Selected Datasets and Justification

"USFR_ReconcNetOpCostBudgDfct_20121001_20220930.csv":

- Ideal for analyzing budget deficits and understanding the components contributing to net operating costs.

"USFR_StmtChgCashBal_20121001_20220930.csv":

- Useful for cash flow analysis, assessing the government's financial position, and identifying trends in cash balance changes.

"USFR_StmtNetCost_20121001_20220930.csv":

- Focuses on the net cost of government agencies, enabling agency cost analysis and fiscal responsibility assessment.

Roadmap for Analysis

Data Preparation:

- Clean and prepare the selected datasets by handling missing values, outliers, and ensuring data types are appropriate for analysis.

Exploratory Data Analysis (EDA):

- Summarize key variables in each dataset.
- Use visualizations (like time-series plots, scatter plots) to identify trends and relationships.

Fiscal Policy Analysis:

- Analyze components of budget deficits and net operating costs.
- Compare budget allocations vs. actual expenses over different fiscal years.

Cash Flow and Budget Analysis:

- Investigate the flow of cash in and out of different budget activities.
- Assess trends in budget receipts, outlays, and deficits.

Agency Cost Analysis:

- Evaluate gross and net costs incurred by different government agencies.
- Analyze the efficiency of agencies based on their net costs.

Statistical and Machine Learning Analysis:

- Apply Logistic Regression, LDA, QDA, or KNN for classification tasks using financial indicators.
- Use Lp-norms and Minkowski distances for financial variables to measure similarity.

Addressing Prof. Bozdog's Feedback:

- Narrowing Research Questions: Focus on specific aspects such as budget deficits, agency costs, and cash flow.
- Fiscal Policy Comparisons: Include comparisons in the EDA to assess fiscal policy changes.
- Time-Series Analysis: Utilize ARIMA or other relevant methods.
- Detailed Tables and Graphs: Create comprehensive visualizations to support your findings.
- Correct Data Preparation: Pay special attention to fields like 'Net.Cost..in.Billions'.

- Scatterplots for Cost vs Earned Revenue: Use relevant datasets to analyze this relationship.

Presentation and Reporting:

- Develop a comprehensive report with detailed findings, visualizations, and interpretations.
- Ensure the report aligns with the narrowed scope and addresses all aspects of your project proposal and feedback received

Project Report

Financial Analysis of US Economy and Government

Naveen Mathews Renji
Anmol Agrawal
Pradyumn Pundir
FA-582 - Fall Semester

Introduction

The financial activities of the U.S. government are vast and multifaceted. Analyzing this financial data can provide significant insights into the economic health and operational efficiency of the government. This study aims to delve into the financial data released by the U.S. government, focusing on data collection, preparation, feature extraction, cleaning, and analytical processing to discern trends and relationships

Research Question

1. What are the prominent trends in the U.S. government's financial data over the years?
2. How do different financial variables interact with each other?
3. Can any correlations or patterns be identified that could inform fiscal policy?

Datasets

Source - U.S. Treasury Fiscal Data. Financial Report of the U.S. Government.

The record dates in the datasets are uniformly set in the **same month each year**, reflecting the annual financial reporting cycle of the U.S. government. This consistency aligns with the end of the federal government's fiscal year on **September 30th**. The uniform record dates represent the time when annual financial statements are finalized, following the completion of necessary accounting and auditing processes. These dates signify the provision of consolidated fiscal year-end figures, ensuring a standardized timeframe for financial reporting and analysis
Ideal for analyzing budget deficits and understanding the components contributing to net operating costs.

1. USFR_ReconcNetOpCostBudgDfct_20121001_20220930.csv

Headers and Descriptions

- Record Date: The date when the financial record was documented.
- Statement Fiscal Year: The fiscal year associated with the financial statement.
- Restatement Flag: Indicator showing whether the data was restated or not (Yes/No).
- Account Description: General description or category of the account.
- Component Description: Detailed description of the account component.
- Line Item Description: Specific details of the line item in the financial statement.
- Position Amount (in Billions): Monetary value of the line item, expressed in billions.
- Fiscal and Calendar Date Details: Include fiscal year, quarter number, and corresponding calendar year, quarter number, month number, and day number.

2. USFR_StmtChgCashBal_20121001_20220930.csv

Useful for cash flow analysis, assessing the government's financial position, and identifying trends in cash balance changes.

Headers and Descriptions

- Record Date: The date of the financial record entry.
- Statement Fiscal Year: The fiscal year for which the statement is relevant.
- Restatement Flag: Indicates if the data has been restated since its initial release.
- Account Description: General category or type of the account in question.
- Component Description: More specific details or subcategories of the account.
- Line Item Description: Detailed information regarding the particular line item.
- Position Amount (in Billions): The value of the line item, measured in billions of dollars.
- Fiscal and Calendar Date Details: Details like fiscal quarter number, calendar year, and more to provide date context.

3. USFR_StmtNetCost_20010930_20220930.csv

Focuses on the net cost of government agencies, enabling agency cost analysis and fiscal responsibility assessment.

Headers and Descriptions

- Record Date: Date when the record was entered or documented.
- Statement Fiscal Year: Fiscal year pertaining to the financial data.
- Restatement Flag: A flag indicating whether the statement has been restated.
- Agency Name: The name of the government agency associated with the financial data.
- Gross Cost: The total cost before deductions or adjustments.
- Earned Revenue: Revenue earned by the agency during the fiscal period.
- Subtotal: A subtotal figure, often calculated before final totals.
- Change in Assumptions: Any changes in financial assumptions from previous statements.
- Net Cost: The net cost after accounting for revenues and changes in assumptions.
- Source Line Number: A reference number for the line in the original document.
- Fiscal and Calendar Date Details: Includes detailed date information like fiscal year, quarter number, and corresponding calendar dates.

Exploratory Data Analysis and Preprocessing

Perform EDA (Exploratory Data Analysis) on the dataset and provide some summary statistics and identify possible relationships between the features

Checking the Summary of the 3 Datasets -

```
> summary(df_reconcNetOpCostBudgDfct)
 Record.Date        Statement.Fiscal.Year Restatement.Flag   Account.Description Component.Description Line.Item.Description Position.Amount..in.Billions.
 Length:408         Min.   :2015         Length:408         Length:408          Length:408            Length:408            Min.   :-4170.90
 Class :character   1st Qu.:2017         Class :character   Class :character    Class :character      Class :character      1st Qu.:  -22.27
 Mode  :character   Median :2019         Mode  :character   Mode  :character    Mode  :character      Mode  :character      Median :    4.35
                    Mean   :2019                                                                                          Mean   :  -25.01
                    3rd Qu.:2020                                                                                          3rd Qu.:   45.23
                    Max.   :2022                                                                                          Max.   : 2629.00
 Source.Line.Number  Fiscal.Year   Fiscal.Quarter.Number Calendar.Year  Calendar.Quarter.Number Calendar.Month.Number Calendar.Day.Number
 Min.   : 1.00      Min.   :2016   Min.   :4            Min.   :2016   Min.   :3               Min.   :9            Min.   :30
 1st Qu.: 8.00      1st Qu.:2017   1st Qu.:4            1st Qu.:2017   1st Qu.:3               1st Qu.:9            1st Qu.:30
 Median :15.00      Median :2019   Median :4            Median :2019   Median :3               Median :9            Median :30
 Mean   :15.14      Mean   :2019   Mean   :4            Mean   :2019   Mean   :3               Mean   :9            Mean   :30
 3rd Qu.:22.00      3rd Qu.:2021   3rd Qu.:4            3rd Qu.:2021   3rd Qu.:3               3rd Qu.:9            3rd Qu.:30
 Max.   :32.00      Max.   :2022   Max.   :4            Max.   :2022   Max.   :3               Max.   :9            Max.   :30
```

```
> summary(df_stmtChgCashBal)
 Record.Date        Statement.Fiscal.Year Restatement.Flag   Account.Description Component.Description Line.Item.Description Position.Amount..in.Billions.
 Length:314         Min.   :2017          Length:314         Length:314          Length:314            Length:314            Length:314
 Class :character   1st Qu.:2018          Class :character   Class :character    Class :character      Class :character      Class :character
 Mode  :character   Median :2020          Mode  :character   Mode  :character    Mode  :character      Mode  :character      Mode  :character
                    Mean   :2020
                    3rd Qu.:2021
                    Max.   :2022
 Source.Line.Number  Fiscal.Year   Fiscal.Quarter.Number Calendar.Year  Calendar.Quarter.Number Calendar.Month.Number Calendar.Day.Number
 Min.   : 1.00      Min.   :2018   Min.   :4            Min.   :2018   Min.   :3               Min.   :9             Min.   :30
 1st Qu.: 8.00      1st Qu.:2019   1st Qu.:4            1st Qu.:2019   1st Qu.:3               1st Qu.:9             1st Qu.:30
 Median :16.00      Median :2020   Median :4            Median :2020   Median :3               Median :9             Median :30
 Mean   :16.27      Mean   :2020   Mean   :4            Mean   :2020   Mean   :3               Mean   :9             Mean   :30
 3rd Qu.:24.00      3rd Qu.:2021   3rd Qu.:4            3rd Qu.:2021   3rd Qu.:3               3rd Qu.:9             3rd Qu.:30
 Max.   :34.00      Max.   :2022   Max.   :4            Max.   :2022   Max.   :3               Max.   :9             Max.   :30

> summary(df_stmtNetCost)
 Record.Date        Statement.Fiscal.Year Restatement.Flag   Agency.Name        Gross.Cost..in.Billions. Earned.Revenue..in.Billions. Subtotal..in.Billions.
 Length:1697        Min.   :2000          Length:1697        Length:1697        Length:1697              Length:1697                 Length:1697
 Class :character   1st Qu.:2006          Class :character   Class :character   Class :character         Class :character            Class :character
 Mode  :character   Median :2012          Mode  :character   Mode  :character   Mode  :character         Mode  :character            Mode  :character
                    Mean   :2011
                    3rd Qu.:2017
                    Max.   :2022
 Change.in.Assumptions..in.Billions. Net.Cost..in.Billions. Source.Line.Number  Fiscal.Year   Fiscal.Quarter.Number Calendar.Year  Calendar.Quarter.Number
 Length:1697                         Length:1697            Min.   : 1.00      Min.   :2001   Min.   :4             Min.   :2001   Min.   :3
 Class :character                    Class :character       1st Qu.:10.00      1st Qu.:2007   1st Qu.:4             1st Qu.:2007   1st Qu.:3
 Mode  :character                    Mode  :character       Median :20.00      Median :2012   Median :4             Median :2012   Median :3
                                                            Mean   :19.88      Mean   :2012   Mean   :4             Mean   :2012   Mean   :3
                                                            3rd Qu.:29.00      3rd Qu.:2017   3rd Qu.:4             3rd Qu.:2017   3rd Qu.:3
                                                            Max.   :42.00      Max.   :2022   Max.   :4             Max.   :2022   Max.   :3
 Calendar.Month.Number Calendar.Day.Number
 Min.   :9             Min.   :30
 1st Qu.:9             1st Qu.:30
 Median :9             Median :30
 Mean   :9             Mean   :30
 3rd Qu.:9             3rd Qu.:30
 Max.   :9             Max.   :30
```

Checking for Missing Values in the Datasets

a. I was unable to find any missing value when I checked for NA
b. So I tried other possibilities of missing entries like -
    i. Null
    ii. null
    iii. NULL
    iv. " "
    v. ""
c. After adding these checks in the missing value criteria, I then created a dataframe of the results and printed it.

	Column	NA_Count	Empty_String_Count	Null_String_Count	Whitespace_Count	Dataset
Position.Amount..in.Billions.	Position.Amount..in.Billions.	0	0	2	0	USFR_StmtChgCashBal
Gross.Cost..in.Billions.	Gross.Cost..in.Billions.	0	0	39	0	USFR_StmtNetCost
Earned.Revenue..in.Billions.	Earned.Revenue..in.Billions.	0	0	193	0	USFR_StmtNetCost
Subtotal..in.Billions.	Subtotal..in.Billions.	0	0	730	0	USFR_StmtNetCost
Change.in.Assumptions..in.Billions.	Change.in.Assumptions..in.Billions.	0	0	1316	0	USFR_StmtNetCost
Net.Cost..in.Billions.	Net.Cost..in.Billions.	0	0	12	0	USFR_StmtNetCost

The above output depicts that the 3rd dataset - USFR_StmtNetCost has the most missing values with the most of them being in the Change in Assumptions in Billions Column.

Handling the missing Values

Now we know that the total count of rows is 1697. That means only 381 rows had a value for that column. Now this is not necessarily missing data as this is due to the fact that a missing value indicates no change in assumptions.

Since this is the case, I will be replacing all the missing values with 0 to represent it more accurately for our analysis.

Checking if the Data types are appropriate for every column

Upon analysis of the summaries, it was obvious that the date columns are not in the Date format and that all the columns that were ".. in Billions" were of character type when a numeric data type would more accurately help our analysis.

Therefore, I converted

1. The Dates to a Date format
2. The columns that contained "in Billions" to numeric data type.

```
Structure of USFR_StmtNetCost:
> str(df_stmtNetCost)
'data.frame':   1697 obs. of  16 variables:
 $ Record.Date                          : Date, format: "2022-09-30" "2022-09-30"
 $ Statement.Fiscal.Year                : num  2022 2022 2022 2022 2022 ...
 $ Restatement.Flag                     : chr  "N" "N" "N" "N" ...
 $ Agency.Name                          : chr  "Department of Veterans Affairs" "D
...
 $ Gross.Cost..in.Billions.             : num  414 1813 980 1294 595 ...
 $ Earned.Revenue..in.Billions.         : num  5.4 154.1 47.6 0.3 55.6 ...
 $ Subtotal..in.Billions.               : num  408 1659 932 1294 540 ...
 $ Change.in.Assumptions..in.Billions.: num  1526.5 1.4 527 0 0 ...
 $ Net.Cost..in.Billions.               : num  1935 1660 1459 1294 540 ...
 $ Source.Line.Number                   : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Fiscal.Year                          : num  2022 2022 2022 2022 2022 ...
 $ Fiscal.Quarter.Number                : num  4 4 4 4 4 4 4 4 4 4 ...
 $ Calendar.Year                        : num  2022 2022 2022 2022 2022 ...
 $ Calendar.Quarter.Number              : num  3 3 3 3 3 3 3 3 3 3 ...
 $ Calendar.Month.Number                : num  9 9 9 9 9 9 9 9 9 9 ...
 $ Calendar.Day.Number                  : num  30 30 30 30 30 30 30 30 30 30 ...
> c
```

Similarly with the other two datasets as well.

Checking for outliers within these datasets

Having outliers seem to be highly unlikely as this data was produced by the US government for their annual reporting and publishing. But let us be sure ourselves, I have decided to use InterQuartile Range analysis to find the outliers. Once we find them, we can choose how to deal with them.

Outliers in USFR_ReconcNetOpCostBudgDfct

```
Outliers in USFR_ReconcNetOpCostBudgDfct :
$Position.Amount..in.Billions.
  [1] -4170.9   774.1  1662.8   176.3  2629.0  -182.0  2485.8  2621.5   156.7   255.2 -1375.5 -3094.9   282.0   439.2   767.5   283.1  1061.2   950.8  -150.7  -389.3
 [21]  -156.7  -242.9 -2775.6 -3094.9   282.0   439.2   767.5   283.1  1061.2   950.8  -150.7  -362.9  -156.7  -269.3 -2775.6 -3841.4   160.1   733.3   975.2  1078.9
 [41]  1025.9  -150.6  -203.5 -3131.9 -3828.8   160.1   733.3   969.0  1078.8  1025.8  -150.7  -198.5 -3131.9 -1446.3   183.1   173.5   458.0   555.6   513.2  -984.4
 [61] -1445.1   183.1   173.5   458.0   555.6   513.2  -984.4 -1159.0   282.2   376.7   423.2  -779.0 -1159.0   282.2   376.8   423.3  -779.0 -1153.6   180.5   313.7
 [81]   490.7   537.5   567.0  -665.7 -1156.7   180.5   313.7   490.7   537.6   607.0  -137.4  -665.7 -1051.7   477.7   437.0   500.9   541.2  -587.4 -1047.4   477.7
[101]   437.0   490.1   509.3  -587.4  -514.2   211.8   173.6  -438.9
```

Outliers in USFR_StmtChgCashBal

```
$Position.Amount..in.Billions.
 [1]   4896.1  -6271.6  -1375.5   1057.1  17457.5 -15701.2   1520.9   4046.0  -6821.6  -2775.6   -823.0  20375.7 -19194.0   1326.7  -1451.9   1926.9   4046.0  -6821.6
[19]  -2775.6   -823.0  20375.7 -19194.0   1326.7  -1451.9   1926.9   3420.0  -6551.9  -3131.9   1066.5  18969.1 -14822.4   4114.2   3449.9   1402.3   1926.9   3420.0
[37]  -6551.9  -3131.9   1066.5  18969.1 -14822.4   4114.2   3449.9   1402.3   1926.9   3462.2  -4446.6   -984.4  11813.4 -10732.1   1020.9   3462.2  -4446.6   -984.4
[55]  11813.4 -10732.1   1020.9   3328.7  -4107.7   -779.0  10080.1  -8993.5   1031.9   3328.7  -4107.7   -779.0  10080.1  -8993.5   1031.9   3314.9  -3980.6   8700.8
[73]  -8222.9
```

