MGT6203 GROUP PROJECT FINAL REPORT

Impact of Key Macroeconomic Indicators on the US Retail and Food Services Industry

Team 66: Grace Suji Han, Somya Agarwal, Vu To Uyen Nguyen (Katherine), Jinhwan Kim

BACKGROUND INFORMATION

The retail and food services industry are an essential part of the US economy, representing a significant share of the country's GDP. In recent years, the industry has experienced significant growth, fueled by factors such as increasing disposable income, changing consumer preferences, and technological advancements. However, the COVID-19 pandemic has had a profound impact on the industry, with many businesses struggling to stay afloat amid shutdowns, restrictions, supply chain disruptions, and increasing inflation. According to a recent FT survey, economists predict a 68% chance of a recession in 2023, with 2% expected by year-end and 30% not until 2024.The outbreak has forced the global population to adopt to a new normal, following safety measures in public, purchasing goods and services online versus offline, and interacting with those around them in ways different from pre-COVID days.

LITERATURE REVIEW

The literature review we conducted has extensively explored the effects of macroeconomic indicators on the retail industry. Notably, the Issah & Antwi (2017) research paper describes the role of macroeconomic conditions in predicting the performance of firms in the U.K. The authors employ regression analysis to identify that real GDP, employment rate, CPI, financial markets, interest rates have a noteworthy impact on the retail industry. By utilizing the same methodology in this study, we would perform regression analysis on US data and draw valuable conclusions.
Moreover, Aziz N, He J, Raza A, Sui H, Yue W. (2021) examine the relationship between macroeconomic factors and undernourishment in South Asian Countries. The authors employ a regression analysis to highlight the close ties between the explanatory variables such as food production, economic growth, food prices and the response variable, prevalence of undernourishment to help policymakers develop comprehensive strategies to combat the underlying issue.

PROBLEM STATEMENT

Given the current economic climate and the challenges facing the US retail and food services industry, we hope to understand the impact of various macroeconomic indicators on its revenues. This study is particularly relevant given the ongoing COVID-19 pandemic and its far-reaching economic consequences. We want to explore the following:

- What are the key macroeconomic factors that impact retail and food services sales?
- Which explanatory variable has the largest impact on this industry?

By understanding the impact of macroeconomic indicators, businesses can make informed decisions that can help them adapt to the challenges they face. The study's results can also inform policymakers and other stakeholders, helping them develop strategies that can support the industry's growth and resilience.

Initial Hypothesis

We hypothesize that the real GDP, personal income, and consumer sentiment will have a significant positive correlation with retail and food sales, while unemployment will have a negative correlation. We base this hypothesis on the understanding that these macroeconomic indicators play crucial roles in shaping consumer behavior and determining the overall health of the economy.

Furthermore, we anticipate that the COVID-19 pandemic will have a significant impact on the retail and food services industry, with a sharp decline in sales during the pandemic year. However, we expect to see a gradual recovery in the post-pandemic years as the economy continues to stabilize.

We recognize that this hypothesis is subject to change as we analyze the data and delve deeper into the relationships between these macroeconomic indicators and the retail and food services industry. Nonetheless, we believe that this initial hypothesis provides a solid foundation for our project and will guide our analysis as we seek to gain a better understanding of the impact of macroeconomic indicators on the industry.

DATA SOURCES

For this project, we used two primary data sources to analyze the impact of macroeconomic indicators on the US retail and food services industry.

The first data source we used is the Monthly Retail Trade - Sales Report, which is published by the US Census Bureau. This report provides monthly retail sales data for various industries, including food services and drinking places, general merchandise stores, and clothing and clothing accessories stores, among others. This data set allows us to analyze trends and patterns in retail sales over time and investigate the impact of macroeconomic indicators on the industry.

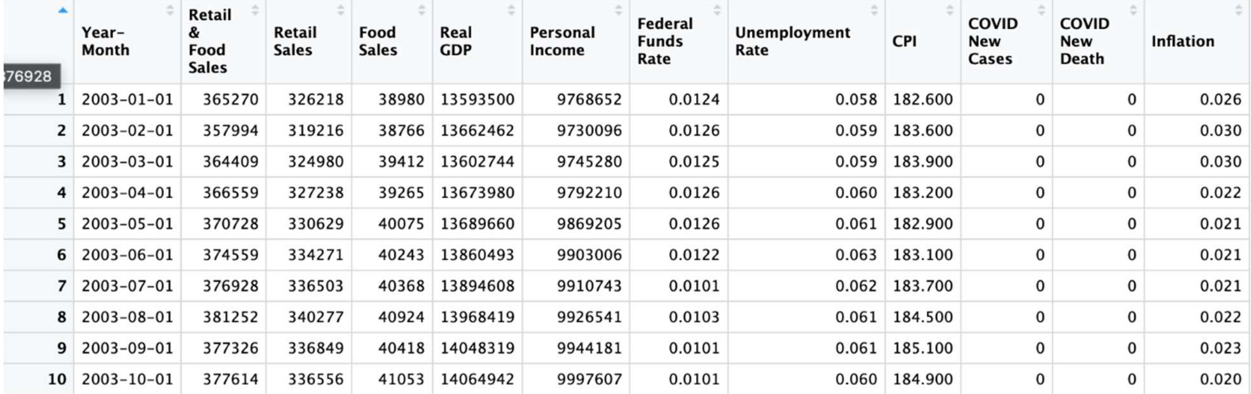
The second data source we used is the Personal Income and its Disposition report, which is also published by the US Census Bureau. This report provides data on personal income, disposable personal income, and personal consumption expenditures, among other variables. We used this data to examine the relationship between personal income and retail sales and investigate how changes in personal income can impact consumer behavior.

