# 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)**

   There were 6 categorical variables in the dataset.

   Box plot is used to study their effect on the dependent variable ('cnt').

   The inference that We could derive here is:

   - **season**: Almost 32% of the bike booking were happening in season3 with a median of over five thousand booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
   - **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
   - **weathersit**: Almost 68% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
   - **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
   - **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
   - **workingday**: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

   During the dummy variable creation, it is very important to use the 'drop_first=True' because it helps in reducing the extra column. Hence it reduces the correlations created among dummy variables.

   If we consider the bike sharing case study analysis, we created dummy variables for season it created season_1, season_2, season_3, season_4 variables but here season_1 is unnecessary so if we use the drop_first=True it will drop first dummy variable that is season_1.

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

From the pair plot we can consider ‘temp' and 'atemp' have highest correlation with the target variable 'cnt'.

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

Validating the assumption of Linear Regression Model:

- Linear Relationship: Linear regression assumes that there exists a linear relationship between the dependent variable and the predictors. During the analysis we draw the pair plot which concludes that there is a linear relation between temp and atemp variable with the predictor ‘cnt’.

- Homoscedasticity: Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable. In our case, there is no visible pattern in residual values, thus homoscedastic is well preserved.

- Absence of Multicollinearity: Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case). While it may not be important for non-parametric methods, it is primordial for parametric models such as linear regression. In our case, All the predictor variables have VIF value less than 5. So, we can consider that there is insignificant multicollinearity among the predictor variables.

- Normality of Errors: If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased. In our case, Based on the histogram, we can conclude that error terms are following a normal distribution

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

Top 3 predictor variables that influences the bike booking are temp, weathersit_3, yr.

- Temperature (temp) - A coefficient value of 0.5613’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5613 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.3021’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by -0.3021 units.
- Year (yr) - A coefficient value of 0.2309’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2309 units.

# General Subjective Questions

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

   Linear regression is a statistical regression method that is used to predict the analysis and it shows the relationship between the independent variables and dependent variables. linear regression model gives the sloped straight line which describes the relationship between the variables

   There are two types of pf linear regression:

   1. simple linear regression: With simple linear regression when we have a single input, we can use statistics to estimate the coefficients

   2. multiple linear regression:

   If two input variables are there, then such linear regression is called as multiple linear regression (MLR)

7. **Explain the Anscombe's quartet in detail. (3 marks)**

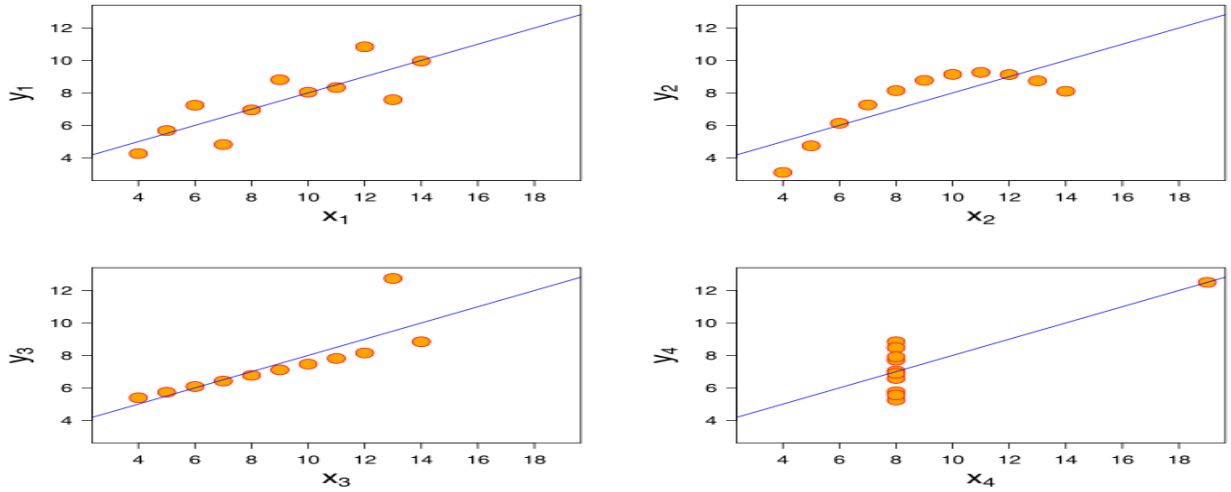
   Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

   • The first scatter plot (top left) appears to be a simple linear relationship.

