Invasive Insects and Effects on US Agriculture Analysis

By: Ali Ugur (agu6)

Motivation

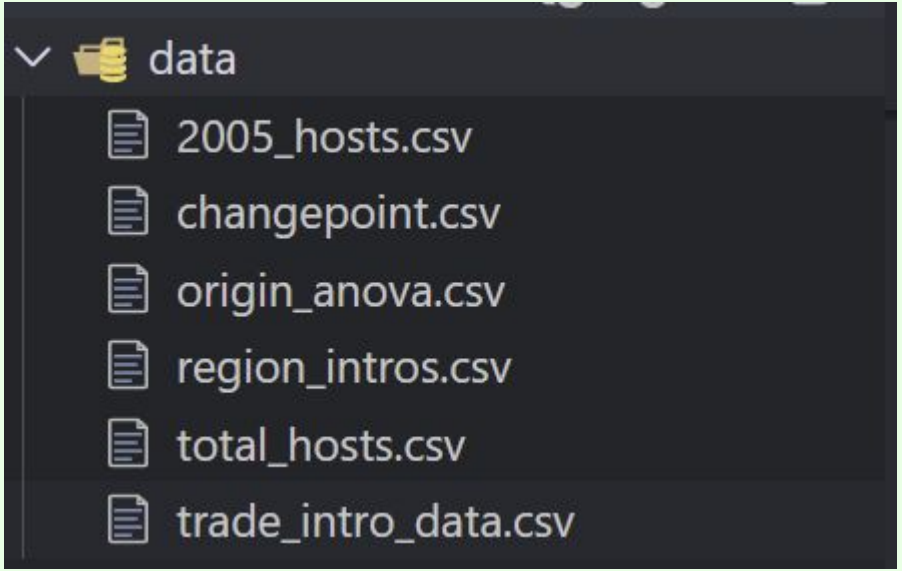
The reason as to why I did this project was:

- Invasive insects behave differently across ecosystems, best example is European honey bees produce honey, but most native North American bees do not
- These differences show how introducing species into new environments can create unexpected ecological impacts
- This project investigates non-native scale insects in the United States and identifies ecological and economic factors that predict high-risk introduction years

Datasets Used

Six ecological datasets were combined:

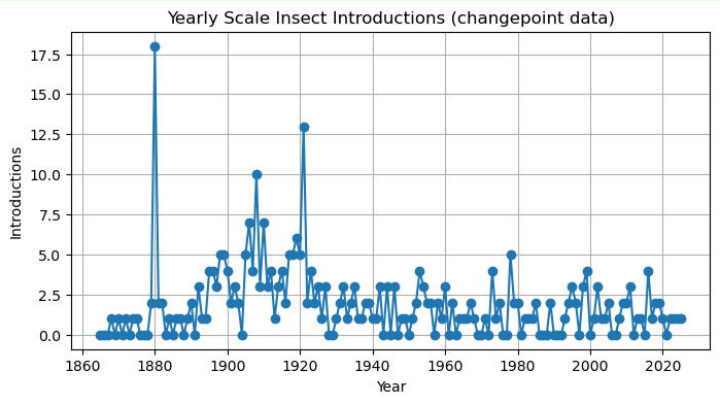
- **changepoint.csv** – yearly introduction counts
- **origin_anova.csv** – species origin and introduction year
- **region_intros.csv** – first region where each species was detected
- **trade_intro_data.csv** – annual trade quantity
- **2005_hosts.csv** – host plant categories present in 2005
- **total_hosts.csv** – all host plant records across species



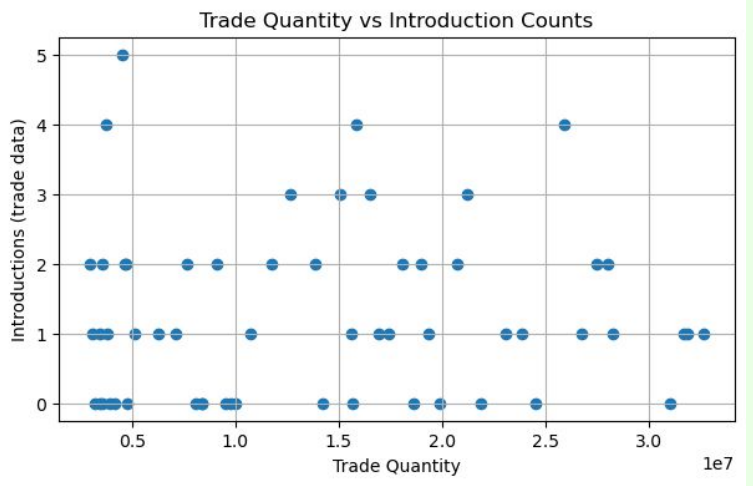
These datasets cover species-level and year-level information, allowing us to connect trade, origin, host plants, and geographic spread to patterns of invasive introductions.

EDA (Exploratory Data Analysis)

After Data Cleaning and Merging, the following charts show that introduction counts rise over time with some distinct peaks.

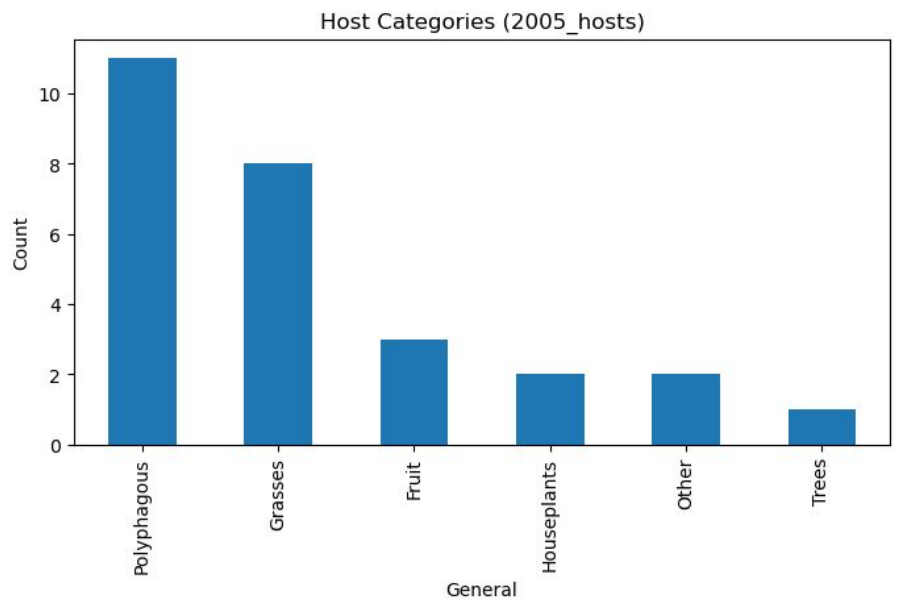


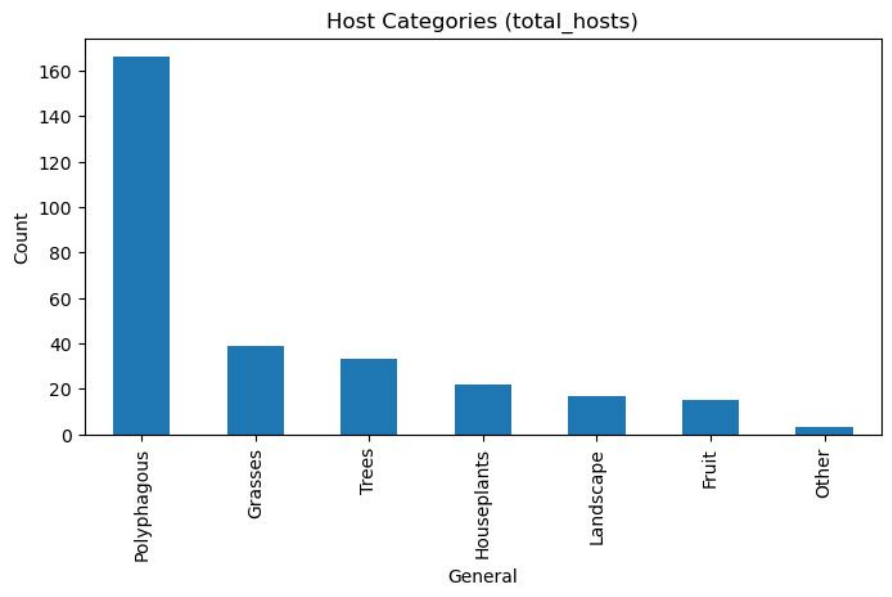
- Higher trade volume aligns with more introductions.
- Vulnerable hosts include fruits, grasses, and multi-host species.



EDA reveals that both ecological and economic factors shape introduction patterns.

EDA (continued)





SQL Analysis

Using the engineered dataset stored in **insect_risk.db**, SQL queries computed decade-level averages for introduction counts and trade quantity.

Takeaways:

- Both introductions and trade steadily increase across decades.
- Modern decades consistently show the highest ecological risk.
- SQL confirms the same trends observed in EDA and modeling.

SQL serves as a secondary verification tool for long-term patterns.

```
import sqlite3

# creating a SQLite database file to store our table
conn = sqlite3.connect(r"C:\Users\aligo\OneDrive\Desktop\210 Final Project\database\insect_risk.db")

# saving the engineered features_df table into the database as "year_data"
features_df.to_sql("year_data", conn, if_exists="replace", index=False)

# SQL query: compute decade averages for introductions and trade
sql_query = """
SELECT
    decade,
    AVG(num_intro_changepoint) AS avg_intros,
    AVG(trade_quantity)        AS avg_trade,
    COUNT(*)                   AS n_years
FROM year_data
GROUP BY decade
ORDER BY decade;
"""

# running the SQL query and loading the results as a DataFrame
sql_summary = pd.read_sql_query(sql_query, conn)
conn.close()

# printing summary statistics calculated through SQL
print("SQL summary by decade using from SQLite")
print(sql_summary)
```

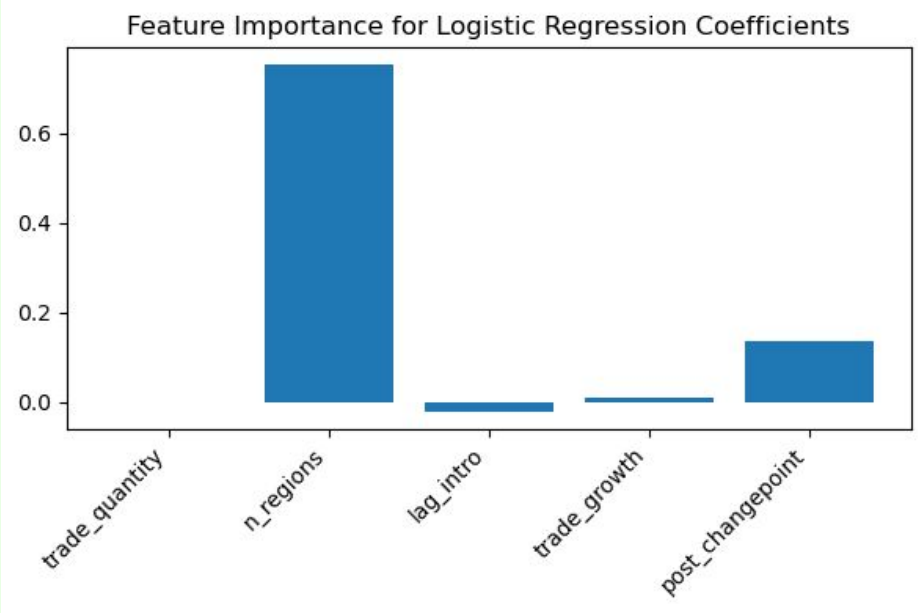
Machine Learning Models & Plots

Models Applied:

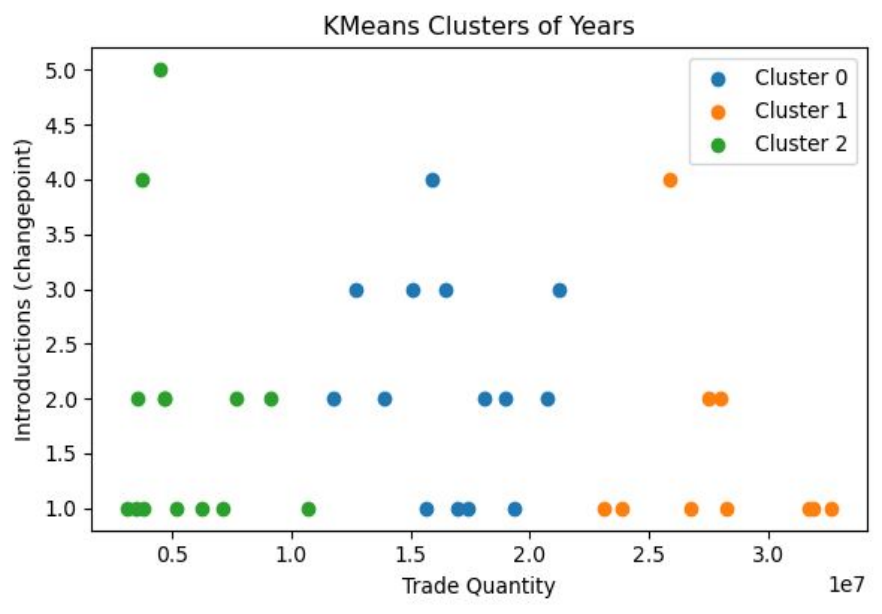
- Linear Regression for predicting introduction counts
- Logistic Regression for classifying high vs. low risk years
- K-Means Clustering for identifying ecological and trade-based groups
- K-Nearest Neighbors for finding similar historical years

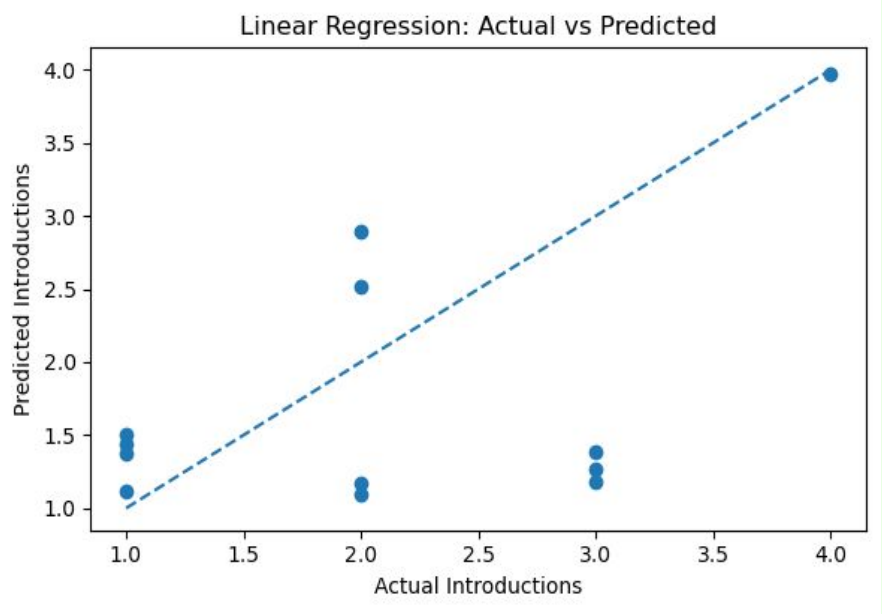
Key takeaways:

- Linear regression predictions are usually within one species of actual values.
- Logistic regression shows **n_regions** and **trade_growth** are strong predictors of high-risk years.
- K-Means separates years into low, moderate, and high-risk ecological groups.
- KNN shows recent years are most similar to one another due to modern trade patterns.



Machine Learning Models & Plots





Conclusion

Trade activity, geographic spread, and ecological momentum work together to shape yearly invasion risk. High-risk years often occur when trade is high, multiple regions detect new species, and the previous year had elevated introductions.

Thank you!