Statystyka opisowa

Martyna Śpiewak

Bootcamp Data Science

Statystyka

Statystyka jest zbiorem metod służących

- gromadzeniu,
- prezentacji,
- analizie,
- interpretacji

danych, w celu podejmowania decyzji.

Termin statystyka obejmuje:

- statystykę teoretyczną;
- statystykę stosowaną.

Statystyka opisowa — gałąź statystyki stosowanej, zajmująca się

- wstępną analizą danych, oraz
- wnioskowaniem statystycznym, czyli metodologią wyciągania wniosków na temat pewnych właściwości badanej zbiorowości na podstawie dostępnych danych.

Podstawowe pojęcia

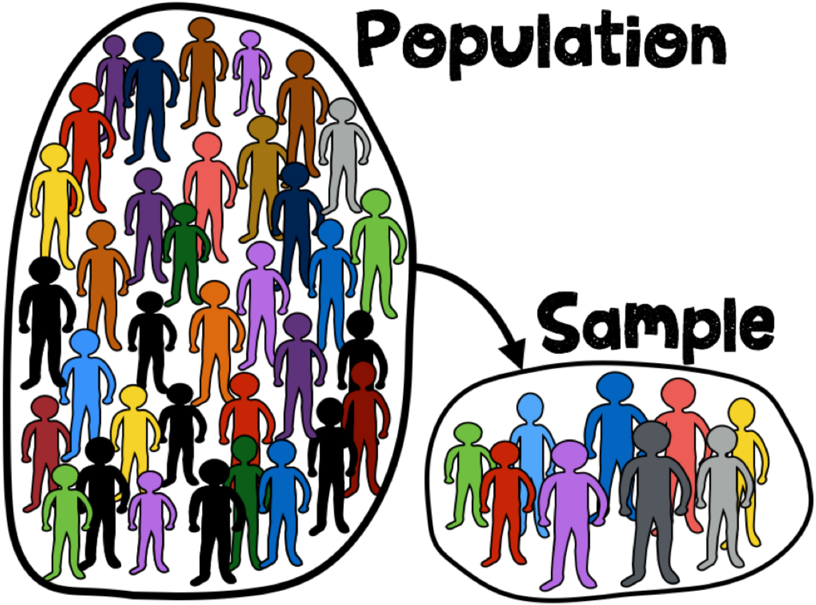
Populacja — zbiór elementów: osób, przedmiotów, zjawisk.

Cecha — własności elementów rozważanej populacji, w praktyce możemy mieć do czynienia z dwojakiego rodzaju cechami:

- cecha jakościowa — właściwość niemierzalna (np. kolor oczu, marka samochodu, stan cywilny);
- cecha ilościowa — właściwość mierzalna (np. waga, wzrost, wielkość zarobków).

Próba — podzbiór populacji.

Obserwacja, pomiar — wartość cechy, wyznaczona dla danego elementu próby.



Źródło 1: https://www.pngkey.com/detail/u2y3q8q8e6o0u2t4$_{p}$*opulation – and – sample –*

Podstawowe pojęcia

W celu uzyskania informacji o rozkładzie interesującej nas cechy można prowadzić badanie:

- **pełne** — badanie wszystkich elementów populacji,
- **częściowe** — badanie pewnego podzbioru populacji (próby).

Próba reprezentatywna — rozkład cechy w próbie nie powinien różnić się istotnie od rozkładu cechy w całej populacji.

Próba losowa — próba utworzona z elementów, które zostały wylosowane z elementów populacji.

Próba losowa prosta — wszystkie elementy populacji mają jednakowe szanse dostania się do próby.

Rozkład empiryczny cechy

Punktem wyjścia każdego wnioskowania statystycznego i podstawą wszelkich analiz statystycznych dotyczących badanej cechy jest **rozkład empiryczny** tej cechy.

Pojęciem tym określamy przyporządkowania poszczególnym wartościom cechy, obserwowanym w próbie, liczności lub częstotliwości ich występowania.

