Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Below inferences were drawn from EDA on Categorical variables:

- People prefer bikes in Fall the most and least in spring
- On Sunday's the bike rentals are comparatively less than other days
- Demand increases between May to October and it falls from November to January
- People prefer bikes the most when weather is clear
- During Holiday's the bike usage falls
- Not much impact on bike usage on weekends but it's low as compared to working days(people might be using bikes to commute to work)

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

The reason to use drop_first=True is because we need only k-1 dummy variable columns, where k is the level of categorical variable. In our dataset season had 4 levels so we created 3 dummy columns for it as the first one can be determined by other 3. It’s simply efficient to drop 1 column during creating dummy variable columns.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

atemp and temp both has high correlation of 0.63 with target variable cnt

---

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

By plotting a histogram for residuals(errors), residual was obtained by substracting actual values of cnt by predicted values of cnt on training data. The histogram plotted followed a normal distribution and it was centered around 0, which meant it has 0 mean, proving that the assumptions made before creating the model are correct.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)
Below are the 3 most significant features:
atemp with coefficient: 0. 4088
yr with coefficient: 0. 2370
season_winter with coefficient: 0. 0901

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression model is part of Supervised learning methods of machine learning it's of 2 types:

Simple Linear Regression: Model with only 1 independent variable.

Multiple Linear Regression: Model with more than 1 independent variable.

It's based upon Equation of a straight line: Y=mX+c

Where m=slope, c=intercept

The Linear regression model is given by y_i=β0+β1X1+ β2X2 +..+ βiXi +e

Where β0=constant coeficient, β.. βi=coeficients of independent variables used for model building, e=error, y_i=target/dependent variable

The steps involved in building linear regression model are as follows:

1. Reading and Understanding the Data
2. Preparing the Data for Model
3. Training the model
4. Residual Analysis
5. Predictions and evaluating model on the test set

We make certain assumptions while creating the model which are part of Residual analysis and these assumptions are:

1.There is a linear relationship between X and y:X and y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.

2.Error terms are normally distributed with mean 0(not X, y):

3.Error terms are independent of each other: The error terms should not be dependent on one another (like in a time-series data wherein the next value is dependent on the previous one).

4.Error terms have constant variance (homoscedasticity):

-The variance should not increase (or decrease) as the error values change.

-Also, the variance should not follow any pattern as the error terms change.

The model is build using Gradient descent, OLS in python

After model building the important parameters to look at are:

R_sqaure: How well the model explain the variance in the target variable

p value: signifies whether the independent variable is significant or now

VIF: To check the multicollinearity

AIC: Lower the better.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Abscombe's quartet is a set of 4 datasets which are identical statistically with respect to mean, variance, R_square and correlations but upon plotting these datasets with scatter plots we can find out that they have very distinct features from each other that can only be observed upon examining them visually through plotting the scatter plot.

Abscombe's quartet explains the drawbacks of just depending upon the summary statistics of the dataset and emphasize on the importance of looking at the data visually.

The 4 datasets are as follows:

1. Dataset 1: Linear relationship between x(independent) and y(dependent): Upon plotting scatter plot between x and y we can observer linear relationship between the 2 variables
2. Dataset 2: Non-Linear Relationship between x(independent) and y(dependent): Upon plotting scatter plot between x and y we can observe the relationship between x and y is non-linear and curved.
3. Dataset 3: Linear relationship between x(independent) and y(dependent) except for outlier:Upon ploting scatter plot and between x and y we can observer that x and y follows linear relationship but there will be an outlier in the data which won't follow the linear relationship
4. Dataset 4 : A high value datapoint point in the dataset: Upon plotting scatter plot between x and y we can observe that all the datapoints will be on one vertical lines but one datapoint with very high values will change the correlation coefficient of complete dataset.

In all above cases the correlation coeficients will be identical and we can only detect that these datasets are having different properties upon plotting the scatter plot between x and y.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R is a correlation coefficient which signifies how strong is the relationship between x(independent) and y(dependent) it's values always lies between -1 to 1.

$r = (\sum (x - \bar{x})(y - \bar{y})) / (\surd(\sum (x - \bar{x})^2)(\sum (y - \bar{y})^2));$
where r=correlation coefficient, $\bar{x}$=mean of x, $\bar{y}$=mean of y

We can understand how strong a linear relatiosnship is as follows:

1>=r>0: Positive correlation between x and y, the higher the value towards 1 the stronger the correlation.Meaning when x increases, y also increases.

R=0: no correlation between x and y, meaning no change in y with respect to change in x

-1<=r<0: negative correlation between x and y, the higher the value towards -1 the stronger the correlation. Meaning when x increases y decreases.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
Scaling: Scaling is part of pre-processing step before building the linear regression model. Scaling is the process of bringing all the features including target variable of the dataset to same scale.
Scaling is performed because our dataset might have features which will be categorical and continuous, we can treat the categorical variables either by dummy variables or by encoding in both cases the values in dataset becomes either 1 or 0. For continuous variables there can be values in the dataset which will have values in millions so their coefficients will be very very low, on contrary the categorical variables will acquire high coefficient values, this might impact our model and significant and non-significant variables will have wrong values for coefficients. Sor this reason scaling is performed.
There are 2 types of scaling that we can perform:
Normalized scaling: it's given by (x-x_min)/(x_max-x_min) and it compresses all the datapoints between 0 and 1.
Standardized scaling:For this we keep mean=0,sigma=1, it centers the data around mean 0.
The difference between the 2 is as follows:

Normalized Scaling (Min-Max Scaling): min max is the preferred one as it handles the outliers.
Standardized Scaling: The advantage of Standardization over the other is that it doesn't compress the data between a particular range as in Min-Max scaling.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF calculates how well one independent variable is explained by all the other independent variables combined and it's formula being:
VIF_i=1/1-(Ri_sqaure)^2
If for one variable in our dataset it was holiday and few more, is being explained by all other variables so good that 100% of it's variance is being explained by other independent variables then we will have high multicollinearity and we will end up having that variable's
VIF=1/0 which will give the VIF of that variable as infinity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Q-Q plot is used to determine whether a dataset follows a theoretical distribution, which is basically normal distribution.
Importance of Q-Q plot:
As we know that while building a linear regression model we make certain assumptions and one

of which is that the residuals(errors) are normally distributed. To validate our linear regression model plotting a Q-Q plot is the most important step that we performe and we check whether the residuals which is the difference between y_true(actual value of y in training dataset) and y_pred(predicted value of y by model) is normally distributed or not.

If we observer that the residuals follows a normal distribution then we can be confident about our model.