Exploratory Data Analysis
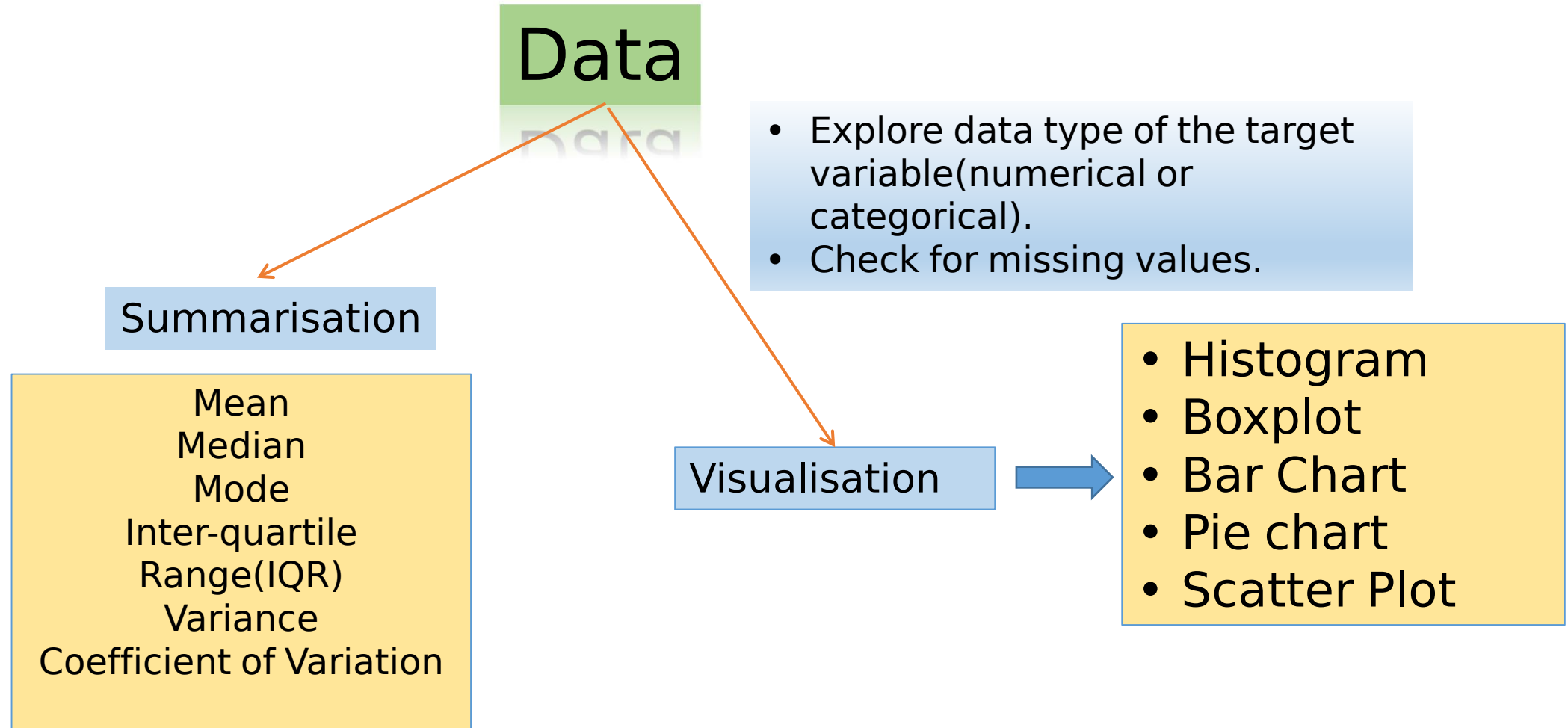
Tutorial

Course: Statistical Data Analysis

IIIT Sricity,2020

Exploratory data analysis

Data

- Explore data type of the target variable(numerical or categorical).
- Check for missing values.

