Panel Data Models
With
Application To Macroeconometrics
(Inflation Forecasting)

Submitted By:

Mohamed Abd Al-mgyd

Supervised By:

Dr. Amal Abd-Elfatah

Table of Contents

1. Introduction And Review

Panel data (longitudinal or cross-sectional time-series data):

that combines temporal depth with cross-sectional breadth, offering a powerful framework that enhances the precision and richness of econometric inference (Hsiao, 2003).

The evolution of panel methodologies began in the mid-20th century when researchers recognized the limitations of using only cross-sectional or time-series data:

- Verbeek & Nijman (1992) and Moulton (1990) introduce fixed/random-effects and temporal-dependency models.
- Arellano & Bond (1991) develop dynamic GMM estimators to address endogeneity.

Panel data methods offer several crucial advantages over purely cross-sectional or time-series analyses:

- **Controls unobservables:** Fixed/Random Effects capture time-invariant factors (e.g., institutions, culture) (Baltagi, 2008). For example, Fischer (1993) showed how inflation behaves differently under various policy regimes (inflation inertia).
- **Improved efficiency:** Pooling over time increases sample size, reducing variance inflation (Wooldridge, 2010).
- **Captures dynamics:** Lagged variables model persistence in macro indicators (Arellano & Bond, 1991).

2. Study Objectives

Compare Panel :

Evaluate Pooled OLS, Fixed Effects, Random Effects, and Dynamic GMM models—assessing their performance via diagnostic tests (e.g., Hausman, AR(2), Sargan).

Empirical Application:

Use a balanced panel (1980–2024) of annual macroeconomic indicators for 77 countries—both developed and developing—to analyze and forecast inflation dynamics.

Recommendations:

Formulate clear recommendations for selecting the most appropriate panel data technique based on data characteristics and diagnostic outcomes.

3. Dataset Overview

The dataset includes annual data for 77 countries from 1980 to 2024, sourced from the IMF's World Economic Outlook (WEO).

- **The target variable is PCPIPCH** (Inflation, average consumer prices).

Explanatory variables used are:

Public Finance:

- **GGSB_NPGDP:** General government structural balance (% of GDP).
- **GGXWDG_NGDP:** General government gross debt (% of GDP).

Economic Output, Productivity & PPP:

- **PPPPC:** GDP per capita based on purchasing power parity (international dollars).

International Trade & Balance:

- **TX_RPCH:** Volume of exports of goods and services (% change).
- **TM_RPCH:** Volume of imports of goods and services (% change).

Savings & Investment:

- **NID_NGDP:** Total investment (% of GDP).

4.1. Missing Values (MICE-BR)

Iterative Imputation with Bayesian Ridge (MICE-BR)

Treats each missing variable **as a regression** on the other observed variables within a **Bayesian** framework.

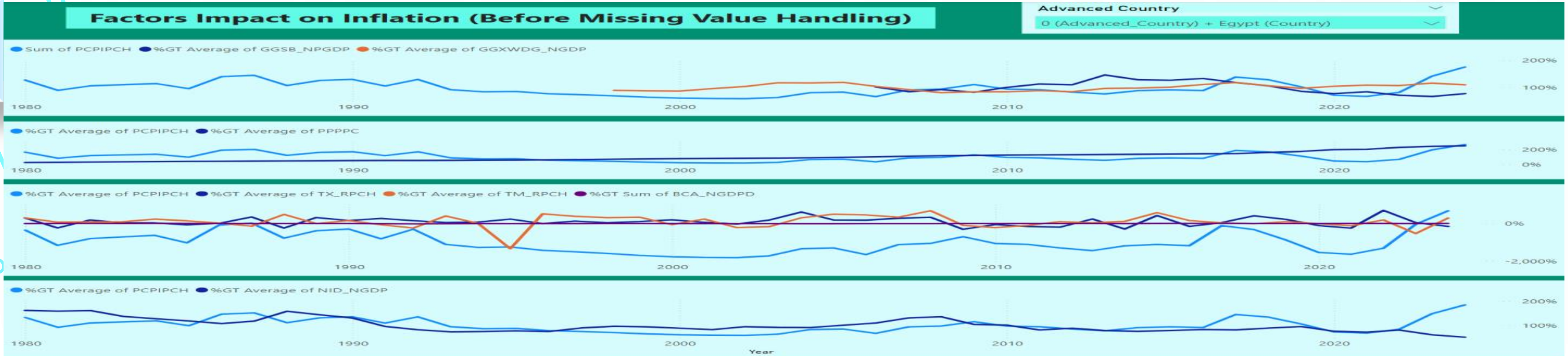
Applied separately for **each country** to preserve country-specific **heterogeneity**.

