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?

   Answer:

   The following are the categorical values in the dataset.
   Season, yr,mnth,holiday,weekday,workingday and weathersit.
   Based on EDA, we can say that the cnt is high during the middle of the year and reduces during the beginning and end of the year. It also gives us a notion that the deman is high during summer.
   Based on our final model,
   yr has a coefficient of 0.23, i.e. the dependent variable increases by 0.23 times.
   The dependent variable increases for the months Sep and Mar, but decreases for Nov and Dec.
   The working day also has the positive correlation with dependent variable. For weathersit where there is light snow or thunderstorm, the demand for bikes reduces.
   The rest of the categories were either less significant or represented by the above categories.

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

   Answer:

   The general rule while creating a dummy variable is, for N categorical levels, we can represent the same with N-1 dummy variables. To achieve this, we use **drop_first=True,** as this would reduce the dummies by 1 column and also reduces the correlation between those variables.

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

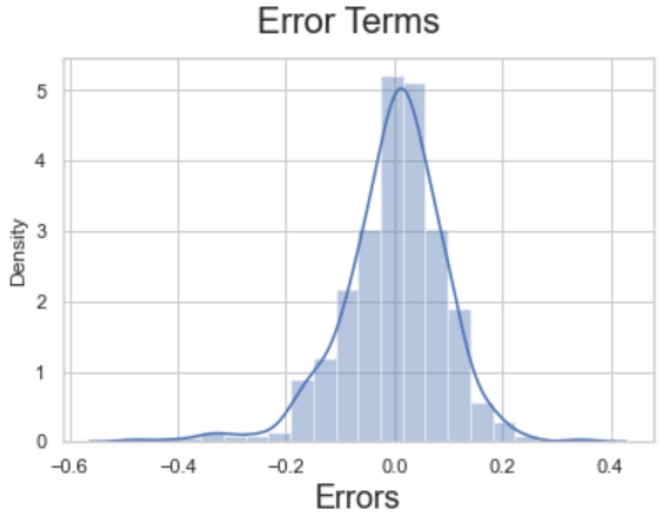
   Answer:

   temp and atemp (0.63) have the highest correlation with target variable

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

   Answer:

   We validate the assumptions by checking if the residuals are normally distributed.



   Also there is no multicollinearity among the individual variables. Their VIF are less than 5.

	Features	VIF
2	temp	4.53
1	workingday	4.48
8	winter	2.35
0	yr	2.03
5	Nov	1.77
7	spring	1.74
9	Sat	1.73
11	Mist	1.53
3	Dec	1.35
4	Mar	1.21
6	Sep	1.17
10	Light	1.06

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

   Answer:

   The top 3 features are temp, weathersit with light snow/rain/thunderstorm and year.

- As temperature increases, more bikes are being rented
- If there are any light weather impacts, the bike demand goes down
- Over the time, i.e. by year the demand increases by 0.23 times

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression algorithm is a regression technique to identify the linear relationship between the independent variables and the target variable.
It is based on the Supervised Learning. It is primarily used for predicting the target/dependent variable based on its relationship with independent variables using a straight line.

Formula for the best fit line is Y = B0 + B1X

Where Y is the dependent variable, B0 is the intercept and B1 is the slope.

Properties of Linear Regression Algorithm

- Regression shows only the correlation and not the causality.
- Linear regression guarantees interpolation and not extrapolation

Assumptions

- There is a linear relationship between X and Y
- Error terms are independent to each other
- Error term follow normal distribution
- There is no multicollinearity among the independent variables

2. Explain the Anscombe's quartet in detail.

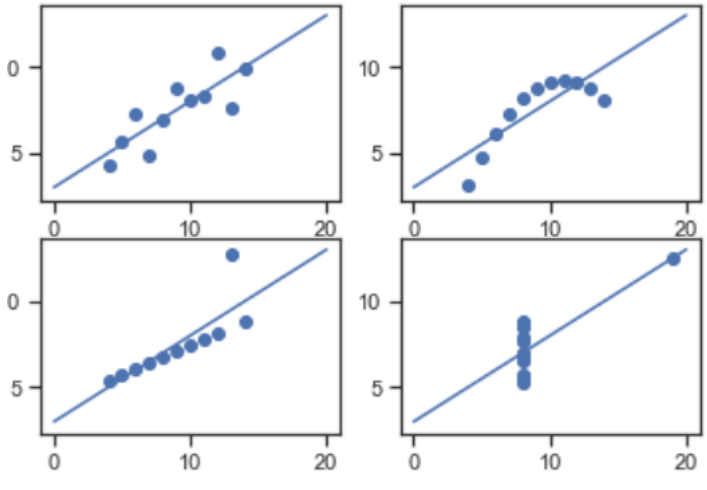
Answer:

Anscombe's quartet consists of 4 data sets with similar statistical descriptions like mean, variance and standard deviation, but have different graphical representations.

The datasets contain 11 (x,y) points.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

For the given data, average of x is 9 and y is 7.5. But when we plot it, it gives us new insight on the data.

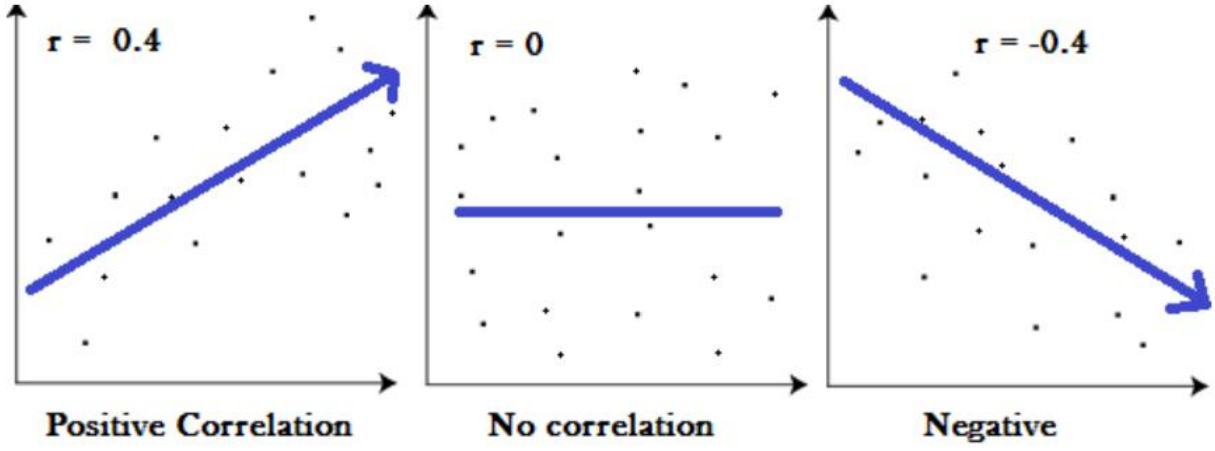


This signifies the importance of plotting the data before coming to any conclusions just based on the statistical data.

3. What is Pearson's R?

Answer:

Pearson's R is a correlation coefficient used in linear regression.
The R determines how strong the variables are dependent on each other.
The value of R lies between -1 and 1. Here a negative value indicates a negative correlation where one dependent increases, the value of the other decreases and vice versa. On the other hand, a positive value indicates a positive correlation that is when once value increases, the dependent also increases and vice versa.
A zero value says there is no relation between the variables.



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

Answer:

Scaling is a technique applied on the independent features to bring their values into a fixed range either using a normalized scaling or standardized scaling. It is a data preprocessing step performed while developing linear regression model.

If the features we have, have different magnitude range or units, the algorithm will consider the magnitude or bigger numerical value and does not consider the scale. Hence it is very important to bring all the features into same level by scaling them

A normalized scaling bring all the values into the range of 0 and 1. The standardization on the other hand bring all the data into a standard normal distribution with mean 0 and standard deviation 1. The normalization loses some information about the outliers in the data where a the standardization doesn't.

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

Answer:

The value of VIF turns infinite when two variables are perfectly correlated. That is their correlation coefficient is equal to 1. When the value of R2 is 1, the equation for VIF becomes 1/(1-1) and hence we get infinite values are result. Hence it is required to remove either one of the variables from the model to avoid multicollinearity.

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

Answer:

Q-Q plot or Quantile-Quantile plots are plots of two quantiles against each other. The purpose of a QQ plot is to identify whether the 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.

In linear regression, it is used to check if the points lie approximately on the line, and if they don't, the residuals do not follow normal distribution and thus your errors aren't either.