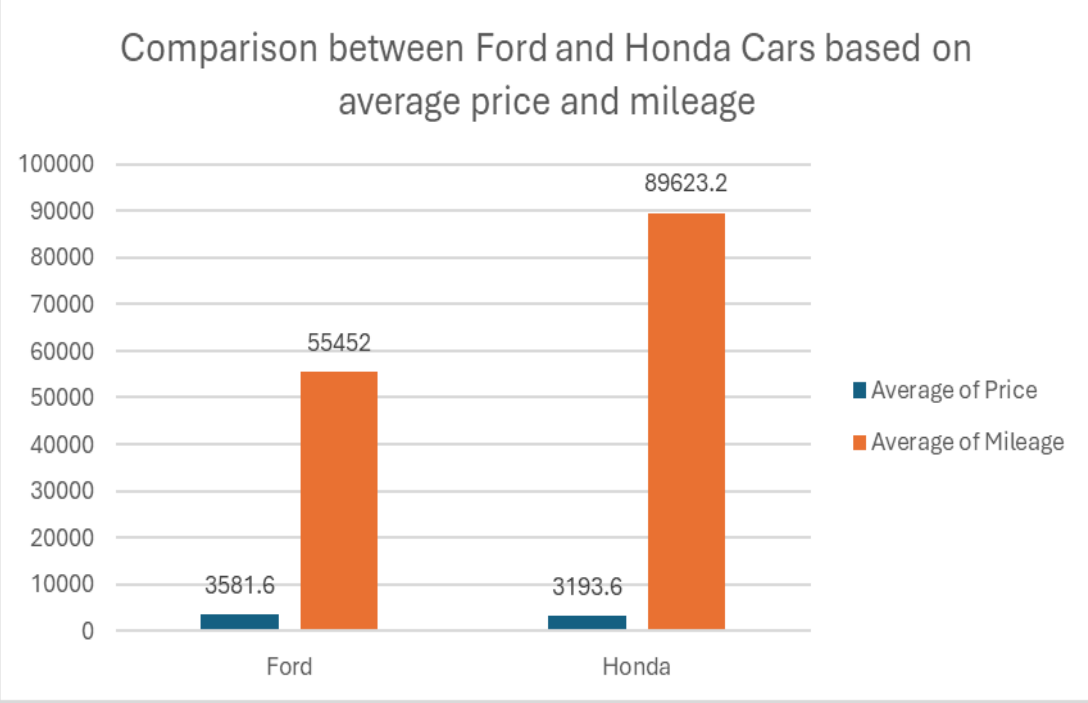
Comprehensive Analysis of Car Attributes: Insights from a Car Collection Dataset



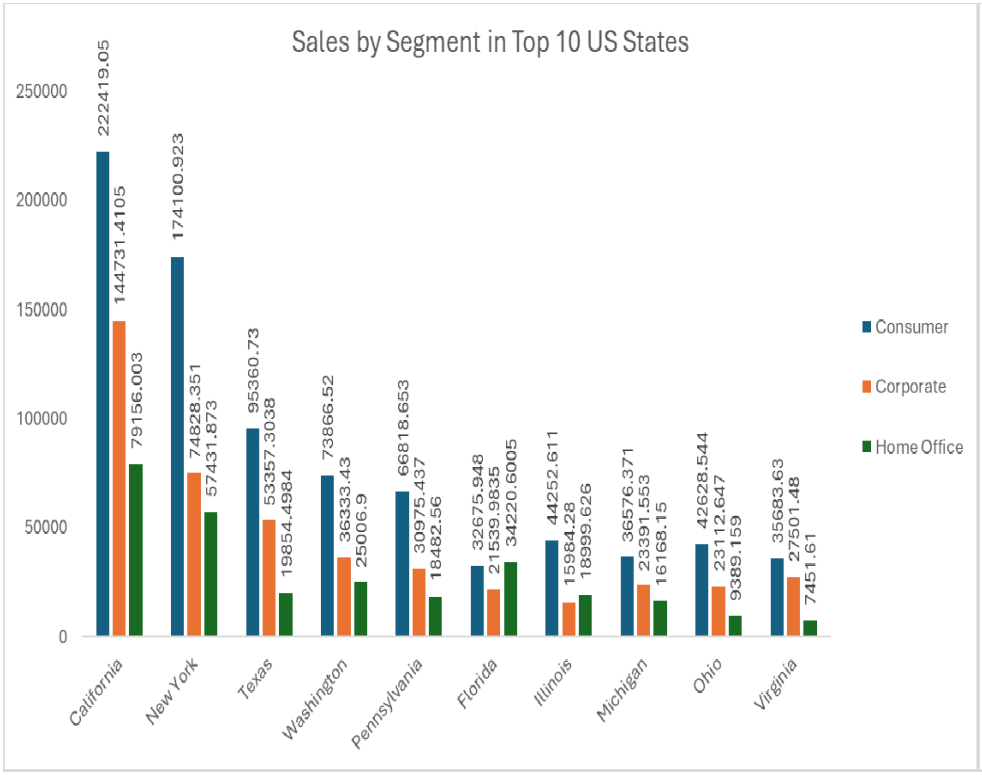


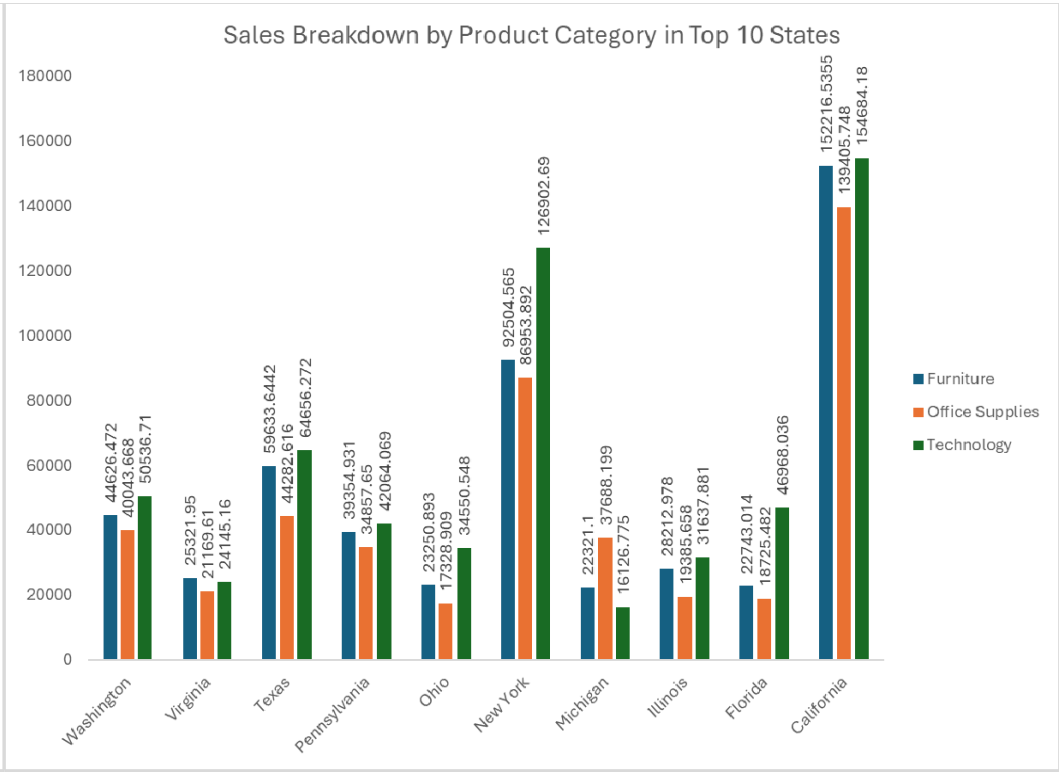


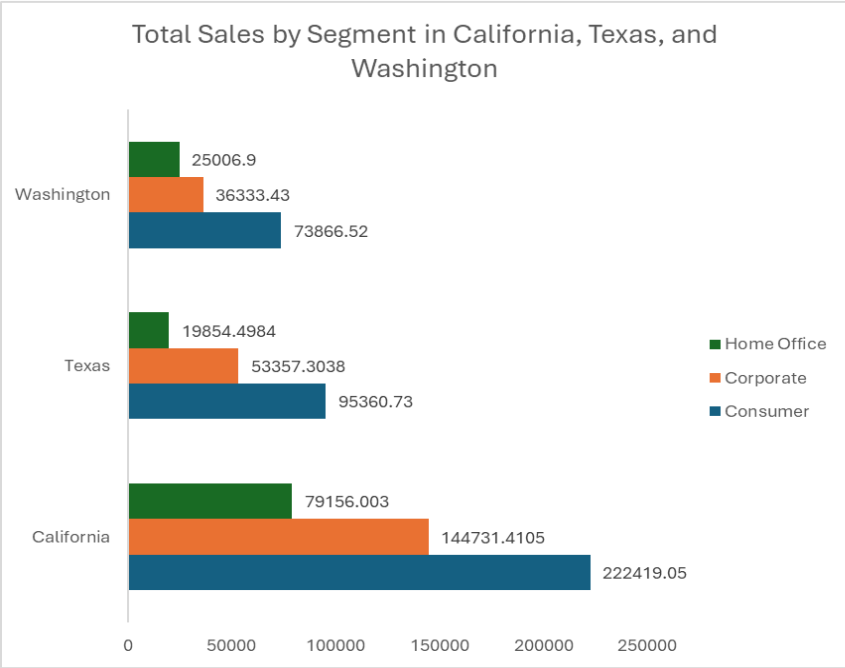

Comparison of Mileage between Chevrolet Impala and Toyota Corolla
120000
100000
80000
60000
40000
20000
0
92377
114243
Corolla
Impala

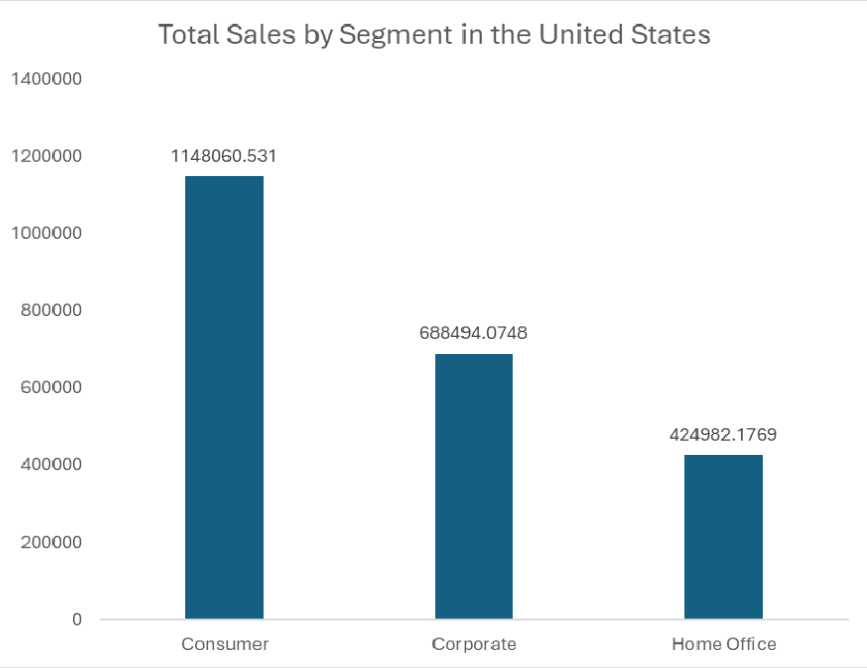

Comparison between Ford and Honda Cars based on average price and mileage
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
3581.6
55452
3193.6
89623.2
Ford
Honda
Average of Price
Average of Mileage

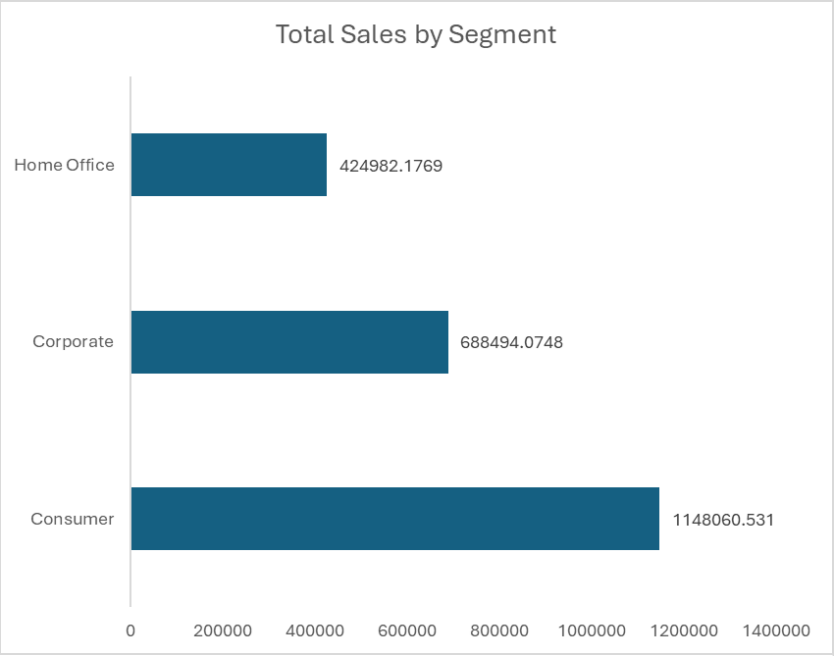
Understanding Sales: Orders, Regions, and Segments

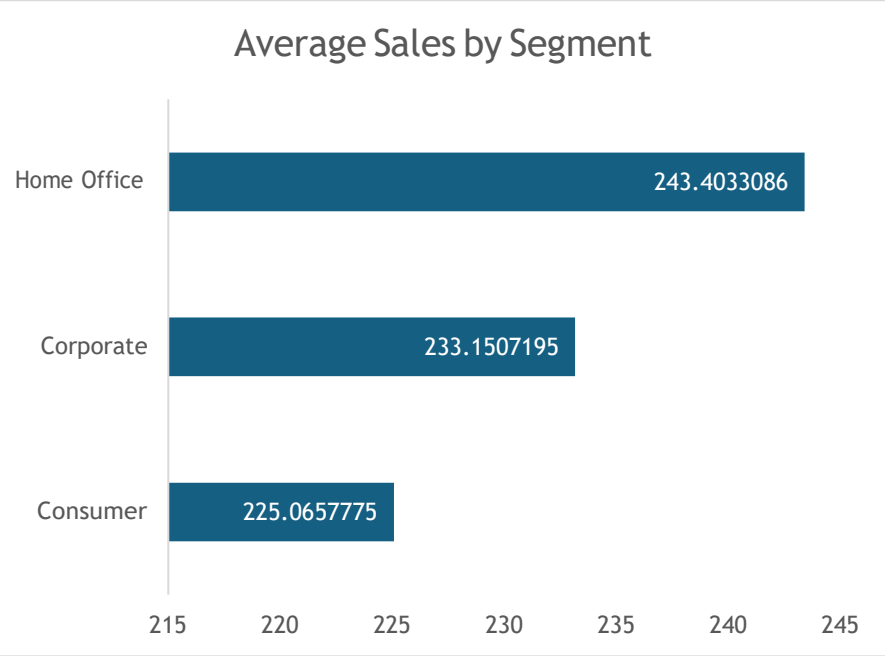


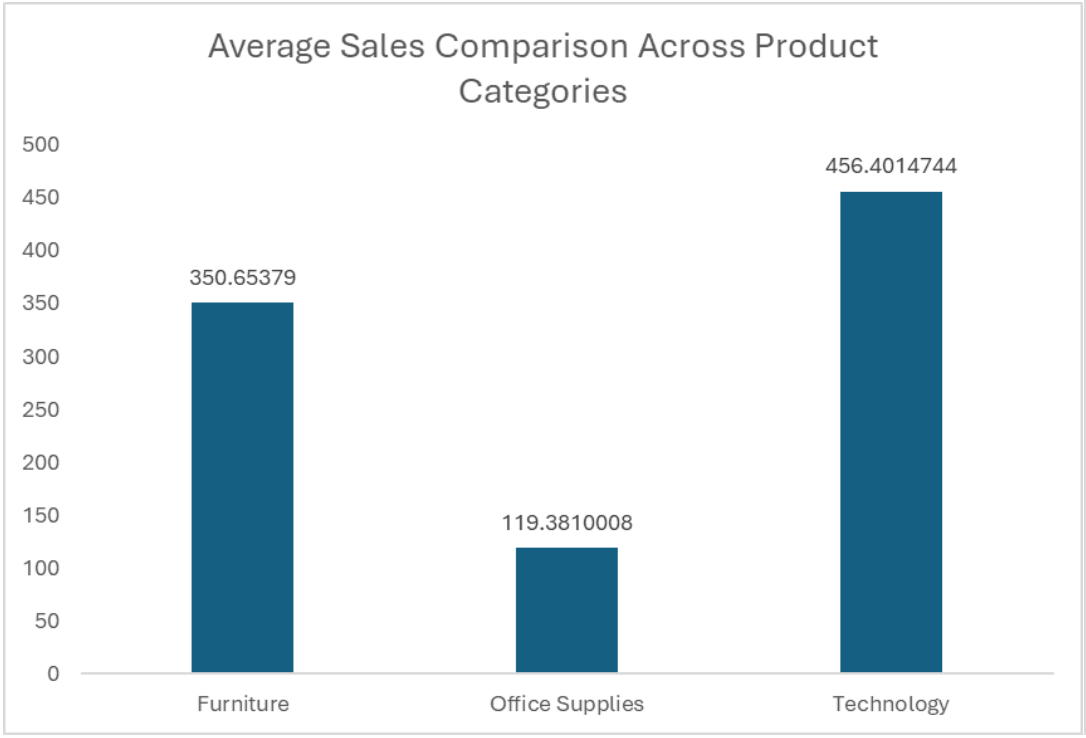


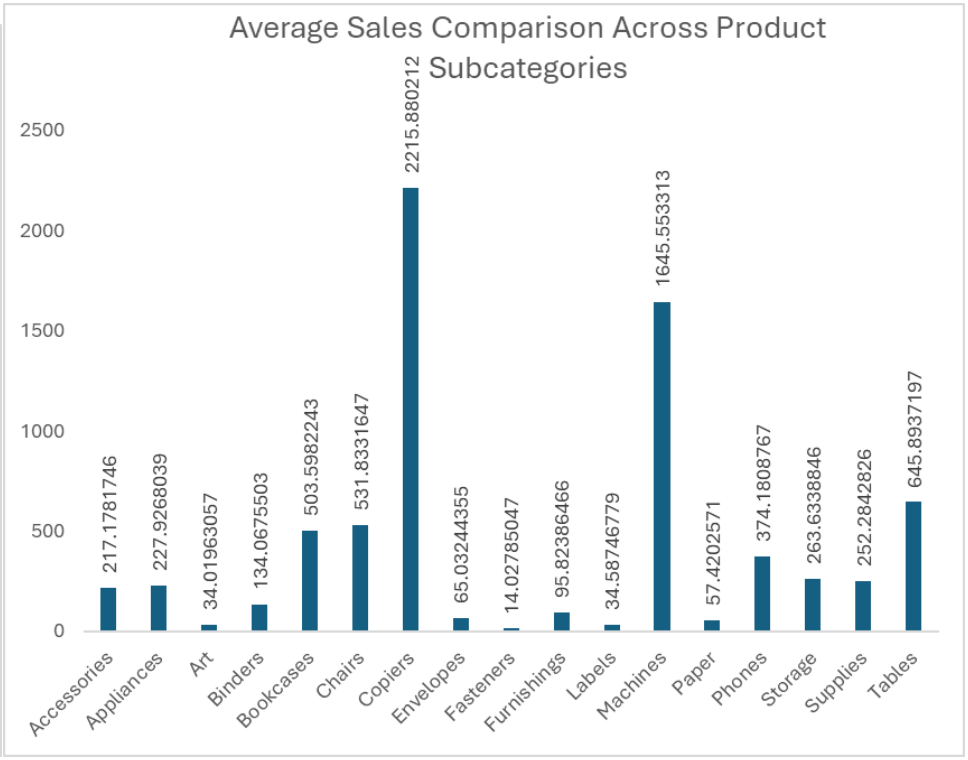

Total Sales by Segment in California, Texas, and Washington
Washington
25006.9
36333.43
73866.52
Texas
19854.4984
53357.3038
95360.73
California
79156.003
144731.4105
222419.05
Home Office
Corporate
Consumer
0
50000
100000
150000
200000
250000


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

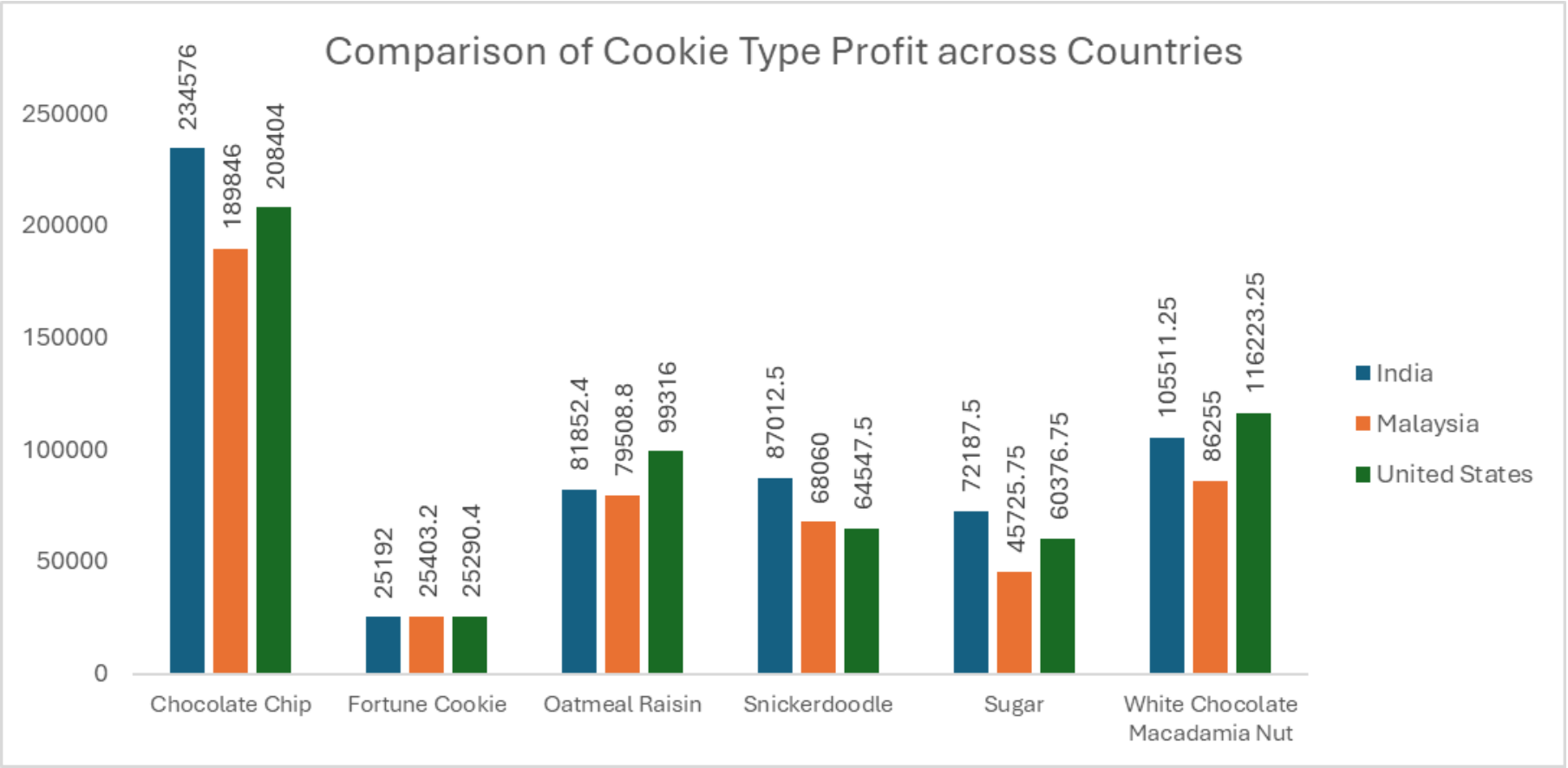

Total Sales by Segment
Home Office
424982.1769
Corporate
688494.0748
Consumer
1148060.531
0
200000
400000
600000
800000
1000000
1200000
1400000

