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?**

- Spring season is when the demand is low as compared to all other seasons, while Fall and Summer account for top 2 in demand.
- The year 2019 was significantly a good year, which shows thate the demand is increasing with time.
- Demand is at all time low for the month January (probably because of snowfall), whereas September is when the demand is at its peak.
- Seasons like Fall and Summer are good for business.
- Weather conditions like light snow and rain are bad for business.
- On weekends and holidays there is a slight increase in demand as compared to weekdays.

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

When dealing with categorical variables/predictors in regression (the output variable to be predicted is a continuous variable) we create dummy variables, the logic behind dummy variables is to create 'n-1' variables, 'n' being the number of levels related to that categorical variable. While creating the dummy variables and not setting 'drop_first=True' we'll get 'n' dummy variables which is not good for model as the dummy variables created will exhibit collinearity, so it's better to set 'drop_first=True' as it drops the extra column during dummy variable creation.

E.g., Let's say we have a categorical variable (Status) with 3 levels Single, Married and Divorced. The dummy variables creation for this would look like:

Status	**Single**	**Married**	**Divorced**
Single	1	0	0
Married	0	1	0
Divorced	0	0	1

But to describe ‘n’ levels ‘n-1’ columns are sufficient:

Status	**Married**	**Divorced**
Single	0	0
Married	1	0
Divorced	0	1

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

Looking at the pair-plot among the numerical variables ‘registered’ has the highest correlation with the target variable ‘cnt’.

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

While going for Linear Regression we must consider following assumptions:

- There is a Linear relationship between X and Y
- Error terms are normally distributed with mean = 0
- No multicollinearity
- Error terms have constant variance (homoscedasticity)

To check for linear relationship, we can use pair plot between the independent/predictors variables and the dependent variables.

To check if the errors are normally distributed or not, we do the residual analysis where we do a histogram plot of error terms and check if it is normally distributed with mean = 0 or not.

To check for not multicollinearity we use the method called VIF (Variance Inflation Factor). A VIF a predictor explains if that predictor can be explained with the combination of others or not, if yes, then we can drop them. A VIF > 10 means there is multicollinearity among the variables. We can also use a correlation matrix to check multicollinearity.

To check for homoscedasticity, we plot a scatter plot to see if the residuals are equal along the regression line.

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 explaining the demand of shared bikes are temp, winter season and month of September.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

In the area of predictive analysis Linear Regression is very common practice, the basic idea behind it is to analyse the relationship between dependent and independent variables by fitting a linear equation over the data. Linear Regression helps us in examining how good the predictors (independent variables) can predict the outcome (dependent variable), it also helps in defining that which dependent variables are best suited to make the predictions.

First, we need to check if linear regression is suitable for a given data, for that we can use a scatter plot, if the relationship seems linear, we can opt for a linear model. But if not, we need to apply some transformations to make the relationship linear.

The representation of a Liner Regression model is in the form a linear equation which the represents the equation of the fitted straight line:

$$Y = B_0 + B_1 X_1 \ldots\ldots + B_n X_n$$

Y: dependent variable

B_0: intercept/constant

B_1: Slope

X: independent variables/predictors

If the value of X_1 increases by 1 unit, keeping other variables constant, the total increase in the value of Y will be B_1. Mathematically, the intercept term (B_0) is the response when all the predictor terms are set to zero or not considered.

While going for Linear Regression we must consider following assumptions:

- There is a Linear relationship between X and Y
- Error terms are normally distributed with mean = 0
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

To check for linear relationship, we can use pair plot between the independent/predictors variables and the dependent variables.

To check if the errors are normally distributed or not, we do the residual analysis where we do a histogram plot of error terms and check if it is normally distributed with mean = 0 or not.

To check for not multicollinearity we use the method called VIF (Variance Inflation Factor). A VIF a predictor explains if that predictor can be explained with the combination of others or not, if yes, then we can drop them. A VIF > 10 means there is multicollinearity among the variables. We can also use a correlation matrix to check multicollinearity.

