ACROPOLIS INSTITUTE OF TECHNOLOGY & RESEARCH,INDORE

DEPARTMENT OF COMPUTER SCIENCE

CS-605 Data Analytics Lab

3rd Year 6th Semester

2023-2024

SUBMITTED BY-

ANUSHKA PATERIYA

(0827CS211031)

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. Car Collection Dataset iv. Understanding Sales: Orders, Regions, and Segments v. Analysis of Cookie Sales vi. Analysis of Loan Applicants vii. Analysis of Sales Performance	
3.	Reports: i. Store Data Analysis ii. Sales Data Analysis iii. Car Collection Dataset iv. Orders Data Report v. Analysis of Cookie Data vi. Analysis of Loan Applicants vii. Sales Data Analysis	
4.	Analysis of Forecasted Trends in Netflix'sClosing Stock Prices	

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

Data Analysis Principles

Data analysis principles are fundamental guidelines and methodologies that guide the process of extracting meaningful insights from datasets.

- **Data Quality:** Ensuring the accuracy, reliability, and completeness of data. This involves data validation, verification, and cleansing to remove errors, inconsistencies, and missing values.
- **Data Cleaning:** The process of identifying and correcting errors, inconsistencies, and outliers in the dataset to enhance data quality and ensure accurate analysis results.
- **Exploratory Data Analysis (EDA):** Utilizing statistical and visualization techniques to explore and summarize the main characteristics of the dataset. EDA helps in understanding data distributions, patterns, trends, and relationships, guiding further analysis and hypothesis generation.
- **Data Visualization:** Graphical representation of data to facilitate understanding, analysis, and decision-making. Techniques include charts, graphs, and dashboards to present complex datasets in an intuitive and visually appealing manner.
- **Reproducibility:** The ability to replicate data analysis processes and results. Documenting the analysis methodology, code, and assumptions ensures that other researchers can verify and reproduce the findings, enhancing transparency and credibility.

Statistical Analytics Concepts

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

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

Hypothesis Testing

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

1. **Null Hypothesis (H0):** The default assumption that there is no significant difference or effect in the population being studied.
2. **Alternative Hypothesis (H1):** A statement contradicting the null hypothesis, suggesting that there is a significant difference or effect in the population.
3. **Hypothesis Testing:** A statistical method used to make inferences about population parameters based on sample data. It involves specifying a null hypothesis, selecting a significance level, collecting data, and determining whether the evidence supports rejecting or failing to reject the null hypothesis.

Regression and Its Types

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

1. **Linear Regression:** Models the relationship between the dependent variable and one or more independent variables using a linear equation. It is commonly used for predicting continuous outcomes.
   Formula: $y=\beta 0+\beta 1x+\varepsilon y=\beta 0+\beta 1x+\varepsilon$
2. **Logistic Regression:** Models the probability of a binary outcome using the logistic function. Suitable for predicting categorical outcomes with two levels.
   Formula: $p=11+e-(\beta 0+\beta 1x)p=1+e-(\beta 0+\beta 1x)1$
3. **Polynomial Regression:** Models the relationship using a polynomial equation to capture non-linear relationships between variables.
   Formula: $y=\beta 0+\beta 1x+\beta 2x2+...+\beta nxn+\varepsilon y=\beta 0+\beta 1x+\beta 2x2+...+\beta nxn+\varepsilon$
4. **Ridge and Lasso Regression:** Regularization techniques used to prevent overfitting in regression models by penalizing large coefficients.

Correlation

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

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

Analysis of Variance (ANOVA)

ANOVA is a statistical technique used to compare means across multiple groups.

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

7. 5 V's OF BIG DATA:

The concept of the 5Vs of Big Data is a framework to understand the key characteristics that define big data and differentiate it from traditional data. Here's a detailed and in-depth look at each of the 5Vs:

1. Volume

Volume refers to the vast amount of data generated every second from various sources such as social media, sensors, transactions, logs, and more.

Key Points:

- Data Scale: The scale of data is enormous, often measured in terabytes, petabytes, and even exabytes.
- Storage Solutions: Requires advanced storage solutions like distributed file systems (e.g., Hadoop HDFS) and cloud storage (e.g., AWS S3) to handle large datasets efficiently.
- Data Sources: Includes data from various sources like social media posts, IoT sensors, transactional databases, multimedia content, and more.
- Impact: High volume necessitates robust data processing and storage capabilities, often leading to the development of new technologies and infrastructure to manage and utilize the data effectively.

2. Velocity

Velocity refers to the speed at which data is generated, processed, and analyzed. It emphasizes the real-time or near-real-time nature of data handling.

Key Points:

- Real-Time Processing: Technologies like Apache Kafka, Apache Storm, and Spark Streaming enable the processing of data in real-time.
- Data Streams: Continuous data streams from sources like financial markets, social media feeds, sensor networks, and more.
- Impact on Decision Making: Faster data processing allows for timely insights and decision-making, which is crucial for applications like fraud detection, stock trading, and personalized marketing.

- Challenges: Managing and analyzing data at high speed can be challenging, requiring optimized algorithms and robust infrastructure.

3. Variety

Variety refers to the different types of data available. This includes structured, semi-structured, and unstructured data from a wide range of sources.
Key Points:

- Structured Data: Organized data in databases, e.g., relational databases (SQL).
- Semi-Structured Data: Data with some organizational properties but not fully structured, e.g., JSON, XML.
- Unstructured Data: Data without a predefined structure, e.g., text, images, videos, and social media posts.
- Data Integration: Combining different types of data for comprehensive analysis can be complex but essential for gaining deeper insights.
- Impact on Analysis: Tools and techniques such as NoSQL databases (e.g., MongoDB), text analytics, image recognition, and natural language processing (NLP) are used to handle and analyze diverse data types.

4. Veracity

Veracity refers to the accuracy, quality, and trustworthiness of the data. It addresses the uncertainties and inconsistencies present in data.
Key Points:

- Data Quality: Involves ensuring data accuracy, completeness, and reliability.
- Data Cleansing: Techniques to clean and preprocess data to improve its quality.
- Source Reliability: Evaluating the trustworthiness of data sources to ensure data integrity.
- Impact on Insights: Poor data quality can lead to incorrect insights and flawed decision-making.
- Management: Implementing data governance policies, validation checks, and anomaly detection to enhance veracity.

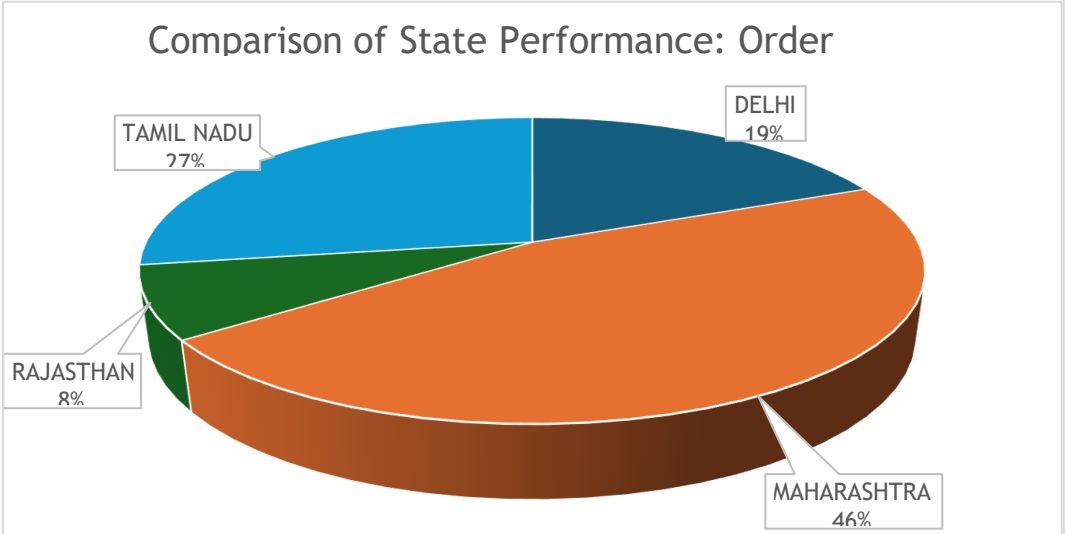
5. Value

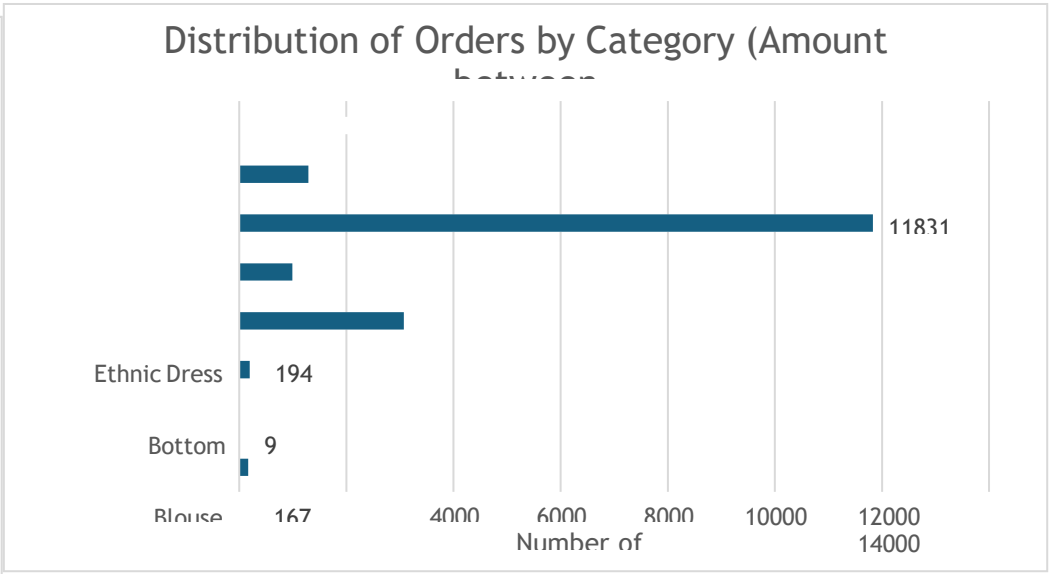
Value refers to the potential insights and benefits that can be derived from analyzing the data. It highlights the importance of transforming data into actionable insights.
Key Points:

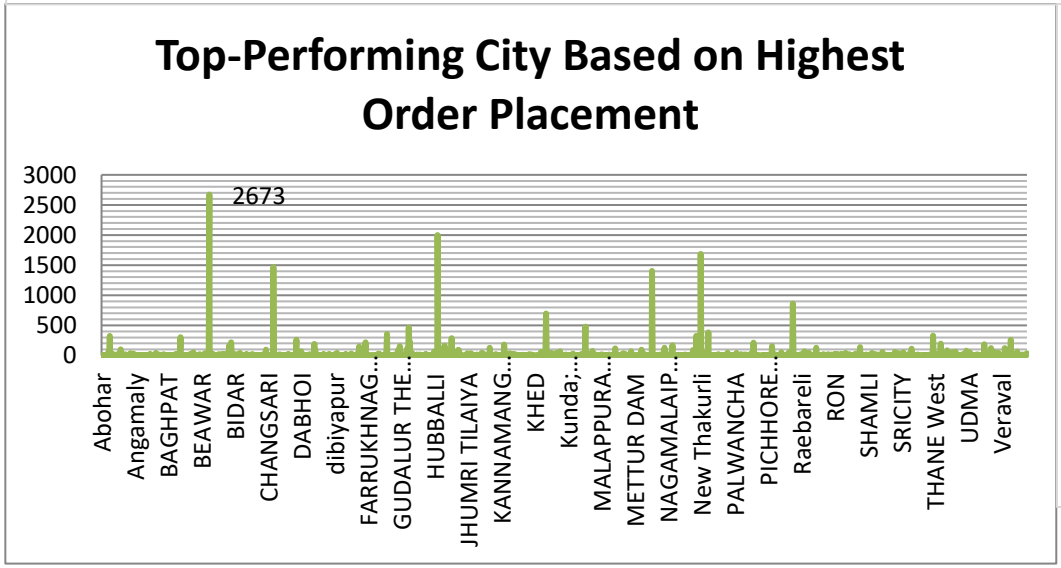
- Insight Generation: Using data analytics to uncover patterns, trends, and correlations that can inform business decisions.
- Business Impact: Data-driven decisions can lead to improved operational efficiency, better customer experiences, and competitive advantages.
- ROI: Assessing the return on investment from data initiatives to ensure that the benefits outweigh the costs involved in data processing and analysis.

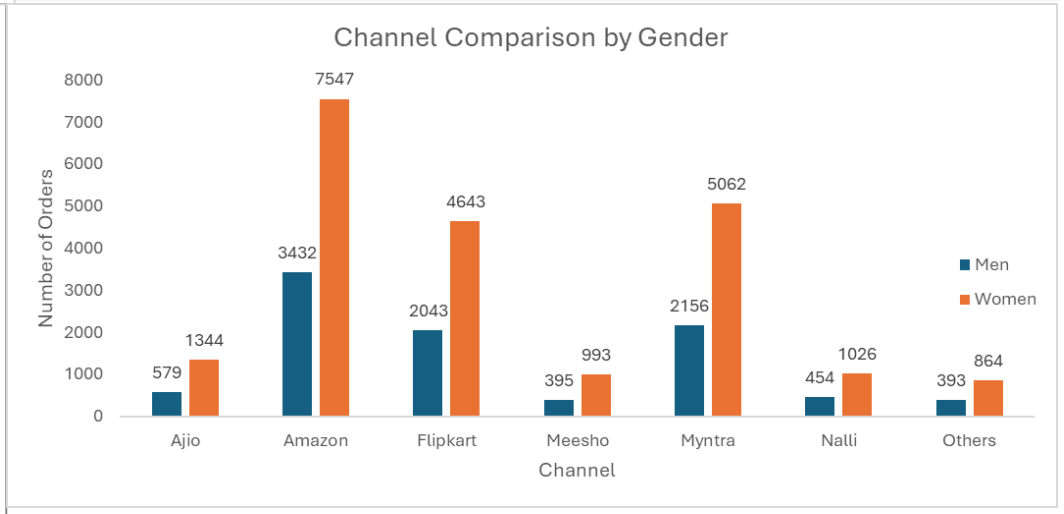
Data Monetization: Strategies to monetize data, such as selling data insights or using data to enhance products and services.

