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.

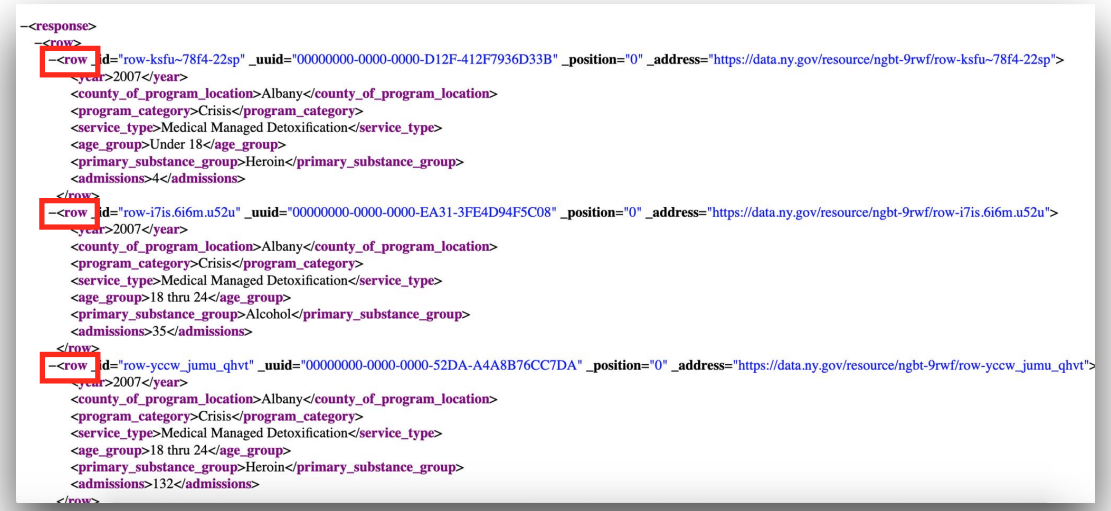
If you are working in groups, you can self-signup in Canvas or notify me via email by Thursday and I will create the group for you. Ensure that you include your name and the name(s) of your group member(s) in the email and cc them.

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. Load the data, directly from the URL, into your R environment. Here is an example of the XML data for your reference:



```
−<response>
  −<row>
    −<row _id="row-ksfu~78f4-22sp" _uuid="00000000-0000-0000-D12F-412F7936D33B" _position="0" _address="https://data.ny.gov/resource/ngbt-9rwf/row-ksfu~78f4-22sp">
        <year>2007</year>
        <county_of_program_location>Albany</county_of_program_location>
        <program_category>Crisis</program_category>
        <service_type>Medical Managed Detoxification</service_type>
        <age_group>Under 18</age_group>
        <primary_substance_group>Heroin</primary_substance_group>
        <admissions>4</admissions>
    </row>
    −<row _id="row-i7is.6i6m.u52u" _uuid="00000000-0000-0000-EA31-3FE4D94F5C08" _position="0" _address="https://data.ny.gov/resource/ngbt-9rwf/row-i7is.6i6m.u52u">
        <year>2007</year>
        <county_of_program_location>Albany</county_of_program_location>
        <program_category>Crisis</program_category>
        <service_type>Medical Managed Detoxification</service_type>
        <age_group>18 thru 24</age_group>
        <primary_substance_group>Alcohol</primary_substance_group>
        <admissions>35</admissions>
    </row>
    −<row _id="row-yccw_jumu_qhvt" _uuid="00000000-0000-0000-52DA-A4A8B76CC7DA" _position="0" _address="https://data.ny.gov/resource/ngbt-9rwf/row-yccw_jumu_qhvt">
        <year>2007</year>
        <county_of_program_location>Albany</county_of_program_location>
        <program_category>Crisis</program_category>
        <service_type>Medical Managed Detoxification</service_type>
        <age_group>18 thru 24</age_group>
        <primary_substance_group>Heroin</primary_substance_group>
        <admissions>132</admissions>
    </row>
```

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). When creating the county codes, you can use the data from the NYS County Codes in the **Useful Resources** section or create your own unique code** for each county. For example:

county_code	**county_name**
AL	Albany

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

** If you provide your own code, you must justify the logic for the IDs/Codes that you selected.

   - ***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 program codes can be alphanumeric.

   - **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 substance codes can be alphanumeric.

   - ***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 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 variables should be substituted with their respective 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. (10 pts) Analyze the percentage of admissions for each county and visualize the results for the top 10 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.

7. (20 pts) Using the "rehab" data from question 6 above, perform a detailed analysis to identify any patterns or trends with respect to the admission to rehab facilities in certain **counties** and **substance groups**. Explain your observations. Note: ensure that you join any related dataframes/tibbles.

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 and any other supporting files 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.**

Useful Resources

- Chemical Dependence Treatment Program Admissions: dataset | data dictionary
- You can obtain NYS county codes from the following resource and select the one that represents your data the best:
  - Option 1: NYS County Codes from NYS Open Data Portal.
  - Option 2: NYS County Codes from the DOT
- 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