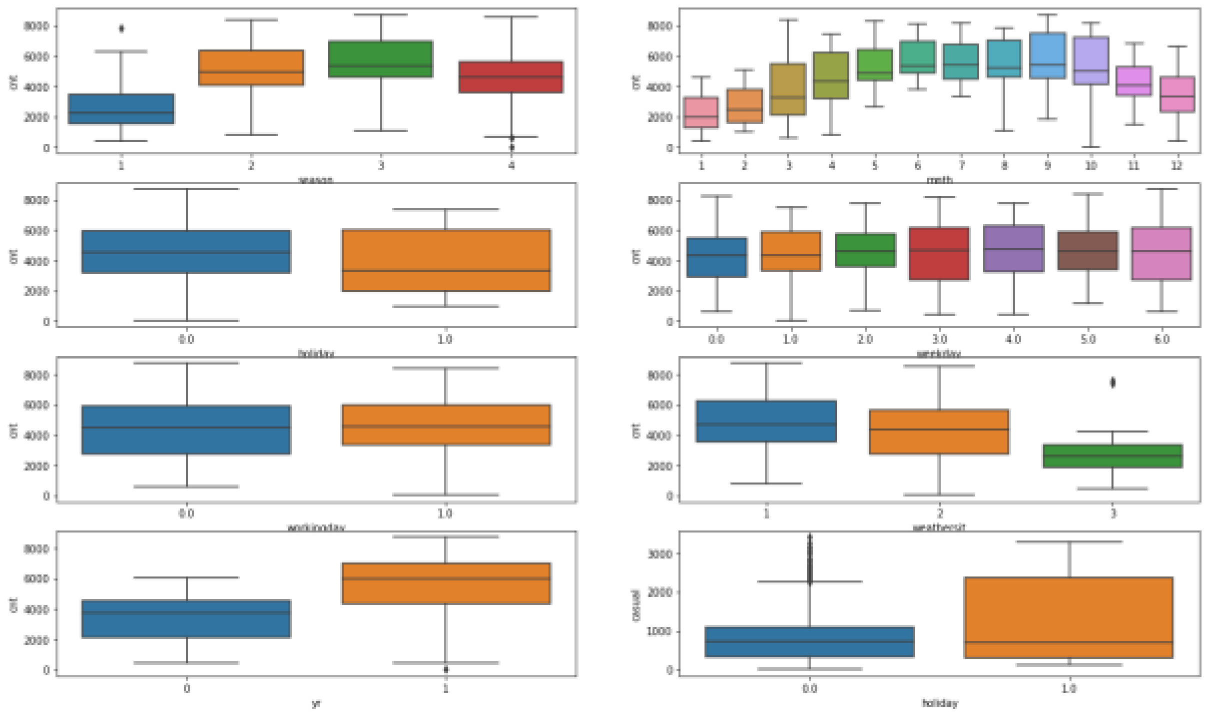
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   While analysing, I shifted cnt column one place up so as to make it like time series data wherein prior day data will always be available for next day analysis.
   Here are key findings about categorical variables after the shift i.e., as consequence of earlier day conditions:

   a. Season: For fall, the count of users is higher compared to other segments. This is followed by summer, winter spring being the lowest. Though the count is higher, there is still overlap that reduces statistical significance.
   b. Weather situation: The count decreases as weather situation category increases from 1 to 3. When weather situation is 3, count is lowest and also has lesser spread. Tried treating weather situation as ordinal variable for linear regression however RFE rejected it (maybe because of shift). (Note: didn’t kept the code in notebook)
   c. Weekday: Though on some weekdays, spread of count was higher and upper quartiles are higher, all weekdays have more or less similar relationship to count.
   d. Year: The count of users was higher in 2019 compared to 2018, though there was overlap, lower quartile of 2019 is higher than median of 2018 indicating that popularity of service was increasing.
   e. Holiday and Working day: On holidays, spread is more and median and Q1 are lower. This might mean that casual users have impact on holiday.
   f. Working day: On working days, spread is more and so is count. This might mean that on working days users use to commute to workplace.
   g. Month: The median count of users was highest in September and lowest in January. The spread was higher in March and October, March having longer head (upper shadow) and October having longer tail. In general, the count increased from January to September and slightly decreased from Oct to December.

