Acropolis Institute of Technology and Research

B. Tech. VI Semester
Jan – June 2024

Lab Assignment On

Data Analytics [CS 605]

Submitted To:
Prof. Anurag Punde

Submitted By:
Isha Gupta
(0827CS211100)

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.

Questionnaires:

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

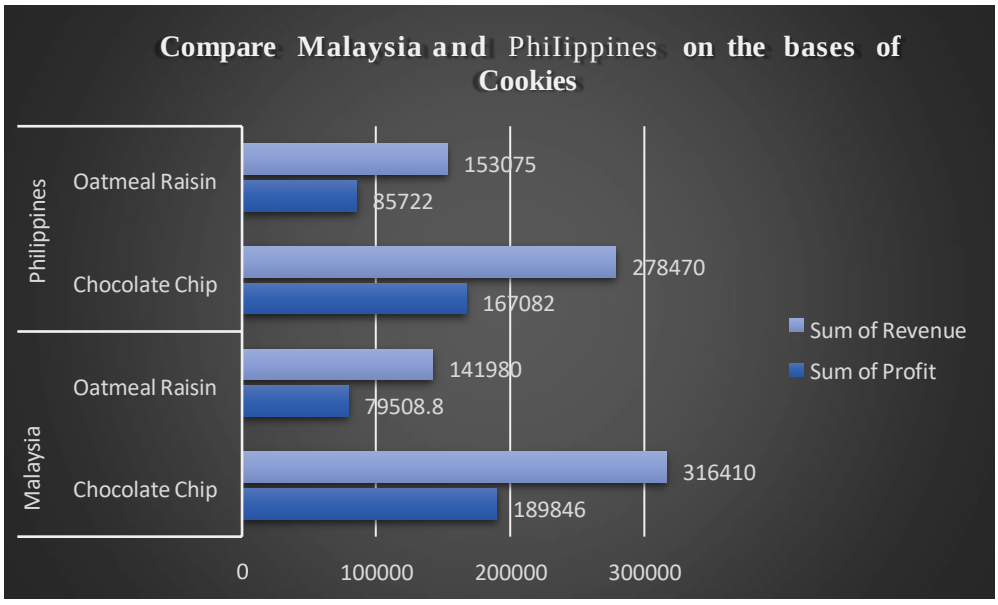
2. What is the performance of Choco Chips Cookies in all Country Which Competes the best.

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,

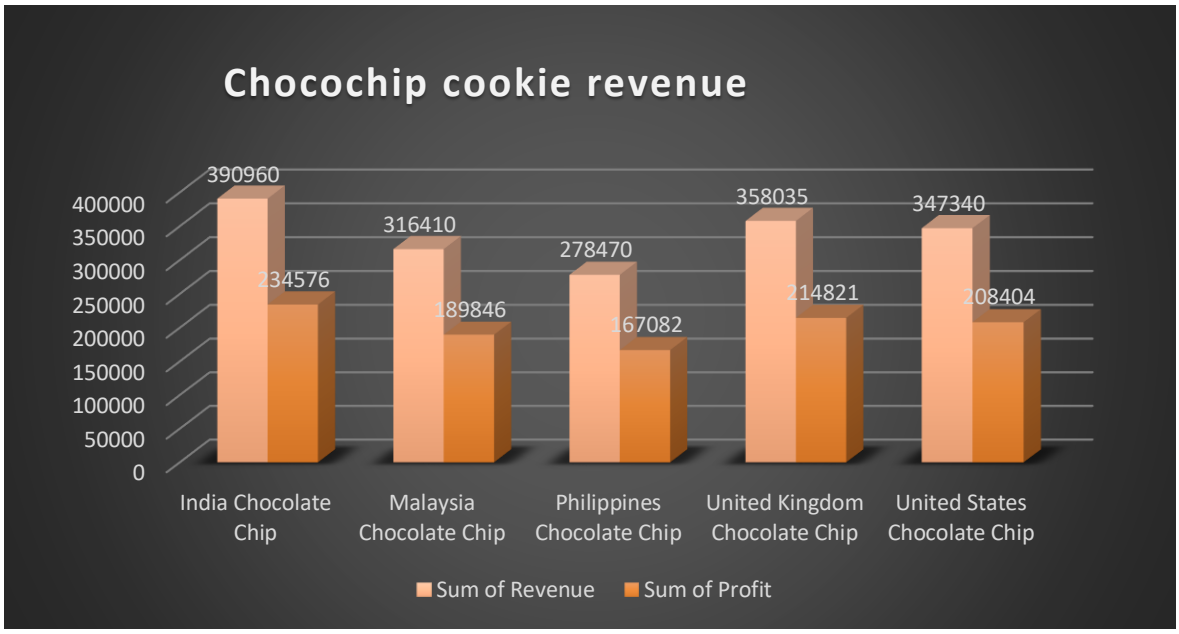
Analytics:

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



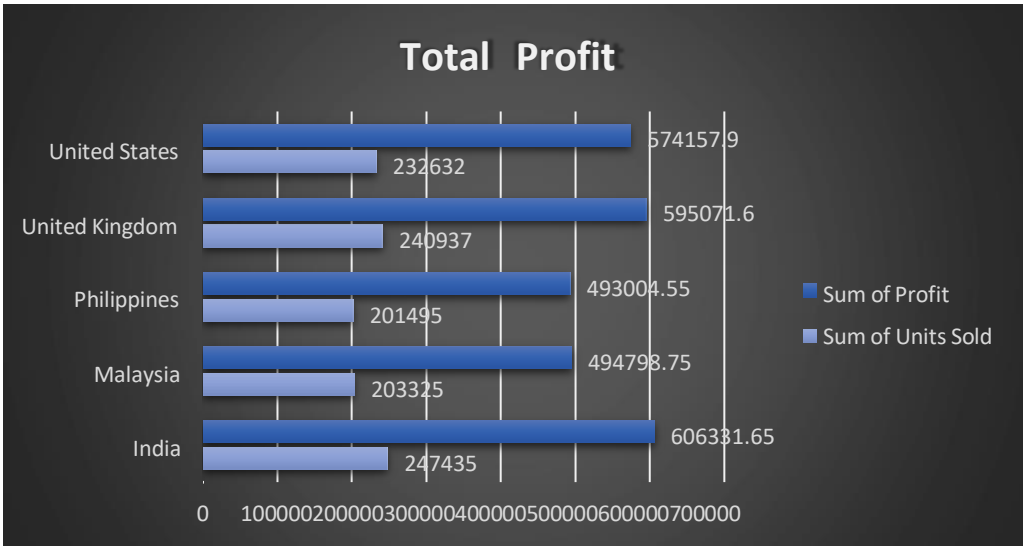
Ans: -The comparison of Malaysia and Philippines on bases of Chocolate chip and White Chocolate Macadamia nut is given above.

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



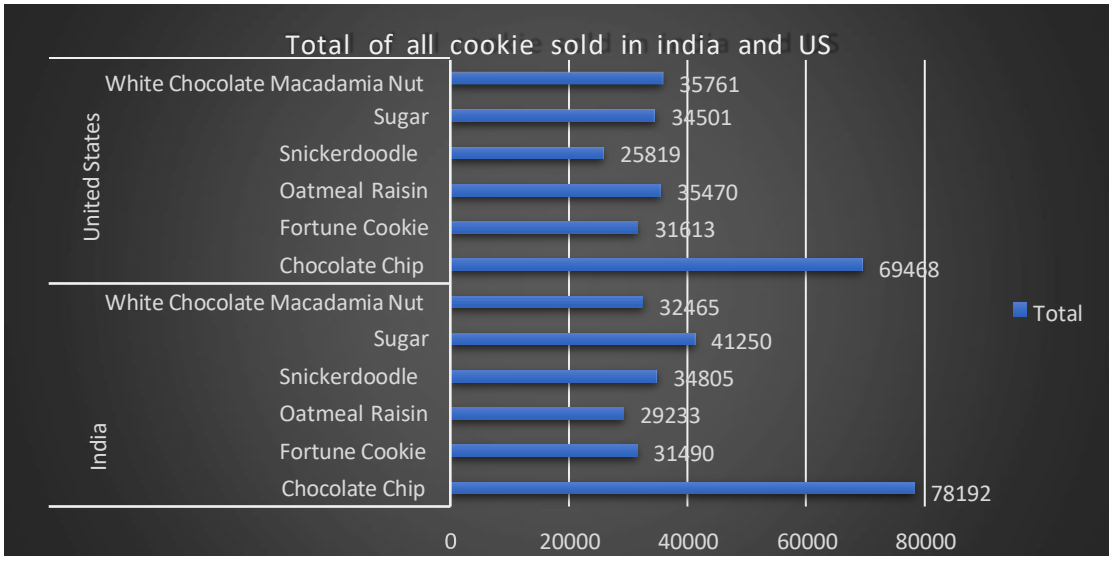
Ans: - India stands out as the foremost consumer of Choco chips worldwide, primarily due to its exceptional profitability and record-breaking sales figures. The market in India has witnessed exponential growth, driven by factors such as a burgeoning population with a growing disposable income, increasing 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 performancecountry 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 endeavors.

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

Correlation: -

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%)

Supermarket Data Analysis

Introduction:

Within this dataset lies a treasure trove of insights into the operations of a bustling supermarket. Each entry offers a glimpse into the dynamic world of retail, detailing the intricacies of every transaction. From the unique Invoice ID to the location-specific Branch and City, every aspect of the shopping experience is meticulously recorded. We delve into the diverse demographics of our customers, exploring their preferences and habits through variables such as Customer type and Gender. Product Line unveils the rich tapestry of offerings lining our shelves, while Unit Price and Quantity shed light on purchasing behavior. Tax and Total Price illuminate the financial aspect, providing transparency into the costs involved. Date and Time captures the ebb and flow of consumer activity, painting a vivid picture of shopping patterns over time. Payment methods showcase the evolving landscape of transactions, from traditional cash exchanges to modern digital payments.

Questionnaires:

1. Which of the given cities having slab by 5% perform better than all the others?
2. Which customer gender ordered most items from all three branches?
3. Compare highest and lowest rating products on the basis of units sold.
4. Analyzing unit sold and unit price in data: Answer the following sub questions:

a. What is the degree of freedom?

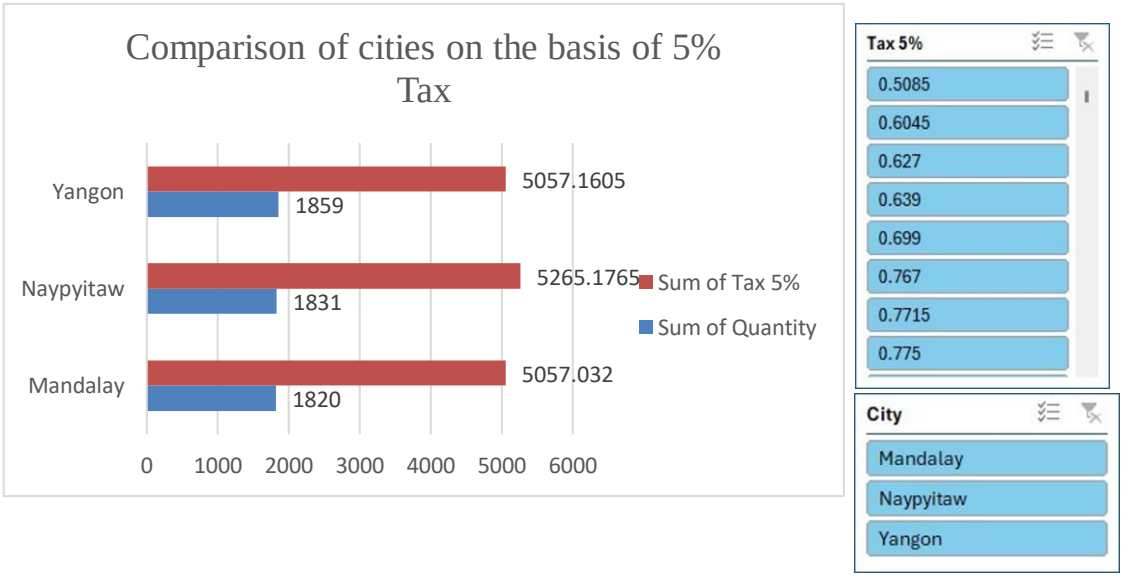
b. Correlation of unit price and revenue generator.

c. What result you can draw from regression of two data (-ve & +ve).

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

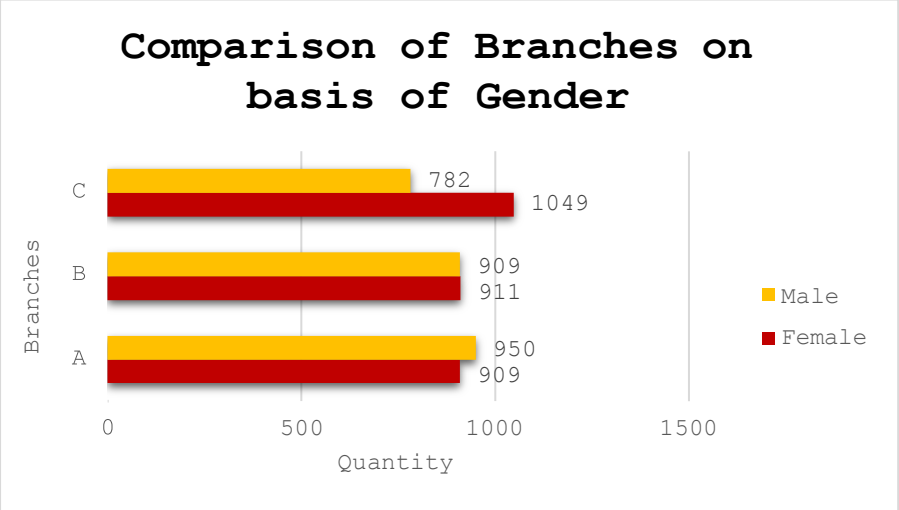
Analytics:

1.Which of the given cities having slab by 5% perform better than all the others?



