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 we do not have 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 get the primary data set.

* + Census Business Builder – Primary Variable = State, Secondary Variable = Consumer Spending, and Secondary Categories = Percent working age (25 to 64 years)
    - <https://cbb.census.gov/cbb/#view=map&industries=00&clusterName=Custom+Industries&geoType=state&dataVariable=179&dashboardVars=15-17-33-64&centerX=-10802692&centerY=4568679&level=4&theme=default&dynHeader=Custom+Region>
* The database includes a total of xxx records. (5 points)
  + Census Records…
  + Yelp Records…
  + Visual Crossing Records…
* In addition to the libraries covered in class the team also imported xxx. Xxx was used in notebook xxx leveraging data sourced from xxx to capture xxx (10 points)
  + import matplotlib.pyplot as plt
  + import requests
  + import pandas as pd
  + from pprint import pprint
  + from datetime import datetime, timedelta
* 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)

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)