Documentation of the second phase of the data mining project

Faridreza Momtazzandi 9812762601

Alireza Noorbaksh 9812762496

1 .Introduction:

After the knowledge we found with our data in the last phase, in this phase we want to get information and proper evaluation regarding data quality. The main goal is firstly to evaluate the quality of this data and in the next step to find the problems of the data and finally to propose some solutions to improve its quality.

In line with this goal, we will use five characteristics of the inherited qualitative model, which we will learn more about in the following, and of course, we will examine the problems of this dataset in two different categories. Finally, after completing this phase, we can say that we not only know the dataset, but we can also provide a complete expert opinion regarding its quality.

1-2 :Getting to know the ISO 25012 quality model:

The ISO 25012 data quality model categorizes the quality characteristics into fifteen groups, of which we will use five inherited groups in this phase. These five categories are as follows:

•**Accuracy**: The proportion in which the data has features that correctly represent the true value of the desired features about a concept or event in a specific context of use.

•**Completeness**: The proportion in which the data has complete values for certain attributes and there are no deficiencies or defects in the expected cases.

•**Consistency**: the ratio in which the data has characteristics that are free of contradictions and are consistent with other characteristics of the data.

•**Validity**: the ratio in which the characteristics of the investigated data have correct values and can be believed by users, which includes the concept of authenticity.

•**Up-to-dateness**: the ratio in which the characteristics of the examined data have a mandatory standard in terms of age.

This data quality model has other quality features that we will not discuss.

2-2 :Dataset review based on ISO 22012 quality model:

Now that we have gained a correct understanding of this quality model and its five inherited characteristics, we want to apply these characteristics to our datasets.

It is necessary to know that these quality features will not work the same for all datasets, and as we saw in phase one, only two datasets, INOUTLINE and PRODUCTINSTANCE, have features that are capable of detailed checks.

For the other three datasets, unfortunately, we do not have precise features that can determine their quality, and even if we find the total number of records and the number of their nulls, we cannot make useful use of this information. So it is better to skip these three datasets.

For the two datasets INOUTLINE and PRODUCTINSTANCE, first by finding the details of these datasets, we will find the total number of records as well as the number of non-nulls and nulls, and finally, we will check each of the five qualitative models for the selected features.

This information is more complete in the HTML file of the second phase notebook, where we have partially analyzed all these five qualitative characteristics for each of the data sets.

3 :Data quality problems:

In the second part of phase two, we want to find some examples of data quality problems that are classified into single-source and multi-source categories.

1-3 :Single-Source Problems:

The data quality of a source depends largely on the ratio you control and the allowed values. Therefore, data quality problems at your level occur widely due to the lack of model-specific integrity constraints, which is a good example of a limited data model or a weak model.

Level-specific problems are examples of errors and inconsistencies that cannot be avoided at your level, such as misspellings.

2-3 :Multi-Source problems:

The problems with single sources are exacerbated when multiple sources want to be integrated, so that each source may have dirty data, and the data in the sources may be different, overlapping, or even contradictory in the final dataset. The reason for this is the special conditions of data storage and independent development in each source, which has made us able to meet specific needs.

We have fully explored these two categories of problems in the second phase notebook HTML file and complete examples are also provided for each case.

4 suggestions for improving data quality:

Finally, with a complete and detailed review of the quality of these datasets, we will provide three suitable suggestions for improving the quality and we will review each one in detail.