Food Claims Process Analysis

Background

Vivendo is a fast-food chain in Brazil with over 200 outlets.

Customers often claim compensation from the company for food poisoning.

The legal team processes these claims. The legal team has offices in four locations.

The legal team wants to improve how long it takes to reply to customers and close claims. The head of the legal department wants a report on how each location differs in the time it takes to close claims.

Data

The dataset was provided by datacamp and can be downloaded from here.

It contains one row for each claim and has 2000 records.

Column Name	Criteria
claim_id	Nominal. The unique identifier of the claim. Missing values are not possible due to the database structure.
time_to_close	Discrete. The number of days to close the claim. Any positive value. Replace missing values with the overall median time to close.
claim_amount	Continuous. The initial claim requested in the currency of Brazil, rounded to 2 decimal places. Replace missing values with the overall median claim amount.
amount_paid	Continuous. Final amount paid. In the currency of Brazil. Rounded to 2 decimal places. Replace missing values with the overall median amount paid.
location	Nominal. Location of the claim, one of “RECIFE”, “SAO LUIS”, “FORTALEZA”, or “NATAL”. Remove missing values.
individuals_on_claim	Discrete. Number of individuals on this claim. Minimum 1 person. Replace missing value with 0.
linked_cases	Nominal. Whether this claim is linked to other cases. Either TRUE or FALSE. Replace missing values with FALSE.
cause	Nominal. Cause of food poisoning. One of “vegetable”, “meat” or “unknown”. Replace missing values with ‘unknown’.

Tasks

1.For every column in the data:

a. State whether the values match the description given in the table above.
b. State the number of missing values in the column.
c. Describe what you did to make values match the description if they did match.

2.Create a visualization that shows the number of claims in each location.

Use the visualization to:

a. State which category of the variable location has the most observations.
b. Explain whether the observations are balanced across categories of the variable location

3. Describe the distribution of time to close for all claims. Your answer must include a visualization that shows the distribution.

4. Describe the relationship between time to close and location. Your answer must include a visualization to demonstrate the relationship.

Analysis

1. Column description and number of missing values

claim_id: There are 2000 unique values that match the description given. There are no missing values. No changes were made to this column.

time_to_close: The values in this column range from 76 to 518. All values are positive. There were no missing values, so no changes were made to this column.

claim_amount: The values are in Brazil currency but in a text format, so I split the currency symbol and the numbers using the text-to-column feature. I then converted the column with the numbers only to Brazil currency and round it to 2 decimal places. There were no missing values, so no changes were made to this column.

amount_paid: The values were not in Brazil currency, so I convert it to Brazil currency and round it to 2 decimal places. There were 36 missing values, so I calculated the median which was 20,105.70 and filled in the median for the missing values.