DATA CLEANING AND EXPLORATORY DATA ANALYSIS

The response variable is *Real BEA Retail and Food Services Sales* and below are our 8 explanatory variables:

1. Real GDP
2. Personal Income and Its Disposition in Chained Dollars - *various sources of personal income such as wages, rental income, investment income, and social benefits net of taxes.*
3. Federal Funds Rate
4. CPI
5. Unemployment Rate
6. COVID New Cases
7. COVID New Death
8. Inflation Rate

To ensure the accuracy and validity of our analysis, we conducted a comprehensive data formatting and cleaning process. The process included collating the data from multiple CSV sources into a single file, followed by time-series analysis to identify any underlying trends and seasonality. The first 10 rows of our dataset look like below:

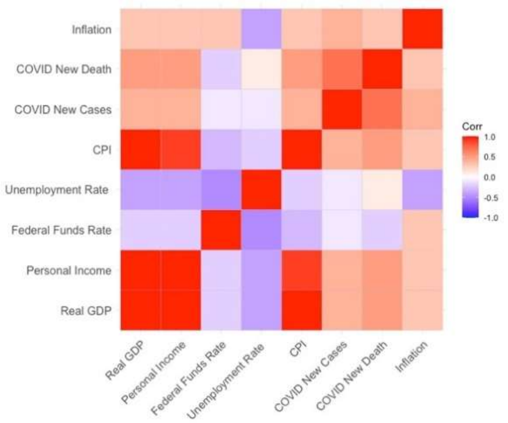


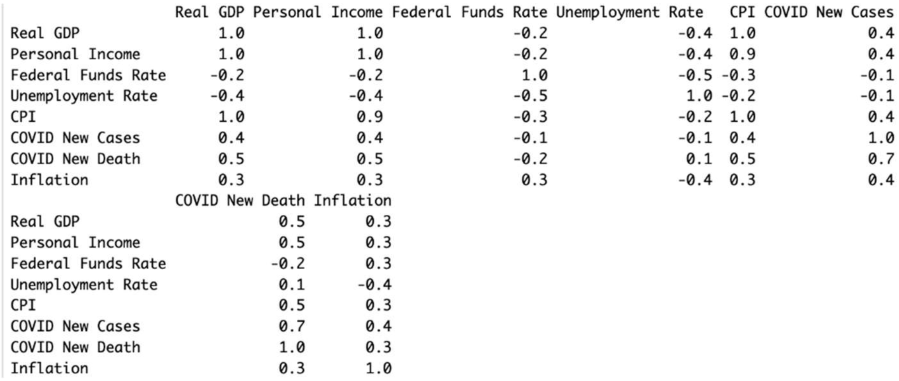
	Year–Month	Retail & Food Sales	Retail Sales	Food Sales	Real GDP	Personal Income	Federal Funds Rate	Unemployment Rate	CPI	COVID New Cases	COVID New Death	Inflation
1	2003-01-01	365270	326218	38980	13593500	9768652	0.0124	0.058	182.600	0	0	0.026
2	2003-02-01	357994	319216	38766	13662462	9730096	0.0126	0.059	183.600	0	0	0.030
3	2003-03-01	364409	324980	39412	13602744	9745280	0.0125	0.059	183.900	0	0	0.030
4	2003-04-01	366559	327238	39265	13673980	9792210	0.0126	0.060	183.200	0	0	0.022
5	2003-05-01	370728	330629	40075	13689660	9869205	0.0126	0.061	182.900	0	0	0.021
6	2003-06-01	374559	334271	40243	13860493	9903006	0.0122	0.063	183.100	0	0	0.021
7	2003-07-01	376928	336503	40368	13894608	9910743	0.0101	0.062	183.700	0	0	0.021
8	2003-08-01	381252	340277	40924	13968419	9926541	0.0103	0.061	184.500	0	0	0.022
9	2003-09-01	377326	336849	40418	14048319	9944181	0.0101	0.061	185.100	0	0	0.023
10	2003-10-01	377614	336556	41053	14064942	9997607	0.0101	0.060	184.900	0	0	0.020