Store Data Analysis

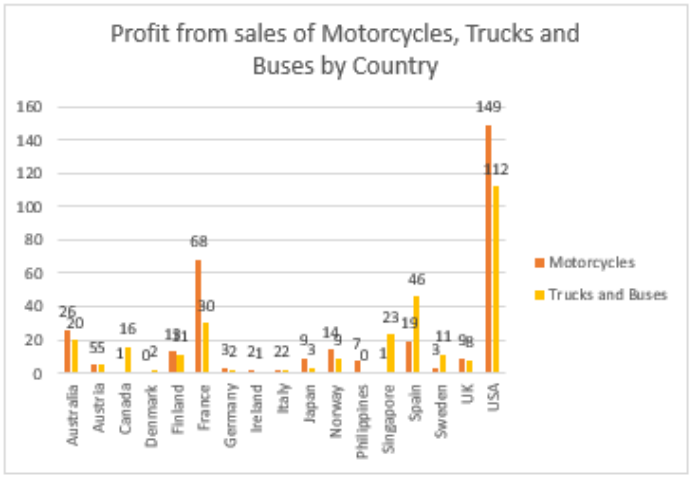


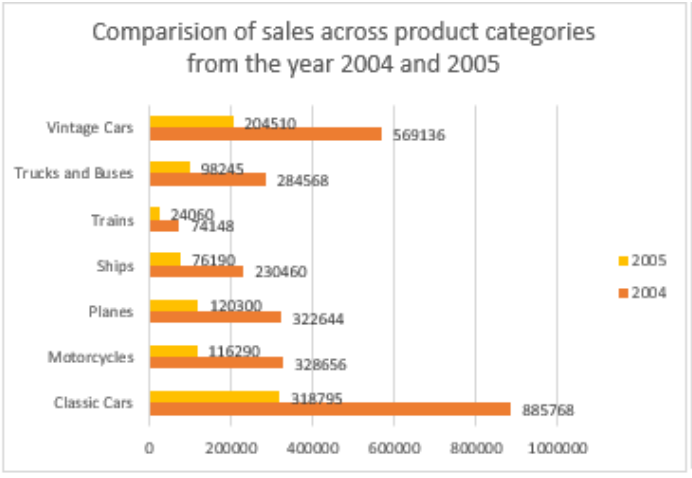


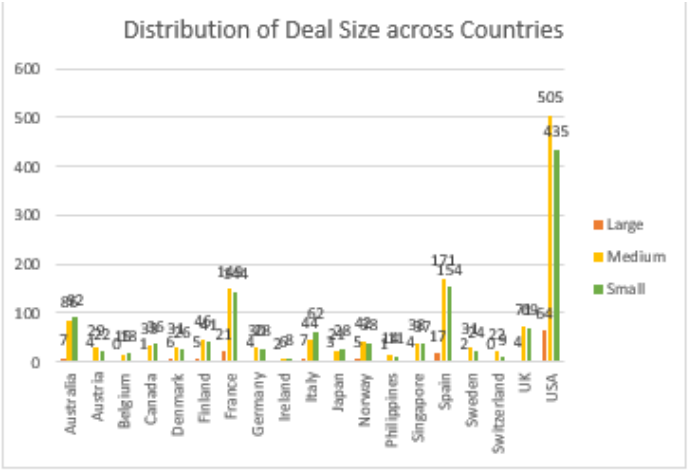


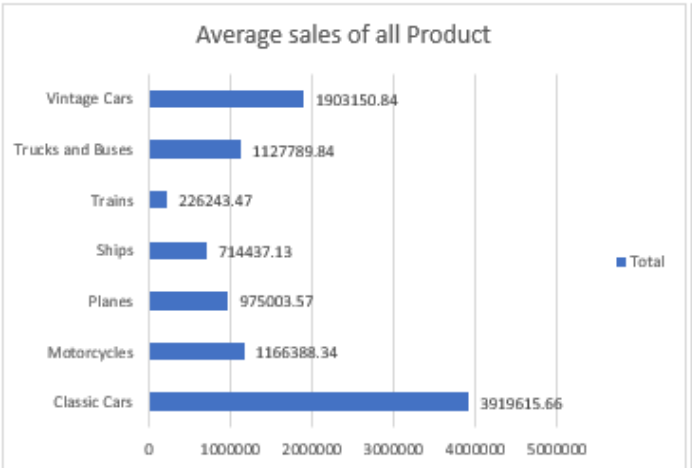


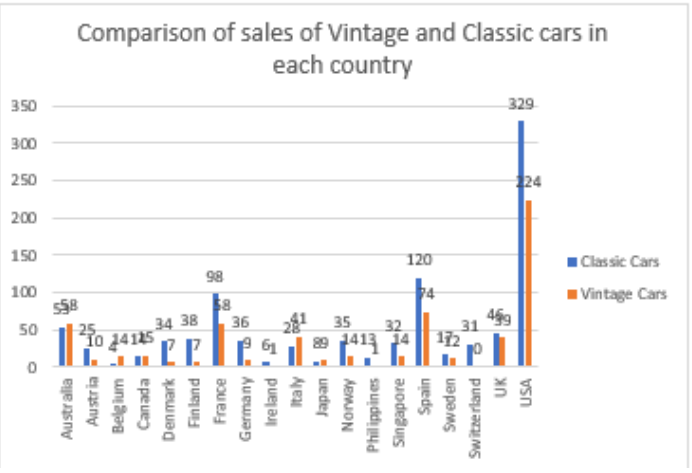
Sales Data Analysis



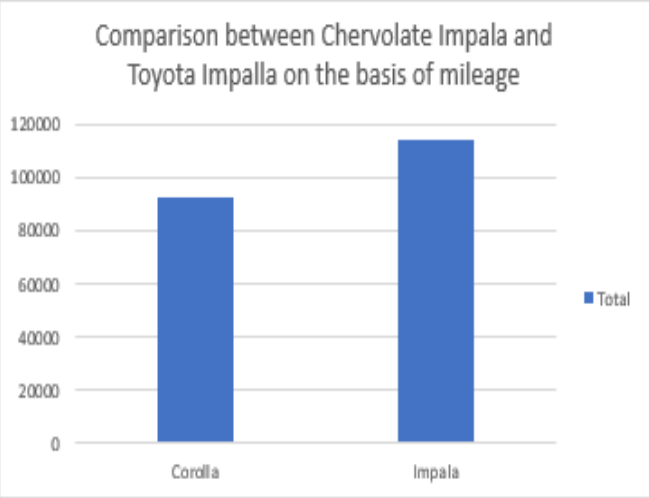


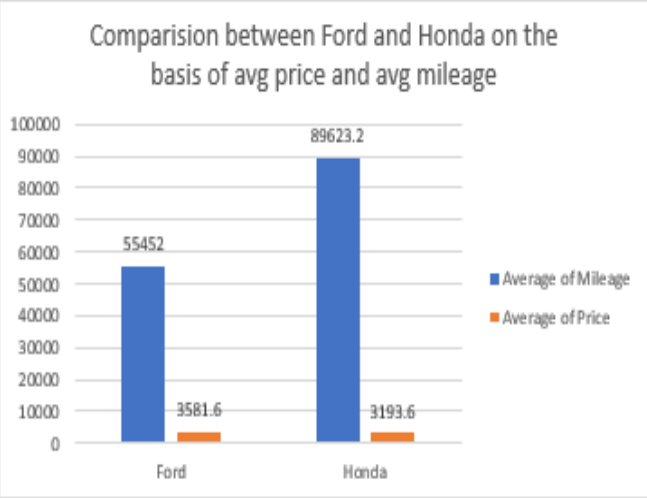


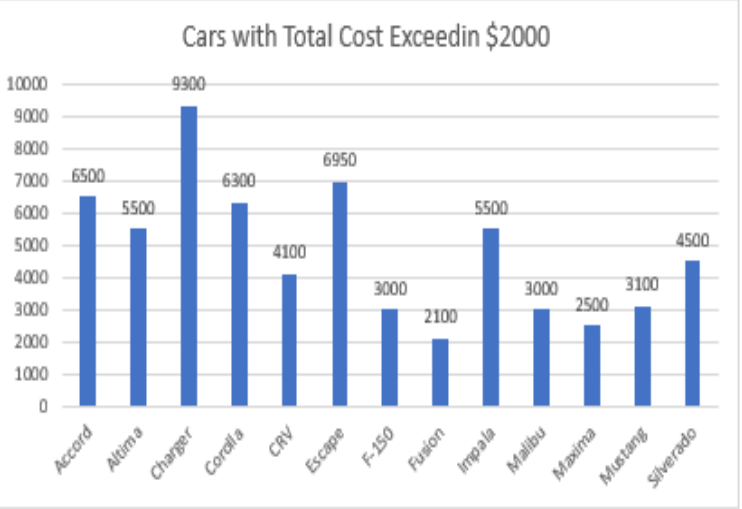


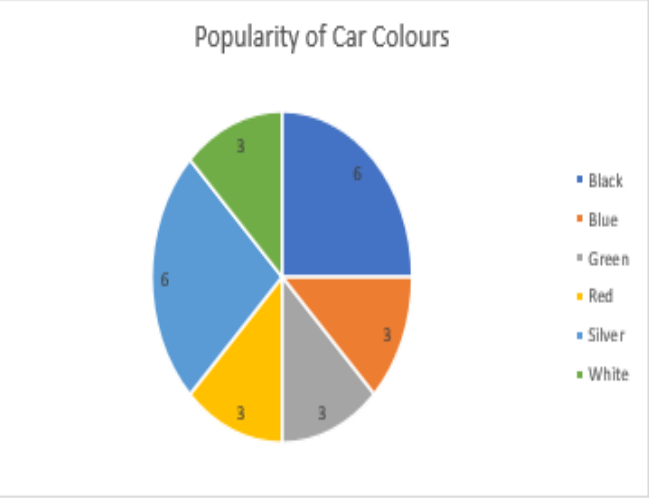


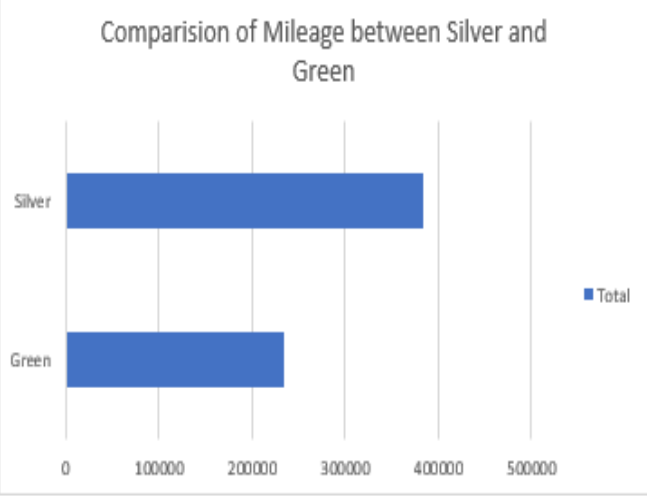
Car CollectionDataset



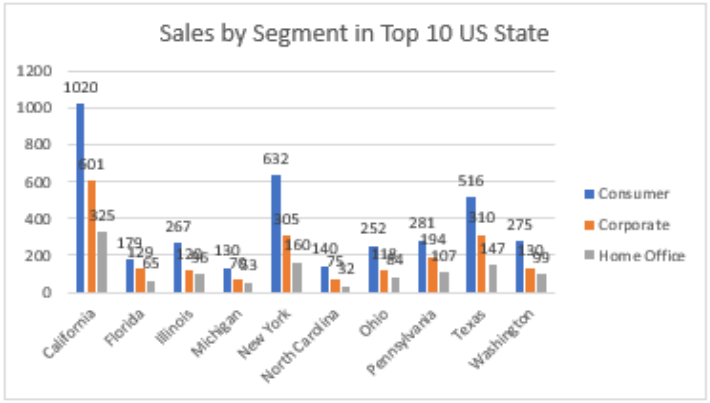


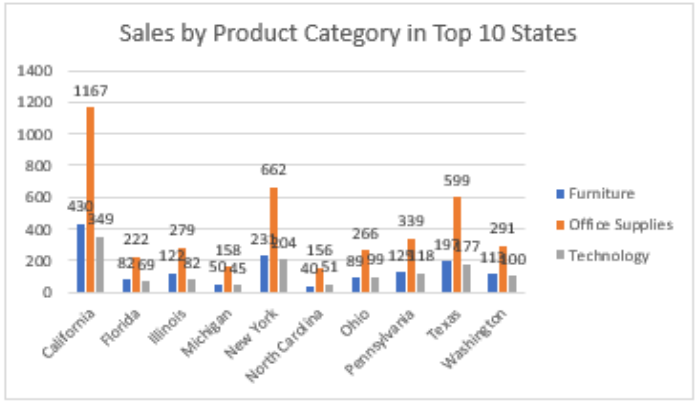


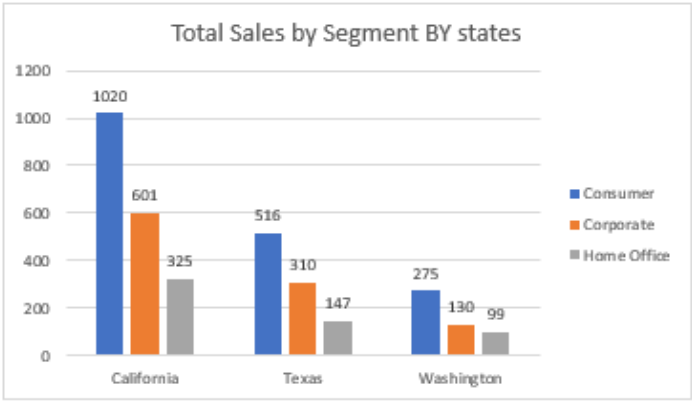


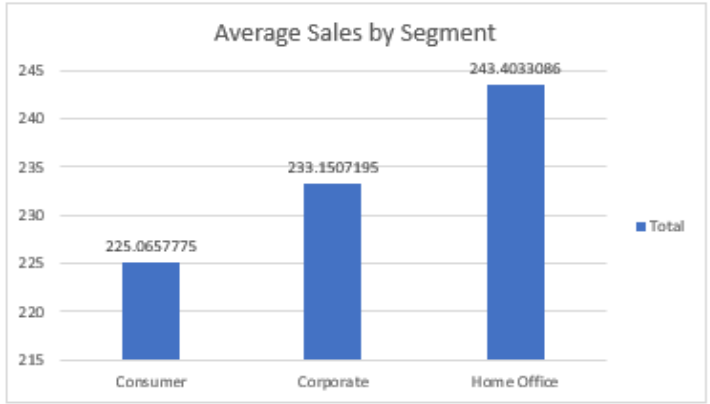


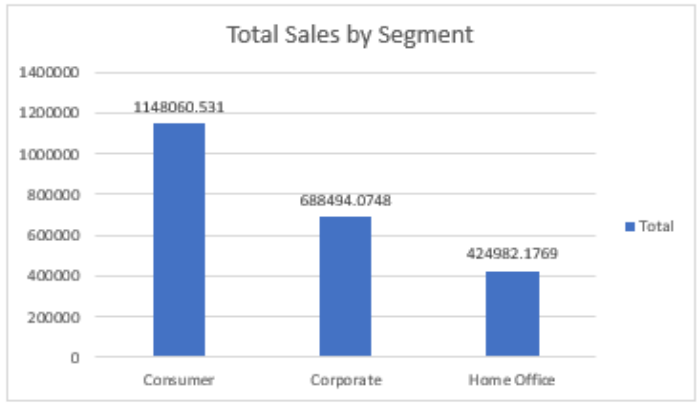
Orders Data Analysis

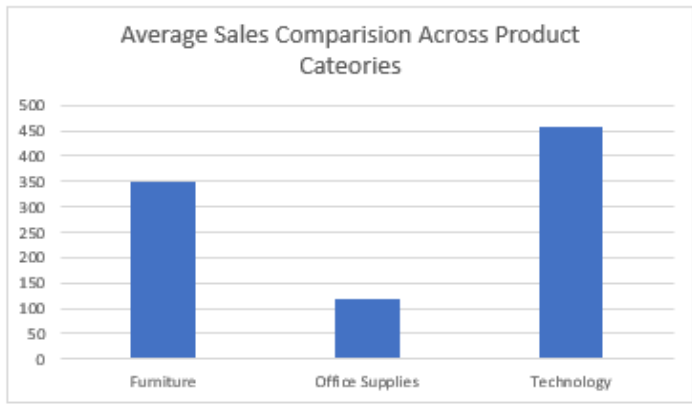


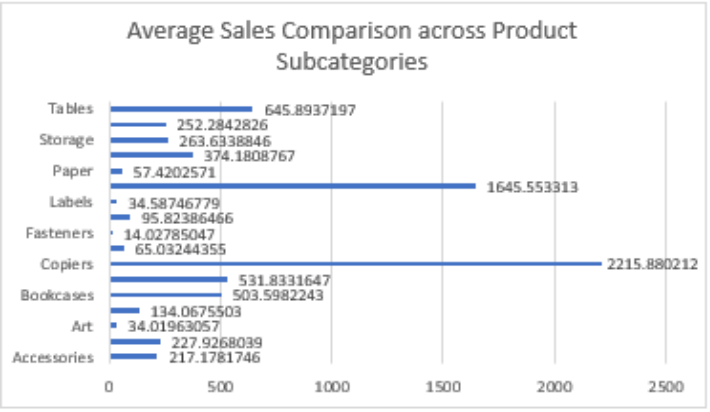


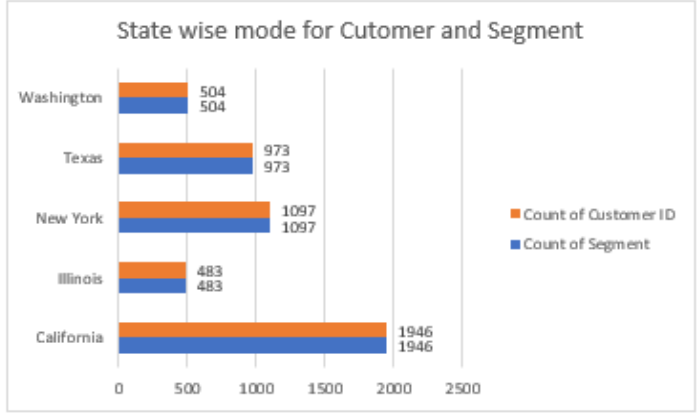




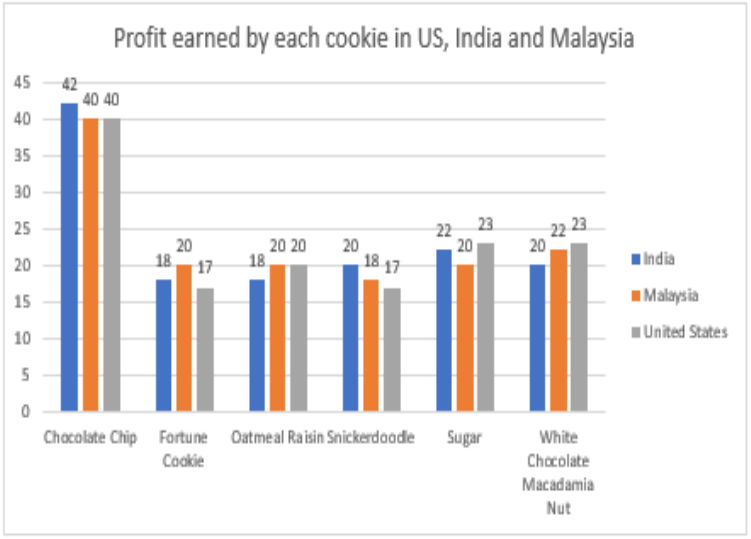


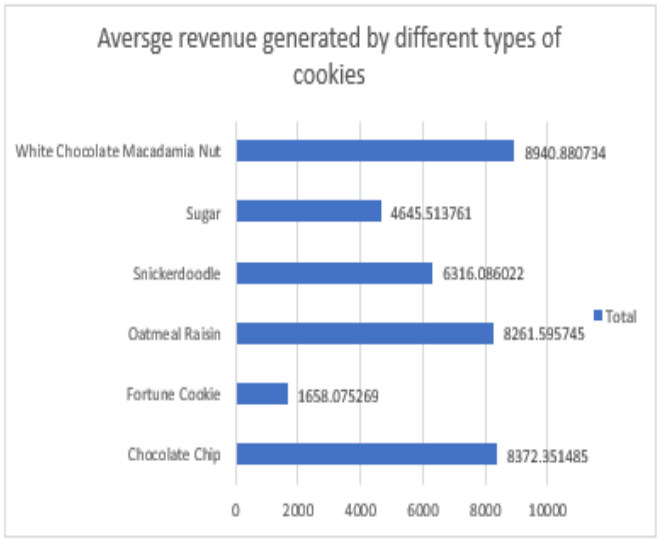


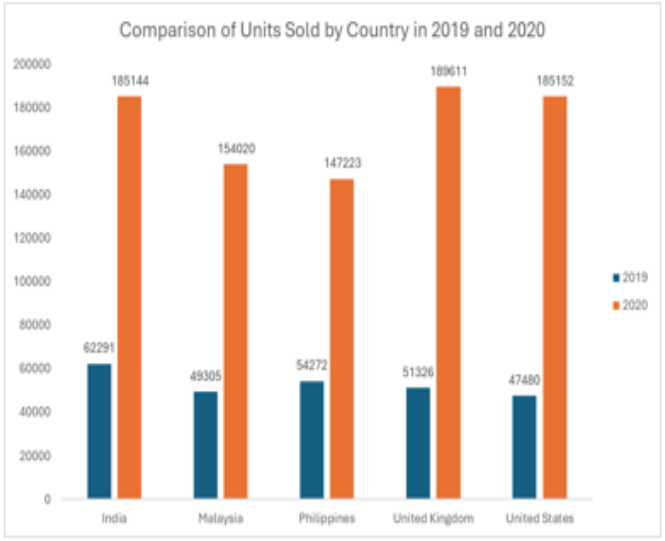


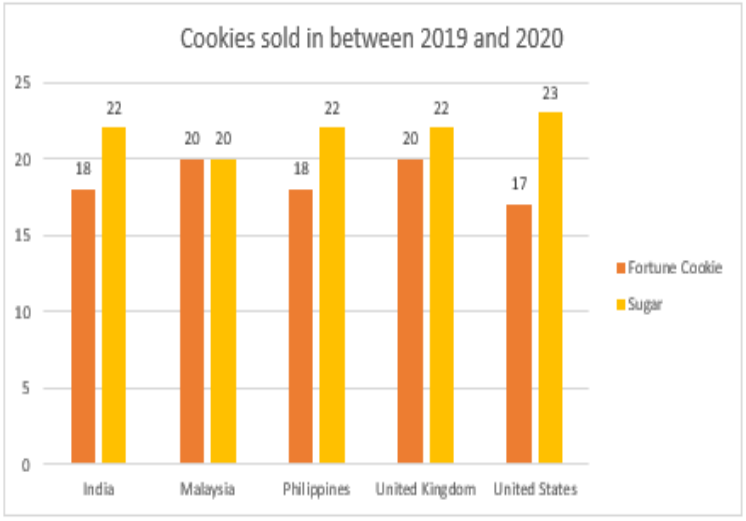


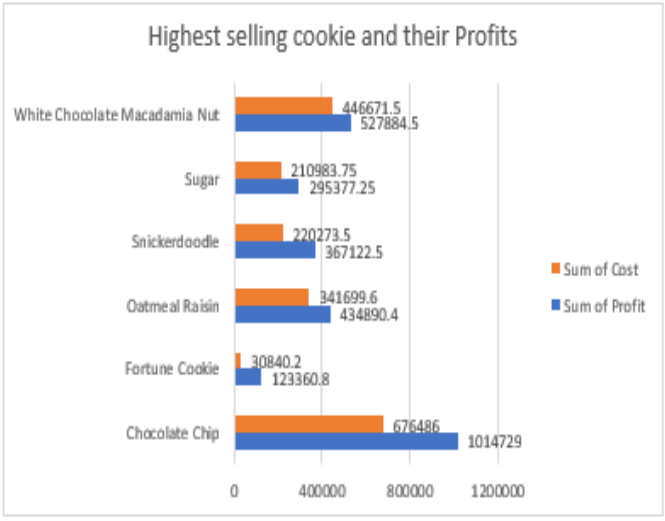
Cookie Data Report



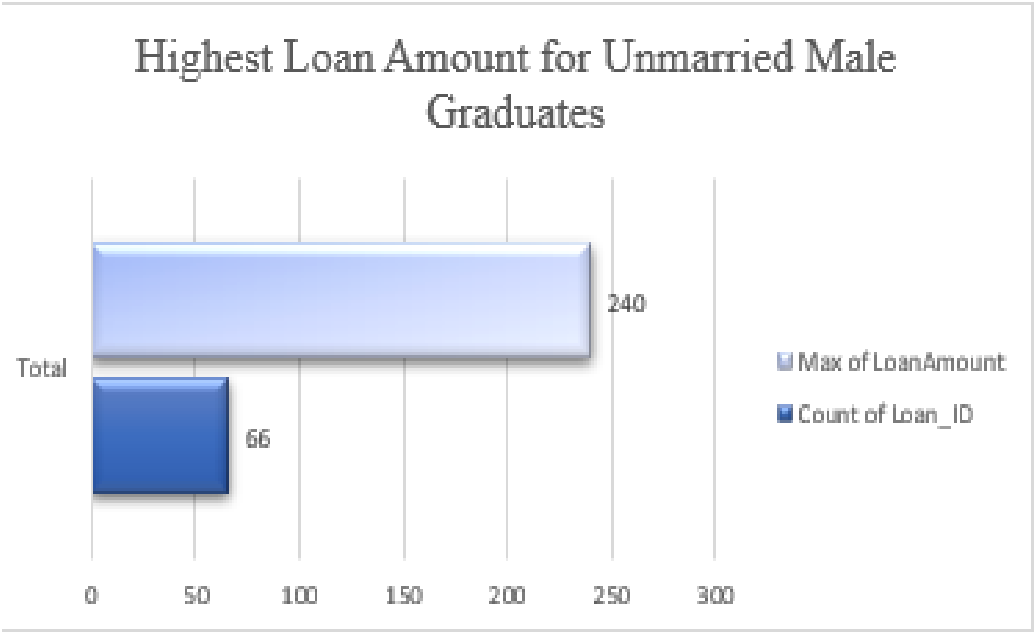


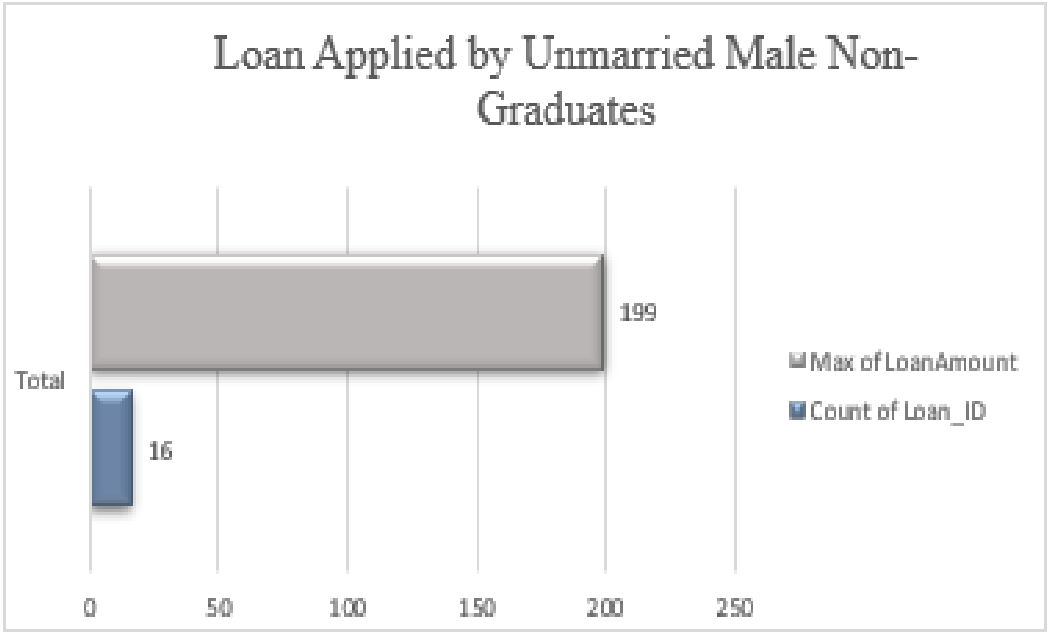


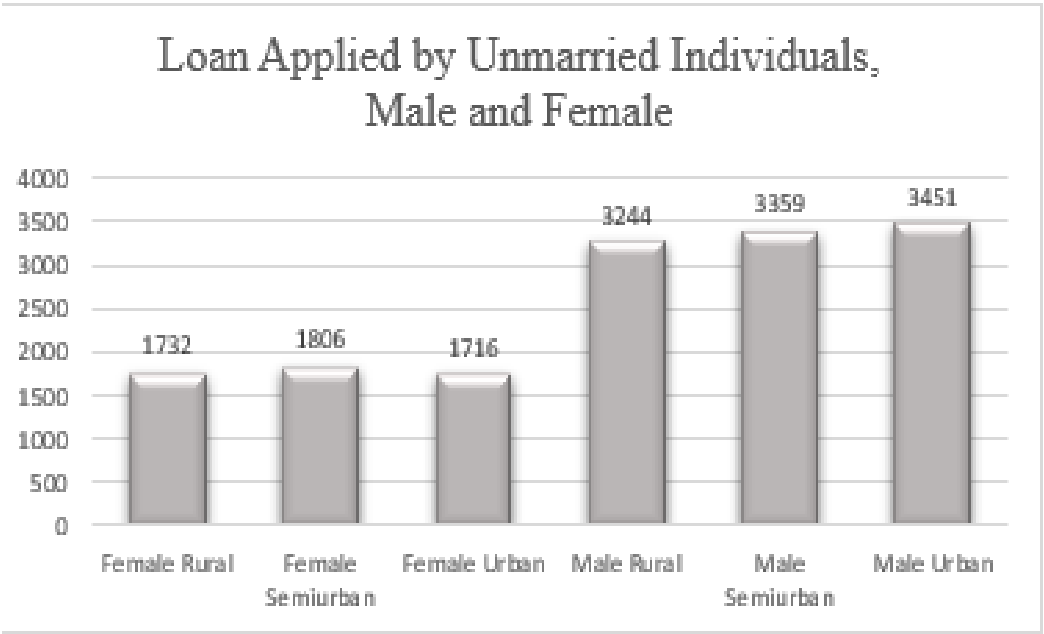


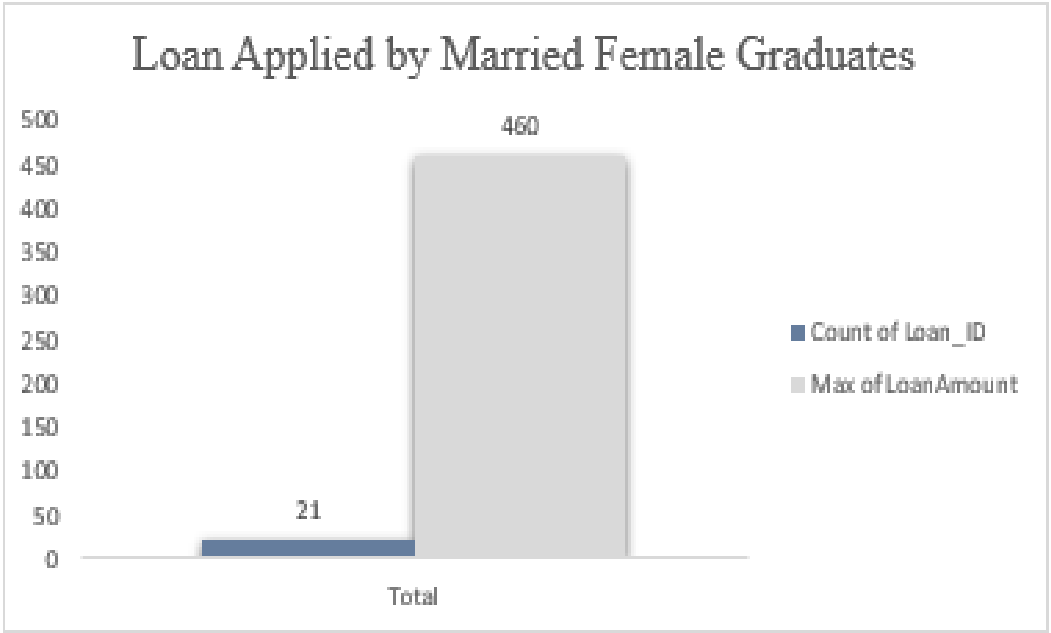


Analysis of Loan Applicants

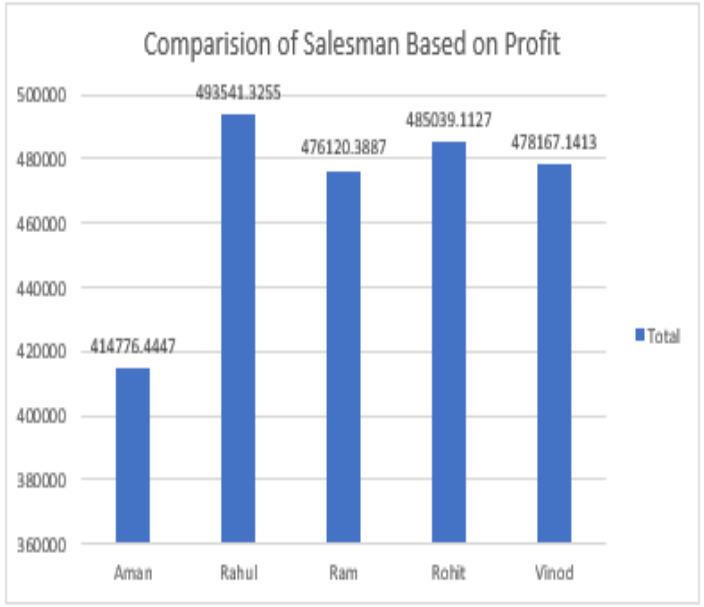


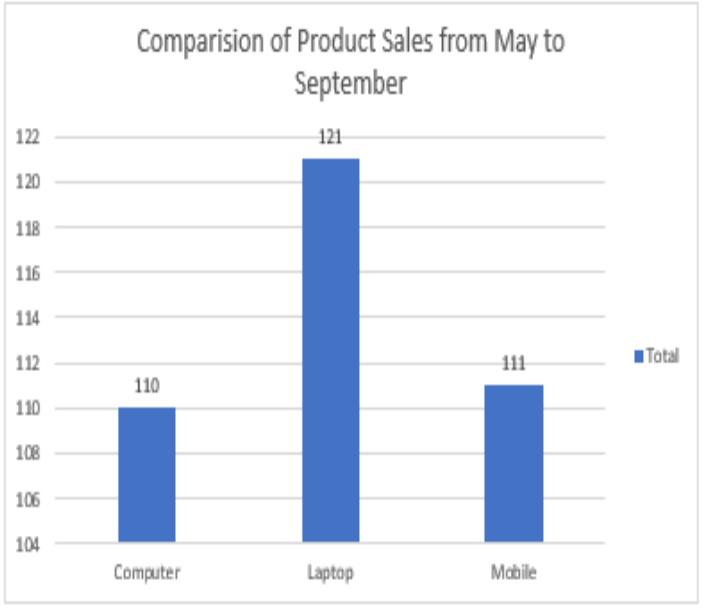


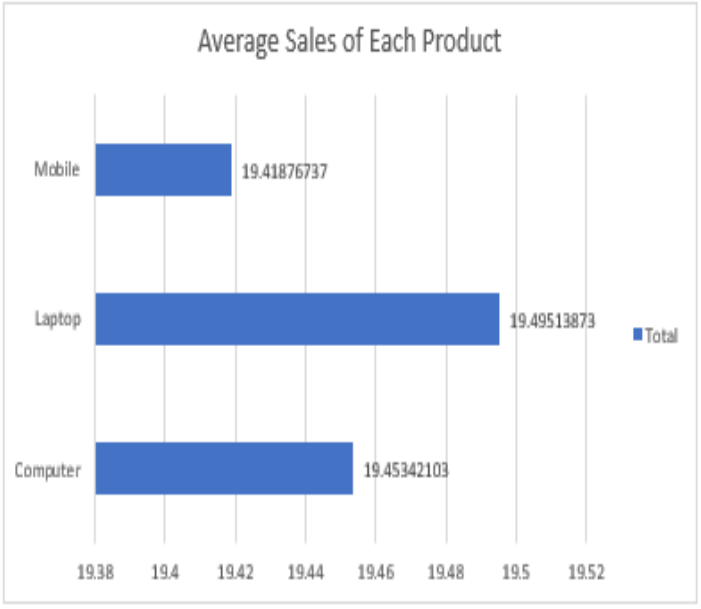


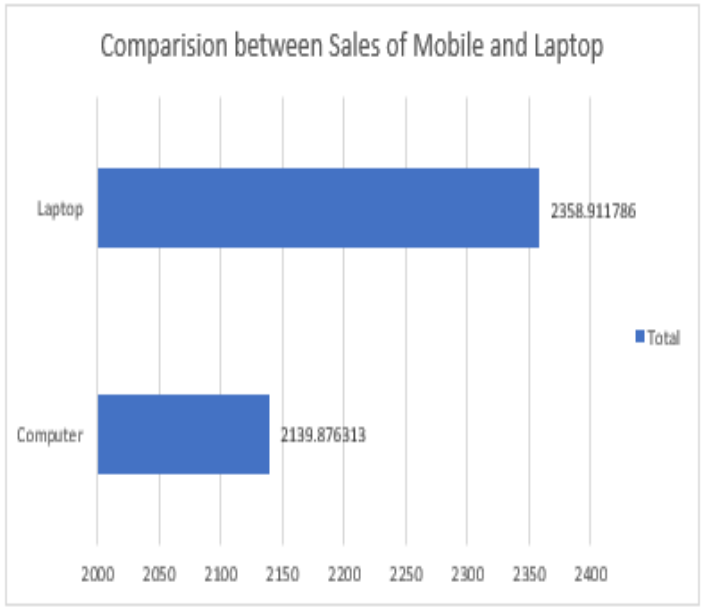


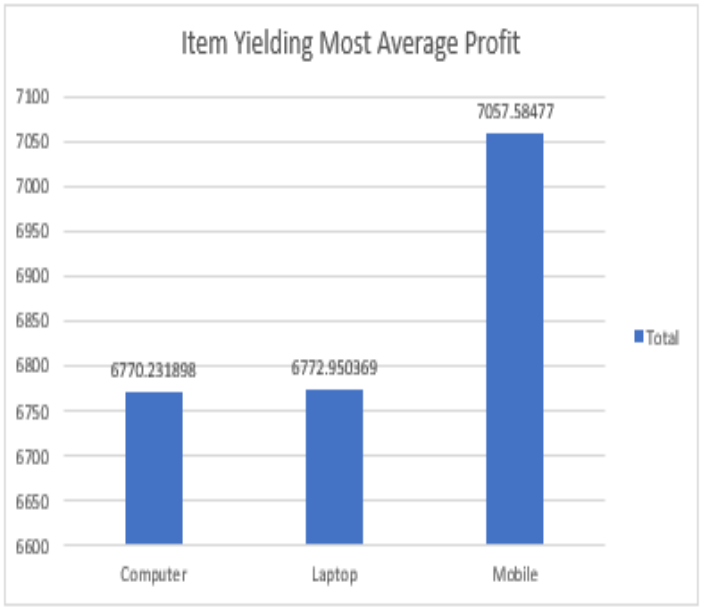
Sales Data Analysis











Store Data Analysis

