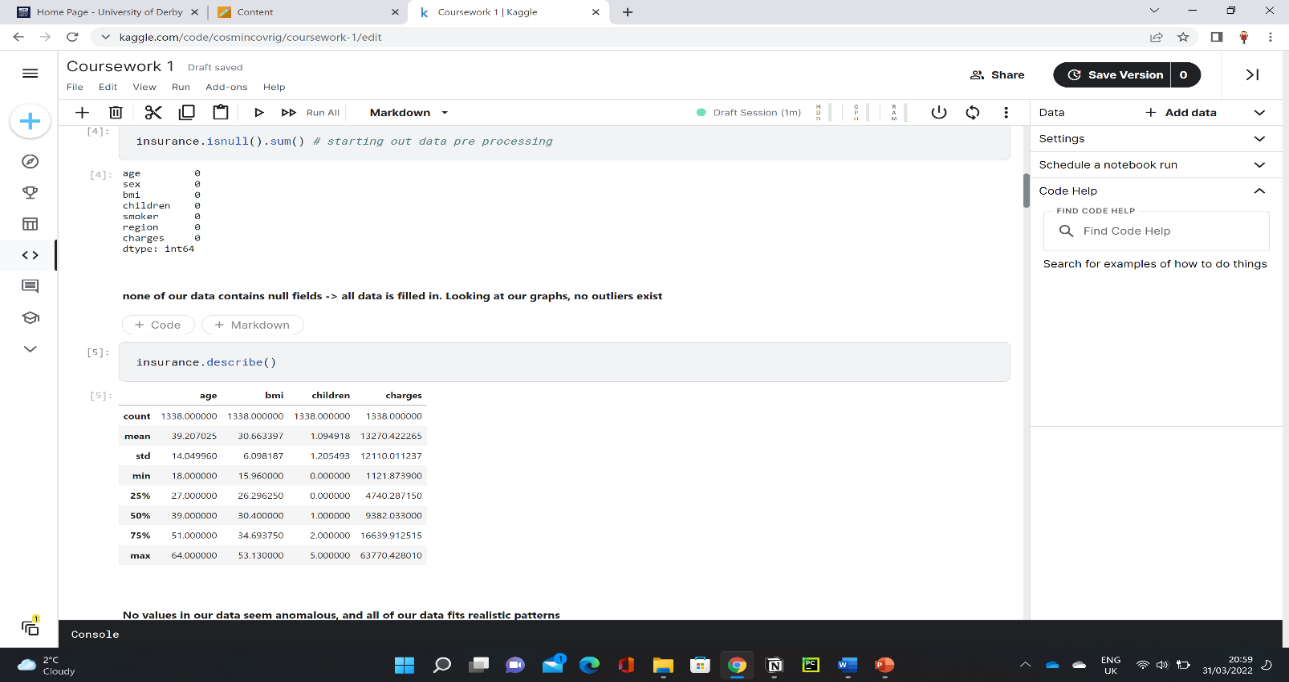
# Introduction

We were given a dataset on medical costs for a person and provided with different features that we had to analyse to determine different factors that affected our data by doing an exploratory data analysis and use various algorithms to provide an insurance cost. This report will explore these findings

# EDA Evaluation and Data Pre-Processing



Figure

By visual analysis of our data, most of the data seemed coherent with everything else, so not much pre-processing was required. However, some pre-processing was still analysed. Looking at *Figure 1,* we can determine that our data has no values of null. This means that all our data has been filled out, so no values have been left blank or missing. We also look further into our data which shows us that our max and min values for our values fit a logical standpoint, so we don’t have any values within any abnormal range.