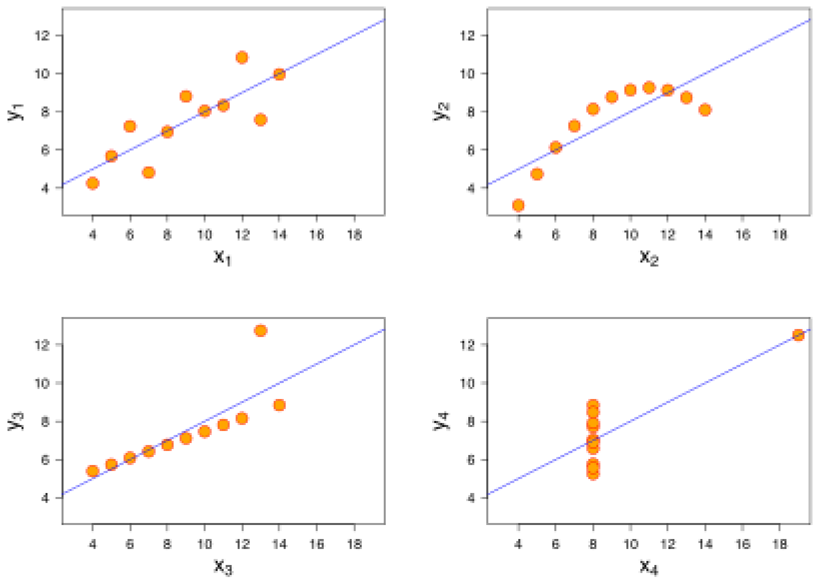
The significance of a model and goodness of fit is done via the F-statistic. Whereas the p-values of betas help in determining the significant predictors for a good model. After the selection of significant predictors and building of model the R-squared and the Adjusted R-Squared values help us determine the goodness of model for simple and multiple linear regression models.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet refers to the 4 different datasets of 11 points which somehow look similar with respect to descriptive statistics like mean , variance, etc. It was suggested by Francis Anscombe, that even if datasets exhibit similar statistics, they may turn out to be different on visualisation with graphs.

Thus, Anscombe’s quartet laid emphasis on not only relying on statistical information but also on visualisation of datasets for the analysis as both can tell a different story. Anscombe’s quartet also laid emphasis on role of outliers in a dataset as just by statistically observing the data we cannot see the effect an outlier can play but on plotting the data we will be able to see its effect.

The Anscombe’s quartet graph is given below:



We can see how datasets have similar statistical information can still be different if we plot them. The effect of outliers can be seen in plot 3 and 4. In plot 3 the data set tends to have a linear relationship apart from the single outlier whereas in plot 4, X tends be constant but the graph shows it otherwise because of the presence of the extreme outlier.

3. What is Pearson’s R?

In statistics correlation is used to check if the two variables have some relationship or not maybe liner or non-linear. Thus, if two variables are correlated then it doesn’t mean that they might have a linear correlation therefore to check if the relationship is linear or not for the two correlated variables, we use Pearson’s R. Its value ranges from -1 to +1 higher the value the higher liner relationship exists.

Pearson’s R can only be taken in consideration if and only if there exists any liner relationship because sometimes the Pearson’s R can show a significant value for two variables which are non-linearly related, in such case we neglect Pearson’s R values, as it’s only useful for linearly correlated variables.

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

Sometimes the datasets may contain features each in a different magnitude and scale thus making it difficult for us to interpret them or compare them, so it is necessary to scale them and bring them on a same level for better understanding of data. Thus, this process to changing the scale of different features into one similar scale is called scaling.

E.g., If a dataset contains weights of male and female in kilograms and pounds it would be difficult to compare them at all without bringing them to a similar scale.

Scaling is important to compare the data. It is also useful in model building, especially in multiple linear regression where we use coefficient values to determine the importance of features in prediction, if the feature scaling is not performed then the coefficient of feature with high magnitude will weigh in a lot while some of the coefficients as obtained by fitting the regression model might be very small as compared to the other coefficients and its effect can be seen at the time of model evaluation. Therefore, feature scaling is important.

The difference between min-max/normalized scaling and standardized scaling is that in min-max scaling the data is scaled in the range of 0 to 1 while in the standardized scaling replaces the values with their Z-score thus having mean as 0 and standard deviation as 1. The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there is are extreme data point (outlier).

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

Variance Inflation Factor or VIF, gives a basic idea of how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. The formula for calculating VIF is:

$$VIF_i = \frac{1}{1-{R_i}^2}$$

Therefore, a VIF of '5' means that 80% of the variance of that column/variable/predictor is explained by all the other columns

Therefore, an infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of all other variables (which show an infinite VIF as well).

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

A Q-Q plot or a quantile-quantile plot in statistics is a graphical approach used for comparison of two probability distributions by plotting their quantiles along with a reference line. In other words, its is used to verify if the two datasets belong to the same population or not. It is a scatter plot with 45^{o} degree reference line, this reference line here is used to determine if the two data sets belong to a population with common distribution. If the data points fall along the reference line, then its concluded that they belong the population with common distribution if not then they came from population with different distributions.

In linear regression model we have a training dataset and a testing dataset then just to confirm if both the datasets belong to the population with same distribution or not, then we use Q-Q plot.