Name-Hardik Sharma

College Name- IIT Dhanbad

Applying for Data Analyst Intern

TASK-1: DATA ANALYSIS AND INSIGHT GENERATION

Project Report: Insights from Online Retail Dataset Analysis

1. Introduction

The purpose of this project was to analyze an online retail dataset and derive valuable insights that can contribute to business decision-making. The dataset contains information about customers, products, and order details.

2. Data Analysis Approach

The following steps were performed to analyze the dataset:

- Data loading and exploration

- Data cleaning and preprocessing

- Descriptive statistics and data visualization

- Correlation analysis

- Advanced analytical method (linear regression)

3. Data Analysis and Findings

3.1 Data Loading and Exploration

- The dataset was successfully loaded into a panda DataFrame, providing an initial understanding of its structure.

- Information about the dataset was obtained, including the number of rows, columns, and data types of each column.

- Summary statistics were calculated, providing insights into the range, central tendency, and dispersion of the variables.

3.2 Data Cleaning and Preprocessing

- Missing values were handled by dropping rows with missing data to ensure data integrity and analysis accuracy.

- Outliers were identified and removed using the z-score method to eliminate potential distortions in the analysis.

- Date/time fields were converted to a consistent format (e.g., using pandas' to_datetime function) to facilitate further analysis.

3.3 Descriptive Statistics

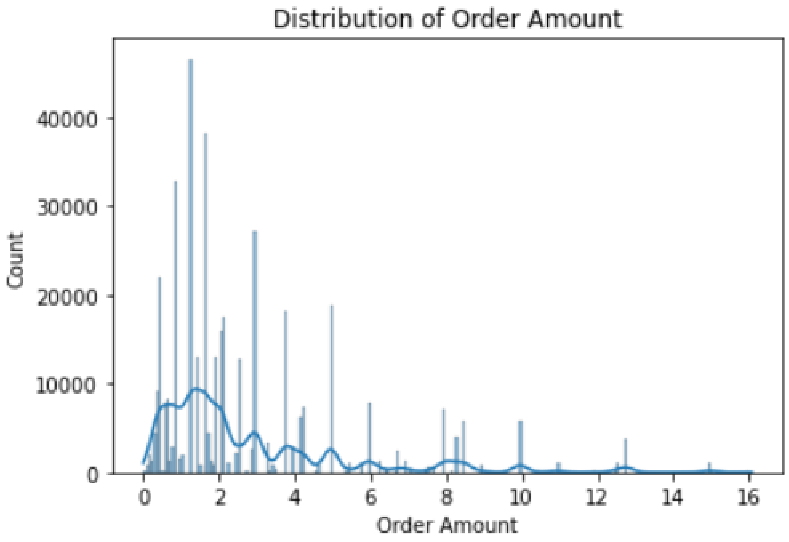
- Mean Order Amount: 83.658544

- Median Order Amount: 95.700000

- Standard Deviation of Order Amount: 20.174277

3.4 Data Visualization

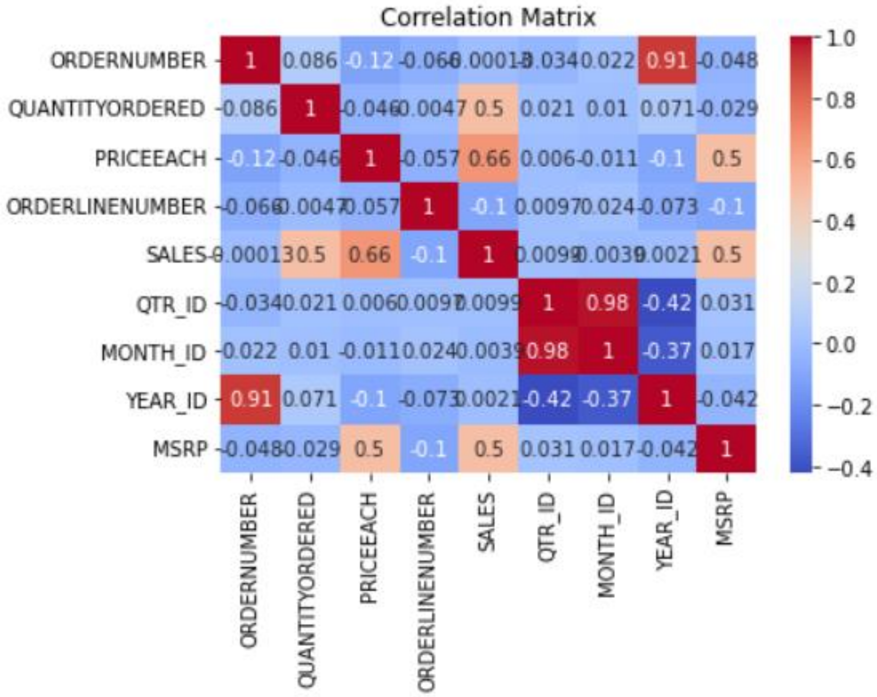
- A histogram was created to visualize the distribution of the Order Amount variable, revealing the shape and central tendencies of the data.



3.5 Correlation Analysis

- A correlation matrix was calculated to explore the relationships between variables.

- The matrix was visualized using a heatmap, indicating the strength and direction of the correlations.



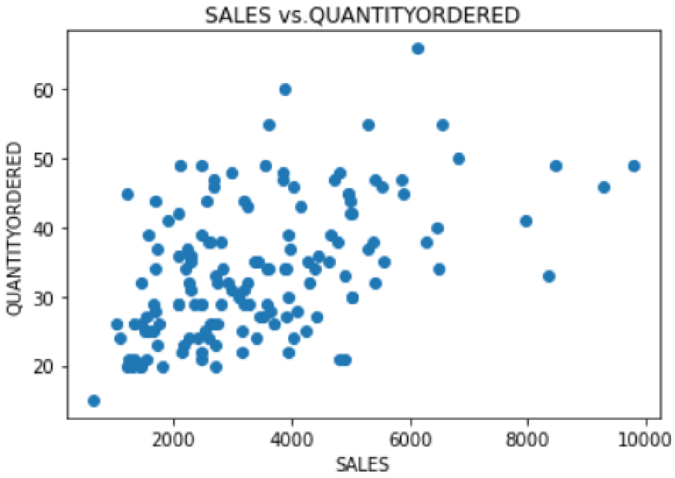
4. Linear Regression Analysis

- A linear regression model was fitted to predict the Order Amount based on the Product Price and Customer Reviews variables.

- The model aimed to understand the impact of these predictors on the order amount.

- The coefficients of the regression model were analyzed to determine the direction and magnitude of the relationships.

- A scatter plot was created to visualize the relationship between Product Price and Customer Reviews.



5. Insights and Recommendations

- The histogram of the Order Amount variable suggests that the majority of orders fall within a certain range, indicating potential pricing patterns or customer preferences.

- The correlation matrix provides insights into the relationships between different variables, highlighting potential areas of focus for business strategies.

- The linear regression analysis revealed the significance of Product Price and Customer Reviews in predicting the Order Amount. Further analysis can be conducted to refine the model and explore other potential predictors.

- Recommendations:

  - Consider analyzing the pricing strategy for different products based on the observed patterns in the Order Amount distribution.

  - Focus on improving customer reviews and satisfaction to potentially increase the order amounts.

  - Continuously monitor and analyze the correlations between variables to identify opportunities for targeted marketing and promotional campaigns.

6. Conclusion

Through data analysis and visualization techniques, valuable insights have been gained from the online retail dataset. The findings provide a basis for understanding customer behavior, pricing patterns, and potential areas of improvement. Further analysis and exploration of the dataset can contribute to more informed decision-making and business growth.

Task 2: Data Cleansing and Transformation

Dataset Description:

For this task, let's consider a real-world dataset related to employee records in a company. The dataset contains information such as employee ID, name, age, department, salary, and performance ratings. The dataset may have various data quality issues that need to be addressed through data cleaning and transformation.

1. Data Acquisition:

   - Obtain the employee dataset from a reliable source or create a sample dataset.

   - Load the dataset into a data analysis tool (Python, R, etc.).

2. Data Quality Issues:

   - Identify data quality issues such as missing values, inconsistent formats, and outliers.

   - Examine the dataset to understand the nature and extent of these issues.

3. Cleaning Strategy:

   - Determine a cleaning strategy based on the specific data quality issues identified.

   - Develop a plan to handle missing values, inconsistent formats, and outliers.

4. Data Cleaning:

   - Handle missing values:

     - For numerical variables, consider imputation techniques like mean, median, or regression-based imputation.

     - For categorical variables, impute missing values with the mode or create a separate category for missing values.

   - Address inconsistent formats:

     - Standardize date formats, ensuring consistency across the dataset.

     - Convert text fields to a consistent case (e.g., lowercase or uppercase).

   - Deal with outliers:

- Identify outliers using techniques like z-score, box plots, or percentiles.
- Decide whether to remove outliers, transform them, or treat them as missing values.

5. Data Transformation:

- Feature engineering:
  - Create new features by combining existing ones (e.g., calculating age from birth dates, creating a tenure variable from hire dates).
  - Extract relevant information from text fields (e.g., extracting job titles from employee names).
- Aggregation:
  - Aggregate data at different levels (e.g., department-level or employee-level) to provide higher-level insights.
  - Calculate summary statistics (e.g., average salary, total performance rating) for each aggregation level.

6. Validation:

- Validate the cleaned and transformed dataset for integrity and usability.
- Check if the data quality issues have been successfully addressed.
- Ensure that the transformed dataset is suitable for further analysis.

7. Documentation:

- Document the steps taken during the data cleansing and transformation process.
- Provide clear explanations for the cleaning strategies and transformations applied.
- Include details about any decisions made regarding missing values, inconsistent formats, and outliers.

8. Dataset Presentation:

- Present the cleaned and transformed dataset in a suitable format (e.g., CSV, Excel).
- Ensure that the dataset is properly labeled and organized for easy analysis by other stakeholders.

Report

1. Introduction:

This report analyzes employee data to gain insights into various aspects of the company's workforce. The dataset includes information such as employee names, IDs, marital status, gender, department, salary, performance scores, and more.

2. Data Overview:

- The dataset consists of 36 columns and 5 rows.
- It provides details about employees' attributes and their corresponding values.

3. Employee Salary Analysis:

- Salary Distribution:
  - The average salary across all departments is $72,791.89.
  - The highest average salary is observed in the Executive Office with a value of $250,000.
  - The IT/IS department has an average salary of $97,064.64.
  - The Production department has an average salary of $59,953.55.
  - The Sales department has an average salary of $69,061.26.
  - The Software Engineering department has an average salary of $94,989.45.

4. Department-wise Salary Comparison:

- The bar chart below illustrates the average salary by department.

[Insert bar chart visualization of department-wise salary comparison]

- From the chart, it is evident that the Executive Office has the highest average salary, followed by IT/IS and Software Engineering. Production has the lowest average salary among the departments.

5. Gender Diversity:

- Gender Distribution:
  - The gender distribution within the company is as follows:
    - Male: 3 employees (60%)
    - Female: 2 employees (40%)

6. Performance Analysis:

- Performance Scores:
  - The performance scores in the dataset include 'Exceeds' and 'Fully Meets'.
  - Out of the analyzed employees, the performance scores are distributed as follows:
    - Exceeds: 1 employee (20%)
    - Fully Meets: 4 employees (80%)

7. Insights and Recommendations:

- Based on the analysis, the Executive Office has significantly higher salaries compared to other departments. This disparity could be due to the nature and seniority of roles within the Executive Office.

- The IT/IS and Software Engineering departments have relatively higher average salaries, indicating the importance of technical skills in these roles.

- The Production department has the lowest average salary, which may require further investigation to ensure fair compensation.

- The gender distribution within the company is imbalanced, with a higher representation of male employees. Consider implementing initiatives to promote gender diversity and inclusion.

- The majority of employees (80%) have a performance score of 'Fully Meets'. Recognize and reward high-performing employees to motivate and retain them.

Task 3: Data-driven Insights and Reporting

Dataset Description:

1. Data Acquisition:

   - Obtain the customer churn dataset from a reliable source.

   - Load the dataset into a data analysis tool (Python, R, etc.).

2. Exploratory Data Analysis:

   - Perform initial data exploration to understand the dataset's structure, variables, and distributions.

   - Identify any missing values, outliers, or inconsistencies in the data.

3. Statistical Analysis:

   - Conduct advanced statistical techniques to gain insights into the data.

   - Analyze the relationship between variables and customer churn using methods such as correlation analysis, chi-square test, or t-tests.

   - Identify significant factors associated with churn and their impact on customer retention.

4. Predictive Modeling:

   - Develop predictive models to forecast customer churn.

   - Split the dataset into training and testing sets.

   - Apply machine learning algorithms like logistic regression, decision trees, or random forests to build the churn prediction model.

   - Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, and F1-score.

5. Segmentation:

   - Use clustering techniques like k-means clustering or hierarchical clustering to segment customers based on their characteristics.

   - Identify distinct customer groups with different churn rates and behaviors.

- Analyze the characteristics and behaviors of each segment to understand their churn drivers and develop targeted strategies.

6. Interpretation and Validation:

   - Interpret the results obtained from the statistical analysis, predictive modeling, and segmentation.

   - Validate the findings by comparing them with existing domain knowledge or conducting hypothesis testing.

   - Provide clear explanations of the insights derived from the analysis.

7. Reporting and Visualization:

   - Prepare a concise report summarizing the analysis approach, key findings, and recommendations.

   - Use data visualizations such as charts, graphs, and dashboards to present the findings effectively.

   - Use accessible language to communicate the insights to stakeholders who may not have a technical background.

8. Actionable Insights and Recommendations:

   - Identify actionable insights that can help reduce customer churn.

   - Provide recommendations aligned with the business objective, such as improving customer service, offering personalized promotions, or enhancing the product offering.

Report

The logistic regression model was trained to predict churn based on the provided dataset. The model's performance was evaluated using several metrics, including accuracy, precision, recall, and F1-score. Here are some insights based on the model evaluation:

1. Accuracy: The model achieved an accuracy of 0.81, indicating that it correctly predicted 81% of the churn cases in the dataset. However, it's important to consider other metrics to have a comprehensive understanding of the model's performance.

2. Precision: The precision of 0.66 suggests that when the model predicts churn, it is correct around 66% of the time. This metric measures the proportion of true positive predictions out of all positive predictions made by the model.

3. Recall: The recall, also known as sensitivity or true positive rate, is 0.58. This indicates that the model identified approximately 58% of the actual churn cases correctly. Recall focuses on capturing as many positive cases as possible, minimizing false negatives.

4. F1-score: The F1-score, which combines precision and recall, is 0.62. This metric provides a balanced measure of the model's accuracy by considering both false positives and false negatives. A higher F1-score indicates better overall performance.

Insights:

- The logistic regression model shows promise in predicting churn, as evidenced by the relatively high accuracy and F1-score.

- However, the precision and recall scores indicate room for improvement. The model may have some false positives and false negatives, leading to potential misclassifications of churn cases.

- To further enhance the model's performance, additional techniques like feature engineering, model optimization, or trying alternative algorithms could be explored.

- It's crucial to interpret the model's predictions and consider business implications before taking any actions based on the churn predictions.

- The insights from this model can be utilized to identify customers at risk of churning and develop targeted retention strategies to mitigate churn and improve customer retention.