**Integrated CA 2**

**Louis Wilkie**

**Table of Contents**

* **Introduction**
* **Objectives**
* **Results: Data Preparation**
* **Results: Statistical Analysis**
* **Results: Machine Learning**
* **Discussion/Conclusion**
* **References**
* **Appendix**

**Introduction**

This project will analyse a dataset: Employee\_attrition.csv.

This project will follow the CRISP-DM framework, and a

**Objectives**

This assignment has a few objectives under a few key areas: data preparation, statistical techniques and machine learning models.

The main objective of this assignment is to investigate employee satisfaction and productivity for a company, which are expressed by two random variables of the provided dataset. This translates to modelling the independent variables that contribute to these metrics and takes the task into a machine learning problem.

The company has asked for specific evaluation of data preparation, statistical techniques, and machine learning. This is provided in the results under the appropriate headings. Tables and graphs are included in the appendix, and also in the attached jupyter notebook file.

**Results: Data Preparation**

**1. Characterisation of the data set: size; number of attributes; has/does not have missing values,**

**number of observations etc.**

For data preparation outcomes of this assignment a few tasks are identified.

* There is considerable missing data, which is identified as randomly missing.
* The dimensions of the data indicate that the *curse of dimensionality* may be a limiting factor to modelling.
* Data is out of scale for machine learning modelling.
* Categorical features are not encoded.

From these the following strategies are employed.

* Missing data is imputed in trials.
* Dimensionality reduction techniques are trialled.
* Different scaling techniques are trialled.
* Different encoding techniques are trialled.

The strategies above are cumulated in Data Preparation Experiment 1, which is shown in the table below.

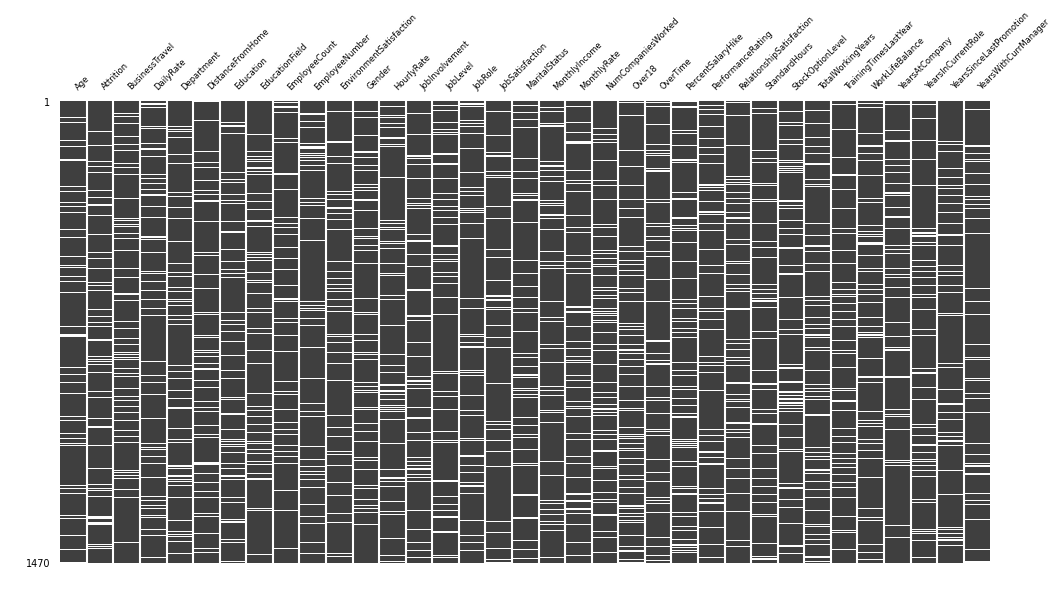
While modelling increased from the established baseline model

if data is largely uncorrelated what does it mean for data analysis

These tasks are identified with data exploration techniques, and throughout the assignment the framework of CRISP-DM recommends re-examining data preparation techniques, and investigation into possible employable stategies.

Data exploration is a key part of data preparation

There is considerable missing data in the dataset. Each feature has 10% missing values. The values are missing at random with no discernable pattern. Only 40 rows contain no missing values and thus trailing imputation strategies is recommended.

