MissionWired

Client Communications

SOP for Constituent Project

The goal of this report is to provide a step-by-step instructions and techniques used to achieve the desired results. There are 3 underlying data sources used in this project as per request:

1- Constituent Information,
2- Constituent Email Addresses,
3- Constituent Subscription Status.

The Data Tools being used for this project is Python in combination with different library (see Requirement.txt file). And the computing platform is Jupyter Notebook. First, I started to import all the required libraries as needed (see pdf Input [136]). Then, I imported the 3 data files (se pdf Input [117]). I started the project by first scanning at the metadata level in these three tables to understand the data being used. I looked for various elements: datatypes, fields, attributes, primary keys, foreign keys and importantly the data quality (see pdf Input [118 to 120]). After this step, we need to JOIN the 3 datable. So, I started with the Constituent Information table did a LEFT JOIN to the Constituent Email Address with cons_id (key). Now using the Constituent Subscription Status table, did an OUTER JOIN to the previously joined table with email_id (key). After this step, since we were stated to filter at the data source level. I filtered the data for subscription status where chapter_id =1 OR if an email blank. Then, I wanted to see if there are any missing values, duplicates records, datatypes or any outliers that can be skewing the data. To fix those issues, I treated the data and scrubbed the dataset. Finally, after this step and validation, I outputted the file to a folder in the saved directory.

For the second exercise, I used the saved output file as the underlying data source to provide analytics. For the acquisition_date, it made sense to use the created_date. One thing we need to notice is that created_date is in DATETIME format. So, I converted created_date to DATE format. Then I used created_date and used GROUPING to get the aggregation counts based on each date. Then I outputted the file to the folder.

Assumption:

1- I assumed that email entry can not be duplicate since it will not make sense that on person create same email for same time so I took out duplicate.
2- Another assumption is that other column entry on the table can allow duplicate for our analysis
3- Finally on the Source(code) column I used Sklearn to impute missing value.
4- The Script save fille worked in the folder without deleting old existing file

Further analysis can be done on this data as example to find what source perform better so we can use the platform for advertisements etc.