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)**
   Answer:
   Below are the points that I can infer after analyzing the categorical variables (season, year, month, holiday, weekday, workingday and weathersit) using boxplot:
   - **Season**: Bike demands is highest during Fall
   - **Year**: There is a significant increase in bike demand during 2019 as compared to 2018, which represents the increased popularity of bike sharing
   - **Month**: Highest bike demand is high in May, June, July, August, September and October.
   - **Holiday**: Bike demand is more during holidays compared to non-holidays. This implies people want to spend time during non-holidays with family and friends
   - **Weekday**: Bike demand is consistent throughout the days, implying regular usage during the week.
   - **Working day:** Bike demand is almost equal either on working day or non-working day
   - **Weathersit**: Clear weather drew more booking due to advantageous condition.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Answer:
   drop_first = True: It helps to reduce the extra column created during creation of dummy variable. Hence, it reduces the correlations created among dummy variables.
   Syntax -
   drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

   Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Answer:
   The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   Answer:
   Linearity of relationship between response and predictor variables.
   Normality of the error distribution (Normal distribution of error terms).
   Constant variance of the errors or Homoscedasticity.
   Less Multi-collinearity between features (Low VIF).

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Answer:
   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –
   Temperature
   Year
   Season

General Subjective Questions

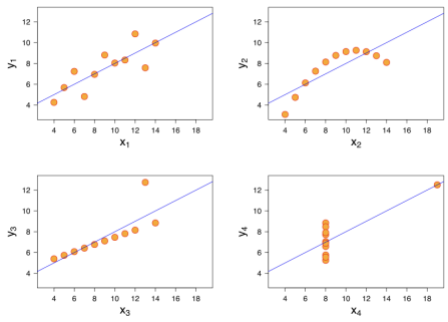
1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer:
   Linear regression is a form of predictive analysis modeling which define the relationship between the dependent (target variable) and independent variables (predictors). Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

   Mathematically the relationship can be represented with the help of following equation −
   Y = mX + c
   Here, Y is the dependent variable we are trying to predict.
   X is the independent variable we are using to make predictions.
   m is the slope of the regression line which represents the effect X has on Y
   c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

   Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

   A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error.
   In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   Answer:
   Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing, it with statistical properties.
   It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these datasets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representations. Each graph plot shows the different behavior irrespective of statistical analysis.
   However, the statistical analysis of these four datasets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represents a different behavior.



   Data Set 1: fits the linear regression model pretty well.
   Data Set 2: cannot fit the linear regression model because the data is non-linear.
   Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model

3. **What is Pearson's R? (3 marks)**
Answer:
In Statistics, The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:
    - Pearson's r
    - Bivariate correlation
    - Pearson product-moment correlation coefficient (PPMCC)
    - The correlation coefficient

    It is a statistic that measures the linear correlation between two variables. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	**Correlation type**
Between 0 and 1	Positive correlation
0	No correlation
Between 0 and –1	Negative correlation

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
Answer:
Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling. Difference between Normalizing Scaling and Standardize Scaling:

S.NO.	**Normalized scaling**	**Standardized scaling**
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
Answer:
VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:
A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform. distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression:
In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also.
- Many distributional aspects like shifts in location, shifts in scale, changes in
- symmetry, and the presence of outliers can all be detected from this plot.

Q-Q plot use on two datasets to check.

- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- *If both datasets have tail behavior*