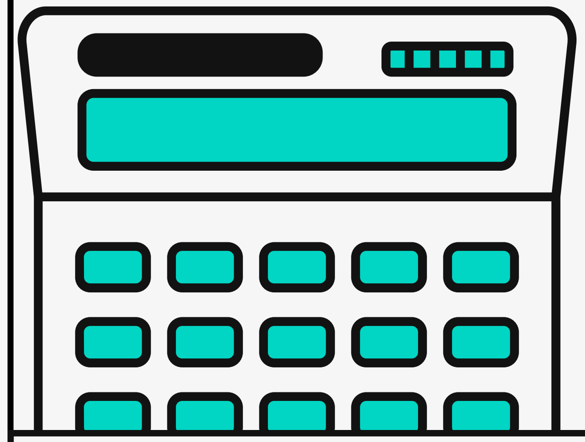
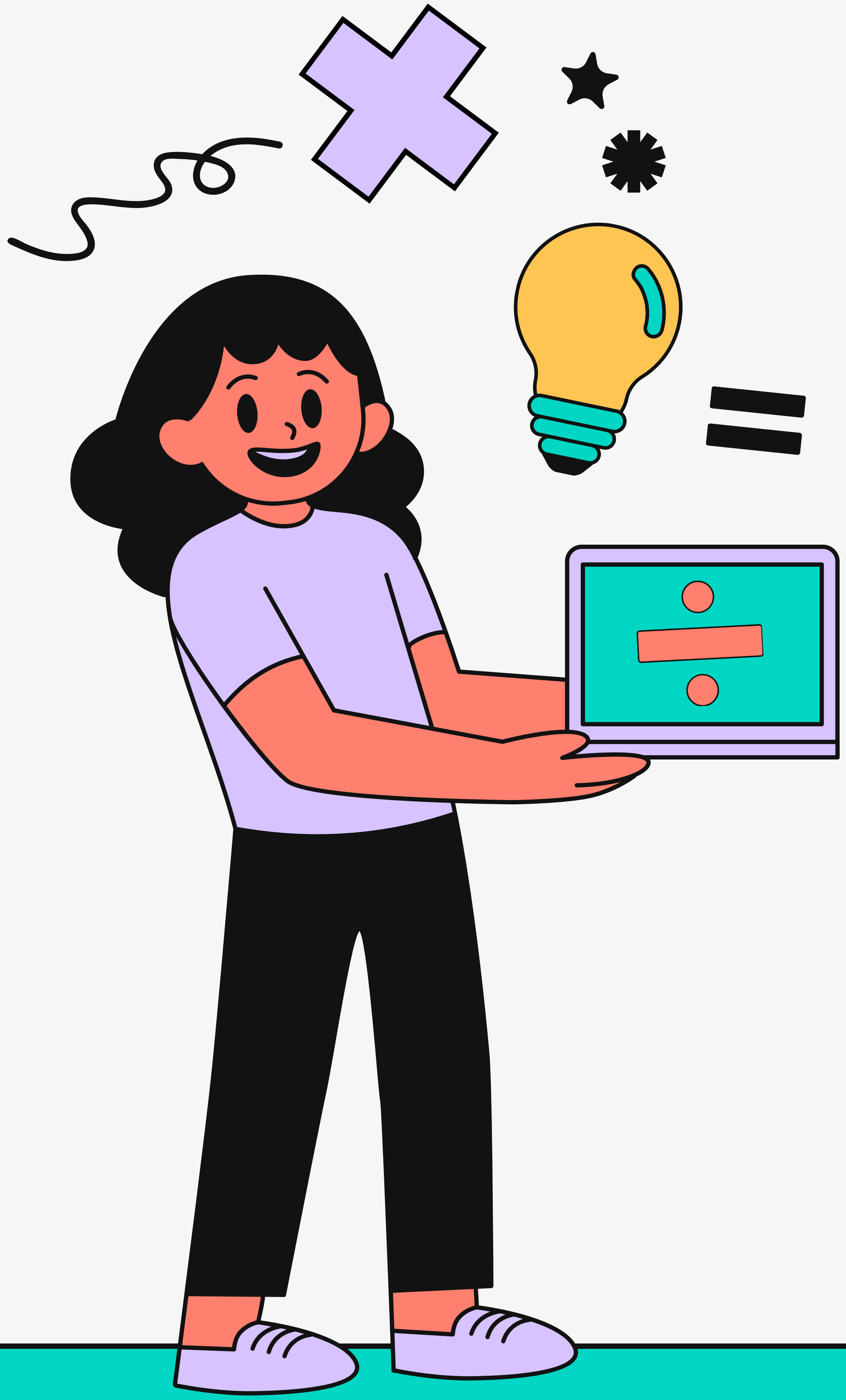
Introduction to Statistical Learning

with Applications



By: Duc Huy Nguyen

Mentor: Patrick Campbell

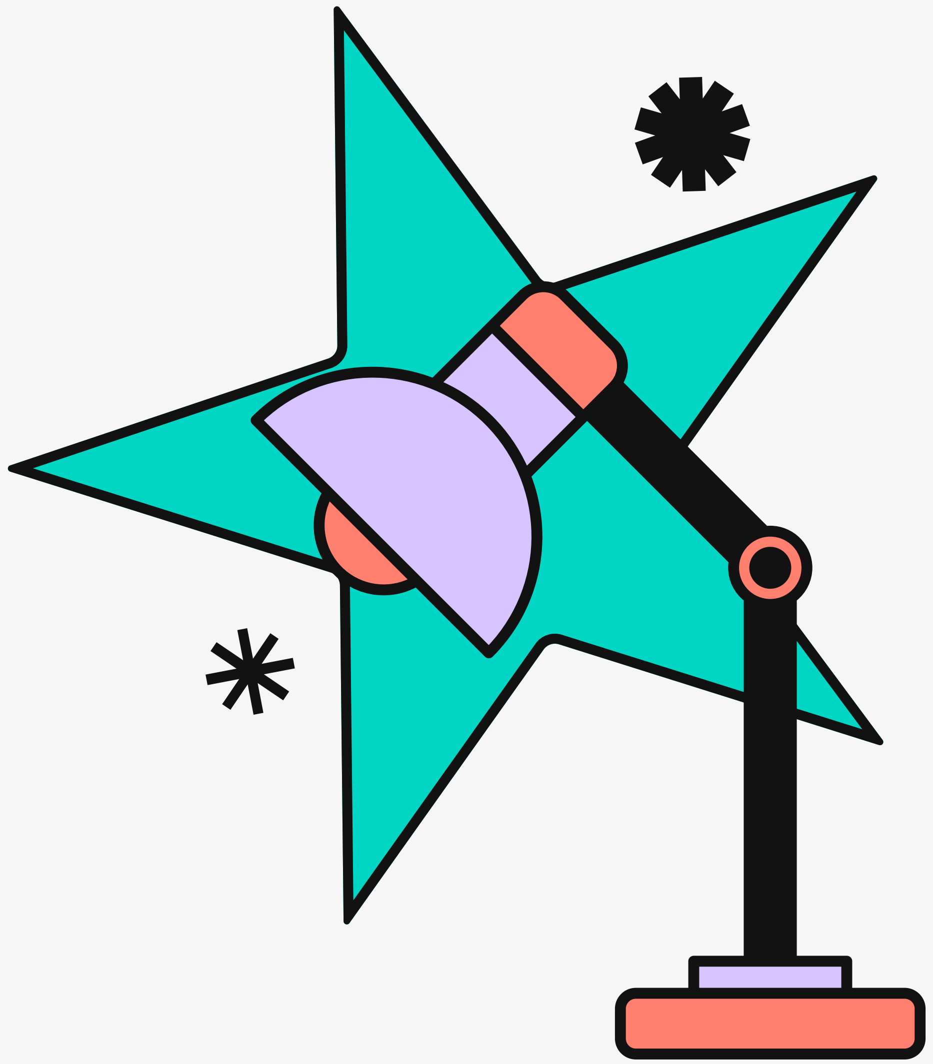


Agenda

- Regression Models
- Applications
- Classification Models
- Applications
- Conclusion

Regression

Predict continuous values
(e.g. prices, life expectancy, etc.)



