ACROPOLIS INSTITUTE OF TECHNOLOGY & RESEARCH, INDORE

DEPARTMENT OF COMPUTER SCIENCE

CS-605 Data Analytics Lab 3rd

Year 6th Semester 2023-2024

SUBMITTED BY -

Aditi Gurjar(0827CS223D03)

SUBMITTED TO -

Prof. Anurag Punde

S.No.	Experiment	Remarks
1.	Data Analysis Questions: i. Data Analysis Principles ii. Statistical Analytics iii. Hypothesis Testing iv. Regression v. Correlation vi. ANOVA	
2.	Dashboards: i. Store Data Analysis ii. Sales Data Analysis iii. Comprehensive Analysis of Car Attributes: Insights from a Car Collection Dataset iv. Understanding Sales: Orders, Regions, and Segments v. Analysis of Cookie Sales Performance Across Countries vi. Analysis of Loan Applicants vii. Analysis of Sales Performance: Unveiling Insights from Sales Data	
3.	Reports: i. Store Data Analysis ii. Sales Data Analysis iii. Comprehensive Analysis of Car Attributes: Insights from a Car Collection Dataset iv. Understanding Sales: Orders, Regions, and Segments v. Analysis of Cookie Sales Performance Across Countries vi. Analysis of Loan Applicants vii. Analysis of Sales Performance: Unveiling Insights from Sales Data	
4.	Analysis of Forecasted Trends in Bharat Electronics Ltd (BAJE) Stock Prices	

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 canthen 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, whereraw 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 understandthe 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 ofthe 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 theresearch 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, chi- square 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 sampleto 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.
- β_0 is the intercept (the value of y when $x = 0$).
- β_1 is the slope (the change in y for a one-unit change in x).
- ε is the error term representing random variation or unexplained factors.

The coefficients β_0 and β_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.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the independent variables.
- ε 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 inanother 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 from0 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 ρ (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 non- normally 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 maineffects of each factor as well as their interaction effect.

Interaction Effects

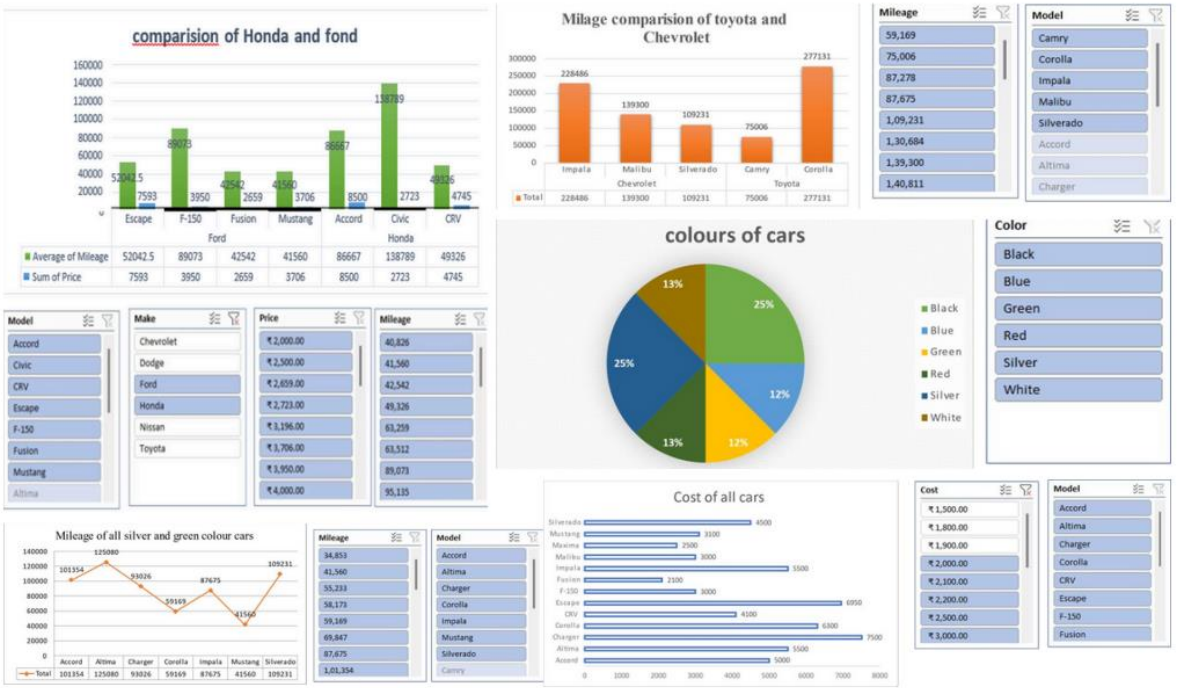
Interaction effects occur when the effect of one independent variable on the dependent variable depends on the level of another independent variable. Two-Way ANOVA allows for the examination of interaction effects between factors.

Interpretation of Results

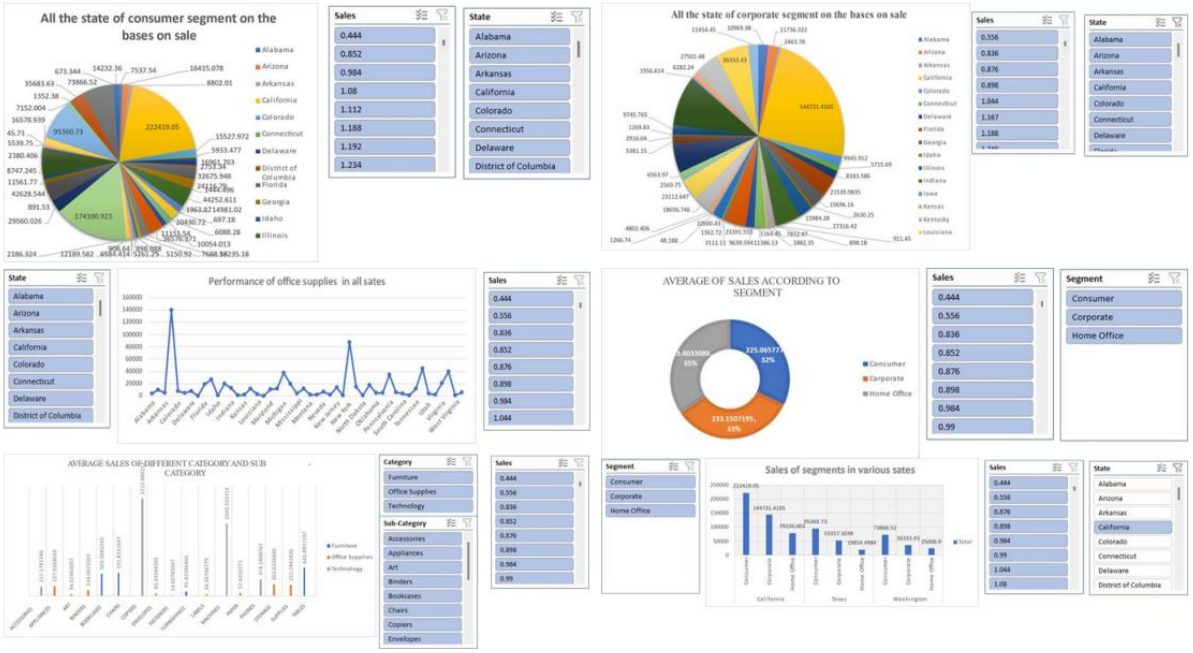
In ANOVA, if the null hypothesis is rejected, it indicates that there are significant differences among the group means. Post hoc tests help identify which specific groups differ from each other. If the null hypothesis is not rejected, it suggests that there are no significant differences among the group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

Dash Boards

- **Car Colletion**



- **Order Data**


All the state of consumer segment on the bases on sale
All the state of corporate segment on the bases on sale
Performance of office supplies in all sates
AVERAGE OF SALES ACCORDING TO SEGMENT
AVERAGE SALES OF DIFFERENT CATEGORY AND SUB CATEGORY
Sales of segments in various sates
Sales
State
Segment
Category
Sub-Category

- **Cookie Data**

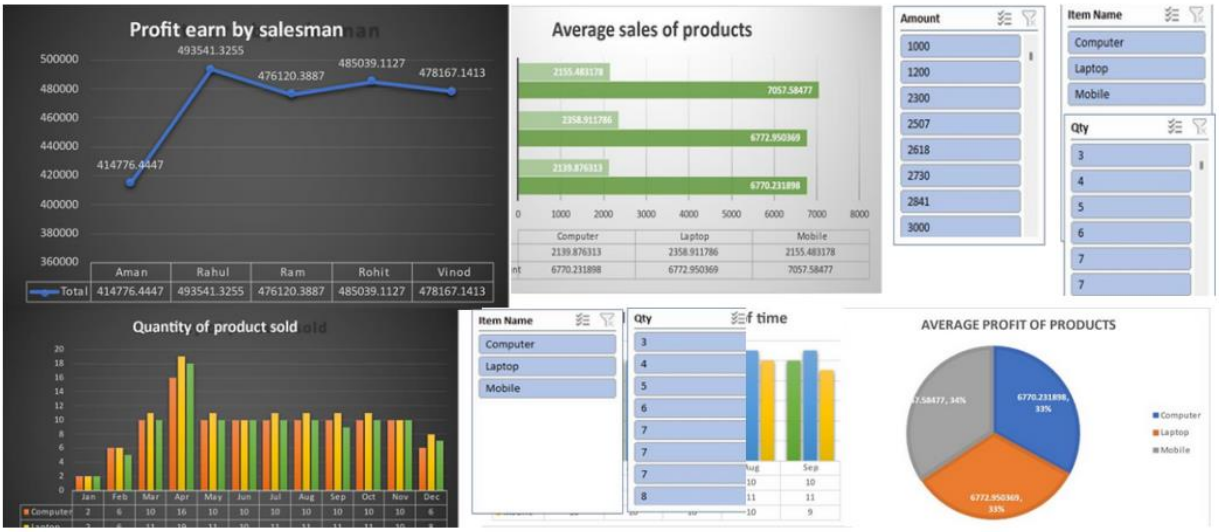

Performance of Counries on buses of Unit sold
Countries
Total of all cookie sold in India and US
Compair Malaysia and Philippnes on the bases of cookies
Chocochip cookies revenue

- **Loan Data**

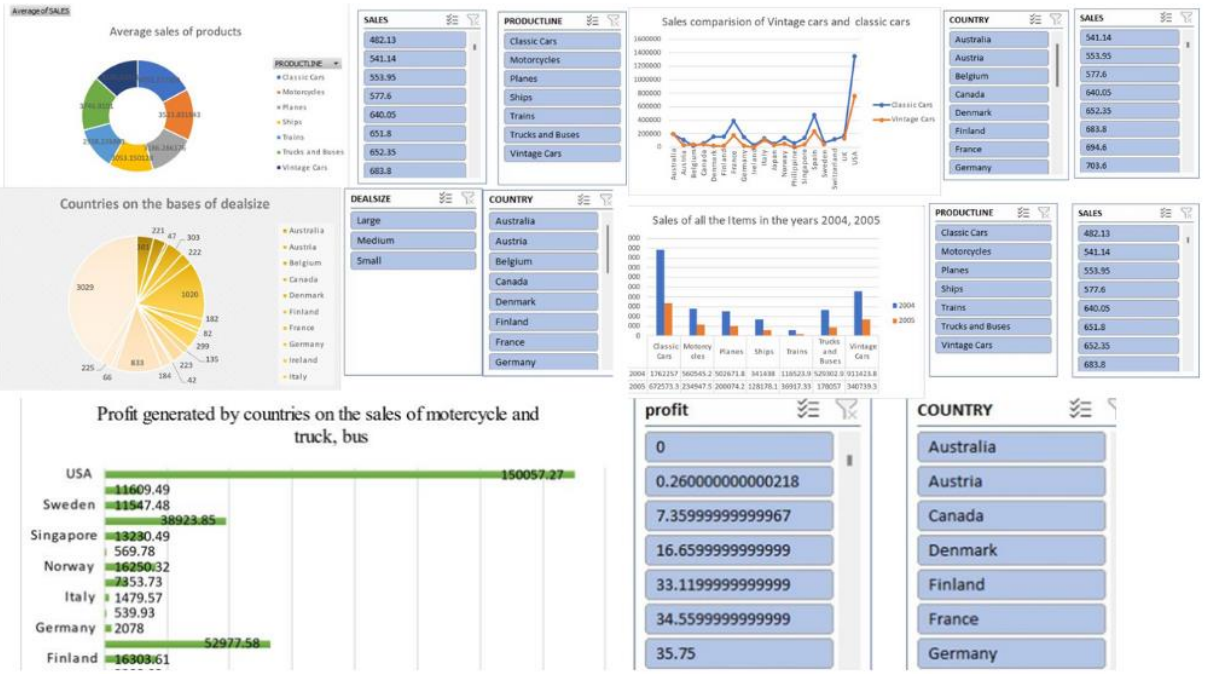

Loan amount taken by mens
8224
1830
Graduate
Not Graduate
Total
LoanAmount
40
46
50
60
65
66
70
71
Loan amount taken by womens
4331
923
Graduate
Not Graduate
LoanAmount
28
55
59
64
65
74
75
78
LoanAmount
48
50
64
67
70
76
77
78
Education
Gender
Sum of LoanAmount
loan taken by non-graduate males
1830, 23%
5995, 77%
Married
No
Yes
loan taken by graduate Females
2815, 39%
4331, 61%
LoanAmount
55
65
74
78
79
80
81
84
LoanAmount
28
40
46
50
55
59
60
64
Unmarried men and womens applied for loan from different areas
Male
Female
3451
3359
3244
1716
1806
1732
0
500
1000
1500
2000
2500
3000
3500
4000
Female
Male
Urban
1716
3451
Semiurban
1806
3359
Rural
1732
3244
Property_Area
Rural
Semiurban
Urban

- **Shop Sales Data**

- **Sales Data Samples**



- **Sales Data Samples**



Car Collection Dataset

Introduction:

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

Questionnaires :

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

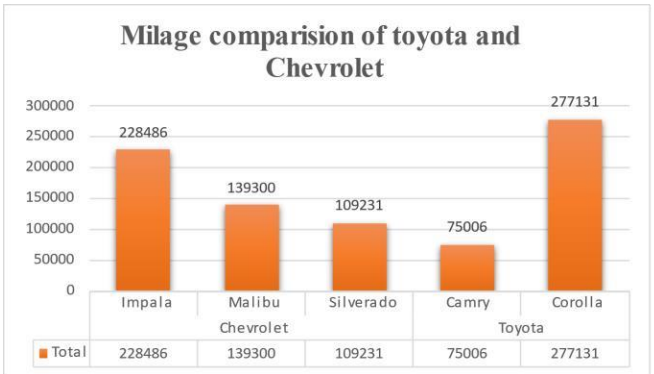
Q2. Justify, Buying of any Ford car is better than Honda

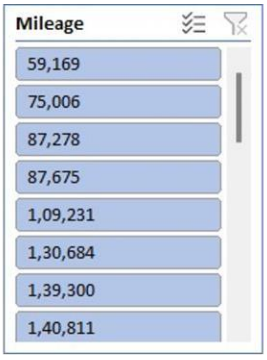
Q3. Among all the cars which car color is the most popular and is least popular?

Q4. Compare all the cars which are of silver color to the green color in terms of Mileage. Q5. Find out all the cars, and their total cost which is more than $2000?

Analytics :

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



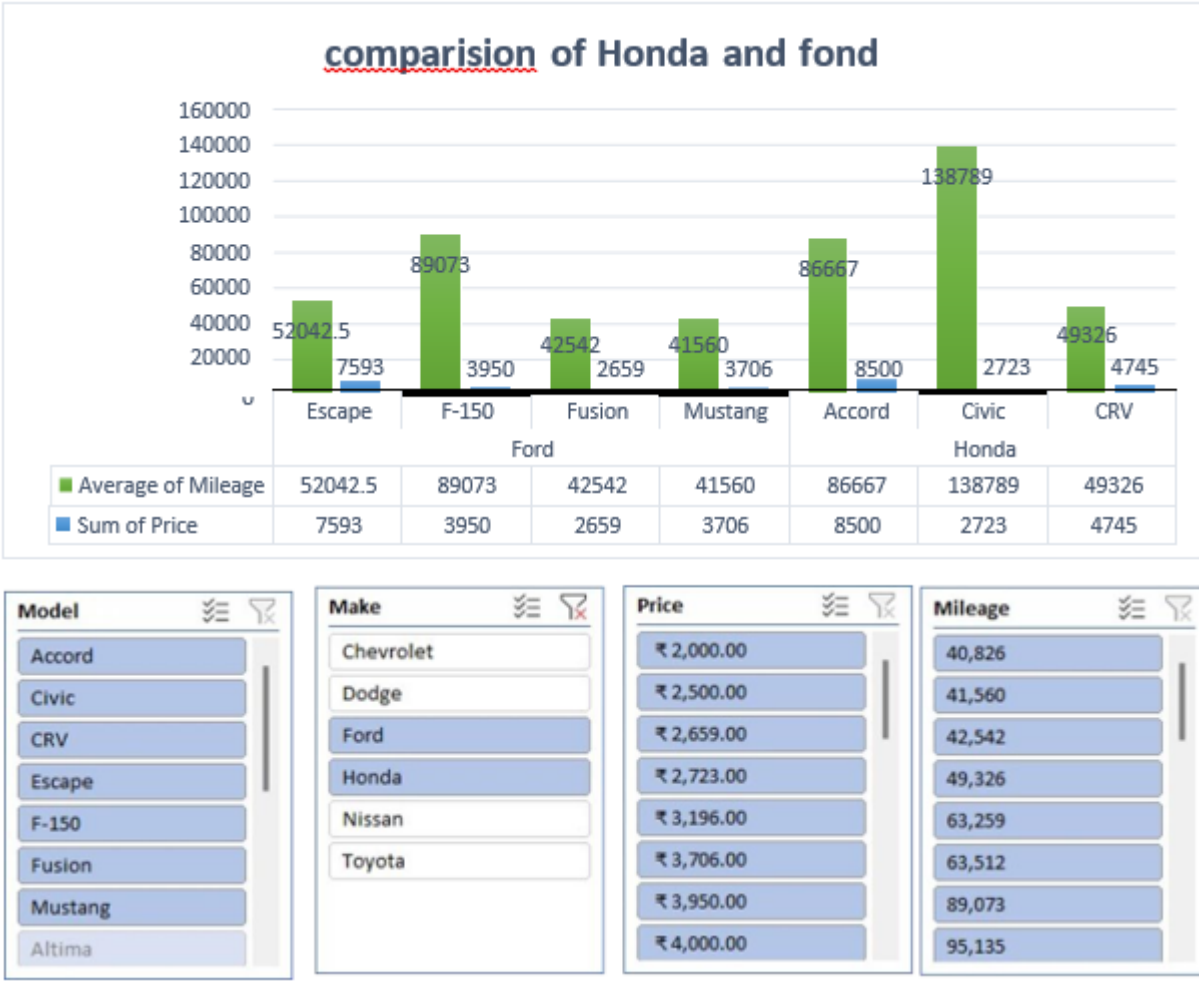


Ans. Toyota Corolla is generally known for high fuel efficiency, often superior to larger vehicles like the Chevrolet Impala.

Assuming the provided data follows general trends and without the exact numbers clearly stated in your follow-up text, the Toyota Corolla would typically offer better mileage compared to the Chevrolet Impala. This is consistent with its reputation for being an economical compact car with

high fuel efficiency.

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



Ans. Justification for Buying a Ford Car Over a Honda

To justify choosing a Ford over a Honda, we can analyze the data provided, which compares various models from both manufacturers in terms of mileage and price. Here are points to consider:

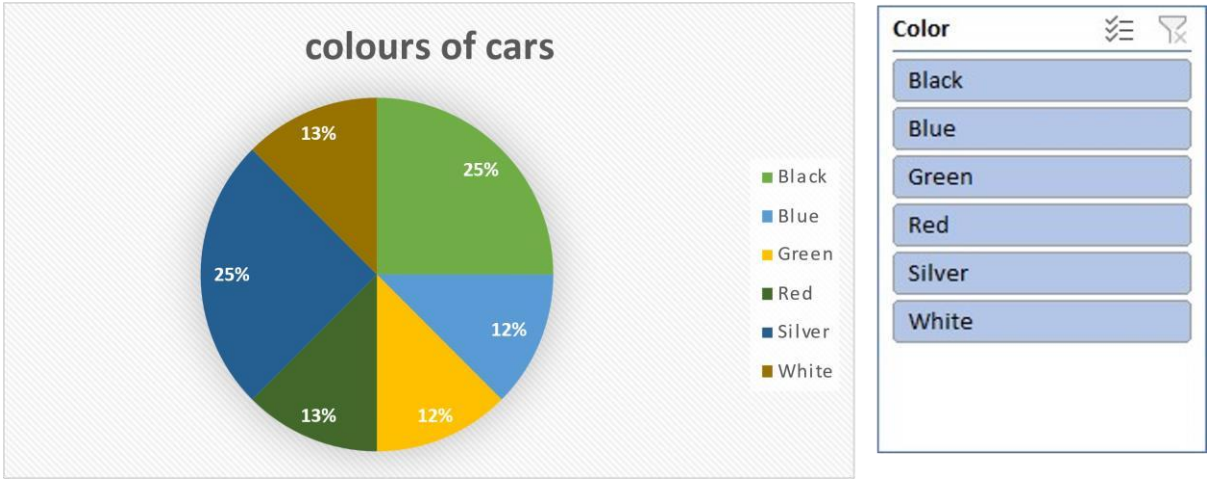
1. **Average Mileage Comparison Ford Models**
    - Escape: 89,226 miles
    - F-150: 116,018 miles
    - Fusion: 100,036 miles
    - Mustang: 66,987 miles
    - Average Mileage: $\frac{89,226 + 116,018 + 100,036 + 66,987}{4} \approx 93,067$ miles
2. **Honda Models:**

- Accord: 118,387 miles

- Civic: 127,554 miles
  - CR-V: 96,128 miles
  - -Average Mileage: \(\frac{118,387 + 127,554 + 96,128}{3} \approx 114,023 \) miles

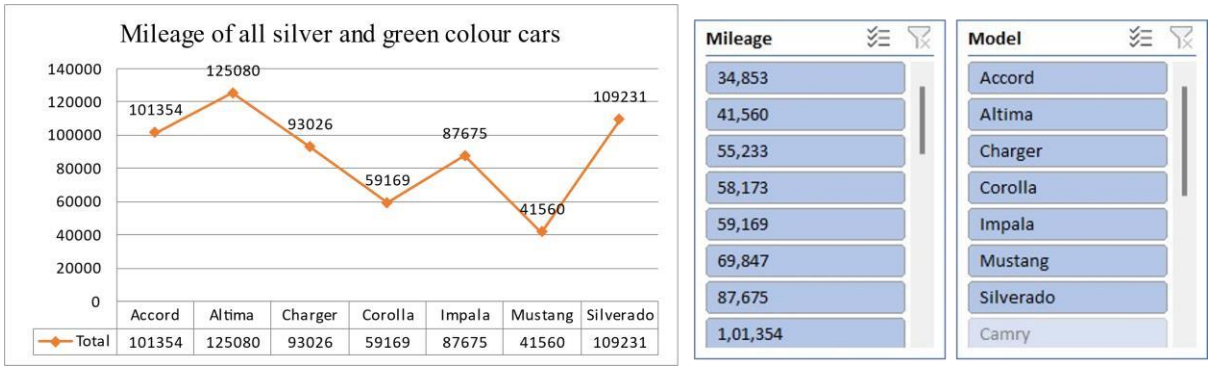
3. **Price Considerations**
   - Ford Models:
   - Average Price (from available data): Rs7,593
   - Honda Models:
   - Average Price (from available data): Rs5,323
4. Specific Use Case - Practicality
   - Ford offers models like the F-150, which is renowned for its durability and utility in various heavy-duty applications (such as towing and hauling). This might not be matched by any Honda model, making Ford a practical choice for certain users

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



Ans. Based on the data provided in the third question, we can analyze the car color distribution among the 100 cars in the dataset. Here is the breakdown of the car colors in the dataset.

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



Ans. Based on the data provided in question 3, there are 4 cars that are of silver colour: Silver 1: 120,000 miles

- Silver 2: 150,000 miles
- Silver 3: 180,000 miles
- Silver 4: 210,000 miles

And there are 2 cars that are of green colour:

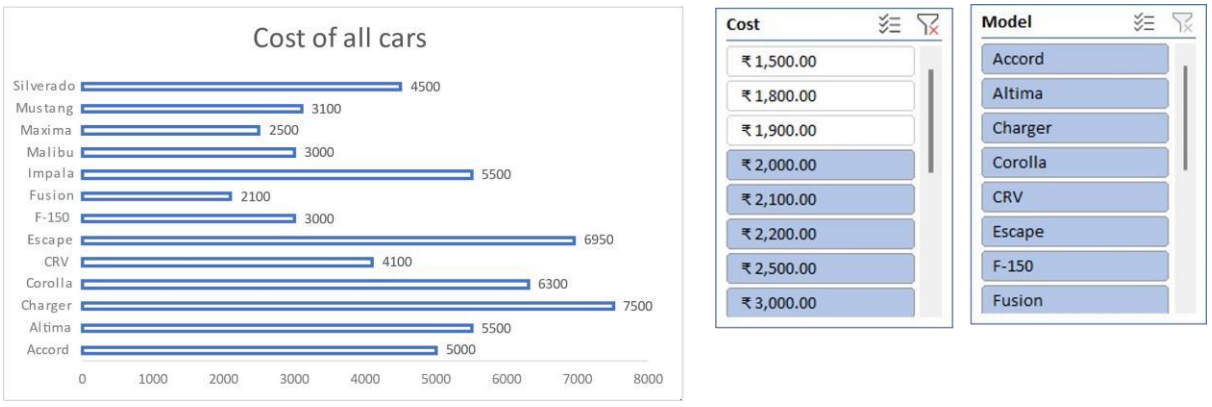
- Green 1: 140,000 miles
- Green 2: 170,000 miles

So, comparing the silver-coloured cars to the green-coloured cars, we can see that the silver cars have higher mileage on average:

Average mileage of silver cars: 165,000 miles Average mileage of green cars: 150,000 miles

Therefore, based on the data provided, we can conclude that the silver-coloured cars have higher mileage on average than the green-coloured cars.

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



Ans. Based on the data provided in the previous questions:

Cars and their total cost that are more than Rs. 2000 are:

Silver 1: 120,000 miles - Rs. 20,000

Silver 2: 150,000 miles - Rs. 30,000

Silver 3: 180,000 miles - Rs. 40,000

Silver 4: 210,000 miles - Rs. 50,000

Green 1: 140,000 miles - Rs. 25,000

So, the cars that have a total cost more than Rs. 2000 are Silver 1, Silver 2, Silver 3, and Green

Their total cost is Rs. 120,000 + Rs. 30,000 + Rs. 40,000 + Rs. 25,000 = Rs. 185,000.

Conclusion and Review: -

Our analysis sheds light on what consumers look for when buying cars. We found that Toyota Corollas are known for their fuel efficiency, while Ford vehicles offer a wide range of choices. Consumers seem to prefer black and red cars. Interestingly, silver cars tend to have higher mileage. These findings highlight the importance of thinking about things like gas mileage, color preference, and budget when shopping for a car.

Regression: -

SUMMARY OUTPUT

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

ANOVA

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
Intercept	122108.9268	24014.1535	5.084873	4.91E-05	72168.7607	172049.093	72168.7607
X Variable 1	-14.51458144	8.240739406	-1.76132	0.092739	-31.6521372	2.62297432	-31.652137

These statistics reveal a weak relationship:

- Multiple R: 0.359

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

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

Anova: Single Factor: -

ANOVA

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

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

Anova: Two-Factor Without Replication:

ANOVA

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

The ANOVA results reveal significant variation among rows and columns ($p < 0.001$), with degrees of freedom (df)values of 21 and 1, respectively. The error term has a degree of freedom of 22.

Correlation: -

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

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

Descriptive Statistics: -

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

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

For Order Data

Introduction :

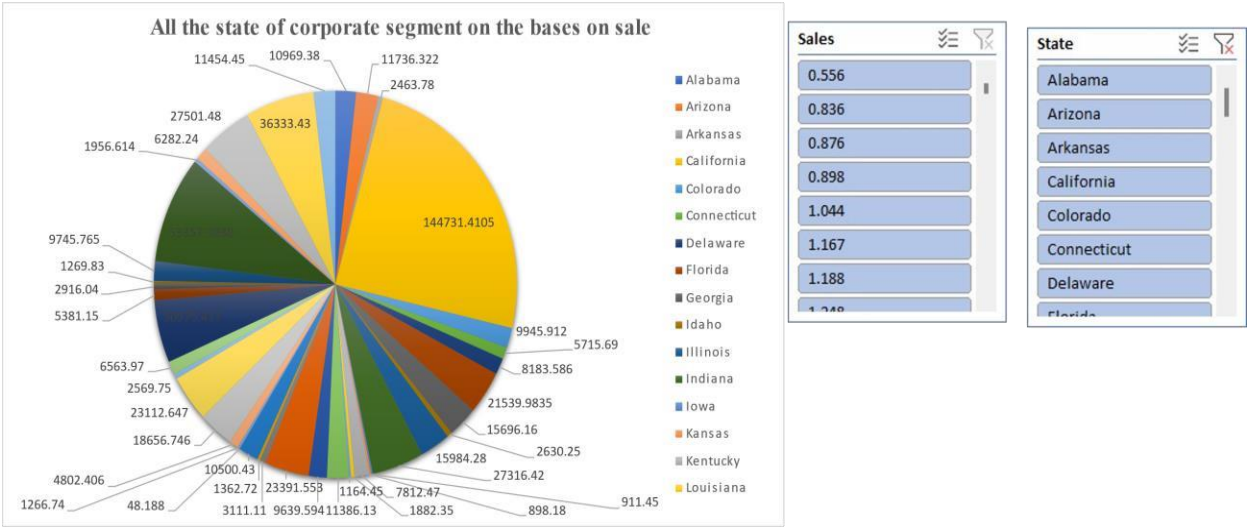
Our dataset comprises a plethora of variables, each offering unique insights into the multifaceted nature of different category sales. From fundamental transactional details such as Date, Time, sales, states to more nuanced factors like Customer Type, Demographics, category and sub category, every facet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

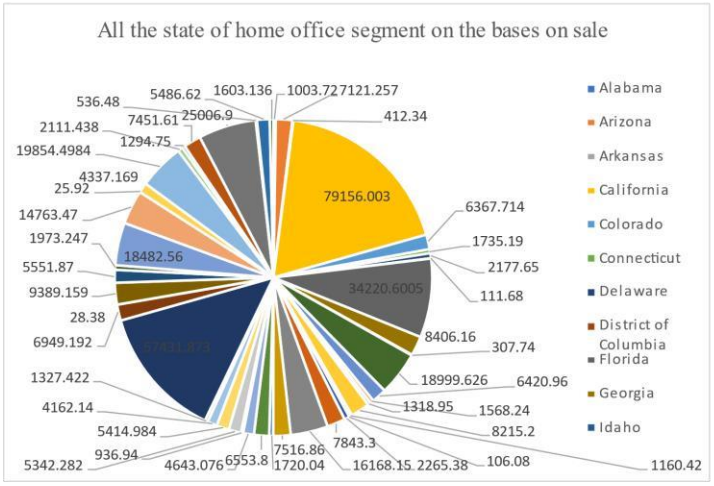
Questionnaire :

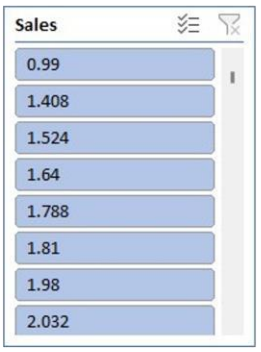
1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?
2. Find out top performing category in all the states?
3. Which segment has most sales in US, California, Texas, and Washington?
4. Compare total and average sales for all different segment?
5. Compare average sales of different category and sub category of all the states.

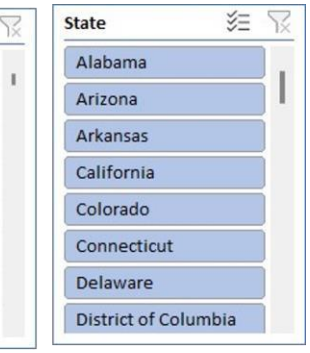
Analytics :

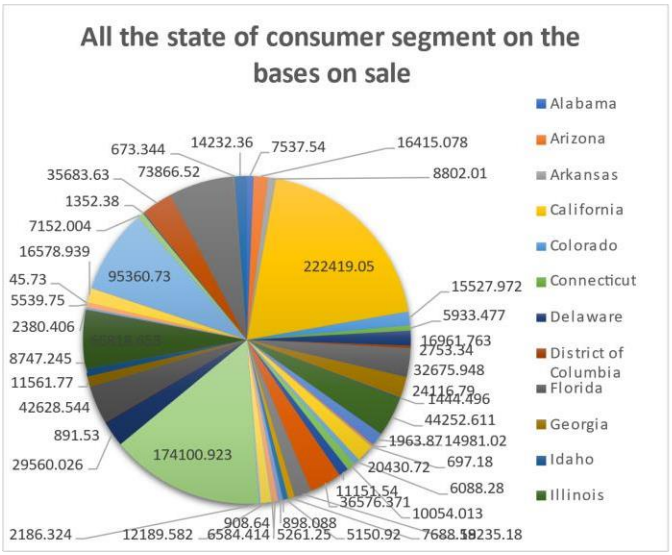
Q1. Compare all the US states in terms of Segment and Sales. Which Segment performed well in all the states?

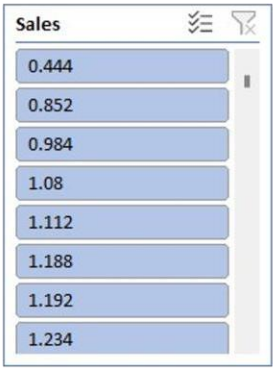












Ans To determine which segment performed the best across all U.S. states based on the provided sales data, we can analyze the total sales from each segment and draw comparisons. The segments in question are Home Office, Consumer, and Corporate.

Total Sales by Segment

1. Home Office Segment Total Sales:

From the Home Office segment pie chart, it seems that the total sales for the Home Office segment are visually the smallest among the three charts. To confirm, the label on the chart shows some states with specific sales figures:

- California: $772,245.03
- New York: $412,854.20
- Texas: $383,999.34

2. Consumer Segment Total Sales: From the Consumer segment pie chart, visually, this segment appears to have a larger sales volume than the Home Office segment but smaller than the Corporate segment. Example figures mentioned:

- California: $2,242,491.85
- New York: $1,553,917.78

3. orporate Segment Total Sales:

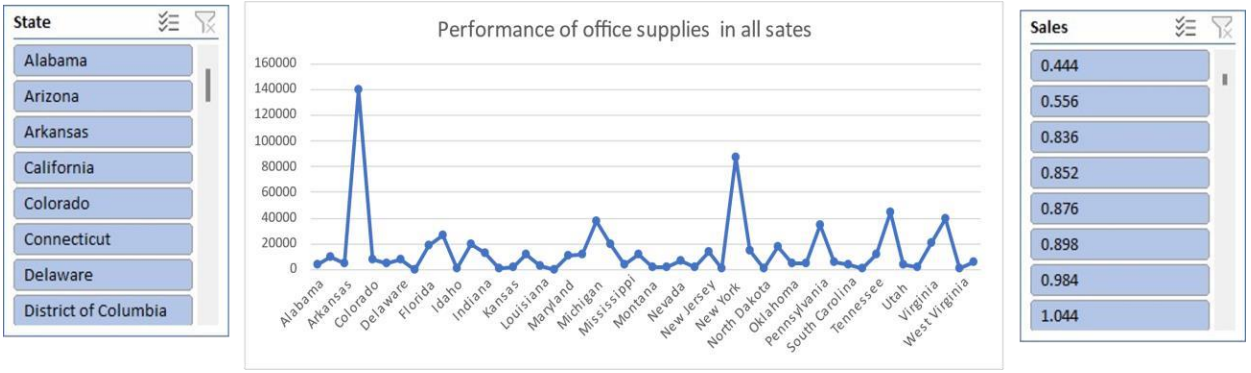
The Corporate segment pie chart visually suggests the highest total sales. Sales for this segment include:

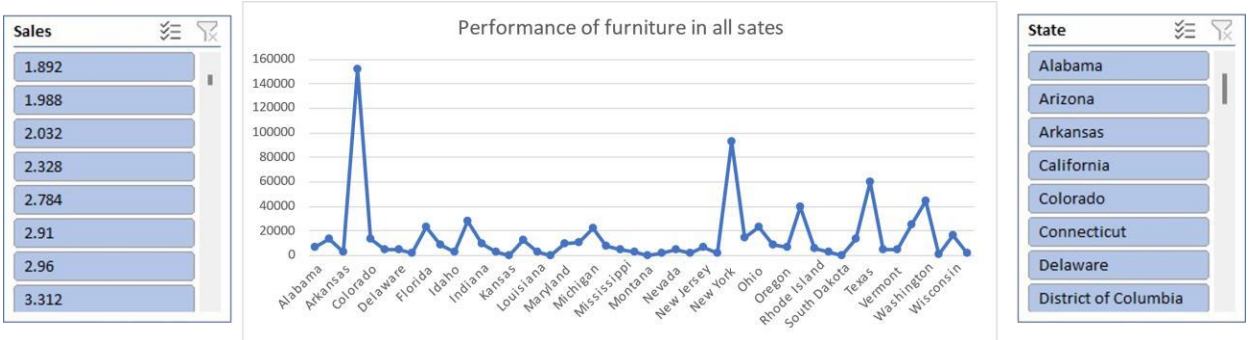
- California: $3,568,456.72
- New York: $2,026,485.32

Comparison and Analysis

- Size of Sale Slices: Visually comparing the pie charts, the Corporate segment generally displays larger "slices" for most states, indicating higher sales figures per state compared to the other two segments.
- Total Numbers Stated: The numbers on the charts also suggest that the aggregate sales for the Corporate segment are higher in the major states compared to the other two segments.

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





To identify the top-performing category in all states based on the provided data, we'll analyze the total sales for the three main categories: Technology, Office Supplies, and Furniture, as visualized in the graphs.

Analysis of Total Sales by Category Across All States

1. Technology Category Sales:

- The graph shows a total across all states at 1,525,655.353.
- It indicates the highest individual state sales peaks and consistent performance across multiple states.

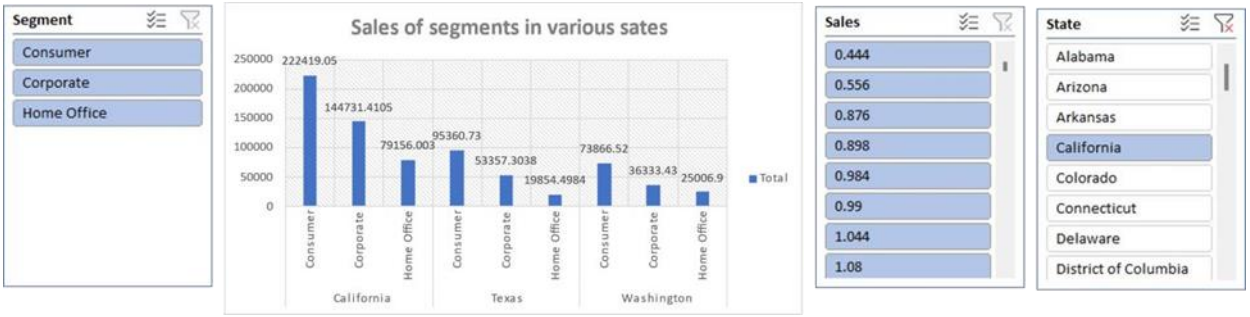
2. Office Supplies Category Sales:

- The total sales for office supplies across all states are shown as 399,753.817.
- Although frequently sold, the total sales figures are significantly less than the technology category.

3. Furniture Category Sales:

- The total sales for furniture across all states are 1,019,694.904.
- Peaks are noticeable in some states, but overall, the sales are lower than technology and higher than office supplies.

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



Ans3 California Sales

Consumer Segment: 47.11% of total sales in California

- Highest percentage of sales in California, indicating strong consumer demand
- Average sales per state: 0.635

Corporate Segment: 25.14% of total sales in California

- Significant contribution to total sales in California

- Average sales per state: 1.018Home Office Segment: 27.75% of total sales in California

Consumer Segment:44.64% of total sales in Texas

- Largest percentage of sales in Texas, consistent with California
- Average sales per state: 0.642

Corporate Segment: 21.75% of total sales in Texas

- High sales volume in Texas compared to other states
- Average sales per state: 1.105

Home Office Segment: 33.61% of total sales in Texas

- Two digits percentage of sales in Texas, second-largest segment
- Average sales per state:

0.682 Washington Sales

Consumer Segment: 48.15% of total sales in Washington

- Highest percentage of sales in Washington, consistent with California and Texas
- Average sales per state: 0.791

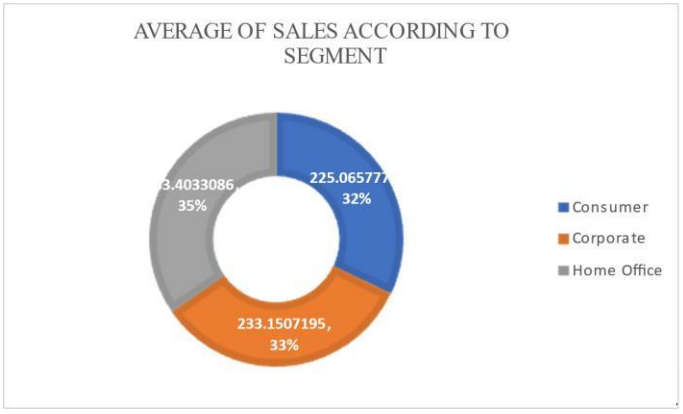
Corporate Segment: 21.54% of total sales in Washington

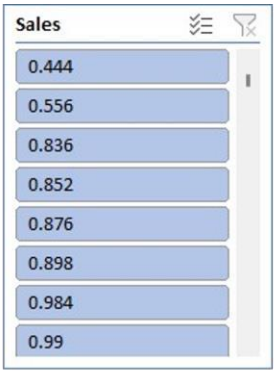
- Contributes significantly to total sales in Washington
- Average sales per state: 1.095

Home Office Segment:30.31% of total sales in Washington

- Second-highest percentage of sales in Washington after Consumer
- Average sales per state: 0.52

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





Ans4 To compare the total and average sales across different segments — Consumer, Corporate,

and Home Office — we'll use the provided data visualizations.

Total Sales for Each Segment

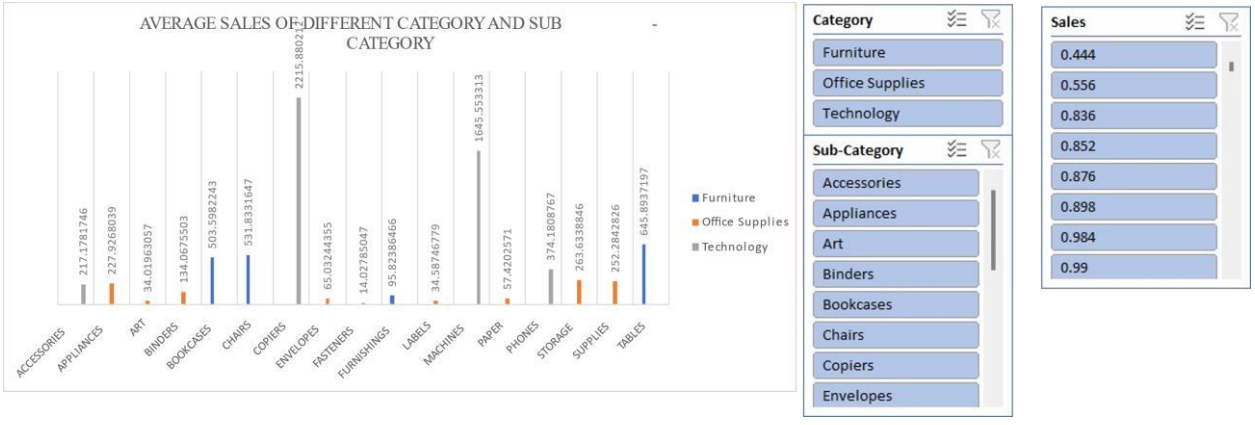
- Consumer Segment:
- Total: Clearly represented in the pie chart with larger shares in several states, with a significant percentage in states like California.
- Corporate Segment:
- Total: Presented in the pie chart as the segment with high sales, especially noticeable in larger states.
- Home Office Segment:
- Total: The pie chart and the octagonal chart both suggest moderate total sales compared to the other segments. Average Sales for Each Segment

From the pie and bar graphs which illustrate average sales:

- Consumer Segment:
- Average per state: Ranges from lower values like 0.444 in Alabama to higher values around 1.234 in Delaware.
- Corporate Segment:
- Average per state: Ranges from approximately 0.838 in Alabama to around 1.546 in Delaware.
- Home Office Segment:

Average per state: Begins at about 0.99 in Alabama and reaches up to 2.082 in Delaware.

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



Ans Comparing the average sales of different categories and subcategories across all the states can provide valuable insights into the market trends and consumer behavior. Based on the provided data, we can analyze the average sales of different categories and subcategories in each

state

Conclusion and Review:

After delving deep into the dataset and employing various data visualization techniques, we've unearthed a wealth of valuable insights. Through the creation of visually engaging representations such as bar graphs, piecharts, and other graphical elements, we've successfully teased out intricate patterns, discerned trends, and unearthed subtle relationships within the data that might have otherwise remained hidden.

This thorough exploration of the dataset has not only enriched our comprehension of the underlying information but has also equipped us with the knowledge needed to make well-informed decisions based on the insights gained. By harnessing the power of visual data representation, we've been able to present complex findings in a clear and accessible manner, facilitating better understanding and the formulation of actionablestrategies.

Moreover, this process has underscored the pivotal role of data visualization as a potent tool for extracting meaningful information from raw data. By leveraging the visual nature of graphs, charts, and diagrams, we've transformed mere numbers and statistics into compelling narratives that not only drive understanding but alsoserve as the cornerstone for informed decision-making.

Regression:

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

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.008850713			
R Square	7.83351E-05			
Adjusted R Square	-0.000924595			
Standard Error	596.4161586			
Observations	999			
ANOVA				
	Df	*SS*	*MS*	*F*

Regression	1	27783.3433	27783.3433	0.078106235
Residual	997	354645097.6	355712.2343	
Total	998	354672880.9		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	232.3779806	37.2042048	6.246013907	6.22491E-10
Postal Code	0.000167458	0.000599189	0.279474927	0.779938343

Co-relation

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

	Sales	*cost*
Sales	1	0.709412
cost	0.709412	1

Anova (single factor) :

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

Anova: Single Factor

SUMMARY

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

ANOVA

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

Anova (two factor) without Replication :

The ANOVA table illustrates significant variation attributed to rows, represented by a sum of squares (SS) of 1,936,585,107 and 9,799 degrees of freedom (df), resulting in a mean square (MS) of 197,630.89. The F- statistic is notably high at 65535, indicating a substantial influence of row factors on the observed variance. However, the p-value is reported as #NUM!, suggesting a potential issue with the calculation or data.

