Subjective Questions

Assignment-based Subjective Questions:

1.) From your analysis of the categorical variables from the dataset, whatcould you infer about their effect on the dependent variable?

A: 1. season: As per the boxplot, fall had around 50% of its' cnt entries more than 5000, spring had around 50% of its' cnt entries less than 2000.

This means that ***fall*** is having high number of users followed by ***summer*** and ***winter*** with not a big difference and ***spring*** is having less number of users

1. mnth: As per the boxplot, ***Jan*** had less number of users followed by ***Feb.*** There are more users coming in the middle of the year and first two months and last two months of two years saw less number of users when compared to other months

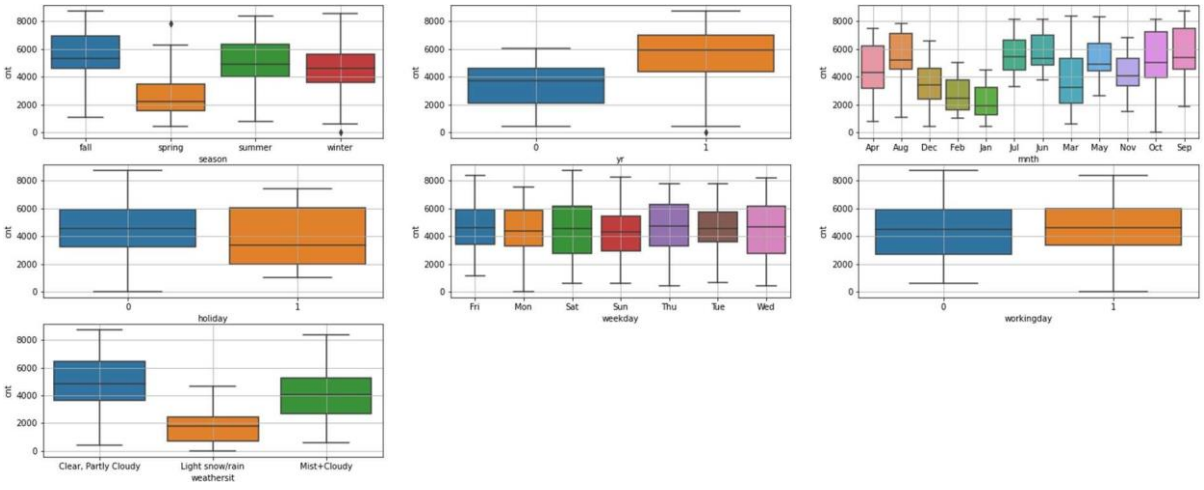
2. weathersit: As per the boxplot, there are no users when there is ***heavy rain/ snow*** which is highly unfavourable situations for bike riding. So, absence of any entries makes sense

when weather is ***Clear, Partly Cloudy*** number of users are high followed by ***Mist+Cloudy***

weather and ***Light snow/rain*** weather has less number of users. All this makes sense

3. weekday: As per the boxplot, weekday doesn't have much effect on number of users

4. yr: As per the boxplot, second year (2019) saw more users than the previous year(2018) which is expected as the time passes by, the company got some exposure and users



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

A: Now when we see this example,

Gender	**Female**
Male	**0**
Female	**1**

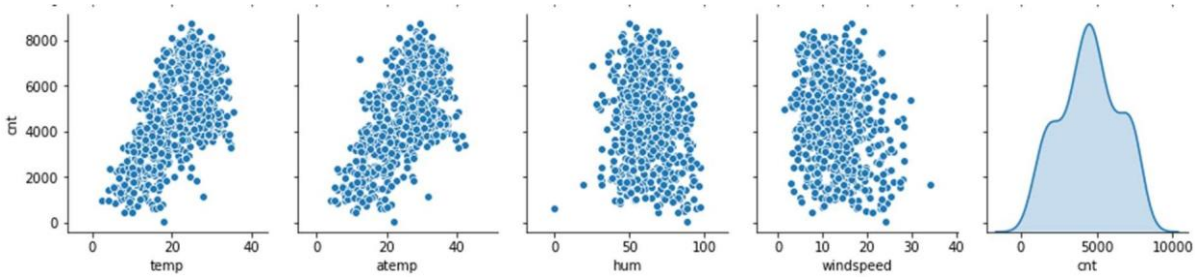
When Female column had 0 value, it obviously means that the value is male. So, having two columns for two values is redundant

Another reason is, if we have all dummy variables it leads to Multicollinearity between thedummy variables. To keep this under control, we lose one column

- So, if there are “n” levels in a column. We need only “n-1” dummy variables and there’s no rule that we need to drop the first dummy variable. We can drop any one dummy variable. But “drop_first” is inbuilt. So, it is generally preferred. Still, we can drop one dummy variablemanually if we want.

