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)

- More people rent bikes during working days.
- Business is more in 2019 compared to 2018.
- There is more people using boom bikes when the weather is "Clear, Few clouds, partly cloudy, partly cloudy"
- Most of users are in fall season followed by summer and winter

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

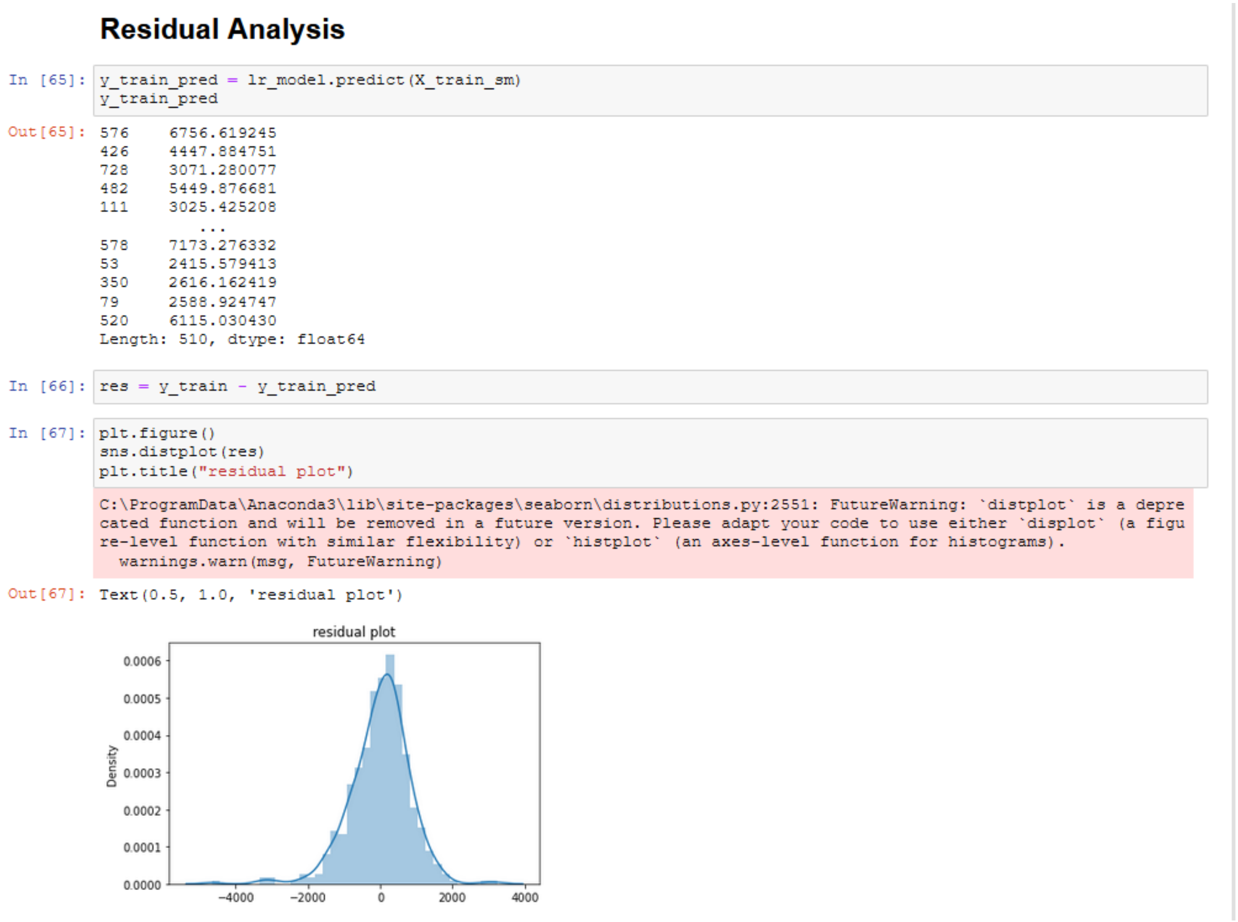
If the first dummy variable of is not dropped then dummy variables created will be redundant and correlated, it is advisable to drop the first dummy variable when the levels in the variables are less for big levels it is okay to not drop the first dummy variable.

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

Temperature in Celsius (temp) and feeling temperature in Celsius (atemp) has highest correlation of 0.63

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

We validate the assumptions of linear regression after building the model on the training set by doing residuals analysis, which concludes residuals are normally distributed. The pictures below the center of distribution is zero and shape is normal distribution.



Residual Analysis

```
In [65]: y_train_pred = lr_model.predict(X_train_sm)
         y_train_pred

Out[65]: 576    6756.619245
         426    4447.884751
         728    3071.280077
         482    5449.876681
         111    3025.425208
                   ...
         578    7173.276332
         53     2415.579413
         350    2616.162419
         79     2588.924747
         520    6115.030430
         Length: 510, dtype: float64

In [66]: res = y_train - y_train_pred

In [67]: plt.figure()
         sns.distplot(res)
         plt.title("residual plot")

         C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a depre
         cated function and will be removed in a future version. Please adapt your code to use either `displot` (a figu
         re-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
           warnings.warn(msg, FutureWarning)

Out[67]: Text(0.5, 1.0, 'residual plot')
```

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

Top 3 features contributing to demand in shared bikes are:

- Weather Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds which has strong negative correlation
- Windspeed also negative correlation
- yr has strong positive correlation

General Subjective Questions

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

Linear regression is a supervised machine-learning algorithm, where there is linear relationship between independent variables(x1,x2 etc) and dependent variable(y).

Based on the data points machines learns and plots a line that models the line best. The line is modelled as below linear equation

y = b0 + b1X1

For multiple linear regression where multiple independent variables are there, the equation is as follows:

y = b0 + b1X1 + b2X2 + ….+ bnXn

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

Anscombe's quartet comprises four data sets with eleven data points that have similar descriptive statistics, yet have very different distributions and appear very different when graphed. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

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

Pearson's correlation coefficient also known as Person's r is a statistic that measures the strength of linear correlation between two variables.

The more inclined the value of Pearson's r towards +1 or – 1, the stronger is the association between two variables.

+1 indicates perfect positive relationship, -1 indicates a perfect negative relationship and 0 indicates no relationship exists.

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

Scaling is bringing all the variables to a common scale. If the scale of the variables is different or larger then the coefficients of the variables also varies accordingly. If the variables are at same scale then the coefficients are also comparable.

Scaling is performed to make the minimization routine faster and much effective.

Normalized scaling converts the data of the variable between 0 and 1 whereas standardized scaling changes the mean of the data to 0 and standard deviation to 1.

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

In case of perfect correlation, R square is 1 leading to VIF infinite value.

To solve this we can drop one of the variable from the dataset, which is causing this perfect multicollinearity.

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

Quantile-Quantile (Q-Q) plot is a graphical tool, which show the quantiles two of sample distribution to determine if the two sets of data come from same distribution.

This helps to confirm the training and test set of data for our linear regression are from populations with same distributions, have common location and scale, have similar distributional shapes (normal, exponential or uniform) and have similar tail behavior.