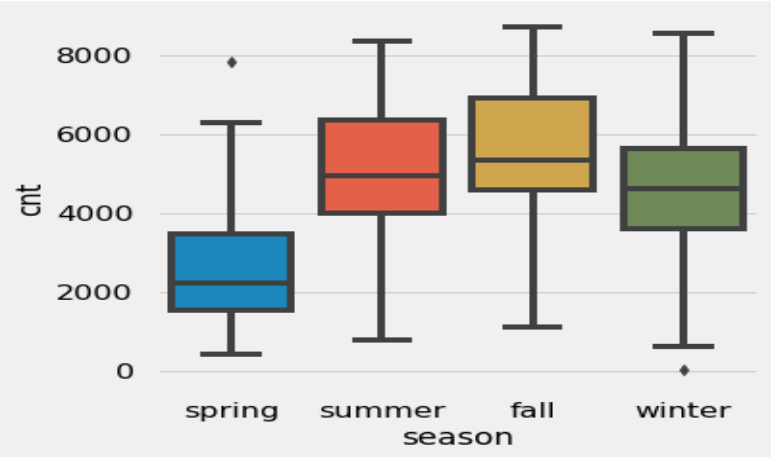
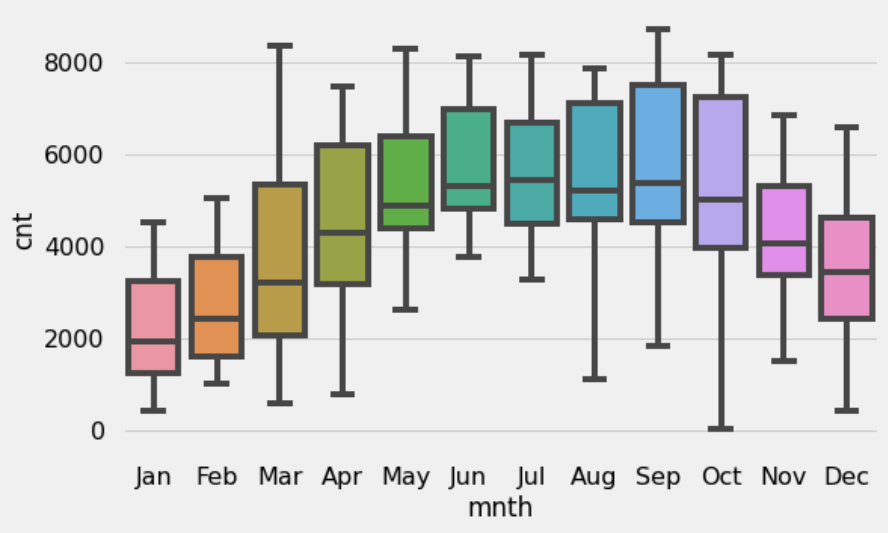
Assignment-based Subjective Questions

