Data Journal

Date: 28/11/2021	**Course/topic: Case study How can a wellness tech company play it smart?**
Prompt:	Scenario
Journal Entry:	You are a junior data analyst at a tech company named Bellabeat. They make women’s health-related designer products that can give users feedback about their activity, stress, sleep, menstrual cycle or hydration levels throughout the day. Seren(co-founder) thinks that analysis of customer usage will uncover much potential growth.
Other thoughts or questions:	You are tasked with providing insights on trends that can help guide the decision making process for marketing strategy.

Date: 28/11/2021	**Course/topic:ASK Phase**
Prompt:	What are the smart device usage trends in the dataset? Which of these trends apply to the Bellabeat customers? How can these trends influence marketing strategy?
Journal Entry:	Business TASK:-we are tasked with finding insights and presenting them with high-level recommendations. that can add value to the decision-making process for marketing strategy. We are trying to find any trends that can help in the growth of overall product sales. The Task that I am presented with can help the decision-making process by providing key insights into the behaviour of consumers and how they use other smart products other than Bellabeat. :-stakeholders expectation is to uncover hidden patterns of usage that can fuel the growth of product sales via the use of marketing. Stakeholders:- Two founders, Market analyst team
Other thoughts or questions:	Seren(co-founder) has an art background and she is a chief creative officer. Sando Mur has a mathematics background and data analyst team under the name of market analyst team. You have to tailor your presentation in a way that Seren and the Mur can understand as they have a non-technical background.

Date:	**Course/topic:PREPARE PHASE**
Prompt:	Identifying sources of data, pre-processing and preparing that for cleaning.
Journal Entry:	Sršen encourages you to use public data that explores smart device users' daily habits. She points you to a specific data set: ● FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius): This Kaggle data set contains a personal fitness tracker from thirty Fitbit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits. Sršen tells you that this data set might have some limitations, and encourages you to consider adding another data to help address those limitations as you begin to work more with this data.
Other thoughts or questions:	-Exploring Dataset to familiarize it to me.

Date:	**Course/topic:PREPARE PHASE**
Prompt:	Guiding questions:- ● Where is your data stored? ● How is the data organized? Is it in long or wide format? ● Are there issues with bias or credibility in this data? Does your data ROCCC? ● How are you addressing licensing, privacy, security, and accessibility? ● How did you verify the data's integrity? ● How does it help you answer your question? ● Are there any problems with the data?
Journal Entry:	Ans:- ● Data is stored on the Kaggle platform and it is of Fitbit's Fitness tracker data made available by Mobius. ● Data is organized in various CSV files with different kinds of variables. There are CSV files of both types of formats some are in long format and some are in wide-format ● There are no missing values that were checked in google sheets by the use of a filter on the range(A1:0941). But, there are issues of bias and credibility as this dataset is not comprehensive and current .there were no inconsistencies and missing values thus making the dataset somewhat

credible, this dataset should be used only for primary hints because the sample size is also low which makes it unreliable.ROCCC(Reliable, Original, Comprehensive, Current, Cited) was partially followed by this dataset.

- **Dataset Origin:** FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available by Mobius). Click here to access the dataset –> link: This Kaggle dataset contains a personal fitness tracker from thirty FitBit users. Thirty eligible FitBit users *consented* to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' behaviours and habits. These datasets were generated by respondents to a distributed survey via Amazon Mechanical Turk between April 2016 to May 2016.
- From the whole dataset, I am using Sleep Day and DailyActivity_merged CSV whiles for getting high-level insight into the fitness regime of the 30 users in this dataset.
- The SleepDay file had an invalid data format for the SleepDay column and it has been corrected.
- In, Big query cloud platform a new project is created and a dataset is created which is populated by the CSV files that I mentioned earlier.
- There are 940 observations in the daily activity CSV file and 413 observations inSleepday CSV.
- Below shown Query was used to find unique values of ID columns for both tables.

```
SELECT
    COUNT(DISTINCT (Id)) AS unique_IDs
FROM
`case-study-bellabeat-333505.Fitbit_Fitness_data.dailyActivity_merged`
```

RESULT:-33

```
SELECT
    COUNT(DISTINCT (Id)) AS UNIQUE_ID
FROM
`case-study-bellabeat-333505.Fitbit_Fitness_data.SleepDay_merged`
```

RESULT:-24

Other thoughts or questions:	Key tasks 1. Download data and store it appropriately. 2. Identify how it's organized. 3. Sort and filter the data.

