**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**)

Ans:

* Felling temperature is affecting the number of renting bikes as it is good at good temperature.
* During weathersit is Mist or Light Snow number of bike rented is reduced
* Bikes are demanded well during the whole week.
* Compared to 2018 demand has increased in 2019
* During Spring demand is very low compared to another season

**2. Why is it important to use drop first=True during dummy variable creation?**

**(2 mark)**

Ans:

To avoiding multicollinearity between categorical variables, two dummy columns create the third reference column that can explain the details of 3 dataset as shown below hence n-1 formula is justified

Eg. Summer ‘00’, Winter’01’, Rainy’10’

|  |  |  |
| --- | --- | --- |
| **Season** | Col1 | Col2 |
| Summer | 0 | 0 |
| Winter | 0 | 1 |
| Rainy | 1 | 0 |

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

Ans:

Temp and Atemp as highest correlation

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

Ans:

* Checking R-squared value with target variable getting constant
* Checking Adjusted R-Squared value for target variable constant
* Checking P-value less than 0.05
* Checking VIF less than 5
* Divided data into 2parts, training set and test set and validating both data set to nearby R-squared value

This confirms for the homoscedasticity, linearity of the data. Below is the example tested for Count of rented bikes affected by dependent variable with above points:

CNT = 0.247  **X**  Year **+** 0.254   **X**   Season Summer **+** 0.314  **X**  Season\_Fall **+** 0.227  **X**  Season Winter **-**0.093  **X**  Holiday **-**0.176  **X**  WindSpeed **-**0.087  **X**  Weathersit 2

**-** 0.291 **X** Weathersit 3

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

Ans:

* During season fall, one day increment may affect count by 0.314 units.
* During summer season, one day increment may affect count by 0.254 units.
* Every Year demand is increasing by 0.247

======================================================================

**General Subjective Questions**

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

* Ans:
* The line drawn among two points is denoted by Y= m\*X +c
* So to draw a bestfit line from multiple points with the equal distance is denoted by slope βn and the predicted points to get the intercept for straight line β0
* The Simple linear regression formula becomes y=β0​+β1​X for one independent and one dependent variable
* The multilinear regression formula becomes y=β0​+β1​X1+β2​X2+………βn​Xn for for more independent and one dependent variable with a constant value that rectifies the predicted data on best fit line.
* This dat points are confirmed using R – squared value = (Yi-Ypred)^2
* For R-squared =1 is best fitted.

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

**3. What is Pearson’s R? (3 marks)**

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

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

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