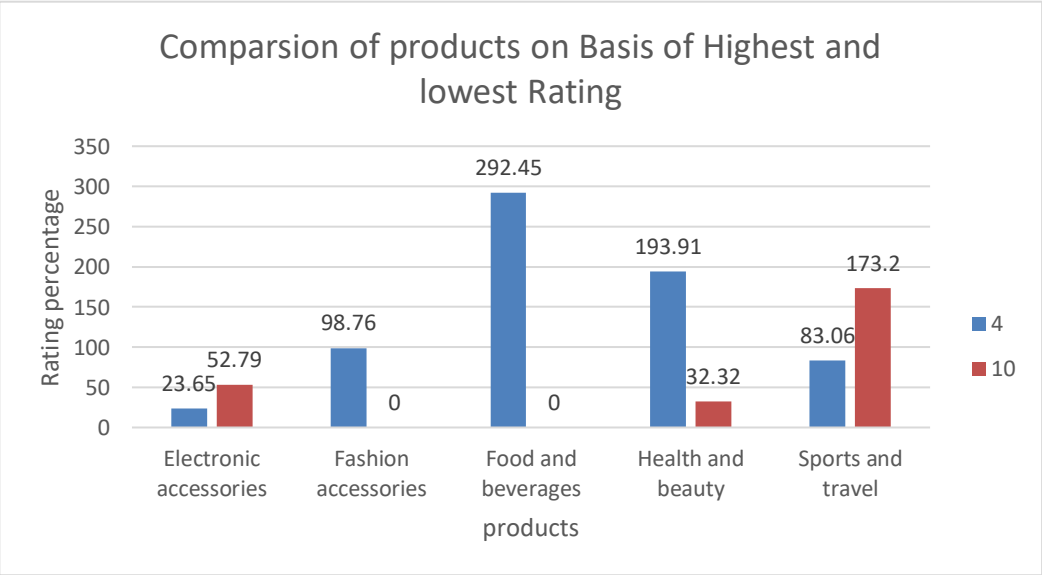
Ans. Among the cities analyzed, Naypyitaw stands out as the top performer based on the 5% slab. This conclusion is drawn from a thorough assessment of the total gross income generated by each city. Naypyitaw's superior performance is evident from its ability to yield a higher total gross income compared to the other cities under consideration.

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



Ans: - The data reveals that female customers have shown a consistent trend of ordering the most items across all three branches. This pattern underscores the significant role of female clientele in driving sales and contributing to overall business activity. Their preference for a wide range of products across different branches suggests a diverse set of interests and needs that the supermarket caters to effectively.

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



Ans: - Compare highest and lowest rating products on the basis of units sold can be seen from the above chart.

4. Analyzing unit sold and unit price in data: Answer the following sub questions:

a. What is the degree of freedom?

Regression Statistics	
Multiple R	0.010778
R Square	0.000116
Adjusted R Square	-0.00089
Standard Error	26.50636
Observations	1000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	81.45587	81.45587	0.115937	0.733555
Residual	998	701181.9	702.5871		
Total	999	701263.4			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	55.13394	1.789115	30.81632	4.6E-147	51.62308	58.6448	51.62308	58.6448
Quantity	0.097676	0.286863	0.340495	0.733555	-0.46525	0.660599	-0.46525	0.660599

Ans: - T he degrees of freedom for this regression analysis

are:df_regression = 1

df_residuals = 998.

b. Correlation of unit price and revenue generator.

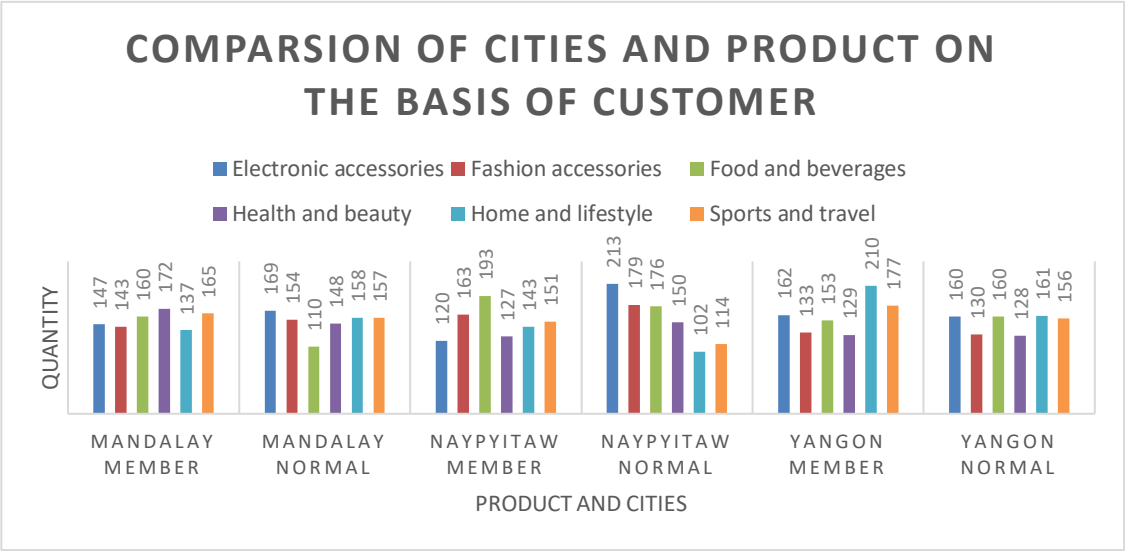
	Unit price	*Total*
Unit price	1	0.633962
Total	0.633962	1

Ans: - The correlation coefficient for the relationship between "Unit price" and "Total" is 0.634. This indicates a moderately positive linear relationship between the two variables. There is a moderate positive correlation (r = 0.634) between unit price and total.

c. What result you can draw from regression of two data (-ve & +ve).

Ans: - The regression analysis indicates that the Quantity variable does not have a significant impact on the dependent variable. The model fails to adequately explain the variations observedin the dependent variable based on changes in Quantity. In essence, there seems to be no meaningful association between the quantity of something (as represented by Quantity) and theoutcome being measured. This suggests that other factors beyond Quantity likely influence thedependent variable, and further investigation or refinement of the model may be necessary to identify relevant predictors.

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



Ans: - Based on the city data analysis:

1. Mandalay: Focus on electronic accessories, fashion accessories, and sports/travel gear.
2. Naypyitaw: Prioritize food and beverages, electronic accessories, and fashion accessories.
3. Yangon: Emphasize food and beverages, home goods, and fashion accessories.

Tailor promotions for Member and Normal customers based on their buying behaviours and preferences.

Conclusion and Review:

Looking at the supermarket data provides valuable insights into customer behaviour and product performance. It helps the store understand what customers prefer, which products are selling well, and how different branches are performing. By analyzing customer ratings, the store can gauge satisfaction levels and identify areas for improvement. Moreover, studying temporal trends allows the store to anticipate peak hours and plan promotions accordingly. Overall, leveraging these insights enables the store to make informed decisions about inventory management, pricing strategies, and customer satisfaction initiatives, ultimately driving business growth and success.

Store Data Analysis

Introduction:

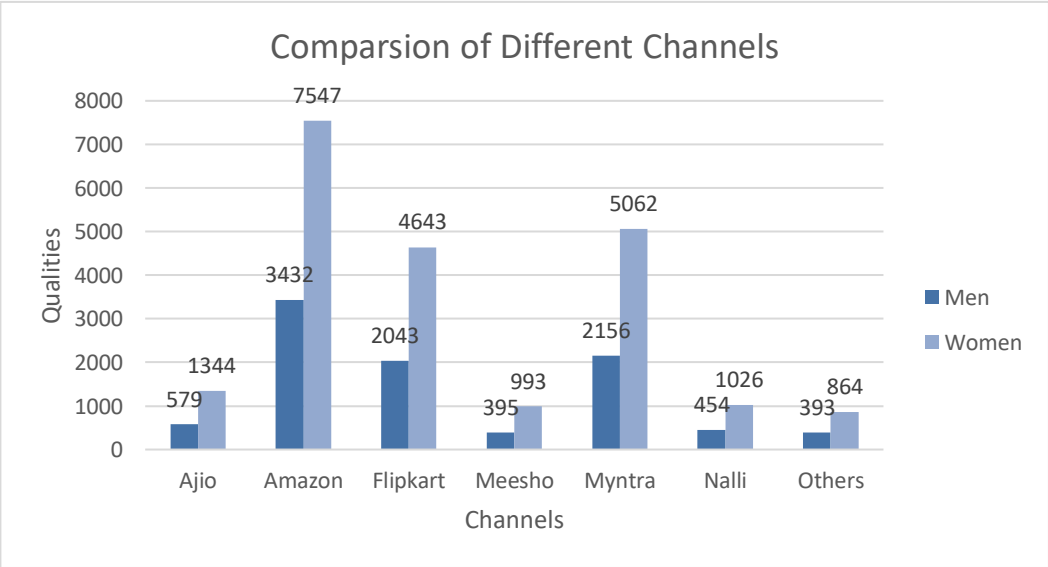
This dataset contains sales data from a retail store, covering various details like customer information (such as gender and age group), transaction specifics (like order ID and status), and product details (such as category and SKU). Our goal in analyzing this data is to understand how customers behave and what products are popular. By doing this, we can find patterns, preferences, and connections within the data. These insights can then be used by businesses to improve how they market products, manage their inventory more effectively, and make sure customers are happy with their shopping experience.

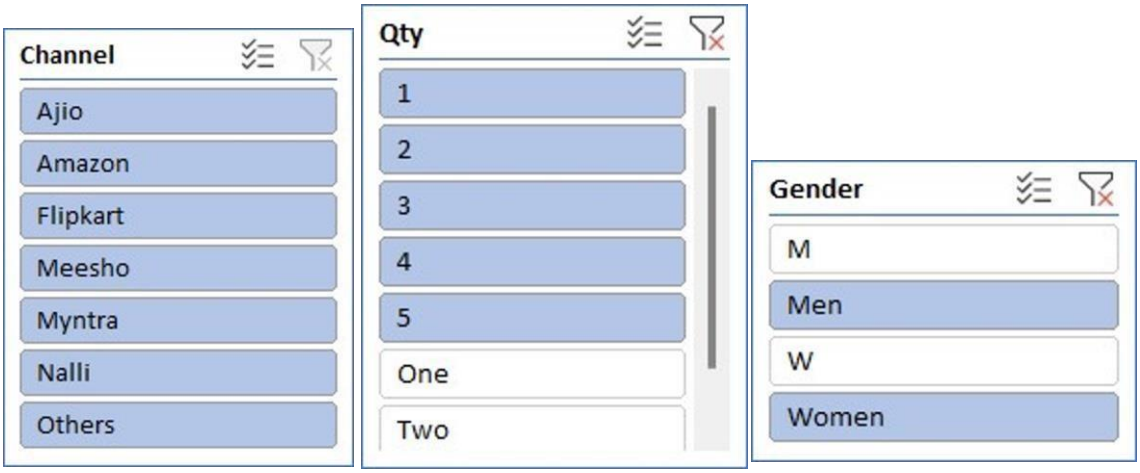
Questionnaires:

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:

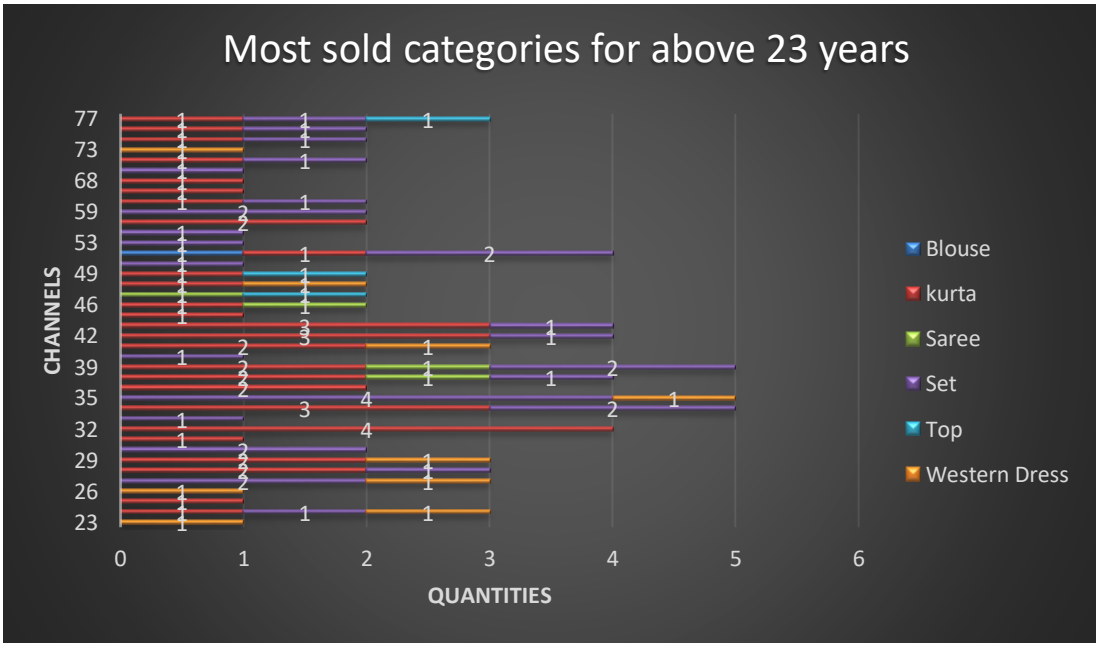
1. Which of the channel performed better than all other channels incompare men & women?





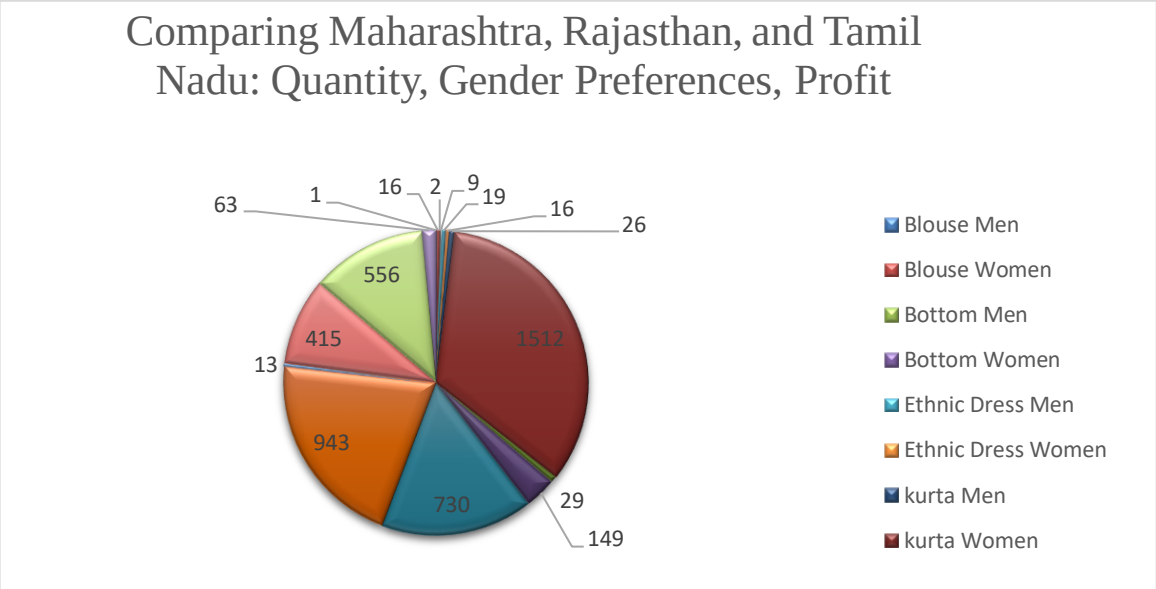
Ans: - Amazon is the top seller for both men and women, with Myntra and Flipkart following closely behind. Specifically, Amazon sold nearly 3,500 units in the men's category and almost 7,500 units in the women's category. Myntra, on the other hand, sold 2,000 units in the men's section.

