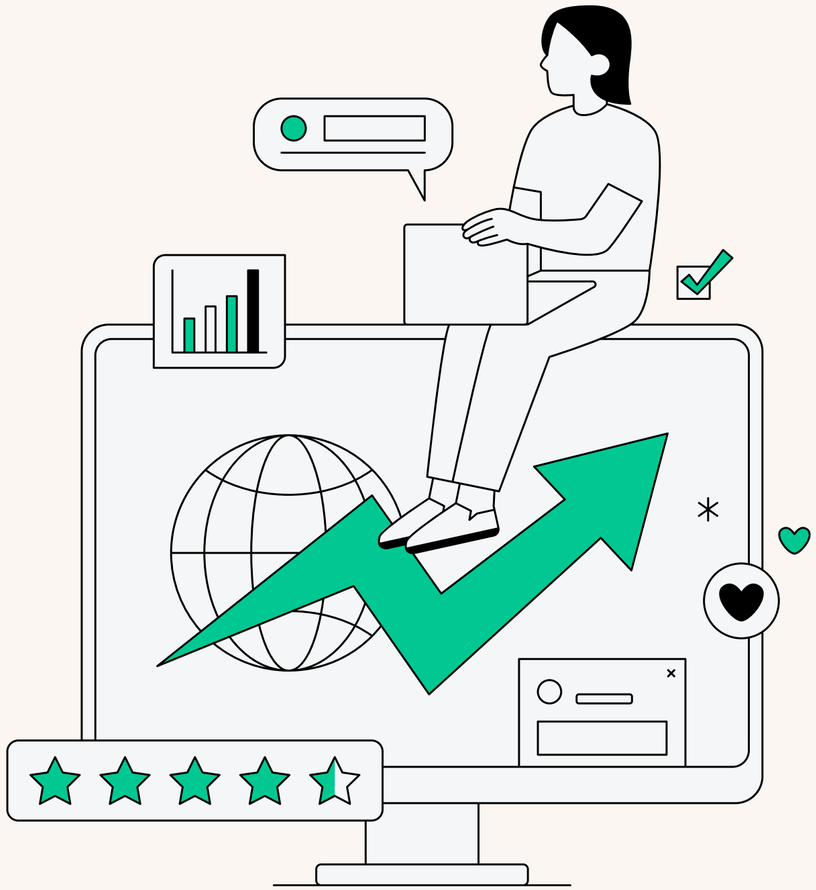
Presented by Team Ninja Turtles

Alvin Yao, Benjamin Swenson, Chris Korabik, Danylo Sovgut, and Paul Pham

Analysis Challenge

Labor Underutilization Rate

Statistical Modeling



What is U-3 Labor Underutilization Rate?

The U-3 Labor Underutilization Rate is the official unemployment rate measured by the Bureau of Labor Statistics (BLS). It represents the **total number of unemployed people as a percentage of the civilian labor force.**

This metric is crucial because it:

- **Serves as a key economic indicator for policymakers**
- **Influences monetary and fiscal policy decisions**
- **Reflects the overall health of the labor market**
- **Acts as a benchmark for economic performance**

Our Goal: **Predict U-3 Labor Underutilization Rate for December 2024**

Our primary goal is to **develop a statistical model to predict the U-3 Labor Underutilization Rate for December 2024**, providing stakeholders with actionable insights for strategic planning.

Dataset Used: **The Bureau of Labor Statistics (from 01/2022 - 09/2024)**
33 rows of data

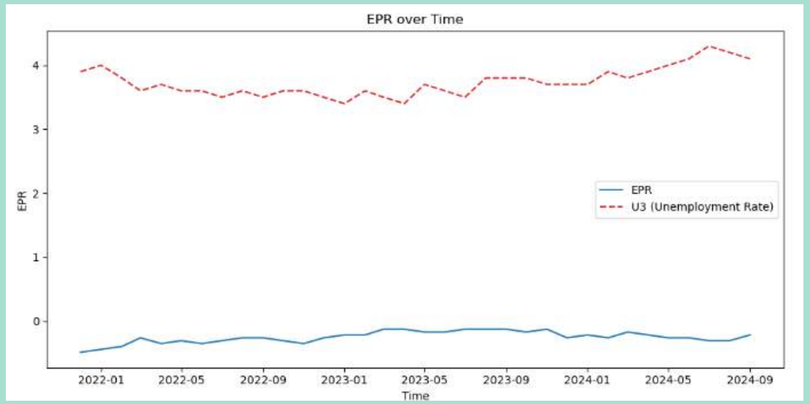
Employment and Labor Force Metrics:

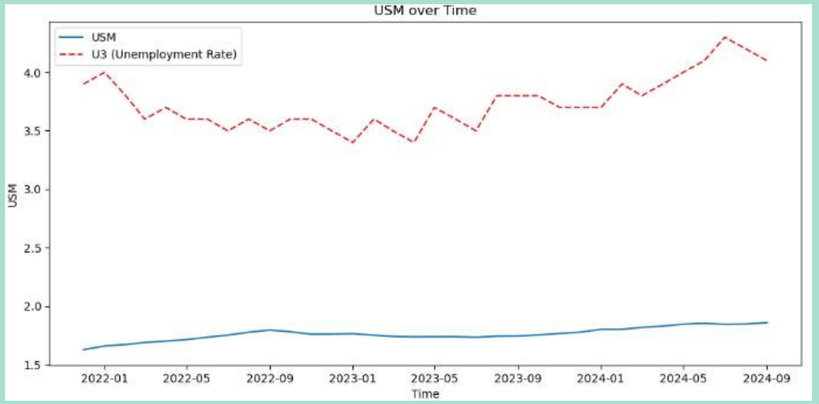
- **AWU (Average Weeks Unemployed):** Measures the mean duration of unemployment
- **CLF (Civilian Labor Force):** Total number of employed and unemployed individuals
- **EPR (Employment Population Ratio):** Percentage of working-age population employed
- **LFPR (Civilian Labor Force Participation Rate):** Percentage of working-age population in labor force
- **TNE (Total Nonfarm Employment):** A comprehensive count of paid U.S. workers, excluding farm workers, private household employees, unpaid volunteers, military service members, self-employed individuals, non-profit employees)
- **U3 (U-3 Unemployment Rate):** Target variable - official unemployment rate
- **U3_lag1 (We created this):** Lag variable for U3, is the previous months unemployment rate

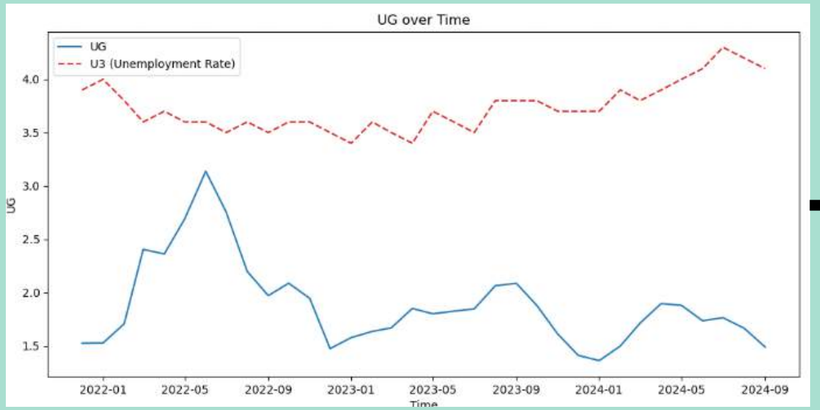
Consumer Price Index (CPI) Components

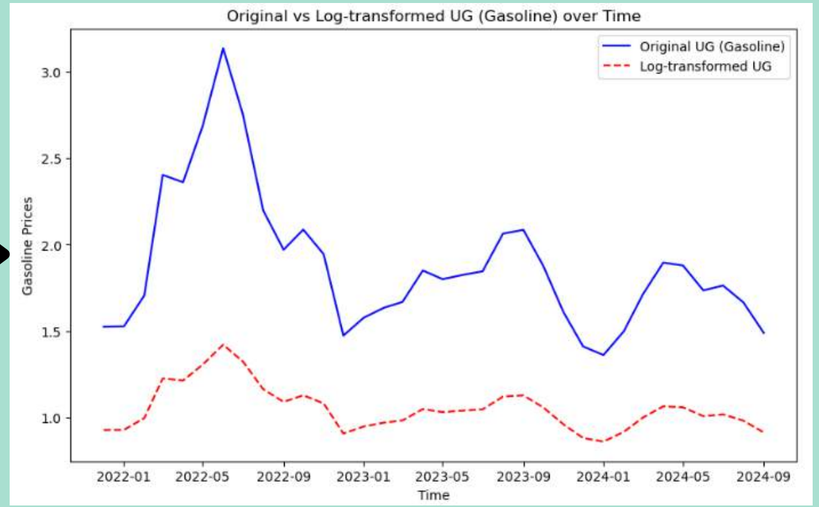
- **UA (All Items):** Overall consumer price index covering all goods and services
- **UFH (Food at Home):** Price index for grocery store food items
- **UG (Gasoline):** Retail gasoline prices across U.S. cities
- **USH (Housing):** Shelter costs including rent, utilities, and housing operations
- **USM (Medical Care):** Healthcare costs including services and medical supplies
- **UST (Transportation):** Vehicle-related costs and public transportation expenses

