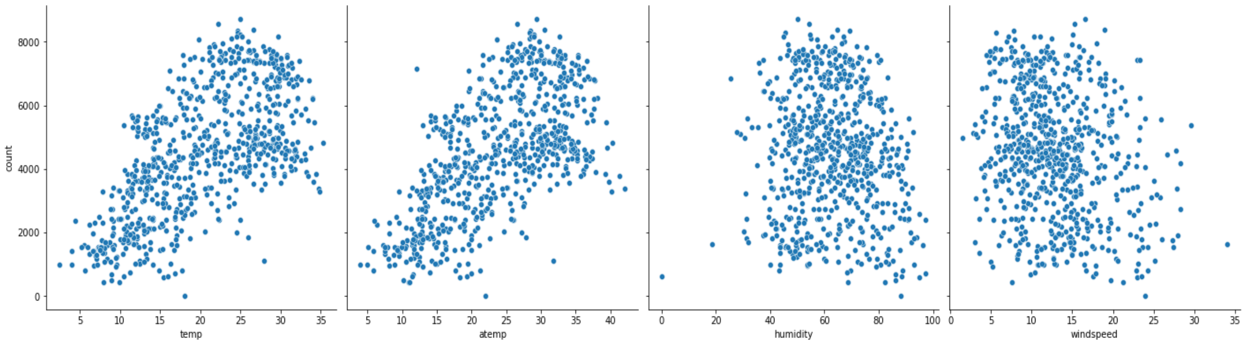
Assignment-based Subjective Questions

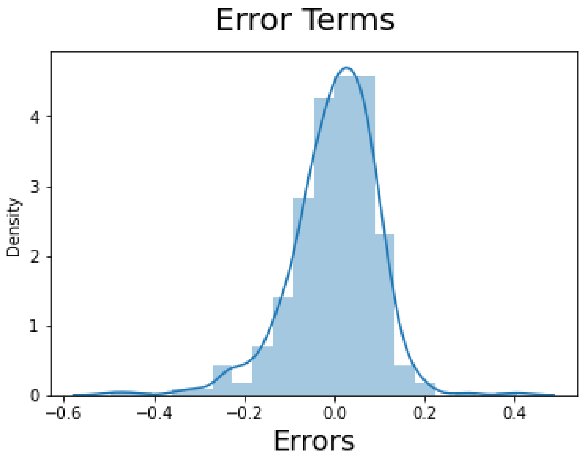
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
    - *Categorical variables are year, month, holiday, spring, month and weathersit.*
      Count = 0.236735+(year×0.239033)-(holiday×0.098796)+(temp×0.386439)-(spring×0.137289)+(month_9×0.075080)+(month_10×0.084182)−(Mist×0.066907)−(Light Snow×0.295987)
    - *From this equation we can conclude that-*
        - *"Year" has coefficient value of* **0.239033** *which indicates that a unit increase in "year" variable, the bike demand will increase by* **0.239033** *units.*
        - *"holiday" has coefficient value of* **-0.098796** *which indicates that a unit increase in "holiday" variable, the bike demand will decrease by* **0.098796** *units.*
        - *"spring" has coefficient value of* **-0.137289***, with respect to "fall", which indicates that a unit increase in "spring" variable, the bike demand will decrease by* **0.137289** *units.*
        - *"month_9" and "month _10" have coefficients value of* **0.075080** *and* **0.084182***, with respect to "month_1", which indicates that a unit increase in variables, the bike demand will increase in both cases by* **0.075080** *and* **0.084182** *units respectively.*
        - *"Mist + Cloudy" and "Light Snow + Rain" has coefficient value of* **0.066907** *and* **0.295987***, with respect to "clear" weather, which indicates that a unit increase in these variables, the bike demand will decrease by* **0.066907** *and* **0.295987** *units respectively.*

2. **Why is it important to use drop_first=True during dummy variable creation?**
    - *It will drop the first dummy variable and helps in reducing the unnecessary columns created due to dummy variable creation. If we include a separate dummy variable for each category, we will introduce multicollinearity in the regression. Thus, if we have N categories, we have to create N-1 dummies. For dropping one dummy variable we use drop_first=True.*

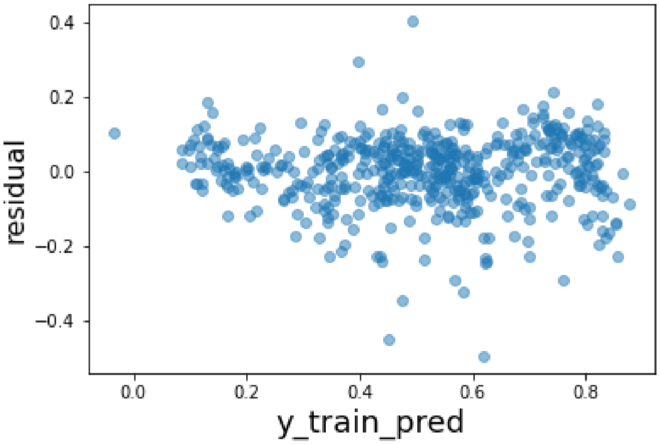
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    - *Here target variable "count" is highly correlated with "temp" i.e temperature.*



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - *From pair-plot, linearity can be checked.*
   - *Error should be normally distributed. For checking it, distribution plot is used.*



   - *Heteroskedasticity can be identified using scatter plot for residual vs predicted values. It should not follow any pattern.*



   - *VIF value should be less than 5 to avoid multicollinearity.*

- *Independence of residuals can be checked by* **Durbin Watson***. The* **Durbin Watson** *(DW) statistic is a test for autocorrelation in the residuals. A value of 2.0 means that there is no autocorrelation detected in the data.*

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - *"temp", "year" and "Light Snow" affect the bike demand as the value of the beta coefficients are high for these three features.*
     - *There is positive correlation between "temp" and target variable "count"*
     - *There is positive correlation between "year" and target variable "count"*
     - *But there is negative correlation between "Light Snow" and target variable "count"*

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- *Linear regression algorithm is a machine learning algorithm which is based on supervised learning. Linear regression is a method to find a best fitted line that gives best fit linear relationship between dependent(target) variable and independent variables(predictors).*

- *We train a model to find the best fitted line and predict the behavior of the data based on predictors. In regression, the target or output variable, to be predicted by linear regression method, is a continuous variable. Linear relation helps in understanding the relationship between data. After building the model, the prediction of output can be done based on the given input variables.*

- *When we are predicting model with single predictor, then it is called simple linear regression. The standard equation of the simple linear regression line is given by –*

$$\mathbf{y=\beta_0+\beta_1 x} \qquad \textit{Single linear regression model}$$

where ***y*** *is the output variable,* ***x*** *is the input variable or predictor,* $\boldsymbol{\beta_0}$ *is the intercept and* $\boldsymbol{\beta_1}$ *is the slope or coefficient of* ***x****.*

- *When we are predicting model with multiple predictors, then it is called multiple linear regression. The standard equation of the multiple linear regression line is given by –*

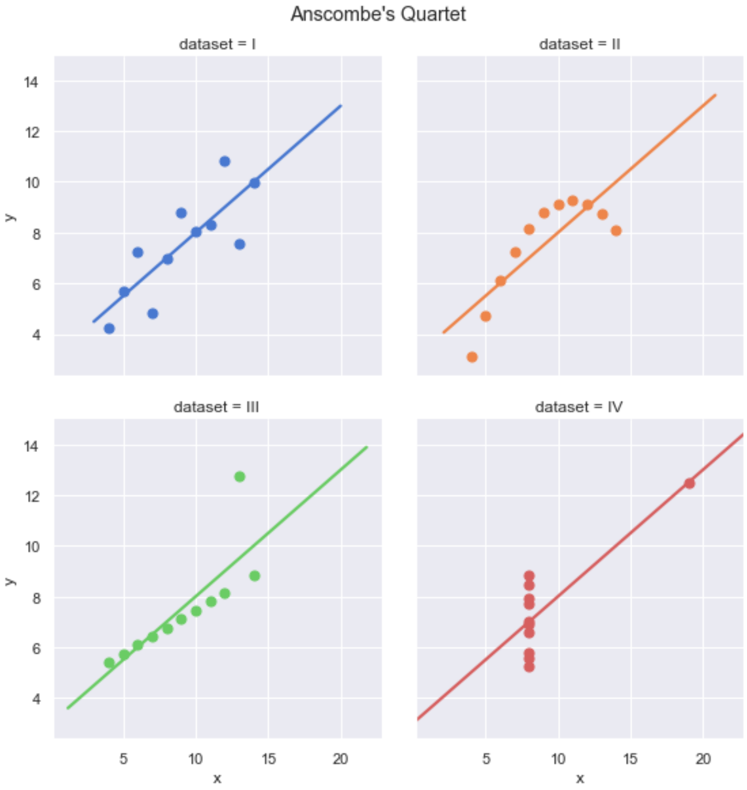
$$\mathbf{y= \beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n} \qquad \textit{Multiple linear regression}$$

where ***y*** *is the output variable,* $\boldsymbol{x_1, x_2,\ldots,x_n}$ *are the inputs or predictors ,* $\boldsymbol{\beta_0}$ *is the intercept and* $\boldsymbol{\beta_1, \beta_2,\ldots, \beta_n}$ *are the slope or coefficients of* ***x****.*