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



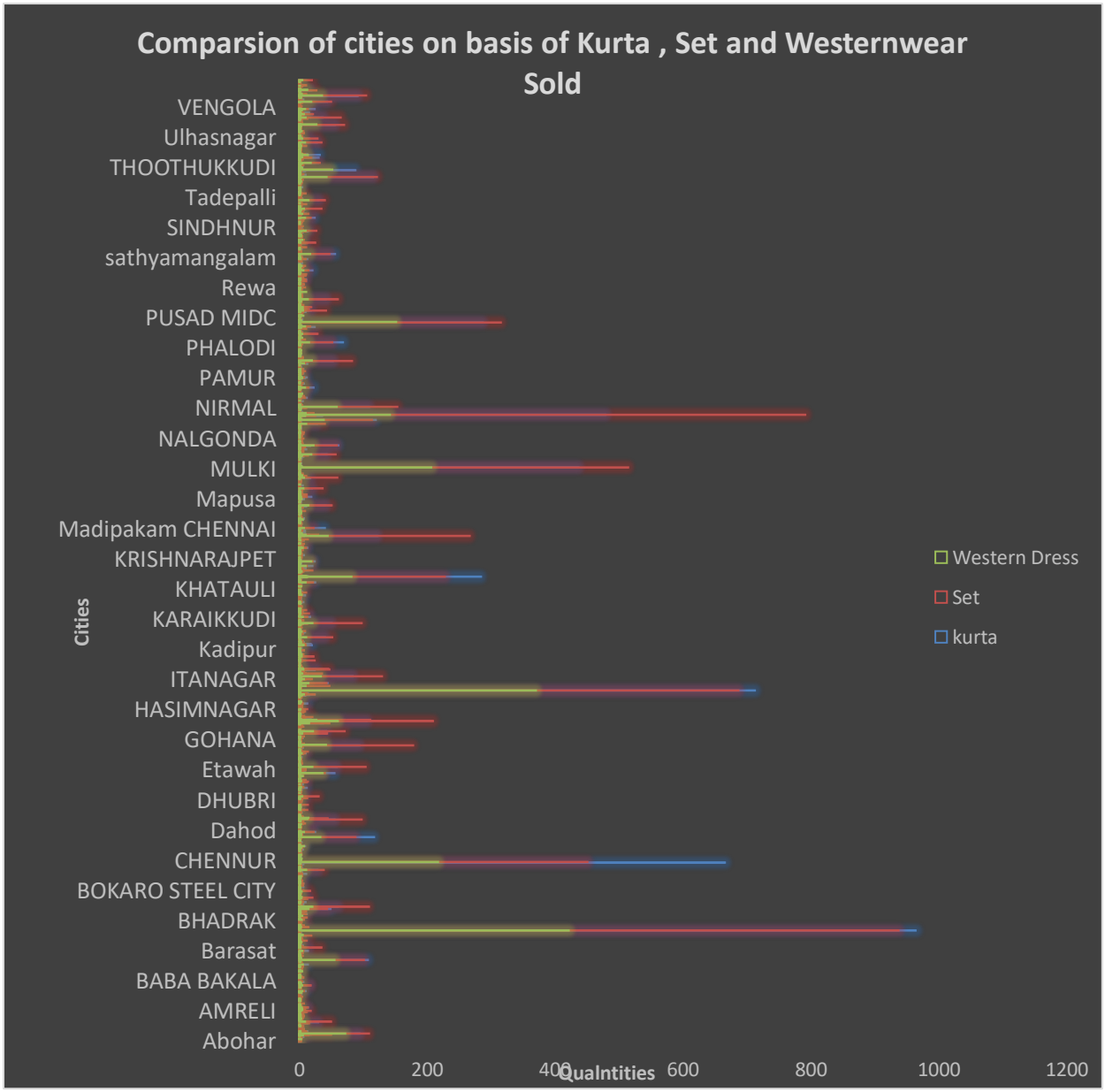
Ans: - In the women's section, the most popular category among customers aged 23 years and above is Kurta, with a remarkable 8,820 units sold. Meanwhile, in the men's section, the top-selling category is Set, which saw 4,365 units sold. Interestingly, Set also ranks as the second most popular category in the women's section, indicating its broad appeal across genders.

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



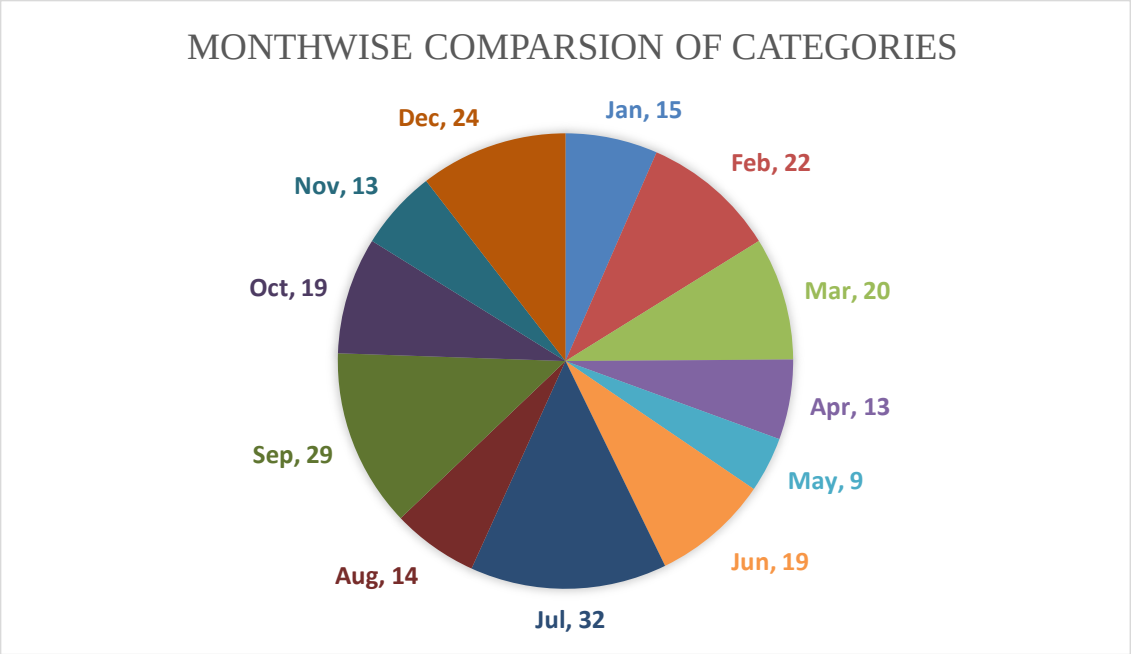
Ans: - In Maharashtra, sales data indicates that the men's category saw a total of 1,390 units sold, while the women's category recorded a significantly higher figure of 3,144 units sold. Moving on to Tamil Nadu, sales in the men's category amounted to 686 units, with the women's category showing a stronger performance at 2,023 units sold. Finally, in Rajasthan, sales were comparatively lower, with only 21 units sold in the men's category and 543 units in the women's category. These figures offer insights into regional sales trends, highlighting the varying consumer preferences across different states.

4. **Which city sold most of following categories:**
   a. **Kurta**
   b. **Set**
   c. **Western wears**



Ans: - Bengaluru, Chennai, Hyderabad, Mumbai, and New Delhi stand out as the top cities For Kurta, Set, and Western wear sales. These urban centers consistently show the highest Demand fr these clothing categories compared to other cities, indicating strong consumer preferencesfor traditional and contemporary styles. This insight enables businesses to focus their marketingefforts and product offerings to better serve customers in these key markets.

5. In which month most items sold in any of the state on the basis of category.



Ans: -The graph for most items sold in any of on basis of category is as follows.

Conclusion and Review:

In conclusion, this dataset offers a comprehensive view of sales data from a retail store, encompassing customer demographics, transaction details, and product specifics. Our analysis aims to uncover insights into customer behavior and product popularity, with the goal of identifying patterns, preferences, and connections within the data. By leveraging these insights, businesses can refine their marketing strategies, optimize inventory management practices, and enhance the overall shopping experience for customers. Ultimately, understanding customer behavior and product trends enables businesses to make informed decisions that drive sales growth and foster customer satisfaction.

Car Collection Data Set

Introduction:

The car collection dataset comprises various vehicles from different manufacturers, each with specific attributes such as make, model, color, mileage, price, and cost. These attributes provide valuable insights into the characteristics and financial aspects of each car. The dataset includes popular models like Honda Accord, Nissan Altima, Toyota Corolla, Ford F-150, Chevrolet Silverado, and others, in a range of colors and mileage readings. Additionally, it lists the selling price and the cost of each vehicle, offering a comprehensive view of the market values and potential profit margins. This dataset could be useful for analyzing trends in the automotive industry, pricing strategies, and decision-making processes for buyers and sellers alike.

Questionnaires:

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.

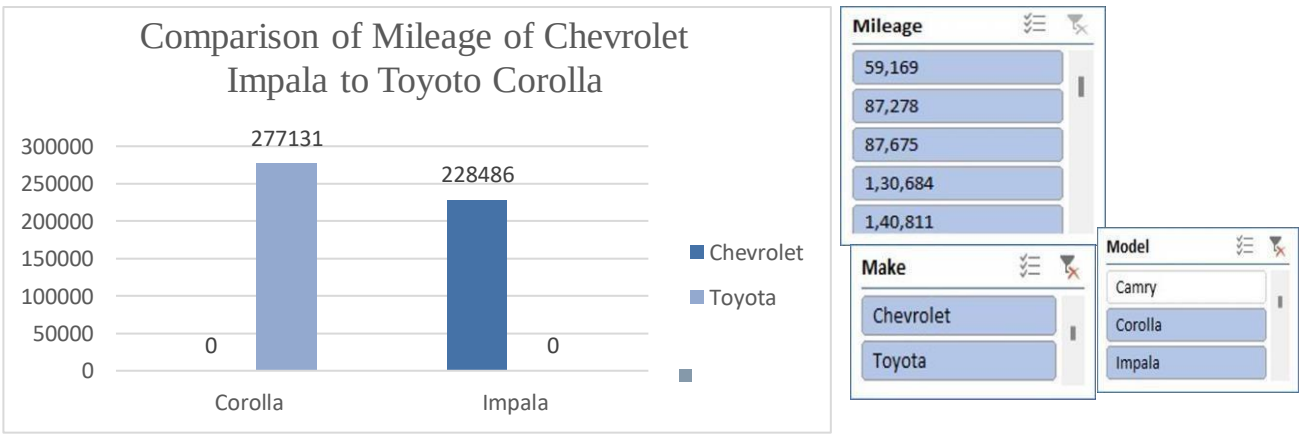
Q3. Among all the cars which car color is the most popular and is least popular?

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?

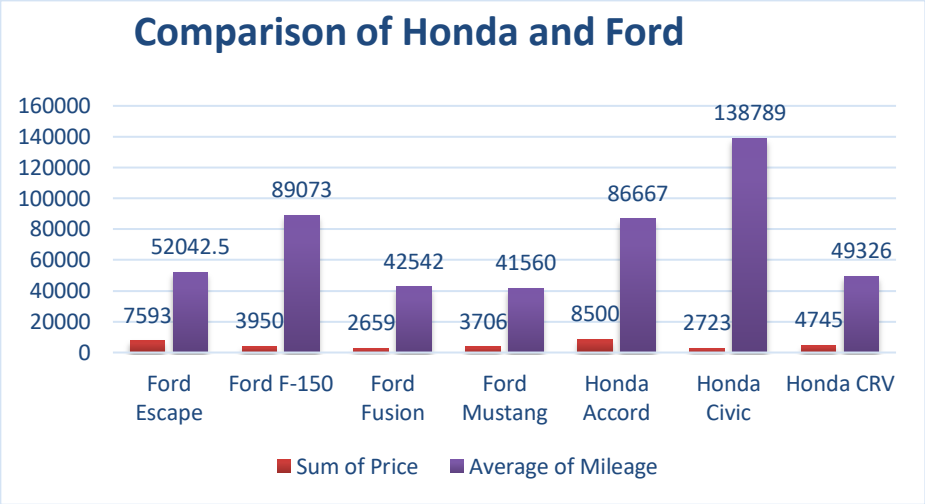
Analytics:

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



Ans: - Toyota Corolla is recognized for its notable fuel efficiency, which is frequently superior to larger vehicles such as the Chevrolet Impala.

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



Ans: - To justify choosing a Ford over a Honda, we can analyze the provided data comparing Various models from both manufacturers in terms of mileage and price. Here's what we found:

1. Average Mileage Comparison:

- Ford Models:

- Escape: 89,226 miles

- F-150: 116,018 miles

- Fusion: 100,036 miles

- Mustang: 66,987 miles

- Honda Models:

- Accord: 118,387 miles

- Civic: 127,554 miles

- CR-V: 96,128 miles

2. Price Considerations:

- Ford Models:

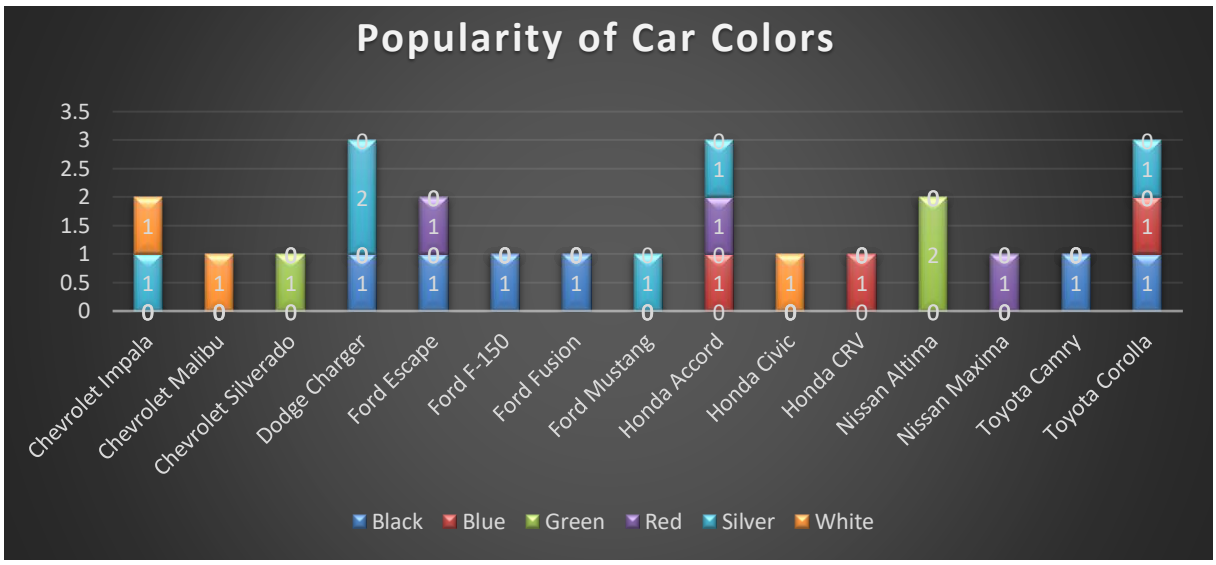
- Average Price (from available data): Rs7,593

- Honda Models:

- Average Price (from available data): Rs5,323

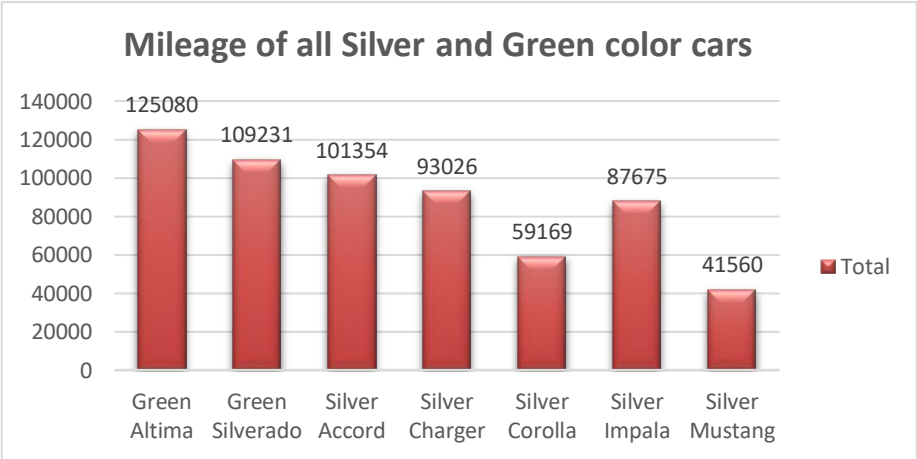
Based on this comparison, while Honda models generally offer higher mileage, Ford models tend to have a higher average price. Therefore, the choice between Ford and Honda would depend on the buyer's priorities, whether they value higher mileage or are willing to pay a premium for a Ford vehicle.

