Tests for Normality (Shapiro wilk test)

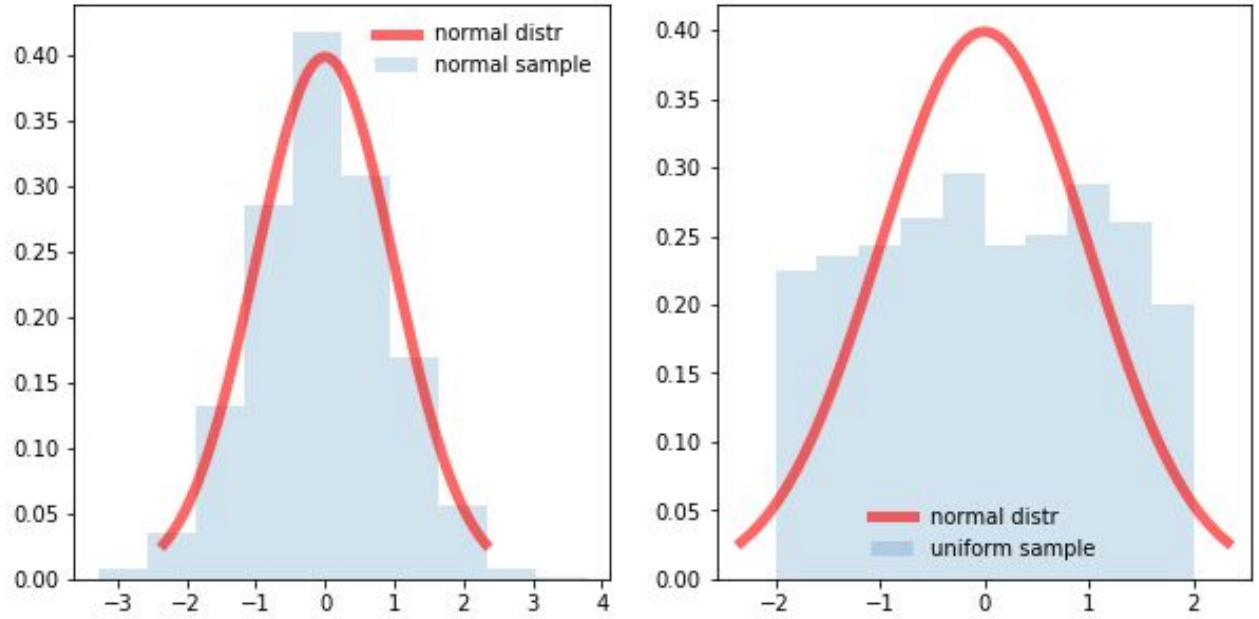
By-
Animesh Sahu
20161028

Visualizing Techniques

There are some techniques that exists to help us visualize our data and further use them to back up the results from the tests.

Histogram

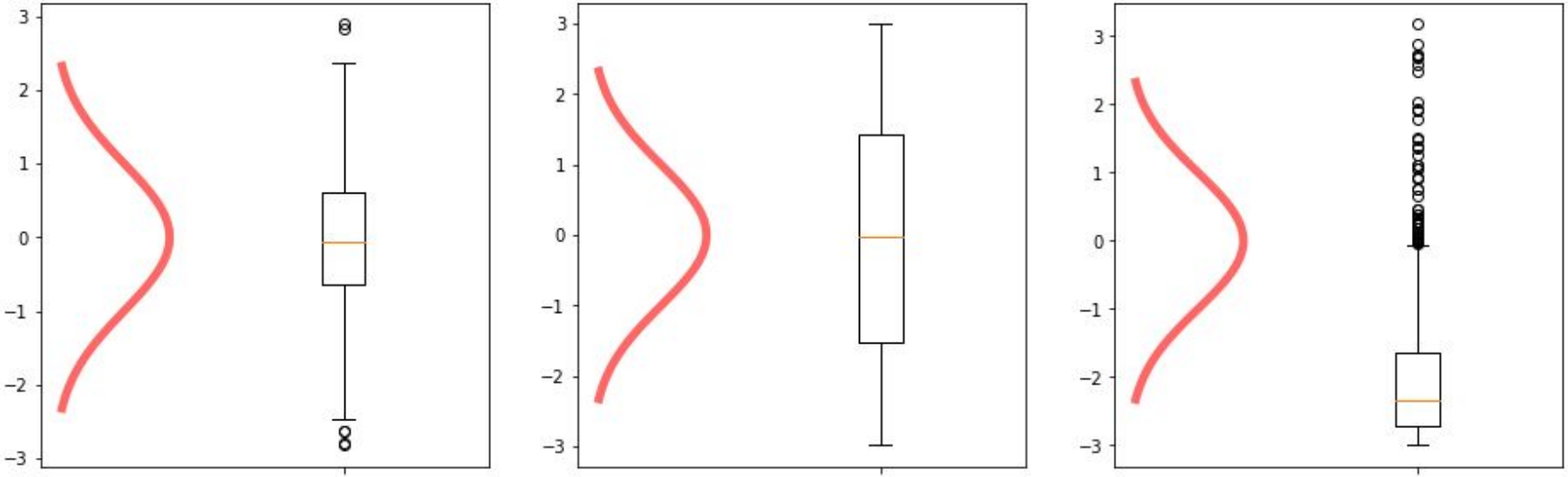
The histogram is a data visualization that shows the distribution of a variable. It gives us the frequency of occurrence per value in the dataset, which is what distributions are about.



- On the left, there is very little deviation of the sample distribution (in grey) from the theoretical bell curve distribution (red line).
- On the right, we see quite a different shape in the histogram, telling us directly that this is not a normal distribution.

Box Plot

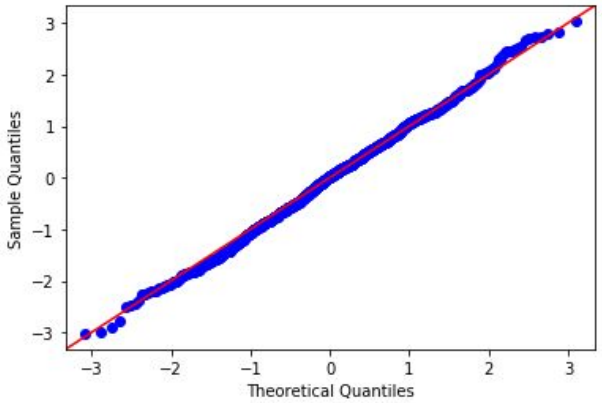
The Box Plot plots the 5-number summary of a variable: minimum, first quartile, median, third quartile and maximum.

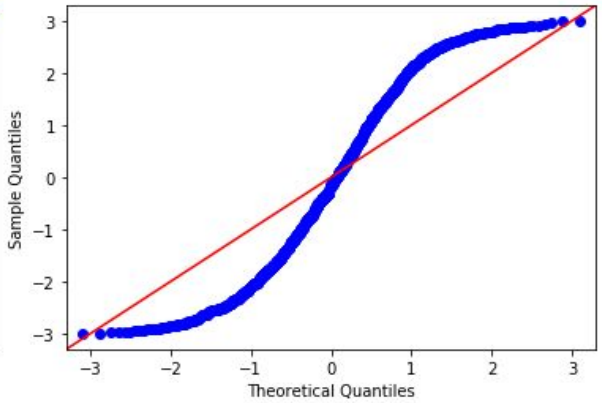


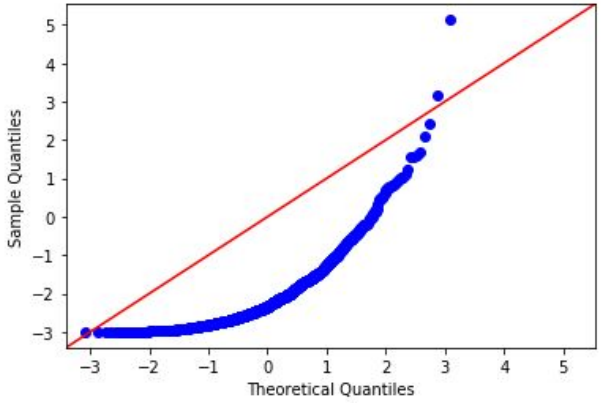
Normal (left), Uniform (middle) and exponential (right) boxplots vs normal bell curve

QQ Plot

QQ Plot stands for Quantile vs Quantile Plot, which is exactly what it does: plotting theoretical quantiles against the actual quantiles of our variable.







- The uniform distribution has too many observations in both extremities (very high and very low values).
- The exponential distribution has too many observations on the lower values, but too little in the higher values.

Shapiro wilk test

The Shapiro Wilk test is the most powerful test when testing for a normal distribution. It has been developed specifically for the normal distribution and it cannot be used for testing against other distributions

$\mathbf{H_0}$:The sample is drawn from normally distributed population.

$\mathbf{H_1}$:The sample is drawn from a population that is NOT normally distributed.

- Compare p-value to some pre-defined threshold.
- We will use significance level of α=0.05.
- We compare p-value to 0.05
- If p value is less than threshold we reject the null hypothesis.
- If p value is greater than threshold we FAIL to reject the null hypothesis.

Basic Approach

- Rearrange the data in ascending order so that $x1 \leq \ldots \leq xn$.
- Calculate SS as follows: $SS = \sum_{i=1}^{n}(x_i - \bar{x})^2$
- If n is even, let $m = n/2$, while if n is odd let $m = (n-1)/2$
- Calculate b as follows, taking the ai weights from the Table 1 (based on the value of n) in the Shapiro-Wilk Tables. Note that if n is odd, the median data value is not used in the calculation of b.

$$b = \sum_{i=1}^{m} a_i(x_{n+1-i} - x_i)$$

Basic Approach

- Calculate the test statistic $W = b^2/SS$
- Find the value in the Table 2 of the Shapiro-Wilk Tables (for a given value of n) that is closest to W, interpolating if necessary. This is the p-value for the test.

For example, suppose $W = .975$ and $n = 10$. Based on Table 2 of the Shapiro-Wilk Tables the p-value for the test is somewhere between .90 ($W = .972$) and .95 ($W = .978$).

Table 1 – Coefficients

n	2	3	4	5	6	7	8	9	10	11	12	13	14
a1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251
a2			0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318
a3					0.0875	0.1401	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460
a4							0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802
a5									0.0399	0.0695	0.0922	0.1099	0.1240
a6											0.0303	0.0539	0.0727
a7													0.0240

Table 2 – p-values

n \ P	0.01	0.02	0.05	0.1	0.5	0.9	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988