Exploratory Data Analysis



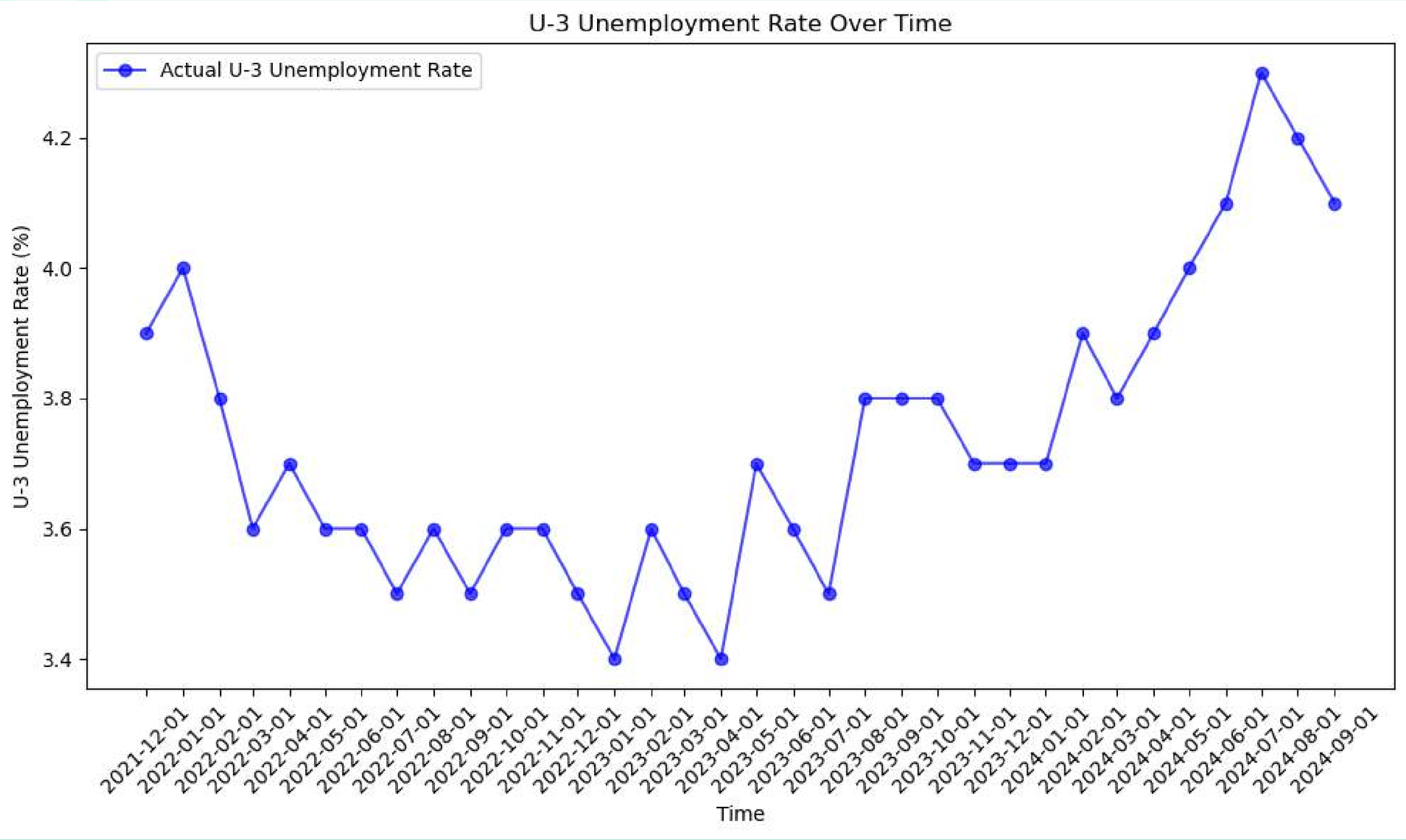






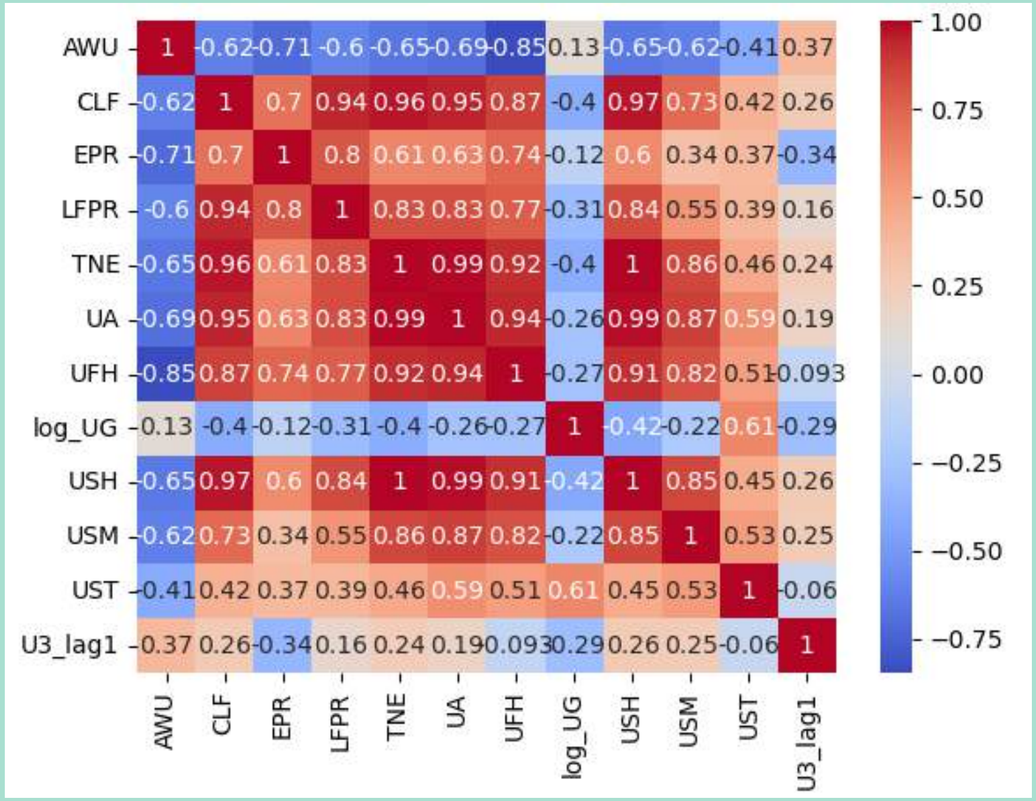
- Most of our variables were flat like the two graphs on top
- UG had much higher irregularity so we applied a **log transform** to flatten it out.

What does the u3 data look like?



Initial VIF and Correlation Matrix

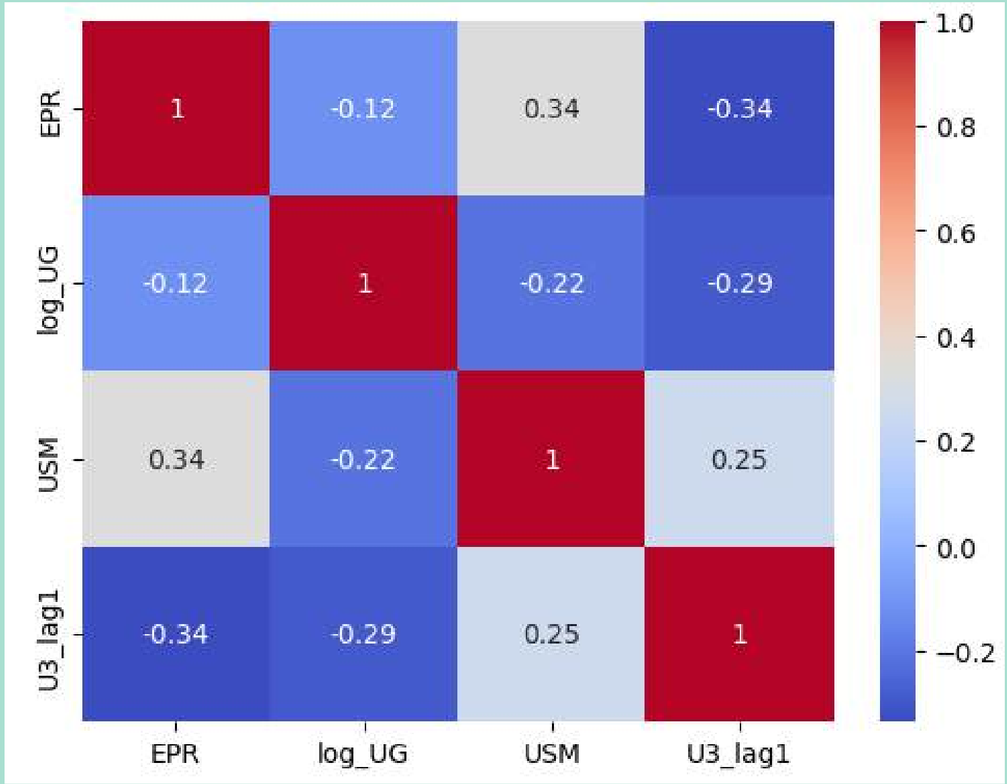
	Feature	VIF
0	const	48710.778991
1	AWU	8.481231
2	CLF	193.451093
3	EPR	10.435184
4	LFPR	39.995687
5	TNE	680.182946
6	UA	4719.111320
7	UFH	103.142113
8	log_UG	49.535124
9	USH	4031.570703
10	USM	19.490590
11	UST	125.770625
12	U3_lag1	6.407183



- Very High VIF on variables like **UA** and **USH**
- Correlation Matrix shows variables being highly correlated from **0** to **1** from low to high, respectively.

Dropping variables based on VIF

	Feature	VIF
0	const	1489.861336
1	EPR	1.491930
2	log_UG	1.163048
3	USM	1.364354
4	U3_lag1	1.505048



- Dropped highest VIF variable at each stage until all variables have a VIF of **5** or less.
- Correlation Matrix shows fixed correlation issue

Data Preprocessing & Timeline

- **Dataset was naturally clean with no missing values**
- **All variables were scaled using StandardScaler**
- **Data timeframe: January 2022 to September 2024**
    - **Chosen to avoid COVID-19 related spikes in the data***

See Appendix for reasoning

Model Selection Process

We trained a Linear, Gamma, Negative Binomial, Poisson, and Tweedie Regression model on the data and chose the best performing one to continue on with the analysis.

Akaike Information Criterion (AIC)

- The **AIC** is a metric often used for **Model Selection.**
- Balances **goodness-of-fit** with **model complexity**.
- Best case **AIC** would be close to **0**.
- k is # of parameters in the model. This **penalizes inclusion of more predictors**.
- ln($\hat{L}$) is the **log-likelihood** of the model.

$$AIC = 2k - 2ln(\hat{L})$$

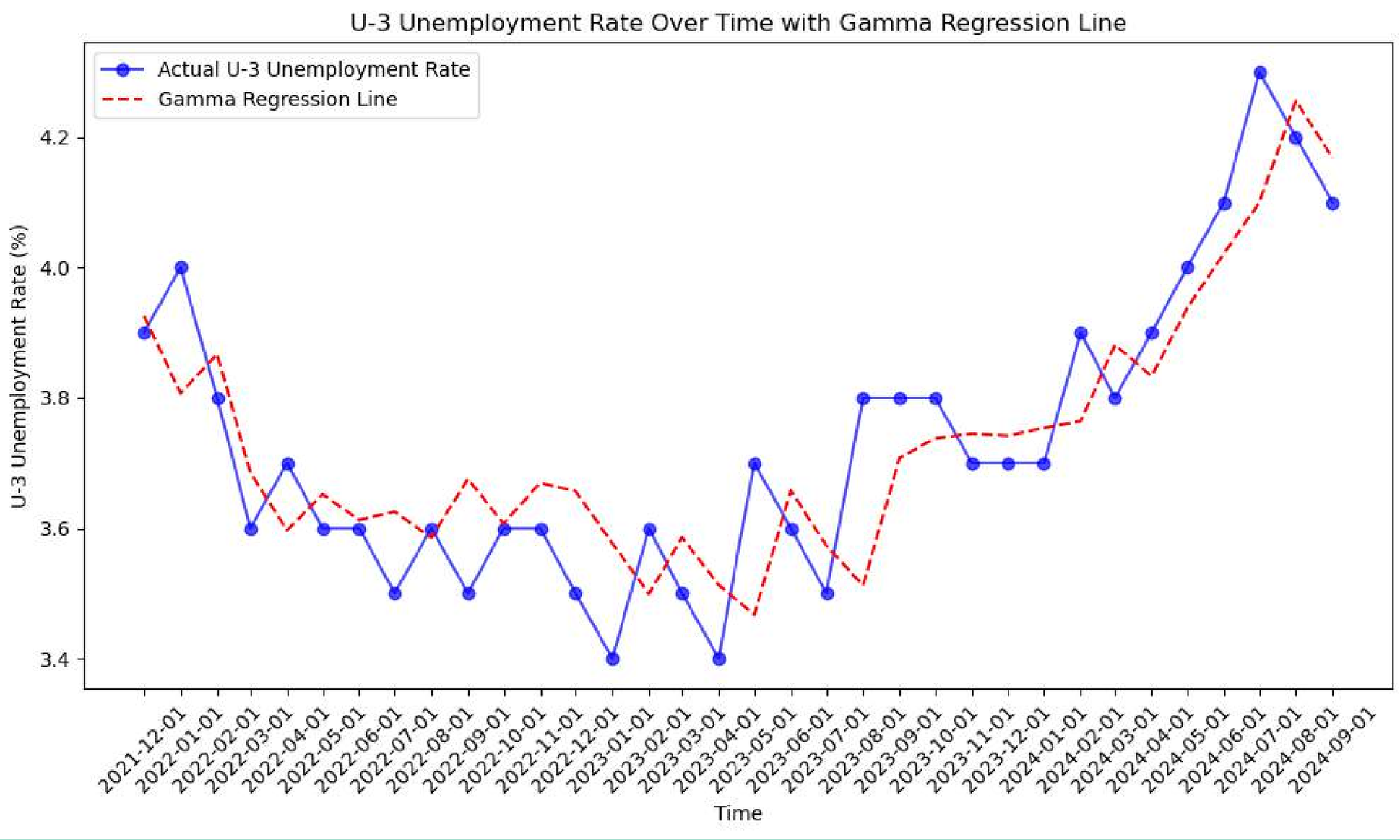
We Chose Gamma Regression Due to Low AIC Score

Model Type	Linear Regression	Gamma Regression	Negative Binomial Regression	Poisson Regression	Tweedie Regression
What is it best suited for?	Continuous outcomes with normal distribution	Positive, continuous outcomes with skewed distribution	Over-dispersed count data	Count data with low variance	Mixed distributions (e.g., continuous + discrete)
MSE	.0075	0.0077	0.0077	0.0076	0.0075
AIC	-26.39	-25.59	141.92	96.55	17379.90

