USA Economic Data Analysis

Priyanka Sharma(m. 141620), Aashka Vadodaria(m. 145252)

Table of contents

1 Introduction **1**

2 Data Preparation **2**
2.1 Loading the dataset 2
2.2 Dataset Summary 2
2.3 Data Cleansing 3

3 Descriptive Analysis **4**
3.1 Histogram Analysis 4
3.2 Time Series Analysis 6
3.3 Box Plot Analysis 9
3.4 Correlation Analysis 11

4 Predictive Analysis **12**
4.1 Linear Regression Model Analysis 12
4.2 Time Series Forecasting 14

5 Conclusion **15**

1 Introduction

To understand the trends and patterns that drive the national and global economics, data analysis plays a very crucial role. With thorough examination of key economic indicators, we can gain insights into the financial health of a country, and predict future economic conditions. With that in mind, in this report we aim to perform a thorough descriptive and predictive analysis of a dataset related to U.S. economy. For our analysis, we have chosen the `Economics` dataset from the `ggplot2` package in R.

The objective of this analysis are twofold: Descriptive Analysis and Predictive Analysis. Through descriptive analysis, we aim to generate summary statistics and visualizations to explore central tendencies, variations, and correlations of the data. And with predictive analysis, we intend to derive the relationship between the selected indicators by using a linear regression model.

2 Data Preparation

2.1 Loading the dataset

For analysis, we need to prepare the dataset correctly. First we load the necessary libraries, then we load the dataset.

```
# Load necessary packages
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
# Load the dataset
data <- ggplot2::economics
```

2.2 Dataset Summary

Now that we have loaded the dataset, we can see the summary of the dataset as follows:

```
# View the first few rows of the dataset
head(data)
```

```
# A tibble: 6 x 6
  date         pce    pop psavert uempmed unemploy
  <date>     <dbl>  <dbl>   <dbl>   <dbl>    <dbl>
1 1967-07-01  507. 198712    12.6     4.5     2944
2 1967-08-01  510. 198911    12.6     4.7     2945
3 1967-09-01  516. 199113    11.9     4.6     2958
4 1967-10-01  512. 199311    12.9     4.9     3143
5 1967-11-01  517. 199498    12.8     4.7     3066
6 1967-12-01  525. 199657    11.8     4.8     3018
```

```
# Summary statistics
summary(data)
```

```
      date                 pce               pop            psavert
 Min.   :1967-07-01   Min.   :  506.7   Min.   :198712   Min.   : 2.200
 1st Qu.:1979-06-08   1st Qu.: 1578.3   1st Qu.:224896   1st Qu.: 6.400
 Median :1991-05-16   Median : 3936.8   Median :253060   Median : 8.400
 Mean   :1991-05-17   Mean   : 4820.1   Mean   :257160   Mean   : 8.567
 3rd Qu.:2003-04-23   3rd Qu.: 7626.3   3rd Qu.:290291   3rd Qu.:11.100
 Max.   :2015-04-01   Max.   :12193.8   Max.   :320402   Max.   :17.300
    uempmed          unemploy
 Min.   : 4.000   Min.   : 2685
 1st Qu.: 6.000   1st Qu.: 6284
 Median : 7.500   Median : 7494
 Mean   : 8.609   Mean   : 7771
 3rd Qu.: 9.100   3rd Qu.: 8686
 Max.   :25.200   Max.   :15352
```

As we can see, the primary variables used in the dataset are: Date(`date`), Personal Consumption Expenditures(`pce`), Personal Savings Rate(`psavert`), Median Duration of Unemployment(`uempmed`), Total Unemployment(`unemploy`), and Population(`pop`).

2.3 Data Cleansing

Before we proceed with any analysis, it is important that we cleanse the data and prepare the dataset properly for analysis.

First we check if there are any missing values in the data.

```
sum(is.na(data))
```

```
[1] 0
```

Since there is no missing data, we are good. To further aid in our analysis, we convert the `date` column to `Date` data type.

```
data$date <- as.Date(data$date)
```

