B3T1101 – Data Engineering and Management

Assignment #3

Due Date: March 31st, 2022, @ 23:59 (Hard Deadline due to block end dates)

Objectives:

- Develop confidence on Data Analytic Skills
- Get your hands dirty with a real-life dataset
- Apply the skills learnt so far in a practical problem setting
- Experience with Data Visualization
- Experience possible problems that you may encounter when dealing with data
- Experience developing a workflow

Introduction

In this assignment, you will work with a real-life dataset, develop an understanding of the data, and produce a document summarizing your findings. We hope that by working with a real-life dataset, you will develop a better understanding of how to deal with data.

For this purpose, you will work with a Diabetes dataset in this assignment. The dataset you will work on is a publicly available machine learning dataset, accessible via the following website:

https://archive.ics.uci.edu/ml/datasets/diabetes

Alternatively, you can also access the dataset via Kaggle:

https://www.kaggle.com/ealtintas/uci-machine-learning-repository-diabetes-data-set

Both Websites contain README documents, and I suggest you look at those documents for developing an understanding of what is in the dataset. **There is also a ".zip" version of the same dataset on Canvas, under Modules/Assignment Related Documents tab** since the dataset is stored in a ".tar.z" format (which is more easily extractable for Mac and Linux machines, but not on Windows machines). However, I strongly suggest you also look at the Webpages for more insight, regardless of your choice.

The dataset contains three different document collections: the data of 70 distinct patients stored in separate files, a readme documentation providing in-dept details about the diabetes and its effects, and a Data-Codes file, which is required to understand the numbers that you will be seeing in the data files. I suggest you read the documentation, look at the codes, and open at least one of the patient data files on a word processor and look at the data before starting: It is straightforward to understand, but quite different than what we have used in the course sessions. **For one, the columns in the file are Tab-separated**.

The entries in the data columns are stored in a form that different measurements are combined into two columns, one representing the values, and one representing data codes for the corresponding values; something we have encountered in the course before. (However, we will approach it slightly differently this time)

Preliminaries

In this assignment, you will need several R packages from the past modules. These are:

- `dplyr`
- `reshape2`
- `Ggplot2`

In this assignment, you will also need to deal with dates and times; something that we have not done before. Apart from common data types, R programming language treats date and time as if it is a data type (so that you can compare, or order them). So, you will need to convert strings into a date and time format to compare them. For this purpose, I suggest you look at three functions:

- `as.date( )`
- `as.POSIXct( )`
- `format ( )`

Also, to combine date and time fields, you may need the `paste ()` function for strings, which combines/concatenate two strings.

- `Paste ()`

Finally, as the data collection is a real-life collection, you may (will) encounter several problems. To tackle those, you will need the distinct( ) function from the dplyr package.

- `distinct ()`

You can find extensive explanatory documentation for these functions in R, or on the Web.

Some Domain Knowledge

To perform analysis on a subject, we need to have some background knowledge about the subject. In this section you can find a brief description of some underlying concepts about diabetes.

What is Diabetes?

"Diabetes is a progressive and life-long disease due to the decline of the insulin hormone secreted by pancreas or deficiency of the utilization of the insulin hormone. The effect of diabetes manifests itself as a reason of cells being unable to process the glucose in the bloodstream. As a deficiency having a close relationship with the glucose levels in the bloodstream, the diabetes usually increases overall level of blood sugar in the body over long periods of time, and thus causing many additional illnesses. Additionally, the low blood sugar levels can also distort bodily functions of human beings. To this day, there is no known cure for diabetes.

One of the biggest problems in Diabetes treatment is that the patients need to constantly monitor their blood sugar levels. The diabetes patients are forced to take lifestyle choices such as what and how much to eat, or how much medication to take before measurement. Predictive algorithms that can represent the ramifications of different choices would improve the life expectancy and quality of a patient considerably. "

How does Blood Glucose Levels Behave for a Healthy Person?

The Blood Glucose level is the prime indicator to identify Diabetes patients. In the medical terminology, the concentration of blood glucose is measured by mmol/L (of glucose in a Litre of blood). The values in the dataset also indicate blood glucose measurements according to this style of measurement. Figure 1 represents a graphical illustration of how blood glucose changes during a day.