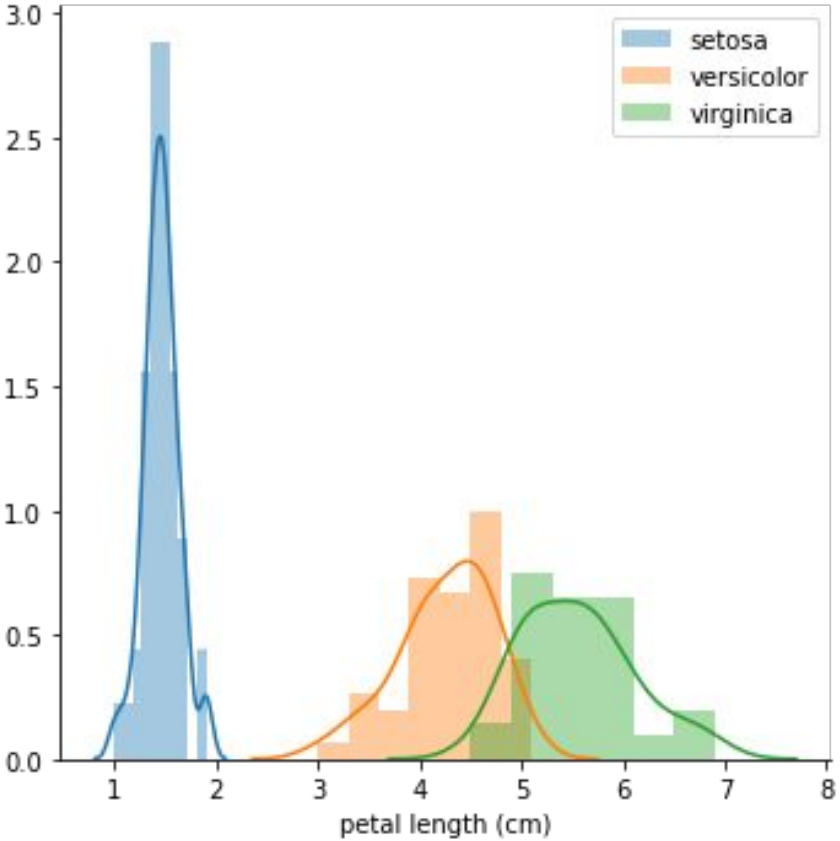
Example 1

- I have used IRIS dataset for normality check, it is used to classify three species(dependent variable) setosa, versicolor, virginica.
- We have 4 independent variable (petal length, petal width, sepal length,sepal width).
- I will do normality test on one of the independent variable (petal length) for each species.
- I will use various data visualization techniques to backup my results.

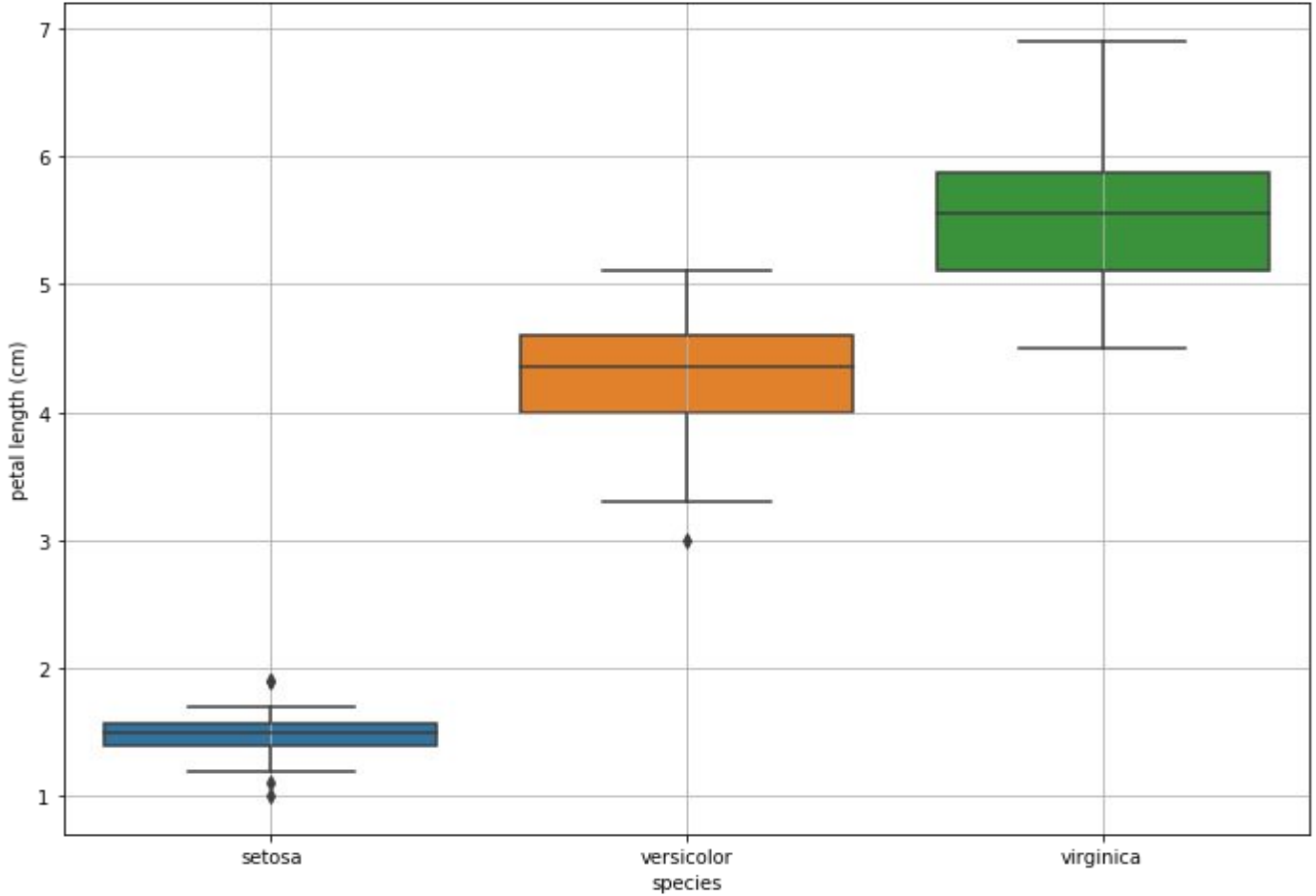
	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa

Dataset

- P value for petal length in setosa species.
  - `p value: 0.05481043830513954`
  - Since p value is > threshold we fail to reject null hypothesis
  - Normal distribution
- P value for petal length in virginica species.
  - `p value: 0.10977369546890259`
  - Since p value is > threshold we fail to reject null hypothesis
  - Normal distribution
- P value for petal length in versicolor species.
  - `p value: 0.1584833413362503`
  - Since p value is > threshold we fail to reject null hypothesis
  - Normal distribution

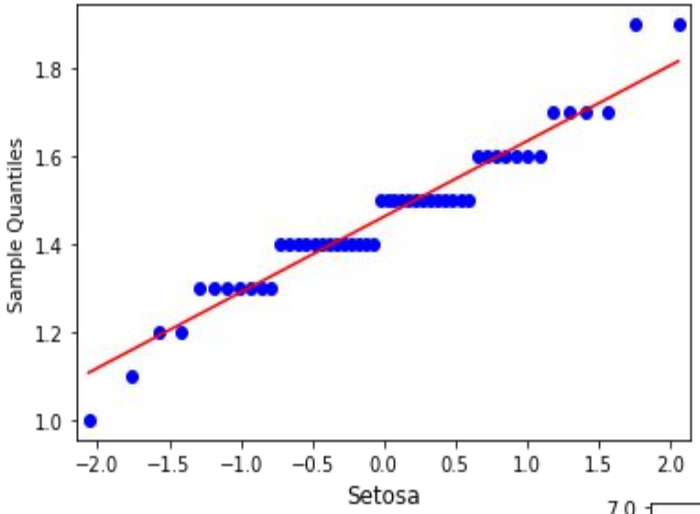

setosa
versicolor
virginica
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8
petal length (cm)

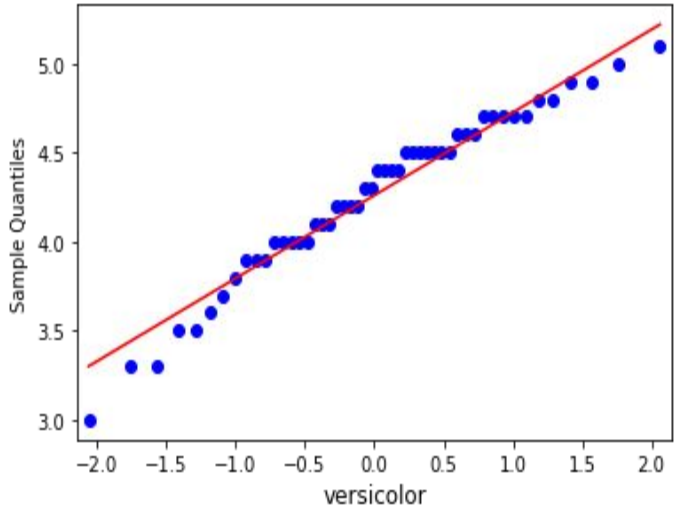
Histogram

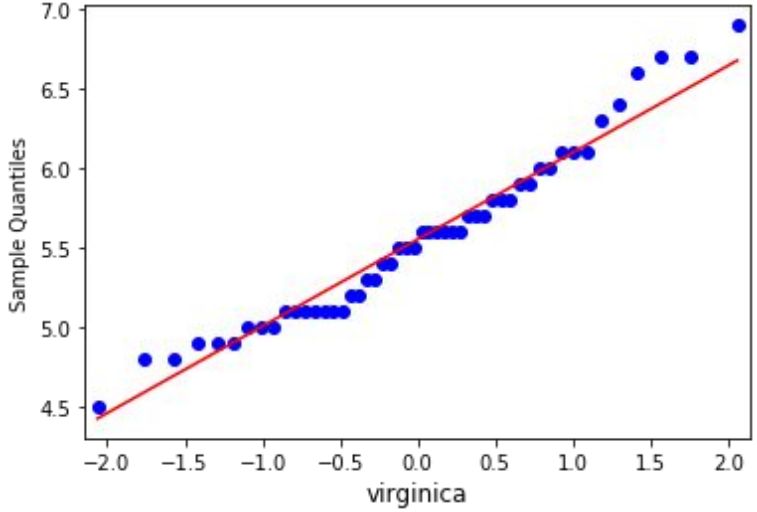

7
6
5
4
3
2
1
petal length (cm)
setosa
versicolor
virginica
species

Box Plot

QQ-Plot


Sample Quantiles
Setosa


Sample Quantiles
versicolor


Sample Quantiles
virginica

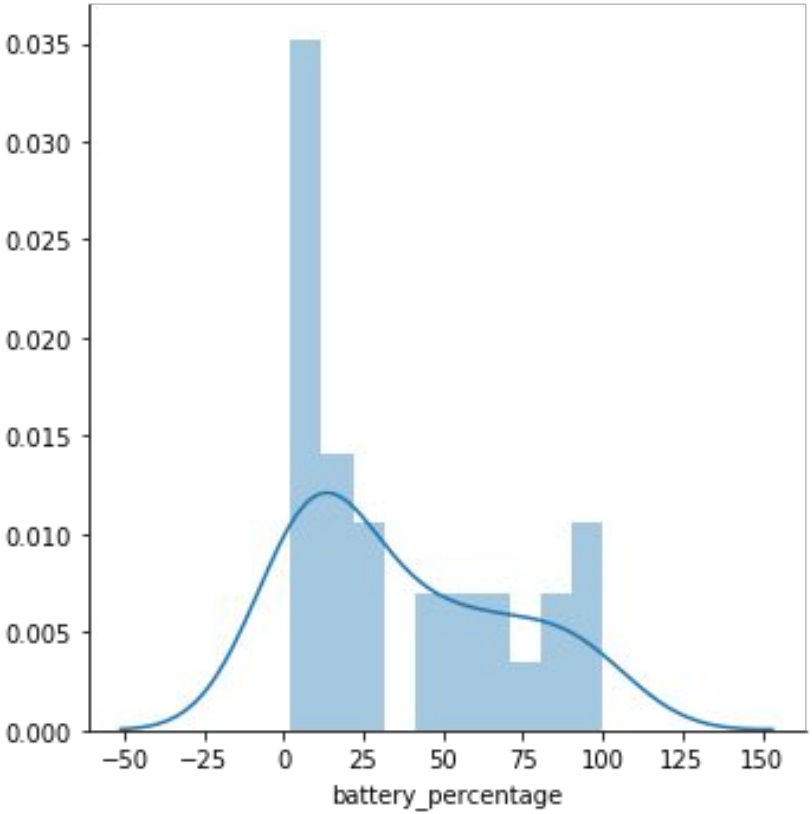
Example 2

- I have used dataset time vs battery percentage(for xiaomi mobiles) for normality check.
- On one column we have time in mins and on other column corresponding battery percentage.
- I will do normality test on the battery percentage.
- I will use various data visualization techniques to backup my results.

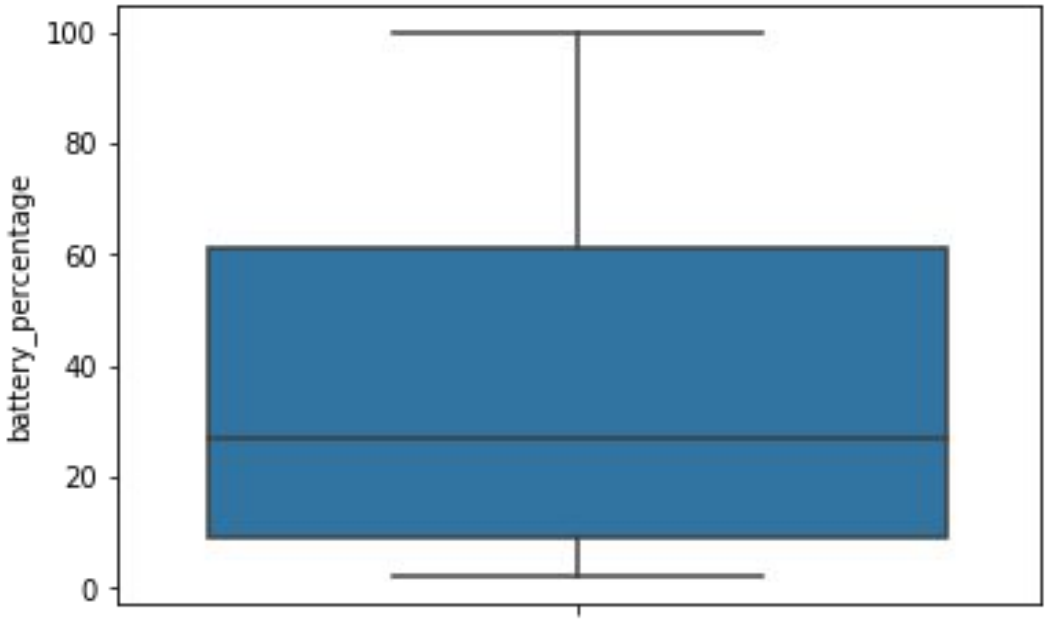
	Time(min)	battery_percentage
0	0	100
1	10	97
2	20	91
3	30	87
4	40	81
5	50	78
6	60	66
7	70	61
8	80	57
9	90	51

