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?**

Ans: As per my analysis of the categorical variables of the dataset I can draw below observations:

a. The greatest number of bookings happen in the fall season and there is a significant increase in the bookings in the year 2019 when compared to 2018.
b. The mid-year time frame (June – September) shows higher bookings as compared to other months of the year and this data also shows significant increase in the bookings in 2019 when compared to 2018.
c. Sunday has the lowest booking as compared to other days of the week probably due to people spending time at home due to weekly off. This data also shows huge increase in the bookings in 2019 when compared to 2018.
d. When the weather is clear then there are maximum bookings. Here also there is an increase in the bookings in 2019 when compared to 2018.
e. There are lesser number of bookings on a holiday.
f. Booking is almost same whether it's a working or non-working day. This data also shows increase in booking numbers in 2019 compared to 2018.

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

Ans: drop_first=True is used to reduce the extra columns that are created during dummy variable creation in case of categorical data analysis. If there are 3 states for any variable, then we don't need 3 columns to represent it. This can be done using 2 columns only. So to generalize, if we have categorical variable with n-levels then we need n-1 columns to represent the dummy variables.

For eg: If we have two categories to denote gender (male & female) then there are 2 states and this is denoted as :

Male	Female
0	1
1	0

But this can also be represented as below (Female as 1)and if its not 1 then its 0 which is male so only 1 state is actually required.

Female
1

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

Ans: 'temp' column has the highest correlation with the target variable.

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

Ans: The assumptions for Linear Regression Model have been validated on training data as below:

a. Linear Relationship: There exists a linear relationship between independent variable and predictor variables.

b. Homoscedasticity: The variance of residuals is constant along the dependent variable.

c. Absence of Multicollinearity: The variables shouldn't have multicollinearity.

d. Normal distribution of Errors: Residual error have normal distribution.

e. No Autocorrelation

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

Ans: Based on the final model the top 3 features that contribute towards explaining demand of shared bikes are:

a. temp

b. year (yr)

c. Sep (September month data)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans : Linear regression is a statistical method that is used for predictive analysis. It can be used to analyse the linear relationship between a set of predictor variables and an independent variable. It can be used to make predictions for continuous or categorical variables.

Linear relationship between variables means that when the value of a dependent variable changes linearly as per the change in one or more independent variables. This linear relationship can be either positive or negative i.e. the value of dependent variable can increase with increase in independent variable or vice versa.

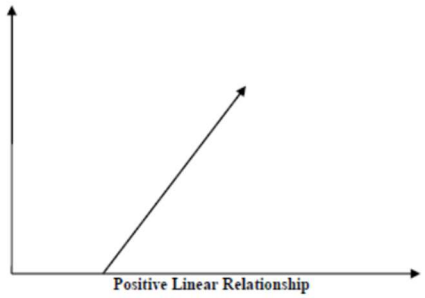
Mathematically the relationship can be represented with the help of following equation -
$Y = mX + c$
Here, Y is the dependent variable.
X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

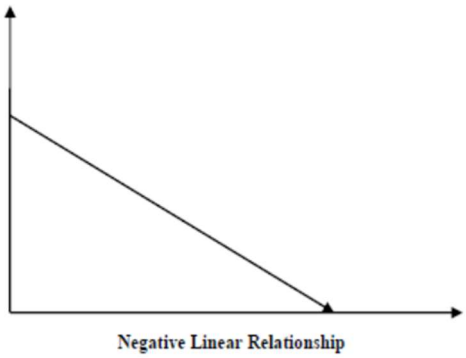
The linear relationship can be either positive or negative –

Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases or decrease in sync with each other.



Positive Linear Relationship

Negative Linear relationship: A linear relationship will be called negative if there is a negative relationship between independent variable and dependent variable i.e. when one variable increases the other decreases and vice-versa.



Negative Linear Relationship

Linear regression is of the following two types -
Simple Linear Regression: If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Eg: y = a0 +a1 x (here there is only one independent variable x)

Multiple Linear Regression: If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Eg: y = a0 +a1 x1 + a2 x2 + a3 x3 (here there are multiple independent variable x1, x2, x3)

There are certain assumptions associated with Linear regression as below:
Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Multi-collinearity occurs when the features have dependency in them.
Auto-correlation – There is very little or no auto-correlation in the data.

Linear relationship between variables – The relationship between independent & dependent variables must be linear.
Normality of residual error – Residual error should be normally distributed
Homoscedasticity – The variance of residuals is constant along the dependent variable

2. Explain the Anscombe's quartet in detail.

Ans:

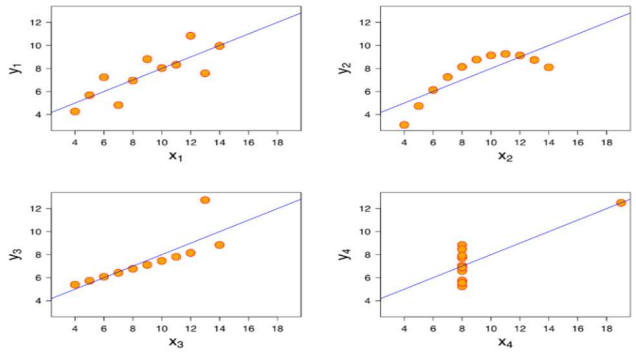
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises of four datasets that have nearly identical simple statistical properties and each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. The data is shown in below table.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics as show in the below part of the table are all identical for all 4 datasets

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Std deviation of x is 3.32 and y is 2.03 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset can be inferred differently:



- Dataset I shows almost linear relationship and is a well-fitted linear models.
- Dataset II is not distributed normally.
- Dataset III has linear dependency but there is a clear outlier.
- Dataset IV shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset

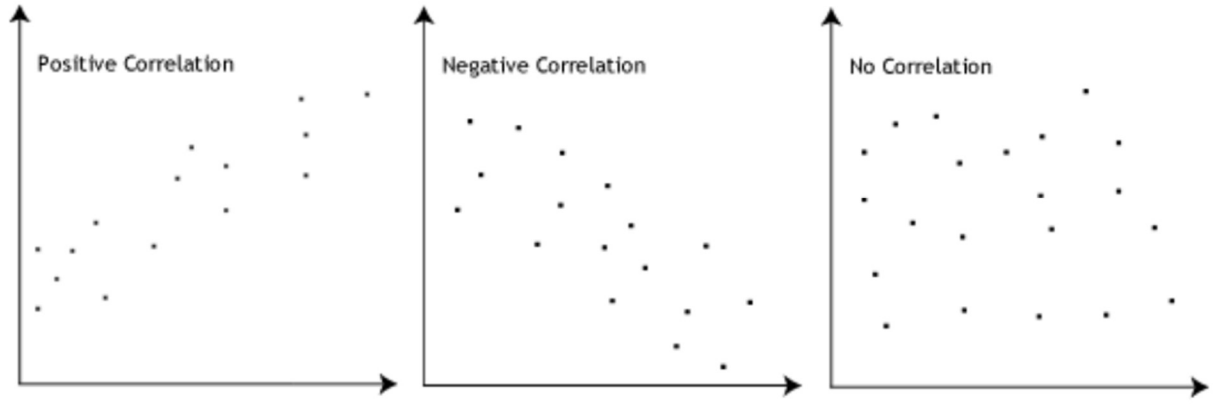
3. What is Pearson's R?

Ans:

Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. It is a numerical summary of the strength of the linear association between the variables. The correlation coefficient will be positive if the variables increase/decrease together. The correlation coefficient will be negative If the variables increase/decrease in opposition with low values of one variable associated with high values of the other.

In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Ans : Scaling is a technique that is used to normalize the range of independent variables or features of a data, generally done to bring the data distribution in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. Generally the collected data varies

highly in terms of magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling happens. 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.

Normalized Scaling:

Also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [-1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x is an original value, x' is the normalized value and xmin and xmax are the max and min values in the dataset.

To rescale between a specific range of values [a, b], the formula becomes:

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

where a, b are the min-max values.

Standardized Scaling:
Standardized scaling is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

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

Where x is the original feature vector, $\bar{x}$ = average (x) is the mean of that feature vector, and σ is its standard deviation. In this case, the values are not restricted to a particular range.

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

Ans: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a high correlation between the variables. When the value of VIF is infinite it shows a perfect correlation between the independent variables.

$$VIF = \frac{1}{1 - R^2}$$

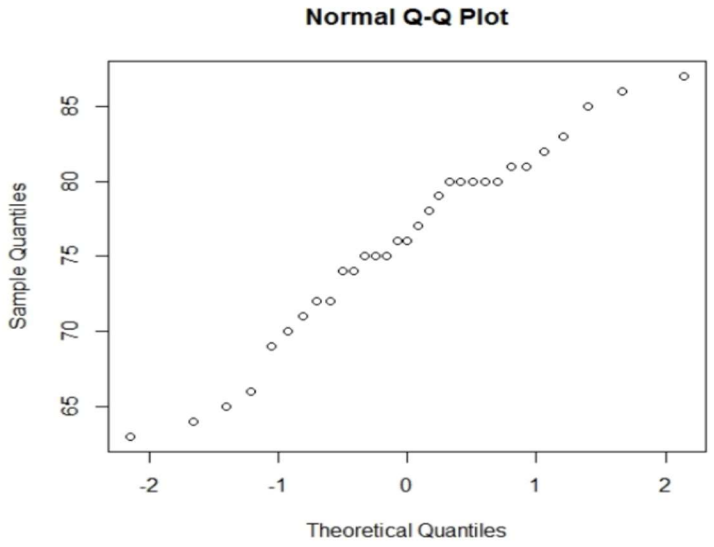
In the case of perfect correlation, we get R-squared (R2)=1, which lead to 1/ (1-R2) as infinity. To overcome this situation multicollinearity among the variables should be analyzed and the required variables should be dropped the dataset.

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

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot of 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. It determines how many values in a distribution are above or below a certain limit. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Q-Q plots take the sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of the sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.

Q-Q plot is very useful to determine:

- If two populations are of the same distribution
- If residuals follow a normal distribution.
- Skewness of distribution