Mathematics For Computer Science Engineers
UE23MA242A

Teaching Assistants : Archishman VB, Suchir M Velpanur, Neha Bhaskar

COVID-19 Case Study

Introduction

The two sections below, Background and Case Study provide context for the data science hackathon. This exercise will allow you to test your skills in using the Python programming language to effectively explore the characteristics of a dataset and analyze the features using descriptive statistics such as summary statistics, tables, and graphs. Happy coding!

Background

The COVID-19 pandemic emerged as one of the most impactful global health crises of the 21st century. First identified in late 2019, the virus quickly spread worldwide, resulting in millions of infections and a substantial loss of life. Governments around the world implemented stringent measures, including lockdowns, travel restrictions, and vaccination drives, to curb the spread of the virus. The pandemic not only highlighted the importance of public health systems but also underscored the role of data in tracking infection rates, understanding virus variants, and managing healthcare resources.

Case Study

Researchers have since gathered vast datasets on COVID-19, documenting cases, recoveries, fatalities, and vaccination rates, providing invaluable insights into managing future pandemics. A researcher, having acquired a certain dataset, would like to view it from the lens of statistical analysis to analyze various factors which affected those infected by the disease and also those who succumbed to the same.

Dataset Description

The different variables involved in the dataset include :

1) UID : A unique identifier assigned to each record in the dataset
2) iso2 : A two-letter ISO (International Organization for Standardization) code representing the country or region
3) iso3 : A three-letter ISO (International Organization for Standardization) code representing the country or region

4) code3 : An ISO 3166-1 numeric code, which is a three-digit numerical code associated with each country
5) FIPS : Federal Information Processing Standards (FIPS) code, a numeric identifier for geographical entities in the United States.
6) Admin2 : The second-level administrative division, which may refer to a county or district within a state or province
7) Province_State : The name of the province or state within a country
8) Country_Region : The name of the country or region
9) Lat : The latitude coordinate of the geographical location
10) Long_ : The longitude coordinate of the geographical location
11) Combined_Key : A concatenated string that combines geographic identifiers (such as state and country) to create a unique key for each row
12) Date : The date of the reported data entry for cases(confirmed/deaths)
13) Confirmed : The total number of confirmed cases of COVID-19 reported for the specific date and geographical location
14) Deaths : The total number of deaths reported for the specific date and geographical location

Problem Set

Unit 1

1) Classify features in the given COVID-19 dataset into corresponding data types(ordinal, nominal, interval, or ratio). Provide a rationale for your classification.
2) A summary statistic provides a numerical summary of a specific feature within the dataset.There are two commonly used categories of summary statistics: those that indicate the central tendency and those that indicate the spread of the data. Identify the most appropriate measure of central tendency for each attribute in the dataset and state its corresponding value. Additionally, calculate the standard deviation and range of values for each column.
3) Identify and describe any data quality issues or inconsistencies within the COVID - 19 dataset. What steps would you take to clean and preprocess the data to ensure its accuracy and reliability for further analysis?
4) Plot histogram and box plot for 'Confirmed' and 'Deaths' variables. From this,
   i) Identify the type of distribution each of the variables follow
   (Hint : limit scale of visualizations for both histograms and box plots)
   ii) Identify number of outliers for each variable

iii) Bonus question : After identifying type of distribution followed by both variables, try to plot a new histogram and boxplot for 'Confirmed' variable after correcting any anomalies noticed in your visualizations

(Hint : Try to rescale histogram and box plot according to distribution noticed from previous visualization)

5) Explain how you identified presence of outliers and how to overcome the same
6) Examine the normal probability plot (Q-Q plot) for the 'Deaths' variable in the dataset. Based on the shape and trend of the plot, what conclusions can be drawn? Provide a rationale for your conclusions.
7) Calculate the correlation between 'Deaths' and other numerical variables. Set a correlation threshold and create a heatmap to visualize the relationships. Identify variable having highest correlation with 'Deaths'
8) Sample 10,000 rows randomly and generate a pairplot that includes the variables 'Deaths,' 'Confirmed,' and 'code3' while using 'iso3' as the hue in the dataset. What insights can be gained from the pair plot, and how does it help in visualizing the relationships between these variables?

Unit 2

9) Use hypothesis testing to answer the following:

   Define a null and alternative hypothesis to investigate whether the number of confirmed cases has a significant impact on the median number of deaths in a given dataset. Use a T-test to analyze the relationship between these two variables. Plot a histogram to analyze your hypothesis and its results. Assume significance level as 0.05.

10) Calculate the margin of error to quantify the precision of the analysis done previously and what you can infer from the results.

Unit 3

11) Perform linear regression to predict Deaths using code3 , FIPS, Lat, Long_ and Confirmed. Plot the predicted vs actual number of deaths. Also calculate RMSE, MSE and R^2. Explain what do each of these metrics signify
12) Given the variables 'code3', 'FIPS', 'Lat', 'Long_', and 'Confirmed', which represent geographical and case-related information, what additional features could be engineered from these variables to improve the prediction of 'Deaths'? For instance, consider how combining or transforming these variables could

reveal new insights into the spread and severity of the pandemic. Give 2 such feature aggregations for the same.