Rozkładem empirycznym liczności dla danych jakościowych nazywamy zbiór par uporządkowanych

$$\{(C_i, n_i) : i = 1, \ldots, k\},$$

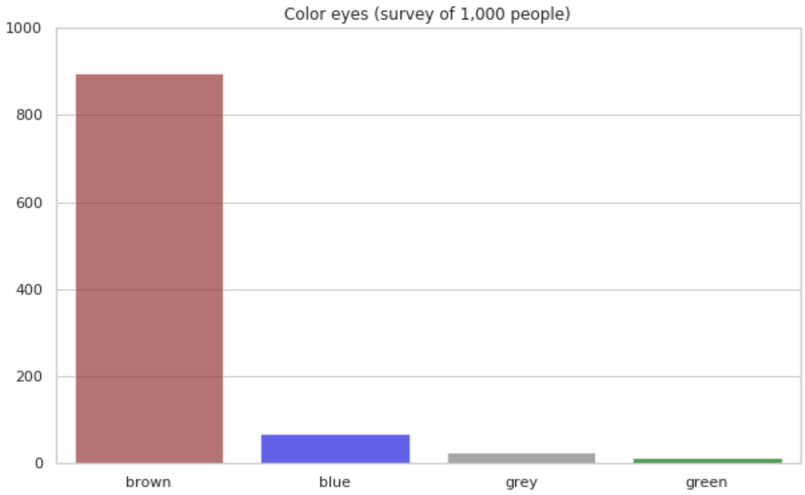
gdzie $C_1, \ldots, C_k$ oznacza rozłączną i kompletną listę kategorii odpowiadających możliwym wynikom obserwacji, natomiast $n_1, \ldots, n_k$ są licznościami obserwacji dla odpowiednich kategorii.

Rozkładem empirycznym częstości dla danych jakościowych nazywamy zbiór par uporządkowanych

$$\{(C_i, f_i = \frac{n_i}{n}) : i = 1, \ldots, k\},$$

gdzie $C_1, \ldots, C_k$ oznacza rozłączną i kompletną listę kategorii odpowiadających możliwym wynikom obserwacji, natomiast $f_1, \ldots, f_k$ są częstościami obserwacji dla odpowiednich kategorii, przy czym $n = n_1 + \ldots + n_k$ jest licznością próby.

Przykład wizualizacji rozkładu empirycznego liczności przy użyciu wykresu słupkowego



Założmy, że próba liczby n obserwacji. Przyjmijmy, że w próbie występuje k różnych wartości $w_1, \ldots, w_k$, przy czym $1 \leq k \leq n$.

Założmy dodatkowo, że wartość w_i występuje w próbie dokładnie n_i razy, przy czym oczywiście zachodzi $n_1 + \ldots + n_k = n$.

Dystrybuantą empiryczną nazywamy funkcję

$$F_n(x) = \begin{cases} 0 & \text{dla} \quad x < w_1 \\ \frac{1}{n}\sum_{j=1}^{i} n_j & \text{dla} \quad w_i \leq x < w_{i+1} \\ 1 & \text{dla} \quad x \geq w_k, \end{cases}$$

gdzie $i = 1, \ldots, k-1$.

Szereg rozdzielczy

W praktyce dystrybuantę empiryczną wyznacza się zwykle jedynie dla prób o stosunkowo małej liczności lub dla prób składających się z obserwacji przyjmujących niewielką liczbę różnych wartości.

Przy dużej liczności próbki, w celu ułatwienia analizy, dane grupuje się w przedziały klasowe (klasy), tworząc tzw. szereg rozdzielczy.

Liczbę klas k dobiera się w zależności od liczności próbki n. Zazwyczaj zaleca się, aby

$$\frac{3}{4}\sqrt{n} \leq k \leq \sqrt{n}$$

Szereg rozdzielczy

Jeżeli przyjmuje się, że klasy będą miały jednakową długość, wówczas długość klasy b wyznacza się ze wzoru

$$b = \frac{X_{n:n} - X_{1:n}}{k},$$

gdzie $X_{1:n}$ i $X_{n:n}$ oznaczają, odpowiednio, najmniejszą i największą obserwację w próbie.

Szereg rozdzielczy

Jeżeli przyjmuje się, że klasy będą miały jednakową długość, wówczas długość klasy b wyznacza się ze wzoru

$$b = \frac{X_{n:n} - X_{1:n}}{k},$$

gdzie $X_{1:n}$ i $X_{n:n}$ oznaczają, odpowiednio, najmniejszą i największą obserwację w próbie.

number klasy	granice klas	x_i^0	n_i	f_i	cn_i	cf_i
1	[19.5, 20.6)	20.05	6	0.15	6	0.15
2	[20.6, 21.7)	21.15	7	0.175	13	0.325
3	[21.7, 22.8)	22.25	12	0.3	25	0.625
4	[22.8, 23.9)	23.35	10	0.25	35	0.875
5	[23.9, 25.0]	24.45	5	0.125	40	1.0

Szereg rozdzielczy

Tradycyjny szereg rozdzielczy jest tablicą zawierającą:

