**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? (3 marks)**

- Year 2019 has the highest number of bike rentals
- Fall has highest average followed by summer
- Months in Q3 has almost same number of bike rentals
- Bike rentals are more in working days compared to holidays

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

It is Important to use frop_first = True mainly due to multicollinearity. Multicollinearity occurs when 2 or more predictor variables are correlated to each other. By using Drop_first = True, We create k-1 dummy variables dropping 1 category.by doing so , the dropping category becomes the baseline against other categories to be compared.

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

The highest Correlation matrix is for ''Registered"category and this is expected as our target variable is sum of Casual and Registered.

```
                temp     atemp       hum  windspeed    casual  registered  \
temp        1.000000  0.991696  0.128565  -0.158186  0.542731    0.539436
atemp       0.991696  1.000000  0.141512  -0.183876  0.543362    0.543678
hum         0.128565  0.141512  1.000000  -0.248506 -0.075211   -0.089212
windspeed  -0.158186 -0.183876 -0.248506   1.000000 -0.167995   -0.217914
casual      0.542731  0.543362 -0.075211  -0.167995  1.000000    0.394137
registered  0.539436  0.543678 -0.089212  -0.217914  0.394137    1.000000
cnt         0.627044  0.630685 -0.098543  -0.235132  0.672123    0.945411

                cnt
temp        0.627044
atemp       0.630685
hum        -0.098543
windspeed  -0.235132
casual      0.672123
registered  0.945411
cnt         1.000000
The variable with the highest correlation with the target variable is: registered
```

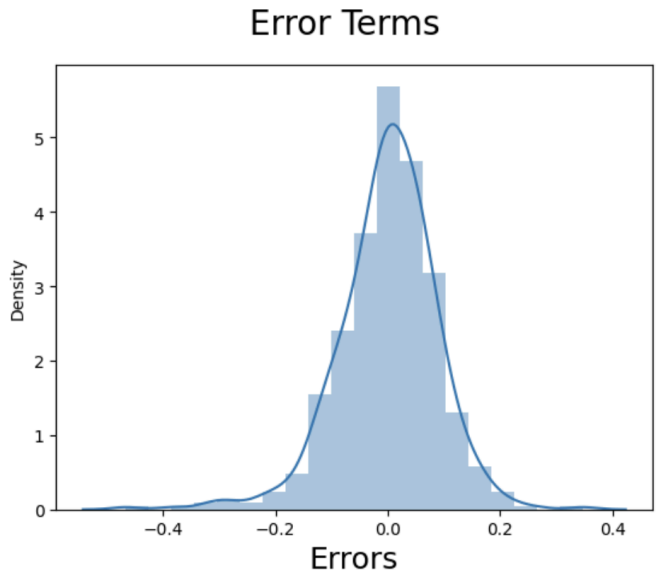
If we remove Casual and Registered, then it is **atemp**

```
                temp     atemp       hum  windspeed    casual  registered  \
temp        1.000000  0.991696  0.128565  -0.158186  0.542731    0.539436
atemp       0.991696  1.000000  0.141512  -0.183876  0.543362    0.543678
hum         0.128565  0.141512  1.000000  -0.248506 -0.075211   -0.089212
windspeed  -0.158186 -0.183876 -0.248506   1.000000 -0.167995   -0.217914
casual      0.542731  0.543362 -0.075211  -0.167995  1.000000    0.394137
registered  0.539436  0.543678 -0.089212  -0.217914  0.394137    1.000000
cnt         0.627044  0.630685 -0.098543  -0.235132  0.672123    0.945411

                cnt
temp        0.627044
atemp       0.630685
hum        -0.098543
windspeed  -0.235132
casual      0.672123
registered  0.945411
cnt         1.000000
The variable with the highest correlation with the target variable is: registered
```

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

- Plotted a histogram to see if the residuals are normally distributed.




- Durbin watson test - Values were close to 2. It means no autocorelations

**Durbin Watson test**

```
from statsmodels.stats.stattools import durbin_watson

dw_test = durbin_watson((y_train - y_train_price))
print(f'Durbin-Watson Test: {dw_test}')
```

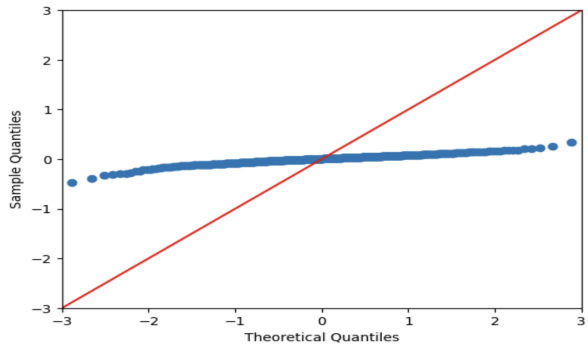
```
Durbin-Watson Test: 2.008166091681963
```

- Q-Q test - Residuals follows a straight line means that it is normally distributed

**Q-Q test**

```
: import scipy.stats as stats

# Q-Q plot
sm.qqplot((y_train - y_train_price), line='45')
plt.show()
```




- Calculated VIF value for each of the variables and made sure that it is within the permissible limit.