Figure 1: Missigno Missing Data Matrix

Sampling data...

**2. Application of Data preparation/evaluation methods (Cleaning, renaming, etc) and EDA (Exploratory Analysis) visualizations (plural), including a clear and concise explanation of your rationale for you are doing with the data and why you are doing it.**

**3. Apply encoding, scaling and feature engineering as and if required, detailing how and why you used techniques and the rationale for your decisions.**

**4. Explore the possibility of using dimensional reduction on the dataset. Employ both LDA (Linear Discriminant Analysis) and PCA (Principal Component Analysis) and compare the separation of classes through visualization. Explain the difference between both techniques in your own words and discuss in detail how your results may affect your analysis of classifying or clustering the normal compared to anomalous biddings.**

**Results: Statistical Techniques**

**1. Use descriptive statistical analyses to explore and evaluate the data set, including measures of**

**central tendency and dispersion and frequency distributions. Correlation matrices are also**

**accepted. Provide a summary of your findings.**

Descriptive statistics are given in the figures at the end of this section. Frequency distributions are included in the appendix. In lieu of a correlation matrix a seaborn correlation heatmap has been included.

Using the describe function yields a lot of information with 35 columns. And while it can seem like a lot of information there is still value in examining it and interesting information about the data can be understood quickly. We will discuss some of this below.

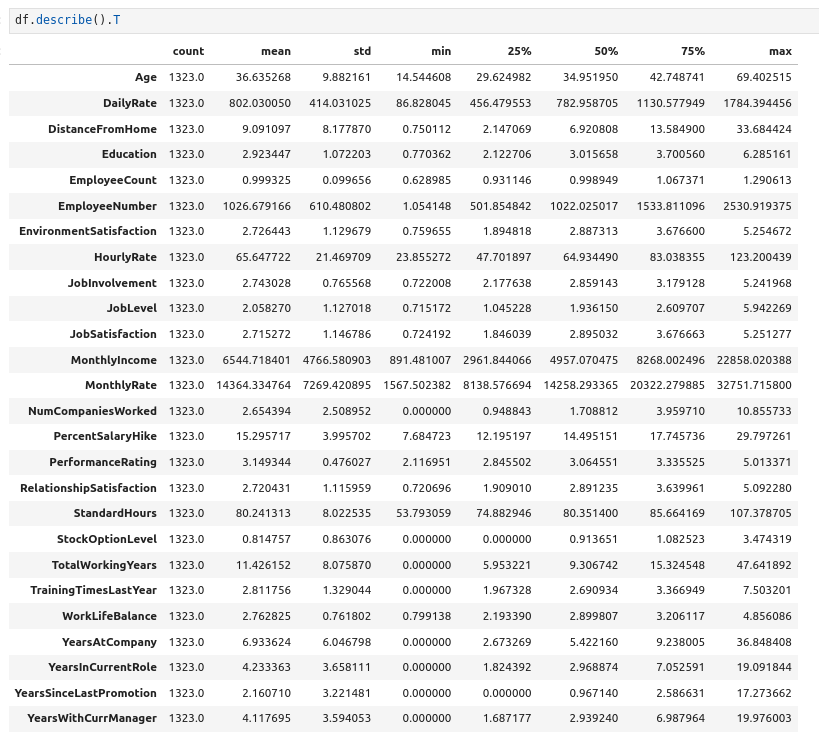
Looking at the minimum value for age seems low for a dataset of employees; 14.544608, and it is at a level of precision unusual for recording a persons age. It would indicate a likely error, or encoded or dirty data. All the data seems to be recorded in this level of precision, and here by using descriptive statistics we can formulate a line of enquiry in data preparation.

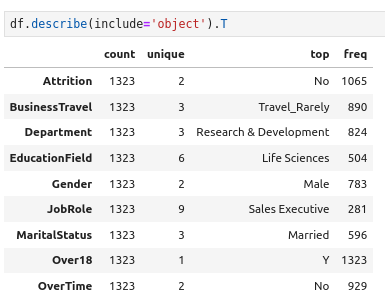
We can understand the shape and spread of the data by examining the descriptive statistics as a whole.

