Data Analysis Report

Title:

Comprehensive Analysis and Dashboard Creation on Multiple Datasets

Subtitle :

Insights and Trends Across Diverse Data Sets Using Excel

Abstract:

This report presents an in-depth analysis of various datasets, focusing on uncovering key insights through the creation of interactive and insightful dashboards using Excel. The analysis spans across diverse domains, including automotive data, sales trends, loan data, and regional market performance in the United States. Each section is designed to provide actionable insights that can aid in strategic decision-making and business optimization.

Prepared by:
Aman Ahmed Khan
Data Analyst

Contact Email:
iamaman.230230@gmail.com

Date:
August 2024

Overview

This report presents a comprehensive analysis of seven distinct datasets, each meticulously examined to extract meaningful insights and trends. The dashboards created in Excel offer a visual representation of the data, enabling a clear understanding of the underlying patterns and helping in data-driven decision-making. Below is a brief overview of each dataset and the focus of its corresponding analysis:

1. **Exploring Car Dataset**
   This analysis delves into various aspects of car attributes such as price, mileage, and brand popularity. The dashboard provides insights into market trends, helping to identify key factors that influence car prices and consumer preferences.

2. **Cookie Data: Trends and Analysis Report**
   Focused on consumer behavior and sales trends within the cookie market, this dashboard highlights patterns in product popularity, seasonal variations, and potential growth opportunities for cookie manufacturers.

3. **Exploring Loan Dataset**
   This section analyzes loan data, offering insights into approval rates, borrower demographics, and factors affecting loan defaults. The dashboard aids in understanding risk management and optimizing loan offerings.

4. **Exploring Sales on Different States of the US**
   A detailed analysis of sales performance across various states in the US, this dashboard reveals regional trends, top-performing states, and areas with potential for market expansion.

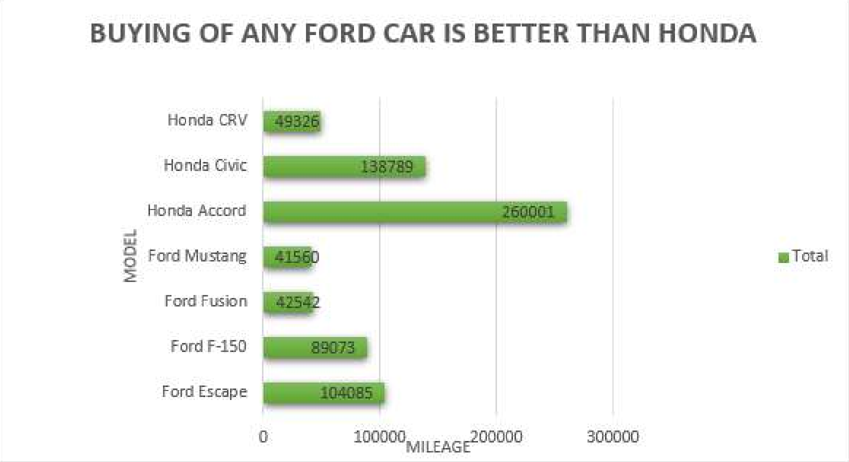
5. **Store Data**
   This dataset examines store operations, focusing on sales efficiency, inventory management, and customer demographics. The dashboard provides a clear picture of store performance and areas for operational improvement.

6. **Shop Sale Data**
   An analysis of sales data from multiple shops, this dashboard highlights key metrics such as revenue, customer footfall, and product popularity, helping to optimize shop management strategies.

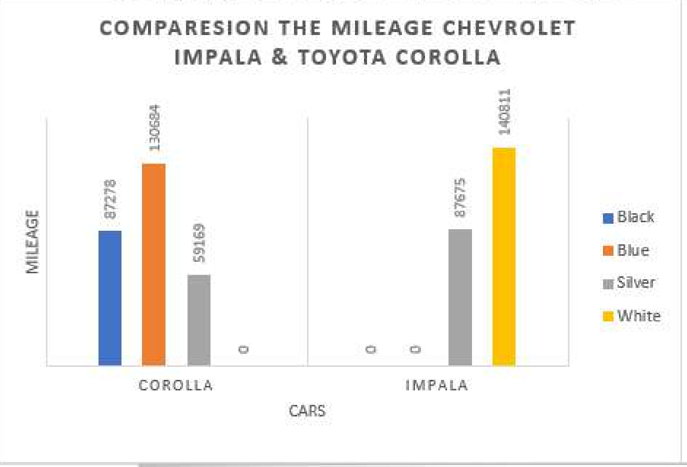
7. **Sale Samples**
   This dataset offers a snapshot of sales samples, analyzed to identify trends in customer preferences and product performance. The dashboard provides actionable insights to enhance sales strategies and improve customer satisfaction.

S.No.	Experiment	Remarks
1.	Data Analysis Questions: i. Data Analysis Principles ii. Statistical Analytics iii. Hypothesis Testing iv. Regression v. Correlation vi. ANOVA vii. 5 v's of Big Data	
2.	Dashboards: i. Exploring Car Dataset ii. Cookie Data: Trends and Analysis Report iii. Exploring Loan Dataset iv. Exploring sales on different states of US v. Store Data Analysis vi. Shop Sale Data Report vii. Sale Samples: A Detailed Report	
3.	Reports: i. Exploring Car Dataset ii. Cookie Data: Trends and Analysis Report iii. Exploring Loan Dataset iv. Exploring sales on different states of US v. Store Data Analysis vi. Shop Sale Data Report vii. Sale Samples: A Detailed Report	
4.	Remote Ratio Forecast Analysis (2020-2026)	

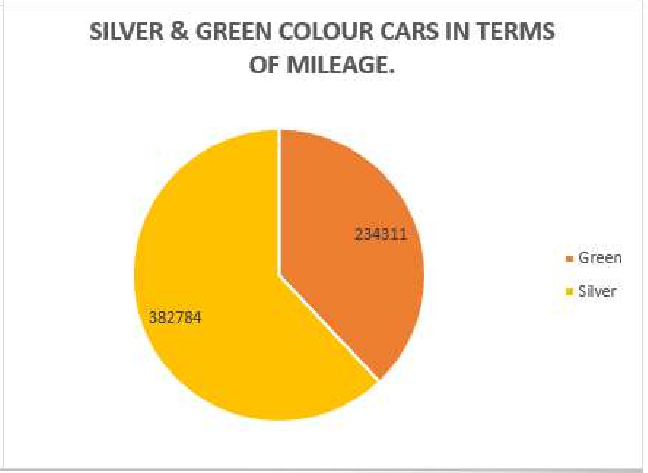
CAR DASHBOARD



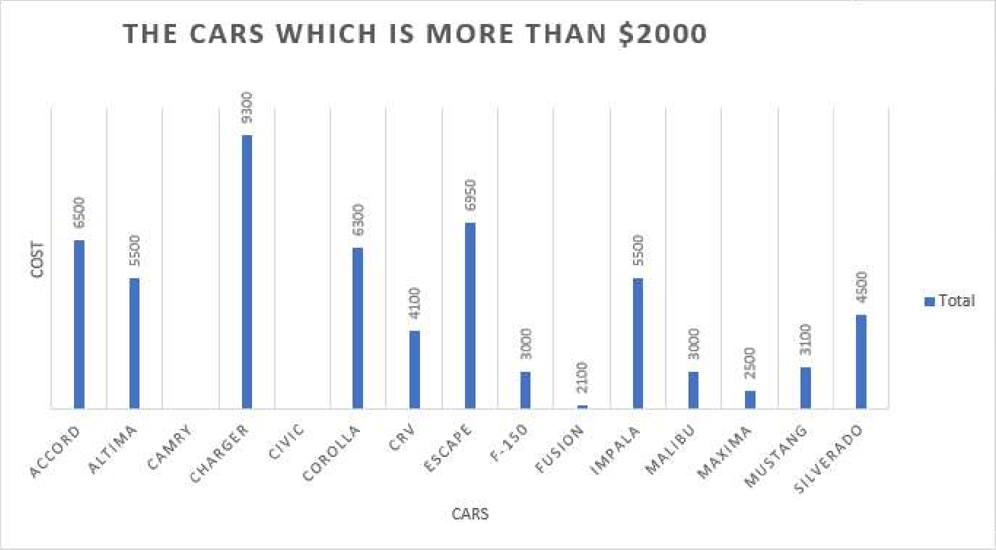
BUYING OF ANY FORD CAR IS BETTER THAN HONDA



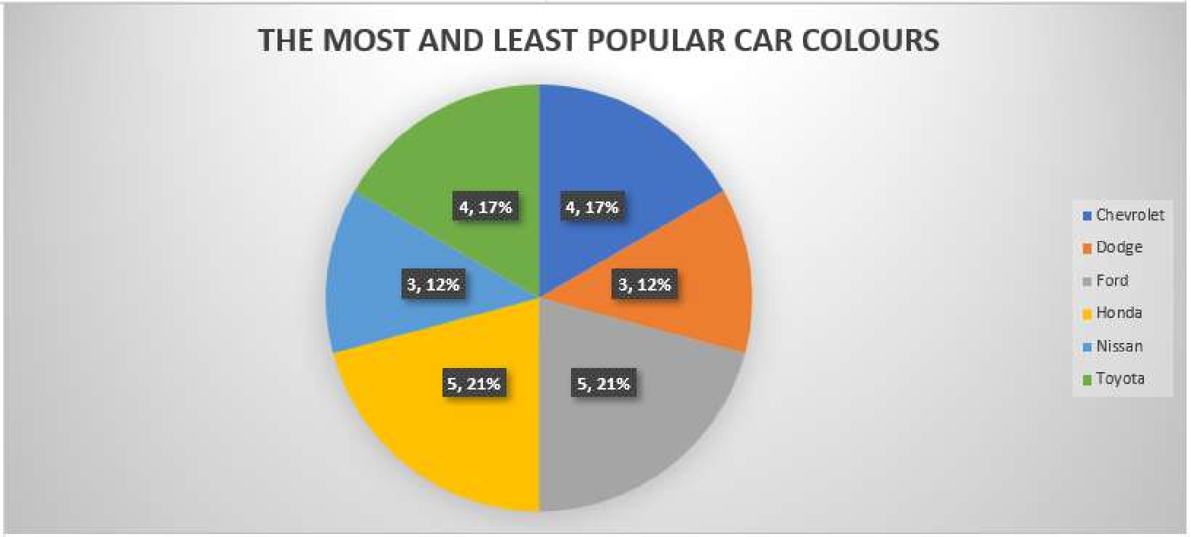
COMPARESION THE MILEAGE CHEVROLET IMPALA & TOYOTA COROLLA



SILVER & GREEN COLOUR CARS IN TERMS OF MILEAGE.

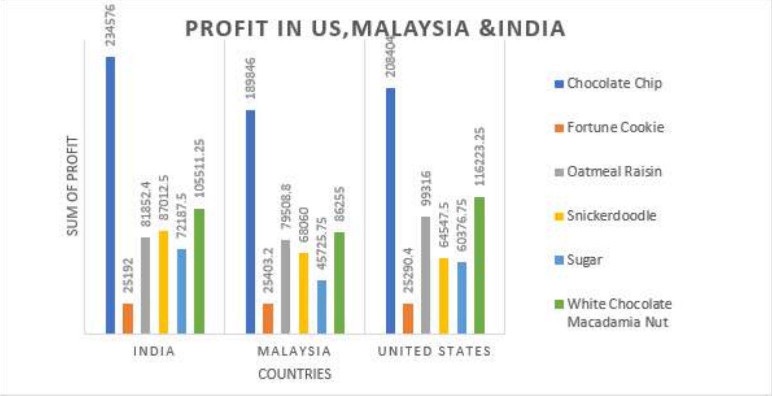


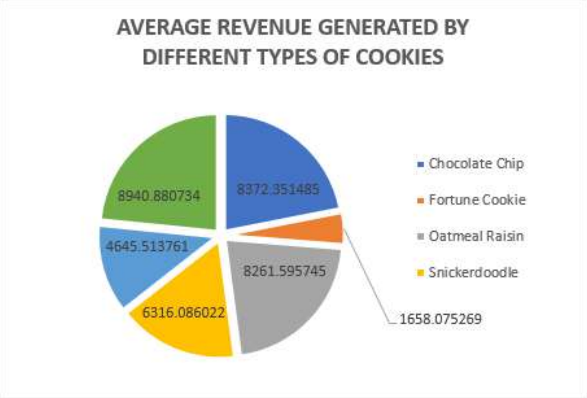
THE CARS WHICH IS MORE THAN $2000

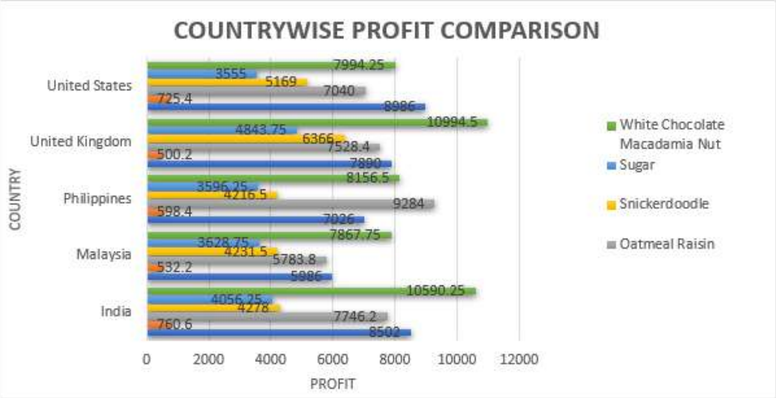


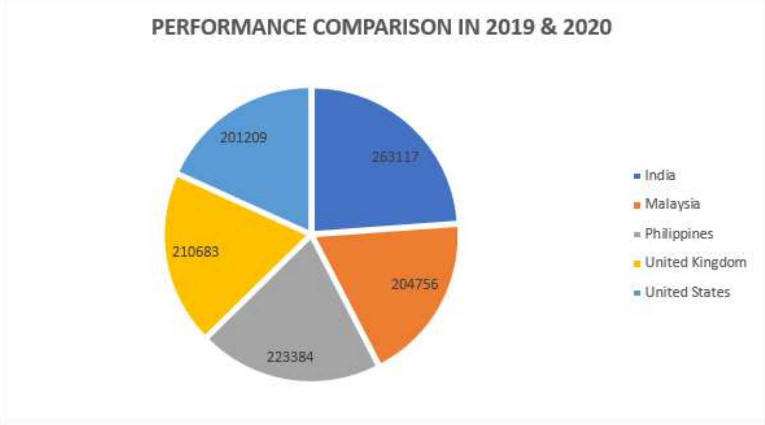
THE MOST AND LEAST POPULAR CAR COLOURS

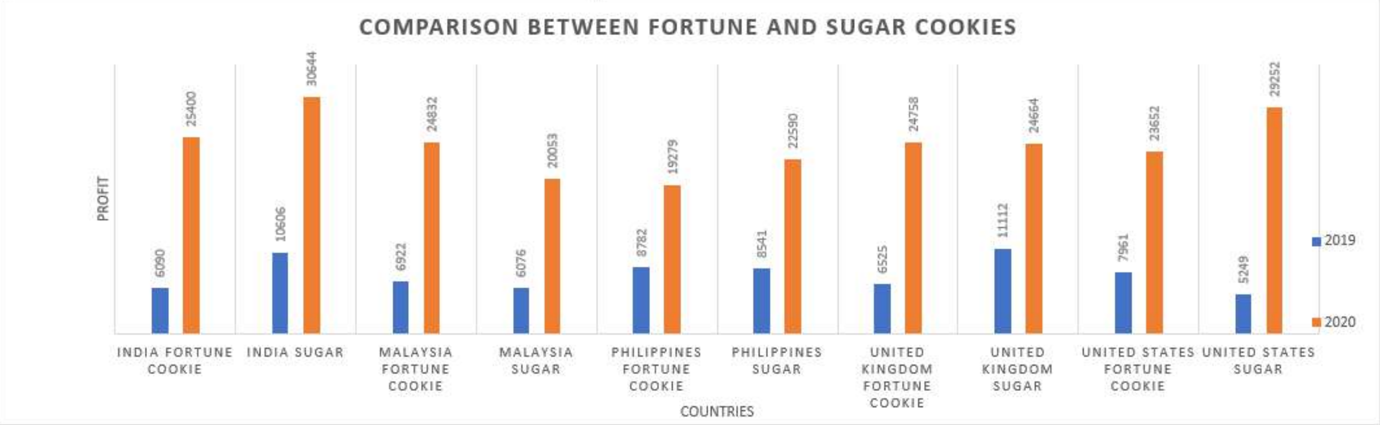
DASHBOARD FOR COOKIES DATA ANALYSIS


PROFIT IN US,MALAYSIA &INDIA


AVERAGE REVENUE GENERATED BY DIFFERENT TYPES OF COOKIES

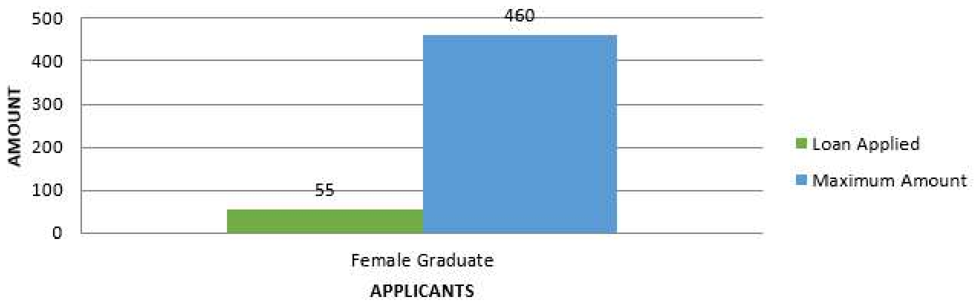

COUNTRYWISE PROFIT COMPARISON


PERFORMANCE COMPARISON IN 2019 & 2020

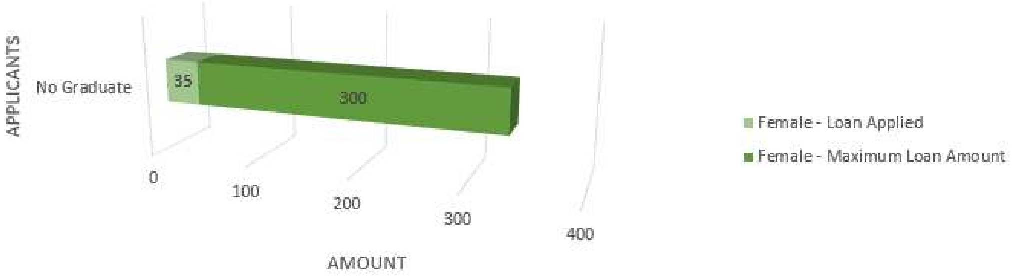

COMPARISON BETWEEN FORTUNE AND SUGAR COOKIES

DASHBOARD FOR LOAN DATASET

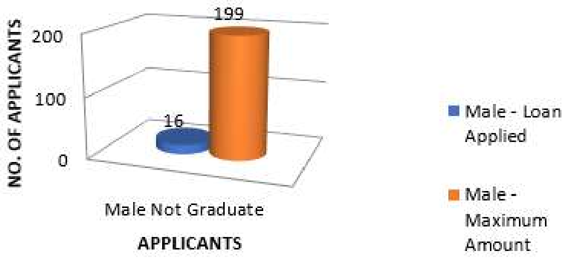
FEMALE GRADUATES WHO ARE MARRIED APPLIED FOR LOAN & THE HIGHEST AMOUNT?



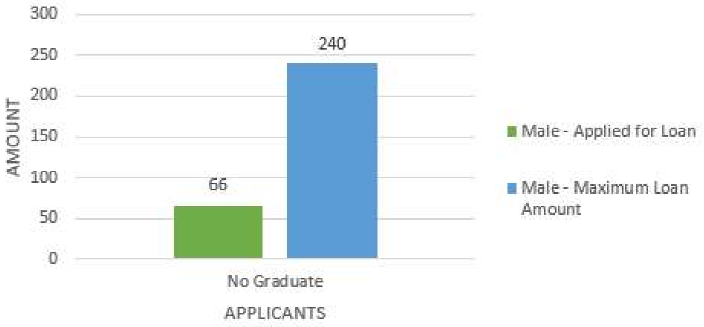
FEMALE GRADUATES WHO ARE NOT MARRIED APPLIED FOR LOAN & THE HIGHEST AMOUNT



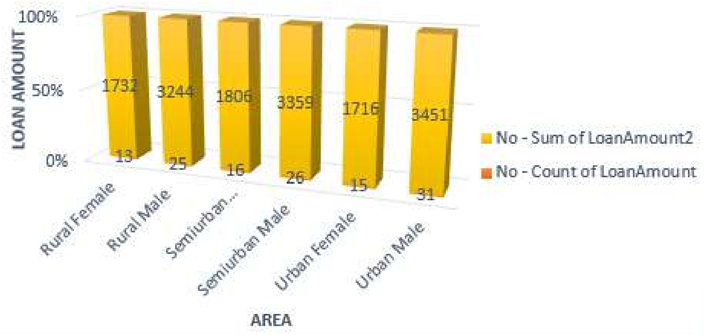
MALE NON-GRADUATES WHO ARE NOT MARRIED APPLIED FOR LOAN & THE HIGHEST AMOUNT.



