Data Quality Report

Table of Contents

[Data Sources and Feature Engineering 2](#_Toc19270259)

[Raw Data Sources 2](#_Toc19270260)

[Provide details about the Raw Data Sources shared by the client. 2](#_Toc19270261)

[Processed Data 3](#_Toc19270262)

[Feature Sets 4](#_Toc19270263)

[General summary of the data 4](#_Toc19270264)

[Data Quality Issues 4](#_Toc19270265)

[Data Transformation Steps 5](#_Toc19270266)

[Data quality summary 5](#_Toc19270267)

[Identification of Target & independent variables 8](#_Toc19270268)

[Variable ranking 8](#_Toc19270269)

[Insight learned 8](#_Toc19270270)

[Project Execution 8](#_Toc19270271)

[Data science / Engineering 9](#_Toc19270272)

[Domain 9](#_Toc19270273)

[Product 9](#_Toc19270274)

[Any specific challenges 9](#_Toc19270275)

[Next Steps 9](#_Toc19270276)

The purpose of this report is to log key findings from the "Data exploration and Pre Processing" exercise conducted as a part of CRISP DM process. Ideally a separate report ought to be generated for each individual data file.

The report includes Key parameter related to data Source, metadata, data inconsistencies, Data summary, statistical tests conducted, transformations performed and visualization techniques employed to describe and analyze the data.

# Data Sources and Feature Engineering

Provide brief description about the supplied Raw Data Sources before processing.

1. Raw data sources,
2. the processed/transformed data,
3. and feature sets.

For each dataset, an individual report describing

* Schema of the Raw Data.
* The meaning of each data field
* and other information that is helpful for understanding the data is provided.
* If the dataset is the output of processing/transforming/feature engineering existing data set(s), the names of the input data sets, and the links to scripts that are used to conduct the operation are also provided.

.

## Raw Data Sources

## Provide details about the Raw Data Sources shared by the client.

| **Dataset Name** | **Original Location** | **Destination Location** | **Data Movement Tools / Scripts** | **Link to Report** |
| --- | --- | --- | --- | --- |
| DatasetX | Brief description of its original location | Brief description of its destination location | [Details](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/python/script/file/in/Code) of the methodology used for ingesting and processing data. This could be a python file or similar. | [This](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/report1) should be a link to where this data is stored |

* **Dataset X Parameters**. For each dataset provide following information:
  + Size/Volume of the data
  + How is the data accessed such as via API or from SQL dumps, excel, CSV files. Any login credentials utilized must also be logged here.
  + Meta Data details.
  + Data Schema developed or existing.
  + Use separate headings for each of the above.

## Processed Data

Provide details about the output from the processing exercise. Basically all relevant details related to how data was processed from its raw format to a format or shape that is in needed shape for the Data mining requirements.

| **Processed Dataset Name** | **Input Dataset(s)** | **Data Processing Tools/Scripts** | **Link** |
| --- | --- | --- | --- |
| Processed Dataset 1 | Name of the Raw Dataset in previous section | [Links](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/python/script/file/in/Code) to tools & techniques used to generate this output. This may be code or script or a GUI based flow. | Provide a link to this output. |
| Processed Dataset 2 | Name of the Raw Dataset in previous section | [Same](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/R/script/file/in/Code) as above. | Provide a link to this output. |

* Processed Data1 summary. Provide summary of rationale behind performing data processing steps in this step.

## Feature Sets

This section is meant to provide details of Feature Engineering (If any) performed on the Raw Data Set

| **Feature Set Name** | **Input Dataset(s)** | **Feature Engineering Tools/Scripts** | **Link** |
| --- | --- | --- | --- |
| Feature Set 1 | DatasetX, [Pro](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/dataset2/report)cessed DatasetX | [Links](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/python/script/file/in/Code) to tools & techniques used to generate this output. This may be code or script or a GUI based flow. | Provide a link to this output. |

* Provide meaning of each feature generated as a part of this exercise.
* Any dummy variables created.
* Techniques used to deal with categorical data.

# General summary of the data

This should include atleast following:

* Get Data Shape
* Get Attribute Description and Variable Type
* Get Summary for each Attribute/Feature such as 5 point summary, IQR etc.
* Get ols (ordinary least squares) results for each feature.
* Generate linear and Pair Wise plots for each feature. This should include Histograms and scatter plots. This should be both univariate and bi variate plots.
* Generate Correlation and Variance Covariance matrices for each variable.
* In general all major attributes from either of the following tools can be inserted here:
* Microsoft IDEAR
* RATTLE , Rcommander
* Pandas profiling

# Data Quality Issues

* Determine missing records for each Feature.
* Determine records with Errors, Div0, NaN etc.
* Determine outliers in the data.
* Determine the records where value data type is unexpected i.e. different from majority values.
* Determine records where unexpected value is found.

# Data Transformation Steps

List details related to:

* Standardization Performed on the Raw Data.
* Feature Scaling Performed and the type of scaling used.
* Data Normalization
* Data Discretization Performed.
* Any feature Reduction steps taken such as PCA etc. Refer using multicollinearity and correlation metrics. Similarly columns with only a single value can be removed.

# Data quality summary

Get following statistics related to the source data:

|  |  |  |
| --- | --- | --- |
| **Data Quality Dimensions** | |  |
| **Dimensions of data quality** | **Definitions** | **Comments** |
| **Accuracy** | Also known as validity. The degree to which the measures conform to defined business rules or constraints. The data measure what they are intended to measure. Following data constraints should be considered .   * *Data-Type Constraints* – e.g., values in a particular column must be of a particular datatype, e.g., Boolean, numeric (integer or real), date, etc. * *Range Constraints:* typically, numbers or dates should fall within a certain range. That is, they have minimum and/or maximum permissible values. * *Mandatory Constraints:* Certain columns cannot be empty. * *Unique Constraints:* A field, or a combination of fields, must be unique across a dataset. For example, no two persons can have the same social security number. * *Set-Membership constraints*: The values for a column come from a set of discrete values or codes. For example, a person's gender may be Female, Male or Unknown (not recorded). * Regular expression patterns: Occasionally, text fields will have to be validated this way. For example, phone numbers may be required to have the pattern (999) 999-9999. * Cross-field validation: Certain conditions that utilize multiple fields must hold. For example, in laboratory medicine, the sum of the components of the differential white blood cell count must be equal to 100 (since they are all percentages). In a hospital database, a patient's date of discharge from hospital cannot be earlier than the date of admission. | Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Reliability** | Check and confirm whether the data generated by a program’s information system is based on protocols and procedures that do not change according to who is using them and when or how often they are used. The data are reliable because they are measured and collected consistently. | Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Precision** | This means that the source data has sufficient detail. For example, an indicator requires the number of individuals who received HIV counseling & testing and received their test results by sex of the individual. An information system lacks precision if it is not designed to record the sex of the individual who received counseling and testing. | Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Completeness** | Completeness means that an information system from which the results are derived is appropriately inclusive. It represents The degree to which all required measures are known. | Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Timeliness** | Data are timely when they are up-to-date (current), and when the information is available on time. Timeliness is affected by: (1) the rate at which the program’s information system is updated; (2) the rate of change of actual program activities; and (3) when the information is actually used or required. | (*May or may not be applicable*) Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Integrity** | Integrity issues arise when the system used to generate them are not protected from deliberate bias or manipulation for political or personal reasons. | Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Confidentiality** | Confidentiality means that clients are assured that their data will be maintained according to national and/or international standards for data. This means that personal data are not disclosed inappropriately, and that data in hard copy and electronic form are treated with appropriate levels of security (e.g. kept in locked cabinets and in password protected files. | Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Uniformity** | This metric specifies whether all concerned methods of data collection, including units of measure are uniform across all system. This implies that test data unit must be normalized to a common unit for all tests. | Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Consistency:** | Consistency Issues occur when a single attribute is recorded in different ways such as a patient name being recorded differently due to either spelling mistakes or otherwise against a common MR number or different. | Add your comments and qualifying remarks about each of the Quality Dimensions. |
| **Data Management Processes and Data Quality Controls** | * Does clear documentation of collection, aggregation and manipulation steps exist? * Are data quality challenges identified and are mechanisms in place for addressing them? * Are there clearly defined and followed procedures to identify and reconcile discrepancies in reports? * Are there clearly defined and followed procedures to periodically verify source data? | Add your comments and qualifying remarks about each of the Quality Dimensions. |

# Identification of Target & independent variables

* Identify all probable target columns
* Visualize all classes manifested in each Target Variable.
* Identify all independent variables.
* Briefly explain the meaning and significance of each variable.
* Discuss and visualize relationship between Target and independent variables. This could be done using Bi variate scatter matrix plots or Variance Covariance matrix as well as correlation matrix plot.

# Variable ranking

Perform variable importance ranking for all independent variables.

This can be achieved using Correlation between a feature and the define target variable or using PMI or pointwise mutual information content concepts. Other methods include RFE (recursive feature elimination), Random Forest feature importance etc.

# Insight learned

This section is meant to summarize key findings related to various aspects of the workflow

## Project Execution

What did we learn in relation to the customer engagement process as far as data collection and access is concerned. How could this be improved.

## Data science / Engineering

Learnings related to data science/engineering, tips/tricks, etc>

## Domain

Any particular insights and observations related to the Specific domain of healthcare which may be helpful in pattern finding in the future.

## Product

What are the different products or services that have been utilized to deal with the data. Key findings here.

## Any specific challenges

Specific issues or setup, unique things, specific challenges that had to be addressed during the engagement and how that was accomplished.

# Next Steps

This should include a brief plan of activities planned following those outlined in this document. This should include milestones, resources etc.