Assignment-based Subjective Questions
Q1). From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans 1)
Bike demand in the fall is the highest.

•

Bike demand takes a dip in spring.

•

Bike demand in year 2019 is higher as compared to 2018.

•

Bike demand is high in the months from May to October.

•

Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.

•

The demand of bike is almost similar throughout the weekdays.

•

Bike demand doesn’t change whether day is working day or not.

Q2)Why is it important to use drop_first=True during dummy variable creation?
Ans 2) It is important to use drop_first=True during dummy variable creation to avoid the "dummy variable trap". The dummy variable trap refers to a situation where one or more dummy variables can be perfectly predicted by the other dummy variables. In other words, there is multicollinearity among the dummy variables. This can lead to unstable and unreliable estimates of the regression coefficients, and can also lead to incorrect statistical inferences. By setting drop_first=True, we drop one of the dummy variables for each categorical variable. This removes the perfect multicollinearity among the dummy variables, and allows us to estimate the regression coefficients more accurately. Dropping one of the dummy variables also makes the interpretation of the coefficients more meaningful, as it provides a baseline category against which the other categories can be compared.

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans 3) TEMP HAD THE HIGHEST CORRELATION COEFFICIENT OF 0.63.

Q4). How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: By plotting the residuals distribution. It came out to be a normal distribution with a mean value of 0.

Q5).Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: The Following are the top 3 features contributing significantly towards explaining the demands of the shared bikes:

- atemp (0.412)
- yr (0.236)
- weathersit Light rain (-0.275)

General Subjective Questions

-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x

General Subjective Question

Q1) Explain the linear regression algorithm in detail.

Ans1): A linear regression algorithm tries to explain the relationship between independent and dependent variable using a straight line. It is applicable to numerical variables only. Following steps are performed while doing linear regression:

- The dataset is divided into test and training data
- Train data is divided into features(independent) and target (dependent) datasets
- A linear model is fitted using the training dataset. Internally the api's from python uses gradient descent algorithm to find the coefficients of the best fit line. The gradient descent algorithm works by minimising the cost function. A typical example of cost function is residual sum of squares.
- In case of multiple features, the predicted variable is a hyperplane instead of line.

The predicted variable takes the following form:

$Y= \beta0+\beta1x1+\beta2x2+\beta3x3+\cdots+ \beta nxn$

- The predicted variable is than compared with test data and assumptions are Checked.

Q2). Explain the Anscombe’s quartet in detail.

Ans2): Anscombe’s quartet comprises of four data sets that have nearly identical simple descriptive statistics but have quite different distribution when visualized graphically. The simple statistics consist of mean, sample variance of x and y, correlation coefficient, linear regression line and R-Square value. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. The graphs are shown below:

Image source - https://en.wikipedia.org/wiki/Anscombe%27s_quartet

3. First plot (top left) appears to be simple linear relationship
4. The second plot (top right) is not distributed normally and correlation coefficient is irrelevant as it shows a nonlinear relationship
5. The third plot (bottom left) is linear but has different regression line. This is happening because of the outliers present in the data
6. The fourth plot (bottom right) does not show linear relationship however due to outliers the statistics got adjusted.

In a nutshell, it is a better practice to visualize data and remove outliers before analysing it.

Q3). What is Pearson’s R?

ANS3): PEARSON’S R MEASURES THE STRENGTH OF ASSOCIATION OF TWO VARIABLES. IT IS THE covariance of two variables divided by the product of their standard deviation. It has a value

from +1 to -1.

- A value of 1 means a total positive linear correlation. It means that if one variable increase then other will also increase

- A value of 0 means no correlation

- A value of -1 means a total negative correlation. It means that if one variable

increase then other will decrease

Q4)What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

ANS4) SCALING OF A VARIABLE IS PERFORMED TO KEEP A VARIABLE IN CERTAIN RANGE. SCALING IS A pre-processing step in linear regression analysis. The reason we scale a variable is to make

the computation of gradient descent faster. The step size of gradient descent are generally

low for accuracy, if the data has some small variables (values in the range of 0-1) and some

big variables (values in the range of 0 -1000) than the time taken by the gradient descent

algorithm will be huge.

NORMALISED SCALING STANDARDIZED SCALING

Called min max scaling, scales the variable Values are centred around mean with a unit

such that the range is 0-1 standard deviation

Good for non- gaussian distribution Good for gaussian distribution

Value id bounded between 0 and 1 Value is not bounded

Outliers are also scaled Does not affect outliers

Q5). You might have observed that sometimes the value of VIF is infinite.

Why does this happen?

ANS5): THE FORMULA FOR VIF IS
1 VIFi = 1− *R*2 *i*

Basically, if R square is 1 than VIF becomes infinite. It means that there is perfect correlation

between the features.

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

Ans6): A Q-Q plot is a scatter plot of two sets of quantiles against each other. Its purpose is to check if the two sets of data came from the same distribution. It is a visual check of data. If the data is from same source than the plot will appear as a line.