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. From the box plot graph for all categorical variables with cnt variable we can infer that:

1. The demand was most in Fall season and least in Spring season. Also, the 75 percentiles of the demand in spring season lies below the 25 percentiles of summer, fall and winter seasons.
2. In 2019, the demand for bikes were much higher than demands in 2018. 75 percentiles demand in 2018 approximately equals 25 percentiles of demand in 2019. Also, the demand increased between August and October in 2019 and declined sharply after that in comparison to 2018. Rest of the pattern seems same.
3. For mnth variable, in Jan, Feb, Nov and December, the demand was comparatively less than rest of the months showing less demand in winter.
4. Weekday does not make much of a difference in bikes demand. The 50 percentiles are approximately equal in all scenarios.
5. In weathersit feature, the demand was least in "lightRain" weather and most in “clear” weather which is expected behavior. In Light rain weather situation, the 75 percentiles demand lies below 25 percentile demand in Mist or clear weather situation.
6. In holidays, the 50-percentile demand equals 25 percentile demand when there are no holidays showing that the demand is more when there are no holidays.
7. Season, weathersit features influence demand considerably.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. While creating dummy variables for categorical features, the number of features that can explain the feature is n – 1 where n is the number of classes in a feature.

drop_first= True feature removes the first column after dummy variable creation and hence the bare minimum number of features required for that feature is met.

It 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. “atemp” feature has the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. I performed below analysis to validate validate the assumptions of Linear Regression after building the model on the training set:

1. No or little multicollinearity.
   Could see that the VIF for all variables for the chosen model was less than 5 with maximum value 4.53 for windspeed feature.

2. Multivariate normality

   Could see the distplot for Errors vs Density followed normal distribution with mean around 0. Hence, error distribution follows normal distribution.

3. Homoscedasticity

   The residuals are distributed evenly above and below the 0 residual so the model is trained correctly and Homoscedasticity is achieved.

4. No auto-correlation

   The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample. Values from 0 to less than 2 point to positive autocorrelation and values from 2 to 4 means negative autocorrelation.

   The value for Durbin Watson (DW) statistic in lr_4 model that we chose is 1.990 hence there is no autocorrelation between the variables.

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

Ans. Independent variables with lowest P values are most significant. Variables are:

1. Temp (P value: 0.00)
2. Yr (P value: 0.00)
3. Windspeed (P value: 0.00)

Note: winter and mist also have P values equal to 0.00

General Subjective Questions on next page.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a supervised machine learning algorithm. Linear regression is one of the very basic forms of machine learning in the field of data science where we train a model to predict the behaviour of your data based on some variables. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

Linear regression is used to predict a quantitative response Y from the predictor variable X. Mathematically, linear regression equation can be expressed as:

$y = mx + c$

m = Slope of the line.

c = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

To create a model, algorithm must "learn" the values of these coefficients (m and c). And once we have the value of these coefficients, we can use the model to predict the target variable (y). The main aim of the regression is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the points to the regression line.

Linear regression can be classified into 2 types:

1. Simple / univariate Linear regression: Where the dependent variable is dependent on only one independent variable.
2. Multiple Linear regression: Where the dependent variable is dependent multiple independent variable.

