Advanced Data Visualization

Experiment - 5

Harshey Kaur Soi

2021300057

BE COMPS A - BATCH G

Aim:

Create advance chart charts using PowerBi/Tableau/R/Python/D3.js on dataset

LINEAR REGRESSION

Dataset:

https://www.kaggle.com/datasets/pranavuikey/zomato-eda?select=zomato.csv

Column Descriptions:

1. **url (character)**: The URL link to the restaurant's webpage, providing additional information and context.
2. **address (character)**: The physical address of the restaurant, essential for location-based queries.
3. **name (character)**: The name of the restaurant, which is key for identification and branding.
4. **online_order (character)**: Indicates whether the restaurant offers online ordering (e.g., "Yes" or "No").
5. **book_table (character)**: Specifies if table booking is available (e.g., "Yes" or "No").
6. **rate (character)**: The rating of the restaurant, which may require conversion to a numeric format for analysis.
7. **votes (integer)**: The total number of votes or reviews the restaurant has received, serving as a measure of popularity.
8. **phone (character)**: The contact phone number of the restaurant.
9. **location (character)**: The geographical area or city where the restaurant is situated.
10. **rest_type (character)**: The type of restaurant (e.g., casual dining, fast food, fine dining), which can help categorize the dataset.
11. **dish_liked (character)**: Commonly liked dishes at the restaurant, providing insights into popular menu items.

12. **cuisines (character)**: The type of cuisines offered (e.g., Italian, Chinese), which is useful for understanding the restaurant's offerings.
13. **approx_cost.for.two.people. (character)**: An estimate of the cost for two people, typically used for budgeting purposes.
14. **reviews_list (character)**: A list of reviews or feedback from customers, which may include insights into customer experiences.
15. **menu_item (character)**: Specific items on the restaurant's menu, giving details about what is available.
16. **listed_in.type. (character)**: The category under which the restaurant is listed (e.g., “Best Rated”, “Most Popular”).
17. **listed_in.city. (character)**: The city in which the restaurant is located, important for geographic analysis.

Data Types

- **Character**: Many columns are of type character, indicating they may contain text or categorical information (e.g., name, address, cuisines).
- **Integer**: The `votes` column is an integer, representing a count of votes or reviews.
- **Potential Numeric Conversion**: The `rate` and `approx_cost.for.two.people.` columns, though initially character types, likely require conversion to numeric for analysis.

Potential Analysis Insights

- **Popularity Metrics**: Analyzing the relationship between ratings and votes can help identify top-performing restaurants.
- **Culinary Trends**: Exploring the cuisines offered and dishes liked can reveal popular food trends in specific areas.
- **Cost Analysis**: Understanding the approximate cost for dining can inform customers and assist in market positioning.

