DQA

DATA QUALITY ANALYZER

INTRODUCTION

PRESENTATION & DEVELOPMENT ENVIRONMENT

DOMAIN PROBLEM

We have a large amount of data generated from a data stream of a transport company. In it there are informations regarding each validated ticket, on which bus and on which line.

GOAL

The goal is to analyze the quality of this dataset and make some analysis on it and comparisons on portions of it.

We want to give users the opportunity to analyze these datasets based on their data quality dimensions and to make queries based on dataset attributes. The entire amount of data will be grouped in different datasets per day or week.

Additionaly, since the dataset is almost perfect, it will be "dirtied" in different ways on the different groups in order to analyze more cases of data quality problems.

FEATURES

○○○

1.The user can choose one of the provided dataset he wants to analyze and build a query for it. “Upon selecting a dataset, the application will profile it and plot some informations to the user in order to provide a first understanding of what the dataset is about.”

When provided with the query, the application will return the portion of the dataset which respects the requirements, and the user will be able to download it in CSV format. Some of the data will also be printed in order to provide feedback to the user.

2.In addition, the user will also receive as suggestions other portions of the dataset that perhaps only respect some of the requirements imposed by him and he will be able to decide whether to download even those or not.

FEATURES

○○○

3.We also give the user the opportunity to see further details regarding each portion of the dataset obtained from the query (a sort of profiling) before downloading it.

4.The user can also compare two dataset using the profiling informations or compare two sections of them based on a query of his choice when possible.

QUERY REQUIREMENTS

Completeness > value : the user can search for data that has a completeness greater than the value entered

Consistency > value : the user can search for data that has a consistency greater than the value entered

Timeliness > value : the user can search for data that has a timeliness greater than the value entered

Group by attributes : the user can group the data based on a selected attributes

Attributes: value : the user can search for data that has a certain v alue for the attributes chosen

PROFILING REQUIREMENTS

Upon selecting a dataset, the application will show metadata about each column, including:

- The type (categorical, numerical, datetime etc.)
- The number and percentage of missing values
- The number and percentage of distinct values - A plot of the values (when possible)

And other information about the dataset itself, including:

- Simple metadata of the dataset (number of variables, size in memory, number of tuples, duplicate rows etc.)
- Mean and standard deviation of the completeness, timeliness and consistency
- Matrix plot of the missing values
- A random sample of the data

COMPARE REQUIREMENTS

- The profiling information will be visible for both selected datasets
- The stats of the dataset with higher quality will be highlited in some way (either by color or direct comparison)

DEVELOPMENT ENVIRONMENT

DIRTY DATASET

HOW WE DUMPED THE DATASET

DIRTY DATAFRAME

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

DIRTY DATAFRAME

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

DIRTY DATAFRAME

Parte Uno

○○○

Completeness

Parte Due

○○○

Consistency

Parte Tre

○○○

Conformity

DATA QUALITY DIMENSIONS

CHECK DATA QUALITY DIMENSIONS

DATA QUALITY DIMENSIONS

Parte Uno

○ ○ ○

Completeness

Parte Due

○ ○ ○

Consistency

Parte Tre

○ ○ ○

Conformity

DATA QUALITY DIMENSIONS

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

DATA QUALITY DIMENSIONS

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

PROFILING

...