ViewPoint Preliminary Questionnaire

Part 1

Insert the data into our sql tables:

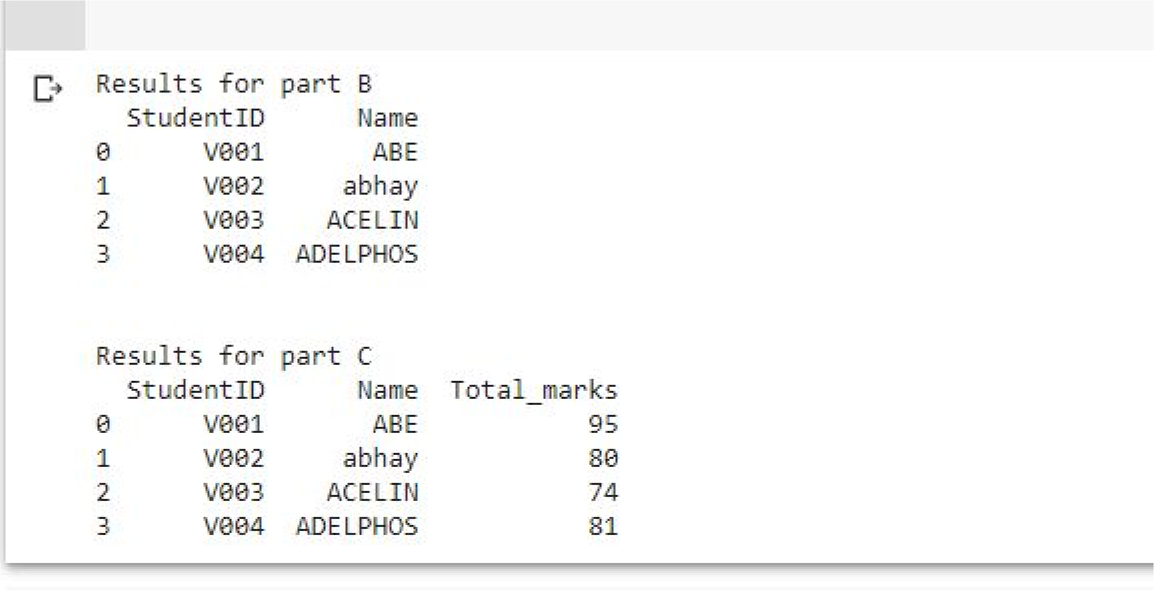
```
Insert the data:

INSERT INTO `name_table` (`StudentID`, `Name`) VALUES ('V001',
'Abe'),('V002', 'Abhay'),('V003', 'Acelin'),('V004', 'Adelphos');

INSERT INTO `mark_table` (`StudentID`, `Total_marks`) VALUES ('V001',
'95'), ('V002', '80'),('V003', '74'), ('V004', '81');
```

(a) **Answer:**

```
SELECT name_table.StudentID, name_table.Name
FROM name_table
INNER JOIN mark_table ON name_table.StudentID = mark_table.StudentID
WHERE mark_table.Total_marks > 80
```

(b) **Answer:** See google collab

(c) **Answer:** See google collab
Image: of my results



```
Results for part B
  StudentID      Name
0      V001       ABE
1      V002     abhay
2      V003    ACELIN
3      V004  ADELPHOS


Results for part C
  StudentID      Name  Total_marks
0      V001       ABE           95
1      V002     abhay           80
2      V003    ACELIN           74
3      V004  ADELPHOS           81
```

Question 2. Was placed in the git repository for this project. I would say I have little experience with data cleaning.

Question2.1.

Editing

[3]

```
# This is a Python 3 file that generates data for
# prospective data engineer candidates.

import pandas as pd
from random import random
nums = list()

for i in range(0,10000):
    ui = 0;
    for x in range(0,12):
        ui = ui + random()
    nums.append(5 + 3*ui)

df = pd.DataFrame(nums)
df.to_csv("data.csv", header=0)
print("done.")
```

done.

