Assignment-based Subjective Questions

QUESTION 1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Categorical variables are those variables can't be used directly in the model since they are non-numeric. One way is to create dummy/indicator variables i.e. variable would take a value of 0 or 1 when it belongs to a certain class. Let us analyze these categorical variables i.e. season, month, weekday and weathersituation.

Season – bike rentals are high in fall and least in spring

Months – Rentals are less in the winter months and highest in summer/fall months

Weekday – Mondays and Fri marks the most bike rentals

Weather situation - Rentals are not preferred during snow and rain

QUESTION 2 - Why is it important to use drop_first=True during dummy variable creation?

There is no need of defining different levels. If you drop a level, you would still be able to explain the other levels. It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

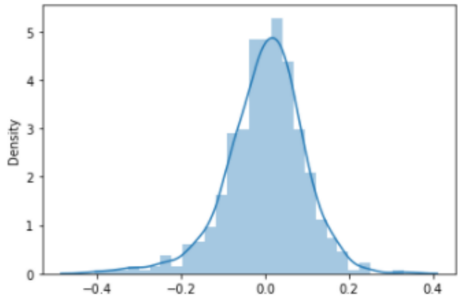
For example, If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

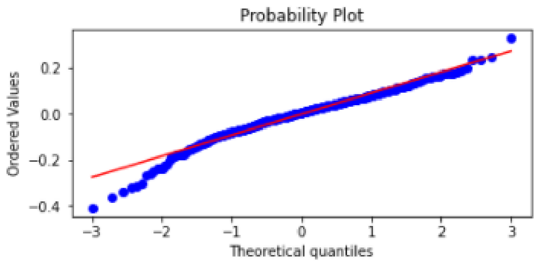
QUESTION 3 - Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temperature in Celsius variable has the highest correlation with the target variable

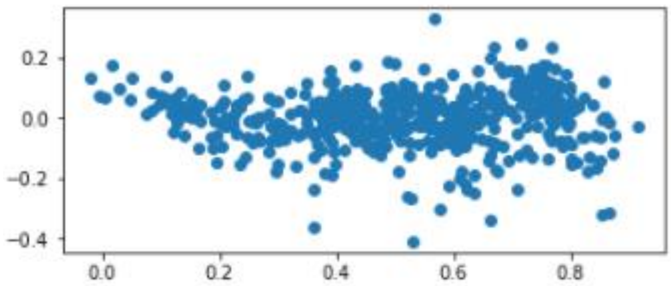
QUESTION 4 - How did you validate the assumptions of Linear Regression after building the model on the training set

- If error terms are normally distributed with mean 0

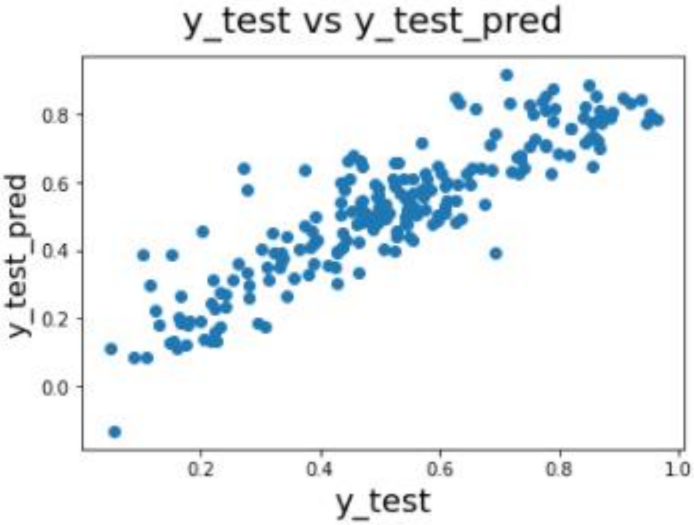




- pattern means all the residuals are scattered around y = 0. but here we are confident that model is good that there is no easily idetifiable patterns. Error values are statistically independent



- Homoscedasticity - We can observe that variance of the residuals (error terms) is constant across predictions. i.e error term does not vary much as the value of the predictor variable changes.



QUESTION 5 - Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top 3 features are

- Temperature in Celsius
- Season
- Month

General Subjective Questions

QUESTION 1 - Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. In predictive analytics, we build a model using training data, and later we use the model parameters to predict for test data i.e. using the training data we learn the model and then we use the model to give label to the test data. In a regression model, we first identify the variables and see if there is a relationship between them. We have simple linear and multiple linear regression with one independent and multiple independent variables respectively.

__The steps/structure of building a linear regression model are__

1. Reading, understand and visualising the data – knowing the target and predictor variables. Also Visualising will help in interpreting the data well, to get an intuitive understanding of the distributions and identifying the variables that can turn out to be useful in building the model.

2. Prepare the data for modelling(train-test split, rescaling etc)

DATA PREPARATION STEPS are

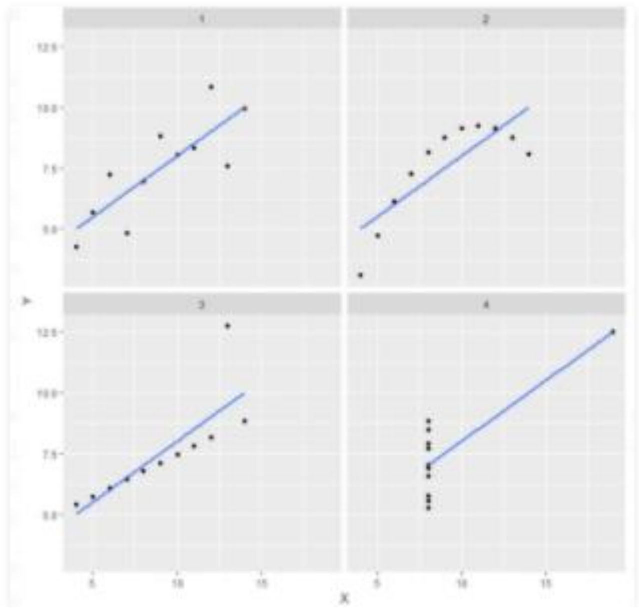
- __Encoding__
    - Converting binary vars to 1/0
    - other categorical vars to dummy vars
- __Splitting into train and test__
- __Rescaling of variables__

3. Training the model

4. Residual analysis - The fundamental assumption in linear reg is that residulas should be normally distributed. This is the step where we validate the assumption

5. Predictions and evaluation on the test set - in this step, we make sure R2 on test is around the same value as R2 on training set

QUESTION 2 - Explain the Anscombe's quartet in detail.



Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. This demonstrates both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. This was found by statistician Francis Anscombe.

He took 4 data sets of 11 data points and analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y. That is, four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

QUESTION 3 - What is Pearson’s R?