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

A: Pairplot involving target variable (cnt)



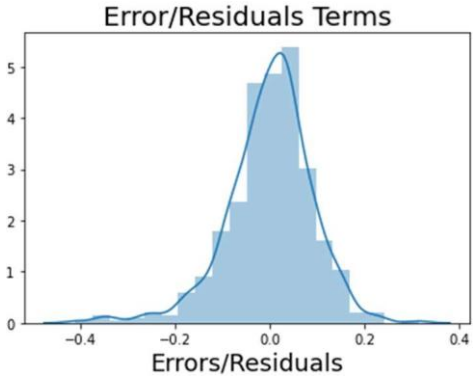
"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt)

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

A: 1. Residuals distribution should follow normal distribution and centred around 0.0.

Residual/error is the difference between y(actual value) and y(predicted value) and when a distplot is plotted for residuals, then that should be a normal distribution and it's mean should be around 0.0

We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not.

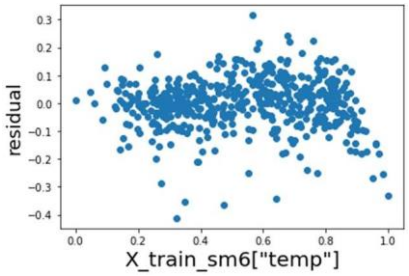


As you can see, error terms are following a normal distribution with its mean at 0.0. Hence, assumption is satisfied

2. There should be no pattern visible when residuals is plotted with an independent variable (predictor) in scatter plot

When residuals and an independent variable is plotted in a scatterplot, then those values should not follow any pattern whatsoever.

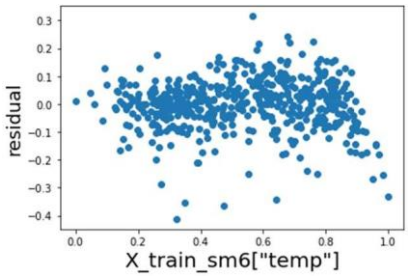
We validate this assumption, by plotting a scatter plot with residuals and an independent variable and check whether there is a pattern or not.



As you can see, there is no visible pattern. Hence, the assumption is satisfied.

3. Error/ residual terms should have a constant variance (no heteroscedasticity)

We validate this assumption, by plotting a scatter plot with residuals and an independent variable (predictor) and check whether the points are equi - distant from the mean or 0.0



As you can see, all points are scattered around 0.0 and are almost equidistant from the y = 0.0 line.

There is no change in variance. Hence, the assumption is satisfied

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

A: These are the 3 top features contributing

Features	**Coefficient**
Temp	**0.5499**
weathersit_Light snow/rain	**-0. 288**
Yr	**0.2331**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A: Linear regression is a machine learning algorithm based on "Supervised learning". It performs a regression task. Regression is one type of supervised machine learning algorithms whose target variable will be continuous.

Linear regression is based on "slope intercept form" which is "y = mx + c"

Where "m = slope/ rate of change" and "c = y-intercept (y value when x = 0)"

Linear regression types:

1. Simple Linear Regression:

Simple linear regression performs the task of predicting a dependent variable based on a single independent variable.

The output model will be a straight line

Assumptions of Simple Linear Regression:

1. Dependent variable is linearly dependent on independent variable
2. Error/ Residual terms are normally distributed with mean around zero
3. Error/Residual terms are independent of each other. (No correlation with each other)
4. Error/Residual terms should have constant variance.

