TU Dublin Tallaght Campus Department of Computing	Data Analysis and Programming	
	Contribution: 25%	
Group Project	Lecturer	Rajesh Jaiswal
	Project Submission Date: 8th March 2023, 6 pm	

For this project, you are required to submit a zip file containing

1) One executable notebook (.ipynb or .py) with file name “last_nameCA1.ipynb”
2) One csv file that contains the cleaned data set with file name “last_nameCA1.csv”
3) One html file generated from the notebook
4) ReadMe.txt if required (Optional)

Important Instructions on the notebook

- Your notebook should contain relevant codes/instructions to install and run any libraries required for your source code.
- **All tasks must be completed using python**. Separate cells and text blocks must be created for any relevant output and plots demonstrating the key points of the project.
- You must include relevant comments for codes and texts for discussions.
- Cells with errors must be commented on appropriately

Task 1 – Descriptive Statistics (univariate plots) 35 Marks

- Download and import data **“project_sales.csv”** from Moodle into your notebook **(1)**
- Give a summary of the data (description, number of variables, size, types of variables) **(6)**
- Perform cleaning of data using good data management techniques and create a new cleaned dataset **(10)**
  - Check and fix the data for inconsistency or non-entry.
  - Assign appropriate attribute names
  - Check if you need to create new attribute(s) or remove redundant attribute(s)
- Create appropriate univariate plots for each attribute from the cleaned dataset. Use frequency tables for categorical data. **(10)**
- Select appropriate summary statistics to describe the center, spread for each numerical variable, and provide justification of the same. **(8)**

Task 2 – Descriptive Statistics (bivariate and multivariate plots) 30 marks

- Plot a correlation matrix for all the numerical variables and interpret the results **(5)**
- Based on univariate analysis and correlation matrix, generate (at least 10) and discuss appropriate bivariate and multivariate plots/tables/proportions. **(20)**
- Pick any two feature pairs that are highly correlated and test their distributions using normal Quantile and QQ plots to determine if the distribution is close to normal. Also, measure the kurtosis and skew of the features chosen. **(5)**

Task 3 – Statistical Inference 35 marks

- State and carry out **two** 1-sample hypothesis test for a numerical variable (use mean or median)

- State and carry out **one** 2-sample hypothesis test for numerical variables (use mean or median)
- State and carry out **one** 2-sample hypothesis test that involves categorical variable(s) (use proportions)

(4,4,9,10)

- Provide conclusive remarks for all the tests performed **(8)**