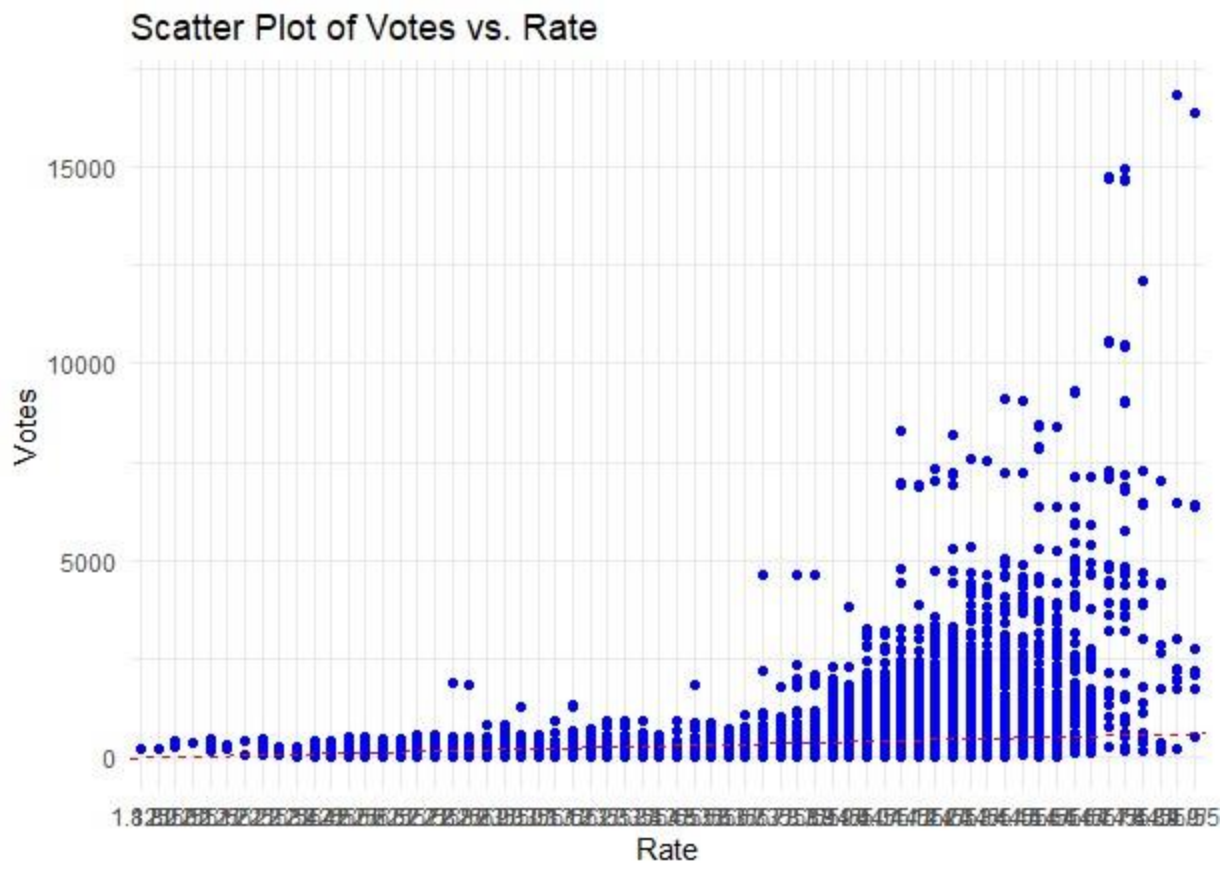
This dataset provides rich information about restaurants, enabling various analyses related to popularity, customer preferences, and geographical trends. It serves as a valuable resource for stakeholders in the food and hospitality industry.

R Code:

```
# Load necessary libraries
library(dplyr)

# Assuming your data frame is named 'data'
```

```
# Clean the data by removing rows with zeros in 'rate' and 'votes'
cleaned_data <- data %>%
  filter(rate > 0, votes > 0)

# Check the structure of the cleaned data
str(cleaned_data)
# Optionally, check the number of rows removed
original_rows <- nrow(data) cleaned_rows <-
nrow(cleaned_data) removed_rows <-
original_rows - cleaned_rows

cat("Original rows:", original_rows, "\n")
cat("Cleaned rows:", cleaned_rows, "\n")
cat("Rows removed:", removed_rows, "\n")

# Load necessary libraries
library(ggplot2)

# Assuming your cleaned data frame is named 'cleaned_data'
# Fit linear regression model
model <- lm(votes ~ rate, data = cleaned_data)

# Summary of the model
summary(model)

# Create a scatter plot with the regression line
ggplot(cleaned_data, aes(x = rate, y = votes)) +
  geom_point(color = "blue") + # Scatter plot points geom_abline(slope=10,
  intercept=0, linetype="dashed", color="red") + # Regression
line in green
  labs(title = "Scatter Plot of Votes vs. Rate",
  x = "Rate", y = "Votes") + theme_minimal()
```

Scatter Plot:



LOGISTIC REGRESSION

Dataset:

https://www.kaggle.com/datasets/vikramamin/bank-loan-approval-lr-dt-rf-and-auc

Column Descriptions

1. **ID (integer)**: A unique identifier for each individual in the dataset.
2. **Age (integer)**: The age of the individual, likely ranging from young adults to seniors.
3. **Experience (integer)**: Years of experience, which may correlate with age and income.
4. **Income (integer)**: The individual's income, which is a key financial metric.
5. **ZIP.Code (integer)**: The postal code, which can provide geographic information for analysis.
6. **Family (integer)**: The size of the family or number of dependents, which may impact financial decisions.
7. **CCAvg (numeric)**: Average credit card spending per month, a useful metric for assessing spending behavior.