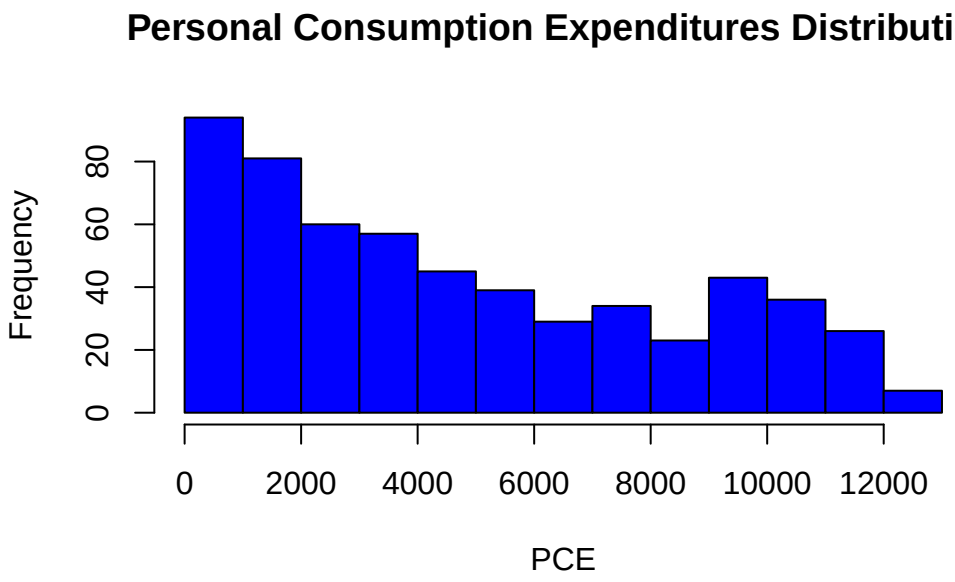
Now that we have cleansed and prepared the data, we are ready for actual analysis.

3 Descriptive Analysis

3.1 Histogram Analysis

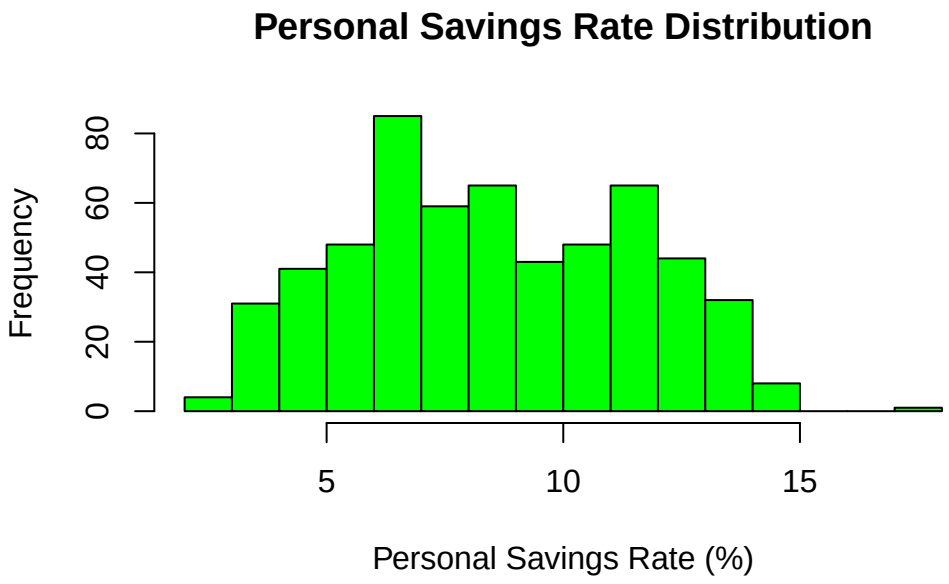
In this section, we will do histogram analysis for the key economic indicators in our dataset. We have chosen Personal Consumption Expenditures (`pce`), Personal Savings Rate (`psavert`), and Median Duration of Unemployment (`uempmed`) as our indicators.

```
hist(data$pce, main = "Personal Consumption Expenditures Distribution",
     xlab = "PCE", col = "blue")
```



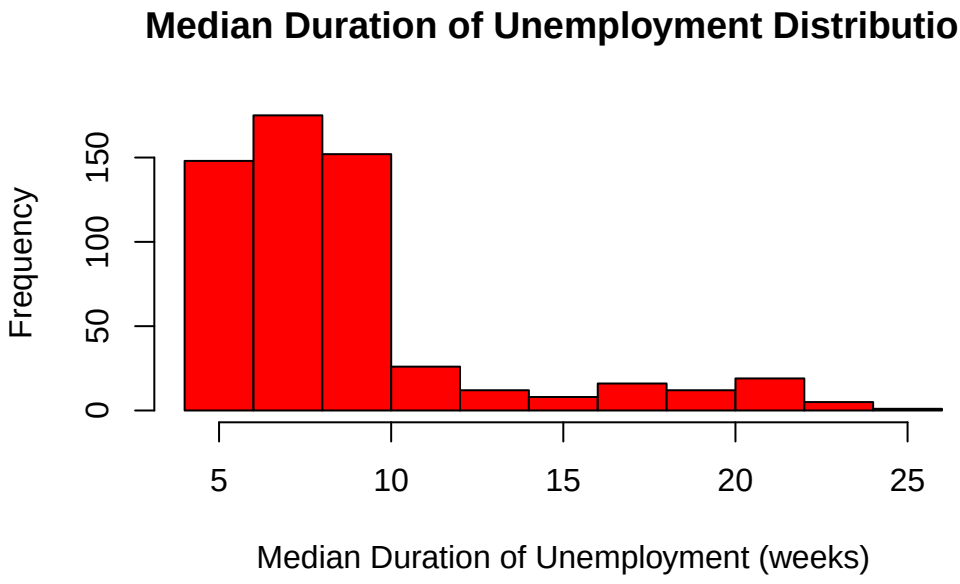
As we can see, the histogram for PCE is skewed towards right. This indicates that PCE values are concentrated on the lower end. Also, it can be seen that PCE values span a wide range which indicates that there has been significant growth in personal consumption expenditures over the years.

```
hist(data$psavert, main = "Personal Savings Rate Distribution",
     xlab = "Personal Savings Rate (%)", col = "green")
```



