A. Assignment based Subjective Questions.

Answers

Ans-1: Following inferences about the effect of given categorical variables on the dependent variable were made based on Univariate/Bi-variate and Multivariate analysis carried out to understand the data:

a. Seasonal user demand increases from spring-summer-fall-winter with lowest in spring and highest in fall.
b. Average user demand in year 2019 is much higher than year 2018 in all seasons and throughout the year.
c. Month wise user demand is increasing steadily from January to May-June-July-Aug (Fall season) and then decreased till december.
d. User demand dips on weather holidays whereas average demand is almost same during week days & working days (weekend+holiday).
e. User demand is max in Clear weather days and min in Light Snow/Rain days, almost nil on Heavy Rain/Ice Pallets days.

Ans-2: It is important to use **drop_first = True** during the dummy variable creation step, as we need to create only n-1 dummy variables (where n is the max number of levels) for a categorical variable. So , 1st dummy variable is dropped as default to ensure that we are not creating redundant variable in the process of creating the dummy variables.

Ans-3: The highest correlation with target variable i.e. Total User Count (Registered + Casual) is with 'temp' and 'atemp' predictor variables.

Ans-4: Some of the Linear Regression assumptions were checked during the model building process viz linear relationship between the Dependent and predictor variables, use of independent predictor variables in modeling process dropping the insignificant predictor variables. After model building using training dataset, error terms distribution for residuals is checked for normal distribution and zero mean.

Ans-5: Top three predictor variables or features significantly affecting the demand are:

a. Temperature
b. Weather
c. Season

B. General Subjective Questions.

Answers

Ans-1: In statistical or predictive modeling, linear regression is a process of estimating the relationship among dependent and independent variables. The focus here is to establish the relationship between a dependent or response variable and one or more independent or predictor variable(s). Regression helps you understand how the values of dependent variable changes as you change the values of 1 predictor, holding the other predictors static (or same).

Two types of linear regression are used:

1. Simple Linear Regression (SLR): It is the most elementary type of regression model, which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.
2. Multiple Linear Regression (MLR): MLR is used to understand the relationship between one dependent variable and several independent variables (explanatory variables).

Following LR assumptions are made in the analysis:

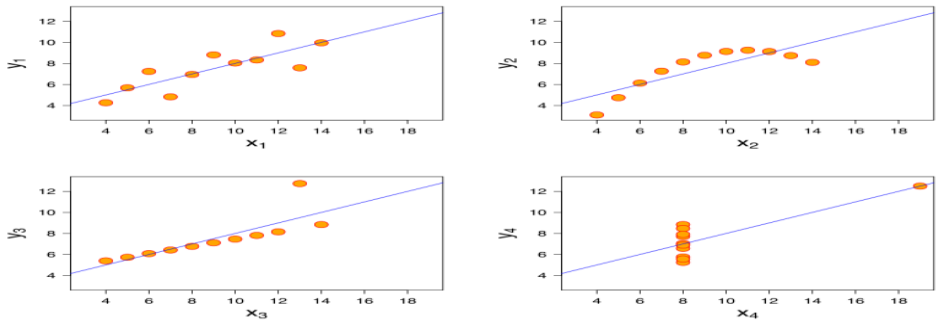
- Linear relationship between independent and dependent variable (s).
- Error terms are normally distributed with zero mean.
- All independent variables are un-correlated or have either no or little multicollinearity.
- Variance of error terms is constant or Homoscedasticity.

In LR method:

- Data is cleaned and prepared for the analysis and modeling.
- Data pre-processing is carried out including null values treatment, data type errors, outlier treatment etc.
- Test and train datasets are prepared and scaling (if required) is also carried out.
- Model Building stage is a very crucial step for model development process with variable selection for the model. It is not a good call to consider all variables in the model because a variable may or may not impact the results of the model. Thus, you have to remove variables based on multicollinearity (VIF) and p-values. Ideally, the model should have a limited number of variables which explain the outcome well.
- A best-fit Regression line is found as defined with intercept constant and slope coefficient (s).
- Residuals or error is calculated at each data point by subtracting predicted value of dependent variable from actual value of dependent variable, followed with calculation of the sum of squares of the residuals.
- Regression line coefficients are calculated by minimizing the cost function or expression of RSS (Residual Sum of Squares) using Ordinary least squares method.
- Residual Analysis is carried out for checking the error distribution.
- Model Validation: The strength of the linear regression model is assessed using R2 or Coefficient of Determination which provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.
- Model Validation is also done by comparing the R-square value between the predicted value and the actual value in the test set, which should be high. In general, it is desired that the R-squared on the test data be high, and as similar to the R-squared on the training set as possible.
- You should note that R-squared is only one of the metrics to assess accuracy in a linear regression model. There are many other metrics.

Ans-2: Anscombe's Quartet was constructed by statistician Francis Anscombe in year 1973 to illustrate the importance of plotting the graphs before analyzing and model building. It was explained by taking four data sets (comprising one dependent and one predictor variable) with nearly identical simple descriptive statistics, however the inter-relationships between the variables was very different otherwise not visible on analyzing the statistical information or observations . Due to existence of these peculiarities in the relationships between the variables not distinctly explained by statistical information only, it was not possible to build the proper regression model until these datasets are plotted graphically for analysis. In fact, these sample datasets taken in the instant case actually have different distributions clearly evident in the graphical plots on scatter plots.

The statistical information for all these four datasets is approximately similar however their graphical scatter plots are showing different distribution:



Ans-3: Pearson's R is also known as Pearson's Correlation Coefficient was given by Karl Pearson. It calculates the linear relationship between two variables with values ranging from +1 to -1. The magnitude tells us the strength of the relationship while the sign suggests the direction. It is the most commonly used correlation coefficient in statistics, however it cannot differentiate between dependent and independent variables. Pearson's R can also be defined as the covariance of the two variables divided by the product of their standard deviations.

Ans-4: Scaling is carried out during data pre-processing stage on independent or predictor variables having data values varying from very large ranges (max & min values) variables to very small ranges to (max & min values) variables requiring normalization of the data within a particular range. If no scaling is carried out, coefficients will be weighted in magnitude and would lead to incorrect modeling. Scaling does not affect the parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Generally two types of scaling methods are used:

a. **Normalization or Min-Max Scaling**: All data values are normalized to the range of 0 and 1 as all data values (x) are replaced as **$(x-X_{min})/(X_{max} - X_{min})$.**

b. **Standardization Scaling**: All data values (x) are replaced as $(x- \mu)/ \sigma$ (z-score) in the data range. It brings all of the data into a standard normal distribution having zero mean (μ) and unit standard deviation (σ).

Ans-5: VIF can be infinite for a predictor variable, when such predictor variable can be completely defined by other related predictor variables. Such predictor variable is basically redundant with very high correlation with other predictor variables and needs to be dropped from its usage in the model building. Mathematically speaking, VIF infinite means corresponding R^2 is equal to one.

Ans-6: Quantile-Quantile (Q-Q) plot is used for carrying out the assessment about the population data in linear regression cases, where train and test datasets of a population are provided distinctly instead of obtaining them by splitting the single dataset. Q-Q plot is a graphical tool and also used to check the statistical distribution type of source population data viz Normal, exponential or Uniform distribution. In simple words, a Q-Q plot is a plot of the quantiles of the one data set versus the quantiles of the second data set.