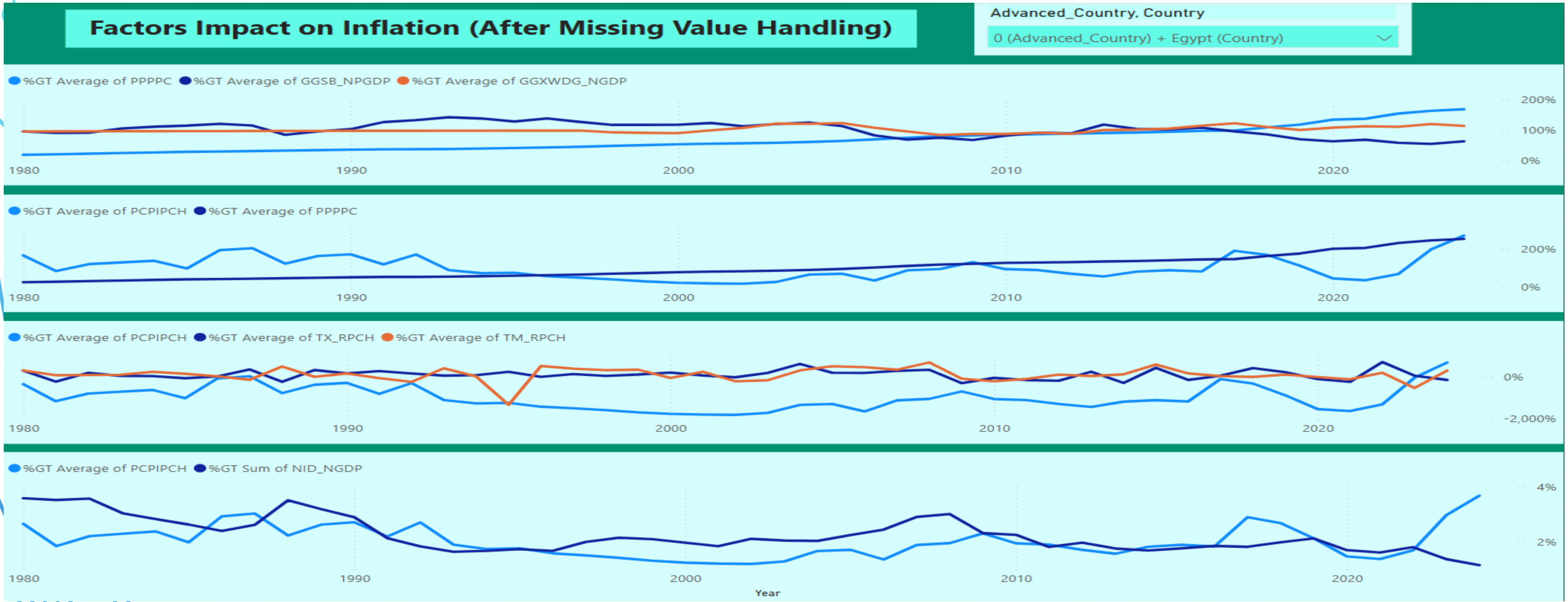
Imputes missing values without discarding the rich **relationships** among variables.

Python Code

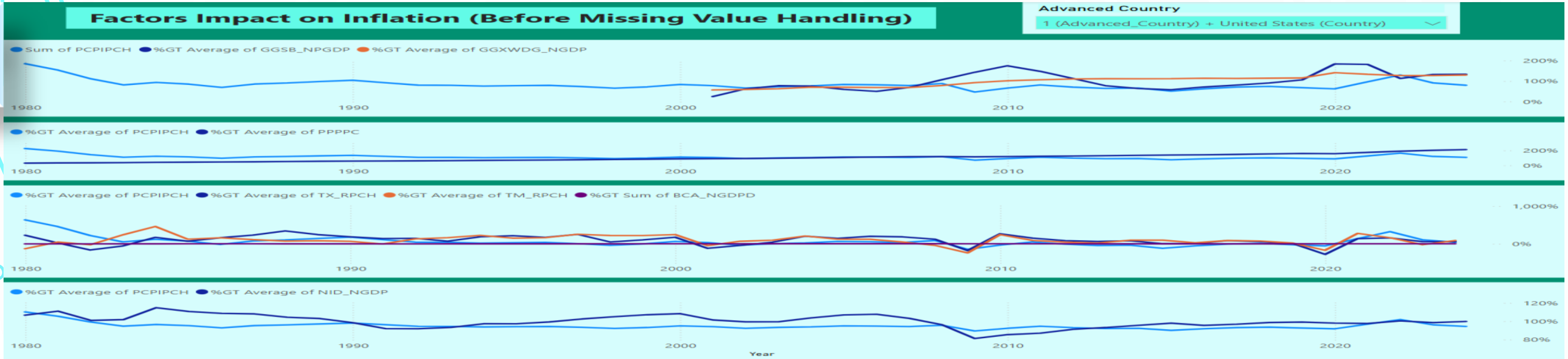
```
# Handle missing values for each country
def handle_missing_country(country):
    imputer = IterativeImputer(estimator=BayesianRidge(), max_iter=20,
random_state=0, verbose=2)
    country[cols_with_na] = imputer.fit_transform(country[cols_with_na])
    return country
# Apply the imputation process by country
df_interpolated = df_panel_b.groupby('Country',
group_keys=False).apply(handle_missing_country)
```

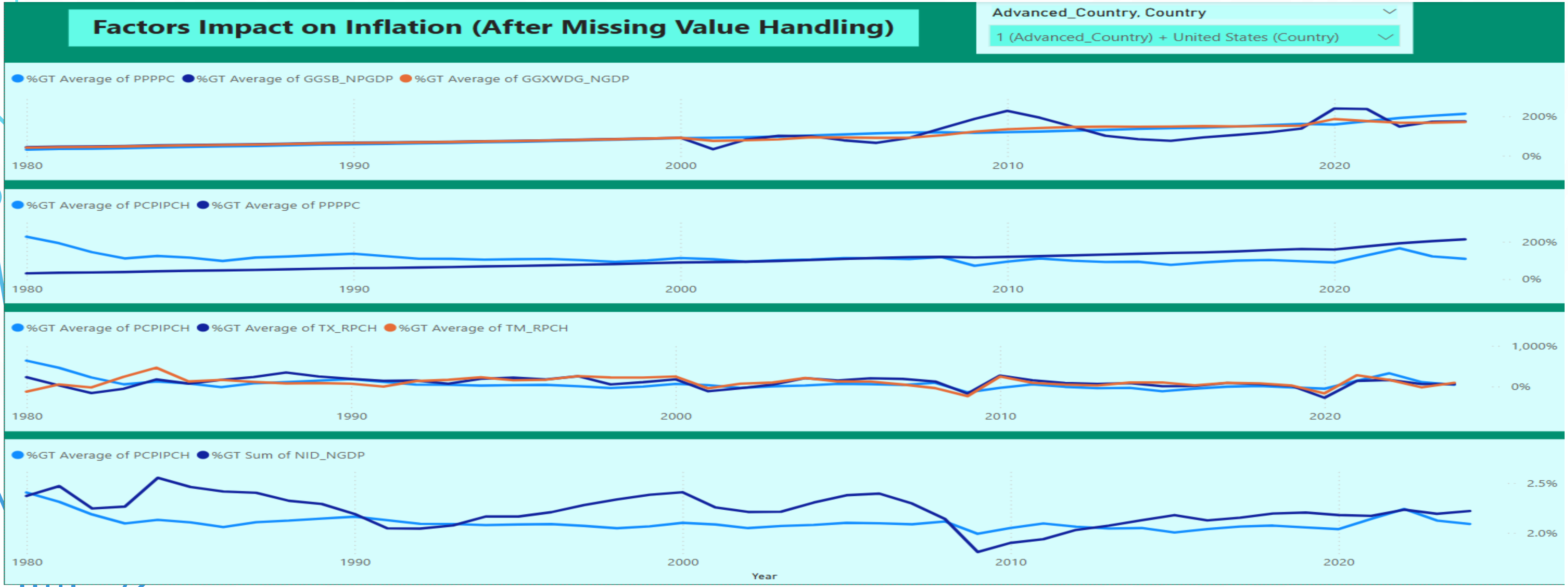
Dashboard (Egypt)



