ECE5464 SP24 - Prof. Jones – HW 1 UPDATED

Due Thursday, Feb 8, 2024 – 11:59 PM via Canvas

In this assignment you will perform some exploratory data analysis on a data file of diabetes patients. You will use Tableau to explore and understand the data, and to derive and show some data-driven insights. Here are the steps that you should follow.

PART I

1. Download the file "diabetic_data.csv" from the Files > Data folder on Canvas. This is a list of just over 107,000 encounters (health visits of some type) with patients who have diabetes. The columns are named. A few tips:
   a. One patient (with one Patient Nbr) can have more than one Encounter Id.
   b. Diagnosis fields (Diag 1, 2 and 3) contain (usually) ICD-9 diagnosis codes, which are standard codes in medicine for what a patient is diagnosed with (for example, 410.* is an acute myocardial infarction – a heart attack). At any encounter a patient can be diagnosed with more than one thing.
   c. There are lots of fields indicating what prescription drug a patient is reported taking: metformin, repaglinide, etc. The fields contain No, Steady, Up or Down, indicating dose.
   d. A csv file is a text file where rows of the dataset are on separate lines of the file, and column entries are separated by commas. For more information, take a look at: https://docs.fileformat.com/spreadsheet/csv/
2. IF YOU HAVE EXCEL: Read the data file into Excel and save as an Excel workbook. To do this, open Excel, and browse and open the .csv file. It will prompt you to choose how to import the file; it's delimited using commas. Once you open the file, save it as an Excel workbook (.xlsx extension).
   IF YOU DON'T HAVE EXCEL: Upload the data file into Google Sheets. If necessary, tell it how to import the file; it's delimited using commas.
3. Examine the columns in the data and identify each as either ID or feature, and determine the data type (numeric, categorical, etc.). Put this information in a new sheet in column format; here is a brief example of how yours should look:

FIELD	ROLE	TYPE
encounter_id	ID	numeric
patient_nbr	ID	numeric
race	feature	categorical
…	…	…

4. Sometimes (as in this file), impossibly large values are used to indicate missing values (where no real data is available). For analysis, we would rather have missing values in the data. So, replace any "impossibly large" (positive or negative) values with empty cells. Make note in your submission of how you defined "impossibly large", which values you replace, and how many replacements are made.
5. Calculate and display the (univariate only!) descriptive statistics for all of the appropriate columns in the data set. Do this in Excel, on a separate sheet of the workbook; you should calculate the following statistics (here is a sample):

Statistics	encounter_id	…	time_in_hospital	…
Mean	N/A	…	4.3959	…
Min	Your data here	…	1	…
Max		…	14	…
Range		…	13	…

Median		…	4	…
Mode		…	3	…
Variance		…	8.910868	…
Std Deviation		…	2.985108	…
Quartile 1		…	Your data here	…
Quartile 2		…		…
Quartile 3		…		…
# Missing		…		…

See the hints below for a couple of tips. Copy these statistics and paste into your submission. Note: you may need to split the table of summary stats up to paste it into your report and still have it be readable!

PART II

6. Save your spreadsheet, and open Tableau. Use your (modified) spreadsheet as the data source (the sheet with all of the data on it, not the summary statistics).
7. In Tableau, you are to answer the following questions; for each, paste the appropriate chart or table, along with appropriate explanation, into your submission. For each chart or table, be sure to include any color legends or other labels. Where questions are posed below, include a chart or graph to support your answer.
    a. Which three patients spent the most time in the hospital, and how many encounters did each have?
    b. Produce a histogram of the number of occurrences of each length of hospital stay.
    c. For each racial group, what is the percentage breakdown among the different A1C results (None, Normal, >7 and >8). Note that I want percentages of patients, not encounters. Produce a table showing the results.
    d. What percentage of the patients were readmitted (to the hospital) in less than 30 days, and in more than 30 days?
    e. What percentage of encounters resulted in a readmission in less than 30 days, and in more than 30 days?
    f. Produce a table showing the average length of time in hospital for each different A1C result, broken down by age group ([0-10], [10-20], etc.

PART III

8. Write a simple Python program to load the data from the spreadsheet into a Pandas dataframe. Print out the shape (the number of rows and columns) of the resulting dataframe to the console; copy and paste this console output into your submission. Paste your results as plain text – no screenshots or dark mode text.
9. Create a simple Data Quality Report using the approach shown in lecture 4. Your DataQuality Report should include the same fields as my example in lecture 5, slide 37! (You use it in the following way:
    a. include the file "StatsReport.py" in your project folder
    b. add the line **`from StatsReport import StatsReport`** at the top of your Python program
    c. create the report object: **`myreport = StatsReport()`**
    d. for each column in the data frame, get its label (the column name in text) and add the column to the report: **`report.addCol(thisLabel, df[thisLabel])`**
    e. note that you will need to add some functionality to the addCol() function to get the full set of statistics that I am looking for!
    f. when the report is done, write it to a file: **`report.writetofile(dirname + "filename.xlsx")`**
10. Paste all of your Python code at the end of your Word submission; paste it in as plain text (no screenshots or dark mode text).

SUBMISSION:

For your submission, paste all of the required information into a single Word (or pdf) file. Submit four separate files: your Word or pdf submission, your Tableau workbook, the .py file for the code you wrote, and the resulting data quality report. Submit your file using Canvas. Do NOT put your files into a zip file for submission; submit them as separate files. Do NOT submit your massive spreadsheet.

HINTS:

- In Excel, to calculate the mean of column A1 on the sheet called "data", where the dataset has 400000 rows, use the following formula:
  `=MEAN(data!A$2:A$400000)`
- For this size of a dataset, it doesn't matter much whether you calculate variance for a sample or the population – but use VAR.S() to be consistent with mine.
- In Tableau, you can generate maps by dragging longitude and latitude to the proper axis (column and row).
- Don't hesitate to use "Show Me" in Tableau to see which graph formats are available.
- Sometimes, to get the format you need, a Measure needs to be converted to a Dimension, and vice versa.
- When using Python to operate on large datasets, I find it useful to create a smaller file (I use the top 1000 rows) to debug my code. Once everything seems to be working, I operate on the large dataset for my final results.