AN

Data Analytics Final Project and Presentation: Description

PROJECT OVERVIEW

Regardless of your company's products or industry, it's difficult to know how to best serve your customers if you don't understand them. How is your product really being used? What does your customer base look like in terms of demographics or spending habits? In today's world of abundant data, there are often these types of insights about your user/customer base waiting to be found; but along with these insights comes the problem of sifting through and making sense of the data.

After considering the Analytics Workflow, students should work with a large dataset of their choosing, exploring that data for business and real-world understanding, and then prepare their results, findings, and recommendations for a stakeholder presentation.

This will be a summation of all the skills that students have learned throughout Analytics, with an emphasis on packaging findings for a non-analytic audience. Employers — not to mention family and friends — will appreciate the extra effort and skill involved in communicating results with a focus on clarity and action.

DELIVERABLES AND TIMELINE

- **Dataset Approval -** due by beginning of Lesson 17:
    - Approval of project plan by instructional team.
    - Format: Short meeting with your instructional team by appointment to review:
        - Identification of business/product objectives
        - Goals and criteria for success
        - The "right" dataset

- **Database Creation** (optional)- due by beginning of Lesson 17
    - Create a PostgreSQL database running locally on your computer.
    - Load your dataset(s) as tables into your database.

- **Exploratory Analysis** - due by beginning Lesson 19
    - Preliminary, cleaned spreadsheet of data
    - Format: Excel file with two worksheets:
        - Sheet 1: Master data
        - Sheet 2: Data overview:
            - Statistical analysis

            - Graphs
            - Pivot tables
            - Sorted lists
        - Sheet 3: Copy/paste all queries you used to sort, filter, or modify your data before exporting to Excel.

- **Final Presentation** - given to class during Lesson 20
    - Create a 10-minute presentation that delivers the most important insights to key stakeholders within your company, and which goes over your goals, methods, findings, and recommendations.
    - Format: Google Slides or PDF (Keynote/PPT need to be exported)

**A detailed description of each deliverable can be found below in the "Requirements" section.*

REQUIREMENTS

1. Dataset Approval (*note, this step will not be graded but will act as a checkpoint):

- Get the data. If using your own data, present it to your instructor for approval. Things to consider if using your own data:
    - There needs to be a sufficient amount of data, with two distinct datasets/tables.
        - Check with your instructional team if you are only using one dataset.
    - One dataset/table contains timestamped "events," with several variables associated with each timestamped "event."
    - Second dataset/table should contain information related to one of the variables in the timestamped dataset (e.g., if your first dataset is user actions in your software, the second table should contain information about the user).
    - Some open-source datasets.
    - Make sure your dataset adheres to General Assembly's guidelines.

2. Database Creation (optional)

- Load the data into a database.
    - Think about each column/variable in the dataset and its type (e.g. integer, floating point value, character, etc).
    - Load each dataset into a different table in the database.

- Exclude obviously misreported data.
    - Comb through the data for obviously misreported values.

- Combine datasets into a new table.
    - Determine the common columns in the different datasets that will allow you to join all of your datasets.

  - Each row of data should contain one “event” with the other associated values.

- Export the data.
  - Export all of the data from the new table.
  - Format: Clean data in the database and export to a CSV.

3. Cleaning and Exploratory Analysis:

- Open the data in Excel.
  - Pull your previously exported data from the CSV you created.
  - Put the data into an Excel workbook.

- Deal with missing values.
  - Whenever faced with missing values, you can replace them with new values (“impute” new values) such as the mean or median of that column, or remove the rows containing missing values.
  - Find and evaluate the rows containing missing values.
  - Use Excel to either remove or input missing values.
  - Be prepared to defend your reasoning for choosing one option over the other.
    - IMPORTANT! Keep a record of choices you make when cleaning or transforming data, and make sure to briefly mention this in your presentation.

- Clean the data.
  - Look for any outliers or incorrectly recorded data.
  - With outliers, determine if it is better to remove or include them in your analysis.
  - Correct any data-formatting issues. Do columns need to be joined or split?
  - Apply any other normalization or cleaning techniques needed.

- Explore the data.
  - The goal is to understand the data well enough to begin to see distinct segments within it.

