Seasonal Produce

November 12th, 2020

Tigist Keneni & Fang Yu Chang

Overview

The project will aggregate data from different sources to provide users an overview of the cost of seasonal produce in the southwest USA and highest-producing countries for seasonal produce.

Sources

1. USDA Department of Agricultural https://snaped.fns.usda.gov/seasonal-produce-guide/
2. USDA Agricultural Marketing Service https://www.seasonalfoodguide.org/state/california
3. Wikipedia https://www.wikipedia.org

Documentation

■Datasets used and their sources.

■Types of data wrangling performed - Data cleaning, joining, filtering, and aggregating.

■The schemata used in the final production database.

Technical Report

* Extract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

- Our data sources were from a CSV file that was downloaded from the USDA Agricultural Marketing Service website and Wikipedia and USDA Department of Agricultural websites.

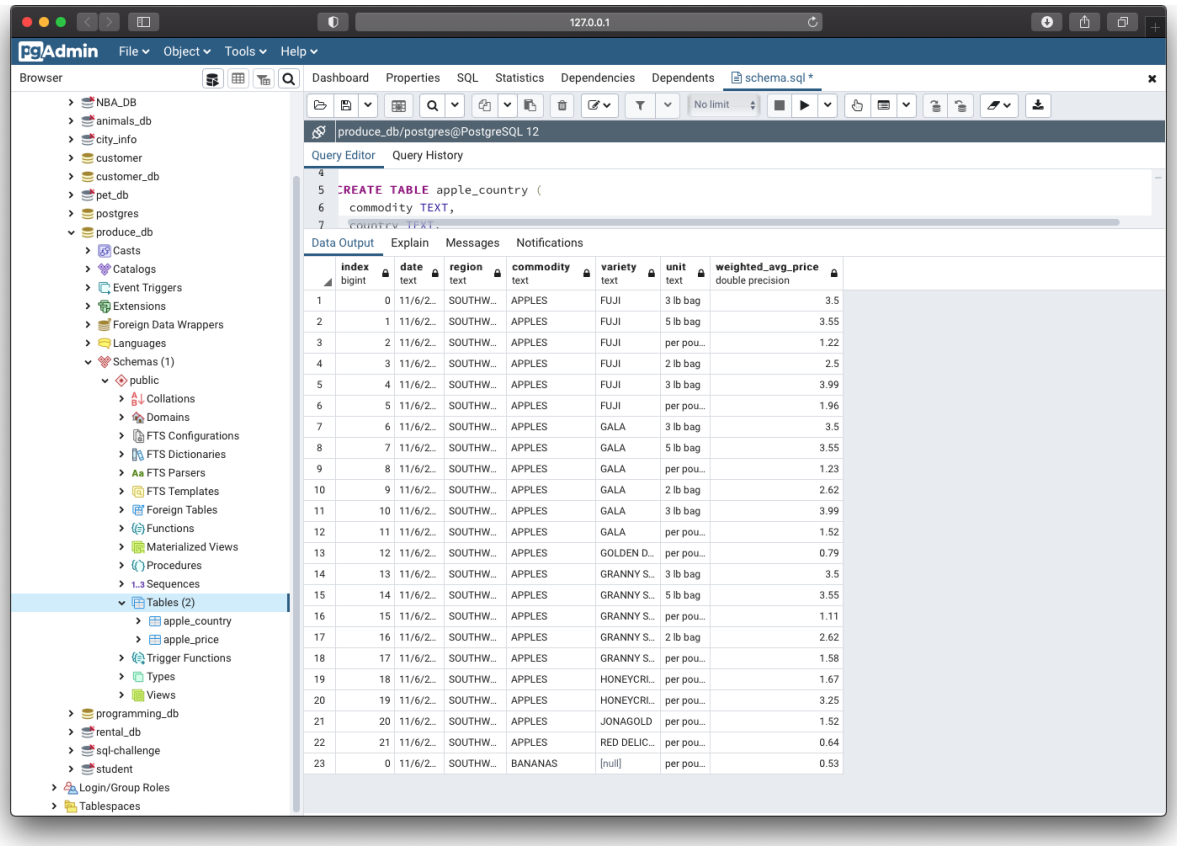
- USDA Department of Agricultural: The five produce lists were scraped from this website in order to get the fall seasonal produce lists. Reference file: "fall_produce_scraper.ipynb".
- USDA Agricultural Marketing Service: A CSV file for price in the southwest US was downloaded and cleaned and transformed using Jupyter Notebook. Reference file: "main_script.ipynb".
- After transforming, the data were imported into pgAdmin.

