ECONOMIC DATA AND INVASIVE SPECIES

Nicolas Saenz

DSC 530

INVASIVE SPECIES HAVE A NEGATIVE IMPACT

- They negatively impact the natural biodiversity of an area.
- May cause serious damage to agriculture.
- They may cause damage to other income producing industries and property.

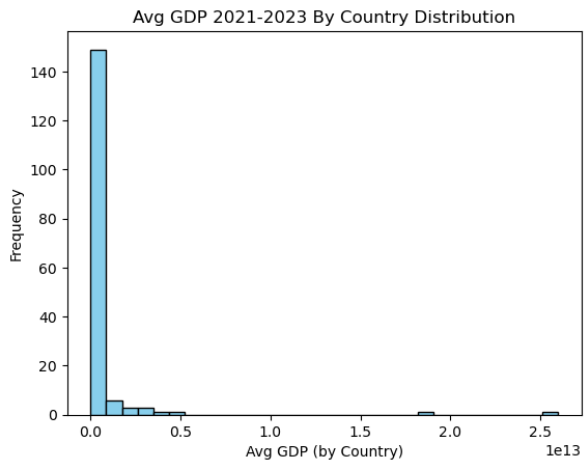
CAN WE USE ECONOMIC DATA TO BUILD A MODEL FOR INVASIVE SPECIES LIKELIHOOD?

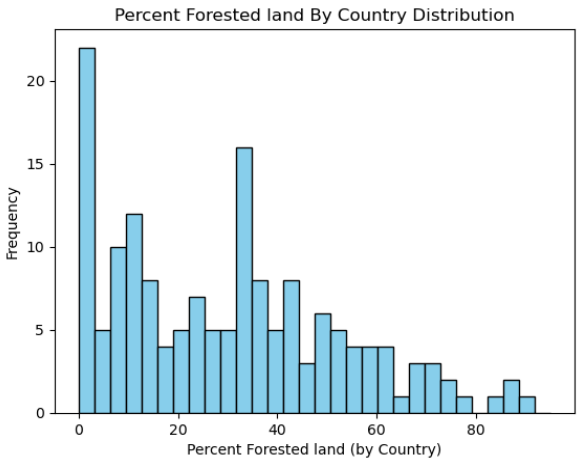
First, we can explore the data comparing several economic variables to the total number of invasive species within each country.

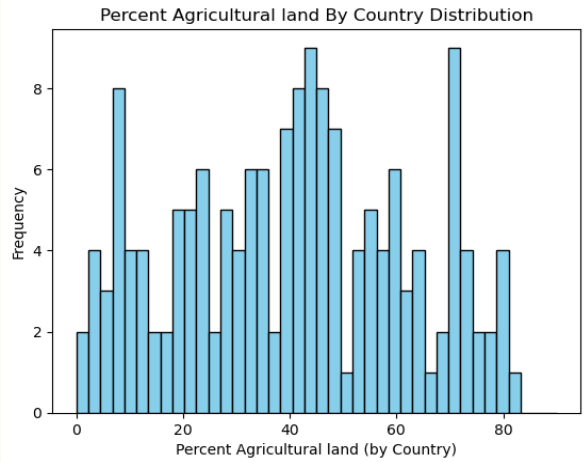
OUR VARIABLES BY COUNTRY

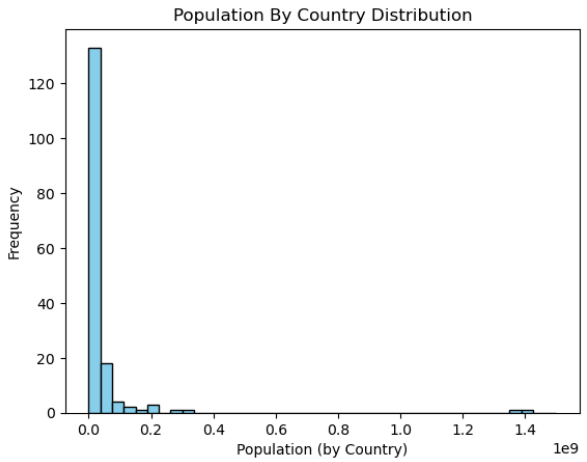
- Percent of total country which is Agricultural Land
- Percent of total country which is Forested land
- Avg GDP 2021-2023 (To get a good estimate while avoiding 2020 economic downturn.)
- Population
- The Value of Total Imports in 2023
- Total Invasive species count (data recorded 2021)
- Whether or not a country is landlocked has no borders on the ocean. (This is used later as a separating variable in EDA)

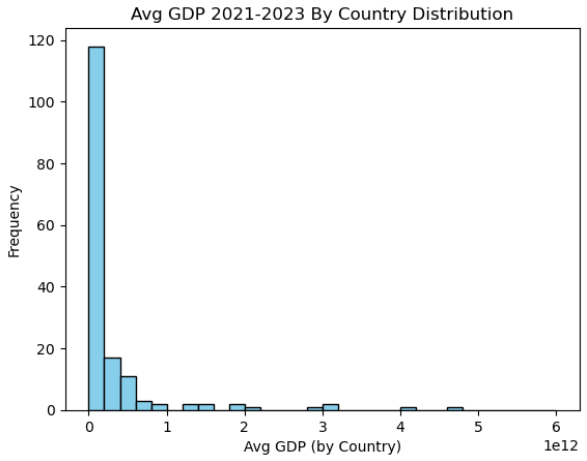
EXPLORATORY DATA ANALYSIS

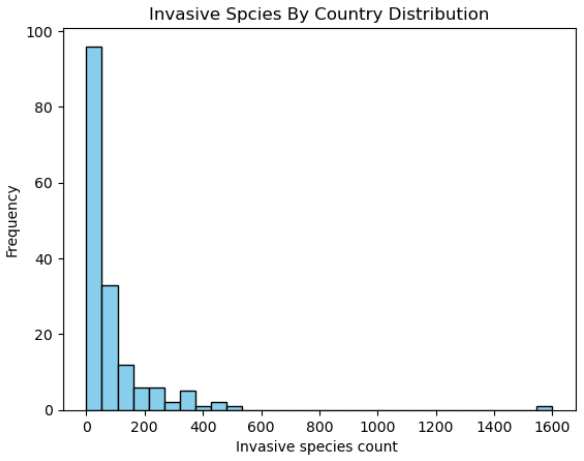


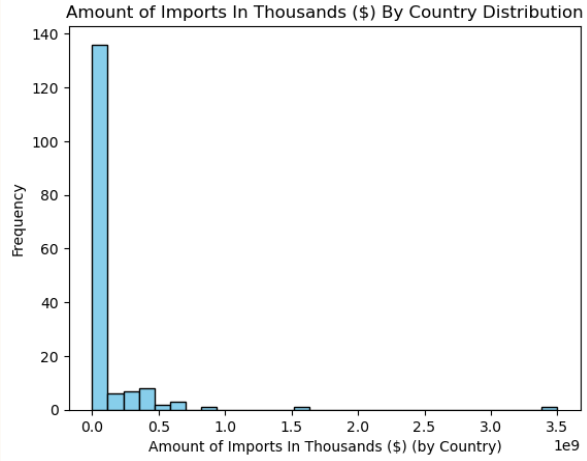


