Certificate in Data Analytics for Business

Project Report

GitHub URL

https://github.com/natmonge/UCDPA\_NataliaMongelos

Abstract

The Nobel prize-winning economist Simon Kuznets is said to have remarked that there were four types of countries: the developed, the underdeveloped, Japan, and Argentina. (https://www.batimes.com.ar/news/opinion-and-analysis/quo-vadis-argentina.phtml, n.d.)

The aim of this project is to examine the Consumer Confidence Index data for Argentina and compare it to a developed country, such as the USA, for the period 2015-2020 with a specific focus on the pre- COVID pandemic year and the first year of the pandemic. The hypothesis is that, if Argentina is truly an outlier among countries, the CCI data for both countries will show significant differences.

Additionally, the CCI data for Argentina will be compared to the OECD average over a 7-year period to further validate this hypothesis.

Introduction

As an Argentinian, I sought to explore and compare the economic development of Argentina with that of other countries, particularly in relation to consumer confidence, using the methods and tools learned during Data Analytics course.

The Consumer Confidence Index (CCI) is one of the most accurate and closely watched [economic indicators](https://www.investopedia.com/terms/e/economic_indicator.asp). It asks consumers about their spending plans in the near-term future and is a great measure of demand. Banks look to the CCI to forecast lending, and retailers and manufacturers use it to help set inventories.

As part of this project, I gathered CCI data for Argentina, USA, and OECD countries from 2 different sources.

When comparing the USA to the OECD, the US CCI trend turned out to be similar to that of the OECD, with noticeable fluctuations in April 2020, suggesting that the COVID pandemic had a significant impact on the CCI values of these countries . When comparing Argentina to the OECD and the USA, Argentina's CCI values were consistently more volatile, reflecting the economic challenges the country has faced in recent years.

Dataset

To conduct this project, I decided to gather CCI data from reliable sources, as follows:

1. Consumer Confidence Index for Argentina from an Argentinian Government website. (https://datosgobar.github.io/series-tiempo-ar-api/quick-start/, n.d.). Searched for the correct time-series code in the time series database and performed API call.

Graphical user interface, application

Description automatically generated

I also used Postman to obtain the same information in a json file.

A screenshot of a computer

Description automatically generated with medium confidence

1. Consumer Confidence Index for OECD (Total), US, UK from the OECD page. <https://data.oecd.org/leadind/consumer-confidence-index-cci.htm>. Downloaded data to a CSV file.

Graphical user interface, application

Description automatically generated

Implementation Process

* SET UP

1. Set up Git-Hub and Google Colaboratory.
2. Imported all the relevant packages.
3. Imported Consumer Confidence Data for Argentina using an API call. I then converted the API call in CSV to a Data Frame (CCI\_ARG) using pandas. (https://datosgobar.github.io/series-tiempo-ar-api/python-usage/, n.d.)

Graphical user interface, text, application

Description automatically generated

1. Created a Data sets folder in MyDrive. Uploaded the Consumer Confidence for OECD (total), UK and USA CSV file.
2. Read the CSV files from the drive and imported into a Data Frame (CCI\_OECD). (https://www.youtube.com/watch?v=VCllZKM7Njk)., n.d.)

Graphical user interface, application

Description automatically generated

* PREPARING DATA

1. CCI\_ARG Data Frame:
   1. Renamed columns from Spanish to English
   2. Inserted a Location column and populated it with value = ARG
   3. Checked for Null values using is.null () method.

Text

Description automatically generated

1. CCI\_OECD Data Frame:
   1. Checked for Null values using is.null () method.
   2. Used .fillna() Pandas method to convert “NaN” to “Unknown”

Graphical user interface, application, table

Description automatically generated

1. ALL\_CCI Data Frame:
   1. Used concatenation to merge the CCI\_\_ARG AND THE CCI\_OECD Data Frames. The objective is to have all the CCI information for OECD, US, UK and ARGENTINA in one Data frame.
   2. Used an ‘inner’ join as I only want the matching columns between the tables.
   3. Used ignore\_index to do not use the index labels.

Graphical user interface, text

Description automatically generated with medium confidence

* 1. Changed the “TIME” column format so that the dates would be in the datetime64 format. (https://towardsdatascience.com/clean-a-messy-date-column-with-mixed-formats-in-pandas-1a88808edbf7, n.d.)
  2. Checked for missing data using .isnull() Pandas method.
  3. Used the.drop\_duplicates() Pandas method to remove duplicate rows from the Data Frame (if any).

