**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.

**Objectives**

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

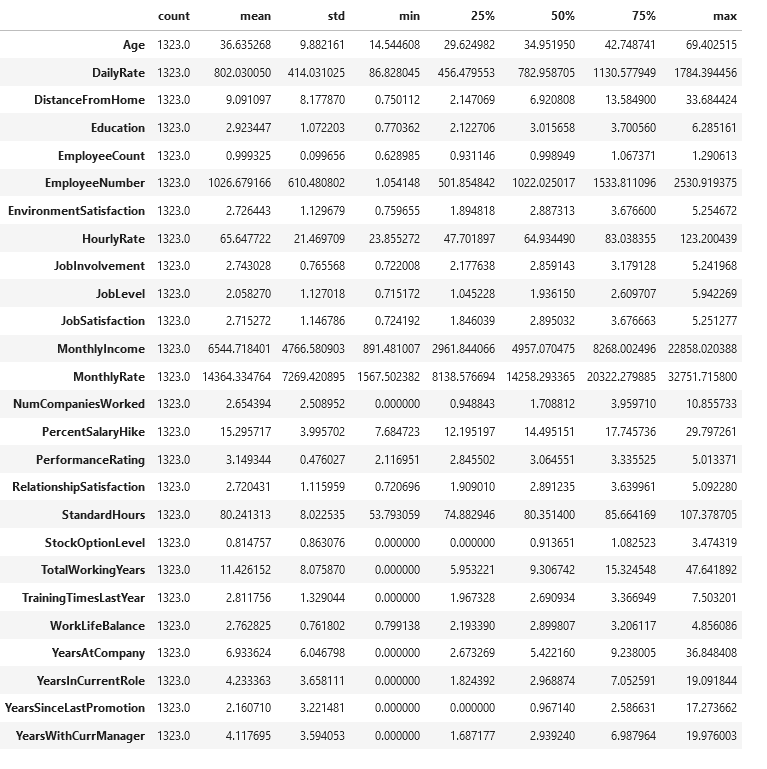
The main objective of this assignment is to investigate employee satisfaction and productivity for a company, which are expressed by two random variables of the provided dataset. This translates to modelling the independent variables that contribute to these metrics and takes the task into a machine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

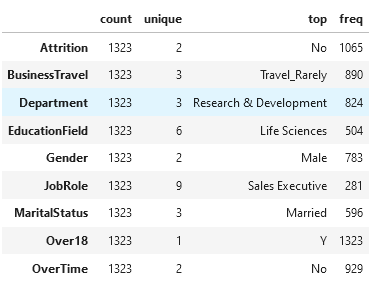
**Results: Data Preparation**

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

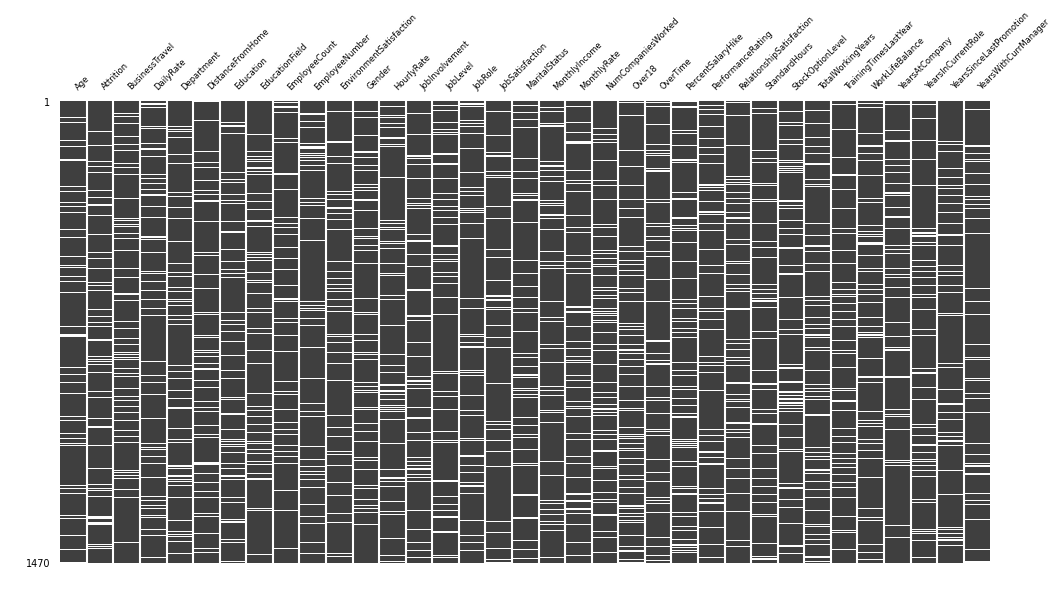
The dataset for this task is called *Employee\_attrition.* It has 35 features and 1470 observations. It is a mixture of categories and numerical data, with 9 categories and 26 numerical features. The numerical variables are continuous features, while the categorical variables are a mixture of categories of different number of unique values. Below are the summary statistics for the dataset. The numerical data is of different scales and are all floating point numbers.

Figure 1: Summary statistics for Employee\_attrition dataset

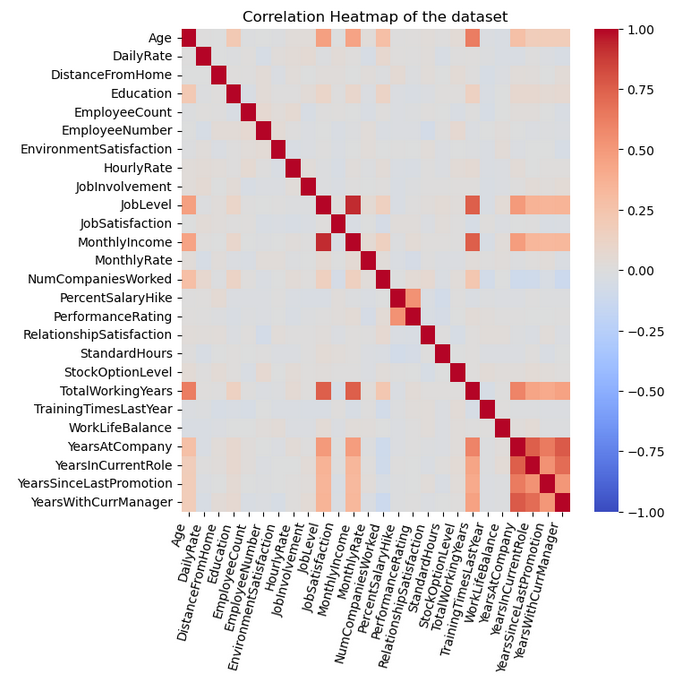
The categories below are largely unbalanced.

Figure 2: Summary statistics for objects

There is considerable missing data in the dataset. Each feature is missing 10% of its values, amounting to 147 for each feature. The values are missing at random with no apparent pattern. Only 40 rows contain no missing values, and below there is a *missingno* missing data matrix which highlights the randomness in the missing data as no discernable pattern can be seen in the data.

Figure 3: Missingno Missing Data Matrix

The numerical data shows mostly neutral correlation. Although there are some variables that show a positive linear relationship, which are mostly concerning time lengths related to the employees work history; employees that have the highest amount of working years tend to be the employees that have worked with the company the longest, etc.

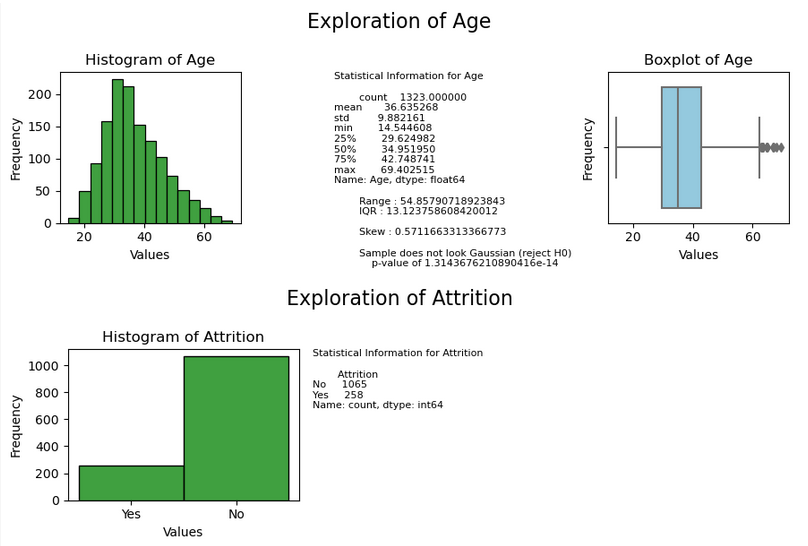
Figure 4: Correlation Heatmap

This was a quick characterisation of the data. From which strategies are developed for data preparation. The data has missing data that will need to be dealt with, low numerical correlations, a wide range of numerical data ranges, and unbalanced categories.

**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.**

The data is first looked at as a whole, looking at correlations of the data and the characterisations which are discussed in the previous section. This gives an overview of the data, but also presents more questions. To articulate these issues further exposure of the data is needed.

In this case, there are a number of variables that make individual exploration of them possible with the time constraints given. If this wasn’t the case, groupings of the variables can be already seen that could logically be processed in the same manner. The features are first visually explored as follows.

Figure 5: Visual exploration of numeric and categorical variables

The full visual exploration of the variables is included in the appendix.

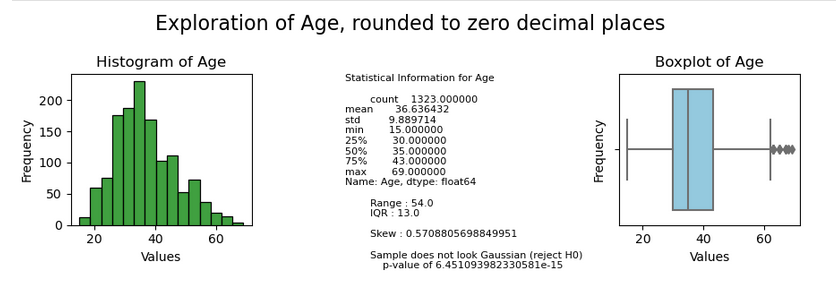
Each set of data has its own characteristics. Two motifs are prevalent in this data; precision and noise. Each of the numerical features are processed with this in mind. The steps to understand the noise and precision issues are discussed next throughout the exploration.