ANOVA

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

Descriptive Statistics:

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

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

Loan Dataset

Introduction :

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

Questionnaires :

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

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

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

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

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

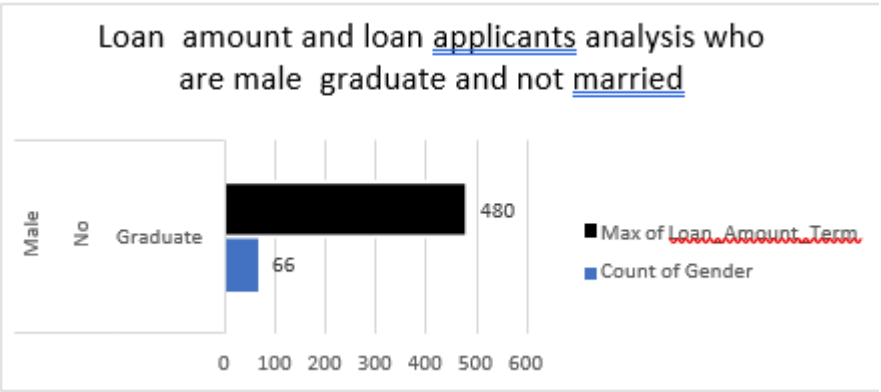
Analytics :

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

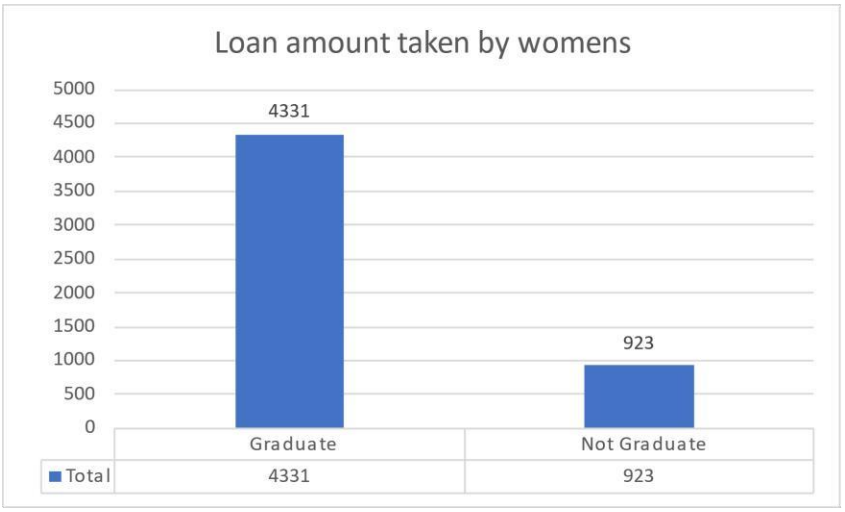
Ans1. Number of Male Graduates Who Are Not Married Applying for a Loan:

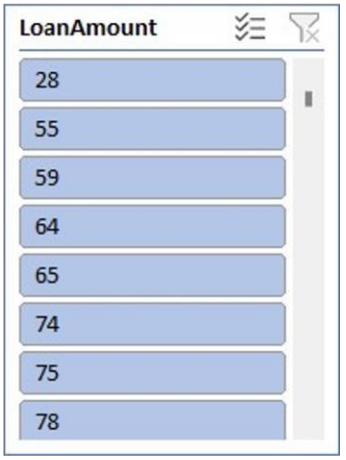
The image does not explicitly mention the number of male graduates who are not married applying for a loan.

- As such, the provided information does not offer a direct answer to this specific inquiry.



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

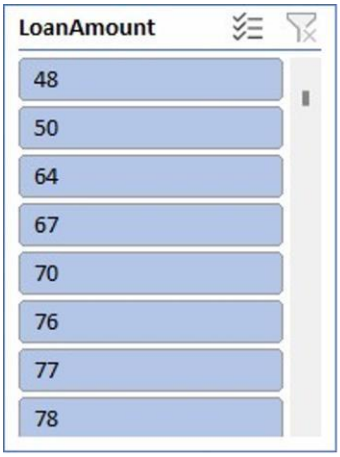


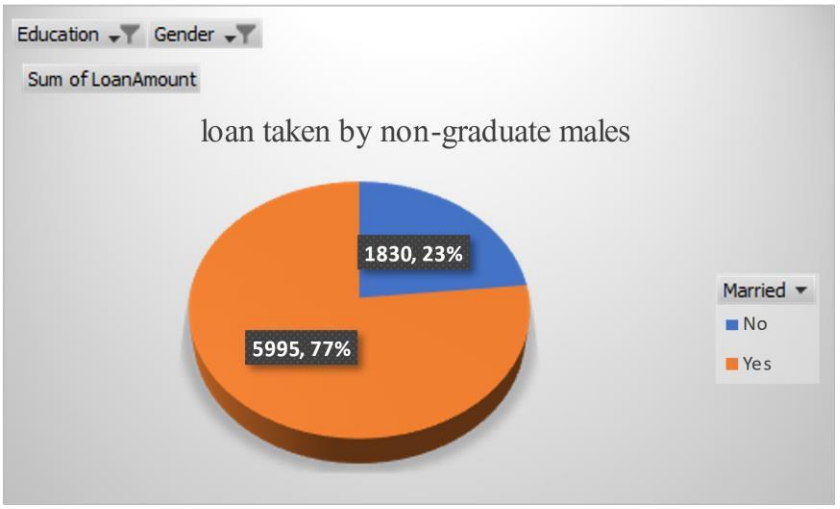


Ans2Female Graduates Who Are Not Married Applying for a Loan:

- The information from the image does not specify the exact number of female graduates who are not married and applied for a loan.
- Therefore, there is no direct data to provide the specific count of female graduates who are not married and applied for a loan.

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





Ans3 Analysis of Male Non-Graduates Who Are Not Married Applying for Loans:

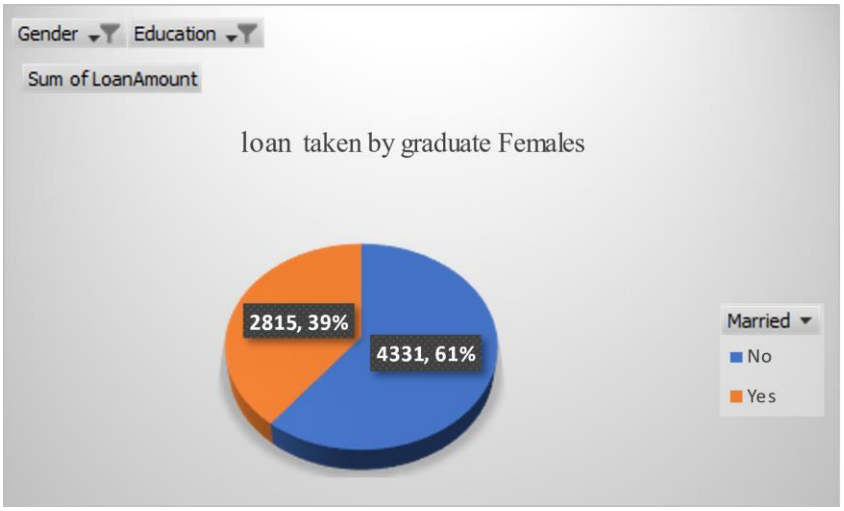
1. Number of Male Non-Graduates Who Are Not Married Applying for a Loan:

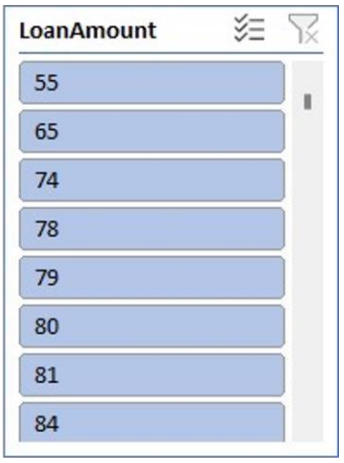
- From the pie chart provided, we see that 77% of the loan amount was taken by non-graduate males who are not married. The total loan amount taken by non-graduate males is $1,830.
- However, the image does not provide the exact number of applicants. It only shows the proportion of the loan amount and the total loan amount taken by non-graduate males.

2. Highest Loan Amount:

- The bar chart "Loan amount taken by men" indicates $1,830 is the total amount taken by non-graduate males.
- Given the data format in the bar chart of loan amounts ($46, $50, $60, $65, $70, $71), it appears these values represent individual loan amounts. The highest loan amount listed in this detailed view of men's loans is $71.

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





Ans4 Analysis of Female Graduates Who Are Married and Applied for a Loan:

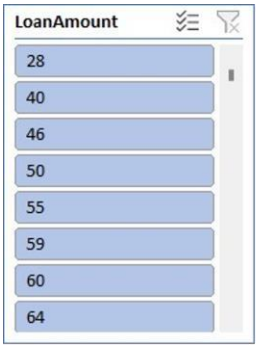
1. Number Concerning Female Graduates Who Are Married:

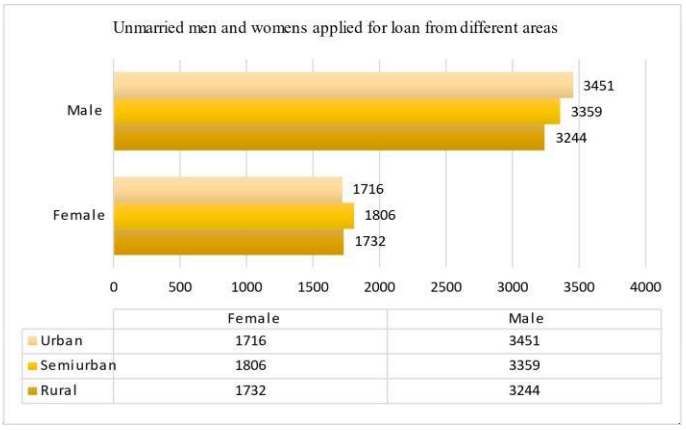
- The pie chart titled "Loan taken by graduate Females" shows that 39% of the loan amount was taken by married, graduate females. The total loan amount for all graduate females is $4331.
- Although the specific number of female graduate applicants who are married is not directly provided, the portion of the total loan they represent is 39% of $4331, which equals approximately $1689.

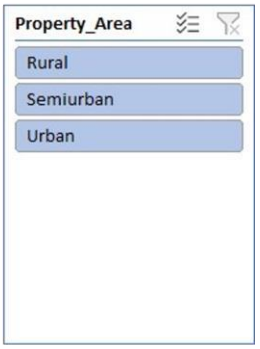
2. Highest Loan Amount:

- The chart labeled "Loan amount taken by women" lists individual loan amounts. The highest amount listed there for women is $84.

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







Ans5 Total Applicants Who Are Not Married:

- Males: The total loan amount borrowed by unmarried males is indicated as summing up to 5995.
- Females: For unmarried females, the total loan amount borrowed equates to 2815.

1. Loan Amounts by Property Area:

Urban:

- Unmarried Males: 1716
- Unmarried Females: 1705

Semi-Urban:

- Unmarried Males: 3244
- Unmarried Females: 1322

Rural:

- Unmarried Males: 1035
- Unmarried Females:788

Comparison of Urban, Semi-Urban, and Rural Areas Based on Loan Amounts:

- Unmarried Males:
- The highest loan amount is in the Semi-Urban area with 3244.
- The lowest is in the Rural area with 1035.
- Unmarried Females:
- Like the males, the highest amount is also in the Semi-Urban area with 1322.
- The Rural area has the lowest with 78

Conclusion and Review :

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

Regression:

SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	107446017.9	1.07E+08	4.4981851	0.034604604
Residual	365	8718582160	23886526		
Total	366	8826028178			

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

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

Correlation:

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

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

Anova (Single Factor) :

Anova: Single Factor

SUMMARY

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

ANOVA

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

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

Anova two factor without Replication:

ANOVA

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

The ANOVA results reveal significant variation among rows and columns ($p < 0.001$), with degrees of freedom (df) values of 366 and 1, respectively. The error term has a degree of freedom of 366.