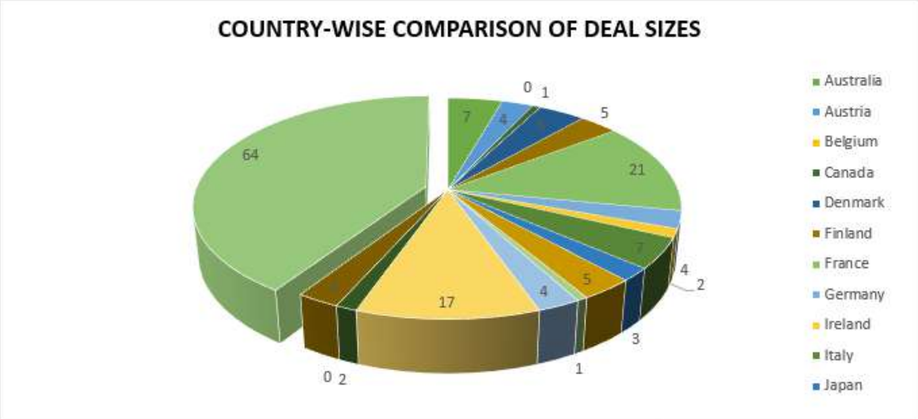
DESCRIPTIVE STATISTICS OF APPLICANT INCOME AND LOAN AMOUNT

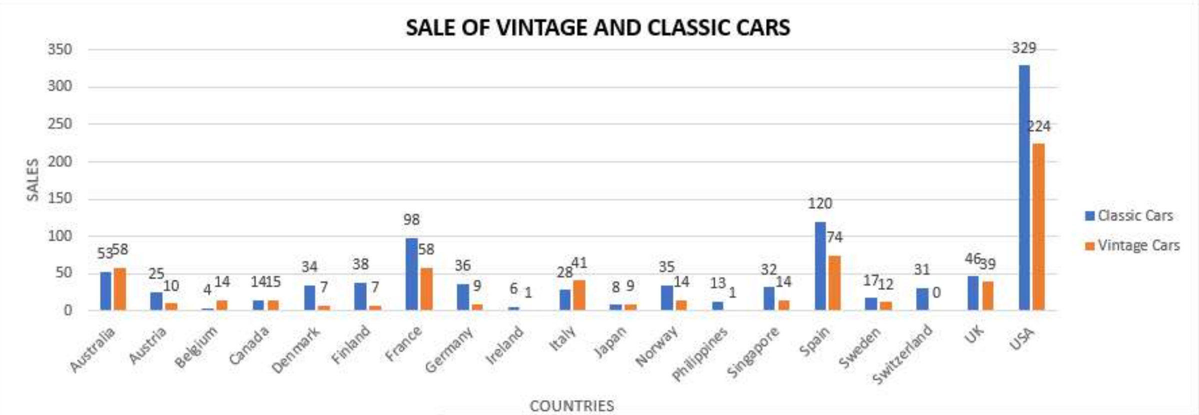


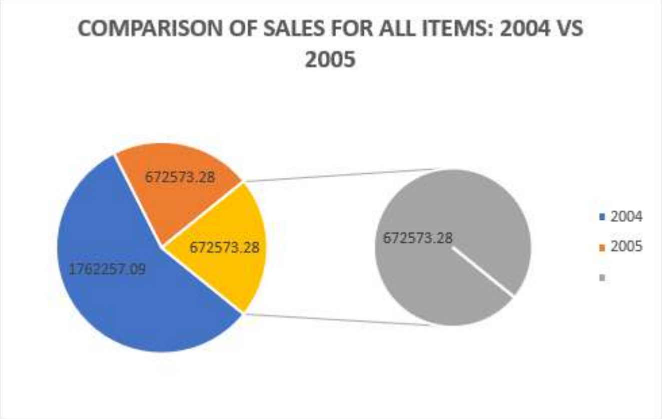
LOAN APPLICATIONS: UNMARRIED APPLICANTS BY GENDER AND URBAN-RURAL COMPARISON

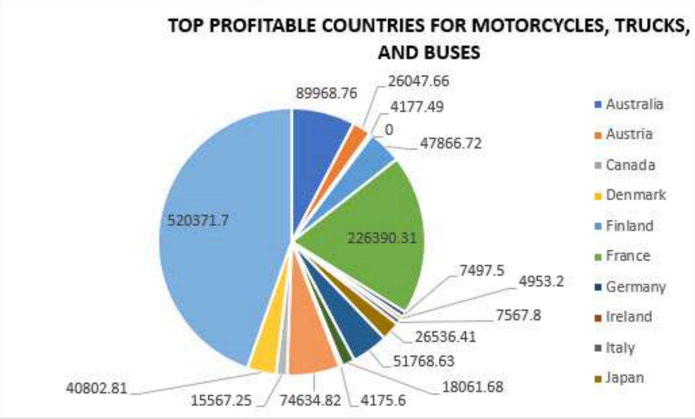


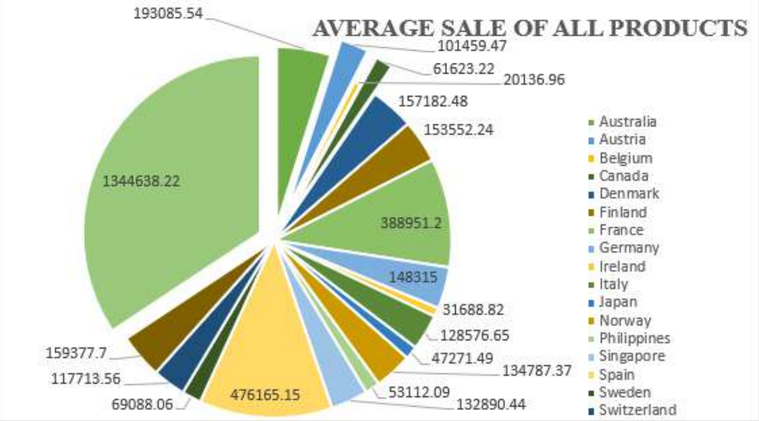
DASHBOARD FOR SALES DATASET


COUNTRY-WISE COMPARISON OF DEAL SIZES


SALE OF VINTAGE AND CLASSIC CARS

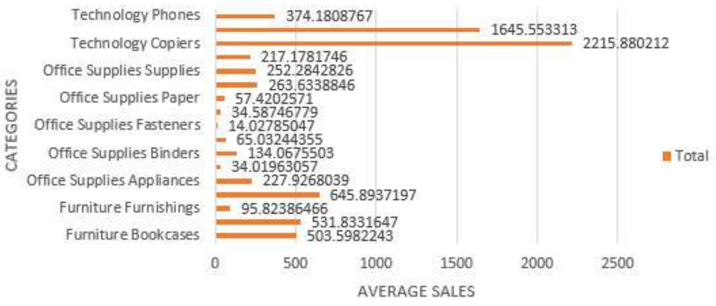

COMPARISON OF SALES FOR ALL ITEMS: 2004 VS 2005


TOP PROFITABLE COUNTRIES FOR MOTORCYCLES, TRUCKS, AND BUSES

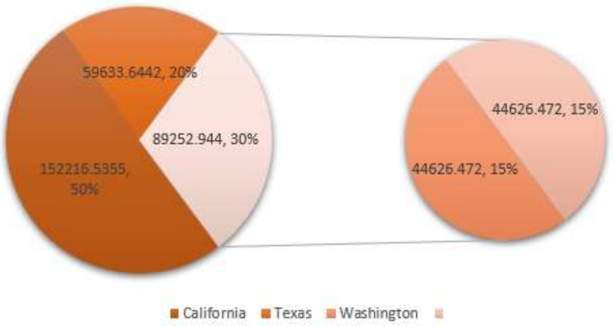

AVERAGE SALE OF ALL PRODUCTS

DASHBOARD FOR ORDER DATASET

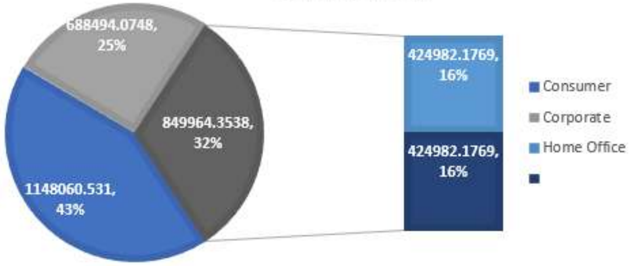
AVERAGE SALES COMPARISON ACROSS CATEGORIES AND SUBCATEGORIES IN ALL STATES



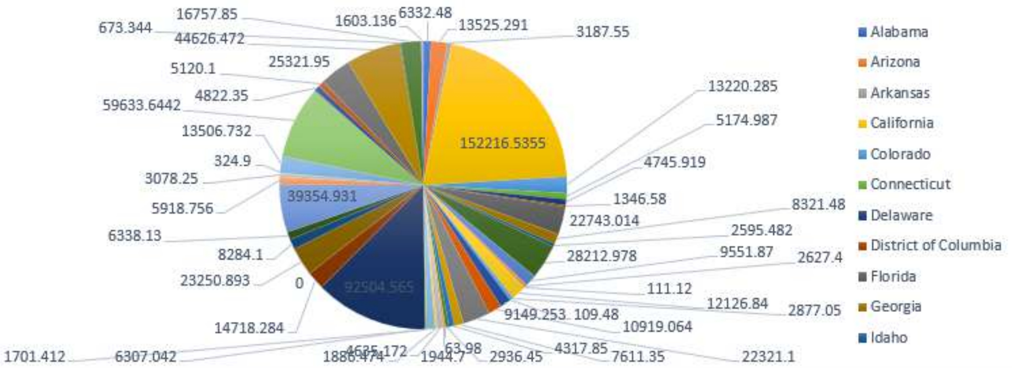
TOTAL VS. AVERAGE SALES BY SEGMENT



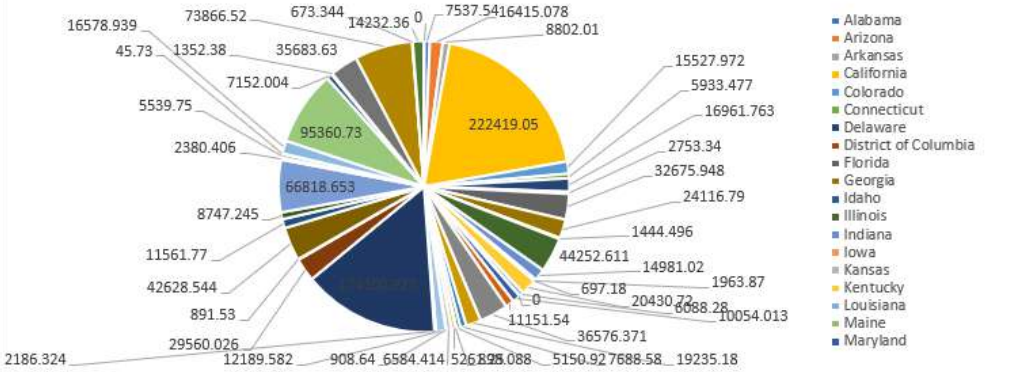
SEGMENT-WISE TOTAL AND AVERAGE SALES COMPARISON

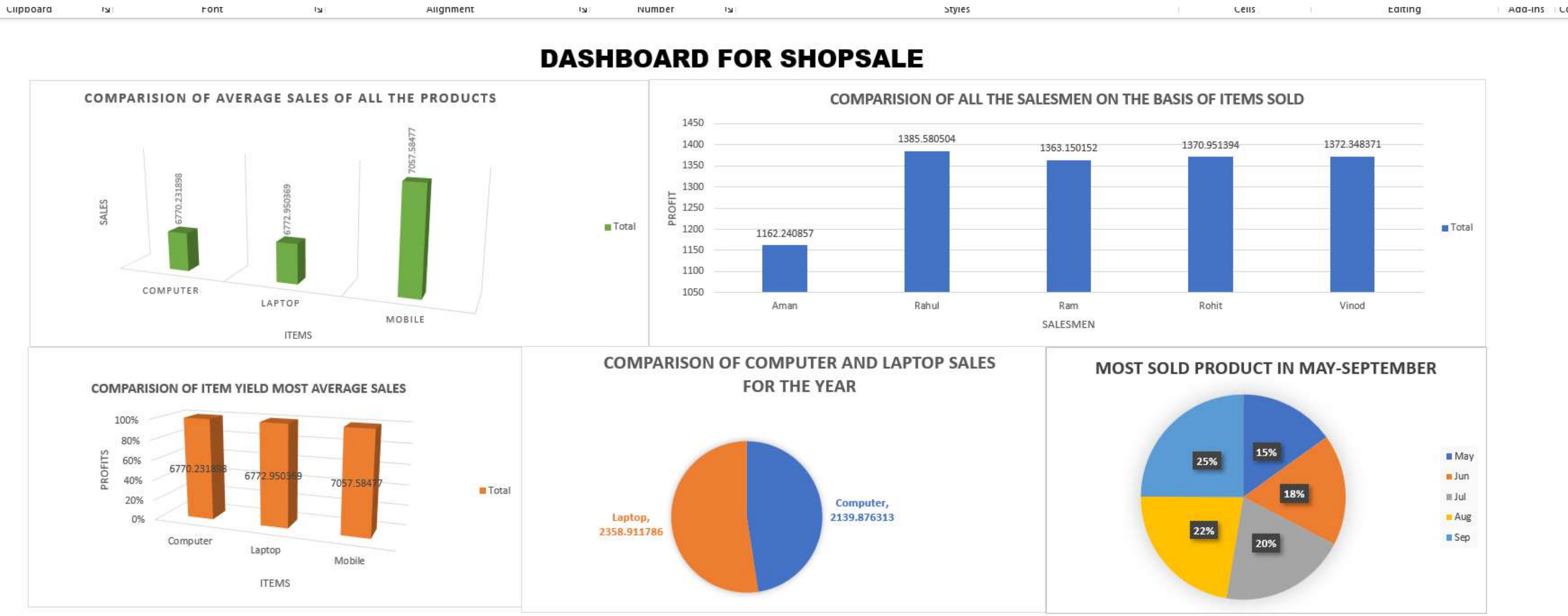


TOP PERFORMING CATEGORY IN ALL STATES



COMPARISON OF US STATES BY SEGMENT AND SALES





DASHBOARD FOR SHOPSALE

COMPARISION OF AVERAGE SALES OF ALL THE PRODUCTS

COMPARISION OF ALL THE SALESMEN ON THE BASIS OF ITEMS SOLD

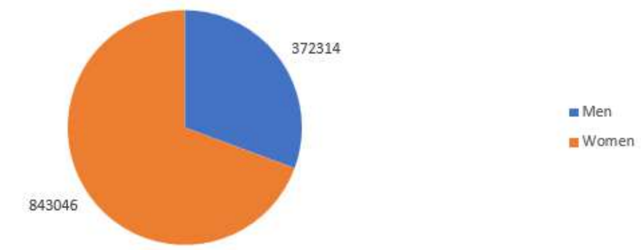
COMPARISION OF ITEM YIELD MOST AVERAGE SALES

COMPARISON OF COMPUTER AND LAPTOP SALES FOR THE YEAR

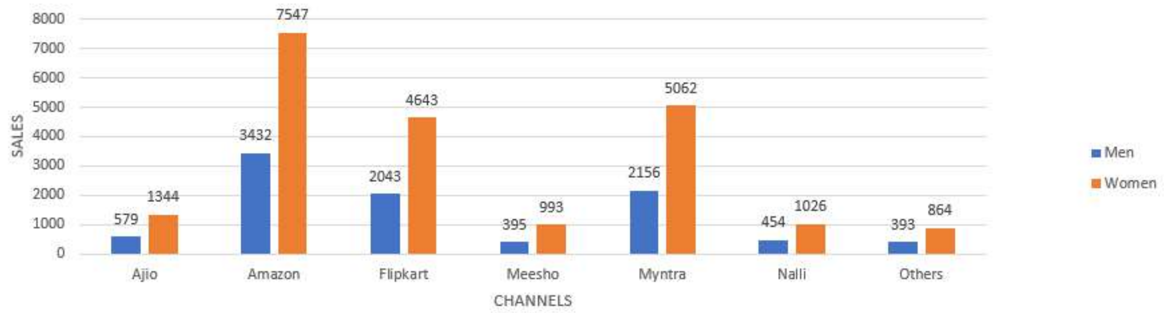
MOST SOLD PRODUCT IN MAY-SEPTEMBER

DASHBOARD FOR STORE DATASET

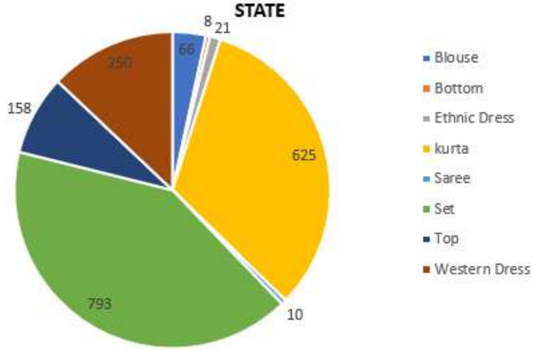
MOST SOLD CATEGORY ABOVE 30 YEARS IN DELHI



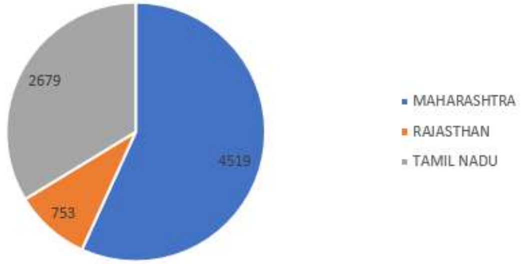
BEST PERFORMING CHANNEL BY GENDER



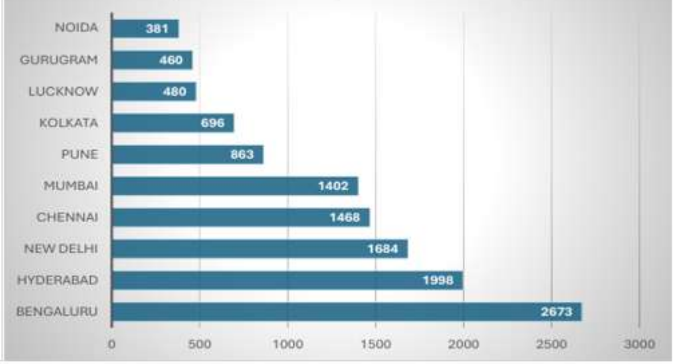
MONTH WITH THE HIGHEST SALES OF ITEMS BY CATEGORY IN ANY STATE



COMPARISON OF MAHARASHTRA, RAJASTHAN, AND TAMIL NADU: QUANTITY AND PROFIT



Top 10 Cities by Order Volume



Exploring Car Dataset

INTRODUCTION:

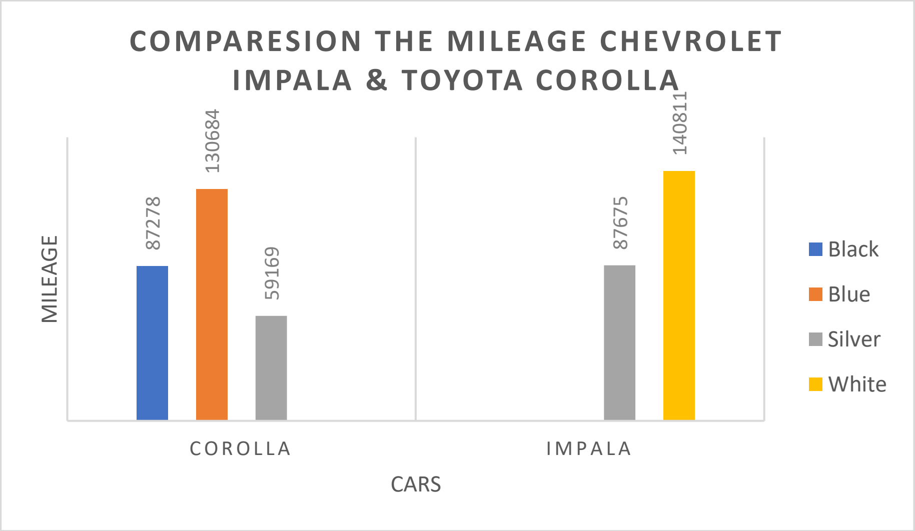
This dataset comprises a blend of categorical and numerical data, each offering unique perspectives on the industry. Categorical data, such as make, model, and colour, encapsulates the diversity of vehicles and consumer preferences. Meanwhile, numerical attributes like mileage, price, and cost provide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

QUESTIONNAIRES:

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

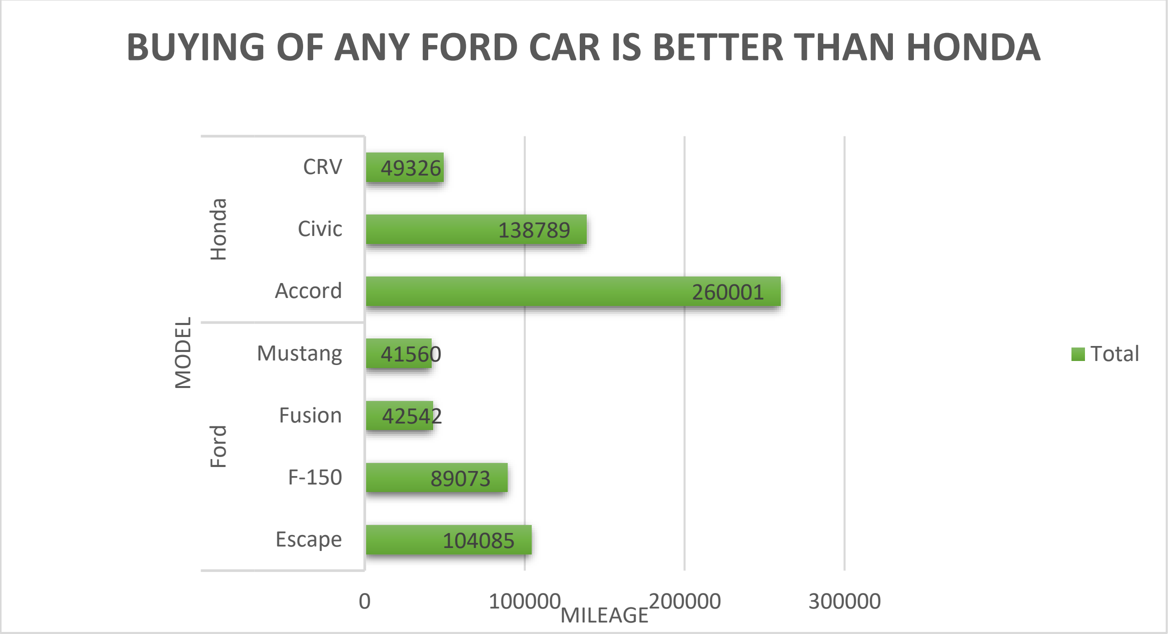
ANALYTICS:

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



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

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

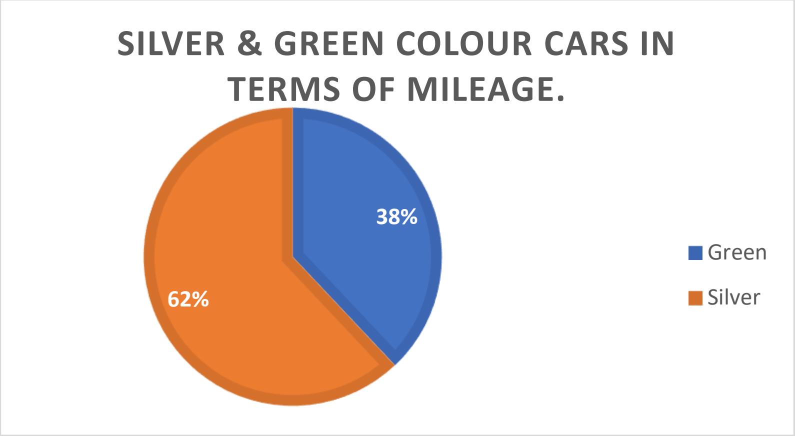


Ans: it appears that Honda cars collectively have a higher total mileage compared to Ford cars. However, it's essential to remember that mileage alone may not be the sole determining factor in choosing a car. Here are some additional points to consider:

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

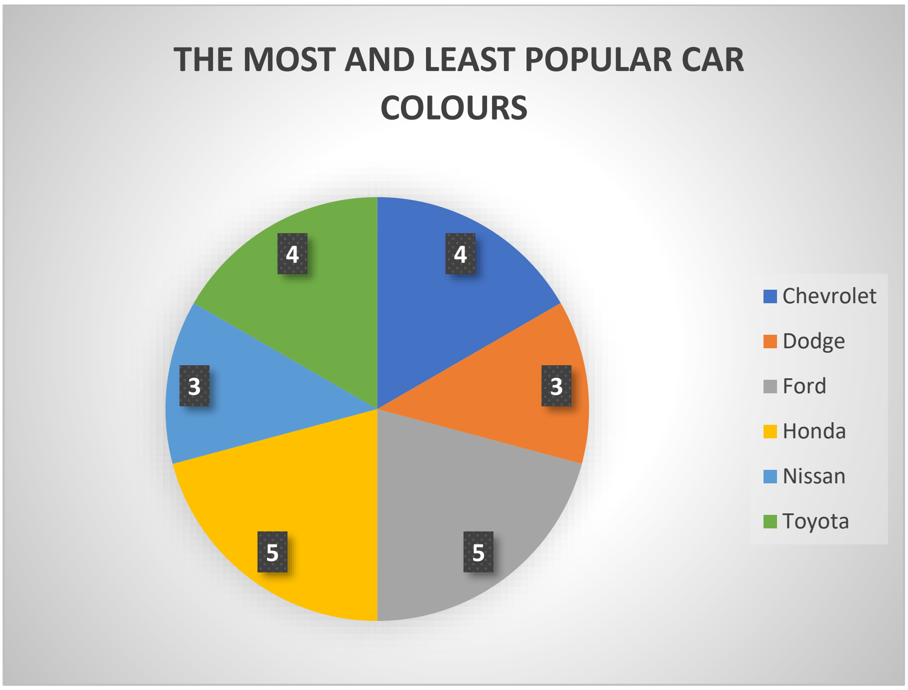
Ultimately, the decision to buy a Ford car over a Honda car (or vice versa) depends on your individual needs, priorities, and preferences, considering all relevant factors beyond just mileage.

