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?

The categorical variables that have been identified in the dataset and their effect on the dependent variable, 'cnt' are as follows:

1. **'season'** – We see mainly four seasons in our dataset: 'spring', 'summer', 'fall' and 'winter'. It is visible from our analysis with the help of bar plot that during 'fall', the bike demand is the highest. 'spring' has the lowest demand of bikes out of all the seasons. 'summer' and 'winter' have almost similar demand for bikes; however, after having created the dummy variables and from our final model, it is quite visible that these two are very significant.
2. **'yr'** – The bike demand increases with the year as we saw an increase in business from 2018 to 2019. This depicts that there is an increasing demand for such a business.
3. **'month'** – The highest demand of bikes is between the months of June and September, outside which, the demand begins to drop. We also saw from our final model later, how August and September were quite significant in depicting the bike demand.
4. **'holiday'** – Here, we can see that the bikes demand is higher when there is no holiday. This probably means that most of the customers are working professionals who use bikes for daily commute to & from office.
5. **'weekday'** – We see how the bike demand is the lowest on Tuesdays while there is not much difference in bike demand on the other days of the week.
6. **'workingday'** – As we concluded earlier from the 'holiday' variable, the bike demand is higher on working days
7. **'weathersit'** – From the bar plot used to analyze the 'weathersit' variable, we can infer that the highest demand for bike is on the days when the weather is Clear or there are Few clouds while it is the lowest when there is Light Snow or Light Rain & Thunderstorm & Scattered clouds or Light Rain and Scattered clouds.

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

When we have a categorical variable with say 'n' levels, it is a logical approach to consider 'n-1' dummy variables. Suppose we have a categorical variable 'Reviews' with n = 4 levels, namely, 'Poor', 'Average', 'Good' and 'Best'. We will be creating the dummy variables as per below table,

Reviews	Poor	Average	Good	Best
Poor	1	0	0	0
Average	0	1	0	0
Good	0	0	1	0
Best	0	0	0	1

Although we created separate dummy variable for each level; however, it is not really necessary to have equal number of dummy variables. For example, if we remove the 'Poor' dummy variable, we will still be able to explain all the 04 levels.
As we see from the below table, if 'Average', 'Good' & 'Best' all are 0, that means the review is 'Poor.
If 'Average' is 1 while 'Good' and 'Best' are 0, it means the review is 'Average'.
Similarly, if 'Average' and 'Best' are 0 while 'Good' is 1, it means the review is 'Good'.
And finally, if 'Average' and 'Good' are 0 while 'Best' is 1, it means the review is 'Best'.

Reviews	Average	Good	Best
Poor	0	0	0
Average	1	0	0
Good	0	1	0
Best	0	0	1

Hence, we use drop_first=True to drop this extra dummy variable which we don't really need.
This reduces the total number of variables and helps in our analysis.

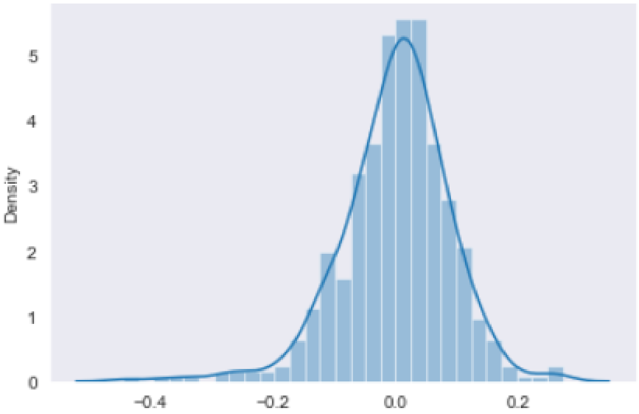
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The variables 'temp' and 'atemp' have the highest correlation with the target variable which is 'cnt'. The correlation (as seen from the heatmap) is 0.63. The pair plot further helped us realize that we could use either of these 02 variables for modelling. We finally chose 'atemp' as this is the feeling temperature for the customers and makes more sense to the business.

