# Introduction to Data Science — Final Project Phase 1

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh**

TAs: محمد امانلو، محمد امین یوسفی، محمدرضا محمدهاشمی، حمید سالمی، متین بذرافشان، امیرمهدی فرزانه

Deadline 1: بیست‌و دو اسفند 1403

Deadline 2: پنجم فروردین 1404

## Introduction

In this project, we aim to have a complete data science project and implement all the principles we have learned theoretically in practice! So we will proceed step by step. In this phase, you should choose the data you want to analyze and investigate. This section constitutes 20% of your project. You must submit the specifications of the data you want to work on in the system by the deadline and wait for the approval of teaching assistants. For this, you need to enter information in four sections.

- In the first section, your data information should be entered into this sheet to ensure that other groups are also informed. Please note that teaching assistants will check this sheet regularly. You are required to provide the name of your dataset (please indicate if you have crawled the dataset yourself), the URL of the website from which you collected the data, a link to a sample of the extracted data, and the target in the sheet.
- In the second section, after the teaching assistants have approved your dataset for the project, a mentor will be assigned to your team as the primary project guide. This mentor will accompany you through various stages of the project and provide guidance to ensure that your team executes the project correctly. Please feel free to ask any questions you may have at different stages of the project to your teaching assistant.
- In the third section, you are required to fully collect the dataset you have selected and prepare your final dataset. This preparation is essential for ensuring that you can effectively proceed with the subsequent stages of the project.
- Ultimately, you are required to create a data storytelling presentation using Power BI that is intuitive and visually accessible. The various visualizations and the insights derived from these visualizations will significantly impact your final score for this phase of the project. It is important to note that if your project involves complex data types, such as audio or video data, the feasibility of using this tool may be somewhat reduced (although it will never be completely eliminated). This variability is inherent to the nature of your dataset, and there is no issue with scaling down this aspect of the

project; you will not lose any points as a result. In fact, the extent to which this part of the project is executed may vary depending on your dataset. For more specific guidance on how to effectively tell a story with your particular dataset, please consult your mentor.

## Important Notes

1. In this project, the data of each group must be unique. Therefore, if two groups submit similar data, only the group that registers the specifications earlier can use the data. The criterion for registration is the website
2. Checking and confirming the data by teaching assistants may take time and may take several days, so please be patient. If your work isn't checked for more than 3 days, you may notify teaching assistants via Telegram or Email.
3. If your data is not approved by the teaching assistants, you'll have to resubmit the specifications and update the sheet as well. Therefore, don't delay registering the specifications until the last days.
4. If you perform the data collection or crawling process yourself and do not use an existing dataset, you will receive a scoring advantage! **(10% bonus score on the first phase mark)**
5. To find datasets, you can use websites such as Kaggle and Hugging Face, or other similar websites.

## Required Data Specifications

1. The dataset you select must possess the necessary complexity. This complexity may include the use of image data, audio data, timeseries data, textual data, the application of large language model agents, video data, various signals, or any data that you consider to be complex. The difficulty level of your dataset will be assessed by the team of teaching assistants. Collected datasets are categorized into three levels:
    a. Rejected: This level indicates that the dataset is too simplistic and not suitable for use in the final project of the course.
    b. Acceptable: This level signifies that your dataset is deemed acceptable for the course project.
    c. Advanced: This level indicates that your dataset is considered advanced and will receive a scoring advantage(**10% bonus score on the overall project mark**)
    In general, employing advanced datasets, complex pipelines, sophisticated pre-processing techniques, and higher-level processing methods can earn you up to a 10% bonus score. Moreover, this bonus may be applied again in later phases if your work continues to demonstrate complexity. Please ensure that

you take the requirements of future phases into account when selecting your dataset, so that it supports all upcoming tasks. The complexity criteria will be determined by the TAs, so kindly upload your dataset to the shared Sheet as soon as possible. If your current dataset does not meet the necessary complexity standards, you will have the opportunity to select a new one.

2. In the final phases of the project, you must implement a data-driven task that leverages the relationships within your dataset. While a common approach might involve predicting one data column using the others—by selecting a target variable and ensuring that a logical and correlated relationship exists—this requirement is intentionally flexible. You are free to choose other tasks, such as developing a recommendation system, generating text, performing image segmentation, or any other application that fits your dataset.
    - Regardless of your chosen approach, your methodology should capitalize on the inherent structure and interconnections within your data. Ensure that your selected task is well-supported by your dataset's characteristics and that your approach is logically justified by the relationships present in the data.

## Task

1. You need to submit the following information in the sheet specified before:
    - Dataset name(if data collection is done by yourself, mention this in the name)
    - Dataset link(if data collection is done by yourself, provide the link of the site you intend to crawl)
    - Target Task
    - Data sample(sending just five samples is sufficient)

    The deadline for completing this phase is 23:59 on Esfand 22th. After this deadline, a penalty of 10% will be deducted from your score for each day of delay in uploading the information to the sheet. If you submit your dataset earlier than other groups, you will receive approval sooner and can proceed with the subsequent stages. Additionally, this will reduce the likelihood of another group selecting a dataset similar to yours. Therefore, please make sure to choose your dataset and enter it into the sheet as soon as possible. The teaching assistants will continuously respond to the datasets submitted in the sheet.
2. After your team receives final approval from the teaching assistants, a teaching assistant will be assigned as a mentor for your team. Following this, you should work closely with this mentor to carefully gather the dataset specified in the sheet. Once you have collected the data, you will proceed to create visualizations and derive insights from the dataset using Power BI. Focus on formulating a problem or question based on the data you have gathered and create a Power BI report that addresses this

practical issue or real-world problem. Aim to present your findings in a clear and impactful manner through your visualizations.

## Overview of the Next Phases of the Project

The project consists of three main phases.

Phase 1: Topic Selection, Data Collection, and Initial Visualization

In this phase, you will focus on selecting a topic, gathering data, and creating initial visualizations of your dataset.

Phase 2: Data Storage and Processing Pipeline

In the second phase, you are required to store your data in an appropriate environment and establish a proper pipeline, which will be covered in the subsequent lessons. This will enable you to read and process the data safely, allowing you to extract the necessary results. Additionally, during this phase, you should perform some preprocessing tasks and address questions related to the nature of your dataset.

Phase 3: Final Conclusions and Model Development

In the third phase, you will arrive at final conclusions and develop a model to do your task.

These phases are designed to guide you systematically through the project, ensuring that you build a solid analytical foundation and derive valuable insights from your data.

## Notes

- Upload your work in this format on the website: DS_Project_P1_[Std numbers].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- This phase will not be accepted after its deadline i.e. there will be no late and grace policy!

**Good luck!**