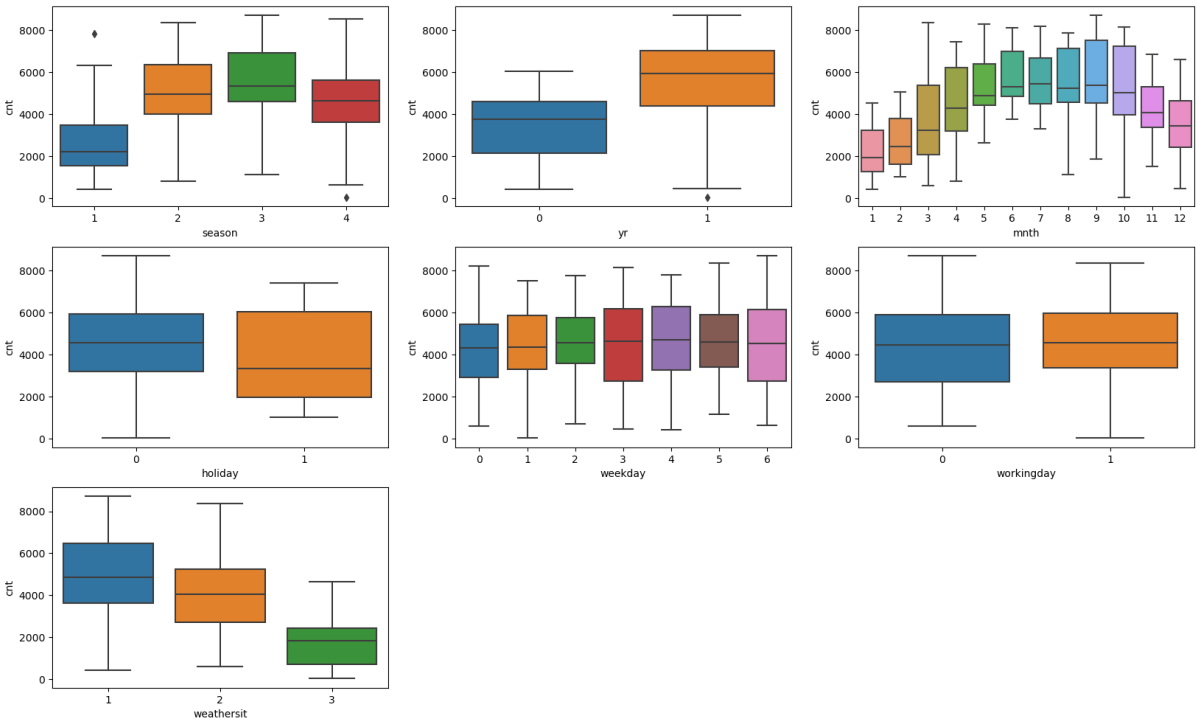
Assignment-based Subjective Questions

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 the box plots of total no of bike rentals with respect to different categorical level (shown below), we can say that the independent variables like season (1: spring, 2: summer, 3: fall, 4: winter), yr (0: 2018, 1:2019), month (1 to 12), weather situation (1: misty, 2: mist_cloudy, 3: light_snow) have impact on number of rentals. It also looks like, on holidays (1 if holiday, else it is 0) no of bike rentals increases but the effect is not much.



Similar results have also been displayed by the fitted model.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0995	0.032	3.148	0.002	0.037	0.162
yr	0.2380	0.009	26.796	0.000	0.221	0.255
temp	0.4799	0.036	13.505	0.000	0.410	0.550
windspeed	-0.1724	0.027	-6.393	0.000	-0.225	-0.119
spring	-0.0530	0.022	-2.371	0.018	-0.097	-0.009
summer	0.0653	0.015	4.263	0.000	0.035	0.095
winter	0.0873	0.018	4.869	0.000	0.052	0.123
September	0.0848	0.018	4.844	0.000	0.050	0.119
clear	0.0941	0.009	10.202	0.000	0.076	0.112

As per the above model parameters, effect of season, month, yr and weather condition on number of rentals is significant. Whatever, impact of holidays or working days we saw in the box plots, is by random chance.

Also, from the above coefficients, we can tell that the impact of categorical variables are pretty less when compared with temperature and year; however, their impact is still not just by chance and can be stated as below.

Season: With respect to Fall, the no of rentals on average decreases by 0.0530 units in Spring. With respect to Fall, the no of rentals on average increases by 0.0653 units in Summer. With respect to Fall, the no of rentals on average increases by 0.0653 units in Winter.