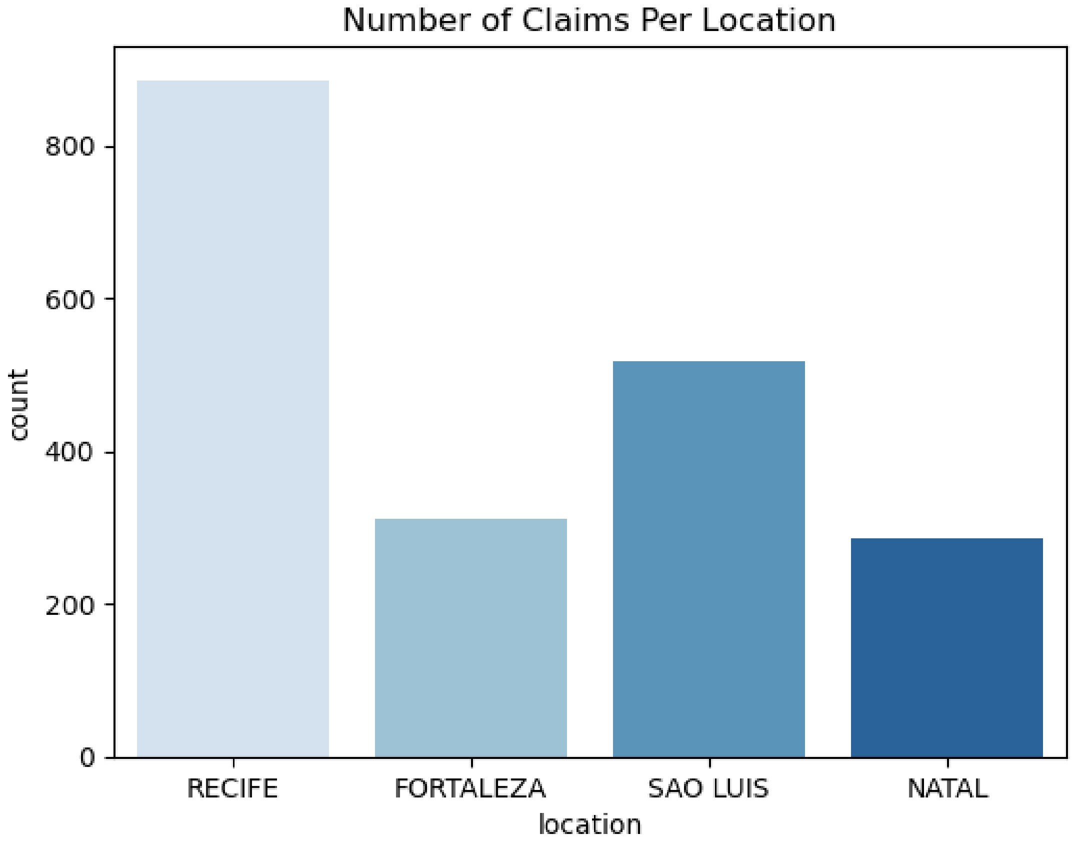
location: This column contains locations of the claim which are “RECIFE”, “SAO LUIS”, “FORTALEZA”, or “NATAL”, which is consistent with the description given. There were no missing values, so no values were removed.

individuals_on_claim: This column has the number of individuals on this claim. The values range from 1 to 15, minimum being 1 person as described. There were no missing values, so no changes were made.

linked_cases: The values in this column were TRUE, FALSE or NA which is not as described. There were 26 missing values so all missing values were replaced with FALSE.

causes: This column contains causes of food poisoning which should be “vegetable”, “meat” or “unknown”. But other values included " Meat" (14 values) and "VEGETABLES" (16 values). I used the find and replace feature in Excel to replace all values of " Meat" to "meat" and "VEGETABLES" to “vegetable”. There were no missing values.

