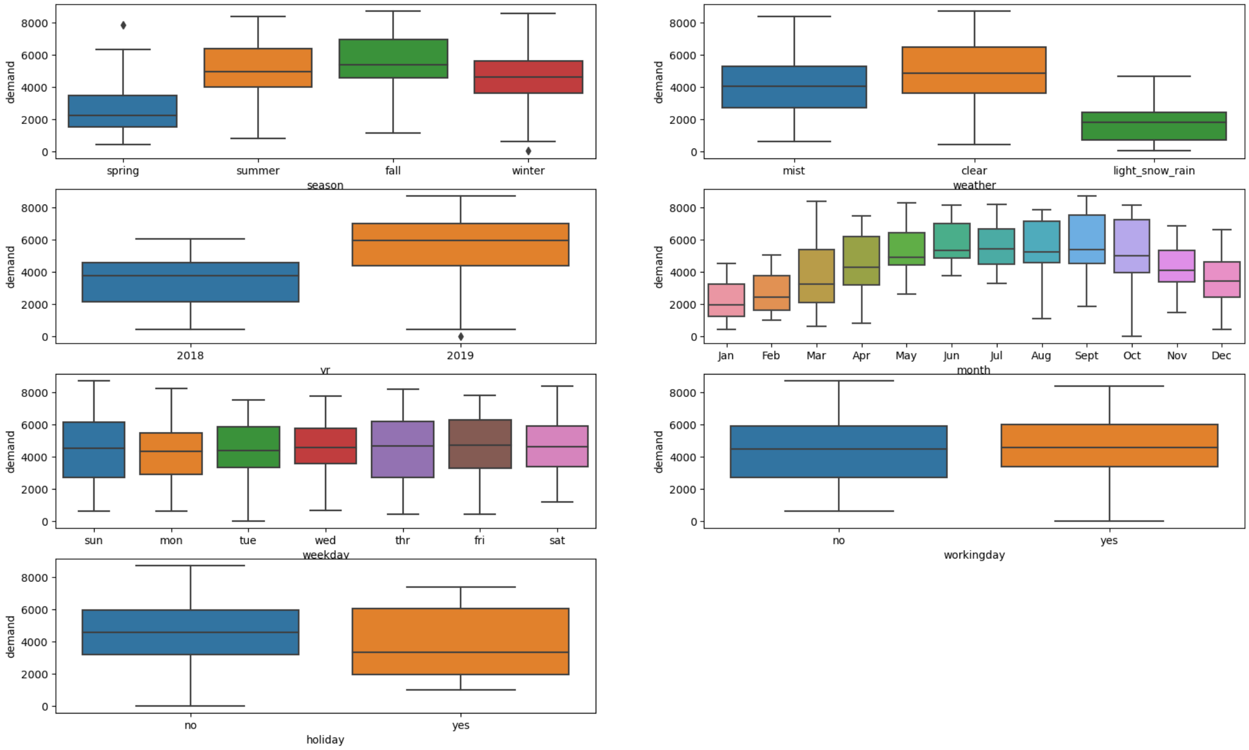
Assignment-based Subjective Questions

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

Answer: I have considered the following variables as categorical:

- Season
- Weather
- Yr
- Month
- Weekday
- Working Day
- Holiday

Let’s look at a box plot of that.



Here are some of the observations:

- Year 2019 has a significant growth
- People seems to be using the service most in the following order:
  - Fall
  - Summer
  - Winter
  - Spring

- Weather also has an effect on the demand for the service according to the plot, people prefer to use the service more when the weather is clear or mist.
- Some Months of the year shows more demand than the others.
- Non Holiday days have a higher mean for the service compared to holidays, seems like many people use the service for their office commute.

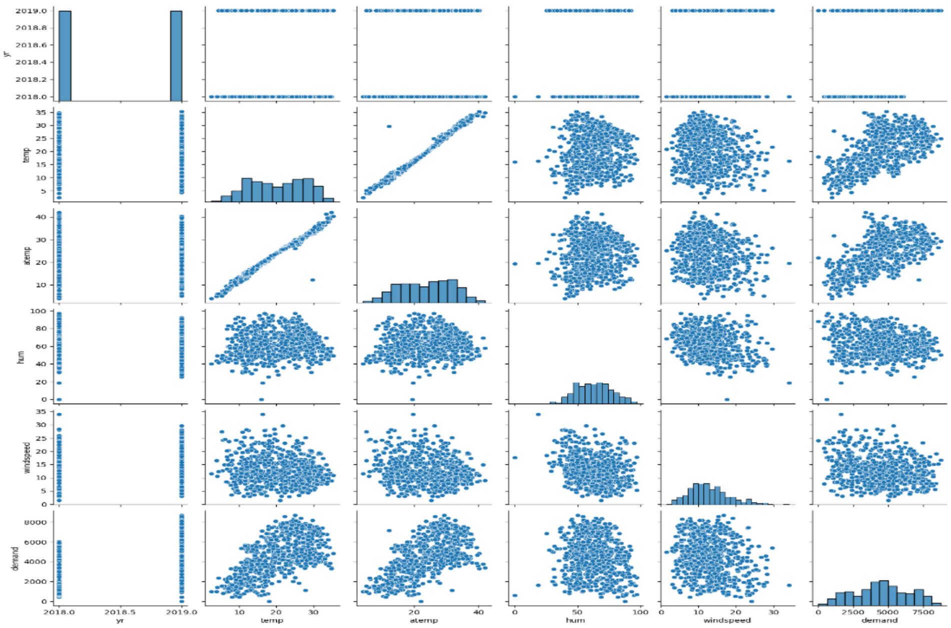
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

ANSWER: it is important to use drop_first=True since we need k -1 dummy variables for a categorical variable with k level. If it is not set to True it will create k dummies for l level variable.

