# 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:

Love Kalra

0827CS211137

# Exploring Car Data Report

## Introduction:

### Dataset Overview:

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.

### Key Attributes:

1. Make: This attribute denotes the brand or manufacturer of the vehicle, offering insights into brand preferences and market share.
2. Model: The specific model of the car, providing granularity in understanding consumer choices and preferences within each brand.
3. Color: Reflects the color of the vehicle, which can influence consumer perception and aesthetic preferences.
4. Mileage: Indicates the distance traveled by the vehicle, a crucial factor influencing its value and pricing.
5. Price: Represents the listed price of the vehicle, serving as a key determinant in consumer purchasing decisions and market competitiveness.
6. Cost: Denotes the cost associated with acquiring the vehicle, which includes factors such as production costs, dealer margins, and other expenses.

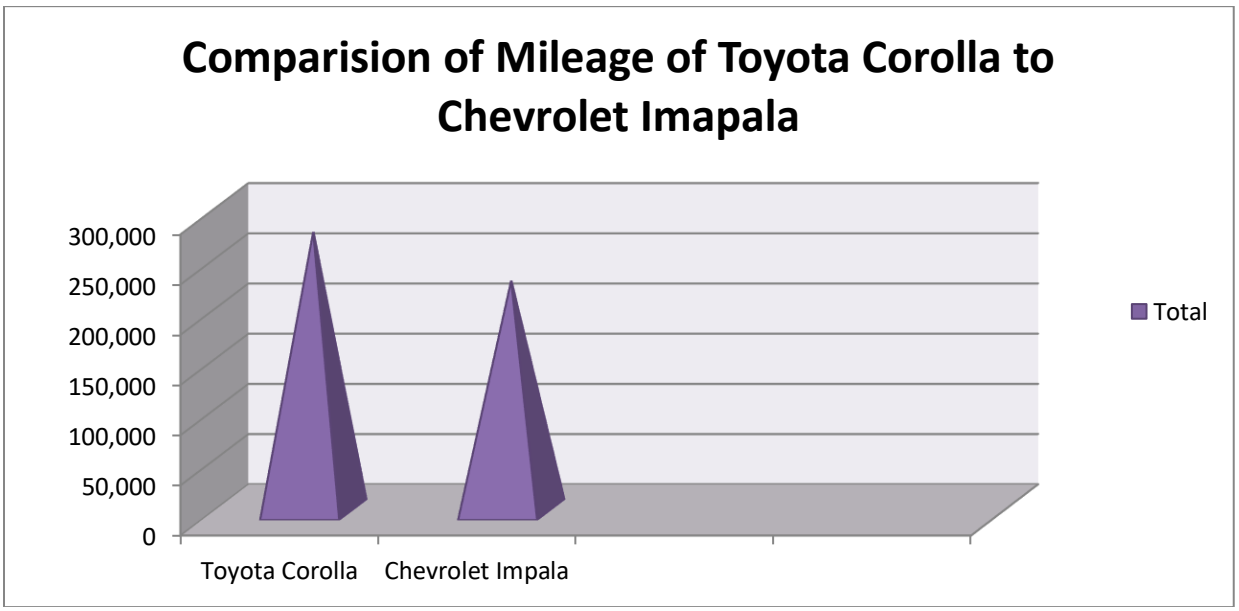
## Questionnaire:

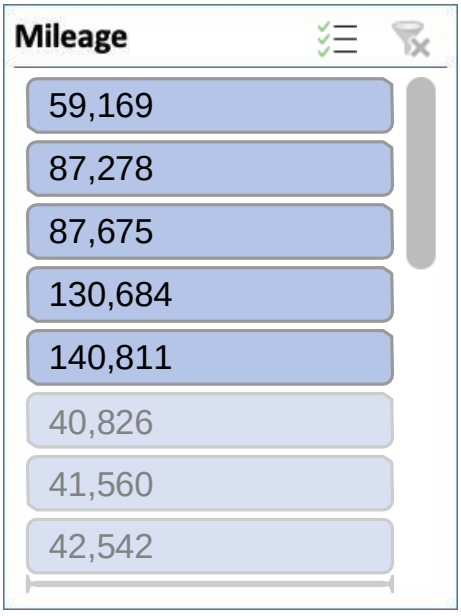
Q1. Compare the mileage of Chevrolet Impala to Toyota Corolla. Which of the two is giving best mileage?
Q2. Justify, Buying of any Ford car is better than Honda
Q3. Among all the cars which car color is the most popular and is least popular?
Q4. Compare all the cars which are of silver color to the green color in terms of Mileage.
Q5. Find out all the cars, and their total cost which is more than $2000?

# Analytics:

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

Ans. Toyota Corolla gives better mileage than Chevrolet Impala.

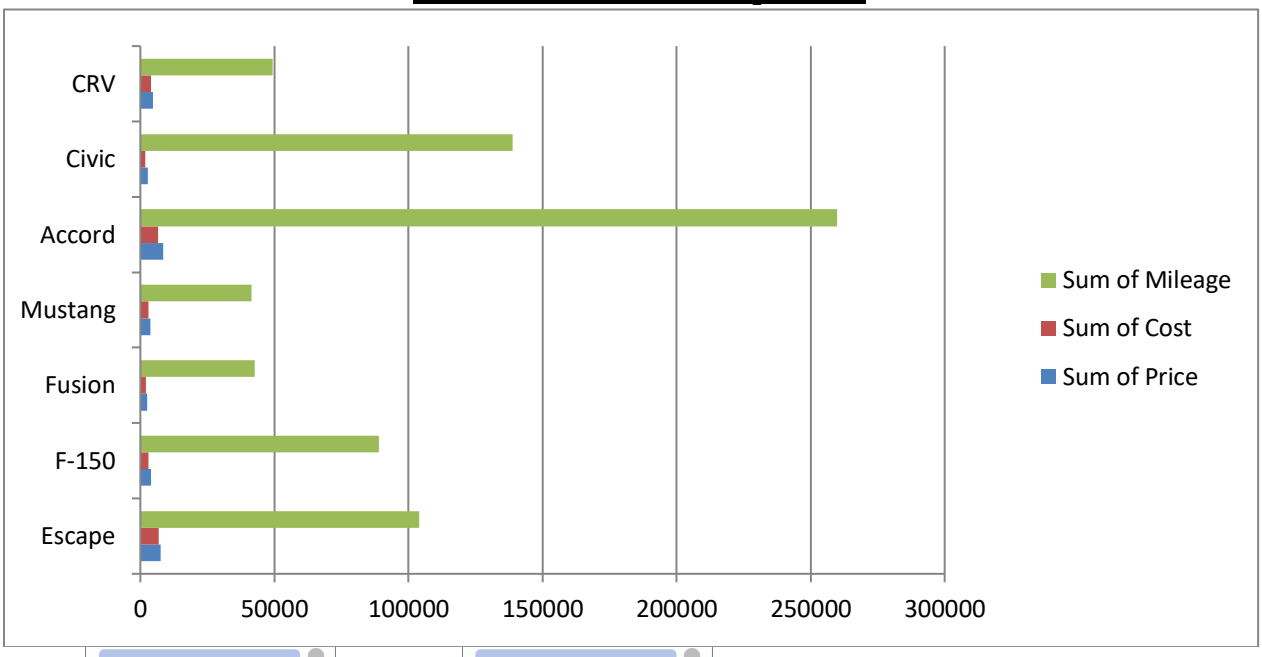


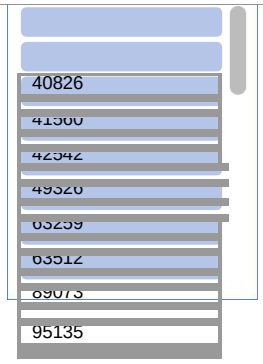



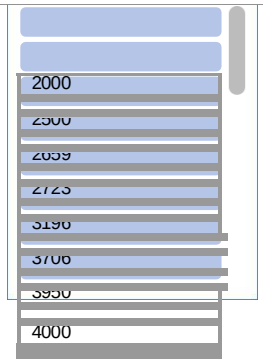
Q2.Justify buying of any Ford car is better than Honda.

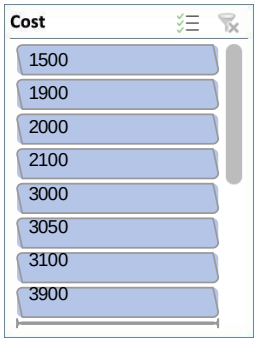
Ans. Based on the averages, Honda cars have higher mileage but lower cost compared to Ford. Therefore, the choice depends on whether the buyer values mileage or cost but if we compare on mileage ford car has low mileage and cost so Buying ford car is better than Honda.

**Ford vs Honda Car Comparision**



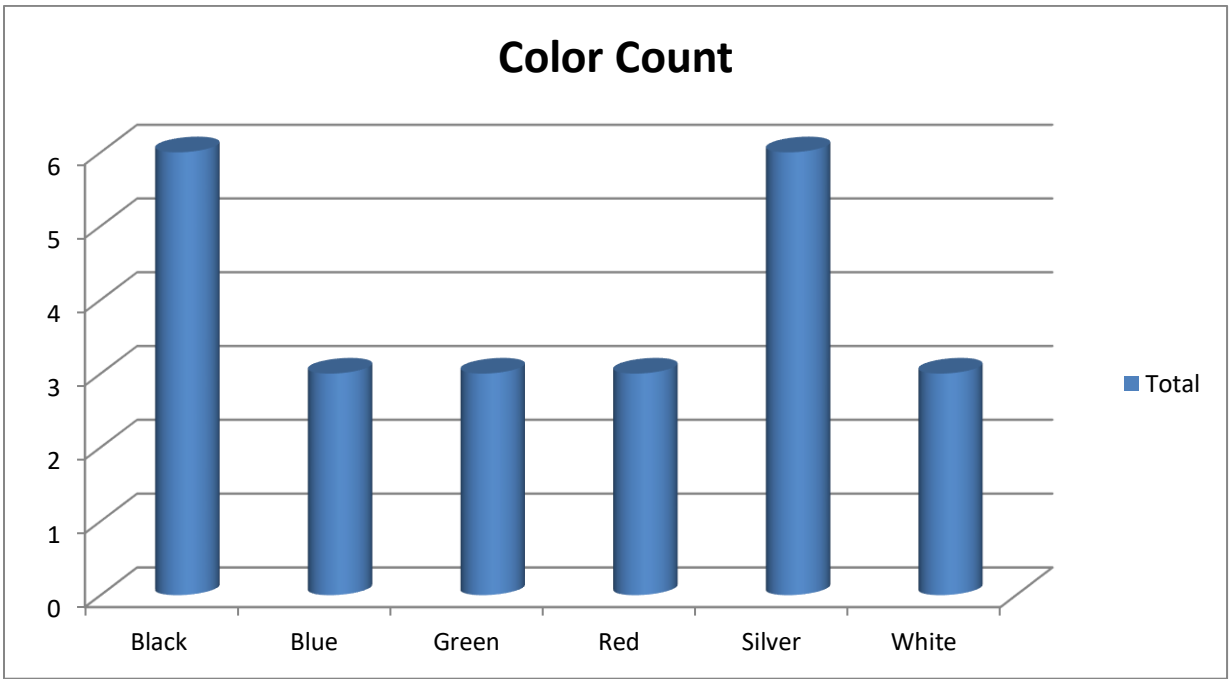



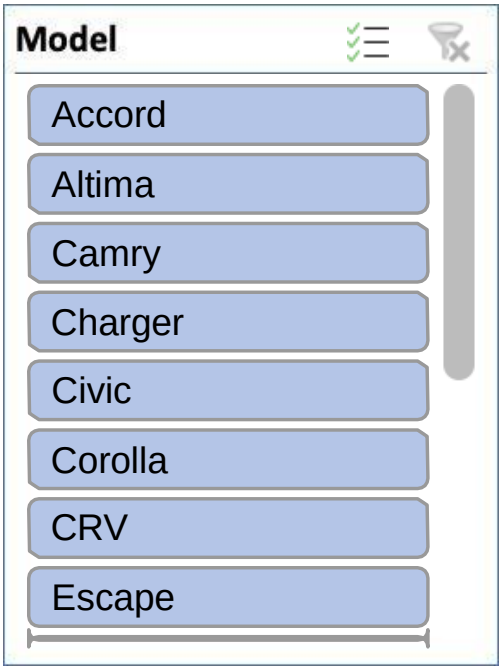



Q3.Among all the cars which car color are the most popular and are least popular?

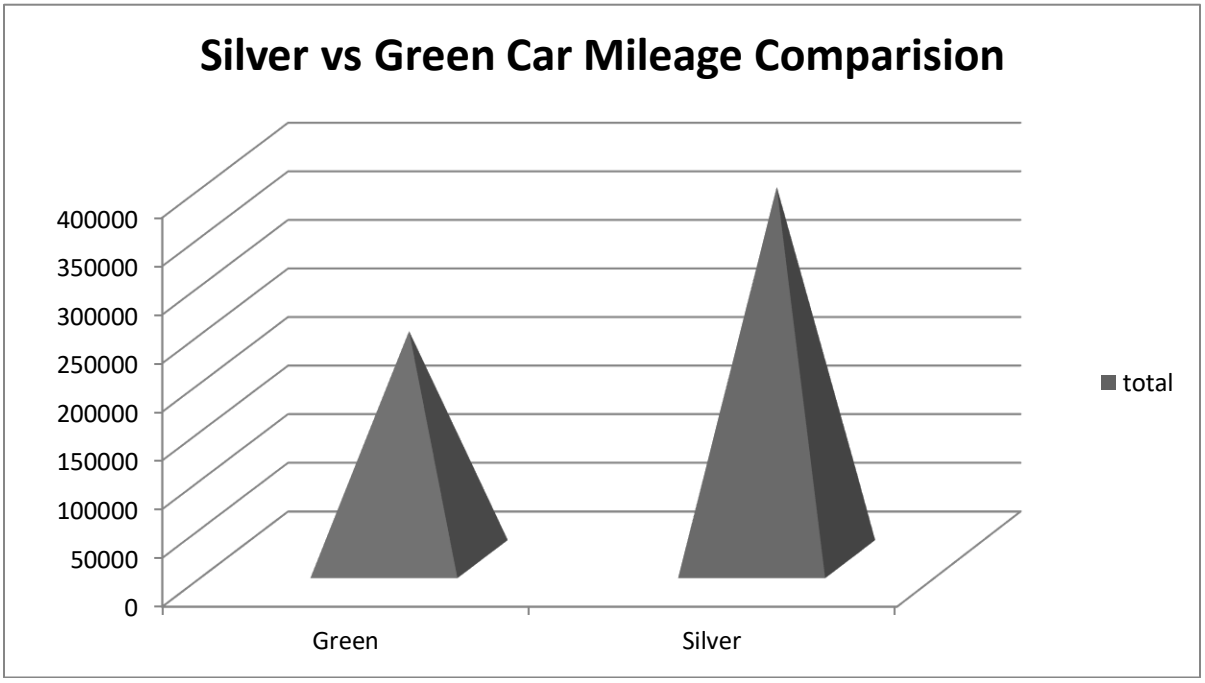
Ans. Most popular color is Silver and Black as each appear 6 times and least appearing Colors are Blue, Green, Red, and White they all appear 3 times.

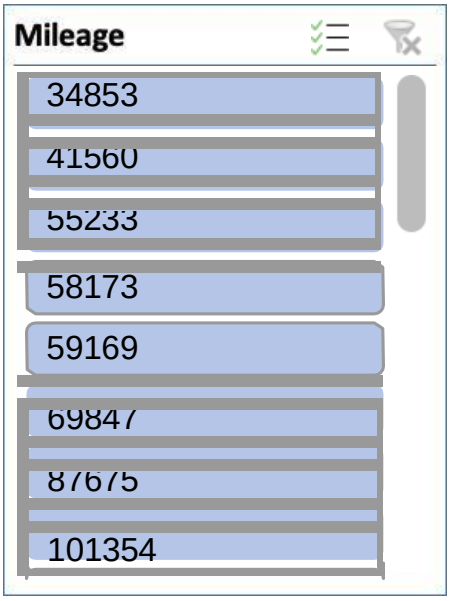



Q4. Compare all the cars which are of silver color to the green color in terms of Mileage.

Ans. Silver color car mileage is more than green color car mileage if we compare their average.

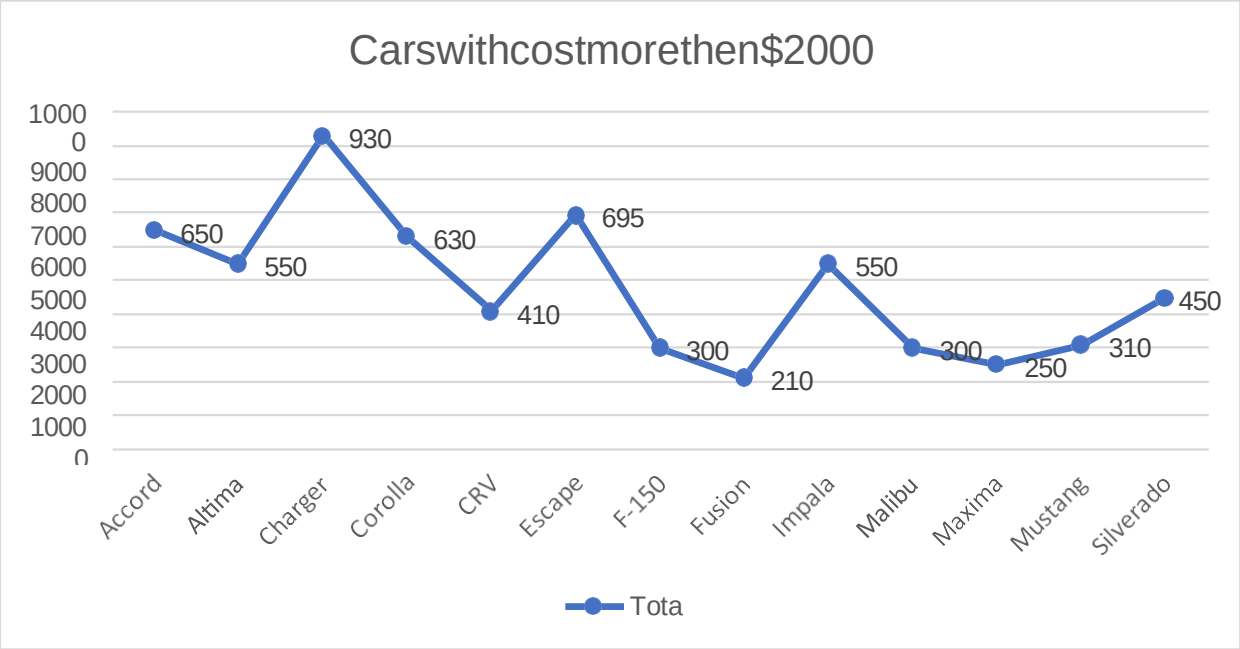


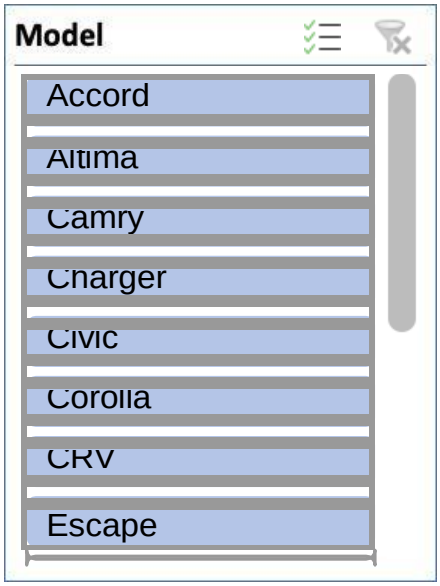



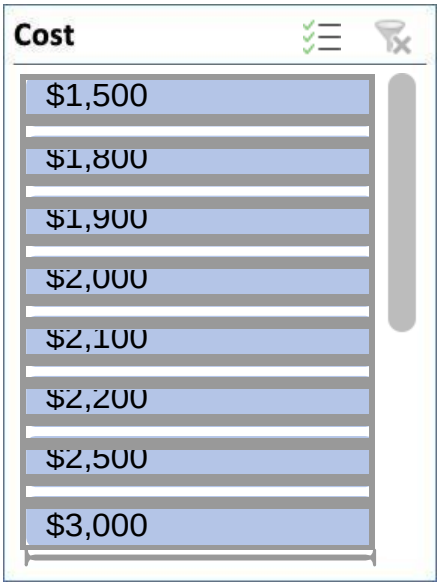

Q5. Find out all the cars, and their total cost which is more than $2000?

Ans. All the car mention below cost is more than $2000

Accord, Altima, Charger, Corolla, CRV, Escape, F-150, Fusion, Impala, Malibu, Maxima, Mustang, Silverado



# Regression

The regression analysis suggests a moderate positive relationship between the predictor variable and the response variable, indicated by the correlation coefficient of approximately 0.40. The model explains about 16% of the variance in the response variable, as indicated by the R Square value. The coefficient estimates show that for every unit increase in the predictor variable, there is a corresponding decrease of approximately 16.66 in the response variable, with a p-value of 0.056, indicating a marginally significant effect.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.40404555 |
| R Square | 0.1632528 |
| Adjusted R Square | 0.1234077 |
| Standard Error | 33099.5397 |
| Observations | 23 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 4488793099 | 4488793099 | 4.09718598 | 0.05586127 |
| Residual | 21 | 2.3007E+10 | 1095579531 | | |
| Total | 22 | 2.7496E+10 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 130438.919 | 23634.1932 | 5.51907645 | 1.7789E-05 | 81288.9236 | 179588.914 | 81288.9236 | 179588.914 |
| 3000 | -16.664135 | 8.23265547 | -2.0241507 | 0.05586127 | -33.784879 | 0.45660911 | -33.784879 | 0.45660911 |

# Co-relational

The correlation matrix indicates a moderate negative correlation (-0.411) between Mileage and Price. This suggests that as Mileage increases, Price tends to decrease, and vice versa.

| | | *Price* |
|---|---|---|
| Mileage | 1 | |
| Price | -0.4110586 | 1 |

# ANOVA: Single Factor

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

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

ANOVA

| *SourceofVariation* | *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 | | | | |

# ANOVA: Two-Factor Without replication

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.

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 | | | | |

# Descriptive Statistics

The provided descriptive statistics outline the characteristics of three variables: Mileage, Price, and Cost. Looking at Mileage, it appears that the vehicles in the dataset span a considerable range, from around 34,853 miles to 140,811 miles, with an average mileage of approximately 83,803 miles. Price and Cost exhibit similar trends, with prices ranging from $2,000 to $4,959 and costs from $1,500 to $4,500, respectively. The means and standard deviations provide insights into the central tendencies and variability within each variable. Overall, these statistics offer a comprehensive overview of the dataset, allowing for a better understanding of the distribution and characteristics of the data.

| *Mileage* | | *Price* | | *Cost* | |
|---|---|---|---|---|---|
| Mean | 83802.7917 | Mean | 3254.5 | Mean | 2756.25 |
| Standard Error | 7112.65205 | Standard Error | 186.751181 | Standard Error | 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 |

# Conclusion & Review

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.

# EXPLORING SALES OF DIFFERENT SEGMENT IN US STATES

## 1. 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:

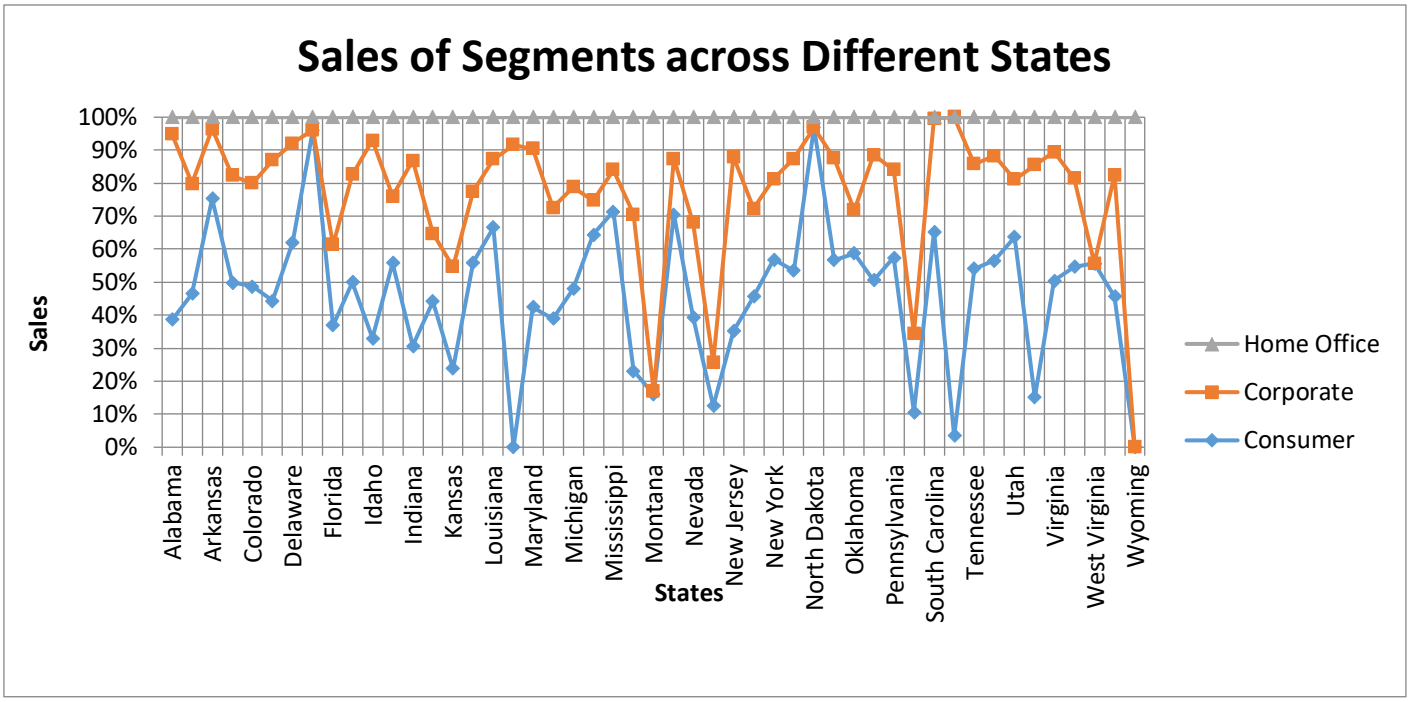
1. ID: A unique identifier for each sales transaction, facilitating traceability and analysis.
2. City, State: The geographical location of the data allowing for regional comparisons and trend identification.
3. Product Line (furniture, Electronic Accessories, appliances, Home and Lifestyle): Categorization 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.

# 1. ANALYTICS:

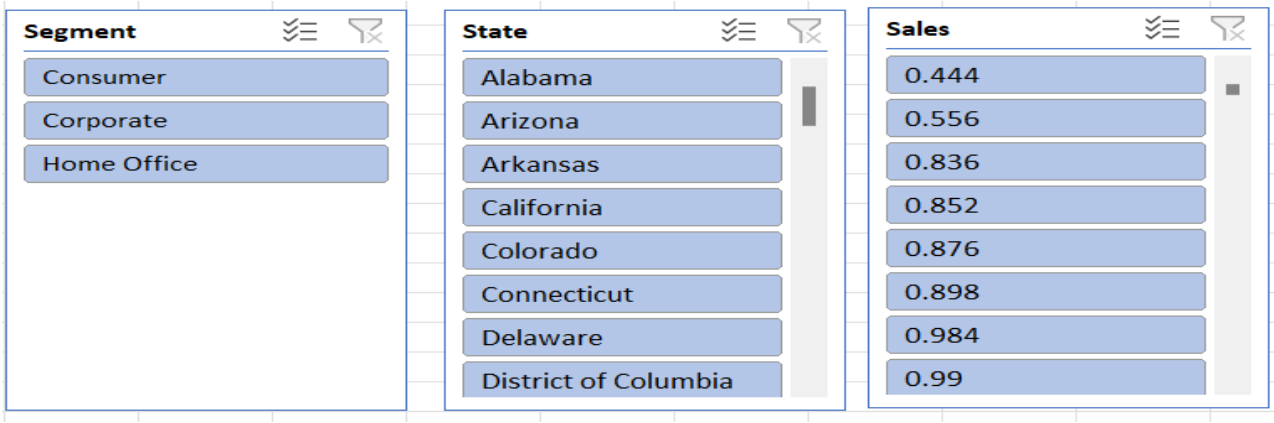
**Q1.** Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?



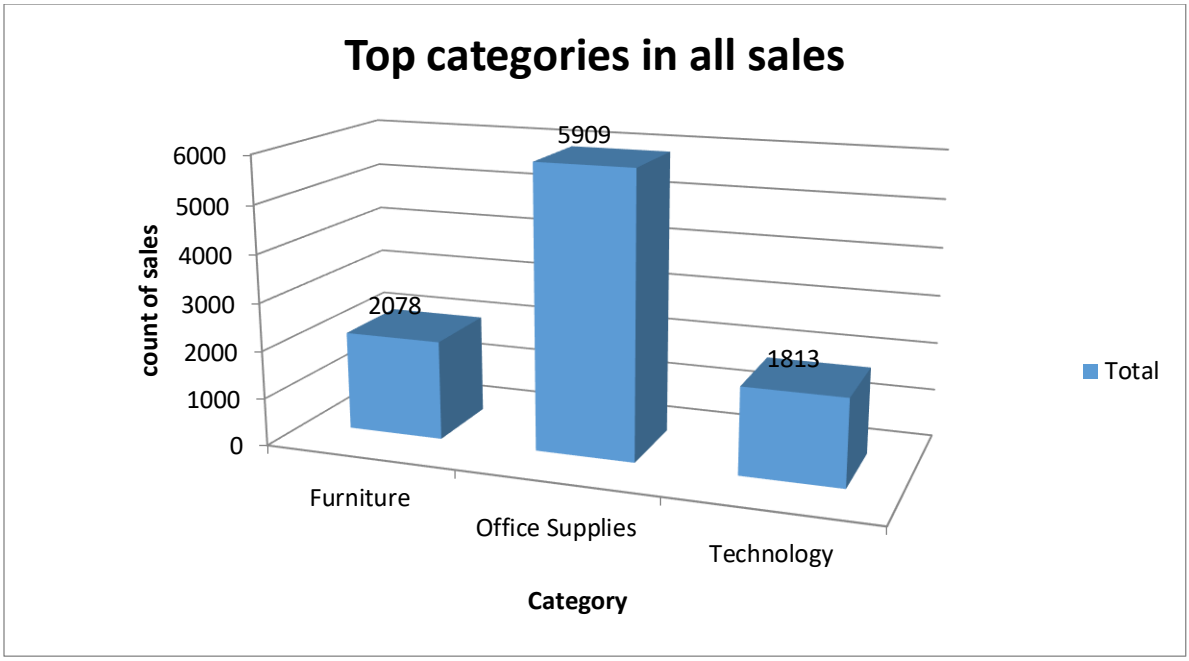

Ans. After comparing all the states in terms of segment and sales, California emerged as the state with the highest amount of sales

Consumer segment performed well in all the states

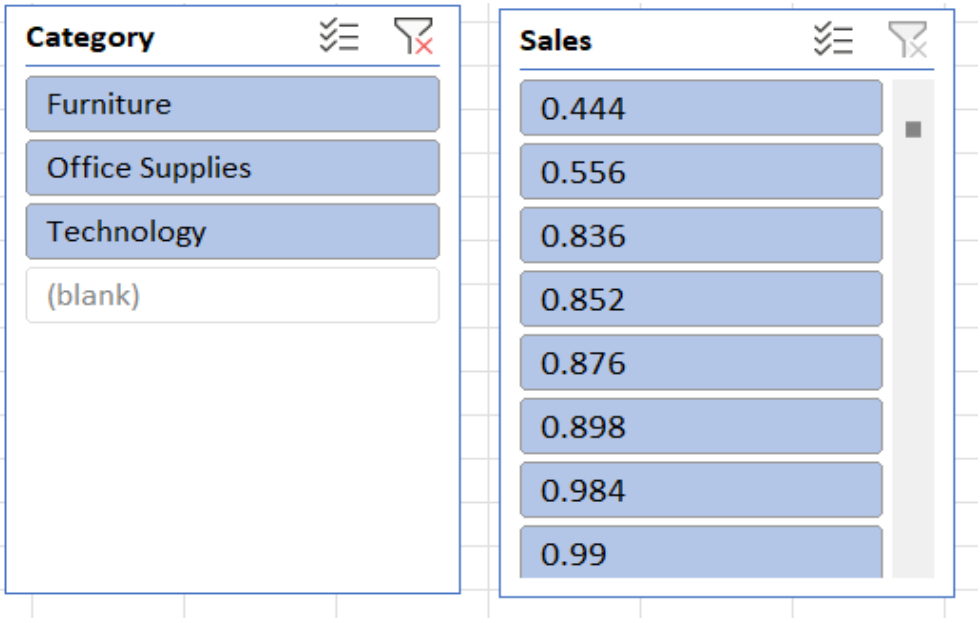
## Slicers:



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

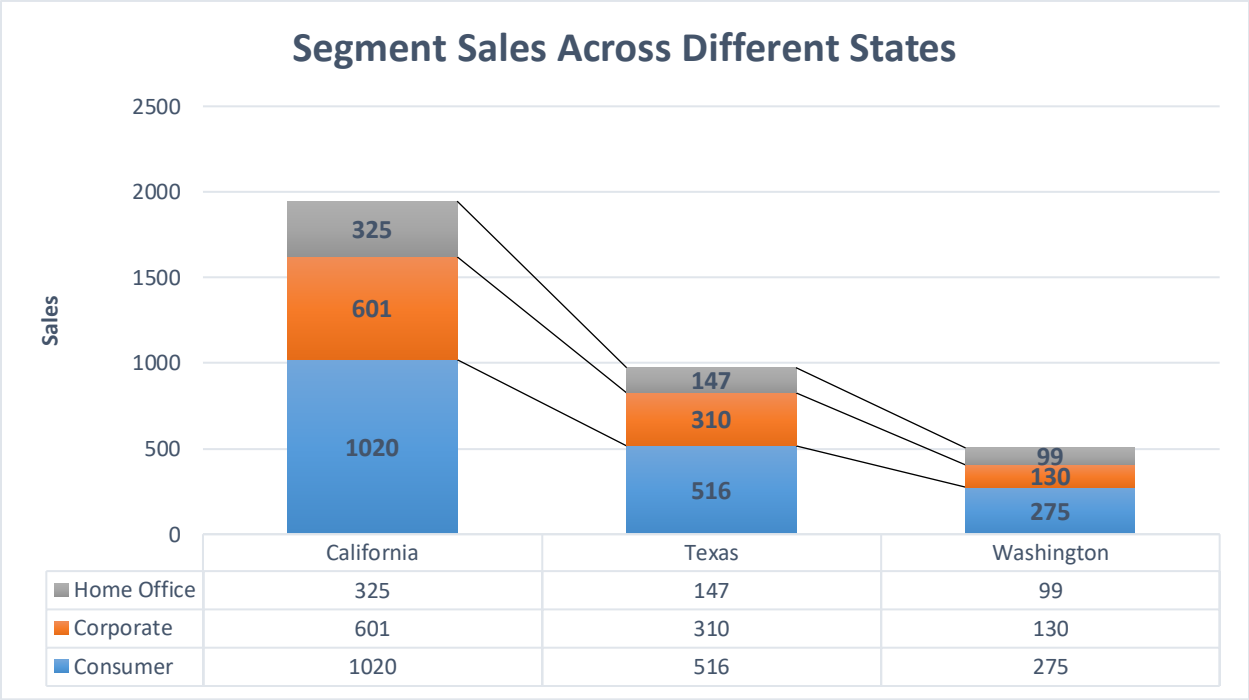



## Slicers:

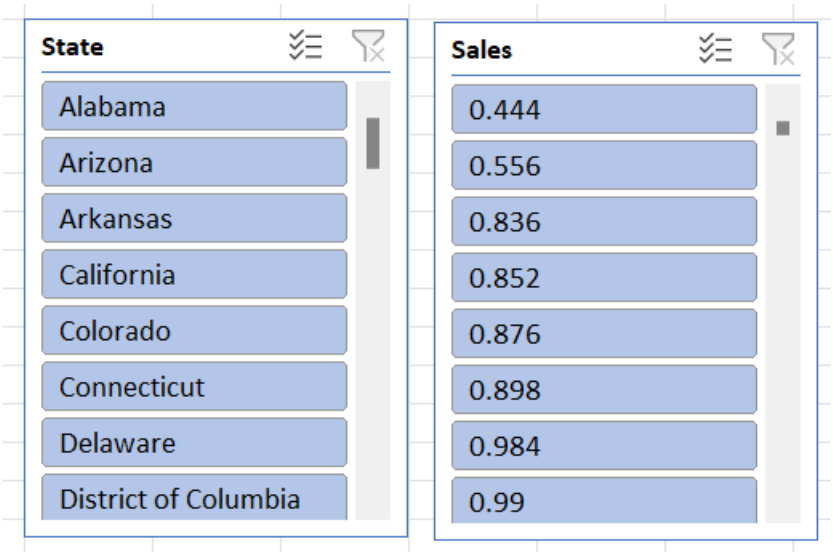



Ans. Office Supplies is the top performing category in all the states

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

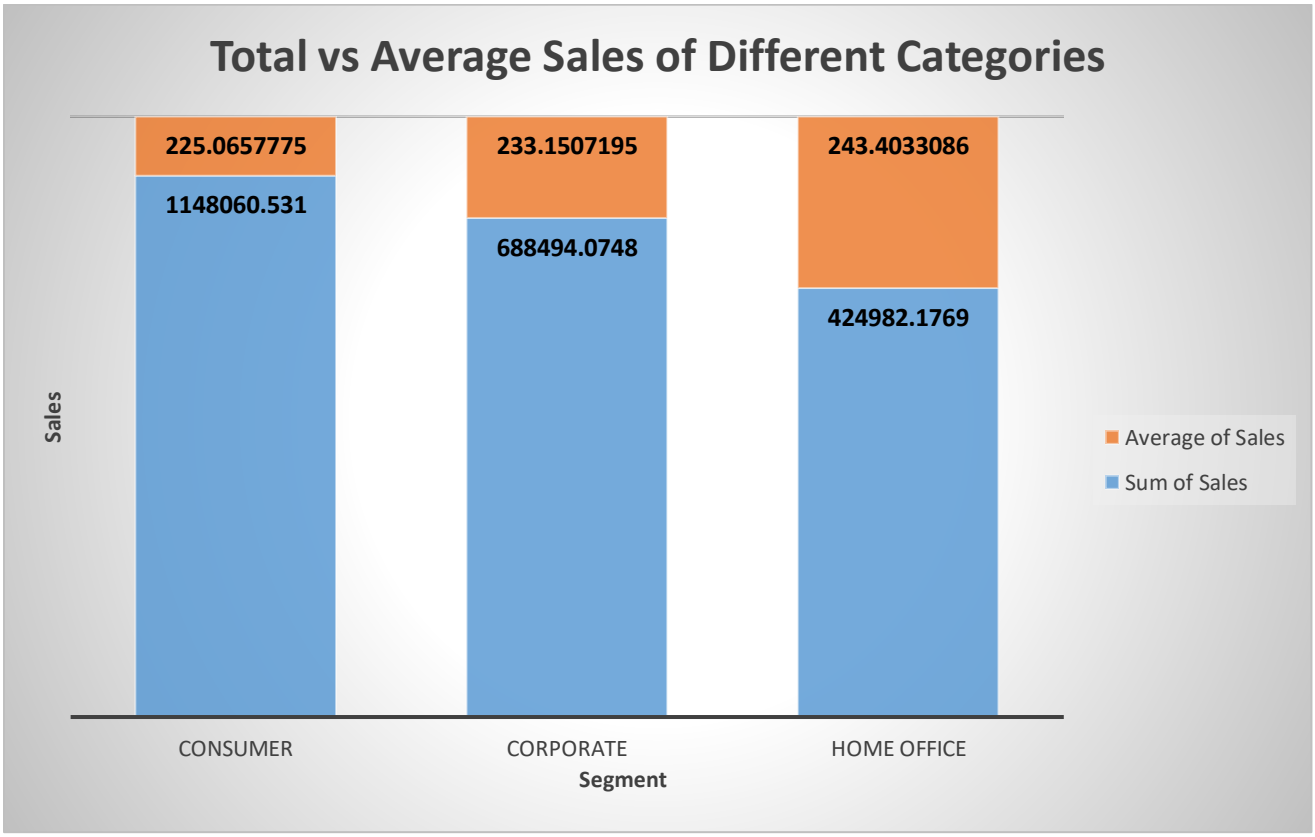



## Slicers:

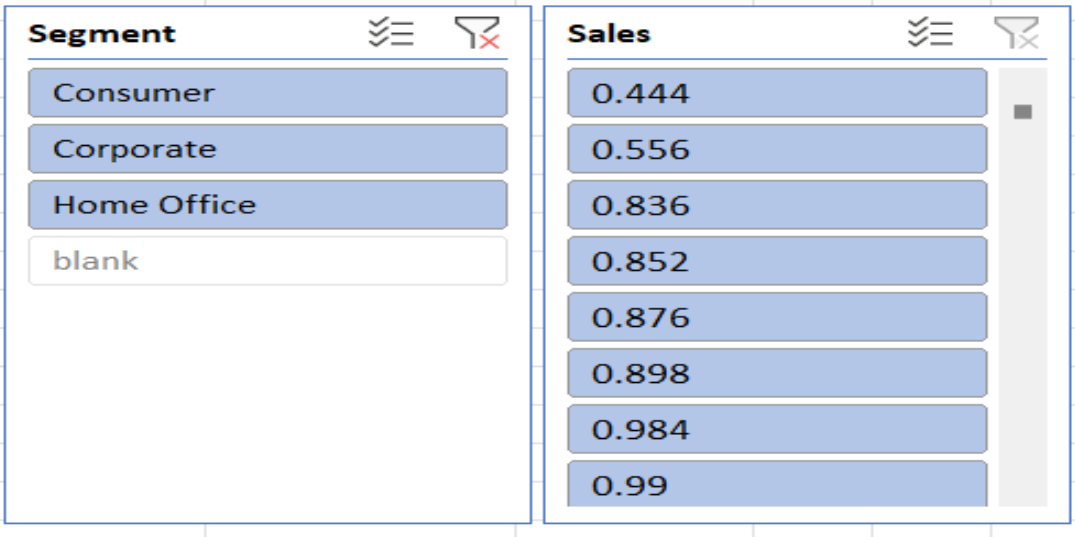



**Ans.** Consumer segment has the most sales in US, California, Texas, and Washington

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

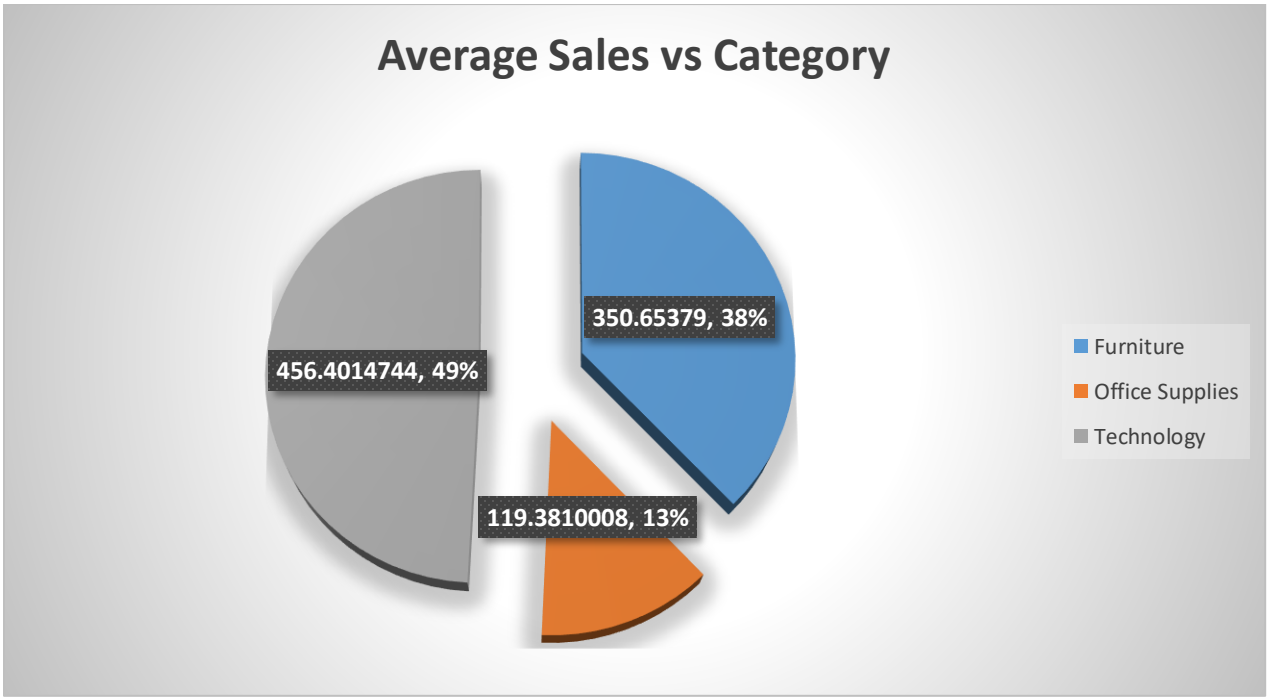


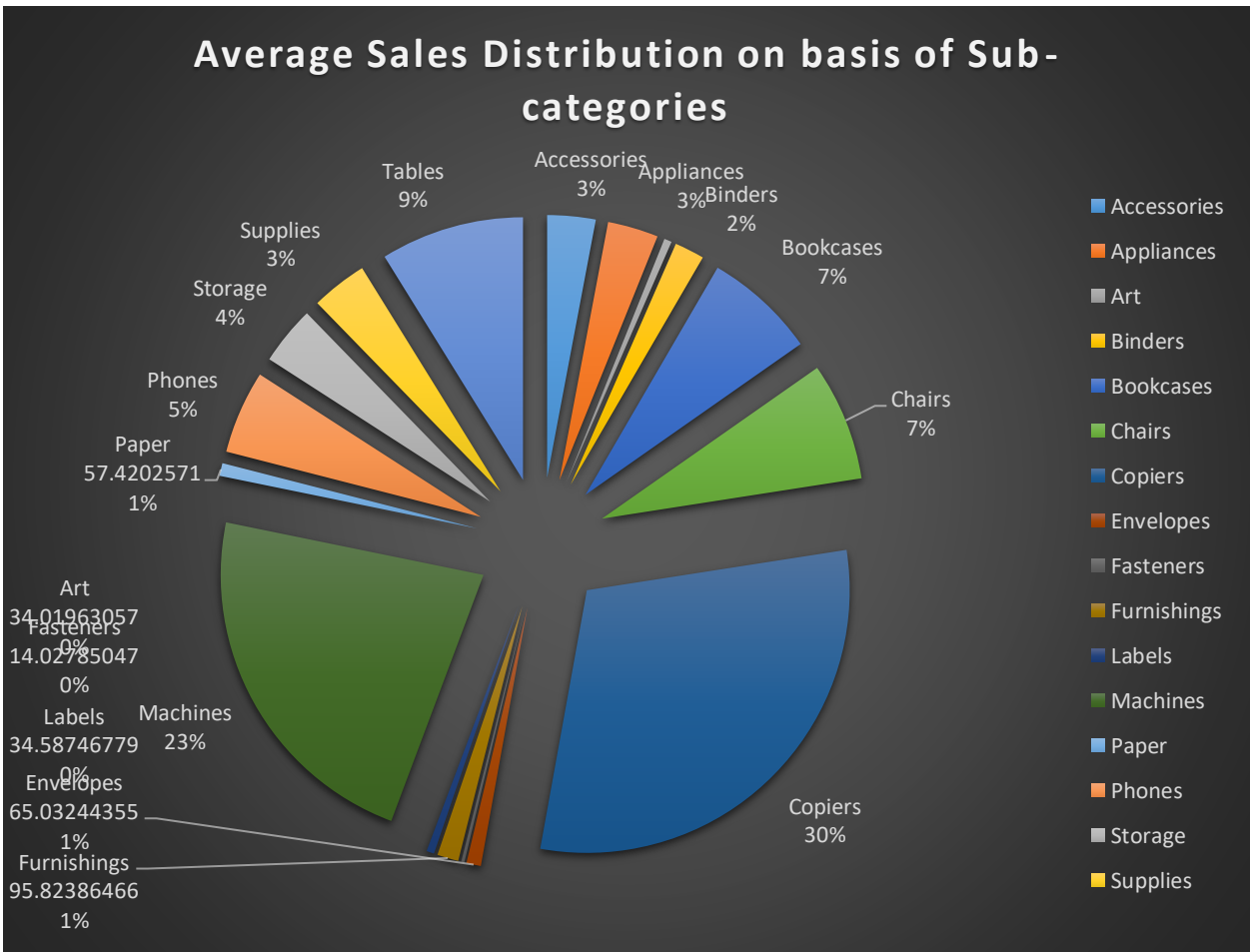

## Slicers:



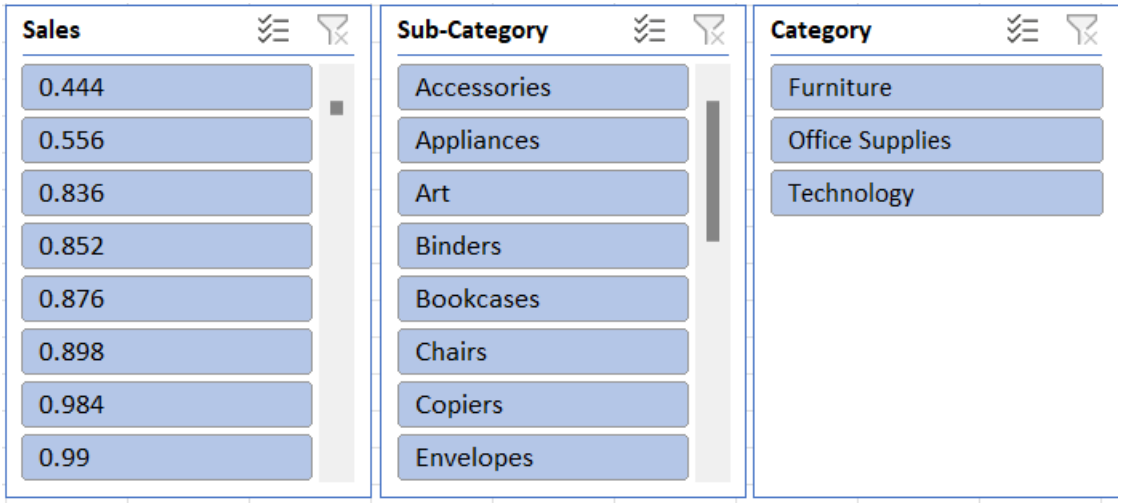
Ans. By Analysis, we can find out that the total sales were a lot higher than the average Sales for each segment.

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



## Slicers:




Ans. In Average Sales vs Category we can observe that the category Technology has the maximum sales.

And in Average Sales Distribution on basis of Sub-Categories, Copiers had the highest contribution of 30%.

**Q6.** Find out state wise mode for Customer and Segment.California, Illinois, New York, Texas, Waashington

Ans.

For California: Customer- William Brown, Segment-Consumer

For Illinois: -Customer- William Brown, Segment-Consumer

For New York: Customer- William Brown, Segment-Consumer

For Texas: Customer- William Brown, Segment-Consumer

For Washington: Customer- William Brown, Segment-Consumer

# Regression and ANOVA:

| SUMMARY OUTPUT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| *Regression Statistics* | | | | | | | | |
| Multiple R | 0.003989285 | | | | | | | |
| R Square | 1.59144E-05 | | | | | | | |
| Adjusted R Square | -0.000484829 | | | | | | | |
| Standard Error | 525.2842121 | | | | | | | |
| Observations | 1999 | | | | | | | |
| | | | | | | | | |
| ANOVA | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| Regression | 1 | 8769.275428 | 8769.275428 | 0.031781546 | 0.85852604 | | | |
| Residual | 1997 | 551019236.5 | 275923.5035 | | | | | |
| Total | 1998 | 551028005.8 | | | | | | |
| | | | | | | | | |
| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
| Intercept | 232.5093842 | 22.96786412 | 10.12324799 | 1.59291E-23 | 187.4658976 | 277.5528709 | 187.4658976 | 277.5528709 |
| Postal Code | -6.54575E-05 | 0.000367174 | -0.178273794 | 0.85852604 | -0.000785542 | 0.000654627 | -0.000785542 | 0.000654627 |

This regression output provides information about the relationship between the predictor variable (Postal Code) and the response variable (which is not explicitly mentioned in the output). Let's break down each section of the output:

**Regression Statistics:**
- **Multiple R:** This is the correlation coefficient, which measures the strength and direction of the linear relationship between the predictor and response variables. In this case, it's very close to zero (0.00399), indicating a very weak linear relationship.
- **R Square:** This is the coefficient of determination, which represents the proportion of the variance in the response variable that is explained by the predictor variable. A value close to zero (1.59144E-05) indicates that the predictor variable explains very little of the variance in the response variable.

Adjusted R Square: This is a modified version of R Square that adjusts for the number of predictor variables in the model. A negative value (-0.000484829) suggests that the model may be overfitting or that the predictor variable is not adding any explanatory power to the model.
Standard Error: This represents the average deviation of the observed values from the regression line. A higher standard error (525.2842121) indicates greater variability in the data points around the regression line.
Observations: This indicates the number of data points used in the regression analysis.

## ANOVA (Analysis of Variance):

df: Degrees of freedom represent the number of independent pieces of information available in the data.
SS: Sum of squares measures the total variation in the response variable.
MS: Mean square is the average variation within groups or between groups.
F: The F-statistic tests the overall significance of the regression model. A low F-value relative to the critical value indicates that the model is not significant.
Significance F: This is the p-value associated with the F-statistic. A high p-value (0.85852604) suggests that the model is not statistically significant.

## Coefficients:

Intercept: This is the y-intercept of the regression line, representing the predicted value of the response variable when the predictor variable is zero. The coefficient (232.5093842) indicates the average value of the response variable when the predictor variable is zero.
Postal Code: This is the coefficient for the predictor variable. It represents the change in the response variable for a one-unit change in the predictor variable. The coefficient (-6.54575E-05) suggests a very small negative effect of the Postal Code on the response variable, but it is not statistically significant given the high p-value (0.85852604).

Overall, based on this regression output, the model does not appear to provide a meaningful explanation of the variation in the response variable, and the predictor variable (Postal Code) does not appear to have a significant effect on the response variable.

## 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 AND REVIEWS:

## Conclusion:

In delving into the sales data across various segments in different US states, our analysis has unearthed valuable insights. The dataset provided a comprehensive view, encompassing crucial variables such as geographical location, product categorization, transactional details, and performance metrics. Through meticulous scrutiny, we addressed pertinent questions and gleaned actionable conclusions.

California emerged as a focal point, exhibiting the highest sales volume among the states analyzed. Notably, the consumer segment showcased consistent performance across all states, underscoring its significance in the market landscape.

Across categories, office supplies emerged as the top performer in all states, indicating a universal demand for these products. Furthermore, the consumer segment demonstrated dominance in terms of sales across the US, including California, Texas, and Washington.

Analysis of total versus average sales per segment revealed significant disparities, with total sales outweighing average sales across the board. This highlights the presence of outlier transactions or high-value sales within each segment.

Delving deeper into category and sub-category analysis, technology emerged as the category with the highest average sales, suggesting a strong market demand for technological products. Subsequently, copiers emerged as the top contributor to average sales distribution, emphasizing their importance within the technology category.

## Reviews:

1. Thorough Analysis with Actionable Insights: The exploration of sales data across US states provides a comprehensive understanding of market dynamics. The inclusion of key attributes and performance metrics ensures a robust analysis, enabling stakeholders to derive actionable insights for strategic decision-making.

2. Clear Presentation of Findings: The presentation of findings is concise and structured, facilitating easy comprehension of complex data. The use of slicers enhances the visual representation, aiding in the interpretation of results and facilitating informed decision-making.

3. Insightful Conclusions: The conclusion succinctly summarizes key findings and draws meaningful conclusions from the analysis. By highlighting overarching trends and significant observations, it provides valuable guidance for market strategies and future research endeavors.

Overall, the exploration of sales data offers valuable insights into market trends and consumer behavior, serving as a foundation for informed business strategies and market interventions.

# Cookie Data Analysis

## Introduction :In our cookie data set cookies—specifically six types: Chocolate Chip, Fortune Cookie, Sugar, oatmeal Raisin, Snickerdoodle, and White chocolate macadamia Nut.

We've got a treasure trove of data on these cookies, covering how many units were sold, their costs, the money they brought in (revenue), and the profits they made. And we're not just looking at one place or time; we're exploring different countries and dates to see how things vary.

This report isn't just about cookies; it's about understanding what people like, how much they're willing to pay, and where these treats are most popular. So, get ready to uncover some fascinating insights into the cookie world and what it means for businesses like yours.

## Questionaries :

1 . Compare Malaysia and Philippines on the bases of two types of Cookies

2. What is the performance of Choco Chips Cookies in all Country Which Competes the best.

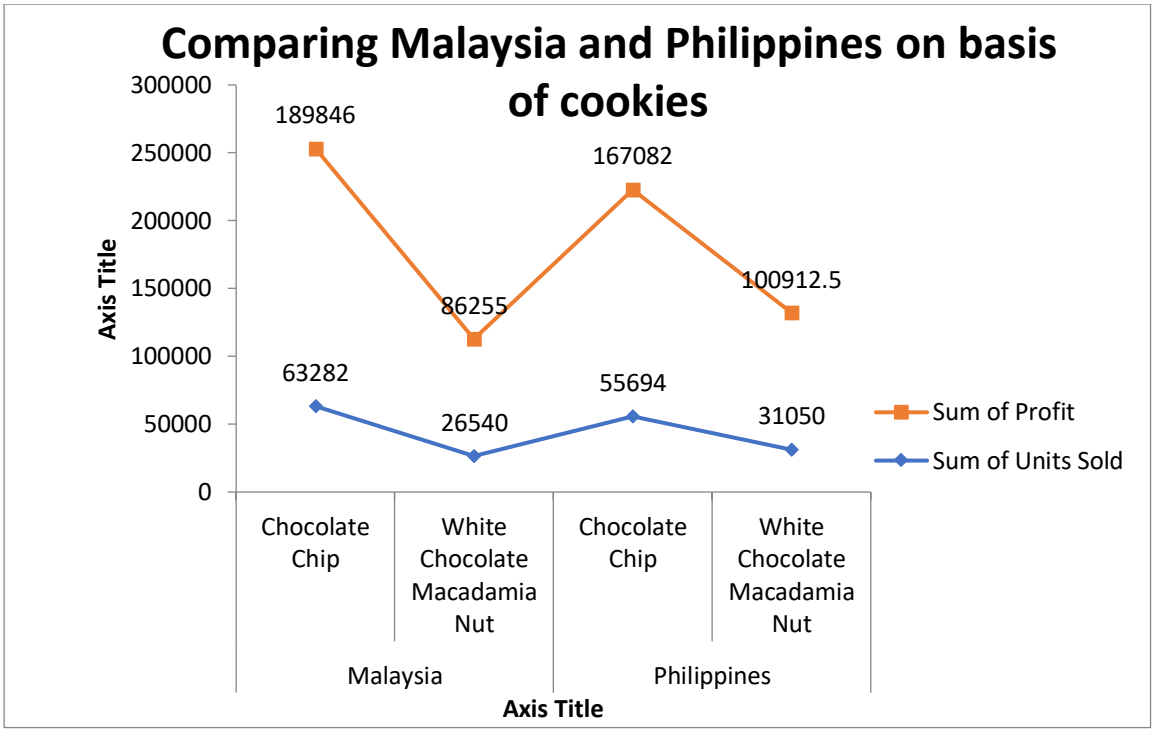
3. Compare all the countries on the bases of profit and unit sold, which is the best performance country on the basis of profit.

4. which Cookie is the best Selling Cookie in India and US in year 2019,

## Analytics :

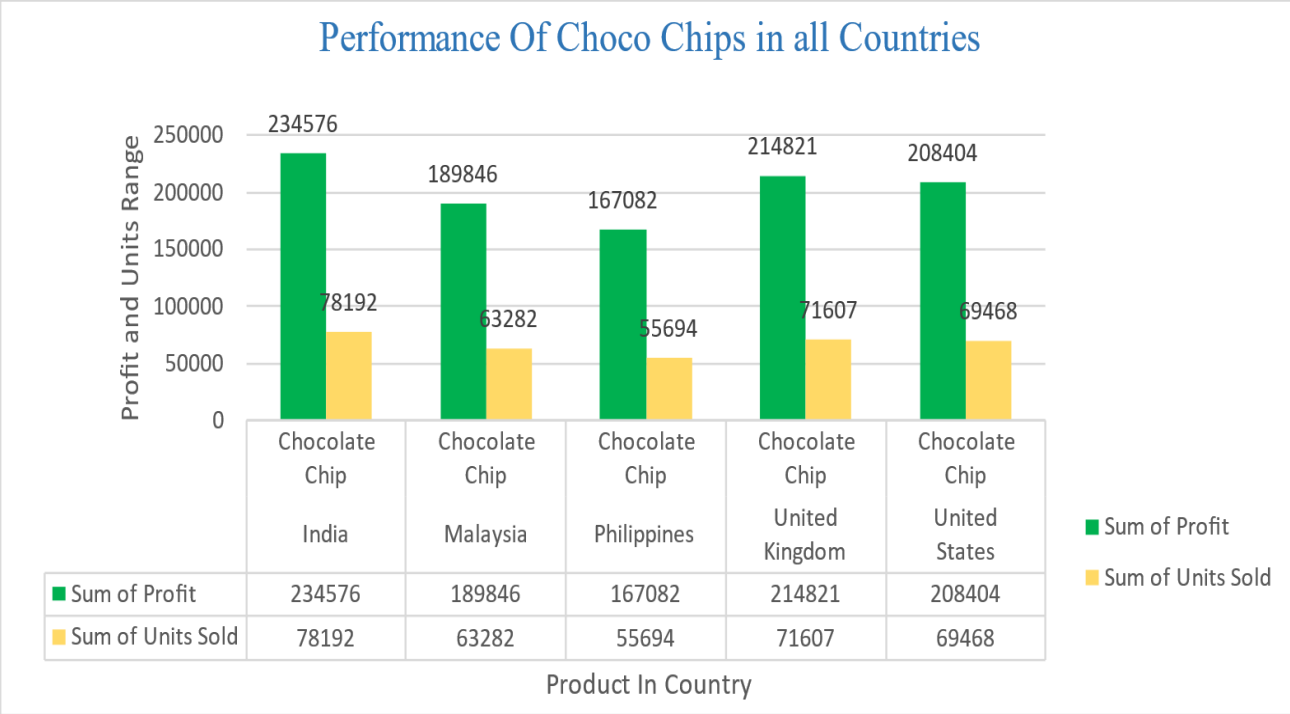
1 . Compare Malaysia and Philippines on the bases of two types of Cookies.