Introduction

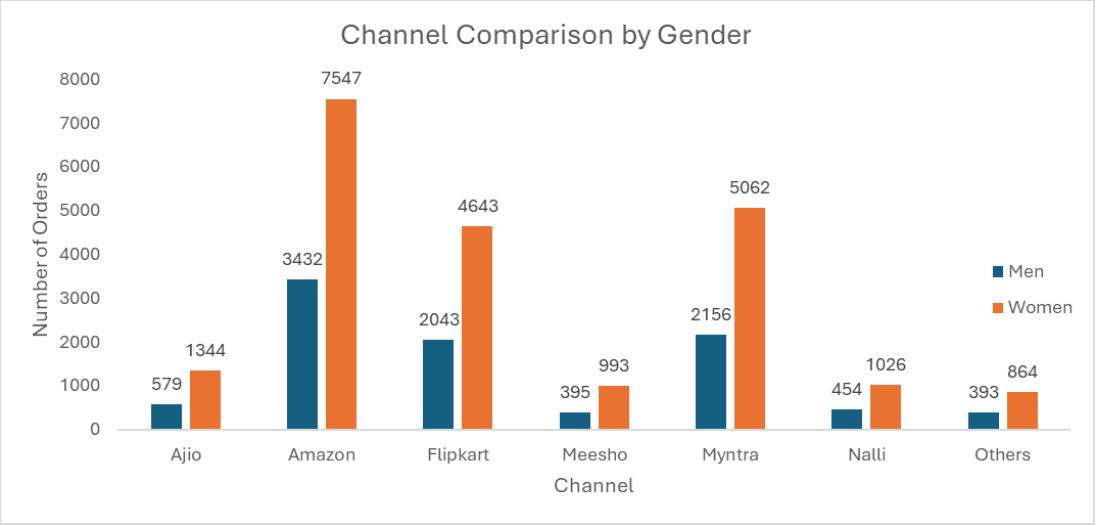
The "Store Data" dataset comprises transactional data from a store, encompassing various attributes such as order ID, customer ID, demographic information, product details, and shipping information. This dataset holds substantial value in uncovering insights related to customer behavior, product preferences,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

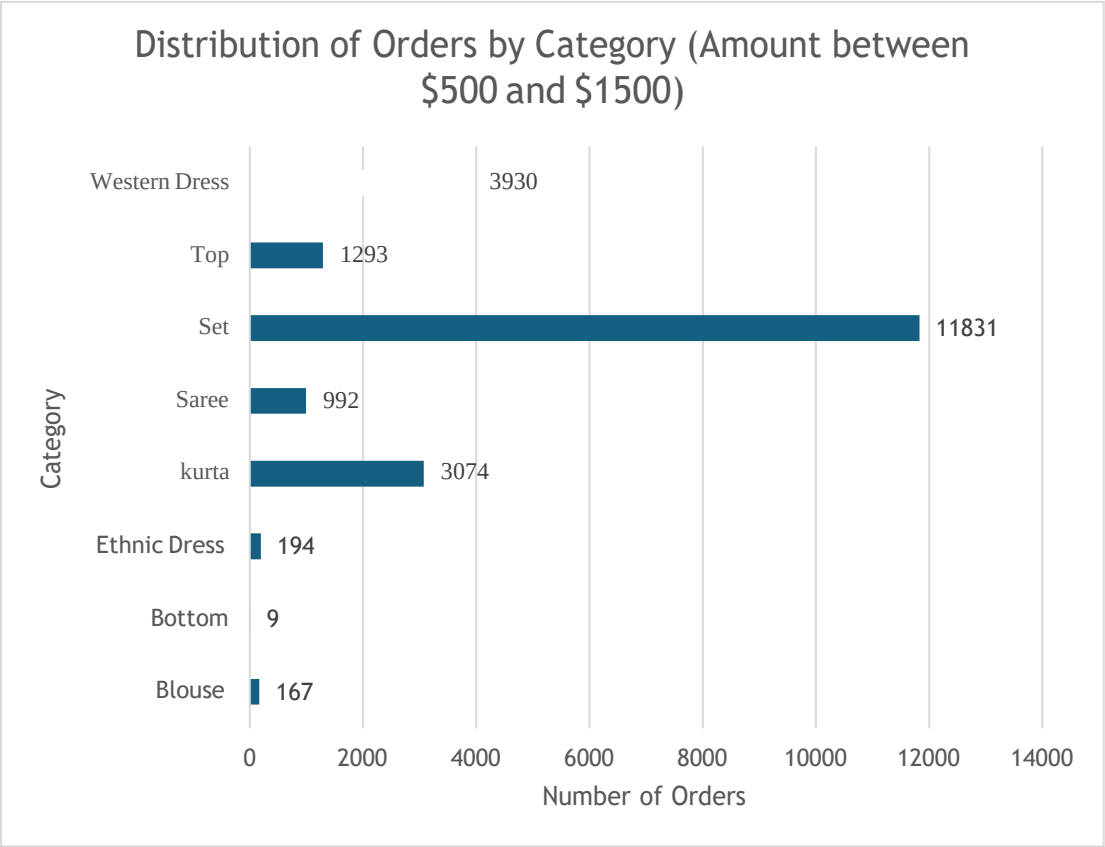
Analytics

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



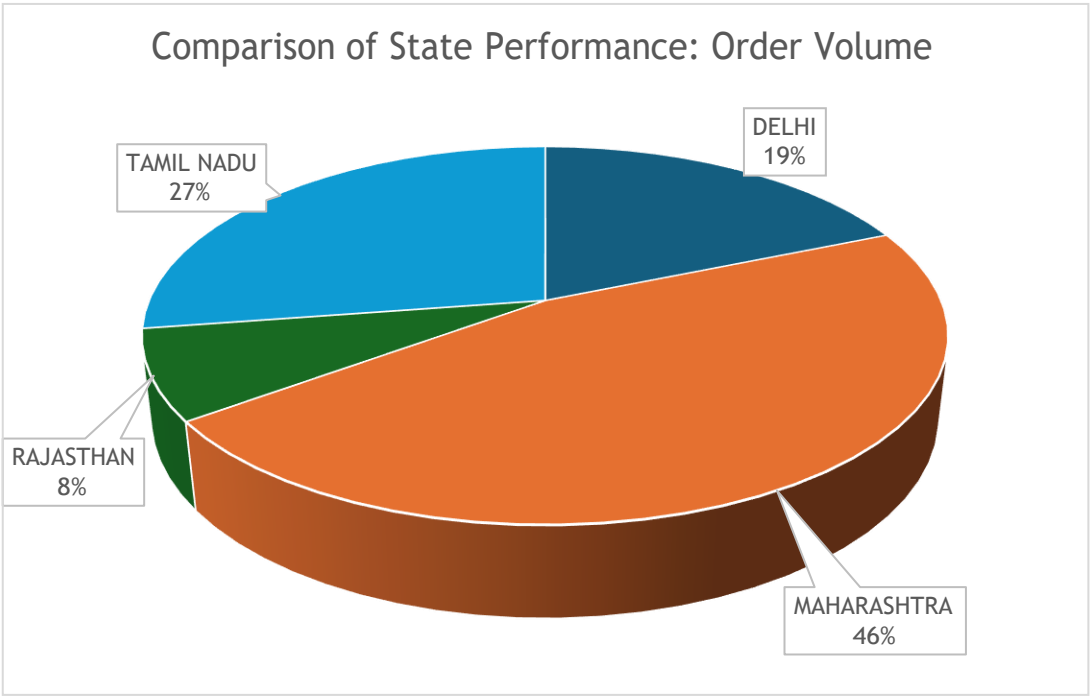
After analyzing the distribution of orders across different sales channels by gender as shown in below graph, it is evident that men have a higher purchase frequency in all channels. This finding highlights the importance of understanding and catering to gender-specific preferences in marketing strategies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

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



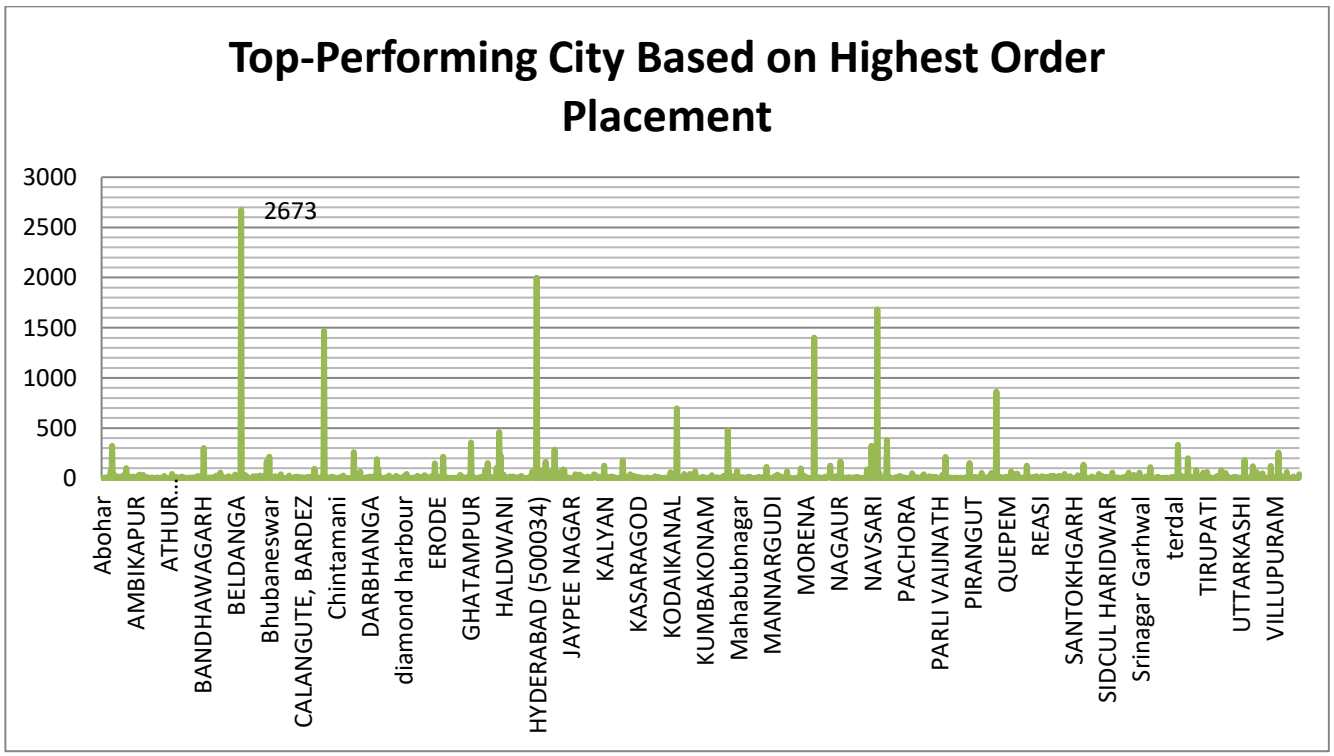
The analysis of orders within the $500 to $1500 range as shown in below graph indicates a diverse distribution across different product categories. Notably, categories such as 'Set' and 'Kurta' emerge as the most popular choices among customers, with 11,831 and 3,074 orders respectively. These findings 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. **Which of the following state perform better than others: Delhi, Tamil Nadu, Maharashtra, Rajasthan?**



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.

