Project 3 – Team 1– Data Engineer Track

Group Members:

Kayla Blais

Andrew Bullard

Daniel Pulliam

Savi Rahiman

Madeline Rondino

Data and Delivery (35 points)

* The GitHub repo has a README.md that includes the following: (10 points)
  + Team 1 has chosen the path of Data Engineering. The industry we have selected is Leisure & Travel. The objective of this project is to explore the most populated cities of the US ranked by the top 5 and bottom 5 affluent States through the lens of household spending.
  + The goal is that the result of the database design will be useful for future users to access and plan their very own vacation by leveraging the table views from SQL to preform analysis, create stimulating visualizations, and have informative conversations.
  + In selecting our dataset the team was careful to select reputable sources erring on the side of caution. We wanted to ensure there was no bias in our data such as Algorithmic bias which refers to situations in which automated systems have a bias either in favor of or against certain groups of people. Lastly, we were careful to not include data that would have any PII (Personally Identifiable Information) to maintain privacy and prevent identity theft.

With these items top of mind the team selected census.gov to obtain the primary data set.

* For this project the team selected Pandas DataFrame. It was selected as the data collected from our sources are Two-Dimensional Structure with rows and columns. It will also allow users interacting with the data to utilize Labeled Axes to access, manipulate and perform analysis. (10 points)
  + Folder Name - Census Data
    - Census Business Builder – Primary Variable is State, Secondary Variable is Consumer Spending, and Secondary Categories is Percent working age (25 to 64 years)
      * MapData.csv is a list of all US states, estimate of consumer expenditures per household as well as the percent of the working population age 25-64. This is the list from where the team captured the top 5 and bottom 5 affluent states. This was derived from the estimate of consumer expenditures per household. Top 5 States in order is DC, Maryland, Massachusetts, New Jersey & Washington. Bottom 5 States in order is West Virginia, Mississippi, Louisiana, Arkansas & Alabama. (link: https://cbb.census.gov/cbb/#view=map&industries=00&clusterName=Custom+Industries&geoType=state&dataVariable=45&dashboardVars=15-17-33-64&centerX=-10802692&centerY=4568679&level=4&theme=default&dataVariable1=291&dynHeader=Custom+Region)
      * City and Town Population Totals 2020-2023 for the United States from census.gov was downloaded to get the population for each city in the US. The highest populated city from the top 5 and bottom 5 affulent States were noted. Washington city, District of Columbia, Boston city, Massachusetts, Newwark City, New Jersey, Baltimore city, Maryland, Seattle city, Washington, Huntsville city, Alabama, Little Rock city, Arkansas, Charleston city, West Virginia, New Orleans city, Louisiana and Jackson city, Mississippi. (link:https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html)
      * For each of the State a detailed view report was exported to a csv file to include demographic & socioeconomic characteristics. These 10 csv files were clean and prepped using python. (link: <https://cbb.census.gov/cbb/#view=report&industries=00&clusterName=Custom+Industries&geoType=state&dataVariable=179&dashboardVars=15-17-33-64&centerX=-8597980&centerY=4694599&level=7&theme=default&geoId=24&dynHeader=Custom+Region>)
  + Folder Name – Virtual Crossing/Weather Data
    - Weather data was for each of the city was obtained and downloaded via a csv file on <https://www.visualcrossing.com/>.
    - The 10 csv files were cleaned, prepped and merged using python to
  + Folder Name – Yelp Data
    - Yelp Fusion API was leveraged to obtain local restaurants within each of the 10 cities. The results were filtered for name, location, categories, ratings, price and reviews.
    - The process of creating the DataFrame and exporting to a csv file was done using python.
* The database includes a total of 9,429 records. (5 points)
  + Top5.csv has 5 rows of data for state, average household income, dining out (per household), entertainment/recreation (per household) & median household income.
  + Bottom5.csv has 5 rows of data for state, average household income, dining out (per household), entertainment/recreation (per household) & median household income.
  + all\_city\_weather.csv has 7,300 rows of data for city name, datetime, tempmin, tempmax, sunrise, sunset, icon, daily\_avg\_temp & month.
  + Yelp\_data.csv has 2,119 rows of data for city name, location, categories, rating, price & reviews.
* In addition to the libraries covered in class the team also imported The Great Expectations library for Python. The Great Expectation Library is a powerful tool for data validation, profiling, and documenting. This library was used in city\_weather.ipynb to help verify that the data we imported was in Fahrenheit and not Celsius with a range such as “weather\_ge.expect\_column\_values\_to\_be\_between('tempmax', -40, 50)”. Because the values were outside of the ordinary range for Celsius this was proven successful to indicate the data imported was in Fahrenheit.
  + import matplotlib.pyplot as plt
  + import pathlib import Path
  + import requests
  + import pandas as pd
  + from pprint import pprint
  + from datetime import datetime
  + import great\_expectations as ge

Database Design (40 points)

* The project uses ETL workflows to ingest data into the database. (10 points)
* The original dataset(s) are transformed prior to storing it in the database. (5 points)
  + Census Cleaning…
  + Yelp Cleaning…
  + Visual Crossing Cleaning…
* The database is used to house the data in SQL. (5 points)
* The database has xxx tables in SQL. (5 points)
* SQL (Structured Query Language) was the choice selected for building the necessary data base for several reasons. SQL is known for its proven stability, it is widely used, is standardized by ANSI(American National Standards Institute, is structured data in Tabular Format and scalability . (5 points)
* The project includes documentation of the ETL workflow with diagrams or ERD. (10 points)