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? (3 marks)

Ans: Observations from above boxplots for categorical variables:

- The year box plots indicates that more bikes are rent during 2019 may be due to growing popularity and people becoming more aware about environment.
- The season box plots indicates that more bikes are rent during fall and summer season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays may be people prefer to travel in personal vehicle while spending time with family.
- The month box plots indicates that more bikes are rent during september month and mid range month which is reflection of season boxplot.
- The week plot shows Median values for weekday almost remains same, demand not highly correlated with weekday, we do see slight increase in saturday and wednesday though
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather i.e good weather condition.

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

Ans: drop first=True is necessary to utilize since it reduces the additional column produced when creating dummy variables. As a result, the correlations between dummy variables are reduced. n-1 dummy variables can be used to represent a variable with n labels, for example. The speed variables 'Low, Medium, and High' can have columns Medium and High, and when both are 0, it automatically becomes Low, therefore only two dummy variables are required.

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

Ans : By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'. (Variables post feature selection (we have drooped regestered and casual))

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans Validated assumptions of linear regression by checking below points:

- Multicollinearity check - There should be insignificant multicollinearity among variables (checking VIF).
- Normality of error terms - Error terms should be normally distributed

- Linear relationship validation - Linearity should be visible among variables (Pair and reg plots)
- Homoscedasticity - There should be no visible pattern in residual values (plotting scatter plot between residual and y train).
- Independence of residuals - No auto-correlation

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

Ans: The Top 3 features contributing significantly towards the demands of share bikes are:

a. temp(Positive correlation).

b. yr (Positive correlation).

c. Windspeed(negative correlation).

d. WeatherSituation Good Clear (Positive correlation) (Wind speed and Good weather have close coeff)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is a supervised learning machine learning technique. It carries out a regression job. Based on independent variables, regression models a goal prediction value. It is mostly utilised in predicting and determining the link between variables. Different regression models differ in terms of the type of relationship they evaluate between dependent and independent variables, as well as the number of independent variables they employ.

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.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship: A linear relationship will be called positive if both independent and

dependent variable increases.

- Negative Linear relationship: A linear relationship will be called negative if independent increases and dependent variable decrease

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions:

- Multicollinearity check - There should be insignificant multicollinearity among variables (checking VIF).
- Normality of error terms - Error terms should be normally distributed
- Linear relationship validation - Linearity should be visible among variables
- Homoscedasticity - There should be no visible pattern in residual values.
- Independence of residuals - No auto-correlation

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet consists of four datasets with virtually similar elementary statistical features that, when graphed, look radically different.

There are eleven (x,y) points in each dataset. They were created by statistician Francis Anscombe in 1973 to show the significance of charting data before studying it, as well as the impact of outliers on statistical features.

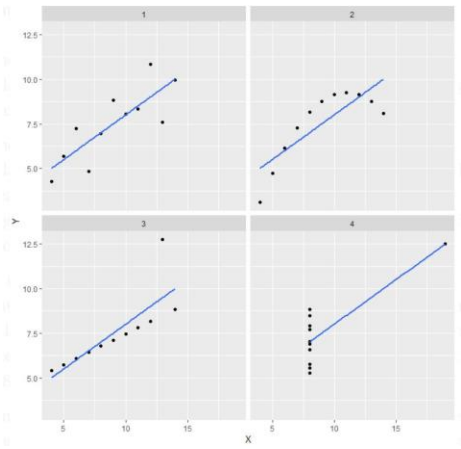
Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

the mean, standard deviation, and correlation between x and y looked similar.

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817



If you look at the scatter plot in the first one (top left), you can see that x and y appear to have a linear connection.

If you look at the second one (top right), you can see that there is a non-linear relationship between x and y.

In the third (bottom left), you may argue that all of the data points have a perfect linear relationship except one, which appears to be an outlier and is highlighted as being far away from the line.

Finally, the fourth one (bottom right) demonstrates how one high-leverage point may yield a large correlation coefficient.

Usage:

The quartet is still frequently used to demonstrate the significance of visually inspecting a set of data before beginning to analyse it according to a certain sort of connection, as well as the insufficiency of fundamental statistical features for characterizing genuine datasets.

3. What is Pearson's R? (3 marks)

Ans: Pearson's r is a numerical representation of the strength of the linear relationship between variables. The correlation coefficient will be positive if the variables tend to go up and down together. The correlation coefficient will be negative if the variables tend to go up and down in opposite directions, with low values of one variable correlated with high values of the other.

The Pearson correlation coefficient, r, can be anything between +1 and -1. A value of 0 implies that the two variables have no relationship.

A positive connection is defined as when the value of one variable grows, the value of the other variable increases as well.

A negative relationship is indicated by a value less than 0; that is, when the value of one variable rises, the value of the other variable falls.

The graphic below depicts this.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: What is scaling?

It is a data pre-processing technique that is used to independent variables in order to normalise the data within a given range.It also aids in the acceleration of algorithm computations.

Why used?

The majority of the time, the acquired data set comprises features with a wide variety of magnitudes, units, and ranges.

If scaling is not done, the algorithm will only consider magnitude rather than units, resulting in erroneous modelling. To address this problem, we must scale all of the variables to the same magnitude level.

Scaling only impacts the coefficients and not the other parameters like the t-statistic, F-statistic, p-values, R-squared, and so on.

Normalization or Min-Max Scaling: is a technique for transforming features into a similar scale.

The new point is determined as follows:

X_new = (X - X_min)/(X_max - X_min)

This reduces the range to [0, 1] .Transformation compresses n-dimensional data into an n-dimensional unit hypercube geometrically.When there are no outliers, normalisation is advantageous since it cannot cope with them.

Z-Score or standardisation:

The process of normalisation involves subtracting from the mean and dividing by the standard deviation.

This is known as the Z-score.

X_new = (X - mean)/Std

When the data has a Gaussian distribution, standardisation might be beneficial. This, however, does not have to be the case.Geometrically, it squishes or expands the points if std is 1 and translates the data to the mean vector of the original data to the origin.

We can see that we're simply altering the mean and standard deviation to a standard normal distribution, which is still normal, therefore the distribution's form is unaffected. Outliers have no effect on standardization since there is no specified range of converted characteristics.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

ANS: VIF = infinite occur when there is perfect correlation. This demonstrates that two independent variables have a perfect correlation.

We get R2 = 1 in the event of perfect correlation, which leads to 1/(1-R2) infinite.

To overcome this issue, we must remove one of the variables that is producing the perfect multicollinearity from the dataset. An infinite VIF value suggests that a linear combination of other variables may perfectly represent the related variable (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. (3 marks)

Ans The Quantile-Quantile (Q-Q) plot is a graphical tool that may be used to determine if a collection of data is likely to have come from a theoretical distribution such as a Normal, exponential, or uniform distribution.

It also aids in determining whether two data sets are from populations with similar distributions.

This is useful in a linear regression scenario where the training and test data sets are obtained separately and the Q-Q plot is used to demonstrate that both data sets are from populations with similar distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

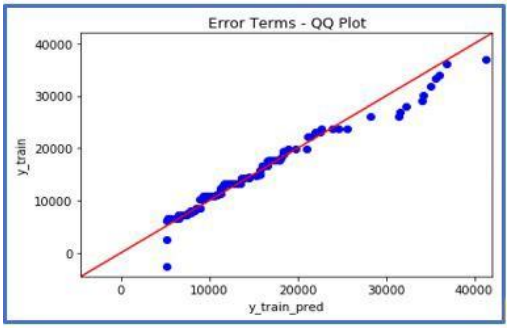
iii. have similar distributional shapes

iv. have similar tail behavior

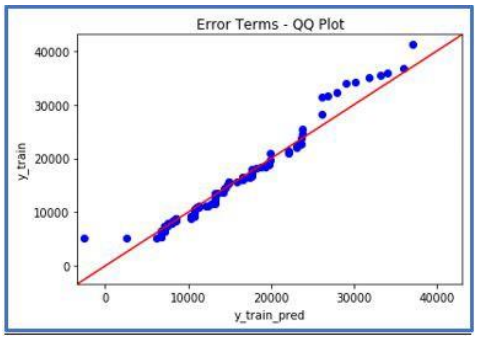
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.