Name: Aaron Gyapong

**Data Analyst Skills Assessment**

The attached Excel file contains a sample dataset of participants of a certain program in 2018. It contains 7 columns, including Participant\_ID – a unique identifier for each individual participant which must be formatted as six numeric characters. You should assume that this dataset is in its raw form and may require you to check for formatting issues, duplicates, data entry errors, spelling inconsistencies, or other necessary data cleaning prior to relying on it for analysis.

Additional instructions:

* In this exercise, assume that there is no other background information or colleague that you can consult about the dataset. When asked to produce a table or graph, you should decide the most appropriate format to convey the information requested.
* Please limit each response to a paragraph or less. Your answers should be written out in complete sentences, unless otherwise stated. You should note any limitations or assumptions behind your response.
* Recognizing that you may not have access to various software programs, this exercise does not require the use of programs besides Excel. At your discretion, you may use other software programs (e.g., SQL, R, Python) to complete this exercise, but **you must note all programs used in the space below:**

Programs used: \_Python Packages: Pandas, Numpy, seaborn, xlrd(excel file reader), Tableau, Jupyter Lab\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* ***Please include any code you wrote as part of your submission.***

Please answer the following questions using the sample dataset provided.

1. What steps, if any, would you taketo clean this dataset? Briefly explain how you might carry out each step. (Bullet points are acceptable. Your response does not need to be exhaustive.)
2. How many people participated in the program in 2018?
3. Create (1) a table and (2) a graph to depict the five most frequent nationalities among program participants in 2018. Include a brief description to explain the figures.
4. The director of the program asks you for a summary of the number of referrals made by program staff for participants in 2018 (as indicated by the variable *Number\_of\_Referrals*).
   1. What is the average number of referrals made per participant?
   2. Report any other summary statistics that will help the program director understand the number of referrals made for participants.
   3. What percent of program participants had at least one referral?
5. Imagine the sample dataset of 2018 program participants (the attached Excel file) is representative of a larger program dataset (“program data”) with 20,000 rows and the same 7 columns. Your supervisor asks you to merge the program data with another dataset (“administrative data”), matching on the same unique identifier for each participant (*Participant\_ID*). The table below shows the number of rows and the list of variables contained in each dataset:

|  |  |  |
| --- | --- | --- |
|  | Program Data | Administrative Data |
| Rows | 20,000 | 2,300,000 |
| Variables | Participant\_ID | Participant\_ID |
| Participant\_First\_Name | Name |
| Participant\_Last\_Name | Date\_of\_Birth |
| Nationality | Address |
| Gender | City |
| Date\_Participated\_in\_Program | State |
| Number\_of\_Referrals | Zip |
|  | Gender |
|  | Nationality |

The two datasets should have a one-to-one relationship, and you expect all of the participants in the program data will be contained within the administrative data. The new merged dataset will be used to explore regional effects of the program based on program participants’ zip codes.

For the following questions, you may reference specific programming languages or functions therein if it helps to clarify your answer, but you are not required to do so. If you mention any hypothetical issues that may arise, you do not have to explain exactly how you would resolve them.

* 1. What steps, if any, would you take *prior* to merging the two datasets to clean the data or check data quality? (Your response does not need to be exhaustive.)
  2. Optional: Explain how you would merge the two datasets. If you do not know how to merge the two datasets, you may leave this question blank.
  3. What steps, if any, would you take *after* merging the two datasets to clean the data or check data quality before it is used for analysis? (Your response does not need to be exhaustive.)

Answer Key

Please answer the following questions using the sample dataset provided.

1. What steps, if any, would you taketo clean this dataset? Briefly explain how you might carry out each step. (Bullet points are acceptable. Your response does not need to be exhaustive.)
   1. Open the file with a **read\_excel function in Jupyter Lab**, I can explore the data set **using Pandas and Numpy in python**.
   2. **Dataframe .Info()** allows me to explore data types of the data set.
   3. Change all necessary numerical values into proper data type **(P**articipant\_Id, 🡺 string, 🡺 int) and (Number\_of\_Referrals 🡺 int)
   4. Convert floats into ‘int’ types, and changed all Null values to ‘0’ int type in referral column
   5. Use **dropna() function** to drop nan in **Participant\_Id column**
2. There are **95 Participant**. Used tableau to display/count participant in 2018
3. Created a figure in tableau that shows the top 5 nationalities that was counted by participant
4. Number\_of\_Referrals. Mean enables to get an average of **1.6 referrals/participant**
   1. **1.6 is the Average # of referrals**
   2. **n/a**
   3. **46%**
5. The second data set will get loaded and clean with the same respect as the first data set.
   1. Analyze columns for proper data types. Dates to proper date columns, and numerical values are converted from strings and floats.
   2. **Use the Melt function in Python to merge both data sets**

df = pd.merge(df, df2, on='d', how='left') df = pd.merge(df, df2s, col1, col2,, etc..on=[how='left')

melt>>>>>>df = pd.melt(sales, id\_vars=['id', col1, col2, etc…..'], var\_name='d').dropna()

* **This allows me to merge on Participant\_ID as a reference ID, and merge all columns form data frame 2**
* **Reset the index of Participant\_ID as the primary key.**
* **Df1.head will allow me to see the entire first 5 rows of the newly merged data set.**