BIKE SHARING

PROBLEM STATEMENT:

A bike sharing system is a service in which electric bikes are made available for shared use on short term basis for a price. Many bike sharing systems allow people to borrow a bike from anywhere in the city where is it available where the user uses the mobile application to pay the charges and to unlock the bike. The bike can then be return to any specified location in the city.

Here a bike - sharing provider has recently suffered losses in their revenue due to COVID-19. So they want to understand the factors on which the demand for these bikes depends.

ANALYSIS:

- Here we have a data with 730 Rows and we have 16 columns of variables out of which 15 are independent and 1 is dependent.
- Except one column, all other are float or integer type.
- There are no missing/null values in either the columns or rows
- After running duplicate command the shape is same as the original which means no duplicate data is present.

DATA CLEANING:

- Firstly we are checking values_counts for the entire dataframe. It helps us to identify any Unknown/Junk values present in data.

REMOVAL OF REDUNDANT & UNWANTED COLUMNS:

1. Instant: It is only the index values
2. dteday: We remove this because we have 'year' & 'month' data and 'dteday' contains only the date.
3. Casual & registered: Both these columns contains bike's count of bike bookings by different categories of customers but our objective is to find out the total count of bikes and not by specific category, so we will ignore these two columns of data.

CREATING DUMMY VARIABLES:

- We will create dummy variable for 4 categories because it helps us to use a single regression equation to represent multiple groups of data. This means that we don't need to write out separate equation models for each subgroup.
- Like here we have 4 different **seasons** so we create dummy variable for each of the **'season'** and we do same for the '**mnth', 'weekday' & 'weathersit'.**

RESCALING :

- It is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range.
- Scaling/Rescaling of the data helps us to neglect the effect of some dominating column in the data. Like here all values are 1,2,3…etc, but the columns **(temp, atemp, hum, windspeed, cnt)** have some large values so it will dominate the data but rescaling helps us to scale the entire data in same range so the dominating factor removes.

EXPLORATORY DATA ANALYSIS :

- Here the pair plot tells us that there is a linear Relation between **'temp', 'atemp'** and **'cnt'**
- Here we make box plot for 6 categorical variables in the dataset

Season: Around 30% of the bike booking were happening in season_3 with a median of over 5000 bookings. This is followed by season_2 and season_3 with 28% & 27%. This indicates that season can be a good predictor for the dependent variable.

Mnth: Almost 10% of the bookings were happening in the month 5,6,7,8 & 9 with a median of over 4000 bookings per month. This indicates that **mnth** has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Almost 66% of the bike booking were happening during **'weathersit1'** with a median of close to 5000 bookings.

holiday: Almost 97% of bike bookings were happening when it is not a holiday which means this data is clearly biased. This indicates holiday cannot be a good predictor for the dependent variable.

Weekday: weekday variable shows very close trend of around 14% of total bookings, having medians between 4000 to 5000 bookings. It have some or no influence towards the predictors.

Workingday: Almost 69% of the bike booking were happening in 'workingday' with a median close to 5000 booking.

Correlation Matrix:

- The heatmap clearly shows which all variable are multicollinear in nature, and which variable have high collinearity with the target variable.
- Here 'temp' and 'atemp' have the highest multicollinearity, then we can remove either one of them if needed.

Building a Linear Model:

- Here we will Split the data into train-test, then we will apply linear regression on it.
- Then we will build a linear model using 'STATS MODEL'. Then we will check R^2 Value and adjusted R^2 which should be near to 1 and then check for P-Value and VIF and remove those who have high values one by one.
- After removing all these variables our model looks good with a very low multicollinearity.

Final Words:

- Here we got top 3 predictors:

  Temperature: With a value of 56.3%, means with a unit increase in **temperature** variable increase the bike hiring number by 56.3%.

  Weather Situation 3: Here the value is – 30.7%, means with a unit increase in **weather situation 3** variable decrease the bike hiring number by 30.7%.

  Year: Here the value is 23.08%, means with a unit increase in **year** variable increase the bike hiring number by 23.08%.