DATA ANALYSIS and MODELLING

It is important to deploy descriptive and graphical statistics to look for potential problems, patterns, classifications, correlations and comparisons in the dataset. In addition, data categorization (i.e. qualitative vs quantitative) is also important to understand and select the correct hypothesis test or data model. Below is a descriptive statistics table of the data.

:

	AGE	DAYS OF EX	conc (ng/ul)	E260/280	E260/230	sec conc (ng/ul)	sec E260/280	sec E260/230	ENV COND_B	ENV COND_NE	ENV COND_OV	ENV COND_S
count	44.000000	44.00000	44.000000	44.000000	44.000000	44.000000	44.000000	44.000000	44.000000	44.000000	44.000000	44.000000
mean	33.068182	17.50000	138.977273	0.911818	0.492273	187.590909	0.965000	0.451364	0.250000	0.250000	0.181818	0.318182
std	18.382749	7.17716	236.897435	0.741385	0.379129	174.140056	0.429025	0.317253	0.438019	0.438019	0.390154	0.471155
min	7.000000	4.00000	-21.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	17.750000	15.75000	0.000000	0.000000	0.000000	67.500000	1.097500	0.195000	0.000000	0.000000	0.000000	0.000000
50%	33.000000	18.00000	17.500000	1.110000	0.625000	163.500000	1.120000	0.480000	0.000000	0.000000	0.000000	0.000000
75%	48.250000	21.00000	145.500000	1.300000	0.797500	254.250000	1.150000	0.780000	0.250000	0.250000	0.000000	1.000000
max	72.000000	29.00000	867.000000	3.110000	1.010000	774.000000	1.370000	0.890000	1.000000	1.000000	1.000000	1.000000

MEAN, MODE, MEDIAN AND SKEWNESS

Let us go ahead and explain/describe the above table:

Taking Age for example:
mean Age = 33.15
median Age = 33
mode Age = 35

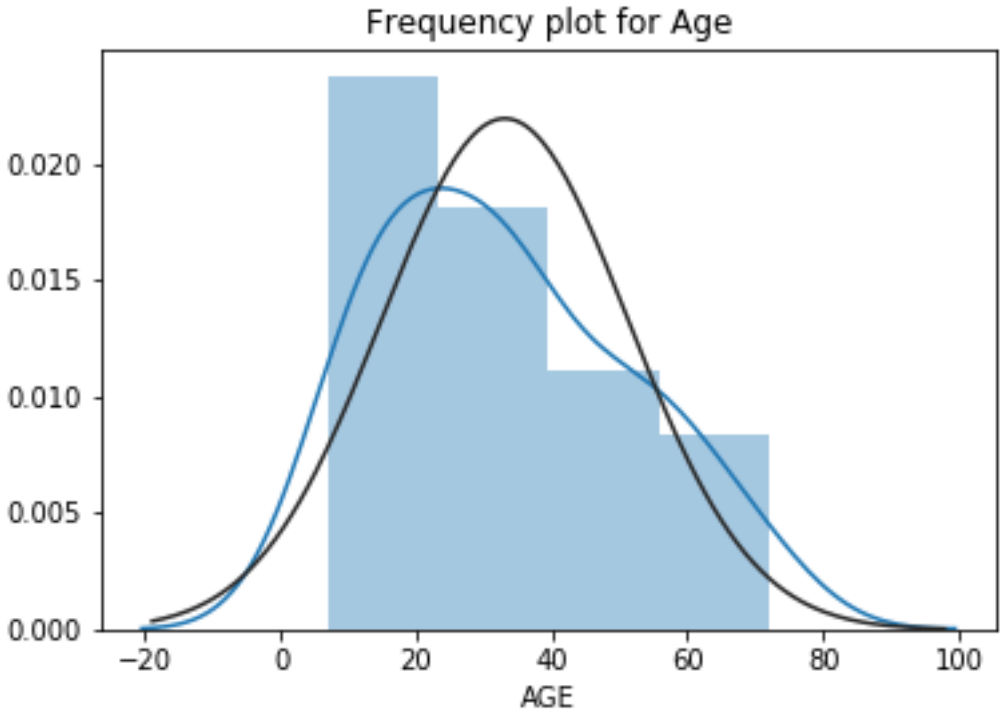
1. Count means the number of observation or rows in that column that are not empty

2. The mean, mode, median, do a nice job in telling where the center of the data set is, but often we are interested in more. For example, a pharmaceutical engineer develops a new drug that regulates iron in the blood. Suppose she finds out that the average sugar content after taking the medication is the optimal level. This does not mean that the drug is effective. There is a possibility that half of the patients have dangerously low sugar content while the other half have dangerously high content. Instead of the drug being an effective regulator, it is a deadly poison. What the pharmacist needs is a measure of how far the data is spread apart. This is what the variance and standard deviation do. First we show the formulas for these measurements. Then we will go through the steps on how to use the formulas.

- From the above data the mean of the Age is 33, mode is 35 and median is 33. This means that the centre of your data as far as age is concerned is around 33 age group. Mode also implies that most of your data population is around the age group of 33years.
- Seeing that the mean, median and mode are quite very close to each other, its an indication that there are no outliers in your dataset.

3. Skewness is a measure of Assymetry and it indicates whether a data is concentrated on one side. The rule states that is the median of a particular feature or column is greater than the mean then that distribution is left or negative skewed and this means that it has some outliers(extremely large values capable of exagerating or negatively affecting your overall computations/analysis) to the left else when mean is greater than median, it is right or positively skewed. But when median=mean=mode we can say that the distribution is zero skewed or is a normal distribution(two tailed) and has no outliers. Hence, in the case of Age column, our data is slightly tailed to the right as mean is slightly higher than median.Its skewness value reads 0.4(slightly positively skewed) hence this can be avoided or estimated as a normal distribution. Note that as skewness approaches zero, it is said to approach a normal distribution.

so, you can



notice from the above that our distribution is skewed or tailed at(from) age 55 upward hence Age 55 upward can be considered as outliers if we must make our distribution normal or zero skewed. But since we have a small dataset and we do not want to loose information, we could not be dropping age rows from 55 upward.

