# Project & Report Assignment

## How-To Guide

This assignment represents 100% of the overall course grade.

## Instructions

Develop a Python project to analyse real world scenarios and generate valuable insights by visualising information. The project aims to analyse data from different data sources, manipulate information and visualise to generate insights.

You can use any open-source dataset available online for analytics. Each bullet point for every learning outcome is a milestone to be achieved.

The project should be submitted on the Learn Site under the Assessments section. You will need to include two files, as described below.

There are three deliverables contained in two files:

1. Project ZIP
    - Create a ZIP file of your entire Python project along with all the code and data files and upload as part of your submission
    - The project should cover all milestones in each learning outcome to gain full marks (see below)
2. Project Report
    - A document containing between 1,500 and 2,000 words
    - Please use the template provided (see Assessments section to download)
    - The report describes your process, dataset, different sources, graphs and insights
    - Justify the use of each learning outcome concept, for example: Why did you use list over dictionary?
    - Upload the document file along with the ZIP file
3. GitHub repository URL
    - Create a new repository on GitHub as [UCDPA_yourname]
    - Keep committing to the repository
    - Remember to include the URL of your repository at the beginning of your Project Report document

The goal of the assignment is to demonstrate how you are thinking about putting course concepts and learning into practice to demonstrate the course learning outcomes:

1. Gain insight into scoping in Python and be able to write functions with multiple parameters and multiple return values, along with default arguments and variable-length arguments.
2. Have a clear understanding of iterators, objects, list comprehensions and generators.
3. Identify, diagnose, and treat a variety of data cleaning problems in Python, ranging

from simple to advanced and deal with improper data types, validate that data is in the correct range, handle missing data and perform record linkage.
4. Understand string manipulation using regular expression and work with datasets containing movie reviews or streamed tweets that can be used to determine opinion, as well as with raw text scraped from the web.
5. Understand the two principles of statistical inference, parameter estimation and hypothesis testing and work on real world datasets to solve real inference problems.
6. Use tools a data scientist needs to clean and validate data, to visualize distributions and relationships between variables, and to use regression models to predict and visualize.
7. Understand and build supervised predictive models, tune their parameters, and determine how well they will perform with unseen data on real-world datasets.
8. Work on unlabeled datasets using unsupervised clustering algorithms to transform, visualize and extract insights and build a recommender system on a real-world usecase.
9. Use deep learning to optimize natural language processing, image and speech recognition, robotics and many more.
10. Understand the advantages and shortcomings of trees and demonstrate how ensembling can alleviate these shortcomings.

## How You Will Be Assessed

The following list describes the areas being assessed, for a total of 150 points (points awarded are indicated in brackets).

1. Real-world scenario
    - The project should use a real-world dataset and include a reference of their source in the report (5)
2. Importing data
    - Your project should make use of one or more of the following: Relational database, API or web scraping (10)
    - Import a CSV file into a Pandas DataFrame (10)
3. Analysing data
    - Your project should use Regex to extract a pattern in data (10)
    - Replace missing values or drop duplicates (10)
    - Make use of iterators (5)
    - Merge DataFrames (5)
4. Python
    - Define a custom function to create reusable code (5)
    - NumPy (5)
    - Dictionary or Lists (5)
5. Machine Learning (30)
    - Predict a target variable with **Supervised** or **Unsupervised** algorithm
    - You are free to choose any algorithm
    - Perform **hyper parameter tuning** or **boosting**, whichever is relevant to your model. If it is not relevant, justify that in your report and Python comments
6. Visualise
    - Present two charts with Seaborn or Matplotlib (10)
7. Generate valuable insights

    - 5 insights from the project (10)
8. Report (30)

The final grade is indicated by a scale as follows:

| No attempt | Clear fail | Fail | Pass | Merit | Distinction |
|---|---|---|---|---|---|
| 0 to 15 | 16 to 38 | 39 to 74 | 75 to 96 | 97 to 119 | 120 to 150 |