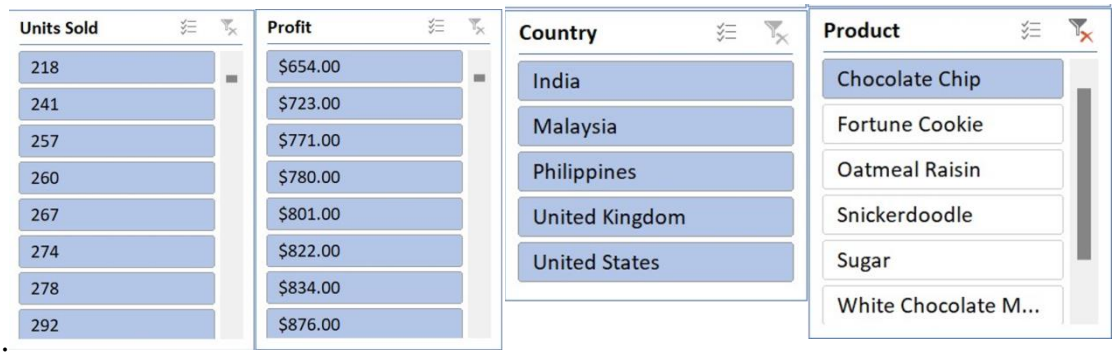
Ans:-The comparsion of Malaysia and Philippines on bases of Chocolate chip and White Chocolate Macadmia nut is given below:-




2. What is the performance of Choco Chips Cookies in all Country Which Competes the best.

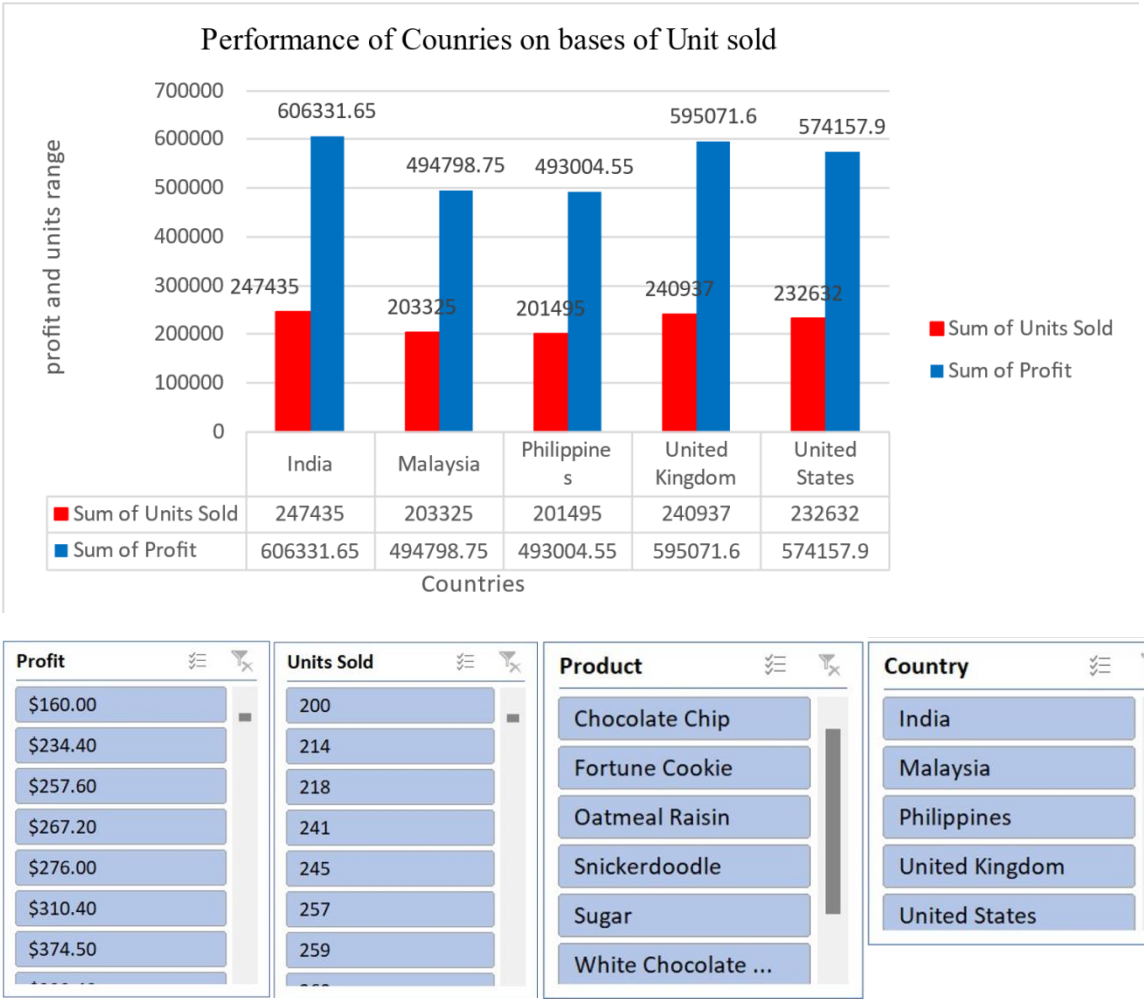
Ans:- India stands out as the foremost consumer of Choco chips worldwide, primarily due to its exceptional profitability and record-breaking sales figures. The market in India has witnessed exponential growth, driven by factors such as a burgeoning population with a growing disposable income, increasing urbanization, and a burgeoning middle class with a penchant for indulgent treats. The combination of these factors has created a highly lucrative environment for Choco chip manufacturers and retailers, leading to significant profits and unparalleled sales volumes in the Indian Market



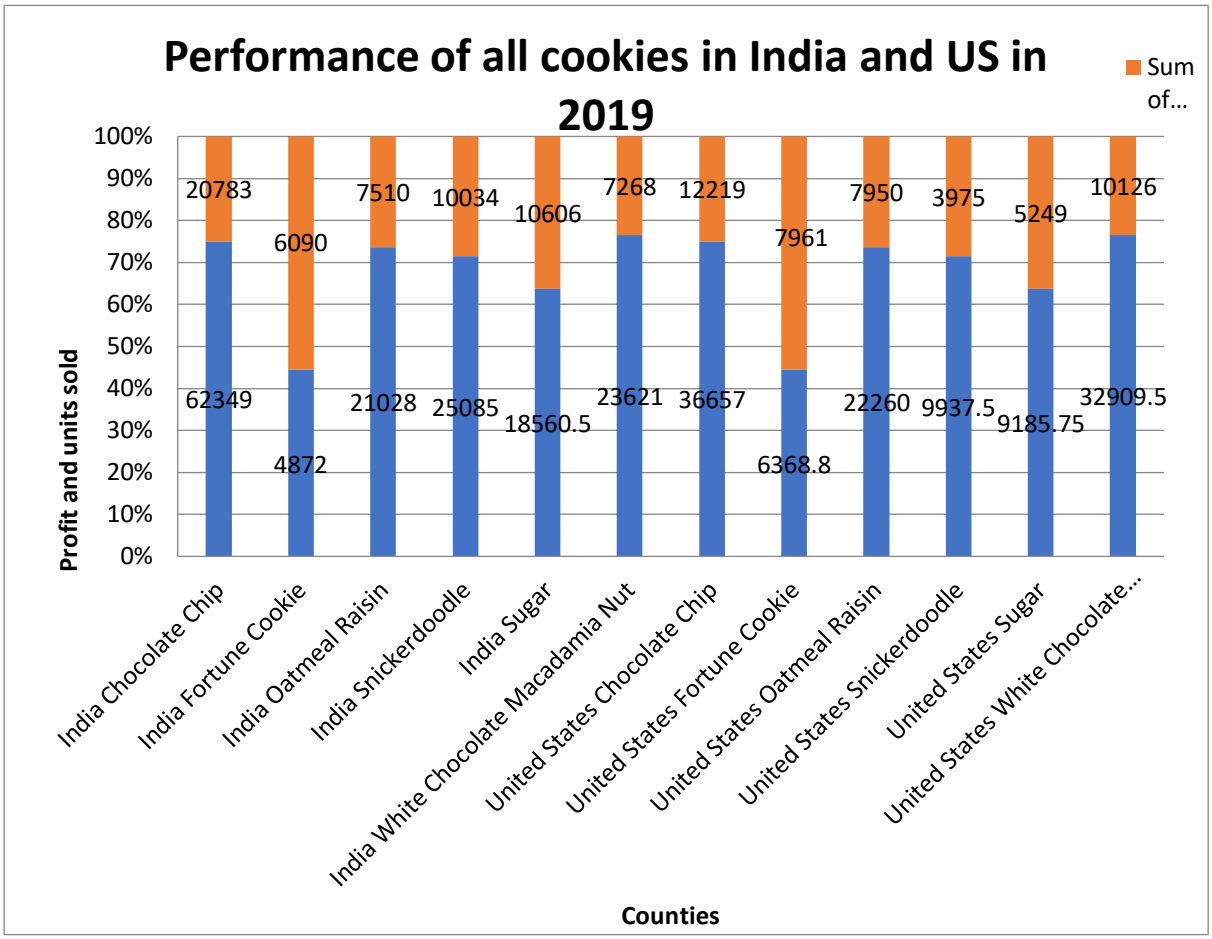

3. Compare all the countries on the bases of profit and unit sold, which is the best performance country on the basis of profit.

Ans:-India stands out as the leading performer globally when it comes to both profit generation and units sold in the Choco chip market.




4 .which Cookie is the best Selling Cookie in India and US in year 2019,

Ans:-In the year 2019, chocolate chip cookies emerged as the top-selling cookie in both India and the United States.




# Conclusion and Review :

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

# 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

| *Regression 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 |

# Co-relation:

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

| | *Units Sold* | *Revenue* |
|---|---|---|
| Units Sold | 1 | 0.796298 |
| Revenue | 0.796298 | 1 |

## 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 | | | | |

## 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.

| *1725* | | *8625* | | *3450* | | *5175* |
|---|---|---|---|---|---|---|
| Mean | 1608.153 | Mean | 6697.702 | Mean | 2751.795 | Mean |
| Standard Error | 32.83303 | Standard Error | 174.9955 | Standard Error | 77.09541 | Standard Error |
| Median | 1540 | Median | 5868 | Median | 2422.2 | Median |
| Mode | 727 | Mode | 8715 | Mode | 3486 | Mode |
| Standard Deviation | 868.0597 | Standard Deviation | 4626.638 | Standard Deviation | 2038.295 | Standard Deviation |
| Sample Variance | 753527.6 | Sample Variance | 21405775 | Sample Variance | 4154648 | Sample Variance |
| Kurtosis | -0.31828 | Kurtosis | 0.463405 | Kurtosis | 0.807696 | Kurtosis |
| Skewness | 0.436551 | Skewness | 0.869254 | Skewness | 0.931429 | Skewness |
| Range | 4293 | Range | 23788 | Range | 10954.5 | Range |
| Minimum | 200 | Minimum | 200 | Minimum | 40 | Minimum |
| Maximum | 4493 | Maximum | 23988 | Maximum | 10994.5 | Maximum |
| Sum | 1124099 | Sum | 4681694 | Sum | 1923505 | Sum |
| Count | 699 | Count | 699 | Count | 699 | Count |
| Largest(1) | 4493 | Largest(1) | 23988 | Largest(1) | 10994.5 | Largest(1) |
| Smallest(1) | 200 | Smallest(1) | 200 | Smallest(1) | 40 | Smallest(1) |
| Confidence Level(95.0%) | 64.46334 | Confidence Level(95.0%) | 343.5807 | Confidence Level(95.0%) | 151.3667 | Confidence Level(95.0%) |

# 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:

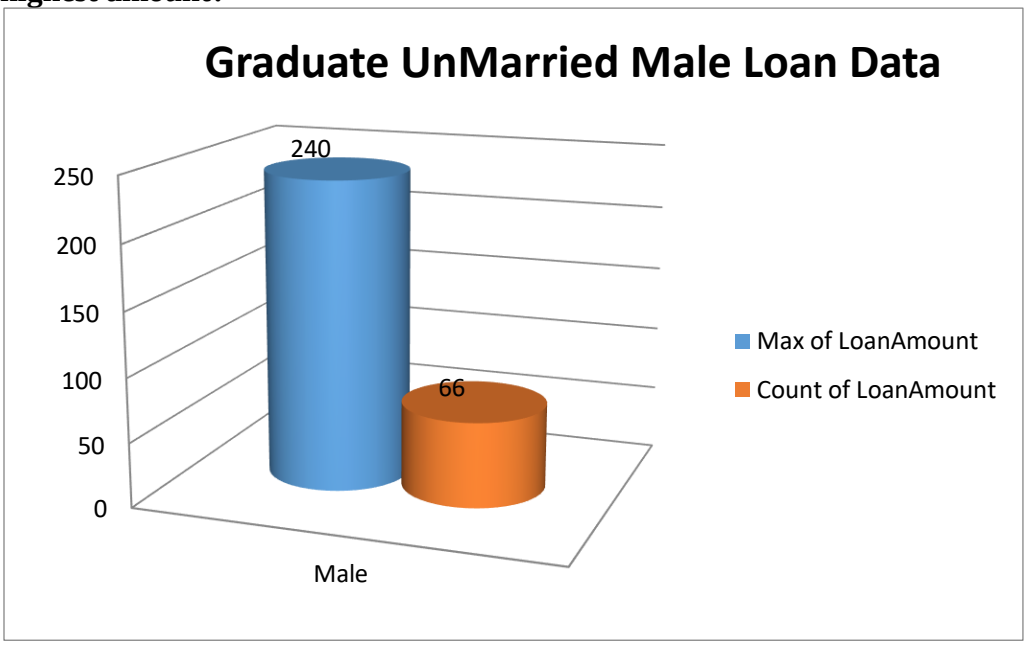
1. Gender: A demographic identifier providing insights into the gender distribution among loan applicants.
2. Marital Status (Married, Not Married): Categorization based on marital status aiding in demographic segmentation.
3. Education (Graduate, Non-graduate): Classification based on educational background 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.

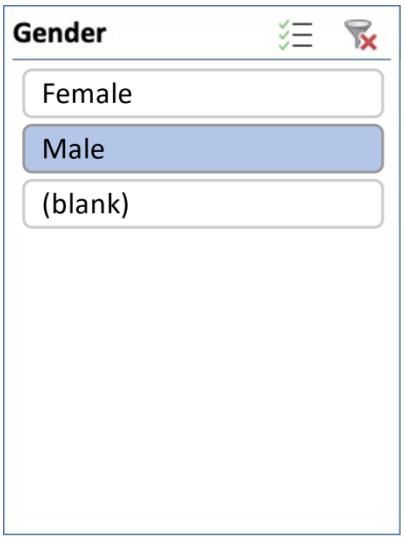
## 2. QUESTIONNAIRE:

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.

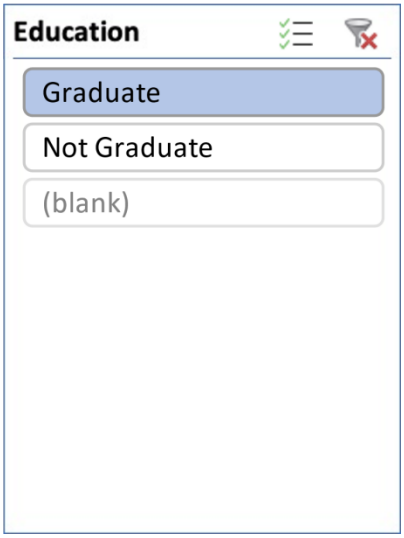
# 3. ANALYTICS:

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

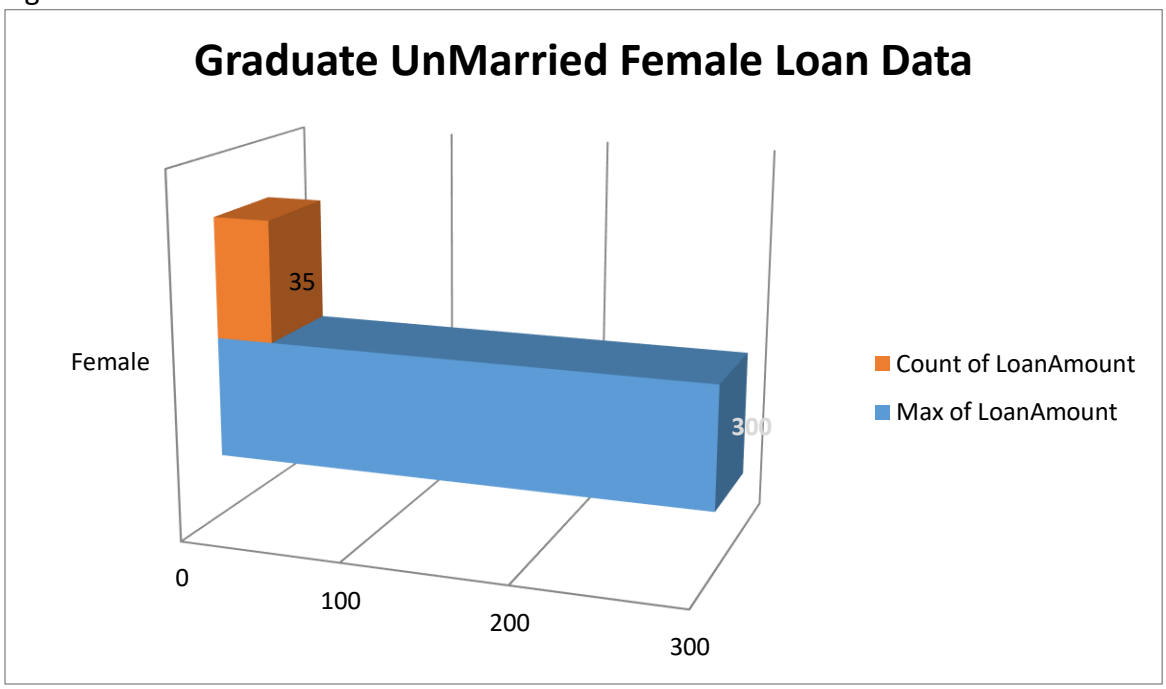




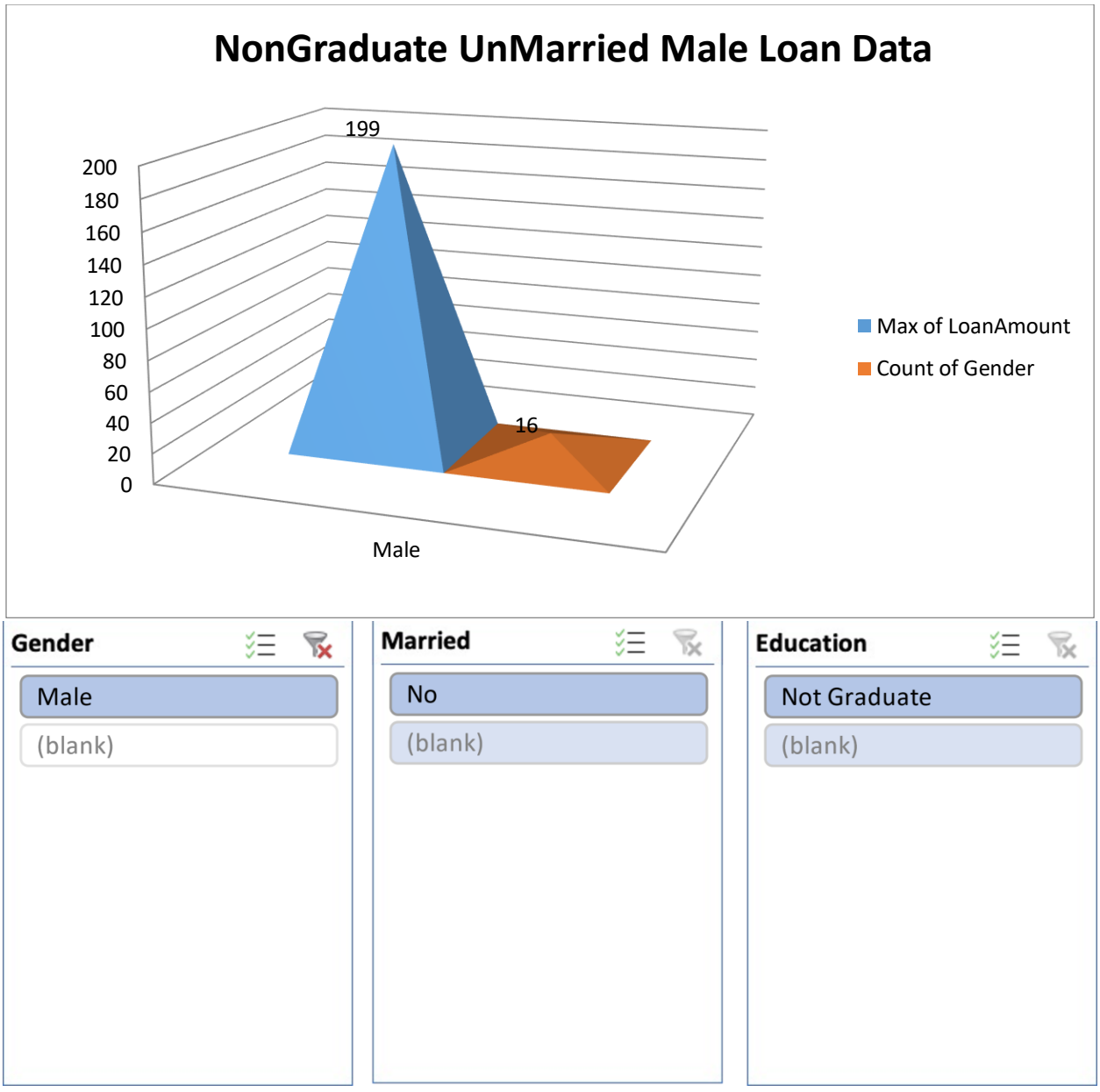


Q2. How many female graduates who are not married applied for Loan? What was the highest amount?

