FINAL ASSESEMENT. Introduction to data Science with Python

General Instructions

This is the final assessment for the course. You need to download the datasets provided to answer the questions.

The 5 datasets named 'World_Happiness_Report' (there is one for each year of data) are used for Part A of this Assessment which corresponds to Modules 3 to 5, and account for 50% of the grade of the assessment.

The dataset 'Air Passengers' is used in the final part of the assessment, which correspond to module 6 and account to 20% of the grade.

Modules 1 and 2 are implicitly evaluated as part of the assessment, as you are expected to use Python and Jupyter notebooks for data analysis and presentation. The correctness, quality and clarity of your code will account for 20% of the grade.

You are supposed to complete the assessment using a Jupyter Notebook. You should use the notebook to load the data, perform the calculations, and present the results using Markdown or text cells. The clarity and presentation of the notebook will also be part of the grading criteria (10% of the grade).

When you finish the assessment, please submit the .ipynb notebook. Make sure that all cells work correctly, as the professor will re-run all cells before evaluating the assessment. To simplify the correction process, please place the data files in the same directory as the Jupyter notebook. This way, the cells loading the data files will work directly in the professor's environment without the need to change the file path.

IMPORTANT
Put your name and your student code (if you know it) at the beginning of your jupyter notebook, and is possible, also as the title of the notebook. For instance

Final_Assesment_Planas_Bielsa_240784.ipynb

==========

PART A- Understanding the Sources of Happiness

(**datasets used**: all 5 CSV files - World Happiness Report_yyyy.csv)

Context:

The 'World Happiness Report' dataset is a comprehensive collection of data that measures happiness levels in different countries worldwide. It provides valuable insights into the factors that contribute to happiness and offers a global perspective on well-being. The dataset includes various variables that assess economic indicators, social support, health, freedom, trust, and generosity, among other factors. With

observations spanning multiple years, it allows for temporal analysis and identification of happiness trends across different countries.

Each entry in the dataset represents a specific country and contains a wealth of information for comparative analysis. Researchers, policymakers, and individuals interested in understanding happiness levels can leverage this dataset to explore the underlying factors and their variations across different regions. By analysing the World Happiness Report dataset, one can gain insights into the determinants of happiness and uncover potential strategies to promote well-being and enhance the quality of life worldwide.

To perform the analysis, you should load all five CSV files corresponding to the years 2015 to 2019 into your Jupyter Notebook. These datasets will provide the necessary data to answer the questions and gain a comprehensive understanding of happiness trends and influencing factors across different years.

QUESTIONS PART A

Descriptive Statistics

1) Identify and describe the types of variables present in the World Happiness Report dataset. Categorize each variable as discrete, categorical, ordinal, or continuous.

2) Calculate the mean, median, variance, and standard deviation for each continuous variable in the dataset for each year from 2015 to 2019.

3) Create a separate line plot for each continuous variable, showing the temporal evolution of the mean from 2015 to 2019. Each plot should have the year on the x-axis and the mean on the y-axis.

4) Generate two histograms in the same plt.figure, one for the Happiness Score in 2015 and another for the Happiness Score in 2019. Discuss any observed differences or similarities.

5) Calculate the correlation coefficients between the Happiness_Index and the continuous variables (economy, family, health, freedom, trust, generosity). For each year, which variable has the highest positive correlation with happiness and which variable has the highest negative correlation with happiness.

Modelling *(Use LinearRegression model from the scikit-learn library*)

1) For each year in the dataset, create a unidimensional linear model where the independent variable is one of the factors (economy, family, health, freedom, trust, generosity), and the dependent variable is the happiness score. Report the coefficients and intercept for each linear model and calculate the coefficient of determination (R^2) for each model.

2) Identify which variable explains the happiness score better by itself for each year. Is this variable the same every year, or does it vary?

3) Build a multidimensional linear model using all the variables (economy, family, health, freedom, trust, generosity) as independent variables and the happiness score as the dependent variable for each year. Then, compare the accuracy of the multidimensional model with the unidimensional models. Does the inclusion of all variables improve the accuracy of the model?

Inference and Hypothesis Testing

1) Consider the population to consist of all the countries listed. For the 2019 data, randomly select a sample of 30 countries from the population. Compute a 95% confidence interval for the sample mean of The Happiness score.

2) When constructing the dataset, various variables were considered, including the annual sunshine duration in hours for each country. Although the data was not ultimately included, we know that the standard deviation of the values was 200 hours. The newspaper claimed that the population average was 2500 hours or less. To investigate this claim, a new measurement was conducted last year on a small sample of 50 countries randomly selected, and the sample average was found to be 2580 hours. Formulate the null and alternative hypotheses to test the claim and determine whether this should be a one-sided or two-sided test. Based on the sampled data, state whether the null hypothesis should be rejected or not.

==========

PART B - Forecasting air passengers.

(**dataset used**: AirPassengers.csv)

Context:

Part B of this Assessment focuses on the famous 'Air Passengers' dataset, which provides monthly data on the number of air passengers from 1949 to 1960. This dataset is commonly used in time series analysis and forecasting tasks and it presents a valuable opportunity to study the trends and patterns in air travel over a specific time period. allowing for the exploration of various time series analysis techniques. This dataset is particularly useful in understanding the growth, seasonality, and long-term trends in air travel demand.

By analysing the 'Air Passengers' dataset, we can gain insights into the factors that influence air travel, identify patterns in passenger behaviour, and make predictions about future passenger numbers. This dataset is widely used in the field of transportation analysis, helping airlines, airports, and policymakers make informed decisions about capacity planning, marketing strategies, and resource allocation.

QUESTIONS PART B

Time Series

1. load the dataset and plot the time series. Which decomposition method do you think will be more appropriate for this dataset? Additive or multiplicative? reason your answer.

2 Perform a seasonal decomposition of the "Air Passengers" dataset using the method that you selected as more appropriate in your previous answer. Visualize the trend, seasonality, and residual components. make a plot(or several plots) to show all different components of the time series.

3 Can you estimate a prediction for the number of passengers in December 1961?

==========END ASSESMENT ==========

Acknowledgement of data sources.

The “Happiness Report dataset”, which is an open source dataset freely available to the public. The World Happiness Report is an annual publication by the Sustainable Development Solutions Network, and the dataset can be accessed from their official website or platforms like Kaggle.

The 'Air Passengers' dataset is a publicly available dataset widely used in time series analysis. The 'Air Passengers' dataset provides monthly data on the number of air passengers from 1949 to 1960 and is often used as a benchmark dataset for forecasting and analysing time series data. The dataset can be accessed from various sources, including the 'datasets' package in Python's seaborn library or other online data repositories."