**ACROPOLIS INSTITUTE OF TECHNOLOGY & RESEARCH, INDORE**

**DEPARTMENT OF COMPUTER SCIENCE**

# CS-605 Data Analytics Lab 3rd Year

# 6th Semester 2023-

# 2024


**SUBMITTED BY -**

**Akshat Sethi**

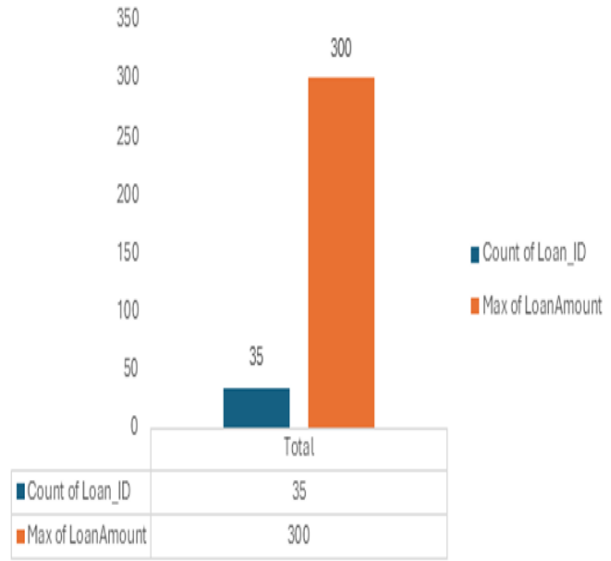
**(0827CS211016)**

**SUBMITTED TO -**

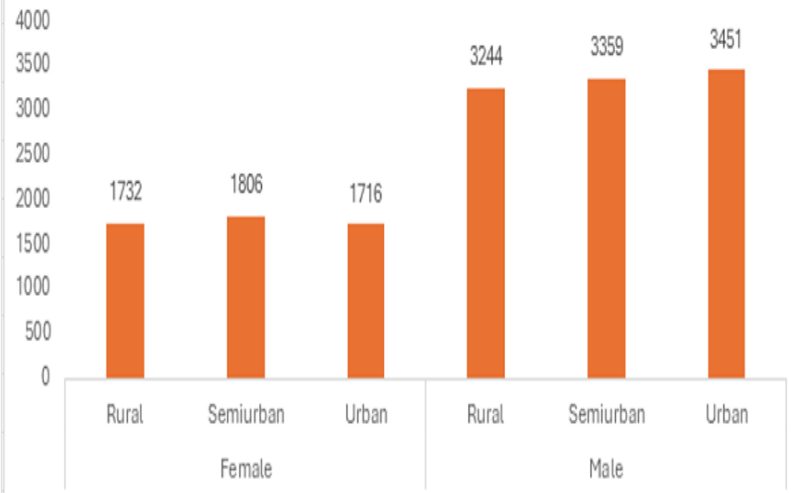
**Prof. Anurag Punde**

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

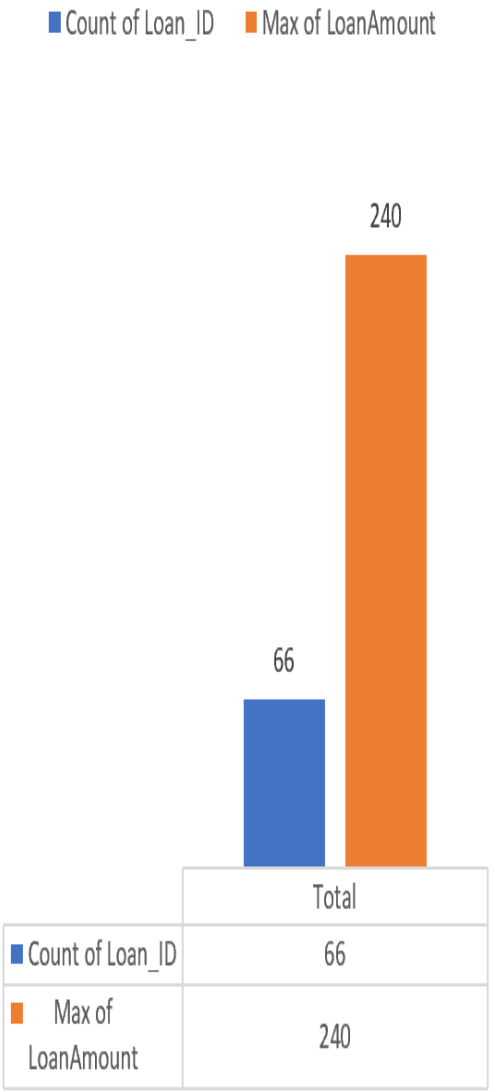
## Female graduates who are not married applied for a Loan and the highest amount



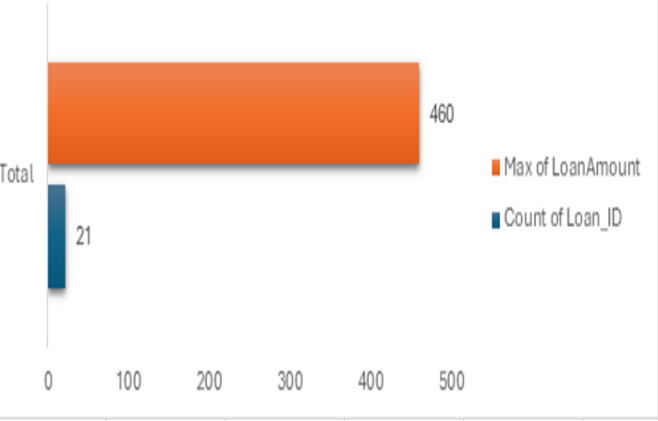

## males and females who are not married applied for a Loan, Compare Urban, Semi-urban, and Rural on the basis of the loan amount



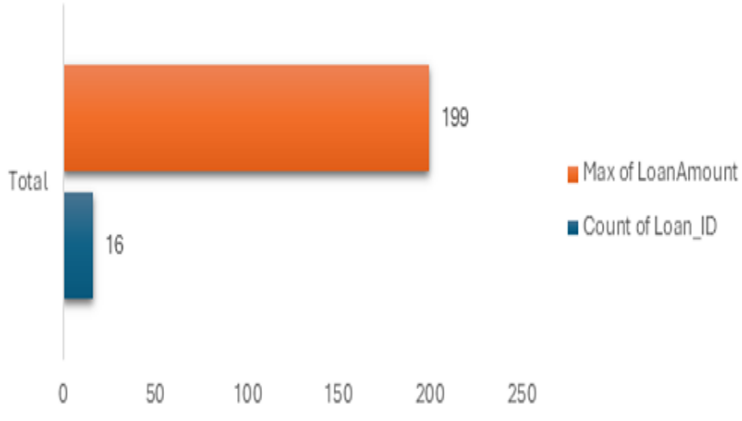

## Male graduates who are not married applied for a Loan and the highest amount

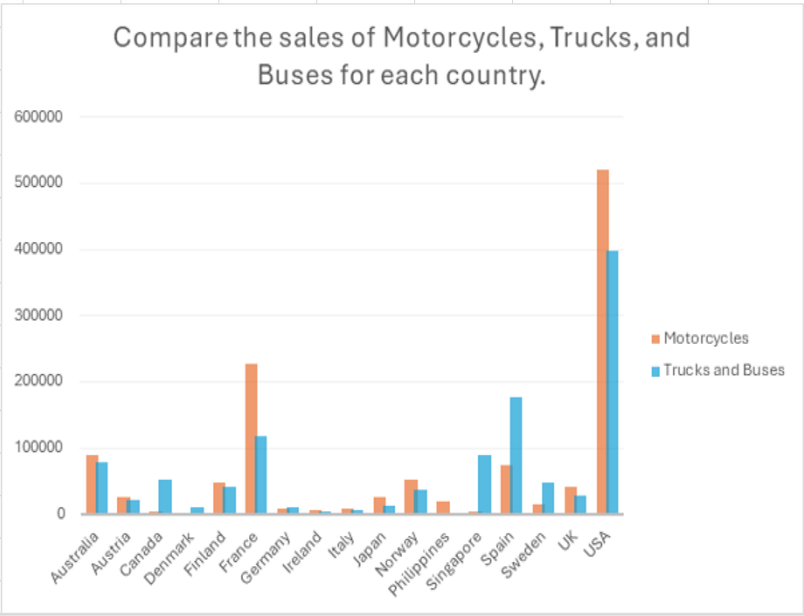


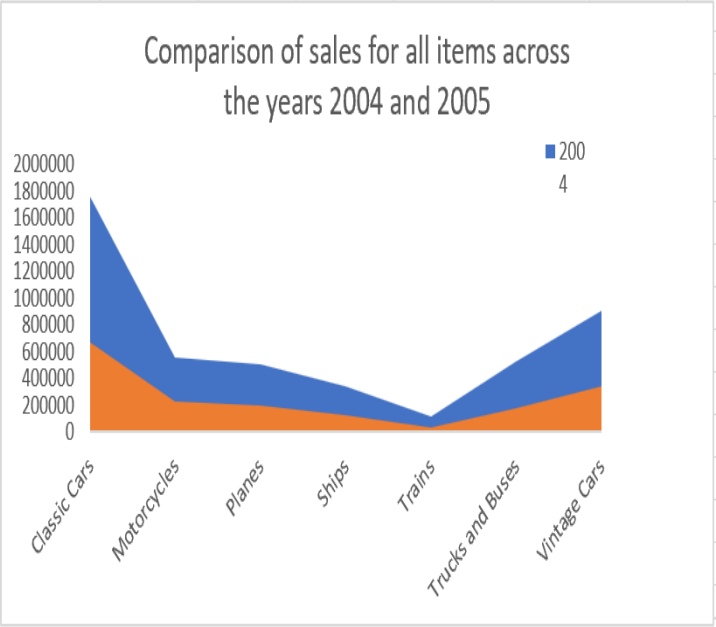

## female graduates who are married applied for a Loan and the highest amount

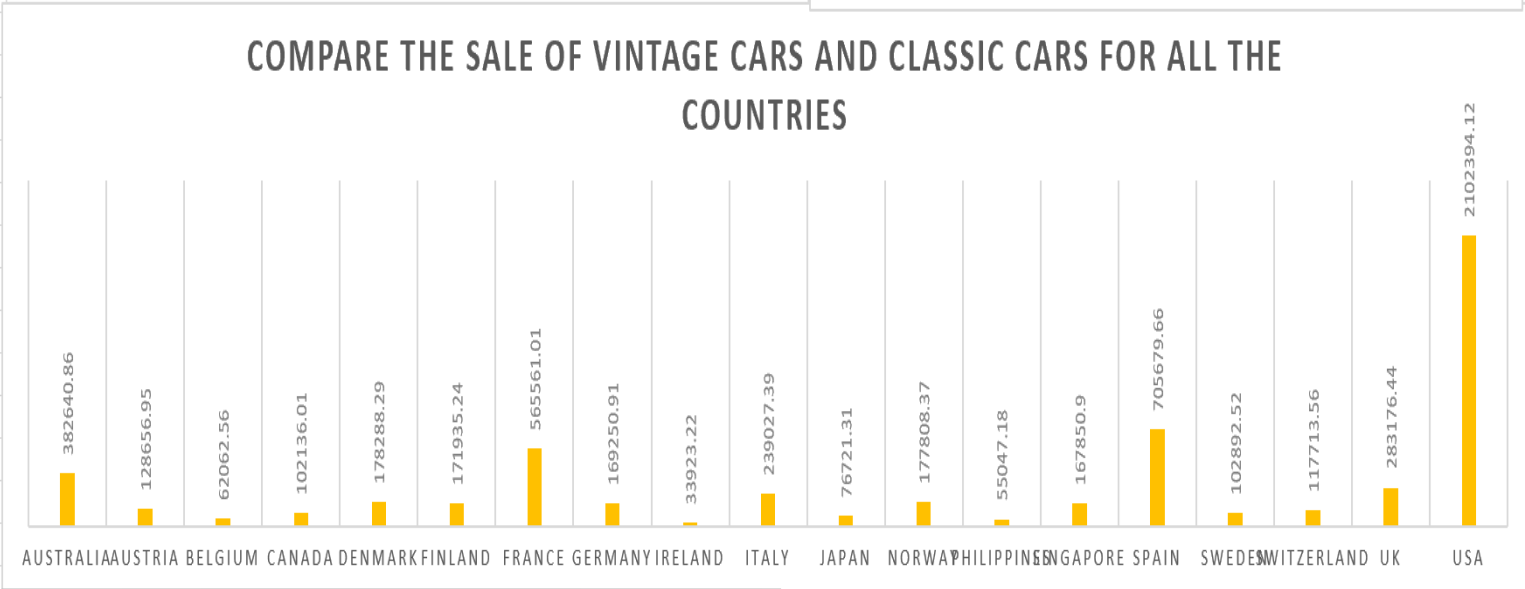


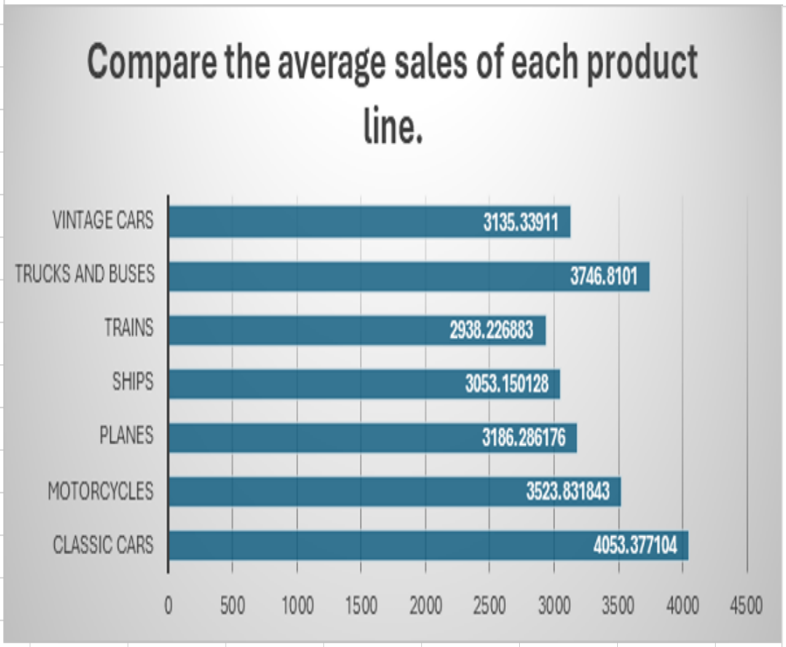

## Male non-graduates who are not married applied for a Loan and the highest amount

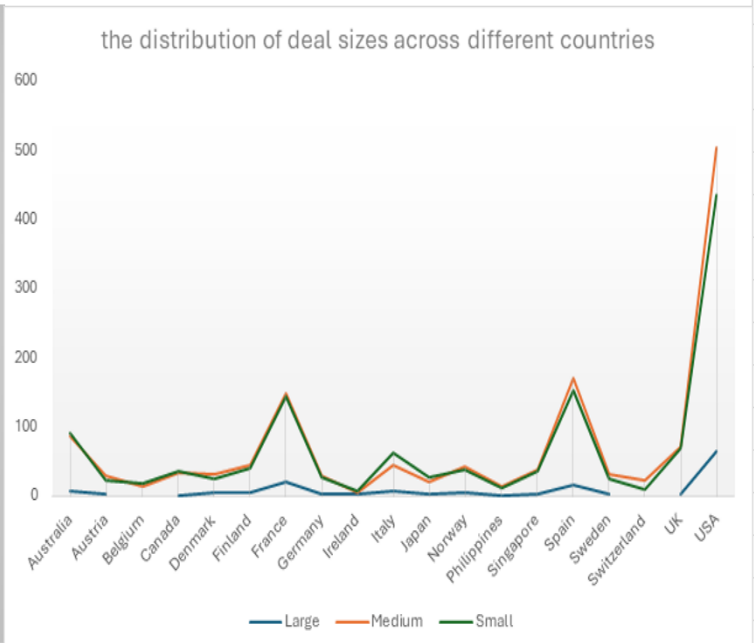



Compare the sales of Motorcycles, Trucks, and Buses for each country.
600000
500000
400000
300000
200000
100000
0
Australia
Austria
Canada
Denmark
Finland
France
Germany
Ireland
Italy
Japan
Norway
Philippines
Singapore
Spain
Sweden
UK
USA
Motorcycles
Trucks and Buses

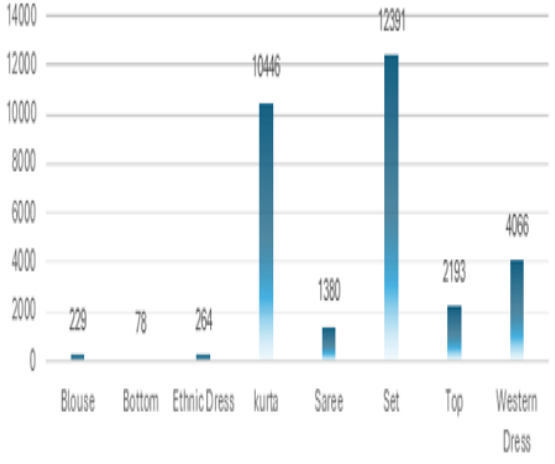

Comparison of sales for all items across the years 2004 and 2005
2000000
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
2004
Classic Cars
Motorcycles
Planes
Ships
Trains
Trucks and Buses
Vintage Cars


COMPARE THE SALE OF VINTAGE CARS AND CLASSIC CARS FOR ALL THE COUNTRIES
382640.86
128656.95
62062.56
102136.01
178288.29
171935.24
565561.01
169250.91
33923.22
239027.39
76721.31
177808.37
55047.18
167850.9
705679.66
102892.52
117713.56
283176.44
2102394.12
AUSTRALIA
AUSTRIA
BELGIUM
CANADA
DENMARK
FINLAND
FRANCE
GERMANY
IRELAND
ITALY
JAPAN
NORWAY
PHILIPPINES
SINGAPORE
SPAIN
SWEDEN
SWITZERLAND
UK
USA

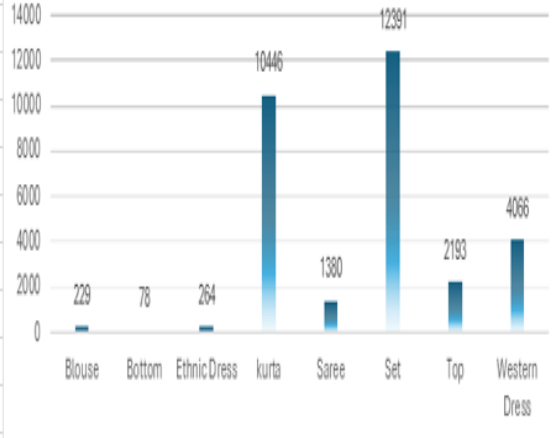

Compare the average sales of each product line.
VINTAGE CARS
3135.33911
TRUCKS AND BUSES
3746.8101
TRAINS
2938.226883
SHIPS
3053.150128
PLANES
3186.286176
MOTORCYCLES
3523.831843
CLASSIC CARS
4053.377104
0
500
1000
1500
2000
2500
3000
3500
4000
4500


the distribution of deal sizes across different countries
600
500
400
300
200
100
0
Australia
Austria
Belgium
Canada
Denmark
Finland
France
Germany
Ireland
Italy
Japan
Norway
Philippines
Singapore
Spain
Sweden
Switzerland
UK
USA
Large
Medium
Small

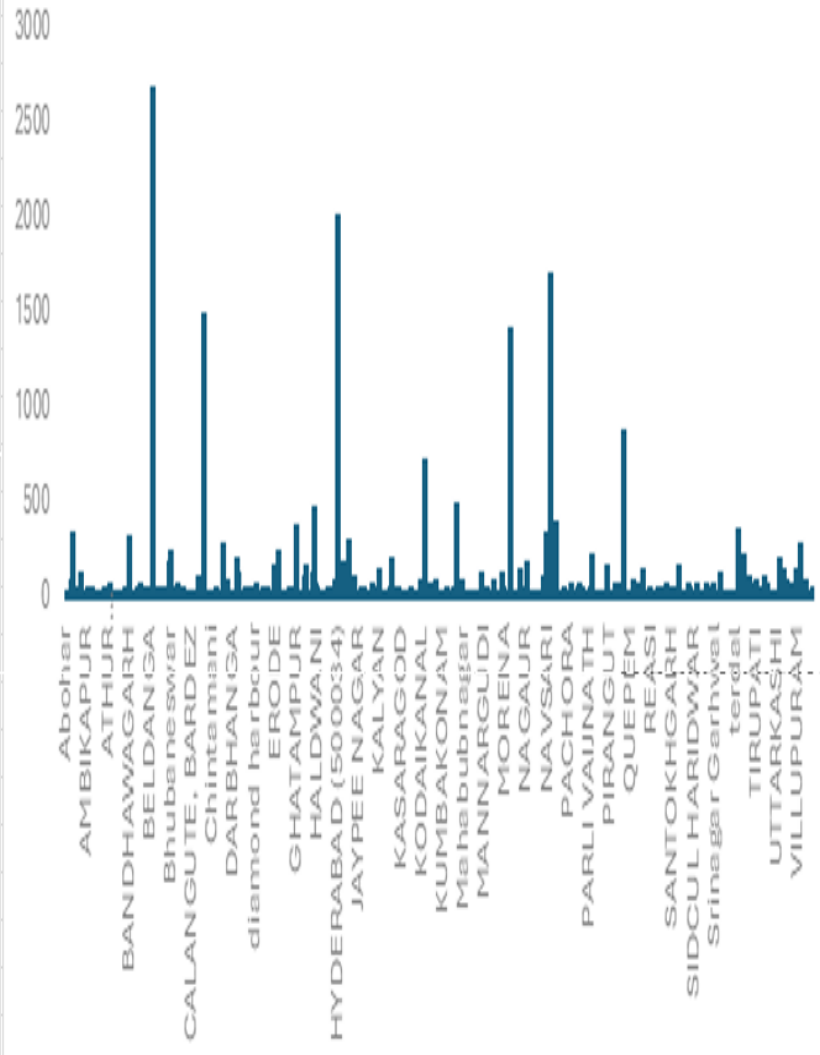
## COMPARE ALL CATEGORIES OF ORDERS WHERE THE AMOUNT IS LESS THAN 1500 AND GREATER THAN 5000