Graphical user interface, text, application

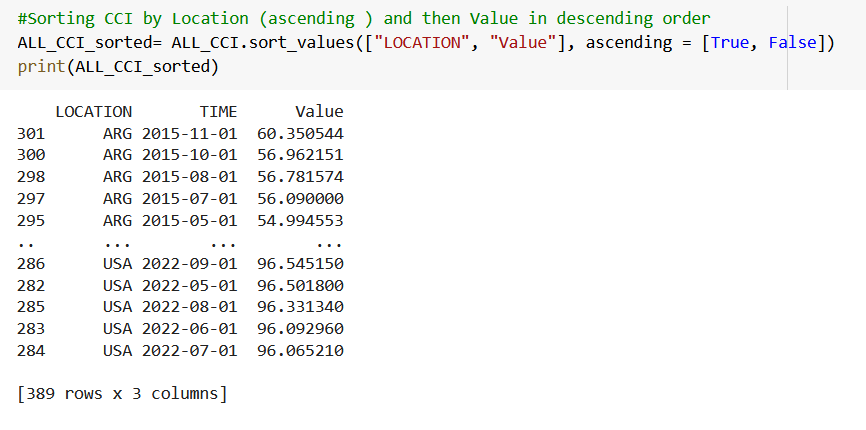
Description automatically generated

* 1. The ALL\_CCI concatenated Data Frame has 3 columns and 388 entries as follows:

Text

Description automatically generated

1. ALL\_CCI \_sorted :
   1. Used sort\_values() method on ALL\_CCI with multiple variables (Location and Value). I wanted the “Location” in ascending order and the Values from higher to lower. Therefore, I set ascending = True for Location and ascending = False for “Value”.



1. ALL\_CCI \_grouped :
   1. Used .groupby() method to form groups based on the “TIME” variable.
   2. Used .first() to select the first row for each group.
   3. I then used .get\_group () to find the entries contained in a specific date group. (https://www.geeksforgeeks.org/python-pandas-dataframe-groupby/, n.d.)

Text

Description automatically generated

* ANALYSIS

COMPARING CCI Values - ARGENTINA vs USA

1. Defined a custom function to calculate the minimum and maximum value in the “Value” column. I called this function “information”. When this function is called, it returns the following phrases: “The minimum value in column ‘Value’ is: “ and “The maximum value in the column ‘Value’ is:” together with the corresponding calculated values.

Text

Description automatically generated

1. Performed 2 queries on ALL\_CCI using the .query() method as follows:
   1. Query to obtain all values for Argentina up to the 31 Jan 2023. The resulting subset of data is stored in the variable “ARG\_CCI”.
   2. Query to obtain all values for US up to the 31 Jan 2023. The resulting subset of data is stored in the variable “USA\_CCI”.
2. Called the “information” function to obtain the minimum and maximum CCI values for both locations.

Graphical user interface, text, application, email

Description automatically generated

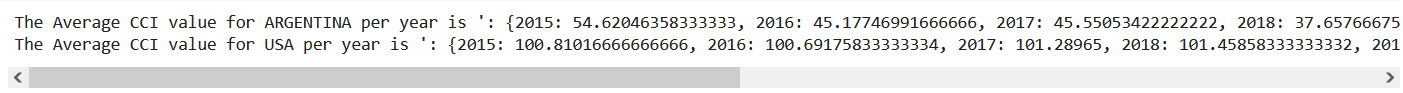
1. A for loop is used to calculate the average CCI value for Argentina. The CCI values for each year are stored in the ‘mean\_dict’ dictionary . I used the iterrows() method to loop through each row of ARG\_CCI . The year is extracted from the ‘TIME’ column of the row and the CCI value from the ‘Value’ column is appended to the list of values associated with that year in the ‘mean\_dict’ dictionary. If the year does not exist in ’mean\_dict’, it will create a new list for that year.

The empty dictionary “result\_ARG” is used to store the average CCI for each year. The code loops through the values in ‘mean\_dict’ using the items() method. It calculates the average of the list of CCI values for each year using “sum(values)/len(values)’. The result is stored in the ”result\_ARG” dictionary.

The same code was then used to get the average CCI value for the USA. (https://sparkbyexamples.com/pandas/pandas-filter-dataframe-rows-on-dates/#:~:text=You%20can%20also%20use%20df,method%20to%20filter%20by%20year, n.d.)

