xploring Anscombe’s Quartet: A Lesson in Exploratory Data Analy

Author(s): Andy Wu

Date: November 09, 2025

A detailed analysis and visualization of Anscombe's Quartet demonstrating the importance of graphical analysis in statistics.

Abstract / Executive summary

Anscombe's Quartet consists of four datasets with nearly identical summary statistics (means, variances, correlations, regression lines), yet markedly different distributions and patterns when plotted. This report demonstrates exploratory data analysis (EDA) on the quartet, emphasizing why visualization is essential. The report includes statistical formulas, summary tables, visual diagnostics, interpretation, and code for reproducibility.

Introduction

Anscombe's Quartet was constructed by Francis Anscombe in 1973 to highlight the importance of graphing data before analyzing. The quartet contains four datasets that have similar summary statistics but different underlying distributions — showing that numerical summaries alone can be misleading. This analysis performs both numeric and visual EDA on each dataset.

Data

Data: The report uses the canonical Anscombe's Quartet datasets. Data were loaded into a pandas DataFrame within this notebook and are shown below.

Data preview (first 6 rows)

x1	y1	x2	y2	x3	y3	x4	y4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.1	14.0	8.84	8.0	7.04

Methods

Statistics computed: mean (x,y), sample variance, standard deviation, covariance, Pearson correlation (r),

linear regression coefficients (intercept and slope), and coefficient of determination (R^2). Visualizations:

scatter plots with regression lines, residual plots, overlaid dataset comparison, violin and box plots for distributions.

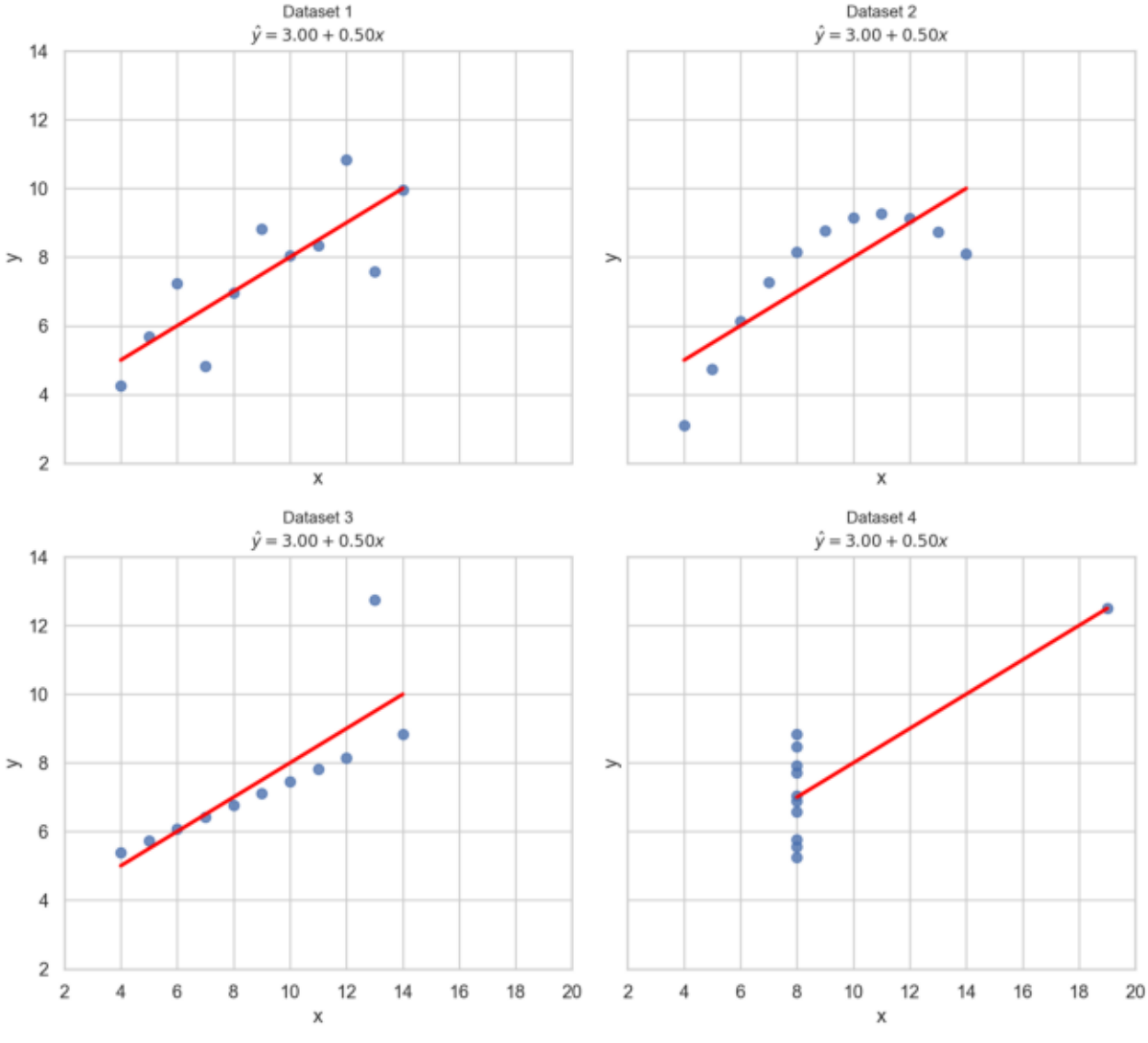
Summary statistics (rounded)