The form of the model will be like this **y = B0 + B1*x+e**

where B0 is intercept and B1 is the slope in general terms

B0 is the value of y(dependent variable) when there is no x (independent variable)

B1 is the rate at which y(dependent variable) increases when x (independent variable) changes by a unit

"e" is nothing but the error or residual and it should follow the above mentioned assumptions

Coefficients are obtained by minimizing sum of squared error (Least Squares criterion)

2. Multiple Linear Regression:

Multiple Linear Regression performs the task of predicting a dependent variable based on a set of independent variables.

The output model will be hyperplane instead of a line.

Additional Assumptions of Multiple Linear Regression:

1. Overfitting: This means that the model is good at training dataset but performs poorly intest dataset. This is caused when there is not enough data to train on or when the modelis too complex with all available features present in the model

2. Multicollinearity: This means that one independent variable can be explained by other independent variables. In other words, if an independent variable is having high correlation with other independent variable/variables

The form of the model will be like $y_i = B_0 + B_1 x_{i1} + B_2 x_{i2} + \ldots B_p x_{ip} + \varepsilon_i$ for $i = 1,2, \ldots n$.

Where B0 is intercept and B1, B2, … are the coefficients for respective independent features

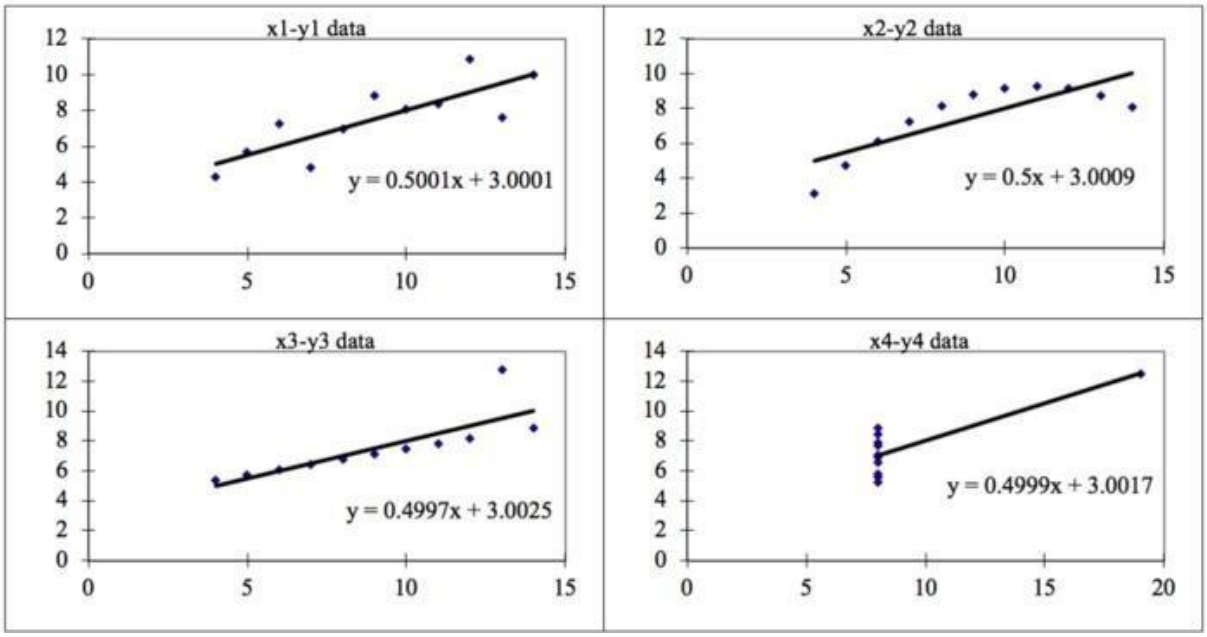
Interpretation of coefficients: B1 is the coefficient of independent variable X1 if all other variables are held constant.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of four datasets which have similar descriptive statistics like mean , R2 , standarddeviation etc. but when plotted as a scatter plot, they all show a different distribution.

This was observed by a statistician **Francis Anscombe** in 1973 where he showed a special dataset as below split into 4 sets.

