**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?**

* Season: 3 has highest demand for rental bikes.
* I see that demand for next year has grown.
* Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing.
* When there is a holiday, demand has decreased.
* Weekday is not giving clear picture about demand.

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

* In your regression model, if you have k categories you would include only k-1 dummy variables in your regression because any one dummy variable is perfectly collinear with remaining set of dummies.
* **drop\_first=True** is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

* Temp variable has the highest correlation with the target variable.

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

* Error terms are normally distributed with mean 0.
* Error Terms do not follow any pattern.
* Multicollinearity check using VIF(s).
* Linearity Check.
* Ensured the overfitting by looking the R2 value and Adjusted R2.

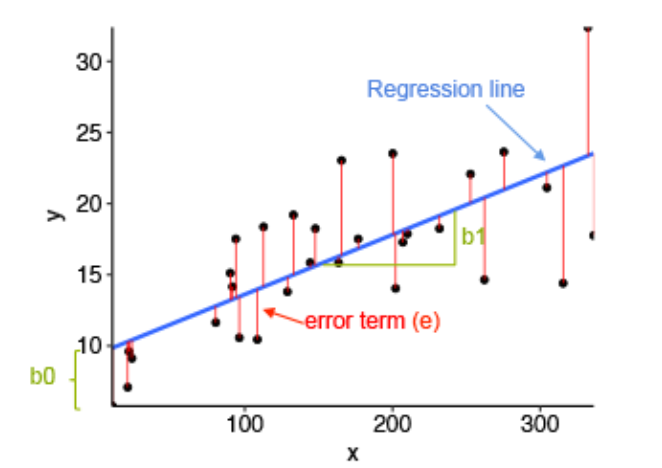
1. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

* According to model if temp increase by 1 unit than 0.51 unit increase in demand of bikes. For windspeed if value increase by 1 unit than bikes demand decrease by 0.14 unit. Same as if weather is of type 3 than bikes demand decrease by 0.28 unit.

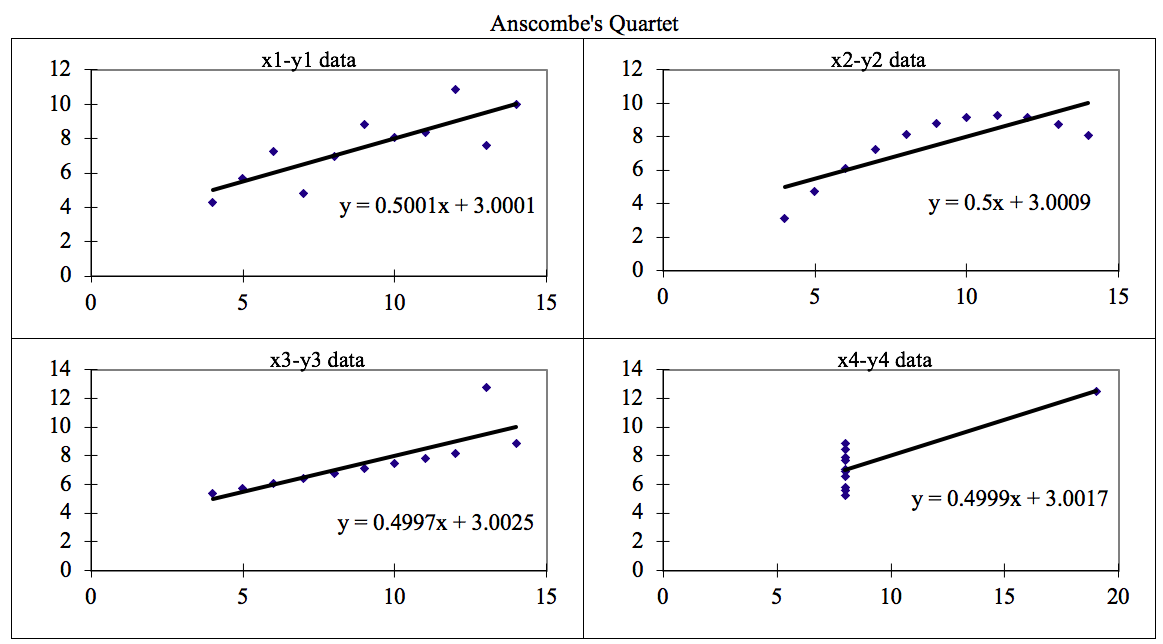
**General Subjective Questions**

1. **Explain the linear regression algorithm in detail.** 
   * Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.
   * Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.
   * Example for that can be let’s say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.
   * Mathematically, we can write a simple linear regression equation as follow y ~ b0 + b1\*x Where y is the predicted variable (dependent variable), b1 is slope of the line, x is independent variable, b0 is intercept(constant). It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.

|  |
| --- |
| Here, x and y are two variables on the regression line. |
| b1 = Slope of the line. b0 = y-intercept of the line. x = Independent variable from dataset.  y = Dependent variable from dataset |

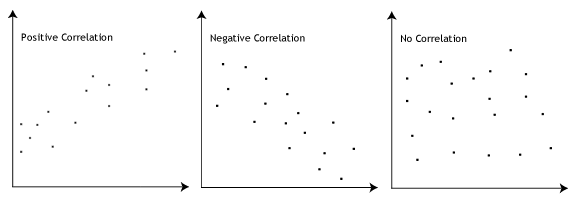
****

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

* Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Each dataset consists of eleven (x,y) points. 
* The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset4: shows the outliers involved in the dataset which cannot be handled by linear regression model
5. **What is Pearson’s R?**

* The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the [linear relationship](https://www.calculushowto.com/types-of-functions/linear-function/#relationships) between two sets of data. Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter “r” for a sample.
* The PPMC is not able to tell the difference between [dependent variables](https://www.statisticshowto.com/dependent-variable-definition/) and [independent variables](https://www.statisticshowto.com/independent-variable-definition/). For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of (.8). However, you could also get the same result with the variables switched around. In other words, you could say that diabetes causes a high calorie diet. That obviously makes no sense. In addition, the PPMC will not give you any information about the [slope of the line](https://calculushowto.com/what-is-a-slope/); it only tells you whether there is a relationship.
* Positive correlation indicates the both the variable increase and decrease together. Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.



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

* It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
* Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
* 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 Network.
* 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.

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

* When there is a perfect relationship then VIF = Infinity whereas if all the independent variables are orthogonal then to each other then VIF = 1.0. Means if a variable is expressed exactly by a linear combination of other variable then it is said that VIF is infinite.

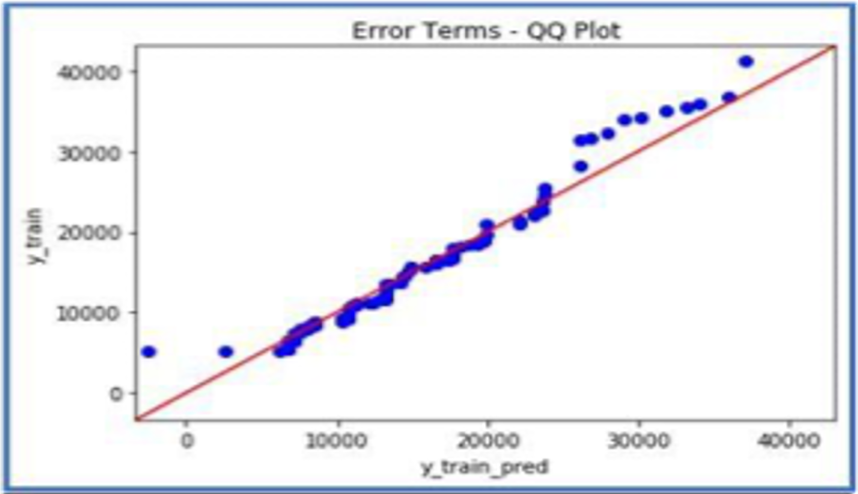
1. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

* Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
* Few advantages:

1. It can be used with sample sizes also
2. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

* It is used to check following scenarios:

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behaviour



* Interpretation:

1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.
2. Y-values < X-values: If y-quantiles are lower than the x-quantiles.
3. X-values < Y-values: If x-quantiles are lower than the y-quantiles.
4. Different distribution: If all point of quantiles lies away from the straight line at an angle of
5. 45 degree from x -axis.