D214 – Data Analytics Graduate Capstone

WGU M.S. Data Analytics

Lyssa Kline

April 20, 2025

A, Research Question

The research question in scope for this Capstone project is “Can tariff rate changes and trade restrictions imposed in 2025 be used to forecast short-term changes in U.S. GDP and inflation?”

This question stems from the broader economic context of the 2024–2025 U.S. election cycle, during which trade policy has resurfaced as a pivotal issue. Proposed and implemented tariff measures have reignited debates over protectionism, global trade dynamics, and their downstream effects on domestic economic indicators.

Trade tariffs have historically influenced economic performance, particularly their impact on import prices, consumer costs, and trade balances. Given the reemergence of aggressive tariff policies during the 2024 campaign and into 2025, examining whether such policy decisions can be statistically linked to measurable changes in macroeconomic indicators is timely and relevant. This analysis could provide valuable foresight for both policymakers crafting future trade strategies and businesses navigating economic uncertainty.

The dataset compiled from Kaggle integrates information on tariff announcements, effective tariff rates, affected trade volumes, and corresponding economic outcomes such as GDP and consumer price inflation. It includes both imposed and threatened tariffs, their dates, the goods and countries targeted, and their modeled or observed economic impacts. The data allows for tracking these tariff events alongside short-term changes in GDP and inflation, enabling a data-driven exploration of cause-and-effect relationships.

The working hypothesis is that higher tariff rates and broader trade restrictions are associated with short-term negative impacts on U.S. GDP and increases in inflation. Specifically, tariffs that increase the cost of imported goods may contribute to upward pressure on consumer prices, while trade disruptions may reduce economic output, especially in globally integrated sectors. The project aims to test whether such relationships are statistically significant and if they can be used for short-term forecasting of macroeconomic trends.

B, Data Collection

For this project, the dataset was obtained in spreadsheet format from Kaggle, a well-known public data repository. By searching for keywords related to "tariffs," a relevant dataset was identified that compiles key data points surrounding U.S. tariff policies proposed or implemented in 2024 and 2025. The dataset includes the dates of tariff announcements, countries and goods affected, effective tariff rates, the estimated impact on trade volume, and key economic indicators such as GDP and consumer price inflation (CPI). This structure enabled the alignment of macroeconomic data with specific tariff-related events to support comparative and predictive analyses.

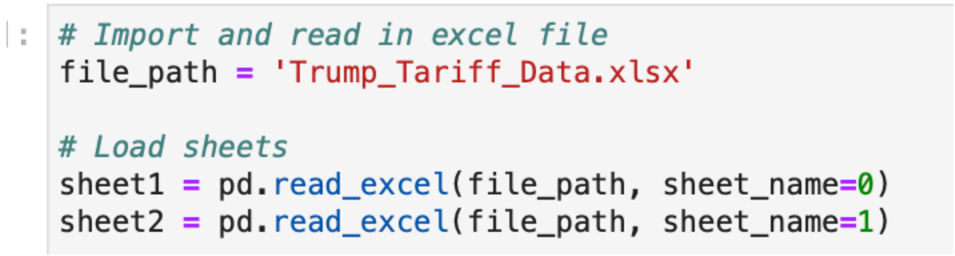
One significant advantage of using Kaggle as the data source is accessibility. The platform provides curated datasets that are already partially cleaned, organized, and often contain relevant documentation or community discussion. This saved time and allowed me to quickly identify a dataset that met the needs of my research without having to scrape or compile raw data from multiple government or international sources.

A disadvantage of using precompiled datasets from public sources like Kaggle is the lack of full transparency into the original data collection methods. The data may have inconsistencies, incomplete documentation, or assumptions made by the original compilers that are not always disclosed. This introduces potential bias or gaps in data reliability.

During the data review process, I reviewed the data for any issues such as vague country references, unspecified tariff statuses, and missing economic metrics in some entries. There were no issues identified in the data that needed to be manually validated and filtered. This review ensured that the dataset was robust enough to support accurate event-based analysis and predictive modeling.

