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 1. Analysis of categorical columns was done using the box plot. Below are the few points we can infer from the visualization :

- More bikes are rented in the summer than in the fall.
- 2019 was a good year for bike rentals compared to 2018.
- Rentals tend to go up in clear and partly cloudy weather.
- Renting a bike is more popular on Saturdays, Wednesdays, and Thursdays.
- Bike renting trend raised in the starting of the year stayed constant during mid of the year and then it increased and as we approached the end of the year it started decreasing.
- When it's a holiday, bike booking seems to be less probably because people might want to spend time at home with their family.

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

Ans 2. To remove the extra column which was created while creating a dummy variable, the drop_first = True is used. This helps in reducing the correlations created among dummy variables.

Syntax – drop_first: bool, default False, which implies whether to get k-1
Dummies out of k categorial levels by removing the first level.

If we have 3 types of values in the categorial column and we want to create a dummy variable for that column. If one variable is not X and Y, then it is obvious Z. So, we do not need 3rd variable to identify the Z.

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

Ans. 3. The numerical variables which have the highest correlation with the target variable were ‘temp’ and ‘atemp’. These two were highly correlated to each other as well.

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

Ans 4. To validate the assumptions of Linear Regression following was done:

❖ Normality of error terms
- Error terms should be normally distributed.

❖ Multicollinearity check
- There should be insignificant multicollinearity among variables.

❖ Homoscedasticity
- There should be no visible pattern in residual values.

❖ Independence of residuals
- No auto-correlation.

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

Ans. 5. Top 3 features contributing significantly towards explaining the demand for the shared bikes are –

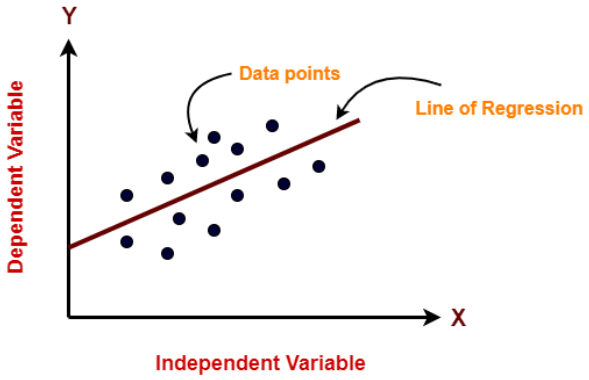
1. Temperature (temp) – with coefficient value - 0.4720
2. Year(yr) – with coefficient value - 0.2342
3. Fall season(Sep) – with coefficient value - 0.0889

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail.

A linear regression algorithm represents a linear relationship between a dependent (y) and one or more independent (y) variables, that's why it is called linear regression. Linear regression displays the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line that represents a relationship between the variables.



For example, we could use the equation to predict weight if we knew an individual's height.

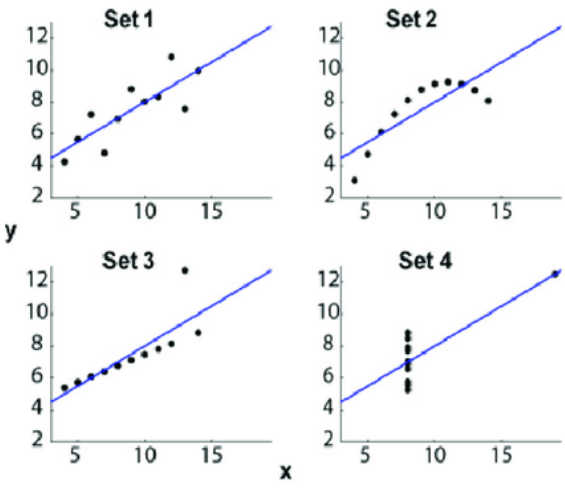
Q 2. Explain the Anscombe's quartet in detail.

Ans 2. There are four datasets in Anscombe's Quartet, which in simple descriptive statistics are nearly identical, but there are some anomalies that lead to errors in a regression model if developed. Scatter plots show differences in their distribution and appearance.

Francis Anscombe constructed it in 1973 as a way of illustrating how important it is to plot graphs before analyzing and building models and to demonstrate the effects of other observations on statistical properties. Four plots with nearly identical observations are shown below, and each plot contains almost the same statistical information, including variance, and mean values for all x, and y points.

Visualizing data can make it easier to determine if there are any anomalies present in the data, such as outliers, diversity, linear separability, etc. To identify anomalous patterns in the data, you must plot the data features to determine the distribution of the samples.

All datasets produce a different kind of plot when plotted on a scatter plot, one that will be unintelligible to any regression algorithm caused by these peculiarities. This can be seen as follows:



Mean of Y	**7.50 in all 4 XY plots**
Sample variance of Y	**4.122 or 4.127 in all 4 XY plots**
Correlation (r)	**0.816 in all 4 XY plots**
Linear regression	**y = 3.00 + (0.500 x) in all 4 XY plots**

Data sets for the 4 XY plots

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

The description of the four datasets is as follows:

1. **Set 1:** Fits well with linear regression.
2. **Set 2:** There was no fitting linear regression model quite well due to the non-linearity of the data.
3. **Set 3:** It represents the outliers that are involved in the dataset. These outliers cannot be handled by linear regression.
4. **Set 4:** It represents the outliers that are involved in the dataset. These outliers cannot be handled by linear regression.

Q 3. What is Pearson's R?

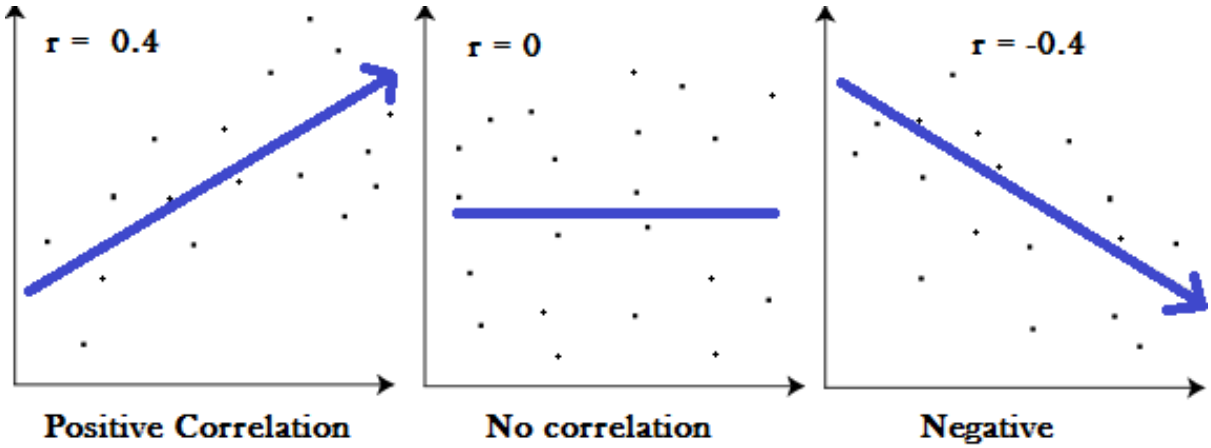
Ans 3. Pearson's R measures how strong is the linear relationship between 2 variables.

Pearson's R always lies between -1 and +1.

The R sign indicates the direction-

1. Positive Correlation – Directly proportional variables. An increase in one variable will lead to an increase in other.
2. Negative Correlation – Inversely proportional variables. An increase in one of the variables will lead to a decrease in other.

The strength of the correlation is by the size of R.
R having a value close to +1 or -1 will have a strong correlation.
R having a value close to 0 will have a weak correlation.



Pearson's correlation coefficient formula -

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

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

Ans 4. It is one of the important steps while processing data, which is used on independent variables to bring the data within a particular range. It also speeds up the calculations while working on any algorithm.

Scaling is performed to bring all the variables to the same level. If scaling is not done, then algorithms will only take up magnitude, which will lead to incorrect modeling. As, we know any data contains features varying in range, unit, and magnitude. So, scaling helps In taking up all the variables together. Scaling does not affect any other parameters like t-statistics, p-values, R-squared, etc.

S. No	Normalization	Standardization
1	It's called Minimum-maximum Scaling	Mean and standard deviation scaling
2	When features are of different scales, we used them to bring them to the same level.	Standardization is used when we want to make sure zero mean and unit standard deviation.

3	Outliers show the effect on normalized scaling	Outliers show less effect on standardized scaling.
4	Scales values in between [0,1] or [-1,1]	No certain range is define
5	Library Sci-kit Learn comes up with a transformer called as Min-Max Scaler for Normalized scaling.	Library Sci-kit Learn comes up with a transformer called Standard Scaler for Standardized Scaling.

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

Ans 5. Sometimes, the value of VIF is infinity when there is a perfect correlation. This means that two independent variables are perfectly correlated. The formula for VIF = $1/1-R_i^2$, when there is a perfect correlation then, $R_i^2=1$, which leads to infinity. So, to solve this problem, we have to finish this multicollinearity by dropping one of the variables from the dataset.

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

Ans 6. When a plot is made from two quantiles against each other then that is a Q-Q Plot (Quantile-Quantile plots). A fraction where values lie underneath that quantile, then that is a quantile.

Use of a Q-Q plot

The Q-Q plot is used for the following:

- Determine whether two samples are from the same population. To check whether 2 samples are from the same population.
- Either both samples have the same tail.
- Whether the distribution shape of both the samples is the same
- Whether both samples have common behavior.

Importance of Q-Q plot

- These Q-Q plots are like probability plots, there is no need to have an equal sample size.
- We don’t have to worry about the dimensions of values to normalize the dataset.