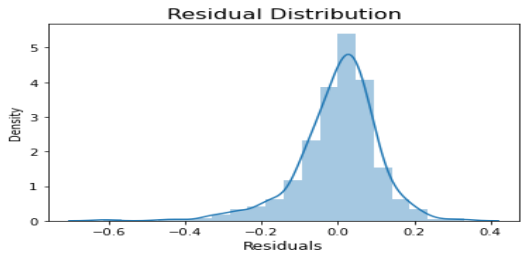
Assignment-based Subjective Questions - Answers

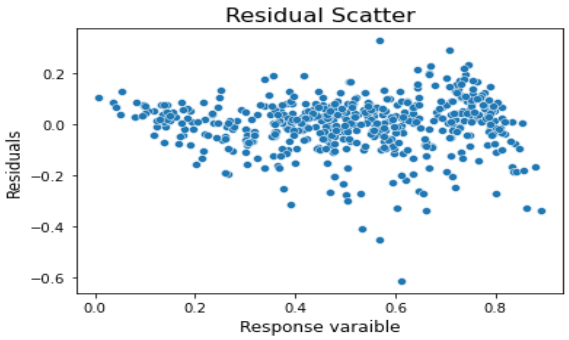
1. Based on 'weathersit' and 'season' we have derived features are which correspond to the values in those columns. Now to analyse the categorical variables it is prudent to look at box plots for each feature against the target variable. From this we have inferred the following:
    a. There is a huge increase in users from year 2018 to 2019, indicating that this might have a strong correlation with the target
    b. Count of total users are highest during months 6-10, which is essentially summer and autumn
    c. This is further verified by the individual graphs for summer, spring and winter which show that no of users is lower in spring, and winter as compared to summer
    d. Count of users is also dependent on weather which we can see is higher for clear days and lower for misty/cloudy days
    e. The spread of users remains the same for working days and weekend however we can see that the lower limit for users reduces on holidays; meaning less users go out cycling on holidays
2. During dummy variable creation we use binary encoding to encode the given levels for that categorical variable. Therefore, if a certain variable has 3 levels of information, we can drop one level without losing any information since it can be inferred from the values of the other 2. For example:
    a. A variable called season has 4 levels: winter, spring, summer, autumn. We can encode this in the following way.

Winter	Spring	Summer	Autumn
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

    b. Based on the above we can safely drop any seasons, since it can be inferred from the other 3. If winter, spring, summer are = 0 then autumn must be = 1. If any of the other 3 are = 1 then autumn must be 0.
    c. So, when we drop_first=True, we are simply deriving k-1 dummies from k levels without losing information
3. Looing at all the pair plots and ignoring ['casual' and 'registered'] since they perfectly define the target variable, we can see the that the strongest correlation to target variable is temperature (temp or atemp). This is to be expected since it makes practical sense that user decision to cycle is heavily dependent on the temperature or felt temperature.
4. The assumptions and steps taken to validate them ae given below:
    a. Linear relation between dependent and independent variable – verified using pair plots to check linearity between the target and other features
    b. Error terms are distributed normally – verified by plotting a histogram of the residuals (error terms). This histogram and its KDE are shown to be in a Gauss distribution.



c. Error terms are independent and have constant variance – verified by plotting a scatterplot of all residuals against the target variable (response). From this we can see that the errors are randomly distributed around mean zero. They have no pattern and have almost equal variance other than some outliers.



5. Based on the correlation between the variables and target as well as the absolute values of their respective coefficients, we see that top 3 predictor features are: 'temp', 'yr' and 'windspeed'. These predictors also make practical sense since temperature and windspeed will directly affect user behaviour and decision making to cycle or not.

