Parsh Gandhi

21th July 2022

ATDP secondary division

Data Science Lab with R & Python, Section 1

Reflection #4

A dataset usually consists of at minimum 2 variables. These are the x, or independent, variable and the y, or dependent, variable. In data we try to form some sort of a relationship between these two variables. We can do this by using linear regression to make a regression line, or line of best fit. This line of best fit is in the form y = mx + b where b is the y intercept of the line and m is the slope of the line. Linear regression analysis also includes finding the errors between the predictions(fitted values) and the residuals(prediction errors). This error can be found subtracting the predicted values from the line of best fit and the actual errors. We can also perform linear regression with multiple variables by simply adding them and the residual error value to the equation like this: $y = b + ax + bx + \ldots + m_1x_1 + e$. Linear regression models are important in data analyses as they can inform us of the relationship between the variables, such as the relationship being negative, positive, or even having no relationship. Another way linear regression is helpful is that it can be used to predict future values in the model. This can be done by plugging in new values to the best fit equation, as this equation is determined using already existing data. The current pace of the class is good given the time constraints that are there to finish all the material in less than a week now although I would like to see more of the conceptual mixed in with the technical side of things. I also have a question on what types of machine learning models are unsupervised? What's the difference between an unsupervised and

supervised machine learning model? Also what is the most quick and effective way of cleaning data to make a dataset to suit your research?