Analysis of Inflation Dynamics

Ying Gao(yg2804), Qinhui Kong(qk2126), Peixuan Song(ps3193), Wenhao Wang(ww2605),

Rong Xu(rx2180), Jingyi Ye(jy3179)

Columbia University

STATGR5291: Advanced Data Analysis

Professor David Rios

May 06, 2022

Part One: Introduction

Inflation is a dynamic process that describes how quickly prices increase over time (Oner, 2022). High inflation rate would reduce purchasing power, which means people will spend more on their daily expenditures. An excessive increase in the inflation rate will also hurt the economic system.

Inflation is one of the most essential macroeconomic indicators to track. For example, the U.S. central bank treats reducing inflation as the primary goal of monetary policy. Moreover, inflation attracts attention because of its detrimental influence on people's well-being. It can also have a significant impact on consumer psychology and purchasing behavior (Zhang, 2013).

Inflation can be influenced by a variety of factors, including external inputs. Increasing price of goods in the market, such as oil and food, and the U.S. federal funding release would lead to inflationary pressure. COVID-19 caused an increase in the unemployment rate and in order to help support people's normal life and combat the pandemic, the government decided to provide funds. The larger amount of money in the market circulation increased the inflation rate. Due to pandemic-related supply disruptions and increased consumer demand, inflation in the United States is at its highest level in 40 years (Guilford, 2022).

Since the inflation rate is closely related to people's living qualities, and many factors may interact with it, this project mainly focuses on observing the relationship between these features and the inflation rate in the United States. Also, time series analysis can be used to explore the existing trend and predict the direction in which the inflation rate is changing over time.

In order to analyze inflation dynamics, linear regression and time series analysis are applied in this project. The report begins with data preprocessing, followed by model fitting

including linear regression, principal component analysis, and ARIMA for analyzing and forecasting.

Part Two: Exploratory Data Analysis

2.1 Data Description

In this part, we apply the linear regression model to the dataset to find the relationship between the inflation rate and other metrics, and figure out the main influencing factors to elaborate the regression model. The dataset contains 23 economic indicators such as GDP, CPI, inflation rate, unemployment rate, retail money funds, etc. from January 2019 to December 2021 in the United States. However, those factors vary in periods. After a preliminary analysis, 7 factors measured on a monthly basis are considered in our model:

- 1. Unemployment Rate: The number of unemployed people as a percentage of the labor force.
- 2. Federal Funds (rate): Excess reserves that commercial banks and other financial institutions deposit at regional Federal Reserve banks.
- 3. Retail Money Funds (billion): Value of certain financial assets held by households, businesses, nonprofit organizations, and state and local governments.
- 4. Durable Goods (million): Total number of consumer goods that have a long life span and are used over time. For example, cars, home appliances, consumer electronics, and furniture.
- 5. 15-Year Fixed-Rate Mortgage: The interest rate of mortgages with the same monthly repayment throughout 15 years. For example, a home loan option with a specific interest rate for the entire term of the loan for 15 years in a total payback period.

- 6. Retail Sales (million): Amount of spending in sales of final goods by businesses to end consumers, such as durable (cars, furniture, etc.) and perishable (groceries, foods, etc.) goods.
- 7. Consumer Sentiment: The index aids in measuring how optimistic consumers feel about personal finances and the state of the economy.

2.2 Data Preprocessing

The inflation rate is a daily based data excluding weekends from 2019 to 2021 in our dataset. To check the consistency of our data in time scale and time unit, the monthly inflation rate is calculated by the monthly CPI using the following equation (1):

$$\text{Inflation (\%)} = \left(\frac{\text{Current CPI} - \text{Initial CPI}}{\text{Current CPI}} \right) \times 100 \quad (1)$$

The features selected are spread out in different datasets. To merge them, we convert them to perform a monthly average. Therefore, our final dataset contains seven features that would be used in future analysis and the inflation rate as the target (Figure 1).

Figure 1

Sample of cleaned data

retail_funds	retail_sales	federal_funds_rate	unemployment_rate	durable_goods	mortgage_fixed_rate	consumer_sentiment	inflation_rate
970.9	565690	0.08	4.2	270039	2.35	67.4	0.699229907
973	562296	0.08	4.6	261728	2.3075	71.7	0.866476548
976.1	552617	0.08	4.7	261353	2.184	72.8	0.410850556
977.9	546946	0.09	5.2	262317	2.145	70.3	0.33359786
981.8	543355	0.1	5.4	258846	2.18	81.2	0.453580853
988.6	550638	0.08	5.9	257663	2.27	85.5	0.877143995
997.8	547116	0.06	5.8	255529	2.28	82.9	0.701841208

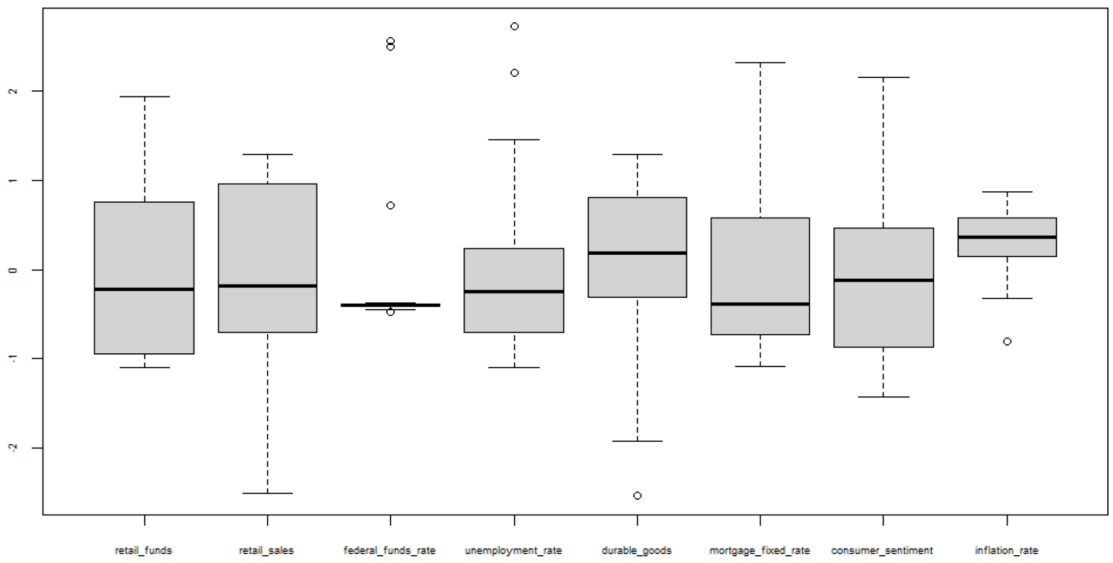
2.3 Normality and Outliers

The distribution of features can be presented in multiple boxplots, which provides statistical details of the data for choosing models. Figure 2 indicates that retail funds and

inflation rates are approximately following the normal distribution and have no more than one outlier. The durable goods and retail sales are slightly skewed to the left. The unemployment rate and federal funds rate have several outliers and are right-skewed. So are the consumer sentiment and mortgage fixed rate.

Figure 2

Boxplot



To further check normality, the Q-Q plot and Shapiro–Wilk test are applied. From the Q-Q plot (Figure 3 and Figure 4), the retail sales and inflation rates are distributed normally.