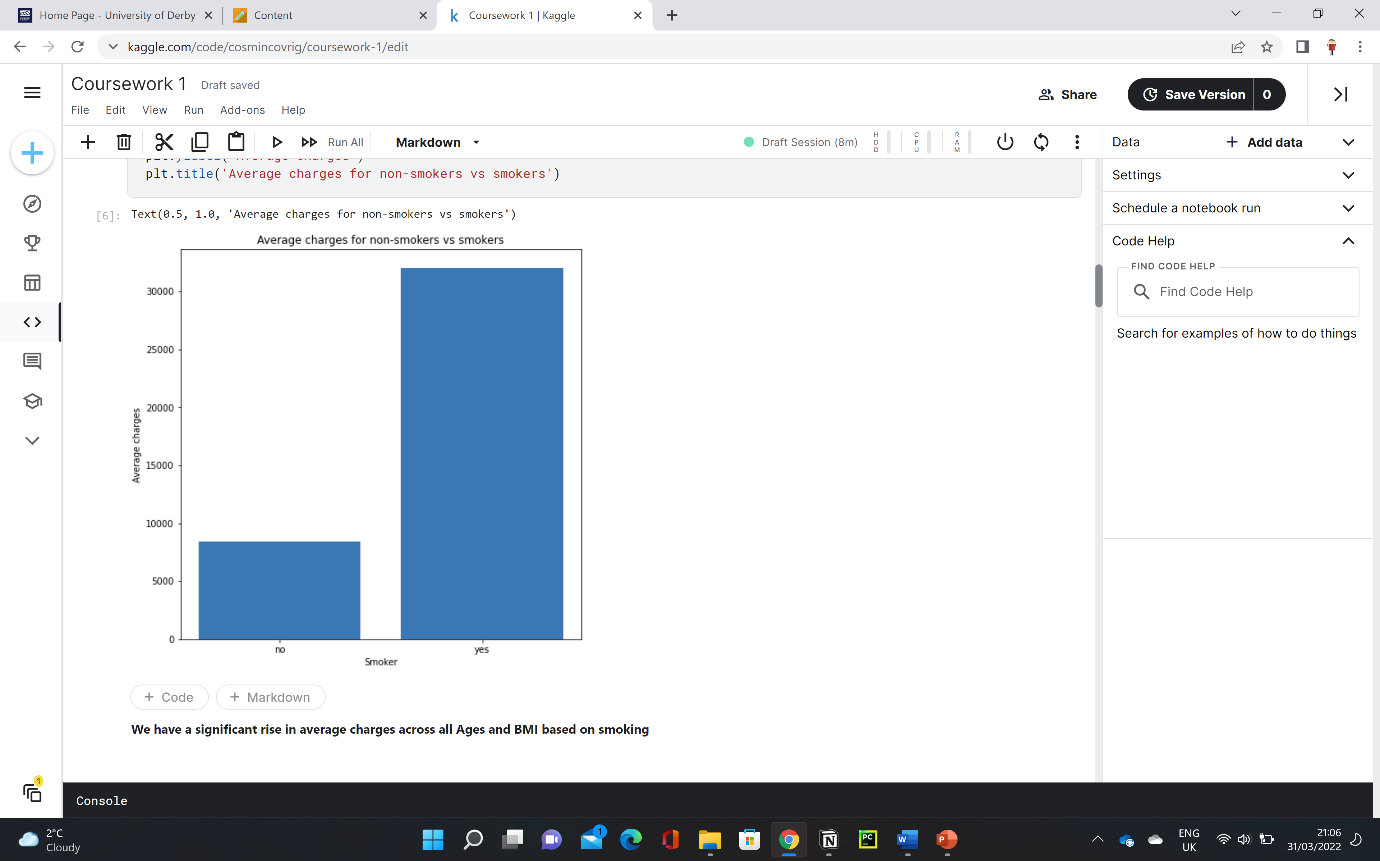


Figure 2

We first looked at the average charge across all our data within only the smoking parameter. We found that on average smokers have £23615.96 more charges than non-smokers!

This is due to the health risks associated with smoking, such as cancer, heart disease, stroke and the problems associated with the immune system (CDC, 2022), which would all need treating. This is regardless of age, BMI and other factors.

Due to such a big difference in average insurance cost, I will be splitting up most of my other data parameters based on smokers vs non-smokers.

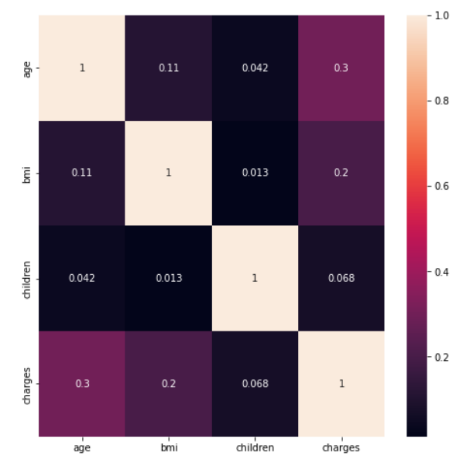


Figure 3

I firstly wanted to investigate all numerical parameters collectively and how they affect our data and its relationships. For this I used a heatmap, as this could show us a 2d matrix of across all our dimensions (GeekforGeeks, 2020). From this I concluded that most numerical values have weak correlations, with age and bmi having the biggest impact on insurance cost.

Graphical user interface, chart

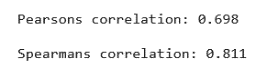
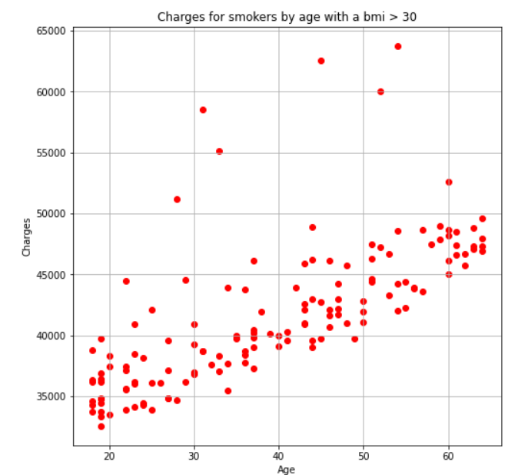
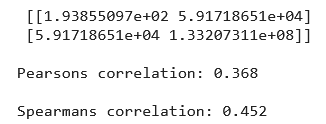
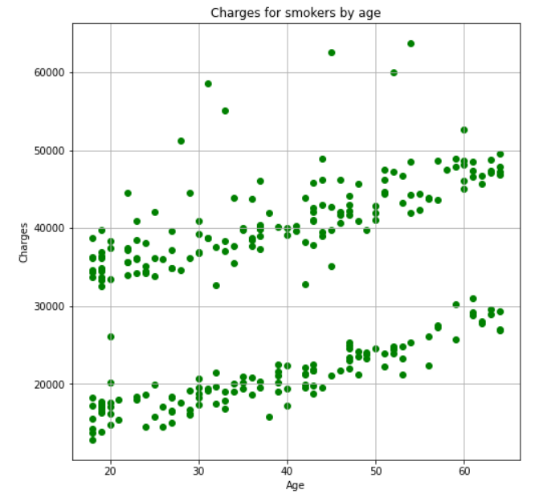
Description automatically generated

Graphical user interface, application, Word

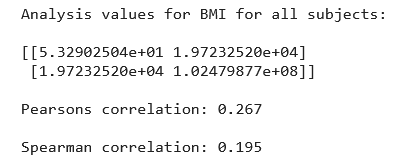
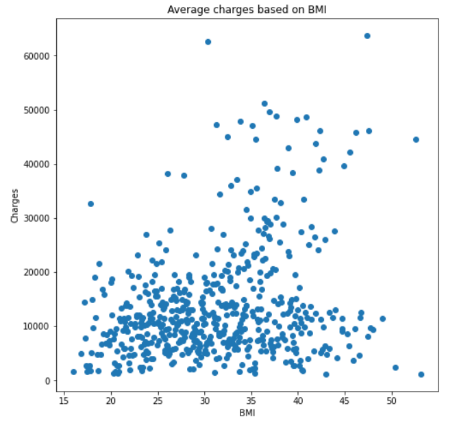
Description automatically generated

I decided to split the data for age by smoker and non-smoker and consider if a high bmi would also increase this number. There is positive correlation of 0.628 between age and insurance cost. This is likely due to requiring more medical care as you get older (Townsend, 2022).