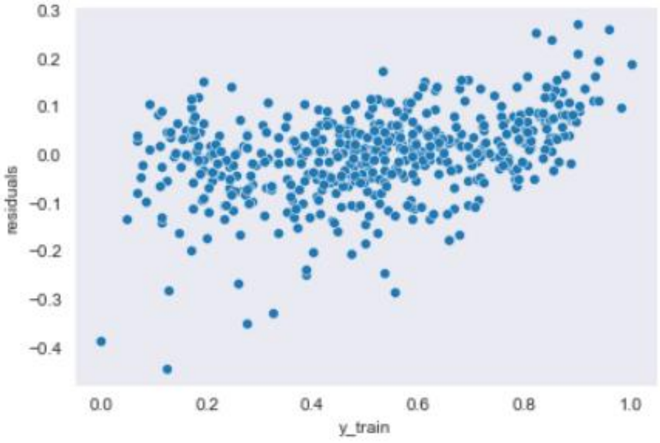
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The assumptions that we had to validate were as follows:

1. **Assumption-1: Error terms are normally distributed**
   We computed the y_train_pred for the X_train using our final model lr_model_05. Then, we calculated the residuals by taking the difference of y_train and y_train_pred. After plotting a distribution plot for the residuals (refer to the below plot), we saw that it is normally distributed around mean = 0.
   This helped us validate the first assumption that the error terms are normally distributed.



2. **Assumption-2: Error terms have constant variance (homoscedasticity)**
   By creating a scatter plot between the y_train and the residuals, it was quite evident that there was no special pattern between the two. The residuals are distributed near the 0 and hence, we can say that mean is around 0 which is also visible from the distribution plot above.



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

1. **atemp:** It has the highest coefficient value of 0.441271 that indicates that this variable has the highest impact on the target variable 'cnt'. A unit increase in 'atemp' variable, increases the bikes demand by 0.441271 units (as per the scale considered by us).

2. **Light_Snow_Light_Rain:** It has the highest negative coefficient value of 0.293286 that indicates that this variable has the second highest impact on the target variable 'cnt'. A unit increase in 'Light_Snow_Light_Rain' variable, decreases the bikes demand by 0.293286 units.

3. **yr:** It has the second highest positive coefficient value of 0.235210 that indicates that this variable has the third highest impact on the target variable 'cnt'. A unit increase in 'yr' variable, increases the bikes demand by 0.0.235210 units.

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

Linear Regression is a type of supervised machine learning model in which we have historical data with labels which can be used for prediction or projection purposes. Linear Regression can only be used when the target variable is a continuous variable. **For example,** temperature, humidity, prices of a particular commodity, energy consumption of an AC equipment etc.
Linear Regression finds its applications in wide range of industries where we intend to predict future outcomes based on historical data, for example, finance, education, medicine, engineering, sports, economics etc.

a. Simple Linear Regression

This is the most elementary type of regression model where we predict the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

The standard equation for this straight line (also know as regression line) is,

$Y = \beta_0 + \beta_1 X$

where,

β_1 – slope of the regression line (describes the increase in y with unit increase in X)

β_0 – intercept of the regression line (value of y when X=0)

X – predictor (or independent) variable

y – target (or dependent) variable

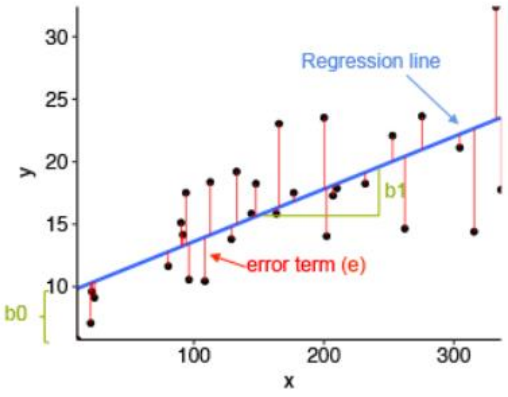


Image Source: Google Images

Best fit regression line is found by minimizing the term RSS (Residual Sum of Squares) which is equal to the sum of the squares of the residuals for all the data points in the plot. Basically, a residual is the difference of the actual value of ‘y’ and the predicted value of ‘y’.

This method of minimizing the RSS is also know as OLS (Ordinary Least Square). Below are the basic steps for computing RSS:

Residual (ϵ_i) $= y_i – y_{pred}$

Residual Sum of Squares (RSS) $= \epsilon_1^2 + \epsilon_2^2 + \ldots\ldots + \epsilon_n^2$

Or

Residual Sum of Squares (RSS) $= (y_1 – \beta_0 - \beta_1 X_1)^2 + (y_2 – \beta_0 - \beta_1 X_2)^2 + \ldots\ldots + (y_n – \beta_0 - \beta_1 X_n)^2$

Or

Residual Sum of Squares (RSS),

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i)$$

Strength of linear regression model can be assessed using the metric R-squared or R^2. It explains what proportion of the given data variation is explained by the model. It takes in a value between 0 and 1. Mathematically, R^2 can be written as,

$$\mathbf{R^2 = 1 - \frac{RSS}{TSS}}$$

where,
RSS –Residual Sum of Squares
TSS – Total Sum of Squares (or variance in y)

Another method to assess the strength of the linear regression model is **Residual Square Error (RSE).** Mathematically it can be expresses as,

$$RSE = \sqrt{\frac{RSS}{df}}$$

where,
df – degrees of freedom (n-2)
n – number of data points

b. Multiple Linear Regression

Multiple linear regression is the type of linear regression method which is used to understand the relationship between one dependent variable and several independent variables (predictor variables).
The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.
Mathematically, a standard multiple linear regression equation can be written as,

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p + \epsilon}$$

where,
β_p – Coefficient of pth predictor variable
X_p – pth predictor variable
ϵ – Error

β_p can be described as the change in the value of 'y' per unit increase in the predictor variable 'X_p' when other predictor variables are kept constant.

The model now fits on a 'hyperplane' instead of a line as in simple linear regression model.

Various Steps used when working with Linear Regression Algorithm

a. Data Preparation
   - Encoding of the data (Converting binary variables into 0 or 1 and Creating dummy variables for other categorical variables)
   - Splitting the data into Train set and Test set
   - Feature Scaling the data set so as the bring all the variables to a common scale

b. Data Modelling
   - Create X and y
   - Feature Selection using RFE method or manual elimination method or a balanced approach
   - Building model using statsmodels or sklearn library.
   - Calculating the VIF (variance inflation factor) to verify the model.
   - Rebuilding the model with all the significant coefficients and less VIF.

c. Residual Analysis
   - Checking for the assumption of error terms being normally distributed
   - Checking for the assumption of error terms being homoscedastic in nature.

d. Model Evaluation
   - Making predictions on the test set using the final model.
   - Comparing the performance metrics of the trained model and the test set.

Assumptions of a Linear Regression Algorithm

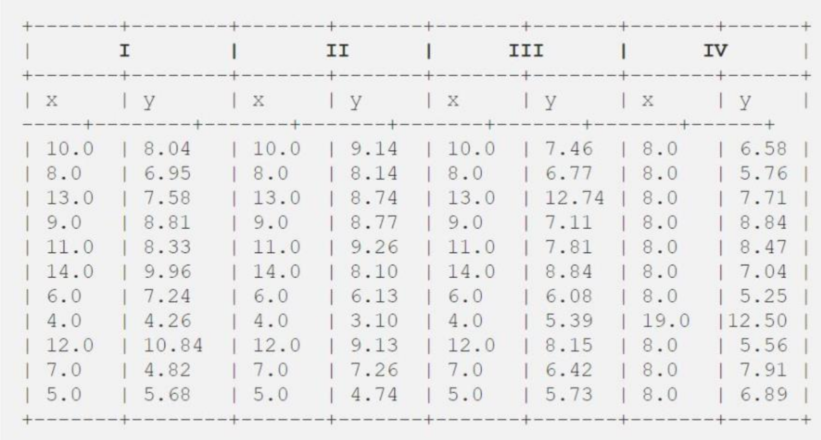
a. There is a linear relationship between X and y.
b. Error terms are normally distributed.
c. Error terms are independent of each other.
d. Error terms have constant variance (homoscedasticity)
e. There is no association between the predictor variables (multicollinearity)

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x, y) points.

The quartet was constructed by the statistician Francis Anscombe to demonstrate **the importance of data visualization in addition to the statistical summary** of the data.

