Data Validation

Describe the validation tasks you performed and what you found. Have you made any changes to the data to enable further analysis? Remember to describe what you did for every column in the data.

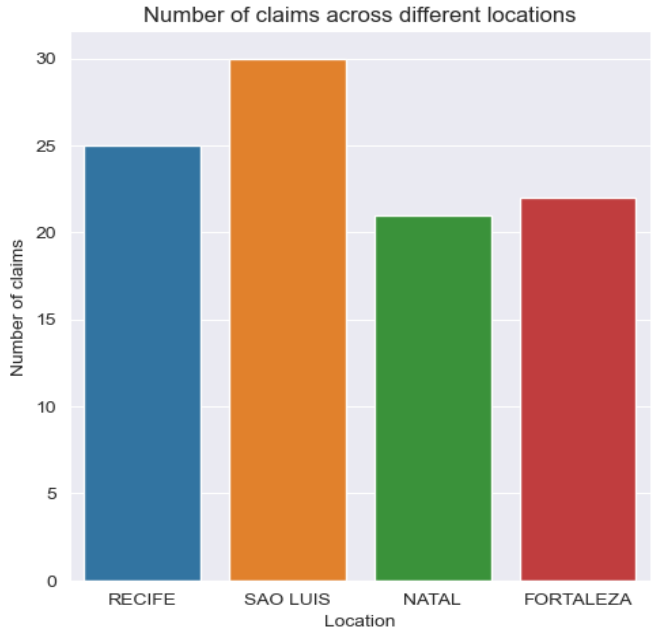
The original data is **98** rows and **8** columns. First, I checked the datatypes of all the columns, they were the same as expected in the description except 'Claim Amount' which was text originally that made sense, to change it from R$50,000.00 to 50000 (as per the description) I changed it from text to numeric. I replaced all the missing values (which were 78 rows) in the ‘Cause’ column with 'unknown' as a result there were no more missing values in the dataset. I noticed that the 61st row in the 'Time to Close' column had a value of '-57' as time cannot be negative, I changed it to 57. Looking at the remaining columns:

- There were 98 Unique Claim IDs.
- There were 4 different locations, as expected.
- There were some rows in the 'Individuals on Claims' column that were 0, suggesting that individuals on a claim cannot be zero, this should be confirmed with the team providing the data
- There were 2 Linked-case options (either case is linked with other cases or not) that are True/False, as expected.
- There were three claim causes. Meat, vegetables, or unknown as expected.

Data Discovery and Visualization

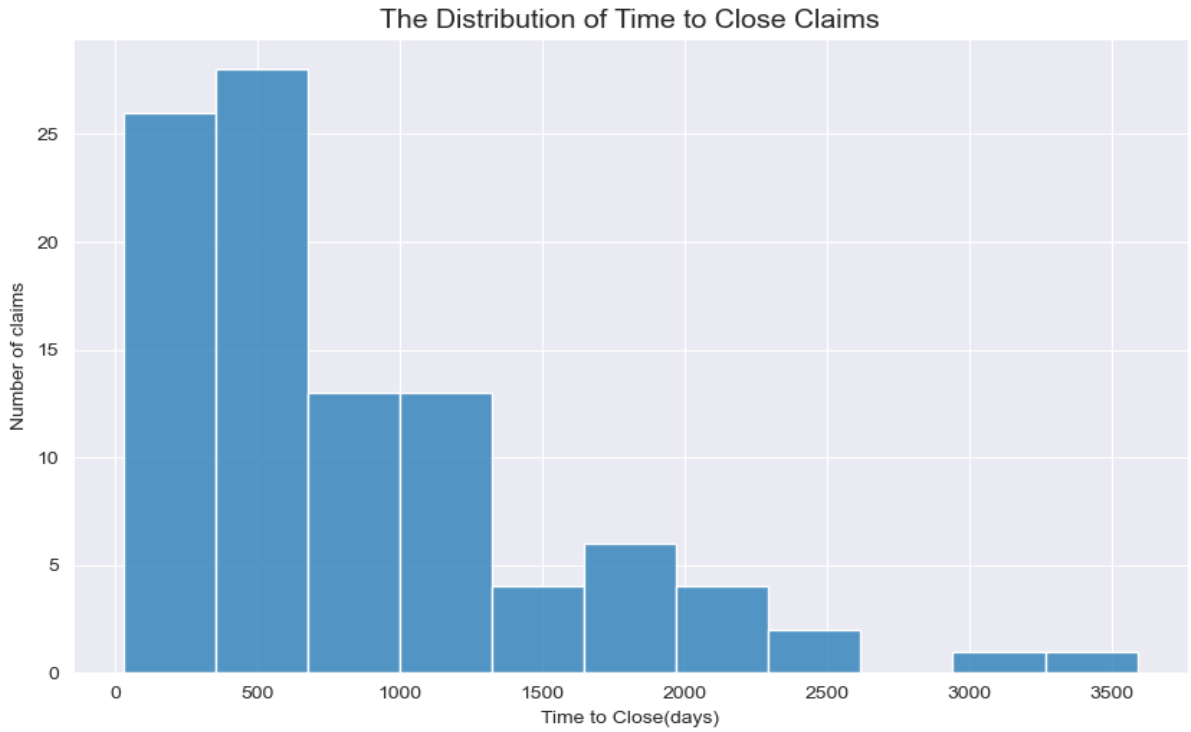
Describe what you found in the analysis and how the visualizations answer the customer questions in the project brief.

