

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? (3 marks)

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

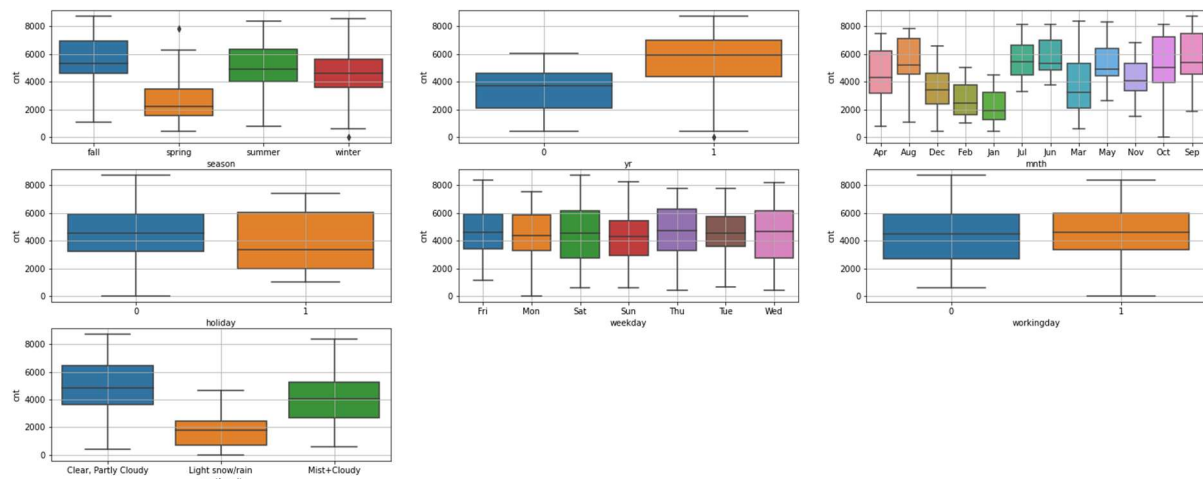
2. mnth: As per the boxplot, **Jan** had less number of users followed by **Feb**. There are more users coming in the middle of the year and first two months and last two months of two years saw less number of users when compared to other months

3. 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

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

5. 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? (2 mark)

A: Now when we see this example,

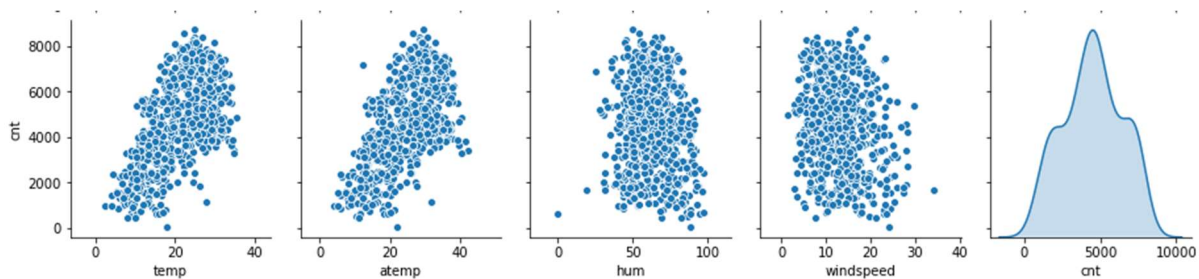
Gender	Female
Male	0
Female	1

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 the dummy 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 variable manually if we want.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

A: Pairplot involving target variable (cnt)



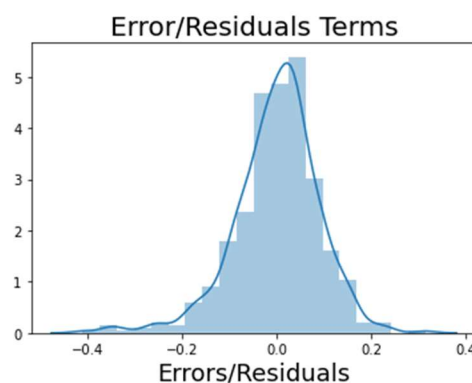
“temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

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.

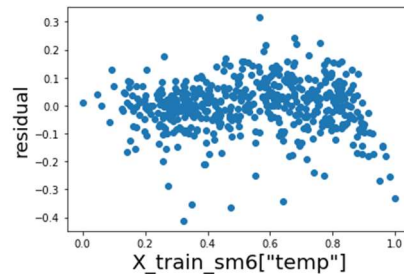


As you can see, error terms are following a normal distribution with it’s mean at 0.0. Hence, assumption is satisfied

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.