Summarisation

Mean
Median
Mode
Inter-quartile
Range(IQR)
Variance
Coefficient of Variation

Visualisation

- **Histogram**
- **Boxplot**
- **Bar Chart**
- **Pie chart**
- **Scatter Plot**

Data used for this exploration

- Daily corona virus spreading cases in india.
- Daily deaths in India.
- Age and sex of the corona infected patients.
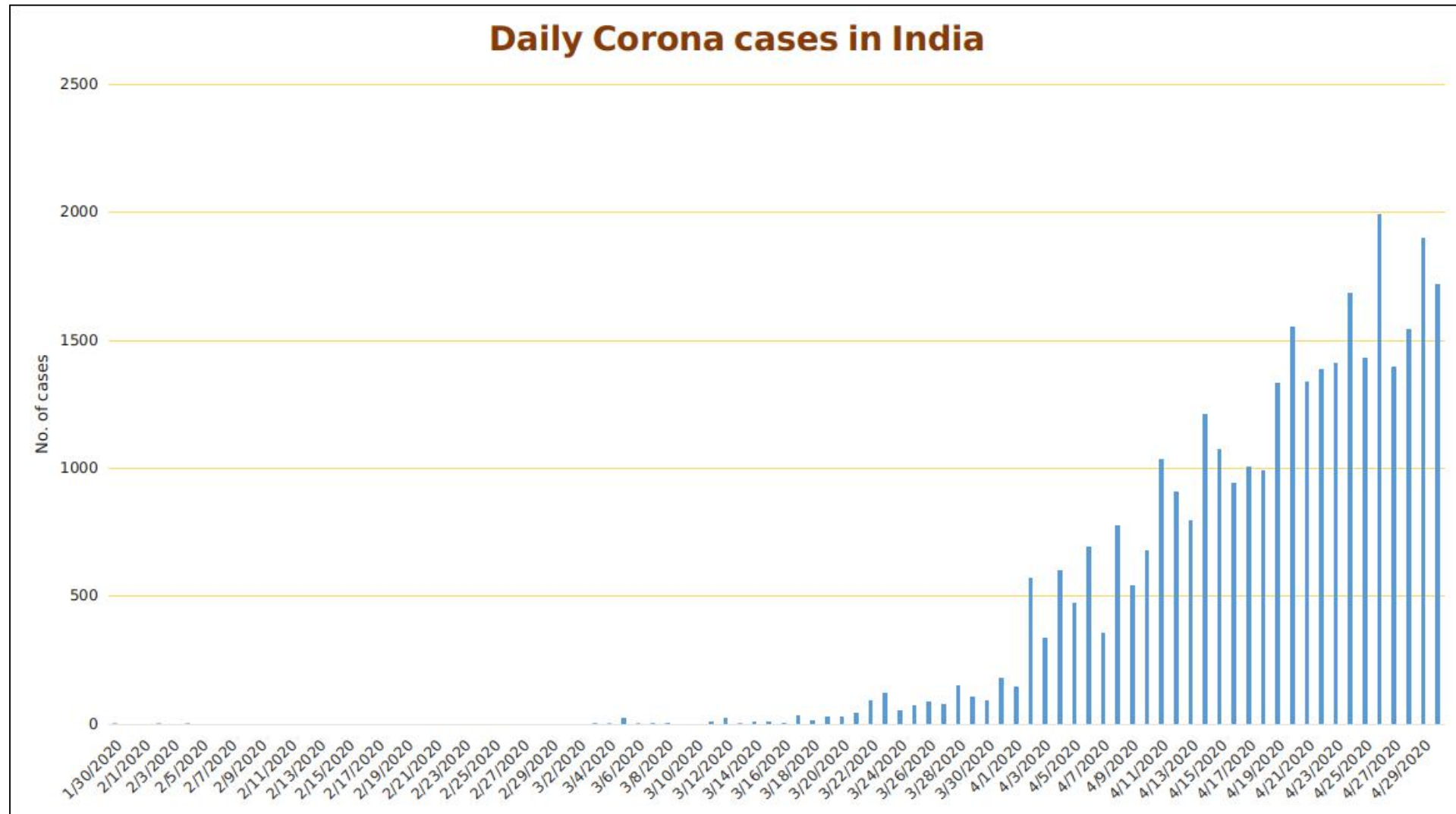- Availability of hospital beds statewise.
- Statewise positive cases.

Data we have considered from:
 30-1-2020 to 30-4-2020

Objectives:
- Summarisation
- Visualisation of aforementioned events

Date: (30-1-2020 to 30-4-2020)
Source(https://data.europa.eu)

Daily corona cases

Daily Corona cases in India

Corona cases Data head

Date	Day	Month	Cases	Deaths	Country
2020-01-30	30	1	1	0	India
2020-01-31	31	1	0	0	India
2020-02-01	1	2	0	0	India
2020-02-02	2	2	1	0	India
2020-02-03	3	2	0	0	India
2020-02-04	4	2	1	0	India

Patient information data head

Diagnosed date	Age	Gender	Detected city	Nationality	Current status
30/1/20	20	F	Thissur	India	Recovered
2/2/20	NA	NA	Alappuzha	India	Recovered
3/2/20	NA	NA	Kasaragod	India	Recovered
2/3/20	45	M	East Delhi	India	Recovered
2/3/20	24	M	Hyderabad	India	Recovered
3/3/20	69	M	Jaipur	Italy	Recovered
4/3/20	55	NA	Gurugram	Italy	Recovered

Summarisation of a univariate variable

- Target variable : Age of the infected patients

• Min	0.3	
• 1st Qu.	26	
• Median	36	(in years)
• Mean	38.50	
• 3rd Qu.	50	
• Max	98	