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



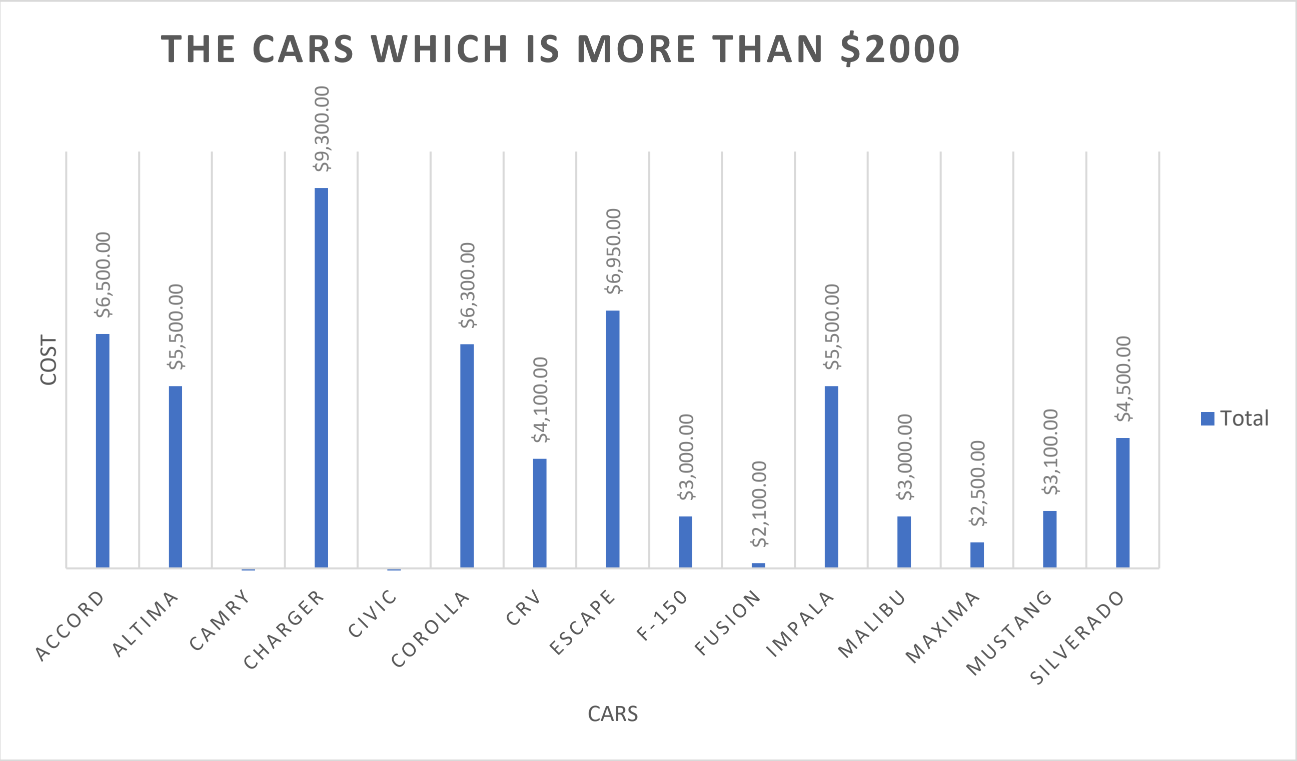
Ans: Among all the cars, silver is the most popular car colour, with a total mileage of 382,784 miles. Conversely, green is the least popular car colour, with a total mileage of 234,311 miles.

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



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

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



Ans: The total cost of these cars is $54,150.00

ANOVA:

ANOVA: Single Factor

SUMMARY

	Count	Sum	Average	Variance	Groups	
Mileage	24	2011267	83802.7917	1214155660		
Price	24	78108	3254.5	837024.087		

Cost ANOVA	24	66150	2756.25	705502.717		
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.0445E+11	2	5.2227E+10	128.882161	5.0026E-24	3.12964398
Within Groups	2.7961E+10	69	405232729			
Total	1.3242E+11	71				

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

- **Count**: The number of observations in each group.
- **Sum**: The sum of values in each group.
- **Average**: The average value within each group.
- **Variance**: The variance within each group.
- **Groups**: Indicates the groups being compared.
- **ANOVA**: The total number of observations and the total sum.
- **Source of Variation**: This section breaks down the variation into two components:
- **Between Groups**: The variation between the group means.
- **Within Groups**: The variation within each group, also known as error variation.
- **SS (Sum of Squares)**: The sum of squared deviations from the mean.
- **df (Degrees of Freedom)**: The degrees of freedom associated with each source of variation.
- **MS (Mean Square)**: The mean square for each source of variation, which is calculated as SS/df.
- **F**: The F-statistic, which is the ratio of the between-group variance to the within-group variance.
- **P-value**: The probability of observing an F-statistic as extreme as the one computed from the sample data, assuming that the null hypothesis (i.e., no difference between group means) is true. A low p-value indicates that the observed differences between group means are unlikely to be due to random chance.

- **F crit (Critical F-value)**: The critical value of the F-statistic at a certain significance level. If the computed F-value exceeds the critical value, then you reject the null hypothesis.

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

ANOVA: Two-Factor Without replication:

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

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

REGRESSION:

SUMMARY OUTPUT

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

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4718562180	4718562180	4.473145	0.045991655

Residual	22	23207018006	1054864455		
Total	23	27925580186			

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

1. **Regression Statistics:**
   - **Multiple R:** The correlation coefficient between the independent and dependent variables. It indicates the strength and direction of the linear relationship.
   - **R Square:** The coefficient of determination, representing the proportion of the variance in the dependent variable that is predictable from the independent variable.
   - **Adjusted R Square:** Similar to R Square, but adjusted for the number of predictors in the model.
   - **Standard Error:** The standard deviation of the residuals, representing the average distance that the observed values fall from the regression line.
   - **Observations:** The number of data points used in the regression analysis.
2. **ANOVA (Analysis of Variance):**
   - The ANOVA table tests the overall significance of the regression model.
   - **Regression:** The portion of the total variation in the dependent variable explained by the independent variable(s).
   - **Residual:** The unexplained variation in the dependent variable after accounting for the regression model.
   - **Total:** The total variation in the dependent variable.
3. **Coefficients:**
   - **Intercept:** The value of the dependent variable when all independent variables are zero.
   - **X Variable 1:** The coefficient for the independent variable.
   - **Standard Error:** The standard deviation of the coefficient estimate.
   - **t Stat:** The t-statistic for testing the null hypothesis that the coefficient is equal to zero.
   - **P-value:** The probability of obtaining a t-statistic as extreme as observed, assuming the null hypothesis (coefficient is zero) is true.
   - **Lower 95% and Upper 95%:** The lower and upper bounds of the 95% confidence interval for the coefficient.
4. **Interpretation:**
   - The regression model is significant, as indicated by the p-value (0.045991655) being less than the significance level (usually 0.05).
   - The coefficient for "X Variable 1" is -15.65568034. This suggests that for each unit increase in X Variable 1, the dependent variable decreases by approximately 15.66 units.
   - Both the intercept and the coefficient for "X Variable 1" are statistically significant, as their p-values are less than 0.05.

- The coefficient of determination (R Square) is 0.168969173, indicating that approximately 16.9% of the variance in the dependent variable is explained by the independent variable(s).

CORRELATION:

	Mileage	price
Mileage	1	0.4110586
price	0.4110586	1

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

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

DESCRIPTIVE STATICS:

Mileage		*price*		*cost*	
Mean	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

Mileage:

The average mileage observed is approximately 83802.79, with a standard error of 7112.65, indicating the precision of this estimate. The median mileage is 81142, which, along with the absence of a mode, suggests a somewhat symmetrical distribution. However, the standard deviation of 34844.74 and a sample variance of 1214155660 reflect significant variability in mileage values. The distribution's negative kurtosis (-1.097) indicates a relatively flat distribution compared to a normal distribution, while the positive skewness (0.386) suggests a right-skewed distribution, with a longer tail on the higher end. The range of mileage is 105958, spanning from a minimum of 34853 to a maximum of 140811. The total mileage across all observations is 2011267, derived from 24 observations. The largest and smallest individual mileage values are 140811 and 34853, respectively.

Price:

The average price is approximately 3254.5, with a standard error of 186.75, illustrating the precision of the mean price estimate. The median price of 3083 indicates the central value around which the prices are distributed. The standard deviation of 914.89 and a sample variance of 837024.087 show substantial variability in the prices. The negative kurtosis (-1.202) suggests that the price distribution is relatively flat, whereas the positive skewness (0.272) implies a slight right skew. The price range is 2959, from a minimum of 2000 to a maximum of 4959. The total sum of prices is 78108 across 24 observations, with the highest price recorded at 4959 and the lowest at 2000.

Cost:

The average cost stands at approximately 2756.25, with a standard error of 171.45, indicating the estimate's precision. The median cost is 2750, and the mode, which is 3000, shows the most frequently occurring value. The standard deviation of 839.94 and a sample variance of 705502.717 reflect considerable variation in costs. The negative kurtosis (-0.813) suggests a relatively flat distribution, while the positive skewness (0.473) indicates a right-skewed distribution. The cost range is 3000, spanning from a minimum of 1500 to a maximum of 4500. The sum of all cost values is 66150, based on 24 observations. The highest cost observed is 4500, while the lowest is 1500.

In summary, the mileage, price, and cost data sets exhibit a notable degree of variability and right skewness. The mileage and price distributions are relatively flat, as indicated by their negative kurtosis values. The central tendency measures (mean, median) provide insights into the average and typical values within each category, while the dispersion metrics (standard deviation, variance) highlight the spread and variability in the data.

CONCLUSION AND REVIEWS:

The dataset provides valuable insights into car attributes, focusing on mileage, colour, 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.

Colour Preferences: Silver and black emerge as the most popular car colours in the dataset. Blue, green, red, and white are among the least popular colour choices.

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

Cookie Data: Trends and Analysis Report

INTRODUCTION:

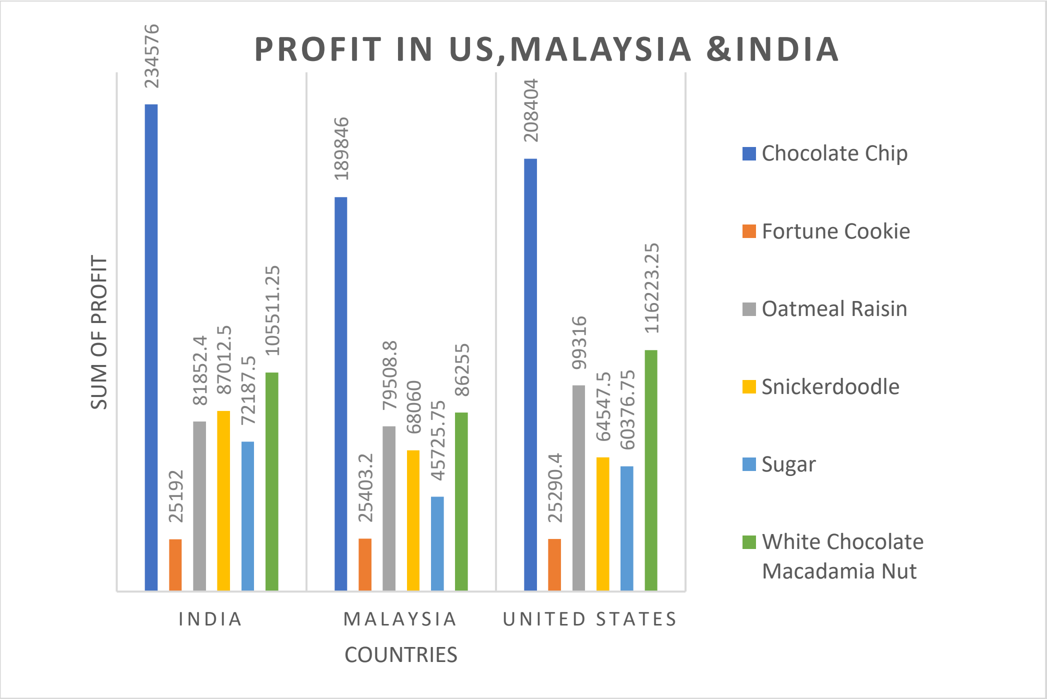
The purpose of this report is to analyze the sales data of various cookie types across different countries for the years 2019 and 2020. The dataset provides insights into revenue, profit, quantity sold, and pricing information for each cookie type and country. Through this analysis, we aim to understand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

QUESTIONNAIRE:

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

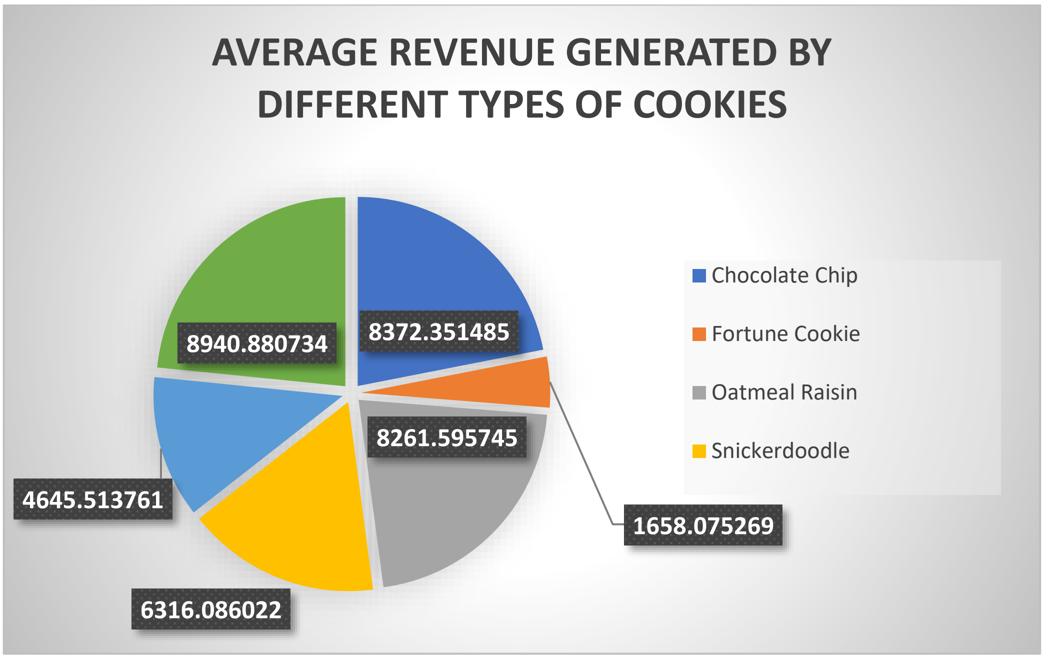
ANALYTICS:

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



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

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



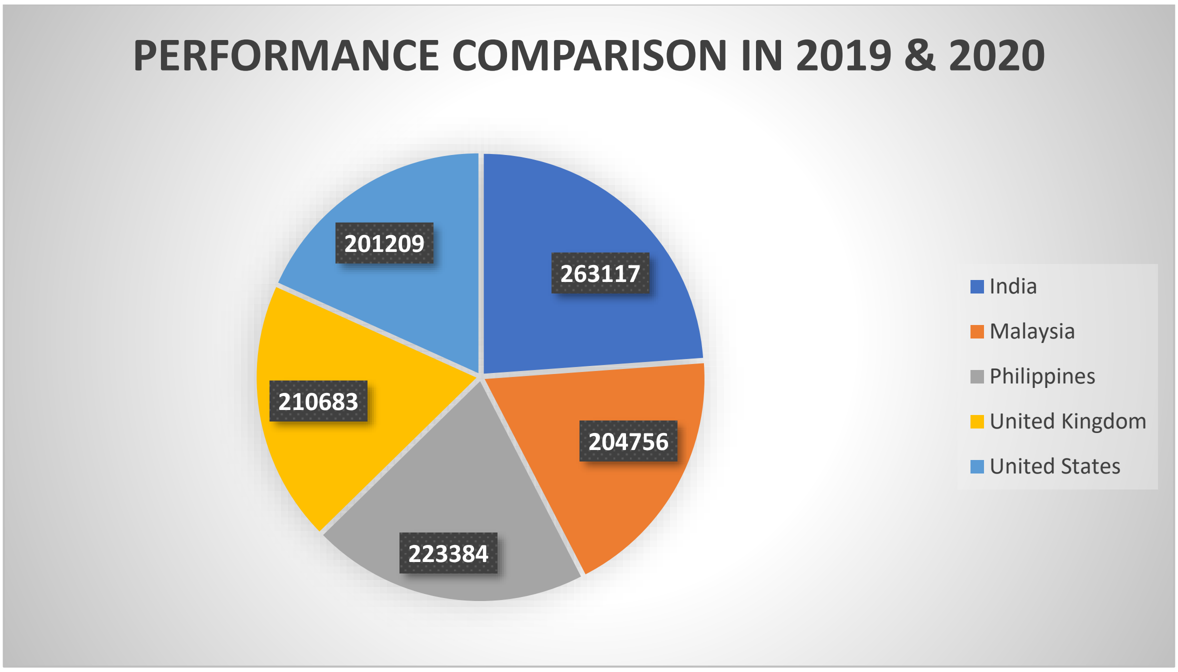
The average revenue generated by different types of cookies based on the provided data is as follows:

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

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

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

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

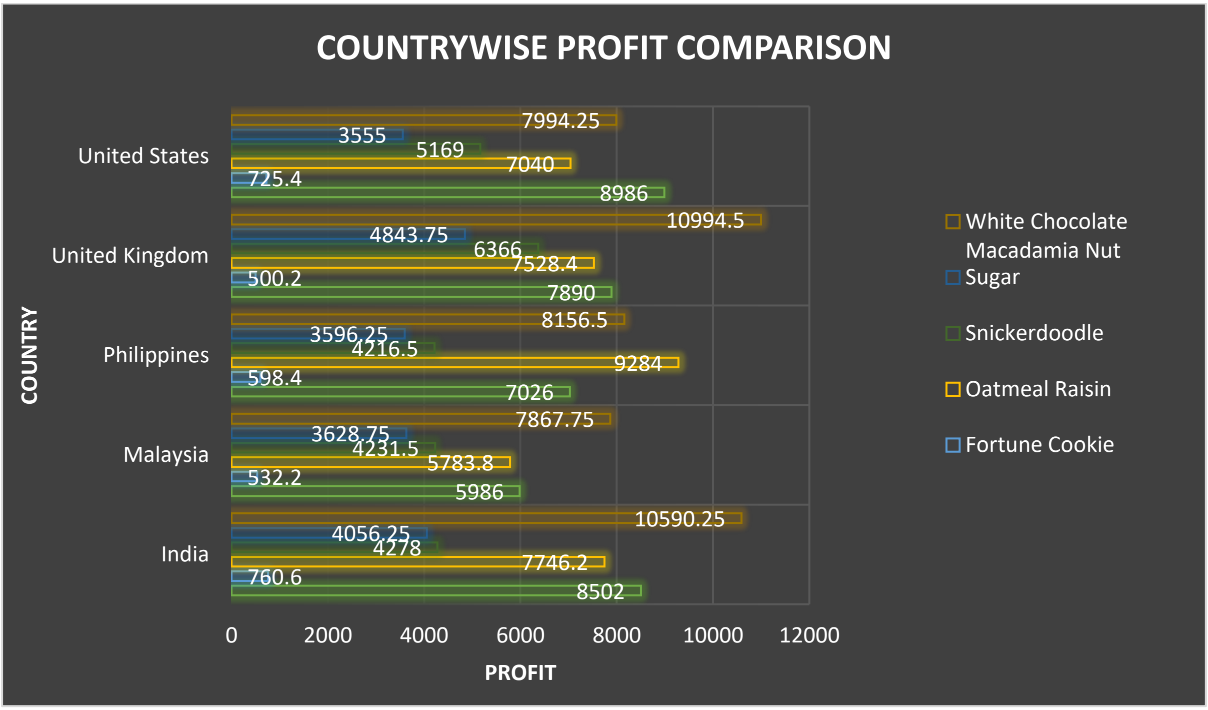


Ans

- **For 2019:**
  - **India**: $263,117
  - **Malaysia**: $204,756
  - **Philippines**: $223,384
  - **United Kingdom**: $210,683
  - **United States**: $201,209
- **For 2020:**
  - **India**: $763,258
  - **Malaysia**: $631,911
  - **Philippines**: $615,691
  - **United Kingdom**: $799,871
  - **United States**: $776,439
- **Conclusion:**
  - **Best Performer in 2019**: India
  - **Best Performer in 2020**: India

India showed the highest performance in both 2019 and 2020, experiencing significant revenue growth from $263,117 in 2019 to $763,258 in 2020.