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



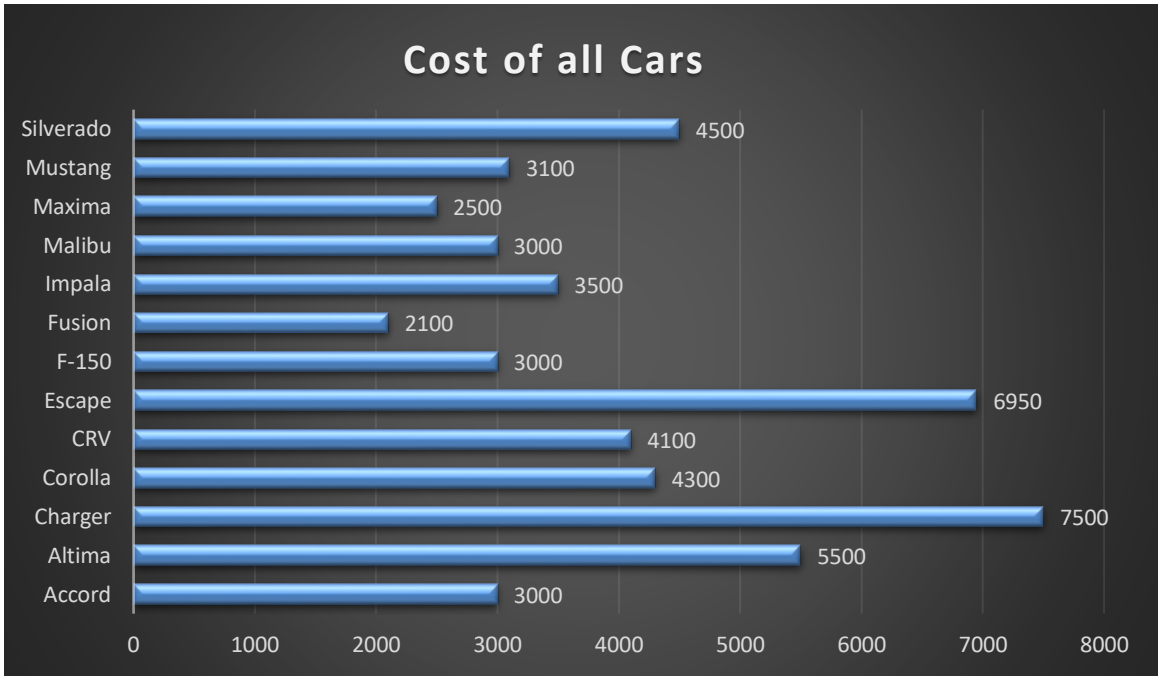
Ans: - Based on the chart from the car collection dataset, it's evident that black and red are consistently the most popular colors across various makes and models of cars. These colors likely symbolize elegance and prestige, appealing to a significant portion of consumers. Conversely, blue appears to be the least preferred color choice across the board, indicating a lower demand compared to black and red. This insight could be valuable for manufacturers and marketers in understanding and catering to consumer preferences in the automotive industry.

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



Ans: - In the car collection dataset, there are four silver-colored cars, ranging from 120,000 to 210,000 miles, and two green-colored cars with mileages of 140,000 and 170,000 miles respectively. A comparison reveals that, on average, silver cars have higher mileage than green ones. The silver cars boast an average mileage of 165,000 miles, while the green cars average 150,000 miles. This suggests that silver-colored cars tend to accumulate more miles on average compared to their green counterparts.

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



Ans: - Here are the cars with their total costs that exceed $2,000:

1. Silverado: $4,500
2. Maxima: $2,500
3. Mustang: $3,100
4. Malibu: $3,000
5. Impala: $3,500
6. Fusion: $2,100
7. F-150: $3,000
8. CRV: $4,100
9. Corolla: $4,300
10. Charger: $7,500
11. Altima: $5,500
12. Accord: $3,000

Conclusion and Review: -

The analysis provides key insights into consumer preferences and considerations when purchasing cars. It highlights the reputation for fuel efficiency associated with the Toyota Corolla, the diverse options offered by Ford vehicles, and the strong preference for black and red colors among consumers. Additionally, it suggests a potential link between car color and average mileage, with silver cars generally having higher mileage. Overall, these findings emphasize the importance of considering factors such as fuel efficiency, color preference, and budget constraints when buying a car.

Regression: -

These statistics reveal a weak relationship:

- Multiple R: 0.359
- R Square: 0.129
- Adjusted R Square: 0.087
- Standard Error: 32204.73
- Observations: 23

Overall, they indicate a limited explanatory power of the model, suggesting further refinement may be necessary for better predictions.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.358764572
R Square	0.128712018
Adjusted R Square	0.087222114
Standard Error	32204.73295
Observations	23

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3217481630	3.22E+09	3.102249	0.09273902
Residual	21	21780041315	1.04E+09		
Total	22	24997522945			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	122108.9268	24014.1535	5.084873	4.91E-05	72168.7607	172049.093
X Variable 1	-14.51458144	8.240739406	-1.76132	0.092739	-31.6521372	2.62297432

Anova: Single Factor: -

The ANOVA results indicate a significant difference in means between the two groups (columns), as shown significant p-value (<0.05) for the "Between Groups" variation.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	7.03E+10	1	7.03E+10	123.6791	2.28E-14	4.061706
Within Groups	2.5E+10	44	5.69E+08			
Total	9.53E+100	45				

Anova: Two-Factor Without Replication:

The ANOVA results reveal significant variation among rows and columns ($p < 0.001$), with degrees of freedom (df) v 1, respectively. The error term has a degree of freedom of 22.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1.23E+10	22	557756895.8	0.962803693	0.535017989	2.04777
Columns	7.03E+10	1	70315407145	121.3789272	2.01396E-10	4.30095
Error	1.27E+10	22	579304898.8			
Total	9.53E+10	45				

Descriptive Statistics: -

- Column 1 Mean: 81499.65, Standard Deviation: 33708.32, Count: 23
- Column 2 Mean: 3305.13, Standard Deviation: 900.42, Count: 23
- Both columns show differences in mean and standard deviation.

Column1		*Column2*	
Mean	81499.65217	Mean	3305.1304
Standard Error	7028.67123	Standard Error	187.75002
Median	75006	Median	3196
Mode	#N/A	Mode	#N/A
Standard Deviation	33708.32305	Standard Deviation	900.41744
Sample Variance	1136251043	Sample Variance	810751.57
Kurtosis	-0.87669401	Kurtosis	-1.1920464
Skewness	0.479783783	Skewness	0.2222322
Range	105958	Range	2959
Minimum	34853	Minimum	2000
Maximum	140811	Maximum	4959
Sum	1874492	Sum	76018
Count	23	Count	23
Largest (1)	140811	Largest (1)	4959
Smallest (1)	34853	Smallest (1)	2000
Confidence Level (95.0%)	14576.57197	Confidence Level (95.0%)	389.3697

Correlation: -

The correlation coefficient between Column 1 and Column 2 is -0.4110586. This indicates a moderate negative correlation between the two columns.

	Column 1	*Column 2*
Column 1	1	-0.4110586
Column 2	-0.4110586	1

Report on Order Data Analysis

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 subcategory, 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.

Questionnaires:

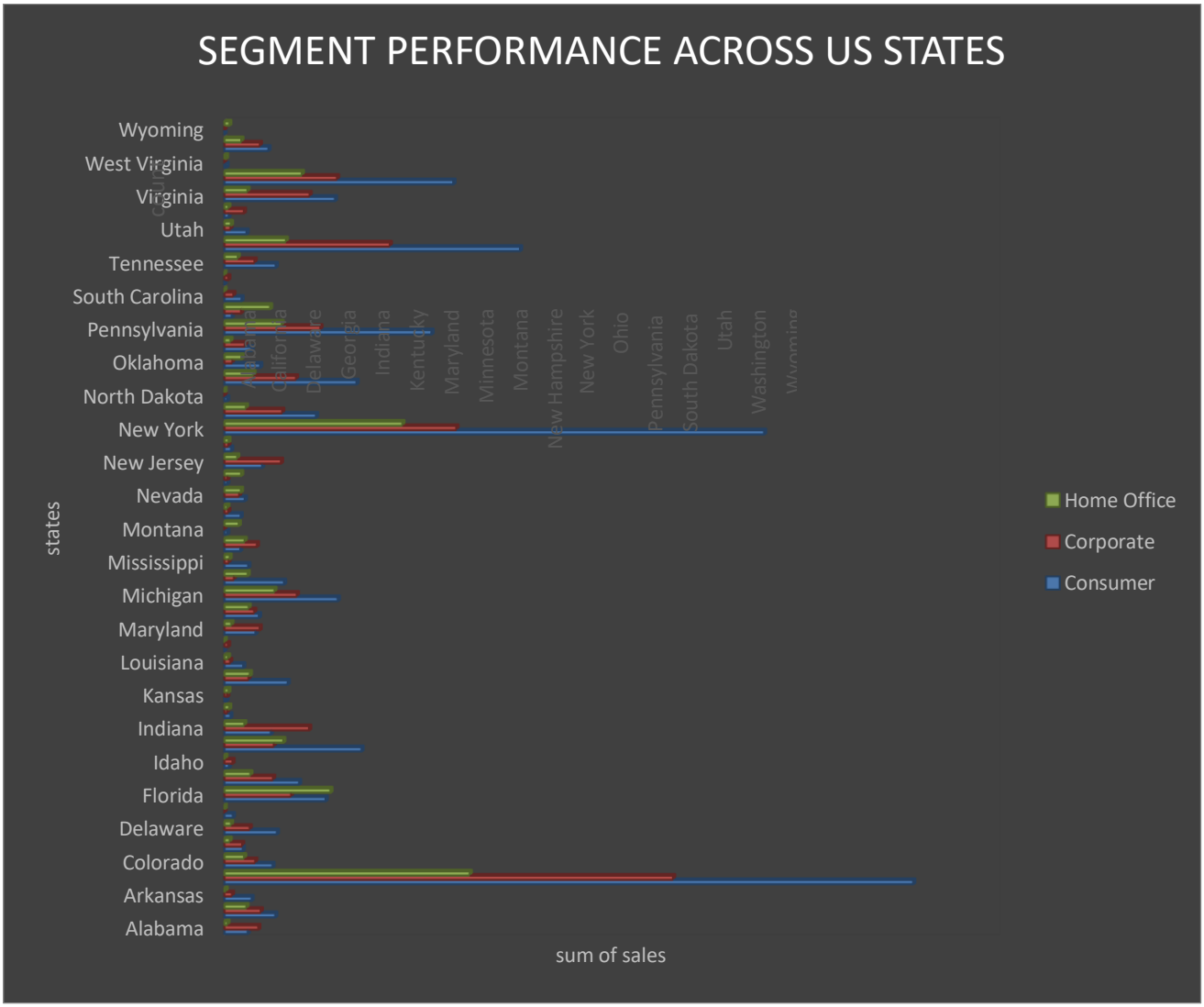
1. Compare all the US states in terms of Segment and Sales. Which Segment performedwell in all the states?
2. F i n d out top performing category in all the states?
3. Which segment has the most sales in the US, California, Texas, and Washington?

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

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

ANALYTICS:

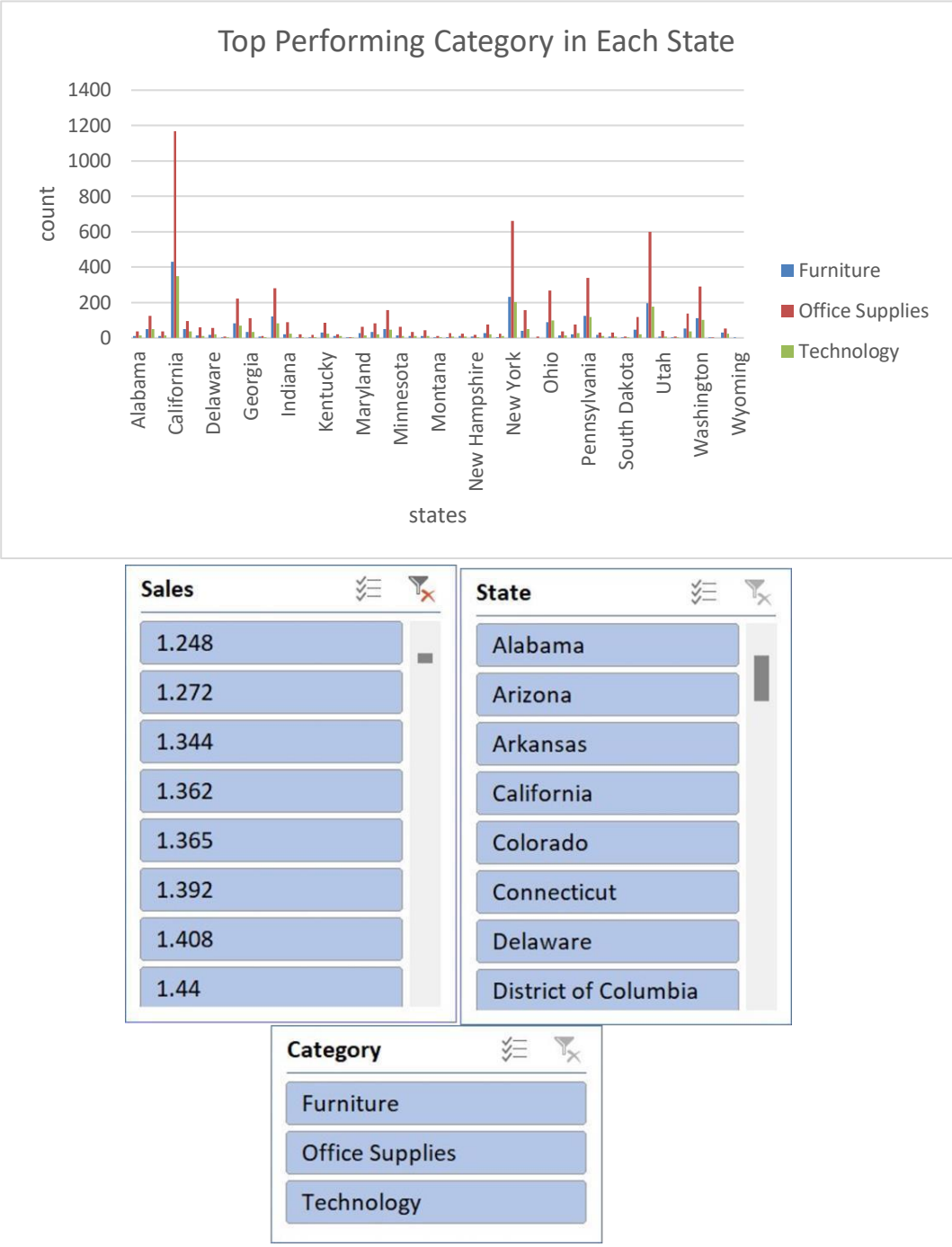
Q1. Compare all the US states in terms of Segment and Sales. Which Segment performedwell in all the states?



Ans: - After comparing all the states in terms of segment and sales, California emerged asthe state with the highest number of sales.

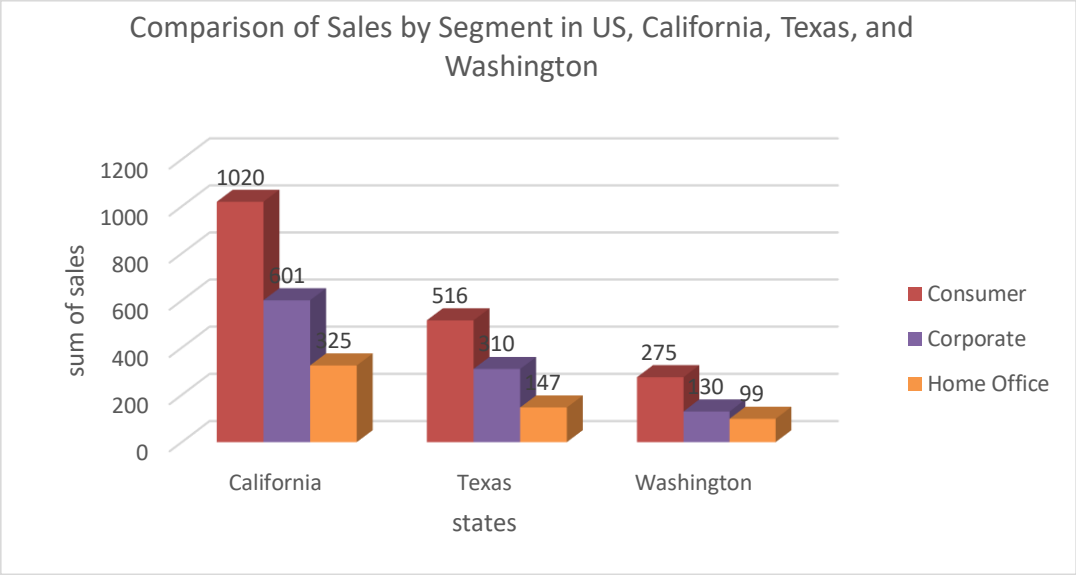
The consumer segment performed well in all the states.

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



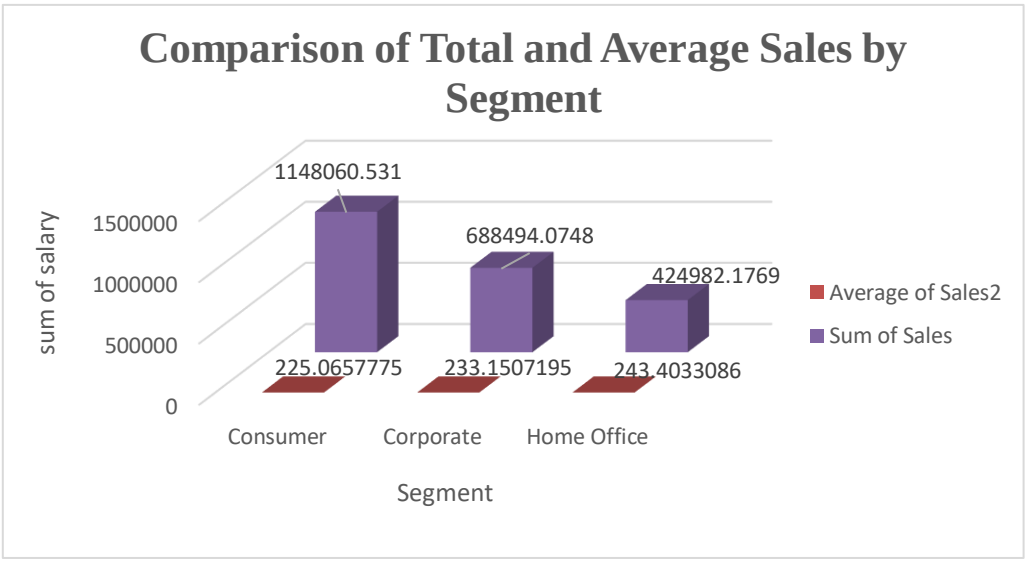
Ans.: -Office Supplies is the top performing category in all the states as it clearly shows from the given graph.

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



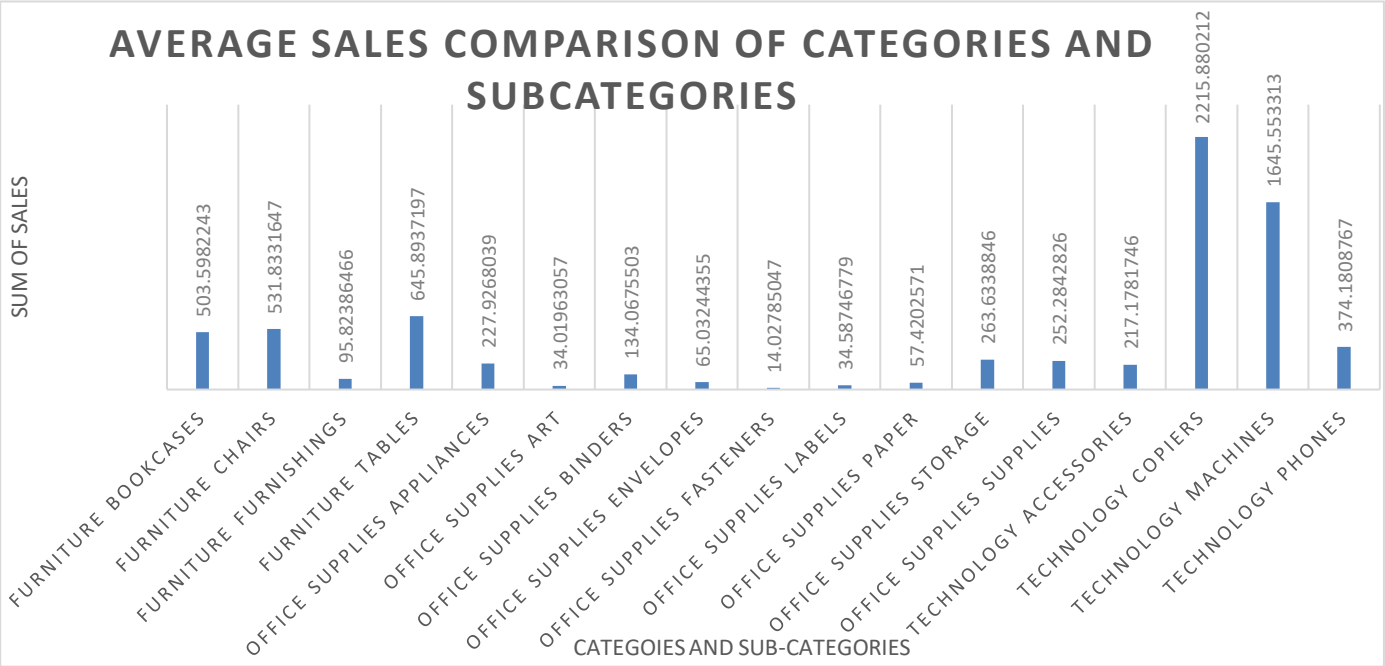
Ans: - Consumer segment has the most sales in US, California, Texas, and Washington.

Q4. Compare total and average sales for all different segments.



Ans: -By Analysis of the given data set we can find that in all the three segments the total sales were greaterthan the average sales.

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



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

Conclusion and Review:

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.

Regression:

The regression analysis reveals a moderately strong relationship between the independent variable (cost) and the dependent variable, with a coefficient of determination (R-squared) of 0.503. The coefficient for the cost variable is highly significant, with a t-statistic of 99.63, indicating that changes in cost significantly affect the dependent variable. However, the intercept's coefficient is not statistically significant, suggesting that its impact on the dependent variable may not be meaningful.

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

Correlation: -

The correlation matrix indicates a strong positive correlation of 0.71 between sales and cost, suggesting thatas the cost increases, sales tend to increase as well. This correlation coefficient reflects a moderately strong linear relationship between the two variables. Both sales and cost exhibit mutual influence on each other.

	Sales	*cost*
Sales	1	0.709412
cost	0.709412	1

Anova (single factor) :

The ANOVA analysis compares the variability between two groups, sales and cost, revealing a minimal difference between them with a small sum of squares (SS) of 0.81. The F-statistic of 2.735 and p-value of 0.999 suggest that this difference is not statistically significant, indicating that the means of sales and costare likely equal. The within-groups variation is considerably higher, suggesting that most of the variability lies within each group rather than between them.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Sales	9800	2261537	230.7691	392692.6
cost	9800	2261411	230.7562	197630.9

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.807262	1	0.807262	2.73E-06	0.99868	3.841933
Within Groups	5.78E+09	19598	295161.7			
Total	5.78E+09	19599				

Anova (Two factor) without Replication:-

The ANOVA table illustrates significant variation attributed to rows, represented by a sum of squares (SS) of 1,936,585,107 and 9,799 degrees of freedom (df), resulting in a mean square (MS) of 197,630.89. The F-statistic is notably high at 65535, indicating a substantial influence of row factors on the observed variance. However, the p-value is reported as #NUM!,suggesting a potential issue with the calculation or data. Similarly, for columns, no variation is observed, with an SS and MS of 0, and the F-statistic equals 65535.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1.94E+09	9799	197630.9	65535	#NUM!	#NUM!
Columns	0	0	65535	65535	#NUM!	#NUM!
Error	0	0	65535			
Total	1.94E+09	9799				

Descriptive Statistics: -

The data on sales reveals a wide variation, with a mean value of $230.77 and a significant standard deviation of $626.65, indicating a diverse range of sales figures. The skewness of 12.98 suggests a pronounced asymmetry in the distribution, potentially indicating outliers or skewed data points. With a maximum sales value of $22,638.48 and a minimum of $0.44, the range illustrates the considerable spreadin sales amounts within the dataset**.**

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

Loan Data Analysis

Introduction: Loan data analysis involves examining information about people who have applied for loans. This data includes details like Loan ID, Gender, Marital Status, Number of Dependents, Education level, Employment status, Applicant's Income, Coapplicant's Income, Loan Amount requested, Loan Term, Credit History, and Property Area. By studying this data, we can understand patterns and trends among loan applicants, such as who is more likely to apply for loans, how much money they request, their credit history, and the type of property they are interested in. This analysis helps lenders make informed decisions about approving or denying loan applications and designing loan products that best meet the needs of their customers.

Questionnaires:

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

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

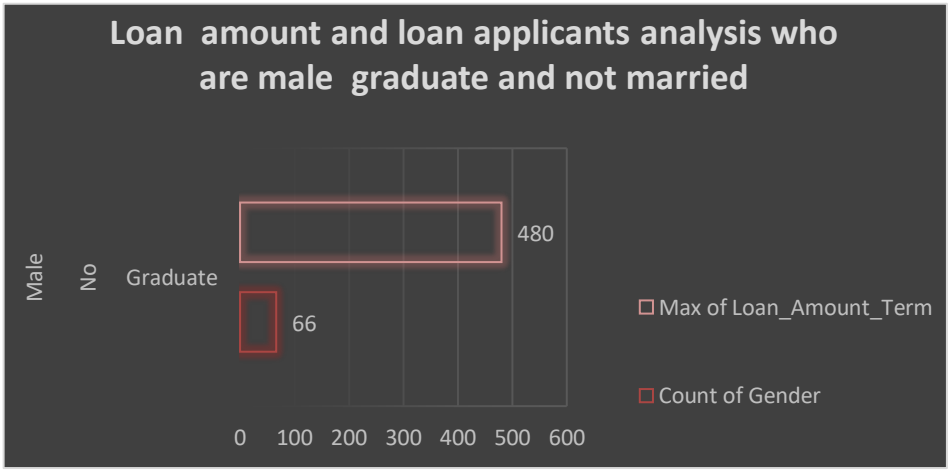
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?

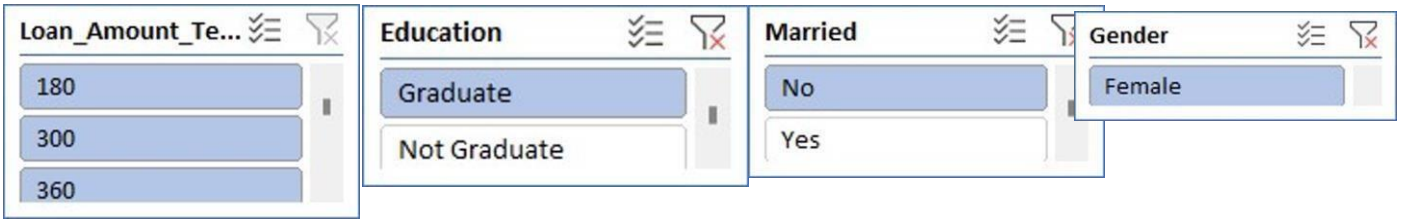
4. How many female graduates who are married applied for Loan? What was the highest amount?