Figure 3

QQ Plot for the inflation rate

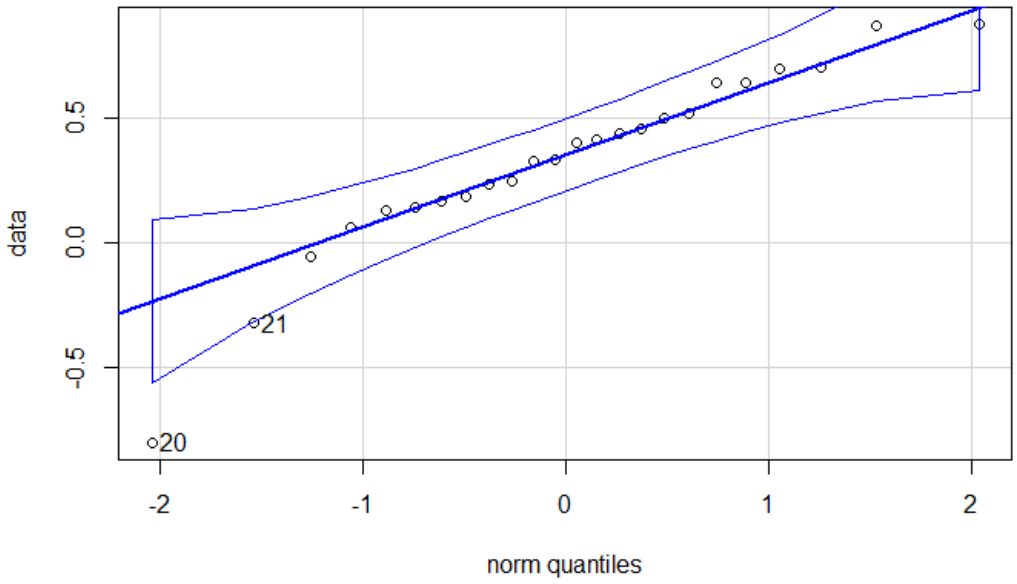
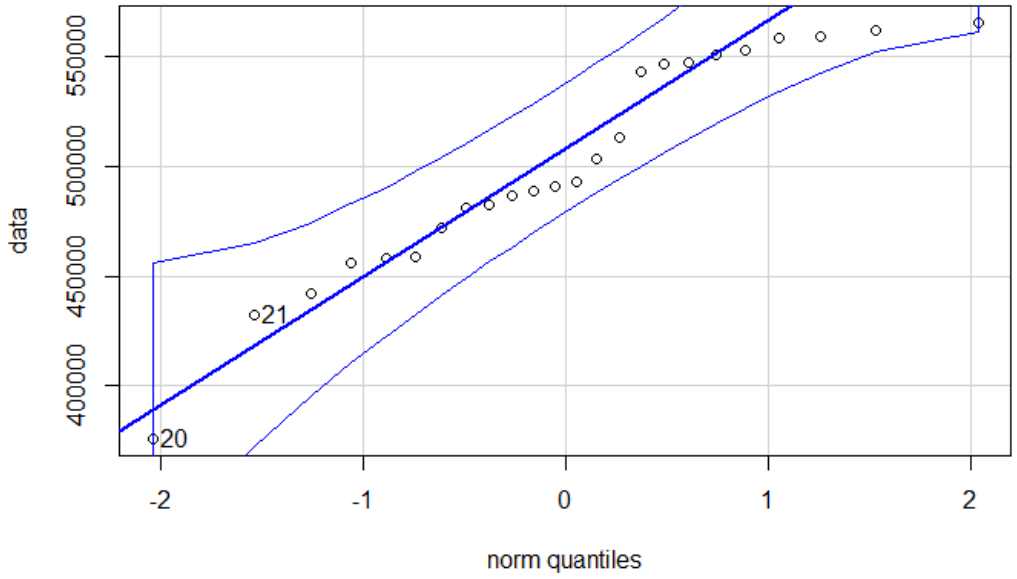


Figure 4

QQ Plot for retail sales



Shapiro–Wilk test (Figure 5) also indicates the same results as the previous method does with the p-value for each feature. If the p-value > 0.05, then we cannot reject the null hypothesis that the data is normal enough.

Figure 5

Shapiro–Wilk test results and p-value

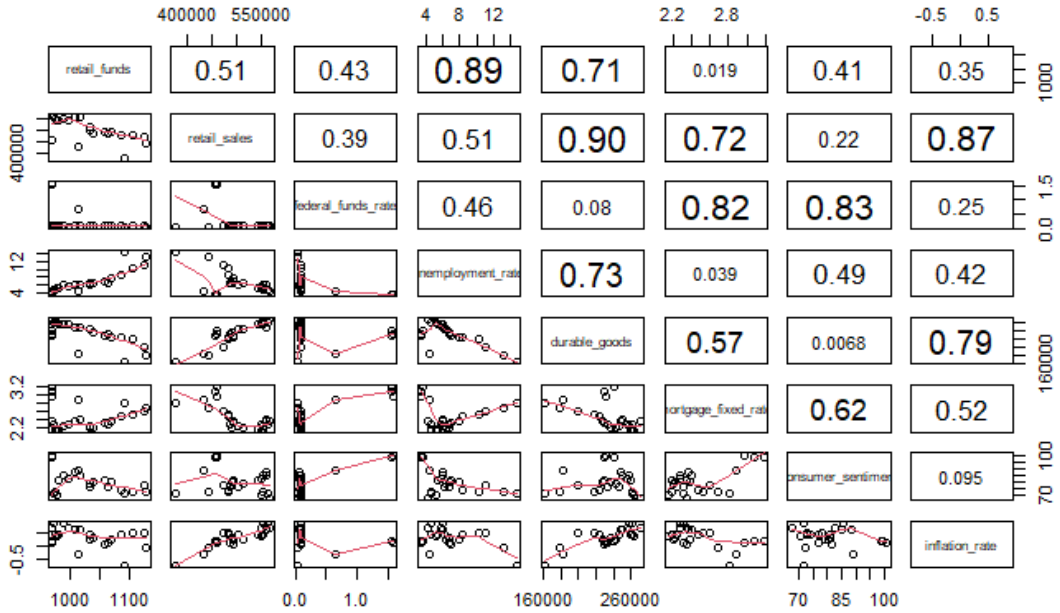
	data	p_value	shapiro_test_result
1	retail_funds	2.053948e-02	not normal
2	retail_sales	8.138535e-02	normal
3	federal_funds_rate	3.408561e-08	not normal
4	unemployment_rate	2.601010e-03	not normal
5	durable_goods	1.700541e-02	not normal
6	mortgage_fixed_rate	3.315925e-03	not normal
7	consumer_sentiment	3.714887e-02	not normal
8	inflation_rate	7.699671e-02	normal

2.4 Relationship between variables

Figure 6 shows the correlation among variables. The inflation rate has greater correlations with retail sales, durable goods, and mortgage fixed rates. Some predictors have a high correlation between each other. For example, the inflation rate is relatively highly correlated with the durable goods and retail sales and their correlations are 0.79 and 0.87 respectively. In addition, the correlation between durable goods and retail sales equals 0.90, which indicates that multicollinearity may exist.

Figure 6

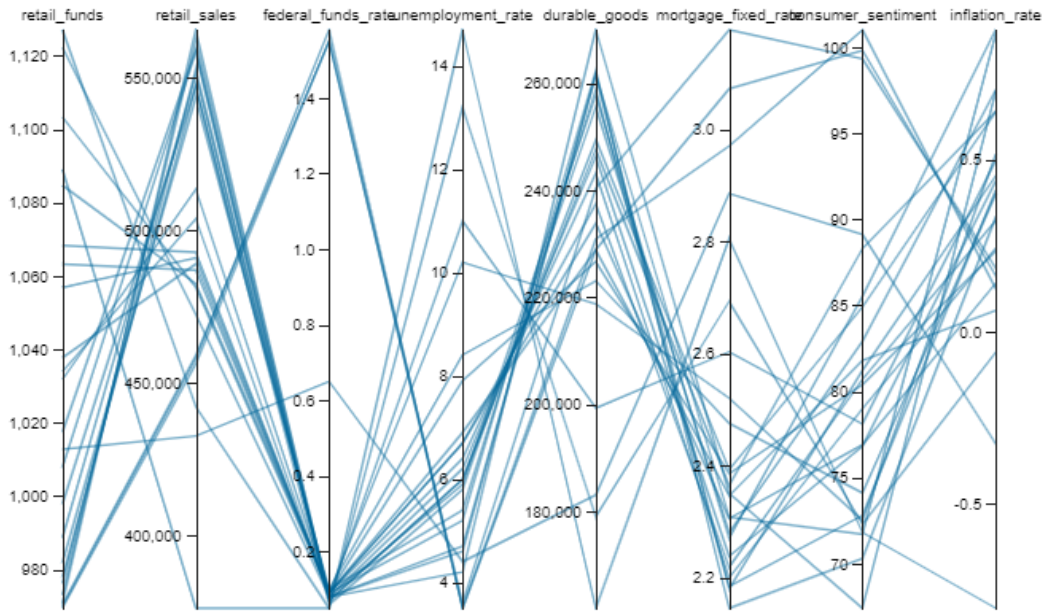
Correlation Matrix



An interactive parallel coordinate plot (Figure 7) is created to observe the relationship between each feature and the target. If the lines within two variables are parallel, then there is a positive association between them. On the contrary, the lines are twisted, they are more likely to have negative relationships. For example, the lines between the retail funds and retail sales are twisted, which indicates that there is a negative relationship between them. On the other hand, the retail funds decreased when the retail sales increased. By moving the vertical axis of the inflation rate, the plot shows that it is positively correlated with the retail sales, and durable goods while having a negative association with the unemployment rate, mortgage fixed rate, retail funds, and consumer sentiments.

Figure 7

Parallel coordinate plot



2.5 Linear Regression Model

Apply the linear regression model to the data. Use all 7 factors as explanatory variables and inflation rate as a response variable. However, after performing the ANOVA test, all the predictors except intercept and retail funds have a p-value>0.05, which indicates that the linear model is not the optimal model for analyzing relationships between inflation rate and 7 predictors. There may exist a non-linear relationship between those variables.

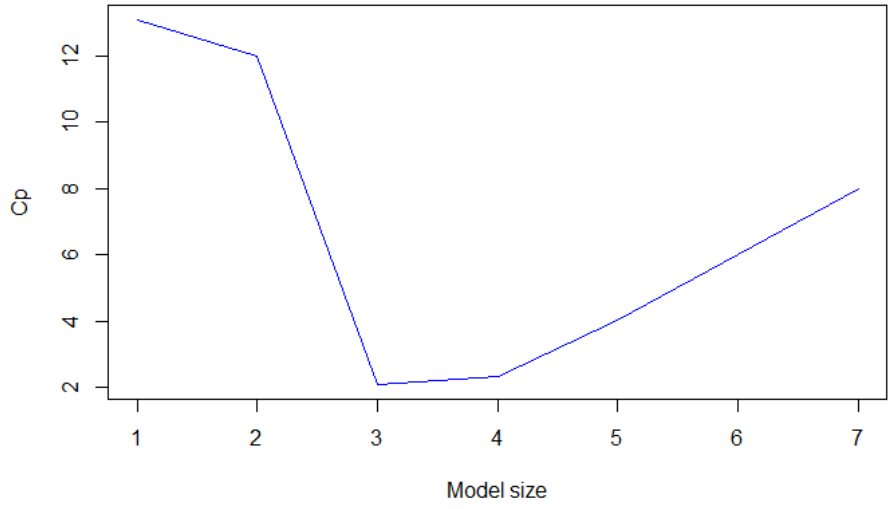
We further developed our model using a power transformation to create transformed data, in which the optimal powers (lambda) equal to -4, 1, 0, -1, 8, -2, 1 for retail funds, retail sales, federal funds rate, unemployment rate, durable goods, mortgage fixed rate, consumer sentiment respectively. However, after applying the ANOVA test to transformed data, p-values for all variables except the retail sales are larger than 0.05, which indicates that only the retail sales affect the inflation rate. Since we find collinearity among the variables in the correlation matrix in part 2.4, it is possible that collinearity highly influences the results of power transformation.

To eliminate collinearity between features and establish a linear model, we use exhaustive, forward selection, and backward elimination to discard redundant factors.

After performing the exhaustive searching, we find that Mallow's Cp has the lowest value at 3. A low Cp value indicates that the model is fairly accurate (Wikimedia Foundation, 2022), which suggests that the best linear regression model has three prediction variables.

Figure 8

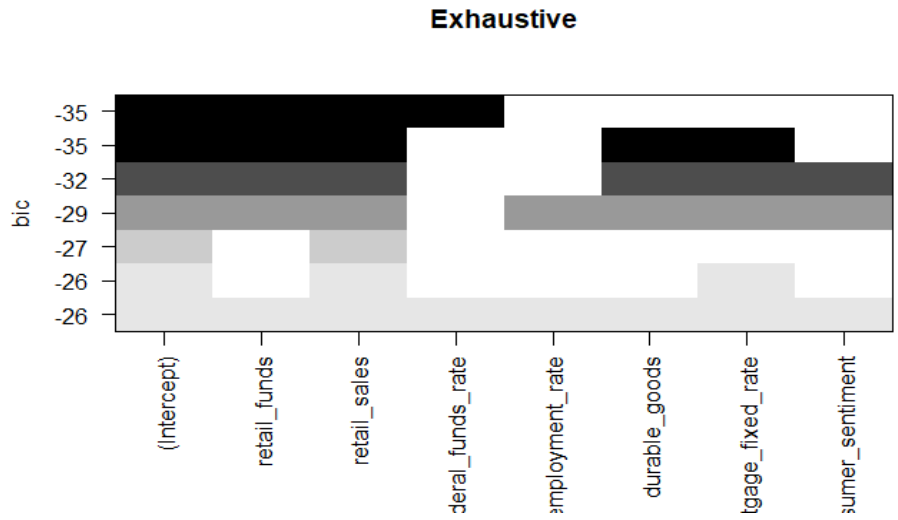
Cp metric



To find out factors that have the lowest BIC values, we visualize the BIC values for all 7 factors. In Figure 9, the intercept, retail funds, retail sales, and mortgage fixed rates have the lowest BIC values. The BIC plot shows that it is enough to include the above three factors as predictors in the linear model. However, the BIC method tends to choose models that are too simple for the smaller and less representative datasets (Brownlee, 2019). Therefore, AIC is used to help select a best-fitting model.

Figure 9

BIC values



AIC statistic penalizes complex models less, so it will put more emphasis on model performance on the dataset and select more complex models. Forward selection and backward elimination are used to find models that have the lowest AIC. Both methods give the same result with the lowest AIC = -88.67 including four factors, retail funds, retail sales, durable goods, and mortgage fixed rates, which are also the same as the exhaustive method. Thus, we will consider retail funds, retail sales, durable goods, and mortgage fixed rates as predictors in the linear regression model.

Fitting the linear regression model with the above four variables, we find that all the p-values are smaller than 0.05, and we are confident enough to conclude that retail funds, retail sales, durable goods, and mortgage fixed rates have effects on inflation rate. Our final linear model becomes:

$$\textit{Inflation rate} = -11.57 + 0.004326 * \textit{retail funds} + 0.0000077 * \textit{retail sales} + 0.0000081 * \textit{durable goods} + 0.69 * \textit{mortgage fixed rate} \quad (2)$$

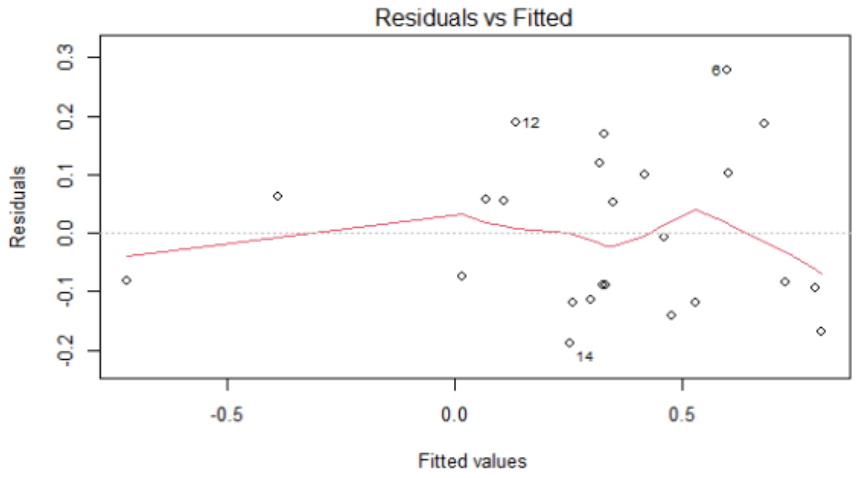
2.6 Check Assumptions

Linearity of the data (Residuals vs Fitted):

The line is approximately horizontal at zero and the residual plot shows no fitted pattern, which indicates that the linear model fits well (Figure 10).