	Skewness
AGE	0.467603051988398
DAYS OF EX	-0.385112154253831
conc (ng/ul)	1.95294339202916
E260/280	0.397021142979223

From the above skewness table, it turns out some are negative and some are positively skewed(1 tailed) but we wont try to make it normal because we would have to drop some observations for that to be possible. Hence, We would be using them that way since they are all close to zero.

VARIATION AND COVARIANCES

Variance simply measures the dispersion of a set of datapoints around their mean values.It is more like the calculation of distances between 2 or more datapoints.

formula for sample variance comes in here.

From the above formular, the closer a number is to the mean, the lower the result we would obtain and the farther a result is to the mean, the larger the result we would obtain.i.e, the larger the variance.

A small variance indicates that the data points tend to be very close to the mean, and to each other. A high variance indicates that the data points are very spread out from the mean, and from one another. Variance is the average of the squared distances from each point to the mean.

AGE	330.245351239669
DAYS OF EX	50.3409090909091
conc (ng/ul)	54844.9313016529
E260/280	0.537160330578513
E260/230	0.140472107438017

From our variance table above, we can see that the 2 concentration features, Days of exposure and Age have incredibly high variances, which means those features has possible noise/outliers that may increase the risk of making wrong statistical inferences and predictions. To deal with this, we would do a mathematical computation on these features called 'standardization or scalling'. This is done by simply dividing each affected column by their highest value, this shrinks the values between 0 and 1 without loosing any information thereby bringing more closer each observation to the mean. See the new table below after applying the above computation.

AGE	0.063704735964442
DAYS OF EX	0.059858393687169
conc (ng/ul)	0.072962264050229
E260/280	0.537160330578513
E260/230	0.140472107438017

Now you can see that all variances are lower or less than 1 and more closer to zero. This has successfully brought observations closer to the mean thereby reducing risks of wrong inferences.

STANDARD DEVIATION

Standard deviation is simply the square root of variance .In fact it same as variance and its used when variance values are very high.Standard deviation is also the measure of the amount of variation or dispersion of a set of values from its mean.

AGE	0.25239797139526
DAYS OF EX	0.244659750852421
conc (ng/ul)	0.270115279186923
E260/280	0.732912225698625
E260/230	0.374796087810447

COVARIANCE

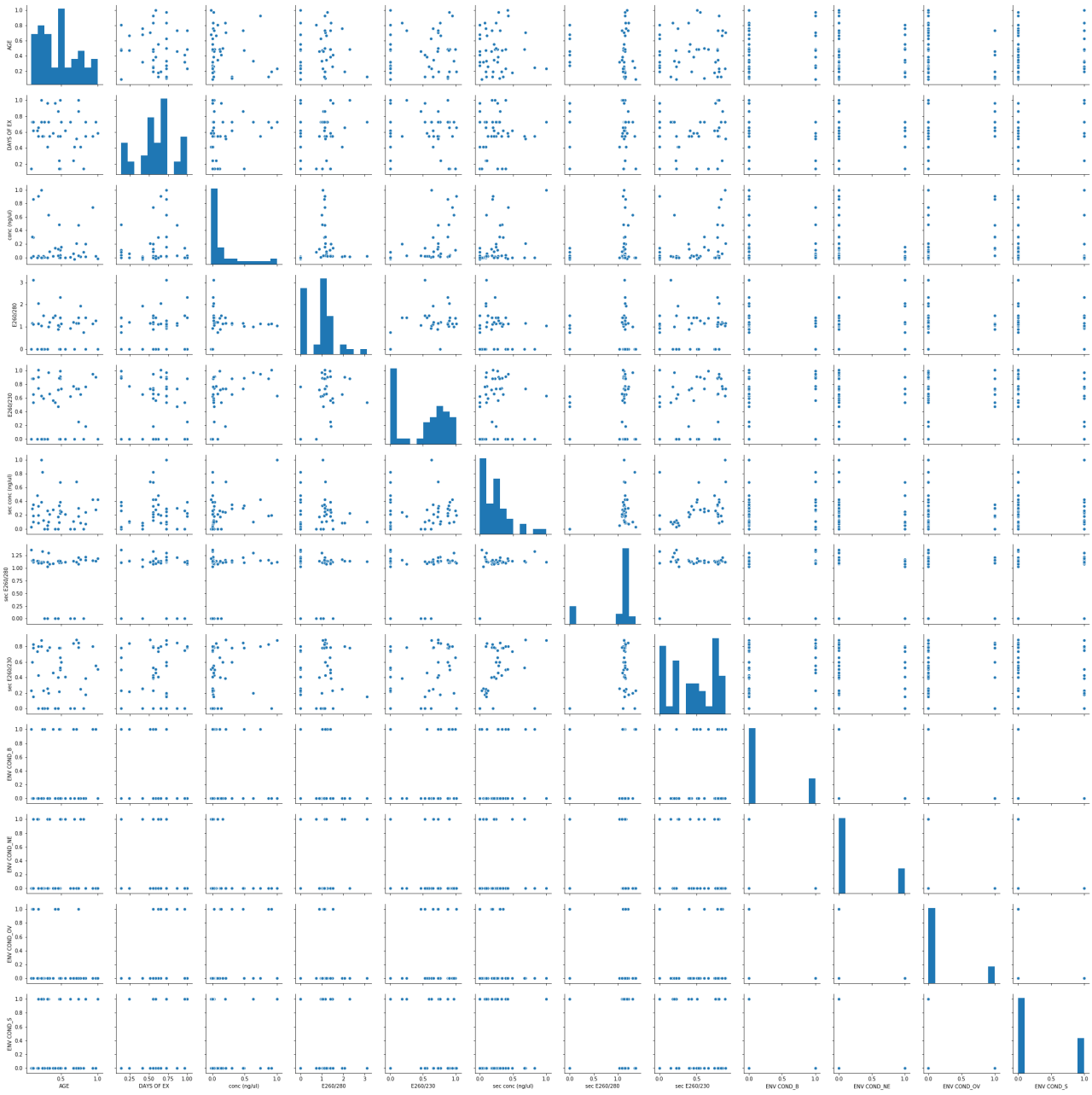
formular

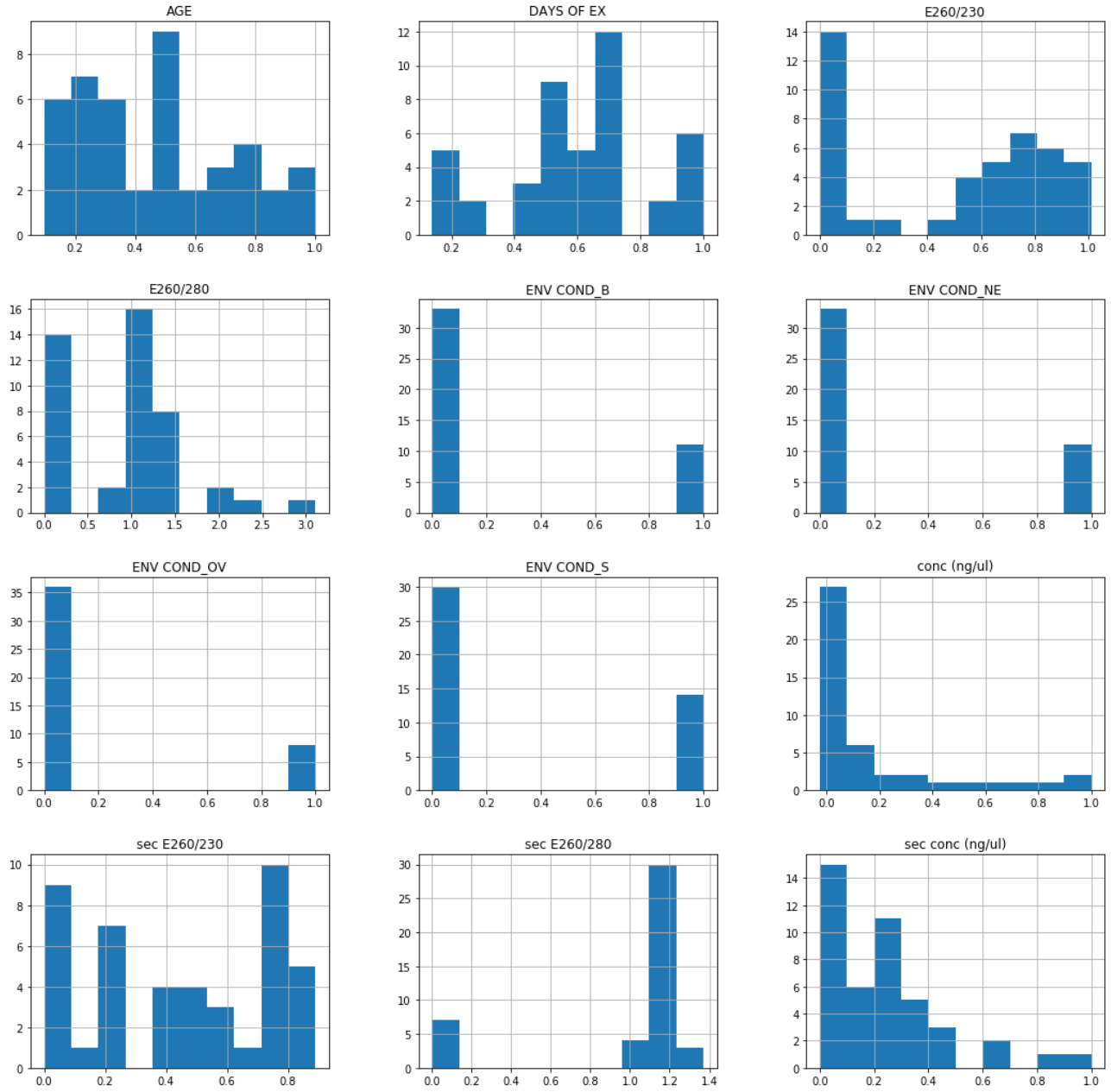
Simply explores and measures the relationship between variables. We can say that 2 variables/features are corelated and the main statistic to measure this corelation is called covariance Unlike variance, covariance could be positive, equal to zero or negative. Covariance gives a sense of direction

- greater than 0 the 2 variables moves together i.e positive covariance
- less than 0 the 2 variables moves in opposite direction, i.e negative covariance
- equal 0 the 2 variables have no relationship, i.e they are independent

One problem with covariance is that it can take on large amounts of values up to millions hence the correlation coeeficient comes in handy .The corelation coefficient has values between 0 and 1.Much better!.

Before we go further,lets take a look at a general scatterplot and histogram plot for all our variables


AGE
DAYS OF EX
conc (ng/ul)
E260/280
E260/230
sec conc (ng/ul)
sec E260/280
sec E260/230
ENV COND_B
ENV COND_NE
ENV COND_OV
ENV COND_S



Correlation

Correlation coefficient is just an adjustment of covariance so that the relationship between the two variables becomes easy and intuitive to interprete. Note that correlation coefficient is between 0 and 1 and could be negative as well. Given by: Formular

Let's take a look at a correlation. I'll then focus on the top most strongly correlated features with the target feature. Note: when the variables are not normally distributed or the relationship between the variables is not linear (as is the case here), it is more appropriate to use the Spearman rank correlation method rather than the default Pearson's method.

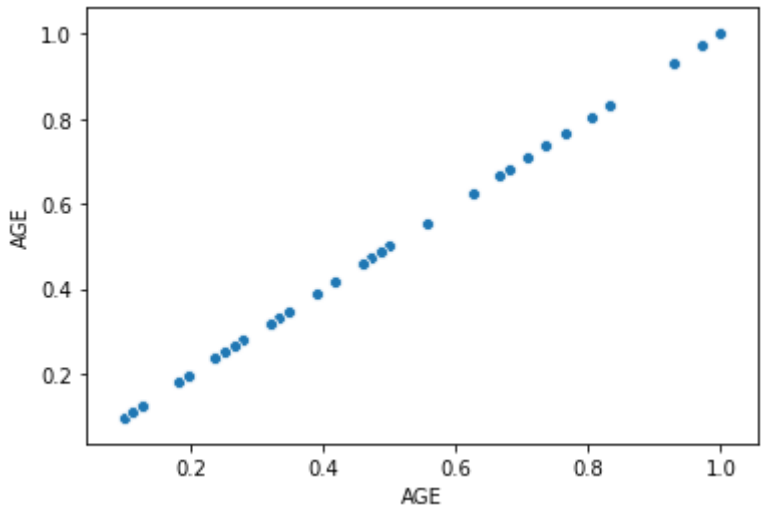
	AGE	DAYS OF EX	conc (ng/ul)	E260/280	E260/230	sec conc (ng/ul)	sec E260/280	sec E260/230	ENV COND_B	ENV COND_NE	ENV COND_OV	ENV COND_S
AGE	1.000000	-0.152515	-0.038791	0.077525	0.061434	-0.017670	0.030775	0.029374	0.132392	0.020686	-0.276366	0.086541
DAYS OF EX	-0.152515	1.000000	0.026294	0.047741	-0.021575	-0.070845	-0.128696	0.056439	-0.432220	-0.217154	0.290679	0.362999
conc (ng/ul)	-0.038791	0.026294	1.000000	0.497395	0.724497	0.211854	0.060841	0.359100	0.070706	-0.284903	0.478617	-0.197200
E260/280	0.077525	0.047741	0.497395	1.000000	0.586770	0.024266	0.008364	0.255317	0.086147	-0.054630	0.108510	-0.119155
E260/230	0.061434	-0.021575	0.724497	0.586770	1.000000	0.185647	0.217723	0.357907	0.237463	-0.277391	0.235925	-0.158245
sec conc (ng/ul)	-0.017670	-0.070845	0.211854	0.024266	0.185647	1.000000	0.338595	0.626350	0.254710	-0.169807	-0.069746	-0.021177
sec E260/280	0.030775	-0.128696	0.060841	0.008364	0.217723	0.338595	1.000000	0.245538	0.068636	-0.155991	0.000000	0.081211
sec E260/230	0.029374	0.056439	0.359100	0.255317	0.357907	0.626350	0.245538	1.000000	0.137033	-0.076821	-0.016317	-0.042465
ENV COND_B	0.132392	-0.432220	0.070706	0.086147	0.237463	0.254710	0.068636	0.137033	1.000000	-0.333333	-0.272166	-0.394405
ENV COND_NE	0.020686	-0.217154	-0.284903	-0.054630	-0.277391	-0.169807	-0.155991	-0.076821	-0.333333	1.000000	-0.272166	-0.394405
ENV COND_OV	-0.276366	0.290679	0.478617	0.108510	0.235925	-0.069746	0.000000	-0.016317	-0.272166	-0.272166	1.000000	-0.322031
ENV COND_S	0.086541	0.362999	-0.197200	-0.119155	-0.158245	-0.021177	0.081211	-0.042465	-0.394405	-0.394405	-0.322031	1.000000

Explaining the above corelation map/matrix

1. You would notice that the features within the dataset are aligned on 2 sides of the table. The intuition behind this is to depict an x-y axis plot(though it may look inverted on the above table) between 2 features.SO, consider the horizontal alignment to our x-axis and the vertical to be the y-axis.

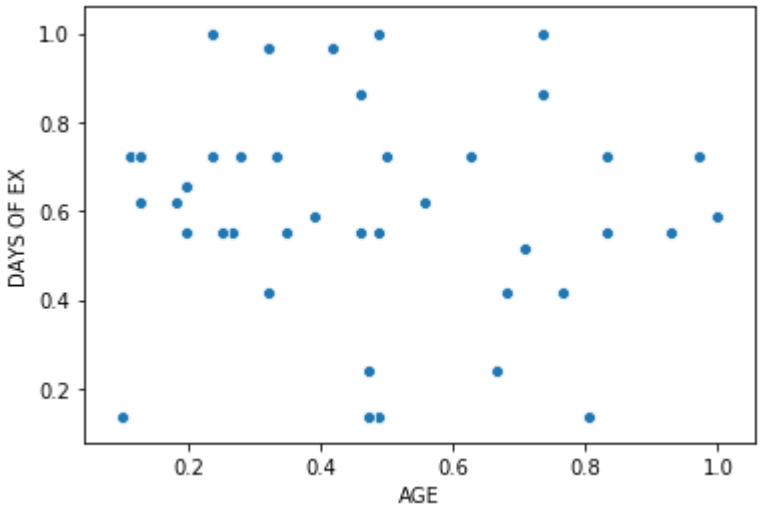
2. Having the above in mind, let us start from the top to read it. The first is 1.0000 which is depicts the x-y
plot of AGE against AGE .Notice?? Yes now you see the reason why it gave 1.0,because it is plotted against itself hence a perfect linear relationship.Note that 1.000 depicts a perfect corelation or relationship strength between 2 features or variables. Note that everytime a feature or variable is compared with itself, it gives a perfect correlation co-efficient of 1.0.

#Scatterplot of AGE against itself showing a perfect linear relationship



3. Next(going down vertically) is AGE against DAYS OF EX. This gave us a correlation coefficient of -0.124371.This depicts a negative correlation but quite small.Negative here means that as one variable goes up ,the other comes down. This could also mean that DOE has little linear importance to predicting or computing AGE.But even though there may not be an existing strong linear relationship, but then, there might exist anoda form of relationship between them. lets take a look at the plot below

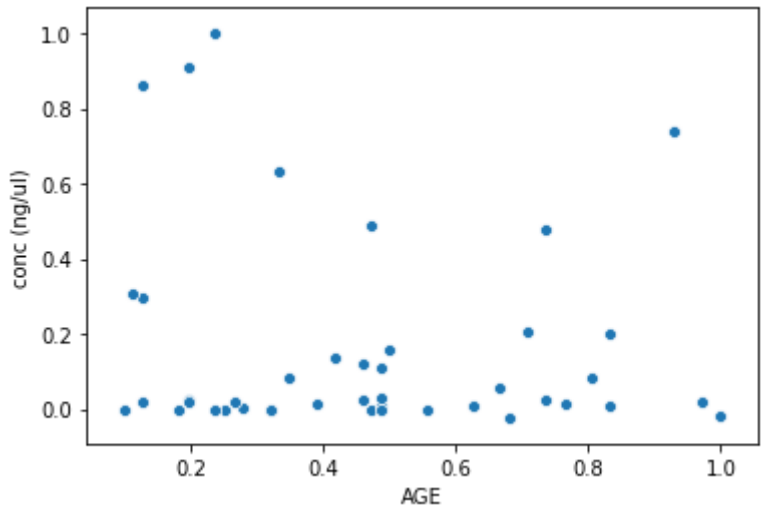
#Scatterplot of AGE against DAYS OF EX



Notice from the above plot that there is little or no linear relationship between them which explains why the low correlation coefficience of -0.124371. A cluster-based relationship may rather exist.

