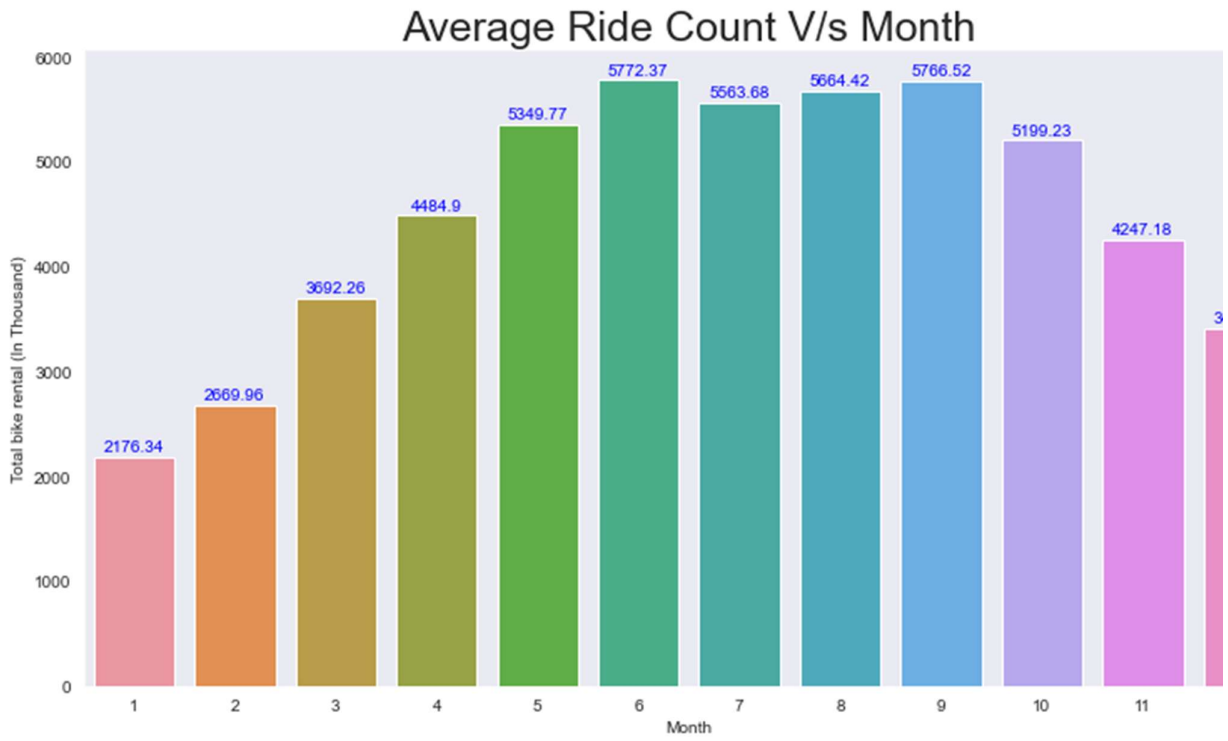
Assignment-based Subjective Questions

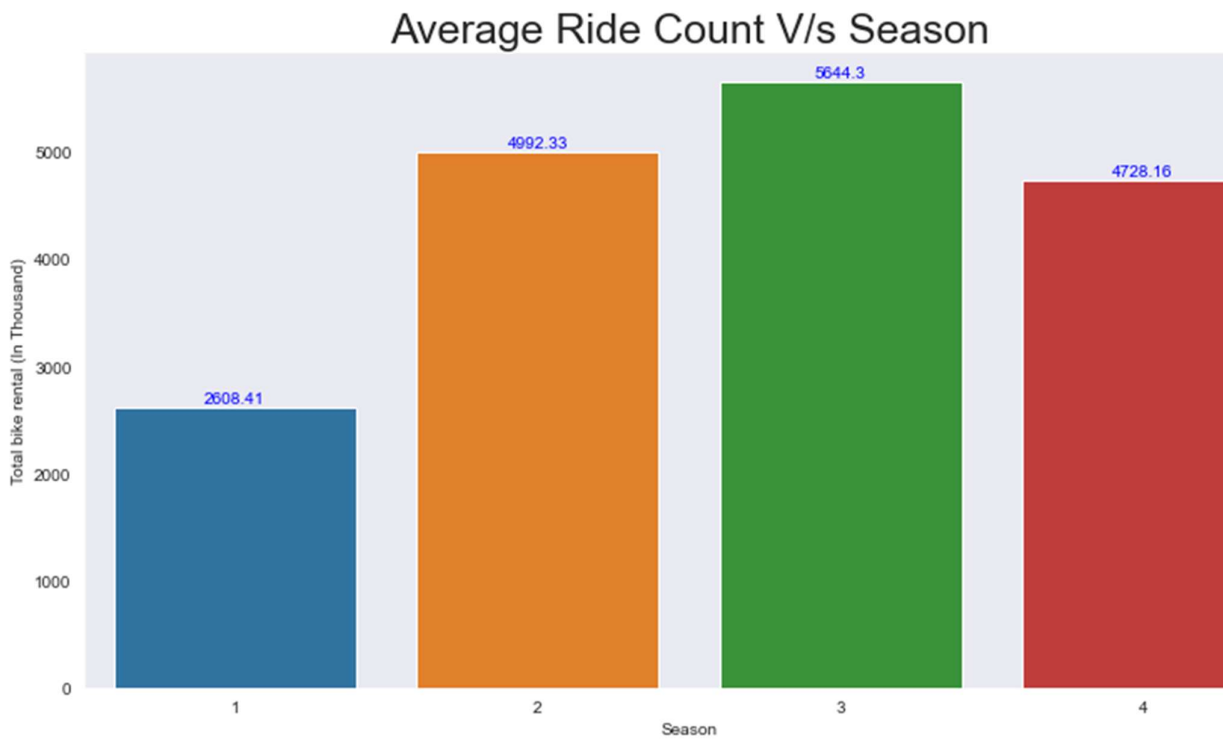
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Please find the below answer supported by charts.

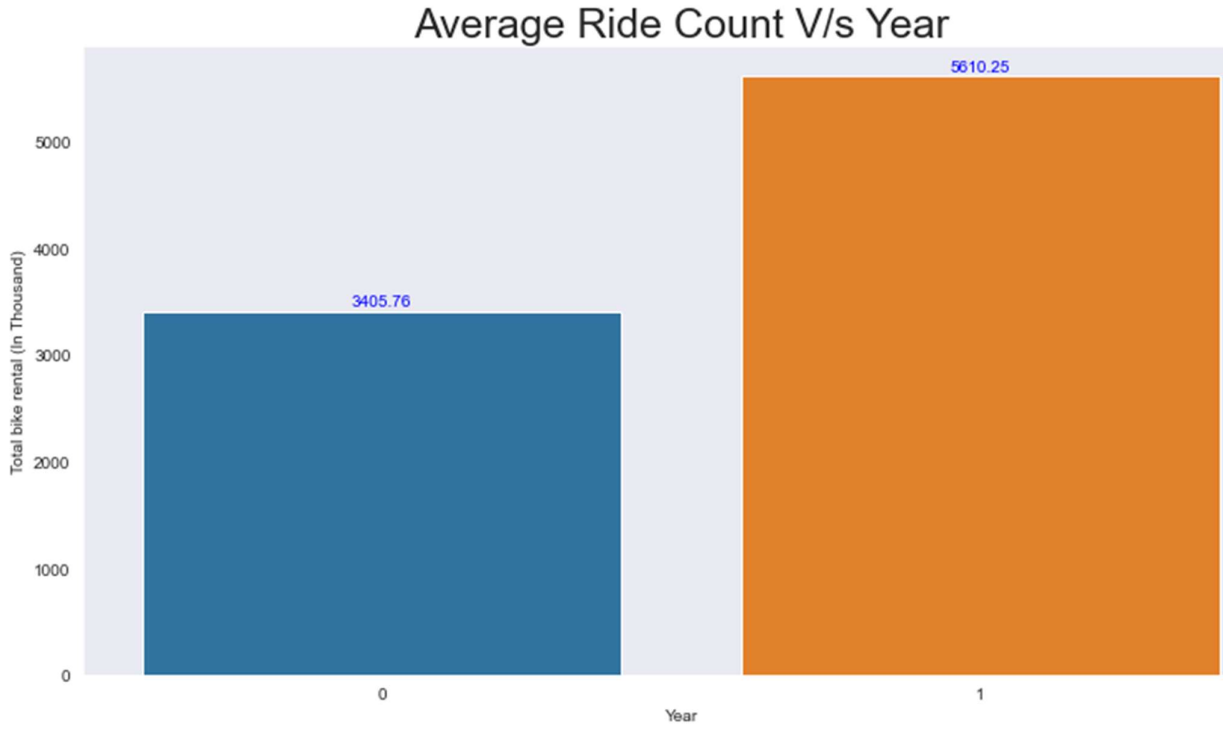
1. Average Ride Count V/s Month: Months 5, 6, 7, 8, 9 and 10 have a good bike rentals relative to others months which are May, June, July, Aug, Sep and October



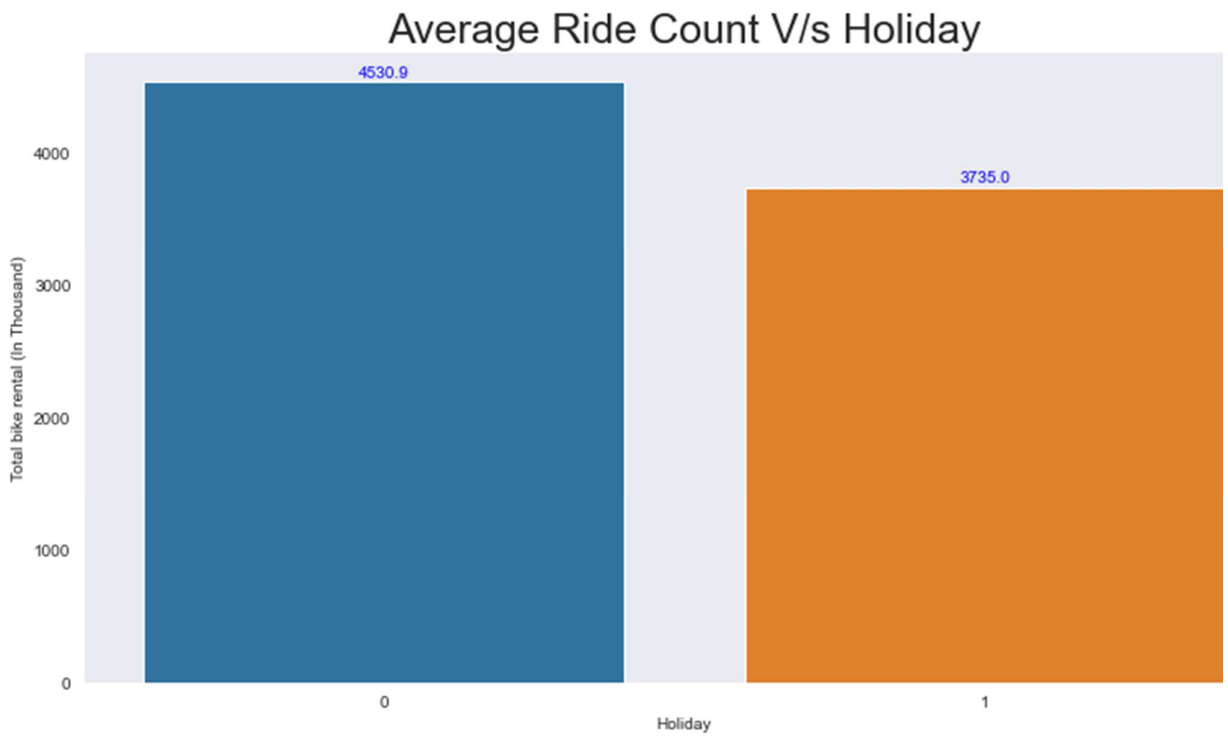
2. Average Ride Count V/s Season : Season 2, 3, 4 which is summer, fall and winter respectively has a positive impact on number of bikes rented, while in season 1 which is spring the number of bike rentals are less



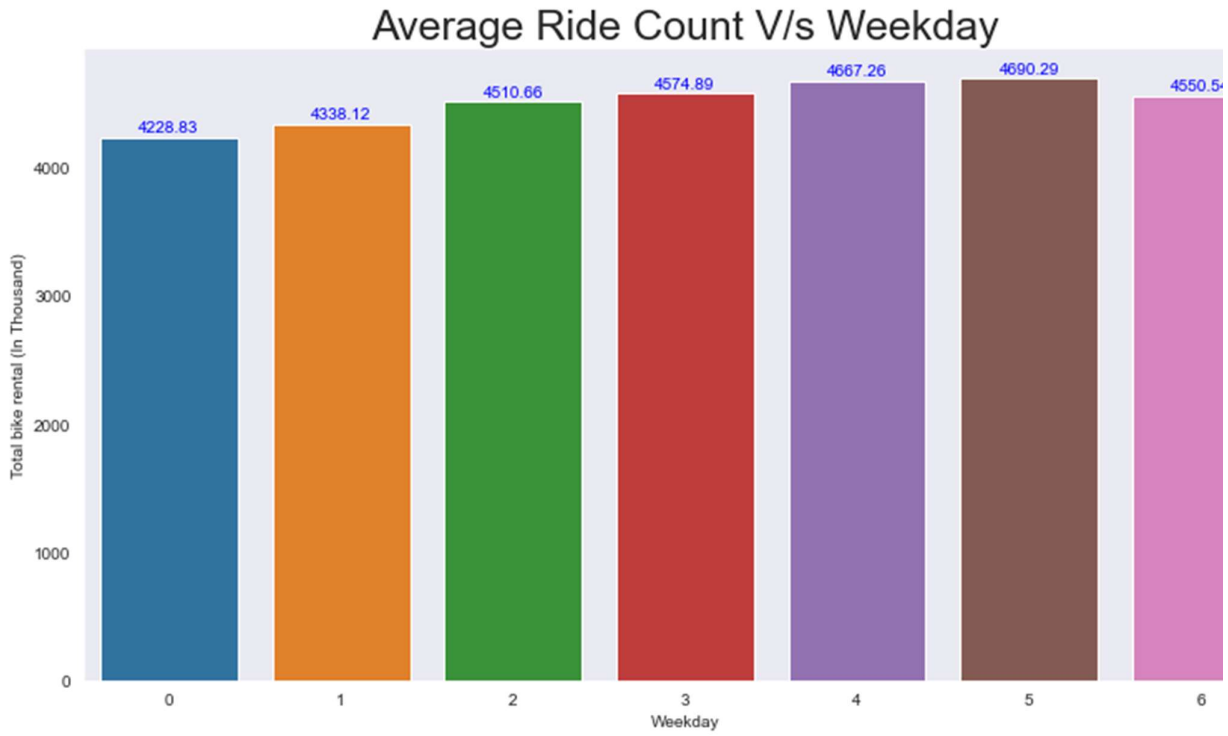
3. Average Ride Count V/s Year: Year 1 has a good increase in Total number of rental bikes as compared to Year 0, which is 2019 and 2018 respectively.



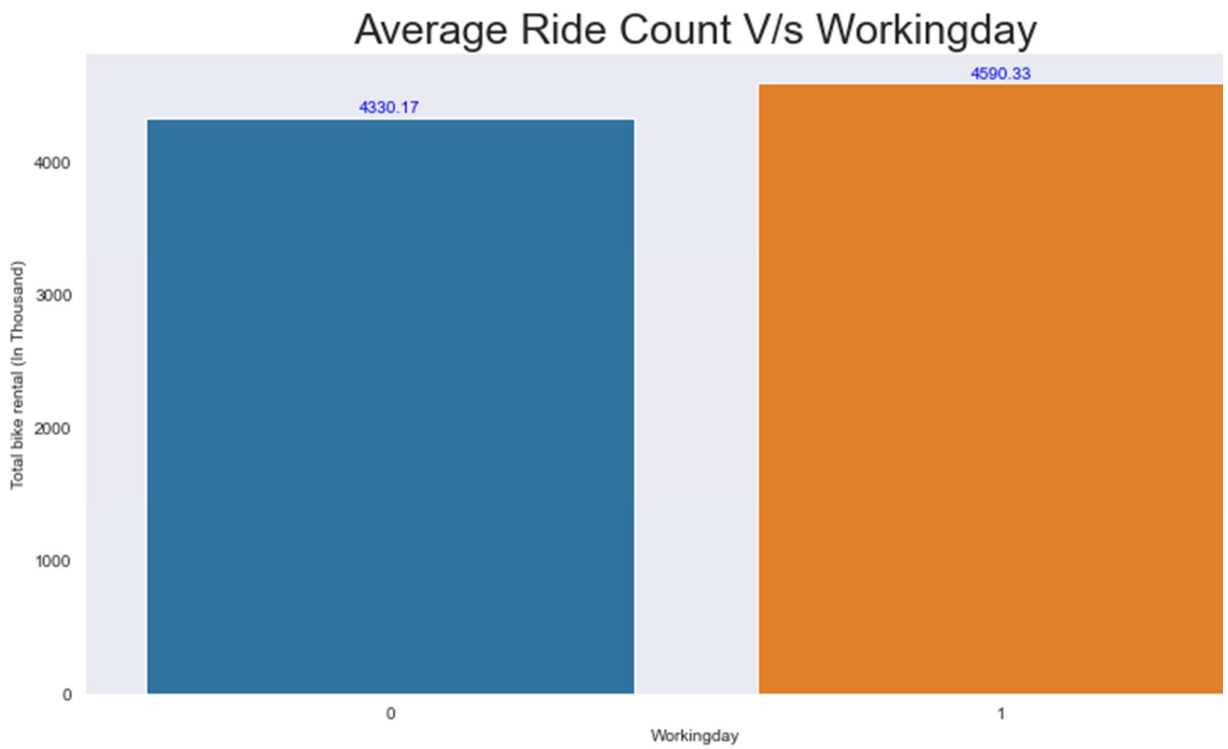
4. Average Ride Count V/s Holiday: Bike rentals are not good when holiday=1, thus on a holiday the bike rentals really drop



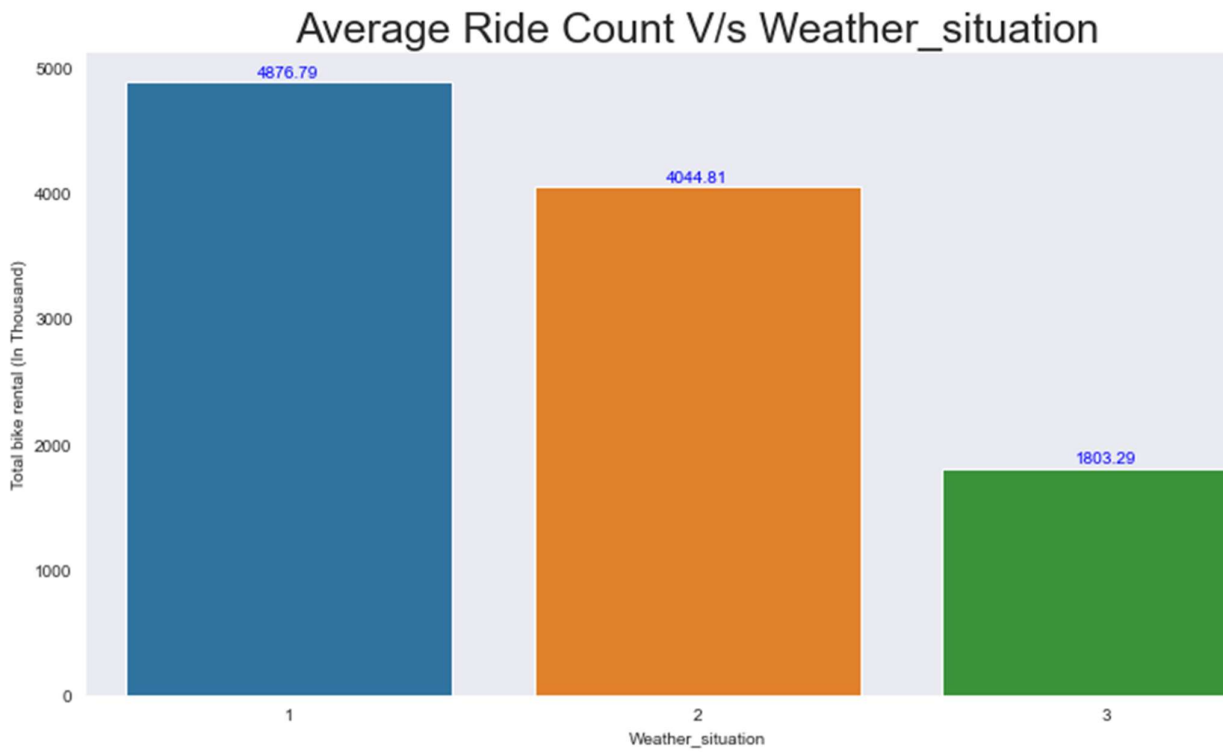
5. Average Ride Count V/s Weekday: Bike rentals drop when Weekday = 0 and pick up steady over the other days until Weekday =5, again on Weekday=6 it starts to drop



6. Average Ride Count V/s Workingday: Bike rentals increase when Workingday = 1 and drop when Workingday = 0 which means people avail bike rentals more on a working day. This also matches with "Average Ride Count V/s Holiday" results



7. Average Ride Count V/s Weather_situation: Weather situation as 1 (which is Clear day) has highest bike rentals followed by 2 (Which is Misty, coludy day). Weather situation as 3 (light snow, light rain etc.,) has least registrations while weather situation 4 has no registrations at all



