Project 2, Group 10 Writeup

1. Project Overview

The project's goal is to analyze crowdfunding campaign data, design a relational database schema for storing and querying the data efficiently, and perform exploratory data analysis (EDA) with the help of SQL queries and visualizations. The dataset used included contacts and campaign information from a crowdfunding website.

---

2. Part 1: Data Transformation and Import

Step 1: Loading and Preparing the Data

We loaded the datasets (crowdfunding.xlsx and contacts.xlsx) using **Pandas and Numby** in Jupyter Notebook. The primary goal was to transform the data into clean, structured tables for relational database storage.

Step 2: Data Transformations

1. **Campaign Data**:
    - The category & sub-category column in the crowdfunding.xlsx file was split into two separate columns: category and subcategory.
    - Unique values from these columns were extracted to create separate category_df and subcategory_df DataFrames, each with their corresponding IDs.
    - Before the merge, the launch_date and end_date columns had their columns adjusted to datetime format.
2. **Campaign Data Merging**:
    - The campaign_df DataFrame was created by standardizing the column names and then merging the campaign_df with subcategory_df on their respective columns (category, subcategory). This allowed us to associate each campaign with a specific category and subcategory.
3. **Contacts Data**:
    - Using Pandas to create the contacts Data file was loaded, and the contact information (ID, name, email) was extracted using regular expressions. The names were further split into contact_id, email, first_name, and last_name. Once the columns are cleaned, the CSV file is exported.

Step 3: Saving Data as CSVs

Once the data was transformed, the resulting DataFrames (category_df, subcategory_df, campaign_df, and contacts_df) were saved as CSV files to facilitate future use and data loading into a relational database system.

---

3. Part 2: Database Design

Objective:

Design a normalized relational database schema for the crowdfunding campaign data, ensuring minimal redundancy, scalability, and efficient querying.

Design Decisions:

1. **Category Table**:
    - This table stores unique campaign categories. The category_id serves as the primary key. By referencing this table in the campaign table, we eliminate redundant category data.
2. **Subcategory Table**:
    - This table stores unique subcategories related to campaigns. It uses subcategory_id as the primary key and links to the category_id in the category table.
3. **Campaign Table**:
    - This table holds detailed campaign information. The campaign_id serves as the primary key, and the table references the category_id and subcategory_id from the category and subcategory tables, respectively.
4. **Contacts Table**:
    - This table stores contact information for individuals involved in the campaigns. The contact_id serves as the primary key.

Relational Integrity:

Foreign keys were used to maintain relationships between tables (e.g., category_id in the campaign references category table). The schema was designed to ensure data integrity and normalization.

Diagram Creation:

The database schema was visualized using **QuickDBD**, a tool that allowed us to create the Entity-Relationship Diagram (ERD) for the database design. The ERD clearly shows the relationships between tables, including foreign keys and constraints (such as NOT NULL for essential fields).

- **QuickDBD Diagram**: The ERD was saved as a PNG image for submission.
- **QuickDBD TXT File**: A text file containing the physical ERD code was also provided for documentation.

---

4. Part 3: Data Loading into PostgreSQL

Data was imported to Posgres by creating the crowdfunding_db into a new server, we connected the server and created the database. After, the import tool was used to import the CSV for the DataFrames.

Part 4: Exploratory Data Analysis (EDA)

For the EDA, we used SQL queries to derive insights from the dataset. Below are the key findings from the three queries we executed:

1. Query 1 – Combined the Company Name, Description, Category, and Subcategory columns to visualize a more consolidated picture of these columns.
2. Query 2 – Combined Company Name, Description, Outcome, and Goal to compare which companies made their goals or did not see success.
3. Query 3 – Combined Last Name, Email, Total Pledged, and viewed who pledged the most with a baseline of more than $10,000.

Visualizations

Two visualizations were generated to complement the analysis:

1. **Distribution of Funding Goals for Failed Campaigns**: A histogram showing the number of failed campaigns and how much money they were anticipating raising.
2. **Total Pledges by Contact Last Name**: A bar chart displaying the most pledged per individual.

Conclusion

Conclusion

The analysis of the crowdfunding dataset uncovered several valuable insights into the dynamics of campaign performance and contributor behavior. Key findings include:

Campaign Success Factors: A significant number of campaigns failed to meet their funding goals, with failure often correlating to higher funding targets. This suggests that campaigns with overly ambitious goals may be at a greater risk of underperformance.

Contributor Patterns: Most contributors were individuals with relatively short names, suggesting a potential trend or bias in the dataset toward certain types of users. Additionally, the most active contributors tended to pledge larger amounts, with several individuals pledging amounts above $10,000.

Email Domain Insights: The analysis of email domains revealed a clear dominance of Gmail users, which could indicate a larger user base from specific regions or demographics using the platform.

Normalization and Data Integrity: The relational database design effectively minimized redundancy and ensured data integrity, providing a solid foundation for further analysis and future expansions of the database.

Overall, this project highlighted both the power of structured database design for handling large datasets and the potential insights that can be gained through exploratory data analysis in the context of crowdfunding platforms. The visualizations and SQL queries successfully revealed trends that can inform future campaign strategies and platform improvements.