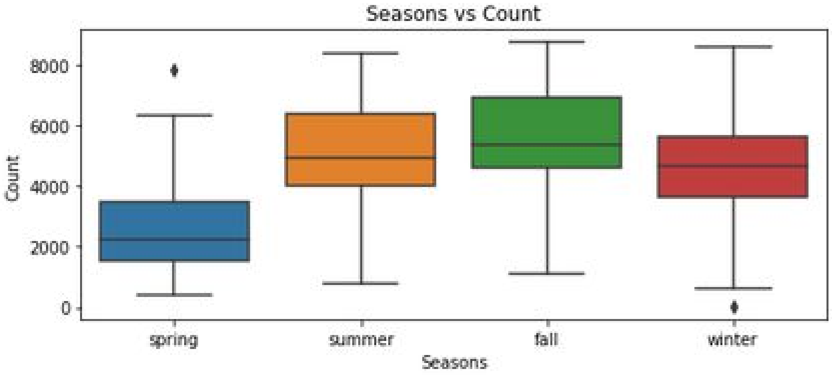
Assignment-based Subjective Questions

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

The categorical variables do have significant impact on the dependent variables. For instance the seasons play a crucial role for the business, months in a year play an important role for business counts.

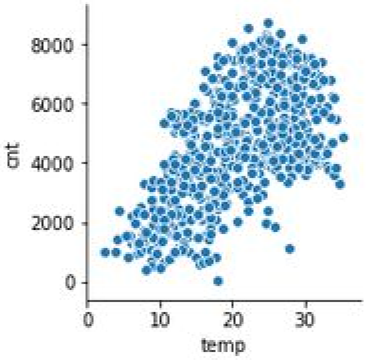


2. Why is it important to use drop_first=True during dummy variable creation?

This avoids multicollinearity since N-1 dummy variables are enough to represent N levels of a categorical variable. The state of dummy variables with all zeros shall act as a variable by itself.

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

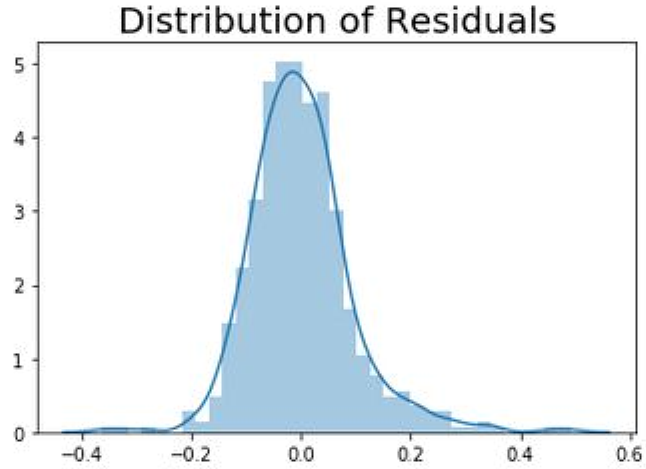
‘Temp’ column has the highest correlation with ‘cnt’. Its correlation coefficient is 0.63.



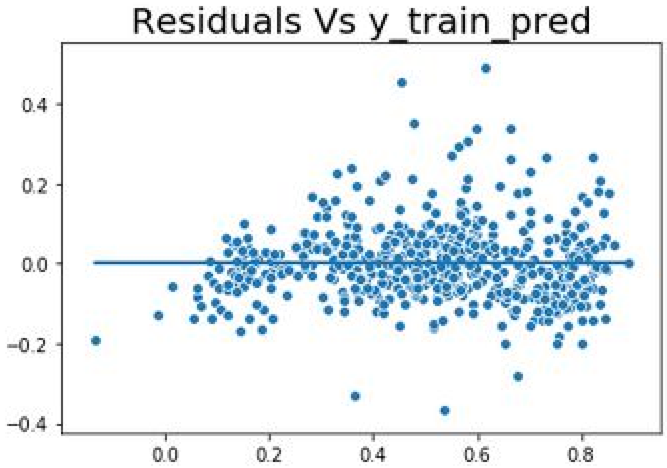
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The assumptions of Linear regression can be validated through Residual analysis.

- **The error terms must be normally distributed with mean as 0.**
  For this we could do a distplot from seaborn for residuals which gives the distribution. QQ plots can help in comparing our data quantiles against the theoretical distribution quantiles.



- **The error terms must be independent of each other**
  For this do a scatter plot between predictions vs its residuals. There must be no pattern.
- **Homoscedasticity**
  The variance of residuals must be constant around zero in the predictions vs residual scatter plot.



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

Temperature, year and winter are the significant contributors since their coefficients are as follows

winter	**0.077431**
const	**0.174331**
yr	**0.238245**
temp	**0.391961**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is used for modelling relationships between a dependent variable which is continuous in nature and one or more independent variables, in a linear way (line or hyperplane) with some assumptions in place. The proper relationship model can be obtained through

1) closed form solution (analytical way)
2) Normal Equations (Linear algebra)
3) Iterative method (Gradient Descent which also has same objective as MLE)

- The model must explain the variance in the data well and must not have multicollinearity.
- Typically we use a train-test split to evaluate our models in an unbiased way.
- To improve the model performance we may use t-statistics, f-statistics, R2. adjusted R2 for evaluation.
- We may also employ feature engineering for this.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a set of four datasets which was used by Francis Anscombe in 1973 to explain the importance of visualisation and flaws in summary statistics.
- The four datasets have similar descriptive statistics but exhibits entirely different characteristics when plotted.
- It also stipulates the importance of outliers analysis.

3. What is Pearson's R?

- Pearson's R or Pearson' correlation coefficient measures the linear correlation between two variables in a dataset.
- It is a normalised covariance given by covariance divided by product of standard deviations.
- Its value ranges from -1 and 1.

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

- Scaling of data is limiting the data to a range of values.
- It plays a very important role in modelling for ease of interpretation.
- It is definitely needed to have faster convergence in iterative algorithms like Gradient descent.
- There are two types of scaling 1) MinMax scaling 2) Standardisation
- Min max scaling brings any data to the range 0-1 given by

scaled_data = (data - min) / (max-min)

- Standardisation makes mean of data as 0 and standard deviation as 1

Scaled_data = (data-mean) / (std_deviation)

- It is recommended to do standardisation in many situations since normalising causes loss of outlier info.
- In situations with image data we must prefer minmax scaling

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

- Having large VIF indicates Multicollinearity
- If a predictor column can be perfectly modelled or predicted with other predictor columns then VIF takes the maximum value which is infinity.
- It is better to remove any predictors with VIF > 5 (implies that more than 80% variance of the predictor variable in question can be explained with other predictor variables).

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

- Q-Q plots are quartile-quartile plots used for plotting quartiles of two variables which could be used for analysing their distributions
- It can be widely used for visualising and comparing theoretical distributions with data sample distribution.
- In Linear regression we may compare the distribution of residuals with theoretical normal distribution.