CMSC320 - Introduction to Data Science (Section 0201)

Assignment 1 Instructions

*Due Date: Feb 20, 2024 by 11:59 pm*

*Spring 2024 – Dr. Fardina Alam*

-- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --

\*\* This is the first assignment of the course, consisting of tasks related to GitHub, Pandas, and SQL. I suggest **starting it as soon as possible**, as it may require some time to complete.\*\*

All Files can be found in the following folder: <https://drive.google.com/drive/folders/1cFj-VERiDexRyAoivBvodVWUXnaaNXBS?usp=sharing>

This assignment consists of the following sections.

1. **Github Assignment**
   1. The assignment and its instructions are linked [here](https://github.com/georgew79/CMSC320_GithubAssignment1_Spring2024_DrAlam).

| **Submission Instructions:** You need to submit a **PDF** file to Gradescope with a link to your  GitHub project here. I.e. submit a ***PDF*** file that contains only the link to ***your completed*** repo. |
| --- |

1. **Pandas Assignment**
   1. The assignment is provided as a notebook (.ipynb) file named “CMSC320-Homework01-Part01-Pandas” with this assignment . Complete it using Jupyter Notebook, or Google Colab, and **submit the .ipynb** to gradescope. The notebook ***MUST*** have your code output shown!! ***NO CREDIT*** will be given to **notebooks** that do not show code output.

# 1 additional CSV file named “costs\_ingredients.csv” is attached to the assignment, you must download them for use in your Pandas notebook!

* + - **Update: Feb 07:** 
      * Due to some requests, we have provided an already cleaned version of the costs (“sanitized\_costs.csv”) for your use. You are free to use either the original or the new data. However, regardless of which one you choose, you must explain your reasoning for any adjustments made to the data. As a good practice, please detail how you handle duplicated column names (if using the old data) and ingredients that do not appear in the cost dataset.
      * Also, we made a slight modification to the second question of Section E: Graph Visualization to make it clearer. We are requesting students to download the updated version again.
  1. Remember, for this Pandas assignment, you have to use two dataset ( The "Pizza Sales Dataset" from Kaggle.com and costs\_ingredients.csv). More details can be found in the notebook\_file.

| **Submission Instructions:** You need to submit a **.ipynb** file to Gradescope, renaming the files as follows: 'CMSC320-Homework01-Part1-Pandas\_UID\_LastName'. |
| --- |

1. **SQL Assignment**
   1. The assignment is provided as a notebook (.ipynb) file named “CMSC320-Homework01-Part2-SQL-final” with this assignment. Complete the **.ipynb** file and **submit that to gradescope**. The **notebook *MUST*** have your code output shown!! ***NO CREDIT*** will be given to **notebooks** that do not show code output.
   2. The assignment has additional reading and problems graded through Zybook. Go to Zybook and complete **sections 4.1 – 4.3**. **You must do this to get full credit.**

| **Submission Instructions:** You need to submit a **.ipynb** file to Gradescope, renaming the files as follows: 'CMSC320-Homework01-Part2-SQL\_UID\_LastNam'.  **For Zybook, you don’t have to submit anything in Gradescope. Just complete the zybook problems.** |
| --- |

# Jupyter Notebook:

If you do not know how to handle Jupyter notebook files, please refer to this [link](https://www.dataquest.io/blog/jupyter-notebook-tutorial/)