# **Data Science Project**

Research Question:

***Will the Corruption Perception Index Score for the United Kingdom over next 5 years be higher on mean average than France or Germany?***

Contents

[**Data Science Project** 1](#_Toc175750669)

[Table of Figures 3](#_Toc175750670)

[Executive Summary 4](#_Toc175750671)

[Introduction 4](#_Toc175750672)

[Data Source and Preparation 4](#_Toc175750673)

[Exploratory Data Analysis 6](#_Toc175750674)

[Time Series Analysis of the CPI Scores 7](#_Toc175750675)

[Time Series Forecast Model 8](#_Toc175750676)

[Evaluating the ARIMA Models 9](#_Toc175750677)

[The Result 10](#_Toc175750678)

[Future Recommendations 11](#_Toc175750679)

[Conclusion 11](#_Toc175750680)

# Table of Figures

[Figure 1 - Snippet of bar chart available on Statista which demonstrates GDP per millions euros 4](#_Toc175750304)

[Figure 2 - This is CPI data that was loaded back into Kaggle after being read in by Pandas 5](#_Toc175750305)

[Figure 3- This shows code about changing the Year data type from int64 to a datetime. 5](#_Toc175750306)

[Figure 4 - filtering the data frame to each separate country. 6](#_Toc175750307)

[Figure 5 - the code used in python to calculate the central tendencies. 6](#_Toc175750308)

[Figure 6 - a for loop that was created to show all three countries together in a time series. 7](#_Toc175750309)

[Figure 7 - the results of the code in Figure 6 7](#_Toc175750310)

[Figure 8 - the code used to input data into the UK model and fit it to the data. 8](#_Toc175750311)

[Figure 9 - Summary of the ARIMA model for the UK. 9](#_Toc175750312)

[Figure 10 - Summary of the ARIMA model for France. 9](#_Toc175750313)

[Figure 11- Summary of the ARIMA model for Germany. 10](#_Toc175750314)

## Executive Summary

The UK is facing concerns around its corruption levels in the upcoming year, and it is essential that the future results can be predicted to estimate the severity of the possible increase so government can be in place measures to prevent this. To research this, a time series forecasting model was built for the UK as well as France and Germany so that the UK could be cross examined with economically competitive countries. The outcome of the project was that based on the forecasting the UK will have a higher mean average however, there is reasons this should not be accepted at face value and more historical data and other metrics should be used to strengthen the case.

## Introduction

“Corruption undermines trust, weakens democratic institutions, impedes economic development, and exacerbates inequality, poverty, social division, and environmental crises.” Transparency (2024). Despite its significant impact, corruption remains a taboo and sensitive topic that is not widely discussed in the media. The Corruption Perception Index (CPI), defined as the perceived level of public sector corruption in each country, assigns an annual score to countries worldwide based on at least three sources. This score facilitates a comparative analysis of corruption levels across nations. A low corruption score is indicative of a higher quality of life, characterised by low inflation and substantial investment in education.

In 2024, Transparency International UK reported a major headline: “Concerns of corruption at an all-time high as the UK falls to its lowest ever score on the global corruption perceptions index.” Transparency International UK (2024). This development is alarming, highlighting the necessity of predicting future CPI scores using time series forecasting to assess the severity of required changes to improve the UK’s perceived corruption. Comparative analysis with other countries is essential in this context. France and Germany have been selected for comparison due to their similar GDP figures in million euros, as illustrated in a chart by Statista (2022) in Figure 1.

A blue and white bars

Description automatically generated

Figure 1 - Snippet of bar chart available on Statista which demonstrates GDP per millions euros

## Data Source and Preparation

The data source was obtained from Kaggle, which aggregated information from multiple origins and compiled it into an Excel file. Instead of performing pivoting and editing operations within Kaggle, it was deemed more efficient to download the Excel file, filter, and pivot it using Excel software. This approach facilitated the removal of unnecessary columns that would otherwise be included in the data frame, thereby enhancing computational efficiency. Subsequently, the pivoted Excel file was re-uploaded to Kaggle for further analysis. This decision was influenced by Kaggle’s capability to support Python coding, which is preferable for time series analysis due to its integrated statistical modelling packages. Python is extensively used for such statistical modelling, removing the need to create numerous equations and functions within Excel. Additionally, this method ensures that the model and visualisations are consolidated within a single workbook, as opposed to being dispersed across various data tools.

A table with numbers and a number on it

Description automatically generatedOnce the data was read into Kaggle it was important to investigate if any transformation needed to occur so the time series analysis would be successful.

Figure 2 - This is CPI data that was loaded back into Kaggle after being read in by Pandas

After looking at the data it was important to check the data types of the columns as shown in Figure 3 and change them if required.

A screenshot of a computer

Description automatically generated

Figure 3- This shows code about changing the Year data type from int64 to a datetime.

As the chosen statistical model involves time it is important the time element of the data is an appropriate data type.

## Exploratory Data Analysis

Firstly using. loc() , three separate data frames were created each filtered to the UK , France, and Germany.

A close up of a white background

Description automatically generated

Figure 4 - filtering the data frame to each separate country.

Figure 5 shows the results of different central tendencies for each country of interest.

A screenshot of a computer code

Description automatically generated

Figure 5 - the code used in python to calculate the central tendencies.