FIGURE 1: Data Example

We then looked into the correlation matrix to gain a better understanding of the relationships between the variables and identify multicollinearity. This analysis enabled us to examine the correlation coefficients between the explanatory variables, as well as their correlation with the response variable. Our findings indicated that:

- Retail and Food Services Sales had a positive correlation with Real GDP, Personal Income and CPI, while showing a negative correlation with the Unemployment Rate.
- However, we did not observe any significant correlation between Retail and Food Services Sales and Federal Funds Rate, COVID new cases, COVID new deaths and Inflation.





```
                   Real GDP Personal Income Federal Funds Rate Unemployment Rate  CPI COVID New Cases
Real GDP                1.0             1.0               -0.2              -0.4  1.0             0.4
Personal Income         1.0             1.0               -0.2              -0.4  0.9             0.4
Federal Funds Rate     -0.2            -0.2                1.0              -0.5 -0.3            -0.1
Unemployment Rate      -0.4            -0.4               -0.5               1.0 -0.2            -0.1
CPI                     1.0             0.9               -0.3              -0.2  1.0             0.4
COVID New Cases         0.4             0.4               -0.1              -0.1  0.4             1.0
COVID New Death         0.5             0.5               -0.2               0.1  0.5             0.7
Inflation               0.3             0.3                0.3              -0.4  0.3             0.4
                   COVID New Death Inflation
Real GDP                       0.5       0.3
Personal Income                0.5       0.3
Federal Funds Rate            -0.2       0.3
Unemployment Rate              0.1      -0.4
CPI                            0.5       0.3
COVID New Cases                0.7       0.4
COVID New Death                1.0       0.3
Inflation                      0.3       1.0
```

FIGURE 2: Correlation Matrix between Different Variables

We also then plotted the below scatter plots to understand the relationship between the explanatory variables and the response variables and the takeaways remain the same.

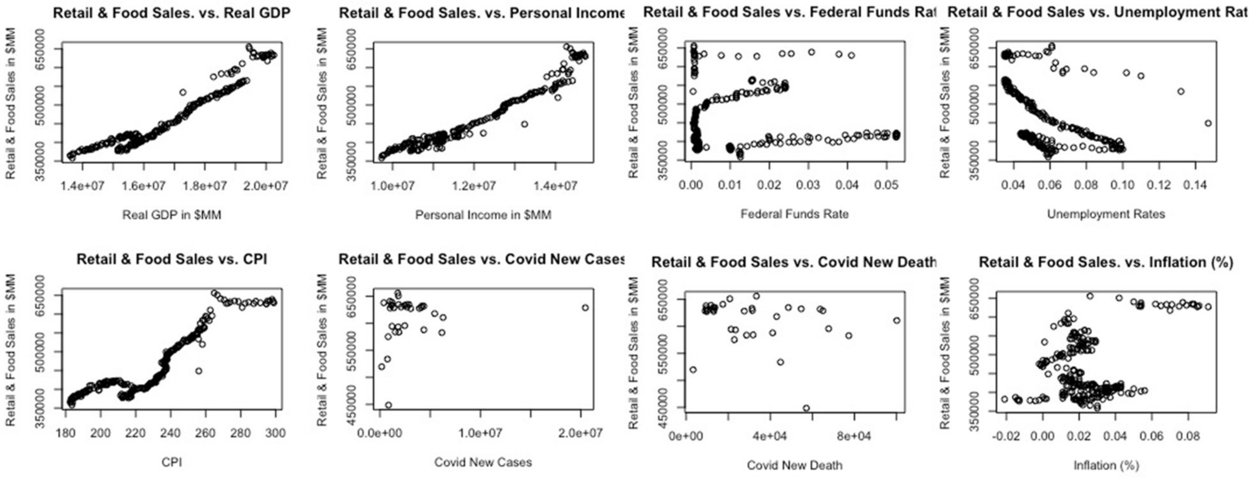
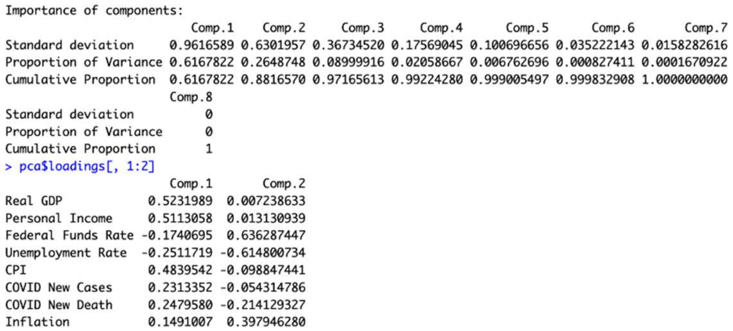


FIGURE 3: Relationship between retail and food sales and different economic indicators

We then carried out Principal Component Analysis (PCA) to see how each explanatory variable captures the variance of the data and reduce to a small number of components to aid in our analysis and visualization. The first component (Comp.1) has the highest standard deviation of 0.96 and explains the largest proportion of variance (0.62) in the data set. The proportion of variance explained by each subsequent component decreases. The first two components (Comp.1 and Comp.2) together explain about 88% of the total variance in the data.