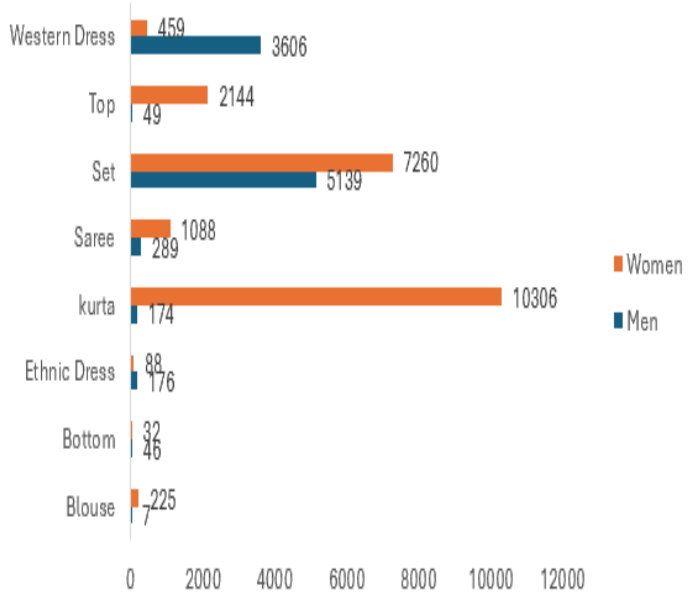

## COMPARE ALL CATEGORIES OF ORDERS WHERE THE AMOUNT IS LESS THAN 1500 AND GREATER THAN 5000



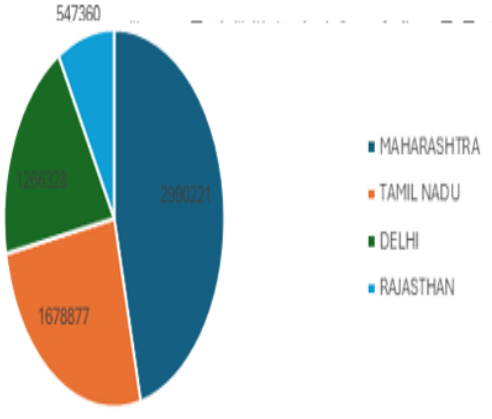

## the city that performed better than all others based on the highest order placed

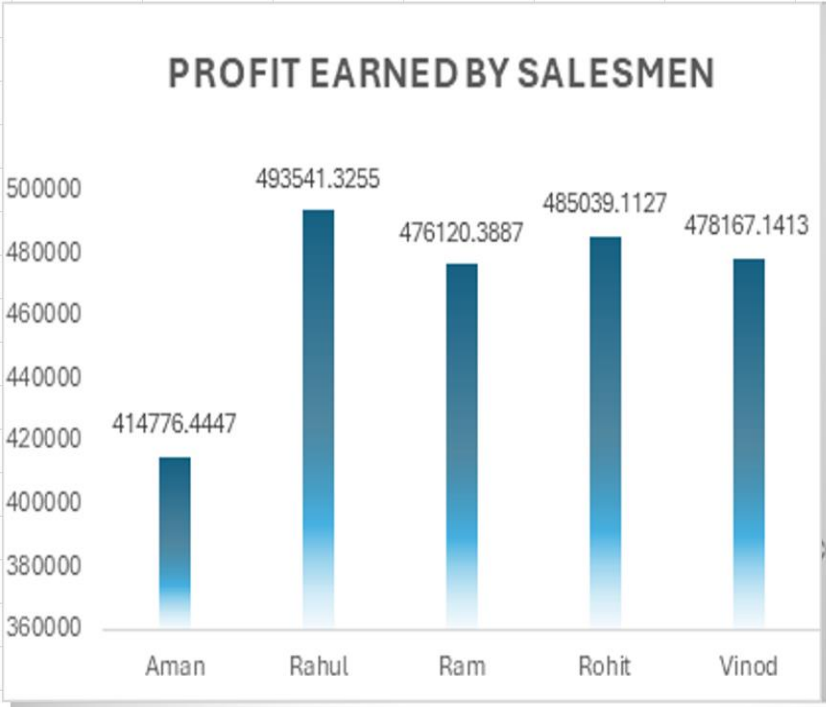


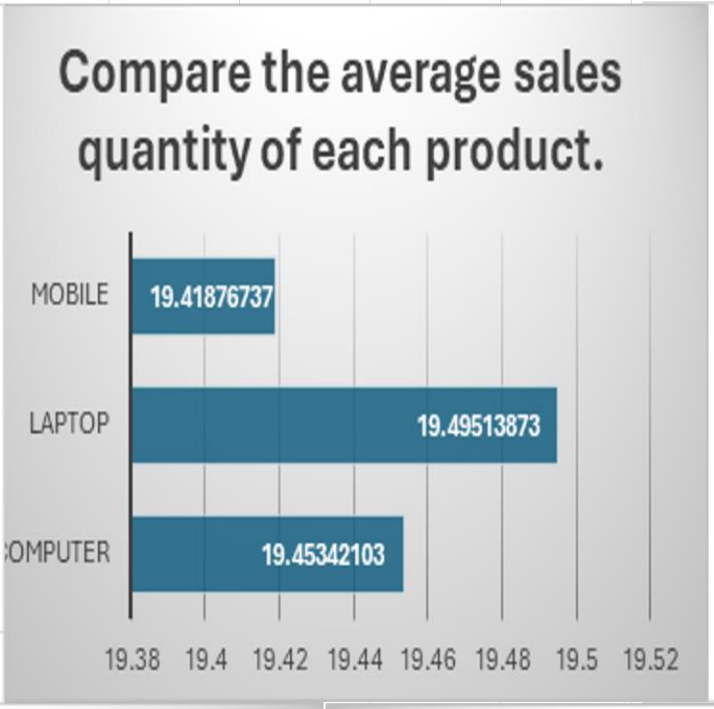

## Compare various categories of items based on the most quantity sold and show which gender buys the most category

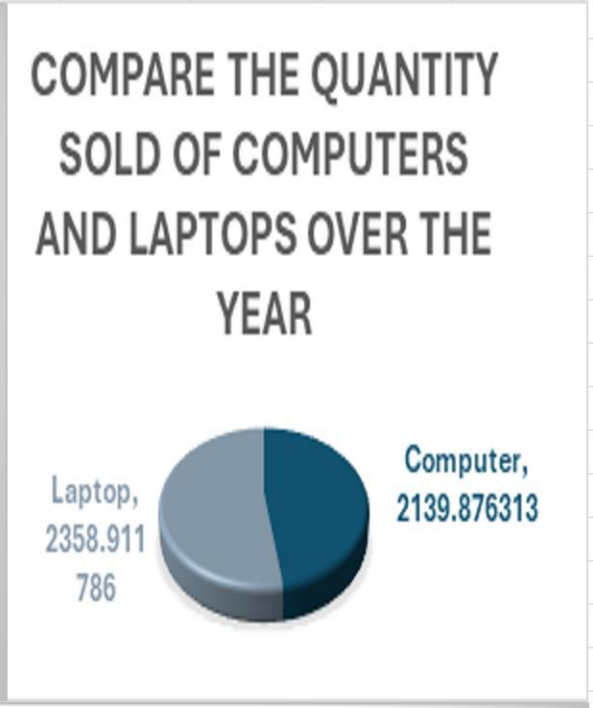



## Sum of Amount

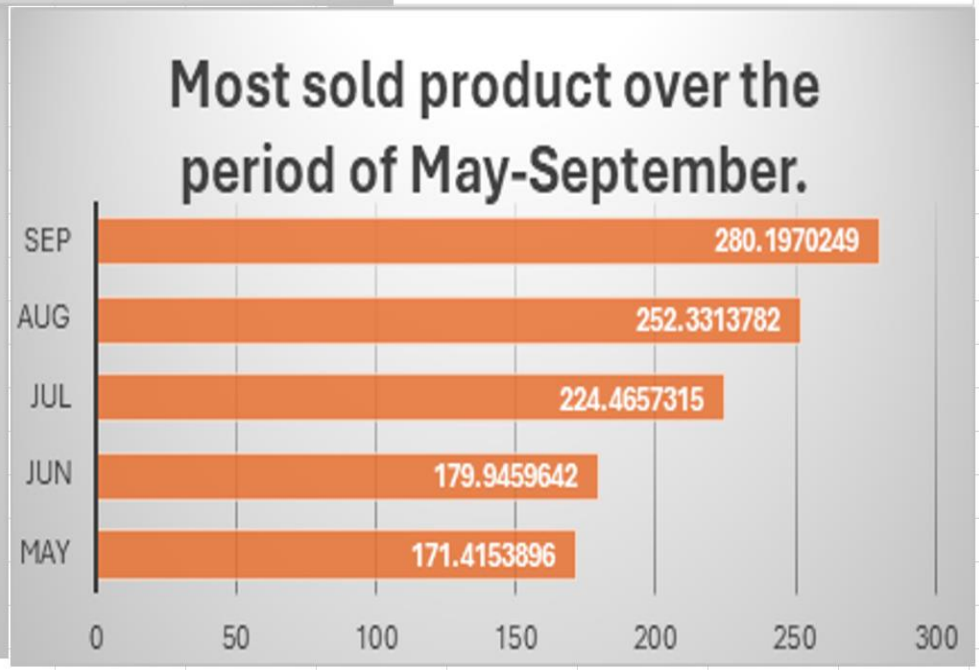


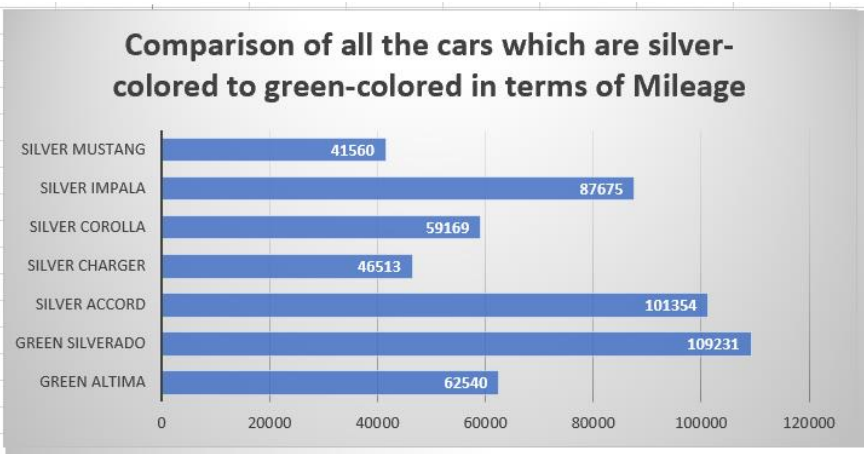

PROFIT EARNED BY SALESMEN
500000
480000
460000
440000
420000
400000
380000
360000
414776.4447
493541.3255
476120.3887
485039.1127
478167.1413
Aman
Rahul
Ram
Rohit
Vinod

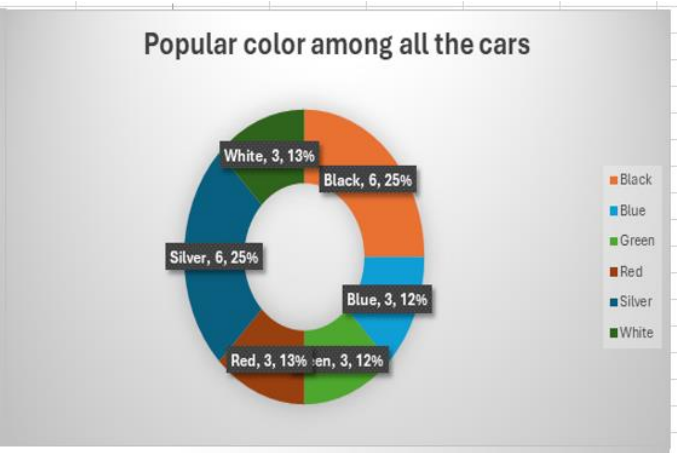

Compare the average sales quantity of each product.
MOBILE
19.41876737
LAPTOP
19.49513873
OMPUTER
19.45342103
19.38
19.4
19.42
19.44
19.46
19.48
19.5
19.52

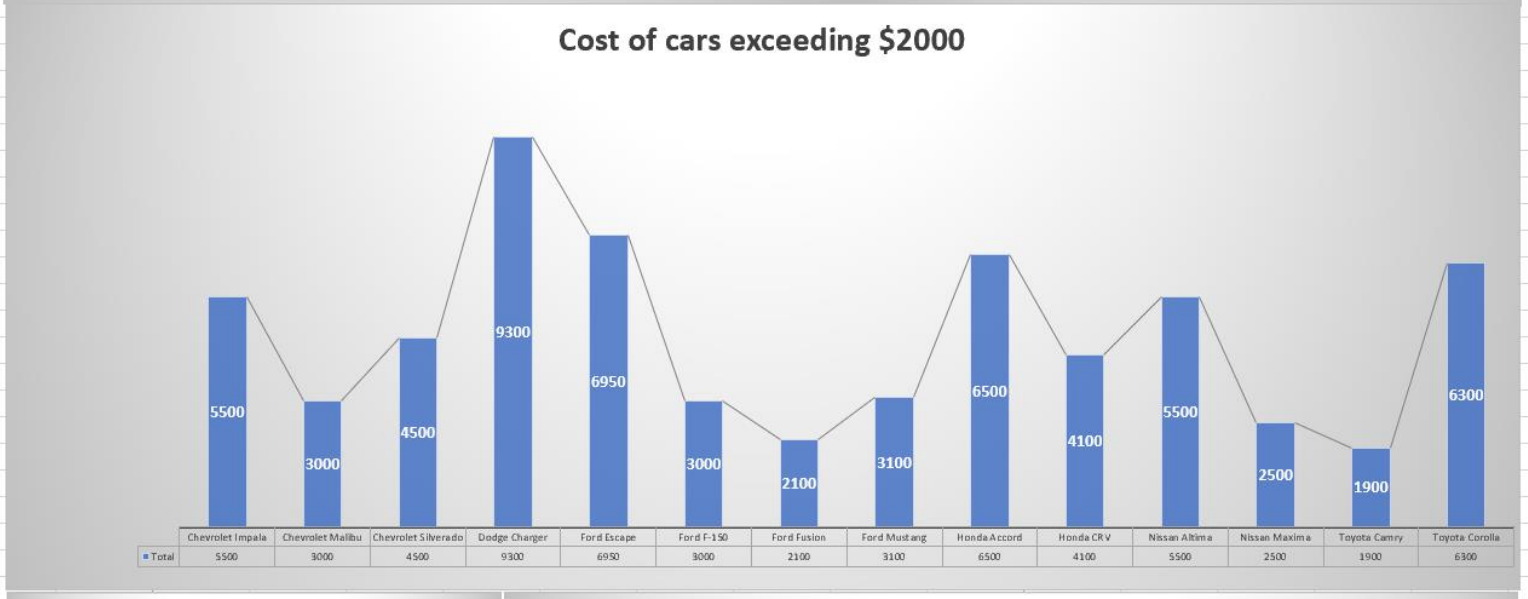

COMPARE THE QUANTITY SOLD OF COMPUTERS AND LAPTOPS OVER THE YEAR
Laptop, 2358.911786
Computer, 2139.876313

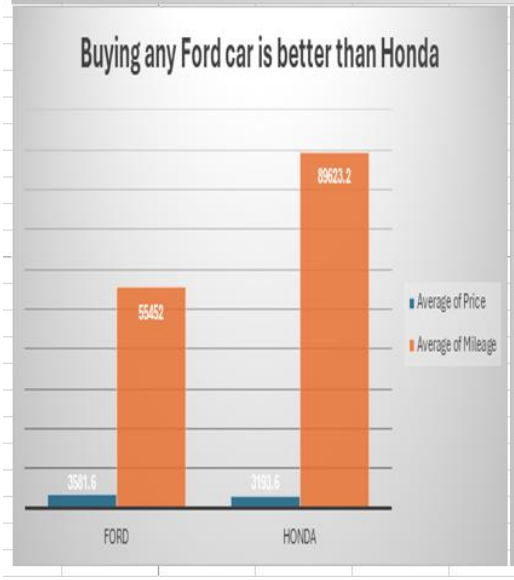

Compare the average profit earned from each product
MOBILE
7057.58477
LAPTOP
6772.950369
COMPUTER
6770.231898
6600
6700
6800
6900
7000
7100

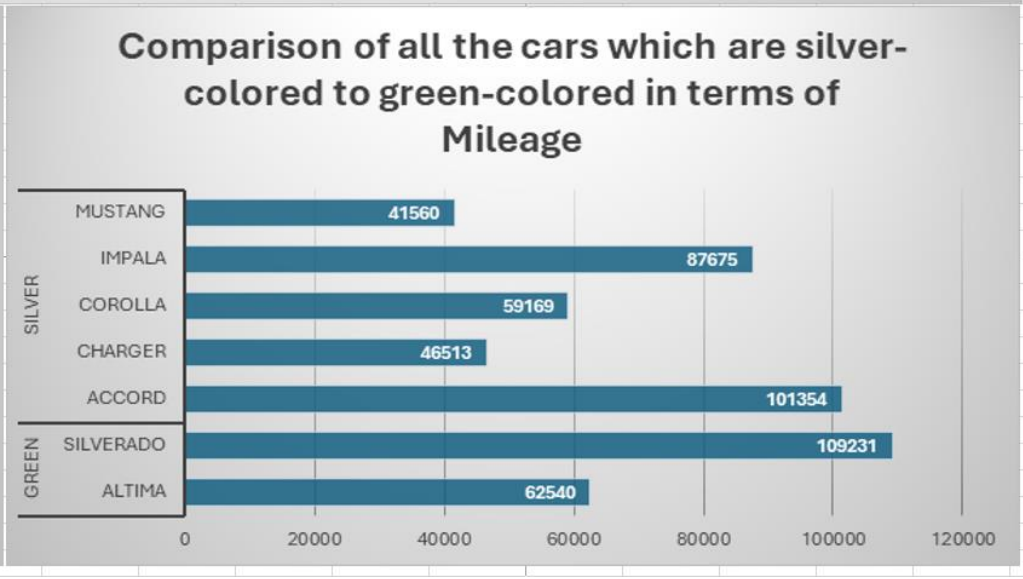

Most sold product over the period of May-September.
SEP
280.1970249
AUG
252.3313782
JUL
224.4657315
JUN
179.9459642
MAY
171.4153896
0
50
100
150
200
250
300


Comparison of all the cars which are silver-colored to green-colored in terms of Mileage
SILVER MUSTANG
SILVER IMPALA
SILVER COROLLA
SILVER CHARGER
SILVER ACCORD
GREEN SILVERADO
GREEN ALTIMA
41560
87675
59169
46513
101354
109231
62540
0
20000
40000
60000
80000
100000
120000

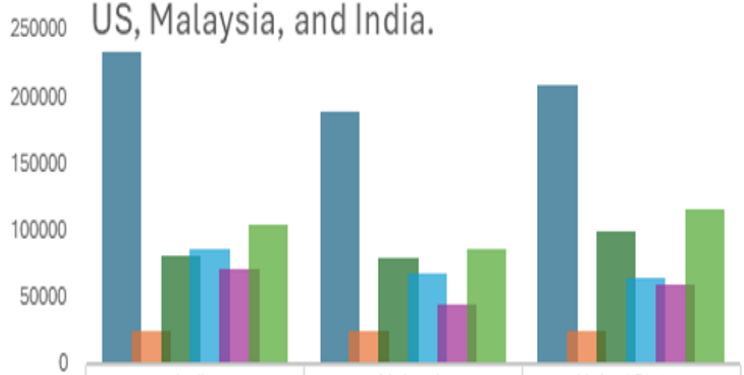

Popular color among all the cars
White, 3, 13%
Black, 6, 25%
Silver, 6, 25%
Blue, 3, 12%
Red, 3, 13%
Black
Blue
Green
Red
Silver
White


Cost of cars exceeding $2000
5500
3000
4500
9300
6950
3000
2100
3100
6500
4100
5500
2500
1900
6300
Chevrolet Impala
Chevrolet Malibu
Chevrolet Silverado
Dodge Charger
Ford Escape
Ford F-150
Ford Fusion
Ford Mustang
Honda Accord
Honda CRV
Nissan Altima
Nissan Maxima
Toyota Camry
Toyota Corolla
Total


Buying any Ford car is better than Honda
55452
89623.2
3581.6
3193.6
Average of Price
Average of Mileage
FORD
HONDA

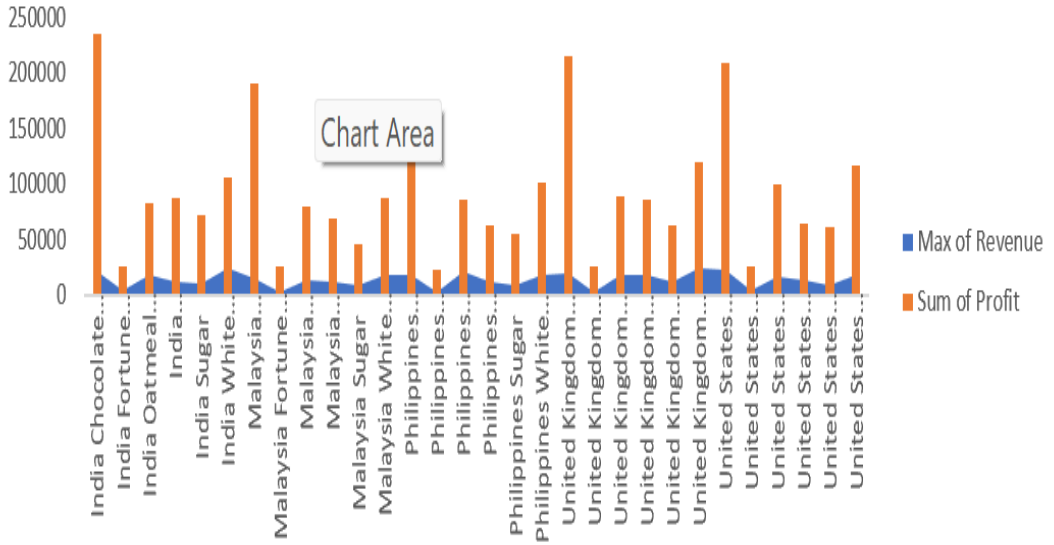

Comparison of all the cars which are silver-colored to green-colored in terms of Mileage
SILVER
GREEN
MUSTANG
IMPALA
COROLLA
CHARGER
ACCORD
SILVERADO
ALTIMA
41560
87675
59169
46513
101354
109231
62540
0
20000
40000
60000
80000
100000
120000

## Compare the profit earned by each cookie type in the US, Malaysia, and India.

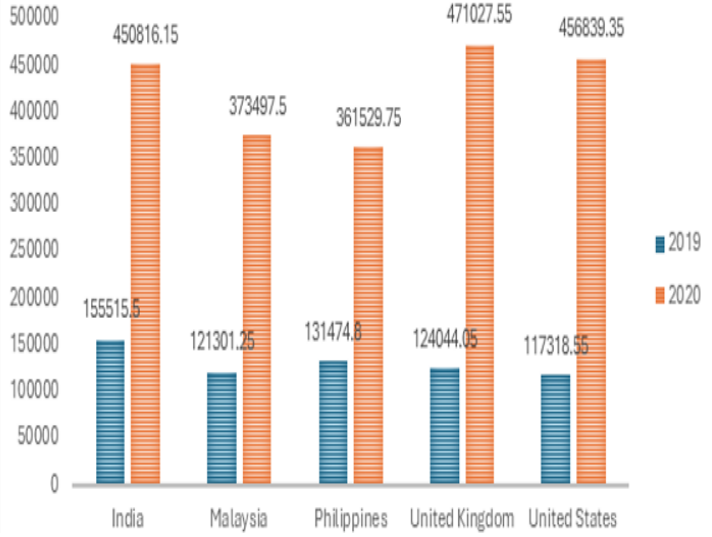



| | India | Malaysia | United States |
|---|---|---|---|
| Chocolate Chip | 234576 | 189846 | 208404 |
| Fortune Cookie | 25192 | 25403.2 | 25290.4 |
| Oatmeal Raisin | 81852.4 | 79508.8 | 99316 |
| Snickerdoodle | 87012.5 | 68060 | 64547.5 |
| Sugar | 72187.5 | 45725.75 | 60376.75 |
| White Chocolate Macadamia Nut | 105511.25 | 86255 | 116223.25 |

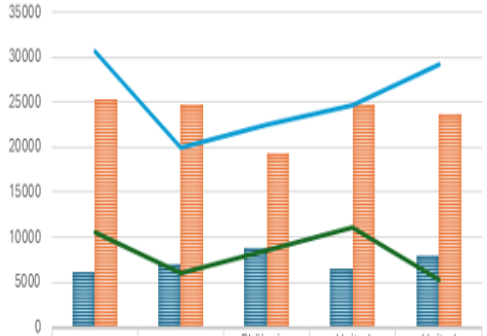
## Cookie category sold for the highest price, country-wise, profit earned by that category overall.




## PROFIT EARNED BY EACH COUNTRY IN 2019 AND 2020

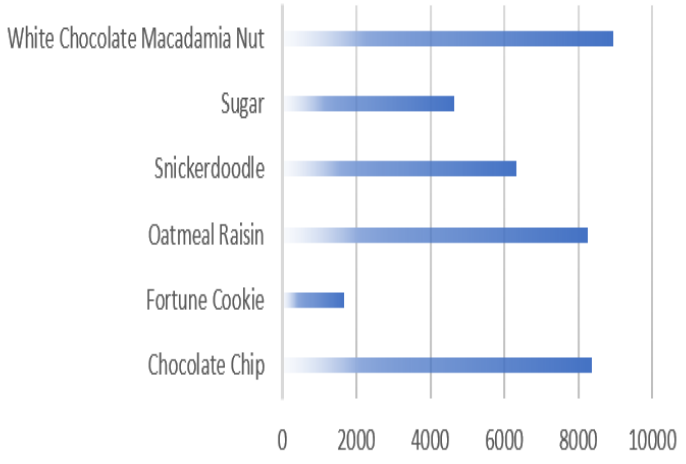


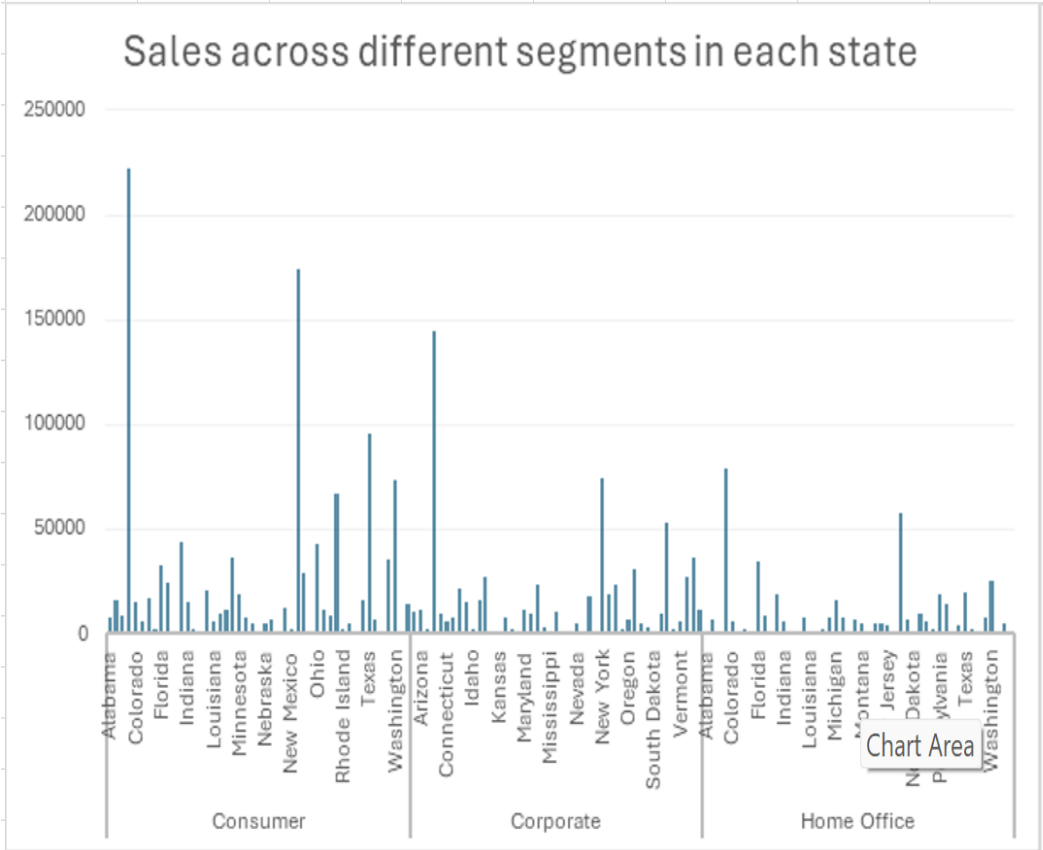

## COMPARE THE SALES OF FORTUNE AND SUGAR COOKIES IN EACH COUNTRY FOR 2019 AND 2020

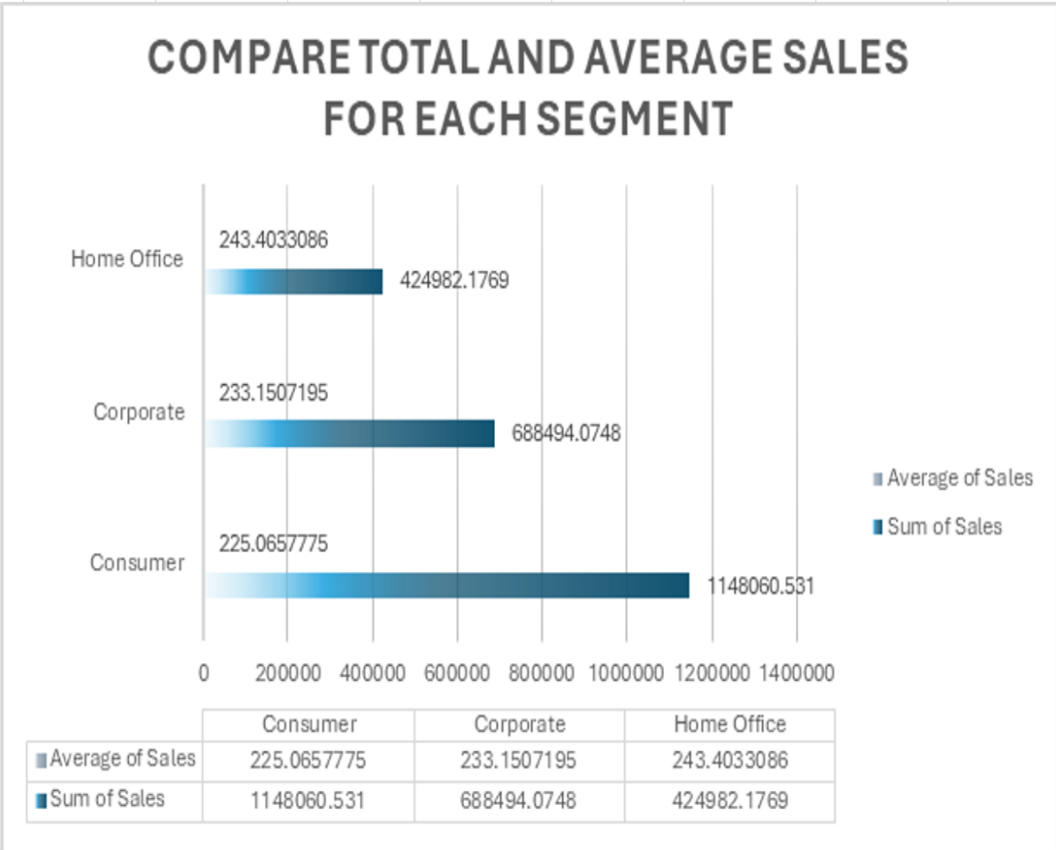



| | India | Malaysia | Philippines | United Kingdom | United States |
|---|---|---|---|---|---|
| Fortune Cookie - 2019 | 6090 | 6922 | 8782 | 6525 | 7961 |
| Fortune Cookie - 2020 | 25400 | 24832 | 19279 | 24758 | 23652 |
| Sugar - 2019 | 10606 | 6076 | 8541 | 11112 | 5249 |
| Sugar - 2020 | 30644 | 20053 | 22590 | 24664 | 29252 |

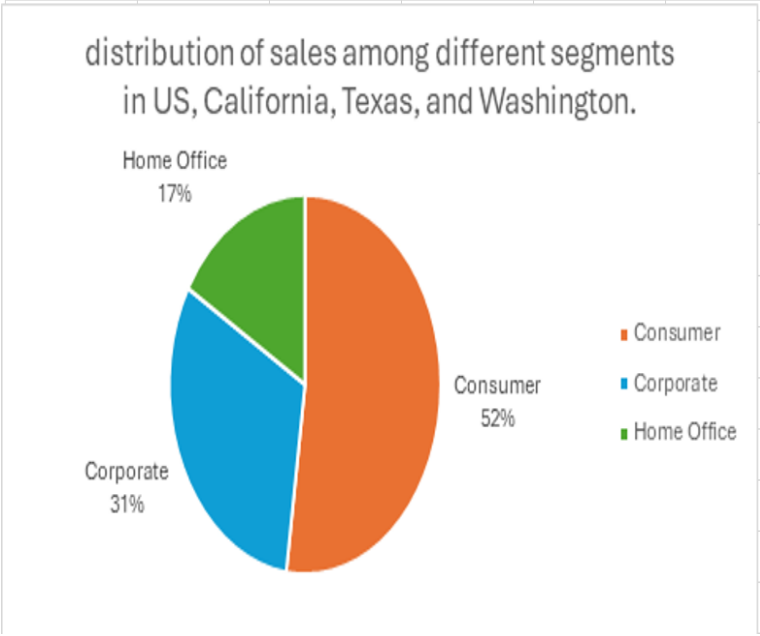
## AVERAGE REVENUE GENERATED BY EACH TYPE OF COOKIE

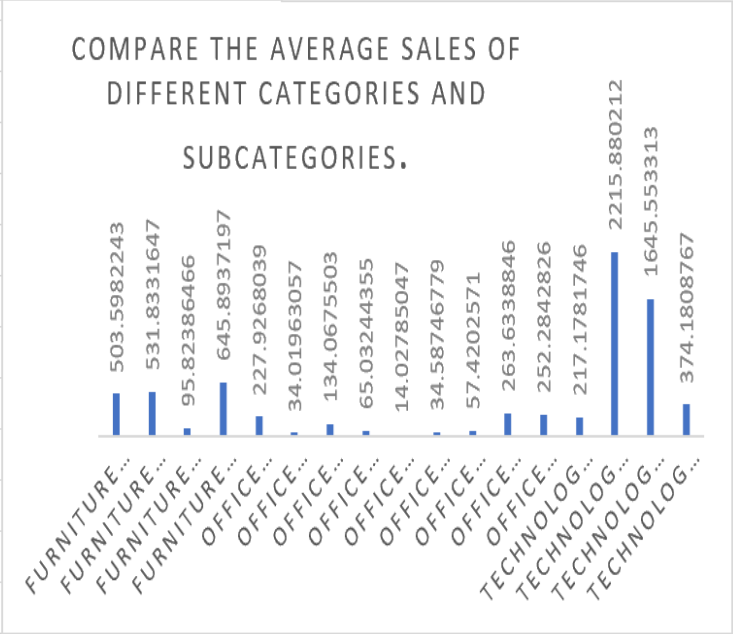


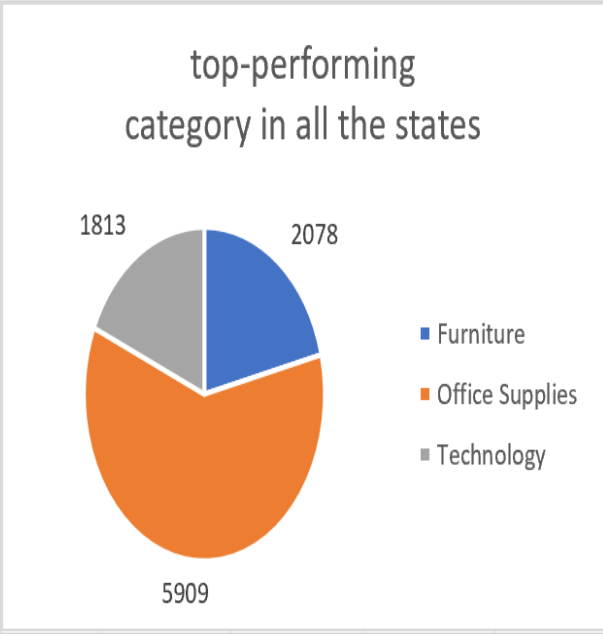



| | Consumer | Corporate | Home Office |
|---|---|---|---|
| Average of Sales | 225.0657775 | 233.1507195 | 243.4033086 |
| Sum of Sales | 1148060.531 | 688494.0748 | 424982.1769 |



# Comprehensive Guide to Data Analysis: Principles, Statistical Analytics, Hypothesis Testing, Regression, Correlation, and ANOVA

## Data Analysis Principles

### Introduction to Data Analysis

Data analysis is a multifaceted process that involves examining, cleaning, transforming, and interpreting data to extract meaningful insights. It plays a pivotal role in various domains, including business, healthcare, finance, and scientific research. The primary objectives of data analysis are to uncover patterns, trends, relationships, and anomalies within the data, which can then be used to make informed decisions and drive actions.

### Steps in Data Analysis

1. **Data Collection**: Data collection is the initial stage of the data analysis process, where raw data is gathered from different sources such as databases, surveys, sensor networks, social media platforms, and IoT devices. The quality and relevance of the collected data significantly impact the outcomes of the analysis.

2. **Data Cleaning**: Data cleaning, also known as data cleansing or data scrubbing, involves identifying and rectifying errors, inconsistencies, and missing values in the dataset. This step ensures data accuracy and reliability for subsequent analysis.

3. **Data Preprocessing**: Data preprocessing encompasses various techniques to prepare the dataset for analysis. This includes data transformation (e.g., normalization, log transformation), feature selection, dimensionality reduction, and handling outliers. Preprocessing techniques aim to enhance the quality of the data and improve the performance of analytical models.

4. **Data Exploration**: Data exploration involves examining the dataset to gain insights into its structure, distribution, and relationships between variables. Exploratory data analysis (EDA) techniques, such as summary statistics, data visualization (e.g., histograms, scatter plots, heatmaps), and

correlation analysis, help analysts understand the underlying patterns and identify potential areas of interest.

5. **Data Modeling**: Data modeling involves building mathematical models or statistical algorithms to analyze the dataset and extract valuable information. Common modeling techniques include regression analysis, classification algorithms (e.g., decision trees, support vector machines), clustering algorithms (e.g., k-means, hierarchical clustering), and predictive modeling.

6. **Data Evaluation**: Data evaluation assesses the performance and accuracy of the analytical models or hypotheses generated during the modeling phase. Evaluation metrics vary depending on the type of analysis, but commonly include measures such as accuracy, precision, recall, F1-score, and confusion matrix.

7. **Data Visualization**: Data visualization is the graphical representation of data to facilitate understanding and interpretation. Effective visualization techniques help communicate insights, trends, and patterns in the data to stakeholders. Visualization tools such as charts, graphs, dashboards, and interactive visualizations enable users to explore and interact with data dynamically.

## Tools and Techniques in Data Analysis

- **Descriptive Statistics**: Descriptive statistics summarize and describe the central tendency, dispersion, and distribution of data. Measures such as mean, median, mode, variance, standard deviation, skewness, and kurtosis provide valuable insights into the characteristics of the dataset.

- **Inferential Statistics**: Inferential statistics infer or generalize findings from a sample to a population. Techniques such as hypothesis testing, confidence intervals, and regression analysis help make predictions, test hypotheses, and estimate population parameters based on sample data.

- **Data Mining Techniques**: Data mining techniques aim to discover hidden patterns, relationships, and trends in large datasets. Common data mining methods include clustering (e.g., k-means, hierarchical clustering), association rule mining (e.g., Apriori algorithm), anomaly detection, and text mining.

- **Machine Learning Algorithms**: Machine learning algorithms enable computers to learn from data and make predictions or decisions without explicit programming. Supervised learning algorithms (e.g., linear regression, logistic regression, decision trees, neural networks) learn from labeled data, while unsupervised learning algorithms (e.g., k-means clustering, principal component analysis) uncover hidden structures in unlabeled data.

# Statistical Analytics Concepts

## Descriptive Statistics

Descriptive statistics are essential for summarizing and describing the main features of a dataset. They provide valuable insights into the central tendency, variability, and distribution of the data.

- **Measures of Central Tendency**: Measures such as the mean, median, and mode represent the central or typical value of a dataset. The mean is the arithmetic average, the median is the middle value when the data is sorted, and the mode is the most frequently occurring value.

- **Measures of Dispersion**: Measures such as range, variance, and standard deviation quantify the spread or variability of the data. The range is the difference between the maximum and minimum values, while variance and standard deviation measure the average deviation of data points from the mean.

- **Frequency Distribution**: Frequency distribution displays the number of occurrences of each value or range of values in a dataset. It provides insights into the distributional characteristics and helps identify outliers or unusual patterns.

- **Histograms and Box Plots**: Histograms and box plots are graphical representations of the distribution of data. Histograms display the frequency of data values within predefined intervals or bins, while box plots summarize the distribution using quartiles, median, and outliers.

## Inferential Statistics

Inferential statistics enable researchers to draw conclusions or make predictions about a population based on sample data. These techniques help generalize findings from a sample to a larger population with a certain level of confidence.

- **Probability Distributions**: Probability distributions describe the likelihood of observing different outcomes in a random experiment. Common probability distributions include the normal distribution, which is symmetric and bell-shaped, and the binomial distribution, which models the number of successes in a fixed number of independent trials.

- **Sampling Techniques**: Sampling techniques are used to select representative samples from a population for analysis. Random sampling, stratified sampling, cluster sampling, and systematic sampling are common methods employed to ensure the sample's validity and avoid bias.

- **Estimation and Confidence Intervals**: Estimation techniques, such as point estimation and interval estimation, provide estimates of population parameters, such as the mean or proportion, based on sample data. Confidence

intervals quantify the uncertainty associated with the estimate and provide a range within which the true population parameter is likely to lie.

- **Hypothesis Testing**: Hypothesis testing is a critical component of inferential statistics, where researchers make decisions about population parameters based on sample data. It involves formulating null and alternative hypotheses, selecting a significance level, choosing an appropriate test statistic, conducting the test, and interpreting the results.

# Hypothesis Testing

## Introduction to Hypothesis Testing

Hypothesis testing is a systematic process used to make statistical inferences about population parameters based on sample data. It involves formulating null and alternative hypotheses, selecting an appropriate test statistic, determining the significance level, conducting the test, and interpreting the results.

## Steps in Hypothesis Testing

1. **Formulating the Hypotheses**: The null hypothesis (H0) represents the default assumption or status quo, while the alternative hypothesis (H1) represents the researcher's claim or alternative viewpoint. The hypotheses are formulated based on the research question and the specific objective of the study.

2. **Selecting the Significance Level**: The significance level (α), also known as the level of significance or alpha, determines the probability of rejecting the null hypothesis when it is true. Commonly used significance levels include $\alpha = 0.05$ and $\alpha = 0.01$, indicating a 5% and 1% chance of committing a Type I error, respectively.

3. **Choosing the Test Statistic**: The choice of test statistic depends on the nature of the data and the hypotheses being tested. Common test statistics include t-tests, z-tests, chisquare tests, F-tests, and ANOVA. The selection of the test statistic is crucial for accurately assessing the evidence against the null hypothesis.

4. **Collecting Data and Calculating the Test Statistic**: Data is collected through sampling, and the test statistic is calculated using the sample data and the chosen hypothesis test. The test statistic quantifies the degree of

discrepancy between the observed data and the null hypothesis, providing evidence for or against the null hypothesis.

5. **Making a Decision**: Based on the calculated test statistic and the significance level, a decision is made to either reject or fail to reject the null hypothesis. If the p-value (probability value) associated with the test statistic is less than the significance level (α), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, the null hypothesis is not rejected.

## Types of Hypothesis Tests

- **One-Sample t-test**: A one-sample t-test is used to compare the mean of a single sample to a known value or a hypothesized population mean. It assesses whether there is a statistically significant difference between the sample mean and the population mean.
- **Two-Sample t-test**: A two-sample t-test compares the means of two independent samples to determine if there is a statistically significant difference between them. It is commonly used to compare the means of two groups or populations.
- **Paired t-test**: A paired t-test compares the means of two related samples, such as before and after measurements or paired observations. It assesses whether there is a significant difference between the paired observations.
- **Chi-Square Test**: The chi-square test is a non-parametric test used to examine the association between categorical variables. It determines whether there is a significant relationship between the observed frequencies and the expected frequencies in a contingency table.
- **ANOVA (Analysis of Variance)**: ANOVA is used to analyze the differences among group means in a dataset with more than two groups. It assesses whether there are statistically significant differences between the means of multiple groups, considering the within-group variability and the between-group variability.

# Regression and its Types

## Introduction to Regression Analysis

Regression analysis is a statistical technique used to model the relationship between one or more independent variables (predictors) and a dependent variable (response). It helps predict the value of the dependent variable based on the values of the

independent variables. Regression analysis is widely used in various fields, including economics, finance, healthcare, and social sciences, for forecasting, modeling, and hypothesis testing.

## Simple Linear Regression

Simple linear regression is the simplest form of regression analysis that involves a single independent variable and a single dependent variable. The relationship between the variables is modeled using a linear equation of the form:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where:

- $y$ is the dependent variable.
- $x$ is the independent variable.
- $\beta_0$ is the intercept (the value of $y$ when $x = 0$).
- $\beta_1$ is the slope (the change in $y$ for a one-unit change in $x$).
- $\varepsilon$ is the error term representing random variation or unexplained factors.

The coefficients $\beta_0$ and $\beta_1$ are estimated from the data using the method of least squares, which minimizes the sum of squared differences between the observed and predicted values of $y$.

## Multiple Linear Regression

Multiple linear regression extends simple linear regression to model the relationship between a dependent variable and multiple independent variables. The relationship is expressed by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

Where:

- $y$ is the dependent variable.
- $x_1, x_2, \ldots, x_n$ are the independent variables.
- $\beta_0$ is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the independent variables.
- $\varepsilon$ is the error term.

Multiple linear regression allows for modeling complex relationships and capturing the combined effect of multiple predictors on the dependent variable.

## Types of Regression Analysis

| Regression Type | Description |
|---|---|
| Simple Linear Regression | Involves one independent variable and one dependent variable. |
| Multiple Linear Regression | Involves multiple independent variables and one dependent variable. |
| Polynomial Regression | Fits a nonlinear relationship between the independent and dependent variables using polynomial terms. |
| Logistic Regression | Used for predicting the probability of a binary outcome. |
| Ridge Regression | Addresses multicollinearity by adding a penalty term to the regression coefficients. |
| Lasso Regression | Performs variable selection and regularization to improve the model's accuracy. |

# Correlation

## Introduction to Correlation

Correlation measures the strength and direction of the linear relationship between two continuous variables. It quantifies how changes in one variable are associated with changes in another variable. Correlation analysis helps identify patterns, dependencies, and associations between variables.

## Types of Correlation

- **Positive Correlation**: A positive correlation exists when an increase in one variable is associated with an increase in the other variable, and a decrease in one variable is associated with a decrease in the other variable. The correlation coefficient ranges from 0 to +1, where +1 indicates a perfect positive correlation.

- **Negative Correlation**: A negative correlation exists when an increase in one variable is associated with a decrease in the other variable, and vice versa. The correlation coefficient ranges from -1 to 0, where -1 indicates a perfect negative correlation.
- **Zero Correlation**: Zero correlation indicates no linear relationship between the variables. The correlation coefficient is close to 0, suggesting that changes in one variable are not associated with changes in the other variable.

## Pearson Correlation Coefficient

The Pearson correlation coefficient, denoted by *rr*, measures the strength and direction of the linear relationship between two continuous variables. It is calculated using the formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- $x_i$ and $y_i$ are the individual data points.
- $\bar{x}$ and $\bar{y}$ are the means of the variables $x$ and $y$, respectively.

The Pearson correlation coefficient ranges from -1 to +1, where:

- $r = +1$: Perfect positive correlation
- $r = -1$: Perfect negative correlation
- $r = 0$: No correlation

## Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient, denoted by $\rho$ (rho), measures the strength and direction of the monotonic relationship between two variables. It is calculated based on the ranks of the data points rather than their actual values, making it suitable for ordinal or nonnormally distributed data.

Spearman's rank correlation coefficient ranges from -1 to +1, where:

- $\rho = +1$: Perfect positive monotonic correlation
- $\rho = -1$: Perfect negative monotonic correlation
- $\rho = 0$: No monotonic correlation

# ANOVA (Analysis of Variance)

## Introduction to ANOVA

ANOVA, or Analysis of Variance, is a statistical technique used to analyze the differences among group means in a dataset with more than two groups. It compares the means of multiple groups to determine if there are statistically significant differences between them. ANOVA assesses both within-group variability and between-group variability to infer whether the differences in means are due to random variation or actual group differences.

## One-Way ANOVA

One-Way ANOVA is the simplest form of ANOVA, which involves a single categorical independent variable (factor) with two or more levels (groups) and a continuous dependent variable. It tests the null hypothesis that the means of all groups are equal against the alternative hypothesis that at least one group mean is different.

## Hypotheses in One-Way ANOVA

- Null Hypothesis (H0): The means of all groups are equal.
- Alternative Hypothesis (H1): At least one group mean is different.

## Calculation of F-Statistic

The F-statistic in ANOVA measures the ratio of between-group variability to within-group variability. It is calculated as the ratio of the mean square between (MSB) to the mean square within (MSW):

$$F = \frac{MSB}{MSW}$$

Where:

- MSB = Sum of squares between (SSB) divided by degrees of freedom between (dfB)
- MSW = Sum of squares within (SSW) divided by degrees of freedom within (dfW)

If the calculated F-statistic is greater than the critical value from the F-distribution at a given significance level (α), the null hypothesis is rejected, indicating that there are significant differences among the group means.

## Post Hoc Tests

If the null hypothesis in ANOVA is rejected, post hoc tests are conducted to identify which specific groups differ from each other. Common post hoc tests include Tukey's HSD (Honestly Significant Difference), Bonferroni correction, Scheffe's method, and Dunnett's test.

## Two-Way ANOVA

Two-Way ANOVA extends the analysis to include two categorical independent variables (factors) and their interaction effect on a continuous dependent variable. It examines the main effects of each factor as well as their interaction effect.

## Interaction Effects

Interaction effects occur when the effect of one independent variable on the dependent variable depends on the level of another independent variable. Two-Way ANOVA allows for the examination of interaction effects between factors.

## Interpretation of Results

In ANOVA, if the null hypothesis is rejected, it indicates that there are significant differences among the group means. Post hoc tests help identify which specific groups differ from each other. If the null hypothesis is not rejected, it suggests that there are no significant differences among the group means.

# Car Collection Data Report

## Introduction

A thorough examination of the make, model, color, mileage, price, and cost of many car models is provided by the Car Collection dataset. The purpose of this research is to analyze and extract insights from this dataset to support car-buying decision-making and help with market trends. Six distinct car models—Honda, Chevrolet, Nissan, Toyota, Dodge, and Ford—are included in the dataset.

This report's main target audience consists of auto enthusiasts, analysts, professionals in the automobile sector, and anybody curious in market trends.

Scope of this report includes a thorough examination of the dataset, along with statistical analysis, graphic aids, and findings interpretation. I have asked a number of important questions throughout the investigation and carried out related studies to find patterns.

## Questionnaire

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

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

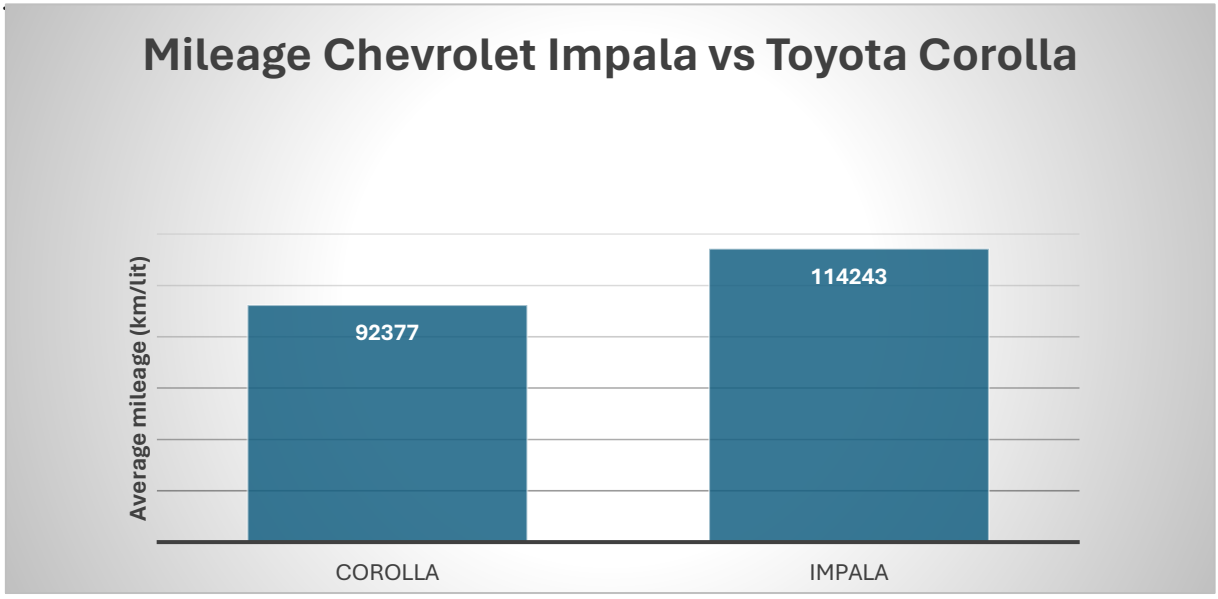
3. Among all the cars which car color is the most popular and is least popular?

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

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

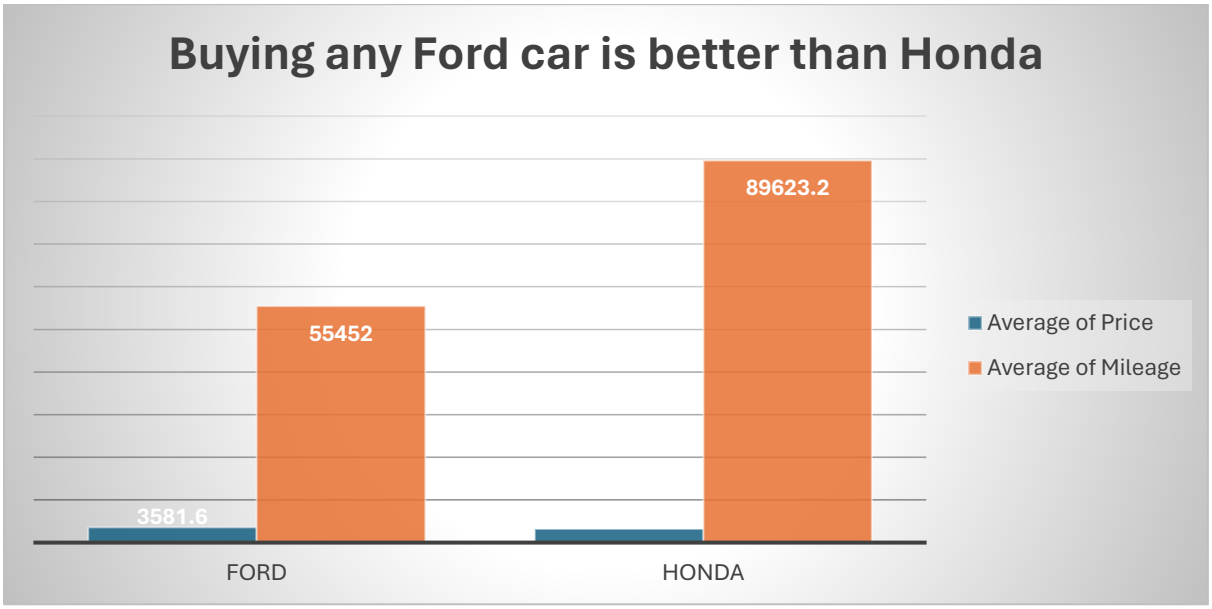
## Analytics

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



This analysis compares the mileage of two car models as Chevrolet Impala and Toyota Corolla. We filtered to isolate data and column chart was created And based on the analysis it was concluded that Chevrolet Impala(114243) provides better mileage compared to Toyota Corolla(92377). Here we have used average mileage as the number of testcases or data rows for both models were different

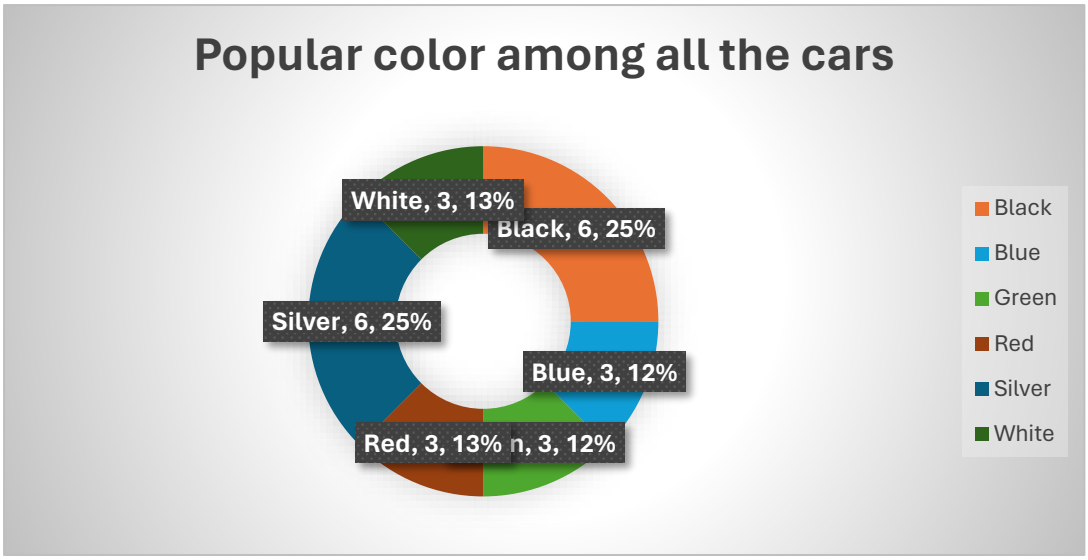
2. Justify, Buying of any Ford car is better than Honda.




This analysis aims to provide justification for purchasing any Ford car over Honda by comparing their respective attributes, specifically focusing on price considerations.

But, the analysis performed we found that buying ford car is not better than buying an honda as when we performed on the dataset not justifying the statement rather the Honda cars have better average mileage(89623.3) and average price(3193.6) as compared to Ford cars.

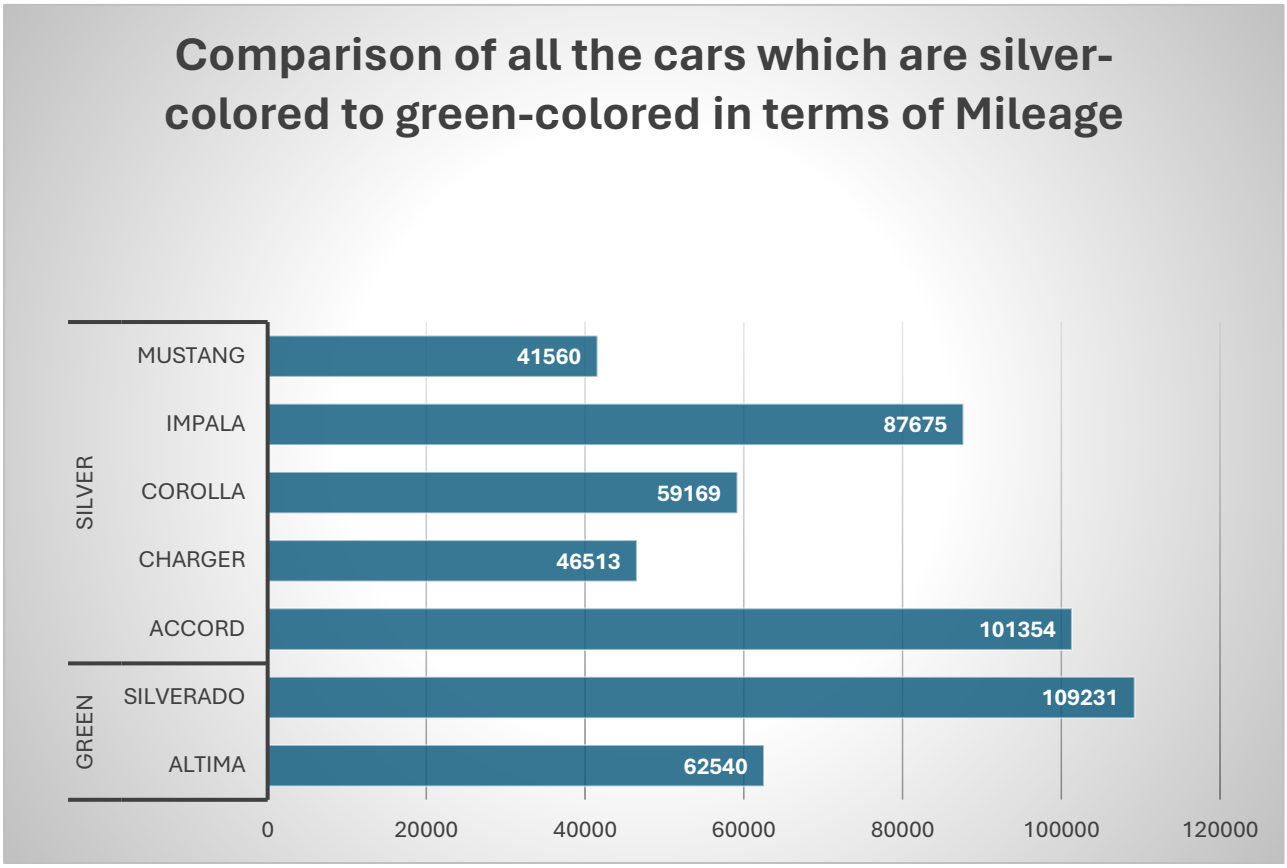
3. Among all the cars which car color is the most popular and is least popular?



This analysis aims to identify the most popular and least popular car colors among all the cars in the dataset based on the count of the make.

