**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: According to the univariate analysis done wrt to all the categorical variables, the following observations can be inferred:

1. **season vs cnt**: It can be seen that during the Fall season the demand of shared bikes is the most and its lowest during the Spring season
2. **yr vs cnt**: The demand for shared bikes is higher in the year 2019 than in 2018
3. **holiday vs cnt**: median value is more when it is not a holiday but the IQR is wider in case its a holiday
4. **workingday vs cnt**: median is higher when it is a working day but the IQR is wider in case when its not a workingday
5. **weathersit vs cnt:** demand is higher in cloudy weather(category 1)
6. **mnth vs cnt:** demand is higher during the month of September
7. **weekday vs cnt:** demand is higher on wednesday and thursday

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

Ans: It is important to use drop\_first=True during dummy variable creation. This is because otherwise it will produce redundant data which may affect our model. We generally use drop\_first=True while dummy encoding and drop\_first=false while using one-hot encoding. Suppose we want to do a dummy encoding for a column Gender which has 'male' , 'female' and 'transgender' as its category. If drop\_first=False i.e. if the first column is not dropped then it will create three different columns for 'male', 'female' and 'transgender'. This may cause multicollinearity to exist. But if drop\_first=True then only two columns will be made which consists of 'male' and 'female'. When 'male' column has 1 and 'female' as 0 it denotes that the gender is 'male', when 'male' column has 0 and 'female' as 1 it denotes that the gender is 'female' and if both are 0 then it denotes 'transgender'.

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

Ans: Looking at the pair plot, it can be clearly seen that 'atemp' has the highest correlation with the target variable 'cnt'.

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

Ans: One of the major assumptions of Linear Regression was that the **error terms should be normally distributed** after building the model on the training set. This was validated by performing the Residual analysis of the training data. The histogram of the error terms was plotted and it was observed that the error terms follow the normal distribution. We have also checked for **multicollinearity** and removed the variables which displayed a strong linear relationship amongst them by checking on the VIF.

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

Ans: Based on the final model the top 3 features are: **Spring**(season), **yr** and **Light\_Snow/Light\_Rain** (weathersit). **'yr'** shows a positive correlation where as the **'Spring'** and **'Light\_Snow/Light\_Rain '** have a negative correlation with 'cnt'.

**GENERAL SUBJECTIVE QUESTIONS**

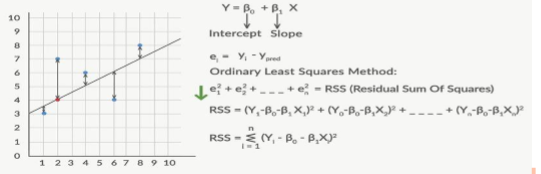
**1. Explain the linear regression algorithm in detail.**

Ans: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. There are two types of Linear Regression:

* Simple Linear Regression: relationship between only one predictor variable and one target variable.
* Multiple Linear Regression: relationship between more than one predictor variable and a target variable

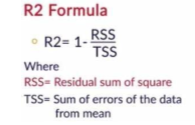
In the first step a scatter plot is plotted between the dependent and the independent variable. A best fit is tried to fit in the scatter plot. The standard equation of this line for simple linear regression is given by the expression of **y = β₀ + β₁x**, where β₀ is the y-intercept and β₁ is the slope of the line that actually means the change in y that takes place due to the unit change in x. For multiple linear regression, the equation of the line is **y = + + + …... +** , where , ,…, are the predictor variables.

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



The strength of the linear regression model can be assessed using 2 metrics: 1. R² or Coefficient of Determination 2. Residual Standard Error (RSE)

R2 provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. So, the higher the R-squared, the better the model fits your data.



**PreProcessing of the Data:** This step is required before the splitting of the data. The Data is first cleaned, all unnecessary columns are removed. Categorical variables are treated with dummy encoding and all the continuous variables are scaled: either by the method of standardisation or normalisation.

**Splitting of Train and Test Sets:** The entire dataset is divided two parts- Training and Test Set. The data is divided in the ratio of 70% Training and 30% Test data or in the ratio of 80-20 percent ratio.

**Adjusted R2:** In case of multiple linear regression, we check the adjusted R-squared value instead of the simple R-squared as adjusted R-squared penalizes the model if it keeps on adding redundant variables whereas R-squared only increases as we add more variables.

**Model Building:** Model is built using feature selection. The various methods for optimal feature selection:

1. Try all possible combinations (2p models for p features)
   * Time consuming and practically unfeasible
2. Manual Feature Elimination
   * Build model
   * Drop features that are least helpful in prediction (high p-value)
   * Drop features that are redundant (using correlations, VIF, usually >5)
   * Rebuild model and repeat
3. Automated Approach
   * Recursive Feature Elimination (RFE)
   * Forward/Backward/Stepwise Selection based on AIC

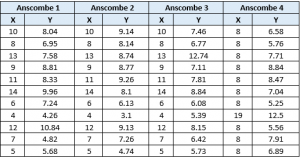
It is generally recommended that we follow a balanced approach, i.e., use a combination of automated (coarse tuning) + manual (fine tuning) selection in order to get an optimal model.

After performing these steps, we need to do a residual analysis of the train data where by the assumptions of linear regression, we need to make check if the error terms are normally distributed and homoscedastic.

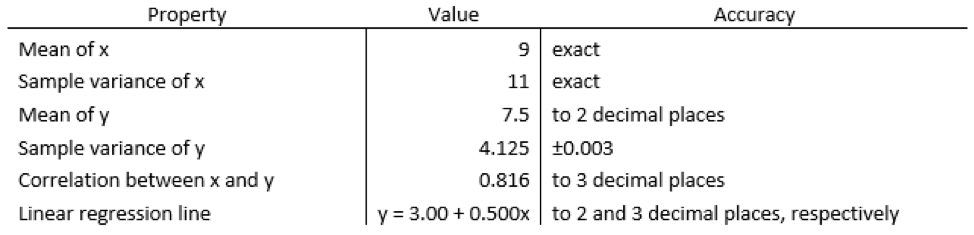
Next, We make predictions on the test set using the final model that we have built with the train set. If we get a R-squared value for the plot of the actual Y and the predicted Y, around the R-squared value we got in our final regression model using the training set, we can say that our model is a fine predictor of our unknown test set and we can go forward with that model. The coefficients we use for our model equation are the coefficients we obtain from the final regression on our training set.

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

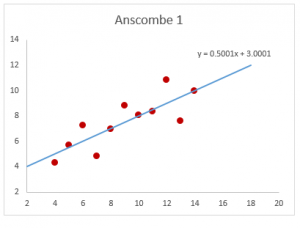
Ans: The statistician France Anscombe constructed the Anscombe dataset in 1973. Anscombe created the dataset to show the importance of visualizing data and also to highlight how outliers can have an effect on the statistical findings of a dataset. Anscombe’s Quartet consists of four data sets, that when examined have nearly the identical statistical properties, yet when graphed the datasets tell a very different story. Each of the datasets in the quartet consists of 11 (x,y) points:



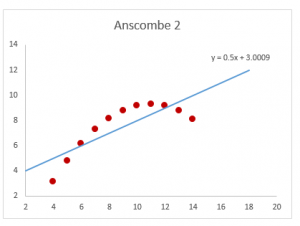
Each Dataset in the quartet consists of the following statistical analysis:



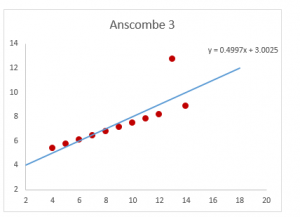
**Anscombe 1** – This graph shows a simple linear positive relationship. It is what we would expect to see, assuming a normal distribution.



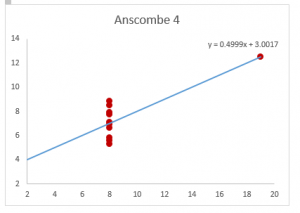
**Anscombe 2** – This graph does not appear to be normally distributed. We can however see a relationship between the 2    variables, it appears to be quadratic or parabolic, but it is not linear.



