DTI5126

Assignment # 1: can be solved individually or in groups of 2.

Part 1: Answer the following questions in a doc/pdf file

Source: Data Mining: Concepts and Techniques 3rd Edition Solution Manual Jiawei Han, Micheline Kamber, Jian Pei (Ch2.)

1) Data quality can be assessed in terms of several qualities (e.g., accuracy). Enumerate some of these qualities and discuss how the assessment of data quality can depend on the intended use of the data, giving examples.
2) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.
3) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (a) What is the mean of the data? What is the median? (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). (c) What is the midrange of the data? (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data? (e) Give the five-number summary of the data. (f) Show a boxplot of the data. (g) How is a quantile-quantile plot different from a quantile plot?

Part 2: Work with your own data set:

For this part, start by picking a dataset either from your own research or something interesting available online. You will then use this dataset to practice what you've learned with R so far. There are no restrictions on the data set as long as you can use it to solve the following questions.

1) Describe the dataset you are using, both in terms of the **content** (what is this data measuring? how was it collected? what kinds of research questions are you hoping to use it to answer?) and in terms of its **format** (what type of file is it saved in? what if it is in a flat file, is it fixed width or delimited? if it is delimited, what is the delimiter? if it is binary, what is the program that would normally be used to open it?).

2) Include code that reads the data into R and assigns it to a dataframe object that you can use later in the document. Explain in the text which R function you used to read in the data (e.g., `read_csv`) and which package it came from (if it was not a base R function). If there were any special options you needed to use (e.g., `skip` to skip some rows without data), list those and explain why you used them. Next, include some code to clean the data (e.g., rename columns, convert any dates into a "Date" format). You can filter to certain rows if you would like, but do **not** filter out missing values, as we'll want to learn more about those later. (5 points)

3) Describe the dataframe you just read in. How many rows does it have? How many columns? What are the names of the columns? What does each row measure (i.e., what is the unit of observation).

4) Pick three columns of the dataframe. Use the summarize function to get the following summaries of these columns: (1) minimum value; (2) maximum value; (3) mean value; (4) number of missing values. If there are missing values, make sure you use the appropriate options in summarizing these values to exclude those when calculating the minimum, maximum, and mean. Assign the result of this summarize call to a new R object, and print it out, so these summaries show up in your final, rendered Word document.
5) Create two plots of your dataframe. One should use a "statistical" geom (e.g., histogram, bar chart, boxplot) and one a "non-statistical" geom (e.g., scatterplot, line plot for time series). Explain why these plots help you learn more about this data and about the interesting research questions you're hoping to explore with the data. Be sure to customize the final size of each plot in the Word document using the fig.width and fig.height commands. For each plot, also be sure to customize the x- and y-axis labels. Finally, explain how each plot is following at least two of the principles of "good graphics" covered in week 4 of the course (Chapter 4 of the book)—if necessary, use ggplot functions and options to make the plots comply with some of these principles.
6) Show a count of all missing values in each column. If you don't have missing values in your data set, artificially add some by erasing some values in some rows.
7) For every numeric column with a missing value, replace the value with the column mean. For every categorial column, replace the missing value with the most frequent category.
8) Pick one numeric column and normalize it.
9) Pick one categorial column and convert it into several dummy columns.
10) Perform three different data preparation operations from those discussed in class.
11) Discretize another numeric column using binning.
12) Produce a **matrix of scatter plots** (use pair function).
13) Plot two other graphs of your choice and discuss you selected them.

Part 3 read the following paper:

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2201-2206).

Q1. Write two paragraphs discussing a number of error detection and error repairing approaches.

Q2. Write two paragraphs about the need of data cleaning for machine learning approaches. You need to provide two references that you used.

Deliverables:

- a word/pdf doc (file name St1lastname_St2lastname_A1) with all answers and snapshots.

- Data file (CSV) for part 2 A1.csv

- R script (for parts 1 and 2) A1.r