Linear Regression

Kedar Lachke

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A.1.

There are a total of 7 categorical variables for the bike sharing data.

yr: There is a slight increase in the year 2019 than in 2018.

mnth: There in an increase in bike sharing from month March to October.

holiday: There is a slight decrease bike sharing on holidays.

weekday: After observing the chart, bike sharing is almost the same throughout the week.

Q2. Why is it important to use drop_first=True during dummy variable creation

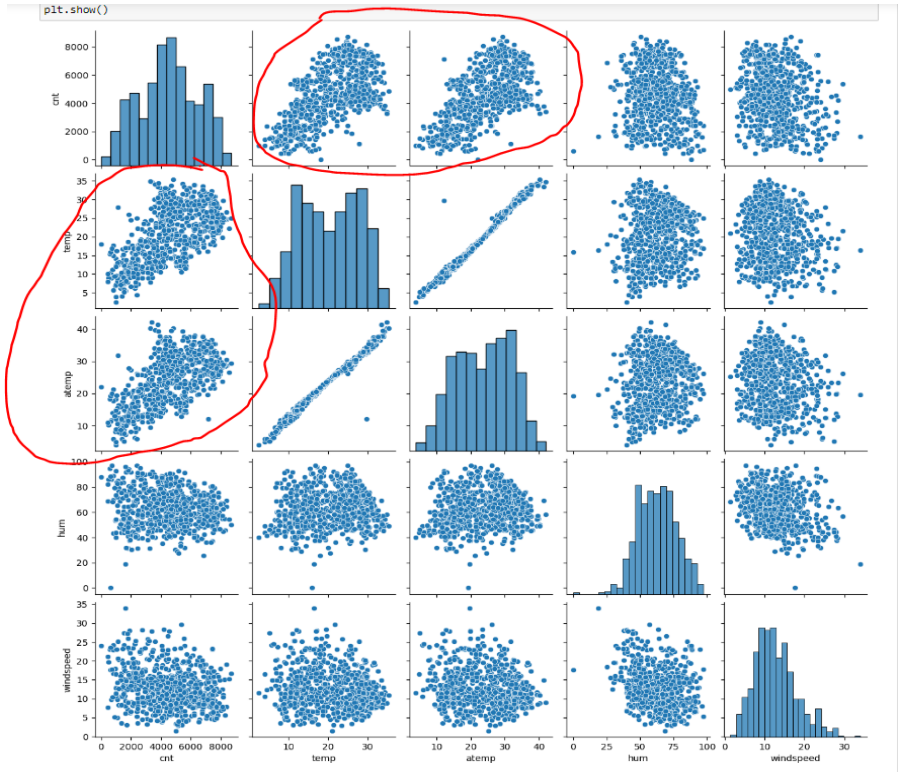
- A2.
- If the category variable has 'n' variations, then n-1 dummy variable are needed
- For example, consider there is a categorical variable blood group having three variations 'A','B' and .C'.
- So the number dummy variables will be (n-1) **3 - 1 = 2** (2 dummy variabbles)
- A B
- 1 0 => will indicate A Group
- 0 1 => will indicate B Group
- **0 0 => will indicate O group**

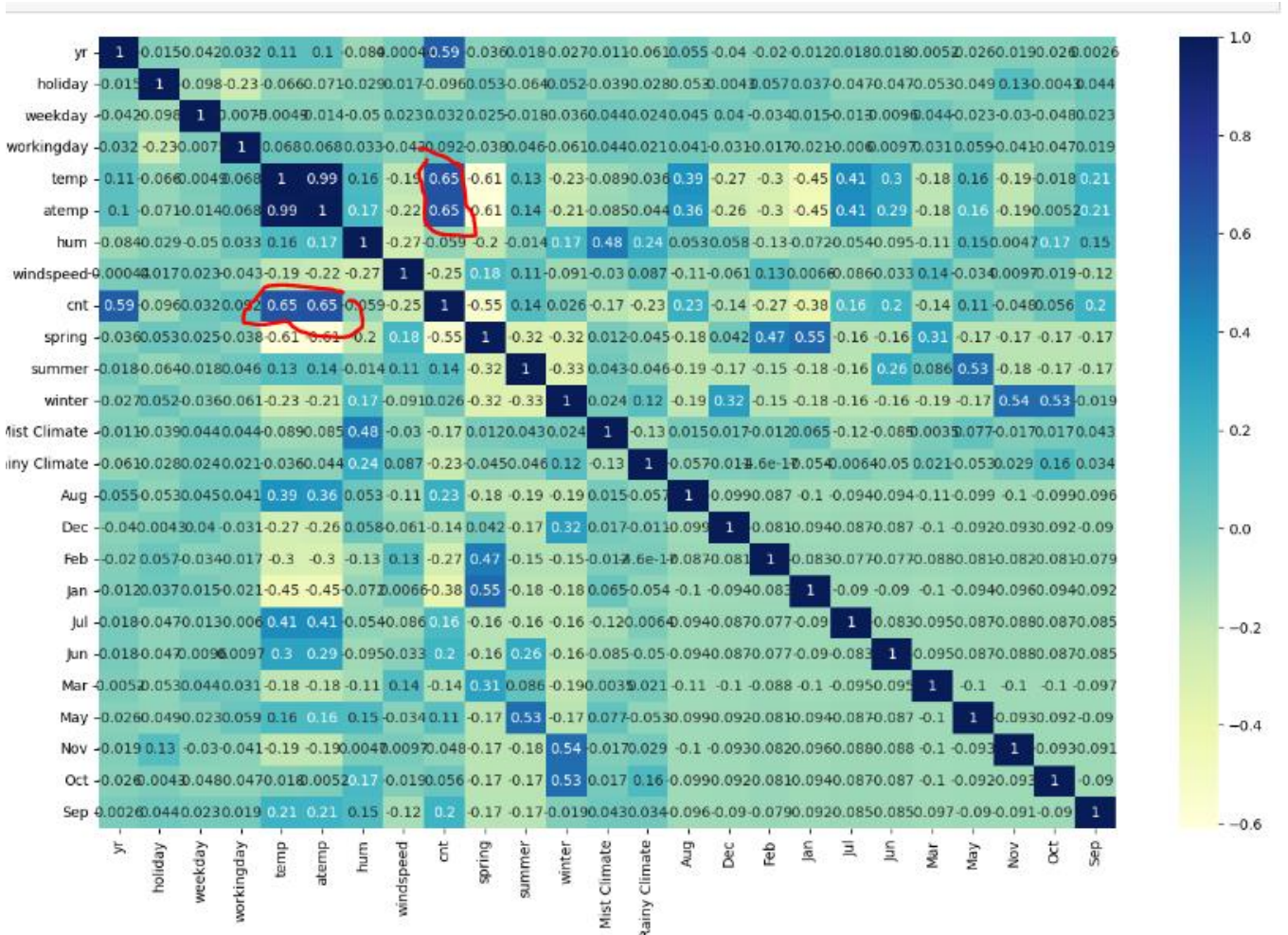
So there will be only two columns for 'A' and 'B' Group in excel. 'O' group column will not be there.

We'll consider that if 'A' and 'B' columns has '0' values, it will indicate 'O' blood Group.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- Looking at the pair plot, temp and atemp have the highest correlation with the target variable cnt (0.65)



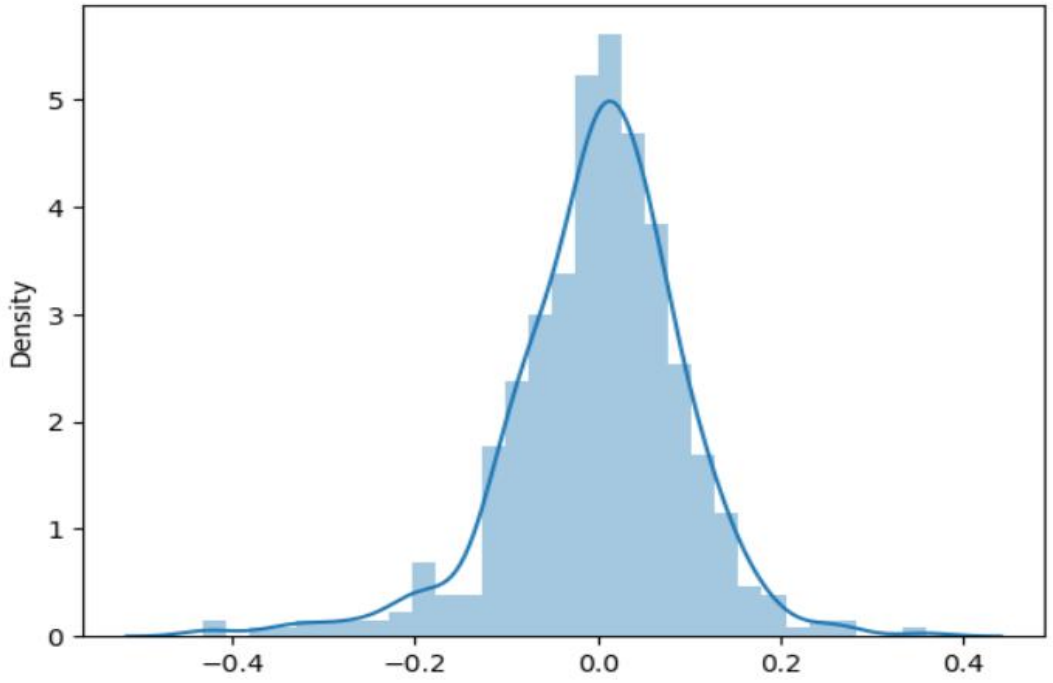


Q.4
How did you validate the assumptions of Linear Regression after building the model on the training set

A4.

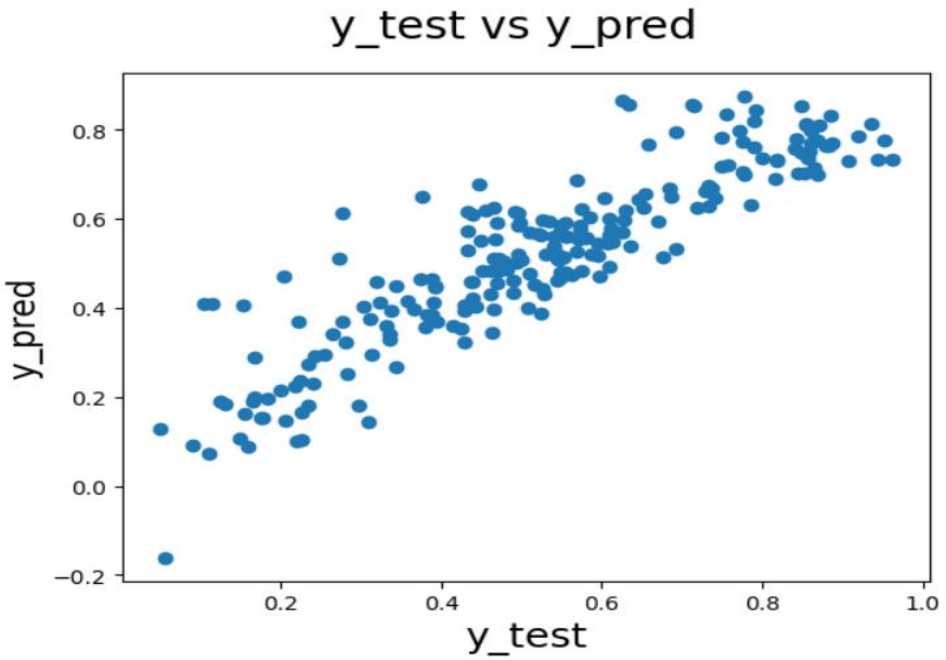
1. First, we check we checked if the error terms are normally distributed

   For that, we have taken the difference between y_train values and y_pred (predicted) and then plotted on the distribution graph



2.

A4. Continued.....



2. Check if and patter is getting formed in the graph

As it is observed that no patterns are getting formed and there are some outliers are present

A4. Continued.....