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	



Let's explain each of the quartets:

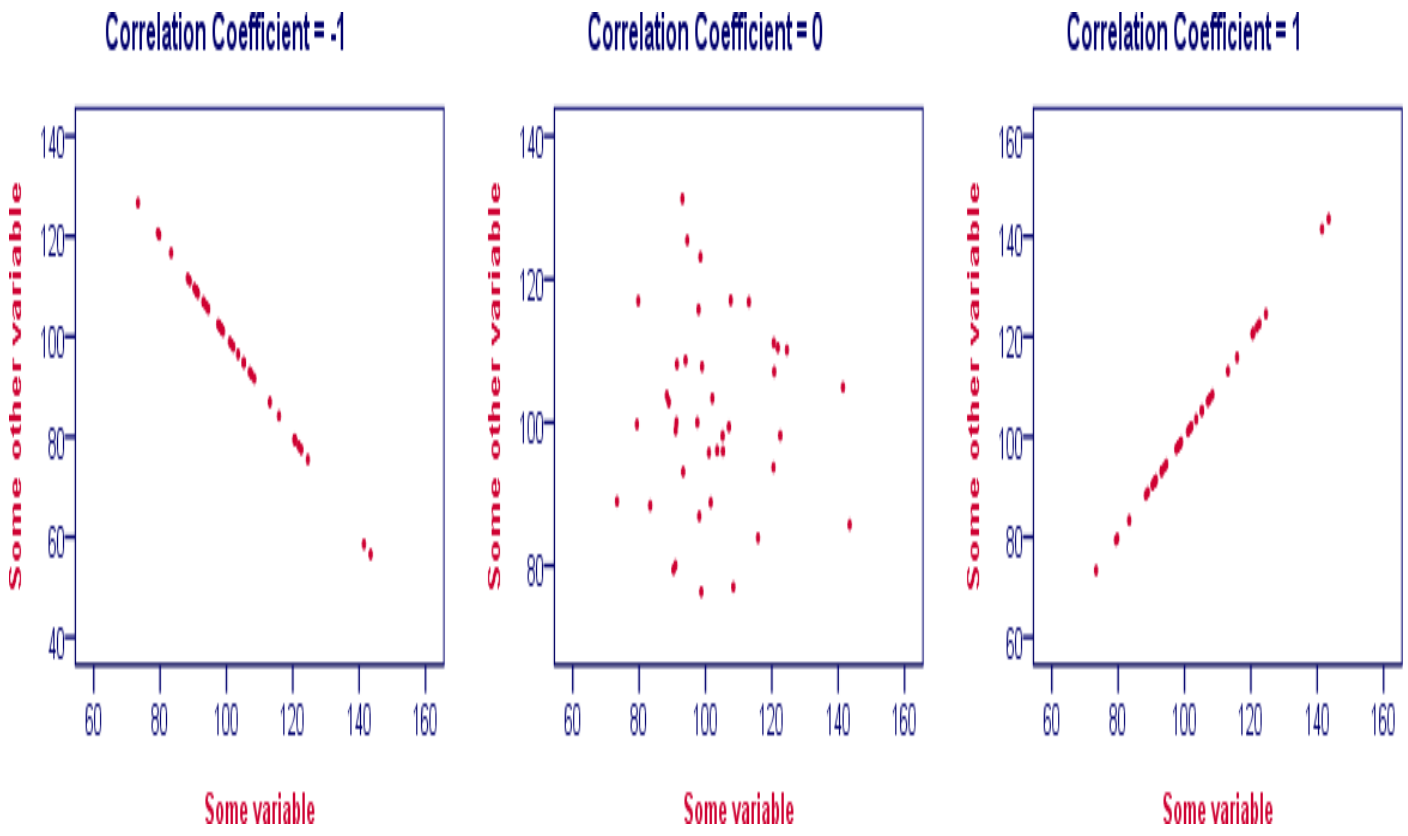
1. The first data looks like it's a good fit for linear regression.
2. The second plot looks like it's not a good idea to use linear regression as the data looks to be non-linear.
3. The third plot if we see, it was almost a very good linear fit, but due to outliers there a little realignment of the line.
4. The fourth shows outliers in the dataset which the model couldn't handle though there is regression line is being fit.

With these above descriptions, we were able to see how the linear regression model was fooled with the data being selected. This way Anscombe's quartet gives a very good analysis on the data by which we can decide upon choosing linear regression as a good model or not!