The loadings table shows the correlation between each variable and each component. A positive loading indicates that the variable is positively correlated with the component, while a negative loading indicates a negative correlation. The larger the absolute value of the loading, the stronger the correlation between the variable and the component.

In this case, the variables Real GDP and Personal Income have similar high loadings Comp.1, indicating that they are strongly correlated with each other. This component can be interpreted as an overall economic indicator. Comp.2 is mainly driven by the variables "Federal Funds Rate" and "Unemployment Rate", which have high negative and positive loadings, respectively. This component may represent monetary policy and its impact on employment. The other components have lower proportions of variance explained and loadings, suggesting they are less important in explaining the overall variance in the data.



```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4      Comp.5      Comp.6       Comp.7
Standard deviation     0.9616589 0.6301957 0.36734520 0.17569045 0.100696656 0.035222143 0.0158282616
Proportion of Variance 0.6167822 0.2648748 0.08999916 0.02058667 0.006762696 0.000827411 0.0001670922
Cumulative Proportion  0.6167822 0.8816570 0.97165613 0.99224280 0.999005497 0.999832908 1.0000000000
                       Comp.8
Standard deviation          0
Proportion of Variance      0
Cumulative Proportion       1
> pca$loadings[, 1:2]
                       Comp.1      Comp.2
Real GDP            0.5231989  0.007238633
Personal Income     0.5113058  0.013130939
Federal Funds Rate -0.1740695  0.636287447
Unemployment Rate  -0.2511719 -0.614800734
CPI                 0.4839542 -0.098847441
COVID New Cases     0.2313352 -0.054314786
COVID New Death     0.2479580 -0.214129327
Inflation           0.1491007  0.397946280
```

