Midterm Project

ME 364: Data-Driven Problem Solving in Mechanical Engineering

Spring 2022

This document provides a short description about the midterm project for the course Data-Driven Problem Solving in Mechanical Engineering for Spring 2022. The project has two parts. The final submission should be two notebooks, one for each part. The deadline to submit the notebooks is Wednesday March 30^{th}. **Make sure you read the Submission section at the end of this document**.

Part I [40 points]

For this part, you will use the latest version of the United States Wind Turbine Database (USWTDB), which you can download from `https://eerscmap.usgs.gov/uswtdb/assets/data/uswtdbCSV.zip`. You can find out more about the data and the dictionary file describing the parameters on this page `https://www.sciencebase.gov/catalog/item/6001e327d34e592d8671fae0` (dictionary file name is "EntityandAttribute_DataDictionary.csv") Use this data set to answer the following questions:

1. How many rows do we have in this data set? Find the number of missing values in the data set. Find and report the number of missing values in each column.
2. In which states and which counties the highest capacity turbines are installed?
3. Based on the data available, where was the first wind turbine project installed in the US? (provide the state and county and the year)
4. How many projects were installed and become operational during 2020? Which state had the most number of projects during 2020? How many? Which had the least number of projects? How many?
5. How many turbines have been installed in California? How many turbines have been installed in New York?
6. How many projects have been installed in California? How many projects have been installed in New York? What are the projects with most number of turbines in each of these states? How many turbines are installed in these projects?
7. Provide one histogram showing the distribution of turbines' rotor diameters and another histogram showing the distribution of turbines' capacities.
8. Visualize the annual installed wind turbine capacity in the US. Use the appropriate type of plot that can effectively show the trend and communicate it well.
9. Create a pivot table showing total turbine capacity for CA, IA, KS, OK, TX during years 2017, 2018, 2019, and 2020.
10. What variables in the data set are correlated with a Pearson correlation coefficient greater than 0.87? Do the correlations between these variables make sense to you?

11. If you want to use one parameter to predict turbine capacity, what would it be? How did you decide about it? If you can use two parameters to predict turbine capacity, what would they be? How about three parameters? Explain how you decide about these parameters.

Part II [60 points]

The dataset for this part of the project and the detailed explanations of the variables in the dataset can be found here: `https://github.com/MasoudMiM/ME_364/tree/main/Manufacturing_Industry_Database`. The data set is the NBER-CES Manufacturing Industry Database, containing the annual data from the United States manufacturing sector for the period from 1958 to 2018. The data set page provides all the information needed to become familiar with the variables and how they were collected or calculated.

Your goal in this part is to use this data set and develop a Python notebook, containing the codes and texts, focused on exploratory data analysis. The main goal should be to perform an exploratory data with all the steps required. Following is a list of tasks that (at least) you should perform and their corresponding weights in the project grade:

- Data Cleaning & Handling Missing Data (10 points):
  Once the data is imported to the notebook, focus on the data cleaning to make sure that there are no missing values. If there are, find the way to handle them. Further, make sure that everything is in the correct format. Removing all the rows with missing values with no investigation is not a proper way of handling missing data.

- Descriptive Statistics (5 points):
  Calculate and report the basic statistics for the variables of your interests. Remember, every statistical calculations and reported values need to be explained in text cells. If you are calculating the statistics for specific variables, explain what are they representing and what do they tell us about the data.

- Data Visualisation (15 points:)
  Use data visualization techniques you learned in the course to investigate the various aspects of the data. Look for interesting relationships and/or plots that can convey your messages or help you investigate the answer to a possible question you might have and think the data can help you to answer. Using graphs and plots, you should be communicating something specific maybe about a trend or some type of relationship between variables. The plots should not be the goal. Use them to make a point or investigate a possible question in mind. All graphs should be properly labeled, properly sized, and sufficiently explained if they are representing a special trend.

- Grouping (10 points):
  Use grouping to look at various aspects of the data. For instance, use grouping as a tool in combination with other tools, if you can, to look at various industries, years, or employment numbers based on other variables in the data set.

- Correlation Analysis (5 points):
  Perform sets of correlation analysis between different variables in the dataset. You might need to combine grouping and correlation analysis to look at a specific aspect of the data. Remember to clearly explain the outcome of your correlation analysis.

- Insights and Interpretations (10 points):
  Every correlation analysis, data visualization, and grouping should come with some explanations and possible insights about what to make of the results and how can it give us some information about the status of manufacturing in the US.

- Overall Quality of the Notebook and Analysis (5 points):
  Make sure the notebook is well organized. It has a title, sections, and subsections. Make sure that text cells and code cells are logically ordered and organized, and the explanations and comments are clear.

Remember, you are doing exploratory data analysis, so explore and analyze the data! Try to ask questions and use the data to answer those questions using the tools and methods you have learned. Show your work step-by-step with comments in your code lines and texts between the code cells in the notebook environment. One possible option to find some interesting information from this data is to look at different trends and explore their possible relations with historical events. Your submission should be a Python notebook with all of the texts, codes, comments, and results.

Submission

Your final submission will be two notebooks, one for part I and another for part II. When creating your notebooks:

- Make sure the notebook is well-organized and it includes the codes and the outputs.
- Make sure the answers to the questions asked are explicitly either printed or displayed in the notebook in the same order as the questions. Take advantage of the "print" options to provide an organized output.
- Make sure your name is in the notebook, preferably as a text in your notebook.
- Make sure you do not print very large chunks of data and variables with so much data. They just make your code unclean and hard to understand. They also mostly serve no purpose.
- Use proper variable names that help anyone reading your code understand what each variable might be representing.
- Comment your code so others can understand your thought process and debug your code in case of an error. Also, it helps you be able to easily use your code in the future.
- Once you are done writing your code, your notebook most probably includes code cells that you used for experimenting with some functions or testing some code lines. Organize your notebook and clean it up. Remove those unneeded sections/cells, restart the kernel, and then run the notebook from the beginning. Then submit the notebook with the results.
- Keep the length of code lines in your code short. Break the long lines into multiple lines. The suggested length of a code line is 79 characters.