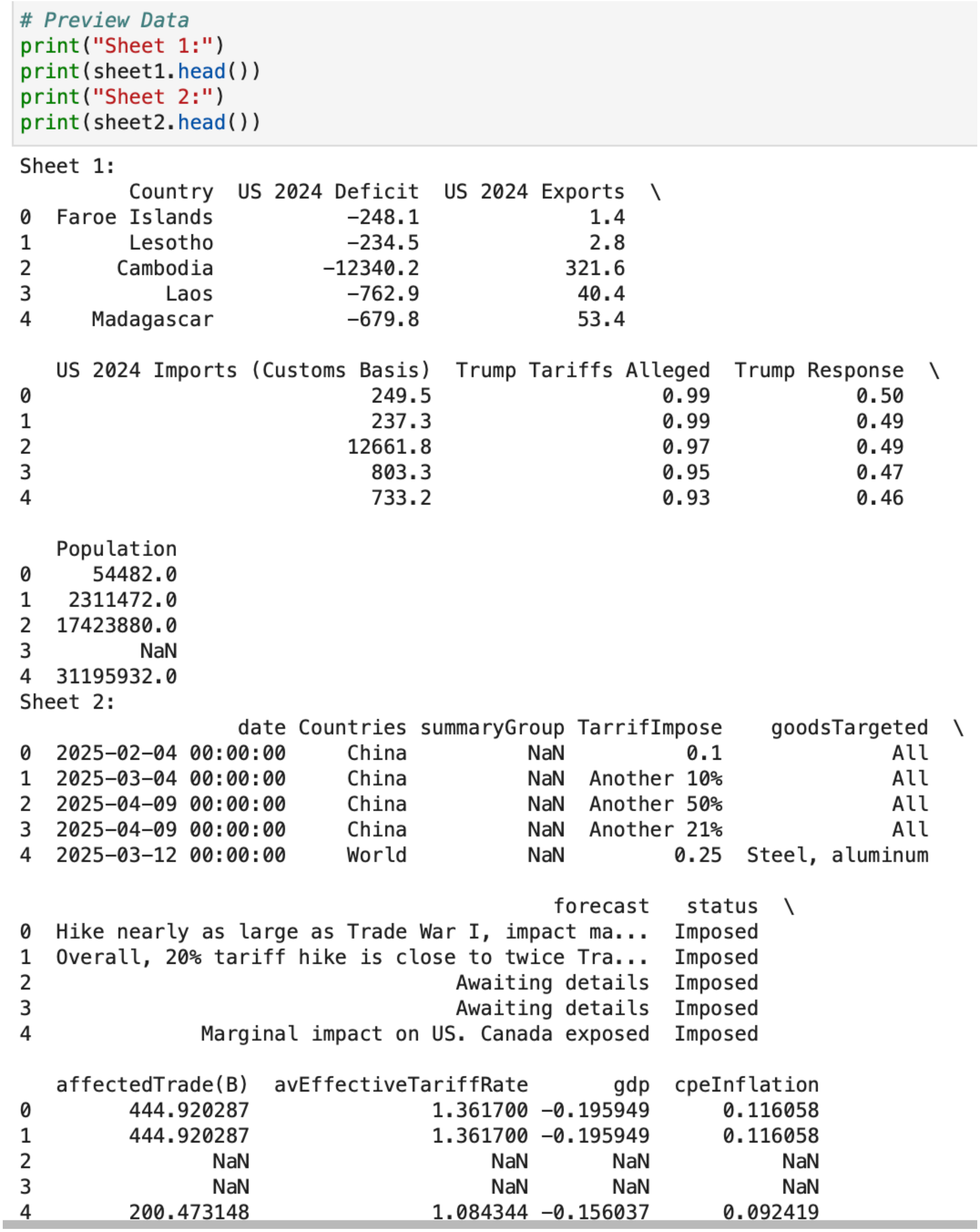
C, Data Extraction and Preparation

The dataset used in this project was extracted from Kaggle, a public data repository. The dataset was downloaded in Excel format and included two tabs of publicly available tariff and GDP data. Sheet 1 contains per-country trade statistics and tariffs. Sheet 2 includes tariff event dates, affected trade, imposed rates, and macroeconomic results (GDP & CPI). The model will be built from sheet 2 since it contains the economic targets (GDP, CPEInflation) and relevant predictors (TariffImpose, EffectiveTariffRate, affectedTrade(B), etc). Using pandas, both sheets were extracted from Excel into the model using the read_excel() function.



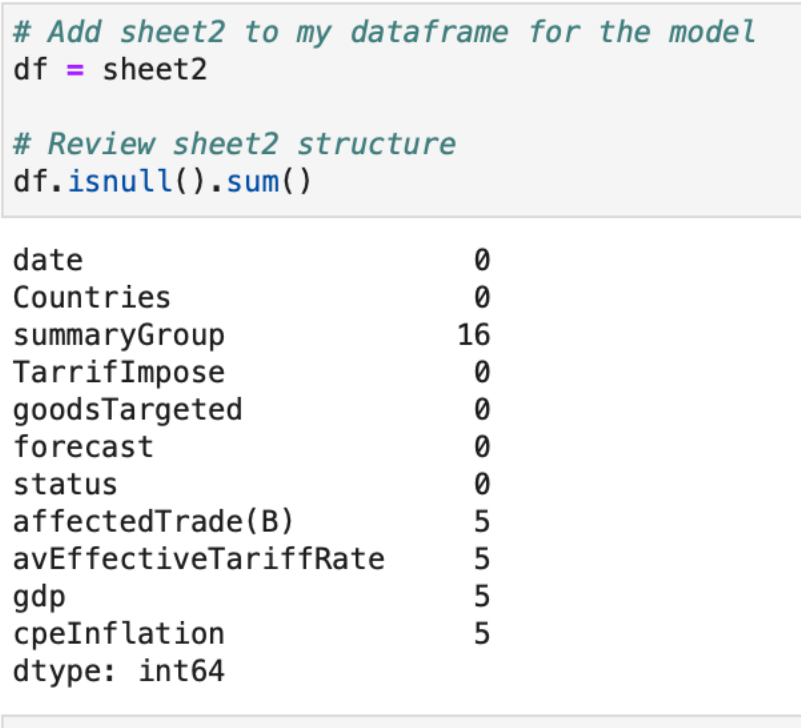
```
# Import and read in excel file
file_path = 'Trump_Tariff_Data.xlsx'

# Load sheets
sheet1 = pd.read_excel(file_path, sheet_name=0)
sheet2 = pd.read_excel(file_path, sheet_name=1)
```

Once extracted, the data was first viewed by using a few functions to observe the data and see where the discrepancies were. Then prepared and cleaned using standardization to convert date and numeric columns, and filtering on the relevant columns/rows. See code below outlining the overview of the data, identifying any anomalies, and the process to clean the data.