Histogram of age

Age of the infected patients

Mean
Median
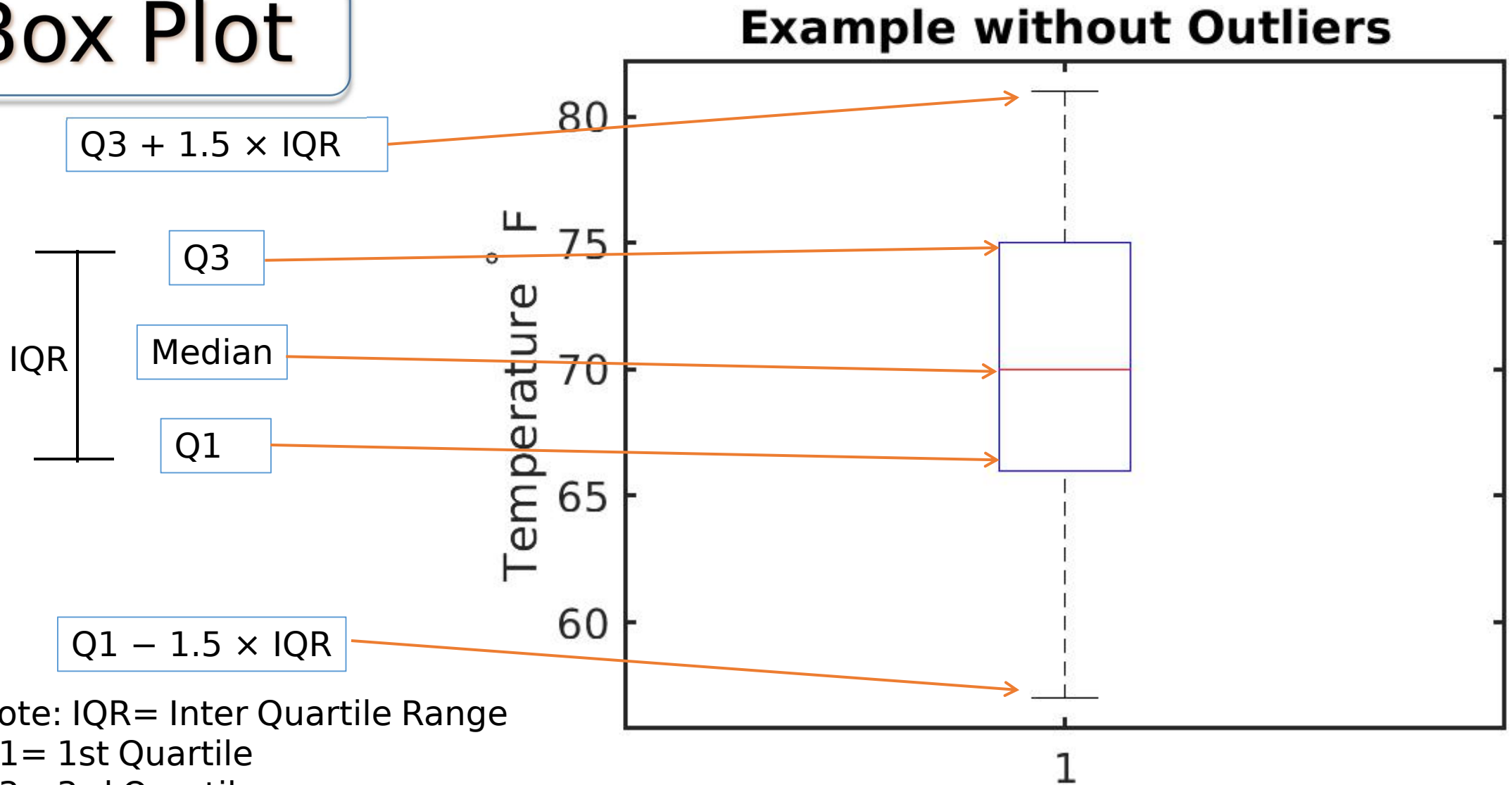3rd Quartile
1st Quartile

Frequency

Age
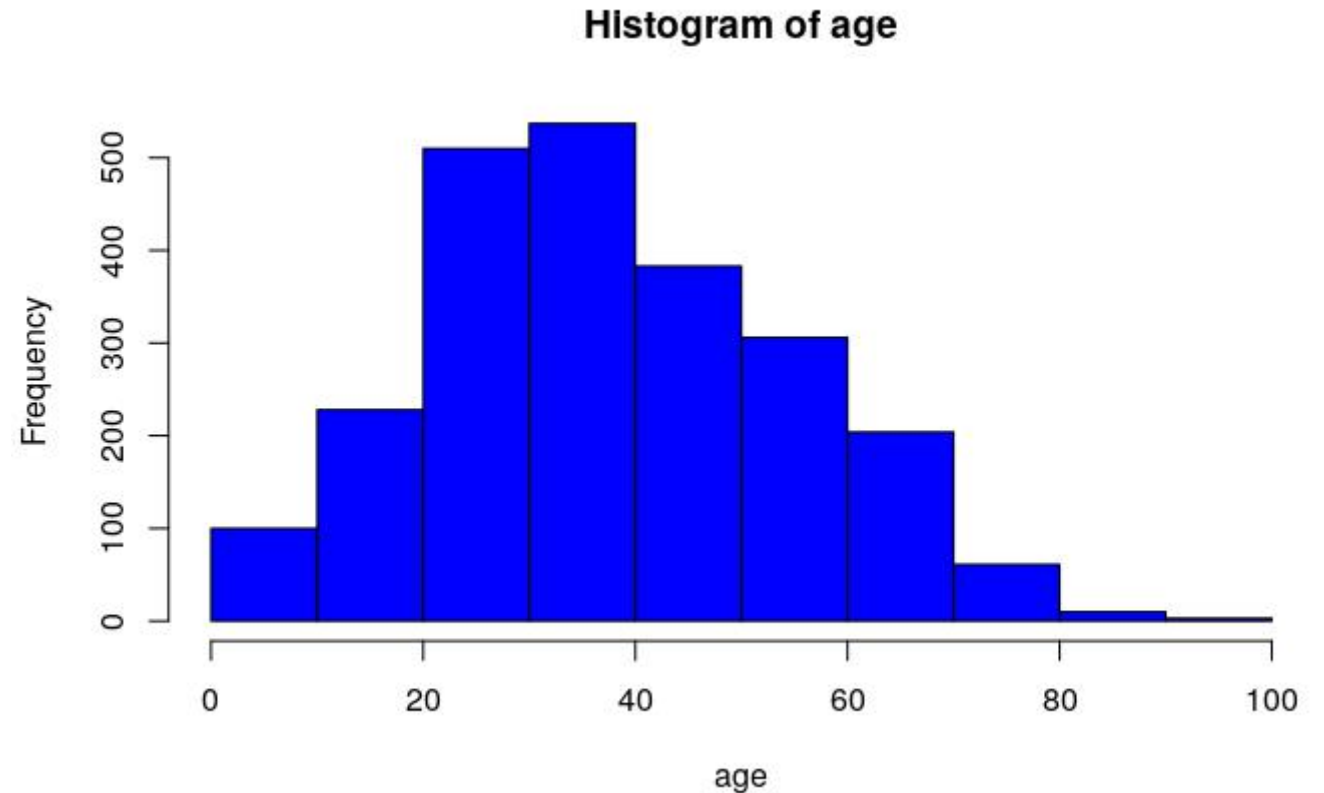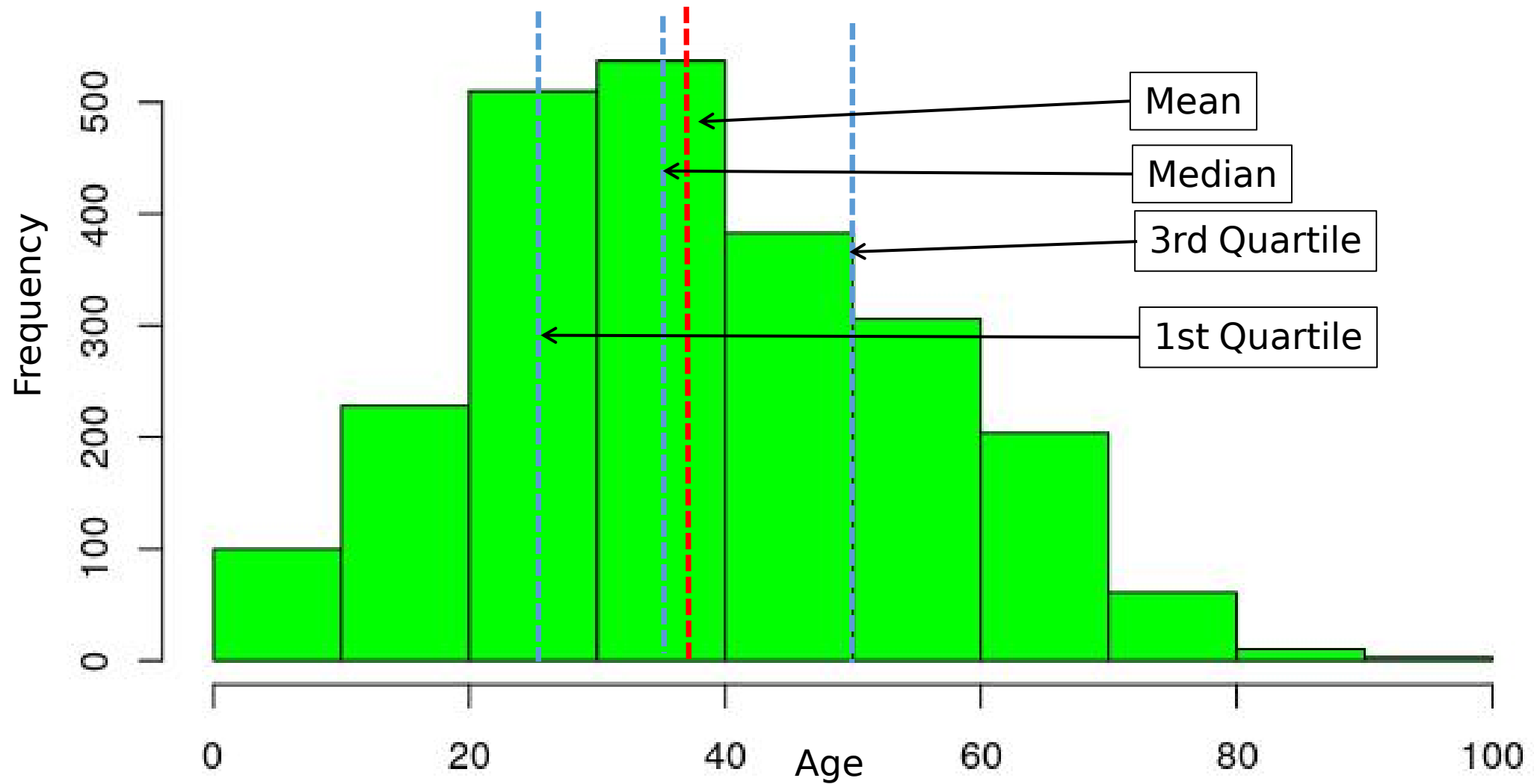
Summarisation

- Inter Quartile Range = 3rd quartile- 1st quartile =24 years (for patient's age)

- Sample Variance= $$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$ = 298

- Standard Deviation (SD)= sqrt(variance)= 17.26
- Coefficient of Variation= (SD/mean) = 0.45

  - Refers to a statistical measure of the distribution of data points.
  - Distributions with a coefficient of variation to be less than 1 are considered to be low-variance, whereas those with a CV higher than 1 are considered to be high variance.