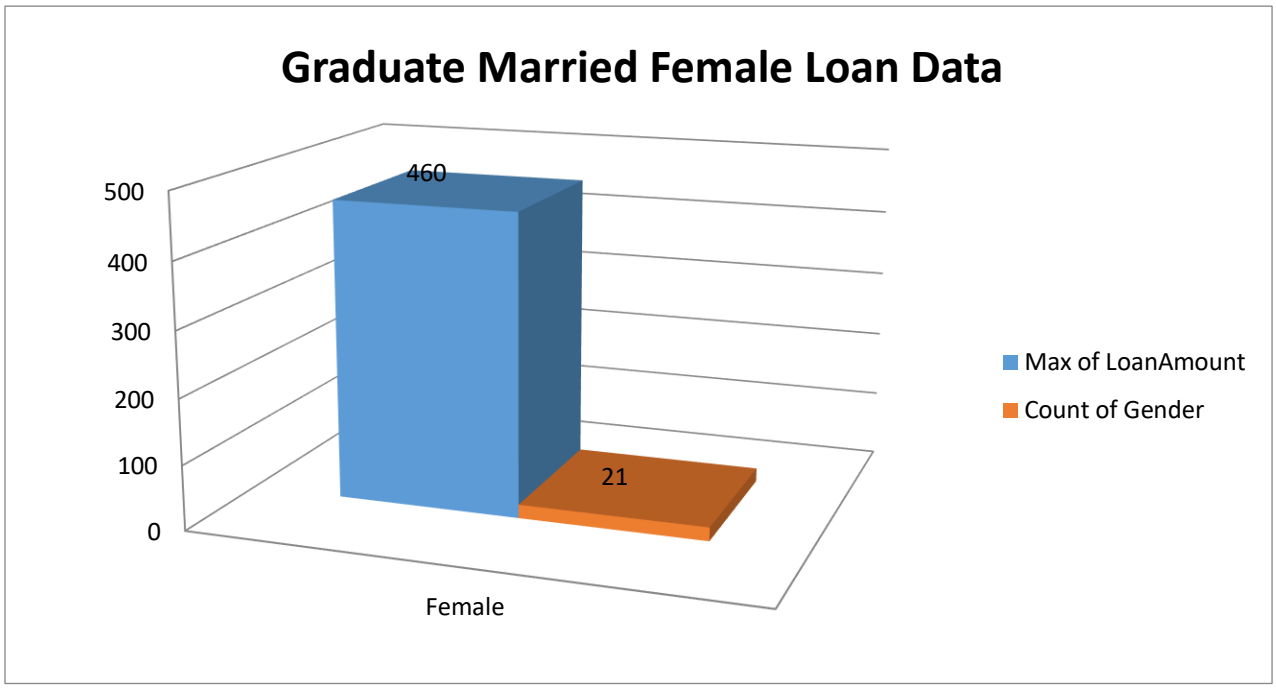


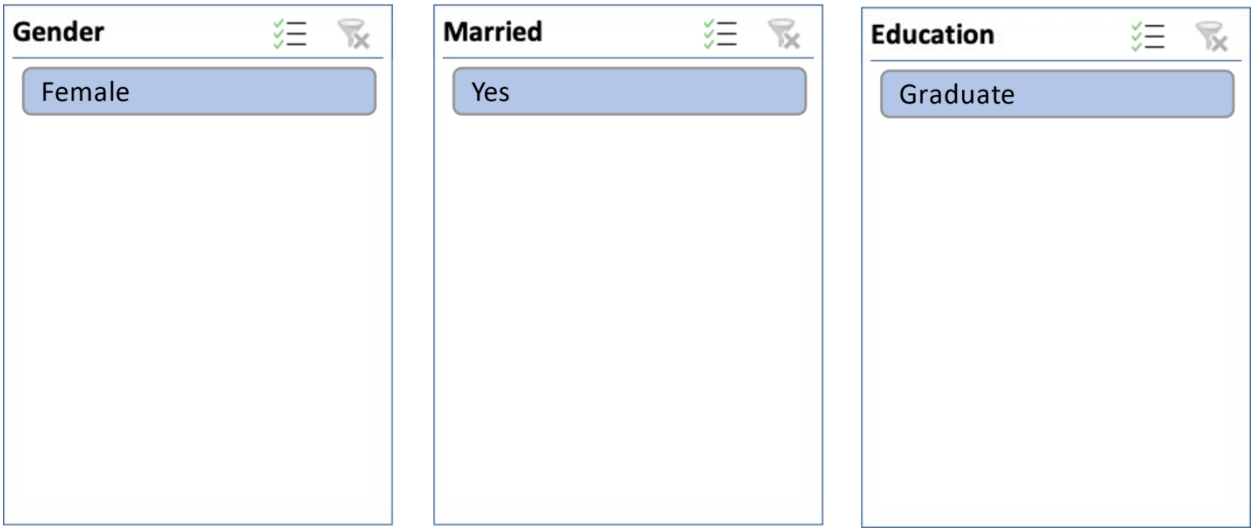

| Gender | Married | Education |
|---|---|---|
| Female | No | Graduate |
| Male | Yes | Not Graduate |
| (blank) | (blank) | (blank) |

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

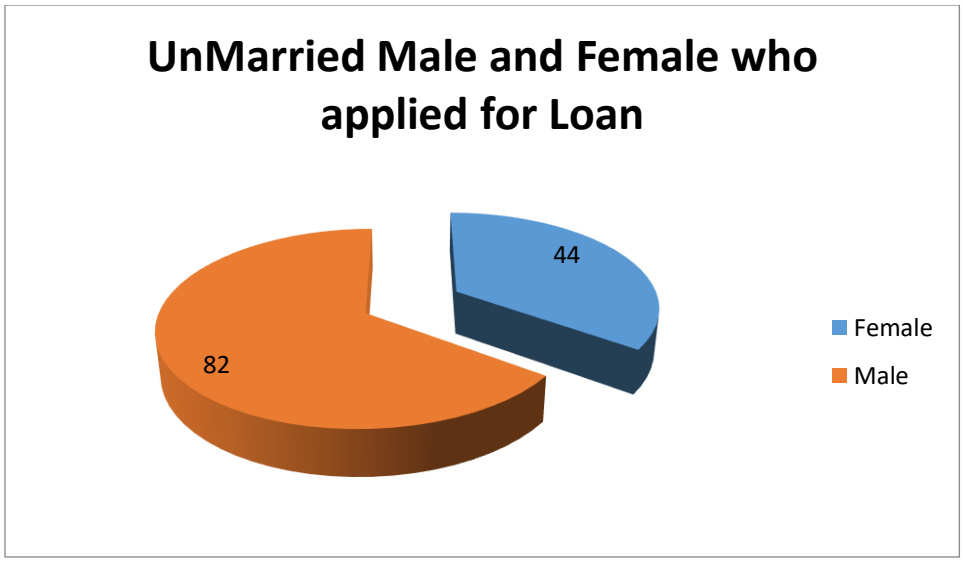


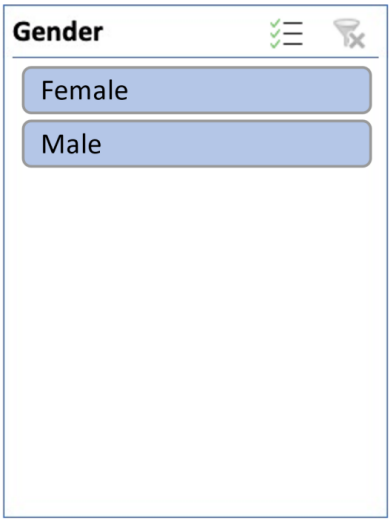

Q4. How many female graduates who are married applied for Loan? What was the highest amount?

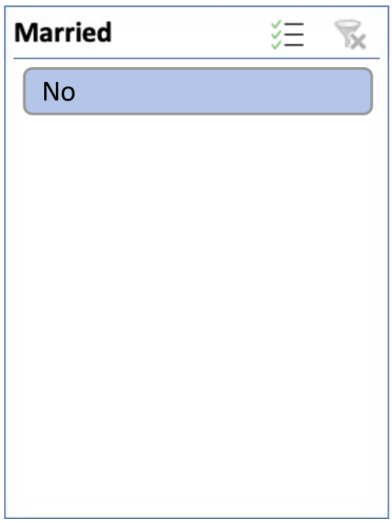


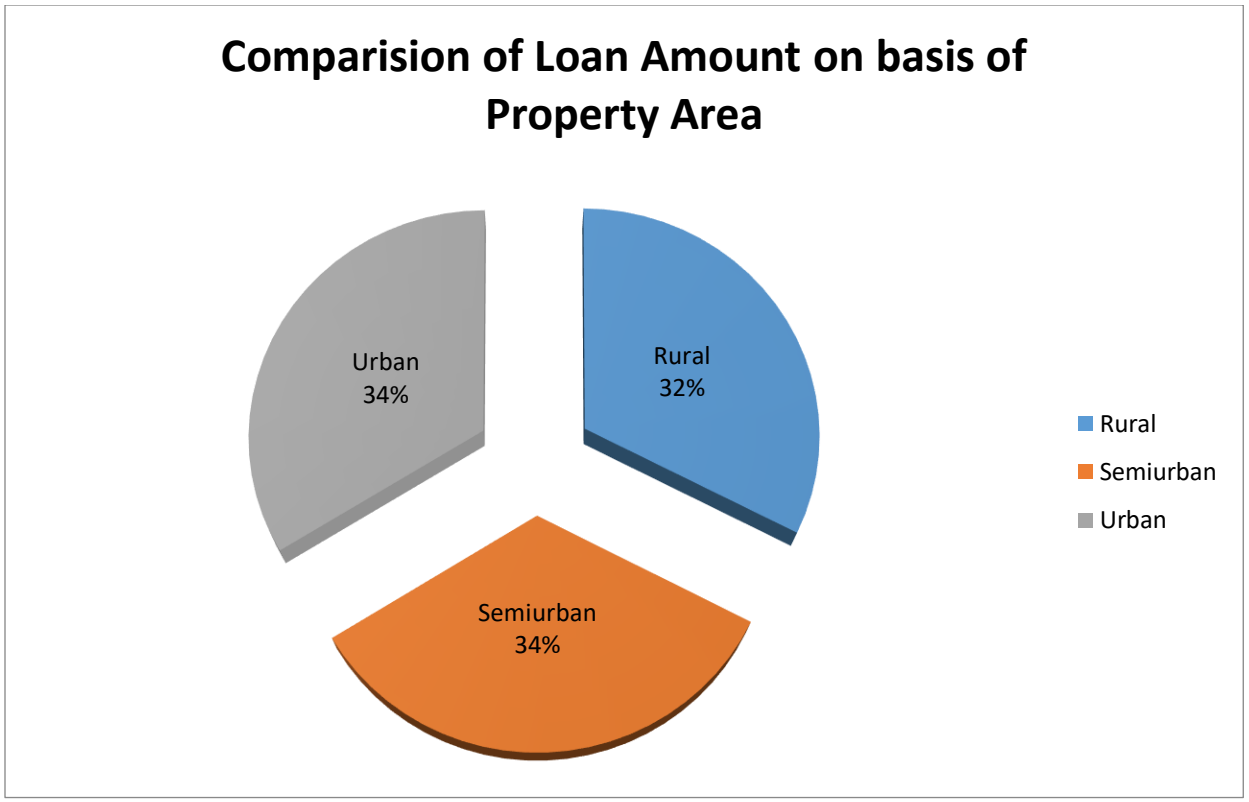



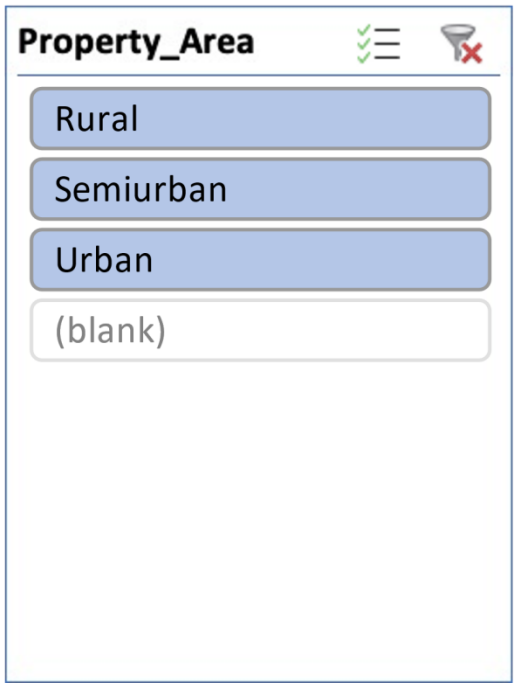

Q5. How many male and female who are not married applied for Loan? Compare Urban, Semi-urban and rular on the basis of amount.

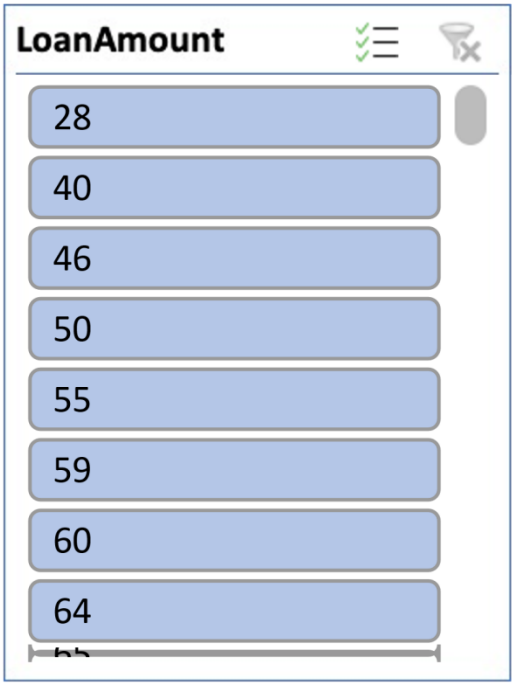

UnMarried Male and Female who applied for Loan
44
82
Female
Male


Gender
Female
Male


Married
No




# 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.* |
|---|---|---|---|---|---|---|---|
| Intercept | 106.07753 | 4.10024098 | 25.8710478 | 1.7585E-84 | 98.014396 | 114.140665 | 98.01439 |
| 5720 | 0.0058851 | 0.00059692 | 9.85902887 | 1.7676E-20 | 0.00471125 | 0.00705895 | 0.004711 |

# 6. CO-RELATION :

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 | 1763655 | 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 | 10821681107 | 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) | 550 |
| Smallest(1) | 0 | Smallest(1) | 0 | Smallest(1) | 0 |
| Confidence | 504.0756 | Confidence | 239.6059 | Confidence | 6.462910 |
| Level(95.0%) | 067 | Level(95.0%) | 543 | Level(95.0%) | 219 |

# REPORT

# SHOP SALES DATA

## 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.

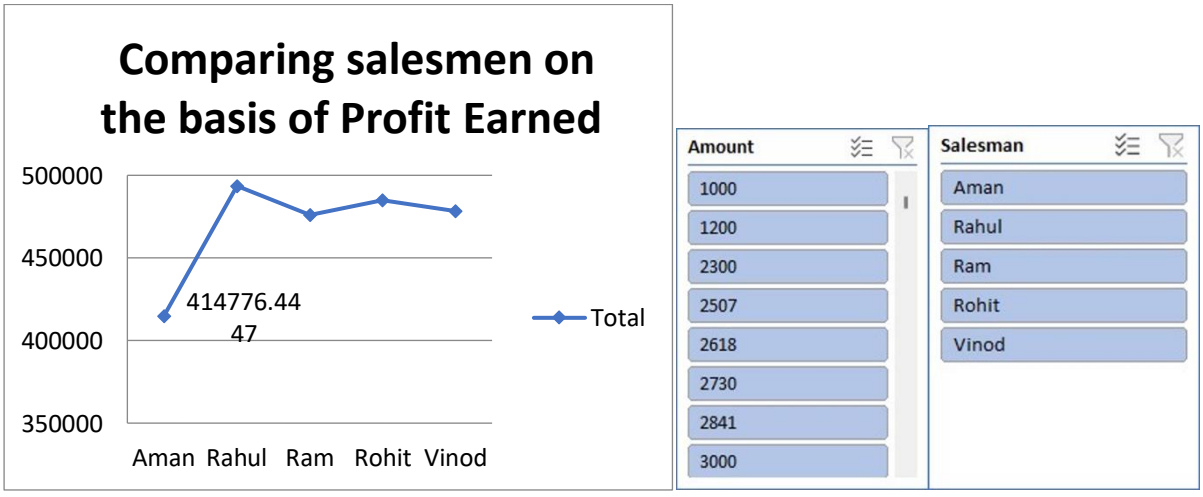
## Questionaries :

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.

## Analytic :

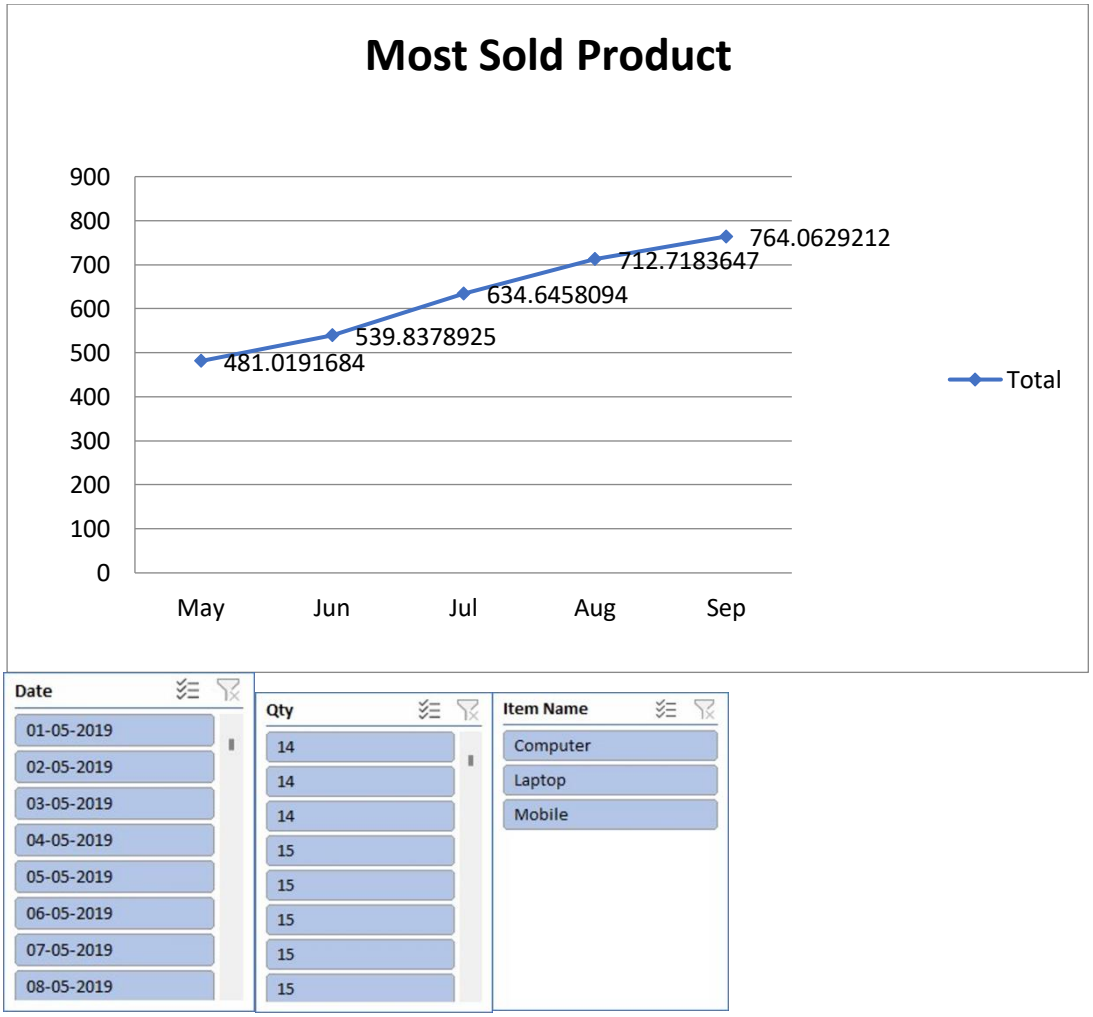
1. Compare all the salesmen on the basis of profit earn.

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



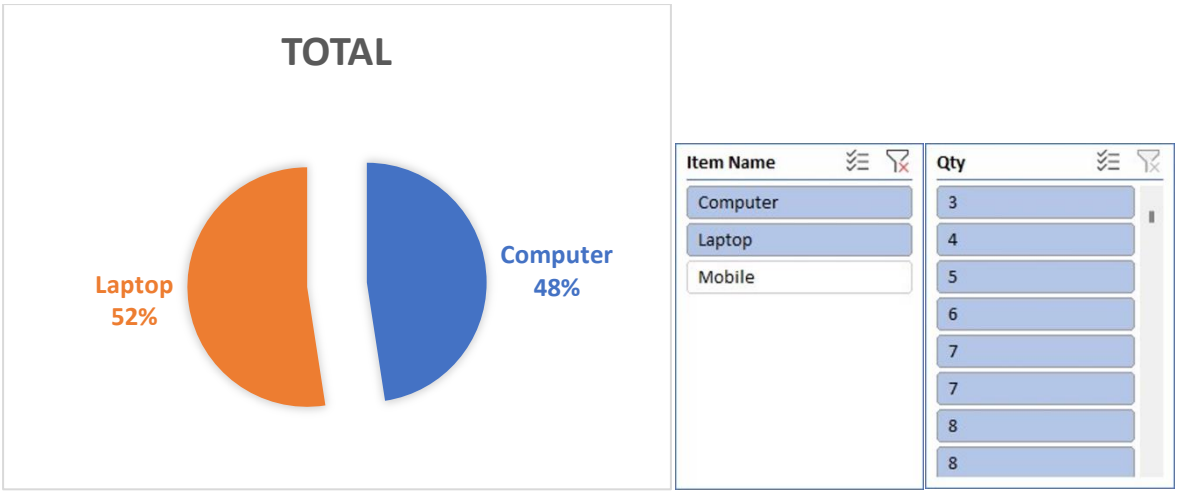
2. Find out most sold product over the period of May-September.