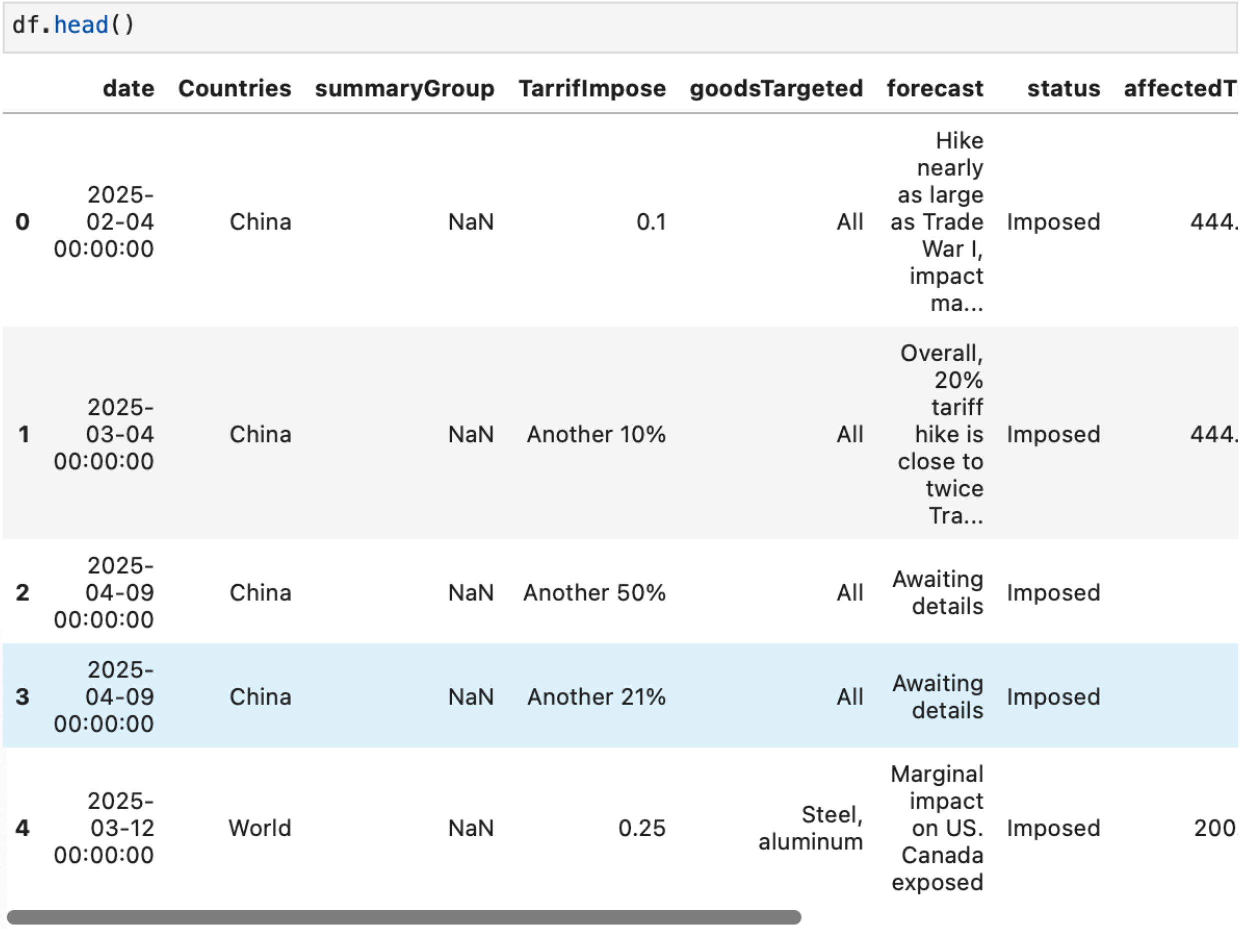
```python
# Preview Data
print("Sheet 1:")
print(sheet1.head())
print("Sheet 2:")
print(sheet2.head())
```

```
Sheet 1:
         Country  US 2024 Deficit  US 2024 Exports  \
0  Faroe Islands           -248.1              1.4   
1        Lesotho           -234.5              2.8   
2       Cambodia         -12340.2            321.6   
3           Laos           -762.9             40.4   
4     Madagascar           -679.8             53.4   

   US 2024 Imports (Customs Basis)  Trump Tariffs Alleged  Trump Response  \
0                            249.5                   0.99            0.50   
1                            237.3                   0.99            0.49   
2                          12661.8                   0.97            0.49   
3                            803.3                   0.95            0.47   
4                            733.2                   0.93            0.46   

   Population  
0     54482.0  
1   2311472.0  
2  17423880.0  
3         NaN  
4  31195932.0  
Sheet 2:
                  date Countries summaryGroup TarrifImpose    goodsTargeted  \
0  2025-02-04 00:00:00     China          NaN          0.1              All   
1  2025-03-04 00:00:00     China          NaN  Another 10%              All   
2  2025-04-09 00:00:00     China          NaN  Another 50%              All   
3  2025-04-09 00:00:00     China          NaN  Another 21%              All   
4  2025-03-12 00:00:00     World          NaN         0.25  Steel, aluminum   

                                            forecast   status  \
0  Hike nearly as large as Trade War I, impact ma...  Imposed   
1  Overall, 20% tariff hike is close to twice Tra...  Imposed   
2                                   Awaiting details  Imposed   
3                                   Awaiting details  Imposed   
4              Marginal impact on US. Canada exposed  Imposed   

   affectedTrade(B)  avEffectiveTariffRate       gdp  cpeInflation  
0        444.920287               1.361700 -0.195949      0.116058  
1        444.920287               1.361700 -0.195949      0.116058  
2               NaN                    NaN       NaN           NaN  
3               NaN                    NaN       NaN           NaN  
4        200.473148               1.084344 -0.156037      0.092419  
```



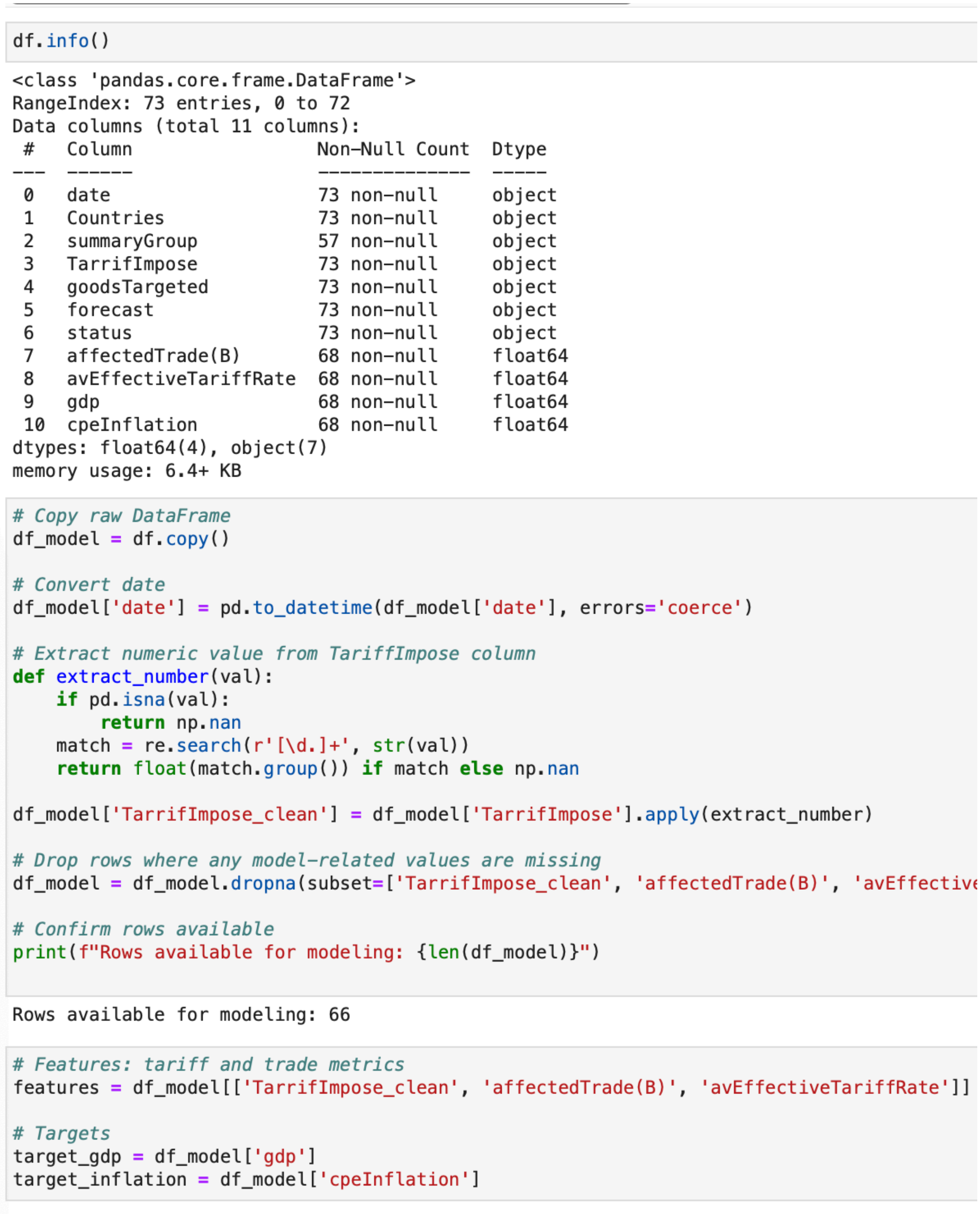
```
# Add sheet2 to my dataframe for the model
df = sheet2

# Review sheet2 structure
df.isnull().sum()
```

```
date                   0
Countries              0
summaryGroup          16
TarrifImpose           0
goodsTargeted          0
forecast               0
status                 0
affectedTrade(B)       5
avEffectiveTariffRate  5
gdp                    5
cpeInflation           5
dtype: int64
```



```
df.head()
```

	date	Countries	summaryGroup	TarrifImpose	goodsTargeted	forecast	status	affectedT
0	2025-02-04 00:00:00	China	NaN	0.1	All	Hike nearly as large as Trade War I, impact ma...	Imposed	444.
1	2025-03-04 00:00:00	China	NaN	Another 10%	All	Overall, 20% tariff hike is close to twice Tra...	Imposed	444.
2	2025-04-09 00:00:00	China	NaN	Another 50%	All	Awaiting details	Imposed	
3	2025-04-09 00:00:00	China	NaN	Another 21%	All	Awaiting details	Imposed	
4	2025-03-12 00:00:00	World	NaN	0.25	Steel, aluminum	Marginal impact on US. Canada exposed	Imposed	200