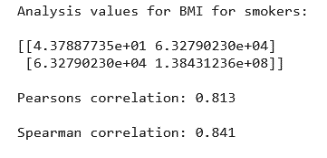
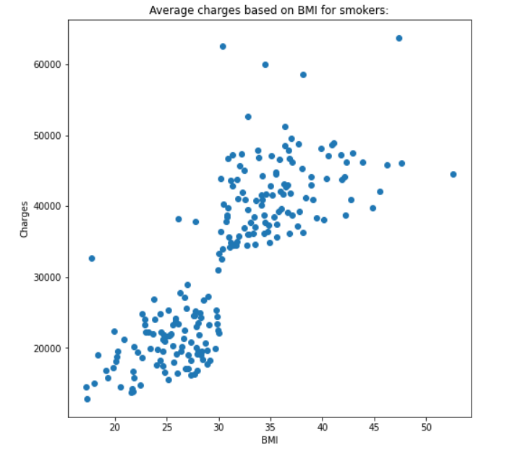
However, there is a weak increase in insurance cost for those with a high bmi. Average difference in insurance cost is only about £420 more for a person of bmi over 30. From these I can conclude that age affects insurance cost but bmi has a weak impact within the rise of costs. Some outliers within data could be interpreted as health conditions or injuries that require treatment, thus a significant rise in insurance charges.



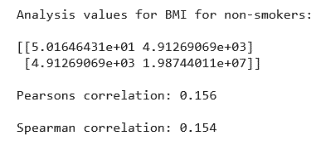
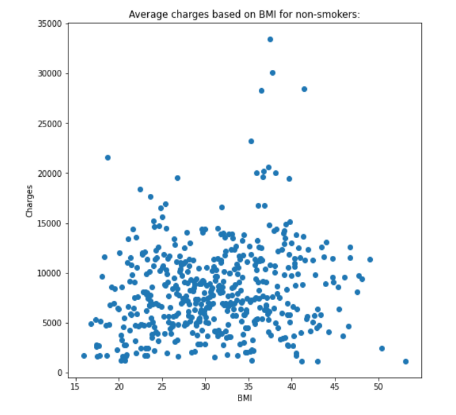
There is also a positive correlation within the smoker group for age. However, there are two distinct correlations, and these are divided by a high vs low BMI. For this reason, our Pearsons correlation is only 0.368. Smokers on average have £9642.58 increase in insurance cost with a BMI over 30. There is also a stronger positive correlation hence our correlation values are 0.698 and 0.811. This could be explained from studies showing that long term smoking increases chances of obesity and lowers blood sugar, which makes the body crave sugar and calories (Hogan, 2022). Some of the outliers could be predicted by other health conditions in addition to smoking.



Our final feature of insurance cost is BMI. I used the average of BMI values, so I could get a range across all insurance charges, and have a less dense scatter graph. This parameter brought interesting results upon further analysis. Firstly, the insurance cost is weakly positively correlated with our BMI. Our Pearsons correlation is only 0.267 with regards to some outliers too. This is due to BMI not having a direct link to health issues e.g., high muscle mass (Schimdt and Michelson, 2021). However, having a high BMI and fat percentage could increase your risk of diseases (NIH, 2022) hence the small relationship.



The charges for smokers showed a strong positive correlation with a coefficient of 0.813. After an average BMI bigger than 30, we have a big increase in charges. This proves that most of the correlation in BMI from our previous graph is led on by smokers. This relationship could be proven by a general higher health risk with smoking with a BMI over 30, and a bigger chance of higher insurance costs.



For non-smokers, there is generally a much weaker correlation with Pearsons correlation only being 0.156

These results show us that BMI will only have a major link to insurance cost if you are smoker.

Chart, bar chart, histogram

Description automatically generated

The last parameter we tested is number of children. There is very weak positive correlation with having **0 ≤ x ≤ 3 children. However, the correlation decreases afterwards. This could be explained by a much lower frequency with > 3 children.**

Critical Analysis on ML Algorithms

I used three different algorithms to make predictions on insurance costs. These were my results.

## Random Forest

Text

Description automatically generatedI attempted to use regression with 1000 decision trees as I thought it’d produce the most consistent results. I calculated the mean absolute error to find how far off the data it was.

