Assignment-based Subjective Questions

- **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks) .**

Answer: Based on analysis season ,month ,year, weathersit and weekday impact the count of cycles registered.

Effect:

1. The bike demand was higher in 2019 than in 2018
2. Bike demand didn't have significant difference based on weekday of week
3. The demand for bike was highest in fall season (Sept and Oct)
4. The demand for bikes was higher on clear few days.

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

Answer : drop_first=True during dummy variable creation deletes extra column created and reduces corelation

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

Answer: The target variable has highest corelation with temperature.

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

Answer:

1. I did residual analysis by plotting the error terms to check for normal distribution
2. I dropped variables based on VIF and p-value to avoid overfitting problem.

- **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes**

The top 3 variables contributing to the demand are year, temperature and winter.

General Subjective Questions

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

Answer: Linear regression is a machine learning which analyses relationship between dependant and given set of independent variables. It is represent by Y=mx+c where y is the dependant vaiable , x is the independent variable, m is the slope of regression line and c is the constant.

2 types of linear regression : simple linear regression and multiple linear regression. Simple linear regression is used when you have a single independent variable (predictor variable) that predicts the value of a dependent variable.It is represented by Y=m0+m1x+c where m0 is y intertcept , m1 is model parameter,c is error coefficient. Multiple linear regression extends simple linear regression to include more than one independent variable. It is used when there are two or more independent variables that predict the value of a dependent variable. It is represented by Y=m0+m1x+m2x2+m3x3+c where m0 is y intertcept , m1 is model parameter,c is error coefficient, x1,x2 and x3 are independent variable

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

Anscombe's quartet include four data sets that have identical statiscal summaries such as means, variances, and correlation coefficients—the datasets have very different distributions and patterns when visualized. It helps to understand limitations of summary statistics and to determine effect of outliers and other observations.

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

Pearson's R is used to measure linear corelation ship between two variables.

It ranges from -1 to +1. If Pearson's R is a positive number, it means that as one variable increases, the other variable tends to increase as well. If Pearson's R is a negative number, it means that as one variable increases, the other variable tends to decrease If Pearson's R is close to zero, it means there's little to no linear relationship between the two variables.

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

Scaling is process of changing data to a common scale without distorting differences in the range of values. It is done to

1. prevent feature dominance: When features have different scales , some features may lead to biased models dominating others.
2. Improve model performance : Machine learning algorithms, such as gradient descent perform better when features are on a similar scale.

Normalized Scaling

Min-Max Scaling, rescales the data to a fixed range, usually [0, 1] or [-1, 1]. It is particularly useful when all features contribute equally within a range.

Standardized scaling

Standardized Scaling, rescales the data to have a mean of 0 and a standard deviation of 1. It is particularly useful when data has varying range/ distributed data.

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

When R2= 1, since VIF=1/(1-R2). R2 =1 shows perfect corelation between independent variables. To solve this we need to drop the variable to reduce multi collinearity.

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

Q-Q plot is scatterplot created by plotting two sets of quantiles against one another.It is used to assess if a dataset follows a particular theoretical distribution. In linear regression it is used to check if the residuals follow a normal distribution.

a QQ plot can be used to decide if train and test data came from same population. If both sets came from same distribution , the plotted points will form a roughly straight line.