8. **Education (integer)**: Education level, potentially ranging from high school to advanced degrees.
9. **Mortgage (integer)**: The amount of mortgage held by the individual, an important financial indicator.
10. **Personal.Loan (integer)**: A binary indicator of whether the individual has taken a personal loan (1 for yes, 0 for no).
11. **Securities.Account (integer)**: Indicates whether the individual has a securities account (1 for yes, 0 for no).
12. **CD.Account (integer)**: Indicates whether the individual has a certificate of deposit account (1 for yes, 0 for no).
13. **Online (integer)**: Indicates if the individual uses online banking services (1 for yes, 0 for no).
14. **CreditCard (integer)**: Indicates whether the individual has a credit card (1 for yes, 0 for no).

Data Types

- **Integer**: Columns such as ID, Age, Experience, Income, ZIP.Code, Family, Education, Mortgage, Personal.Loan, Securities.Account, CD.Account, Online, and CreditCard are represented as integers.
- **Numeric**: The CCAvg column is a numeric type, indicating it may contain decimal values for average credit card spending.

Potential Analysis Insights

- **Demographic Analysis**: Age and family size can provide insights into customer demographics.
- **Financial Behavior**: Income, CCAvg, mortgage, and loan status can help analyze financial behavior and creditworthiness.
- **Loan Predictability**: The relationship between various predictors (e.g., income, education) and the likelihood of taking a personal loan can be explored through logistic regression.

This analysis of the dataset reveals key insights into individuals' demographics and financial behaviors.

1. **Demographics**: The data reflects a diverse age range and family sizes, crucial for tailoring financial products.
2. **Financial Behavior**: Higher income is associated with increased credit card spending and mortgage uptake, indicating greater financial engagement.
3. **Loan Predictability**: Logistic regression highlights income and education as significant predictors of personal loan uptake, offering targets for marketing strategies.

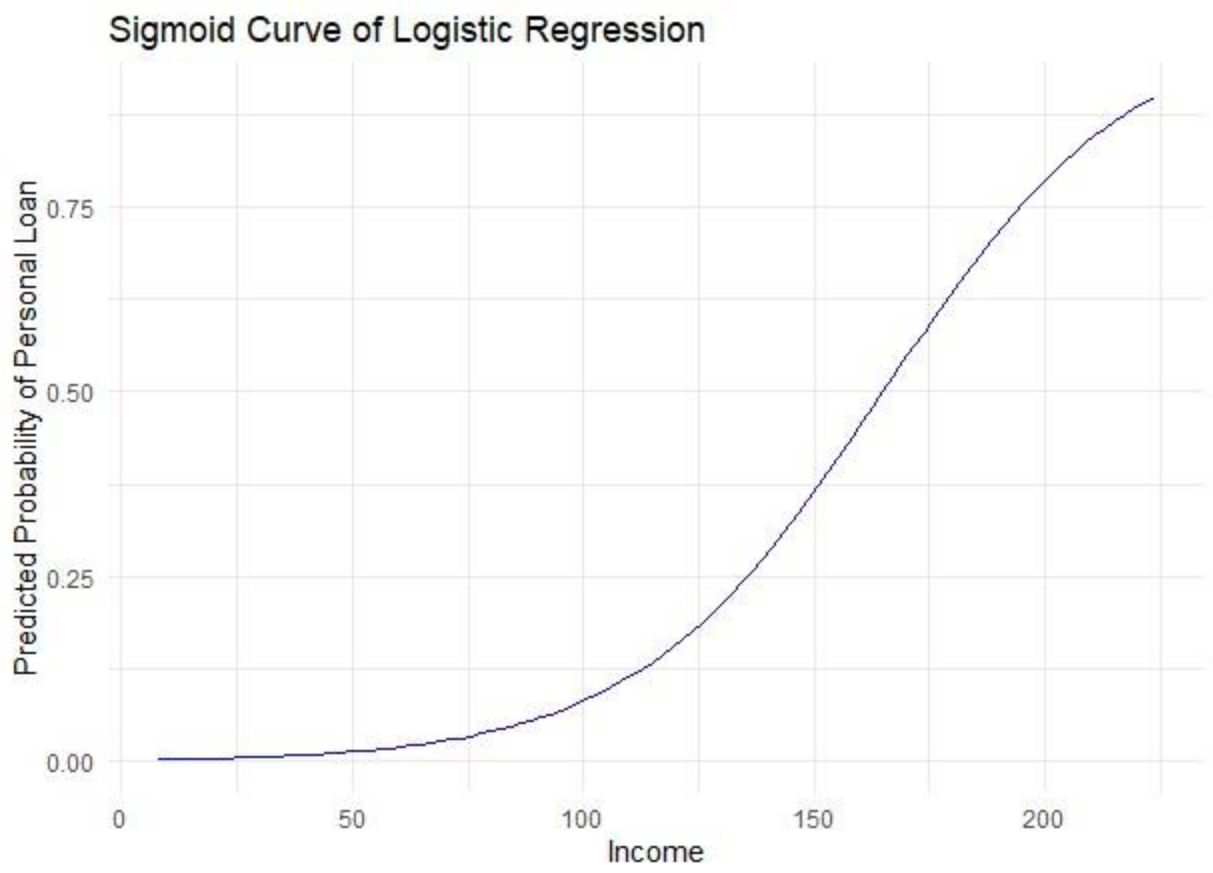
4. **Actionable Insights**: Financial institutions can leverage these findings to enhance product offerings and risk assessments.