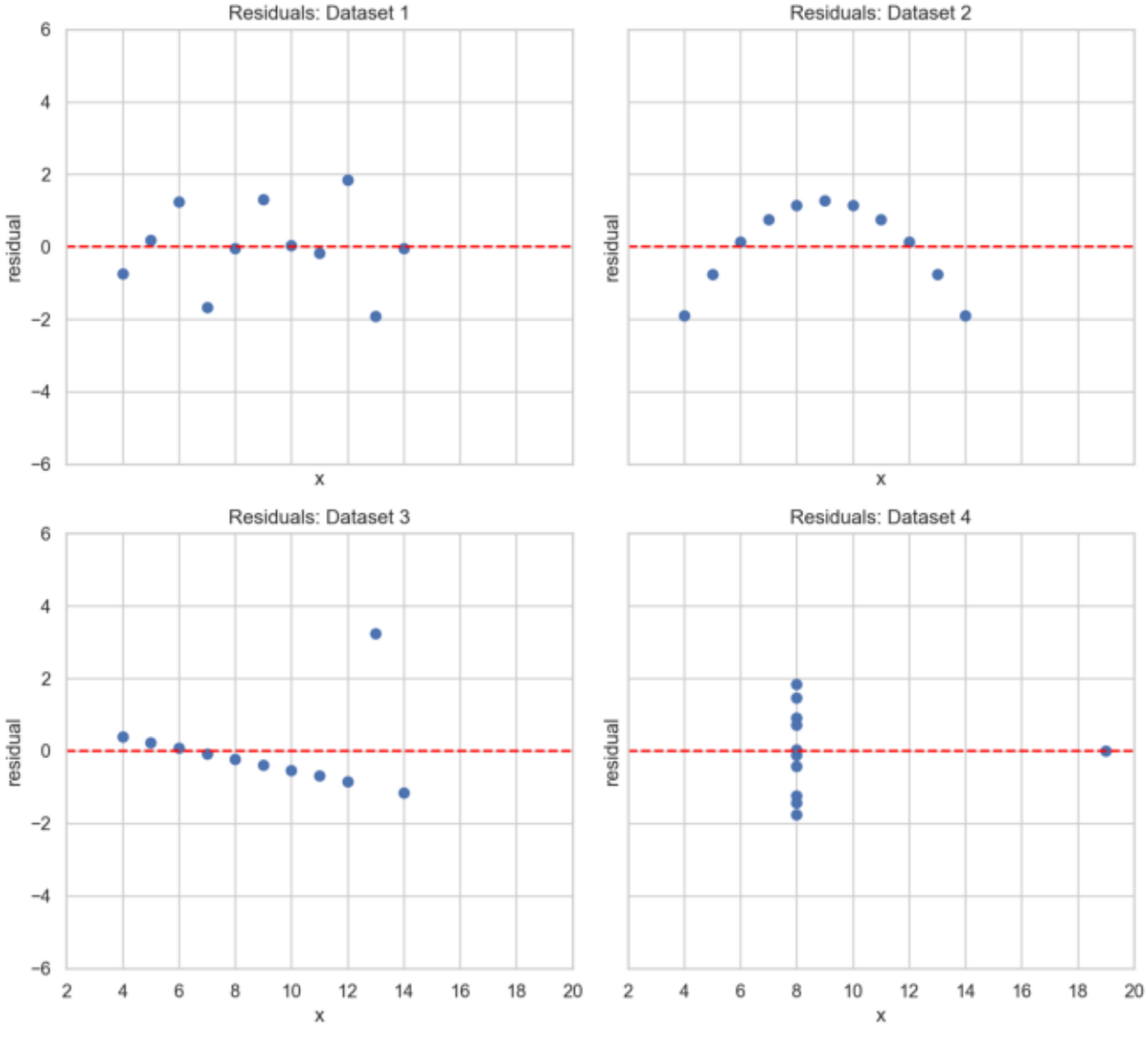
Summary statistics for Anscombe's Quartet (rounded)

