Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1. The following observations were made during the analysis of categorical variables:

1. Bike demand is more in 2019 as compared to 2018
2. Bike demand is highest during the fall season and lowest during the spring season
3. During clear weather, the bike demand was the most and during light snow and rain, it was the least.
4. Bike demand is high between May to October and starts decreasing from November onwards. Bike demand is the lowest in January month.
5. There is not much difference between the bike demand during holidays/weekends and working days with demands during the working day being just slightly higher.
6. The bike demand throughout the week remains almost similar.

Q2. Why is it important to use drop_first=True during dummy variable creation?

A2. If a column has 3(n) dummy variables, then it can be represented by 2 dummy variables(n-1). Drop_first in get_dummies is used to drop the extra dummy variable.

For Example, we have three variables: Clear, Misty, and Light_Snow&Rain. We can only take 2 variables as Clear will be 1-0, and Misty will be 0-1, so we don't need Light_Snow&Rain as we know 0-0 will indicate Light_Snow&Rain.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A3. temp and atemp variable has the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A4. The assumptions of the linear regression model can be validated by Residual Analysis of the Train data. A histogram or a distplot can be used to check if the Residual or Error (i.e difference between the y_train and y_train_pred) is normally distributed or not with the mean equal to zero.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5. The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. temp (temperature) – It has a modal coefficient value of 0.597749 which indicates that a unit increase in temp variable, increases the bike count by 0.597749 units.
2. yr (year) – It has a modal coefficient value of 0.227954 which indicates that a unit increase in yr variable, increases the bike count by 0.227954 units.
3. Light_Snow&Rain(Light Rain and Snow)- It has a modal coefficient value of -0.231830 which indicates that a unit increase in yr variable, decreases the bike count by 0.231830 units.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

A1. Linear regression is a form of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression, the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as: $y = mx + c$

Where m = Slope of the line

c = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

Linear Regression is a type of supervised learning model. In the Supervised Learning model, we use training data to build the model and then use test data to test its accuracy. Linear Regression shows the relationships between a set of independent variables to that of the dependent variable.

If our model is well-trained using Linear Regression then, in that case, the predicted point will lie on the regression line.

R-square and adjusted R-square is generally used to determine if the model is good or not. R-squared values range from 0 to 1 and are commonly stated as percentages from 0% to 100%. An R-squared of 100% means that 100% of the dependent variable are completely explained by changes in the independent variable.

Q2. Explain the Anscombe's quartet in detail.

A2. Anscombe's quarter consists of four datasets which has very identical statistical description but has very different distribution when the datasets are plotted on graphs. Each dataset consists of eleven data points.

Statistician Francis Anscombe constructed the quartet in 1973 to demonstrate the importance of plotting the graphs of datapoints while analyzing the data and the effect of outliers on statistical properties.

The quartet was intended to counter the impression among statisticians that numerical calculations are exact, but graphs are rough.

Q3. What is Pearson's R?

A3. The Pearson correlation coefficient also known as Pearson's R, 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.

An absolute value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1. A value of 0 implies that there is no linear dependency between the variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4. Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

A dataset may contain features varying highly in magnitude, units, or range. If scaling is not done then the algorithm only takes the magnitude into consideration and not units which may lead to incorrect modelling. Scaling helps to bring all features or variables to the same level of magnitude.

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). sklearn.preprocessing.scale helps to implement standardization in python.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5. If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) equals infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6. 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 falls 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.