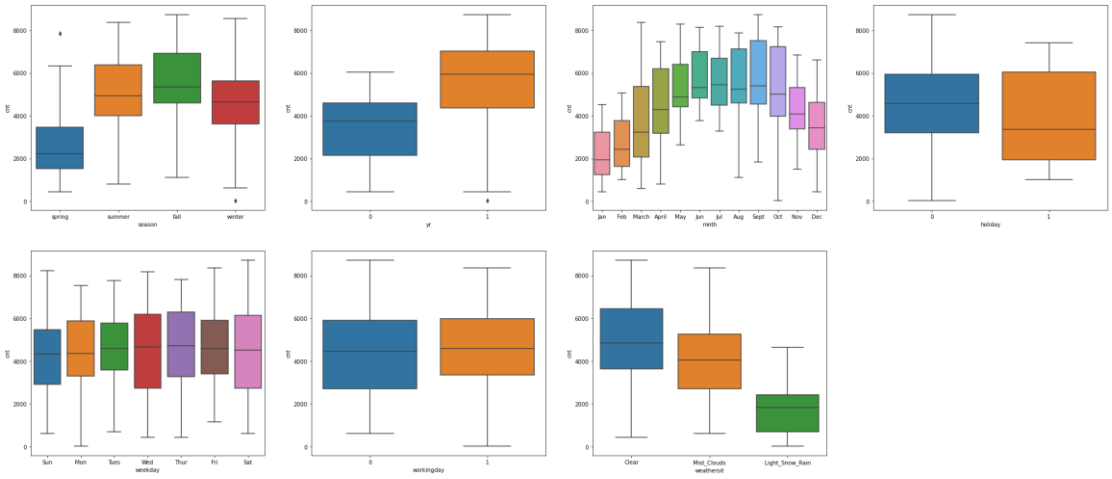
Assignment-based Subjective Questions:

1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



1. Answer:

   Season: Spring shows the least effect on count with a median as low of 2,100 while all other season summer fall and winter have similar effect on count with an approximate median of 5000 counts. Fall shows the highest median and count.

   Yr: The number of people booking bikes increased in 2019 rather than 2018 we can see that the lower quartile in year 2019 is approximately equal to the upper quartile of 2018 as a reason behind the service is being more popular thus year has an impact on the count of bookings made.

   Mnth: September records the highest count of bike bookings and months from may to September are considered the season for bike booking counts while Jan shows the lowest count.

   Holiday: During holidays the median count of booking is less than non-holidays both holidays and non-holidays share an upper limit of 6,000 counts of bike bookings the median for non-holidays is 5,000 compared to 3,000 in holidays.

   Weekday: Across all weekdays from Sunday to Saturday the median of bike booking counts is almost the same days with highest count of bookings are Thursday and Saturday.

   Workingday: Weather it's a working day or not the the average count of bike bookings is the same the counts is more spread across the non-workings days and non- working days show slightly higher counts.

   Weathersit: Clear weather has the highest count of bike bookings while Light_Snow_Rain weather shows the lowest count of bookings no bookings were made in the extreme weather of heavy snow which shows it's unfavorable for all users.

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

2.Answer:

drop_first=True: It helps in reducing the extra column created during dummy variable creation, thus reducing the correlations (Multicollinearity) among dummy variables.

For example, in the season column, we have 4 types of values, and we want to create a dummy variable for that column. If the first variable is summer and the second is fall and the third is winter, then the last is obviously spring. So, we do not need 4th variable to identify the spring season. Therefore, if we have a categorical variable with n-levels, we must use n-1 columns to represent the dummy variables.

3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

3. Answer:

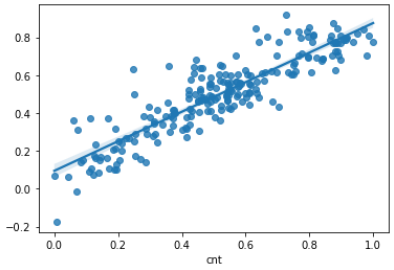
Temp and Atemp has the highest correlation with cnt column (Target).

