Assignment-based Subjective Questions

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

Categorical variables in the dataset are:

Season, Year(yr), Month (mnth), Holiday, Weekday, Workingday and Weather Situation (weathersit).

Dependent variable is Count of total user (cnt) .

1.**Season**: Fall season seems to have attracted more booking, followed by summer. We notice a median of about 5000 during the fall season.

2. **Year**: 2019 showed more bike sharing counts as compared to 2018.

3. **Month**: Most of the bookings has been done during the month of May, June, Aug, Sep and Oct. Trend increases starting from March till October and then it started decreasing as we approached the end of year. The highest cnt is obtained in the month of September because US experiences a clear weather.

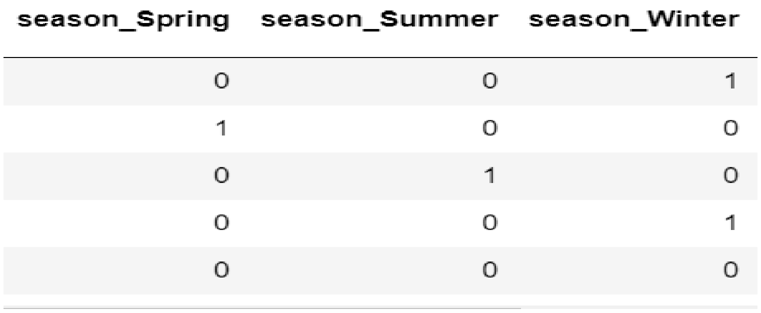
4. **Weekday**: Thursday, Friday, Saturday, and Sunday have a greater number of bookings as compared to the start of the week.

5. **Workingday**: Booking seemed to be almost equal either on working day or non-working day.

6. **Weather Situation**: Clear weather situations have high cnt values for bike sharing.

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

During dummy value creation it is advisable to use drop_first=True, because otherwise we will get a redundant feature i.e., a dummy variable might be correlated because the first column becomes a reference group during dummy encoding. For example:

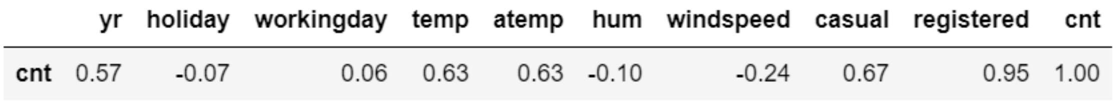


season_Spring	season_Summer	season_Winter
0	0	1
1	0	0
0	1	0
0	0	1
0	0	0

We can understand that Season status with just 2 columns like a status of 000 – will match to status type Fall, 001 will correspond to Winter and 100 to Spring.

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

The highest correlation with the target variables in the dataset was for registered variable followed by casual, atemp and temp. Registered had a correlation of 0.95 with cnt, casual had a correlation of 0.67, followed by 0.63 for temp and atemp.

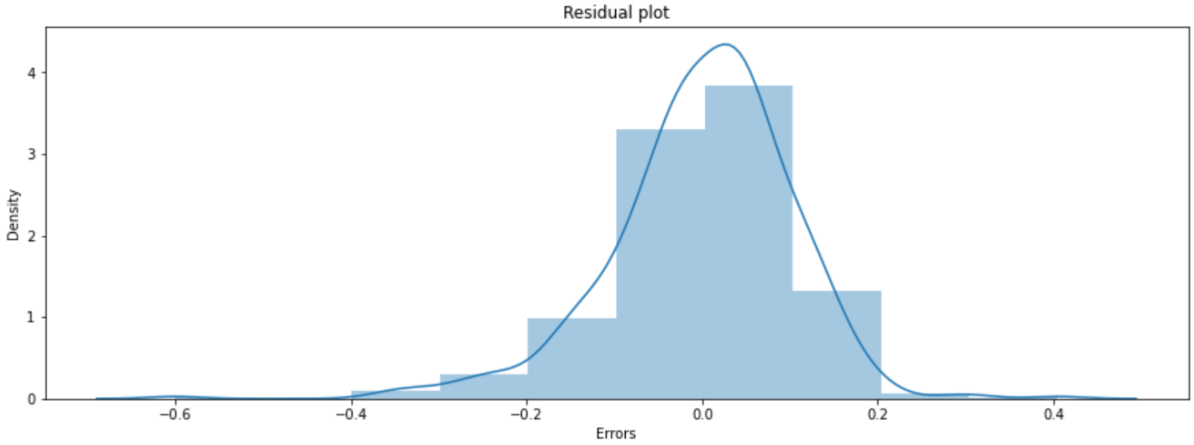


	yr	holiday	workingday	temp	atemp	hum	windspeed	casual	registered	cnt
cnt	0.57	-0.07	0.06	0.63	0.63	-0.10	-0.24	0.67	0.95	1.00

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

I have validated the assumption of Linear Regression Model based on below 5 assumptions –

1. Normality of error terms
   - Error terms should be normally distributed.



2.Multicollinearity check

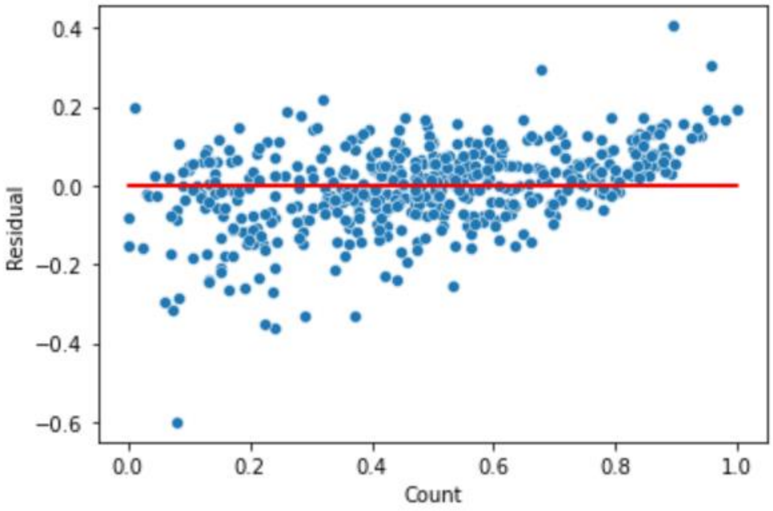
- There should be insignificant multicollinearity among variables. (VIF)

3.Linear relationship validation

- Linearity should be visible among variables.

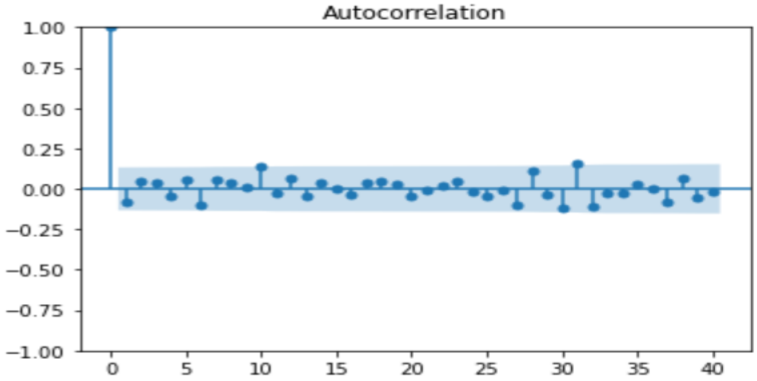
4.Homoscedasticity

- There should be no visible pattern in residual values.



5.Independence of residuals

- No autocorrelation.



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

- **year**- A coefficient value of "0.26" indicated that a unit increase in yr variable increases the bike hire numbers by "0.26" units.
- **weathersit_Light_Rain** - A coefficient value of "-0.29" indicated that, w.r.t weathersit_Light_Rain, a unit increase in weathersit_Light_Rain variable decreases the bike hire numbers by "0.29" units.
- **season Spring** - A coefficient value of "-0.22" indicated that a unit increase in season spring variable increases the bike hire numbers by "-0.22" units.