4. Next(as we move down continually) is AGE against conc(ng/ul). -0.038791. This is low of course but then, there might exist other possible non-linear relationships between features. Let us take a look at the actual plot below.

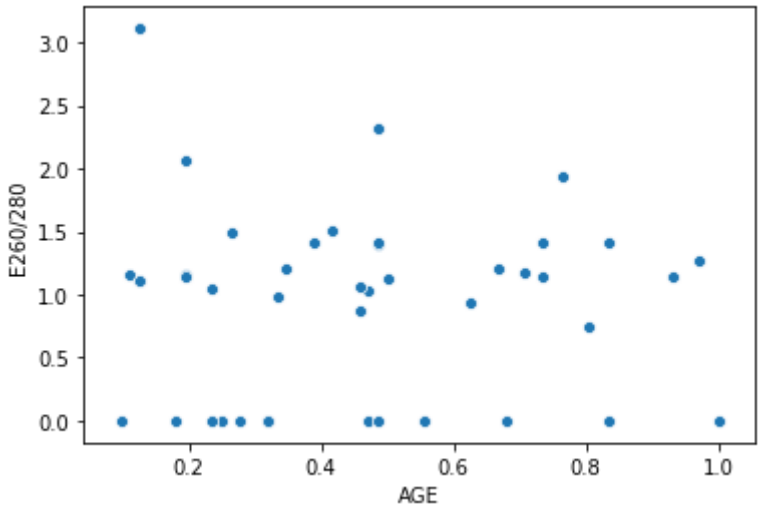
Scatterplot of AGE against conc(ng/ul)



Again from the above plot we see that there is little or no linearity between the 2 variables/features hence the low correlation coefficient

5. Next is AGE against E260/280 which gives us 0.077525. See plot below

#Scatterplot of AGE against E260/280

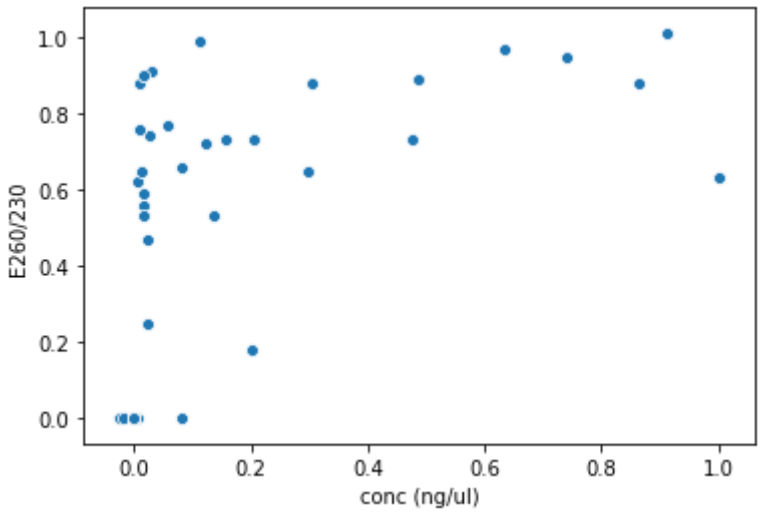


And yet again we can see little or no linearity between the two variables above.

The correlations goes on and on throughout the correlation map/table. Now lets go ahead and view some very high correlated features.Say for e.g

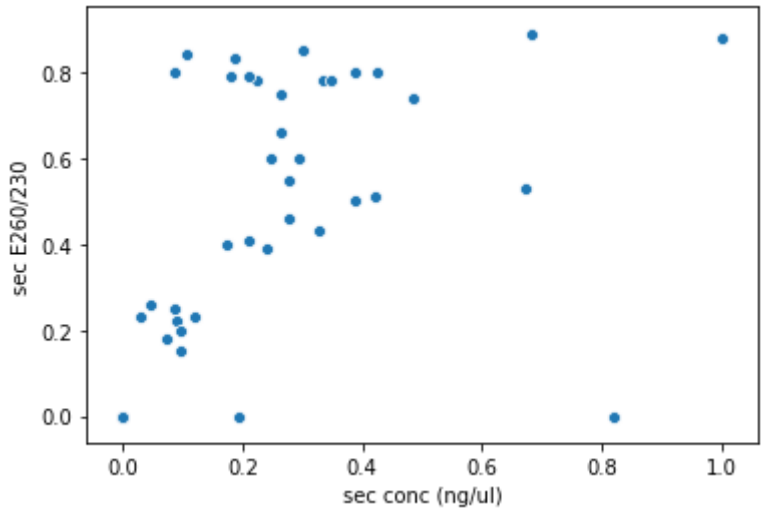
- conc vs E260/230 = 0.724497,
- sec conc vs sec E260/230=0.626350.
- E260/280 vs E260/230 =0.586770 ,
- secE260/280 vs secE260/230 =0.245538

#Scatterplot of concentrations against E260/230



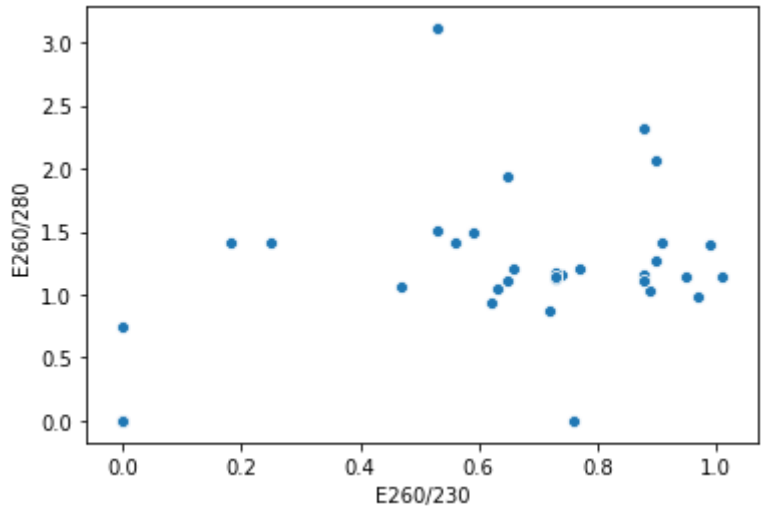
From the above you can notice some high level linearity at the top of the graph/plot which explains their correlation coefficient of 0.7244

#Scatterplot of concentrations against sec E260/230



We can as well notice some good linearity here where you can draw lines of best fit. This is a positive correlation(0.626350) which indicates that there is a possibility that the increase in one variable would cause an increase as well in the other.