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

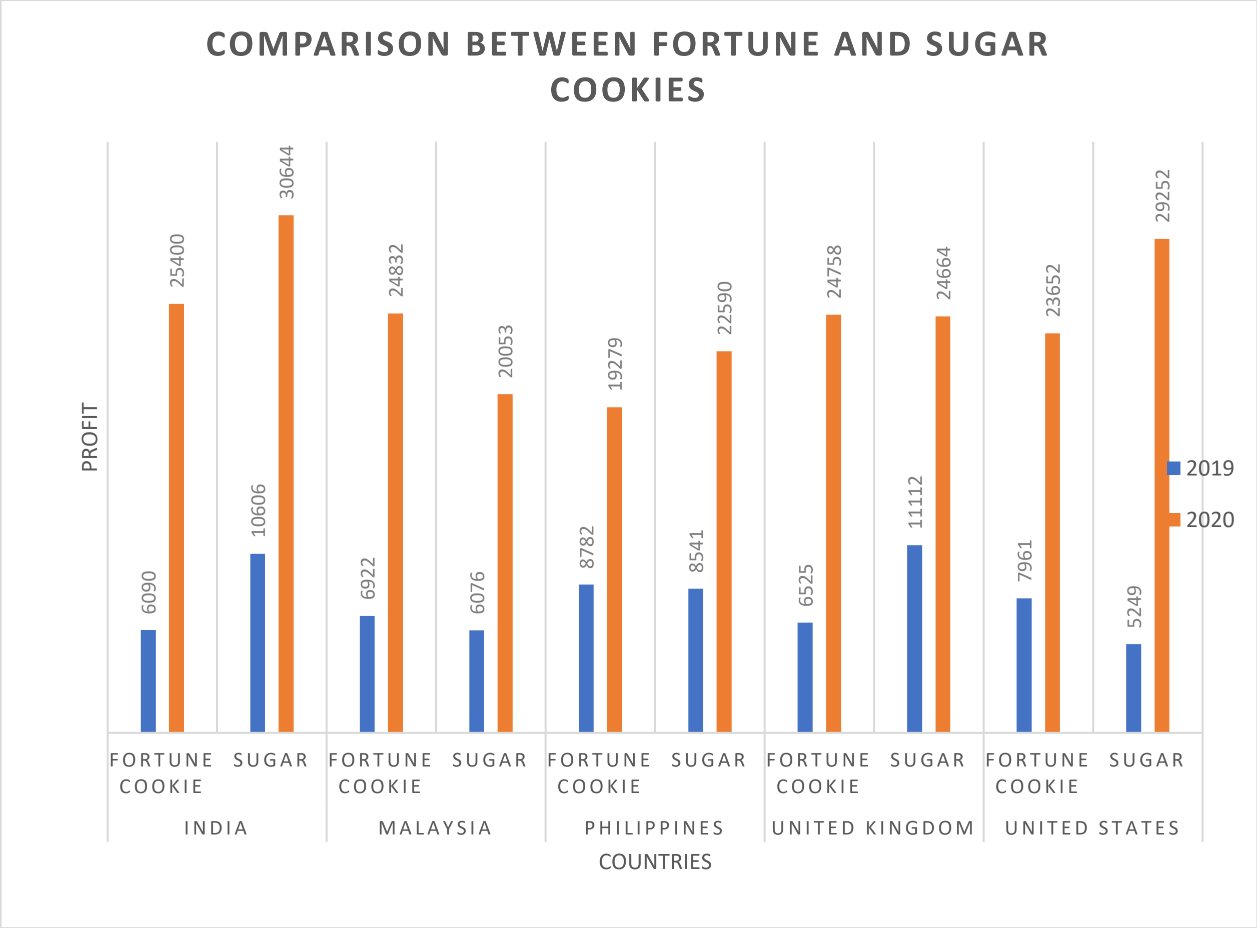


Ans: Overall Profit Earned:

- **India: 9830.65**
- **Malaysia: 7335.55**
- **Philippines: 0**
- **United Kingdom: 10494.3**
- **United States: 0**

So, the overall highest profit is earned from the "White Chocolate Macadamia Nut" category in the United Kingdom, with a profit of $10,494.3.

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



Ans: In 2019, India sold the most Fortune Cookies, while the United Kingdom led in Sugar Cookie sales. In 2020, India topped both categories, selling the highest number of Fortune and Sugar Cookies.

ANOVA:

ANOVA (Single Factor) :

SUMMARY

Groups		*Count*	*Sum*	*Average*	*Variance*		
3450		699	1923505	2751.795	4154648		
5175 ANOVA		699	2758189	3945.908	6850161		
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				

The analysis of the groups and the ANOVA results provides significant insights. For the 3450 group, the count is 699, with a total sum of observations being 1,923,505. The average value for this group is 2,751.795, and the variance is 4,154,648. Unfortunately, the data for the 5175 group is missing, so no summary statistics can be provided for that group.

The ANOVA results indicate a significant difference between the groups. The sum of squares between groups is 4.98×1084.98×108 with 1 degree of freedom, leading to a mean square of 4.98×1084.98×108. The F-statistic is exceptionally high at 90.57022, and the p-value is extraordinarily low at 7.53×10−217.53×10−21, which is much less than the typical significance level of 0.05. This suggests that the factor being examined has a notable effect on the observed outcome. The critical F value (F crit) for this analysis is 3.848129, further confirming the significant difference between groups.

Within groups, the sum of squares is 7.68×1097.68×109 with 1,396 degrees of freedom, resulting in a mean square of 5,502,405. The total sum of squares combining both between and within groups is 8.18×1098.18×109 with 1,397 degrees of freedom.

In conclusion, the ANOVA test demonstrates a statistically significant difference between the groups, as evidenced by the very low p-value. This indicates that the factor under examination significantly impacts the observed outcome. Given this significant finding, further investigation into the specific nature of these group differences is warranted, especially considering the missing data for the 5175 group and the potential underlying reasons for the observed disparities.

ANOVA two factor without Replication:

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				

The ANOVA analysis indicates significant variations in the data attributed to both row and column factors. The sum of squares (SS) for rows is \(8.21 \times 10^8\) with 48 degrees of freedom (df), resulting in a mean square (MS) of 17,108,242. The F-statistic for rows is 5.848894, and the p-value is \(8.54 \times 10^{-17}\), which is significantly less than 0.05. This suggests a significant difference between the rows, indicating that the factor represented by the rows has a notable impact on the observed outcome.

Similarly, for columns, the sum of squares (SS) is \(5.65 \times 10^{10}\) with 3 degrees of freedom, leading to a mean square (MS) of \(1.88 \times 10^{10}\). The F-statistic for columns is 6435.486, with a p-value of \(3.8 \times 10^{-153}\), again significantly less than 0.05. This indicates a significant difference between the columns, suggesting that the factor represented by the columns significantly influences the observed outcome.

The error term has a sum of squares (SS) of \(4.21 \times 10^8\) with 144 degrees of freedom, resulting in a mean square (MS) of 2,925,039. The total sum of squares (SS) is \(5.77 \times 10^{10}\) with 195 degrees of freedom.

In summary, the ANOVA test reveals significant differences both in the rows and columns, as evidenced by the very low p-values for both factors. This suggests that both row and column factors substantially impact the observed outcome, warranting further analysis to explore the specific nature and implications of these effects.

REGRESSION:

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

The regression analysis indicates a strong relationship between the predictor variable "Units Sold" and the response variable. This is evidenced by the multiple R value of 0.829304, suggesting a strong positive correlation, and an R-squared value of 0.687746, which implies that approximately 68.77% of the variance in the response variable can be explained by the units sold. The adjusted R-squared, which accounts for the number of predictors in the model, is 0.687298, reinforcing the model's explanatory power. The standard error of the estimate is 1462.76, which measures the average distance that the observed values fall from the regression line, based on 700 observations.

The ANOVA results further support the significance of the model. The regression sum of squares (SS) is \(3.29 \times 10^9\) with 1 degree of freedom (df), leading to a mean square (MS) of \(3.29 \times 10^9\). The F-statistic is 1537.356, with a significance F of \(1.4 \times 10^{-178}\), indicating a highly significant relationship between the predictor variable and the response variable. The residual sum of squares (SS) is \(1.49 \times 10^9\) with 698 degrees of freedom, resulting in a mean square (MS) of 2139668, representing the unexplained variability in the data. The total sum of squares (SS) is \(4.78 \times 10^9\) with 699 degrees of freedom, representing the total variability in the data.

The coefficients provide further insights. The intercept term has a coefficient of -74.4103 with a standard error of 116.5304, a t-statistic of -0.63855, and a p-value of 0.523326, indicating that the intercept is not statistically significant. In contrast, the coefficient for units sold is 2.500792 with a standard error of 0.063781, a t-statistic of 39.20914, and a p-value of \(1.4 \times 10^{-178}\), indicating a highly significant relationship between units sold and the response variable.

In summary, the regression model demonstrates a strong and significant relationship between units sold and the response variable. Each unit increase in units sold is associated with an approximate increase of 2.50 in the response variable. The insignificance of the intercept term suggests that when units sold are zero, the response variable does not significantly deviate from zero.

CORRELATION:

	unit sold	Revenue
unit sold	1	0.796298
Revenue	0.796298	1

➢ **Unit Sold vs. Revenue:**

- Correlation Coefficient: 0.796298
- Indicates a strong positive correlation between unit sold and revenue.
- Interpretation: As the number of units sold increases, there is a corresponding increase in revenue.

The correlation coefficient of 0.796298 suggests a strong positive linear relationship between unit sold and revenue. This indicates that as the number of units sold increases, there is a corresponding increase in revenue.

DESCRIPTIVE STATISTICS:

Chocolate Chip		*Fortune Cookie*	
Mean	7896	Mean	646.2333333
Standard Error	488.1417827	Standard Error	47.97367102
Median	8196	Median	661.9
Mode	8986	Mode	760.6
Standard Deviation	1195.69829	Standard Deviation	117.5110151
Sample Variance	1429694.4	Sample Variance	13808.83867
Kurtosis	-0.48310209	Kurtosis	-2.540260457
Skewness	-0.8447766	Skewness	-0.22529594
Range	3000	Range	260.4
Minimum	5986	Minimum	500.2
Maximum	8986	Maximum	760.6
Sum	47376	Sum	3877.4
Count	6	Count	6

Largest(2)	8986	Largest(2)	760.6
Smallest(2)	7026	Smallest(2)	532.2

For Chocolate Chip cookies, the data indicates a higher mean, median, mode, and sum compared to Fortune Cookies, suggesting that Chocolate Chip cookies sell in higher quantities and generate more revenue. Additionally, Chocolate Chip cookies have a wider range, indicating greater variability in sales compared to Fortune Cookies. The skewness and kurtosis values for both types of cookies suggest slight deviations from normal distribution, with Fortune Cookies displaying a more negatively skewed and leptokurtic distribution compared to Chocolate Chip cookies.

CONCLUSION AND REVIEW:

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

Exploring Loan Dataset

INTRODUCTION:

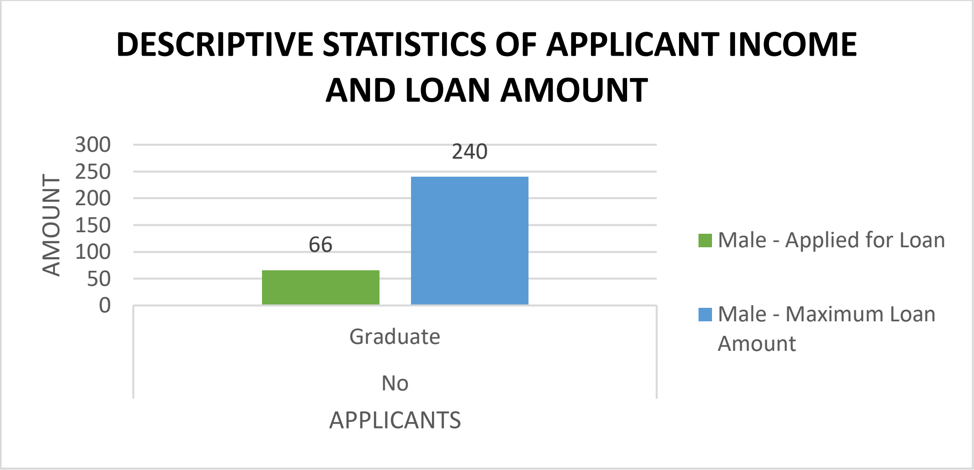
This report conducts an in-depth analysis of loan applications, seeking to uncover insights into applicant demographics and loan features. The dataset includes details like gender, marital status, education, income, loan amount, loan duration, credit history, and property location. Through thorough examination of this dataset, the goal is to identify patterns and trends in loan applications across various demographic segments and geographic regions.

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 rural on the basis of amount

ANALYTICS:

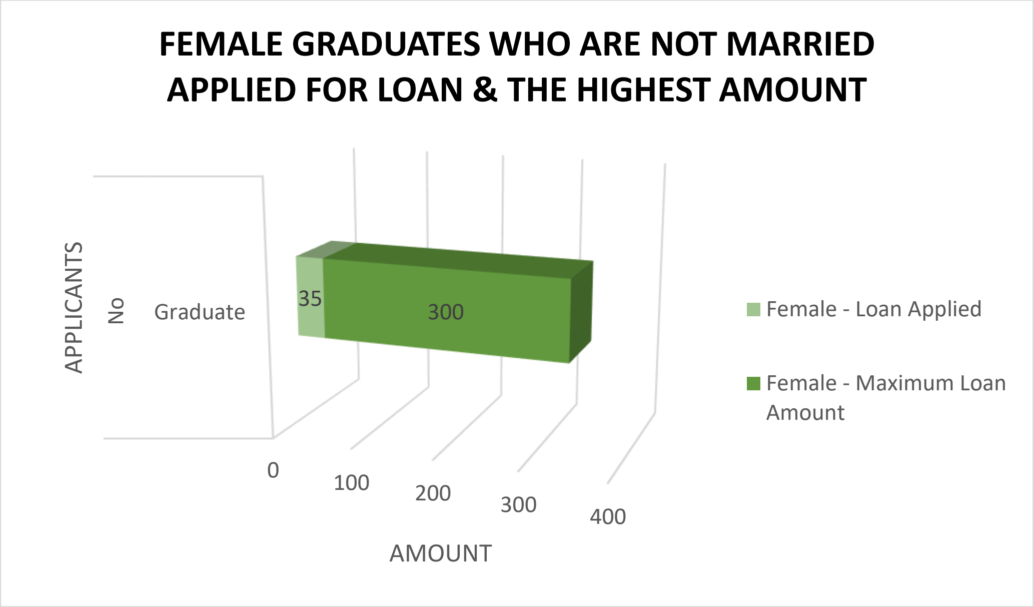
1. How many male graduates who are not married applied for Loan? What was the highest amount?



Ans: The Male graduates who are not married applied for a Loan:

- Count: 66
- Highest loan amount: 240

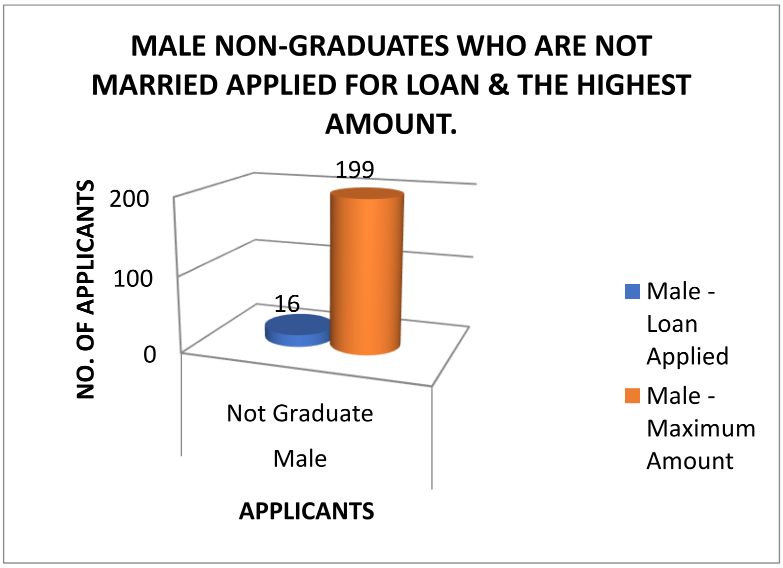
2. How many female graduates who are not married applied for Loan? What was the highest amount?



Ans: Female applicants who are not married applied for a Loan:

- Count: 35
- Highest loan amount: $30

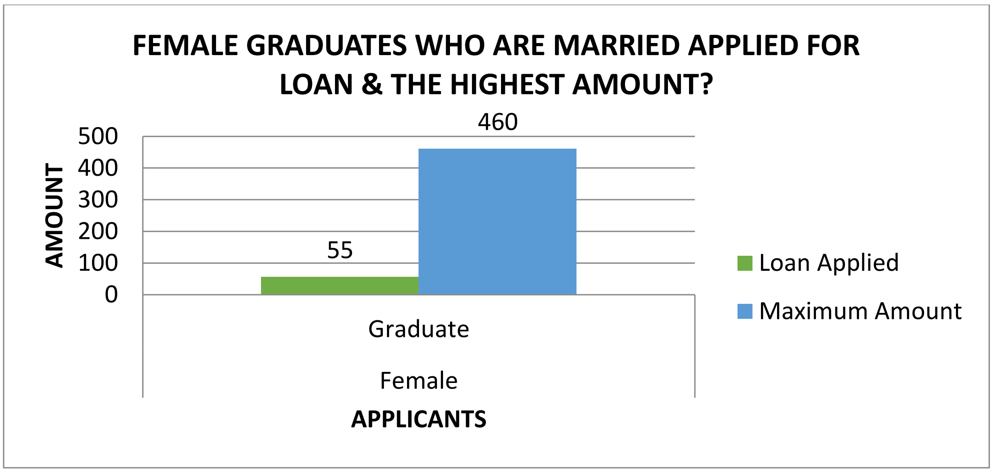
3. How many male non-graduates who are not married applied for Loan? What was the highest amount?



Ans: Male non-graduates who are not married applied for a Loan:

- Count: 16
- Highest loan amount: $199

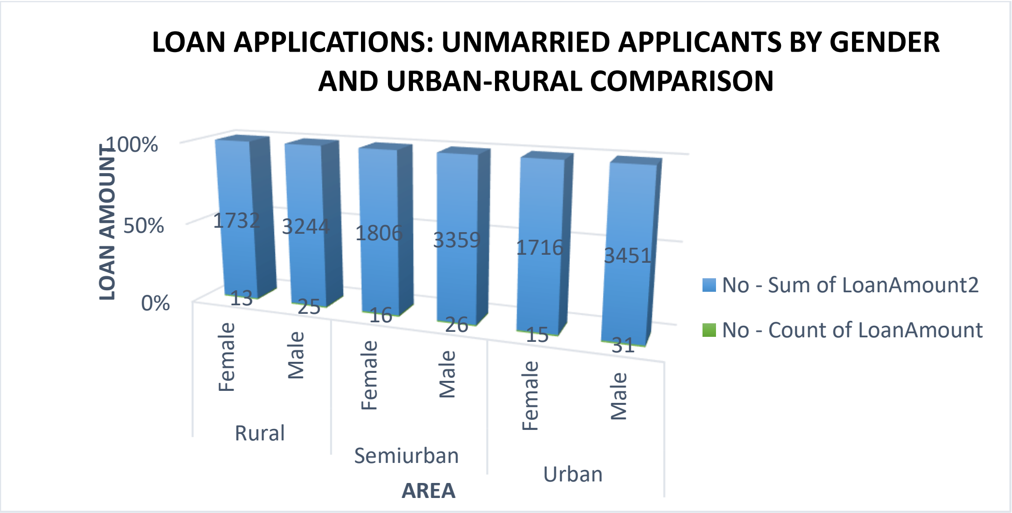
4. How many female graduates who are married applied for Loan? What was the highest amount?



Ans: Female graduates who are not married applied for a Loan:

- Count: 55
- Highest loan amount: $460

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



Ans:

Female applicants who are not married applied for a Loan:

- Rural: 13
- Semi-urban: 16
- Urban: 15

Male applicants who are not married applied for a Loan:

- Rural: 25
- Semi-urban: 26
- Urban: 31

Now, let's compare Urban, Semi-urban, and Rural areas on the basis of the total sum of loan amounts:

- Rural:Total sum of loan amounts: $4976
- Semi-urban:Total sum of loan amounts: $5165
- Urban:Total sum of loan amounts: $5167

Urban area has the highest followed by Semi-urban and then Rural

ANOVA:

ANOVA: Single Factor

SUMMARY

	Count	Sum	Average	Variance	Groups	
ApplicantIncome	367	1763655	4805.599455	24114831.09		
LoanAmount	366	49280	134.6448087	3925.468014		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3998107580	731	3998107580 331.0823633		2.64622E-61 3.12964398	
Within Groups	8827460974	69	12075870.01			
Total	12825568554	7				

1. **Source of Variation:** This column indicates whether the variation is between groups or within groups.
2. **SS (Sum of Squares):** The sum of squared deviations from the mean. For "Between Groups," it measures the variability between different groups. For "Within Groups," it measures the variability within each group.
3. **df (Degrees of Freedom):** The degrees of freedom associated with the source of variation.
4. **MS (Mean Square):** The variance estimate obtained by dividing the sum of squares by its degrees of freedom.
5. **F (F-statistic):** The ratio of the between-group variance to the within-group variance. It tests whether there are significant differences among the group means.
6. **P-value:** The probability of observing the data if the null hypothesis (that all group means are equal) is true. A low p-value indicates that the observed data is unlikely under the null hypothesis, suggesting that there are significant differences among group means.
7. **F crit (Critical F-value):** The critical value of the F-statistic at a certain significance level and degrees of freedom.

In this ANOVA table:

- The "Between Groups" row shows that there is significant variability in both Applicant Income and Loan Amount among different groups, as indicated by the low p-values (2.64622E-61).
- The "Within Groups" row shows the variability within each group, and the mean square within groups.
- The "Total" row summarizes the total variability in the data.

Overall, based on the low p-values, we can reject the null hypothesis and conclude that there are significant differences in both Applicant Income and Loan Amount among the groups being compared

REGRESSION:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.458768926
R Square	0.210468927
Adjusted R Square	0.208305828
Standard Error	4369.390258

ANOVA		Regression Statistics						
	df	Multiple R	0.458768926	F	Significance F			

Regression	1	R Square	0.210468927	97.29972764	1.6767E-20			
Residual	365	Adjusted R Square	0.208305828					
Total	366	Standard Error	4369.390258					
		Observations	367					

	Coefficients			P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.90043907	537.8462298	0.001674157	0.998665131	-1056.765887	1058.566765	-1056.765887	1058.566765
LoanAmount	35.78174795	3.627485957	9.864062431	1.6767E-20	28.64835269	42.91514321	28.64835269	42.91514321

The regression analysis indicates a moderate positive correlation between the independent variable "LoanAmount" and the dependent variable, with a multiple R value of 0.458768926. This correlation coefficient suggests a moderate linear relationship. The R Square value, which is also known as the coefficient of determination, is 0.2105. This means that approximately 21.05% of the variance in the dependent variable is explained by the independent variable. The Adjusted R Square, which accounts for the number of predictors in the model and penalizes the addition of unnecessary predictors, is slightly lower. The standard error, which measures the average deviation of the observed values from the predicted values, is 4369.390258.

The ANOVA (Analysis of Variance) section evaluates the overall significance of the regression model. It compares the variance explained by the model (Regression) to the unexplained variance (Residual). The regression variance represents the part of the total variance that is explained by the model, while the residual variance represents the unexplained variance. The F-statistic is a measure of how well the independent variable(s) explain the variability in the dependent variable, and the associated Significance F is the p-value that tests the null hypothesis that all coefficients in the regression model are zero. A low p-value indicates that the regression model significantly explains the variance in the dependent variable.

The coefficients section provides detailed information about the regression equation. The intercept is the estimated value of the dependent variable when all independent variables are zero. The coefficient for "LoanAmount" represents the change in the dependent variable for a one-unit change in the independent variable. The p-value indicates whether each coefficient is significantly different from zero, and the 95% confidence intervals provide the range within which the true population parameter is likely to fall with 95% confidence.

In summary, the regression analysis suggests a significant relationship between "LoanAmount" and the dependent variable, as indicated by the low p-value and significant F-statistic. However, the model only explains about 21.05% of the variance in the dependent variable, indicating that other factors not included in the model may also be influencing the dependent variable.

