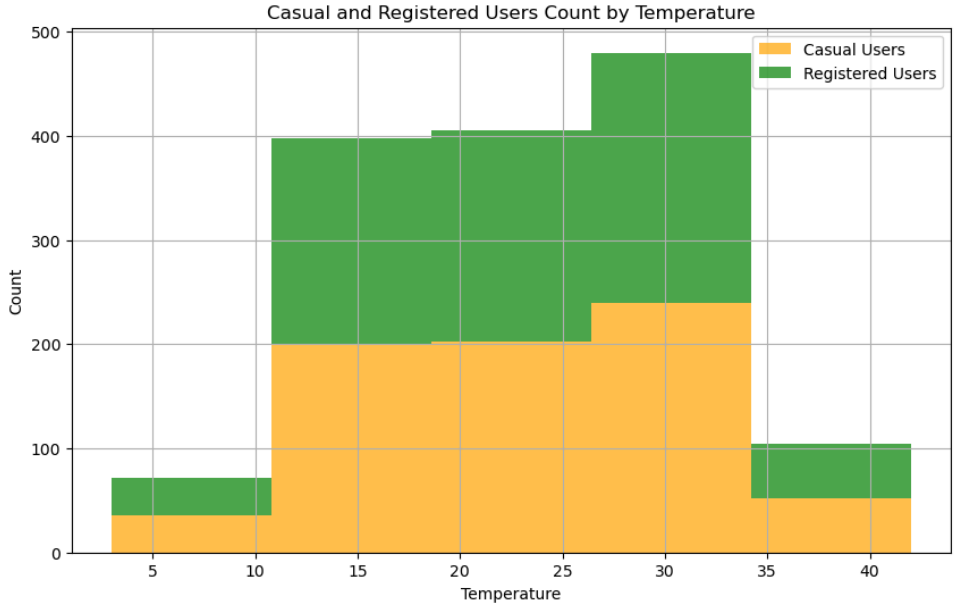
Assignment-based Subjective Questions

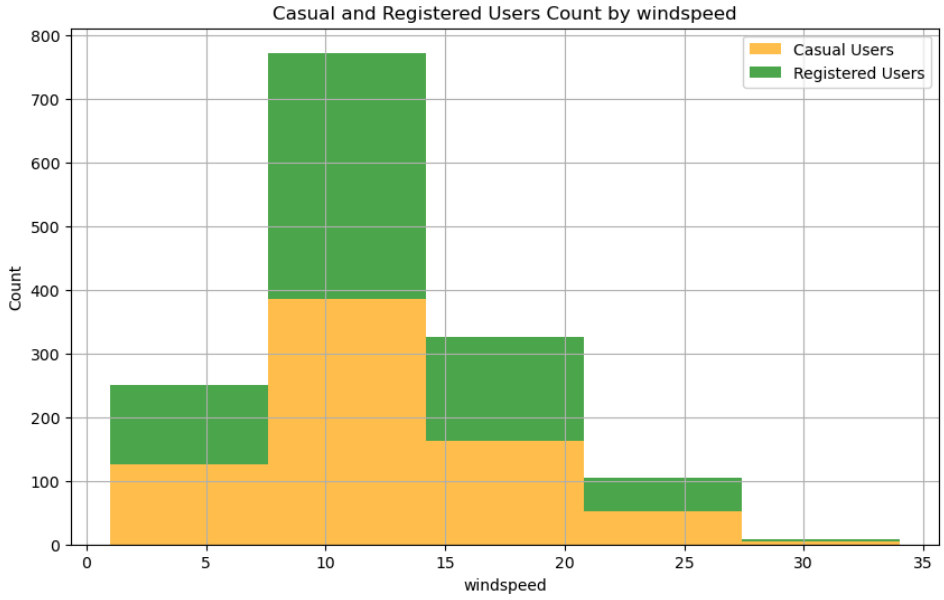
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans: After analyzing categorical variable with dependent one we found some of them had significant effect on target variable, like

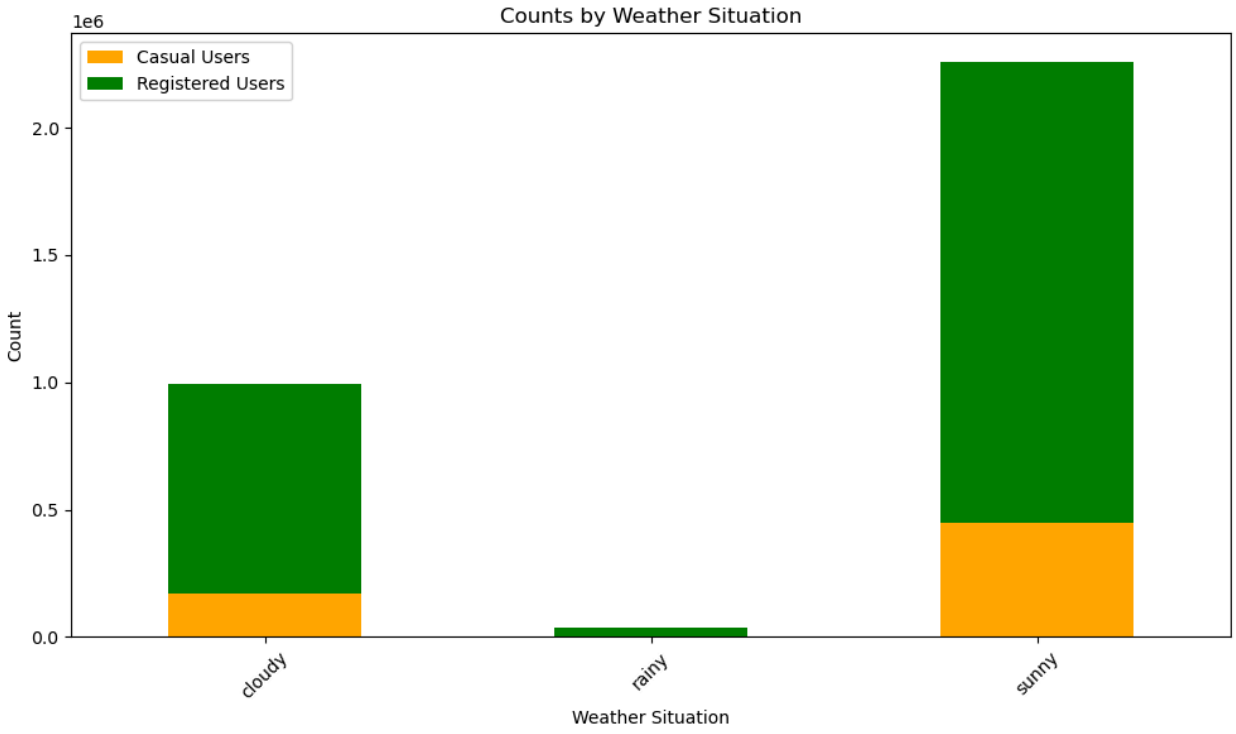
- **Temp**: If the temperature is not suitable to be outside (too cold or too hot) than it reduces the usage of bikes and it affects both type of users.



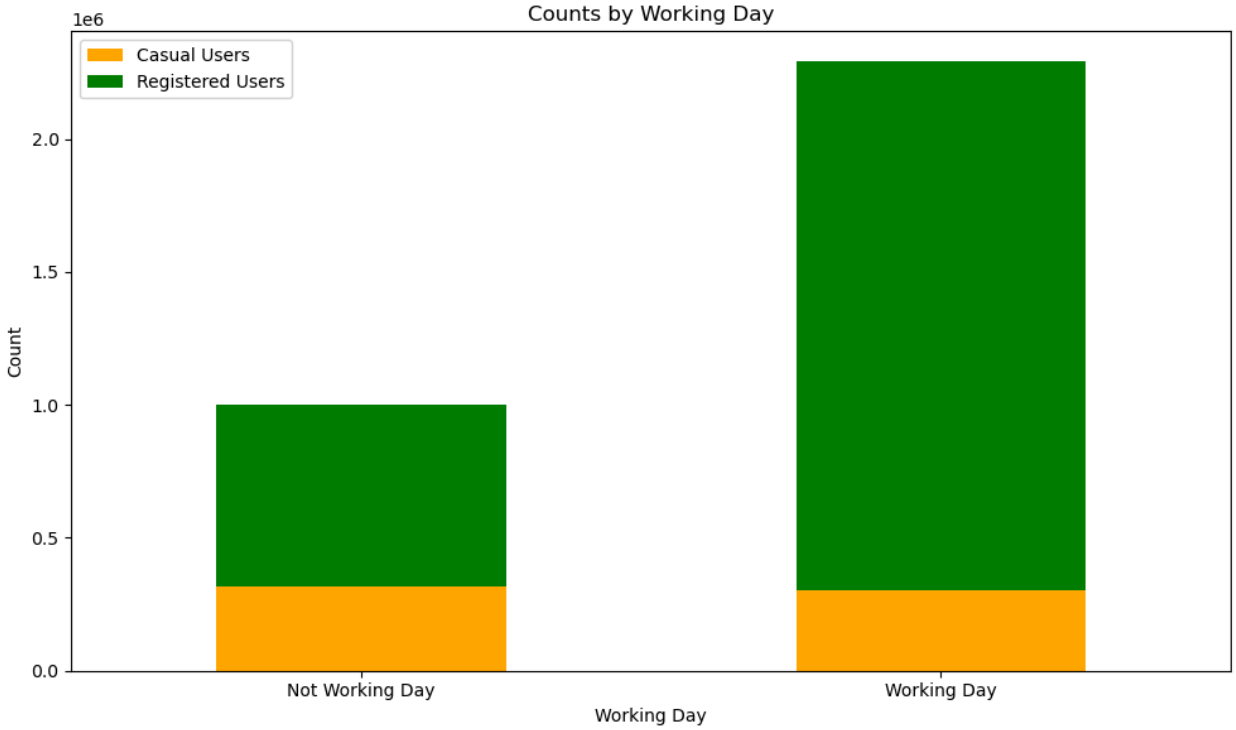
- **Windspeed**: If the wind speed is too high, it significantly reduces the users of bikes as it might be dangerous to do so.



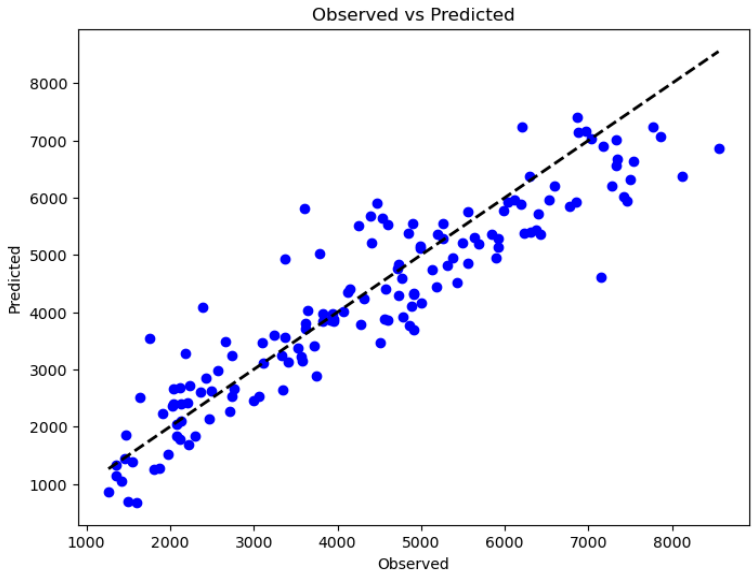
- **Weather**: One of the most important factor as if weather gets works people avoid riding bikes. There are almost no users for **stormy(4)** weather.

