Project Guidelines

The spirit behind the project is to get practice conducting a data science project end-to-end. This includes thinking about what question(s) to answer, gathering the right dataset, data quality assessment, EDA, drawing inferences from data, and building models. In addition, we want you to get some practice in story telling and communicating your data science work clearly and crisply. Project is open ended by design, you are expected to improvise and come up with ideas on what to do and how to do. We have designed four check-points during the semester to help guide your exploration:

Checkpoint	**Deadline**	**Deliverable**
Finalize dataset / problem	17th October	Submit two preferences
Finalize Project Scope	30th October	5-7 min presentation to your TA
Mid Progress	15th November	Blog Post
Final Presentation	1st week of December	Blog Post + 10-15 min presentation to instructors

Checkpoint 1: **Dataset Finalization**

Deadline: 17th October

Due: Two preferences of dataset

You are required to finalize two preferences of your dataset (or idea). You can find a list of datasets in the following spreadsheet:

https://pern-my.sharepoint.com/:x:/g/personal/20100212_lums_edu_pk/EYe2JreoF7dMlKCFLYtQVRgBC-7OmsxflbCYneE_3ud6Qg?e=ET1fo3

You are encouraged to find datasets of your own preference. You should make sure that your suggested dataset is wide enough to answer enough questions.

After you have submitted your preferences, a dataset will be finalized by course staff for you and 1 TA will be assigned to your group. You will be reporting your progress to assigned TA at the end of each checkpoint.

Enter your preferences in the following list by 17th October Midnight:

https://pern-my.sharepoint.com/:x:/g/personal/20100212_lums_edu_pk/ES76IpttI8lJq1xxO6xLXokB5LTvzJw7B3sEuk_qBw90SQ?e=mVdc51

Checkpoint 2: **Project Scope**

Deadline: 30th October

Due: 2-3 slides to be presented to assigned TA

In this phase, you will be expected to come up with questions that can be answered from your chosen dataset. You will prepare 2-3 slides and present them to your assigned TA. Slides must briefly describe dataset and present a set of questions that can be answered and inferences that can be drawn from the dataset.

Checkpoint 3: **EDA/ Mid report**

Deadline: 15th November

Due: A comprehensive blog post

In this phase, you will be preparing your dataset using data cleaning and EDA techniques. You will prepare a write up where you will have to explain what steps you took through out the process and their reason. For reference you can see some sample articles attached at the end of this document. Purpose of this exercise is to get you to learn art of storytelling with the data.

Checkpoint 4: **Statistical Inferences/Machine Learning/Final report and Presentation**

Deadline: 1st week of December

Due: A comprehensive blog post, 10-15 minutes presentation

In this phase, you will use the dataset you had prepared in the previous phase to draw statistical inferences or train a machine learning model to answer questions you had proposed in checkpoint 2. You will again prepare a comprehensive report following the outline of a blog post. You will also have to prepare a 10-15 minutes long presentation to present your key results and method used to get them.

Sample Blog-spots:

1) https://medium.com/analytics-vidhya/modeling-chicago-crime-data-set-a90f8eafecb2
2) https://towardsdatascience.com/i-was-looking-for-a-house-so-i-built-a-web-scraper-in-python-part-ii-eda-1effe7274c84
3) https://medium.com/mlreview/spotify-analyzing-and-predicting-songs-58827a0fa42b