DA5020 – Practicum I

Practicums provide you with an opportunity to dive deeper into a data analytics problem. In this practicum you will practice data loading, parsing, shaping, and exploration. This is more than an assignment and requires creative problem solving and patience, a skill that is essential to data analytics.

It is anticipated that the average student will spend 8-12 hours on this practicum. Use this practicum as an opportunity to practice your coding skills on solving real-world problems (without the confines of an assignment). This is very close to an actual problem you might encounter in a data science or data engineering job. The questions are guidelines as to the minimum expectations. However, you are encouraged to explore the data and integrate data from other sources to support your stance.

This is a group practicum which means that you may choose to work in groups of up to three students. **You may fully collaborate and submit the same work**. However, you must put all students' names on all submitted work. If a group member is not adequately contributing, the remaining team members may "vote to eject" the student from the team by emailing me the reason. In such an event, the team member who was "fired" must still complete the project individually by the due date.

Useful Resources

- Chemical Dependence Treatment Program Admissions: dataset | data dictionary
- NYS County Codes
- Kumar, S. (2017, Oct. 2). The Art of Story Telling in Data Science and how to create data stories? Analytics Vidhya.
- Line chart annotation with ggplot2

Part 1: Answer the following questions.

1. (2.5 pts) Create a dataframe with **4 variables.** The 4 variables should be doctor_type, doctor_lastname, location, AVG_Rating .

The variable doctor_type should contain 4 inputs (PCP,

Psychiatrist, Surgeon, Anesthesia)

The variable doctor_lastname should contain 4 inputs (Smith,

Dame, Jones, Zayas)

The variable location should contain 4 inputs (MA, ME, NH,

VT)

The variable AVG_Rating should contain 4 inputs (7,9,8,9)

Print the dataframe and include a screenshot

2. (2.5 pts) Using the dataframe above…

Select row 1 in column 2, what was selected?

Select rows 2 through 4, what was selected?

Select the last column, what was selected?

3. (2.5pts) Using the dataset Mtcars create a scatterplot showing the relations between any two variables of your choosing. Explain why you picked those variables. Each point should also be based on a color scheme of your choosing.
4. (2.5pts) Using the dataset MTcars perform a summary statistic on the dataset and calculate the pearson coefficient of the correlation R picking two variables of choice. Explain why you picked those variables and explain the purpose of a pearson coefficient.

Part 2: Practicum Tasks

The Office of Addiction Services and Support publishes a dataset on reported admissions of people in certified chemical dependence treatment programs throughout New York State (NYS). This dataset includes the number of admissions to certified treatment programs aggregated by the program category, county of the program location, age group of client at admission, and the primary substance of abuse group. For more information on the dataset, visit the following website.

You are given the task of performing a comprehensive analysis of the admission statistics from 2007 to 2019 and summarize your findings with an accompanying narrative that explains your process-flow.

1. (5 pts)Load the data, directly from the URL, into your R environment.

2. (10 pts) Evaluate the dataset to determine what data preparation steps are needed and perform them. At a minimum, ensure that you discuss the distribution of the data, outliers and prepare any helpful summary statistics to support your analysis.

3. (30 pts) Structure the data relationally, at a minimum, you should have four tibbles or data frames as follows:
   - ***county*** which contains the name of all counties and their respective county code (which is the primary key). For example:

county_code	county_name
AL	Albany

Note: ensure that your data frame does not contain duplicate counties and ensure that your dataframe contains all counties in the data.

- ***program_category***: which contains a unique identifier and the name of the program category. For example:

program_code	program_category
CR	Crisis

Note: ensure that your data frame does not contain duplicates. The codes can be numeric (e.g. auto incremented).

- **primary_substance_group**: which contains a unique identifier and the name of the substance. For example:

substance_code	primary_substance_group
H	Heroin

Note: ensure that your data frame does not contain duplicates. The codes can be numeric (e.g. auto incremented).

- ***admissions_data*** which contain the details on the reported number of admissions — excluding the data that resides in the **county,** ***program_category*** and **primary_substance_group** tibbles/data frames; you should instead include a column with their respective foreign keys. For example, if this was your original dataframe:

year	county_of_program_location	program_category	service_type	age_group	primary_substance_group	admissions
2007	Albany	Crisis	Medical Managed Detoxification	Under 18	Heroin	4

The names should be substituted with their respective foreign keys as follows:

year	county_of_program_location	program_category	service_type	age_group	primary_substance_group	admissions
2007	AL	CR	Medical Managed Detoxification	Under 18	H	4

4. (15 pts) Create a function called **annualAdmissions()** that derives the total number of reported admissions that transpired each year, for the entire state of NY and displays the results using a line chart. Annotate the chart to show the year with the highest number of admissions. Note: the year should be on the x-axis and the number of admissions on the y-axis. Explain the chart.

5. (15 pts) Analyze the percentage of admissions for each county and visualize the results for the top 5 counties using a bar chart. Explain the results. Note: ensure that you join any related dataframes/ tibbles.

6. (15 pts) Filter the data, using a regular expression, and extract all admissions to the various "**Rehab**" facilities; i.e. your regex should match all facilities that include the word rehab, rehabilitation, etc. Using the filtered data, identify which substance is the most prominent among each age group. Visualize and explain the results.

All charts should have the following:

- An informative title (and subtitle if applicable)
- Labels on the x-axis and y-axis that indicate the units of measurement.
- A caption that indicates the purpose of the chart.

Submission Details

- Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd file, DA5020.P1.FirstName.LastName.Rmd and your PDF/HTML DA5020.P1.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.

  • The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it.) Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.

- Not submitting a knitted PDF or HTML will result in reduction of 30 points.
- Not submitting the .Rmd file (or both) will result in a score of 0.
- **Include the names of all group members in your RMD file.**