**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?**

**Ans:**

* Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019 which means this business showed progress
* Most of the bookings has been done during the month of Q2, Q3 and Oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
* Clear weather attracted more booking. There are no bookings for high weather conditions.
* Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
* When it’s not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
* Bookings are equal on working day and non-working day as well.

1. **Why is it important to use drop\_first=True during dummy variable creation?**

**Ans:** The purpose of the “drop\_first=True” argument is to remove excess columns while creating dummy variables of the categorical data. Which implies whether n-1 dummies out of n categorical levels by removing the first level.

It is more than sufficient to have n-1 dummies which gives the same meaning. For example, let’s say there are totally three unique values in a categorical column. If we want to create dummy variables for that column. If one variable is not equal to A or B, then it’s obviously equals to C. So, we don’t need the third variable.

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

**Ans:** ‘temp’ and ‘atemp’ variables have highest positive correlation with the target variable.

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

**Ans:** Validated the assumptions of linear Regression model based on 5 assumptions.

* Linear Relationship Validation: linear relationship should be directly visible with independent variable.
* Normality or error terms: Error terms are normally distributed.
* Homoscedasticity: Variance of the residual in model should be constant.
* Independence of Residuals: No auto correlation.
* Multicollinearity check: There should not be inter relation among independent variables.

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

**Ans:** Here are the top 3 features explaining the demand of shared bikes

* Holiday
* temp
* sep

**General Subjective Questions**

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

Ans: Linear regression is a statistical model that analyses the linear relationship between a dependent variable with one or more than one independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (either increase or decrease), the value of dependent variable will also change accordingly (either increase or decrease). the relationship between dependant and independent variable can be explained with the help of following equation −> Y = mX + c Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

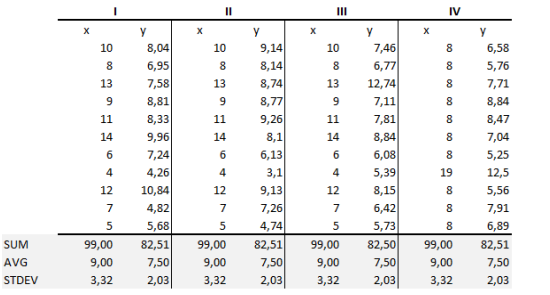
Linear Relationship can be positive or negative in nature.

Positive Linear Relationship: Predictor variable increases with respect to independent variable.

Negative Linear Relationship: Predictor variable decreases with respect to independent variable.

**2. Explain the Anscombe’s quartet in detail.**

Ans: Anscombe’s Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. These four datasets depicts different story while plotting in graph though they have different summary statistics. This illustrates importance of graphing data is statistical analysis to avoid misleading conclusions solely based on summary statistics.



The summary statistics show that the means and the variances were identical for x and y across the groups:

* Mean of x is 9 and mean of y is 7.50 for each dataset.
* Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
* The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:

A group of graphs with dots

Description automatically generated

* Dataset I appears to have clean and well-fitting linear models.
* Dataset II is not distributed normally. In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
* Dataset IV shows that one outlier is enough to produce a high correlation coefficient. This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

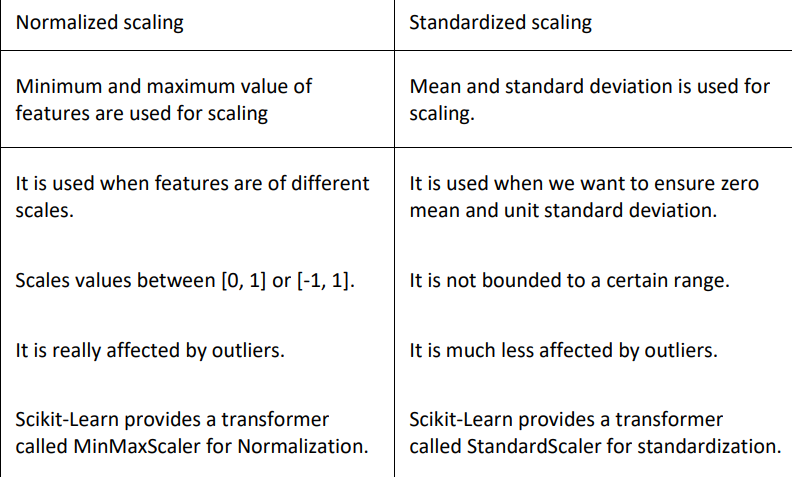
**3. What is Pearson’s R?**

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. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:

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

Ans:

Scaling: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Example: If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that’s actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.



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

Ans: 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. When the value of VIF is infinite, it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests