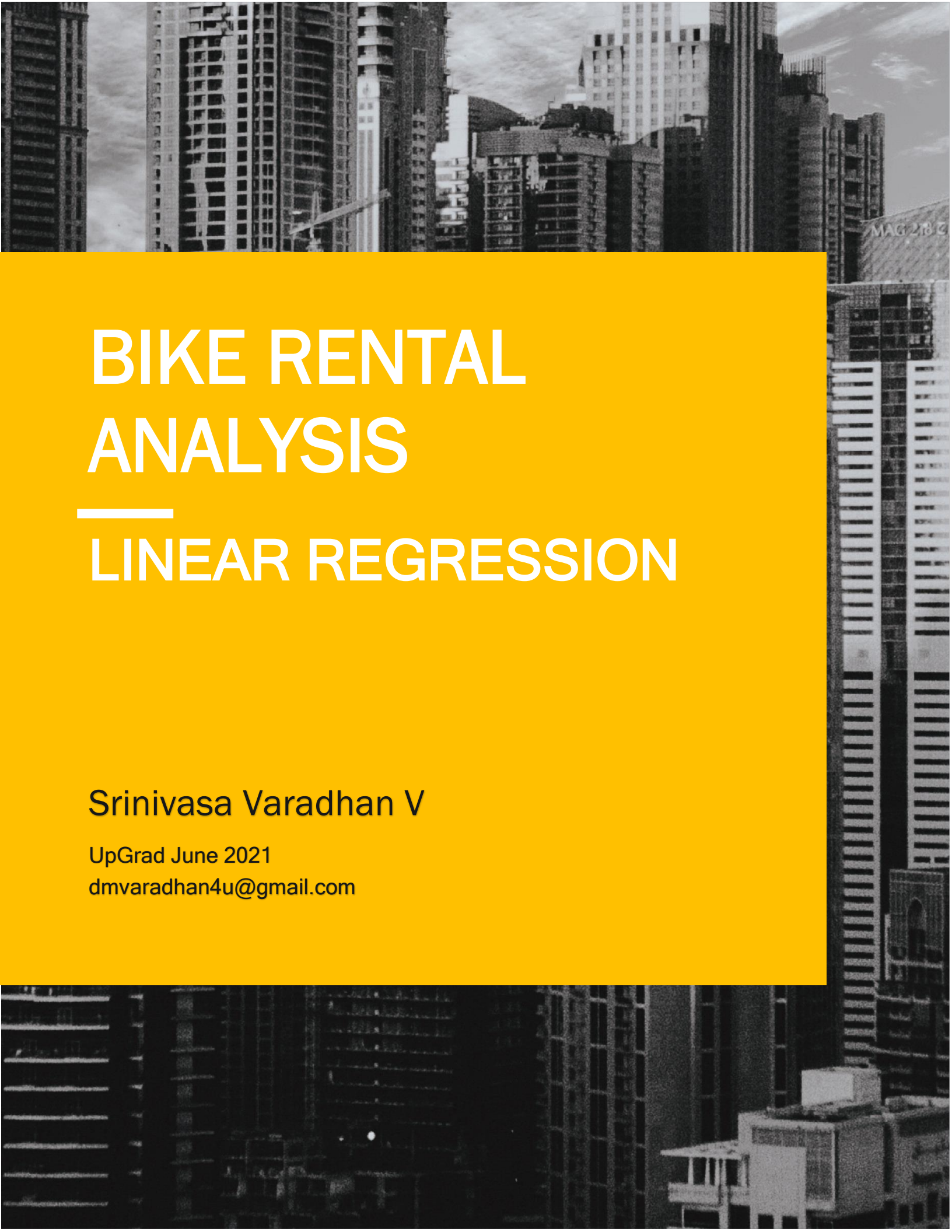


BIKE RENTAL ANALYSIS

LINEAR REGRESSION

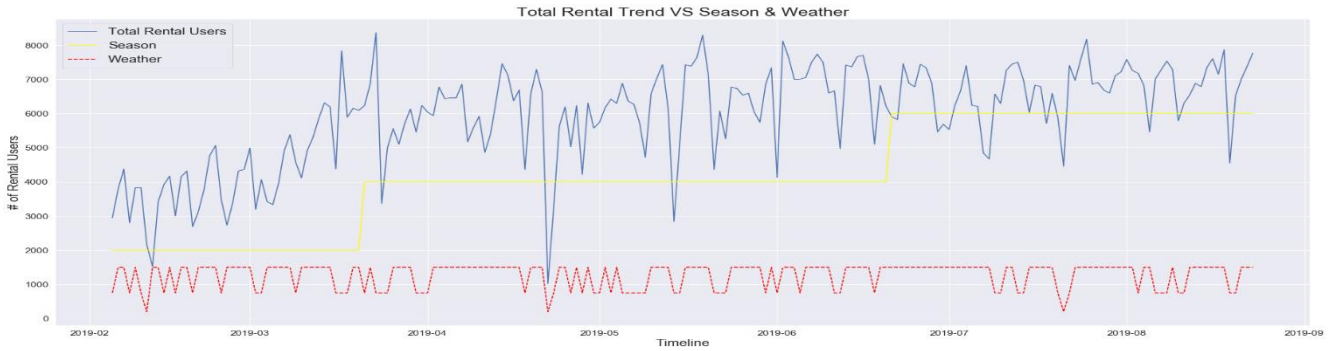
Srinivasa Varadhan V

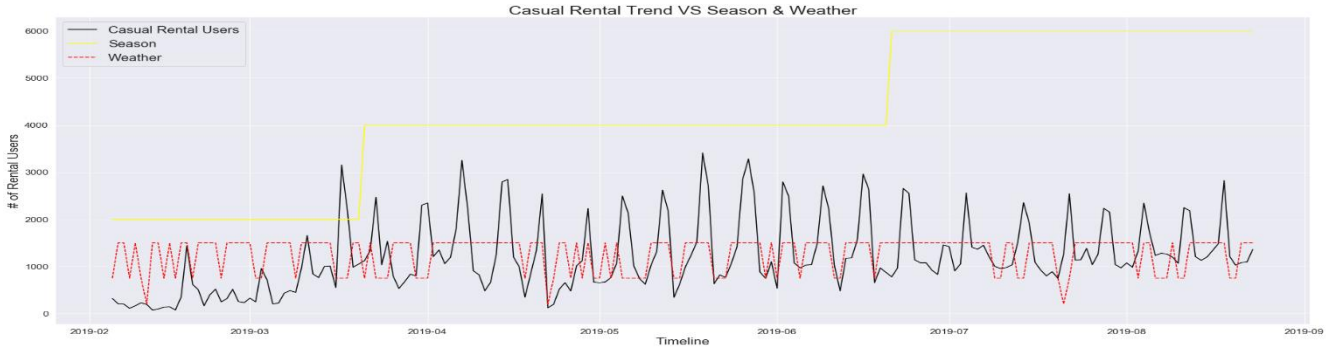
UpGrad June 2021

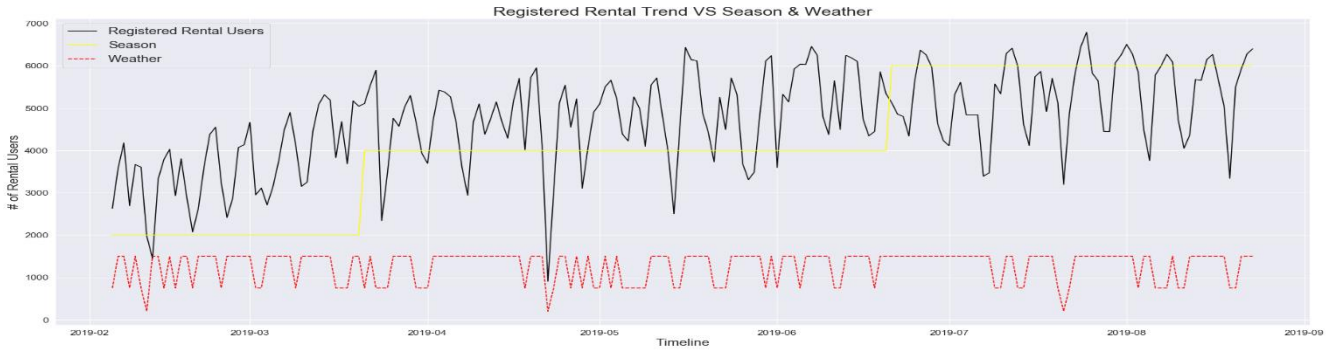
dmvaradhan4u@gmail.com

Assignment based Subjective Question

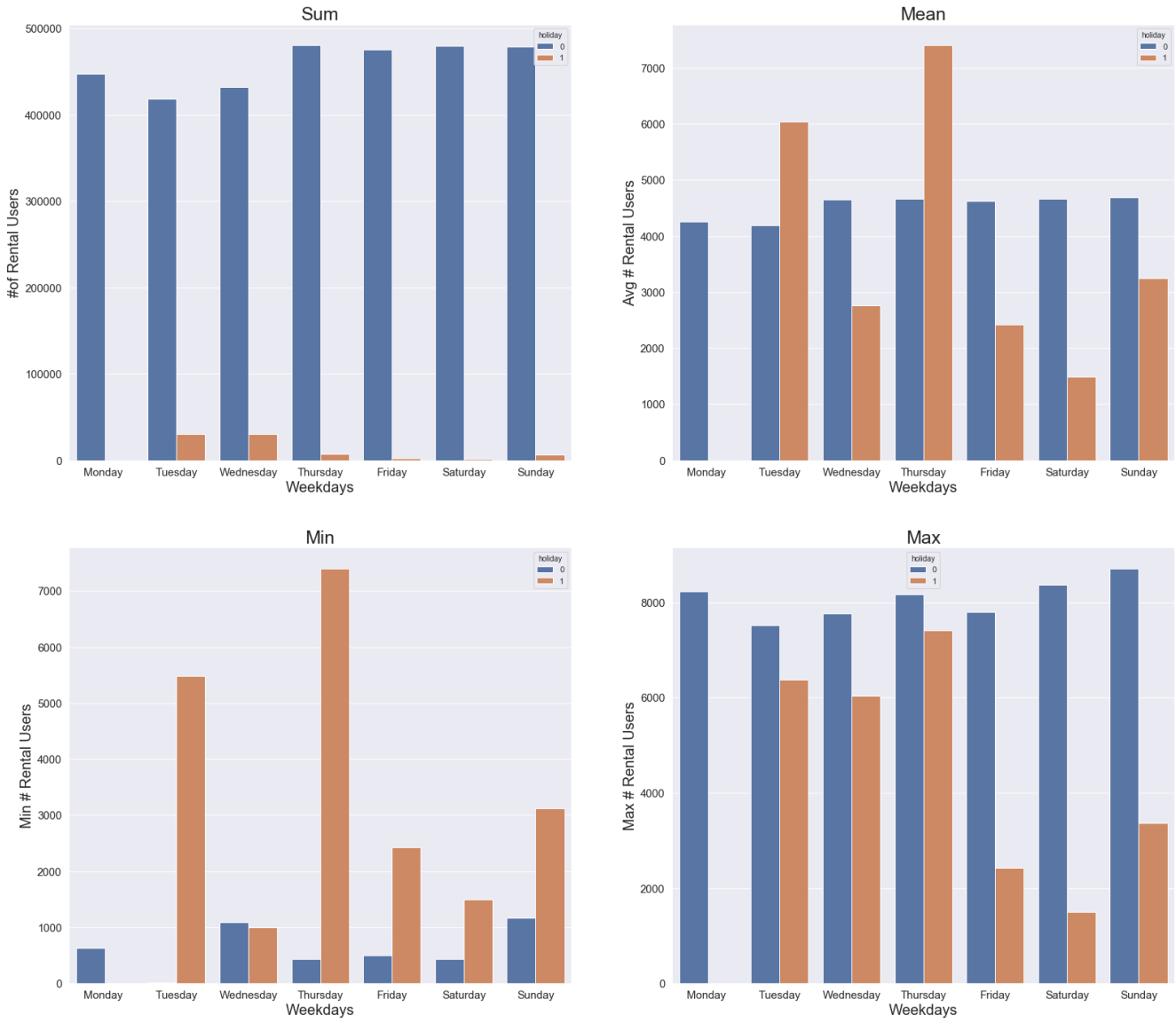
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?







- Significant negative relationship with `Rain/Snow and Fog`
- Significant positive relationship with `Clear Cloud`
- Significant positive relationship with `Summer`
- Significant negative relationship with `Winter`
- `cnt` tends to rise when continuous days have clear weather post rain



- `Thursday`, `Friday`, `Saturday` and `Sunday` have high rentals
- In these two years, there has been one holiday on `Thursday` and `Tuesday` and rentals have peaked as it was long holiday
- Average ride on a given day is around `4000 - 4200 rides`

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

- In modelling, the dummy variables are created to incorporate the categorical variables in form of numerical columns. While creating a dummy variable, it creates individual Boolean value columns against the unique values of the original categorical column.
- For eg:

Column Name
Option A
Option B
Option C

Table 1 - Original Column

Column Name	Option A	Option B	Option C
Option A	1	0	0
Option B	0	1	0
Option C	0	0	1

Table 2 - df with dummy variables

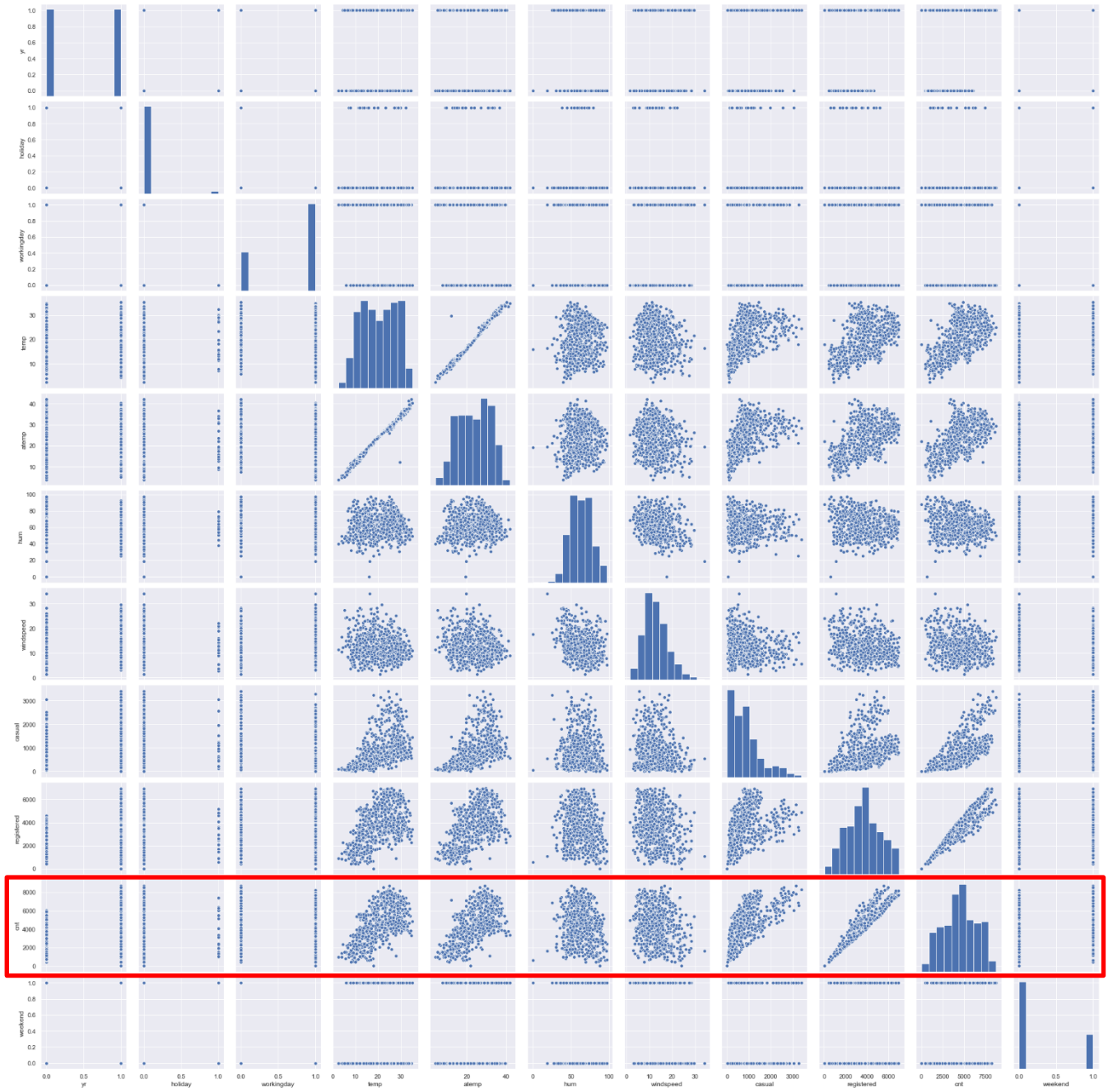
But in order to reduce the number of columns for a model, we use drop_first, because the required detail is still communicated.

Column Name	Option B	Option C
Option A	0	0
Option B	1	0
Option C	0	1

Table 3 - df with drop_first

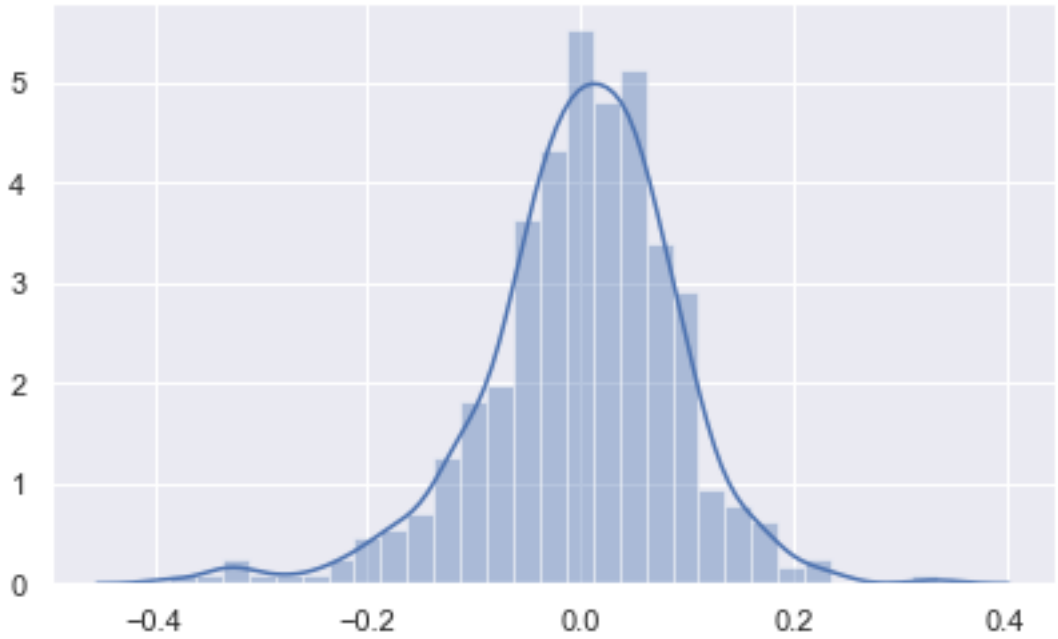
- 00 - `Option A`
- 10 - `Option B`
- 01 - `Option C`

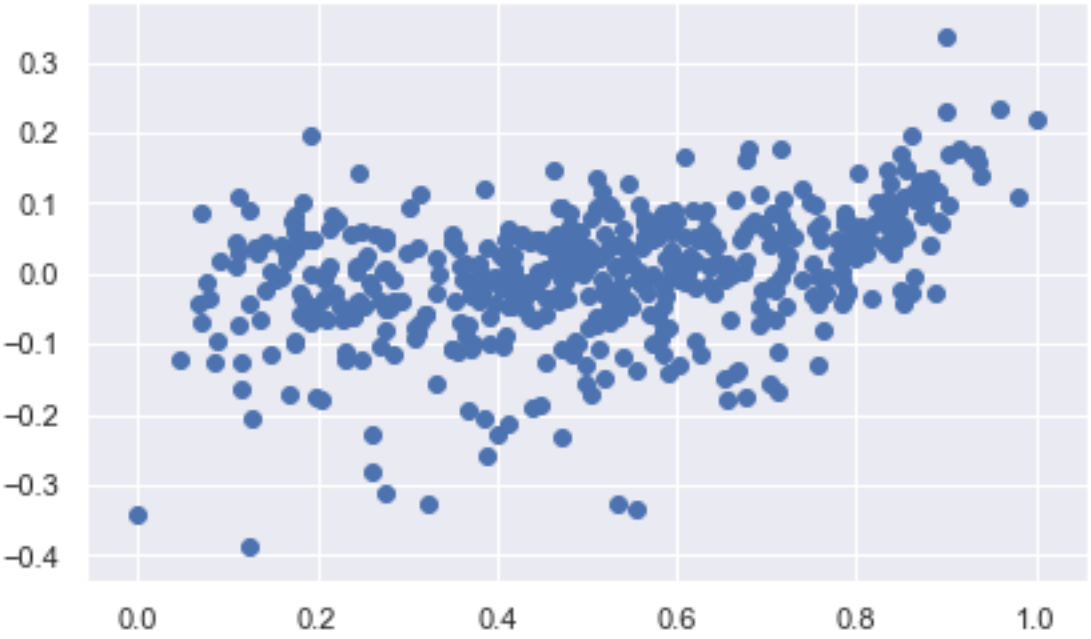
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



- **temp & atemp** have highest correlation with cnt

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





- The residue (y_true – y_pred) has mean around zero and is normally distributed. Therefore, linear regression is a right approach
- There are no visible patterns in the error terms

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

- Temperature
- Year (in this case years of operation)
- Light Rain/Light Snow

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear Regression is one among the methods of machine learning. The regression is used to predict an output or a target variable provided only one of the predictor variable varies and the other remains a constant.
- The regression model is built on the concept of best fit line and the equation of the same is y = mx + c
- Here m is the slope and the c is the intercept.
- For a simple linear regression it is y = b1x + b0
- In case of multiple regression model, there are multiple predictor variable
- Linear regression can only be interpolated, not extrapolated.
- Linear regression not necessarily determines the causation

2. Explain the Anscombe's quartet in detail.

- Anscombe quartet consist of four datasets
- The mathematical statistics (mean, standard deviation, variance, R2) of all these four datasets are the same but they have different distribution when plotted in a graph
- It was developed to say the importance of graphing before concluding and how statistics can easily misguide without graphical visualization.

3. What is Pearson's R?

- It is a statistic measure of linear relationship between two variables
- The value ranges from -1 to 1
- In case the correlation value is 1, then with every increase in a X, there is an equal increase in Y
- In case the correlation value is -1, then with every increase in X, there is an equal decrease in Y
- In case of correlation value 0, with increase in X, there is no change in Y

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

- Scaling is a method implemented on continuous numerical values each of being in different range
- The scaling method is important as different range of values, some functions in linear regression does not performs as anticipated.
- It is also performed, as linear regression uses gradient descent, it assist in faster convergence.
- The Normalized scaling, maps the max value to 1 and min value to -1 and the remaining is proportional to the max value

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- In the standardization method, the value is subtracted with the feature mean and divided by the feature standard deviation. In this method the value need not necessarily between -1 and 1.

$$x' = \frac{x - \bar{x}}{\sigma}$$

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

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

- Here the R2, is the proportion of variance in the predictor variable.
- If the correlation between two predictor variable is perfect 1, then VIF tends to be infinity
- VIF is a measure of multicollinearity that exists amongst the predictor variable
- This is important as the basic assumption of a linear regression is, keeping all value except one predictor variable constant, change in this, changes the target variable. But due to multicollinearity, there will be change in one another variable

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

- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
- It helps to determine if two data sets come from populations with a common distribution.
- Aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.