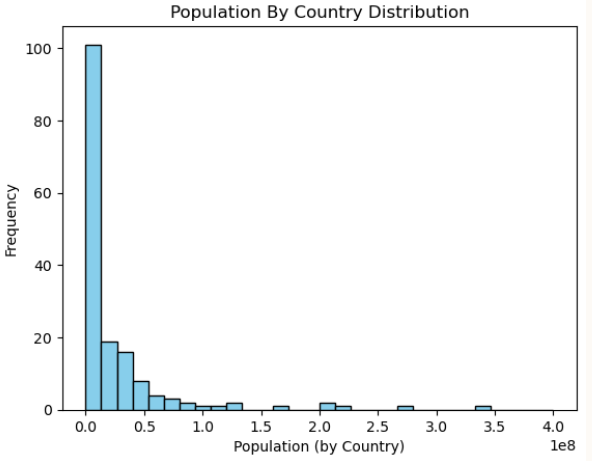










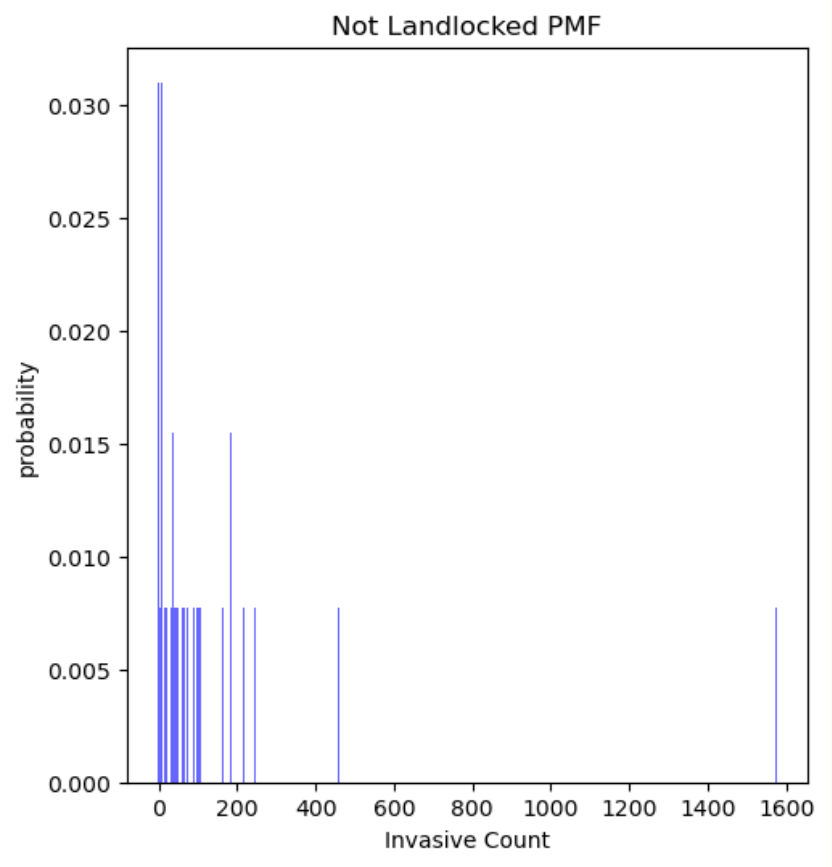


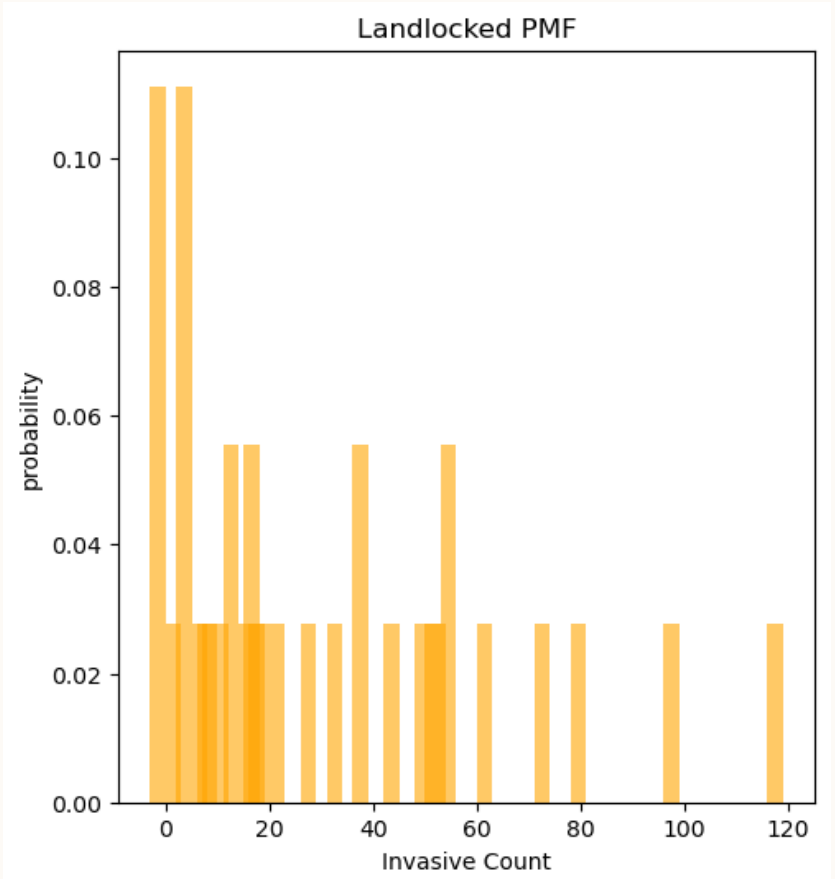
Avg GDP top all, bottom adjusted for outliers and resolution.

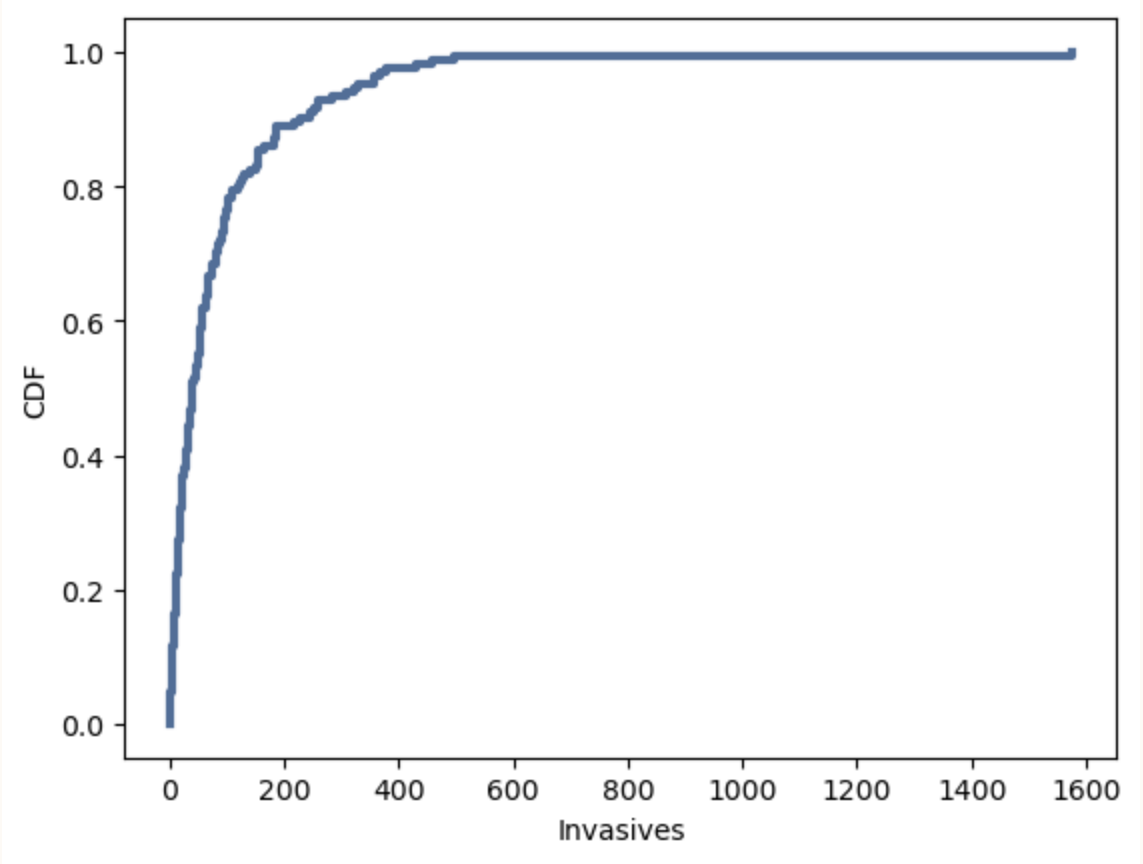
We can see invasive count, value of imports, population, and avg gdp are all highly left skewed.

Population top all, bottom adjusted for outliers and resolution.

We separate the data using whether or not a country is landlocked, and build two separate PMF's.



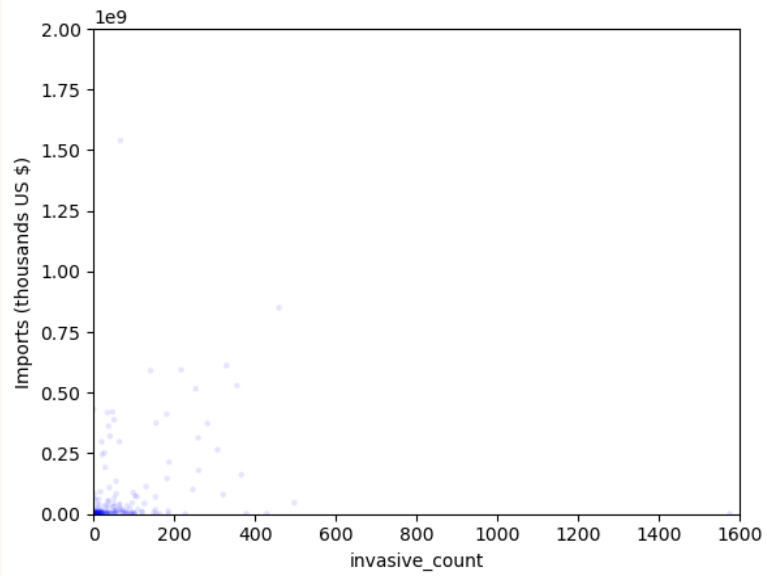


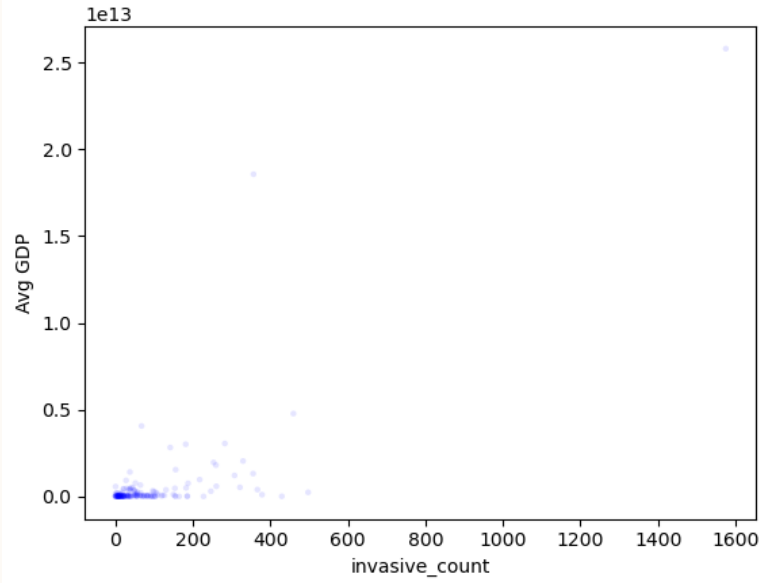


CDF OF ALL COUNTRIES PLOTTED FOR INVASIVES

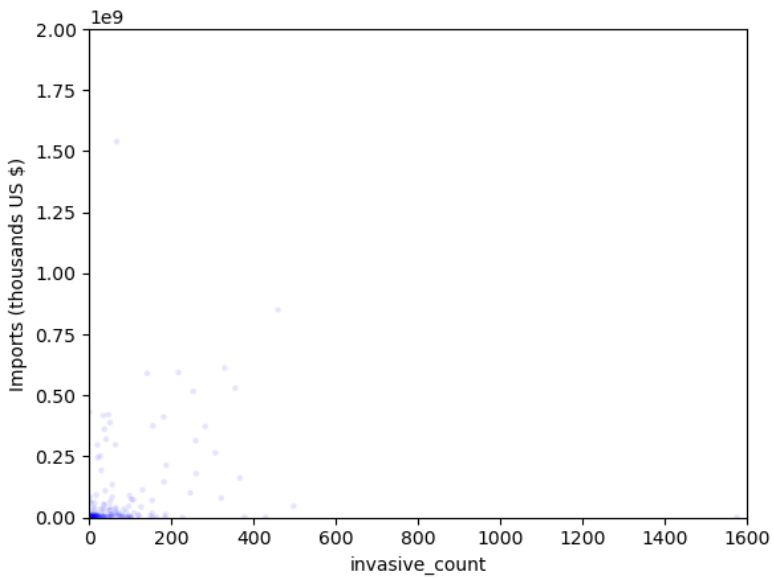
Notice the small uptick at the end of the function. This is the United states at 1575 invasive species.