Text

Description automatically generated with medium confidence



1. Calculated the average CCI value for Argentina over the period Jan 2015 and Jan 2023. The variable “Average\_ARG” is assigned the “Value”column of the ARG\_CCI data frame. This is then converted to a NumPy array using np.array(). The mean of the NumPy array is calculated using np.mean() and it is assigned to the Average\_ARG\_mean variable.

Graphical user interface, text, application, email

Description automatically generated

1. **VISUALIZATION #1 –** ARG CCI Period Jan 2015-2023
2. **VISUALIZATION #2** – USA CCI Period Jan 2015-2023
3. **VISUALIZATION #3** – ARGENTINA vs USA CCI Period JAN2015/JAN 2023
4. **VISUALIZATION #4** – ARGENTINA CCI Period Year 2019-2020
5. **VISUALIZATION #5** – USA CCI Period Year 2019-2020
6. Two data frames (ARG\_CCI1 and USA\_CCI2) are created as follows:
   1. Performed 2 queries on ALL\_CCI using the .query() method:
      1. Query to obtain all values for Argentina for the period 01-12-2018 to 01-12-2020. The resulting subset of data is stored in the variable “ARG\_CCI1”.
      2. Query to obtain all values for USA for the period 01-12-2018 to 01-12-2020. The resulting subset of data is stored in the variable “USA\_CCI1”.
   2. The index for both data frames is set to the “TIME” column. This is the best option when performing time-based operations on the data.
   3. 3 new columns are created for both data frames as follows:
      1. “CHANGE” column, which calculates the difference between consecutive values in the “Value” column. The method used for this calculation is .diff()
      2. “PCT\_CHANGE” column, which calculates the percentage change in the “Value” column between consecutive periods. The method used for this calculation is .pct\_change().
      3. “PCT\_CHANGE\_ABS” column, which calculates the absolute percentage change in the “Value” column between consecutive periods. The function used to obtain the absolute value is .abs(). This column can be used to compare the magnitude of percentage changes, regardless of whether the change is negative or positive.

These new columns provide different ways of analysing the change in CCI over time.

Text

Description automatically generated

1. The ARG\_CCI1 and USA\_CCI1 data frames are manipulated and merged to create a new data frame : USA\_ARG as follows:
   1. The pd.merge\_ordered() function is used to merge the two data frames on the “TIME” column and using an “outer” join which means all rows from both data frames will be included in USA\_ARG.
   2. This data frame will be sorted by the “TIME” column.
   3. Suffixes are used to deal with overlapping columns (‘\_ARG’, ‘\_USA’)
   4. A subset of the columns from the merged data frame is selected (TIME, CHANGE\_ARG, CHANGE\_USA, PCT\_CHANGE\_ARG, PCT\_CHANGE\_USA)
   5. Rows from index position 1 are selected as the position 0 contains a NaN value (used .loc function to achieve this)

A picture containing text

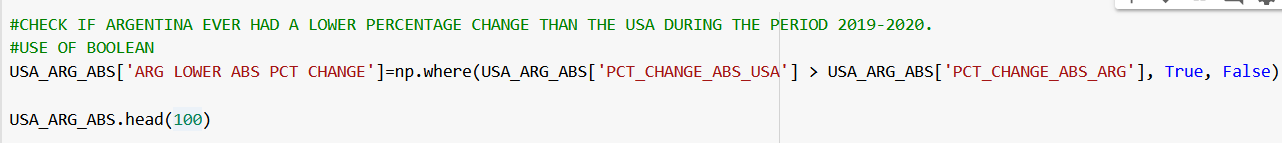
Description automatically generated

1. **VISUALIZATION #6** – CCI ARG vs USA – PERCENTAGE CHANGE 2019-2020
2. The ARG\_CCI1 and USA\_CCI1 data frames are manipulated and merged to create a new data frame : USA\_ARG\_ABS as follows:
   1. The pd.merge\_ordered() function is used to merge the two data frames on the “TIME” column and using an “outer” join which means all rows from both data frames will be included in USA\_ARG.
   2. A subset of the columns from the merged data frame is selected (TIME, CHANGE\_ARG, CHANGE\_USA, PCT\_CHANGE\_ABS\_ARG, PCT\_CHANGE\_ABS\_USA)

A picture containing text

Description automatically generated

1. Used NumPy “where()” method to create a new column “ARG LOWER ABS PCT CHANGE” in USA\_ARG\_ABS data frame. This column is populated with True or False depending on whether the values in “PCT\_CHANGE\_ABS\_USA” are greater than “PCT\_CHANGE\_ABS\_ARG”. The result is FALSE for all rows.