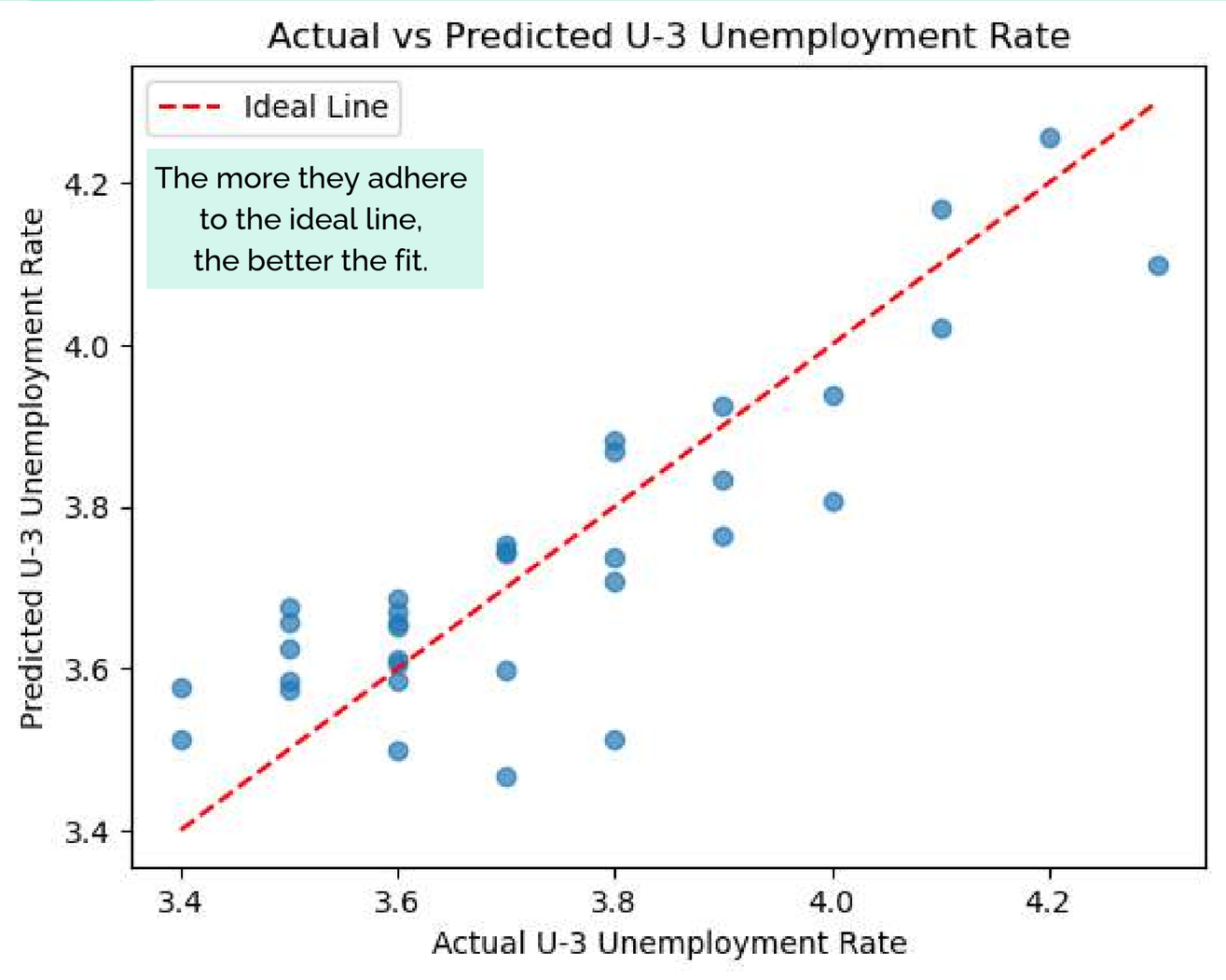
Model Output

Variable	Coefficient	Std Error	95% CI	Description
Const	0.1520	0.226	[-0.291, 0.595]	Baseline Rate
EPR	-0.0884	0.078	[-0.241, 0.065]	Employment Population Ratio
log_Ug	-0.0240	0.049	[-0.120, 0.072]	log transformed gasoline prices
USM	0.2829	0.124	[0.041, 0.525]	medical care costs
u3_lag1	0.1790	0.032	[0.116, 0.242]	u3 lag variable

How does the Model fit?



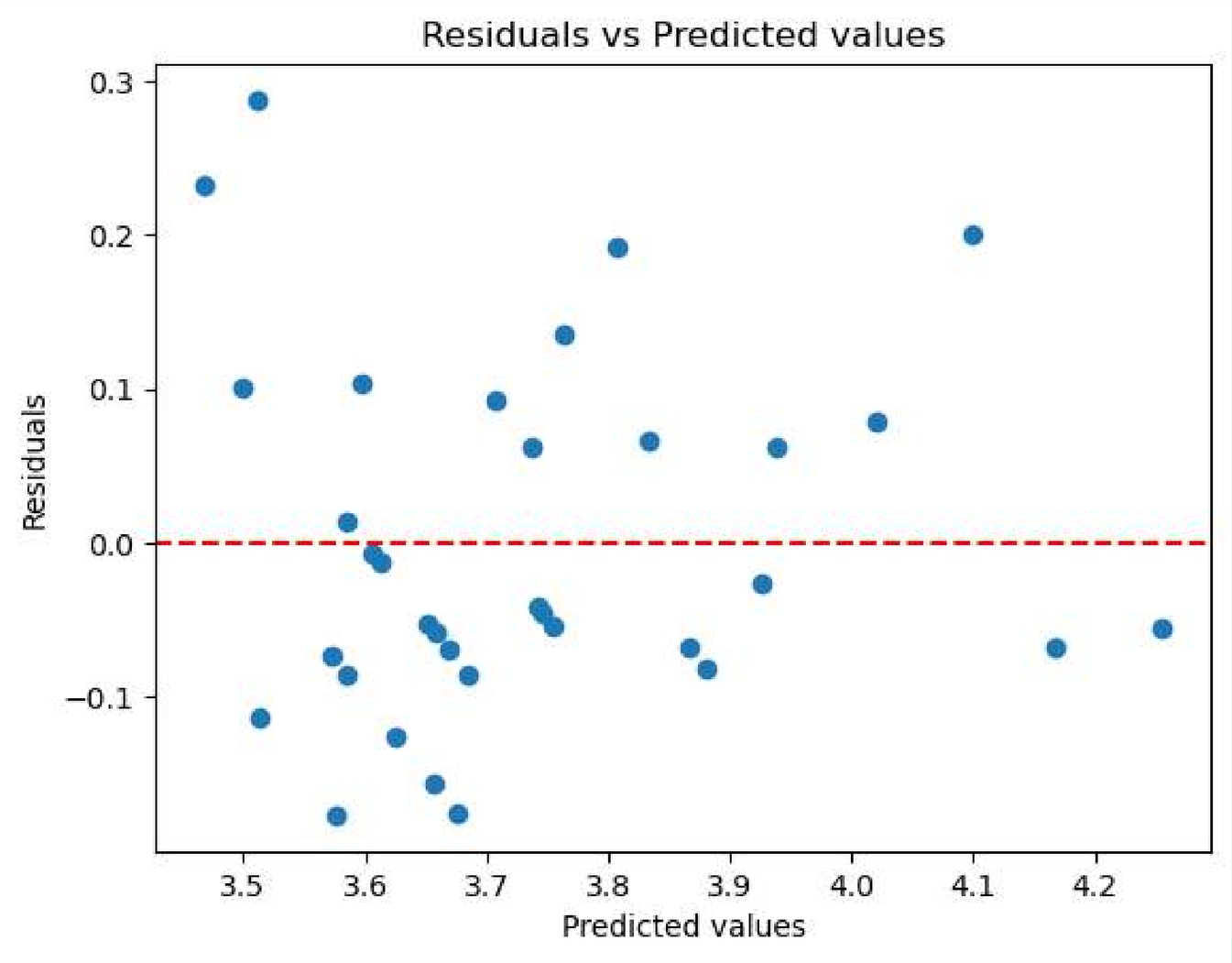
Another way to look at it ...



