Assignment based Answers:

1. Analysis on Categorical Variables:

- **Season** - One-third of the bike booking were done in fall with a median of over 5200 booking, followed by summer & winter respectively of total booking. Season should definitely an impact on dependent variable.
- **Year** - Bike rentals increased by 75% in the year 2019, this can be due to the awareness of new company being launched and marketing.
- **Weather** - Clear, few clouds, partly cloudy weather is most favorable for bike bookings, harsh weather has a very negative impact on the rentals. Weather can be a good predictor for the dependent variable.
- **Holiday** - Almost 97% of the bike booking were made on a non-holiday which means this variable cannot be considered for prediction.
- **Weekday** - weekday variable has almost uniform distribution of bike rentals with Saturday, Wednesday and Thursday having a median around 4500 bookings. This variable can have some or no influence towards the predictor.
- **Working Day** - Almost 69% of the bike booking were happening in workingday with a median of close to 5000 booking. This implies, workingday can be a good predictor for the dependent variable.
- **Month** - May, Jun, Jul, Aug, Sept, and Oct have the greatest number of bike rentals with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

2. Use of DROP_FIRST in Dummy Variable Creation

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. In the exercise where we had the Blood Group with A, B, AB, and O values in Categorical column, while creating dummy variable if we use drop_first as True then we will need only 3 columns instead of 4 as the 4th one can be expressed as negation of either of one variable.

3. Most correlated variable

Temperature

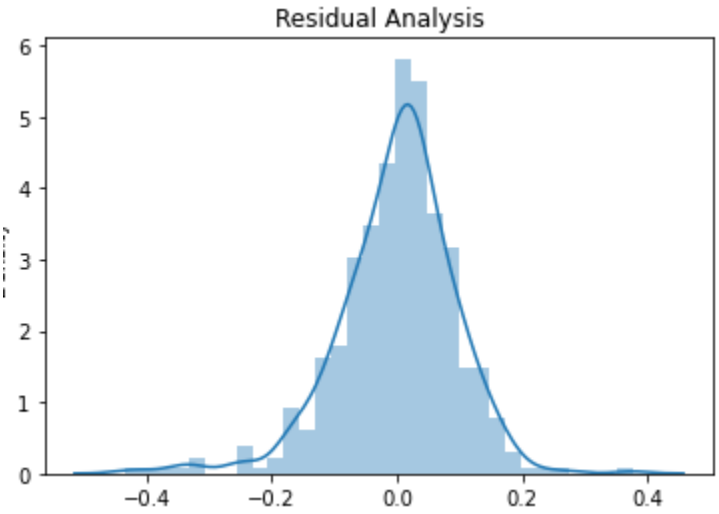
4. Validate Linear Regression Assumptions

There should be a linear and additive relationship between dependent variable and independent variable(s). The model fits a hyperplane with 10 variables and target variable count.

cnt = 0.0902 - 0.0650 * spring + 0.0527 * summer + 0.0970 * winter + 0.0916 * sep + 0.0645 * sat - 0.3041 * lightsnow - 0.0786 * mistcloudy + 0.2334 * yr + 0.0566 * workingday + 0.4914 * temp

Absence of multicollinearity in Independent variables, the VIF values are all less then 5 hence there is no independent variables are not correlated.

Error terms are independent of each other and doesn't follow heteroskedasticity, if it would have been heteroskedastic the plot would exhibit a funnel shape pattern.



The error terms are normally distributed with a mean of 0.

5. Top 3 features contributing to Model

temperature, season, and weather.

General Subjective Answers:

1. Linear Regression

Linear regression is machine learning algorithm used for Supervised learning. Linear regression is used to predict a dependent variable based on given independent variable(s). It attempts to find a linear relationship between target variable and independent variable(s). The motive of the linear regression algorithm is to find the best values for weights and coefficient, which is done through techniques like Cost function. Linear regression can lead to overfitting and should be accompanied with dimensionality reduction techniques (e.g., RFE), regularization techniques and cross validation. Outliers affect linear algorithm badly and this algorithm is oversimplified for most of the complex real-world problems.

2. Pearson's R

Pearson's R (Pearson correlation coefficient) is a measure of linear correlation between two sets of data, this is represented as the ratio between the covariance of two variables and the product of their standard deviation, it's a normalized measurement of the covariance. Value of Pearson's r always lie between -1 and 1, where 1 means for every positive increase in one variable, there is a positive increase of fixed proportion in the other and -1 means that for every positive increase in one variable there is a negative decrease of fixed proportion in the other. it's only used to define linear correlation between two sets of data.

3. Anscombe's Quartet

Anscombe's quartet comprises four data sets that have nearly identical summary statistics (same mean, variance, correlation of variables X and Y, same slope, and intercept of the line) when a least square regression line is fit through them. But when we plot these data points their distribution varies a lot. Anscombe's quartet proved that we should always plot/ graph the data before analyzing it, it also tells us the effect of outliers and other influential statistics properties. – 'Always plot the Data'

4. Use and type of Scaling

Data being collected for training can have different magnitude, units, and range for each variable, while assigning the weights (coefficients) to these variables the algorithm will not know about this diversity and variables with large magnitudes will have a huge significance on the model. To avoid this, scaling is performed to bring all the variables to the same magnitude (scale). Also, in many algorithms, to converge them faster (less iterations) scaling is very important. Normalized scaling is used when we want to bring our values between two values i.e. [0,1] or [-1,1] or any other desired interval. On the other hand, Standardization transforms the data with zero mean and a variance of 1. If we want to retain the information on outliers Standardization should be used over Normalization.

5. Infinite VIF

Infinite VIF indicates that a dependent variable can be exactly derived from the given variables with a linear relationship, this will be a perfect fit line expressed very high collinearity among variables.

6. Q-Q Plot

Q-Q(quantile-quantile) plot is graphical representation of two probability distributions with quantiles against each other. If the two distributions being compared are equal, then the points on the graph will lie on the same line. Q-Q plots are used to find the type of distribution for a random variable. They also help us locate the skewness and kurtosis of a distribution, Purpose in linear regression is to find if two sets of data come from the same distribution.