#Scatterplot of E260/230 against E260/280

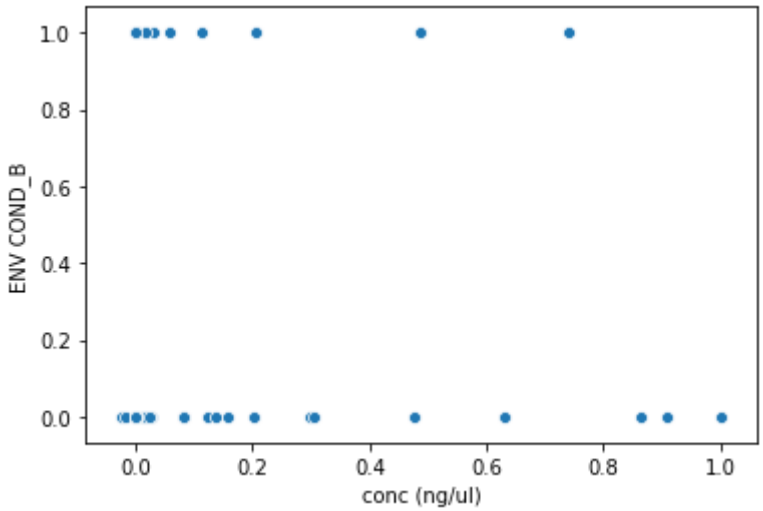


The above plot shows a correlation coefficient of 0.58

Basically, if you compare the different correlation coefficient values for each concentration level,you would discover that the first conc readings is more correlative and significant to the AGE than the second conc readings. Hence it can be conclusive that the first conc readings has more importance than the second.Which means that if we were to choose a set of conc reading to go with, it is the first.

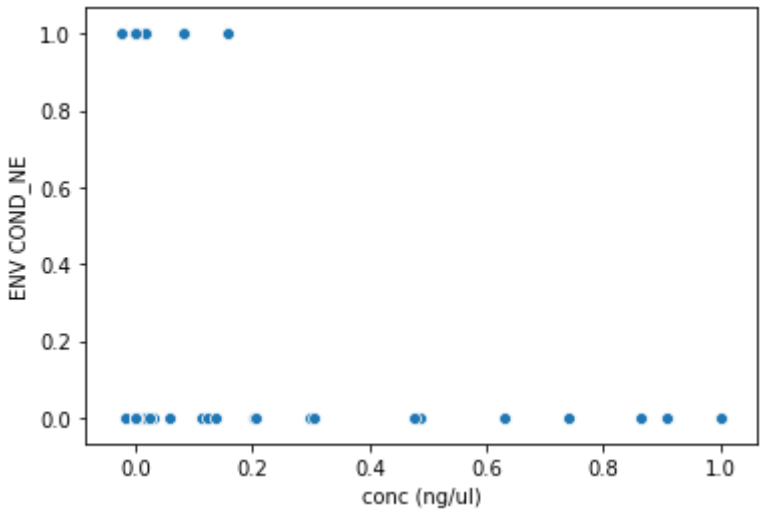
let us go ahead and examine the correlations between the two concs and environmental conditions.

#Correlation between conc and ENV_COND_B



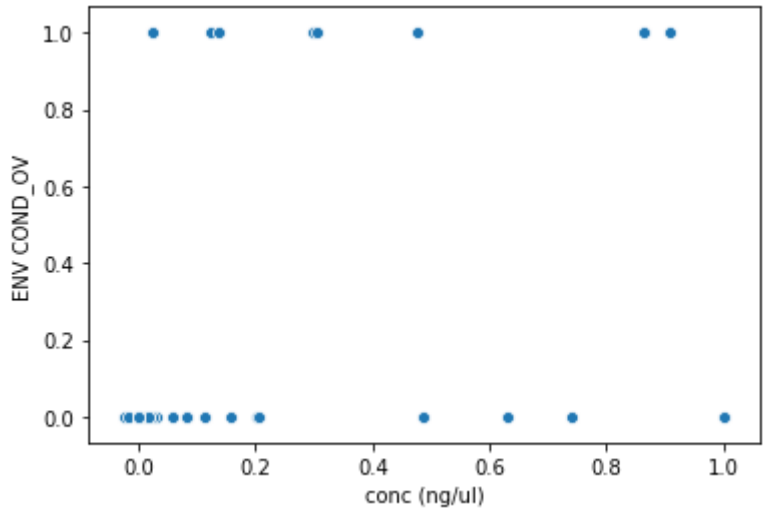
The reason why the above plot is appearing that way is because one variable is categorical(S,B,NE,OV) in nature hence it tends to form a kind of cluster or seperation(s) when plotted against another feature/variable.Nonetheless, such features can also have high correlations or importance in the dataset. In this case, we see a correlation coefficient of as low as 0.070706. It can be infered that this particular environmental condition does not have much impact on the level of concentration.

#Correlation between conc and ENV_COND_NE



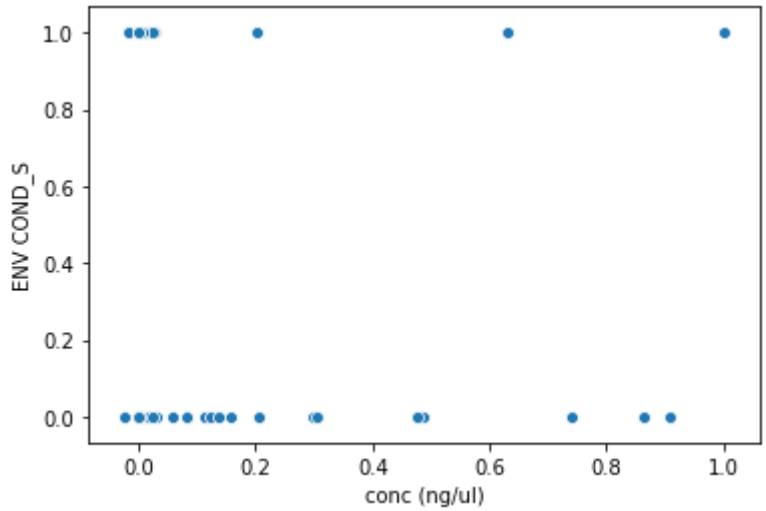
Just as the same case as the previous environmental condition above, just a slight improvement this time as we have Negative correlation coefficient of -0.284903. This is negative linearity which means that the a slight reduction in one variable might lead to some increase in the other variable. We can as well infer that a slight reduction or increase in this particular environmental condition might lead to a little alteration in the concentration.

#Correlation between conc and ENV_COND_OV



This particular environmental condition seems to be kind of perculiar as shows quite a high positive correlation value of 0.478617 with the organic conc values. Which means they both have a strong dependence on each other hence it can be interpreted that a slight increase in one might lead to another increase in the other variable.

#Correlation between conc and ENV_COND_S



this one has a negative correlation of -0.197200 which is also worthy of note. Basically, environmental conditions seems to have some dependence and relationship with the concentration values.

The following are the correlation coefficients of the KIT conc values against environmental conditions

- sec conc(ng/ul) vs ENV COND_B = 0.254710

- sec conc(ng/ul) vs ENV COND_NE = -0.169807

- sec conc(ng/ul) vs ENV COND_OV = -0.069746

- sec conc(ng/ul) vs ENV COND_S = - -0.021177

Again we see how low these coefficients for the KITS are compared to the Organic readings hence We can say that there is more impact/dependence of env conds on the organic conc than it is for the kit conc. Hence we can go with the organic conc for further experiments and inferences.

COMPUTING POISSONS VALUE(p-value) FOR EACH FEATURE

OLS Regression Results

Dep. Variable:	AGE	**R-squared:**	0.303
Model:	OLS	**Adj. R-squared:**	-0.070
Method:	Least Squares	**F-statistic:**	0.8132
Date:	Wed, 18 Dec 2019	**Prob (F-statistic):**	0.656
Time:	17:32:28	**Log-Likelihood:**	6.0989
No. Observations:	44	**AIC:**	19.80
Df Residuals:	28	**BIC:**	48.35
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.035	0.965]
const	0.2769	0.094	2.957	0.006	0.101	0.453
x1	-0.0908	0.231	-0.393	0.697	-0.526	0.345
x2	-0.1030	0.217	-0.474	0.639	-0.512	0.306
x3	-0.0127	0.088	-0.144	0.886	-0.178	0.152
x4	0.1382	0.206	0.670	0.508	-0.250	0.527
x5	0.0132	0.269	0.049	0.961	-0.494	0.521
x6	-0.1475	0.144	-1.022	0.316	-0.420	0.125
x7	0.0846	0.182	0.464	0.646	-0.259	0.428
x8	0.0133	0.099	0.134	0.894	-0.173	0.200
x9	0.0164	0.088	0.185	0.854	-0.150	0.183
x10	0.2040	0.170	1.199	0.240	-0.116	0.524
x11	0.0432	0.092	0.468	0.644	-0.131	0.217
x12	0.0042	0.107	0.039	0.969	-0.197	0.205
x13	0.2770	0.148	1.878	0.071	-0.001	0.555
x14	0.1817	0.096	1.893	0.069	0.001	0.363
x15	-0.2871	0.183	-1.566	0.129	-0.632	0.058
x16	0.1011	0.087	1.160	0.256	-0.063	0.265
x17	0.2180	0.068	3.209	0.003	0.090	0.346
x18	0.0588	0.069	0.854	0.400	-0.071	0.189

Omnibus:	2.752	**Durbin-Watson:**	1.738
Prob(Omnibus):	0.253	**Jarque-Bera (JB):**	2.352
Skew:	0.563	**Prob(JB):**	0.309
Kurtosis:	2.877	**Cond. No.**	4.15e+16

Explaining the above OLS statistical table

1. Standard error are same as standard deviation which measures the variability of datapoints from the mean i.e
how far are my datapoints/observations from the mean. Note also that standard error decreases when sample size
increases.It turns out to be that we have very low standard errors(closer to zero) which informs the abscence
outliers or noisy data that can wrongly affect our experiments.It is also a measure of the statistical accuracy of an estimate, equal to the standard deviation of the theoretical distribution of a large population of such estimates.

2. where :

- x1 = 'DAYS OF EX',
- x2 = 'conc (ng/ul)',
- x3 = 'E260/280',
- x4 = 'E260/230',
- x5 = 'sec conc (ng/ul)',
- x6 = 'sec E260/280',
- x7 = 'sec E260/230',
- x8 = 'ENV COND_B',
- x9 = 'ENV COND_NE',
- x10 = 'ENV COND_OV',
- x11 = 'ENV COND_S',
- x12 = 'TYPE OF TEETH_C',
- x13 = 'TYPE OF TEETH_IN',
- x14 = 'TYPE OF TEETH_LM',
- x15 = 'TYPE OF TEETH_PM',
- x16 = 'TYPE OF TEETH_UM',
- x17 = 'SEX_F',
- x18 = 'SEX_M'

3. Assessing model performance. Both the Multiple R-Squared and Adjusted R-Squared values are measures of model performance. Possible values range from 0.0 to 1.0. The Adjusted R-Squared value is always a bit lower than the Multiple R-Squared value because it reflects model complexity (the number of variables) as it relates to the data, and consequently is a more accurate measure of model performance. Adding an additional explanatory variable to the model will likely increase the Multiple R-Squared value, but decrease the Adjusted R-Squared value. Suppose you are creating a regression model of residential burglary (the number of residential burglaries associated with each census block is your dependent variable, y). An Adjusted R-Squared value of 0.84 would indicate that your model (your explanatory variables modeled using linear regression) explains approximately 84% of the variation in the dependent variable, or said another way: your model tells approximately 84% of the residential burglary "story".

