**Thomas Parmentier**

Zypressenstrasse, 49

8004 Zürich

Thomas.parmentierlux@gmail.com

**Stefan Reinhard**

Sandstrasse 13

3302 Moosseedorf

Stefan.reinhard@me.com

**Data Science Project**

**The impact of macroeconomic variables on the Nasdaq composite index price**

**Conceptual Design Report**

**31 October 2021**

# Abstract

The dynamic relationship among macroeconomic indicators such as gross domestic product, interest rate, unemployment, and inflation are crucial to understand stock market performance. Any unexpected or abrupt changes in economic indicators can lead to high market volatility and shift in market behavior. Actor of financial market needs to understand how these variables interacts with each other. Several studies have been conducted to determine the relationship between macroeconomic variables and stock prices in the past, with mixed results. The objective of this study is to discover the impact of changes in macroeconomic variables on the Nasdaq composite index price performance. For this purpose, we have performed a unit root test as well as a Multivariate Regression Model computed on the OLS (Ordinary Least Square) method. The time-period analyzed range from 2012 to 2019 with monthly data. Our results show that higher inflation, unemployment rate and interest rates lead to lower equity markets, but only inflation has a significant influence on equity markets. While a higher GDP leads to higher equity markets, but is not significant based on our model.

# Table of Contents

[Abstract 0](#_Toc86169134)

[Table of Contents 1](#_Toc86169135)

[1 Project Objectives 2](#_Toc86169136)

[2 Methods 3](#_Toc86169137)

[3 Data 4](#_Toc86169138)

[4 Metadata 7](#_Toc86169139)

[5 Data Quality 8](#_Toc86169140)

[6 Data Flow 9](#_Toc86169141)

[7 Data Model 10](#_Toc86169142)

[7.1. Conceptual model 10](#_Toc86169143)

[7.2. Logical 10](#_Toc86169144)

[7.3 Physical data model 11](#_Toc86169145)

[8 Risks and limitations 11](#_Toc86169146)

[8.1 Risks 11](#_Toc86169147)

[8.2 Limitations 12](#_Toc86169148)

[9 Preliminary Studies 13](#_Toc86169149)

[10 Conclusions 20](#_Toc86169150)

[Appendix A: Link to GitHub 20](#_Toc86169151)

[Appendix B: Decomposition analysis 20](#_Toc86169152)

[References and Bibliography 22](#_Toc86169153)

# 1 Project Objectives

The objective of this report is to evaluate the relationship between four macroeconomic variables (gross domestic product, inflation, unemployment rate, interest rate) and the Nasdaq composite index. Stock market participants are looking for “trading signals” to define if they must buy or sell the markets, economic indicators are among the most famous signals market participants are using. The trading strategies based on macroeconomics have even become a whole industry among Hedge Fund. Global Macro Hedge Funds aim to actively managed funds from markets move caused by economic indicators and political events, this strategy manages around 800 billion USD and has made famous some Hedge Funds managers like George Soros or Steve Cohen [1], which have bet on currency movements.

With this project, our objective is to look through four independent variables, which will be described in more details in the section 3.

**We will answer the following research question: Can we see a meaningful relationship between the four independent variables selected and the Nasdaq composite index?**

To achieve this objective, we will perform the following tasks:

* Extract, clean and combine information of time-series on macro-variables with stock market time-series,
* Visualize the data and run some inferential statistics to understand our data set,
* Run the Augmented Dick Fuller (ADF) test to check for non-stationary,
* Run the OLS (Ordinary Least Square) analysis,
* Analysis the residuals (Autocorrelation, Heteroscedasticity and Normality) to check for white noise.

# 2 Methods

**Database:** We used the Saint Louis FRED (Federal Reserve Economic Data) database [2] to extract macroeconomic data. It is one of the most comprehensive and trustful sources of economic indicators in the US. This database is widely used and reported in the media and financial market participants look closely at it every time new data is released. The database is also publicly available. From the FRED we have extracted time-series for our four independent variables, which will be described in the section 3 (US GDP, US CPI, US interest rates and US unemployment rate).

For our dependent variable, we used data from the Nasdaq website [3], which give the daily time-series of the price of the Nasdaq composite index.

**Tools:** We use Python and the Jupyter notebook environment to write a program to analyze our data set. The program will import, clean, consolidate, arrange, and merge the data, allowing for statistical and visualization analyses. The data are downloaded in CSV and then loaded into Jupyter notebook.

