# CW01 Introduction to Data Science Assessment ‘Exploratory Data Analysis’

Write a report, using Google Collaboratory, documenting the ‘Exploratory Data Analysis’ process for the **‘student\_data\_2. csv’** dataset on GCU learn.

**The aim of this EDA is to examine factors which may affect student dropout.**

The data set has been cut down from the publicly available version. However, it still contains extraneous information for this aim. You should select 3-4 potentially interesting columns focusing on the aim. If the analysis doesn’t support an interesting column(s) affect on student dropout that is a useful finding.

It is important that you explain your analysis – see marking scheme.

The analysis techniques (python code) should be appropriately applied. For example, in the titanic data set, mean ‘age’ is appropriate, mean ‘class’ is not appropriate.

The \*.ipynb file should be uploaded using the assessment submission link.

Introduction

* Brief explanation of the **student\_data\_2. csv** data set.
* Overview of the assessment objectives.
* Description of key questions, given the aim, your report aims to answer through EDA.

Data Loading and Initial Inspection

* Load the **student\_data\_2. csv** Data Set.
* Display basic information about the dataset (e.g., number of rows and columns).
* Examine the first few rows of the dataset to get a glimpse of the data.

Data Cleansing and Pre-processing

* Check for and address any errors, removing rows if/where appropriate.
* Examine the numeric columns (not categorical) listing the number of outliers at a threshold of 3 SD
* Explain the initial steps referencing the aim of the report.

Univariate Analysis

* Examine the data set using; basic statistics, measures of central tendency, standard deviations, and value counts
* Visualize the distributions of key numerical variables using a histogram
* Analyse the frequency distribution of a categorical variable using a bar chart
* Explain how the techniques work and your findings referencing the aim of the report.

Bivariate Analysis

* Explore the correlation between two meaningful numerical variables, using a conditional if statement to show the strength of the correlation and another to show whether or not it is statistically significant
* Explore either a cross tabulation or a pivot table between two columns
* Explain how the techniques work and your findings referencing the aim of the report.

Conclusions

* Summarize the main findings and insights from the exploratory data analysis.
* Draw conclusions and insights based on the EDA results.

## Marking Scheme

**Introduction and Preliminary Findings: 10 marks**

|  |  |
| --- | --- |
| A  7-10 | The student's introduction is exceptionally well-crafted, providing a clear and concise overview of the dataset and the objectives of the analysis. The preliminary findings are insightful and effectively set the stage for the subsequent analysis. The introduction demonstrates a deep understanding of the data and sets clear expectations for the rest of the exploratory data analysis. |
| B  6-6.9 | The student's introduction is solid, offering a clear overview of the dataset and the objectives of the analysis. The preliminary findings provide meaningful insights, though some aspects could be further emphasized. The introduction effectively introduces the dataset and its context, setting the tone for the rest of the analysis. |
| C  5-5.9 | The student's introduction is satisfactory, but it lacks some clarity or conciseness. The overview of the dataset and objectives might be adequate, but the preliminary findings may be limited in depth or clarity. The introduction provides a basic understanding of the data but could benefit from further refinement. |
| D  4-4.9 | The student's introduction is weak, with limited or unclear information about the dataset and the objectives of the analysis. The preliminary findings are inadequate, lacking valuable insights or context. The introduction does not effectively set the stage for the exploratory data analysis. |
| F  1-3.9 | The student's introduction and preliminary findings are missing or significantly inadequate, failing to provide any meaningful information or context about the dataset or the objectives of the analysis. |

**Data Cleansing: 10 marks**

|  |  |
| --- | --- |
| A  7-10 | The student demonstrates a thorough understanding of data cleansing techniques and effectively addresses outliers, missing values, and inconsistencies. The data cleansing process is meticulously executed, resulting in a high-quality dataset. Comprehensive explanation showcases a clear record of the steps taken. |
| B  6-6.9 | The student performs a solid data cleansing process, effectively handling most errors, outliers, and missing values. The resulting dataset is of good quality, and the steps taken are reasonably well explained, reflecting a sound understanding of data cleansing principles. |
| C  5-5.9 | The student applies basic data cleansing techniques, but some errors or missing values might remain unattended. The resulting dataset is satisfactory, and the explanation of the process is acceptable, though it could be more comprehensive. |
| D  4-4.9 | The student's data cleansing efforts are rudimentary, with significant errors or missing values still present in the dataset. The resulting dataset's quality is subpar, and there is limited evidence of explanation or understanding of data cleansing concepts. |
| F  1-3.9 | The student's data cleansing attempts are minimal or non-existent, leading to severely compromised data quality. The resulting dataset is unreliable, and there is no evidence of explanation or understanding of data cleansing principles. |

**Univariate Analysis: 30 marks**

|  |  |
| --- | --- |
| A  21-30 | The student demonstrates an excellent grasp of univariate analysis techniques, utilizing excellent visualizations and uncovering meaningful patterns and insights. The explanation of the analysis is highly informative and clearly presented. |
| B  18-20.9 | The student performs a solid univariate analysis, utilizing appropriate visualizations and providing valuable insights into key variables. The explanation of the analysis is well-executed and effectively communicates important findings. |
| C  15-17.9 | The student conducts a satisfactory univariate analysis, but the insights provided are limited or somewhat unclear. The use of visualizations is basic, and the presentation could be improved. The explanation of the analysis highlights some key aspects. |
| D  12-14.9 | The student's univariate analysis is rudimentary, with limited visualizations and weak insights. The explanation of analysis lacks depth and fails to provide a clear understanding of the data's characteristics. |
| F  0-11.9 | The student's univariate analysis is inadequate or non-existent, providing no meaningful insights or visualizations. The analysis does not contribute to the understanding of the dataset. |

**Bivariate Analysis: 30 marks**

|  |  |
| --- | --- |
| A  21-30 | The student demonstrates exceptional skills in bivariate analysis, employing advanced techniques to explore complex relationships between key variables. The explanation of the analysis is insightful and effectively communicates nuanced findings. |
| B  18-20.9 | The student conducts a competent bivariate analysis, using appropriate techniques to examine relationships between variables. The explanation of the insights provided are valuable, but some aspects could be further explored. |
| C  15-17.9 | The student performs a satisfactory bivariate analysis, but the exploration of relationships lacks depth, and some key connections might be missed. The explanation of the analysis could benefit from additional context and clarity. |
| D  12-14.9 | The student's bivariate analysis is rudimentary, with limited exploration of relationships between variables. The explanation of the analysis lacks coherence, and important connections are overlooked. |
| F  0-11.9 | The student's bivariate analysis is inadequate or non-existent, providing no meaningful insights into relationships between variables. The explanation of the analysis does not contribute to the understanding of the dataset. |

**Conclusions: 20 marks**

|  |  |
| --- | --- |
| A  14-20 | The student draws exceptional conclusions, providing valuable insights and actionable recommendations supported by the analysis. The conclusions are well-structured and demonstrate a comprehensive understanding of the dataset. |
| B  12-13 | The student's conclusions are solid, offering clear insights and relevant recommendations based on the analysis. The conclusions are well-presented, but some aspects could be further emphasized. |
| C  10-11 | The student's conclusions are satisfactory but may lack depth or fail to address all relevant findings from the analysis. The recommendations could be more explicit and connected to the insights. |
| D  8-9 | The student's conclusions are weak, providing limited insights or failing to draw meaningful connections from the analysis. The conclusions lack coherence and fail to offer actionable insights. |
| F  1-7 | The student's conclusions are missing or severely inadequate, providing no meaningful insights or recommendations based on the analysis. |