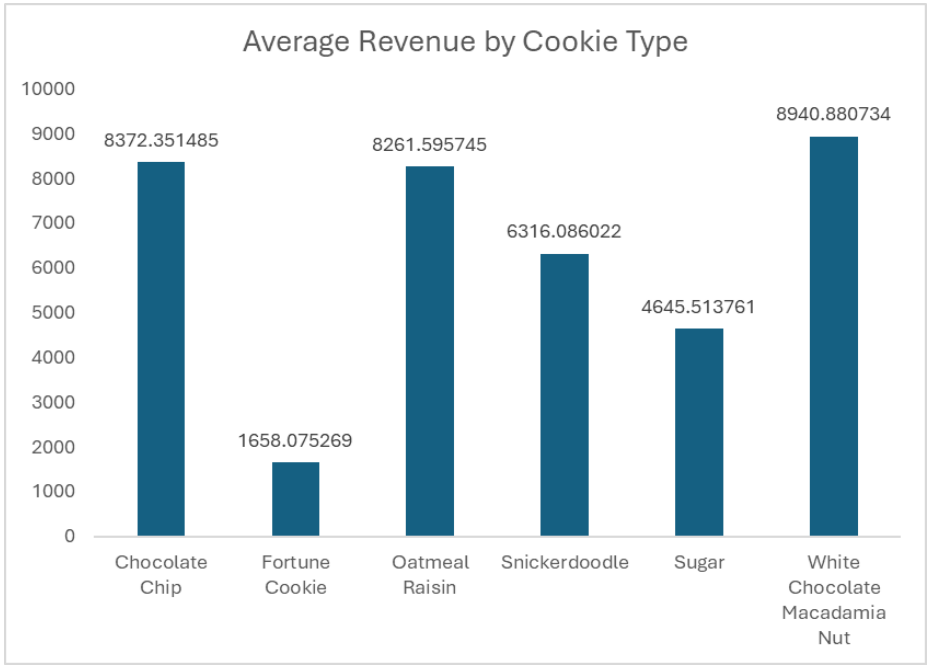

Average Sales by Segment
Home Office
243.4033086
Corporate
233.1507195
Consumer
225.0657775
215
220
225
230
235
240
245

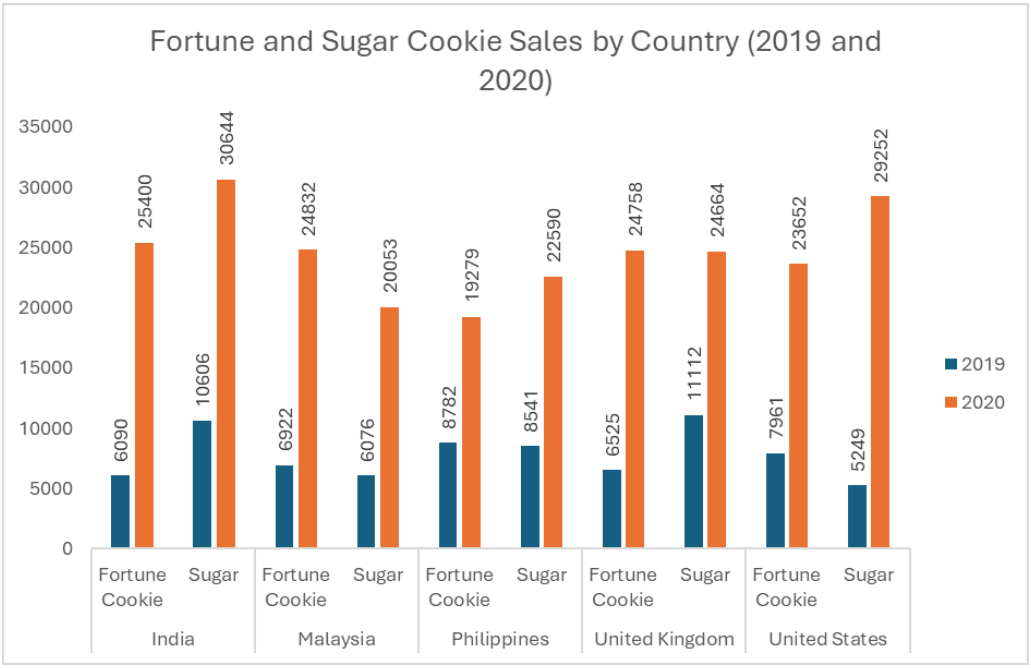

Average Sales Comparison Across Product Categories
500
450
400
350
300
250
200
150
100
50
0
350.65379
119.3810008
456.4014744
Furniture
Office Supplies
Technology

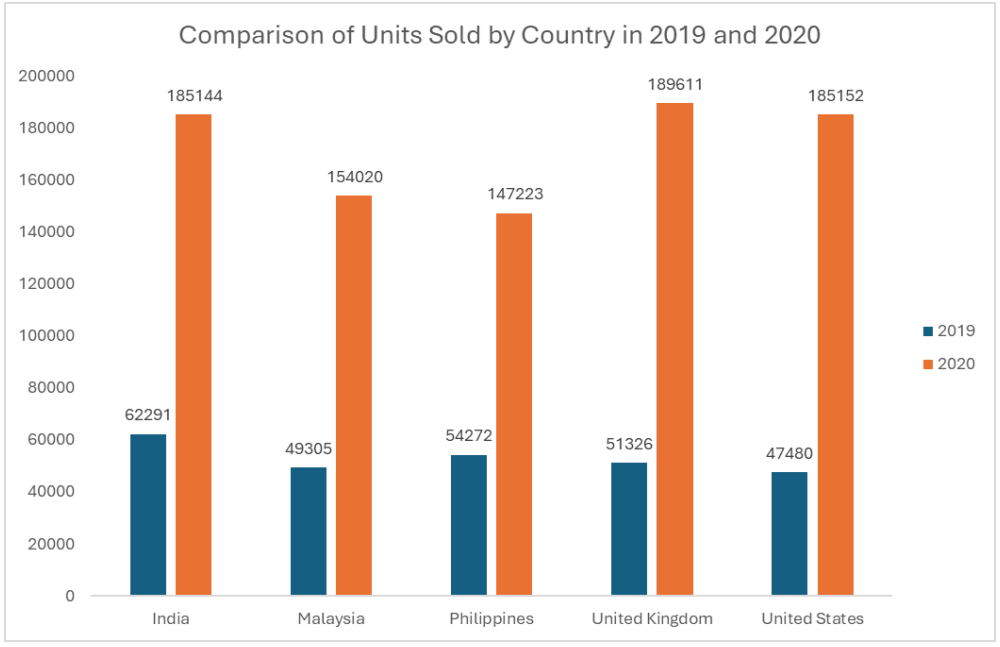

Average Sales Comparison Across Product Subcategories
2500
2000
1500
1000
500
0
217.1781746
227.926839
34.01963057
134.0675503
503.5982243
531.8331647
2215.880212
65.03244355
14.02785047
95.82386466
34.58746779
1645.553313
57.4202571
374.1808767
263.6338846
252.2842826
645.8937197
Accessories
Appliances
Art
Binders
Bookcases
Chairs
Copiers
Envelopes
Fasteners
Furnishings
Labels
Machines
Paper
Phones
Storage
Supplies
Tables

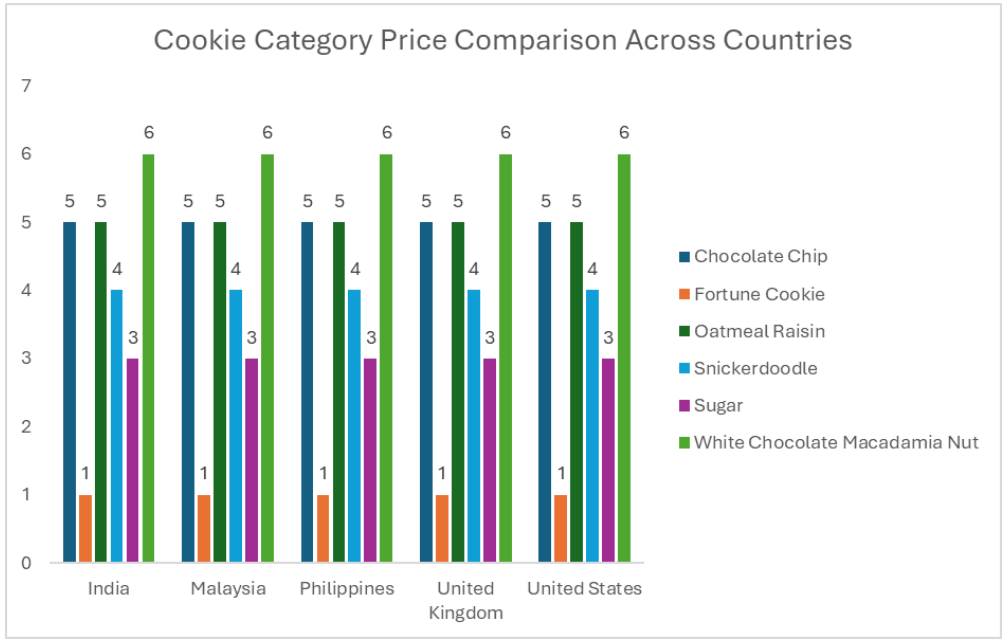
Analysis of Cookie Sales Performance Across Countries



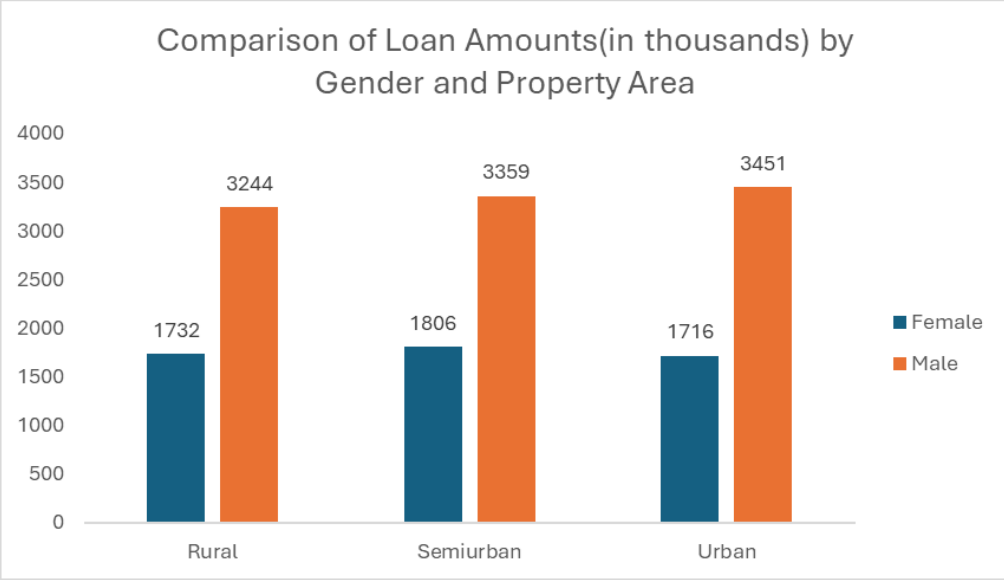

Average Revenue by Cookie Type
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
8372.351485
1658.075269
8261.595745
6316.086022
4645.513761
8940.880734
Chocolate Chip
Fortune Cookie
Oatmeal Raisin
Snickerdoodle
Sugar
White Chocolate Macadamia Nut

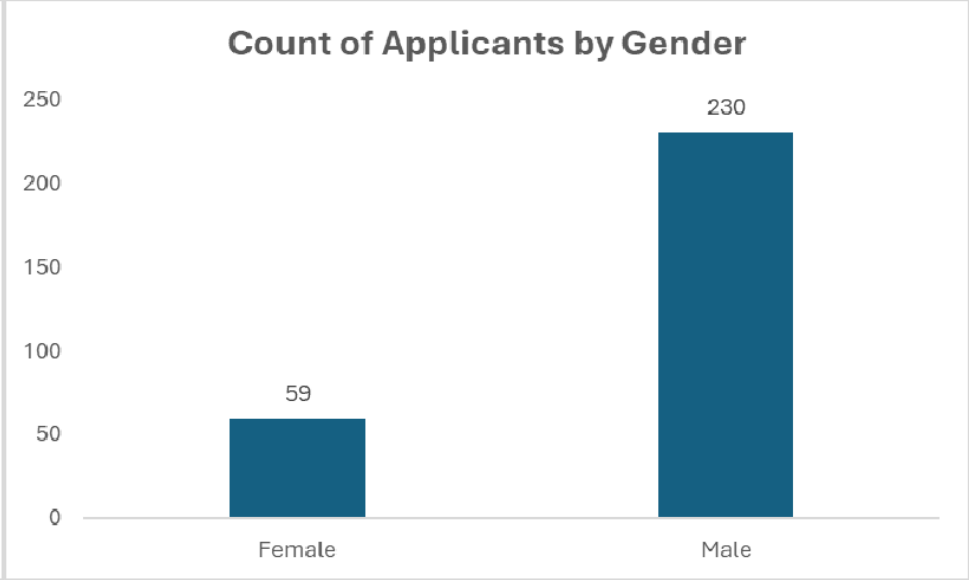

Fortune and Sugar Cookie Sales by Country (2019 and 2020)
35000
30000
25000
20000
15000
10000
5000
0
6090
25400
10606
30644
6922
24832
6076
20053
8782
19279
8541
22590
6525
24758
11112
24664
7961
23652
5249
29252
Fortune Cookie
Sugar
India
Malaysia
Philippines
United Kingdom
United States
2019
2020

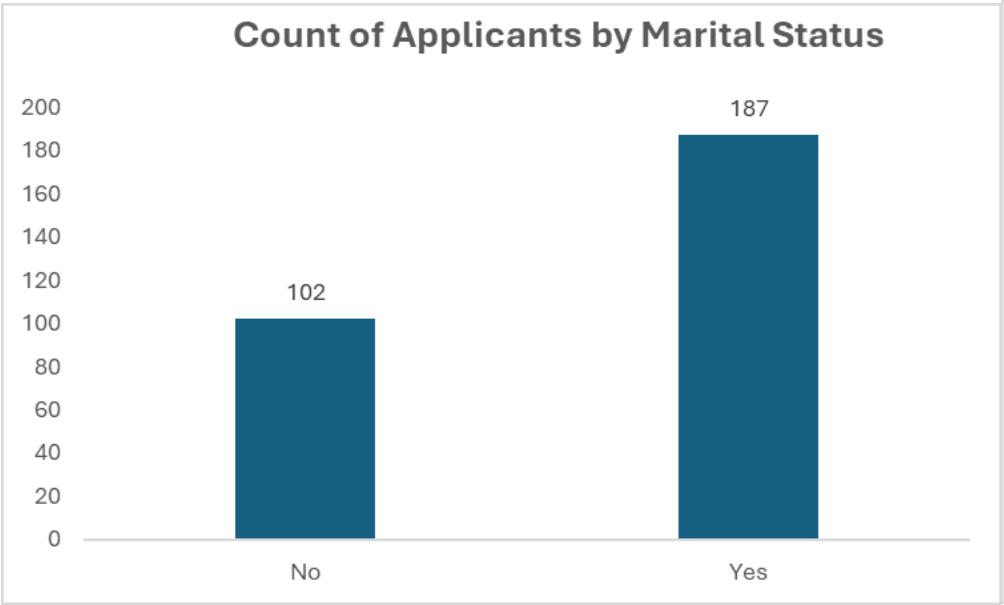

Comparison of Units Sold by Country in 2019 and 2020
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
62291
185144
49305
154020
54272
147223
51326
189611
47480
185152
India
Malaysia
Philippines
United Kingdom
United States
2019
2020

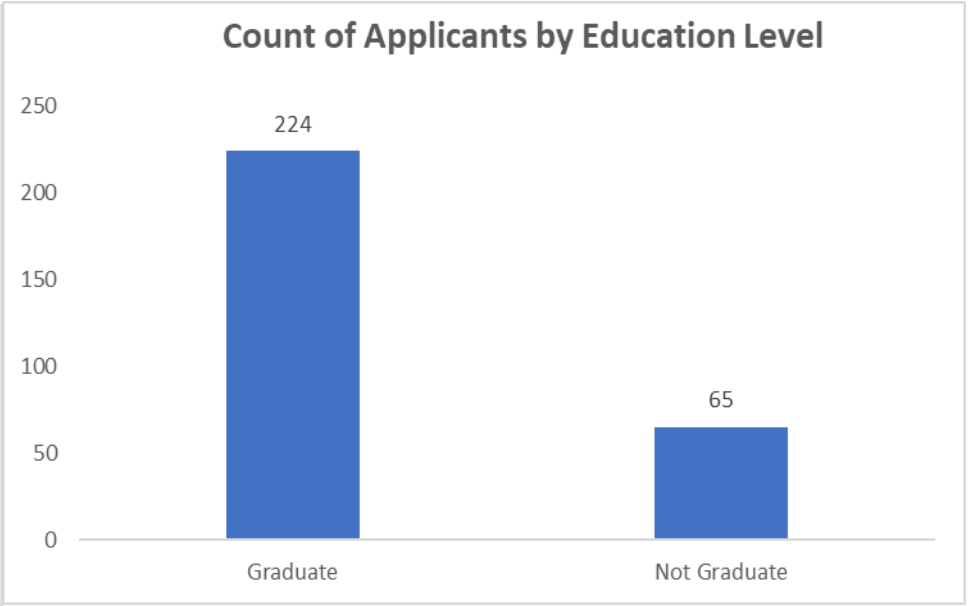

Cookie Category Price Comparison Across Countries
7
6
5
4
3
2
1
0
India
Malaysia
Philippines
United Kingdom
United States
Chocolate Chip
Fortune Cookie
Oatmeal Raisin
Snickerdoodle
Sugar
White Chocolate Macadamia Nut

Analysis of Loan Applicants

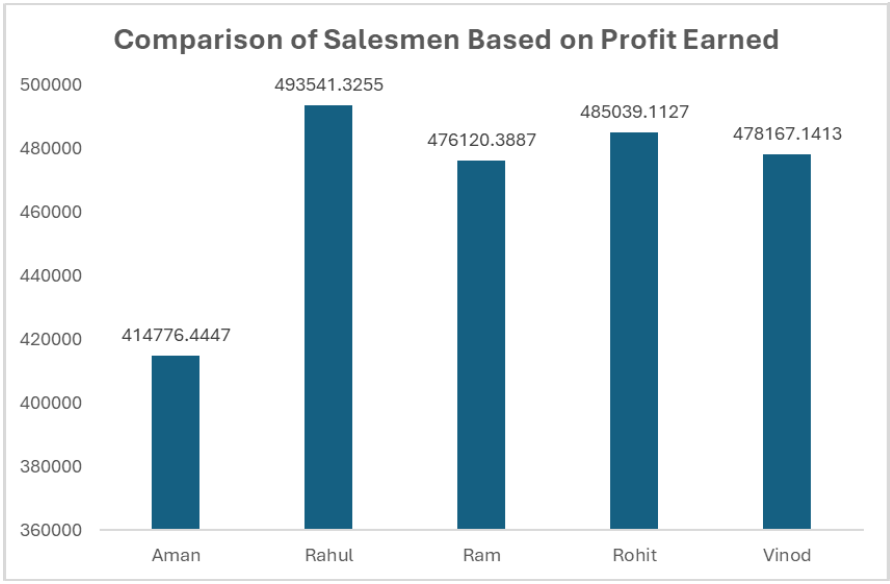


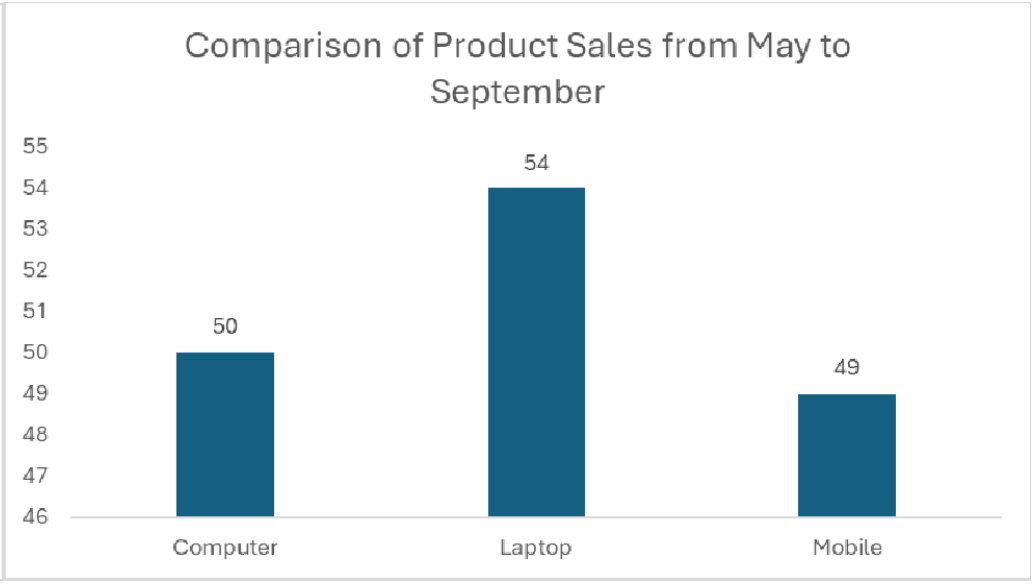


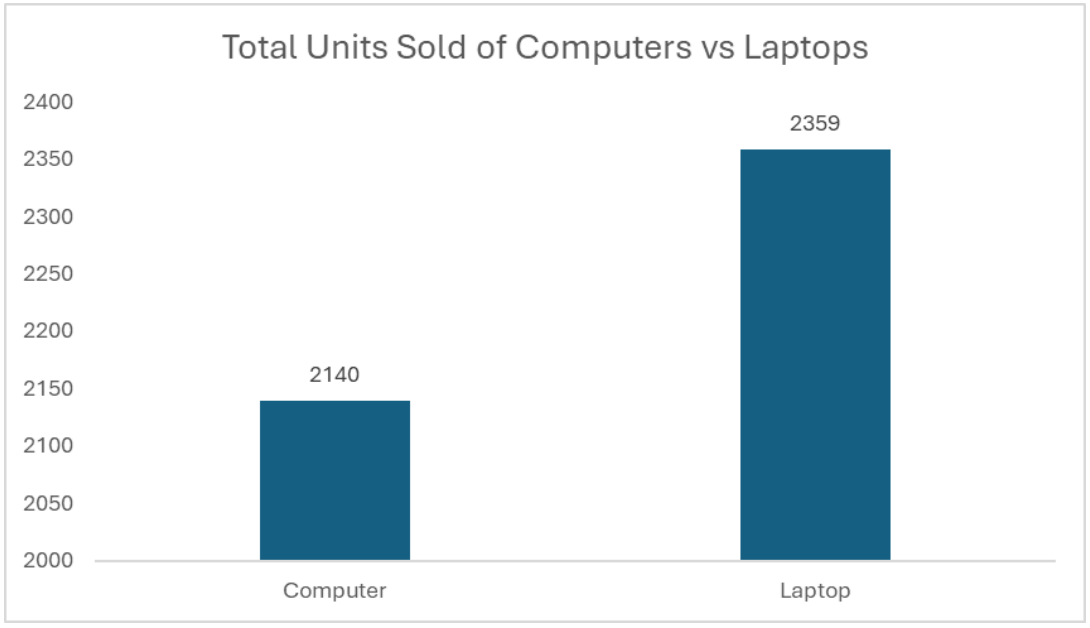


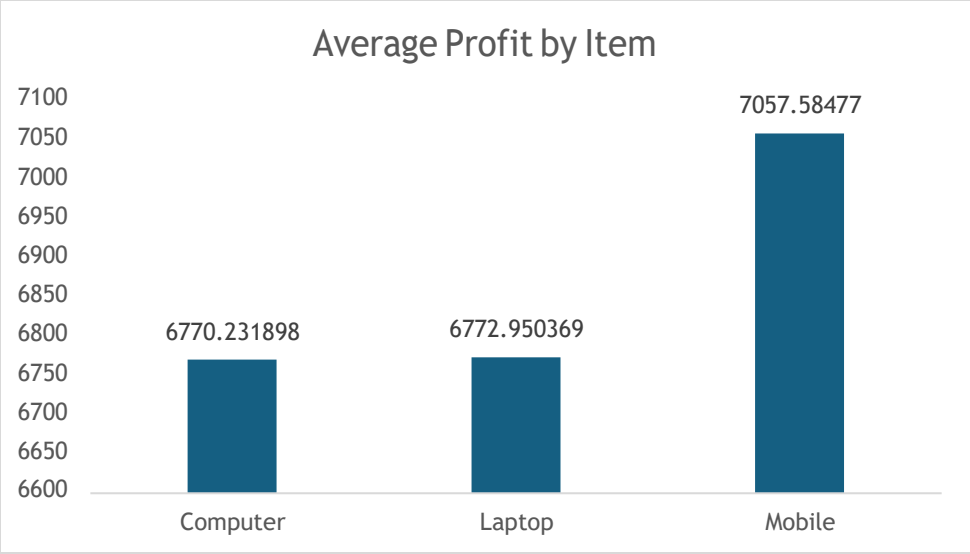


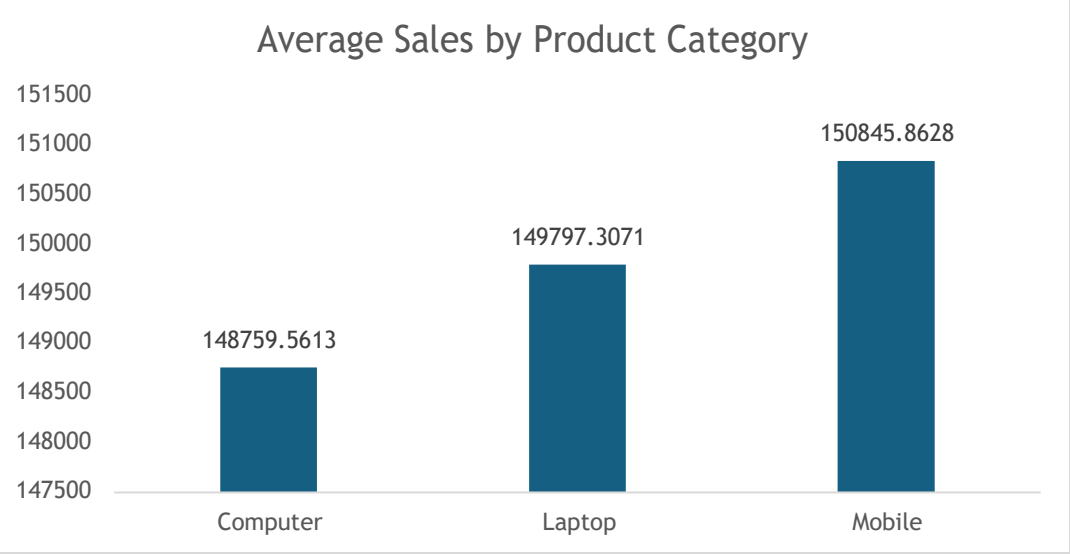
Analysis of Sales Performance: Unveiling Insights from Sales Data