3.)What is Pearson's R?

A:Pearson's Correlation Coefficient **R** is a value ranging from -1 to +1 explaining how two numeric variables are related to each other

5. **R value cannot be below 1:** If the R value is as close to -1 then it means that if one variable value increases then the other variable value strictly decreases.
6. **R Value being 0 :** This indicates that both the variables are independent of each other
7. **R value cannot be above 1:** If the value of R is close to +1 then it means that if one variable increases then the other variable value also increases



The increase/decrease of one variable with other depends on how close the R value is to -1 and +1. R value is very sensitive to outliers as well. The Person R formula is given by:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

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

A: Scaling is a concept of bringing all the data values in a similar range usually 0-1. Scaling is performed for variousreasons,

1. It will help us understand the Beta values comparatively much better.
2. It will help the gradient descent algorithm to converge quickly towards minima. This is because we can avoid different step sizes for different variable ranges.

If values are in different ranges and units like for example , 5000 grams and 5Kgs are one and the same, but if not scaled properly, variation in the y-ypred value will be more for grams than kgs which will end up showinga greater Beta value for this variable, which might mislead if other variables are on a shorter scale like Kgs.

Scaling Types:

Normalized scaling: This will make sure that the values after scaling when plotted, show a normal distribution. One popular normalized scaling technique is **Min-Max Scaling** where the formula looks like,

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

Standardized scaling: This will make sure that the distribution of the standardized values has a mean of 0 anda standard deviation of 1.

$$X' = \frac{X - \mu}{\sigma}$$

Normalized scaling is affected with outliers as it is dependent on max(x), whereas standardized values aren't affected. We used normalization when we want to get a normal bell curve in the data, and if already have a bell curve kind of data we proceed to standardize, if possible, to make the data more focused.

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

A: VIF is the short form for "Variance Inflation Factor". VIF calculates how well one independent variable is explained by the other independent variables combined.

VIF= 1/(1-R2)

Where R-squared is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables

- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1

So, VIF = 1/(1-1) which gives VIF = 1/0 which results in “infinity”

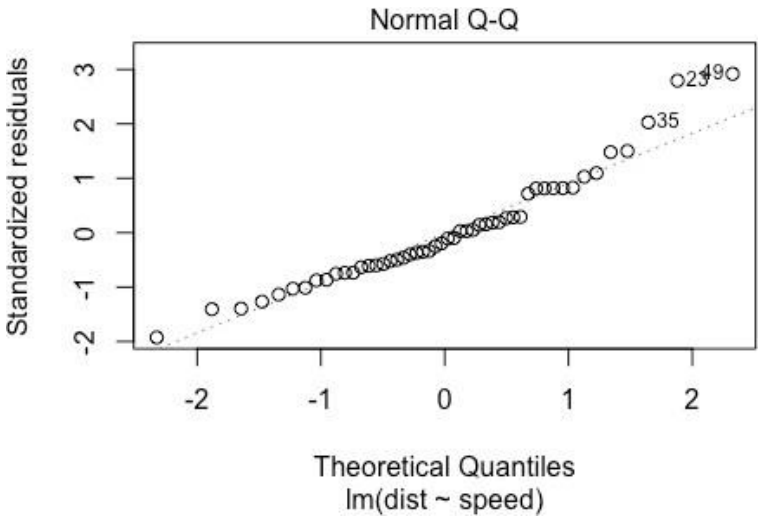
6.) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A: Q-Q plot will help us understand if two data sets have come from a similar kind of distribution. When we receiveTrain and test data sets separately, these plots will help us understand if we obtained data from a similar kind ofdistribution.

How is Q-Q plot formed?

If we have 9 data points. We take a normalized data and starting from the mean, we create half no. of quantiles to the left and other half to the right of it. Totally creating 10 theoretical quantiles with equal width.

Further, we calculate the z-scores of all these data points and again create 10 actual quantiles at these z-scores. Now plotting these values on a graph with theoretical quantiles on the x-axis and actual quantiles on the y-axis will give us the Q-Q plot. An example



Using the plot, we can observe if two datasets,

- Come from a common population with common distribution
- Have a common location and scale
- Have similar distribution shapes
- Similar tail behavior