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?**

Solution:

From the box plots we can infer that

- In Fall season there are high rentals compared to other seasons.
- Counts are high at September month but low in January, February.
- Fridays and Thursdays there are high count compared to other weekdays
- Clear weather is having high rental counts to other weather situations.

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

Solution:

Creation of dummy variables is to get n-1 categorical variables out of total n variables, which reduces the complexity and make the model simple and inferable. The first level can be explained by combination of remaining dummy variables created.

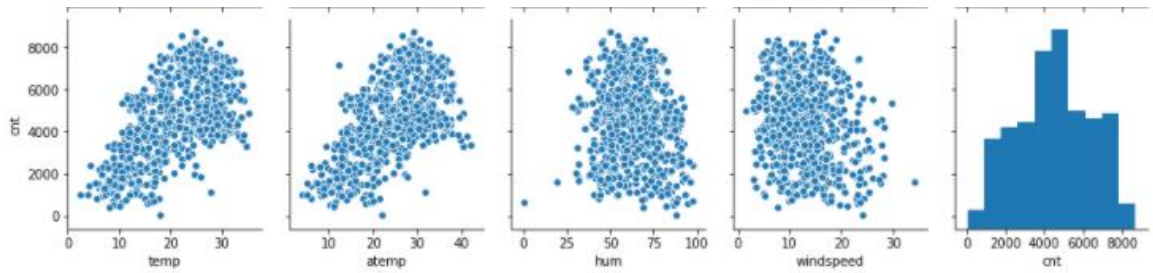
```
#creating the dummy variable for weather situation
weathersit_dummy=pd.get_dummies(bike_rental['weathersit'],drop_first = True)
weathersit_dummy
```

	Light snow rain	Mist
0	0	1
1	0	1
2	0	0
3	0	0
4	0	0
...	...	...
725	0	1
726	0	1
727	0	1
728	0	0
729	0	1

730 rows × 2 columns

Here, if the weather situation is Mist, it's value will be 1 and others will be 0;
if the weather situation is Light snow rain, it's value will be 1 and others will be 0;
Clear is explained by other 2 variables when both of them are zero.

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



Here the numerical variables 'temp' and 'atemp' has highest correlation with the target variable.

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

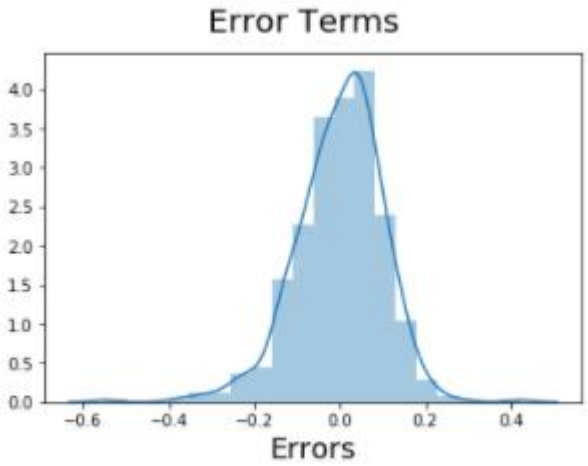
1. R-squared value is 79.2% and adjusted R-squared value is 78.7% which tells exactly how much variance in the data has been explained by the model. Higher the value, better the model.
2. VIF: This tells whether multicollinearity is present between the variables of dataset or not. (All VIFs are < 5 in the final model)

	Features	VIF
2	windspeed	3.22
3	Spring	2.36
0	yr	1.74
4	Summer	1.69
6	Jan	1.62
10	Mist	1.47
7	Nov	1.23
8	Sep	1.16
5	Dec	1.13
1	holiday	1.06
9	Light snow rain	1.06

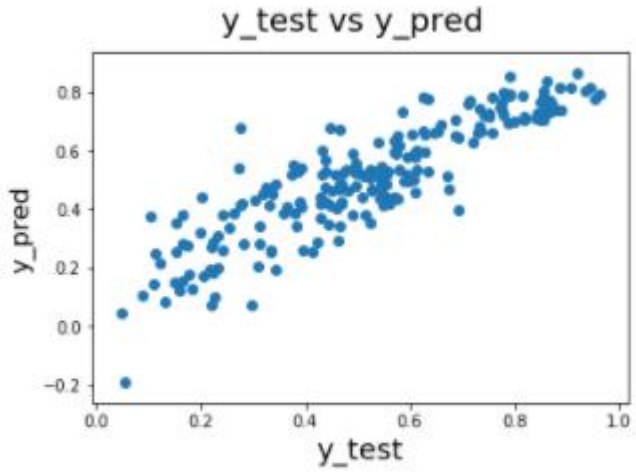
3. p-value of coefficients: This tells whether the coefficient is significant or not. (All p-values are < 0.05 in the final model)

	coef	std err	t	P>|t|
const	0.5845	0.013	45.172	0.000
yr	0.2456	0.009	26.656	0.000
holiday	-0.0855	0.030	-2.892	0.004
windspeed	-0.1910	0.028	-6.706	0.000
Spring	-0.2372	0.015	-16.294	0.000
Summer	-0.0392	0.013	-3.087	0.002
Dec	-0.1169	0.017	-6.725	0.000
Jan	-0.1215	0.020	-6.150	0.000
Nov	-0.1107	0.018	-6.167	0.000
Sep	0.0602	0.018	3.287	0.001
Light snow rain	-0.3160	0.028	-11.416	0.000
Mist	-0.0884	0.010	-9.008	0.000

4. Distribution of error terms: Histogram of distribution of error terms is normalised curve around 0.



5. Graph between y_test and y_pred: This scatterplot is almost linear.



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

Year (0.2456), **spring** (-0.2372) and **light snow rain** (-0.3160) are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Regression is the most commonly used predictive analysis model.

Explaining Algorithm stepwise include:

Step !: Read in the data after importing required dataset and libraries.

Step 2: Then analyse the dataset by performing some EDA techniques, find null columns and initiate methods to sort it, check similar datatypes, observe pattern within the independent variables by using correlation techniques. Build a basic understanding of the dataset.

Step 3: Preparing X and y which include putting feature variable to X and putting response variable to y.

Step 4: Splitting data into train and test

Step 5 : Performing Linear Regression

import LinearRegression from sklearn

Representing LinearRegression as lr(Creating LinearRegression Object) lr = LinearRegression()

There is no need to specify an object to save the result because 'lr' will take the results of the fitted model. lr.fit(X_train, y_train)

Step 6: Coefficients calculation

Print the intercept and coefficients

Step 7: Making predictions on the testing set

step 8: . Model evaluation (Plot Actual vs Predicted)

Step 9: Model evaluation (Plot Error terms)

step 10 : Checking mean square error and R square

You can now check the mean square error and r square value of your model. Your model is getting a mean square error of 7.9 which means the model is not able to match 7% of the values only, which is good. The r square value is about 60% which means our model is able to explain 60% of the variance which is also good.

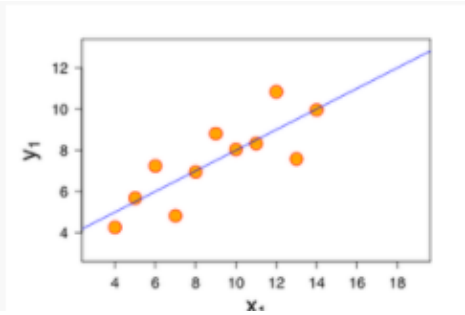
2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet contains four distinct datasets, each with statistical properties that are essentially identical: the mean of the x values is 9.0, mean of y values is 7.5**, they all have nearly identical variances, correlations, and regression lines (to at least two decimal places).**

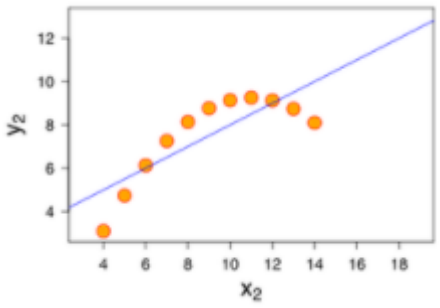
Dataset of Anscombe's Quartet

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

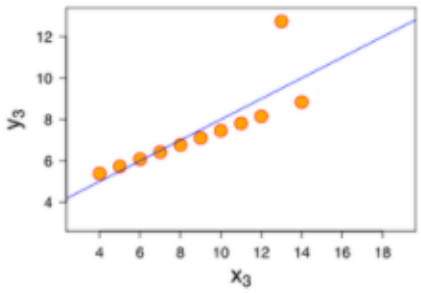
But when plotted they suddenly appear very different

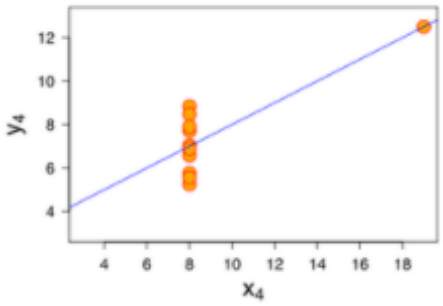


Dataset 1 appears like many well-behaved datasets that have clean and well-fitting linear models



Dataset 2 does not have a linear correlation.





Dataset 3 does but the linear regression is thrown off by an outlier. It would be easy to fit a correct linear model if only the outlier was spotted and removed before doing so

Dataset 4 does not fit any kind of linear model, but the single outlier makes keeps the alarm from going off.

3) What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

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

Feature 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.

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, regardless of the unit of the values.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale.

X_new = (X - X_min)/(X_max - X_min)

It is used when features are of different scales. Scales values between [0, 1] or [-1, 1].

It is really affected by outliers. Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

It is used when we want to ensure zero mean and unit standard deviation. It is not bounded to a certain range.

It is much less affected by outliers. Scikit-Learn provides a transformer called StandardScaler for standardization.

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

VIF = infinity is when R-squared is equal to 1(from the formulae)

This is the perfect correlation condition between 2 variables where we should drop one of perfectly correlated variable causing multicollinearity. It also tells that corresponding variable may be expressed exactly by a linear combination of other variables which shows infinite VIF as well.

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

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.

In the 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.