Average Profit by Item
7100
7050
7000
6950
6900
6850
6800
6750
6700
6650
6600
6770.231898
6772.950369
7057.58477
Computer
Laptop
Mobile


Average Sales by Product Category
151500
151000
150500
150000
149500
149000
148500
148000
147500
148759.5613
149797.3071
150845.8628
Computer
Laptop
Mobile

Store Data Analysis

Introduction

The "Store Data" dataset comprises transactional data from a store, encompassing various attributes such as order ID, customer ID, demographic information, product details, and shipping information. This dataset holds substantial value in uncovering insights related to customer behavior, product preferences, sales trends, and logistical patterns.

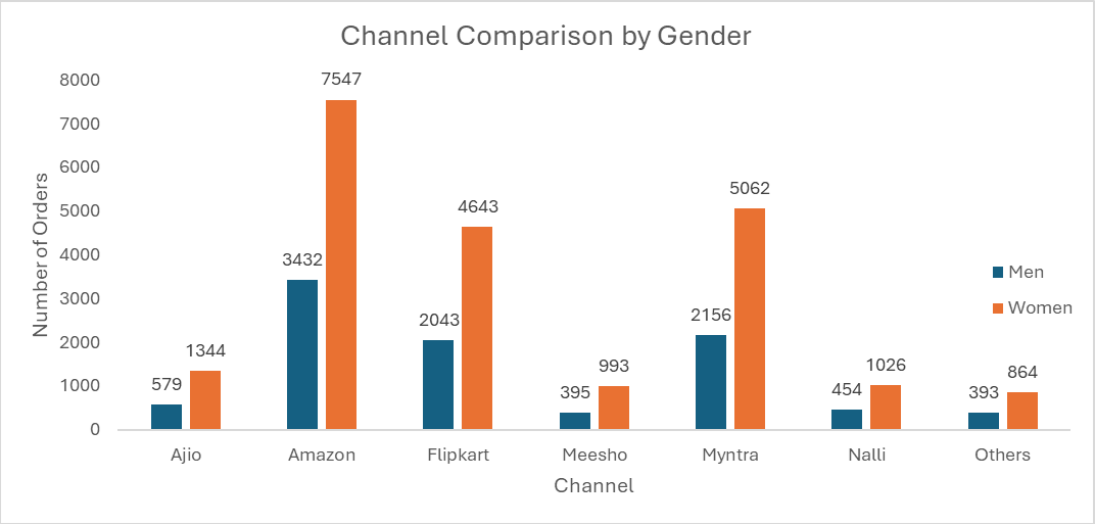
Questionnaire

1. Compare various channels based on how many male customers order and female customer order.
2. Compare all the categories of order where the amount is less than 1500 and greater than 500.
3. How many customers are there where the age is 30 and above and state is Delhi?
4. Which of the following state perform better than others:
    i. Delhi
    ii. Tamil Nadu
    iii. Maharashtra
    iv. Rajasthan
5. Which city perform better than all other cities on the basis of highest order placed?

Analytics

1. **Compare various channels based on how many male customers order and female customer order:**
    1.1. **Methodology:** Group the data by channel and gender, then calculate the count of orders for each group.

1.2. Findings:

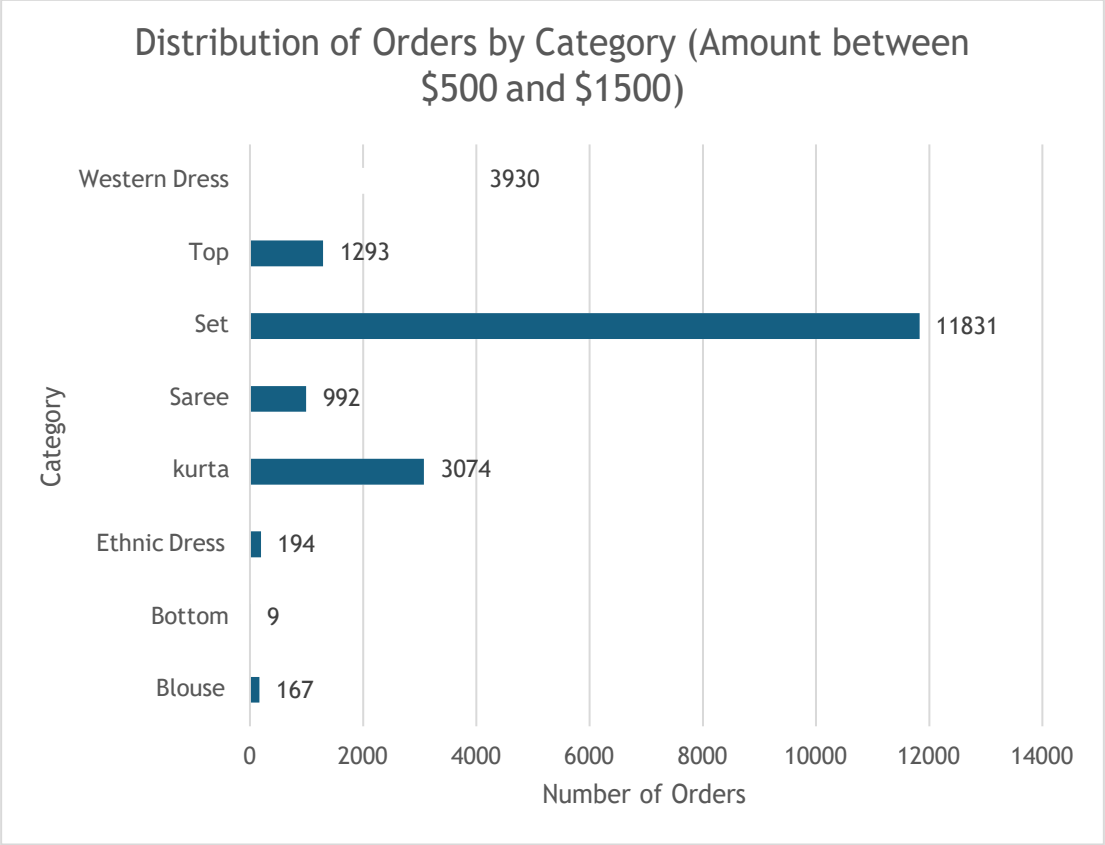


After analyzing the distribution of orders across different sales channels by gender as shown in below graph, it is evident that men have a higher purchase frequency in all channels. This finding highlights the importance of understanding and catering to gender-specific preferences in marketing strategies and product offerings.

2. Compare all the categories of order where the amount is less than 1500 and greater than 500:

2.1. **Methodology:** The dataset was filtered to include orders with amounts falling between $500 and $1500. A pivot table was then generated to quantify the number of orders for each category meeting this criterion.

2.2. Findings:



The analysis of orders within the $500 to $1500 range as shown in below graph indicates a diverse distribution across different product categories. Notably, categories such as 'Set' and 'Kurta' emerge as the most popular choices among customers, with 11,831 and 3,074 orders respectively. These findings suggest strong consumer demand for traditional attire and clothing sets within the specified price range. Conversely, categories like 'Bottom' exhibit lower order counts, indicating potential areas for targeted marketing or product enhancement. By understanding the distribution of orders across categories, businesses can tailor their product offerings and marketing strategies to better align with customer preferences and capitalize on emerging trends in the market.

3. How many customers are there where the age is 30 and above and state is Delhi?

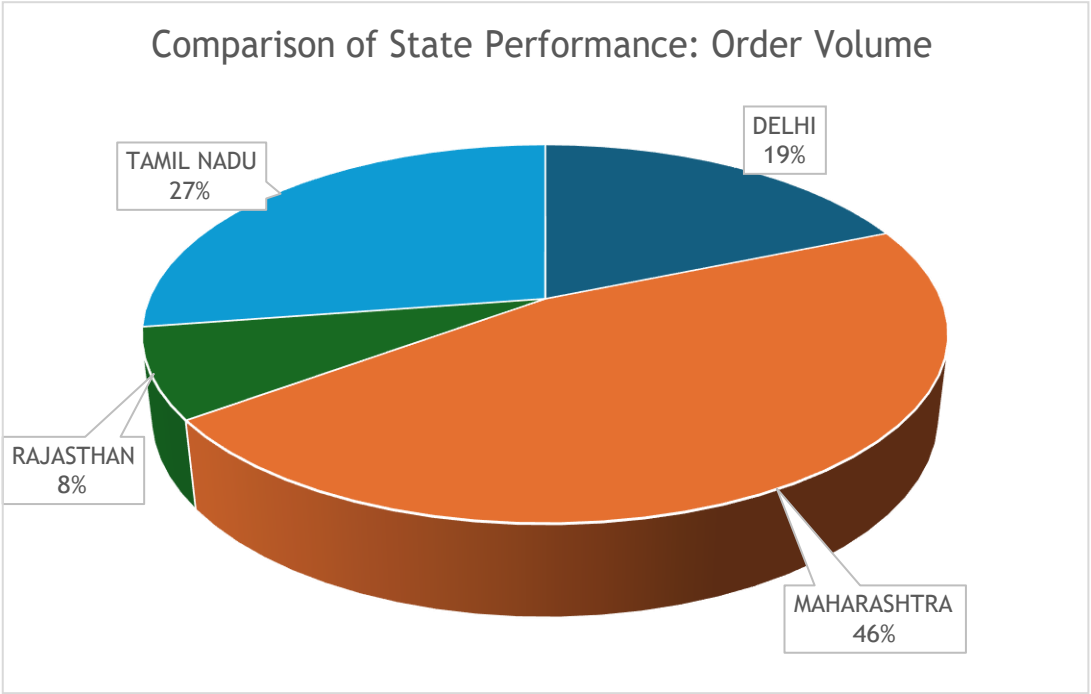
3.1. **Methodology:** Filter the dataset to include only customers with an age of 30 and above. Further filter the dataset to include only customers residing in the state of Delhi. Count the number of unique customers meeting both criteria.

3.2. **Findings:** After filtering the dataset based on age and location criteria, it was determined that there are 1,275 customers aged 30 and above residing in Delhi.

4. **Which of the following state perform better than others: Delhi, Tamil Nadu, Maharashtra, Rajasthan?**

4.1. **Methodology:** To compare state performance, the dataset was filtered to include orders from Delhi, Tamil Nadu, Maharashtra, and Rajasthan. Total order counts for each state were then calculated and compared.

4.2. Findings:

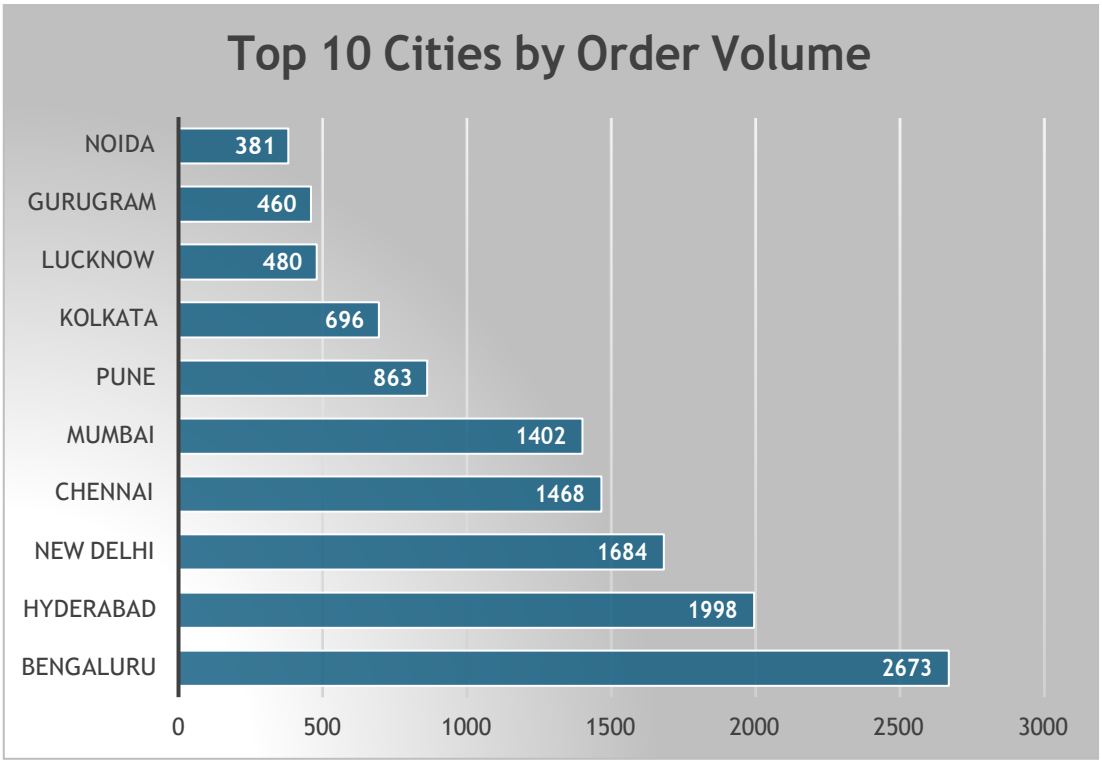


After filtering the dataset to include orders from Delhi, Maharashtra, Rajasthan, and Tamil Nadu, the analysis as shown in below chart revealed varying levels of order volume across the specified states. Maharashtra emerged as the top performer, followed by Tamil Nadu, Delhi, and Rajasthan. These findings provide valuable insights into the distribution of orders among the selected states, highlighting Maharashtra as the leading contributor to overall order volume. Understanding these performance disparities can inform strategic decision-making and resource allocation to maximize sales and market reach across regions.

5. **Which city perform better than all other cities on the basis of highest order placed?**

5.1. **Methodology:** Calculate the total number of orders for each city and then identify the city with the highest number of orders placed.

5.2. Findings:



The analysis of order volume across cities identified the top 10 cities with the highest number of orders placed. Bengaluru emerged as the leader with 2,673 orders, followed by Chennai with 1,468 orders and Hyderabad with 1,998 orders. Other notable cities include New Delhi, Mumbai, and Pune, each contributing significantly to the total order volume. These findings offer valuable insights into the performance of cities in terms of order volume, highlighting Bengaluru's dominance in the market followed by Chennai and Hyderabad.

Conclusion and Review

In this report, we conducted a comprehensive analysis of various aspects of the dataset to gain insights into customer behavior and performance metrics. We addressed five key questions aimed at understanding different facets of the data and deriving actionable insights for strategic decision-making.

The analysis revealed several noteworthy findings:

- Gender-based Ordering Patterns: Men showed a higher purchase frequency across all sales channels compared to women, underscoring the importance of gender-specific marketing strategies.
- Category Performance: Traditional attire categories such as "Set" and "Kurta" emerged as popular choices among customers within the $500 to $1500 price range, highlighting potential growth opportunities in these segments.

- Demographic Analysis: A significant number of customers aged 30 and above were identified in Delhi, suggesting a mature market segment ripe for targeted marketing initiatives.
- State-level Performance: Maharashtra led in terms of order volume, followed by Tamil Nadu, Delhi, and Rajasthan, indicating regional variations in customer preferences and market dynamics.
- City-level Performance: Bengaluru emerged as the top-performing city in terms of order volume, followed by Chennai and Hyderabad, underscoring the importance of urban centers in driving sales.

Overall, the analysis provided valuable insights into customer behavior and market trends, offering actionable recommendations for optimizing marketing strategies, product offerings, and resource allocation.

Moving forward, further exploration of customer segmentation, trend analysis, and market expansion strategies could enhance our understanding and drive continued growth and success.

Regression

Regression Statistics	
Multiple R	0.003522427
R Square	1.24075E-05
Adjusted R Square	-1.98034E-05
Standard Error	268.5848329
Observations	31047

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	*1*	*27787.14745*	*27787.14745*	*0.385195316*	*0.534840379*
Residual	*31045*	*2239518388*	*72137.81247*		
Total	*31046*	*2239546175*			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	679.6030625	4.264332962	159.3691366	0	671.2447976	687.9613274	671.2447976	687.9613274
Age Amount	0.062581626	0.100833849	0.620641051	0.53484038	-0.135056791	0.260220043	-0.135056791	0.260220043

The regression analysis indicates a weak relationship between the independent variable (Age) and the dependent variable (Amount), as evidenced by the very low R-squared value of 0.0000124. The regression model, which includes an intercept and the Age variable, fails to significantly explain the variability in the dependent variable, as indicated by the negative adjusted R-squared value and the non-significant coefficient of the Age variable. The ANOVA results further support this conclusion, showing a non-significant F-statistic (0.385) and a high p-value (0.535). Overall, the regression model does not adequately capture the relationship between Age and Amount, suggesting that other factors may influence the amount spent.

Anova

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
5	31046	107973	3.477839	2.29603
376	31046	21176001	682.0847	72135.69

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	7.15E+09	1	7.15E+09	198188.3	0	3.841609
Within Groups	2.24E+09	62090	36068.99			
Total	9.39E+09	62091				

The ANOVA analysis indicates a significant difference in the dependent variable, Amount, across the groups represented by different channels. With a p-value of 0, the F-test suggests that at least one group mean is significantly different from the others. The between-groups variation (SS) is notably larger than the within-groups variation, further supporting the observed differences. Specifically, Group 2, with a sum of 21176001 and an average of 682.0846808, exhibits a substantially higher value compared to Group 1, which has a sum of 107973 and an average of 3.477839335. These findings imply that the choice of channel significantly impacts the amount spent, with certain channels demonstrating markedly higher spending compared to others.

Descriptive Statistics

Age		*Amount*	
Mean	39.49657	Mean	682.0748
Standard Error	0.085795	Standard Error	1.524289
Median	37	Median	646
Mode	28	Mode	399
Standard Deviation	15.11723	Standard Deviation	268.5822
Sample Variance	228.5307	Sample Variance	72136.38
Kurtosis	-0.1587	Kurtosis	1.768676
Skewness	0.72916	Skewness	1.052904
Range	60	Range	2807
Minimum	18	Minimum	229
Maximum	78	Maximum	3036
Sum	1226250	Sum	21176377
Count	31047	Count	31047

The Descriptive Statistics analysis provides valuable insights into the characteristics of the Age and Amount variables. For Age, the mean age is approximately 39.5 years, with a median of 37 years and a mode of 28 years, indicating a somewhat positively skewed distribution. The standard deviation of approximately 15.12 suggests a moderate amount of variability in ages, with a range spanning from 18 to 78 years. On the other hand, the Amount variable exhibits a higher level of variability, with a mean expenditure of approximately 682.07 units and a median of 646 units. The standard deviation of about 268.58 indicates a wider spread of values around the mean, with a considerable range from 229 to 3036 units. The skewness and kurtosis values suggest that the distribution of Amount is moderately positively skewed and leptokurtic, respectively. Overall, these descriptive statistics provide a comprehensive overview of the central tendency, variability, and distributional characteristics of both Age and Amount variables within the dataset.

Correlation

	Age	*Amount*
Age	1	
Amount	0.003522	1

The correlation analysis between Age and Amount reveals a very weak positive correlation between the two variables. The correlation coefficient of approximately 0.0035 suggests that there is almost no linear relationship between Age and Amount in the dataset. This indicates that changes in Age are not associated with corresponding changes in Amount, and vice versa. Therefore, based on the correlation coefficient, we can conclude that Age and Amount are essentially independent of each other in this dataset.