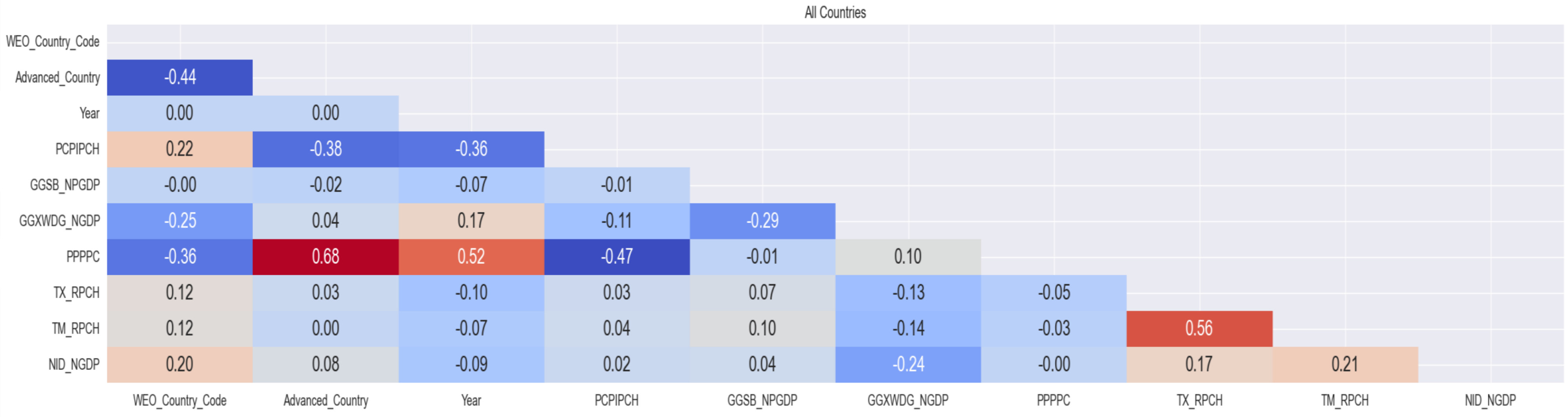


Dashboard (US)



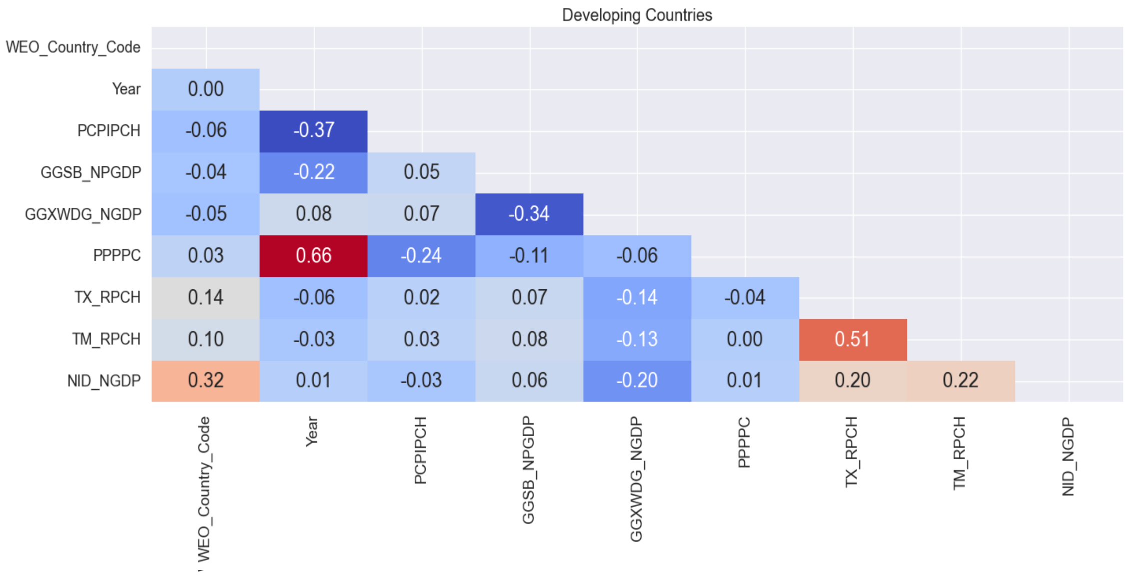


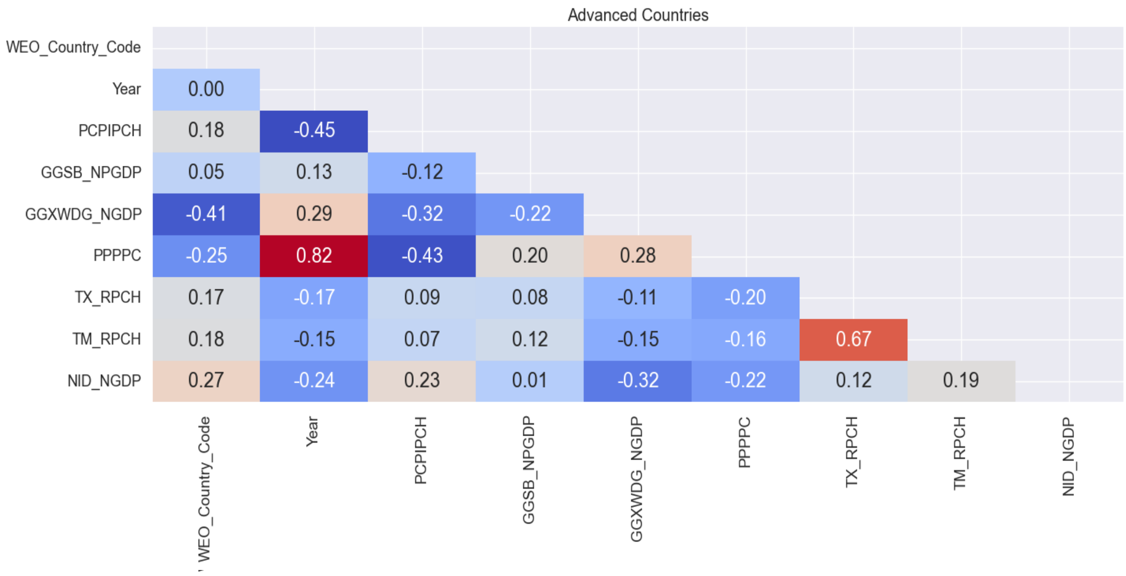
4.2 Correlation And (VIF)

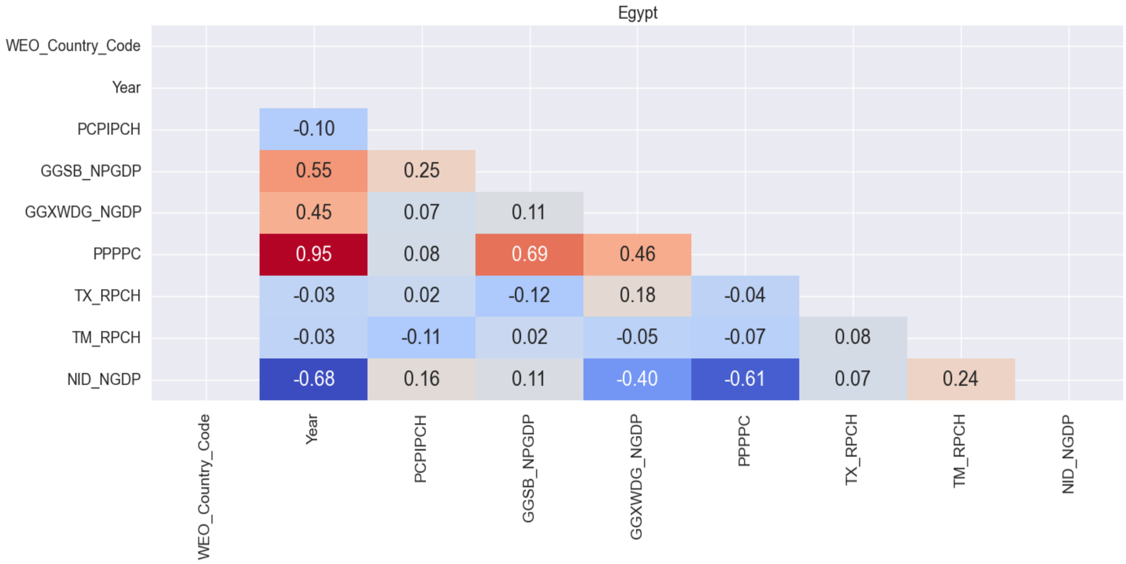


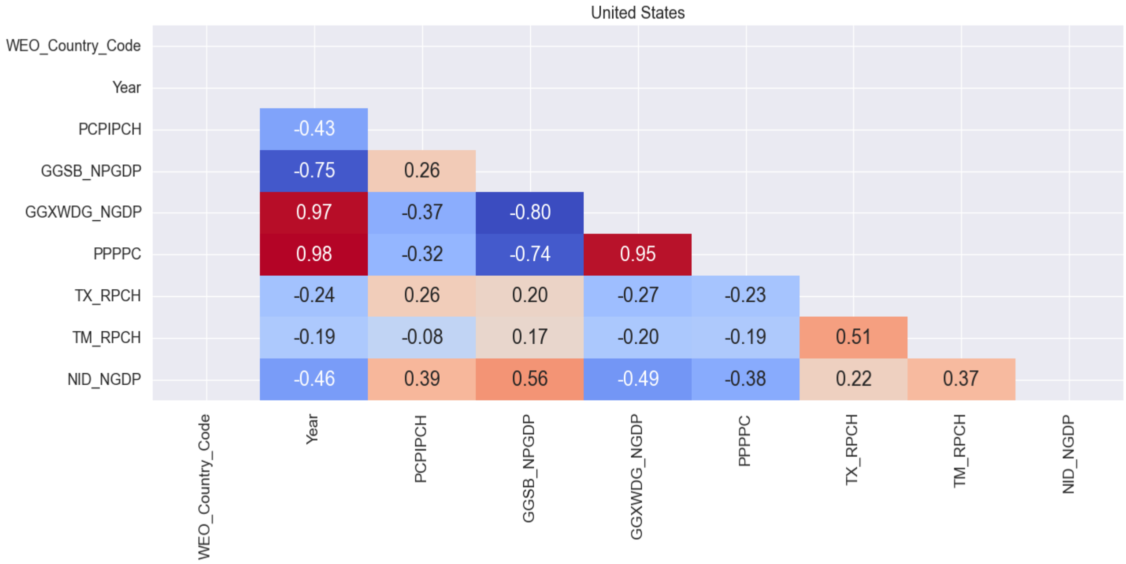
Feature	VIF	Feature	VIF	Feature	VIF
GGSB_NPGDP	1.752526	GGXWDG_NGDP	3.403836	PPPPC	3.457779
TX_RPCH	2.082648	TM_RPCH	2.119305	NID_NGDP	4.473428

Correlation









5. Panel Data Models

Pooled OLS: Ignores both country-specific and time-specific heterogeneity.

Fixed Effects (FE): Assumes country effects are time-invariant, country-specific characteristics (intercept for each country).

Random Effects (RE): Assumes country effects are randomly distributed (treated as random parameters).

Two-Step Difference GMM (Arellano–Bond):

- Addresses **endogeneity** using internal instruments (e.g., past values).
- Adds **lagged** dependent variables to model persistence over time.
- Suitable for panels with many countries (N) and a few years (T) **(N≫T).**
- Robust to **autocorrelation** and **heteroskedasticity** with proper diagnostics.

5.1 (POLS), (FE), and(RE)

The Pooled OLS model explains approximately 45% of the variation in inflation rates across the sample, as indicated by the R-squared value of 0.4525.

The overall model is statistically significant ($F = 454.25$, $p < 0.001$), suggesting that the selected regressors jointly explain inflation variation across countries and time.

The Fixed Effects model explains approximately 46.9% of the within-country variation in inflation, as indicated by the R-squared value.

The model is statistically significant overall ($F = 535.95$, $p < 0.001$), confirming that the selected regressors jointly explain meaningful within-country variation in inflation rates.

The Random Effects model explains approximately 46.5% of the variation in inflation across countries and over time.

The model is statistically significant overall ($\chi^2 = 3{,}262.87$, $p < 0.001$), indicating strong joint explanatory power of the selected regressors.

5.2.A. Wald test: Pooled OLS vs Fixed Effects

H_0: All individual (country) effects are jointly equal to zero → Pooled OLS is sufficient.