Outliers in USFR_StmtNetCost

```
$Subtotal..in.Billions.
  [1]  408.4 1658.6  932.0 1294.1  539.5  526.4  496.5  147.2  240.9  115.4   87.4   82.8 6888.9  347.1 1507.2  807.8 1193.8  208.6  830.8  392.0  107.7  230.6  101.9
 [24]   88.1   78.5  347.4  396.8  -79.1 6832.4 1507.2 1193.8  807.8  830.8  347.1  396.8  392.0  347.4  230.6  208.6  107.7  101.9   88.1   78.5  -79.1 6832.4 1407.0
 [47] 1157.3  763.2  560.7  382.5  493.2  371.1  559.0  187.9  156.9   72.1  107.7  110.5  112.0 6733.5 1407.0 1157.3  382.3  762.4  560.7  559.0  493.2  371.1  187.9
 [70]   72.1  156.9  110.5  112.0  107.7 6732.5 1222.3 1100.9  359.6  769.4  155.7  403.6  140.8   94.4  122.0   81.1 4868.8 1222.3 1100.9  769.4  359.6  403.6  155.7
 [93]  140.8  122.0   94.4   81.1 4868.8 1142.1 1038.3  681.6  267.7  357.3  128.6  130.4   77.9   77.9  140.9 4415.7 1142.1 1038.3  681.6  357.3  267.7  140.9  130.4
[116]  128.6   77.9   77.9 4415.7 1085.7  998.8  641.3  296.3  250.0  134.8  141.6   67.8   78.7   65.0   67.4 4174.3 1085.7  998.8  641.3  250.0  296.3   67.8  141.6
[139]  134.8   78.7   69.2   65.0 4177.4 1073.9  981.8  666.8  271.6  273.0  129.3  133.6   79.8   73.2 4131.8 1073.9  981.8  271.6  666.8  273.0  133.6  129.3   79.8
[162]   73.2 4131.1 1029.6  944.7  188.3  589.4  250.8  138.6  116.7   75.3   71.9   84.6 3872.6 1029.6  944.7  601.1  250.8  182.1  138.6  116.7   84.6   75.3   71.9
[185] 3878.1  951.4  906.4  655.4  260.0  203.1  141.2  103.0   65.8   76.3 3833.5  951.4  906.4  655.4  260.0  203.1  141.2  103.0   65.8   76.3 3833.5  895.5  867.0
[208]  640.2  247.6  238.8  140.1   76.4   80.1 3525.4  895.5  867.0  640.2  238.8  247.6  140.1   80.1   76.4 3525.4  854.6  822.6  713.6  205.8  245.4  146.8  102.6
[231]   76.5   72.5 -180.4 3494.1  856.2  825.1  728.7  209.5  245.4  149.0  107.3   78.2   73.0 -177.5 3494.1  876.9  782.5  749.0  119.0  250.9  144.4  132.7   77.0
[254]   84.2 3632.7  877.0  782.5  750.7  250.9  119.6  144.8  132.8   84.6   77.2 3632.7  857.7  753.9  889.2  214.8  235.5  130.6  179.0  372.9   79.8   89.5 4163.1

$Change.in.Assumptions..in.Billions.
  [1] 1526.5    1.4  527.0  148.2    1.2    3.4    0.1    0.1 2207.9  346.3    0.7   82.8   84.9    1.6    1.9    0.1    0.1  518.4    0.7   82.8  346.3   84.9    1.6
 [24]    1.9    0.1    0.1  518.4    0.1 -17.4  602.7   89.9    3.1    1.1  679.5    0.1  602.7 -17.4   89.9    3.1    1.1  679.5   58.0  139.0    0.3    0.9    0.7
 [47]  198.9  139.0   58.0    0.3    0.9    0.7  198.9    0.4   16.8   79.2   26.2    1.1    1.5  125.2    0.4   16.8   79.2   26.2    1.1    1.5  125.2    0.4   24.1
 [70]  229.7  102.5   -0.5    0.3  356.5    0.4   24.1  229.7  102.5   -0.5    0.3  356.5    0.4  -57.6  377.5  -47.1    0.2   -0.1  273.3    0.4  377.5  -57.6    0.2
 [93]   -0.1  -47.1  273.3   -0.1  -13.0  -27.5    4.1    0.1   17.1  -19.3   -0.1  -27.5  -13.0   17.1    4.1    0.1  -19.3    0.1    6.9  -22.1   21.3   -1.3   -1.4
[116]    3.5    0.1    6.9  -22.1   21.3   -1.3   -1.4    3.5    0.2  -62.8  114.1   81.9   -2.4    0.4   -0.2  131.2    0.2  -62.8  114.1   81.9   -2.4    0.4   -0.2
[139]  131.2    0.3   70.4  149.3   98.9    0.4    0.8    0.1  320.2    0.3   70.4  149.3   98.9    0.4    0.8    0.1  320.2    0.1  -32.0   58.9    0.3    0.4    0.4
[162]   28.1    0.1  -32.0   58.9    0.4    0.3    0.4   28.1   -0.1  -58.8  101.4    5.7   84.1    0.6  132.9

$Net.Cost..in.Billions.
  [1] 1934.9 1660.0 1459.0 1294.1  539.5  526.4  496.5  295.4  240.9 9096.8  693.4 1507.9  890.6 1193.8  208.6  830.8  392.0  192.6  230.6  347.4  396.8 7350.8 1507.9
 [24] 1193.8  890.6  830.8  693.4  396.8  392.0  347.4  230.6  208.6  192.6 7350.8 1407.1 1157.3  745.8  560.7  985.2  493.2  371.1  559.0  187.9 7413.0 1407.1 1157.3
 [47]  985.0  745.0  560.7  559.0  493.2  371.1  187.9 7412.0 1222.3 1100.9  417.6  908.4  403.6 5067.7 1222.3 1100.9  908.4  417.6  403.6 5067.7 1142.5 1038.3  698.4
 [70]  346.9  357.3 4540.9 1142.5 1038.3  698.4  357.3  346.9 4540.9 1086.1  998.8  665.4  296.3  479.7 4530.8 1086.1  998.8  665.4  479.7  296.3 4533.9 1074.3  981.8
 [93]  609.2  649.1  273.0 4405.1 1074.3  981.8  649.1  609.2  273.0 4404.4 1029.5  944.7  561.9  250.8 3853.3 1029.5  944.7  573.6  250.8 3858.8  951.5  906.4  662.3
[116]  260.0 3837.0  951.5  906.4  662.3  260.0 3837.0  895.7  867.0  577.4  247.6  352.9 3656.6  895.7  867.0  577.4  352.9  247.6 3656.6  854.9  822.6  784.0  355.1
[139]  245.4 -180.4 3814.3  856.5  825.1  799.1  358.8  245.4 -177.5 3814.3  877.0  782.5  717.0  250.9 3660.8  877.1  782.5  718.7  250.9 3660.8  857.6  753.9  830.4
[162]  214.8  336.9  372.9 4296.0  889.2  857.7  753.9  372.9  235.5  214.8 4163.1 4296.0  682.8  806.9  736.2  235.2  189.1 3434.7 3434.7  806.9  736.2  682.8  235.2
[185]  189.1 3434.7  712.7  663.6  740.8  184.6  241.6  430.4 3640.7  740.8  712.7  663.6  430.4  241.6  184.6 3640.7  664.5  666.8  626.1  238.9 2909.5  666.8  664.5
[208]  626.1  238.9 2909.5  627.4  633.9  592.8  221.5 2901.3  633.9  627.4  592.8  221.5 2901.3  677.0  583.8  574.1  273.2 2949.8  677.0  583.8  574.1  273.2 2949.8
[231]  649.8  550.5  532.3 2524.9  649.8  550.5  532.3 2524.9  549.7  512.6  512.3 2488.1  549.9  512.9  512.3 2485.5  406.5  472.9  215.8  492.6 2259.7  406.5  472.9
[254]  215.8  492.6 2259.7  764.2  434.5  217.7  193.3  465.0 2545.8  764.2  434.5  217.7  193.3  465.0 2545.8  382.2  387.2  230.2  444.8 2000.4
```

```
$Gross.Cost..in.Billions.
  [1]  413.8 1812.7  979.6 1294.4  595.1  557.7  496.5  250.5 7420.0  351.0 1644.9  851.8 1194.1  243.0  966.5  392.0  239.4  350.5  396.8 7406.6 1644.9 1194.1  851.8
 [24]  854.6  351.0  396.8  392.0  350.5  239.4  243.0 7294.7 1537.0 1157.6  803.0  581.3  386.5  493.2  371.1  562.1  198.0 7195.1 1537.0 1157.6  386.3  802.2  581.3
 [47]  562.1  493.2  371.1  198.0 7194.1 1341.4 1101.2  364.7  813.6  403.6 5287.2 1341.4 1101.2  813.6  364.7  403.6 5287.2 1252.6 1038.5  719.8  272.5  357.3 4808.5
 [70] 1252.6 1038.5  719.8  357.3  272.5 4808.5 1186.8  999.1  718.7  296.3  254.8 4606.2 1186.8  999.1  718.7  254.8  296.3 4609.3 1170.0  982.1  721.9  276.5  273.0
 [93] 4515.7 1170.0  982.1  276.5  721.9  273.0 4507.7 1130.9  945.0  193.1  634.9  250.8 4248.2 1130.9  945.0  646.6  250.8 4253.7 1029.5  906.8  716.9  260.0  207.4
[116] 4251.4 1029.5  906.8  716.9  260.0  207.4 4251.4  968.4  867.4  685.1  247.6  243.3 3940.9  968.4  867.4  685.1  243.3  247.6 3940.9  922.4  822.9  769.6  209.9
[139]  245.4 -152.9 3844.9  924.0  825.4  784.7  213.6  245.4 -150.0 3844.9  943.3  782.9  827.0  250.9 3998.3  943.4  782.9  828.7  250.9 3998.3  920.4  754.2  929.0
[162]  214.8  240.2  402.3 4472.3  929.0  920.4  754.2  402.3  240.2  214.8 4472.3 4605.2  718.4  866.4  736.6  254.6 3735.6 3735.6  866.4  736.6  718.4  254.6 3735.6
[185]  769.1  663.9  767.6  197.0  241.6  434.6 3891.6  767.6  769.1  663.9  434.6  241.6  197.0 3891.6  689.6  718.6  626.4  238.9 3157.3  718.6  689.6  626.4  238.9
[208] 3157.3  678.8  658.0  593.1  221.5 3127.7  658.0  678.8  593.1  221.5 3127.7  703.9  623.4  572.1  276.6 3174.6  703.9  623.4  572.1  276.6 3174.6  672.1  583.9
[231]  534.9 2732.0  672.1  583.9  534.9 2732.0  562.2  542.3  512.6 2652.9  562.4  542.6  512.6 2650.3  420.4  499.9  218.4  492.9 2416.3  420.4  499.9  218.4  492.9
[254] 2416.3  776.8  459.2  217.7  196.0  465.3 2705.8  776.8  459.2  217.7  196.0  465.3 2705.8  393.5  410.2  230.2  448.7 2157.3

$Earned.Revenue..in.Billions.
  [1] 154.1  47.6  55.6  31.3  27.3  22.4  76.2 531.1 137.7  44.0  34.4 135.7  26.3  75.7  22.6 574.2 137.7  44.0  23.8  34.4  26.3  75.7  22.6 462.3 130.0  39.8  20.6
 [28]  33.3  25.2  42.8  71.7  22.0 461.6 130.0  39.8  20.6  25.2  33.3  42.8  71.7  22.0 461.6 119.1  44.2  25.3  24.1  31.9  70.2  27.8 418.4 119.1  44.2  25.3  31.9
 [55]  24.1  27.8  70.2 418.4 110.5  38.2  21.9  31.4  23.4  69.7 392.8 110.5  38.2  21.9  23.4  31.4  69.7 392.8 101.1  77.4  37.9  22.3  30.6  68.7 431.9 101.1  77.4
 [82]  22.3  37.9  30.6  68.7 431.9  96.1  55.1  21.3  29.9  70.4 383.9  96.1  55.1  29.9  70.4  21.3 376.6 101.3  45.5  29.3  26.6  67.9  20.3 375.6 101.3  45.5  29.3
[109]  20.3  26.6  67.9 375.6  78.1  61.5  79.9  24.8  66.9 417.9  78.1  61.5  79.9  24.8  66.9 417.9  72.9  44.9 103.5  23.2  66.3 415.5  72.9  44.9  23.2  66.3 103.5
[136] 415.5  67.8  56.0  20.0  64.2  27.5 350.8  67.8  56.0  20.0  64.2  27.5 350.8  66.4  78.0  64.6  30.6 365.6  66.4  78.0  30.6  64.6 365.6  62.7  39.8  29.4  65.7
[163] 309.2  39.8  62.7  29.4  65.7 309.2 309.2  35.6  59.5  67.1 300.9 300.9  59.5  35.6  67.1 300.9  56.4  26.8  73.7 250.9  26.8  56.4  73.7 250.9  25.1  51.8  73.7
[190] 247.8  51.8  25.1  73.7 247.8  51.4  24.1  71.6 226.4  24.1  51.4  71.6 226.4  26.9  39.6  19.9  68.9 224.8  26.9  39.6  19.9  68.9 224.8  22.3  33.4  68.0 207.1
[217]  22.3  33.4  68.0 207.1  29.7  67.6 164.8  29.7  67.6 164.8  27.0  66.4 156.6  27.0  66.4 156.6  24.7  65.6 160.0  24.7  65.6 160.0  23.0  64.5 156.9
```