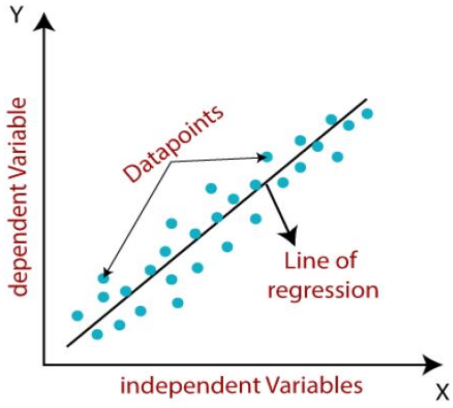
General Subjective Questions

Explain the linear regression algorithm in detail.

Linear regression is defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.
Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
Mathematically the relationship can be represented with the help of following equation −

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

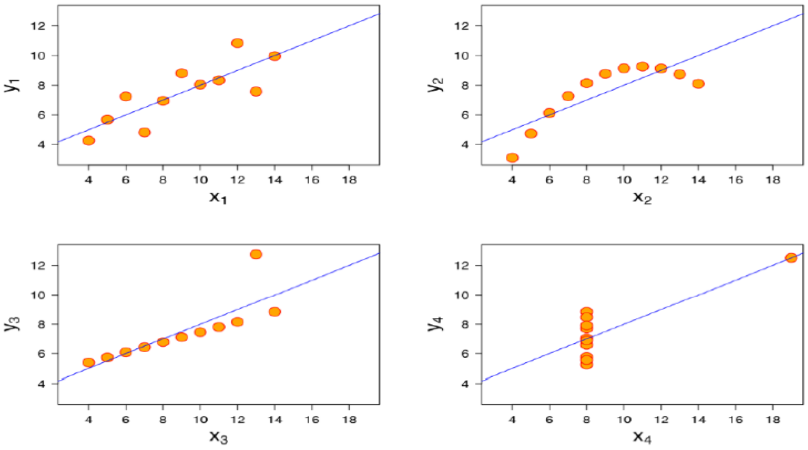


Explain the Anscombe’s quartet in detail.

Developed by statistician Francis Anscombe, Anscombe’s quartet comprises of four datasets that have nearly identical simple statistical properties yet appear very different when graphed generally using scatter plots. Each dataset consists of eleven (x, y) points.

The essential thing to note about these datasets is that they share the same descriptive statistics.

They have very different distributions and appear differently when plotted on scatter plots.



As you can see in the image above:
1. Plot1: fits the linear regression but the other 3 plots do not.
2. As in plot 2 we can see data is non-linear.
3. In plot 3 and 4 we see outliers which cannot be handled by linear regression model.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

What is Pearson's R?

Pearson's R is a numerical summary of the strength of the linear association between the variables.
Pearson's R is also referred to as Pearson's coefficient. The Pearson's correlation coefficient varies between -1 and +1 where:
r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association

In Summary, If the variables tend to go up and down together, the coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Scaling commonly known as Feature scaling is a method used to normalize the range of independent variables or features of data. When we have a lot of independent variables in a model, a lot of them might be on different scales which will lead a model with very weird coefficients that might be difficult to interpret. So, to avoid this we need to scale features and mainly scaling is done because of two reasons:

1. Ease of interpretation.
2. Faster convergence for gradient descent methods.

Normalized Scaler	**Standard Scaler**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scaling is done using MinMaxScaler	Scaling is done using StandardScaler
$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$	$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.
For Example, If the VIF is 5, this means that the variance of the model coefficient is inflated by a factor of 5 due to the presence of multicollinearity.

5 = 1/ (1-R2)

R2 = 80%

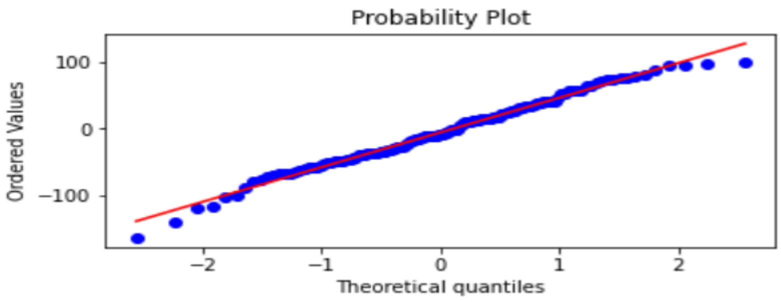
When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence

Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.


Probability Plot
Ordered Values
Theoretical quantiles