- Check for RSME value which should be close to 0

In the module the that has been finalized is **0.09923852498067227**

```
In [178]: #Returns the mean squared error; we'll take a square root
          np.sqrt(mean_squared_error(test_y, test_y_pred))

Out[178]: 0.09923852498067227
```

- Check the R_squared value on the test data

Its value should be close the R_squared value of train data

Checking R_square value on test data

```
In [179]: r_squared = r2_score(y_true=test_y, y_pred=test_y_pred)
          r_squared

Out[179]: 0.7940304665931015
```

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	282.1
Date:	Wed, 09 Aug 2023	Prob (F-statistic):	4.11e-180
Time:	20:24:09	Log-Likelihood:	473.83
No. Observations:	510	AIC:	-929.7
Df Residuals:	501	BIC:	-891.6
Df Model:	8		
Covariance Type:	nonrobust		

Q5.
Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Based on the model,
- 'Temp', 'Year' and 'Windspeed' are three features contributing to the demands of the bike sharing.

General Subjective Questions

Q6.
Explain the linear regression algorithm in detail.

A6.

Linear regression is used to find the relation between a dependent variable and Independent variables. It is a process of estimating the relationship between variables. It explains the change in a dependent variable with the changes in the values of predictors. Regression analysis is mainly used are Forecasting and predictions. Linear regression shows a correlation between independent variables and dependent variable and not causation. Linear Regression is a type of parametric regression.

There are two type of linear regression

1. Simple Linear Regression

In simple linear regression, only one independent variable value is changed, and the other independent variables values are kept constant, and then changes are observed in the independent variable.

2. Multiple linear regression

In multiple linear regression, only many independent variable value is changed and then changes are observed in the independent variable.

Linear Regression guarantees interpolation and not extrapolation.

Interpolation means using the model to predict the value dependent variable on independent values that lies within the data ranges that are already available.

There are some assumptions in linear regression that we need to keep in mind while building a regression model.

- There are some assumptions in linear regression that we need to keep in mind while building a regression model.

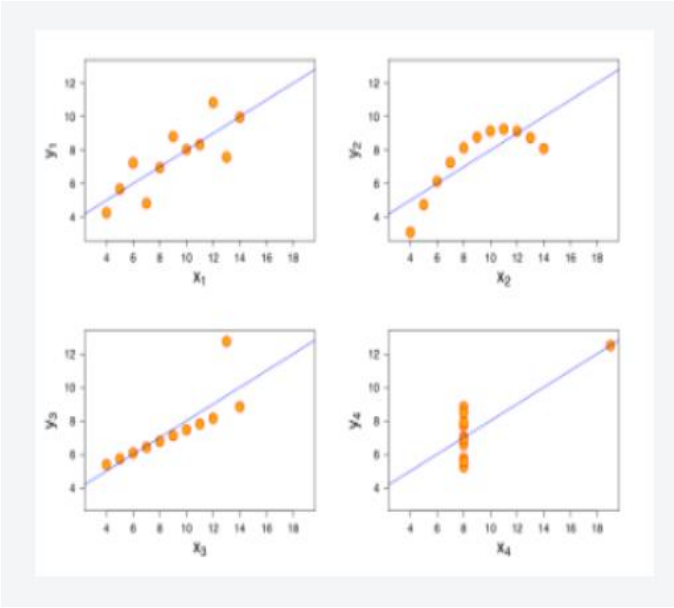
1. There is linear regression between Dependent variable and Independent valiables
2. Error Terms or Residual are normally distributed
3. There is no multi collinearity between dependent variables
4. The residual term has the same variance

Q7. Explain the Anscombe's quartet in detail.

- For linear regression to work, we should first have a good visualization of data as linear regression has some shortcomings

1. It is sensitive to outliers
2. It is modeled on data that has a linear relationship
3. A few assumptions are required to make inferences

- For explaining these shortcoming there is a concept called Anscombe's Quartet.



As we see in the graph, the linear regression is exactly the same. But there are some records in the data that are fooling the Linear regression line.

1. First graph it working fine, there seem to no issue with it.
2. Second Graph, there is no linear relationship between the independent variable and dependent variables
3. Third and fourth graphs, there are some outliers that are affecting the linear regression. If the outliers would have not been there, the best-fit line would have passed perfectly through the points.
4. That Is why we should have a good look at the data before going forward with linear regression model.

Q8.
What is Pearson’s R?

- The Pearson's R-value tell us if the two variable are moving in the same direction or opposite direction
- It values ranges between –1 and +1
- -1 indicate the two variables are negatively correlated, which means if one variables value increases the other variable's value decreases.
- 1 indicate the two variables are positively correlated, which means if the value of one variable increases the other variable's value also increases
- 0 indicate that there is no correction between them

Q9.
What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- A.9
- Scaling or feature scaling is the process of normalizing the range of a data set.
- In actual data sets, often there might be features that have varying degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

1. Normalized Scaling:

In normalized scaling the scaled values are between 0 and 1.

Whenever the data is not normally distributed, use Normalized Scaling

values	Feature scaling
1234	0.87
335	0.3
56	0.1

2. Standardization Scaling:

In standardization scaling, after data rescaling the mean is 0 and standard deviation is 1

Whenever the data is not distributed, use Standardization Scaling

Q.10
You might have observed that sometimes the value of VIF is infinite. Why does this happen

- When the correlation between two variables is R=1, then the VIF Value goes to zero.

VIF = 1/1-1

VIF = 1/0 = Infitity

$$VIF_i = \frac{1}{1 - R_i^2}$$

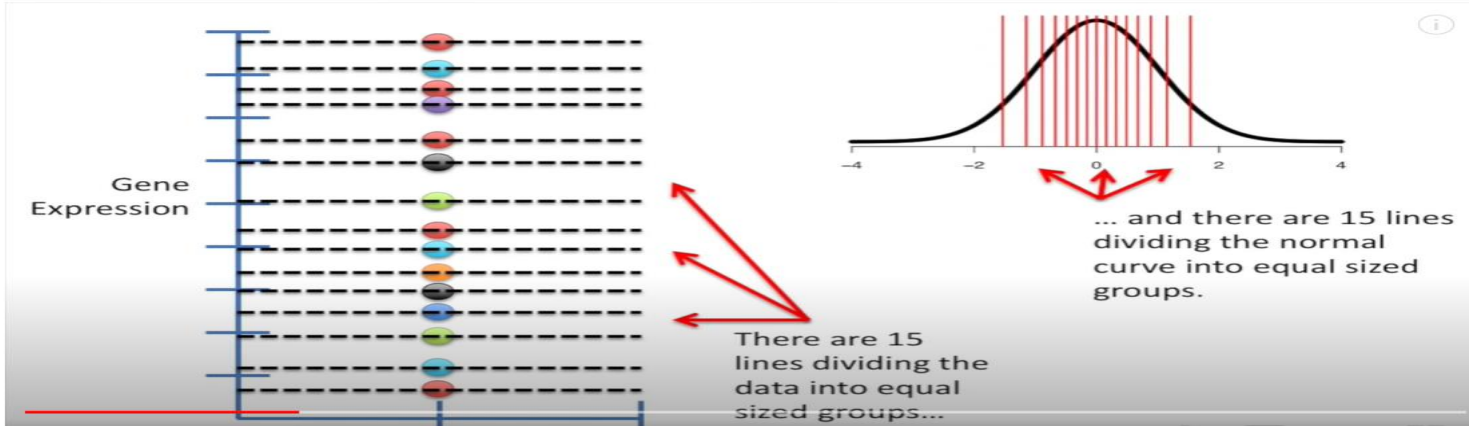
Q11.
What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- To check which distribution fits our data, we use Q-Q plot graph or Quantile-Quantile plot.
- There are two types of data distribution

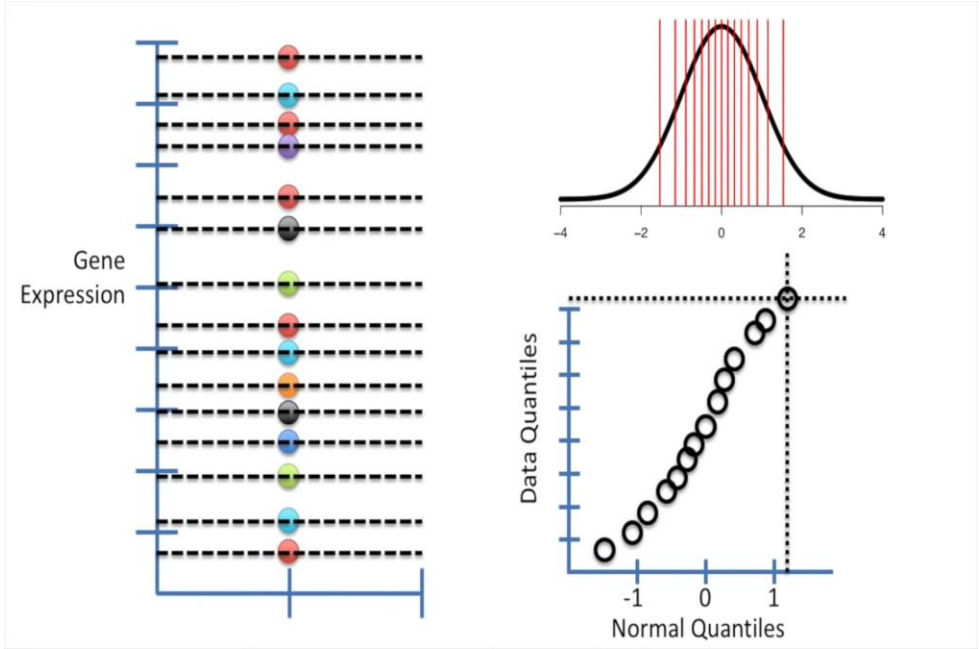
1. Normal Distribution
2. Standard distribution

Quantile is nothing but a grouping of data within a range. In all, there can be max 100 equal quantiles. Each quantile can be converted into a percentile.

Consider that each data point is quantile in our data. Now plot the normal distribution for our data and the same number of quantiles on the normal distribution as shown in the image below.



Now plot all the quantile distribution on the y-axis and the corresponding normal distribution quantile on the x-axis as shown in the first image

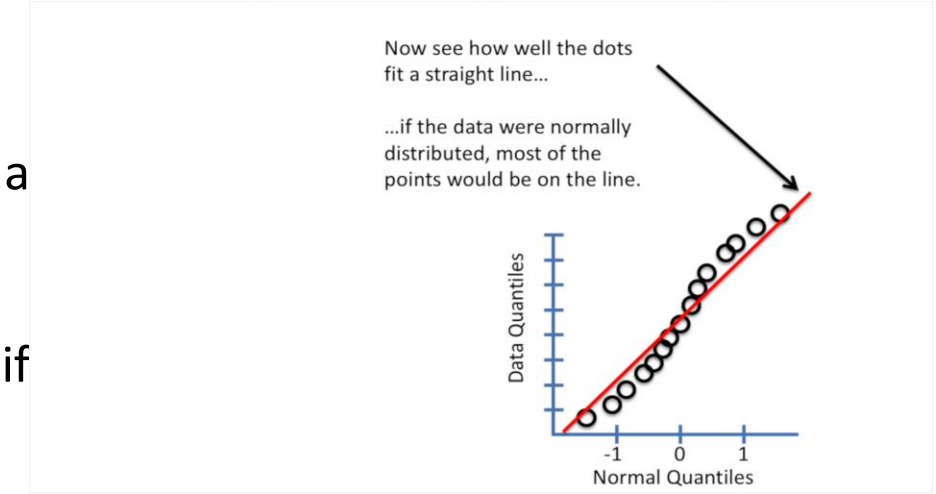


Now draw a best-fit line for the graph

We observe that the best-fit line is not passing through a plotted points.

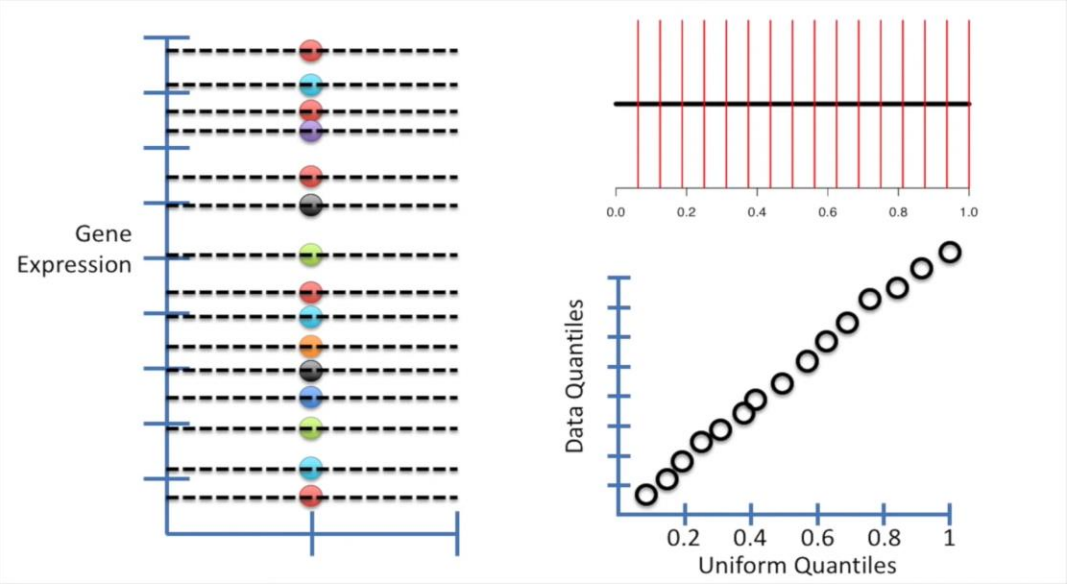
In this case, the best fit line is not awesome or not signif

So we should compare our data in another distribution



- Let compare the data on standard distribution
- Consider that each data point is quantile in our data. Now plot the Standard distribution for our data and the same number of quantiles on the standard distribution as shown in the image below.

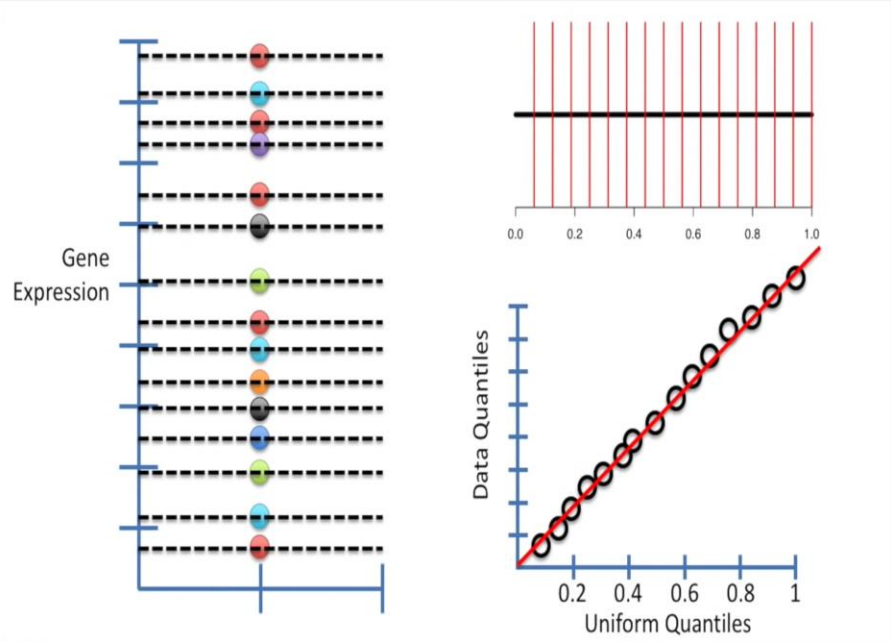
- Now plot all the quantile distribution on the y-axis and the corresponding normal distribution quantile on the x-axis as shown in the image



Now draw a best-fit line for the graph

We observe that the best-fit line is passing through all the plotted points.

In this case, the best-fit line is awesome or significant.



After observing the graph we observe that our data bits that uniform distribution better.