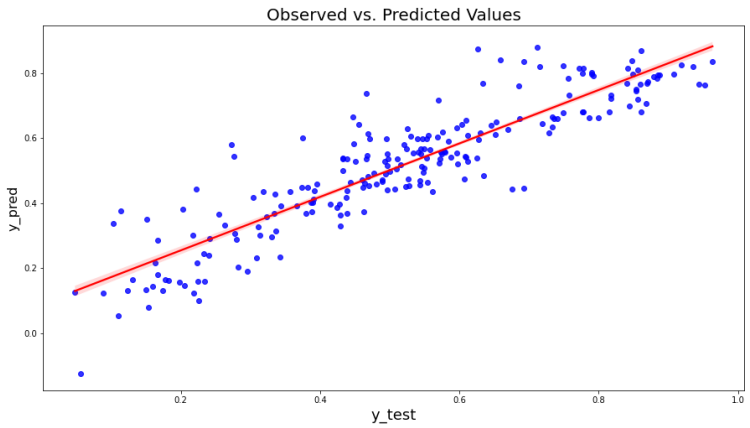
Assignment-based Subjective Questions

1. *From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)*
   Season, month, and weather all point towards the notion that pleasant temperatures and warmer months lead to more bikes being rented. These three points appear to be highly significant in how many bikes are likely to be rented. There also seems to be a general growth from 2018 to 2019 of the company in terms of bikes being rented. It looks like variables related to temperature (season, month, weather) have a much greater impact on the number of bikes rented than day of the week, holidays or working day.

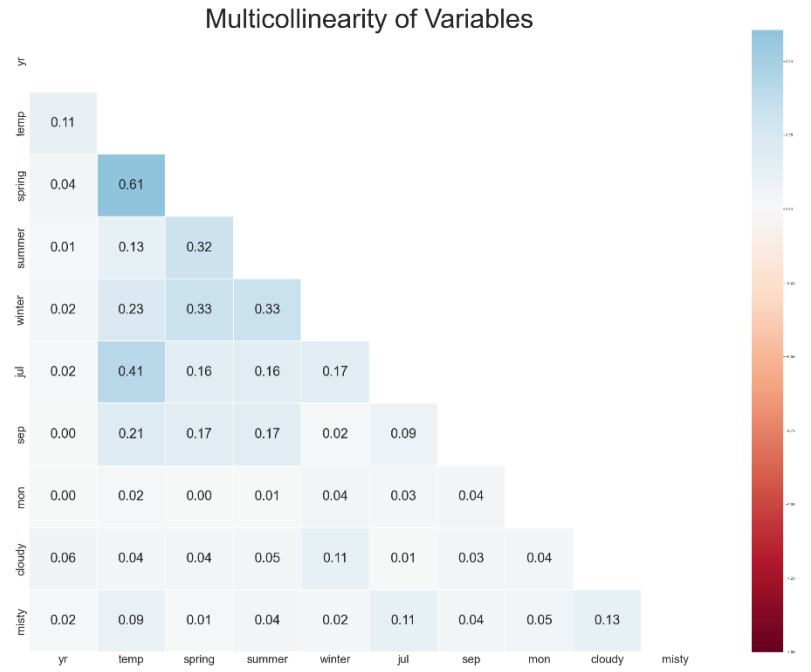
2. *Why is it important to use drop_first=True during dummy variable creation? (2 mark)*
   Dummy variables are created as n-1, that is one variable less than the data set contains. This is the case, because it is possible to still express all of the variables with one less variable. Dropping the first column allows it to remove the extra column that will be created during dummy variable creation and therefore reduces the correlations created among the variables. This is also known as “one hot coding” where categorical data is transformed into machine readable format using 1 and 0. Creating dummy variables is useful, because the variable is more flexible and allows multiple comparisons to be made. If the formula n-1 is not followed, it can create what is known as the “Dummy variable trap”. In this case, more categories than necessary (n) are created which means that there will be attributes that are highly correlated (multicollinear) with the result that one variable will predict another. To avoid this, dummy variables are created by excluding one dummy variable.

3. *Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)*
   Temperature and the related variable, perceived temperature.

