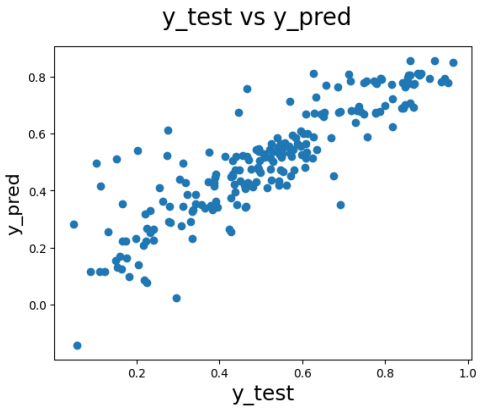
Assignment based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   (3 marks)
   - The ‘month’ variables appear to have a strong and positive effect on the dependent variable. As the months progress from March to September, there seems to be an increase in bike rentals, possibly suggesting a seasonality effect where warmer months lead to more bike rentals.
   - The ‘season’ variables show that different seasons have different impacts, While fall has the highest positive affect, spring has the least
   - The ‘weekday’ has almost no impact on the count as the median across weekdays is pretty much the same
   - The ‘yr’ (year) variable has a huge impact, although there are just 2 values, the YoY increase is quite substantial (~50% when comparing medians)
   - The 'weathersit' variables are significant and have negative coefficients, suggesting that worse weather conditions are associated with a decrease in bike rentals.
   - The ‘holiday’ variable has a significant impact too, if the given day is a holiday the rentals drop a lot, suggesting people use this mostly for traveling to work or for leisure on weekends
2. **Why is it important to use drop_first=True during dummy variable creation?**
   (2 mark)
   Using drop_first=True helps in avoiding the dummy variable trap, which is a scenario where the independent variables are highly correlated. By dropping one dummy variable (making it a reference category), multicollinearity issues are mitigated, which ensures that the model is not redundant and is more interpretable.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   (1 mark)
   Based on the regression summary output, the ‘yr’(year) variable seems to have a strong positive effect on the target variable, implicating there is a very strong growth in business YoY
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   (3 marks)

- Linearity: This can be checked using scatter plots between the predictors and the response, and also by plotting the residuals vs. the fitted values.
- Independence of Residuals: The Durbin-Watson test, which is close to 2 in your output, suggests that the residuals are not autocorrelated.
- Homoscedasticity: This assumption can be verified by plotting the residuals vs. the fitted values and looking for a constant variance.
- Absence of Multicollinearity: VIF values, all below 5 in your model, suggest low multicollinearity.



	Features	VIF
0	yr	1.96
13	weekday_1	1.81
15	weekday_3	1.75
18	weekday_6	1.69
14	weekday_2	1.63
16	weekday_4	1.62
17	weekday_5	1.61
19	weathersit_2	1.56
8	month_8	1.49
3	month_3	1.48
9	month_9	1.41
5	month_5	1.40
11	month_11	1.39
10	month_10	1.38
12	month_12	1.37
4	month_4	1.37
7	month_7	1.33
6	month_6	1.32
2	month_2	1.28
1	holiday	1.16
20	weathersit_3	1.11

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

   Based on the coefficients in the model summary, the top three features appear to be
   - **yr**(year): there is a massive increase YoY
   - **weathersit_3**(heavy rain): this is adversely affecting the count of rentals
   - **month_X** (month_9 has the highest impact in September): this is a positive affect

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Linear regression is a linear approach for modeling the relationship between a dependent variable and one or more independent variables. The algorithm assumes a linear relationship between the input variables (X1, X2…) and the single output variable (Y). More specifically, the algorithm calculates the best-fitting line that predicts Y from X.
   The best-fitting line is calculated using the Least Squares method, which minimizes the sum of the squares of the differences (the residuals) between the observed and estimated values.
   In general this can be achieved using various libraries, notably Scikit-learn and statsmodel in the python ecosystem
2. **Explain the Anscombe's quartet in detail. (3 marks)**
   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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough
3. **What is Pearson's R? (3 marks)**
   Pearson's R, also known as Pearson's correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. The value ranges between -1 and 1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Scaling is the process of transforming the range of numerical features in a dataset. It ensures that each feature contributes equally to the computation of distances between data points or the gradient in optimization algorithms. In other words, scaling ensures that no variable is artificially inflated due to differences in units or scale, which makes the algorithms work better and converge faster.
   Normalized Scaling (Min-Max Scaling):
   a. It transforms features by scaling each feature to a specific range, usually between 0 and 1.

   Standardized Scaling (Z-score Normalization):
   b. It transforms the features by subtracting the mean and dividing by the standard deviation, resulting in features with zero mean and unit variance.
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

Infinite VIF usually happens when one variable is a perfect linear combination of other variables, indicating perfect multicollinearity. In such a scenario, the variables are redundant, and it becomes impossible to determine the individual effect of each variable.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   In statistics, a Q–Q plot (quantile-quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.