Outlier Analysis in U.S. Government Financial Data

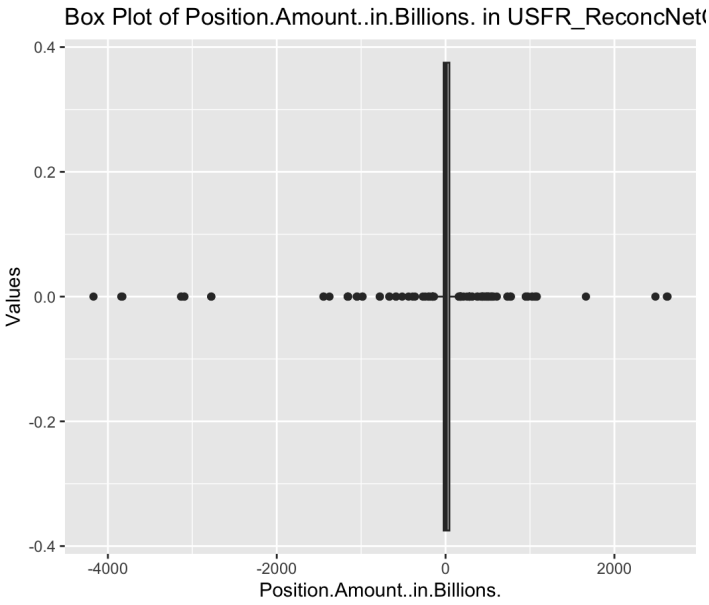
An Interquartile Range (IQR) analysis was conducted on three key financial datasets to identify outliers. This method revealed several outlier values across various financial metrics. Notably, these outliers were not uniformly classified as data errors. Given the governmental context, many outliers correspond to significant fiscal events, policy changes, or unusual but legitimate financial transactions. For instance, substantial deviations in "Position Amount (in Billions)" and "Net Cost (in Billions)" may reflect specific budgetary adjustments or extraordinary government activities. These findings underscore the importance of contextual interpretation in financial data analysis. Outliers, while initially appearing as anomalies, can provide critical insights into governmental fiscal behavior and policy implications.

Data Visualization and Analysis of Trends

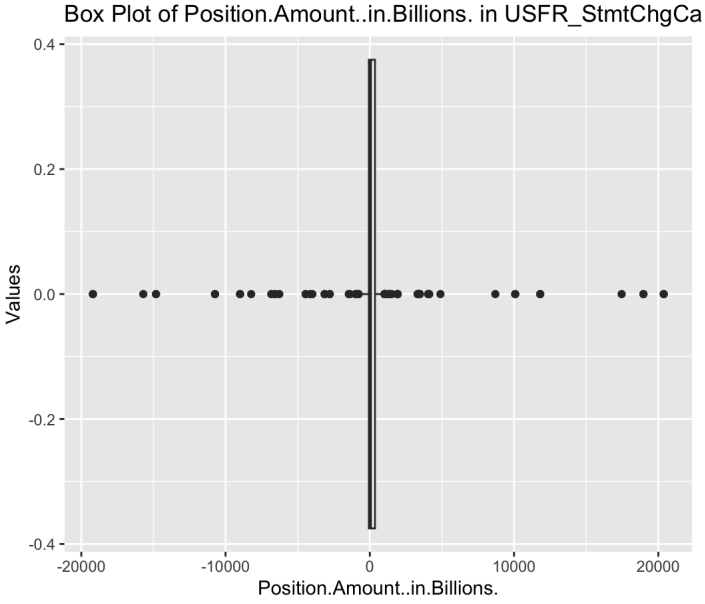
Box Plot for the Numeric Columns

The box plot n three datasets important numerical columns will give us insights on the range of data and the concentration of the data

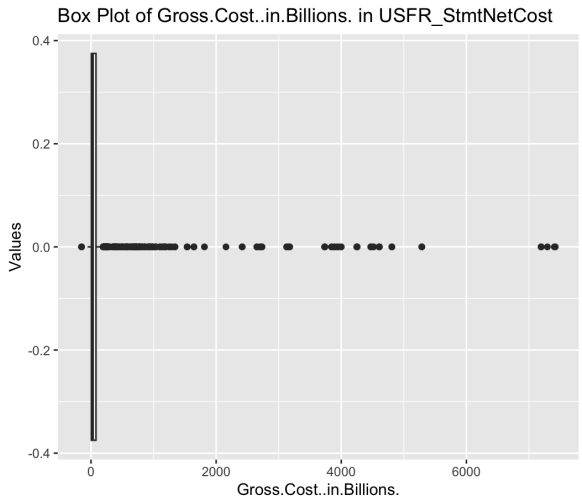
1. **Box plot of Position Amount in Billions in USFR_ReconcNetOpCostBudgDfct**

