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?**

Ans: 1>As by seeing the boxplot in the jupyter notebook , we can infer that in Summer and Fall count increases in both years (2018 as well as 2019). Count has been increased from 2018 to 2019.

2>Summer and Fall count is high and we can also conclude that count has been pretty high in Friday and Sat.

3> count has been increased from 2018 to 2019. Count has also been increased in Summer and Fall. Holiday or working day doesn'tmatter much

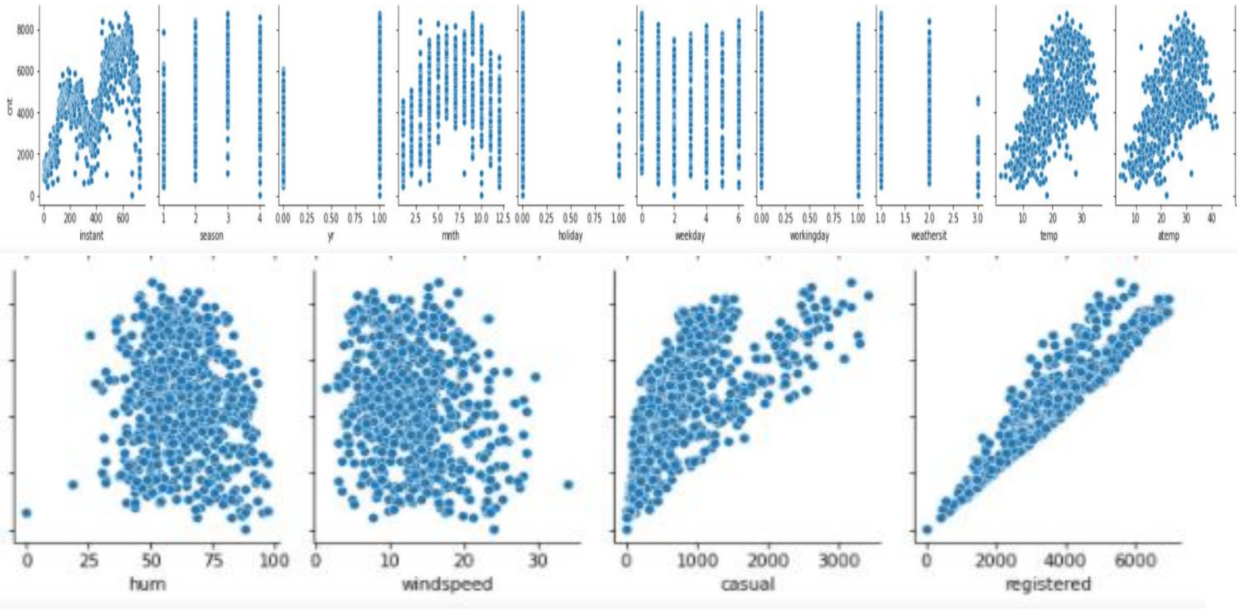
2. **Why is it important to use drop_first=True during dummy variable creation?**

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

This may affect some modelsadversely and the effect is stronger when the cardinality is smaller.

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

Ans:



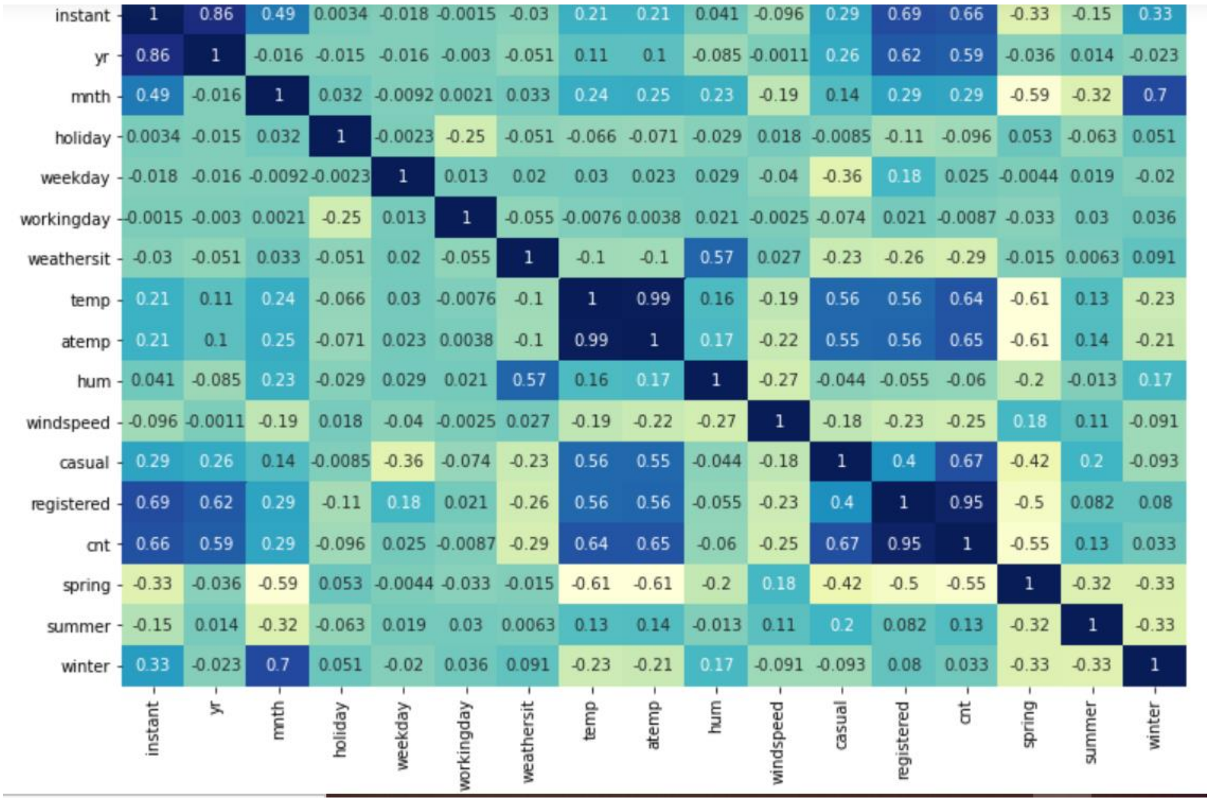
From the pairplot given above we can say that atemp ,temp , registered variables has good correlation with target variable.

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

Ans: There are certain assumptions of Linear Regression after building the model on the training set which are as follows:

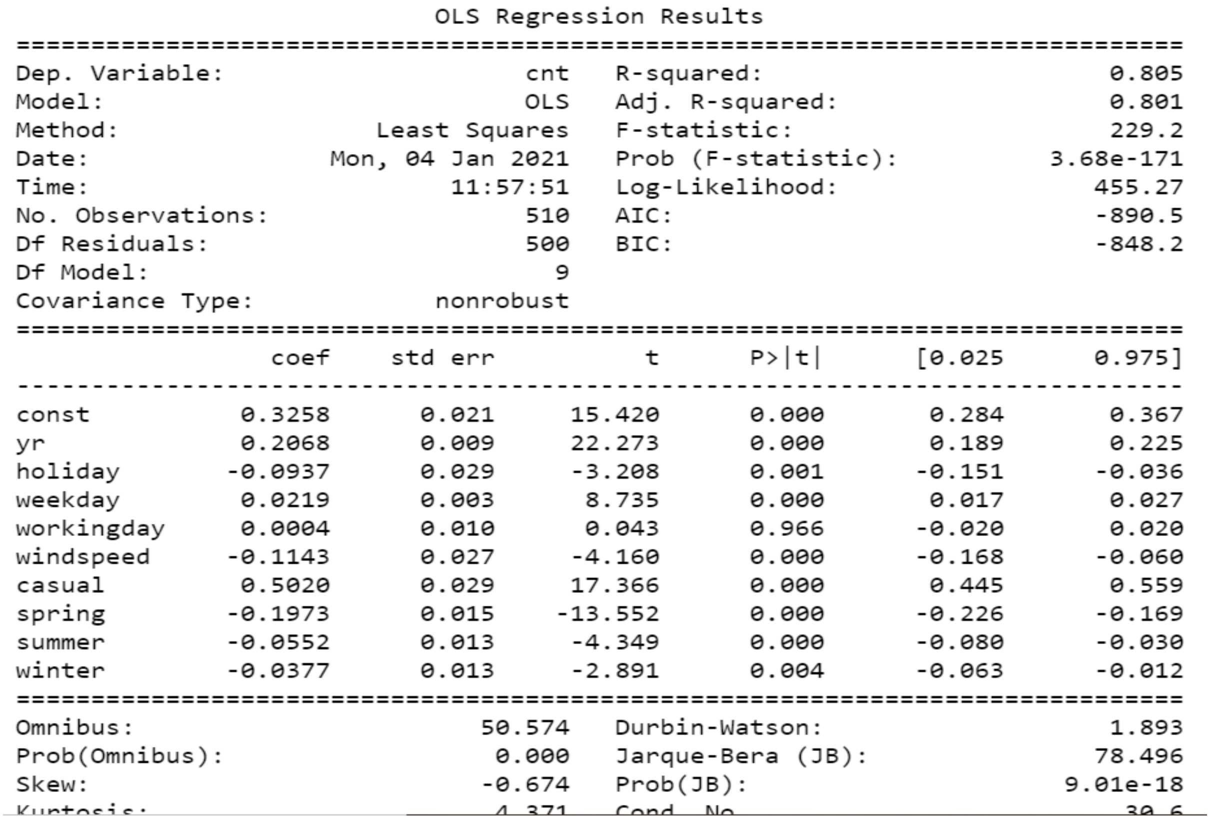
1. Linear Relationship: Linear regression assumes that there exists a linear relationship between the dependent variable and the predictors. A partial residual plot that represents the relationship between a predictor and the dependent variable while taking into account all the other variables may help visualize the “true nature of the relationship” between variables.
2. Homoscedasticity: It means that the residuals have constant variance no matter the level of the dependent variable. To verify homoscedasticity, one may look at the residual plot and verify that the variance of the error terms is constant across the values of the dependent variable.

3. Absence of Multicollinearity: Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case). While it may not be important for non-parametric methods, it is primordial for parametric models such as linear regression. Often, a tell-tale sign of multicollinearity is the fact that some of the estimated coefficients have the “wrong” sign (i.e. the coefficient related to the size of a being negative in a model attempting to predict house prices). Pairwise correlations could be the first step to identify potential relationships between various independent variables.



	instant	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt	spring	summer	winter
instant	1	0.86	0.49	0.0034	-0.018	-0.0015	-0.03	0.21	0.21	0.041	-0.096	0.29	0.69	0.66	-0.33	-0.15	0.33
yr	0.86	1	-0.016	-0.015	-0.016	-0.003	-0.051	0.11	0.1	-0.085	-0.0011	0.26	0.62	0.59	-0.036	0.014	-0.023
mnth	0.49	-0.016	1	0.032	-0.0092	0.0021	0.033	0.24	0.25	0.23	-0.19	0.14	0.29	0.29	-0.59	-0.32	0.7
holiday	0.0034	-0.015	0.032	1	-0.0023	-0.25	-0.051	-0.066	-0.071	-0.029	0.018	-0.0085	-0.11	-0.096	0.053	-0.063	0.051
weekday	-0.018	-0.016	-0.0092	-0.0023	1	0.013	0.02	0.03	0.023	0.029	-0.04	-0.36	0.18	0.025	-0.0044	0.019	-0.02
workingday	-0.0015	-0.003	0.0021	-0.25	0.013	1	-0.055	-0.0076	0.0038	0.021	-0.0025	-0.074	0.021	-0.0087	-0.033	0.03	0.036
weathersit	-0.03	-0.051	0.033	-0.051	0.02	-0.055	1	-0.1	-0.1	0.57	0.027	-0.23	-0.26	-0.29	-0.015	0.0063	0.091
temp	0.21	0.11	0.24	-0.066	0.03	-0.0076	-0.1	1	0.99	0.16	-0.19	0.56	0.56	0.64	-0.61	0.13	-0.23
atemp	0.21	0.1	0.25	-0.071	0.023	0.0038	-0.1	0.99	1	0.17	-0.22	0.55	0.56	0.65	-0.61	0.14	-0.21
hum	0.041	-0.085	0.23	-0.029	0.029	0.021	0.57	0.16	0.17	1	-0.27	-0.044	-0.055	-0.06	-0.2	-0.013	0.17
windspeed	-0.096	-0.0011	-0.19	0.018	-0.04	-0.0025	0.027	-0.19	-0.22	-0.27	1	-0.18	-0.23	-0.25	0.18	0.11	-0.091
casual	0.29	0.26	0.14	-0.0085	-0.36	-0.074	-0.23	0.56	0.55	-0.044	-0.18	1	0.4	0.67	-0.42	0.2	-0.093
registered	0.69	0.62	0.29	-0.11	0.18	0.021	-0.26	0.56	0.56	-0.055	-0.23	0.4	1	0.95	-0.5	0.082	0.08
cnt	0.66	0.59	0.29	-0.096	0.025	-0.0087	-0.29	0.64	0.65	-0.06	-0.25	0.67	0.95	1	-0.55	0.13	0.033
spring	-0.33	-0.036	-0.59	0.053	-0.0044	-0.033	-0.015	-0.61	-0.61	-0.2	0.18	-0.42	-0.5	-0.55	1	-0.32	-0.33
summer	-0.15	0.014	-0.32	-0.063	0.019	0.03	0.0063	0.13	0.14	-0.013	0.11	0.2	0.082	0.13	-0.32	1	-0.33
winter	0.33	-0.023	0.7	0.051	-0.02	0.036	0.091	-0.23	-0.21	0.17	-0.091	-0.093	0.08	0.033	-0.33	-0.33	1

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



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.801
Method:                 Least Squares   F-statistic:                     229.2
Date:                Mon, 04 Jan 2021   Prob (F-statistic):          3.68e-171
Time:                        11:57:51   Log-Likelihood:                 455.27
No. Observations:                 510   AIC:                            -890.5
Df Residuals:                     500   BIC:                            -848.2
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3258      0.021     15.420      0.000       0.284       0.367
yr             0.2068      0.009     22.273      0.000       0.189       0.225
holiday       -0.0937      0.029     -3.208      0.001      -0.151      -0.036
weekday        0.0219      0.003      8.735      0.000       0.017       0.027
workingday     0.0004      0.010      0.043      0.966      -0.020       0.020
windspeed     -0.1143      0.027     -4.160      0.000      -0.168      -0.060
casual         0.5020      0.029     17.366      0.000       0.445       0.559
spring        -0.1973      0.015    -13.552      0.000      -0.226      -0.169
summer        -0.0552      0.013     -4.349      0.000      -0.080      -0.030
winter        -0.0377      0.013     -2.891      0.004      -0.063      -0.012
==============================================================================
Omnibus:                       50.574   Durbin-Watson:                   1.893
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               78.496
Skew:                          -0.674   Prob(JB):                     9.01e-18
Kurtosis:                       4.371   Cond. No.                         30.6
```

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans: The reason is because linear regression has been around for so long (more than 200 years). It has been studied from every possible angle and often each angle has a new and different name.

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.
Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called **Ordinary Least Squares**. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.
Now that we know some names used to describe linear regression, let's take a closer look at the representation used.

Linear Regression Model Representation

Linear regression is an attractive model because the representation is so simple.
The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

Now that we understand the representation used for a linear regression model, let’s review some ways that we can learn this representation from data.

2. **Explain the Anscombe’s quartet in detail.**
   Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough

3. **What is Pearson’s R?**
   Ans: The correlation between two variables reflects the degree to which the variables are related. The most common measure of correlation is the Pearson Product Moment Correlation (called Pearson's correlation for short). When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho (ρ). When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between variables.
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Ans: Scaling is a personal choice about making the numbers feel right, e.g. between zero and one, or one and a hundred. For example converting data given in millimeters to meters because it's more convenient, or imperial to metric.

While normalisation is about scaling to an external 'standard' - the local norm - such as removing the mean value and dividing by the sample standard deviation, e.g. so that your sorted data can be compared with a cummulative normal, or a cummulative Poisson, or whatever.

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

   Ans: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

   Ans: *Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.*