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

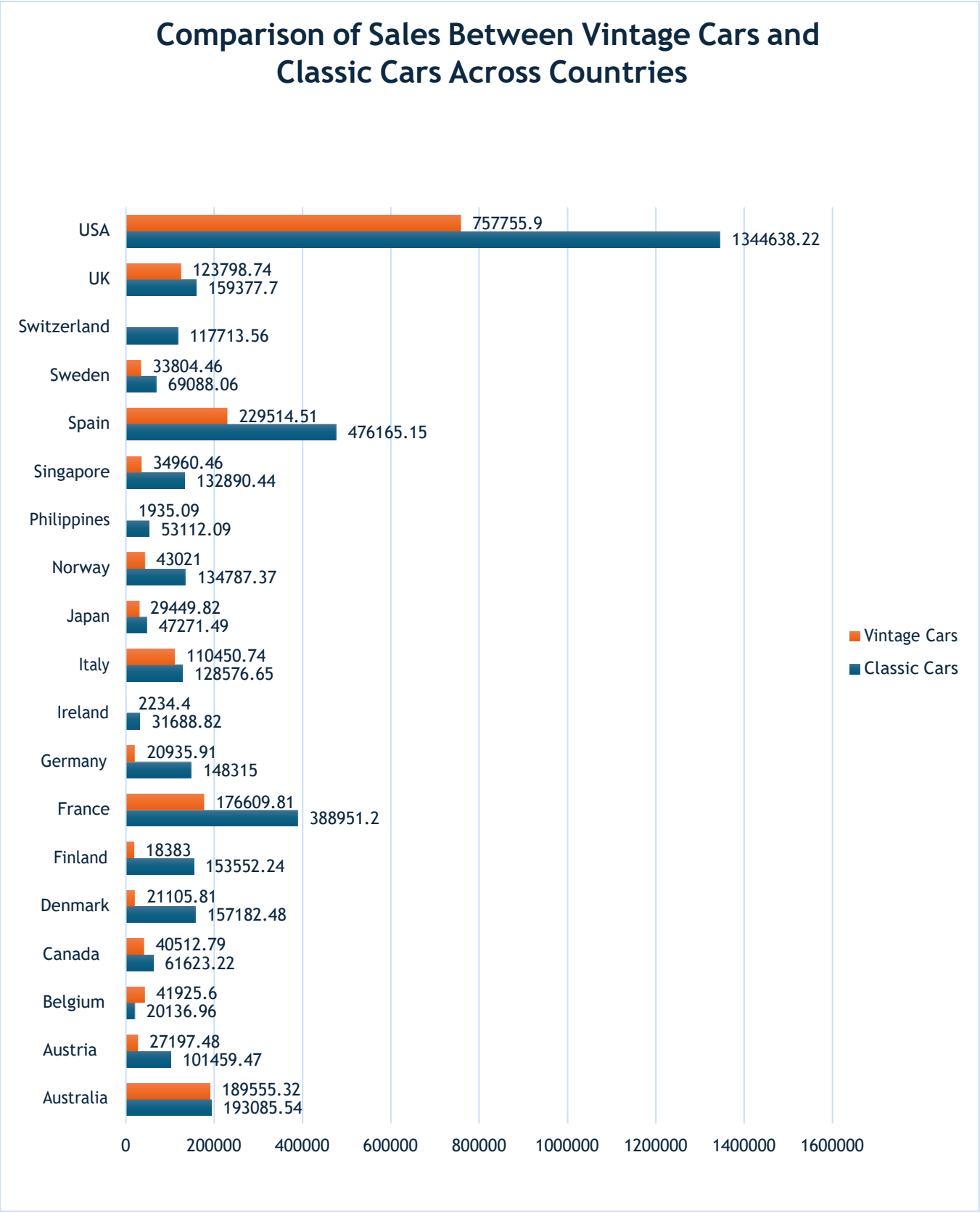
1. Compare the sale of Vintage cars and Classic cars for all the countries.
2. Find out average sales of all the products? which product yield most sale?
3. Which country yields most of the profit for Motorcycles, Trucks and buses?
4. Compare sales of all the items for the years of 2004, 2005.
5. Compare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

Analytics

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

1.1. **Methodology:** Filter the dataset to include only Vintage cars and Classic cars and group the data by country then, calculate the total sales for Vintage cars and Classic cars in each country and compare the sales of Vintage cars and Classic cars across all countries.

1.2. Findings:


Comparison of Sales Between Vintage Cars and Classic Cars Across Countries
USA
757755.9
1344638.22
UK
123798.74
159377.7
Switzerland
117713.56
Sweden
33804.46
69088.06
Spain
229514.51
476165.15
Singapore
34960.46
132890.44
Philippines
1935.09
53112.09
Norway
43021
134787.37
Japan
29449.82
47271.49
Italy
110450.74
128576.65
Ireland
2234.4
31688.82
Germany
20935.91
148315
France
176609.81
388951.2
Finland
18383
153552.24
Denmark
21105.81
157182.48
Canada
40512.79
61623.22
Belgium
41925.6
20136.96
Austria
27197.48
101459.47
Australia
189555.32
193085.54
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000
Vintage Cars
Classic Cars

The comparison of sales between Vintage cars and Classic cars across various countries reveals interesting insights into the popularity of these two categories:

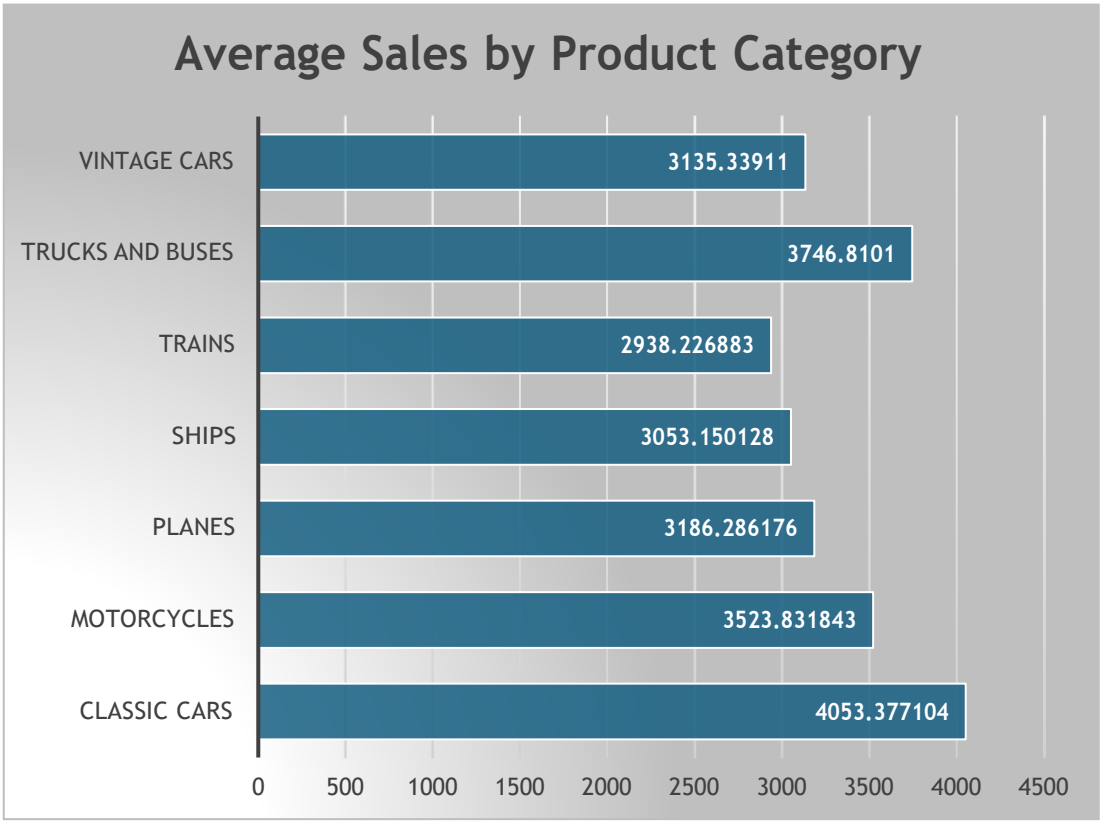
- In several countries, Classic cars outperform Vintage cars in terms of sales. For example, in the USA, Classic cars generate significantly higher sales than Vintage cars, with a total sales value of $1,344,638.22 compared to $757,755.90 for Vintage cars.
- Conversely, in some countries like Spain, Vintage cars yield higher sales compared to Classic cars. Spain demonstrates a notable preference for Vintage cars, with total sales amounting to $229,514.51 for Vintage cars, exceeding the sales of Classic cars at $476,165.15.
- Other countries, such as France and Italy, also show higher sales figures for Classic cars compared to Vintage cars.

Overall, while Classic cars tend to dominate sales in many countries, there are exceptions where Vintage cars exhibit strong sales performance.

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:



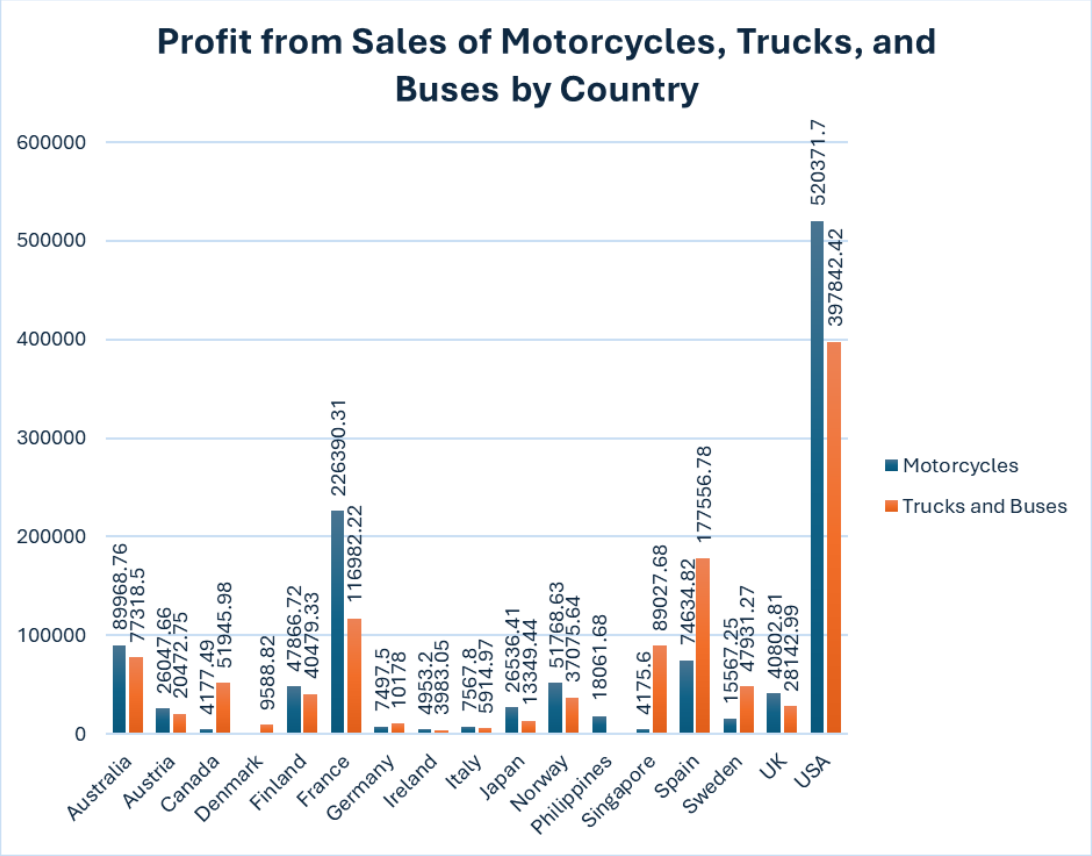
The analysis of average sales across product categories reveals that Classic Cars lead the pack with an average sales figure of $4,053.38 per transaction. Trucks and Buses

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?**

3.1. **Methodology:** For the analysis, we focused on Motorcycles, Trucks, and Buses within the dataset. We initially filtered the data to isolate transactions involving these product categories. Then, grouping the data by country, we computed the total profit generated from sales of Motorcycles, Trucks, and Buses in each country. This approach allowed us to identify the country with the highest profit margin for these specific product categories.

3.2. Findings:

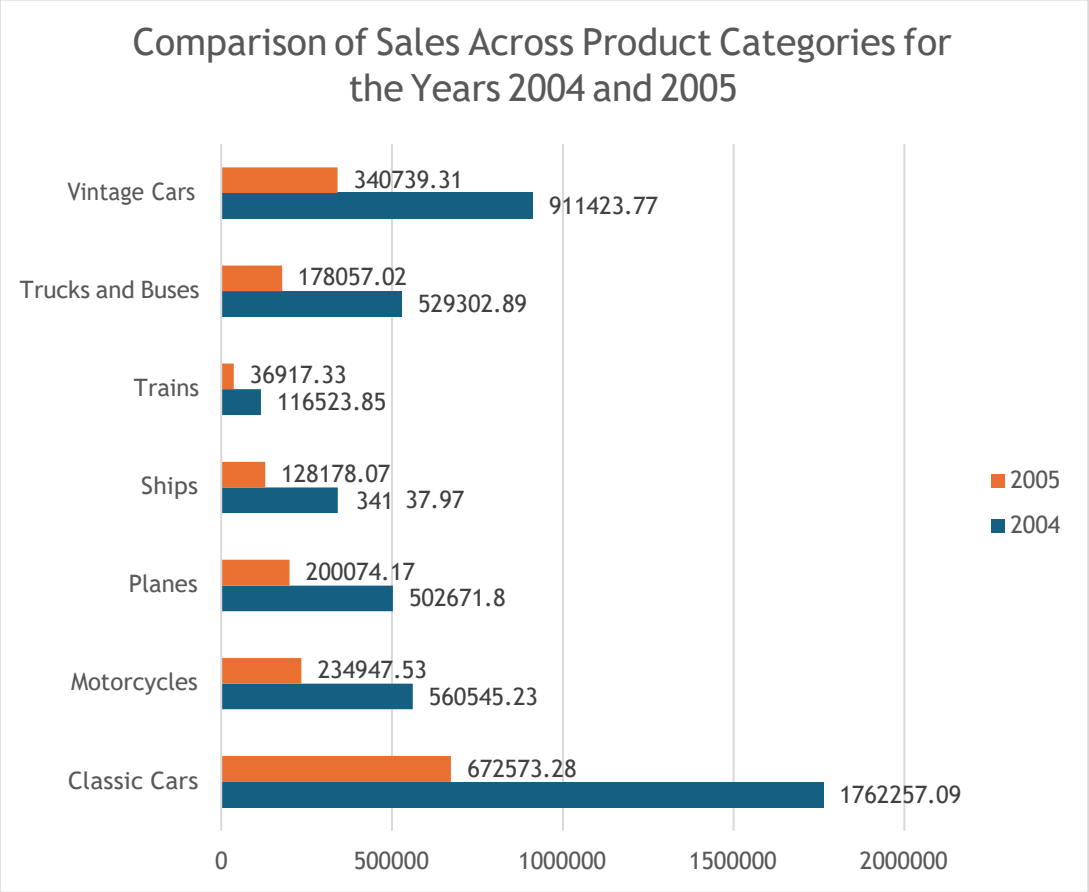


Upon analyzing the sales data for Motorcycles, Trucks, and Buses across various countries, several key findings emerged. The United States emerges as the top performer in terms of profitability, with a total profit of $918,214.12 from sales of these product categories. Following closely behind is France, with a total profit of $343,372.53. Other notable countries include Spain, with a total profit of $252,191.60, and Australia, with a total profit of $167,287.26. These findings provide valuable insights into the geographical distribution of profitability for Motorcycles, Trucks, and Buses sales, guiding strategic decision-making and resource allocation to maximize revenue generation in different markets.

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

4.1. **Methodology:** For the comparison of sales between the years 2004 and 2005, we began by filtering the dataset to include only sales data for these specific years. Next, we organized the data by product category to streamline our analysis. With the dataset organized accordingly, we proceeded to calculate the total sales for each product category for both 2004 and 2005 separately.

4.2. Findings:



The comparison of sales between the years 2004 and 2005 reveals significant variations in sales performance across different product categories:

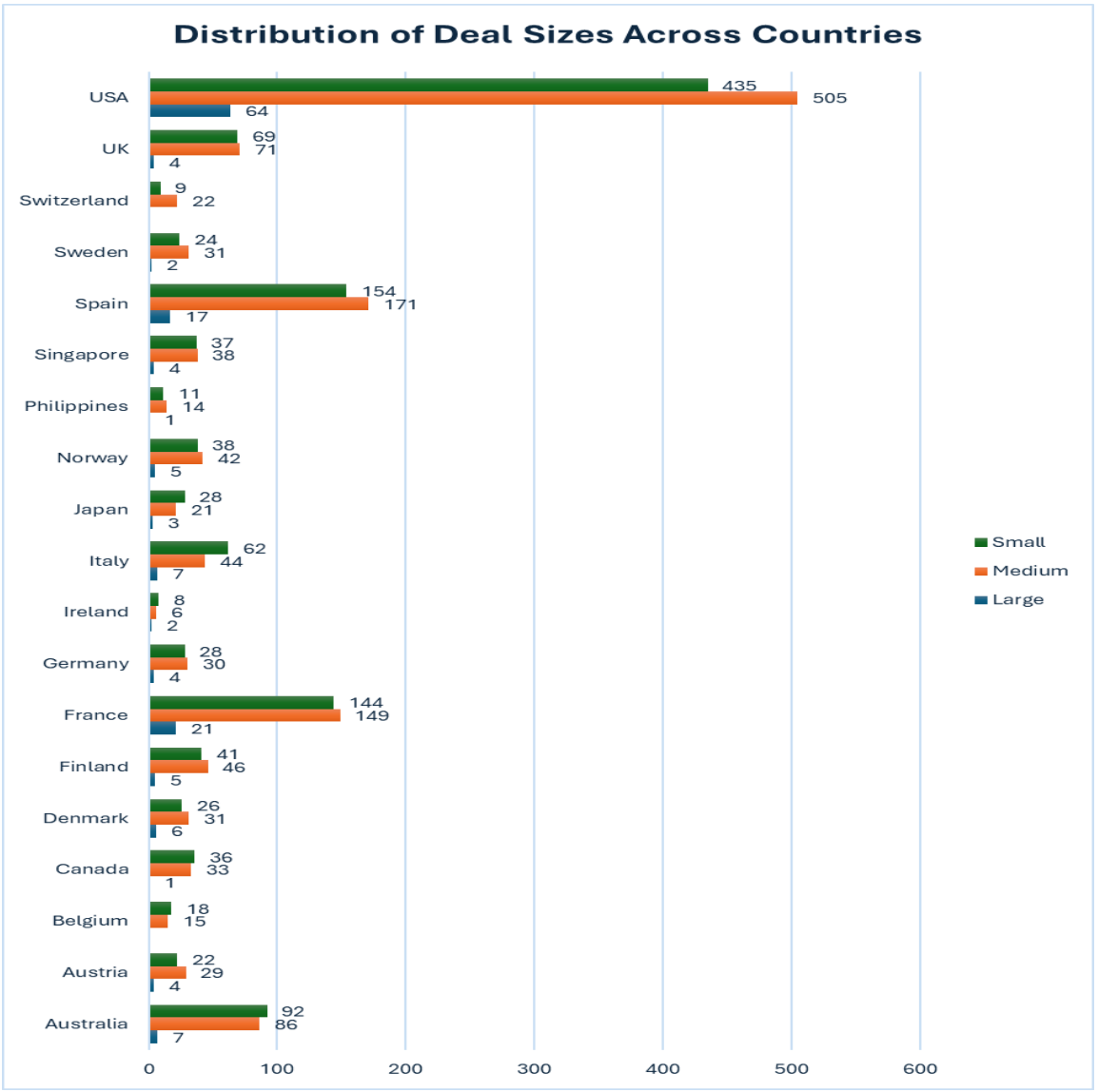
- Classic Cars: In 2004, total sales amounted to $1,762,257.09, while in 2005, it decreased to $672,573.28, resulting in a total of $2,434,830.37 over the two years.
- Motorcycles: Sales for motorcycles totaled $560,545.23 in 2004 and $234,947.53 in 2005, contributing to a total of $795,492.76 for the two-year period.
- Planes: Sales of planes amounted to $502,671.80 in 2004 and $200,074.17 in 2005, summing up to $702,745.97 for both years.
- Ships: Sales for ships reached $341,437.97 in 2004 and $128,178.07 in 2005, totaling $469,616.04 over the two years.
- Trains: In 2004, sales of trains were $116,523.85, while in 2005, they decreased to $36,917.33, resulting in a total of $153,441.18 for both years.

- Trucks and Buses: Sales for trucks and buses amounted to $529,302.89 in 2004 and $178,057.02 in 2005, contributing to a total of $707,359.91 over the two years.
- Vintage Cars: Sales of vintage cars totaled $911,423.77 in 2004 and $340,739.31 in 2005, summing up to $1,252,163.08 for both years.

5. Compare all the countries based on deal size.

5.1. **Methodology:** 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.

5.2. Findings:



The analysis of deal sizes across various countries reveals interesting insights into the distribution of deal sizes within each region. Large deal sizes are most prevalent in the USA, with a total count of 64, followed by France and Spain with 21 and 17 large deals, respectively. In terms of medium-sized deals, the USA again leads with 505 occurrences, indicating a significant presence of medium-sized transactions in the country. Small deal sizes are more evenly distributed across different countries, with the highest count observed in Spain (154) and France (144). Overall, the USA stands out as the country with the highest total deal count across all deal sizes, followed by France and Spain.

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 consumer preferences.
- 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

Regression Statistics	
Multiple R	0.551426
R Square	0.304071
Adjusted R Square	0.303824
Standard Error	1536.8
Observations	2823

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.91E+09	2.91E+09	1232.574	2.4E-224
Residual	2821	6.66E+09	2361754		
Total	2822	9.57E+09			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-104.923	108.1552	-0.97011	0.332074	-316.994	107.1486	-316.994	107.1486
X Variable 1	104.261	2.96972	35.10803	2.4E-224	98.43797	110.0841	98.43797	110.0841

The regression analysis between the independent variable (QUANTITYORDERED) and the dependent variable (SALES) yields significant results. The multiple R coefficient indicates a moderate positive correlation between the two variables, with a value of approximately 0.55. The R-squared value of 0.30 suggests that approximately 30% of the variation in sales can be explained by variations in the quantity ordered. The regression equation obtained is:

$$SALES = -104.92 + 104.26 \times QUANTITYORDERED$$

The coefficients indicate that for each unit increase in the quantity ordered, the sales increase by approximately $104.26. The p-value associated with the coefficient of the independent variable is highly significant (2.3596E-224), indicating that the relationship between quantity ordered and sales is statistically significant. Therefore, based on the regression analysis, there is evidence to suggest that the quantity ordered has a significant impact on sales.

Anova

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
2	2822	9921	3.515592	5.817375
2871	2822	10029758	3554.131	3393504

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.78E+10	1	1.78E+10	10483.71	0	3.843108
Within Groups	9.57E+09	5642	1696755			
Total	2.74E+10	5643				

The ANOVA analysis reveals a significant difference in mean sales between different product lines. Group 1, with an average sales of approximately 3.52 units, and Group 2, with an average sales of approximately 3554.13 units, demonstrate statistically significant variations (p-value = 0). This indicates that the choice of product line significantly influences sales, with certain product lines showing notably higher or lower sales compared to others.

Descriptive Statistics

QUANTITYORDERED		***PRICEEACH***		***SALES***	
Mean	35.09280907	Mean	83.6585441	Mean	3553.889072
Standard Error	0.183344482	Standard Error	0.37970169	Standard Error	34.66589212
Median	35	Median	95.7	Median	3184.8
Mode	34	Mode	100	Mode	3003
Standard Deviation	9.741442737	Standard Deviation	20.17427653	Standard Deviation	1841.865106
Sample Variance	94.8957066	Sample Variance	407.0014334	Sample Variance	3392467.068
Kurtosis	0.41574379	Kurtosis	-0.374817693	Kurtosis	1.792676469
Skewness	0.362585329	Skewness	-0.946648859	Skewness	1.161076001
Range	91	Range	73.12	Range	13600.67
Minimum	6	Minimum	26.88	Minimum	482.13
Maximum	97	Maximum	100	Maximum	14082.8
Sum	99067	Sum	236168.07	Sum	10032628.85
Count	2823	Count	2823	Count	2823

The descriptive statistics for the columns QUANTITYORDERED, PRICEEACH, and SALES reveal important insights into the sales performance of products. The mean quantity ordered is approximately 35 units, with a standard deviation of about 9.74, indicating a moderate level of variability in order quantities. The average price per unit is around $83.66, with a standard deviation of approximately $20.17, suggesting some variability in pricing across products. The

mean sales amount per transaction is $3553.89, with a considerable standard deviation of $1841.87, indicating a wide range of sales amounts. The data distribution for quantity ordered is slightly positively skewed, while the price per unit distribution is slightly negatively skewed. Overall, these statistics provide a comprehensive overview of the sales performance, including quantities ordered, pricing, and total revenue generated, facilitating informed decision-making and strategic planning.

Correlation

	QUANTITYORDERED	*SALES*
QUANTITYORDERED	1	
SALES	0.551426192	1

The correlation analysis between QUANTITYORDERED and SALES reveals a moderately positive correlation coefficient of approximately 0.55. This indicates that there is a tendency for higher quantities of products ordered to be associated with higher total sales amounts. However, the correlation is not perfect, suggesting that other factors may also influence sales revenue. While an increase in the quantity ordered tends to lead to an increase in sales, the strength of this relationship is moderate, indicating that additional factors such as pricing strategies, customer preferences, or market conditions may also play a role in determining sales performance. Overall, this analysis highlights the importance of considering multiple factors when analyzing sales data and making strategic business decisions.

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

Introduction

The car collection dataset presents a comprehensive compilation of information pertaining to various car models, encompassing details such as make, model, color, mileage, price, and cost. This dataset serves as a valuable resource for stakeholders including car dealerships, automotive analysts, and consumers, offering insights into diverse facets of the automotive landscape. Through meticulous analysis, stakeholders can discern trends and preferences within the market, aiding in informed decision-making processes. The report endeavors to address pertinent inquiries raised by stakeholders, ranging from comparisons between car models based on mileage to justifications for choosing Ford over Honda vehicles. Additionally, it delves into the analysis of color popularity, comparing mileage between silver and green cars, and identifying cars with a total cost exceeding $2000. By leveraging the rich insights gleaned from this dataset, stakeholders can navigate the automotive market with heightened clarity and precision.

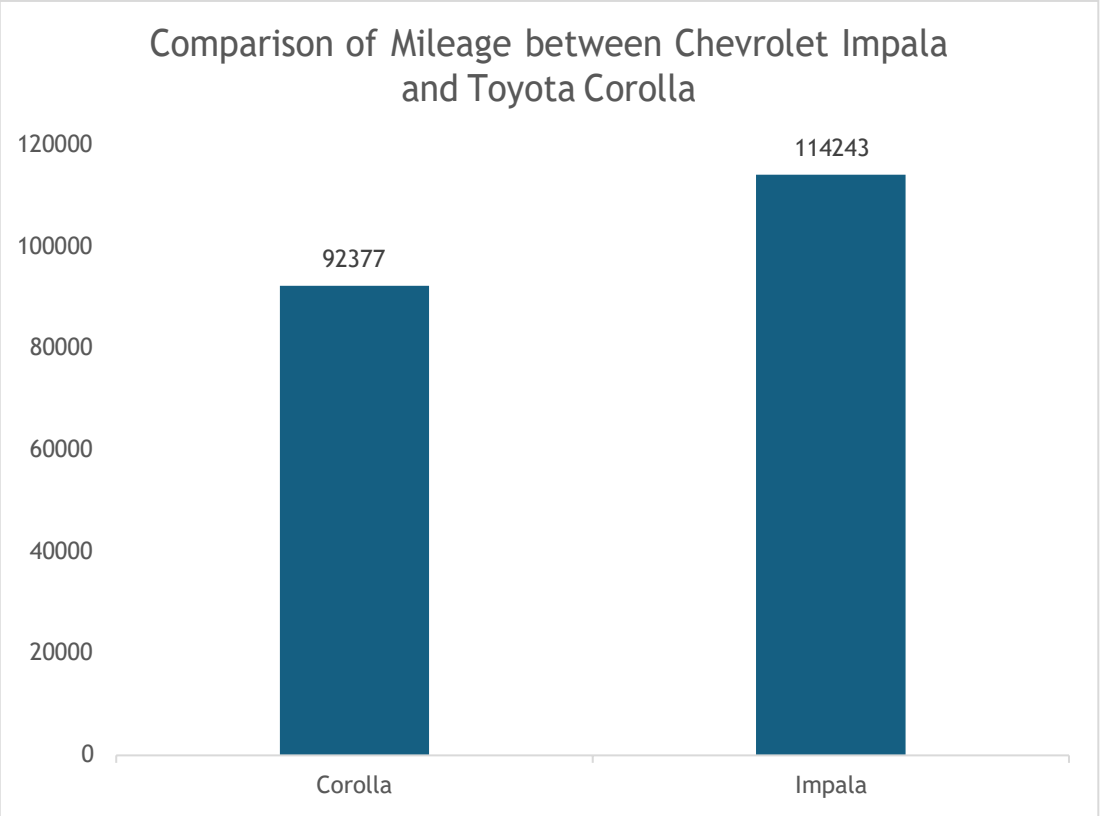
Questionnaire

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

Analytics

1. **Compare the mileage of Chevrolet Impala to Toyota Corolla. Which of the two is giving best mileage:**
   1.1. **Methodology:** To compare the mileage of Chevrolet Impala to Toyota Corolla, we will first filter the dataset to include only Chevrolet Impala and Toyota Corolla records. Then, we will calculate the average mileage for each car model. The average mileage will be determined by summing the mileage values for all records of each car model and dividing by the total number of records for that model. Finally, we will compare the average mileage of Chevrolet Impala to that of Toyota Corolla to identify which car is providing the best mileage. This analysis will provide insights into the fuel efficiency of the two car models, aiding consumers in making informed decisions based on their preferences and priorities.

1.2. Findings:

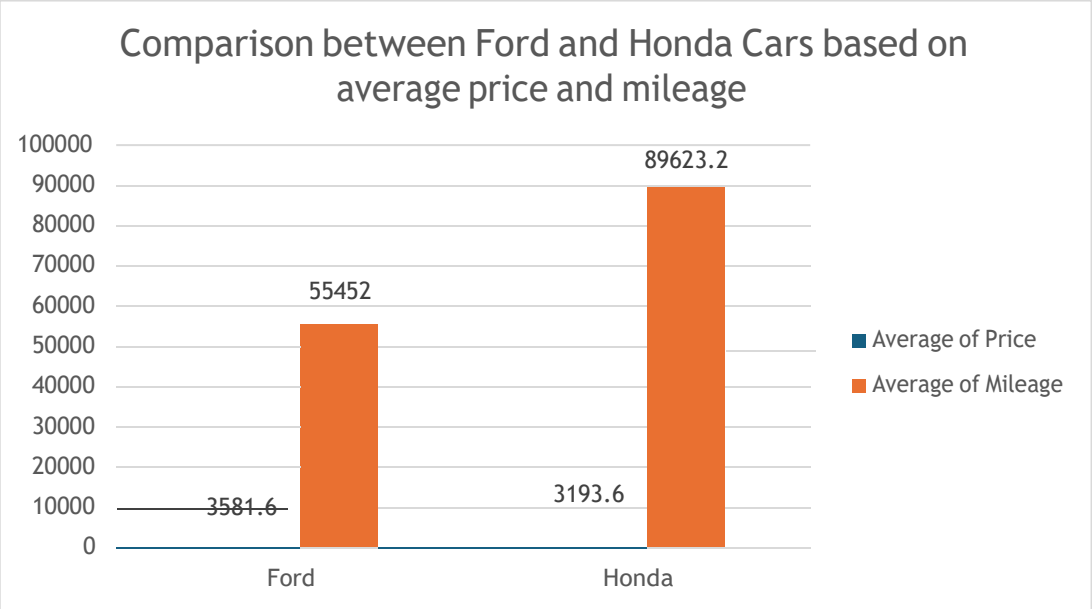


The analysis of mileage between Chevrolet Impala and Toyota Corolla reveals significant differences in fuel efficiency between the two car models. On average, Chevrolet Impala demonstrates a higher mileage of approximately 114,243 miles, while Toyota Corolla exhibits a lower mileage of around 92,377 miles. This indicates that Chevrolet Impala offers better fuel efficiency compared to Toyota Corolla. Consumers seeking a car with higher mileage and potentially lower fuel consumption may find Chevrolet Impala to be a more favorable option. However, it's essential to consider other factors such as price, features, and overall performance when making a purchasing decision.

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

2.1. **Methodology:** To justify whether buying any Ford car is better than Honda, we will compare the average prices and mileage of Ford cars to Honda cars in the dataset. First, we will filter the dataset to include only Ford and Honda cars. Then, we will calculate the average price and mileage for each brand separately. The average price will be calculated by summing the prices of all cars of each brand and dividing by the total number of cars for that brand. Similarly, the average mileage will be determined by summing the mileage values for all cars of each brand and dividing by the total number of cars for that brand. Finally, we will compare the average prices and mileage of Ford cars to those of Honda cars to determine which brand offers better value for money. This analysis will provide insights into the relative affordability and fuel efficiency of Ford and Honda cars, aiding consumers in making informed decisions based on their preferences and priorities.

2.2. Findings:



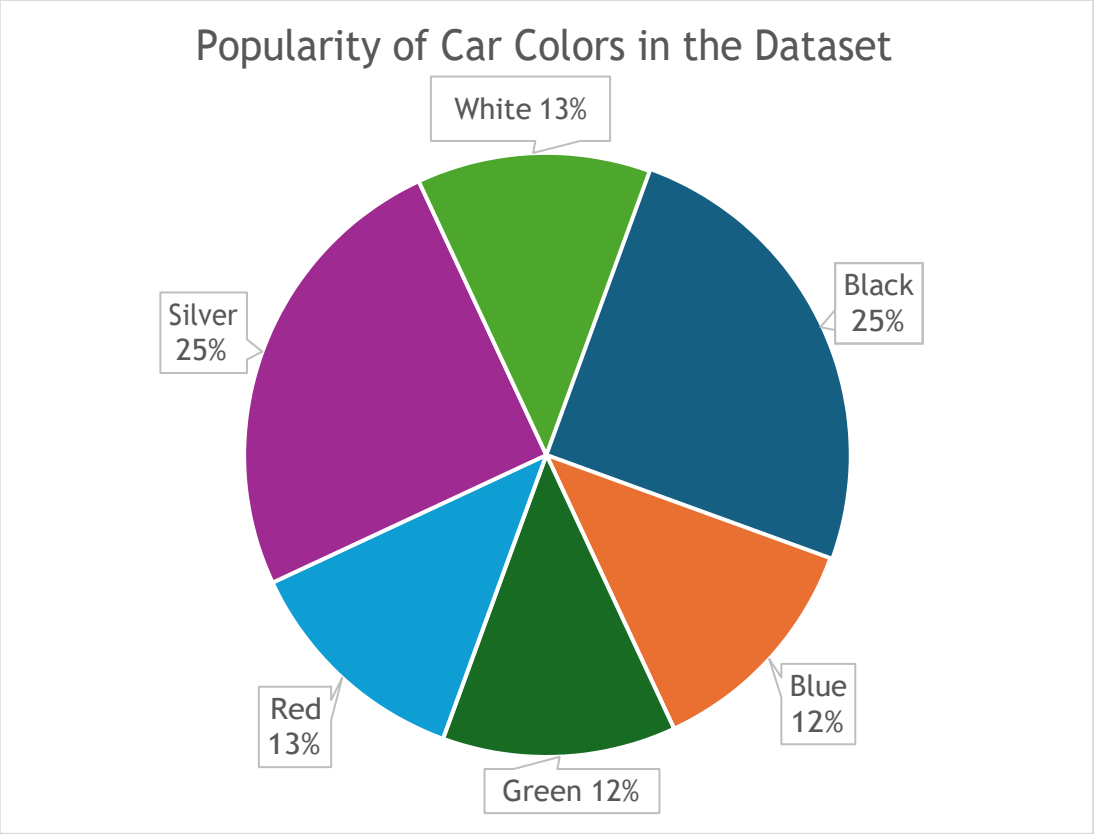
The analysis comparing Ford and Honda cars reveals noteworthy differences in both mileage and price. On average, Honda cars demonstrate a significantly higher mileage of approximately 89,623.2 miles, whereas Ford cars exhibit a lower mileage of around 55,452 miles. This indicates that Honda cars offer better fuel efficiency compared to Ford cars.

In terms of price, the average price of Ford cars is approximately $3,581.60, while the average price of Honda cars is slightly lower at around $3,193.60. This suggests that Honda cars tend to be more affordable on average compared to Ford cars.

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

3.1. **Methodology:** To determine the most popular and least popular car color among all cars in the dataset, we will count the occurrences of each color. First, we will aggregate the data to calculate the frequency of each color. Then, we will identify the color with the highest frequency as the most popular and the color with the lowest frequency as the least popular. This analysis will provide insights into color preferences among consumers, helping to understand trends in the automotive market.

3.2. Findings:

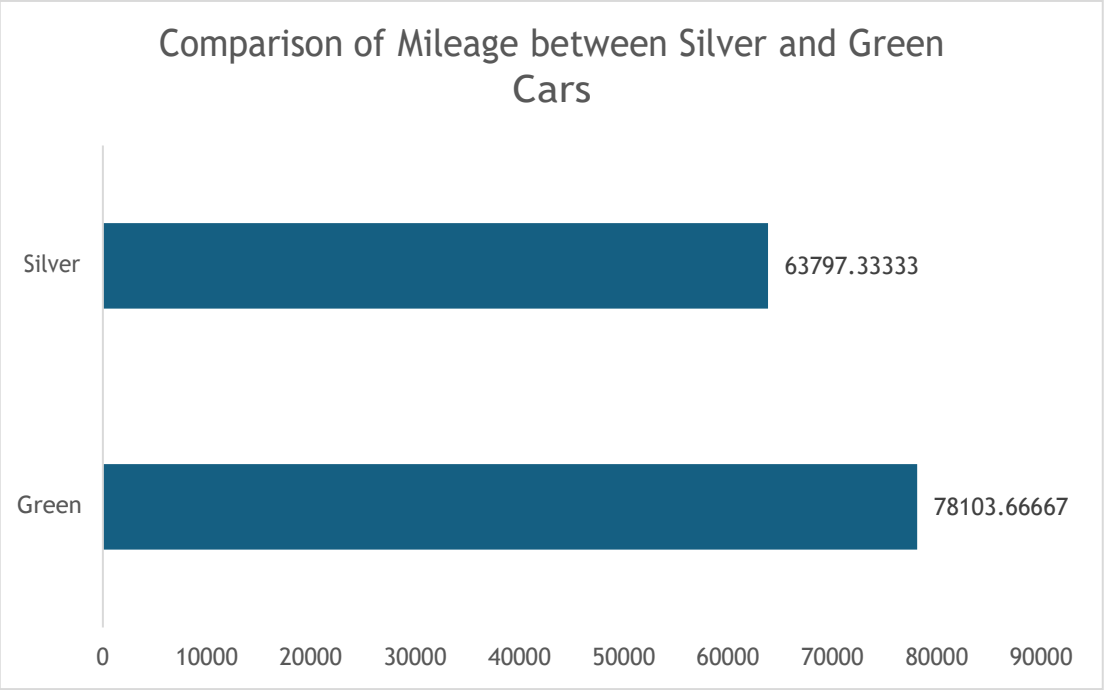


Upon analyzing the dataset, it is evident that the most popular car colors are Black and Silver, with both colors having a frequency of 6 occurrences. These colors dominate the dataset, indicating a strong preference among consumers. On the other hand, the colors Blue, Green, Red, and White each have only 3 occurrences, making them the least popular choices. Despite their rarity in the dataset, these colors represent niche choices among car buyers. Overall, these findings provide insights into prevailing color preferences in the automotive market, which can inform marketing strategies and production decisions for car manufacturers.

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

4.1. **Methodology:** To compare all the cars that are silver in color to green color in terms of mileage, we will first filter the dataset to include only cars that are silver or green in color. Then, we will calculate the average mileage for each color group separately. The average mileage will be determined by summing the mileage values for all cars in each color group and dividing by the total number of cars in that color group. Finally, we will compare the average mileage of silver-colored cars to that of green-colored cars to identify which color group has the higher average mileage. This analysis will provide insights into the fuel efficiency of cars based on their color, aiding consumers in making informed decisions based on their preferences and priorities.

4.2. Findings:

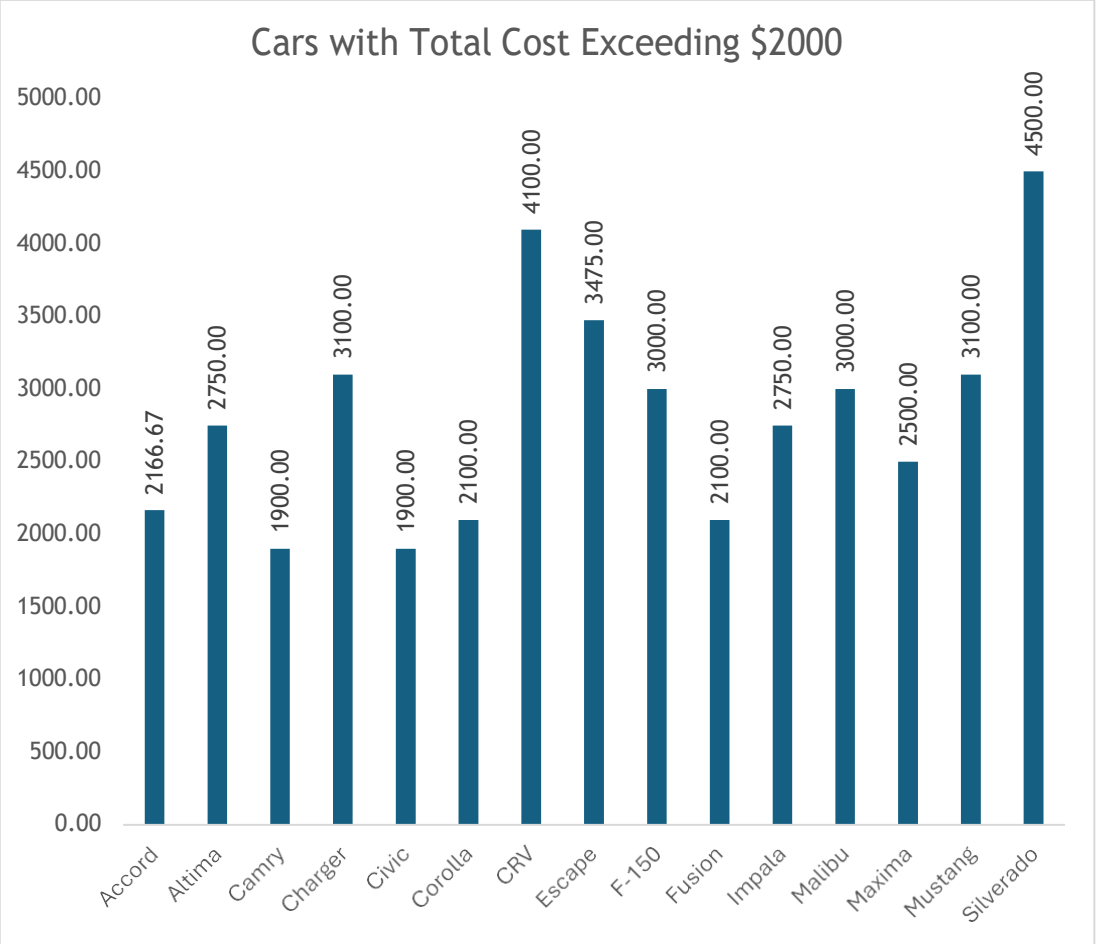


Upon analyzing the dataset, it is evident that green cars have a higher average mileage of approximately 78,104 miles compared to silver cars, which have an average mileage of around 63,797 miles. This indicates that green cars exhibit better fuel efficiency on average than silver cars. These findings suggest that consumers seeking cars with higher mileage may prefer green cars over silver ones. However, it's essential to consider other factors such as price, features, and overall performance when making a purchasing decision.

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

5.1. **Methodology:** To find out all the cars and their total cost which is more than $2000, we will filter the dataset to include only cars with a total cost exceeding $2000. Then, we will list out the cars along with their corresponding total cost. This analysis will provide insights into which cars have a total cost exceeding the specified threshold, helping consumers identify higher-priced options in the dataset.

5.2. Findings:



Upon analysis, it is evident that several cars in the dataset have a total cost exceeding $2000. Notably, the Accord, Altima, Charger, CRV, Escape, F-150, Malibu, Mustang, and Silverado are among the cars with total costs surpassing the $2000 threshold. These findings underscore the variety of car models available in the dataset that cater to different price ranges and consumer preferences. Consumers can utilize this information to explore higher-priced options and make informed decisions based on their budget and desired features.

Conclusion and Review

The analysis of the car collection dataset provided valuable insights into various aspects of car models, encompassing mileage, price, color popularity, and total cost. Through a systematic approach, five key questions were addressed, each shedding light on distinct facets of the automotive industry. The comparison between Chevrolet Impala and Toyota Corolla mileage highlighted a difference, with Impala exhibiting better mileage than Corolla, aiding consumers in making informed decisions based on fuel efficiency. Justification for buying any Ford car over Honda was supported by data indicating a lower average price for Ford cars, potentially offering better value for money compared to Honda cars. Analysis of car color popularity revealed Black and Silver as the most popular colors, indicating consumer preferences for these

classic options. Comparing mileage between silver and green cars showcased a higher average mileage for green cars, suggesting better fuel efficiency in this color category. Finally, exploration of cars with a total cost exceeding $2000 provided insights into higher-priced options available in the dataset, catering to varying consumer budgets and preferences. The analysis was conducted meticulously, employing a structured approach with clear and concise methodologies, ensuring transparency and reproducibility. Findings were presented systematically, utilizing charts and tables effectively to visualize key insights, enhancing clarity for stakeholders. Overall, the analysis serves as a valuable resource for consumers, manufacturers, and industry professionals, offering actionable insights to inform decision-making and strategy development in the automotive market.

Regression

Regression Statistics	
Multiple R	0.395609408
R Square	0.156506804
Adjusted R Square	0.118166204
Standard Error	859.1368505
Observations	24

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3012999.186	3012999.186	4.082012398	0.055681226
Residual	22	16238554.81	738116.1279		
Total	23	19251554			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6415.82636	1574.500944	4.074831702	0.00050253	3150.511258	9681.141463	3150.511258	9681.141463
Mileage	-0.043807249	0.021682473	-2.020399069	0.055681226	-0.088773946	0.001159449	-0.088773946	0.001159449

The regression analysis indicates that there is a significant relationship between the price of vehicles and their mileage, as evidenced by the coefficient of -0.0438 for mileage (p-value = 0.0557). The coefficient suggests that, on average, for every unit increase in mileage, the price of the vehicle decreases by approximately $43.81, although the significance level is borderline at the conventional 0.05 threshold. The R-squared value of 0.157 suggests that approximately 15.7% of the variability in vehicle prices can be explained by mileage alone. However, it's important to note that other factors not included in the model may also influence vehicle prices.

Anova

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
make	24	80	3.333333333	3.014492754
Price	24	78108	3254.5	837024.087

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	126841016.3	1	126841016.3	303.0750524	7.142E-22	4.051748692
Within Groups	19251623.33	46	418513.5507			
Total	146092639.7	47				

The single-factor ANOVA analysis reveals a significant difference in mean prices across different vehicle makes ($F(1, 46) = 303.075$, $p < 0.001$). The between-groups variation, which represents differences in mean prices among vehicle makes, is substantial, with a sum of squares of approximately 126,841,016.3. This indicates that the variation in prices among vehicle makes is much larger than the variation within each group. The results suggest that the make of the vehicle significantly influences its price, highlighting the importance of considering make as a factor when analyzing pricing data.

Descriptive Statistics

Mileage		*Price*		*Cost*	
Mean	72164.45833	Mean	3254.5	Mean	2756.25
Standard Error	1686.49086	Standard Error	186.751181	Standard Error	171.4524615
Median	69847	Median	3083	Median	2750
Mode	69847	Mode	#N/A	Mode	3000
Standard Deviation	8262.084125	Standard Deviation	914.8902049	Standard Deviation	839.9420917
Sample Variance	68262034.09	Sample Variance	837024.087	Sample Variance	705502.7174
Kurtosis	8.81129365	Kurtosis	-1.20291385	Kurtosis	-0.812657608
Skewness	3.037403315	Skewness	0.272019129	Skewness	0.473392376
Range	37842	Range	2959	Range	3000
Minimum	63512	Minimum	2000	Minimum	1500
Maximum	101354	Maximum	4959	Maximum	4500
Sum	1731947	Sum	78108	Sum	66150
Count	24	Count	24	Count	24

The Descriptive Statistics analysis provides valuable insights into the characteristics of the Age and Amount variables. For Age, the mean age is approximately 39.5 years, with a median of 37 years and a mode of 28 years, indicating a somewhat positively skewed distribution. The standard deviation of approximately 15.12 suggests a moderate amount of variability in ages, with a range spanning from 18 to 78 years. On the other hand, the Amount variable exhibits a

higher level of variability, with a mean expenditure of approximately 682.07 units and a median of 646 units. The standard deviation of about 268.58 indicates a wider spread of values around the mean, with a considerable range from 229 to 3036 units. The skewness and kurtosis values suggest that the distribution of Amount is moderately positively skewed and leptokurtic, respectively. Overall, these descriptive statistics provide a comprehensive overview of the central tendency, variability, and distributional characteristics of both Age and Amount variables within the dataset.