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

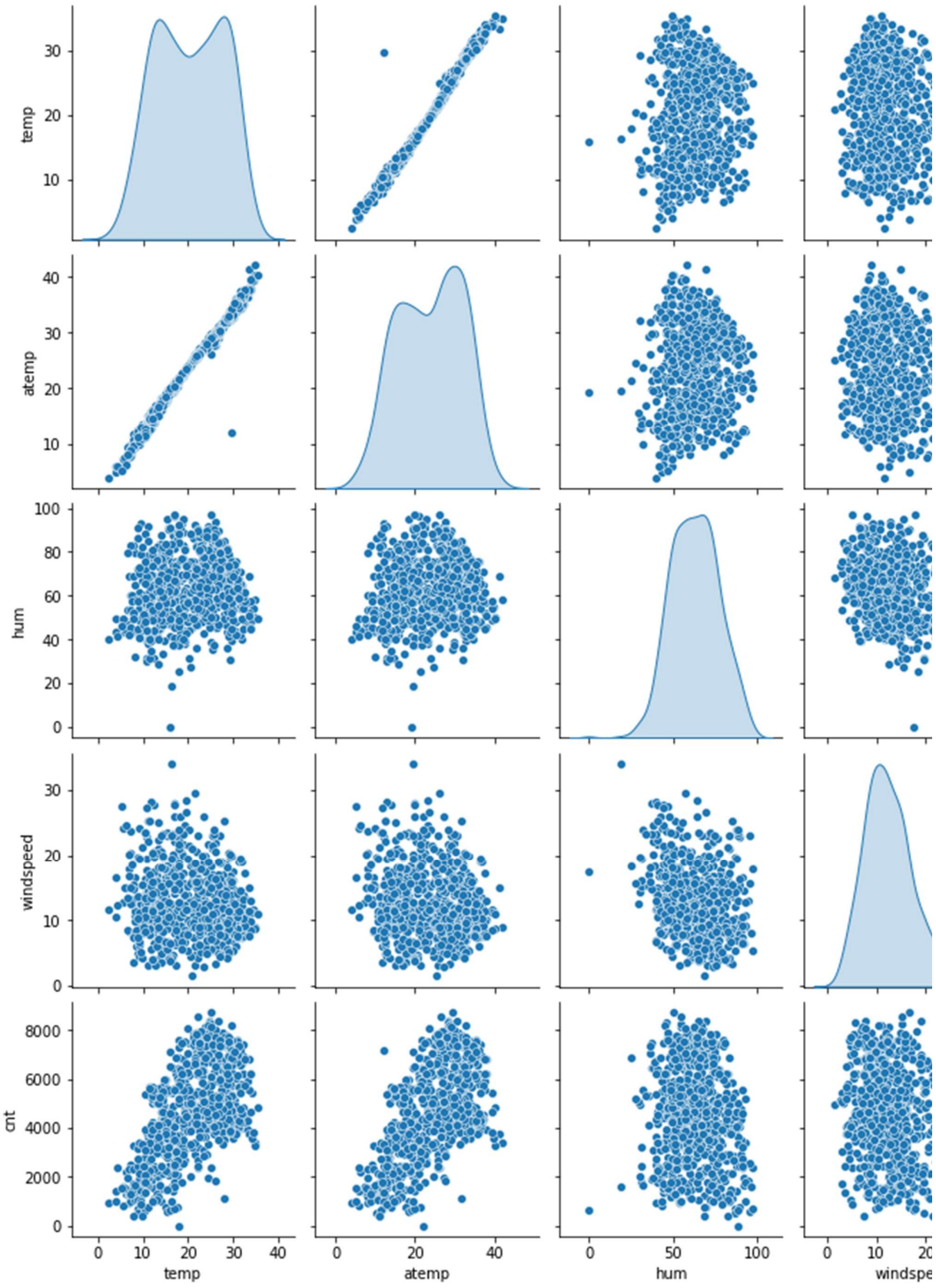
Answer:

drop_first=True helps in reducing extra column that is created during dummy variable, when we create dummy variables for continuous variables. This is done to reduce correlations created among dummy variables and have optimum dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

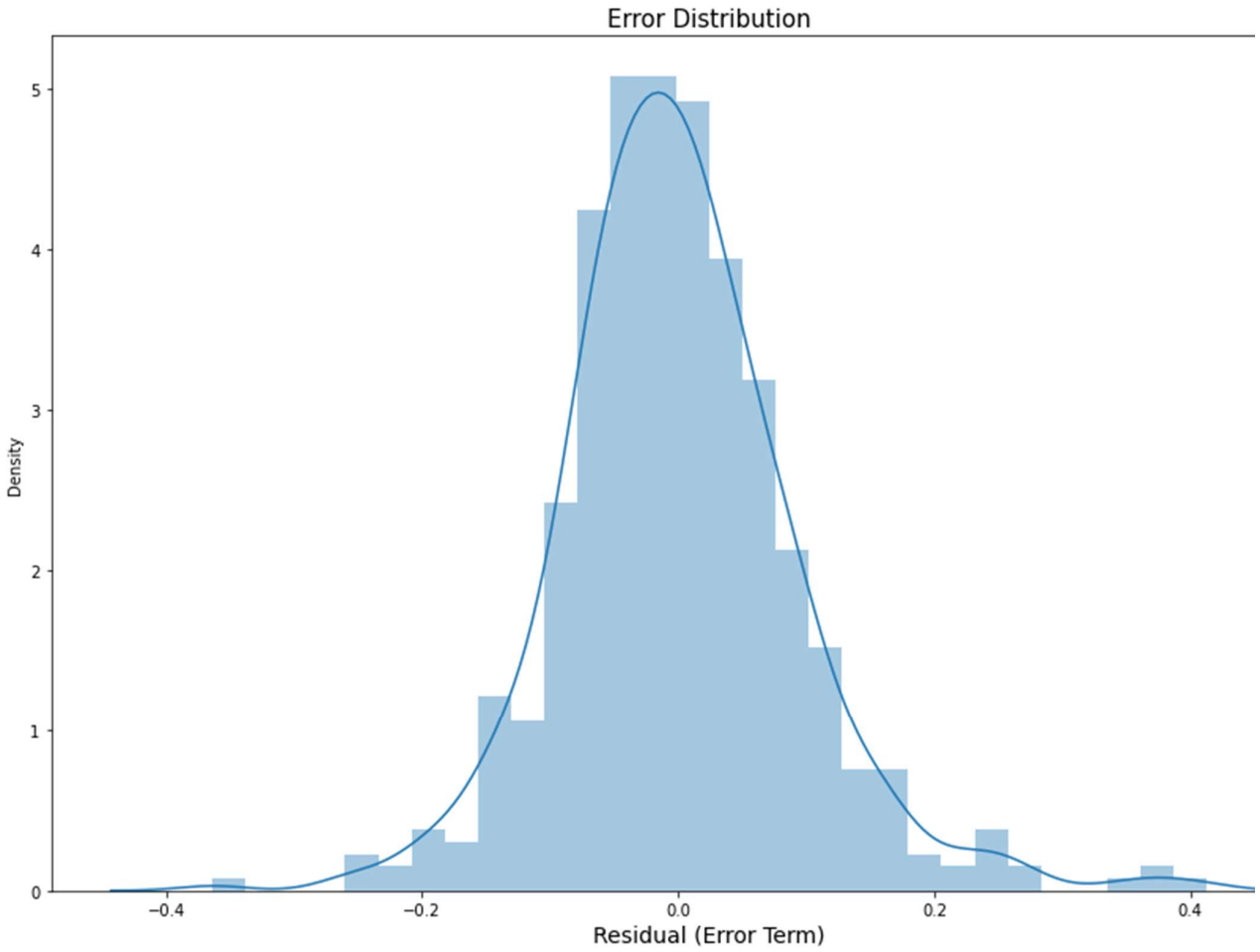
Temp and atemp which represent actual temperature and temperature felt have highest correlation with targe variable. Pls find the plot for reference.


temp
atemp
hum
windspeed
cnt
temp
atemp
hum
windspe

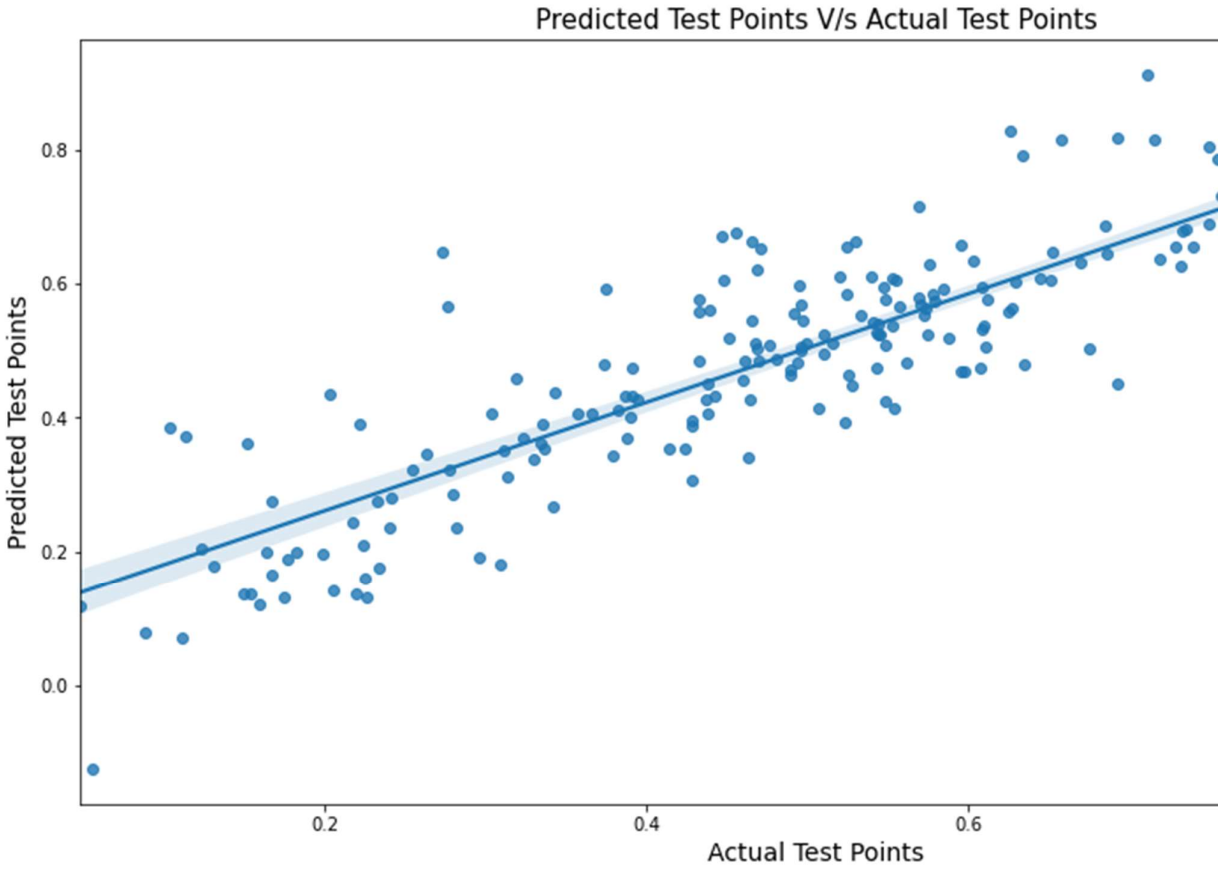
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: Pls find the details below, the ideas behind LR model is satisfied in the model

1. **Error Distribution Is Normallly Distributed Across 0 in our model**



2. **Residual v/s Predicted values**



From the above graph, we see that there is almost no relation between Residual & Predicted Value. This is what we had expected from our model to have no specific pattern. This describes homoscedasticity

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

Answer:

Final model parameters:

const	0.111262
yr	0.232250
workingday	0.023135
temp	0.519176
windspeed	-0.148850
season_2	0.101969
season_4	0.137260
mnth_8	0.054916
mnth_9	0.113041
weathersit_2	-0.080691
weathersit_3	-0.280261

- As per our final Model, the top 3 predictor variables that influences the bike booking are: **temp, yr, season_2**
    - A unit increase in temp (Temperature) variable increase the bike hire numbers by 0.519176 units.
    - A unit increase in yr(Year) variable increase the bike hire numbers by 0.232250 units.
    - A unit increase in season_2(summer) variable increases the bike hire numbers by 0.101969 units.

So it recommended to give maximum consideration to above features to achieve maximum bike rental demand

General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer: Linear regression is a statistical technique to understand the relationship between one dependent variable and several/one independent variables. The objective of regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Linear regression algorithm is machine learning algorithm that is based on supervised learning category.

The assumptions of simple linear regression are

1. Linear relationship between independent and dependent variables
2. Error terms are normally distributed
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

LR algorithm is built using the below concept:

$$y = a + bx$$

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet they have 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 analysing it, and the effect of outliers and other influential observations on statistical properties.

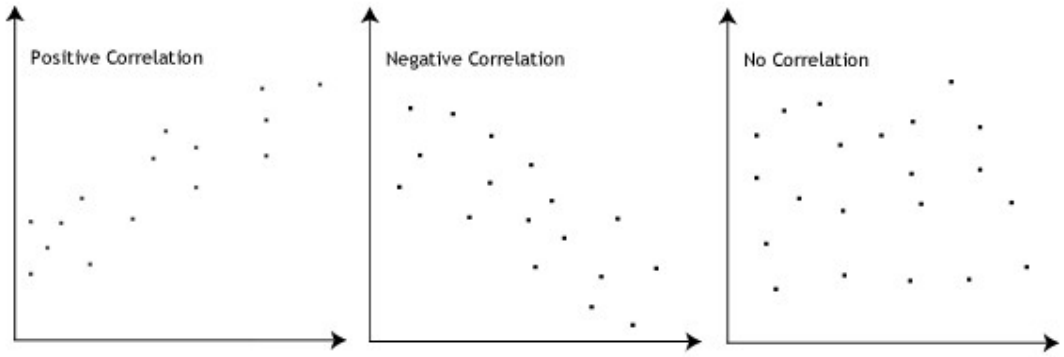
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

3. What is Pearson's R? (3 marks)

Pearson correlation coefficient (PCC), also referred to as Pearson's r is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Formula

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling of variables in model building is one of the pre-processing steps, this helps to speed up the algorithm calculation.

Pls find the difference between Normalized and standard scaling as below

	Normalised scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero me and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
7.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: VIF basically helps explaining the relationship of one independent variable with all the other independent variables.

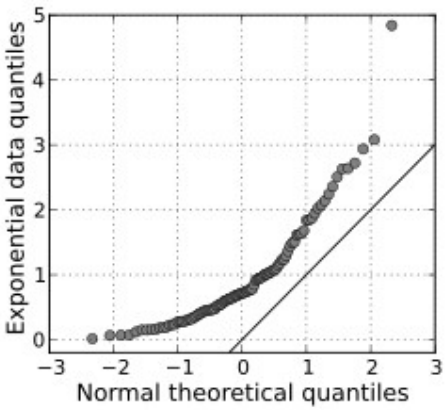
VIF = 1/(1-R2). When VIF is infinity means R2 = 1. This happens when 2 independent variables are perfectly correlated.

This means that specific variable can be explained by linear combination of other variables. Thus, to overcome this, we will need to drop one of the variables which is causing perfect collinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



For Linear distribution of variables considered in linear regression, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots is also used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.