|  |  |
| --- | --- |
| A picture of a winding road and trees  Forecasting MODELS: PCE Consumption  [Document subtitle] | Abstract  This study analyses **U.S. Personal Consumption Expenditure (PCE) data** using three forecasting models: **Drift Method, Holt’s Exponential Smoothing, and Auto ARIMA**. After **data cleaning, missing value imputation, and trend analysis**, each model was trained and tested for accuracy using **RMSE and MAE metrics**. Results indicate that **Holt’s Exponential Smoothing** outperformed other models, providing the most reliable forecasts. The study highlights the **effectiveness of time series forecasting techniques** in predicting future consumption trends, aiding economic planning and decision-making.  Shreshth Sharma |

Table of Contents

**PART 1:**

**1. Introduction Executive Summary3**

**2. Data Preparation** 3

2.1 Data Cleaning and Data Visualization 3

**3. Method Selection & Training** 10

**4. Model Development & Selection11**

4.1 Simple Forecasting 11

4.2 Exponential Smoothening 12

4.3 ARIMA 13

**5. Data Evaluation15**

5.1 Forecast accuracy 15

5.2 Data Validation 15

5.3 Forecast with Rolling Origin & Comparison 16

5.4 Forecasts Results for OCT 2024 16

**PART 2:**

**6. Introduction17**

**7. Data Preparation17**

**8. Sentiment Analysis21**

8.1 BING Method21

**9. Topic Modelling And Corpus Formation 21**

9.1 Corpus Creation22

9.2 Positive Corpus22

9.4 Negative Corpus 25

**10. Data with Topic Tables28**

**11. Top 3 Factors29**

**12. Analysis using Clustering in Json’s Plot30**

**13. References36**

**14. Appendices34**

**Note:** Check For explanation under Appendices Section for yellow highlighted data.

1

# US Personal Consumption Data – Forecasting Models 1. Introduction

In this analysis, we use the seasonally-adjusted PCE data from the United States, sourced from a CSV file named "PCE.csv," to compare the predictive capabilities of three distinct forecasting models.

2. Data Preparation:The analysis begins with loading the PCE data from the "PCE.csv" file. Initial steps include:

2.1 Data Cleaning and Data Visualization:Checking for and handling missing values, outliers, or erroneous entries. Plotting the data to understand its characteristics, such as trends and seasonality.

**Summary for Dataset**

A screenshot of a computer

Description automatically generated

Summary of Personal Consumption Expenditure states the data has **779 observations** with 2 columns as “DATE” (Month of Observation) and “PCE” (Personal Consumption Expenditure).

A graph with numbers and lines

Description automatically generated

Box Plot – PCE for US consumers.

In the Box-Plot above we can see the median value is close to 5000 and it is evident that more than 50% of data lies below 5000.

In column PCE the minimum and maximum recorded value is **306.1** and **18858.9** with a mean value at **5792.1**. This raise suspicion about the major data being near to initial values and in later years the consumption Expenditure have experienced a dramatic increase.

Upon Critical Observation we can observe there are no outliers. And majorly the data is spread in the lower section of Box-Plot.



A graph with a blue square

Description automatically generated

A diagram of a graph

Description automatically generated with medium confidence

### Missing Data:

From the summary, section we can see that there are **43 *NA’s*** values in the PCE column.

Blue text on a white background

Description automatically generated

A graph with a blue line

Description automatically generated

There are many ways to treat missing data, like deletion and imputation. Deletion would account to **loss of 5.5%** of data which would influence our analysis negatively.

Imputing missing values in time series data is a crucial step in data preprocessing to ensure accurate analysis and forecasting. In R Impute-Package specializes in univariate time series imputation, providing various techniques such as linear/nonlinear interpolation, decompositions, and Kalman filtering to fill irregularly spaced series gaps (Nickolas & Shobha, 2021).

A graph with red and blue squares

Description automatically generated

Pre-Processing

Analysing from the gap size plot we can conclude that there are majorly 1-gap i.e. 39 NAs at a time between data and with just two 2-gap missing data resulting in a total gap of 4 NAs.

Also, It can be observed that under ACF plots values are way above the significance boundaries. Hence, can be predicted well enough even after 35-lags. This shows there’s a significant correlation between values.

With Autocorrelation Plot we can observe that every data is closely co-related to next coming data and there is a linear trend of corelation between the data points.

As seen in all these plots as well just like above PCE has exhibited quite a linear trend with respect to months showing no seasonality.

A graph of a number of data

Description automatically generated with medium confidence

**Handling Missing Data:** For linear trends we can use simple interpolation or kalman. na\_kalman() is used when you need a more sophisticated approach that can capture complex patterns particularly for high-frequency time series data.

Handling Missing Data

**Method Chosen: Simple Interpolation**

**Reason:**

1. As the trend is quite linear and there seems to be no seasonality.

2. Missing values are infrequent and not systematic.

3. Upon comparison Simple-Interpolation offers best imputations for missing values.

4. Save Computational-Overheads comparing to other methods.

**Imputed Values** stored in the dataset fits well in the graph.

A table with text on it

Description automatically generated

A table with numbers and a few letters

Description automatically generated with medium confidence A graph with a line going up

Description automatically generated

### Season Plot.

A graph with different colored lines

Description automatically generatedA chart with colorful circles with Royal Albert Hall in the background

Description automatically generated A table of numbers with different colored lines

Description automatically generated

Seasonality

PCE has exhibited quite a linear trend with respect to seasons and months.

So, we can declare the data has **no seasonality** over the given time-period.

Sub-Series & ACF Plots**:**

Cyclicity

For PCE over the years there is a linear trend across all months with no major fluctuation showing no sign of cyclicity in the data.