**Anscombe 3** – This graph is showing a clear outlier in the dataset. The data points, with the exception of the outlier are showing what appears to be perfect linear relationship, but because of the outlier the value of the correlation coefficient has been reduced from 1 to 0.816.

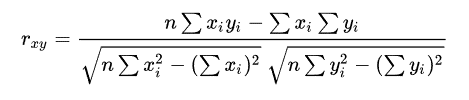


**Anscombe 4**– In this graph we can see that the value of x stays constant with the exception of one outlier. This outlier has created the same correlation coefficient as the other datasets, which is a high correlation, however the relationship between the two variables is not linear.



**3. What is Pearson’s R?**

Ans: Pearson's R also known as the Pearson's Correlation constant, measures the strength of the linear association between two continuous variables X and Y. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.  It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. The coefficient varies between -1 and +1. The formula is as follows:



rxy= Pearson r correlation coefficient between x and y  
n = number of observations  
xi = value of x (for ith observation)  
yi= value of y (for ith observation)

**Assumptions:**

1. Both variables should be normally distributed i.e the normal distribution describes how the values of a variable are distributed

2. There should be no significant outliers as Pearson's R is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient.

3. Each variable should be continuous.

4. The two variables have a linear relationship.

5. The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.

6. Homoscedasticity:  the residuals scatterplot should be roughly rectangular-shaped

**Properties:**

1. Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..

2. Pure number: It is independent of the unit of measurement.  For example, if one variable’s unit of measurement is in inches and the second variable is in quintals, even then, Pearson’s correlation coefficient value does not change.

3. Symmetric: Correlation of the coefficient between two variables is symmetric.  This means between X and Y or Y and X, the coefficient value of will remain the same.

**Degree of correlation:**

1. Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).

2. High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

3. Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.

4. Low degree: When the value lies below + .29, then it is said to be a small correlation.

No correlation: When the value is zero.

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

Ans: Scaling is a technique to standardize the independent features present in the data in a fixed range. In [data processing](https://en.wikipedia.org/wiki/Data_processing), it is also known as data normalization and is generally performed during the data preprocessing step.

**Reason:** While working with a model, it is important to scale the features to a range which is centered around zero. This is done so that the variance of the features are in the same range. If a feature’s variance is orders of magnitude more than the variance of other features, that particular feature might dominate other features in the dataset.

**Normalized Scaling vs Standardised Scaling**:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here’s the formula for normalization:

[Normalization equation](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/Norm_eq.gif)

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0

On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1. If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1.

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here’s the formula for standardization:

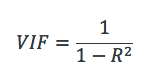
[Standardization equation](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/Stand_eq.gif)

Feature scaling: Mu is the mean of the feature values and Feature scaling: Sigma is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

* Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
* Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Ans: **Variance inflation factor***(VIF)*signifies the extent of correlation between one predictor and the other predictors in a model. It is used for checking collinearity/multicollinearity among the predictor variables. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model. The formula for VIF is as follows:

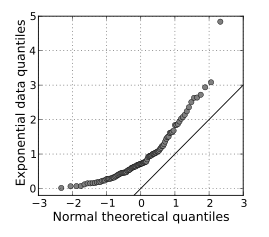


where R-squared value signifies extent to which a predictor is correlated with the other predictor variables in a linear regression. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well). This happens when R2 value becomes 1.

**6. 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. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, it is meant that the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

**Use:** A 45-degree reference line is plotted. If the two sets come from a population with the same distribution, the points 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.



**Importance:** A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as [location](https://en.wikipedia.org/wiki/Central_tendency), [scale](https://en.wikipedia.org/wiki/Statistical_dispersion), and [skewness](https://en.wikipedia.org/wiki/Skewness" \o "Skewness) are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or [theoretical distributions](https://en.wikipedia.org/wiki/Statistical_population). The use of Q–Q plots to compare two samples of data can be viewed as a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) approach to comparing their underlying distributions. A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing [histograms](https://en.wikipedia.org/wiki/Histogram) of the two samples, but requires more skill to interpret. Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary. Q–Q plots are also used to compare two theoretical distributions to each other.  Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a [scatter plot](https://en.wikipedia.org/wiki/Scatter_plot), or even for the numbers of values in the two groups being compared to be equal.