How does the number of claims differ across the locations?
There are four different locations included in this data along with 98 claims. As the graph shows, the number of claims is different across each location with the greatest number of claims in **Sao Luis** with **30** claims, followed by **Recife** with **25**, and **Natal** and **Fortaleza** with **21** and **22** cases respectively. This would suggest that the larger team should be working in those locations with the highest number of claims.
The number of these claims across different locations could be directly connected to the causes, but nothing can be said with certainty because most of the causes are 'Unknown'.



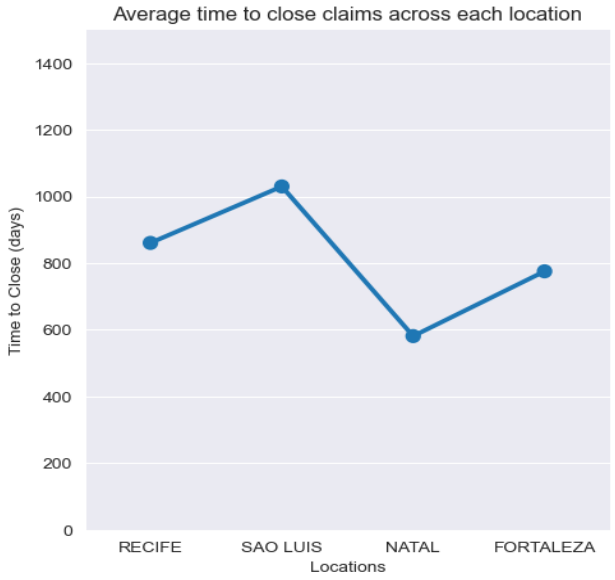
What is the distribution of time to close claims

The distribution of time to close claim shows that more than half of the cases took less than **500** days to close from the time the claim was made, there are some outliers that took **3000-3500** days to close a claim, but that is very uncommon. The legal team needs to focus on those locations where claims took more than **500 days** to close.

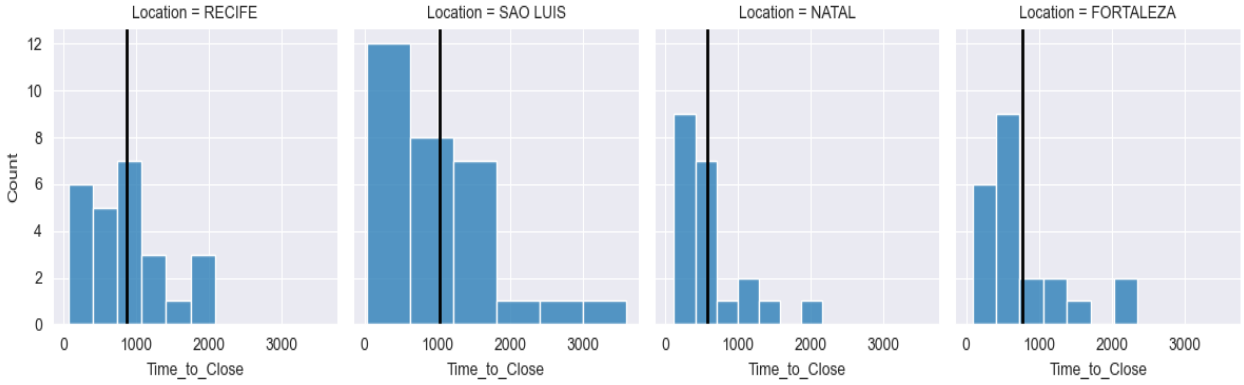


How does the average time to close differ across each location?

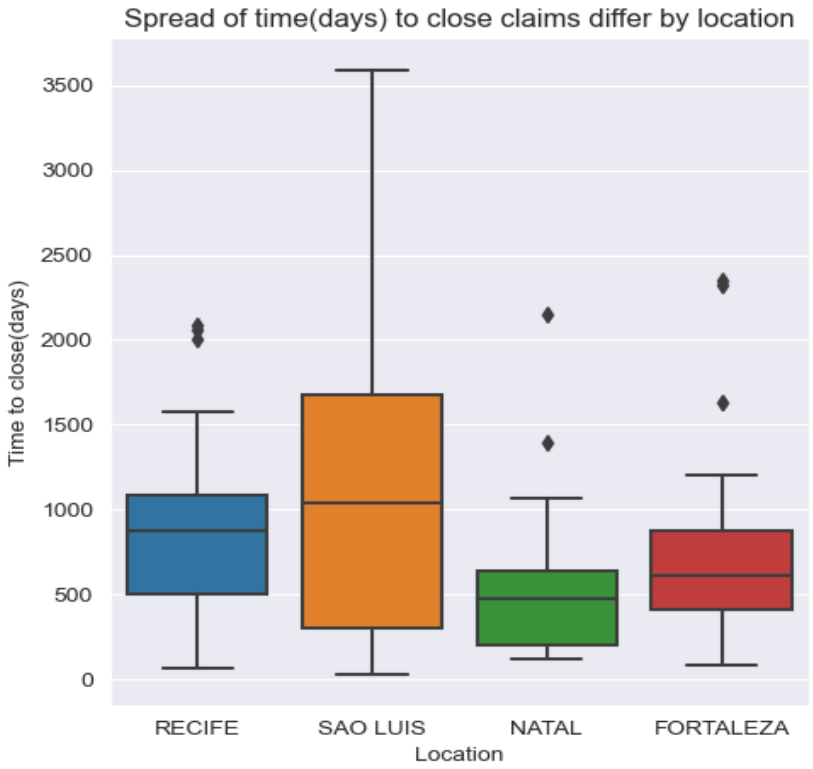
The difference in average time to close claims depends on the number of claims being made in a location. The location with the higher number of claims has a higher average time to close a claim than other locations. This suggests that the locations with a higher number of claims need a bigger legal team to deal with claims.



From the plot below the distribution and average time to close a claim suggests that **Fortaleza** and **Natal** have almost the same no of claims but there is more difference in the average time to close, this suggests that the legal team in **Fortaleza** needs to improve its efficiency in lowering down the average time to close a claim.



The plot below shows that **Sao Luis** has a larger spread of time to close a claim than other locations which could be due to the reason that couple of claims took more than **3000** days which is stretching its spread, legal team in **Sao Luis** can work on this area as well which will also help to lower the average of **Sao Luis** as most of the claims were closed before **500** days.



Based on all the above analysis, we would recommend that the legal team should focus on the locations that have the highest average time to close a claim as compared to the total number of claims. The fast-food company should also consider why there were a high number of claims in a certain location and it can be done by further analysis of the cause of a claim.