SIT112 - TASK COMPLETION REPORT

Important note 1: Please **do not** include Python code in this report. It would be acceptable though to make brief references to function names or different parts of the task's notebook when needed.

Important note 2: Please **do not** leave any response field empty; if not applicable, enter NA.

1. TASK SUMMARY

TASK NAME (ABBREVIATION)	YOUR NAME (STUDENT ID)	YOUR DEAKIN EMAIL
Distinction (D1)	Hoang Long Tran (s223128143)	s223128143 @deakin.edu.au

2. TASK DESCRIPTION

ITEM	RESPONSE
What was the objective(s) of the task?	The objective of the task was to analyze YouTube videos on COVID 19. Specifically, my task was to analyze all comments. From its distribution, it changes overtime in general and on specific channels. I also touched on the total likes and views to further prove my point in the analysis.
What kind of data did you work with?	There was initially no data and I had to scrape. Thankfully, Deakin provided the code to scrape all the information about the YouTube videos, and code to clean the data after scraping.
Briefly describe the data science task you worked on.	Deakin's cleaned and prepared the data for me, so I just have to analyze. I analyze the distribution of comments, likes, and views for all channels. Find out why the distribution is skewed and look for patterns that cause the skewness. Then, understand why those outliers appear, and use that as a reference for the next analysis. Plot two time series for the total comments for each month of all channels and for individual ones. Then, analyze the time series.

3. TECHNICAL SKILLS

ITEM	RESPONSE
What technical skills did you use during the task?	The technical skills used are the following. First is to be able to read the plots. Understand what skewed distribution is, how to deal with those, what is outlier, how to remove them and so on. Use python to merge, group, resample, and plot data.
list any challenges or obstacles you faced while working on the task and how you overcame them.	The challenge I faced is understanding the code provided by Deakin. Understand how to use the API to scrape using the code, understand how to code is cleaned and prepared. Research about statistics, correlation, distribution, time series, outliers. Don't know what to

ITEM	RESPONSE
	present, then go through all of the questions and find out what is best to present. Those are the challenges I faced throughout this task. To overcome each one, I just have to research via Google and ChatGPT.

4. DATA CLEANING AND PREPARATION (ENTER NA WHEN NOT APPLICABLE)

ITEM	RESPONSE
What steps did you take to clean and prepare the data?	The only cleaning that I can say I did was to drop duplicates. In some tasks, while grouping the data, I saw some duplicates, so I just called in the function from pandas.
Did you encounter any issues with the data during this process? How did you address these issues with the data?	Like I said, I address it by searching. On technical terms, I use those new knowledge gain to address those issues. Like, confirming the skewness by looking at the tail, remove the outlier using upper and lower bound, drop the duplicates for more interpretable data. It is a trial-and-error process.

5. DATA ANALYSIS (ENTER NA WHEN NOT APPLICABLE)

ITEM	RESPONSE
How did you analyze the data?	I analyzed the data by going through all the questions. Understand each question and fill my knowledge gaps. Once done, I re-analyze my analysis on the selected questions to present. I analyze the distribution, understand why there are outliers and the meaning of it. Plot the time series to explain the outliers.
Did you use any visualization techniques to better understand the data?	Yes, I used histogram to look at the distribution overall, boxplot to look at outlier, line plots to understand time series data.
What insights did you gain from this analysis?	Most video views about Covid19 are between 44990 and 456817. "Bill Gates makes a prediction about when coronavirus cases will peak" is the most viewed video. Most video likes are about covid 19 is between 396 and 4102. Most video comments are around 133 and 1729. Most of the outliers in the likes, comments and views are in the early to mid of 2020, peaked COVID 19 time. Time series plot on the major events of COVID 19 shows that. Also, outlier removal is not suitable for time series because it will mess up the natural variations. That has been proven in the time series plot for the total comments per channel group by month.

6. BASIC REQUIREMENTS FOR THE TASK

ITEM	RESPONSE (YES/NO)
Are you confident to execute the Python code in this task and explain the output?	Yes
Are you confident to explain what each line of code does and how it contributes to the solution(s)?	Yes
Are you confident to rewrite or modify the code after completing this task? • For pass tasks: with guidance, no time limit. • For credit tasks: with limited guidance, no time limit. • For distinction tasks: independently, no time limit. • For high distinction tasks: independently, in a limited time.	Yes

7. CODE ATTACHMENT (NOT APPLICABLE TO THE PASS TASKS – ENETER NA)

ITEM	RESPONSE (YES/NO/NA)
Have you attached the notebook file that contains your solutions (Python code) for this task?	Yes
Have you executed all the cells in your attached notebook and ensured there is no error? *Please note your submission will not be flagged as complete if your attached notebook contains any error.*	Yes

8. VIDEO ATTACHMENT (NOT APPLICABLE TO THE PASS/CREDIT TASKS: NA)

ITEM	RESPONSE (THE VIDEO LINK/NA)
Provide the link to the video recording that presents your completed task. This is only for Distinction and High Distinction tasks. Enter NA for Pass/Credit tasks.	https://www.youtube.com/watch?v=hDp0AyfAZwA

9. AKNOWLEDGEMENT

BY SUBMITTING THIS REPORT, I ACKNOWLEDGE THAT:

- MY RESPONSES ARE ACCURATE AND ARE MY OWN WORDS.
- I HAVE MET ALL THE BASIC REQUIREMENTS OF THE TASK (LISTED IN SECTION 6).
- I HAVE READ AND FULLY UNDERSTOOD THE ASSESSMENT GUIDELINE OF THE UNIT.
- THIS REPORT DOES NOT EXCEED 3 PAGES.
- THIS REPORT DOES NOT INCLUDE CODE EXCEPT BRIEF REFERENCES TO FUNCTION NAMES OR DIFFERENT PARTS OF THE TASK'S NOTEBOOK.
- MY SUBMISSION DOES NOT CONTAIN ANY CREDENTIALS (E.G., PASSWORD, API KEY, ETC) OR PERSONAL INFORMAIOTN.

IMPORTANT NOTE 3: IF YOU HAVE ANSWERED NO TO ANY OF THE QUESTIONS IN SECTIONS 6, PLEASE RECONSIDER SUBMITTING YOUR REPORT; ASK HELP FROM YOUR TUTOR.

ADD YOUR NAME AND SIGNATURE HERE: HOANG LONG TRAN

Long