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)

    - Using Boxplot of season vs cnt , we found bike demand is highest in fall and lowest in spring
    - Using Boxplot of year vs cnt, we found that Year 2019 has higher demand for bikes as compared to 2018. It means demand for bikes is increasing
    - Using Boxplot of weathersit vs cnt we found , whenever the weather is clear or partly cloudy bike demand is highest whereas demand is lowest in snow and rainy weather
    - Using Boxplot of holiday vs cnt we found that bike demand is highest when it is holiday
    - Bike demand is not affected by the weekday which means it remains same

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

    - When we create dummy variables using pd.get_dummies() and pass a categorical column in it , we get as many dummy variables as many categories we have.
    - Consider a categorical column with 3 categories.
    - cat_1 will be denoted by 1 when cat_2,cat_3 will be 0
    - cat_2 will be denoted by 1 when cat_1,cat_3 will be 0
    - cat_3 will be denoted by 1 when cat_1,cat_2 will be 0
    - Thus we can see a pattern and say that when cat_2 and cat_3 are 0, at that time cat_1 will be always 1
    - Thus for the C number of categories we need C-1 columns only.
    - Also if we consider all C columns , it will increase multicollinearity which will be bad for model

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

    - From the heatmap built using numerical variables we can see variables temp and atemp have the highest positive correlation with target variable cnt of 0.63 .

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

    - After building a model on a training set, we assume that dependent feature and independent feature vary linearly
    - Absence of Multicollinearity
        - Variance Inflation Factor was calculated and VIF score of variables less than 5 were only considered
        - Out of 15 features selected 10 were shortlisted with VIF less than 5
    - Residual Analysis
        - We plot distplot from seaborn library of error terms
        - Error terms are considered to follow normal distribution with mean equal to zero
    - Linear Relationship
        - We built pairwise scatterplot to help in validating the linearity assumption
        - We plot regplot from seaborn library to find best fit line of our model

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

    - Since we used MinMax Scaling and everything is scaled between 0 and 1
    - To check top 3 features contributing significantly towards explaining the demand we will check features who have highest coefficient value
    - Feature 1 = temp denotes temperature noted on a particular day
        - coefficient = 4906.61
        - Correlation between temp and cnt is 0.66
        - Denotes as temp increases bike demand increases
    - Feature 2 = yr denotes year
        - coefficient = 2035.78
        - Correlation between yr and cnt is 0.59
        - Denotes demand increases as year increases
    - Feature 3 = weathersit_3 denotes light snow,light rain or scattered clouds
        - coefficient = -2259.61
        - Correlation between weathersit_3 and cnt is -0.21
        - Denotes demand decreases in bad weather

General Subjective Questions

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

   - Regression tells how dependent variable changes as independent variable changes
   - It is a statistical model that attempts to show relationship between two variables with the linear equation
   - It is supervised learning machine learning algorithm
   - It helps in finding best fit line between independent and dependent variables such that the error difference between predicted value and true value is as less as possible
   - To minimize errors we must minimize cost function using Gradient Descent
   - It is used either to Forecast Demand or Predict a future value based on conditions
   - Although Linear Regression is
     - Easier to implement and interpret the model
     - Prone to noise and overfitting
     - Quite sensitive to outliers
     - Prone to multicollinearity
   - Two Types
   - Simple Linear Regression
     - Denoted by equation : y = B1x + B0
     - y=Dependent Variable
     - x=Independent Variable
     - B0=Intercept of the line
     - B1=Linear Regression Coefficient
   - Multiple Linear Regression
     - Denoted by equation
     - Y=B0 + B1x1 +B2x2 + B3x3 ….

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

   - Francis Anscombe developed Anscombe’s quartet who was statistician
   - Take 4 datasets with same descriptive statistics that is mean,variance and correlation is identical for (x,y) across groups
   - Each dataset contains eleven (x,y) pairs
   - When we plot these 4 datasets, four scenarios can occur
     - Plot 1 has clean well fitting linear model
     - Plot 2 is not distributed normally
     - Plot 3 distribution is linear but regression calculated is thrown of by wrong data
     - Plot 4 has a single outlier enough to produce a high correlation coefficient

    - Thus, Anscombe’s Quartet makes Data visualization important step and reminds of outliers in data

3. What is Pearson’s R? (3 marks)

    - Karl Pearson made a correlation coefficient known as Pearson’s R.
    - Correlation tells how strongly are two variables dependent on each other
    - -1 denotes highest negative correlation
    - +1 denotes highest positive correlation
    - 0 denotes no correlation
    - Formulae for Pearson correlation :
        - Covariance of two variables divided by product of standard deviations

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

    - We have encoded categorical variables as 0 or 1 but our numerical variables have values different from 0 or 1.
    - Scaling is performed on numerical variables so that all variables have values on same scale.This helps in identifying significant features based on coefficient value
    - Normalization(min-max scaling) scales data between 0 and 1 using formulae : (x - xmin)/(xmax - xmin)
    - Standardization is performed using : (x-mean)/sigma. Data is not scaled to particular range.It is scaled such that data has mean 0 and standard deviation 1

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

    - If there is no correlation between all the independent variables then VIF = 1.0.
    - If there is a complete correlation, then VIF = infinity.
    - Thus the Variance Inflation Factor or VIF tells how much the feature variables are correlated with each other.
    - If VIF is 9, this means that the variance coefficient of the model is increased by 9 factors due to the presence of multicollinearity.
    - This would mean that the normal error of this coefficient is magnified by 3 factor (square root of the difference deviation standard).
    - Formula for VIF : (1 / (1 - R2))
    - Here, ‘i’ refers to the variability of ith.If the square root of R is equal to 1 then the denominator of the above formula becomes 0 and the whole number becomes infinite.
    - It means a complete correlation of variables.

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

    - Another way to test the distribution of continuous features graphically is with a Q-Q structure.
    - A Quantile-Quantile plot or Q-Q plot is a scattering structure created by arranging 2 different quantiles against each other.
    - The first quantile is the variable that you test for hypothesis and the second is the actual distribution by which you test.
    - If the two quantiles of the sample are in the same distribution, they should almost fall into a straight line.
    - Use of Q-Q plot in Linear Regression:
        - To check if the dataset lies on a straight line, a Q-Q plot is used.
        - If they don’t form a straight line, it means residuals and errors are not Guassian(Normal)
    - Importance of Q-Q plot:
        - Sample sizes can be unequal
        - Aspects such as shifts in location, shifts in scale and changes in symmetry can be simultaneously tested.
        - Q-q plot can provide more insight than analytical methods