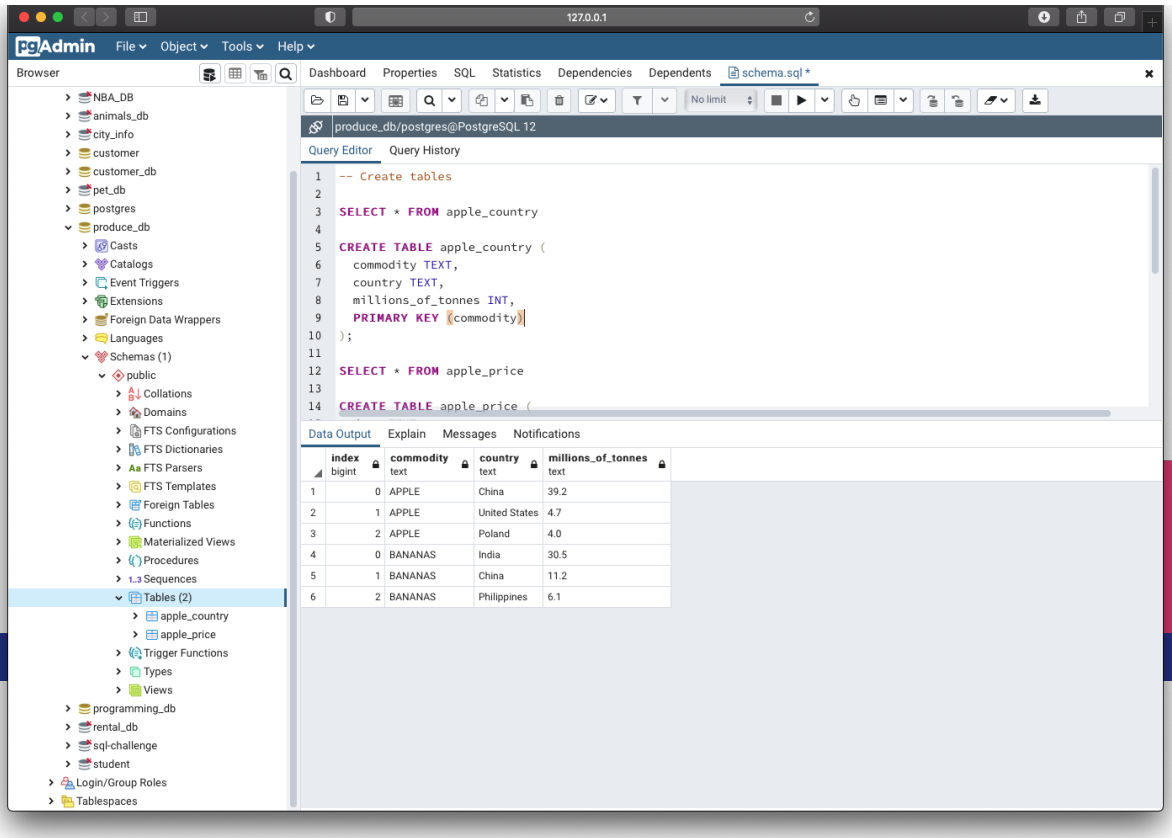
* Transform: required steps for cleaning and transformation of the data

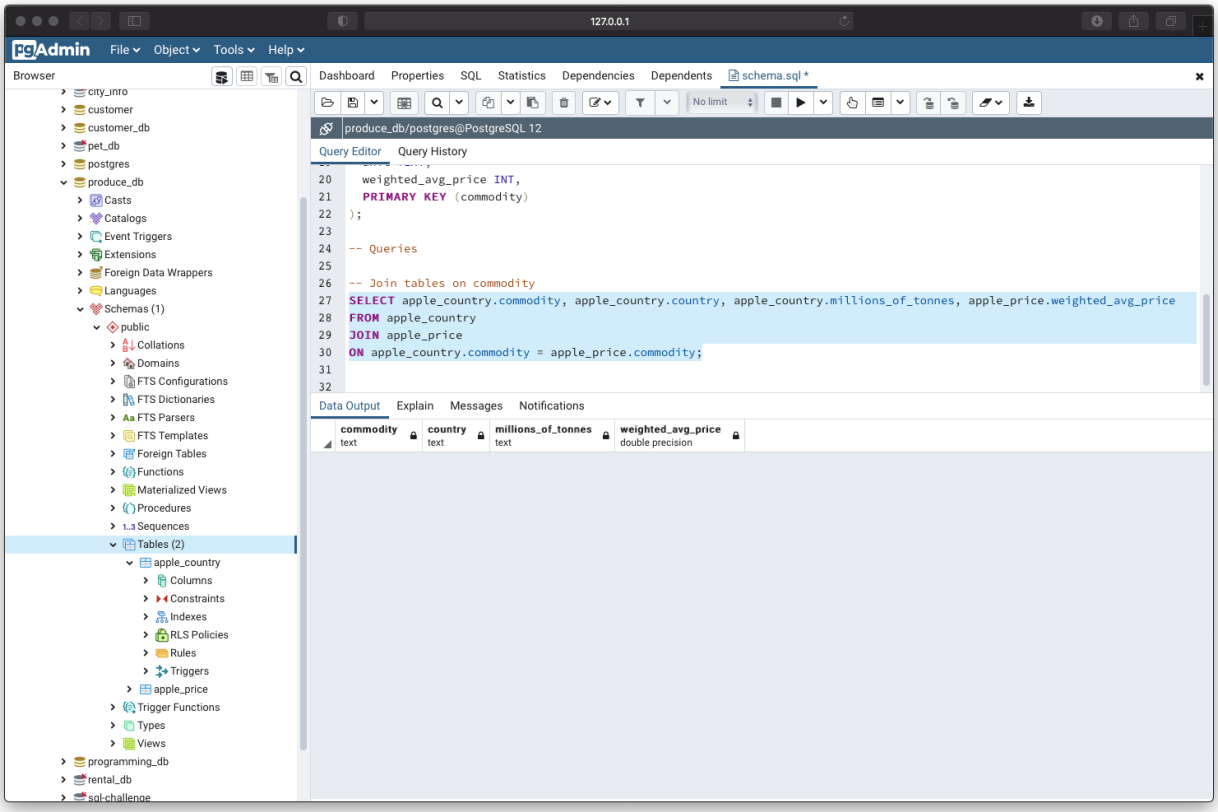
- For Wikipedia data, getting the appropriate columns, renaming the columns, and adding the column in the apple_country table were performed.
- Data cleaning was done during this project by selecting specific produce that is popular in the southwest USA. More data cleaning was done when we selected data from Wikipedia.

