Assignment-based Subjective Questions

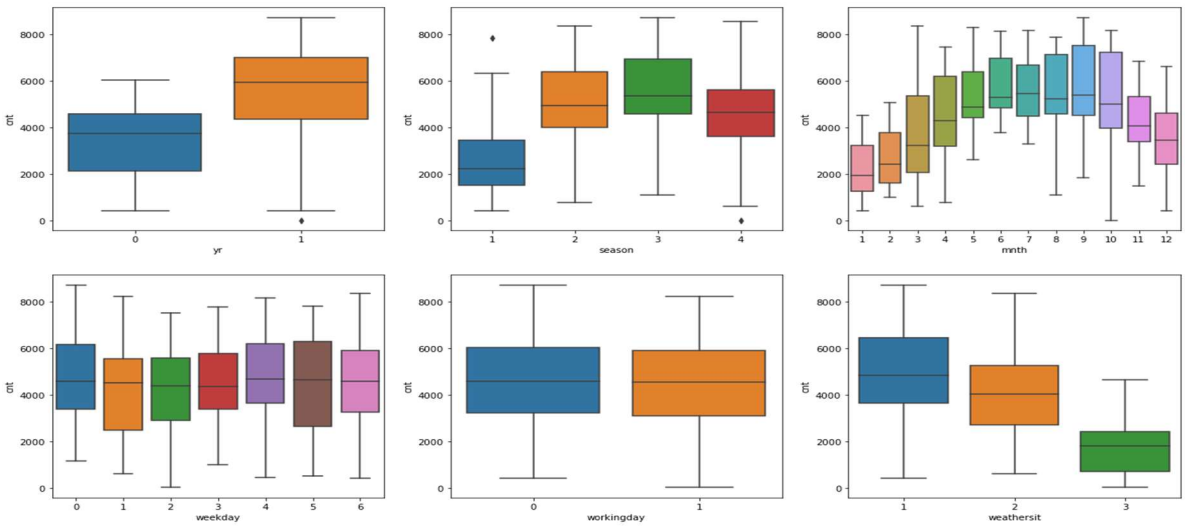
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans :

As per my analysis , I have observed that categorical variables (year ,season, month, weekday, working day, weather situation and holiday) have a major effect on dependent variable “Count”. Variables such as weekday and 'working day' doesn’t showed much inference. It is quiet tough to infer anything from 'workingday' and weekday, so I decided to drop them. Whereas ‘year’, ‘seasons’,’month’ and ‘Weather situation showed a lot of variance.

Code :

```
plt.figure(figsize=(20, 12))
plt.subplot(2,3,1)
sns.boxplot(x = 'yr', y = 'cnt', data = bike_data)
plt.subplot(2,3,2)
sns.boxplot(x = 'season', y = 'cnt', data = bike_data)
plt.subplot(2,3,3)
sns.boxplot(x = 'mnth', y = 'cnt', data = bike_data)
plt.subplot(2,3,4)
sns.boxplot(x = 'weekday', y = 'cnt', data = bike_data)
plt.subplot(2,3,5)
sns.boxplot(x = 'workingday', y = 'cnt', data = bike_data)
plt.subplot(2,3,6)
sns.boxplot(x = 'weathersit', y = 'cnt', data = bike_data)
plt.show()
sns.boxplot(x = 'holiday', y = 'cnt', data = bike_data)
plt.show()
```



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

Ans:

drop_first = True, is used to drop the first column of the dummy variable dataframe , so that we only left with **n-1** number of variable in dummy_dataframe , where n is number of categories. If I didn't had drop the first column then my dummy variables will be correlated and it can affect some models adversely and the effect is stronger when the cardinality is smaller.

For example, we have variables ' Male' and 'Female' , we can convert them in one categorical variable as gender where 0 represents Male and 1 represent Female.

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

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

e.g.

```
In [28]: # Get the dummy variables for the feature 'furnishingstatus' and sto

         df_season = pd.get_dummies(bike_data['season'], drop_first = True)

         # Check what the dataset 'status' looks like
         df_season.head()
```

Out[28]:

	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

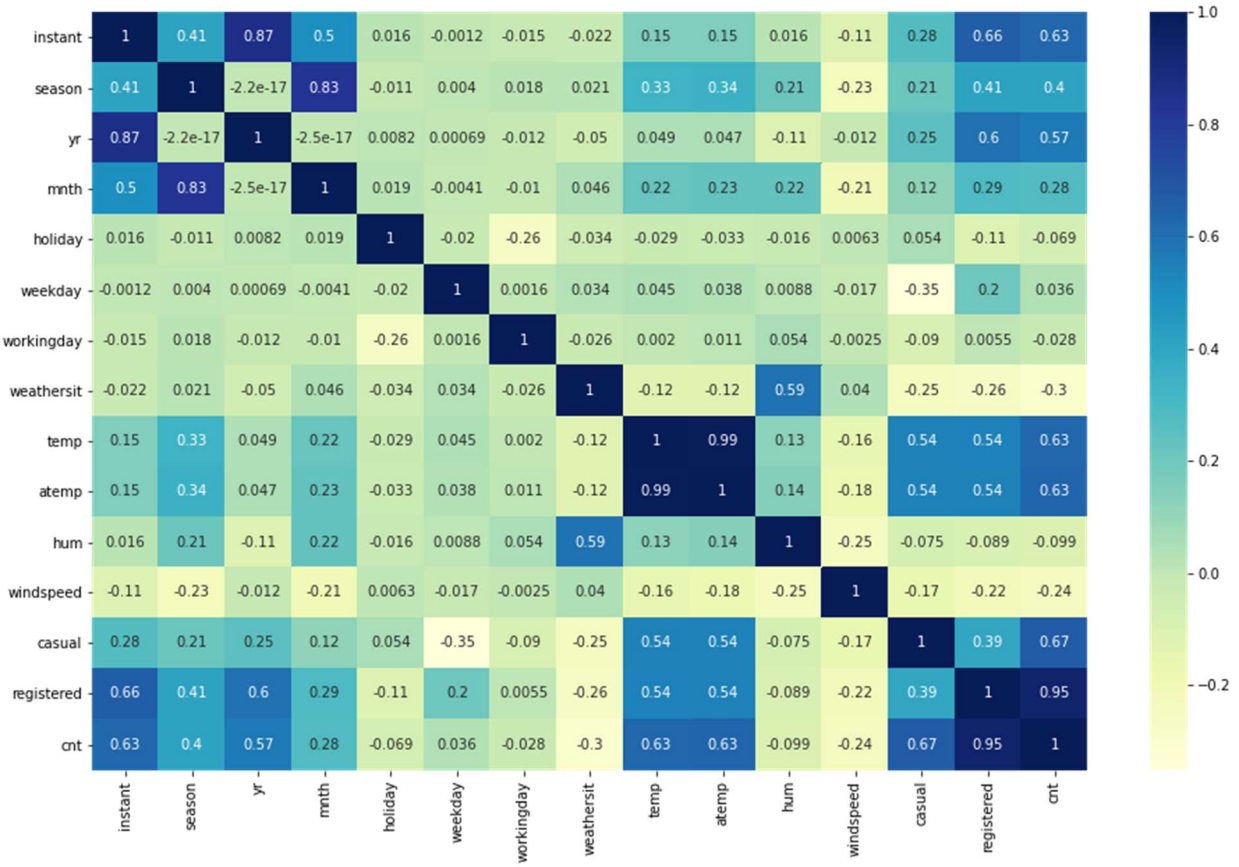
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Ans:

Registered variable is highly co-related with count variable . We can see below :

Code :

```
plt.figure(figsize = (16, 10))
sns.heatmap(bike_data.corr(), annot = True, cmap="YlGnBu")
plt.show()
```

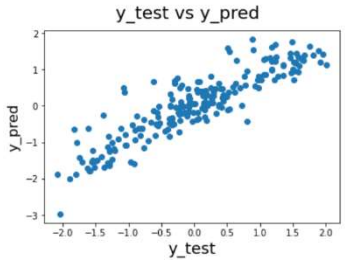


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?** Ans : Checking for R2 square values and Normal Distribution curve .

Model Evaluation ¶

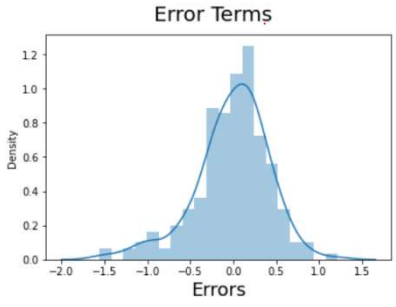
```
]: # Plotting y_test and y_pred to understand the spread.
   fig = plt.figure()
   plt.scatter(y_test,y_pred)
   fig.suptitle('y_test vs y_pred', fontsize=20)              # Plot heading
   plt.xlabel('y_test', fontsize=18)                          # X-Label
   plt.ylabel('y_pred', fontsize=16)                          # Y-Label
```

```
]: Text(0, 0.5, 'y_pred')
```




```
In [986]: # Plot the histogram of the error terms
          fig = plt.figure()
          sns.distplot(res, bins = 20)
          fig.suptitle('Error Terms', fontsize = 20)
          plt.xlabel('Errors', fontsize = 18)
```

```
Out[986]: Text(0.5, 0, 'Errors')
```



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

   Ans : Based on the final model temp:temprature , windspeed and Good weather are the three majorly affect the demand of bikes.

General Subjective Questions and Answers ::

1. Explain the linear regression algorithm in detail.

Ans :

Linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

It is a machine learning algorithm based on **supervised learning,** where we check the **data is continuous manner**. It performs a **regression task**. Regression models a target dependent variable based on independent variables. It is mostly used for finding out the relationship between variables. 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).
The regression line is the best fit line for our model.

$$y = \theta_1 + \theta_2.x$$

Hypothesis function for Linear Regression :

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x

Regression analysis is used for three types of applications:

- Finding out the effect of Input variables on Target variable.
- Finding out the change in Target variable with respect to one or more input variable.
- To find out upcoming trends.

2. **Explain the Anscombe's quartet in detail**.

Ans :

Anscombe's Quartet is defined as a group of four dataset which are **almost identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on

scatter plots. It is used illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**. These will be four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x, y points in all four datasets.
The importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

3. What is Pearson Correlation?

Ans :

Correlation coefficients are used to measure how strong a relationship is between two variables. Pearson Correlation is defined as **Pearson Product Moment Correlation (PPMC)**. It shows the linear relationship between two sets of data. In simple terms, it answers the questionTwo letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

The PPMC is not able to tell the difference between dependent variables and independent variables. For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also get the same result with the variables switched around. In other words,

Real Life Example

Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson's correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans : 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.

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.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization 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 **outliers**.

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

Ans:

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). If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity

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

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.

Advantages of Q-Q plot:

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.

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis