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)

    - Count of rental bike users have increased in 2019 compared to 2018. Business is expanding.
    - More number of people rent bikes during 'fall' followed by 'summer' and 'winter'. During 'spring', the count is low. Season plays a major role in predicting the dependent variable.
    - During clear weather a greater number of people rent bikes compared to cloudy weather conditions. During moderate rainfall, the rental count is very less. Weather is important factor impacting the dependent variable.
    - Average number of users are more during weekdays in comparison to holidays/weekends.

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

    - During dummy variable creation for categorical variable having n-values, n variables are created. But it requires only n-1 variables to define the categorical variable.
    - So, using 'drop_first = True' ensures additional column is dropped.

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

    - 'temp' & 'atemp' has the highest correlation (0.63) with the target variable.

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

    - X & Y have linear relationship
        - Main predictor variables (temp, weather, and year) have linear relationship with the target variable count of users.
    - Error term is normally distributed with mean zero
        - The residual analysis done by plotting the (actual – predicted) values. The error terms are normally distributed with mean close to zero
    - Error term has constant variance
        - Fitted value against residual error is checked to confirm homoscedasticity
    - Error term is independent of each other (autocorrelation check done)

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

    - 'Temp' has the highest co-efficient value of 0.57. This is aligned with the strong correlation between 'cnt' & 'temp'. When temp is more, demand is high
    - Next to 'temp', we have weather (moderate rains) with coefficient of -0.25. So, during rains there is negative impact on the demand.
    - Year with coefficient of 0.23 is next feature which affects the demand.

General Subjective Questions:

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

Linear Regression is supervised ML in which output is continuous and has a constant slope. It shows the linear relationship between the independent & dependent variables.
It is used in forecasting, predicting the output based on the past data. E.g., score of students
Types:

- Simple Linear Regression (has one independent variable & dependent variable)
  $y = \beta0 + \beta1*x$ ($\beta0$ – intercept, $\beta1$ – slope/co-efficient)
- Multiple Linear Regression (has multiple independent variables & one dependent variable)
  $y = \beta0 + \beta1*x1 + \beta2*x2 + \ldots \beta n*xn$ ($\beta0$ – intercept, $\beta1$, $\beta2$ …βn – coefficients of independent variable x1, x2… xn)

It uses Ordinary Least Square method to reduce the error terms

- Residual (ei) = Actual value (Yi) – Predicted variable (Y-pred)
- Residual sum of squares (RSS) – sum of squares of residuals.
- Total sum of squares (TSS) – sum of squares of distance of actual point from the mean (Yi – Y-mean)
- R square = 1 – (RSS/TSS)

Assumptions:

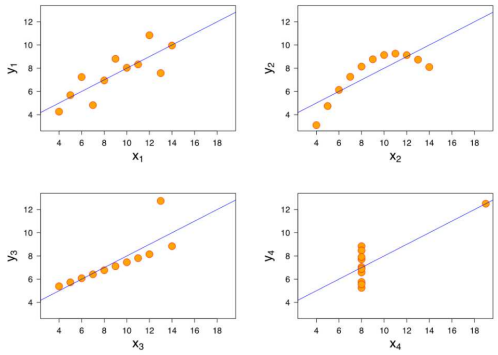
- Independent variable X & dependent variable Y have linear relationship
- Error term is normally distributed with mean zero
- Error term has constant variance
- Error term is independent of each other

Statistics:

- Significance of the co-efficient is defined by the lower p-value.
- R square shows the variance explained by the model.
- Fit of the model is explained using F-statistics & P(F-statistic)
- Adjusted R-square penalizes the model for considering additional variables. Used in comparing the models.

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

- Anscombe's quartet has four data sets which have almost identical descriptive statistics but when plotted as graph shows different distributions.
- It was constructed to illustrate the importance of visualizing the data before analyzing the statistical measures



- Above data sets have same statistical measures but we can see the linear relationship only in the 2nd and 3rd quadrant.
- Linear Regression model can be applied only if the independent and dependent variable have linear relationship. The summary statistics measures indicating the mean & variance may not explain the linear relationship. It is important to graphically validate the linear relationship.

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

- Also known as Pearson's Correlation Coefficient (PCC) is measure of strength of linear relationship between two data sets.
- It attempts to draw bets fit line through two data sets. It does not represent the slope of the best fit.
    - ρ (x,y) = covariance of (X,Y)/ SD of X* SD of Y
- It lies between -1 to 1 (1 - strong positive relationship; -1 - strong negative relationship; 0 – no relationship)
- It cannot capture non-linear relationships and cannot differentiate dependent and independent variables.
- It is not affected by the units of measurement (X & Y can have different units of measure)
- It is also known as Pearson's Product Moment Correlation Coefficient (PPMCC) or bivariate correlation

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

<u>What?</u>

Scaling is the normalization of the independent variable. It is used in data preprocessing step to normalize the data within a particular range

It affects only the coefficients and has no impact on the p-values and accuracy.

<u>Why?</u>

The variables in data set may differ in wide range in magnitudes and units. The algorithm considers only the magnitude and not the unit. So, it is essentially to normalize the data within the same range for reliable model calculations.

It helps is easy understanding/comparison of the data and also increases the speed of computation of the model

Types:

- Normalization or Min-Max scaling brings the data in the range within 0 to 1
  - x = (x - min(x)) / (max(x) – min(x))
- Standardization brings the data into standard normal distribution with mean zero and standard deviation 1
  - x = (x - mean(x))/ S.D. of x
- Normalization loses some information on the outliers as all the data is tried to fit in the range of 0 to 1.

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

- If there is a perfect correlation between two independent variables (R square =1), then VIF will be infinity,
  - VIF = 1/(1-R2)
- If there is perfect correlation between two variables, it will lead to multicollinearity which will affect the model. So, it is necessary to drop one of the variable to get accurate results.
- An infinite VIF also means that the respective variable is expressed as linear combination of another variable.

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

- Q-Q (quantile – Quantile) plot is a plot of quantiles of first dataset against the quantiles of second dataset
- Quantile is the fraction of points below the given value. E.g., 0.4 quantile means 20% of data fall below the given value and 60% above the given value
- A 45-degree reference line is plotted. If two data sets are from the population with same distribution, then points fall along the reference line (y=x). Greater the distance of the point from the reference line higher the evidence suggesting the data sets from different distribution
- If distributions are linearly related, then points lie on the same line but not necessarily on the line y=x
- Sample sizes need not be equal & many distributional functions like shape, location, scale and skewness can be simultaneously tested