|  |
| --- |
| **King Saud University**  **College of Computer and Information Sciences**  **Information Technology department**  **IT 326: Data Mining**  **Course Project** |
| **Project Title** |
|  |

**Project Report: Data Summarization and Preprocessing**

|  |  |  |
| --- | --- | --- |
| **Group#:** |  | |
| **Section#:** |  | |
| **Group Members** | Name | ID |
|  |  |
|  |  |
|  |  |
|  |  |

|  |
| --- |
| [Pick the date] |

# Problem

Introduce the problem. What do you want to solve? Why do you think it is important?

# Data Mining Task

Formalize the problem as a data mining task. i.e., classiﬁcation, and clustering. In the case of classification specify the class attribute. What is the goal of both data mining tasks?

# Data

Describe the dataset include: Source, Number of objects, Number of attributes, and the main characteristics of attributes (e.g. data type, possible values, missing values, etc..) using statistical measures and graphical presentation you learned.

You should calculate the statistical measures for numeric attributes as five number summary, outliers and boxplots. Also, you should apply *three more different plotting methods* that would help you to understand your dataset, such as scatter, histogram (numeric attributes) and bar plot (nominal attributes) using R tool. For each plot, *provide a clear description* of the plot and what information it provides you about the dataset and how this information helped you to decide that your data needs preprocessing.

# Data preprocessing

This section explains why your dataset needs data preprocessing. You can apply either data cleaning or data transformation or both. For each, you should describe why you applied it and how and on which attributes. You need to provide a snapshot of raw dataset, as well as preprocessed dataset. If your data did not require preprocessing, you should justify why and illustrate that using code or pictures.

**NOTE**: you should submit *at least four preprocessing different tasks* (not including: removing attributes or splitting the data set).

# References

Use IEEE format <https://libraryguides.vu.edu.au/ieeereferencing/gettingstarted>

You can use any resource for reference generating such as <https://www.citethisforme.com/citation-generator/ieee>

# Tasks Distribution

|  |  |  |
| --- | --- | --- |
| ID | Name | Responsibilities |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |