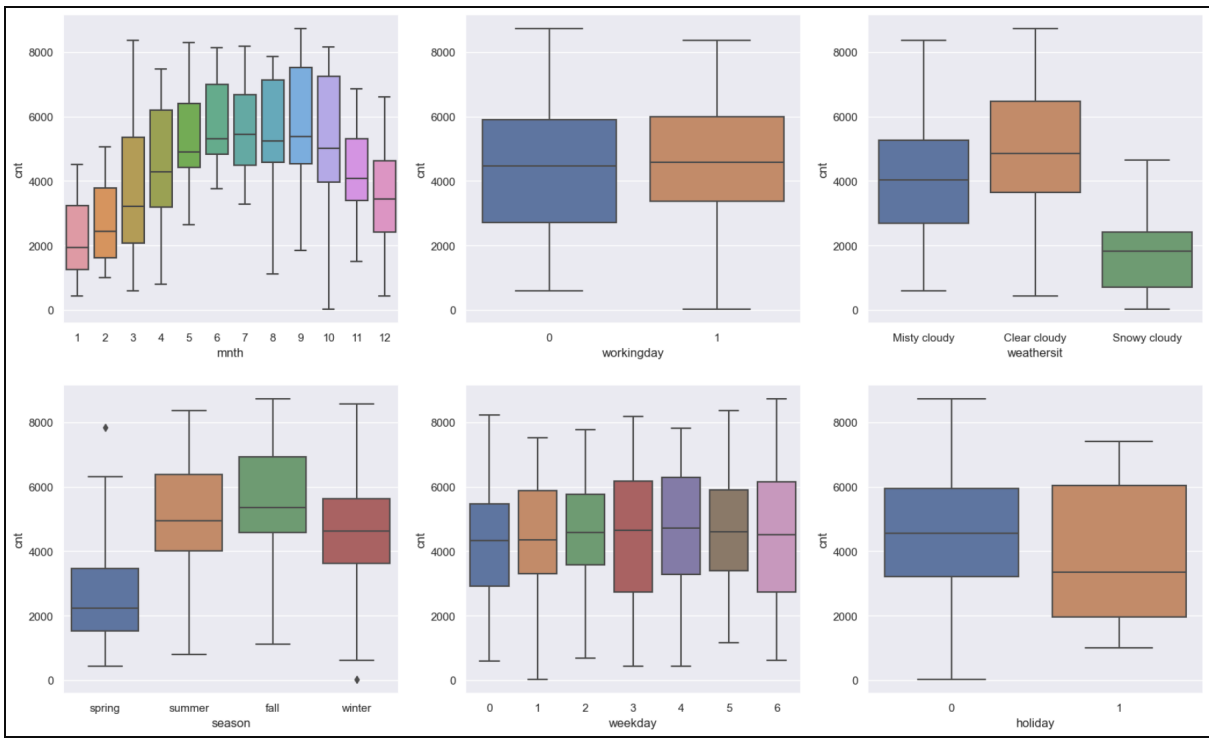
Assignment-based Subjective Questions

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

I created the below box plots to analyse the relationship between categorical variables and dependent variable.



Based on this, we can conclude that demand for rental bikes is generally high:

1. when weather is clear or partially cloudy, makes total sense as people wouldn't want to use bikes in misty or rainy weather.
2. in summer or fall season. Its lowest in spring as there is direct sunlight and temperatures are generally higher and winters as temperature is low.
3. in months which fall in summer or fall season.
4. on weekdays and working days.

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

If there is a categorical variable with n categories, we create n-1 dummy columns with values as 0 and 1. This is because n categories can be represented with n-1 dummy columns.

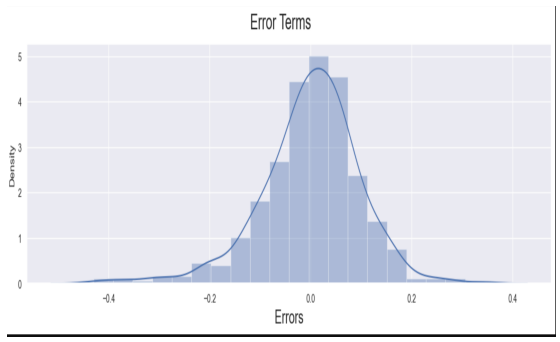
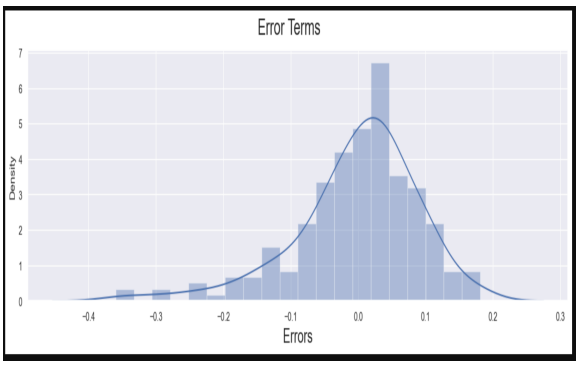
That is where drop_first=True comes in handy, It drops the first category of the categorical so that we end up with n-1 dummy columns. If we don't use it, then n columns will be created.

