# Dataset Exploration Project Part 4 Rubric

BDAT1005-22F – Mathematics for Data Analytics

## Dataset Exploration Project Part 4 (DE pt4) – Inferential Techniques & Final Report

**Assignment Date:** Week 9

**Due Date:** End of day Sunday after **Week 13**

**Evaluation Weight:** 12% of course

## Project Description – Please read the entire rubric before proceeding!

This is the fourth and last part of your four-part individual term project. If possible, you should use the same data set for all parts – otherwise, you will have to re-do previous work. **In DE pt 4, you will continue to work on step VI, update your previous work as appropriate, and complete your final report covering all steps.** Your final report should be comprehensive – your professor will not look at previous submissions in grading this one.

I. Define your research question(s)
II. Define your ideal dataset
III. Define or discover what data could be available
IV. Obtain the data
V. Perform univariate data analysis, including examination, visualization, and preliminary cleaning and coding
VI. Develop hypotheses and perform appropriate bi- and multi-variate statistical analysis
VII. Visualize and interpret results (including tables, charts and diagrams)
VIII. Verify / challenge results
IX. Communicate results
X. Create reproducible instructions for others to recreate results

We have studied all of these steps; you should be able to proceed with confidence.

You've got your dataset; you've described it, looked for inappropriate values, performed univariate descriptive statistics on many of the variables, developed and refine FINER research questions and applied a selection of hypothesis tests. **You must continue to use MS Excel for data manipulation, analysis and visualization, and your discussions (including appropriate visualizations) must be in MS Word format.** Here are the steps of DE pt 4:

1) Continue to use the same appropriate data source, with a minimum size of 1,000 records across at least 8 separate variables, with a good mix of qualitative and quantitative data types. **If your dataset was deemed inappropriate** in feedback from parts 1-3, you won't be able to resubmit those, but you will need to go back and re-do that work on a new dataset in order to proceed.
2) **Update your part 3 content** as suggested by feedback from parts 1-3 and/or according to your own further work.
3) Develop inferential techniques you have learned through Lesson 7, which you believe can help answer one or more of your research questions. Apply the techniques and describe the results, including appropriate visualizations (tables, charts, graphs, etc.). **You must apply at least two inferential techniques; at least one must be predictive**, and they can be either interpolation (regression analysis) or extrapolation (forecasting). NOTE: Your report (and spreadsheet) must include full univariate analysis (inspection for inappropriate data, discussion and visualization of frequency distribution, etc) on every variable you are using for any of these techniques as well as for your hypothesis tests; if you didn't do that already, you must do it now.
4) Reconsider your FINER questions, your assumptions, and your need for extra data according to your further analysis.
5) Continue to track your analysis –record your thoughts about what to study; what data you found and where; what questions you're asking and what assumptions you're making, and why. *Tracking is not about recording times or dates, or who said or did what – it's about enabling other analysts to replicate your work.*
6) **Write your final report.** Once your data analysis is complete, you should know whether or not you have answers for your original (and developed, if any) questions, and you're going to share and present them.
    a. Provide an introduction to the topic you are investigating. Describe the industry or enterprise or topic of investigation, and the problem or opportunity or question you are trying to address. Why are you analyzing

this topic? What need are you addressing? Why would your audience care? What internal and/or external drivers are you trying to satisfy? What kinds of questions did you chose to address in your analysis? Did your focus change as you analyzed the data? You may describe this through SMART indicators if appropriate. What were your expected outcomes? How would your analysis further the need you are addressing, in the context of the industry or enterprise or topic described above?

b. Describe your data set. Where did you get it? Does it satisfy the project requirements for numbers and types of variables and number of records? How did you think the data could help you to achieve the goals laid out in your introduction?

c. State your FINER research questions. How did they change over time? Did you add focusing questions?

d. Describe your data analyses – this is very important, but DO NOT DISCUSS THE BROADER CONTEXT OR COME TO CONCLUSIONS YET! That comes later.

   i. Explain how you had to clean and manipulate your data to help you address each of your research questions. Did you delete, infer or ‘correct’ values, and/or add categorized and/or binned variables? If so, why?

   ii. Does your data set have any shortcomings that might make it difficult to address your questions?

   iii. Provide one or more statistical analyses specific to each of your research questions. For each analysis, provide all relevant discussion, tables, figures and/or charts, with clear, meaningful titles, headings and labels as appropriate. Explain your methodology as you go, with reference to the methods we’ve learned in class this term, and describe the results of each analysis. (Don’t overdo it here – the trick is to provide enough, but not too much!

e. Discuss your overall study results – You may need to do some research here.

   i. What do the data and analyses tell you and your audience, in relationship to the problems or opportunity or question you are trying to address?

   ii. Did your analyses answer your questions? If so, how? If not, why not?

   iii. Did you run into issues with your data? Were there missing elements, or problems with data quality (wit reference to cleaning discussed above, or discovered later)?

   iv. Did you discover new questions that could or should be investigated, with this or other data?

   v. Can you find any other research, studies or analyses that supports or conflicts with your findings? Describe your search, and provide references.

f. Tie it all together with a short conclusion - What answers did you find for your analysis questions? What are the consequences in terms of the original research question, problem or opportunity? This should be a clear statement that leaves the audience in no doubt about the outcome of the analysis.

g. In appendices, provide your analysis tracking and references – the reader should understand your thoughts about what you studied; what data you found and where; what descriptive statistics you developed and why; how you cleaned and categorized the data; what questions you asked; what assumptions you made, and why; how you chose your techniques and why; and any other materials or research you used (and cited). **Using your tracking, another competent data analyst should be able to replicate your analysis.**

**Your MS-Word report should stand on its own as a complete and thorough final report of your complete analysis. You will submit your Excel spreadsheet as well, so your instructor can examine how you got your results and conclusions.**

## Evaluation

*Part 4 is marked out of 120, and worth 12% of the course.*

- **Appropriate Data Set (6 marks)** – Does your dataset follow the prescriptions above?
- **Univariate Descriptive Statistics (6 marks)** – Do your univariate descriptive statistics thoroughly and appropriately describe at least eight variables and a mix of qualitative and quantitative data types? Did you present them well (including charts and tables as appropriate)?
- **Cleaning & Coding (6 marks)** – Did you identify suggested outliers or other unusable information in your dataset? Did you clearly explain appropriate decisions about whether and how to ‘clean’ them in your modified data? Did you create and describe codes and/or categories for at least one variable that could help with your analyses?
- **Hypothesis Testing (12 marks)** – Did you develop at least five hypotheses according to the selection requirements, explain their relevance to your research question (from part 2), perform the tests (with whatever data rearrangement or manipulation is required for each), and describe the results, including appropriate visualizations (tables, charts, graphs, etc.)?
- **Inferential Techniques (45 marks)** – Did you develop and apply at least two inferential techniques, at least one of them predictive? Did you describe the results, including appropriate visualizations (tables, charts, graphs, etc.)?
- **Written Discussion (45 marks)** – Did you address all of the (three or more) research questions (hypotheses, business problems) that you raised in part 1 about your dataset? Did you accurately describe what you found in the data? Does your written description include an introduction and correctly describe the data, questions, analyses and results? Are your conclusions from all of that supported? Could another competent data analyst replicate your work?

## Submission

In the ‘Assignments & Tests’ folder in our Blackboard shell for this course there are two places for you to submit materials for each of the four parts of your Dataset Exploration project.

1) Your submission should include three files:
   a. a complete and thorough written discussion of your analysis, in MS-Word format, including charts, graphs and/or tables as appropriate to illustrate all points;
   b. your analysis spreadsheet, in MS-Excel format, including your original raw data in one tab, and all of your arranged and manipulated data and analysis, in other tabs in the same Excel file; and
   c. a Turnitin Report for your discussion file, including Turnitin’s marked-up version of your work, the Originality Report, and the Final Review
2) A day or two before the due date, you should submit your Word file in the **‘DE Turnitin’** assignment. This will not be marked, and does not appear in the Grade Centre – it’s there only to allow you to get Turnitin reports to help with your spelling, grammar and syntax and to see if you’ve cited and referenced everything correctly and whether you’ve used too much borrowed material. Do not submit CSVs or spreadsheets in Turnitin – it doesn’t work out well!

You must submit all of your work in the **‘Dataset Exploration Part 4’** assignment by the given due date and time for your section, and it must be accompanied by your Turnitin report. **If your work is not submitted on time, there is a penalty of 10% of the entire assignment per week late. If your work is submitted without a Turnitin report, there is a penalty of 5% on the entire assignment and your professor may submit your work in Turnitin regardless.**

***NOTE:*** *If your Turnitin report is not ready before the due date and time, you should submit the Turnitin receipt with your other files, then submit ONLY the Turnitin report separately as soon as it is available. If you submit your Word and Excel files more than one time, only the last submission will be graded, and late penalties will be applied as appropriate.*