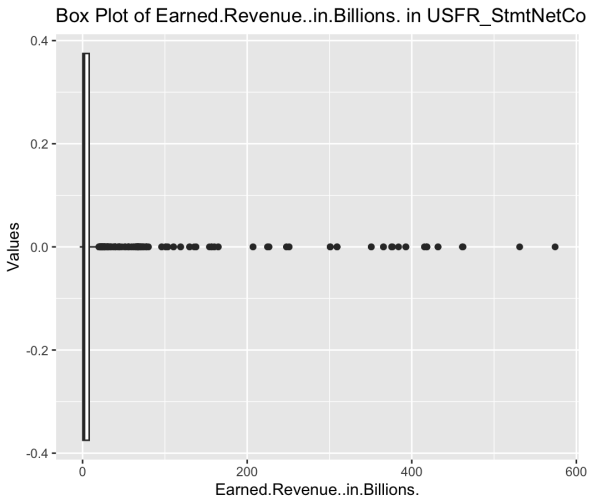


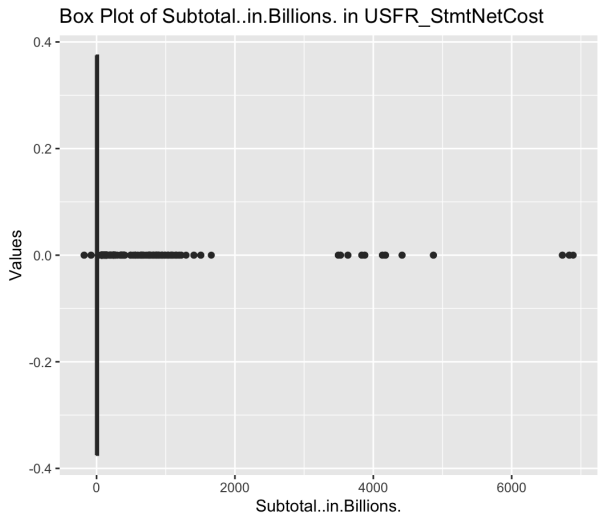
2. **Box plot of Position Amount in Billions in USFR_StmtChgCashBal**

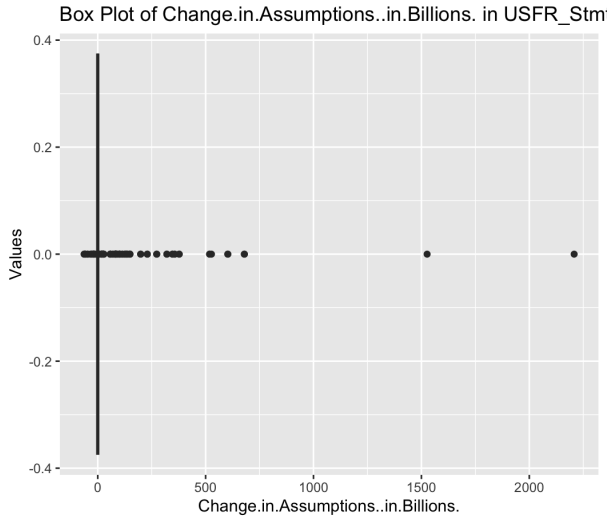


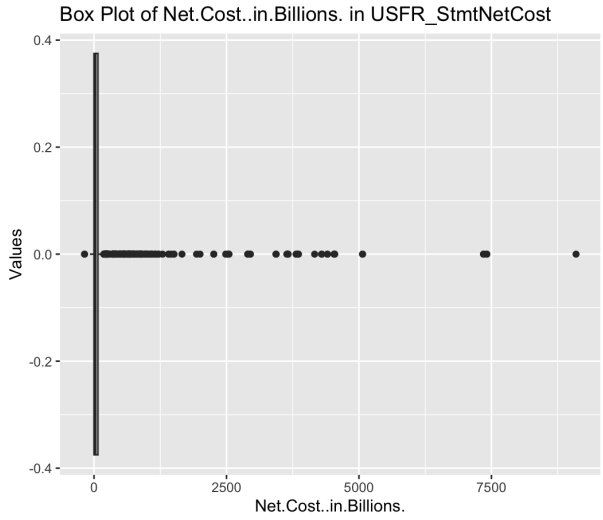
3. **Box plot of Position Amount in Billions in USFR_StmtNetCost**


Box Plot of Gross.Cost..in.Billions. in USFR_StmtNetCost
Values
0.4
0.2
0.0
-0.2
-0.4
0
2000
4000
6000
Gross.Cost..in.Billions.


Box Plot of Earned.Revenue..in.Billions. in USFR_StmtNetCo
Values
0.4
0.2
0.0
-0.2
-0.4
0
200
400
600
Earned.Revenue..in.Billions.

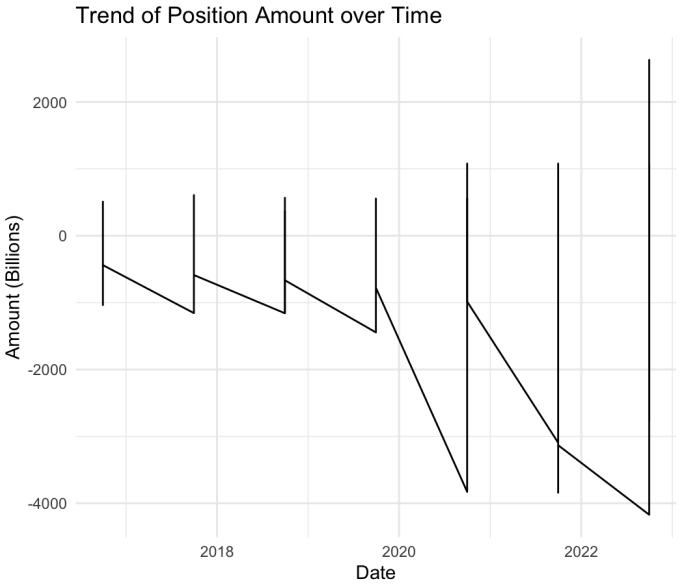

Box Plot of Subtotal..in.Billions. in USFR_StmtNetCost
Values
0.4
0.2
0.0
-0.2
-0.4
0
2000
4000
6000
Subtotal..in.Billions.

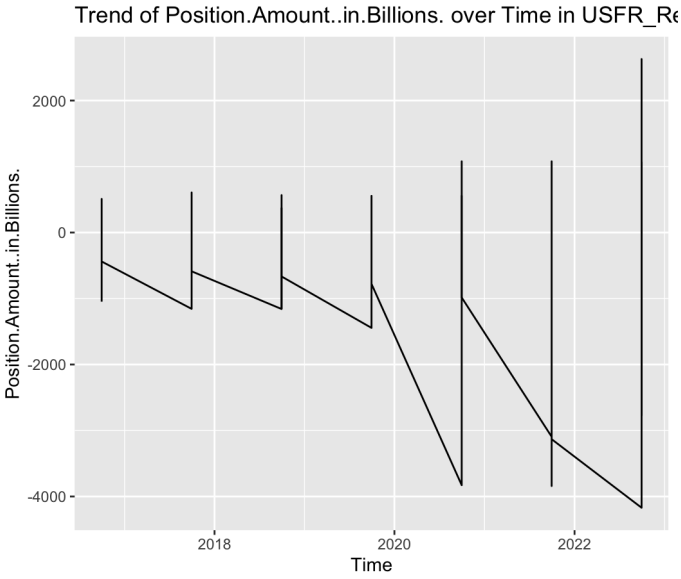

Box Plot of Change.in.Assumptions..in.Billions. in USFR_Stm
Values
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Change.in.Assumptions..in.Billions.

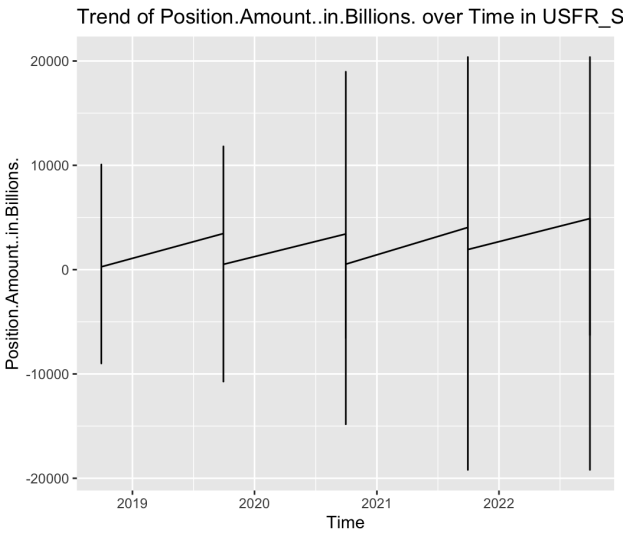

Box Plot of Net.Cost..in.Billions. in USFR_StmtNetCost
Values
0.4
0.2
0.0
-0.2
-0.4
0
2500
5000
7500
Net.Cost..in.Billions.