**Libraries:** We have used numerous libraries to perform our analysis, here we will list the most important.

* **Pandas** [4]is a library for Python allowing data manipulation and analysis. Pandas offers data structures and operations for manipulating numerical tables and time-series,
* **Numpy** [5]allows for manipulation of matrix and multidimensional tables. It also allows to perform a wide range of mathematical operations on these tables,
* **Matplotlib** [6]is a library used to make graphics and visualization of data,
* **Statsmodels** [7]is a python library providing a wide range of statistical models to conduct statistical data exploration and statistics,
* **Scikit-learn** [8]is a machine learning library built on Numpy, Scipy and matplotlib and is used to perform predictive data analysis.

We also used other libraries for data manipulation such as request and OS.

**Statistical methods**: At first, we will use descriptive analysis and graphics to understand our data set. Secondly,when working with time-series it is important to check if the data is stationary. For that purpose, we will use the ADF (Augmented Dickey Fuller) test and then perform first-difference (or more) if necessary, on our data set. Then, we will use the OLS method to evaluate the impact of macro variables on the Nasdaq composite index. A last we will analyze the residuals for white noise.

# 3 Data

As describe above, the data were obtained through FRED database [2] and the Nasdaq exchange [3]. We will work with monthly data which represent 95 observations for four independent variables from January 2012 until December 2019:

- **US GDP:** We use the US GGP in percentage change from preceding period, monthly, seasonally adjusted on an annual rate. The US GDP refers to the total amount of goods and services produce in the US. It is a snapshot of the economy at a certain point in time. The US GDP has moved around 4% over the time-period selected with a robust growth between 2016 and 2018 following looser regulation and a favorable tax environment.

**Figure 1: Historic of the GDP over the observation period**

Chart

Description automatically generated

- **US** **CPI (Consumer Price Index):** We use the US CPI in percentage change, from the same period previous year, monthly. The US CPI gives the current price of a selected basket of goods and services that is updated periodically. The US CPI has moved around 2% to 2.5% over the time-period selected.

**Figure 2: Historic of the CPI percentage change over the observation period**

Chart, line chart

Description automatically generated

- **US Interest rates:** We will use effective Federal Funds Rate in percentage, monthly. Interest rates have been extremely low following the global financial crisis, up to 2016 where the FED decided to tighten its monetary policy. The easing of the monetary policy in 2019 was due to the market crash of December 2018 following the trade war between China and the US and the slowdown of the US economy.

**Figure 3: historic of the FED funds rate over the observation period**

Chart, line chart

Description automatically generated

- **US Unemployment rate** in percentage, monthly, seasonally adjusted. The [unemployment rate](https://www.investopedia.com/terms/u/unemploymentrate.asp) shows how many people from the available pool of labor are not able to find a job. US unemployment rate constantly went down following the global financial crisis to reach a low point of 3.8%, end 2019.

**Figure 4: historic of the unemployment rate over the observation period**

Chart, scatter chart

Description automatically generated

As a first observation we can already see that some of the series are not stationary. CPI looks stationary as the mean and the standard deviation look constant and there is no clear trend, while for the unemployment rate we can see that the mean is not constant. We will confirm all of that in the statistical analysis when we test for unit root.

Our **independent variable** is the Nasdaq composite index. The Nasdaq composite index had an impressive performance over the selected period, with a drop in December 2018 and a catch back beginning 2019 following the decrease of the interest rates and the rebound of the US economy.

**Figure 5: performance of the Nasdaq index over the observation period**

Chart, line chart

Description automatically generated

To put everything on the same scale we have converted the daily Nasdaq composite index price in monthly return by using the following code:

**Table 1: Excerpt from our python code**

Text

Description automatically generated

Our dependent variable (y) will be **‘Nasdaq composite index return’**

Below is the data set, we will work with to perform our statistical analysis. Data quality will be asset in the section 5 of this report.

**Table 2: Excerpt of the dependent and independent variables**

Graphical user interface, text

Description automatically generated

# 4 Metadata

All data including metadata is stored on a GIT repository and publicly accessible. Metadata and other descriptions are briefly described in a readme file (readme.md) on the repository and also detailed below. The results, plots and tables can be created and/or reproduced using the supplied script (Jupyter Notebook).

We use the FRED database, which is the database from the Federal Reserve of Saint Louis. The database use filters that have been parsed out to their most finite terms. These metadata allowed us to sort through the information of the database quickly. For example, there are 78,211 series in FRED that are associated with GDP. The FRED allows to break down the data in eight sections:

- Concepts

- Geography

- Frequencies

- Sources

- Releases

- Seasonal Adjustments

- Citation & copyright

Below we describe the filters used for the four independent variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metadata filters** | **US GDP** | **US CPI** | **US Interest rate** | **US Unemployment rate** |
| Concepts | Gross Domestic Product | Consumer price inflation | FED fund rate | Unemployment rate |
| Geography types | Country | Country | Country | Country |
| Geographies | US | US | US | US |
| Frequencies | Monthly | Monthly | Monthly | Monthly |
| Sources | n.a. | n.a. | n.a. | n.a. |
| Releases | Bureau of Economic Analysis | Bureau of Economic Analysis | Bureau of Economic Analysis | Bureau of Economic Analysis |
| Seasonal Adjustements | Seasonally adjusted | Seasonally adjusted | No | Seasonally adjusted |

We have saved the metadata in xls files available on our github, alternatively the metadata are available on the website of the FRED of St Louis by using the below process:

Step 1: reach out to research.stlouisfed.org and select FRED  
Graphical user interface, website

Description automatically generated  
Step 2: Input the name of the independent variable

Graphical user interface, text, application

Description automatically generated

# 5 Data Quality

The quality from the point of view of the plausibility of the source data is already guaranteed by the providers. The data are initially examined in a descriptive analysis to identify possible quality deficiencies.

For the present analysis, the integrity of the time-series is the highest quality criteria. Therefore, it was examined whether there was a data point in all tables for each month in the study period and whether there were no NAs. The descriptive analysis ensured the validity of the data quality. No further steps are required to improve the quality of the data. We can see below that the quality of our data is high.

**Table 3: Excerpt from “.info” in python**

Text

Description automatically generated

# 6 Data Flow

In the first step, the data is retrieved from the provider's website [2], [3], downloaded and saved in the project as a csv file. The data sets are cleaned up, simple adjustments and transformations (e.g., date) are performed, and the sources are merged into one data frame. Next, the descriptive analysis can be performed, and first outputs are generated. Before the regression can be performed, the independent variables must be tested for unit root and transformed, if necessary (first/secondary difference), resulting in further figures. Afterwards, the regression is performed with its output and finally the residuals are examined and plotted.

**Figure 6: Data flow**

Diagram

Description automatically generated

# 7 Data Model

Presentation of the data model at the conceptual level, the logical level, and the physical level.

## 7.1. Conceptual model

The conceptual data model shows the Nasdaq index as the dependent variable and the relation to the independent variables (US GDP, US CPI, US Interest rates, US Unemployment rate). (Fed fund = interest rates).

**Figure 7: Conceptual data model**Diagram

Description automatically generated

## 7.2. Logical

The logical data model shows the connection of the dependent variable (Nasdaq index) and the common column (index) the date. The one-to-one connection thereof allows to join the individual variables to a 2-dimensional data frame. the data type of the date variable is date, alle all other variables are floats

**Figure 8: Logical model**

Diagram

Description automatically generated

## 7.3 Physical data model

Since the data sets are in the kilobyte segment and the analysis does not require complex calculations, no significant infrastructure is required, a standard laptop is sufficient. For the initial download of the data sets an internet access is required.

# 8 Risks and limitations

## 8.1 Risks

For the present analysis, we were able to identify three risk factors:

1. Availability of data sets at the data provider. The data sets were downloaded from publicly available providers. More recent data sets may no longer be available or might be subject to restrictions (e.g., change of license model).
2. Change of structure/formatting by the data provider. If the data service provider modifies the submitted data in future versions, e.g. different date format or transformation of percentage values to decimal values, the validity of the data model cannot be guaranteed.
3. Systemic errors caused by vague domain knowledge of underlying relationships, etc. Insufficient knowledge of the authors.

Consequences steps were taken to minimize the risks: The first two risks could be mitigated by a one-time download of the data but must be considered when updating the data. Risk three was minimized by peer review of the two authors.

