1. Research Objective

Food expenditure is a crucial indicator of household budgets and has a significant impact on the economy. The trend of food prices is a matter of great concern to the public, as evidenced by the recent 49% increase in egg prices, which has affected people's daily lives. To shed light on the factors influencing food expenditure, this project aims to identify and analyze key variables such as GDP, CPI, and S&P 500. By using these variables as explanatory factors, our project aims to provide insights that can help people predict future food prices and enable the government to make informed decisions on policies related to residents' food expenditure. Our ultimate goal is to help ensure that appropriate measures are taken to address any challenges that may arise in the food market and to promote sustainable and affordable access to food for all.

2. Data Proprocessing

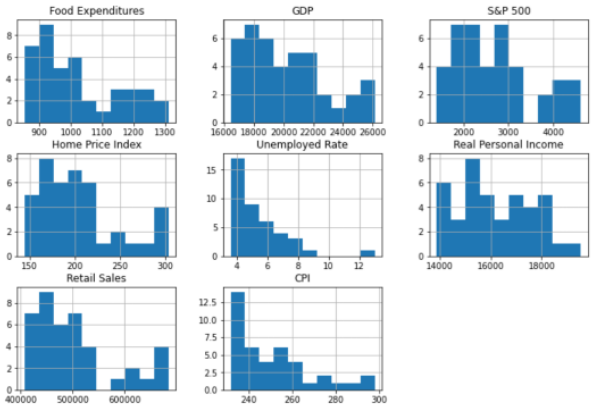
This project takes *Personal consumption expenditures: Food* from Federal Reserve Bank as dependent variable, and seven macroeconomic variables as independent variables. We take quarterly data from October 2012 to October 2022. The data analysis job recruitment data set “project.csv” consists 40 rows of data and 8 variables.

Variable	Explanatory	Source
Food Expenditures	Billions of Dollars,Seasonally Adjusted Annual Rate	U.S. Bureau of Economic Analysis
GDP	Billions of Dollars,Seasonally Adjusted Annual Rate	U.S. Bureau of Economic Analysis
S&P 500	Real Time Price. Currency in USD	Yahoo Finance
Home Price Index	Index Jan 2000=100, Seasonally Adjusted	S&P Dow Jones Indices LLC
Unemployed Rate	Percent, Seasonally Adjusted	U.S. Bureau of Labor Statistics
Real Personal Income	Billions of Chained 2012 Dollars, Seasonally Adjusted Annual Rate	U.S. Bureau of Economic Analysis
Retail Sales	Millions of Dollars, Seasonally Adjusted	U.S. Census Bureau
CPI	Measures the quarterly change in prices paid by U.S. consumers	U.S. Bureau of Labor Statistics

First, we input missing values with mean, and do the descriptive analysis to have a general understanding of our dataset.

```
DATE                    0
Food Expenditures       0
GDP                     0
S&P 500                 0
Home Price Index        0
Unemployed Rate         0
Real Personal Income    0
Retail Sales            0
CPI                     0
dtype: int64
       Food Expenditures           GDP      S&P 500  Home Price Index  \
count          41.00000     41.000000    41.000000         41.000000
mean         1019.28439   20159.171463  2700.867805        204.158911
std           133.69554    2648.926849   895.806103         43.966488
min           852.90000   16420.390000  1412.160000        144.564333
25%           917.37000   18193.710000  2018.050000        170.881000
50%           989.60000   19636.730000  2575.260000        195.969000
75%          1127.70000   21538.030000  3225.520000        217.263333
max          1310.57000   26144.960000  4605.380000        303.489667

       Unemployed Rate  Real Personal Income   Retail Sales         CPI
count        41.000000             41.000000      41.000000   41.000000
mean          5.326829          16188.739512  503686.179024  251.007878
std           1.865077           1463.827630   81462.264523   17.604223
min           3.600000          13882.260000  407313.000000  231.638000
25%           3.900000          15163.610000  445353.670000  237.498000
50%           4.900000          16051.050000  477474.670000  246.626000
75%           6.200000          17621.190000  522398.670000  258.278000
max          13.000000          19539.380000  685023.330000  297.987000
```



3. Correlation

To gain a better understanding of the relationship between each variable, we calculated the correlation matrix. The results indicated that all variables, with the exception of the Unemployment Rate, exhibited a high positive correlation with Food Expenditures. Moreover, all independent variables, except for the Unemployment Rate, also showed a strong positive correlation with each other. These findings align with common sense and suggest that the chosen variables are relevant to the analysis and likely to have an impact on food expenditures. This result also suggests that our model has strong multicollinearity.

	Food Expenditures	DATE	GDP	S&P 500
Food Expenditures	1.000000	0.959797	0.967718	0.967858
DATE	0.959797	1.000000	0.962011	0.950977
GDP	0.967718	0.962011	1.000000	0.960462
S&P 500	0.967858	0.950977	0.960462	1.000000
Home Price Index	0.984239	0.944131	0.983817	0.964754
Unemployed Rate	-0.224710	-0.321067	-0.442434	-0.314800
Real Personal Income	0.912266	0.956119	0.868517	0.900390
Retail Sales	0.968568	0.919483	0.981261	0.962242
CPI	0.973996	0.927074	0.981258	0.934073

	Home Price Index	Unemployed Rate	Real Personal Income
Food Expenditures	0.984239	-0.224710	0.912266
DATE	0.944131	-0.321067	0.956119
GDP	0.983817	-0.442434	0.868517
S&P 500	0.964754	-0.314800	0.900390
Home Price Index	1.000000	-0.334121	0.859585
Unemployed Rate	-0.334121	1.000000	-0.134606
Real Personal Income	0.859585	-0.134606	1.000000
Retail Sales	0.989643	-0.378948	0.832237
CPI	0.987794	-0.338865	0.829793

	Retail Sales	CPI
Food Expenditures	0.968568	0.973996
DATE	0.919483	0.927074
GDP	0.981261	0.981258
S&P 500	0.962242	0.934073
Home Price Index	0.989643	0.987794
Unemployed Rate	-0.378948	-0.338865
Real Personal Income	0.832237	0.829793
Retail Sales	1.000000	0.978597
CPI	0.978597	1.000000

4. Regression

To better understand how each predictor variable affects Food Expenditures, we conducted a linear regression analysis. Based on the t-value and p-value of the regression coefficients, we found that the variables S&P 500, Home Price Index, Real Personal Income, and CPI had a significant effect on Food Expenditures at the 10% level.

- S&P 500 is positively associated with Food Expenditures. When S&P 500 index increase one unit, the Food Expenditures will increase 0.0270 billion USD.
- Home Price Index is positively related to Food Expenditures. When Home Price Index increases one unit, the Food Expenditures will increase 0.8567 billion USD.
- Increases in the Real Personal Income is reflected by corresponding increases in Food Expenditures. When Real Personal Income increases one USD, the Food Expenditures will increase 0.0197 USD.

- CPI is positively associated with Food Expenditures. When CPI increases 1%, the Food Expenditures will increase 0.035 billion USD.

However, we also observed a high R-squared value of 0.994, meaning that explain 99.4% of the variation in the dependent variable, which indicates that there may be strong multicollinearity in the model. The VIF also gives the same result. To address this issue, we plan to implement Ridge Regression and Principal Component Analysis (PCA) methods to decrease the multicollinearity in our regression analysis.

OLS Regression Results

Dep. Variable:	Food Expenditures	R-squared:	0.994
Model:	OLS	Adj. R-squared:	0.993
Method:	Least Squares	F-statistic:	777.0
Date:	Sat, 25 Mar 2023	Prob (F-statistic):	1.00e-34
Time:	22:04:09	Log-Likelihood:	-153.62
No. Observations:	41	AIC:	323.2
Df Residuals:	33	BIC:	336.9
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-352.4315	170.300	-2.069	0.046	-698.909	-5.955
GDP	-0.0074	0.013	-0.580	0.566	-0.033	0.019
S&P 500	0.0270	0.012	2.241	0.032	0.002	0.052
Home Price Index	0.8567	0.465	1.844	0.074	-0.088	1.802
Unemployed Rate	4.8120	2.955	1.628	0.113	-1.200	10.824
Real Personal Income	0.0197	0.005	3.599	0.001	0.009	0.031
Retail Sales	8.234e-05	0.000	0.435	0.666	-0.000	0.000
CPI	3.5342	1.355	2.608	0.014	0.777	6.292