Box Plot

Example without Outliers

Q3 + 1.5 × IQR

Q3

IQR

Median

Q1

Q1 − 1.5 × IQR

Note: IQR= Inter Quartile Range
Q1= 1st Quartile
Q3= 3rd Quartile

Temperature °F

80

75

70

65

60

1

Boxplot of corona infected patient's age
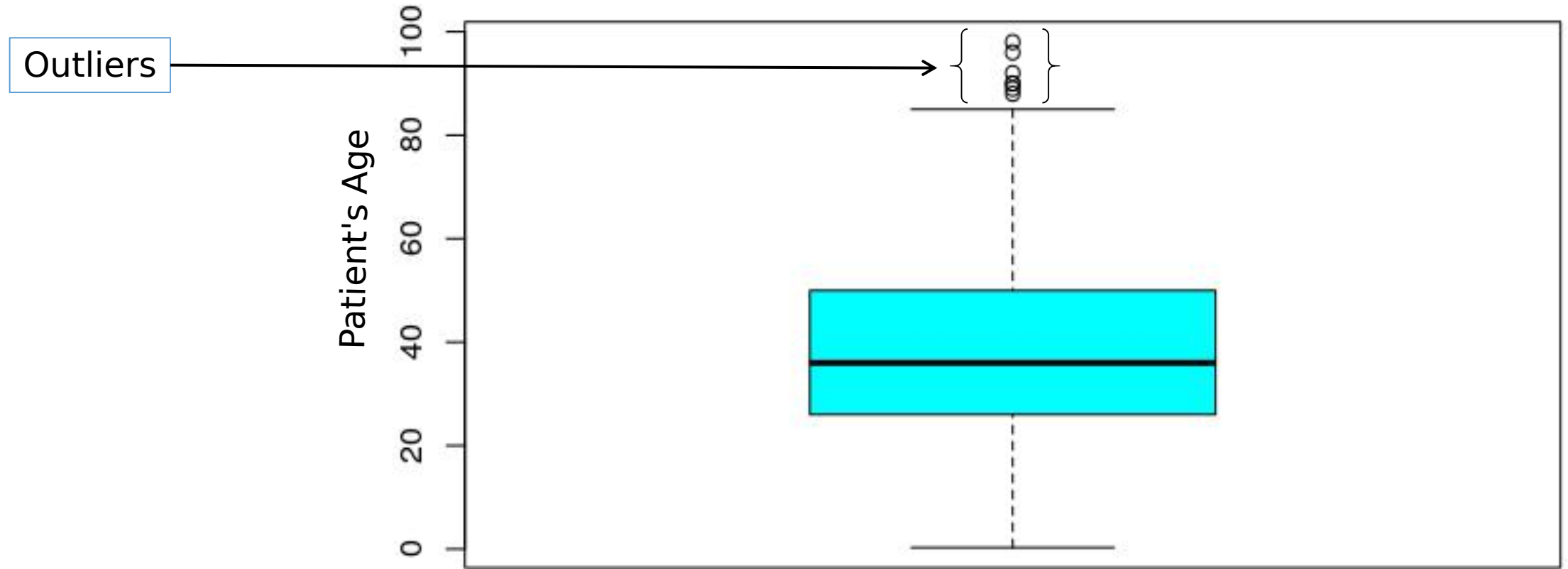
Box plot of corona infected patient's age

