Re: Sprocket Data Quality Issues

Dear Manager,

Thank you for reaching out to KPMG for your data needs. I am Firdaws Yahya, your analyst for the project. With regards to the provided dataset, here is a summary of data quality issues raised based on our benchmarks. In this email, alongside the issues, I have also made recommendations on how to approach them in the future.

1. Missing Values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Transaction | Customer list | Customer’s Demographic | Customer’s Address |
| Missing cells | 1542 | 317 | 1763 | 0 |
| Missing cells(%) | 0.6 | 1.8 | 3.4 | 0.0 |
| Number of records | 20000 | 1000 | 4000 | 3999 |

Missing values are a huge hindrance to data quality of your datasets. This affects the completeness of your data. In handling the missing values, columns like that in transaction was dropped. The customer job category was filled in the customer list table while the remaining column with null values were dropped. Moving further, I would make three suggestions in regards to tackling this

* Define a data quality threshold. A threshold will help define an allowable percentage of missing data allowed for each column. This will help reduce the volume of missing cells and help further define action for imputations
* Based on the method of data entry, Mandatory fields should be used. This help differentiate from truly missing data and data not entered into the database. An option can then be made for data actually missing.
* Defining an imputation method for various columns. After a threshold is defined, it is also advised to have an imputation method for column deemed relevant. An example can be seen in the case of Job category where I had null values filled based on the job title given or giving the last name column a uniform name rather than dropping them.

1. **Unique ID**

All data provided had no duplicates but the lack of an id to uniquely identify some columns proved to be a problem when dealing with the data. I will recommend the following:

* I will advise a master table containing all your data or rather, several master tables for specifics. The data can then be divided into a customer’s master table, product master table and if it turns out necessary, locations master table. The transaction table on its own is alright.
* The customer list should have an id linked. Considering that the customer list table in a way, serves a customer’s master table, lack of an id makes it harder to join with other tables. Without a unique id for each customer, we cannot link each customer to other data.

1. **OUTDATED DATABASE**

**It is very important that sprocket data is up to date. This helps in making business decisions that are relevant to the present day as well as historically correct. The data sent contains**

* People whose age can be deemed inaccurate. One of such is 180 years old
* Contains dead people in the data. Remove dead people to avoid skewing customer data.

1. **INACCURATE ENTRIES**

There seem to be errors in the data itself. On of which is the labelling of genders. In the customer’s list table, one can see here categories of genders namely: Male, Female and U’. U was later replaced with unknown. Issues with how gender was input was a common inaccuracy across the data. To further prevent this error and improve data quality I suggest that:

* Setting Data constraint
* Data validation for fields

1. **INCONSISTENCY**

Due to lack of a master table, Data are spread in different sheet. This has proved to be an issue since same data have appeared in different format. The gender and state values in the customer list differed greatly from the data in the customer demographic table. I advise to either keep the customer list table or merge the demographics and address table before working on them. Although when merged, data is similar to that of the customer list table but despite coming from the same source. Both seem to be inconsistent with the other. I suggest paying attention to creating a master table as well as including strict data validation.

1. **Wrong Data types**

Not The data sent had a lot of inconsistent data type. A lot of which had to be declared while working on the project. Date objects like first\_sold\_date in the transaction sheet was not in the roper date format. This had to be converted. In database it is very important that constraint on the data types is given. This is what I recommend sprockets do.

1. **POOR VALIDITY**

For every data sprocket generates, I advise metadata for each should be created. This metadata will help those working on the database understand what it is the company is trying to communicate as well as why the data exists. Columns like default in the customer demographic table are an example of why your data needs a metadata. It helps understand each data instead of. Moving forward, I advise that the documentation of data should be included in sprockets data process.

This is a summary of the data issues I have come across while working on sprocket’s data. There is still a lot to do in terms of transforming the data for further analysis. I will be available to answer your questions or give you a walkthrough our process if need be.

Kind Regards

Firdaws Yahya

Data Analyst,KPMG