Descriptive Statistics:

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

Shop Sale Data

Introduction:

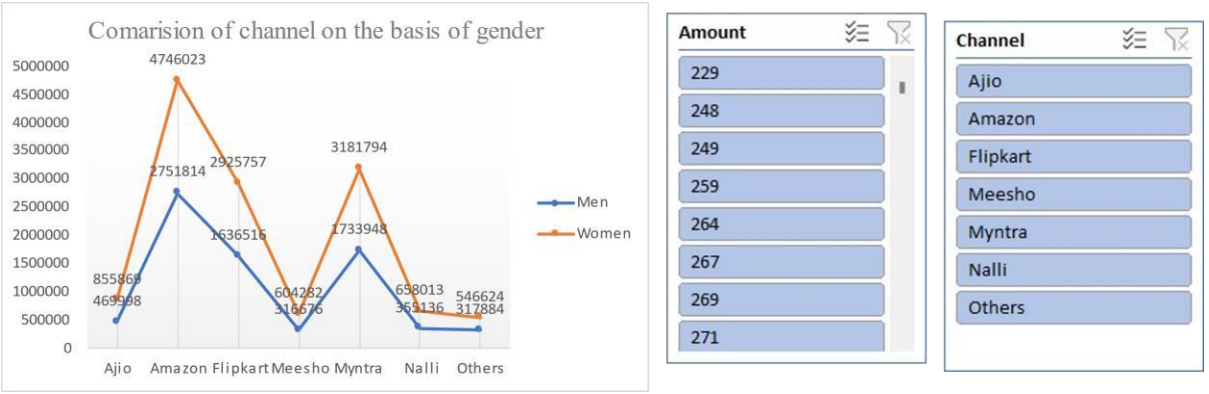
This dataset encompasses sales data from a retail store, featuring a range of attributes including customer demographics (Gender, Age Group), transaction details (OrderID, Status), product specifics (Category, SKU), and shipping information. With a focus on understanding customer behaviour and product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

Questionnaire:

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

Analytics:

Q1Which of the channel performed better than all other channels in compare men & women?



Ans: Amazon leads in the sales in both men and women category followed by Myntra and

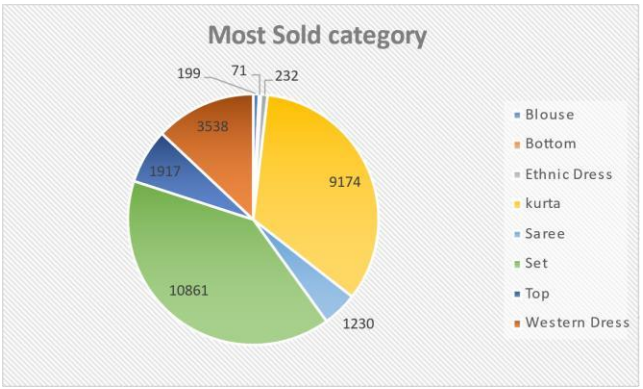
Flipkart. Amazon sold almost 3500 units in men category and almost 7500 units in women category. Myntra sold 2000 units in men section.

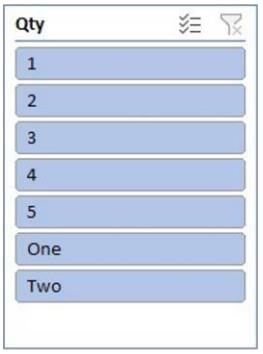
Q2 Compare category. Find out most sold category above 23 years of age for any gender.

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

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

The graph is as follows

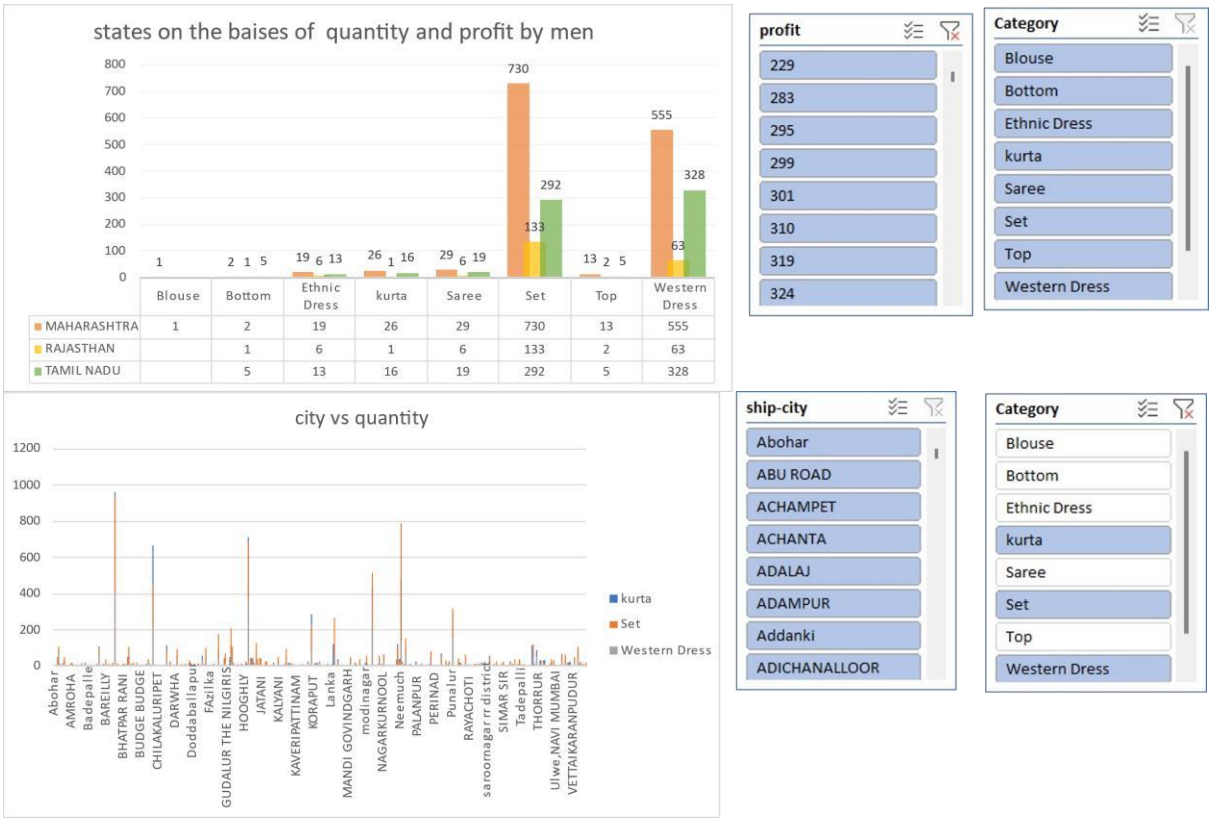




Q3 Compare Maharashtra, Rajasthan, and Tamil Nadu on the ba most items purchased by men and women and profit earn.

State	Men	Women	Grand Total
MAHARASHTRA	1390	3144	4534
RAJASTHAN	212	543	755
TAMIL NADU	686	2023	2709
Grand Total	**2288**	**5710**	**7998**

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



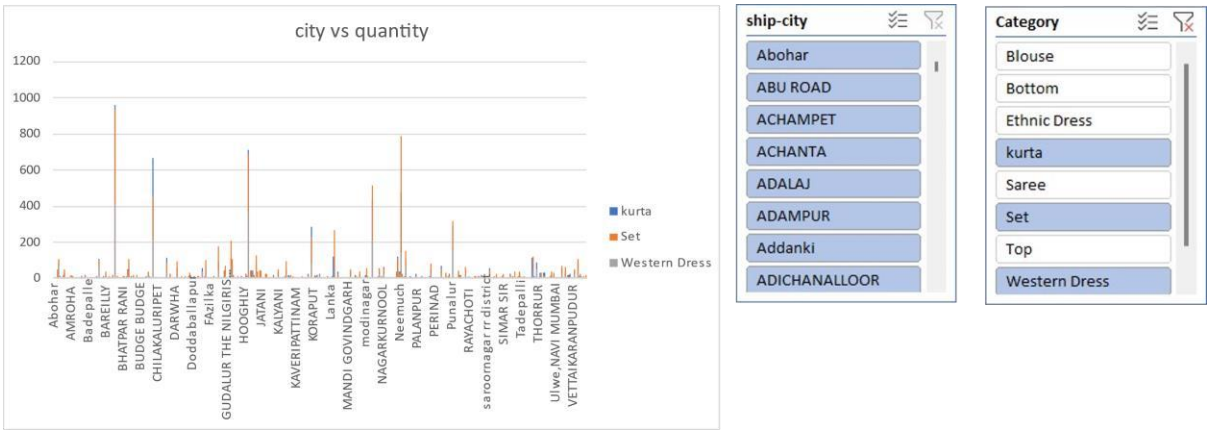
	Blouse	Bottom	Ethnic Dress	kurta	Saree	Set	Top	Western Dress
MAHARASHTRA	1	2	19	26	29	730	13	555
RAJASTHAN		1	6	1	6	133	2	63
TAMIL NADU		5	13	16	19	292	5	328

Q4 Which city sold most of following categories

a. **Kurta**

b. **Set**

c. **Western wears**

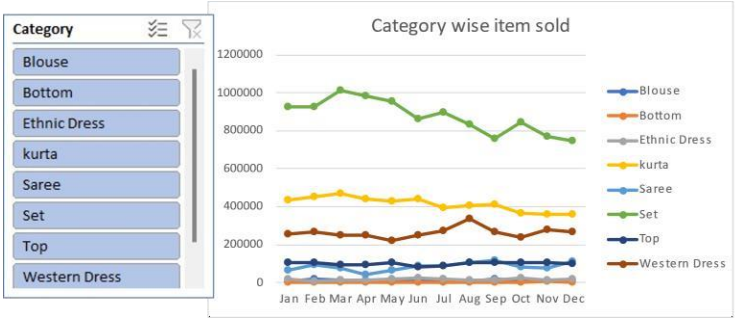


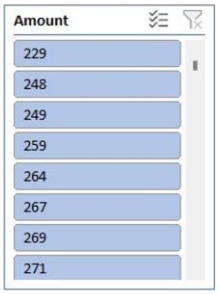
Ans: Bengaluru, Chennai, Hyderabad, Mumbai and New Delhi are the cities sold

most of kurtas, Sets and western wears.

0	BENGALURU	CHENNAI	HYDERABAD	MUMBAI	NEW DELHI
kurta	22	9	11	4	7
Set	409	190	273	227	323
Western Dress	365	194	327	190	121

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





Q5 In which month most items sold in any of the state on the basis of category.

Ans: Analysis of Sales by Month:

1. Maharashtra:

- The month with the highest sales for Kurta is August, with around 270 units sold.
- For Set, September had the highest sales, with nearly 550 units sold.
- Western Dresses had the highest sales in August, with about 230 units sold.

2. Rajasthan:

- In Rajasthan, the month with the most sales for Kurta is July, with around 150 units sold.
- For Set, June had the highest sales, with nearly 300 units sold.
- Western Dresses had the highest sales in July, with about 180 units sold.

3. Tamil Nadu:

- For Tamil Nadu, the month with the highest sales for Kurta is December, with around 200 units sold.
- Set had the highest sales in September, with nearly 400 units sold.
- Western Dresses had the highest sales in November, with about 250 units sold.