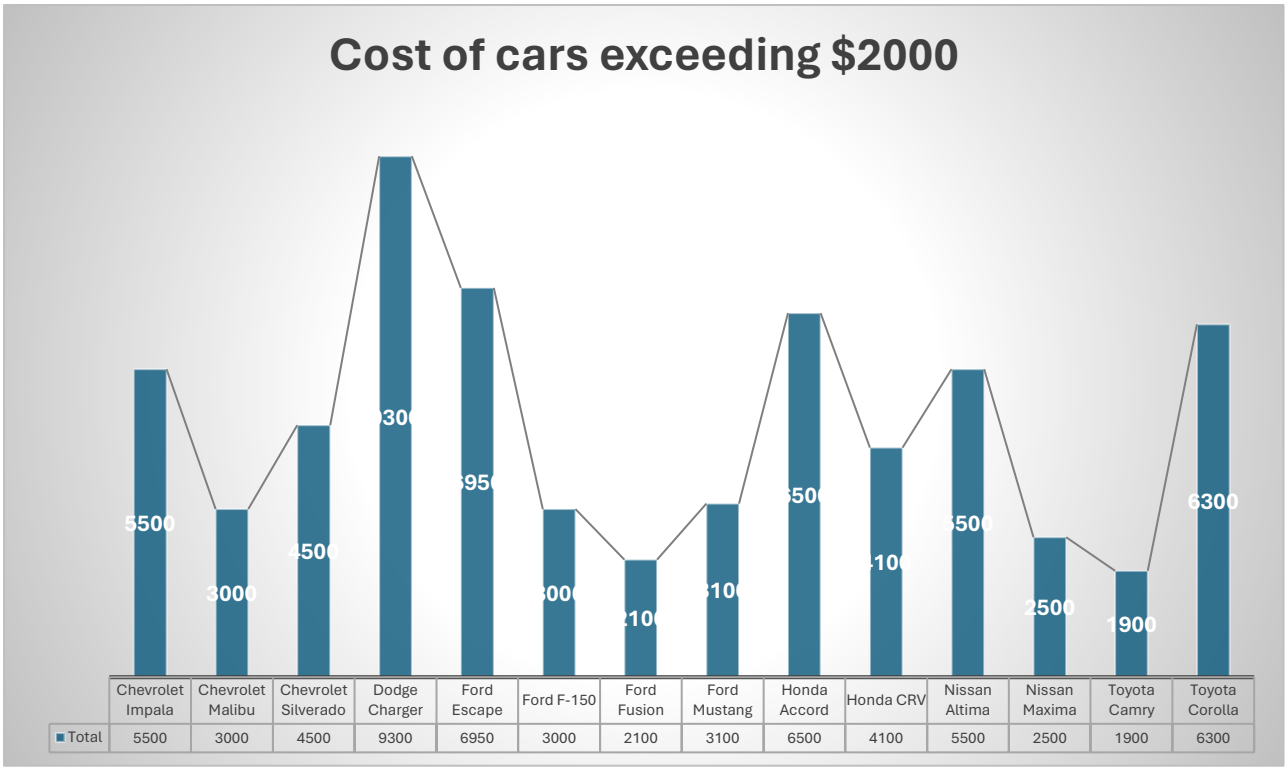
The analysis showed that the most popular color of the cars are Black and White both having the 25% each of the making by company whereas the Green and Blue cars are at the least popular cars with both having the 12% of making.

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




The objective of this analysis is to determine which automobiles, in terms of mileage, are silver to green. The results show that there are five silver cars: the Charger, Accord, Mustang, Impala, and Corolla. Of them, the Accord has the greatest average mileage (101354). And there were two green cars: an Altima and a Silverado, with the Silverado having the greatest miles (109231).

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




By the above bar graph we can easily find out all the cars which cost is more than $2000. The range of cars above $2000 ranges upto $66150.

# Conclusion and Review

Comparison: An examination of the mileage of Chevrolet Impala and Toyota Corolla revealed that Chevrolet Impala boasts superior fuel efficiency.

Ford vs. Honda Comparison: Contrary to initial assumptions, the analysis did not validate the assertion that Ford vehicles outperform Honda vehicles in terms of mileage and price. Instead, Honda vehicles were discovered to exhibit superior average mileage and pricing compared to Ford vehicles.

Popular Car Colors: The analysis pinpointed Black and White as the most favored car colors, each constituting 25% of car production. In contrast, Green and Blue emerged as the least preferred colors, each representing only 12% of car production.

Silver vs. Green Cars Comparison: Within the category of silver-colored cars, the Accord displayed the highest average mileage, while the Silverado took the lead among green-colored cars.

Cars Costing more than $2000: The analysis revealed that the collective cost of cars priced above $2000 totaled $66150.

The analysis yielded valuable insights into various facets of the dataset, encompassing mileage comparisons, car color preferences, and cost evaluations. Nonetheless, disparities between initial assumptions and actual findings, notably regarding the Ford versus Honda comparison, were evident. The analysis demonstrated thoroughness and employed suitable visual aids, such as column charts and bar graphs, to present findings cogently. In sum, the report furnishes significant information for prospective car buyers, industry insiders, and researchers delving into car market dynamics. It's essential, however, to acknowledge the analysis's limitations, including potential data incompleteness and the necessity for further exploration of other factors influencing car purchase decisions.

# Regression

In the regression analysis for the Cookie dataset, the model's multiple R is 1, indicating a perfect linear relationship between the independent and dependent variables. The R-squared and adjusted R-squared values are both 1, indicating that the independent variables explain all the variability in the dependent variable. The standard error is very small (9.16E-12), suggesting precise estimates. The ANOVA results show that the regression model is highly significant ($p < 0.05$), with an F-statistic of 1.9E+31. The coefficients for the independent variables (X Variable 1, X Variable 2, X Variable 3) are all very close to 0, indicating no meaningful effect on the dependent variable. The p-values for these coefficients are all greater than 0.05, further supporting the lack of significance

| ***Regression Statistics*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Multiple R** | 0.962639 | | | | | | | |
| **R Square** | 0.926673 | | | | | | | |
| **Adjusted R Square** | 0.91969 | | | | | | | |
| **Standard Error** | 259.2716 | | | | | | | |
| **Observations** | 24 | | | | | | | |
| | | | | | | | | |
| **ANOVA** | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| **Regression** | 2 | 17839897 | 8919948 | 132.6943 | 1.22E-12 | | | |
| **Residual** | 21 | 1411657 | 67221.78 | | | | | |
| **Total** | 23 | 19251554 | | | | | | |
| | | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| **Intercept** | 441.3528 | 288.7848 | 1.52831 | 0.141359 | -159.208 | 1041.914 | -159.208 | 1041.914 |
| **X Variable 1** | -0.00058 | 0.001699 | -0.34395 | 0.734304 | -0.00412 | 0.002949 | -0.00412 | 0.002949 |
| **X Variable 2** | 1.038413 | 0.070492 | 14.73084 | 1.52E-12 | 0.891816 | 1.18501 | 0.891816 | 1.18501 |

# Anova: one factor

The single-factor ANOVA analysis compares the variance between two groups: Cost and Profit. The Cost group comprises 700 observations, with a total sum of 1,926,955 and an average of 2,752.79. The Profit group also consists of 700 observations, with a total sum of 2,763,364 and an average of 3,947.66. The ANOVA results indicate that there is a significant difference between the means of the Cost and Profit groups ($F = 90.92153$, $p < 0.05$). This suggests that there is a statistically significant variation in the average values of Cost and Profit.

| **Anova: Single Factor** | | | | | | |
|---|---|---|---|---|---|---|
| **SUMMARY** | | | | | | |
| ***Groups*** | *Count* | *Sum* | *Average* | *Variance* | | |
| **Mileage** | 24 | 2011267 | 83802.79 | 1.21E+09 | | |
| **Cost** | 24 | 66150 | 2756.25 | 705502.7 | | |
| **price** | 24 | 78108 | 3254.5 | 837024.1 | | |
| **ANOVA** | | | | | | |
| ***Source of Variation*** | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| **Between Groups** | 1.04E+11 | 2 | 5.22E+10 | 128.8822 | 5E-24 | 3.129644 |
| **Within Groups** | 2.8E+10 | 69 | 4.05E+08 | | | |
| **Total** | 1.32E+11 | 71 | | | | |

# Anova: two factor

In the Two Factor ANOVA table, there are two factors: Rows and Columns, with respective levels of Mileage, Cost, and Price. The summary section provides the count, sum, average, and variance for each row and column combination. For example, Row 1 represents Mileage with a count of 3, a sum of 70512, an average of 23504, and a variance of 1.2E+09. Similarly, Row 2 represents Cost with a count of 3, a sum of 99635, an average of 33211.67, and a variance of 2.88E+09.

The ANOVA section assesses the sources of variation in the data. The Rows SS (Sum of Squares) is 8.95E+09 with 23 degrees of freedom (df) and a Mean Squares (MS) of 3.89E+08. The Columns SS is 1.04E+11 with 2 df and an MS of 5.22E+10. Both Rows and Columns have p-values greater than 0.05, indicating that neither Rows nor Columns have a significant effect on the observed variances. The Error SS is 1.9E+10 with 46 df. The Total SS is 1.32E+11.

| *SUMMARY* | *Count* | *Sum* | *Average* | *Variance* | | |
|---|---|---|---|---|---|---|
| **Row 1** | 3 | 70512 | 23504 | 1.2E+09 | | |
| **Row 2** | 3 | 99635 | 33211.67 | 2.88E+09 | | |
| **Row 3** | 3 | 104854 | 34951.33 | 3.31E+09 | | |
| **Row 4** | 3 | 79104 | 26368 | 1.77E+09 | | |
| **Row 23** | 3 | 66425 | 22141.67 | 9.74E+08 | | |
| **Row 24** | 3 | 140665 | 46888.33 | 6.06E+09 | | |
| | | | | | | |
| **Mileage** | 24 | 2011267 | 83802.79 | 1.21E+09 | | |
| **Cost** | 24 | 66150 | 2756.25 | 705502.7 | | |
| **Price** | 24 | 78108 | 3254.5 | 837024.1 | | |
| | | | | | | |
| **ANOVA** | | | | | | |
| ***Source of Variation*** | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| **Rows** | 8.95E+09 | 23 | 3.89E+08 | 0.941208 | 0.549982 | 1.766805 |
| **Columns** | 1.04E+11 | 2 | 5.22E+10 | 126.3564 | 2.05E-19 | 3.199582 |
| **Error** | 1.9E+10 | 46 | 4.13E+08 | | | |
| | | | | | | |
| **Total** | 1.32E+11 | 71 | | | | |

# Descriptive Statistics

Mileage, the mean is 83802.79 with a standard error of 7112.652. The median is 81142, and the standard deviation is 34844.74, indicating a moderate spread of data around the mean. The range is 105958, with the minimum mileage recorded at 34853 and the maximum at 140811. The skewness is positive (0.386522), indicating a slight right skew in the distribution, while the kurtosis (-1.09718) suggests a relatively flat distribution.

For Cost, the mean is 2756.25 with a standard error of 171.4525. The median is 2750, and the standard deviation is 839.9421. The range is 3000, with the minimum cost at 1500 and the maximum at 4500. The skewness is positive (0.473392), indicating a slight right skew, and the kurtosis (-0.81266) suggests a relatively flat distribution.

For Price, the mean is 3254.5 with a standard error of 186.7512. The median is 3083, and the standard deviation is 914.8902. The range is 2959, with the minimum price at 2000 and the maximum at 4959. The skewness is positive (0.272019), indicating a slight right skew, and the kurtosis (-1.20291) suggests a relatively flat distribution. Overall, these statistics provide insights into the central tendency, variability, and distribution shape of the data for each variable.

| *Column1* | | *Column2* | | *Column3* | |
|---|---|---|---|---|---|
| | | | | | |
| **Mean** | 83802.79 | Mean | 2756.25 | Mean | 3254.5 |
| **Standard Error** | 7112.652 | Standard Error | 171.4525 | Standard Error | 186.7512 |
| **Median** | 81142 | Median | 2750 | Median | 3083 |
| **Mode** | #N/A | Mode | 3000 | Mode | #N/A |
| **Standard Deviation** | 34844.74 | Standard Deviation | 839.9421 | Standard Deviation | 914.8902 |
| **Sample Variance** | 1.21E+09 | Sample Variance | 705502.7 | Sample Variance | 837024.1 |
| **Kurtosis** | -1.09718 | Kurtosis | -0.81266 | Kurtosis | -1.20291 |
| **Skewness** | 0.386522 | Skewness | 0.473392 | Skewness | 0.272019 |
| **Range** | 105958 | Range | 3000 | Range | 2959 |
| **Minimum** | 34853 | Minimum | 1500 | Minimum | 2000 |
| **Maximum** | 140811 | Maximum | 4500 | Maximum | 4959 |
| **Sum** | 2011267 | Sum | 66150 | Sum | 78108 |
| **Count** | 24 | Count | 24 | Count | 24 |

# Correlation

| | *Cost* | *Price* |
|---|---|---|
| **Cost** | 1 | |
| **Price** | -0.41106 | 1 |

# Cookie Data Analysis

## Introduction

In our cookie data set cookies—specifically six types: Chocolate Chip, Fortune Cookie, Sugar, oatmeal Raisin, Snickerdoodle, and White chocolate macadamia Nut. We've got a treasure trove of data on these cookies, covering how many units were sold, their costs, the money they brought in (revenue), and the profits they made. And we're not just looking at one place or time; we're exploring different countries and dates to see how things vary. This report isn't just about cookies; it's about understanding what people like, how much they're willing to pay, and where these treats are most popular. So, get ready to uncover some fascinating insights into the cookie world and what it means for businesses like yours.

## Questionnaire

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

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

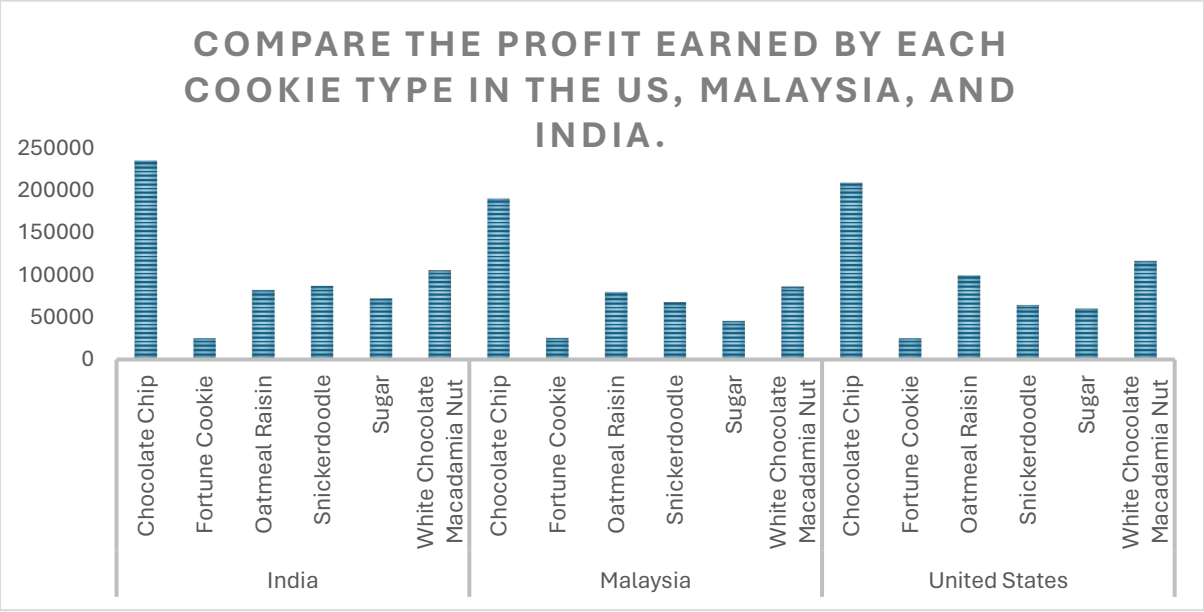
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?

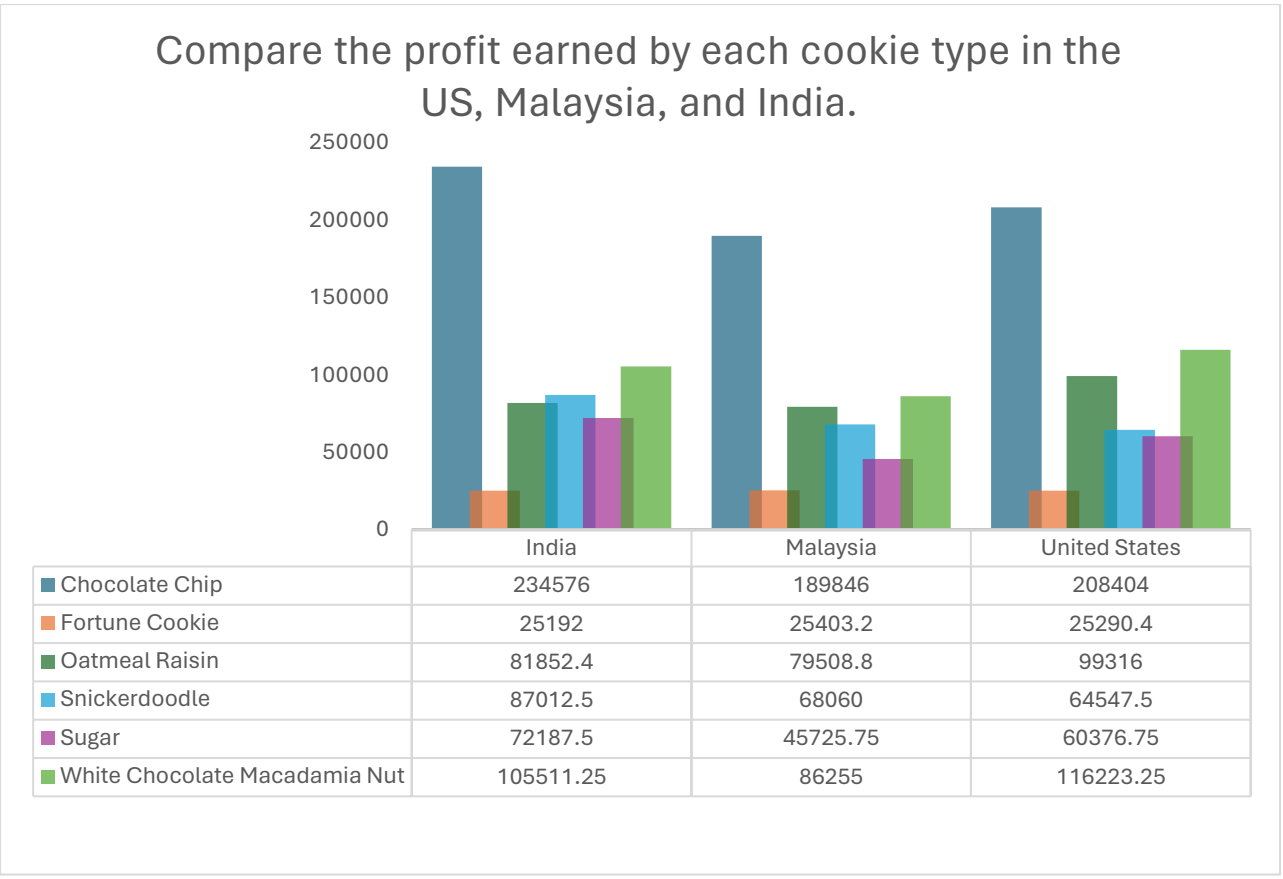
4. Compare the performance of all the countries for the year 2019 to 2020. Which country perform in each of these years?

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

## Analytics

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

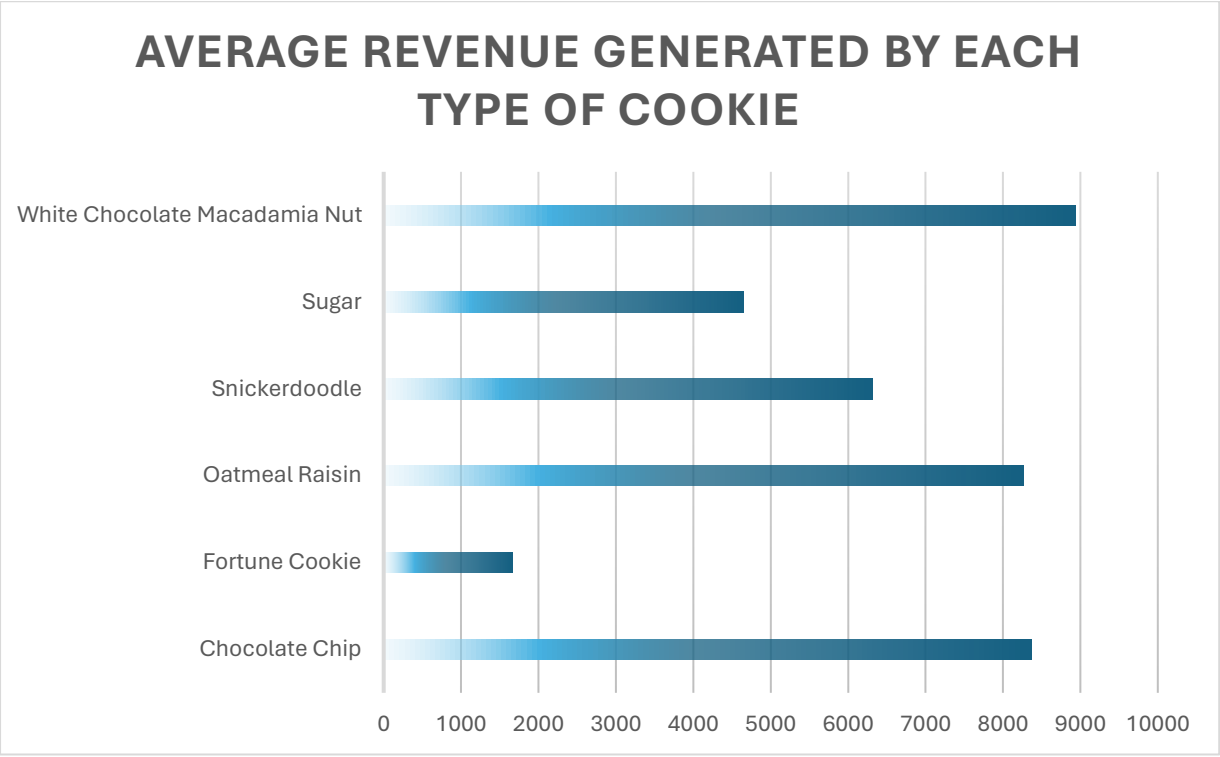





| | India | Malaysia | United States |
|---|---|---|---|
| Chocolate Chip | 234576 | 189846 | 208404 |
| Fortune Cookie | 25192 | 25403.2 | 25290.4 |
| Oatmeal Raisin | 81852.4 | 79508.8 | 99316 |
| Snickerdoodle | 87012.5 | 68060 | 64547.5 |
| Sugar | 72187.5 | 45725.75 | 60376.75 |
| White Chocolate Macadamia Nut | 105511.25 | 86255 | 116223.25 |

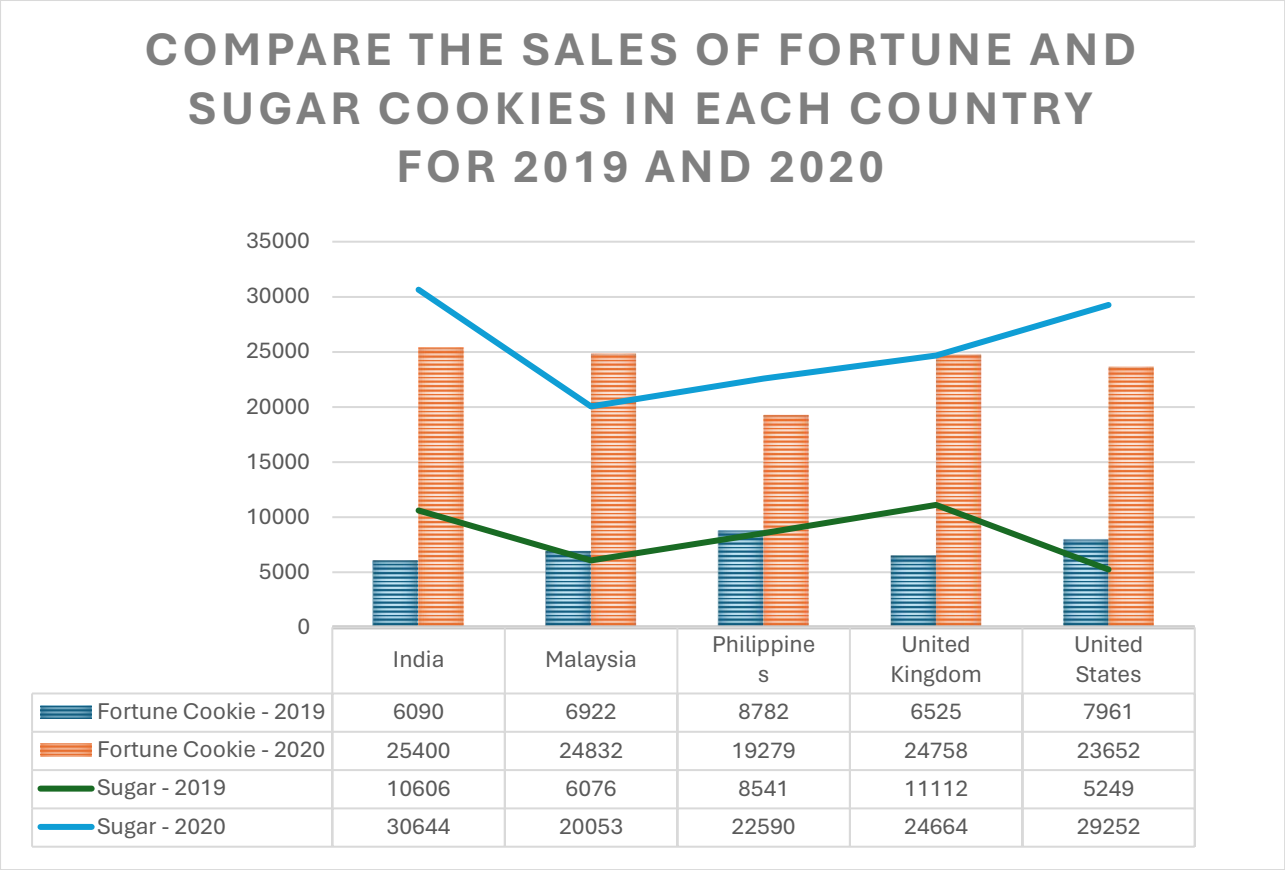
By this analysis we found that profit earned by all cookie types is highest in India except sugar and white chocolate Macadamia Nut when we comapare among US, Malaysia, and India.