## 8.2 Limitations

On the top of the above risks, we have also identified three limitations to our analysis:

1. Financial markets are affected by a wide range of factors that go beyond what we have included in our model. Our four metrics are important, but far from being the only ones relevant when one wants to understand stock market moves. At macroeconomic level, geopolitics (the trade war between US and China), weather (the earthquake in 1906 that trigger a market crash at wall street in 1907 or the impact of hurricanes on oil production), fiscal policy or presidential elections are factors that also affect the market. At microeconomic level, companies financial result is also a relevant indicator of market performance. Finally market psychological behavior also plays a lot in the performance of financial markets (e.g., herd behavior). Markets can also exhibit cycles that we have considered in our analysis [9].
2. A second essential element that matter is the methodology of calculation for the difference macroeconomic data. They are not all “perfect” and subject to emotional discussions among economists, they also change over time. For our four independent variables with have the following pitfalls [10],[11]:
   * **US GDP**: The one issue of this metric is that data must be collected within a specific period, a figure for the GDP now would have to be an estimation. The metric is backward looking.
   * **US CPI**: There is nothing less that eleven metrics of calculation to determine consumer price indexes (some do not include the cost of healthcare for example), but the CPI index is one of the most well-known and used by the Federal Reserve to dictate its monetary policy. For example, a strong increase in healthcare cost could be missed in this metric which would lead to an understatement of inflation.
   * **US interest rates:** US interest rates are set by the monetary authority, a wrong evaluation of the current state of the economy could lead to a mistake in the monetary policy. For example, an increase in interest rates too early following the COVID crisis could lead to an economic slowdown as the cost of borrowing would go up.
   * **US Unemployment rate:** This metric only includes people looking actively for a new job, others would be excluded, which could lead to an understatement of the “real” unemployment rate.
3. Last pitfall, most of these metrics are backward looking and could already miss the current state of the economy.

To solve the two limitations, we could include more macroeconomic indicators in the model. Below a list of macroeconomic indicators [12], we could evaluate our model with:

- Government Debt to GDP

- Balance of Trade

- Credit Rating

- Minimum Wages

- Non-Farm Payrolls

- Consumer Confidence

- Consumer Credit

- Corporate Profits

- Personal Spending

- Retail Sales YoY

The second limitation on the methodology calculation is tricky. The metrics we have selected are the most used by macroeconomics. We could cross-check the selected metrics by the ones using another methodology.

Regarding the last limitation, one can use alternative data to assess if one of the metrics is currently over-or under-estimating the current state of the economy. Some of these metrics are credit card issuance, traffic on the road or satellite image of activity.

The project could suffer some additional costs as collecting alternative data is usually expensive. The computing power would also need to increase to digest these data.

# 9 Preliminary Studies

Literature has widely studied the impact of macroeconomic variables on broader stock markets such has the S&P 500 or Dow Jones index, but few looks specifically at the Nasdaq composite index, which focus on technology stocks in the US. The Nasdaq has emerged as one of the big winners of the COVID-19 crisis [13]. Overall, studies output has been mixed on the impact and the strength of economic variables and financial markets. For example, even if it is commonly admitted that inflation is negative for stock markets, as demonstrated in 1979 by Fama and Schwert who found that stock returns were negatively correlated with the expected component of inflation rate [14], other researchers like Gjerde and Saettern [15], in 1999 found that inflation might not have a direct impact on stock performance. The same discussion can be made on other variables such as the GDP, interest rate level and unemployment rate. In 2021 a study [16] found that there is no ‎causal relation between macro factors and stock returns.

A first look at the descriptive statistics, that we already described a bit above, show us that we deal with 95 observations. The first step is to understand if we have any outliers and if we should use any smoothing methods (moving average, or exponential moving average) to better mitigate the effects of outliers.

**Table 4: Metrics of our dataset**

Table

Description automatically generated

To achieve this, we have used boxplot charts. As we can see on the below charts there does not seem large outliers. All the data stay within 1.5\*IQR (Interquartile range), as a result we will not exclude data from the data set.

**Figure 9: Boxplots for our independent variables and our dependent variable**

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Before running any analysis on time-series we need to ensure they are stationary and fulfill the following conditions:

* Mean is constant
* Standard deviation is constant
* No specific trend

A chart analysis can be useful to identify any seasonality in a time-series. We have use “decompose” from the package “statsmodel” to better understand what we are dealing with, the charts are in appendix B. A close look at the charts show us that the data exhibit a strong seasonality, and we will have to deal with this issue. We have used the resources highlighted in the section “References” [17] to [20] for the below analysis.