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



The analysis of order volume across cities identified the top 10 cities with the highestnumber of orders placed. Bengaluru emerged as the leader with 2,673 orders, followedby 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 byChennai 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

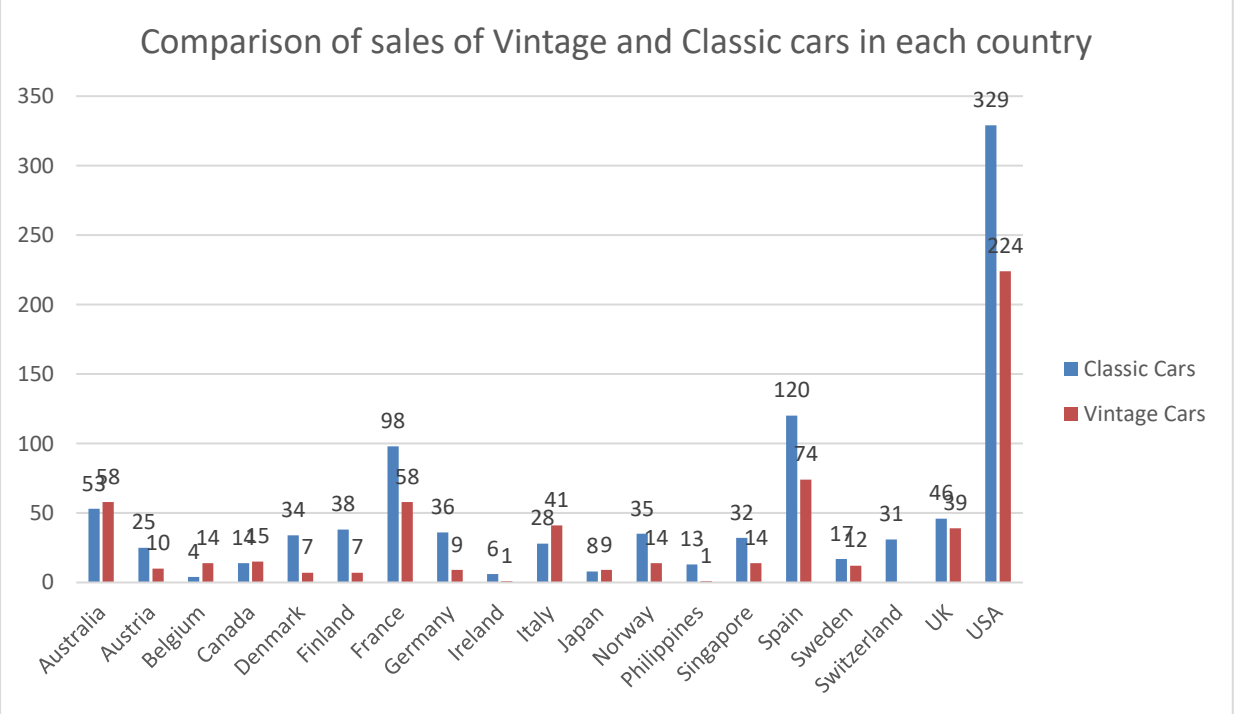
An assortment of attributes, including order number, quantity ordered, price per, order line number, sales revenue, order date, status, product line, MSRP, product code, customer name, phone number, address details, city, state, postal code, country, territory, contact names, and deal size, are included in the "Sales Dataset" which is a collection of transactional records from a store. This dataset is very valuable for revealing patterns in sales, the performance of products, the behaviour of customers, 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:

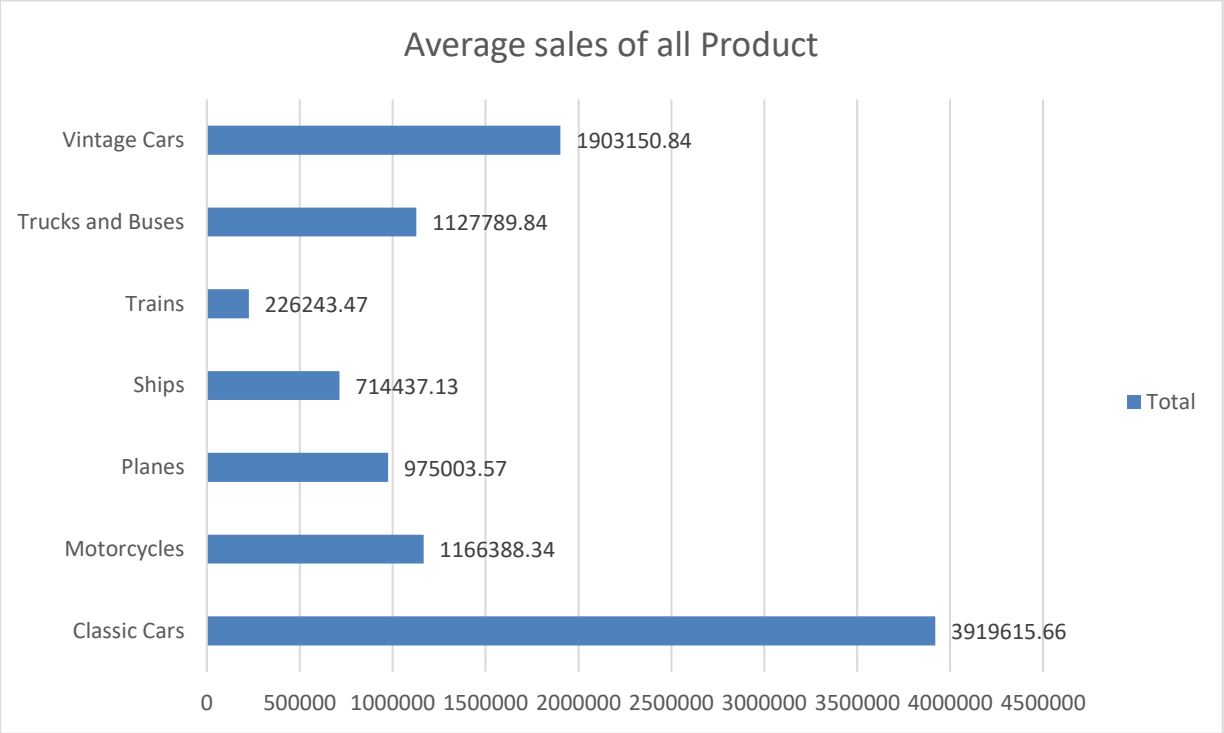


The comparison of sales between Vintage cars and Classic cars across various countries reveals interesting insights into the popularity of two categories:

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, demonstrating a notable preference for Vintage cars with total sales amounting to $229,514.51, 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.

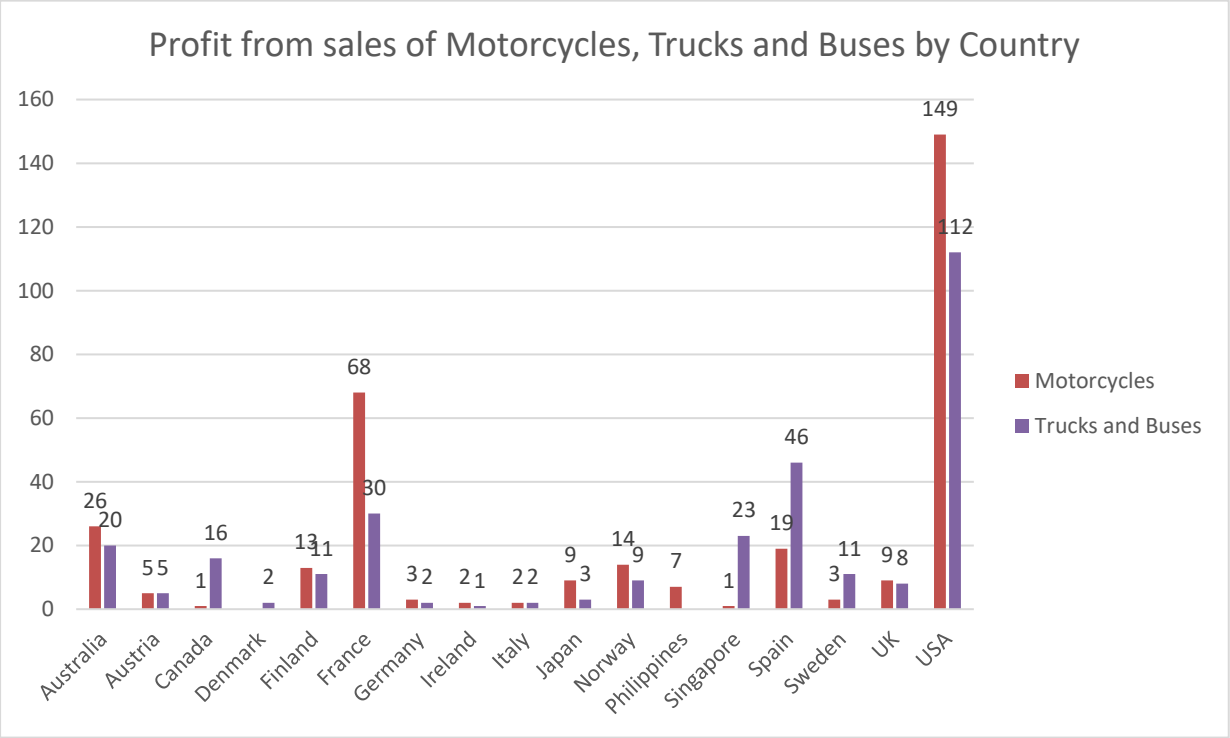
2. **Find out average sales of all the products? which product yield most sale?**



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

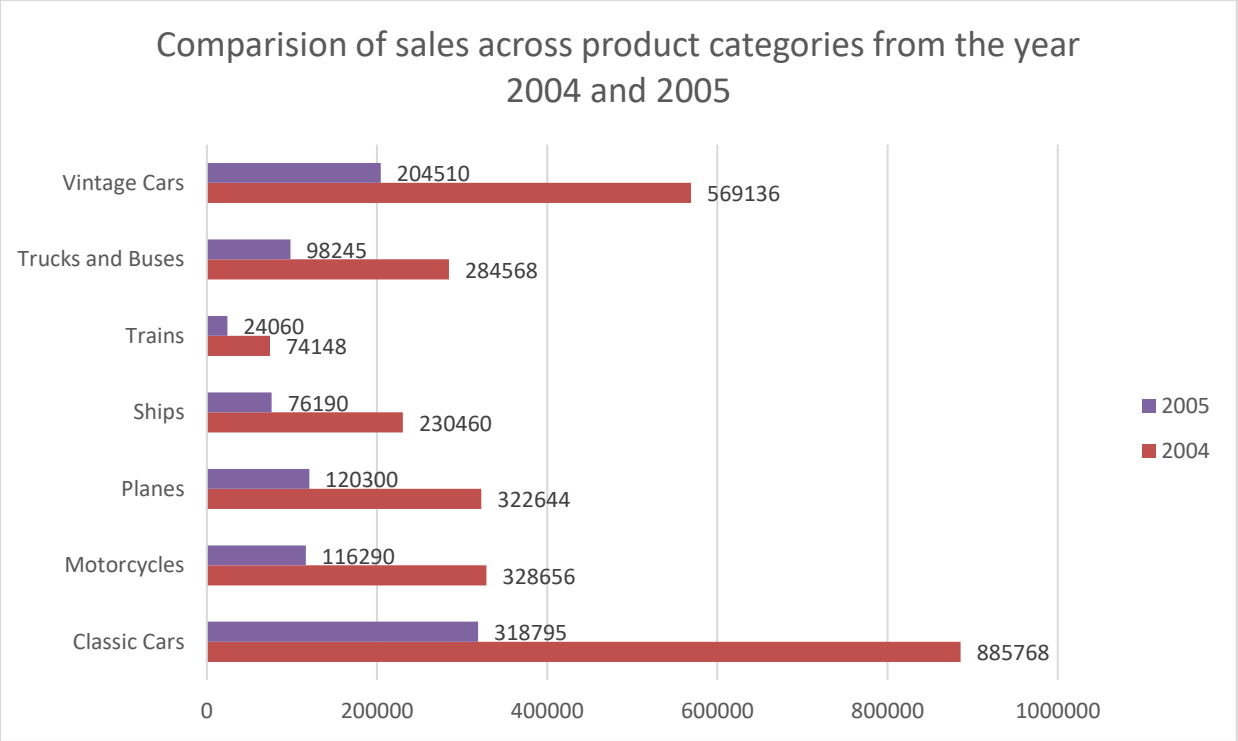
follow closely behind with an average sales value of $3,746.81. This data provides valuable insights into the relative sales performance of different product categories, enabling targeted resource allocation and strategic decision-making.

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



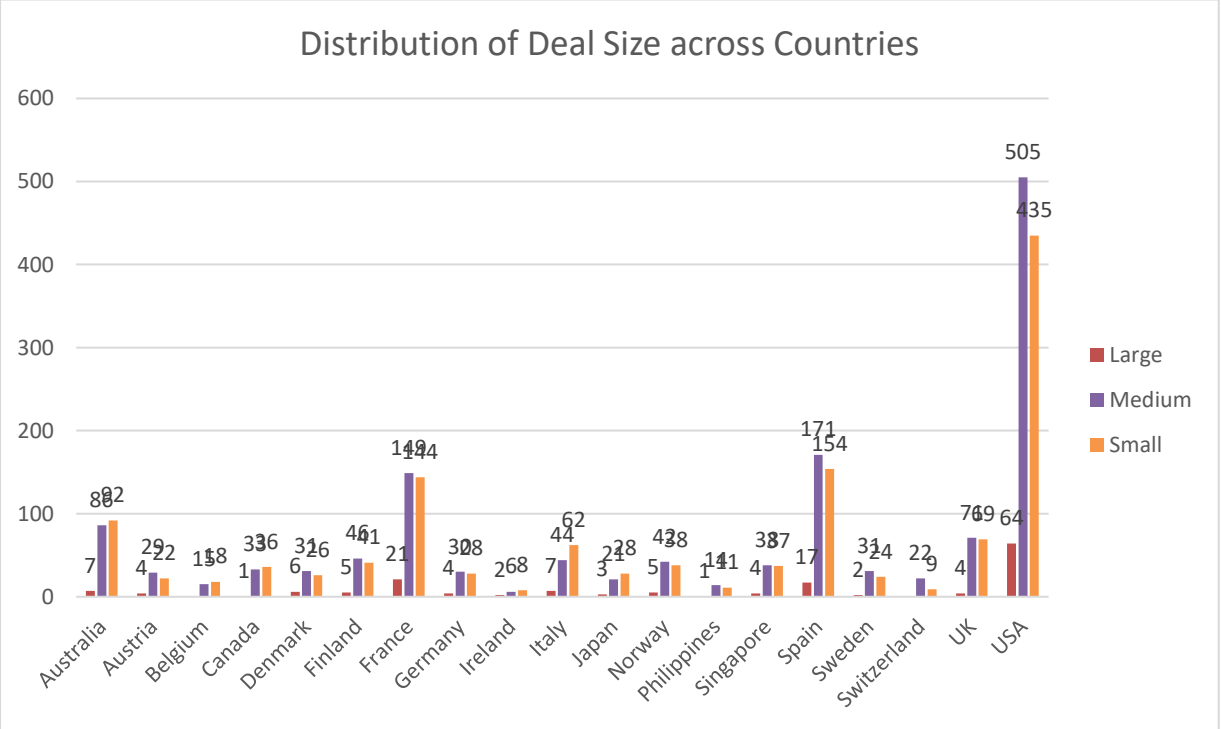
Finding the country that generates the most revenue from trucks, buses, and motorbikes is the aim of this investigation. A bar graph shows that the USA leads the globe in motorcycle sales (520371.7), followed by truck and bus sales (397842.42) and motorcycle sales (France and Spain).

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



Comparing the sales of each item in 2004 and 2005 is the aim of this investigation. With the exception of historic cars, which had the highest sales of any category in both years—1762257.09 in 2004 and 672573.28 in 2005—the line chart demonstrates how quickly the sales of every item are changing.

5. Compare all the countries based on deal size.



The examination of transaction sizes in different nations provides fascinating information about how deal sizes vary within each area. With 64 large agreements overall, the USA leads

the world in large deal sizes, followed by France and Spain with 21 and 17, respectively. The United States of America maintains its top spot in terms of medium-sized transactions with 505 occurrences, suggesting a substantial volume of medium-sized transactions in the nation. Small deal sizes are more evenly dispersed throughout the world, with Spain and France having the highest numbers (154) and 144, respectively. The United States of America leads the world in the total number of deals made, regardless of deal size. France and Spain come in second and third.

Conclusion and Review

We thoroughly examined the sales dataset for this study in order to obtain knowledge about consumer behaviour, product performance, and market trends. To help strategic decision-making, we sought to extract actionable insights by analysing a range of important metrics and patterns.

Several significant discoveries emerged from the analysis:

- Product Category Performance: Different product categories showed varying sales performance. In terms of sales income, "Classic Cars" and "Vintage Cars" stood out as the top-performing categories, indicating a high desire for vintage and classic car models.
- Yearly Sales Comparison: Sales performance over time showed variations when comparing sales from 2004 and 2005. Further exploration into the factors driving these variations could provide insights into market dynamics and consumer preferences.
- Profitability by Country: Differences in profitability were found after sales data from several nations was analysed. France and Spain trailed the USA as the top-performing nation in terms of sales revenue. Comprehending the elements that contribute to these markets' performance could help with resource allocation and expansion plans.
- Deal Size Distribution: Analysing deal sizes across national boundaries provided information on transactional trends. Large and medium-sized deal frequency was highest in the USA, suggesting that there may be room in the market for high-value transactions.
- Client Demographics: Information about client preferences and segmentation was discovered through demographic analysis. More research into consumer demographics may make it possible to develop individualised services and focused marketing campaigns that will increase client loyalty and satisfaction.

Important insights into consumer behaviour, product performance, and market trends were obtained from the investigation.

We advise the following actions to take advantage of these insights:

- More research into consumer segmentation to better target marketing campaigns.
- Constantly observing sales patterns in order to modify strategy as necessary.

- Focused initiatives to increase income by utilising markets and product categories that are performing well.

All things considered, the research lays the groundwork for strategic planning and data-driven decision-making, which helps the company negotiate the complexities of the market and promote long-term growth in the sales sector.

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

Significant results are obtained from the regression analysis between the dependent variable (SALES) and the independent variable (QUANTITYORDERED). With a value of roughly 0.55, the multiple R coefficient shows a moderately positive correlation between the two variables. The R-squared value of 0.30 indicates that changes in the quantity ordered can account for about 30% of the variation in sales.

SALES=−104.92+104.26×QUANTITYORDERED is the regression equation that was found.

According to the coefficients, sales rise by about $104.26 for every unit increase in the quantity ordered. The statistical significance of the link between quantity ordered and sales is indicated by the extremely significant p-value (2.3596E-224) associated with the independent variable's coefficient.

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 results of the ANOVA analysis show that the mean sales of the various product lines differ significantly. Group 2 shows statistically significant variances (p-value = 0), with an average sales of roughly 3554.13 units, and Group 1, with an average sales of about 3.52 units. This suggests that selecting a product line has a big impact on sales, with certain product lines having noticeably higher or lower sales than 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

Descriptive statistics for the columns SALES, PRICEEACH, and QUANTITYORDERED provide valuable information about how well things sell. There is a substantial degree of variability in the order quantities, with a mean of about 35 units and a standard deviation of roughly 9.74. With a standard deviation of roughly $20.17 and an average price per unit of about $83.66, there may be some variation in product prices.

With a significant standard deviation of $1414.87, the mean sales amount per transaction is $3553.89, suggesting a broad range of sales amounts. While the price per unit distribution is slightly negatively skewed, the quantity ordered data distribution is slightly favourably

skewed.

Correlation

	QUANTITYORDERED	*SALES*
QUANTITYORDERED	1	
SALES	0.551426192	1

Approximately 0.55 is the moderately positive correlation coefficient between QUANTITYORDERED and SALES, according to the correlation analysis. This suggests that there is a propensity for larger orders of products to be linked to larger sums of revenue. The imperfect association, however, raises the possibility that additional variables could possibly affect sales income. Although a higher order quantity often translates into higher sales, this association is only moderately strong, suggesting that other factors like pricing tactics, consumer preferences, or market conditions can also affect sales performance. All things considered, this analysis emphasises how critical it is to take into account a variety of aspects when examining sales data and formulating strategic business decisions.

Car Collection Data Report

Introduction

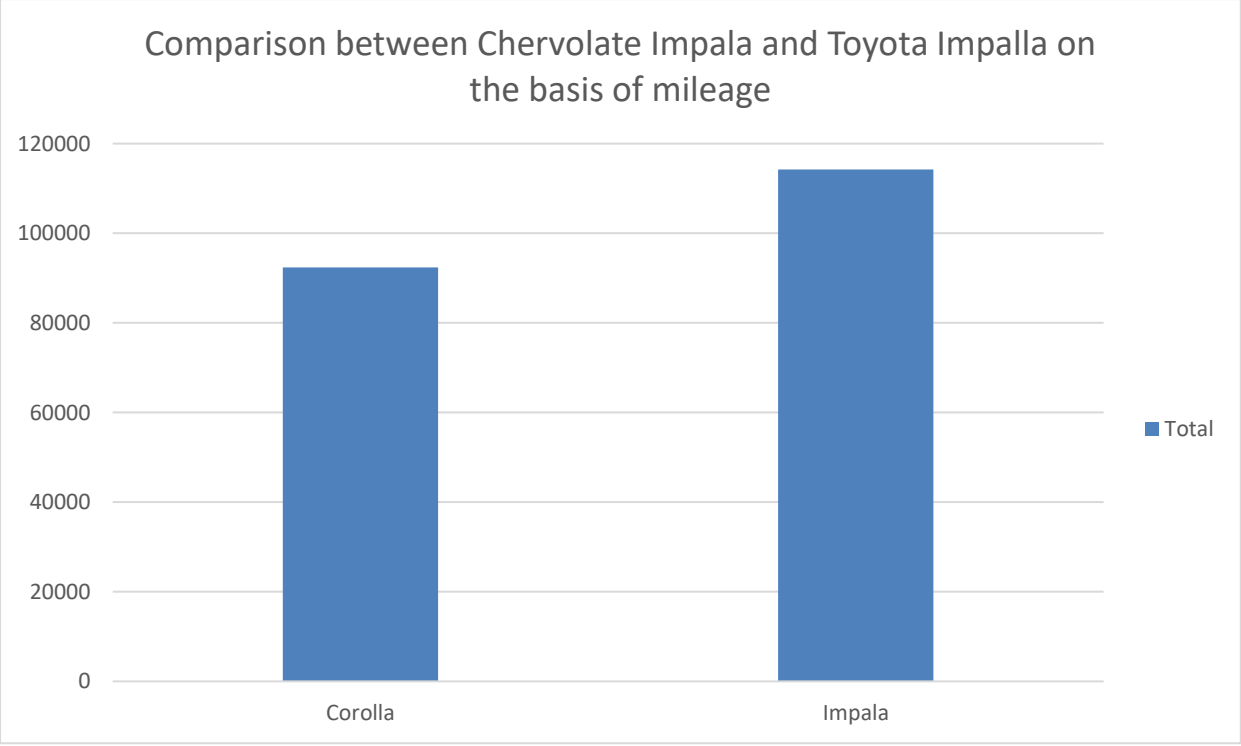
The automobile collection dataset offers an extensive aggregation of data related to different car models, including make, model, colour, mileage, price, and cost. With its insights into a variety of aspects of the automobile industry, this dataset is a useful tool for stakeholders such as auto dealerships, automotive analysts, and consumers. Stakeholders can identify market trends and preferences by carefully examining the data, which facilitates well-informed decision-making. The paper attempts to respond to relevant questions posed by interested parties, such as mileage comparisons across car models and arguments in favour of Ford over Honda. It also explores colour popularity analysis, comparing the mileage of silver and green cars, and identifying cars that cost more than $2000 in total. By utilising the deep insights.

Questionnaire

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

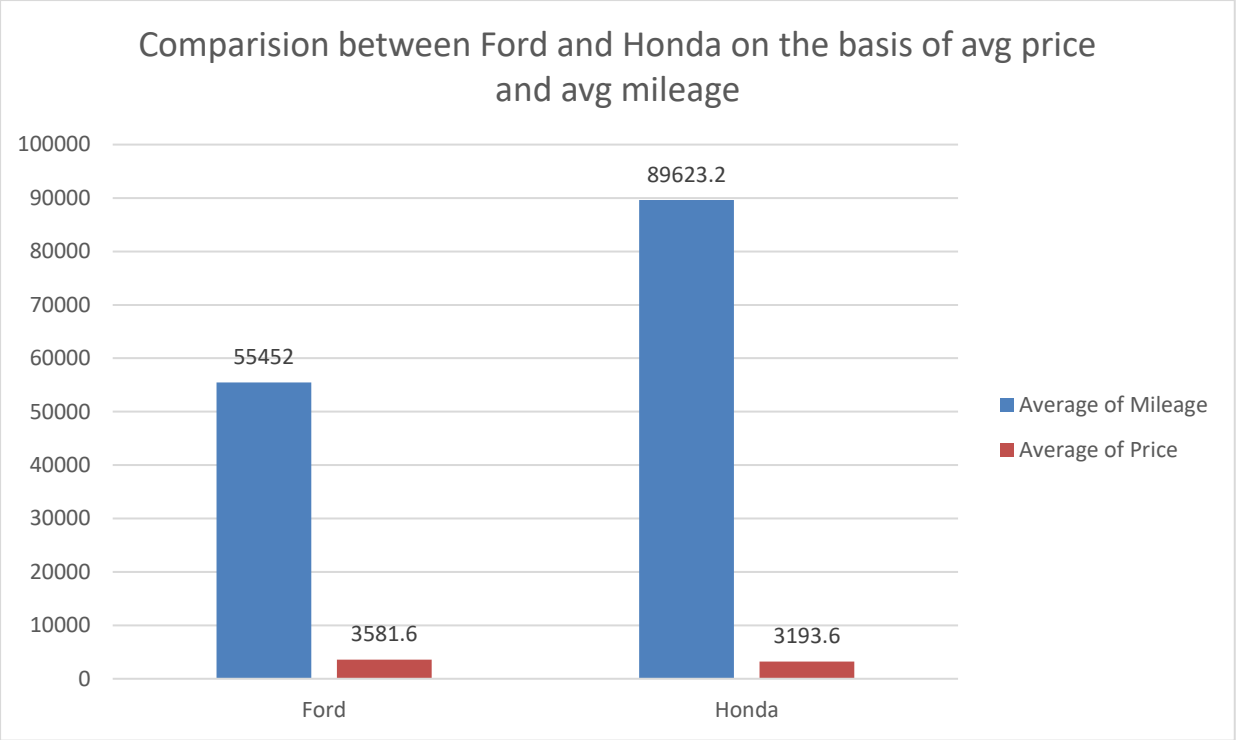
Analytics

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



Significant variations in fuel efficiency between the Chevrolet Impala and Toyota Corolla are shown when comparing their mileage. The Toyota Corolla has a lower average mileage of around 92,377 miles, whilst the Chevrolet Impala has a higher average mileage of about 114,243 miles.

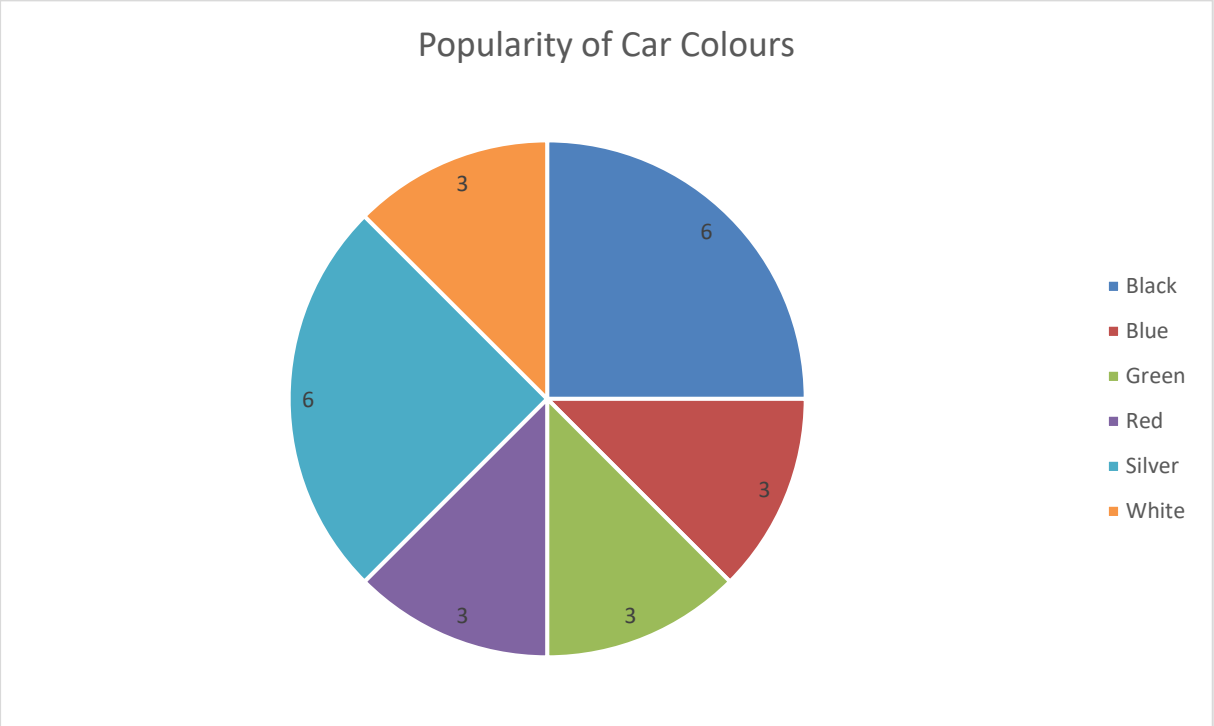
2. **Justify, buying of any Ford car is better than Honda:**



This study tries to support the purchase of any Ford vehicle over a Honda by comparing their respective attributes and with a particular emphasis on price.

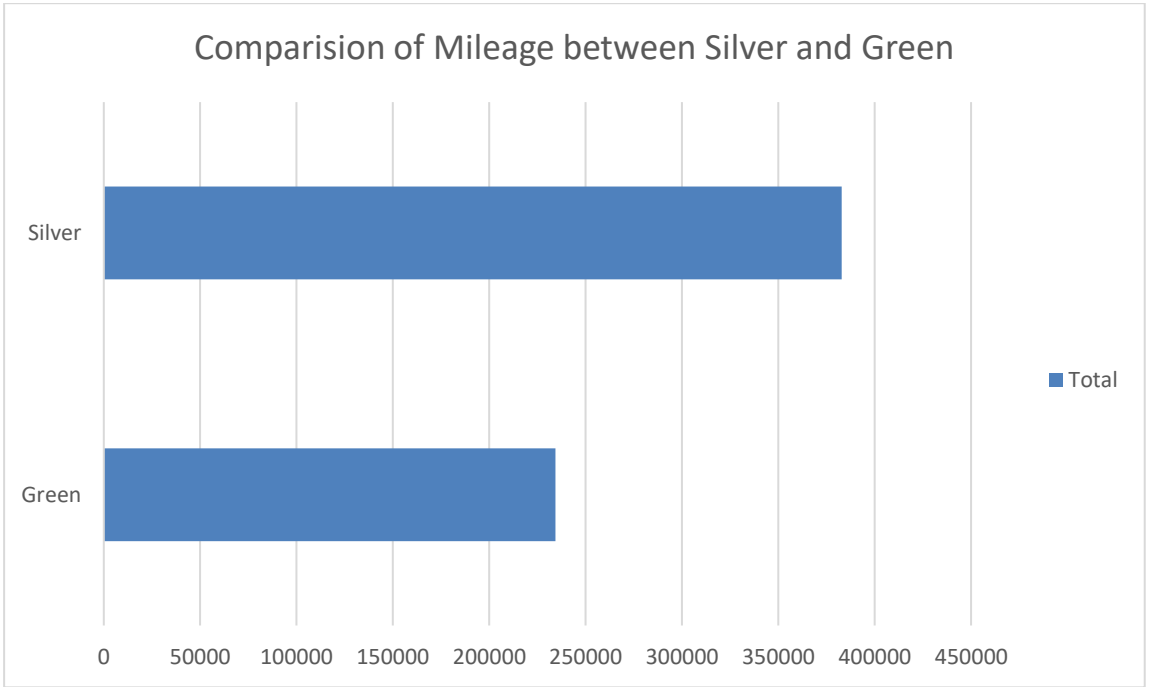
The assertion was refuted by the dataset analysis, which showed that Honda cars perform better than Ford cars in terms of average price and mileage.

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



Based on the count of the make, this study seeks to determine which car colours are the most and least common among all the cars in the dataset. The data indicates that the two most popular car colours are silver and black, which make up 25% of the company's manufacturing, and blue and green cars, which make up 12% of the total.

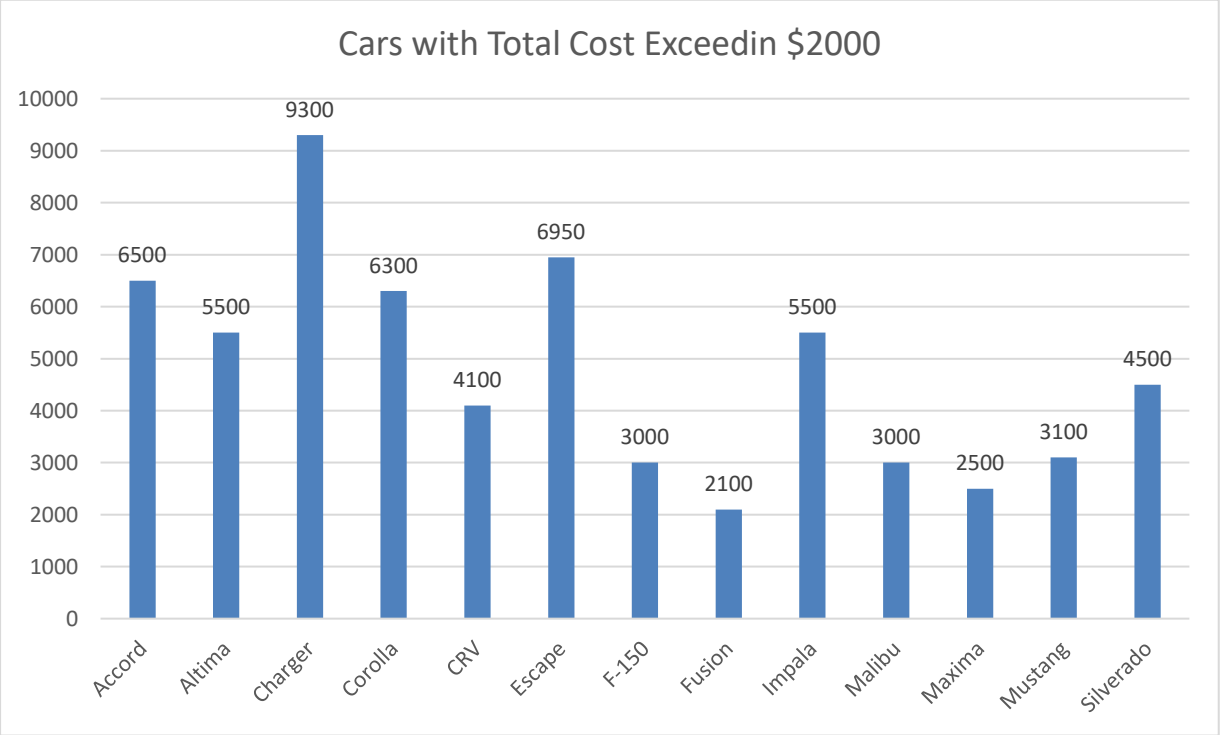
4. Compare all the cars which are of silver color to the green color in terms of Mileage.



Finding out which cars are silver to green in terms of mileage is the aim of this investigation. There are five silver automobiles, according to the results: the Charger, Accord, Mustang, Impala, and Corolla. The Accord has the highest average mileage (101354) out of all of them. An Altima and a Silverado, with the latter having the most miles (109231), were the two green

vehicles.

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



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

Conclusion and Review

The examination of the automobile collection dataset yielded significant findings for a range of car model attributes, including mileage, cost, popularity of colour, and overall expenditure. Five main questions were answered in a methodical manner, each of which provided insight into different aspects of the automobile sector. The mileage of the Toyota Corolla and Chevrolet Impala was compared, and the results showed that the Impala got better mileage than the Corolla, which let customers make selections based on fuel efficiency. Data showing a lower average price for Ford cars—possibly providing greater value for money than Honda cars—supported the argument for purchasing any Ford vehicle over a Honda vehicle. Black and silver were shown to be the most common car colour combinations, showing consumer preferences for these traditional choices.

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 study shows that there is a significant correlation (p-value = 0.0557) between car price and mileage. The coefficient for mileage is -0.0438. The coefficient indicates that the price of the car drops by about $43.81 on average for every unit increase in mileage, even if the significance level is barely significant at the traditional 0.05 threshold. With an R-squared of 0.157, mileage alone appears to be able to explain 15.7% of the variation in car pricing. It's crucial to remember, though, that costs of cars could also be influenced by other variables not taken into account by the model.

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				

A substantial variation in mean pricing between car makes is revealed by the single-factor ANOVA analysis ($F(1, 46) = 303.075$, $p < 0.001$). The between-groups variation, which sums up to roughly 126,841,016.3, indicates significant disparities in mean pricing among different car makes. This suggests that there is a far greater price variance between car manufactures than there is within each group. The findings indicate that the vehicle's make has a major impact on its price, underscoring the need of taking make into account when examining 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 features of the Age and Amount variables are clarified by the descriptive statistics analysis. Age shows a slightly positively skewed distribution, with a mean of roughly 39.5 years, a median of 37 years, and a mode of 28 years. The standard deviation, which is roughly 15.12, indicates that there is a substantial degree of age variability, with ages ranging from 18 to 78. Conversely, the Amount variable displays a greater degree of fluctuation, with a median of 646 units and a mean expenditure of roughly 682.07 units. The data are more widely distributed around the mean, with a significant range from 229 to 3036 units, as indicated by the standard deviation of roughly 268.58. The skewness and kurtosis values suggest that the distribution of Amount is moderately positively skewed and leptokurtic,

respectively.

Correlation

	Price	*Mileage*
Price	1	
Mileage	-0.39561	1

Price and mileage have a moderately negative association coefficient of about -0.396 ($p < 0.05$), according to the correlation analysis. This suggests that the price of cars and their mileage have a statistically significant inverse connection. The price of the car usually drops as the mileage rises. Although the association is not very significant, it does indicate that cars with higher miles typically have cheaper prices, which is in line with what most consumers expect from the automobile industry. This research emphasises how crucial it is to take mileage into account when determining car costs because it has a big influence on how much a vehicle is thought to be worth and how marketable it is.

Order Data Report

Introduction

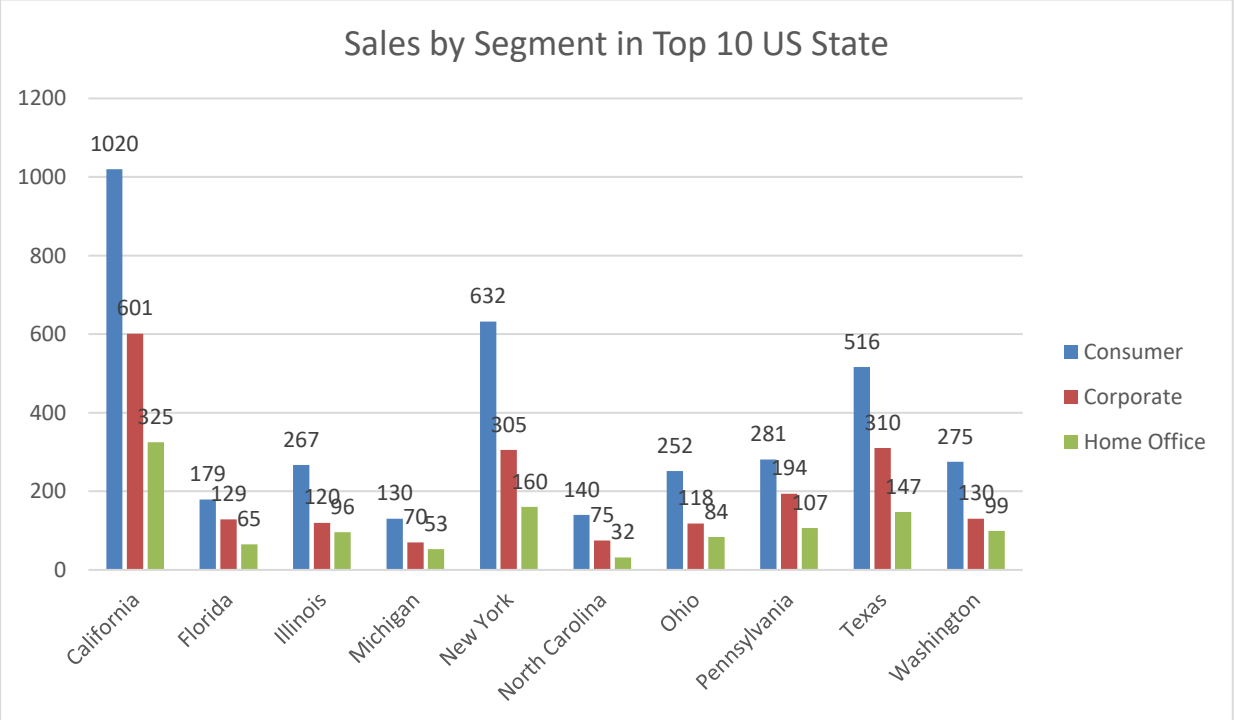
This dataset includes important details about our orders, including Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, and Sales. This dataset is an invaluable tool for strategic decision-making and market optimisation, as it focuses on comprehending geographical variances and segment-specific trends.