---

[4]

```
# Hmmm...  I wonder what this does?
import statistics
from matplotlib import pyplot as plt
%matplotlib inline

plt.hist(nums, bins=100, range=[0,50])
print("Mean: {}".format(statistics.mean(nums)))
print("Std:  {}".format(statistics.stdev(nums)))
```

1.Talking about the descriptive statistics in regards to the data that was provided above. First off let us define what a descriptive statistic is; A general definition for the term descriptive statics could be the following: A summary statistic that summarizes (quantitatively) features in regards to the collection of information. A descriptive Statistic is essentially a means of analysing our data collection.

Given:

1. Histogram: In general this shows the approximate distribution of our numerical data set. (A means of graphical representation).
2. Mean: The average of our data set (can be found by adding up all data points and dividing this value by the number of values in our set).
3. Standard Deviation(std): This is a measure of variation/dispersion in the given data set.

Thoughts (just looking at the data given):

1. We have an average value of 23.028 (Based off of the csv given this is for arbitrary data)
2. Based on a standard deviation value of 3 we have a decent spread in the data that was collected and synthesized. 99.7% of our data lies within 3 standard deviations of our mean(23.028). This could mean we have 99.7% confidence that would depend on the threshold of the project. (near certainty and certainty depend on the contrants of the project). EX: In certain fields of science only 2 standard deviations away from the mean are considered statistically significant. (just depends)
3. Just from looking at this histogram I don't see many outliers that show up visually. It might be smart to create a python script in-order to find the number of outliers that occur outside of a given range.
4. All and all from the data.csv looks like it has a solid distribution.

Final Thoughts:

5. Probably need more Quantitative figures + actual understanding of the data set to make further claims.

links: link text https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule

Question 3: If you were asked to impute null values in a column of a file that was 365 Gigabytes, what would you do? What tools would you use? What tools would you NOT use?

Missing data is common amongst large datasets and filling in these columns with null values is common in-order to make these data sets usable. In-order to conduct this task I would most likely want a tool that has low impact on the cpu and gpu of the machine. The reason for this is because we are working with an extremely large dataset that is 365 Gigabytes in size. Most likely due to the requirement of obtaining a tool that can carry out this task the fidelity will be lower in comparison to a tool that is extremely slow and gpu intensive.

Some methods:

- Use the C programing language and create a script to insert null characters in the col that we want to conduct impute data. I would not use python because the language is very smart but slow and our data set is extremely large. (I have no idea if this method would and have little experience with data cleaning + editing via a file the size of 365gb). Most likely this would not work.
- Attempt to find a 3rd party application that could conduct this sort of imputation on a file that is extremely large in size.

- Run some tests with a couple different methods such as Multiple Imputation: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4638176/ and see if I can apply some method for handling this missing data task.
- Could be a bad way but use python data frames (most likely would crash) and read the file line by line (to help combat the crash) and check to see if that line col needs a null character and add it if it is needed.

Would not use for this size:

- DataWig python library in-order to conduct this task (would be very slow)
- Prediction of missing values using Sklearn library. (would be very slow)
- Most likely would not use any python libraries for this task and would attempt to achieve this outside of python.
- Any tool that would try and read the entire file at once.

Question 4: What would you do if you were asked to do the above task every Thursday morning at 2:00am?

- If it was part of my job I would wake up every morning at 2:00am and run a theoretical script that I created in-order to automate this task; or run the c program that was created in-order to carry this task out.
- Wake up and utilize a 3rd party program in-order to conduct this task hopefully it can be automated so I don't need to wake up at 2:00am.
- Somehow create a script that can do this task every Thur morning at 2:00am so I don't need to wake up.

Question 5: Who is your favorite mathematician, statistician or computer scientist and why?

My favorite mathematician(theoretical physicists) would have to be Max Planck, one of the founders of quantum theory. The reason being is because without his discoveries I would not have been able to read about Quantum Harmonies:(the idea that the universe is built around harmonies).