Ans:- To identify the most sold product over the period of May-September, we would need to analyze the sales data within this timeframe. By aggregating the quantity sold for each product across all transactions during this period and then determining which product has the highest total quantity sold, we can pinpoint the most popular item.



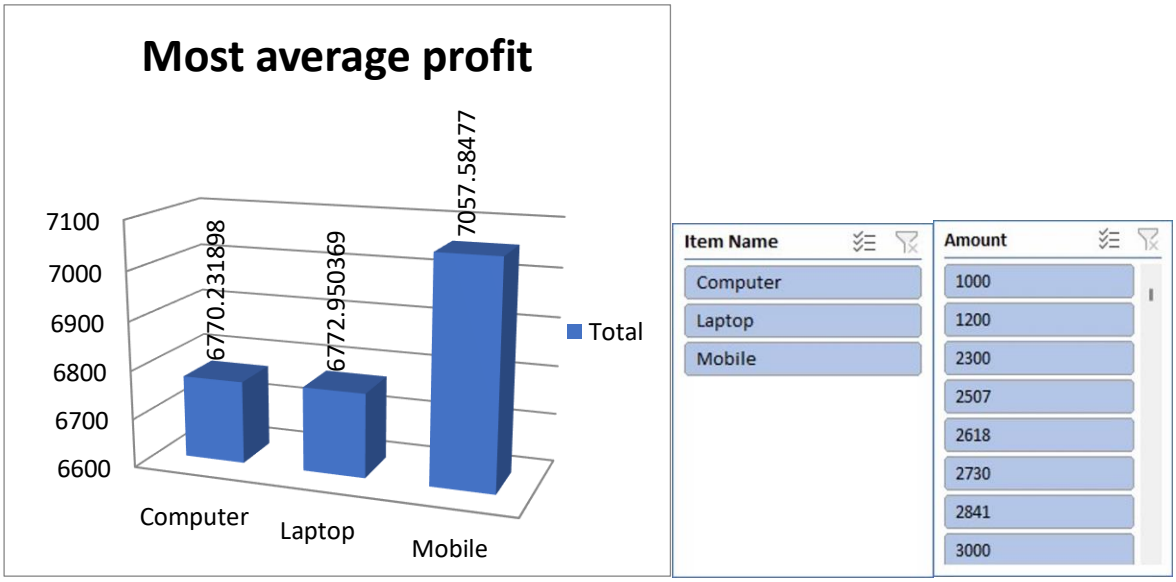

3. Find out which of the two product sold the most over the year Computer or Laptop?

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



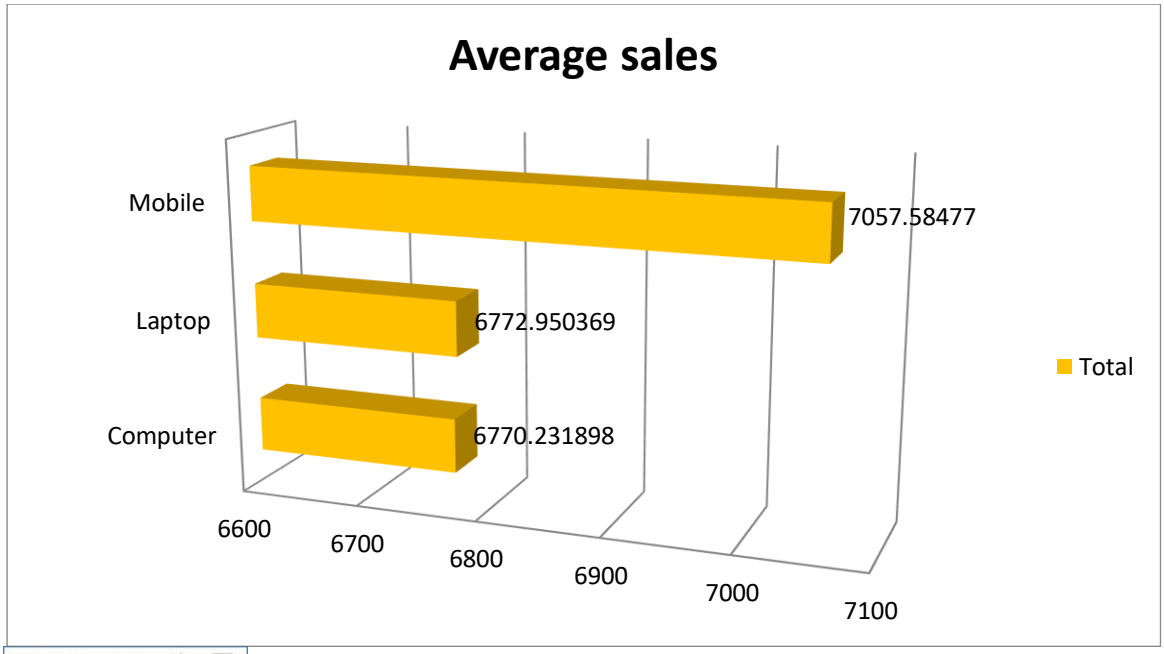

4 . Which item yield most average profit?

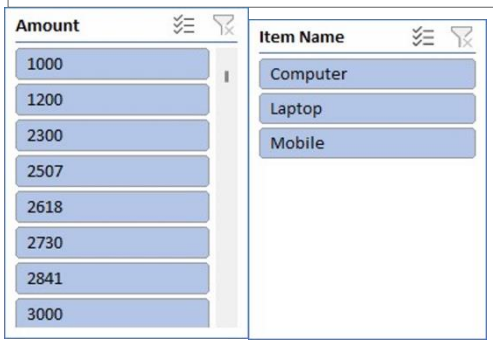
Ans:-The item that yields the most profit between laptop, computer and mobile is :




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

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




# 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 |

## Co-relation:

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 | 47825420 | 28 | 170851 | | | |

| | | |
|---|---|---|
| Group | | |
| 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 Samples

## 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.

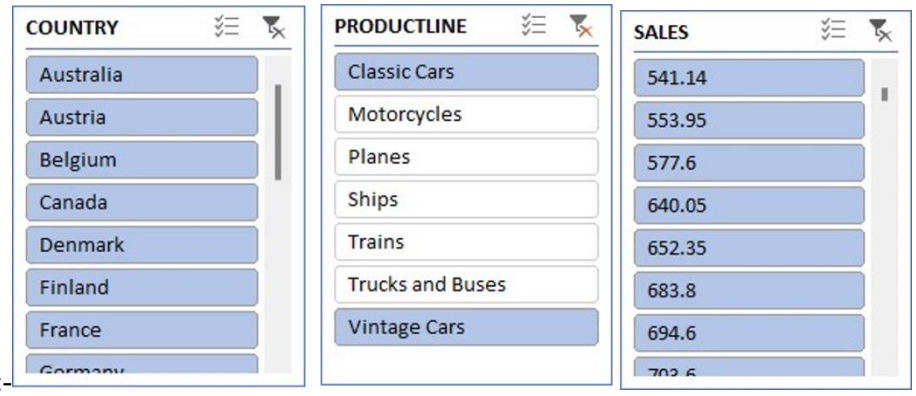
## Questionaries:

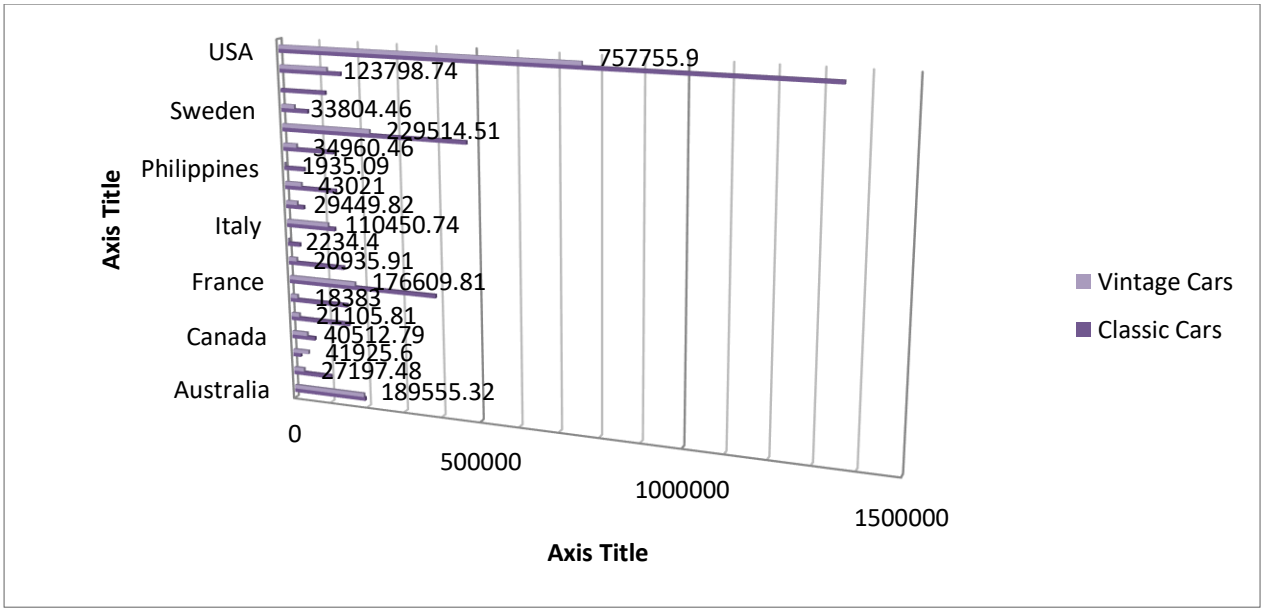
1. Compare the sale of Vintage cars and Classic cars for all the countries.
2. Find out average sales of all the products? which product yield most sale?
3. Which country yields most of the profit for Motorcycles, Trucks and buses?
4. Compare sales of all the items for the years of 2004, 2005.
5. Compare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

## Analytics:

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

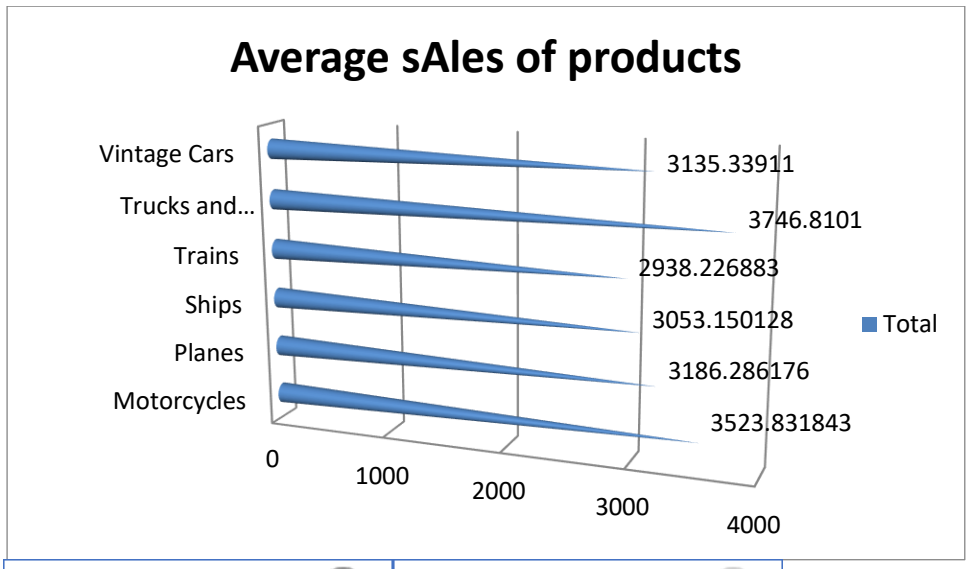
Ans:-The comparsion of sale of Vintage cars and Classic cars for all the countries is given below:-






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

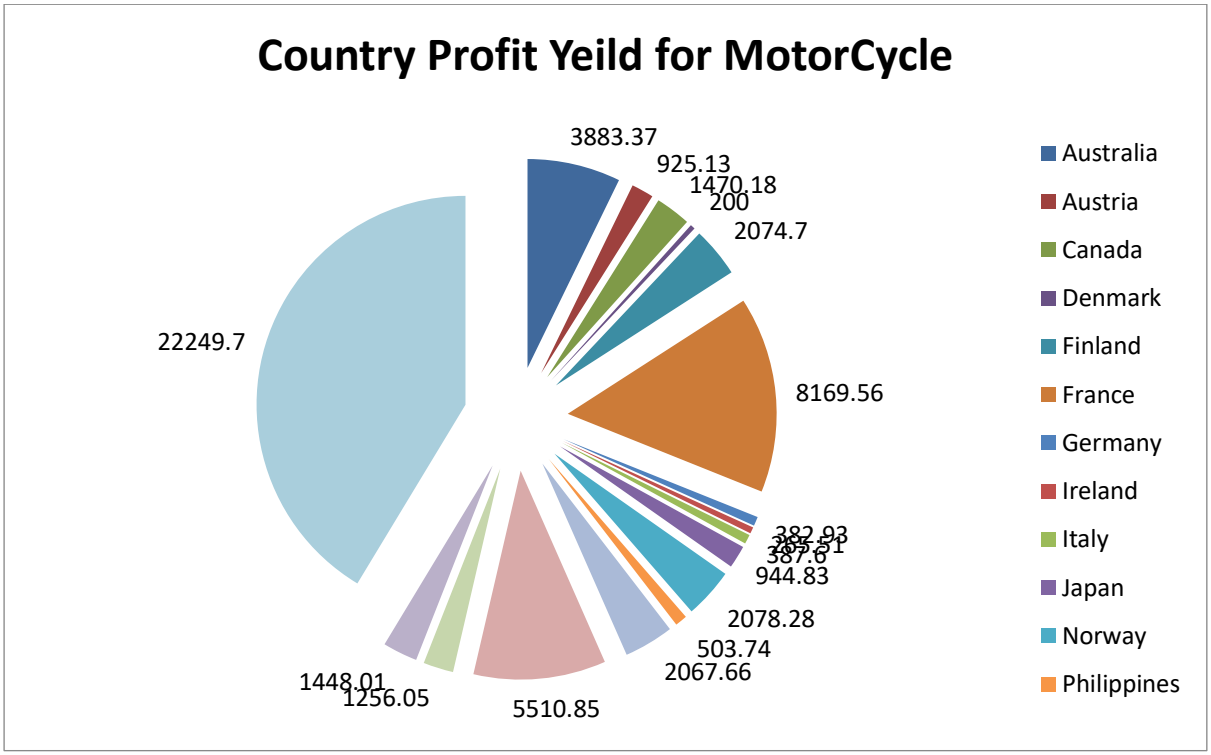
Ans:




| PRODUCTLINE | SALES |
|---|---|
| Classic Cars | 482.13 |
| Motorcycles | 541.14 |
| Planes | 553.95 |
| Ships | 577.6 |
| Trains | 651.8 |
| Trucks and Buses | 652.35 |
| Vintage Cars | 683.8 |
| | 694.6 |

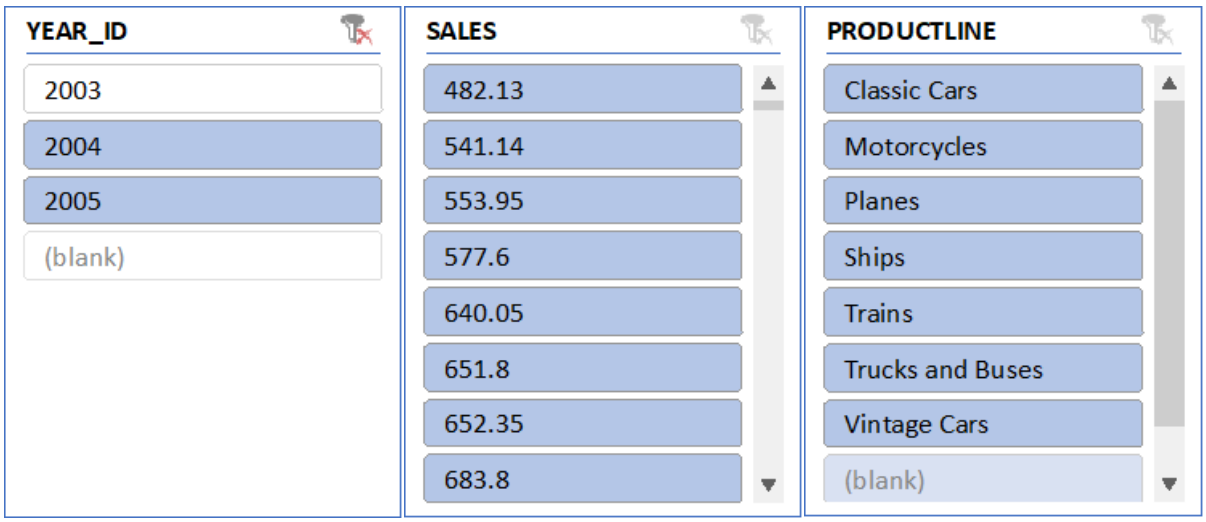
3. Which country yields most of the profit for Motorcycles, Trucks and buses?

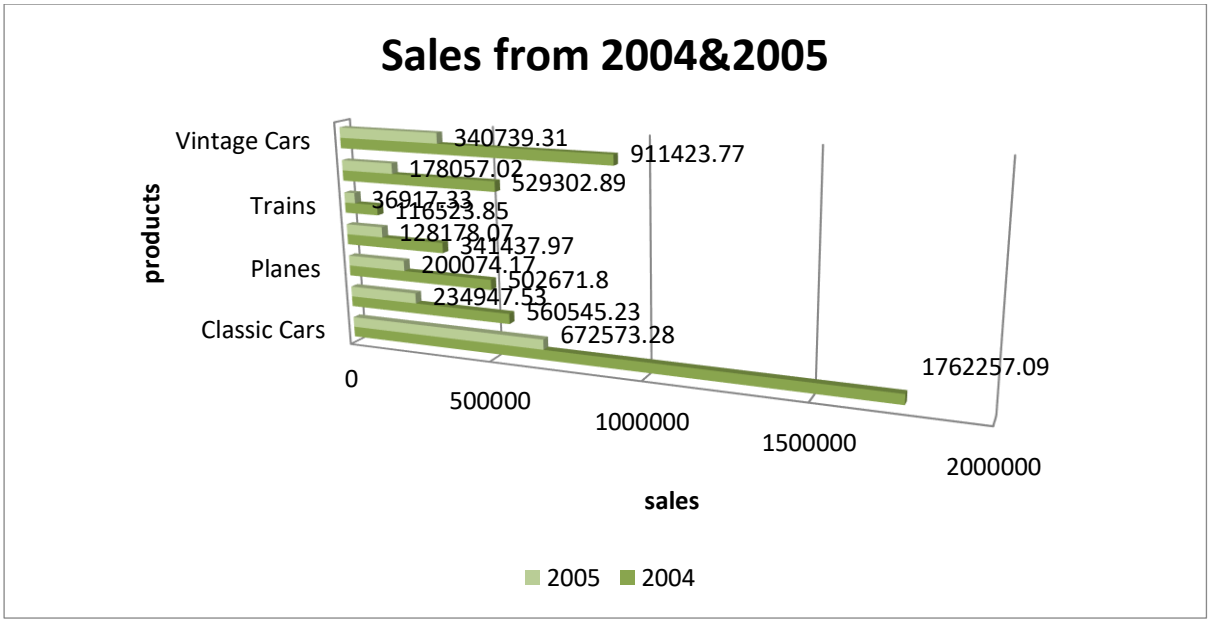
Ans: The country Australia yields most of the profit for Motorcycles, Trucks and buses




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

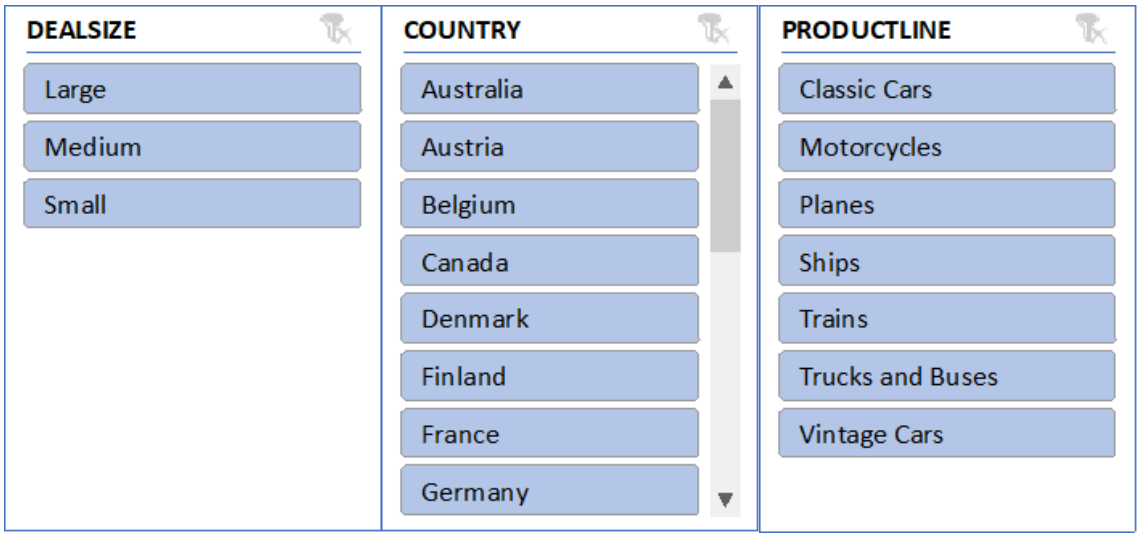
Ans: - The following is the sales of all the items for the years of 2004, 2005 and as graph represents the sales has grown down from 20024 to 2005.