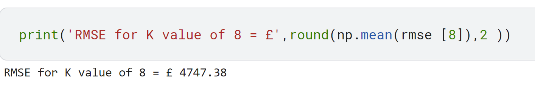
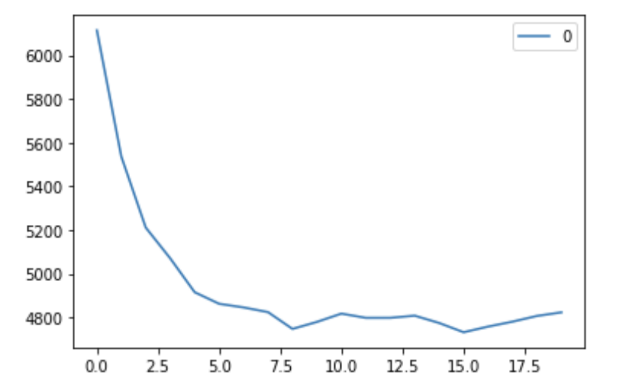
Our average absolute error was £2616.60. This seemed like a decent prediction error. However, upon analysing I found using 900 decision trees produced a slightly lower error of £2613.53 and used less memory so I adapted my model to this instead. I also took into consideration that our data didn’t have a direct linear regression since our values differed so much for smokers and non-smokers.

## KNN Algorithm

This algorithm uses its nearest neighbours and n-dimensional space to calculate predictions. I decided to firstly create an elbow curve so I could find the optimal k value. I created a for loop to find all the different values and by looking at my elbow curve and values produced, k was most optimal at 8.

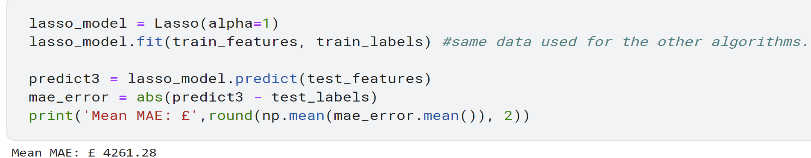
Graphical user interface, text, application, email

Description automatically generated



I analysed my data using the root mean square error since this would tell me how concentrated my predicted data was compared to my actual data. My value was £4747.38

## Lasso Regression



This algorithm uses shrinkage towards a mean value that it predicts its values from. It deals very well with overfitting (Tiwari, 2020). Using a value of alpha = 1.0, our average mean squared error was £4261.28

# Conclusion

Our data correlation was mostly reflected by the insurance cost for smokers. This provided us with the biggest correlation. Age had a moderate positive correlation too on our data too, followed by BMI. However, the biggest impact was only for those who smoked. The most suited algorithm for the prediction was random forest as it gave us the most accurate predictions in our dataset and the least average error.

# References

1. CDC. 2022. *Smoking & Tobacco Use*. [online] Available at: <https://www.cdc.gov/tobacco/basic\_information/health\_effects/index.htm#:~:text=Smoking%20causes%20cancer%2C%20heart%20disease,immune%20system%2C%20including%20rheumatoid%20arthritis.> [Accessed 31 March 2022].
2. Townsend, R., 2022. *How Age Affects Health Insurance Costs*. [online] ValuePenguin. Available at: <https://www.valuepenguin.com/how-age-affects-health-insurance-costs> [Accessed 31 March 2022].
3. Hogan, M., 2022. *How does smoking affect your weight?*. [online] spunout. Available at: <https://spunout.ie/health/smoking/how-does-smoking-affect-weight> [Accessed 31 March 2022].
4. GeeksforGeeks. 2022. *How to create a seaborn correlation heatmap in Python? - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/how-to-create-a-seaborn-correlation-heatmap-in-python/> [Accessed 31 March 2022].
5. Schimdt, A. and Michelson, A., 2021. *7 signs you don't have overweight or obesity, regardless of BMI*. [online] Insider. Available at: <https://www.insider.com/bmi-accuracy-signs-youre-not-overweight-2018-11#your-muscle-mass-makes-up-most-of-your-weight-4> [Accessed 31 March 2022].
6. Nhlbi.nih.gov. 2022. *Assessing Your Weight and Health Risk*. [online] Available at: <https://www.nhlbi.nih.gov/health/educational/lose\_wt/risk.htm> [Accessed 31 March 2022].
7. Tiwari, H., 2020. *“I do address Over-fitting problem” says Lasso and Ridge Regression.*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/i-do-address-over-fitting-problem-says-lasso-and-ridge-regression-638e30cbce24> [Accessed 1 April 2022].