With Autocorrelation-Plot we can observe that every data is closely co-related to next coming data and there is a linear trend of correlation between the data points.

A graph of a number of months

Description automatically generated with medium confidence

A graph showing the growth of a company

Description automatically generated

A graph with lines and numbers

Description automatically generated

### 2.2 Method Selection & Training:

Here PC\_SI is the personal consumption expenditure time series with missing values imputed using function na\_interpolation().

**Method Selection Process:**

* **Simple Forecasting:** The drift method is applied to time-series data where the trend is continuous, and it helps in capturing the underlying linear pattern in the data (Zulkifle et al., 2022).
* **Exponential Smoothening:** Holt's method is effective for forecasting trends in time series data and is widely used in practical fields for its ability to handle trended time series (Chatfield, 1978).
* **ARIMA models:** Auto ARIMA has been shown to outperform manual ARIMA in terms of determining the appropriate ARIMA parameters(p, d, q), based on measures such as root mean square error (RMSE), mean absolute error (MAE) without a manual intervention of an expert data scientist (Al-Qazzaz & Yousif, 2022).

**Train Set** represents data trained for the models using **80%** of the initial data of the imputed time series.

Train set is represented by **“train”.** Inside variable **“train”** we have stored a subset of our imputed time series **“PC\_SI”. Here “end = 620”** denoting the subset has first 620 observations out of 779 observations.

**Test Set:**

Using next **20% observations** would be stored in the test set over which model accuracy will be tested. The observation under test sets would be compared with the imputed time series to find the best performing models out of Drift, Holt’s and auto.arima. **“test\_d”** denotes the test data stored using drift method whereas **“test\_h”** denotes holt’s and **“test\_a”** represents data stored using auto.arima method.

A screenshot of a computer code

Description automatically generated

## 3. Model Development and Selection

### 3.1 Simple Forecasting: Drift method

**Reason:**

As our data-trend is quite linear and The drift method of forecasting is beneficial for capturing linear trends in data *Pwasong & Sathasivam (2015).*

**Observation:**

In the first plot we can see the forecast done for next 40 periods are quite acceptable and somehow tries to justify the trend.

**Testing:**

While testing drift after training model it **doesn’t capture trend properly.**

**Accuracy:**

Model accuracy can be measured by **RMSE and MAE values** in the green cubical shapes.

**Simple Forecasting: Drift’s Method**

A graph of a graph showing the growth of a number of people

Description automatically generated

A graph with a blue line

Description automatically generated

A number of numbers and letters

Description automatically generated with medium confidence

3.2 Exponential Smoothening Method: Holt Method**:** Holt's method involves two smoothing parameters, α and γ, which allow for capturing both the level and trend in the data *Trull et al. (2020).*

A graph showing the growth of a number of individuals

Description automatically generated

**Reason:**

Holt’s Method is commonly used for forecasting trends in time series data that exhibit a trend component. *(Oni & Akanle, 2018).*

**Observation:**

In the first plot we can see the forecast done for next 40 periods are better than drift and captures trend better.

**Testing:**

While testing we can observe actual trend lies in the trained model range with **95% confidence interval.**

**Accuracy:**

**RMSE and MAE** values in the yellow cubical shapes below seems to be better than “drift’s”.

**Exponential Smoothening: Holt’s Method**

A graph with a line

Description automatically generated

A number and numbers on a white background

Description automatically generated

### 3.3 Arima Models: Auto Arima – [Auto Regressive order = 2, Differencing Order = 2, Moving Average = 1]

**ARIMA: Auto.arima Method**

A graph showing the growth of a number of years

Description automatically generated

**Reason:**

Auto Arima performs better than normal ARIMA models.

**Observation:**

In the first plot we can see the forecast done for next 40 periods are fairly capturing the trend similar to Holt’s model. To make series stationary it has been differenced 2 times.

**Testing:**

After Training we can observe the auto Arima model seems to capture less trend with the same confidence interval than the Holt’s

**Accuracy:**

**RMSE and MAPE** in the purple cubical shapes below seems to be more than ones in Holt’s.

A graph with a line

Description automatically generated

A number with numbers and letters

Description automatically generated with medium confidence

**Checking Residuals for test set after using ARIMA.**

A screenshot of a graph

Description automatically generatedA screenshot of a computer

Description automatically generated

Arima there are just 2 spikes out of the significance level in ACF plot and data is also normally distributed in PACF. It almost represents a white noise pattern and **P-value is significant** so the null hypothesis will be rejected i.e. there is **no correlation between residuals**. Hence, ARIMA models fits good with the data and evidently will perform well while forecasting future values.

## 4. Model Evaluation: Compare Models

### 4.1 Forecast Accuracy

**Selection Criteria:** RMSE (Root of the mean squared errors) and MAE (mean of absolute errors) scores can help in identifying the best model. *Kouadri et al. (2021)* emphasized the significance of RMSE as a predictive numerical index for measuring model performance in time series forecasting. Hence, we are looking for **the models with least values of MAE and RMSE**.

**A table with numbers and text

Description automatically generated**

**Holt’s Model** stand out and performs best with the **least RMSE and MAE** values with the test set.

4.2 Model Validation:Graphical analysis to compare the forecasts from each model against the actual data.

A graph showing the growth of a model

Description automatically generated

### 4.3 One Step Ahead Rolling Forecast:

**A computer screen shot of a program

Description automatically generated**

**Results:**

**A table with numbers and text

Description automatically generated**

**Holt’s Model** stand out and performs best with the **least RMSE and MAE**.

### 4.4 Forecast For OCT 2024:

In both cases, Holt’s seems to perform better and forecast for

A table with numbers and a black text

Description automatically generated