Name: Ganesh Visweswaran
Module 2: Machine Learning Part 1

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>
   (a) Year 2019 has customer improved across all seasons and weather
   (b) Customer demand was higher during clear weather
   (c) During Holidays, customer demand is less
   (d) Summer and Fall has higher bike rental demand than winter and spring

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

   <Answer>
   For a category variable with n orders, it is sufficient to create n-1 dummies. By reducing the dummies, no of features for modelling gets reduced which has significant effort is creating model easier.
   Drop_first=True allows to create n-1 dummies for variable with n orders by removing the first order.

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

   <Answer>

   Temperature (temp/atemp in dataset) has highest co-relation with target variable (cnt in dataset)

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

   <Answer>
   (a) Linear relationship between dependent variable and target variable (by pairplot)
   (b) No Multicollinearity among features (by VIF check)
   (c) No auto-correlation in the residuals1234
   (d) Homoscedasticity of the residuals
   (e) Mean value of residual error is zero (binomial distribution)

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

<Answer>
   (a) Temperature (b) Year (c) Windspeed

General Subjective Questions

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

<Answer>
Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
Mathematically the relationship can be represented 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.

It can be a positive linear relationship or negative linear relationship

Linear regression is of the following two types – Simple Linear Regression and Multiple Linear Regression

Assumptions - The following are some assumptions about dataset that is made by Linear Regression model

- Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no autocorrelation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms – Error terms should be normally distributed
- Homoscedasticity – There should be no visible pattern in residual values.

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

<Answer>
Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

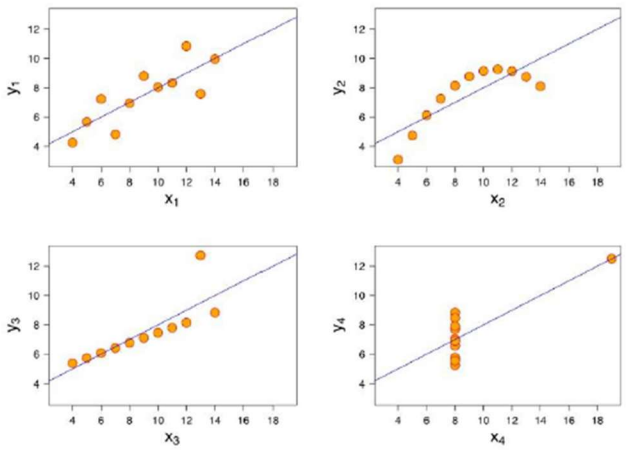
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

• 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:



• Dataset I appear 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?

<Answer>

Pearson's correlation coefficient, also known as **Pearson's r**, is a statistic that measures the linear correlation between two variables. It is a number between –1 and 1 that indicates the strength and direction of the relationship. A value of 0 means no association, a value greater than 0 means a positive association, and a value less than 0 means a negative association. Pearson's r is the most common way of measuring a linear correlation.

The formula for Pearson's r is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where:

- (x_i) and (y_i) are the individual data points for the two variables.
- (\bar{x}) and (\bar{y}) are the means of the two variables, respectively.

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

<Answer>

(a) Scaling is the process of transforming data so that it fits within a specific range. It's commonly used in machine learning and data analysis to bring multiple variables onto a similar scale.

(b) Purpose: Scaling is performed for several reasons:

Equalizing Variables: When different features (variables) have different units or ranges, scaling ensures that they are comparable.

Improving Model Performance: Many machine learning algorithms work better when features are scaled.

Gradient Descent Convergence: In optimization algorithms, scaling helps gradients converge faster.

(c) Two popular methods are Normalization and Standardization.

(d) **Standardization**: Have a mean of 0 and a standard deviation of 1. Less sensitive to outliers. Works well for algorithms that assume normally distributed data. But, alters the original distribution. Sckit- StandardScaler library is used

(e) **Normalization**: Normalized values lie between 0 and 1. Preserves the original distribution. Useful when features have varying ranges. But, sensitive to outliers. Sckit-MinMaxScaler library is used

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

<Answer>

VIF shows the magnitude of variance of the coefficient estimate is being inflated by collinearity.

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

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 (R^2) =1, which lead to 1/ (1-R^2) infinity.

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

<Answer>

- 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.