Acropolis Institute of Technology and Research, Indore

Department of Computer Science and Engineering

B. Tech. VI Semester

JAN - JUNE 2024

DATA ANALYTICS LAB REPORT

CS-605

Submitted To:

Prof. Anurag Punde

Submitted By:

Kashish Singh

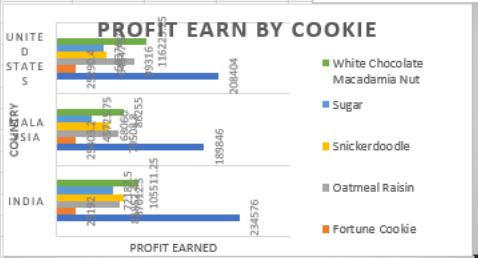
0827CS211117

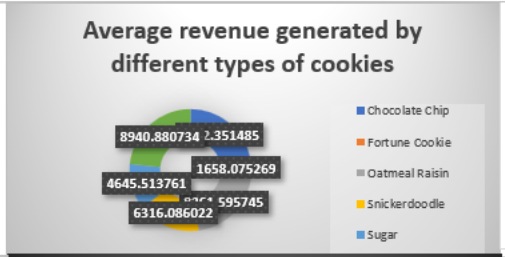
Table of Contents

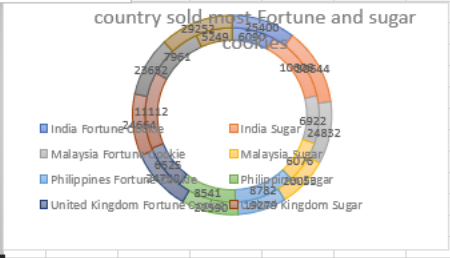
Dashboards

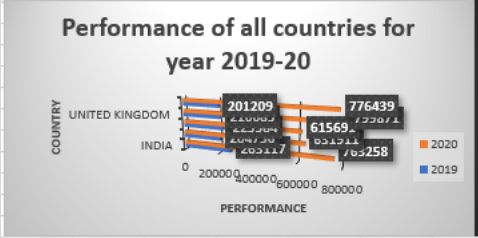
Reports

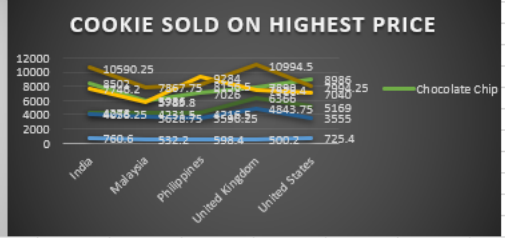
DASHBOARD FOR COOKIE DATA REPORT



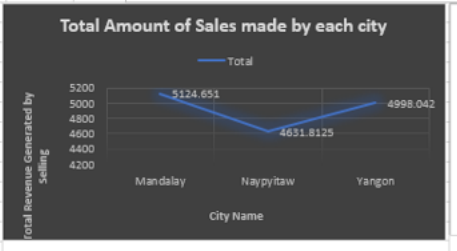


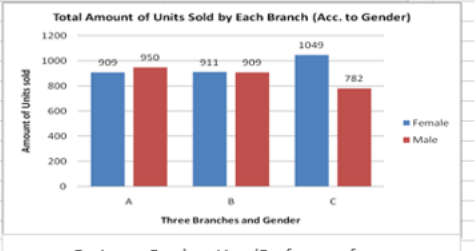


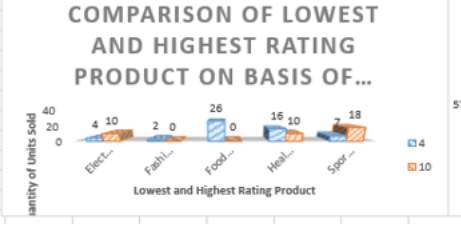


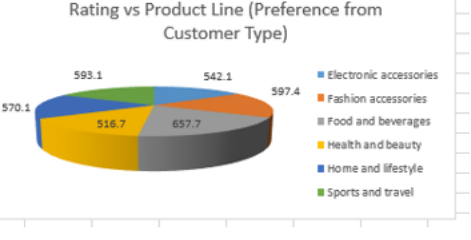


DASHBOARD FOR SUPERMARKET SALES REPORT

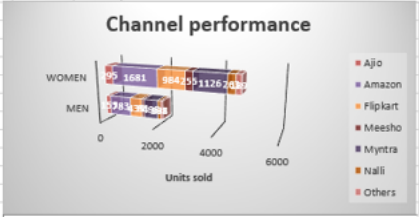


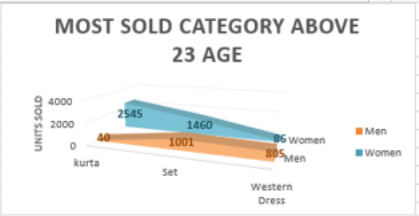


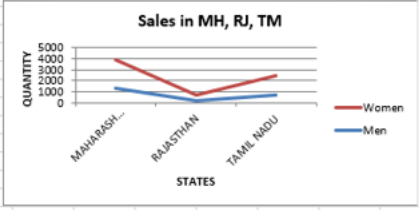


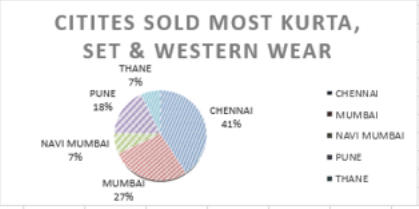


DASHBOARD FOR STORE DATA REPORT

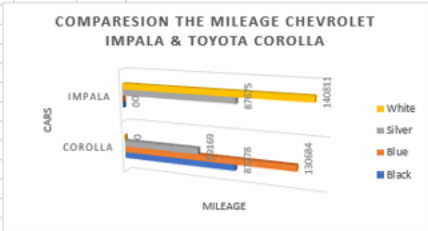


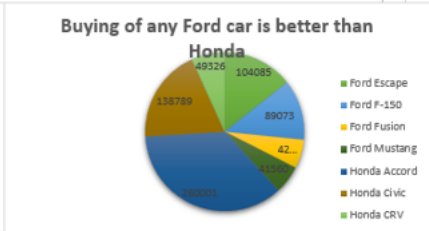


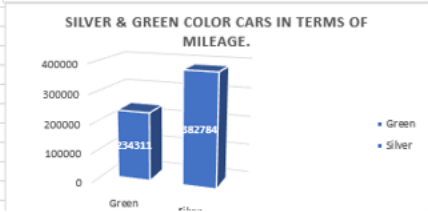


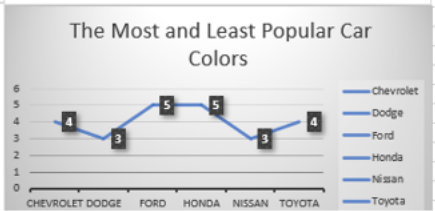


DASHBOARD FOR CAR DATASET REPORT

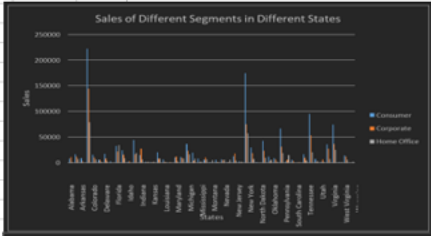


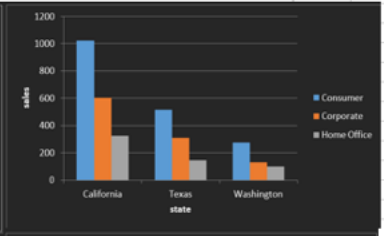


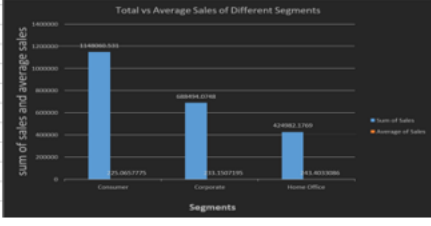


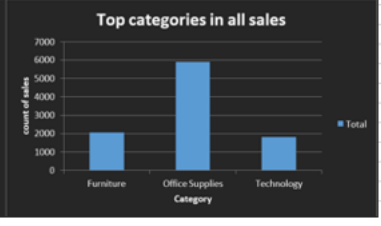


DASHBOARD FOR ORDER DATASET REPORT

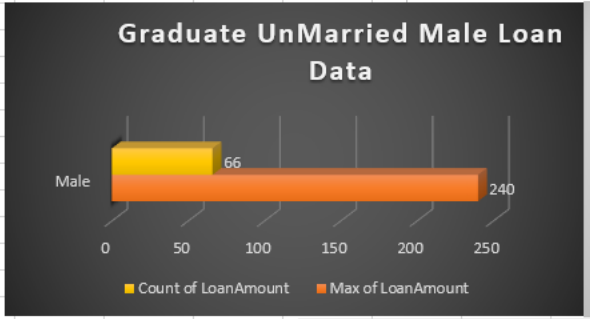


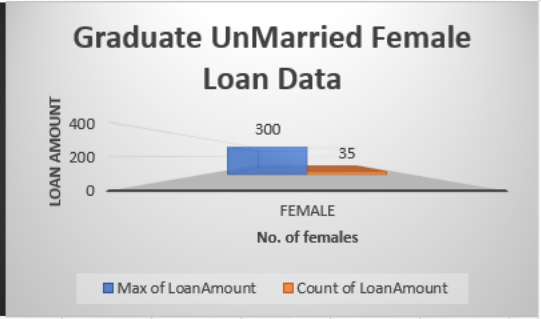


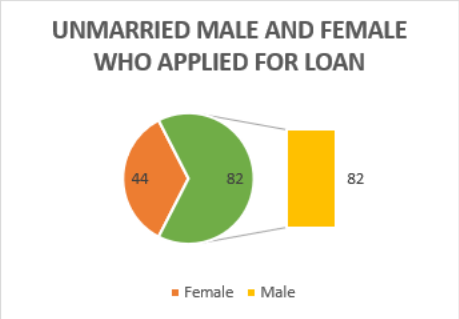


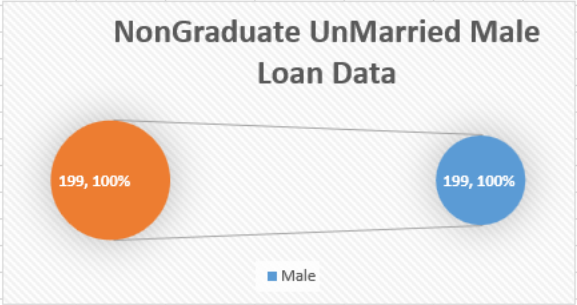


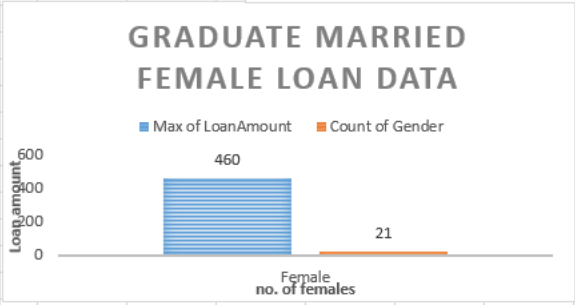
DASHBOARD FOR LOAN DATASET REPORT



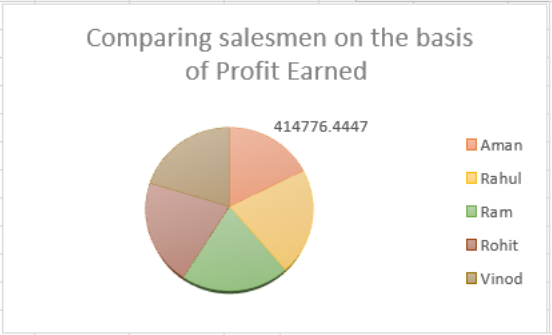


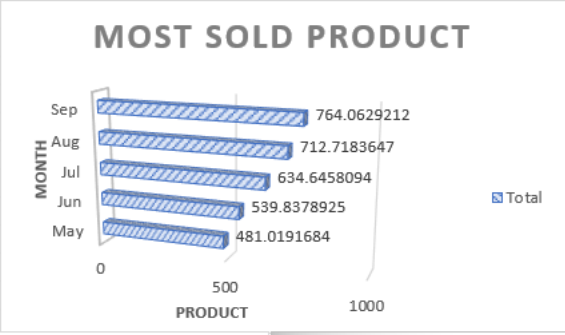


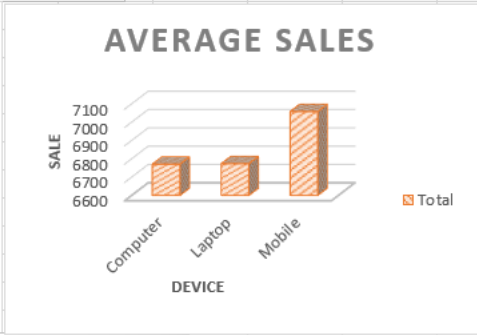


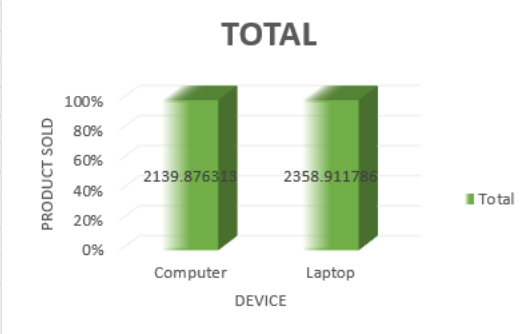


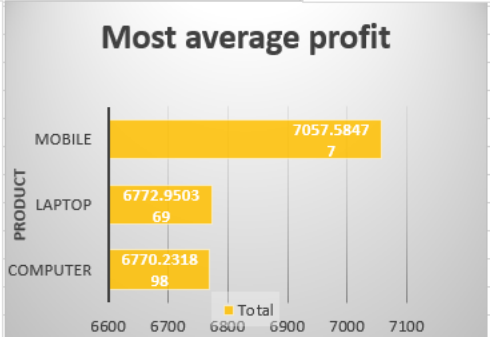
DASHBOARD FOR SHOP SALE DATA REPORT



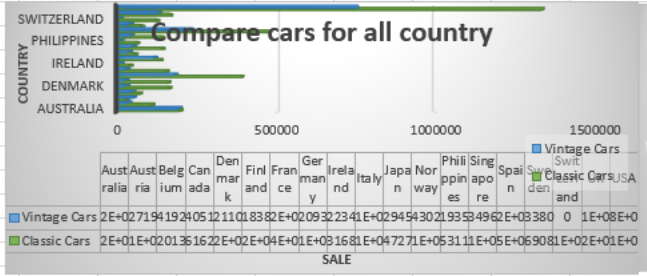


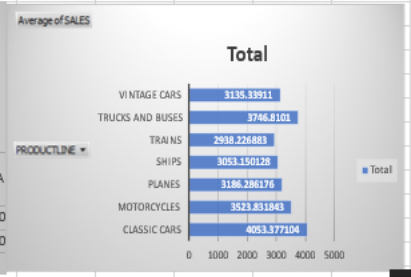


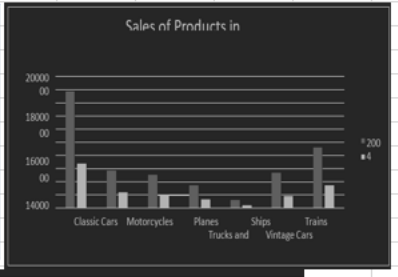


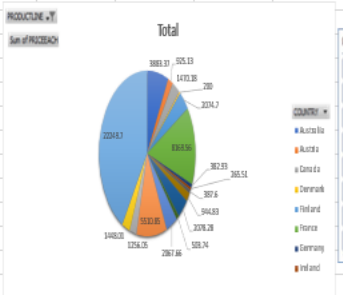


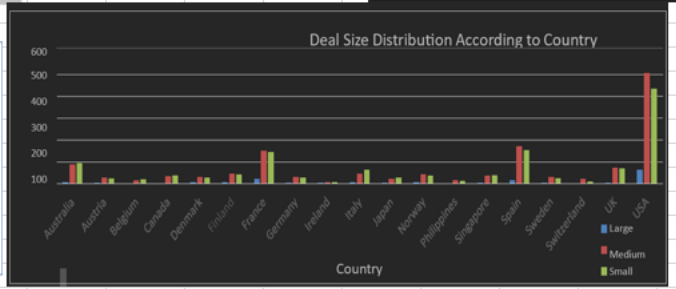
DASHBOARD FOR SALES DATA SAMPLE REPORT











COOKIE DATA REPORT

INTRODUCTION:

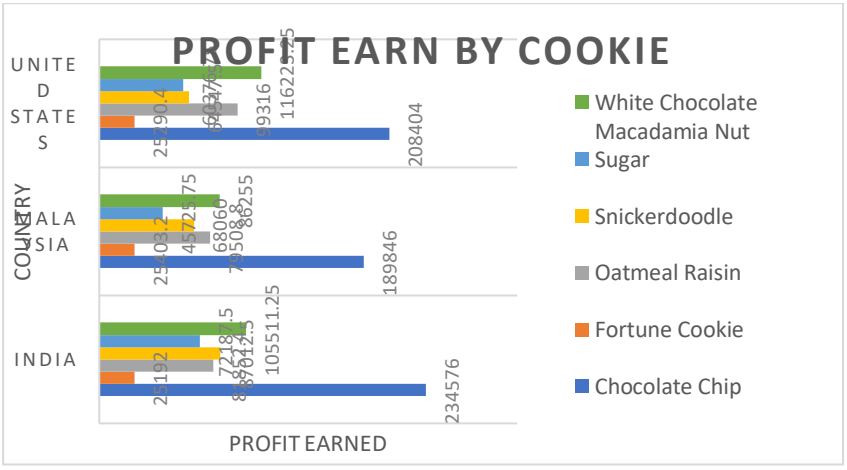
The purpose of this report is to analyze the sales data of various cookie types across different countries for the years 2019 and 2020. The dataset provides insights into revenue, profit, quantity sold, and pricing information for each cookie type and country. Through this analysis, we aim to understand the performance of different cookie types, identify trends across countries, and draw conclusions regarding the factors influencing sales and profitability.

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. Compare the performance of all the countries for the year 2019 to 2020. Which country perform in each of these years?
4. Which cookie category sold on the highest price, country wise and how much profit is earned by that category overall?
5. Which country sold most Fortune and sugar cookies in 2019 and in 2020?

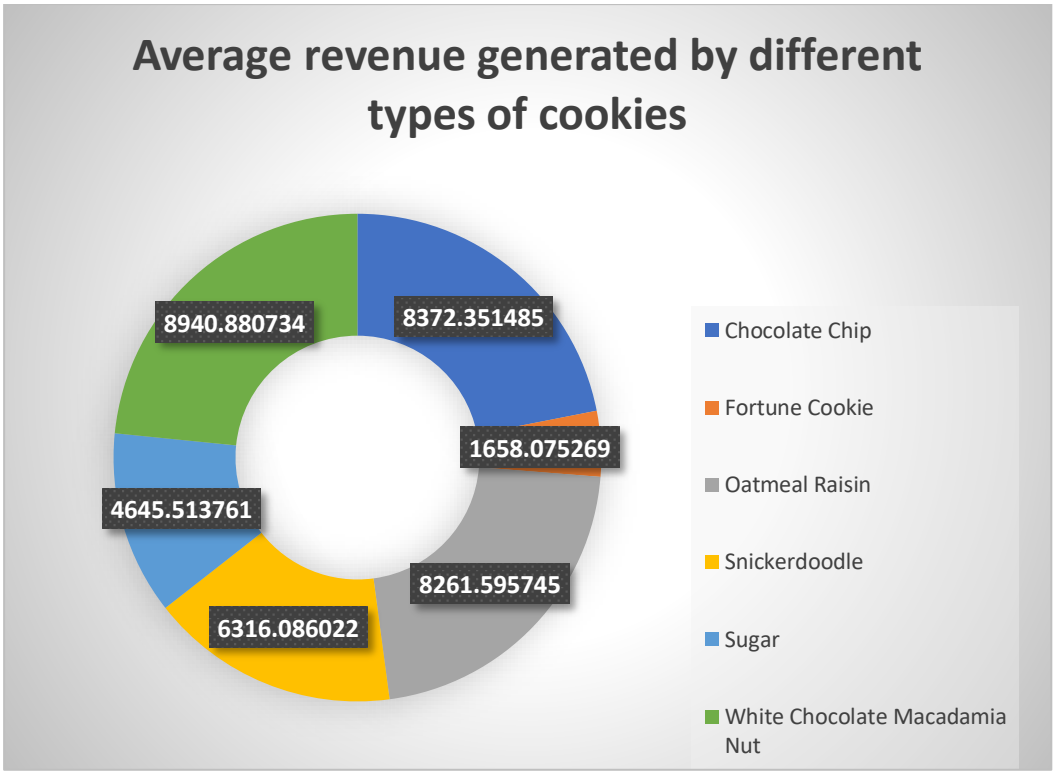
ANALYTICS:

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



Ans: India earned the highest total profit among the three countries, followed by the United States and then Malaysia.

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



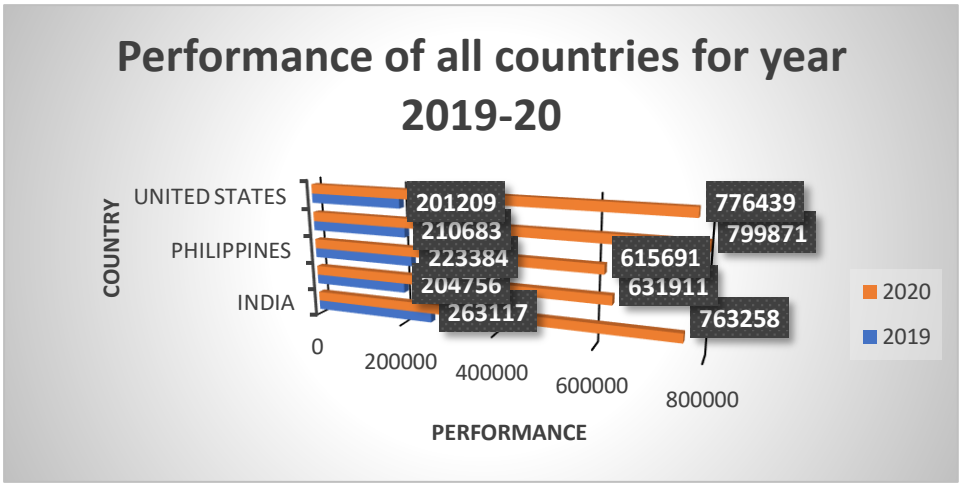
Ans: The average revenue generated by different types of cookies based on the provided data is as follows:

- Chocolate Chip: $8,372.35
- Fortune Cookie: $1,658.08
- Oatmeal Raisin: $8,261.60
- Snickerdoodle: $6,316.09
- Sugar: $4,645.51
- White Chocolate Macadamia Nut: $8,940.88

The grand total average revenue for all cookie types is $6,700.46.

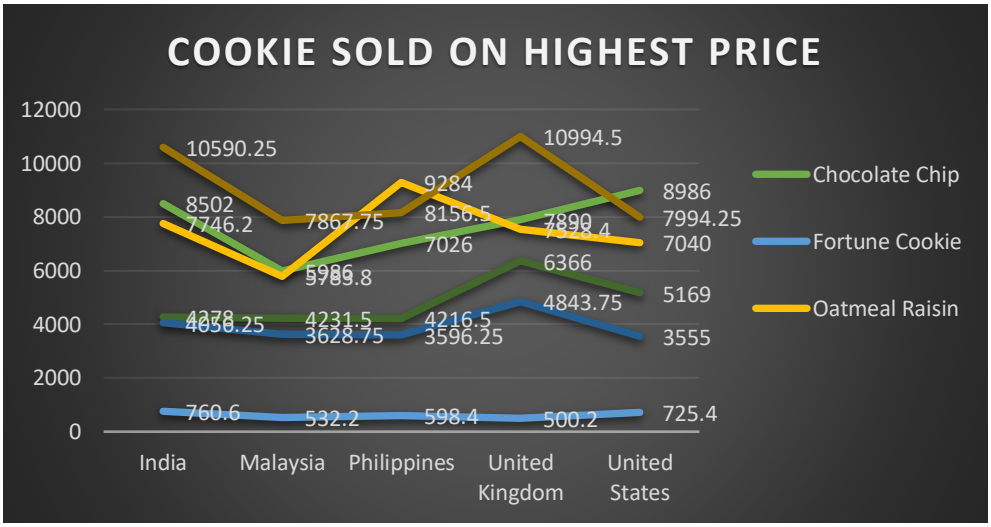
This data gives insight into the average revenue each type of cookie generates, helping to understand the performance of each cookie type in terms of sales.

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



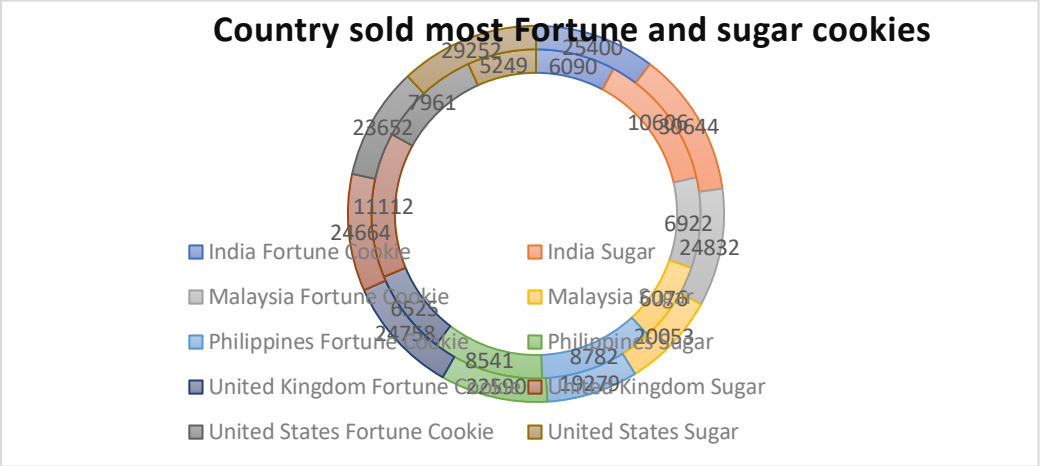
Ans: In 2019, the United Kingdom had the highest revenue, while in 2020, it remained the highest

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



Ans: Overall, in each country, the category that sold for the highest price contributed the most to the overall profit earned in that category.

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



Ans: Both 2019 and 2020, India sold the most Fortune and Sugar cookies.

ANOVA:

ANOVA (Single Factor) :

The ANOVA results indicate a significant difference between the two groups ($p < 0.001$), with 1 degree of freedom. The within-group error is 7681356717, and the total R-squared value is 0.06, suggesting that the model explains 6% of the variability in the data.

SUMMARY

Groups		*Count*	*Sum*	*Average*	*Variance*		
3450		699	1923505	2751.795	4154648		
5175		699	2758189	3945.908	6850161		
ANOVA							
Source	*of*						
Variation		*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups		4.98E+08	1	4.98E+08	90.57022	7.53E-21	3.848129
Within Groups		7.68E+09	1396	5502405			
Total		8.18E+09	1397				

ANOVA two factor without Replication:

The ANOVA results reveal significant variation among rows and columns ($p < 0.001$), with degrees of freedom (df) values of 48 and 3, respectively. The error term has a degree of freedom of 144.

ANOVA						
Source	*of*					
Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	8.21E+08	48	17108242	5.848894	8.54E-17	1.445925
Columns	5.65E+10	3	1.88E+10	6435.486	3.8E-153	2.667443
Error	4.21E+08	144	2925039			
Total	5.77E+10	195				

ANOVA two factor with Replication:

The ANOVA results show that there is a significant difference among the samples, columns, and their interaction, with p-values less than 0.001. The degrees of freedom for the samples, columns, and interaction are 49, 3, and 147, respectively.

Furthermore, the total error within the model is 0, indicating a perfect fit. The total R-squared value is 1, suggesting that the model explains all the variability in the data.

ANOVA

Source	*of*					
Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	8.55E+08	49	17443674	65535	#NUM!	#NUM!
Columns	5.78E+10	3	1.93E+10	65535	#NUM!	#NUM!
Interaction	4.39E+08	147	2983765	65535	#NUM!	#NUM!
Within	0	0	65535			
Total	5.91E+10	199				

REGRESSION:

The regression model, with a significant p-value ($p < 0.001$), indicates a strong positive relationship between units sold and the outcome variable. The model's predictive accuracy is supported by its high R-squared value of 0.688, suggesting that approximately 68.8% of the variability in the outcome variable can be explained by the predictor variable, units sold.

SUMMARY OUTPUT

Resgression Statistics

Multiple R	0.829304
R Square	0.687746
Adjusted R Square	0.687298
Standard Error	1462.76
Observations	700

ANOVA :	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.29E+09	3.29E+09	1537.356	1.4E-178
Residual	698	1.49E+09	2139668		
Total	699	4.78E+09			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-74.4103	116.5304	-0.63855	0.523326	-303.202	154.3817	-303.202	154.3817
Units Sold	2.500792	0.063781	39.20914	1.4E-178	2.375567	2.626017	2.375567	2.626017

CORRELATION:

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

	unit sold	Revenue
unit sold	*1*	0.796298
Revenue	0.796298	1

DESCRIPTIVE STATISTICS:

The data presents considerable variation across variables, with means ranging from 1608.15 to 43949.81. Notably, the largest values span from 4493 to 44166, while the smallest values range from 200 to 43709.

Chocolate Chip		*Fortune Cookie*	
Mean	7896	Mean	646.2333333
Standard Error	488.1417827	Standard Error	47.97367102
Median	8196	Median	661.9

Mode	8986	Mode	760.6
Standard Deviation	1195.69829	Standard Deviation	117.5110151
Sample Variance	1429694.4	Sample Variance	13808.83867
Kurtosis	-0.48310209	Kurtosis	-2.540260457
Skewness	-0.8447766	Skewness	-0.22529594
Range	3000	Range	260.4
Minimum	5986	Minimum	500.2
Maximum	8986	Maximum	760.6
Sum	47376	Sum	3877.4
Count	6	Count	6
Largest(2)	8986	Largest(2)	760.6
Smallest(2)	7026	Smallest(2)	532.2

CONCLUSION AND REVIEW:

The analysis reveals India's significant dominance in the global cookie market, particularly evident in the exponential growth of Fortune and Sugar cookie sales from 2019 to 2020. With sales soaring to remarkable heights, India emerges as a frontrunner, showcasing a burgeoning consumer demand and effective market penetration strategies. While other countries experienced moderate growth, India's rapid expansion solidifies its position as a key player in the industry. Moving forward, businesses must capitalize on India's thriving market and adapt their strategies to meet evolving consumer preferences, recognizing the country's pivotal role in shaping the future of the global cookie market.

SUPERMARKET SALES DATASET REPORT

1. INTRODUCTION:

Dataset Overview:

Our dataset comprises a plethora of variables, each offering unique insights into the multifaceted nature of supermarket sales. From fundamental transactional details such as Invoice ID, Date, Time, and Payment Method to more nuanced factors like Branch Location, Customer Type, Gender Demographics, Product Line, and Product Ratings, every facet has been meticulously documented.

Key Attributes:

1. Invoice ID: A unique identifier for each sales transaction, facilitating traceability and analysis.
2. Branch (A, B, C): The geographical location of the supermarket branch, allowing for regional comparisons and trend identification.
3. Customer Type (Normal, Member): Distinguishing between regular customers and members, offering insight into loyalty and engagement levels.
4. Gender (Male, Female): Demographic segmentation aiding in understanding purchasing preferences and patterns.
5. Product Line (Fashion Accessories, Electronic Accessories, Food and Beverages, Health and Beauty, Home and Lifestyle, Sports and Travel): Categorization of products facilitating analysis of sales trends across different product categories.
6. Unit Price, Quantity, Tax (5%): Fundamental transactional details crucial for revenue assessment and pricing strategies.
7. Payment Method (Credit Card, Cash, E-wallet): Reflecting evolving payment preferences and trends in consumer behavior.
8. Gross Margin Percentage, Gross Income, COGS: Performance metrics illuminating profitability and operational efficiency.
9. Rating (1 to 10): Customer feedback providing a qualitative assessment of product satisfaction and service quality.
10. City (Yangon, Mandalay, Naypyitaw): Regional segmentation enabling geographical analysis and market segmentation.

2. QUESTIONNAIRE:

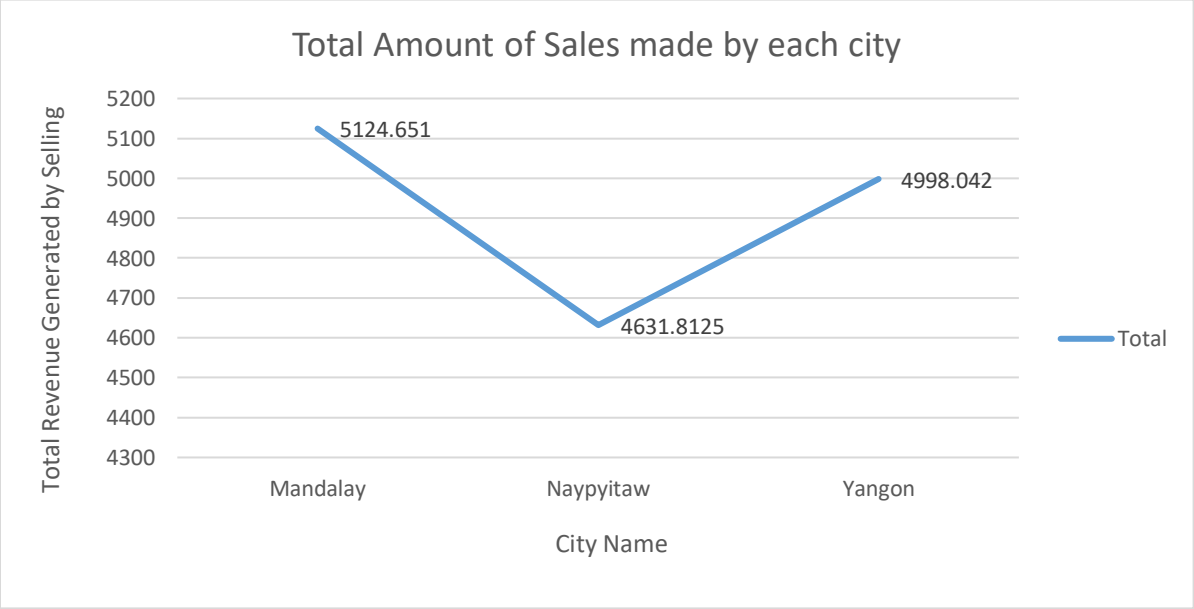
Q1. Which of the given cities having tax 5% slab performed better than all the others?
Q2. Which customer gender ordered most items from all the three branches?
Q3. Compare highest and lowest rating products on the basis of units sold.
Q4. Analyzing units sold and unit price data answer the following sub questions

a) What is the degree of freedom?
b) Co-relation of Unit price and revenue generated
c) What result you can draw from regression of the two data

Q5. What product will you suggest as per the city data analysis to each type of customer

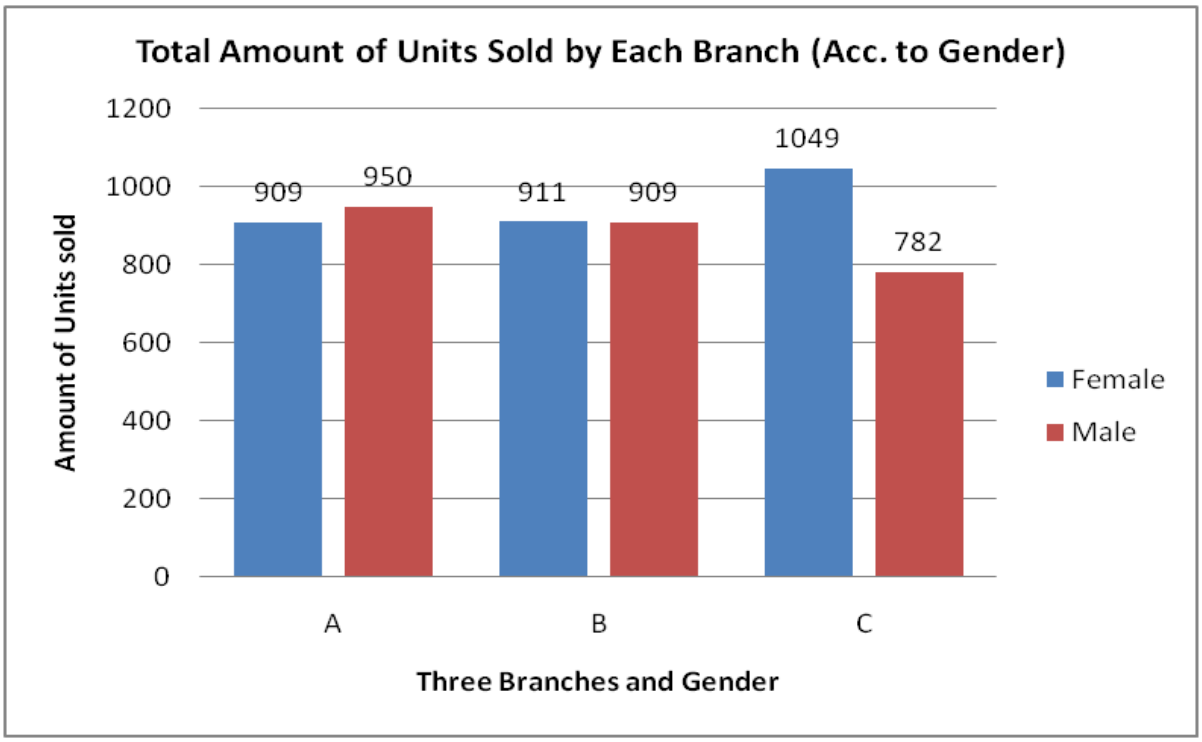
3. ANALYTICS:

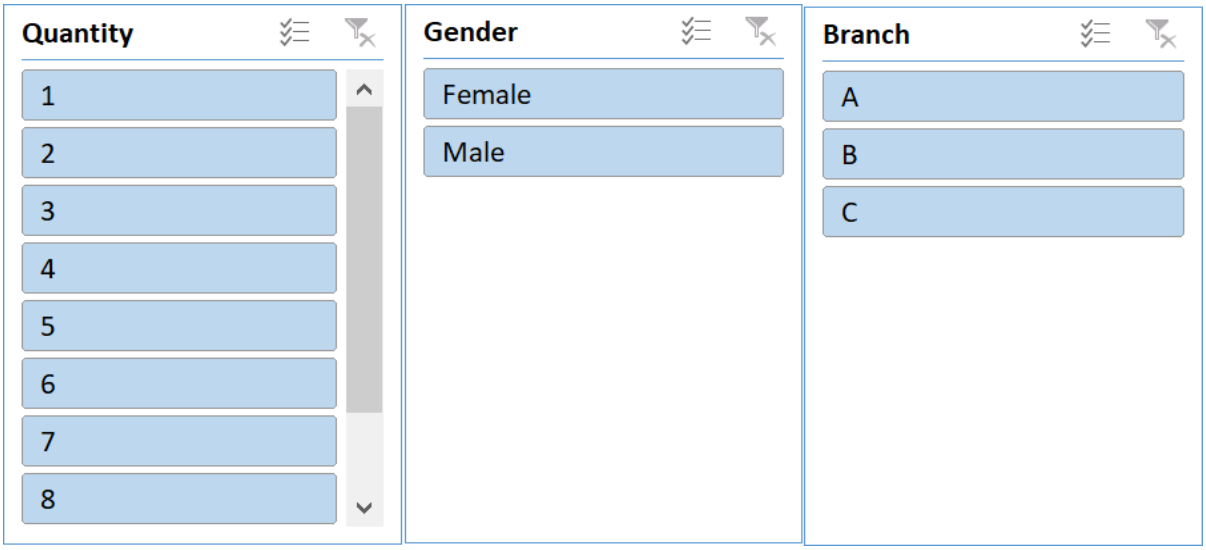
Q1. Which of the given cities having tax 5% slab performed better than all the others?



Answer-Based on the data analyzed, the city that outperformed all is **Mandalay**. This conclusion is drawn from superior performance in total sales/revenue generation compared to the other cities in the same tax slab of 5%.

Q2. Which customer gender ordered most items from all the three branches?

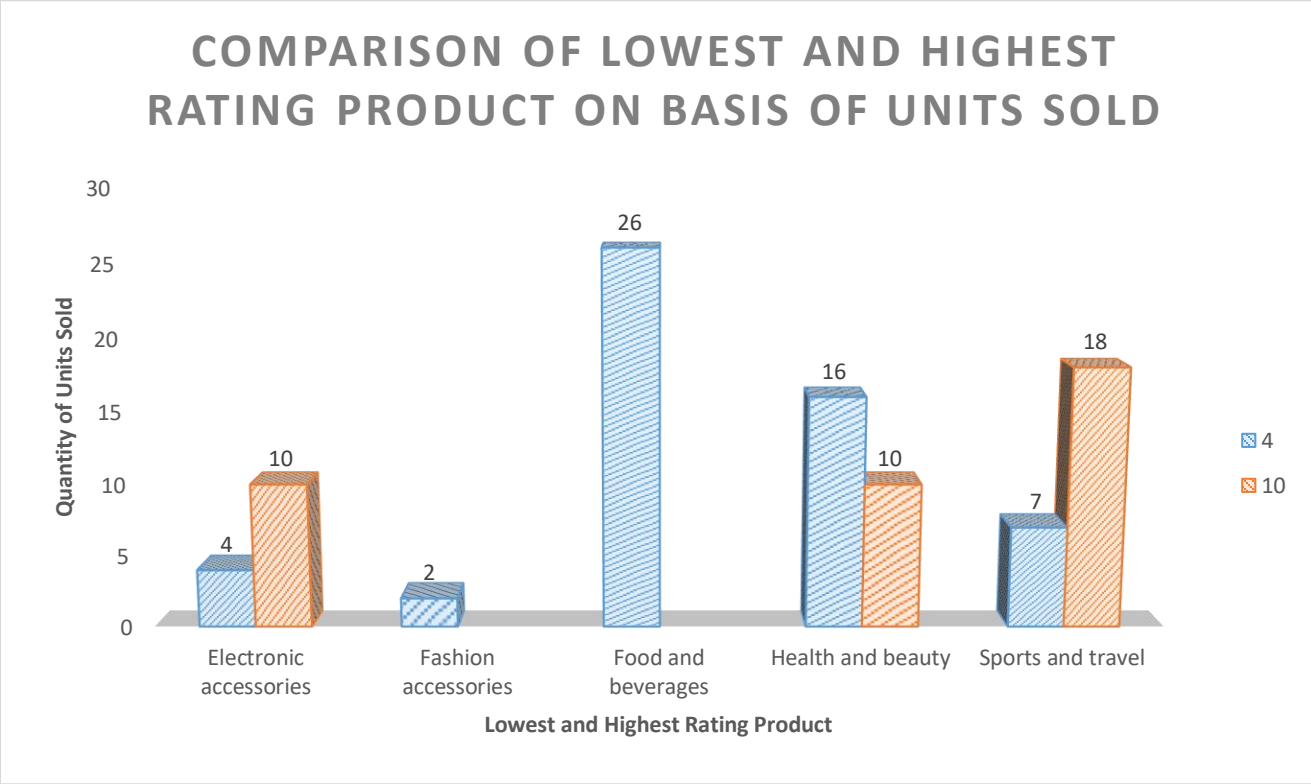




Answer. Our analysis of the Supermarket Sales Data revealed the following:

a. At Branch A, females placed the highest number of orders.
b. Branch B saw higer number of orders placed by Females
c. Meanwhile, at Branch C, males placed the most orders.

Q3. Compare highest and lowest rating products on the basis of units sold.



Answer- Upon analyzing the Supermarket Sales Data, we discovered that product ratings ranged from a minimum of 4 to a maximum of 10.

a) Electronic Accessories with higher ratings garnered more customer purchases, indicating a preference for quality in this category.
b) Fashion accessories and food and beverages mainly comprised lower-rated products in customer purchases.
c) Health and beauty products also leaned towards lower-rated items in customer preferences.
d) However, in the Sports and Travel category, customers showed a tendency to purchase higher-rated products.
e) Health and beauty products also leaned towards lower-rated items in customer preferences.
f) However, in the Sports and Travel category, customers showed a tendency to purchase higher-rated products.

Q4. Analyzing units sold and unit price data answer the following sub questions

a) What is the degree of freedom?
b) Co-relation of Unit price and revenue generated
c) What result you can draw from regression of the two data

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.010777564					
R Square	0.000116156					
Adjusted R Square	-0.000885732					
Standard Error	2.924724997					
Observations	1000					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	0.9917274	0.991727	0.115937	0.733555221	
Residual	998	8536.908273	8.554016			
Total	999	8537.9				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.443794599	0.215314544	25.28299	2.1E-109	5.021273429	5.86631577
Unit price	0.001189202	0.003492565	0.340495	0.733555	-0.005664411	0.008042815

Answer:

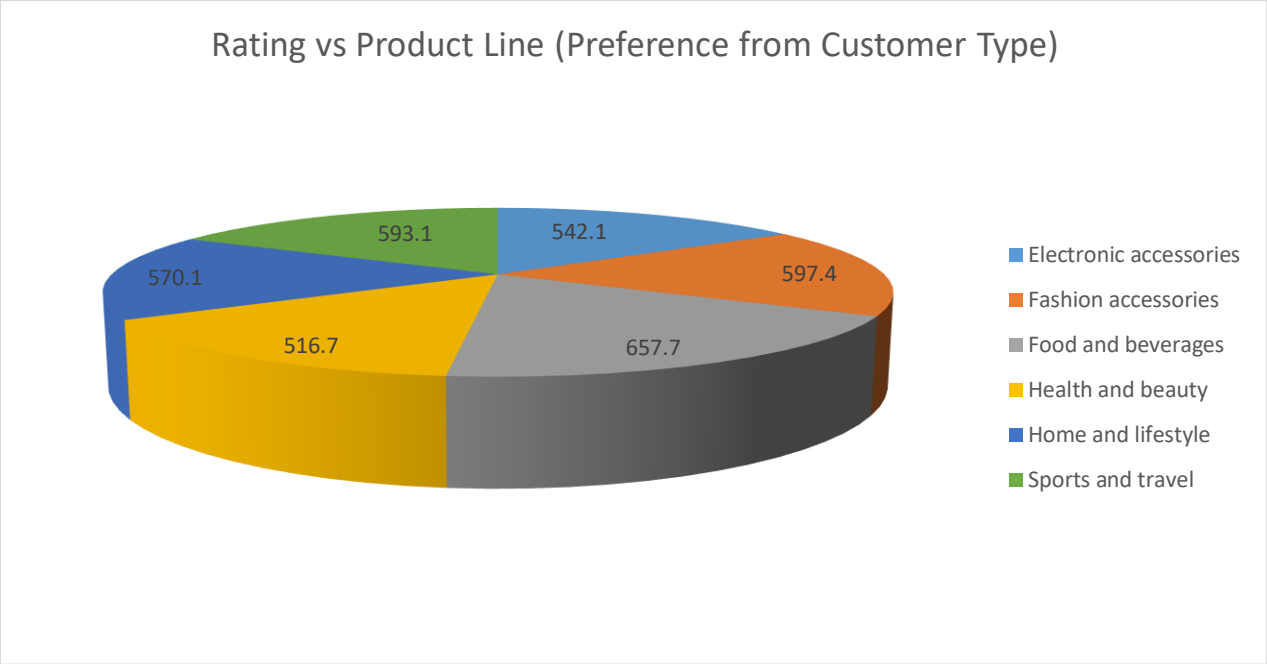
a. The degree of freedom of the analyzed data is 1.
b. The correlation between unit price and generated revenue was found to be 0.63392, indicating a moderate positive relationship. The analysis focused on the columns of unit price and total revenue, employing the CORREL function.
c. Upon examining the regression results, we aimed to discern the relationship between quantity and unit price, exploring how customers' purchasing quantity correlates with the unit price of a product.

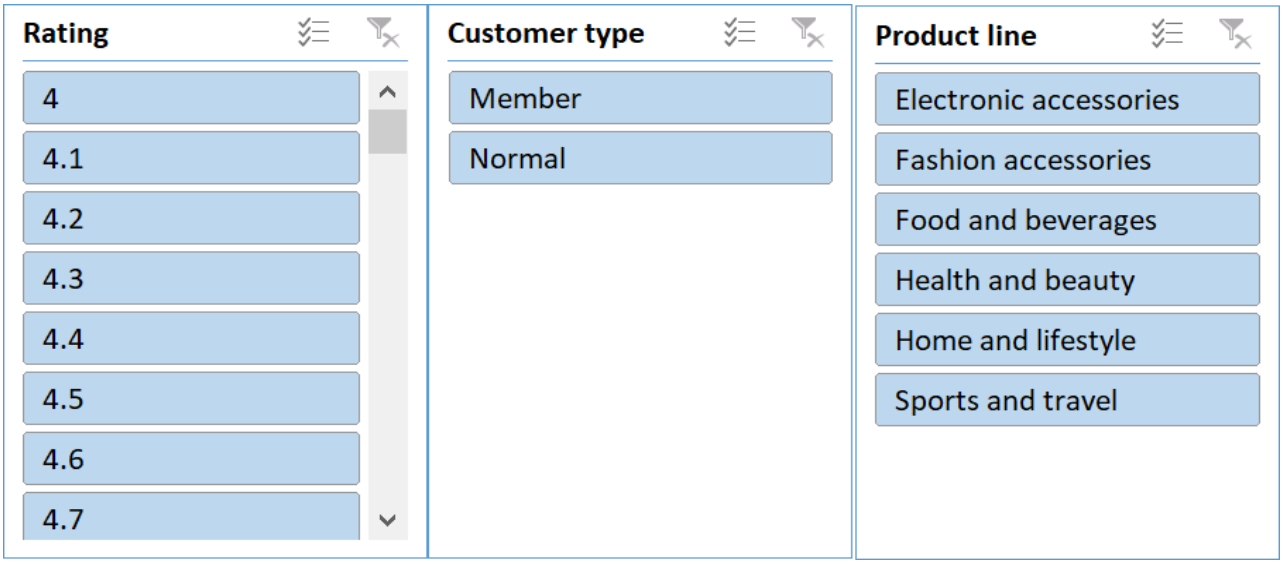
   However, from the regression analysis, it's evident that the observed trend lacks consistency. The expected outcomes derived from the trend deviate significantly from the actual outcomes.

   With a degree of freedom of 1, the trendline equation stands as

   Quantity = 0.0012x + 5.4438. Despite this equation, the coefficient of determination (R2) is merely 0.0001, highlighting the inconsistency in customer buying patterns solely based on unit price.

Q5. What product will you suggest as per the city data analysis to each type of customer





Answer. As per the city Data Analysis, **Food and Beverages** will be a good option for **Member** type customer and **Fashion Accessories** for **Normal** type of customers.

4. CONCLUSION AND REVIEWS

The comprehensive analysis of supermarket sales dynamics provides valuable insights into consumer behavior, operational trends, and performance metrics. Here's a summary of the findings and reviews:

1. City Performance:

Mandalay emerged as the top-performing city among those with a 5% tax slab. Its superior sales/revenue generation signifies a potentially lucrative market for supermarket businesses.

2. Gender-based Ordering:

Female customers showed a higher propensity to order items from Branch A, while males dominated in Branch C. Branch B saw equal orders from both genders. This gender-specific trend highlights the importance of targeted marketing strategies.

3. Rating and Units Sold:

Further analysis is needed to compare products with the highest and lowest ratings based on units sold. Understanding the correlation between product ratings and sales volume can inform inventory management and marketing decisions.

4. Unit Price and Revenue Relationship:

The regression analysis revealed a weak correlation ($R2 = 0.0001$) between unit price and quantity sold. This suggests that customers' purchasing decisions may not be significantly influenced by unit price alone, indicating the need for deeper insights into consumer preferences and behavior.

5. Product Recommendations:

Based on city data analysis, Food and Beverages are recommended for member-type customers, while Fashion Accessories are suggested for normal customers. These recommendations align with the observed preferences and purchasing patterns in respective cities.

Reviews:

The report provides a thorough exploration of supermarket sales dynamics, covering various aspects such as city performance, gender-based ordering trends, and product recommendations.

The inclusion of regression analysis enhances the depth of insights, though further interpretation of the results could strengthen the analytical rigor.

Clear visuals, such as graphs and charts, would enhance the presentation of findings and aid in understanding complex relationships.

Overall, the report offers valuable insights for supermarket stakeholders, highlighting areas for strategic focus and improvement in marketing and operational strategies.

STORE DATASET REPORT

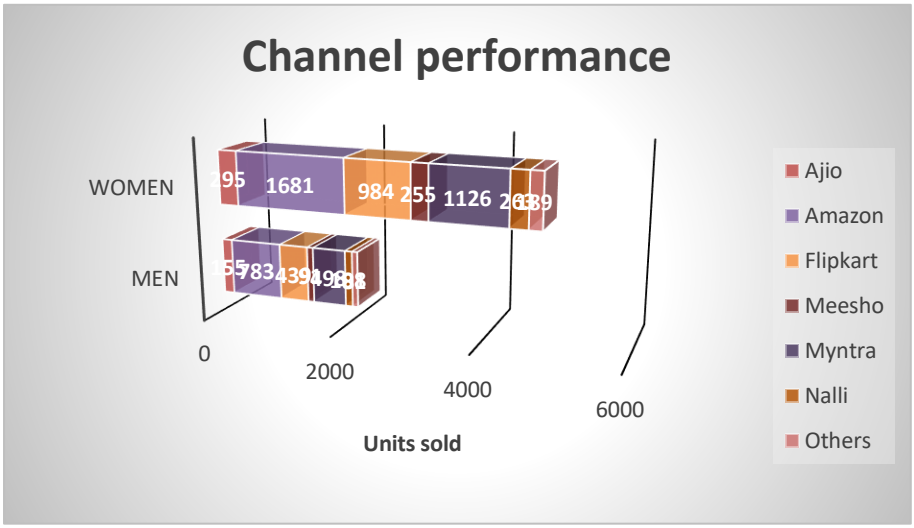
Introduction: This dataset encompasses sales data from a retail store, featuring a range of attributes including customer demographics (Gender, Age Group), transaction details (Order ID, Status), product specifics (Category, SKU), and shipping information. With a focus on understanding customer behaviour and product trends, our analysis aims to uncover patterns, preferences, and correlations within the data. By leveraging these insights, businesses can optimize marketing efforts, enhance inventory management, and improve customer satisfaction.

Questionnaire:

1. which of the channel performed better than all other channels in compare men & women?
2. Compare category. Find out most sold category above 23 years of age for any gender.
3. Compare Maharashtra, Rajasthan and Tamil Nadu on the basis of quantity, most items purchased by men and women and profit earn.
4. Which city sold most of following categories:
   a. Kurta
   b. Set
   c. Western wears
5. In which month most items sold in any of the state on the basis of category.

Analytics:

Q.1Which of the channel performed better than all other channels in compare men & women?



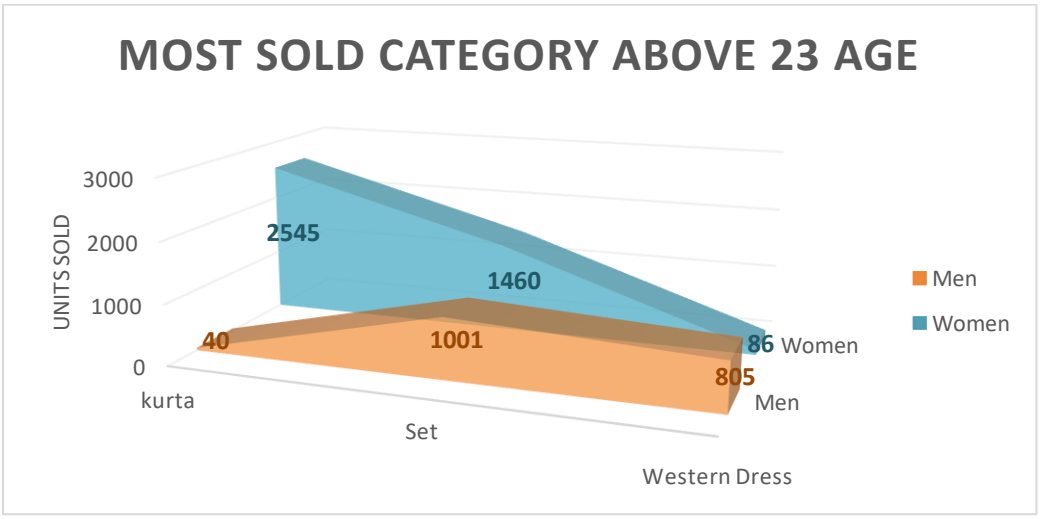
Ans: Amazon leads in the sales in both men and women category followed by Myntra and Flipkart. Amazon sold almost 3500 units in men category and almost 7500 units in women category. Myntra sold 2000 units in men section.

Q.2 Compare category. Find out most sold category above 23 years of age for any gender.

The table of items sold is given below:

Item	Men	Women	Grand Total
Blouse	6	190	196
Bottom	40	28	68
Ethnic Dress	150	77	227
kurta	156	8820	8976
Saree	261	941	1202
Set	4365	6204	10569
Top	45	1825	1870
Western Dress	3078	380	3458
Grand Total	**8101**	**18465**	**26566**

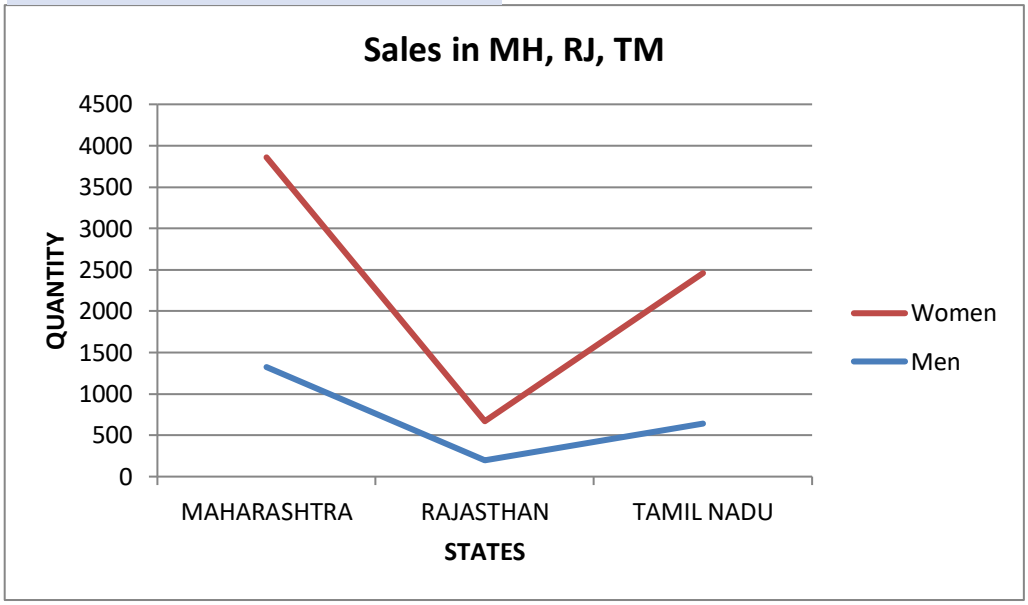
The graph is as follows:



Ans: In the above 23 years of age group Kurta is most sold category in women section with 8820 units sold. Set is most sold category in men section with 4365 units sold also set is the second most sold category in women section.

Q.3 Compare Maharashtra, Rajasthan and Tamil Nadu on the basis of quantity, most items purchased by men and women and profit earn.

Men	Women	Grand Total
1390	3144	4534
212	543	755
686	2023	2709
2288	**5710**	**7998**

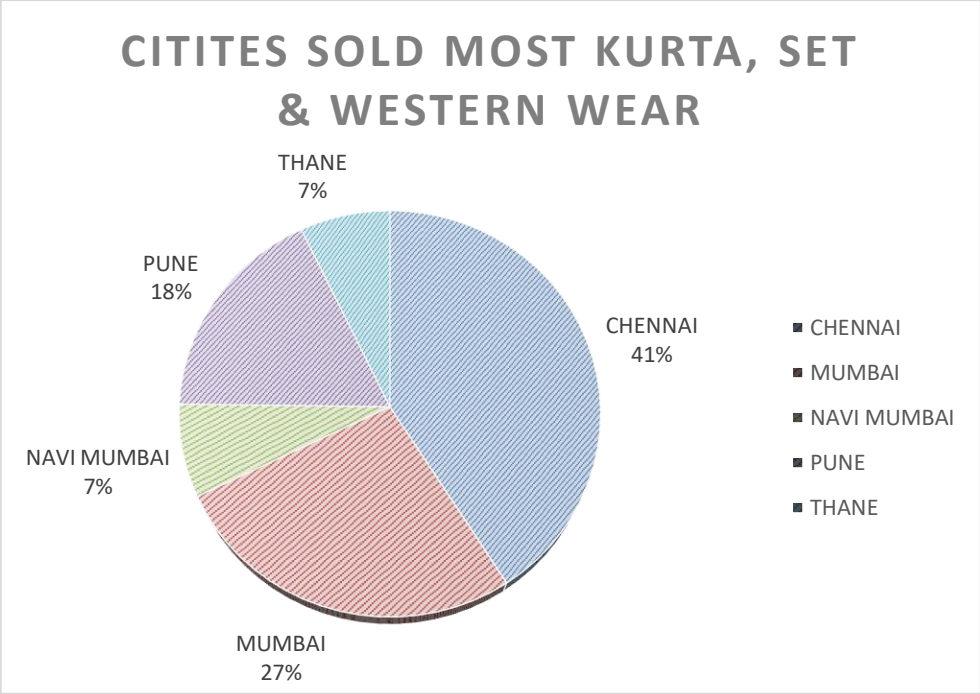


Ans:

- In Maharashtra: Sales in men category=1390, Sales in women category= 3144
- In Rajasthan: Sales in men category=21, Sales in women category=543
- In Tamil Nadu: Sales in men category=686, Sales in women category= 2023

Q.4 Which city sold most of following categories

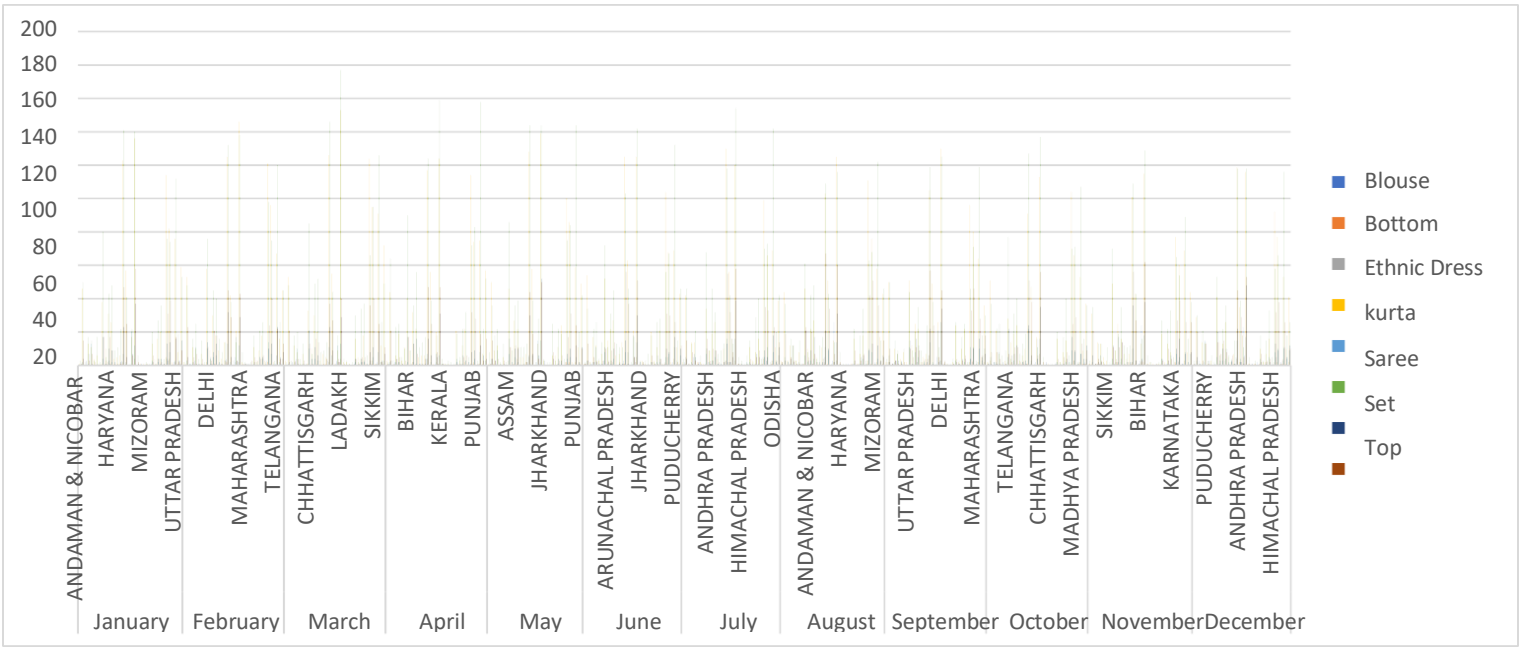
a. Kurta
b. Set
c. Western wears



City	kurta	Set	Western Dress	Grand Total
BENGALURU	964	938	422	2324
CHENNAI	666	451	217	1334
HYDERABAD	713	687	370	1770
MUMBAI	437	515	207	1159
NEW DELHI	479	792	142	1413
Grand Total	**3259**	**3383**	**1358**	**8000**

Ans:-Bengaluru, Chennai, Hyderabad, Mumbai and New Delhi are the cities sold most ofkurtas, Sets and western wears.

Q.5 In which month most items sold in any of the state on the basis of category. The graph for most items sold in any of stats on basis of category is as follows:



Ans: Most items sold are in month of march.

Conclusion and Review:

After thorough analysis of the store data, it is evident that there are notable trends and insights to be gleaned. By examining key metrics such as units sold, state wise analytics, geographic, and sales across different stats and products, we can draw valuable conclusions about market demand, sales and overall profitability. This comprehensive understanding will enable informed decision-making to optimize resources, target specific markets, and maximize profits in future store sales endeavours.

EXPLORING CAR DATADET REPORT

INTRODUCTION:

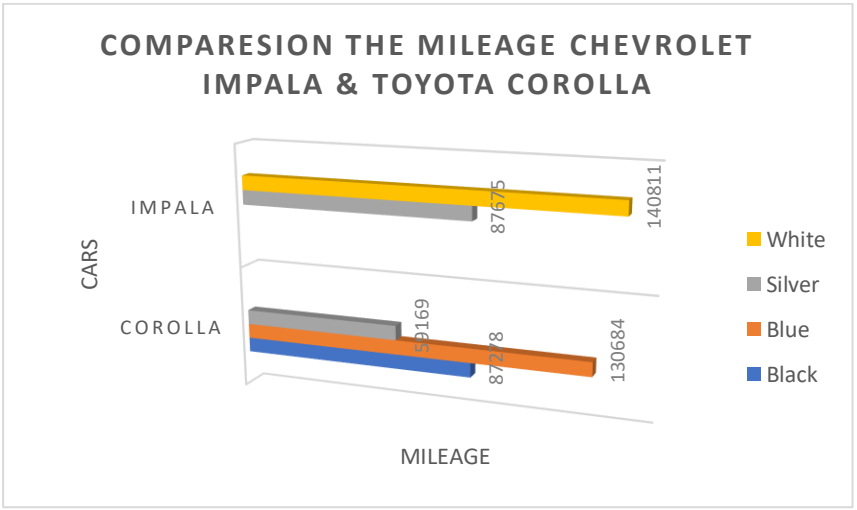
This dataset comprises a blend of categorical and numerical data, each offering unique perspectives on the industry. Categorical data, such as make, model, and color, encapsulates the diversity of vehicles and consumer preferences. Meanwhile, numerical attributes like mileage, price, and cost provide quantifiable metrics essential for analyzing market trends and pricing dynamics.

QUESTIONNAIRES:

1. Compare the mileage of Chevrolet Impala to Toyota Corolla. Which of the two is giving best mileage?
2. Justify, Buying of any Ford car is better than Honda
3. Among all the cars which car 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?

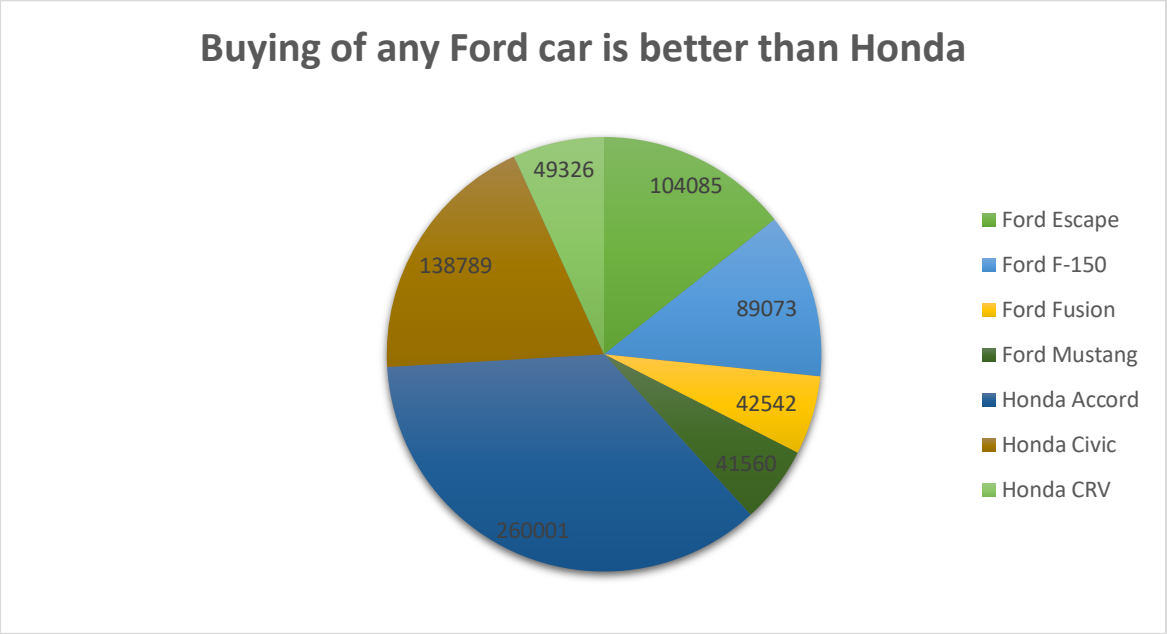
ANALYTICS:

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



Ans: The Toyota Corolla has a total mileage of 277,131 miles across all available models, while the Chevrolet Impala has a total mileage of 228,486 miles. Comparing the two, the Toyota Corolla offers better overall mileage.

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

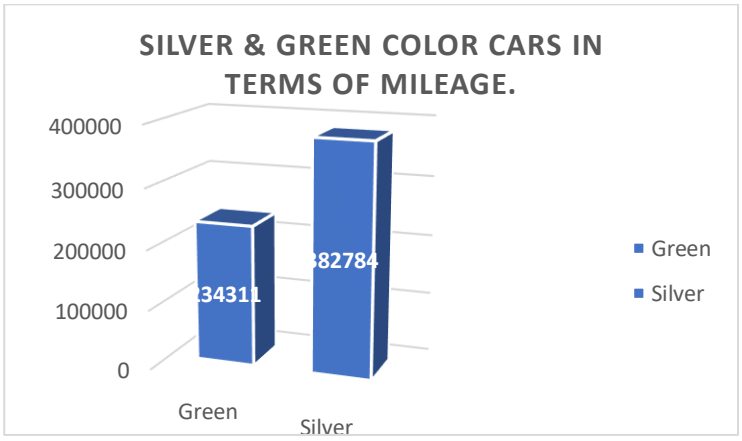


Ans: it appears that Honda cars collectively have a higher total mileage compared to Ford cars. However, it's essential to remember that mileage alone may not be the sole determining factor in choosing a car. Here are some additional points to consider:

1. **Performance:** Compare the performance metrics such as acceleration, handling, and engine power of Ford and Honda models you are considering.
2. **Reliability:** Look into the reliability ratings, recalls, and customer reviews for both Ford and Honda vehicles.
3. **Features:** Evaluate the features offered in both Ford and Honda cars, such as infotainment systems, safety features, comfort amenities, and driver-assistance technologies.
4. **Safety Ratings:** Check the safety ratings and crash test results from organizations like the National Highway Traffic Safety Administration (NHTSA) and the Insurance Institute for Highway Safety (IIHS).
5. **Price:** Compare the prices of comparable Ford and Honda models, including the initial purchase price, maintenance costs, and resale value.
6. **Personal Preference:** Consider your personal preferences, including styling, brand loyalty, and any specific requirements or preferences you have for your vehicle.

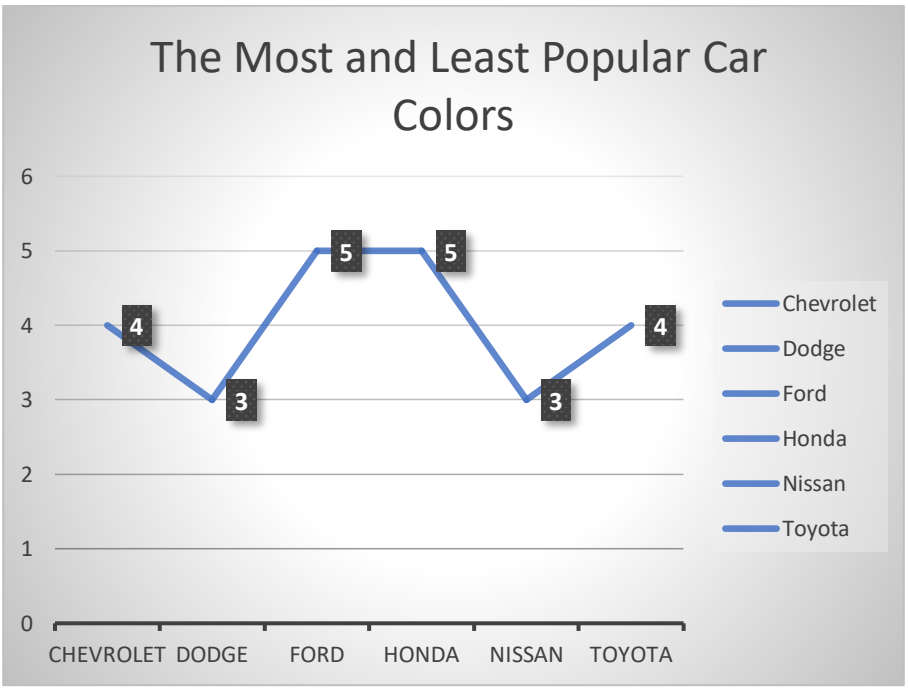
Ultimately, the decision to buy a Ford car over a Honda car (or vice versa) depends on your individual needs, priorities, and preferences, considering all relevant factors beyond just mileage.

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



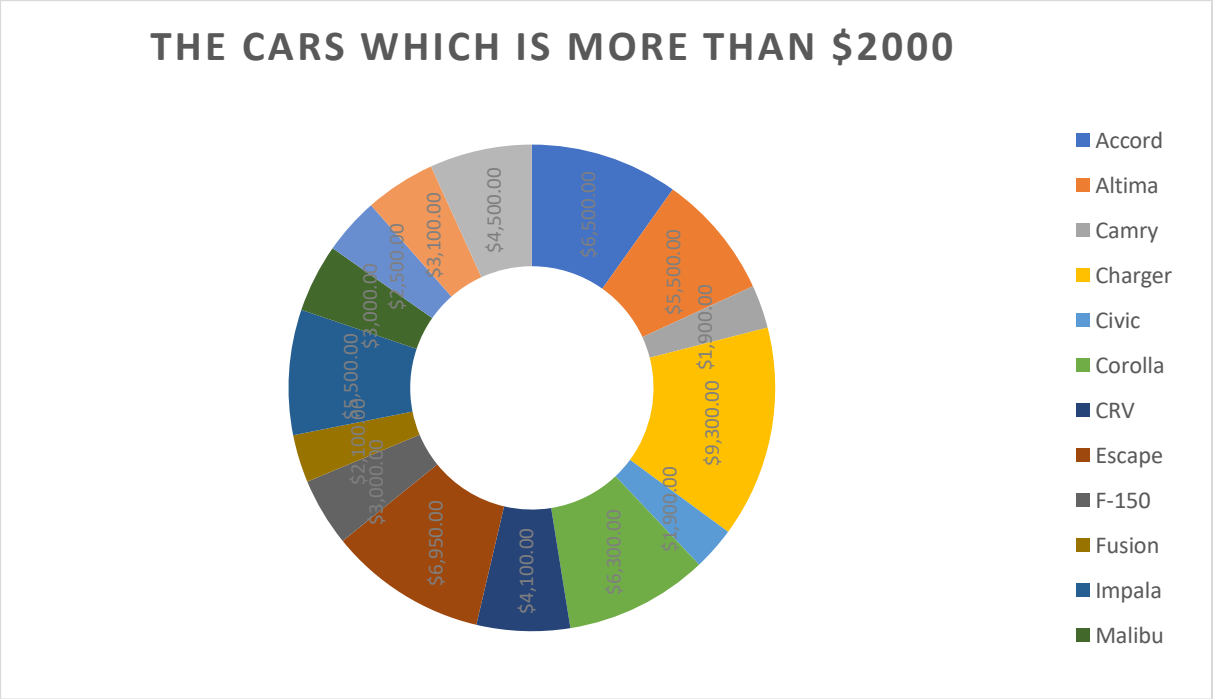
Ans: Among all the cars, silver is the most popular car color, with a total mileage of 382,784 miles. Conversely, green is the least popular car color, with a total mileage of 234,311 miles.

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



Ans: In terms of mileage, cars with silver color collectively have a higher count compared to cars with green color. Specifically, Chevrolet, Ford, and Toyota each have 4 models in silver color, while Dodge and Honda have 3 models each in silver color. Conversely, all other brands have no models in green color, resulting in a total count of 0 for green-colored cars. Therefore, silver-colored cars generally offer more options and potentially higher mileage compared to green-colored cars.

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



Ans:- All the mentioned cars, including Accord, Altima, Charger, Corolla, CRV, Escape, F-150, Fusion, Impala, Malibu, Maxima, Mustang, and Silverado, have prices exceeding $2000

ANOVA:

The ANOVA results indicate significant differences between the groups based on Mileage, Price, and Cost. The F-statistic is large (128.88), with a very low p-value (5.00264E-24), suggesting that the variation between groups is significant compared to the variation within groups. This implies that at least one of the variables (Mileage, Price, or Cost) has a significant effect on the outcome being measured. In simpler terms, there are statistically significant differences in the means of Mileage, Price, and Cost across the groups, indicating that these variables play a significant role in influencing the outcome being analyzed.

ANOVA: Single Factor

SUMMARY

	Count	Sum	Average	Variance	Groups	
Mileage	24	2011267	83802.7917	1214155660		
Price	24	78108	3254.5	837024.087		
Cost	24	66150	2756.25	705502.717		

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.0445E+11	2	5.2227E+10	128.882161	5.0026E-24	3.12964398
Within Groups	2.7961E+10	69	405232729			
Total	1.3242E+11	71				

This ANOVA table summarizes the results of a single-factor ANOVA test. The test compares the means of three different groups (presumably categorized by Mileage, Price, and Cost) to determine if there are statistically significant differences between them. Let's break down the table:

- **Count**: The number of observations in each group.
- **Sum**: The sum of values in each group.
- **Average**: The average value within each group.
- **Variance**: The variance within each group.
- **Groups**: Indicates the groups being compared.
- **ANOVA**: The total number of observations and the total sum.
- **Source of Variation**: This section breaks down the variation into two components:
- **Between Groups**: The variation between the group means.
- **Within Groups**: The variation within each group, also known as error variation.
- **SS (Sum of Squares)**: The sum of squared deviations from the mean.
- **df (Degrees of Freedom)**: The degrees of freedom associated with each source of variation.

- **MS (Mean Square)**: The mean square for each source of variation, which is calculated as SS/df.
- **F**: The F-statistic, which is the ratio of the between-group variance to the within-group variance.
- **P-value**: The probability of observing an F-statistic as extreme as the one computed from the sample data, assuming that the null hypothesis (i.e., no difference between group means) is true. A low p-value indicates that the observed differences between group means are unlikely to be due to random chance.
- **F crit (Critical F-value)**: The critical value of the F-statistic at a certain significance level. If the computed F-value exceeds the critical value, then you reject the null hypothesis.

In this case, the p-value (5.0026E-24) is much smaller than the significance level of 0.05, indicating strong evidence against the null hypothesis. Therefore, you would reject the null hypothesis and conclude that there are statistically significant differences between at least one pair of group means.

ANOVA: Two-Factor Without replication:

ANOVA						
Source	*of*					
Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	34749383.3	23	1510842.75	47.6846408	2.2236E-14	2.01442484
Columns	2979036.75	1	2979036.75	94.023218	1.3629E-09	4.27934431
Error	728733.25	23	31684.0543			
Total	38457153.3	47				

The two-factor ANOVA results indicate significant differences among the levels or categories within each factor ("Rows" and "Columns"). Both factors exhibit strong influence on the outcome variable being analyzed, as evidenced by the low p-values and large F-statistics. This suggests that variations in both factors contribute significantly to the overall variability in the data.

REGRESSION:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.4110586
R Square	0.168969173
Adjusted R Square	0.131195044
Standard Error	32478.67693
Observations	24

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4718562180	4718562180	4.473145	0.045991655
Residual	22	23207018006	1054864455		
Total	23	27925580186			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	134754.2033	24986.30198	5.393123138	2.04E-05	82935.7846	186572.6221	82935.7846	186572.6221
X Variable 1	-15.65568034	7.402278713	-2.1149812	0.045992	-31.00706681	-0.304293877	-31.00706681	-0.304293877

1. **Regression Statistics:**
   - **Multiple R:** The correlation coefficient between the independent and dependent variables. It indicates the strength and direction of the linear relationship.
   - **R Square:** The coefficient of determination, representing the proportion of the variance in the dependent variable that is predictable from the independent variable.
   - **Adjusted R Square:** Similar to R Square, but adjusted for the number of predictors in the model.
   - **Standard Error:** The standard deviation of the residuals, representing the average distance that the observed values fall from the regression line.
   - **Observations:** The number of data points used in the regression analysis.
2. **ANOVA (Analysis of Variance):**
   - The ANOVA table tests the overall significance of the regression model.
   - **Regression:** The portion of the total variation in the dependent variable explained by the independent variable(s).
   - **Residual:** The unexplained variation in the dependent variable after accounting for the regression model.

- **Total:** The total variation in the dependent variable.

3. **Coefficients:**
    - **Intercept:** The value of the dependent variable when all independent variables are zero.
    - **X Variable 1:** The coefficient for the independent variable.
    - **Standard Error:** The standard deviation of the coefficient estimate.
    - **t Stat:** The t-statistic for testing the null hypothesis that the coefficient is equal to zero.
    - **P-value:** The probability of obtaining a t-statistic as extreme as observed, assuming the null hypothesis (coefficient is zero) is true.
    - **Lower 95% and Upper 95%:** The lower and upper bounds of the 95% confidence interval for the coefficient.
4. **Interpretation:**
    - The regression model is significant, as indicated by the p-value (0.045991655) being less than the significance level (usually 0.05).
    - The coefficient for "X Variable 1" is -15.65568034. This suggests that for each unit increase in X Variable 1, the dependent variable decreases by approximately 15.66 units.
    - Both the intercept and the coefficient for "X Variable 1" are statistically significant, as their p-values are less than 0.05.
    - The coefficient of determination (R Square) is 0.168969173, indicating that approximately 16.9% of the variance in the dependent variable is explained by the independent variable(s).

CORRELATION:

	Mileage	price
Mileage	1	0.4110586
price	0.4110586	1

The correlation matrix provided shows the correlation coefficients between two variables: Mileage and Price. Here's the interpretation:

- The correlation coefficient between Mileage and Mileage is 1, which is the highest possible correlation coefficient. This is because it's the correlation of a variable with itself, so it's perfectly correlated.
- The correlation coefficient between Mileage and Price is approximately 0.4110586. This indicates a moderate positive correlation between Mileage and Price. In other words, there is a tendency for higher mileage values to be associated with higher price values, but the correlation is not extremely strong.

This correlation coefficient suggests that there is a moderate positive relationship between Mileage and Price: as Mileage increases, Price tends to increase as well, but the relationship is not extremely strong.

DESCRIPTIVE STATICS:

Mileage		price		cost	
Mean Standard Error	83802.7917 7112.65205	Mean Standard Error	3254.5 186.751181	Mean Standard Error	2756.25 171.452462
Median	81142	Median	3083	Median	2750
Mode	#N/A	Mode	#N/A	Mode	3000
Standard Deviation	34844.7365	Standard Deviation	914.890205	Standard Deviation	839.942092
Sample Variance	1214155660	Sample Variance	837024.087	Sample Variance	705502.717
Kurtosis	-1.0971827	Kurtosis	-1.2029138	Kurtosis	-0.8126576
Skewness	0.38652215	Skewness	0.27201913	Skewness	0.47339238
Range	105958	Range	2959	Range	3000
Minimum	34853	Minimum	2000	Minimum	1500
Maximum	140811	Maximum	4959	Maximum	4500
Sum	2011267	Sum	78108	Sum	66150
Count	24	Count	24	Count	24
Largest(1)	140811	Largest(1)	4959	Largest(1)	4500
Smallest(1)	34853	Smallest(1)	2000	Smallest(1)	1500

Mileage:

- **Mean:** The average mileage is approximately 83802.79.
- **Standard Error:** The standard error of the mean is approximately 7112.65.
- **Median:** The median mileage is 81142.
- **Mode:** The mode is not available.
- **Standard Deviation:** The standard deviation of the mileage is approximately 34844.74.
- **Sample Variance:** The sample variance of the mileage is approximately 1214155660.
- **Kurtosis:** The kurtosis indicates the peakedness or flatness of the distribution. A negative value (-1.097) suggests a relatively flat distribution.

- **Skewness:** The skewness measures the asymmetry of the distribution. A positive value (0.386) suggests a right-skewed distribution.
- **Range:** The range of the mileage is 105958, indicating the difference between the maximum and minimum values.
- **Minimum:** The minimum mileage is 34853.
- **Maximum:** The maximum mileage is 140811.
- **Sum:** The sum of all mileage values is 2011267.
- **Count:** There are 24 observations for mileage.
- **Largest(1):** The largest value of mileage is 140811.
- **Smallest(1):** The smallest value of mileage is 34853.

Price:

- **Mean:** The average price is approximately 3254.5.
- **Standard Error:** The standard error of the mean is approximately 186.75.
- **Median:** The median price is 3083.
- **Mode:** The mode is not available.
- **Standard Deviation:** The standard deviation of the price is approximately 914.89.
- **Sample Variance:** The sample variance of the price is approximately 837024.087.
- **Kurtosis:** The kurtosis indicates the peakedness or flatness of the distribution. A negative value (-1.202) suggests a relatively flat distribution.
- **Skewness:** The skewness measures the asymmetry of the distribution. A positive value (0.272) suggests a right-skewed distribution.
- **Range:** The range of the price is 2959.
- **Minimum:** The minimum price is 2000.
- **Maximum:** The maximum price is 4959.
- **Sum:** The sum of all price values is 78108.
- **Count:** There are 24 observations for price.
- **Largest(1):** The largest value of price is 4959.
- **Smallest(1):** The smallest value of price is 2000.

Cost:

- **Mean:** The average cost is approximately 2756.25.
- **Standard Error:** The standard error of the mean is approximately 171.45.
- **Median:** The median cost is 2750.
- **Mode:** The mode is 3000.
- **Standard Deviation:** The standard deviation of the cost is approximately 839.94.

- **Sample Variance:** The sample variance of the cost is approximately 705502.717.
- **Kurtosis:** The kurtosis indicates the peakedness or flatness of the distribution. A negative value (-0.813) suggests a relatively flat distribution.
- **Skewness:** The skewness measures the asymmetry of the distribution. A positive value (0.473) suggests a right-skewed distribution.
- **Range:** The range of the cost is 3000.
- **Minimum:** The minimum cost is 1500.
- **Maximum:** The maximum cost is 4500.
- **Sum:** The sum of all cost values is 66150.
- **Count:** There are 24 observations for cost.
- **Largest(1):** The largest value of cost is 4500.
- **Smallest(1):** The smallest value of cost is 1500.

CONCLUSION AND REVIEWS:

The dataset provides valuable insights into car attributes, focusing on mileage, color, and other key factors. Here's a simple conclusion based on the data:

Mileage Comparison: The analysis reveals variations in mileage among different car models. Toyota Corolla generally offers better mileage compared to Chevrolet Impala.

Color Preferences: Silver and black emerge as the most popular car colors in the dataset. Blue, green, red, and white are among the least popular color choices.

Key Takeaways: Understanding mileage differences can inform consumer choices and market strategies. Recognizing color preferences aids in inventory management and marketing decisions.

ORDER DATASET REPORT

INTRODUCTION:

Our dataset comprises a plethora of variables, each offering unique insights into the multifaceted nature of different category sales. From fundamental transactional details such as Date, Time, sales, states to more nuanced factors like Customer Type, Demographics, category and sub category, every facet has been meticulously documented.

Key Attributes:

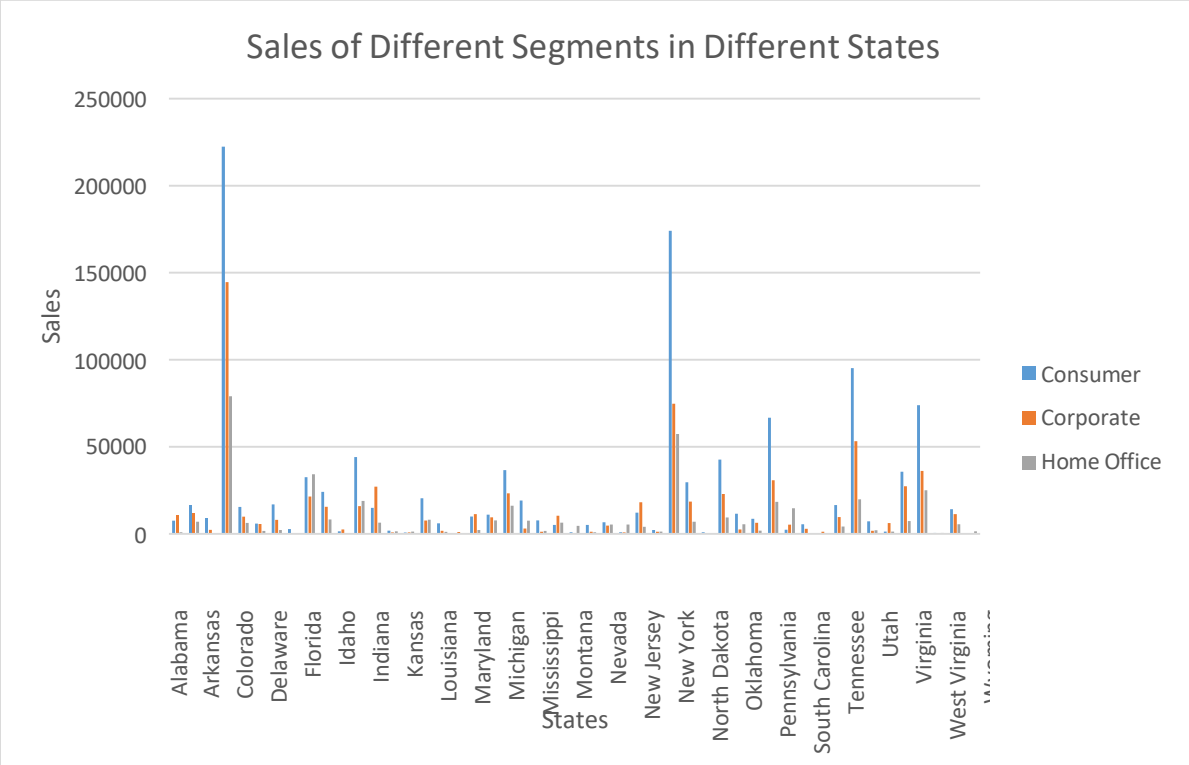
1. ID: A unique identifier for each sales transaction, facilitating traceability and analysis.
2. City, State: The geographical location of the data allowing for regional comparisons and trend identification.
3. Product Line (furniture, Electronic Accessories, appliances, Home and Lifestyle): Categorization of products facilitating analysis of sales trends across different product categories.
4. Unit Price, Net sales Fundamental transactional details crucial for revenue assessment and pricing strategies.
5. Net sales of different category, category performing well in different states: Performance metrics
6. Rating: different product performing well in different state
7. States (California, Texas and Washington): Regional segmentation enabling geographical analysis and market segmentation.

2. 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 segment?
5. Compare average sales of different category and sub category of all the states.

3. ANALYTICS:

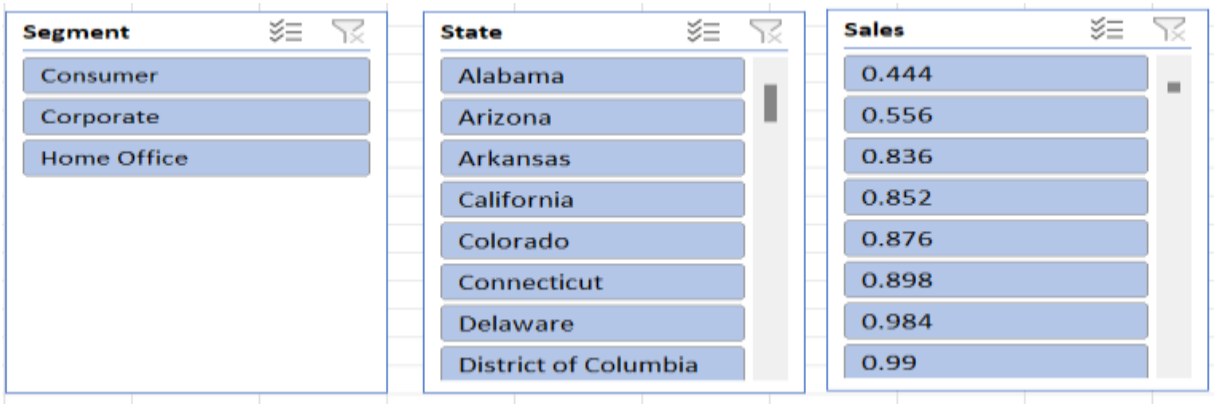
Q1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?



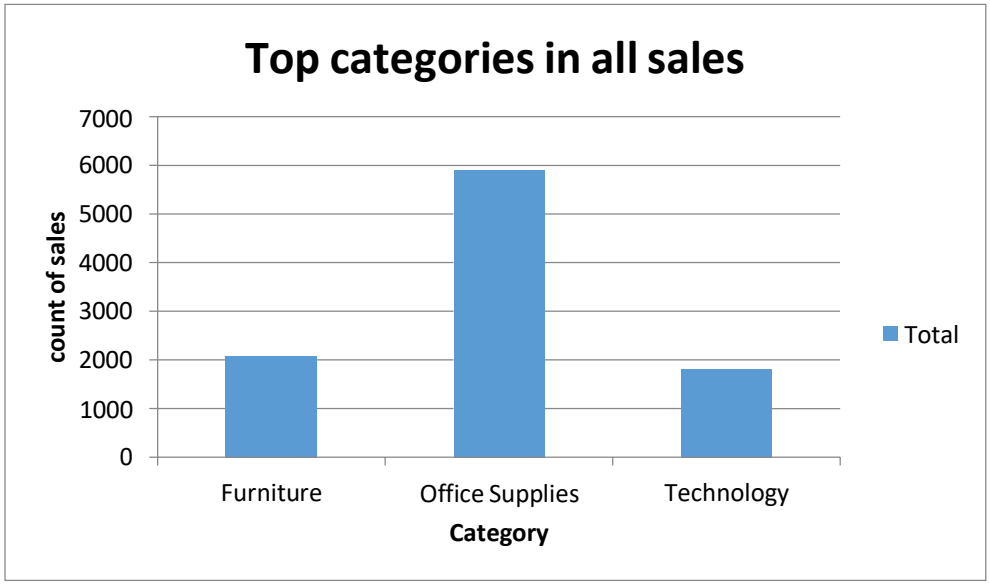
Ans:-

- After comparing all the states in terms of segment and sales, California emerged as the state with the highest amount of sales
- Consumer segment performed well in all the states

Slicers:

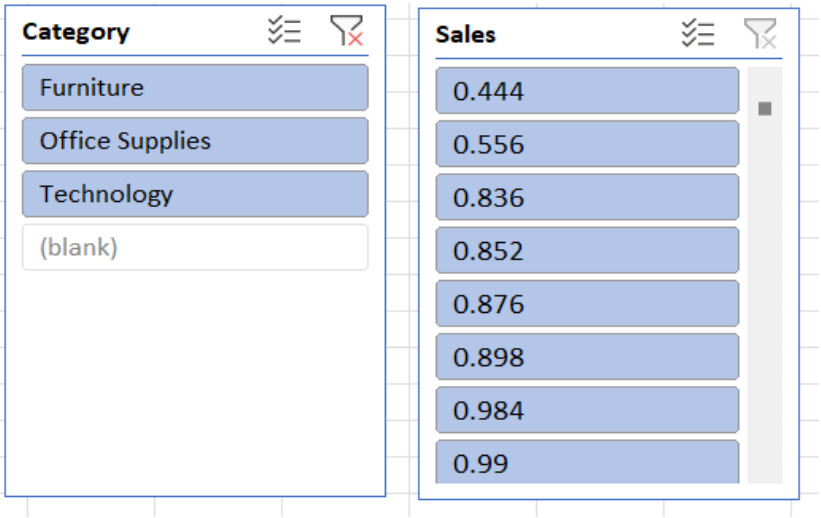


Q2. Find out top performing category in all the states?

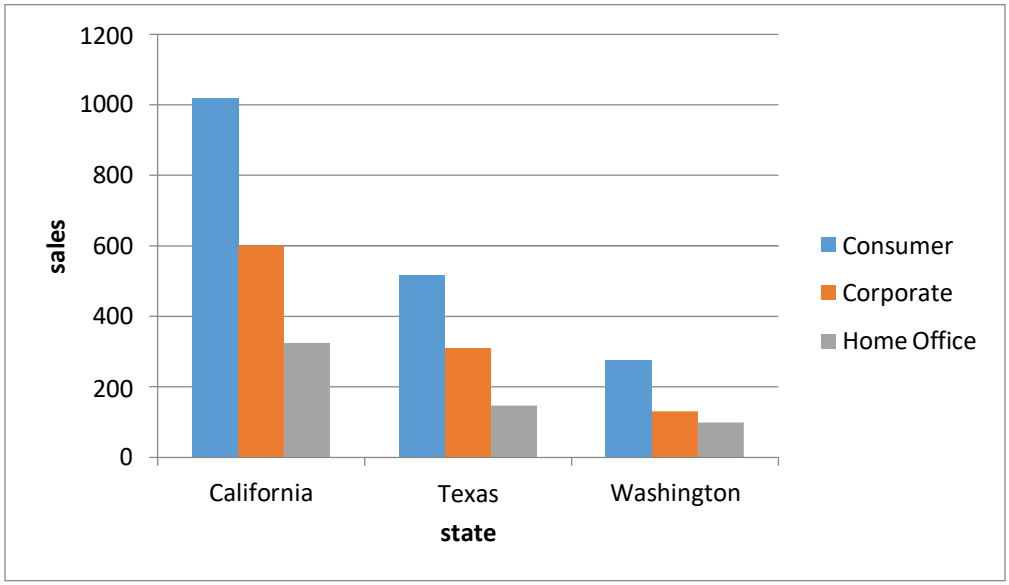


Ans. Office Supplies is the top performing category in all the states

Slicers:

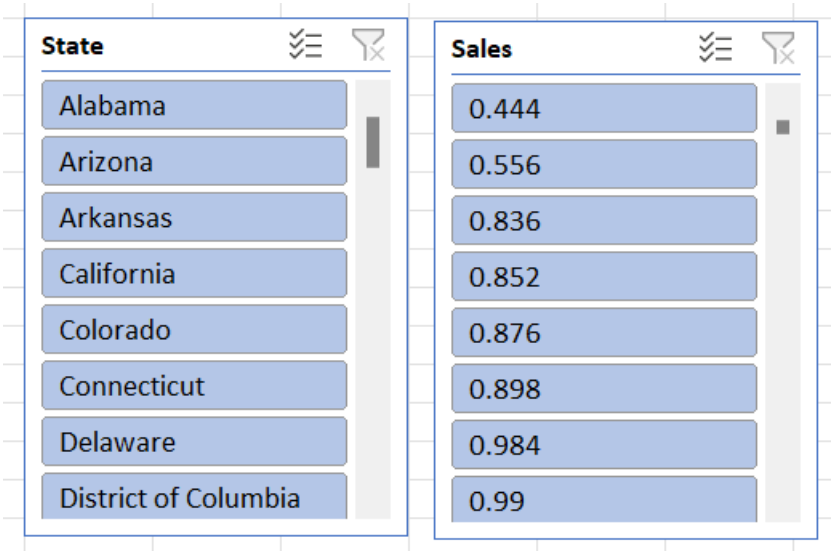


Q3. Which segment has most sales in US, California, Texas, and Washington?

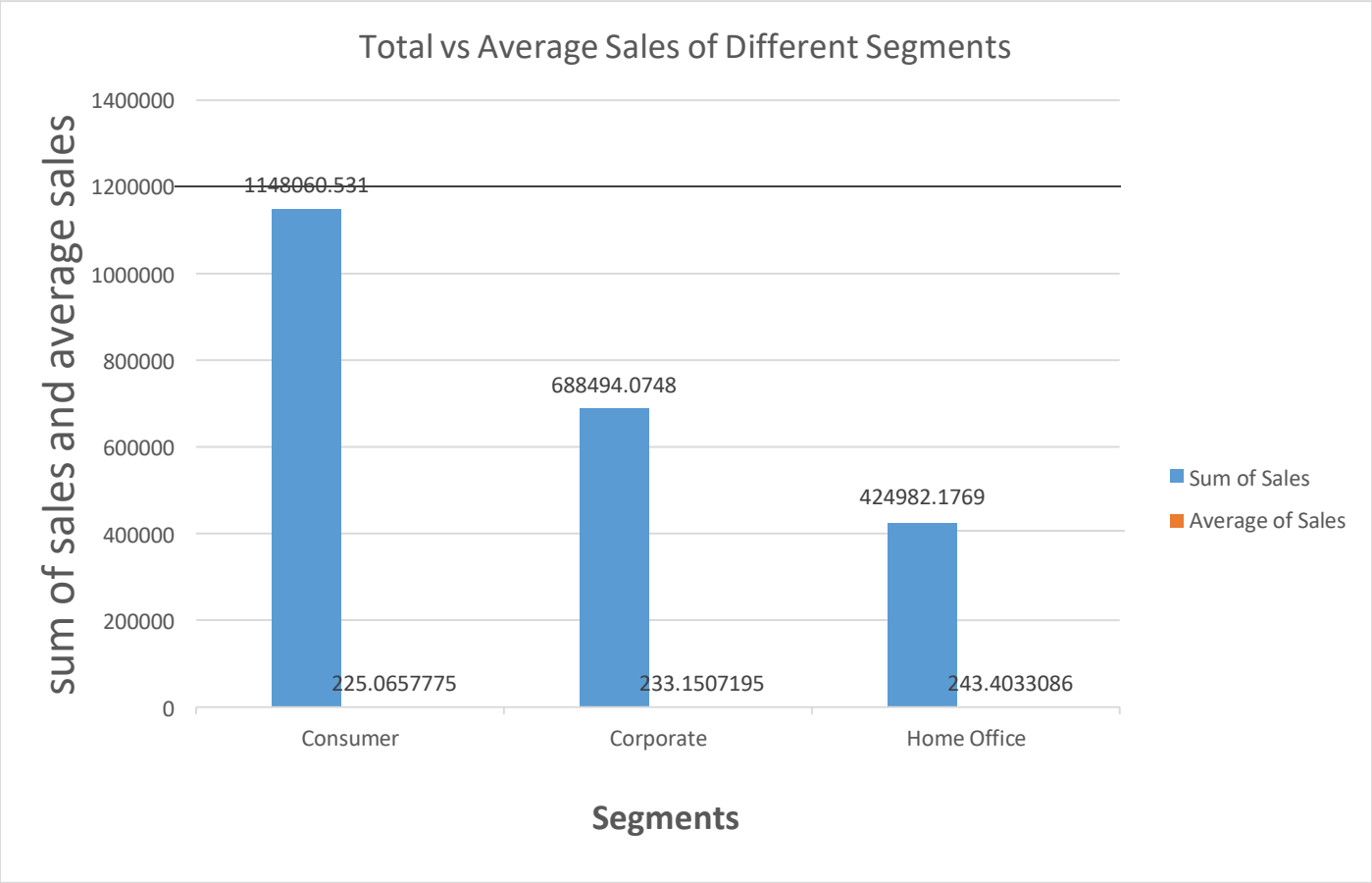


Ans. Consumer segment has the most sales in US, California, Texas, and Washington

Slicers:

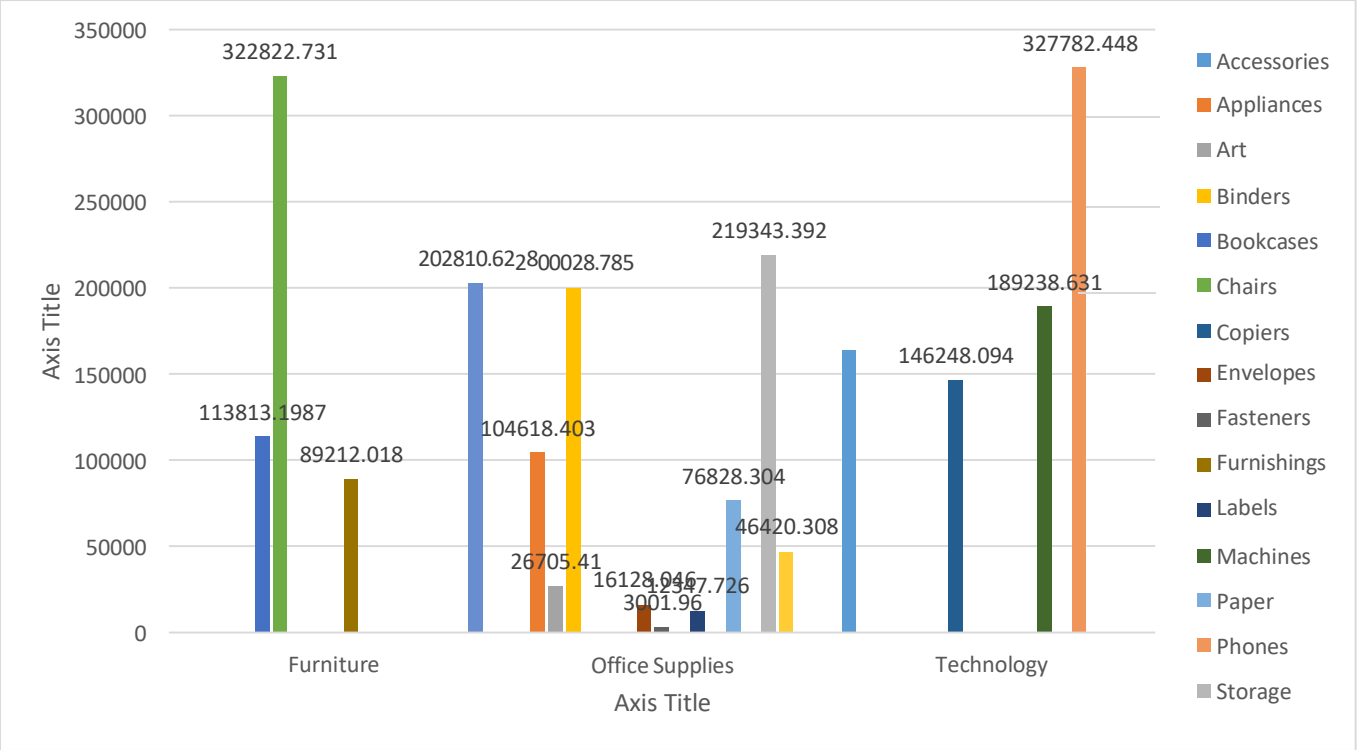


Q4. Compare total and average sales for all different segment?



Ans. By Analysis of the given data set we can found that in all the three segments the total sales were greater than the average sales.

Q5. Compare average sales of different category and sub category of all the states.



Ans. By doing analysis of the given Order Sales dataset we were able to observe that, average sales of Technology was far greater than rest of the categories

Regression and ANOVA:

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.008850713			
R Square	7.83351E-05			
Adjusted R Square	-0.000924595			
Standard Error	596.4161586			
Observations	999			
ANOVA				
	Df	*SS*	*MS*	*F*
Regression	1	27783.3433	27783.3433	0.078106235
Residual	997	354645097.6	355712.2343	
Total	998	354672880.9		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	232.3779806	37.2042048	6.246013907	6.22491E-10
Postal Code	0.000167458	0.000599189	0.279474927	0.779938343

This regression analysis aims to examine the relationship between two variables: an independent variable represented by "Postal Code" and a dependent variable (not explicitly mentioned in the output). Here's an explanation of the key components:

1. Regression Equation:

The regression equation is of the form:

Y = 232.38 + 0.000167458* (Postal Code)

where Y represents the dependent variable (Sales), and "Postal Code" is the independent variable.

2. Interpretation of Coefficients:

The intercept coefficient (232.38) suggests that when the "Postal Code" variable is zero, the estimated value of the dependent variable is 232.38. However, the interpretation of this intercept may not be meaningful since postal codes are unlikely to be zero.

The coefficient for "Postal Code" (0.000167458) suggests that for every one-unit increase in the postal code, the estimated value of the dependent variable increases by approximately

0.000167458 units. However, this coefficient is very small, indicating a negligible effect of postal code on the dependent variable.

3. Statistical Significance:

The p-value associated with the coefficient for "Postal Code" is 0.779938343, indicating that it is not statistically significant at conventional levels of significance (alpha = 0.05). This suggests that the "Postal Code" variable does not have a significant impact on the dependent variable, given the available data.

4. Goodness of Fit:

- The R-squared value (0.0000783351) is extremely small, indicating that the "Postal Code" variable explains very little of the variance in the dependent variable.

- The Adjusted R-squared value (-0.000924595) is negative, which can happen when the model is over fit or when the independent variable is not relevant. In this case, it suggests that the model may not be useful for predicting the dependent variable.

5. ANOVA:

- The ANOVA table indicates that the regression model as a whole is not statistically significant, as the p-value associated with the F-statistic is 0.779938343.

6. Standard Error:

- The standard error (596.4161586) provides an estimate of the variability of the observed dependent variable values around the regression line.

7. Observations:

- The analysis is based on a sample of 999 observations.

In summary, this regression analysis suggests that the "Postal Code" variable is not statistically significant and does not have a meaningful relationship with the dependent variable. Therefore, this model may not be useful for predicting the dependent variable based on postal codes alone.

Correlation:

The absolute value of the correlation coefficient (0.024067424) is close to zero. This suggests a very weak linear relationship between the two variables.

Descriptive Statistics:

Sales	
Mean	230.7691
Standard Error	6.33014
Median	54.49
Mode	12.96
Standard Deviation	626.6519
Sample Variance	392692.6
Kurtosis	304.4451
Skewness	12.98348
Range	22638.04
Minimum	0.444
Maximum	22638.48
Sum	2261537
Count	9800

4. CONCLUSION:

Our comprehensive analysis of the provided dataset through various data visualization techniques has yielded valuable insights. Through the creation of bar graphs, pie charts, and other visual representations, we've been able to discern patterns, trends, and relationships within the data that might have otherwise remained obscured.

Our deep dive into the dataset has not only enhanced our understanding of the underlying information but has also empowered us to make informed decisions based on the insights gained. By visually depicting the data, we've been able to communicate complex findings in a clear and accessible manner, facilitating better comprehension and actionable strategies.

Furthermore, this process has underscored the importance of data visualization as a powerful tool for extracting meaningful information from raw data. By harnessing the visual nature of graphs and charts, we've transformed numbers and statistics into compelling narratives that drive understanding and inform decision-making.

LOAN DATASET REPORT:

1. INTRODUCTION:

Dataset Overview:

Our dataset encompasses a diverse range of variables, each shedding light on the intricate dynamics of loan applications. From fundamental applicant details such as Gender, Marital Status, and Education to more nuanced factors like Employment Status, Loan Amount, and Residential Type, every aspect has been meticulously recorded.

Key Attributes:

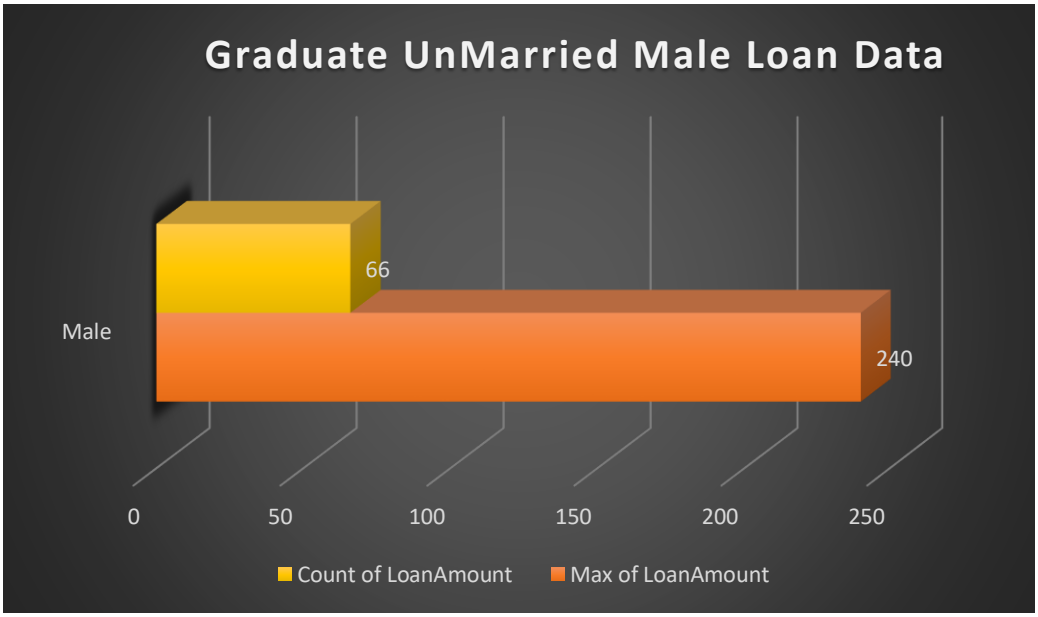
1. Gender: A demographic identifier providing insights into the gender distribution among loan applicants.
2. Marital Status (Married, Not Married): Categorization based on marital status aiding in demographic segmentation.
3. Education (Graduate, Non-graduate): Classification based on educational background for further analysis.
4. Employment Status (Employed, Unemployed): Distinction between employed and unemployed applicants, crucial for risk assessment.
5. Loan Amount: The principal amount applied for, providing a measure of financial need and capacity.
6. Residential Type (Urban, Semi-urban, Rural): Geographic classification enabling analysis across different residential areas.

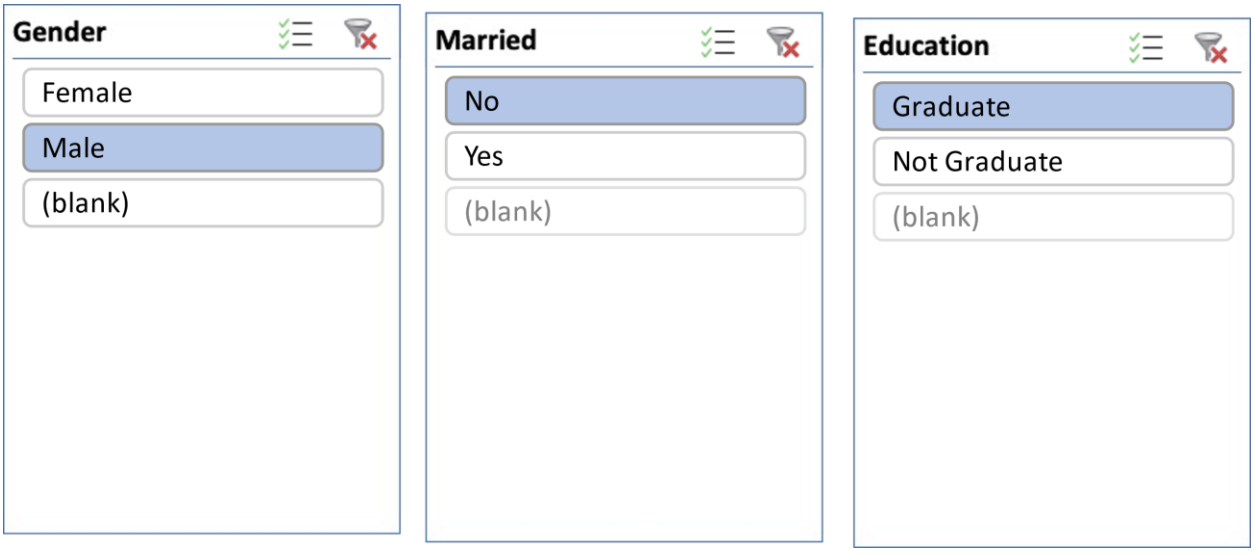
2. QUESTIONNAIRE:

Q1. How many male graduates who are not married applied for Loan? What was the highest amount?
Q2. How many female graduates who are not married applied for Loan? What was the highest amount?
Q3. How many male non-graduates who are not married applied for Loan? What was the highest amount?
Q4. How many female graduates who are married applied for Loan? What was the highest amount?
Q5. How many male and female who are not married applied for Loan? Compare Urban, Semi-urban and rular on the basis of amount.

3. ANALYTICS:

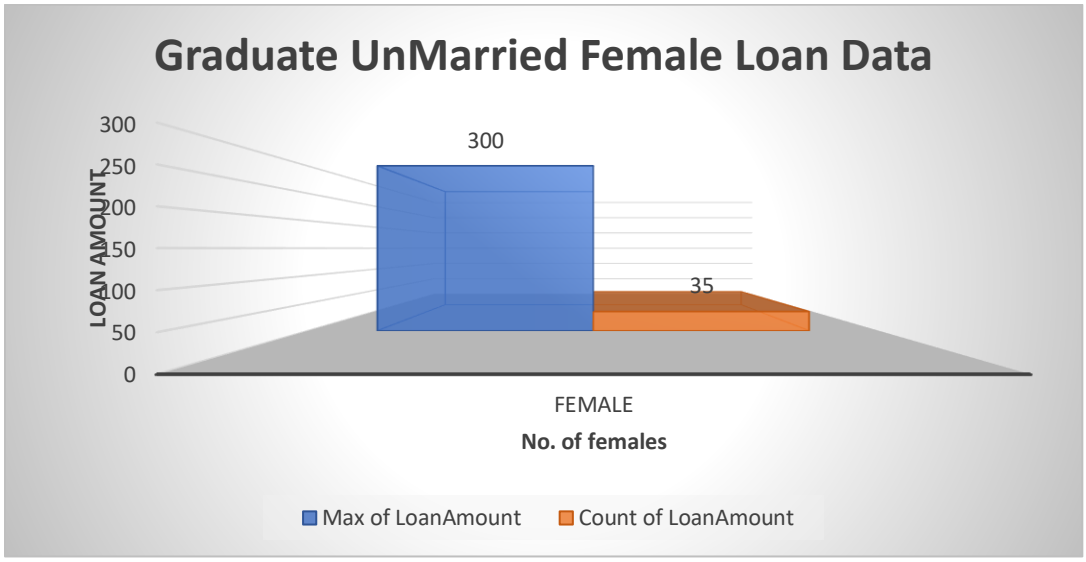
Q1. How many male graduates who are not married applied for Loan? What was the highest amount?





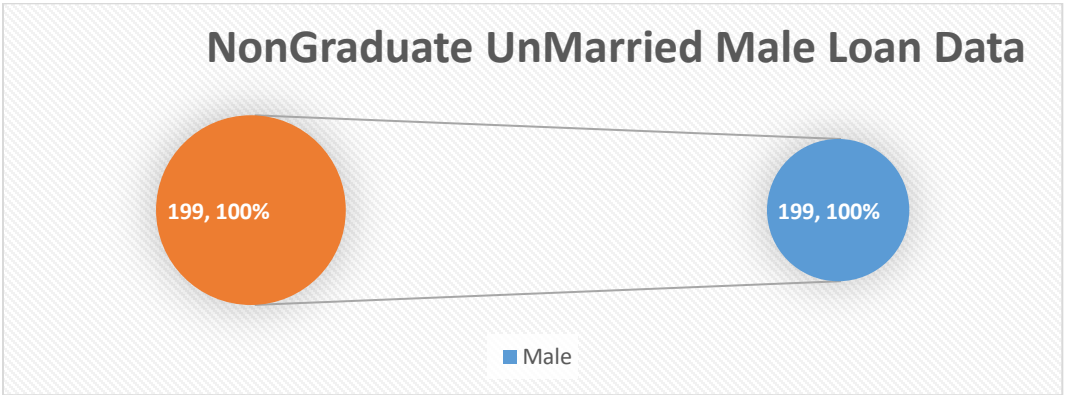
Answer- Out of 240 loan applicants who were unmarried graduates and males, the highest loan amount applied for was $66.

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



Ans:- Among the 300 unmarried female loan applicants who were graduates, the highest loan amount applied for was $35.

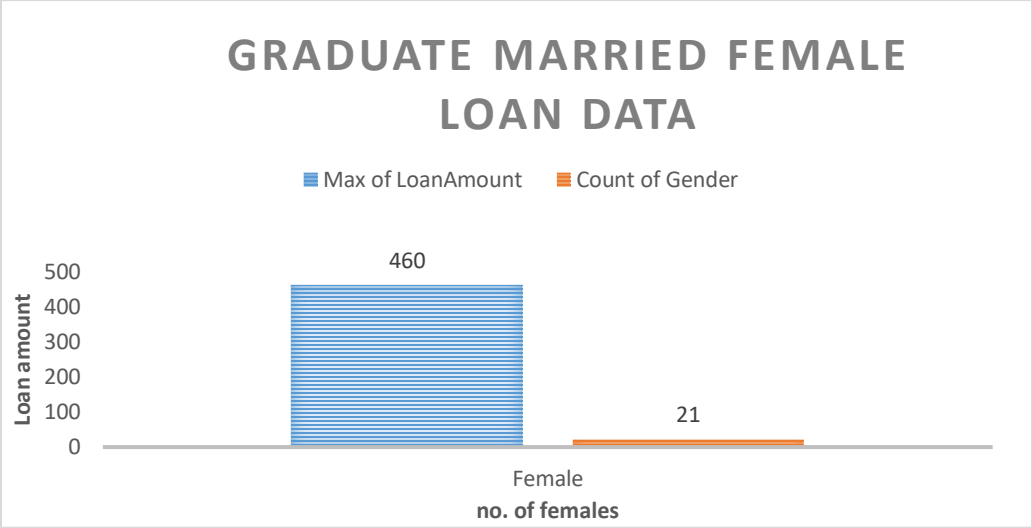
Q3. How many male non-graduates who are not married applied for Loan?
What was the highest amount?



Ans:- Out of the 199 loan applicants who were unmarried males and non-graduates, the highest loan amount applied for was $16.

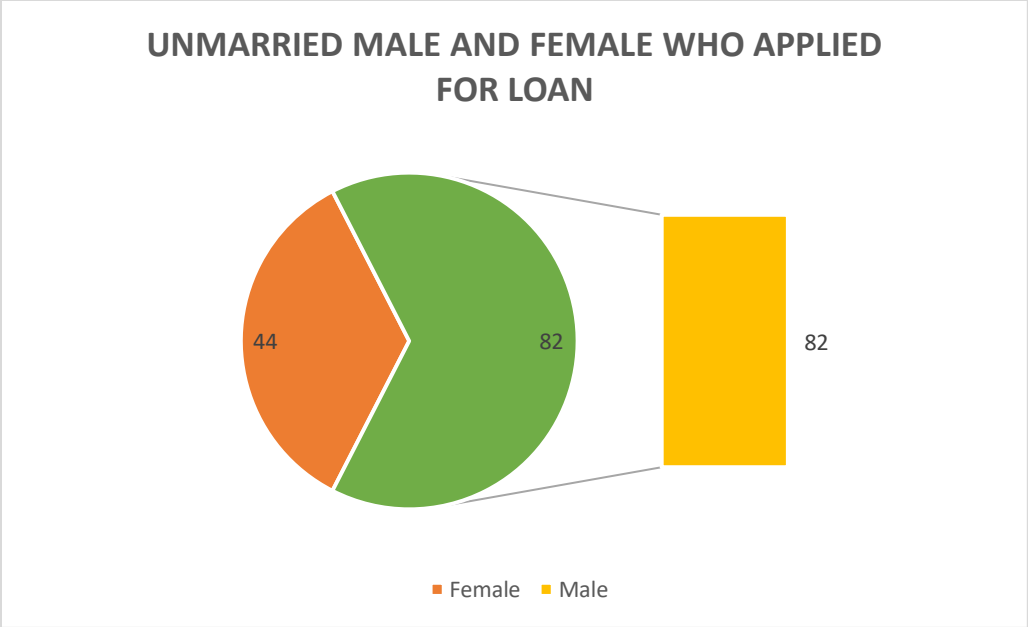
Q4. How many female graduates who are married applied for Loan?

What was the highest amount?



Ans:- Among the 460 married female loan applicants who were graduates, the highest loan amount applied for was $20.

Q5. How many male and female who are not married applied for Loan? Compare Urban, Semi-urban and rular on the basis of amount.



Ans:- The number of loan applications from unmarried males exceeded those from females by 38 requests. The average loan amount in rural areas is $131.182, in semi-urban areas is $134.04, and in urban areas is $136.22.

4. CONCLUSION:

Our analysis, using varied visualization techniques, revealed valuable insights, enhancing comprehension and decision-making. Visualizing data clarified complex findings, facilitating actionable strategies. This highlights the pivotal role of data visualization in extracting meaningful insights and informing decisions effectively.

5. REGRESSION:

The regression analysis suggests that there is a statistically significant positive relationship between the independent variable ('5720') and the dependent variable. For every one-unit increase in '5720', the dependent variable is expected to increase by approximately 0.0059 units. However, it's important to note that the model only accounts for about 21.1% of the total variance in the dependent variable.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.45908096
R Square	0.21075532

Adjusted R Square	0.20858707
Standard Error	56.0766111
Observations	366

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	305655.205	305655.205	97.2004502	1.7676E-20
Residual	364	1144629.42	3144.58631		
Total	365	1450284.62			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
Intercept	106.07753	4.10024098	25.8710478	1.7585E-84	98.014396	114.140665	98.0143
5720	0.0058851	0.00059692	9.85902887	1.7676E-20	0.00471125	0.00705895	0.00471

6. CORRELATION :

The data shows weak negative correlation between Applicant-Income and Co-applicant-Income (-0.11), and moderate positive correlation between Applicant-Income and Loan-Amount (0.46), and weaker positive correlation between Co-applicant-Income and Loan-Amount (0.14).

	ApplicantIncome	*CoapplicantIncome*	*LoanAmount*
ApplicantIncome	1		
CoapplicantIncome	-0.110334799	1	
LoanAmount	0.458768926	0.144787815	1

7. Anova (Single Factor) :

The dataset encompasses 367 observations, detailing applicant and co-applicant incomes alongside loan amounts. On average, applicants possess a higher income, averaging around $4805.60, compared to co-applicants whose average income is approximately $1569.58. Loan amounts vary widely, averaging $134.28. ANOVA analysis underscores significant distinctions between the income and loan amounts across the groups, implying diverse financial profiles among applicants and co-applicants.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*

ApplicantIncome	367	176365	4805.599455	24114831.09
CoapplicantIncome	367	576035	1569.577657	5448639.491
LoanAmount	367	49280	134.2779292	3964.141124

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4202537452	2	2101268726	213.2009841	5.87569E-79	3.003920577
Within Groups	1082168110	1098	9855811.573			
Total	1502421856	1100				

8. Anova two factor without Replication:

The ANOVA results indicate significant variation both within rows ($p = 0.441$) and between columns ($p < 0.001$). This suggests that there are meaningful differences among the row categories and column categories in the dataset, warranting further investigation into the factors influencing these variations.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1004340909	365	2751618.93	1.015674698	0.440986529	1.1881716
Columns	379216841.8	1	379216841.8	139.9761235	1.47092E-27	3.867061668
Error	988841123.7	365	2709153.763			
Total	2372398875	731				

9. Descriptive Statistics:

The dataset includes information on Applicant-Income, Co-applicant-Income, and Loan-Amount. The largest Applicant-Income recorded is $72,529, while the smallest is $0. For Co-applicant-Income, the largest value is $24,000, and the smallest is $0. Additionally, the Loan-Amount ranges from a maximum of $550 to a minimum of $0. Confidence levels for these variables at a 95.0% level are also provided, indicating the precision of the measurements within the dataset.

Largest(1)	72529	Largest(1)	24000	Largest(1)	
Smallest(1)	0	Smallest(1)	0	Smallest(1)	

Confidence Level(95.0%)	504.0756067	Confidence Level(95.0%)	239.6059543	Confidence Level(95.0%)	6.46

SHOP SALES DATA REPORT

Introduction :

This dataset encapsulates a wealth of information regarding sales transactions, providing valuable insights into the dynamics of retail operations. With columns meticulously crafted to capture key facets of each transaction, including Date, Salesman, Item Name, Company, Quantity, and Amount, analysts and businesses alike gain access to a treasure trove of actionable data.

Whether it's uncovering trends, optimizing inventory management, or refining sales strategies, this dataset serves as an invaluable resource for driving informed decision-making and unlocking new avenues for growth.

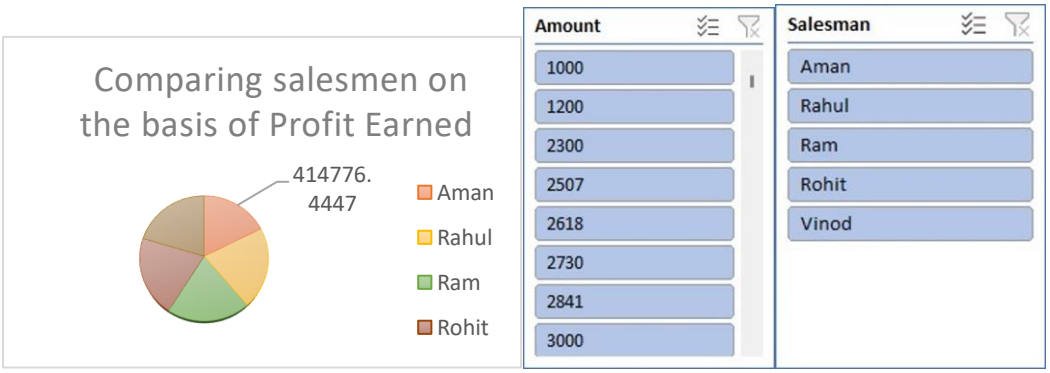
Questionaries :

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

Analytics :

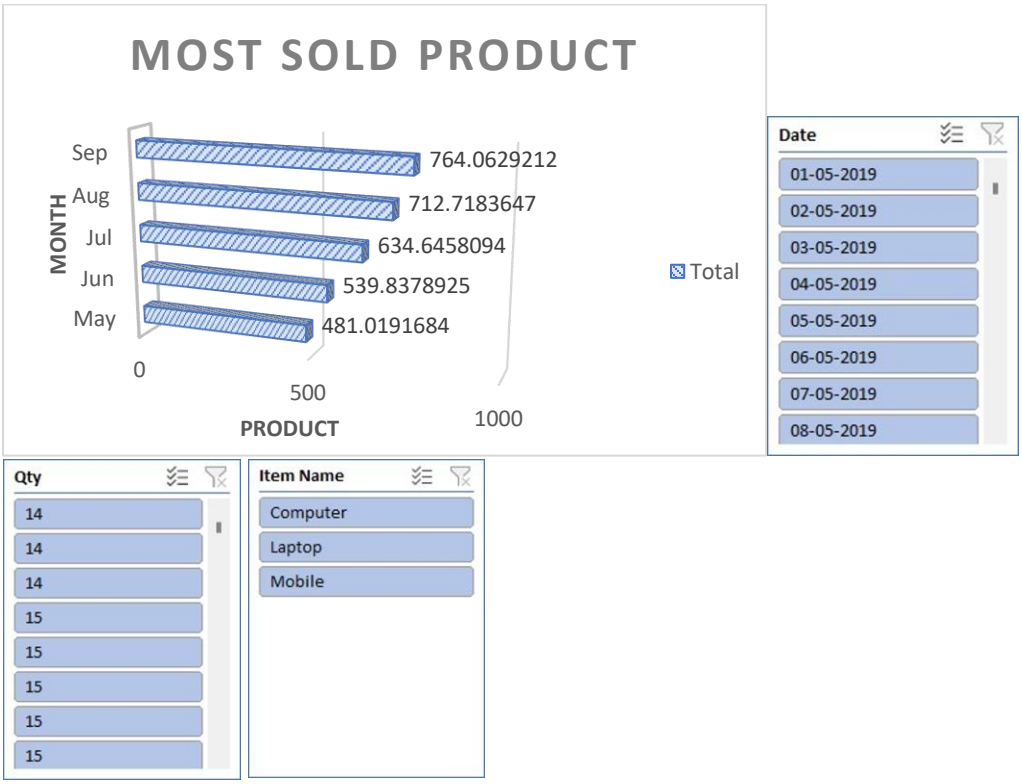
Q.1. Compare all the salesmen on the basis of profit earn.

Ans:- The comparison of all the salesmen on the basis of profit earned is given below:



Ans:-Aman earned a profit of $43,476.4, Ram earned -$476,120.38 (indicating a loss), Rohit earned $485,039.11, and Vinod earned $478,167.14. Rahul earned $493,541.37, making him the salesman with the highest profit.

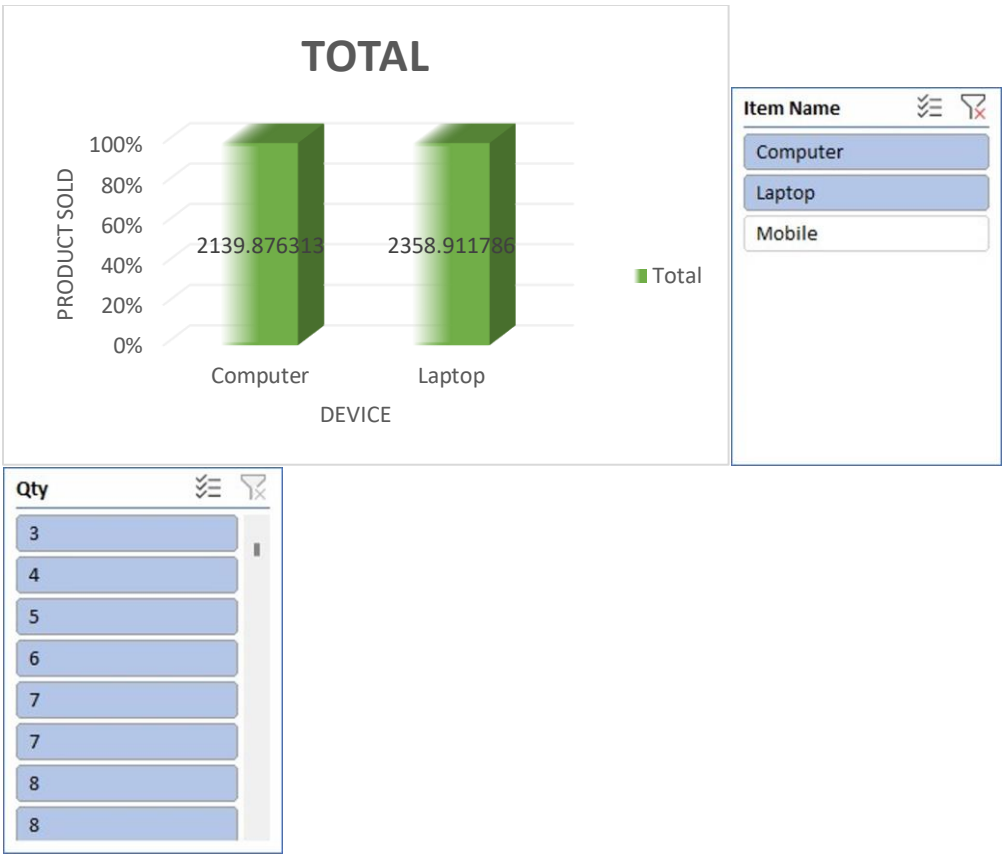
Q.2. Find out most sold product over the period of May-September.



Ans:- After filtering out the months from May to September, laptops recorded the highest sales.

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

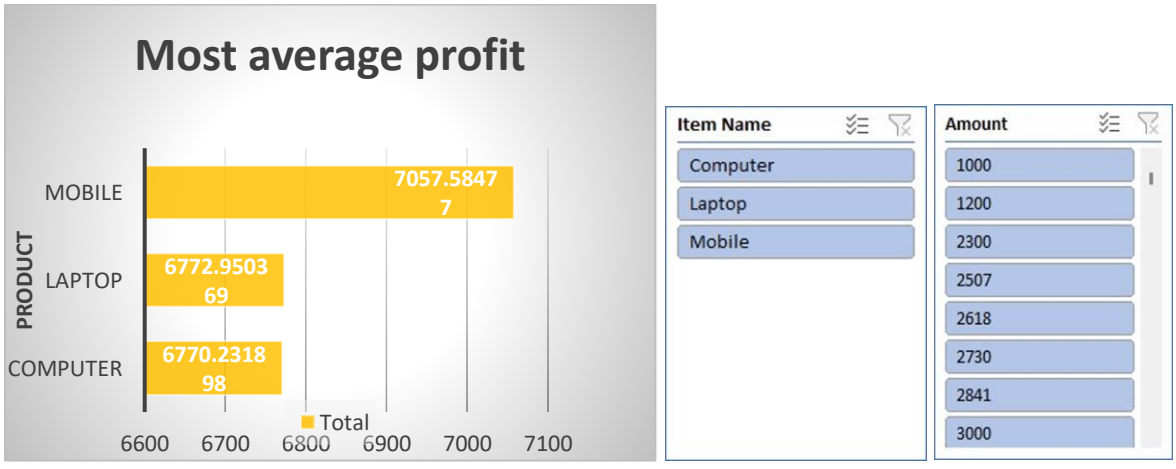
The two product sold the most over the year between computer or laptop :



Ans:- Laptop was sold more than Computer over the year.

Q.4 . Which item yield most average profit?

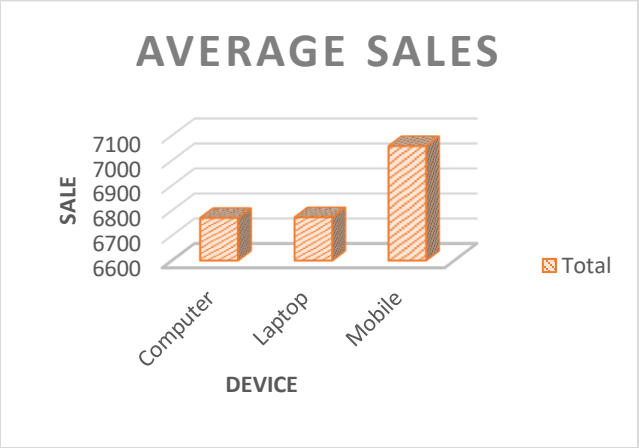
The item that yields the most profit between laptop, computer and mobile is :

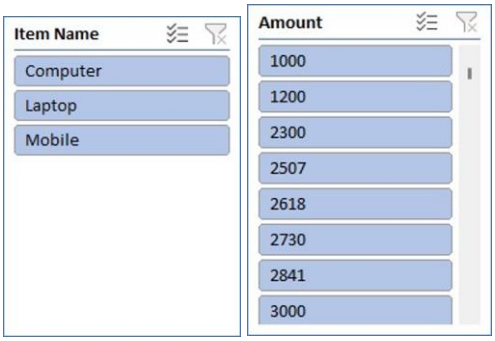


Ans:-Mobiles have the highest average profit among all other products.

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

The average sales of all the products with their respective comparison is :





Ans:-The average sales for mobiles were $7,057.584, for laptops were $6,772.95, and for computers were $6,770.23. Mobiles had the highest sales.

Conclusion and Review :

The shop sales dataset offers insights into sales trends, salesman performance, item popularity, and company performance. Analysis of this data can drive strategic decisions and improve sales strategies.

The dataset is well-structured and provides comprehensive information on sales transactions. It allows for various analyses, but could benefit from additional variables for deeper insights. Overall, it's a valuable resource for understanding sales dynamics and informing business decisions.

Regression:

The regression model, with a significant p-value indicates a strong positive relationship between Amount and the profit earned and the outcome variable. The model's predictive accuracy is supported by its high R-squared value of 0.660.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.812617
R Square	0.660347
Adjusted R Square	0.629469
Standard Error	1215.119
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	31576697	31576697	21.38598	0.000753
Residual	11	16241653	14776514		
Total	12	47818350			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	244.7062	754.0557	0.32452	0.751632	-1414.96	1904.372
X Variable	0.190729	0.041243	4.624498	0.000735	0.099954	0.281505

Correlation:

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

	Qty	*Amount*
Column 1	1	
Column 2	#DIV/0!	1

ANOVA (Single Factor) :

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

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	15	78.56643	5.237762	2.766871
Column 2	15	50419.05	3361.27	3416099

ANNOVA

Source of Variance	SS	df	MS	F	P-Value	F crit
Between Group	84472135	1	84472135	49.45528	1.2E-07	4.195972
Without Group	47825420	28	170851			

Total	1.32E+08	29

ANOVA two factor with Replication:

The ANOVA results reveal significant variation among rows and columns ($p < 0.001$), with degrees of freedom (df) values of 10 respectively. The error term has a degree of freedom of 0

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	841600745	10	4160074	65535	#NUM!	#NUM!
Columns	0	0	65535	65535	#NUM!	#NUM!
Error	0	0	65535			
Total	41600745	10				

ANOVA two factor without Replication:

Summary	Count	Sum	Average	Variance		
4	*1*	*7800*	*7800*	*#DIV/0!*		

5	*1*	*3000*	*3000*	*#DIV/0!*		
4	1	2300	2300	*#DIV/0!*		
3	1	7000	7000	*#DIV/0!*		
3	1	1200	1200	*#DIV/0!*		
4	1	2506.667	2506.667	*#DIV/0!*		
5	1	2618.095	2618.095	*#DIV/0!*		
6	1	2729.524	2729.524	*#DIV/0!*		
7	1	2840.952	2840.952	*#DIV/0!*		
6	1	4500	4500	*#DIV/0!*		
7	1	3063.81	3063.81	*#DIV/0!*		
1000		39559.05	3596.277	4160074		

Descriptive Statistics:

Column1	
Mean	1000
Standard Error	0
Median	1000
Mode	#N/A
Standard Deviation	#DIV/0!
Sample Variance	#DIV/0!
Kurtosis	#DIV/0!
Skewness	#DIV/0!
Range	0
Minimum	1000
Maximum	1000
Sum	1000
Count	1

SALES DATA SAMPLE REPORT

Introduction:

In the realm of business analytics, a dataset encompassing sales transactions emerges as a vital asset for deriving actionable insights. With columns detailing ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and more, it offers a comprehensive view of sales dynamics. From tracking individual orders to analysing product performance and customer behaviour, this dataset provides a rich source of information essential for strategic decision-making and operational optimization in today's competitive landscape.

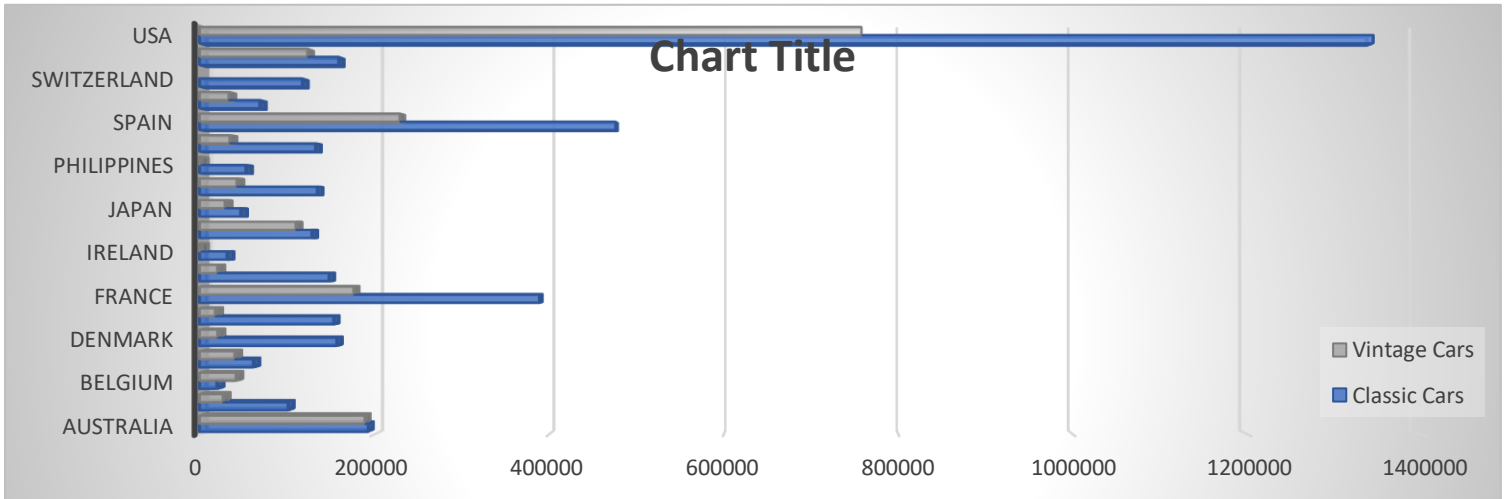
Questionaries:

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.

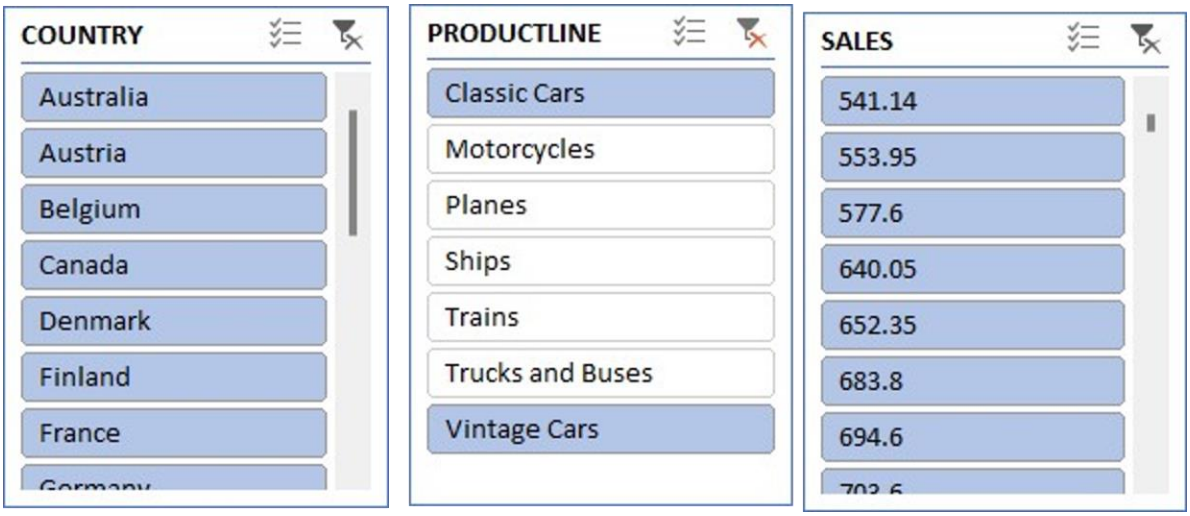
Analytics:

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

Ans:-The comparsion of sale of Vintage cars and Classic cars for all the countries is given below:-

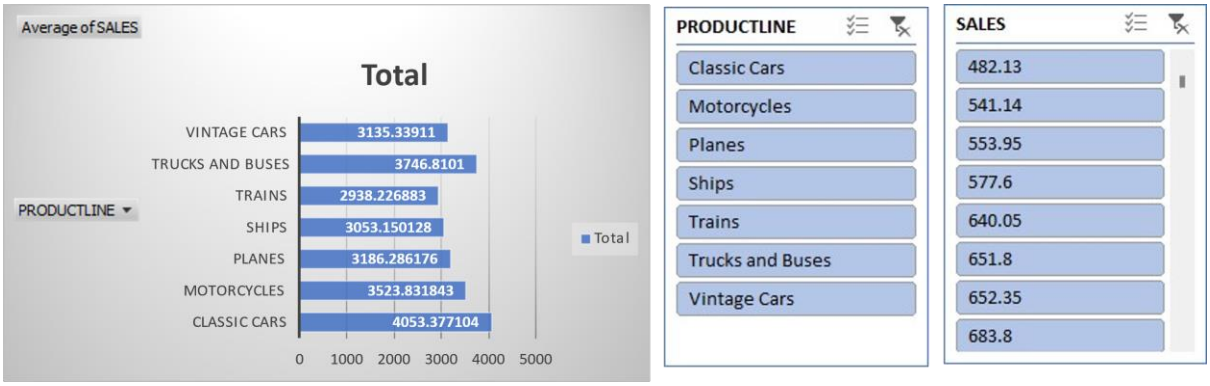


	Australia	Austria	Belgium	Canada	Denmark	Finland	France	Germany	Ireland	Italy	Japan	Norway	Philippines	Singapore	Spain	Sweden	Switzerland	UK	USA
Vintage Cars	189555	27197	41926	40513	21106	18383	176610	20936	2234.4	110451	29450	43021	1935.1	34960	229515	33804		123799	757756
Classic Cars	193086	101459	20137	61623	157182	153552	388951	148315	31689	128577	47271	134787	53112	132890	476165	69088	117714	159378	1E+06



Ans:-The USA leads in sales of Vintage and Classic cars, followed closely by Spain.

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



Ans:-Average Sales of Classic Cars = 4053.3, Motorcycles = 3523, Planes = 3186, Ships = 3053, Trains = 2938.22, Trucks and Buses = 3746.8, Vintage Cars = 3135. Classic Cars had the most sales.

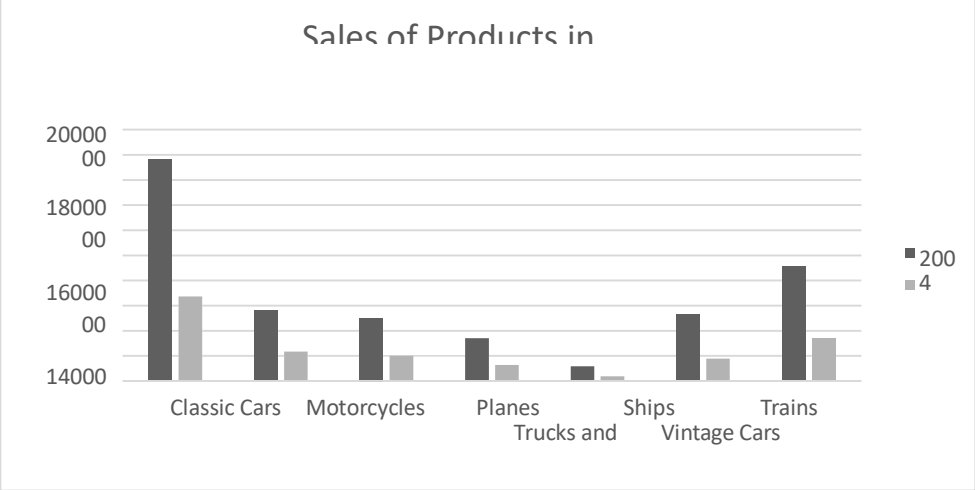
Q.3. Which country yields most of the profit for Motorcycles, Trucks and buses?

The country Australia yields most of the profit for Motorcycles, Trucks and buses

Ans:-The USA garnered the highest profit, totalling $147,340.15 (44%), from the sales of Motorcycles, Trucks and Buses

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

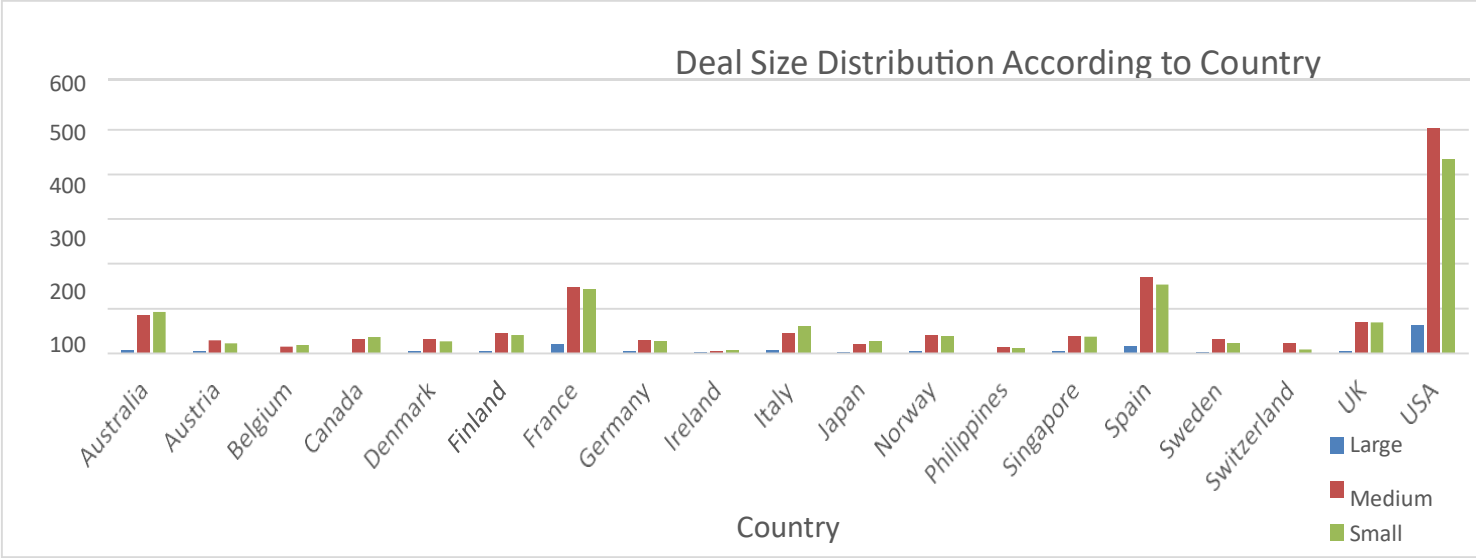
The following is the sales of all the items for the years of 2004, 2005 and as graph represents the sales has grown down from 20024 to 2005.



Ans:-In both 2004 and 2005, Classic Cars topped sales, while Trains saw the least. Motorcycle sales were 560,545 and 234,947 in 2004 and 2005 respectively, with Planes at 502,671 and 200,074, and Ships at 341,437 and 128,178. Trains had 116,523 and 36,917 sales, Trucks and Buses sold 529,302 and 178,057, and Vintage Cars reached 911,423 and 340,739. Classic Cars hit 1,762,257 sales in 2004 and 672,573 in 2005.

Q.5. Compare all the countries based on deal size.

The comparison of all the countries based on deal size are:



Ans:- The USA led in all three deal categories, while Switzerland had the fewest deals, with zero large deals.

Conclusion and Review:

In conclusion, the analysis of the provided sales dataset offers a window into the intricacies of business operations, shedding light on customer preferences, product performance, and market trends. By leveraging the insights gleaned from this dataset, businesses can make informed decisions, streamline processes, and drive growth. As the landscape of data analytics continues to evolve, harnessing the power of such datasets remains instrumental in staying competitive and responsive to the ever-changing demands of the market.