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)**
   - Working day/Holiday has significant affect on target variable
   - Day of the week has no significant affect on target variable
   - casual riders seem to be more fair-weather riders.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   - we encode each category with a different binary feature. In statistics, it is common to encode a categorical feature with k different possible values into k–1 features (the last one is represented as all zeros). This is done to simplify the analysis (more technically, this will avoid making the data matrix rank-deficient).

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   - Registered users have high correlation with target variable , this is expected as cnt is derived from registered and casual.
   - If we exclude registered and casual users, temp has highest correlation with cnt and this will be used for model.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - checked if the error terms are also normally distributed
   - Plotting y_test and y_pred to understand the spread.
   - Checked Error terms are independent of each other and checked if they have constant variance

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   - Temp
   - Year
   - weathersit_3 (negatively but bigger coefficient value)

General Subjective Questions

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

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. gender,season). There are two main types:

Simple regression

Simple linear regression uses traditional slope-intercept form, where mm and bb are the variables our algorithm will try to "learn" to produce the most accurate predictions. xx represents our input data and yy represents our prediction.

y=mx+b

Multivariable regression

A more complex, multi-variable linear equation might look like this, where ww represents the coefficients, or weights, our model will try to learn.

f(x,y,z)=w1x+w2y+w3zf(x,y,z)=w1x+w2y+w3z

The variables x,y,zx,y,z represent the attributes, or distinct pieces of information, we have about each observation.

Cost Function

The prediction function is nice, but for our purposes we don't really need it. What we need is a cost function so we can start optimizing our weights.

Let's use MSE (L2) as our cost function. MSE measures the average squared difference between an observation's actual and predicted values. The output is a single number representing the cost, or score, associated with our current set of weights. Our goal is to minimize MSE to improve the accuracy of our model.

Given our simple linear equation y=mx+by=mx+b, we can calculate MSE as:

MSE=1N∑i=1n(yi−(mxi+b))2

Gradient descent

To minimize MSE we use Gradient Descent to calculate the gradient of our cost function.

Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

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

- People tended to ignore visualizations in favor of summary statistics. Everyone said it was too much effort to plot the data and they are I a view of mean is enough to know about data.
- Frank old statitioner sent to the Council(fellow practitioners) a parchment with 4 sets of 11 data-points and requested the council as his last wish to plot those points.

graphs were completely different even though the summary was exactly similar.

- The first scatter plot (top left) appeared to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.

- The second graph (top right) was not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Master PoorPoetrix realized the folly of the Council's **ways and rectified them, this data-set came to be known as Anscombe's quartet.**

Reference: https://towardsdatascience.com/fables-of-data-science-anscombes-quartet-2c2e1a07fbe6

3. **What is Pearson's R? (3 marks)**
   In statistics, the Pearson correlation coefficient also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

   The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).
   The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

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

   Feature scaling is an important technique in Machine Learning and it is one of the most important steps during the preprocessing of data before creating a machine learning model. This can make a difference between a weak machine learning model and a strong one. They two most important scaling techniques is Standardization and Normalization.

   Normalization
   Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

   Here's the formula for normalization:

   Normalization equation: X-Xmin/(Xmax-Xmin)

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1.

Standardization

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

Standardization equation X'= (X-Mu)/sigma

Feature scaling: Mu is the mean of the feature values and Feature scaling: Sigma is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

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).

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.