Month: With respect to April, the no of rentals on average increases by 0.0848 units in September.

Yr: Compared to year 2018, the no of rentals on average decreases by 0.238 units in year 2019.

Weather situation: Compared to weather with light snow and rain, the no of rentals on average increases by .0941 units when it is clear weather. Although, the year also looks like a promising factor which implies that with coming year the demand in bike rentals gets increased, but the dataset contains only data of year 2018 and 2019, more recent data should be analysed to know current trend.

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

Ans. –

The drop_first=True is an argument in get_dummies() function of Pandas library. Where function get_dummies() creates binary dummy variables based on the levels of input categorical level, the drop_first=True argument drops the first dummy variable created from the level which comes first in alphabetic order. Hence, if a categorical variable has n levels, this argument will create n-1 dummies as shown below.

Season is categorical variable with 4 levels and 5 rows as shown below.

Season
Spring
Fall
Summer
Winter
Summer

If drop_first=False, dummy variables will be created with binary levels (0/1) as below for above 5 rows. If the row has value Spring, then the “Spring” dummy variable has value 1 for same row, for other rows it will have 0. Same goes for other dummy variables or level.

Spring	Fall	Summer	Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	1	0

Now, we can clearly see that if any n^{th} variable is 1, the other n-1 variables are having 0 only. We can extract the same information from n-1 variables also as shown below.

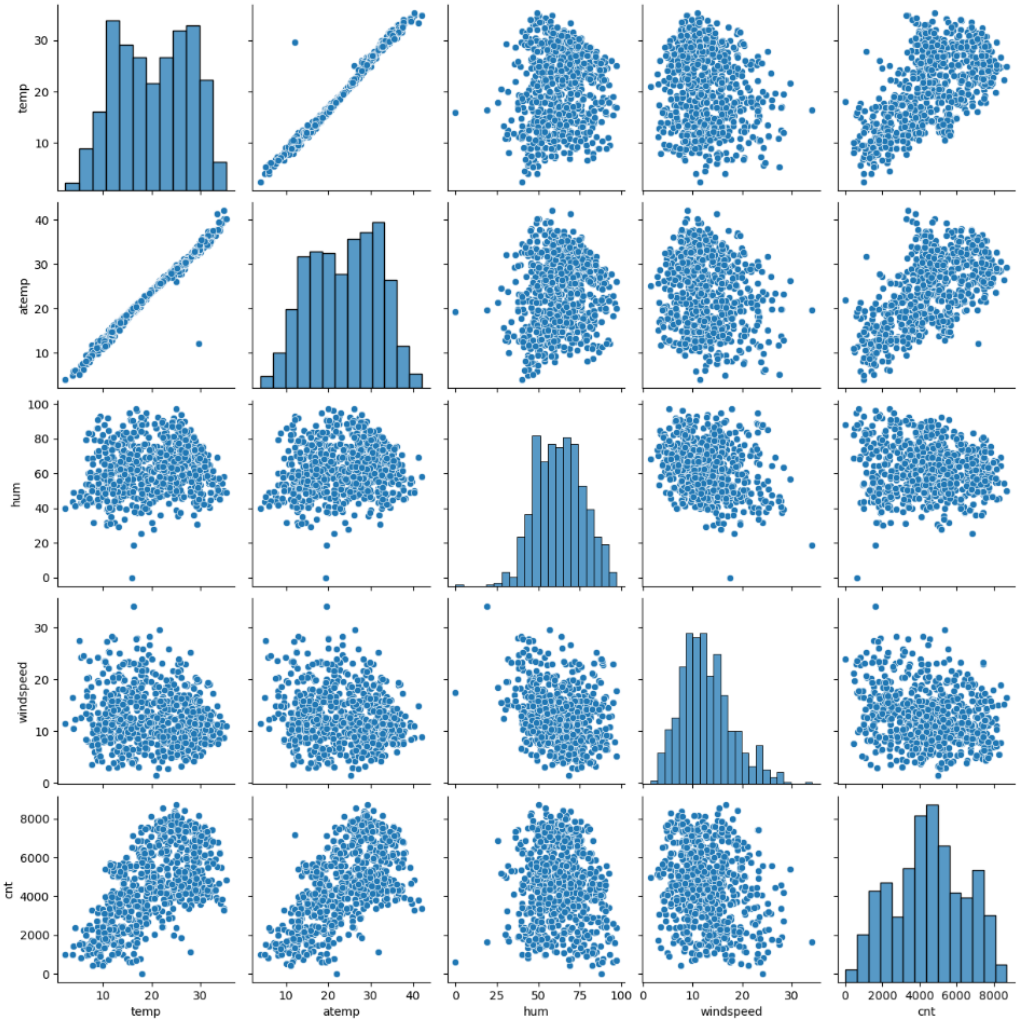
Spring	Summer	Winter
1	0	0
0	0	0
0	1	0
0	0	1
0	1	0