COMPARING CCI Values

1. Three data frames (ARG\_CCI2, USA\_CCI2 and OECD\_CCI) are created as follows:
   1. Performed 3 queries on ALL\_CCI using the .query() method.
   2. 3 columns are created for both data frames as follows: CHANGE, PCT\_CHANGE and PCT\_CHANGE\_ABS.

Text

Description automatically generated

1. The ARG\_CCI2 and OECD\_CCI data frames are manipulated and merged to create a new data frame : OECD\_ARG\_ABS . Dropped the NaN value using .dropna(axis=0)

Graphical user interface, text

Description automatically generated with medium confidence

1. Compare absolute percentage change for Argentina and the OECD. The objective is to check if Argentina ever had a lower absolute percentage change than the OECD. The “result” data frame is returned that contains only the rows where the value in the “ARG LOWER ABS PCT CHANGE” column is TRUE.

Graphical user interface, text, application, email

Description automatically generated

1. Created a custom function that plots time series.

Text

Description automatically generated

1. **VISUALIZATION #7**- CCI ARG vs OECD – Absolute Percentage Change – 2015-2023
2. The USA\_CCI2 and OECD\_CCI data frames are manipulated and merged to create a new data frame : OECD\_USD\_ABS . Dropped the NaN value using .dropna(axis=0) OECD\_USD\_ABS2
3. **VISUALIZATION #8**  CCI USA vs OECD – Absolute Percentage Change – 2015-2023

Results

* **VISUALIZATION #1** ARG CCI Period Jan 2015-2023 and **VISUALIZATION #2** USA CCI Period Jan 2015-2023.
  + Created 2 separate visualizations of the Consumer Confidence Index (CCI) of Argentina and USA for the period January 2015 to January 2023 using the Matplotlib library in Python.
  + I used the following line: plt.rcParams['figure.figsize'] = (9, 6) to make the plot bigger. (https://colab.research.google.com/drive/1wqIUyStkKbUc4MYPvJgIExPOFvC5cUs2#scrollTo=SJ\_K4UjsNDcR, n.d.)
  + The x-coordinate is TIME and the y-coordinate is the CCI value.
  + A circle was used as a marker (marker=”o”) and a dashed line was used to connect the data points (linestyle=”- -"). The “color” argument specifies the colour of the line and markers (blue for Argentina and red for USA). Labels and title were also added to the visualizations using set\_title (), set\_xlabel () and set\_ylabel()

Graphical user interface, text

Description automatically generated

Chart, scatter chart

Description automatically generated

Text

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

* **VISUALIZATION #3** – ARGENTINA vs USA CCI Period JAN2015/JAN 2023
  + When trying to compare visualizations #1 and #2 I realized that the scales were different, so I decided to plot them in the same sub-plot, using two different y-axis scales. The twinx method was used so that the 2 Axes share the same x-axis but the y-axes are separate. (Data Camp -Introduction to Data Visualization with Matplotlib, Chapter 2)

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

* **VISUALIZATION #4 –** ARGENTINA CCI Period Year 2019-2020and **VISUALIZATION #5** USA CCI Period Year 2019-2020
  + Zoomed on ARGENTINA period 2019-2020 and USA period 2019-2020 by slicing into the corresponding data frames.

Text

Description automatically generated

Text

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

* **VISUALIZATION #6 –** CCI ARG vs USA – PERCENTAGE CHANGE 2019-2020
  + To compare the CCI percentage change between ARG and USA, I created a line plot of 2 time series data (PCT\_CHANGE\_USA and PCT\_CHANGE\_ARG) on a common x-axis (USA\_ARG.index). I used .ylim () to set the y-axis limit between -0.09 to 0.12. (How to change y axis scale: https://pythonguides.com/matplotlib-set-y-axis-range/, n.d.)

Text

Description automatically generated

Chart, line chart

Description automatically generated

* **VISUALIZATION #7 –** CCI ARG vs OECD – ABSOLUTE PERCENTAGE CHANGE 2015-2023
  + To compare the CCI percentage change between ARG and USA, I created a line plot of 2 time series data (PCT\_CHANGE\_USA and PCT\_CHANGE\_ARG) on a common x-axis

A picture containing calendar

Description automatically generated

Chart, histogram

Description automatically generated

* **VISUALIZATION #8 –** CCI USA vs OECD – ABSOLUTE PERCENTAGE CHANGE 2015-2023

**Chart, histogram

Description automatically generated**

Insights

* **Argentina vs US** 
  + USA shows an upward trend in its CCI values from 2015 up to 2020, which coincides with the start of the COVID pandemic. The US CCI reached its highest points in March 2018, December 2019, and January 2020 (101.6). From January 2020 to July 2020, the CCI drops by 3.09% to 98.54. The CCI then goes up again until April 2021, when it takes a drop of 3.86% to 96.06 in July 2022. These fluctuations do not follow the trend of previous years. The pandemic had a significant impact on the US economy and the fluctuations in the CCI over the past few years most likely reflect this. [Visualizations #2, #3,#5]
  + In contrast, Argentina displays a downward trend from November 2015, when it had its maximum CCI value (60.35), until November 2018, when it displays its lowest CCI value (32.09). However, the average CCI goes up in 2019 and again in 2020. The COVID pandemic did not have the same effect in Argentina as in the US. [Visualizations #1, #4]
  + The CCI fluctuations on a month-to-month basis are significantly larger in Argentina than in the USA. Even during the period 2019-2020, when USA experienced larger CCI fluctuations than in previous years, Argentina continued to experience even greater month-to-month fluctuations in its CCI.[Visualization # 6]
* **US vs OECD**
  + The US CCI trend is extremely like that of the OECD countries. The fluctuations tend to follow a very similar pattern, with a noticeable peak of 0.01 absolute percentage change for both the US and the OECD in April 2020 which coincides with the beginning of the COVID pandemic. [Visualization #8]
* **Argentina vs OECD**
  + Argentina’s CCI fluctuation on a month-to-month basis is consistently higher than that of the OECD countries for the period 2015-2023. There were only two occasions during this period (in November 2017 and August 2018) when the absolute percentage change was lower for Argentina than the OECD. Argentina has experienced significant economic challenges in recent years, including high inflation, a volatile currency, and political instability. As a result, it is expected that the CCI in Argentina would be highly volatile.[Visualization # 7]

Machine Learning

* **Predictions that could be performed using machine learning:**
  + **Future trend of CCI based on historical data.** This information can be used for economic forecasting and decision making. For example, the Federal Reserve in the USA uses it when considering interest rate change and it also affects stock market prices. (https://chernoffnewman.com/news/using-the-the-consumer-confidence-index-to-forecast-business/, n.d.)
  + **Identify factors that impact CCI for each country**. Factors such as Inflation, GDP, unemployment could be analysed to determine their correlation with the CCI. (https://www.economicshelp.org/blog/6544/economics/uk-consumer-confidence-2/, n.d.)
  + **Identify groups of countries with similar CCI trends.**
* **Methods:**
  + As CCI can only be observed over time, a time series regression model would be an appropriate method to use. Autoregressive models, such as ARIMA can be used to create a linear equation that forecasts the time series data. (https://www.sas.upenn.edu/~fdiebold/Teaching104/Ch14\_slides.pdf, n.d.)

# References

Data Camp -Introduction to Data Visualization with Matplotlib, Chapter 2. (n.d.).

Data Camp -Introduction to Data Visualization with Matplotlib,Annotating time-series data. (n.d.).

*How to change y axis scale: https://pythonguides.com/matplotlib-set-y-axis-range/*. (n.d.).

*https://chernoffnewman.com/news/using-the-the-consumer-confidence-index-to-forecast-business/*. (n.d.).

*https://colab.research.google.com/drive/1wqIUyStkKbUc4MYPvJgIExPOFvC5cUs2#scrollTo=SJ\_K4UjsNDcR*. (n.d.).

*https://datosgobar.github.io/series-tiempo-ar-api/python-usage/*. (n.d.).

*https://datosgobar.github.io/series-tiempo-ar-api/quick-start/*. (n.d.).

*https://sparkbyexamples.com/pandas/pandas-filter-dataframe-rows-on-dates/#:~:text=You%20can%20also%20use%20df,method%20to%20filter%20by%20year*. (n.d.).

*https://towardsdatascience.com/clean-a-messy-date-column-with-mixed-formats-in-pandas-1a88808edbf7*. (n.d.).

*https://www.batimes.com.ar/news/opinion-and-analysis/quo-vadis-argentina.phtml*. (n.d.).

*https://www.economicshelp.org/blog/6544/economics/uk-consumer-confidence-2/*. (n.d.).

*https://www.geeksforgeeks.org/python-pandas-dataframe-groupby/*. (n.d.).

*https://www.youtube.com/watch?v=VCllZKM7Njk).* (n.d.).