Multiple Linear Regression

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

We would want to minimize the sum of squared residuals to minimize our error when we are fitting Linear Regression

$$\begin{aligned}
\mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2.
\end{aligned}$$

Shrinkage Models

Ridge Regression

$$\underset{\hat{\beta}_R}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right) + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

The shrinkage penalty is squared the magnitude of coefficient

Coefficients converges towards (but not) 0 as the parameter gets larger

Reduce the effects of irrelevant predictors

Lasso Regression

$$\underset{\hat{\beta}_L}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right) + \lambda \sum_{j=1}^{p} \beta_j = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j$$

The shrinkage penalty is based on the absolute value of the coefficients

Coefficients converges towards and might get to 0 as the parameter gets larger

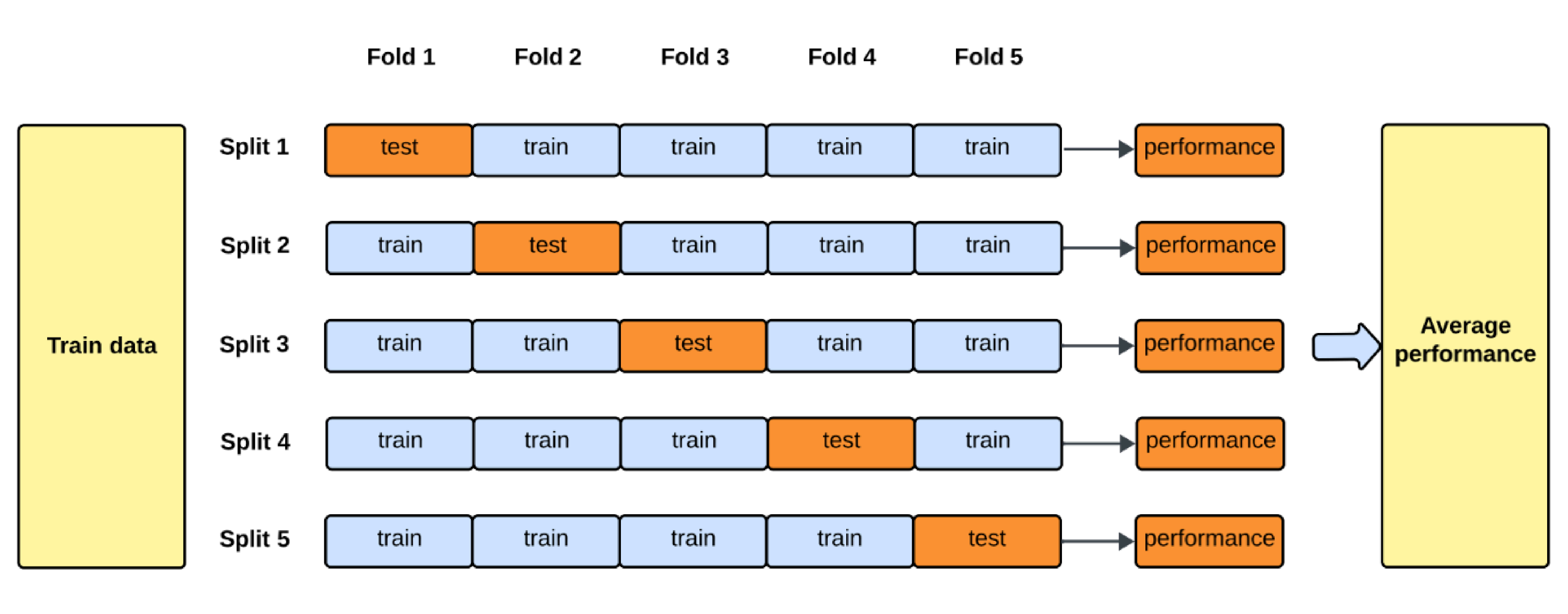
Can potentially selection important features to our linear and omit the irrelevant noises

Scaling of Predictors

Since we are trying to minimize our coefficients here, the scale of our predictors would matter in our model. So we need to standardize our data before model fitting.

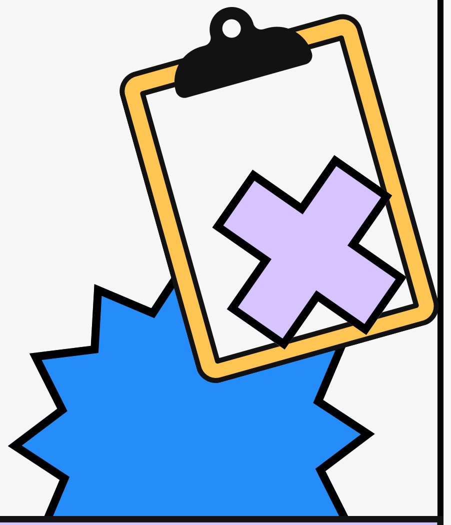
$$z = \frac{x - \mu}{\sigma}$$

How do we select our tuning parameters



Cross - Validation (CV): split the data randomly into k parts and use k - 1 of them for training and the other for validation (we usually use k = 5 or 10)

We want cross-validation to have a robust estimate of our model performance, mitigate overfitting, utilize data, and have effective hyperparameters tuning (for models like SVM, etc.)



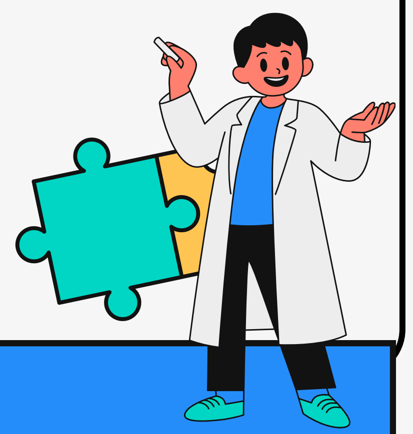
Our problems

Datasets

- Monthly data on GDP, CPI, S&P500, job postings, unemployment claims, crime data, etc.
- Mostly retrieved from the Federal Reserve Economic Data (FRED)
- Note that a few of this are interpolated