One way to determine whether if we see stationary and if differencing is required is to use a unit root test. These are statistical hypothesis tests of stationary that are designed for determining whether differencing is required. We tested for stationary using the Augmented Dickey-Fuller test.

**H0**: p-value > 0.05: Fail to reject the null hypothesis, **the data has unit root and is non-stationary.** It has some time dependent structure.

**H1**: p-value <= 0.05: Reject the null hypothesis**, the data has no unit root and is stationary.**

**Table 5: Results of the ADF test**

Text, letter

Description automatically generated Text, letter

Description automatically generatedText, letter

Description automatically generated Text, letter

Description automatically generated

Text, letter

Description automatically generated

The results suggest that we can reject the null hypothesis for the Nasdaq composite index return and CPI. Data are non-stationary for interest rate, unemployment, and GDP. We need to take the first difference of those variables.

A seasonal difference is the difference between an observation and the previous observation from the same season.



m= the number of seasons.

The charts show the first difference analysis, we can see that now the time-series exhibit at least a constant mean, and maybe a constant standard deviation? We will verify that.

**Figure 10: Charts after first differencing**

Graphical user interface, application

Description automatically generated

Once we have first difference, we can run again the Augmented Dickey Fuller test.

**Table 6: Results of the ADF test (first differencing)**

Text, letter

Description automatically generated Text, letter

Description automatically generated Text, letter

Description automatically generated

We use the same H0 / H1 hypotheses The results suggest that we can reject the null hypothesis for the GDP. Unit root is still present for Fund rate, Unemployment

We need to take the second difference of those variables and we re-run again the Augmented Dickey Fuller test

**Table 7: Results of the ADF test (second differencing)**

Text

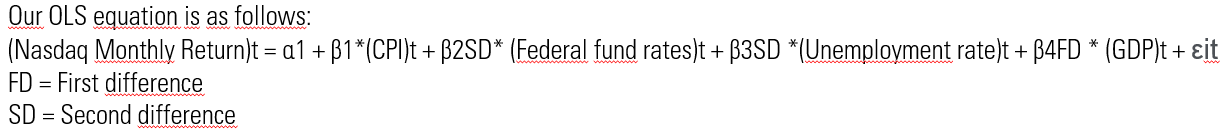
Description automatically generated Text

Description automatically generated

This time it looks like we have achieved our goal. We can reject the null hypothesis for all of the variables.

1. Multivariate Regression Model computed on the OLS

We are going to use our new variables (after differencing) in the following equation:



Using statsmodel we have computed the OLS regression.

**Table 8: OLS Regression Results**

Table

Description automatically generated

Based on the above we can conclude that CPI, Interest rate and unemployment rate are negatively correlated with the Nasaq composite index, but the GDP is positively correlated. The T-test show us that only independent variable statistically significant is the CPI. We also learn from this analysis that the R-squared and adjusted R-squared are extremely low, while the F-statistic is not satisfactory. We would not use this model for investment purpose it would have to be refine with other macroeconomic variables and also more advance machine learning methods.

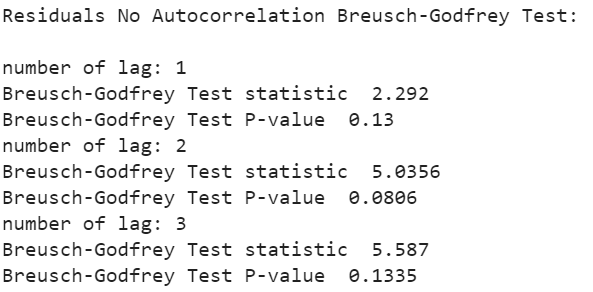
To confirm and trust the T-test results from our OLS regression, we must make sure that the residuals are white noise. Residuals from a regression should never contain any information, since this is a sign that this information is not included in the regression model. We will perform three tests: Autocorrelation, Heteroscdasticity and Normaly.

**Autocorrelation**: The presence of serial correlation is examined by Breusch-Godfrey Test. Residuals for OLS output is tested for serial correlation, using the following hypothesis:

**H0 : No autocorrelation**

**H1 : Autocorrelation**

**Table 9: Result of Breusch-Godfrey Test**



The p-value is 0.13 which is greater than critical value at 5%. We cannot reject the null hypothesis and we can conclude for the absence of autocorrelation.

**Heteroscedasticity**: This test is important to confirm the robustness of the OLS output since we cannot rely on them in the presence of heteroscedasticity:

**H0: No heteroscedasticity**

**H1: Heteroscedasticity**

**Table 10: Result of Breusch-Pagan Test**

Text

Description automatically generated

The p-value is 0.3 which is greater than critical value at 5%. We cannot reject the null hypothesis and we can conclude there is no heteroscedasticity is present, and thus OLS t-test results can be trusted.

Normality: This test is important to find out whether the error term follows normal distribution:

**H0: Residuals are normally distributed**

**H1: Residuals are not normally distributed**

**Table 11: Result of Jarque bera Test**

Text

Description automatically generated

The p-value is 0.79 greater than the critical value at the 5% level. So, the null hypothesis cannot be rejected.

# 10 Conclusions

Based on our statistical analysis, we can conclude that one out of the four selected macroeconomic variables are relatively significant and likely to influence monthly return of the Nasdaq composite index, which is the inflation. While we see no significant impacts from the other variables, the sign of direction is the one we would expected as explained in section 9. The evidence of this study is consistent with most of the other similar studies. However, the results from this empirical research should not be a conclusive indicator for investment. We could further expend the analysis, by choosing different macro-variables, using other models (Granger causality testor more complicated machine learning codes).

# Appendix A: Link to GitHub

Link to GitHub: <https://github.com/parment1/Project-M1-M2>

# Appendix B: Decomposition analysis

Graphical user interface, application, Word

Description automatically generated

**CPI:**

Chart, histogram

Description automatically generated

**Unemployment rate:**

Chart, histogram

Description automatically generated

**Interest rate:**

Chart, histogram

Description automatically generated

**GDP:**

Chart, histogram

Description automatically generated

# References and Bibliography

[1] https://hedgelists.com/top-25-global-macro-hedge-funds-2021/

[2] https://fred.stlouisfed.org/

[3] <https://www.nasdaq.com/>

[4] <https://pandas.pydata.org/>

[5] https://numpy.org/

[6] https://matplotlib.org/

[7] <https://www.statsmodels.org/stable/index.html>

[8] <https://scikit-learn.org/stable/>

[9] Finance and Macroeconomics: The Preponderance of the Financial Cycle [Michel Aglietta](https://www.cairn.info/publications-de-Michel-Aglietta--6668.htm), [Revue de l'OFCE](https://www.cairn.info/revue-de-l-ofce.htm) [2018/3 (N° 157)](https://www.cairn.info/revue-de-l-ofce-2018-3.htm), 197-224

[10] BIS Working Papers No 677 Macroeconomic implications of financial imperfections: a survey by Stijn Claessens and M Ayhan Kose Monetary and Economic Department November 2017

[11] https://www.investopedia.com/articles/investing/031413/economic-indicatiors-affect-us-stock-market.asp

­[12] <https://tradingeconomics.com/indicators>

[13] <https://www.schroders.com/en/insights/economics/covid-19-why-the-tech-giants-have-emerged-as-winners/>

[14] Asset returns and inflation, Eugene F.Fama and G.William Schwert, November 1977, Journal of Financial Economics , volume 5, issue 2

[15] Causal relations among stock returns and macroeconomic variables in a small, open economy [Øystein Gjerde](https://econpapers.repec.org/RAS/pgj6.htm) and [Frode Sættem](https://econpapers.repec.org/RAS/pst120.htm) [Journal of International Financial Markets, Institutions and Money](https://econpapers.repec.org/article/eeeintfin/), 1999, vol. 9, issue 1, 61-74

[16] The impact of macroeconomic variables on Stock ‎market in United Kingdom NEIFAR, MALIKA and Dhouib, Salma ‎ and Bouhamed, Jihen ‎ and Ben Abdallah, Fatma ‎ and Arous, Islem ‎and Ben Braiek, Fatma ‎ and Mrabet, Donia ‎ (2021)

[17] <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>

[18] <https://machinelearningmastery.com/time-series-data-stationary-python/>

[19] <https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/>

[20] <https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/>