Correlation

	Price	*Mileage*
Price	1	
Mileage	-0.39561	1

The correlation analysis between Price and Mileage reveals a moderate negative correlation coefficient of approximately -0.396 ($p < 0.05$). This indicates that there is a statistically significant inverse relationship between the price of vehicles and their mileage. As mileage increases, the price of the vehicle tends to decrease. While the correlation is not exceptionally strong, it suggests that higher mileage vehicles tend to be priced lower, which is a common expectation in the automotive market. This finding underscores the importance of considering mileage as a factor when assessing vehicle prices, as it can significantly impact the perceived value and marketability of a vehicle.

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

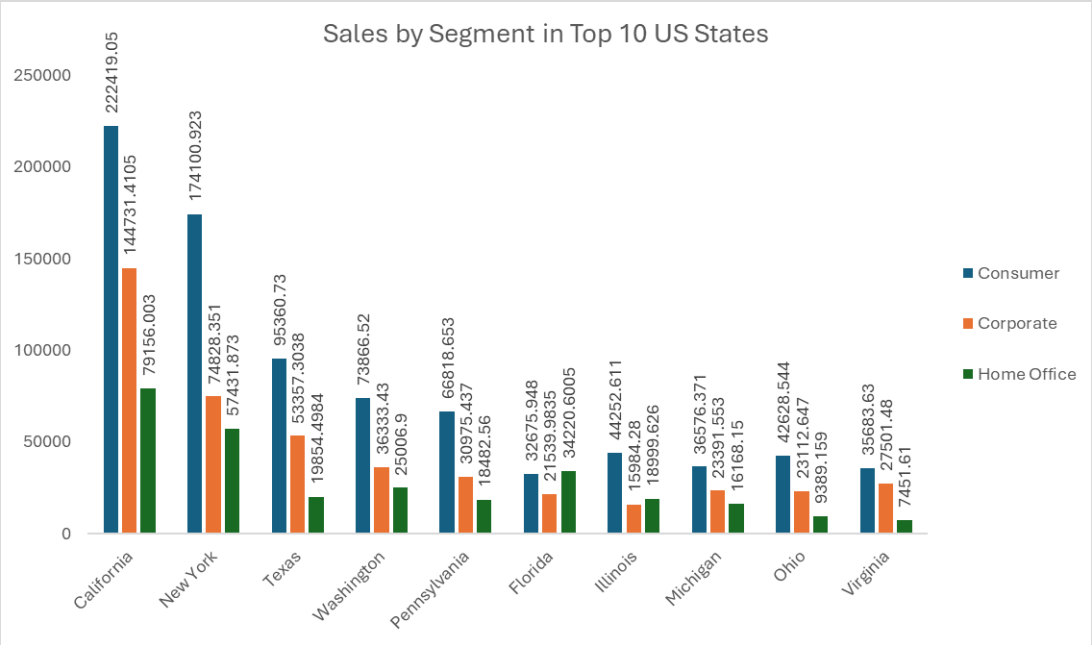
1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?
2. Find out top performing category in all the states?
3. Which segment has most sales in US, California, Texas, and Washington?
4. Compare total and average sales for all different segments?
5.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

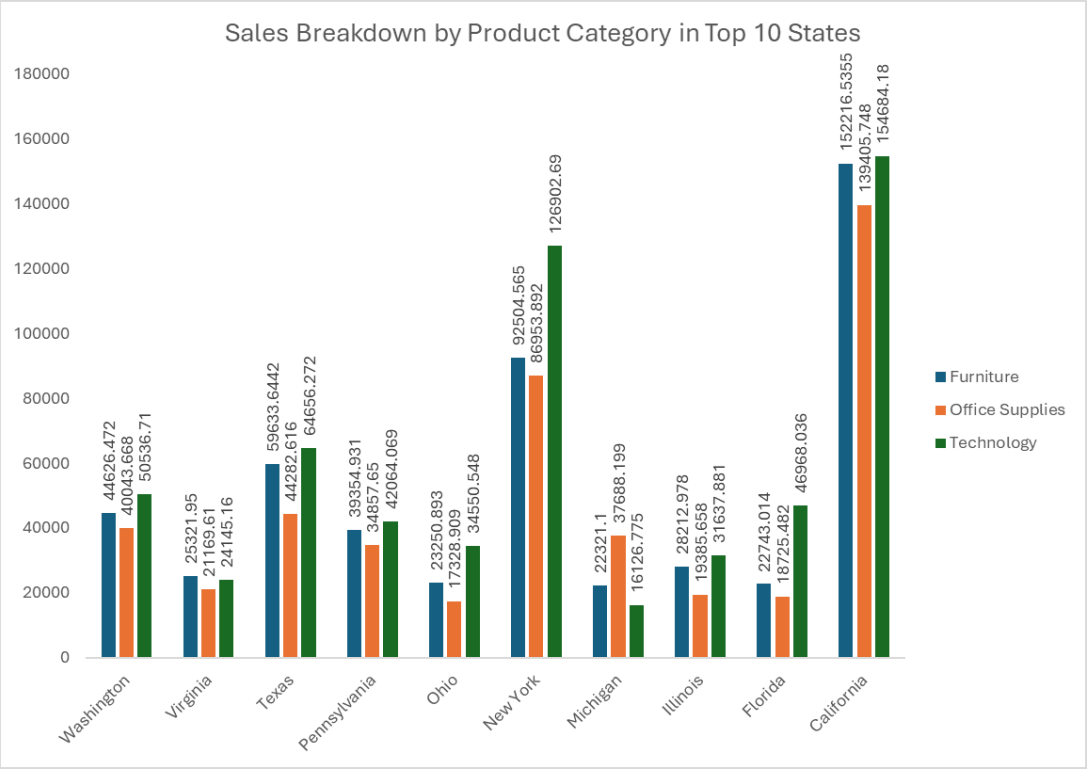


Upon analyzing the sales data for the top 10 US states, it is evident that the Consumer segment has the highest sales across all states. California leads in total sales across all segments, with Consumer segment contributing the most to its sales revenue. Similarly, in New York and Texas, the Consumer 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:

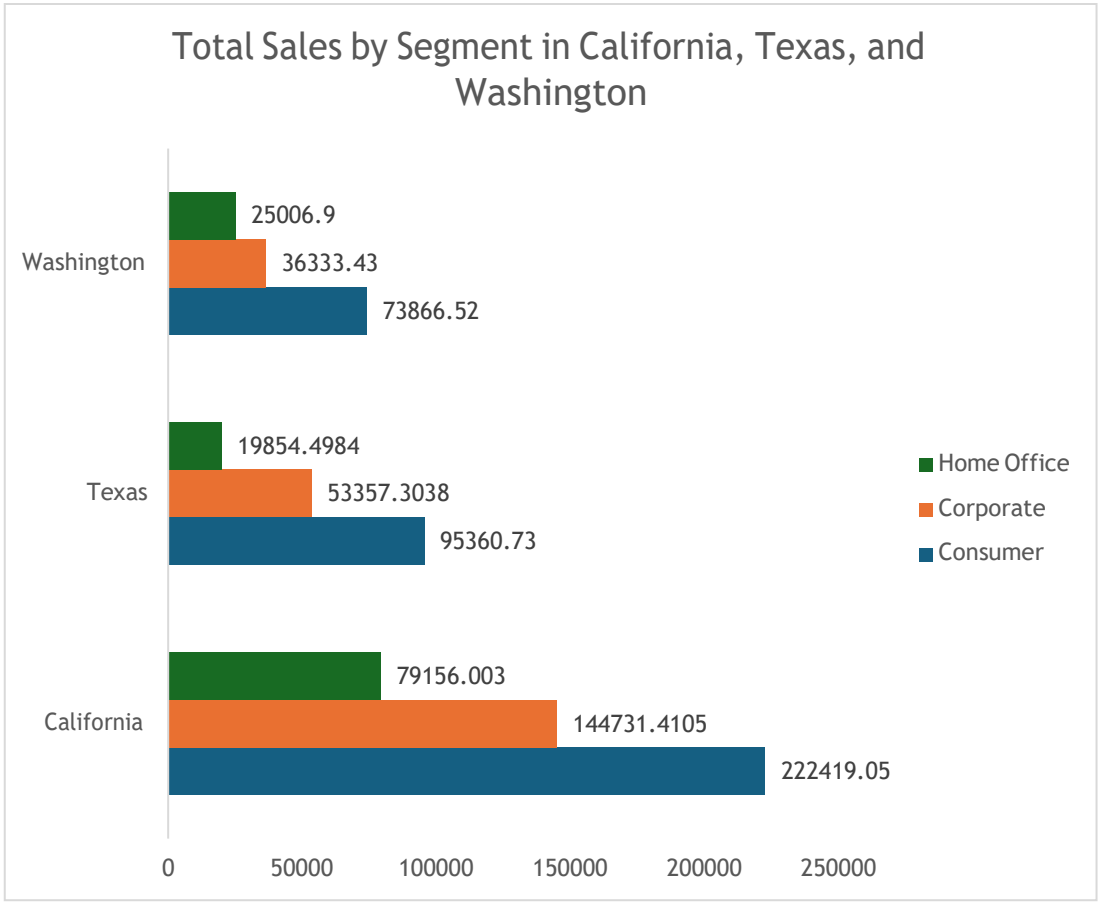


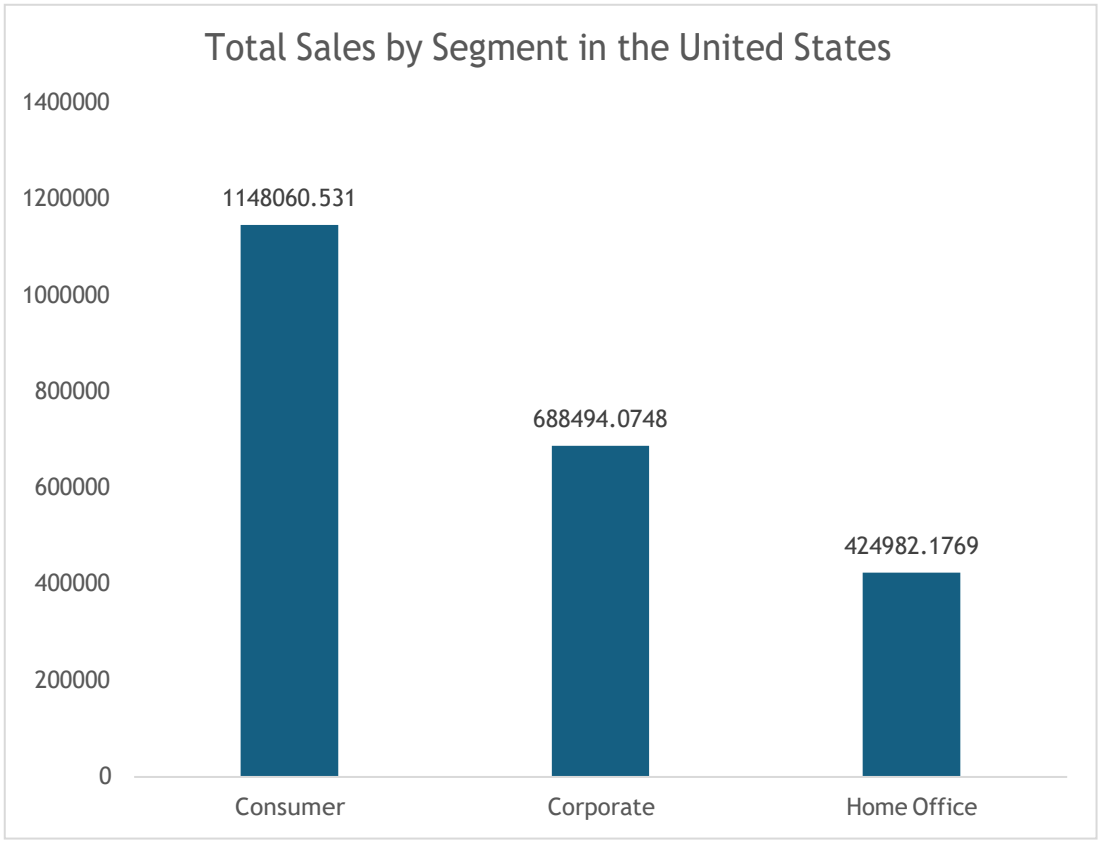
Upon analyzing the sales breakdown by product category in the top 10 states, several notable insights emerge. In Washington, California, and New York, the Technology category stands out as the highest-selling category, contributing significantly to the overall sales figures. Meanwhile, in states like 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 grand 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 most sales in 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
Consumer
0
50000
100000
150000
200000
250000

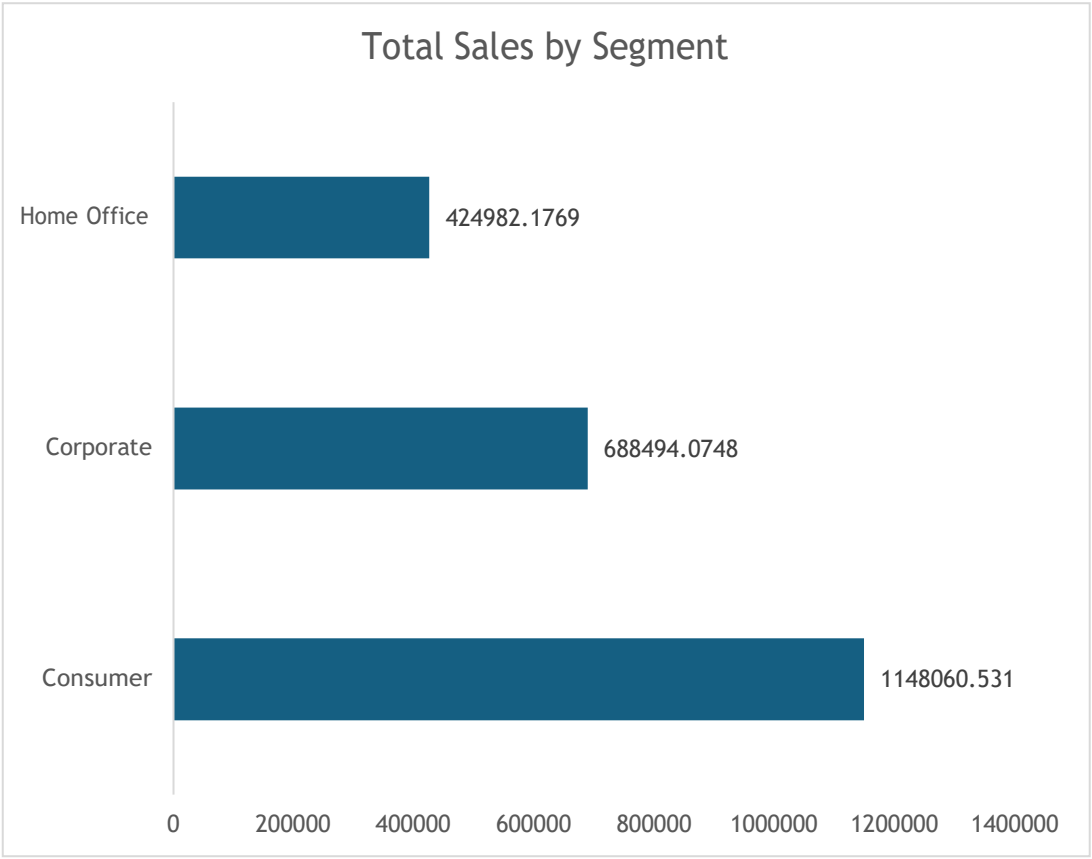

Total Sales by Segment in the United States
1400000
1200000
1000000
800000
600000
400000
200000
0
1148060.531
688494.0748
424982.1769
Consumer
Corporate
Home Office

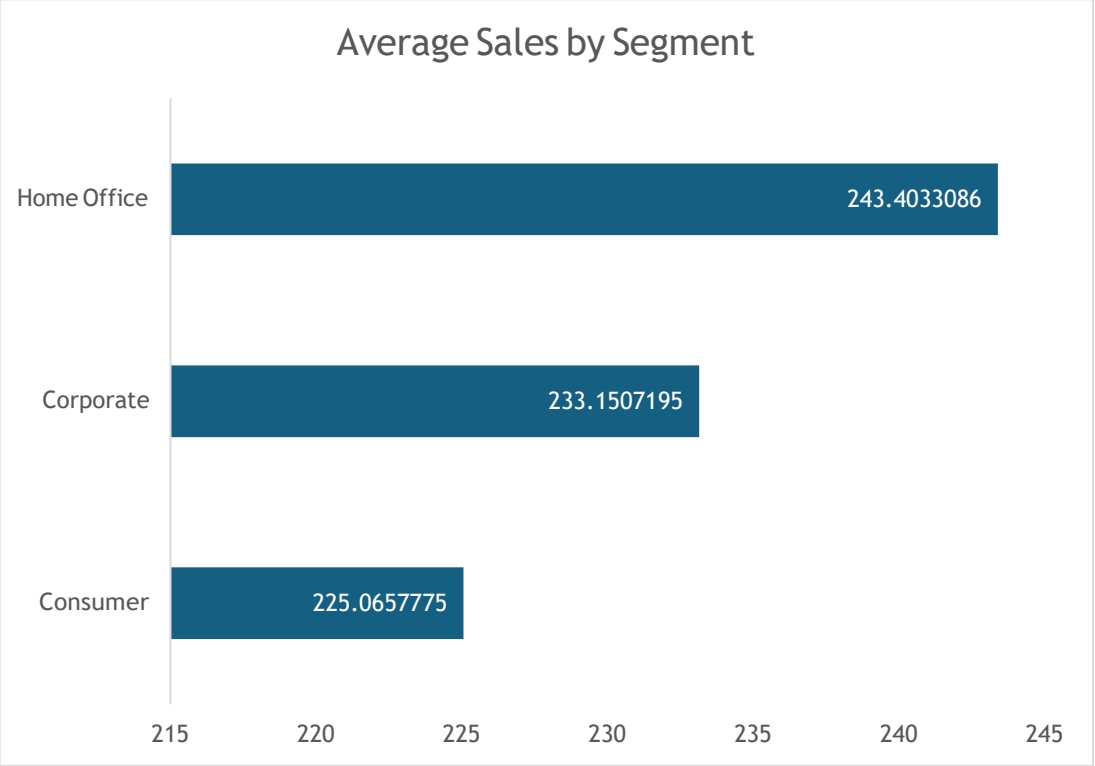
The analysis of total sales by segment in California, Texas, and Washington illustrates consistent trends across the three states. In all three states, the consumer segment dominates sales, with California leading in total sales followed by Texas and Washington. Specifically, the consumer segment exhibits 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:



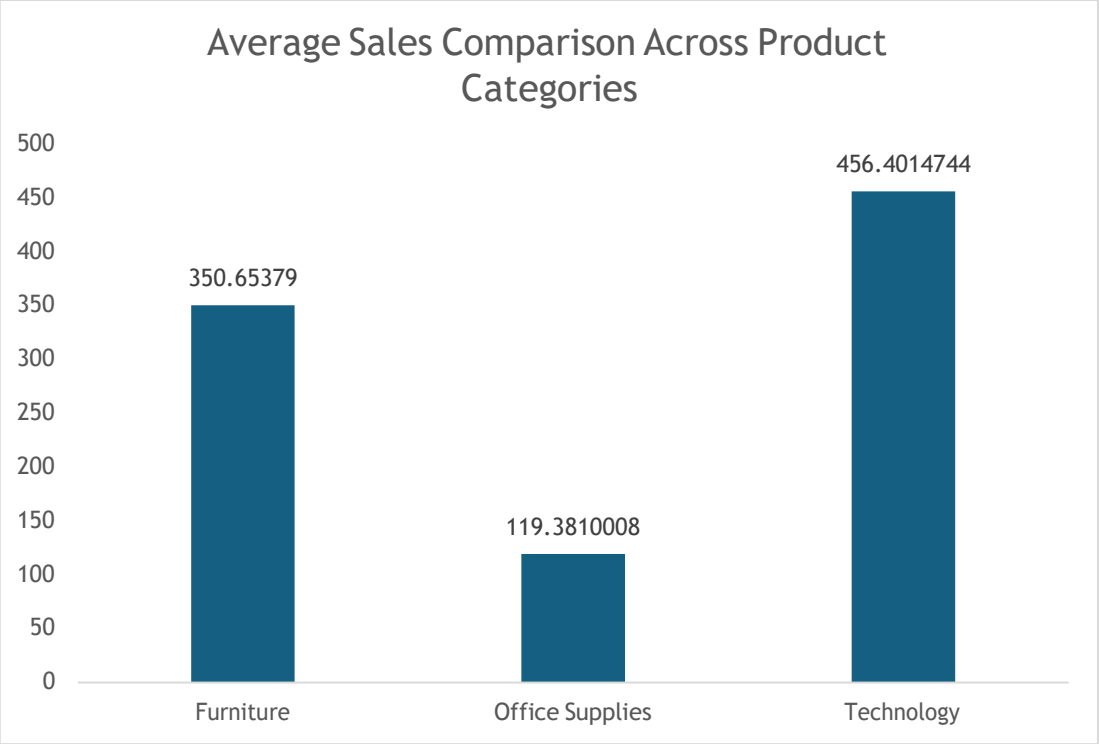


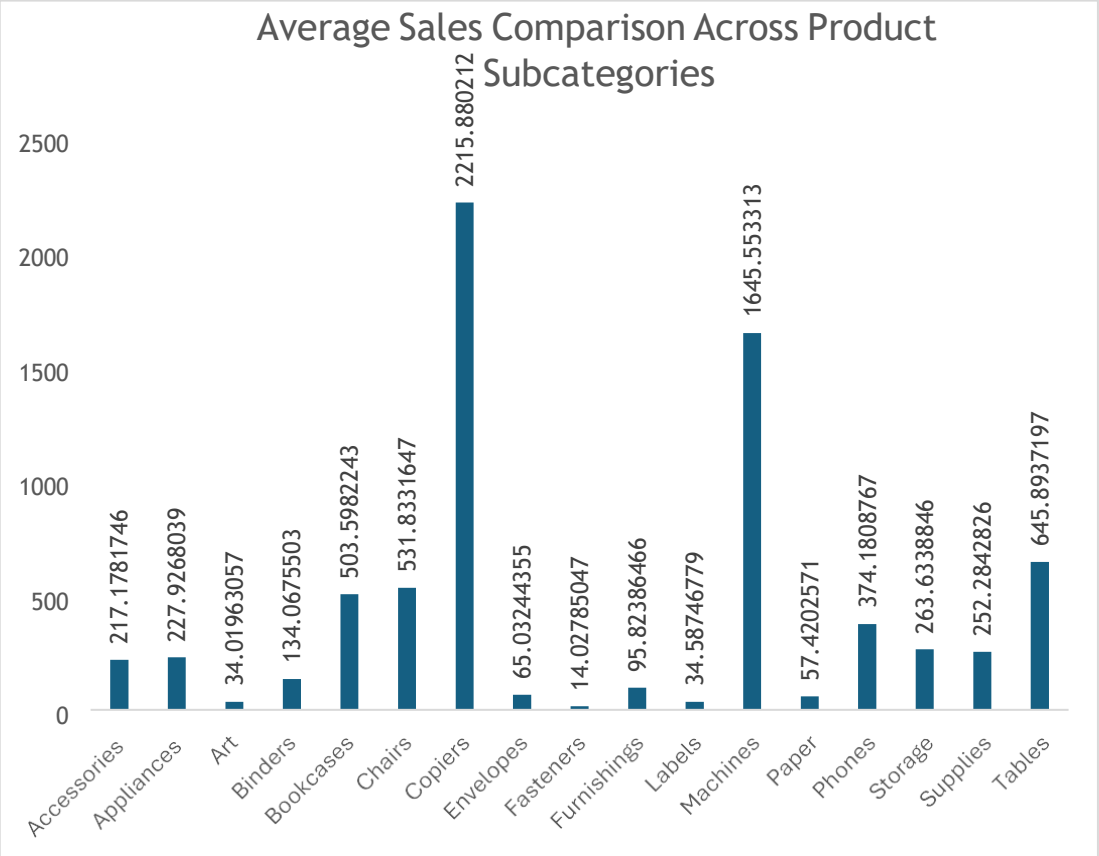
The analysis of total sales by segment reveals that the consumer segment leads with a total sales figure of $1,148,060.53, followed by corporate ($688,494.07) and home office ($424,982.18) segments. On the other hand, when considering average sales per segment, the home office segment emerges with the 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 sales of $350.65, suggesting a steady market for home and office furniture. Meanwhile, the Office Supplies category lags behind with an average sales of $119.38, indicating comparatively lower demand for office supplies products.
Analyzing the average sales across various product subcategories reveals significant disparities in performance. The highest average sales are observed in categories such as Copiers ($2215.88), Machines ($1645.55), and Tables ($645.89), indicating strong demand for these high-value products. In contrast, categories like Fasteners ($14.03) and Art ($34.02) exhibit much lower average sales, suggesting relatively lower consumer interest or niche market segments. Understanding these differences in average sales can inform product pricing strategies, inventory management, and marketing efforts to optimize revenue generation and meet customer demands effectively.