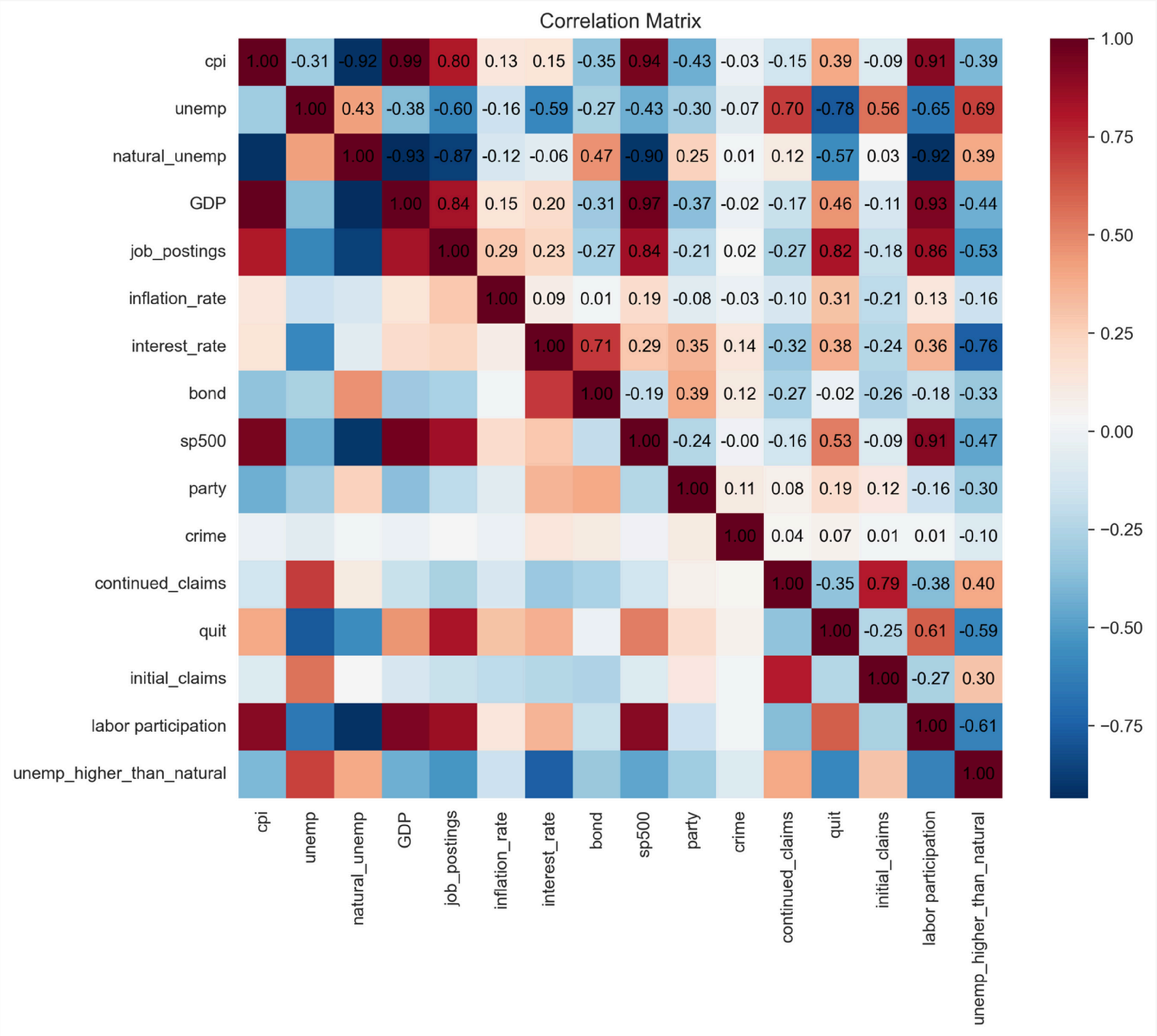
Questions

1. How can we predict unemployment using various economics predictors (potentially job postings)
2. When would the unemployment raise over the natural unemployment rate*



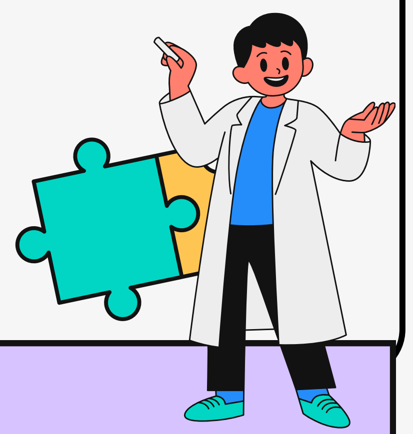
*I estimated this using the NAIRU, read more about that here: https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/nairu-data-set

Correlation Matrix



A notable relation here is between GDP and CPI with correlation up to 0.99 and GDP and S&P500 with correlation up to 0.97

Note that I also have Pairs Plot visualization for these variables so feel free to check the Github repo for that



Adjustments for Multicollinearity

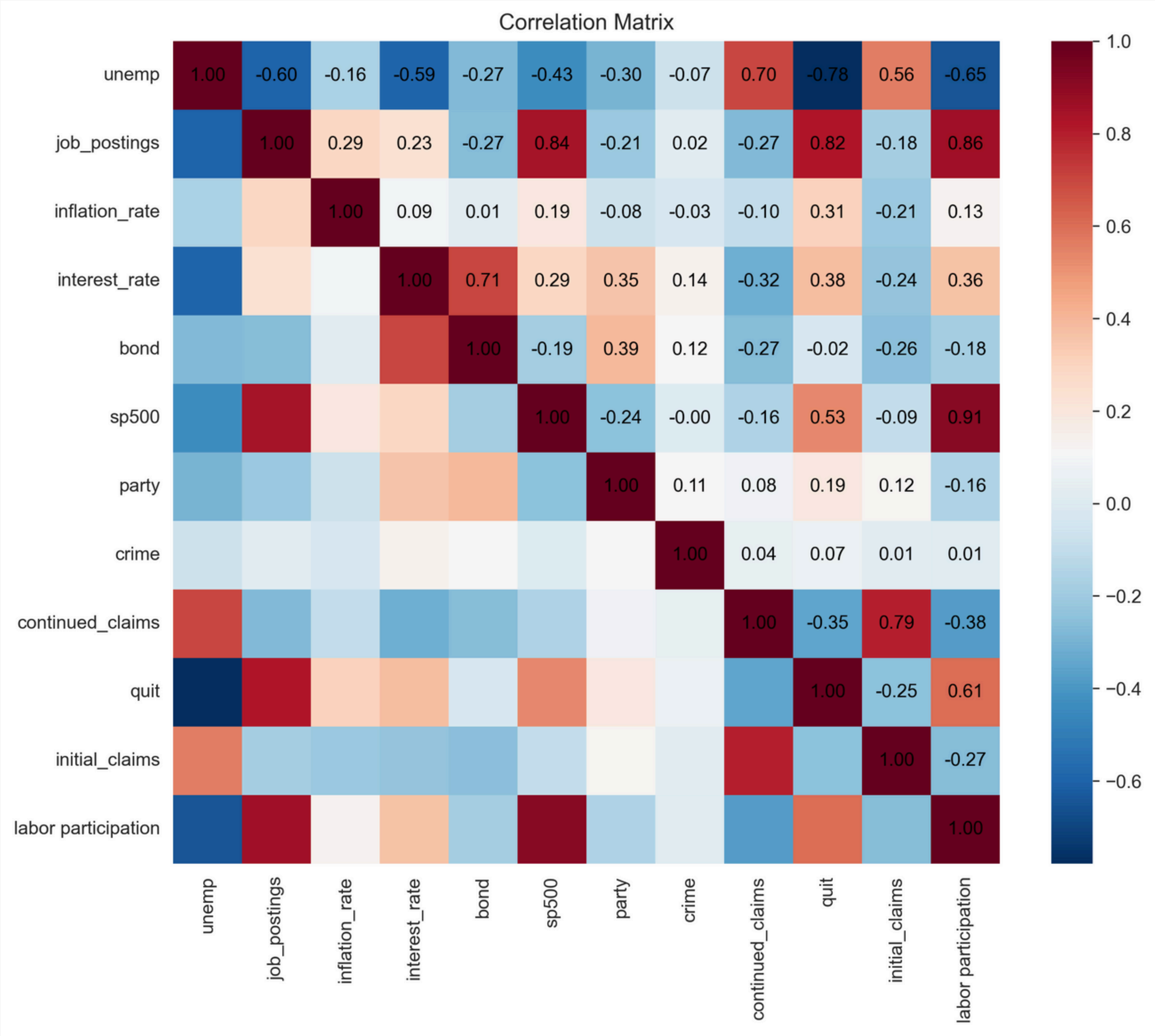
Why do we need to adjust for Multicollinearity: to increase our interpretability as we can identify direct relationship between predictors and response

