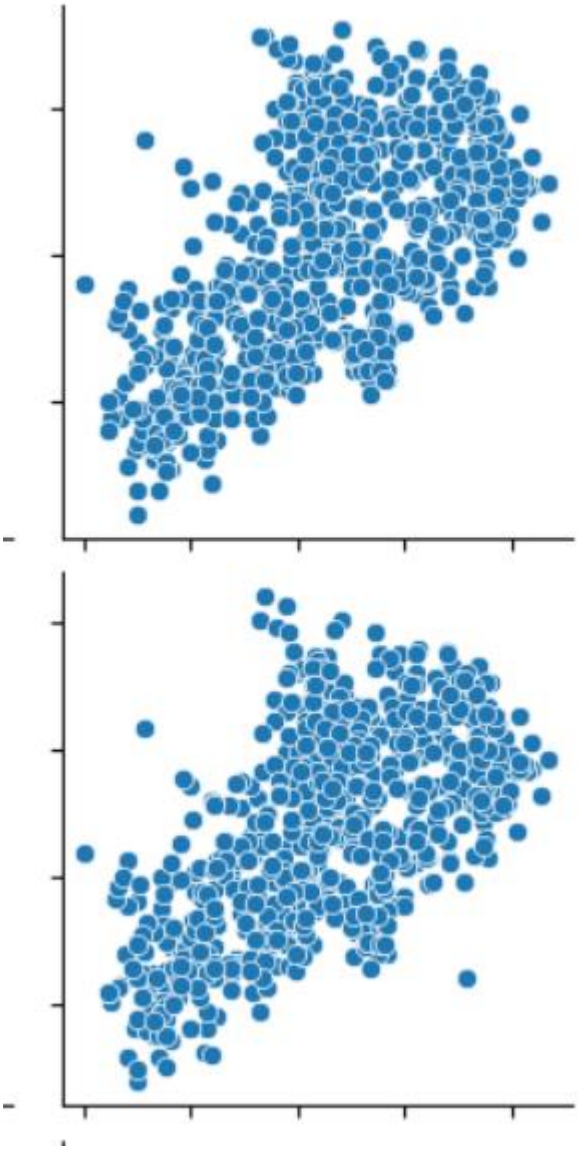
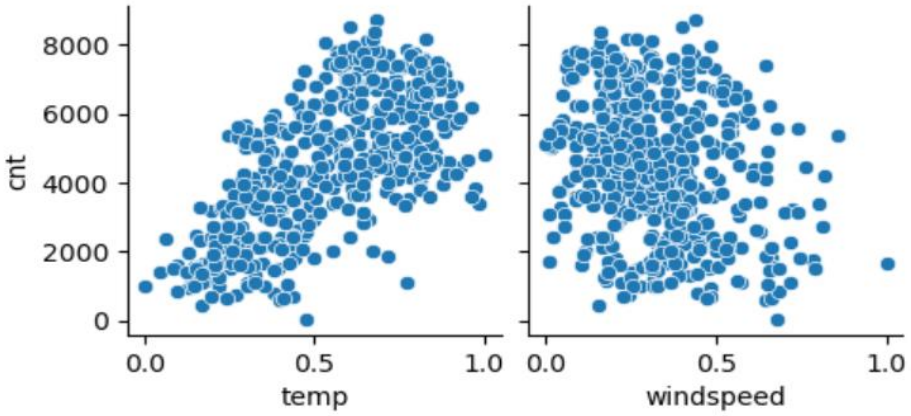
Assignment-based Subjective Questions

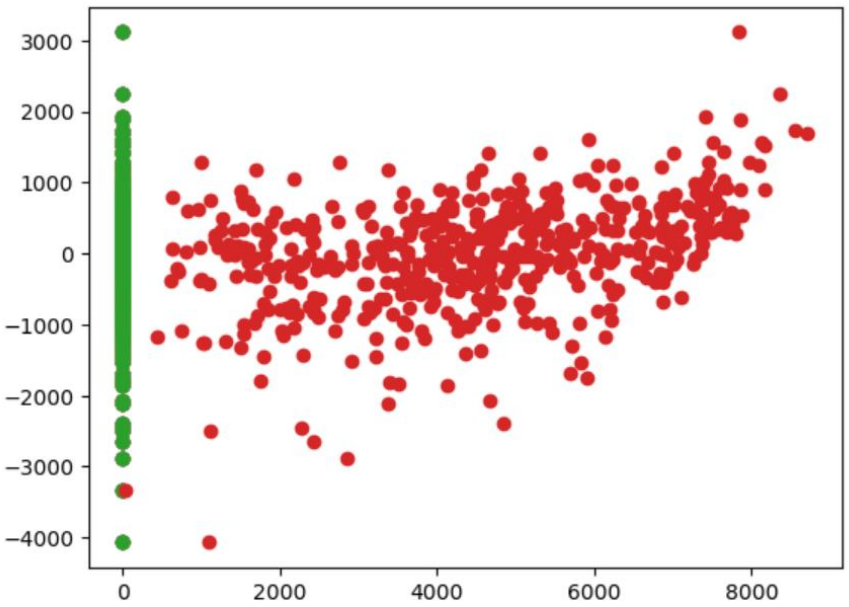
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
    a. We can infer that bike rentals are high during spring and Fall. They are also high during clear and cloudy days as compared to rainy, snowy days. We see that bike rentals are highest during months of September and October.
2. **Why is it important to use drop_first=True during dummy variable creation?**
    a. drop_first=True is needed to minimize the number of columns created during dummy variable creation, this in turn helps minimize the corelation between the variables.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    a. Temp and atemp have the highest correlation with target variable rental count.

