Data Science Fall 2026 - CSCI 4350/6350, ITWS 4350/6350, ERTH 4350/6350

Assignment 3: Data Analysis Project (written – 30%) and (presentation - 10%)

Due: Slides are due December 2nd 2026 by 11:00 am ET by email (eleisa2@rpi.ed)

Written report is due December 11th 2026 by 11:00 pm ET on LMS

Please use the following file naming for electronic submission for documents: DataScience_A3_YOURGROUPNUMBER_report.xxx

Late submission policy:

The written report is due 12/11/2026 by 11:00 pm ET. If your group submission is more than 2 days late, it is likely that you will not have your grade for this assignment included in your final grade before they need to be submitted.

Note: Your report for this assignment should be the result of group work. All members of the group will receive the same grade. Take care to avoid plagiarism ("copying"), include all web resources, text, and class presentations. You are expected to work within a group setting contributing equally and with complementary skills.

Assignment instructions: In this group project you will be working with published data. The group chooses a data science investigation, then finds, acquires, analyzes a relevant dataset, and manages the resulting products. The topic can be from any domain discussed or referenced in previous lectures including scientific research, economics, business, finance, healthcare, sports, etc. The dataset should be professionally produced and published. This means government (data.gov, data.nasa.gov, noaa.gov/data, etc.) and academic (e.g. UC Irvine ML) data portals, international organizations (e.g. data.un.org), analytics competitions (e.g. kaggle.com), etc. Ask me if you are unsure whether a specific dataset is suitable.

The weighting score for each question is included below. Please use the question numbering for your written report.

Written submission: Font size 12, New Times Roman, Double Spaced. Include each group members full name and email address at the top of the written document.

1. Abstract, Introduction, Literature Review and Project Workflow (5%):

a) Write an Abstract 300 – 500 words. **(1 %)**

b) Introduction: write an introduction describing the project description, background and project goals. **Minimum 1.5 page (1.5%)**

c) Literature Review: Include 2-3 research papers (or technical articles or blog posts by professionals) that are relevant to your project work. Summarize the papers and discuss how they are relevant. **Minimum 1 page (1.5%)**

d) Project Workflow Diagram: Create a workflow diagram that shows the steps followed throughout the project, from data acquisition to conclusions. Explain each step of the workflow. **Minimum 4-5 sentences + diagram (1%)**

2. Data Description and Analysis Methodology (10%):

Choose an investigation and identify a pre-existing source of data that can address the data science project goal. Describe the dataset, data format(s), and any special characteristics of the data.

a) Concisely state the goal of the investigation and the reasons why the dataset was chosen, how it was found and describe the data source (website/data portal) as well as how the dataset was managed, including its physical/logical organization (even if it's a couple of files). **Min 8 sentences. (2.5%)**

b) Document the data formats and any metadata standards/conventions in use, and how they helped or hindered the process. If there are no or very little metadata with the dataset, consider how you may record new metadata. **Min 5 sentences. (2.5%)**

c) Develop and state two particular questions/hypotheses related to the goal of the investigation that can be answered using the dataset(s) under consideration. Design an analysis study (exploratory analysis and predictive modeling) to answer the project questions and document the analysis design. **Min 3-4 sentences per question/hypothesis (5%)**

3. Exploratory Data Analysis (5%)

a) Conduct Exploratory Data Analysis to examine the distribution of interesting variables, associations between pairs of variables using suitable figures (histograms, boxplots, scatter plots, etc.). Compute summary statistics for the variables that you plan to use in predictive models. Apply any data cleaning/transformations (scaling, removing outliers, etc.) as needed to obtain informative plots/stats. Describe the patterns you found and any early insights. **Minimum 0.5 page + stats/plots (5%)**

4. Predictive Model Development and Application (5%):

a) Implement a minimum of two predictive models for each question/hypothesis you mentioned in Q2. Experiment with different algorithms and different sets of features. Describe the types of models you used to analyze the data (regression, classification, clustering, etc.), describe the variables in the models you decided to use, other parameter

choices or settings for the models (e.g. distance metrics in clustering algorithms, or a type of kernel, etc.). **(5%)**

b) Evaluate the chosen models' predictive performance using a cross-validation strategy and report on model outputs using suitable figures (residuals plots, confusion matrices, etc.) and evaluation measures (R^2, rmse, recall, precision, etc.). Describe the final results, explain how you validated your models, and discuss your confidence in the results. **(2%)**

Minimum 3 pages text + graphics for Question 4.

5. Conclusions and Discussion (5%):

Describe your conclusions; interpret the results, predictions you made using the models, and the overall project outcome. Give a summary of what changed as you went through the project (data management, model choices, etc.), what you would do next, or do differently in a subsequent exploration.

Minimum 1 page text + graphics (optional) (5%)

References – websites, papers, packages, data sources, etc. You should properly cite your references in the body of the document and include the cited references at the end. You may choose any citation format (Chicago, MLA, etc.,), just be consistent. Make sure all the tables/figures/plots have numbers and captions in your written report. The final written document should be a minimum of 10 pages (but can be more).

Include your code scripts! (e.g. codes in a zipped folder with the written part of the assignment and upload the zipped folder to LMS).

6. Oral Presentation (10%). Presentations will take place during class on December 2nd and 9th. Please submit your slides by email using the same naming scheme mentioned above.

a). Title (followed by team number/name + group members names) and project name.

b). Problem area – what you wanted to explore/solve/predict and why.

c). The data – where it came from, why it was applicable, and any related background info. Include the questions/hypotheses and a sample of data.

d). Briefly discuss your exploratory data analysis:

How you conducted your analysis: what distributions, patterns/relationships you found, any interesting observations/insights.

e). Summarize your predictive modeling: What models have you applied? How did you evaluate them and what were the most informative results?

f). Conclusions: What is the outcome regarding the research questions? What decisions (prescriptions) would you suggest based on what you learned? What would you do differently/next?

g) A *thank you* or *questions?* slide at the end.

EOF