Figure 10

Residuals vs Fitted

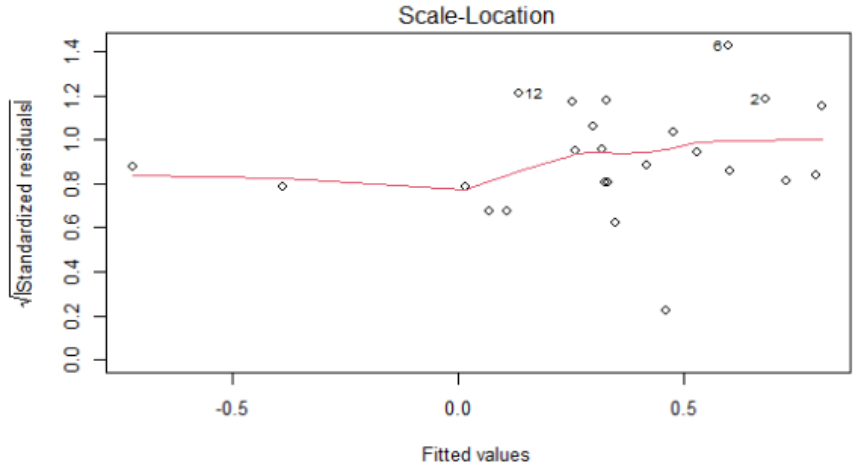


Homogeneity of variance (Scale-Location):

Though residuals are scattered slightly wider as the fitted value increases, there is a horizontal line with spread points in Figure 11, which suggests an equal variance.

Figure 11

Scale-Location

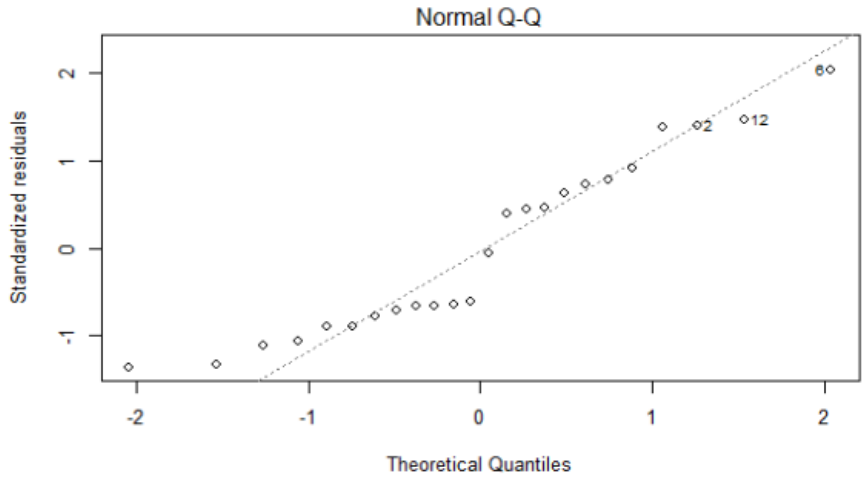


Normality of residuals (Normal Q-Q):

The points approximately follow the straight line and it presents that the residuals are normally distributed.

Figure 12

Normal Q-Q Plot



Outliers and high leverage points (Residuals vs Leverage):

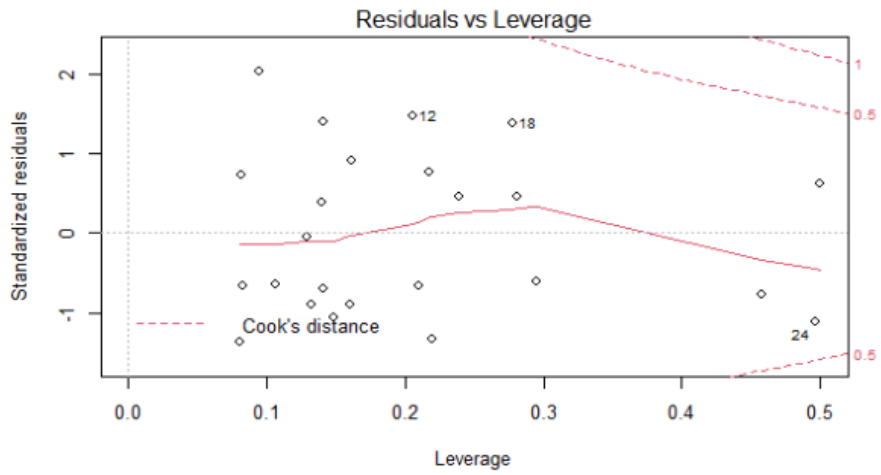
The Residuals vs Leverage plot (Figure 13) highlights the top three most extreme points, 12, 18, and 24. The standardized residuals of points 12 and 18 are around 1.5, and those of point

24 are slightly below -1. There are no outliers that differ by twice the standard deviation from other observations.

Moreover, there is no high leverage point in the data. All data points have a leverage statistic below $2(p + 1)/n = 10/120 = 0.083$.

Figure 13

Residuals vs Leverage



From the analysis above, the model we select (Equation (2)) satisfies all the assumptions associated with a linear regression model, including linearity between independent and dependent variables, independence, homoscedasticity, and normality of residuals.

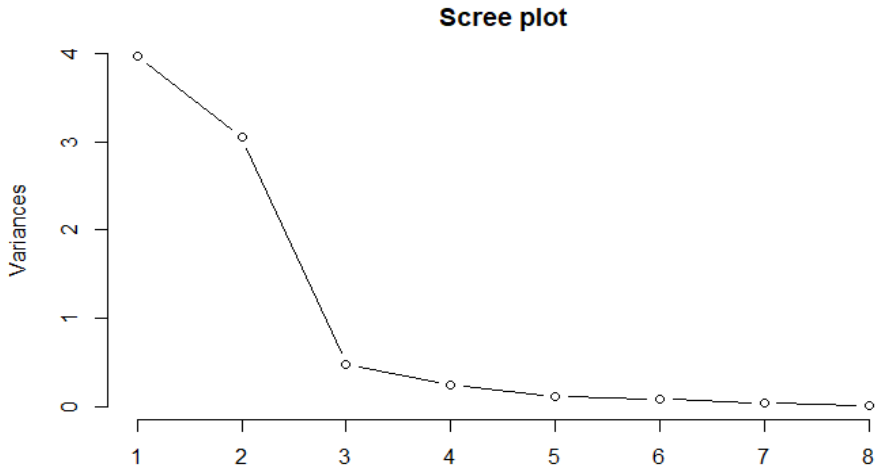
2.7 Principal Component Analysis

In an alternative approach to observing the correlation between features and the target, we looked at principal component analysis and its plots. PCA helps reduce the dimensionality of datasets and computes the principal components which contain most of the information of the data. If multicollinearity appears among the features that should have been gathered to analyze and predict the target, PCA could be used to solve the problem.

By using PCA, we find that the first three principal components capture 94% of the total variance in the dataset, which means using PC1, PC2, and PC3 would be enough to fit in a principal component regression model. Figure 14 suggests the same results that unexplained variance in the dataset decreases quickly when adding PC1, PC2, and PC3 while having a smoother slope after adding PC4. Thus, only three principal components can explain the majority of the variance in the dataset.