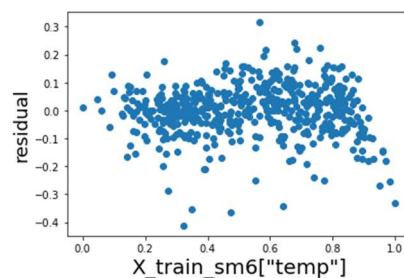
We validate this assumption, by plotting a scatter plot with residuals and an independent variable and check whether there is a pattern or not.



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? (2 marks)

A: These are the 3 top features contributing

<u>Features</u>	<u>Coefficient</u>
Temp	0.5499
weathersit_Light snow/rain	-0. 288
yr	0.2331

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

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 $y = B_0 + B_1x + e$

where B_0 is intercept and B_1 is the slope in general terms

B_0 is the value of y (dependent variable) when there is no x (independent variable)

B_1 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 in test dataset. This is caused when there is not enough data to train on or when the model is 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 $y_i = B_0 + B_1x_{i1} + B_2x_{i2} + \dots B_px_{ip} + \epsilon_i$ for $i = 1, 2, \dots n$.

Where B_0 is intercept and B_1, B_2, \dots are the coefficients for respective independent features

ϵ_i is the error term which still have to follow all the assumptions from simple linear regression.

Interpretation of coefficients: B_1 is the coefficient of independent variable X_1 if all other independent variables are held constant.

2. Explain the Anscombe's quartet in detail. (3 marks)

A: Anscombe's quartet is the detailed demonstration of the pitfalls when summary statistics are used to get an overall idea on the dataset.

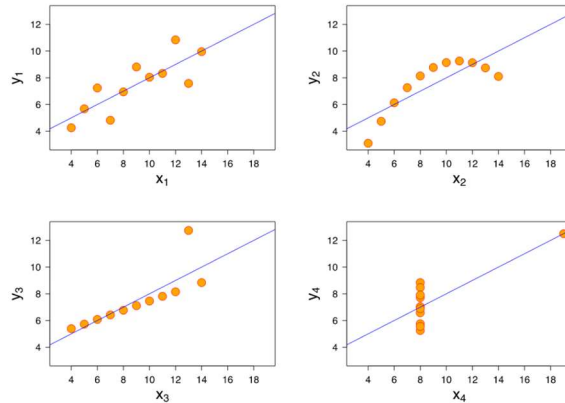
Anscombe's quartet is a group of four datasets that appear to be similar when using typical summary statistics, yet four datasets are very different from each other when graphed

Each dataset consists of 11 (x, y) pairs as follows:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All the summary statistics we think of using are very close:

- The average x value is 9 for each dataset
 - The average y value is 7.50 for each dataset
 - The variance for x is 11 and the variance for y is 4.12
 - The correlation between x and y is 0.816 for each dataset
 - A linear regression (line of best fit) for each dataset follows the equation $y = 0.5x + 3$
- But when we plot these four datasets,



Now, each dataset is showing it's own real relationship. Dataset 1 consists a set of values following a rough linear relationship with some variance. Dataset 2 fits a neat curve but that doesn't follow linear relationship. Dataset 3 is following a strict linear relationship with an outlier and Dataset 4 looks like x remains constant for all y values except one outlier

Through this, we can understand how important data visualization in understanding the clear understanding of the data.

3. What is Pearson's R? (3 marks)

A: Pearson's R is a statistic that measures the linear correlation between two variables.

It value ranges between -1 to +1




It's formula is

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Where x(i) is X-value, x-bar is mean of "x" variable

y(i) is Y-value, y-bar is mean of "y" variable

Values of Pearson's correlation coefficient:

Values of Pearson's correlation coefficient		
Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:		
r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables
r = +1		data lie on a perfect straight line with a positive slope

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

A: The problem while building a model,

Gradient descent takes a bit more time when it have to go through every variable and it's specific range. To overcome this, scaling is done

Scaling is a process in which all numerical variables are brought into one range so that model building would be way faster and efficient

There are mainly two types of Scaling. They are:

Standardized scaling

This means that transforming the data so that it fits within a specific scale, like 0–100 or 0–1. You want to scale data when you're using methods based on measures of how far apart data points, like support vector machines, or SVM or k-nearest neighbors, or KNN.

