Assignment 2: Data Analytics using Python

Nai-Wen, Chung

Assignment activity 1: Prepare your GitHub repository

My Github Link: https://github.com/feinwc/LSE_DA201_assignment

Assignment activity 2: Import and explore the data

The dataset provides detailed information about the pandemic. Interestingly, there are some missing values and unusual patterns in the values of “deaths, recovered and hospitalised” columns, which might be caused by human errors or miscalculations during the time as the pandemic was unexpected.
There is a default index in DataFrame starting with the number zero. I chose not to create a new set of indexes as I used other functions to filter the relevant data. The initial insights that I have discovered are the missing values and the max/min values. By filtering and sorting the shape, type, and max/min of each column, I have developed a deep insight into the story over time. Firstly, the number of vaccinated individuals has fluctuated over time. The vaccinated number reached the highest in April and May 2021 (maximum: 69,619) and slightly went down afterwards. From 2020-01-22, there was zero vaccinated individual until January 2021, showing that the vaccination scheme may start to roll out from the month. Moreover, there are some missing values in the columns of “Deaths, Cases, Recovered and Hospitalised” on 2020-09-21 and 2020-09-22 in cases. The missing values all belong to "Bermuda". Fortunately, there is no missing value in the "vaccinated" file.
Some unusual findings are discovered in DataFrame “Gilbraltar”. During the time appearing the highest death and case rate, the hospitalised values are showing as zero on both 2021-10-14 and 2021-10-13, which can be human errors or miscalculations. Also, some recovered rate seems missing as the values are showing as zero. The data reflects that there is certain relevance among death, cases, recovered and hospitalised growth rates. When the rate of death and case increase, the recovered rate tends to decrease, and the hospitalised number grows simultaneously.

Assignment activity 3: Merge and analyse the data

Firstly, the “First Dose, Second Dose and Vaccinated” column was filtered to understand the province/state with the highest full vaccination rate. “Groupby()”, “sum()” and “sort_values()” functions were used. The result shows that **Gibraltar** has the highest first, and second dose and full vaccination numbers, followed by **Montserrat and British Virgin Islands**. Interestingly, January to March is the period that people have the highest vaccination number for the first dose, whilst April and May are the two months with the highest vaccinated number for the second dose. The time difference can infer that the first dose application may start in January and the second one in April. **Gibraltar** is found to be the country with the highest number of people who took the first dose only. A new DataFrame “uk_perc” was created and a “use user-defined” function was applied to understand the percentage of the vaccination rate. The result demonstrates that **Turks and Caicos Islands** have the largest percentage of people who had the first dose only, followed by **Isle of Man and Anguilla**. Over time, the growth of the first dose rate raised from January and reached the top in March; while the second one grew from April and reached the top in May. In short, **Gibraltar** has the best performance in terms of the largest number of

vaccinations. The government can implement a marketing campaign in **Turks and Caicos Islands** to take the second dose to improve the percentage.

Assignment activity 4: Visualise and identify initial trends

Based on the plot, the percentage of full vaccinated percentage is much higher than partially vaccinated across regions, inferring that most people in the UK have been fully vaccinated.
Regarding the death rate, there are some outliers found when running the sum of deaths across provinces/states, which caused the distribution to be skewed. Thus, groups “Anguilla, Falkland Islands (Malvinas), Saint Helena, Ascension and Tristan da Cunha and others” were removed. The line chart shows that **Bermuda and Isle of Man** seem to have not reached the peak in October 2021, whilst other regions are presented as more stable with their death numbers. Also, the percentage of the full vaccination displays that none of the regions has reached 100%, explaining that the government can initialise the first marketing promotion in **Bermuda and Isle of Man** to stabilise the death numbers.
Regarding recoveries, **Channel Islands and Gibraltar** embrace the highest number. The recoveries across regions are not consistent as the number dropped rapidly approximately from August/September 2021, which can assume that fewer individuals tested positive; thus, fewer recovered numbers were reported.
Regarding visualisation, visualisation can help the government to determine the target regions for a marketing campaign. The bar chart demonstrates the comparison of partially and fully vaccinated rates across regions and the line charts inform the trend of death and recovery over time. The government can easily identify the relevant information such as maximum, minimum in numbers and regions with different colours. The error line appears after the date is converted into the month, showing that the number contains a certain level of the confidence interval. The smoothness of the lines, the accuracy of the raw data and more annotation on the charts can increase the quality of this visualisation.

Assignment activity 5: Analyse the Twitter data

I decided to keep my and Norah’s approaches as they can be applied to different scenarios. I chose to use “string.extractall()” function to extract all hashtags and converted the series to a DataFrame. The result displays the top 30 hashtags. To filter the hashtags with covid-related strings, I used “str.contains()” function and some hashtags were excluded as the count is relatively low. Only the top 10 were displayed. Finally, the chart demonstrates that **“#COVID19, #CovidIsNotOver, #coronavirus, #covid19/Covid19 and #COVID/covid”** are the hottest hashtags, which #COVID19 was tweeted for nearly 2,000 times, following by #CovidIsNotOver. The hashtag “#CovidIsNotOver” is quite interesting as people who tagged this might have a different view towards the current vaccination scheme or ideas towards post-pandemic lives. I would apply sentiment analysis to understand cyber behaviour more.
Norah’s approach provides the content of each tweet, which can be used to investigate the relevant information or relationships between tweets. I would investigate the content of the tweets to analyse the correlation/causation. For example, “Pregnant women have a low risk of having severe symptoms” shows how “risk” is related to “cases” and the sentiment should be relatively positive. In this approach, we could

understand human reaction toward the government scheme and predict the future in post-pandemic lives for different groups.

Assignment activity 6: Perform time-series analysis

The consultant tried to achieve...
plot_moving_average() function can be used to understand the upper bound and lower bound of data points. The plot below demonstrates the time-series forecasting between 2021-01 to 2021-11 to visualise the standard deviation of the rolling mean for each day.

The mean_absolute_error() function can be used to understand the mean absolute error between the actual and seven-day rolling average. Firstly, a copied DataFrame was created. Secondly, define "s_rolling" to calculate the mean of each day. Thirdly, create a new column "error" in DataFrame "s" and use mean_absolute_error() function to calculate the mean absolute error for each day. Finally, sort the values and display the top 3 values of mean absolute errors.

Additional Question 3.1
A qualitative method is an approach that achieves assumptions based on the opinions or judgements of customers and professions. This method is usually used when there is a lack of historic data. In business predictions, a qualitative method can build a deep understanding of an occurrence, identify variables and create hypotheses for quantitative research. On the other hand, quantitative research can analyse the data statistically by quantifying the data. In business predictions, quantitative forecasting can “test hypotheses, make predictions and generalize results to the population of interest” from large samples (Chrysochou, 2017, p.412; Malhotra et al., 2012). It is better in terms of efficiency and accuracy. Examples of quantitative forecasting methods are “causal, correlational or observational, and time-series methods”.

Additional Question 3.2
Continuous improvement is important as it can ensure constituency and accuracy to optimise a process. Continuous improvement can help a business to envision better operations and make better decisions to achieve business goals. We are unable to just implement the project and move on to other matters as continuous improvement is not only about “Plan and Do” but also about “Check and Act”. The process of “Check and Act” can help to monitor and review the implementation as well as update and improve the operations. If we neglect “Check and Act”, the project is likely to be out of track and ends up not aligning with business goals. Thus, continuous improvement is vital in business operations.

Additional Question 3.3
Data ethics should be seen as a serious topic as it affects people and society. Aligning with data ethics can not only reduce the impact on people and society but also prevent the risk of “fraud, data misuse, unauthorised use, and phishing schemes”. Although we don’t expose any personal details, the risk mentioned above might cause a negative influence on a business in the community. In worse-case scenarios, a business can cost a lot if a competitor starts a lawsuit against the company. Thus, data ethics must be seen as essential in businesses.