**MIT 5032: Analytics Programming**

**Fall 2023**

**ASSIGNMENT 3**

The due date for assignment 3 is 12/3/2023 (11:59 pm). It must be submitted through the “Assignment 3” folder provided on Canvas.

**Part 1**

The goal of this part of the assignment is to help you further orient towards your final project. By this time, you must have finalized your set of specific hypotheses (with reasoning) and identified key attributes in the dataset(s) you have chosen for the project. You must have made progress by thinking through the programming and analytics techniques learned so far to analyze the data, derive meaningful information, and provide further insights into your list of hypotheses.

**Key Deliverables**

As part of this submission, please submit a rev 0.5…/0.6/… version of your analysis (in relation to your business hypothesis) and the relevant dataset. This will provide a chance for to instructor to get a better understanding of the business problem, hypotheses, specific approach taken, and provide necessary help if needed. While all of you are working extremely hard to learn the new material(s) and apply your knowledge in the programming and data analytics domain, the instructor wants to make sure he is available to provide prompt assistance, if needed.

Please plan on including the following items in the work-in-progress version of your report: *(Keep in mind this is a working version of the final report and is not expected to be a comprehensive version of the final product)* (15 points)

* Problem Statement
* Revised Hypotheses (approx. five) and Reasoning/Intuition
* Data Source(s) and Description
* Data Preprocessing Steps (so far)
* Specific Approach/Analysis (so far)
* Specific Outcomes/Insights (so far)
* Summary and Conclusions (so far)
* Key Learning (so far)
* References (so far)
* Specific Next Steps

Please specify your name (first and last name) in this part one of the assignment submission documents/attachments. The attachments name should follow the following convention:

MIT5032\_ASSIGNMENT3\_PART1\_YOURNAME.DOCX (detailed report including appropriate pictures)

MIT5032\_ASSIGNMENT3\_PART1\_YOURNAME.R (R script file with comments and ready to execute code)

**The report will be graded according to the following criteria:**

* Exhibits convincing range and quality of knowledge, having done appropriate research.
* Identifies and thoughtfully analyzes and evaluates major points of view using the items listed above in the previous section.
* The writing style is understandable and organized in explaining research investigation results and supporting conclusions (so far).
* The dataset used for the project is in conformance with the requirements listed in the project guidelines.

**Part 2**

The data in the file - BillingInfoAssignment.csv represent billing information of patients who visited the Highland Urgent Care Clinic. The file contains each patient’s first name, last name, age, and sex. After their last visit, the patients were mailed a survey to gather patient satisfaction scores. These are represented in the opinion column. A score of 1 is very poor, 2 is poor, 3 is neutral, 4 is good, and 5 is very good. The billing department has supplied the insurer and the charges billed for the last visit.

The clinic administrator uses the VitalInfoAssignment.csv file to enter the number of prior visits, date visit, the time spent in the clinic, systolic blood pressure, diastolic blood pressure, along with the first name and the last name for each patient.

You have been recently hired on as an intern in Highland Urgent Care Clinic and will spend the next few hours looking at this data.

1. You would like to create a new dataset (using R code) that contains the unique patients who appear in both datasets. Note that both datasets have two columns (A and B) in common i.e. columns representing the first name and the last name. However, the spellings/text of these column names don’t match. (3 points)
2. You would like to create a new dataset (using R code) that contains all the patients appearing in either dataset. (2 points)
3. Write R code to find patients records where insurance starts with characters “Se" in the merged dataset generated in part a. (2 points)
4. Write R code to compute overall (combined) mean of diastolic and systolic blood pressure using the merged dataset generated in part a. (2 points)
5. Write R code to delete the last row in the merged dataset generated in part a. (Don’t use any hard-coded index in the R code) (2 points)
6. Write R code to create another column “YearVisited”, extract the year information (for example 14 from 4/18/14) from the “date” column, and store the year information in the newly created column “YearVisited”. Use the data generated in part e. (2.5 points)
7. Sort the data by ChargesInDollars (in descending order) using the data generated in part f. (2 points)
8. Find unique values of the insurance type using the data generated in part g. (2 points)
9. Write R code to delete the columns – PriorVisits and Opinion in the dataset generated in part g. (Don’t use any hard-coded index in the R code) (2.5 points)

It is important that you specify your name in this part of the assignment documents/attachments. The attachments name should follow the following convention:

MIT5032\_ASSIGNMENT3\_PART2\_YOURNAME.DOCX (detailed report including appropriate pictures and output of R command)

MIT5032\_ASSIGNMENT3\_PART2\_YOURNAME.R (R script file with comments and ready to execute code)

The following criteria will be used to grade part two of this assignment:

* Data are explored and analyzed in many different ways.
* Calculations are detailed, accurate, and answers are correct.
* Graphs/pictures are labeled.
* A rigorous approach is used to solve a specific problem.
* Good coding practices are followed. For example, code are functional, readable, commented, and error-free.
* Convincing conclusions are drawn.
* The writing style is understandable and organized in explaining investigation results and supporting conclusions.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* NOTE \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Please feel free to consult your instructor by email or phone, if you have questions or need further assistance!!