2. The number of claims in each location

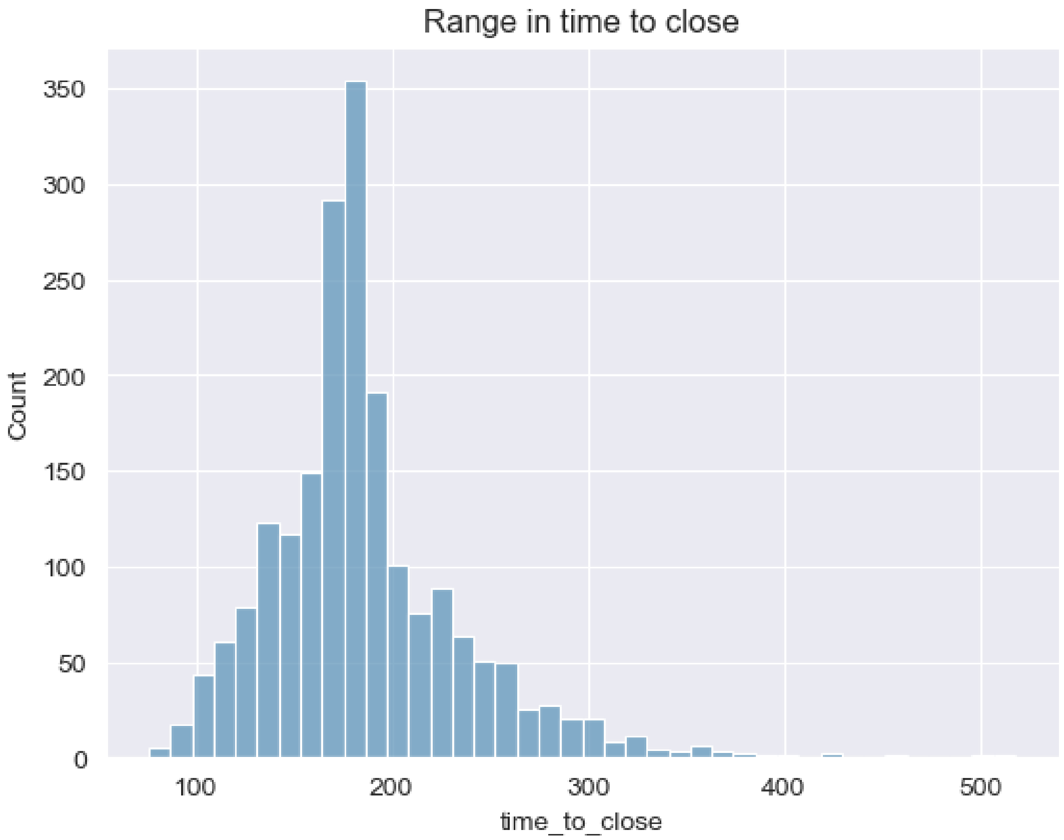


time_to_close

There are four locations included in this data. The most common type listed is RECIFE, with over 800 number of claims and the least common being NATAL with less than 300 number of claims. The categories are unbalanced, with most claims being from RECIFE or SAO LUIS.

The legal team should focus on these locations to check for the claims and improve how long it takes to reply to customers and close claims.

3. The distribution of time to close for all claims



The time to close for all claims ranges from 76 to 518. As we can see from the plot above, the highest time to close was between 90 to 300 and after 300, we see low values which we can be categorized as outliers.

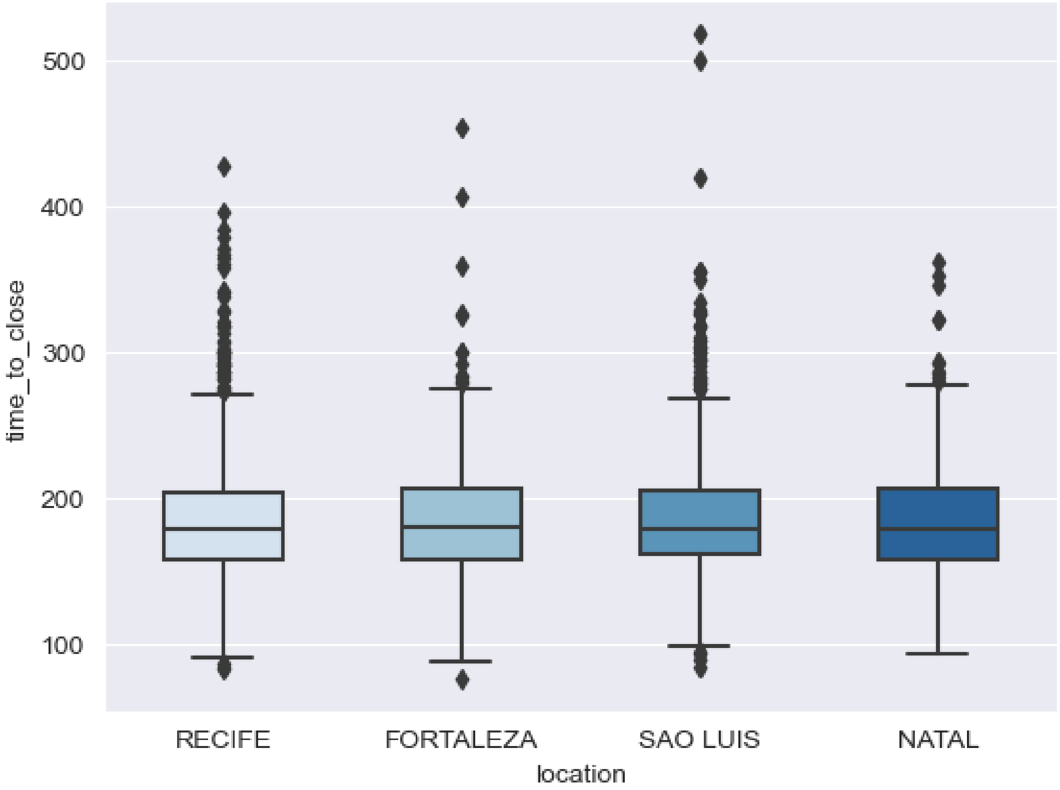
The distribution of time to close is right skewed since the values tend to be populated more on the left of the histogram and the tail extends to the right.

