Kevin Zhao (zhaokevin79@gmail.com)

Section 4: Writing an email to a stakeholder explaining the results of EDA on the dataset

When talking to non-technical people, it is vital that you get your findings and concerns across without using any technical words or patterns that they would not understand. Putting yourself in their shoes and understanding that they won't be able to reason through your points like you do is incredibly important.

A good analogy to follow for talking to non-technical people is that if your parents can understand what you're saying, then they can. This is the way that I introduce myself and what I do when I meet people for the first time who don't work in computer/data science, and I have found that it works wonderfully.

Dear [Product or Business Stakeholder],

I hope you are doing well. I have completed my analysis on the data sets you sent me last week, and wanted to provide you a brief overview of my findings, as well as any questions I have to address concerns and next steps.

Data Quality Issues:
Using SQL queries and data processing scripts, I explored the data sets and did some initial insight evaluations. Here are some issues I found with the data, which make it significantly harder to derive concrete insights from the dataset:

- Key metrics have many missing values for some fields
- Some data is duplicated and causes issues when loading into our database
- Data fields are being assigned the wrong type (e.g. Receipt price totals should be recorded as a number, but are instead being recorded as a string of text)

Data Scaling Considerations:
While this data set was relatively small and easy to manage, it can quickly overflow with unnecessary information if I do not optimize and structure the data efficiently. As the dataset grows larger, we could run into the following problems:

- Increased time and cost to find insights and summaries about the data itself

- Increased expense to maintain and restore the dataset in the case of a power outage or program failure
- Increased number of fields each row will have, which could complicate the data

Next Steps:
To help scale and manage the database from my end, I will:

- Remove redundant/non-important fields from each data point to reduce storage load
- Create indices for commonly used tables/fields, which helps speed up performance
- Re-organize the database to optimize commonly used tables/fields for query times

Additionally, to move forwards efficiently, I would appreciate responses to the following:

- A list of sources contributing to this dataset
- Contact information for the data scientists responsible for data sourcing and management
- What different data sources and data fields do you anticipate needing the most amount of insight on, and how frequently do you believe you will need these insights?
- What is the rate of data you anticipate adding to the data set per day?
- What are the different places this data is sourced from, and do you anticipate any new streams of data entering this data set in the foreseeable future?

Let me know a convenient time to discuss these findings further. Thank you for your input and thoughts on this!

Best,
Kevin Zhao
[Email Signature]