- Create new variables (as needed).
  - Create new variables aimed at distinguishing the distinct segments within your data.
  - Use Excel’s conditional logic and the previously created aggregations to create these new variables.

- Create distinct customer/user segments.
  - Create two to five distinct data segments that partition the majority of your data. This range can vary, but it needs to be at least two, and more than five becomes a bit difficult to track.

    - The data segments should be a combination of several different, newly created variables.

- Summarize each data segment.
    - Just as people have distinct personalities, your data segments should also have distinct characteristics and attributes.
    - Use Excel's statistical and aggregation functions, pivot tables, plots, graphs, and any other methods to summarize and describe your data segments.

- Organize your insights.
    - Organize your insights so that you can quickly navigate them.
    - You might want to put all of your newly created tables and graphs on a separate worksheet in your workbook.
    - Your tables and graphs should have the appropriate titles and labels. They don't need to be absolutely finalized, but they should make sense to someone who isn't familiar with your data. Make sure to include titles, axis labels, column names, and row names where appropriate.

4. Final presentation

- Identify problem statement or goal of the analysis:
    - Think about the original question you were trying to answer.
    - With that in mind, identify the insights most important to answering that question.
        - For example, if my original goal was to study my customers' purchasing behavior to better market to them, I might focus on the attributes of those who purchase often vs. those who don't.
- Describe the datasets you worked with.
- Describe the presence of null values and how you handled them.
- Describe your cleaning methods.
- Provide a brief overview of analysis methods.
    - This can be mentioned through the presentation of your insights.
- Describe your data segments.
- Point out the most important insights from your data segments:
    - Focus on the most important insights and the implications.
    - People's time is valuable; make the most of it.
- Suggest next steps:
    - Include appropriate tables, charts, graphs, and aggregations to support your insights.
    - The presentation should contain an appropriate level of technical depth. You don't want to give a step-by-step recounting of your process, but mentioning high level concepts is important.
    - Make sure to be familiar with all aspects of your analysis, so you can be ready for questions. Though you might not spend a lot of time talking about all of the

intricacies of the data, it's important to have them in mind if someone asks a question.

WHERE TO GET STARTED

- When choosing a dataset, think about what you want to get out of this course, and how you will demonstrate that newfound knowledge to others.
    - Are you more interested in developing a specific technical skill?
    - Are you more interested in expanding your knowledge of a topic or industry via data analysis?
- As a starting point, consider the following prompts regarding publically-available datasets:
    - NYC School Test Scores: How do different districts in New York City schools compare on English Language Arts (ELA) and Math results? Are there areas of the city that stand out as high-performing and low-performing? Which areas are performing well with English Language Learners (ELL) and Students with Disabilities (SWD)?
        - Datasets:
            - English Language Arts data: http://schools.nyc.gov/NR/rdonlyres/06F00EDA-3204-4773-9E36-13D5AE812DBE/0/DistrictELAResults20132016Public.xlsx
            - Math data: http://schools.nyc.gov/NR/rdonlyres/1268C6F3-80FC-4CBF-8F3D-6721E54EE9DA/0/DistrictMathResults20132016Public.xlsx
    - FEC Expenditure Data: What do candidates spend the most money on? Do some candidates get a spending lead early, while others save for the end? How do spending patterns differ?
        - Dataset: http://www.fec.gov/disclosurep/PDownload.do

RESOURCES

- Review all class materials to date.
- Watch examples of short, powerful presentations (e.g., TED talks).
- SQL queries for creating a database and loading in data from a CSV: https://docs.google.com/document/d/1AtsdFT5Et8PdiR_oDwjknhB3j7-RVKXaRwrU8P49ExA/edit?usp=sharing
- List of publically-available data sources to consider: https://docs.google.com/document/d/1Vx3cBgTkiiA8-Rdq6VWS9lLB7p28WcTfqXKlm8lGRn0/edit?usp=sharing

CITATIONS

- https://blog.modeanalytics.com/five-public-dataset/

EVALUATION

- Your presentation will be graded using the requirements section above as a rubric.
- The Excel workbook containing the master data and data overview will be evaluated using the requirements section above as a rubric.