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 : Categorical variables are firstly to be converted to dummy variables for ease. The categorical variables, as in our assignment dataset has shown impact on variations in p-values and VIFs as weel.

For example : With high p-value for 'weekday_5', we eliminate it and thus shows changes in VIF of other variables from inf value to numeric. We keep on improving the model, by removing the further dummies for 'weekday4/3/2/1'. But the removal of these variables maintains the Homoscedasticity of the model, by having very low variance of data. The R square and Adjusted-R square value does not fluctuate with these changes. Thus, giving a decent model for further evaluations.

---

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

Answer: The drop_first=True during dummy variable creation is to eliminate the redundant variables. For example here, there are four seasons, but all four do not need dummy variables to be recognized. So, only 3 season dummy variables are kept instead, as third season can be identified, with just other two season variables.

---

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

Answer : 'temp' and 'atemp' variables have highesht correlation with target variable.

---

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

Answer:

1. The relation between the dependent and independent variables should be almost linear.
2. The data is homoscedastic, meaning the variance between the results should not be too much.
3. The results obtained from an observation should not be influenced by the results obtained from the previous observation.
4. The residuals should be normally distributed. This assumption means that the probability density function of the residual values is normally distributed at each independent value.

---

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 : The top 3 features contributing significantly towards explaining the demand of the shared bikes are : 'weathersit_3', 'mnth_9' and 'holiday'

****

General Subjective Questions

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

Answer : While building a Linear Regression model, you assume that the target variable and input variables are linearly dependent.

Assumptions of Linear Regression: The assumptions allow us to make inferences.

- Linear relationship between X and y.
- Error terms are normally distributed
- Error terms are independent of each other.
- Error terms have constant variance i.e. termed as homoscedasticity

Linear regression has to be based on following steps:

Step 1: Reading and understanding the data

- Perform EDA and check the correlation between the variables

Step 2: Data preparation

- Creation of dummy variables
- Divide into train-test split
- Perform scaling
- Divide the data into X and y.

Step 3: Training the model

- Create Linear regression model using mixed approach:
    a. Manual approach
    b. RFE or automated approach

Step 4: Residual analysis

- Perform residual analysis of the error terms

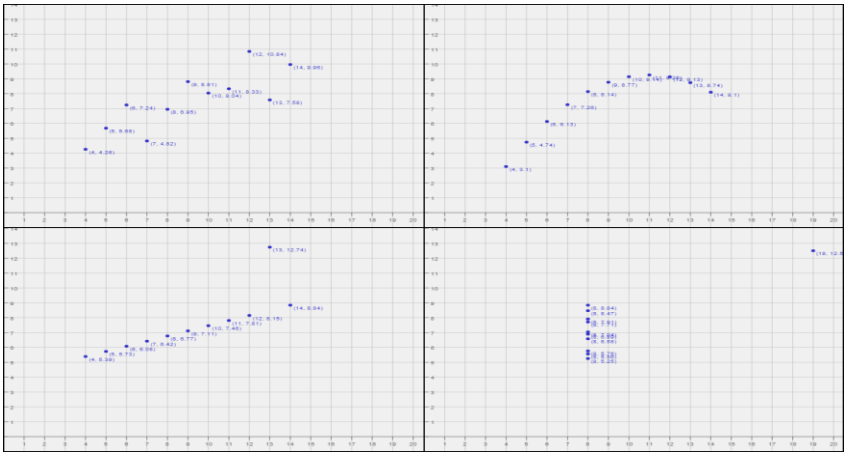
Step 5 : Predictions and evaluations

- Evaluate the model based on the predictions.

---

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

Answer: Anscombe's Quartet is a great demonstration of the importance of graphing data to analyze it. Given simply variance values, means, and even linear regressions can not accurately portray data in its native form. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.



- Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. Think about it: if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict.
- How to analyze your data? For example, while all four data sets have the same linear regression, it is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature. Conversely, the top left graph probably should be analyzed with a linear regression because it's a scatter plot that moves in a roughly linear manner. These observations demonstrate the value in graphing your data before analyzing it.
- Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

---

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

Answer: **The Pearson's correlation** coefficient varies between -1 and +1. It is a statistic that measures linear correlation between two variables *X* and *Y*. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation:

1. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

2. r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
3. r = 0 means there is no linear association
4. r > 0 < 5 means there is a weak association
5. r > 5 < 8 means there is a moderate association
6. r > 8 means there is a strong association

- Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or

  the *population Pearson correlation coefficient*. Given a pair of random variables

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

  where:

- cov is the covariance
- Sigma of X is the standard deviation of X
- Sigma of 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)

Answer:

- **Scaling** is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
- **Need of Scaling :** If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
- **Example: T**he given data set contains 3 features – Age, Salary, BHK Apartment. Consider a range of 10- 60 for Age, 1 Lac- 40 Lacs for Salary, 1- 5 for BHK of Flat. All these features are independent of each of other and are within different units, so once scaling is performed on this data points, all the variables fall under same range and helps in building model effectively.

There are two ways to perform scaling

1. **Standardized scaling :** It brings all the scaled data into standard normal distribution, with mean around zero and standard deviation as 1.
   x = x-mean(x) / SD(x)
2. **Normalized scaling :** or **MinMax Scaling** brings all the data between the range of 0 to 1.
   x = x-min(x) / max(x)-min(x)

---

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

Answer : The **variance inflation factor** (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

- Multicollinearity was measured by variance inflation factors (**VIF**) and tolerance. If **VIF** value exceeding 4.0, or by tol- erance less than 0.2 then there is a problem with multicollinearity
- The numerical value for **VIF** tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient.

- **A rule of thumb for interpreting the variance inflation factor:**
    - 1 = not correlated.
    - Between 1 and 5 = moderately correlated.
    - Greater than 5 = highly correlated.

- An **infinite VIF value** indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an **infinite VIF** as well)
- These VIFs tell you there is perfect collinearity and you have completely redundant variables. The very first thing you should do to address collinearity is to *think about what the variables mean.*

---

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 Plots (Quantile-Quantile plots) are plots of two quantiles against each other. By a quantile, we mean the fraction (or percent) of points below the given value. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
-
- The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
-
- The assumption of normality is an important assumption for many statistical tests; you assume you are sampling from a normally distributed population. The normal Q Q plot is one way to assess normality. However, you don't have to use the normal distribution as a comparison for your data; you can use any continuous distribution as a comparison, as long as you can calculate the quantiles. In fact, a common procedure is to test out several different distributions with the Q Q plot to see if one fits your data well.