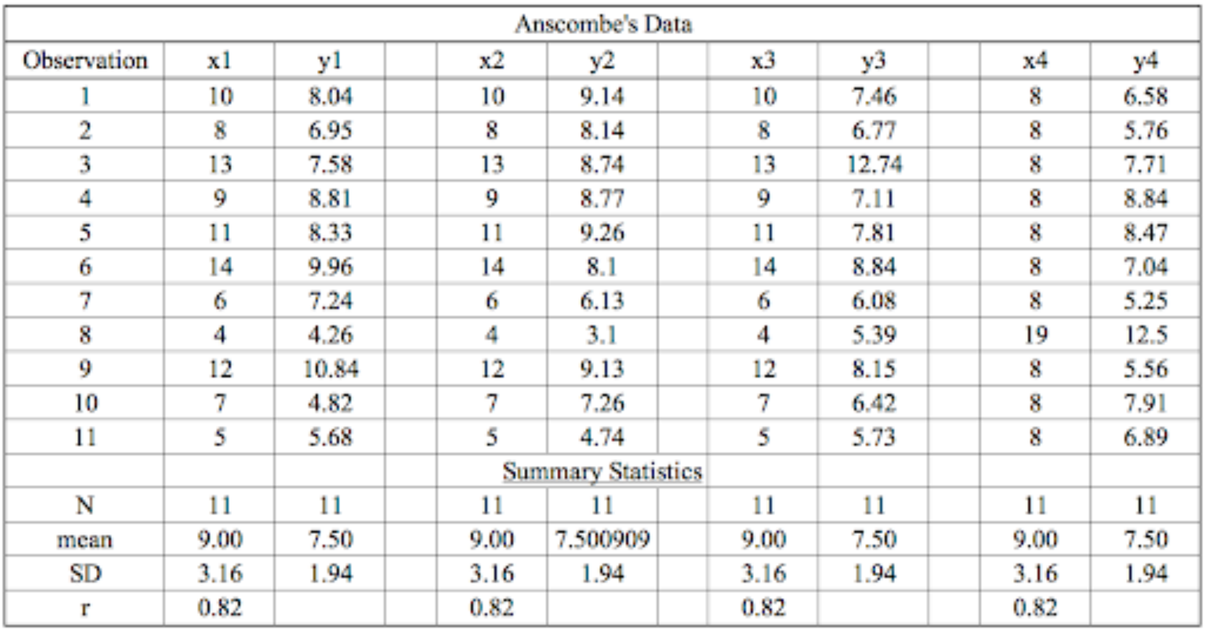
The regression has below key assumptions:

- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

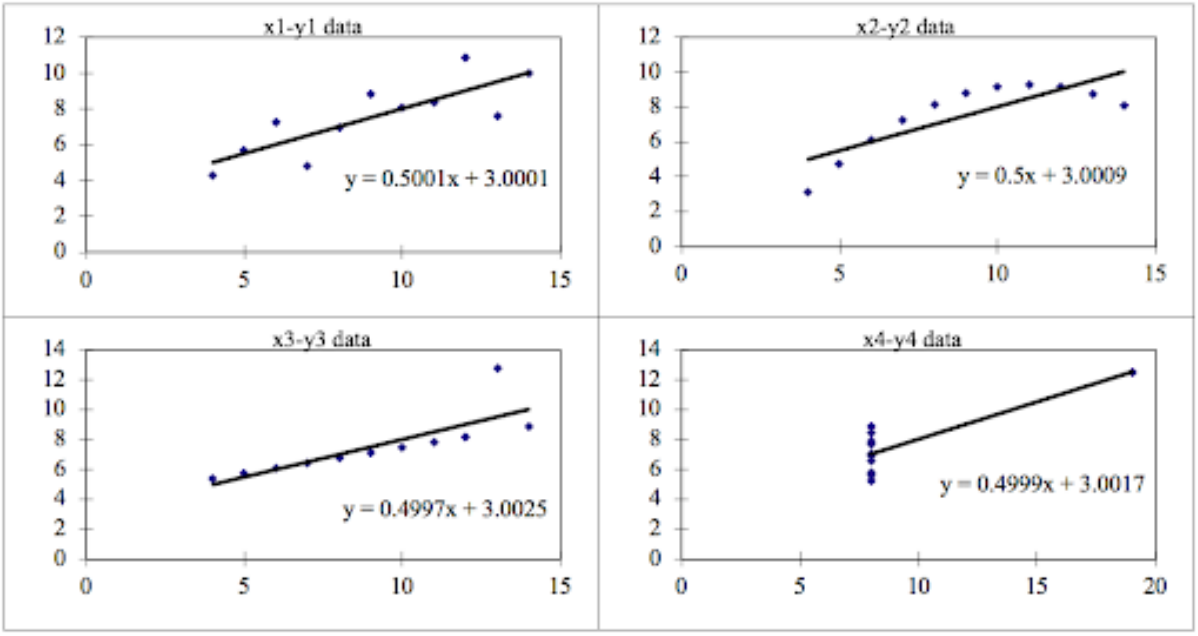
The statistical information for these four data sets are approximately similar. We can compute them as follows:



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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm.

Q3. What is Pearson's R?

