Assignment-Based Subjective Questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Inferences obtained from the model:

1. Count of total rental bikes (cnt) in year 2019 is 23.5% higher than that in 2018
2. Cnt is 8.8% lower during the holidays
3. Cnt is 10.9% lower in Spring
4. Cnt is 5.8% higher in Winter season
5. Cnt is 5.3% lower in the month of December
6. Cnt is 5.6% higher in the month of January
7. Cnt is 5.9% lower in the month of July
8. Cnt is 5.0% lower in the month of November
9. Cnt is 5.5% higher in month of September
10. Cnt is 29.1% lower when there is light snow or rain
11. Cnt is 8.2% lower when the weather is misty and cloudy

Q2: Why is it important to use drop_first=True during dummy variable creation?
Ans: drop_first=Trueis important to use, asit helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: Feeling temperature (atemp) has the highest correlation with the target variable, unit increase in feeling temperature increases count by 41.2%which is the highest correlation in the model.

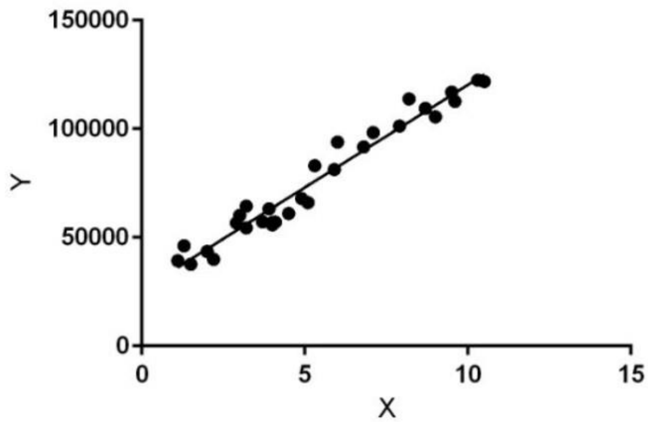
Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans: According to this assumption there is linear relationship between the features and target. This can be validated byplotting a scatter plot between the features and the target, checking the VIF & by error distribution of residuals.

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans:
1. Feeling temperature (atemp): demand increase by 41.2% with unit increase
2. Weather: demand decreases by 29.1% in light snow or rain
3. Year: demand hasincreasedfrom2018to2019by23.5%

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.
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 getting 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.In the figure below, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.



Hypothesis function for Linear Regression :

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

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 θ1and θ2values.**θ1:**intercept**θ2:**coefficient of x
Once we find the best θ1and θ2values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Q3: Explain the Anscombe's quartet in detail.

Ans:Anscombe's Quartetcan 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. There are these four data set plots which have nearly**same statistical observations**, which provides same statistical information that involves**variance**, and**mean**of all x,ypoints in all four datasets.

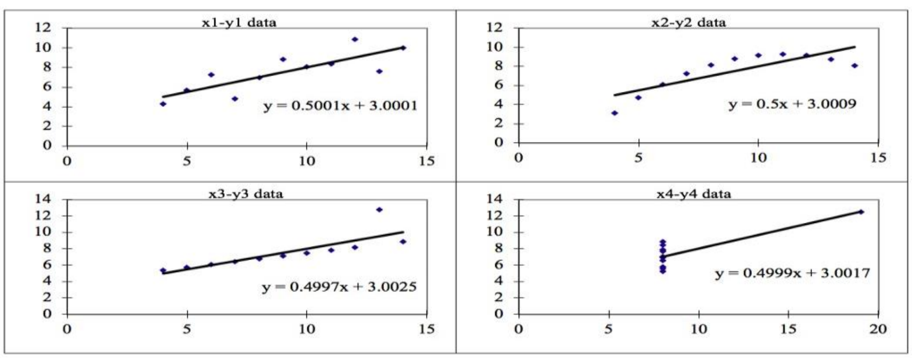
Also, the Linear Regression can be only be considered a fit for the**data with linear relationships**and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

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

The statistical information for all these four datasets are approximately similar and can be computed 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 generates 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:

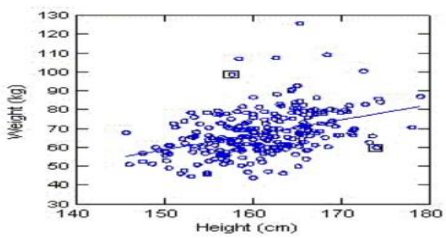
1. **Dataset 1:**this**fits**the linear regression model pretty well.
2. **Dataset 2:**this**could not fit**linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:**shows the**outliers**involved in the dataset which**cannot be handled**by linear regression model
4. **Dataset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

Conclusion:

We have described the four datasets that were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

Q3:What is Pearson's R?

Ans: 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.
"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.



Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: 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.MinMaxScaler**helps to implement normalization in python.

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

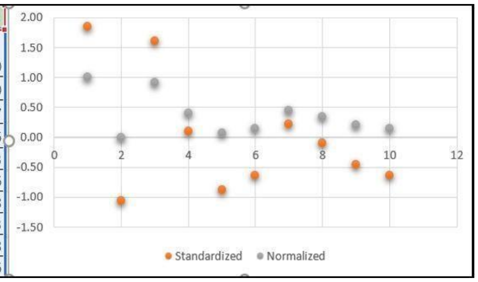
Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)**zero and standard deviation one (**σ**).

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

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**.
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**).

Example:
Below shows example of Standardized and Normalized scaling on original values.

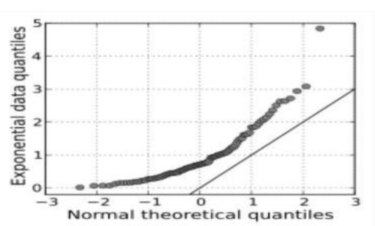
Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans:Ifthereisperfectcorrelation,thenVIF=infinity.Thisshowsaperfectcorrelationbetweentwoindependentvariables.In thecaseofperfectcorrelation,wegetR2=1,whichleadto1/(1-R2)infinity.Tosolvethisproblemweneedtodroponeofthevariablesfromthedatasetwhichiscausingthisperfectmulticollinearity.
AninfiniteVIFvalueindicatesthatthecorrespondingvariablemaybeexpressedexactlybyalinearcombinationofothervariables(whichshowaninfiniteVIFaswell).

Q6:WhatisaQ-Qplot?ExplaintheuseandimportanceofaQ-Qplotinlinearregression.
Ans:Q-QPlots(Quantile-Quantileplots)areplotsoftwoquantilesagainsteachother.Aquantileisafractionwherecertainvaluesfallbelowthatquantile.Forexample,themedianisaquantilewhere50%ofthedatafallbelowthatpointand50%lieaboveit.ThepurposeofQQplotsistofindoutiftwosetsofdatacomefromthesamedistribution.A45degreeangleisplottedontheQQplot;ifthetwodatasetscomefromacommondistribution,thepointswillfallonthatreferenceline.

A Q Qplot showing the 45 degree reference line:



- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q– Q plot will approximately lie on a line, but not necessarily on the line y = x.
- Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.