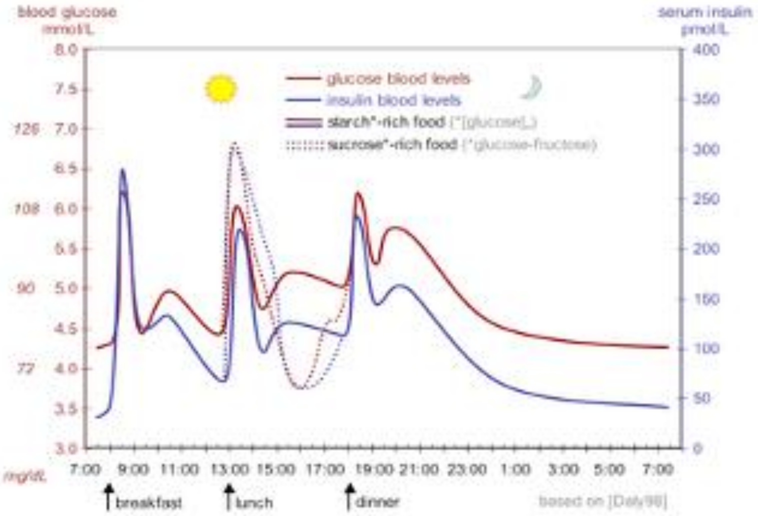

Figure 1: From Daly et al. (1998), American Journal of Clinical Nutrition

The blood glucose levels for a healthy person show variation throughout a day: It is usually lower during inactive periods (i.e., night-time) and relatively higher during active periods (i.e., daytime). Food intake during the day also have a significant effect on measurements, as the blood glucose usually requires some time to return to normal levels, typically 2 hours after food intake.

Many studies agree that the following blood glucose values are considered normal for a healthy person:

- In the morning (early, without any food intake): 70 – 100 mmol/L
- 2 hours or more later than a meal during the day: 125 – 150 mmol/L
- Within 2 hours of a meal during the day: < 200 mmol/L

Values over 200 mmol/L usually indicate a medical condition, and values under 70 indicate a condition called hypoglycaemia which is a different, but a life-threatening condition.

There are of course many things that effect the measurements: Fresh food usually have a healthier effect, having exercise also have positive effects, and drinking sugary beverages may change the blood sugar to very high levels instantly. Hence doing analysis over a data collection is complicated. However, we will not go into such analysis in this assignment.

Note: Of course, these values are true only for healthy persons. The dataset we are dealing with is collected from hospitalized patients, and you will see that (almost) none of the patients will exhibit such regular behaviour. You may often need to give decisions with unknowns.

Are we going to work with all the patients?

No**, each group will be assigned 5 patients**. The assignment of patients to groups will be performed in a round robin fashion. That is, Group 1 will deal with patients 1-5, Group 2 with patients 6-10 and so forth. Group 14 will deal with patients 66-70 and then for Group 15, the assignment returns to the beginning and continue (patients 1-5 again). Each group should determine the patients that are assigned to them and work with those patients.

Note: Since the collected data involves patients, many of them exhibit different blood glucose levels and behaviour. Hence, different groups may have different conclusions depending on the data they are dealing with.

Note II: Some patients may have missing data, or missing data for specific codes (or may have problematic records). That is why each group will be working on 5 patients.

What is our task in this assignment?

Assume that you are working as the analysis team in a company, and the company is aiming to build predictive models for Diabetes. **One of the research questions was whether to build generic models for all patients, or to build individual models for each patient**.

The machine learning team is assigned to work with this data collection. However, there are many unknowns, and the data is not ready for fast development. As the data analysis team of the company, your task is to create a cleaner format for the data and provide some insight about the quality of the dataset. For this purpose, you are tasked with the following:

Importing and cleaning the data:

- Make sure that all fields are of an appropriate type. For values, we want a numeric format, for codes we want a numeric format (for easy subsetting), and for comparing dates we want to have an additional field which combines Date and Time as a date format.

- For some patients, there may be misspellings in the Value field (although it should have been numeric), which may lead to the Value field being read as a character type. Convert such cases into numeric type and delete any rows that result in a missing value from the dataset. (You should also report how many instances you need to delete in the report)

Note: As such problems does not necessarily exist in all patients, you may not encounter such a problem at all. That does not mean you are missing something.

- Provide meaningful names to the columns of the data frames.

Derive individual summaries for patient data:

- min, max, average, and standard deviation will give us a good indicator of whether patients’ blood sugar values are similar in the long term or not. Using subsetting with the codes, derive these values for each patient.

Derive collective summaries for the patient data:

- For this purpose, you will merge the patient data of all the patients into a single data frame and derive min, max, average, and standard deviation of the merged data.

- Before merging, create a new column representing patient Ids in your combined dataset. This would serve two purposes. First, it allows you to identify different patients. Second, to visualize patient data on the same plot, this ID field will provide a categorical information. Note that patient id needs to have a factor data type, so that it can be used as categorical information. Convert it into a factor. (Alternatively, you may leave it as a numeric field and plot the graph to see what the effect of factors are)

- Comparing the collective and individual summaries, write your initial conclusions.

Perform a validation study on the data for the research question:

For this purpose, we would like to see whether the collected data makes sense (That is, whether measurements are somewhat exhibiting a similar behaviour to that of in Figure 1). Although it is known that the data is collected from patients, blood glucose levels should still rise after meals, and it should be relatively low during the night.

- In order to plot the daily behaviour of blood glucose measurements throughout a day, we need to first categorize each measurement with respect to the time they are taken during a day. For this purpose, create a new column named Hours. By using the mutate function, extract the hour from the second column of the datasets and assign it as the value of the new Hours column. (You will need to use both format, and as.POSIXct functions for this purpose). Convert the Hours data type into numeric so that it is displayed in an orderly fashion in the plots.

https://www.geeksforgeeks.org/how-to-extract-time-from-datetime-in-r/

provides a short description of how this transformation can be performed.

- The Hours column (attribute) can be used to group all measurements taken at a particular hour, so that we can draw a graph representing the general behaviour of blood glucose levels for a patient or for multiple patients. I also suggest getting the average of all measurements taken in a particular hour to represent multiple measurements as a single data point in the graph.

- Draw a 24-hour blood glucose change graph for each patient and collective data. Can you observe relatively low measurements during night-time? Can you observe possible deviations due to meal intake? Report your findings along with supporting graphs.

- Examine patient data. You will notice that blood glucose measurements are marked with different codes in the dataset (Ranging from 48 to 64). Note that you should only work with the blood glucose measurements in this part of the assignment. Thus, before building your graphs, make sure that you are subsetting the measurements only.

Individual data or collective data?

With the graphs that you have built above, refine your conclusions about whether collective data represent individual patients or not.

Cleaning/preparing the data:

- Notice that, there are codes other than blood glucose measurements. This is contradictory to the definition of tidy data. Hence, such codes need to be moved to columns.

- Replace all blood glucose measurement codes (48...64) with one common code. (For example, I used an unused code, 40, in my solution)

- Replace all insulin intake measurement codes (33…35) with one common code.

- Now your code column represents different categories, and value column represent their respective values. Divide/Cast different codes into columns.

Important note: The data collection you are dealing with is not a “sanitary/clean” dataset. For some patients, duplicate entries are stored during the documentation process. This may impede your casting process by introducing error messages/producing faulty columns. It is possible to identify such problems using the distinct ( ) function before you start your analyses.

More important note: Save your progress regularly. More importantly, always validate your progress after each operation, no matter how simple it looks. The assignment is longer than what we have covered in the Workshops or lectures, so it is easy to get lost track.

Also Draw a blood glucose change graph for each patient/collective data for Dates

That is, draw a graph where, x axis shows the Dates and y axis represent blood glucose measurements.

Is that useful in any way? Can you derive some conclusions? Describe in your report.

WHAT TO SUBMIT

For your submission you need:

- To prepare a report with your findings.
- Your report should describe with which part of the data you are working with.
- Your report should contain at least one visualization (figures).
- You need to use at least 2 geoms, 2 aesthetics, and 1 theme related customization in your figures.
- Your figures should be annotated and clear.
- You need to provide conclusions in your report, and the conclusions should be supported with the information you are providing with your report.
- In order to support your findings, you can create tables, have discussions that include numbers derived from your programs.
- Your report should be in the form of a formal report: With a short introduction and written in an understandable manner. It should not be merely a document of responses to the tasks in the assignment.

Submit your R scripts and Rdata files along with the report. However, note that we will primarily grade the report, and only check your code only for correctness of your final data frame structure and annotation. Rest of the code is appreciated in case there is a misunderstanding but is not required for submission nor will be graded.

Save the above document as a zip file. Name the zip file as

GROUP<GID>.zip where, <GID> is your Group ID, or group number.

HOW TO SUBMIT

One submission on behalf of the whole group. That is, each group can select a person to submit, and that person can submit on behalf of the group

IN CASE THERE ARE COMMUNICATION PROBLEMS

Please note that although this is a group assignment, and intra-group communication is the responsibility of the students, it is possible for groups not being able to communicate with all the group members. Although the students are responsible to attempting communication with their peers, in case there is no response from other parties (it is possible for students to withdraw, or other reasons), groups should proceed with the assignment and document such group related problems in their report. We will take the condition of each group into consideration while grading.

Most importantly, we hope that you will have fun with the assignment!