# 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)

**Answer:** The data contains the following 7 categorical variables:

- season
- weathersit (renamed to ‘weather)
- yr (renamed to ‘year’)
- mnth (renamed to ‘month’)
- weekday
- workingday
- holiday

Their effects on target-variable ‘cnt’ (renamed to ‘bike_rental_count’)as observed through boxplot visualisations are as follows:-

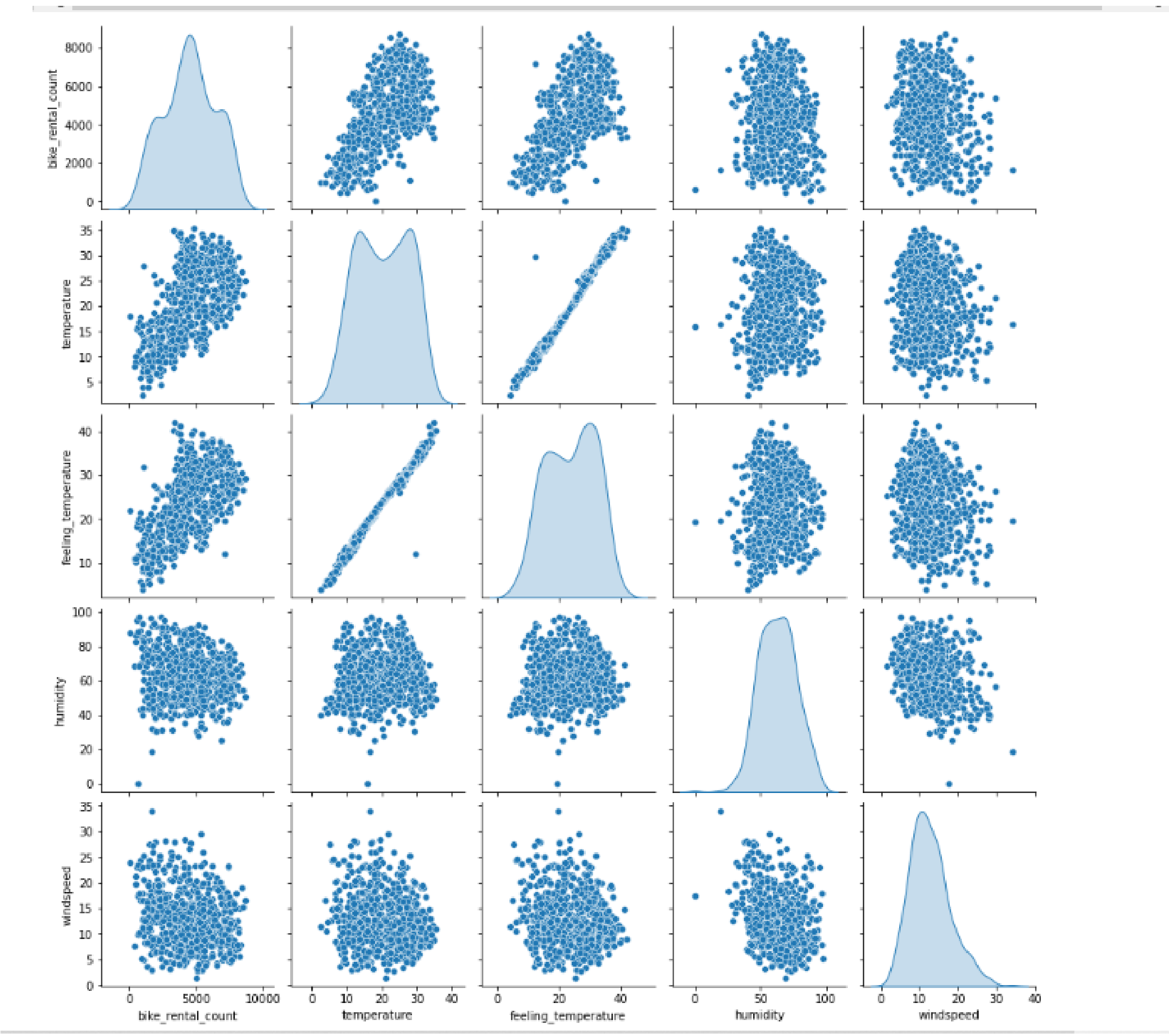
- **season** - The cnt has been recorded the lowest in the 'Spring' seasonand highest in ‘Fall’. Summer and Winter seasons showed intermediate values of cnt.
- **weathersit** - There are was no demand during 'Heavy Rain & Thunderstorm' which means it is extremely unfavourable. The cnt recorded highest for 'Clear' weather.
- **yr** - The cnt increased in the year '2019' as compared to '2018'
- **mnth** - The cnt has been recorded the highest for ‘September’ and lowest ‘ December’. Increased in months during 'Summer' and 'Fall' seasons which tally with the favourable and unfavourable weather-situations in September and December respectively.
- **weekday** – The median cnt does not experience much variation as per ‘weekday’.
- **workingday** – The cnt does not get affected much by 'workingday'
- **holiday** – The bikes rentals recorded less on 'holidays'.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

**Answer:** The argument **drop_first=True** is used to drop the first-dummy column created for a categorical variable. If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example: Iterative models may have trouble converging and lists of variable significance may be distorted. Another reason is, if we have all dummy variables it leads to multicollinearity between the dummy variables. To keep this under control, we remove one column.