MinMax Scaling is most used standardized scaling method:

This is done like this: $X_{sc} = (X - X_{min}) / (X_{max} - X_{min})$

Where X is that original value in that X column and X(min) and X(max) are the minimum and maximum values of that column. This brings the values to a range in between 0 – 1.

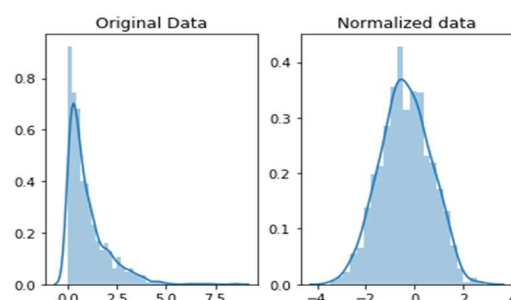
Normalization

Scaling just changes the range of your data. Normalization is a more radical transformation. The point of normalization is to change your observations so that they can be described as a normal distribution.

Normal distribution: Also known as the “bell curve”, this is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean. The normal distribution is also known as the Gaussian distribution.

It's done like this:
$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

In general, you'll only want to normalize your data if you're going to be using a machine learning or statistics technique that assumes your data is normally distributed. Some examples of these include t-tests, ANOVAs, linear regression, linear discriminant analysis (LDA) and Gaussian naive Bayes.



Notice that the shape of our data has changed. Before normalizing it was almost L-shaped. But after normalizing it looks more like the outline of a bell (hence “bell curve”).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

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.

It is done like this

$$VIF = \frac{1}{1 - R^2(x_1)}$$

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 it's R-squared value will be equal to 1

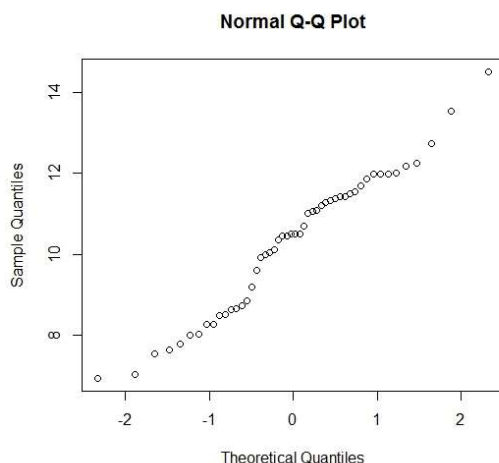
So, $VIF = 1/(1-1)$ which gives $VIF = 1/0$ which results in “infinity”

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A: 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 dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. 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 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. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



“Quantiles” are points in your data below which a certain proportion of your data fall. For example, imagine the classic bell-curve standard Normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That’s the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64.

It is used to check following scenarios:

If two data sets —

- i. come from populations with a common distribution
- ii. have common location and scale
- iii. have similar distributional shapes
- iv. have similar tail behavior

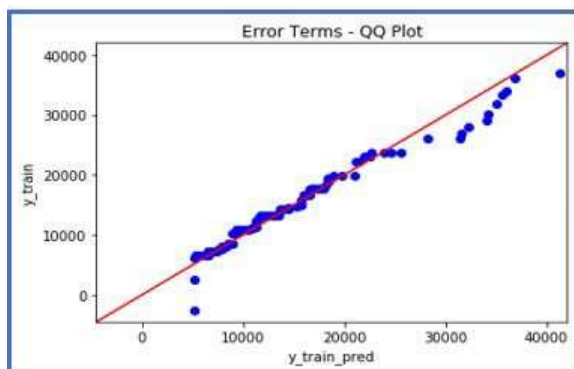
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

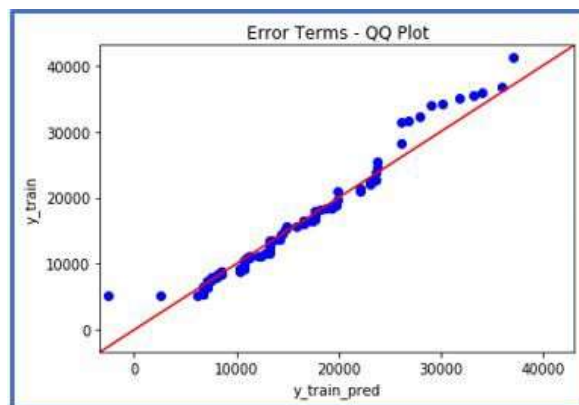
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis