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)
    a. **Weather** for sure has an impact on demand, looks like demand is reasonably high even in December and Jan winter.
    b. Demand growth **increased by 60% plus** in the second year.
    c. Irrespective of year and season, only weather drives demand.
    d. Day of week doesn't have any impact on demand.
    e. Holidays see no demand.
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
    a. **Answer**: For a column with 2 categorical values, the number of columns (dummy columns) required is 1. Say Male and Female can be represented by a single column (dummy variable) with values 1 and 0. If there are 3 categorical values, we need 2 columns (dummy variables), say, single, married, divorced. 1 Column (dummy variable) can handle single vs (married + divorced). The second can handle married vs divorced. Hence the number of dummy variables (columns) required is N – 1, where N is the distinct number of values for the categorical variable. By default, a category column is converted to N dummy variable columns and we need to drop 1 column. **drop_first=True**, by default drops the (alphabetically) first column generated out of the category values. **If 1 dummy variable column is not dropped, it creates multi-collinearity affecting inference with coefficients swinging wildly, signs inverting and unreliable p-values**.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    a. **Answer**: **'temp' has the highest correlation of 0.63 with 'cnt'**. Since 'atemp' was correlated with 'temp' I have dropped it.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. **Answer:** The assumptions of LR are about the errors. I did (a) Distribution plot to check if the error terms are normally distributed (b) Scatter plot of errors to know if there are any patterns or independence (c) Regplot (yellow line), to check if the error terms are having uniform variance (distance) from it – Homoscedasticity (d) Scatter plot (between y_train and y_train_pred) that shows that they are linear (Note that the net effect of all features X is y_pred), which implies linear relationship between Xs and 'cnt'.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    a. **Answer: temp**, **weathersit** (**lightsnowrain**) and **yr** have the highest magnitude of coefficients impacting bike demand. temp and yr have positive and weathersit has negative impact respectively.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    a. **Answer:** Linear regression algorithm is used to do predictive analytics between continuous variables that have a reasonable linear relationship. From data, we fit a straight line that explains the relationship between them. It has the following steps.

i. Since a straight line is y=mx + c, where m is slope and c is intercept for a pair of data points x and y, we want to plot the best y (called y_predicted), such that the distance between y_predicted and observed y is minimal. This difference is the error.

ii. To achieve this, we find minima using differential calculus in simple cases. The methods of differential Calculus is called **Ordinary Least Squares / Residual Sum of Squares / Closed Form. The RSS =** $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_i x_i)^2$**.** We would minimize this to find the right β_0 and β_i which are the intercept and coefficients for each value of a single feature x. This is the typical cost function we want to minimize.

iii. The above algorithm gets complicated for multiple linear regression and we use a methods called **Gradient Descent** which is a numerical method. We get the slope of the equation above and step down towards zero value of slope with a learning factor. In a few iterations when the cost function converges to a fixed value and slope is close to zero, we stop and call that the RSS is minimal.

iv. With multiple linear regression, we have more features (Xs) for a target prediction variable y.

v. We represent the difference between RSS and TSS (Total sum of squares) which is sum of squares of difference of observed y values and $\bar{y}$ (the mean) as a percent of TSS. This is called R^2. **Higher this value, better the predictive power of the model**.

vi. As features increase we need a way of penalizing uncorrelated features with the help of **Adjusted** $\boldsymbol{R^2}$. We also apply F-statistic and Probability (F-Statistic) with p-values of coefficients.

vii. At the core of this is the Gradient Descent algorithm, that helps us to converge to a point where value of the cost function RSS doesn't fall beyond a point and the slope of RSS is infinitely small and close to zero. This is done by fixing other features when gradient descending a given feature X.

viii. The different coefficients for the features and the intercept are used to represent the optimal linear function that helps in interpolating Y and to some extent extrapolating.

ix. To conclude we get the Multiple Linear Regression line as

y = $\boldsymbol{\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ + \ldots + \beta_n x_n}$

We calculate the R2 on the test data and compare it with train data R2 and if they are within a 5% range, we conclude that the LR model is stable.

References – Upgrad Material.

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

a. **Answer:** Anscombe's quartet represents 4 data sets that have **similar statistical interpretation, but the distributions are very different and unique**. The core of this is to understand that effect of outliers that could sway the statistic representation compared against visual. In the context of Linear Regression, for 4 quartets that have same line fitting, if outliers are fixed, would produce a different set of straight lines for each data set. The first quarter has a reasonably right fix, the second has a convex shape that has a wrong straight line, the third has an outlier that seems right but the outlier seems to be outweighed by the rest of data, fourth is not linear but still a linear fit was done wrongly. **The summary of this is that we need to do an EDA and interpret the data visually before**

moving to predictive analytics. This would tell if Linear Regression is the right fit for the data at all.

Reference - https://en.wikipedia.org/wiki/Anscombe%27s_quartet

3. What is Pearson's R? (3 marks)
    a. **Answer: Pearson's R represents the correlation between a pair of numerical datasets. The value can be between -1 to 1**. A value of 0 indicates no correlation, where as a value close to 1 means high correlation indicating when one increases, the other does and when its close to -1, it indicates negative correlation where if one increases, the other decreases. The Pearson's R works best if the data sets are Gaussian or Gaussian like distributions. Say the data set is X and Y, the Pearson's R is given as **Covariance(X,Y)/(Standard Deviation(X) x Standard Deviation(Y))**

    Reference - https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    a. **Answer: Scaling is a process where numerical data in different scales is converted to a uniform scale that enables clean interpretation of their relationships**. Data that is in the order of Millions can be scaled down to a narrow range for better range. Percentage is a good scale rather than saying I got 7898 marks out of 9989. The type of scaling depends on what your application is. **The main objective behind scaling is that the numeric values should be comparable to each other. If they are not scaled, the coefficients will be too large and skewed. Another reason is that, in Linear Regression the underlying Gradient Descent algorithm converges faster.** The 2 common scaling algorithms are Standard Scaling and Normalized scaling.
        i. **Standard scaling** – The values x, in the data set are converted to $(x - \mu)/\sigma$, where μ is mean and σ is Standard deviation. This becomes a Normal distribution with Mean 0 and SD 1. Applying the general Normal distribution rules, 65% of values will be within 1σ , 95% will be within 2σ and 97% within 3σ . This is easy to interpret and apply statistical algorithms as its Gaussian now!
        ii. **Normalized scaling** – The values x, in the data set are converted to $(x - Min_x)/(Max_x - Min_x)$. So the entire data is squeezed between 0 and 1.

        Reference – Upgrad material.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
    a. **Answer:** The formula for $VIF_i = 1 / (1 - R_i^2)$, says that when a very high R value is sighted for the feature, its VIF shoots up. This is an indication of very high multi-collinearity of the feature with others. A value above 5 means high collinearity. But when we sight infinitely high values of VIF, say in the order of 100s, it's an indication that **the feature can be represented as a linear combination of other features that in turn have infinite VIFs as well.**

    References - http://www.imagelab.at/help/vif_descriptors.htm#:~:text=An%20infinite%20VIF%20value%20indicates,an%20infinite%20VIF%20as%20well).
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
    a. **Answer:** Q-Q plot is a plot of the quantiles of two distributions. The pattern that comes out is used to compare the two distributions. Depending on the slope, we can identify the

distribution that is more disperse or with heavier tails. The general application of Q-Q plot is to check if the (a) Distributions are common like Normal or Uniform, (b) Distributions have similar shapes (c) Distributions have similar tail behavior. **In the context of Linear Regression, consider we get the train and test data from two different sources, we would like to know if they are from the same distribution. This would help us validate their distribution types before validating the model with the test data. Another application could be that we want to know if a distribution is Normal. A distribution being Normal is foundational to do any statistical analysis.**

References: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot, https://medium.com/@premal.matalia/q-q-plot-in-linear-regression-explained-ab040567d86f#:~:text=Quantile%2DQuantile%20(Q%2DQ)%20plot,populations%20with%20a%20common%20distribution.