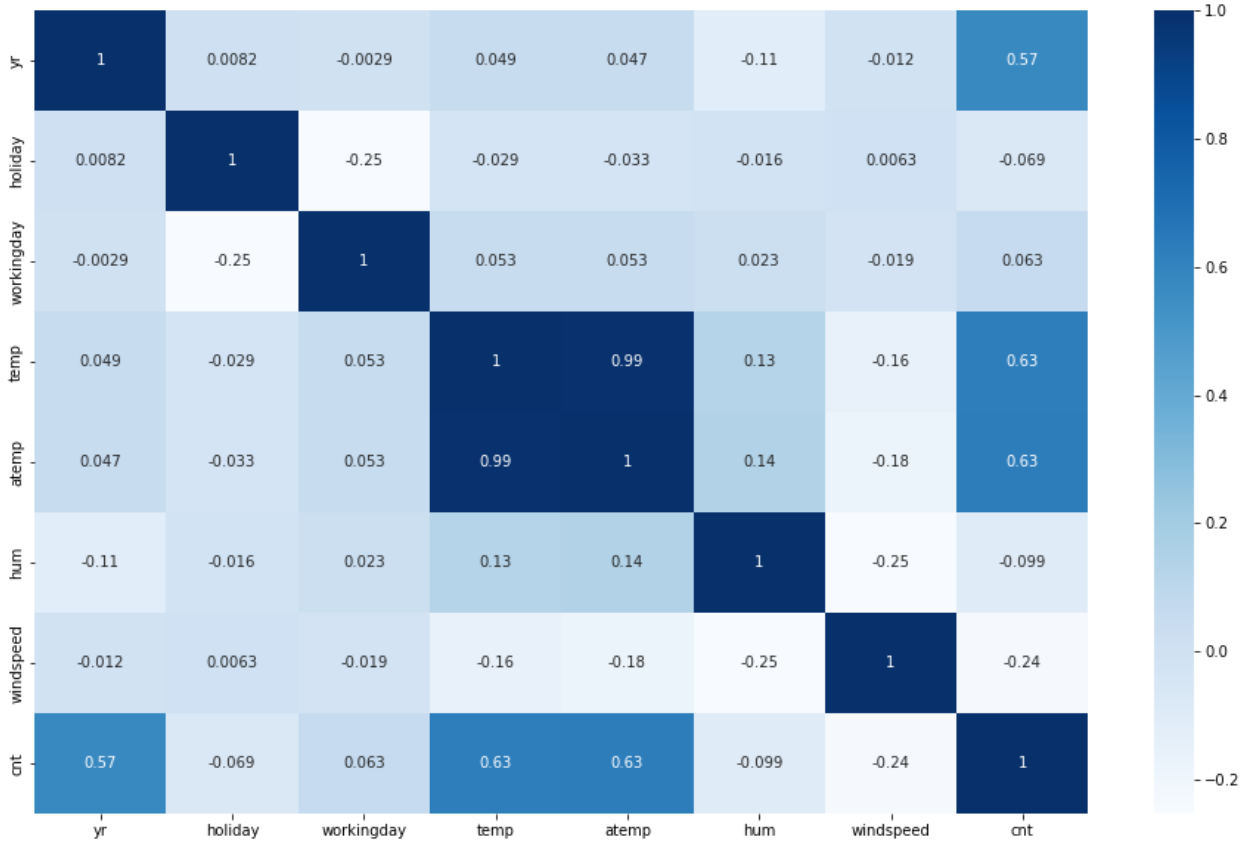
1. 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
   - Categorical variables like holiday, workingday and year are not highly correlated with other Dependent variables as can be seen in the correlation heatmap below:



2. 2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   - Drop_first= True is important during dummy variable creation as it eliminates the unnecessary variable from the dataset. Usually dropped value can be represented by other variables, as general rule P values can be represented by P-1 variables.

3. 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - Temp/Atemp has highest correlation with target variable, it almost follows straight line. This can be seen in heatmap shown in answer 1.

4. 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - After building the model, we checked the error codes is thy are normally distributed or not.

Additionally we check for using scatter plot that residuals are randomly distributed and not following any pattern.

5. 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   - Top three features are:
     1. Temperature
     2. Weather is moderate or not
     3. September month has highest demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   - Linear regression algorithm is approach for modelling relationship between a target variable(dependent variable) and one or more independent variables. In-case there is one independent variable it is called simple linear regression and where there are more than one independent variables, it is called multilinear regression.

     Linear regression is one of the most common algorithm used in machine learning for predicting the values of target variables using dependent variables if linear relationship can be established between the variables. It is represented by equation:

     $y = B0 + B1*X$

2. Explain the Anscombe's quartet in detail. (3 marks)
   - Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

3. What is Pearson's R? (3 marks)
   - Pearson's correlation coefficient or Pearson's R is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

     The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   - Scaling is data pre-processing step which is used to normalize the range of independent variables. It is performed as most of the time data collected has features/dependent variables in varying units/magnitudes and if scaling is not then algorithm taken magnitude in accounts. This results into incorrect modelling. Scaling helps to being all variables in same level of magnitude.
   - Normalization is used to transform features to be on similar scale and this scale ranges between -1 and 1. Normalization is useful when there are no outliers in data as it cannot cope with them.

     Standardization is transformation of features by subtracting mean and diving by standard deviation. This is also called as Z-score. Standardization does not get affected by outliers because there is no predefined range of transformed features.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   - VIF is infinite when there is perfect correlation between two independent variables. To solve this problem, we need to drop variable which is causing perfect collinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   - Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. In linear regression it is used to check if distribution of residuals are normally distributed.