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)
   - ➔ For seasons - Summer and Fall is when the demand for the bikes were the highest.
   - ➔ In clear weather conditions again the demand for the bikes is highest whereas in light rain condition the demand falls very much.
   - ➔ And finally, in misty conditions the demand is half of what we have in clear weather conditions.
   - ➔ In the year 2019 the demand for the bikes went up to an extent as compared to what we had in 2018.
   - ➔ The demand for the bikes is lowest in month of Jan and subsequently goes up in mid of the year and then finally drops again.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   - ➔ drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. In the given dataset, categorical variable yr has only two values - 2018 or 2019, so it makes no sense to have two separate columns to define the same variable, whether the year was 2018 or 2019. Similarly, the categorical variable season and weathersit has 4 and 3 values respectively. By dropping the first, we will reduce the number of columns created using the get_dummies function.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - ➔ From the pair plot, it is observed that dependent variable 'cnt' has the highest positive correlation with variables 'registered' where the correlation is 0.95 as derived from the correlation plot above.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - ➔ There should be no correlation between the residual (error) terms. To verify that the observations are not auto-correlated, we can use the Durbin-Watson test. The test will output values between 0 and 4. The closer it is to 2, the less autocorrelation there is between the various variables (0–2: positive autocorrelation, 2–4: negative autocorrelation). Auto-correlation would lead to spurious relationships between the independent variables and the dependent variable.
   - ➔ Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable. To verify homoscedasticity, one may look at the residual plot and verify that the variance of the error terms is constant across the values of the dependent variable.
   - ➔ There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). To validate this - we check the pair wise scatter plots and see that there is some linear relationship between the dependent variable and the independent variables.
   - ➔ The independent variables should not be correlated. From the correlation plot and heatmap above, we can see that there are independent variables that show collinearity. For example - variables "atemp" and "temp" show collinearity. However, both the variables describe avg temperature on a given day. It is a business' call to consider one of these variables as both are not required for the Linear Regression Model. However we can look at the Variance Inflation Factors (VIF) in such cases. It is calculated by regressing each independent variable on all the others and calculating a score as follows: VIF = 1 / (1 - R2). Hence, if there exists a linear relationship between an independent variable and the others, it will imply a large R-squared for the regression and thus a larger VIF. As a rule of thumb,

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   - ➔ The statistical output displays the coded coefficients, which are the standardized coefficients. Casual has the standardized coefficient with the largest absolute value i.e., 0.4809. This measure suggests that casual is the most important independent variable in the regression model followed by weekday and workingday. However, variables like season_Spring & weathersit_Light Rain has a negative relationship with the dependent variable which signifies that on light rains and in spring season the demand for shared bike rides decreases. VIFs scores above 5 are generally indicators of multicollinearity (above 10 it can definitely be is an issue)

General Subjective Questions

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

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used.

- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

- While training the model we are given:
  x: input training data (univariate – one input variable(parameter))
  y: labels to data (supervised learning)

- When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
  θ1: intercept
  θ2: coefficient of x

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

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

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

- Pearson's is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. Basically it is used to measure how strong a relationship is between 2 variables. It is commonly used in linear regression.
- 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
- 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. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.
- It is commonly represented by the Greek letter ρ

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

- Scaling is one of the most important data pre-processing steps in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.
- Tree-based algorithms are insensitive to the scale of the features. Also, feature scaling helps machine learning, and deep learning algorithms train and converge faster.
- There are some feature scaling techniques such as Normalisation and Standardisation that are the most popular
- Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:
- X_new = (X - X_min)/(X_max - X_min)
- This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes, but the age is close to uniform.
- Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as
- X_new = (X - mean)/Std
- Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.

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

- If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:
- One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
- Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

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

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. 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.
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.
- Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.