Assignment-based Subjective Question

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

Ans: From the analysis of the categorical variables from the dataset it could be inferred the bike rental rates are likely to be higher in summer and the fall season, are more prominent in the months of September and October, more so in the days of Sat, Wed and Thurs and in the year of 2019. Additionally we could discern that bike rental is higher on holidays.

Used Box plot (refer the fig above) to study their effect on the dependent variable ('count') more deeply.

Season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

month: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

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

Ans: - 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?

Ans: There is linear relationship between temp and atemp. Both of the parameters cannot be used in the model due to multicolinearity. We will decide which parameters to keep based on VIF and p-value w.r.t other variables

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

Ans: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

The Coefficient values from the model of all the variables are not equal to zero which means we are able to reject Null Hypothesis.

F-Statistics is used for testing the overall significance of the Model: Higher the F-Statistics, more significant the Model is.

F-statistic: 249.2

The F-Statistics value of 249 (which is greater than 1) and the p-value of '~0.0000' states that the overall model is significant.

The Residuals were normally distributed after plotting the histogram. Hence our assumption for Linear Regression is valid.

VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5.

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

The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and the Light Rain & Snow variables.

I. Temperature (Temp) A coefficient value of ‘0.478177’ indicated that a temperature has significant impact on bike rentals
II. Light Rain & Snow (weathersit =3) A coefficient value of ‘-0.286002’ indicated that the light snow and rain deters people from renting out bikes
III. Year (yr) A coefficient value of ‘0.234060’ indicated that a year wise the rental numbers are increasing.

So, it's suggested to consider these variables utmost importance while planning, to achieve maximum Booking.

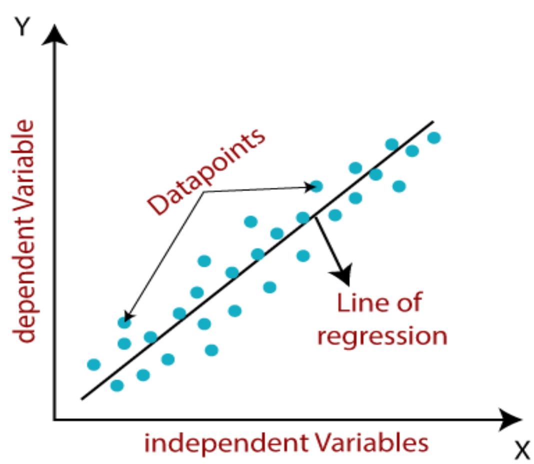
General Subjective Questions

1) Explain the linear regression algorithm in detail.

Ans - 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.

It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. 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. For example, X (input) is the work experience and Y (output) is the salary of a person.



The regression line is the best fit line for our model.

$y = \theta 1 + \theta 2.x + \varepsilon$

Here,

$\theta 1$: intercept of line.

$\theta 2$: Line regression coefficient of x

x: input training data(Independent Variable).

y: labels to data(Dependent Variable).

ε: random error

Once we find the best $\theta 1$ and $\theta 2$ values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

We use Cost Function to update the value of $\theta 1$ and $\theta 2$ to get the best fit line Cost function (J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression**:
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear Regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

2) Explain the Anscombe's quartet in detail.

Ans: - 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.

Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear reparability of the data, etc.

The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3) What is Pearson's R?

Ans:- Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatter plot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrates.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction).

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions).

r = 0 means there is no linear association.

r > 0 < 5 means there is a weak association.

r > 5 < 8 means there is a moderate association.

r > 8 means there is a strong association.

The figure below shows some data sets and their correlation coefficients. The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: - 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.

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 unit's hence incorrect modeling. 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.

1- 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.

2- 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.
   One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans - The variance inflation factor (VIF) quantifies the extent of correlation between one Predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

VIF = 1/1-R^2

If there is perfect correlation, then VIF = infinity. 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. This would mean that that standard error of this coefficient is inflated by a factor of 2. The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant 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.

Ans – 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.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The slope tells us whether the steps in our data are too big or too small. For example, if we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes (a.k.a. quantiles) compare between our data and the normal distribution.

A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. One example cause of this would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).

(Submitted by Md Irshad Ali)