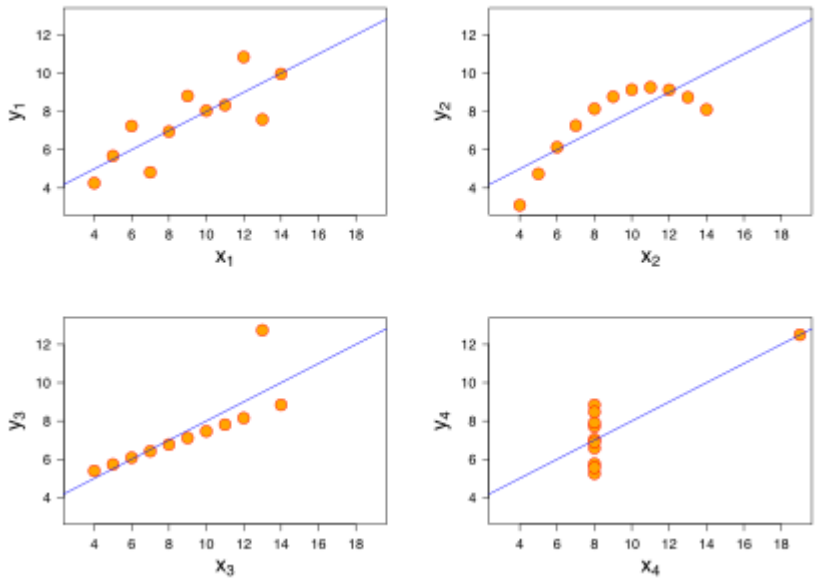
General Subjective Questions – Answers

1. Linear regression is a method for modelling the relationship between a dependent variable and one or more independent variables. When independent variable is only one it is called simple linear regression; for more than one, it is called multiple linear regression. In essence we are trying to fit a straight-line y=mx+c to the data set, where m=slope and c= intercept. To apply this kind of regression modelling there are certain assumptions:
    a. Dependent or response variable is continuous
    b. Linear relationship between independent and dependent variable
    c. Error terms are normally distributed
    d. Error terms are independent of each other
    e. Error terms have constant variance (homoscedasticity)

    Algorithm:
    - After data cleaning and prepping perform EDA. This is common for all data models
    - Try to discern the kind of relationship between response and other variables. There should be some linearity otherwise the linear will be a poor predictor
    - Data can be split into train and test sets; this can be done for sufficient existing data, but it is good practice for testing the model after it has been trained
    - Perform feature scaling on the training set using any of the given scaling methods:
        - MinMax scaling or Normalization
        - Standardization
        - Mean Normalization

- Perform feature selection using automated (RFE, etc.) or manual methods or a combination of both. This will reduce the number of features used as predictors in our model and make the final analysis easier to explain
- The best practice is to use some automated method to reduce the number of features (coarse selection) and from here use statistical methods and domain knowledge to manually add or remove further features (fine-tuning)
- We also look at Variance Inflation Factor in the case of multiple linear regressions to check multicollinearity between the features variable (excluding the target). As a general rule if the VIF for a variable is > 5 we should remove it from the model
- We also look at the R-squared and adjusted R-squared value. R2 value tells us how much of the variance is explained by the model and the adjusted R2 value penalizes the model for using more predictors than necessary
- Once the model is built, we verify the assumptions of the linear regression by residual terms to check if they are normally distributed. When plotted against the target they have no discernible pattern and are randomly distributed around the mean=0
- Now we evaluate the model by using it to predict on the test set and plot the predictions against the actual values. The plot should have a decent overlap. We also check the R2 value for test set, it should be like the training set otherwise our model could be underfitted or overfitted

2. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. Each dataset consists of eleven points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

- The second graph (top right); while a relationship between the two variables is not linear. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line.
- Finally, the fourth graph (bottom right) shows an example when one outlier is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance data visualization and looking at a dataset graphically before analysis based on simple statistical properties, as they can be inadequate for describing realistic datasets.

3. The Pearson correlation coefficient, also known as Pearson's r, is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
   The correlation coefficient ranges from −1 to 1. An absolute value of exactly 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1. A value of 0 implies that there is no linear dependency between the variables.
   In the case of linear regression using ordinary least squares method, it estimates the fraction of the variance in the dependent variable that is explained by the independent variables.

4. Scaling is used to bring all the features to a comparable scale using certain methods. E.g., a feature ‘bedrooms’ will have single digit number (min=1 and max=5) and another feature ‘area’ can have values in the range of thousands. For a machine which has no context this can cause the model to assign higher priority to larger number which may not be accurate. Scaling also makes gradient descent perform faster in the background and converge to a minima with higher efficiency.
   Normalization – Scaling the feature by subtracting the min and dividing by the range

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

   Standardization – Scaling the feature by subtracting the mean and dividing by the standard deviation

$$x_{new} = \frac{x - \mu}{\sigma}$$

5. VIF can only be infinity if R2 =1. This is the case of perfect correlation between two independent variables. This causes 1-R2 =0 thus VIF to be infinite. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6. 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.
   If the two distributions being compared are similar, the points in the Q–Q plot will approximately 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 necessarily on the line y = x