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}$$

$$\det(A - \lambda I) = 0$$

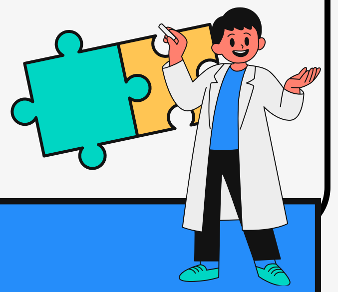
1. Compute the eigenvalues and the eigenvectors for the correlation matrix
2. Take the ratio of the max eigenvalue to all other eigenvalues elements-wise
3. Identify which element of the ratio vector is the highest
4. Choose the corresponding eigenvector for the highest element in the previous step
5. Identify which two elements in this eigenvectors are the highest in value

Correlation Matrix After Adjustments



GDP and CPI are omitted to deal with multicollinearity

We can see that though there are still some correlations between predictors, the overall correlations significantly decreased



Regression Model Results

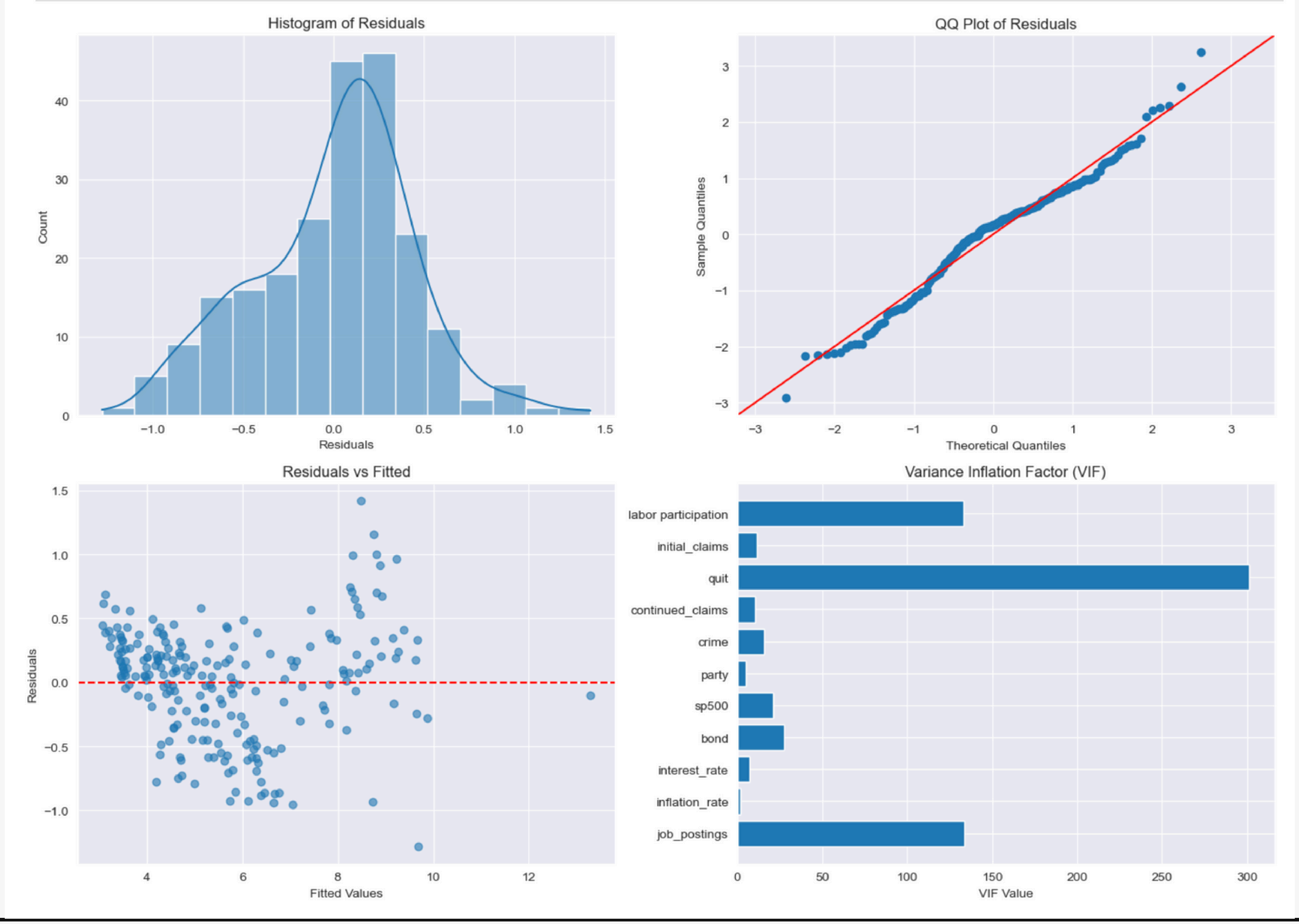
Multiple Linear Regression (Without Scaling)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  unemp   R-squared:                       0.948
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                     346.0
Date:                Tue, 02 Dec 2025   Prob (F-statistic):          5.67e-128
Time:                        13:33:44   Log-Likelihood:                -131.98
No. Observations:                 222   AIC:                             288.0
Df Residuals:                     210   BIC:                             328.8
Df Model:                          11
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                  31.0544      2.098     14.800      0.000      26.918      35.191
job_postings            0.0002    6.7e-05      3.378      0.001    9.43e-05       0.000
inflation_rate          0.1184      0.110      1.073      0.284      -0.099       0.336
interest_rate          -0.0229      0.035     -0.655      0.513      -0.092       0.046
bond                   -0.1991      0.056     -3.581      0.000      -0.309      -0.089
sp500                   0.0031      0.001      4.191      0.000       0.002       0.005
party                  -0.5710      0.106     -5.371      0.000      -0.781      -0.361
crime               -2.327e-06   1.76e-06     -1.325      0.187   -5.79e-06    1.13e-06
continued_claims     6.409e-08    6.7e-09      9.561      0.000    5.09e-08    7.73e-08
quit                   -3.1322      0.285    -10.983      0.000      -3.694      -2.570
initial_claims      -6.754e-08   5.47e-08     -1.234      0.219   -1.75e-07    4.04e-08
labor participation    -0.0001   1.45e-05    -10.068      0.000      -0.000      -0.000
==============================================================================
Omnibus:                        2.231   Durbin-Watson:                   2.062
Prob(Omnibus):                  0.328   Jarque-Bera (JB):                1.864
Skew:                          -0.191   Prob(JB):                        0.394
Kurtosis:                       3.234   Cond. No.                     1.16e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.16e+09. This might indicate that there are
strong multicollinearity or other numerical problems.

