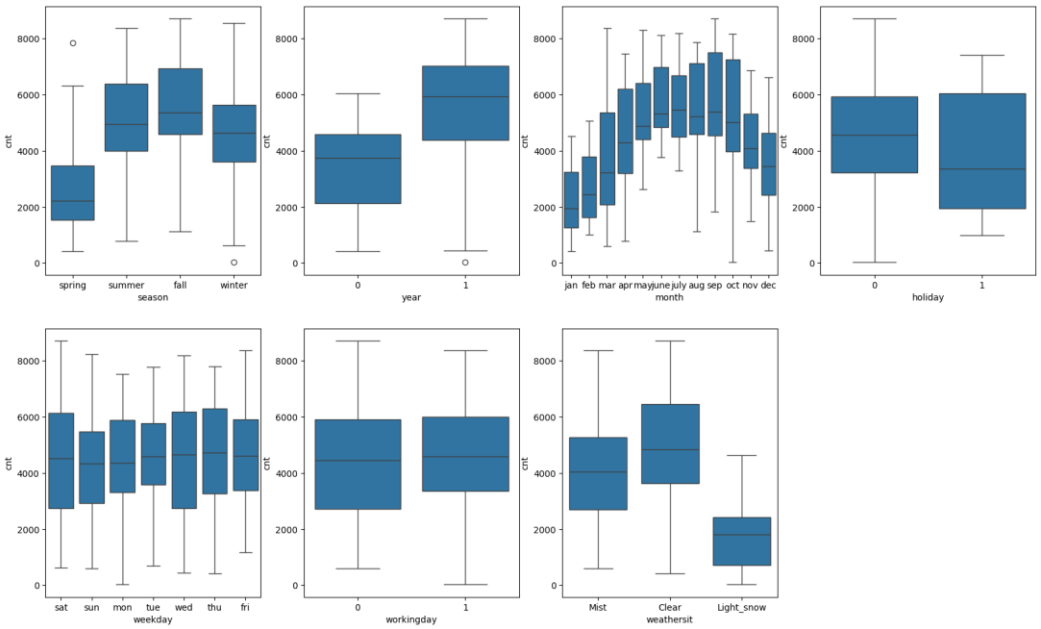
## Assignment-based Subjective Questions

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




I have done analysis on categorical variables using box plot. Below are the points I can infer from the visualisation:

- Fall season has more bookings on bike rentals comparing with other seasons. Similarly, summer season is next highest bookings.
- Booking count has been increased double in 2019 than 2018.
- Trend of booking increases when year starts and ends at end of year nearer to start level. May, June, July, August, September has more bookings than other months and declines nearing end of year.
- Bookings seems to be more in holidays and less in working days implies people spending time with family during holiday / vacation.
- Bookings seems to be equal in both working and non-working days. There is not much difference.
- Similarly, bookings are equal in case of weekdays and weekends. Not much difference on bookings count
- It is obvious that, more bookings during clear weather season as people can book and ride their rental bikes.

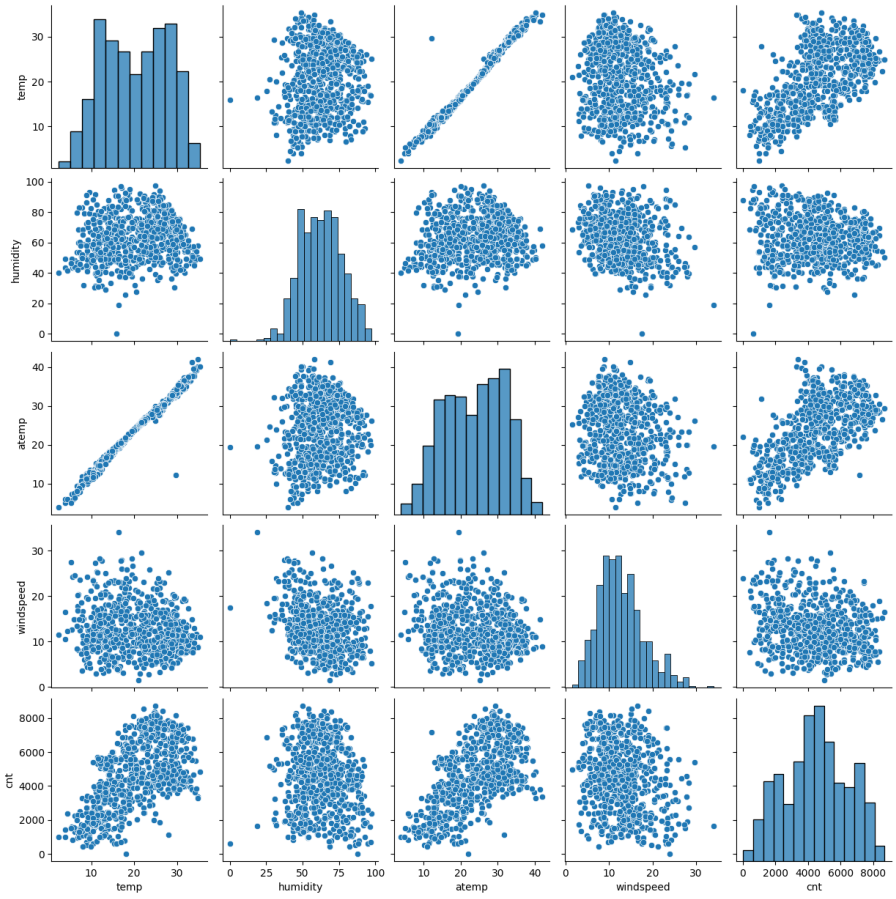
2. Why is it important to use drop_first=True during dummy variable creation?

   Option drop_first=True is not a mandatory field to create dummy variables but its good and required when creating a dummy variable. It helps to reduce columns during dummy variable creation. These extra columns either unnecessary for building the model or can be duplicate or meaning less.

   For ex: in sample housing data, we have three categories called furnished, semi-furnished and unfurnished. Here either “furnished or unfurnished” or “semi-furnished and unfurnished” can be considered and no meaning in keeping 3rd column which gives same meaning as furnished/semi furnished

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

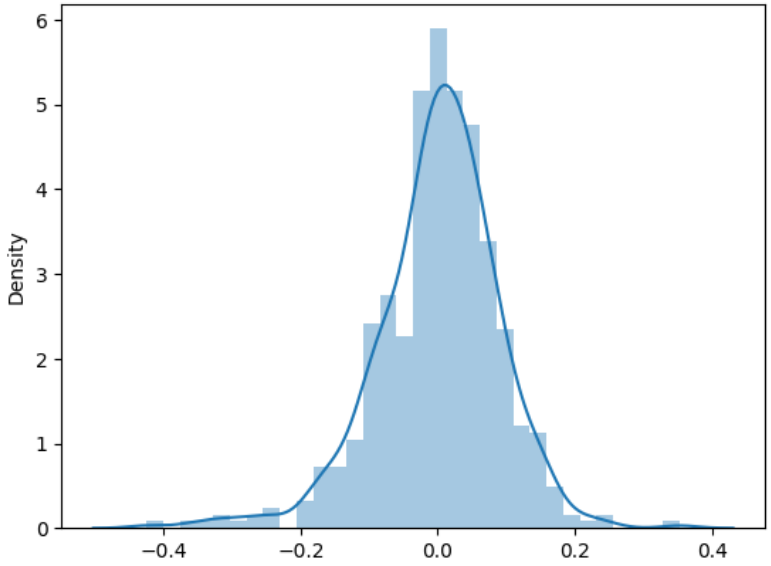
   Variable “temp” has the highest correlation with the target variable.



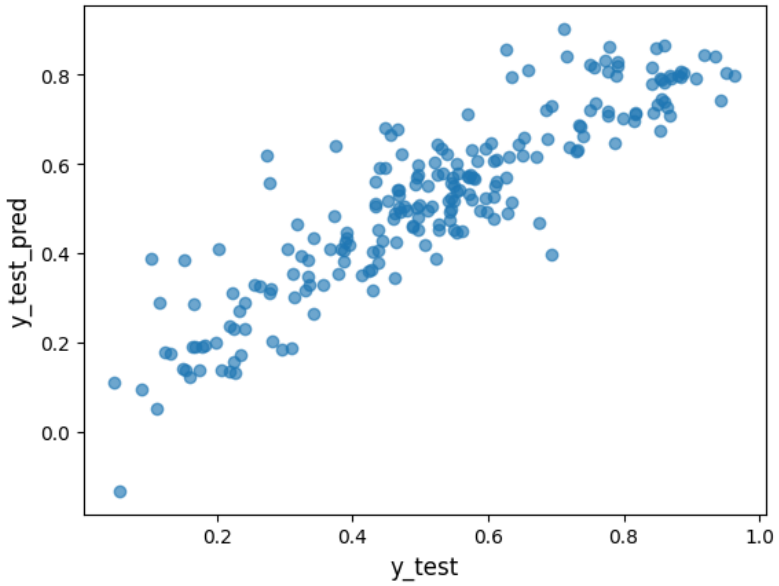
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Assumptions are validated using below steps,

- Residual analysis to identify the error terms distribution. Residual histogram should canter around 0, which is mean=0. This plot will show whether residuals are following normal distribution.




- Evaluating the model by plotting using scatter plot and evaluated the Linear relationship between dependent variables and independent variables.



- Verified to make sure no multicollinearity in the data. Calculated VIF, p-value to make sure the model built is significance.