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

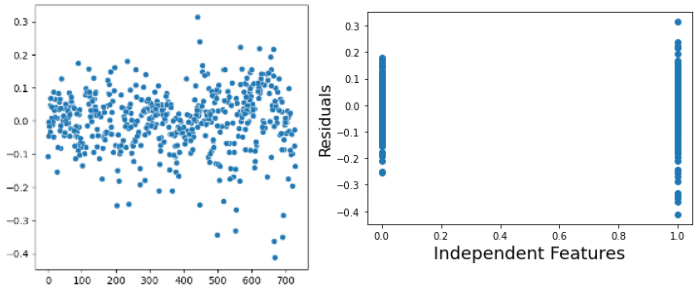
4. Answer:

I validated the assumptions after building the model on the following categories:

4.1 Linear Relationship of variables with the target variable.



4.2 Constant Variance of residual plot homoscedasticity.



4.3 Absence of Multicollinearity by checking VIF's of all variables below 5:

Features	VIF
temp	4.76
workingday	4.04
windspeed	3.44
yr	2.02
Sat	1.69
summer	1.57
Mist_Clouds	1.53
winter	1.40
Sept	1.20
Light_Snow_Rain	1.08

4.4 No autocorrelation checked using durbin_watson:

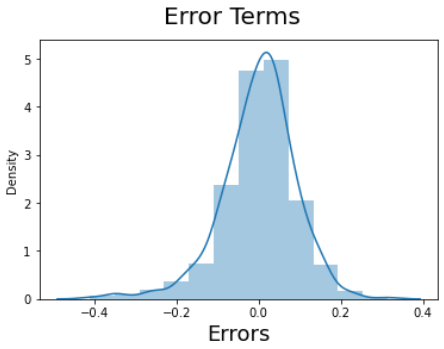
```
1 autocorrelation_assumption(lm17,X_test_new,y_test)
```

```
Assumption 4: No Autocorrelation


Performing Durbin-Watson Test
Values of 1.5 < d < 2.5 generally show that there is no autocorrelation in the data
0 to 2< is positive autocorrelation
>2 to 4 is negative autocorrelation
-------------------------------------
Durbin-Watson: 2.0885340299289723
Little to no autocorrelation

Assumption satisfied
```

4.5 Normality of errors: residuals are normally distributed with mean of 0.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

5. Answer:
5.1 temp with a coefficient of 0.55
5.2 weather condition of Light Snow-Heavy rain -0.29 in bad weather bike booking count decreases.
5.3 yr with a coefficient of 0.23 as the older the app or service the more popular it becomes.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

1. Answer:

Linear Regression could be a variety of supervised Machine Learning algorithms that are used for the prediction of numeric values. Linear Regression is the simplest type of regression analysis. Regression is the most ordinarily used predictive analysis model. Linear regression is predicated on the favoured equation “y = mx + c”.

It assumes that there's a linear relationship between the dependent variable(y) and therefore the predictor(s)/independent variable(x). In regression, we calculate the most effective fit line which describes the relationship between the independent and variable.

Regression is performed when the variable quantity is of continuous data type and Predictors or independent variables might be of any data type like continuous, nominal/categorical etc.

The regression method tries to search out the simplest fit line which shows the connection between the dependent variable and predictors with least error.In regression, the output/dependent variable is that the function of an experimental variable and therefore the coefficient and also the error term.

Regression is broadly divided into simple and multiple linear regression.

1. Simple statistical regression: SLR is employed when the variable is predicted using only one variable quantity.

2. Multiple regression: The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The equation for MLR will be:

β1 = coefficient for X1 variable

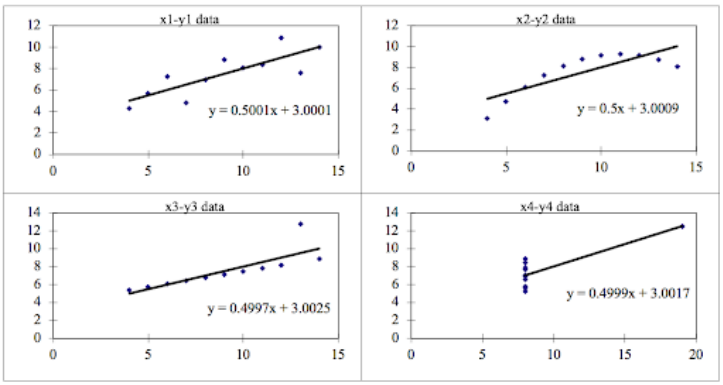
β0 is that the intercept (constant term).

2.Explain the Anscombe's quartet in detail. (3 marks)

2.Answer:

Anscombe's quartet speaks of the importance of visualizing data before applying various algorithms to make models. This means the variables- features must be plotted to work out the distribution of the samples which will facilitate you to identify the varied anomalies present within the data (outliers, diversity of the information, linear separability of the information). Moreover, the regression toward the mean can only be considered suited to the variables with linear relationships and is incapable of handling the other reasonably data set.

Anscombe's quartet takes four data sets that have nearly identical simple descriptive statistics yet to possess very different distributions and appear very different when graphed. When these models are visualized on a scatter plot, each data set generates a unique reasonably plot that doesn't meet or is described by any regression algorithm, as you'll see below



Plot 1: fits the linear regression model well.

Plot 2: cannot fit the linear regression model because the data is non-linear.

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

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

As seen, Anscombe's quartet helps us to know the importance of doing visualizations and how easy it's to fool a regression algorithm. So, before attempting to interpret and model the information or implement any machine learning algorithm, we first must visualize the information set to assist build a well-fit model.

3.What is Pearson's R? (3 marks)

3. Answer:

The Pearson correlation coefficient (r) measures the linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson R	Type of Correlation	Meaning	Example
Between 0 and 1	Positive correlation	A variable increases the other increases both change in the same direction.	Car model and price as the model is newer the price is higher
0	No correlation	Variables show no relationship	Car price and the size of the mirror
Between 0 and –1	Negative correlation	A variable increases the other decreases both change in opposite direction.	Weather situation and count of booked bikes if the weather is snowy with heavy rains the count of booked bikes drops.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

4.Answer:

Scaling is performed to prepare the data before building the linear model as the model would prefer higher values over lower values ex: temperature 20 and 30 thus misleading the coefficients and not considering the units of temperature and humidity for example thus scaling is done on numerical values with varied range to have a better model.

Scaling does not change the distribution p-values or R^2 it only changes the coefficients.

Standardization: Brings all the data into a standard normal distribution with mean =0 and std. deviation of 1 $\frac{x-mean(x)}{std.dev(x)}$ thus it deals with data following Gaussian distribution no upper range in standardization thus it does not deal with outliers.

Min- Max Scaling: Brings all the data in range of 0 to 1 $\frac{x-min(x)}{max(x)-min(x)}$ dealing with outliers and deals with data not following a Gaussian distribution useful for data's not following any type of distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

5. Answer:

VIF or variance inflation factor $VIF = \frac{1}{1-R^2}$ simply calculates how well one independent variable is explained by all other independent variables based on the videos values above 10 should be eliminated values >5 should be inspected and values<5 are okay to remain. It's used to detect multicollinearity with other independent variables thus having an infinity VIF means that variable x1 can be represented by another variable or feature or another variable is representing X1 perfectly thus having a really high correlation in this study temp and atemp showed 0.99 correlation thus leaving both in the model would lead to biased model ex 1/(1-1) is infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

6. Answer:

Q-Q Plots or Quantile-Quantile plots are used to summarize the any 2 quantiles or distributions you have visually. They determine 3 important factors:

- If both populations have the same distribution
- If the residuals are following a normal distribution which is one of the assumptions made in linear regression.
- Compare skewness of the distribution in terms of upper and lower limits and their tail behavior.

While building a linear regression model, we shall ensure that the assumptions made are met thus checking the distribution of the error terms or prediction errors while using a Q-Q plot. If the plot shows a significant deviation from the mean, we will need to check the distribution of our feature variable and work into transforming them to normal shape.

Learner Email: khalidomarasu@gmail.com
UpGrad ML& AI program
Multiple Linear Regression Subjective Questions.