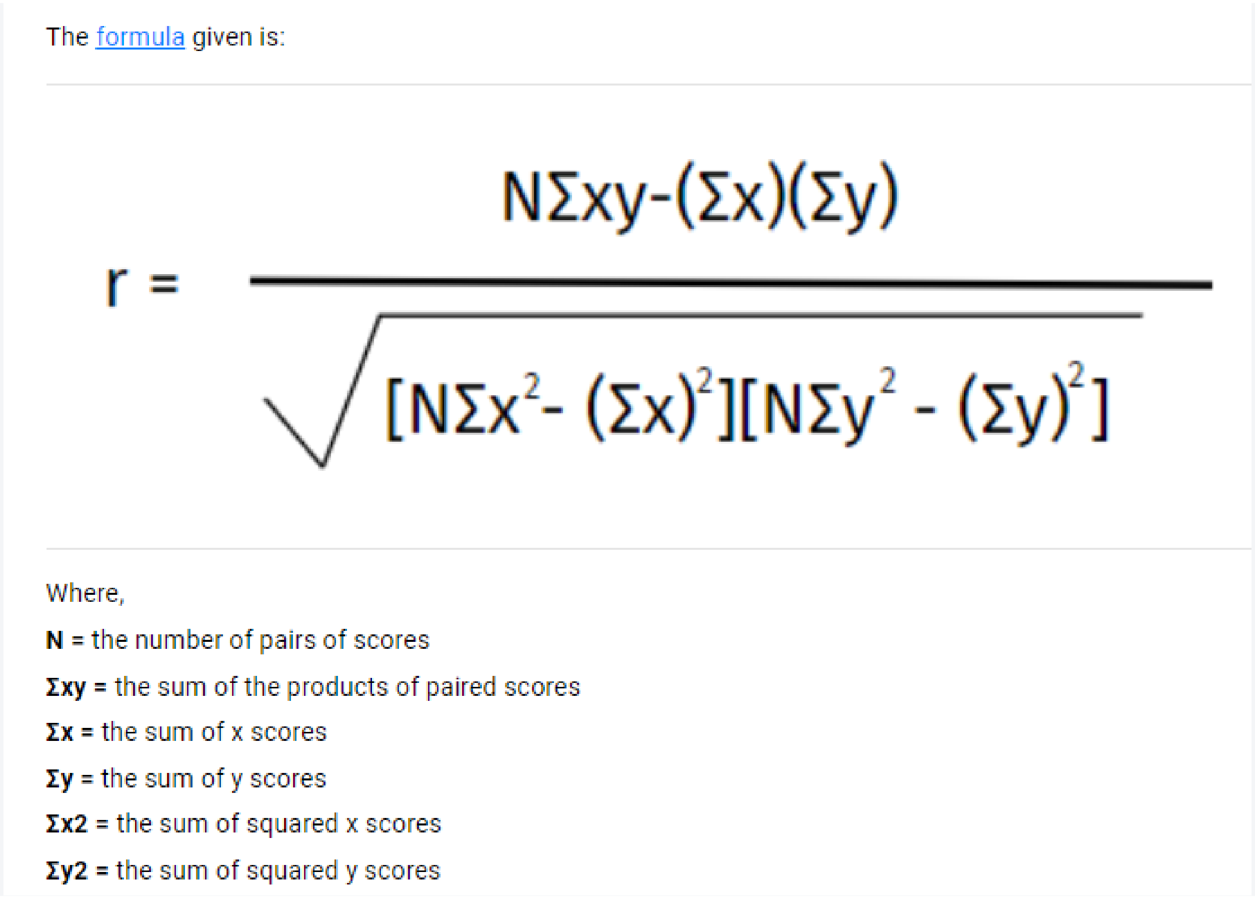
Pearson correlation coefficient Correlation measures the strength of association between two variables as well as the direction. There are mainly three types of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables.

Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and

+1.0. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data



The formula given is:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

QUESTION 4 - What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling does not affect the overall prediction or fit of the model. It just changes the coefficients. How to scale? Methods to Scale

1. Standardization – subtrating mean and diving by SD such that its centred at 0 and has SD of 1
2. MinMax scaling – getting value between 0 and 1

None of these methods change the shape or distribution of the original variable, it just scales them and shifts. Relation it had with outcome variable will not change.

This is to make sure we have all variables on the same scale.

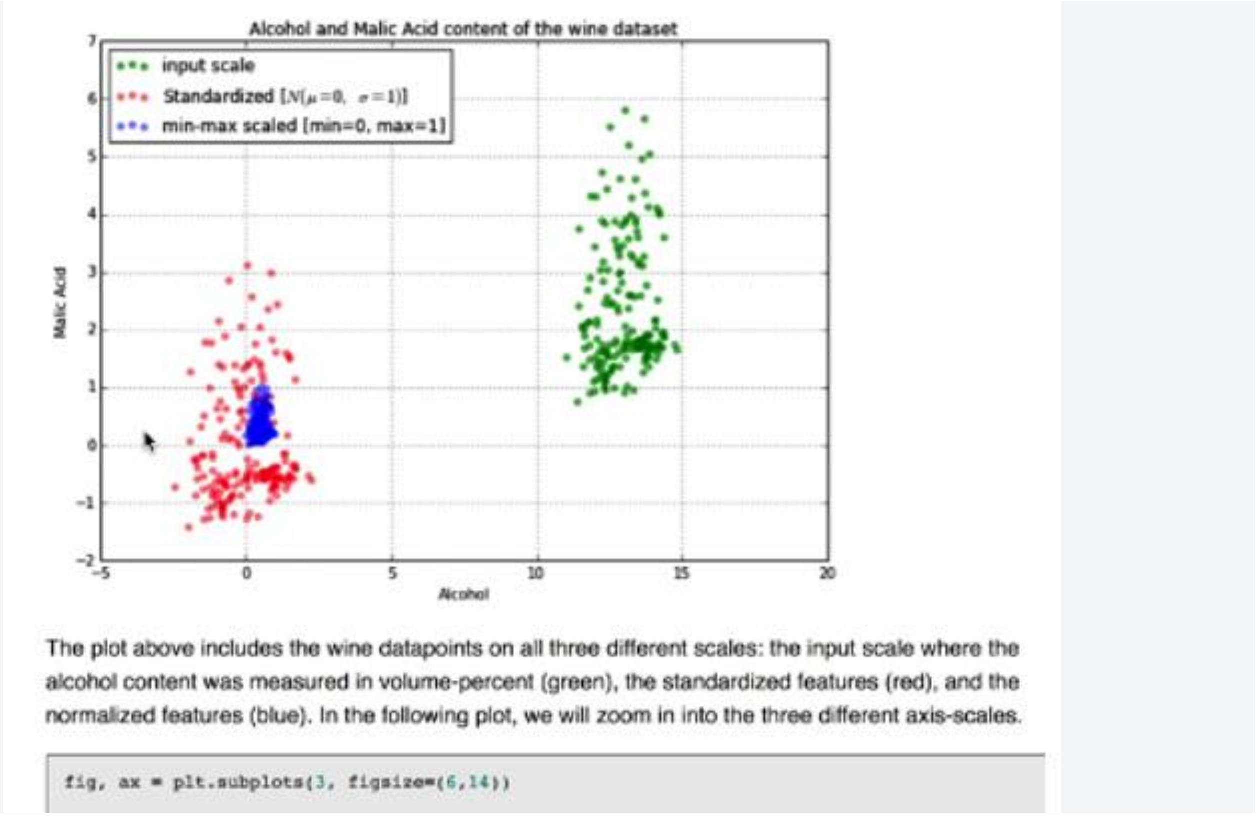
There are two major methods to scale the variables, i.e. standardisation and MinMax scaling. Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:

- **Standardisation:** $\mathbf{x = x - mean(x) / sd(x)}$
- **MinMax Scaling:** $\mathbf{x = x - min(x) / max(x) - min(x)}$

In the below graph, blue shows normalization where all values are compressed between 0 and 1

Standardized red data , data is spread is more but ,mean is 0 and std dev is 1.

As a general rule of thumb, use min-max scaling as it takes care of outliers. If there was an outlier, it would be mapped to 1 whereas in standardization, spread would be more in data because of outliers.


Alcohol and Malic Acid content of the wine dataset
input scale
Standardized [$\mathcal{N}(\mu=0, \sigma=1)$]
min-max scaled [min=0, max=1]
Malic Acid
Alcohol
7
6
5
4
3
2
1
0
−1
−2
−5
0
5
10
15
20

The plot above includes the wine datapoints on all three different scales: the input scale where the alcohol content was measured in volume-percent (green), the standardized features (red), and the normalized features (blue). In the following plot, we will zoom in into the three different axis-scales.

```
fig, ax = plt.subplots(3, figsize=(6,14))
```

QUESTION 5 - You might have observed that sometimes the value of VIF is infinite. Why does this happen

Variance Inflation Factor (VIF) - VIF calculates how well one independent variable is explained by all the other independent variables combined

In correlation coefficient, we see pair wise associations/correlation but this may not be enough as a variable may be associated with just not one variable but 2 to 3 other variables . eg X1 is completely defined by X2, X3, X4 . so how do we access that? i.e. some associations can get undetected as a variable can depend on more than just 1 variable. To quantify this association, we have a feature called VIF.

Once we have r2 for that particular variable, VIF is 1/1-R2 where i would stand for ith variable and Ri would be the model for ith variabke using all other variables excluding the outcome variable itself.

So when we do VIF, we do VIF for all variables so for each variable we will have VIF i.e how associated it is with other variables

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

QUESTION 6 - What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Steps

Arrange the items in ascending order.

Create a normal distribution curve

Calculate z-value

Plot the data set values against normal distribution cut off

Through Q Q plots, we can test the normality of the distribution.