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?

Below are the categorical variables present in the given dataset,

a. season
b. mnth
c. yr
d. weathersit
e. weekday
f. workingday
g. holiday

season: Fall has highest number of bookings, the bookings are in good numbers for summer and winter as well.

mnth: Most of the bike bookings were happening in the months 5 to 10 months with a median above 5000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

yr: Booking has increased year over year

weathersit: There are more bookings when whether is clear with a median close to 5000 booking, followed by other weathersit's. This indicates, weathersit also can be a good predictor for the dependent variable.

weekday: There is not much difference in bookings for weekdays. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday: Bike rental median for working/non-working days are almost close to each other

holiday: Most of the bookings happen during non-holiday and hence it cannot be a good predictor

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

drop_first=True is important to us as, it helps in reducing the extra column created during dummy variables which in turn creates dummy variable trap. Dummy variable trap might lead to multicollinearity issues among dummy variables. This may lead to violation of assumptions of linear regression.

Hence if we have categorical variables with n levels, we only use n-1 levels while dummy variables encoding.

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

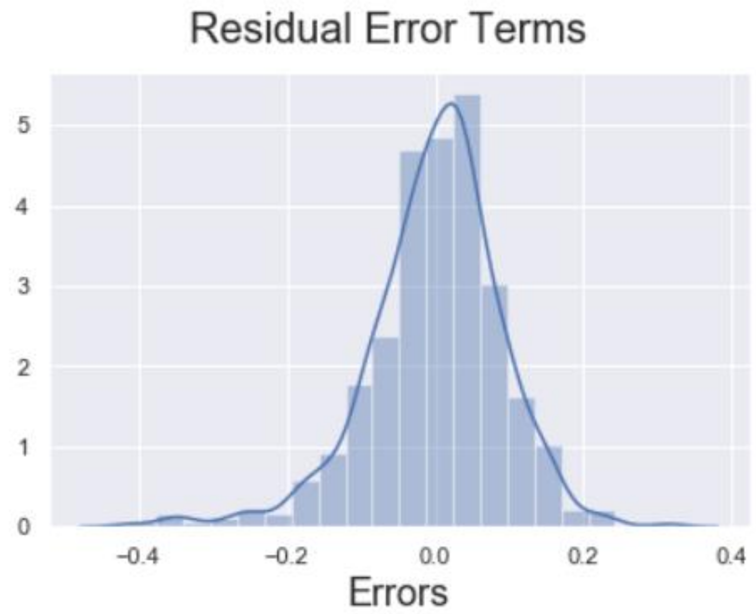
temp and atemp are the two numerical variables which are highly correlated with the target variable.

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

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

Residual Analysis:

We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms and this is what it looks like:



The residuals are following the normally distribution with a mean 0.

Linear relationship between predictor variables and target variable:

This is happening because all the predictor variables are statistically significant.

Also, R-Squared value on training set is 0.836 and adjusted R-Squared value on training set is 0.832. This means that variance in data is being explained by all these predictor variables.

Error terms are independent of each other:

Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

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

According to the final Model, the top 3 predictor variables that influences the bike booking are:

temp: A coefficient value of '0.5499' indicated that a unit increase in temp variable, increases the bike hire numbers by 0.5499 units.

yr: A coefficient value of '0.2331' indicated that a unit increase in yr variable, increases the bike hire numbers by 0.0.2331 units.

season_winter: A coefficient value of '0.1318' indicated that w.r.t season_spring, a unit increase in season_winter variable increases the bike hire numbers by 0.1318 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Mathematically the relationship can be represented with the help of following equation –

$$\boxed{Y = mX + c}$$

Here,

Y is the dependent variable we are trying to predict.

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.

Furthermore, the linear relationship can be positive or negative in nature

The assumptions of linear regression are:

a. There is a linear Relationship between dependant and independent variables
b. Error values (ε) are normally distributed for any given value of X that means Error terms are normally distributed around zero.
c. Constant variance assumption: It is assumed that the residual terms have the variance, σ2, this assumption is also known as the assumption of homogeneity or homoscedasticity.
d. Independent error assumption: residual terms are independent of each other, i.e. their pair-wise covariance is zero.
e. The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

Use Cases of Linear Regression:

Prediction of trends and Sales targets: To predict how industry is performing or how many sales targets industry may achieve in the future.

