San José State University
Department of Applied Data Science

DATA 220
Mathematical Methods for Data Analysis

Spring 2021
Instructor: Ron Mak

Assignment #3

Assigned: Thursday, February 11
Due: Thursday, February 18 at 5:30 pm
Team assignment, 100 points max

Analysis of a dataset

The purpose of this assignment is to practice using Python's data science modules to perform data analysis and data visualization.

Find an interesting dataset on the internet. Use the analysis of the Titanic Survival dataset shown in class as an example and analyze your dataset using descriptive statistics.

Your analysis should include:

- Data wrangling
  - filtering out bad data
  - filling in missing values
  - reformatting the data
  - loading data into Python
  - *etc*., as necessary
- Measures of central tendency
- Measure of variability
- Suitable charts

Common ways to fill in missing values are to replace each one with a default value or with the mean or median of all the good values.

Do not calculate a measure simply because you can. In your notebook, explain why you did the calculation and what insight you gained from the result.

Markdowns

Jupyter notebooks are an important form of communication among data scientists to share data and analyses. Therefore, be sure to include useful markdowns and comments in your notebook.

Your markdowns should explain:

- What is your dataset and why do you think it’s interesting?
  Include a link to the dataset.
- What data cleanup was required?
- Before you did the analysis, what did you hope to discover?
- What was your overall approach to analysis
  (*e.g*., what data categories did you use)?
- What insights did you discover from your analysis?
  Will it support good decision-making?

What to submit to Canvas

Before submitting your notebook into Canvas, make sure it works!

1. Clear all the output.
2. Restart the kernel.
3. Run all cells.

Submit your notebook to **Assignment #3: Analysis of a Dataset**

Rubric

Your analysis and report will be scored according to these criteria:

Criteria	Max points
Analysis	**75**
• Data wrangling	• 15
• Measures of central tendency	• 15
• Measure of variability	• 15
• Suitable charts	• 15
• Overall quality (useful comments, good choice of measures, etc.)	• 15
Markdowns	
• What is your dataset and why do you think it’s interesting?	**25**
• What data cleanup was required?	• 5
• Before you did the analysis, what did you hope to discover?	• 5
• What was your overall approach to analysis?	• 5
• What insights did you discover from your analysis? Will it support good decision-making?	• 5 • 5