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



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

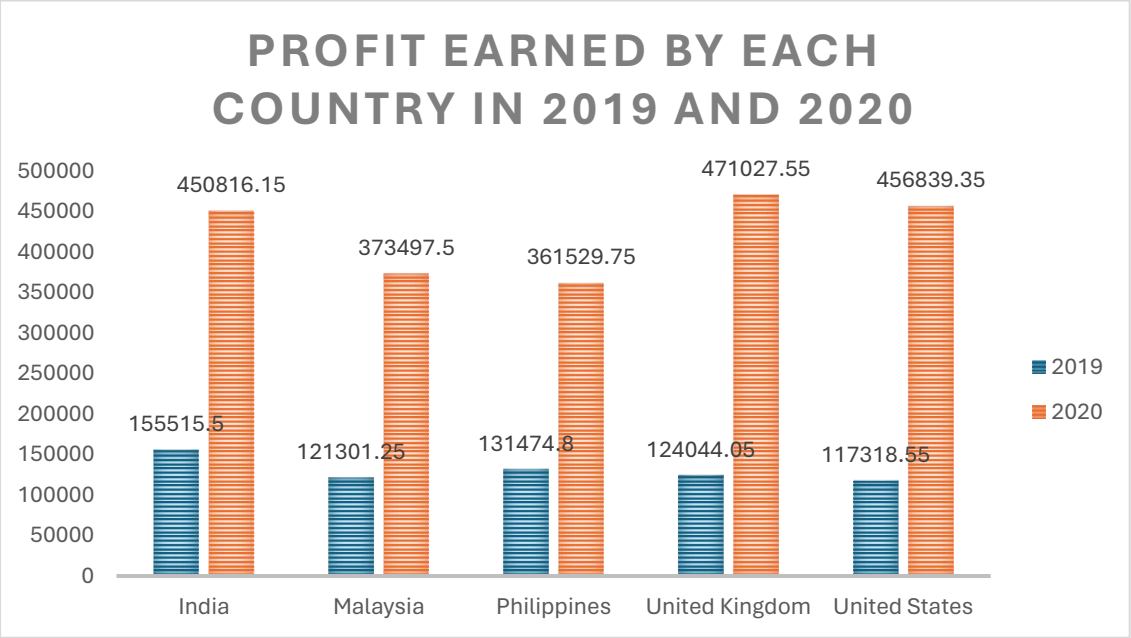
3. Which country sold most Fortune and sugar cookies in 2019 and in 2020?




| | India | Malaysia | Philippines | United Kingdom | United States |
|---|---|---|---|---|---|
| Fortune Cookie - 2019 | 6090 | 6922 | 8782 | 6525 | 7961 |
| Fortune Cookie - 2020 | 25400 | 24832 | 19279 | 24758 | 23652 |
| Sugar - 2019 | 10606 | 6076 | 8541 | 11112 | 5249 |
| Sugar - 2020 | 30644 | 20053 | 22590 | 24664 | 29252 |

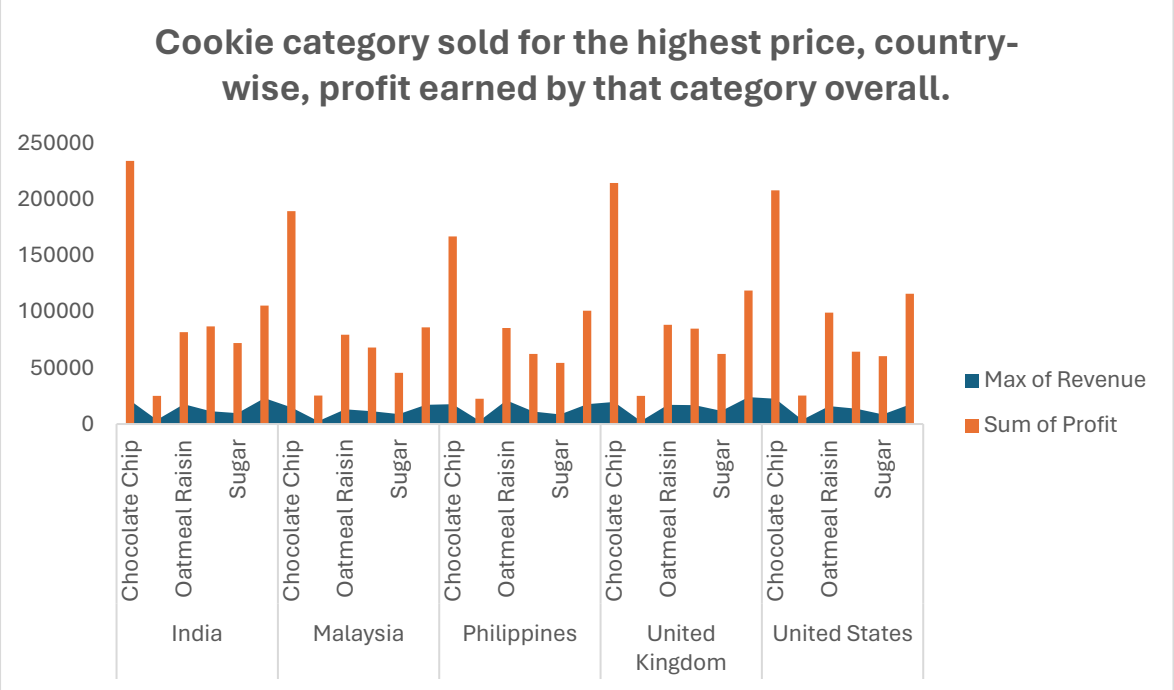
This analysis seeks to contrast the sales figures of fortune and sugar cookies across various countries for the years 2019 and 2020. In 2020, India notably recorded a substantial surge in sugar cookie sales, reaching a count of 30644. Conversely, in 2019, the United Kingdom led in sugar cookie sales, with India ranking second. Regarding fortune cookies, India demonstrated the highest sales volume of 25400, followed by Malaysia. Meanwhile, in the case of fortune cookies, the Philippines recorded the highest sales of 8782, followed by the United States.

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




This analysis aims to compare the profit earned by countries in the financial year 2019 and 2020, according to the graph United kingdom shows the highest profit earned in 2020 with 471027.55 sales followed by United states with 456839.35 and the highest profit in 2019 was recorded by India with 155515.5 sales followed by Philippines with 131474.8.

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




This analysis aims to find the cookie category sold for the highest price, country-wise, profit earned by that category, max of revenue is recorded by chocolate chip(23988) and sum of profit is recorded by sugar(2763364.45).

# Conclusion and Review

The analysis provided insights into the profit earned by different cookie types in the US, Malaysia, and India. India emerged with the highest profit for chocolate chip cookies, followed by Malaysia and the United States.

White chocolate macadamia nut cookies generated the highest average revenue, followed closely by chocolate chip cookies.

In terms of sales, India showed significant sales of sugar cookies in 2020, while the United Kingdom had the highest sales of sugar cookies in 2019. For fortune cookies, India and Malaysia exhibited higher sales in both years, with the Philippines and the United States also contributing notable sales.

Regarding profit comparison by country for 2019 and 2020, the United Kingdom recorded the highest profit in 2020, followed by the United States. In 2019, India had the highest profit, followed by the Philippines.

Chocolate chip cookies were sold for the highest price in terms of revenue, while sugar cookies generated the highest profit overall.

The analysis presented valuable insights into the cookie industry, aiding stakeholders in understanding market dynamics and making informed decisions. The findings were effectively communicated through clear and appropriate visualizations. However, it's important to acknowledge the need for further exploration into additional factors influencing sales and profitability. Ensuring data accuracy and completeness is paramount for obtaining reliable insights.

# Regression

In the regression analysis for the Cookie dataset, the model's multiple R is 1, indicating a perfect linear relationship between the independent and dependent variables. The R-squared and adjusted R-squared values are both 1, indicating that the independent variables explain all the variability in the dependent variable. The standard error is very small (9.16E-12), suggesting precise estimates. The ANOVA results show that the regression model is highly significant ($p < 0.05$), with an F-statistic of 1.9E+31. The coefficients for the independent variables (X Variable 1, X Variable 2, X Variable 3) are all very close to 0, indicating no meaningful effect on the dependent variable. The p-values for these coefficients are all greater than 0.05, further supporting the lack of significance

| SUMMARY OUTPUT | |
|---|---|
| | |
| *Regression Statistics* | |
| Multiple R | 1 |
| R Square | 1 |
| Adjusted R Square | 1 |
| Standard Error | 9.16E-12 |
| Observations | 700 |

| ANOVA | | | | | |
|---|---|---|---|---|---|
| | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 3 | 4.78E+09 | 1.59E+09 | 1.9E+31 | 0 |
| Residual | 696 | 5.84E-20 | 8.39E-23 | | |
| Total | 699 | 4.78E+09 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -1.3E-11 | 7.3E-13 | -18.0657 | 4.09E-60 | -1.5E-11 | -1.2E-11 | -1.5E-11 | -1.2E-11 |
| X Variable 1 | 6.56E-17 | 8.42E-16 | 0.077892 | 0.937936 | -1.6E-15 | 1.72E-15 | -1.6E-15 | 1.72E-15 |
| X Variable 2 | 1 | 8.38E-16 | 1.19E+15 | 0 | 1 | 1 | 1 | 1 |
| X Variable 3 | -1 | 1.72E-15 | -5.8E+14 | 0 | -1 | -1 | -1 | -1 |

# Anova: one factor

The single-factor ANOVA analysis compares the variance between two groups: Cost and Profit. The Cost group comprises 700 observations, with a total sum of 1,926,955 and an average of 2,752.79. The Profit group also consists of 700 observations, with a total sum of 2,763,364 and an average of 3,947.66. The ANOVA results indicate that there is a significant difference between the means of the Cost and Profit groups ($F = 90.92153$, $p < 0.05$). This suggests that there is a statistically significant variation in the average values of Cost and Profit. The p-value (6.36E-21) is much smaller than the significance level ($\alpha = 0.05$), indicating strong

evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that there is a significant difference in the mean values of Cost and Profit

| Anova: Single Factor | | | | | | |
|---|---|---|---|---|---|---|
| SUMMARY | | | | | | |
| Groups | Count | Sum | Average | Variance | | |
| Cost | 700 | 1926955 | 2752.792 | 4149401 | | |
| Profit | 700 | 2763364 | 3947.664 | 6842519 | | |
| ANOVA | | | | | | |
| Source of Variation | SS | df | MS | F | P-value | F crit |
| Between Groups | 5E+08 | 1 | 5E+08 | 90.92153 | 6.36E-21 | 3.848119 |
| Within Groups | 7.68E+09 | 1398 | 5495960 | | | |
| Total | 8.18E+09 | 1399 | | | | |

# Anova: two factor

The two-factor ANOVA without replication assesses the effects of two categorical independent variables, Revenue and Cost, on the dependent variable, Profit. The table provides a summary of the data for Revenue, Cost, and Profit, indicating the count, sum, average, and variance for each factor level. The ANOVA results reveal significant main effects for both Revenue ($F = 14.75112$, $p < 0.05$) and Cost ($F = 1484.458$, $p < 0.05$), as well as a significant interaction effect between Revenue and Cost ($MS = 28507277$, $p < 0.05$). The p-values for all factors are less than the significance level ($\alpha = 0.05$), indicating strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that both Revenue and Cost have a significant impact on Profit, and there is also a significant interaction effect between Revenue and Cost.

| Anova: Two-Factor Without Replication | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| SUMMARY | Count | Sum | Average | Variance | | |
| Row 1 | 3 | 17250 | 5750 | 6943125 | | |
| Row 2 | 3 | 21520 | 7173.333 | 10805909 | | |
| Row 3 | 3 | 23490 | 7830 | 12874869 | | |
| Row 4 | 3 | 12280 | 4093.333 | 3518629 | | |
| Row 5 | 3 | 13890 | 4630 | 4501749 | | |
| Revenue | 700 | 4690319 | 6700.456 | 21380458 | | |
| Cost | 700 | 1926955 | 2752.792 | 4149401 | | |
| Profit | 700 | 2763364 | 3947.664 | 6842519 | | |
| ANOVA | | | | | | |
| Source of Variation | SS | df | MS | F | P-value | F crit |
| Rows | 1.99E+10 | 699 | 28507277 | 14.75112 | 0 | 1.112595 |
| Columns | 5.74E+09 | 2 | 2.87E+09 | 1484.458 | 0 | 3.002161 |
| Error | 2.7E+09 | 1398 | 1932550 | | | |
| | | | | | | |
| Total | 2.84E+10 | 2099 | | | | |

# Descriptive Statistics

The descriptive statistics provide insights into the distribution and characteristics of the variables Unit Sold, Revenue, Cost, and Profit. For Unit Sold, the mean value is 1608.32 units, with a standard error of 32.79 units. The median value is 1542.5 units, indicating the central tendency of the data, and the mode is 727 units, representing the most frequently occurring value. The standard deviation, skewness, and kurtosis values indicate the dispersion, symmetry, and shape of the distribution, respectively. Similarly, for Revenue, Cost, and Profit, the descriptive statistics provide measures of central tendency, variability, and distributional characteristics.

| ***Unit Sold*** | | ***Revenue*** | | ***Cost*** | | ***Profit*** | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| **Mean** | 1608.32 | Mean | 6700.456 | Mean | 2752.792 | Mean | 3947.664 |
| **Standard Error** | 32.78652 | Standard Error | 174.767 | Standard Error | 76.99166 | Standard Error | 98.86874 |
| **Median** | 1542.5 | Median | 5871.5 | Median | 2423.6 | Median | 3424.5 |
| **Mode** | 727 | Mode | 8715 | Mode | 3450 | Mode | 5229 |
| **Standard Deviation** | 867.4498 | Standard Deviation | 4623.901 | Standard Deviation | 2037.008 | Standard Deviation | 2615.821 |
| **Sample Variance** | 752469.1 | Sample Variance | 21380458 | Sample Variance | 4149401 | Sample Variance | 6842519 |
| **Kurtosis** | -0.31491 | Kurtosis | 0.464596 | Kurtosis | 0.810043 | Kurtosis | 0.338621 |
| **Skewness** | 0.43627 | Skewness | 0.867861 | Skewness | 0.930442 | Skewness | 0.840484 |
| **Range** | 4293 | Range | 23788 | Range | 10954.5 | Range | 13319 |
| **Minimum** | 200 | Minimum | 200 | Minimum | 40 | Minimum | 160 |
| **Maximum** | 4493 | Maximum | 23988 | Maximum | 10994.5 | Maximum | 13479 |
| **Sum** | 1125824 | Sum | 4690319 | Sum | 1926955 | Sum | 2763364 |
| **Count** | 700 | Count | 700 | Count | 700 | Count | 700 |

# Correlation

For Unit Sold and Revenue, the correlation coefficient is approximately 0.796, indicating a moderately strong positive correlation. Similarly, Unit Sold and Profit exhibit a correlation coefficient of approximately 0.829, indicating a moderately strong positive relationship. Revenue and Cost demonstrate a correlation coefficient of around 0.992, signifying a strong positive correlation. Additionally, Revenue and Profit show a correlation coefficient of approximately 0.995, indicating a very strong positive relationship.

| | ***Unit Sold*** | ***Revenue*** | ***Cost*** | ***Profit*** |
|---|---|---|---|---|
| **Unit Sold** | 1 | | | |
| **Revenue** | 0.796298 | 1 | | |
| **Cost** | 0.742604 | 0.992011 | 1 | |
| **Profit** | 0.829304 | 0.995163 | 0.974818 | 1 |

# Order Data Report

## INTRODUCTION

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

Key Attributes:

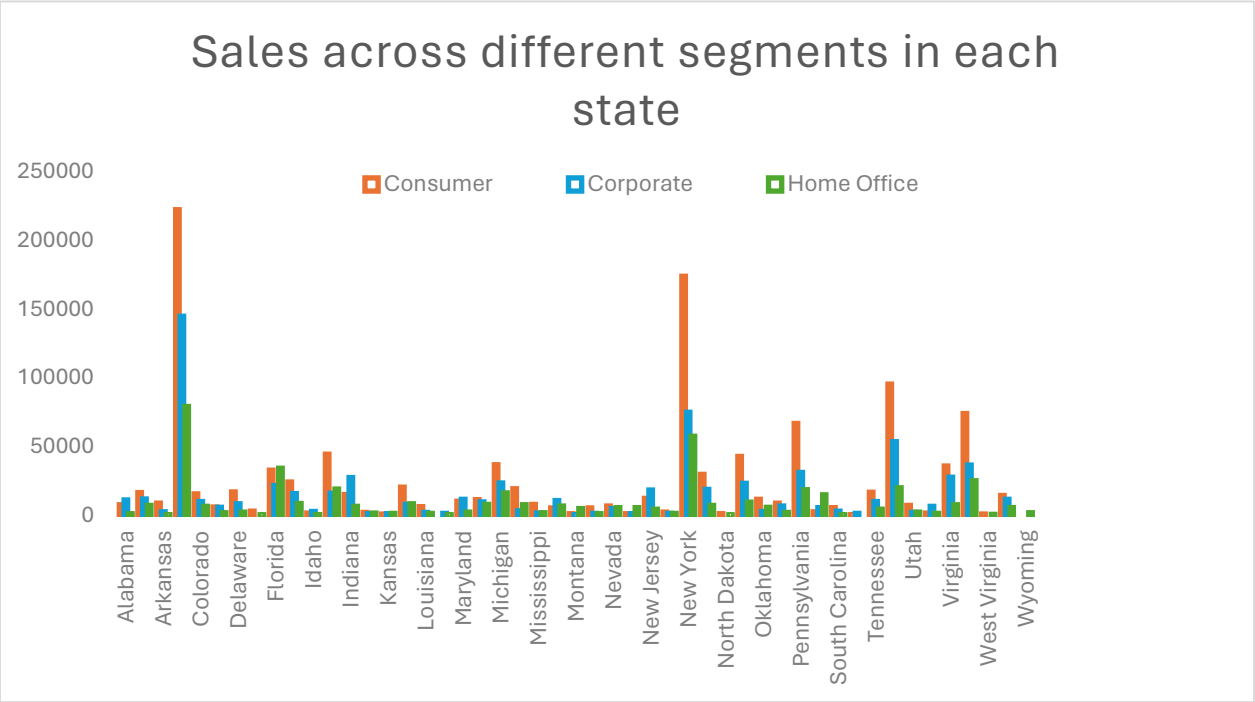
1. ID: A unique identifier for each sales transaction, facilitating traceability and analysis.
2. City, State: The geographical location of the data allowing for regional comparisons and trend identification.
3. Product Line (furniture, Electronic Accessories, appliances, Home and Lifestyle): Categorization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

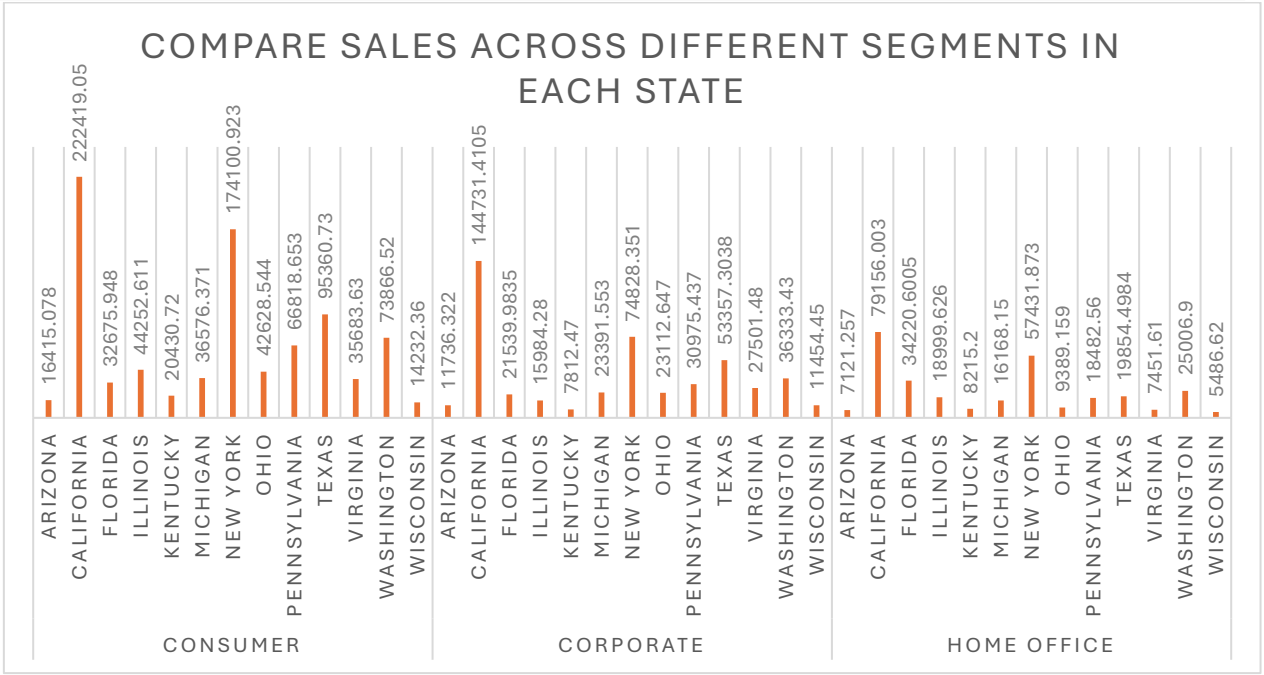
## QUESTIONNAIRE

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.

# ANALYTICS

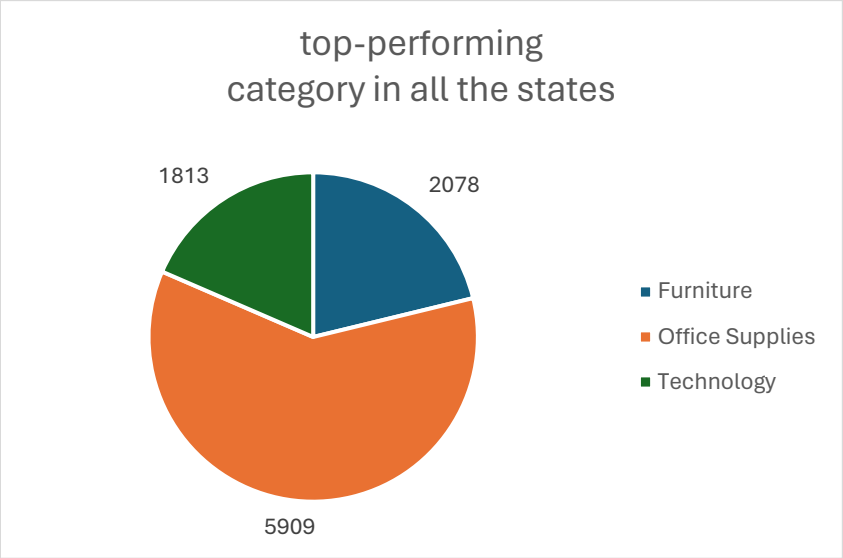
1.Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?



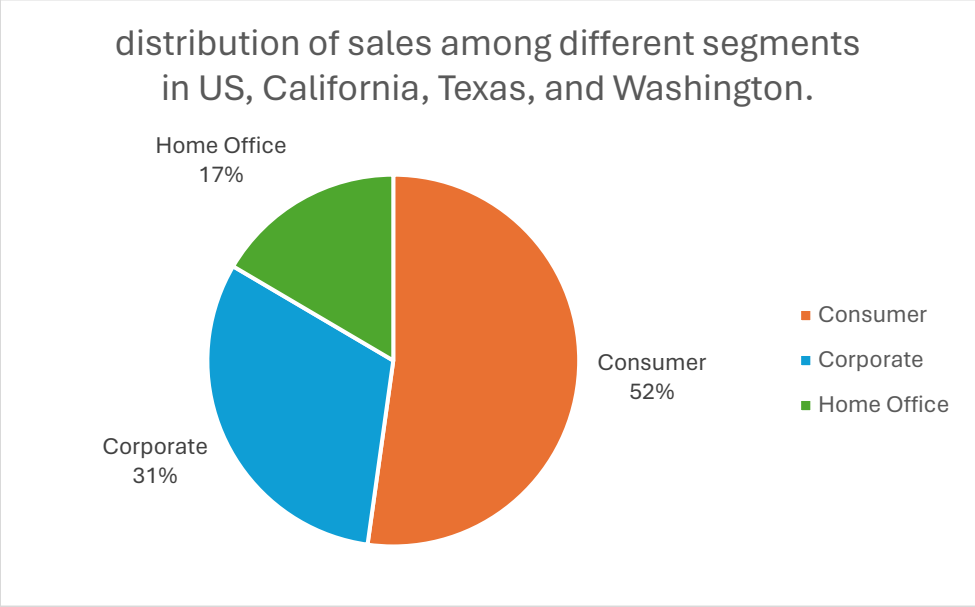

After comparing all the states in terms of segment and sales , California(222419.05) emerged as the state with the highest amount of sales. Consumer(1148060.531) segment performed well in all the states.

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



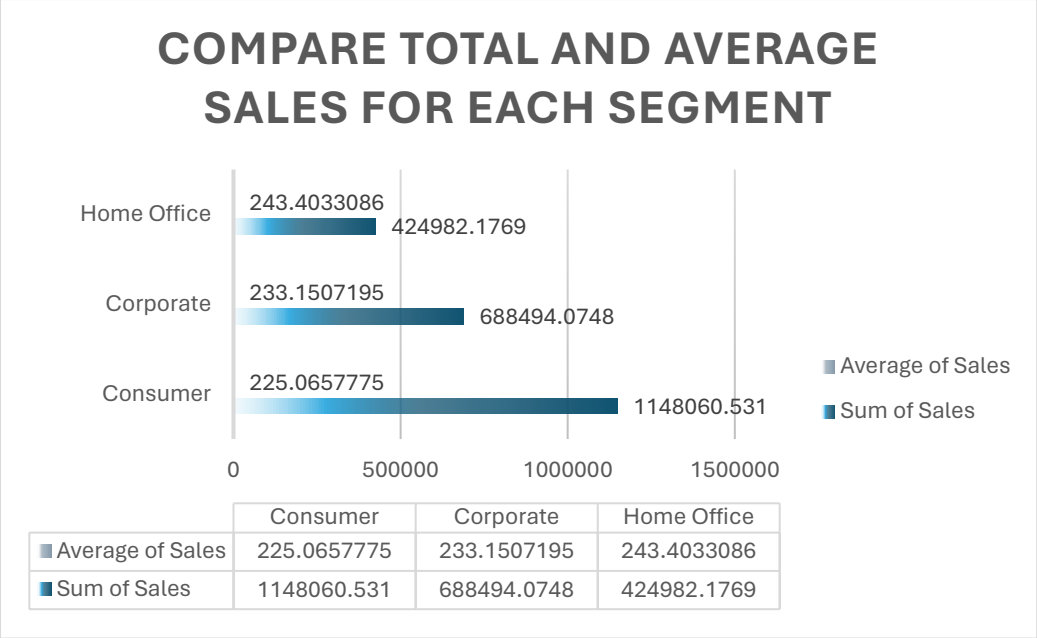

Office Supplies is the top performing category in all the states with total count of sales of 5909 followed by furniture(2078) and technology(1813).