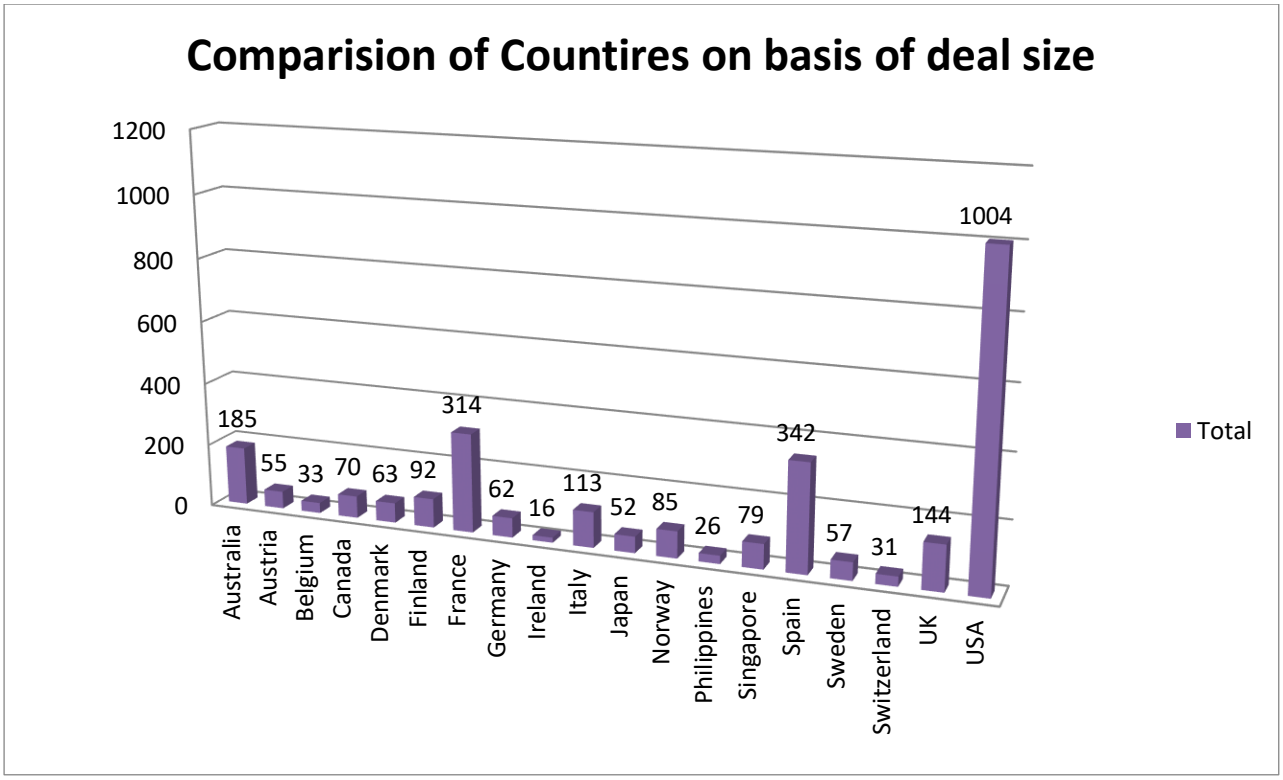





5. Compare all the countries based on deal size.

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






Regression and Anova

| SUMMARY OUTPUT | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| *Regression Statistics* | | | | | | |
| Multiple R | 0.657840928 | | | | | |
| R Square | 0.432754687 | | | | | |
| Adjusted R Square | 0.432553607 | | | | | |
| Standard Error | 1387.45926 | | | | | |
| Observations | 2823 | | | | | |
| | | | | | | |
| ANOVA | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | |
| Regression | 1 | 4142995200 | 4142995200 | 2152.157001 | 0 | |
| Residual | 2821 | 5430546866 | 1925043.199 | | | |
| Total | 2822 | 9573542065 | | | | |
| | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | -1470.590019 | 111.4099971 | -13.19980305 | 1.20143E-38 | -1689.043329 | -1252.13671 |
| PRICE EACH | 60.05936566 | 1.294624334 | 46.39134619 | 0 | 57.52085944 | 62.59787188 |

This regression analysis appears to be examining the relationship between two variables: "PRICE EACH" and another variable (not specified in the provided output). Here are the results:

1. **Regression Equation:** The regression equation can be written as: Y=−1470.59 ( PRICE EACH)+60.06where:
    - *Y* represents the dependent variable Quantity.
    - *X* represents the independent variable "PRICE EACH".
2. **Interpretation of Coefficients:**
    - The intercept coefficient (-1470.59) suggests that when the "PRICE EACH" variable is zero, the estimated value of the dependent variable is -1470.59. However, depending on the context, this interpretation might not make sense practically.
    - The coefficient for "PRICE EACH" (60.06) suggests that for every one-unit increase in "PRICE EACH", the estimated value of the dependent variable increases by 60.06 units.
3. **Statistical Significance:**
    - The p-value associated with the coefficient for "PRICE EACH" is 00, indicating that the coefficient is statistically significant at conventional levels of significance (typically $\alpha$=0.05).
    - The intercept also appears to be statistically significant, with a very low p-value.
4. **Goodness of Fit:**
    - The R-squared value (0.433) indicates that approximately 43.3% of the variance in the dependent variable is explained by the independent variable "PRICE EACH".
    - The adjusted R-squared value (0.433) adjusts the R-squared value for the number of predictors in the model.
5. **ANOVA:**
    - The ANOVA table indicates that the regression model as a whole is statistically significant, as the p-value associated with the F-statistic is 00.
6. **Standard Error:**
    - The standard error (1387.46) gives 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 2823 observations.

These results suggest that there is a statistically significant positive relationship between "PRICE EACH" and the dependent variable, as indicated by the coefficient and its associated p-value. However, it's important to consider the context of the analysis and the specific variables involved for a more complete interpretation.

CORELATION:

The correlation coefficient you calculated (0.657840928) represents the strength. Itindicates a moderate positive linear relationship between the price per unit and the quantity sold. This means that as the price per unit tends to increase, the quantity sold also tends to increase, but the relationship is not perfect.

Descriptive Statistics:

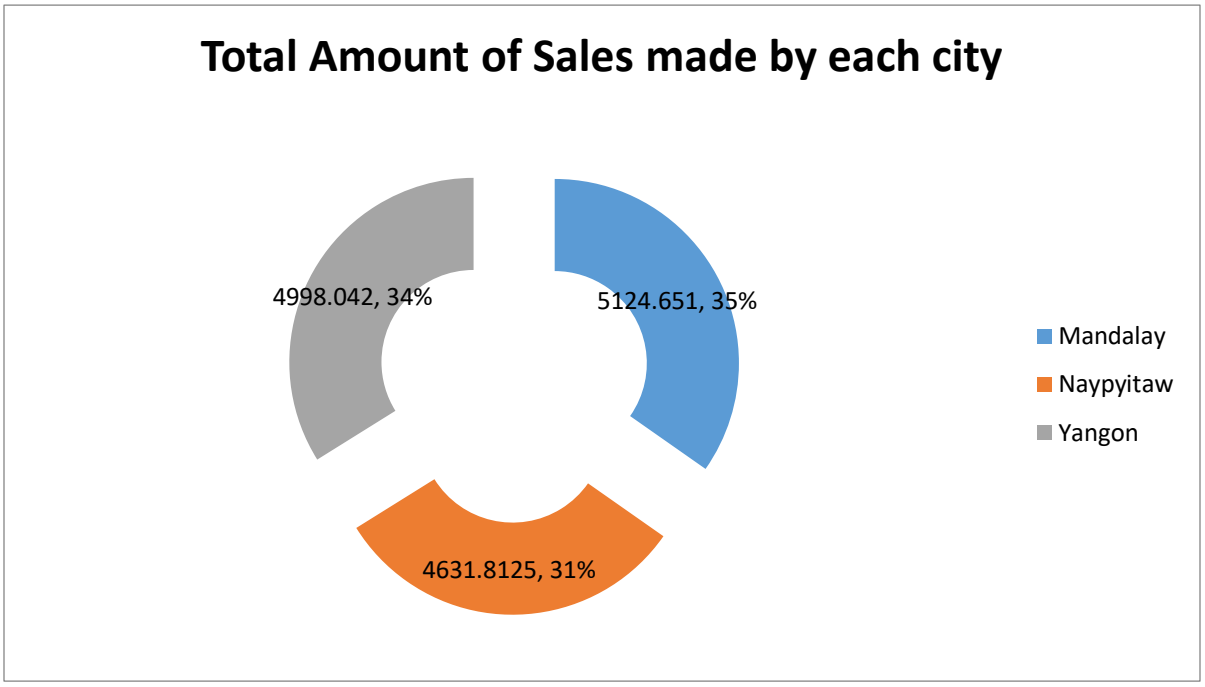
| *SALES* | |
|---|---|
| Mean | 3553.889072 |
| Standard Error | 34.66589212 |
| Median | 3184.8 |
| Mode | 3003 |
| Standard Deviation | 1841.865106 |
| Sample Variance | 3392467.068 |
| Kurtosis | 1.792676469 |
| Skewness | 1.161076001 |
| Range | 13600.67 |
| Minimum | 482.13 |
| Maximum | 14082.8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

# 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.

# Supermarket Sales

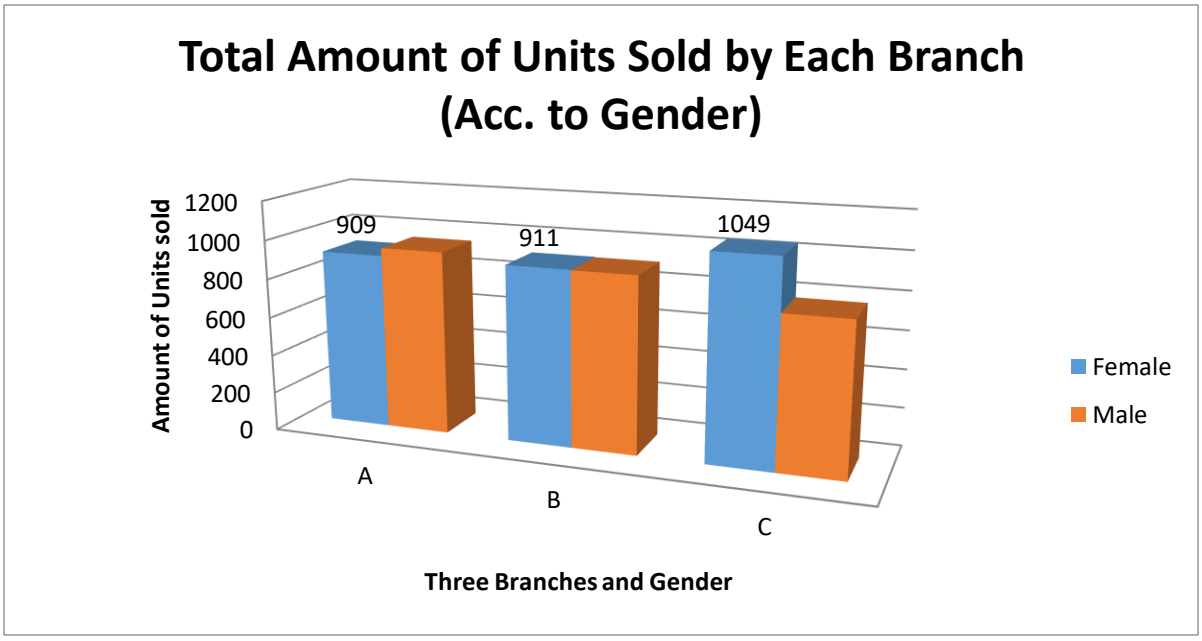
**Q1.** Which of the given cities having tax 5% slab performed better than all the others?

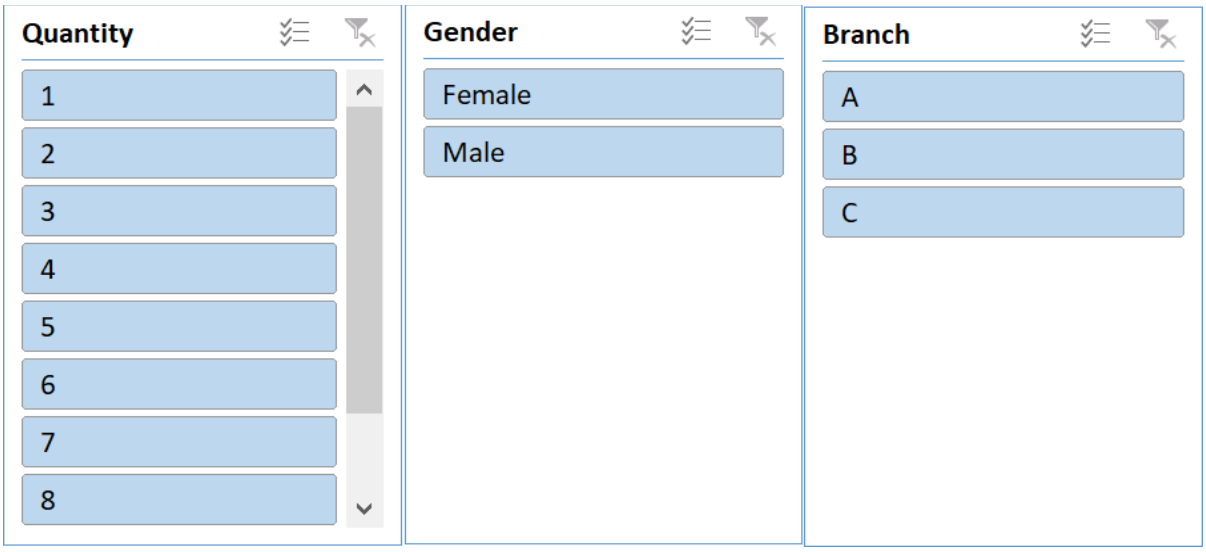



| Total | Tax 5% | City |
|---|---|---|
| 10.6785 | 0.5085 | Mandalay |
| 12.6945 | 0.6045 | Naypyitaw |
| 13.167 | 0.627 | Yangon |
| 13.419 | 0.639 | |
| 14.679 | 0.699 | |
| 16.107 | 0.767 | |
| 16.2015 | 0.7715 | |
| 16.275 | 0.775 | |

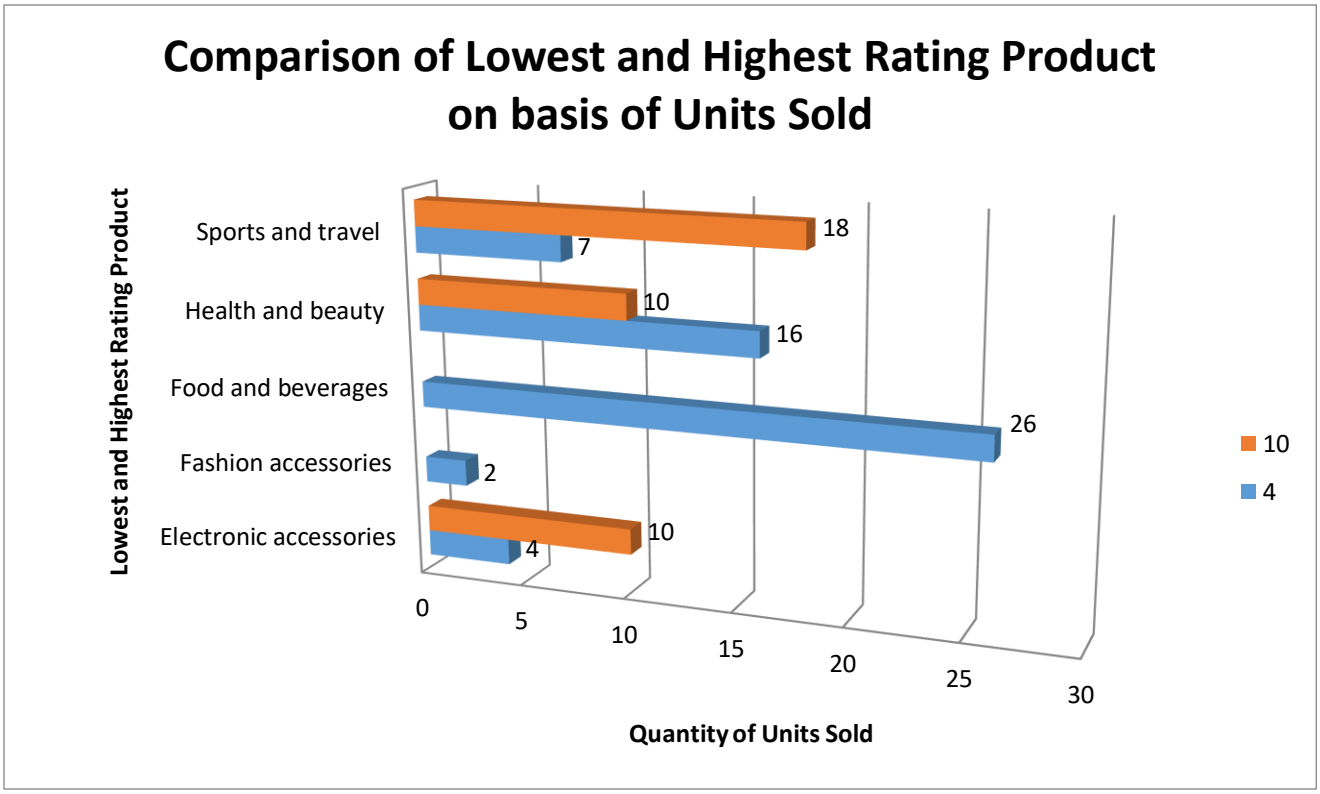
Based on the data analyzed, the city that outperformed all is **Mandalay**.This conclusion is drawn from superior performance in total sales/revenue generation compared to the other cities in the same tax slab of 5%.

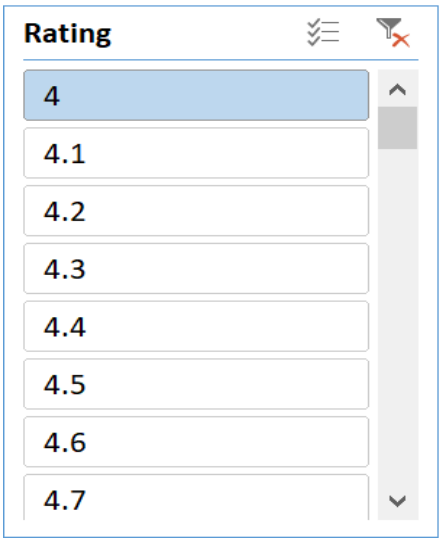
**Q2.** Which customer gender ordered most items from all the three branches?

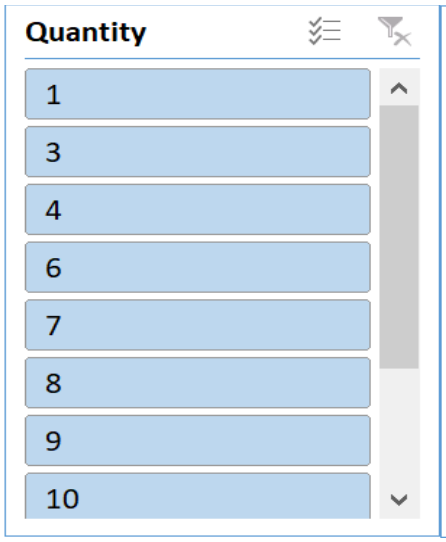


**Q3.** Compare highest and lowest rating products on the basis of units sold.




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

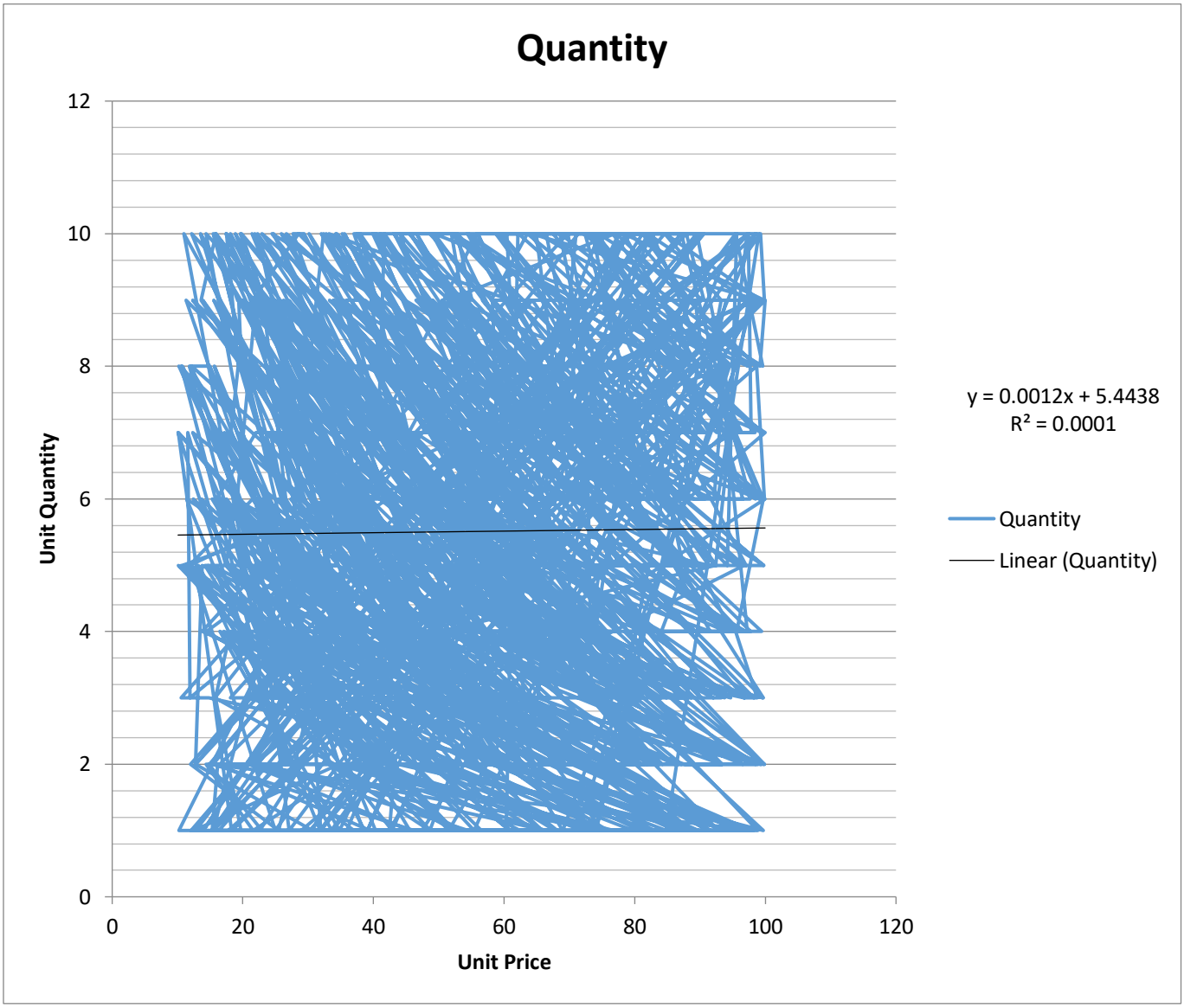
a) What is the degree of freedom?
b) Co-relation of Unit price and revenue generated
c) What result you can draw from regression of the two data

| 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 |

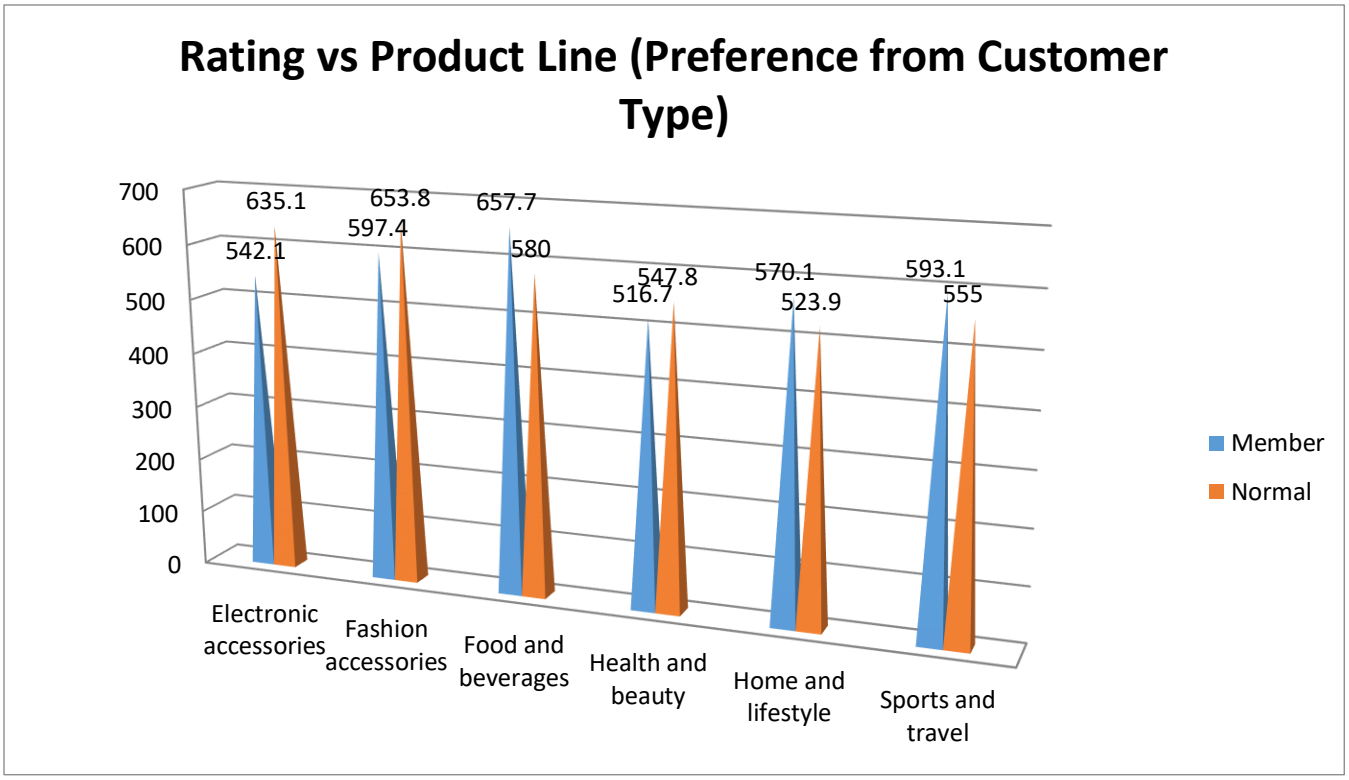
**Solution:**

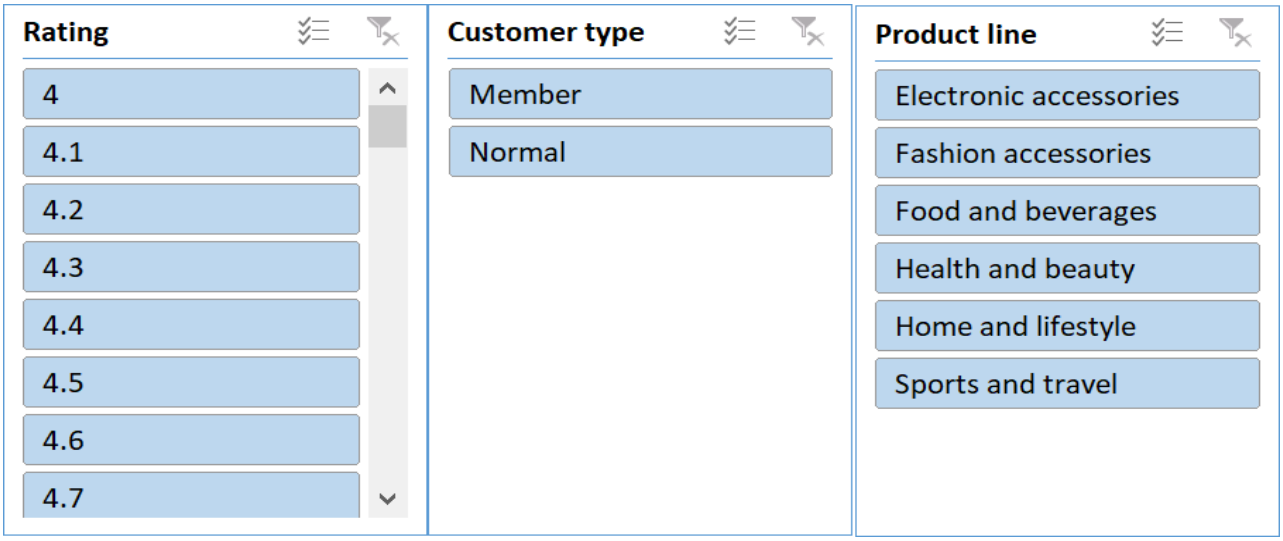
a. The degree of freedom of the analyzed data is 1.
b. The Co-relation of Unit Price and Revenue generated turned out to be 0.63392.
   The two columns considered in the calculation were Unit Price and the Total.
   Function Used: =CORREL

**c.** From the regression result we can plot the following graph:



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



# StoreDatasetReport

## Introduction:

This dataset encompasses sales data from a retail store, featuring a rangeofattributesincludingcustomerdemographics(Gender,AgeGroup),transactiondetails(OrderID, Status), product specifics (Category, SKU), and shipping information. With a focus onunderstanding customer behaviour and product trends, our analysis aims to uncover patterns,preferences, and correlations within the data. By leveraging these insights, businesses canoptimizemarketingefforts,enhanceinventorymanagement,andimprovecustomersatisfaction.

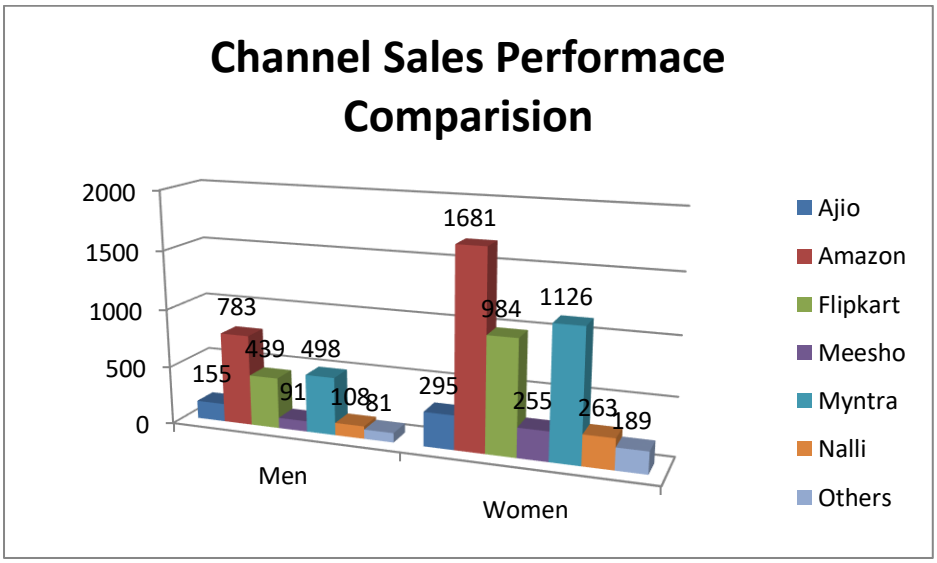
## Questionnaire:

1. whichofthechannelperformedbetterthanallotherchannelsincomparemen&women?
2. Comparecategory.Findoutmostsoldcategoryabove23yearsofageforanygender.
3. Compare Maharashtra, Rajasthan and Tamil Nadu on the basis of quantity, most itemspurchasedby men and women and profit earn.
4. Whichcitysoldmostoffollowingcategories:
   a. Kurta
   b. Set
   c. Westernwears
5. Inwhichmonthmostitemssoldinanyofthestateonthebasisofcategory.

## Analytics:

1. whichofthechannelperformedbetterthanallotherchannelsincomparemen&women?

Ans: Amazon leads in the sales in both men and women category followed by Myntra andFlipkart. Amazon sold almost 3500 units in men category and almost 7500 units in womencategory.Myntrasold 2000unitsin mensection.



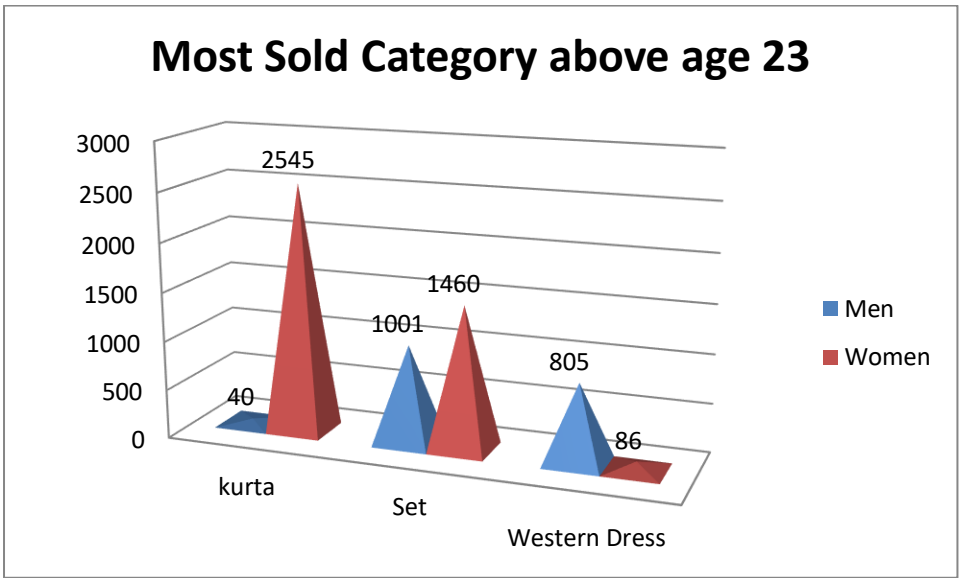
2. Comparecategory.Findoutmostsoldcategoryabove23yearsofageforanygender.

Ans: In the above 23 years of age group Kurta is most sold category in women section with8820 units sold. Set is most sold category in men section with 4365 units sold also set is thesecondmost sold category in womensection.

Thetable ofitems sold is givenbelow:

| Item | Men | Women | GrandTotal |
|---|---|---|---|
| Blouse | 6 | 190 | 196 |
| Bottom | 40 | 28 | 68 |
| EthnicDress | 150 | 77 | 227 |
| kurta | 156 | 8820 | 8976 |
| Saree | 261 | 941 | 1202 |
| Set | 4365 | 6204 | 10569 |
| Top | 45 | 1825 | 1870 |
| WesternDress | 3078 | 380 | 3458 |
| **GrandTotal** | **8101** | **18465** | **26566** |

Thegraphis asfollows:

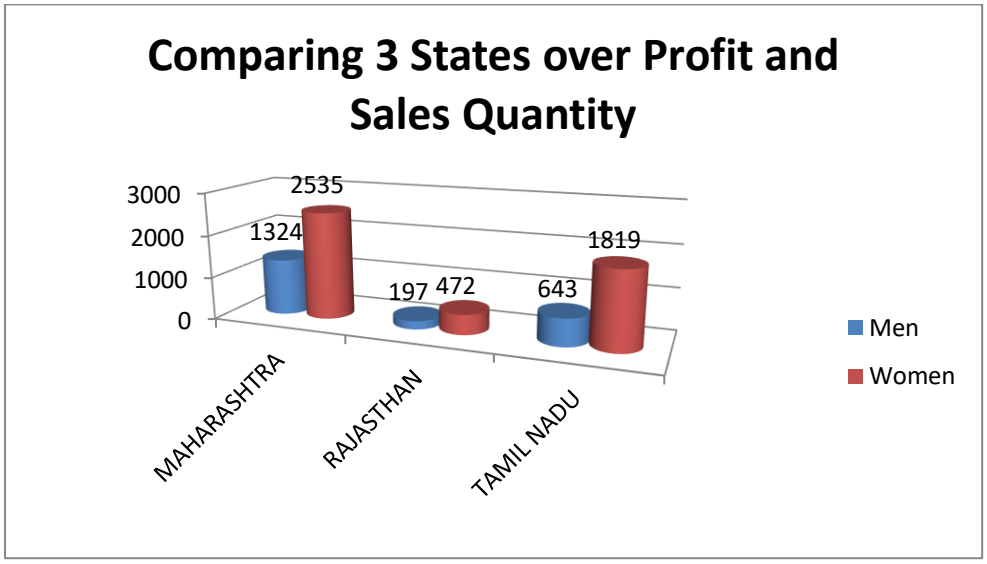



3. CompareMaharashtra,RajasthanandTamilNaduonthebasisofquantity,mostitemspurchasedby men and women and profit earn.

Ans: In Maharashtra: Sales in men category=1390,Sales in women category=

3144In Tamil Nadu: Sales in men category=686, Sales in women category=

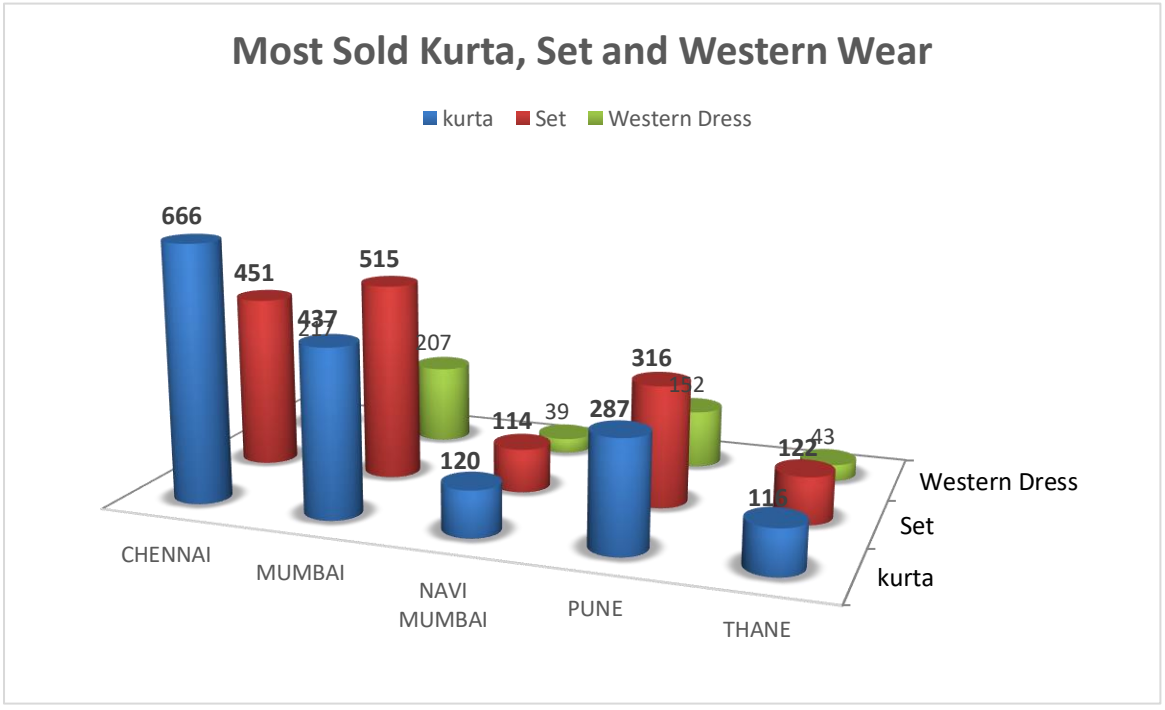
2023InRajasthan:Salesinmen category=21,Salesinwomencategory=543



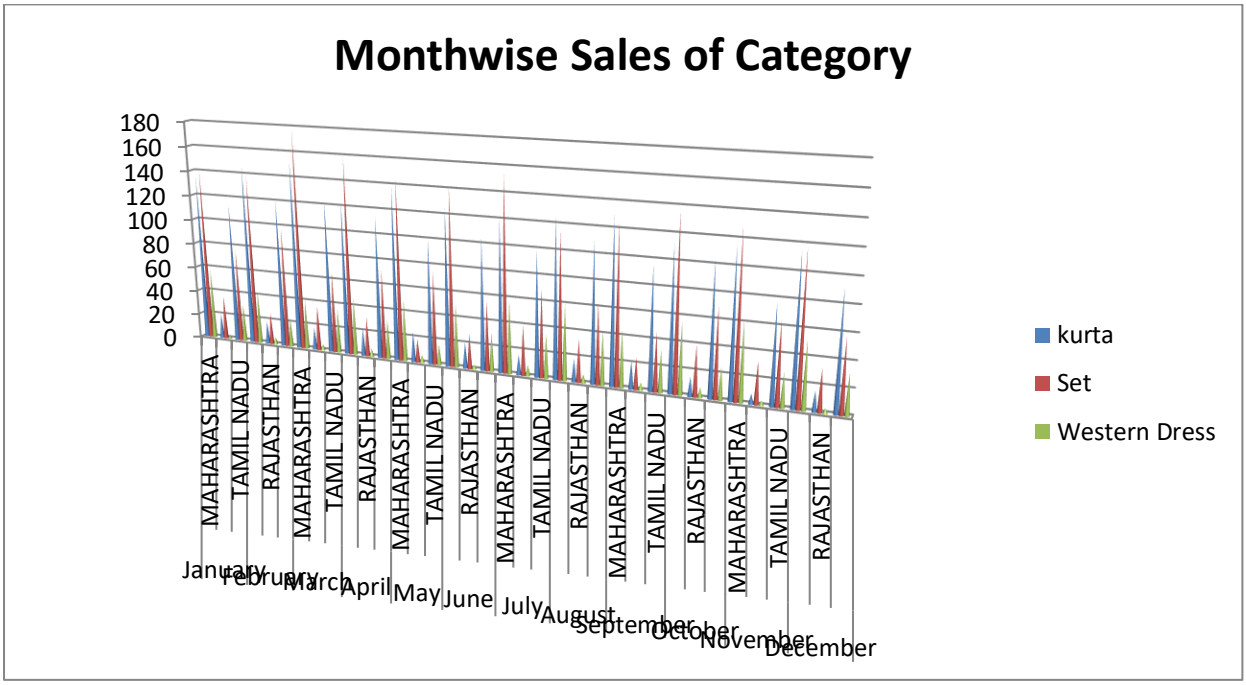

4. Whichcitysoldmostoffollowingcategories
   a. KurtaSet
   b. Westernwears

Ans:

Bengaluru,Chennai,Hyderabad,MumbaiandNewDelhiarethecitiessoldmostofkurtas,Sets andwesternwears.




5. In which month most items sold in any of the state on the basis of category.Ans:Thegraphformost itemssold inanyofstatson basisofcategoryisasfollows:




# ConclusionandReview:

Afterthoroughanalysisofthestoredata,itisevidentthattherearenotabletrendsandinsightsto 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 marketdemand,salesandoverallprofitability.Thiscomprehensiveunderstandingwillenableinformeddecision-makingtooptimizeresources,targetspecificmarkets,andmaximizeprofitsinfuturestoresalesendeavours.