Pearson Correlation

0.856

Residuals look mostly random



Using 5-Fold Cross Validation

RMSE

0.1281

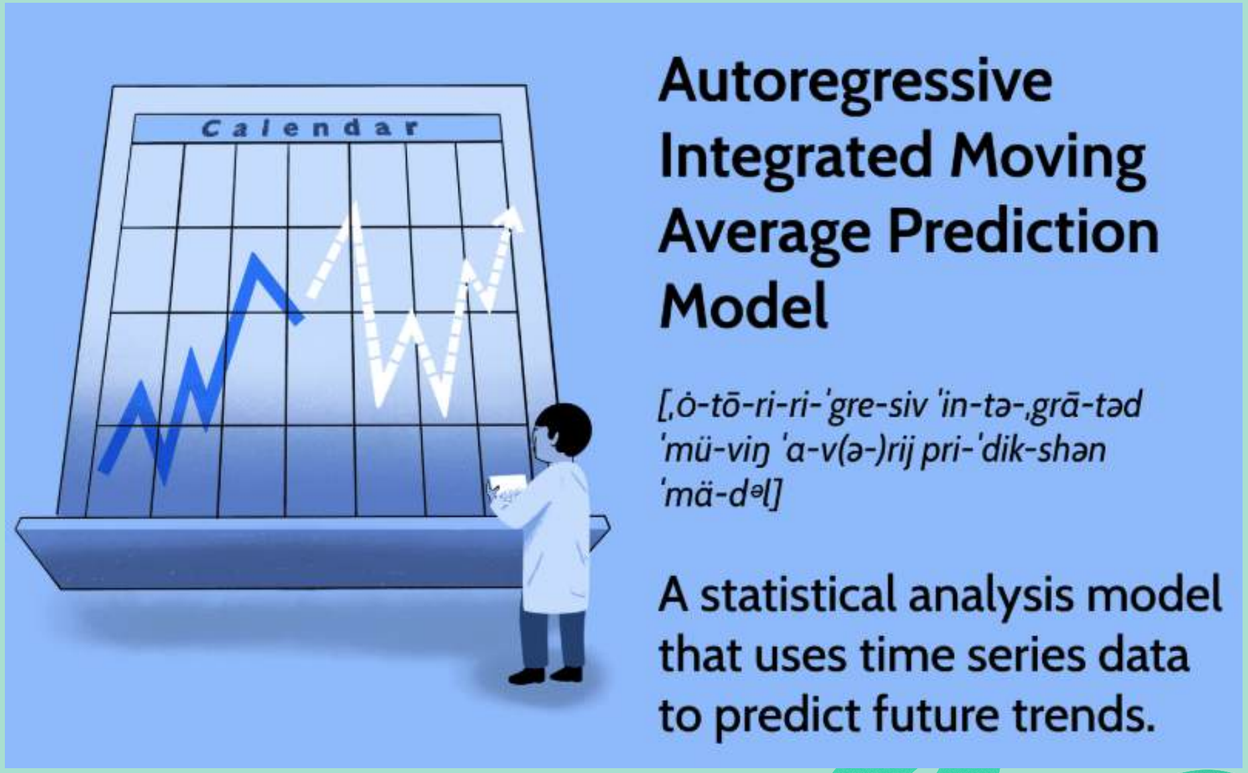
Pseduo-R2

0.8840

Predicting the Future

- Since our data goes up until September 2024, we need to predict the U3 Unemployment rate for October, and November due to the U3 lag variable we implemented. We need to have estimate values for the other variables as well.