Several inferences can be made on the results of this:

* Mean = The UK and Germany have a similar score which implies that from this data they are likely to have a similar range and values over time which is important to know for forecasting.
* Median = France has the lowest median meaning that it could be inferred they are likely to have a smaller range and will have more consistent forecasted cpi scores.
* Mode = Germany has the highest mode count and the highest modal value which is also higher than it is mean score which suggests the data is negatively skewed towards higher CPI scores whereas the UK and France have a lower mode in comparison to their means suggesting they are positively skewed.

## Time Series Analysis of the CPI Scores

A screenshot of a computer program

Description automatically generated

Figure 6 - a for loop that was created to show all three countries together in a time series.

Figure 6 shows the code used to create a time series which unearthed some interesting patterns in the data.

A graph of different colored lines

Description automatically generated

Figure 7 - the results of the code in Figure 6

As discovered with the central tendencies Germany and the UK are closer in scores than France and this is only enhanced in the visualisation. If you look at the overall trends of all the countries 2020 is a significant year for a change in corruption scores.

From a qualitative perspective, this could be down to the global pandemic which reinvented the normalcy of many public sector organisations globally which has led to a very out of pattern score for each country with Germany been more positively affected in terms of corruption than France. However, as there is no qualitative research included in this dataset and the year 2020 has had a drastic impact on all the countries in the time series, it should be kept in the data as there is no evidence to suggest this is an anomaly and that it isn’t likely to happen again.

To improve this time series and enrich the understanding of the drastic difference research involving Covid-19 could be added to the data to improve the understanding of the reasons behind the fluctuations.

## Time Series Forecast Model

To answer the research question, a time series forecasting model needed to be selected to input the data into to get the desired output. There are several important decisions that led to the decision that an Autoregressive Integrated Moving Average (ARIMA) time series model would be the best suited. These are:

* The data is univariate – which meant that a machine learning algorithm was not necessary for the project to get an outcome unlike if the data was multivariate.
* ARIMA is the most popular classical method which meant even as a starting point this a good place to begin exploring the question at hand.

A white rectangular object with a black border

Description automatically generated

Figure 8 - the code used to input data into the UK model and fit it to the data.

To see the full code used to build each ARIMA models for each country it is posted in the following GitHub link: <https://github.com/alice-hayes/Data-Science-Projects-/blob/main/cpi-forecasting.ipynb>

The crucial decision regarding ARIMA is to decide the d,p,q for each use case. This was done using Auto Correlation Graphs and an Augmented Dickey-Fuller (ADF) test which allows for testing against the p – value to find the optimal number.

These metrics are key in improving the model’s accuracy and susceptible to change.

## Evaluating the ARIMA Models

**A screenshot of a data

Description automatically generated**

Figure 9 - Summary of the ARIMA model for the UK.

A screenshot of a computer

Description automatically generated

Figure 10 - Summary of the ARIMA model for France.

A screenshot of a computer

Description automatically generated

Figure 11- Summary of the ARIMA model for Germany.

One key metric is the Akaike Information Criteria (AIC) which looks at how good the model fits the data, meaning the lower the score the better. In the examples above, the model fits the UK the best and Germany the worst. This is likely due to the bigger peaks and troughs experienced by Germany especially around 2020. The aim is to have the lowest AIC score possible which is done by fine tuning the d,q and p parameters of the back of this evaluation fine tuning seems a good option to improve the accuracy for the model for future predictions. This reasoning is backed up by the log likelihood value which reflects the same.

However, after looking at this summary ARIMA may not have been the optimal choice and therefore in the future, other models should be tried like Exponential Smoothing which is also suitable for univariate data.

The variance has been measured to understand the spread of the values from the mean to see if it is better at predicting the forecasted or the actual values. The UK had similar variances across both, but France and Germany had huge variance differences between the actual and the forecasted with this being lower. It suggests that there is too much weighting on values when fitting models which is cause for concern in terms of accuracy and the face validity of the model.

## The Result

Upon calculating the mean averages of the forecasted values for each country, the models indicate that the United Kingdom will exhibit a higher mean average compared to its competitors. However, this conclusion should not be accepted solely based on these univariate ARIMA models. Additional data and further fine-tuning are necessary to achieve a higher level of accuracy.

## Future Recommendations

Currently, the dataset is univariate, consisting of a “single time-dependent variable” (Moez Ali, 2022) and predictions are based solely on the year. The project could evolve into a multivariate analysis by incorporating additional data related to the CPI score, such as GDP, political stance, or happiness score, enhancing predictive accuracy.

Global changes, like those in 2020, which are unpredictable and cannot be modelled, will impact future CPI scores. Thus, results should be communicated as accurate forecasts assuming no significant changes.

To refine the project, comparing the UK against all countries in the CPI dataset would provide a more comprehensive assessment of the UK’s global standing and its relationship with public sector corruption.

## Conclusion

The result of research suggests that the UK need to implement severe changes to its public sector to ensure it stays competitive with France and Germany. The research uses time series forecasting to predict this outcome yet tis cannot be accepted as the definitive result. The problem with this project is the models do not have a high enough accuracy to allow a confident answer and there are several factors that affect the CPI Score which have not been incorporated into the model, so they will not count towards the score. To further improve this project, these data points should be included to create a well-rounded and accurate score. The year 2020 breaks the pattern and trend for each country and will have a massive impact on the results it could be considered to remove this ear from the data and see how drastically this impacts the score.