Line chart of Daily corona cases in India

Daily Corona infection cases in India
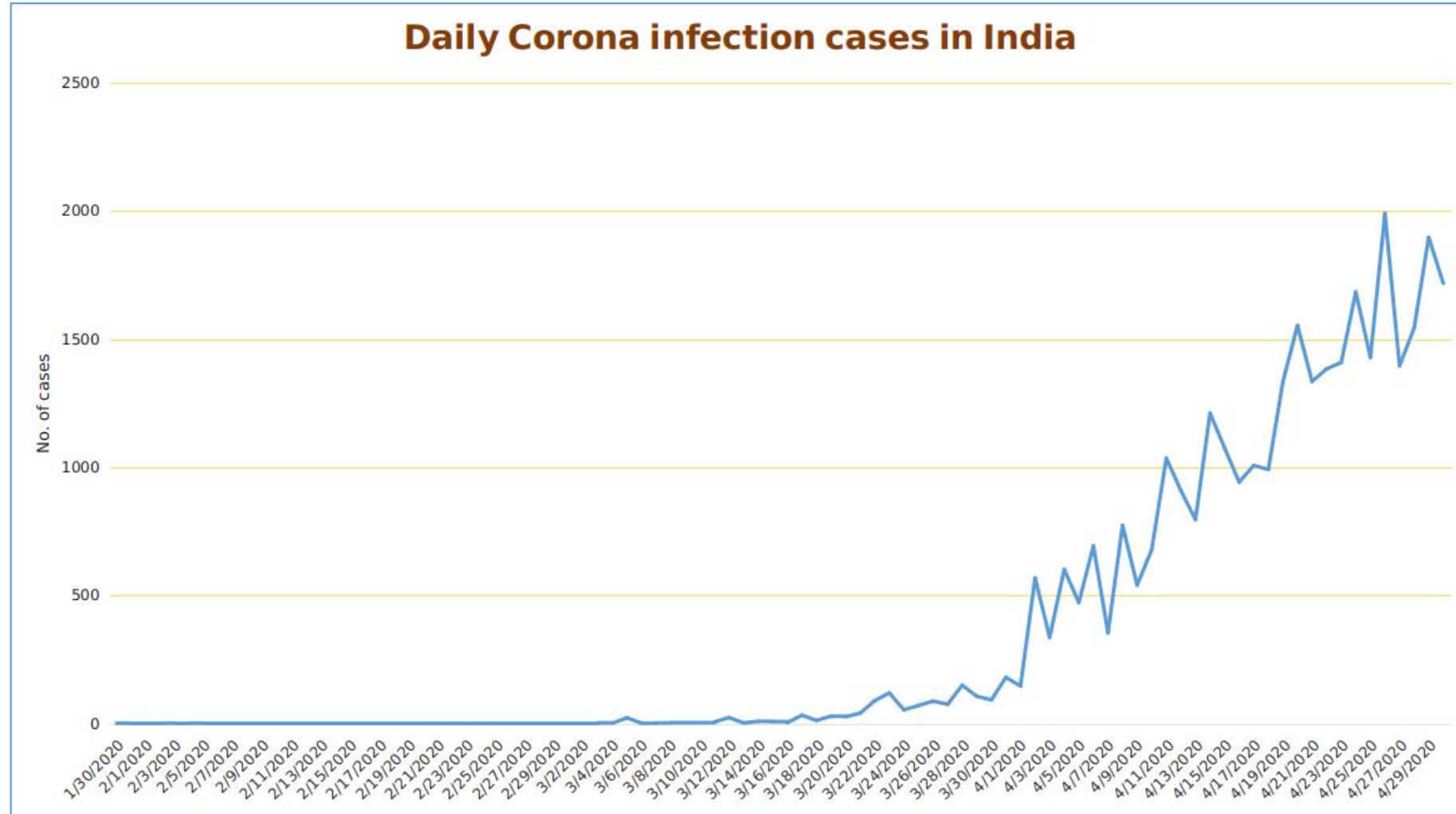
Line chart of daily death cases in India

Daily no. of deaths in India
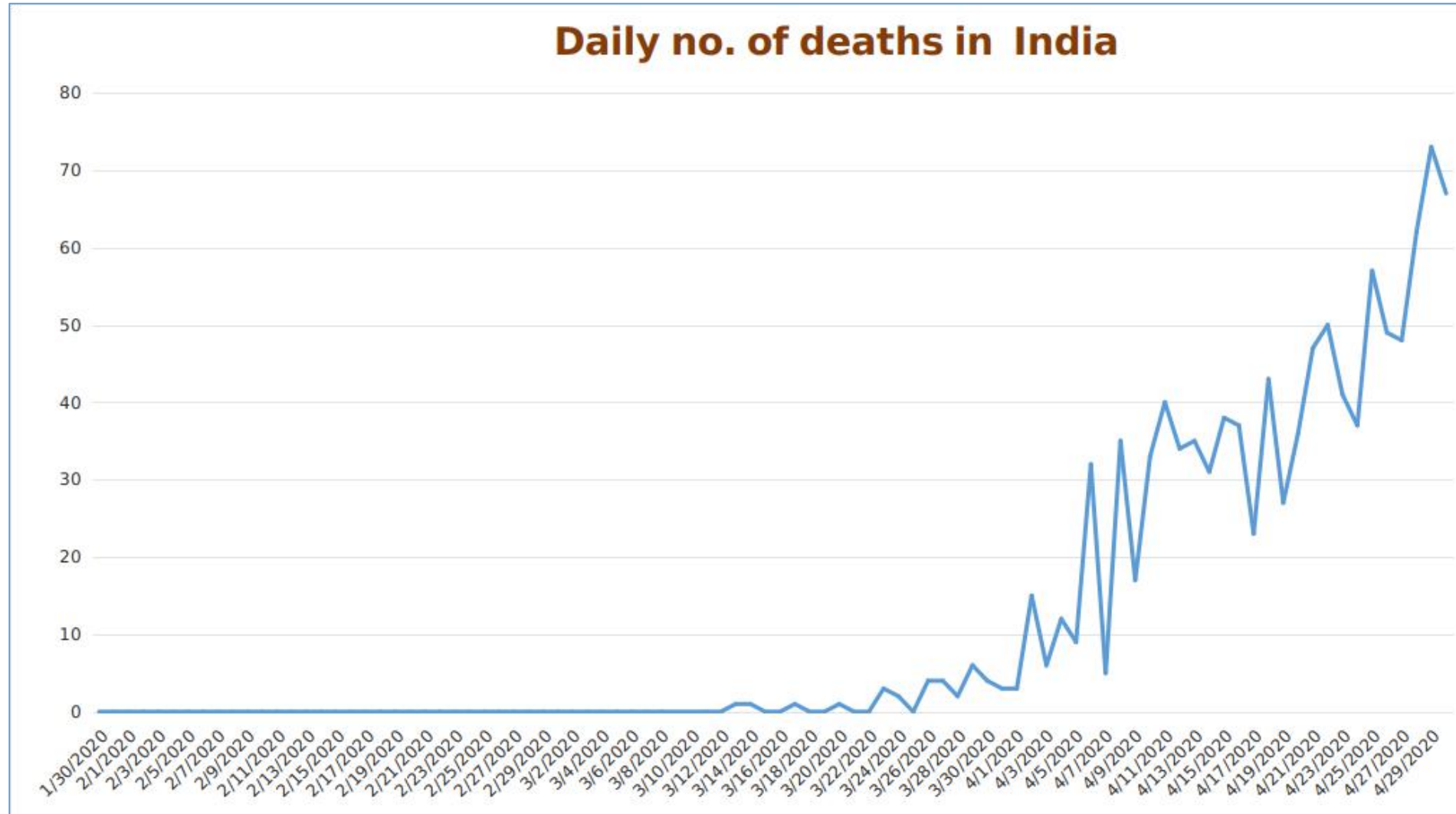
Scatter Plot of Corona cases VS deaths in India