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

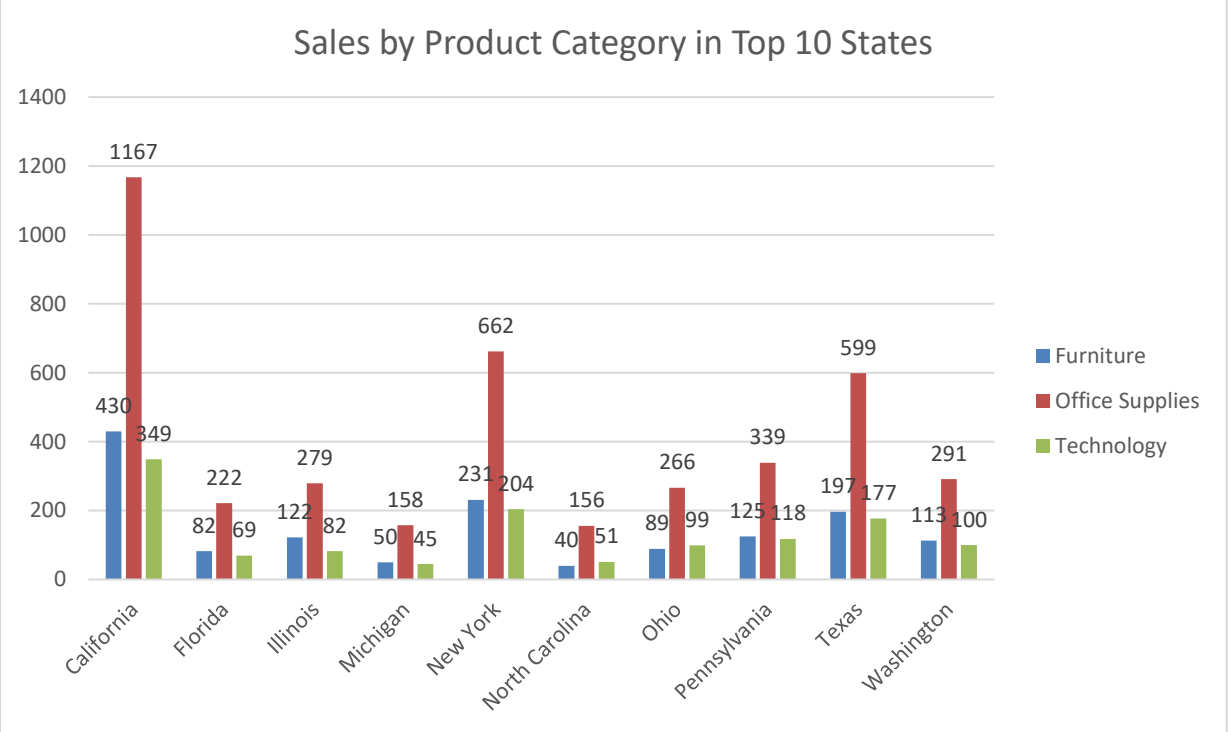
1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states:



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

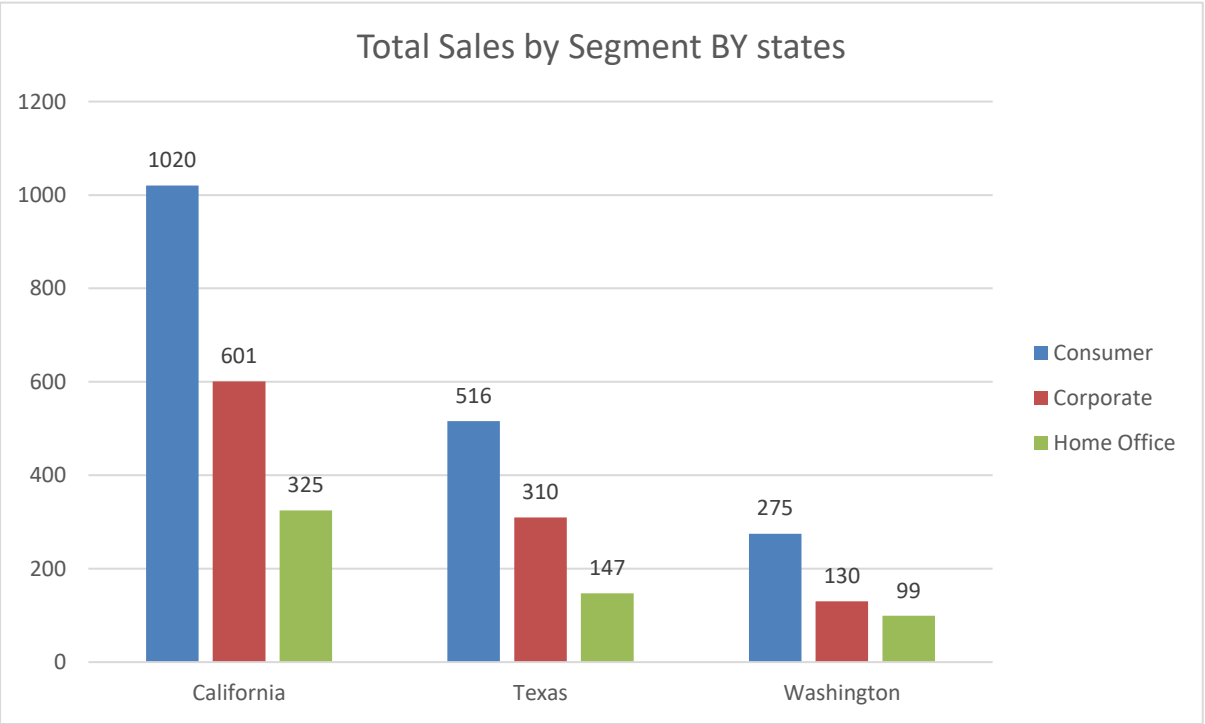
across all states.

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



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

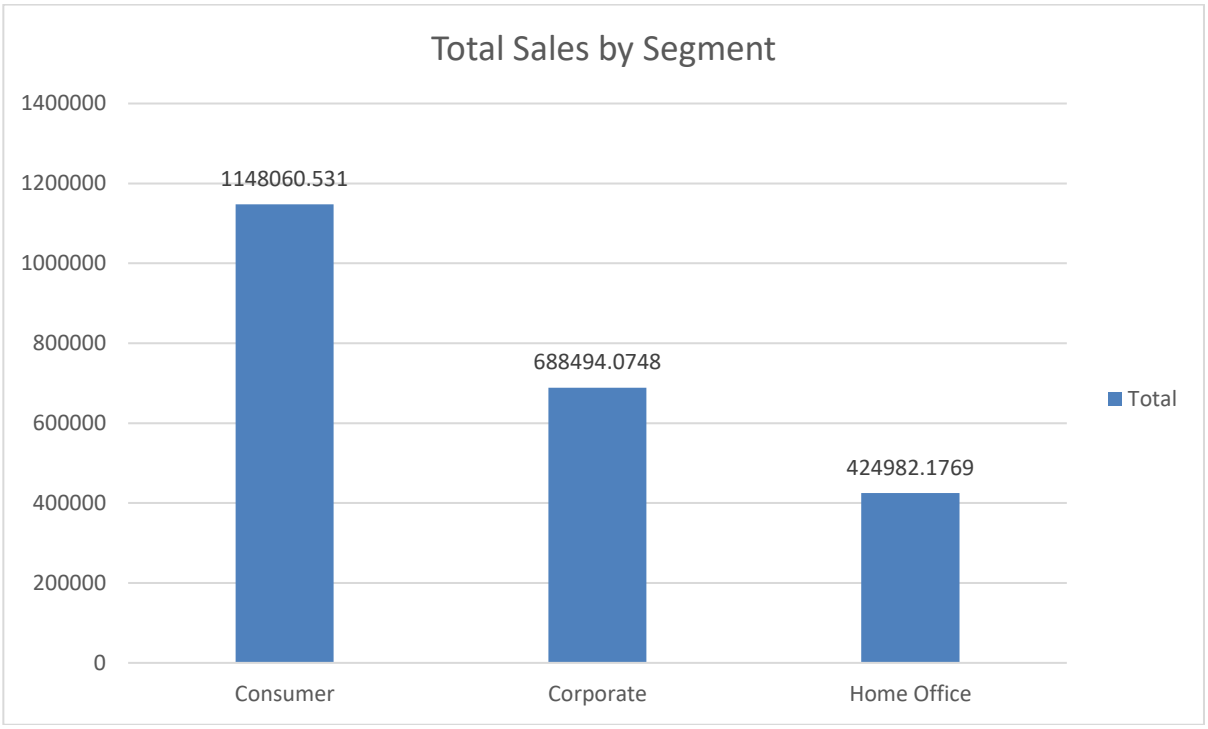
3. **Which segment has most sales in US, California, Texas, and Washington?**

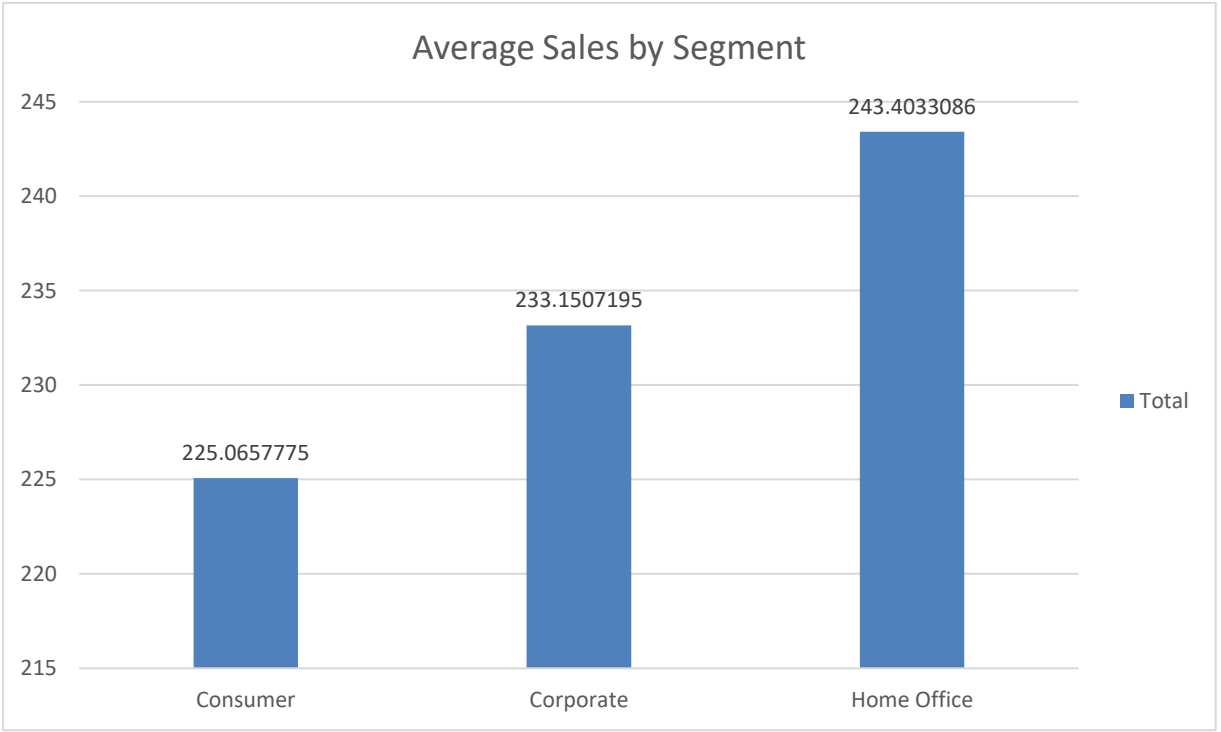


Using a bar chart to display the proportion of distribution and filtering the states for the overall sales count. The US, California, Texas, and Washington have the highest sales in the consumer

category.

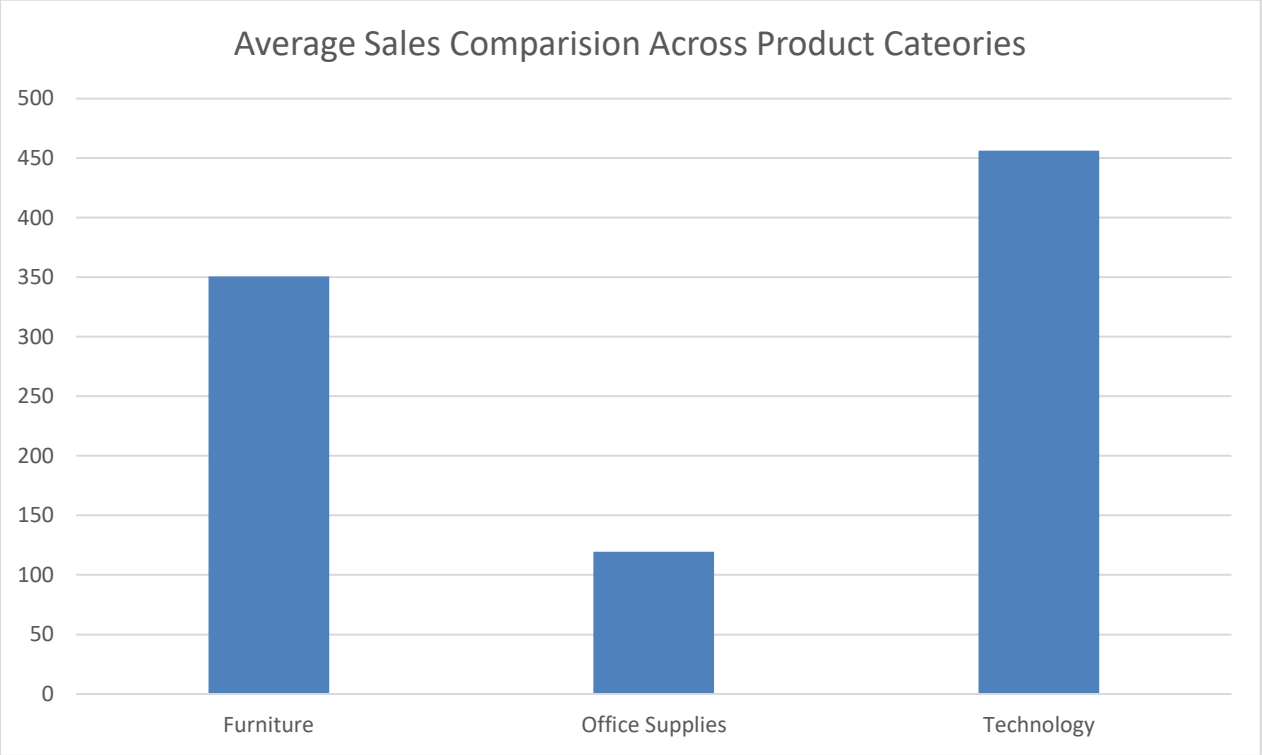
4. **Compare total and average sales for all different segments?**

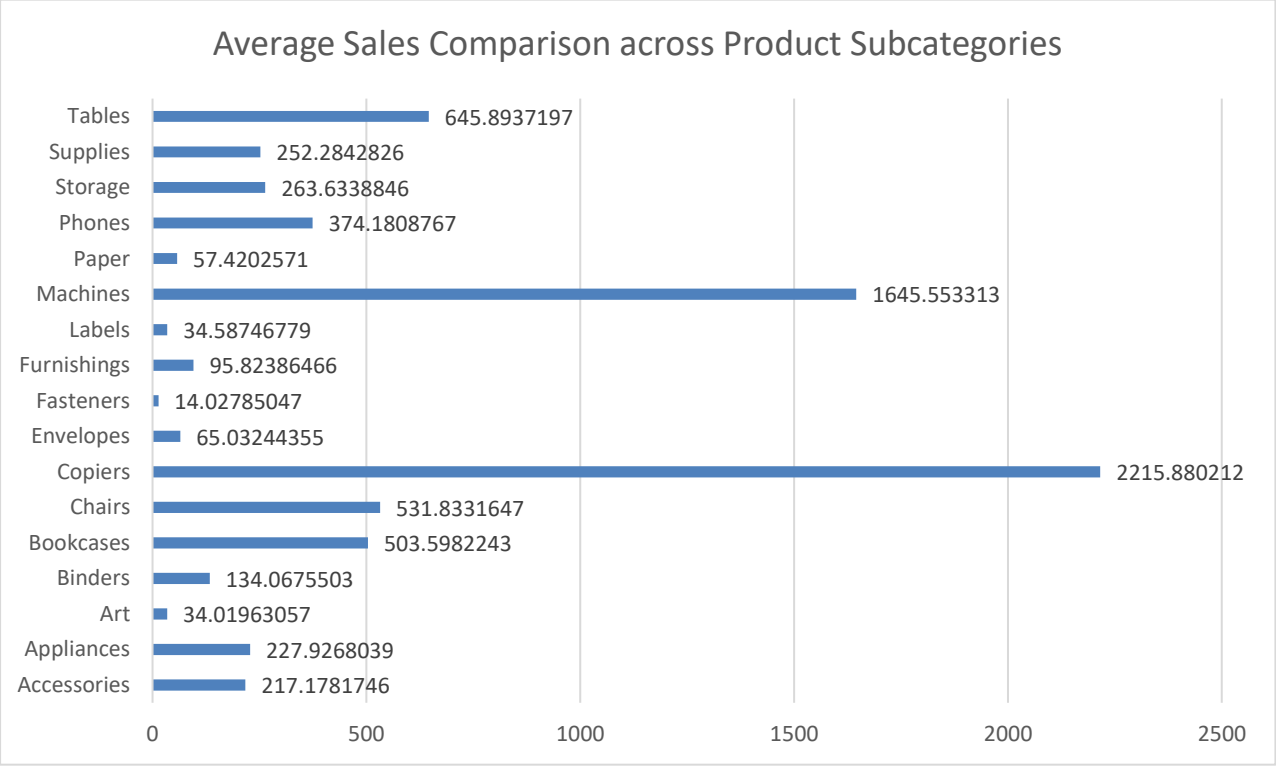




It is clearly visible that the consumer segment has higher average sales with 1148060.531 and home office segment has total sales of 243.40.

5. **Compare the average sales of different categories and subcategories in all states.**





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

Conclusion and Review

Finally, our research provided important new information about product category preferences, segment dynamics, and regional sales performance. Two notable themes that emerged were

the countrywide domination of the Consumer market and the robust demand for Technology products. Businesses can improve their plans to efficiently target particular market segments by comprehending these tendencies. Our research provides practical advice for enhancing sales tactics and spurring expansion in the retail sector. With open procedures and a lucid display, our study offers insightful information for wise choices and long-term success in a cutthroat environment.

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				

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 dataset, which is likely made up of sales data, is thoroughly summarised by the descriptive statistics that are given. The data's central tendency is indicated by the mean (average) sales value of 230.1162. With a substantially lower value of 54.384, the robust measure of central tendency—the median—indicates a skewed distribution. The value that occurs the most frequently, or the mode, is 12.96. The data's high 625.3021 standard deviation indicates that sales numbers vary significantly. 391002.7 is the sample variance, which is a dispersion measure. The distribution is characterised by a large right tail and several outliers, as indicated by the strong positive skewness (13.05363) and extreme kurtosis (307.3056). The sales values exhibit a wide range, as seen by the range of 22638.04, which is the difference between the maximum (22638.48) and minimum (0.444) values.

Cookie Data Report

Introduction

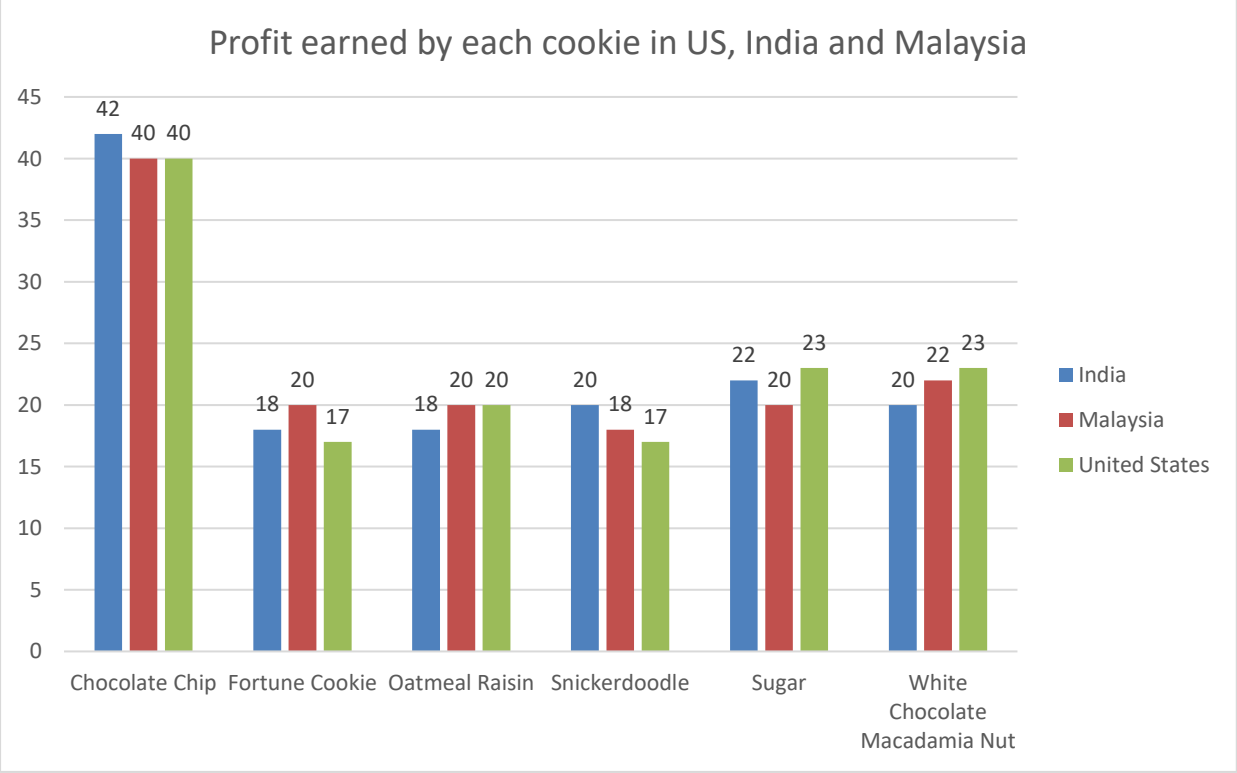
The dataset being examined includes comprehensive data on cookie sales, such as the selling country, cookie kind, units sold, revenue, cost, profit, and sale date. With an emphasis on profitability, income creation, and product trends, this report seeks to provide light on how cookie sales perform in various nations. This paper looks at important data and trends in order to provide practical advice on how to maximise profitability and improve sales tactics in the cookie sector.

Questionnaire

1. Compare the profit earn by all cookie types in US, Malaysia and India.
2. What is the average revenue generated by different types of cookies?
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?
4. Compare the performance of all the countries for the year 2019 to 2020. Which country perform in each of these years?
5. Which cookie category sold on the highest price, country wise and how much profit is earned by that category overall?