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

**Answer:** “temp” and “atemp” (renamed to ‘temperature’ and ‘feeling_temperature’, respectively) are the two numerical variables which are highly correlated with the target-variable ‘cnt’ (renamed to bike_rental_count). The pair-plots are as follows:

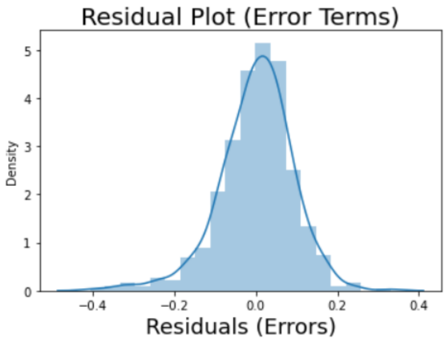


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

**Answer:** The assumptions of simple linear regression are:
1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

The assumptions of multiple linear regression are:
1. Model now fits a 'hyperplane' instead of a line
2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion)
3. For inference, the assumptions from from Simple Linear Regression still hold
4. The model should be free from multicollineraity and should not be overfitted

While building a model, the Residuals distribution should follow normal distribution and centred around 0.(mean = 0). Below is the distribution-plot for the Residual Analysis we performed on our training data-set.:-



Here, we can see the following:-

1. The center of the distribution is around 0.0
2. The errors are normally distributed i.e. the shape of the plot is a good normal-distribution

Thus, the assumption is validated.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: The top 3 features are contributing significantly towards the target-variable are:-

1. **temp** - coefficient : 0.5480 (renamed, to ‘temperature’)
2. ***Light_Snow*&_Rainy_weather** - coefficient : *0.2838*
3. **yr** - coefficient : 0.2328 (renamed, to ‘year’)

# General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

**Answer:** Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation “y = mx + c”.

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc.

Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression:** SLR is used when the dependent variable is predicted using only one independent variable.
2. **Multiple Linear Regression:** MLR is used when the dependent variable is predicted using multiple independent variables.

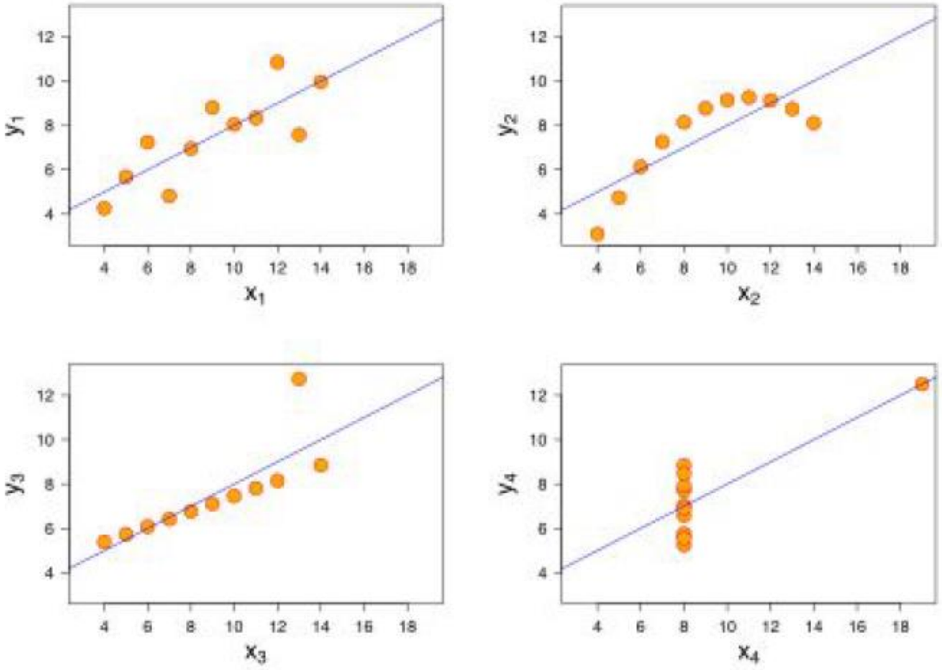
The equation for MLR will be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

β1 = coefficient for X1 variable
β2 = coefficient for X2 variable
β3 = coefficient for X3 variable and so on…
β0 is the intercept (constant term).

2. Explain the Anscombe’s quartet in detail. (3 marks)

**Answer:** Anscombe’s Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. The following plots can be used to explain this:



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them, it’s not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson’s R? (3 marks)

**Answer:** Pearson's r is a numerical summary of the strength of the linear association between the variables. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us whether we can draw a line graph to represent the data or not.

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association

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

**Answer:** Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

**Answer:** The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-R_1^2 ). If there is perfect correlation, then VIF = infinity, where,

R-1 is the R-square value of that independent variable which we want to check, as in, how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity".

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

**Answer:** A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?