Summary statistics (that consists of parameters like mean, mode, median, quantiles, mini-max values, etc.) does help in understanding the range of the values, however, it doesn’t really give any idea about the shape or distribution of the data.
Below is how an Anscombe’s quartet looks like:

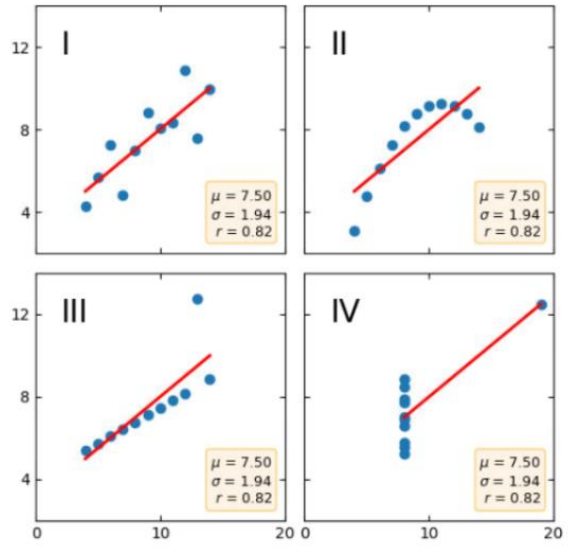


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

The summary statistics show that mean and variance for all the groups is same:

1) Mean of x values in each data set = 9.00
2) Standard deviation of x values in each data set = 3.32
3) Mean of y values in each data set = 7.50
4) Standard deviation of x values in each data set = 2.03
5) Pearson’s Correlation coefficient for each paired data set = 0.82
6) Linear regression line for each paired data set: y = 0.500x + 3.00

As seen below, when the graphs for all the groups are plotted, even though regression line is same for all the groups, however, each graph has a different distribution of data.



Graph I:
The dataset does fit a linear regression and we can further use the line of best fit for making predictions on the data.

Graph II:
Although we could fit in a linear regression on this dataset, however, this clearly is not the best regression line for the data. The distribution of the data is non-linear.

Graph III:
In this graph, we see how for the 10 points, it is a perfect linear regression and then there is one outlier. For our prediction purposes, we may investigate this outlier (checking if it does conform to the mathematical definition of an outlier) and then continue doing our regression with this removed.

Graph IV:
In this graph, we again see the effect of a single outlier. Again, the line of best fit is not appropriate for this dataset.

Hence, we may conclude that it is always a good idea to preform graphical analysis alongside statistical analysis. Without plotting, we might not be able to judge whether the regression line is the best fit or not.

3. What is Pearson's R?

Pearson's r is also referred to as the correlation coefficient that basically measures the relationship between the two continuous variables. It does give information on the magnitude and the direction of the relationship between the variables. Below is the mathematical formula of Pearson's r,

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

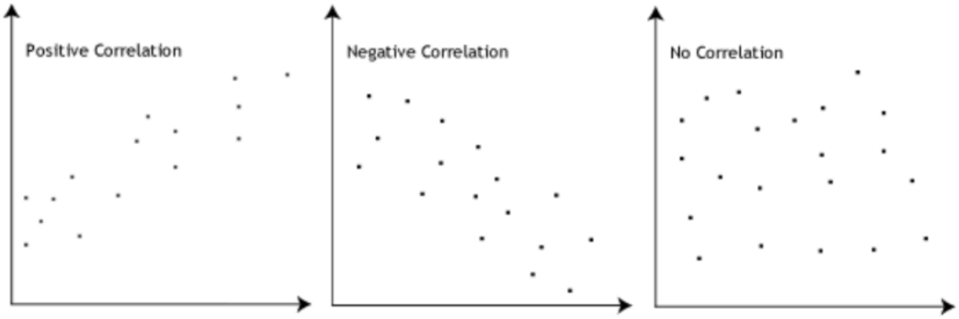
$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Properties of Pearson's r:

i. ***Range:*** The coefficient's values range between -1 and +1, where +1 tells that there is a perfect positive correlation between the two variables, -1 tells that there is perfect negative correlation between the two variables while 0 indicates that there is no relation between the two variables at all. This can be understood from the below figure,



ii. ***Independent of the Units of the variables:*** The Pearson's r doesn't depend on the units of measurement of either of the variables. That means if the units of the both the variables are different (one variable has unit m/s while the unit of other variable is changed to Km/h), the correlation coefficient doesn't change.

iii. ***Symmetric nature:*** This means that correlation between X and Y will be same as that between Y and X.

