**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: There are **4 important categorical columns** which are having **effect on Target variable**.

* **Year** = as per data dictionary only 2 years are there. 2018 is represented as 0 and 2019 as 1. This categorical column lies in top 3.
* **Season**\_4 = this is also a column which comes under category Season. We created dummy variable from the season column as this column contained more than 2 options. This column also lies in top 3 variables.
* **Mnth\_9** = this is also a column which comes under category Months. We created dummy variable from the month column as this column contained more than 2 options. This column lies in top 5 variables.
* **Season\_2** = this is also a column which comes under category Season. We created dummy variable from the season column as this column contained more than 2 options. This column also lies in top 5 variables.

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

Ans: drop\_first True deletes the 1st dummy variable after creation of dummy columns and also deletes the original columns. It prevents **Multicollinearity** and **Overfitting** in the Model.

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

Ans: **temp and atemp** are having highest correlation with the **Target variable ‘cnt’**. Here, temp and atemp represents similar type of variables. As per the data dictionary **temp** represents **original temperature** and **atemp** represents **feeling like temperature**. So, both are having almost similar values. Hence**, highest correlation with Target variable is temp**.

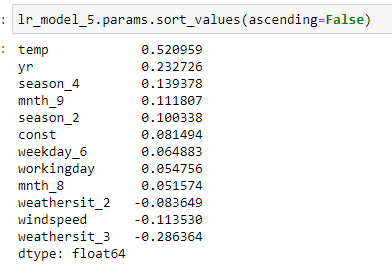
1. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans: Checked 3 things before validating the assumptions.

* P-value = I selected the model where p-value for all variables is lower than 0.05 because more less the p-value more accurate you model is.
* R² value = also, found the R²value and found that the R²value for both and Train and Test data are close to value 1. R² measures how well a statistical model predicts an outcome.
* Multicollinearity = Also checked that the model is having almost null multicollinearity. Hence, overfitting won’t affect the model.
* Error Distribution = Graph for Residuals/Errors gave a Normal Distribution Bell shaped curve which proves that all the errors or residuals are equally distributed all over the data. Which is a best feature for selecting a model.

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

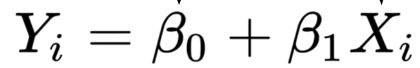
Ans: Top 3 features which are important to this model.

1. **temp** = Target variable is **positively correlated** with the temp feature. A unit increase in temp (Temperature) variable **increase** the bike hire numbers by **0.520959** units.
2. **yr** = **Positively correlated** with Target variable. A unit increase in yr (Year) variable increase the bike hire numbers by **0.232726** units.
3. **season\_4** = this is also having **positive relationship** with Target variable. A unit increase in season\_4 variable increases the bike hire numbers by **0.139378** units.

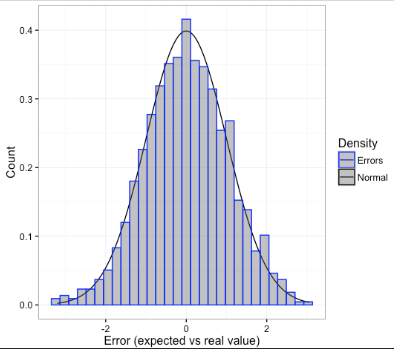
**General Subjective Questions**

1. **Explain the linear regression algorithm in detail.**

Ans: Explaining Linear Regression Algorithm with few words is not possible. Hence, noting down important features in points.

* It is a Supervised Learning.
* It represents the relationship between dependent variables and independent target variable with a single line.
* Linear regression is of 2 types.
  1. Simple linear regression which represents with this formula.  
     
  2. Multiple linear regression and formula is:  
     
* The strengths of the linear regression model are:

1. R² or Coefficient Determination
2. Residual Standard Error (RSE)

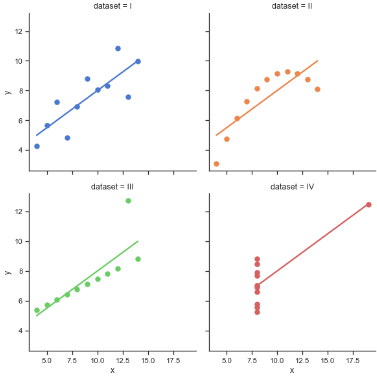
* If R2 value is near by 1 then we can say that our Model is the best fit for the business problem.
* The difference between Actual points and Predicted points are called as residuals. And after plotting the residuals if the distribution is normal distribution (bell shaped curve) then it is a perfect fit model and errors are distributed in the data well.  
  

1. **Explain the Anscombe’s quartet in detail.**

Ans:Anscombe's quartet comprises a set of four dataset, having **indistinguishable** **expressive** **factual** properties in terms of implies, fluctuation, R-Squared, relationships, and direct relapse lines but having distinctive representations when we **scramble plot on chart.**

The four datasets that make up Anscombe's quartet each incorporate 11 x-y sets of information. When plotted, each dataset appears to have a special association between x and y, with interesting changeability designs and unmistakable relationship qualities. In spite of these varieties, each dataset has the same outline measurements, such as the same x and y cruel and change, x and y relationship coefficient, and direct relapse line.

Anscombe's quartet is utilized to demonstrate the significance of exploratory information examination and the downsides of depending as it were on rundown insights. It too emphasizes the significance of utilizing information visualization to spot patterns, exceptions, and other significant subtle elements that might not be self-evident from outline measurements alone.

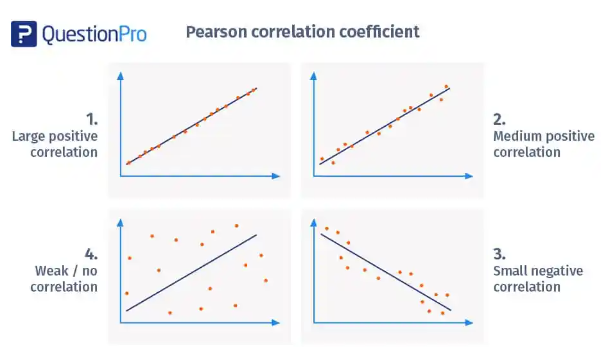


1. **What is Pearson’s R?**

Ans:Pearson's Relationship Coefficient is additionally alluded to as Pearson's r or bivariate correlation. It measures the **linear** **relationship** between **two** **factors**. It contains a numerical esteem that lies between **-1.0 and +1.0**.

At whatever point we examine relationship in measurements, it is for the most part Pearson's relationship coefficient. It cannot capture nonlinear connections between two factors and cannot separate between subordinate and free factors.

Pearson's relationship coefficient is the covariance of the two factors partitioned by the item of their standard deviations.

****

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

Ans: It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations.

In most cases, collected datasets contain a huge number of features. If no scaling is performed, inaccurate modelling will occur because the algorithm only considers size and not units. To solve this problem, we need to perform scaling so that all variables have the same number of digits. It is important to note that scaling only affects coefficients, not other parameters such as t-values, p-values, r-squared, etc.

|  |  |
| --- | --- |
| Normalization | Standardization |
| When the dataset is standardized, it is rescaled to have a mean of 0 and a standard deviation of 1. This is achieved using the following formula: | When a dataset is standardized, it is rescaled to have a mean of 0 and a standard deviation of 1. It accomplishes this using the formula below. |

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

Ans: The variance inflation factor, or VIF, indicates how much collinearity has inflated the variance of the coefficient estimate. (VIF) =1/ (1-R\_1^2). VIF = infinity if there is perfect correlation. A high VIF score denotes a strong connection between the variables. The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4.

VIF displays a complete correlation between two independent variables when its value is infinite. If the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. In order to resolve this, we must remove the variable from the dataset that is producing this perfect fit.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Uses:

* Identify the population that two samples of come from.
* Whether the tails of two samples are identical.
* Whether the distribution shapes of two samples are the same.
* Whether there is common location behaviour between two samples.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.