October 15, 2023

**DATA 607 PROJECT 3: MOST VALUED DATA SCIENCE SKILLS**

1. **GROUP MEMBERS**
2. STEPHEN PHILLIPS
3. DIRK HARTOG
4. LWIN SHWE
5. **COMMUNICATION**

For Communication, we will be using slack to keep in touch on our daily progress and ZOOM for when we need to communicate in person.

1. **CODE SHARING**

We will be pushing all our code to a GitHub repo that is shared between the three of us.

1. **PROJECT DOCUMENTATION**

We will document our changes in our GitHub pushes, as well as in the comments that we leave on our R markdown code.

1. **DATA SOURCE**

We will be sourcing the data from https://coinmarketcap.com/currencies where the time series data with crypto curries prices varying in time for a month on top five coins are readily available in csv formatting.

1. **RELATIONAL DATABASE MODEL**

Our model for the database will be multiple data sets which contain time series data for different categories within each of them. They will be related to each other by the time in which the price occurred. A normalized RDB was created using MySQL workbench software on a local server and connecting to R studio. The five most popular cryptocurrencies were included in our data set; Bitcoin (BTC), Ethereum (ETH), Tether (USDT), Dogecoin (DOGE) and Cardana (ADA). Datasets of csv files were separately downloaded from the link as stated above and those were combined into one excel sheet. A new column called “CryptoName” was added to connect to the “DailyPrice” table that was created in SQL as a child table. In addition, a parent table with primary id’s was built with two main columns called “CryptoName” and “CryptoDescription”. The values in this table are declared once but used many times in the “DailyPrice” table and prevents from entering invalid data names. Utilizing 1NF (first normalization) when creating our relational database will reduce or eliminate data redundancy. Finally, we connected to the database in a R markdown file, queried and loaded the data into a data frame. We exported the data frame into a csv file to Github to be read into R again for further analysis.

1. **ENTITY-RELATIONSHIP (ER)DIAGRAM**

# A diagram of a blockchain Description automatically generated

We hope to answer the question “Which are the most valued data science skills? by working with real data using technical and non-technical skills to gather, store and analyze the data.  Our exploratory data analysis of the cryptocurrencies will include identifying and visualizing trends in the prices over time, relationships between the variables, and we hope to provide insights to events that may have impacted the price of cryptocurrencies (ex. did the collapse of FTX have an impact on the coin price and if so, have they recovered since that time).  Along with this practical application of data science skills we also hope to summarize in a document the skills that were most valuable to us as a team during this project.

Exploratory Data Analysis (EDA) is a method used to analyze and investigate datasets and summarize their main characteristics, often using visual methods. Here's a procedure to perform EDA on a cryptocurrency dataset:

1. **Data Collection**

Collect the data from a reliable source in the form of CSV file and build a database and data frame. The data include features like the Date, opening price, closing price, highest and lowest prices, volume, and market Cap.

1. **Data Cleaning**

Check for missing values and outliers in the dataset. If there are any, decide how to handle them. This could involve removing rows with missing data, filling in missing values with the mean or median, or using a method like regression to predict missing values.

1. **Data Transformation**

Convert the data into a format that's easier to work with. This could involve converting dates into a standard format or scaling numerical values.

1. **Data Visualization**

Use graphs and charts to visualize the data. This could include line graphs to show the price of the cryptocurrency over time, bar charts to compare the volume of different cryptocurrencies, or scatter plots to show the relationship between different variables.

1. **Statistical Analysis**

Perform statistical tests to understand the data better. This could involve calculating the mean, median, and mode of the prices, or performing a correlation analysis to understand the relationship between different variables.

1. **Insights Conclusions**

Based on the visualizations and statistical analysis, extract insights from the data. This could involve identifying trends in the price, understanding the factors that influence the price, or predicting future prices.