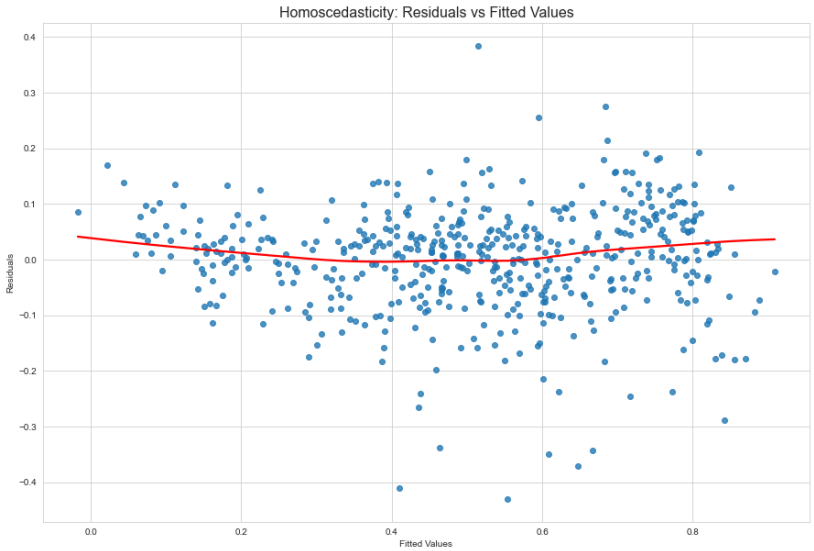
4. *How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)*
   I tested the assumptions of linear regression using different methods (all pictures in this section are from the analysis I conducted for the assignment).
   - *Linearity*: using regression plot of actual and predicted values to see if the values are closely scattered around the regression line



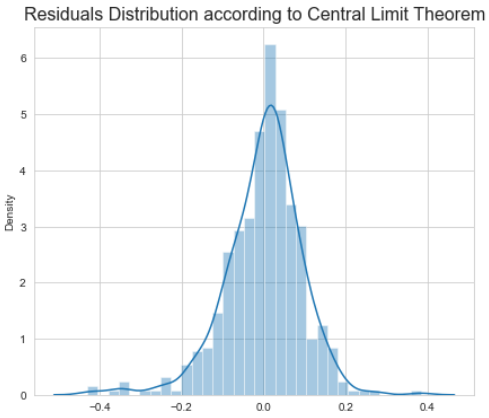
- *Multicollinearity*: correlation of 0.80 and above between variables is linked to multicollinearity. The model shows that the highest value is 0.60

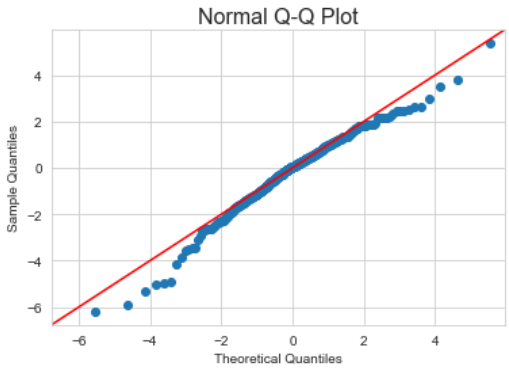


- *Autocorrelation*: Autocorrelation of the regression residuals can occur in incorrectly specified models. It violates the assumption that observations should be independent of each other. In the present data, this could have been a problem, because various variables describe points in time in different ways, and temperature is also linked to different times of the year. A result of two, like in the model created here, indicates no autocorrelation.
- *Homoscedasticity*: Assumes that the residuals variance is constant across the target. This is the case in the model.



- *Normality of residuals:* The residuals are normally distributed, shown as curve and Q-Q plot.





5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   In model 4, `temp` has the highest coefficient, with 0.50. This means, that when the temperature increases by one unit, the count of bike rentals also increases by 0.50. The next highest coefficient is `yr` with 0.23. There are also negative coefficients which show a relation in the opposite direction. The largest negative coefficient is `cloudy` with -0.30.