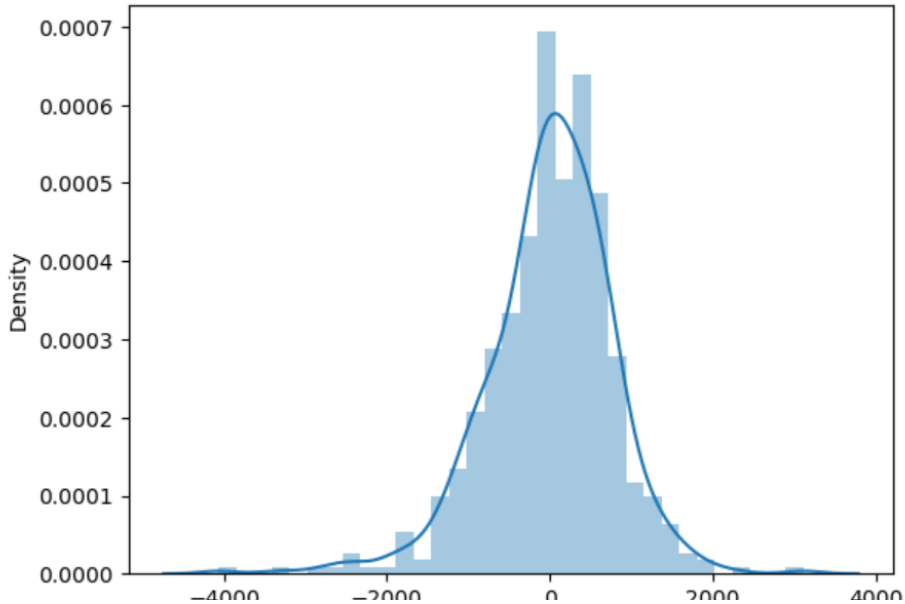


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

a. I confirmed that the error terms were normally distributed using residual analysis.
b. I removed the variables with high VIF value to take care collinearity so that the variables are independent of each other.
c. Checked the residual scatter plot to confirm that there is no obvious pattern in the residuals.
d. The residual variance mostly remains constant (slightly higher at higher values) proving homoscedasticity.







5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    a. The top 3 features contributing significantly are temperature, year and lightsnow. (they have the highest coefficient)

const	2765.6556	166.371	16.623	0.000	2438.778	3092.534
yr	2054.4942	70.558	29.118	0.000	1915.866	2193.122
holiday	-871.5045	223.588	-3.898	0.000	-1310.798	-432.211
temp	3296.6465	225.930	14.591	0.000	2852.751	3740.542
windspeed	-1368.0927	214.054	-6.391	0.000	-1788.655	-947.530
spring	-1053.8093	115.296	-9.140	0.000	-1280.338	-827.281
Sun	-427.8081	100.580	-4.253	0.000	-625.423	-230.193
Jan	-446.6869	151.787	-2.943	0.003	-744.910	-148.464
July	-533.8562	149.896	-3.562	0.000	-828.365	-239.348
Oct	496.6336	132.209	3.756	0.000	236.877	756.390
Sept	553.8025	136.187	4.067	0.000	286.230	821.375
cloudy	-714.1491	75.287	-9.486	0.000	-862.068	-566.230
lightsnow	-2599.7775	214.050	-12.146	0.000	-3020.331	-2179.224

General Subjective Questions

1. **Explain the linear regression algorithm in detail:**
    a. Linear regression algorithm is type of machine learning algorithm that can be used to compute linear relationship between one predictor variable and one or more independent variables by fitting a linear equation on the observed data. This equation can then be used to determine the importance of each feature in deriving the result. This information can then be used to make predictions about continuous variables. Linear regression works on the following assumption on the dataset and model:
        i. The variables are independent of each other.
        ii. Error terms are normally distributed.
        iii. There is no pattern in residuals.
        iv. There is constant variance in the residual.
2. **Explain the Anscombe's quartet in detail:**
    a. Anscombe quartet shows four dataset that have same statistical values but different representations on scatter plot. It is used to emphasize the importance of exploratory data analysis before building the model.
3. **What is Pearson's R?**
    a. Pearson's R is a correlation coefficient used to determine the linear relationship between two variables. It holds value between -1 and 1. -1 denotes negative corelation, 0 denotes no corelation and 1 denotes positive corelation.
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
    a. Scaling in general means increasing or decreasing the value of variable mostly to fit within certain limits. This is done to ensure that all variable values are within a certain range so we can make accurate predictions while deriving coefficients/weight of each feature.
        i. Normalized scaling can be derived by subtracting mean from each value and dividing with difference of max and min value. This is done so that the values are cantered around the mean.
        ii. Standardized scaling can be derived by subtracting the mean from each value, but we divide by the standard deviation. This is done to achieve a normal distribution in an already normal but skewed data with mean 0 and standard deviation 1.
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    a. VIF of infinity shows perfect corelation between variable and predictor variable. The formula for VIF is 1/(1-R2). If R2 is 1 then 1-R2 will be 0. This suggests that there exists high degree of collinearity (will lead to overfitting). We should analyse the data and drop those columns.
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
    a. A Q-Q plot is a quantile-quantile plot that can be used to determine if the residuals follow a normal distribution or not. If the points on the plot are on a straight diagonal line then we can say that the residuals have a normal

distribution. It can also be used to observe outliers and variance in the residuals. Hence it is used to determine whether linear regression can be used for prediction using the model or not.