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

Temperature(temp) or feeling temperature(atemp) has the highest correlation with the target variable among numerical variables

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

1. I did residual analysis on the training data set. I created a distribution of error terms on both training data and test data and it followed normal distribution.

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>Training data error distribution</td>
<td>Test data error distribution</td>
</tr>
</table>

2. For homoscedasticity, I created a scatter plot of error terms and they seemed to be normally distributed with a mean of zero.

3. For Independence/No Multicollinearity, I calculated VIF at each step of model building and removed all features one by one that have a value greater than 5.

4. To check autocorrelation, I used the **Durbin-Watson** test. The statsmodel gave the value of this test as 1.985 which is very close to 2. Hence, we can conclude that there is no autocorrelation between the error terms.

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

If we look at the box plots and feature coefficients the top 3 features are:

1. **Temperature -** If temperature is favourable, demand increases.
2. **Year -** Demand for bikes has increased 0.2 times from 2018 to 2019.
3. **Weather situation** - If weather is favourable, then demand for bikes will increase. Conversely, if its unfavourable it drops significantly.

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 linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x).

Linear Regression is of two types: **Simple** and **Multiple**.

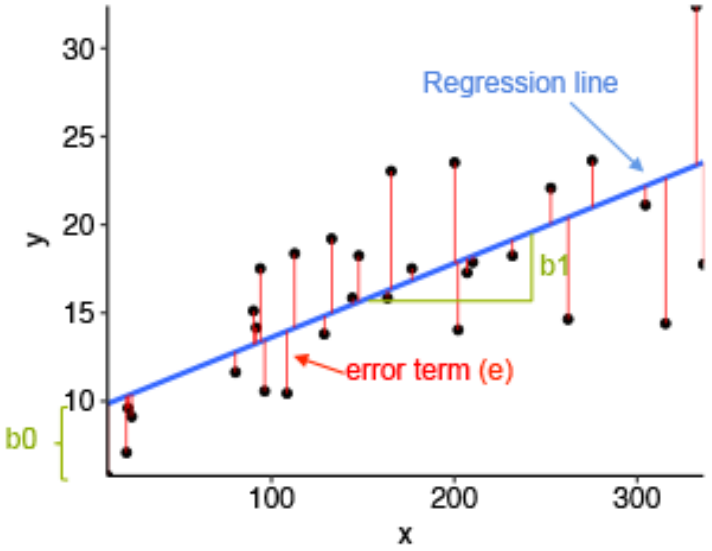
Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable
Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.
Equation of Simple Linear Regression, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

Equation of Multiple Linear Regression, where bo is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.
Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.



In the above diagram,

- x is our independent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- b_0 is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.
- The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

Mathematical Approach:

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

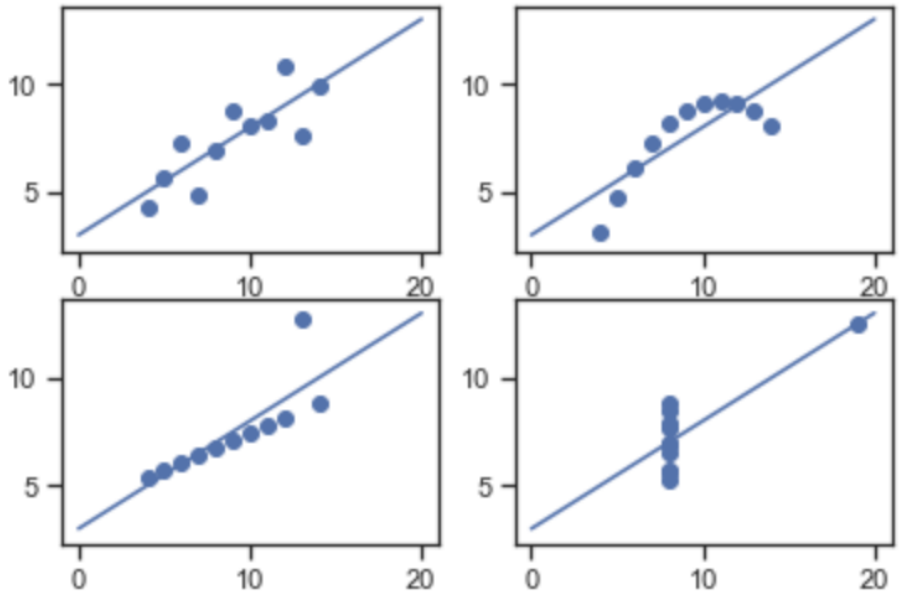
Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2

i.e

$$\sum {e_i}^2 = \sum (Y_i - \hat{Y}_i)^2$$

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician *Francis Anscombe* in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.



- Data Set 1: fits the linear regression model pretty 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.

As you can see, Anscombe's quartet helps us to understand the importance of data visualisation and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualise the data set in order to help build a well-fit model.

3. What is Pearson's R ?

Pearson's correlation coefficient is the test statistics that measures the statistical relationship between two continuous variables. It gives information about the magnitude of the correlation, as well as the direction of the relationship.

Properties:

1. **Limit:** Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.

2. **Pure number:** It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
3. **Symmetric:** Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Different values of Pearson coefficient:

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

When to use the Pearson correlation coefficient

- Both variables are quantitative.
- The variables are normally distributed.
- The data have no outliers.
- The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line.

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

Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set Income is having values in thousands and age in just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

It also helps the machine learning algorithm to converge faster while minimizing the residual errors through gradient descent algorithm.

Normalized Scaling:

Also known as MinMax scaling, it is a scaling technique used to reduce redundancy by bringing all of the data in the range of 0 and 1. It is used to remove the unwanted characteristics from the dataset, and it is useful when there are no outliers as it can not handle them. Mathematically it is represented as:

$$x' = (x - x_{min})/(x_{max} - x_{min})$$

where x_{min} is the minimum x value, x_{max} is the maximum x value.

Standardized Scaling:
It is a scaling technique that is used to convert the feature data in such a way that mean and standard deviation of the feature distribution are 0 and 1 respectively. It is usually applied when the data has a bell curve i.e. it has gaussian distribution. Mathematically, it is represented as:

$$x' = (x - \mu)/\sigma$$

where μ is the mean of the distribution and σ is the standard deviation.

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

VIF for a feature is calculated as:

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

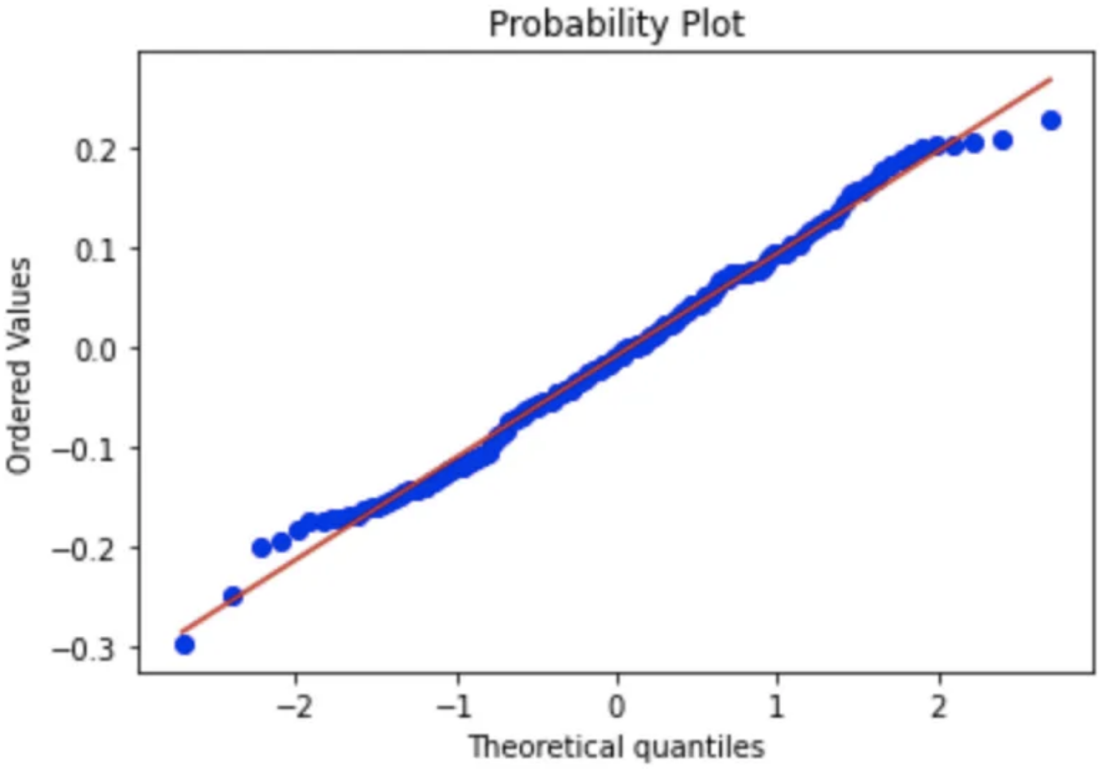
If VIF value for a particular feature is infinite, it means that it has a perfect correlation with rest of the feature variables. When there is a perfect correlation, R^2 is 1 and hence VIF approaches towards infinity.

This means that this feature doesn't add any meaningful information value to the model and should be dropped from the feature list.

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

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in ML algorithms like linear regression, where we split data into train-validation-test to see if the distribution is indeed the same.



As you can observe, the data points lie approximately in a straight line. Thus, we can say the data point is normally distributed.