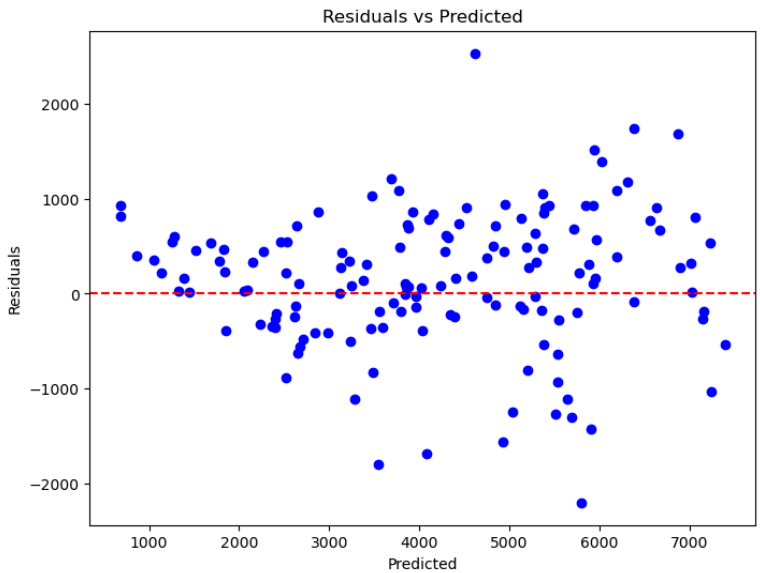


- **Working Day:** There are a lot of registered users who use bike service to go to their daily work. Therefore counts can significantly differ if its a working day or not.



2. **Why is it important to use drop_first=True during dummy variable creation?**
   Ans: By using **drop_first=True** we can avoid Multicollinearity and make sure the models are properly specified.By dropping the first dummy variable, we avoid perfect multicollinearity between the dummy variables and the intercept term in the regression model.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Ans: "atemp" has the highest correlation i.e. **"0.631"**.
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Ans: We can use r2 to calculate variance or we can use the following more advance techniques one of them is Linearity check by plotting the observed values against the predicted values or by using partial residual plots.





5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Ans:
   a. atemp
   b. weathersit
   c. Windspeed
      (by number ‘yr’ is also significant but we only have 2 years)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Ans:
   a. First of all we will assume that the data relationship between independent and target variables is linear, we can plot it and check if all points are around line.
   b. $y = \beta0 + \beta1*x1 + \beta2*x2 + ... + \beta n*xn + \varepsilon$
      i. y is the dependent variable (target).
      ii. β0 is the intercept term (the value of y when all predictors are zero).
      iii. β1,β2,...,βn are the coefficients of the independent variables (the slopes of the regression line).
      iv. x1,x2,...,xn are the independent variables (features).
      v. ε is the error term (the difference between the observed and predicted values).
   c. In linear regression our motive is to reduce the sum of squared differences between the observed values and the predicted values as much as we can.
   d. Estimation of Coefficients β0,β1,...,βn to fit the most dots on line
   e. Divide all the data that we have into train and test data
   f. Do training and testing
   g. Do verification like Linearity check
2. **Explain the Anscombe’s quartet in detail.**
   Ans: Anscombe's quartet is a classic dataset in statistics that consists of four sets of x and y values. Each set of data appears to have similar statistical properties when examined using simple summary statistics such as mean, variance, correlation coefficient, and linear regression line. However, when graphed, the datasets reveal very different relationships between the variables.
3. **What is Pearson’s R?**
   Ans: Pearson’e r is used to find linear relationship between two variables. It quantifies the strength and direction of the linear association between the variables.

a. Formula of pearson r is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Ans: Scaling is basically changing the range of values so they can better fit with other values. No by changing we dont mean replacing but using proper type of scaling processes. Normalized scaling preserves the original range of values but may be sensitive to outliers. Standardized scaling centers the data around the mean and scales it by the standard deviation, making it robust to outliers but not preserving the original range of values.
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Ans: This might happen due to multicollinearity, It can be treated by removing redundant variables or changing the model specification for better.
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Ans: QQ plot tells weather a data is normally distributed or not. It plot the values on plot alongside normal distribution and if points follows the line, it means the data is normally distributes, Below is an example