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




Filtering the states for the total sales count and showing the percentage of distribution through pie chart Consumer segment has the most sales in US, California, Texas, and Washington

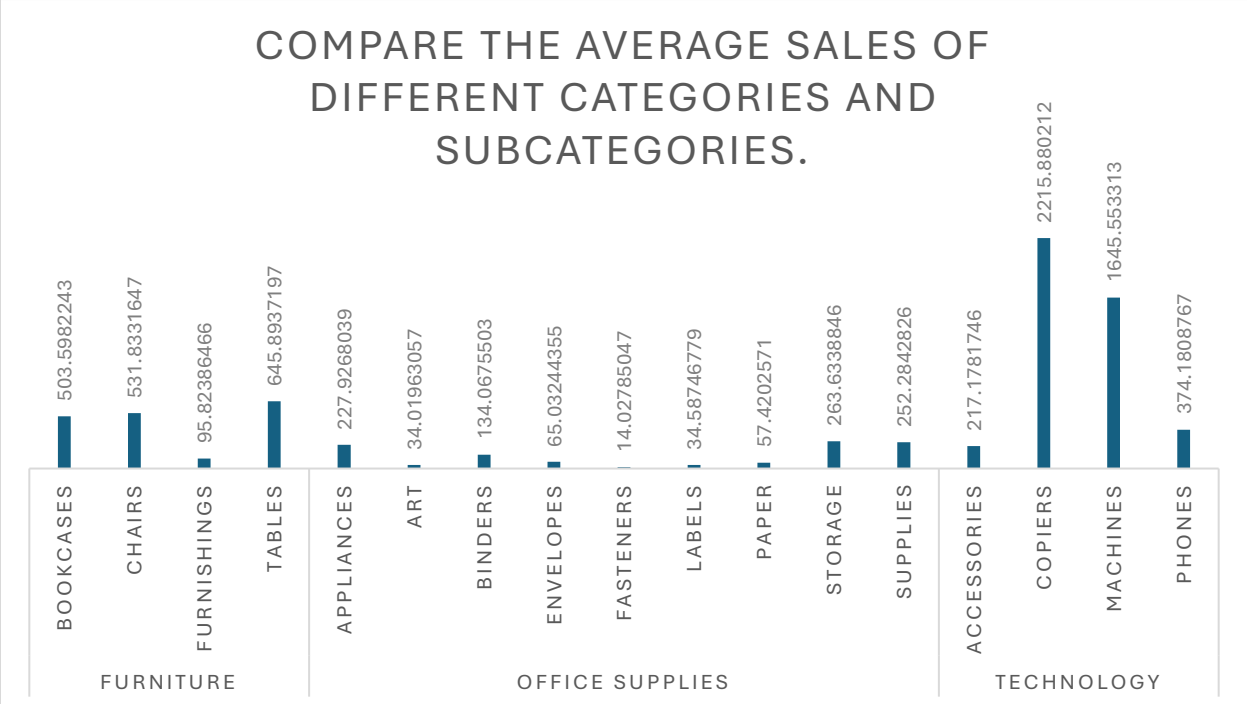
4. Compare total and average sales for all different segment?




| | Consumer | Corporate | Home Office |
|---|---|---|---|
| Average of Sales | 225.0657775 | 233.1507195 | 243.4033086 |
| Sum of Sales | 1148060.531 | 688494.0748 | 424982.1769 |

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

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




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

# Regression

In this regression analysis for the Order dataset, there is almost no relationship between Order ID and Sales, as indicated by the very low multiple R and R-squared values (0.000434 and 1.88E-07, respectively). The coefficient for Order ID is not statistically significant, with a p-value of 0.965747. This suggests that Order ID does not predict Sales. Similarly, the ANOVA test confirms the lack of significance, with an F-statistic p-value of 0.965747.

| **SUMMARY OUTPUT** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| ***Regression Statistics*** | | | | | | | | |
| **Multiple R** | 0.000434 | | | | | | | |
| **R Square** | 1.88E-07 | | | | | | | |
| **Adjusted R Square** | -0.0001 | | | | | | | |
| **Standard Error** | 625.334 | | | | | | | |
| **Observations** | 9789 | | | | | | | |
| **ANOVA** | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| **Regression** | 1 | 721.1637 | 721.1637 | 0.001844 | 0.965747 | | | |
| **Residual** | 9787 | 3.83E+09 | 391042.6 | | | | | |
| **Total** | 9788 | 3.83E+09 | | | | | | |
| | | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| **Intercept** | 230.5863 | 12.63999 | 18.24261 | 3.83E-73 | 205.8093 | 255.3633 | 205.8093 | 255.3633 |
| **X Variable 1** | -9.6E-05 | 0.002235 | -0.04294 | 0.965747 | -0.00448 | 0.004286 | -0.00448 | 0.004286 |

# Descriptive Statistics

In the Sales dataset, the mean sales amount is 230.1162, with a standard error of 6.320053. The median sales value is 54.384, while the mode is 12.96. The standard deviation is 625.3021, indicating considerable variability in sales amounts. The data is highly positively skewed, with a skewness value of 13.05363, and exhibits high kurtosis at 307.3056, indicating heavy-tailed distribution. The range of sales values spans from 0.444 to 22638.48, with a total sum of 2252607 across 9789 observations

| ***Sales*** | |
|---|---|
| | |
| **Mean** | 230.1162 |
| **Standard Error** | 6.320053 |
| **Median** | 54.384 |
| **Mode** | 12.96 |
| **Standard Deviation** | 625.3021 |
| **Sample Variance** | 391002.7 |
| **Kurtosis** | 307.3056 |
| **Skewness** | 13.05363 |
| **Range** | 22638.04 |
| **Minimum** | 0.444 |
| **Maximum** | 22638.48 |
| **Sum** | 2252607 |
| **Count** | 9789 |

# CONCLUSION

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

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

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

# Shop Sales Data

## Introduction

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

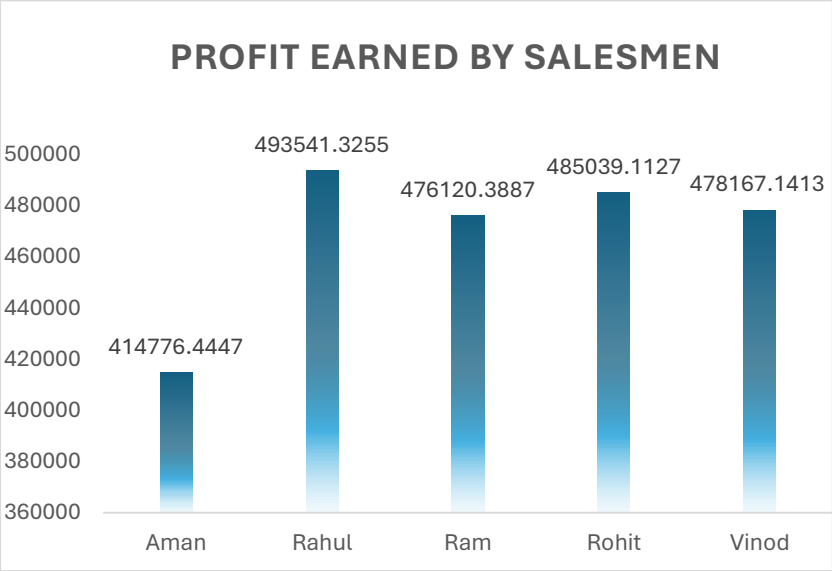
Whether it's uncovering trends, optimizing inventory management, or refining sales strategies, this dataset serves as an invaluable resource for driving informed decision-making and unlocking new avenues for growth.

## Questionaries

1. Compare all the salesmen based on 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.

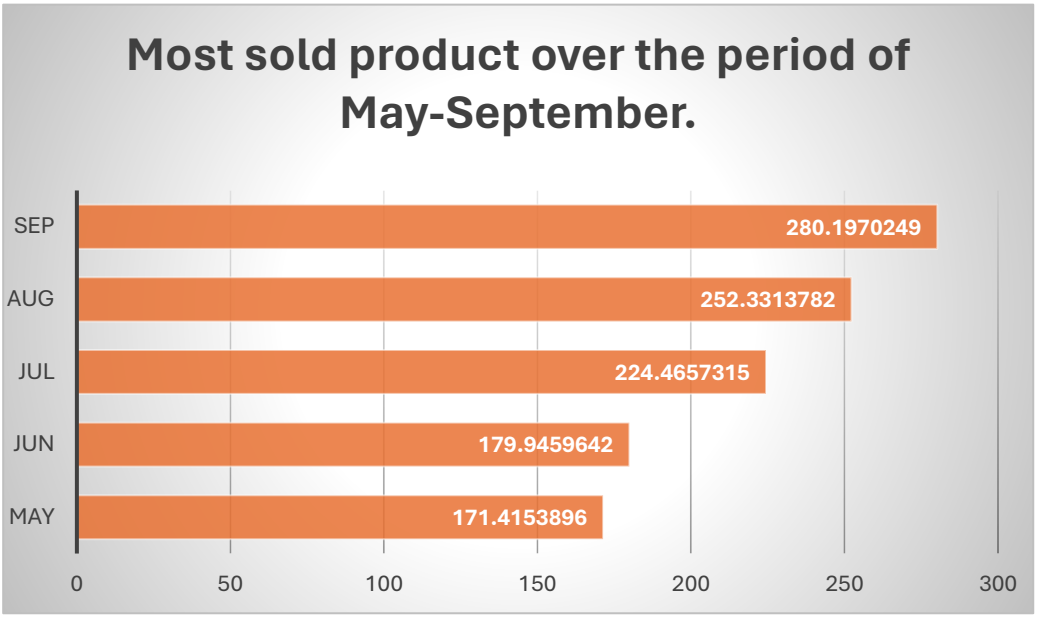
## Analytics

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




The comparison of all the salesmen on the basis of profit earned from above bargraph shows that Rahul has earned highest profit while Aman the lowest.

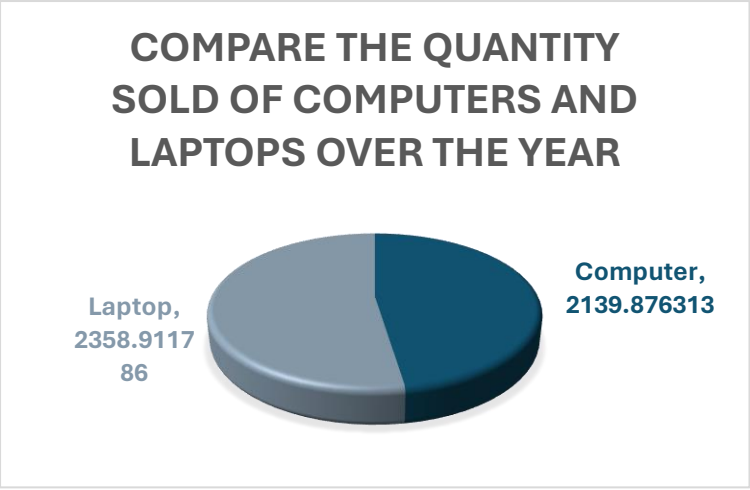
2. Find out most sold product over the period of May-September.




To ascertain the best-selling product between May and September, we must scrutinize the sales data within this time span. By consolidating the quantity sold for each product across all transactions during this period and subsequently identifying the product with the highest total quantity sold, we can pinpoint the most favored item. According to the analysis, the laptopemerges as the top-selling product during the May-September period.

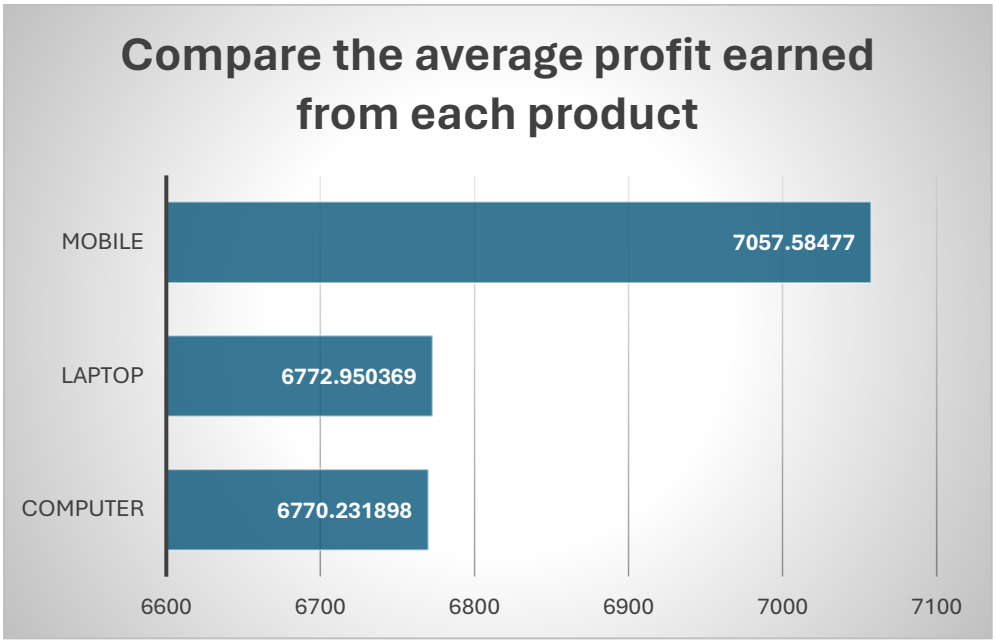
3. Find out which of the two product sold the most over the year Computer or Laptop?

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



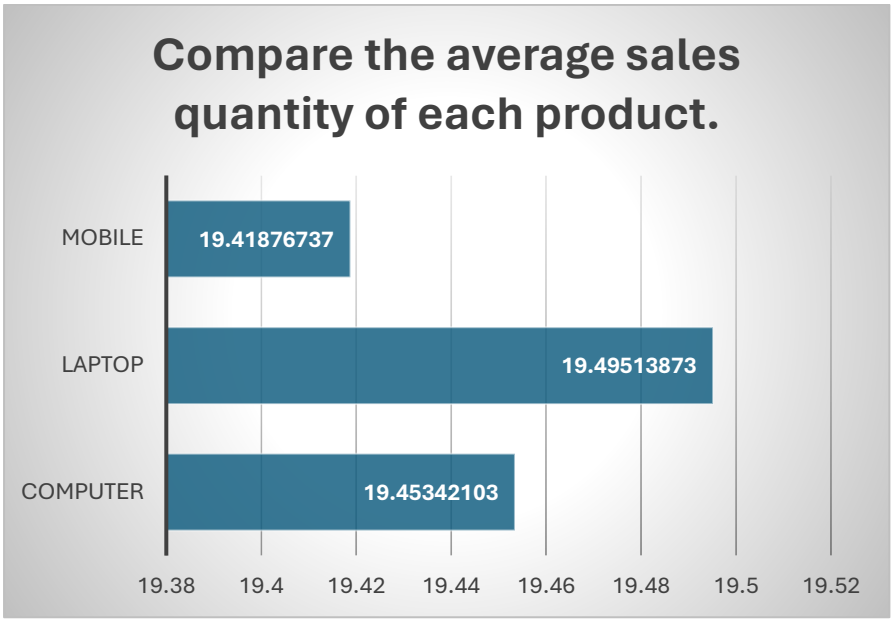

The two product sold the most over the year between computer or laptop where Computer has the sold quantity of 2140 units and laptop has 2359 units sold quantity

4 . Which item yield most average profit?




The item that yields the most profit between laptop, computer and mobile is given in above graph.

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




The average sales of all the products with their respective comparison is showed in above result of analysis where we found that laptop has the highest sales over quantity.

# Conclusion and Review:

The analysis reveals significant insights into sales performance and product trends among salesmen. Rahul emerges as the top performer, earning the highest profit compared to all other salesmen. Moreover, the most sold product over the period of May-September is identified as the Laptop, with the highest sales recorded in September. Between computers and laptops, laptops outperform computers in terms of units sold throughout the year. Additionally, mobile phones exhibit the highest average profit among mobiles, laptops, and computers. Lastly, laptops demonstrate the highest average sales quantity compared to mobiles and computers.

# Regression:

This regression analysis demonstrates a robust relationship between the quantity of items sold (X Variable 1) and the corresponding sales amount (Y). With a high R Square value of approximately 0.91, it indicates that about 91% of the variability in sales amount can be accounted for by changes in the quantity of items sold. For each additional unit increase in the quantity sold, there's an average increase of approximately $246.47 in sales amount. Both the intercept and the coefficient of the X Variable 1 are statistically significant, with t-stats of 23.38 and 58.73, respectively, and very low p-values (nearly zero), affirming the reliability of these coefficients. Thus, we conclude that the quantity of items sold serves as a strong predictor of sales amount in this dataset.

| **SUMMARY OUTPUT** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ***Regression Statistics*** | | | | | | | | |
| **Multiple R** | 0.954076972 | | | | | | | |
| **R Square** | 0.910262868 | | | | | | | |
| **Adjusted R Square** | 0.909998936 | | | | | | | |
| **Standard Error** | 630.0595983 | | | | | | | |
| **Observations** | 342 | | | | | | | |
| **ANOVA** | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| **Regression** | 1 | 1.37E+09 | 1.37E+09 | 3448.844 | 4.6E-180 | | | |
| **Residual** | 340 | 1.35E+08 | 396975.1 | | | | | |
| **Total** | 341 | 1.5E+09 | | | | | | |
| | | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| **Intercept** | 2068.993161 | 88.47952 | 23.38387 | 9.14E-73 | 1894.957 | 2243.029 | 1894.957 | 2243.029 |

# Anova : Single Factor

The single-factor ANOVA conducted on the quantity (Qty) and sales amount (Amount) reveals a significant difference between the groups. The analysis indicates that there's a substantial variance between the groups (SS = $8.01E+09) compared to within the groups (SS = $1.5E+09), resulting in a high F-statistic of 3632.879 with a very low p-value (nearly zero). This implies that the difference in means between the quantity and sales amount is unlikely to have occurred by chance. Therefore, we reject the null hypothesis and conclude that there's a significant difference in sales amounts attributed to different quantities sold.

| **Anova: Single Factor** | | | | | | |
|---|---|---|---|---|---|---|
| **SUMMARY** | | | | | | |
| ***Groups*** | *Count* | *Sum* | *Average* | *Variance* | | |
| **Qty** | 342 | 6654.271 | 19.45693 | 66.0952 | | |
| **Amount** | 342 | 2347644 | 6864.457 | 4410782 | | |
| | | | | | | |
| **ANOVA** | | | | | | |
| ***Source of Variation*** | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| **Between Groups** | 8.01E+09 | 1 | 8.01E+09 | 3632.879 | 2.1E-275 | 3.85513 |
| **Within Groups** | 1.5E+09 | 682 | 2205424 | | | |
| | | | | | | |
| **Total** | 9.52E+09 | 683 | | | | |

# Anova two factor

In the two-factor ANOVA analysis without replication, we observe both rows and columns contribute significantly to the variance. The sums of squares (SS) for rows and columns are $7.58E+08 and $8.01E+09, respectively. The high F-statistics for both rows (1.014883) and columns (3659.913) with low p-values (nearly zero) indicate that the differences observed in both factors are statistically significant. Therefore, we reject the null hypothesis and conclude that both the quantity sold (Qty) and sales amount (Amount) significantly affect the variance in the dataset.

| **Anova: Two-Factor Without Replication** | | | | |
|---|---|---|---|---|
| | | | | |
| ***SUMMARY*** | *Count* | *Sum* | *Average* | *Variance* |
| **Row 1** | 2 | 1003 | 501.5 | 497004.5 |
| **Row 2** | 2 | 7804 | 3902 | 30388808 |
| **Row 3** | 2 | 3005 | 1502.5 | 4485013 |
| **Row 4** | 2 | 2304 | 1152 | 2635808 |
| **Row 5** | 2 | 7003 | 3501.5 | 24479005 |
| **Row 339** | 2 | 10252.82 | 5126.411 | 51884342 |
| **Row 340** | 2 | 10272.93 | 5136.467 | 52087770 |
| **Row 341** | 2 | 10293.05 | 5146.523 | 52291595 |
| **Row 342** | 2 | 10313.16 | 5156.58 | 52495819 |
| | | | | |
| **Qty** | 342 | 6654.271 | 19.45693 | 66.0952 |

| | | | | |
|---|---|---|---|---|
| **Amount** | 342 | 2347644 | 6864.457 | 4410782 |

| **ANOVA** | | | | | | |
|---|---|---|---|---|---|---|
| ***Source of Variation*** | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| **Rows** | 7.58E+08 | 341 | 2221714 | 1.014883 | 0.445792 | 1.195299 |
| **Columns** | 8.01E+09 | 1 | 8.01E+09 | 3659.913 | 2.1E-184 | 3.868873 |
| **Error** | 7.46E+08 | 341 | 2189134 | | | |
| | | | | | | |
| **Total** | 9.52E+09 | 683 | | | | |

## Descriptive Statistics:

For the quantity sold (Qty), the mean is approximately 19.46 with a standard error of 0.44. The data is moderately positively skewed (skewness = -0.10) and shows a slight negative kurtosis (-0.999), indicating a slightly flatter distribution compared to a normal distribution. The range of quantity sold spans from 3 to 33.31.

For the sales amount (Amount), the mean is approximately 6864.46 with a larger standard error of 113.57. The data is also moderately positively skewed (skewness = -0.36) and has a slightly negative kurtosis (-0.508). The range of sales amount is much larger, ranging from 1000 to 10279.85.

| ***Qty*** | | ***Amount*** | |
|---|---|---|---|
| Mean | 19.45693 | Mean | 6864.457 |
| Standard Error | 0.439614 | Standard Error | 113.5651 |
| Median | 19.45693 | Median | 6984.647 |
| Mode | 3 | Mode | 1000 |
| Standard Deviation | 8.129896 | Standard Deviation | 2100.186 |
| Sample Variance | 66.0952 | Sample Variance | 4410782 |
| Kurtosis | -0.99883 | Kurtosis | -0.5078 |
| Skewness | -0.09948 | Skewness | -0.36449 |
| Range | 30.30852 | Range | 9279.851 |
| Minimum | 3 | Minimum | 1000 |
| Maximum | 33.30852 | Maximum | 10279.85 |
| Sum | 6654.271 | Sum | 2347644 |
| Count | 342 | Count | 342 |

## Correlation

The correlation coefficient between quantity sold (Qty) and sales amount (Amount) is approximately 0.954. This strong positive correlation suggests that there is a significant relationship between the quantity of items sold and the corresponding sales amount, indicating that as the quantity sold increases, the sales amount also tends to increase.

| | ***Qty*** | ***Amount*** |
|---|---|---|
| **Qty** | 1 | |
| **Amount** | 0.954077 | 1 |

# Store Dataset Report

## Introduction

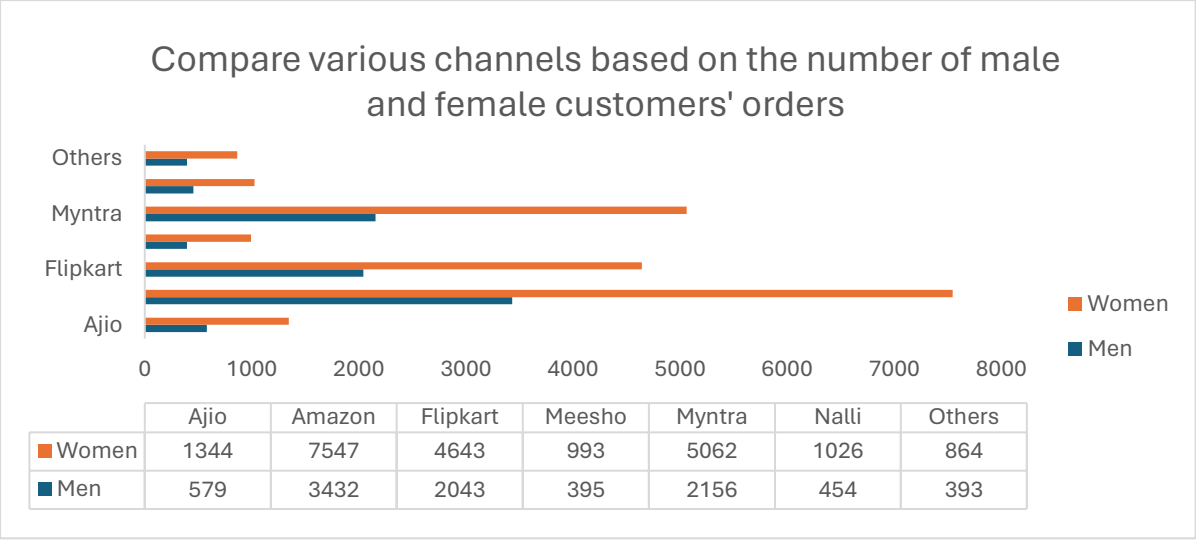
This dataset encompasses sales data from a retail store, encompassing a multitude of attributes including customer demographics (such as Gender and Age Group), transaction details (like Order ID and Status), product specifics (including Category and SKU), and shipping information. Our analysis is focused on illuminating customer behavior and product trends, with the aim of unveiling patterns, preferences, and correlations embedded within the data. By leveraging these insights, businesses can fine-tune their marketing strategies, optimize inventory management processes, and enhance overall customer satisfaction levels.

## Questionnaire

1. Compare various channels based on how many male customers order and female customer order.
2. Compare all the categories of order where amount is less than 1500 and greater than 5000.
3. How many Customers are there whose age is 30 and above and state is Delhi.
4. Which of the following state perform better than other, Delhi, Tamil Nadu, Maharashtra, Rajasthan.
5. Which city performed better than all other cities based on highest order placed.
6. Compare various categories of items based on most quantity sold and show which gender buys the most category.

## Analytics

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

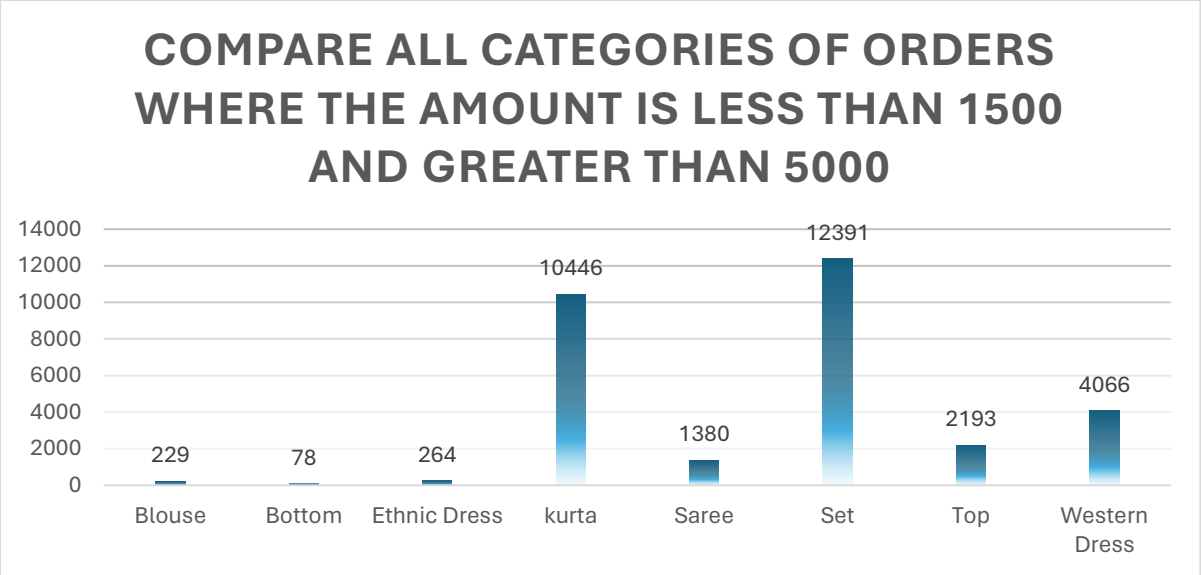



| | Ajio | Amazon | Flipkart | Meesho | Myntra | Nalli | Others |
|---|---|---|---|---|---|---|---|
| Women | 1344 | 7547 | 4643 | 993 | 5062 | 1026 | 864 |
| Men | 579 | 3432 | 2043 | 395 | 2156 | 454 | 393 |

Amazon dominates sales in both the men's and women's categories, with Myntra and Flipkart following closely behind. In the men's category, Amazon sold approximately 3432 units, while in the women's category, it sold nearly 7547 units. Myntra, on the other hand, sold 2156 units

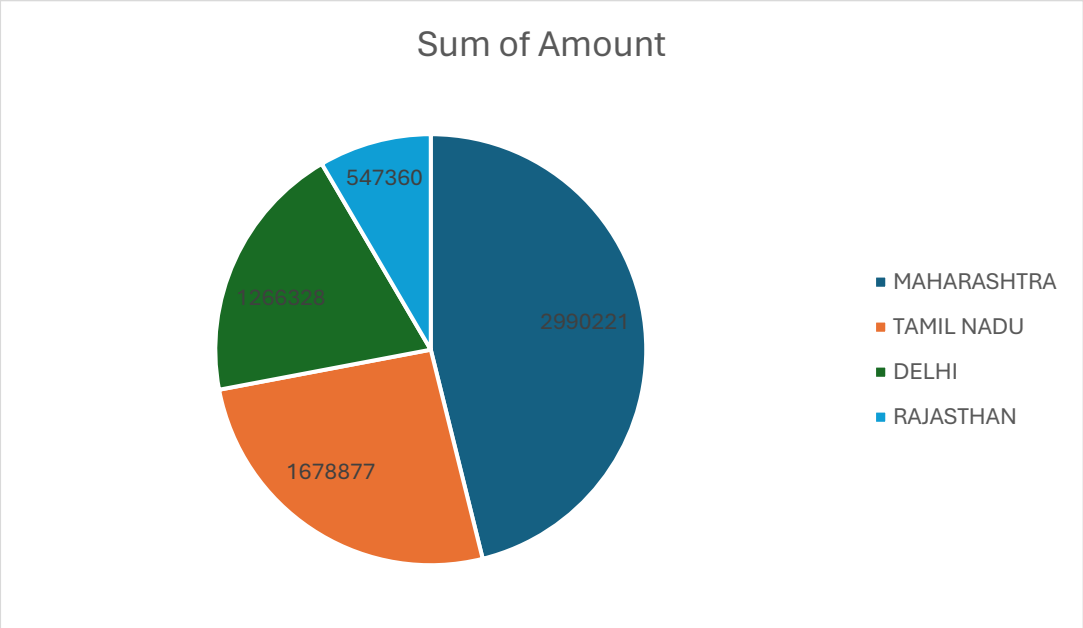
in the men's section and 5062 units in the women's section.

2. Compare all the categories of order where amount is less than 1500 and greater than 5000.



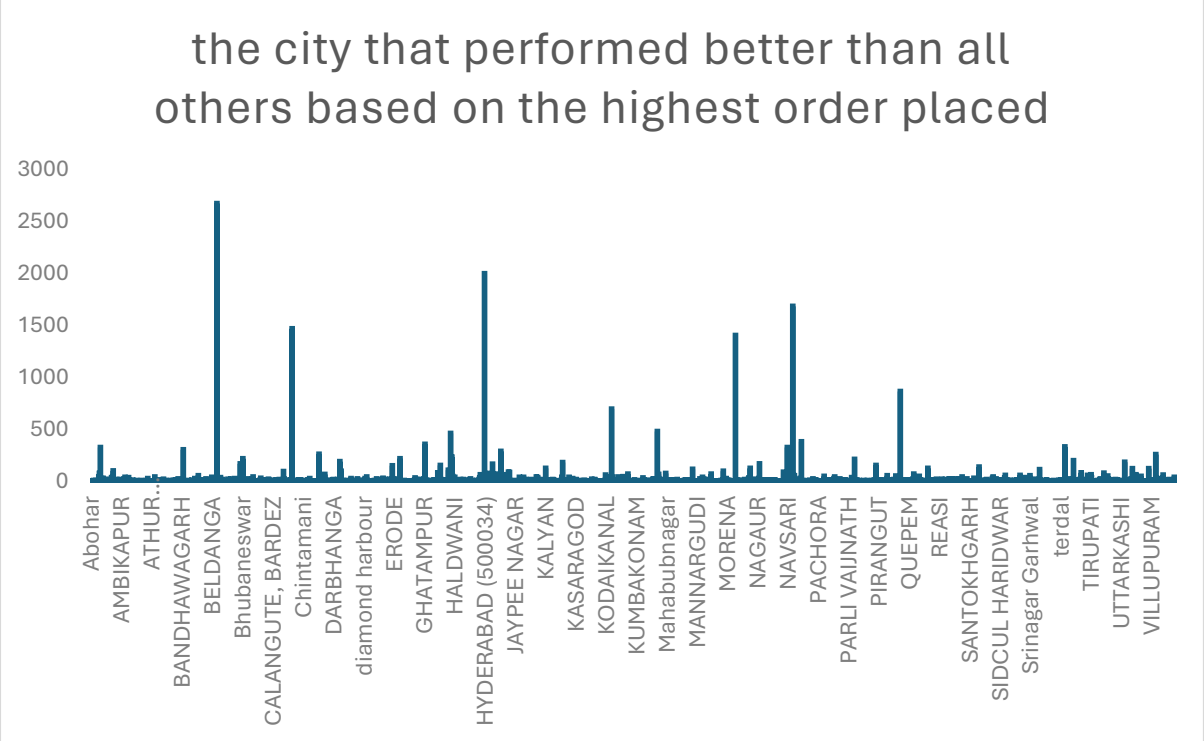

This analysis helps in comparing the categories of order where amount is less than 1500 and greater than 5000. Showing the kurta(12391) and set(10446) with highest count of the orders followed by western dress, top and saree.