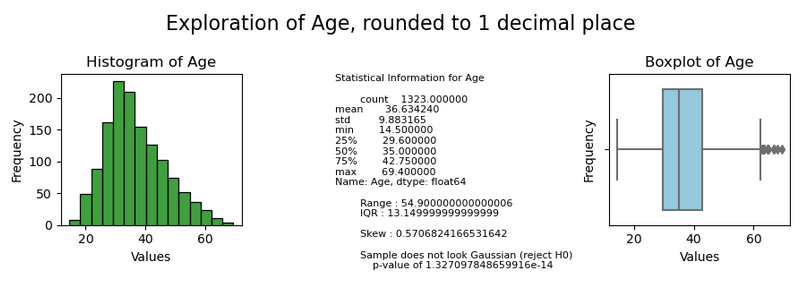
The figure below shows *Age* where the employee is under 18, with the *Over18* feature. It a discrepancy that demands attention, and indicates that the data is at least misaligned in places. This is the first indication of several for the dataset.

Figure 6: Employees are labelled as both under and over 18 in the data.

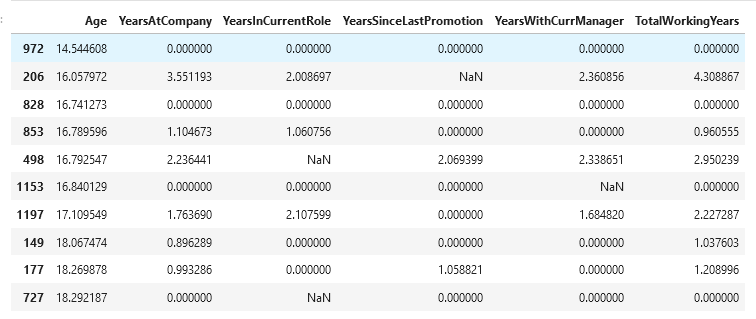
*Over18* will not be used for analysis as it has *Y* as the only value and therefore lacks signal, but to remove it and ignore the issue would be hasty. Several options were considered but the discrepancies in *Age* were set to a new minimum value of 18 for the feature, this would change seven values of the data and would add an assumption to the analysis: it is assumed that every employee is over 18, based on the values recorded for the removed Over18 feature. Although with encoding later this assumption can be removed.

*Age* has an unrealistic level of precision, but the range doesn’t seem unusual. Their are outliers of employees over 60, and this seems to make a kind of sense to work with. Logically the precision in *Age* can only seem to be noise, and is unrealistic for any record keeping. The feature could be rounded to a whole number but in this the data is shape is kept best at 1 decimal place, while also representing the majority of the noise present in the data.

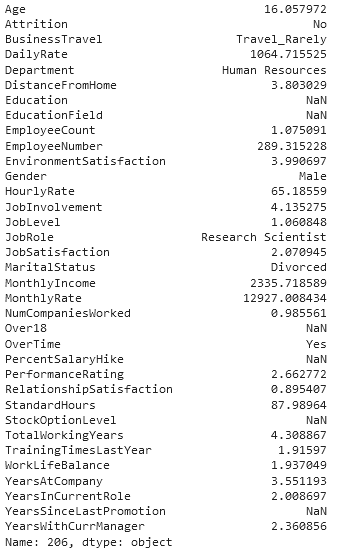
Figure 7: Age variable rounded to a whole number. (see Figure above for original shape)

Figure 8: Age rounded to one decimal place.

The year features have the same issues. Some employees are not recorded as working yet are listed as employees. This could be omitted data entered as zeros. Some employees have *TotalWorkingYears* less than some of the metrics which shouldn’t be.

Figure 9: The working years of the youngest employees

From the figure above, one sixteen year old has already been working for 4 years. The figure below looks at the entries for him. The data isn’t just sparse but quite dirty.

Figure 10: A 16 year old divorced research scientist that works in human resources, in his fourth working year.

To counteract the noise some manual cleaning of the year features is done and the years are binned into categories so a broader representation of them can be used for modelling, the opposite of the precision the original data offered.

More data preparation is done as follows: features that represent monetary values are rounded to two decimal places, which is logical but also preserves the integrity of the feature as before with *Age.*

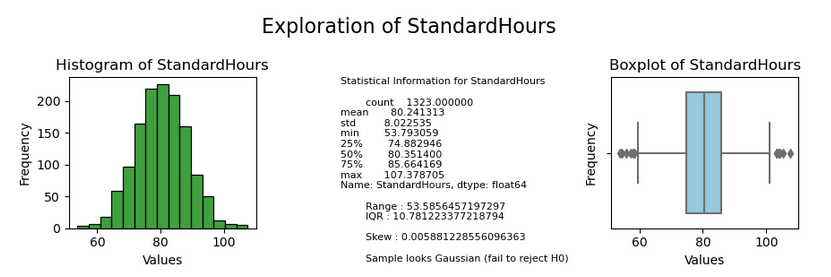
Some features are described in the data dictionary as “coded as integers”, and when inspecting them they look like they could better be represented as categories.

Missing data is handled differently,

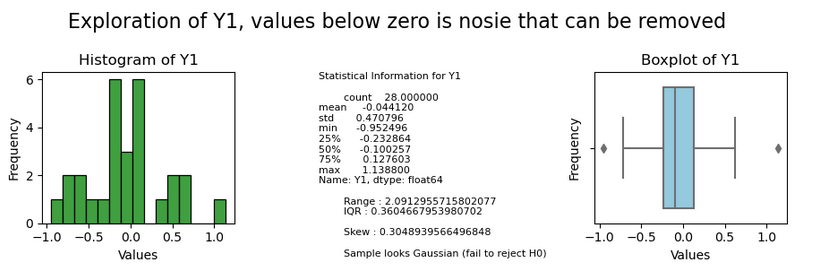
At the end of data preparation there are now 32 features, 20 numerical features and 12 categories

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

Some feature were engineered for the data, which increased dimensionality. These features were the continuous features representing timescales. These binned categories had missing values replaced with “Unknown”. As per the section above, this was done to try and capture the signal in the data by examining the data more broadly as questions had been raised on the nature of some of the noise in the data. One feature listed as a constant value in the data dictionary exhibited a normal distribution.

Figure 11: StandardHours feature was listed as a constant value in the data dictionary.

And a sample of errors in some of the year features seemingly comes from a normal distribution. The figure below shows *TotalWorkingYears* minus *YearsAtCompany.* With the values below zero representing some of the noise.



This was the justification for converting some of the numerical features to categories. While a model is the sum of the actual value plus error, effort has been made to alleviate the impact of what is shown above to be an oscillating Gaussian error, but is expected to be exhibited in most of the features of the data.

The data was encoded, scaled and imputed in trials. For these trials a Logistic Regression model of *Attrition* was created to refine the choices. The results of those trails are shown in table 1.

For these trials the data was under sampled to balance the *Attrition* class.

Table 1: Data preparation trials with under sampling of Attrition class

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trial | Imputer | Scaler | Encoder | Accuracy |
| 1 | SimpleImputer | StandardScaler | OneHotEncoder | 0.625 |
| 2 | SimpleImputer | MinMaxScaler | OneHotEncoder | 0.653 |
| 3 | SimpleImputer | RobustScaler | OneHotEncoder | 0.625 |
| 4 | SimpleImputer | StandardScaler | OrdinalEncoder | 0.615 |
| 5 | SimpleImputer | MinMaxScaler | OrdinalEncoder | 0.634 |
| 6 | SimpleImputer | RobustScaler | OrdinalEncoder | 0.615 |
| 7 | KNNImputer | StandardScaler | OneHotEncoder | 0.625 |
| 8 | KNNImputer | MinMaxScaler | OneHotEncoder | 0.653 |
| 9 | KNNImputer | RobustScaler | OneHotEncoder | 0.625 |
| 10 | KNNImputer | StandardScaler | OrdinalEncoder | 0.615 |
| 11 | KNNImputer | MinMaxScaler | OrdinalEncoder | 0.634 |
| 12 | KNNImputer | RobustScaler | OrdinalEncoder | 0.615 |
| 13 | IterativeImputer | StandardScaler | OneHotEncoder | 0.625 |
| 14 | IterativeImputer | MinMaxScaler | OneHotEncoder | 0.653 |
| 15 | IterativeImputer | RobustScaler | OneHotEncoder | 0.625 |
| 16 | IterativeImputer | StandardScaler | OrdinalEncoder | 0.615 |
| 17 | IterativeImputer | MinMaxScaler | OrdinalEncoder | 0.634 |
| 18 | IterativeImputer | RobustScaler | OrdinalEncoder | 0.615 |

With under sampling a pattern emerges, and the imputation method seems to not impact the results. The trial is conducted again with over sampling of the *Attrition* class and the results are blow in Table 2.

Table 2: Data preparation trials with over sampling.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trial | Imputer | Scaler | Encoder | Accuracy |
| 1 | SimpleImputer | StandardScaler | OneHotEncoder | 0.718 |
| 2 | SimpleImputer | MinMaxScaler | OneHotEncoder | 0.706 |
| 3 | SimpleImputer | RobustScaler | OneHotEncoder | 0.711 |
| 4 | SimpleImputer | StandardScaler | OrdinalEncoder | 0.714 |
| 5 | SimpleImputer | MinMaxScaler | OrdinalEncoder | 0.725 |
| 6 | SimpleImputer | RobustScaler | OrdinalEncoder | 0.714 |
| 7 | KNNImputer | StandardScaler | OneHotEncoder | 0.720 |
| 8 | KNNImputer | MinMaxScaler | OneHotEncoder | 0.718 |
| 9 | KNNImputer | RobustScaler | OneHotEncoder | 0.721 |
| 10 | KNNImputer | StandardScaler | OrdinalEncoder | 0.721 |
| 11 | KNNImputer | MinMaxScaler | OrdinalEncoder | 0.723 |
| 12 | KNNImputer | RobustScaler | OrdinalEncoder | 0.723 |
| 11 | IterativeImputer | StandardScaler | OneHotEncoder | 0.702 |
| 14 | IterativeImputer | MinMaxScaler | OneHotEncoder | 0.709 |
| 15 | IterativeImputer | RobustScaler | OneHotEncoder | 0.702 |
| 16 | IterativeImputer | StandardScaler | OrdinalEncoder | 0.699 |
| 17 | IterativeImputer | MinMaxScaler | OrdinalEncoder | 0.704 |
| 18 | IterativeImputer | RobustScaler | OrdinalEncoder | 0.695 |

With over sampling the imputation method is expressed and instead of a repeated pattern of results the trials are different, and the accuracy has increased to reflect this.

OrdinalEncoder performs better for mean imputation and KNN imputation, but not for multiple imputation. Min-max scaler performs the best out of any of the scaling methods in each batch. KNN imputation performs best, mean imputation next, and multiple imputation has lowest performance.

There may be logical justification to use any of the methods above, but the methods identified as highest performing will be used for machine learning modelling; OrdinalEncoding, KNN Imputation, and multiple imputation.

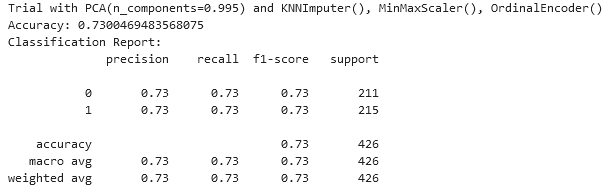
**4. Explore the possibility of using dimensional reduction on the dataset. Employ both LDA (Linear Discriminant Analysis) and PCA (Principal Component Analysis) and compare the separation of classes through visualization. Explain the difference between both techniques in your own words and discuss in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

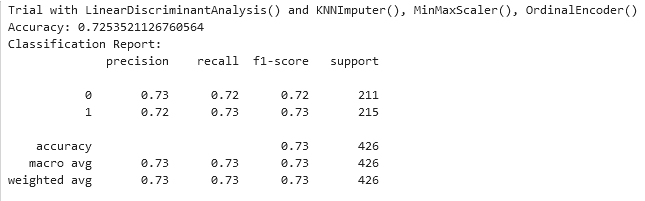
Dimensionality reduction is a technique to aid the processing of data by increasing performance and reducing computational resource needs. When data is modelled the resulting model is nominally agnostic to feature names or data points. Thus data can be processed in different manners that may change the data points but retain the variance, distributions and relationships in the data. Two methods of dimensionality reduction are trialled in this assignment, Principal Component Analysis and Linear Discriminant Analysis.

PCA is concerned with keeping the variance in the data and identifying where the largest amount of variance lies in the data by identifying eigenvectors and eigenvalues of the original data. PCA transforms the data into a dataset of new principal components that represent this variance. A subset of the principal components can be selected to reduce dimensionality while keeping significant information. In this case 99.5% of the variance is kept.

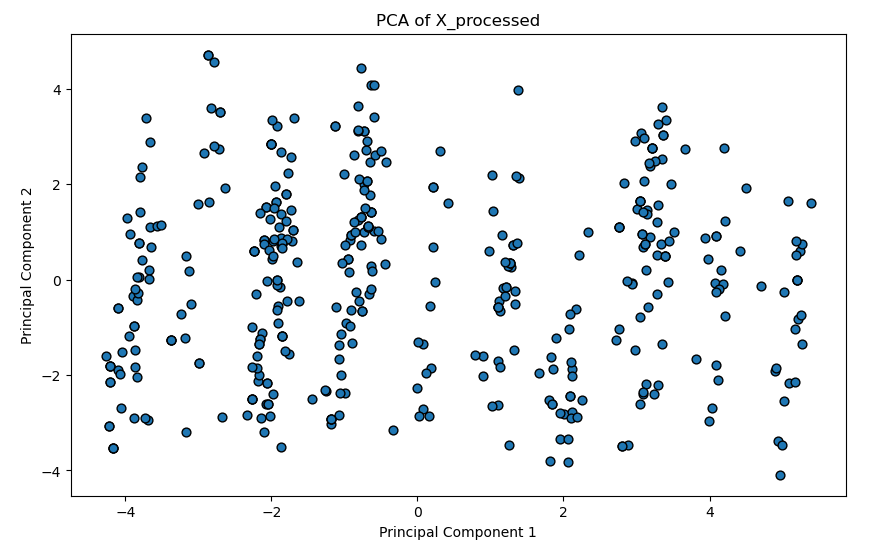
LDA is also a dimensionality reduction technique. It is different to PCA in that it is focused on linear separation of classes. LDA is tailored for supervised learning scenarios, particularly classification tasks. LDA identifies feature space that maximizes the separation between different classes, the core idea is projecting the data onto lower-dimensional space where the distance between classes is maximised. This results in linear combination of features.

Both PCA and LDA were employed on the dataset. First they are included in a Logistic Regression model with the highest performing attributes from the previous trials. The results is models that perform slightly better than previous trials in both cases. Both models generalise well and the dimensionality reduction is successful in both cases.

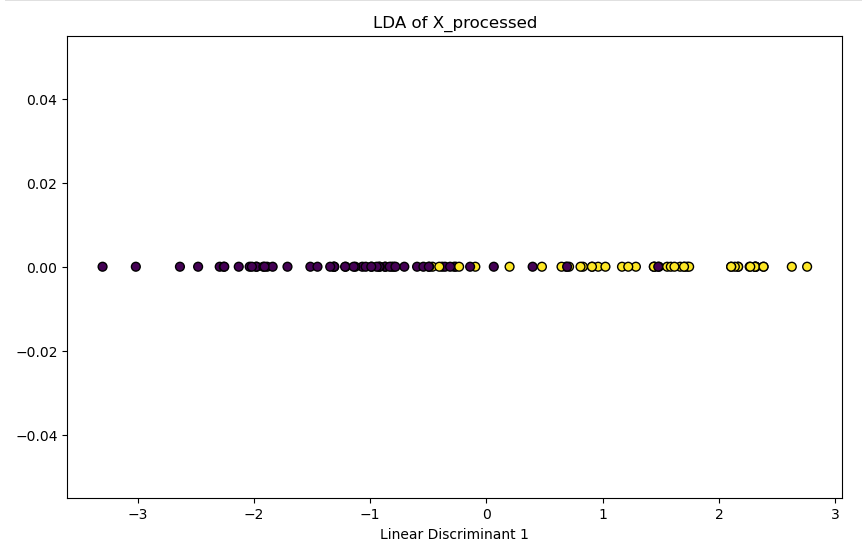
Figure 12: PCA applied with best performing processors.

Figure 13: LDA applied with best performing processors.

While they both seem equally performing for the task some differences are seen when they are plotted.

Figure 14: PCA first two components

For PCA, 99.5% of the variance is held in 29 principal components and the plot of the first two principal components does not seem to hold the variance in a linear relationship.

Figure 15: LDA with class separation.

LDA separates the classes in linear space, although it isn’t a perfect separation.

The lack of linear relationships in the data could be a reason that PCA requires more features to capture the variance of the data. For the classification task LDA is effective at linearly separating the classes, and reduces the dimensions to only one feature making it the ideal choice for use in modelling.

**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.

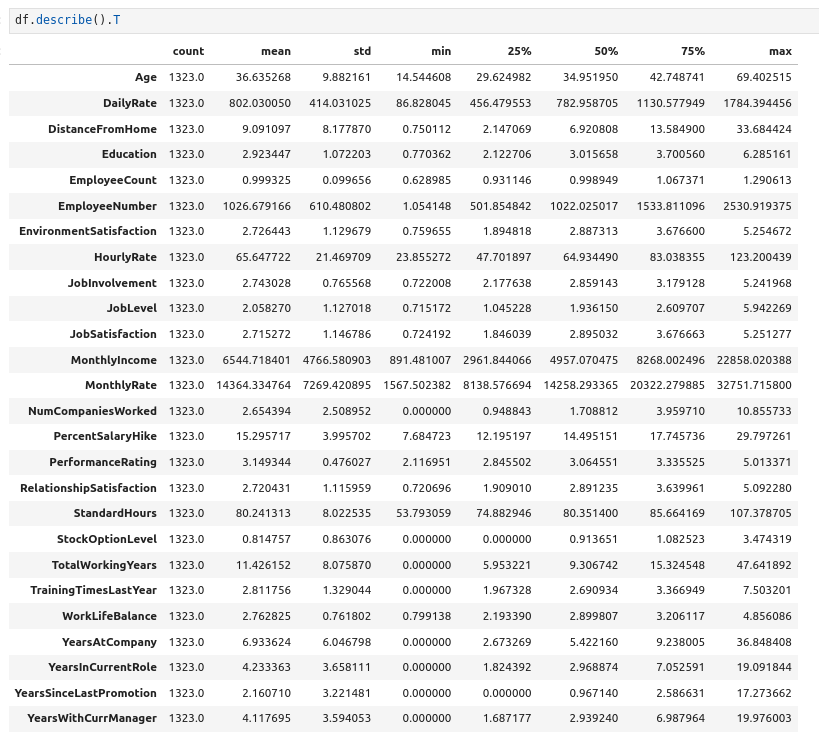
Looking at the minimum value for age seems low for a dataset of employees; 14.544608, and it is at a level of precision unusual for recording a persons age. It would indicate a likely error, or encoded or dirty data. All the data seems to be recorded in this level of precision, and here by using descriptive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

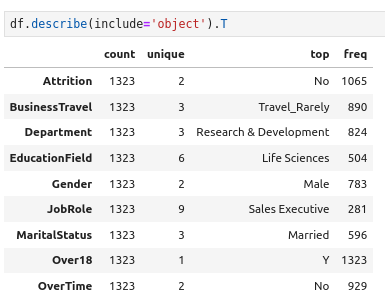
We can understand the shape and spread of the data by examining the descriptive statistics as a whole.

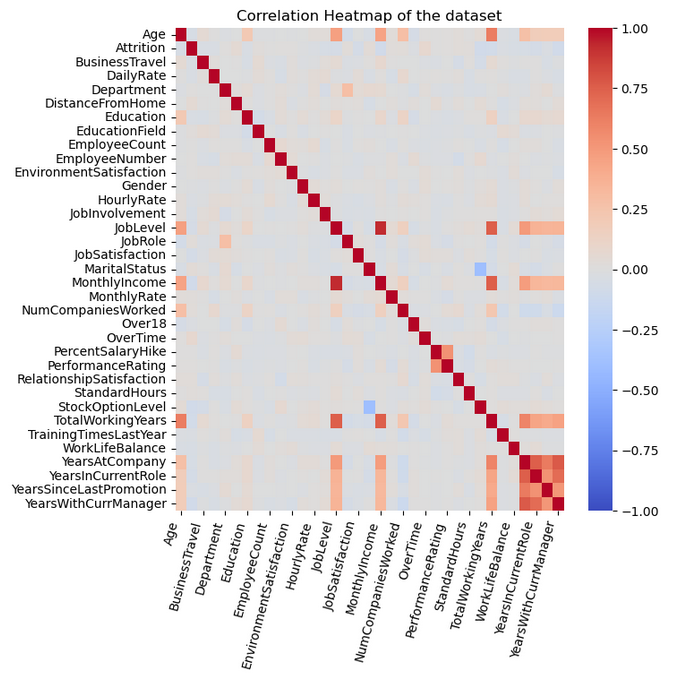
* *DailyRate* stands out as being widely spread over its range, by having a relatively large standard deviation against its mean and range. Its mode is its min value so it could be skewed to the right but likely has a lot of variance.

