# This document explains about ..

# Bike sharing system case study Inferences

# General Subjective question

# Bike sharing system case study Inferences

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

Inference from the analysis of categorical variables is as follows..

* season : season shows 32% of the bike booking were happening in season3 (fall) with a median of over 5000 booking. This was followed by season2 (summer) & season4 (winter) with 27% & 25% of total booking. Hence we can say that season can be a good predictor for the dependent variable.
* month : month shows 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. So, month shows some trend for bookings and it's also a good predictor for the dependent variable.
* weathersit : weathersit shows 69% of the bike booking were happening during ‘weathersit1 (Clear\_Few Clouds) with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% (Mist\_cloudy) of total booking. Weathersit also influence the bike bookings can be a good predictor for the dependent variable.
* holiday : holiday shows 98% of the bike booking were happening when it is not a holiday which means this data is clearly biased. Hence holiday is not a good predictor for the dependent variable.
* weekday : weekday variable shows very close trend . Hence can't say any particular day of the week has good or low number of bookings and also having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. The model should decide if this needs to be added or not.
* workingday: Almost 68% of the bike booking were happening on a workingday with a median of close to 5000 bookings. Hence workingday might be a good predictor for the dependent variable.

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

Importance of drop\_first=True:

* ‘drop\_first = True’ is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
* Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi\_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.
* Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables

Syntax for dummy variable creation **: df = pd.get\_dummies(df, drop\_first=True)**

This code does 3 things:

* + Create Dummy variable
  + Drop original variable for which the dummy was created
  + Drop first dummy variable for each set of dummies created.

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

At the pair-plot among all the numerical variables , atemp and temp has highest correlation with target variable.

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

Linear Regression Assumptions validation

Assumptions of simple linear regression

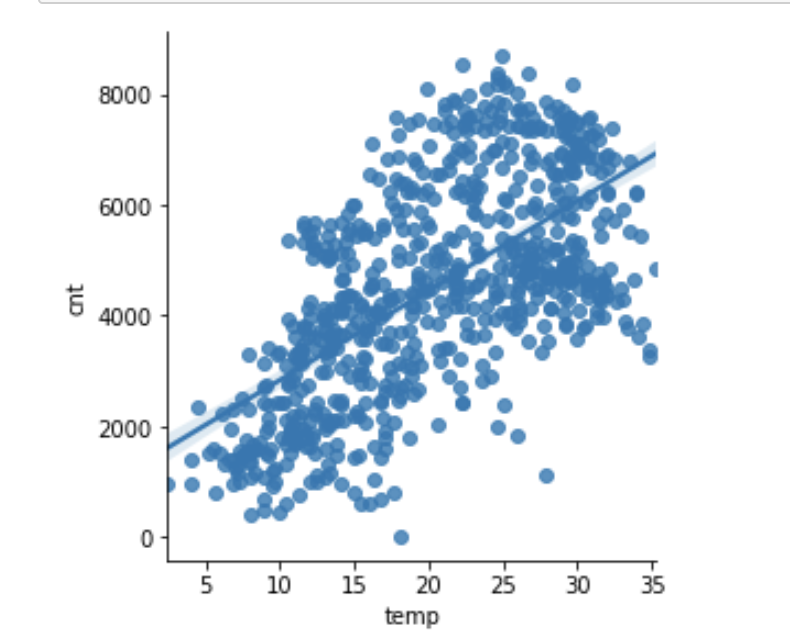
- Linear relationship between X and y.

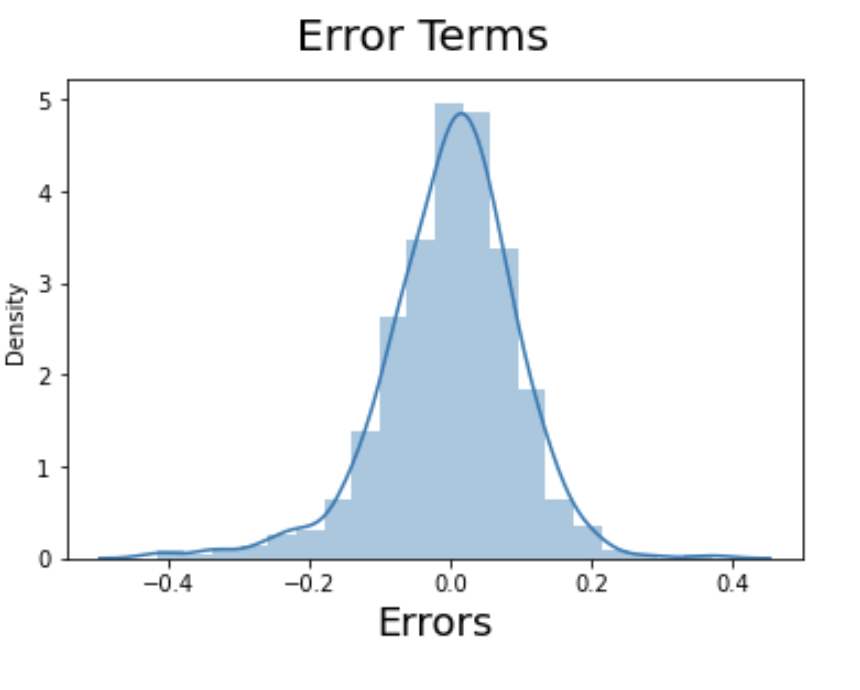
- Normal distribution of error terms.

- Independence of error terms.

- Constant variance of error terms (homoscedasticity).

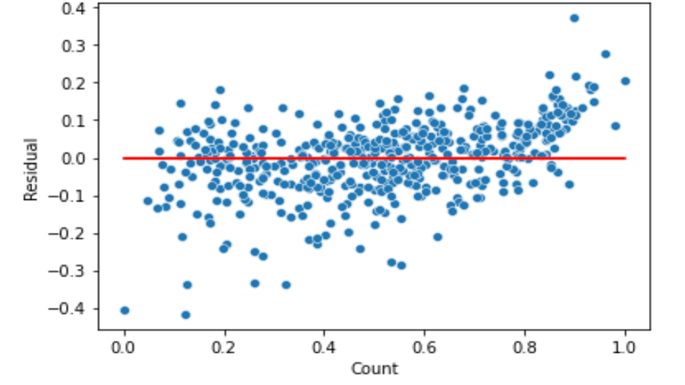
* Linear relationship between X and y
* Using the scatter plot checked the linear relationships of X with y .
* Also used the regression plot to check the relationship with the most important variable temp shows a linear relationship with cnt varaible.



* Normal distribution of error terms : Residual Analysis Of Training Data[¶](http://localhost:8888/notebooks/Documents/ML/Linear%20Regression%20on%20Bike%20Sharing%20data%20set%20.ipynb" \l "Residual-Analysis-Of-Training-Data)
* From the above histogram, we could see that the residuals are normally distributed. Hence our assumption for Linear Regression is valid.

### Independence of residuals

* Autocorrelation refers to the fact that observations’ errors are correlated. To verify that the observations are not auto-correlated, we can use the Durbin-Watson test. The test will output values between 0 and 4. The closer it is to 2, the less auto-correlation there is between the various variables.
* 0 – 2: positive auto-correlation 2 – 4: negative auto-correlation)
* Result : The Durbin-Watson value for Final Model lr 6 is 2.0311
* Constant variance of error terms (homoscedasticity):



* There is no visible pattern in residual values, thus homoscedacity is well preserved.

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

* Temperature (temp) - A coefficient value of ‘0.539040’ indicated that a unit increase in temp variable increases the bike rentals by 0.539040 units
* Weathersit\_Mist\_Cloudy (weathersit\_2) - A coefficient value of ‘-0.080930’ indicated that a unit increase in Weathersit\_Mist\_Cloudy variable decreases the bike rentals by 0.080930 units.
* Weather\_Light rain\_Light snow\_Thunderstorm (weathersit\_3) - A coefficient value of ‘-0.297362’ indicated that a unit increase in weathersit\_Light rain\_Light snow\_Thunderstorm variable decreases the bike rentals by 0.297362 units.
* Year (yr) - A coefficient value of ‘0.231263’ indicated that a unit increase in yr variable increases the bike rentals by 0.231263 units.
* holiday - A coefficient value of ‘0.100177’ indicated that a unit increase in holiday variable decreases the bike rentals by 0.100177 units. Hence bike rentals are more on working days than holidays

# General Subjective Questions

1. Explain the linear regression algorithm in detail.

* Linear Regression is a machine learning algorithm which falls under supervised learning methods - in which you have the previous years’ data with labels and you use that to build the model.
* It finds abest linear-fit relationship on any given data, between independent (Predictors) and dependent (Target) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared ResidualsMethod.

Linear regression is of the 2 types:

* Simple Linear Regression SLR:It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

SLR equation **:** Y=β0+β1X1 +ϵ

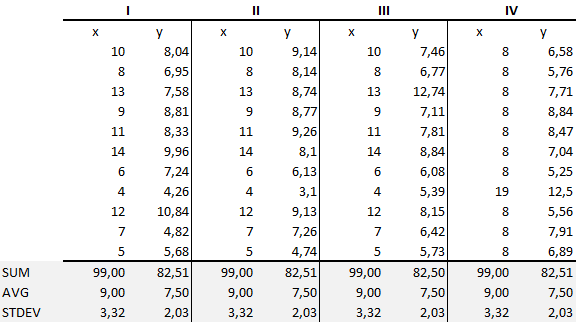
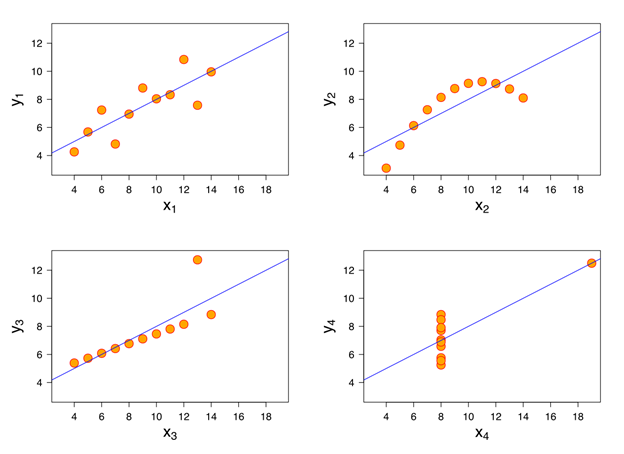
* Multiple Linear Regression MLR: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a ‘hyperplane’ instead of a straight line.

MLR equation **:** Y=β0+β1X1+β2X2+…+βpXp+ϵ

* The equation of the best fit regression line Y = β₀ + β₁X can be found by the following two methods:
  + Differentiation
  + Gradient descent

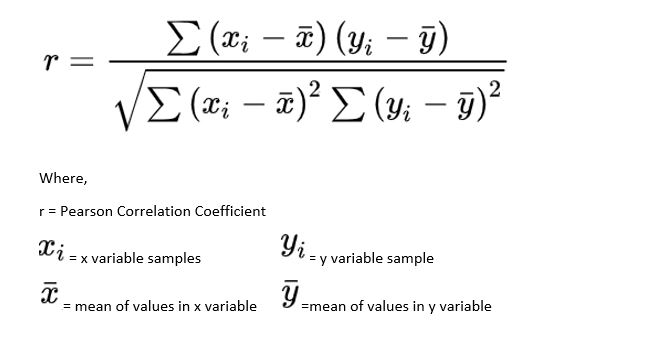
We can use statsmodels or sklearn libraries in python for the linear regression.

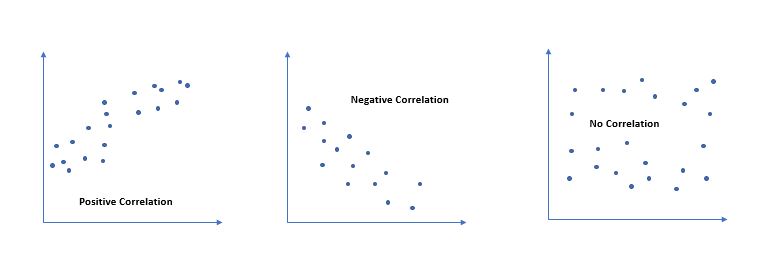
1. Explain the Anscombe’s quartet in detail.

* Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.
* Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.
* 
* 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.
* 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 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.
* Additionally, Anscombe’s Quartet warns of the dangers of outliers in data sets.

3.What is Pearson’s R?

* Correlation means to find out the association between the two variables.
* Correlation coefficients are used to find out how strong the is relationship between the two variables.
* The most popular correlation coefficient is Pearson’s Correlation Coefficient and is represented as ‘r’
* It measures how strong is the linear association between two continuous variables using the formula:



* Value of ‘r’ ranges from ‘-1’ to ‘+1’.
* Value ‘0’ specifies that there is no relation between the two variables.
* A value greater than ‘0’ indicates a positive relationship between two variables where an increase in the value of one variable increases the value of another variable.
* Value less than ‘0’ indicates a negative relationship between two variables where an increase in the value of one decreases the value of another variable.
* 

Assumptions for a Pearson Correlation:

* Data should be derived from random or least representative samples, draw a meaningful statistical inference.
* Both variables should be continuous and normally distributed.
* There should be Homoscedasticity, which means the variance around the line of best fit should be similar.
* Extreme outliers influence the Pearson Correlation Coefficient. You need to consider outliers that are unusual only on one variable, called as ‘univariate variable’ or for both of the variables known as ‘multivariate outliers’. 2 variables are measured independently from each other pairs.
* Example : If we plot age vs amount then, we can certainly, see that there is a correlation between the age of a person and loan the amount is given to that person, as age increases the loan amount given to the person decreases and vice versa. But if we plot the loan amount vs age, it is not possible to draw any conclusion from it. It would violate the assumption.

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

Scaling : It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm

Reason to perform scaling

* Most of the times, collected data set contains features highly varying in magnitudes, units and range.
* If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
* Note : Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization (Min-Max scaling)

* It brings all the data in the range of 0 and 1.
* Formula for normalization : X` = (X – Xmin) / (Xmax – Xmin)

If the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0

* If the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1

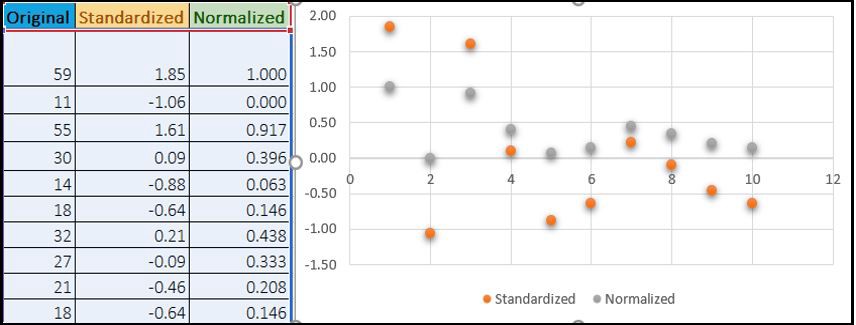
Standardization

* Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
* Formula for standardization: X - mean(X)/sd(X)
* X is the mean of the feature values and sd is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

Major differences between Normalization and Standardization

* Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution.
* This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
* Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution.
* Unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.
* A good practice to fit the scaler on the training data and then use it to transform the testing data. This would avoid any data leakage during the model testing process. Also, the scaling of target values is generally not required.
* One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Example of standardized and normalized data..



