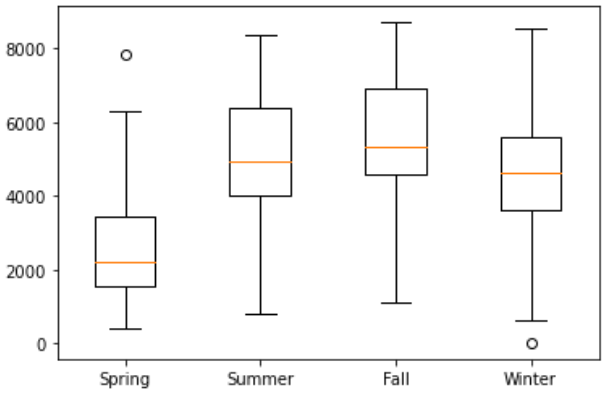
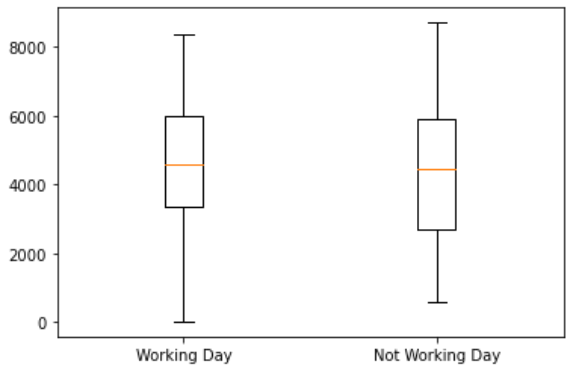
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   - **Season**: The demand changes with season. The demand rises during summer and fall, and then decreases during winter and spring. The demand is highest during fall and lowest during spring.

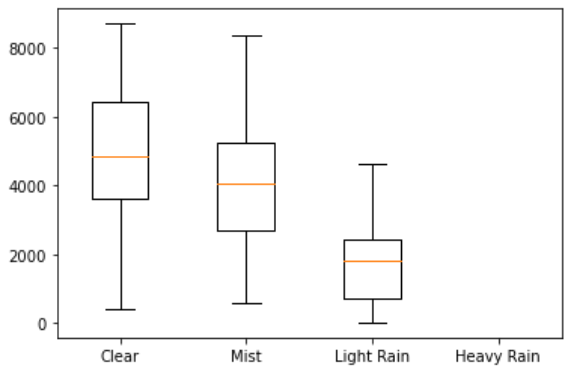


   - **Working Day**: The demand does not seem to depend on the working day. Whether it's a working day or not doesn't have a great impact on the median demand. The spread during non-working days is greater than on working days with the lower hinge being lower than the working days and upper hinge being almost the same.

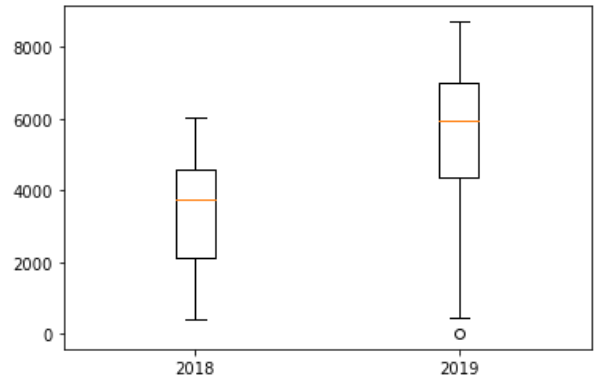
     Since working day is closely related to holidays and days of the week, working day is considered instead of holidays and days of the week.



   - **Weather:** The demand decreases when the weather is misty or rainy. The dataset had no data points during heavy rain. So, predictions during heavy rain cannot be made.



- **Year**: The demand increases from 2018 to 2019 with lower hinge, median and upper hinge for 2019 being higher than those of 2018. This suggests that the demand is increasing year on year.

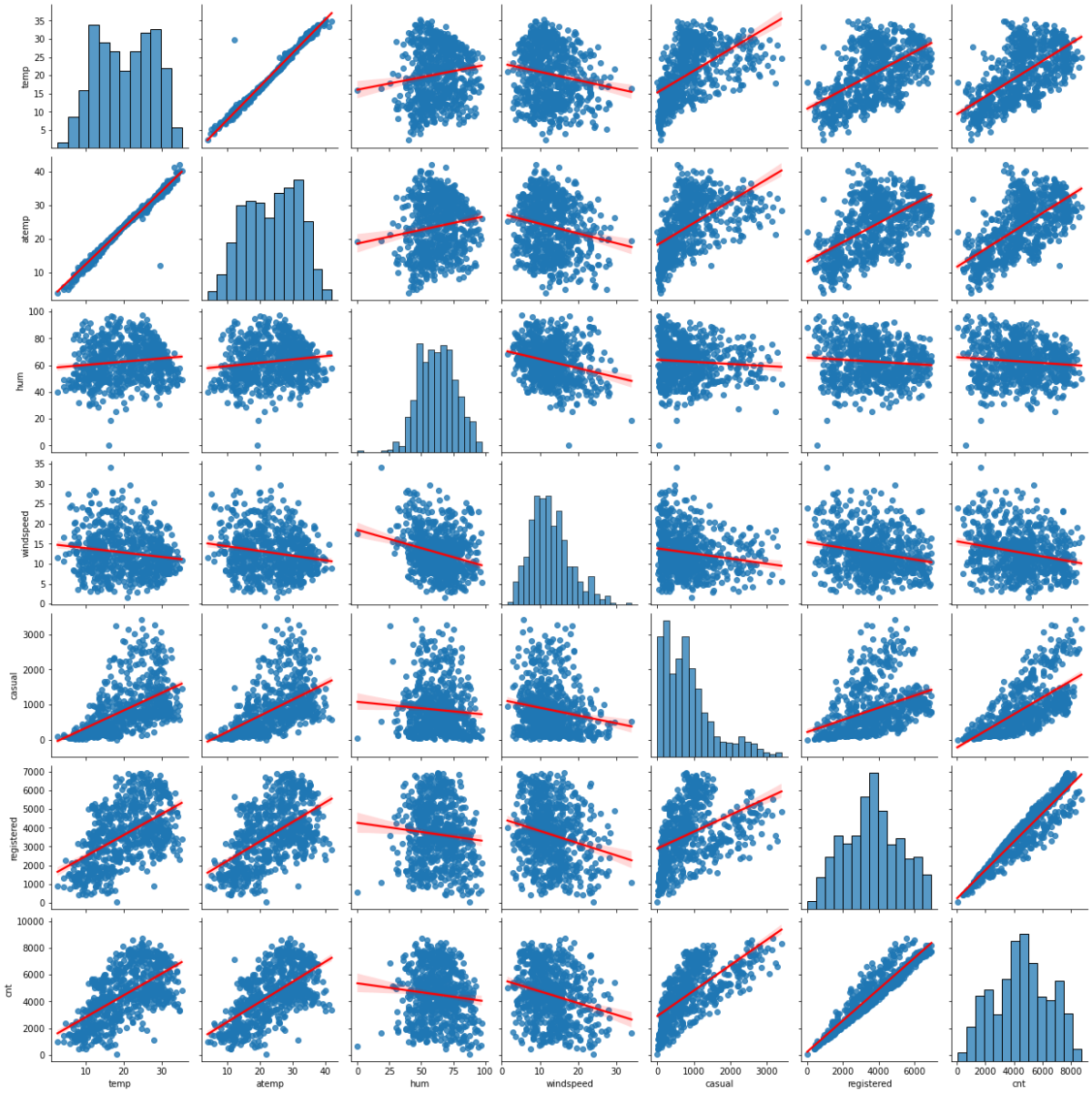


2. **Why is it important to use drop_first=True during dummy variable creation?**

   - The pd.get_dummies() function returns a dummy variable for each of the levels of the categorical variable. But a categorical variable with k levels can be expressed using k – 1 dummy variables, and the lower the number of predictors in the model, the simpler and easy to interpret the model is. The state where all the dummy variables are zero can be used as the base state. Upon using drop_first = True, the function returns only k – 1 dummy variables instead of k.

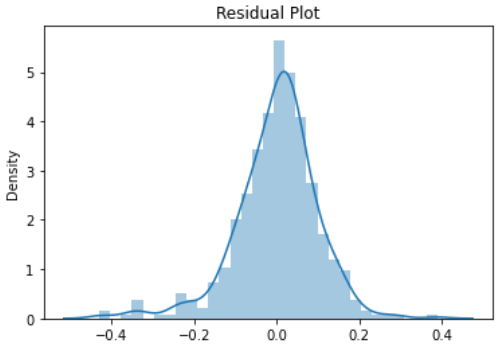
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   - According to the pairplot, registered and casual seems to have the highest correlation with the target variable. But, these cannot be used as a predictor since we do not have the number of registered users in the dataset. And these variables make up the target variable. The variable with next highest correlation is temp and atemp – the demand increases as temp and atemp increase. This is understandable and in line with the earlier observation where we see that the demand is highest during the summers and lowest during the winter. The pairplot is shown on the next page.

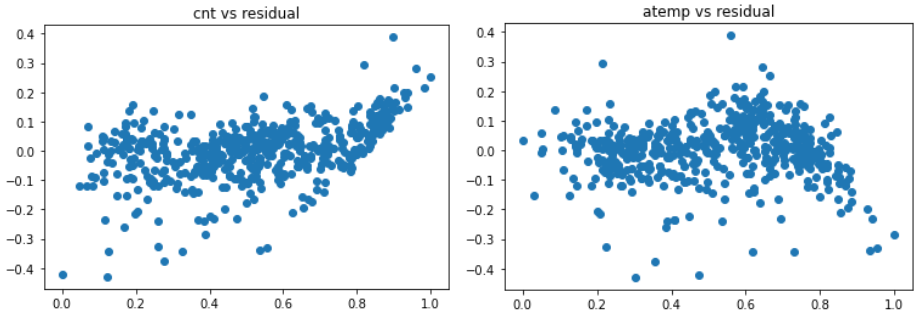


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- For validating the normal distribution of the error terms, we check whether the distribution plot of the errors and they are found to be normally distributed.



- The error terms were then plotted against X and y to check for any pattern. There was no specific pattern found.



- For validating multicollinearity, correlation matrix and VIF was used. None of the correlations were 1 and none of the VIFs were near 10.

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

- Top 3 features contributing to explaining the demand:
  - Atemp
  - weather
  - year

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2436	0.019	12.859	0.000	0.206	0.281
yr	0.2346	0.009	26.328	0.000	0.217	0.252
atemp	0.4852	0.036	13.447	0.000	0.414	0.556
fall	-0.0557	0.017	-3.319	0.001	-0.089	-0.023
spring	-0.1666	0.014	-12.056	0.000	-0.194	-0.139
summer	-0.0396	0.013	-2.947	0.003	-0.066	-0.013
LightRain	-0.2843	0.027	-10.676	0.000	-0.337	-0.232
Mist	-0.0723	0.009	-7.662	0.000	-0.091	-0.054

General Subjective Questions

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

- In linear regression, we try to predict the value of the outcome variable from the features. If we are using one feature to predict the outcome, it is known as simple linear regression, and if we use more than one feature then it is known as multiple linear regression. In linear regression, we basically fit a line through the data which is of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- Here x_1, x_2 etc are the predictor variables and y is the outcome variable. To fit a line through the data, we use Ordinary Least Squares (OLS) as the method. Here our objective is to fit a line through the data so that the sum of squared residuals between the line and the data points is minimized. This will give us the best line that can be fitted through the data.
- Before fitting a line through the data, we divide our dataset into the training data and testing datasets. This is done to ensure that whatever model we build using the training data is performing well against the test data as well.

- Then, we normalize the variables using min-max scaling or standardization. The scales of the different predictors might be very different, and this would cause the coefficients in the final model also to be on different scales and they will be very hard to interpret.
- After preparing the data by removing the outliers and scaling the data, and we fit a line through the data.
- Once we fit a line through the data, we can calculate the variance in the data that is explained by the line. This metric is known as R-squared. The higher the variance in the data that is explained by the model, the higher the value of R-squared will be and the better the model will be.
- R-squared is calculated by SSR(Sum of squares due to regression)/TSS(Total sum of squares).
- Since adding too many predictors may cause overfitting – the case where we memorize the training data so much that we are unable to generalize it on the test data - we use another metric known as adjusted R-squared which penalizes additional predictors if the data points are same.
- We also check the significance of the model using the F-statistic which is a measure of (mean-square of the model)/(mean-square of the residual) and its p-value. If the p-value is less than the level of significance, we can be sure that the association between the predictors and the output variable that we have found is not purely based on chance and that there is a relationship that exists.
- Once we have built the final model, we check whether all the assumptions of linear regression are met:
    - Normal distribution of error terms
    - Homoskedasticity
    - No correlation of the error terms with the predictors and the outcome variable.
- Once we are sure that the assumptions of linear regression are met, we can go ahead and test the model on the test data.
- If the model is performing well against the test data, which can be measured by the r-squared metric, we can go on and make predictions from this model.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet tells us about the importance of data visualization before applying various algorithms to build models. The quartet was developed by F. J. Anscombe and he argues that graphs help us perceive and appreciate some broad features and see what else is there. Most kinds of statistical calculations rest on assumptions about the behaviour of the data. The assumptions may be false, and then the calculations can be misleading. This suggests that the data features must be plotted to see the distribution of the samples that can help to identify the various anomalies present in the data. Linear regression can only be considered a fit for data with linear relationships and is incapable of handling any other kind of dataset.
- Anscombe's quartet is a group of four datasets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once the data set is plotted. In the quartet, the datasets have completely different distributions, but the linear regression model gives the same relationship for all the datasets.
- Data set 1: It fits the linear regression model pretty well.
- Data set 2: The data is non-linear and hence we cannot fit a linear regression model.
- Data set 3: The data points lie on a straight line except an outlier due to which the pearson's R coefficient is less than 1. The outlier cannot be handled by the linear regression model.
- Data set 4: For majority of the points, there is no relationship. But, due to the presence of one outlier, there seems to be a relationship in the data.