Price Prediction: LR is used in price prediction – we can predict the change in price of product.

Risk Management: LR is used in Risk Management in the financial and insurance sector.

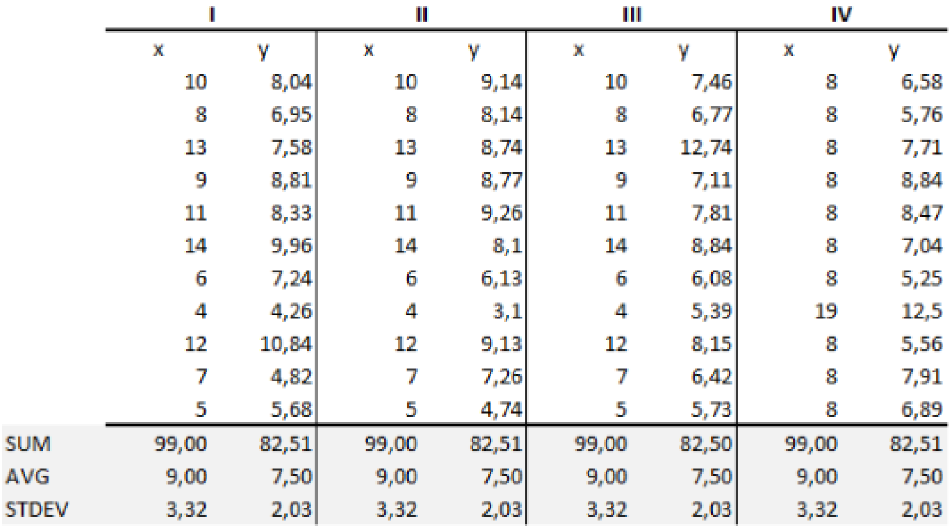
2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

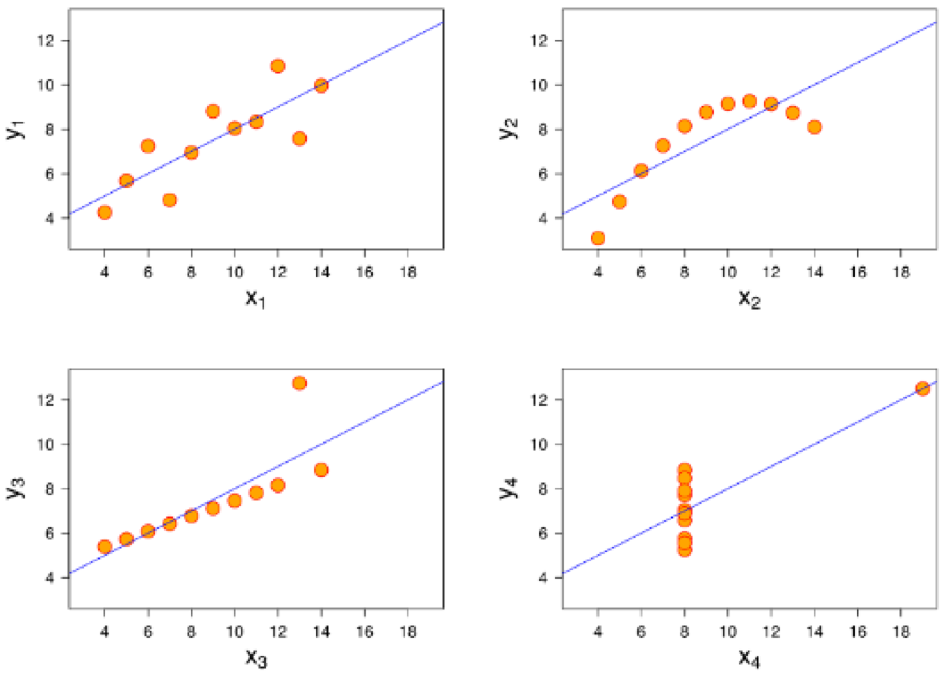
The summary statistics show that the means and the variances were identical for x and y across the groups:

a. Mean of x is 9 and mean of y is 7.50 for each dataset.
b. Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