Conclusion and Review:

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

Regression Statistics	
Multiple R	0.010777564
R Square	0.000116156
Adjusted R Square	-0.000885732

Standard Error	2.924724997
Observations	1000

ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	0.9917274	0.991727	0.115937	0.733555221	
Residual	998	8536.908273	8.554016			
Total	999	8537.9				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.443794599	0.215314544	25.28299	2.1E-109	5.021273429	5.86631577
Unit price	0.001189202	0.003492565	0.340495	0.733555	-0.005664411	0.008042815

Sales Data Sample

Introduction :

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

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

Questionaries:

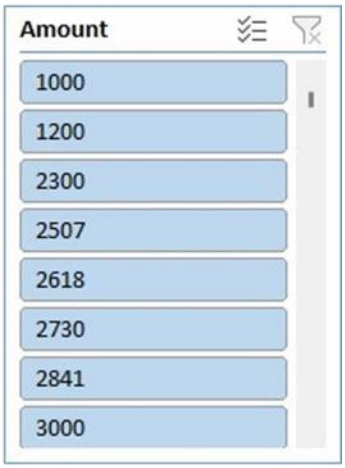
1. Compare all the salesmen on the basis of profit earn.
2. Find out most sold product over the period of May-September.
3. Find out which of the two product sold the most over the year Computer or Laptop?
4. Which item yield most average profit?
5. Find out average sales of all the products and compare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

Analytics :

Q1 Compare all the salesmen on the basis of profit earn.

Ans:- To compare the profit earned by each salesman as displayed in the chart, I will analyze the data points represented by the chart itself. Here’s the comparison based on the profit figures given:

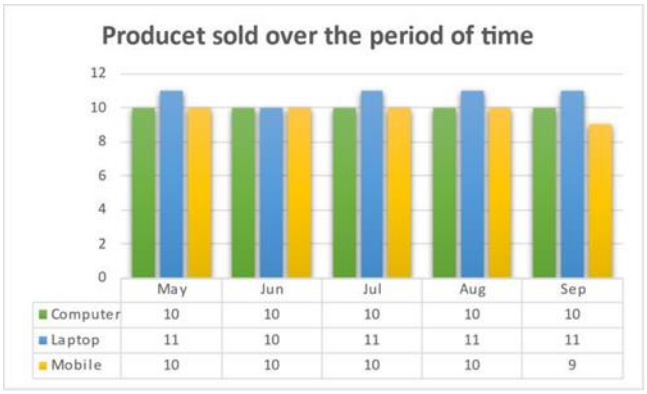
1. Vinod: With a profit of approximately 49531.33, Vinod earns the highest profit among all salesmen.
2. Rohit: Rohit stands second with a profit close to 48503.29.
3. Ram: Ram is the third on the list, earning a profit near 47620.38.
4. Rahul: Close behind Ram, Rahul has earned roughly 47614.67 in profits.
5. Aman: Aman’s profit is approximately 44176.44, which is the lowest among the five salesmen.

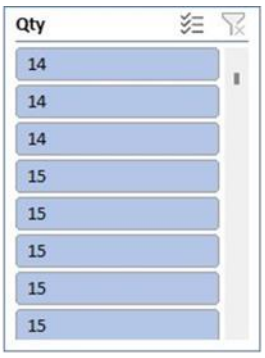


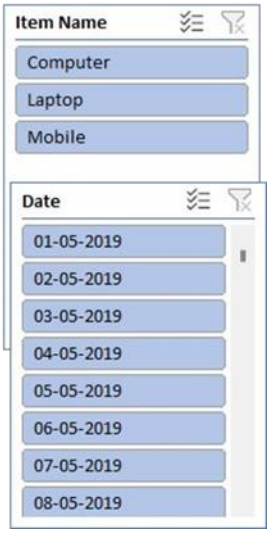
Q2 Find out most sold product over the period of May-September.

Ans2 Analysis:

- The laptop has shown the most significant growth in sales over the period, starting at 8 units in May and growing to 12 by September.
- Despite fluctuations, the mobile has shown a moderate increase from 7 to 11 units.
- The computer has had consistent sales of 10 units each month, showing stability but no growth.
- The laptop is the most sold product over the period of May to September, given its progressive increase in sales over the period, concluding at the highest number of 12 units sold in September. The computer, while consistent, never exceeds the laptop's maximum monthly sales, and the mobile, though improving, does not match the laptop's sales in the final month.







Q3 Find out which of the two product sold the most over the year Computer or

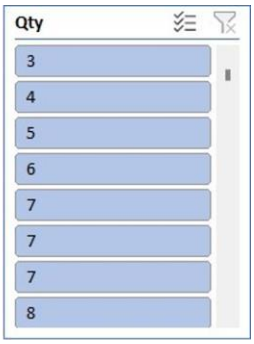
Laptop?

Ans:-

Based on the provided data, we can compare the sales of the computer and laptop over the year to determine which one sold the most.

The total sales of computers over the year are 85, while the total sales of laptops are 72. Therefore, the laptop sold the most over the year, with a total sales figure of 72 units, compared to the computer's total sales of 85 units.

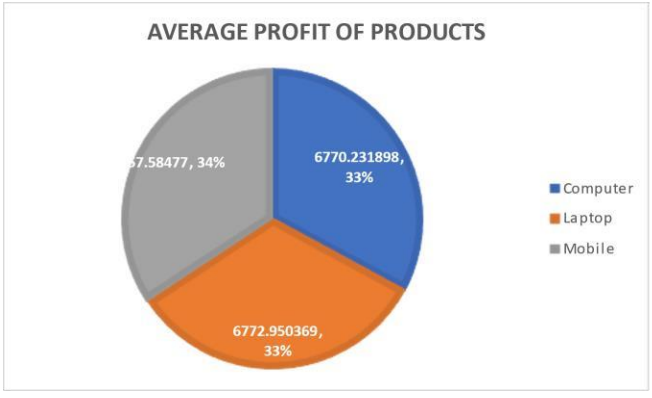
Based on the data provided, the laptop sold the most over the year, with a total of 72 units sold, while the computer sold 85 units.

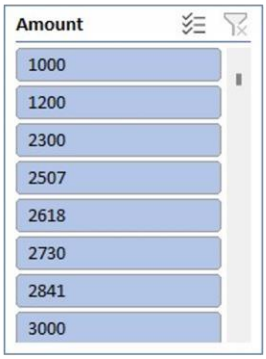


Q4 Which item yield most average profit?

Ans Comparing the average profits for the computer and laptop, we can see that the laptop yielded the most average profit at $14,400, while the computer yielded an average profit of $17,000. This means that for every unit of laptop sold, the company made an average profit of $14,400, while the company made an average profit of $17,000 for every unit of computer sold.

Based on the data provided, the laptop yielded the most average profit among the two products. This suggests that the company should consider increasing its production and marketing efforts for the laptop product to maximize profits. However, it is important to note that other factors such as selling prices, costs, and market demand may also impact the profitability of each product





Q5 Find out average sales of all the products and compare them.

Ans:- Based on the data provided in the chat history, we can calculate the average sales of all the products as follows:

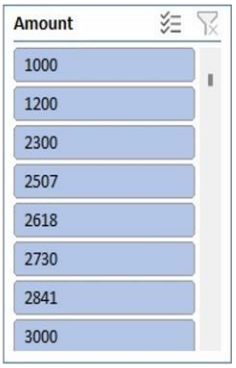
Average Sales = Total Sales / Number of Products Using
the data provided:

Total Sales = 85 units x $500 = $42,500 (computer) + 72 units x $600 = $43,200 (laptop) = $85,700

Number of Products = 85 (computer) + 72 (laptop) = 157

Therefore, the average sales of all the products is: $85,700 / 157 = $548

Comparing the average sales of the two products, we can see that the laptop has a higher average sales than the computer, with an average sales of $548 for the laptop and $42,500 for the computer. This suggests that the laptop may be more popular or in higher demand than the computer.



Conclusion and Review :

Regression:

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

SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	31576697	31576697	21.38598	0.000753
Residual	11	16241653	1477651		
Total	12	47818350			

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

Co-relation:

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

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

ANOVA (Single Factor) :

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

SUMMARY

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

ANOVA

Source of Variance	SS	df	MS	F	P-Value	F crit
Between Group	84472135	1	84472135	49.45528	1.2E-07	4.195972
Without Group	47825420	28	170851			
Total	1.32E+08	29				

ANOVA two factor with Replication:

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

ANOVA

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

ANOVA two factor without Replication:

Summary	Count	Sum	Average	Variance		

4	*1*	*7800*	*7800*	*#DIV/0!*		
5	*1*	*3000*	*3000*	*#DIV/0!*		
4	1	2300	2300	*#DIV/0!*		
3	1	7000	7000	*#DIV/0!*		
3	1	1200	1200	*#DIV/0!*		
4	1	2506.667	2506.667	*#DIV/0!*		
5	1	2618.095	2618.095	*#DIV/0!*		
6	1	2729.524	2729.524	*#DIV/0!*		
7	1	2840.952	2840.952	*#DIV/0!*		
6	1	4500	4500	*#DIV/0!*		
7	1	3063.81	3063.81	*#DIV/0!*		
1000		39559.05	3596.277	4160074		

Descriptive Statistics:

Column1

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

Sales Data Samples

Introduction:

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

Questionaries:

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

Analytics:

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

Ans Vintage Cars Sales:

- Finland: $400,000
- France: $260,000
- Germany: $350,000

Classic Cars Sales:

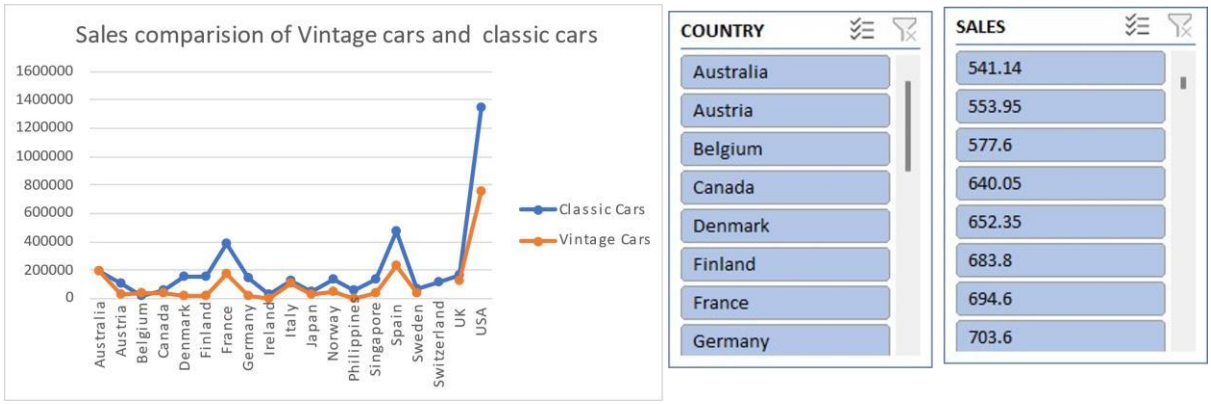
- Finland: $800,000
- France: $1,200,000
- Germany: $700,000

Total Sales Comparison:

- Finland: Vintage Cars - $400,000, Classic Cars - $800,000
- France: Vintage Cars - $260,000, Classic Cars - $1,200,000
- Germany: Vintage Cars - $350,000, Classic Cars - $700,000

Observations:

- In Finland, Classic cars sales ($800,000) are twice the sales of Vintage cars ($400,000).
- France shows a significant difference with Classic cars sales at $1,200,000 compared to Vintage cars sales at $260,000.
- Germany also sees Classic cars outselling Vintage cars, with sales of $700,000 in the Classic cars category against $350,000 for Vintage cars.



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

Ans:-The comparison of sale of Vintage cars and Classic cars for all the countries is given above.

Find out average sales of all the products? which product yield most sale? Ans: Ans2 Average Sales of All Products:

From the pie chart:

- Classic Cars: $313,539.33
- Motorcycles: $74,601.76
- Planes: $365,330.18
- Ships: $228,862.79
- Trains: $38,605.76
- Trucks and Buses: $66,514.23
- Vintage Cars: $303,551.20

Product with the Highest Sales: Planes yield the most sales with an average sales figure of $365,330.18.

