**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?

**Ans:** The final model came up with the following summary:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | 1831.5022 | 260.508 | 7.031 | 0.000 | 1319.679 | 2343.325 |
| **workingday** | 444.2145 | 106.282 | 4.180 | 0.000 | 235.400 | 653.029 |
| **temp** | 5449.7276 | 185.404 | 29.394 | 0.000 | 5085.462 | 5813.993 |
| **hum** | -2047.0984 | 295.989 | -6.916 | 0.000 | -2628.630 | -1465.566 |
| **windspeed** | -1631.0111 | 245.003 | -6.657 | 0.000 | -2112.371 | -1149.651 |
| **winter** | 1010.6281 | 95.893 | 10.539 | 0.000 | 822.226 | 1199.030 |
| **Light Rain** | -1768.6664 | 242.088 | -7.306 | 0.000 | -2244.300 | -1293.033 |
| **Sun** | 505.8591 | 137.137 | 3.689 | 0.000 | 236.425 | 775.293 |
| **year\_of\_operation** | 1975.7590 | 78.748 | 25.090 | 0.000 | 1821.041 | 2130.477 |
|  |  |  |  |  |  |  |

By analysing the co-efficient it is clear if all other variables are kept constant ,then :

* ‘**const’:** Each day is going to by default attract 1831 approx. number of bike rentals despite every other contributing variables**.**

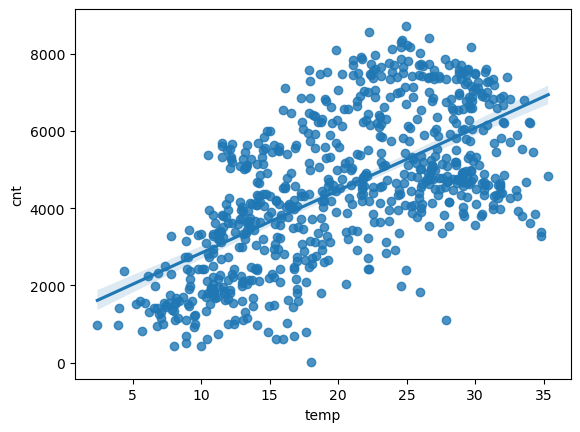
* ‘**workingday’:** Every working day is going to attract approx.. 444 count of bike rentals.
* ‘temp’: Every degree rise of temperature in Celsius is going to attract approx.. 5450 count of bike rentals.
* ‘**hum’:** Every unit rise in humidity is going to decrease approx. 2047 count of bike rentals
* ‘**windspeed’:** Every unit rise of windspeed is going to decrease approx. 1631 count of bike rentals.
* **‘winter’:** Every winters there is an increase in bike rentals by a count of 1011 approx.
* **‘Light Rain’:** In case of light rains bike rentals are likely to decrease by a count of approx. 1764.
* **‘Sun’:** Sundays are going to increase the rentals by a count of 506 approx.

**Q2.** Why is it important to use drop\_first=True during dummy variable creation?

**Ans:** The pandas.get\_dummies() method creates N dummy variables for every categorical variables having N levels . Out of the N new variables one variable is redundant which can be dropped for simplicity as the absence of all other dummy variables automatically ensures the presence of the dropped variable.

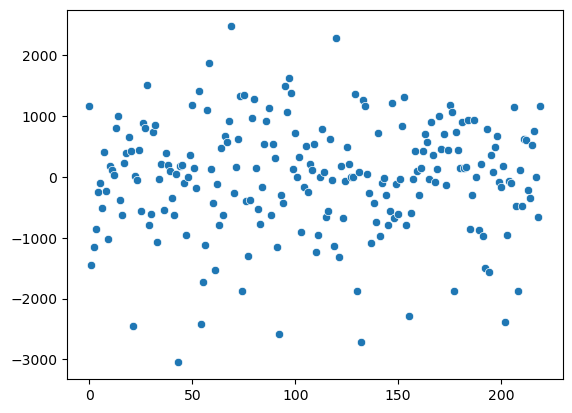
**Q3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

**Ans:** As per the scatter plot rental counts has the most linear relation with temperature.



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

**Ans:** Homoscedacity and no Multicollinearity: When the residuals(difference between predicted and actual target values) were plotted in a scatter plot ,it gets clear that there is no pattern to the errors and the errors are randomly distributed and there is constant variance. The graph below justifies that.



Normalized Errors:

When the residuals were plotted as a histogram we get a near normal distribution. It is evident from the graph obtained.

A graph with blue lines

Description automatically generated

Both these indicate that the assumptions of Linear Regression are maintained in our model and Linear Regression is good fit to our data.

**Q5**. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

**Ans:** Based on the co-efficient obtained from the model it is evident that the top three contributing features are as follows:

1. Temperature
2. Humidity
3. Year of operation.

**General Subjective Questions**:

**Q1.** Explain the linear regression algorithm in detail.

**Q2.** Explain the Anscombe’s quartet in detail.

**Ans:**

A group of graphs with orange dots

Description automatically generated

**Anscombe’s Quartet**

The above quartet is called as Anscombe’s Quartet. Its is evident from the graph that all four data types have the exact same best fit line as given by linear regression algorithm. But while the first graph seems to do a decent job at explaining the data, the rest have obviously fooled the linear regression.

In the second graph it is clear that linear regression cant handle data of polynomial types or ones that don’t possess a linear relationship.

The next two graphs clearly show the sensitivity of linear regression to outliers. Had the outliers been absent we would have got a very good fitting line through the data.

Hence we should always have a good look at the data before applying linear regression.

Anscombe’s quartet is a good reference to watch out for.

**Q.3.** What is Pearson’s R?

**Ans:** Pearson’s R is a co-efficient which is used to check the corelation between two variables.

If two variables are corelated it is very much possible that it is not a linear one. Pearson’s R very much helps in verifying the same. If there is a linear relationship then Pearson’s R gives a value close to 1. For polynomial relationship between variables, the value of Pearson’s R drops as the degree of polynomial increases.

For ,

Pearson’s R ≈ 0.91

While for ,

Pearson’s R ≈ 0.66

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

**Ans:** Scaling is a step in data pre-processing which helps to normalize the data and aids in faster convergence of the algorithm.

**Why:** The data often obtained has wide variety of values ,scales and ranges. In order to bring all the variables into a same level of magnitude and account for their varying units scaling is done so that modelling is correct and proper inferences can be made from the model.

Note: Scaling doesn’t affect the projection capability of a model but does affect the prediction capacity.

Normalized Scaling: It brings all the data n the range of 0 to 1. It is also called as MinMaxScaler in sklearn.

MinMaxScaling ( x )

Standardized Scaling: It brings all the data into a standard normal distribution with mean(μ) =0 and std\_dev(σ) =1. It is also called as the StandardScaler in sklearn.

Standardization ( x )

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

**Ans:** VIF is measure of how well a predictor variable is a linear regression of all other predictor variables. Higher the value of VIF higher is the correlation and chances of multicollinearity. In such cases the coefficients prescribed by the model become unreliable.

A variable having VIF value of infinite means that there exists a perfect correlation between the variable and other predictor variables.

The best remedy would be drop the variable and analyse the model again in such a scenario.

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

**Ans:** A Q-Q plot, short for quantile-quantile plot, is a scatterplot that compares the quantiles of two distributions. One distribution is usually the observed data, and the other is a theoretical or reference distribution, such as the normal distribution. The idea is to see how well the data fit the expected distribution by checking if the points lie on or near a straight line.

**Use and importance:**

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. You can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.