Figure 14

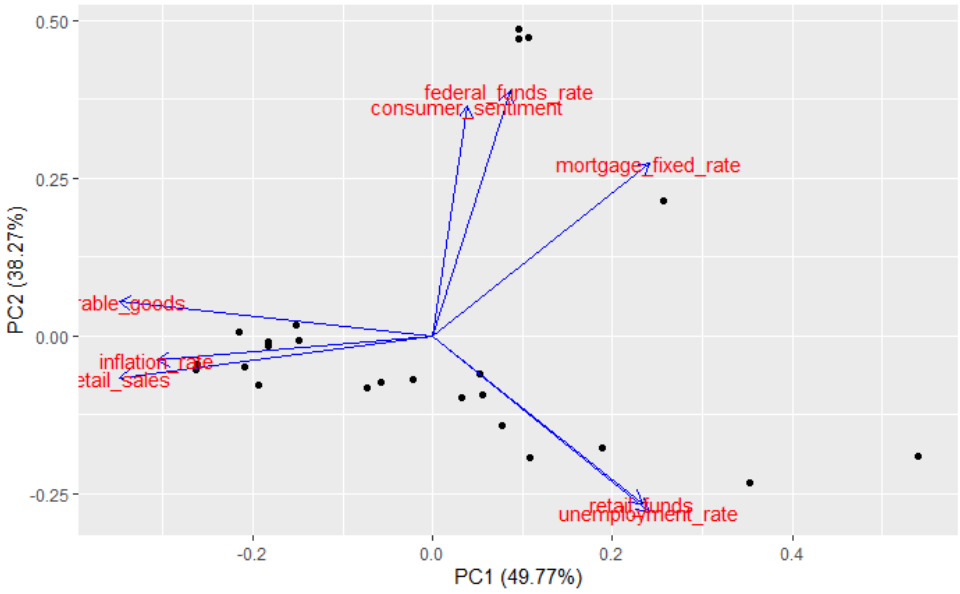
Scree Plot



Based on the biplot (Figure 15), if two vectors form an angle that is less than 90 degrees, they are positively correlated. If the angle is greater than 90 degrees, the corresponding features of these two vectors would have a negative relationship. For instance, the small-angle which is constructed by the retail sales and inflation rate indicates that the amount of money spent in the market increases when the inflation rate increases.

Figure 15

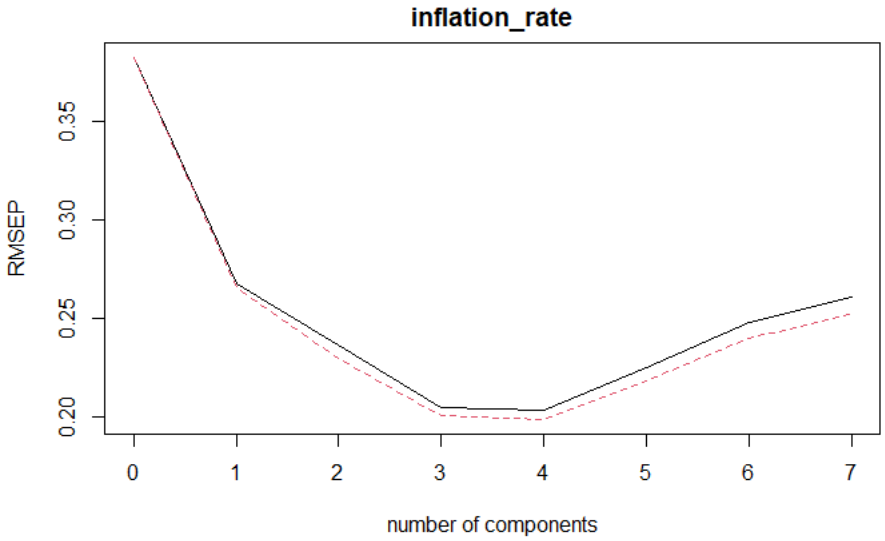
Biplot of PCA



The model performs well enough by adding 3 components, which is consistent with Figure 16 that the root means the square error of prediction increases after 4 principal components are included.

Figure 16

RMSEP Plot



We still want to look into the changes and advancements in inflation rate after having the linear regression model. Time series analysis in the following part helps us understand how inflation rate changes over time. Finally, a prediction is made after fitting the data into time series models.

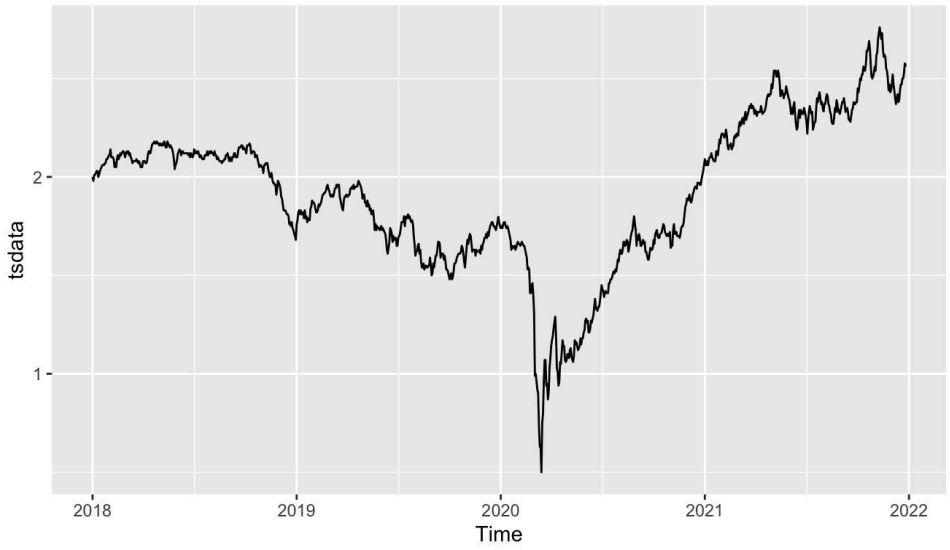
Part Three: Time Series Analysis

3.1 Introduction

In this part, we use time series to extract meaningful statistics and other characteristics of a dataset to understand it. While regression focuses more on the relationships between inflation rate and other metrics, time series analysis digs into the patterns of the inflation rate itself over time. In the first step, past observations are collected and analyzed to develop a suitable mathematical model, trying to capture the underlying data generating process for the series. Stationary tests and differential methods are used to find the lag. Also, we train different time series models to find the most suitable one. In the second step, the future events are predicted using the model.

Figure 17

Data Change Over Time As a Line Graph

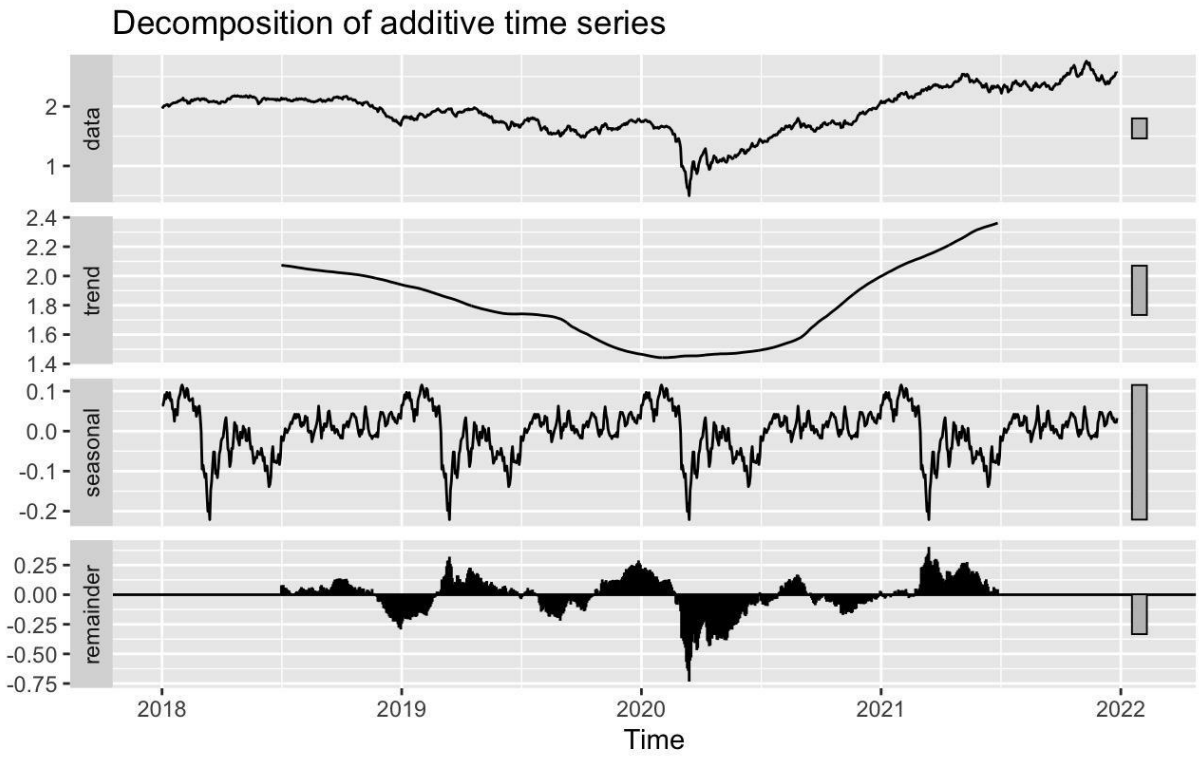


3.2 Decomposing

Time series data can exhibit a variety of patterns, and it is often helpful to split a time series into several components, each representing an underlying pattern category. Usually, three of them attract most of the attention: trend, seasonality, and cycles.

