CRISP-DM stage two – data understanding

Collect initial data

Task

The second stage of the CRISP-DM process requires you to acquire the data (or access to the data) listed in the project resources. This initial collection includes data loading, if necessary for data understanding. For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool. If you acquire multiple data sources integrating these together needs to be considered, either here or in the later data preparation phase.

Output

- **Initial data collection report** – list the dataset(s) / data sources acquired together with their locations, the methods used to acquire them and any problems encountered. Record problems encountered and any resolutions achieved. This will aid with future replication of this project or with the execution of similar future projects.

Describe data

Task

Examine the "gross" or "surface" properties of the acquired data and report on the results.

Output

- **Data description report** – describe the data that has been acquired including the format of the data, the quantity of data (for example, the number of records and fields in each table), the identities of the fields and any other surface features which have been discovered. Evaluate whether the data acquired satisfies the relevant requirements.

Explore data

Task

This task addresses data mining questions using querying, visualization, and reporting techniques. These include distribution of key attributes (for example, the target attribute

of a prediction task) relationships between pairs or small numbers of attributes, results of simple aggregations, properties of significant sub-populations, and simple statistical analyses. These analyses may directly address the data mining goals; they may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed for further analysis.

Output

- **Data exploration report** – describe results of this task, including first findings or initial hypothesis and their impact on the remainder of the project. If appropriate include graphs and plots to indicate data characteristics that suggest further examination of interesting data subsets.

Verify data quality

Task

Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct, or does it contain errors and, if there are errors, how common are they? Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they?

Output

- **Data quality report –** list the results of the data quality verification; if quality problems exist, list possible solutions. Solutions to data quality problems generally depend heavily on both data and business knowledge.