FIGURE 4: PCA Summary

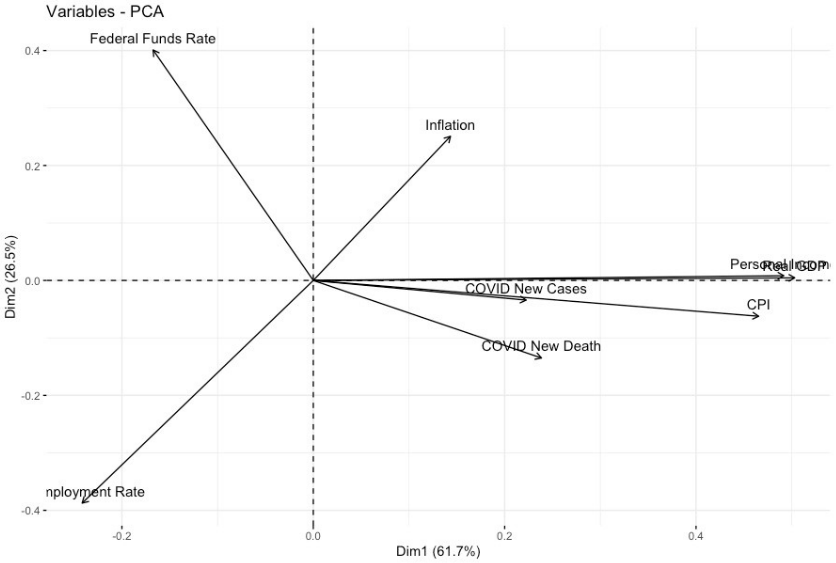


FIGURE 5: Biplot of the attributes

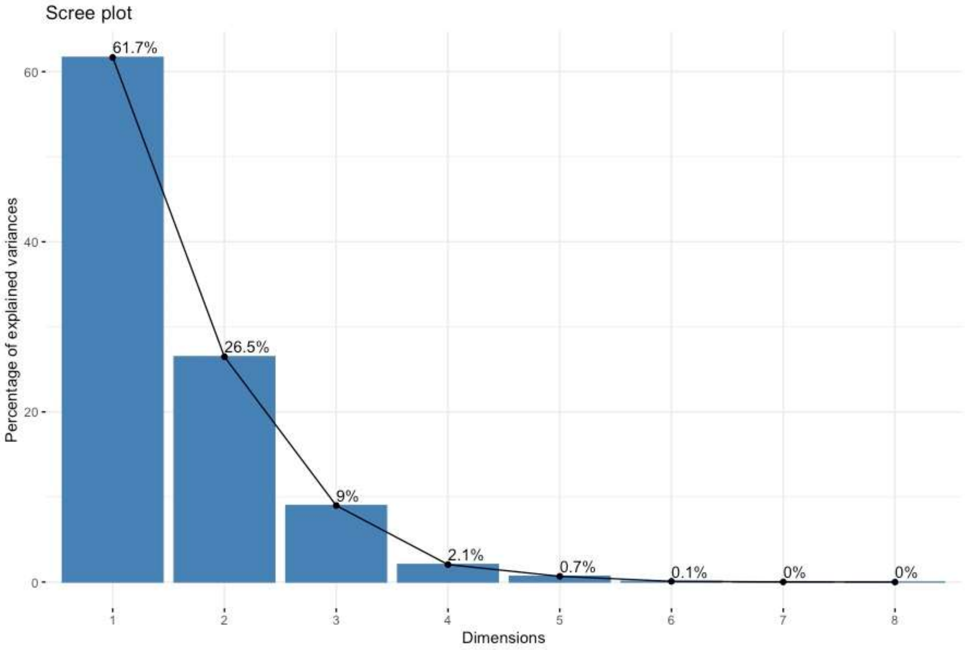


FIGURE 6: Screeplot of the attributes

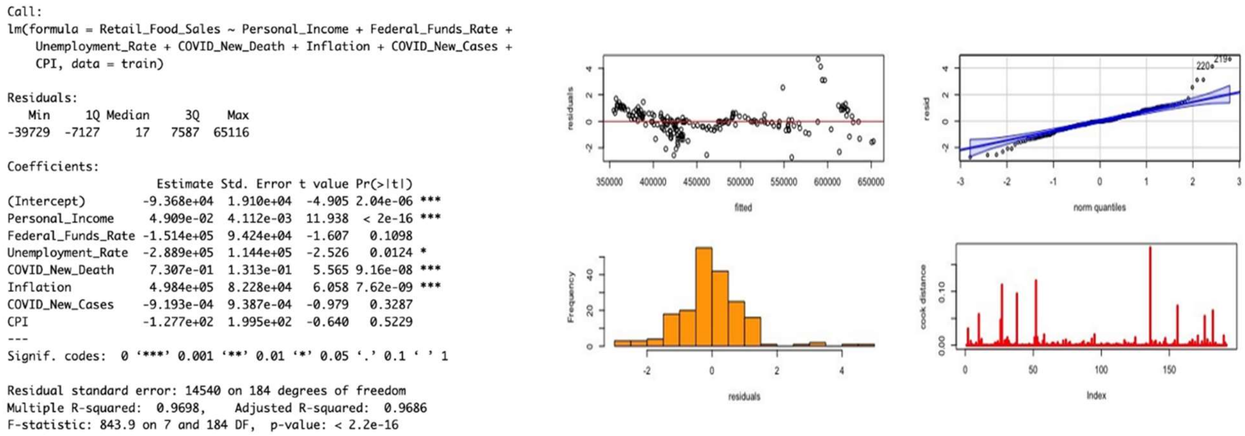
MODEL SELECTION

In this project, we employed multiple linear regression analysis to understand the relationship between the various macroeconomic factors and the retail & food sales. We started with full regression, tested significance of all variables, and did multiple iterations and transformations including the Box Cox transformation to improve Goodness-of-Fit. We also tested LASSO and Ridge regression to do variable selection and model comparison to come up with the best regression model.

While fitting the models, we tested the linear regression assumptions to make sure that linear regression is the right approach for our solution:

- Linearity - the change in the dependent variable is proportional to the change in the independent variable(s).
- Constant Variance - across all levels of the independent variables.
- Normality - the residuals are normally distributed.
- And previously we had already accounted for multicollinearity

In Model 1, we fitted all independent variables against the dependent variable. We can see that COVID new cases and CPI have very low statistical significance and all other variables are statistically significant with Personal Income, COVID New Death ad Inflation around 100% confidence level, and Federal Funds Rate at 90% confidence interval. Model 1 Results and the Residual Analysis can be seen in Figure 7 below.

