ETL Project Report U of O Data Analysis Bootcamp December 2020

PURPOSE: The purpose of this project was to demonstrate our team's ability to Extract, Transform, and Load data. We developed an ETL workflow, first extracting US opioid prescription drug sales data from the web, then transforming the data via Python into a sufficient format for answering our business question, and finally loading the data to Postgres for analysis.

STRUCTURE OF THIS REPO DIRECTORY: The following three items comprise our key deliverables:
- The Jupyter Notebook titled "**Transform_Data(refined).ipynb**" contains our Python code that we leveraged in the TRANSFORM process.
- The CSV files titled "**OR_county_data.csv**" and "**WA_county_data.csv**"are housed in the CSV_files folder, and contain the cleaned data resulting from our TRANSFORM process.
- The SQL file titled "**SQL_code.sql**" contains the code we used to LOAD the final data into Postgres.

BUSINESS QUESTION: Our business question is to compare opioid prescription drug rates across two counties in the Portland, Oregon metropolitan area: Multnomah County in Oregon and Clark County in Washington.

NULL HYPOTHESIS: There is no difference between Multnomah and Clark county opioid prescription rates.

DATA SOURCE & SCOPE: Our data was sourced from Washington Post's website containing publicly available DEA data on US opioid prescription drug sales for 2006-2014. Our original datasets were full-state data for Oregon and Washington.
https://www.washingtonpost.com/graphics/2019/investigations/dea-pain-pill-database/#download-resources?itid=lk_inline_manual_10

Group Members: Aaron Garber-Paul, Sung Choo, Alejandro Barnatan, Brett Williams, & Adam Fancher

ETL Project Report U of O Data Analysis Bootcamp December 2020

EXTRACT: Our first step was to extract the datasets. We extracted full Oregon and Washington state datasets as CSV files from Washington Post’s website.

TRANSFORM: We performed several transformation steps to trim down the data to only the columns and rows needed to compare Multnomah and Clark counties, and to trim the files down to more manageable size. These steps included:

1. Read both CSV data files into a Jupyter Notebook for data exploration and clean-up.
2. Leveraged Pandas for data clean-up:
    a. Dropped superfluous columns, reducing the column count from 42 to 11. We created a Pandas dataframe for each county, keeping only the columns relevant for comparison (REPORTER_DEA_NO,REPORTER_BUS_ACT, REPORTER_NAME, BUYER_DEA_NO, BUYER_BUS_ACT, BUYER_NAME,BUYER_COUNTY, DRUG_NAME, TRANSACTION_ID, Product_Name , MME).
    b. Eliminated all counties except Multnomah County Oregon, and Clark County Washington. We leveraged the Pandas LOC function to keep only rows with MULTNOMAH or CLARK in the BUYER_COUNTY column.
3. Created separate dataframes for the cleaned-up Oregon & Washington data, and exported the dataframes to separate CSV files for loading into Postgres (for end-user analysis).

LOAD: We uploaded both datasets to Postgres, so that end-users (fictional data analysts) can analyze Multnomah & Clark county opioid sales, leveraging SQL in the Postgres environment. Load steps:

1. Created one table for Oregon and one table for Washington, defining the 11 columns and their respective datatypes. (Tables created via SQL CREATE TABLE.)
2. Loaded CSV data files to their respective tables in Postgres.
3. Data is loaded and ready for analysis! 😊

Group Members: Aaron Garber-Paul, Sung Choo, Alejandro Barnatan, Brett Williams, & Adam Fancher