We can see that the histogram for personal savings is approximately normally distributed but slightly skewed towards right. Bulk of the data is centered near the mean.

```
hist(data$uempmed, main = "Median Duration of Unemployment Distribution",
     xlab = "Median Duration of Unemployment (weeks)", col = "red")
```

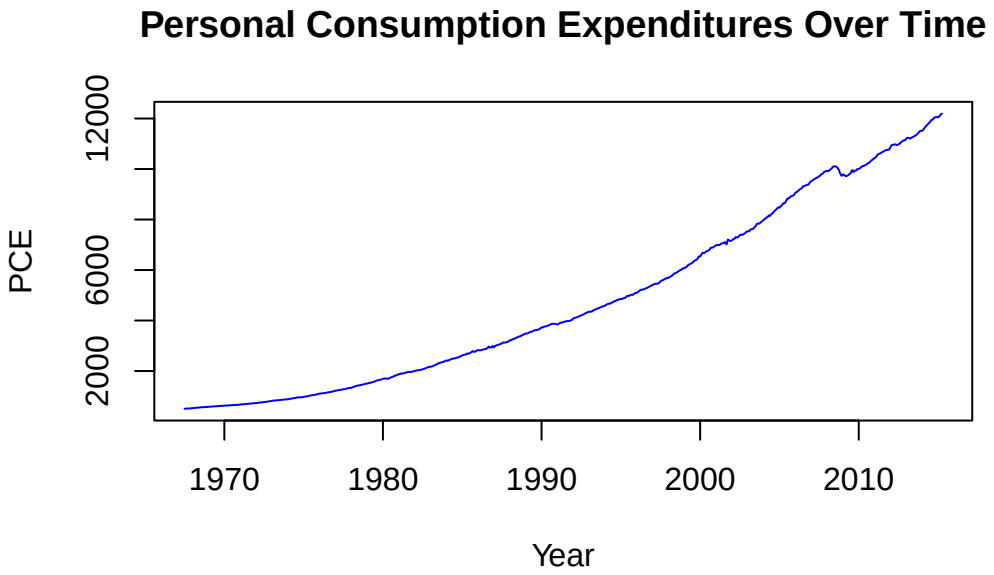


The distribution shape for the median duration of unemployment is right skewed suggesting shorter durations of unemployment are more common while longer durations are less frequent. Also, the range shows that while most individuals experience shorter unemployment periods there are notable instances where the duration extends significantaly.

3.2 Time Series Analysis

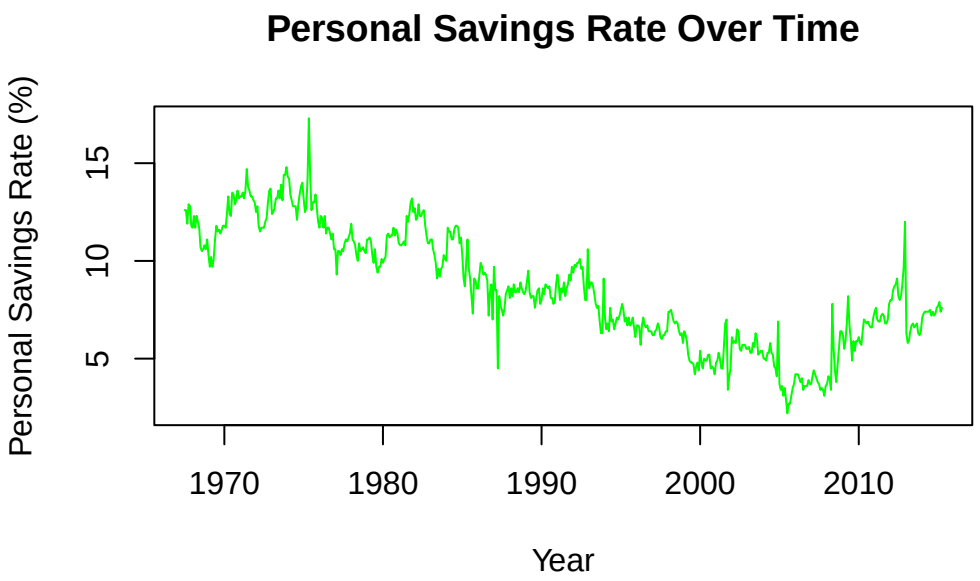
Now let's intrepret the trends seen from the time series plots.

```
plot(data$date, data$pce, type = "l",
     main = "Personal Consumption Expenditures Over Time",
     xlab = "Year", ylab = "PCE", col = "blue")
```



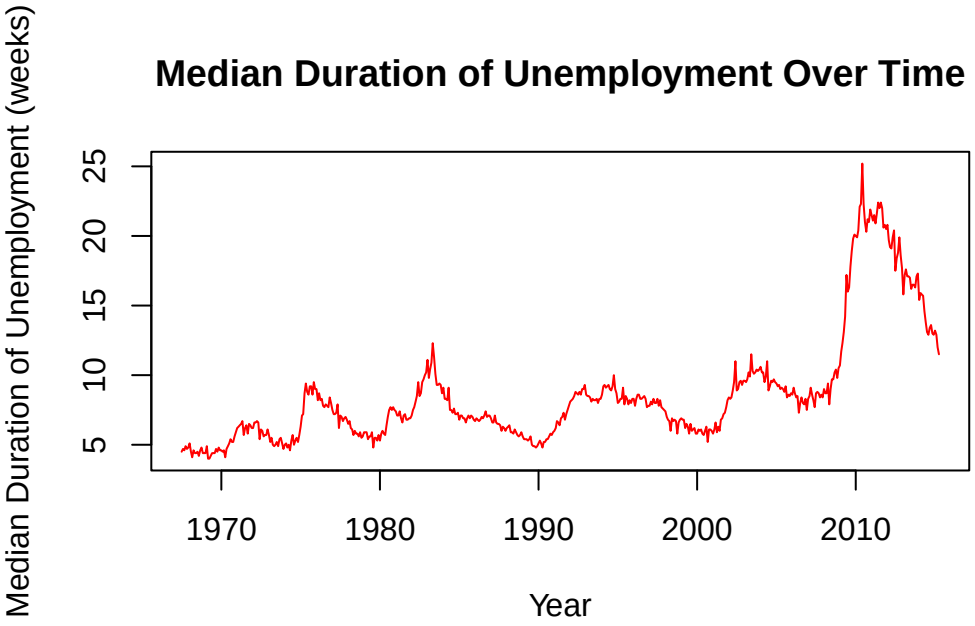
For PCE, a clear upward trend is seen from 1967 to 2015. This indicates that personal expenditures have steadily increased over time, which reflects economic growth and increased consumption power. At around the year 2008, we can see a dip in the trend. This could be due to the heavy recession of the year 2008.

```
plot(data$date, data$psavert, type = "l",
     main = "Personal Savings Rate Over Time",
     xlab = "Year", ylab = "Personal Savings Rate (%)", col = "green")
```



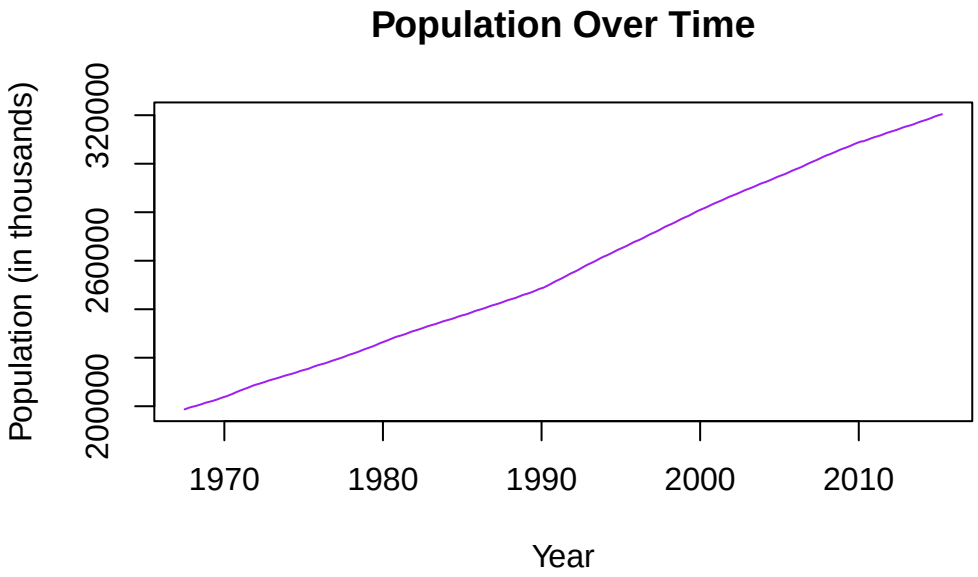
The time series plot for personal savings show considerable variation over the years. From late 1970s to early 2000s, there is a general downward trend. But in the latter part of the series, especially during and after the 2008 financial crisis, there is an observable increase in the personal savings rate. This is likely due to heightened economic uncertainty and precautionary savings behavior.

```
plot(data$date, data$uempmed, type = "l",
     main = "Median Duration of Unemployment Over Time",
     xlab = "Year", ylab = "Median Duration of Unemployment (weeks)",
     col = "red")
```



Median duration of unemployment varies significantly over time. There is no long term trend but periods of increase and decrease corresponding to economic cycles. During recession the median duration of unemployment spikes while it decreases during the times of economic growth. Here also sharp increase can be seen around the 2008 financial crisis.

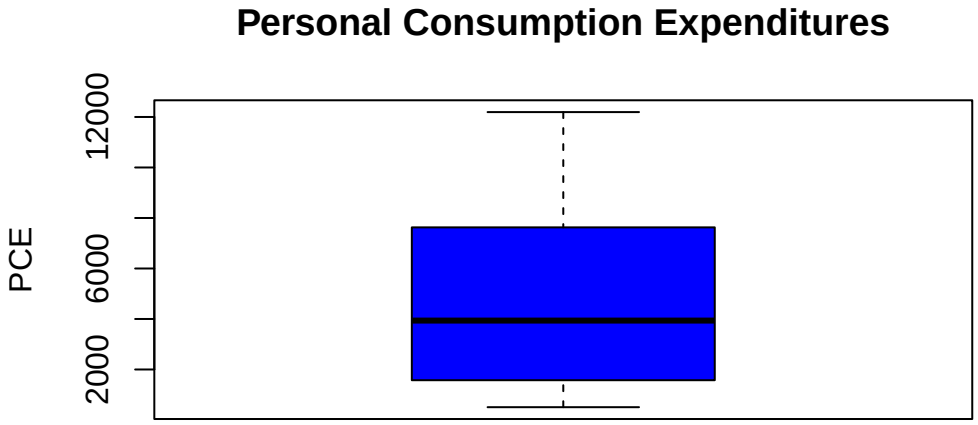
```
# Time series plot to observe trends over time
plot(data$date, data$pop, type = "l",
     main = "Population Over Time", xlab = "Year",
     ylab = "Population (in thousands)", col = "purple")
```



There is a clear and consistent trend in population over time. Also, the population is increasing almost linearly.

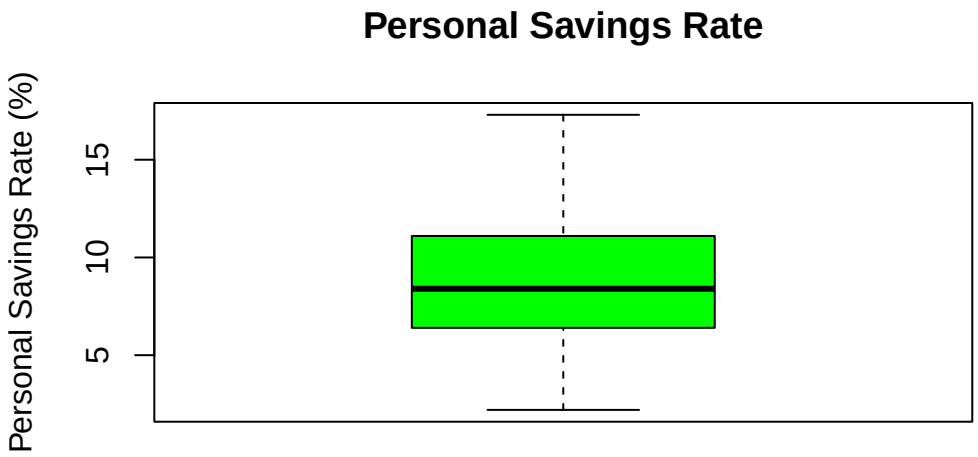
3.3 Box Plot Analysis

```
boxplot(data$pce, main = "Personal Consumption Expenditures",
        ylab = "PCE", col = "blue")
```