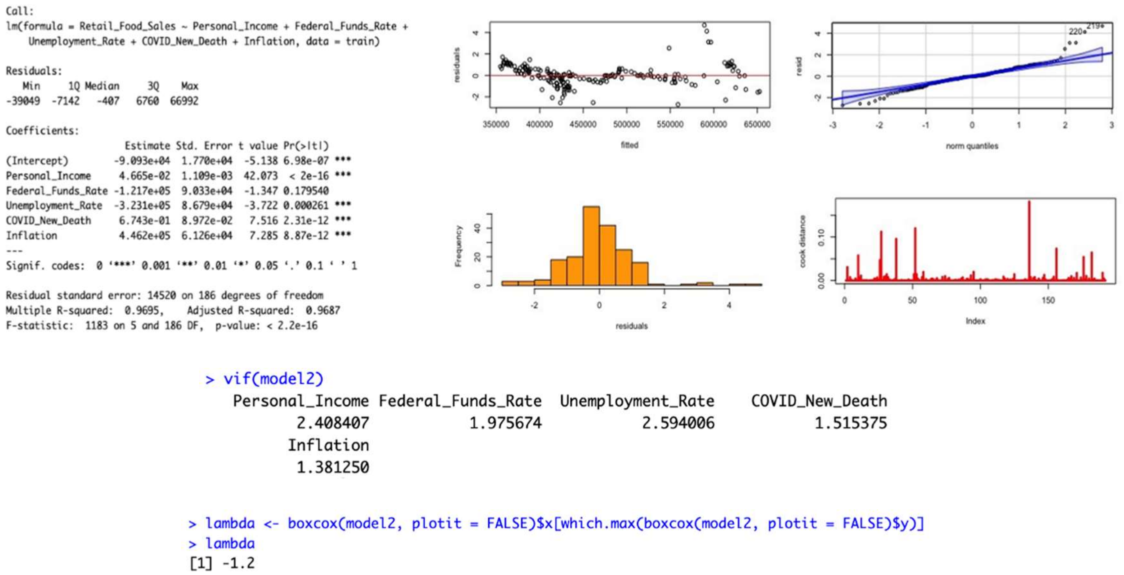


```
Call:
lm(formula = Retail_Food_Sales ~ Personal_Income + Federal_Funds_Rate +
    Unemployment_Rate + COVID_New_Death + Inflation + COVID_New_Cases +
    CPI, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-39729  -7127     17   7587  65116

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -9.368e+04  1.910e+04  -4.905 2.04e-06 ***
Personal_Income     4.909e-02  4.112e-03  11.938  < 2e-16 ***
Federal_Funds_Rate -1.514e+05  9.424e+04  -1.607   0.1098
Unemployment_Rate  -2.889e+05  1.144e+05  -2.526   0.0124 *
COVID_New_Death     7.307e-01  1.313e-01   5.565 9.16e-08 ***
Inflation           4.984e+05  8.228e+04   6.058 7.62e-09 ***
COVID_New_Cases    -9.193e-04  9.387e-04  -0.979   0.3287
CPI                -1.277e+02  1.995e+02  -0.640   0.5229
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14540 on 184 degrees of freedom
Multiple R-squared:  0.9698,    Adjusted R-squared:  0.9686
F-statistic: 843.9 on 7 and 184 DF,  p-value: < 2.2e-16
```

FIGURE 7: Model 1 Results and the Linear Regression Assumption Analysis

The R^2 of model1 is 0.97, meaning that 97% of the response can be explained by the variables included in the model. The Goodness-of-Fit was then analyzed by assessing the linear regression assumption. The residuals vs fitted plot shows that the assumption of Constant Variance and Uncorrelated Error holds true as the data scattered randomly around the zero line without any clusters or trends although there is a slight increase in the variance as the fitted values get larger. The qqPlot shows a tail on the right, which is consistent with the histogram of the residuals, suggesting that the data is skewed to the left. We also plotted all the Cook’s Distance and there seems to be no outliers or influential points.

Since CPI and COVID New Cases were determined to be insignificant from model1, we tried to remove those parameters and run a MLR model again as model2. model2 summary and the Residual Analysis and VIF results can be seen in Figure 8 below.



```
Call:
lm(formula = Retail_Food_Sales ~ Personal_Income + Federal_Funds_Rate +
    Unemployment_Rate + COVID_New_Death + Inflation, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-39049  -7142   -407   6760  66992

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -9.093e+04  1.770e+04  -5.138 6.98e-07 ***
Personal_Income    4.665e-02  1.109e-03  42.073  < 2e-16 ***
Federal_Funds_Rate -1.217e+05  9.033e+04  -1.347 0.179540
Unemployment_Rate -3.231e+05  8.679e+04  -3.722 0.000261 ***
COVID_New_Death    6.743e-01  8.972e-02   7.516 2.31e-12 ***
Inflation          4.462e+05  6.126e+04   7.285 8.87e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14520 on 186 degrees of freedom
Multiple R-squared:  0.9695,	Adjusted R-squared:  0.9687
F-statistic:  1183 on 5 and 186 DF,  p-value: < 2.2e-16
```

```
> vif(model2)
   Personal_Income Federal_Funds_Rate  Unemployment_Rate    COVID_New_Death
          2.408407           1.975674           2.594006           1.515375
         Inflation
          1.381250
```

