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: If we take a look at the correlation coefficients of the feature variable w.r.t target variables obtained from the analysis as shown below:

	Correlation_Coef
temp	0.65
yr	0.59
mnth_January	-0.38
windspeed	-0.25
mnth_August	0.23
weathersit_light	-0.23
mnth_September	0.20
weathersit_mist	-0.17
season_summer	0.14
weekday_sunday	-0.06
season_winter	0.03

We find out that the count(target variable) is negatively correlated to the categorical variables like 'mnth_January', 'mnth_August' etc. and positively and strongly correlated to the categorical variable 'yr' with a value of 0.59.

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

Ans: It is important to use drop_first=True because it ensures the prohibition of multicollinearity by removing the highly correlated dummy columns.

Example: In Bike sharing assignment, the column 'yr' had 2 values, '0' and '1', if we are creating dummy variables and not dropping the column harboring any one of above values, this would result in the 2 dummy columns having 100% correlation.

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

Ans: The column 'temp' has the highest correlation with 'cnt'(target) variable.

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

   Ans: I performed following checks to validate the assumptions of Linear Regression:

   - I checked whether the Residual of Error terms is **Normally Distributed.**
   - I checked and found out that Residuals and Predicted values **are independent of each other.**
   - I checked the **Homoscedasticity of Predicted and Actual target values.**

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

   Ans: Top 3 features are:

   - Temperature
   - Year
   - Month of January

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans: Linear regression algorithm is a Supervised machine learing algorithm, which involves a training model to predict the behaviour of data on the basis of some variables. Linear regression implies a linear relationship between the target and predictor variables.

Mathematical representation of Linear regression looks like the following equation:

$$y = m1.X + c$$

Where y = target variable, X is predictor variable and c is a constant

Procedure of Linear regression algorithm:

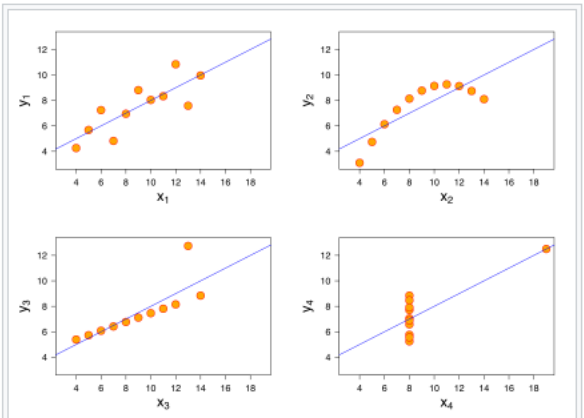
- Data cleaning
- Creating dummy variables for Categorical columns
- Scaling the numerical columns with techniques like MinMaxScaling etc.
- Removing highly correlated columns
- Selection of relevant independent columns using RFE or manual procedures
- Conducting an iterative model building procedure on the basis of significance of the selected independent variables and pruning the ones having least significance.
- Checking the Linear regression assumptions like

- Residual distribution should be normal
- No relationship between Predicted values and Residuals
- Should satisfy Homoscedasticity

- Prediction on test dataset
- Checking the significance by calculating r2_score for test dataset predicted values.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four data sets having nearly same simple descriptive statistical attributes but have stark differences in graphical representation. Below is the example as per Wikipedia.org



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

As one can observe that, the mean, variance are exactly same, and similarly all other stats are almost identical but the graphical distribution is totally different.

3. What is Pearson's R?

Ans: Pearson's R is a bivariate correlation, a measure of linear correlation between two sets of data. It is defined as covariance of two numerical variables, divided by the product of their standard deviations, implying that it essentially is a normalized covariance.

Formula for Population:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where:

cov is the covariance
σ_X is the standard deviation of X
σ_Y is the standard deviation of Y

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

Ans: Scaling is an important step in data pre-processing which is applied to the independent/predictor variables to normalize the data within a particular range. It helps speed up the algorithm.

It is performed to eliminate the redundancies present in data due to wild variations in magnitudes, units and range of different variables. If we don't perform scaling, model spits out incorrect results based on the wide variation in magnitude of un-scaled data. We should also note that scaling just affects the coefficients of independent variables and not the other parameters like t-statistic, f-statistic, p-values etc.

In normalized scaling, all the data values are brought in the range of 0 and 1

In standardized scaling, data values are replaced by their respective z-scores.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
Ans: Infinite value of VIF implies a perfect 100% correlation between two independent variables. It implies one variable can be expressed exactly by a linear combination of other variables.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
Ans: Quantile-Quantile plots help us identify if a dataset is normally distributed. Inferences one can obtain from Q-Q plot:
   - It tells us whether the steps in our data are too big or too small
   - A high slope in Q-Q plot implies observations are far spread out which might imply a presence of high number of outliers
   - A flat Q-Q plot means our data is clustered around a very small range as compared to the distribution of a normally distributed dataset.