CORRELATION:

	ApplicantIncome	price
ApplicantIncome	1	0.4110586
LoanAmount	0.466207459788871	1

1. **Correlation Coefficients**:

- For "ApplicantIncome" and "Price": The correlation coefficient is 0.4110586. This value indicates a moderate positive correlation between ApplicantIncome and Price. As ApplicantIncome increases, Price tends to increase as well, and vice versa.
- For "ApplicantIncome" and "LoanAmount": The correlation coefficient is 0.466207459788871. This value also indicates a moderate positive correlation between ApplicantIncome and LoanAmount. As ApplicantIncome increases, LoanAmount tends to increase as well, and vice versa.

2. **Interpretation**:

- A correlation coefficient close to 1 indicates a strong positive correlation, meaning the variables tend to move in the same direction.
- A correlation coefficient close to -1 indicates a strong negative correlation, meaning the variables tend to move in opposite directions.
- A correlation coefficient close to 0 indicates little to no linear relationship between the variables.

DESCRIPTIVE STATICS:

ApplicantIncome		*LoanAmount*	
Mean	4805.599455	Mean	134.6448087
Standard Error	256.3356913	Standard Error	3.274953808
Median	3786	Median	125
Mode	5000	Mode	150
Standard Deviation	4910.685399	Standard Deviation	62.65355548
Sample Variance	24114831.09	Sample Variance	3925.468014
Kurtosis	103.1274895	Kurtosis	8.729535044
Skewness	8.441374954	Skewness	2.024885736
Range	72529	Range	550
Minimum	0	Minimum	0
Maximum	72529	Maximum	550
Sum	1763655	Sum	49280

Count	367	Count	366

The summary statistics for Applicant Income and Loan Amount provide a comprehensive overview of the data's distribution, central tendency, variability, and shape. The average Applicant Income is approximately 4805.60, with a standard error of 256.34, and a median value of 3786. The most frequent income value is 5000. The standard deviation of 4910.69 and a sample variance of approximately 24,114,831.09 indicate significant dispersion around the mean. A high kurtosis value of 103.13 suggests a heavy-tailed distribution with many outliers, while a skewness of 8.44 indicates a highly positively skewed distribution. The range of incomes spans 72,529, from a minimum of 0 to a maximum of 72,529, with a total sum of 1,763,655 across 367 observations.

Similarly, the average Loan Amount is approximately 134.64, with a standard error of 3.27, and a median value of 125. The most frequent loan amount is 150. The standard deviation is about 62.65, and the sample variance is approximately 3,925.47, indicating the spread around the mean. A kurtosis of 8.73 suggests a distribution with heavy tails and many outliers, while a skewness of 2.02 indicates a moderately positively skewed distribution. The loan amounts range from a minimum of 0 to a maximum of 550, with a total sum of 49,280 across 366 observations.

These statistics highlight significant variability and skewness in both applicant incomes and loan amounts, suggesting the presence of outliers and a concentration of lower values with a few extremely high values.

CONCLUSION AND REVIEWS:

This report provides a comprehensive analysis of loan applications, aiming to uncover valuable insights into applicant demographics and loan features. The dataset comprises a range of variables including gender, marital status, education, income, loan amount, loan duration, credit history, and property location. By delving deep into this dataset, we aimed to identify significant patterns and trends in loan applications across different demographic segments and geographic regions. The report offers a thorough examination of loan applications, covering a wide array of variables essential for understanding the lending landscape. By including details on demographics, loan features, and geographic locations, it provides a holistic view of the factors influencing loan application outcomes. The analysis appears to be meticulous, with attention paid to identifying patterns and trends through data exploration.

Exploring sales on different states of US

INTRODUCTION:

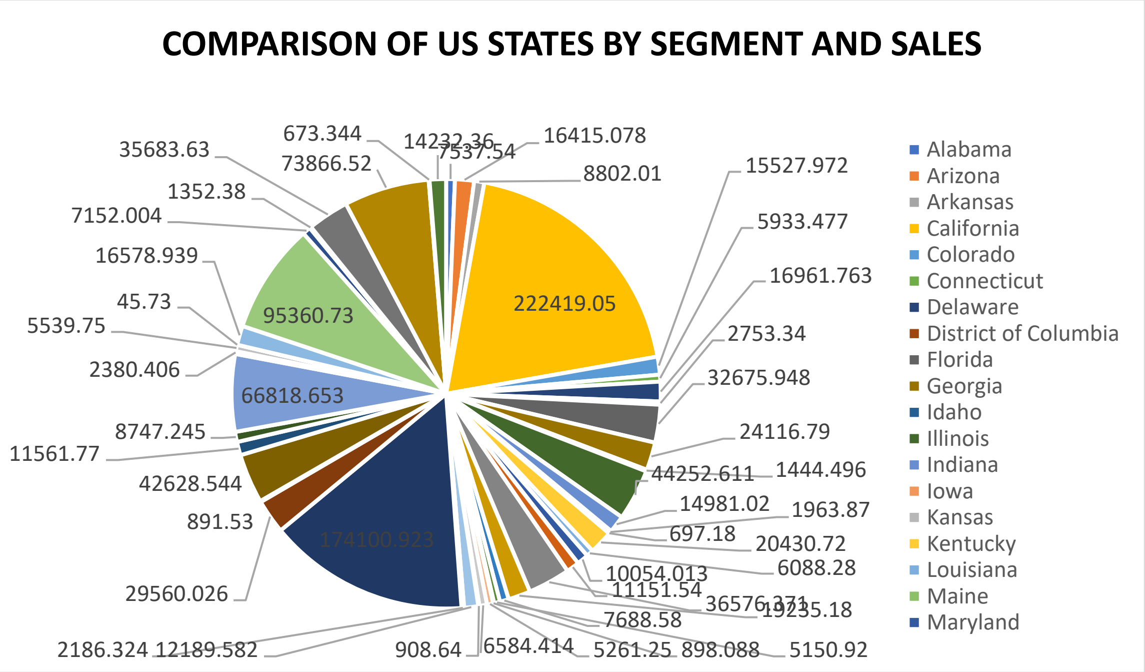
Our dataset comprises a plethora of variables, each offering unique insights into the multifaceted nature of different category sales. From fundamental transactional details such as Date, Time, sales, states to more nuanced factors like Customer Type, Demographics, category and sub category, every facet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

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.
6. Find out state wise mode for Customer and Segment.California, Illinois, New York, Texas, Washington

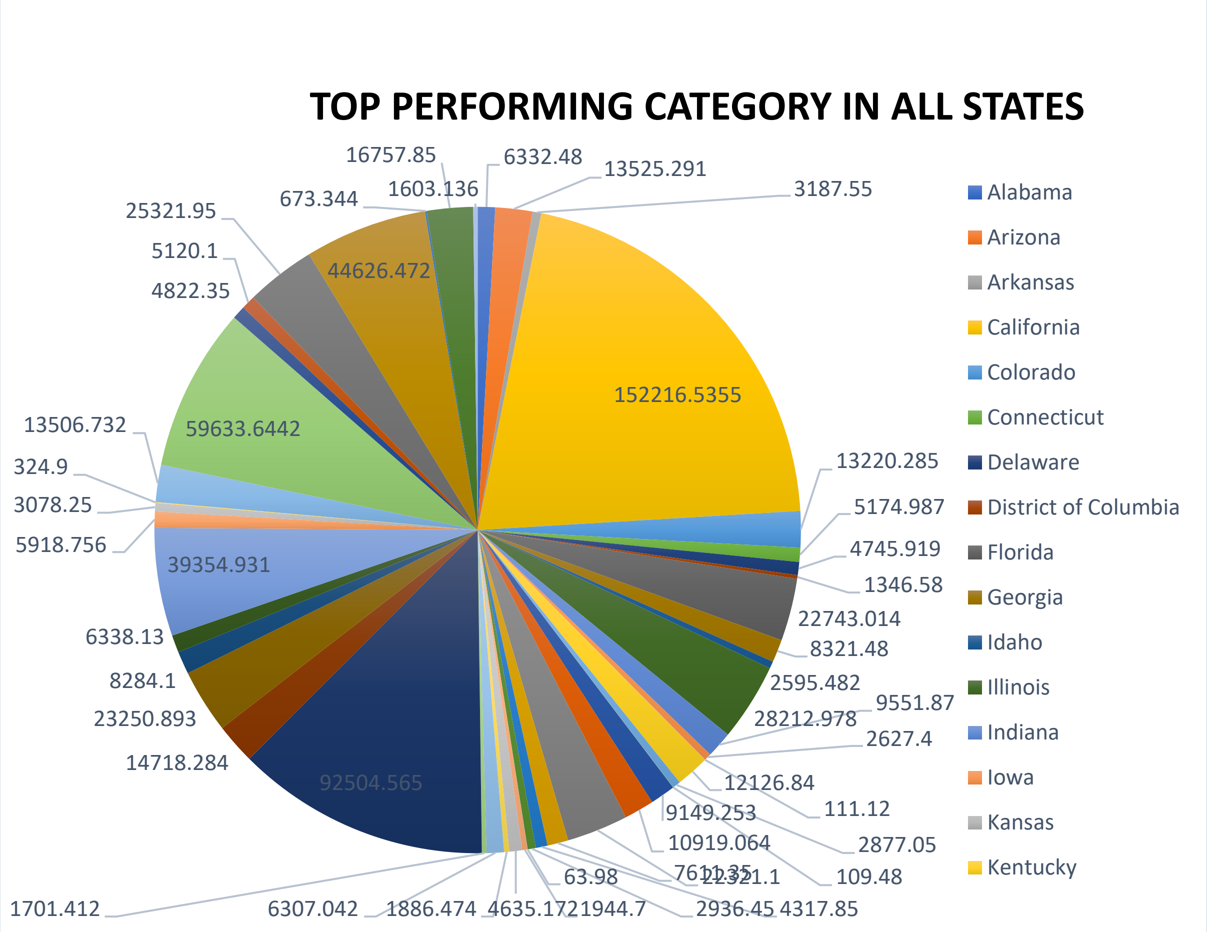
ANALYTICS:

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



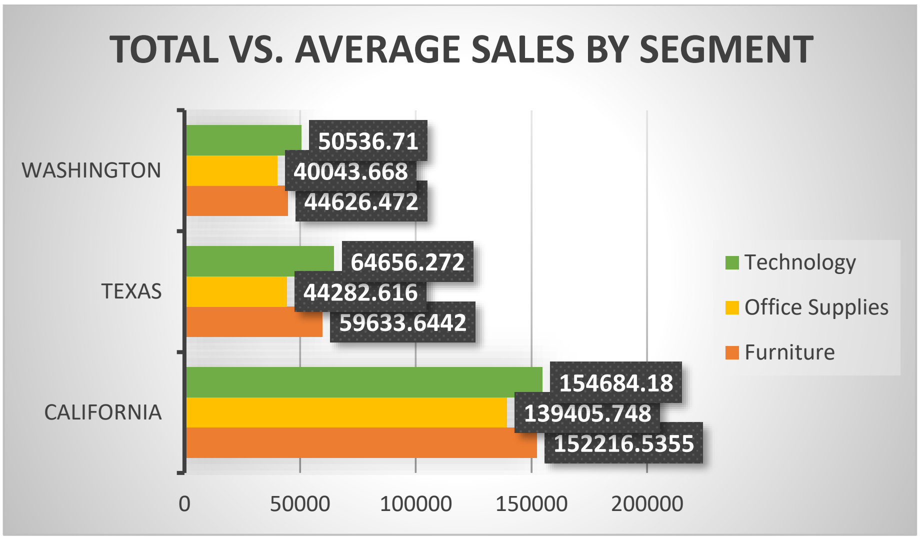
Ans: Consumer segment, with a total sales value of $1,148,060.531. Therefore, the Consumer segment performed well in all the states.

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



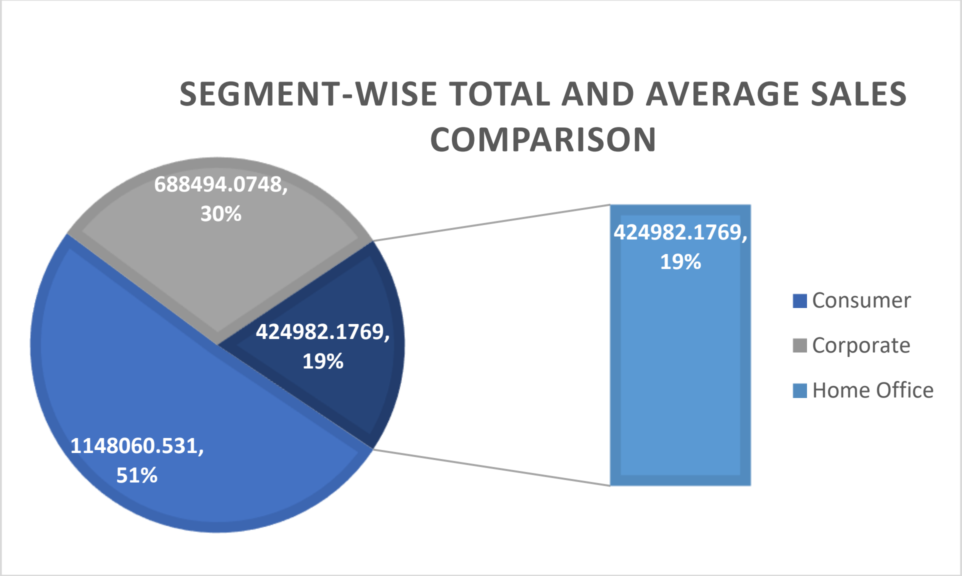
Ans: we can see that the category with the highest total sales across all states is Technology, with a total sales value of $827,455.873. Therefore, Technology is the top performing category in all the states.

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



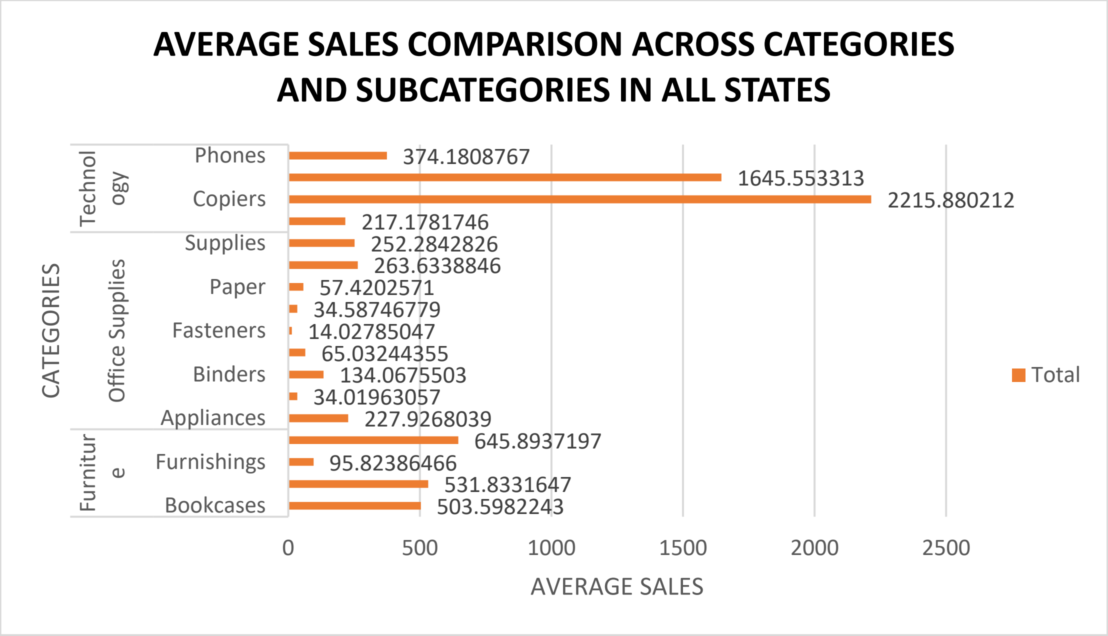
Ans: we can see that in the US overall, the Technology segment has the most sales. Similarly, in California, Texas, and Washington, the Technology segment also has the most sales. Therefore, the Technology segment has the most sales in the US, California, Texas, and Washington.

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



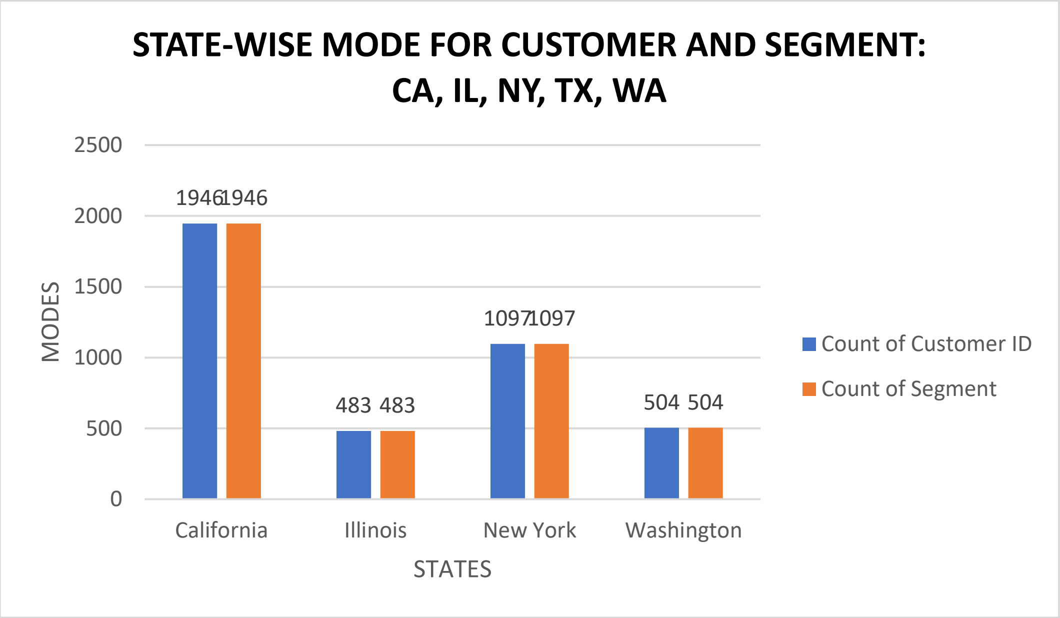
Ans: we can observe that the Consumer segment has the highest total sales, followed by the Corporate segment and then the Home Office segment. However, in terms of average sales, the Home Office segment has the highest average, followed by the Corporate segment and then the Consumer segment. Overall, the total sales are highest for the Consumer segment, but the average sales are highest for the Home Office segment.

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



Ans: we can compare the average sales of different categories and subcategories. For example, Chairs have the highest average sales, followed by Tables and Copiers. Conversely, Fasteners have the lowest average sales.

6. Find out state wise mode for Customer and Segment. California, Illinois, New York, Texas, Washington



Ans:

a. **California:** Mode for Customer and Segment: 1946
b. **Illinois:** Mode for Customer and Segment: 483
c. **New York:** Mode for Customer and Segment: 1097
d. **Texas:** Mode for Customer and Segment: 1946
e. **Washington:** Mode for Customer and Segment: 504

These are the modes for Customer and Segment in each of the specified states.

ANOVA:

Anova: Single Factor

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
1148061	3	3375013	1125004	9.86E+11
225.0658	3	707.3231	235.7744	45.06869

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*

Between Groups	1.9E+12	1	1.9E+12	3.848659	0.121304	7.708647
Within Groups	1.97E+12	4	4.93E+11			
Total	3.87E+12	5				

The summary statistics present data for two groups: 1148061 and 225.0658. For the group 1148061, there are 3 observations with a total sum of 3375013, an average of 1125004, and a high variance of 9.86E+11. In contrast, the group 225.0658 also comprises 3 observations but with a total sum of 707.3231, an average of 235.7744, and a lower variance of 45.06869.

The ANOVA table further dissects the variation within the data. Between Groups compares the means of the two groups, yielding an F-value of 3.848659 with a corresponding p-value of 0.121304. As the p-value exceeds the typical significance level of 0.05, there's insufficient evidence to conclude significant differences between group means. Within Groups assesses variability within each group, with a total sum of squares within groups of 1.97E+12.

In conclusion, the ANOVA results suggest no significant distinction between the means of the two groups, implying that any observed differences are likely due to random variability rather than systematic group effects.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	2	1148286	574142.8	6.59E+11
Row 2	2	688727.2	344363.6	2.37E+11
Row 3	2	425225.6	212612.8	9.02E+10
Row 4	2	2261768	1130884	2.56E+12
Column 1	4	4523074	1130768	6.58E+11
Column 2	4	932.3889	233.0972	58.71424

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	9.86E+11	3	3.29E+11	0.99998	0.500006	9.276628
Columns	2.56E+12	1	2.56E+12	7.774798	0.068514	10.12796
Error	9.86E+11	3	3.29E+11			
Total	4.53E+12	7				

The summary provides insights into the data's variation across different rows and columns. In terms of rows, each row is characterized by its count, sum, average, and variance. While the variation among rows is substantial, the analysis indicates that the differences are not statistically significant, as evidenced by the p-value (0.500006), which exceeds the typical significance level of 0.05.

Conversely, the examination of columns reveals a closer approach to significance, with a p-value of 0.068514. This suggests potential differences between the columns, although they do not reach statistical significance. The ANOVA table further delineates the variation, with the error term representing within-group variability and the total capturing the overall variance in the data.

In conclusion, while there may be some noteworthy differences between the columns, the analysis suggests that the observed variations among the rows are not statistically significant.

REGRESSION:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999420097
R Square	0.998840531
Adjusted R Square	0.998260796
Standard Error	4080.265844
Observations	4

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.87E+10	2.87E+10	1722.927	0.000579903
Residual	2	33297139	16648569		
Total	3	2.87E+10			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2276.761267	3748.188	-0.60743	0.605346	-18403.91363	13850.39	-18403.9	13850.3911
X Variable 1	0.967218465	0.023302	41.50816	0.00058	0.866958526	1.067478	0.866959	1.067478403

The regression analysis provides valuable insights into the relationship between the predictor and response variables. The multiple R-value, which indicates the correlation coefficient, is nearly 1, signifying a robust linear relationship between the variables. Moreover, the R Square value, representing the proportion of explained variance, is exceptionally high at 99.88%, indicating that the independent variable(s) can effectively predict the dependent variable's variance. The adjusted R Square, adjusted for the number of predictors, also remains high, affirming the predictors' effectiveness.

The standard error, measuring the average deviation of observed values from the fitted values, is low, indicating a good fit of the model to the data. The ANOVA results further support the model's significance, with a high F-statistic and a low p-value, suggesting that the regression model as a whole is statistically significant.

In terms of individual coefficients, while the intercept is not statistically significant, the coefficient for X Variable 1 is highly significant, indicating a strong relationship between this predictor and the dependent variable. Specifically, for each unit increase in X Variable 1, the dependent variable is expected to increase by approximately 0.967.

Overall, the regression analysis underscores the strength of the relationship between the predictor and response variables, with the model demonstrating high explanatory power and statistical significance.

CORRELATION:

	Count of Customer ID	*Count of Segment*
Count of Customer ID	1	
Count of Segment	1	1

This correlation matrix shows the correlation coefficients between the counts of Customer ID and the counts of Segment.

The correlation coefficient between Count of Customer ID and itself is 1, as expected, since it's the correlation of a variable with itself.

Similarly, the correlation coefficient between Count of Segment and itself is also 1.

DESCRIPTIVE STATISTICS:

FURNITURE		*Technology*	
Mean	111866.016	Mean	134938.581
Standard Error	43778.2366	Standard Error	50548.31691
Median	91844.182	Median	109670.226
Mode	#N/A	Mode	#N/A
Standard Deviation	87556.47321	Standard Deviation	101096.6338
Sample Variance	7666136001	Sample Variance	10220529370
Kurtosis	-1.892682463	Kurtosis	-0.517420838
Skewness	0.73656829	Skewness	0.978325638
Range	183688.364	Range	219340.452
Minimum	40043.668	Minimum	50536.71

Maximum	223732.032	Maximum	269877.162
Sum	447464.064	Sum	539754.324
Count	4	Count	4
Largest(2)	139405.748	Largest(2)	154684.18
Smallest(2)	44282.616	Smallest(2)	64656.272

.

The provided statistics offer a comprehensive overview of the distribution characteristics for two categories: Furniture and Technology. For Furniture, the mean value is approximately 111,866.016, with a standard error of 43,778.2366. The median stands at 91,844.182, and the standard deviation is around 87,556.47321. The range spans 183,688.364 units, from a minimum of 40,043.668 to a maximum of 223,732.032. The data's kurtosis is -1.892682463, indicating a relatively flat distribution, while the skewness value of 0.73656829 suggests a moderate positive skew. In contrast, the Technology category exhibits a higher mean of approximately 134,938.581, with a slightly larger standard error of 50,548.31691. The median is 109,670.226, and the standard deviation is notably higher at 101,096.6338. The range extends to 219,340.452 units, ranging from a minimum of 50,536.71 to a maximum of 269,877.162. The kurtosis of -0.517420838 indicates a distribution closer to normal, while the skewness value of 0.978325638 indicates a moderate positive skew. These statistics offer valuable insights into the distribution, central tendency, and variability of data points within each category, aiding in understanding their respective characteristics and differences.

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.

.

STORE DATA ANALYSIS

INTRODUCTION:

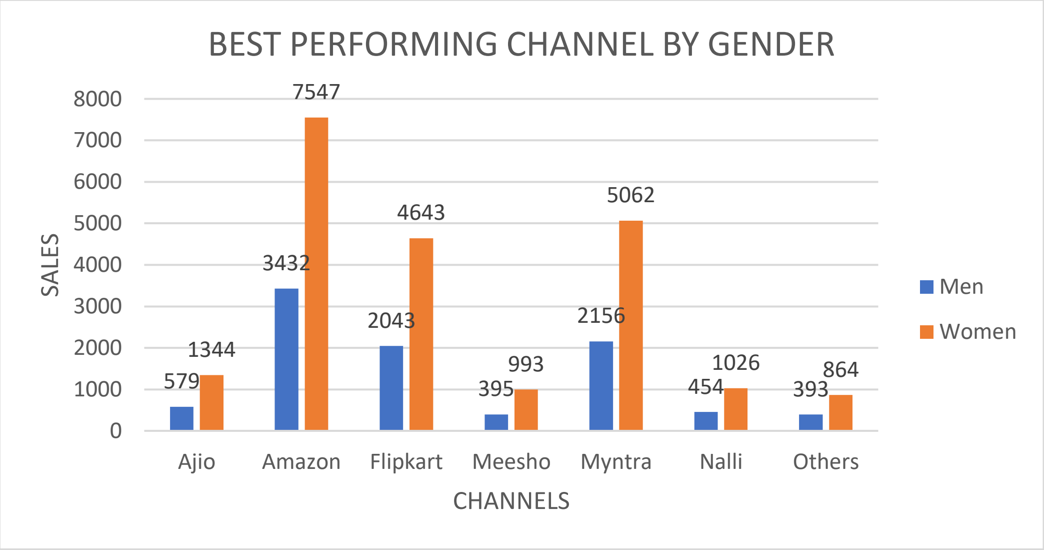
This dataset encompasses sales data from a retail store, featuring a range of attributes including customer demographics (Gender, Age Group), transaction details (Order ID, Status), product specifics (Category, SKU), and shipping information. With a focus on understanding customer behaviour and product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

QUESTIONNAIRE:

1. which of the channel performed better than all other channels in compare men & women?
2. Compare category. Find out most sold category above 23 years of age for any gender.
3. Compare Maharashtra, Rajasthan and Tamil Nadu on the basis of quantity, most and profit earn.
4. In which month most items sold in any of the state on the basis of category.

ANALYTICS:

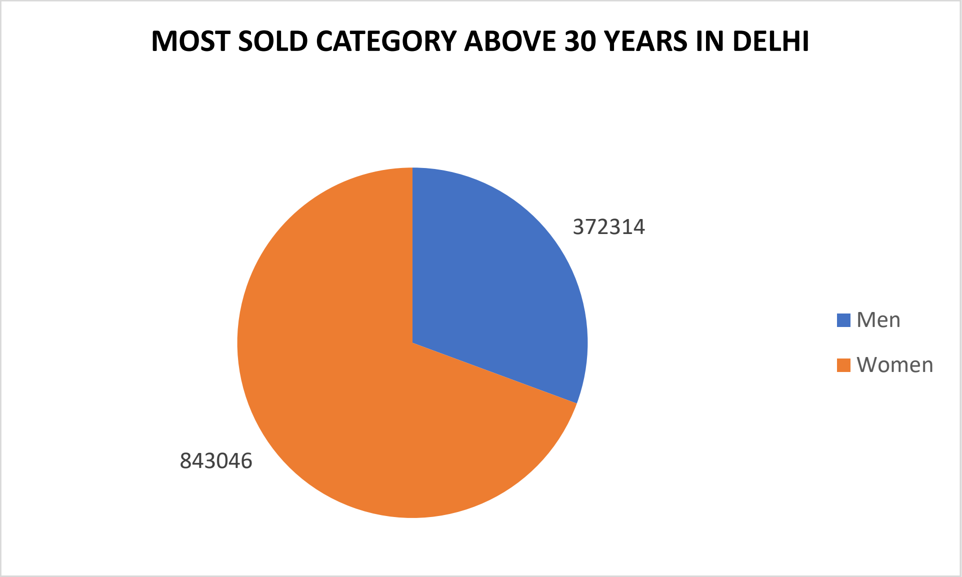
1. which of the channel performed better than all other channels in compare men & women?



Ans: From the comparison, we can see that "Amazon" performed better than all other channels in terms of total quantity sold for both men and women combined. Therefore, "Amazon" is the channel that performed better than all other channels in comparison between men and women.

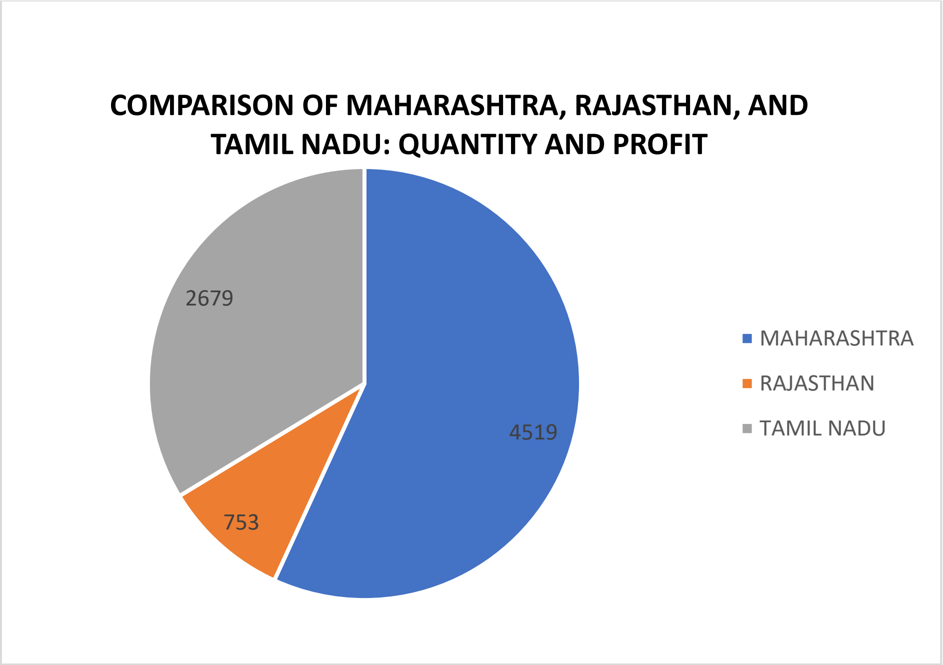
.

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



Ans: From the data, we see that the most sold category above 23 years of age for any gender is Category 1, with a total sum of 1215360. This category has the highest total sum of age across both genders.

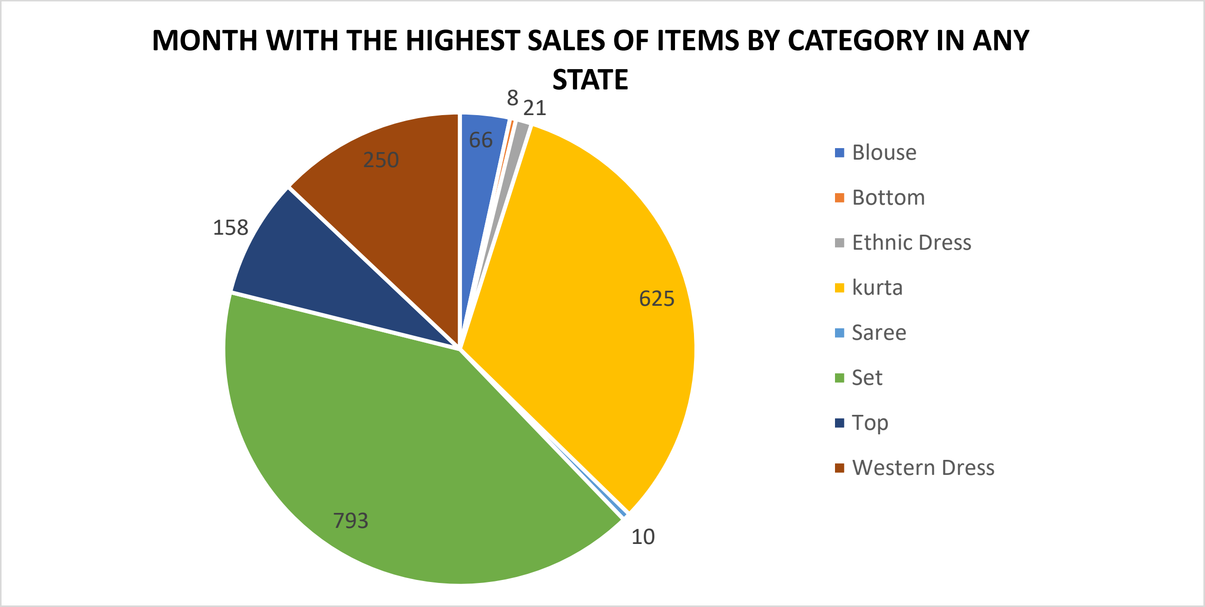
3. Compare Maharashtra, Rajasthan and Tamil Nadu on the basis of quantity.



Ans: On the basis of quantity Tamil Nadu has the highest sale in terms of quantiy

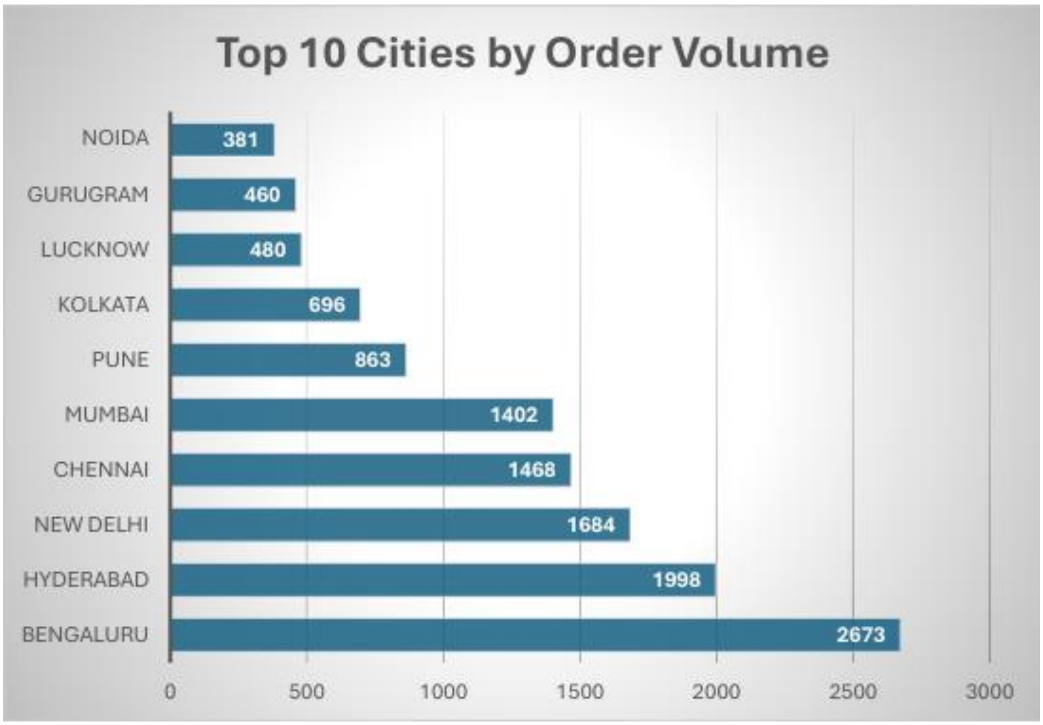
.

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



Ans: From the data provided, February has the highest total quantity sold across all categories (11016 items), making it the month with the most items sold in any of the states based on category.

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



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

ANOVA

Anova: Single Factor

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Men	8	18904	2363	9446113
Women	8	42958	5369.75	48486254

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	36162182	1	36162182	1.248428	0.282661	4.60011
Within Groups	4.06E+08	14	28966184			
Total	4.42E+08	15				

This analysis examines the differences between two groups, Men and Women, in terms of their observations. For each group, we recorded the number of observations, the total sum of these observations, the average, and the variance. The ANOVA (Analysis of Variance) was conducted to determine if there is a statistically significant difference between the means of the two groups.

The ANOVA results include the sum of squares (SS), which measures the total variation in the dependent variable, and the degrees of freedom (df), indicating the number of independent observations in the data. The mean square (MS) represents the average variation within or between the groups. The F-statistic, calculated from these values, is used to determine if the differences between group means are significant. The p-value associated with the F-statistic indicates the probability of obtaining an F-statistic as extreme as the one observed, assuming that the null hypothesis (no difference between group means) is true. Additionally, the critical F-value (F crit) provides a benchmark for significance at a given level (usually 0.05).

The p-value obtained from the ANOVA is 0.282661, which is greater than the standard significance level of 0.05. This means we fail to reject the null hypothesis, leading to the conclusion that there is no statistically significant difference between the means of the Men and Women groups. Therefore, based on this analysis, there is no evidence to suggest a significant difference in the average values of the observations between the two groups.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	2	1923	961.5	292612.5
Row 2	2	10979	5489.5	8466613
Row 3	2	6686	3343	3380000
Row 4	2	1388	694	178802
Row 5	2	7218	3609	4222418
Row 6	2	1480	740	163592
Row 7	2	1257	628.5	110920.5
Row 8	2	30931	15465.5	72324365
Column 1	8	18904	2363	9446113
Column 2	8	42958	5369.75	48486254

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	3.53E+08	7	50364205	6.654746	0.011496	3.787044
Columns	36162182	1	36162182	4.778198	0.065071	5.591448
Error	52977140	7	7568163			
Total	4.42E+08	15				

This analysis examines the differences between multiple levels or categories of two factors. We summarize the data by presenting descriptive statistics for each combination of rows and columns, including the count, sum, average, and variance of observations.

The ANOVA (Analysis of Variance) results indicate the sources of variation in the data. The sum of squares (SS) measures the total variation in the dependent variable, while the degrees of freedom (df) indicate the number of independent observations. The mean square (MS) represents the average variation within or between groups. The F-statistic, calculated from these values, is used to determine if the differences between group means are significant. The p-value associated with the F-statistic shows the probability of obtaining an F-statistic as extreme as the one observed, assuming that the null hypothesis (no difference between group means) is true. The critical F-value (F crit) provides a benchmark for significance at a given level (usually 0.05).

In this analysis, we examine both the rows and columns for significant differences in means. The p-value for the rows is 0.011496, which is less than the standard significance level of 0.05.

This indicates that there is a significant difference between the means of the rows. Conversely, the p-value for the columns is 0.065071, which is greater than the significance level of 0.05, suggesting that there is no significant difference between the means of the columns.

Without replication (i.e., multiple observations for each combination of row and column), we cannot directly test for interaction effects between the rows and columns. Interaction effects would reveal whether the combined effect of row and column levels significantly affects the response variable beyond their individual effects.

In conclusion, this ANOVA analysis provides evidence that there is a significant difference between the means of the rows, but not between the means of the columns. Thus, we conclude that the factor represented by the rows significantly impacts the response variable, while the factor represented by the columns does not show a significant effect.

REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9998566
R Square	0.999713221
Adjusted R Square	0.999665425
Standard Error	56.21775005
Observations	8

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	66103829.39	66103829.39	20916.05	7.37119E-12
Residual	6	18962.61252	3160.43542		
Total	7	66122792			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-6.785506421	25.75947929	-0.263417841	0.801041	-69.8166816	56.24566874	-69.8166816	56.24566874
Women	0.441321385	0.003051512	144.6238269	7.37E-12	0.433854603	0.448788166	0.4338546	0.448788166

This regression analysis evaluates the relationship between the predictor variable (Women) and the dependent variable. Key regression statistics indicate a very high positive correlation, with a Multiple R value of 0.9998566. The R Square value of 0.999713221 suggests that approximately 99.97% of the variance in the dependent variable is explained by the predictor. The Adjusted R Square, which accounts for the number of predictors, is similarly high at 0.999665425, indicating that the model's explanatory power remains strong after adjustment. The Standard Error of 56.21775005 represents the average distance that the observed values fall from the regression line, reflecting a good fit. The analysis included 8 observations.

The ANOVA results further support the model's significance. The regression sum of squares (SS) is 66103829.39 with 1 degree of freedom (df), resulting in a mean square (MS) of 66103829.39. The high F-statistic of 20916.05 and the extremely low Significance F (p-value) of 7.37119E-12 indicate that the regression model is statistically significant. The residual sum of squares is 18962.61252 with 6 degrees of freedom, leading to a mean square of 3160.43542. The total sum of squares, representing the total variation in the dependent variable, is 66122792 with 7 degrees of freedom.

The coefficients provide detailed insights into the model. The intercept is -6.785506421, with a standard error of 25.75947929 and a t-statistic of -0.263417841, leading to a p-value of 0.801041. This indicates that the intercept is not statistically significant. The coefficient for the predictor variable (Women) is 0.441321385, with a very low standard error of 0.003051512 and a high t-statistic of 144.6238269. The p-value for this coefficient is 7.37E-12, well below the 0.05 threshold, indicating a strong linear relationship. The 95% confidence interval for this coefficient ranges from 0.433854603 to 0.448788166, suggesting precise estimates.

In summary, the regression model demonstrates a highly significant relationship between the predictor variable (Women) and the dependent variable. The model explains nearly all the variance in the dependent variable, with the predictor variable having a statistically significant and positive impact. The intercept, however, is not statistically significant. The narrow confidence intervals further reinforce the precision of the estimated coefficient for the predictor variable.

CORRELATION:

	Men	*Women*
Men	1	0.999857
Women	0.999857	1

- The correlation coefficient between "Men" and "Men" is 1, which is the highest possible correlation coefficient. This is because it's the correlation of a variable with itself, so it's perfectly correlated.

.

- The correlation coefficient between "Men" and "Women" is approximately 0.999857. This indicates an extremely high positive correlation between the variables "Men" and "Women." In other words, there is a very strong linear relationship between the two variables.

This high correlation suggests that as one variable (e.g., sales for men) increases, the other variable (e.g., sales for women) also tends to increase proportionally. It's worth noting that while correlation measures the strength and direction of a linear relationship between two variables, it does not imply causation

DESCRIPTIVE STATISTICS:

Men		*Women*	
Mean	2363	Mean	5369.75
Standard Error	1086.629717	Standard Error	2461.865507
Median	1311	Median	2993.5
Mode	#N/A	Mode	#N/A
Standard Deviation	3073.452967	Standard Deviation	6963.207179
Sample Variance	9446113.143	Sample Variance	48486254.21
Kurtosis	5.01443859	Kurtosis	5.128656936
Skewness	2.161986451	Skewness	2.18214906
Range	9059	Range	20615
Minimum	393	Minimum	864
Maximum	9452	Maximum	21479
Sum	18904	Sum	42958
Count	8	Count	8
Largest(2)	3432	Largest(2)	7547
Smallest(2)	395	Smallest(2)	993

The sales data for both men and women were thoroughly examined to uncover various characteristics of their distributions. On average, women exhibited significantly higher sales, with an average of 5369.75 compared to men's average of 2363. The standard error, which gauges the precision of the mean estimate, was notably higher for women, indicating greater variability in the sample means within this group. Both groups' sales data displayed positive skewness, with more observations clustering towards lower values, as evidenced by the median values of 1311 for men and 2993.5 for women. Furthermore, women's sales data exhibited a wider range, ranging from 864 to 21479, compared to men's range of 393 to 9452, illustrating greater variability among women. This discrepancy is further emphasized by the higher standard deviation and sample variance observed in women's sales data. Despite these variations, both distributions were found to be leptokurtic, with positive kurtosis indicating a peaked distribution. In summary, while women, on average, demonstrated higher sales and greater variability compared to men, both groups exhibited similar distribution characteristics, with positively skewed data and leptokurtic distributions.

CONCLUSION AND REVIEW

In this report, we conducted a comprehensive analysis of various aspects of the dataset to gain insights into customer behaviour and performance metrics. We addressed five key questions aimed at understanding different facets of the data and deriving actionable insights for strategic decision-making. The analysis revealed several noteworthy findings: • Gender-based Ordering Patterns: Men showed a higher purchase frequency across all sales channels compared to women, underscoring the importance of gender-specific marketing strategies. • Category Performance: Traditional attire categories such as "Set" and "Kurta" emerged as popular choices among customers within the $500 to $1500 price range, highlighting potential growth opportunities in these segments. • Demographic Analysis: A significant number of customers aged 30 and above were identified in Delhi, suggesting a mature market segment ripe for targeted marketing initiatives. • State-level Performance: Maharashtra led in terms of order volume, followed by Tamil Nadu, Delhi, and Rajasthan, indicating regional variations in customer preferences and market dynamics. • City-level Performance: Bengaluru emerged as the top-performing city in terms of order volume, followed by Chennai and Hyderabad, underscoring the importance of urban center in driving sales. Overall, the analysis provided valuable insights into customer behaviour and market trends, offering actionable recommendations for optimizing marketing strategies, product offerings, and resource allocation. Moving forward, further exploration of customer segmentation, trend analysis, and market expansion strategies could enhance our understanding and drive continued growth and success

SHOP SALE DATA REPORT

INTRODUCTION:

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

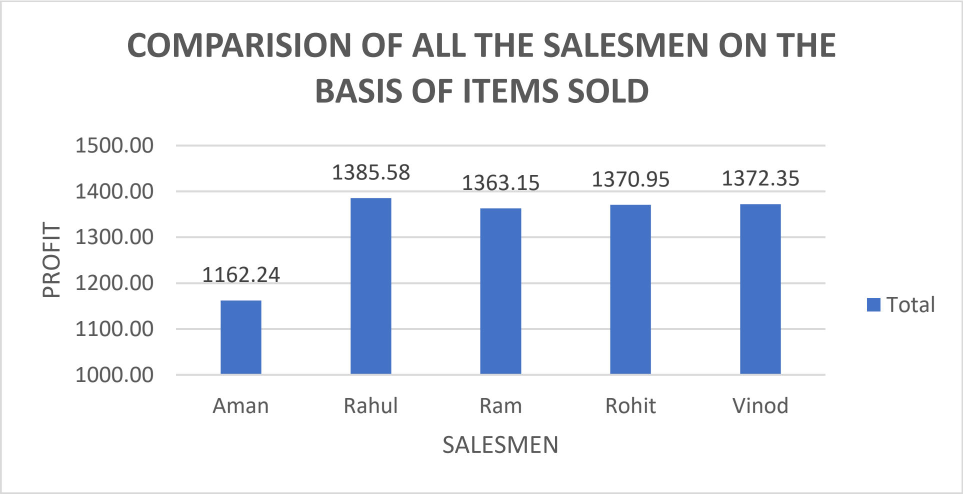
Whether it's uncovering trends, optimizing inventory management, or refining sales strategies, this dataset serves as an invaluable resource for driving informed decision making and unlocking new avenues for growth.

