

# Exploratory Data Analysis (EDA)

Statsomat.com

24 June 2021

## Basic Information

Automatic statistics for the file:

| File                       |
|----------------------------|
| HolzingerSwineford1939.csv |

Your selection for the encoding: Auto

Your selection for the decimal character: .

Observations (rows with at least one non-missing value): 301

Variables (columns with at least one non-missing value): 16

Variables considered continuous: 12

| Variables considered continuous |
|---------------------------------|
| agemo                           |
| id                              |
| V1                              |
| x1                              |
| x2                              |
| x3                              |
| x4                              |
| x5                              |
| x6                              |
| x7                              |
| x8                              |
| x9                              |

Variables considered categorical: 4

| Variables considered categorical |
|----------------------------------|
| ageyr                            |
| grade                            |
| school                           |
| sex                              |

## Results for Numerical Variables

### Descriptive Statistics

Variables are sorted alphabetically. Missings are omitted in stats. CV only for positive variables.

|       | N Obs | N Missing | N Valid | % Complete | N Unique | Mean   | SD     | Median | MAD    | Min  | Max  | Skewness | Kurtosis | CV   |
|-------|-------|-----------|---------|------------|----------|--------|--------|--------|--------|------|------|----------|----------|------|
| agemo | 301   | 0         | 301     | 100        | 12       | 5.38   | 3.45   | 5      | 4.45   | 0    | 11   | 0.09     | -1.21    |      |
| id    | 301   | 0         | 301     | 100        | 301      | 176.55 | 105.94 | 163    | 140.85 | 1    | 351  | -0.01    | -1.35    | 0.6  |
| V1    | 301   | 0         | 301     | 100        | 301      | 151    | 87.04  | 151    | 111.2  | 1    | 301  | 0        | -1.2     | 0.58 |
| x1    | 301   | 0         | 301     | 100        | 35       | 4.94   | 1.17   | 5      | 1.24   | 0.67 | 8.5  | -0.26    | 0.36     | 0.24 |
| x2    | 301   | 0         | 301     | 100        | 25       | 6.09   | 1.18   | 6      | 1.11   | 2.25 | 9.25 | 0.47     | 0.38     | 0.19 |
| x3    | 301   | 0         | 301     | 100        | 35       | 2.25   | 1.13   | 2.12   | 1.3    | 0.25 | 4.5  | 0.39     | -0.89    | 0.5  |
| x4    | 301   | 0         | 301     | 100        | 20       | 3.06   | 1.16   | 3      | 0.99   | 0    | 6.33 | 0.27     | 0.12     |      |
| x5    | 301   | 0         | 301     | 100        | 25       | 4.34   | 1.29   | 4.5    | 1.48   | 1    | 7    | -0.35    | -0.53    | 0.3  |
| x6    | 301   | 0         | 301     | 100        | 40       | 2.19   | 1.1    | 2      | 1.06   | 0.14 | 6.14 | 0.87     | 0.88     | 0.5  |
| x7    | 301   | 0         | 301     | 100        | 97       | 4.19   | 1.09   | 4.09   | 1.1    | 1.3  | 7.43 | 0.25     | -0.27    | 0.26 |
| x8    | 301   | 0         | 301     | 100        | 84       | 5.53   | 1.01   | 5.5    | 0.96   | 3.05 | 10   | 0.53     | 1.24     | 0.18 |
| x9    | 301   | 0         | 301     | 100        | 129      | 5.37   | 1.01   | 5.42   | 0.99   | 2.78 | 9.25 | 0.21     | 0.34     | 0.19 |

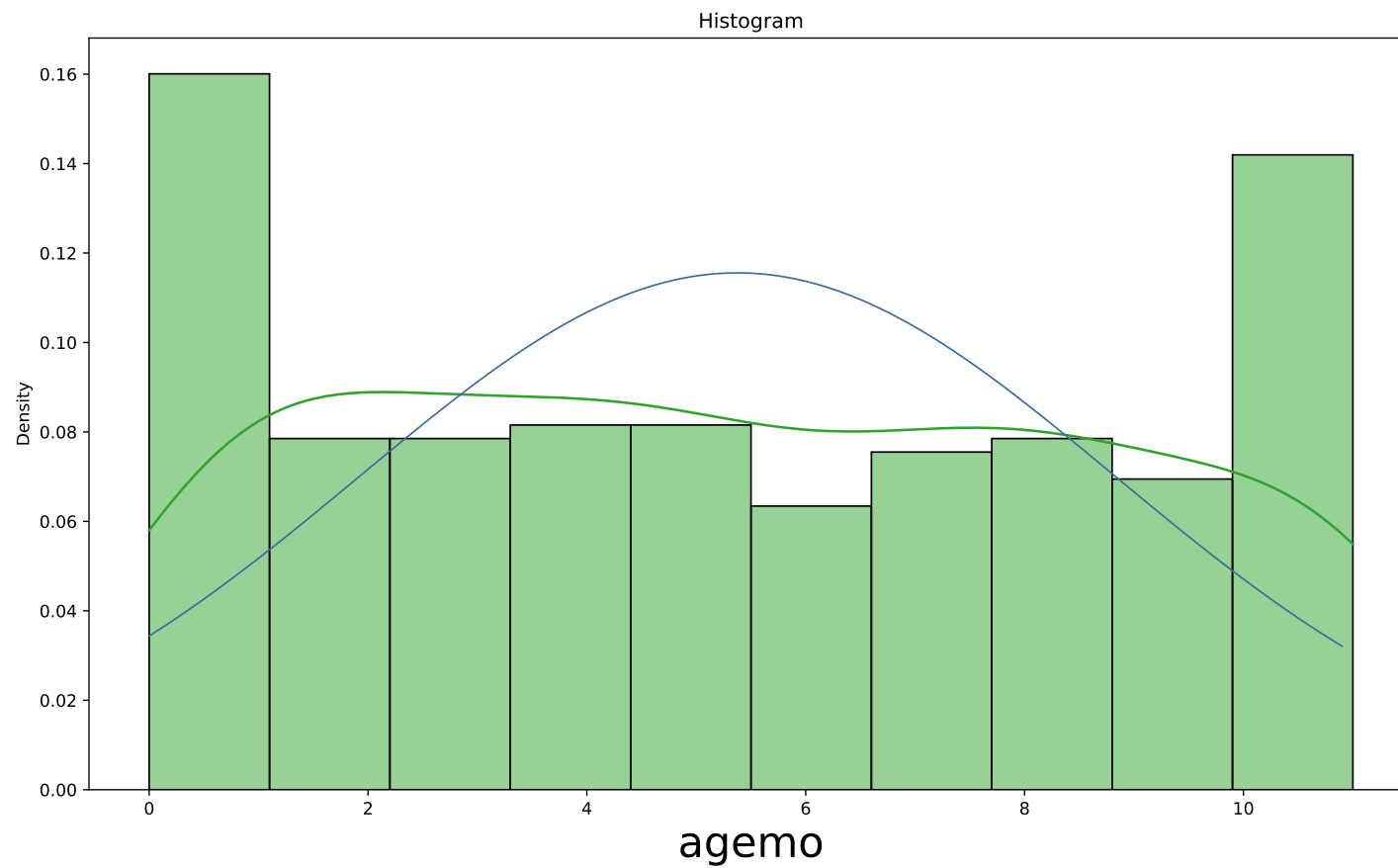
## Graphics

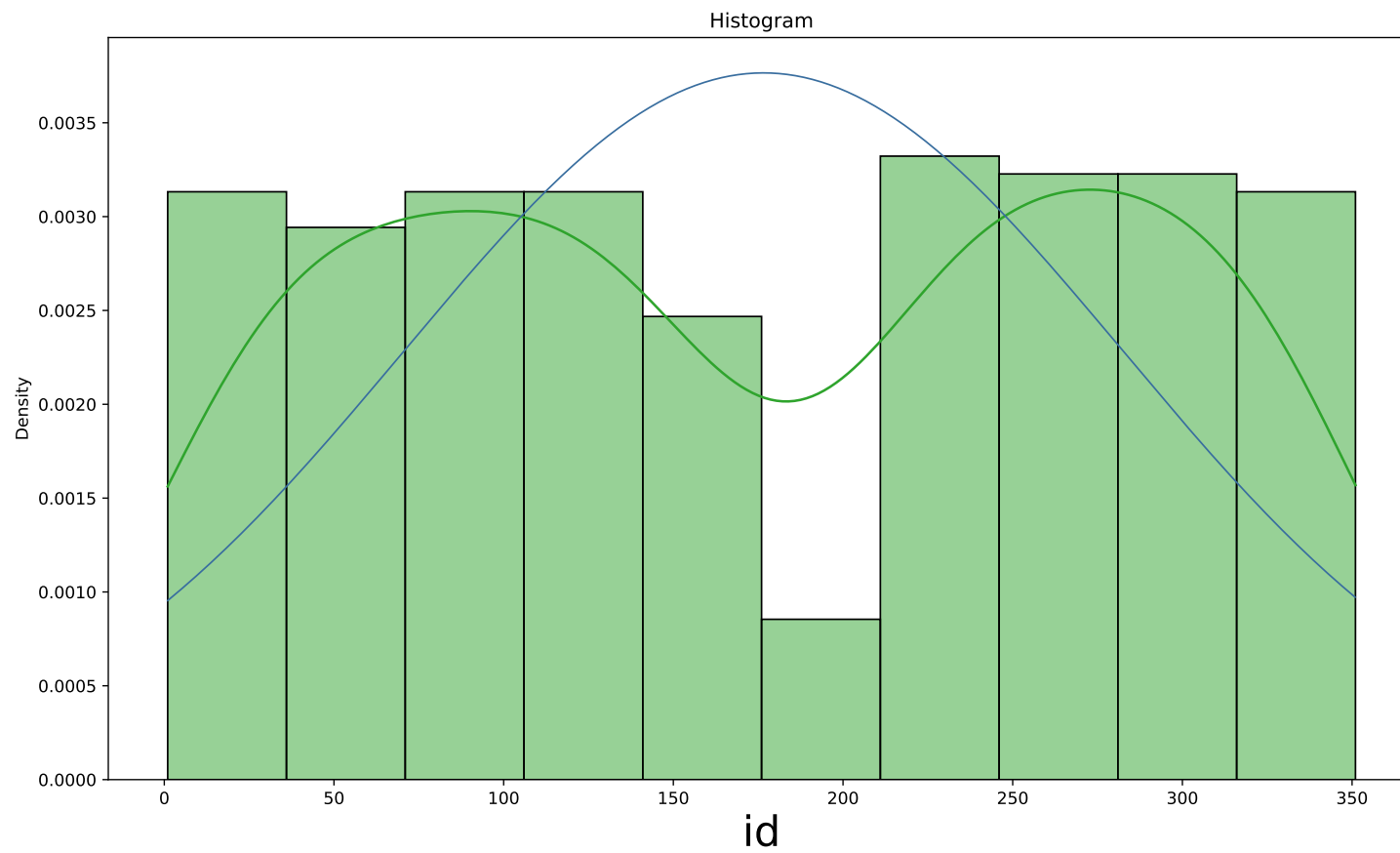
### Histograms

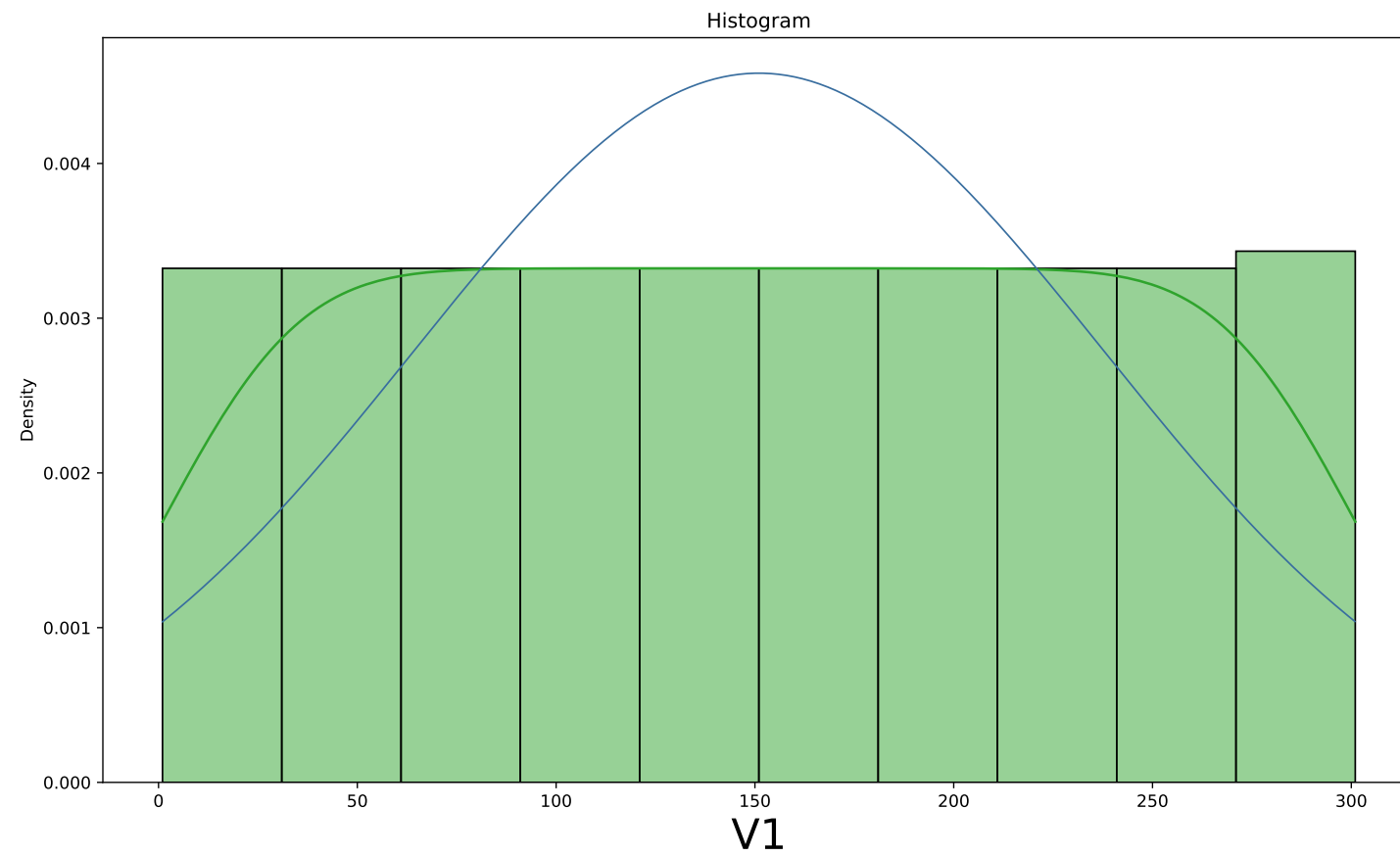
Details: Density Histograms. One large figure per page for each variable, sorted alphabetically. The blue line represents the normal density approximation. The green line represents a special kernel density approximation.

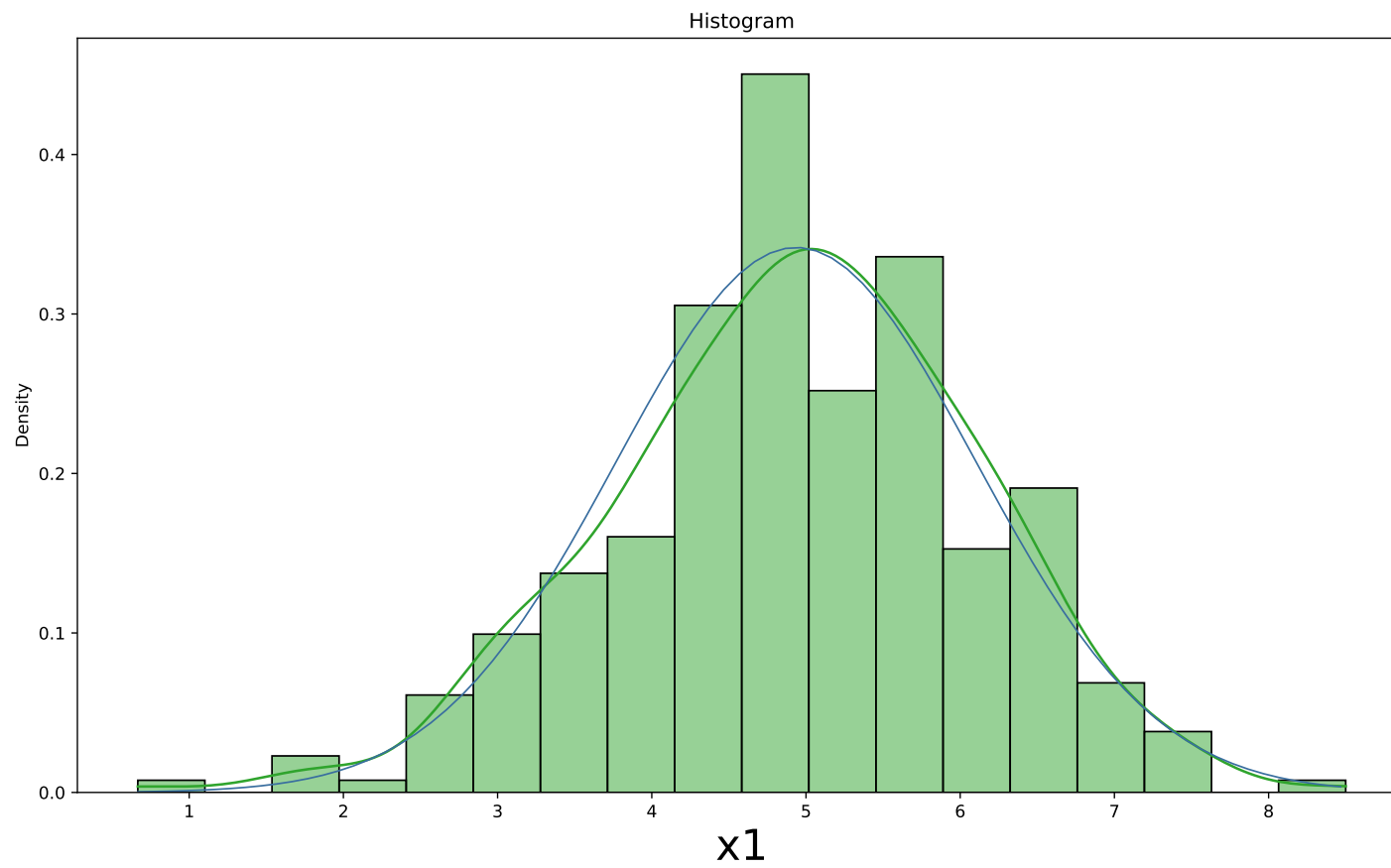
See figures on next page.

