



Semester Project Part 1

Semester Project

► Semester Project Deadlines:

1. Part 1: 02/20 at 9:59 am (10% of final grade).
2. Part 2: 03/20 at 9:59 am (10% of final grade).
3. Part 3: 05/02 at 11:59 pm (10% of final grade).
4. Report: 05/05 11:59 pm (25% of final grade).

► The project is to be done in groups of 4 students.

► The groups have been published on Blackboard.

Semester Project Part 1: Video Presentation

- Deliverable: 6 minutes of video presentation.
- Create a PowerPoint presentation (maximum 7 slides).
- Use MS Teams to create the video of the PowerPoint presentation.
- Each member of the group should introduce him/her-self.
- Each member should speak equally and should be visible when speaking.
- Make sure that your video is saved on SharePoint and is not expiring.
- Make sure that the instructor and both TAs have access to it.
- Submit the link to the video on Blackboard.
Penalty for submitting the video itself on Blackboard: 50%.
- If you submit a video shorter than 5'45'': 25% penalty
If you submit a video longer than 6'15'': 50% penalty.
- Penalty for late assignments: 33.34% per day.

Semester Project Part 1

- Use the UCI repository to find a dataset:
<https://archive.ics.uci.edu/ml/datasets.php>
- Choose your own dataset:
 - Classification or Regression, NO TIME SERIES
 - Number of instances: greater than 999
 - Number of Variables: greater than 9
 - Area: CS/Engineering
- Submit the name of your dataset to skpathan@cougarnet.uh.edu by 02/02 at 11:59 pm — the subject of the email: Group_X Dataset.
- If your dataset has already been selected by another group, you will be requested to find another one. More than 40 datasets from the UCI repository fulfill all the requirements.

What should be in your presentation?

- Answer the following questions:
 - What is the dataset you have selected?
 - What is the problem to be solved?
 - Why did you find this dataset interesting?
 - Is it structured or unstructured? Why?
 - What are the variables? Is it a multimodal dataset?
- Tasks to be performed **and** to be presented:
 - Save the data in a CSV file:
Group_X_data.csv with X the number of your group.
 - Clean the dataset using Pandas
Save the cleaned dataset in a second CSV file:
Group_X_data_cleaned.csv
 - Save your dataset in a NPZ files: Group_X_data_cleaned.NPZ
 - Save your dataframe in a pickle file: Group_X_data_cleaned.pkl
 - Save your data in a SQL database and execute a simple query using both MySQL Workbench and pymysql.
 - Save your work in Group_X_data.ipynb and in Group_X_data.sql
(including code for cleaning, to save the data, to create the database, and code to run the query)

To submit

- The link to your video presentation on SharePoint.
- Group_X_data.csv
- Group_X_data_cleaned.csv
- Group_X_data_cleaned.NPZ
- Group_X_data_cleaned.pkl
- Group_X_data.sql
- Group_X_data.ipynb