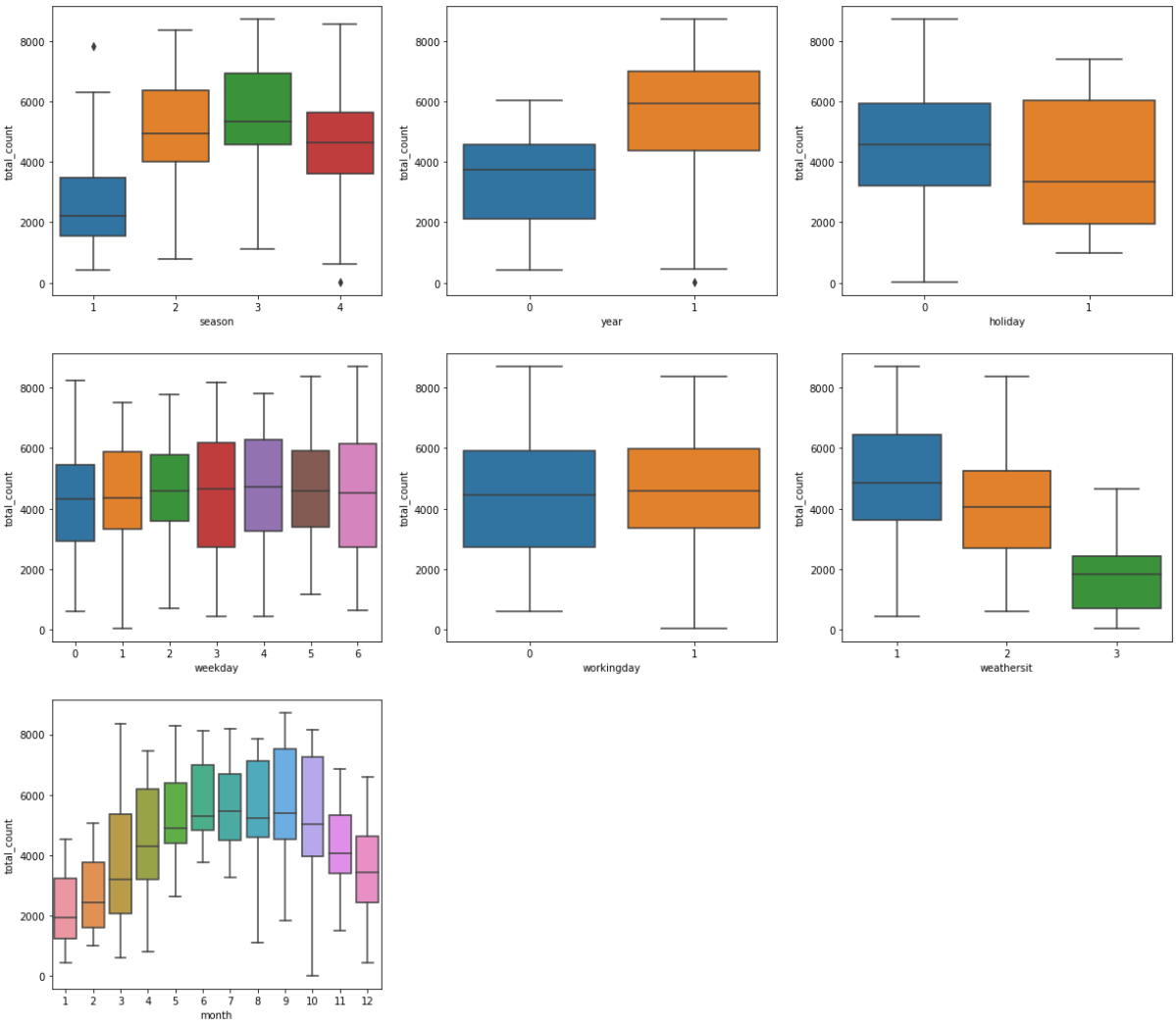
Assignment-based Subjective Questions

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

Below is the visualization of all the categorical variables with the target variable.