4. Which of the following state perform better than other, Delhi, Tamil Nadu, Maharashtra, Rajasthan.



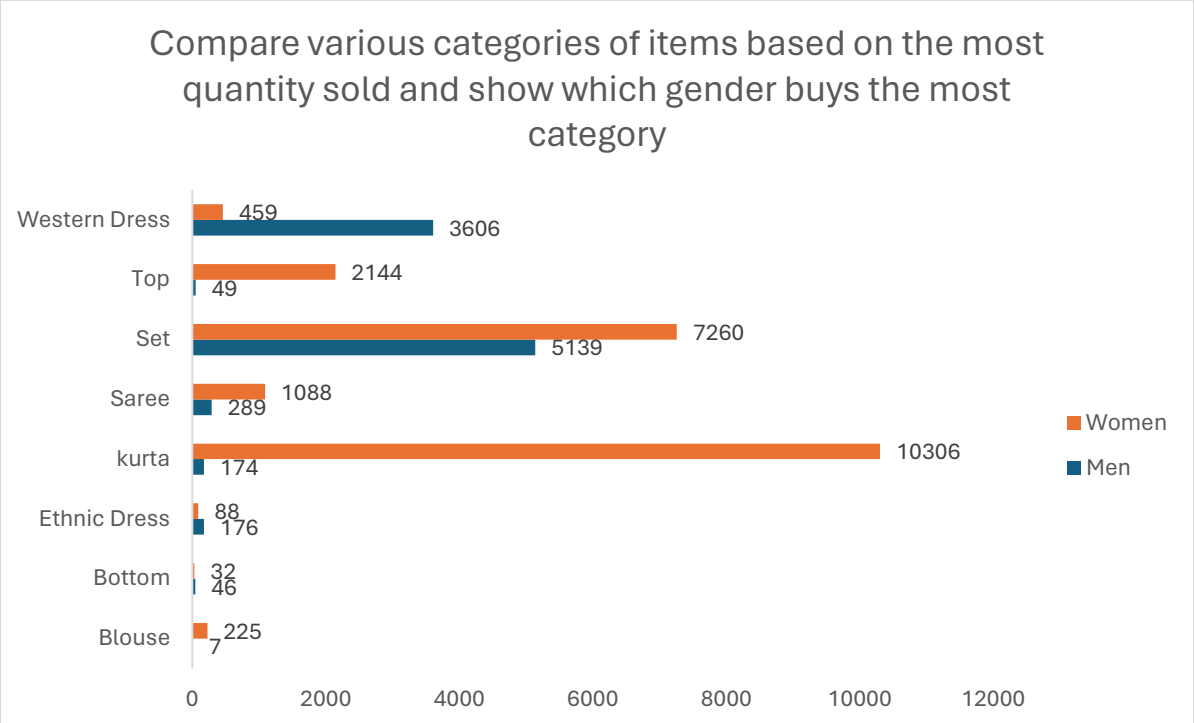

This analysis shows which states performed better in terms of total amount generated where we found that the state of Maharashtra performed the best while rajasthan remained at last.

5. Which city performed better than all other cities based on highest order placed.




Based on the graph recorded we can actually see which city performed better than all other cities based on highest order placed, so according to graph Bangluru has the highest order placed with 2673 orders followed by Hyderabad(1998).

6. Compare various categories of items based on most quantity sold and also show which gender buys the most category.




This analysis shows the comparison of various categories of items based on most quantity sold which is kurta bought by women set bought by women followed by men and western dress followed by top for both men and women.

# Conclusion and Review

The analysis highlights Amazon's dominance in sales across both men's and women's categories, with Myntra and Flipkart following closely behind. Amazon leads in sales for both men's and women's categories, followed by Myntra and Flipkart. Top-selling items include kurta and set, with Karnataka and Bangalore showing the highest sales performance.

The analysis provides valuable insights into sales trends and regional performance, aiding decision-making for retailers. However, further exploration into additional factors influencing sales could enhance the analysis. Overall, the findings offer valuable information for optimizing sales strategies in competitive markets.

# Regression

The regression analysis for the store dataset indicates that the model has a multiple ( R ) value of approximately 0.172, suggesting a weak positive correlation between the independent variables (quantity and size) and the dependent variable (amount). The ( R^2 ) value, which measures the proportion of the variance in the dependent variable explained by the independent variables, is approximately 0.030. This suggests that only about 3% of the variability in the amount can be explained by the quantity and size.

In terms of significance, the ANOVA results show that the regression model is statistically significant, as indicated by the very low p-value of 0 for the regression. However, when looking at the coefficients, it's observed that the coefficient for X Variable 1 (quantity) is not statistically significant, with a p-value of 0.632. On the other hand, the coefficient for X Variable 2 (size) is highly significant, with a very low p-value (approximately 1.3 * 10^{-205}), indicating that size has a substantial impact on the amount.

| **SUMMARY OUTPUT** | |
|---|---|
| ***Regression Statistics*** | |
| **Multiple R** | 0.172398 |
| **R Square** | 0.029721 |
| **Adjusted R Square** | 0.029659 |
| **Standard Error** | 264.5693 |
| **Observations** | 31047 |

| **ANOVA** | | | | | |
|---|---|---|---|---|---|
| | *df* | *SS* | *MS* | *F* | *Significance F* |
| **Regression** | 2 | 66561870 | 33280935 | 475.4629 | 0 |
| **Residual** | 31044 | 2.17E+09 | 69996.92 | | |
| **Total** | 31046 | 2.24E+09 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* |
|---|---|---|---|---|---|
| **Intercept** | 185.155 | 16.57854 | 11.16836 | 6.61E-29 | 152.6604 |
| **X Variable 1** | 0.047626 | 0.099327 | 0.479489 | 0.631594 | -0.14706 |
| **X Variable 2** | 492.0276 | 15.95904 | 30.83065 | 1.3E-205 | 460.7472 |

# Anova-1 factor

The single-factor ANOVA test conducted on the Qty and Amount groups reveals a highly significant result. The between-groups variance, which measures the variability between the Qty and Amount groups, is extremely large (SS = 7.2 * 10^9), resulting in a very high F-statistic (F = 199639.8) and an associated p-value close to zero (p < 0.001). This indicates that there is a significant difference between the Qty and Amount groups concerning their means. The within-groups varian.ce, representing the variability within each group, is also considerable (SS =2.24 * 10^9), reflecting the dispersion of data points around their respective group means. Overall, the ANOVA test suggests a strong statistical significance in the difference between Qty and Amount groups.

| **Anova: Single Factor** | | | | | | |
|---|---|---|---|---|---|---|
| **SUMMARY** | | | | | | |
| ***Groups*** | *Count* | *Sum* | *Average* | *Variance* | | |
| **Qty** | 31047 | 31237 | 1.00612 | 0.008853 | | |
| **Amount** | 31047 | 21176377 | 682.0748 | 72136.38 | | |
| **ANOVA** | | | | | | |
| ***Source of Variation*** | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| **Between Groups** | 7.2E+09 | 1 | 7.2E+09 | 199639.8 | 0 | 3.841609 |
| **Within Groups** | 2.24E+09 | 62092 | 36068.2 | | | |
| | | | | | | |
| **Total** | 9.44E+09 | 62093 | | | | |

# Anova- 2 factor

The two-factor ANOVA conducted on Age, Qty, and Amount reveals interesting insights. Regarding rows, the variability in the data across different groups (SS = 7.49 * 10^8) doesn't show a significant difference between them (p = 0.468). However, the variability between columns is substantial (SS =9.09*10^9), indicating a significant difference among the factors Age, Qty, and Amount (p < 0.001). The error term, representing variability within groups, is also noteworthy (SS = 1.5*10^9), showing dispersion within each combination of factors. Overall, the ANOVA results suggest a statistically significant difference between the factors Qty and Amount concerning their means, but no significant difference across age groups.

| **Anova: Two-Factor Without Replication** | | | | |
|---|---|---|---|---|
| | | | | |
| ***SUMMARY*** | *Count* | *Sum* | *Average* | *Variance* |
| **Row 1** | 3 | 421 | 140.3333 | 42116.33 |
| **Row 2** | 3 | 1479 | 493 | 685648 |
| **Row 3** | 3 | 521 | 173.6667 | 59609.33 |
| **Row 4** | 3 | 750 | 250 | 172171 |
| **Row 5** | 3 | 607 | 202.3333 | 88482.33 |
| **Row 31044** | 3 | 974 | 324.6667 | 283326.3 |
| **Row 31045** | 3 | 1145 | 381.6667 | 403529.3 |
| **Row 31046** | 3 | 446 | 148.6667 | 47506.33 |
| **Row 31047** | 3 | 828 | 276 | 199225 |
| | | | | |

| | | | | |
|---|---|---|---|---|
| **Age** | 31047 | 1226250 | 39.49657 | 228.5307 |
| **Qty** | 31047 | 31237 | 1.00612 | 0.008853 |
| **Amount** | 31047 | 21176377 | 682.0748 | 72136.38 |

**ANOVA**

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| **Rows** | 7.49E+08 | 31046 | 24134.08 | 1.000774 | 0.468198 | 1.016275 |
| **Columns** | 9.09E+09 | 2 | 4.54E+09 | 188446.6 | 0 | 2.995877 |
| **Error** | 1.5E+09 | 62092 | 24115.42 | | | |
| | | | | | | |
| **Total** | 1.13E+10 | 93140 | | | | |

# Descriptive Statistics

The mean age is approximately 39.50 years, with a standard deviation of 15.12. The data's distribution is slightly skewed to the right (skewness = 0.73), and its kurtosis indicates a relatively normal distribution (kurtosis = -0.16). On average, the quantity ordered is about 1.01, with a small standard deviation of 0.09. The mode is 1, indicating that this value appears most frequently in the dataset. However, the data is heavily right-skewed (skewness = 19.45) and exhibits high kurtosis (kurtosis = 475.36), suggesting a heavily tailed distribution. The average amount is approximately 682.07, with a considerable standard deviation of 268.58. The data is moderately skewed to the right (skewness = 1.05) and has a slightly heavier tail (kurtosis = 1.77). The range of values spans from 229 to 3036.

| ***Age*** | | ***Qty*** | | ***Amount*** | |
|---|---|---|---|---|---|
| Mean | 39.49657 | Mean | 1.00612 | Mean | 682.0748 |
| Standard Error | 0.085795 | Standard Error | 0.000534 | Standard Error | 1.524289 |
| Median | 37 | Median | 1 | Median | 646 |
| Mode | 28 | Mode | 1 | Mode | 399 |
| Standard Deviation | 15.11723 | Standard Deviation | 0.094088 | Standard Deviation | 268.5822 |
| Sample Variance | 228.5307 | Sample Variance | 0.008853 | Sample Variance | 72136.38 |
| Kurtosis | -0.1587 | Kurtosis | 475.3566 | Kurtosis | 1.768676 |
| Skewness | 0.72916 | Skewness | 19.4509 | Skewness | 1.052904 |
| Range | 60 | Range | 4 | Range | 2807 |
| Minimum | 18 | Minimum | 1 | Minimum | 229 |
| Maximum | 78 | Maximum | 5 | Maximum | 3036 |
| Sum | 1226250 | Sum | 31237 | Sum | 21176377 |
| Count | 31047 | Count | 31047 | Count | 31047 |

# Correlation

The correlation matrix reveals the relationships between Age, Qty (quantity), and Amount variables. Firstly, Age shows an almost negligible positive correlation with Qty, with a correlation coefficient of around 0.0049, indicating an extremely weak association. Similarly, the correlation between Age and Amount is also very weak, standing at approximately 0.0035. Conversely, there appears to be a slightly stronger positive correlation, though still weak, between Qty and Amount, with a correlation coefficient of about 0.1724. This suggests that as the quantity ordered increases, there's a modest increase in the total amount. Overall, these correlation values signify subtle connections between the variables, with the quantity ordered having the most notable influence on the total amount compared to Age.

| | ***Age*** | ***Qty*** | ***Amount*** |
|---|---|---|---|
| **Age** | 1 | | |
| **Qty** | 0.004884 | 1 | |
| **Amount** | 0.003522 | 0.172377 | 1 |

# Loan Dataset Report

## Introduction

Our dataset has lots of different kinds of information that help us understand how people apply for loans. It includes basic things like whether someone is a man or a woman, if they're married, and how educated they are. But it also includes more detailed stuff like whether they have a job, how much money they want to borrow, and what kind of home they live in. All of this information together helps us see the whole picture of how loans work and who is applying for them.

## Key Attributes:

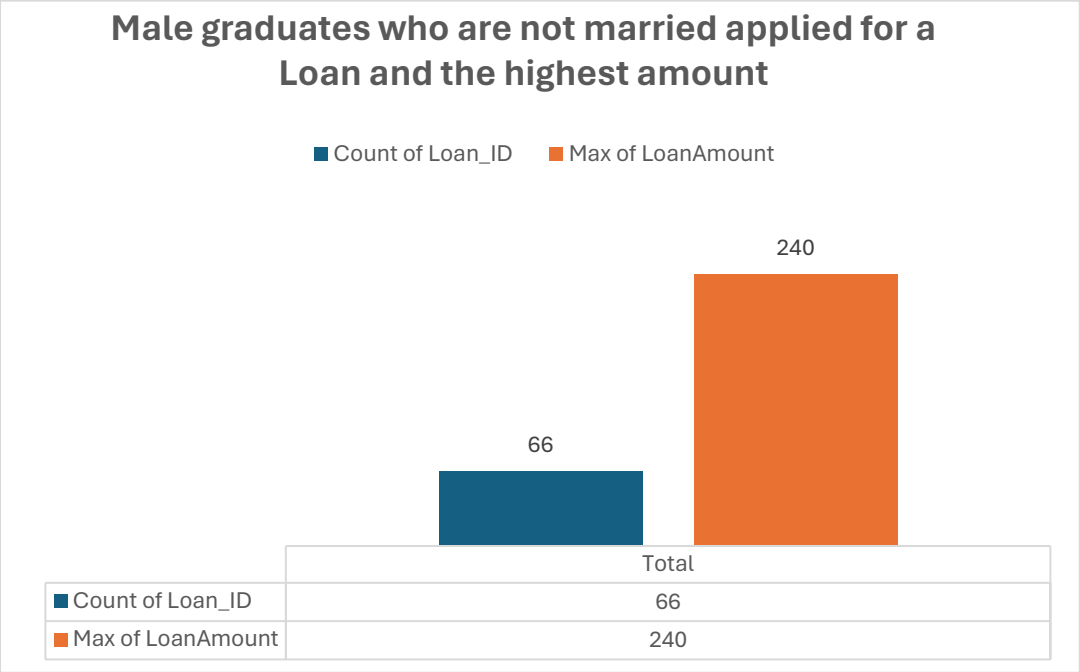
1. Gender: A demographic identifier providing insights into the gender distribution among loan applicants.
2. Marital Status (Married, Not Married): Categorization based on marital status aiding in demographic segmentation.
3. Education (Graduate, Non-graduate): Classification based on educational background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

## Questionnaire

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

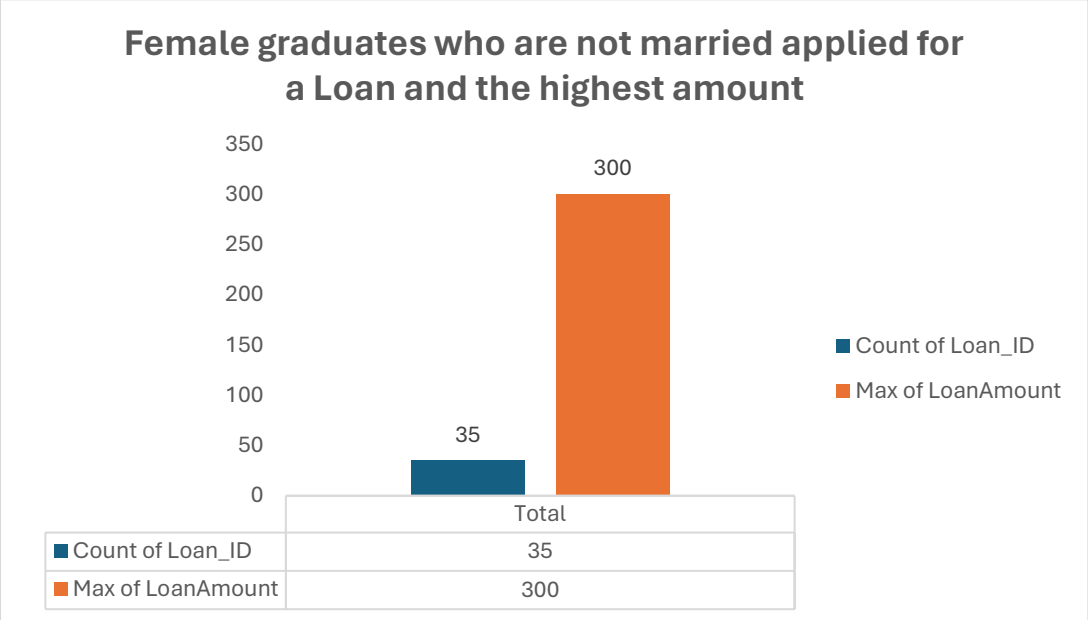
# Analytics

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



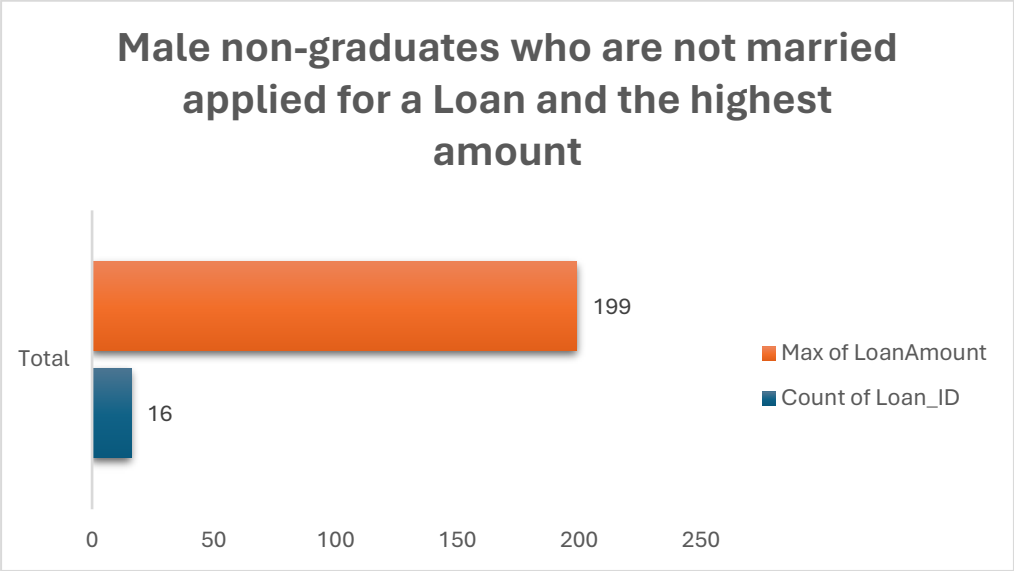

Total 66 male non graduates non married applied for loan while the highest amount carried out among them is 240k.

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



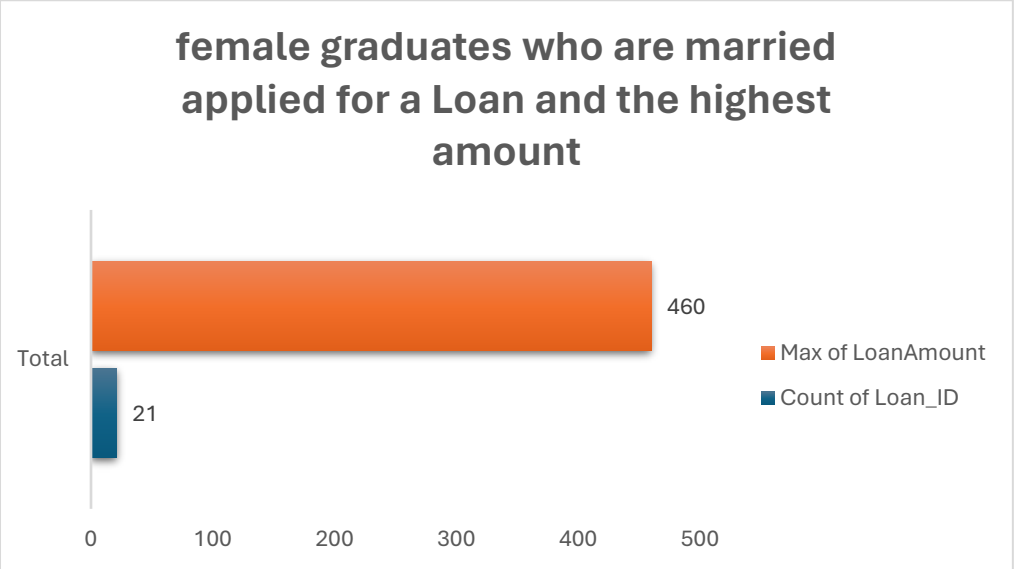

Total 35 female graduates not married applied for loan and the highest amount was noticed to be 300k

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



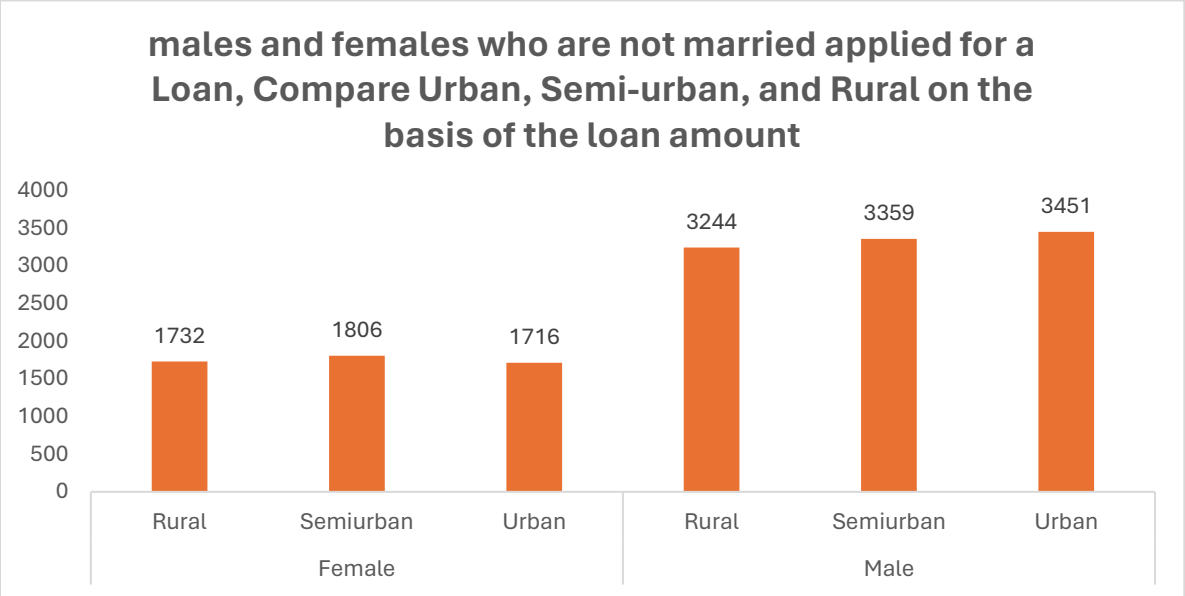

Total 16 male non-graduates who are not married applied for Loan and the highest amount applied for is 189$.

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




This analysis shows the no. of female graduates applied for the loan and are not married with the highest amount. As of analysed the total no. of loan applied is 21 and max loan amount is 460.

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




The bar gragh above clearly shows the number of male and female who are not married and applied for loan along with their region where we observed that both male and female from sumiurban has applied the most for the loan in their gender category.

# Conclusion and Review

The analysis indicates clear gender disparities in loan applications. Male graduates not married dominated the applicant pool, followed by female graduates not married. Both male non-graduates not married and married female graduates also applied for loans, albeit in smaller numbers. Notably, males significantly outnumbered females across rural, semi-urban, and urban areas.

The analysis effectively illustrates gender-based trends in loan applications and provides valuable insights into borrower demographics. Further exploration into factors influencing loan decisions is recommended, along with visual enhancements to improve data presentation. Overall, the report lays a foundation for understanding loan dynamics, with potential for deeper insights.