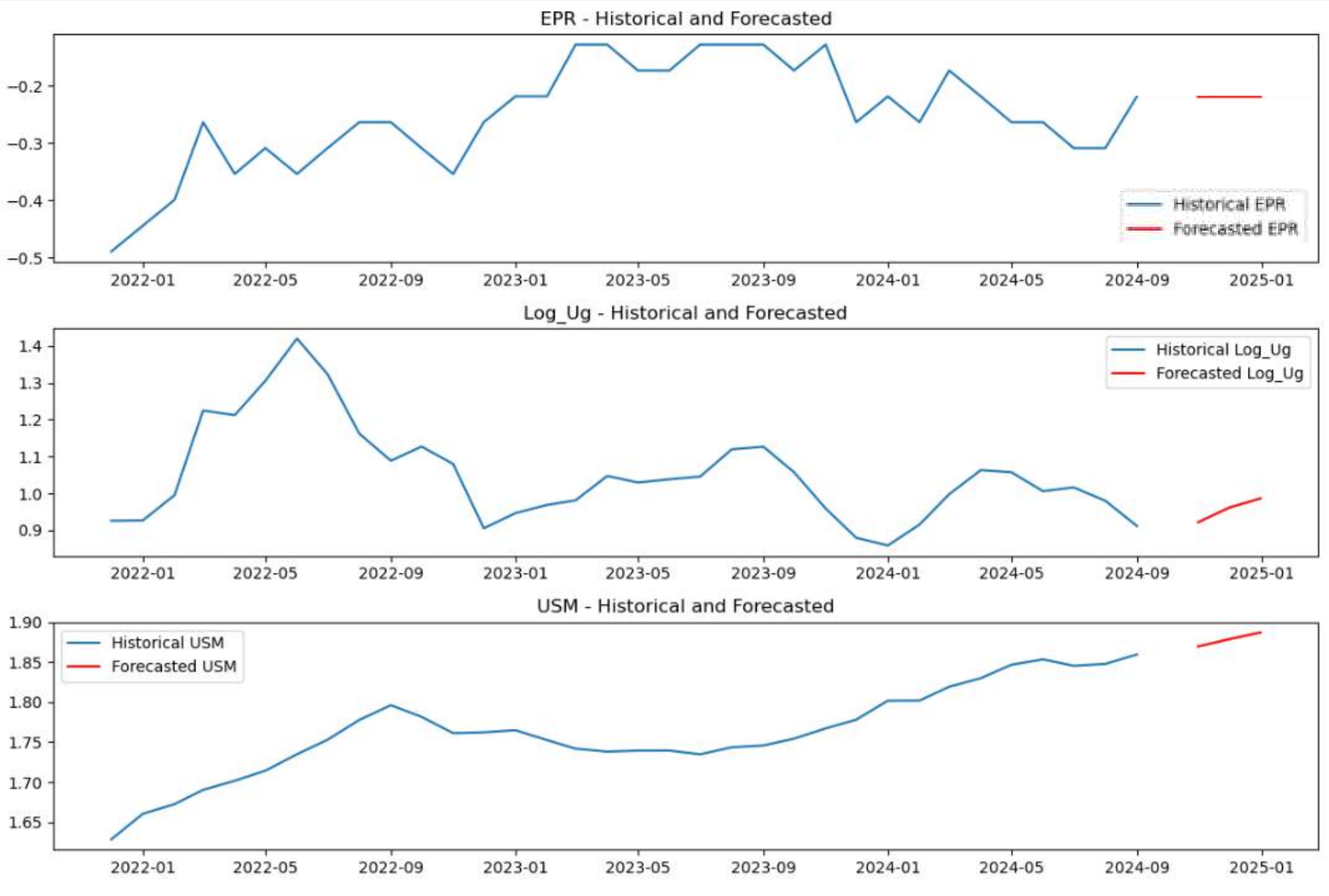
- We are utilizing ARIMA to predict the values of EPR, log UG, and USM for the months of October, November, and December.



Using Python pmdarima for auto finding of p,d,q

EPR	log_UG	USM
ARIMA(0,1,0)	ARIMA(1,0,1)	ARIMA(1,1,0)

ARIMA Value Predictions



Predictions

October 2024 U3 Predicted Rate - 4.178%
November 2024 U3 Predicted Rate - 4.169%

December 2024 U3 Predicted Rate

4.17% +/- .12%

APPENDIX

APPENDIX

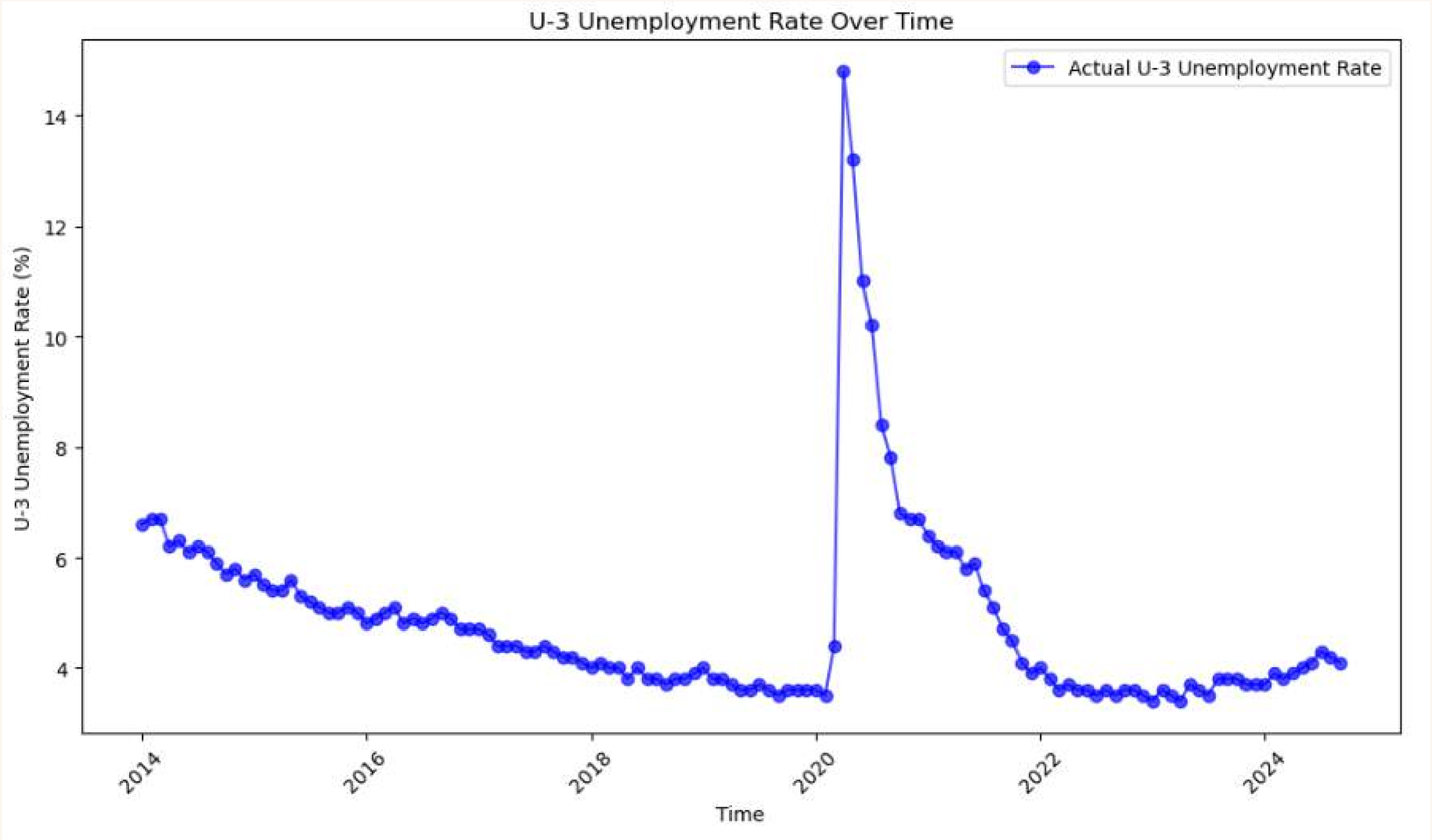
Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	34
Model:	GLM	Df Residuals:	29
Model Family:	Gamma	Df Model:	4
Link Function:	Log	Scale:	0.0011665
Method:	IRLS	Log-Likelihood:	24.520
Date:	Wed, 11 Dec 2024	Deviance:	0.033361
Time:	18:41:21	Pearson chi2:	0.0338
No. Iterations:	6	Pseudo R-squ. (CS):	0.8840
Covariance Type:	nonrobust		

