Project 2 Crowdfunding Summarization

Chris Hicks, Steven Madden, Samantha Schutz

Data is collected and stored in different ways depending on the collection methods utilized. Some of these data collection methods are more standardized than others. This means the data must be “cleaned” in order to utilize it to make any meaningful conclusions from it. An important concept for data analytics is ETL (extract, transform, load). This is the process of cleaning, organizing and storing the data to use in multiple settings of data analytics. This project shows how a set of crowdfunding data is cleaned using pandas, transformed to sequel for storage then loaded to other tools to then analyze the data.

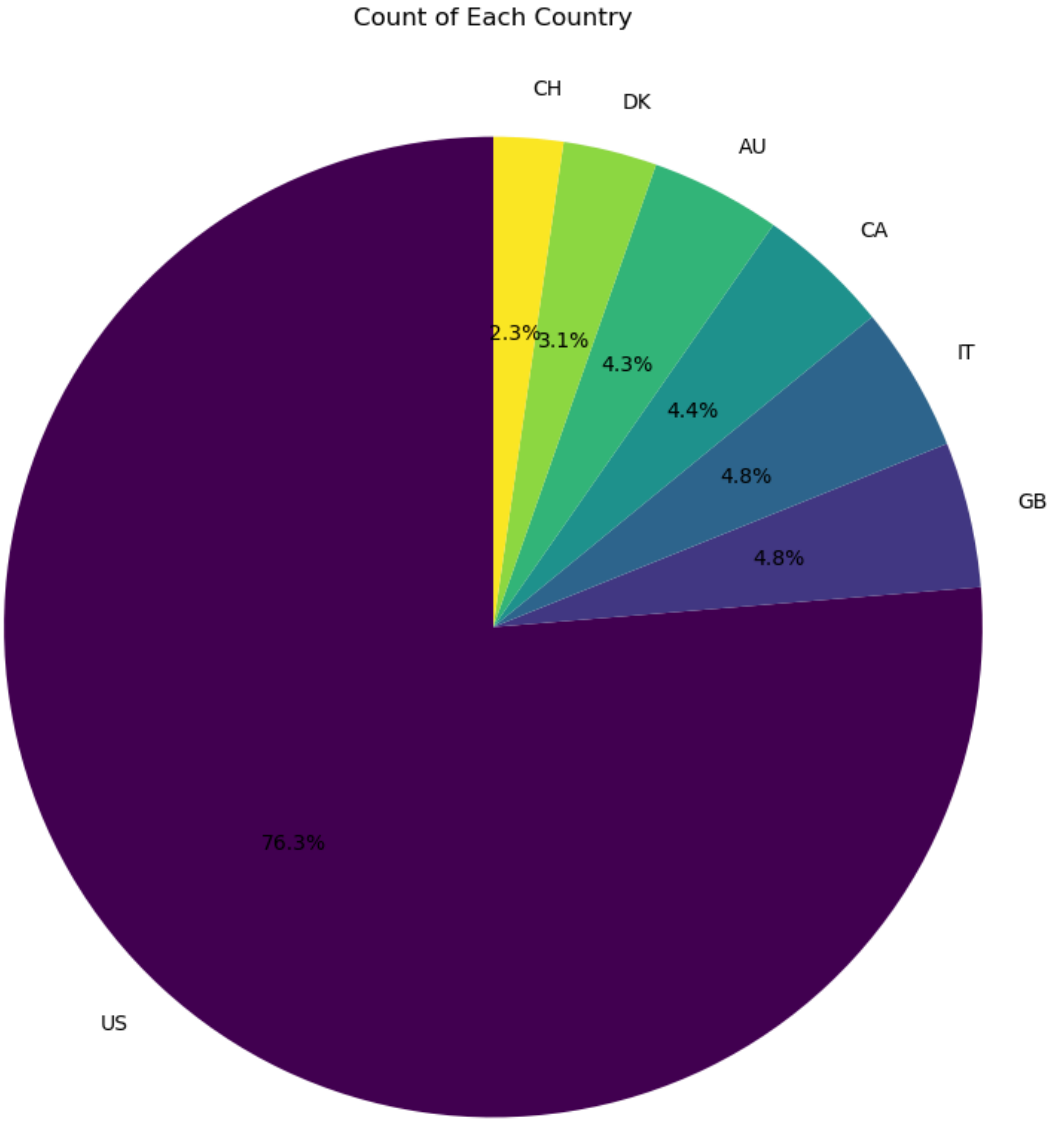
The first step for this project was cleaning the data to ensure our columns were the correct data types. Using pandas, columns were split and dropped, tables created and data types converted. The extracted data was then exported to four CSVs where it could be used.

In order to understand the relationship between the four CSVs, an entity relationship diagram (ERD) was made. The ERD is used as a blueprint to determine if there are any issues with the relational databases enabling the tables to be merged for analysis. Using a free website, quickdatabasediagrams.com, a visual representation of the tables come up and primary keys can be established in addition to the connections between the tables being visually represented for ease of understanding. This can then be loaded into a Postgresql database.

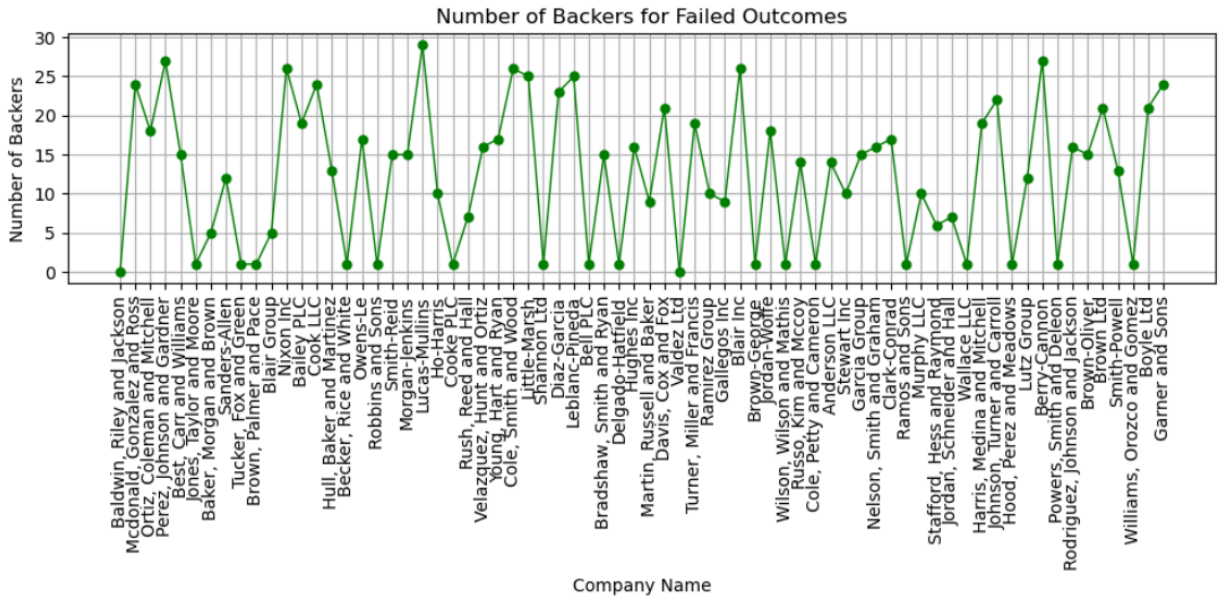
The Postgresql database is an easy way to store and quickly use a large dataframe. The loaded tables will be populated with the CSV files, showing any errors that may be hidden. An easy select from statement can be used to ensure proper representation of the data. While Postgresql has many uses, it lacks the ability to make visualizations of the data.

To use the clean and loaded data from Postgresql, a few transformations can be made to load the data back into a Jupyter notebook. With a few lines of code, queries using raw SQL can be made against the Postgresql database and visualizations can be made using matplotlib.

A couple of visualizations were created to ensure the data was read into the environment and were usable. The first crowdfunding visualization showed the US making up the vast majority of the countries participating with this method of money donation.



The next visualization showed the number of companies that failed with 30 or fewer backers. Obviously a company with 0 backers would fail because no one donated money, however, further analysis would need to be performed to make some sort of conclusion as to why a company with almost 30 backers failed. This could be explained simply by the amount of money required or there were smaller donations being made.



Overall, this was an interesting dataframe to use for an analysis. It could be useful for companies using this form of donation to dictate whether a project is more likely to succeed in the requirements or fail.