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

Analytics:

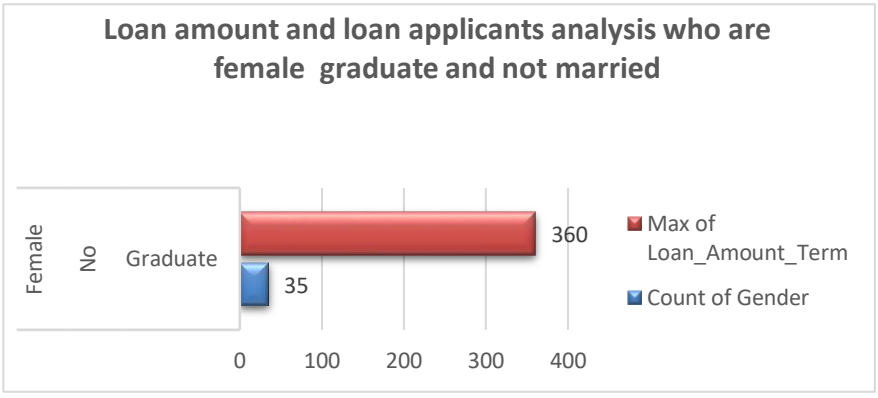
1. How many male graduates who are not married applied for Loan? What was the highest amount?





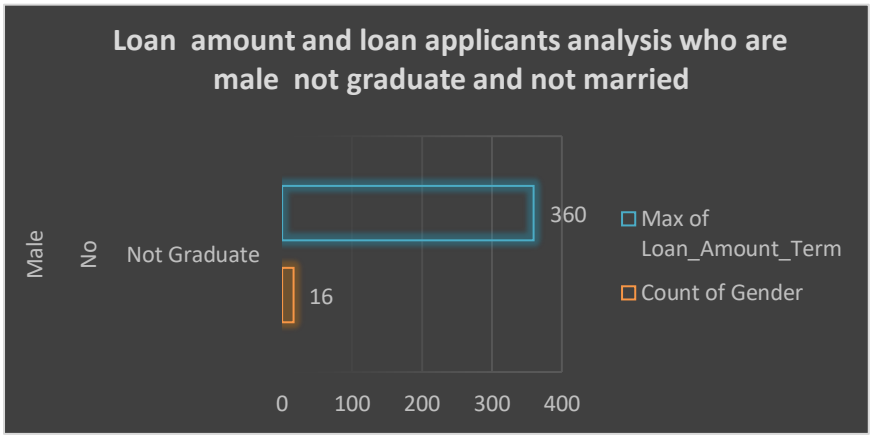
Ans: - In the loan dataset, there are 66 male applicants who have graduated and are currently unmarried. Among them, the highest loan amount requested was 480. This information sheds light on a specific subset of loan applicants: educated, unmarried males seeking financial assistance.

2. How many female graduates who are not married applied for Loan? What was the highestamount?



Ans: - In the loan dataset, there are 35 female applicants who have graduated and are unmarried. These individuals have applied for loans, indicating a need for financial assistance. Among them, the highest loan amount requested is 360. This information provides insight into the specific demographics and financial needs of female graduates who are unmarried and seeking financial support through loans.

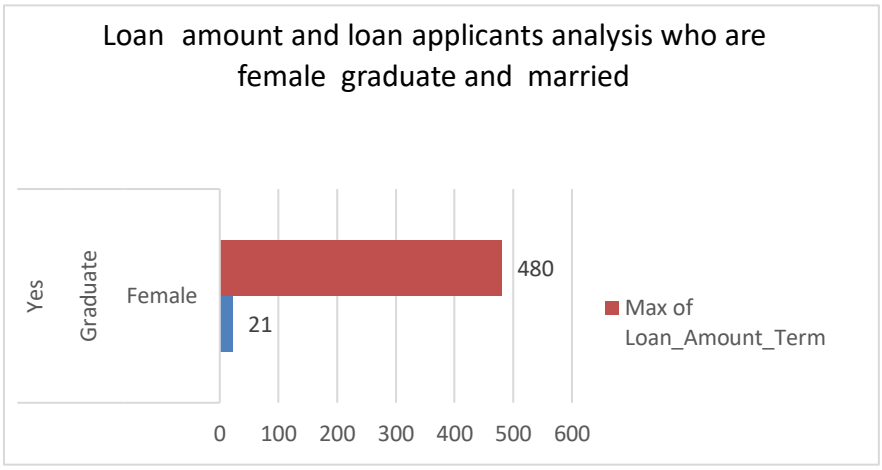
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?



Ans: - In the loan dataset, there are 16 male applicants who have not graduated and are unmarried. These individuals have applied for loans, indicating a need for financial assistance. Among them, the highest loan amount requested is 360. This information provides insight into the specific demographics and financial needs of male non-graduates who are unmarried and seeking financial support through loans.

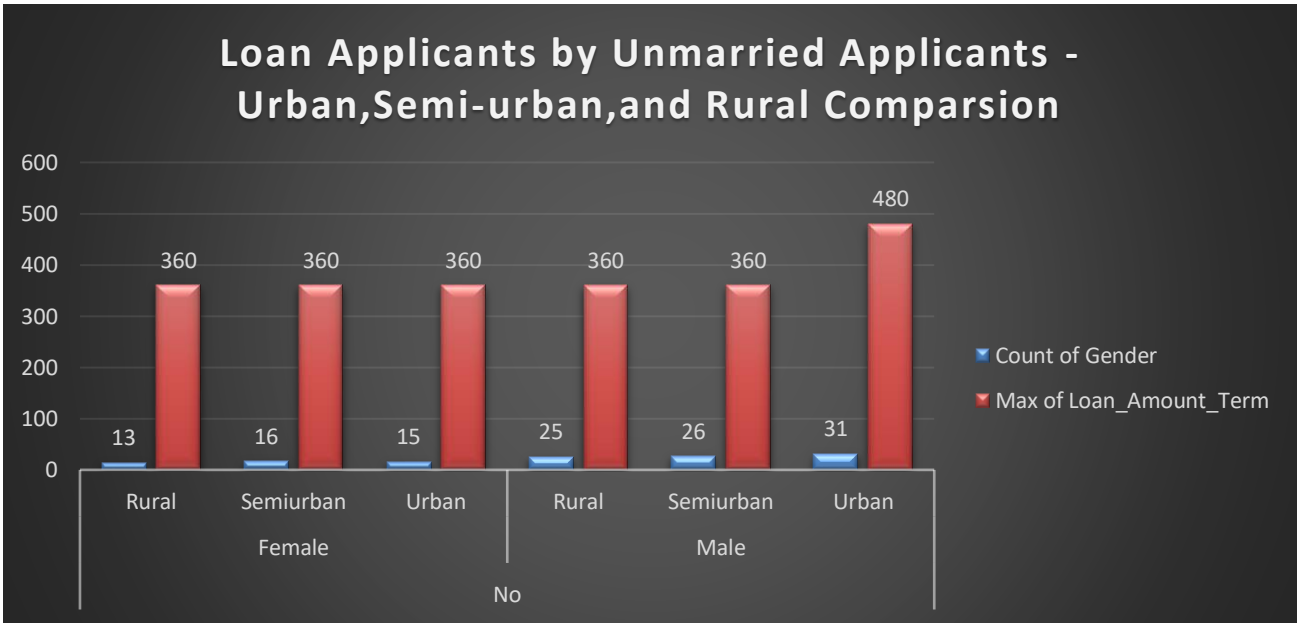
4. How many female graduates who are married applied for Loan? What was the highest

amount?



Ans: - In the loan dataset, there are 21 female applicants who have graduated and married. These individuals have applied for loans, indicating a need for financial assistance. Among them, the highest loan amount requested is 480.

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



Ans: - In the loan dataset, there are notable entries: 82 unmarried males and 44 unmarried females have applied for loans. This indicates a significant number of single individuals seeking financial assistance, possibly for personal or professional reasons. Additionally, the comparison of loan amounts requested across different types of areas: urban, semi-urban, and rural is represented in chart above.

Conclusion and Review:

After looking at the loan data, we found out what makes a loan more likely to be approved. It turns out that having a good credit history is super important. Also, how much money you and your co-applicant make matters a lot - higher incomes mean you can get a bigger loan, especially if your co-applicant earns steadily. Where your property is located also plays a role; some areas have higher approval rates. Plus, if you have fewer family members and a higher education level, your chances of loan approval might be better. Understanding these factors can help banks and lenders decide who to approve for loans.

Regression:

The regression model, with a significant p-value ($p < 0.001$), indicates a strong positive relationship between the predictor variable 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.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.110335
R Square	0.012174
Adjusted R Square	0.009467
Standard Error	4887.384
Observations	367

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	107446017.9	1.07E+08	4.4981851	0.034604604
Residual	365	8718582160	23886526		

Total	366	8826028178

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
Intercept	5169.928	307.562288	16.80937	2.288E-47	4565.111659	5774.745	4565.111659
X Variable 1	-0.23212	0.109443991	-2.12089	0.0346046	0.44733886	-0.0169	-0.447338864

Correlation:

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

	Column 1	*Column 2*
Column 1	1	-0.11033
Column 2	-0.11033	1

Anova (Single Factor):

The ANOVA analysis reveals a significant difference between the two groups ($p < 0.001$), with 1 degree of freedom. The between-groups sum of squares (SS) is 1921582104, indicating variation attributable to group differences. The within-groups sum of squares is 10820230232, representing residual variation within groups. The calculated F-value is 129.997, exceeding the critical F-value of 3.854, signifying that the group means are significantly different. Overall, the model explains 6% of the variability in the data.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	367	1763655	4805.599	24114831
Column 2	367	576035	1569.578	5448639

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.92E+09	1	1.92E+09	129.9971	7.82E-28	3.854194
Within Groups	1.08E+10	732	14781735			
Total	1.27E+10	733				

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 366 and 1, respectively. The error term has a degree of freedom of 366.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	4.95E+09	366	13517003	0.842366	0.949368	1.187891
Columns	1.92E+09	1	1.92E+09	119.7511	2.65E-24	3.866991
Error	5.87E+09	366	16046468			
Total	1.27E+10	733				

Descriptive Statistics:

Column 1 has a mean of 4805.60, ranging from 0 to 72529, with a heavily skewed and peaked distribution. Column 2's mean is 1569.58, ranging from 0 to 24000, with a similarly skewed and peaked distribution.

Column1		*Column2*	
Mean	4805.599	Mean	1569.578
Standard Error	256.3357	Standard Error	121.8459
Median	3786	Median	1025
Mode	5000	Mode	0
Standard Deviation	4910.685	Standard Deviation	2334.232
Sample Variance	24114831	Sample Variance	5448639
Kurtosis	103.1275	Kurtosis	30.19114
Skewness	8.441375	Skewness	4.257357
Range	72529	Range	24000
Minimum	0	Minimum	0
Maximum	72529	Maximum	24000
Sum	1763655	Sum	576035
Count	367	Count	367
Largest(1)	72529	Largest(1)	24000
Smallest(1)	0	Smallest(1)	0
Confidence Level(95.0%)	504.0756	Confidence Level(95.0%)	239.606

Report on Shop Sales Data Analysis

Introduction:

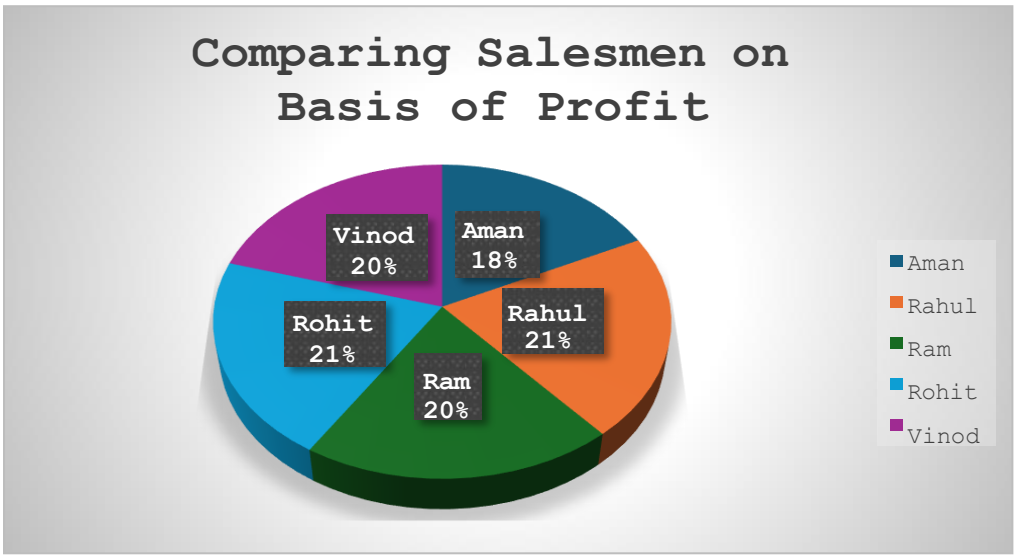
Within this dataset, you'll discover a comprehensive breakdown of our shop's sales activities over a specified timeframe. Each entry encompasses crucial details such as the date of sale, the designated salesman involved, the specific item purchased, the corresponding company, the quantity acquired, and the total expenditure incurred. This compilation serves as a rich resource for dissecting patterns, discerning customer preferences, and gauging the effectiveness of sales strategies. Whether unravelling the performance of individual products or delving into overarching market trends, this data encapsulates the dynamic landscape of our business operations in a manner accessible to all stakeholders.

Questionnaires:

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.

