Improving Claims Processing with Data Analysis

In this project, we aimed to optimize our claims processing by conducting a meticulous analysis of a dataset comprising 2000 rows and 8 columns. We focused on ensuring data accuracy, gaining insights, and identifying areas that may benefit from additional attention or resources.

Task 1: Data Validation

1.1 Data Import & Data Validation

I started by importing the necessary libraries for data analysis. I loaded the dataset from the 'food_claims.csv' file into a Pandas DataFrame.

Our data validation process revealed the following key findings

- claim_id: The column serves as the unique identifier for each claim and requires no cleaning.
- time_to_close: The column contains 256 unique values and no missing values.
- claim_amount: The column does not contain any missing values.
- amount_paid: The column has 36 missing values ("NA") which have been replaced with the median amount paid across all instances.
- location: The column has four unique locations without any missing values.
- Individuals_on_Claim: The column contains 15 unique values without missing values.
- linked_cases: The column has 26 missing values ("NA") which have been replaced with FALSE.
- cause: The column has two categories, meat and vegetables, and 713 missing values ("unknown"). Requires no cleaning as per the criteria.

1.2 Data Cleaning

During data validation, I uncovered some key findings, such as missing values in 'amount_paid' and 'linked_cases,' which I subsequently addressed.

Task 2 Data Discovery and Visualization

2.1 Identification of the Most Observed Location Category

I identified the most observed location categories using a pie chart, which allowed me to visualize the distribution of claims across different locations.

Our findings were as follows:

- Recife: Comprised the highest percentage of total observations at 44.2%.
- Sao Luis: Ranked second with 25.8% of the observations.
- Fortaleza: Followed with 15.5% of the total observations.
- Natal: Had the lowest number of observations at 14.3%.

2.2 Assessment of Observation Balance across Location Categories

I also assessed the balance of observations across location categories using a bar chart, highlighting variations in claims frequency by location.

Recife had the highest number of claims (857), while Natal had the fewest (278). Sao Luis and Fortaleza fell in between, with 499 and 304 observations, respectively, indicating variations in claims frequency by location.

Task 3: Distribution of Time to Close for Claims

3.1 Histogram

In this task, I visualized the distribution of time it takes to close claims using a histogram, offering insights into the efficiency of our claims process and potential areas for improvement.

3.2 Scatter Chart

I utilized a scatter chart to display the average time it took to close claims by claim ID, helping me identify any outliers or variations in claim processing times.

Task 4: Investigating the Correlation between Time to Close and Location

4.1 Box Plot

To explore the relationship between claim closure times and location, I used a box plot, providing insights into central tendencies, variability, and potential outliers.

4.2 Line Chart

I also created a line chart to depict the average claim closure times across different locations, revealing which location had the longest average closure times.

In conclusion, our data analysis has provided valuable insights into our claims process. We have identified areas of strength and potential improvement, particularly in the context of location-specific processing times. By leveraging these findings, we can enhance the efficiency and effectiveness of our claims processing system, ultimately providing better service to our customers and optimizing resource allocation. Thank you for your attention. ✅