Getting ready for data mining with basic statistics

It is important to understand your data in various ways to help first preprocess and later perform data mining.

I.e.

Types of attributes
Values each attribute is having
Whether attribute values are discrete or continuous
Data central tendency / center of distribution
Dispersion/spread
Data similarity and dissimilarity wide proximity measures
Noise, outliers

Know that plots/visualization can also help identify relations, trends, biases, skewness, etc.

This also is part of basic statistics. R is widely used for basic statistics and data analytics especially in research.

Central tendency of data

Mean (average value)

data = c(30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110)
mean(data)

```
> data
 [1]  30  36  47  50  52  52  56  60  63  70  70 110
> mean(data)
[1] 58
>
```

Weighted mean

Median (middle value)

median(data)

```
> data
 [1]  30  36  47  50  52  52  56  60  63  70  70 110
> median(data)
[1] 54
>
```

Mode (most common value)

Skew (Asymmetricity)

Positively skewed

Negatively skewed

Dispersion of data

Range i.e. min and max

Quantiles

Quartiles

Five number summary and box plots

Min-q1-meadian-q3-max

```
> data = c(30, 36, 47, 50, 52, 52, 56,
+ 60, 63, 70, 70, 110)
> data
 [1]  30  36  47  50  52  52  56  60  63  70  70 110
> summary(data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.00   49.25   54.00   58.00   64.75  110.00
>
```

```
> data=c(30, 36, 47, 50, 52,
+ 52, 56, 60, 63, 70, 70, 110)
> data
 [1]  30  36  47  50  52  52  56  60  63  70  70 110
> fivenum(data)
[1]  30.0  48.5  54.0  66.5 110.0
>
```

data = c(30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110)
summary(data)
fivenum(data)