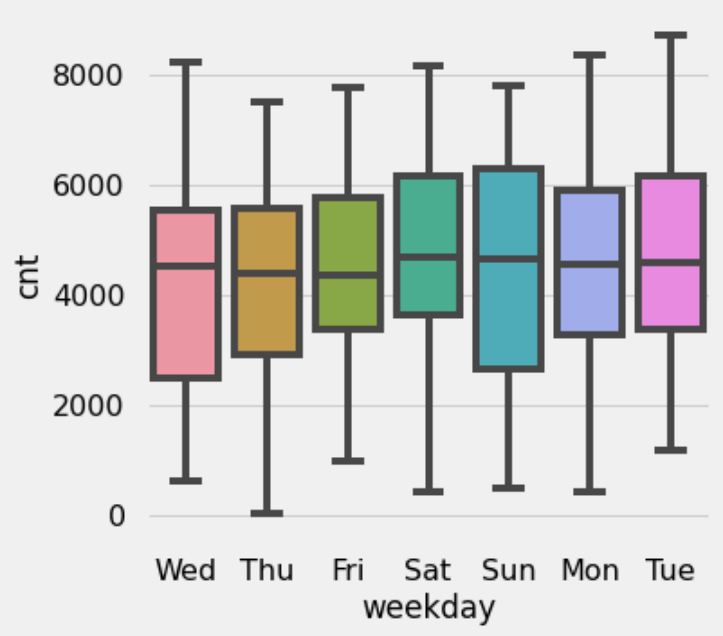
Answer

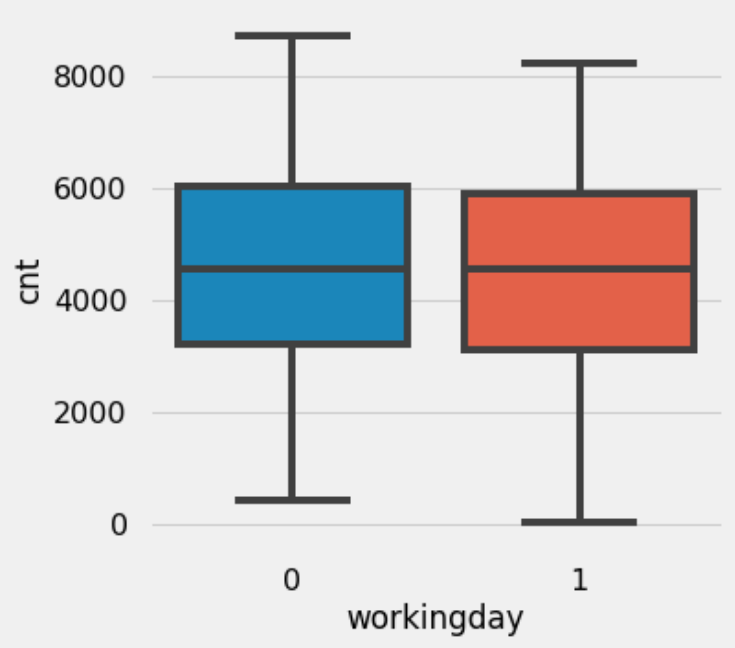
1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

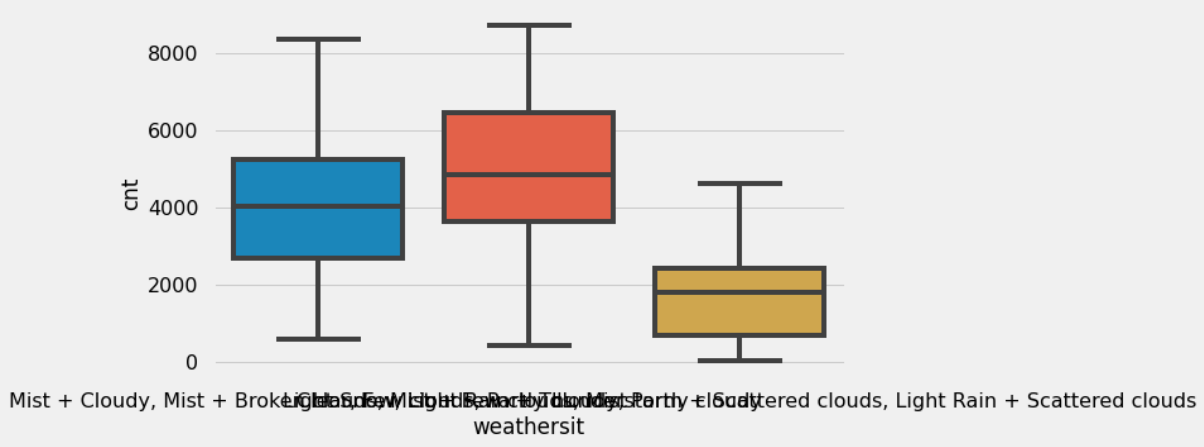
Ans:

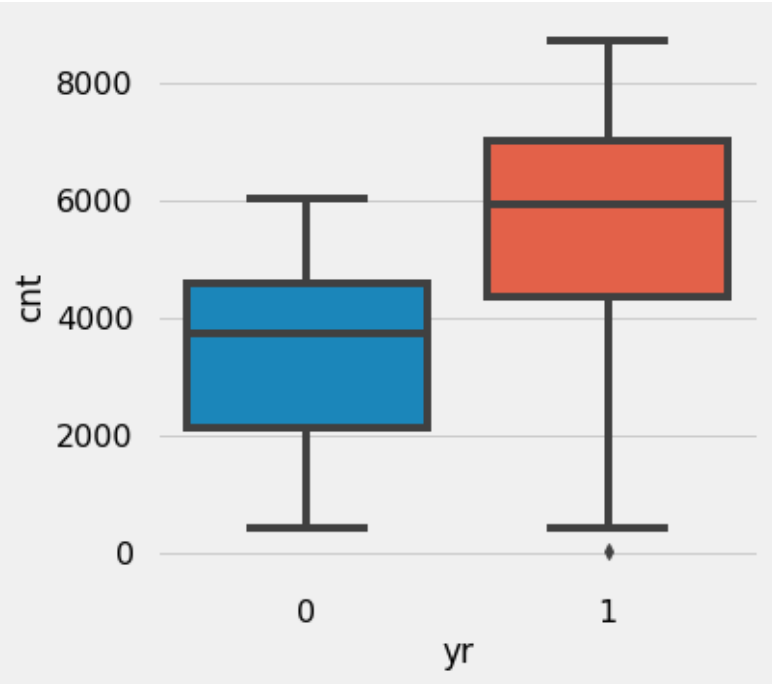





8000
6000
4000
2000
0
cnt
Wed
Thu
Fri
Sat
Sun
Mon
Tue
weekday


8000
6000
4000
2000
0
cnt
0
1
workingday


8000
6000
4000
2000
0
cnt
Mist + Cloudy, Mist + Broken clouds, Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
weathersit