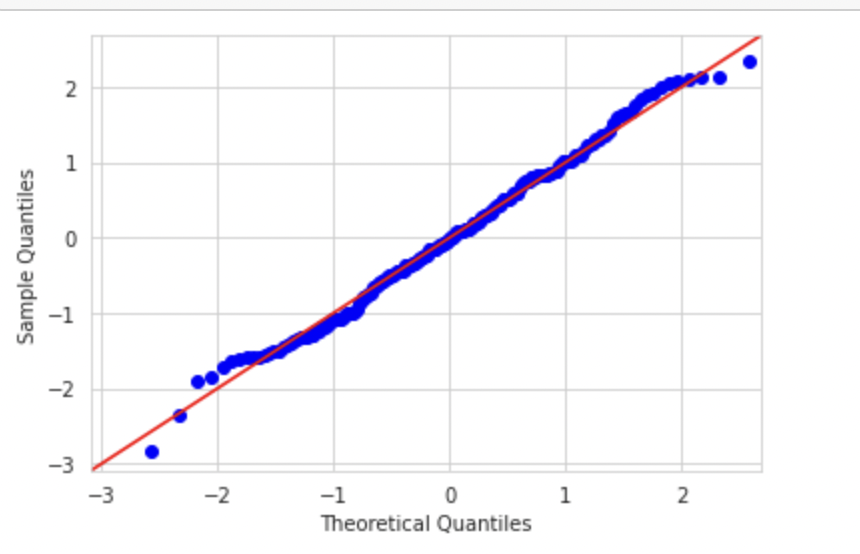
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

* VIF stands for Variance inflation factor . VIF = 1 / (1 - R^2))
* It is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.
* VIF score of an independent variable represents how well the variable is explained by other independent variables .
* VIF starts at 1 and has no upper limit.
* VIF = 1, no correlation between the independent variable and the other variables.
* VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others

VIF = Infinity

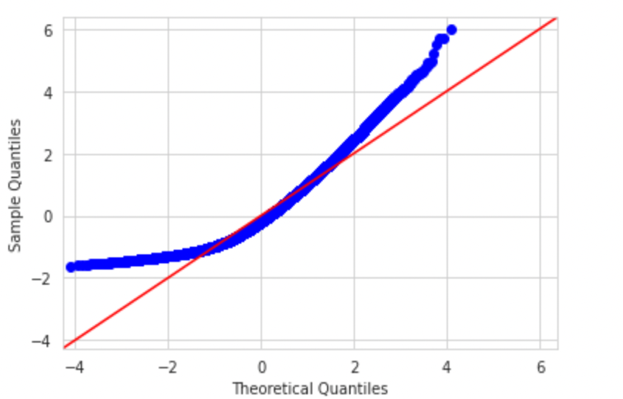
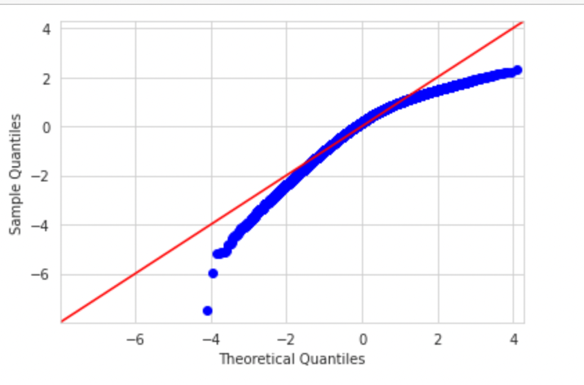
* If there is perfect correlation, then VIF = infinity.
* R^2 value is determined to find out how well an independent variable is described by the other independent variables.
* High value of R^2 means that the variable is highly correlated with the other variables.
* In the case of perfect correlation between two independent variables, we get R2 =1, which lead to 1/(1-R2) infinity.
* To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
* Hence an infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (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.

* Q-Q plots are also known as Quantile-Quantile plots.
* They plot the quantiles of a sample distribution against quantiles of a theoretical distribution.
* Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.
* If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.
* 

Use and importance :

* Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.
* The power of Q-Q plots lies in their ability to summarize any distribution visually.
* If two populations are of the same distribution
* If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
* Skewness of distribution .If the see the left side of the plot deviating from the line, it is left-skewed. When the right side of the plot deviates, it’s right-skewed.



Left skewed Right skewed