The legal team should focus on time to close between 90 and 300 to check for the claims and improve how long it takes to reply to customers and close claims.

4. The relationship between time to close and location

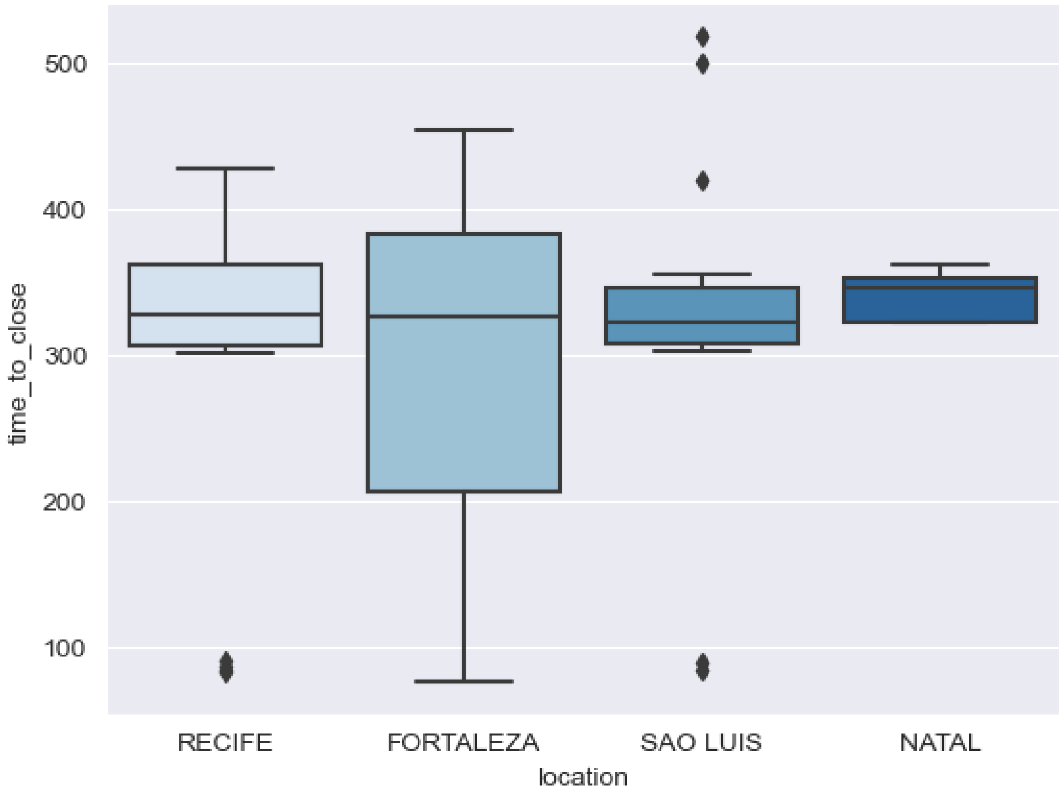
Now looking at the relationship between time to close and location, we want to see the variation in time to close for each location. The time to close can be taken from 90 to 100 but we need to look at the two variables together to see if this is realistic.

We plot a box plot to show the relationship between time to close and location. We want to look at the range of values in the time to close by location including all outliers in the data. In the plot below, we can see that the outliers are dominating the data and making comparison difficult. High outliers are the data points that indicate long time to close for specific claims in a particular location while low outliers are data points that indicate short time to close for specific claims in a particular location. To make it easier to compare the rest of the data, we will remove this outlier.



In the plot below, we have removed the outlier and can now focus on the main data range.

As can be seen, SAO LUIS has the highest number of time to close but its interquartile range of time to close is lower than that of RECIFA and FORTALEZA. This would suggest that the highest time to close may be lower than other types. However, this could also be an effect of having the largest number of locations, so the large number of low time to close locations brings the median down.



Suggestions

Based on all the above, we would recommend that the legal team:

- focus on locations with time to close between 90 and 300.
- focus on RECIFE and FORTALEZA with high time to close.
- consider locations with low time to close to check how they differ with other locations in terms of how claims are closed shortly.

Further analysis should be done to understand if location really does impact the time to close.