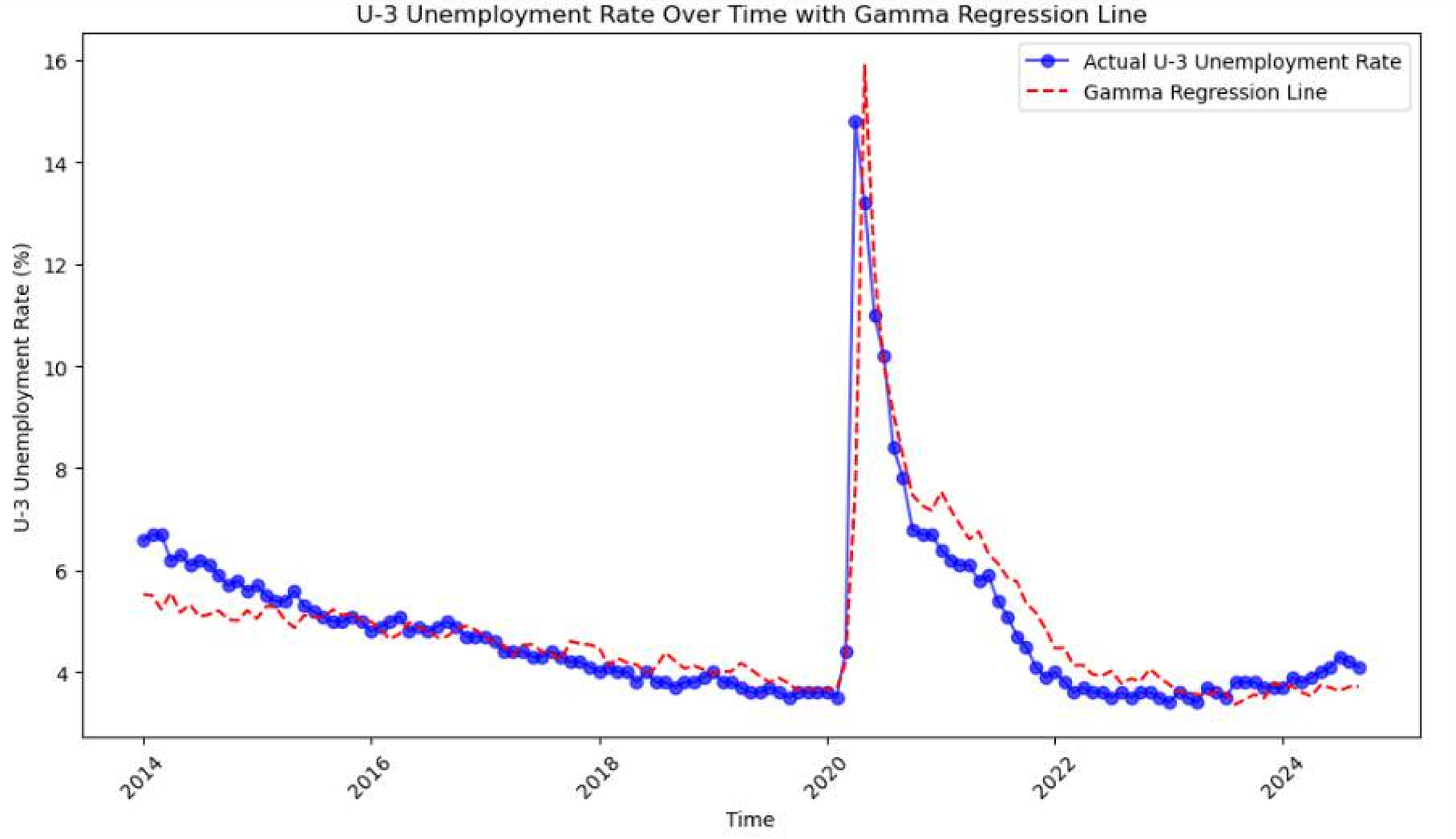
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1520	0.226	0.672	0.501	-0.291	0.595
x1	-0.0884	0.078	-1.132	0.258	-0.241	0.065
x2	-0.0240	0.049	-0.490	0.624	-0.120	0.072
x3	0.2829	0.124	2.290	0.022	0.041	0.525
x4	0.1790	0.032	5.567	0.000	0.116	0.242

Pearson Correlation: 0.8562271

APPENDIX


U-3 Unemployment Rate Over Time
Actual U-3 Unemployment Rate
U-3 Unemployment Rate (%)
14
12
10
8
6
4
2014
2016
2018
2020
2022
2024
Time

APPENDIX


U-3 Unemployment Rate Over Time with Gamma Regression Line
Actual U-3 Unemployment Rate
Gamma Regression Line
U-3 Unemployment Rate (%)
16
14
12
10
8
6
4
2014
2016
2018
2020
2022
2024
Time