- *For achieving the best fit regression line, the model tries to predict the values of y for which the error difference between true value of y and predicted value of y is minimum.*

2. Explain the Anscombe's quartet in detail.

- ***Anscombe's quartet*** *is a group of four datasets which are having nearly identical descriptive statistical observations i.e. mean, standard deviation and coefficient of correlation, but they have very different distribution and appear very different when plotted on graph. Each dataset consists of 11 data points. There are some abnormalities in the dataset that tricks the linear regression model.*



- *The quartet is often used to demonstrate the importance of visualizing the data graphically before starting to analyze according to a particular type of relationship, and the basic statistical properties are insufficient to describe the actual dataset. This shows us the importance of visualizing data before applying various algorithms to model it. It is recommended to plot the features of the data to see the distribution of the samples. This can help you identify various anomalies in the data like outliers, data diversity, linear data separability, etc.*

3. What is Pearson's R?

- *The degree to which the two continuous variables are correlated is reflected by correlation coefficient between the variables. This coefficient of correlation is called*

*“***Pearson’s R”***. Pearson’s R correlation is a method of measuring the degree of linear relationship between the two variables based on the method of covariance.*

- *Its value lies between -1 to +1. The correlation coefficient of -1 indicates perfect negative correlation which means if one variable increase then other will decrease. The correlation coefficient of +1 indicates perfect positive correlation which means if one variable increase then other will also increase. If the correlation coefficient is 0, then it indicates no linear relationship between two variables.*

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

- *Feature scaling is a method of standardizing the independent features that exist in the data within a fixed range and are executed during the data preprocessing to handle highly variable quantities, values or units. It helps in easy interpretation and fast convergence for gradient descent method.*
- *If the features are not scaled, then machine learning algorithm tends to give high weightage to the features with greater magnitudes and treat smaller magnitudes with lower values, regardless of the units of the values. In other words, the algorithms that are used to calculate the distance between the features become biased towards numerically high values.*
- *There are two methods of scaling. One is “normalized scaling” in which values are scaled between 0 and 1. Second is “standardized scaling” in which values are scaled around zero mean with unit standard deviation. Difference between normalized scaling and standardized scaling are-*
    1. *Minimum and maximum values are used in normalized scaling while mean and standard deviation is used in standardized scaling.*
    2. *Unlike normalized scaling, whose scale values lies between 0 and 1, standardized scaling has no fixed range.*
    3. *Normalized scaling is more affected by outliers as compared to standardized scaling.*
    4. *Normalized scaling is used when the distribution of feature is unknown while standardized scaling is used when the distribution of feature is normal.*

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

- *In multiple regression analysis, the variance inflation factor (VIF) is a measure of multicollinearity. Multicollinearity exists when two or more input variables are linearly correlated with each other. VIF is calculated by the ratio of the variance of all betas of a given model by the variance of a single beta if it were fit alone. This can negatively affect the regression model.*
- *A high value of VIF shows that there is high correlation between the variables. An* ***infinite VIF*** *value indicates that the two variables have perfect correlation with each other. In*

other words, one variable can be explained perfectly by a linear combination of another variable.

- *R^2 value is measure to find how well an input variable is represented by the other independent variables. A high R^2 value indicates that the input variable is extremely related with the other variables. The VIF is denoted by*

$$\textbf{\textit{VIF}} = \frac{\mathbf{1}}{\mathbf{1}-\boldsymbol{R}^{\mathbf{2}}}$$

If two variables are perfectly correlated, then the value of R^2 will be 1. Thus,

VIF = 1/(1-1) = 1/0 = infinite

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

- *Quantile-quantile plots (Q-Q plots) is a graphical method to determine whether two sets of datasets are from populations with a common distribution. Q-Q plots are plots of two quantiles against each other. A quantile refers to a fraction or percentage of data below a given value. For example, 0.3 (or 30%) quantile is the point at which 30% of the data fall below and 70% above that value. A 45-degree angle straight line is plotted on Q-Q plot and if the two data sets are from the same distribution, then these points will fall on this reference line**.** The greater the deviation from this reference line**,** the more evidence to support the conclusion that two data of populations with different distributions are drawn**.***
- *Q-Q plot is employed to visualize following situations-*
  1. *If the two datasets are from population with same distribution.*
  2. *If the two datasets have same location and scale.*
  3. *If the two datasets have similar distribution pattern.*
  4. *If the two datasets have similar tail behavior.*
- *The advantages of using Q-Q plots are as follows-*
  1. *It can be used for different sample sizes.*
  2. *Many aspects of the distribution such as shift in location, change in scale, symmetry change and the existence of outliers can be tested at the same time from this plot.*