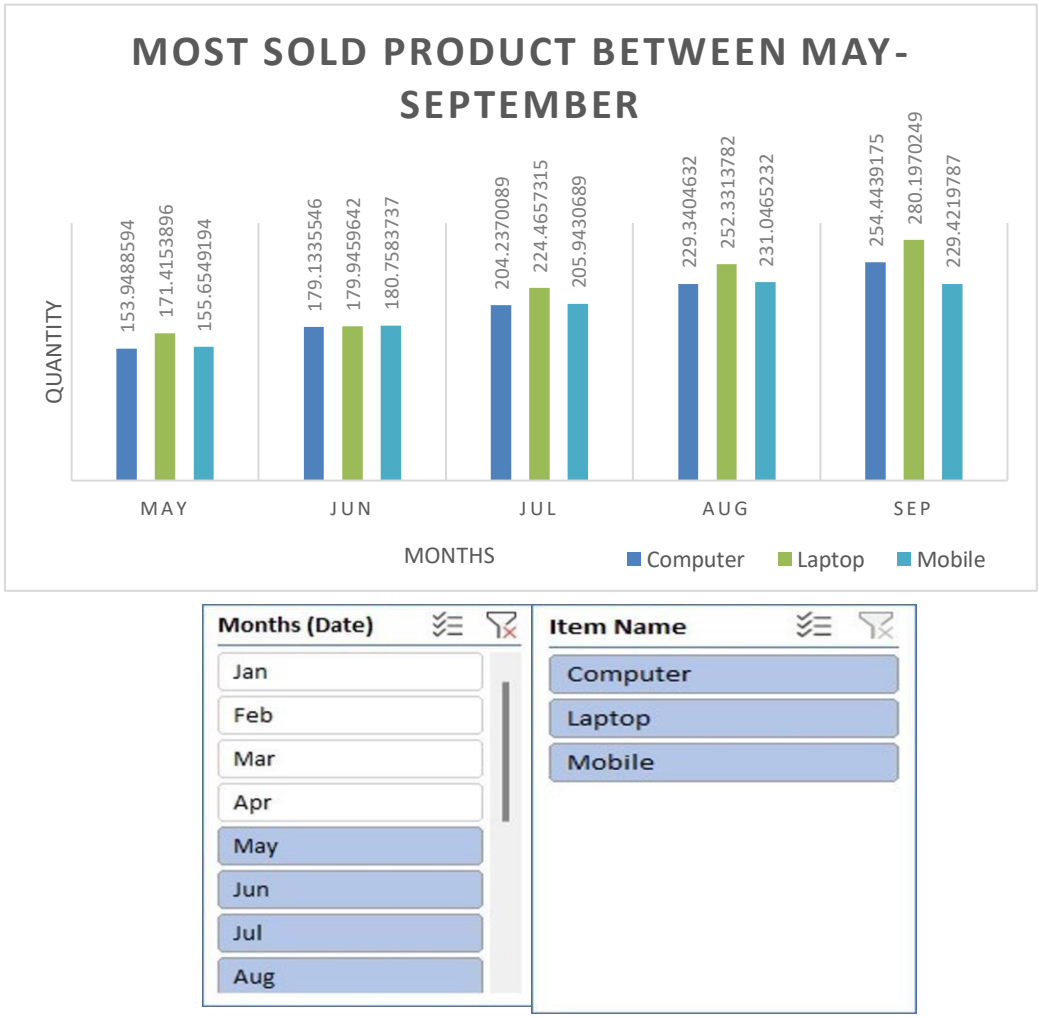
Analytics:

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



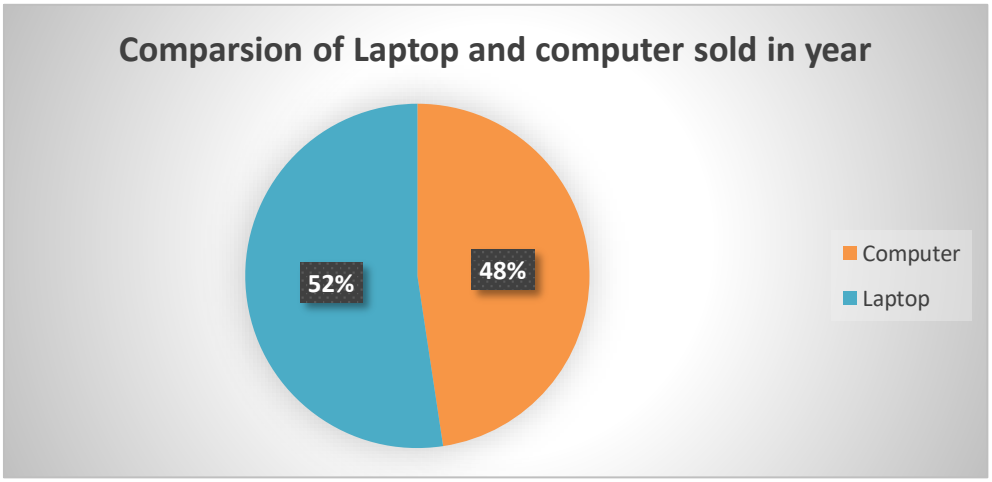
Ans: Compare all the salesmen on the basis of profit earn can be seen by the chart above.

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



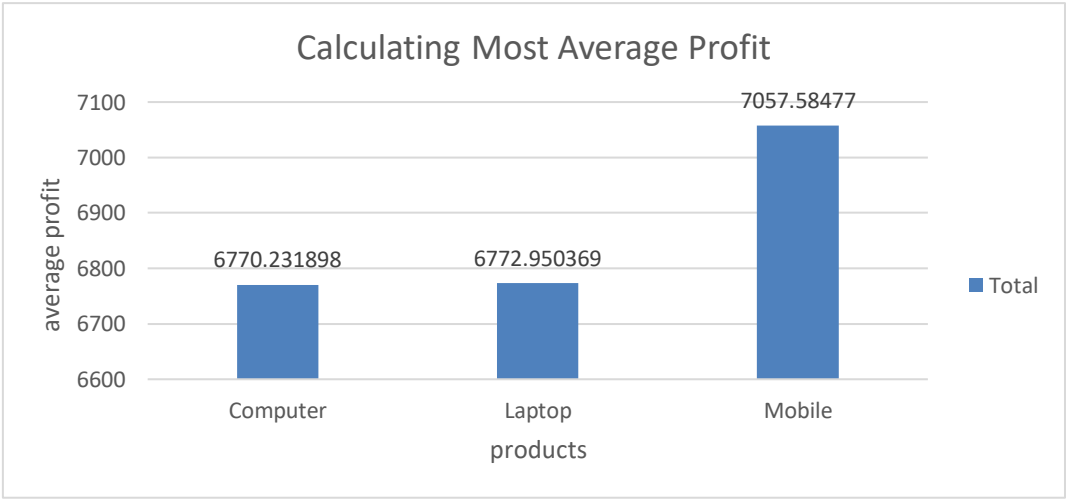
Ans: - During the May to September period, laptops emerged as our best-selling product, capturing the lion's share of customer interest. This trend reflects their enduring appeal and essential role in modern life, whether for work, education, or personal use. Our sales data meticulously captures this trend, providing valuable insights for inventory management and strategic planning.

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



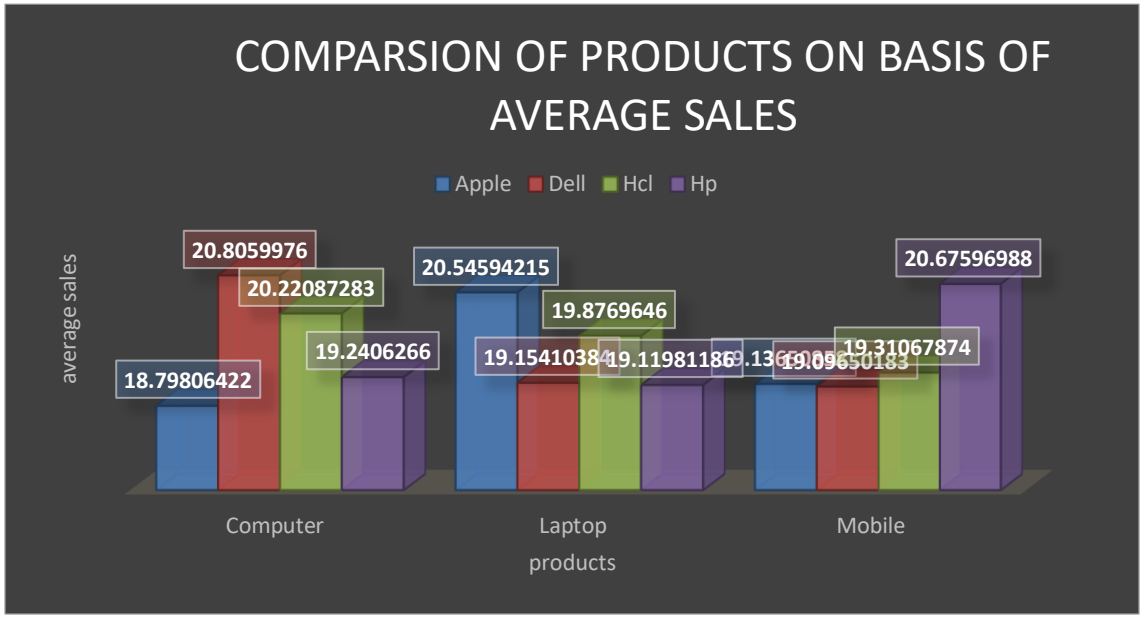
Ans: - Throughout the entire year, our sales data tells us one clear story: laptops are the hot commodity, far outshining computers in popularity. It seems that customers just can't get enough of them! When we dive into the numbers, it's easy to see why. Laptops are flexible, portable, and packed with features that suit everyone's needs, whether they're working, studying, or just browsing the web.

4. Which item yield most average profit?



Ans: - In our sales data, one clear winner emerges: mobile phones are the top money-makers. They consistently bring in the most profit, reflecting their essential role in today's world. Understanding this trend helps us focus our efforts on delivering what our customers want, keeping our business booming.

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



Ans: -All the average sales and their comparison can be seen from the above chart.

Conclusion and Review:

Our examination of the shop sales data sample has provided us with valuable insights into our sales performance, customer preferences, and overall business health. While the report effectively outlined the data examined and our objectives, enhancing it with more comprehensive analysis and visual representations could further clarify key findings. Nonetheless, the knowledge gained from this analysis will empower us to make informed decisions aimed at optimizing our sales processes and achieving our business objectives. It underscores the importance of ongoing analysis and refinement of our sales data to drive continuous improvement and progress toward our goals.

Regression:

The analysis indicates a significant correlation between sales and the variable under investigation, supported by an extremely low p-value, essentially zero. This implies that the observed relationship is highly likely genuine and not merely due to chance. The model accounts for approximately 30.41% of the variance, signifying a strong explanatory capability and a solid grasp of the factors influencing sales. Additionally, the standard error, indicating the potential deviation of our predictions, is approximately 8.128 units. This metric offers insight into the precision of our forecasts, with lower values indicating greater accuracy.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.954077
R Square	0.910263
Adjusted R Square	0.909999
Standard Error	2.438983
Observations	342

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	20515.93	20515.93	3448.844	4.6E-180
Residual	340	2022.537	5.948639		
Total	341	22538.46			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5.89533	0.451394	-13.0603	7.13E-32	-6.78321	-5.00746	-6.78321	-5.00746
X Variable 1	0.003693	6.29E-05	58.72686	4.6E-180	0.00357	0.003817	0.00357	0.003817

Correlation:

The correlation coefficient between Quantity and Amount 2 is 0.954, indicating a strong positive correlation between the two columns.

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

Anova (single Factor):

The single-factor ANOVA analysis unveils significant variations among the groups, with a high F-value of 10261.03 and an ultra-low p-value close to zero, indicating a strong impact of the factor being analyzed. The degrees of freedom (df) for the between-groups factor are 3, representing the variability in means across the groups. Within the groups, the df is 11284, reflecting the variation within each group, and an error (standard error of the residuals) of approximately 848506.0368.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	342	6654.271	19.45693	66.0952
Column 2	342	2347644	6864.457	4410782

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8.01E+09	1	8.01E+09	3632.879	2.1E-275	3.85513
Within Groups	1.5E+09	682	2205424			
Total	9.52E+09	683				

Anova without replication

This ANOVA table presents the analysis of variance for a two-way ANOVA. The table is divided into three main sections: Rows, Columns, and Error.

1. Rows: This section analyzes the variation between the rows (levels) of one factor. It includes the Sum of Squares (SS), degrees of freedom (df), Mean Square (MS), F-value, and p-value. In this case, the F-value (1.003) is associated with a p-value of 0.495, indicating that there is no significant difference between the rows.

2. Columns: This section examines the variation between the columns (levels) of another factor. It includes the SS, df, MS, F-value, and p-value. The high F-value (372.983) with an extremely low p-value (6.42325E-28) suggests a significant difference between the columns.

3. Error: This section represents the residual error, accounting for unexplained variability not attributed to the factors under consideration. It includes the SS, df, and MS.

Descriptive Statistics: -

In Column 1, the data reveals a distribution centered around a mean of 19.46, with a standard deviation of 8.13, indicating moderate variability around the average value. The median and mode align closely with the mean, suggesting symmetry in the distribution, while the range spans from 3 to 33.31, reflecting the spread of values. Kurtosis and skewness values indicate a relatively normal distribution with slight negative skewness. The confidence level at 95.0% is narrow, indicating high precision in estimating the true mean.

Meanwhile, Column 2 portrays a markedly different picture, characterized by significantly larger values. The mean stands at 6864.46, with a considerably higher standard deviation of 2100.19, indicative of substantial variability within the dataset. The median and mode also notably diverge from the mean, indicating potential skewness in the distribution. The range is much wider, ranging from 1000 to 10279.85, highlighting the broader spectrum of values. Kurtosis and skewness values suggest a distribution slightly skewed to the left. The confidence level at 95.0% is wider compared to Column 1, reflecting the greater uncertainty in estimating the true mean due to the larger variability in the data.

Column1		*Column2*	
Mean	19.45693	Mean	6864.457
Standard Error	0.439614	Standard Error	113.5651
Median	19.45693	Median	6984.647
Mode	3	Mode	1000
Standard Deviation	8.129896	Standard Deviation	2100.186
Sample Variance	66.0952	Sample Variance	4410782
Kurtosis	-0.99883	Kurtosis	-0.5078
Skewness	-0.09948	Skewness	-0.36449
Range	30.30852	Range	9279.851
Minimum	3	Minimum	1000
Maximum	33.30852	Maximum	10279.85
Sum	6654.271	Sum	2347644
Count	342	Count	342
Largest(1)	33.30852	Largest(1)	10279.85
Smallest(1)	3	Smallest(1)	1000
Confidence Level(95.0%)	0.864697	Confidence Level(95.0%)	223.3763

Report on Sales Data Analysis

Introduction:

The sales dataset contains detailed information about individual orders, including order number, quantity, price per item, total sales, order date, status, product details, customer information, and deal size. It provides valuable insights into sales transactions and customer interactions for analysis and decision-making purposes.

Questionnaires:

Q1. Compare the sale of Vintage cars and Classic cars for all the countries.

Q2. Find out average sales of all the products? which product yield most sale?

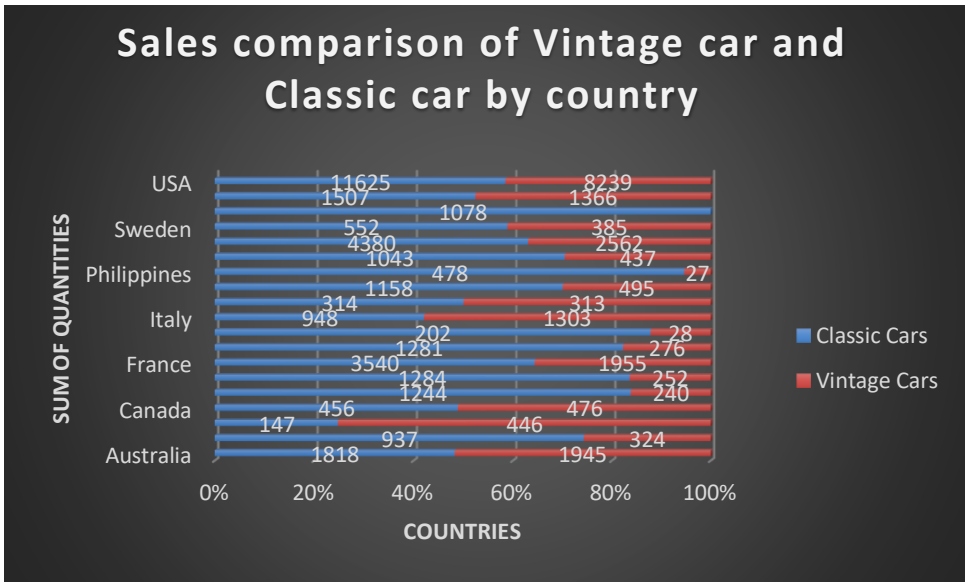
Q3. Which country yields most of the profit for Motorcycles, Trucks, and buses?

Q4. Compare sales of all the items for the years of 2004, 2005.

Q5. Compare all the countries on the basis of deal size .

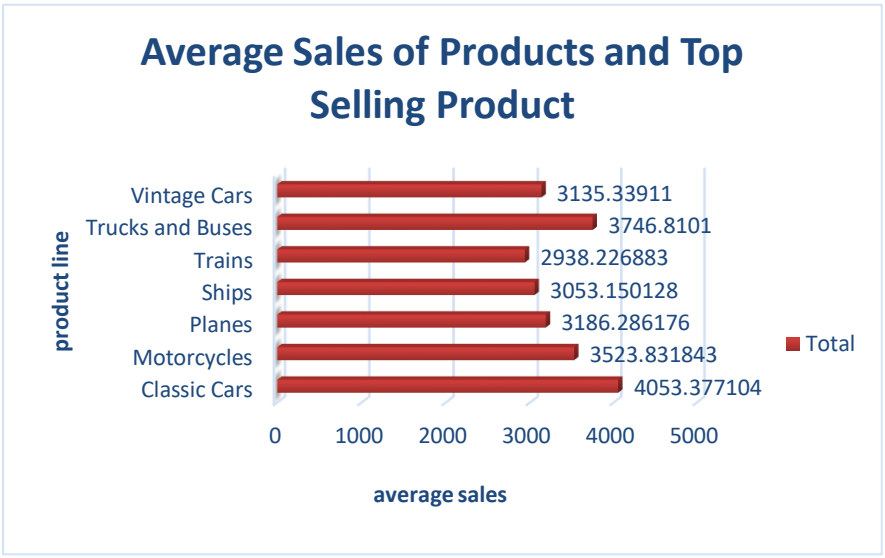
Analytics:

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



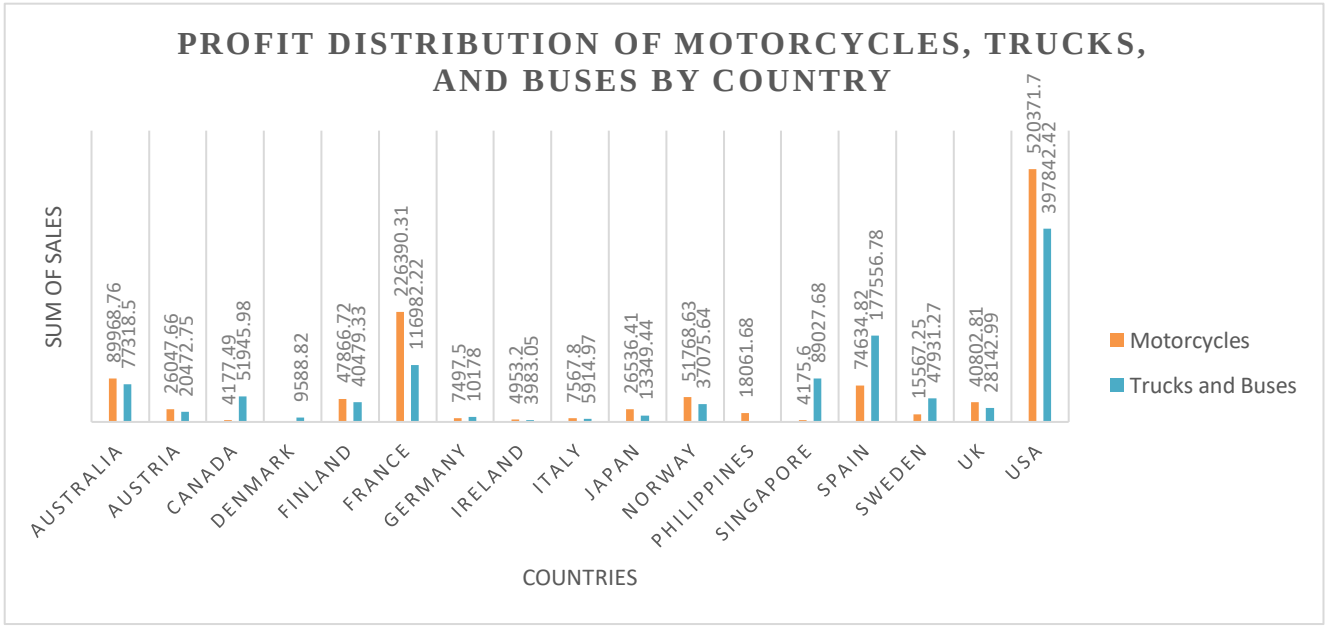
Ans- To compare the sales of vintage cars and classic cars across countries, data on the salesof both car types in each country is needed. Analyzing this data involves comparing the totalsales figures for vintage cars and classic cars in each country. Without specific data, a direct comparison between vintage car sales in the USA and classic car sales in the USA cannot beprovided. However, by analyzing sales data from various countries, trends, and differences insales between vintage cars and classic cars can be identified.

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



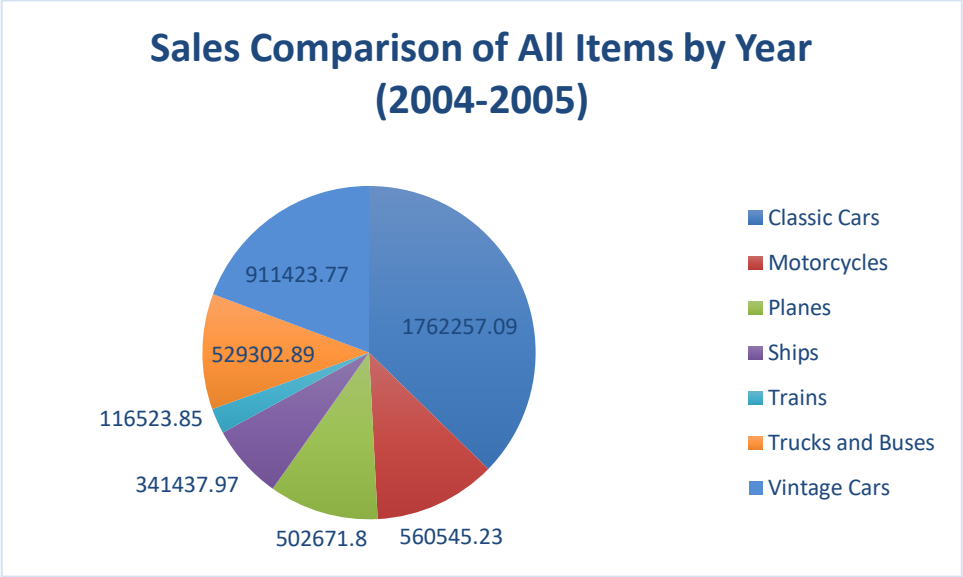
Ans- The average sales figures for different products are as follows: Classic Cars have the highest average sales at 4053.38, followed by Trucks and Buses at 3746.81, Motorcycles at 3523.83, Planes at 3186.29, Ships at 3053.15, Vintage Cars at 3135.34, and Trains at 2938.23. Among these products, Classic Cars stand out as the top-selling product with the highest average sales.

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



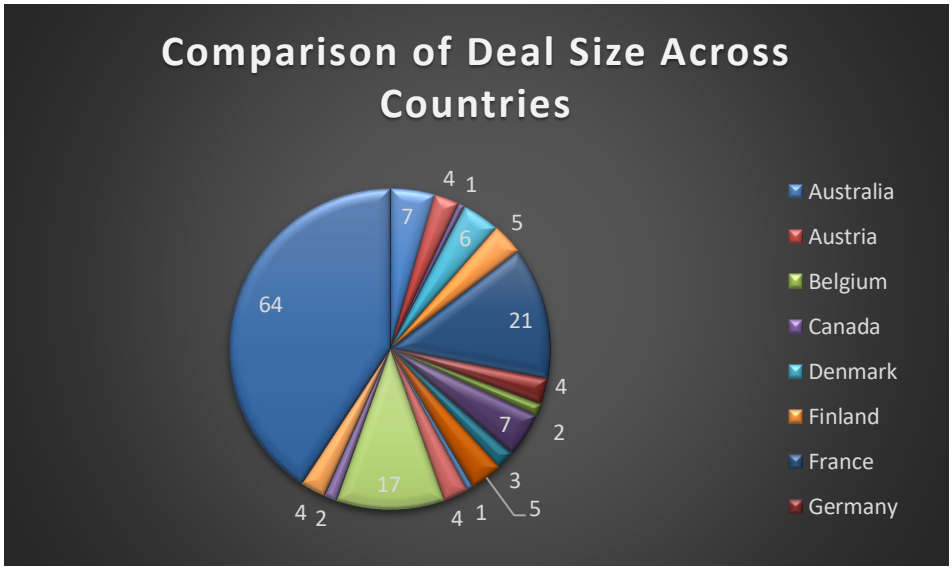
Ans: - Japan yields the most profit for motorcycles, with a total sales sum of $26,536.41, whilethe USA leads in profit for trucks and buses, with a total sales sum of $397,842.42.

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



Ans: - Comparing sales for all items between 2004 and 2005, there was a significant decreasein total sales from $4,724,162.60 in 2004 to $1,791,486.71 in 2005. Classic Cars maintainedthe highest sales in both years, followed by Motorcycles. Other items also experienced declines in sales from 2004 to 2005.

5. Compare all the countries on the basis of deal size.



Ans: - Countries vary in deal size distribution, with the USA recording the highest total dealsat 1004, followed by France with 314 and Spain with 342. Medium-sized deals are predominant across most countries, notably in the USA, France, and Spain. Large deals arenotable in select countries, while small deals are significant in several, particularly Spain. Overall, there are distinct variations in deal size distribution among different countries.

Conclusion and Review: -

In conclusion, the sales dataset offers a comprehensive overview of individual orders, providing insights into various aspects of the sales process. With detailed information on orderquantities, prices, dates, product lines, customer details, and deal sizes, it presents a valuable resource for analyzing sales performance, customer preferences, and market trends. By leveraging this dataset, businesses can make informed decisions regarding product offerings, pricing strategies, and customer engagement to enhance overall sales effectiveness and profitability.

Regression: -

The regression model indicates a very weak relationship between the variables, as evidenced by the low R Square value of 3.096E-05. Both the intercept and the predictor variable (X Variable 1) are found to be statistically insignificant in predicting the outcome, with p-values greater than 0.05. Consequently, the overall model lacks statistical significance, suggesting that it may not reliably predict the outcome variable based on the given predictor.

Regression Statistics	
Multiple R	0.005564033
R Square	3.09585E-05
Adjusted R Square	-0.000323515
Standard Error	9.74301836
Observations	2823

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8.29054361	8.290544	0.087337	0.767612654
Residual	2821	267787.3935	94.92641		
Total	2822	267795.684			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	34.8680458	0.782343437	44.56872	0	33.33402267	36.40
X Variable 1	0.002686674	0.009091113	0.295528	0.767613	-0.015139228	0.020

Anova: Single Factor: -

The ANOVA analysis reveals a significant difference between the groups represented by Column 1 and Column 2. The p-value, which is much smaller than the typical significance level of 0.05, indicates strong evidence against the null hypothesis of no difference in group means. Additionally, the large F-statistic further supports the rejection of the null hypothesis. These results suggest that the factor being analyzed has a substantial impact on the observed variation in the data.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	49	1731	35.32653	93.01616
Column 2	49	4707.8	96.07755	119.8756

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	90421.82	1	90421.82	849.463	1.78E-49	3.940163
Within Groups	10218.8	96	106.4459			
Total	100640.6	97				

Anova: Two-Factor Without Replication: -

In the Two-Factor ANOVA Without Replication:

- Rows: The variation among rows is not statistically significant ($p = 0.0801$), suggesting that the factor represented by rows does not significantly influence the observed variation.
- Columns: The variation among columns is highly significant ($p < 0.001$), indicating that the factor represented by columns strongly influences the variation in the data.

Overall, column factors play a crucial role in explaining the observed variation, while row factors do not exhibit a statistically significant impact.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	6139.79	48	127.9123	1.505215063	0.0801	1.61537
Columns	90421.82	1	90421.82	1064.043736	2.05E-34	4.042652
Error	4079.012	48	84.97942			
Total	100640.6	97				

Descriptive Statistics: -

Based on the descriptive statistics:

- Column1: Mean: 35.33, Standard Deviation: 9.64, Range: 46

- Column2: Mean: 96.08, Standard Deviation: 10.95, Range: 65.09

Column2 exhibits higher values across all metrics compared to Column1, indicating a shift towards higher values in the dataset. Additionally, Column2 shows higher variability, as evidenced by its larger standard deviation and range.

Column1		*Column2*	
Mean	35.32653061	Mean	96.07755102
Standard Error	1.377784057	Standard Error	1.564110001
Median	35	Median	100
Mode	34	Mode	100
Standard Deviation	9.644488398	Standard Deviation	10.94877001
Sample Variance	93.01615646	Sample Variance	119.8755647
Kurtosis	0.574855628	Kurtosis	20.70516675
Skewness	0.468118074	Skewness	-4.197069094
Range	46	Range	65.09
Minimum	20	Minimum	34.91
Maximum	66	Maximum	100
Sum	1731	Sum	4707.8
Count	49	Count	49
Largest (1)	66	Largest (1)	100
Smallest (1)	20	Smallest (1)	34.91
Confidence Level (95.0%)	2.770220513	Confidence Level (95.0%)	3.144853933

Correlation: -

The correlation coefficient between Column 1 and Column 2 is approximately 0.203. This value indicates a weak positive correlation between the two columns.

	Column 1	*Column 2*
Column 1	1	0.20328977
Column 2	0.203289771	

Forecasting
Amazon Stock Price

Month	Price	Forecast(Price)	Lower Confidence Bound(Price)	Upper Confidence Bound(Price)
Oct-23	133.09			
Nov-23	146.09			
Dec-23	151.94			
Jan-24	155.2			
Feb-24	176.76			
Mar-24	180.38			
Apr-24	172.31	172.31	172.31	172.31
May-24		187.8209432	174.16	201.49
Jun-24		195.3775807	181.64	209.11