Linear Regression Assignment

Assignment based subjective questions

Q 1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans Following insights found on the dependence of target variables w.r.t to categorical variables

- There is significant drop in target variable in spring season.
- Majority of bikes are rented in fall season.
- There is increasing trend in bikes rented in the month of January to September with dip in month of November & December.
- Majority of bikes are rented in the month of June to September.
- Majority of bikes are rented when there is no holiday.
- There is pretty slight variation in bikes getting rented among days.
- Bikes are rented more when there is clear weather with few clouds or partly cloudy.
- Bikes are rented when there are light snow and rain with scattered clouds.
- Bikes are not rented when there is heavy rain accompanied by thunderstorms and falling of ice pallets.
- Bikes are rented more in year 2019 compared to 2018

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

Ans While creating dummy variables the function will create variable of all categorical values present in that column. However, we can explain the values using one variable less and hence to carry out the same we use drop_first=True command so that the 1st dummy variable created gets dropped off automatically leaving with n-1 variables with us.

Let's say we have 3 types of values say Poor, Average, Good in Categorical column and we want to create dummy variable for that column. If one variable is not Poor and Average, then It is obvious Good. So, we do not need 3rd variable to identify the Good.

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

Ans Variable 'temp' or 'atemp' have highest correlation with target variable 'cnt' among the numerical variables present in the dataset.

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

Ans The basic assumptions of Linear Regression are: -

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other i.e., no multicollinearity
- Error terms have constant variance (homoscedasticity)

Hence after building the Final Model on training dataset, we do residual analysis where we calculate the errors between y_{actual} and $y_{predicted,}$ then checked the distribution of errors whether it is normal or not. Also, we checked the R-Square value on test dataset and observed whether it is close to training dataset.

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

Ans - As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **temp**: - A coefficient value of '0.5223' indicates that a unit increase in temp variable increases the bike hire numbers by 0.5223 units.

- **yr**: - A coefficient value of '0.2303' indicates that a unit increase in yr variable increases the bike hire numbers by 0.2303 units.

- **light snow and rain(weathersit-3):-** A coefficient value of '-0.3048' indicates that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3048 units

General subjective questions

Q 1 Explain the linear regression algorithm in detail.

Ans Linear Regression Model is a type of machine learning algorithm where the dependent variable is continuous in nature.

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).

It is used to analysis linear relationship between dependent variable and independent variable using a straight line

Mathematically the relationship can be represented with the help of following equation −

$$\mathbf{Y = \beta_0 + \beta_1 X}$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

β_1 is the slope of the regression line which represents the effect X has on Y

β_0 is a constant, known as the intercept. If X = 0, Y would be equal to β_0.

Based on the sign of β_1 we can say that there are there are two types of Linear Relationship: -

1. **Positive Linear Relationship**: - If sign of β_1 is positive means that with increase in dependent variable there is increase in independent variable.

2. **Negative Linear Relationship**: - If sign of β_1 is negative means that with increase in dependent variable there is decrease in independent variable.

Linear Relationship is of two types: -

1. **Simple Linear Relationship**: - When there is only 1 Independent Variable.

2. **Multiple Linear Relationship**: - When there are more than 1 Independent Variable.

Basic Assumptions of Linear Relationship: -

- **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.

- **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**: - Error terms have constant variance

Q 2 Explain the Anscombe's quartet in detail.

Ans It consists of 4 datasets that have nearly identical simple descriptive statistics yet have very different distributions. It was constructed by Francis Statistician Anscombe in Year 1973. The main purpose to create this is to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.

Dataset: -

Anscombe's quartet

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

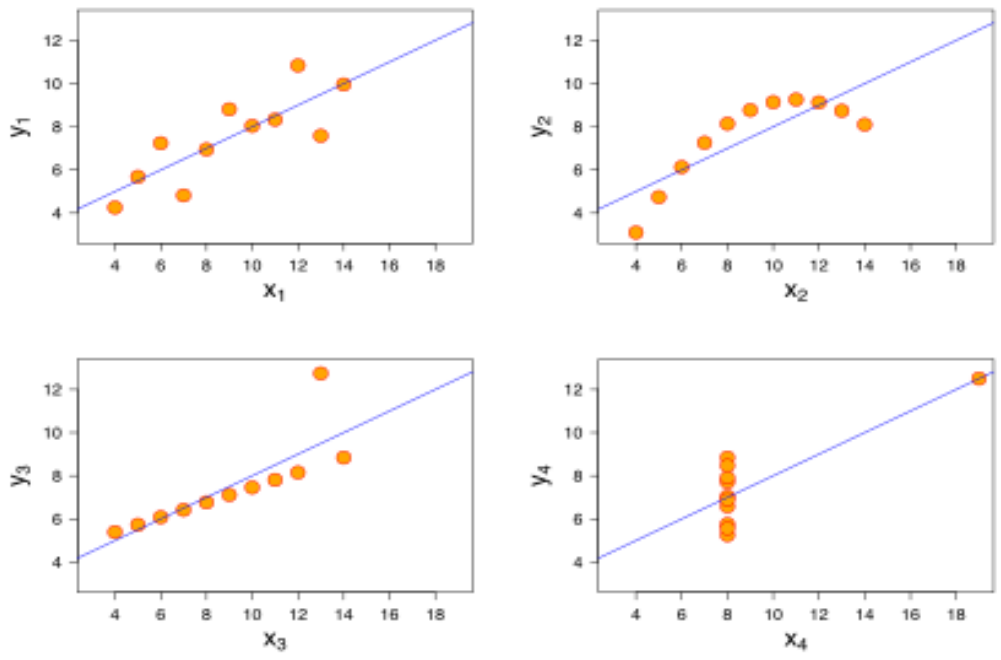
Statistics: -

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Inferences: -

- The Mean of x is 9 and mean of y is 7.5 for each dataset.
- The variance of x is 11 and y is 4.125 for each dataset.
- The correlation coefficient between x and y is 0.816 for each dataset.
- R Squared Value is 0.67 for each dataset.

Graphs: -



Observations on above four Plots: -

1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated.

2. The second graph (top right); while a relationship between the two variables is obvious, it is not linear.

3. In third graph (bottom left) the distribution is linear, but the calculated regression is thrown off by an outlier.

4. Finally, the fourth graph (bottom right) shows that one outlier is enough to produce a high correlation coefficient.

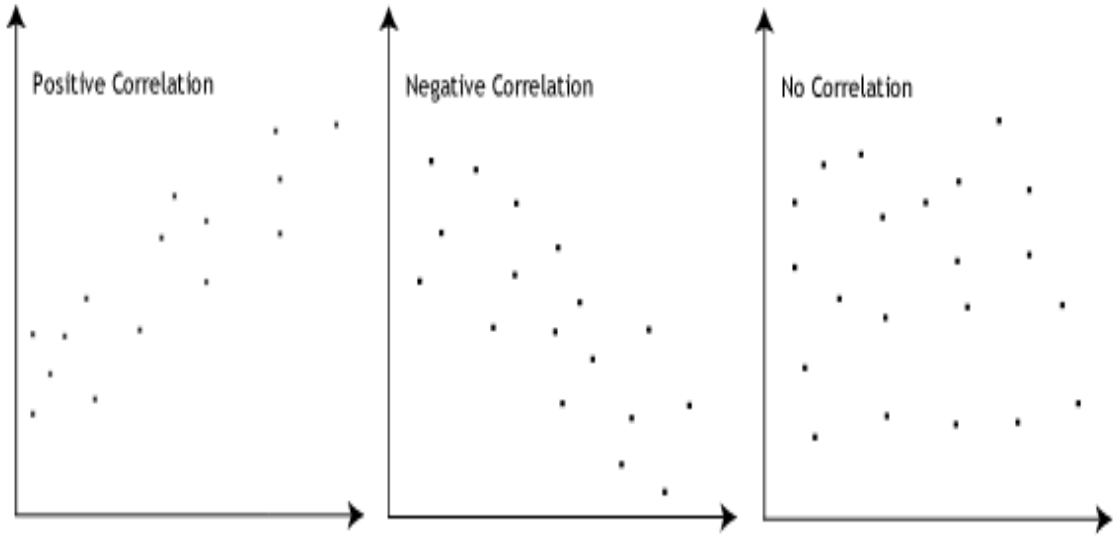
This quartet emphasises the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Q 3 What is Pearson's R?

Ans It is most common way of measuring a linear correlation It is the ratio between the covariance of two variables and the product of their standard deviations. It measures the strength and direction of relationship and usually lies between -1 and +1. It also measures how close the observations are to a line of best fit.

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:



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

Ans Scaling is a technique to standardize the independent features present in the data in a fixed range. When we have a lot of independent variables in a model, a lot of them might be on very different scales having highly varying magnitudes which will lead a model with very weird coefficients that might be difficult to interpret.

Let us say we have values like 5000 watts and 7kW, then the algorithm not using feature scaling will consider the value 5000 greater than 7 but in actual that is not true and in this case algorithm will give wrong predictions. So, we use feature scaling to bring all values to same magnitude and tackle this issue.

So, we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

There are two types of scaling: -

- Standardizing
- Min-Max Scaling

S No	Standardizing	Min-Max Scaling
1	Mean and Standard Deviation are used for Scaling	Maximum and Minimum Values of Dataset are used for Scaling
2	There is no particular range	Values lie between 0 and 1
3	It maintains information about outliers	It does not maintain any information about outliers
4	StandardScaler class of sklearn.preprocessing is used for standardization of features.	MinMaxScaler class of sklearn.preprocessing is used for normalization of features.

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

If there is perfect correlation, then VIF = ∞. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 7, this means that the variance of the model coefficient is inflated by a factor of 7 due to the presence of multicollinearity.

When the value of VIF is ∞, 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) = ∞. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

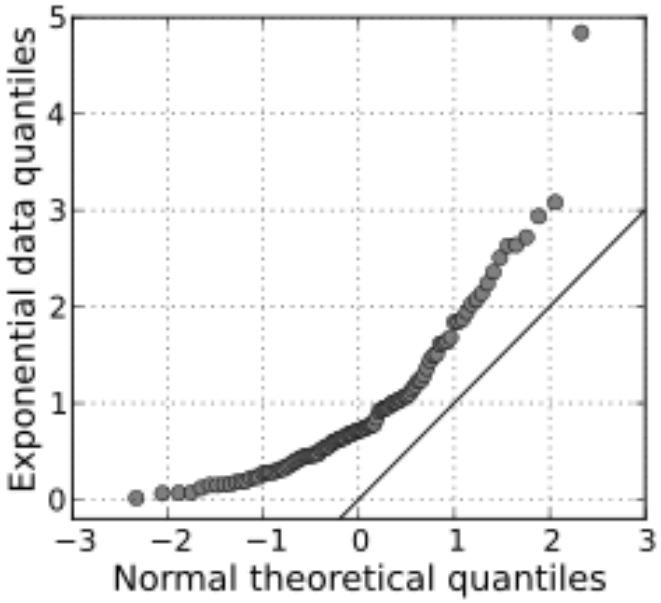
An ∞ VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an ∞ VIF as well).

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

Ans Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, 25% quantile is the point at which 25% percent of the data fall below and 75% fall above that value. The purpose of Q Q plots is to find

out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45-degree reference line:



Importance of Q-Q Plot: -

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.