   • The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear.

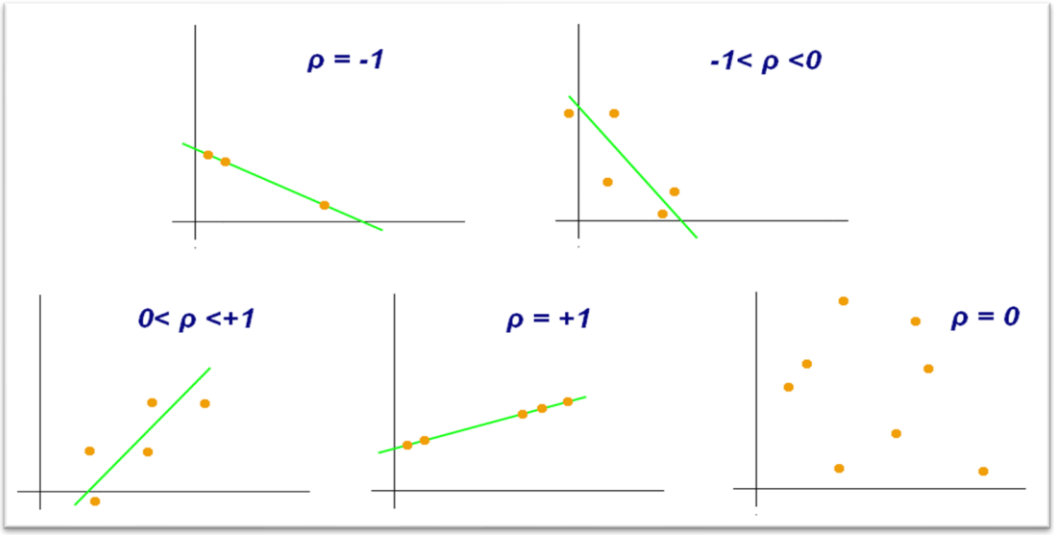
   • In the third graph (bottom left), the distribution is linear, but should have a different regression line.

   • Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient




**8. What is Pearson's R? (3 marks)**

The Pearson correlation coefficient, also referred to as Pearson's r, the Pearson product moment correlation coefficient (PPMCC), or the bivariate correlation is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.



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

Scaling helps to speeding up the calculations in an algorithm. Scaling is used as data pre-processing which can be applied within the particular range to the independent variables to normalize the data

Scaling methos is used to bring all the variables to the same level of magnitude

If scaling is not done then algorithm can lead to incorrect modelling because most of the time collected data set contains features which are highly varying in magnitudes, units and range. If scaling is not done algorithm consider only the magnitude not units. Scaling will affect only the coefficients

Difference between the normalized scaling and standardized scaling is below

- Normalization brigs all the data in range of 0 and 1

  We can use **sklearn.preprocessing.MinMaxScaler** to implement the normalization in python

  Formula: MinMax Scalling: x = x-min(x)/max(x)-min(x)
- Standardization scaling replaces the values by their Z score .

  This brings the data into a standard normal distribution which as mean and standard deviation

  Formula: Standardardization :x= x-mean(x)/Sd(x)

  We can use **sklearn.preprocessing.scale** to implement standardization in python.

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

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. We can consider another case where corresponding variable may be expressed exactly by a linear combination of other variables then that time also it shows infinity VIF. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. The Q-Q plot

use in linear regression in a scenario when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. Some of the uses and importance of a Q-Q plot in Linear Regression are:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios: If two data sets

1. Come from populations with a common distribution
2. Have common location and scale
3. Have similar distributional shapes
4. Have similar tail behavior