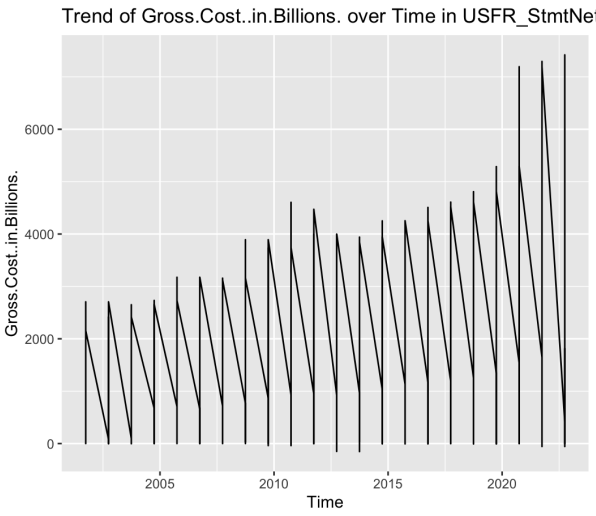
These box plots give us a good idea on the outliers that we had analyzed. We can see that there exists outliers but we cannot call them as incorrect or bad data as we stated in the IQR outlier analysis. There aren't clear medians, we cannot really see much skewing either. The box plots show us how the range of values are spread and it shows us that amounts in billions have a wide range from a couple of billions to thousands of billions.

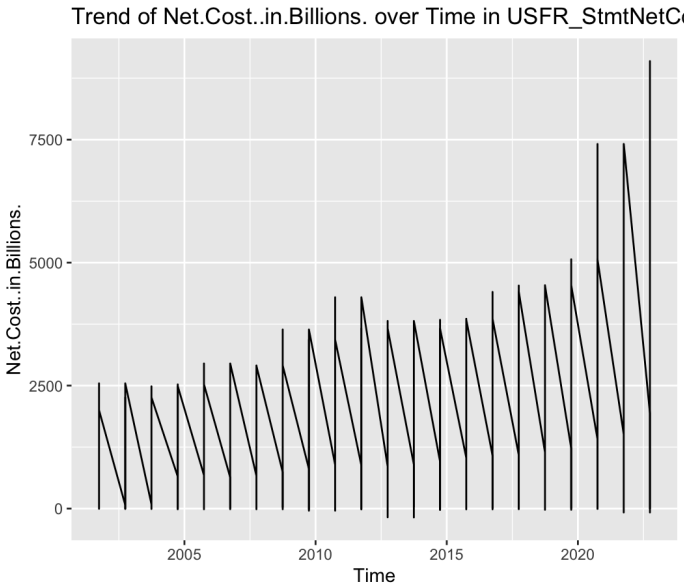
Analyzing Trends over Time






Trend of Position.Amount..in.Billions. over Time in USFR_S
Position.Amount..in.Billions.
20000
10000
0
-10000
-20000
2019
2020
2021
2022
Time


Trend of Gross.Cost..in.Billions. over Time in USFR_StmtNet
Gross.Cost..in.Billions.
6000
4000
2000
0
2005
2010
2015
2020
Time


Trend of Net.Cost..in.Billions. over Time in USFR_StmtNetCo
Net.Cost..in.Billions.
7500
5000
2500
0
2005
2010
2015
2020
Time

Based on the line plot visualizations provided for the U.S. Government financial data, it's clear that the financial position as indicated by the 'Position Amount (in Billions)' has experienced significant fluctuations. The downward trend in recent years suggests increased spending or reduced revenues, which could reflect economic challenges or shifts in fiscal policy.

The 'Gross Cost (in Billions)' indicates a generally increasing trend, which could be attributed to natural growth in government operations, inflationary effects, or new governmental programs. Similarly, the 'Net Cost (in Billions)' also exhibits an upward trend with notable spikes. These spikes could correspond to specific fiscal stimuli or emergency spending measures, possibly in response to economic events.

The visualization of these financial metrics over time is critical in understanding the trajectory of government financial management. The data suggests that while there is an overall increase in costs, there are also periods of significant adjustments. Each peak or trough in the plots could potentially be aligned with major economic events, legislative changes, or shifts in budgetary priorities.

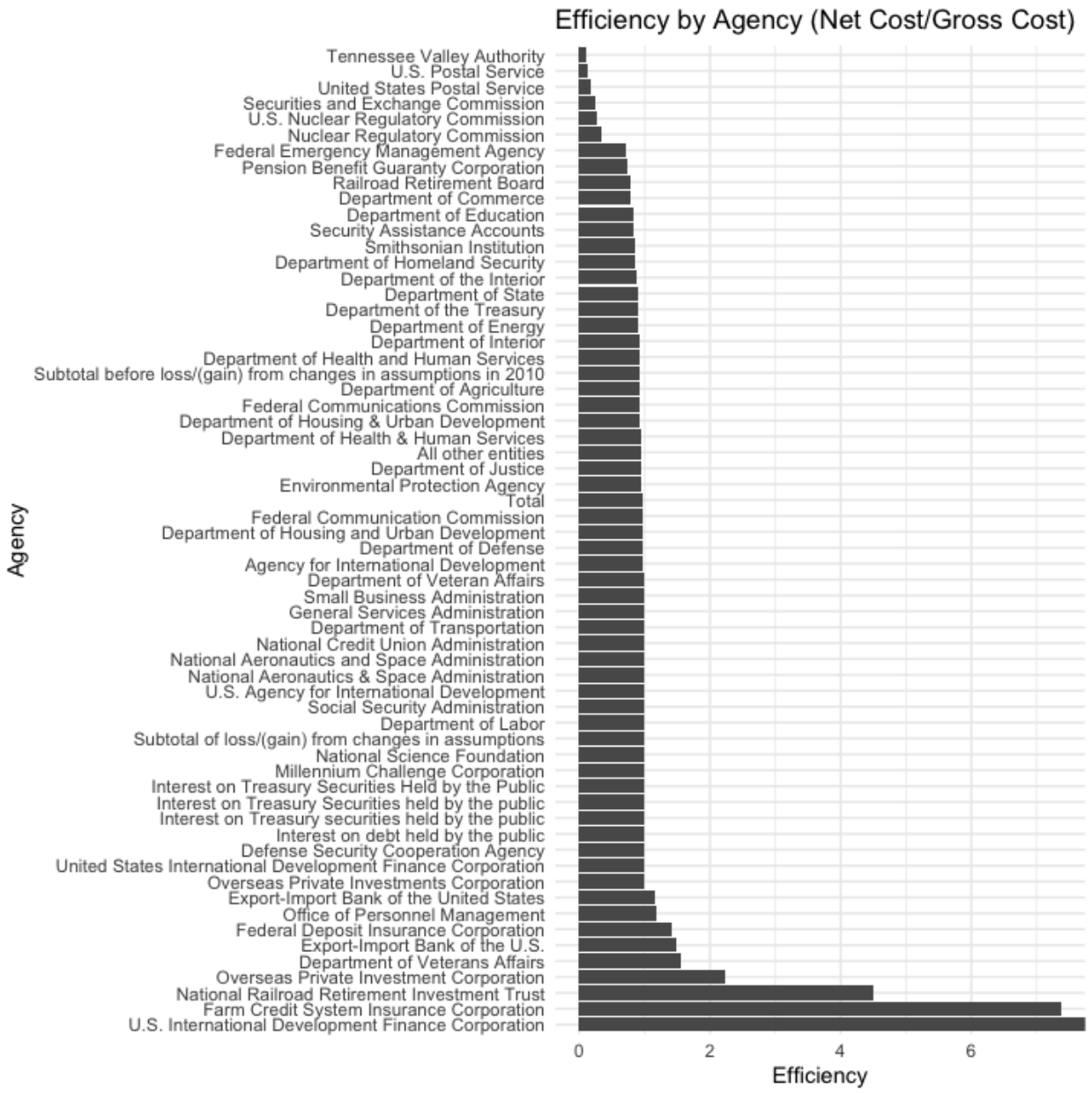
Agency wise Analysis

Agency Total Net Cost wise graph

The bar chart visualization presents an analysis of net costs incurred by various U.S. government agencies. The chart illustrates a wide range of net costs across different agencies, highlighting significant differences in financial outlays. Agencies such as the Social Security Administration and the Department of Health and Human Services demonstrate the highest net costs, likely reflecting their substantial roles in public welfare and healthcare. In contrast, organizations like the United States International Development Finance Corporation exhibit much lower net costs. This disparity in financial figures can serve as a starting point for assessing the efficiency and fiscal impact of each agency, providing valuable insights for evaluating government spending and resource allocation. Further analysis may consider the size of the agency, the scope of their programs, and the specific services they provide to fully understand the context behind these financial figures.