4. Assess each explanatory variable in the model: Coefficient, Probability or Robust Probability, and Variance Inflation Factor (VIF). The coefficient for each explanatory variable reflects both the strength and type of relationship the explanatory variable has to the dependent variable(AGE). When the sign associated with the coefficient is negative, the relationship is negative (e.g., the larger the distance from the urban core, the smaller the number of residential burglaries). When the sign is positive, the relationship is positive (e.g., the larger the population, the larger the number of residential burglaries). Coefficients are given in the same units as their associated explanatory variables (a coefficient of 0.005 associated with a variable representing population counts may be interpretted as 0.005 people). The coefficient reflects the expected change in the dependent variable for every 1 unit change in the associated explanatory variable, holding all other variables constant (e.g., a 0.005 increase in residential burglary is expected for each additional person in the census block, holding all other explanatory variables constant). The T test is used to assess whether or not an explanatory variable is statistically significant. The null hypothesis is that the coefficient is, for all intents and purposes, equal to zero (and consequently is NOT helping the model). When the probability or robust probability is very small, the chance of the coefficient being essentially zero is also small. If the Koenker test (see below) is statistically significant, use the robust probabilities to assess explanatory variable statistical significance. Statistically significant probabilities have an asterisk "*" next to them. An explanatory variable associated with a statistically significant coefficient is important to the regression model if theory/common sense supports a valid relationship with the dependent variable, if the relationship being modeled is primarily linear, and if the variable is not redundant to any other explanatory variables in the model. The variance inflation factor (VIF) measures redundancy among explanatory variables. As a rule of thumb, explanatory variables associated with VIF values larger than about 7.5 should be removed (one by one) from the regression model. If, for example, you have a population variable (the number of people) and an employment variable (the number of employed persons) in your regression model, you will likely find them to be associated with large VIF values indicating that both of these variables are telling the same "story"; one of them should be removed from your model.

5. Assess model significance. Both the Joint F-Statistic and Joint Wald Statistic are measures of overall model statistical significance. The Joint F-Statistic is trustworthy only when the Koenker (BP) statistic (see below) is not statistically significant. If the Koenker (BP) statistic is significant you should consult the Joint Wald Statistic to determine overall model significance. The null hypothesis for both of these tests is that the explanatory variables in the model are not effective. For a 95% confidence level, a p-value (probability) smaller than 0.05 indicates a statistically significant model.

6. Assess Stationarity. The Koenker (BP) Statistic (Koenker's studentized Bruesch-Pagan statistic) is a test to determine if the explanatory variables in the model have a consistent relationship to the dependent variable (what you are trying to predict/understand) both in geographic space and in data

space. When the model is consistent in geographic space, the spatial processes represented by the explanatory variables behave the same everywhere in the study area (the processes are stationary). When the model is consistent in data space, the variation in the relationship between predicted values and each explanatory variable does not change with changes in explanatory variable magnitudes (there is no heteroscedasticity in the model). Suppose you want to predict crime and one of your explanatory variables in income. The model would have problematic heteroscedasticity if the predictions were more accurate for locations with small median incomes, than they were for locations with large median incomes. The null hypothesis for this test is that the model is stationary. For a 95% confidence level, a p-value (probability) smaller than 0.05 indicates statistically significant heteroscedasticity and/or non-stationarity. When results from this test are statistically significant, consult the robust coefficient standard errors and probabilities to assess the effectiveness of each explanatory variable. Regression models with statistically significant non-stationarity are especially good candidates for GWR analysis.

7. Assess model bias. The Jarque-Bera statistic indicates whether or not the residuals (the observed/ known dependent variable values minus the predicted/estimated values) are normally distributed. The null hypothesis for this test is that the residuals are normally distributed and so if you were to construct a histogram of those residuals, they would resemble the classic bell curve, or Gaussian distribution. When the p-value (probability) for this test is small (is smaller than 0.05 for a 95% confidence level, for example), the residuals are not normally distributed, indicating model misspecification (a key variable is missing from the model). Results from a misspecified OLS model are not trustworthy.

8. View the coefficient and diagnostic tables. Creating the coefficient and diagnostic tables is optional. While you are in the process of finding an effective model, you may elect not to create these tables. The model building process is iterative and you will likely try a large number of different models (different explanatory variables) until you settle on a few good ones. You can use the Aiaike Information Criterion (AIC) on the report to compare different models. The model with the smaller AIC value is the better model (that is, taking into account model complexity, the model with the smaller AIC provides a better fit to the observed data). You should always create the coefficient and diagnostic tables for your final OLS models in order to capture the most important elements of the OLS report including the list of explanatory variables used in the model with their coefficients, standard errors, and probabilities, and results for each diagnostic test. The diagnostic table includes a description of each test along with some guidelines for how to interpret test results.

MACHINE LEARNING MODELLING

:

	Algorithm	RMSE Mean	RMSE SD
0	ElasticNet	0.2575	0.0337
1	Lasso	0.2575	0.0337
2	LGBM Regressor	0.2575	0.0337
3	SVR	0.2686	0.0260
4	Random Forest	0.2752	0.0257
5	Ridge Regression	0.3007	0.0169
6	Gradient Boosting Regressor	0.3334	0.0073
7	XGBoost Regressor	0.3370	0.0077
8	Linear Regression	0.4332	0.0559

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. Formally it is defined as follows: image formlar

The RMSE been closer to zero depicts a good model.Which explains that the degree of error of that model in predicting age is low. The goal and aim of every machine learning modelling is not just to train an algorithm on a dataset but to also ensure that this error, RMSE is kept at its barest minimum. So, from the above , we notice the following:

- Having tried a simple linear regression, we discovered that it has the highest error so far of 0.4332.This is understandable because the linear regression is non-sophisticated algorithm/equation that focuses on only distinct linear relationships between the target variable AGE and the predictors and in this case where most of our variable/predictors barely have any linear relationship with Age(based on spearman correlation), we expect our model to learn or perform poorly in predicting Age from an unseen data. Another downside of the simple Linear equation is that it only considers few variables and can hardly predict from a data with multiple variables/features.

- In the case such as this where existing relationships between most of our features are non-linear, we used something in machine learning called a tree-based algorithm/equation approach which is more complex and sophisticated in extracting and learning from other non-linear relationships that exists between AGE and other features. Of course, these tree-based equations turned out to perform better as we witnessed a more reduced RMSE score. These tree-bases algorithms/equations includes.

1. LightGBM
2. RandomForest
3. Lasso Regressor
4. ElasticNet Regressor