**APPENDIX D – EXPLORATORY DATA ANALYSIS**

***(TO BE COMPLETED INDIVIDUALLY)***

**Module Group: Group 2**

**Project Group Number:**

**Team Name:**

Description of Business

• Problem Statement / Hypothesis

*(similar to Appendix C - describe in detail the targeted problem, the proposed hypotheses and projected outcome of the analysis)*

Details of Datasets

• Scope of datasets

o List down all possible data sources that *have been studied* and explain why a dataset is included for further analysis or excluded

*(List down the names and URLs of* ***ALL*** *the possible data sources that you have* ***found*** *– and for each URL, give reasons why you have decided to include OR exclude them)*

• Common or individual datasets

*(For this section, you will give details of* ***ALL*** *the datasets* ***used*** *in the entire project)*

o Description of the dataset

*(Describe the dataset and the kind of data it contains. For example, this dataset contains NYP Student enrolment data from Year 2010 – 2020)*

o Source of the dataset, e.g., URLs

*(Include the URL for each dataset)*

o Type of the dataset, e.g., XML, CSV?

*(Format of your dataset)*

o Is the dataset shared or common?

*(Is this dataset shared/common? If not, list which team member(s) are using it.)*

o How is the dataset being used? e.g., selected fields or all fields?

*(Will the entire dataset be used? If not, list out the names of the column that you will be using.)*

• Individual contribution *(Indicate Team Member name)*

o List datasets being used

*(as specified in the previous section under Common or individual datasets)*

o List datasets being used

*(as specified in the previous section under Common or individual datasets)*

o Describe the data cleaning and transformation done

*(Explain briefly the steps required to clean your data source. For instance, you can state points like (1) Striped symbols and empty spacings from ALL columns, (2) Wrote Excel*

*formulas to transform abbreviations into actual description, etc.)*

o How is the dataset related to problem / hypothesis?

*(Justify how your dataset is relevant to your team’s hypothesis, for example, this dataset contains environmental data collected by NEA which helps us gather insights on…......etc.)*

Data Catalogue

*(Describe the structures and fields of your datasets/ data sources. Draw ER diagrams to show the relationship between the various datasets if the datasets that you are using is related to other datasets. For example, if dataset A contains column B whose meaning can only be understood by interpreting column Z in the dataset)*

Miscellaneous

*(anything else you wish to add about your project can be included here)*