daily deaths (in No.)

Daily infection cases (in No.)

The data we considered lies within the bounding box

Daily infection cases (in No.)

30th Jan-30th April 2020
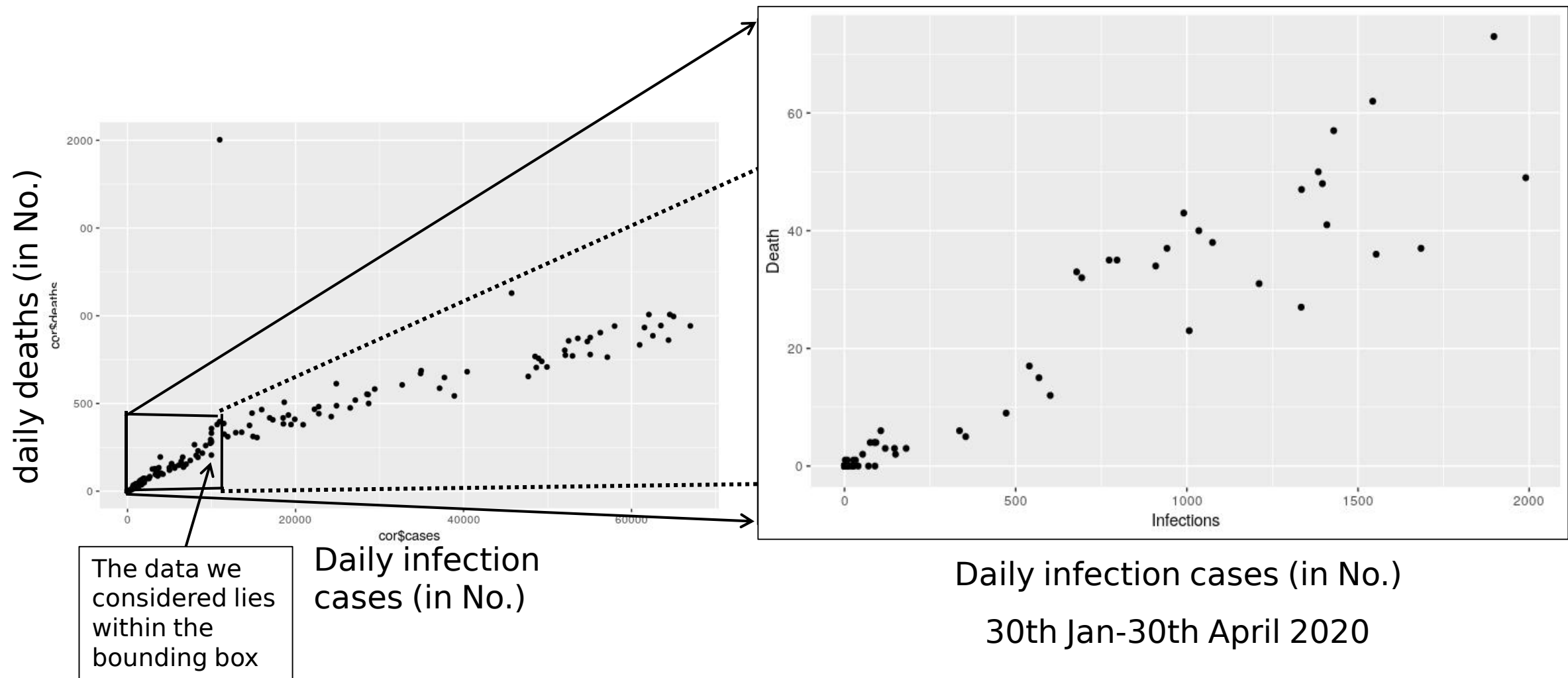
Bar graph

Availability of hospital beds for corona patients (statewise)