Figure 18

Decomposition of additive time series



These components can be added together to reconstruct the original data displayed in the top panel.

The trend component shows the effect of the COVID-19 pandemic on the inflation rate. The indentation around the beginning of 2020 matches the crisis of the virus, breaking the stable raising trend of inflation.

The seasonal component changes slowly over time, so any consecutive year has similar patterns, but years that are far apart may have different seasonal patterns. Usually, the inflation rate goes down at the beginning of each year and raises gradually to a steady level through the year.

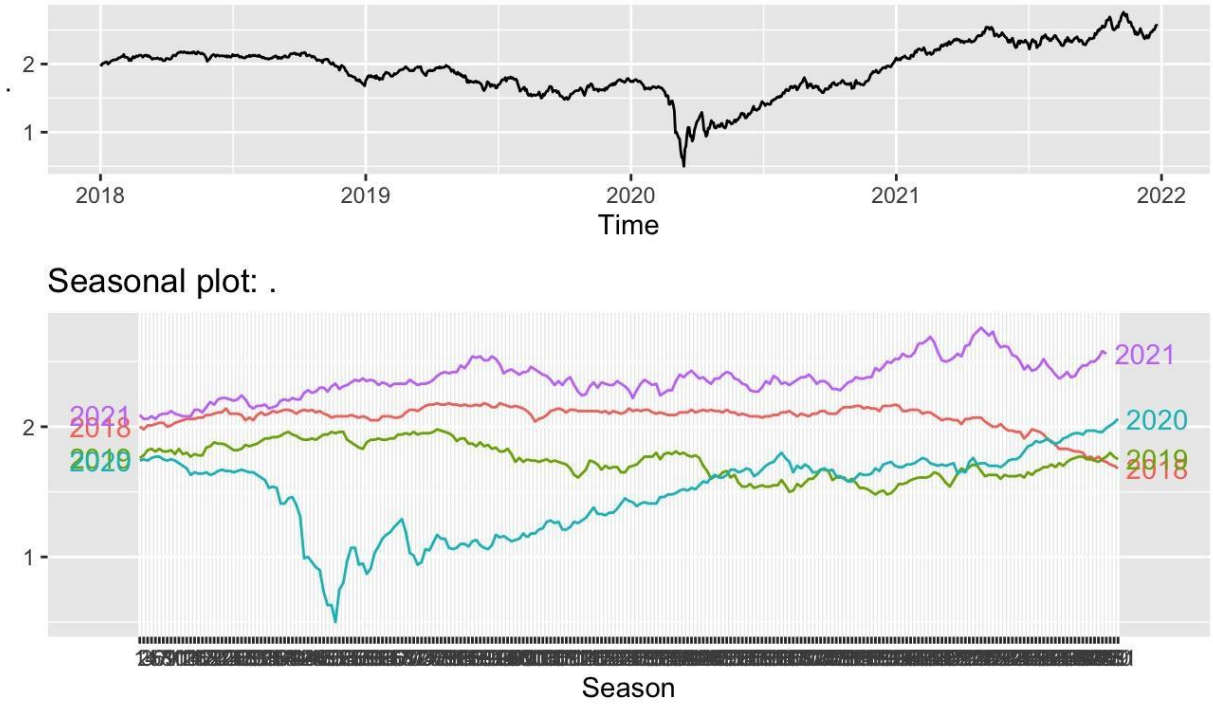
The remainder shown in the bottom panel is what is left after subtracting the seasonal and trend-cycle components from the data. The largest unexplainable remainder concentrates around

the beginning of 2020 as well, demonstrating that COVID-19 does decrease the inflation rate unexpectedly.

We further draw the combination of trend and seasonal components by year to exhibit the influence of the pandemic on inflation (Figure 19) and there isn't a similar or significant trend in the observed four years, so we still include the seasonal components in the following analysis without deducting it as noise.

Figure 19

Seasonal Plot



3.3 Stationarity and Differencing

When forecasting or predicting the future, we'd love to assume that each point in our models is independent of one another. This property makes sure that the patterns of the data do not depend on the time of the observed series. However, usually, most of the time-series data are

not qualified. To check the stationarity and remove the unexpected relations, we carry out the stationarity test and differencing process.

The original data set does not pass the stationary test. So, we differentiate the data to stabilize the mean of a time series by removing changes in the level of a time series, therefore reducing trend and seasonality. After differencing, we performed ADF test and KPSS test, the p-value of ADF tests is less than 0.01 and the p-value of KPSS test is larger than 0.1, which indicates the stationarity is satisfied.

We plot the Autocorrelation Function in Figure 20, which could explain the relationship between current and historical values. Since the ACF has a decreasing trend, the model applies to time series. And we tried a differencing process in Figure 21 to observe ACF and PACF. By removing fluctuations in the level of a time series and so eliminating (or lowering) trend and seasonality, differencing can assist in stabilizing the mean of the time series (Brownlee,2020). By subtracting the past observation from the current observation, differencing is produced (Brownlee, 2020).

Figure 20

Time Series ACF of Original Data

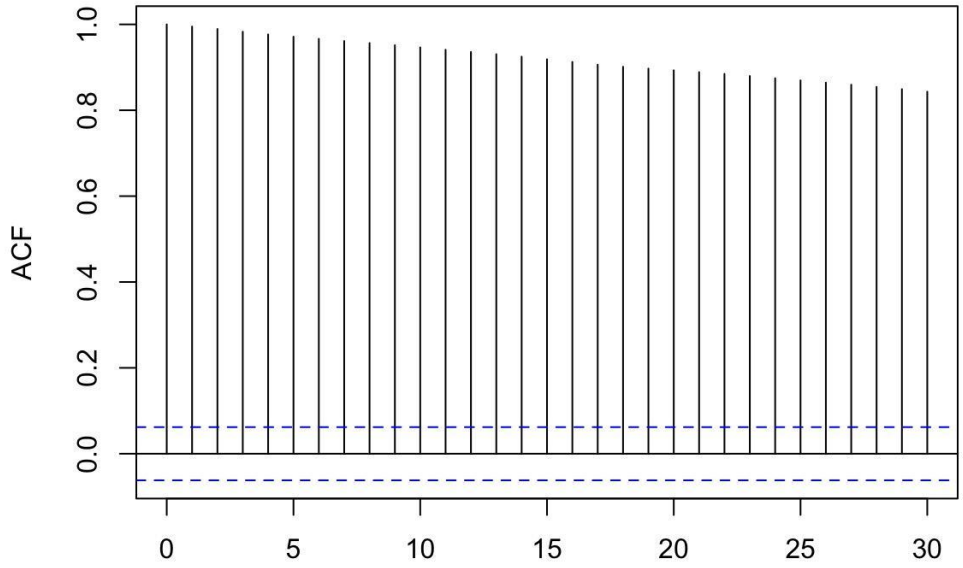
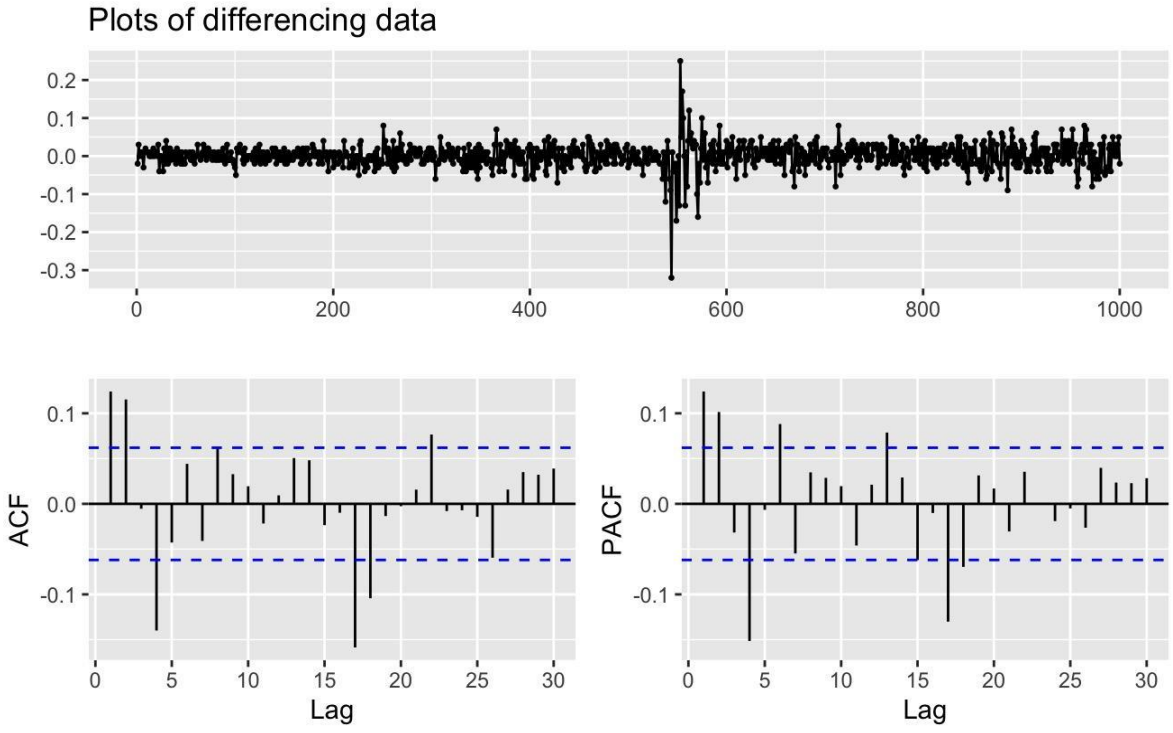


Figure 21

Plots of Differencing Data

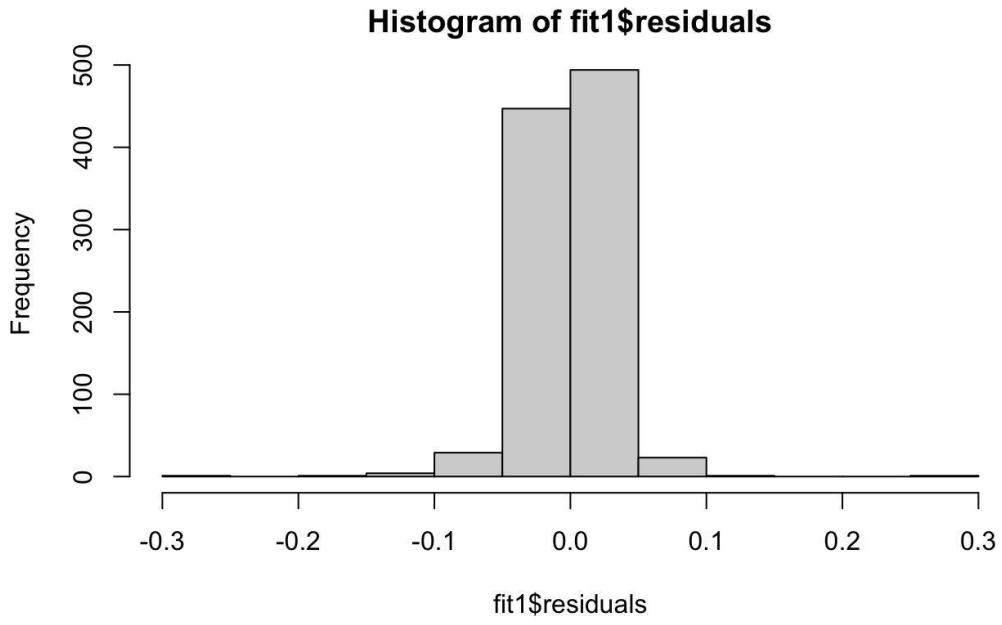


3.4 ARIMA Model

In Figure 21, both ACF and PACF plots are not clear in trailing off or cutting off. So we choose the ARIMA model as our general prediction model. Since the plot does not give out strong suggestions on choosing parameters, we carried out a searching function to pick the best model parameter set. According to the AIC criterion, the ARIMA(8,1,9) has the lowest AIC=-3948.179, which is outstanding.

Figure 22

Histogram Plot

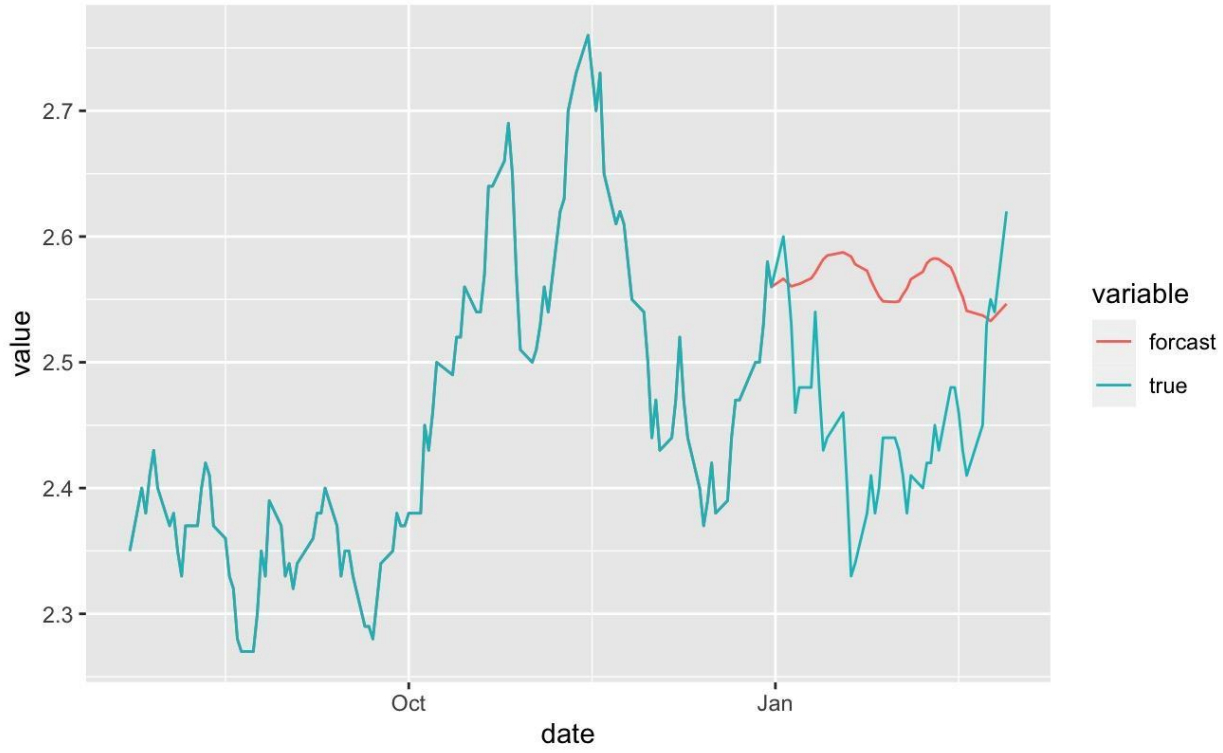


3.5 Prediction

Another important step in time series is to forecast the future. Using the ARIMA(8,1,9) model, we forecast 2 month's inflation rate and plot the prediction and the true value in the same plot (Figure 23).

Figure 23

Inflation rate forecast graph



Although it does not capture certain moves, the model successfully predicts the overall decreasing trend and the local seasonal fluctuation.

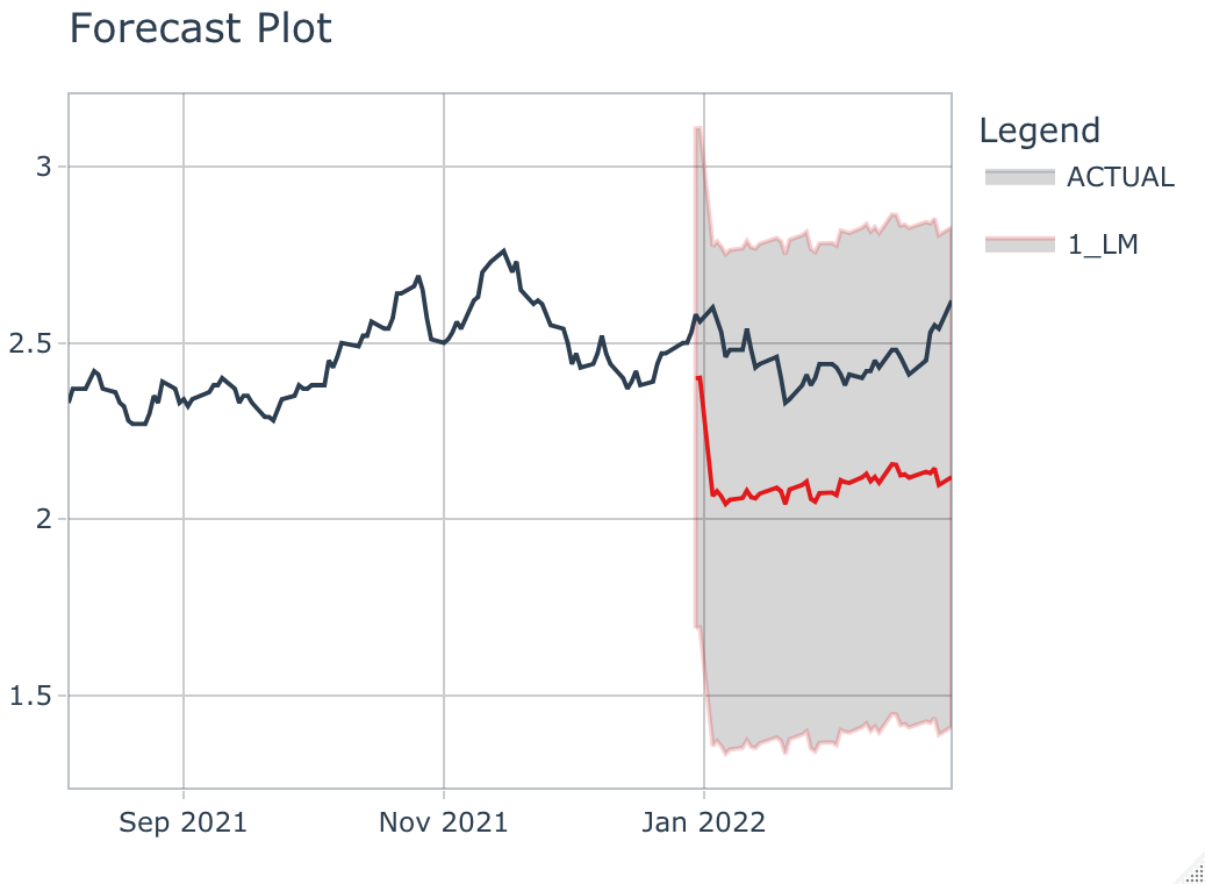
3.6 A Machine Learning Model

The machine learning algorithm is popular in time series analysis. However, it's usually time-consuming and expensive. There is a new package released this month named "timetk", a tool kit for working with Time Series in R. (https://github.com/business-science/timetk). This is a functional package focusing on building High-Performance Forecasting Systems. Since our data set is not too large to carry out the time machine learning algorithm, we apply the package to our dataset to see whether we have better options than the ARIMA model. All the data from 2018-01-01 to 2022-01-01 are used as the train set. This machine learning model projects the

original two columns data set into a wider data set for the model to learn. There are also many basis regression models to choose from. In our training, the regular linear regression model is used.

Figure 24

Forecast Plot



Although this time, the model captures the fluctuation trend more precisely, the value of the output is much lower than expected. Since there are so many parameters that can be tuned and various column process methods can be used, this may be caused by the unsuitable parameter and preprocessing methods choices. Also, a more complicated basis regression model may lead to a better prediction. Since the computation source is limited, we cannot grid search for the best parameters.

This can be our next step to dig more into the time series machine learning model that has better performance.

Part Four: Conclusion

In this project, we visualized and tested data to explore the correlation between variables and the targets in order to support the model and feature selection. After first fitting a multivariate linear regression model with all the variables, the ANOVA test and power transformation are performed to improve the model. Using forward selection and backward elimination to remove multicollinearity among features, we find that retail funds, retail sales, durable goods, and mortgage fixed rate have linear relationships with the inflation rate. The ANOVA test using the above four factors as explanatory variables and the inflation rate as the response variable suggest that the inflation rate increases as retail funds, retail sales, durable goods, and mortgage fixed rate increase.

We utilize time series to investigate the pattern of inflation across time, starting with decomposing the inflation rate into trends, seasonality, and cycles. Finally, we select the best subgroup and plot the 50-days ahead forecast of inflation rate using the ARIMA model and machine learning. Our final result successfully forecasts the general trending and local fluctuations of the inflation rate, and the prediction interval covers the actual inflation value.

Overall, the linear regression model reveals the relationship between the inflation rate and economic indicators that relate to people's daily lives. At the same time, the time series model helps us deeply understand the systematic patterns of the inflation rate over time. Additionally, knowing how the inflation rate would change in the future guides people to adjust their investment and manage their assets. However, the randomness of the inflation rate and the

limit of predictors make it difficult for us to forecast certain moves precisely. The results show that the forecast accuracy still needs to be further improved. More complex and nonlinear models such as random forest models will be considered in the future. A large dataset can also be used to support what we find in this project.

References

Brownlee, J. (2020, June 22). *How to remove trends and seasonality with a difference transform in Python*. Machine Learning Mastery. Retrieved May 6, 2022, from https://machinelearningmastery.com/remove-trends-seasonality-difference-transform-python/#:~:text=Differencing%20is%20performed%20by%20subtracting%20the%20previous%20observation%20from%20the%20current%20observation.&text=Inverting%20the%20process%20is%20required,step%20to%20the%20difference%20value.

Brownlee, Jason. (2019, October 30). *Probabilistic Model Selection with AIC, BIC, and MDL.* Machine Learning Mastery. Retrieved May 5, 2020, from https://machinelearningmastery.com/probabilistic-model-selection-measures/

Gao, P., Jing, L., Shi, D., Wen, G., Zhang, Q., & Zhang, Q. (2013). Promote economic development and curb inflation. In *Macroeconomy, Inflation and Price Reform: Proceedings of the International Symposium on Macroeconomy and Price Reform* (pp. 746–759). essay, Social Sciences Academic Press (China).

Guilford, G. (2022, February 11). *U.S. inflation rate accelerates to a 40-year high of 7.5%*. The Wall Street Journal. Retrieved May 6, 2022, from https://www.wsj.com/articles/us-inflation-consumer-price-index-january-2022-11644452274

Reaser, A. L. (2013). China's Role in the New Economic Reality. In P. Gao (Ed.), *Macroeconomy, Inflation and Price Reform: Proceedings of the International Symposium on Macroeconomy and Price Reform* (pp. 248–275). essay, Social Sciences Academic Press (China).

Oner, C. (n.d.). *F&D article*. Inflation: Prices on the Rise. Retrieved May 2, 2022, from https://www.imf.org/external/pubs/ft/fandd/basics/30-inflation.htm#:~:text=Inflation%20is%20the%20rate%20of,of%20living%20in%20a%20country.

Wikimedia Foundation. (2022, January 30). *Mallows's CP*. Wikipedia. Retrieved May 6, 2022, from https://en.wikipedia.org/wiki/Mallows%27s_Cp#:~:text=It%20is%20applied%20in%20the,the%20model%20is%20relatively%2