Project Report

EXTRACT:

- Imported the original data sources based on the (3) URL links that were provided in our Proposal:
    1. **Tesla Autonomous Deaths data**
        a. contributors.csv
        b. faq.csv
        c. resources.csv
        d. suddenacelaration.csv
        e. tesla-deaths.csv
    2. **Tesla stock data from 2010 to 2020**
        a. Tesla Stock Data.csv
    3. **Elon Musk Tweets 2021-2010**
        a. Tweets 2010.csv
        b. Tweets 2011.csv
        c. Tweets 2012.csv
        d. Tweets 2013.csv
        e. Tweets 2014.csv
        f. Tweets 2015.csv
        g. Tweets 2016.csv
        h. Tweets 2017.csv
        i. Tweets 2018.csv
        j. Tweets 2019.csv
        k. Tweets 2020.csv
        l. Tweets 2021.csv

TRANSFORM:

- Removed the following CSV files; Either the excels were not relevant or there were specific years in which some of the information was not available so in order to be consistent, we elected to only keep data pertaining to the years 2013-2020.
    1. Tesla Auto Deaths Data-contributors.csv; excel file that included a minimal list of sources that contributed to the Tesla Auto Deaths Data information was not needed.
    2. Tesla Auto Deaths Data-faq.csv; excel file with minimal FAQ’s was not needed.
    3. Tesla Auto Deaths Data-resources.csv; excel file with URL listings indicating sources utilized for Tesla Auto Deaths Data information was not needed.
    4. Tesla Auto Deaths Data-suddenacelaration.csv; excel not needed
    5. *Only the Tweets 2020.csv file was utilized.* All other Tweets csv files were discarded, as this file included all tweets from 2010-2020.

- Transformed all relevant datasets from .csv files in Pandas DataFrames
- Cleaned the Tesla Stocks DataFrame

I. Renamed all columns to lowercase headers to maintain consistency with tables created in postgreSQL.
II. Reset index to display DB from Start Date: 01-01-2013 to End Date: 12-31-2020

- ❖ Cleaned the Tesla Tweets DataFrame to reflect the timeframe of our interest.
  I. Dropped unwanted columns.
  II. Separated the date from the datetime column. Replaced datetime column with just a column containing the date, to match our other datasets.
  III. Checked for Null values.
  IV. Dropped all tweets from 2010-2012 as these years are not represented in our other datasets.
  V. Reset index to display DB from Start Date: 01-01-2013 to End Date: 12-31-2020

- ❖ Cleaned the Tesla Deaths DataFrame
  I. Dropped unwanted columns.
  II. Dropped any rows with Null values.
  III. Reset index to display DB from Start Date: 01-01-2013 to End Date: 12-31-2020

LOAD:

- ❖ Created a new database in PostgreSQL titled Tesla_DB
- ❖ Developed an ERD to choose which data columns would need to be labeled as primary keys.
  - ➢ Using the Query Tool we created the following tables:
    - ■ **tesla_deaths** [primary key was set to 'id' as SERIAL]
      - **date** was set as DATE for querying purposes.
      - **country** was set as a VARCHAR with a 30 character limit, no countries were found to have longer names than 30 characters.
      - **description** was set as a VARCHAR with a 500 character limit, as each incident varied in details with none exceeding over 500 characters.
      - **deaths, tesla_driver, tesla_occupant, cyclists_peds, tsla_cycl_peds** explained the number and types of victims involved in each incident was set as INT.
      - **autopilot_claimed** explained if autopilot was blamed as the cause for the incident.
      - **verified_tesla_autopilot_death** explained if the autopilot feature was the root cause based on Tesla's confirmation.
    - ■ **tesla_tweets** [primary key was set to 'id' as BIGINT]
      - **date** was set as DATE for querying purposes
      - **tweet** was set as a VARCHAR with a 1000 character limit to account for variance in tweet length
      - **username** was set as a VARCHAR with a 10 character limit

  - **nlikes, nreplies,** and **nretweets** were all set as INT
- **tesla_stock** [primary key was set to ‘id’ as SERIAL]
  - **date** was set as DATE for querying purposes
  - **open, high, low, close,** and **adj_close** were all set as DEC as these values contained decimal points
  - **volume** was set as INT

- Created a connection into PostgreSQL for **Tesla_DB** database using the create_engine and table_names functions.
- Imported tables using the to_sql function and setting the index as false, due to a new index that will be created once imported into PostgreSQL.
- Confirmed import for all tables: **tesla_stock, tesla_deaths, tesla_tweets -** using pd.read_sql_query function.
- Closed the engine connection.