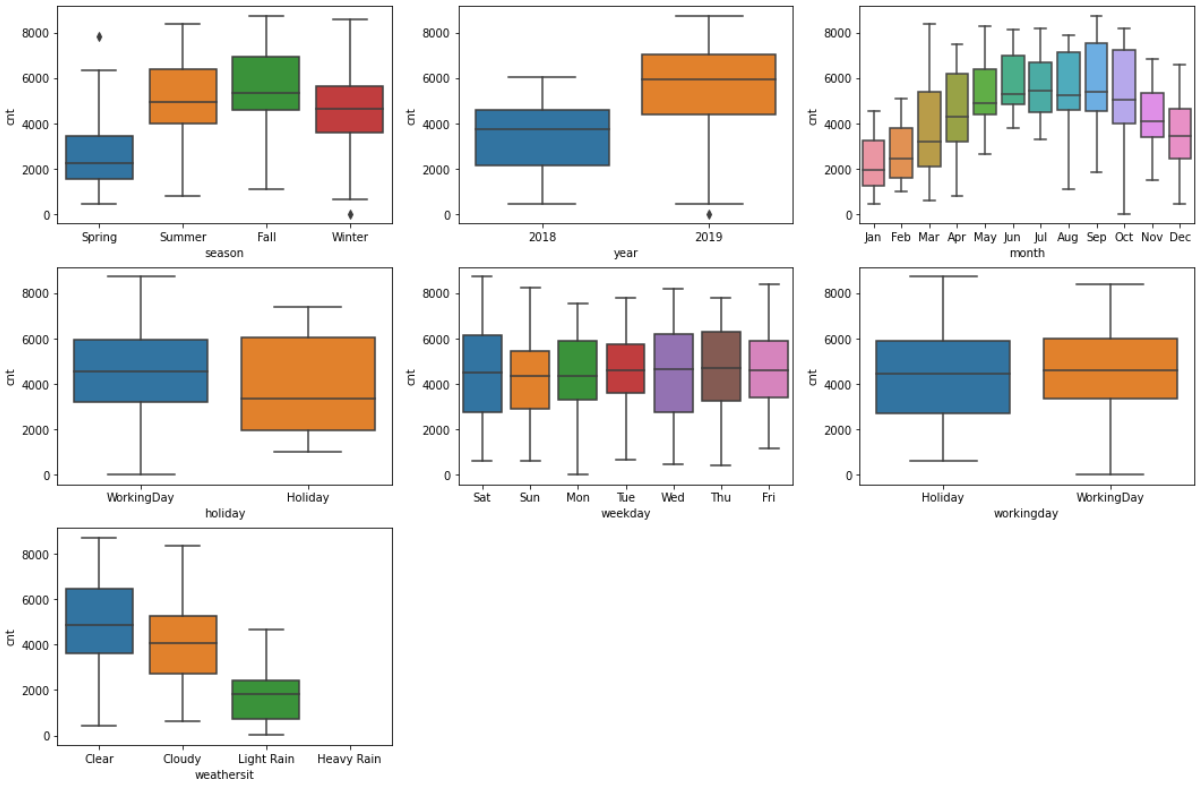
Assignment-based Subjective Questions

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: The plot of categorical variable against the dependable variable “cnt” is as below:



From the above plot we can clearly analyse below points

1. The demand of the bike is maximum in season “Fall” followed by “Summer”, “winter” and “spring”
2. The demand of the bike is maximum in the year 2019 when compared to 2018
3. The month “September” has maximum demand.
4. “WorkingDay” is the preferred day and has high median of bike demand.
5. The demand of the bike is maximum when sky is “Clear” followed by “Cloudy”, “Light Rain” and “Heavy Rain”

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
Ans: Drop_first=True is required to create n-1 new dummy variables for a categorical variable with n levels. This is known as n-1 encoding
Example: Season has 4 levels namely - 1:Spring, 2:Summer, 3:Fall, 4:Winter

```
# Get the dummy variables for the feature variable 'season'
season_dummy = pd.get_dummies(bike_df['season'], drop_first=True)
season_dummy.head()
```

✓ 0.7s

	Spring	Summer	Winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

Where,
Spring = 1 0 0
Summer = 0 1 0
Winter = 0 0 1
Fall = 0 0 0

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans: “temp” (0.64), “atemp” (0.65) and “yr” (0.59) variables have highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans: There are mainly 5 assumptions of Linear regression which are validated after building the model

1. Linear Relationship between the features and target – There must be linear relationship between X (independent variable) and y (dependent variable). This can be checked by plotting scatter plot between X and y.
2. Error terms are normally distributed with mean zero – This can be validated by plotting a histogram of the error term also known as residual and check whether the error terms are normal
3. Error terms are independent of each other – This can be validated by plotting a scatter plot of the error term and check no visible patterns exist.
4. Error terms have constant variance (homoscedasticity) – This can be validated by plotting a scatter plot of predicted and residual values and check if it has constant variance.
5. Little or no Multicollinearity between the features – This can be validated by plotting pair plots and heatmaps and identifying highly correlated features and not included in the model.

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: Top 3 features which contribute significantly towards explaining the demand are “Temperature”, “Year” and “Winter Season”

cnt = 0.2331 * yr + 0.0515 * workingday + 0.6022 * temp - 0.1388 * windspeed + 0.1052 * Winter + 0.0617 * Sat - 0.2546 * Light Rain + 0.0531

General Subjective Questions

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

Ans: Linear regression in machine learning is a supervised algorithm and most used regression algorithm. In this algorithm we train a model to predict the behaviour of data based on the linearity of the input. In simple words, linear regression means fitting the best fit line / hyperplane between independent and target variables with the least mean square error.

There are mainly two types of linear regression

1. Simple linear regression
2. Multiple linear regression

Simple linear regression: This is the most elementary type of regression model which explains relationship between a dependent variable and one independent variable using a straight line

Mathematically, we can write a simple linear regression equation as:

Y = B0 + B1X, where B1 is slope and B0 is intercept

Multiple linear regression: This type of regression model is a statistical technique to understand the relationship between one dependent variable with several independent variables.

The steps involved in creating a linear

Mathematically, we can write a multiple linear regression equation as:

Y = B0 + B1X1 + B2X2 + ……. + Bn+Xn

The general assumptions of linear regression model are:

1. Linear relationship between X and Y
2. Error terms are normally distributed and mean zero
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

In Multiple linear regression we also assume that there is little or no Multicollinearity between the features.

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

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

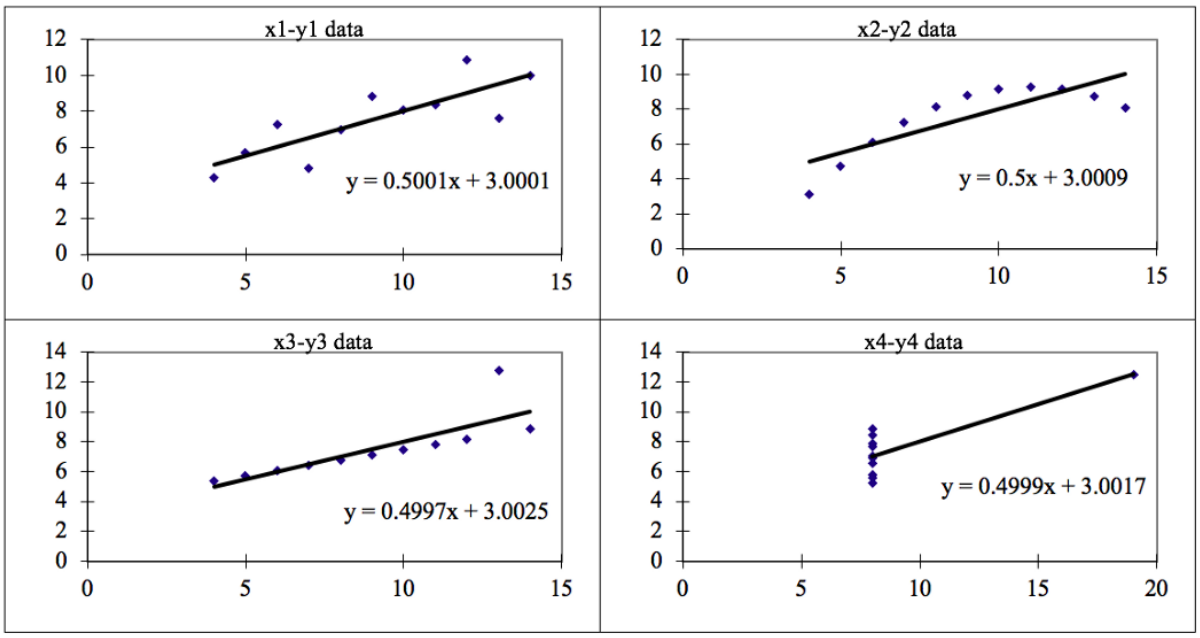
Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

This tells us about the importance of visualising 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 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:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R? (3 marks)
Ans: Pearson's R is defined in statistics as the measurement of linear correlation between two sets of data.
Pearson correlation coefficient is derived as

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

- r = correlation coefficient
- xi = values of the x-variable in a sample
- x(bar) = mean of the values of the x-variable
- yi=values of the y-variable in a sample
- y(bar) = mean of the values of the y-variable

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- 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)
Ans: When there are lot of independent variables in a model, a lot of them have different scales then it leads a model with very weird coefficients that might be difficult to interpret. Bringing all these independent variables on same scale is know as scaling.

Scaling is performed because of two reasons:
1. Ease of interpretation
2. Faster convergence for gradient descent methods

The difference between normalized and standard scaling are as follows

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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

Ans: The value of VIF is determined by below formula

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

The value of VIF becomes infinite when the value of R2 becomes 1 i.e 1 / (1 – 1) = infinity
We get R2 = 1 when there is a perfect correlation between two independent variables. An infinite VIF value indicates that an independent variable may be expressed exactly by a linear combination of other variables.

To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or Q-Q plot. 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. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Q-Q plot functionality is available under statsmodels.api library.
Example:

```
>>> import statsmodels.api as sm
>>> from matplotlib import pyplot as plt
>>> data = sm.datasets.longley.load()
>>> exog = sm.add_constant(data.exog)
>>> mod_fit = sm.OLS(data.endog, exog).fit()
>>> res = mod_fit.resid # residuals
>>> fig = sm.qqplot(res)
>>> plt.show()
```