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 16: Descriptive Statistics with pandas

Figure 17: Descriptive statistics of objects with pandas

Figure 18: 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.**

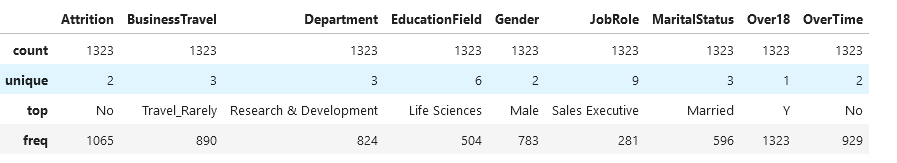
To test hypotheses the original data is used. There are two identified business targets (*JobSatisfaction* and *PerformanceRating*), and first a two sample t test is performed on these features. First we will state the hypothesis:

H0: The samples come from the same population.

H1: The features come from different populations.

The result is a p-value of 1.66e-35. We can conclude with 99% confidence that the target features are independent from each other. The also have a low correlation of 0.02, so while we could have been sure that they are unrelated we have been able to formulate and test a hypothesis to that fact.

Next another interesting feature was modelled; *Attrition.* Chi squared tests were done on *Attrition* and the other categories to examine the expected frequencies and determine if they are based on chance of an underlying relationship.

Figure 19: Categories of unprocessed data.

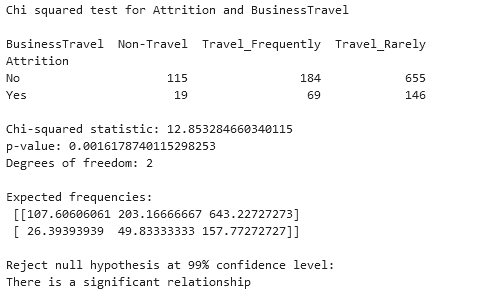
First we will state the hypotheses:

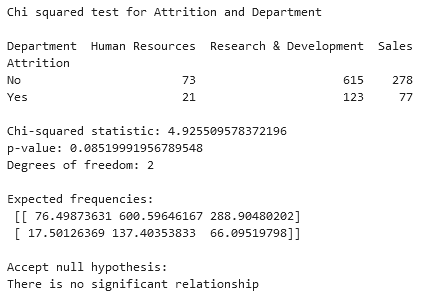
H0: There is no relationship between the features

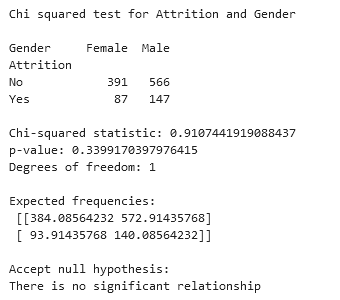
H1: There is a relationship between the features

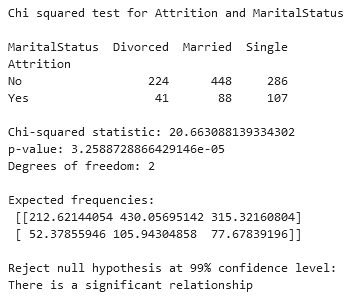
A small p-value, lower than 0.05 or 0.01 would be cause to reject the null hypothesis at 95% oir 99% confidence respectively.

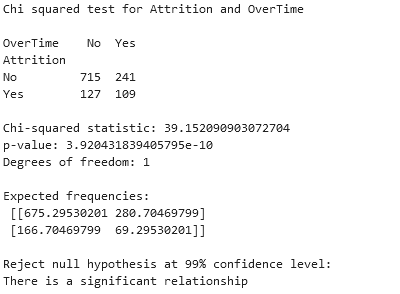
The results of the chi squared test are below. They show that relationships exist between *Attrition* and the some categories that may make modelling appropriate.

Figure 20: Chi square test for Attrition and BusinessTravel

Figure 21: Chi square test for Attrition and Department

Figure 22: Chi square test for Attrition and Gender

Figure 23: Chi square test for Attrition and MaritalStatus

Figure 24: Chi square test for Attrition and OverTime

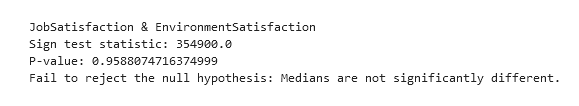
Next the features were checked for normality for modelling with ANOVA. Not all numeric columns exhibited normality for use in ANOVA.

Next a Wilcoxon signed-ranked test is done, which can be used to examine relationships with ordinal data. For this the prepared data is used, which converted several features to ordinal data, First we will state the hypotheses.

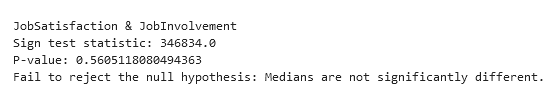
H0: There is no significant difference between the features.

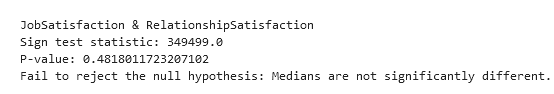
H1: There are significant differences between the features.

The results are included in the appendix, with three that accepted the null hypothesis shown below.

Figure 25: Wilcoxon signed-rank test for JobSatisfaction and EnvironmentSatisfaction

\

Figure 26: Wilcoxon signed-rank test for JobSatisfaction and JobnInvolvement

Figure 27: Wilcoxon signed-rank test for JobSatisfaction and RelatioshipSatisfaction

While there may be limitations to this test it gives an insight that *JobSatisfaction* may have relationships with other ordinal data while *PerformanceRating* lacks these relationships, which may be notable for modelling, and shows an improvement in discovering meaningful relationships in data with data preparation.

**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.**

This assignment attempted to find relationships that exist in the data that relate to job satisfaction and employee performance.

The results of the findings are as follows:

First the data contained features that

**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:

One thing to note about the dataset is it is called *Employee Attrition* and it’s features can be seen as related metrics for uncovering factors of attrition, and while this is the case it’s natural to model attrition and dependent features with machine learning modelling. *Attrition* is a binary category of *Yes* or *No* and it makes this a binary classification (supervised learning) task.

Other interesting features have been identified and these can modelled with machine learning. During processing, these features have been engineered from continuous features to ranked numbers. Investigations into some features during data exploration revealed noise which was particularly visible in the variables concerning time, such as *Age* and *TotalYearsWorked. ADD DIAGRAM*

And it’s with that understanding that the business targets were encoded into ranks, which also thematically better articulate their apparent ordinal nature which they seem to represent. While these variables can be used in a regressive task, the onset of fitting to noise is a looming err, which is somewhat allayed with classification.

So while we may achieve a positive result with regression we are fitting to suspected noise in the data, and also fitting to an unusual level of precision for typical employee scoring metrics. The same scale of precision is seen in all numerical variables of the dataset, and exploration seems to reveal this as noise and not a form of encoding. The distribution of the age feature being close to normal is what is expected, and their values aren’t too discerning bar a few outliers, which helps rationalise that varying levels of noise is impacting *Age* and we extrapolate that logic to the other features.

But these are assumptions, and to get to this point in the data mining process other assumptions on the data have been made. The rounded features are understood as being just that, and when a classification of 5 is achieved its understood that it represents a range of possible outcomes of the original dataset (e.g. 4.74353534). In this manner we tell the client that we can predict their business targets with more integrity after they are encoded into ranks.

On feature selection.

Which models are used and how are they tuned.

Table of hyper parameters?

**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

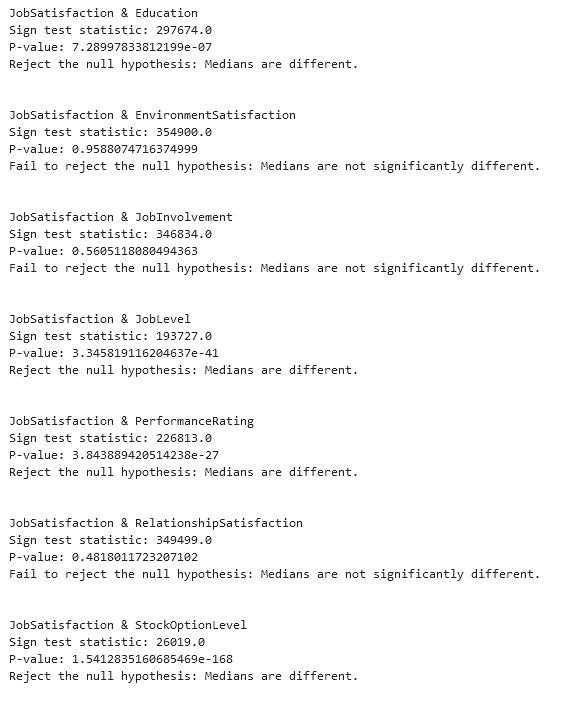
**4. Exhibit a comparison of ML modelling outcomes using a Table or graph visualisation. Identify the similarities and contrast of the Machine Learning modelling outcomes based on chosen metric discuss their statistical understanding.**

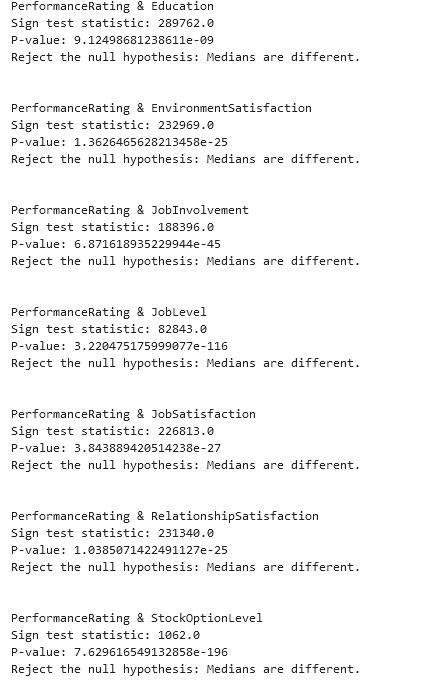
Below is a detailed table on modelling outcomes of this assignment.

**Conclusion**

**References**

**Appendix**

Figure 28: Wilcoxon signed-rank test results for JobSatisfaction

Figure 29: Wilcoxon signed-rank test results for PerformanceRating