From the box plot, we can easily conclude that:

1. Bike sharing count is more in season of summer and fall compared to winter and spring.

2. Bike sharing count is more in month of June to October than other months.

3. Bike sharing count is more when weathersit is Clear, Few clouds, partly cloudy, Partly cloudy.

4. Bike sharing count is more in year 2019

5. Bike Sharing count have no major impact of working day and weekday.

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

Ans: Using drop_first=True is more common in statistics and often referred to as "dummy encoding". When we convert the categorical variables to dummies, indirectly we are giving importance to each value in a categorical column by making each value as a column. If we don't drop the first column then your dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may have trouble converging and lists of variable importance may be distorted. For e.g. if we have a categorical column for gender for Male and Female, then after creating dummy variables, we can easily drop any one of column, as the other columns zero values will represents the deleted column values.

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

Ans: : From pair plots, we can conclude that, Registered and Casual were highly correlated with cnt, that shows that both casual and registered are similar to cnt, hence they needs to be ignored in analysis as they directly relates to cnt.

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

Ans: Assumptions of Linear Regression. There are 5 basic assumptions of Linear Regression Algorithm:

1.Linear Relationship between the features and target: According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.

2. **Little or no Multicollinearity between the features**: Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model. Pair plots and heatmaps (correlation matrix) can be used for identifying highly correlated features.

3.**Homoscedasticity Assumption**: Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a goodway to check for homoscedasticity.

4. **Normal distribution of error terms**: The fourth assumption is that the error(residuals) follows a normal distribution.

5.**Little or No autocorrelation in the residuals**: Autocorrelation can be tested with the help of Durbin-Watson test. From the above summary note that the value of Dur bin-Watson test is 1.908 quite close to 2 as said before when the value of Durbin-Watson is equal to 2, r takes the value 0 from the equation 2*(1-r), which in turn tells us that the residuals are not correlated.

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.788
Model:	OLS	**Adj. R-squared:**	0.784
Method:	Least Squares	**F-statistic:**	186.0
Date:	Mon, 30 Nov 2020	**Prob (F-statistic):**	3.29e-161
Time:	16:30:11	**Log-Likelihood:**	434.57
No. Observations:	510	**AIC:**	-847.1
Df Residuals:	499	**BIC:**	-800.6
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5838	0.013	44.794	0.000	0.558	0.609
yr	0.2459	0.009	26.496	0.000	0.228	0.264
windspeed	-0.1920	0.029	-6.693	0.000	-0.248	-0.136
Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds	-0.3130	0.028	-11.232	0.000	-0.368	-0.258
Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist	-0.0870	0.010	-8.817	0.000	-0.106	-0.068

spring	-0.2398	0.015	-16.381	0.000	-0.269	-0.211
summer	-0.0395	0.013	-3.087	0.002	-0.065	-0.014
Dec	-0.1178	0.018	-6.728	0.000	-0.152	-0.083
Jan	-0.1224	0.020	-6.151	0.000	-0.161	-0.083
Nov	-0.1181	0.018	-6.599	0.000	-0.153	-0.083
Sep	0.0562	0.018	3.054	0.002	0.020	0.092

Omnibus:	68.166	**Durbin-Watson:**	1.908
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	175.873
Skew:	-0.675	**Prob(JB):**	6.45e-39
Kurtosis:	5.541	**Cond. No.**	8.76

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

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

Ans: $ cnt =0.584 + 0.245 \times yr - 0.192 \times windspeed - 0.313 \times Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds - 0.087 \times Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist - 0.239 \times spring - 0.039 \times summer - 0.117 \times Dec - 0.122 \times Jan - 0.118 \times Nov + 0.056 \times Sep .

Standardized coefficients signify the mean change of the dependent variable given a one standard deviation shift in an independent variable. thus, coefficient of predictors can be compared to assess the importance. Thus, based on high coefficient values below factor have highest influence:

1. yr
2. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3. spring