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 73 entries, 0 to 72
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   date                  73 non-null     object
 1   Countries             73 non-null     object
 2   summaryGroup          57 non-null     object
 3   TarrifImpose          73 non-null     object
 4   goodsTargeted         73 non-null     object
 5   forecast              73 non-null     object
 6   status                73 non-null     object
 7   affectedTrade(B)      68 non-null     float64
 8   avEffectiveTariffRate 68 non-null     float64
 9   gdp                   68 non-null     float64
 10  cpeInflation          68 non-null     float64
dtypes: float64(4), object(7)
memory usage: 6.4+ KB
```

```
# Copy raw DataFrame
df_model = df.copy()

# Convert date
df_model['date'] = pd.to_datetime(df_model['date'], errors='coerce')

# Extract numeric value from TariffImpose column
def extract_number(val):
    if pd.isna(val):
        return np.nan
    match = re.search(r'[\d.]+', str(val))
    return float(match.group()) if match else np.nan

df_model['TarrifImpose_clean'] = df_model['TarrifImpose'].apply(extract_number)

# Drop rows where any model-related values are missing
df_model = df_model.dropna(subset=['TarrifImpose_clean', 'affectedTrade(B)', 'avEffective

# Confirm rows available
print(f"Rows available for modeling: {len(df_model)}")
```

```
Rows available for modeling: 66
```

```
# Features: tariff and trade metrics
features = df_model[['TarrifImpose_clean', 'affectedTrade(B)', 'avEffectiveTariffRate']]

# Targets
target_gdp = df_model['gdp']
target_inflation = df_model['cpeInflation']
```

This data extraction and model-building process was all performed in Python using Jupyter as the driver to allow us to write and build the model. Python is a powerful and flexible tool for automating, cleaning, and analyzing data. Pandas specifically allow easy manipulation of complex

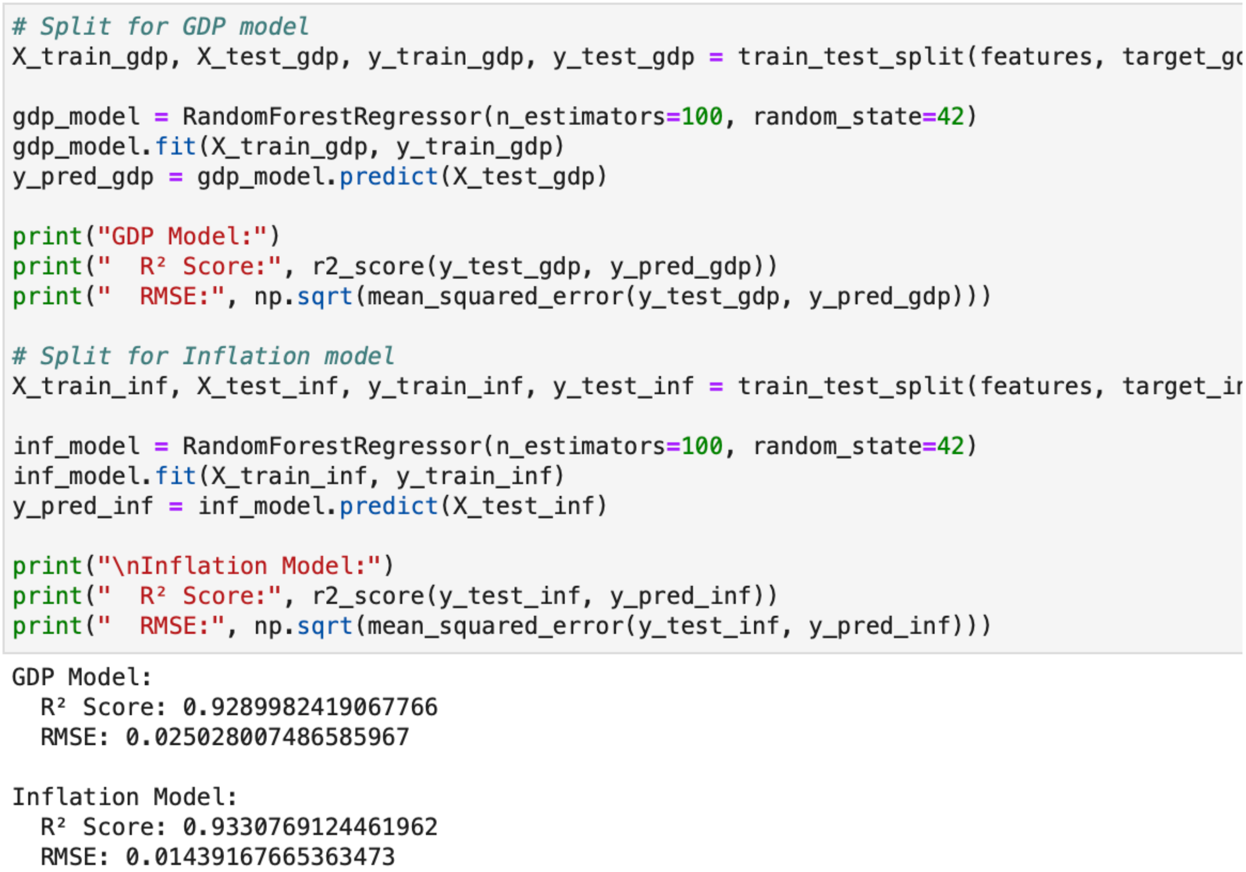
Excel files, especially those with multiple sheets and mixed formats. It also integrates seamlessly into the machine learning pipeline.

One advantage of pandas is that it allows you to clean and prepare large datasets quickly with reusable, reproducible code. This is especially useful when working with multi-tab Excel files or inconsistent formatting (like text + numbers in one column). Whereas one disadvantage of these tools is that they require code proficiency, unlike point-and-click tools like Excel, Python requires programming knowledge. If the data structure is especially messy, debugging scripts to extract numeric values or handle string variations can be time-consuming.

D, Analysis

To analyze the impact of 2025 tariff changes on short-term economic outcomes, a supervised machine learning technique was used to forecast changes in U.S. GDP and U.S. CPI inflation. This technique was applied using the RandomForestRegressor model from the scikit-learn Python library.

First, the dataset was split into an 80% training set and a 20% test set using train_test_split(). This ensured that model performance could be evaluated using data the model had not seen during training. Two separate models were trained, one to predict GDP change and the second to predict the CPI inflation rate. The models were evaluated using an R-squared score to measure how much variance is explained by the model, and a root mean square error (RMSE) to measure average prediction error. See the output of this train-test split and calculations being performed below. Additional details on the output of the prediction model are described under section E below.

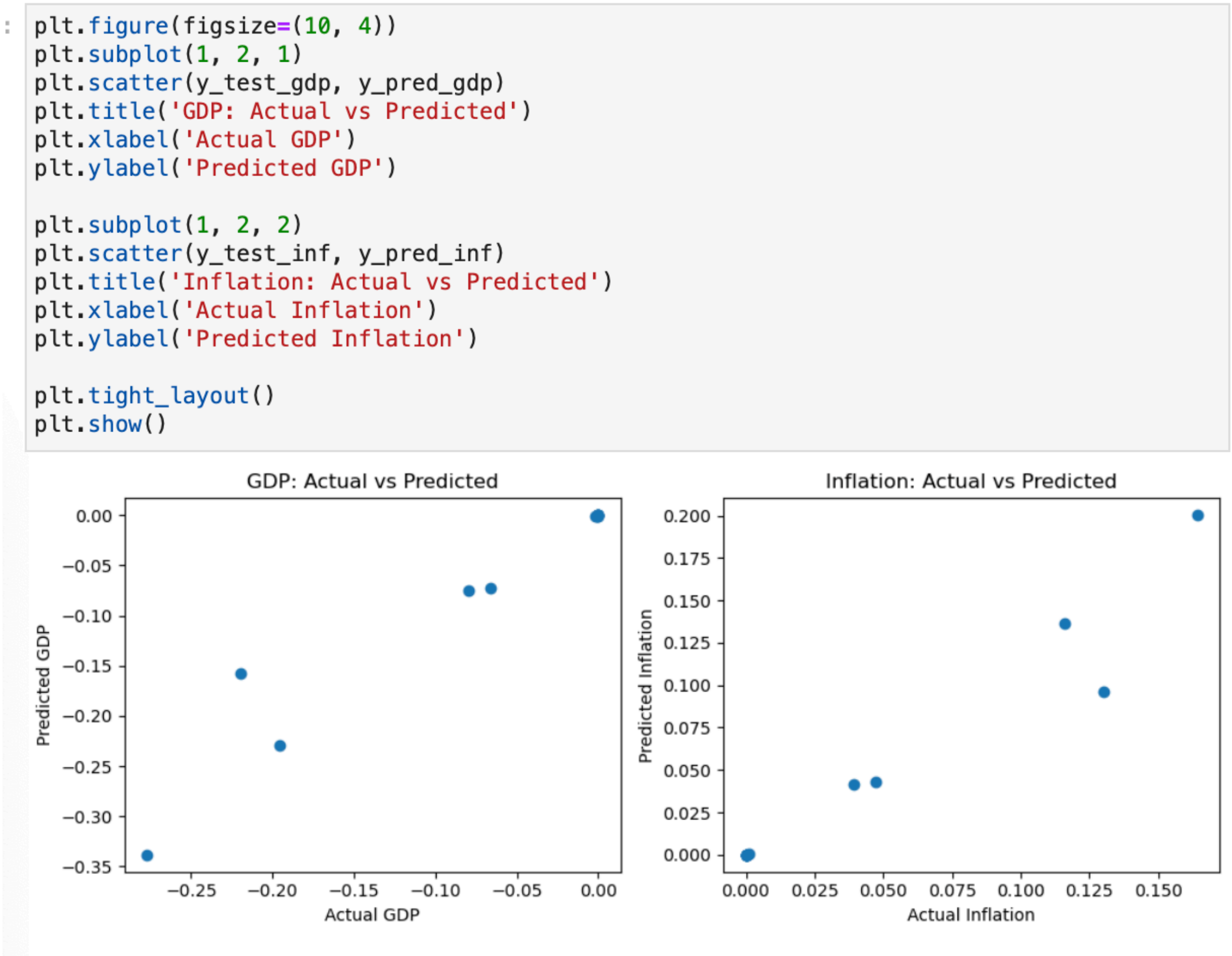


```
# Split for GDP model
X_train_gdp, X_test_gdp, y_train_gdp, y_test_gdp = train_test_split(features, target_gd

gdp_model = RandomForestRegressor(n_estimators=100, random_state=42)
gdp_model.fit(X_train_gdp, y_train_gdp)
y_pred_gdp = gdp_model.predict(X_test_gdp)

print("GDP Model:")
print("  R² Score:", r2_score(y_test_gdp, y_pred_gdp))
print("  RMSE:", np.sqrt(mean_squared_error(y_test_gdp, y_pred_gdp)))

# Split for Inflation model
X_train_inf, X_test_inf, y_train_inf, y_test_inf = train_test_split(features, target_in

inf_model = RandomForestRegressor(n_estimators=100, random_state=42)
inf_model.fit(X_train_inf, y_train_inf)
y_pred_inf = inf_model.predict(X_test_inf)

print("\nInflation Model:")
print("  R² Score:", r2_score(y_test_inf, y_pred_inf))
print("  RMSE:", np.sqrt(mean_squared_error(y_test_inf, y_pred_inf)))
```

```
GDP Model:
  R² Score: 0.9289982419067766
  RMSE: 0.025028007486585967

Inflation Model:
  R² Score: 0.9330769124461962
  RMSE: 0.01439167665363473
```

Random Forest Regression was selected due to how it handles nonlinear relationships well, specifically for the ease of handling the nonlinear relationship between tariff metrics and macroeconomic outcomes. Additionally, random forest is robust to outliers and multicollinearity, which are common in real-world policy data. Lastly, it provides a strong performance even with a moderate dataset size, which is this one with only 66 rows available for modeling.

The plots below compare the predicted values to the actual GDP and inflation values from the test set. The alignment of points along the diagonal trend suggests a high level of accuracy for both models.



```
plt.figure(figsize=(10, 4))
plt.subplot(1, 2, 1)
plt.scatter(y_test_gdp, y_pred_gdp)
plt.title('GDP: Actual vs Predicted')
plt.xlabel('Actual GDP')
plt.ylabel('Predicted GDP')

plt.subplot(1, 2, 2)
plt.scatter(y_test_inf, y_pred_inf)
plt.title('Inflation: Actual vs Predicted')
plt.xlabel('Actual Inflation')
plt.ylabel('Predicted Inflation')

plt.tight_layout()
plt.show()
```

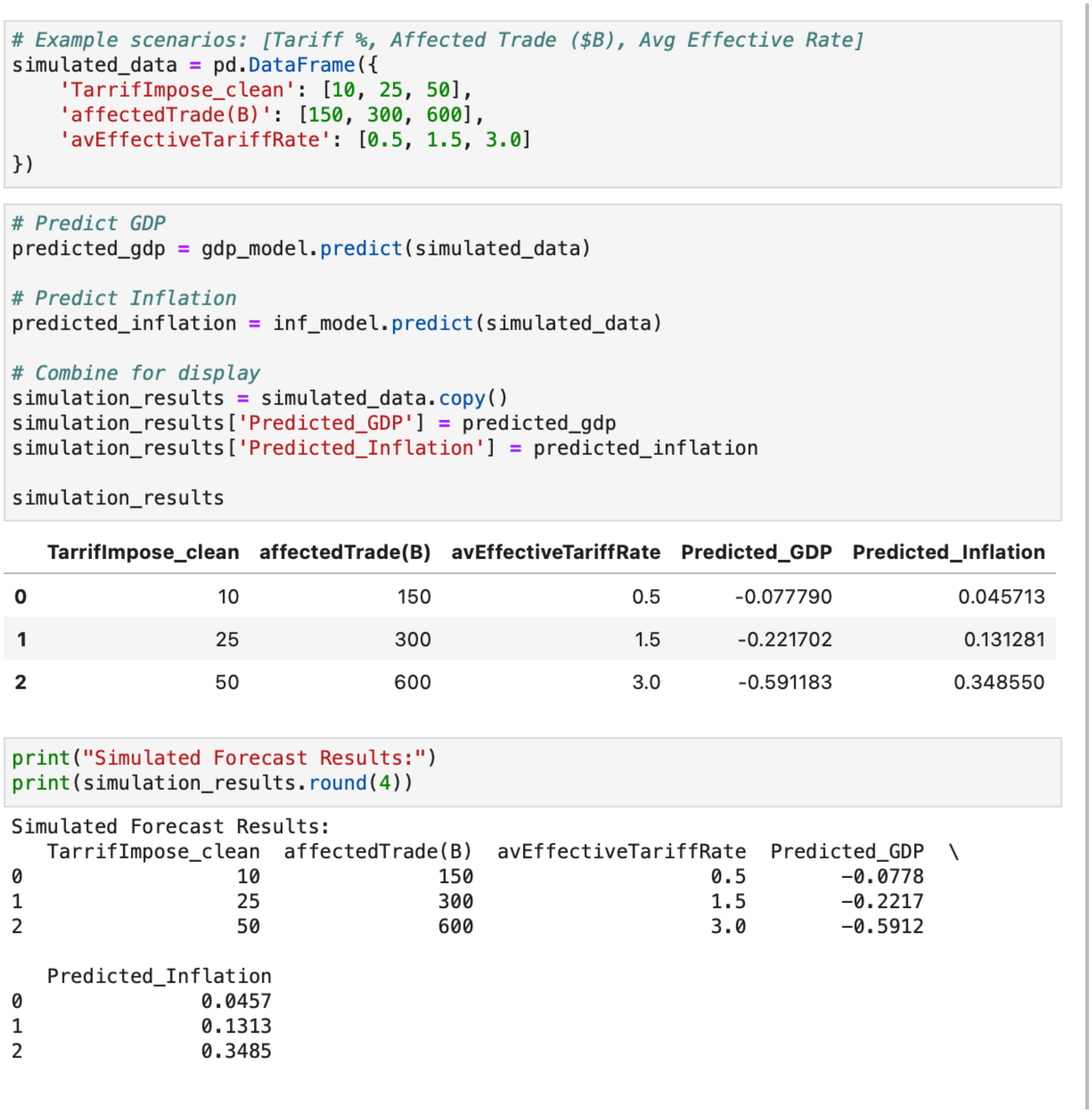
One advantage of this analysis technique is that it has high accuracy with flexible modeling. Random forests can capture complex interactions between variables like tariff size, trade impact, and policy effects, leading to accurate predictions and high R-squared scores in this project. One disadvantage of this analysis technique is that it has a lack of interpretability. While Random Forests are powerful, it is difficult to interpret exactly how much each feature (e.g., effective tariff rate) contributes to a specific prediction without additional tools like feature importance or SHAP values.

E, Data Summary and Implications

The research question asked, “Can tariff changes and trade restrictions imposed in 2025 be used to forecast short-term changes in U.S. GDP and inflation?” The results of this analysis support this hypothesis. Using random forest regression models, the project demonstrated that tariff rate changes the volume of affected trade, and the effective tariff rate can accurately forecast short-term economic changes in both GDP and inflation. The GDP model achieved an R-squared score of .93 and the inflation model scored a .93, indicating that over 90% of the variance in economic outcomes could be explained by these tariff-related inputs. Simulated scenarios further confirmed that higher tariff rates and larger trade disruptions result in greater GDP declines and higher inflation, aligning with macroeconomic theory.

To assess the potential economic impact of various tariff strategies, I simulated three hypothetical policy scenarios by adjusting tariff rates, the volume of affected trade, and the average effective tariff rate. These simulated inputs were then passed through the trained Random Forest models to forecast corresponding changes in U.S. GDP and inflation. Higher tariffs and broader trade impacts are associated with sharper GDP declines. For instance, a 50% tariff affecting $600B in trade could reduce GDP by nearly 0.6%. Inflation scales with trade restrictions, rising from 0.05% to nearly 0.35% across scenarios, suggesting upward pressure on consumer prices. These results support the hypothesis that tariff policies can be strong predictors of short-term economic shifts, reinforcing their importance in macroeconomic planning.

The following table illustrates the model's predictions for three hypothetical 2025 policy scenarios, varying by tariff percentage, trade volume affected, and effective tariff rate. These outputs demonstrate the model's practical application, estimating economic impact before policies are implemented.



```python
# Example scenarios: [Tariff %, Affected Trade ($B), Avg Effective Rate]
simulated_data = pd.DataFrame({
    'TarrifImpose_clean': [10, 25, 50],
    'affectedTrade(B)': [150, 300, 600],
    'avEffectiveTariffRate': [0.5, 1.5, 3.0]
})
```

```python
# Predict GDP
predicted_gdp = gdp_model.predict(simulated_data)

# Predict Inflation
predicted_inflation = inf_model.predict(simulated_data)

# Combine for display
simulation_results = simulated_data.copy()
simulation_results['Predicted_GDP'] = predicted_gdp
simulation_results['Predicted_Inflation'] = predicted_inflation

simulation_results
```

	TarrifImpose_clean	affectedTrade(B)	avEffectiveTariffRate	Predicted_GDP	Predicted_Inflation
0	10	150	0.5	-0.077790	0.045713
1	25	300	1.5	-0.221702	0.131281
2	50	600	3.0	-0.591183	0.348550

```python
print("Simulated Forecast Results:")
print(simulation_results.round(4))
```

```
Simulated Forecast Results:
   TarrifImpose_clean  affectedTrade(B)  avEffectiveTariffRate  Predicted_GDP  \
0                  10               150                    0.5        -0.0778
1                  25               300                    1.5        -0.2217
2                  50               600                    3.0        -0.5912

   Predicted_Inflation
0               0.0457
1               0.1313
2               0.3485
```

A key limitation of this analysis is the size of the dataset, which included approximately 66 complete observations after cleaning. While the models performed well, the limited sample size may affect generalizability and could lead to overfitting, particularly if extreme or rare policy scenarios are introduced. Additionally, the model assumes immediate or short-term economic impacts and does not account for long-term adjustments in trade behavior, global supply chains, or monetary policy reactions.

Given the findings, policymakers and economic advisors should consider using real-time tariff metrics as part of economic forecasting models, especially when preparing for major policy announcements. Additionally, it should be recognized that aggressive or broad-based tariffs are likely to cause significant short-term inflation spikes and GDP slowdowns. One workaround would be to use data-driven tools like these models to simulate policy effects before implementation. These insights could be valuable for anticipating risk, especially during election cycles or trade renegotiations.

Proposed directions for a future study would be to incorporate additional macroeconomic variables. Future models could include features such as interest rates, currency exchange shifts, and stock market reactions. This would give a more holistic view of the economy's short-term response to trade policies.

Additionally, this model could be expanded using a time frame or frequency analysis. Instead of focusing solely on 2025 tariff data, future research could analyze a multi-year time series of tariff events or build a real-time forecasting model using daily or monthly economic indicators. This would improve the model's robustness and allow for tracking of lagged effects from policy decisions.

F, Acknowledge the Sources

Trump Tariff Data 2025
Used to download the dataset and review its structure for the prediction analysis.
Webpage: https://www.kaggle.com/datasets/mesumraza/trump-tarrif-data?resource=download

Raza Hemani, M. (2025). *Trump Tariff Data 2025* [Data set]. Kaggle. https://www.kaggle.com/datasets/mesumraza/trump-tarrif-dataKaggle