Efficiency of Agencies

To calculate the efficiency per agency, we are calculating the Efficiency as Total Net Cost divided by the Total Gross Cost. This gives us a ratio that represents the percentage of gross cost that is actually used to produce outputs.



Key Agencies and Observations -

Agency	Total Gross Cost (billions)	Total Net Cost (billions)	Efficiency Ratio
Department of Defense	$317.22	$307.68	0.970

Department of Health and Human Services	$33.90	$31.18	0.920
Social Security Administration	$35.25	$35.23	1.000
Department of the Treasury	$9.05	$8.10	0.896
Department of Veterans Affairs	$9.61	$15.04	1.570
Department of Agriculture	$5.95	$5.52	0.928
Department of Labor	$4.56	$4.56	1.000
Interest on Treasury Securities Held by the Public	$6.52	$6.52	1.000
All Other Entities	$1.38	$1.30	0.941
Department of Transportation	$3.36	$3.32	0.989
United States Postal Service	$0.67	$0.12	0.180
Tennessee Valley Authority	$0.43	$0.04	0.102
United States International Development Finance Corporation	$0.00	$0.00	1.000
Farm Credit System Insurance Corporation	$0.00	$0.01	7.370
Overseas Private Investment Corporation	$0.00	$0.00	2.220

National Railroad Retirement Investment Trust	$0.01	$0.02	4.510

The data shows the total gross cost, total net cost, and efficiency ratio for 62 US government agencies in the fiscal year. The total gross cost for all agencies was $181.9 billion, while the total net cost was $175.3 billion. The overall efficiency ratio across agencies was 0.963. Main observations

- The Department of Defense incurred the highest total gross cost at $317.2 billion, while also having one of the higher efficiency ratios at 0.970. This indicates that while the DoD spends a significant amount, the difference between gross and net costs is relatively small compared to its total budget.
- The Social Security Administration and the Interest on Treasury Securities categories has perfect efficiency ratios of 1.000. This makes sense given the nature of these expenditures.
- The US Postal Service stands out as having a very low efficiency ratio of 0.128, with gross costs of $2.4 billion and net costs of just $312 million. This indicates the USPS is operating at a significant loss.
- Similarly, the Tennessee Valley Authority has an efficiency ratio of just 0.102, spending far more than it takes in.
- Some smaller agencies like the Export-Import Bank and Farm Credit System Insurance Corporation have efficiency ratios well above 1.000, indicating they generate more revenue than expenses.
- And from the above table, the last 4 agencies are examples of high efficiencies not meaning much as they have very low importance in government expenditure and economical evaluation as their gross cost itself is so low.

In summary, while most major agencies operate at efficiency ratios of 0.8 to 1.0, there is significant variation across the board. **Loss-making organizations stand out as areas of potential cost savings or reform**

Discussion

Data preparation for the model training :

1. Data Set Selection : The data frame that we selected for modeling purposes is the “Statement Net Cost” data set for several reasons.

    A. The data frame contains the relevant information about the government expenditures, which are key indicators for the fiscal policies. Such as, **"Gross.Cost..in.Billions."** and **"Earned.Revenue..in.Billions."** which provides the direct measures of government spending and revenue.

    B. The “**Restatement.Flag**” column gives us an idea whether or not there have been any restatements to the financial data, which could help in estimating the fiscal policy analysis. For example : Restatement flag signifies the increase in the government expenditures for a particular year, it would make it appear that the government was spending more than actually was in that year.

    C. The “**Source.Line.Number**” provides a reference for the original source of the data, which is important for ensuring the data’s integrity and credibility.

    D. The data frame we have selected has the sufficient amount of the data to conduct the meaningful analysis of government fiscal policies. The data also has the Fiscal years, which provides enough time to observe trends and patterns in fiscal policy.

2. Feature Selection : Relevant columns were selected based on their correlation with the target variable and their overall contribution to the model's predictive performance. Irrelevant or redundant features were removed to avoid overfitting and improve the model's generalization ability.

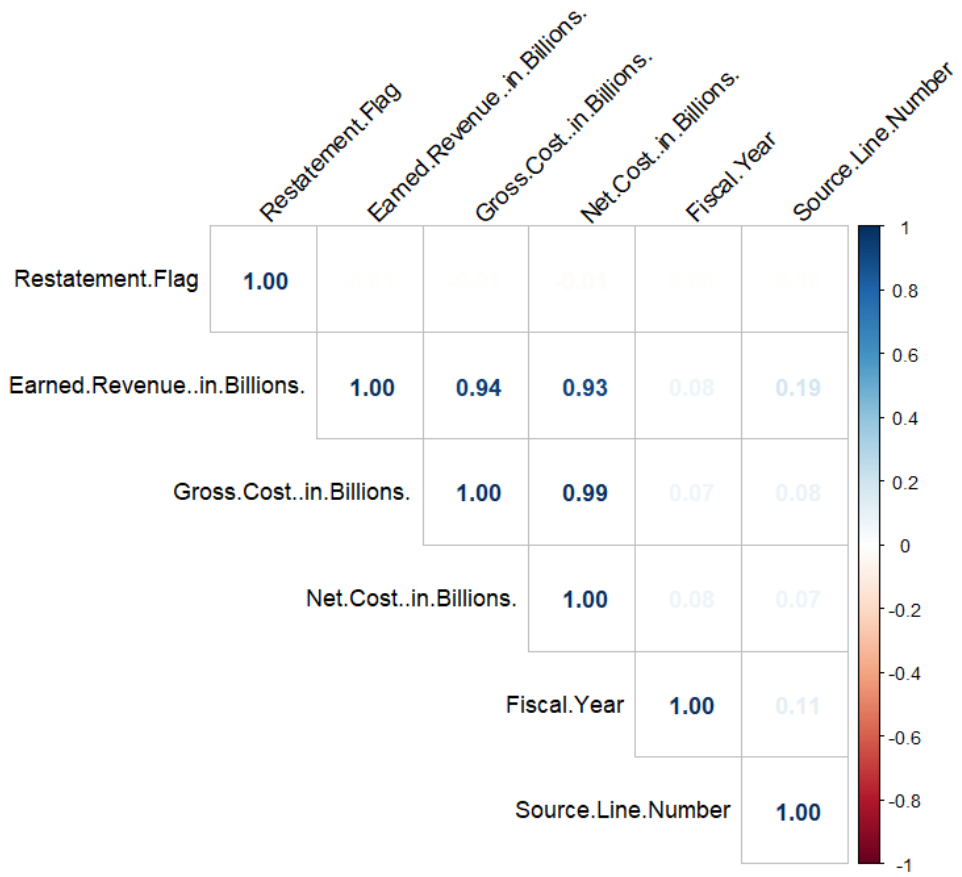


Figure : Correlation Plot

3. Feature Engineering : Removed the null values from the data set and **"Restatement.Flag**" Column has been converted to numerical data "Y" as 1 and "N" as 0.
4. Machine Learning Models : For this project we used three machine learning projects for the regression task as the target column is "**Net.Cost..in.Billions**.". Linear regression, SVM, and Random forest has been used and model performance was evaluated using the appropriate metrics such as **RMSE** (Root Mean Squared Error) and **MAE** (Mean Absolute Error).

5. Model Summary :

```
Call:
lm(formula = Net.Cost..in.Billions. ~ Gross.Cost..in.Billions. +
    Earned.Revenue..in.Billions. + Restatement.Flag + Fiscal.Year +
    Source.Line.Number, data = train_df)

Residuals:
    Min      1Q  Median      3Q     Max
-240.99   -5.57    0.66    7.52 1810.01

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -2.047e+03  5.813e+02  -3.521 0.000445 ***
Gross.Cost..in.Billions.      1.062e+00  7.942e-03 133.718  < 2e-16 ***
Earned.Revenue..in.Billions. -1.134e+00  1.015e-01 -11.172  < 2e-16 ***
Restatement.Flag             -2.551e+00  3.616e+00  -0.705 0.480631
Fiscal.Year                   1.016e+00  2.891e-01   3.514 0.000457 ***
Source.Line.Number            4.499e-02  1.732e-01   0.260 0.795116
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66.6 on 1351 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9918
F-statistic: 3.283e+04 on 5 and 1351 DF,  p-value: < 2.2e-16
```

Figure : Linear Regression Model Summary

```
> summary(rf)
                Length Class  Mode
call               3   -none- call
type               1   -none- character
predicted       1357   -none- numeric
mse              500   -none- numeric
rsq              500   -none- numeric
oob.times       1357   -none- numeric
importance         5   -none- numeric
importanceSD       0   -none- NULL
localImportance    0   -none- NULL
proximity          0   -none- NULL
ntree              1   -none- numeric
mtry               1   -none- numeric
forest            11   -none- list
coefs              0   -none- NULL
y               1357   -none- numeric
test               0   -none- NULL
inbag              0   -none- NULL
terms              3   terms  call
```

