**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?**

There were 6 categorical variables in the dataset.

The inference that I could derive were: - season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable. - month: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, month has some trend for bookings and can be a good predictor for the dependent variable.

**2. Why is it important to use drop\_first=True during dummy variable creation?**

Drop\_first = true drop original variable for which the dummy was created and will drop first dummy variable for each set of dummies created.

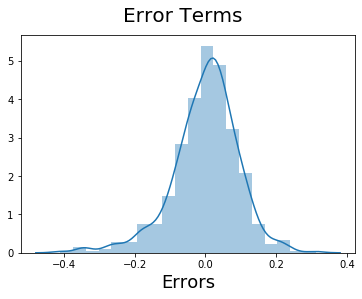
one level of our categorical feature become the reference group during dummy encoding for regression and is redundant. A categorical variable of K categories, or levels, usually enters a regression as a sequence of K-1 dummy variables. This amounts to a linear hypothesis on the level means**.**

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

Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'

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

By performing residual analysis of training data



We can observe that the residuals are normally distributed from the above histogram hence validating the model.

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

The top 3 features contributing towards the demand are:

Temperature (temp) - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.

Weather Situation 3 (weathersit\_3) - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.

Year (yr) - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised learning machine learning algorithm. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression is used for finding linear relationship between target and one or more predictors.

**Algorithm**

The model predicts the final values with minimum error if given the variables

Where error is



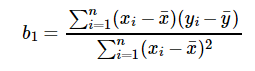
If we don’t square the error, then positive and negative point will cancel out each other.

The predictor can be represented by the formula

Y(pred) = b0 + b1\*x

For model with one predictor,





**Exploring ‘b1’**

* If b1 > 0, then x(predictor) and y(target) have a positive relationship. That is increase in x will increase y.
* If b1 < 0, then x(predictor) and y(target) have a negative relationship. That is increase in x will decrease y.

**Exploring ‘b0’**

* If the model does not include x=0, then the prediction will become meaningless with only b0. For example, we have a dataset that relates height(x) and weight(y). Taking x=0(that is height as 0), will make equation have only b0 value which is completely meaningless as in real-time height and weight can never be zero. This resulted due to considering the model values beyond its scope.
* If the model includes value 0, then ‘b0’ will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible.

**Co-efficient from Normal equations**

Apart from above equation co-efficient of the model can also be calculated from normal equation.



**Optimizing using gradient descent**

Complexity of the normal equation makes it difficult to use, this is where gradient descent method comes into picture. Partial derivative of the cost function with respect to the parameter can give optimal co-efficient value.

**Residual Analysis**

Randomness and unpredictability are the two main components of a regression model.

Prediction = Deterministic + Statistic

Deterministic part is covered by the predictor variable in the model. Stochastic part reveals the fact that the expected and observed value is unpredictable. There will always be some information that are missed to cover. This information can be obtained from the residual information.

Non-random pattern of the residual plot indicates that the model is,

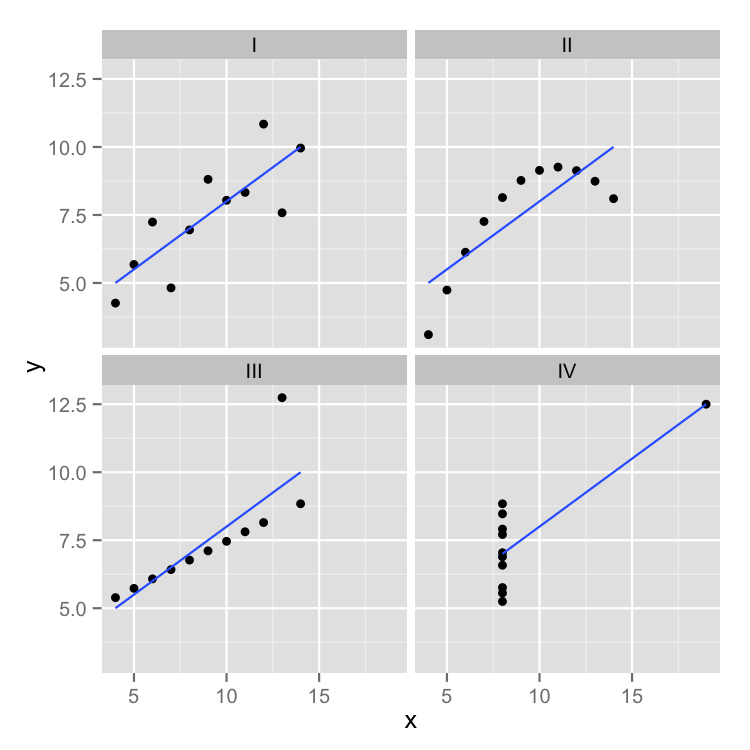
* Missing a variable which has significant contribution to the model target
* Missing to capture non-linearity (using polynomial term)
* No interaction between terms in model

Characteristics of a residue

* Residuals do not exhibit any pattern
* Adjacent residuals should not be same as they indicate that there is some information missed by system.

**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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



**3. What is Pearson’s R?**

The Pearson correlation coefficient, also called Pearson’s R, is a statistical calculation of the strength of two variables’ relationships. In other words, it’s a measurement of how dependent two variables are on one another.

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

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

**Scaling**

In scaling (also called min-max scaling), you transform the data such that the features are within a specific range e.g. [0, 1].

x′=x−xminxmax−xminx′=x−xminxmax−xmin

where x’ is the normalized value.

Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important.

**Normalization (or normalizing)** :

Scaling some data to a confined range.

ex) {1,2,3} → {0, 0.5, 1} (by normalization in the reference of [0.0, 1.0](;from 0.0 to 1.0) range)

For this, each value should follow : result = ( original - 1 ) / (3–1)  
= { original - the\_smallest\_value\_in\_range } / (the\_biggest\_value\_in\_data–the\_smallest\_value\_in\_data)

In this case, 1( The smallest value in range) took the role of the reference point. The reference point depends on data itself. It’s different with scaling.

In image processing and computer vision the range is changed from [0, 255] to [0, 1] for convenience of mathematical operation.

**Standardization** :

It’s quite different from the two operations above. It’s statistical operation.

ex) {1, 2, 3} → {-1/0.816, 0, 1/0.816} (by standardization)

mean(:average) : (1+2+3)/3 = 2, standard deviation: SQRT((1–mean)^2+(2–mean)^2+(3-mean)^2}/3) = SQRT((1+0+1)/3)) = 0.816

result = (original - mean)/standard\_deviation

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

The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation.

When a dummy variable that represents more than two categories has a high VIF, multicollinearity does not necessarily exist. The variables will always have high VIFs if there is a small portion of cases in the category, regardless of whether the categorical variables are correlated to other variables.

**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 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.