H_1: At least one individual effect is non-zero → Fixed Effects model is preferred.

```
# Wald test: Pooled OLS vs Fixed Effects
waldtest(pooled_model, fe_model, test = "F")
```

Results: F-statistic = 9.48 (p-value = 0.0021)

Decision: Since the p-value < 0.05, we reject the null hypothesis.

Conclusion: The test confirms that individual (country-specific) effects are statistically significant, and therefore, the **Fixed Effects** model is preferred over Pooled OLS.

5.2.B. Hausman Test: Fixed Effects vs Random Effects

H_0: Random Effects model is consistent and efficient.

H_1: Random Effects model is inconsistent

```
# Hausman test: fixed vs random effects
phtest(fe_model, re_model)
```

Results: Chi-squared = 2.58 (p-value = 0.8597)

Decision: Since the p-value > 0.05, we fail to reject the null hypothesis.

Conclusion: There is no statistical evidence of correlation between the country-specific effects and the regressors, implying that the **Random Effects model is consistent** and preferred over the Fixed Effects model in this case.

✓ (A sample of countries and a long period of time)

5.3.A. Diagnostic Testing and Model Validity Assessment

Key Assumptions of the Random Effects Model:

A) No heteroskedasticity (Breusch–Pagan and White Test):

H_0: Homoskedasticity (constant variance). H_1: Presence of heteroskedasticity.

Breusch–Pagan Test: Statistic = 13.62 (p-value = 0.0182)

White Test: Statistic = 18.01 (p-value = 0.0062)

Decision: Since both p-values are below 0.05, we reject the null hypothesis of homoskedasticity.

B) No autocorrelation (Breusch–Godfrey Test):

H_0: No serial correlation. H_1: Serial correlation exists.

Results: Chi-squared = 751.38, p-value < 2.2e-16

Decision: With a p-value near zero, we reject the null hypothesis of no serial correlation in the residuals.

5.3.B. Diagnostic Testing and Model Validity Assessment

C) Pesaran's Cross-sectional Dependence (CD) Test:

$\boldsymbol{H_0}$: Cross-sectional independence. $\boldsymbol{H_1}$: Cross-sectional dependence.

Results: z-statistic = 22.65 (p-value < 2.2e-16)

Decision: Reject the null hypothesis significant cross-sectional dependence exists.

D) Levin, Lin, and Chu (LLC) Panel Unit Root (Stationarity) Test:

$\boldsymbol{H_0}$: Non-stationarity (unit root present). . $\boldsymbol{H_1}$: Stationarity.

Results: Overall statistic = -20.16 (p-value < 2.2e-16)

Decision: Reject the null hypothesis the variable is stationary (PCPIPCH rate series is stationary).

- ✓ **We must use robust standard errors because:** (Heteroskedasticity, Serial correlation, and Cross-sectional dependence).
- ✓ **Lagged dependent variables must be included because:** the series is stationary, which supports using a dynamic panel model like Difference GMM.

Correcting Standard Errors: Driscoll–Kraay Robust Estimation

Variable	Estimate	Robust Std. Error	t-value	p-value	Significance
GGSB_NPGDP	-4.1547	4.3543	-0.95	0.3401	
GGXWDG_NGDP	0.3200	0.3864	0.83	0.4076	
PPPPC	-0.00059	0.00020	-2.91	0.0037	**
TX_RPCH	0.0169	0.2415	0.07	0.9442	
TM_RPCH	0.7598	0.2488	3.05	0.0023	**
NID_NGDP	0.0022	0.6100	0.0035	0.9972	

5.4. Two-Step Difference GMM (Arellano-bond)

Two-Step Difference GMM (Arellano–Bond):

Accounts for :

- Endogeneity (lagged inflation).
- Autocorrelation (by differencing and higher-order serial correlation)
- Heteroskedasticity(by a robust weighting matrix)
- Omitted variable bias (by differencing away unobserved fixed effects)
- Efficiency (via robust two-step weighting)

5.4 Two-Step Difference GMM (Arellano-bond)

Variable	Estimate	Std. Error	z-value	p-value	Significance
lag(PCPIPCH, 1)	0.2741	0.0936	2.93	0.0034	**
GGSB_NPGDP	-3.2433	1.4750	-2.20	0.0279	*
GGXWDG_NGDP	0.2870	0.1456	1.97	0.0487	*
PPPPC	-0.00022	0.00019	-1.13	0.2593	
TX_RPCH	0.2467	0.3035	0.81	0.4162	
TM_RPCH	0.5278	0.2599	2.03	0.0423	*
NID_NGDP	-0.5358	0.3329	-1.61	0.1075	

5.5. Diagnostic Tests for the GMM Model

Key Assumptions of the GMM Model:

A) Joint Significance of Coefficients (Wald Test):

$\boldsymbol{H_0}$: All slope coefficients are jointly zero. $\boldsymbol{H_1}$: At least one coefficient is non-zero.

Results: χ^2 (7) = 660.07 (p-value < 0.001)

Decision: Reject the null hypothesis, the explanatory variables are jointly significant.

B) Overidentifying Restrictions (Sargan Test):

$\boldsymbol{H_0}$: All instruments are valid . $\boldsymbol{H_1}$: At least one instrument is invalid.

Results: χ^2 = 11.86 (p-value = 0.457)

Decision: Fail to reject the null hypothesis, Instruments are valid.

C) Arellano–Bond Test for First/ Second -Order Autocorrelation:

$\boldsymbol{H_0}$: No serial correlation. $\boldsymbol{H_1}$: Serial correlation exists.

Results: [AR (1)]: z = −0.99(p-value = 0.3199) & [AR (2)]: z = −0.58 (p-value = 0.5607)

Decision: Fail to reject null hypothesis **, No serial correlation in the residuals.** [AR (1)&(2)]

5.6. Conclusion: Model Comparison and Final Selection

After estimating and evaluating four model specifications—Pooled OLS, Fixed Effects, Random Effects, and Two-Step Difference GMM—we summarize the findings:

- Pooled OLS failed to account for country heterogeneity and produced biased estimates.
- Fixed Effects controlled unobserved heterogeneity but was inconsistent under endogeneity.
- Random Effects passed the Hausman test and provided a consistent structure but ignored dynamics and endogeneity.
- Difference GMM successfully addressed autocorrelation, heteroskedasticity, and endogeneity, and passed all diagnostic checks (Sargan, AR(1), AR(2)).

6. Recommendations and Future Work

Based on the findings, it is recommended that applied macroeconomic research and policy forecasting in inflation contexts prioritize dynamic panel estimators—particularly Two-Step Difference GMM—when dealing with persistent variables and potential endogeneity.

Future Work:

- Expansion of Dataset Coverage.
- Inclusion of Energy Price Indices.
- Extended Diagnostic Testing.
- Development and Integration of Machine Learning (ML) and Deep Learning (DL).

Appendix (Codes)

(A): SQL Codes:-

Dataset: https://github.com/1145267383/Panal_Data_Inflation/tree/main/02-Dataset

Database: https://github.com/1145267383/Panal_Data_Inflation/blob/main/03-Clean_Organize_EDA/04-SQL.ipynb

(B): Python Codes:-

Clean and organize and EDA: https://github.com/1145267383/Panal_Data_Inflation/tree/main/03-Clean_Organize_EDA

Descriptive and Correlation Analysis: https://github.com/1145267383/Panal_Data_Inflation/tree/main/04-Descriptive_Correlation_Analysis

Models: https://github.com/1145267383/Panal_Data_Inflation/blob/main/05-Models_Panel_Data/01-Models_Python.ipynb

(C): R Codes:-

Models: https://github.com/1145267383/Panal_Data_Inflation/blob/main/05-Models_Panel_Data/02-Models_R.ipynb

References

- Hsiao, C. (2003). Analysis of Panel Data. Cambridge University Press.
- Verbeek, M., & Nijman, T. (1992). "Testing for selectivity bias in panel data models." International Economic Review, 33(3), 681–703.
- Moulton, B. R. (1990). "An illustration of a pitfall in estimating the effects of aggregate variables on micro units." Review of Economics and Statistics, 72(2), 334–338.
- Arellano, M., & Bond, S. (1991). "Initial conditions and moment restrictions in dynamic panel data models." Journal of Econometrics, 87(1), 115–143.
- Baltagi, B. H. (2008). Econometric Analysis of Panel Data (4th ed.). John Wiley & Sons.
- Fischer, S. (1993). "The role of macroeconomic factors in growth." Journal of Monetary Economics, 32(3), 485–512.
- Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data (2nd ed.). MIT Press.

END