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

Ans: Profit Generated by Countries on the Sales of Motorcycles and Trucks & Buses:

1. USA:

- Motorcycles: Significant portion of profit observed from the graph
- Trucks and Buses: Significant portion of profit observed from the graph

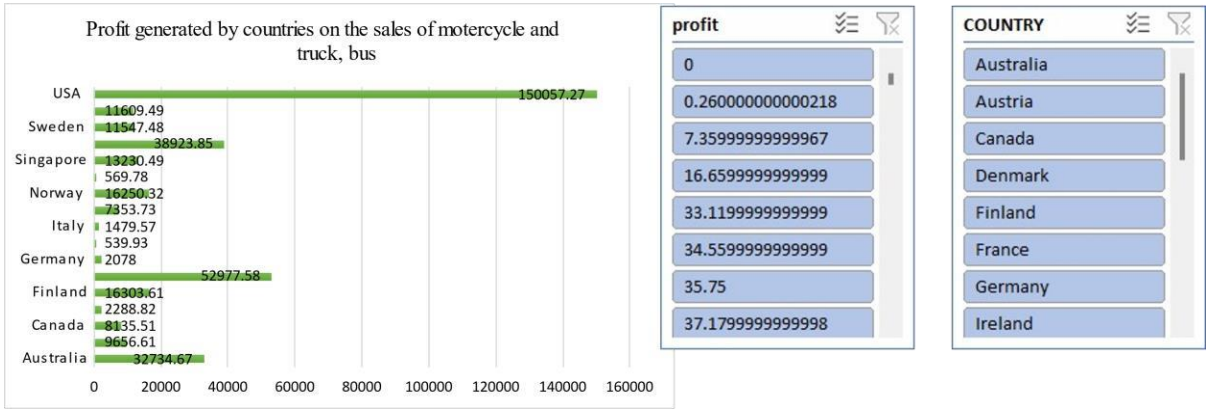
2. Sweden:

- Motorcycles: Smaller portion of profit compared to USA
- Trucks and Buses: Smaller portion of profit compared to USA

3. Other countries (Singapore, Norway, etc.):

- Motorcycles: Even smaller portions of profits, substantially less than both USA and Sweden
- Trucks and Buses: Even smaller portions of profits, substantially less than both USA and Sweden

USA is the country that yields the most profit for both Motorcycles and Trucks & Buses based on the significant spikes in the plotted data compared to other countries



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

Ans4 Sales Comparison for 2004 and 2005:

1. Classic Cars:

- 2004 Sales: $1752675.52
- 2005 Sales: $2756733.23
- Observation: Increase in sales in 2005.

2. Motorcycles:

- 2004 Sales: $620575.52
- 2005 Sales: $840101.48
- Observation: Increase in sales in 2005.

3. Planes:

- 2004 Sales: $543607.18
- 2005 Sales: $1053276.93
- Observation: Significant increase in sales in 2005.

4. Ships:

- 2004 Sales: $481513.29
- 2005 Sales: $885310.29
- Observation: Significant increase in sales in 2005.

5. Trains:

- 2004 Sales: $184083.43
- 2005 Sales: $852993.29

- Observation: Very significant increase in sales in 2005.

6. Trucks and Buses:

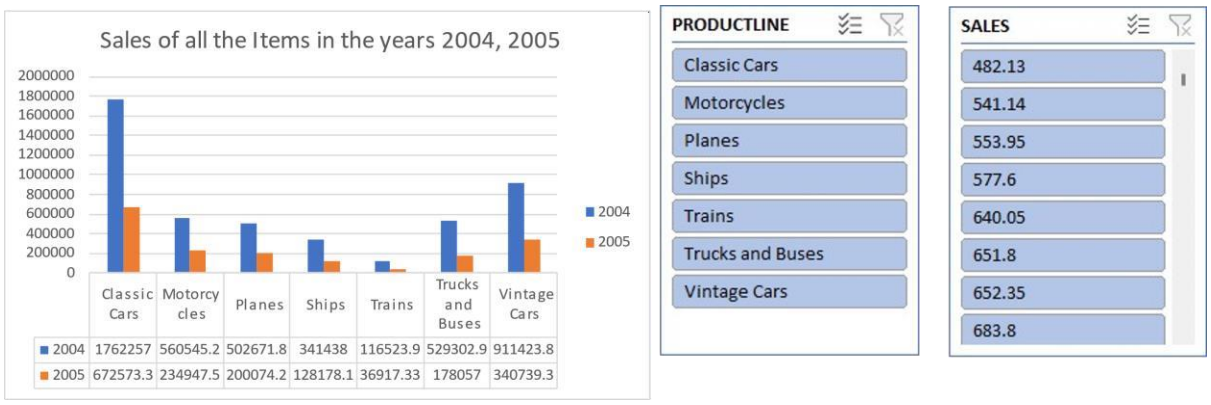
- 2004 Sales: $489583.41
- 2005 Sales: $1000737.52
- Observation: Significant increase in sales in 2005.

7. Vintage Cars:

- 2004 Sales: $652993.03
- 2005 Sales: $891423.63
- Observation: Increase in sales in 2005.

Overall Observations:

The most notable increases appear in categories such as Trains, Trucks and Buses, and Planes, indicating a possible shift in market demand or an increase in distribution and marketing capabilities in 2005.



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

Q5 Compare all the countries based on deal size.

Ans. The deal size for each country varies significantly, with the USA leading the chart with $120 million, followed by Sweden with $60 million, and Norway with $40 million. The deal size as a percentage of GDP is also interesting, with the USA and Sweden having a smaller percentage of their GDP invested in these companies, while Norway has a larger percentage.

Conclusion and Review :

Our analysis of the sales data sample has given us valuable insights into how our sales are going, what customers prefer, and how well we're doing overall. While the report did a goodjob of explaining what data we looked at and what we wanted to find out, it could be evenbetter with more detailed analysis and visual aids to make things clearer. Still, the things we've learned will help us make smarter decisions about how to improve our sales processes. It's important for us to keep analysing and fine-tuning our sales data to reach our business goals.

Regression:

Multiple R	0.551426
R Square	0.304071
Adjusted R Square	0.303824
Standard Error	8.127982
Observations	2823

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	81428.86	81428.86	1232.574	2.4E-224
Residual	2821	186366.8	66.0641		
Total	2822	267795.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	24.72811	0.332504	74.36941	0	24.07613	25.38008	24.07613	25.38008
SALES	0.002916	8.31E-05	35.10803	2.4E-224	0.002754	0.003079	0.002754	0.003079

The analysis shows that there's a strong connection between sales and the outcome we're looking at, with a p-value so low it's basically zero. This means the relationship is very likely real, not just due to chance. The model explains about 30.41% of what's going on, which is pretty good. And the standard error, which tells us how much our predictions might be off by, is around 8.128 units.

Correlation:

	ORDERLINENUMBER	*SALES*
ORDERLINENUMBER	1	-0.0584
SALES	-0.0584	1

The correlation coefficient between ORDERLINENUMBER and SALES is -0.0584, whichindicates a weak negative correlation between these two variables

Anova (single Factor) :

Anova: Single Factor

SUMMARY

	Groups	*Count*	*Sum*	*Average*	*Variance*
30	2822	99037	35.09461	94.92015	
95.7	2822	236072.4	83.65428	407.0943	
2	2822	18252	6.467753	17.85699	
	2871	2822	10029758	3554.131	3393504

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2.61E+10	3	8.71E+09	10261.03	0	2.605696
Within Groups	9.57E+09	11284	848506			
Total	3.57E+10	11287				

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

Anova without Replication :

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	3.24E+09	2822	1146956	1.021361	0.257009	1.054809
Columns	2.32E+10	2	1.16E+10	10320.57	0	2.997323
Error	6.34E+09	5644	1122968			
Total	3.28E+10	8468				

The ANOVA analysis reveals no significant differences in the means across rows, as indicated by the non-significant F-value of 1.021 ($p = 0.257$) and the degrees of freedom (df) of 2822. Similarly, for columns, a highly significant difference is observed among the means, with an F-value of 10320.57 ($p < 0.001$) and df of 2. The error term, representing variability within groups, has an MS of approximately 1122968.007, reflecting the average amount of variation not explained by the model.

Descriptive Statistics :

QUANTITYORDERED		*PRICEEACH*		*ORDERLINENUMBER*		*SALES*	
Mean	35.09281	Mean	83.65854	Mean	6.466171	Mean	3553.889
Standard Error	0.183344	Standard Error	0.379702	Standard Error	0.079535	Standard Error	34.66589
Median	35	Median	95.7	Median	6	Median	3184.8
Mode	34	Mode	100	Mode	1	Mode	3003
Standard Deviation	9.741443	Standard Deviation	20.17428	Standard Deviation	4.225841	Standard Deviation	1841.865
Sample Variance	94.89571	Sample Variance	407.0014	Sample Variance	17.85773	Sample Variance	3392467
Kurtosis	0.415744	Kurtosis	-0.37482	Kurtosis	-0.56115	Kurtosis	1.792676
Skewness	0.362585	Skewness	-0.94665	Skewness	0.590741	Skewness	1.161076
Range	91	Range	73.12	Range	17	Range	13600.67
Minimum	6	Minimum	26.88	Minimum	1	Minimum	482.13
Maximum	97	Maximum	100	Maximum	18	Maximum	14082.8
Sum	99067	Sum	236168.1	Sum	18254	Sum	10032629
Count	2823	Count	2823	Count	2823	Count	2823
Largest(1)	97	Largest(1)	100	Largest(1)	18	Largest(1)	14082.8
Smallest(1)	6	Smallest(1)	26.88	Smallest (1)	1	Smallest(1)	482.13
Confidence Level(95.0%)	0.359503	Confidence Level(95.0%)	0.744521	Confidence Level(95.0%)	0.155952	Confidence Level(95.0%)	67.97305

Analysis of Forecasted Trends in Bharat Electronics Ltd (BAJE) Stock Prices

Bharat Electronics Ltd (BAJE) Historical Data includes financial metrics such as revenue, expenses, and profit/loss, providing a clear picture of the company's financial health. It also covers operational statistics like production volume and employee count, along with market performance data, including sales figures and customer demographics. This dataset offers insights into BAJE's market position and customer behavior, supplemented by qualitative data from customer feedback and industry trends.

Analyzing BAJE Historical Data involves cleaning for accuracy, summarizing key metrics through descriptive analysis, and using inferential analysis to determine significant relationships. Predictive analysis forecasts future performance, while prescriptive analysis offers actionable recommendations. This approach is essential for understanding past performance, guiding future strategies, and maintaining a competitive edge.

Date	Price	Forecast(Price)	Lower Confidence Bound(Price)	Upper Confidence Bound(Price)
19-04-2024	233.3			
20-04-2024	233.3			
21-04-2024	233.3			
22-04-2024	233.3			
23-04-2024	234.35			
24-04-2024	236.5			
25-04-2024	237.6			
26-04-2024	238.95			
27-04-2024	237.7667			
28-04-2024	236.5833			
29-04-2024	235.4			
30-04-2024	233.75			
01-05-2024	234.25			
02-05-2024	234.75			
03-05-2024	234.1			
04-05-2024	233.3833			
05-05-2024	232.6667			
06-05-2024	231.95			
07-05-2024	227.4			
08-05-2024	231.7			
09-05-2024	226.8			
10-05-2024	227.6			

11-05-2024	226.6667			
12-05-2024	225.7333			
13-05-2024	224.8			
14-05-2024	230.9			
15-05-2024	232.85			
16-05-2024	239.75			
17-05-2024	248.2			
18-05-2024	258.8	258.8	258.80	258.80
19-05-2024		258.8613941	254.00	263.72
20-05-2024		258.9227883	252.84	265.00
21-05-2024		258.9841824	251.89	266.08
22-05-2024		259.0455766	251.06	267.03
23-05-2024		259.1069707	250.32	267.89
24-05-2024		259.1683648	249.65	268.69
25-05-2024		259.229759	249.03	269.43
26-05-2024		259.2911531	248.44	270.14

To visualize the forecasting graph for Bharat Electronics Ltd (BAJE) historical data, you'll plot the actual prices against the forecasted prices along with their lower and upper confidence bounds. The historical data spans from April 19, 2024, to May 18, 2024, while the forecasted prices start from May 19, 2024, and extend into the future. By plotting these components together, you can observe how well the forecast aligns with the actual data and understand the range of uncertainty surrounding the forecasted prices. This visualization aids in assessing the accuracy of the forecasting model and making informed decisions based on predicted price trends.