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

   Based on final model, following are features contributing significantly,

   - Temp
   - Winter
   - Sep

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression is a supervised Machine Learning model in which the model finds the best fit between the independent variables and dependent variables. Linear relationship between variables means, when the value of one or more independent variables will change, the value of dependent variables will also change accordingly.

   Linear regression are two types,

   - Simple Linear Regression
   - Multiple Linear Regression

   Mathematical formula: $Y = mX + c$

   Here,

   Y is dependent variable

   X is independent variable to make predictions

   m is the slope of the regression line

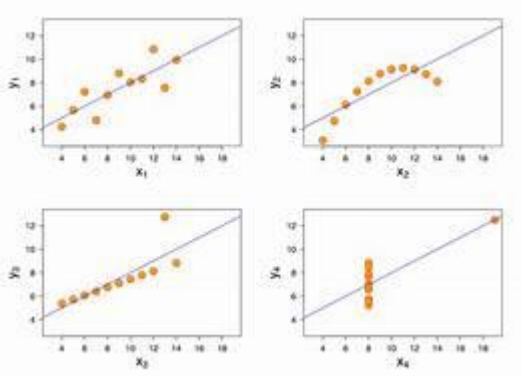
   c is the constant known as the Y-intercept

   Assumptions made in Linear Regression model,

   - Multi-collinearity
   - Autocorrelation
   - Relationship between variables
   - Normal distribution on errors terms
   - Homoscedasticity

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. This means, it looks similar in statistical properties but looks different in graphical representation. Each data set consists of eleven points and was considered to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other data anomalies on statistical properties.



Brief overview on above graphs,

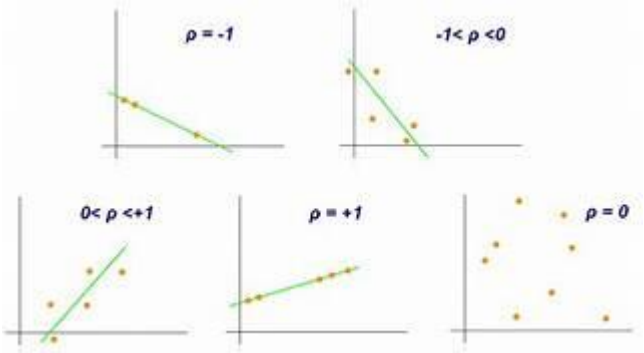
- Dataset I follow a simple Linear relationship between x and y variables, fitting a linear regression model well,
- Dataset II is not suitable for linear regression because it forms a clear curve, indicating a non-linear relationship,
- Dataset III contains an outlier that affects the regression line significantly, despite the rest of the data following a linear trend.
- Dataset IV also has an outlier that, in this case, results in a regression line that is not representative of the data distribution.

The quartet is a powerful illustration of why visual data exploration is essential for a comprehensive understanding of any dataset.

3. What is Pearson's R?

   is a statistical measure that expresses the extent of a linear relationship between two variables. It is denoted as ( r ) and ranges from -1 to 1. Here's what the values indicate:

   1: A perfect positive linear relationship
   -1: A perfect negative linear relationship
   0: No linear relationship




   Values closer to 1 or -1 indicate a stronger linear relationship, while values closer to 0 indicate a weaker linear relationship. It's important to note that Pearson's R only measures linear relationships and may not accurately reflect non-linear relationships.

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

   It is performed during data pre-processing step. Feature scaling is a technique to standardise the independent features present in the data in a fixed range and comparable scale.
   If feature scaling is not done, greater values are considered as highest and smaller values are considered as lowest.
   There are two types of scaling,

   - Normalised Scaling
   - Standardised Scaling

   Normalised Scaling:

   - Minimum and maximum values are used for scaling.
   - Scaling values between 0 & 1 or 1 & -1.
   - Python module Scikit-Learn provides a method called MinMaxScaler for Normalisation
   - It used on the features are of different scales to make it comparable scales.

- It is affected by outliers.

Standardised Scaling:

- Mean and Standard deviation used for scaling.
- It is used to make sure mean and standard deviation is zero.
- It is not bounded to any certain range.
- It is not much affected by outliers.
- Python module Scikit-Learn provides a method called StandardScaler for Normalisation

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

   In case large VIF value, variance of model co-efficient is inflated by the value 4 due to multicollinearity.

   Perfect correlation intends VIF to stay infinite. This means there is perfect correlation between independent variables. This kind of correlation provides R2 as 1 which leads to infinity. We need to drop a variable causing multicollinearity to fix this problem.

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

   It is a scatter plot called as Quantile-Quantile plot. It is a plot of quantiles of the first data set against the quantiles of the second data set. This is used to compare the shapes of distribution. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

   - If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.
   - Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.