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

1. Temp
2. Yr
3. Winter

For these variables p value seems to be around 0 and constants are positive

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 0.2992 | 0.025 | 11.958 | 0.000 | 0.250 | 0.348 |
| **yr** | 0.2350 | 0.008 | 29.150 | 0.000 | 0.219 | 0.251 |
| **holiday** | -0.0989 | 0.026 | -3.836 | 0.000 | -0.149 | -0.048 |
| **temp** | 0.3988 | 0.032 | 12.481 | 0.000 | 0.336 | 0.462 |
| **windspeed** | -0.1539 | 0.025 | -6.200 | 0.000 | -0.203 | -0.105 |
| **spring** | -0.1038 | 0.015 | -6.747 | 0.000 | -0.134 | -0.074 |
| **winter** | 0.0651 | 0.014 | 4.641 | 0.000 | 0.038 | 0.093 |
| **DEC** | -0.0523 | 0.017 | -3.030 | 0.003 | -0.086 | -0.018 |
| **JAN** | -0.0572 | 0.018 | -3.183 | 0.002 | -0.093 | -0.022 |
| **JULY** | -0.0620 | 0.017 | -3.624 | 0.000 | -0.096 | -0.028 |
| **NOV** | -0.0485 | 0.019 | -2.607 | 0.009 | -0.085 | -0.012 |
| **SEP** | 0.0519 | 0.016 | 3.350 | 0.001 | 0.021 | 0.082 |
| **SUN** | -0.0495 | 0.011 | -4.310 | 0.000 | -0.072 | -0.027 |
| **Rain + Scattered clouds** | -0.2996 | 0.024 | -12.314 | 0.000 | -0.347 | -0.252 |
| **Mist + Few clouds, Mist** | -0.0819 | 0.009 | -9.515 | 0.000 | -0.099 | -0.065 |

General Subjective Questions

**1. Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is a supervised learning algorithm used to build relationship between a dependent variable and one or more independent variables.
There are 2 types of Linear regression.

1.**Simple linear regression** - Explains relationship between a dependent variable and independent variable.
Equation -
: $Y = \beta o + \beta_1 X$

2.**Multiple linear regression** - Explains relationship between one dependent variable and one or more independent variable.
Equation -

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Assumptions
Relationship between dependent and independent variable is linear
Variance across the error terms is constant across all levels of independent variable
The error terms are normally distributed
Independent variables are highly correlated with each other.

Once the data is prepared, we will start with the Model.

- We will check if there are columns that have high corelation using heat map and take decision on keeping or dropping it
- First we will create dummy variables for categorical columns and drop the first variable.
- Then join this with the original dataset
- Remove the actual columns where dummy variables are created.
- Once dummy variables are created, We can split the dataset into Train and Test. Usually we give 70 % of the data to train dataset and the other 30 % for testing
- For Numerical variables, We use Min-Max scaler and scale down the variables so that the machine can understand
- Now we can start with the Model building method(OLS)
- First by considering all variables we can try fitting the linear model and check the P value for each variables
- We can also use Recursive Feature elimination method to find the set of variables.
- For this analysis, I have given the number of features as 20.

- Then on the basis of this 20 columns, I validated the P value and VIF for each. Removed all the variables where P value and VIF value.
- Once you have the Final dataset, We can start with the Residual tests to see if the Assumptions are met. I have performed Durbin Watson test, Q-Q test and plotted the histogram to see if the errors are normally distributed
- Once the model is ready, we can transform the numerical variables from the test dataset. Here only transform operation is used.
- After this we can start with the prediction
- Then using R squared method we can perform the model evaluation

```
from sklearn.metrics import r2_score
r2_score(y_test, y_pred)
```

- 
```
0.8194868882771147
```

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

Anscombe's Quartet comprises a set of 4 datasets, having identical descriptive statistical properties in terms of means, variance,R-Squared, Correlations, and linear regression lines but having different representations when we scatter plots on a graph. When plotted the dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strength.

Each dataset has the same mean, same variance for x and y values, same corelation coefficient between x and y, same linear regression line

Here even though the stat properties are the same, when plotted each dataset shows a different set.
Anscombe's quartet explains about the Importance of Graphical Data, Influence of outliers , how shape and distribution of data can affect the results and interpretations of statistical analyses.

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

It is one of the most common ways of measuring a linear correlation. It summarizes the characteristics of a dataset .It measures how close an observation is to the best fit line. It's a number between -1 and 1 and measures the strength and direction of the relationship between 2 variables.
If the value is positive, it means that if one variable changes, the other changes as well.
If the value is negative, it means tha if a variable changes , the other changes in opposite directions.
If its 0, it means no relationship

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

Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. Generally the dataset contains variables of high varying units and range. If scaling is done, it only takes magnitude into account and not units.

Normalized scaling brings all the data in the range of 0 and 1 while standardized scaling replaces the values by their Z score. It brings all the data into a standard normal distribution with mean 0 and standard deviation 1. Normalized scaling are sensitive to outliers and Standardized are less sensitive to outliers.

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

It can happen due to multicollinearity. Multicollinearity happens when independent variables are highly correlated which can cause problems in regression.

An infinite VIF happens when R2 = 1. As per the equation, The denominator becomes zero and it leads to infinite value

I have faced infinite VIF in my dataset because of not removing the Constant column as well.

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

A Q-Q Plot is a plot of the quantiles of the 2 distributions against each other, or a plot based on estimates of the quantiles. It is used to find the type of distribution for a random variable whether it be a Guassian distribution

If the data follows theoretical distribution, it follows a straight line. One of the assumptions in linear regression is the errors are normally distributed. Q-Q plot helps to determine if this holds true.

Below is the observation from Q-Q Plot from our dataset

## Q-Q test

```
:
import scipy.stats as stats

# Q-Q plot
sm.qqplot((y_train - y_train_price), line='45')
plt.show()
```