Date:	**Course/topic:PROCESS PHASE**
Prompt:	Guiding questions:- ● What tools are you choosing and why? ● Have you ensured your data's integrity? ● What steps have you taken to ensure that your data is clean? ● How can you verify that your data is clean and ready to analyze? ● Have you documented your cleaning process so you can review and share those results?

Journal Entry:	● I have corrected the incorrect format of the data in the spreadsheet.
	● After that, In the big_query cloud platform, I identified unique participants on whom the analysis is carried out based on two CSV files that represent daily activity and sleep records.
	● Every column has the same type of value and there are no missing or NULL values.
	● There is some duplication of records though. so we first made a copy of the sheet to have a backup in case something goes wrong, click the sheet button next to its name on the right side on the bottom of the sheet after that select duplicate to make a copy of the sheet in the same spreadsheet. Then, These Duplicate records were removed by the use of the remove duplicate option in the data menu of google sheets.
	● The first row of both CSV files is for column names, thus freezing that row will help in keeping that row at 1st position only in case of sorting data.
	● There are a lot of 0 values where whole rows are made of 0 values for every column. Thus eliminating them will be the best course of action as they can skew results in our aggregation or any other type of technique we use during the analysis phase.
	● Below Code, was used to eliminate 0 value rows.
	● `SELECT`
	`    *`
	`FROM`
	`` `case-study-bellabeat-333505.Fitbit_Fitness_data.dailyActivity_merged` ``
	`WHERE`
	`    TotalSteps != 0 OR   TotalDistance != 0`
	● Data is clean as there are no missing values, data formats are consistent, and there are no duplicate records.

Other thoughts or questions:	Key tasks 1. Check the data for errors. 2. Choose your tools. 3. Transform the data so you can work with it effectively. 4. Document the cleaning process

Date:	**Course/topic:ANALYSIS PHASE_part1**
Prompt:	Guiding questions:- ● How should you organize your data to perform analysis on it? ● Has your data been properly formatted? ● What surprises did you discover in the data? ● What trends or relationships did you find in the data? ● How will these insights help answer your business questions?
Journal Entry:	● I used Rstudios for analysis and visualization of two relationships that I came upon. ● Those two correlations were as follows: 1. Total Distance vs Calories 2. Total time in bed vs TotalMinutesAsleep ● Firstly, I explored the CSV files for dailyActivity and SleepDay to familiarize myself with the data. then I started making some hypotheses as my data was already clean from previous phases. after that I loaded a library of tidyverse for the use of many functions it provides. ● Secondly, As I needed to see for myself if there is a relation between these correlations that I have taken into consideration, I started plotting appropriate columns with ggplot library. I choose a scatter plot for the visualization as it was the clear choice for comparison between two numerical variables with no real-time series involved. ● After Plotting I came to realize that, in both of the correlation considerations I got a positive trend by adding smooth lines to the scatter plot. ● This made me conclude that the variables affecting one another in an upward trend which can be inferred as follows: 1. As you have guessed, the more you cover the ground by walking, the more calories you burn. This was a simple test for practice but I wanted to find what percentage of people are outliers in this dataset but as the sample is small I didn't get any such person. 2. For the participants that consented to provide their data, I have seen that longer they are in bed more they sleep up to a certain point but after that, there is more increase in time on the bed than their sleeping time, This suggests that people are just entertaining themselves at night than usual because there are some records of participants who have more gap between time on bed and sleep time which can hazardous for their health. ● These tasks were only under preliminary analysis. I have planned to explore three more correlations from the dataset. ● As there were no categories to indicate the status of sleep cycles for our participates I added a Dataframe named Status which consists of categories of status made from TotalMinutesAsleep column of the SleepDay.csv.I divided sleep time into 4 categories and showed them with distinctive coloured points on a scatter plot so that audience can parse the

	visual more comprehensively and easily.
Other thoughts or questions:	Key tasks 1. Aggregate your data so it’s useful and accessible. 2. Organize and format your data. 3. Perform calculations. 4. Identify trends and relationships.

Date:	**Course/topic:PREPARE/PROCESS PHASE(Extended)**
Prompt:	What can I do to make my analysis more comprehensive to find new relationships between data points?
Journal Entry:	● To know if there is correspondence between sleep time and steps that a user takes. For that, I combined SleepDay and Activity Data frame to further investigate my question. After that firstly I converted total sleep time in minutes to hours by use of mutate function in R.Lastly, I plotted for Hours of sleep vs Total Steps. ● From the result of that steps, We can conclude that a large chunk of people who sleeps 5 to 10 hours a day is getting To the threshold of 10000 steps recommended by many Health Experts. Also, as sleep time increases people are getting more and more inactive in their Activities which can be seen by the Trend line suggesting a relatively downward trend from the 6-hour mark. But as the sample size is too small we can’t generalize this finding but this is to be true only for these 24 participates. (Note:-As combining data is an Inner join some participants’ records are deleted for the sake of accurate analysis otherwise, there would be many NULL values which are never good for analysis) ● Moreover, I selected the other 3 CSV files to extend my analysis. Those files are hourly steps,hourlyIntensities and hourly calories. ● I used a spreadsheet as a tool for data cleaning and data preprocessing. ● To begin with, I checked if there are any duplicate rows present which resulted in every row being unique. ● After that, I explored to know how many distinct Ids were present in the Id column. For this, I used the UNIQUE function and found out that there are 33 participants. ● Then there were some issues regarding splitting the column which contained the date and time of the Log. This field contained time data in an inconsistent format, some were in a 24-hour cycle and others were in 12 hours day cycle with am and pm indication .so to resolve this inconsistency I formatted the time to 24 hours cycle by first using split to take apart date and time into different columns and then lastly changing the format of time column to a 24-hour cycle.

	• These issues were consistent through all 3 CSV files that I selected and these issues were resolved by iterating the same steps for all CSV files. • Then I loaded these CSV files into the R studio where analysis and visualization is carried out by me.
Other thoughts or questions:	

Date:	**Course/topic: SHARE PHASE(many tasks of this phase are completed in the analysis phase thus only recommendations are to be provided here)**
Prompt:	What are your recommendations?
Journal Entry:	• 1. Based on the activity levels and amount of calories burned, users appear to burn more calories with more exercise which in turn will help them become leaner and lead a healthy life. so For that, Bellabeat can motivate users to exercise more through reminders and tips or insights to stay motivated. They could also offer an app dashboard for Activities they have logged in and a leader-board in which there will be an option for users to compete with their connections which will be fueled by a ranking system. Also, the company can give points according to steps the user have walked which can then be used by the user to get discounted items from the company's website. • 2. The data suggests the majority of users live either light or sedentary lifestyles, which may be due to a lack of motivation Thus there can be a feature of push notifications that daily sends motivational quotes to the user. As there is lacking motivation users should be positively reinforced when they complete a certain threshold of activity that will encourage The user psychologically. • 3. To promote better sleep habits, Bellabeat may combine reminders with an app that informs users of the best time to go to bed and wake up so they can feel refreshed in the morning and get enough sleep which will lead to better performance throughout the day and they will not feel burned out when trying to complete their daily step count. • 4. There is a dire need for a larger sample size in order to improve the Credibility of the analysis. • 5. There is a need to acquire current data in order to better reflect current

	consumer behaviour or trends in smart device usage. Collect data from internal sources to increase the credibility and reliability of the data sets which will produce accurate analysis regularly and also give insights into the changing behaviour patterns of the customers and that will answer the question of In what ways that particular group of people can be attracted towards our product/service by incorporating Cluster analysis.
Other thoughts or questions:	