Project Title: Invasive Insects and Effects on US Agriculture Analysis

Group Details:
Ali Ugur (agu6)

Project Definition:
This project explores how non-native insect species, particularly scale insects, enter the United States and how their arrivals relate to agricultural and ecological factors. My motivation came from learning that not all insects function the same way across ecosystems. For example, European honey bees produce honey while most North American bees do not, and some ant species can only survive by enslaving workers from other colonies. These differences illustrate how introducing a species into a new environment can create unpredictable ecological outcomes. The goal of this project is to understand which factors help predict high risk years for invasive scale insect introductions and how variables such as trade activity, species origin, host plants, and geographic spread contribute to these patterns.

Introduction:
Invasive insects are a growing concern for US agriculture because they can damage fruit trees, grasses, ornamental crops, and other economically important plants. Scale insects in particular can spread rapidly, reduce yields, and disrupt natural ecosystems. As global trade increases, so does the unintentional introduction of non-native species into the country. Accidental movement through imported goods, shipping materials, or plant products plays a major role in the establishment of new pests. This project uses six publicly available ecological datasets to examine how insect introductions relate to trade volume, regional detection patterns, biological origins, and host plant categories. By integrating these datasets into a single analytical workflow, the project aims to reveal long term patterns and identify measurable factors that help explain the arrival of invasive species.

The datasets include yearly introduction counts from changepoint.csv, species origins from origin_anova.csv, first regional appearances from region_intros.csv, trade quantities from trade_intro_data.csv, and host plant categories from both 2005_hosts.csv and total_hosts.csv. These sources cover species level and year level information, allowing the project to connect economic activity with ecological outcomes. The approach taken in this project mirrors typical data science practices by performing cleaning, merging, visualization, SQL querying, feature engineering, and several prediction and clustering models that were demonstrated in the course.

Methodology:
The analysis began by loading all six datasets and inspecting them for formatting issues. Some files required encoding adjustments, such as origin_anova.csv, which contained special characters that needed Latin1 decoding. Numerical fields such as year, introduction counts, trade quantity, and total observations were converted into numeric types to ensure consistency. Species level data in region_intros.csv were aggregated by year to create a summary of how many distinct regions recorded new detections and how many total observations occurred. This produced a clear year based dataset for comparison with introduction and trade data.

After cleaning each dataset, the project merged information from changepoint.csv, trade_intro_data.csv, and the year summarized regional data into a single table organized by year.

This merged table was saved as year_level_table.csv. It contains variables such as yearly introduction counts, trade quantity, number of regions with detected species, and total regional observations. A decade variable was also added to facilitate temporal grouping.

Feature engineering added several new variables that help improve model interpretation. These include a high risk label for years with above median introductions, a lagged introduction variable that represents temporal momentum, a percentage based trade growth metric, and a pre or post changepoint indicator that identifies potential shifts in ecological or trade behavior over time.

To satisfy the database component of the course, the engineered dataset was exported into a SQLite database stored in the file insect_risk.db. SQL queries were used to compute decade level averages for introductions and trade. These queries confirmed long term upward trends and reinforced the visual patterns observed earlier. The database step highlights how relational systems can be used alongside Python to analyze ecological data from multiple perspectives.

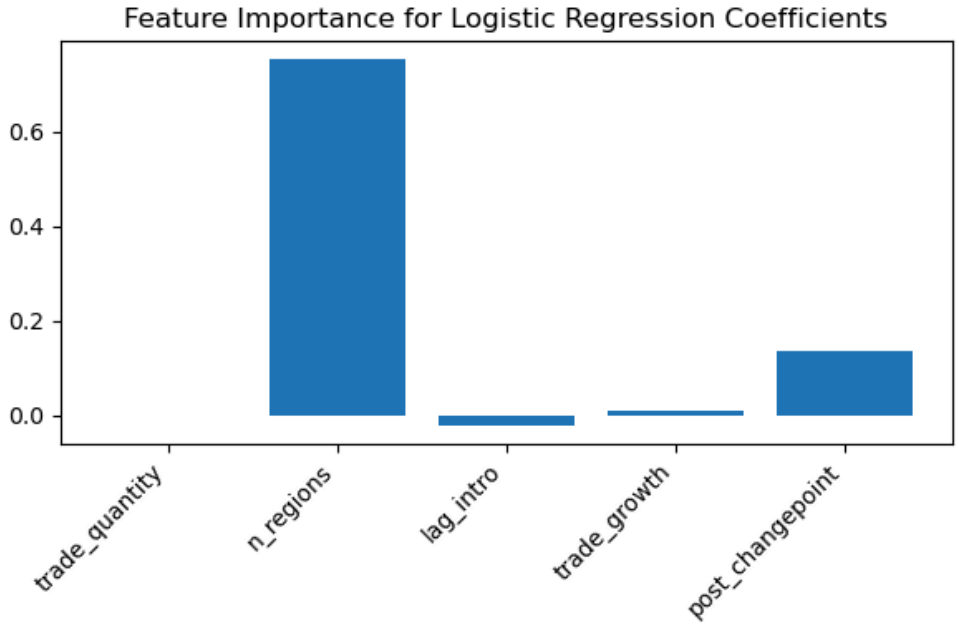
The machine learning phase of the project included linear regression, logistic regression, K-Means clustering, and K-nearest neighbors similarity analysis. Linear regression predicted the number of introductions per year using trade quantity, number of regions, lagged introductions, and trade growth as predictors. Logistic regression classified years as high risk or low risk based on these same engineered features. K-Means clustering grouped years into ecological and economic categories, while KNN identified which years most closely resemble each other.

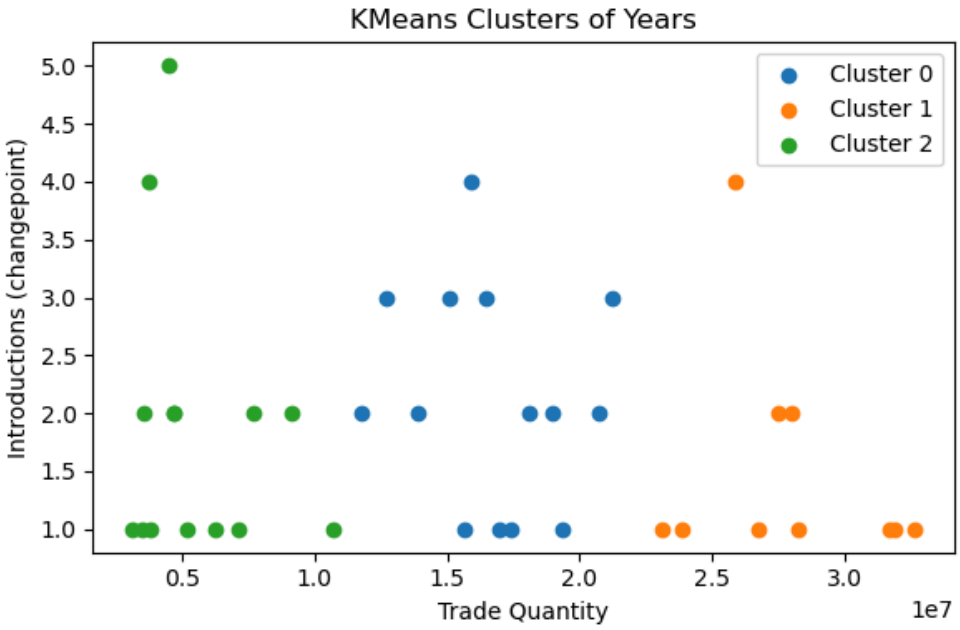
Results:
The visualizations produced early insights into how ecological introductions behave over time. A line plot of annual introductions from changepoint.csv shows long term fluctuations with noticeable increases during certain decades. The scatterplot comparing trade quantity and introduction counts reveals a positive relationship, suggesting that periods of higher trade activity tend to coincide with greater numbers of invasive species arrivals. Host plant bar charts created from 2005_hosts.csv and total_hosts.csv show that fruit producing plants, grasses, and multi host species appear frequently, confirming that these agricultural categories are particularly vulnerable to scale insect invasion.

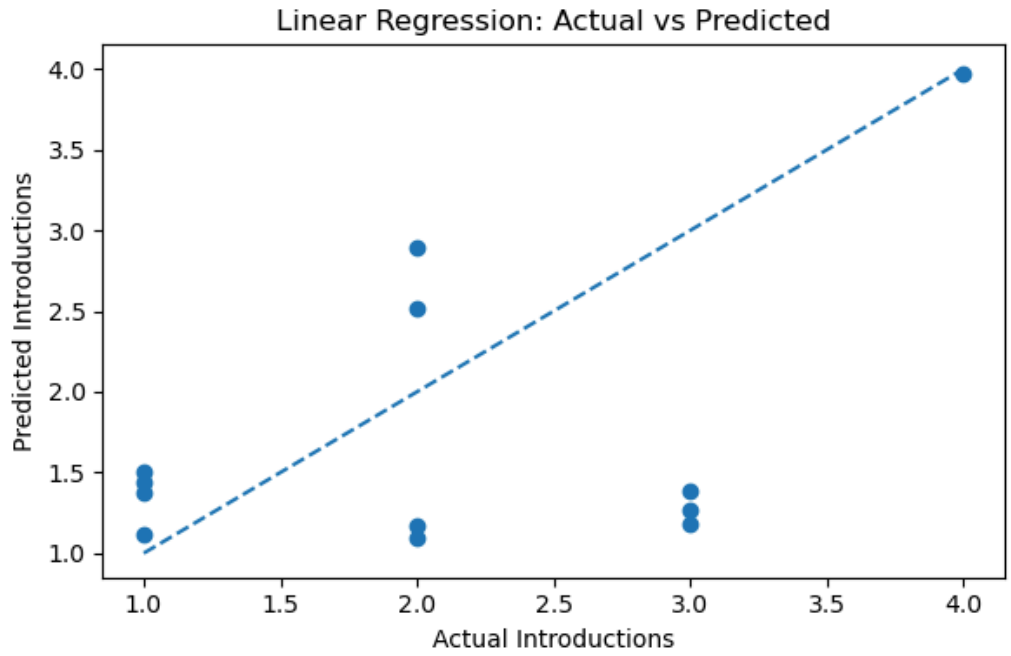
The SQL analysis performed on insect_risk.db shows a steady rise in both introductions and trade volume beginning around the mid-twentieth century. Light numerical patterns indicate that later decades consistently have higher average introductions and larger trade quantities than early decades. This reinforces the idea that globalization amplifies ecological risk through increased movement of goods.

The machine learning results provide additional confirmation. The linear regression model produces predictions that are generally within about one species of the actual introduction values, suggesting that ecological momentum, trade activity, and geographic spread work together to shape yearly outcomes. The logistic regression model achieves moderate accuracy and demonstrates that increases in regional detections and trade growth are strong signals of a high risk year. The feature importance plot shows that the number of regions with detections has one of the strongest positive influences on high risk classification, while lagged introductions and post changepoint values also contribute meaningfully.

K-Means clustering reveals three distinct patterns among the years. One cluster represents low introduction and low trade periods. A second captures transitional periods with moderate trade and moderate introductions. The third cluster contains the highest introduction and highest trade years, emphasizing how these factors align historically. The cluster scatterplot clearly separates these groups. The KNN results show that recent years resemble each other more closely than they resemble earlier decades, which is consistent with modern increases in global trade and ecological connectivity.







References:

https://catalog.data.gov/dataset/data-from-non-native-scale-insects-hemiptera-coccomorpha-of-the-united-states-and-their-im