   Below are box-plots for univariate analysis of categorical variables. Bivariate analysis didn’t add much interesting observations.



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

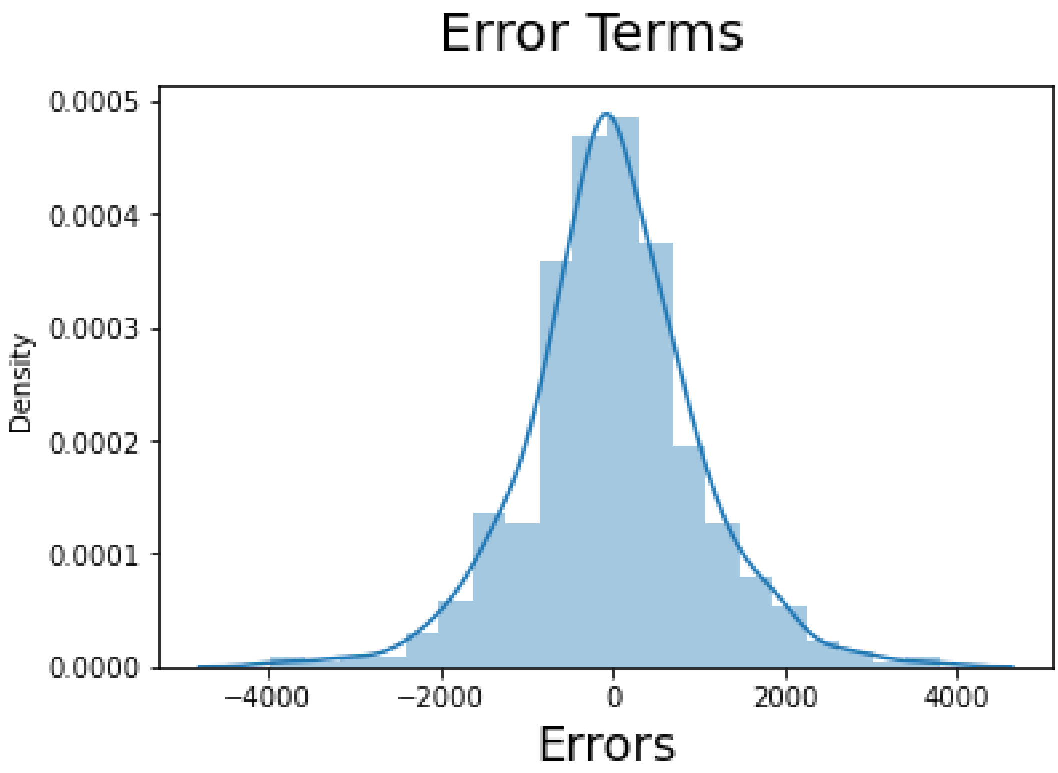
   While dummy variables get created from a categorical column, with one-hot encoding depending on number of categories, many new columns get created. If there are n categories, n-1 columns are enough to represent them. Also, if drop_first = False, an additional column gets created that has big potential to result in collinearity and redundancy and training a model will be more complex without any added advantage. **drop_first=True** ensures that redundant column is not created and helps in keeping model simple.
   e.g. Say Furnishing Status Column has 3 categories: unfurnished, semi furnished and full furnished. While 2 dummy variables get created from this with **drop_first=True,** they are say named semi furnished and furnished. When combined values of these is 00, it means unfurnished, 10 means semi furnished and 01 means furnished.

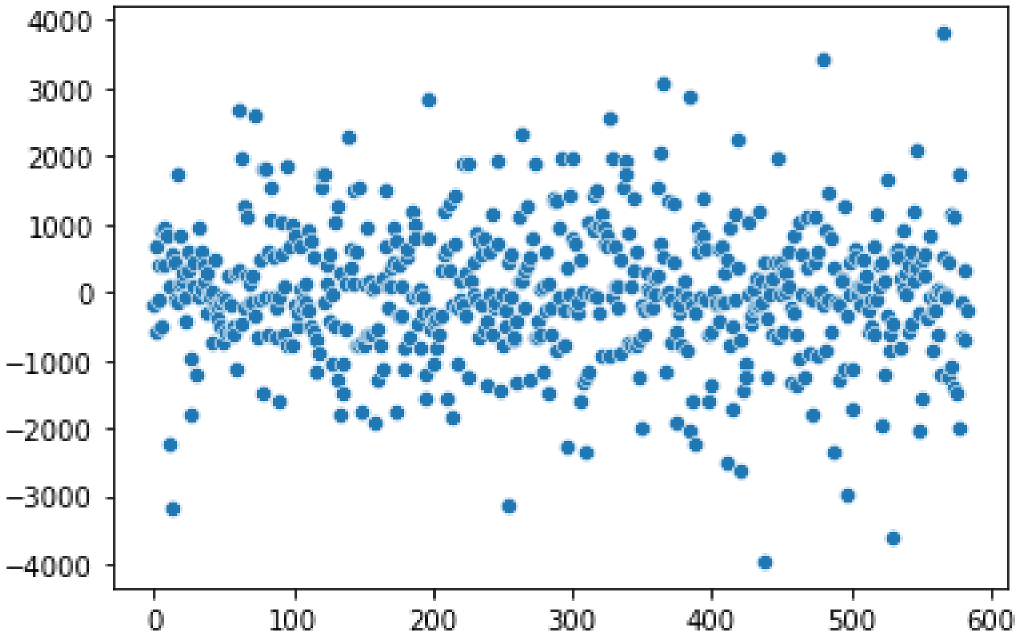
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   ‘registered’ column has highest correlation with target column ‘cnt’.

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

   This was done by plotting histogram for error terms (residuals) and validating that it follows normal distribution. And by plotting a scatterplot of x_train versus error terms and checking the randomness.





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

   ‘registered’ meaning number registered users on earlier day followed by month_6 i.e. month June followed by month_7 i.e. July are top 3 features in order.

General Subjective Questions

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

Linear Regression is supervised machine learning algorithm where output variable is predicted as continuous variable. There are 2 main types of linear regression

A. Simple linear regression: Dependent variable called target/response/outcome variable is function of single independent variable called predictor. And this function is described by a line. A SLR function is of form:
   Y = mx + c
   where c is intercept on y axis i.e. value of y when x = 0
   and
   m is slope of line.
   slope can also be calculated as m = (y2-y1)/(x2-x1) and indicates units by which y changes for unit change in x.
   In layman terms, creating a SLR model involves finding best fit line that minimizes the error where error is difference between actual value of y and predicted value of y
   Meaning e = actual y – predicted y = actual y – (mx+c)
   Practically since error is calculated for each point, this error takes form of RSS (Residual sum of squares) wherein RSS is equal to summation of square of error for all data points.
   The strength of SLR is measured by R-square wherein
   R-square = 1- (RSS/TSS) , TSS being Total Sum of Squares i.e. sum of error of data points from mean of data points for target variable.
   Values of R-square range from 0 to 1 , 1 indicating prefect linear relationship. Lower is the value of R-square , lower is strength of linear regression.
   Linear regression makes below assumptions:
   a. X and y have linear relationship
   b. Error terms follow normal distribution with mean of 0
   c. Error terms are independent of each other
   d. Error terms have constant variance

   Creating a SLR model involves below steps:

   a. Splitting given dataset in training and testing data, training data is used to train model and testing to evaluate
   b. Finding best fit line
   c. Plotting histogram of error terms and scatter plot x terms with error terms to validate assumptions

B. Multiple linear regression: Target variable is function of multiple predictors. A MLR function is of form:
   Y = c + m1x1 + m2x2 + m3x3 …. + mnxn
   Where c is y intercept as in SLR above
   And m1,m2,m3…mn are slopes of lines describing relationship of y with x1,x2,x3….xn in order.
   MLR follows same assumptions of linear regression however MLR model fits hyperplane instead of line and it involves additional things like :

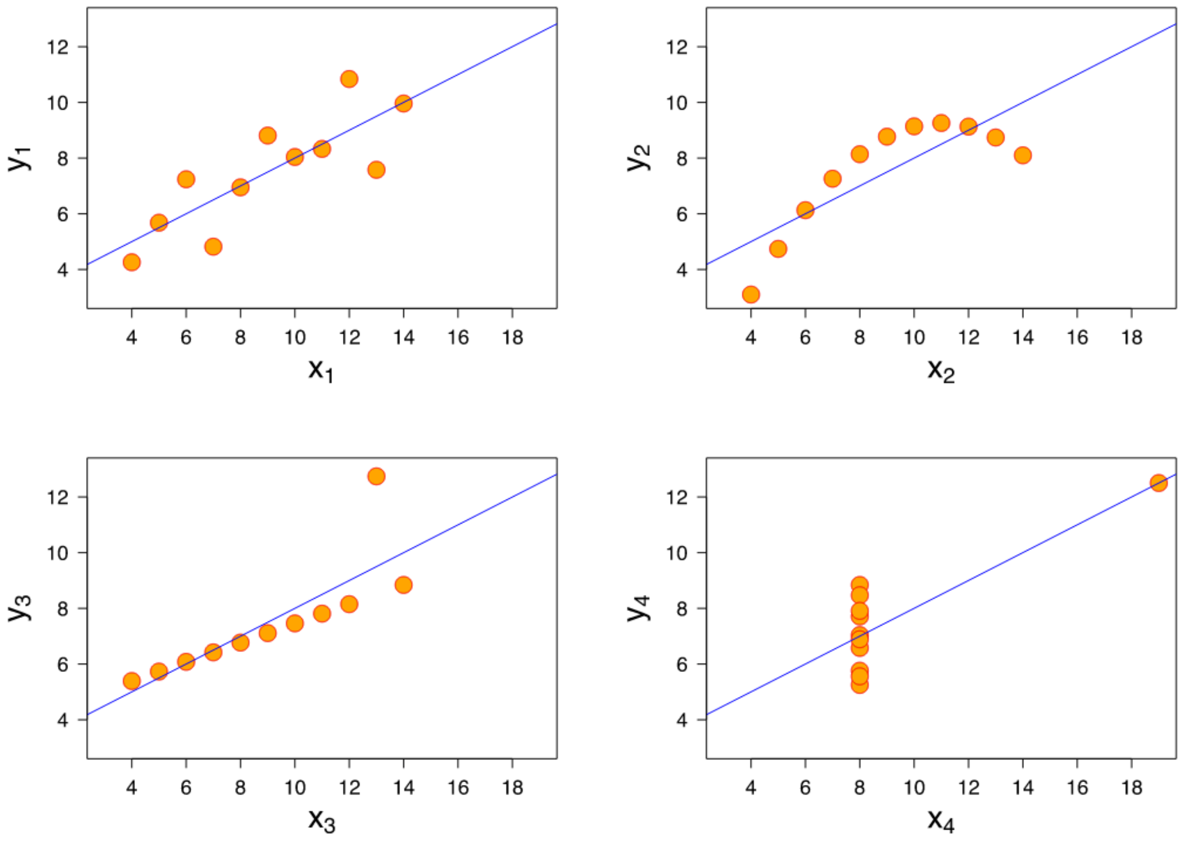
a. Overfitting and Underfitting: Depending on number and slopes of dependent variables, the model may or may not explain the relationship clearly and result in overfitting wherein model might not generalize and behave poorly on testing data and extremely well on training data. Underfitting means the model fails to explain the relationship. Both these can be minimized by proper feature selection.
b. Multicollinearity: One or more independent variables might be interrelated resulting in redundancy. This is normally treated by removing the interrelated variables.
c. P value and VIF: p-value is result of null hypothesis that given variable is insignificant while VIF (variance inflation factor) is how closely one variable is related to multicollinearity. Normally p-value less than 5% is considered good and VIF <=5 is considered good however this can vary depending on use case.
d. Adjusted R-square: This defines strength of MLR. As more and more independent variables get added, normal R-square increases and hence is not best indicator of strength of MLR. Adjusted R-square penalizes R-square for addition of dependent variables.

Though linear regression has predictive power it is more useful in interpolation than in extrapolation where in there are swings beyond range under which model is created.

2. Explain the Anscombe's quartet in detail.

Normally descriptive statistics describes/summarizes a dataset well however Anscombe found a need to plot/visualize the datasets as after plotting one can easily find that same descriptive stats can apply to multiple distributions. Its termed quartet as the findings consisted of 4 datasets that have different distributions but same descriptive stats resulting in same linear regression model.

The dataset distributions are as below (ref: https://heap.io/blog/anscombes-quartet-and-why-summary-statistics-dont-tell-the-whole-story)



And as we can see that in first case linear regression is effective but as we move on it becomes less effective. In second case, we can see non-linear relationship, in 3rd one outliars and in 4th one , all independent values correspond to single dependent value entirely failing linear regression.

3. **What is Pearson's R?**

   Pearson's R or Pearson's correlation coefficient or simply correlation coefficient describes correlation between 2 variables. It is kind of measure of linear relationship between 2 variables. It ranges from -1 to 1, -1 signifies perfectly inverse/negative relationship (y increases as x decreases and vice versa) while 1 signifies positive relationship wherein both x and y increase/decrease together. When this value is 0, it means no relationship.

   It is calculated as ration of covariance of 2 variables divided by product of standard deviations of the same 2 variables.

   Pearson's R just measures linear relationship and cannot determine dependent/independent variables.

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

   During model building, it may happen that different variables have different ranges and/or different units and difference between ranges can be very significant and/or conversion between units also might not be possible. When such a model is being trained, minimizing cost function (e.g. using gradient descent method) , might take lot of

effort. In such cases, scaling is used to bring the variables in similar range by applying a transformation. And it results in faster minimization of cost function.
There are 2 types of scaling:

1. Normalized scaling: Here the variables are brought in range 0 to 1 using transformation as
   (x- minimum(x))/(maximum(x)-minimum(x))
2. Standardized scaling: Here variables maintain their positioning and are converted to standard normal distribution using transformation as
   (x – average(x)/standard deviation of x

Normalized scaling takes care of outliers to a limited extent as values come in same range whereas standardized scaling maintains positions of outliers.

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

VIF is calculated as
1/(1-Rsquare)
When the Rsquare is 1 , the denominator becomes 0 , resulting in infinite VIF.
When VIF is high, normally it means multicollinearity and multicollinearity is minimized by removing the variables. While removing the variables, they are removed 1 by 1 as one might removal might affect the model in significant and objective is to create optional model.

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

Q-Q plot is quantile-quantile plot between 2 datasets (e.g., a sample data set vs a dataset for normal distribution or 2 different datasets) that is used to find out if a dataset follows a particular distribution and/or skewness of the distribution and/or if 2 datasets follow same distribution.
Q-Q plot is a visualization technique where in dataset1 quantiles are plotted against dataset2 quantiles. Along with this a 45 degrees reference line is also plotted. When the deviations from this reference line are smaller, one can conclude that 2 dataset follow similar distributions and if the deviations are larger, it means distributions are different. The deviation from reference can also serve as measure of skewness.
In linear regression one of the important assumptions is normality of error terms.
Q-Q plots can help to determine if error terms are normal also if deviation from reference line is larger, it might indicate that the dependent variables are not significant if mean of error terms is away from 0. And its important to review feature selection for model and/or apply proper scaling method.