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

Answer:

The Plot:



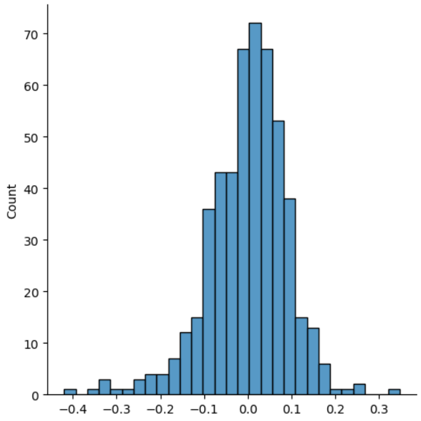
temp, atemp

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

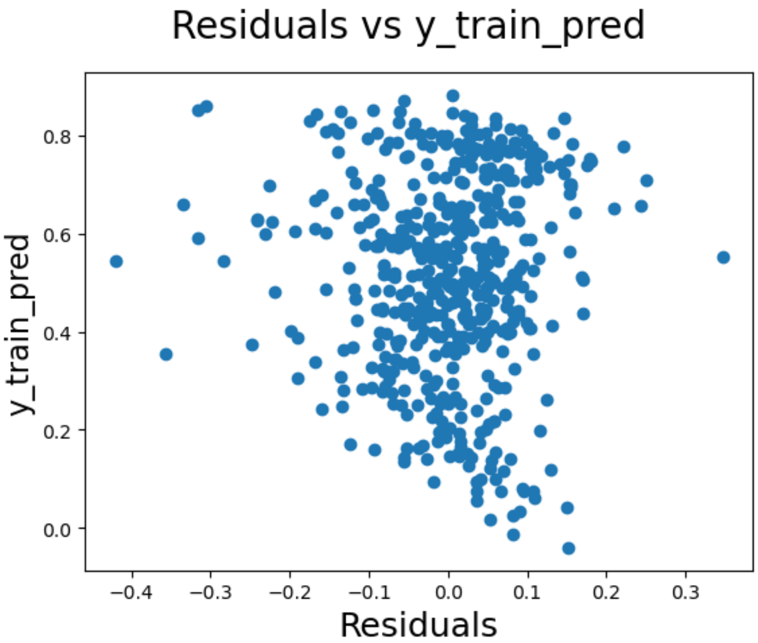
ANSWER: I looked at the following things

- The Pair Plot showed me that the target variable has a linear relationship with some of the independent variables like temp and a temp

- Residual Displot showed that the error terms are normally distributed and close to 0.



- Checked for the residual spread for any visible patterns. There were none and they seemed to be centered at 0.



- VIF of all features selected are < 5 so there is no interdependency among the features variables.

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

ANSWER:

- Temp – 0.458296
- Yr – 0.234864

- Month of sept – 0.090036

General Subjective Questions

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

Answer:
Linear Regression is a supervised learning method where we must have some past data to create a Regression line that follows the equation of a straight line:

Y = mx + c + e
In case of a single Linear Regression and in case of multiple regression, the equation becomes:
Y = M1X1 + M2X2 +…MnXn + c + e
where M1, M2… Mn are coefficients and X1,X2, X3.. Xn are independent variables,
c is a constant and e is the error terms.

Here are the basic assumptions of Linear Regression:

1. There must be a linear relationship between the independent and the dependent variable.
2. Error Terms are normally distributed.
3. There is no pattern in the residuals.
4. The feature variables are independent of each other or have No or Little Multicollinearity.
5. Homoscedasticity

First using the above equation, we form a Regression Line.
After which we try to find the BEST fit line using the gradient decent cost function which in the case of Linear regression is the Least Residual Sum of squared.

Once the Best fit line is found, we find the Coffs of each variable in the equation.

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

Answer: 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.

It explains the importance of Data Visualization.

3. What is Pearson's R? (3 marks)

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation 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.

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

Answer: Scaling is a method to bring all our numerical values to a similar range of scale.
We need to scale our data so that the cost function can run efficiently and ease of interpretation.

There are two types of Scaling

- Normalized (min-max Scaling)
- Standardization Scaling

Let's look at the difference

Normalized	**Standardization**
Formula – x = x- xmin / max(x) – min(x)	Formula – x = x – mean(x) /sd (x)
All values are wrapped between 0 and 1	Values will centralized towards mean 0

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

Answer: VIF might go to infinity if there is a perfect correlation between the variable and the others.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: Q-Q(quantile-quantile) plots play a very vital role in graphically analyze and comparing two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight line y = x.

In terms of Linear Regression, it can help us identify which variables beta1 or coff is 0.