□

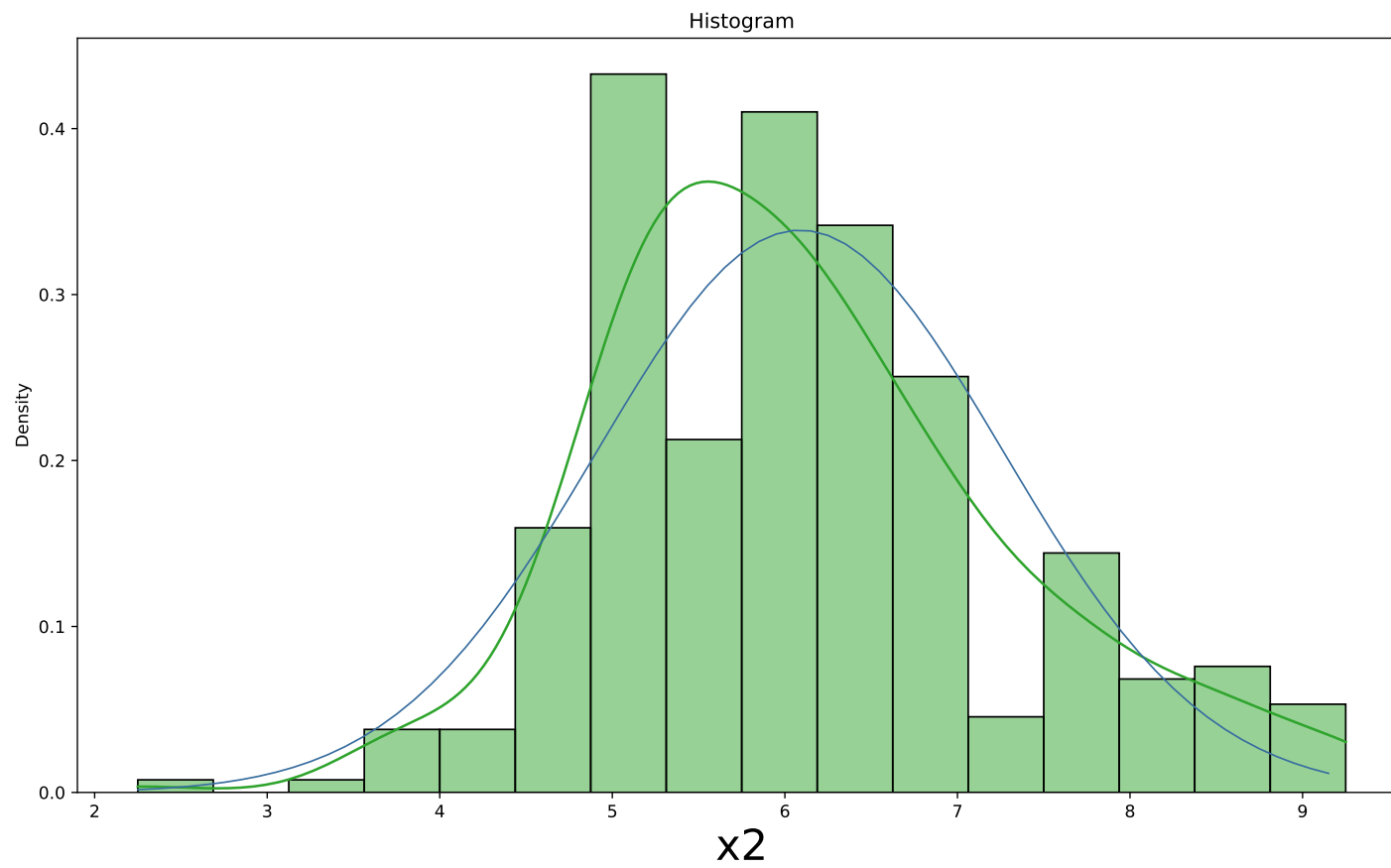


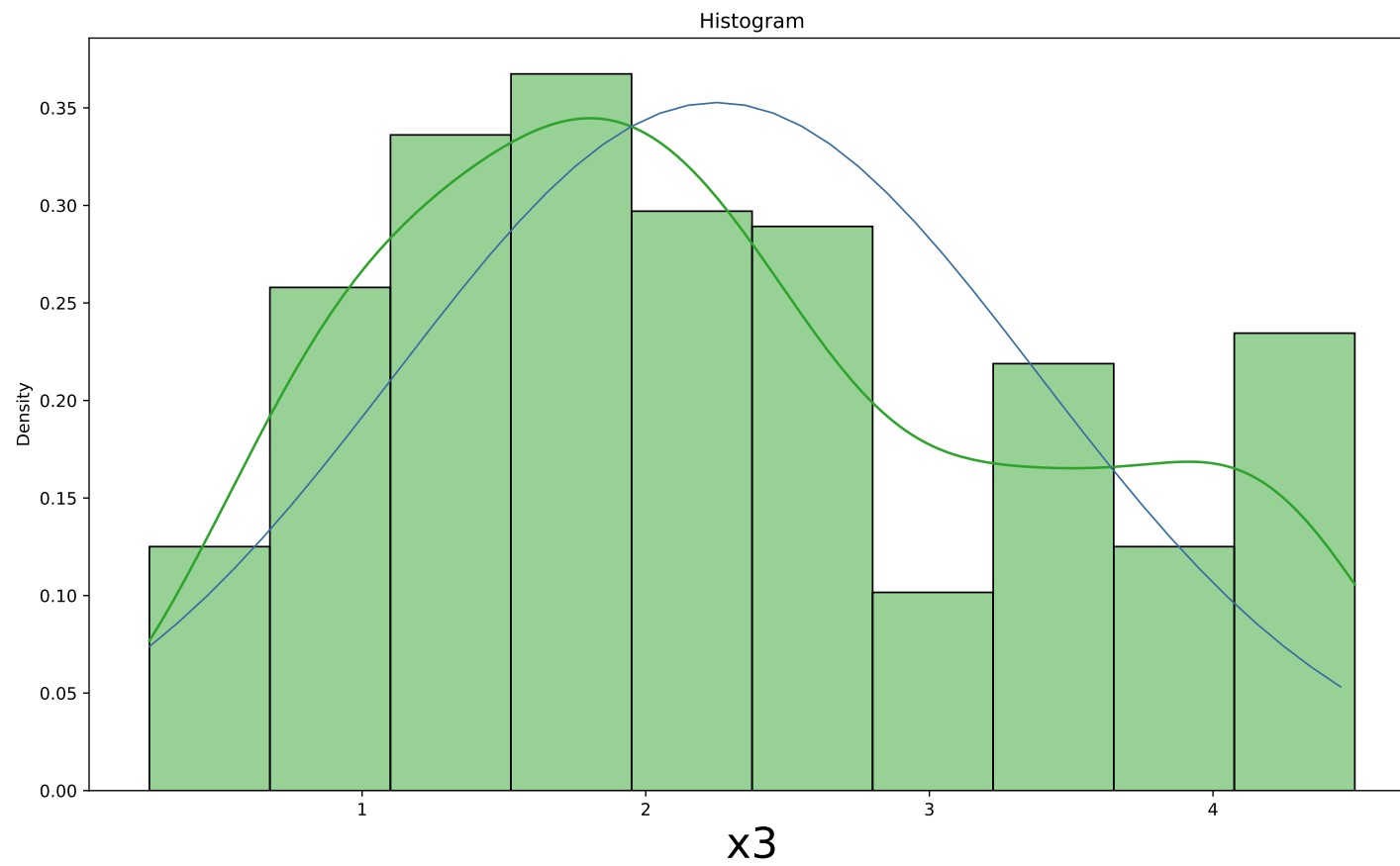


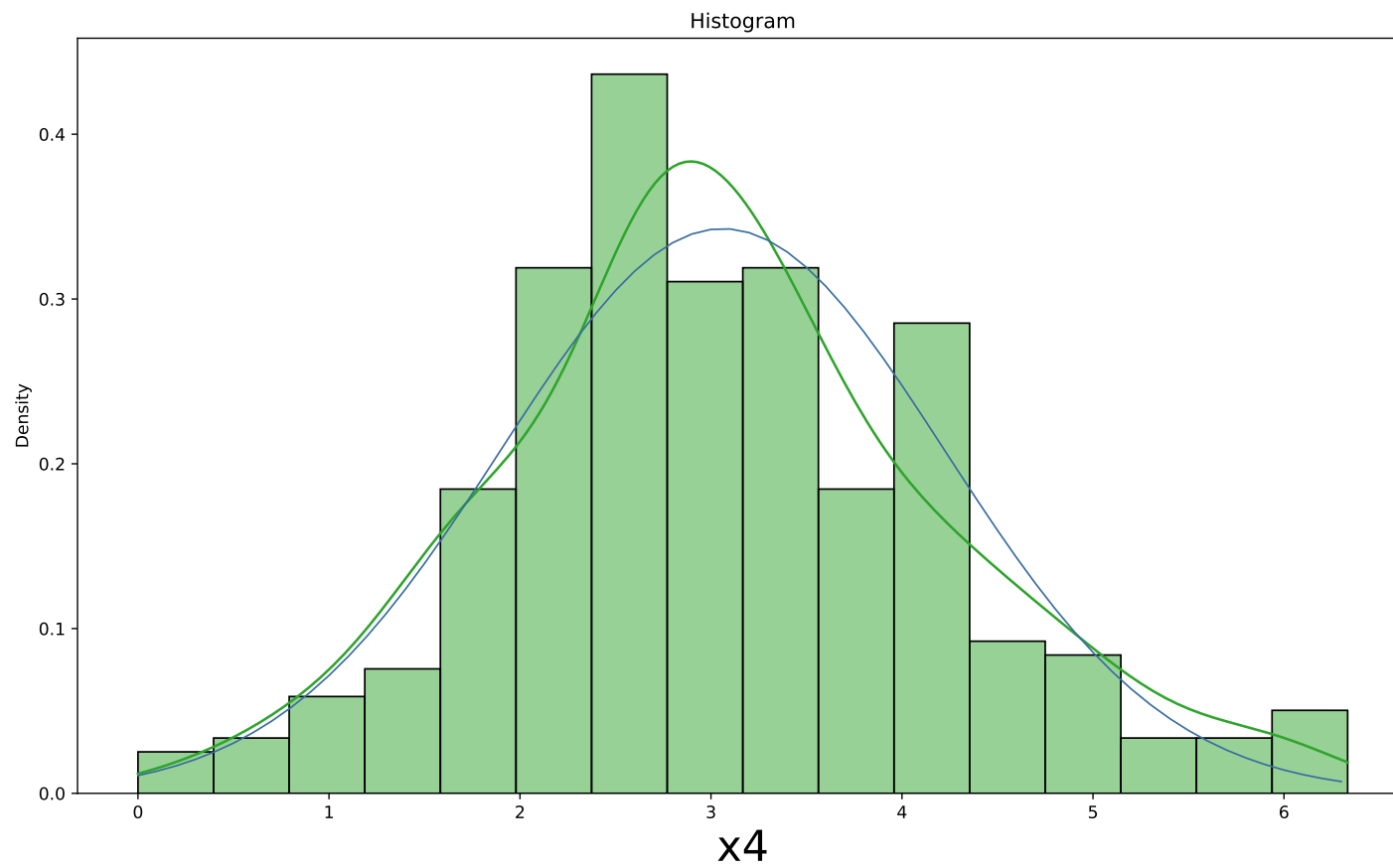


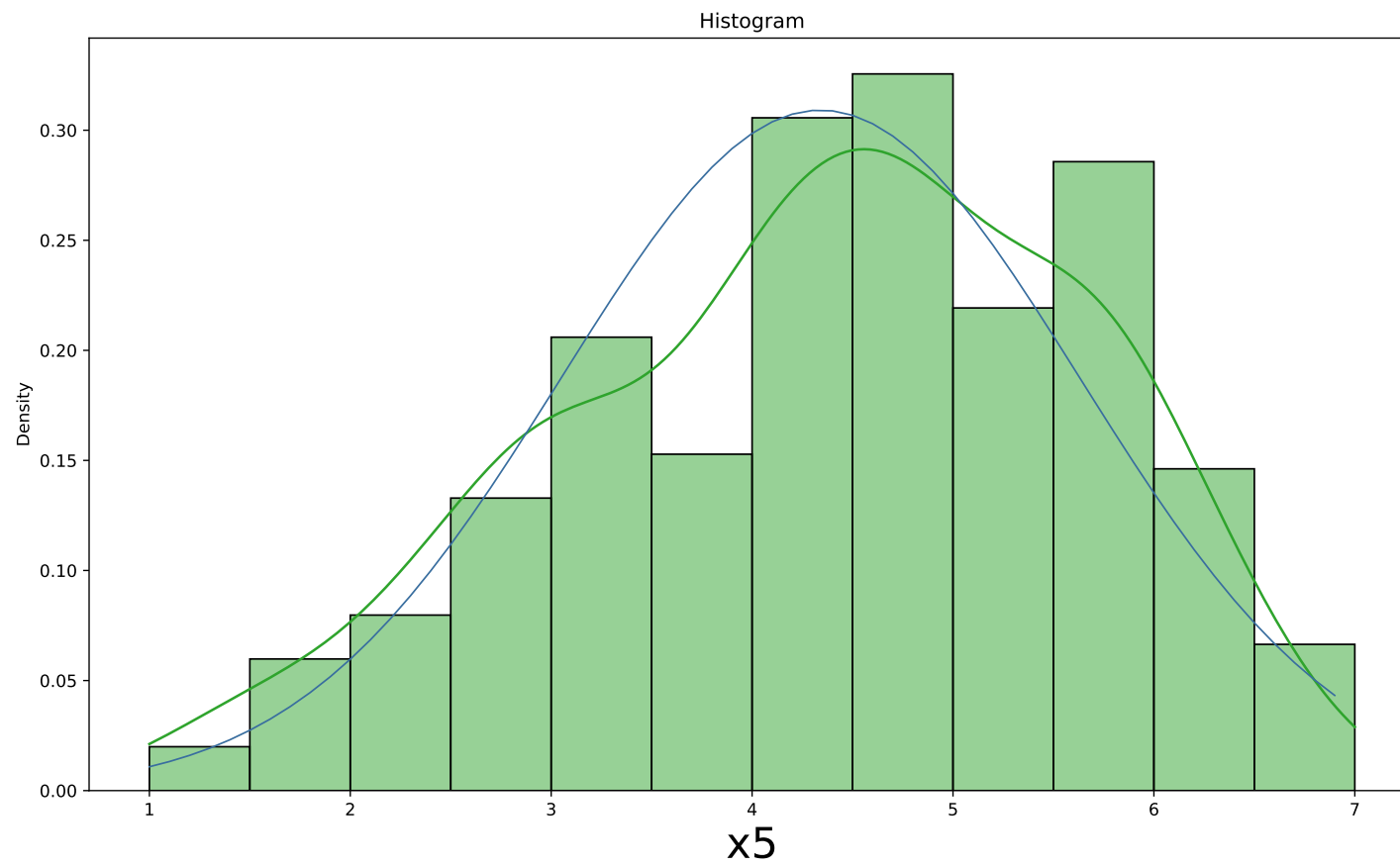


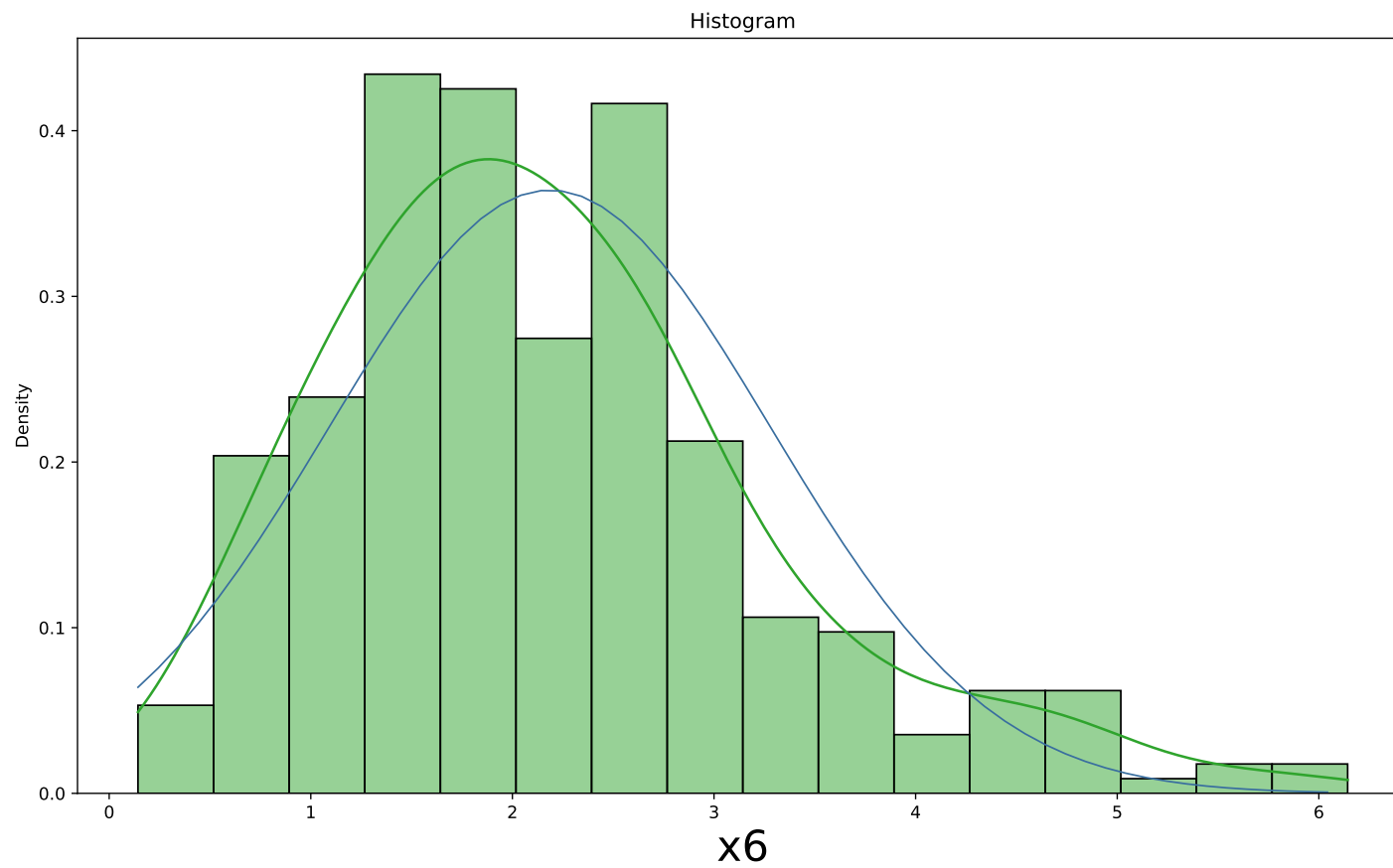


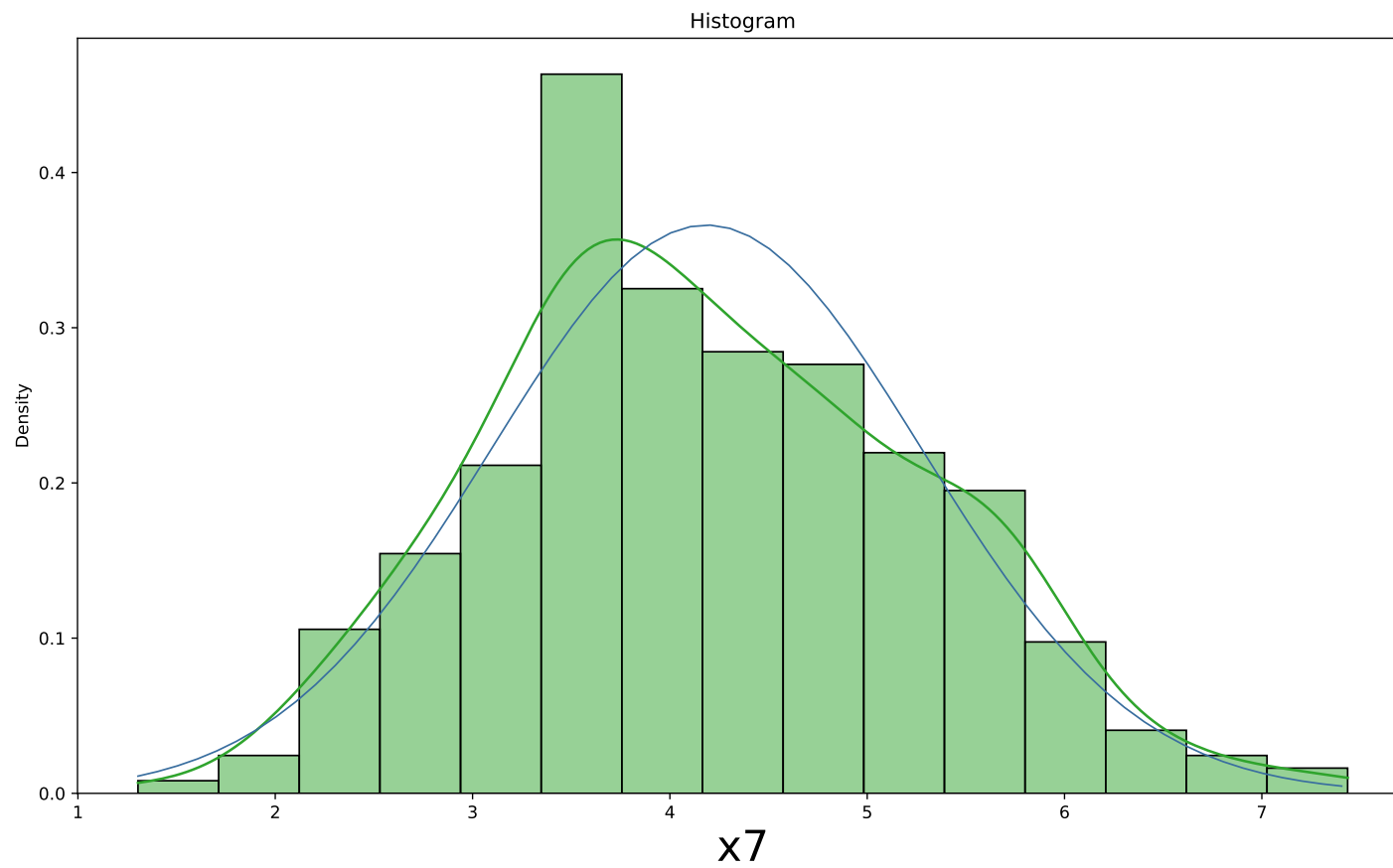


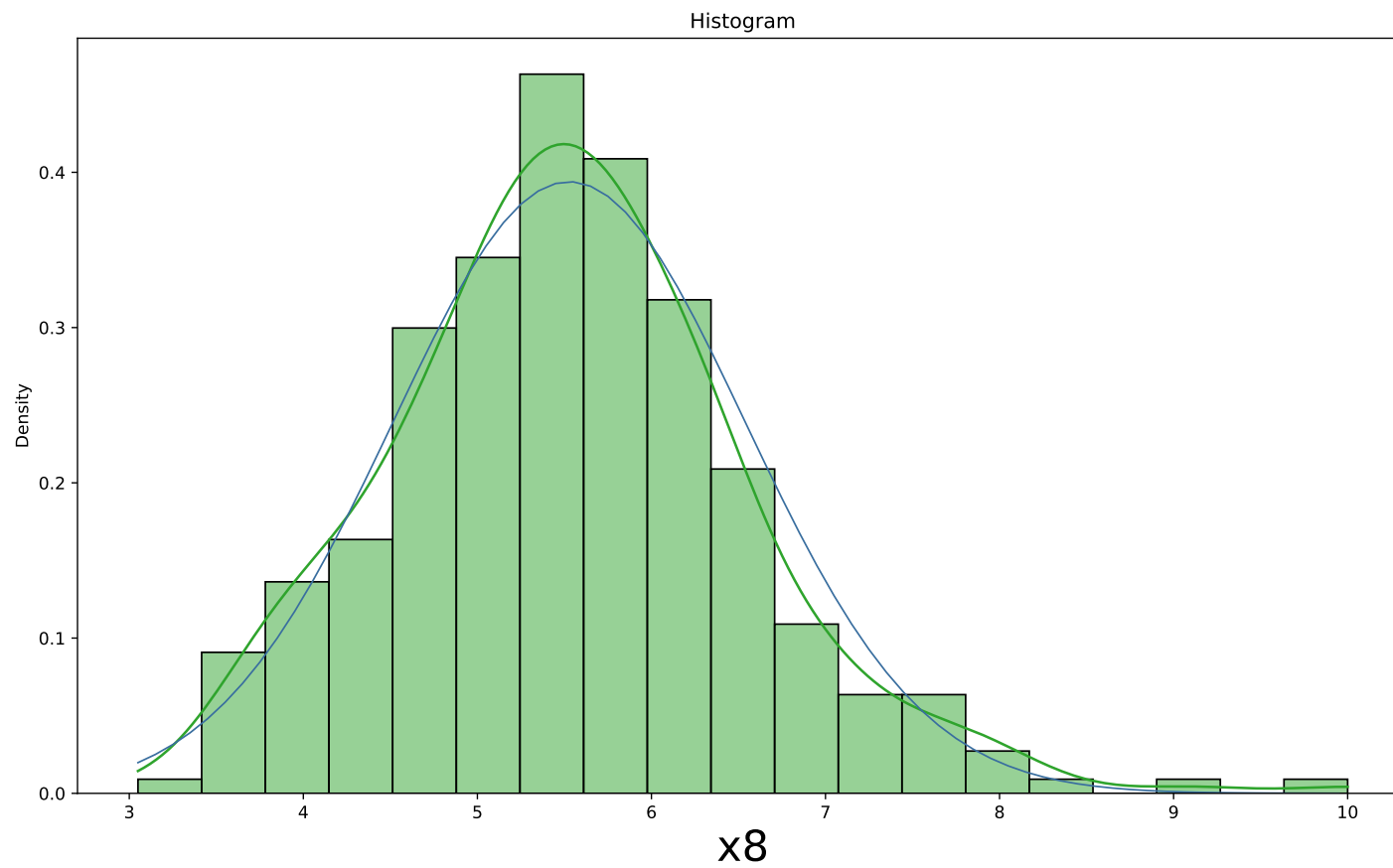


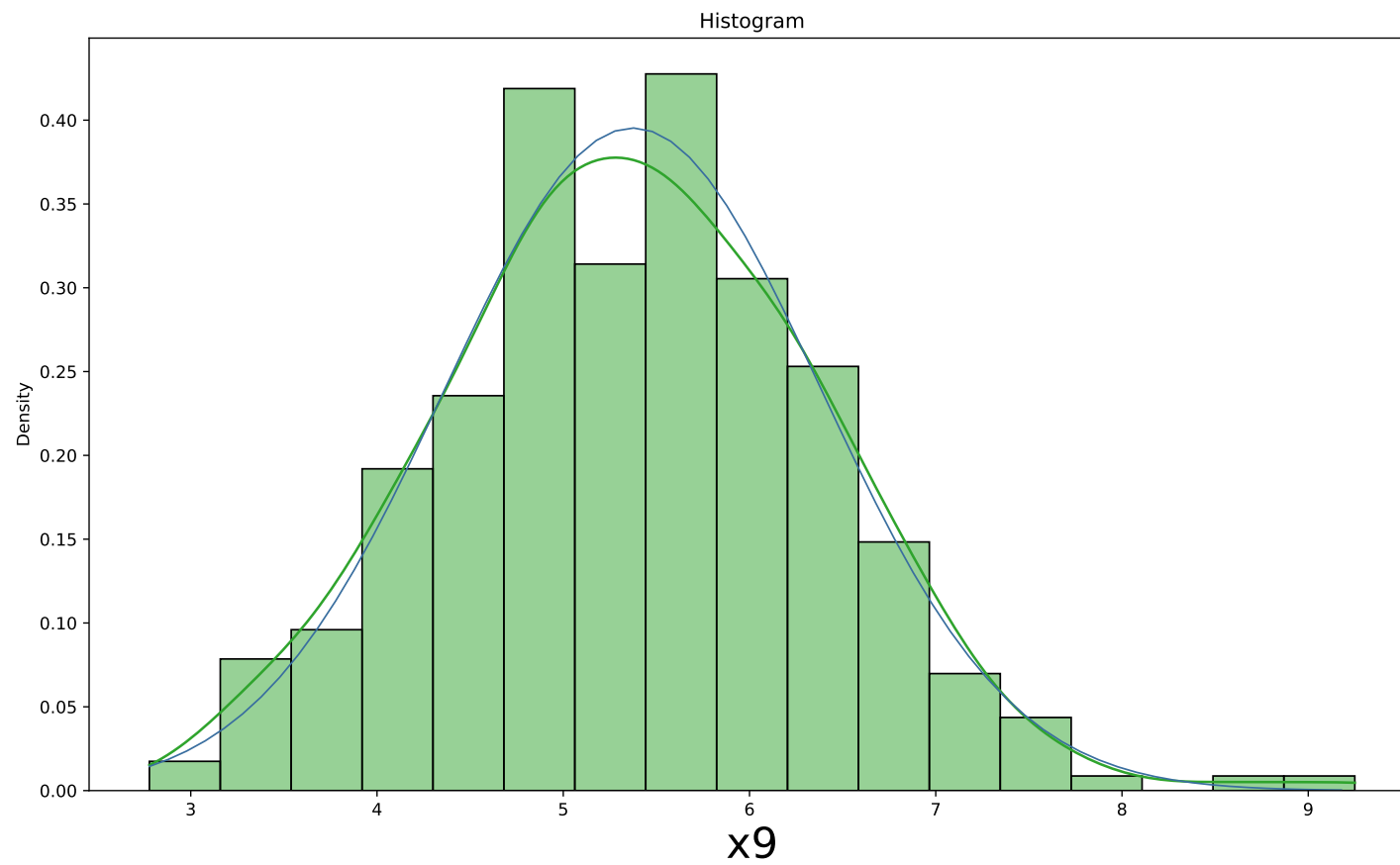








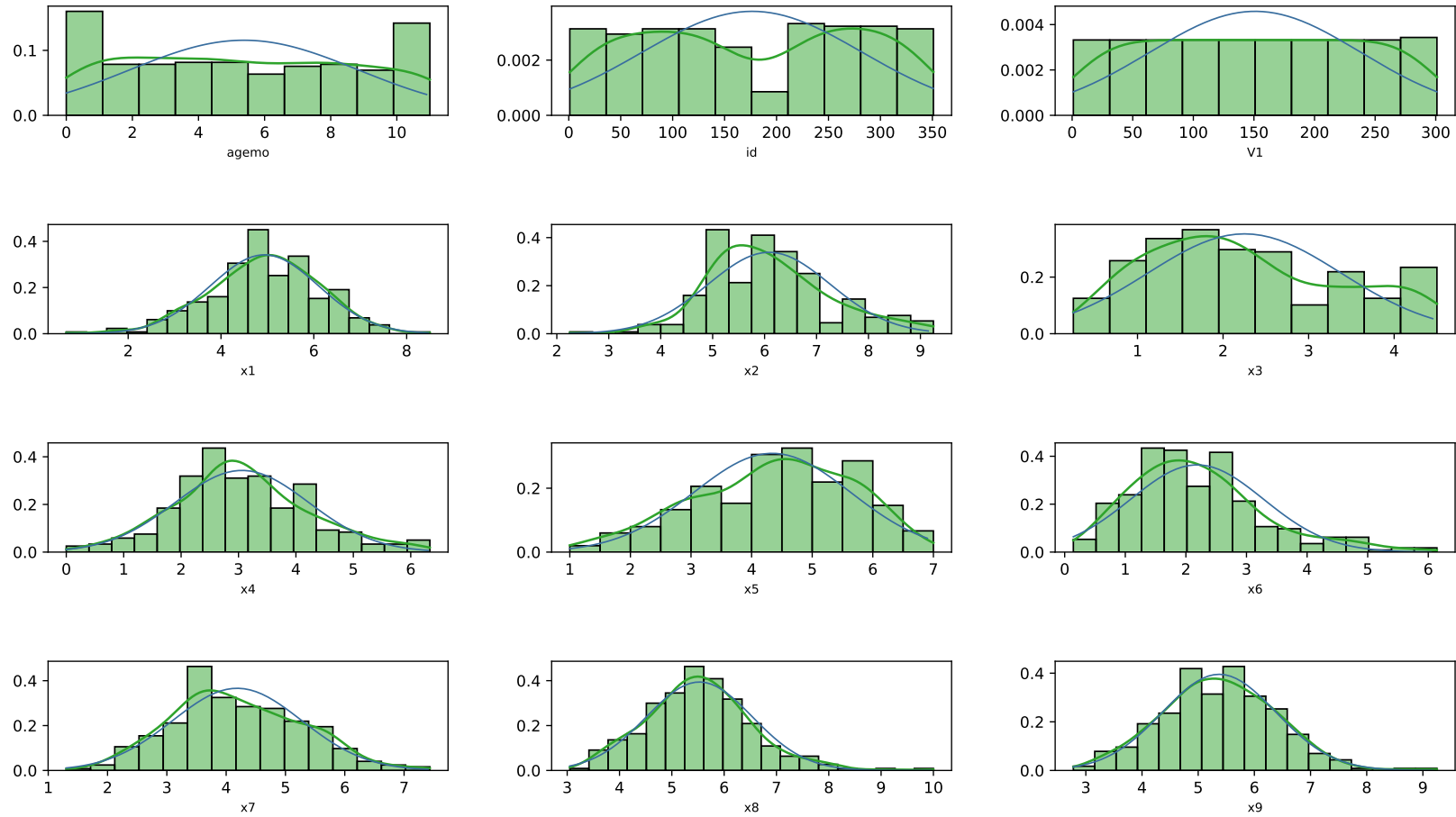






## Histograms Summary

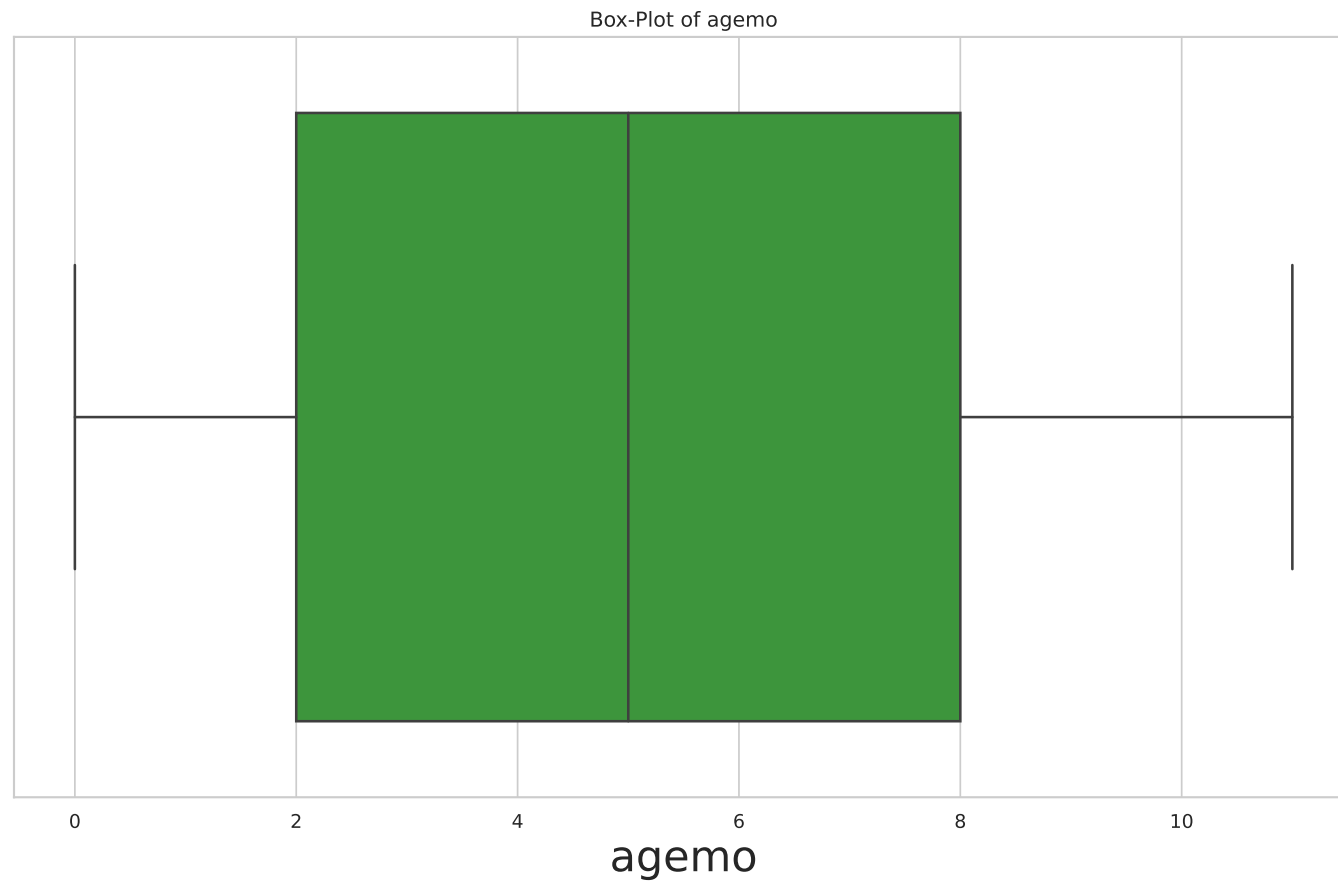
Multiple Relative Frequency Histogram in one figure. Variables are sorted alphabetically. The blue line represents the normal density approximation. The green line represents a special kernel density approximation.

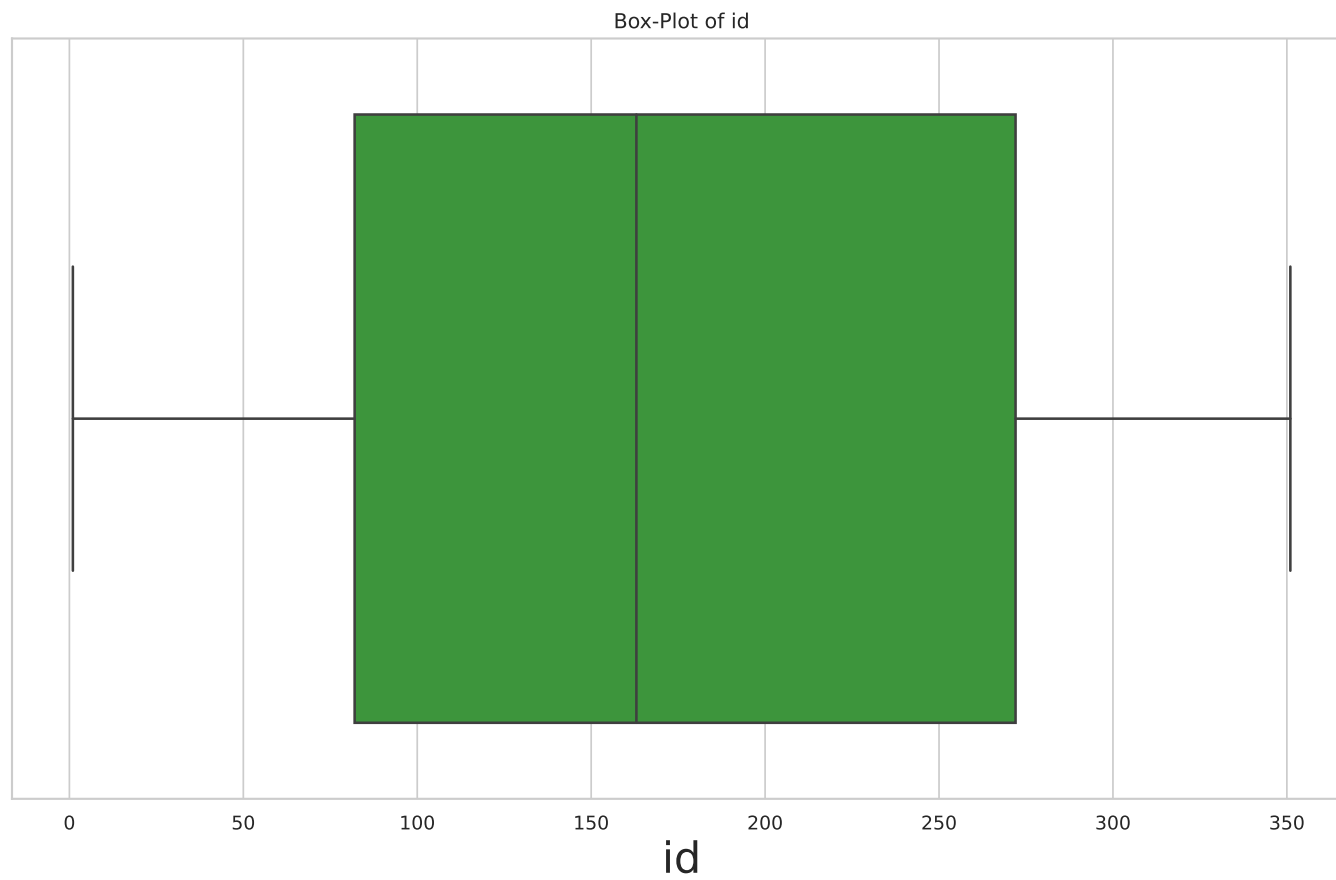


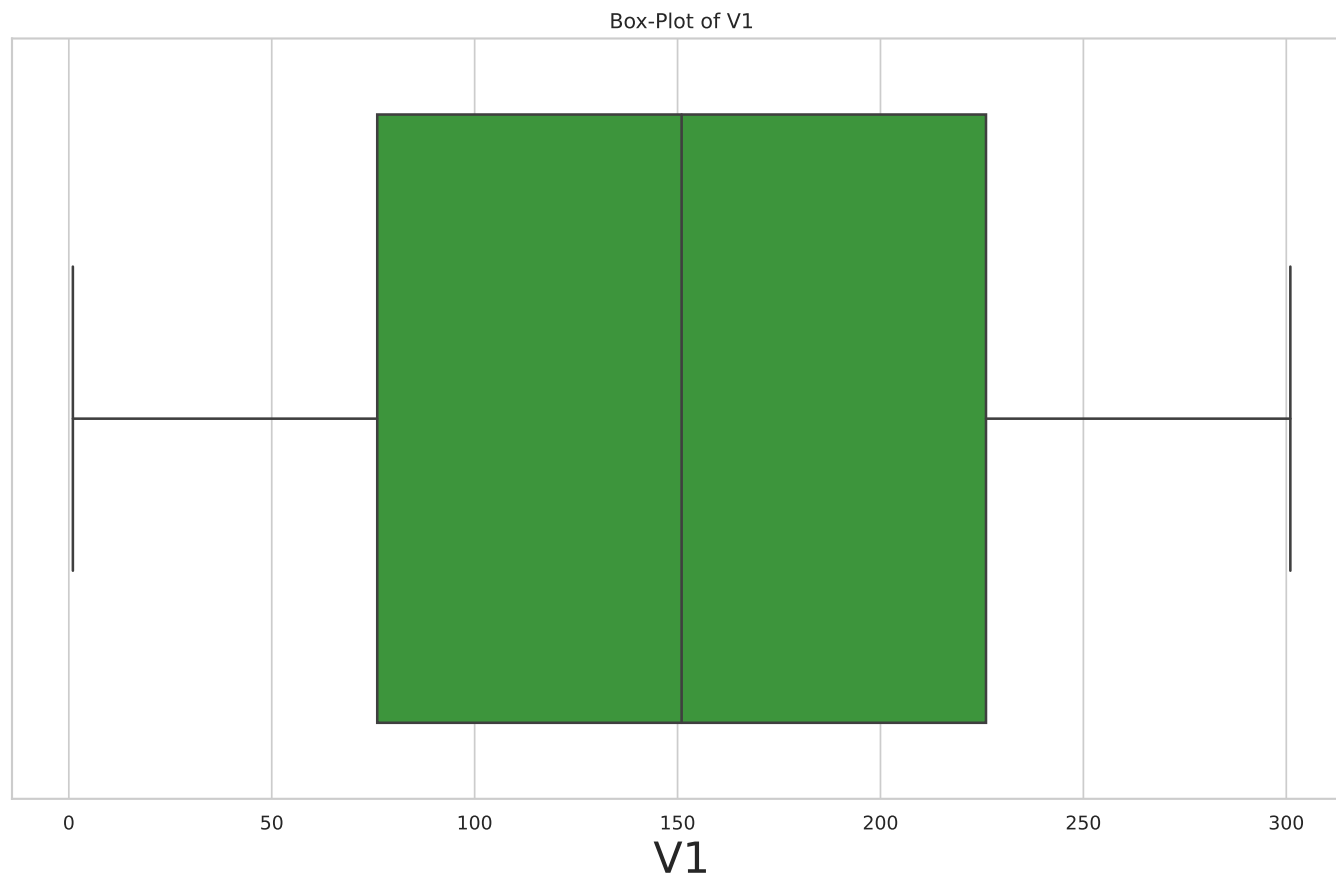
## Box-Plots

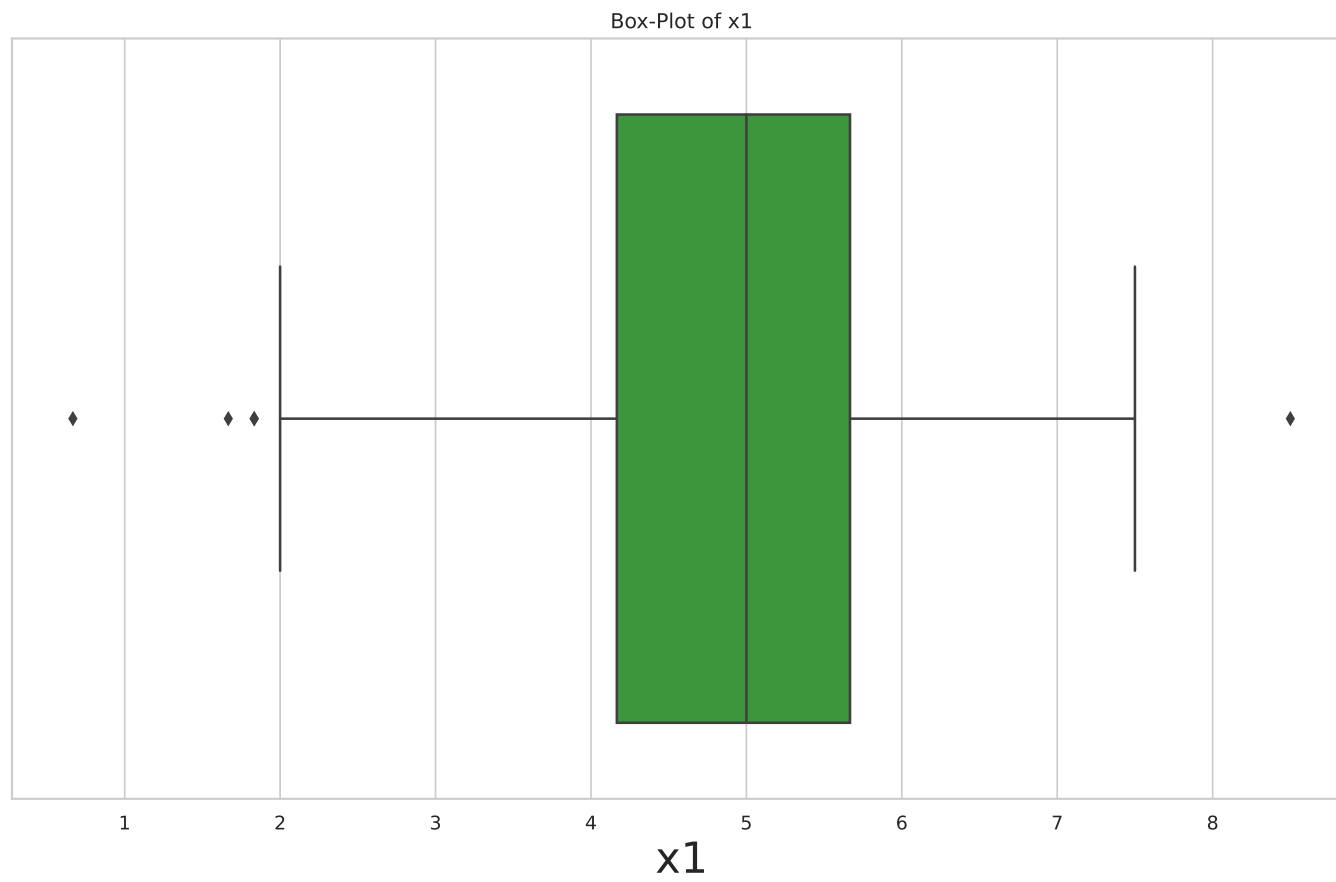
One Box-Plot per page for each variable. Variables are sorted alphabetically.

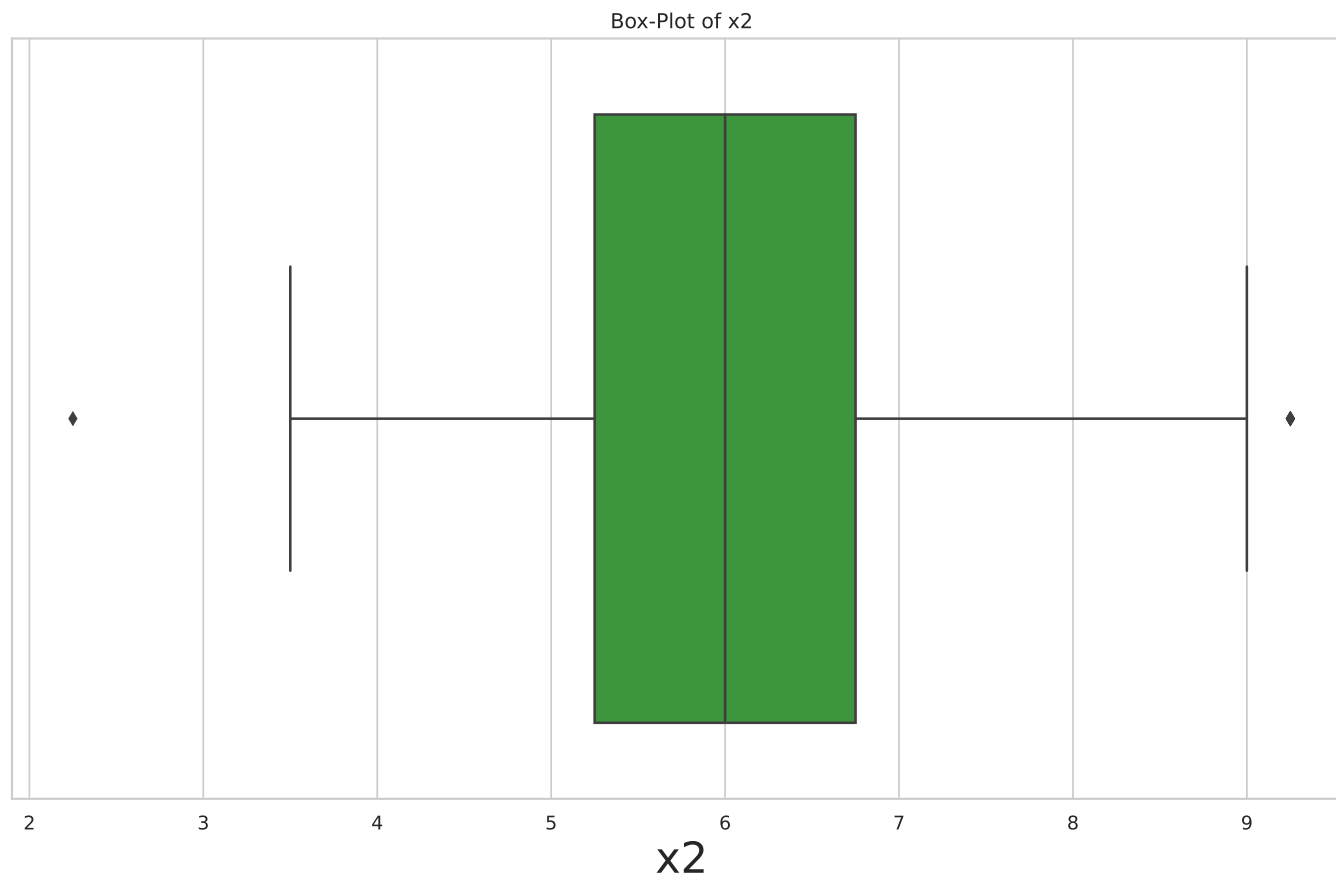
□

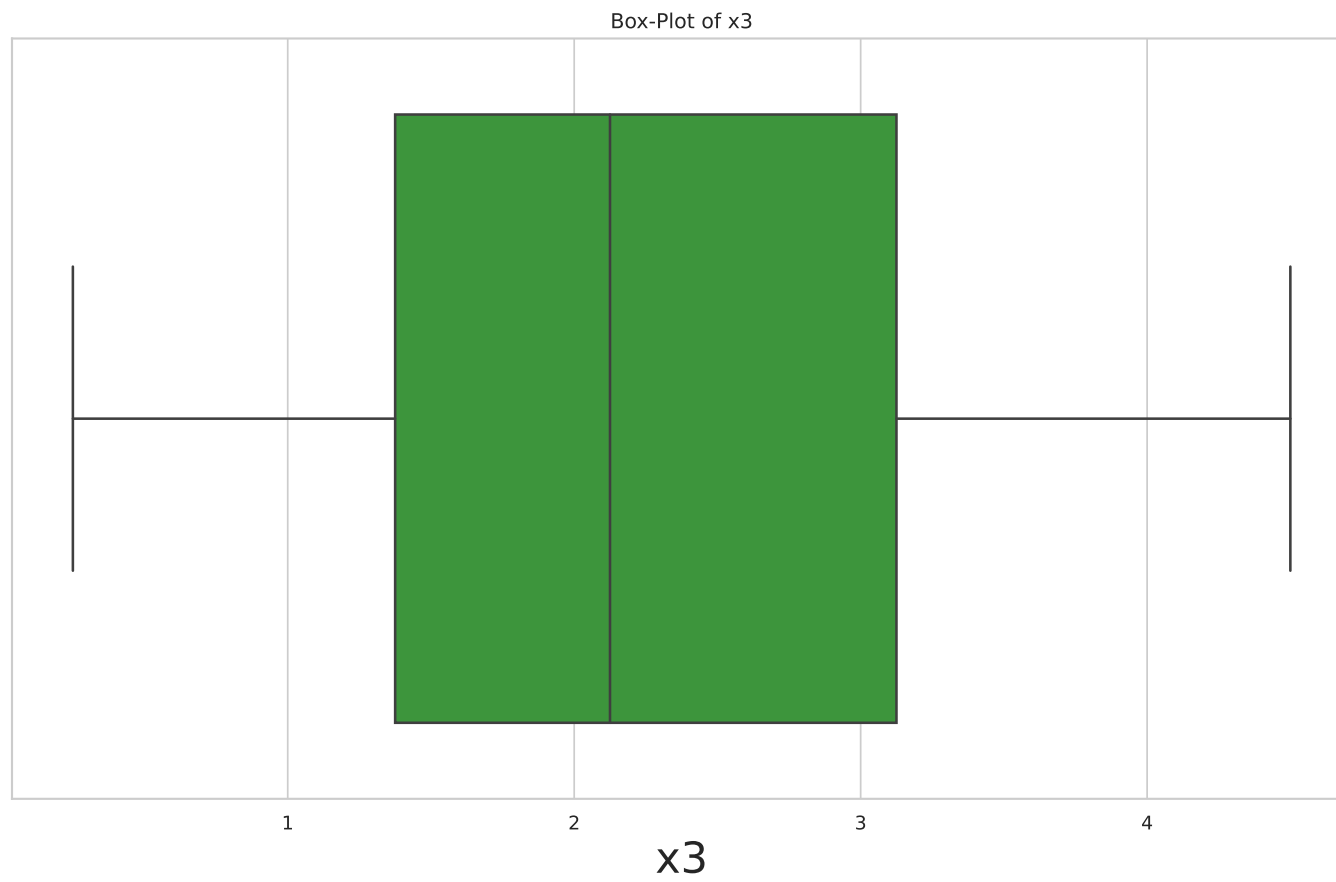


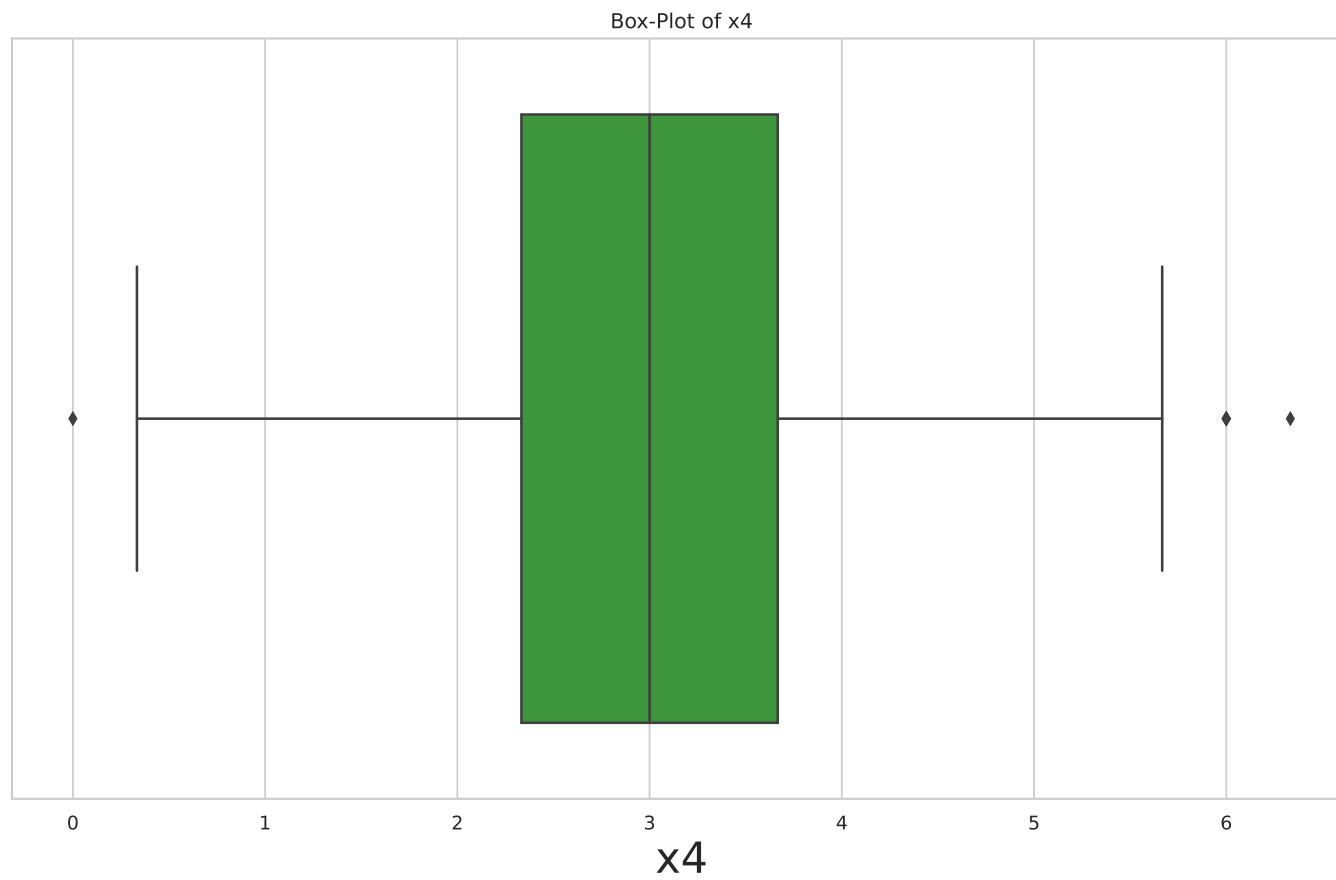




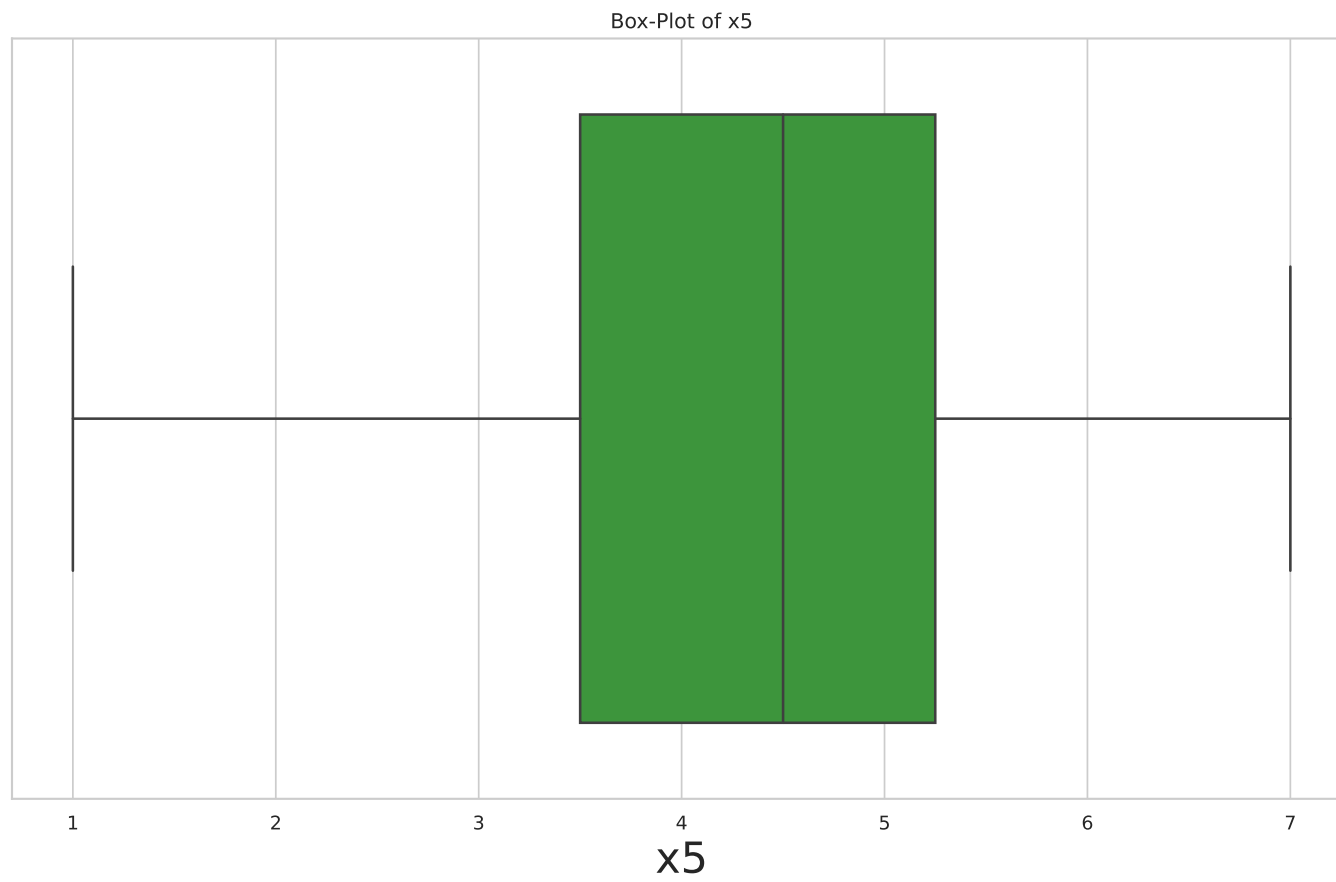


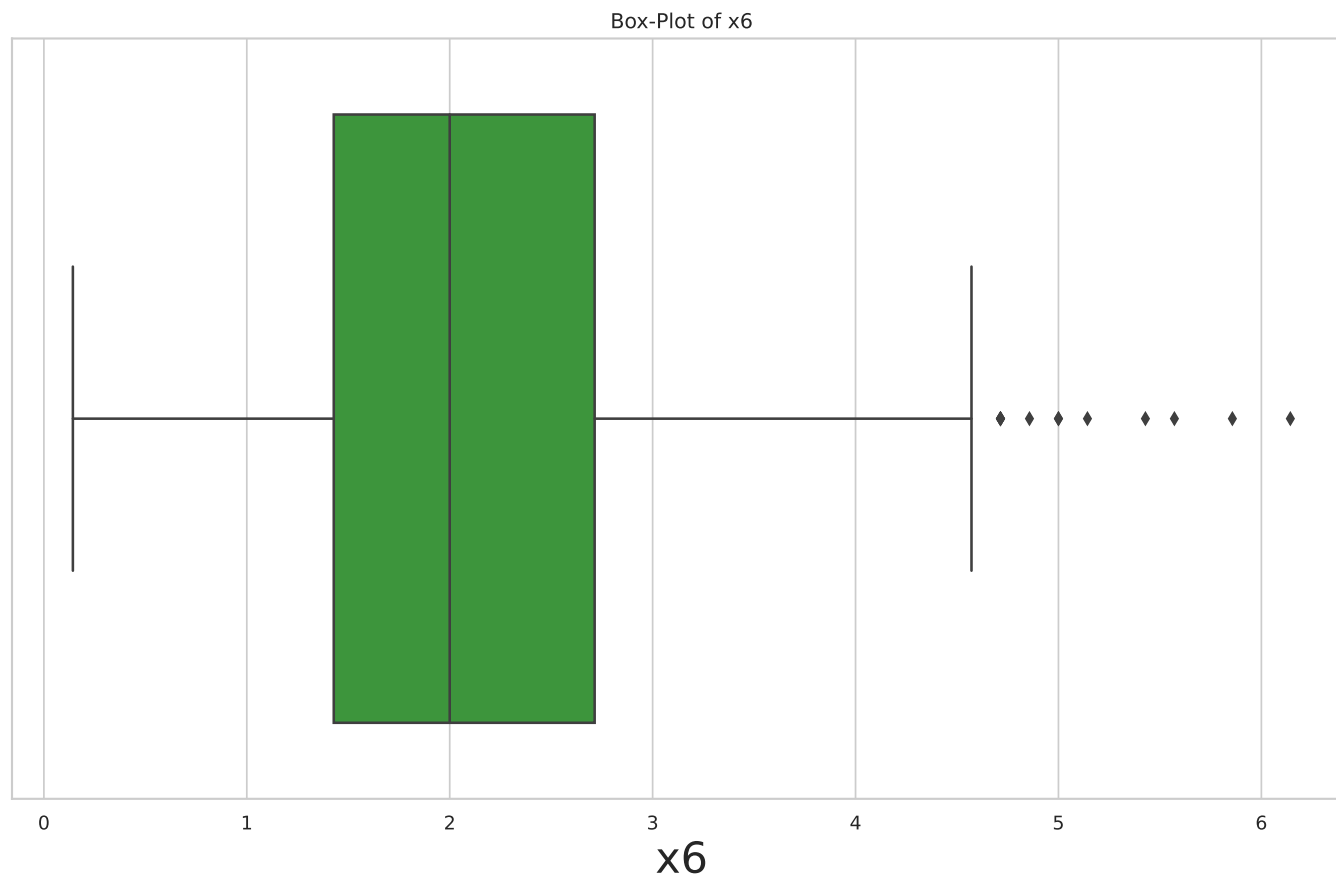


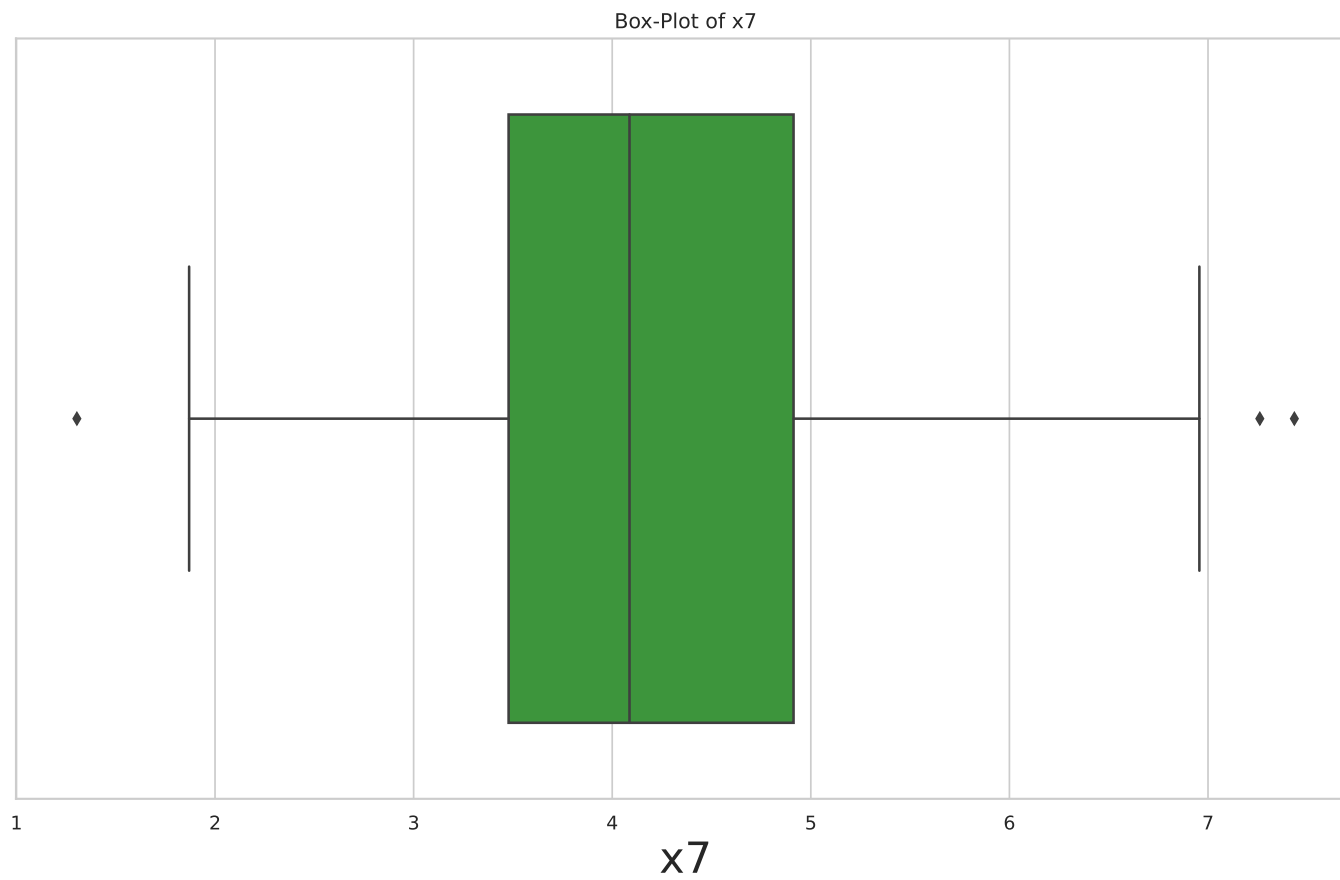


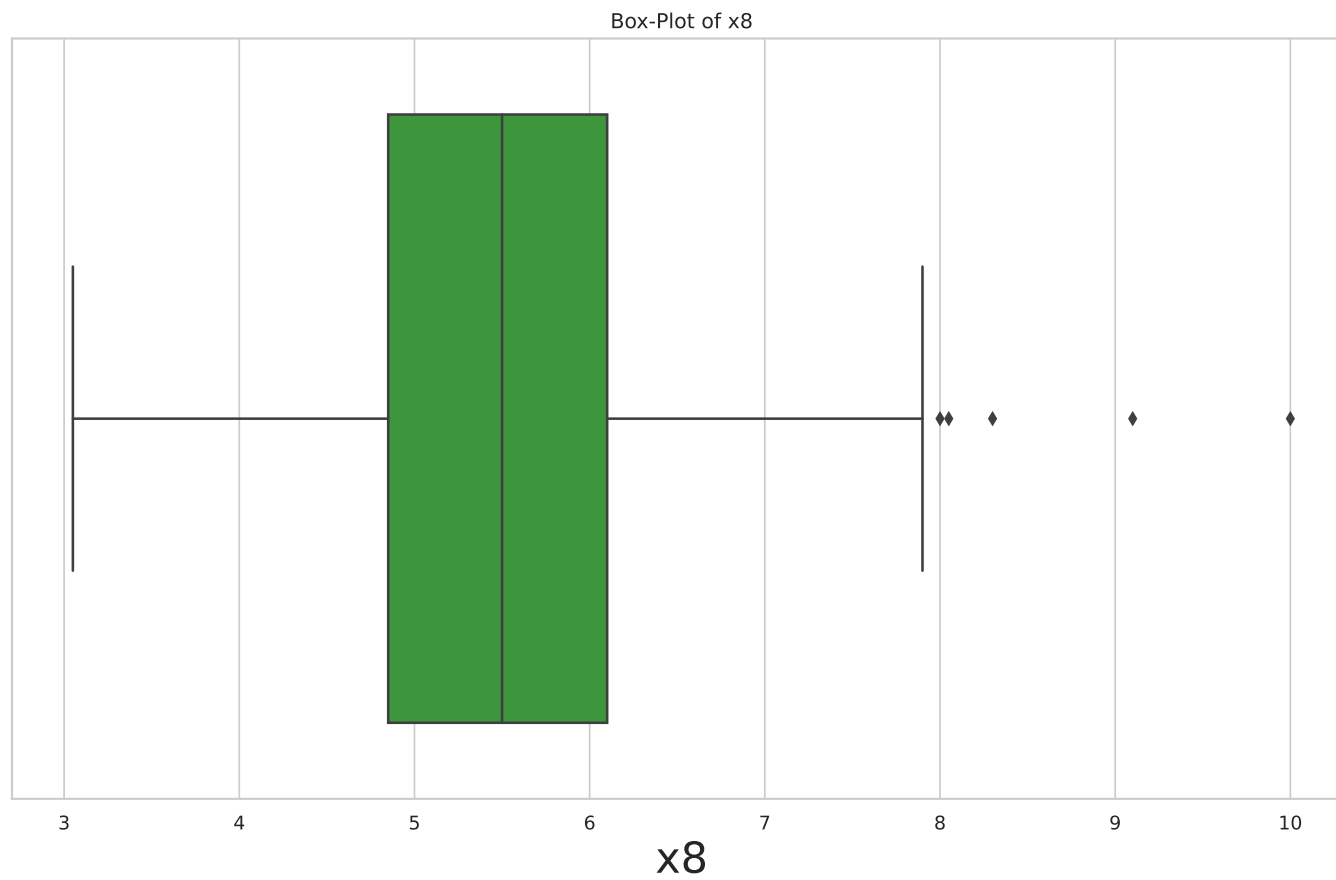


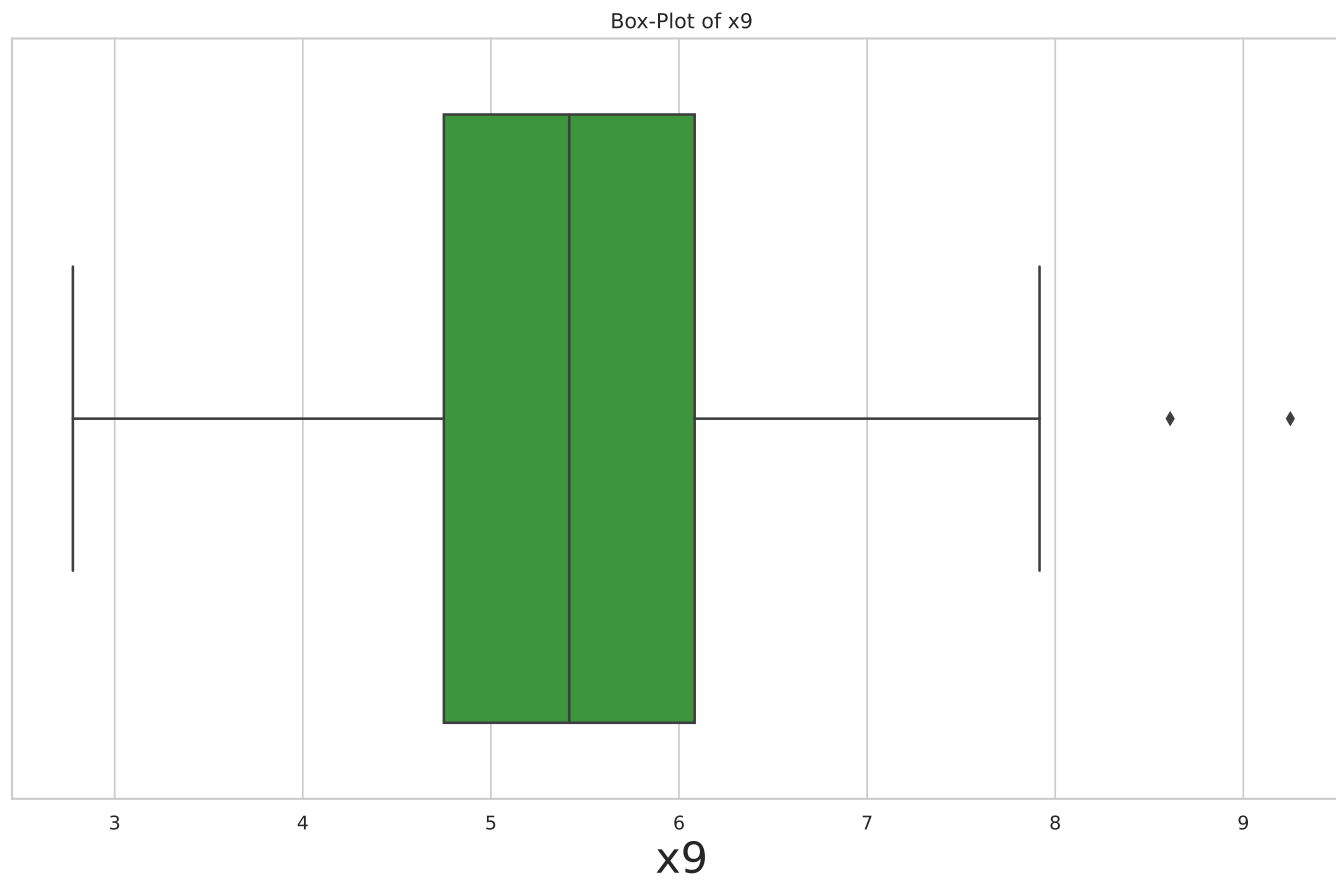






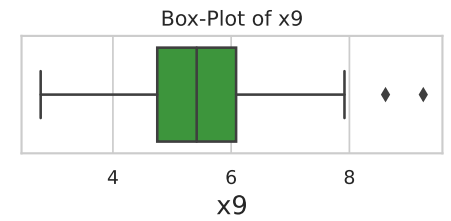
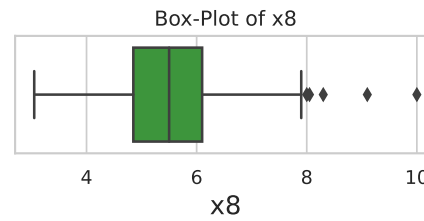
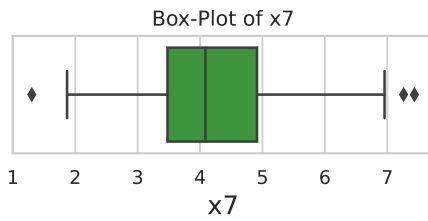
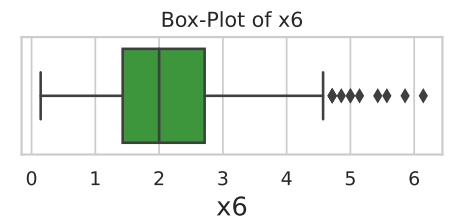
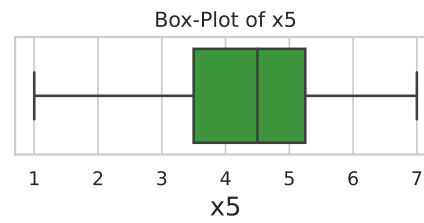
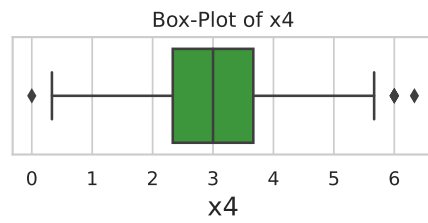
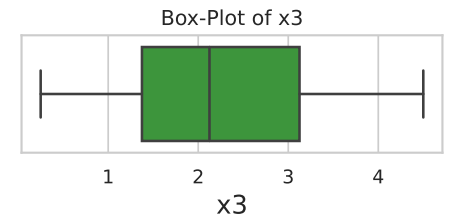
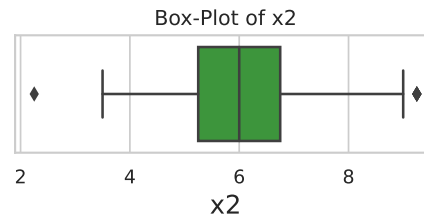
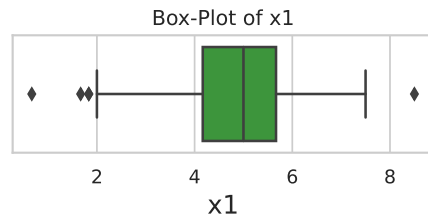
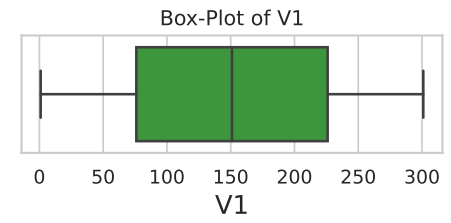
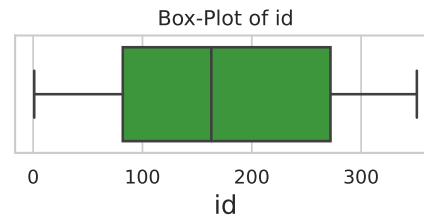
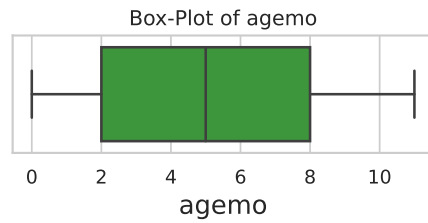






## Box-Plots Summary

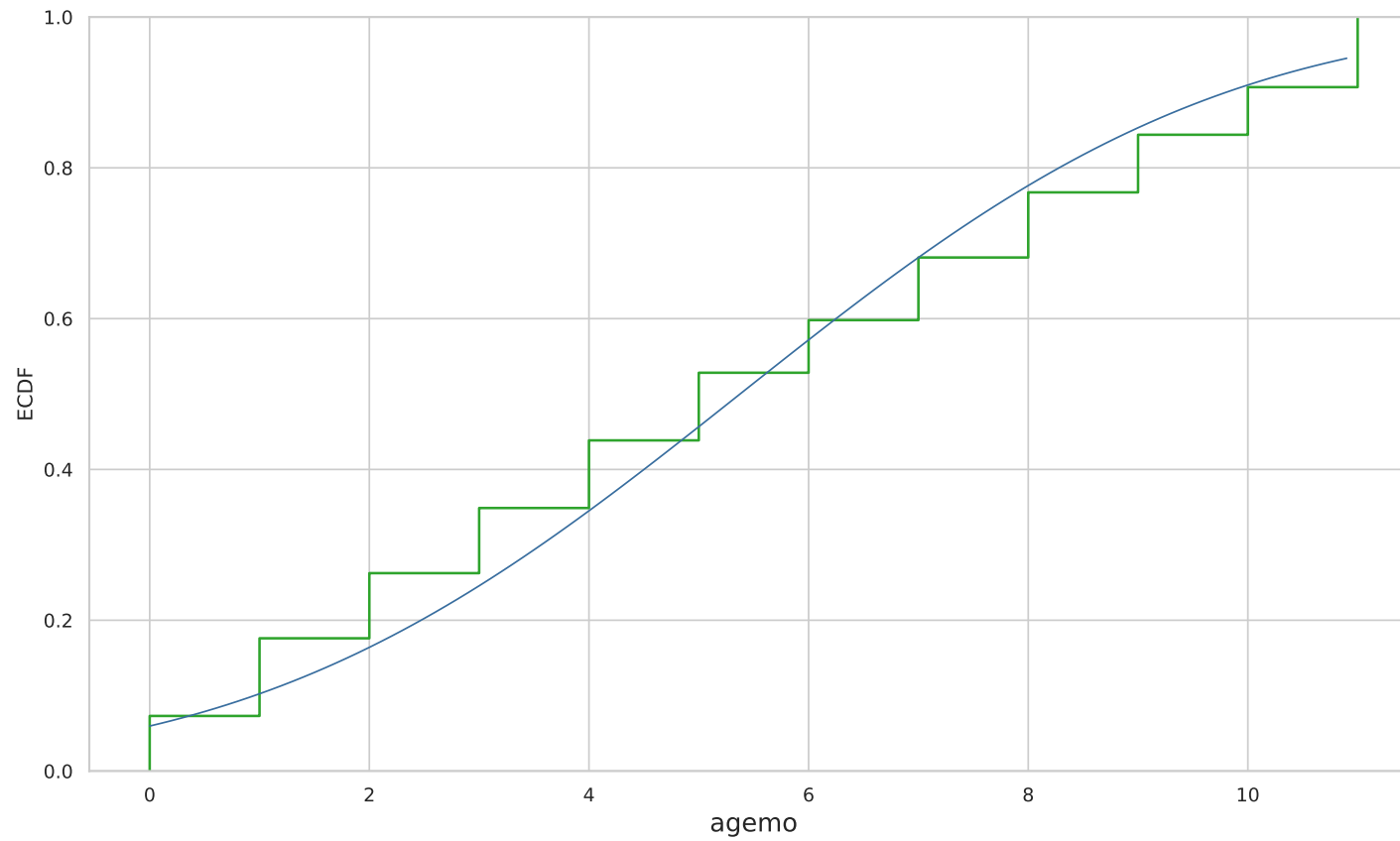
Multiple Box-Plots of variables in one figure. Variables are sorted alphabetically.

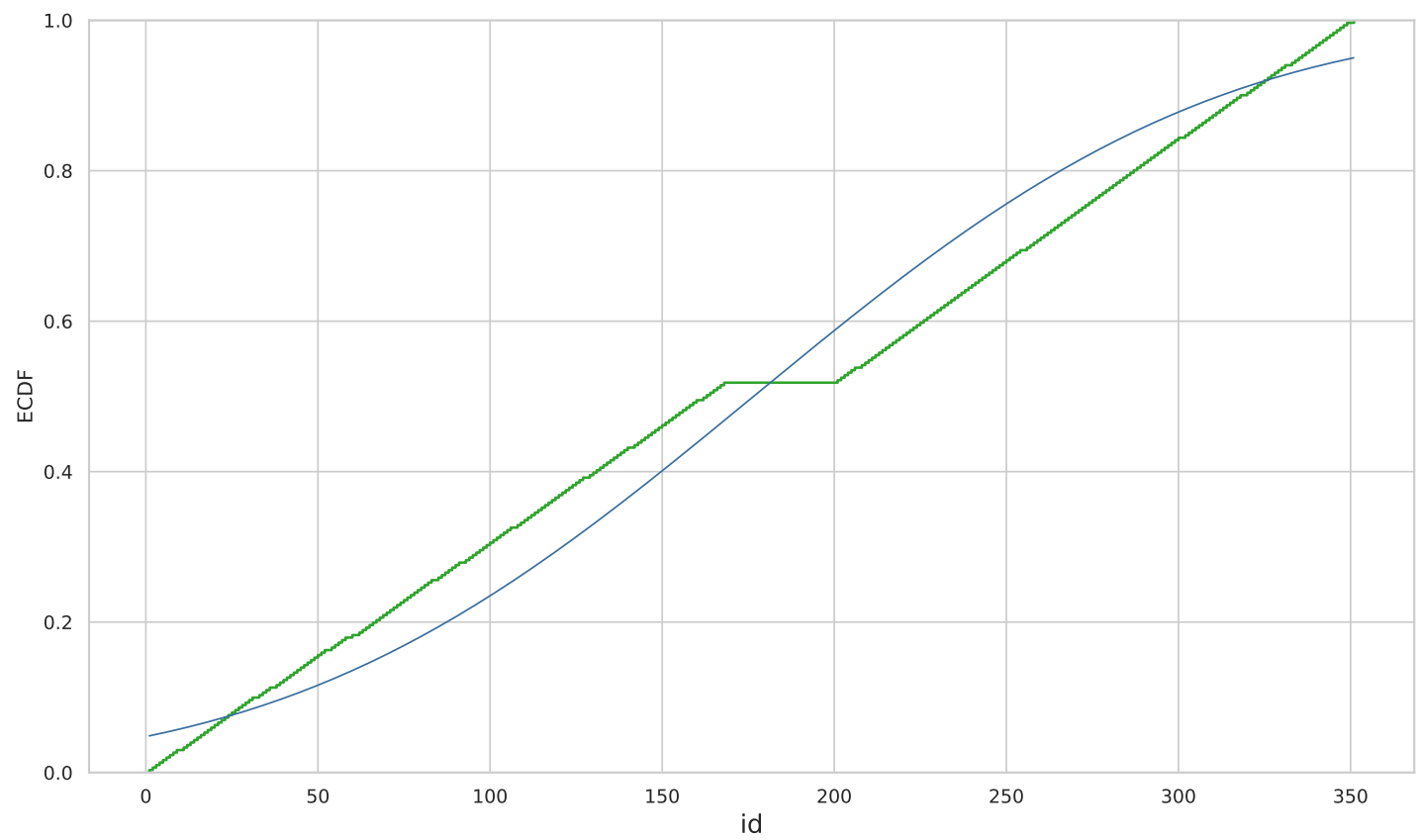


## ECDF Plots

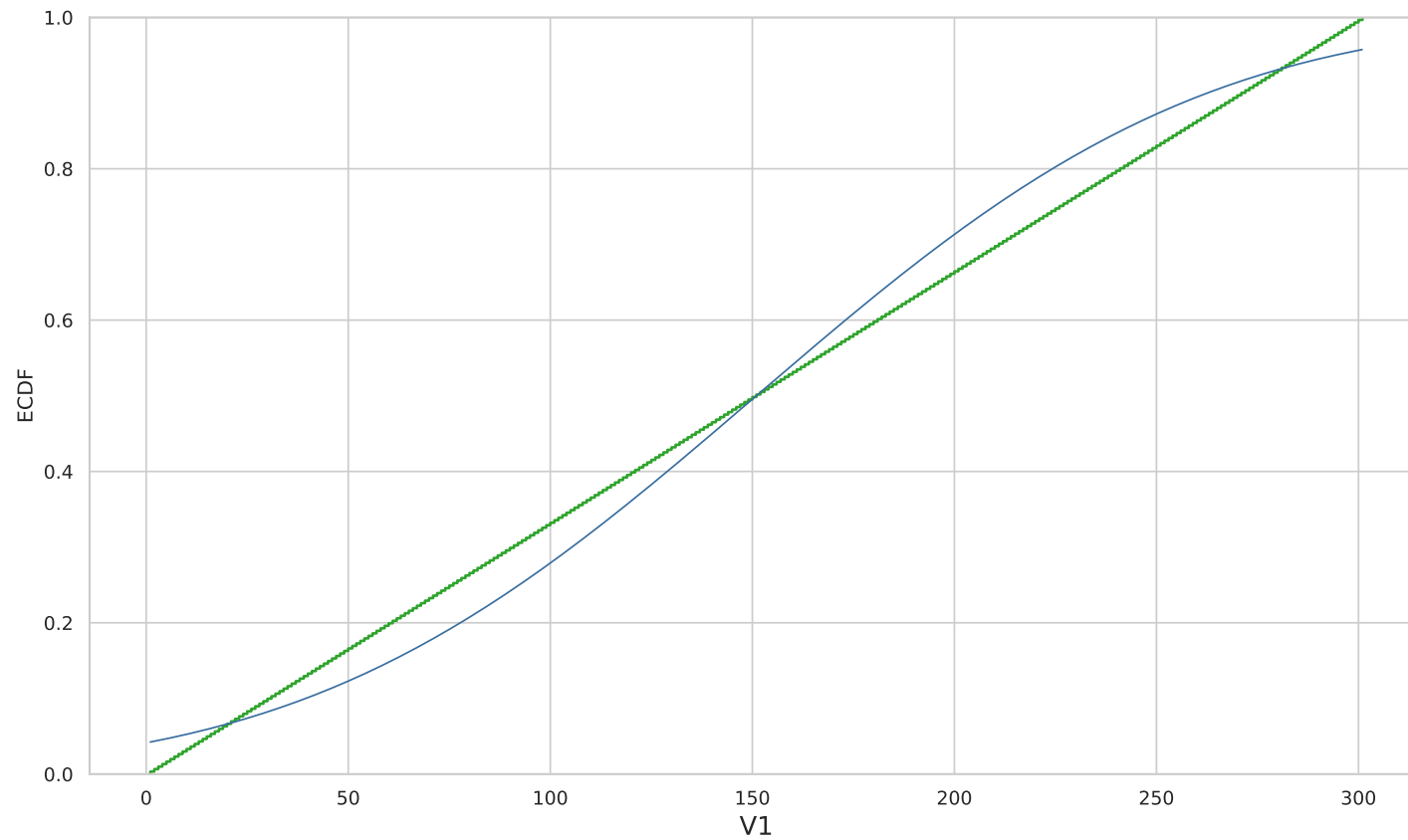
One ECDF (Empirical Cumulative Distribution Function) Plot per page for each variable. Variables are sorted alphabetically. The blue line represents the CDF of a normal distribution. If the variable is normally distributed, the blue line approximates well the ECDF.

