ETL Project Overview

09/03/2020

Butler University

Grant Leach

Tomorrow Rose

Catherine Smith

Introduction

For this project, we analyzed World Happiness data from two csv files found on Kaggle. Our analysis involved extracting, transforming, and loading the data into a database. In this technical summary we will explain these steps in detail.

Summary

We started our project by searching for data to use for our analysis. While searching Kaggle, we came across CSV files that had data on World Happiness for the years 2015 to 2020. Once we saved the files, we began transforming the data. We used the csv files for the years 2018 and 2019. In a jupyter notebook, we dropped any duplicate data, transformed the data into dataframes, changed column names, and dropped unwanted columns. After completing the Jupyter Notebook we used SQL to create desired tables.

Extract

Our group visited Kaggle to look for a data set. When we started out, our goal was to find a data source with multiple csv files relating to the same topic. We came across information on the World Happiness Reports from 2015 to 2020 (<https://www.kaggle.com/mathurinache/world-happiness-report>). Attached were csv files for each year. Over the years, data collection looked slightly different. Therefore, not all years had similar columns. 2018 and 2019 had a very similar set up in terms of information gathered so we decided to compare those two years. We had already made a repository on github and cloned it to our local computers. We saved the csv files inside the local folder and began our transformation on the data.

Transform

For this section, we changed the names of the csv files and saved the csv files into the ETL folder we created. Then we created dataframes for the files in the Jupyter notebook for the years 2018 and 2019. We created a filtered dataframe from the specified columns and included a list of the column names for each year. Then we transformed the data by renaming the column headers for 2018. We cleaned the data by dropping the duplicates and setting the index for the first year by “rank”. Next, we created a filtered dataframe from the specified columns for 2019. Then renamed the column headers in our dataset. We repeated the process of cleaning that data by dropping the duplicates and setting the index by “rank”. We included the connection string for the postgres database. Finally, we created the “[transformed.to](http://transformed.to/)\_sql” functions for each year to create the tables in SQL.

Load

After we transformed the data into 2 dataframes with the desired format, we loaded the data into pgAdmin 4. We ran into some trouble trying to do this. When we tried to make the connection to pgAdmin we got an error that there was “no module named psycopg2.” After consulting with Fred, we learned that we needed to do the following line of code: “!pip install psycopg2.” From there, we ran queries to create 4 separate tables to compare the 2018 data vs the 2019 data. The first time we tried to run a query, there was no information in the tables. This was because we had accidentally loaded the data into the wrong database. When we realized this, we reloaded it into the correct database. We also had to adjust the path in our Jupyter Notebook. Once we fixed these two errors, our queries started working as expected.

We created a table for the Rank, Score, GDP Per Capita and Healthy Life Expectancy. Each of the tables were joined on the country so that we could easily compare the values for the years side by side. Once we queried the tables, we realized that we had lost decimal data and only had integers. This was because we entered columns with number data as INT. We went back and recreated the tables with those columns as FLOAT instead. Then when we reran the queries, we now had the decimals included as desired. The tables for Rank was ordered in ascending order while the GDP Per Capita, Score, and Healthy Life Expectancy tables were displayed in descending order. This allowed the tables to show countries with the most desirable outcomes at the top of the tables. We chose to use tables to display our data because it was the most efficient way to look to get a side by side comparison of scores from 2018 to 2019.

Results and Discussion

We were interested in seeing how the United States matched up with other countries on these measures. According to the tables in our analysis, the United States was ranked 18th overall in 2018 and 19th in 2019. In 2018, the United States was 18th for overall score. The score improved slightly in 2019 from 6.886 to 6.892. Finland had the highest score with 7.63 in 2018 and 7.769 in 2019. The United States ranked 10th for GDP per capita in 2018. The score increased from 1.39 in 2018 to 1.43 in 2019. Lastly, the United States was 33rd in the world for life expectancy in 2018 with a score of .819 and increased slightly in 2019 with a score of .874. The documentation on Kaggle for our data set says that the overall Happiness Score is determined from the following measurements: GDP per capita, Healthy Life Expectancy, Social support, Freedom to make life choices, Generosity, Corruption Perception, Residual error. Keeping an eye on these measures can be a good method for reflection for individual countries and the world as a whole. Seeing these reports can also give direction in terms of where to focus efforts for improvement and which countries to look to for inspiration.