

## 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?

Answer: The demand of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018.

2) Why is it important to use `drop_first=True` during dummy variable creation?

Answer: `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.

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

Answer: the 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?

Answer: Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

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

Answer: temperature, year and season winter is the features with high demand.

## General Subjective Questions

1) Explain the linear regression algorithm in detail.

Answer: 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.

equation  $y=mx+c$

2). Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3) What is Pearson's R?

Answer: Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name

4) What is scaling?

Answer: It 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.

Why is scaling performed?

Answer: Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

What is the difference between normalized scaling and standardized scaling?

Answer: normalized 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 ( $\mu$ ) zero and standard deviation one ( $\sigma$ ).

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

Answer: If there is perfect correlation, then  $VIF = \infty$ . A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Importance:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior