Zoe Erpelding

Final Project

CPSC324 Spring 2024

Description

(a). Describe whether you are developing an “end-to-end” analytics project or diving deeper into a particular technology as well as the services you plan to use. Note that if you decide to develop an “end-to-end” project, you should look at Google Cloud Composer as a tool to help automate the overall process.

I developed an “end-to-end” analytics project. First I loaded in the data using the Cloud Shell into a bucket in Google Storage, then to BigQuery, where I explored the data using queries. Within BigQuery, I joined the tables and cleaned the data. I used BigQuery ML to create a model to predict the heart disease mortality rate based on the combined data’s attributes. I ended up deciding not to use Composer because I did not have streaming data so would not need to repeat queries. I used Looker Studio to create visualizations to better understand the relationship between attributes.

Datasets

(b). Identify and describe the datasets you are using including their location, their general contents, and challenges inherent in the data.

[Medicare Monthly Enrollment - Catalog (data.gov)](https://catalog.data.gov/dataset/medicare-monthly-enrollment)

[US Household Income Statistics (kaggle.com)](https://www.kaggle.com/datasets/goldenoakresearch/us-household-income-stats-geo-locations)

<https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-county-2018-2020-3a2b0>

I originally planned to use the three datasets linked above. Two of the three are from my project last semester, US Household Income and Heart Disease Mortality. I wished to also compare these to Medicare Enrollment, to analyze the connections between enrollment in Medicare in a county, the county’s income, and the heart disease mortality rate. I downloaded the data to my computer from Kaggle and Data.gov before uploading it to Cloud Storage. The Heart Disease Mortality dataset has the location (state and county), the heart disease mortality rate per 100,000, gender, race, and latitude and longitude. The Income data has location (state and county), median income, poverty percentages, and similar statistics. The Medicare data has what type of insurance the group has, the state, and prescription drug coverage. I came across issues with the Income data. While it would load into the Cloud Storage bucket, there were too many errors in the data for it to be loaded into BigQuery. I tried to fix some of the errors manually before uploading it, but the dataset was too large. Instead, I found a different but similar dataset in BigQuery public datasets. This dataset has a location (City and State), unemployment, income, population, and similar attributes. The location is all in one attribute, so I needed to separate the two. Even after separating the city and state, there were still issues with being able to combine the data with the Heart dataset. It would run the query, but then say that the results had no rows. After searching for another avenue and coming up empty, I decided to drop the 3rd dataset and keep with the two.

Goals

(c). Outline the goals of a corresponding demonstration project. The goals should be driven by the datasets you select, including what insights you hope to achieve and/or what data products you hope to produce. Note that the goals will also depend on the services you plan to use.

My goals include using Cloud Storage, Big Query, and Looker Studio. From Cloud Storage I will use buckets to load in the datasets. From there, I will use Cloud Shell to get the datasets from storage to Big Query. In Big Query, I will use SQL statements to query the data and create and evaluate a prediction model. I would like to create Looker reports/dashboards to make information easily accessible to people. Then I will use Looker Studio to create visualizations. Since I have latitude and longitude from the Heart dataset, I want to turn the two attributes to a single geographical attribute using the CONCAT() function so that I can create a bubble map. I think that a visual representation of the heart disease mortality rates by location will give insight into what areas (rural vs urban, the north vs the south) have unusually high rates of heart disease mortality. I also want to determine if there is a correlation between income, medical insurance rates, and heart disease mortality rates. I want to see if I can make a machine learning model to do linear regression to predict heart disease mortality rates off of the attributes of these datasets.