Train RMSE: 0.43847419579909785
Test RMSE: 0.647007010077397
```

Relationship is hard to draw here since predictors are on different scaled which disrupt our interpretation of coefficients

Model Diagnostics

Multiple Linear Regression (Without Scaling)



Regression Model Results

Multiple Linear Regression (With Scaling)

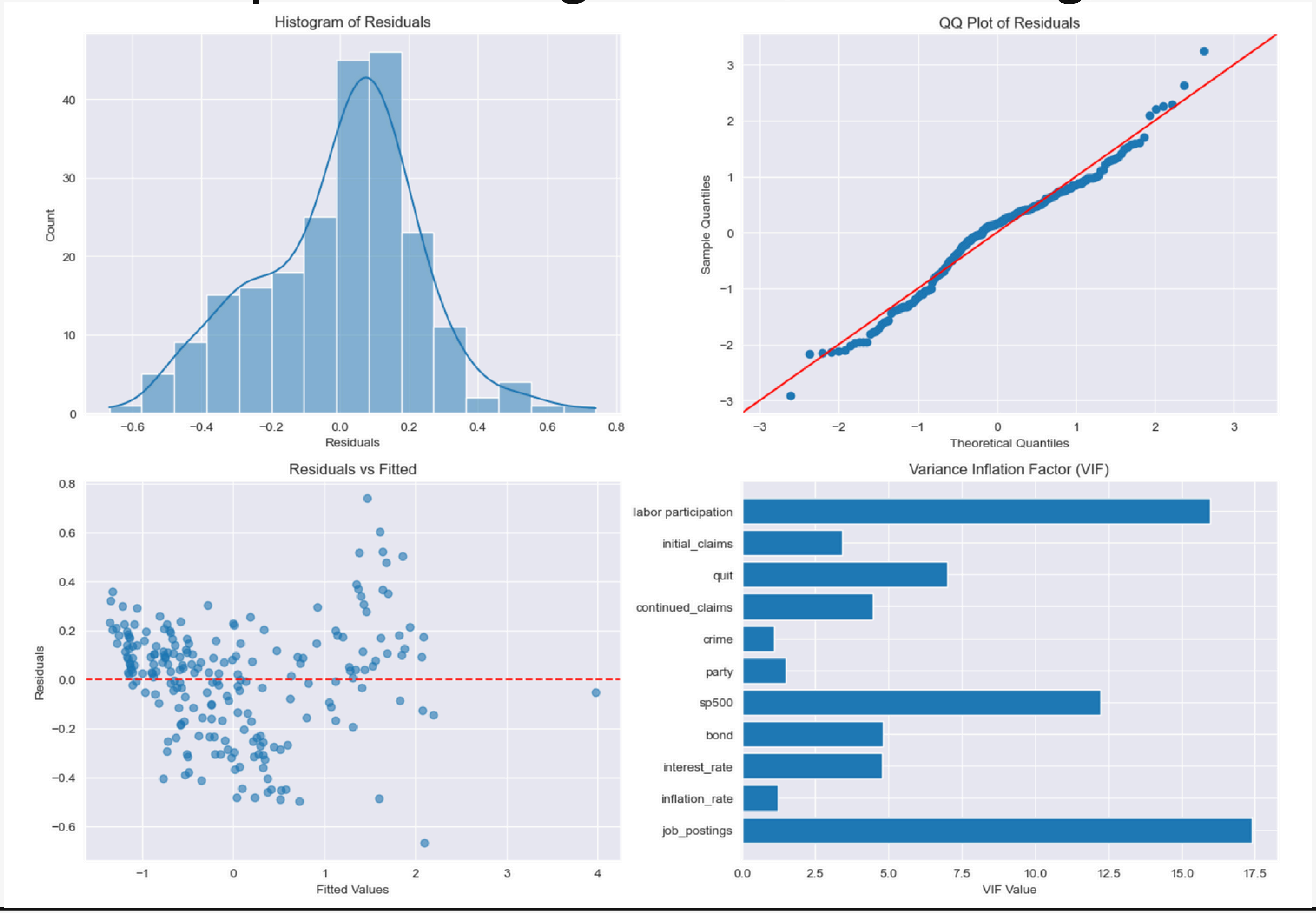
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.952
Model:                            OLS   Adj. R-squared:                  0.950
Method:                 Least Squares   F-statistic:                     380.4
Date:                Mon, 01 Dec 2025   Prob (F-statistic):          4.52e-132
Time:                        16:03:21   Log-Likelihood:                 22.557
No. Observations:                 222   AIC:                            -21.11
Df Residuals:                     210   BIC:                             19.72
Df Model:                          11
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.1352      0.030      4.522      0.000       0.076       0.194
job_postings            0.2439      0.075      3.256      0.001       0.096       0.392
inflation_rate          0.0155      0.017      0.910      0.364      -0.018       0.049
interest_rate          -0.0324      0.034     -0.941      0.348      -0.100       0.035
bond                   -0.1312      0.035     -3.793      0.000      -0.199      -0.063
sp500                   0.2935      0.053      5.570      0.000       0.190       0.397
party                  -0.2679      0.051     -5.238      0.000      -0.369      -0.167
crime                  -0.0064      0.015     -0.417      0.677      -0.037       0.024
continued_claims        0.3237      0.033      9.831      0.000       0.259       0.389
quit                   -0.5801      0.051    -11.272      0.000      -0.682      -0.479
initial_claims         -0.0295      0.029     -1.019      0.309      -0.087       0.028
labor participation    -0.7162      0.064    -11.207      0.000      -0.842      -0.590
==============================================================================
Omnibus:                        0.996   Durbin-Watson:                   2.304
Prob(Omnibus):                  0.608   Jarque-Bera (JB):                0.837
Skew:                          -0.149   Prob(JB):                        0.658
Kurtosis:                       3.044   Cond. No.                         13.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Train RMSE: 0.21859248073714296
Test RMSE: 0.3088854422696523
```

We can see that there is a strong negative relation between labor participation and unemployment and also total number of labor quits

Our Mean Squared Errors are significantly reduced

Model Diagnostics

Multiple Linear Regression (With Scaling)



Regression Model Results

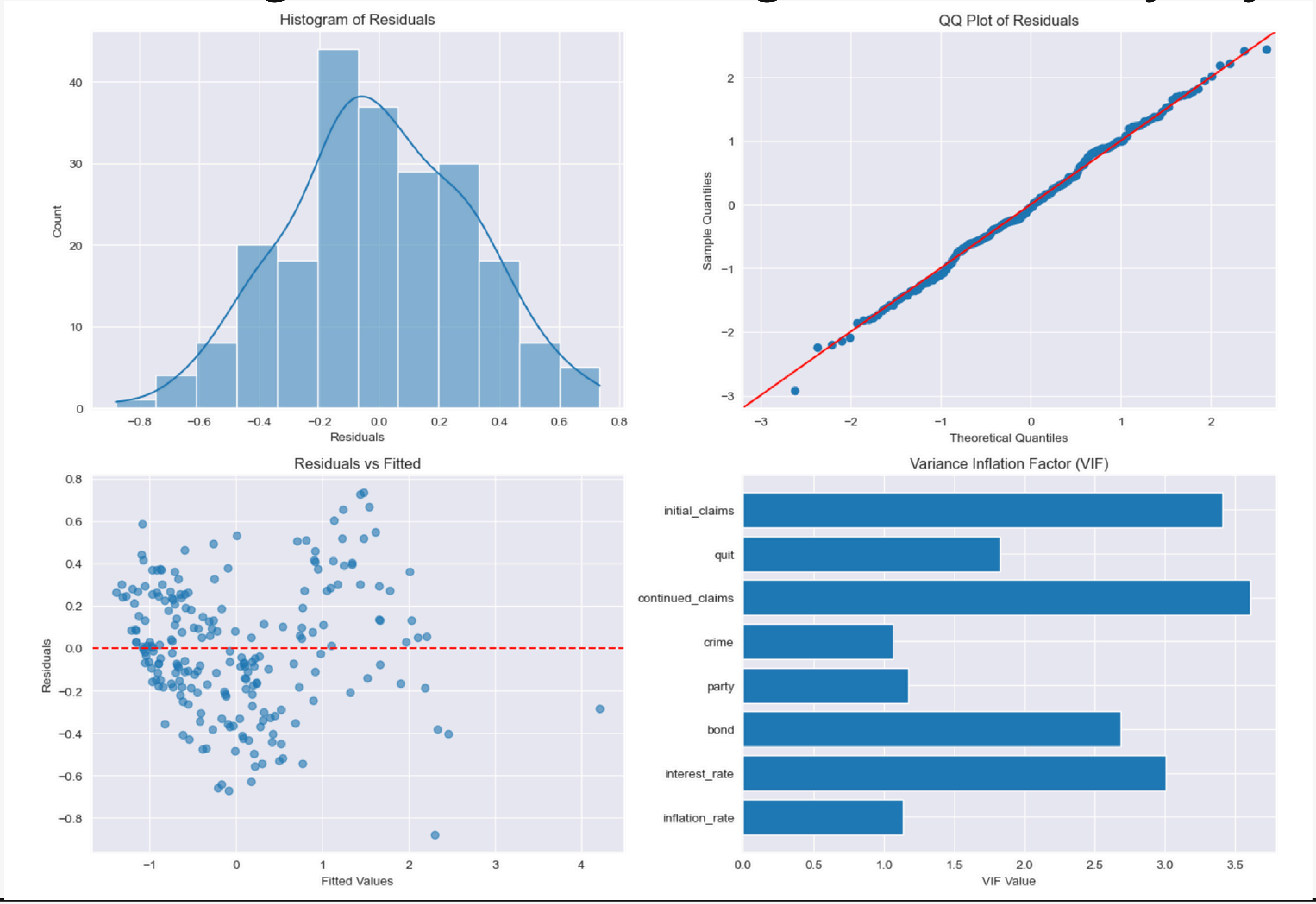
Multiple Linear Regression (With Scaling and Collinearity Adjusted)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.948
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                     346.0
Date:                Tue, 02 Dec 2025   Prob (F-statistic):          5.67e-128
Time:                        13:34:06   Log-Likelihood:                 12.559
No. Observations:                 222   AIC:                            -1.119
Df Residuals:                     210   BIC:                             39.71
Df Model:                          11
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.1542      0.033      4.707      0.000       0.090       0.219
job_postings            0.2653      0.079      3.378      0.001       0.110       0.420
inflation_rate          0.0188      0.017      1.073      0.284      -0.016       0.053
interest_rate          -0.0226      0.035     -0.655      0.513      -0.091       0.046
bond                   -0.1256      0.035     -3.581      0.000      -0.195      -0.056
sp500                   0.2324      0.055      4.191      0.000       0.123       0.342
party                  -0.2977      0.055     -5.371      0.000      -0.407      -0.188
crime                  -0.0218      0.016     -1.325      0.187      -0.054       0.011
continued_claims        0.3254      0.034      9.561      0.000       0.258       0.393
quit                   -0.5825      0.053    -10.983      0.000      -0.687      -0.478
initial_claims         -0.0362      0.029     -1.234      0.219      -0.094       0.022
labor participation    -0.6813      0.068    -10.068      0.000      -0.815      -0.548
==============================================================================
Omnibus:                        2.231   Durbin-Watson:                   2.062
Prob(Omnibus):                  0.328   Jarque-Bera (JB):                1.864
Skew:                          -0.191   Prob(JB):                        0.394
Kurtosis:                       3.234   Cond. No.                         13.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Train RMSE: 0.30034234743688865
Test RMSE: 0.41860694635325
```

The relationship between predictors and responses do not change too radical but our MSE increases

Model Diagnostics

Multiple Linear Regression (With Scaling and Collinearity Adjusted)



Regression Model Results

Ridge Regression

	columns	coefficient
0	job_postings	0.237053
1	inflation_rate	0.015957
2	interest_rate	-0.034675
3	bond	-0.129258
4	sp500	0.290275
5	party	-0.271628
6	crime	-0.006404
7	continued_claims	0.324969
8	quit	-0.575488
9	initial_claims	-0.028760
10	labor participation	-0.708199