If for any row, all n-1 variables are zero, that means it belongs to the dropped variable/level. From this, we can also say that information of n^{th} variable can be extracted from other n-1 variables. Having all n dummy variables is not just redundant but also induce multicollinearity. To avoid this, we use argument drop_first=True, so that the function can only generated n-1 no of variables instead n variables.

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

Ans. –

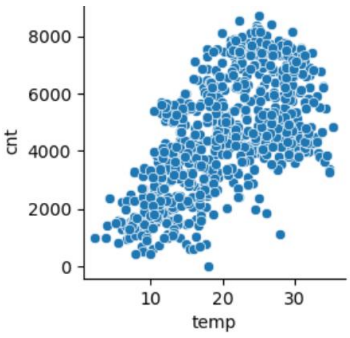
We have plotted the numerical independent variables with the target variables as shown below. Based on this, the weather temperature has highest correlation with the target variable.



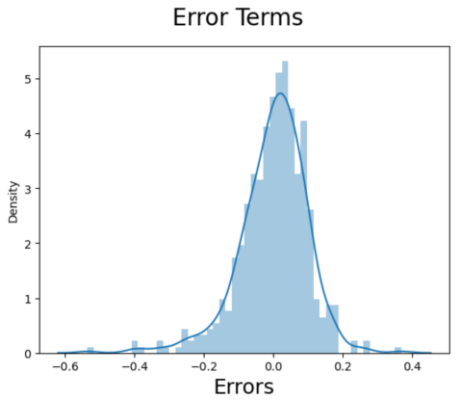
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. –

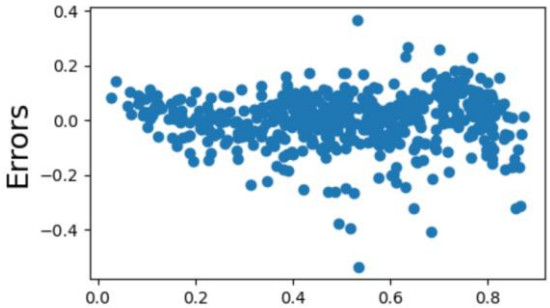
We plotted the feature variables with the dependent variable on a scatter plot and found some of features specially the weather temperature (as shown below) has linear relationship with total no of bike rentals. This proves the assumption of linearity.



Once the final model is fit, we calculated the residual errors (actual – predicted value of response variable) based on training dataset and their distribution is plotted. The plot shows a normal distribution with a mean around zero.



We have also visualized the error terms with respect to predicted value of response. It has been observed that error terms are having slightly increasing variance with increasing predicted value (a funnel like structure). One of the assumptions of linear regression is that the error terms should be homoscedastic (constant variance). However, this is not completely fulfilled, indicating the presence of non-linearity in the data.



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

Ans. –

The top three features, contributing significantly towards explaining the demand of the shared bikes are chosen based on their coefficient's absolute value (in descending order). These are given as below.

Temperature

Year

Windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. –

The linear regression is a supervised machine learning algorithm where the predictor variables are assumed to be linearly related with response variable. The response variable must be a numerical variable in this case. The effect of each feature on the response variable is additive.

The linear regression equation is given as follows,

$y = b + m_1x_1 + m_2x_2 + \ldots\ldots\ldots m_nx_n$

where:

- y is the dependent variable
- x_1, x_2, …, X_n are the independent variables
- b is the intercept coefficient
- m_1, m_2, …, m_n are the slopes or coefficient

The goal of this algorithm is to find the best fit line from a given dataset of dependent and independent variables which gives minimum mean of residual sum of square. This is called ordinary least square method (OLS).

The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

We utilize the cost function to compute the best values in order to get the best fit line since different values for weights or the coefficient of lines result in different regression lines.

The cost function is nothing but the mean of residual sum of squares or Mean Squared Error (MSE), which can be written as

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

m is a matrix of m_1, m_2….m_n.

The coefficients are determined by using gradient descent method. This an optimization technique which finds the values of coefficient where the cost function achieves its minima.

However, a linear regression model is only reliable when the below assumptions are true.

1. The independent variables should be linearly related with dependent variable.
2. The error terms (difference between actual response value and predicted response value) are normally distributed.
3. The error terms should have constant variance (homoscedastic).
4. The error terms should be independent with each other.
5. The independent variables should not be correlated with each other (this has no impact on prediction power of the model; however, this leads to unreliable model parameters or coefficients)

2. Explain the Anscombe's quartet in detail.

Ans. –

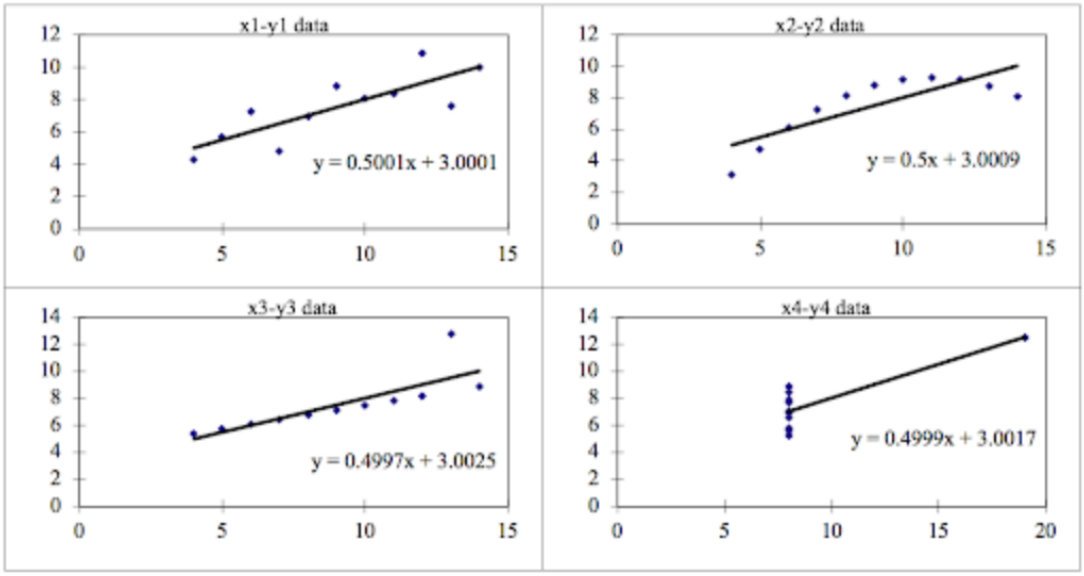
Anscombe's quartet is a set of four datasets (4 pairs of response and feature variables) with similar summary statistics (like mean, variance, correlation coefficient) but having different representations when we draw scatter plots on a graph.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of visualizing data before we analyze and build a model. These four data sets have nearly the same statistical observations for x and y points. However, when we plot these data sets, they look very different from one another. This tells us about the importance of visualizing data before applying various algorithms to build models.

Four different datasets are given below, which has same summary statistics.

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	

If we generate scatter plots using above data as below, we will realize that there is no common regression algorithm that will fit all these datasets.



Hence, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R?

Ans. –

the Pearson correlation coefficient (PCC)(R_{xy}) is a correlation coefficient that measures linear correlation between two sets of data or two variables. It also shows the direction of the correlation. If the both the variables increases or decreases together, then the coefficient have positive value. If one increases and the other decreases or vice versa, the coefficient becomes negative. The value of the same ranges from 0 to 1. This higher the value is, the stronger association between two variables. However, correlation do not define any cause-and-effect relationship.

Also, for a simple linear regression, the coefficient of determination R^2 equals to square of R_{xy} .

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

Ans. –

Scaling a method that squeeze the original data into certain range. It is usually done to bring all the variables of a dataset into same range or scale irrespective of their magnitudes and unit. Scaling changes the value of coefficients, but it does not have any impact on model efficiency or significance (R2, P value, F statistics etc)

In machine learning, scaling helps in ease of interpretation in independent variable. Also, if the variables are scaled, gradient descent convergence (which is used in many machine learning models) is achieved faster when variables are withing certain range.

The two most popular scaling methods are normalized or MinMax scaling and standardized scaling. In case of normalized scaling the data is squeezed into 0 and 1. The minimum value is assigned to 0 and maximum value is assigned to 1. All, in between original values get converted into any number between 0 and 1. The advantage of this method is that it does not change the distribution of the original data so interpretation is easy. Below formula is used for normalization.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

In case of standardization, it brings all data to standard normal distribution with a mean of zero and standard deviation of 1. It changes the distribution of the original data and also loses the outlier information. Due to this, it becomes difficult to interpret the coefficients. However, machine algorithm (eg. Algorithms that use Gradient descent method) works faster on standardized variables.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Hence, if interpretation is main objective of the machine learning model then normalization should be chosen. In case, we are only concerned about prediction by the fitted model, standardization should be selected as scaling method.

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

Ans. –

The VIF of any i^{th} predictor is given $1 / (1 - R_i^2)$. Where R_i^2 is the R^2 of the model when ith predictor is explained by all other predictor variables. If R_i^2 is 1, VIF becomes infinite. The R_i^2 becomes 1 when a variable is completely explained by its predictors. This is called perfect correlation.

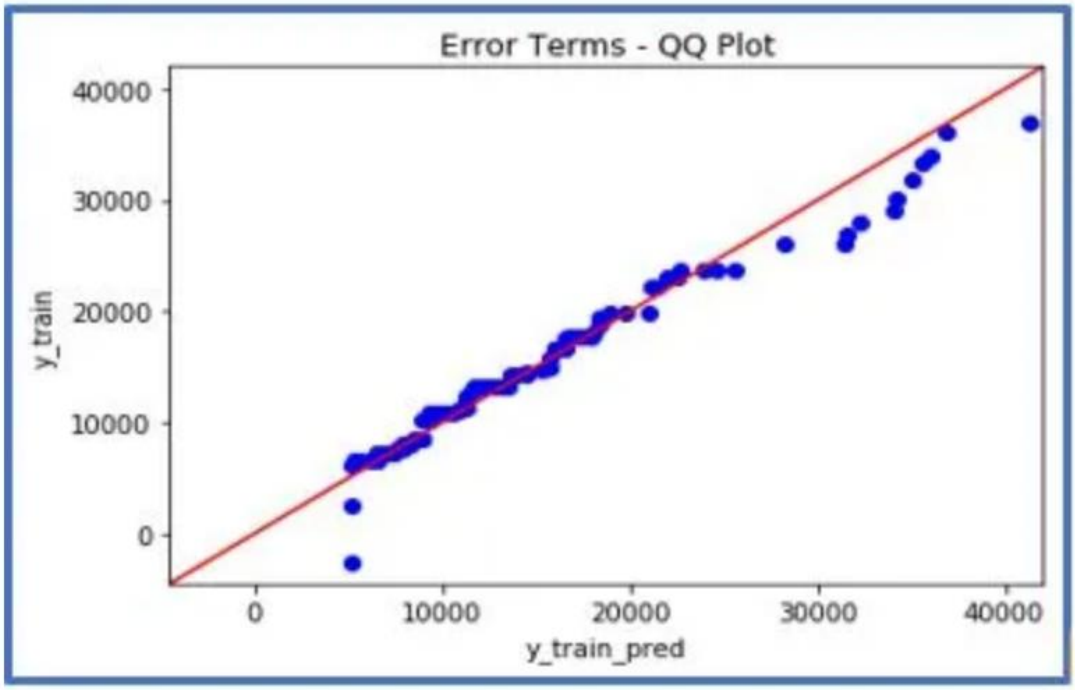
Hence, we can say that if a predictor variable has infinite VIF, that means it has the highest correlation with the other predictors and its always best to remove such variable to avoid multicollinearity as presence of the same makes it difficult to interpret a model with respect its predictors. Also, the coefficients can swing wildly, even the sign of coeff can invert, therefore, p values are also not reliable.

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

Ans. –

Quantile-Quantile (Q-Q) plot is a graphical tool, that can determine whether a particular dataset is part of any theoretical distribution such as a Normal, exponential etc. It also helps to determine if two sample datasets are come from same populations with same distribution.

It basically plots quantile values of two datasets against each other as shown below.



if all points lie on or close to straight line x=y, then it means that both datasets are following same distribution. On the other hand, if the points are lies away from the straight line that means the datasets have different distribution or they are from different population.

Use and importance in linear regression:

In a scenario of linear regression where we have training and test data set received separately, we can confirm using Q-Q plot that both the data sets are from populations with same distributions before making prediction on test data set.

We can also check if the generated errors are following normal distribution using Q-Q plot which is a major assumption behind linear regression.

The predicted and actual response y values can also be compared using Q-Q plot to evaluate the fitted model.

The Q-Q plot also has below advantages

a) the sample datasets do not need to be of same sizes. So it is fine, if the test data set smaller or larger than the train dataset.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can also be detected from this plot.