c. The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset



	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



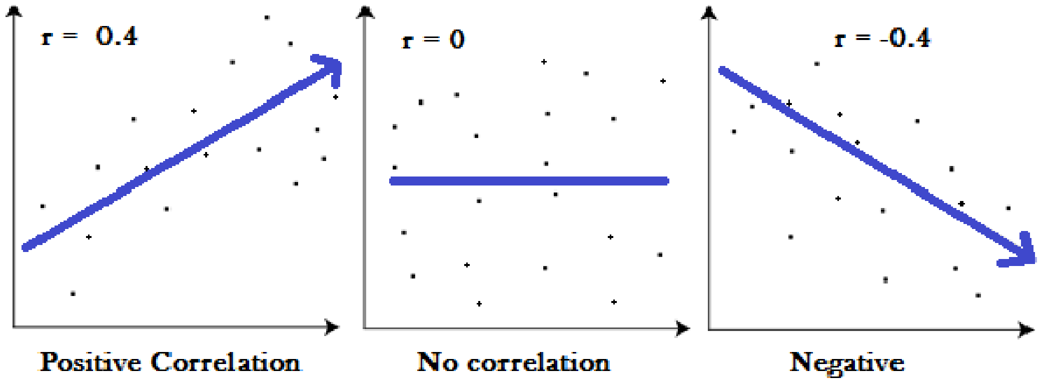
When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well, but each dataset is telling a different story:

a. Dataset I appear to have clean and well-fitting linear models.
b. Dataset II is not distributed normally.
c. In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
d. Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

These datasets are created intentionally 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.

3. What is Pearson's R?

Pearson's R or correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationship or correlation.



a. A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
b. A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.
c. Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The absolute value of the correlation coefficient gives us the relationship strength.

The larger the number, the stronger the relationship. For example, |-.95| = .95, which has a stronger relationship than .55.

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

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

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalization:

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

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

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

Standardization:

Feature 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.

Disadvantage of Min-max scaling over standardization: Normalization lose the information of outliers

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

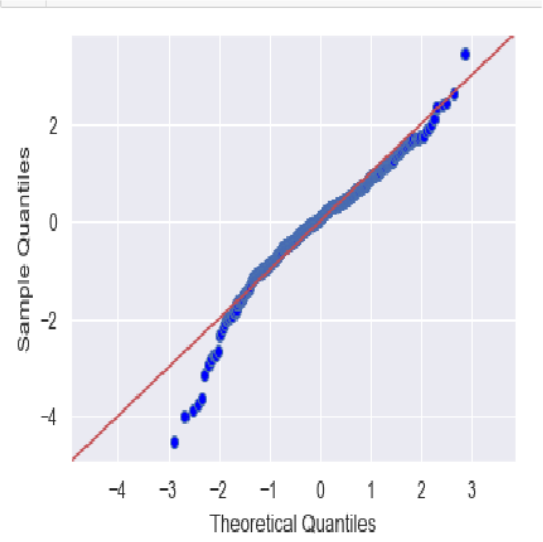
Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

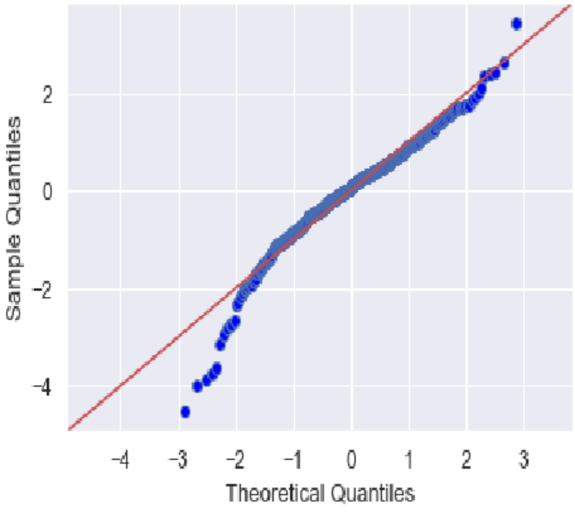
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

A Q-Q plots are used to check the following scenarios:

a. If two data sets, come from populations with a common distribution
b. If two data sets have common location and scale
c. If two data sets have similar distributional shapes
d. If two data sets have similar tail behavior

QQ plot looks like:


Sample Quantiles
Theoretical Quantiles


Sample Quantiles
Theoretical Quantiles