```
=== Cross-Validation Metrics ===
CV RMSE : 0.2455
CV R²   : 0.9339

=== Train/Test RMSE ===
Train RMSE : 0.4675
Test RMSE  : 0.5559
```

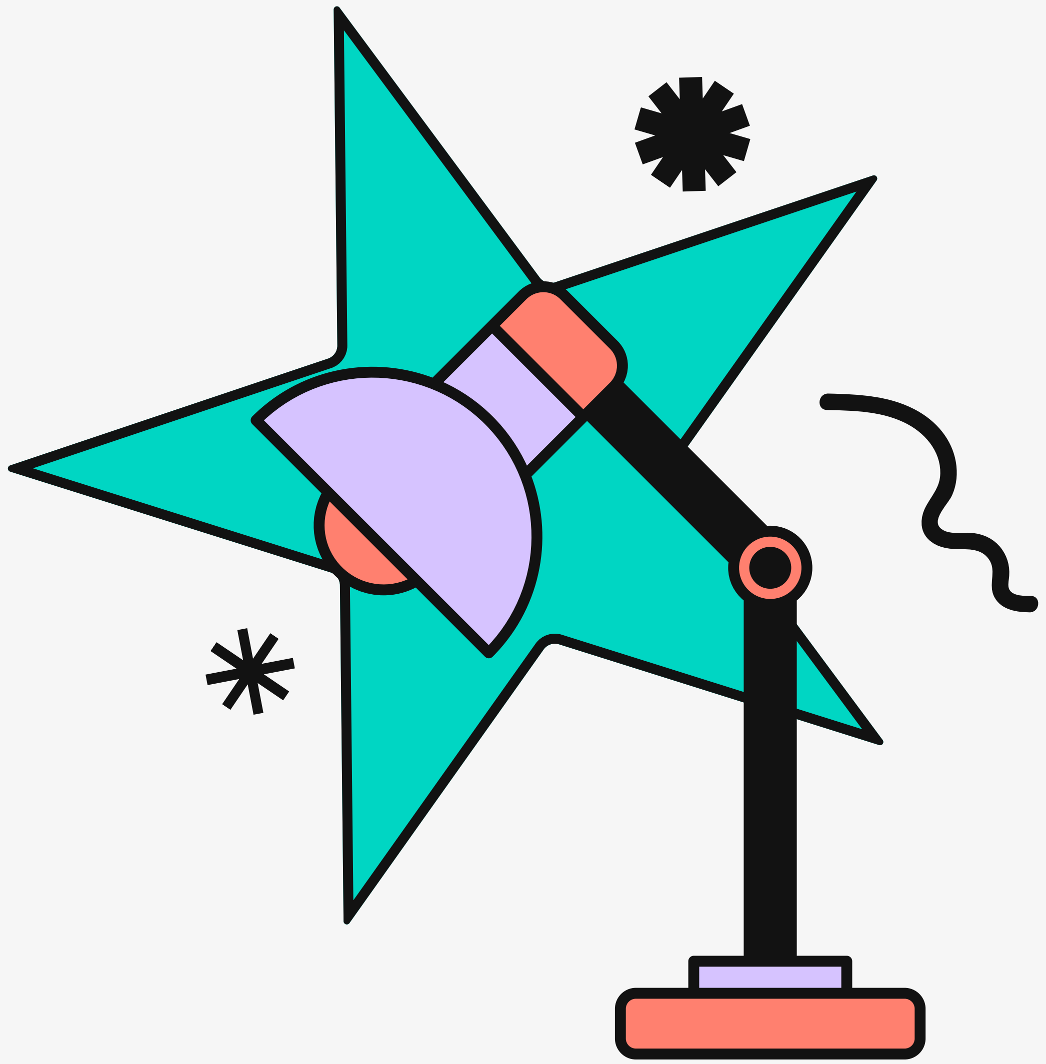
Lasso Regression

	columns	coefficient
0	job_postings	0.220325
1	inflation_rate	0.016650
2	interest_rate	-0.037950
3	bond	-0.126145
4	sp500	0.284944
5	party	-0.279390
6	crime	-0.005559
7	continued_claims	0.324048
8	quit	-0.564672
9	initial_claims	-0.023662
10	labor participation	-0.692688

