|  |
| --- |
| Homework #1: Exploratory Data Analysis  Hide Assignment Information  Instructions  Exploratory analysis and essay  **Pre-work**   1. Visit the following website and explore the range of sizes of this dataset (from 100 to 5 million records): 2. https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/ or 3. (new) https://www.kaggle.com/datasets 4. Select 2 files to download 5. Based on your computer's capabilities (memory, CPU), select 2 files you can handle (recommended one small, one large) 6. Download the files 7. Review the structure and content of the tables, and think about the data sets (structure, size, dependencies, labels, etc) 8. Consider the similarities and differences in the two data sets you have downloaded 9. Think about how to analyze and predict an outcome based on the datasets available 10. Based on the data you have, think which two machine learning algorithms presented so far could be used to analyze the data   **Deliverable**   1. Essay (minimum 500 word document) 2. Write a short essay explaining your selection of algorithms and how they relate to the data and what you are trying to do 3. Exploratory Analysis using R or Python (submit code + errors + analysis as notebook or copy/paste to document) 4. Explore how to analyze and predict an outcome based on the data available. This will be an exploratory exercise,   so feel free to show errors and warnings that raise during the analysis. Test the code with both datasets selected and compare the results.  **Answer questions such as:**   1. Are the columns of your data correlated? 2. Are there labels in your data? Did that impact your choice of algorithm? 3. What are the pros and cons of each algorithm you selected? 4. How your choice of algorithm relates to the datasets (was your choice of algorithm impacted by the datasets you chose)? 5. Which result will you trust if you need to make a business decision? 6. Do you think an analysis could be prone to errors when using too much data, or when using the least amount possible? 7. How does the analysis between data sets compare?   Develop your exploratory analysis of the data and the essay in the following 2 weeks. |
|  |
|  |

Bottom of Form