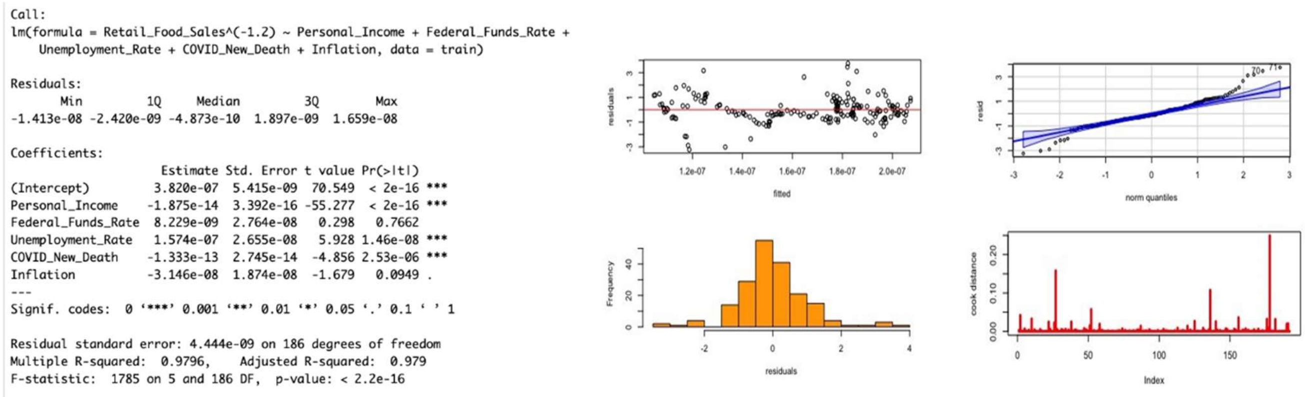
```
> lambda <- boxcox(model2, plotit = FALSE)$x[which.max(boxcox(model2, plotit = FALSE)$y)]
> lambda
[1] -1.2
```

FIGURE 8: Model 2 Results, the Linear Regression Assumption Analysis, VIF, BoxCox Results

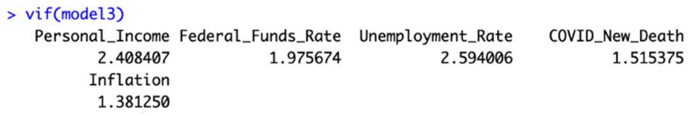
Model2 also displays a high R^2 of 0.97. Assumptions of Uncorrelated Error and Constant Variance still hold true. However, the Normality was not improved by removing CPI and Covid New Case. Therefore, we performed BoxCox to Model2 and determine lambda in order to perform BoxCox Transformation on the response – Retail and Food Sales – with the hope to improve the normality of data. The VIF analysis suggests that multicollinearity is not a concern for model2.

Lambda was determined to be -1.2. Therefore, with model3, Retail & Food Sales was transformed to be Retail & Food Sales ^ (-1.2)

Model3 Summary and Residual Analysis can be seen in Figure 9 below:



```
Call:
lm(formula = Retail_Food_Sales^(-1.2) ~ Personal_Income + Federal_Funds_Rate +
    Unemployment_Rate + COVID_New_Death + Inflation, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-1.413e-08 -2.420e-09 -4.873e-10  1.897e-09  1.659e-08

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.820e-07  5.415e-09  70.549  < 2e-16 ***
Personal_Income   -1.875e-14  3.392e-16 -55.277  < 2e-16 ***
Federal_Funds_Rate 8.229e-09  2.764e-08   0.298   0.7662
Unemployment_Rate  1.574e-07  2.655e-08   5.928 1.46e-08 ***
COVID_New_Death   -1.333e-13  2.745e-14  -4.856 2.53e-06 ***
Inflation         -3.146e-08  1.874e-08  -1.679   0.0949 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.444e-09 on 186 degrees of freedom
Multiple R-squared:  0.9796,	Adjusted R-squared:  0.979
F-statistic:  1785 on 5 and 186 DF,  p-value: < 2.2e-16
```



```
> vif(model3)
   Personal_Income Federal_Funds_Rate  Unemployment_Rate    COVID_New_Death
          2.408407           1.975674           2.594006           1.515375
         Inflation
          1.381250
```

FIGURE 9: Model 3 Results, the Linear regression Assumption Analysis & VIF Results

R^2 of model3 was slightly better than of model1 and model2. The residual vs. fitted plot also shows a slightly better random distribution of the data around the zero line. In addition, the normality improved as the histogram shows a normal distribution and qqPlot shows less tailing at the end. There are also no outliers, and vif results showed that multilinearity was not a concern.

Although model3 displays a high Rsq of 0.98, we realized that linear regression does not include any penalty terms and estimates the coefficients that best fit the data, without considering their size or significance. Therefore, overfitting could be a problem to the linear regression model. In order to assess this situation, we also run LASSO and RIDGE regression models. Ridge regression adds a penalty term to the regression equation to shrink the coefficient estimates towards zero, which reduces the variance of the model. Lasso regression uses a penalty term that shrinks some coefficient estimates to exactly zero, which allows for variable selection and can simplify the model.