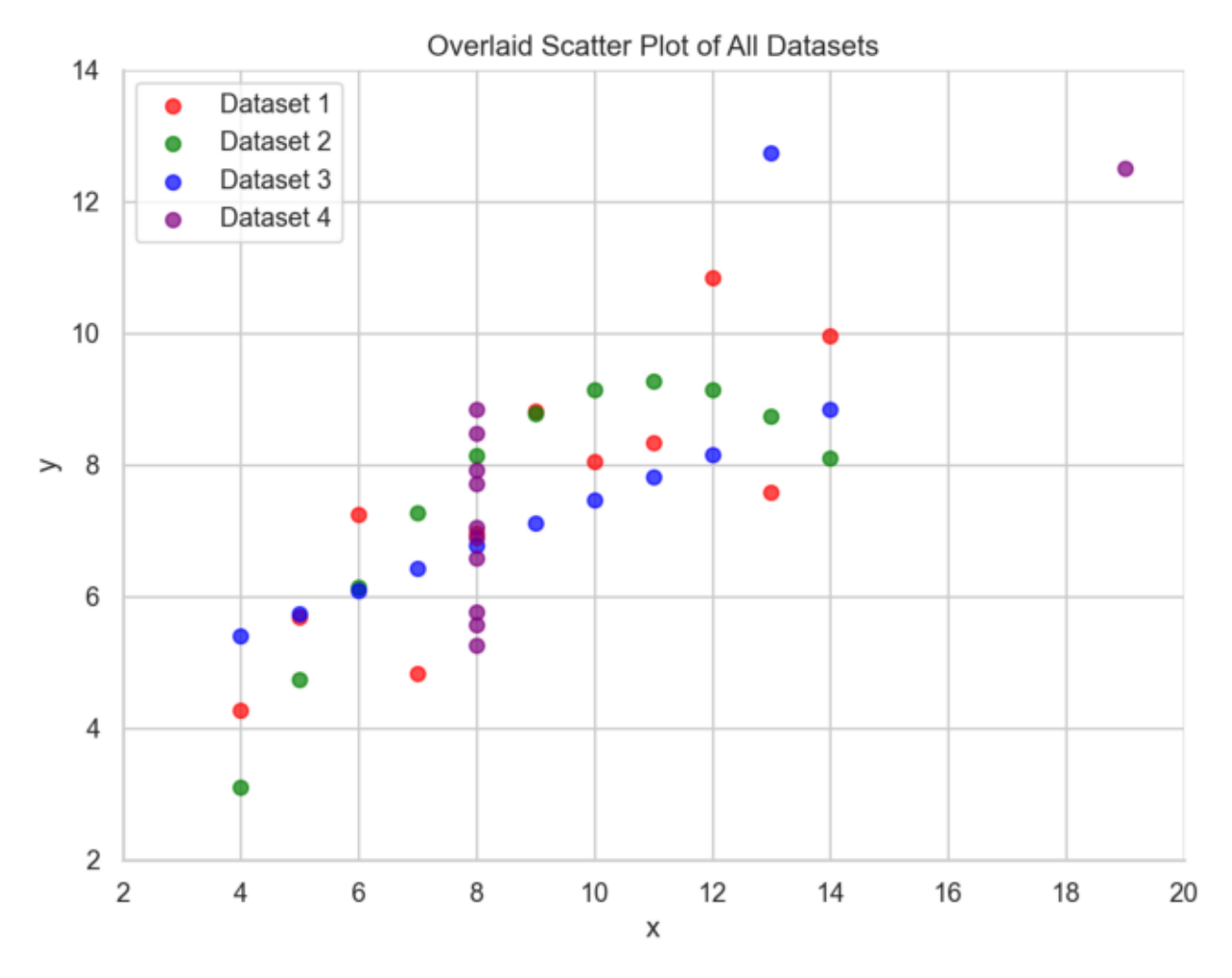
Dataset	n	mean_x	mean_y	var_x	var_y	sd_x	sd_y	cov_xy	r	slope	intercept	r2
Dataset 1	11	9.0	7.500909	11.0	4.127269	3.316625	2.031568	5.501	0.816421	0.500091	3.000091	0.666542
Dataset 2	11	9.0	7.500909	11.0	4.127629	3.316625	2.031657	5.5	0.816237	0.5	3.000909	0.666242
Dataset 3	11	9.0	7.5	11.0	4.12262	3.316625	2.030424	5.497	0.816287	0.499727	3.002455	0.666324
Dataset 4	11	9.0	7.500909	11.0	4.123249	3.316625	2.030579	5.499	0.816521	0.499909	3.001727	0.666707

