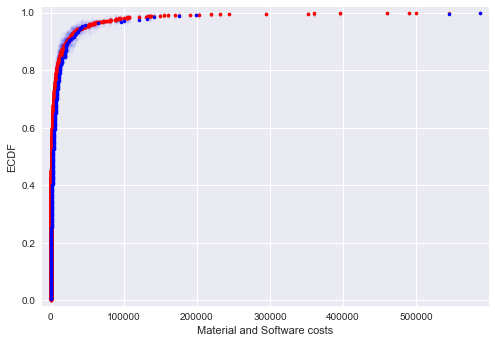
From the data analysis segment of the project, several questions are raised, and the salient features of the data set need to be evaluated. The inferential statistics analysis is needed to delve deeper into our data to explore, and assess the data prior to machine learning. The data wrangling segment of the project has created data sets of qualifying R&D expenditure across all of our clients. These data sets are dependent variables which reflect what each client has invested into the R&D projects. Independent variables from client data also need to be investigated to consider their significance in answering the overall question of the project, how can we predict R&D expenditure?

Exploring the cost brackets is important as these are the variables which R&D expenditure is dependent on, which we require to predict for the project. The cost bracket combination is the total value we are trying to predict for future clients. First the distribution of each cost bracket needs to be explored. The staff expenditure is the largest data set as described in the data story, and has the largest values of the cost brackets. We can see this by the median as described in the data story, but also by examining the distribution. As the data is exponentially distributed, a comparison between the types of expenditure at a high percentile of 80% will help to comprehend the difference in valuation. Staff expenditure is 20% likely to be valued above £82,885.81. Subcontractors at this percentile would be above £36,430.46, material £8,420.15 and software £12,220.16. Note that the types of expenditure are not as numerous as each other, and this comparison of distribution is derived from the data set we have. It is insightful still to consider the likelihood valuation of each type of cost, and would be illuminating when considering to predict expenditure valuations.

The CDF plot is most interesting in showing the similarities between the material and software costs. The software observations are not as numerous, and its distribution reveals more outliers. It would be useful to consider how different these distributions are, and what statistical significance there is between them.

Completing a permutation sample of the material and software expenditure confirms a lot of overlap within the data sets:

When calculating a P value of the difference between the software and material, we have a P test result of 0.958. This is calculating the P value positively, therefore the P value is 0.042; there is statistical significance that the distributions are the same. Having said that; although the distributions are proven to statistically very similar, the costs should not necessarily be treated the same, and more processing needs to be completed when machine learning is applied.

Measuring the correlation of the dependent total cost variable against the different cost bracket is important to assess the contribution of each cost bracket to total expenditure values. From the pearson correlation coefficients, we see that the staff costs has a value of 0.915, subcontractor costs 0.638, material 0.535 and software 0.643. Staff costs are by far the most important variable in predicting the total expenditure. From initial assumptions, it is surprising to see the pearson r being higher for the software costs than the subcontractor and material costs, only being marginally higher than the subcontractor costs.

The apportionment median has performed poorly in predicting R&D expenditure. The median of the apportionments was selected as the data is exponentially distributed. Staff apportionments have a pearson r of 0.137, and the other cost brackets have weaker still associations to the R&D expenditure. Subcontractors have an r value of 0.07, materials 0.713, software 0.038. Apportionments are a poor measure of predicting R&D expenditure. The number of entries per cost brackets performs better, although the non-staff expenditure have far weaker correlations than the number of staff members. This could be as subcontractors can work as individuals or teams, therefore there is not a set number of how many individuals are actually involved, and similarly material and software costs can be evidenced per supplier or per item. By comparison, the number of staff is a stronger correlation, with an r value of 0.62. The number of staff members involved in an R&D project is not incredibly strong, but could be a contributing predictor of R&D expenditure.

The exploration of basic statistical figures show that the expenditure is exponentially distributed; it means that the larger the expenditure the less frequent it would be. There is huge variance between the data, as well as within each industry. Although some industries reflect more expenditure on R&D on average, each industry can have particularly small and large claims. There is no set pattern that one client would be more valuable than another depending on the industry.

As there would be great variance between different industries and the nature of the work completed by our clients, it is important to investigate statistical significances between them. I have selected the 5 most common industries of our clients to explore, under the assumption that as they have the greatest data sets, and they are our most important industries, it would be more representative to specify statistical inference on these industries.

The percentile of each industry with their exponential distributions show the threshold of their top 20% values. Manufacturing is the lowest of the four, £85,613, IT at £111,001.16, Mechanical Engineering £107,659.29 and Mechanical Engineering being £106,167.39. Mechanical Engineering has a very similar percentile to Food and Drink. Looking specifically at the distribution of Mechanical Engineer and Food and Drink, the distribution seems very similar on a CDF plot.

As the Mechanical Engineering and Food industries have similar characteristics in their distribution and central tendencies, the statistical significance between their central tendencies should be investigated. As they are exponentially distributed, a logarithm is used to normalize the data. This normalization is used to apply a P statistical test on the data. First, the difference of the means of the log normalized data is calculated at 0.08, and the 95% confidence interval is calculated. The 95% confidence interval is between -0.257 and 0.436, which means that the mean is 95% most likely to be greater in the mechanical engineering set. This can be expressed as a ratio of 0.589/1 that the mean of the food industry expenditure would be greater than the mechanical engineering expenditure. Utilizing the P value statistic we can present the hypothesis that the difference of mean is lower than 0, proposing that the mean of food industry R&D expenditure is greater than of Mechanical Engineering. The P statistic calculating the difference replicates being lower than 0 (therefore the difference in the mean is greater for the food industry) with a value of 0.49 shows that there is not a statistically significant difference between the two data sets. Therefore we should be careful to assume that one industry is necessarily a better predictor than another.

Furthermore, utilizing a spearman correlation test on the industry related data shows that there is almost no correlation, with a value of -0.0209. This test shows that there is almost no correlation between the two. A disadvantage with this test is that by the data set’s nature, there are anomalies; such as the Funeral Planning and Plastic/Glass Moulding which number only one record each. However, the spearman test is effective, particularly in not needing normal distributions and in correlating within categorical data.

As this is an effective test on the correlation between categorical data and the total R&D costs, it has also been applied on other categorical data, the postcode areas, the accounting month end, and the instance that our clients have worked with us in previous projects. These too are very weak predictors. The postal code areas has a spearman value of -0.011, the accounting month end is -0.0566, and the claim phase is 0.10. Clearly, examining the R&D expenditure brackets is far more effective than the characteristics of the company. However we should not completely discount the additional data sets without putting machine learning into practice.