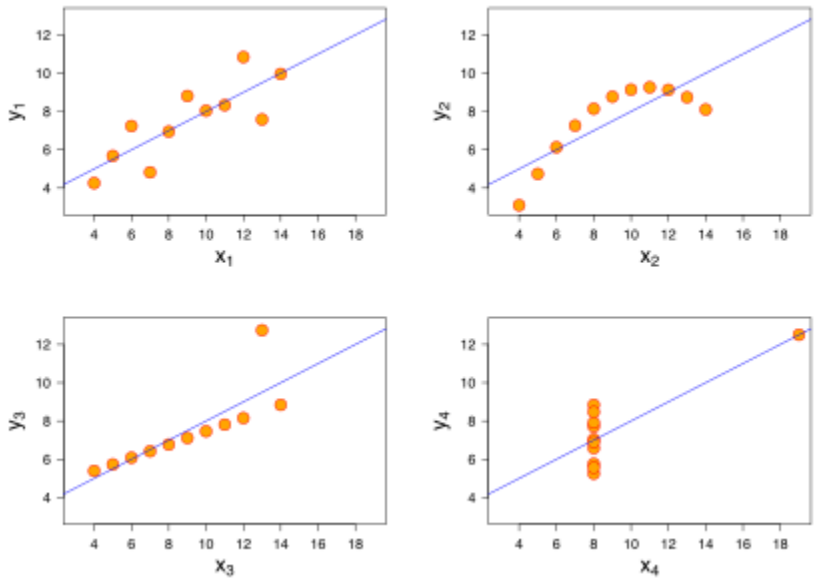
General Subjective Questions

1. *Explain the linear regression algorithm in detail. (4 marks)*
   Regression analysis is the exercise in modelling behaviour of a variable using one or more independent variables. It allows to test theoretical arguments which are causal statements. It aims to find out if there is a linear relationship between x (input) and y (output). A linear function, y=mx+c, is fit to the data in a scatter diagram. This function is called the line of best fit. It is created using the expected value of Y expressed as a straight-line function of X. The aim is to choose the line based on minimising the squared distance from each Y value to the line of best fit. This is achieved with the formula: $\sum_i (y_i - (\beta_0 + \beta_1 X))^2$
   For a relationship between a single dependent variable Y and one predictor, the formula is: $E[Y] = \beta_0 + \beta_1 X$
   In the instance of multiple predictors, the formula changes to describing multiple linear regression as: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \ldots + \beta_k X_k$
   The predicted values are also known as fitted values and describe the values of Y that are created when the X values are entered into the fitted model. These values, minus the actual observed values of Y are known as the residuals.
   Linear regression has four different assumptions that need to be met for the model to be meaningful (see also the questions above in which I provide examples on how I made sure that these assumptions are met in the assignment model.)

1. **Linearity**: relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: variance of residual is the same for any value of X.

3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

2. *Explain the Anscombe's quartet in detail. (3 marks)*
   It contains four data sets that contain nearly identical simple descriptive statistics. They have different distributions and are very different when shown in the graphs. It can be used to visualise the data before applying model building algorithms to identify anomalies in the data, such as outliers. This is important for linear regression, as it is only able to be used on data with linear relationships.
   It consists of a group of data sets that have the same mean, standard deviation and regression line, but differ in quality. It demonstrates why visualising the data is important and relying on basic statistics may not be sufficient to understand the data. It shows that simple linear regression may indicate the same type of estimate of a relationship, when visual inspection demonstrates that this is not the case. The first scatter plot shows a linear relationship, the second graph is not normally distributed, the third one should have a different regression line and the last graph has a high leverage point.



(picture from Wikipedia)

3. *What is Pearson's R? (3 marks)*
   Pearson's r (Pearson's Correlation Coefficient/Pearson product-moment correlation coefficient/bivariate correlation) shows if two variables are linearly correlated. It has a value between -1 and 1. It is the covariance of the two variables divided by the product of their standard deviations. It assumes that the scale of measurement is interval or ratio, normal distribution, linear association, and no outliers in the data. The closer the value is to -1 and 1, the stronger the correlation or association is between the variables. The strength of the obtained value shows the relationship connection between wo variables, showing how when one variable changes because of the adjustment of another. The direction of the line shows positive or negative connection between the variables.

4. *What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)*

Feature scaling describes data normalisation which is used to normalise or standardise either the range of independent variables or the features of the data during the pre-processing stage. It ensures that variables can be compared, because variables of differing scales will create a bias in the model, because they cannot contribute equally.
Normalised scaling changes the original data range to a range of 0 and 1 (MinMax Scaler in sklearn). Standardised scaling calculates each input variable by subtracting the mean and dividing by the standard deviation, so that the distribution has a mean of 0 and a standard deviation of 1. It is useful when data has values with differing scales. It assumes normal distribution of the data (StandardScaler in sklearn).

5. *You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)*
   An infinite VIF indicates perfect correlation. It means that R2=1, so the formula is 1/(1-R2). A perfect correlation between two variables indicates that the corresponding variable could be expressed by another combination of other variables. To solve this case of multicollinearity, one of the variables should be dropped from the data set.

6. *What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)*

   A Q-Q plot compares how close two distributions are. “Quantile-Quantile plots” plot quantiles of a sample distribution against the theoretical quantile values. It is plotted using a 45 degree reference line. They can help to assess if the two sets of data are derived from populations that have a common distribution, if residuals follow a normal distribution, and if the distribution is skewed. In regression models, the mean of the error terms should be zero. If the mean of the error terms is away from zero, it indicates that the features in the model are not having a significant impact on the outcome variable. When to distributions are similar, the points in the Q–Q plot lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not on the line y = x.

   The plot shows how properties such as location, scale, and skewness are similar or different in two distributions. My example from the assignment is available in the questions above.