Min, Quartile1, median, Quartile3 and Maximum

Interquartile range IQR = Q3 - Q1

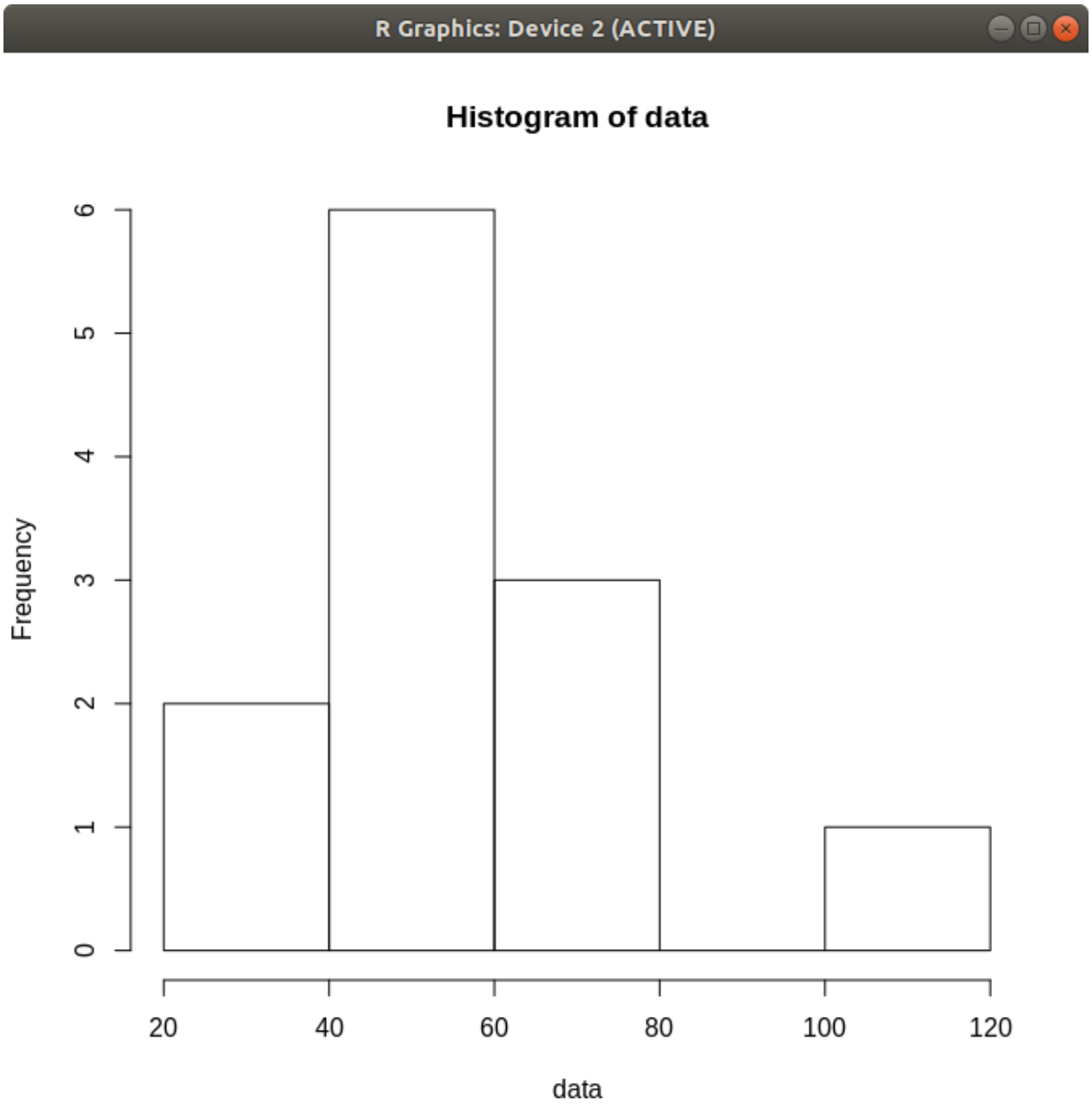
Variance

Standard deviation

Quantile plots

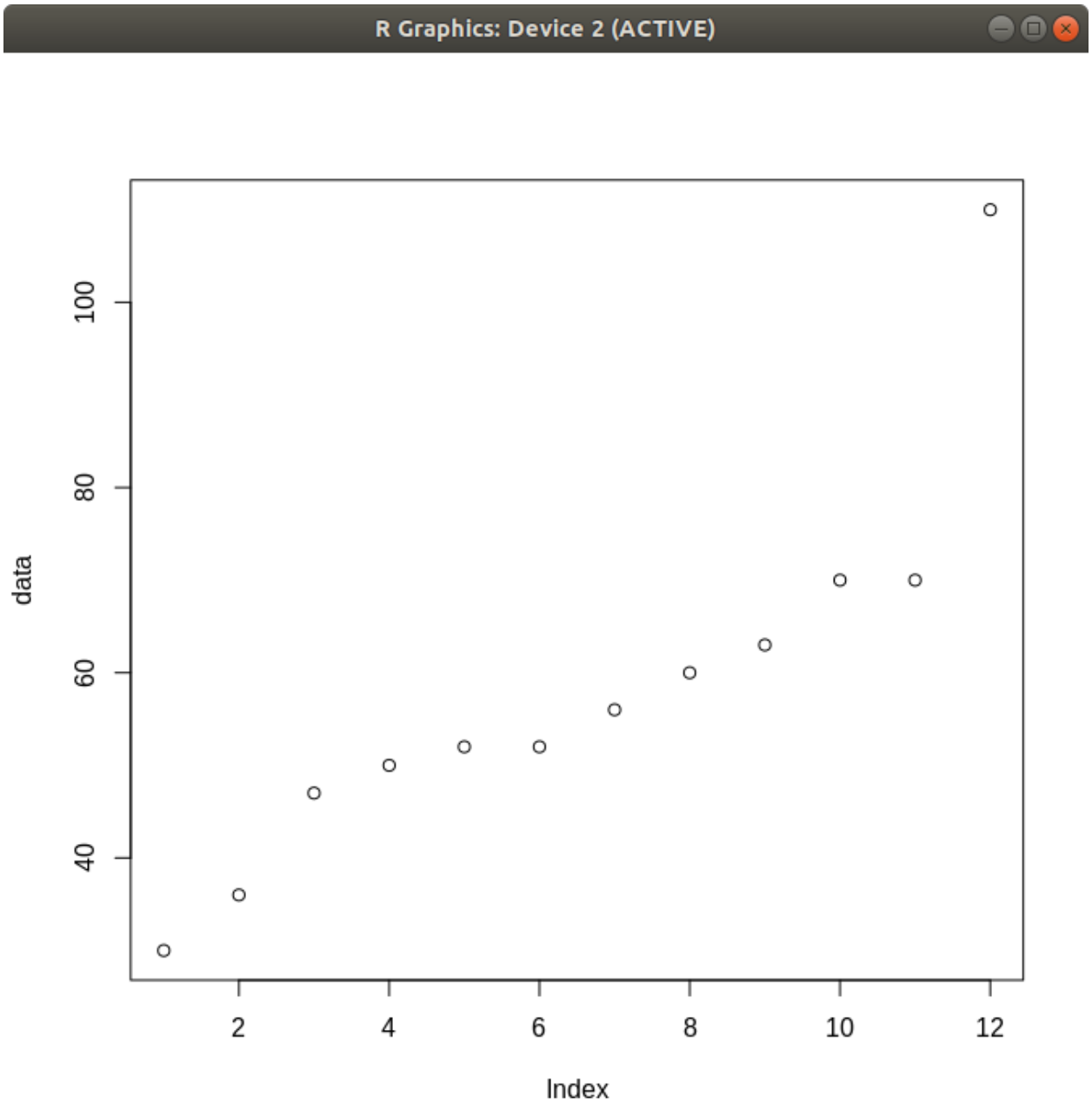
Q-Q plots (Quantile-quantile)

hist(data)


R Graphics: Device 2 (ACTIVE)
Histogram of data
Frequency
0
1
2
3
4
5
6
20
40
60
80
100
120
data

Histograms

plot(data)



Scatter plot

Measures of object similarity and dissimilarity

Data matrix / object by attribute structure

Dissimilarity matrix / object by object structure

Outlier
1.5 x IQR

Asymmetric binary attribute
Jaccard coefficient

Numeric attributes
Euclidean distance

Minkowski Distance

Manhattan Distance

Chebyshev Distance (a.k.a supremum distance)

Sparse numeric data

Term frequency vectors

Cosine similarity / cosine measure

Tanimoto coefficient