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 perform in each of these years?
5. Which cookie category sold on the highest price, country wise and how much profit is earned by that category overall?

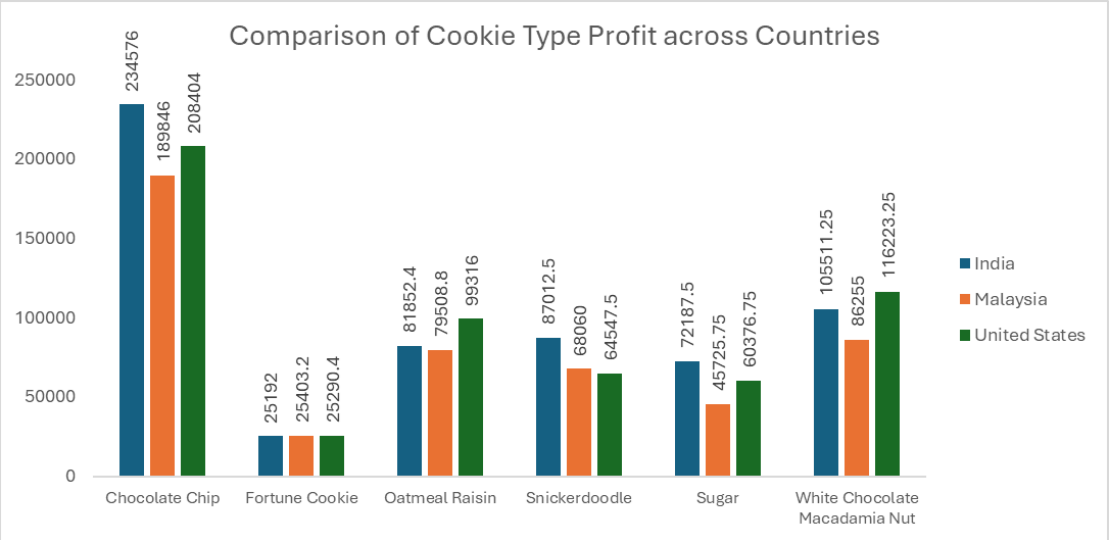
Analytics

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

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 analyse and compare profitability trends across different cookie types in diverse market contexts, aiding strategic decision-making.

1.2. Findings:

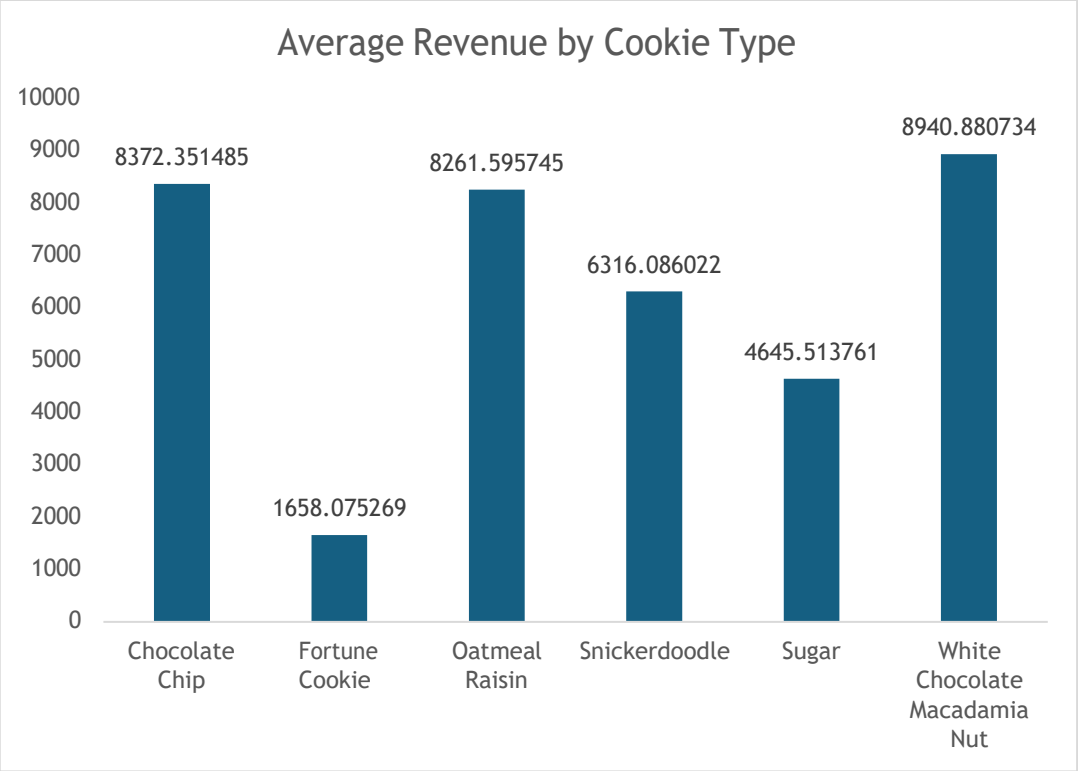


The analysis of cookie profitability across India, Malaysia, and the United States reveals interesting insights into sales performance. Among the various cookie types, Chocolate Chip cookies generate the highest total profits, with India contributing $234,576, Malaysia $189,846, and the United States $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:



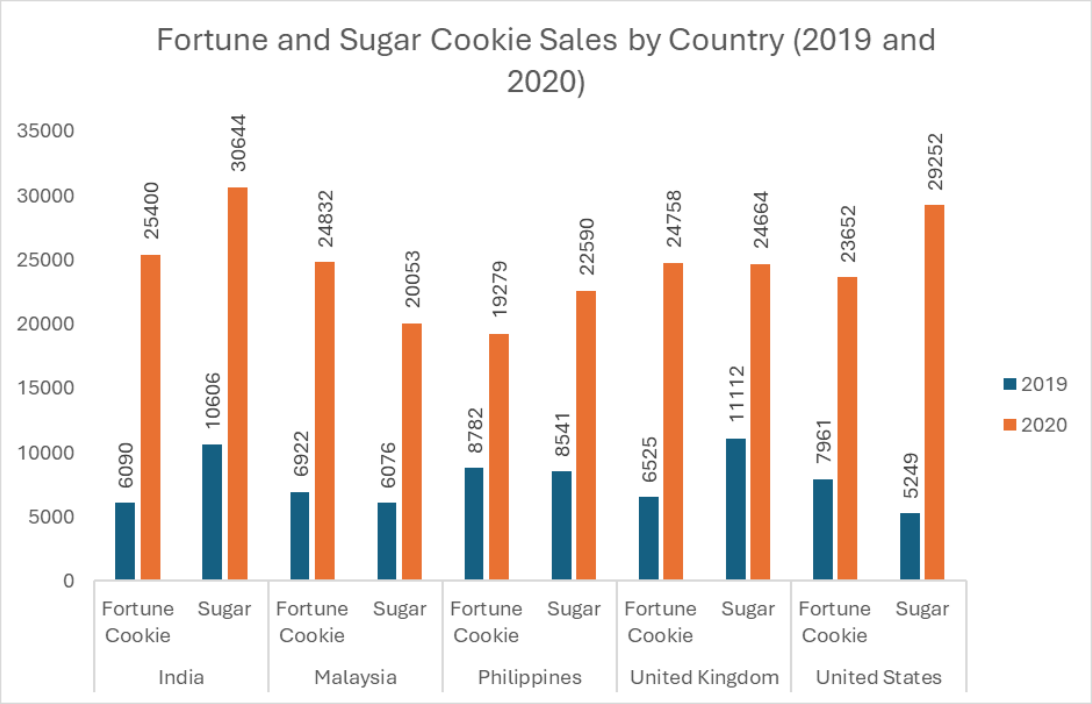
The analysis of average revenue generated by different types of cookies reveals interesting insights into their respective sales performance. Among the various cookie types, the highest average revenue is observed for the White Chocolate Macadamia Nut cookies, with an average revenue of $8,940.88. This indicates strong consumer demand and willingness to pay premium prices for this particular 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 favourites 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 in 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-year period, providing valuable insights for strategic decision-making and targeted marketing efforts.

3.2. Findings:

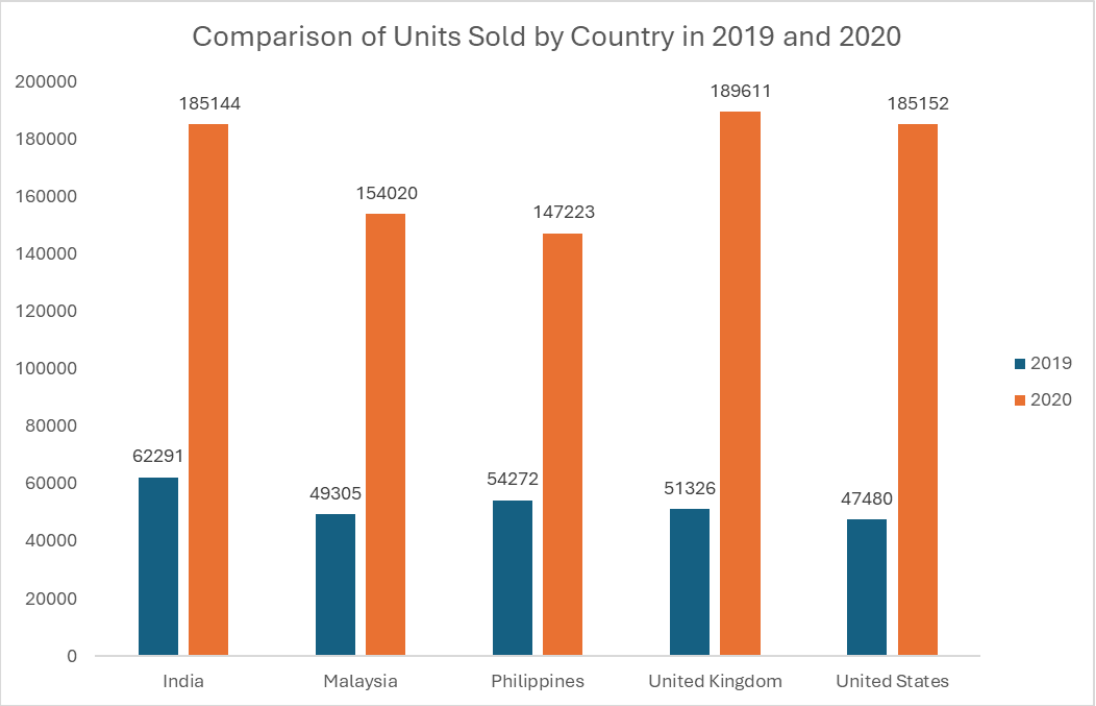


In 2019, the Philippines emerged as the top seller of Fortune cookies, while the United Kingdom led in Sugar cookie sales. This indicates distinct consumer preferences in different regions, with the Philippines showing a preference for Fortune cookies and the UK favoring Sugar cookies. However, in 2020, 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 perform 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:



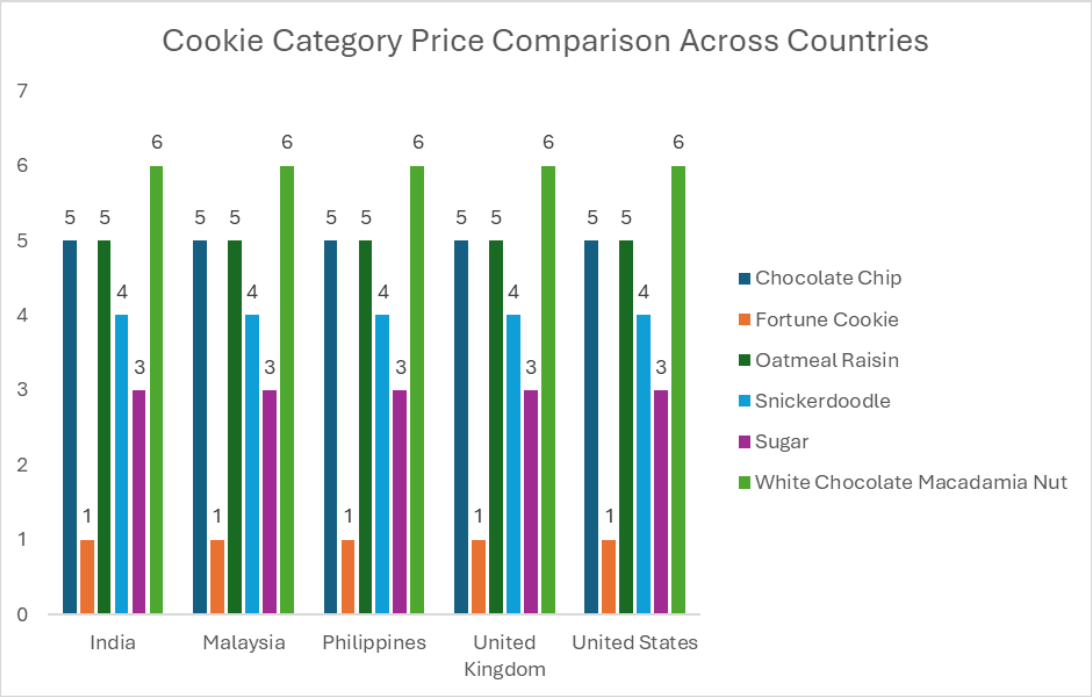
In examining the units sold by each country in 2019 and 2020, notable trends emerge. India demonstrates a substantial increase in units sold from 2019 to 2020, with a rise from 62,291 units to 185,144 units, marking a significant surge in sales. Similarly, Malaysia experiences a considerable uptick in 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-year period.

5. Which cookie category sold on the highest price, country wise and how much profit is 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 particular 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.

Analysis of Loan Applicants

Introduction

The loan dataset contains information about loan applicants, including their demographic details, education, employment status, income, loan amount, loan term, credit history, and property area. This analysis aims to provide insights into loan application trends based on gender, marital status, education level, and property area. By examining key metrics and trends, this report aims to offer actionable recommendations for financial institutions and lenders to optimize their loan approval processes and effectively target different customer segments.

Questionnaire

1. How many male graduates who are not married applied for a Loan? What was the highest amount?
2. How many female graduates who are not married applied for a Loan? What was the highest amount?
3. How many male non-graduates who are not married applied for a Loan? What was the highest amount?
4. How many female graduates who are married applied for a Loan? What was the highest amount?
5. How many male and female who are not married applied for a Loan? Compare Urban, Semi-urban, and Rural areas based on the loan amount.

Analytics

1. **How many male graduates who are not married applied for a Loan? What was the highest amount?**
   1.1. **Methodology**: Filter the dataset for male applicants who are graduates and not married. Count the number of applicants meeting these criteria and identify the highest loan amount among them.

   1.2. **Findings**: Among the male graduates who are not married and applied for a loan, the count is 66 applicants. The highest loan amount among them is $240,000.

2. **How many female graduates who are not married applied for a Loan? What was the highest amount?**
   2.1. **Methodology**: Filter the dataset for female applicants who are graduates and not married. Count the number of applicants meeting these criteria and identify the highest loan amount among them.

2.2. **Findings**: Among the female graduates who are not married and applied for a loan, the count is 35. The highest loan amount among them is $300,000.

3. **How many male non-graduates who are not married applied for a Loan? What was the highest amount?**
   3.1. **Methodology**: Filter the dataset for male applicants who are non-graduates and not married. Count the number of applicants meeting these criteria and identify the highest loan amount among them.

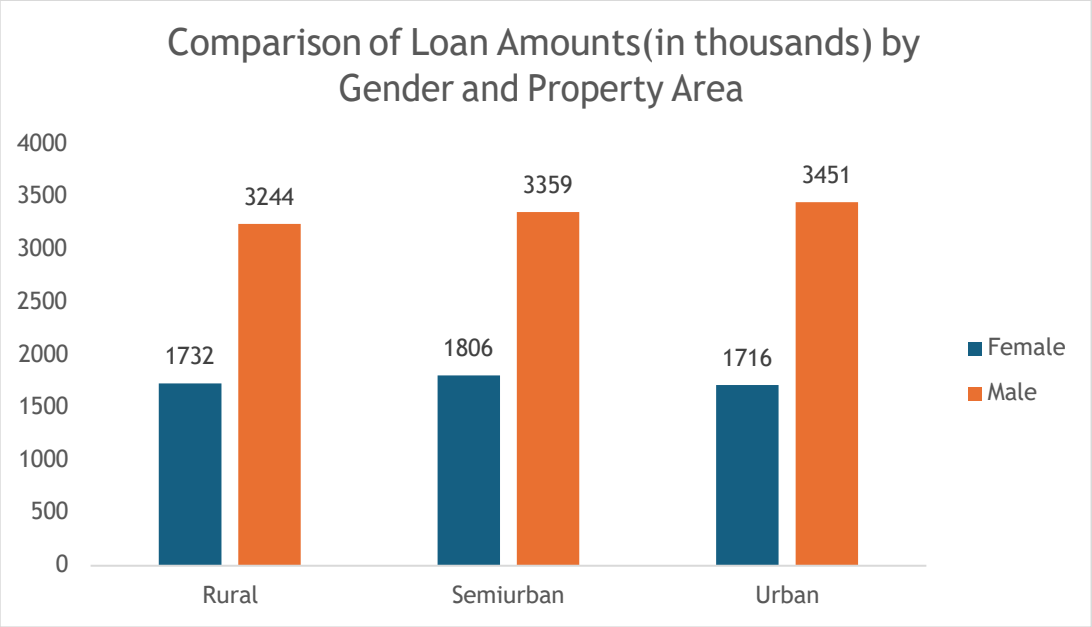
   3.2. **Findings:** Upon analyzing the loan data, it was found that 16 male non-graduates who are not married applied for a loan. The highest loan amount among these applicants was $199,000. This indicates that there is a segment of male non-graduates who are unmarried and seeking financial assistance through loans, with varying loan amounts based on their individual needs and circumstances.

4. **How many female graduates who are married applied for a Loan? What was the highest amount?**
   4.1. **Methodology**: Filter the dataset for female applicants who are graduates and married. Count the number of applicants meeting these criteria and identify the highest loan amount among them.

   4.2. **Findings**: The analysis reveals that 21 female graduates who are married applied for a loan. Among them, the highest loan amount sought was $460,000. These findings indicate a considerable number of married female graduates seeking financial assistance, with a notable demand for higher loan amounts.

5. **How many male and female who are not married applied for a Loan? Compare Urban, Semi-urban, and Rural areas based on the loan amount.**
   5.1. **Methodology**: Filter the dataset for male and female applicants who are not married. Calculate the loan amount statistics (count, mean, median, etc.) for each property area (Urban, Semi-urban, Rural) and compare them.

5.2. **Findings**:



The count of unmarried applicants is 126, with a total loan amount of $15.308 million (in thousands). When comparing loan amounts across different property areas and genders, we observe variations in the loan amounts. In rural areas, male applicants have a higher total loan amount ($3.244 million) compared to female applicants ($1.732 million). Similarly, in semi-urban areas, male applicants also have a higher total loan amount ($3.359 million) compared to female applicants ($1.806 million). Urban areas follow a similar trend, with male applicants having a higher total loan amount ($3.451 million) compared to female applicants ($1.716 million). Overall, male applicants tend to have higher total loan amounts compared to female applicants across all property areas, indicating potential disparities in loan approvals or borrowing capacity between genders.

Conclusion and Review

The analysis of loan applicant data provides valuable insights into the demographics and borrowing behaviour of individuals seeking financial assistance. Several key findings emerge from the analysis:

1. There is a notable demand for loans among both male and female applicants, with various demographic profiles and financial backgrounds.
2. Graduates, particularly females, show a significant interest in obtaining loans, with varying loan amounts based on marital status and other factors.
3. Married female graduates tend to apply for higher loan amounts compared to other demographic groups, reflecting their potentially higher financial needs.
4. Property area also influences loan application trends, with differences observed in loan amounts across urban, semi-urban, and rural areas.

5. Gender disparities exist in loan amounts, with male applicants generally seeking higher loan amounts compared to female applicants across different property areas.

These insights can inform financial institutions and lenders in tailoring their loan products, marketing strategies, and underwriting processes to better serve diverse customer segments and address their financial needs effectively.

The analysis offers valuable insights into loan application trends based on gender, marital status, education level, and property area, aiding financial institutions and lenders in decision-making. The methodology ensures reliable findings, though further exploration of factors influencing loan approval rates could enhance understanding. Overall, the report serves as a valuable resource for optimizing loan operations and customer satisfaction.

Regression

Regression Statistics	
Multiple R	0.445695483
R Square	0.198644464
Adjusted R Square	0.195852284
Standard Error	53.53517246
Observations	289

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	203897.327	203897.327	71.14315487	1.64679E-15
Residual	287	822546.2163	2866.014691		
Total	288	1026443.543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	111.0361659	4.386527877	25.31299675	4.28724E-75	102.4023205	119.6700113	102.4023205	119.6700113
ApplicantIncome	0.005554078	0.000658484	8.434640174	1.64679E-15	0.004258007	0.006850149	0.004258007	0.006850149

The regression analysis indicates a moderate relationship between ApplicantIncome and LoanAmount, with a coefficient of 0.0056 ($p < 0.001$). This suggests that for every unit increase in applicant income, there is an expected increase of approximately $0.0056 in the loan amount. The regression model explains approximately 19.9% of the variance in loan amounts, as indicated by the R-squared value of 0.199. Both the multiple R and adjusted R-squared values are relatively low, suggesting that other factors not included in the model may also influence loan amounts. The ANOVA results are highly significant ($p < 0.001$), indicating that the regression model is a significant predictor of loan amounts. Overall, the regression analysis suggests that applicant income is a significant predictor of loan amounts, with higher incomes associated with higher loan amounts.

Anova

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

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

The single-factor ANOVA analysis indicates a significant difference in mean loan amounts across different groups labeled as "P" ($F(1, 576) = 1474.81$, $p < 0.001$). The between-groups variation, which represents differences in mean loan amounts among the groups, is substantial, with a sum of squares of approximately 2,628,654.74. This suggests that the variation in loan amounts among the groups is much larger than the variation within each group. The results imply that the group variable significantly influences loan amounts, highlighting the importance of considering this factor when analyzing loan data. This finding underscores the potential impact of group-specific factors on loan amounts and the need to account for these factors in loan assessment and decision-making processes.