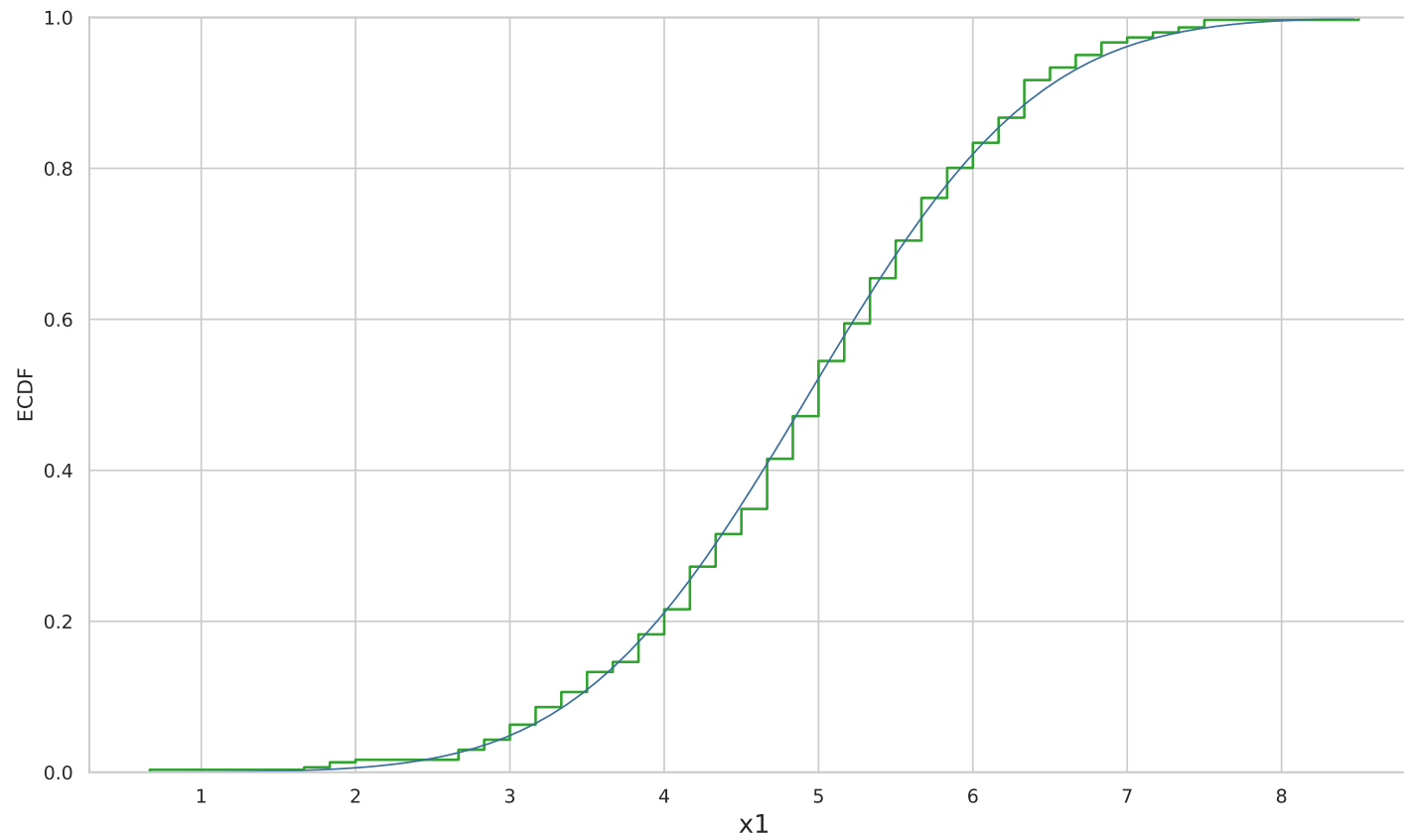
□

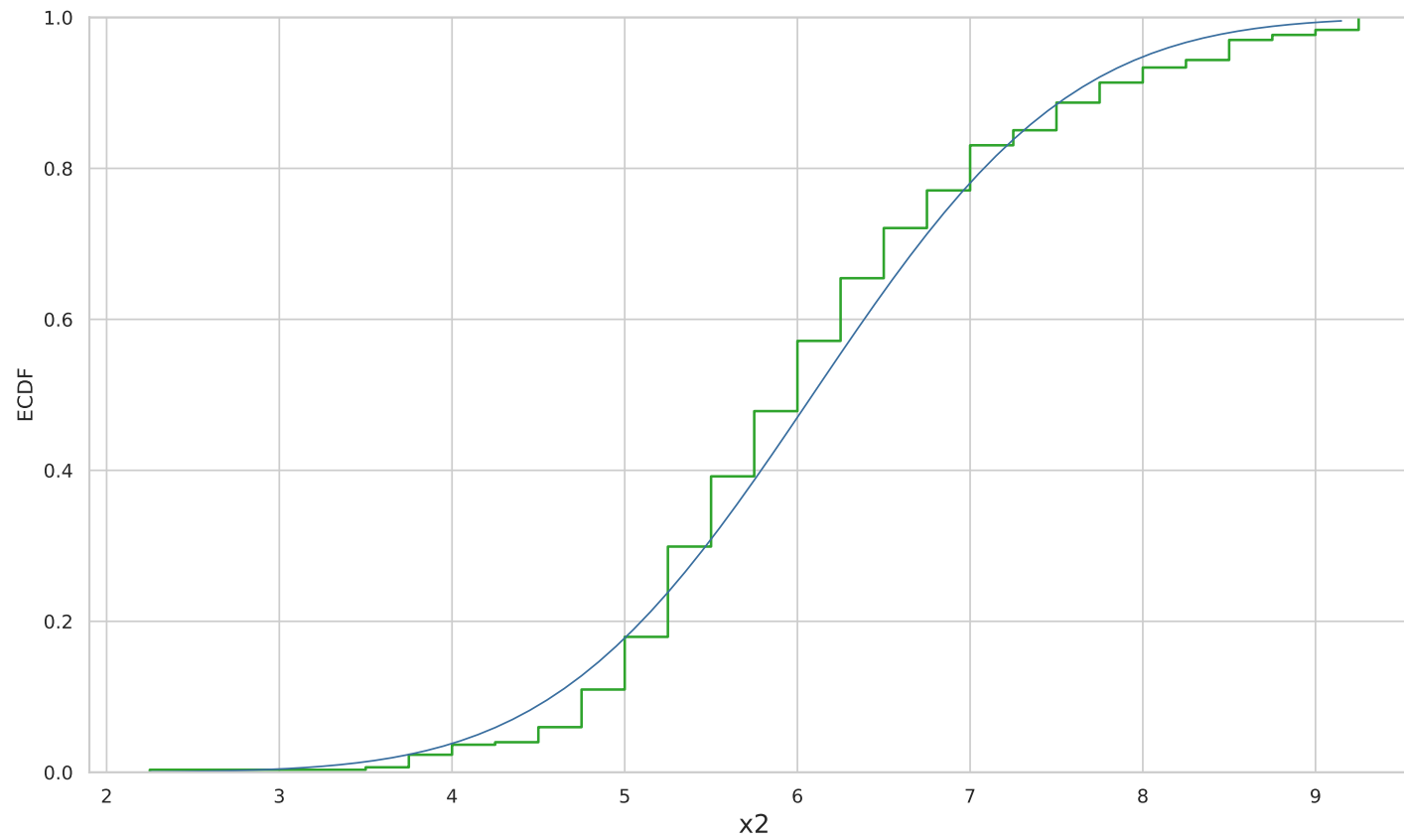


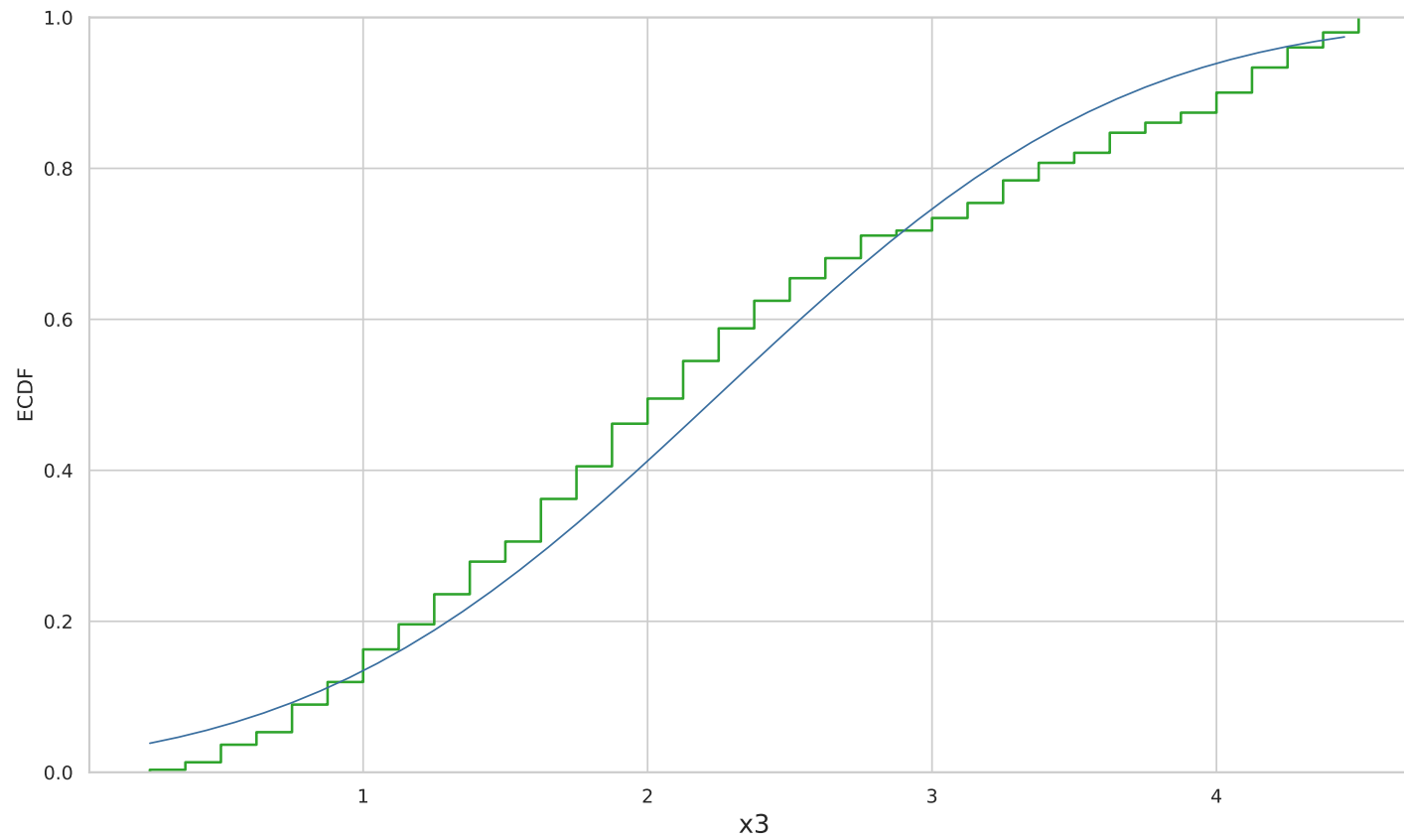


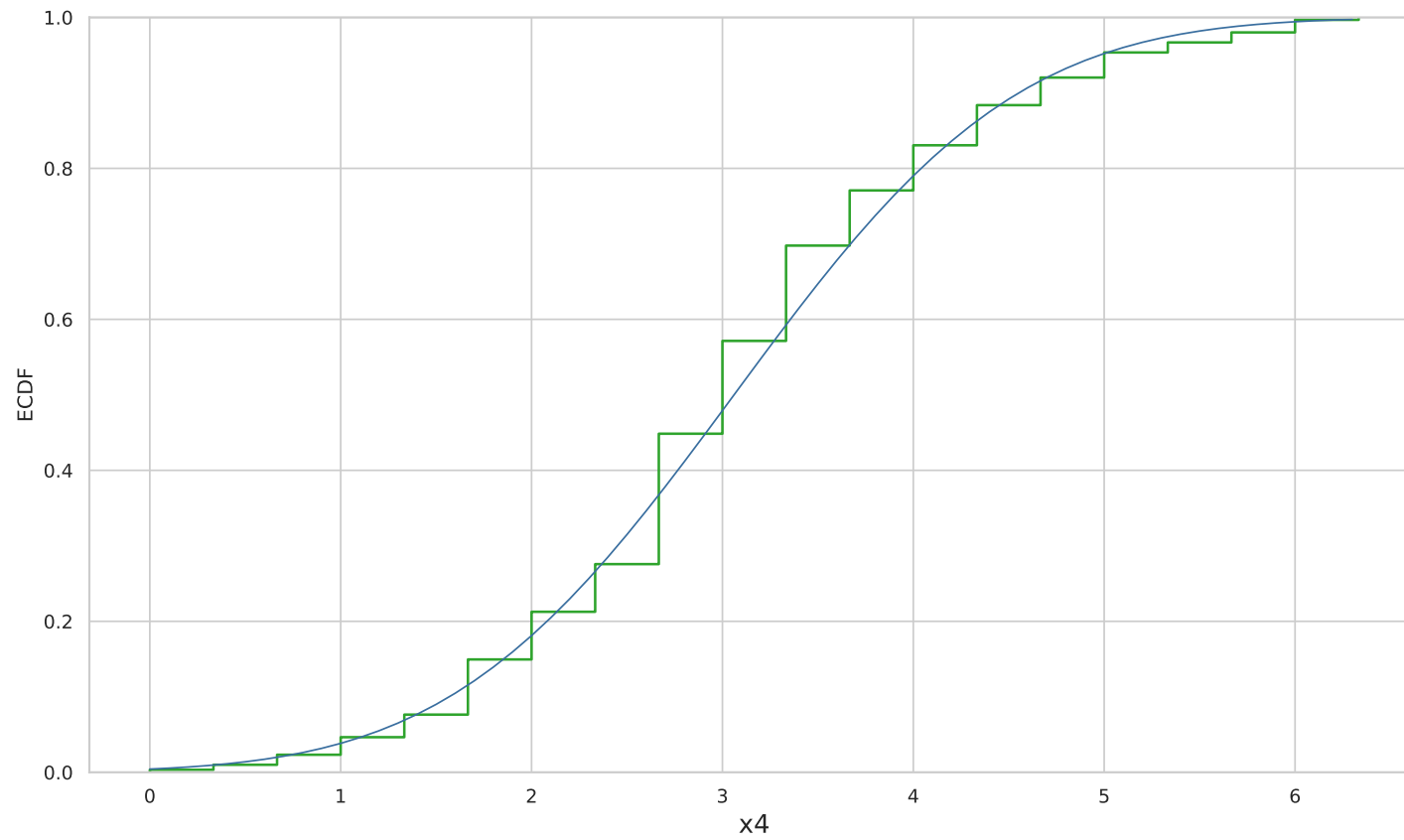


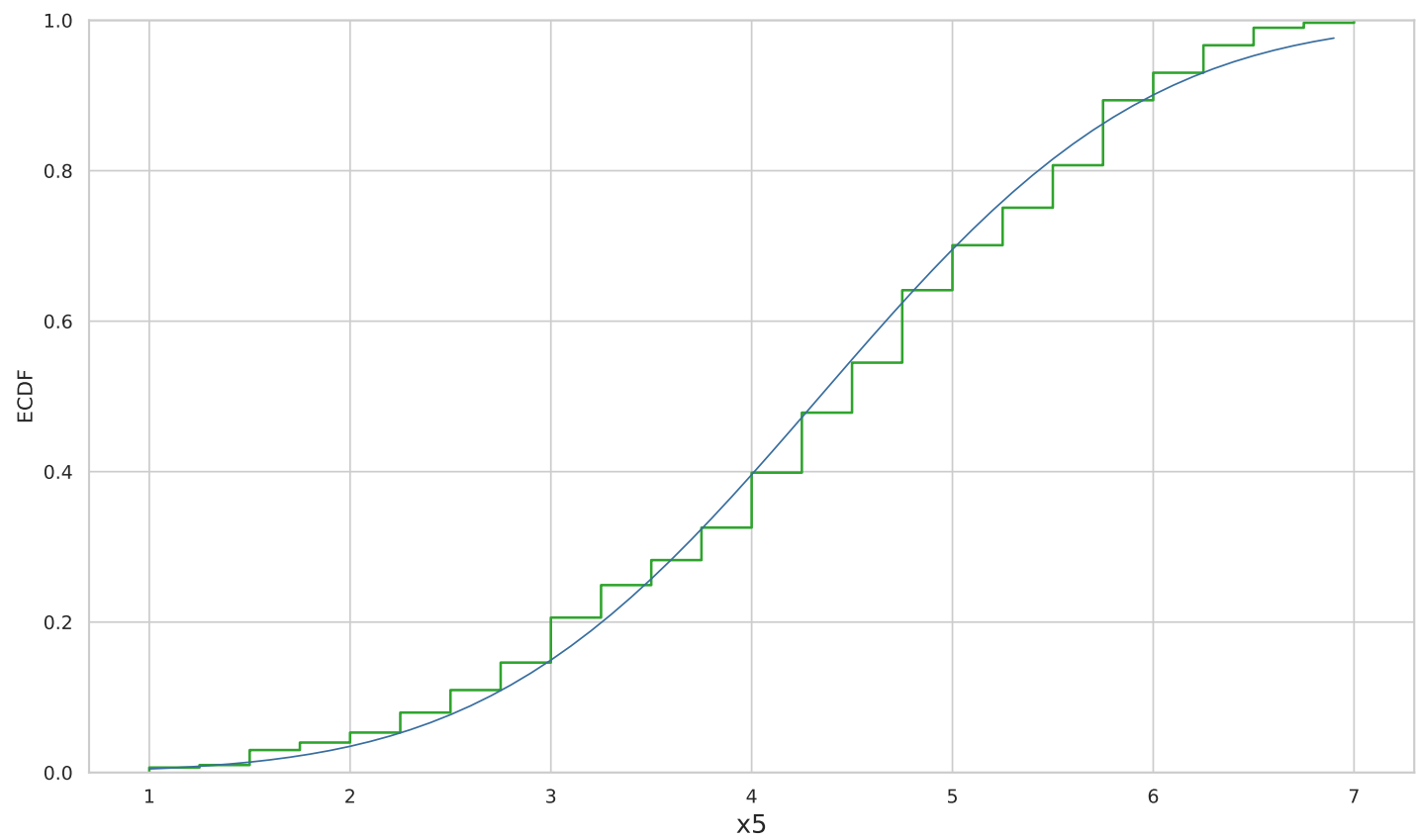


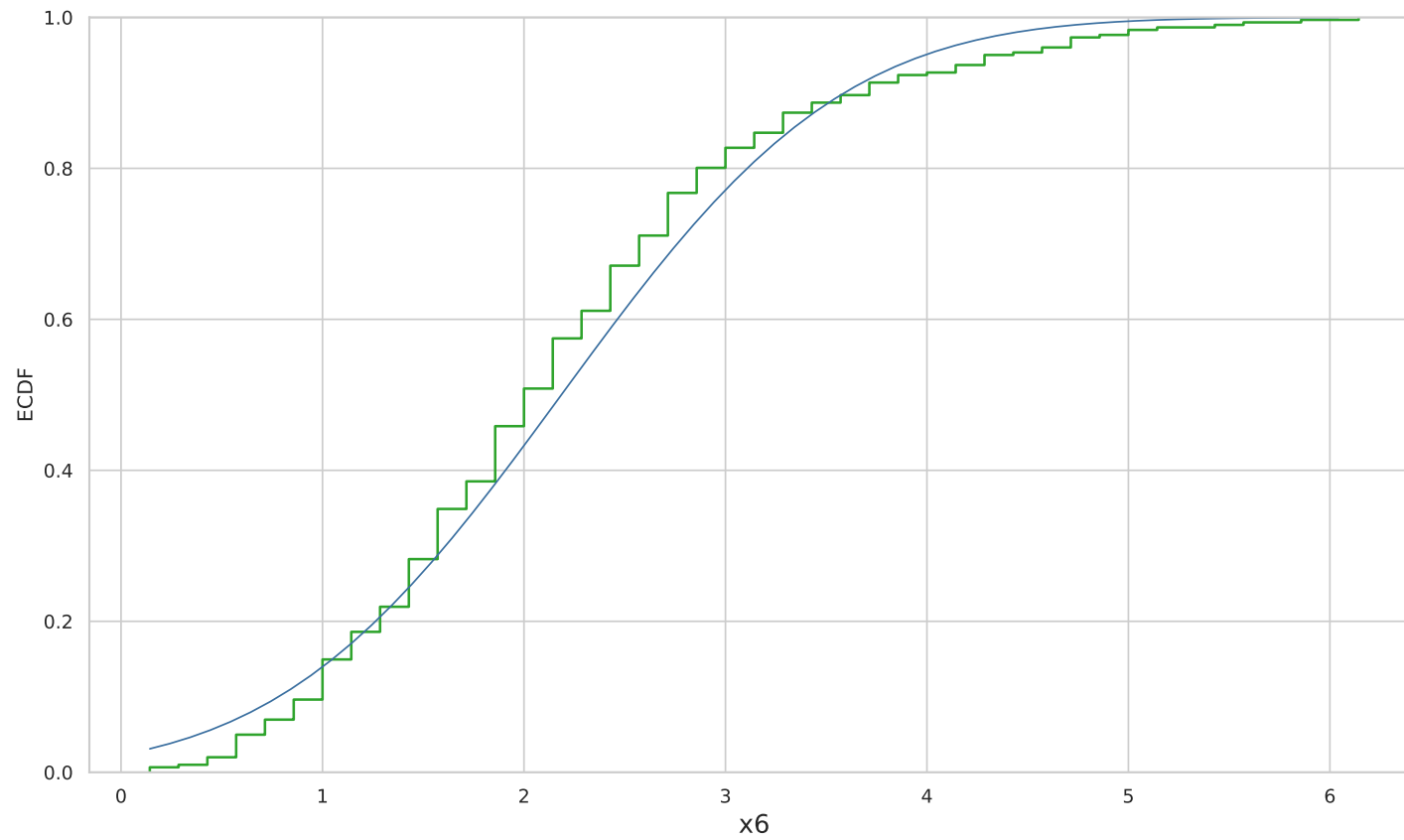


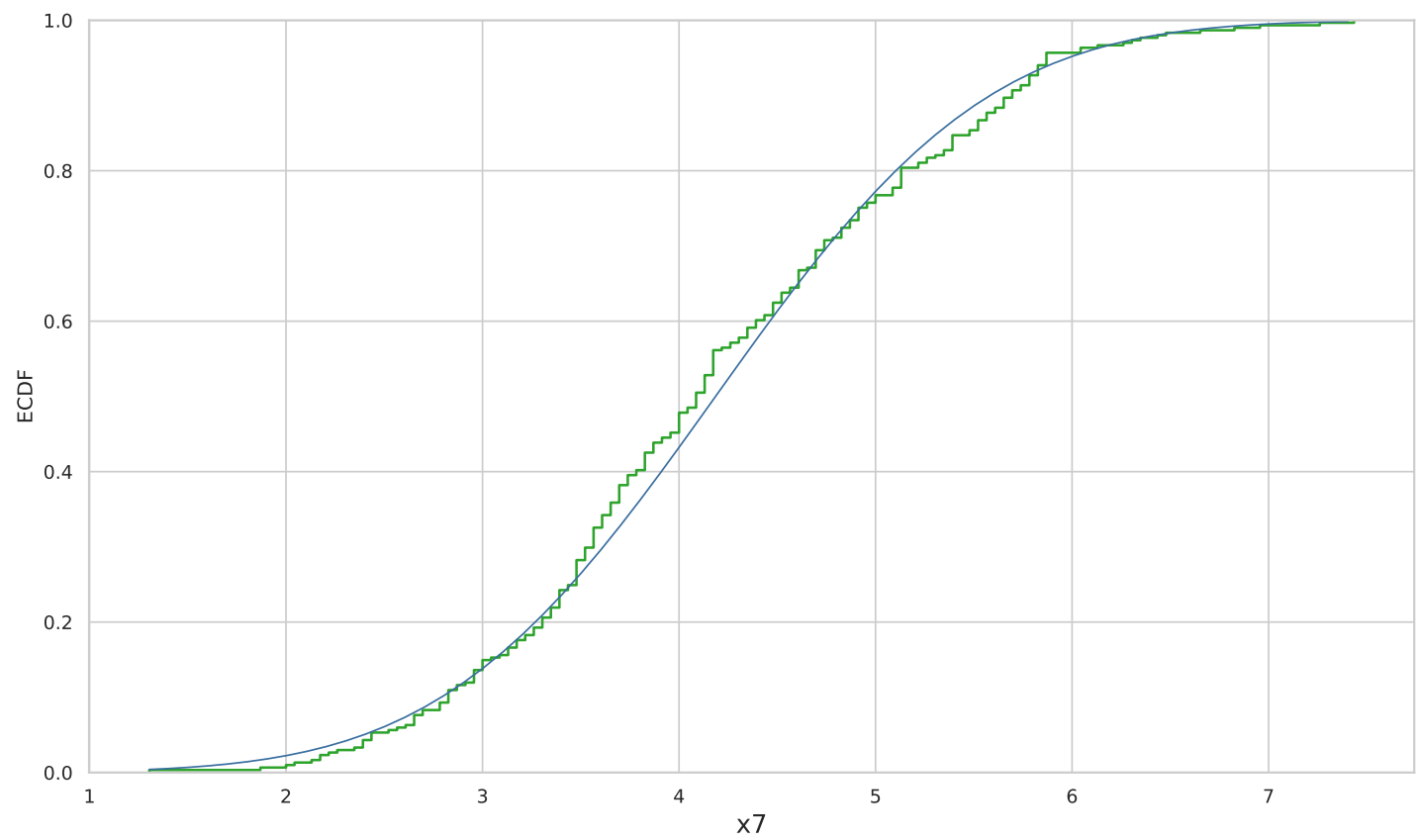




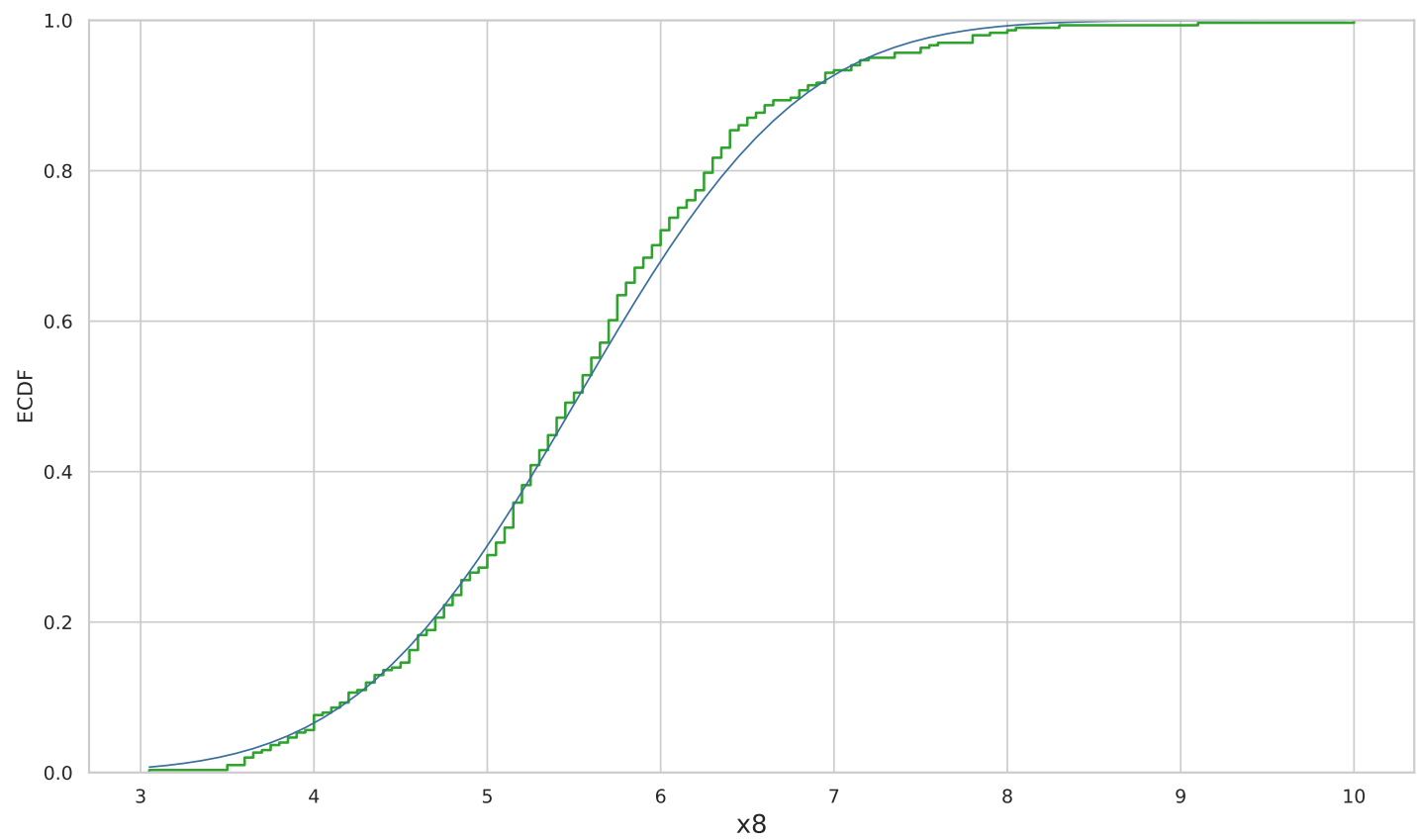


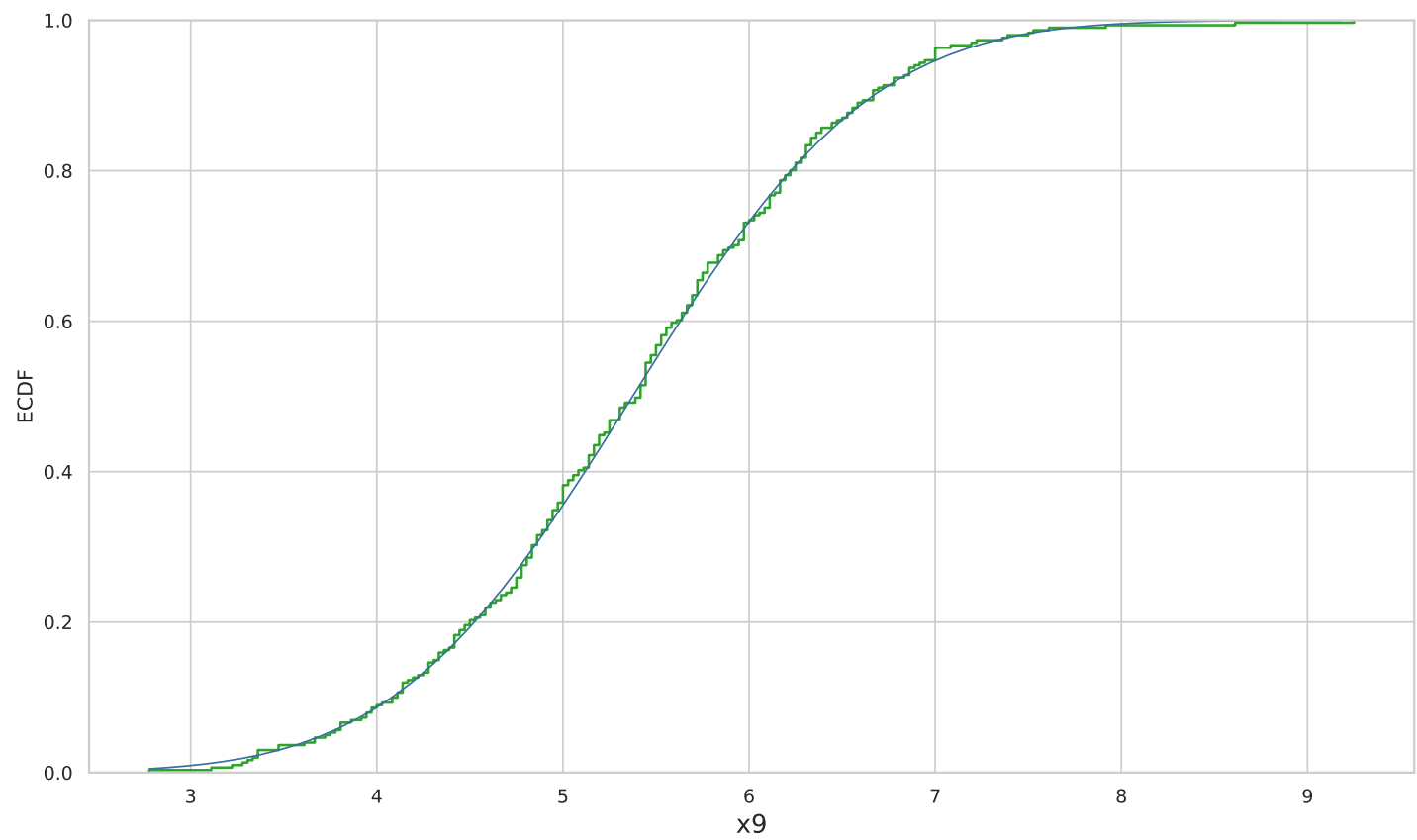






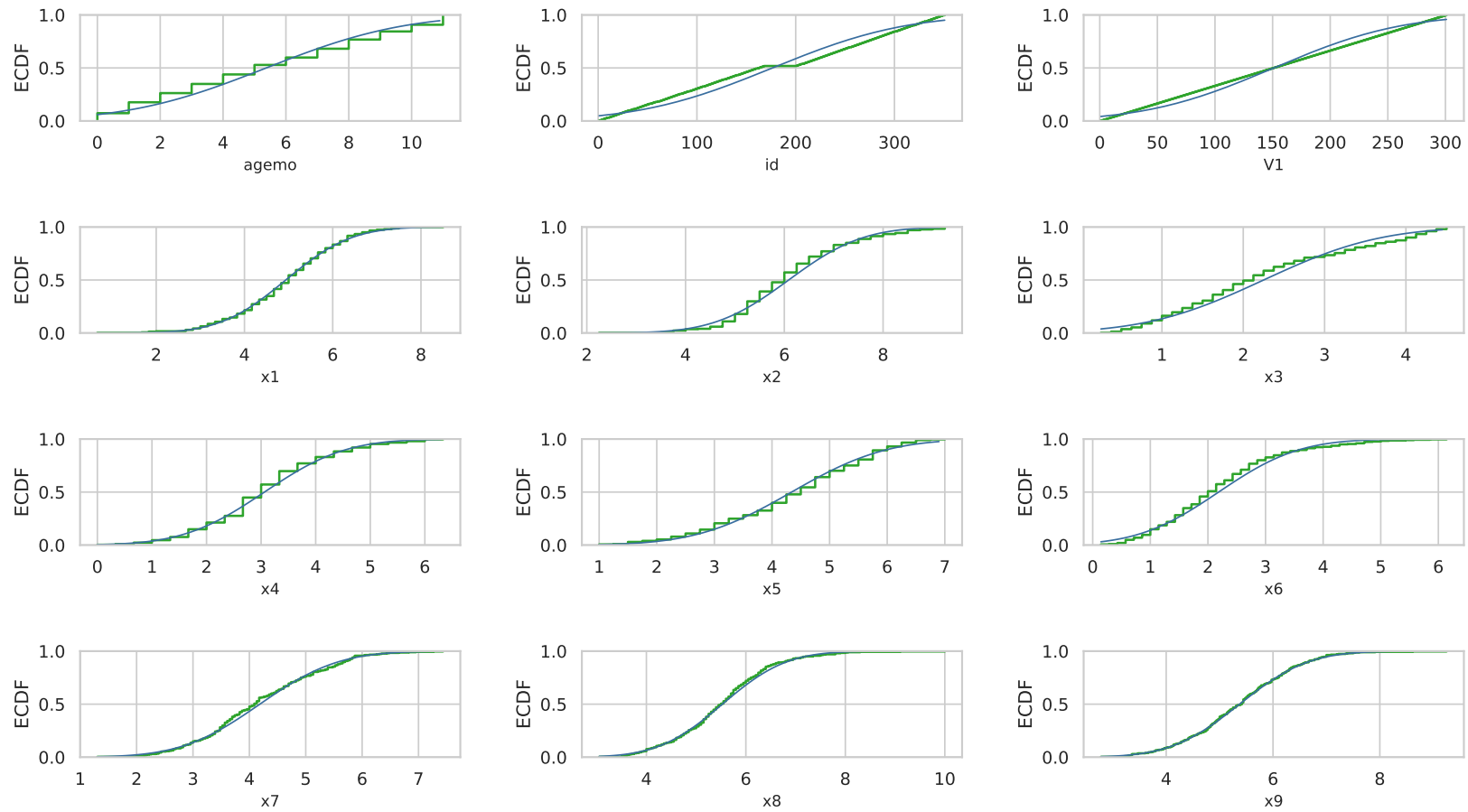






## ECDF Plots Summary

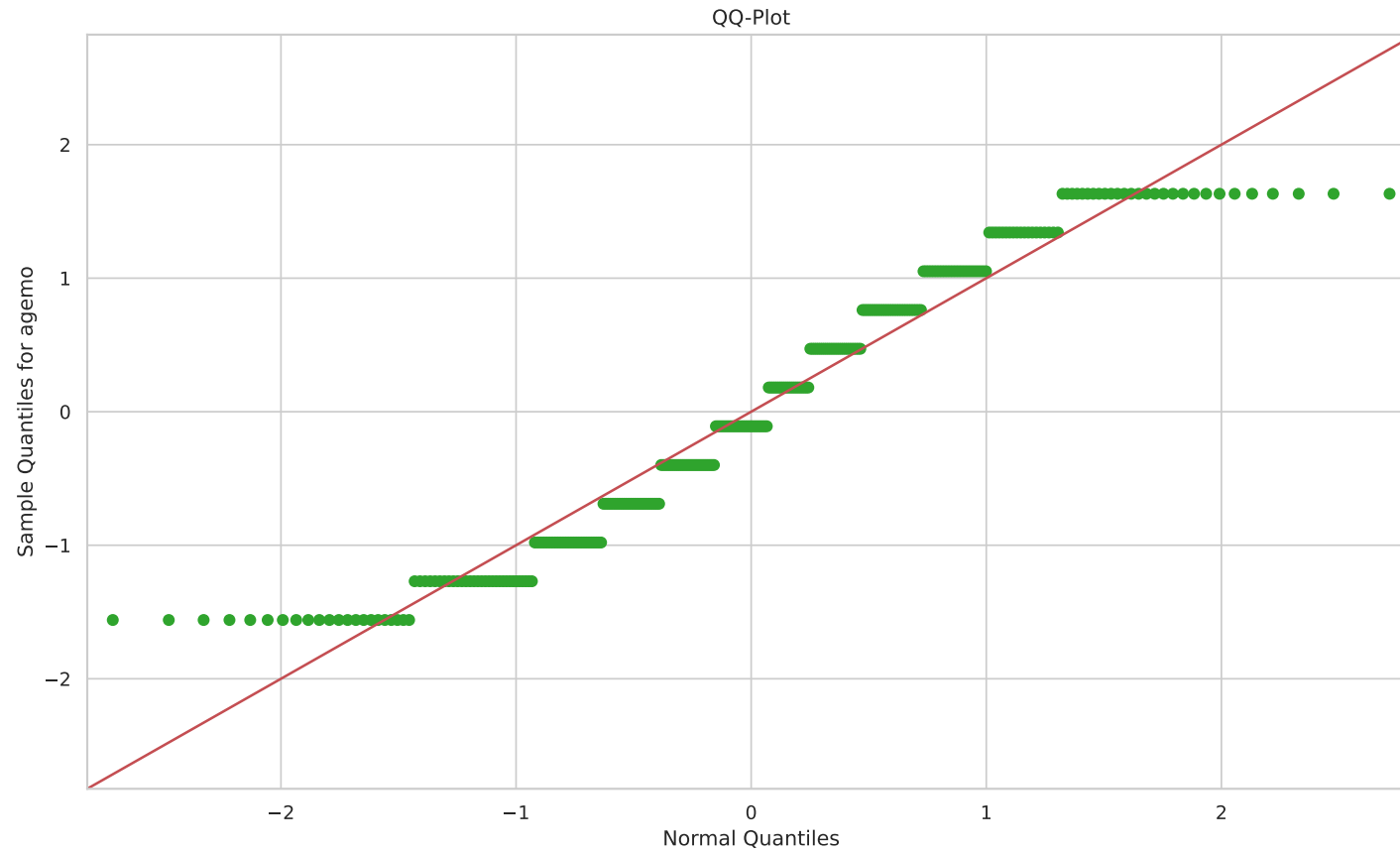
Multiple ECDF Plots of variables in one figure. Variables are sorted alphabetically.

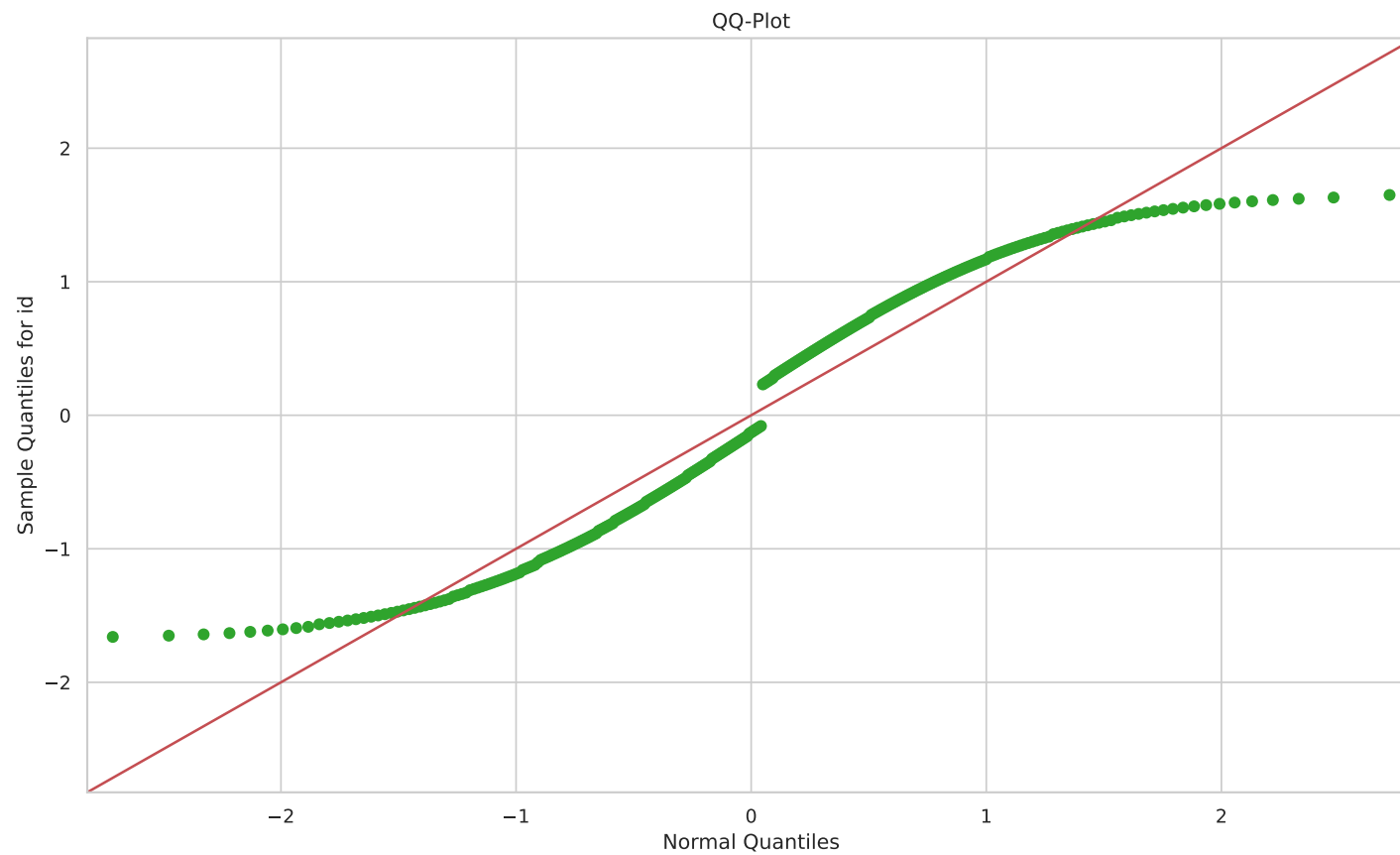


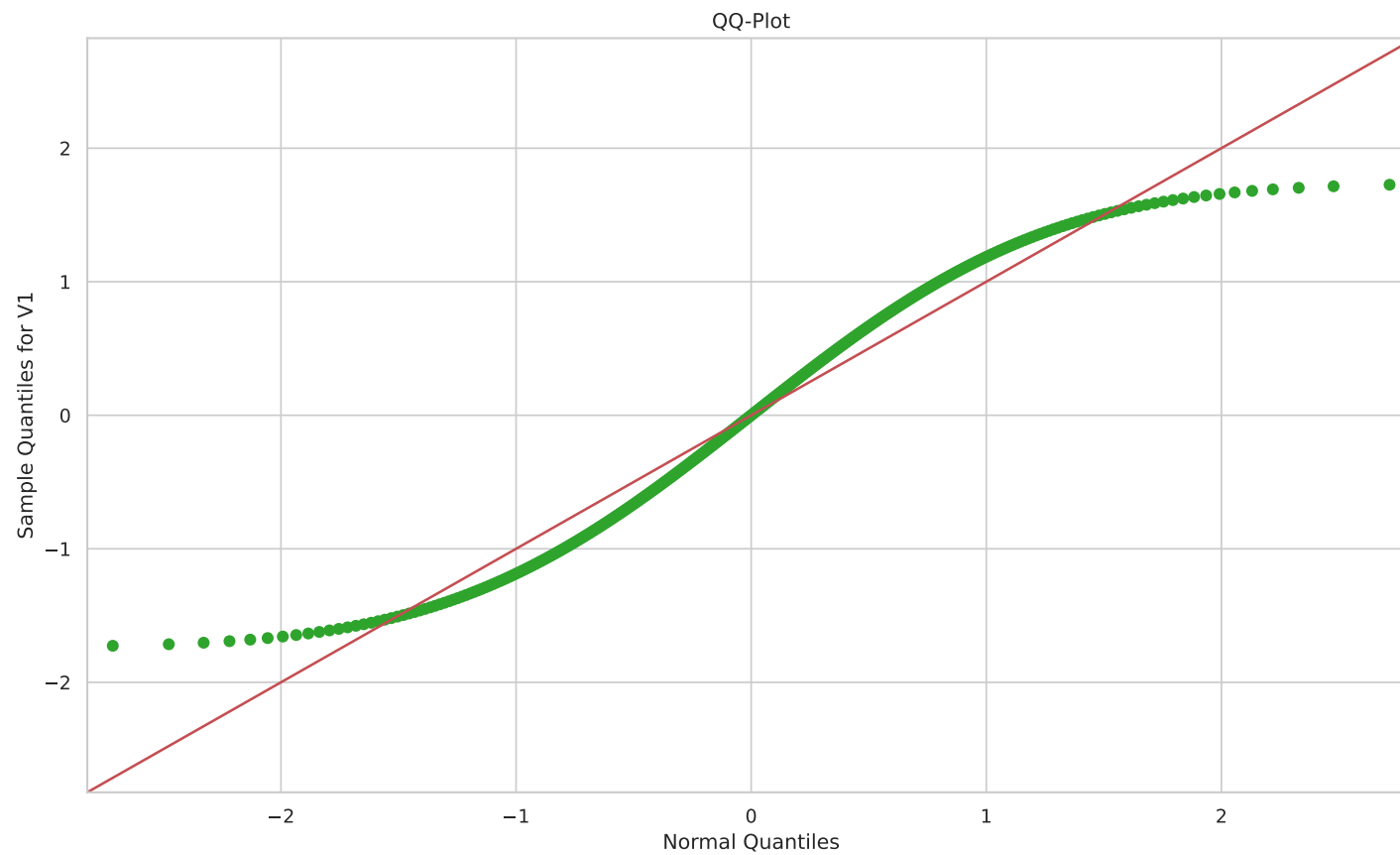
## QQ-Plots

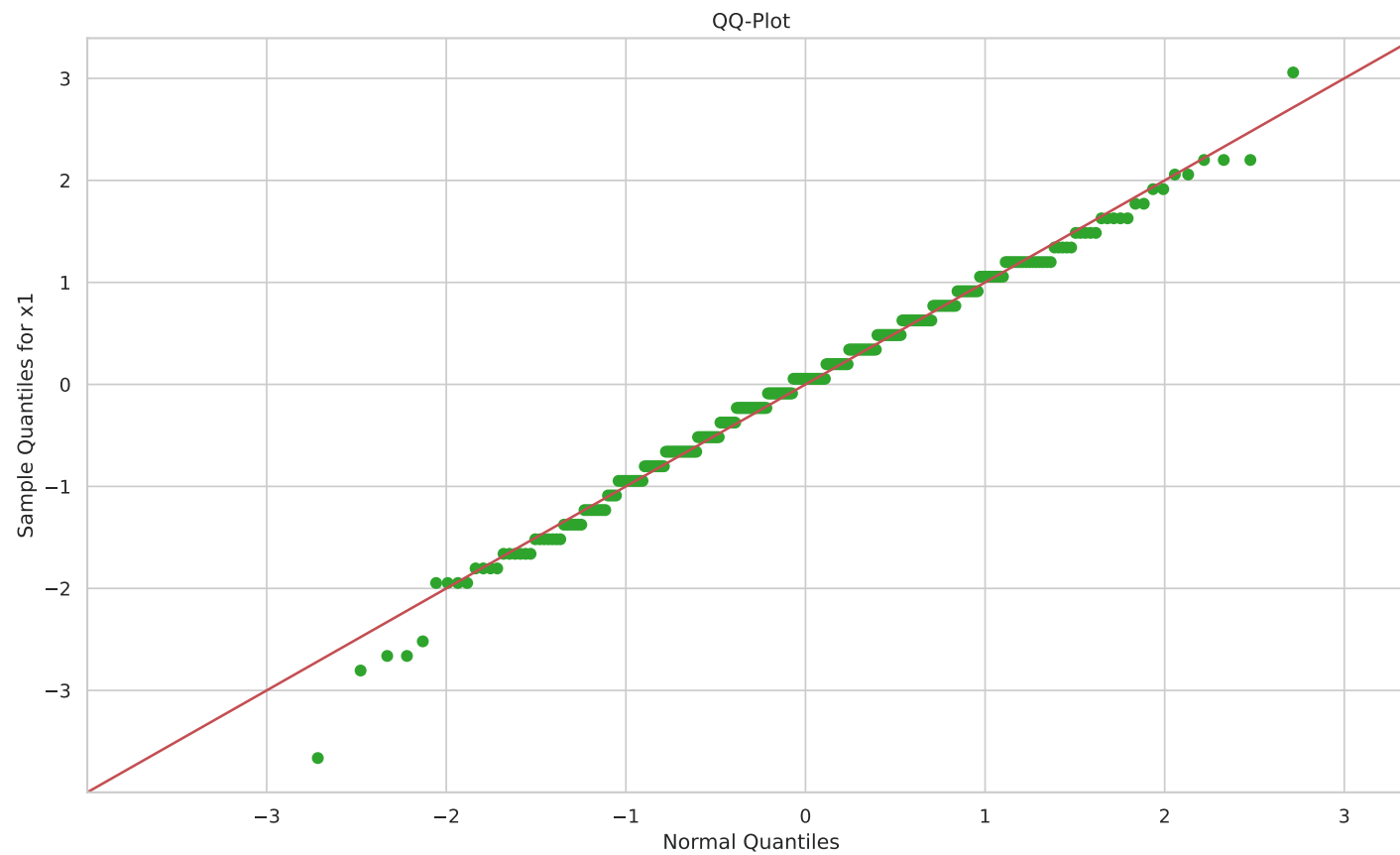
One QQ-Plot per page for each variable. Variables are sorted alphabetically.

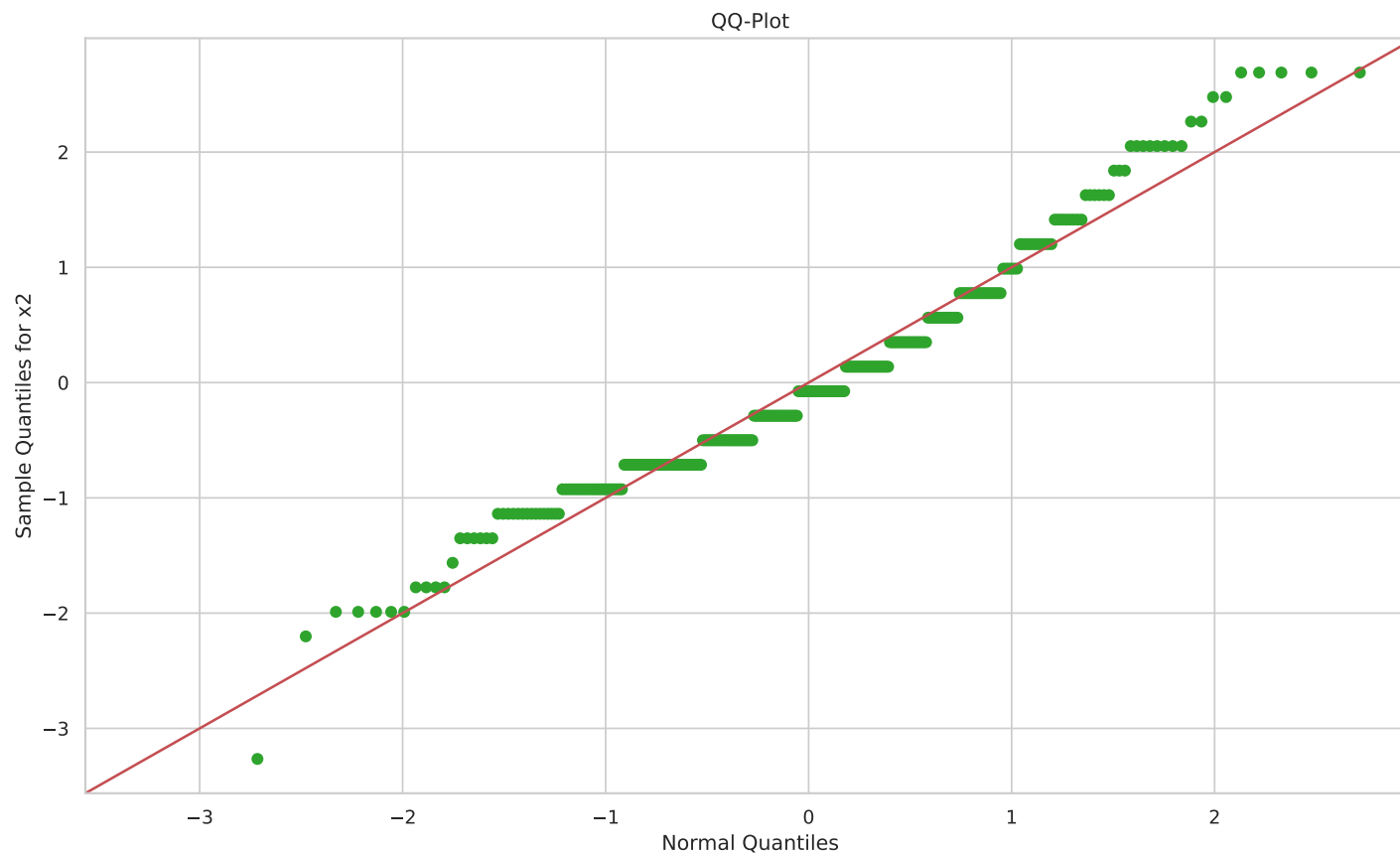
□



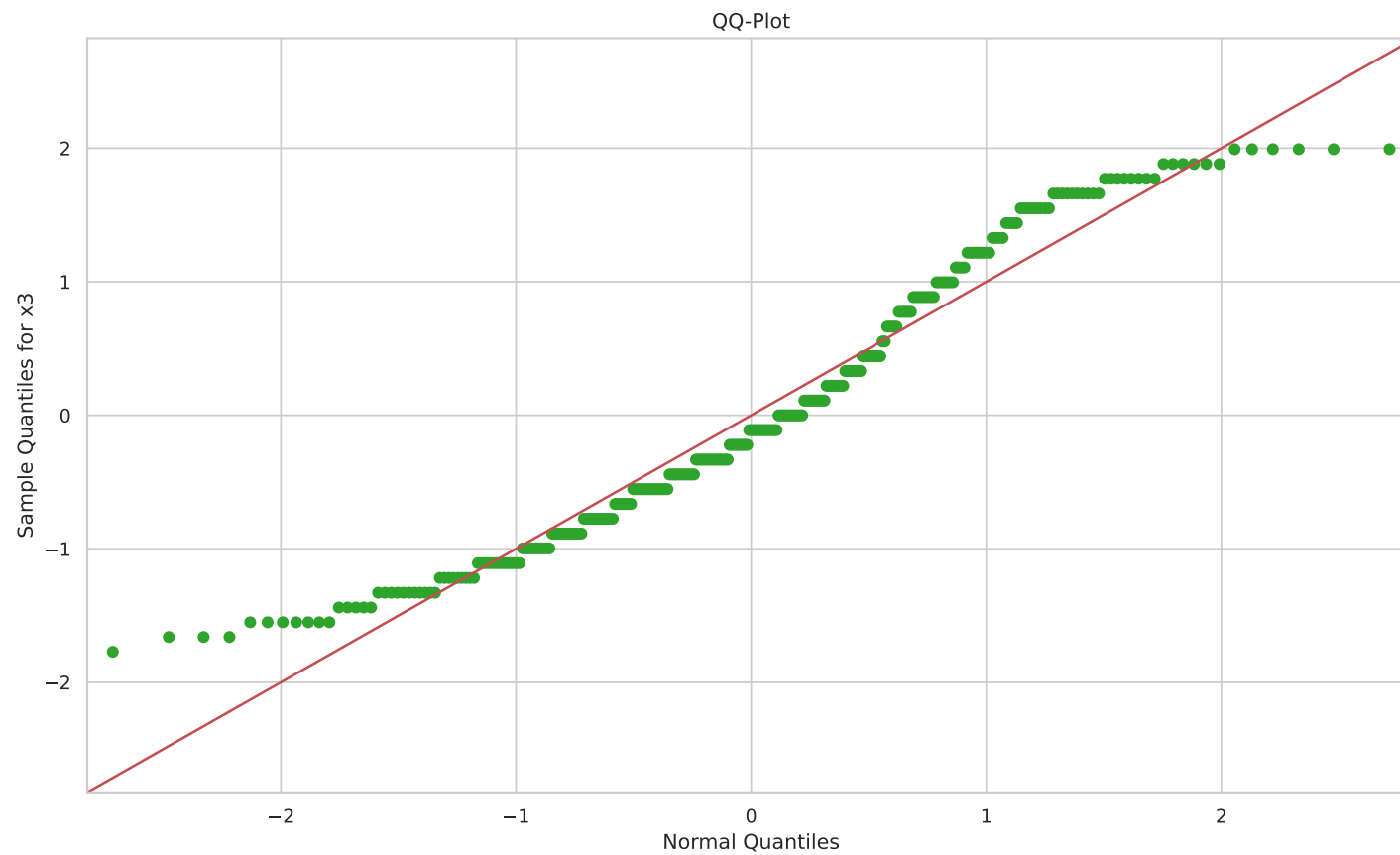


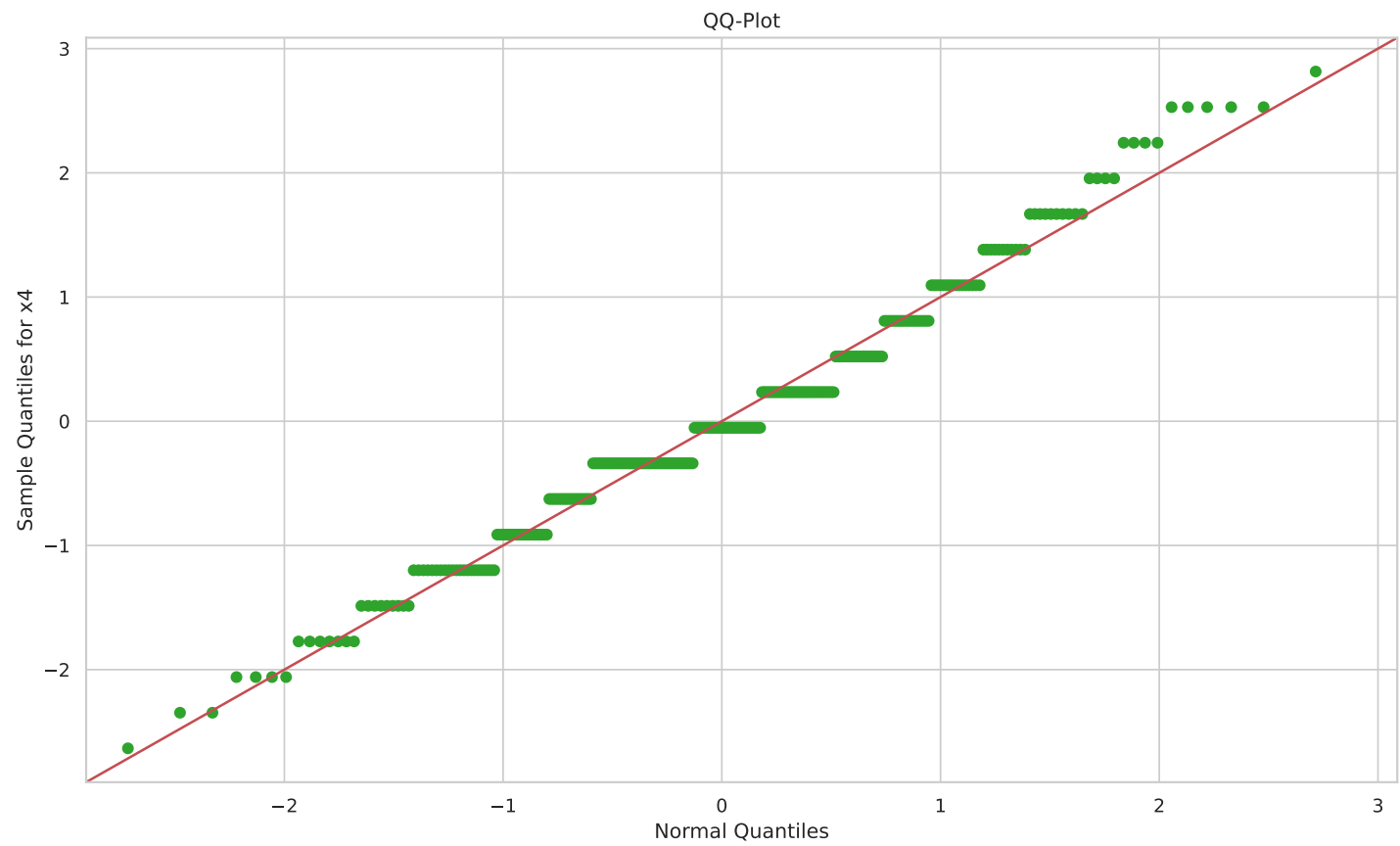


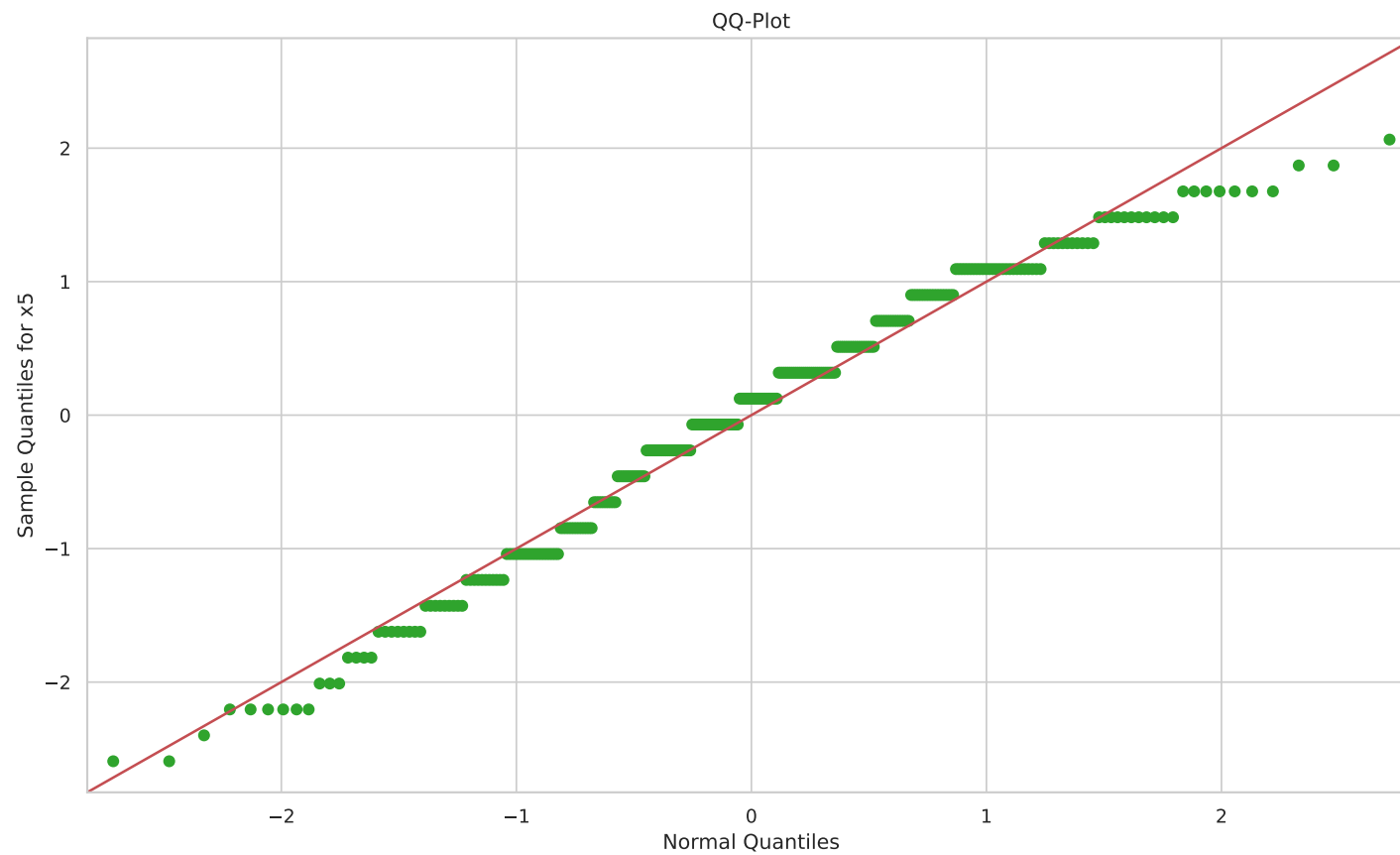


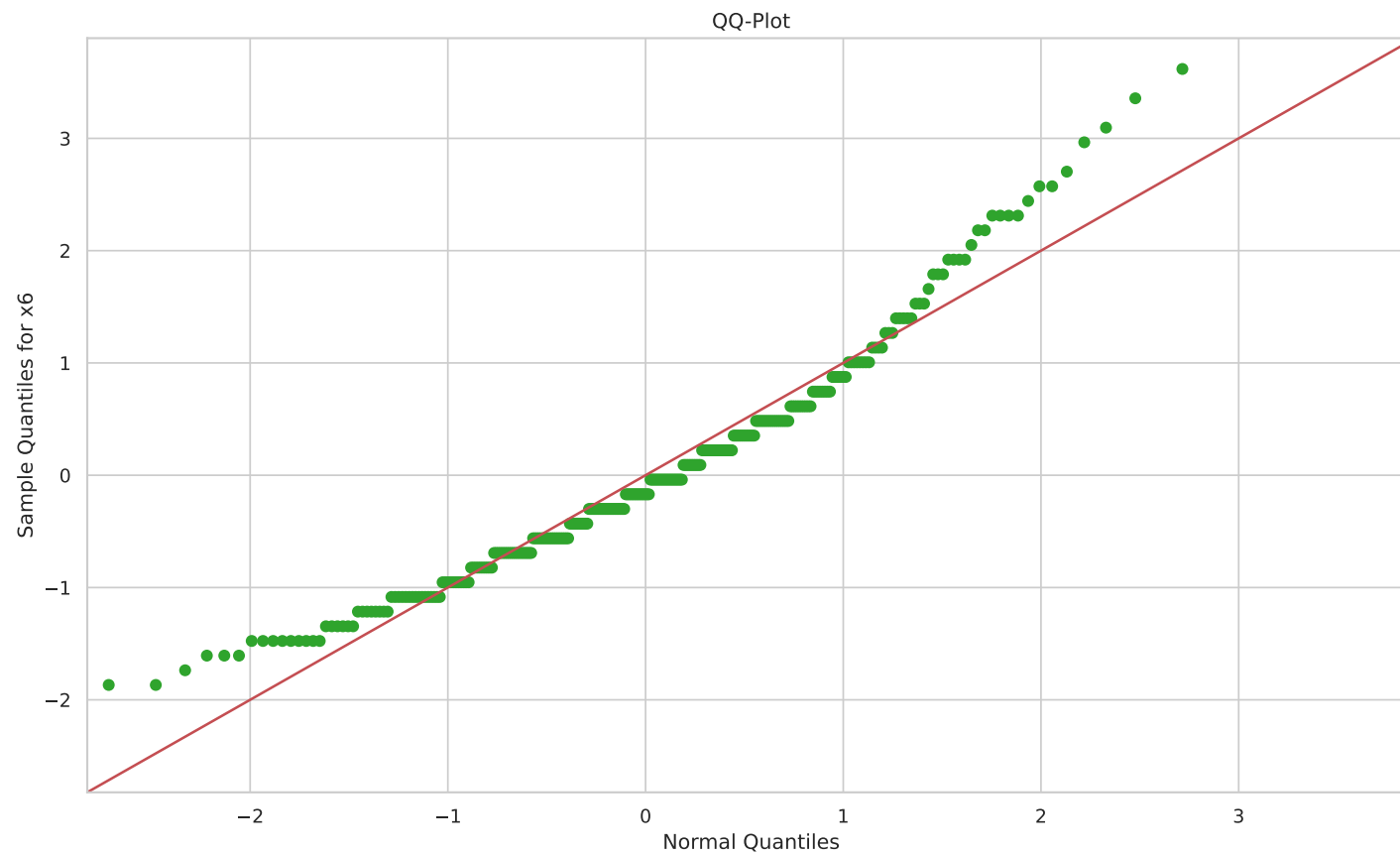


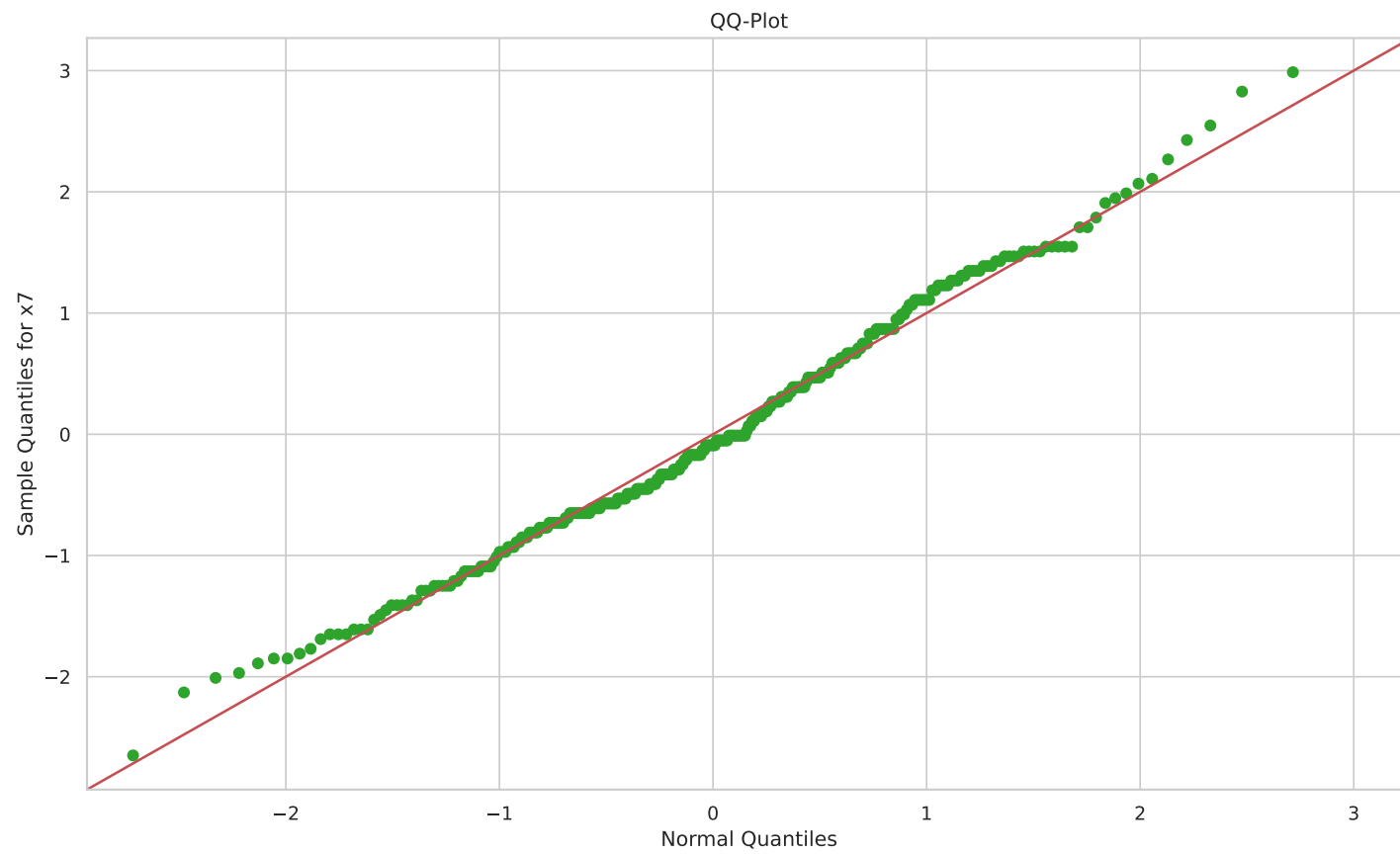


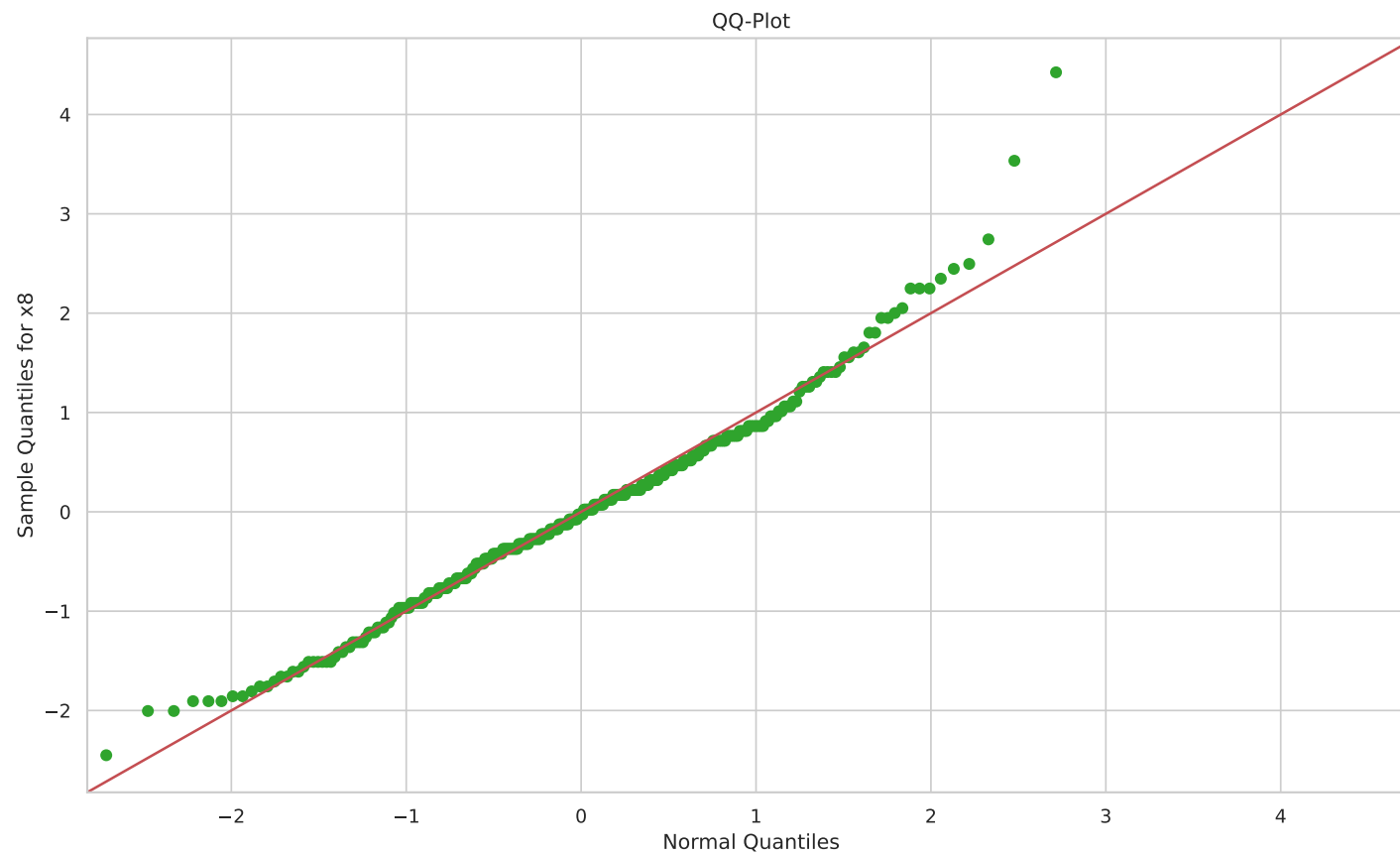


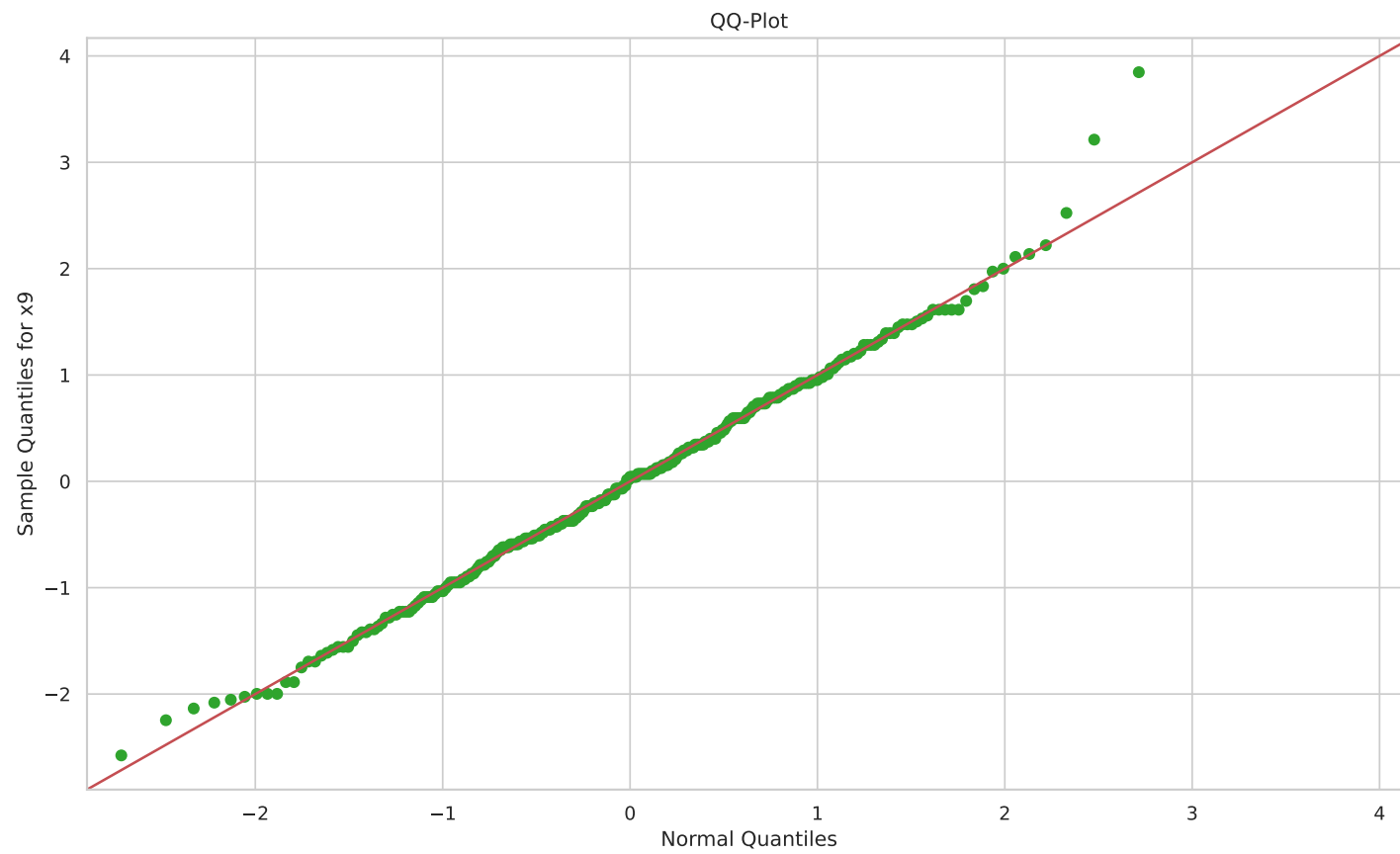






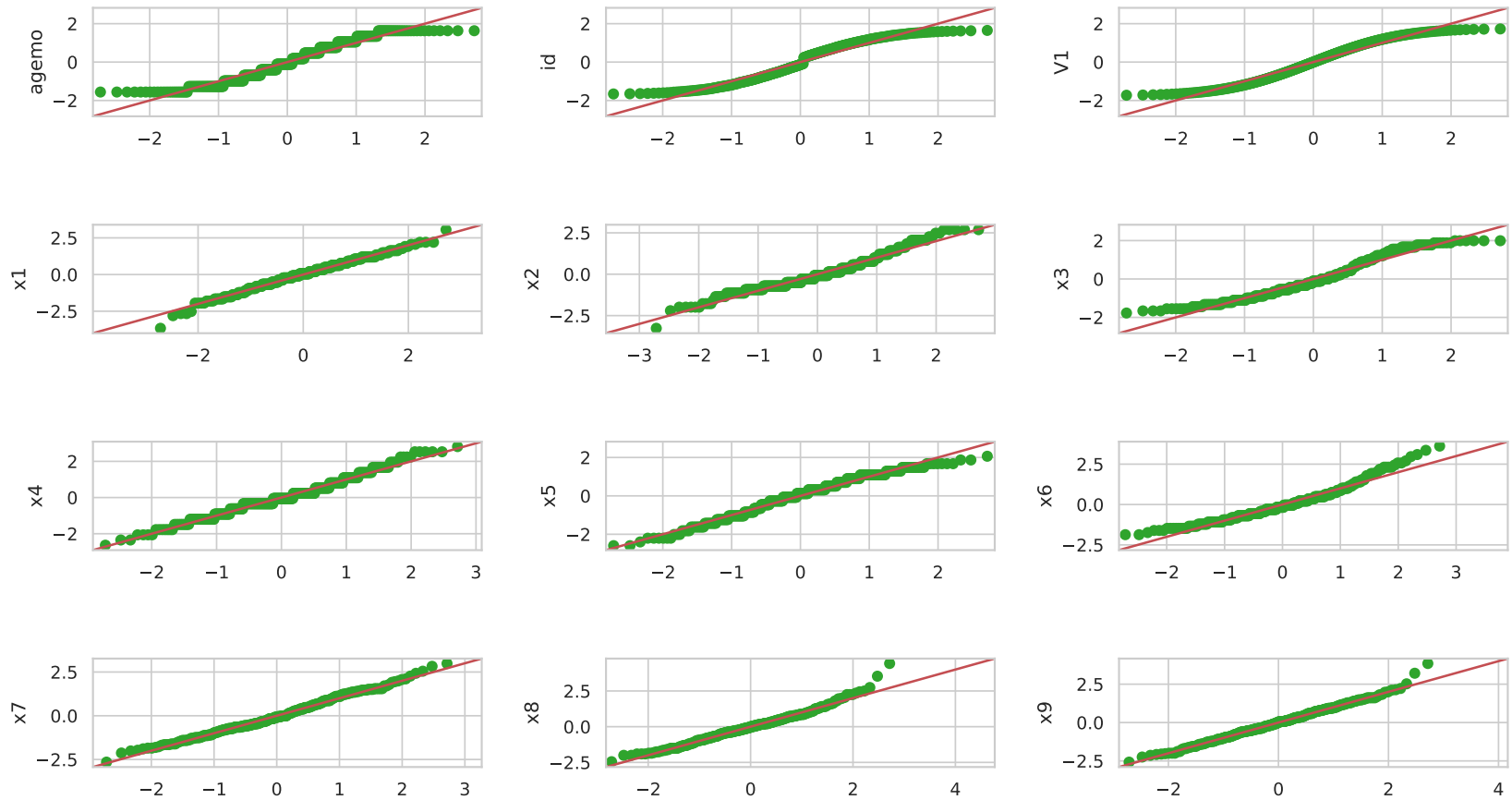






## QQ-Plots Summary

Multiple QQ-Plots of variables in one figure. Variables are sorted alphabetically.