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

Scaling (or feature scaling) refers to a technique of bringing down all the independent variables or features of the data to a same scale. This is done during the data preprocessing step before proceeding with model building. It leads to easier interpretation and faster convergence for gradient descent methods. One thing to note here is that scaling only affects the coefficients and doesn’t affect the parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Most of the time, the data that we have contains variables with mixed units. As a result, their magnitude is completely different. For example, one variable has values of year of work experience that lie between, let’s say 0-60, while there is another variable that has annual income that lie in range of thousands and even higher.

If the scaling is not done, then our model will only take into consideration the magnitude of the values and not the unit of values which will further lead to incorrect modelling; meaning, the model will tend to weigh values with higher magnitude as higher while the values with lower magnitude, lower. This is going to be a disaster for the business for which we are creating a model as the resultant model will be biased towards the higher values.

Thus, in order to eliminate this issue, we bring all the variables to a common scale, either 0 to 1 or centered around 0, depending on the type of scaling technique used.

There are 02 types of scaling techniques widely used: Standardisation and MinMax Scaling. The difference between both these techniques is as follows:

Standardisation	MinMax Scaling
Standardisation basically brings all the data into a standard normal distribution with mean zero and standard deviation one.	MinMax Scaling brings down all the data within a range of 0 and 1.
The formula in the background used for this method is: $x = \frac{x - mean(x)}{sd(x)}$	The formula in the background used for this method is: $x = \frac{x - min(x)}{max(x) - min(x)}$
Standardisation is implemented in Python using sklearn.preprocessing.StandardScaler	MinMax Scaling is implemented in Python using sklearn.preprocessing.MinMaxScaler

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

VIF or Variance Inflation Factor is a measure of multicollinearity between the various feature variables. In a model that has been built using several independent variables, there are higher chances of the variables being correlated with each other which makes the presence of those variables redundant.

The common heuristic for the VIF values is:

- > 10 : Definitely high VIF and the variable should be eliminated.
- > 5 : Might be okay but still needs inspection from business point of view.
- < 5 : Good VIF value. No need to eliminate this variable.

Now, in some cases, the VIF might also come out to be infinite. This means that there is a perfect correlation between the two independent variables. This can be explained by following theory:

Going back to the equation of R-squared which is 1 – (RSS/TSS), when there is a perfect correlation between the variables, RSS becomes 0, which means the sum of all the errors is 0 or there is no error in the predicted and the actual values.

When RSS becomes 0, the value of R-squared (R2) becomes 1.

As we know, VIF = 1/(1 – R2); now, when R2 is 1, the denominator of the equation becomes 0, which makes the VIF as infinite.

Now, that further explains that when there is a perfect correlation between the two variables, the VIF becomes infinite.

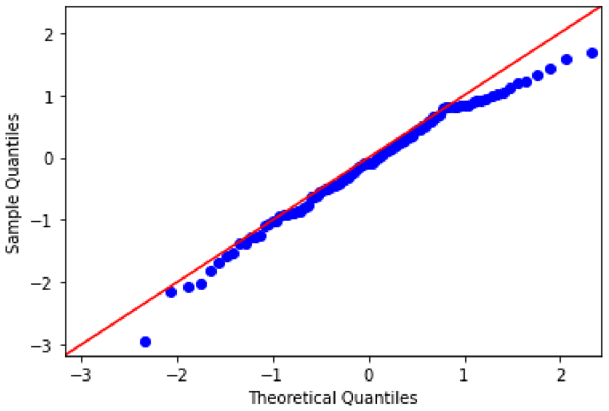
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The Q-Q plot or Quantile – Quantile plot is a graphical method of determining if the two datasets come from populations with a common distribution.

A Q-Q plot is basically a scatterplot which is created by plotting the quantiles of the first dataset against the quantiles of another dataset. If both sets of quantiles belong to the same distribution, we should see the points forming an almost straight line. By a quantile, we mean the fraction (or percent) of points below the given value. For example, the 0.6 (or 60%) quantile is the point at which 60% percent of the data fall below and 40% fall above that value.

A 45-degree reference line is also plotted along with the scatter plot. If the two sets come from a population with the same distribution, the points in that case should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions and vice versa.

Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



The Q-Q plot in linear regression can be used to validate our assumption of normal distribution of error terms of our model.

It uses standardized values of residuals to determine the normal distribution of errors. Ideally, this plot should show a straight line. A curved, distorted line suggests residuals have a non-normal distribution.