Lab1: Exploratory data analysis

Alice Wu. Andrew ID: qichaow

1. Load dataset & check datatypes. I selected the columns with numbers.
2. Find out how many nan values. There are 100 nan values out of 2000 rows, so I decided to remove the rows with nan values, being able to conduct dimensionality reduction.
3. Outliers. Using the df.describe(), the statistics table showed that most features have outliers since the max is far from the mean. I defined the function to remove outliers. If the datum is greater than 2 std from the mean of each column, the number is treated as an outlier, and the entire row of that outlier is removed from the dataset. Since each column has different rows of outliers, I set up the std as 5, so not many rows are removed. Here, 374 rows are removed.
4. Find highly correlated variables. I used correlation matrix to find out correlations across all variables, and build the data frame df1 by removing all the highly correlated variables (>0.95). I found that 259 variables are highly correlated, and so those columns were removed. Here, I used two dataframes df (original) and df1 (remove highly correlated variables) to see the difference when doing dimensionality reduction.
5. Find if the target variable is correlated with any features. I used df.corr() to find out if the experimental property is correlated with any feature. I plotted the two features with the highest correlation in the scatter plot.

Chart, scatter chart

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

1. To conduct dimensionality reduction,

I separated features and the target

and standardized the features, and

then applied PCA. I plotted the number

number of components needed to

explain variance here. Around 49 would

be the optimal number.

1. PCA plots. To be able to color code the target property by groups, I manually divided y into ten groups, and color code by these 10 groups. I defined a plot function that color codes the principal components from dimensionality reduction.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

The PCA plot did not show a very obvious distinction of features. I think if more components are used, the distinction of groups could be shown since they might be located in higher dimensions.

1. Non-linear dimensionality reduction. t-SNE plots.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

tSNE plot removing highly correlated variables shows some distinct clusters of groups such as the purple clusters below. Here, I think tSNE performs better than PCA.