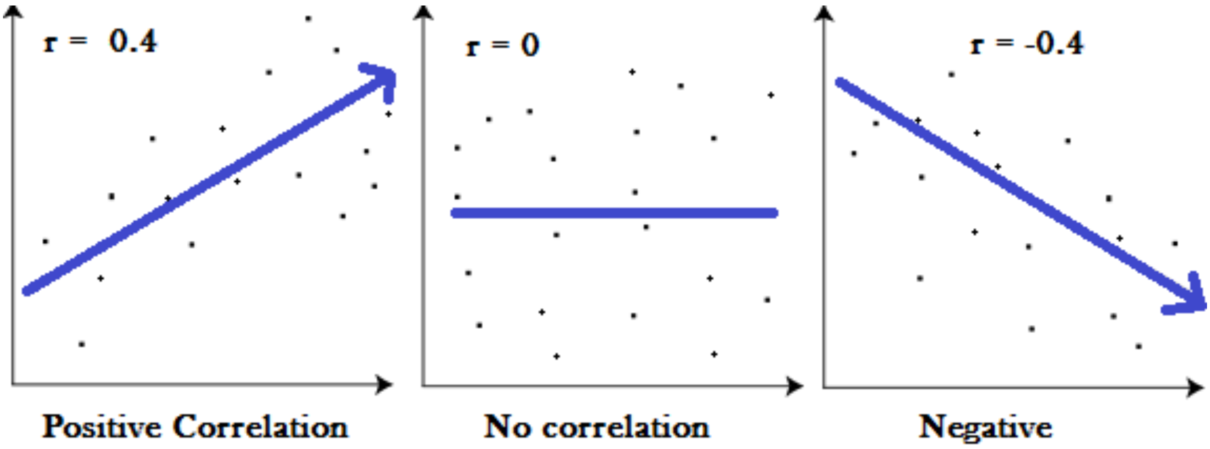
Ans. Pearson's R is a correlation coefficient commonly used in linear regression. Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

Pearson's correlation coefficient formula is :

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

It shows the linear relationship between two sets of data.



Positive Correlation

No correlation

Negative

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a method used to normalize the range of independent variables or features of data.

For example — if you have multiple independent variables like age, height, and weight; With their range as (18–100 Years), (1–2 Meters), and (50-150 Kgs) respectively, feature scaling would help them all to be in the same range, for example- centred around 0 or in the range (0,1) depending on the scaling technique.

Normalization

It is also called min-max scaling. It rescales the range of a feature between 0 and 1. General formula is:

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

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

It gets affected by outliers.

Normalization is good to use when the distribution of data does not follow a Gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours.

Standardization

Standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.

It does not get affected by outliers much.

Standardization **can be** helpful in cases where the data follows a Gaussian distribution. Though this does not have to be necessarily true.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. The formula for VIF is 1/(1- r2), where r2 is the percentage of the variance in the individual independent variable (IV) that the set of IVs explains.

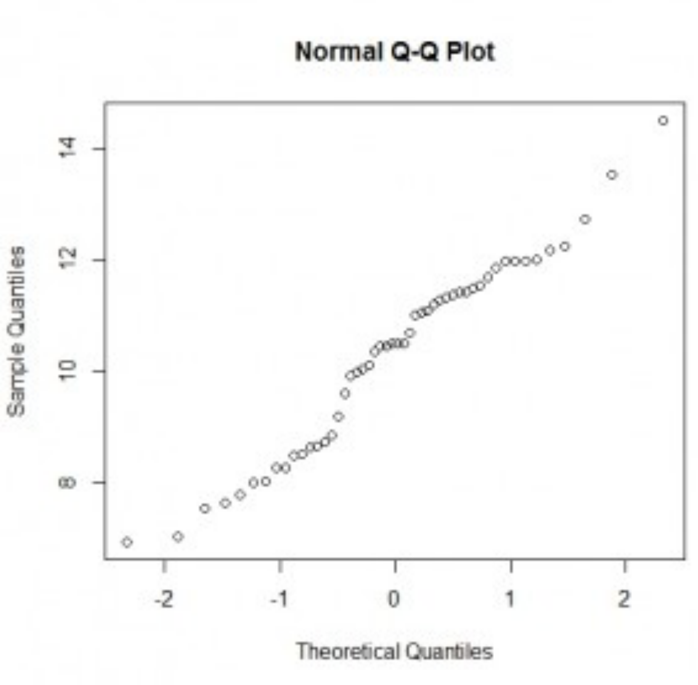
VIF infinity means the value of r2 equals 1. This means that there is a perfect correlation between independent variables. Multicollinearity is definitely present and the IVs are highly co-related to each other.

When this occurs, the variable with infinite VIF should be dropped and then the model should be retrained and the VIF should be calculated again for independent features used to train new linear model.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The points in the Q–Q plot will approximately lie on the line y = x. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal 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.