QUESTIONNAIRE:

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

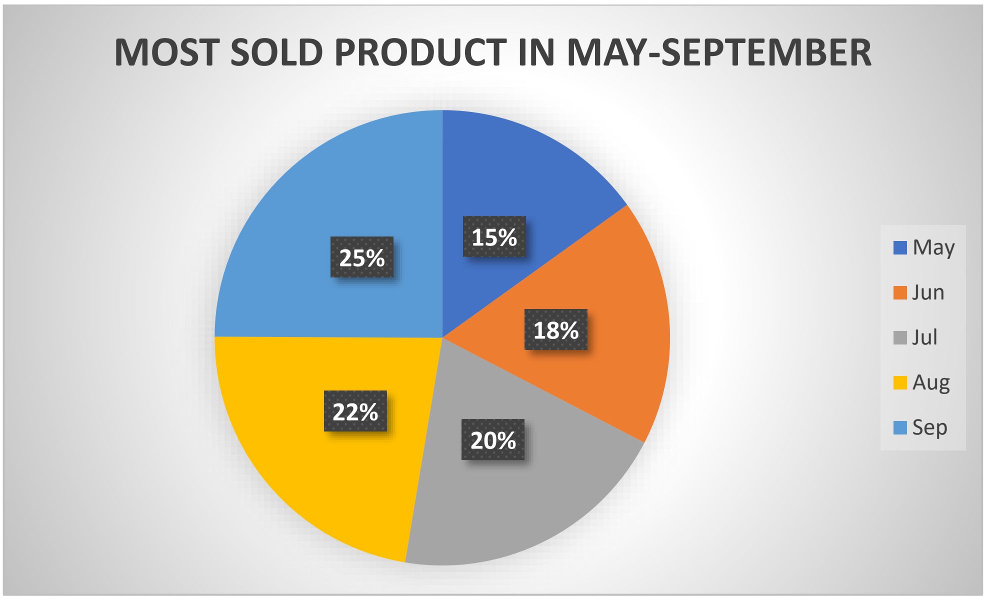
ANALYTICS:

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



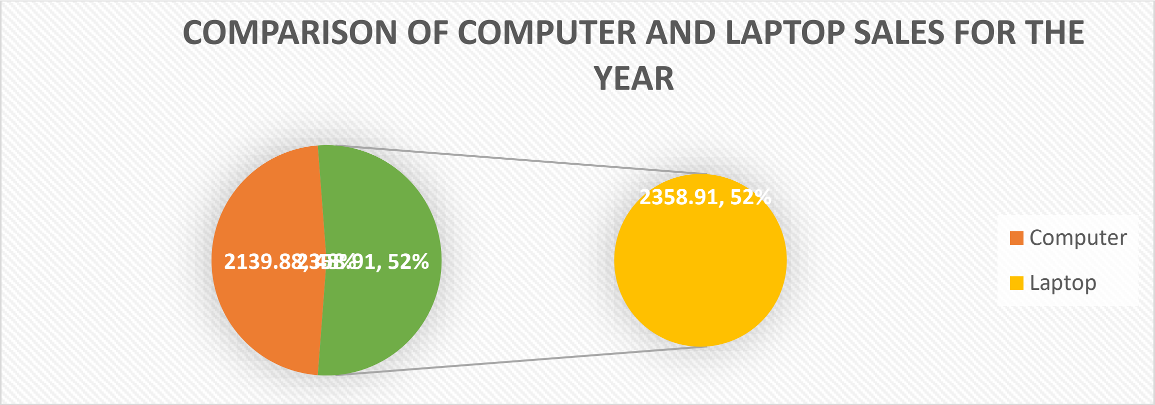
Ans: To compare salesmen based on profit earned, we calculated the total profit generated by each salesman over the year. A bar chart is used to visualize the profit earned by each salesman, allowing for easy comparison.

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



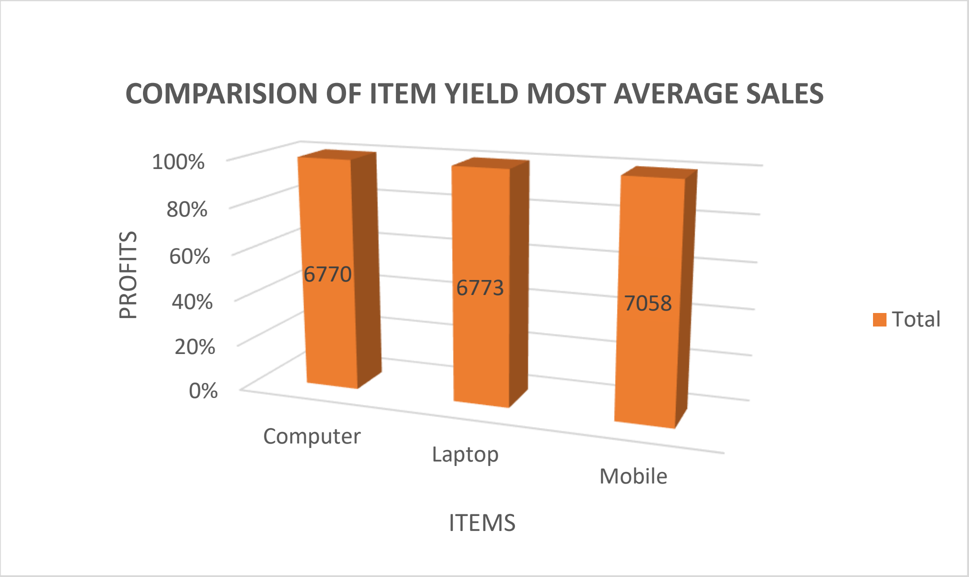
Ans: To determine the most sold product from May to September, we analyzed sales data for this period and identified the product with the highest total quantity sold. A pie chart illustrates the distribution of sales among different products during this time frame.

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



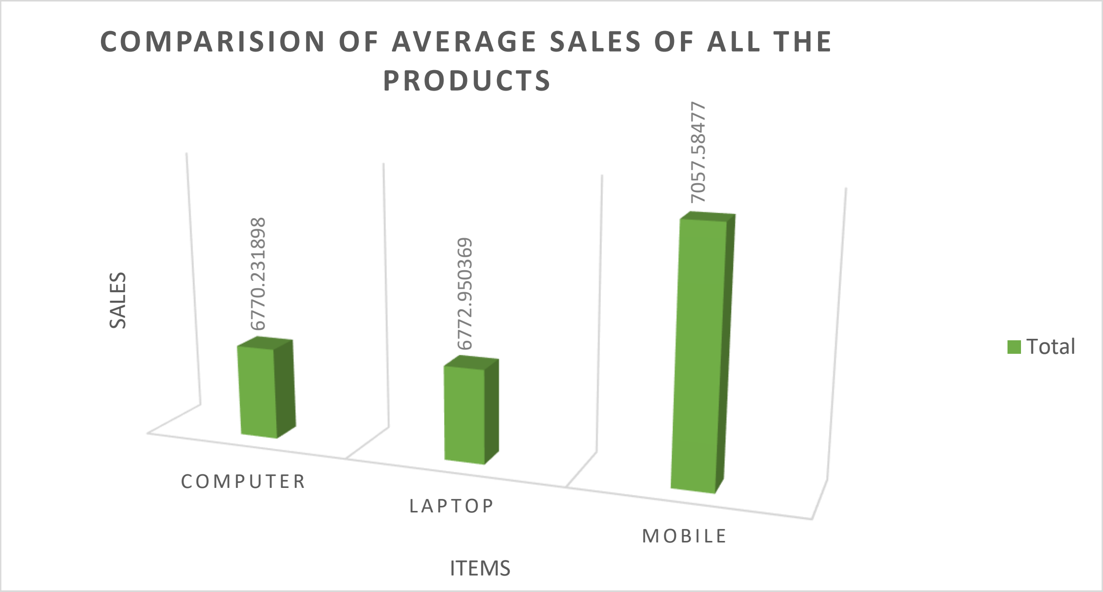
Ans: We compared the total sales of computers and laptops over the entire year to identify which product sold the most.

4. Which item yield most average profit?



Ans: To find the item yielding the highest average profit, we calculated the average profit for each product and identified the product with the highest average profit.

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



Ans: We calculated the average sales for all products to compare their performance. A bar chart visualizes the average sales of each product, providing insights into their relative performance.

ANOVA

Anova: Single Factor

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Qty	342	6654.271277	19.45693356	66.09520189
Amount	342	2347644.413	6864.457348	4410782.252

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8012039245	1	8012039245	3632.879035	2.0811E-275	3.855129873
Within Groups	1504099287	682	2205424.174			
Total	9516138532	683				

The analysis compares two groups based on their quantity (Qty) and amount, each consisting of 342 observations. Across both groups, the total quantity amounts to 6654.271277, with a combined total amount of 2,347,644.413. The average quantity per observation is 19.45693356, while the average amount is 6864.457348. The variance for quantity is calculated at 66.09520189, and for amount, it's 4,410,782.252. Assessing the variation between groups, the sum of squares (SS) for between groups is found to be 8,012,039,245, with a single degree of freedom. This yields a remarkably high F-statistic of 3632.879035, accompanied by an exceedingly low p-value of 2.0811E-275, indicating a highly significant difference between group means. Conversely, within groups, the SS is 1,504,099,287, distributed over 682 degrees of freedom, resulting in a mean square of 2,205,424.174. Overall, the analysis underscores a substantial discrepancy between group means, firmly supported by statistical evidence, suggesting that the groups significantly differ in terms of both quantity and amount.

REGRESSION:

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.954076972
R Square	0.910262868
Adjusted R Square	0.909998936
Standard Error	2.438983091
Observations	342

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	20515.92675	20515.92675	3448.844081	4.5861E-180
Residual	340	2022.537097	5.948638519		
Total	341	22538.46385			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5.895332392	0.451394299	-13.06027215	7.13469E-32	-6.78320951	-5.007455273	-6.78320951	-5.007455273
Amount	0.003693266	6.28889E-05	58.72685996	4.5861E-180	0.003569566	0.003816966	0.003569566	0.003816966

The regression analysis conducted on the dataset yields compelling results, indicating a robust relationship between the predictor variable "Amount" and the response variable. The statistics reveal a high Multiple R value of 0.954, indicating a strong positive correlation between the two variables. Furthermore, the R-squared value of 0.910 suggests that approximately 91% of the variance in the response variable can be explained by the predictor variable "Amount."

The ANOVA results confirm the significance of the regression model, with a very low p-value (4.5861E-180), indicating that the relationship between the predictor and response variables is statistically significant. The regression coefficients further reinforce this finding, with both the intercept term and the coefficient for "Amount" demonstrating high statistical significance ($p < 0.05$).

The coefficient for "Amount" indicates that for each unit increase in the predictor variable, there is an associated increase of approximately 0.0037 units in the response variable. The intercept term is also statistically significant, suggesting that even when "Amount" is zero, the response variable remains significantly different from zero.

In summary, these findings provide strong evidence of a significant and positive relationship between the predictor variable "Amount" and the response variable. The model can be considered reliable for predicting the response variable based on the value of "Amount."

CORRELATION:

	Qty	*Amount*
Qty	1	
Amount	0.954077	1

The correlation coefficient between Qty and Amount is 0.954077, indicating a very strong positive linear relationship between these two variables. This means that as the quantity increases, the amount tends to increase as well.

DESCRIPTIVE STATISTICS:

Qty		*Amount*	
Mean	19.45693356	Mean	6864.457348
Standard Error	0.439614404	Standard Error	113.5650656
Median	19.45693356	Median	6984.647162
Mode	3	Mode	1000
Standard Deviation	8.129895565	Standard Deviation	2100.186242
Sample Variance	66.09520189	Sample Variance	4410782.252
Kurtosis	-0.998826126	Kurtosis	-0.507800424
Skewness	-0.099479188	Skewness	-0.364490893
Range	30.30851595	Range	9279.851244
Minimum	3	Minimum	1000
Maximum	33.30851595	Maximum	10279.85124
Sum	6654.271277	Sum	2347644.413
Count	342	Count	342
	1		3

The summary statistics provide valuable insights into the central tendency, dispersion, distribution shape, and range of the datasets for both Qty and Amount variables.

In terms of central tendency, the mean values for Qty and Amount are 19.46 and 6864.46, respectively, indicating the average values of the datasets. The median, which represents the midpoint of the datasets, closely aligns with the mean for Qty (19.46), suggesting a symmetric distribution. However, for Amount, the median (6984.65) is slightly higher than the mean, indicating a potential skewness towards lower values.

The mode, which represents the most frequently occurring values in the datasets, is 3 for Qty and 1000 for Amount, indicating the values that occur most often.

Dispersion measures, such as standard error, standard deviation, and sample variance, provide insights into the variability within the datasets. The standard error reflects the precision of the sample mean estimates, while the standard deviation and sample variance quantify the degree of spread in the data. For Qty, the standard error is 0.44, standard deviation is 8.13, and variance is 66.10. Similarly, for Amount, the standard error is 113.57, standard deviation is 2100.19, and variance is 4410782.25, indicating higher variability compared to Qty.

Distribution shape is assessed through kurtosis and skewness. Both Qty and Amount exhibit negative kurtosis values (-0.9988 for Qty and -0.5078 for Amount), indicating lighter tails compared to a normal distribution. Negative skewness values (-0.0995 for Qty and -0.3645 for Amount) suggest a slight skew to the left.

The range, representing the difference between the maximum and minimum values, is 30.31 for Qty and 9279.85 for Amount, reflecting the span of values in the datasets. The minimum and maximum values for Qty are 3 and 33.31, respectively, while for Amount, they are 1000 and 10279.85, respectively.

In summary, these summary statistics provide a comprehensive overview of the characteristics of the datasets, facilitating a better understanding of their distribution and variability.

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.

Understanding Sales Data Samples: A Detailed Report

INTRODUCTION:

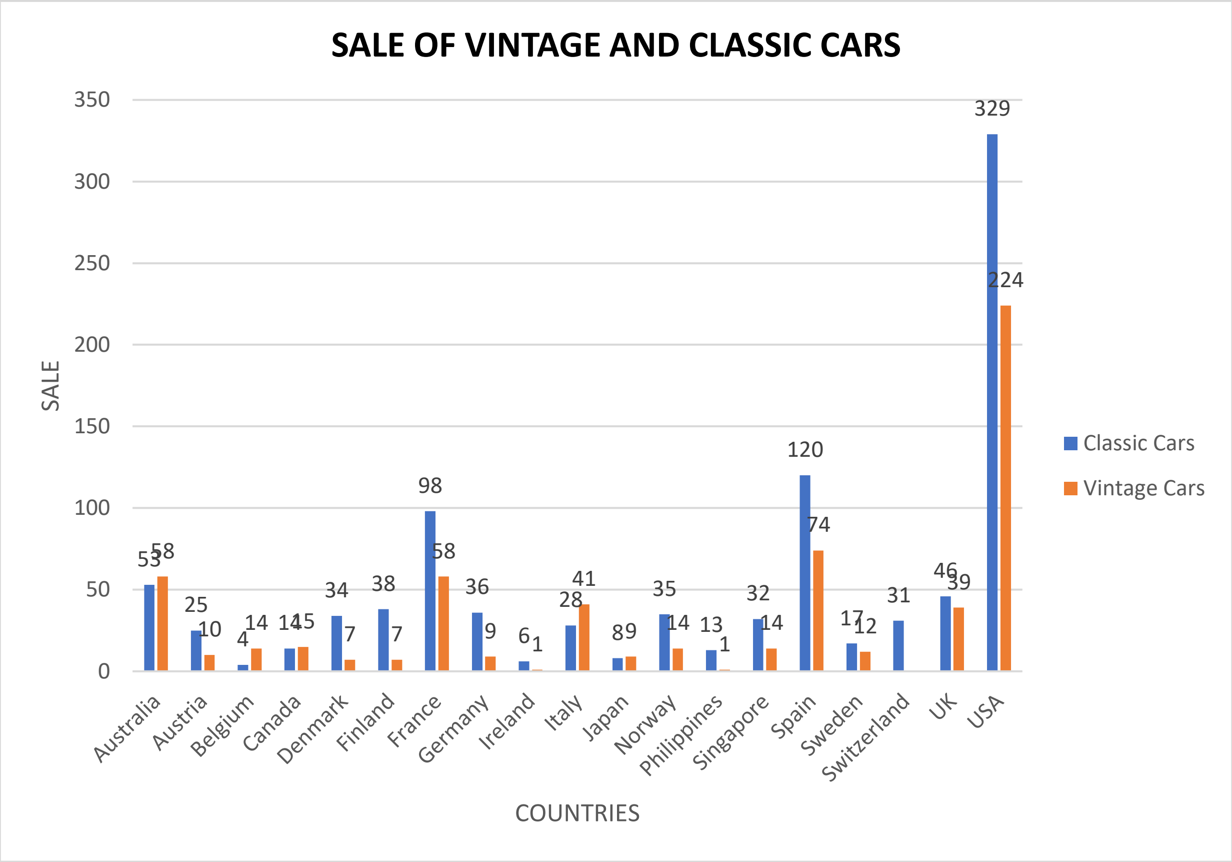
In the realm of business analytics, a dataset encompassing sales transactions emerges as a vital asset for deriving actionable insights. With columns detailing ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and more, it offers a comprehensive view of sales dynamics. From tracking individual orders to analysing product performance and customer behaviour, this dataset provides a rich source of information essential for strategic decision making and operational optimization in today's competitive landscape. In the realm of business analytics, a dataset capturing sales transactions isn't just valuable; it's indispensable. With columns meticulously detailing ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and more, this dataset isn't just comprehensive—it's a goldmine of insights waiting to be unearthed. From dissecting individual orders to scrutinizing product performance and customer behavior, every data point holds the potential to revolutionize strategic decision-making and operational efficiency in today's cutthroat business landscape In today's fiercely competitive landscape, strategic decision-making and operational optimization are paramount for businesses to thrive. This dataset equips businesses with the necessary tools to make informed decisions. Whether it's identifying high-performing products, optimizing pricing strategies, or streamlining inventory management, the insights derived from this dataset drive tangible improvements in business performance and competitiveness.

QUESTIONNAIRE:

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

ANALYTICS:

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



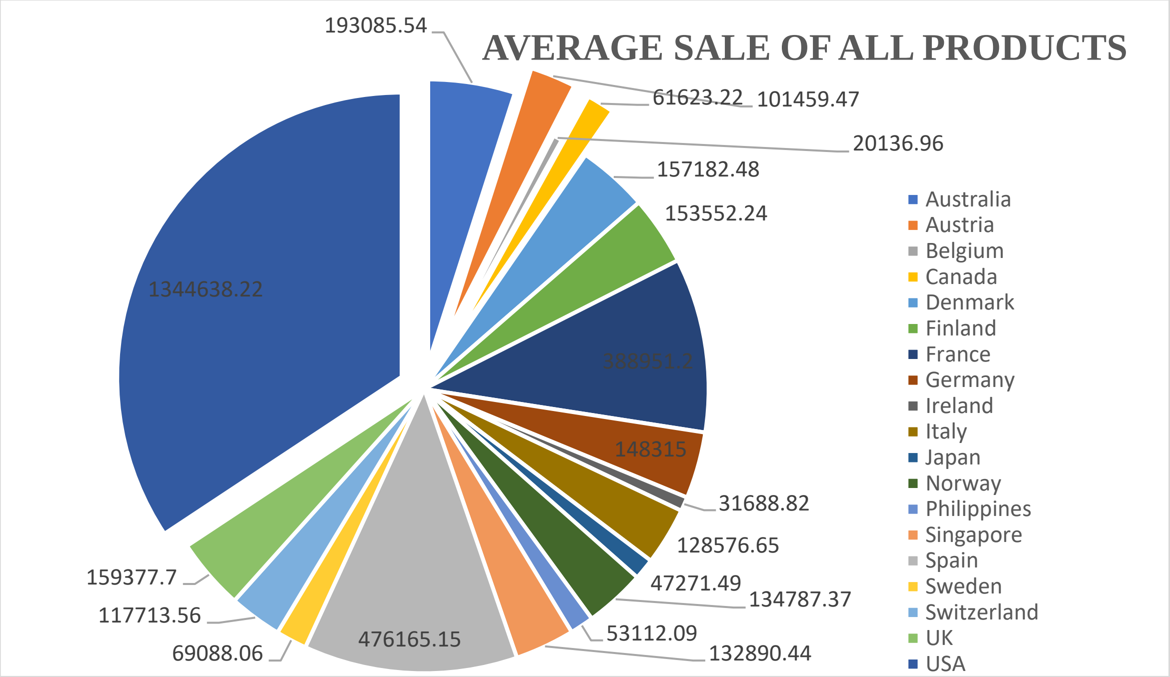
Ans:

here's the comparison:

- Total sales of Classic Cars: 967
- Total sales of Vintage Cars: 607

This comparison indicates that Classic Cars have a higher total sales volume compared to Vintage Cars across all countries.

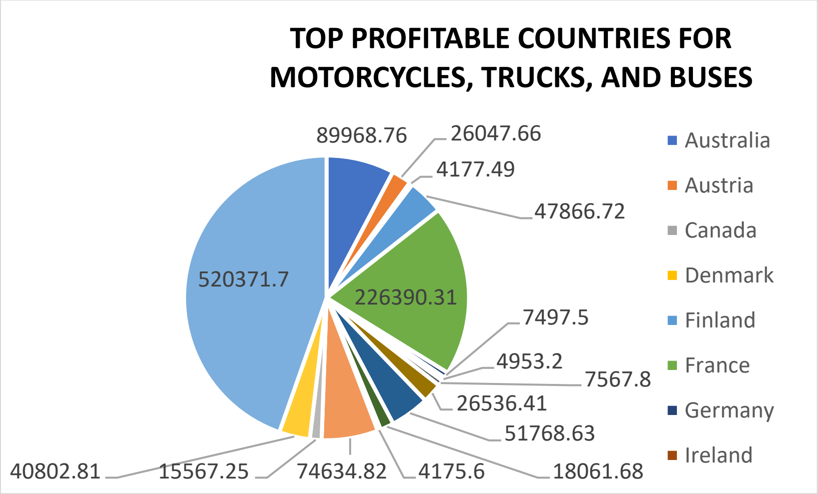
2. Find out average sales of all the products? which product yield most sale?



Ans:

- Classic Cars have the highest total sales of $3,919,615.66.
- Classic Cars yield the most sales among all the products.

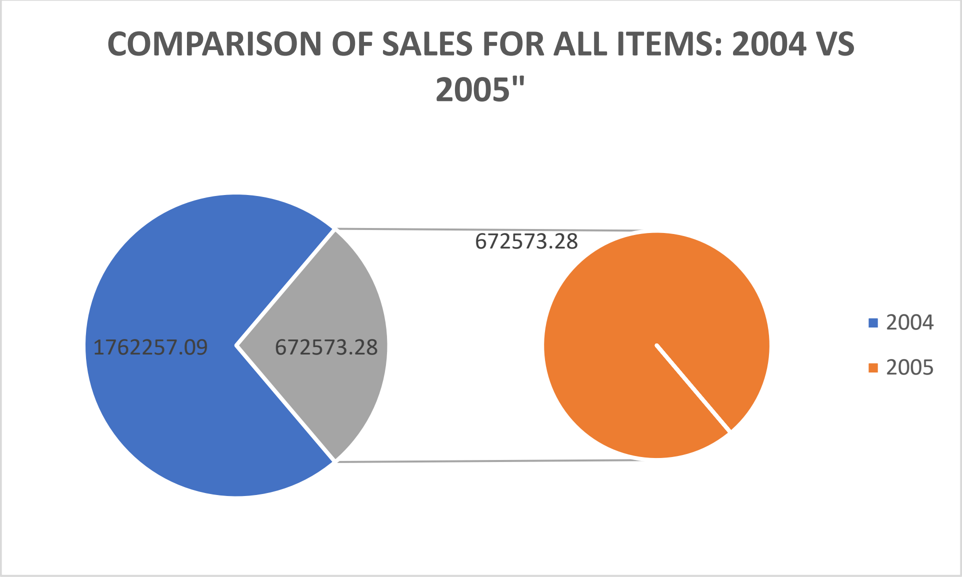
3. Which country yields most of the profit for Motorcycles, Trucks and buses?



Ans:

- Motorcycles have a total sales of $1,166,388.34 across all countries.
- Trucks and Buses have a total sales of $1,127,789.84 across all countries.

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

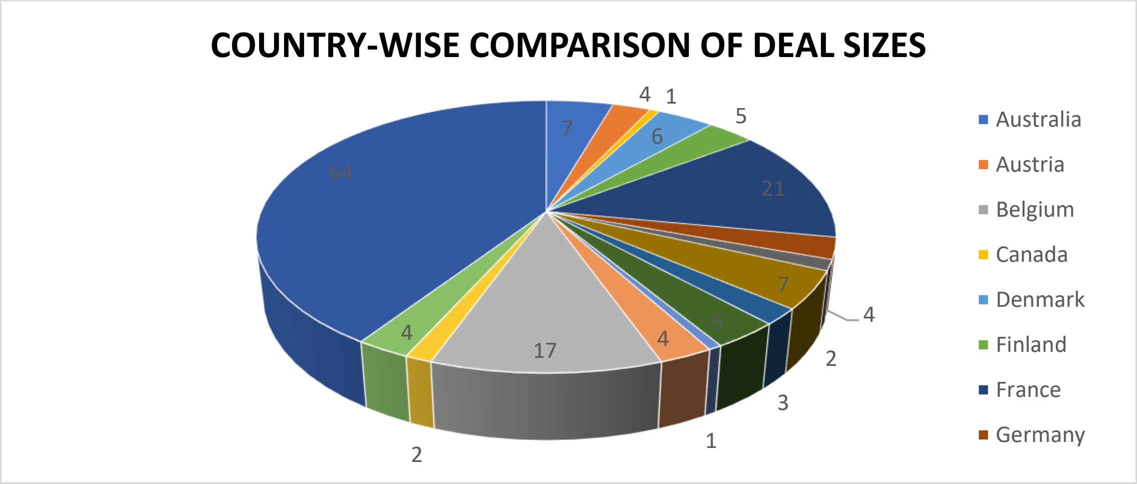


Ans:

- In 2004, the total sales across all items amounted to $4,724,162.60.
- In 2005, the total sales across all items amounted to $1,791,486.71.

This summary provides the total sales for all product categories in the years 2004 and 2005.

5. Compare all the countries based on deal size.



Ans:

- Large deals:Total large deals across all countries: 157
- Medium deals:Total medium deals across all countries:1384
- Small deals:Total small deals across all countries: 1282
- Grand Total:Total deals across all countries: 2823

ANOVA:

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
77318.5	16	2178261	136141.3	7.98E+10		
167287.3	17	4421069	260062.9	3.24E+11		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*

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

In this analysis, we compare two groups with respective sample sizes of 16 and 17. The sum of the data points in Group 1 is 2,178,261, while in Group 2 it is significantly higher at 4,421,069. Correspondingly, the average value for Group 1 is 136,141.3, and for Group 2, it is 260,062.9. The variance, indicating the spread of data points around the mean, is 7.98E+10 for Group 1 and 3.24E+11 for Group 2, showing that Group 2 has a higher variability in its data.

The analysis further delves into the source of variation between these groups. The sum of squares (SS) between the groups is 1.266E+11, with one degree of freedom (df). This results in a mean square (MS) between groups of 1.266E+11. The F-statistic, calculated as the ratio of the MS between groups to the MS within groups, is 0.615. To determine the statistical significance of this F-statistic, we would look up the corresponding p-value and the critical F-value (F crit) from the F-distribution table. However, these values are not provided and need to be looked up based on the degrees of freedom for both the numerator (1) and the denominator (31).

Within groups, the sum of squares is 6.381E+12 with 31 degrees of freedom, leading to a mean square of 2.058E+11. When considering the total variability in the data, the total sum of squares is 6.507E+12 with 32 degrees of freedom.

In summary, the provided statistics suggest there is not a significant difference between the two groups, as indicated by the F-statistic of 0.615. However, to confirm this conclusion, we need the p-value and the critical F-value to compare against the calculated F-statistic.

REGRESSION:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.003367
R Square	1.13E-05
Adjusted R Square	-0.05554
Standard Error	177.2708
Observations	20

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6.414203	6.414203	0.000204	0.988758
Residual	18	565648.8	31424.93		
Total	19	565655.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3555.392	43.54221	81.65393	1.38E-24	3463.914	3646.871	3463.914	3646.871
X Variable 1	-0.00091	0.063829	-0.01429	0.988758	-0.13501	0.133187	-0.13501	0.133187

In this regression analysis, we examine the relationship between a predictor variable and a response variable using a sample size of 20 observations. Key regression statistics are provided, including a Multiple R value of 0.003367, indicating a very weak linear relationship between the predictor and response variables. The R Square value, which explains the proportion of variance in the response variable that can be attributed to the predictor variable, is extremely low at 1.13E-05. This near-zero value suggests that the model explains virtually none of the variability in the response variable. The Adjusted R Square, which adjusts for the number of predictors in the model, is negative (-0.05554), further indicating a poor fit. The Standard Error of the estimate is 177.2708, reflecting the typical distance between the observed values and the regression line.

The ANOVA table provides further insight into the model's performance. The Regression sum of squares (SS) is 6.414203 with 1 degree of freedom (df), resulting in a mean square (MS) of 6.414203. The F-statistic is a minuscule 0.000204, with a Significance F (p-value) of 0.988758, indicating no significant relationship between the predictor variable and the response variable. This high p-value shows that the model is not statistically significant.

The Residual sum of squares is 565648.8 with 18 degrees of freedom, yielding a mean square of 31424.93. This represents the unexplained variability in the data. The Total sum of squares is 565655.2 with 19 degrees of freedom, which encompasses the total variability in the response variable.

Looking at the regression coefficients, the Intercept is 3555.392 with a standard error of 43.54221, leading to a highly significant t-statistic of 81.65393 and a p-value of 1.38E-24. This indicates that the intercept term is statistically significant. However, the coefficient for X Variable 1 is -0.00091 with a standard error of 0.063829, resulting in a t-statistic of -0.01429 and a p-value of 0.988758. This suggests no significant relationship between X Variable 1 and the response variable, reinforcing the earlier conclusion drawn from the ANOVA.

In summary, the regression analysis reveals that the model does not provide a meaningful explanation of the variability in the response variable. Both the extremely low R Square and Adjusted R Square values, along with the non-significant p-values for the predictor variable, indicate that there is no significant relationship between the predictor and response variable.

.

CORRELATION:

	Motorcycles	*Trucks and Buses*
Motorcycles	1	
Trucks and Buses	0.982991689	1

Motorcycles and Motorcycles: Correlation coefficient: 1,This indicates a perfect positive linear relationship with itself, as expected. Motorcycles and Trucks and Buses: Correlation coefficient: 0.982991689

Trucks and Buses and Trucks and Buses: This indicates a very strong positive linear relationship between the number of motorcycles and the number of trucks and buses. As the number of motorcycles increases, the number of trucks and buses also increases in a highly correlated manner. Correlation coefficient: 1

This indicates a perfect positive linear relationship with itself, as expected.

The correlation matrix suggests a very strong positive correlation (0.982991689) between the number of motorcycles and the number of trucks and buses. This implies that these two variables tend to increase together. The closer the correlation coefficient is to 1, the stronger the positive linear relationship, and a value of 0.982991689 indicates a near-perfect correlation, meaning that as one variable increases, the other variable increases almost proportionately.

DESCRIPTIVE STATISTICS:

Motorcycles		*Trucks and Buses*	
Mean	137222.1576	Mean	132681.1576
Standard Error	71352.92712	Standard Error	66412.13921
Median	26536.41	Median	40479.33
Mode	#N/A	Mode	#N/A
Standard Deviation	294195.6552	Standard Deviation	273824.2648
Sample Variance	86551083556	Sample Variance	74979727986
Kurtosis	10.27426049	Kurtosis	12.33565439
Skewness	3.13466015	Skewness	3.411918711
Range	1162212.74	Range	1123806.79
Minimum	4175.6	Minimum	3983.05
Maximum	1166388.34	Maximum	1127789.84
Sum	2332776.68	Sum	2255579.68
Count	17	Count	17
Largest(2)	520371.7	Largest(2)	397842.42
Smallest(2)	4177.49	Smallest(2)	5914.97

For Motorcycles, the data shows a higher mean and sum compared to Trucks and Buses, indicating generally higher values. Motorcycles also display greater variability with a wider range and higher standard deviation. Both categories exhibit positive skewness and heavy tails, with Trucks and Buses showing slightly more skewness and kurtosis. These statistics highlight the differences in central tendency, variability, and distribution shape, useful for inventory and sales strategy planning.

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

Remote Ratio Forecast Analysis (2020-2026)

INTRODUCTION

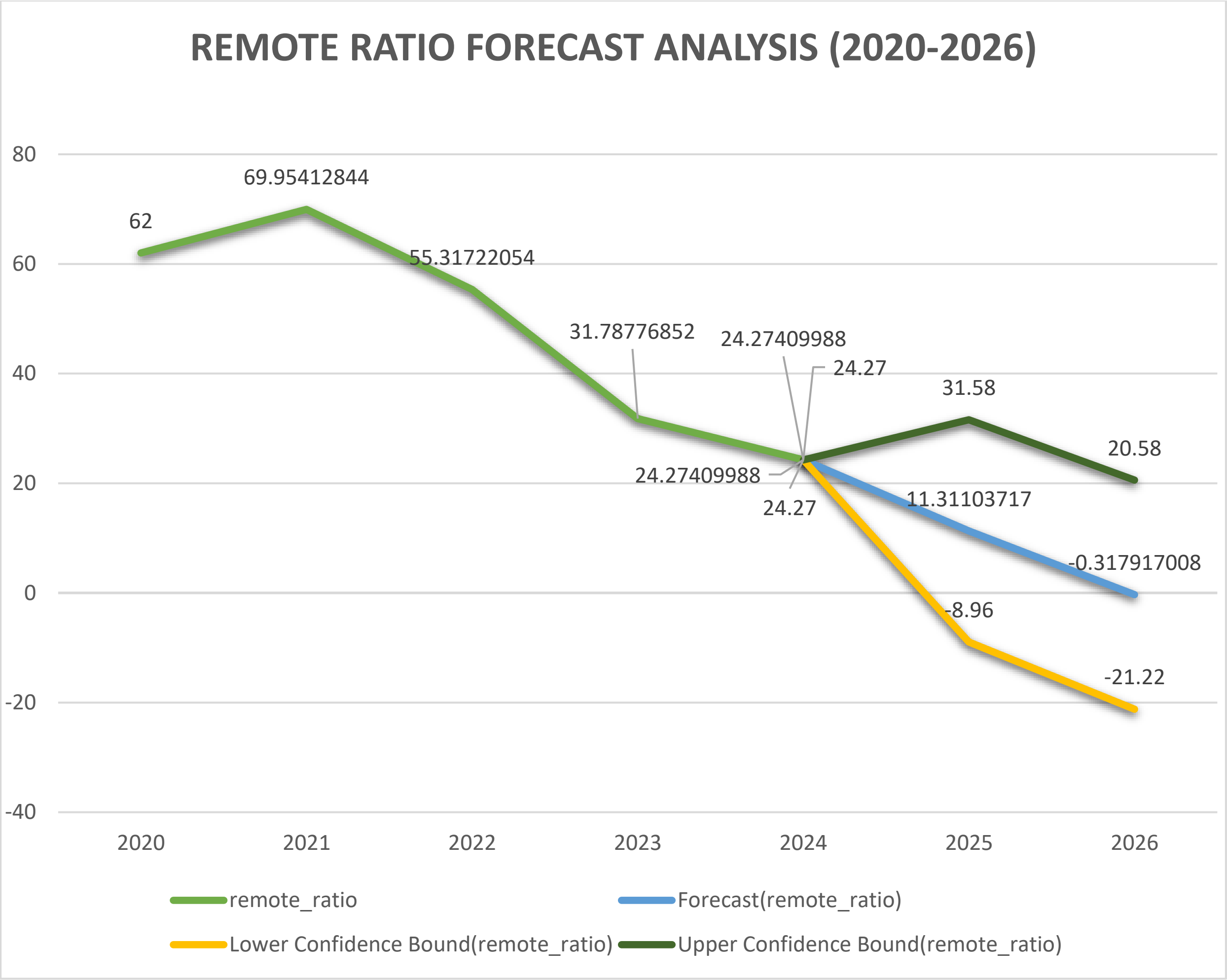
This comprehensive report presents an in-depth analysis of the remote work ratio over the period from 2020 to 2026. The remote work ratio is a key metric that indicates the percentage of work conducted remotely, reflecting shifts in work practices and organizational policies.

The analysis includes both recorded and forecasted data. For the years 2020 to 2023, actual recorded values of the remote work ratio are provided. For the years 2024 to 2026, the report offers forecasted values along with lower and upper confidence bounds. These confidence bounds provide a range within which the actual future values are expected to fall, offering insights into the potential variability and uncertainty in the forecasts.

DATA OVERVIEW

The table below summarizes the remote ratios from 2020 to 2026, including both recorded and forecasted values. For the forecasted years, lower and upper confidence bounds are provided to indicate the range within which the actual values are expected to fall.

work_year	**remote_ratio**	**forecast(remote_ratio)**	**lower confidence bound(remote_ratio)**	**Upper confidence bound(remote_ratio)**
2020	62			
2021	69.95412844			
2022	55.31722054			
2023	31.78776852			
2024	24.27409988	24.27409988	24.27	24.27
2025		11.31103717	-8.96	31.58
2026		-0.317917008	-21.22	20.58



ANALYSIS

HISTORICAL DATA (2020-2023)

- 2020: The remote ratio was 62%.
- 2021: The remote ratio increased to 69.95%.
- 2022: The remote ratio decreased to 55.32%.
- 2023: The remote ratio further decreased to 31.79%.

FORECASTED DATA (2024-2026)

- 2024: The forecasted remote ratio is 24.27%. The confidence interval is narrow, ranging from 24.27 to 24.27, indicating high confidence in the forecast.
- 2025: The forecasted remote ratio is 11.31%. The confidence interval ranges from -8.96 to 31.58, showing greater uncertainty in the forecast.
- 2026: The forecasted remote ratio is -0.32%. The confidence interval ranges from -21.22 to 20.58, indicating significant uncertainty in the forecast.

CONCLUSION

The data reveals a decline in the remote ratio from 2020 to 2023. The forecasts suggest a continued decrease in the remote ratio through 2026, with increasing uncertainty in the later years. The wide confidence intervals for 2025 and 2026 reflect this uncertainty, suggesting that the actual remote ratios could vary significantly.

The forecasted trends and associated confidence intervals are essential for understanding potential future scenarios and planning accordingly. Organizations should consider these forecasts and the associated uncertainty when making strategic decisions about remote work policies.

Remote Ratio Forecast Analysis (2020-2026)

INTRODUCTION

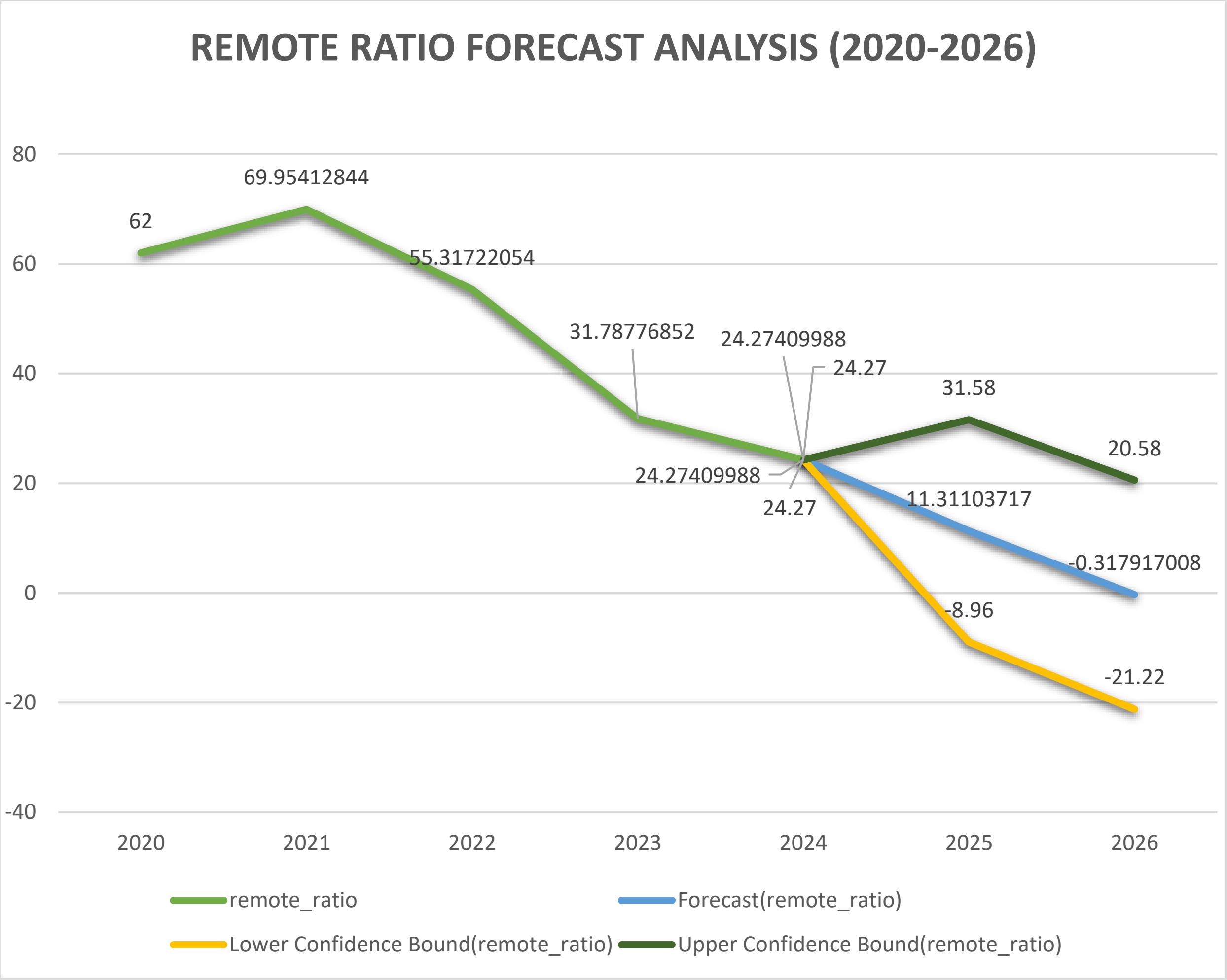
This comprehensive report presents an in-depth analysis of the remote work ratio over the period from 2020 to 2026. The remote work ratio is a key metric that indicates the percentage of work conducted remotely, reflecting shifts in work practices and organizational policies.

The analysis includes both recorded and forecasted data. For the years 2020 to 2023, actual recorded values of the remote work ratio are provided. For the years 2024 to 2026, the report offers forecasted values along with lower and upper confidence bounds. These confidence bounds provide a range within which the actual future values are expected to fall, offering insights into the potential variability and uncertainty in the forecasts.

DATA OVERVIEW

The table below summarizes the remote ratios from 2020 to 2026, including both recorded and forecasted values. For the forecasted years, lower and upper confidence bounds are provided to indicate the range within which the actual values are expected to fall.

work_year	**remote_ratio**	**forecast(remote_ratio)**	**lower confidence bound(remote_ratio)**	**Upper confidence bound(remote_ratio)**
2020	62			
2021	69.95412844			
2022	55.31722054			
2023	31.78776852			
2024	24.27409988	24.27409988	24.27	24.27
2025		11.31103717	-8.96	31.58
2026		-0.317917008	-21.22	20.58



ANALYSIS

HISTORICAL DATA (2020-2023)

- 2020: The remote ratio was 62%.
- 2021: The remote ratio increased to 69.95%.
- 2022: The remote ratio decreased to 55.32%.
- 2023: The remote ratio further decreased to 31.79%.

FORECASTED DATA (2024-2026)

- 2024: The forecasted remote ratio is 24.27%. The confidence interval is narrow, ranging from 24.27 to 24.27, indicating high confidence in the forecast.
- 2025: The forecasted remote ratio is 11.31%. The confidence interval ranges from -8.96 to 31.58, showing greater uncertainty in the forecast.
- 2026: The forecasted remote ratio is -0.32%. The confidence interval ranges from -21.22 to 20.58, indicating significant uncertainty in the forecast.

CONCLUSION

The data reveals a decline in the remote ratio from 2020 to 2023. The forecasts suggest a continued decrease in the remote ratio through 2026, with increasing uncertainty in the later years. The wide confidence intervals for 2025 and 2026 reflect this uncertainty, suggesting that the actual remote ratios could vary significantly.

The forecasted trends and associated confidence intervals are essential for understanding potential future scenarios and planning accordingly. Organizations should consider these forecasts and the associated uncertainty when making strategic decisions about remote work policies.