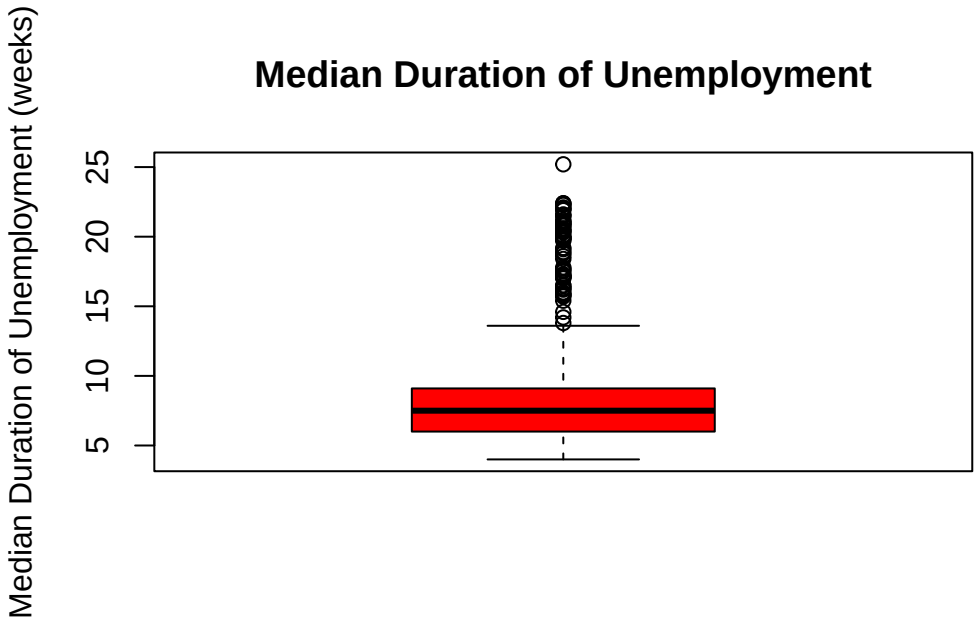
For PCE, median lies on the lower end of the scale. Also, there may be a few outliers indicating significant spikes at certain point of time.

```
boxplot(data$psavert, main = "Personal Savings Rate",
        ylab = "Personal Savings Rate (%)", col = "green")
```



For Personal Savings Rate, median is located centrally indicating a balanced distribution. Also, the Interquartile Range(IQR) is relatively narrow indicating the middle half of the data are close together.

```
boxplot(data$uempmed, main = "Median Duration of Unemployment",
        ylab = "Median Duration of Unemployment (weeks)", col = "red")
```



For Median Duration of Unemployment, we can see outliers at the upper end. This corresponds to economic downturns where unemployment durations are significantly higher.

3.4 Correlation Analysis

```
# Correlation analysis between variables
cor_matrix <- cor(data[, c("pce", "psavert", "uempmed", "unemploy", "pop")])
print(cor_matrix)
```

```
                pce    psavert    uempmed   unemploy        pop
pce       1.0000000 -0.7928546  0.7269616  0.6145176  0.9872421
psavert  -0.7928546  1.0000000 -0.3251377 -0.3093769 -0.8363147
uempmed   0.7269616 -0.3251377  1.0000000  0.8693097  0.6950085
unemploy  0.6145176 -0.3093769  0.8693097  1.0000000  0.6337165
pop       0.9872421 -0.8363147  0.6950085  0.6337165  1.0000000
```

From the table we can see that:

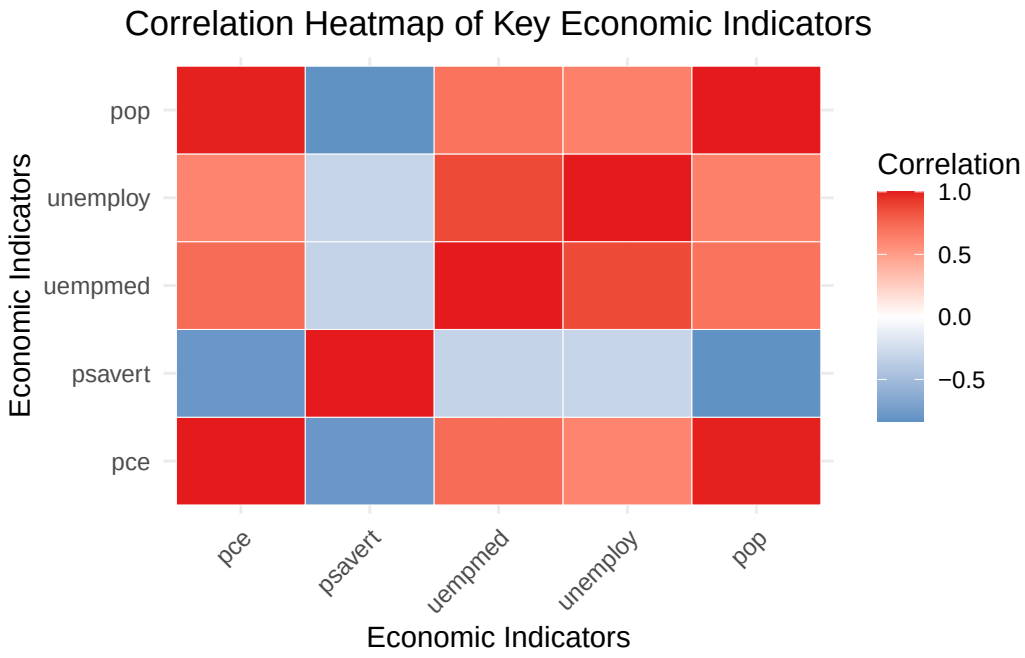
- PCE has a strong positive correlation with Population. This is expected because as population increases, PCE increases as well.
- Personal Savings Rate exhibit moderate negative correlations with PCE. This is also expected because the more you spend, the less you save.

For better visual presentation, we can show the correlation as a heatmap as well.

```
library(reshape2)
# Reshape correlation matrix for plotting
cor_melted <- melt(cor_matrix)

# Create heatmap plot
heatmap_plot <- ggplot(cor_melted, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "#377EB8",
  mid = "white", high = "#E41A1C", midpoint = 0, name = "Correlation") +
  theme_minimal() +
  labs(title = "Correlation Heatmap of Key Economic Indicators",
       x = "Economic Indicators", y = "Economic Indicators") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))

# Display the heatmap plot
print(heatmap_plot)
```



4 Predictive Analysis

4.1 Linear Regression Model Analysis

For predictive analysis, first we develop a linear regression model to predict Personal Consumption Expenditures(PCE) as dependent variable based on other economic indicators(independent variables), namely: Personal Savings Rate (`psavert`), Median Duration of Unemployment (`uempmed`), Total Unemployment (`unemploy`), and Population (`pop`).

We also split the data into training(80%) and testing(20%) sets.

```
library(caret)
```

```
Loading required package: lattice
```

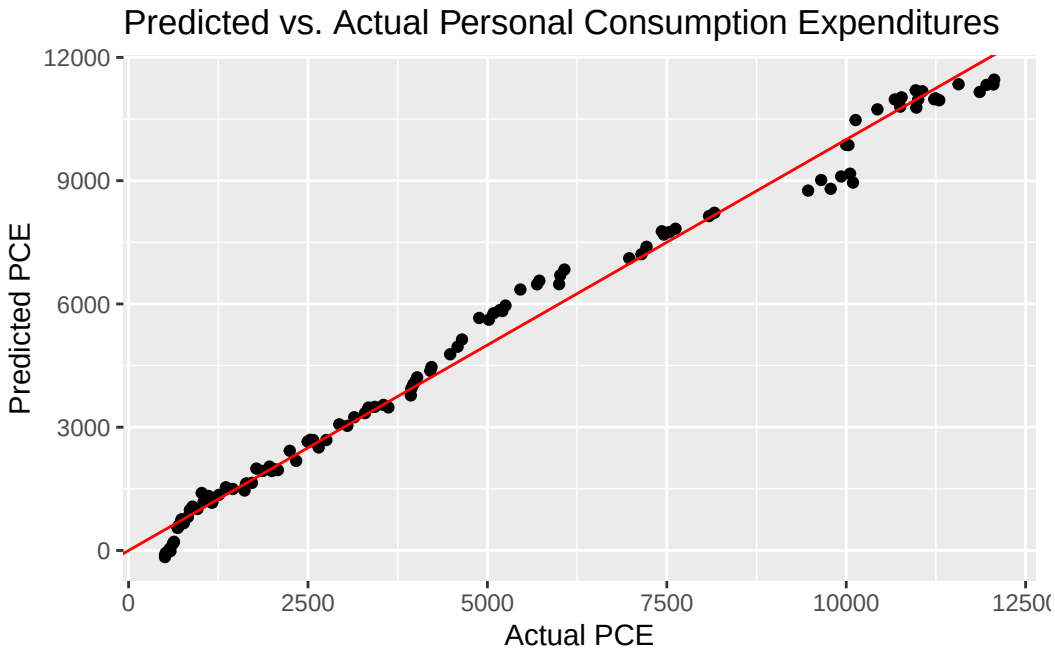
```
# Split the dataset into training and testing sets (80% training, 20% testing)
set.seed(123)
train_index <- createDataPartition(data$pce, p = 0.8, list = FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
```

```r
# Train the linear regression model
lm_model <- lm(pce ~ psavert + uempmed + unemploy + pop, data = train_data)

# Predict PCE on the testing set
predictions <- predict(lm_model, newdata = test_data)

# Evaluate the model performance
mae <- mean(abs(predictions - test_data$pce)) # Mean Absolute Error
rmse <- sqrt(mean((predictions - test_data$pce)^2)) # Root Mean Square Error

# Visualize the predicted vs. actual PCE
ggplot(test_data, aes(x = pce, y = predictions)) +
  geom_point() +
  geom_abline(color = "red") +
  labs(title = "Predicted vs. Actual Personal Consumption Expenditures",
       x = "Actual PCE",
       y = "Predicted PCE")
```




```r
# Print evaluation metrics
print(paste("Mean Absolute Error (MAE):", round(mae, 2)))
```

```
[1] "Mean Absolute Error (MAE): 291.77"
```

```r
print(paste("Root Mean Square Error (RMSE):", round(rmse, 2)))
```

```
[1] "Root Mean Square Error (RMSE): 396.58"
```

Here we can see the plot of Predicted PCE on the y-axis vs Actual PCE on the x-axis. From the plot, we can infer our linear model is actually good as the predicted values are close to the actual values. This is also reinforced by the fact that the Mean Squared Error(MSE) value is also relatively low.

4.2 Time Series Forecasting

We use the `auto.arima` model from the **`forecast`** package in R. The model automatically selects optimal parameters based on the data. Here we try to forecast the PCE values using historical data.

```r
# Load necessary packages
library(forecast)
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

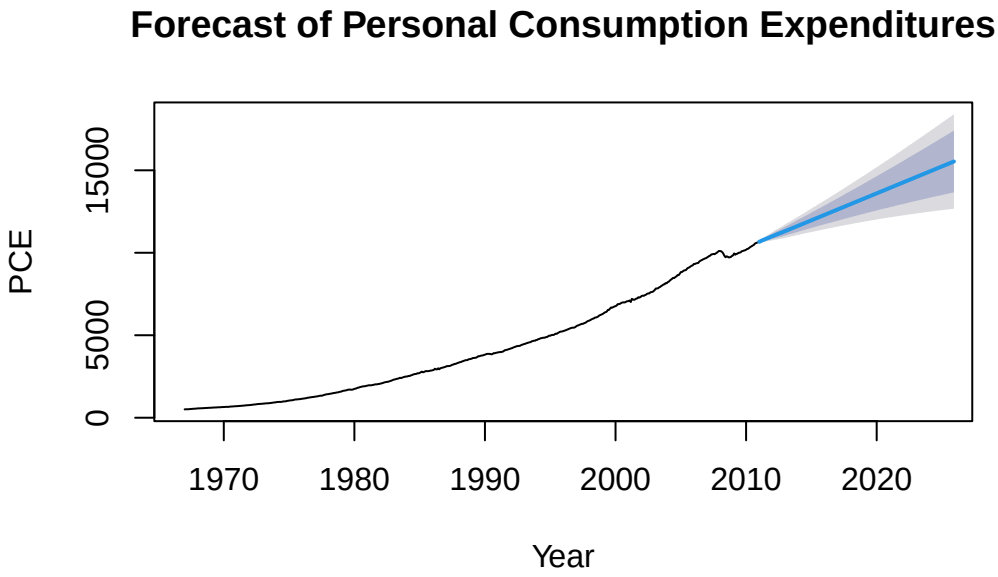
```r
# Prepare time series data for PCE
ts_data <- ts(data$pce, start = c(1967, 1), end = c(2015, 12), frequency = 12)

# Split time series data into training and testing sets
train_ts <- window(ts_data, start = c(1967, 1), end = c(2010, 12))
test_ts <- window(ts_data, start = c(2011, 1), end = c(2015, 12))

# Fit AutoARIMA model
auto_arima_model <- auto.arima(train_ts)

# Forecast future values
forecast_values <- forecast(auto_arima_model,
      h = 180) # Forecasting 15 years (180 months) ahead

# Plot forecasted values
plot(forecast_values, main = "Forecast of Personal Consumption Expenditures",
     xlab = "Year", ylab = "PCE", xlim = c(1967, 2025))
```



As seen in the plot we have a general upward trend. Thus we expect to see an upward trend in the forecast as well which we see in the plot. We trained the model on the data upto the year 2010 and have made the forecast on from the year 2011 to 2025. The shaded area around the forecasted values is the 95% confidence interval that the true future values fall within this range.

5 Conclusion

In this document, we did a comprehensive descriptive and predictive analysis of the U.S Economic Data. We uncovered key patterns and variations in the Personal Consumption Expenditures, Personal Savings Rate, and Median Duration of Unemployment, and Population. Similarly, leveraging linear regression and time series forecasting models, we demonstrated the potential for pediction of future economic conditions. These findings highlight the importance of data analysis in economic planning and policy making.