```
=== Cross-Validation Metrics ===
CV RMSE : 0.2497
CV R²   : 0.9322

=== Train/Test RMSE ===
Train RMSE : 0.4677
Test RMSE  : 0.5537
```

Classification

Predicts discrete categories / classes
(e.g. spam/not spam, gender, etc.)



Logistic Regression

Predicting a binary response using multiple predictors

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Estimated coefficients are chosen to maximize the likelihood function rather than minimizing sum of squared residuals;

$$\ell(\beta_0, \beta_1, \cdots, \beta_p) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=1} (1 - p(x'_i))$$

Linear Discriminant Analysis and Naive Bayes

Based on different assumptions about our datasets and Bayes' Theorem

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{k} \pi_l f_l(x)}$$

LDA

Assuming that predictors are normally distributed

$$f(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Naive Bayes

Assuming that within the k class, the p predictors are independent

$$f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$$

Linear Support Vector Classifier

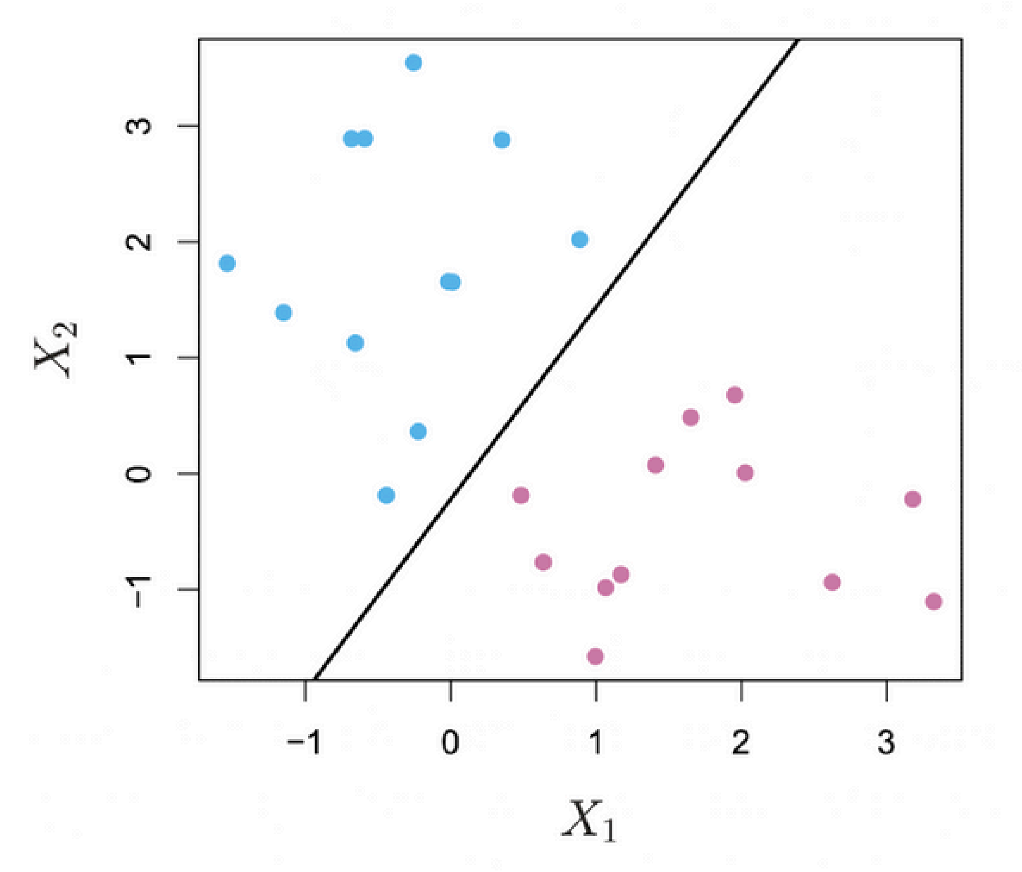
The main idea is that we want to fit a hyperplane seperating classes

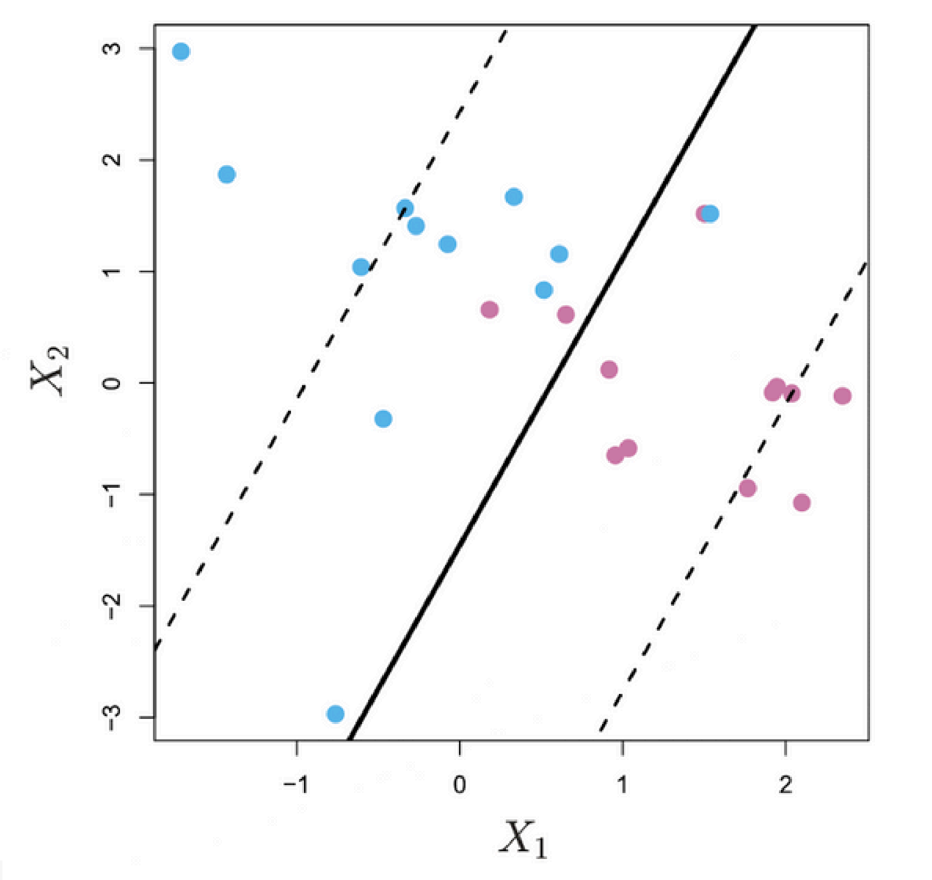
$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,\,M}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$





In which M is the minimal distance between any points and the decision boundary

C is our total budget for errors (ϵi) of how the point violates our margin

Support Vector Machine

An extension from the support vector classifier that enlarge the feature spaces using kernel

Kernel

Generalization of the inner product between prediction and actual

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j},$$

Give us the same linear support vector classifier

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle,$$

S: support vectors - data points closest to the decision boundary

Radial kernel

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

Classification Model Results

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

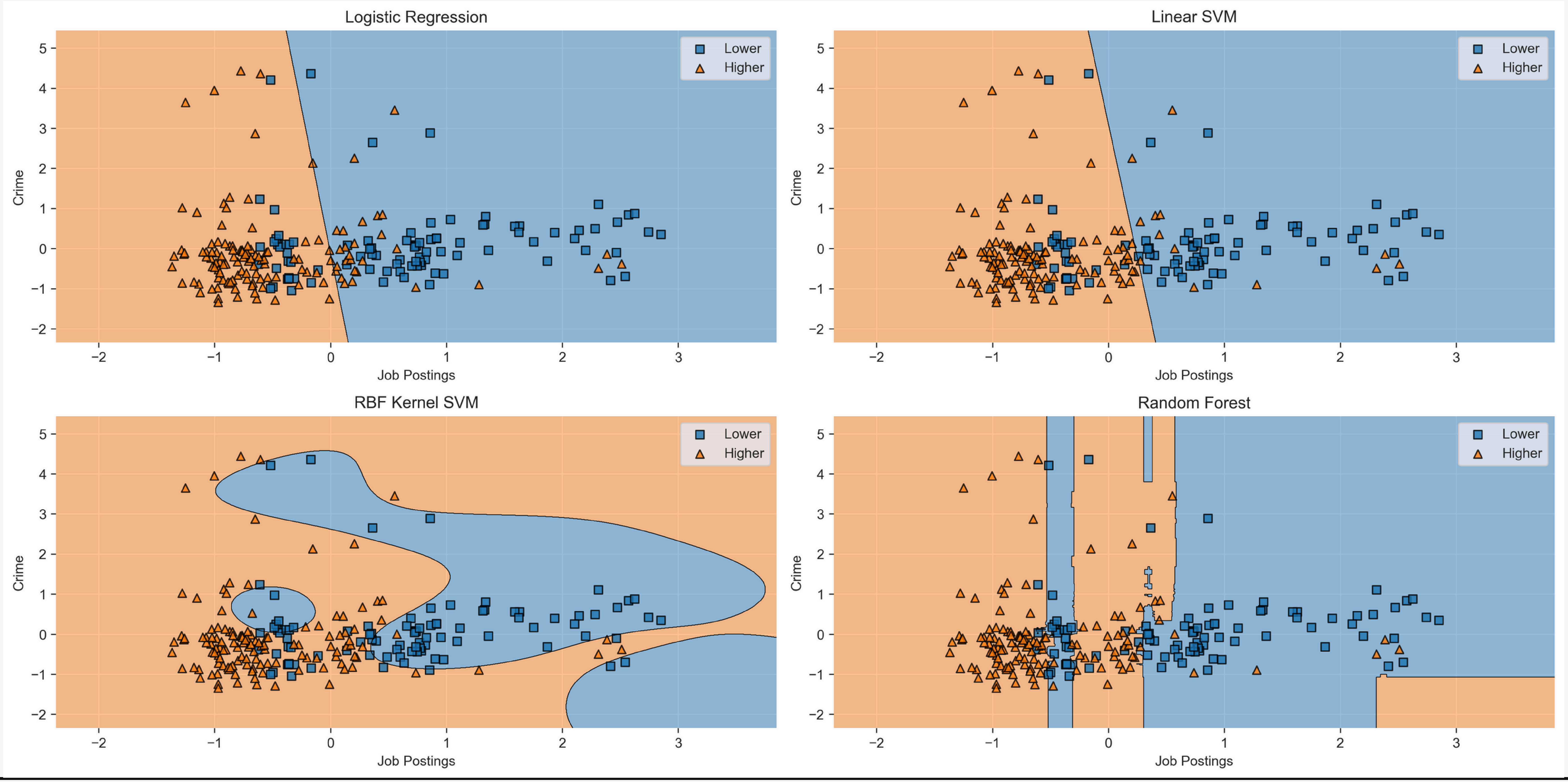
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

	Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC	TN	FP	FN	TP
0	Logistics Regression	0.981982	0.986667	0.977273	1.000000	0.988506	0.984375	31	1	0	43
1	Linear Discriminant Analysis	0.977477	0.933333	0.952381	0.930233	0.941176	0.933866	30	2	3	40
2	Naive Bayes	0.914414	0.906667	0.973684	0.860465	0.913580	0.914608	31	1	6	37
3	Linear SVM	0.981982	0.973333	0.955556	1.000000	0.977273	0.968750	30	2	0	43
4	RBF SVM	0.986486	0.973333	0.955556	1.000000	0.977273	0.968750	30	2	0	43

Disclaimer: This is totally based on the data that I have on hand so the overfitting might be a issue for predicting future data

Visualization of our Classifier



Current shortcomings

- Heteroskedasticity in MLR, can try WLS to fix
- Limited Datasets so overfitting might be a problem
- Data is currently manually pulled from FRED through CSV format then extract into the notebook
- Predictors and Responeses are highly time dependent
- Can fixed this with consider Time Series models but that might not be working as well

Remarks

- Correlation does not result in direct causation
- We would prefer a simpler models when they are yielding comparable results
- Simpler Models are more interpretable than complex ones
- Sometimes, complex models does not yield better results

Thank you

Feel free to check out my Github repo for your reference (QR code below)

Special thank to Patrick
for all his help this quarter