Data/Information Visualization Project

Data/Information Visualization Project Introduction

As part of the requirement for AIT-664, you will perform an end-to-end-data analysis and visualization project. By completing this project, you will gain experience and develop skills in critical data analysis and visualization tasks. This project is designed to simulate a real-life experience with data in a corporate environment; it is not a "clean" academic process.

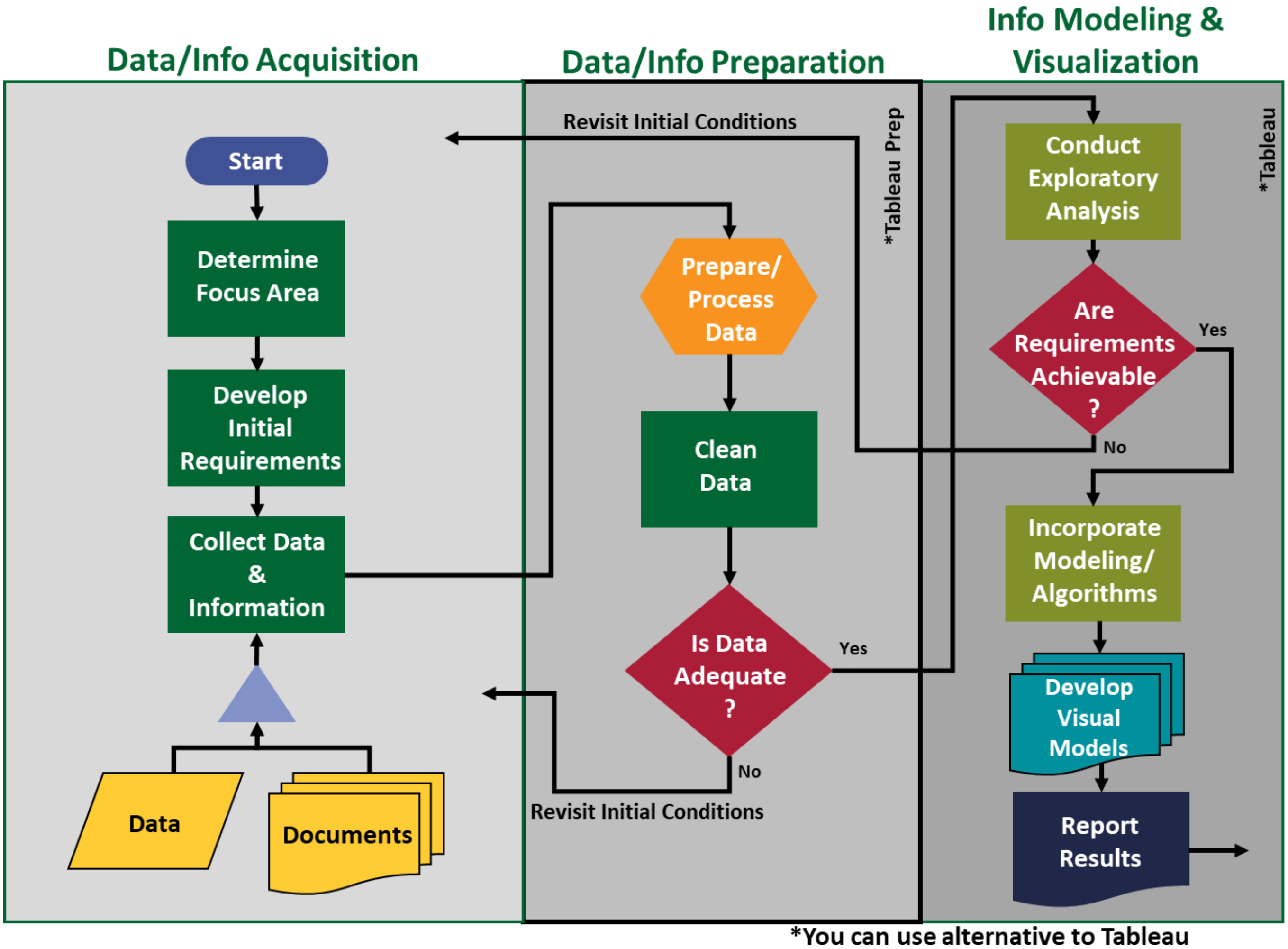
Every student's project will be different since you will find your own data and develop your own hypothesis. You will acquire data, prepare the data and information for processing, perform modeling, and then create visualizations of the information. This process, as shown in the figure below, is iterative. It is possible that you will change the scope of your project at each step.

Submission Requirement

Each project part is to be submitted on time through Blackboard. Late submission will receive a diminished grade based on the number of days late.

Submission can be in standard Office formats such as Word, PowerPoint, and Excel and/or PDF. Reports are to be written using a standard 12pt font using clear font types such as Calibri, Arial, etc. written documents should be single spaced with one-inch margins.

When submitting a report, include any and all documents to support your efforts including report documents, data files, spreadsheets, code, and application files and output, such as Tableau. If a data file is too large for submission, provide a sample.


Data/Info Acquisition
Data/Info Preparation
Info Modeling & Visualization
Start
Determine Focus Area
Develop Initial Requirements
Collect Data & Information
Data
Documents
Revisit Initial Conditions
*Tableau Prep
Prepare/ Process Data
Clean Data
Is Data Adequate ?
Yes
No
Revisit Initial Conditions
*Tableau
Conduct Exploratory Analysis
Are Requirements Achievable ?
Yes
No
Incorporate Modeling/ Algorithms
Develop Visual Models
Report Results

*You can use alternative to Tableau

Part 1, Project Overview Data Acquisition Report (Part 1 of 4)

Due at end of week 2, Sunday at 11:59 p.m. ET

Submit a two-page (minimum) report describing your analysis focus area, your initial set of requirements, your hypothesis, and a description of your data acquisition. Describe your data investigations, sources looked at, initial review of data obtained, and data formats encountered. Provide initial impressions of data validity and quality. Consider the following:

1. Research potential data sources in your area of interest. Many sources are available from the U.S. Government and private organizations. For instance, check out https://www.kaggle.com/ and select “Datasets” from the top.
2. Determine the focus area.