State	Beds
West Bengal	135105
Uttarakhand	16398
Uttar Pradesh	147449
Tripura	9781
Telangana	41068
Tamil Nadu	156527
Sikkim	2802
Rajasthan	90722
Punjab	33626
Puducherry	8151
Odisha	40418
Nagaland	4192
Mizoram	4573
Meghalaya	9559
Manipur	4247
Maharashtra	127633
Madhya Pradesh	71164
Lakshadweep	579
Kerala	81389
Karnataka	135229
Jharkhand	19859
Jammu & Kashmir	24761
Himachal Pradesh	23248
Haryana	27115
Gujarat	78367
Goa	5801
Delhi	46294
Daman & Diu	559
Dadra & Nagar Haveli	1194
Chhattisgarh	26026
Chandigarh	4632
Bihar	35140
Assam	39923
Arunachal Pradesh	5340
Andhra Pradesh	87527
Andaman & Nicobar Islands	2419

Series1

No. of Infections of Male vs Female for a week
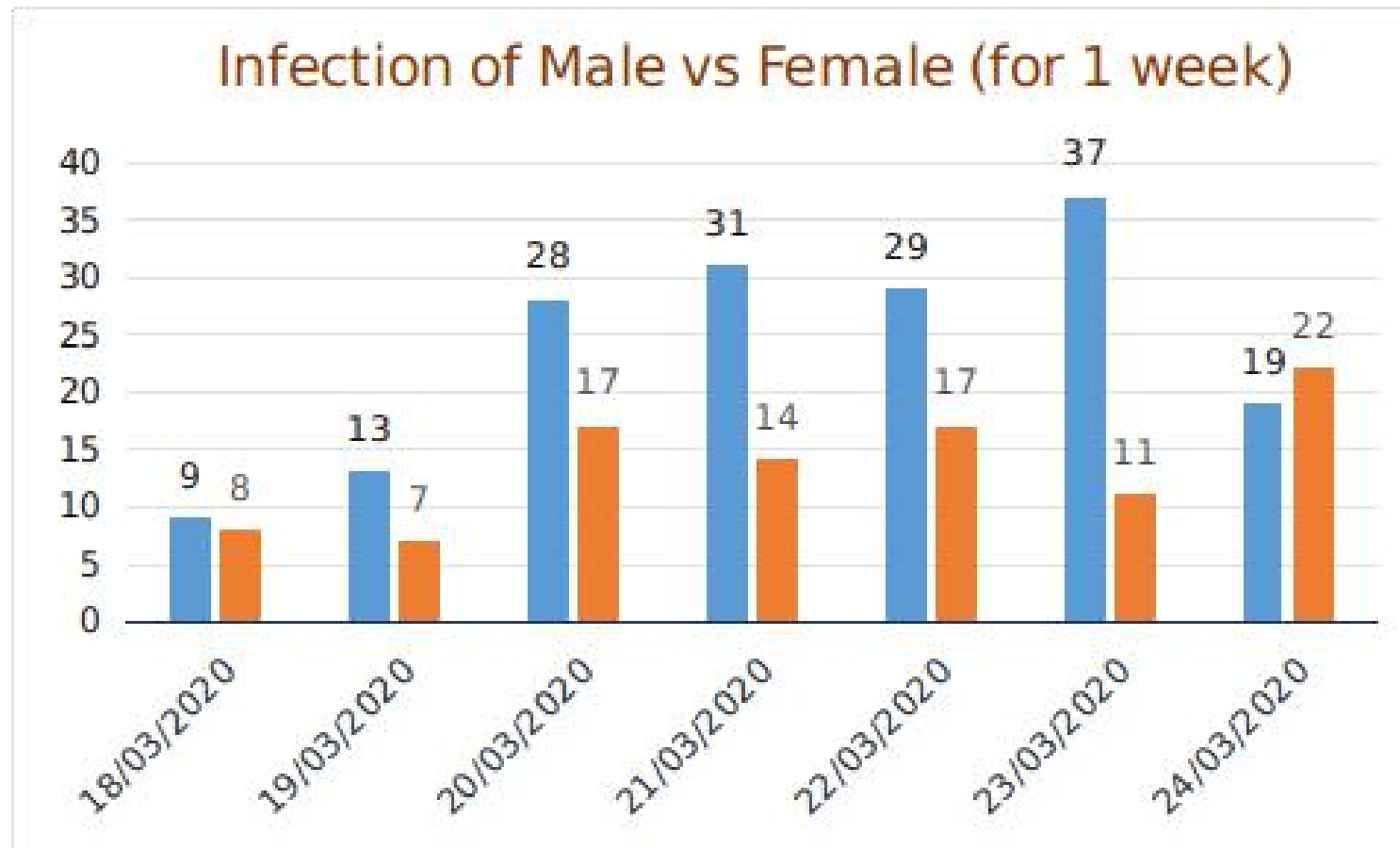
18th March- 24th March 2020

Pie chart

Covid-19 Statewise positive cases in INDIA

data source:https://www.mygov.in/corona-data/covid19-statewise-status/