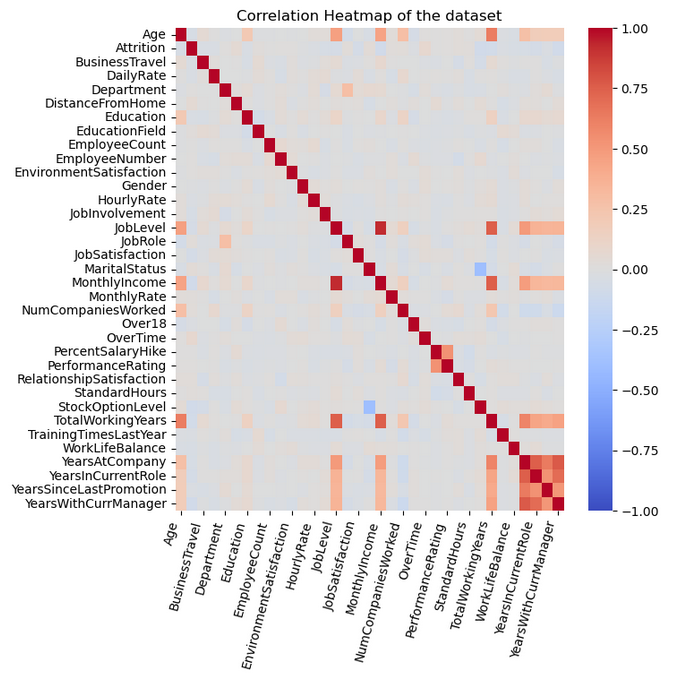
* *DailyRate* stands out as being widely spread over its range, by having a relatively large standard deviation against its mean and range. Its mode is its min value so it could be skewed to the right but likely has a lot of variance.

The descriptive statistics for objects reveal some imbalanced categories as well as one category with a constant value. Again, this informs our data preparation and we can decide to balance these categories and remove the feature with one constant value. We also see categories that contain interesting contextual information that will be useful for ANOVA.

For this data the correlation matrix is interesting. There seems to be a lot of neutral correlations which indicates a lack of linear relationships with the data. This could make linear modelling of the data troublesome. And with some trials of modelling different target features it did.

Figure 2: Descriptive Statistics with pandas

Figure 3: Descriptive statistics of objects with pandas

Figure 4: Correlation Heatmap

**2. Formulate and test hypotheses within a business context using appropriate statistical techniques t-tests or ANOVA to identify significant relationships between variables. Provide a summary of findings. Use at least two statistical tests.**

**3. Use a Jupyter notebook to produce result sets from the provided dataset, such as scatter plots or regression models. Provide a summary of your findings.**

**4. Write the results of the analysis of your findings to stakeholders using clear and concise explanations, visualisations, and appropriate statistical terminology.**

**Results: Machine Learning**

**1. Provide a conceptual understanding and logical justification based on the reasoning for the specific of machine learning approach (supervised/ Unsupervised) for the provided data set. You can the pros and cons of both approaches based on your understanding.**

What is viable? Supervised learning for attrition as it is modelling the data effectively. Unsupervised learning, clustering appropriate to evaluate metrics discussed.. will do if time]]

To model data with supervised learning you need an appropriate target variable with labels. The data does seem to have this in the Attrition feature and the features are thematically related to attrition. It may not be the motif of the dataset, which is information on employee work history and satisfaction, but it is related to these variables in ways and this can be explored with supervised learning modelling.

Other features..

**2. Machine Learning models can be used for Prediction, Classification, and Clustering. You can choose features for the machine learning models based on feature selection methods, such as random or any other method. The selection of hyperparameters for the ML models should be performed using hyperparameter tuning, such as GridSearchCV. Obtain the best accuracy using optimal values the hyperparameters.**

Choose features for machine learning modelling:

Machine Learning models used:

Best model accuracy achieved:

**3. You should train and test the Machine learning models in the case of supervised learning for different (at least 2 splits) and use appropriate metrics for unsupervised learning. Use k-fold (10 or 20 or ) cross-validation to provide authenticity of the modelling outcomes.**

For this section

**4. Exhibit a comparison of ML modelling outcomes using a Table or graph visualisation. Identify the similarities and contrast of the Machine Learning modelling outcomes based on chosen metric discuss their statistical understanding.**

Below is a detailed table on modelling outcomes of this assignment.

**Conclusion**

**References**

**Appendix**