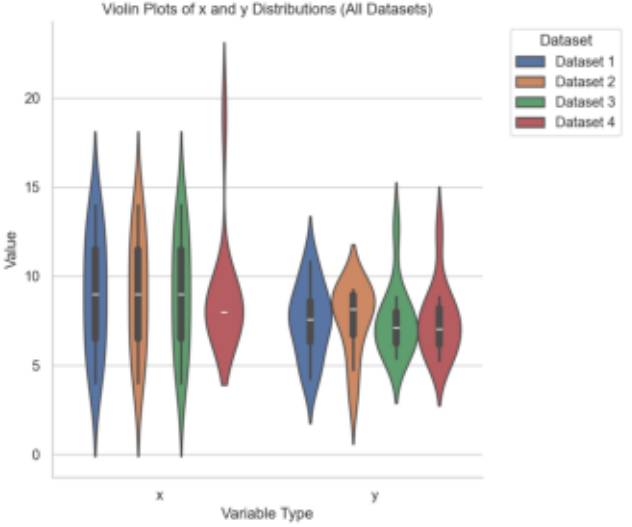
Visualizations

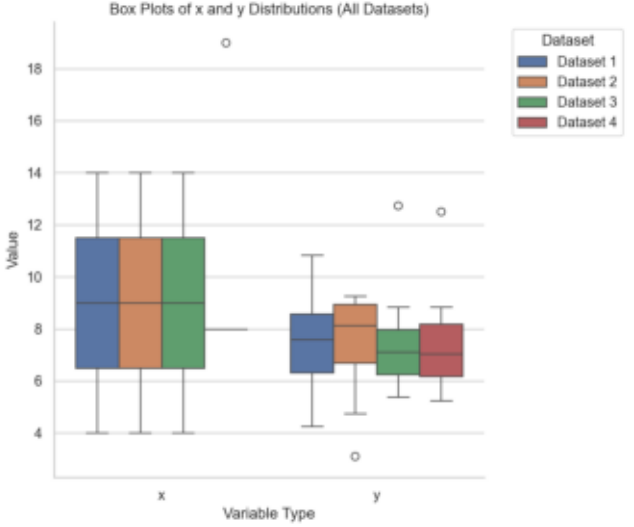
Below are the key visualizations generated during the EDA.


Dataset 1
$\hat{y} = 3.00 + 0.50x$
Dataset 2
$\hat{y} = 3.00 + 0.50x$
Dataset 3
$\hat{y} = 3.00 + 0.50x$
Dataset 4
$\hat{y} = 3.00 + 0.50x$
x
y


Residuals: Dataset 1
Residuals: Dataset 2
Residuals: Dataset 3
Residuals: Dataset 4
residual
x
6
4
2
0
−2
−4
−6
2
4
6
8
10
12
14
16
18
20


Overlaid Scatter Plot of All Datasets
Dataset 1
Dataset 2
Dataset 3
Dataset 4
y
x
14
12
10
8
6
4
2
2
4
6
8
10
12
14
16
18
20


Violin Plots of x and y Distributions (All Datasets)
Dataset
Dataset 1
Dataset 2
Dataset 3
Dataset 4
Value
0
5
10
15
20
x
y
Variable Type


Box Plots of x and y Distributions (All Datasets)
Dataset
Dataset 1
Dataset 2
Dataset 3
Dataset 4
Value
4
6
8
10
12
14
16
18
x
y
Variable Type

Interactive version

Interactive notebook or dashboard can be linked here:
https://andywu784037.github.io/andywu784_Interactive_Graph/

Interpretation

Interpretation: Although the four datasets have nearly identical summary statistics, their scatterplots reveal very different structures: Dataset 1 looks roughly linear, Dataset 2 has a small curve/outlier effect, Dataset 3 contains a high-leverage point, and Dataset 4 contains an outlier that strongly influences the regression.

Therefore, summary statistics alone are insufficient — visualization is required to reveal pattern, leverage, non-linearity, and outliers.

Reproducibility & Code

Reproducibility: Code to reproduce this analysis is saved alongside the notebook. GitHub link : https://github.com/andywu784037/andywu784_Interactive_Graph/tree/main#readme

Collaboration notes

Analysis performed by Andy Wu. Template collaboration notes included; replace with actual PR links and contributor roles if applicable.

Conclusion & next steps

Conclusion: Anscombe's Quartet is a classic demonstration that identical numerical summaries can hide very

different underlying distributions. Next steps: consider robust regression, influence diagnostics (Cook's distance), and interactive dashboards to explore leverage and outliers.

Appendix: Code & Extra Figures

The following pages include the core notebook code and extra figure images.

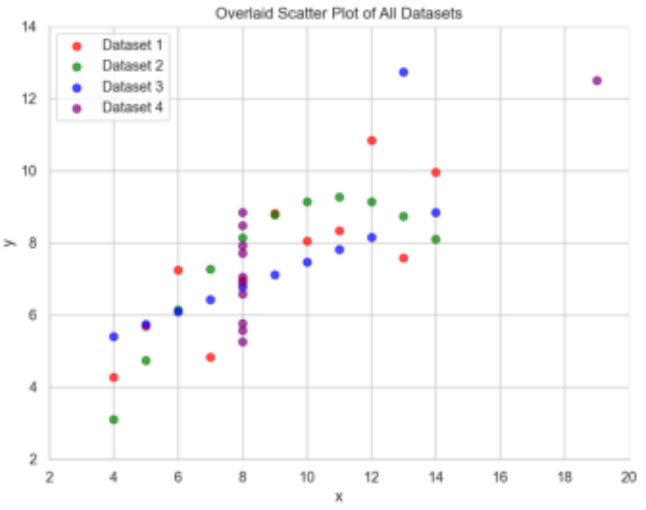
Appendix: Code Location

The full notebook and exported script were saved: C:\Users\wh503\anscombe_report_code.py. Use the notebook itself for full source.