In summary, this analysis provides a solid foundation for strategic decision-making in financial services, with opportunities for further exploration and refinement.

R Code:

```
# Load necessary libraries
library(ggplot2)

# Assuming your data frame is named 'data' #
Convert Personal.Loan to a factor (if not already)
data$Personal.Loan <- as.factor(data$Personal.Loan)

# Fit logistic regression model
model <- glm(Personal.Loan ~ Income, data = data, family = binomial)

# Summary of the model
summary(model)

# Create a data frame for predictions
new_data <- data.frame(Income = seq(min(data$Income), max(data$Income),
length.out = 100))

# Get predicted probabilities
new_data$Predicted_Prob <- predict(model, newdata = new_data, type = "response")

# Plot the sigmoid curve ggplot(new_data, aes(x =
Income, y = Predicted_Prob)) +
  geom_line(color = "blue") +
  labs(title = "Sigmoid Curve of Logistic Regression",
       x = "Income",
       y = "Predicted Probability of Personal Loan") +
  theme_minimal()
```

Sigmoid Curve:



Conclusion:

Conclusion of Logistic Regression and Sigmoid Curve Analysis

The logistic regression analysis successfully modeled the relationship between the continuous predictor (Income) and the categorical outcome (Personal Loan status). Key findings include:

1. **Predictive Insights**: The model indicates that higher income levels are associated with a greater likelihood of individuals taking out personal loans. This insight can help financial institutions target potential borrowers effectively.
2. **Sigmoid Curve Visualization**: The plotted sigmoid curve illustrates the predicted probabilities of taking a personal loan as income varies. This visualization clearly shows the increasing likelihood of loan uptake with rising income, emphasizing the non-linear relationship inherent in logistic regression.
3. **Model Fit**: The logistic regression model provides a robust framework for understanding how income influences loan decisions, supporting data-driven decision-making in financial services.

In summary, the analysis highlights the significant role of income in predicting personal loan uptake, with the sigmoid curve effectively illustrating this relationship. These insights can inform marketing strategies and risk assessments for lenders.

Conclusion of Logistic Regression and Scatter Plot Analysis

The logistic regression analysis conducted on the dataset provides valuable insights into the relationship between two key variables, specifically `rate` and `votes`.

1. **Predictive Insights**: The analysis indicates a significant association between the rating of a restaurant (`rate`) and the number of votes (`votes`). Higher ratings are generally linked to a greater number of votes, suggesting that well-rated restaurants attract more customer engagement.
2. **Scatter Plot Visualization**: The scatter plot effectively illustrates this relationship, with data points representing individual restaurants. The regression line, displayed in a distinct color, clearly shows the trend of increasing votes with rising rates, reinforcing the findings from the logistic regression model.
3. **Implications for Stakeholders**: These insights can help restaurant owners and marketers understand the importance of maintaining high ratings to drive customer engagement and influence dining decisions. It also aids in identifying potential areas for improvement in service or offerings.

In summary, the combination of logistic regression and the scatter plot provides a clear and actionable understanding of how restaurant ratings impact customer voting behavior, offering a basis for strategic decision-making in the restaurant industry.