MODEL COMPARSION

To evaluate the performance of the multiple models, we used the Root Mean Squared Error (RMSE). The RMSE is a measure of the difference between the actual and predicted values of the response variable. A lower RMSE value indicates a better fit of the model. The Mean Square Error across the 3 models (Model 3, LASSO and RIDGE) were computed for comparison and plotted in Figure 9. R squared value of LASSO and RIDGE were 0.95 and 0.94, respectively. Although Linear Regression displays the highest R^2 value among three model, LASSO had the lowest MSE (Figure 9). As a result, LASSO was chosen to be the best model to predict retail and food sales. CPI was determined to be insignificant and was removed out of this model.

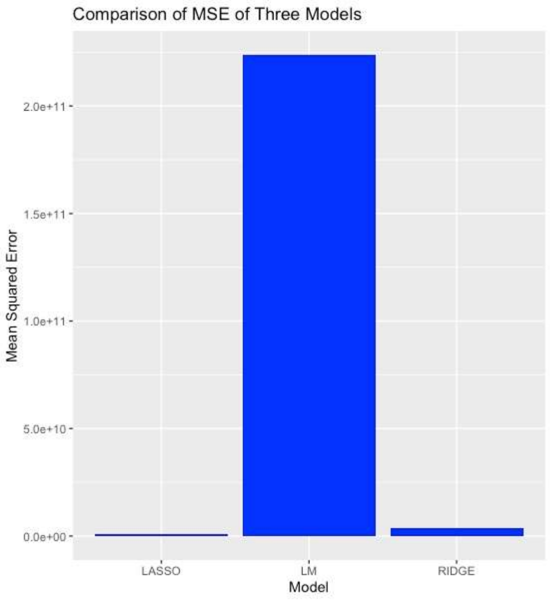


FIGURE 10: MSE Comparison across Model 3, LASSO and RIDGE

CONCLUSION

Based on our decision to choose LASSO as our final model, below is the full equation:

Retail & Food Sales = **-8.457877e+04** + **4.627665e-02** *(Personal_Income) - **1.215702e+05** *(Federal_Funds_Rate) - **3.494865e+05***(Unemployment_Rate) **-5.910178e-04***(COVID_New_Cases) + **7.467228e-01***(COVID_New_Death) + **4.409633e+05***(Inflation)

From the equation, we can see that Retail & Food Services sales have positive correlation with Personal Income, COVID New Deaths and Inflation Rate and have negative correlation with Federal Funds Rate, Unemployment Rate and COVID New cases.

FINDINGS

- Almost all economic indicators such as Retail and Food Sales, Unemployment Rate, Personal Income, and Real GDP exhibited similar patterns/trends during the 2008 financial crisis with sharp declines felt across these variables followed by a gradual recovery.
- CPI appeared to be linearly correlated to Retail and Food Sales and was determined to be insignificant to both Linear Regression, Ridge, and Lasso. This could be because there are other variables in the model that are more strongly associated with the response, and the effect of CPI is being masked or confounded by other variables. It is also possible that CPI truly has a significant relationship with retail and food sales, despite appearing to have a linear relationship.

Some of the challenges we faced during our project included the following:

- Highly correlated macroeconomic factors led to difficulties in the model selection process.
- Variability in the frequency of the independent variables, ie. Real GDP is reported on a quarterly basis, and we needed to find the best way to allocate the quarterly figures in three months per quarter.
- Discussion revolving around whether to remove the outliers such as a spike in COVID numbers in January 2022. We deemed it was appropriate to keep the outlier as it reflected the economic reality.

With additional resources and time, we would have liked to further explore the impact of COVID on by comparing sales across different geographic regions or performing a deeper dive into sales in various sectors within this industry like online versus offline retail, entertainment, and fitness.

WORKS CITED

1. " Role of macroeconomic variables on firms' performance: Evidence from the UK " (2017) https://www.tandfonline.com/doi/full/10.1080/23322039.2017.1405581

2. "Elucidating the Macroeconomic Determinants of Undernourishment in South Asian Countries: Building the Framework for Action" by Noshaba Aziz, Jun He, Ali Raza, Hongguang Sui, and Wang Yue (2021) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8397478/

3. US Set for Recession Next Year, Economists Predict" https://www.ft.com/content/53fcbbf1-39e3-483c-a6f2-b0de432ed5a3

DATA SOURCES

- Real Gross Domestic Product, Expanded Detail, Chained Dollars US Bureau of Economic Analysis
- Personal Income and Its Disposition in Chained Dollars US Bureau of Economic Analysis
- Federal Funds Rate Selected Interest Rates
- CPI CPI for All Urban Consumers (CPI-U)
- Unemployment Rate Labor Force Statistics from the Current Population Survey
- Real BEA Retail and Food Services Sales – US Bureau of Economic Analysis
- Inflation OECD.Stat
- COVID WHO Coronavirus (COVID-19) Dashboard