**Assignment-based Subjective Questions**

**Q1.** **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

**Ans**. a. People prefers Bike sharing less in Holydays, and they share bike more in working days. So, we can say that they prefer bike sharing for travelling to offices or their work.

b. Most of the people want to share the bike in clear weather condition compared to cloudy and raining weather.

c. If we talk about “**Season vs Count**” then we can see that in the season of ‘Fall’ there is a greater number of sharing happened and in ‘spring’ season it is less.

d. For the **“Year vs Count”** say that business is getting improved in the 2019 compared to 2018. It’s means that business was getting popular year by year.

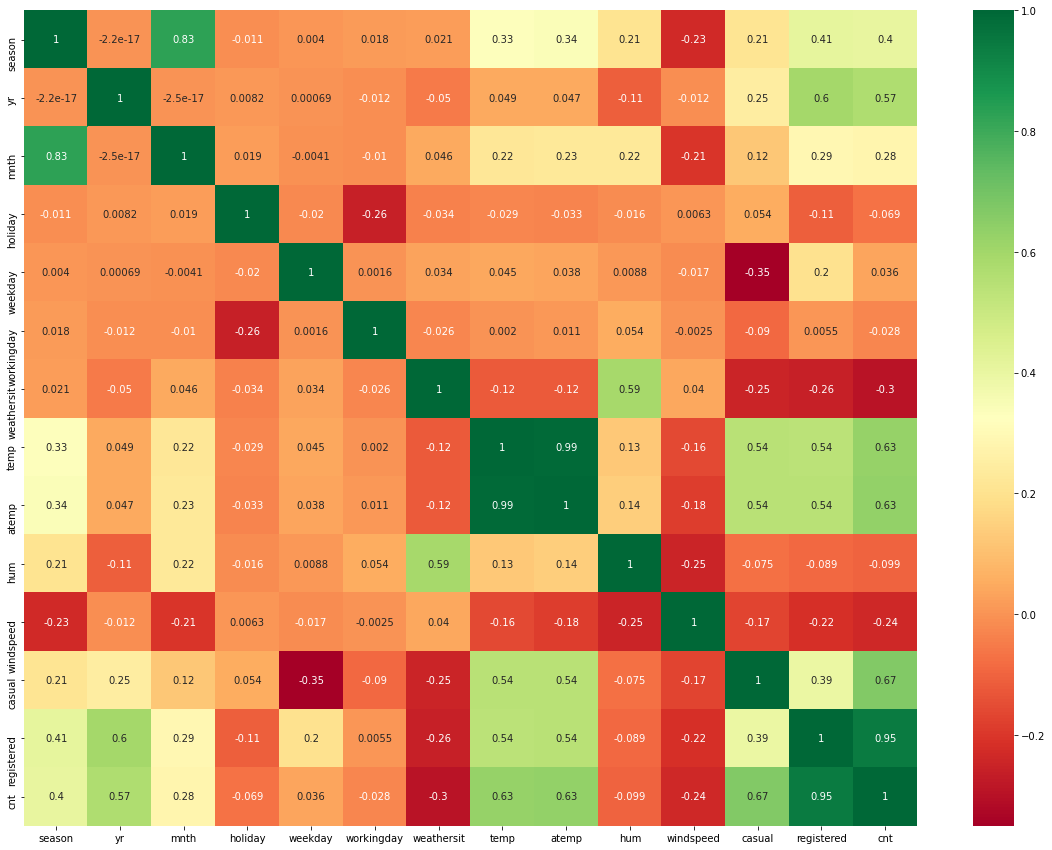
e. **Temperature** and **Ambient Temperature** are linearly correlated with **Count**. That means if temperature increases the number of counts also increase.

**Q2. Why is it important to use drop first =True during dummy variable creation? (2 mark)**

**Ans.** Use of **drop\_first = True** helps in reducing the number of columns in the data set as it drops the 1st column created after use of **dummy**.

**Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

**Ans.**

****

From above heatmap, firstly, it’s obvious that **registered** and **casual** will be highly correlated with the **cnt**. After these **atemp**, **temp** and **season** columns are correlated with respectively.

**Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

**Ans. 1. Linear Relationship –** the regression plot between different predictors and target variable shows the linear relationships among them.

**2. Homoscedasticity -** Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable. To verify homoscedasticity, I plotted the Residual plot which shows 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. To verify this, look at the Variance Inflation Factors (VIF) and R- squared values. Large R-squared for the regression and a larger VIF indicators of multicollinearity. Which is not in my case.

**4. Normality of Errors -** To verify this, I plotted the QQ-plot. Which shows normal distribution which indicates unbiasedness of the residuals.

**5. Independence of residuals (absence of auto-correlation) –** I used **Durbin-Watson test** to verify this.

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

**Ans. yr**, **holiday** and **temp** are my top 3 features contributing significantly towards explaining the demand of the shared bikes.

**General Subjective Questions**

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

**Ans. a.** Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

**b.** Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

**c.** The linear regression model provides a sloped straight line representing the relationship between the variables.

**d.** Linear regression can be further divided into two types of the algorithm:

1. **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

2. **Multiple Linear regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

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

* Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 analysing it and the effect of outliers 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." It has been rendered as an actual musical quartet.

**3. What is Pearson’s R? (3 marks)**

* Correlation coefficients are used to measure how strong a relationship is between two variables.
* There are several types of correlation coefficient, but the most popular is Pearson’s.
* Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression.
* Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

1. 1 indicates a strong positive relationship.
2. -1 indicates a strong negative relationship.
3. A result of zero indicates no relationship at all.

* There are several types of correlation coefficient formulas.

1. One of the most commonly used formulas is Pearson’s correlation coefficient formula.
2. Two other formulas are commonly used: the sample correlation coefficient and the population correlation coefficient.

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

* Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules. In other words, the process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned is called as scaling.
* Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.
* The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

**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. (3 marks)**

* In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
* If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.
* A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.