Omnibus:	18.592	Durbin-Watson:	1.825
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.187
Skew:	1.261	Prob(JB):	4.59e-07
Kurtosis:	6.275	Cond. No.	4.87e+07

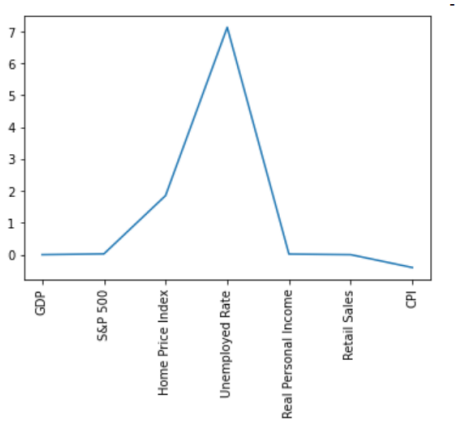
	variables	VIF
0	GDP	12197.642069
1	S&P 500	325.994025
2	Home Price Index	1689.876666
3	Unemployed Rate	57.352612
4	Real Personal Income	1411.197878
5	Retail Sales	2677.502748
6	CPI	5548.297449

4.1 Ridge Regression

Ridge Regression is a powerful method to address multicollinearity in a regression model by adding a penalty term to the regression equation, shrinking the coefficients towards zero. However, it's important to evaluate the performance of the model after implementing Ridge Regression.

In this case, we found that after implementing Ridge Regression, the Mean Square Error increased significantly from 104 to 649. This indicates that the model is not

performing better after using this method. Therefore, we need to consider other methods, such as Principal Component Analysis (PCA), to reduce multicollinearity in our regression analysis.



4.2 Principal Component Analysis (PCA)

PCA can eliminate the multicollinearity by using only the first few principal components as predictor variables, we can reduce the multicollinearity and improve the accuracy of the coefficient estimates.

As a result, we found that the variance ratio of GDP is 0.9999 and S&P 500 is 0.0001, so we input these two variables as principal components. Additionally, we include Unemployment Rate as an important variable in the Ridge Regression, so we also put it as an independent variable in the new regression and got the following conclusion:

- GDP is the most important factor that affects the Food Expenditures, which will positively affect Food Expenditures. When GDP increase one USD, the Food Expenditures will increase 0.0458 USD.
- S&P 500 positively affects Food Expenditures. When S&P 500 index increase one unit, the Food Expenditures will increase 0.0252 billion USD.
- Unemployment rate doesn't significantly affect Food Expenditures.

OLS Regression Results

Dep. Variable:	Food Expenditures	R-squared:	0.990
Model:	OLS	Adj. R-squared:	0.989
Method:	Least Squares	F-statistic:	1187.
Date:	Sat, 25 Mar 2023	Prob (F-statistic):	8.21e-37
Time:	22:10:10	Log-Likelihood:	-164.55
No. Observations:	41	AIC:	337.1
Df Residuals:	37	BIC:	344.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-59.6116	51.369	-1.160	0.253	-163.694	44.471
GDP	0.0458	0.004	12.849	0.000	0.039	0.053
S&P 500	0.0252	0.010	2.531	0.016	0.005	0.045
Unemployed Rate	16.4753	1.485	11.097	0.000	13.467	19.483

Omnibus:	3.459	Durbin-Watson:	1.252
Prob(Omnibus):	0.177	Jarque-Bera (JB):	2.513
Skew:	0.283	Prob(JB):	0.285
Kurtosis:	4.073	Cond. No.	4.79e+05

5. Conclusion

In the end, we still didn't completely eliminate the multicollinearity, we think it's because of these reasons:

- The predictor variables may be highly interrelated in a complex way that cannot be fully captured by simple linear relationships.
- The sample size of the data may be too small relative to the number of predictor variables, which can limit the ability of the analysis to fully capture the complex relationships among the variables.
- There may be unmeasured or unobserved variables that are related to the predictor variables and the outcome variable, which can create residual correlation among the variables even after using methods to reduce multicollinearity.