The insights from the above plot of categorical variables on dependent variable (total_count) are:

season: Majority of the user prefer to rent a bike in fall season (season_3) because fall is the season between summer and winter and the temperature is moderate, so user prefer to rent a bike in fall season

year: Number of user renting bike in 2019 is more compared to 2018. As in the initial year for any new company, the growth increases. So, the count in 2019 has increases as compared to 2018

holiday: More people are renting bikes when it is a holiday to enjoy the time with family and friends

weekday: on day 3, the count of bikes for rent is usually more compared to another weekday because it is the mid-day of the week and user go to office and hence rents a bike

wokingday: when it is not a working day, the count of bikes for rent is usually more compared to a working day

weathersit: when it is Clear, few clouds or Partly cloudy, the bikes for rent are more

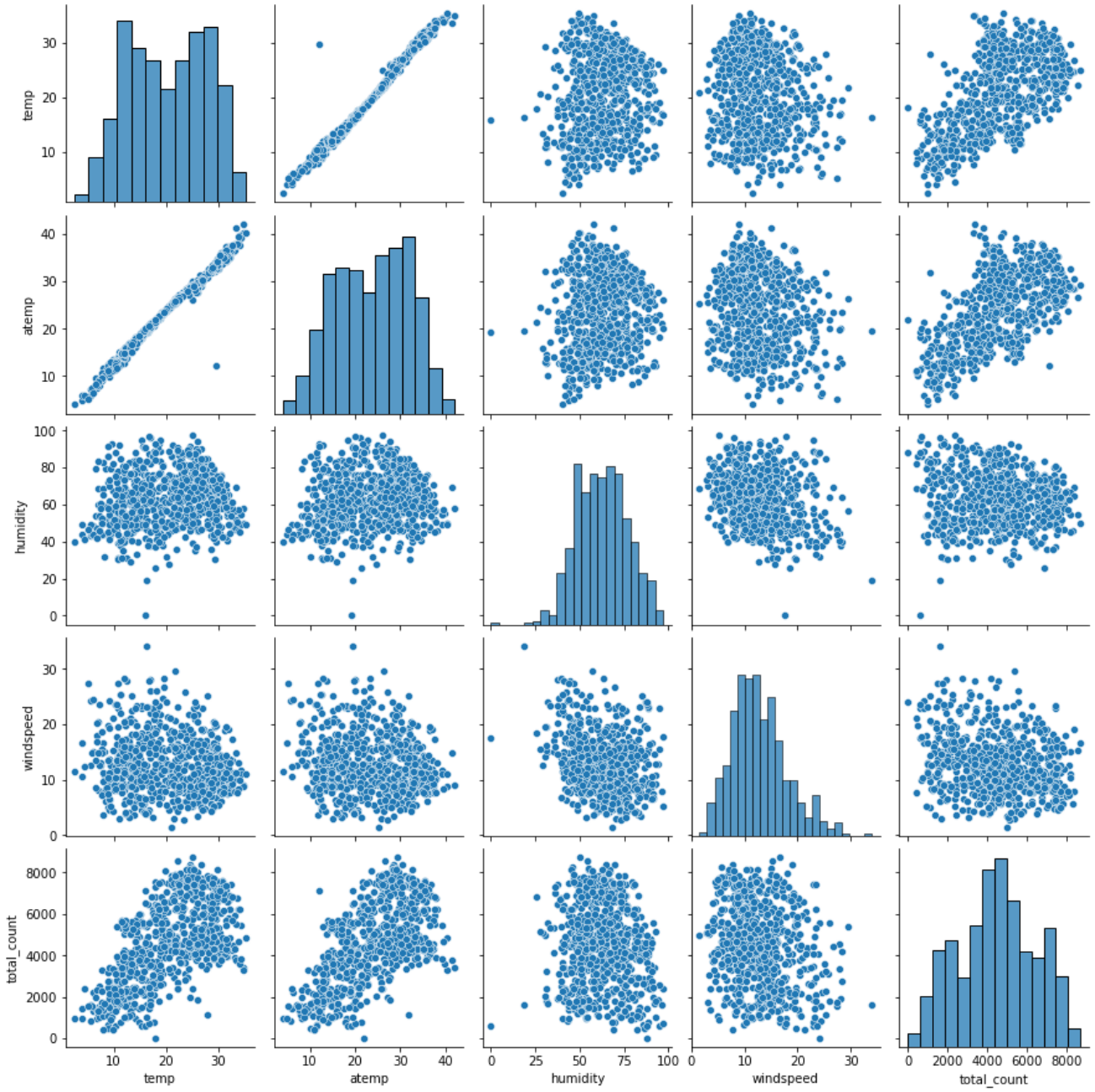
month: Majority of the bikes are being rented on the month of August-September compared to other months

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

During Dummy variable creation, extra column is created for each categorical variable. Using drop_first=True helps in dropping these extra columns which are no more useful. It also reduces the correlation created among dummy variables

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

From the below pair-plot among the numerical variables, we observed that 'temp' and 'atemp' has the highest correlation with target variable, total_count. Both plots are very similar to each other and shows a linear relationship with total_count

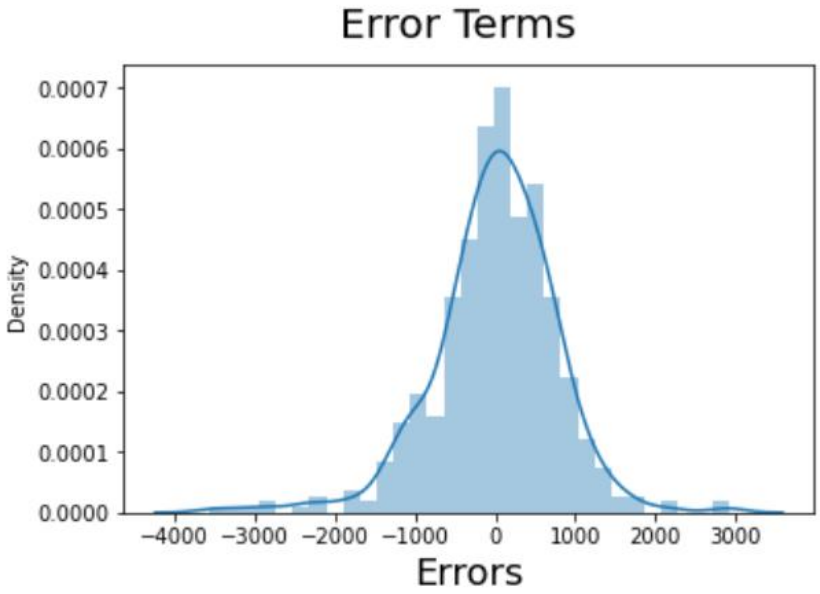


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Validation of the assumptions of Linear Regression after building the model on the training set are as follows:

1. One of the assumptions of Linear Regression is that the Error terms must follow a Normal distribution pattern with mean 0. So, to validate this we perform Residual analysis on the train set. Residual is the difference between the actual y_train and the predicted y_train.

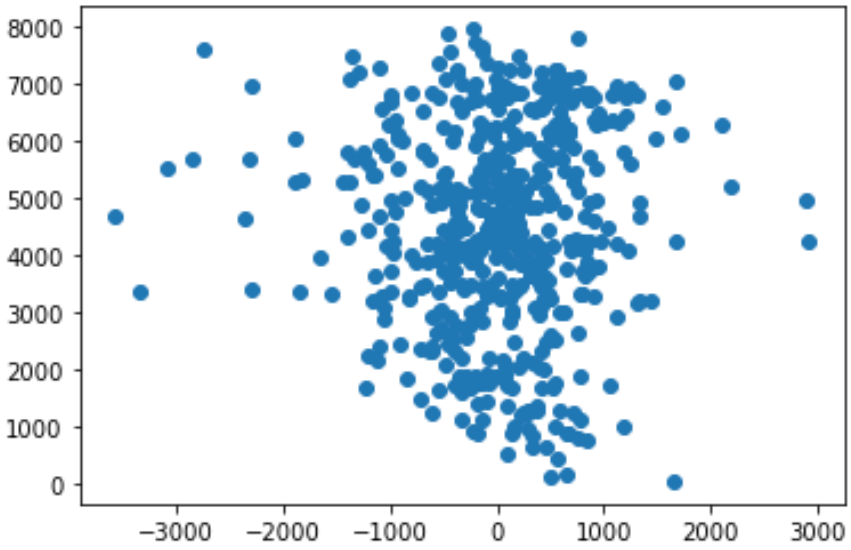
Plotting a distribution plot of residual, we get a normally distributed pattern with mean ~0



2. The next assumption is Multicollinearity i.e., correlation between independent features is very small or there is no multicollinearity between them. This is verified seeing the VIF of the variables in the final model. VIF below 5 is considered as a good VIF.

	Features	VIF
1	workingday	2.76
8	weathersit_2	2.28
6	season_4	2.07
0	year	1.88
3	humidity	1.82
5	season_2	1.64
12	month_10	1.61
2	atemp	1.55
10	month_8	1.49
7	weekday_6	1.41
9	weathersit_3	1.28
11	month_9	1.28
4	windspeed	1.19

3. The third assumption Homoscedasticity which says there should not be any pattern observed when graph is plotted between residual and fitted value. The below plot validates Homoscedasticity assumption.



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

Top three features which are highly significant towards explaining the demand of the shared bikes are:

- year
- workingday
- atemp

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is an ML algorithm used for **supervised learning** for predicting a dependent variable (Y-axis) based on independent variable (X-axis). It shows a linear relationship between a dependent variable and other independent variables. Simple linear regression contains single input variable(X), and Multiple linear regression contains more than one input variable.

Linear regression is represented as: $y = a0 + a1x + \varepsilon$

where, Y = Dependent variable

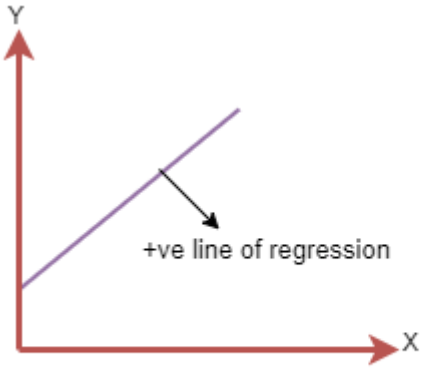
X = Independent variable

a0 = intercept of line

a1 = Linear regression coefficient

ε = Error

A line showing relationship between dependent and independent variable is called **Regression Line**. There are 2 types of RL: +ve line of Regression and -ve line of Regression



The line equation will be: $\mathbf{Y= a_0+a_1x}$

While performing linear regression, our goal is to find the best fit line.

Cost function is used to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

In LR, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error between the predicted values and actual values. To minimize the MSE in LR, Gradient descent is used. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.

The process of finding the best model out of various models is called **optimization** and is done by **R squared method**. Model with highest value of r2 represents a model with less difference between actual value and the predicted value, and hence it represents a good model.

There are few assumptions of Linear Regression:

- There is a linear relationship between target variable and other features
- Multicollinearity i.e., correlation between independent features is very small or there is no multicollinearity between them
- Linear regression assumes that the error term should follow the normal distribution pattern
- There is no correlation in error terms. Error terms are independent of each other

- Homoscedasticity Assumption: Error terms have constant variance and Error term is the same for all the values of independent variables. Also, the variance should not follow any pattern as the error terms change

So basically, in Linear regression we start with reading, understanding, and visualising the data. Then we perform train test split on the cleaned data. Train set contains 70% or 80% and test contains 30% or 20%. Next, we prepare date for modelling. We perform encoding (dummy variables) for all categorical variables and scaling (Min-max scaling) for all numerical variables. Now the model is ready for modelling, so next step is training the model. We fit our model into training set. We drop features having either high p values (>0.05) or high VIF (>5) or both. Then we perform the Residual analysis where we verify the assumption that error terms are normally distributed.

After successfully training our model using the training set, we test the performance of our model using the test set. Here we predict and evaluate the result. We check r2 value of test set and verify if it matched with the r2 value of the training set. There should not be much difference in the r2 values.

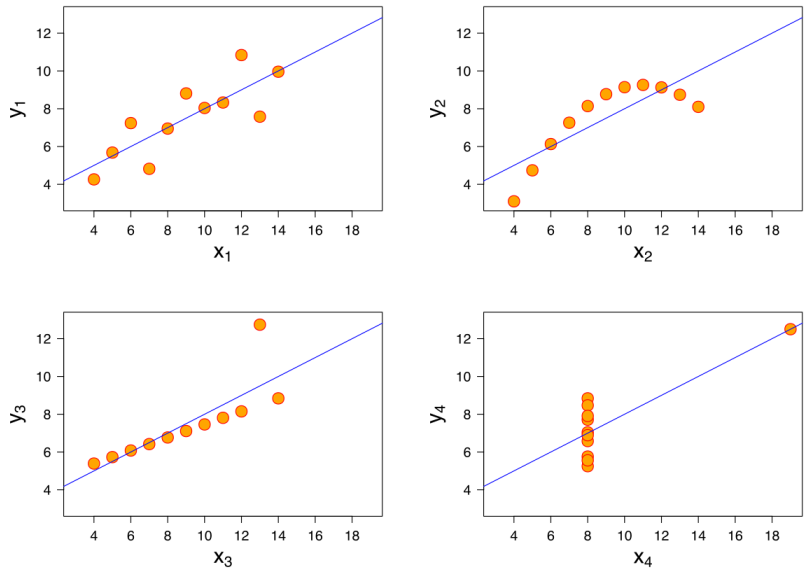
So lastly if the r2 value is verified we conclude the model is good and can find out which are the top features contributing significantly towards explaining the target variable.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

Anscombe's quartet is mainly a group of four data sets which are nearly identical in simple descriptive statistics, but there are some strange behaviours in the dataset. They have very different distributions and appear differently when plotted on scatter plots.

Each dataset consists of eleven (x,y) points.



Observations from above data sets:

Data Set 1: fits the linear regression model well

Data Set 2: cannot fit the linear regression model because the data is non-linear

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model

We observed that Anscombe's quartet helps us to understand the importance of data visualizations to build a well-fit model.

3. What is Pearson's R?

The Pearson's Correlation Coefficient (r) is a statistical measure of the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Pearson's R cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

r = 1 means strong positive relationship. If one increases, the other also increases

r = -1 means strong negative relationship. If one increases, the other decreases

r = 0 means there is no linear relationship

Formula for Pearson r is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

Scaling is nothing but putting the numerical feature values in the same range. Scaling is very important because few variables may have values on a different scale (smaller/higher) compared to other variables. Scaling also helps in speeding up the calculation in an algorithm.

Scaling is performed to bring all the numerical features in the same range. If some feature having value in range of thousands or lakhs while other in the range on tens or hundreds, for example, if salary column has values ranging from 25k to 10L and no of years experienced is in the range 1-25, so the model will take magnitude in account and not the units resulting in wrong or incorrect modelling. Hence, it is very important to scale all numerical feature in the same range before building the model.

There are many types of scaling, but normalized and standardized scaling are popular and widely used.

Normalized scaling: Scaling which makes all numerical feature lie in the range of 0 and 1. One disadvantage of normalization is that it losses some information in the date such as outliers.

Standardized scaling: Standardized scaling replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) 0 and standard deviation (σ) 1.

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

Variance Inflation Factor or VIF tells how much an independent variable is correlated with other independent variables. It detects the multicollinearity in the OLS regression analysis. VIF below 5 is a good VIF. And VIF above 10 shows high correlation and should be removed.

If there is a perfect correlation between two independent features, then the VIF is infinite.

VIF is defined as:

$$VIF = 1 / (1-R2)$$

So, if r2 is 1 then the denominator becomes 0 and hence the VIF i.e., 1/0 is infinite. This means that variable is fully explained by some other variable in the model and hence does not make any sense to keep this feature in the model.

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

Quantile-Quantile or Q-Q plot is a plot of the quantiles of the theoretical set against the quantiles of the sample set. It helps us understand if a sample comes from a known distribution such as normal distribution. In Regression, we use Q-Q plot to check if the data in the sample is normally distributed. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed.

If two distributions being compared are similar, the points on Q-Q plot will lie approx. on the line y=x

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions

Q-Q plot helps us to determine if two population are of the same distribution, if residuals follow normal distribution and if there is any skewness in the distribution.

If date sets, we are comparing are of the same type of distribution type, then plot would be a straight line.