Our mini project focused on building an Extract, Transform, Load (ETL) pipeline using Python. The ETL process is commonly used in data integration and involves extracting data from various sources, transforming it into a suitable format, and loading it into a target system or database. In this project the data we were given two (excel spreadsheets), our priority was to extract the data and format them into functional data frames, once we created individual tables from our data frames, we created an ERD with our tables.

We utilized the python libraries of pandas, numpy, matplotlib , sqlalchemy.

We read in the excel file titled “crowdfunding” and began with cleaning up the data.

Steps to clean the data and create a category and subcategory data frame required us to rename columns and remove spaces and special characters. We had to split a specific column called category & subcategory. The pandas split function helped us achieve the desired outcome. Once all steps in the first part of our ETL project were completed we exported the new data frame to a csv file titled the\_real\_category.csv . Similar steps were taken to create a second data frame called subcategory and in turn exported this data frame to a csv file titled eal\_subcategory.csv.

Our next step was to build a campaign data frame. Similar clean up steps as with the previous data frame where taken . An additional step that needed to be take was to convert the data types for goal and pledge . We achieved this by changing these existing data types of int64 to float with the astype pandas function. Our dates had to be updated as well and the to\_datetime function took care of the conversion. Once the data types were converted columns were removed to create a updated data frame titled campaign\_df. This new data frame was exported and titled real\_cleaned\_campaign\_df.csv.

Our last data frame required the import of the supplied file titled contacts. After reading in this excel file, we created new column names and arranged the data to clean up the date according to the column names. We created an iteration that allowed the data to be entered into the new column names:, contact\_id, name, email. This file was then exported to a file labeled real\_contacts.csv.

Our next deliverable required us to create a ERD design and schema using the quick DBA website. Once our design was

Team 4 ( van, Danny , Carlos ) , stand up.

1. Completed the sqlalchemy code in pandas to read in the csv files . Completed the required queries. We uploaded and organized our files into github . Created our write and added to our github a pdf version.
2. We have uploaded our final project version to github.