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)

Ans.

Inferential analysis of categorical variables from the dataset on cnt is:

- **Yr:** Bike bookings are higher in 2019 as compared to 2018, it might be due to the fact bike rentals are getting popular and people are becoming more aware about environment.
- **season:** Highest booking happening in season3(fall) with a median of over 5000 booking. This was followed by season2(summer) & season4(winter) of total booking.
- **mnth**: Bike booking is quite high 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 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking. This was followed by weathersit2. Clear weather is most optimal for bike renting.
- **holiday**: The bike booking were happening mostly when it is not a holiday.
- **weekday**: weekday variable shows very close trend. 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**: Median is quite close, does not have much impact.

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

Ans.

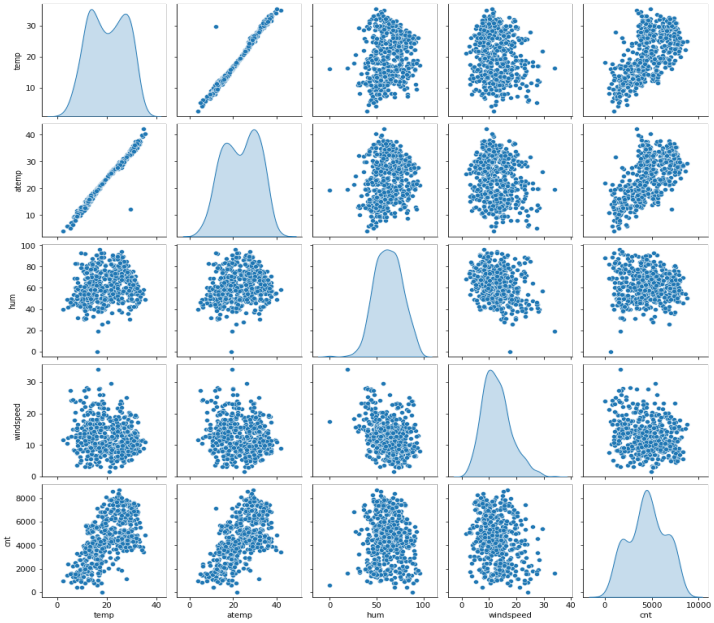
It is important to use **drop_first=True** to :

1) prevent getting a redundant feature as it can be explained as the linear combination of the other dummy features.
2) The approach reduces multi-collinearity in the dataset, which is one of the prime Assumption of Multiple Linear Regression.

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

Ans.

As we can see 'temp' and 'atemp' appear to highly corelated with taget variable 'cnt'.

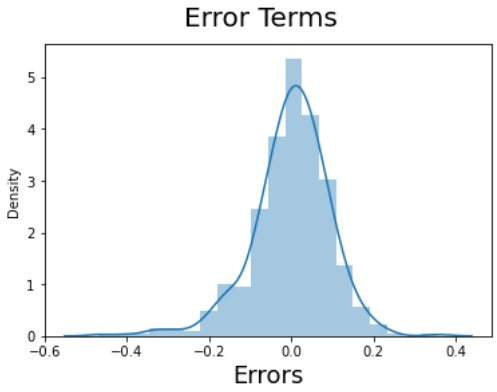


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?** (3 marks)

Ans.

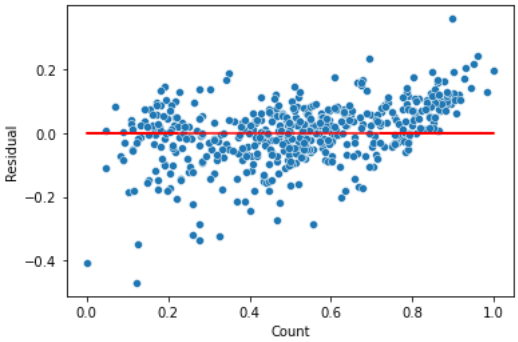
After building the model on the training set, I did the following analysis:

I. Error terms are normally distributed with mean as zero



II. Linear relationship between X and y

III. Residue error follow 'Homoscedasticity'



IV. There is No Multicollinearity between the predictor variables
   a. p-value of all predicting variables is below 0.05
   b. VIF value of all predicting variables is below 5

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

Ans.

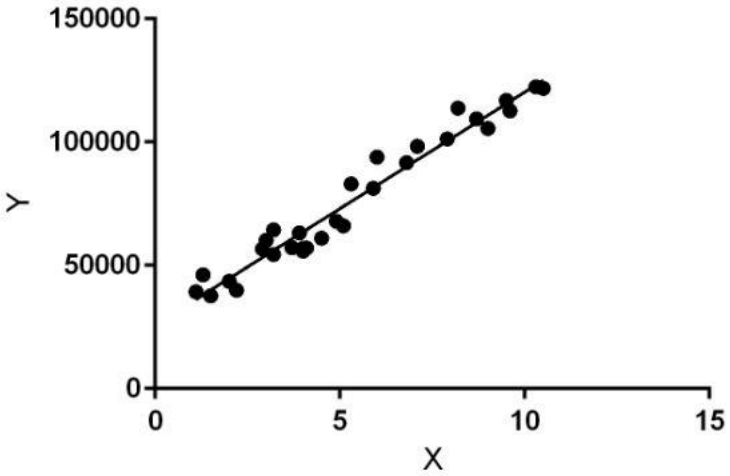
As per our final Model, the top 3 predictor variables that influences the bike booking are:

1) **Temperature (temp**) – A coefficient value of '0.441121'
2) Weather Situation 3 (weathersit_Light_Snow_Rain**)(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered)** – A coefficient value of '-0.328180'
3) **Year (yr)** – A coefficient value of '0.231005'

General Subjective Questions

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

   Ans. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
The linear regression model can be represented by the following equation:
$Y = \theta_o + \theta_1 x_1 + \theta_1 x_1 + \ldots\ldots + \theta n\, xn$
where,
Y is the predicted value
θ_o is the constant term.
$\theta_1,\ldots,\theta n$ are the model parameters
$x_1, x_2,\ldots,xn$ are the feature values.
The goal of regression analysis is to create a trend line based on the data you have gathered. This then allows you to determine whether other factors apart from the number of calories consumed affect your weight, such as the number of hours you sleep, work pressure, level of stress, type of exercises you do etc. Before considering, we need to look at these factors and attributes and determine whether there is a correlation between them. Linear Regression can then be used to draw a trend line which can then be used to confirm or deny the relationship between attributes. If the test is done over a long-time duration, extensive data can be collected and the result can be evaluated more accurately.

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

It is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset contains of eleven (x, y) pairs as follows:-

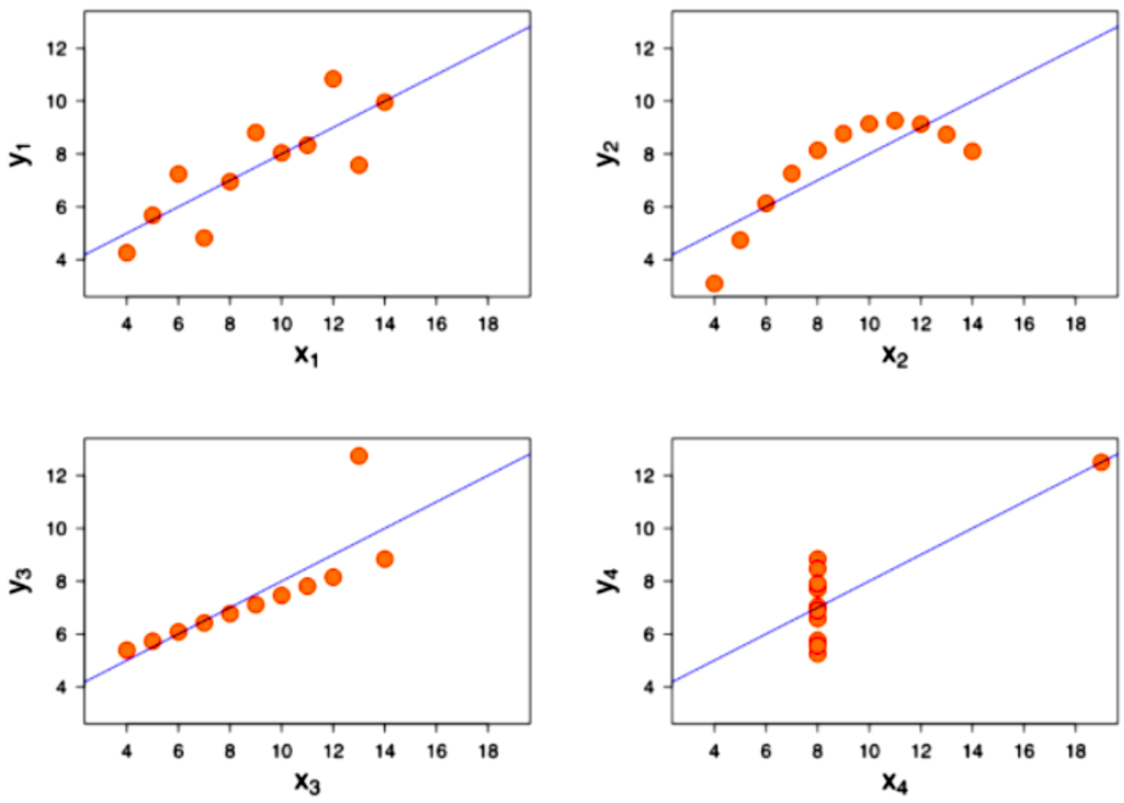
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All the summary statistics for each dataset are identical

1. The average value of x is 9.
2. The average value of y is 7.5.
3. The variance for x is 11 and y is 4.12
4. The correlation between x and y is 0.816
5. The line of best for is y = 0.5x + 3.

But the plots tell a different and unique story for each dataset.

- 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) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.



- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). 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. It values ranges between -1 to +1. It shows the linear relationship between two sets of data.

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

r=1 means the data is perfectly linear with a positive slope
r=-1 means the data is perfectly linear with a negative slops
r=0 means there is no linear association

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

Ans.

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.
Normalized scaling means to scale a variable to have values between 0 and 1, while standardized scaling refers to transform data to have a mean of zero and a standard deviation of 1

What?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.
- sklearn.preprocessing.MinMaxSc $\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$ on

Strandardized Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outlier**

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

Ans.

The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables within a multiple regression. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone.
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression 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.

Few advantages:
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.

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes

iv. have similar tail behaviour

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 i.e.