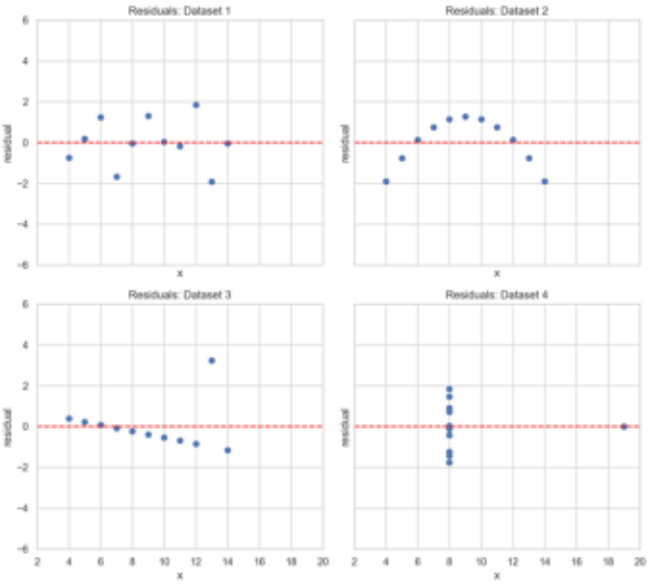
Appendix: Extra Figures

Figures used in the analysis were saved to the 'report_figs' folder. Thumbnails shown below.

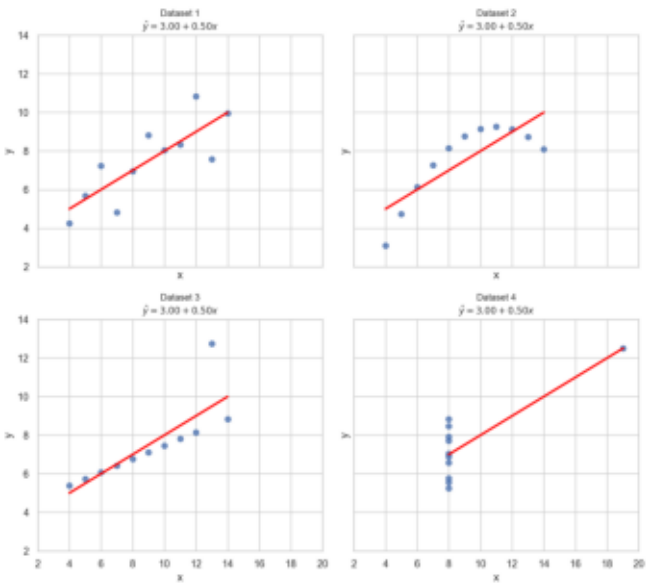
overlaid.png



residuals_4panel.png



scatter_4panel.png



summary_table.png

Summary statistics for Anscombe's Quartet (rounded)

Dataset	n	mean_x	mean_y	var_x	var_y	sd_x	sd_y	cov_xy	r	slope	intercept	r2
Dataset 1	11	9.0	7.500909	11.0	4.127269	3.316625	2.031568	5.501	0.816421	0.500091	3.000091	0.666542
Dataset 2	11	9.0	7.500909	11.0	4.127629	3.316625	2.031657	5.5	0.816237	0.5	3.000909	0.666242
Dataset 3	11	9.0	7.5	11.0	4.12262	3.316625	2.030424	5.497	0.816287	0.499727	3.002455	0.666324
Dataset 4	11	9.0	7.500909	11.0	4.123249	3.316625	2.030579	5.499	0.816521	0.499909	3.001727	0.666707

violin_box.png