**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

**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.

**Tasks:**

* Load the dataset tables into MySQL.
* Create a data model and normalize the database.
* Execute the following queries:
  + The number of samples in the database
  + The number of distinct samples in the database
  + The number of results obtained per sample.
  + The number of tests/results obtained per sample.
  + The country of origin of the distinct samples types
  + Country of origin with most samples
  + States with the highest sample collected.
  + Types of tests carried out on each sample.
  + Lab with the highest number of tests carried out.
  + Commodity with highest pesticide concentrations for each pesticide
* 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.