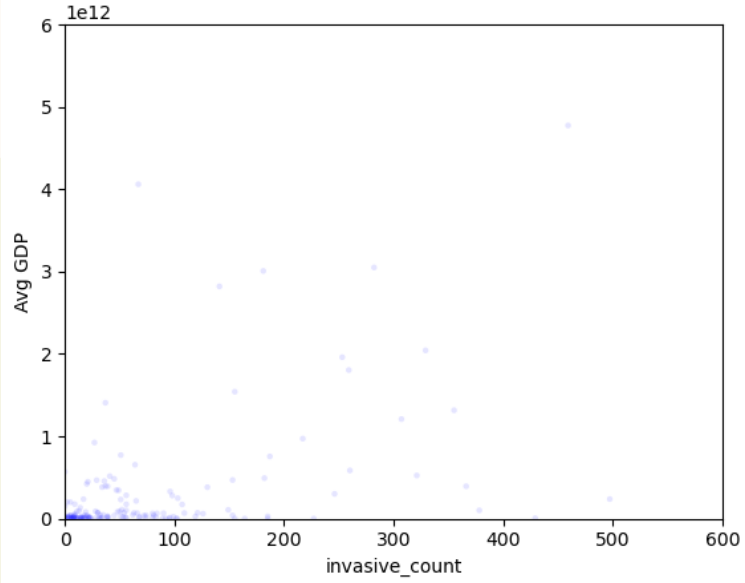
COMPARING VARIABLES TO INVASIVES





Scatterplots are highly skewed. The bottom graphs are the unadjusted versions. Top are adjusted to remove outliers and slightly improve resolution of data.





HYPOTHESIS

Whatever the human caused variables may be. My hypothesis is that landlocked countries will have a lower mean number of invasive species than non-landlocked countries.

Since the distribution of invasive species is nowhere close to normal. I used a Wilcox test to account for this. And ten repeated samples of non-landlocked countries to landlocked countries.

The hypothesis tests below followed the form shown (Statistic, p-value)

```
[(167.5, 0.008164219470927492), (120.0, 0.0005099171248730272), (110.5, 0.00025036028819158673), (146.5, 0.005780522797770957
5), (134.0, 0.0030295430006892702), (111.5, 0.000269419135292992), (177.0, 0.013287985610077158), (169.5, 0.02862055584734788
2), (142.0, 0.0020792766590602696), (201.5, 0.03775702390703373)]
```

Out of 10 wilcox tests comparing means. All 10 showed a large static with nine out of 10 p-values being significant with a cut-off of 0.05.

VARIABLE CORRELATION

```
Agricultural.Land....   -0.000049
Forested.Area....        0.139193
X2023                    0.755195
Population               0.329767
Import..US..Thousand.    0.230383
Avg_21_23                0.757606
Name: invasive_count, dtype: float64
```

The two strongest correlation in the available variables are Avg GDP and Population. So these are the variables we move forward to build our model.

MODEL

OLS Regression Results

Dep. Variable:	invasive_count	R-squared:	0.594
Model:	OLS	Adj. R-squared:	0.589
Method:	Least Squares	F-statistic:	118.5
Date:	Sun, 02 Mar 2025	Prob (F-statistic):	1.95e-32
Time:	11:04:00	Log-Likelihood:	-989.23
No. Observations:	165	AIC:	1984.
Df Residuals:	162	BIC:	1994.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	65.0056	7.929	8.198	0.000	49.348	80.664
Avg_21_23	5.196e-11	3.73e-12	13.915	0.000	4.46e-11	5.93e-11
Population	-1.708e-07	6.04e-08	-2.827	0.005	-2.9e-07	-5.15e-08

Omnibus:	58.107	Durbin-Watson:	2.122
Prob(Omnibus):	0.000	Jarque-Bera (JB):	263.211
Skew:	1.226	Prob(JB):	6.99e-58
Kurtosis:	8.681	Cond. No.	2.69e+12

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.69e+12. This might indicate that there are strong multicollinearity or other numerical problems.

Model results. Predicting based on population and avg GDP.

Adjusted R-squared 0.589

FUTURE STEPS

The goal was to find the best economic predictors of high invasive species count.

Future steps I would like to include other economic variables that might have a better predictive power over the data observed. Some of these may include how large the tourism industry is or how large the exotic pet market is within a country as other possible means for predicting more invasives within a country.

THANK YOU