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?**

   The following where some of the observation:
   Seasons: It turns out that maximum no of deals were made during fall followed by summer, winter and spring.

   Months: Maximum no of bikes were rented during the months of Jun, Sep, Aug and Jul respectively

   Days of Week: On the whole there is no much difference in no of deals that happened in 7 days of week. However, we can say that most deals happened during Sat and Sunday. Least no of deals was observed on Monday.

   Weather Conditions: Its very evident that most deals happened when the weather was clear/partly cloudy. Least no deals were observed when weather was light rain/snow and almost no deals during heavy rain/snow.

   It's also interesting that most deals happened in 2019 compared to 2018
   Also it was observed that most deals were done on a working day compared to weekend or holiday.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   If there are n categories in a categorical variable, while creating dummy variables you need not create dummies with n variables, rather (N-1) variables are enough to represent the categories. In this way correlations created among dummy variables are reduced and also model will be as simple as possible.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   The variable temperature has the highest correlation with target variable

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   The following are some of the criteria used to validate our model:
   - As per assumption there indeed was a linear relationship between inputs and output.
   - The p-value of the coefficients were very low which means the coefficients are significant.
   - The error terms had a normal distribution with mean equals zero.
   - Multicollinearity: Little to no multicollinearity among predictors with **VIF** < 5
   - Homoscedasticity of error terms: Residuals had a constant variance.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Yr, Spring and Light snow/rain

General Subjective Questions

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

- Linear regression is a supervised learning algorithm. It is used to learn the relationship between input/predictor and target variables.

- Given a set of data points inputs (x1, x2, x3…) and output (y). linear regression tries to fit a line on the data, while minimizing the squared distance of the points to the fitted line values (minimizing **sum of squared errors**). In general, it is used for finding the causal effect relationship between the variables.

- In general, a simple regression equation, takes the form of a line; **Y = mX + c**
  Y- Output
  X- Input
  m- Coefficient
  C- constant/Intercept

- The regression equation is of the form,
  y = b0 + b1x + e

- The term bo is the intercept, b1 is the slope of the regression line x is the input variable, e is the error term and y is the predicted value of response variable.

- The slope b1 tells how change in the input causes changes in the output.

- For example, Consider the equation **y = 1.8 + 0.86x** this means for a unit increase in x the y is impacted by 0.86. if equation changes to **y = 1.8 - 0.86x** then unit change in x causes y to decrease by 0.86.

- The goodness of the model is evaluated using the measure R^2 (R squared) which is given by the below equation:

  R squared = **1 – (RSS/TSS)**

  RSS (Residual Sum of Squares) = Unexplained Variation

  TSS (Total Sum of Squares) = Total Variation

  Higher R^2 indicates that the model best explains the variations

There are some assumptions made in regression to validate our model. Following are some of the assumptions:

- linear association between input and output variable

- Feature coefficients should be significant (low P- Value)

- Normally distributed error terms with mean equals zero

- Multicollinearity: Little to no multicollinearity among predictors with (**VIF** < 5) or **(VIF** < 8)

- Homoscedasticity of error terms: Residuals must have a constant variance.

2. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

The variables in dataset might not be of same range. In order to have all the variables in same range without losing the contained information, scaling is used.

Scaling quickens the process of gradient descent. If the variables are of different range, the contours of cost function will be skewed, this in turn will take more time while finding the parameters for best fit. This is the reason why feature scaling is used which is a well-known optimization technique.

There are two popular scaling methods:

Normalized Scaling/Min Max Scaling: This method maps given values between zero and one but retains the original distribution.

Standardized Scaling: This method transforms the data to have a mean equals zero and standard deviation of one. The original distribution is retained without losing any information.

3. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

multicollinearity occurs when predictor variables are correlated. VIF is an Index used to measure it. In order to determine VIF, we fit a regression model between the independent variables. VIF is estimated using the below eq where Rsquared is the accuracy of the regression model:

VIF: **1/ (1-RSquared)**

The general rule of thumb is:

- If VIF =1, it implies No Multicollinearity
- If VIF is between 4 to 5, It implies moderate Multicollinearity
- VIF >= 10, Implies severe Multicollinearity

When the VIF becomes infinite it means there is perfect corelation. That is the regression model is 100% accurately describing the output. This makes the Rsquared equal to 1. Therefore, a zero in denominator **(VIF = 1/ (1-1**)) makes the VIF infinite.

4. **What is Pearson's R?**

In general Correlation coefficient tells how strong a relationship is between two variables. There are many correlations formula, one such correlation is Pearson. It shows the linear relationship between two variables. In simple terms, it checks if we can draw a line graph to represent the data. Two letters are used to represent the Pearson correlation: Greek letter rho **(ρ)** for a population and the letter **"r"** for a sample.

The correlation coefficient **r** takes between -1 and 1, where:

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other variable.

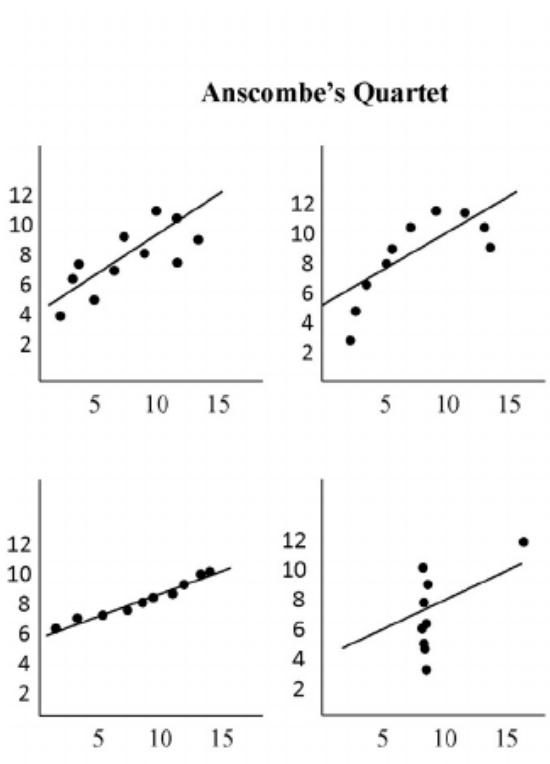
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other variable.

- Zero means that the two variables aren't just related

5. Explain the Anscombe's quartet in detail.

Anscombe's quartet tells us about the importance of visualising the data before applying various algorithms to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Let's consider the classic example of Anscombe's quartet. The fig below shows the Anscombe's quartet of different XY plots of four data sets having identical averages, variances, and correlations. Source: Adapted from Anscombe (1973, pp. 19 – 20).



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

We can see that there are four data sets that have nearly identical simple descriptive statistics, but have different distributions and appear very different when graphed.

The four data set plots which have nearly same statistical observations, that provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

Also, the Linear Regression can be only be considered as a fit for the data with linear relationships and is incapable of handling other kind of datasets.

Therefore, it is very essential to plot the data and analyse them, instead of directly jumping onto model building or any statistical analysis.

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

A Quantile-Quantile (Q-Q) plot is a graphical technique to check if a sample follows a specific distribution. It can also be used to check if two different samples follow the same distribution.

Basically, it is a scatter plot between the quantiles of the two distributions that need to be compared. A Q-Q plot approximately results in the straight-line if the two distributions are identical.

The idea is that if the distributions are identical, then the quantiles should be approximately equal. For example, in the case of quartiles (4-quantiles), the first quartile should be roughly the same for both the distributions, and similarly, the other quartiles. Thus, resulting in the straight-line.

If the points of your data set are near that line, and following that line, and are distributed equally across that line, then, you can say that your dataset is normally distributed.

A normal quantile-quantile (QQ) plot can be used to spot also outliers in a dataset as well.