- Determine an area of interest in which to perform analysis and which has a data source available.

3. Develop Initial Requirements.
    - Access the results of the analysis that are to be achieved and develop specific requirements for outcomes that you expect.
        - Identify data/information necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis or customers (who will use the finished product of the analysis).
        - Develop a set of questions that you will attempt to answer with your analysis.
        - Develop specific variables regarding a “population” that you will attempt to obtained
        - Data may be numerical or categorical.
        - Avoid textual data unless you plan to perform some form of Natural Language Processing or word vectorization. This type of analysis is highly technical and not recommended.
4. Develop a hypothesis of what you expect to determine in your analysis.
    - A hypothesis is a supposition or proposed explanation based on limited evidence as a starting point for further investigation.
    - An initial hypothesis helps to guide your investigation and search for data to support the hypothesis and/or the null hypothesis.
5. Collect Datasets and Information based on your chosen area of interest.
    - Collect data and information from a variety of sources, as required.
    - Evaluate your data sources for validity and quality.
        - Consider the following data characteristics:
            - Defined, Measurable, Unitized, Relatable, Normalized, and Quality.

Part 2, Data Preparation & Information Modeling Report (Part 2 of 4)

Due at end of week 4, Sunday at 11:59 p.m. ET

Submit a two-page report (minimum) describing the process used for preparing your data. Describe data preparation related to format, normalization, unitization, quality, and cleanliness. Describe the process used for exploring and modeling your data. Describe data exploration and modeling related to the techniques and algorithms used. Information modeling is essentially playing with and interpreting the data. Use available tools such as Excel or Tableau to try to find correlations and relationships. Use visualization techniques to see which methods reveal interesting information. In many instances, exploratory analysis will show that additional data is needed, or the data needs additional cleaning or normalization. Consider the following:

1. Prepare/Process Data
    - Data initially obtained must be processed or organized for analysis. Data may require some initial analysis or structuring and relating various data elements.
        - Data normalization, e.g., structure data by date ranges, bring monetary figures into current or future values.
        - Map Reduce, Structuring in Tables/Spreadsheets, Natural Language Processing.
2. Clean Data
    - Once processed and organized, data and information may be incomplete, contain duplicates, or contain errors. These errors should be corrected if possible.
    - Types and methods of data cleaning will depend on the type of data, such as phone numbers, email addresses, employers, etc.

- Quantitative data methods for outlier detection can be used to get rid of potentially incorrectly entered data.
- Textual data spell checkers can be used to lessen the amount of mistyped words.

3. Conduct Exploratory Analysis
    - Apply a variety of techniques referred to as exploratory data analysis to begin understanding the relationships, correlations, and messages contained in the data.
    - The process of exploration may result in additional data cleaning or additional needs for data, so these activities may be iterative.
    - Algorithms or calculations may be employed, such as the average or median, to help understand the data. Data visualization may also be used to examine the data in graphical format, to obtain additional insight into the data and information.
4. Incorporate Modeling & Algorithms
    - Post Exploratory Analysis will aid in defining specific algorithms to be implemented in the analysis.
    - Mathematical formulas or models (algorithms) can be applied to the data to identify relationships among the variables, such as correlation or causation.
        - e.g., Clustering algorithm on numerical or textual feature data, Topic Modeling algorithm on textual data
    - Data and information may require additional "normalization." For instance, interest and equivalence formulas may be needed to normalize data to a reference year.
    - In general terms, models may be developed to evaluate a particular variable in the data based on other variable(s) in the data, with the possibility of some residual error.

NOTE:

- If valuable data is not available for your specific area of interest, you may, at this point, change your topic area and hypothesis. Be sure to describe why you changed your topic and the changes you are making in your report.

- If you changed your topic, develop new requirements and hypothesis and report them as part of Report 2.

Part 3, Information Visualization Report (Part 3 of 4)

Due week 6, Sunday at 11:59 p.m. ET

Submit a report (No minimum or maximum length) showing your visualizations of your data and information. Interpret the visualizations and describe how the results reflect your initial hypothesis. Make sure to establish the foundation for your visualization, such as audience needs and requirements. Also, describe why you selected the visualization techniques you used. Consider the following:

1. Develop Visual Models
    - Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
    - Data visualization tools and technologies are essential to analyze large or complex

data/information sets in support of data-driven decisions.
- Data visualization helps the identification of trends, correlations, and outliers. Many forms of data and information are difficult to interpret without some form of visualization. Multiple visual “views” may aid in identifying data and information relationships.

Part 4, Project Recap & Lessons Learned (Part 4 of 4)

Due week 8, Sunday at 11:59 p.m. ET

Submit a summary of your overall analysis, including your initial requirements and hypothesis, data acquisition, data preparation, information modeling, and information visualization. This final report should encapsulate your full semester's effort. Report on the effectiveness of your analysis and whether you met your initial goal of demonstrating your hypothesis. Explain the challenges you faced, what went right and what went wrong, and what you might have done differently to improve on your analysis if you were to do it again. Consider the following:

1. Report Results
   - Restate your initial hypothesis.
   - Report your analysis results.
   - Describe your overall approach to your analysis effort.
   - Determine how effective your analysis was.
   - Describe if and how you demonstrated your hypothesis or the null hypothesis.
   - Describe any lessons you learned from the analysis project.