Dataset

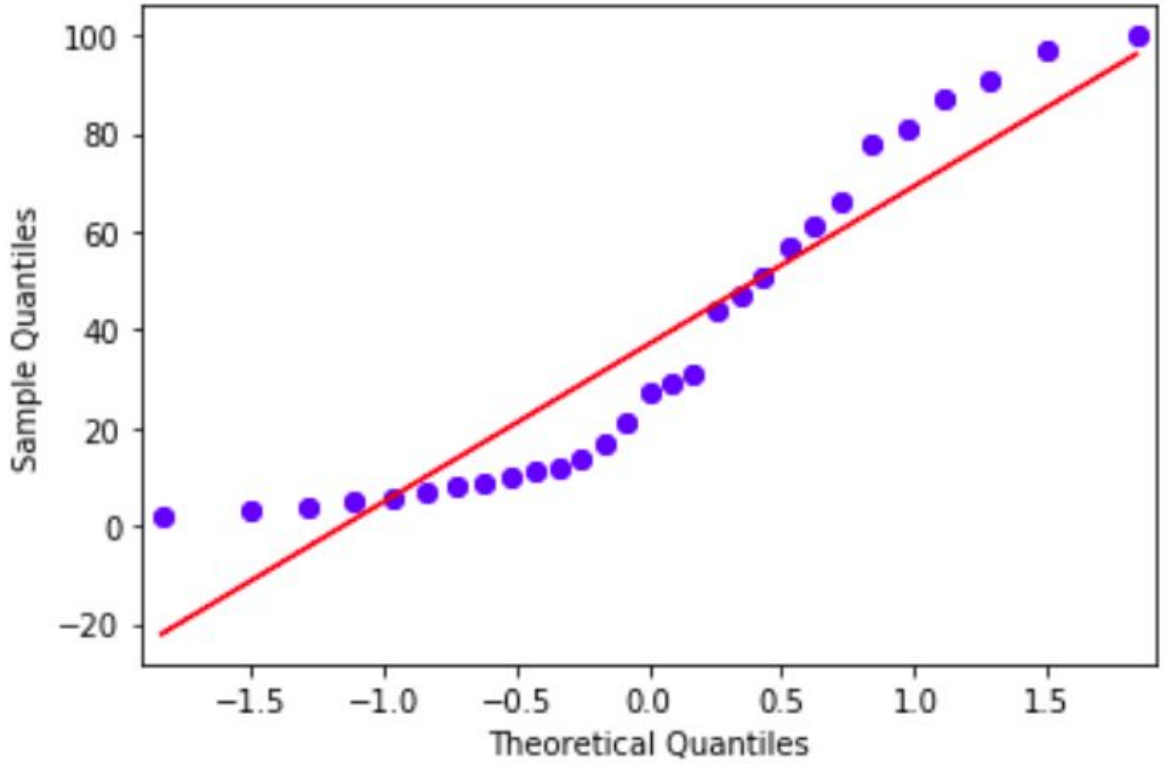
- P value for petal length in setosa species.
  - `p value: 0.0018077816348522902`
  - Since p value is < threshold we reject null hypothesis
  - Not a Normal distribution


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−50
−25
0
25
50
75
100
125
150
battery_percentage

Histogram

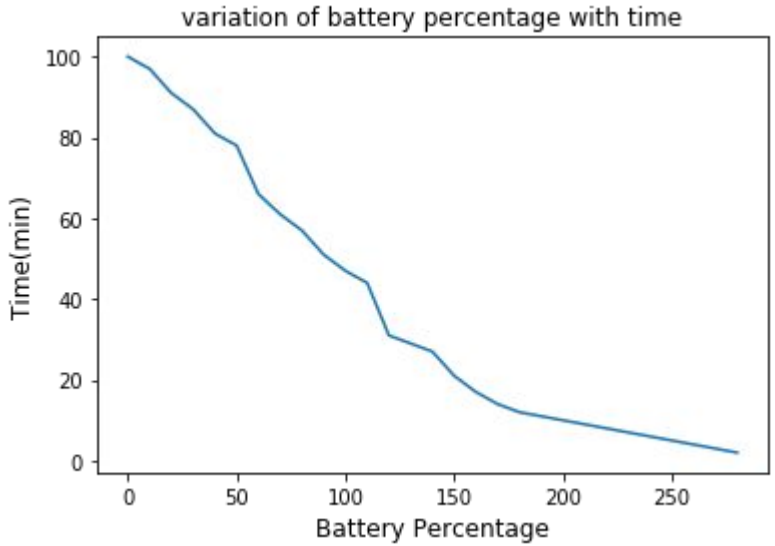


Box Plot


Sample Quantiles
100
80
60
40
20
0
−20
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles

QQ Plot

- From the graphs we can conclude that distribution is not normal.
- It is rather exponential.



Other Tests for normality

- **Kolmogrov-Smirnov**
  The Kolmogorov–Smirnov tests if a sample distribution fits a cumulative distribution function (CDF) of are referenced distribution. Or, if the CDF between of two different samples fit each other.

- **Anderson-Darling**
  The Anderson-Darling tests if data comes from a particular distribution. The null hypothesis — similar to the previous two tests — is the sample comes from a population that follows a particular distribution.

THANK YOU