# Regression

The regression analysis for the loan dataset reveals a multiple R coefficient of approximately 0.531, indicating a moderate positive relationship between the predictors and the loan amount. The R-squared value of around 0.282 suggests that about 28.2% of the variability in the loan amount can be explained by the independent variables. The coefficient for Applicant Income is approximately 0.096, and for Co-applicant Income, it's about 0.0068. These coefficients signify the impact of each predictor on the loan amount. Additionally, the ANOVA table displays a significant F-value of 37.32 ($p < 0.05$), affirming the statistical significance of the regression model.

| **SUMMARY OUTPUT** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ***Regression Statistics*** | | | | | | | | |
| **Multiple R** | 0.531078663 | | | | | | | |
| **R Square** | 0.282044546 | | | | | | | |
| **Adjusted R Square** | 0.274487121 | | | | | | | |
| **Standard Error** | 50.85033905 | | | | | | | |
| **Observations** | 289 | | | | | | | |
| **ANOVA** | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| **Regression** | 3 | 289502.8035 | 96500.93 | 37.32019 | 2.25609E-20 | | | |
| **Residual** | 285 | 736940.7397 | 2585.757 | | | | | |
| **Total** | 288 | 1026443.543 | | | | | | |
| | | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| **Intercept** | 66.690952 | 16.26833015 | 4.099434 | 5.41E-05 | 34.66963005 | 98.71227396 | 34.66963 | 98.71227 |
| **X Variable 1** | 0.095771273 | 0.045649816 | 2.097955 | 0.03679 | 0.005917708 | 0.185624838 | 0.005918 | 0.185625 |
| **X Variable 2** | 0.005807787 | 0.000627861 | 9.250122 | 5.49E-18 | 0.004571955 | 0.007043619 | 0.004572 | 0.007044 |
| **X Variable 3** | 0.006772797 | 0.001264765 | 5.354983 | 1.76E-07 | 0.004283331 | 0.009262263 | 0.004283 | 0.009262 |

# Anova: one factor

The loan dataset is divided into two groups based on the factors Loan Amount and Loan Amount Term. The total sum of squares (SS) is approximately 8392703, with 2267909 within-group SS and 6124794 between-group SS. This results in a mean square (MS) of 3937.343 within groups and 6124794 between groups. The F-value of 1555.565 and the associated p-value of approximately 8.4E-166 indicate that there is a significant difference between the means of the two groups. Therefore, the factor being considered (Loan Amount vs. Loan Amount Term) has a significant impact on the loan dataset.

| **Anova: Single Factor** | | | | | | |
|---|---|---|---|---|---|---|
| **SUMMARY** | | | | | | |
| ***Groups*** | *Count* | *Sum* | *Average* | *Variance* | | |
| **Loan Amount** | 289 | 39533 | 136.7924 | 3564.04 | | |
| **Loan Amount Term** | 289 | 99032 | 342.6713 | 4310.645 | | |
| **ANOVA** | | | | | | |
| ***Source of Variation*** | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| **Between Groups** | 6124794 | 1 | 6124794 | 1555.565 | 8.4E-166 | 3.857654 |
| **Within Groups** | 2267909 | 576 | 3937.343 | | | |
| | | | | | | |
| **Total** | **8392703** | **577** | | | | |

# Anova: two factor

Analyzed based on two factors: Loan Amount and Loan Amount Term. The total sum of squares (SS) is approximately 8392703, with 1264619 SS for rows, 6124794 SS for columns, and 1003290 SS for error. The mean square (MS) for rows is 4391.038, and for columns is 6124794. The F-value for rows is 1.260472, and for columns is 1758.156, with associated p-values indicating significance ($p < 0.05$).

| **Anova: Two-Factor Without Replication** | | | | | | |
|---|---|---|---|---|---|---|
| ***SUMMARY*** | *Count* | *Sum* | *Average* | *Variance* | | |
| **Row 1** | 2 | 470 | 235 | 31250 | | |
| **Row 2** | 2 | 486 | 243 | 27378 | | |
| **Row 3** | 2 | 568 | 284 | 11552 | | |
| **Row 288** | 2 | 518 | 259 | 20402 | | |
| **Row 289** | 2 | 278 | 139 | 3362 | | |
| | | | | | | |
| **Loan Amount** | 289 | 39533 | 136.7924 | 3564.04 | | |
| **Loan Amount Term** | 289 | 99032 | 342.6713 | 4310.645 | | |
| **ANOVA** | | | | | | |
| ***Source of Variation*** | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| **Rows** | 1264619 | 288 | 4391.038 | 1.260472 | 0.024978 | 1.214301 |
| **Columns** | 6124794 | 1 | 6124794 | 1758.156 | 1.2E-124 | 3.87395 |
| **Error** | 1003290 | 288 | 3483.647 | | | |
| **Total** | 8392703 | 577 | | | | |

# Descriptive Statistics

Descriptive statistics were computed for four variables in the loan dataset: Loan Amount Term, Applicant Income, Co-Applicant Income, and Loan Amount. For Loan Amount Term, the mean is approximately 342.67 months, with a standard error of 3.86 months. The median Loan Amount Term is 360 months, with a mode of 360 months as well. The standard deviation is 65.66 months, indicating variability in loan term lengths. Applicant Income has a mean of approximately 4637.35, with a standard error of 281.80. The median and mode are 3833 and 5000, respectively. The standard deviation is high at 4790.68, suggesting significant variability in applicant incomes. Co-Applicant Income has a mean of about 1528.26, with a standard error of 139.86. The median is 879, with a mode of 0, indicating a right-skewed distribution. The standard deviation is 2377.60, highlighting variability in co-applicant incomes. Lastly, Loan Amount has a mean of 136.79, with a standard error of 3.51. The median is 126, with a mode of 150. The standard deviation is 59.70, suggesting variability in loan amounts. These statistics provide insights into the central tendency, variability, and distribution of the loan dataset variables.

| ***Loan Amount Term*** | | ***Applicant Income*** | | ***Co-Applicant Income*** | | ***Loan Amount*** | |
|---|---|---|---|---|---|---|---|
| **Mean** | 342.6713 | Mean | 4637.353 | Mean | 1528.263 | Mean | 136.7924 |
| **Standard Error** | 3.862088 | Standard Error | 281.8049 | Standard Error | 139.8588 | Standard Error | 3.51174 |
| **Median** | 360 | Median | 3833 | Median | 879 | Median | 126 |
| **Mode** | 360 | Mode | 5000 | Mode | 0 | Mode | 150 |
| **Standard Deviation** | 65.6555 | Standard Deviation | 4790.684 | Standard Deviation | 2377.599 | Standard Deviation | 59.69958 |
| **Sample Variance** | 4310.645 | Sample Variance | 22950653 | Sample Variance | 5652978 | Sample Variance | 3564.04 |
| **Kurtosis** | 8.62994 | Kurtosis | 141.612 | Kurtosis | 32.96701 | Kurtosis | 5.739804 |
| **Skewness** | -2.64147 | Skewness | 10.41123 | Skewness | 4.510775 | Skewness | 1.780616 |
| **Range** | 474 | Range | 72529 | Range | 24000 | Range | 432 |
| **Minimum** | 6 | Minimum | 0 | Minimum | 0 | Minimum | 28 |
| **Maximum** | 480 | Maximum | 72529 | Maximum | 24000 | Maximum | 460 |
| **Sum** | 99032 | Sum | 1340195 | Sum | 441668 | Sum | 39533 |
| **Count** | 289 | Count | 289 | Count | 289 | Count | 289 |

# Correlation

The correlation matrix for the loan dataset variables shows the correlation coefficients between Applicant Income, Co-Applicant Income, and Loan Amount. The correlation between Applicant Income and Co-Applicant Income is approximately -0.084, indicating a weak negative correlation between these two variables. The correlation between Applicant Income and Loan Amount is approximately 0.446, suggesting a moderate positive correlation. Similarly, the correlation between Co-Applicant Income and Loan Amount is approximately 0.230, indicating a weak positive correlation.

| | ***Applicant Income*** | ***Co-Applicant income*** | ***Loan Amount*** |
|---|---|---|---|
| **Column 1** | 1 | | |
| **Column 2** | -0.08435 | 1 | |
| **Column 3** | 0.445695 | 0.230355 | 1 |

# Sales Data Sample Report

## Introduction

This report analyzes a detailed sales dataset containing attributes like ORDERNUMBER, QUANTITYORDERED, PRICEEACH, and SALES. Its goal is to uncover insights that can guide sales strategies and improve business performance. The report targets sales managers, marketers, and executives aiming to streamline sales operations and boost revenue. Key analyses include comparing sales of Vintage cars and Classic cars, calculating average sales, identifying top-selling products, assessing profit by country for specific product lines, comparing sales over different years, and evaluating countries based on deal size. By conducting these analyses, the report aims to offer practical insights for driving sales growth and achieving better business outcomes.

The project's scope involves analyzing the extensive sales dataset to extract valuable insights useful for refining sales strategies, optimizing product offerings, and enhancing overall business performance. Analysts and researchers interested in understanding sales dynamics and market trends will find significant value in this project..

## Questionnaire

1. Comparison of sales between Vintage cars and Classic cars across all countries.

2. Determination of the average sales of all products and identification of the highest-selling product.

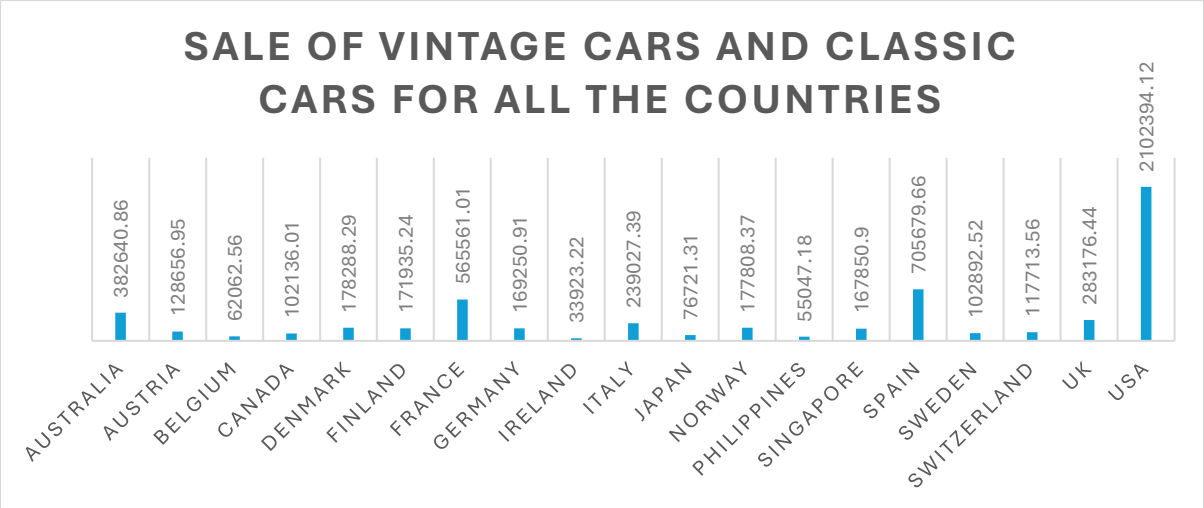
3. Assessment of the country yielding the most profit for Motorcycles, Trucks, and Buses.

4. Comparison of sales for all items across the years 2004 and 2005.

5. Comparative analysis of all countries based on deal size.

## Analytics

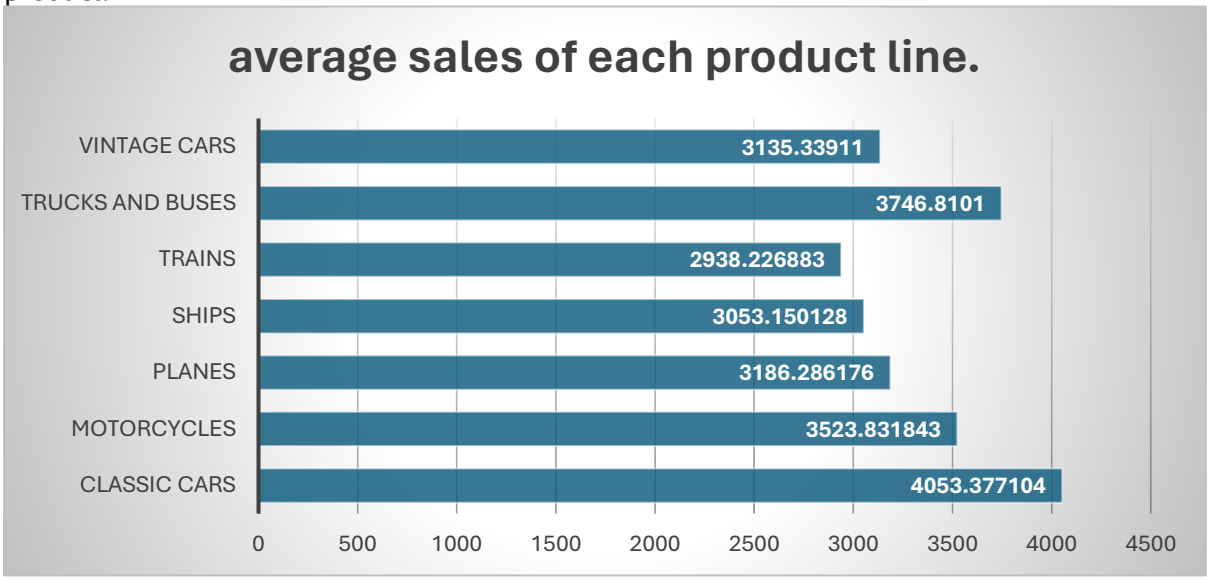
1. Comparison of sales between Vintage cars and Classic cars across all countri




es.

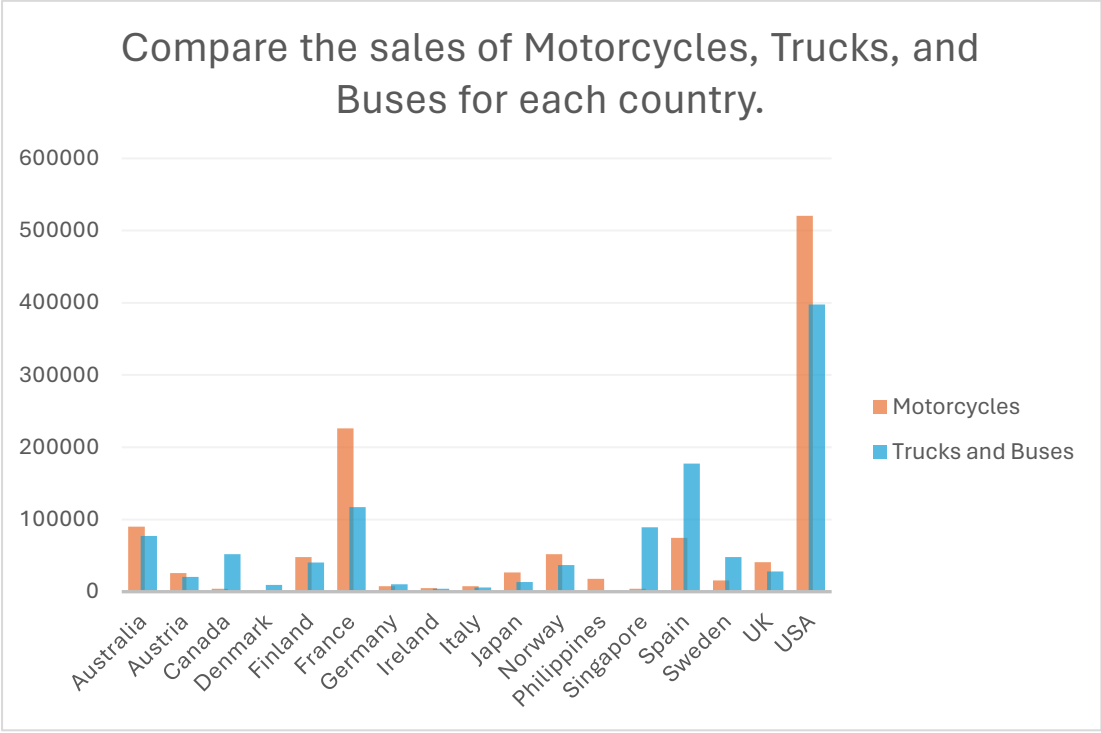
This analysis Compare the sale of Vintage cars and Classic cars for all the countries. Where USA(2102394.02) has the highest sales followed by Spain, France, and Australia.

2. Determination of the average sales of all products and identification of the highest-selling product.



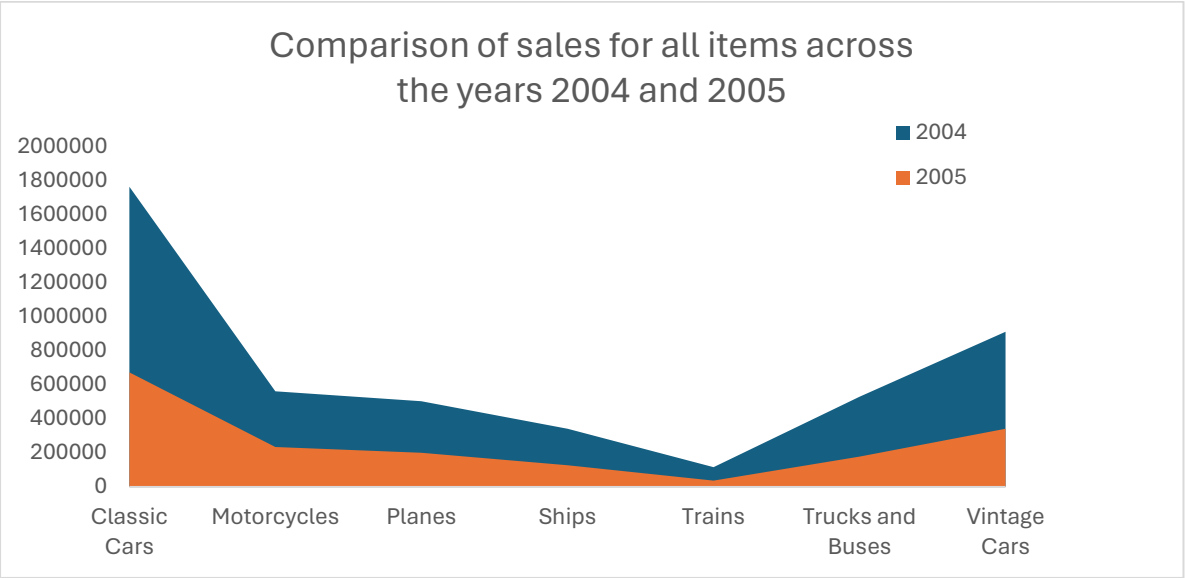

This analysis aims to provide average sales of all products and identification of the highest-selling product. And through the graph we can see that Classic Cars have the highest sales with 4053.377104 average sales followed by Trucks and Buses and Motorcycles.

3. Assessment of the country yielding the most profit for Motorcycles, Trucks, and Buses.



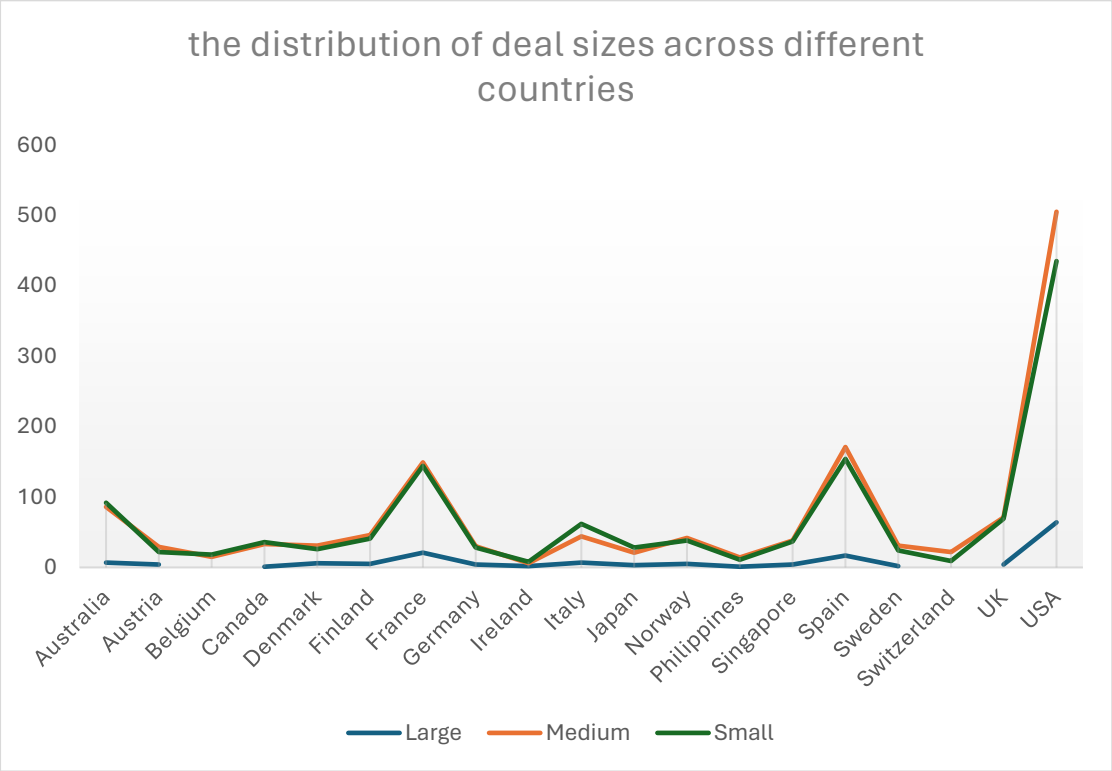

The country yielding the most profit for Motorcycles, Trucks, and Buses. And bar chart shows that the USA has the highest sales with 397842.42 sum of sales for Trucks and Buses while 520371.7 sum of sales for Motorcycles followed by France and Spain.

4. Comparison of sales for all items across the years 2004 and 2005.




This analysis aims to compare sales across all items for the years 2004 and 2005. Using a line chart, we observe that sales for all items fluctuate significantly between these two years. Notably, Classic cars emerge as the top-selling category in both years, with sales totaling $1,762,257.09 in 2004 and $672,573.28 in 2005.

5. Comparative analysis of all countries based on deal size.




This analysis find out the distribution of deal sizes across the different countries. The line chart shows that the deal size in the USA with large deal size of 64, medium deal size of 505, and small deal size of 435 is way higher than all the other countries.

# Conclusion and Review

The analysis uncovers significant insights into sales dynamics and profitability across categories and countries. Notably, the USA emerges as a key market leader, exhibiting strong sales performance in Vintage and Classic cars, Trucks, Buses, and Motorcycles. Classic Cars stand out as the highest-selling product, contributing significantly to overall sales revenue. Moreover, the USA demonstrates exceptional profitability, particularly in the Trucks, Buses, and Motorcycles categories. Sales for Classic cars remain consistently robust throughout the years 2004 and 2005, indicating sustained demand for this category. Additionally, the USA showcases markedly larger deal sizes compared to other countries, underscoring its dominance in sales volume.

While the analysis effectively presents key findings through visualizations, further exploration into factors influencing sales fluctuations and deal size disparities could provide deeper insights. Overall, the report offers valuable insights for optimizing sales strategies and driving business growth.

# Regression

The multiple R value of 0.877 suggests a strong positive linear relationship between the independent variables (MSRP, Quantity Ordered) and the dependent variable (Sales). The coefficient values indicate that for every unit increase in MSRP, there's an increase of approximately $103.08 in sales. Similarly, for every unit increase in Quantity Ordered, sales increase by about $12.82, and for every unit increase in the third independent variable, sales increase by approximately $47.43. The adjusted R-squared value of 0.766 indicates that the model explains about 76.6% of the variance in the sales data

| **SUMMARY OUTPUT** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| ***Regression Statistics*** | | | | | | | | |
| **Multiple R** | 0.877178 | | | | | | | |
| **R Square** | 0.769441 | | | | | | | |
| **Adjusted R Square** | 0.766629 | | | | | | | |
| **Standard Error** | 896.6688 | | | | | | | |
| **Observations** | 250 | | | | | | | |
| **ANOVA** | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| **Regression** | 3 | 6.6E+08 | 2.2E+08 | 273.6567 | 4.62E-78 | | | |
| **Residual** | 246 | 1.98E+08 | 804014.9 | | | | | |
| **Total** | 249 | 8.58E+08 | | | | | | |
| | | | | | | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| **Intercept** | -5271.93 | 322.9166 | -16.326 | 4.32E-41 | -5907.96 | -4635.9 | -5907.96 | -4635.9 |
| **X Variable 1** | 103.0809 | 6.001152 | 17.17685 | 5.42E-44 | 91.26071 | 114.9011 | 91.26071 | 114.9011 |
| **X Variable 2** | 12.81807 | 1.661734 | 7.713668 | 3.04E-13 | 9.545024 | 16.09111 | 9.545024 | 16.09111 |
| **X Variable 3** | 47.42944 | 3.350938 | 14.15408 | 1.13E-33 | 40.82925 | 54.02963 | 40.82925 | 54.02963 |

# Anova: one factor

The ANOVA results indicate that there is a significant difference between the groups, as the p-value is very low (3.1E-113). This suggests that there's strong evidence to reject the null hypothesis, indicating that at least one of the means of the groups (Sales and MSRP) is significantly different from the others. The F-value of 894.0704 further supports this, as it is much greater than 1, indicating that there is a significant difference between the groups

***Anova: Single Factor***

| | | | | | | |
|---|---|---|---|---|---|---|
| *SUMMARY* | | | | | | |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* | | |
| *Column 1* | 250 | 903280.9 | 3613.123 | 3445221 | | |
| *Column 2* | 250 | 25534 | 102.136 | 1664.552 | | |
| | | | | | | |
| | | | | | | |
| *ANOVA* | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| *Between Groups* | 1.54E+09 | 1 | 1.54E+09 | 894.0704 | 3.1E-113 | 3.860199 |
| *Within Groups* | 8.58E+08 | 498 | 1723443 | | | |
| | | | | | | |
| *Total* | 2.4E+09 | 499 | | | | |

# Anova: two factor

This two-factor ANOVA without replication analyzes the impact of Sales, MSRP, and Quantity Ordered on the variance in the dataset. The ANOVA results show that there is no significant difference between the rows (Sales, MSRP, and Quantity Ordered) as the p-value (0.33951) is greater than the significance level (0.05). However, there is a significant difference between the columns (Sales and MSRP) with a very low p-value (1.9E-168), indicating that at least one of the means of the groups is significantly different from the others.

| Anova: Two-Factor Without Replication | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| *SUMMARY* | *Count* | *Sum* | *Average* | *Variance* | | |
| **Row 1** | 3 | 4097.66 | 1365.887 | 5069957 | | |
| **Row 2** | 3 | 2451.12 | 817.04 | 1725170 | | |
| **Row 3** | 3 | 1566 | 522 | 648687 | | |
| **Row 4** | 3 | 5095.24 | 1698.413 | 7507173 | | |
| **Row 5** | 3 | 5140.39 | 1713.463 | 7650609 | | |
| **Row 248** | 3 | 4386.35 | 1462.117 | 5944534 | | |
| **Row 249** | 3 | 2261.6 | 753.8667 | 1546167 | | |
| **Row 250** | 3 | 4176.72 | 1392.24 | 5420980 | | |
| **Sales** | 250 | 903280.9 | 3613.123 | 3445221 | | |
| **MSRP** | 250 | 25534 | 102.136 | 1664.552 | | |
| **QuantityOrdered** | 250 | 8659 | 34.636 | 89.69428 | | |
| **ANOVA** | | | | | | |
| ***Source of Variation*** | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| **Rows** | 2.95E+08 | 249 | 1182944 | 1.044989 | 0.33951 | 1.194432 |
| **Columns** | 2.09E+09 | 2 | 1.05E+09 | 925.2361 | 1.9E-168 | 3.013826 |
| **Error** | 5.64E+08 | 498 | 1132016 | | | |
| | | | | | | |
| **Total** | 2.95E+09 | 749 | | | | |
| | | | | | | |

# Descriptive Statistics

The descriptive statistics for Quantity Ordered, Sales, MSRP, and Price Each reveal valuable insights into the dataset. Quantity Ordered has a mean of 34.636 units, with a standard deviation of 9.470706, indicating moderate variability in the quantity ordered. Sales, on the other hand, show a much higher variability, with a mean of 3613.123 and a standard deviation of 1856.131. The MSRP (Manufacturer's Suggested Retail Price) has a mean of 102.136, with a standard deviation of 40.79892, suggesting moderate variability in the price. In contrast, Price Each, with a mean of 84.45296 and a standard deviation of 20.22993, exhibits less variability compared to MSRP. The skewness and kurtosis values provide insights into the distribution shape and tail behaviour of the variables. Overall, these descriptive statistics offer a comprehensive understanding of the dataset's central tendency, variability, and distribution characteristics for each variable.

| *Quantity Ordered* | | *Sales* | | *MSRP* | | *Price Each* | |
|---|---|---|---|---|---|---|---|
| **Mean** | 34.636 | Mean | 3613.123 | Mean | 102.136 | Mean | 84.45296 |
| **Standard Error** | 0.59898 | Standard Error | 117.392 | Standard Error | 2.58035 | Standard Error | 1.279453 |
| **Median** | 34 | Median | 3263.96 | Median | 99 | Median | 100 |
| **Mode** | 29 | Mode | #N/A | Mode | 118 | Mode | 100 |
| **Standard Deviation** | 9.470706 | Standard Deviation | 1856.131 | Standard Deviation | 40.79892 | Standard Deviation | 20.22993 |
| **Sample Variance** | 89.69428 | Sample Variance | 3445221 | Sample Variance | 1664.552 | Sample Variance | 409.2499 |
| **Kurtosis** | -0.64676 | Kurtosis | 1.127057 | Kurtosis | -0.19836 | Kurtosis | -0.40344 |
| **Skewness** | 0.256745 | Skewness | 1.013489 | Skewness | 0.517104 | Skewness | -0.9678 |
| **Range** | 51 | Range | 10626.85 | Range | 181 | Range | 73.12 |
| **Minimum** | 15 | Minimum | 652.35 | Minimum | 33 | Minimum | 26.88 |
| **Maximum** | 66 | Maximum | 11279.2 | Maximum | 214 | Maximum | 100 |
| **Sum** | 8659 | Sum | 903280.9 | Sum | 25534 | Sum | 21113.24 |
| **Count** | 250 | Count | 250 | Count | 250 | Count | 250 |

# Correlation

The correlation matrix indicates the relationships between Quantity Ordered, Sales, and Price Each. There's a moderate positive correlation of approximately 0.514 between Quantity Ordered and Sales, suggesting that as the quantity ordered increases, sales tend to increase as well. Similarly, there's a weak positive correlation of about 0.664 between Sales and Price Each, indicating that higher-priced items may contribute to higher sales, albeit to a lesser extent. However, there seems to be a negligible correlation of approximately -0.013 between Quantity Ordered and Price Each, implying that changes in the quantity ordered don't significantly impact the individual item price.

| | *Quantity Ordered* | *Sales* | *Price Each* |
|---|---|---|---|
| **Quantity Ordered** | 1 | | |
| **Sales** | 0.513951 | 1 | |
| **Price Each** | -0.01254 | 0.663973 | 1 |