**Project Title:** Pesticide Data Program Database SQL Analysis

**Summary:** The Pesticide Data Program (PDP) is a national pesticide residue monitoring program and produces the most comprehensive pesticide residue database in the U.S. The Monitoring Programs Division administers PDP activities, including the sampling, testing, and reporting of pesticide residues on agricultural commodities in the U.S. food supply, with an emphasis on those commodities highly consumed by infants and children.

**Problem Question:** Are pesticide residues posing any risk for agricultural commodity consumption?

**Dataset Title**: Pesticide Data Program 2021-USDA

See link below for the dataset.

[**https://www.ams.usda.gov/datasets/pdp**](https://www.ams.usda.gov/datasets/pdp)

**Tech Stack:** SQL, Pandas, Power BI

**Tools:** MySQL Workbench, Jupyter Notebook, MS Visio

**Objective(s):**

* Explore the dataset using Pandas and SQL
* Explore cleaning the dataset using Pandas.
* Create queries below to analyse the dataset to generate insights using SQL
* Use Pandas as a control feature for the SQL analysis.
* Visualise the results of applicable queries in Power BI

**Tasks:**

* Load the dataset tables into MySQL.
* Create a data model and normalize the database.
* Execute the following queries:
  + Explore each table in the database
  + The number of samples in the database
  + The number of distinct samples in the database
  + The number of results in the database
  + The number of tests/results obtained per sample.
  + The countries from which samples was collected and number of samples.
  + Commodity sample collected from each country.
  + Total number of distinct countries samples was gotten from.
  + Countries from which test results was gotten.
  + Commodity results from each country
  + Count of results from each country from highest to lowest
  + Each Commodity with number of test results
  + Commodities samples not in test results.
  + Confirm the Concentration/LOD units to confirm if uniform for all the results.
  + Commodity sample with highest Limit of Detection (LOD)
  + Which pesticide has the highest Limit of Detection (LOD) per commodity?
  + What is the average Limit of Detection (LOD) per pesticide per commodity?
  + What is the maximum Limit of Detection (LOD) per pesticide per commodity?
  + How many tests did not detect any residue?
  + How may tests detected pesticide residue?
  + How may tests detected pesticide residue for each commodity?
  + Calculate percentage of the results that detected pesticide residue per commodity.
  + Confirm the Labs from which test results was gotten and the samples.
  + Count of results from each lab from highest to lowest.
  + What is the commodity type of each commodity with test results?
  + What are the most common confirmation methods used by each lab.
  + What is the most common determinative methods used by each lab.
  + What is the origin of each commodity?
  + Are there samples without COUNTRY ?
* Use Pandas to run a similar analysis of each query.
* Execute the query in a Jupyter notebook and save in a CSV file for further use. ( Use SQLAlchemy to connect to the database and run the queries from the Jupyter notebook.