- Inner data joining with the country from Wikipedia and countries were collected from USDA Marketing Services. By joining these sets of data we noticed interesting trends that helped us identify where most of the popular produce was grown at.
- We filtered data that was most important to this project and selected specific produce information to identify which countries had the most production of apples.

* Load: the final database, tables/collections, and why this was chosen.


127.0.0.1
pgAdmin
File
Object
Tools
Help
Browser
Dashboard
Properties
SQL
Statistics
Dependencies
Dependents
schema.sql *
No limit
produce_db/postgres@PostgreSQL 12
Query Editor
Query History
CREATE TABLE apple_country (
commodity TEXT,
Data Output
Explain
Messages
Notifications
index bigint
date text
region text
commodity text
variety text
unit text
weighted_avg_price double precision
NBA_DB
animals_db
city_info
customer
customer_db
pet_db
postgres
produce_db
Casts
Catalogs
Event Triggers
Extensions
Foreign Data Wrappers
Languages
Schemas (1)
public
Collations
Domains
FTS Configurations
FTS Dictionaries
FTS Parsers
FTS Templates
Foreign Tables
Functions
Materialized Views
Procedures
Sequences
Tables (2)
apple_country
apple_price
Trigger Functions
Types
Views
programming_db
rental_db
sql-challenge
student
Login/Group Roles
Tablespaces


127.0.0.1
pgAdmin
File
Object
Tools
Help
Browser
Dashboard
Properties
SQL
Statistics
Dependencies
Dependents
schema.sql *
No limit
produce_db/postgres@PostgreSQL 12
Query Editor
Query History
-- Create tables
SELECT * FROM apple_country
CREATE TABLE apple_country (
commodity TEXT,
country TEXT,
millions_of_tonnes INT,
PRIMARY KEY (commodity)
);
SELECT * FROM apple_price
CREATE TABLE apple_price (
Data Output
Explain
Messages
Notifications
index bigint
commodity text
country text
millions_of_tonnes text
0 APPLE China 39.2
1 APPLE United States 4.7
2 APPLE Poland 4.0
0 BANANAS India 30.5
1 BANANAS China 11.2
2 BANANAS Philippines 6.1
NBA_DB
animals_db
city_info
customer
customer_db
pet_db
postgres
produce_db
Casts
Catalogs
Event Triggers
Extensions
Foreign Data Wrappers
Languages
Schemas (1)
public
Collations
Domains
FTS Configurations
FTS Dictionaries
FTS Parsers
FTS Templates
Foreign Tables
Functions
Materialized Views
Procedures
Sequences
Tables (2)
apple_country
apple_price
Trigger Functions
Types
Views
programming_db
rental_db
sql-challenge
student
Login/Group Roles
Tablespaces



Summary

- These final schemas identify relevant information that gave us insight into the question we started with, how much do popular seasonal fruits cost?

- From this data analysis, we were able to discover some of the most popular fruits and decided to focus on aggregating data from different sources. After looking at the data, we had 5 fruits to work with. We then decided to aggregate relevant data for one specific fruit which was apples.
- Some of our takeaways from working on this project was to ask a specific question and to keep iterating on the project scope as we discovered more information from the data that was available.