# How to do HR analytics data processing the correct way

**Author: Benjamin Sombi . March 2020**

Most of the HR people we deal with know where to get the HR data they may need to make sense of. They also know about what they want to achieve but struggle with the most important part of HR analytics, data processing. How can HR people capture and process data before analysis or sent for analysis.

Data processing is turning raw data into usable data which can be easily interpreted by a person. Data analytics refers to qualitative and quantitative techniques used to study behavioral data and patterns with a view to improving decision making.

Please note that data processing requires a lot of time and care. It may take the greatest amount of time than any other activity in the analytics project. A lot of mistakes can accumulate at this stage leading to an ultimate model that outputs biased results.

The data process stage is divided into the following sub – stages, data capturing, data coding, data understanding, data enrichment and imputation of outliers and missing values

1. Data capturing and coding

Different data sources provide their data in different forms. Some of the sources provide data in forms that cannot be read by computers e.g. a survey may use paper – and – pencil questionnaires. The responses are entered into computer spreadsheets so that subsequent activities can be carried out on the data. This process, known as **data capturing is the changing of so many forms of data into a form that that can be read by computers**. All data are captured from their disparate sources and integrated into one master file.

**Data coding is the transformation of captured data into forms that are understood by the computer software** to be used for analytics e.g. if gender is one of the captured variables, it may need to be coded as “1 for males” and “0 for females” for a computer software such as SPSS to be able to understand the data.

Data capturing and coding is error prone. Great care should be given to this stage.

1. Data understanding

Once the data has been captured and coded, the next stage is to summarise the variables. This is a necessary step as it provides insight as to which variables should be included in the analytics model building process.

Below is a list of some of the variable summaries that can be studied for continuous variables.

- Mean
- Standard deviation or variance
- Skewness
- Kurtosis
- Percentiles
- Interquartile range
- Correlations etc.

The following questions can be asked when auditing a continuous variable.

- Are there strange minimum or maximum values? (Question is related to outlier detection)
- How many missing values are there? Do any variables have mostly or all missing values? (Question is related to imputation of missing values)

- Are there strange mean values or large differences between mean and median?
- Is there large skew or excess kurtosis? (This only matters for algorithms that assume normal distributions in the data.)
- Are there gaps in the distributions, such as bi-modal or multi-modal distributions?

The following questions may be asked when auditing a categorical variable.

- Are there any values in the categorical variables that don't match the dictionary of valid values? (Typing errors may be corrected by answering this question)
- Are there any categorical variables with many levels? (Categorical variable with a huge number levels may present problems when building models since the required computational power is significantly large)
- Are there any categorical variables with large percentages of records having a single value?
- Are any variables highly correlated with each other, possibly indicating redundant variables? (Question related to data enrichment)
- Are there any crosstabs that show strong relationships between categorical variables, possibly indicating redundant variables? (Question related to data enrichment)

Data visualisations at this stage are used to enhance the exploration of relationships which may not be visible when one looks at data tables alone. There are a lot of visualisations types that can be used. Each is best suited for specific kinds of analysis results. The list below gives some of the visualisation techniques which can be used in this exploratory data analysis stage (EDA).

- Histograms (Single variable, overlays etc.)
- Boxplots
- Stem and leaf diagrams
- Line graphs
- Scatterplots etc.

- Bar plots etc.

1. Data enrichment

This stage involves processes used to enhance, refine or improve the raw data in the master file. For example, counting the number of disciplinary hearings per employee for grouped years may provide more insight than studying the number of hearings per year i.e. Study the number of disciplinary hearings in the periods 2009 – 2011, 2012 – 2014 and 2015 as opposed to the periods 2009, 2010, 2011, 2012 etc. The motivation for grouping the years may be major management changes corresponding to the given grouped periods.

Note that data enrichment requires the improvement of raw data by any means necessary. This requires consulting domain experts.

1. Imputation of outliers and missing values

Please note that imputation of outliers and missing values requires great expertise. Not for the faint heartedJ.

After data gathering, capturing and coding, there will be outliers and missing data in the resulting data set. These pose problems when building the machine learning models.

Outliers are observations that are distant from other observations. They may be as a result of the variability in the variable (representative outliers) or as a result of errors in observation (non – representative outliers). (However) the distance at which a value is considered as an outlier is subjective. It may be relaxed depending on the model that needs to be built.

There are several ways of handling outlier and missing values but we will not cover that in this article.

In the article we covered how to do data processing in a nutshell. There are several activities that are done in each step. Most of these activities are technical. This means that they require someone with statistics background.

***Benjamin Sombi is a Data Scientist, Entrepreneur, & Business Analytics Manager at Industrial Psychology Consultants (Pvt) Ltd a management and human resources consulting firm.***

https://thehumancapitalhub.com/articles/HowToDoHRAnalyticsDataProcessingTheCorrectWay