## Results for Discrete Variables

### Descriptive Statistics

#### Totals

The table is sorted by the variable name. If any, N Unique contains the missing category.

|        | N Obs | N Missing | N Valid | % Complete | N Unique |
|--------|-------|-----------|---------|------------|----------|
| ageyr  | 301   | 0         | 301     | 100        | 6        |
| grade  | 301   | 1         | 300     | 99.67      | 3        |
| school | 301   | 0         | 301     | 100        | 2        |
| sex    | 301   | 0         | 301     | 100        | 2        |

## Frequencies

The table is sorted by the variable name. For each variable, a maximum of 20 unique values are considered, sorted in decreasing order of their frequency. If any, missings are counted as a category.

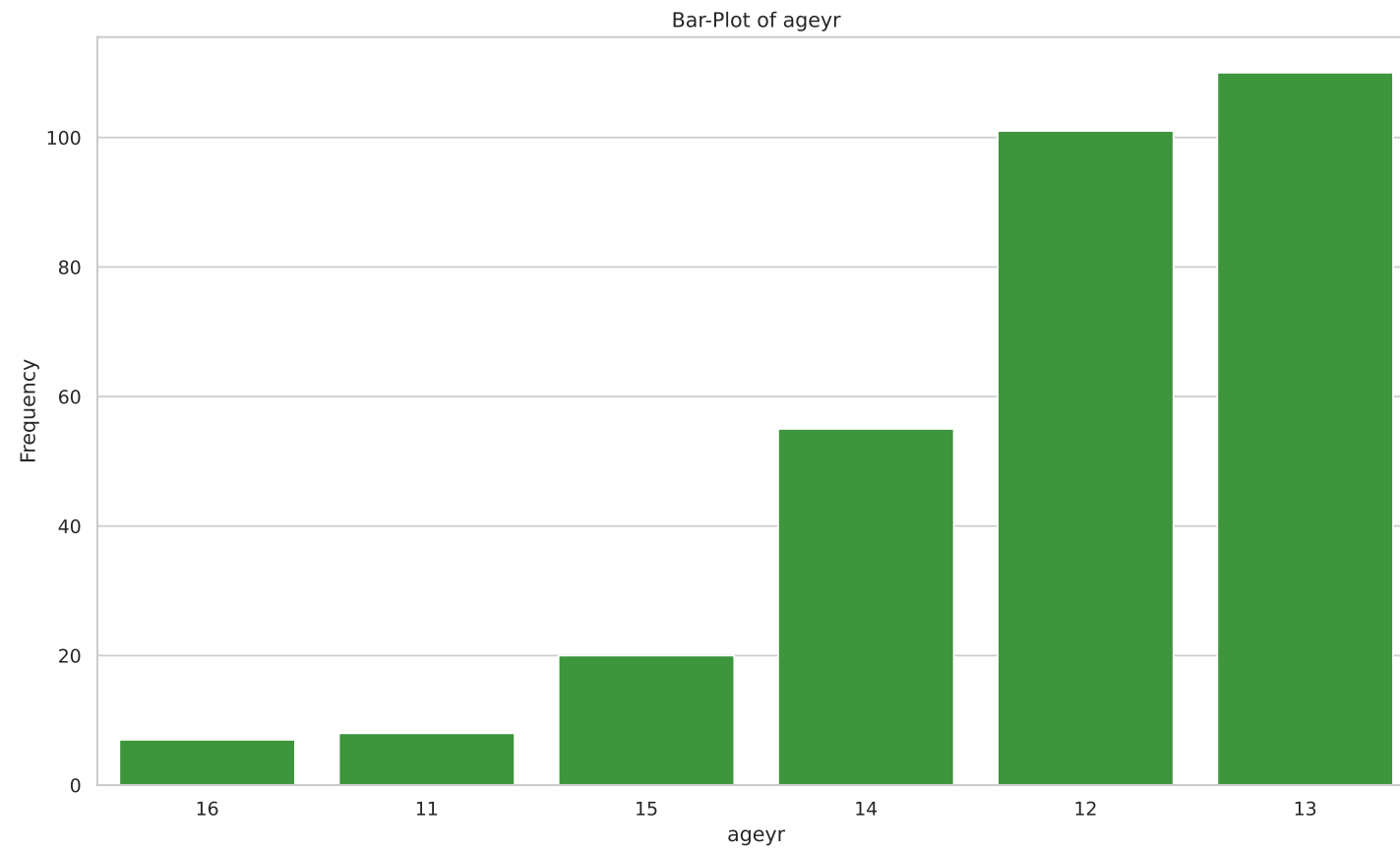
| Variable | Category    | Frequency | Percent    |
|----------|-------------|-----------|------------|
| ageyr    | 13          | 110       | 0.365449   |
| ageyr    | 12          | 101       | 0.335548   |
| ageyr    | 14          | 55        | 0.182724   |
| ageyr    | 15          | 20        | 0.0664452  |
| ageyr    | 11          | 8         | 0.0265781  |
| ageyr    | 16          | 7         | 0.0232558  |
| grade    | 7           | 157       | 0.521595   |
| grade    | 8           | 143       | 0.475083   |
| grade    | Missing     | 1         | 0.00332226 |
| school   | Pasteur     | 156       | 0.518272   |
| school   | Grant-White | 145       | 0.481728   |
| sex      | 2           | 155       | 0.51495    |
| sex      | 1           | 146       | 0.48505    |

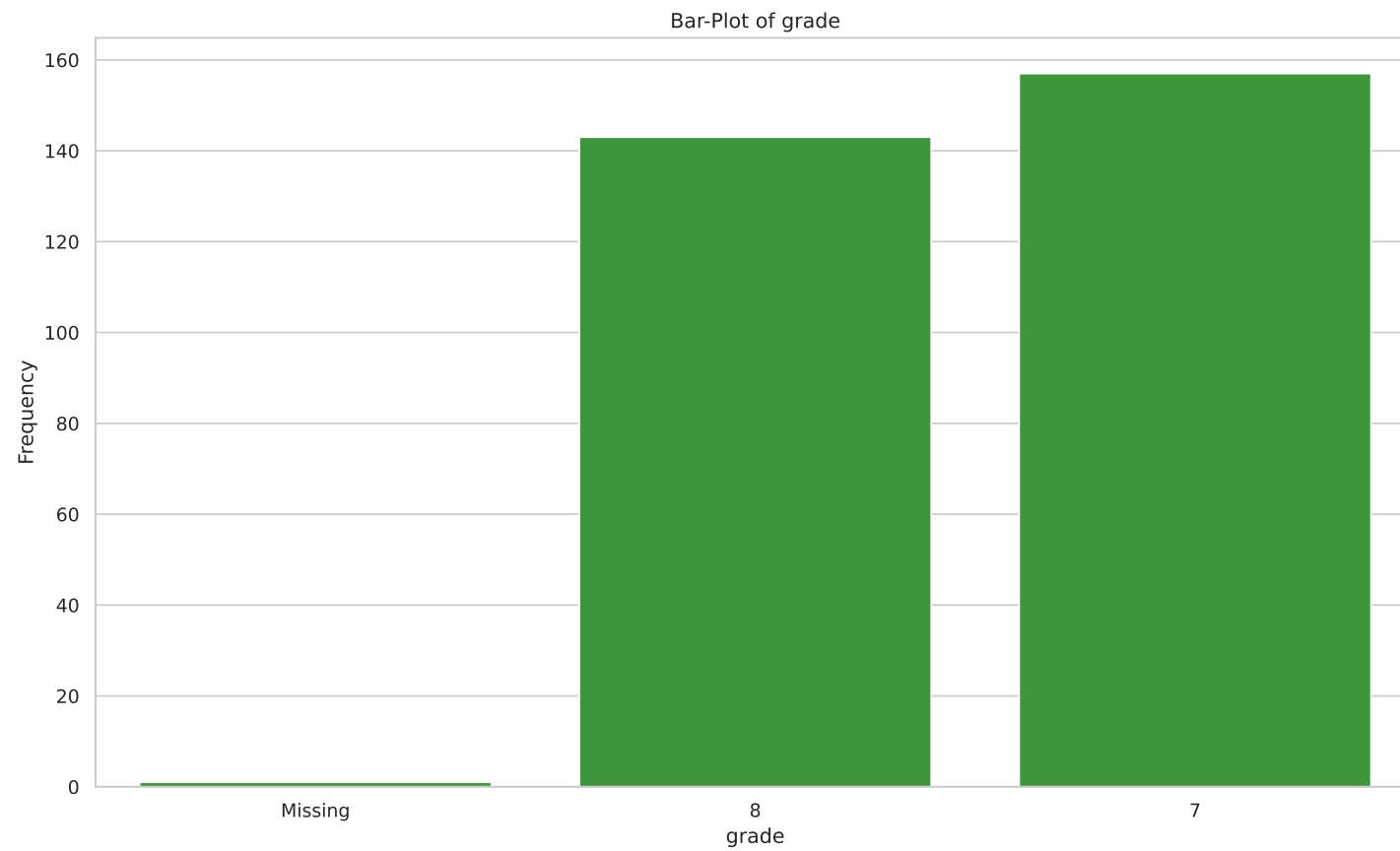
## Graphics

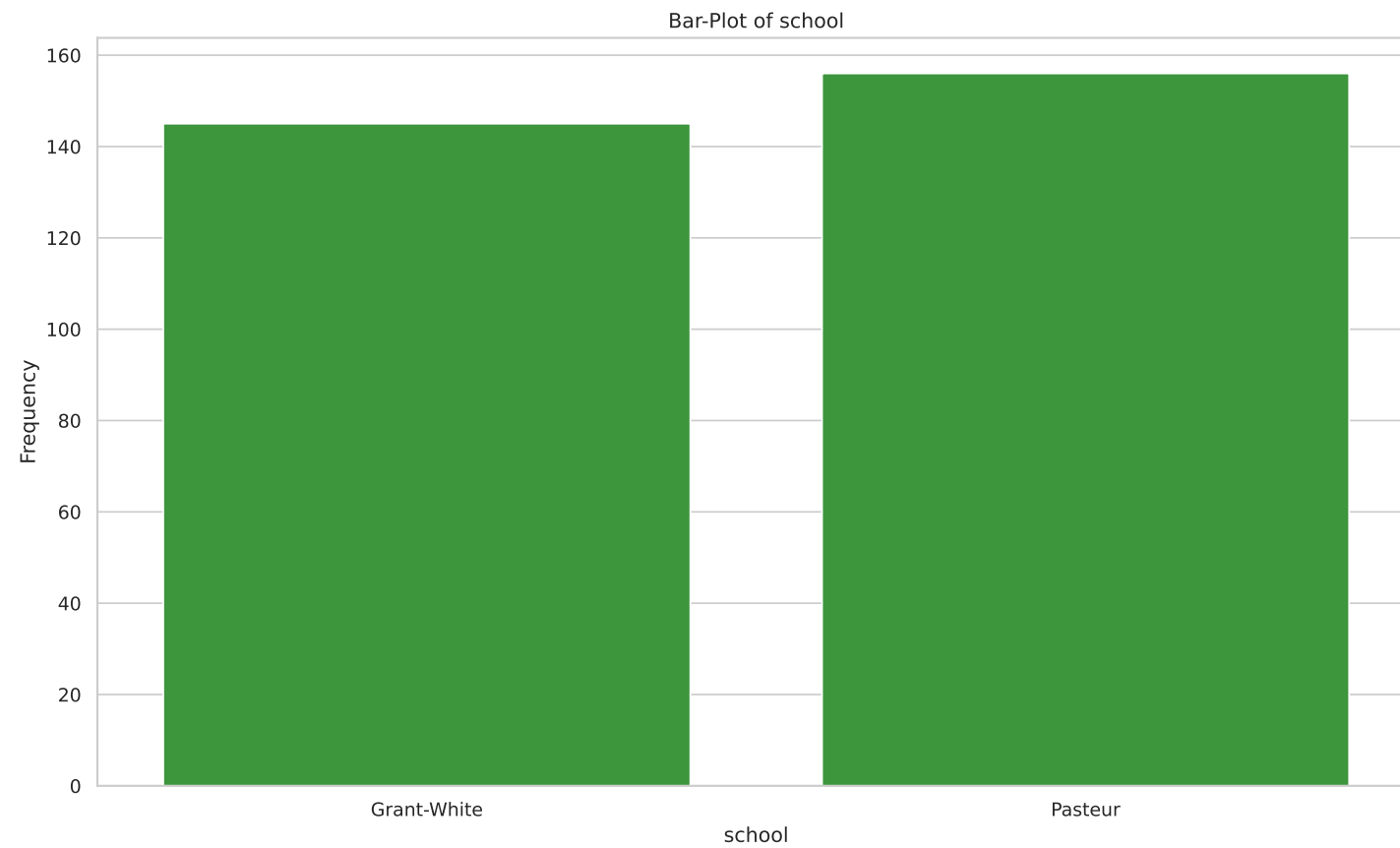
### Bar-Plots

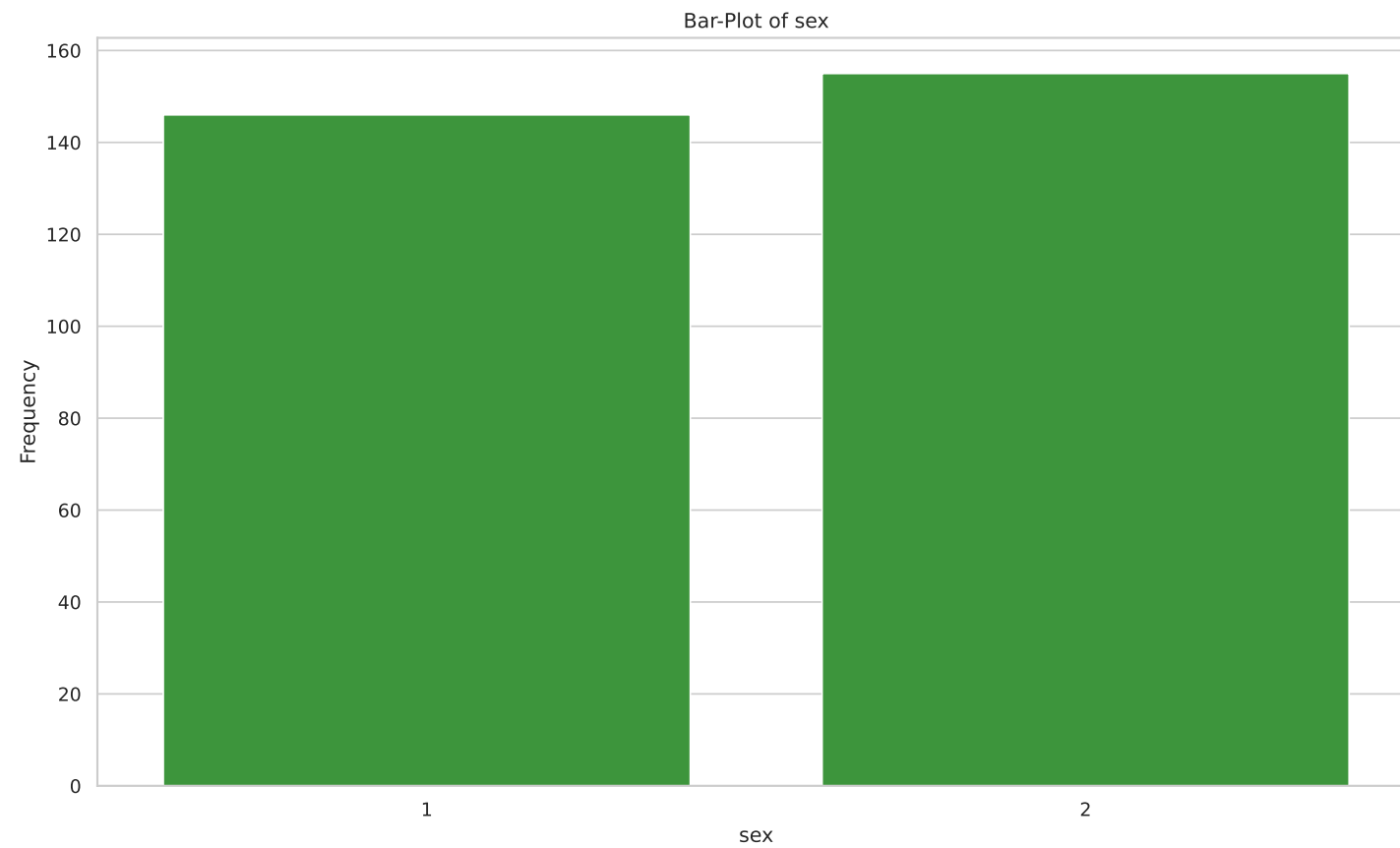
One Bar-Plot per page for each variable. Variables are sorted alphabetically. No labels for variables with more than 40 categories.

□



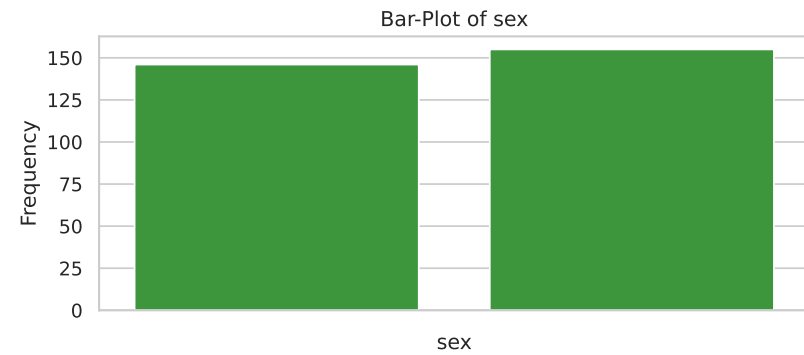
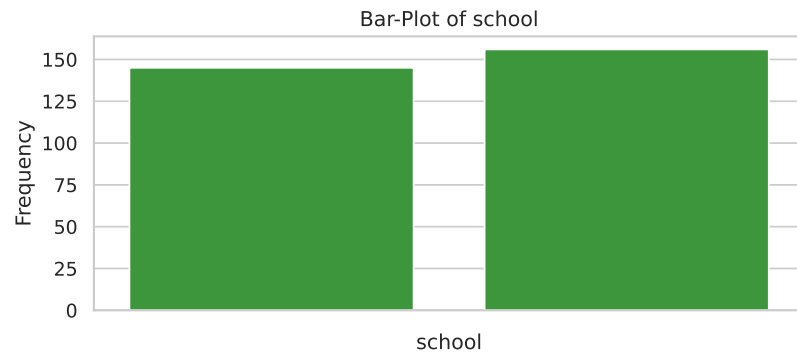
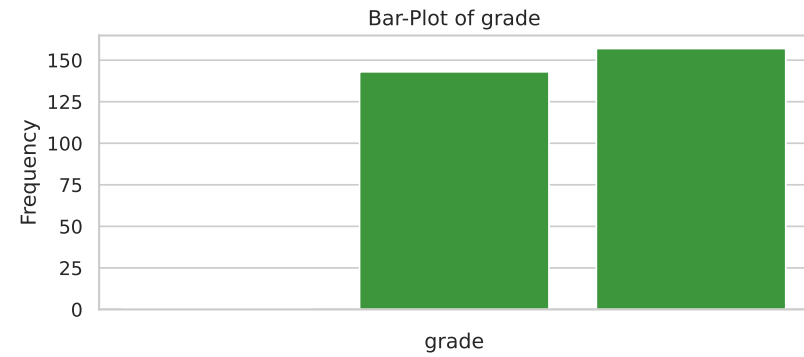
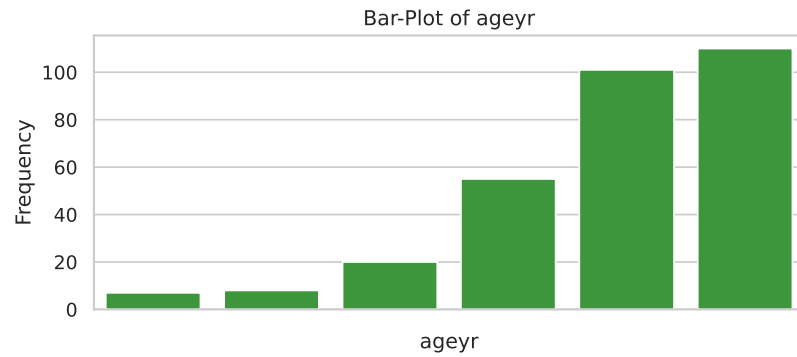






## Bar-Plots Summary

Multiple Bar-Plots of variables in one figure. Variables are sorted alphabetically. No labels displayed.

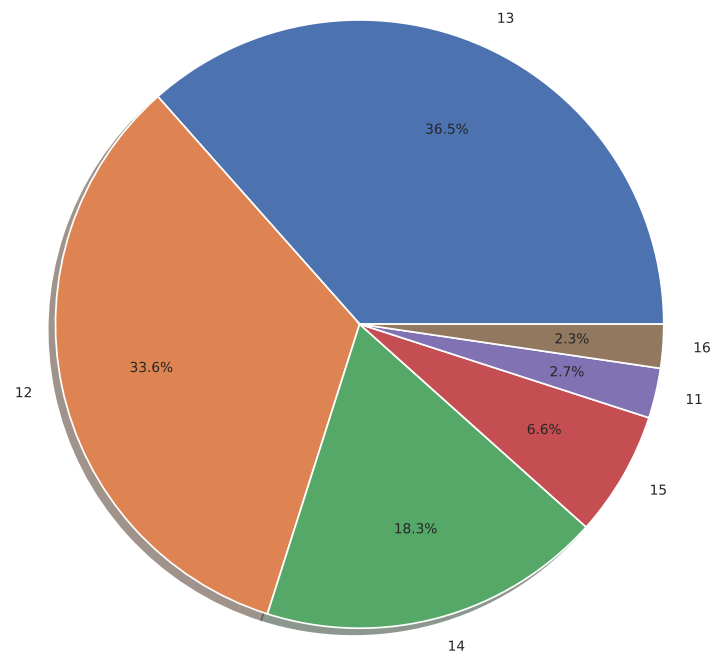


## Pie Plots

One Pie Plot per page for each variable. Variables are sorted alphabetically.

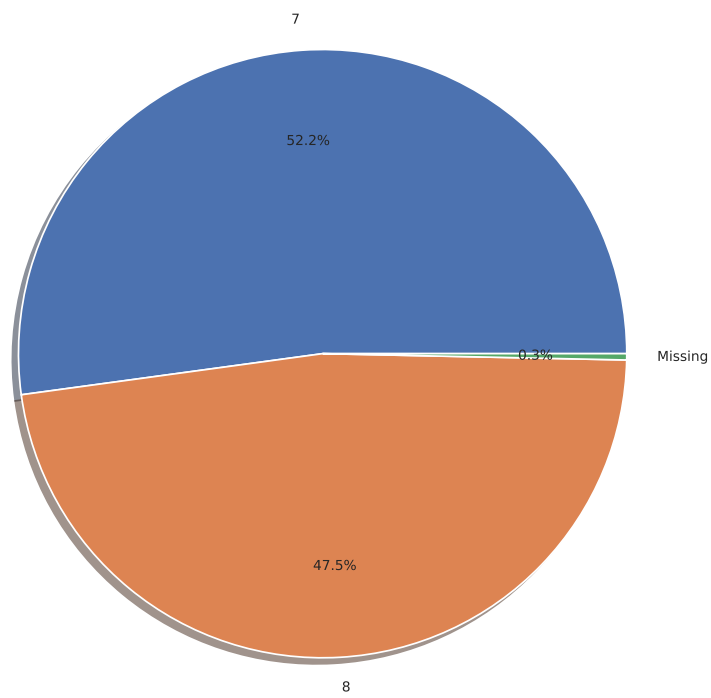
□

Pie-Plot of ageyr

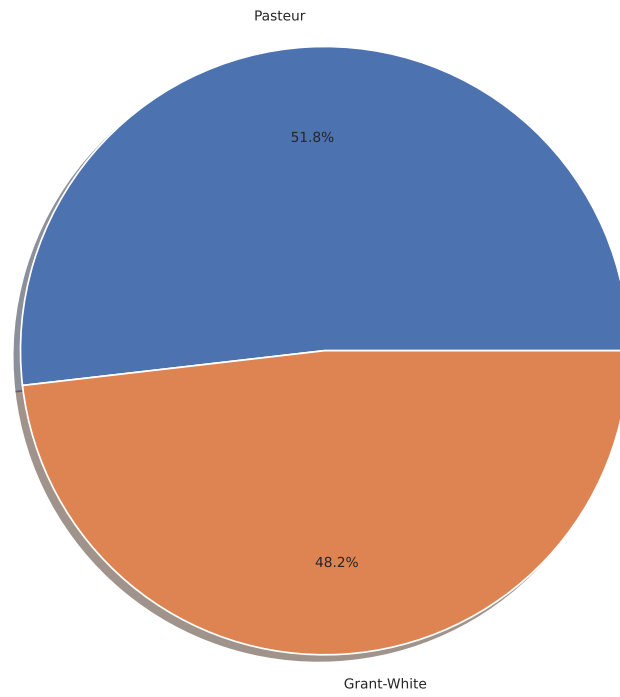




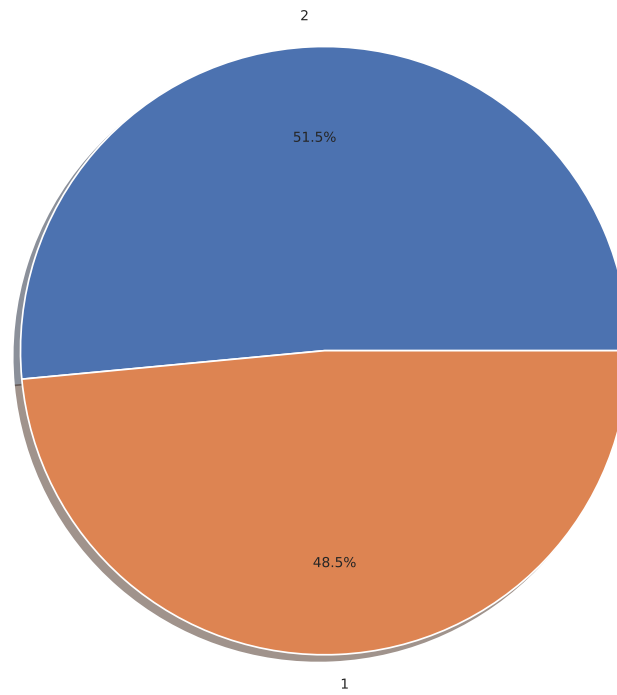
Pie-Plot of grade



Pie-Plot of school



Pie-Plot of sex



### **Pie Plots Summary**

Multiple Pie Plots of variables in one figure. Variables are sorted alphabetically.

See figures on next page.

