**DSC 540: Data Preparation – Final Project Summary**

**Title : WEATHER DATA PRESENATATION & VISUALIZATION**

Author: Venkat Jagadeesh Jampani

Date: August 12, 2022

**DSC 540: Final Project Summary**
**August 12, 2022**

As a part of this class project final project, I have considered three datasets regarding Weather details(temperature, pressure) based on the different cities in USA. The three datasets are linked based on the CITY, STATE of a given LOCATION.

1. **Data source as a flat file** - A Data from the source, is taken that contains multiple CVS files that contains attributes of city names, temperature, pressure, humidity details for the past.

   Data Source details:

   1. CSV data source, Kaggle; https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data

   In the Data source, there are multiple CSV files with names :

   a. City attributes – With details of city names
   b. Temperature – With Temperature details for that city.
   c. Pressure – With Pressure details for that city.
   d. Humidity- With Humidity details for the city.
   e. Wind – With wind details for that city.

   2. API Data Source : Data Pull from API, Openweather
      a. API download: https://openweathermap.org/api

   In this Data source from the API, if we pass a ZIP code or City Name and State , all the weather attributes are displayed for that city. The attributes like : City Name, State, current temperature , High and Low temperature for that date , Pressure , Wind for the given date.

   3. Web Site data : Website Data, factiva.com
      http://www.factiva.com/CP_Developer/ProductHelp/FDK/FDK20/registration/shared_code_tables/state_province_table/united_states_state_province_table.htm

This Website writes the all US state names and codes. This can be used to join the other 2 data sources by state and city.

2. **Milestone : 2, 3, 4**:
    a. As a part of these mile stones, the data cleansing is done from each of the sources. The following tasks are performed from each of the source.
    b. From Flat file :
        i. Data is read from the csv file.
        ii. Loaded into data frame
        iii. Headers are updated with the required names.
        iv. Validated the data for blank values and removed them.
        v. Identified the unnecessary columns and deleted from the data frame.
        vi. Corrected the values if required in each column of the data.
    c. From Website :
        i. Loaded the data from website to the Data Frame.
        ii. Validated if all the required attributes are present in the data.
        iii. Updated the Column headers with the required names.
        iv. Deleted/corrected the unnecessary data.
    d. From API :
        i. Identified the weather api and understood how it is working.
        ii. Validated the data that was written from weather API. There are multiple attributes for a given city or zip code.
        iii. Converted the returned data into JSON format for ease of access.

iv. Identified the required weather attributes from the data written from the API.

v. Loaded the data into a data frame. Updated the headers with the required names.

vi. Validated the attributes for the few examples of weather attributes for zip codes and cities.

vii. Exposed the data set to pandas_profiling to identify the correlation between variables missing values and outliers in the dataset.

3. **Milestone 5**: I have loaded individual data from flat file, website, API data frames to separate tables using SQLite. I have encountered few issues along with way with data extraction from website and API but figured out in the end and loaded the final set of data to the API table. Created few visualizations and mainly focused on Temperature Attribute of the Weather attribute.

**Data Source References:**

Data Source 1: CSV data source, Kaggle
https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data

Data Source 2: Data Pull from API, Openweather
API download: https://openweathermap.org/api

Data Source 3: Website Data, factiva.com
http://www.factiva.com/CP_Developer/ProductHelp/FDK/FDK20/registration/shared_code_tables/state_province_table/united_states_state_province_table.htm