Descriptive Statistics

ApplicantIncome		*CoapplicantIncome*		*LoanAmount*		*Loan_Amount_Term*	
Mean	4637.352941	Mean	1528.262976	Mean	136.7923875	Mean	342.6712803
Standard Error	281.8049373	Standard Error	139.858777	Standard Error	3.511740113	Standard Error	3.862088397
Median	3833	Median	879	Median	126	Median	360
Mode	5000	Mode	0	Mode	150	Mode	360
Standard Deviation	4790.683934	Standard Deviation	2377.599209	Standard Deviation	59.69958191	Standard Deviation	65.65550274
Sample Variance	22950652.56	Sample Variance	5652978	Sample Variance	3564.040081	Sample Variance	4310.64504
Kurtosis	141.6120337	Kurtosis	32.96701001	Kurtosis	5.73980391	Kurtosis	8.629939979
Skewness	10.41122588	Skewness	4.510775295	Skewness	1.780616236	Skewness	-2.641467851
Range	72529	Range	24000	Range	432	Range	474
Minimum	0	Minimum	0	Minimum	28	Minimum	6
Maximum	72529	Maximum	24000	Maximum	460	Maximum	480
Sum	1340195	Sum	441668	Sum	39533	Sum	99032
Count	289	Count	289	Count	289	Count	289

The descriptive statistics reveal key insights into the variables ApplicantIncome, CoapplicantIncome, LoanAmount, and Loan_Amount_Term. ApplicantIncome exhibits a wide range of values, with a mean of $4637.35 and considerable variability indicated by a standard deviation of $4790.68. CoapplicantIncome also shows variability, with a mean of $1528.26 and a wider range from $0 to $24000, suggesting instances of no coapplicant income.

LoanAmount, with a mean of $136.79, displays moderate variability compared to income variables, while Loan_Amount_Term, with a mean of 342.67 months, suggests a relatively consistent duration for loan terms. However, both LoanAmount and Loan_Amount_Term exhibit skewness and kurtosis, indicating departures from normal distribution. Overall, these statistics offer a comprehensive understanding of the central tendency, variability, and distributional characteristics of the dataset, essential for further analysis and decision-making.

Correlation

	ApplicantIncome	*CoapplicantIncome*	*LoanAmount*	*Loan_Amount_Term*
ApplicantIncome	1			
CoapplicantIncome	-0.084353248	1		
LoanAmount	0.445695483	0.230355168	1	
Loan_Amount_Term	0.022726771	-0.000621142	0.115750256	1

The correlation analysis reveals nuanced relationships among the variables. ApplicantIncome shows a moderate positive correlation with LoanAmount, implying that higher incomes tend to be associated with larger loan amounts. However, there's only a weak correlation between ApplicantIncome and Loan_Amount_Term, indicating little influence of income on loan duration. CoapplicantIncome exhibits a weak positive correlation with LoanAmount, suggesting that higher coapplicant incomes are associated with larger loans. Additionally, LoanAmount and Loan_Amount_Term display a weak positive correlation, implying that larger loan amounts tend to have longer terms. Overall, while income influences loan amount, its impact on loan term is minimal, highlighting the complex dynamics involved in loan assessment.

Analysis of Sales Performance: Unveiling Insights from Sales Data

Introduction

The sales dataset under examination offers a wealth of information on transactions conducted over a period, detailing the salesmen, items sold, companies, quantities, and amounts involved. This report embarks on a journey to extract valuable insights from this dataset, shedding light on sales performance, product popularity, profitability, and average sales trends. By delving into key metrics and patterns, this analysis aims to uncover actionable insights to guide strategic decision-making and optimize sales strategies in the future.

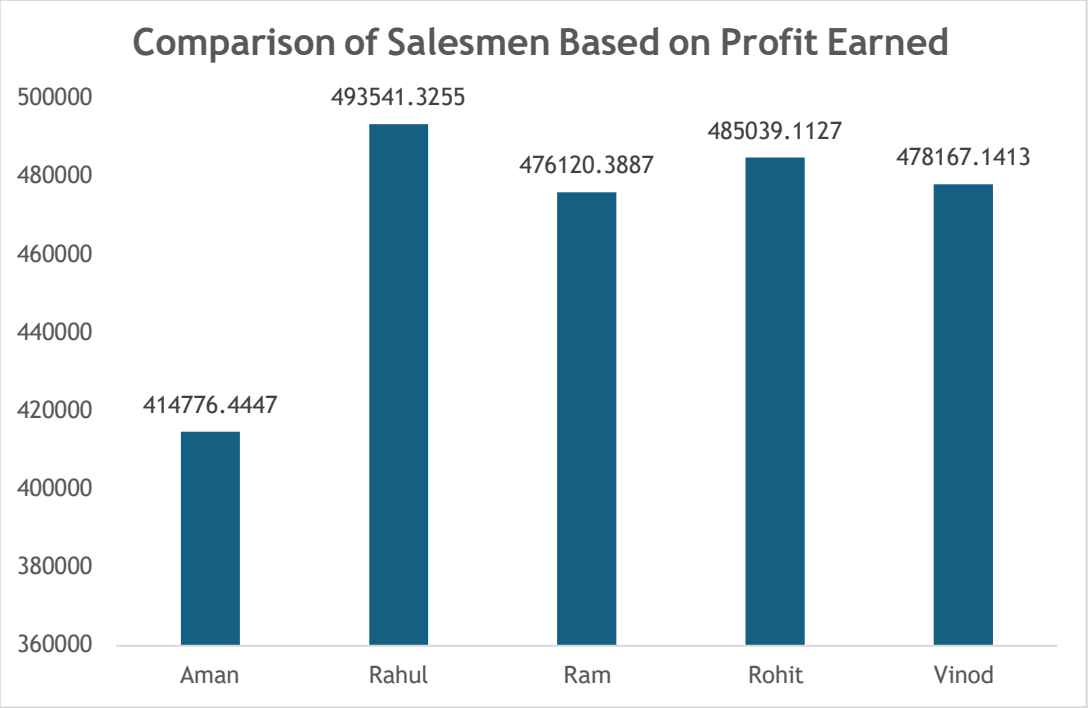
Questionnaire

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 products sold the most over the year Computer or Laptop?
4. Which item yield most average profit?
5. Find out average sales of all the products and compare them.

Analytics

1. **Compare all the salesmen based on profit earn.**
   1.1. **Methodology**: Calculate the total profit earned by each salesman by summing the profit from all their sales. Then, compare the total profits to identify the top-performing salesmen.

1.2. **Findings**:

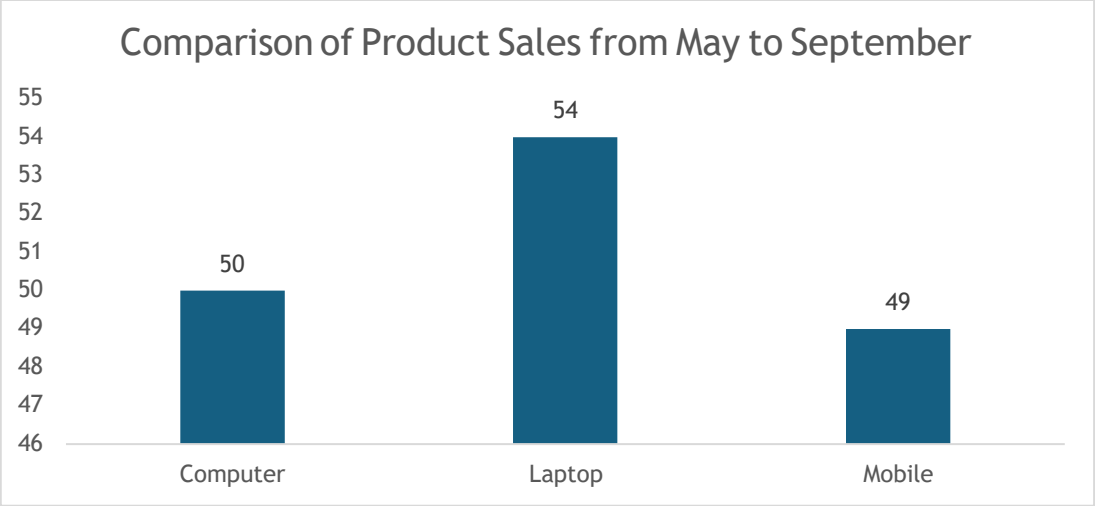


In analyzing the sales data, it's evident that each salesman has contributed significantly to the overall profit earned during the period under consideration. Rahul emerges as the top performer, generating the highest total profit of $493,541.33. Close behind are Rohit, Ram, and Vinod, each demonstrating strong sales performance with profits of $485,039.11, $476,120.39, and $478,167.14, respectively. While Aman's contribution is slightly lower, with a total profit of $414,776.44, it still represents a substantial portion of the overall profit. This comparison underscores the importance of evaluating individual sales performance to identify top performers and areas for potential improvement, ultimately contributing to the company's overall profitability and success.

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

2.1. **Methodology**: To find the most sold product from May to September, we filter the dataset for entries within this period. Then, we aggregate the quantities sold for each product and determine the product with the highest total sales volume.

2.2. **Findings**:

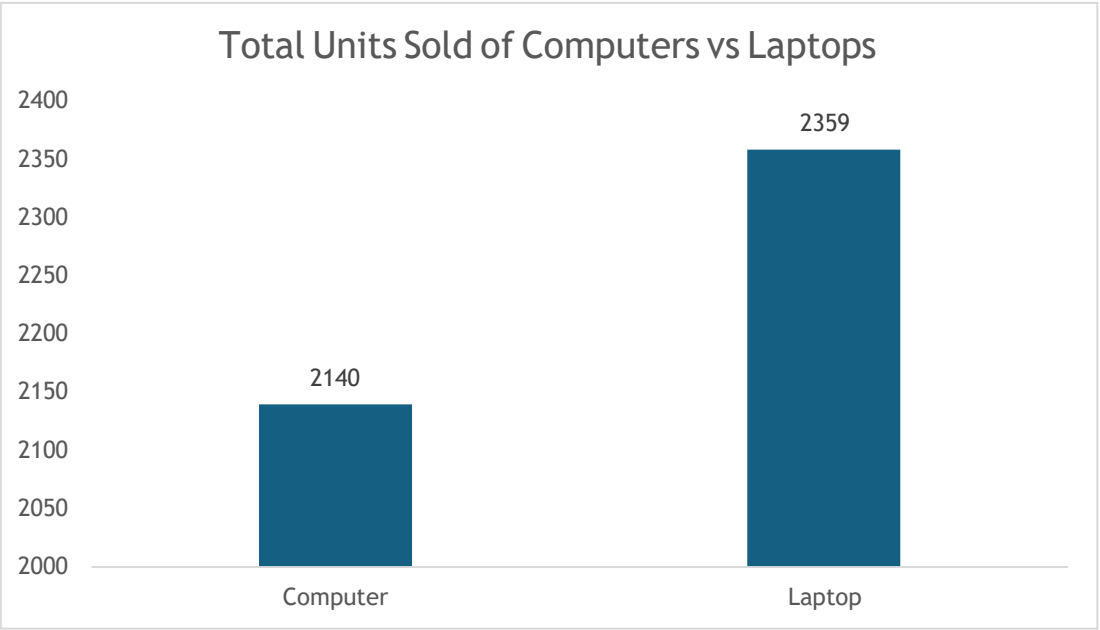


Between May and September, the most sold product was Laptop, with a total count of 54 units sold. Computer closely followed with 50 units sold, indicating consistent demand for both electronic devices during this period. Mobile phones, while still popular, had a slightly lower count of 49 units sold.

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

3.1. **Methodology**: To determine which product, Computer or Laptop, sold the most over the year, we need to aggregate the quantity of each item sold across all sales transactions. Then, we compare the total quantity sold for Computers and Laptops to identify the product with the highest sales volume.

3.2. Findings:

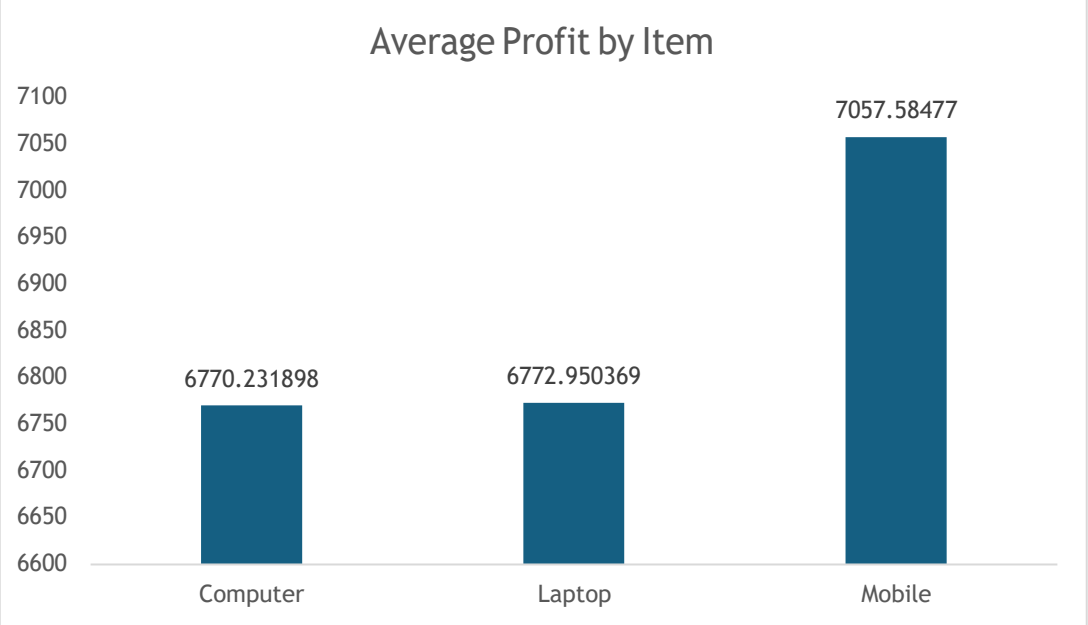


Over the year, a total of 2,140 units of Computers and 2,359 units of Laptops were sold. This indicates that Laptops outsold Computers during the year, with a higher demand for portable computing devices compared to desktop computers. The difference in sales volumes between Laptops and Computers suggests shifting consumer preferences towards more mobile and versatile computing solutions. This finding underscores the importance for companies to adapt their product offerings to meet evolving customer needs and technological trends.

4. **Which item yield most average profit?**

4.1. **Methodology**: To identify which item yielded the most average profit, we calculate the average profit for each item by dividing the total profit generated by the total quantity sold for that item. Then, we compare the average profits across all items to determine the item with the highest average profit.

4.2. **Findings**:

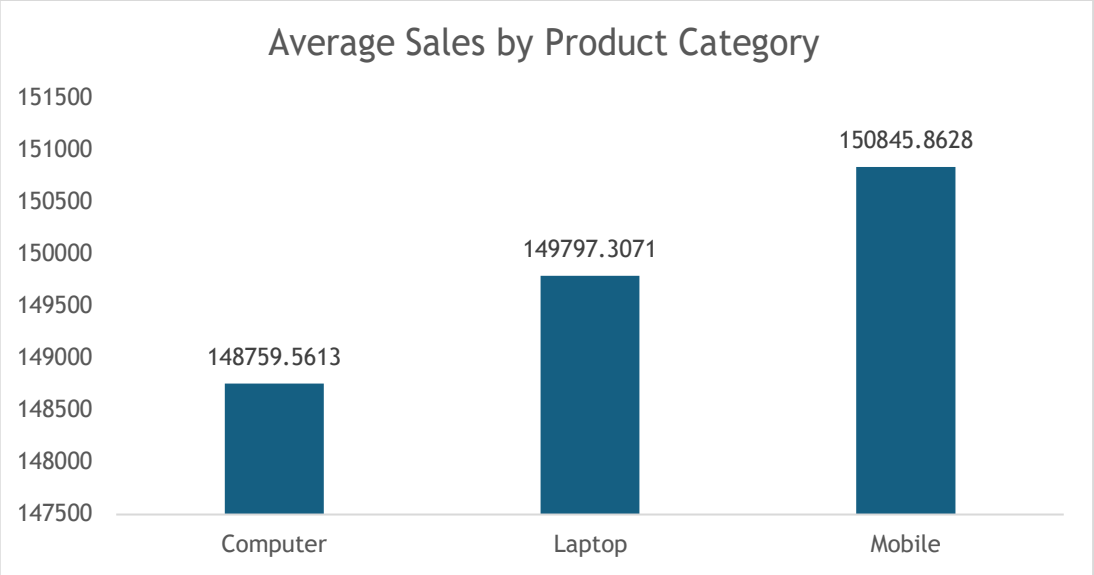


The analysis of average profits reveals that the item "Mobile" yielded the highest average profit, with an average profit of $7,057.58. This indicates that, on average, each unit of the "Mobile" item generated the highest profit compared to the other items. While "Computer" and "Laptop" also had relatively high average profits, the "Mobile" item stood out as the top performer in terms of average profit.

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

5.1. **Methodology**: For finding out the average sales of all the products and comparing them, we calculate the total sales amount for each product and then divide it by the total number of transactions for that product. This gives us the average sales for each product. We then compare these average sales figures across all products to identify any variations in sales performance.

5.2. **Findings**:



The analysis of average sales indicates that the product category "Mobile" has the highest average sales, with an average of $150,845.86. This suggests that, on average, sales of mobile devices were the highest compared to computers and laptops. However, the differences in average sales between the product categories are relatively small, with "Computer" and "Laptop" also exhibiting strong average sales performances, averaging $148,759.56 and $149,797.31, respectively.

Conclusion and Review

In conclusion, the analysis of the sales dataset has provided valuable insights into sales performance, product popularity, profitability, and average sales trends. Through evaluating individual salesmen based on profit earned, identifying the most sold product over a specific period, determining the top-selling product between Computers and Laptops, analyzing the item with the highest average profit, and comparing average sales across all products, we have gained a comprehensive understanding of the sales landscape. These insights can guide strategic decision-making processes and help optimize sales strategies to drive revenue growth and enhance overall business performance.

Review: The analysis conducted on the sales dataset has been thorough and insightful, providing actionable insights that can inform strategic decision-making. By examining key metrics such as profit earned, product sales volume, profitability, and average sales, the report offers valuable information for stakeholders to understand sales performance and identify areas for improvement. The methodology employed in each analysis was clear and systematic, ensuring accurate results and reliable conclusions. Overall, the report effectively fulfills its objective of uncovering insights from the sales data, empowering businesses to make informed decisions and drive success in the competitive market landscape.

Regression

Regression Statistics	
Multiple R	0.984561511
R Square	0.969361369
Adjusted R Square	0.969271256
Standard Error	16609.19129
Observations	342

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.96751E+12	2.96751E+12	10757.10171	2.0342E-259
Residual	340	93794180017	275865235.3		
Total	341	3.0613E+12			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-73454.83602	2332.435331	-31.49276425	7.8806E-103	-78042.65639	-68867.01564	-78042.65639	-68867.01564
Qty	11474.50523	110.6334182	103.7164486	2.0342E-259	11256.89309	11692.11737	11256.89309	11692.11737

The regression analysis indicates a highly significant relationship between the independent variable Qty and the dependent variable (not specified), with an adjusted R-squared value of 0.969. This suggests that approximately 96.9% of the variance in the dependent variable can be explained by Qty. The coefficient for Qty is 11474.51, indicating that for each unit increase in Qty, there is an expected increase of $11474.51 in the dependent variable. Both the intercept and Qty coefficients are highly significant ($p < 0.001$), with t-statistics of -31.49 and 103.72, respectively. Overall, the model demonstrates strong predictive power, suggesting that Qty is a significant predictor of the dependent variable.

Anova

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Companyy	342	753	2.201754386	0.988501312
Amount	342	2347644.413	6864.457348	4410782.252

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8052484363	1	8052484363	3651.27186	4.8775E-276	3.855129873
Within Groups	1504077085	682	2205391.62			
Total	9556561448	683				

The ANOVA analysis indicates a significant difference in mean amounts across different groups labeled as "Companyy" and "Amount" ($F(1, 682) = 3651.27$, $p < 0.001$). The between-groups variation, representing differences in mean amounts among the groups, is substantial, with a sum of squares of approximately 8.05 billion. This suggests that the variation in amounts among the groups is much larger than the variation within each group. The results imply that the group variable significantly influences the amounts, highlighting the importance of considering this factor when analyzing the data. This finding underscores the potential impact of group-specific factors on amounts and the need to account for these factors in further analysis and decision-making processes.

Descriptive Statistics

Qty		*Amount*	
Mean	19.45693356	Mean	6864.457348
Standard Error	0.439614404	Standard Error	113.5650656
Median	19.45693356	Median	6984.647162
Mode	3	Mode	1000
Standard Deviation	8.129895565	Standard Deviation	2100.186242
Sample Variance	66.09520189	Sample Variance	4410782.252
Kurtosis	-0.998826126	Kurtosis	-0.507800424
Skewness	-0.099479188	Skewness	-0.364490893
Range	30.30851595	Range	9279.851244
Minimum	3	Minimum	1000
Maximum	33.30851595	Maximum	10279.85124
Sum	6654.271277	Sum	2347644.413
Count	342	Count	342

The descriptive statistics for the variables Qty and Amount provide a clear overview of their distribution and central tendency. Qty has a mean of approximately 19.46 units, with a standard deviation of 8.13, indicating moderate variability around the mean. The data is approximately symmetrically distributed, as indicated by a skewness value close to zero (-0.10) and a slightly negative kurtosis value (-1.00), suggesting a slightly flatter distribution. The range of Qty spans from 3 to 33.31 units. Amount, on the other hand, has a mean of approximately $6864.46, with a much larger standard deviation of $2100.19, indicating considerable variability in amounts. The data is slightly negatively skewed (-0.36) and has a slightly negative kurtosis (-0.51), suggesting a slightly flatter distribution. The range of Amount spans from $1000 to $10279.85. Overall, these statistics provide insights into the distribution and central tendency of Qty and Amount, essential for understanding their characteristics in the dataset.

Correlation

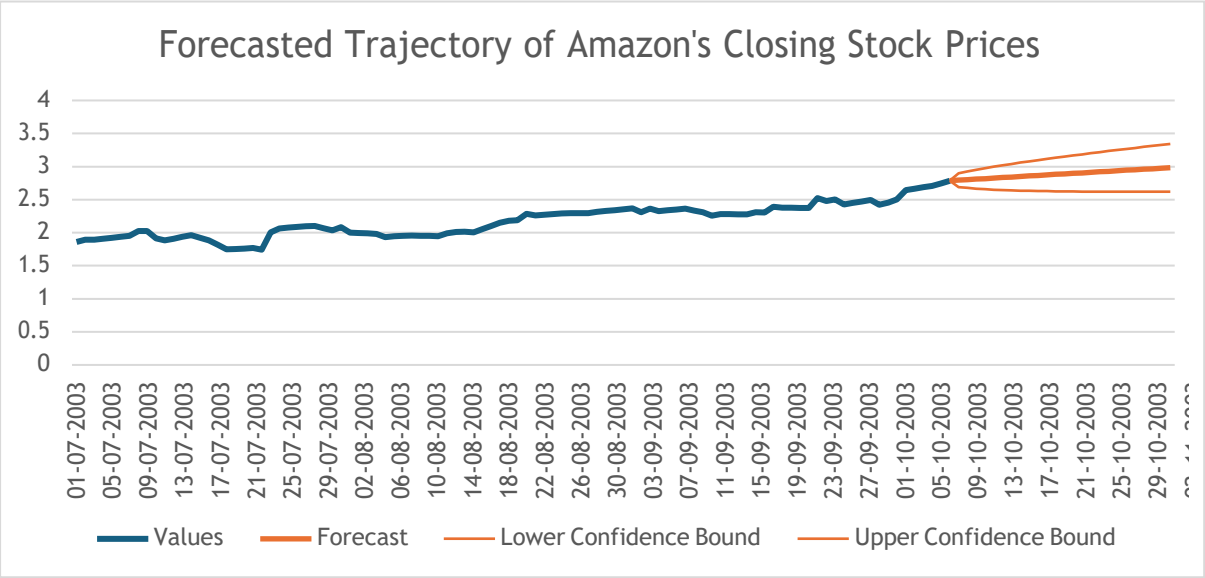
	Qty	*Amount*
Qty	1	
Amount	0.954077	1

The correlation analysis reveals a strong positive relationship between the variables Qty and Amount, with a correlation coefficient of approximately 0.954. This indicates that there is a high degree of linear association between the two variables, suggesting that as the quantity increases, the amount also tends to increase proportionally. This strong correlation suggests that changes in Qty are highly predictive of changes in Amount, highlighting the close connection between the two variables within the dataset.

Analysis of Forecasted Trends in Flipkart'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.