- However, all the above 4 datasets have the same mean, variance, correlation and R-squared.
- This is the reason it is very important to visualize the data before doing any kind of regression analysis.

3. What is Pearson's R?

- Pearson's R is the measure of correlation between two variables. Correlation measures the degree to which two phenomena are related to one another. For example, there is a correlation between summer temperatures and ice cream sales. When one goes up, so does the other. Two values are positively correlated if a change in one is associated with a change in the other variable in the same direction. A correlation is negative when a positive change in one variable is associated with a negative change in the other. Pearson's correlation coefficient lies between -1 and 1.
- A correlation of 1, means that every change in one variable is associated with an equivalent change in the other variable in the same direction. A correlation of -1 means that every change in one variable is associated with an equivalent change in the other variable in the opposite direction.
- The closer the correlation to 1 or -1, the stronger the association.
- Correlation coefficient is calculated as follows:
  - Calculate the mean and standard deviation for both the variables.
  - Convert all the datapoints so that each datapoint is represented by its distance in standard deviations from the mean.
  - Calculate the product of the variables for each set of datapoints and add them up.
  - Divide the product by the number of pairs of observations to get Pearson's coefficient.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Often in our data, the predictors are on different scales. Some of the predictors like population can be in the order of millions and on the other hand, temperature will be in a much smaller scale. So, if we build a linear regression model with the raw values, the coefficients of the predictors will also be on very different scales. This makes the coefficients of the model very difficult to interpret. For example, the coefficient of population, which is in the scale of millions might be in the order of 10^-3 and temperature will be on a scale of 10^1. This will give the impression that population might not have that great an impact on the outcome variable as temperature. But, this could be totally false. To overcome this, we scale the predictors and the outcome variable to bring them on the same scale so that interpretation becomes easy.
- Scaling also helps in faster convergence of the gradient descent methods.
- There are two methods which are used for scaling: normalized scaling and standardized scaling.
- Normalized scaling is also known as min-max scaling, where we scale every data in the range of minimum to maximum value. It is obtained by $(x - \min(x))/(\max(x) - \min(x))$. Normalized scaling ensures that each datapoint lies in the range of 0 to 1.
- Standardized scaling is representing each data point in terms of standard deviations from then mean. It is calculated by: $x = (x - \text{mean}(x))/\text{stdev}(x)$. The range of standardized scaling is not fixed. It is determined by how far the data point is from the mean.

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

- VIF (Variance Inflation Factor) is used to detect multicollinearity. It is the method of trying to predict a predictor using the other predictors. Once we form a model trying to explain a predictor using other predictors, we get an R-squared value for the model.
- The formula for calculating VIF is: 1/(1 – Rsquared).
- So, in case a predictor is totally explained by other predictors, the value of R-squared will be very close to 1. In this case, the VIF value will be infinite.

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

- A Q-Q plot is a probability plot which is the graphical method of comparing two probability distributions by plotting their quantiles against each other. While making statistical inferences, we assume that the sample data is coming from a certain type of underlying distribution. In case of linear regression, we assume that the data is coming from a normal distribution. Q-Q plot is a way to check whether our assumption about the underlying population distribution is correct.
- A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both the sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
- To plot a Q-Q plot, the data is sorted in ascending order, and the underlying assumed distribution curve is also divided into equal sized chunks. For example, if we have a dataset containing 10 datapoints assumed to have come from a normal distribution, we divide the normal distribution curve into 10 equal sized areas. Equal sized areas under a normal distribution curve means equal probability of a datapoint being in a particular section of the curve.
- Then, we plot the datapoints in our sample and the corresponding standard deviations from the normal distribution curve. If the data has indeed come from a normal distribution, we should get one data in each of the ten divisions and our scatter plot will lie on a straight line.
- Q-Q plot is used in linear regression to validate whether the data and the error terms are indeed normally distributed. Because if they are not, then we will not be able to make interpretations from the model.