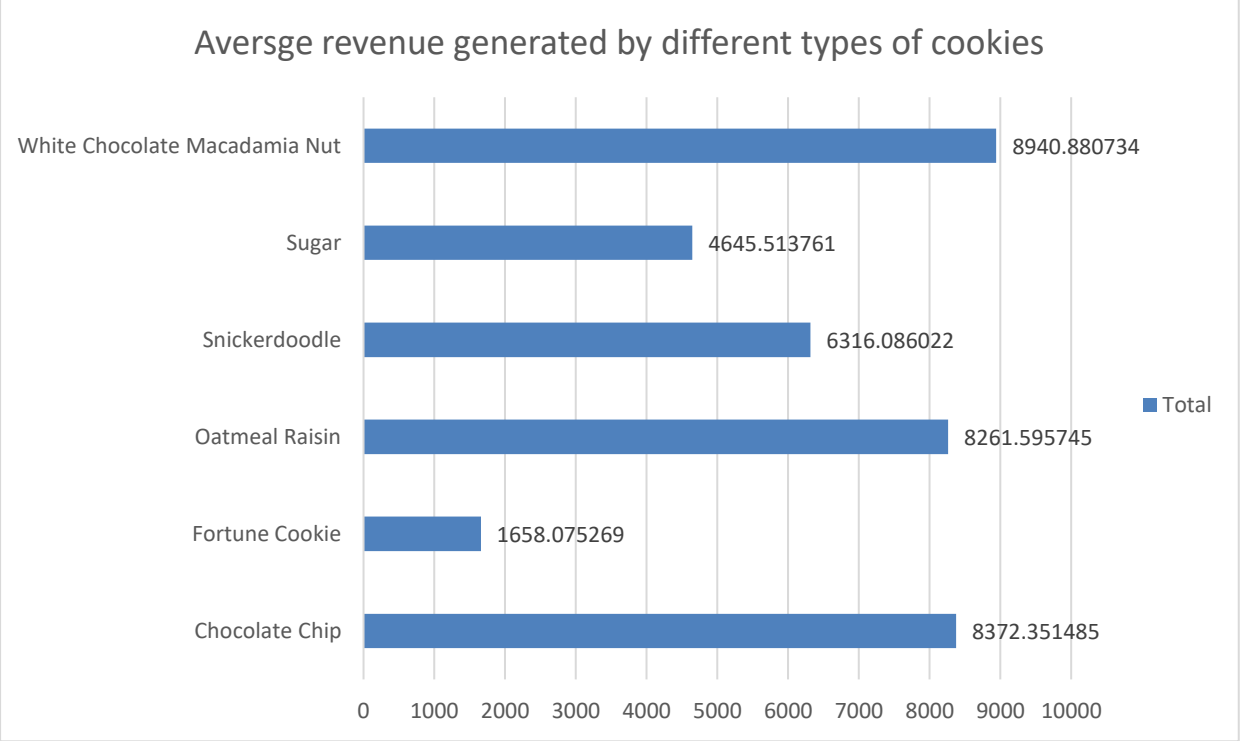
Analytics

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



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

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



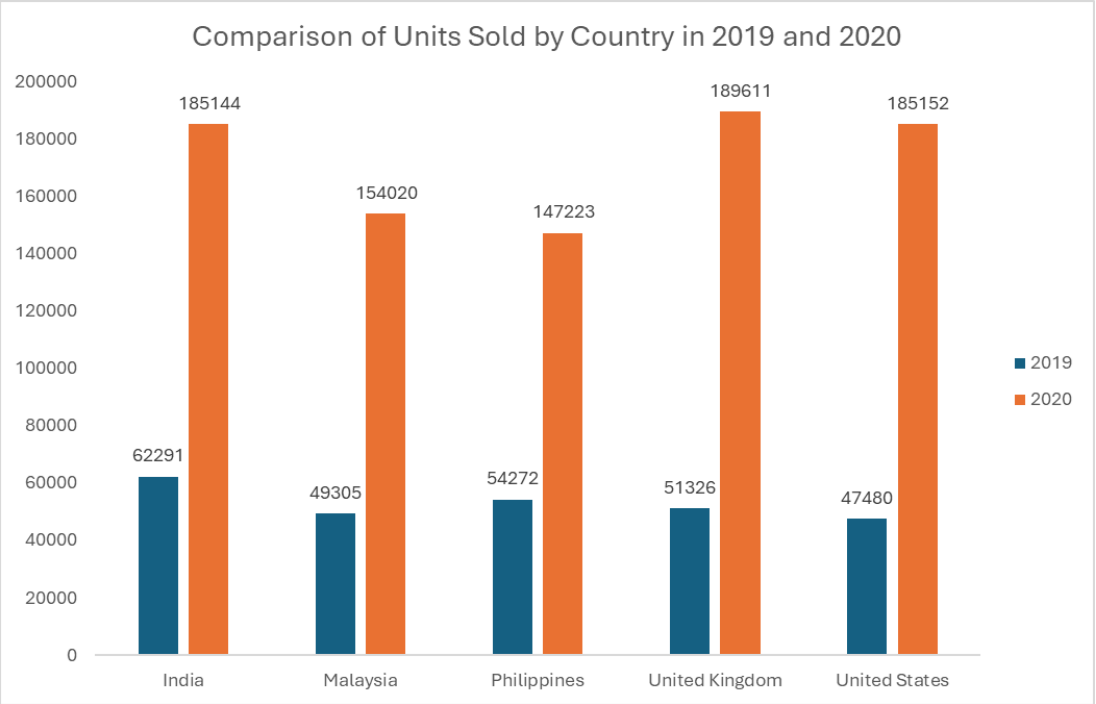
This analysis aims to provide average revenue generated and it's visible that white chocolate macadamia nut with average revenue generate is 8940.88 followed by chocolate chip.

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

The fortune and sugar cookie sales for the years 2019 and 2020 are compared in this analysis across the different countries. With 30644 sales, India leads the world in significant sugar cookie sales for the year 2020; in 2019, the UK led the world in sugar cookie sales. Once more, India tops the fortune cookie sales charts with 25,400, followed by Malaysia; the

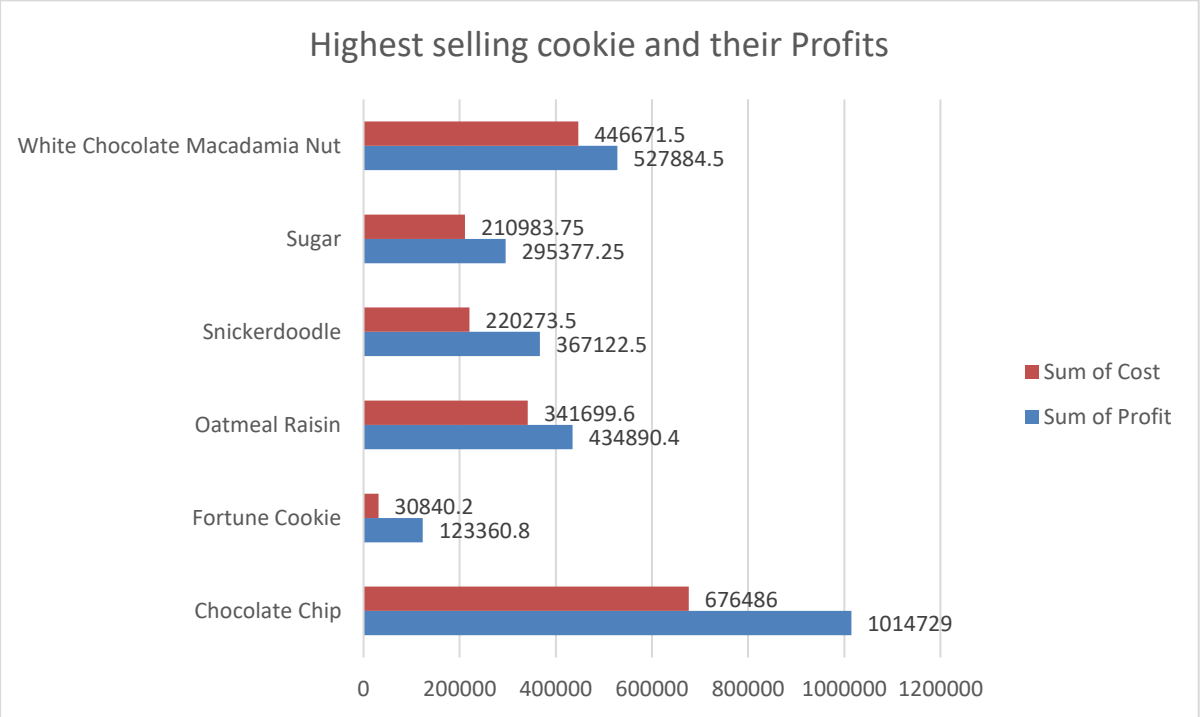
Philippines tops the charts with 8782, followed by the US.

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



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

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



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

Conclusion and Review

To sum up, the examination of cookie sales data from different nations offers insightful information on customer preferences, market trends, and profitability. In terms of profitability, chocolate chip cookies were the best-performing variety in Malaysia, India, and the US, whereas white chocolate macadamia nut cookies had the highest average price and the highest total profit worldwide. Sales of Fortune and Sugar cookies were compared, and the results showed that consumer tastes were changing, with India becoming the top selling country in 2020. Additionally, the analysis of sales performance between 2019 and 2020 revealed a notable increase in the number of units sold in Malaysia and India, suggesting a growing market demand in these regions. These results highlight how crucial it is to modify tactics in response to changing market conditions and customise product offerings to satisfy customer.

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

With a value of 0.275 ($p < 0.001$), the regression analysis shows a strong positive link between the dependent variable and the predictor variable "Profit". This implies that the dependent variable should rise by $0.275 for every unit increase in profit. The dependent variable's volatility may be explained by the regression model to the extent that the R-squared value of 0.688 indicates that roughly 68.8% of the variance is. A high multiple R and adjusted R-squared value suggests that the regression model fits the data well. The regression model appears to be a significant predictor of the dependent variable, as indicated by the highly significant ($p < 0.001$) ANOVA results.

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				

A substantial difference in mean revenues between nations is indicated by the single-factor ANOVA analysis ($F(1, 1398) = 1468.59$, $p < 0.001$). The between-groups variation, which sums up to roughly 15,699,569,566 and indicates variations in mean revenues among nations, is significant. This implies that there is a far greater variance in earnings between nations than there is within each nation. The findings suggest that the nation has a major impact on income generation, underscoring the significance of taking geography into account when examining revenue data. This result emphasises the necessity of customising business strategies to various market contexts and the possible influence of country-specific factors on revenue performance.

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 Units Sold, Revenue, Cost, and Profit variables' characteristics are shown by the descriptive statistics. The average quantity sold is roughly 1608.32, with an 867.45 standard deviation. The revenue ranges widely from $200 to $23988, demonstrating high diversity in sales amounts. The mean revenue is $6700.46. The data indicates that, on average, profits surpass costs, with a mean cost of $2752.79 and a mean profit of $3947.66. Positive skewness in sales and profit both points to a propensity for higher values. The median figures, which show revenue at $5871.5 and units sold at 1542.5, give an idea of central trend. All things considered, these statistics offer a thorough analysis of the sales data, together with distributional features, central tendency, and variability measurements that can help with decision-making.

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

Units Sold, Revenue, Cost, and Profit all show significant positive connections when analysed using correlation analysis. The correlation coefficient between Units Sold and Revenue is roughly 0.796, showing a strong positive relationship. Similarly, the correlation coefficient between Units Sold and Profit is approximately 0.829, suggesting a similarly strong positive association. Furthermore, there is a virtually perfect positive association between revenue and profit, as seen by their correlation coefficient of roughly 0.995. Cost and Revenue have strong correlations; their respective correlation values are roughly 0.992 and 0.975. These results show the interdependence of these variables in the sales process by indicating that there is a proportional increase in Revenue, Cost, and Profit as Units Sold grow.

Analysis of Loan Applicants

Introduction

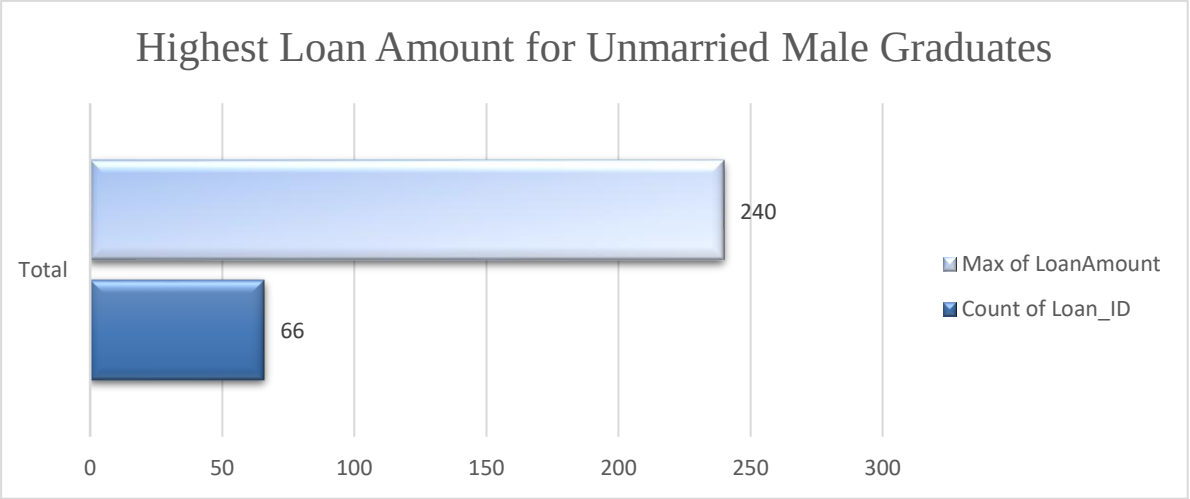
The loan dataset includes data on loan applicants, such as credit history, property area, loan amount, loan length, employment status, education, and demographics. The purpose of this analysis is to provide light on patterns in loan applications according to property area, gender, marital status, and degree of education. This paper looks at important metrics and trends in order to provide lenders and financial institutions with practical advice on how to improve loan approval procedures and successfully target various client 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.

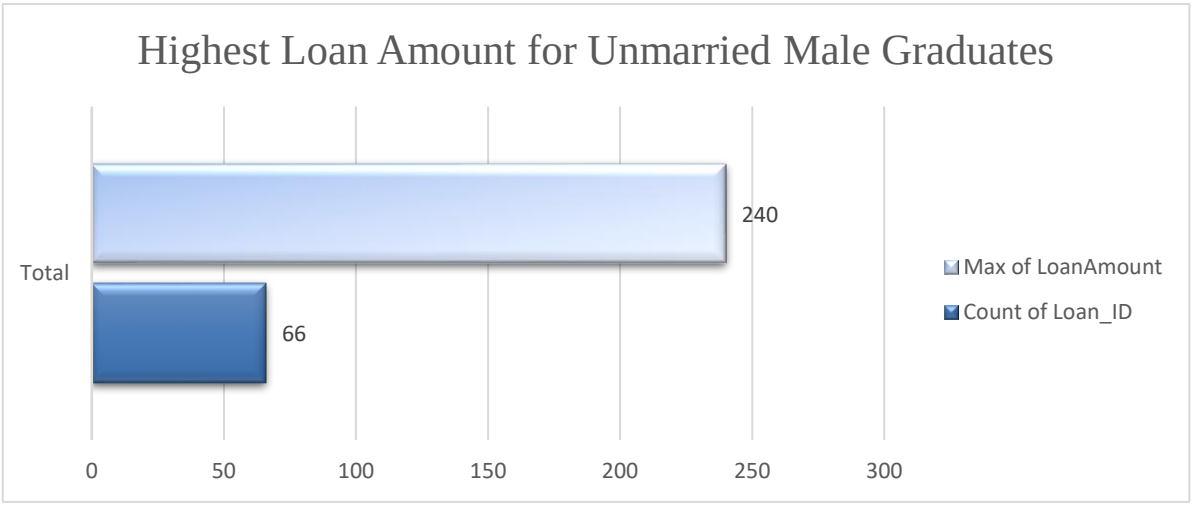
Analytics

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



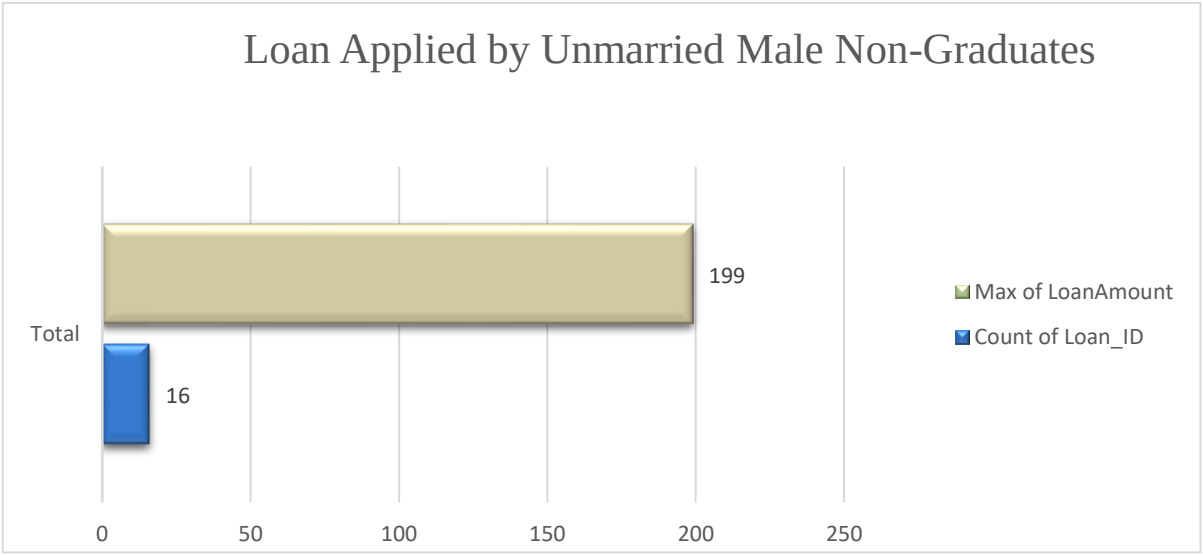
This analysis shows the no. of male graduates applied for the loan and are not married with the highest amount. As of analysed the total no. of loan applied is 66 and max loan amount is 240.

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



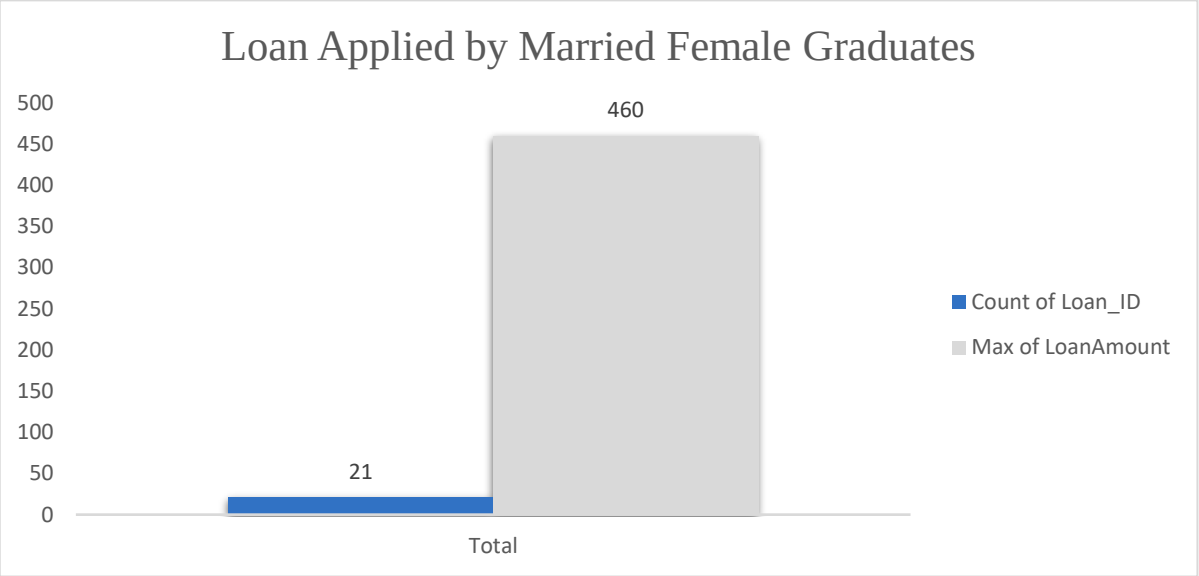
Among the female graduates who are not married and applied for a loan, the count is 35. The highest loan amount among them is $300,000.

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



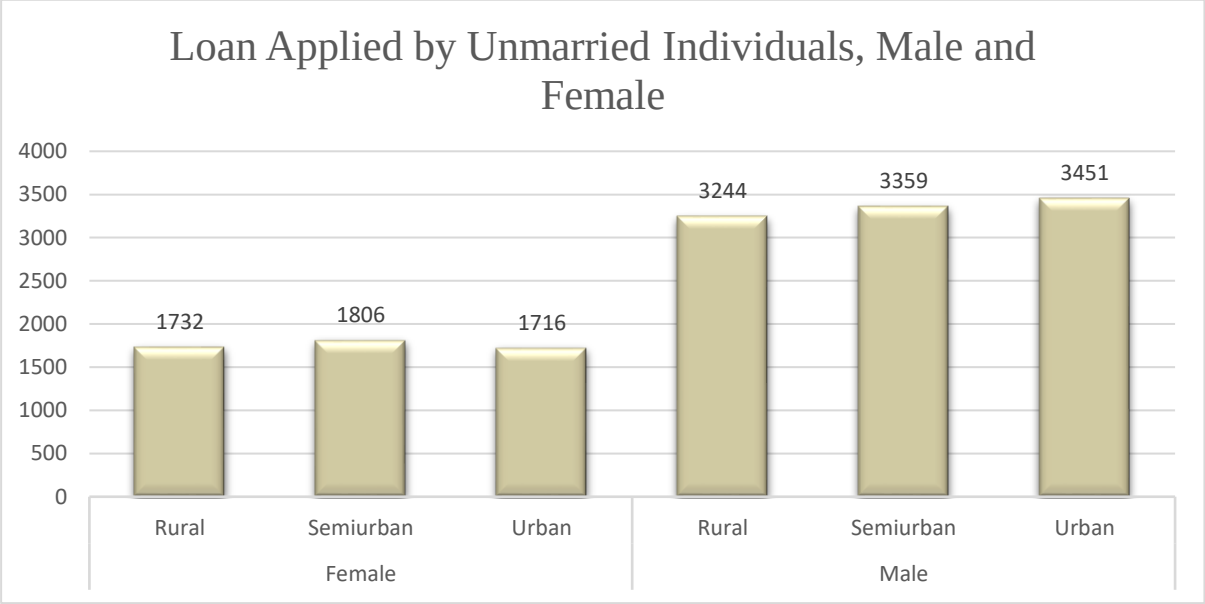
Upon analyzing the loan data, it was found that 16 male non-graduates who are not married applied for a loan. The highest loan amount among these applicants was $199,000.

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



The analysis reveals that 21 female graduates who are married applied for a loan. Among them, the highest loan amount sought was $460,000.

3. **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.**



This research compares unmarried male and female applicants for loans in rural, semi-urban, and metropolitan areas; the number of applications for loans is much larger in males than in females. Loan counts for women are as follows: women's (1732), semi urban (1806), and urban (1716); men's (3244), semi urban (3359), and urban (3451)..

Conclusion and Review

The examination of loan applicant data offers important insights into the characteristics and borrowing patterns of those in need of financial support. The analysis reveals several important findings:

1. Male and female applicants with a range of demographic profiles and financial

backgrounds demonstrate a considerable demand for loans.

2. Graduates—especially women—have a strong desire to borrow money, with different loan amounts depending on criteria like marital status.
3. Compared to other demographic groups, married female graduates typically ask for larger loan amounts, which may indicate that they have greater financial demands.
4. There is a gender gap in loan amounts, with men applicants typically requesting larger.

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.

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

With a coefficient of 0.0056 ($p < 0.001$), the regression analysis shows a moderate association between ApplicantIncome and LoanAmount. This implies that the predicted rise in the loan amount is $0.0056 for each unit increase in the applicant's income. The R-squared score of 0.199 indicates that the regression model accounts for 19.9% of the variance in loan amounts. Given the low values of both the multiple R and the modified R-squared, it is possible that loan amounts are also influenced by factors not included in the model. The regression model is a significant predictor of loan amounts, according to the very significant ($p < 0.001$) ANOVA results. Overall, the regression analysis indicates that the size of the loan is significantly predicted by the applicant's income.

Anova

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

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

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

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

Important information about the variables ApplicantIncome, CoapplicantIncome, LoanAmount, and Loan_Amount_Term is provided by the descriptive statistics. With a standard deviation of $4790.68 and a mean of $4637.35, ApplicantIncome shows a significant degree of variability. With a mean of $1528.26 and a broader range from $0 to $24000, coapplicant income likewise exhibits variability, indicating situations in which coapplicant income is absent.

While Loan_Amount_Term, with a mean of 342.67 months, indicates a somewhat constant

length for loan durations, LoanAmount, with a mean of $136.79, shows moderate fluctuation in relation to income variables. Nonetheless, there are deviations from the normal distribution shown by the skewness and kurtosis of LoanAmount and Loan_Amount_Term.

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

Complex relationships between the variables are shown by the correlation analysis. A moderately positive association has been shown between ApplicantIncome and LoanAmount, suggesting that higher income levels are generally linked to greater loan amounts. On the other hand, ApplicantIncome and Loan_Amount_Term show just a modest association, suggesting that income has minimal impact on loan duration. The data indicates a modest positive association between Coapplicant Income and LoanAmount, indicating a potential relationship between greater coapplicant earnings and larger loans. Furthermore, there is a weak positive association between LoanAmount and Loan_Amount_Term, suggesting that longer terms are typically associated with bigger loan amounts. Overall, income affects loan amount but has little effect on loan length, indicating the intricate dynamics involved in loan appraisal.

Sales Data Analysis

Introduction

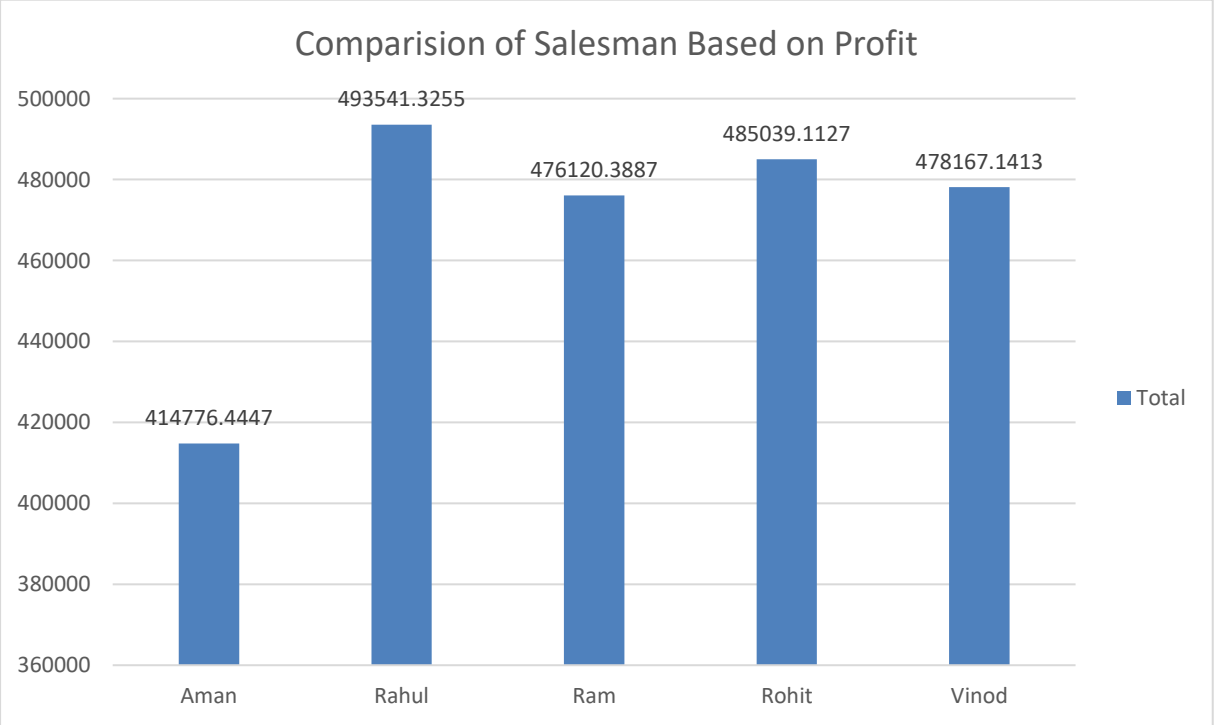
The sales dataset under investigation provides extensive details on transactions carried out over time, including salespeople, products sold, firms, quantities, and money involved. This research sets out to extract useful information from the dataset, providing light on average sales trends, profitability, product popularity, and sales performance. This research looks at important indicators and trends in an effort to find practical information that can inform strategic choices and improve future sales tactics.

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.

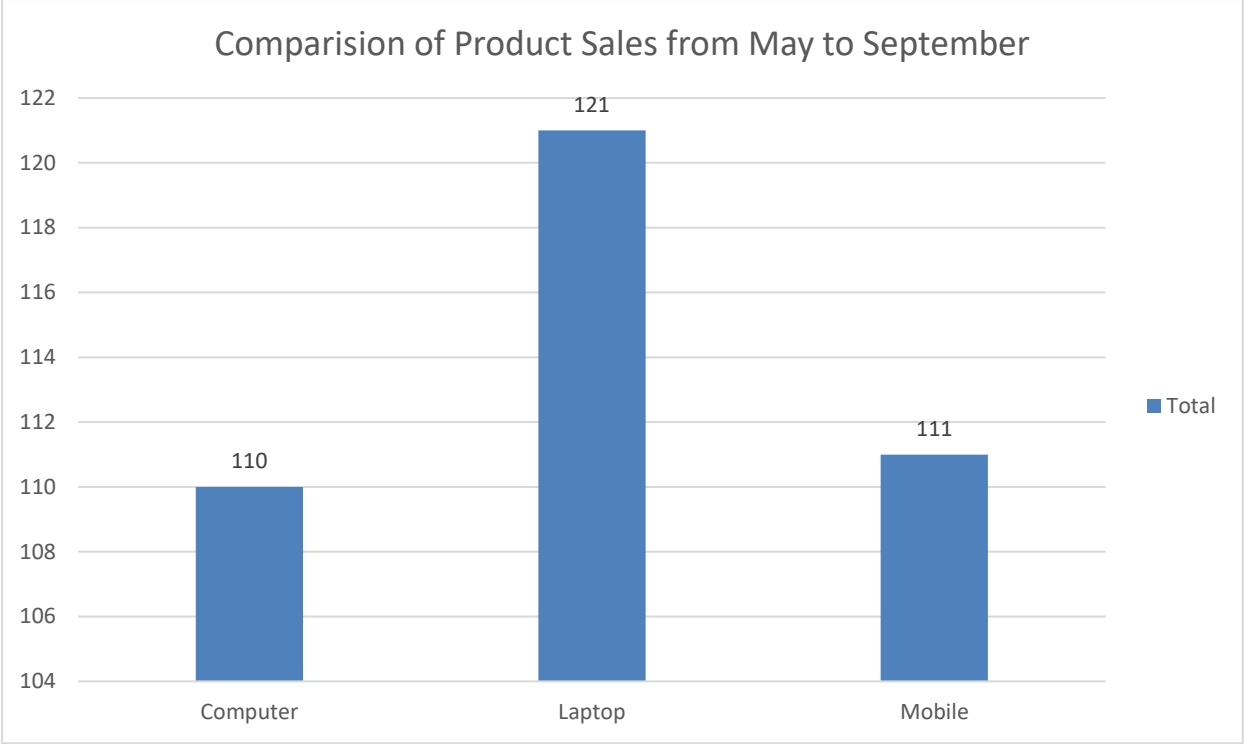
Analytics

1. Compare all the salesmen based on profit earn.



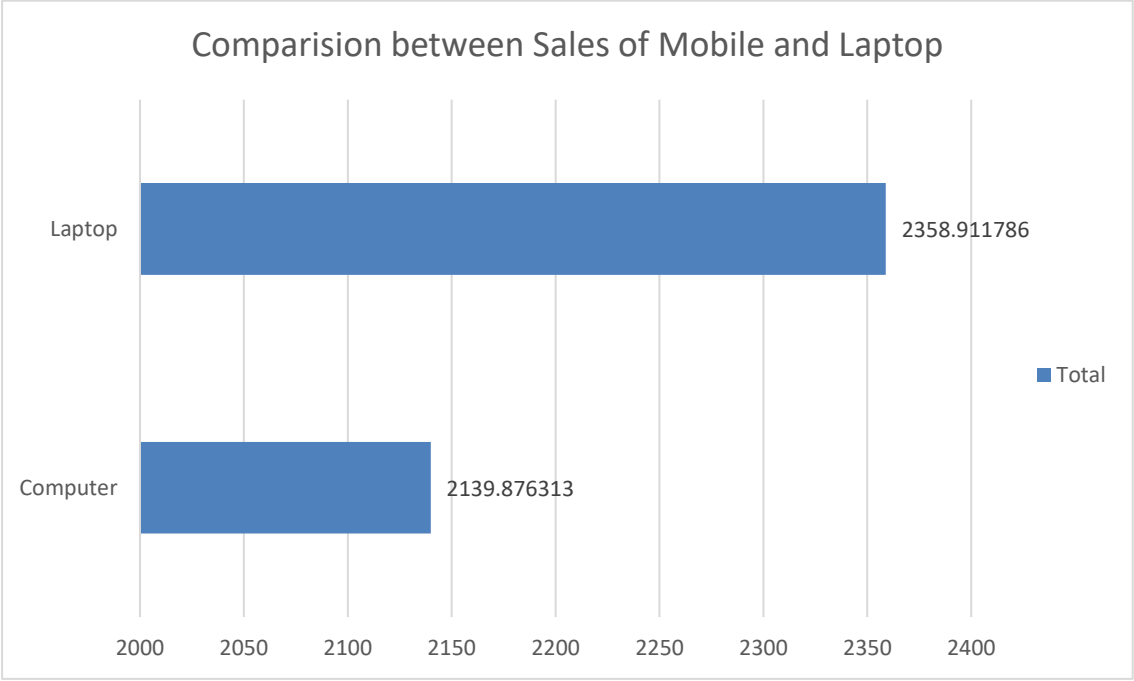
When all of the salesmen are compared based on profit made, as seen by the line chart, Rahul has the most profit earned, valued at 493541.3255.

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



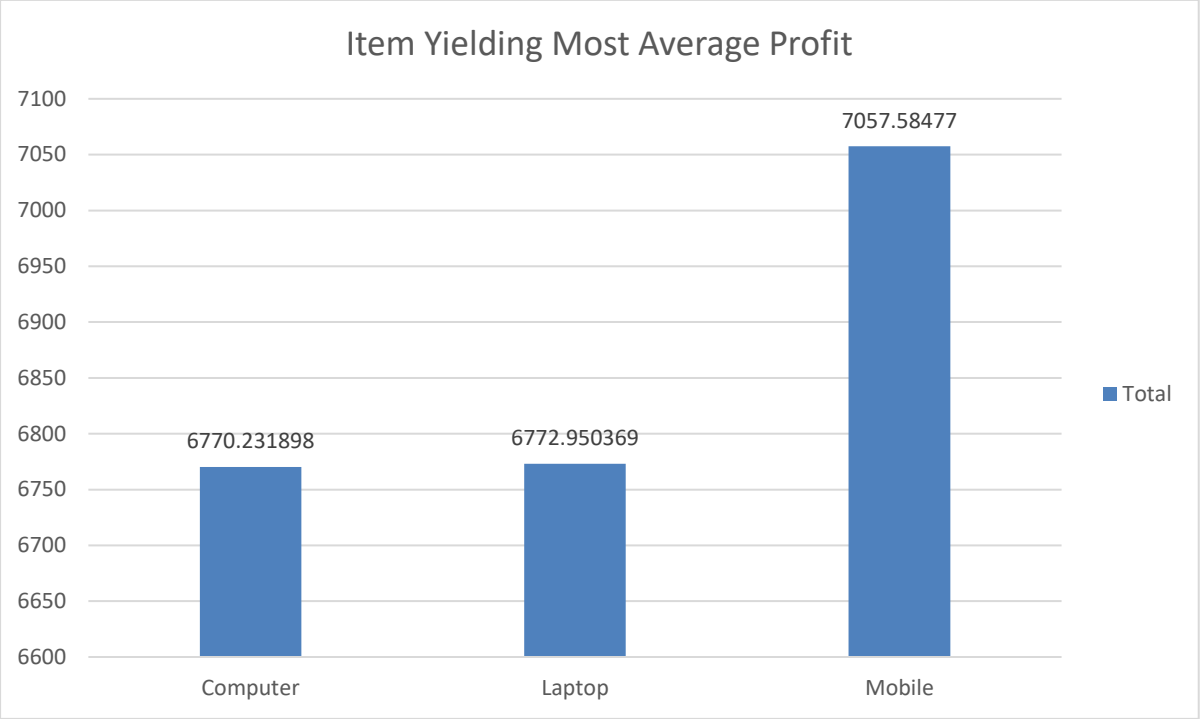
To find out which product sold the most from May through September, we would have to look at the sales data for that entire period. Compiling all transactions within this time period and adding the quantity sold for each product, the laptop is the most sold product from May to September, with September having the most sales, reaching 280.1970249.

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



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

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



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

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

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

Conclusion and Review

The analysis provides valuable insights into sales effectiveness and product trends among sales representatives. Rahul wins by outperforming all other salespeople and generating the largest profit. In addition, the laptop is the best-selling item from May to September, with the highest sales occurring in September. Over the course of the year, laptops outsell PCs in terms of units sold. Additionally, mobile phones have the highest average profit among PCs, laptops, and smartphones. Finally, laptops do better than PCs and mobile devices in terms of average sales quantity.

The study effectively highlights sales performance and product trends while providing useful data for enhancing sales strategy. Understanding enduring trends and popular products is aided by visualisations

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

With an adjusted R-squared value of 0.969, the regression analysis shows a highly significant link between the independent variable Qty and the dependent variable (not specified). This indicates that Qty may account for almost 96.9% of the variance in the dependent variable. The dependent variable is projected to rise by $11474.51 for every unit increase in Qty, according to the coefficient for Qty, which is 11474.51. T-statistics for the intercept and Qty coefficients are -31.49 and 103.72, respectively, indicating that they are both highly significant ($p < 0.001$). In general, the model exhibits robust predictive ability, indicating that Qty is a noteworthy 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 results of the ANOVA analysis show that the mean amounts of the various groups labelled "Companyy" and "Amount" differ significantly ($F(1, 682) = 3651.27$, $p < 0.001$). The variation between the groupings, which indicates the variations in average amounts between the groups, is significant, with an approximate sum of squares of 8.05 billion. This implies that there is a lot more variance in the amounts between the groups than there is within each group. The findings suggest that the group variable has a substantial impact on the numbers, which emphasises how crucial it is to take this into account when analysing the data. This result emphasises how group-specific factors may affect amounts and how important it is to take these elements into consideration when doing additional analysis and making decisions.

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

A concise summary of the distribution and central tendency of the variables Qty and Amount can be obtained from the descriptive statistics. With a standard deviation of 8.13 and a mean of roughly 19.46 units, quantity has significant variability around the mean. A skewness score near to zero (-0.10) and a slightly negative kurtosis value (-1.00), indicating a slightly flatter distribution, indicates that the data is approximately symmetrically distributed. Quantity ranges from 3 to 33.31 units. On the other hand, quantity shows significant diversity in amounts, with a mean of roughly $6864.46 and a much bigger standard deviation of $2100.19. A significantly flatter distribution is suggested by the data's slightly negative kurtosis (-0.51) and slightly negative skewness (-0.36).

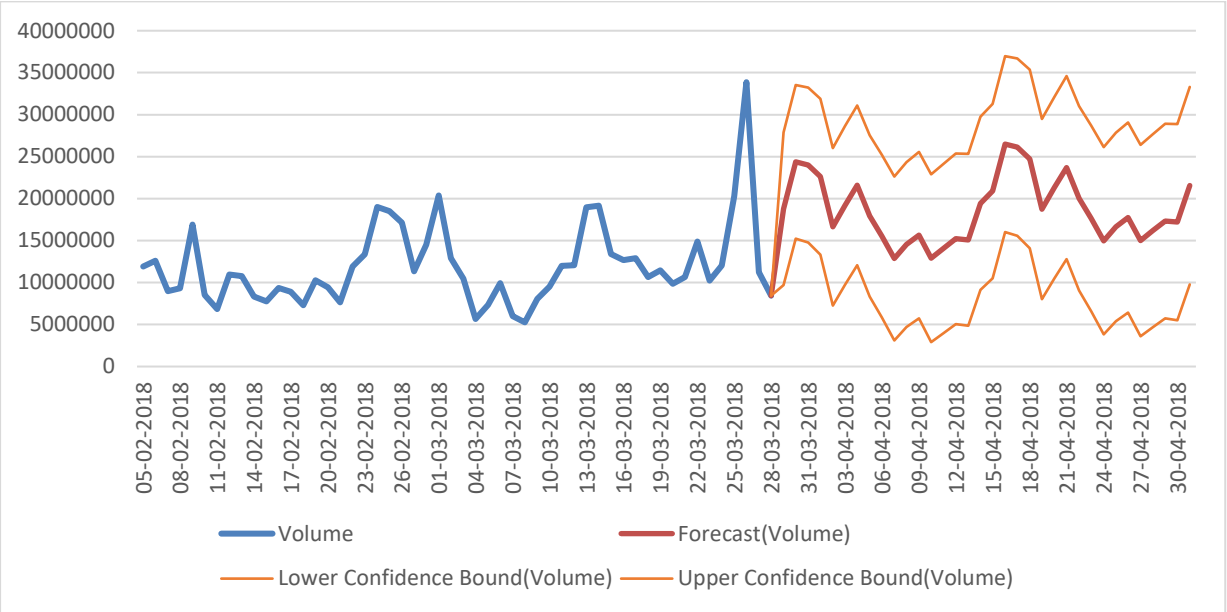
Correlation

	Qty	*Amount*
Qty	1	
Amount	0.954077	1

With a correlation coefficient of roughly 0.954, the correlation study shows a significant positive association between the variables Qty and Amount. This suggests that there is a strong linear relationship between the two variables and that the amount tends to increase proportionately to the increase in quantity. The intimate relationship between the two variables in the dataset is highlighted by this significant correlation, which indicates that changes in Qty are strongly predictive of changes in Amount.

Analysis of Forecasted Trends in Netflix's Closing Stock Prices

Date	YD Gains	Forecast	Lower Confidence Bound	Upper Confiddence bound
06-02-2018	12595800			
07-02-2018	8981500			
08-02-2018	9306700			
09-02-2018	16906900			
10-02-2018	8534900			
11-02-2018	6855200			
12-02-2018	10972000			
13-02-2018	10759700			
14-02-2018	8312400			
15-02-2018	7769000			
16-02-2018	9371100			
17-02-2018	8891500			
18-02-2018	7301800			
19-02-2018	10268600			
20-02-2018	9416500			
21-02-2018	7653500			
22-02-2018	11932100			
23-02-2018	13345300			
24-02-2018	18986100			
25-02-2018	18525800			
26-02-2018	17132200			
27-02-2018	11340100			
28-02-2018	14500200			
01-03-2018	20369200			
02-03-2018	12917200			
03-03-2018	10475100			
04-03-2018	5642900			
05-03-2018	7333700			
06-03-2018	9925200			
07-03-2018	5991900			
08-03-2018	5263900			
09-03-2018	8063300			
10-03-2018	9529900			
11-03-2018	11988300			
12-03-2018	12068600			
13-03-2018	18972900			
14-03-2018	19145500			
15-03-2018	13405800			
16-03-2018	12694900			
17-03-2018	12914000			
18-03-2018	10655200			
19-03-2018	11444800			
20-03-2018	9853600			
21-03-2018	10660500			
22-03-2018	14877400			
23-03-2018	10249400			
24-03-2018	12046600			
25-03-2018	20307900			
26-03-2018	33866500			
27-03-2018	11221100			
28-03-2018	8438800	8438800	8438800.00	8438800.00
29-03-2018		18791261.93	9711259.87	27871263.98
30-03-2018		24386659.14	15233724.31	33539593.97
31-03-2018		24012925.34	14786494.91	33239355.76
01-04-2018		22609466.77	13308982.40	31909951.14
02-04-2018		16649403.86	7274311.63	26024496.09
03-04-2018		19208256.08	9758006.47	28658505.69
04-04-2018		21583939.54	12057987.39	31109891.69
05-04-2018		17930358.38	8328162.83	27532553.93



The forecast depicted in the line graph illustrates the projected trajectory of Netflix's closing stock prices from February 5, 2018, 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 Netflix's stock prices. This predictive analysis equips stakeholders with valuable insights for strategic decision-making in financial markets