COMPUTER SCIENCE AND DATA ANALYTICS

Topic title: Data pipeline monitoring and alerting.

Interim Report

Student: Sokrat Bashirov

GWID: G26315644.

1. Introduction

The purpose of this interim report is to provide an overview of the progress made in the research project "Data Pipeline Monitoring and Alerting". The goal of this project is to design and implement a web application using Flask that allows users to monitor data pipelines and receive real-time alerts for anomaly detection. The web application is planned to have two main pages - one for data distribution visualization and the other for the alerting system.

2. Problem Description

The industry and companies rely heavily on efficient data pipelines for data processing and analysis. Ensuring the smooth operation of these pipelines is critical for accurate decision-making and timely actions. However, data pipeline issues, such as latency, errors, or anomalies, can lead to data inaccuracies and operational inefficiencies. Therefore, there is a need for a reliable monitoring and alerting system that promptly detects and notifies users of any potential problems within the data pipelines.

3. Current Progress

As of the current stage of the project, the progress has been made in developing the "Data Distribution" page of the web application. The primary objective of this page is to visualize the distribution of the input dataset and provide users with insights into the data's characteristics.

3.1 Data Distribution Page

The "Data Distribution" page has been successfully implemented using Flask and HTML. The page allows users to upload their dataset, select a specific column of interest, and visualize its distribution using two main plots - box plot and distribution plot (histogram).

- The box plot provides a graphical representation of the dataset's five-number summary, including minimum, maximum, median, and quartiles. It allows users to quickly identify outliers and gain an understanding of the dataset's spread and skewness.
- The histogram displays the frequency distribution of the chosen column's values across different intervals or bins. Users can observe the data's shape and identify significant peaks or clusters in the distribution.

For illustration of the work done so far I can show box plots and histogram of the test dataset’s selected columns.

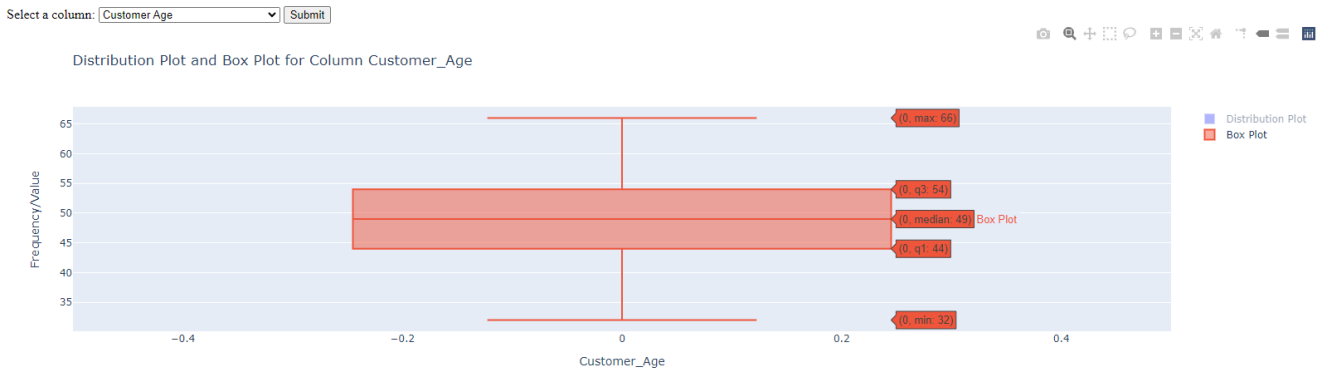

Figure 1. Box plot of customer age column.

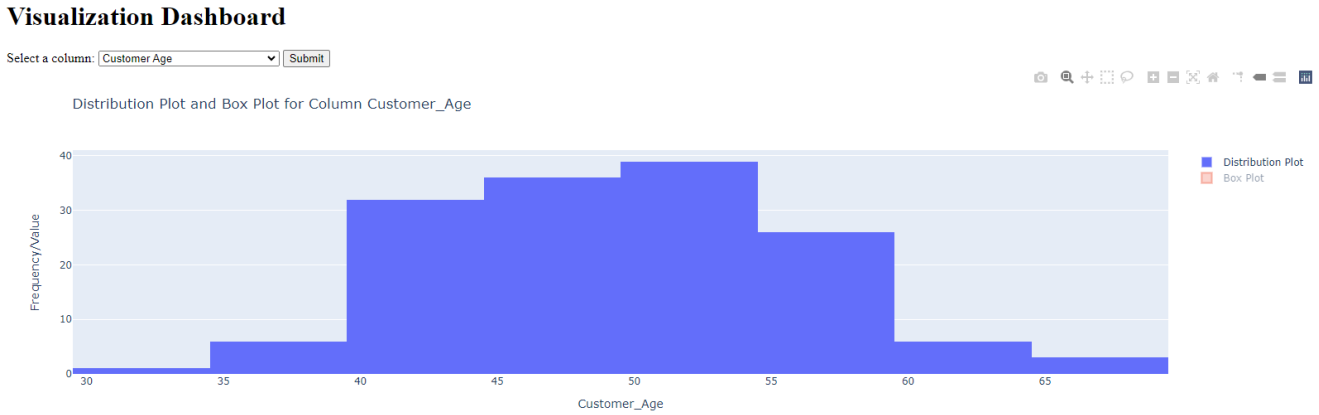

Figure 2. Distribution plot of customer age column.

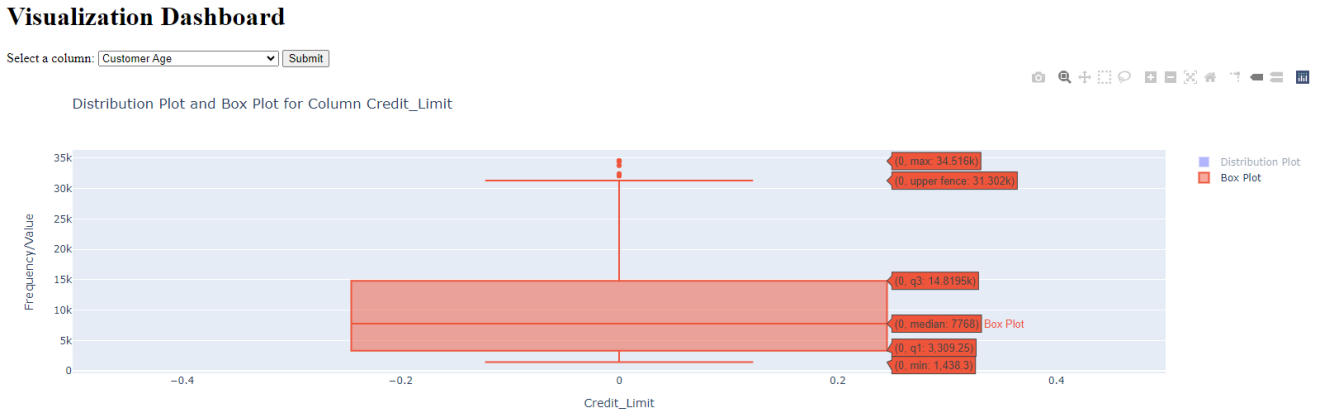

Figure 3. Box plot of Credit Limit column.

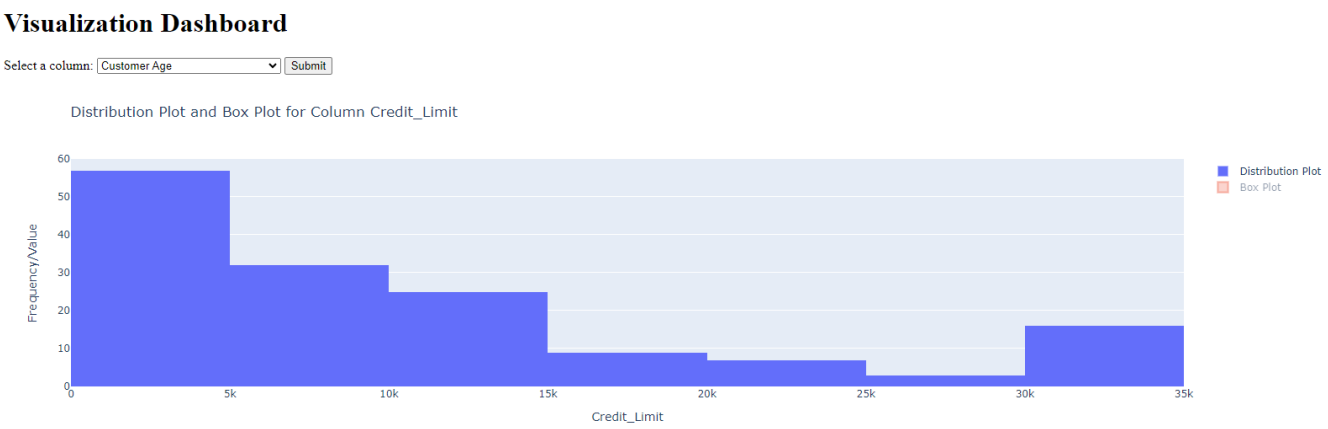

Figure 4. Distribution plot of Credit Limit column.

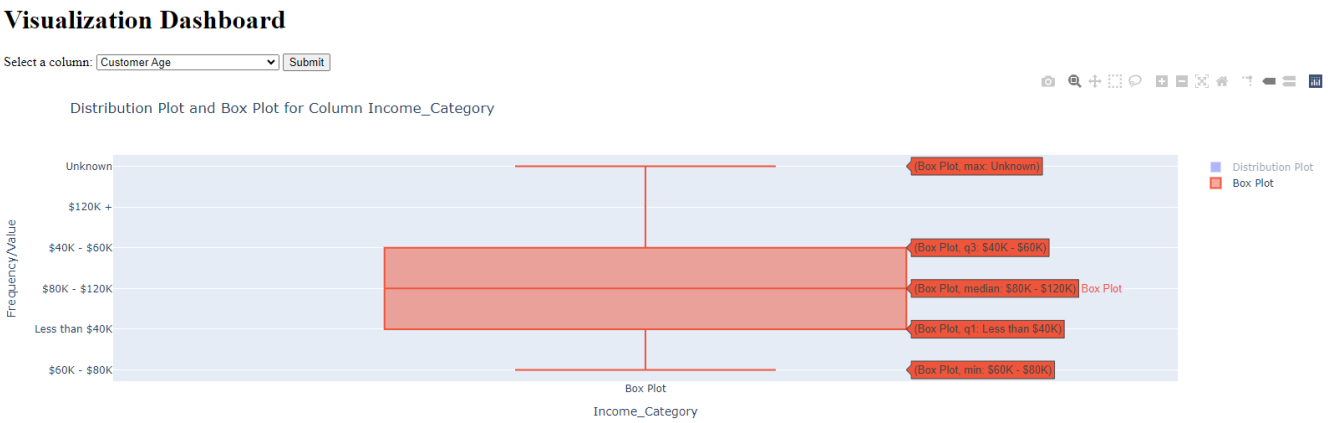

Figure 5. Box plot of Income Category column.

Figure 6. Box plot of Income category column.

4. Future Work

While substantial progress has been made in the "Data Distribution" page, there are several areas for improvement and additional work that will be addressed in the next stages of the project:

4.1 Enhancements to the Data Distribution Page

Include additional visualizations, such as density plots and scatter plots, to provide users with more comprehensive insights into the dataset's distribution.

4.2 Alerting System Page

Design and develop the "Alerting System" page to monitor the data pipelines in real-time and raise alerts when anomalies or issues are detected.

Utilize machine learning techniques and predefined monitoring rules to detect anomalies and trigger alerts.

5. Conclusion

The interim report highlights the progress made in developing the "Data Distribution" page of the web application. The implementation of box plots and distribution plots allows users to explore the dataset's characteristics. The project's future work includes enhancing the "Data Distribution" page, as well as creating the "Alerting System" page for real-time monitoring and anomaly detection.

The web application's ultimate goal is to provide the industry and companies with an efficient and user-friendly tool to monitor data pipelines effectively, detect potential issues promptly, and ensure data quality for informed decision-making.