Data Cleanup

REVIEW

HISTORY

Meets Specifications

Dear Student,

Great work on your submission, you showed a great understanding of the underlying concepts and passed all the rubrics!

You might also find the following links useful to learn more about both the why and how behind the process of data cleaning:

- The Ultimate Guide to Data Cleaning
- How to clean your data to make it analytics ready
- Data Cleaning in 2021: What it is, Steps to Clean Data & Tools

In case you still have any outstanding questions, we are here to help! Make sure to reach out on knowledge where one of our expert mentors will help with any doubts you might have

Best of luck on the following projects and lessons!

All the best,

Business and Data Understanding

✓ The section is written clearly and is concise. The section is written in less than 250 words.

Answers are clear and concise while respecting the word limit, well done!

✓ All the following questions have been accurately answered:

1. What decisions need to be made?
2. What data is needed to inform those decisions?

Both answers are correct, well done!

Q1) We succinctly stated our main goal, clearly illustrating what is the expected outcome of the project

Q2) The idea here is to provide an overview of the necessary information to reach a final recommendation.

- The variables shared help illustrate what is necessary for our analysis and provides context for the reader to better understand the problem we want to tackle

Suggestion:

To provide the reader with even more context, and help them better follow our reasoning, we could also add some comments on how each of the variables might add value to our final decision, for example:

- Total population and population density could be useful as denser areas would have more customers.
- Number of families and families with individuals under 18 are important demographic information that would allow us to detect areas with more potential customers.

In addition to demographic variables, we could also seek out the traffic drivers to our current stores to understand what are the other landmarks or businesses that can drive foot traffic to the store. Understanding how far competitors' stores are from the company's stores can help model any traffic drivers as well.

Competitor sales could also be relevant to understand if there is significant customer demand in the city. It would also be helpful to see if competition might pose a risk in case Pawdacity decides to open a store there.

Finally, we could gather data relative to our current local promotions and marketing budget spent per city on the current stores. We can also try to get information on the expected marketing funds the company will spend to promote the new store.

Building the Training Set

✓ **The averages for each column is correct in the training set**

The objective here is to ensure our data cleaning steps resulted in the expected dataset before moving to step 3.

We did an excellent work combining the multiple sources into a single dataset, both the `Sum` and `Average` values for each of the six variables are correct!

Suggestion:

To improve the "readability" of the numbers I would recommend rounding to the nearest integer value e.g 343,027.64 => 343,028

You can also check the following link for an ebook by Alteryx on the topic of data preparation
Alteryx e-book on data preparation

obs:

The submission template mentions the following:

> In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

So notice the comment above is just a suggestion from my personal experience that business stakeholders generally prefer rounded numbers

✓ **Outliers have been analyzed for each field in the training set.**

The outliers are accurately identified.

The decision to keep, remove, or impute each outlier is well justified.

After pointing out all the possible outlier cities only one city should be decided upon to remove.

Great work addressing all the rubric requirements!

- We clearly identify all the outliers!
- We also outlined the rationale used for keeping or removing each of them, well done!

SOME THOUGHTS ON THE PROCESS OF OUTLIERS REMOVAL

When dealing with outliers the whole process is more art than science. The most important part is deciding whether an outlier is going to add information or noise to our model.

- The rule of thumb would be to *remove any outliers we think will add noise* and *keep any outliers we believe has valuable information*

For example:

- If we are modeling `weight`, we know that this follows a `normal distribution`, so any *outliers* can safely be removed from the analysis as they are *noise and not information.*
- On the other hand if we are modeling `wealth`, which follows a `power-law distribution`, by removing the outliers we might underfit our model as *outliers are an important part of the explanation*, which means they *would add value* and not noise into our model.

I would also recommend going over the following brief article on the topic of outlier removal
outlier removal

The following is a great reference if you ever need to quickly remember the IQR analysis process:
IQR analysis

DOWNLOAD PROJECT

RETURN TO PATH