Figure : Random Forest Model Summary

```
Call:
svm(formula = Net.Cost..in.Billions. ~ Gross.Cost..in.Billions. + Earned.Revenue..in.Billions. + Restatement.Flag + Fiscal.Year +
    Source.Line.Number, data = train_df)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.2
    epsilon:  0.1


Number of Support Vectors:  107
```

Figure : SVM Model Summary

6. Results :

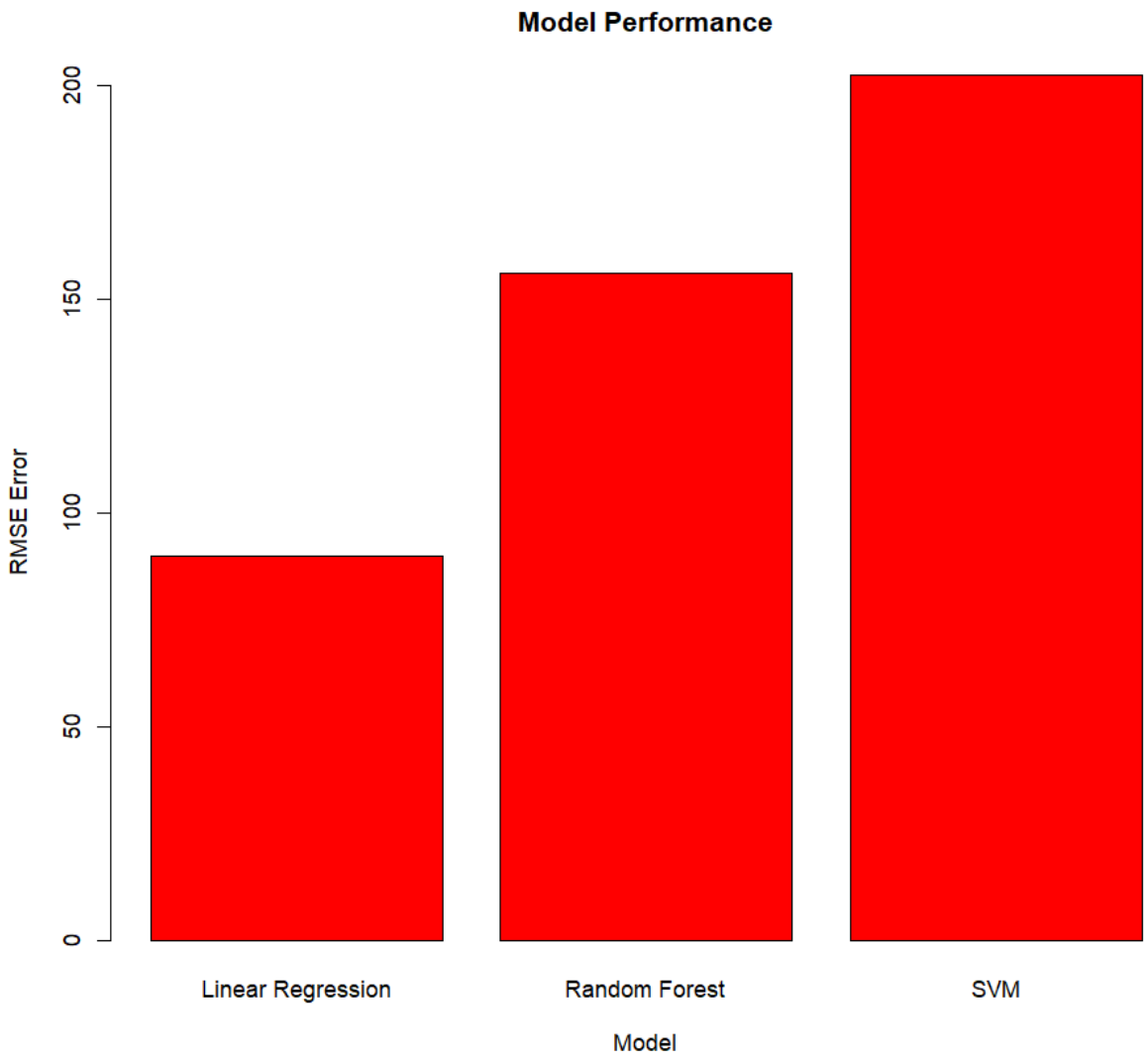


Figure : Root Mean Square Error

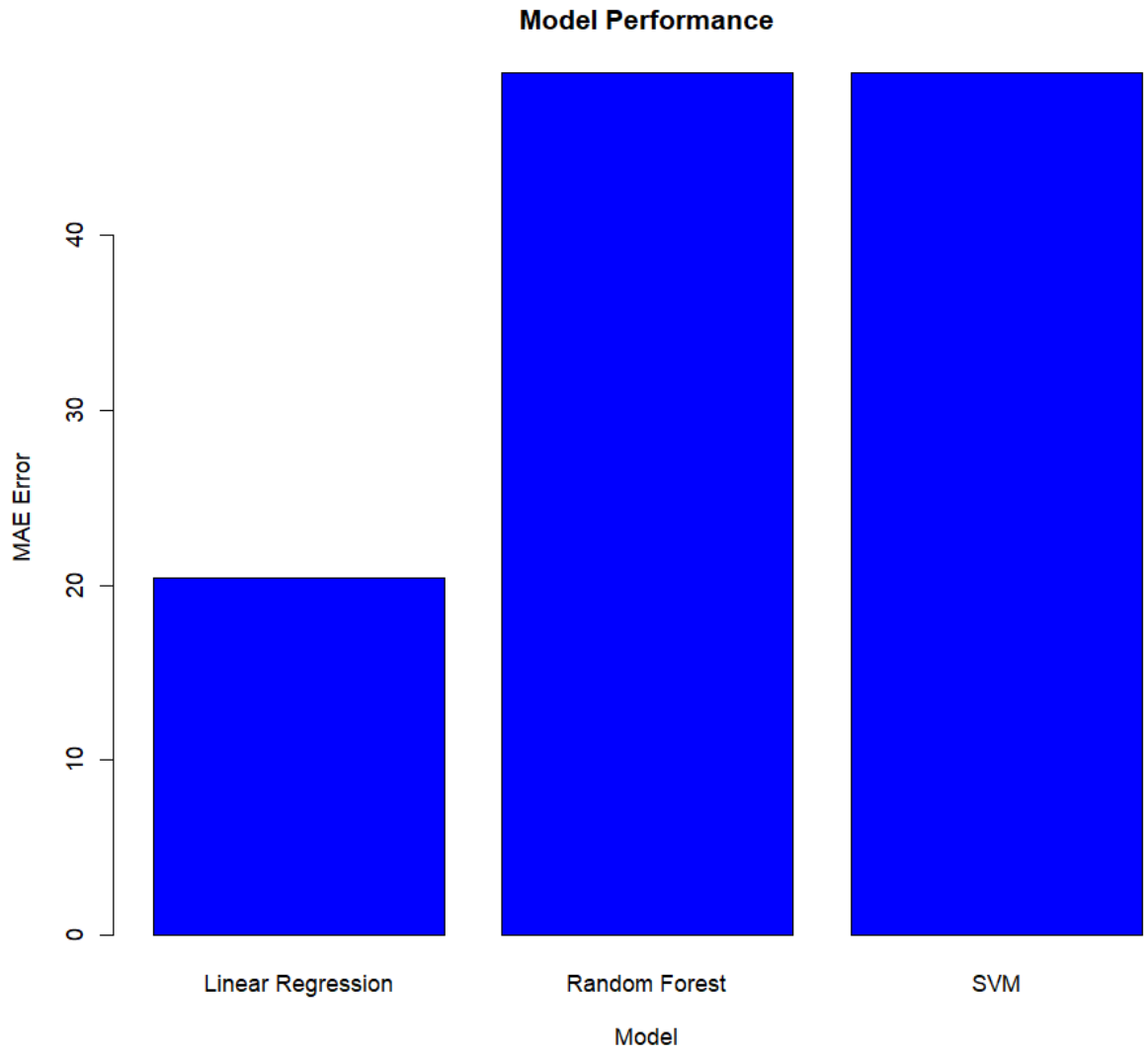


Figure : Mean Absolute Error

Issues Encountered

Present any issues that you encountered during the analysis

Next Steps

More data preparation steps such as capturing the complex relationship among the data, steps such as normalization, scaling, etc. This will help in reducing the error and increased model performance. It will be a crucial step as columns like

“**Restatement.Flag**: is crucial in determining the result but the correlation plot is not showing that much of relevance. This could be the way we encode the data assigning the “Y” as 1 and “N” as 0, which is a bad idea as it can assign more importance to a specific value. In the future we will be exploring more encoding techniques such as one-hot encoding, etc.

More modeling techniques such as ensemble learning, deep learning, and bayesian models can be used and their performance will be analyzed to get the more suitable and best performing model.

References

U.S. Treasury Fiscal Data. Financial Report of the U.S. Government. Retrieved from - link