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?

Below is the equation we finally arrived and we have the below dependent variables.

Some variables -vely impact like holiday, windspeed,spring; some variables are +ve impact like summer, weekday

```
#cnt = 0.2061 + 0.2378 * yr -0.0596 * holiday + 0.0193 * weekday + 0.0193 *
workingday -0.1965 * weathersit
#              + 0.4640 * atemp -0.1498 * windspeed -0.0989 * spring + 0.0282 *
summer + 0.0582 * winter
```

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

We just need n-1 dummy variables instead n – variables since , if all zero the , 1st dummy variable will have 1 or if any of n-1 variables have 1,then 1st variable will have 0

In our project we have season as dummy variable and removed 1 column

```
#print(df.head())
df = df.dropna(axis=1)
df = df.drop(columns=['casual','registered'])
mapping = {1:'spring',2:'summer', 3:'fall', 4:'winter'}
df['season'] = df['season'].map(mapping)
mapping = {0:2018,1:2019}
df['yr'] = df['yr'].map(mapping)
df_dummies = pd.get_dummies(df['season'],drop_first=True).astype(int)

df = pd.concat([df, df_dummies], axis=1)
#print(df.head())
```

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

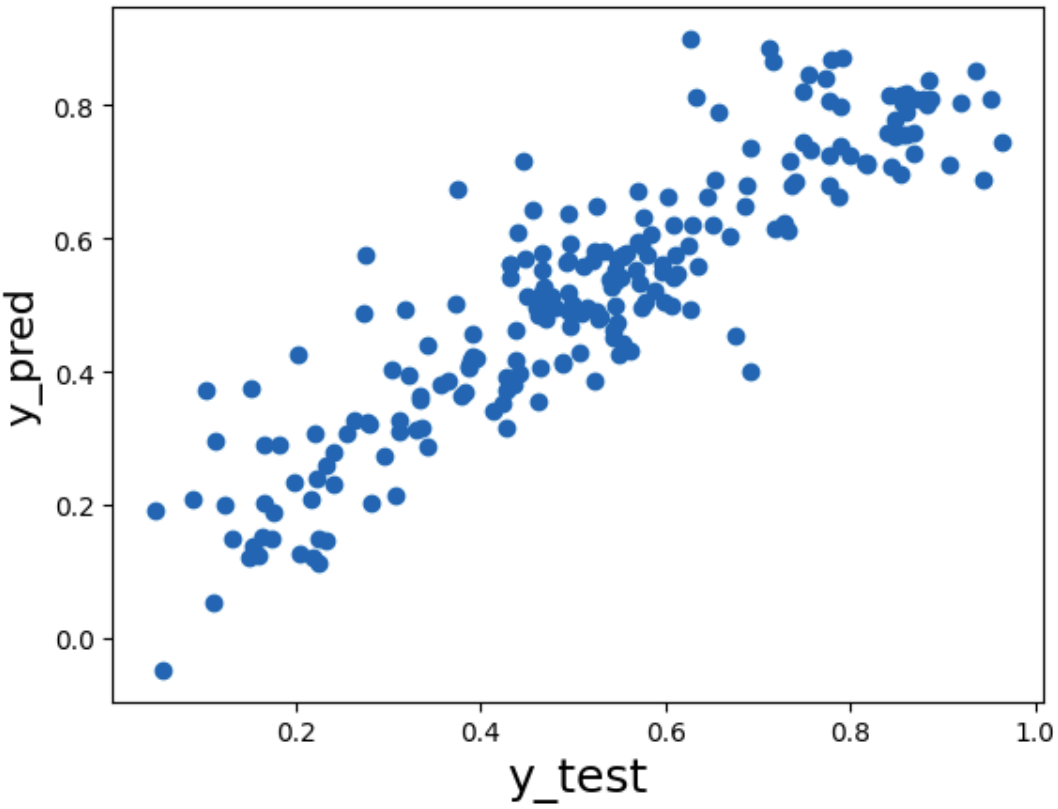
atemp – seems to be more relevant

```
print(correlation_matrix)

temp         0.627044
atemp        0.630685
hum         -0.098543
windspeed   -0.235132
dtype: float64

[ ]:
```

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

Homoscedasticity – with temp,atemp, windspeed.

Linearty -Points and independent variables supports linearty



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

atemp, summer, weekday supports

In [118…

```
print(lr_4.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.821
Model:                            OLS   Adj. R-squared:                  0.817
Method:                 Least Squares   F-statistic:                     228.8
Date:                Wed, 13 Mar 2024   Prob (F-statistic):          3.14e-179
Time:                        22:06:54   Log-Likelihood:                 477.14
No. Observations:                 510   AIC:                            -932.3
Df Residuals:                     499   BIC:                            -885.7
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2061      0.031      6.663      0.000       0.145       0.267
yr             0.2378      0.009     27.724      0.000       0.221       0.255
holiday       -0.0596      0.028     -2.130      0.034      -0.115      -0.005
weekday        0.0550      0.013      4.307      0.000       0.030       0.080
workingday     0.0193      0.009      2.052      0.041       0.001       0.038
weathersit    -0.1965      0.016    -12.453      0.000      -0.227      -0.165
atemp          0.4640      0.035     13.275      0.000       0.395       0.533
windspeed     -0.1498      0.026     -5.732      0.000      -0.201      -0.098
spring        -0.0989      0.020     -4.919      0.000      -0.138      -0.059
summer         0.0282      0.014      2.060      0.040       0.001       0.055
winter         0.0582      0.016      3.604      0.000       0.026       0.090
==============================================================================
Omnibus:                       76.881   Durbin-Watson:                   2.068
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              212.026
Skew:                          -0.736   Prob(JB):                     9.10e-47
Kurtosis:                       5.795   Cond. No.                         19.3
==============================================================================

Notes:
```

General Subjective Questions

Linear regression is a fundamental statistical and machine learning technique used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors)

Assumptions:
Linearity,Independence, Normality,No perfect multicollinearity

Model Representation
For a simple linear regression with one independent variable
Y= c0 + c1*x1 + e

multiple linear regression with 'n' independent variables
Y=co+c1*x1+c2*x2+c3+x3....+cnxn+ x

Ordinary Least Squares (OLS), Mean Squared Error (MSE), Root Mean Squared Error (RMSE)

2.Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a famous example in statistics that demonstrates the importance of visualizing data before drawing conclusions. It consists of four datasets that have nearly identical summary statistics (e.g., mean, variance, correlation) but exhibit vastly different relationships when plotted.

3. **What is Pearson’s R?**

Pearson correlation coefficient, often denoted as Pearson's r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where

1 - +ve linear reln

0 no linear reln

-1 -ve linear reln

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

Scaling is a preprocessing step in data preparation that involves transforming the features of a dataset so that they have a similar scale. This is very crucial step in many machine learning algorithms, especially those that are distance-based or gradient-descent-based since it ensures that all features contribute equally to the model fitting process. Scaling does not change the shape of the distribution of the data; it only changes the scale.

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

It indicates perfect multicorinality between dependent and independent variables.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used for access whether a given set of data follows a specific distribution for comparing it with theoretical distribution such as the normal distribution. In a Q-Q plot, the

quantiles of the sample data are plotted against the quantiles of the theoretical distribution, creating a scatterplot. If the data points fall approximately along a straight line, it suggests that the sample data follows the theoretical distribution.