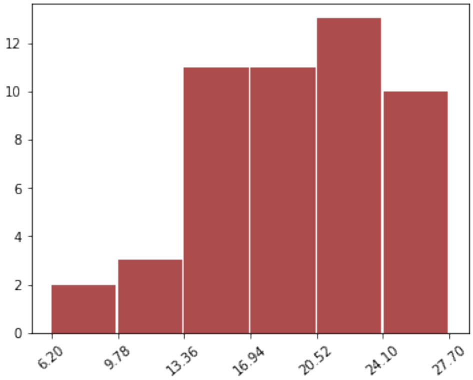
- granice klas,
- środki przedziałów klasowych x_i^0, przy czym x_i^0 wyznacza się wzorem
$$x_i^0 = \frac{\xi_i^- + \xi_i^+}{2},$$
gdzie ξ_i^- i ξ_i^+ oznaczają, odpowiednio, dolną i górną granicę i-tego przedziału klasowego,
- liczności n_i obserwacji należących do kolejnych klas,
- częstości f_i,
- liczności skumulowane cn_i, tzn. licznością skumulowaną i-tej klasy nazywamy łączną liczność danej klasy oraz klas poprzedzających daną klasę,
- skumulowane częstości nf_i.

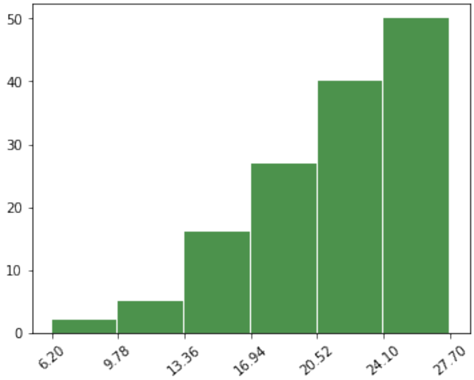
Histogram

Graficzną ilustracją szeregu rozdzielczego jest **histogram**.

- **histogram liczności** — wykres słupkowy, którego podstawę stanowią przedziały klasowe, natomiast wysokości słupków są proporcjonalne do liczności n_i poszczególnych klas;
- **histogram częstości** — wykres słupkowy, którego podstawę stanowią przedziały klasowe, natomiast wysokości słupków są proporcjonalne do częstości $f_i = \frac{n_i}{n}$ poszczególnych klas;

Przykład wizualizacji histogramu liczności i skumulowanych liczności.





Syntetyczne charakterystyki próby

Syntetyczna ocena dotyczy w szczególności

- poziomu cechy,
- jej zróżnicowania,
- kształtu rozkładu cechy.

Statystki opisowe dzieli na

- miary położenia,
- miary rozproszenia,
- charakterystyki kształtu.

Miary położenia

Wśród miar położenia wyodrębnia się

- miary tendencji centralnej — wskazują wartości „typowe” badanej cechy,
- miary pozycji — określają położenie wybranych obserwacji względem innych obserwacji w próbie.

Średnią arytmetyczną z próby losowej $X_1, \ldots, X_n$ nazywamy liczbę określoną wzorem

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Modą, dominantą nazywamy wartość najczęściej powtarzającą się w próbie.

Medianą z próby nazywamy taką wartość cechy, że co najmniej 50% obserwacji przyjmuje wartość nie większą od niej i jednocześnie co najmniej 50% obserwacji ma wartość nie mniejszą od tej wartości.

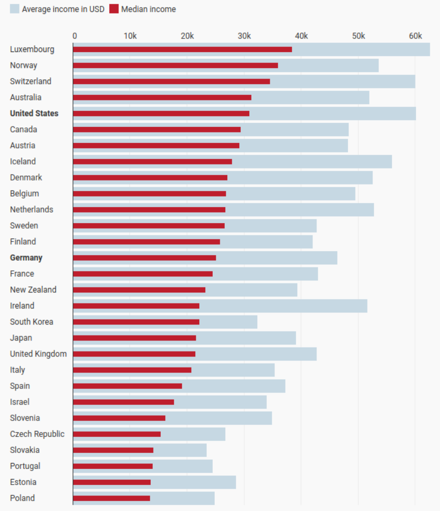
Formalnie, można ująć to wzorem

$$\mathrm{Med} = \begin{cases} X_{\frac{n+1}{2}:n} & \text{gdy } n \text{ jest nieparzyste} \\ \frac{1}{2}(X_{\frac{n}{2}:n} + X_{\frac{n}{2}+1:n}) & \text{gdy } n \text{ jest parzyste,} \end{cases}$$

gdzie $X_{k:n}$ oznacza k-tą statystykę pozycyjną, czyli k-tą obserwację w uporządkowanej niemalejąco próbie

$$X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}.$$

Porównanie miar tendencji centralnej



Źródło 2: https://blog.datawrapper.de/weekly-chart-income/

Miary pozycji — kwartyle

Pierwszy kwartyl (kwartyl dolny) Q_1 — to taka wartość cechy, że co najmniej 25% obserwacji przyjmuje wartość nie większą od niej i jednocześnie co najmniej 75% obserwacji ma wartość nie mniejszą od tej wartości.

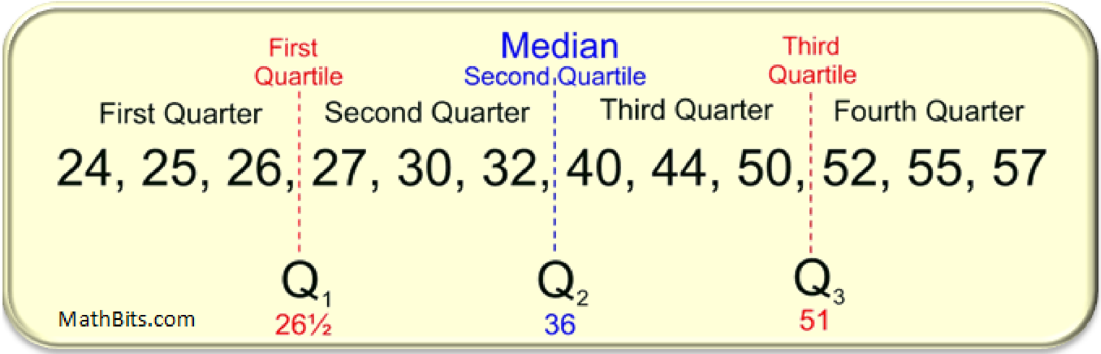
Drugi kwartyl jest równy **medianie**.

Trzeci kwartyl (kwartyl górny) Q_3 — to taka wartość cechy, że co najmniej 75% obserwacji przyjmuje wartość nie większą od niej i jednocześnie co najmniej 25% obserwacji ma wartość nie mniejszą od tej wartości.

Oprócz kwartyli w statystyce opisowej stosuje się

- **decyle** — dzielące uporządkowaną niemalejąco próbę na 10 równych części,
- **percentyle, centyle** — dzielące uporządkowaną niemalejąco próbę na 100 równych części.

Przykład



Źródło 3: https://mathbitsnotebook.com/Algebra1/StatisticsData/STboxplot.html

Miary rozproszenia

Rozstęp — odległość między najmniejszą największą obserwacją w próbie

$$R = X_{n:n} - X_{1:n}.$$

Rozstęp międzykwartylowy — odległość między pierwszym a trzecim kwartylem

$$\mathrm{IGR} = Q_3 - Q_1.$$

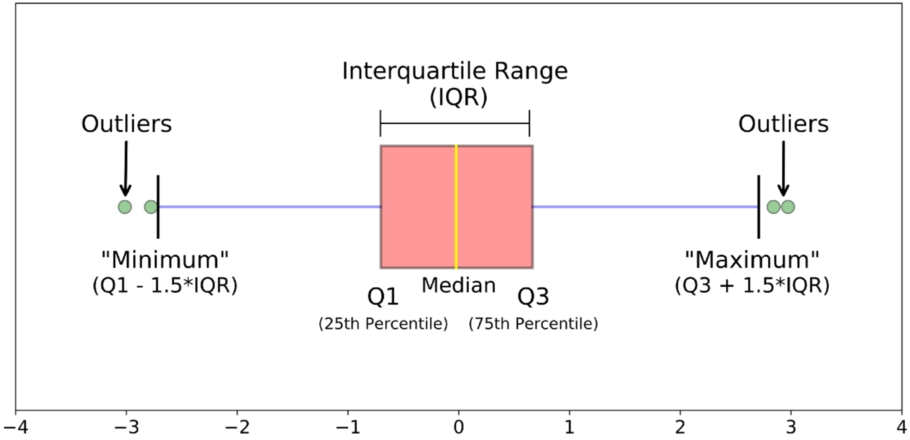
Wariancją z próby nazywamy liczbę

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Odchylenie standardowe — pierwiastek kwadratowy z wariancji

$$S = \sqrt{S^2}.$$

Wykres skrzynkowy (ang. *boxplot*)



Źródło 4: Understanding Boxplots

Charakterystyki kształtu

O rozkładzie empirycznym powiemy, że jest

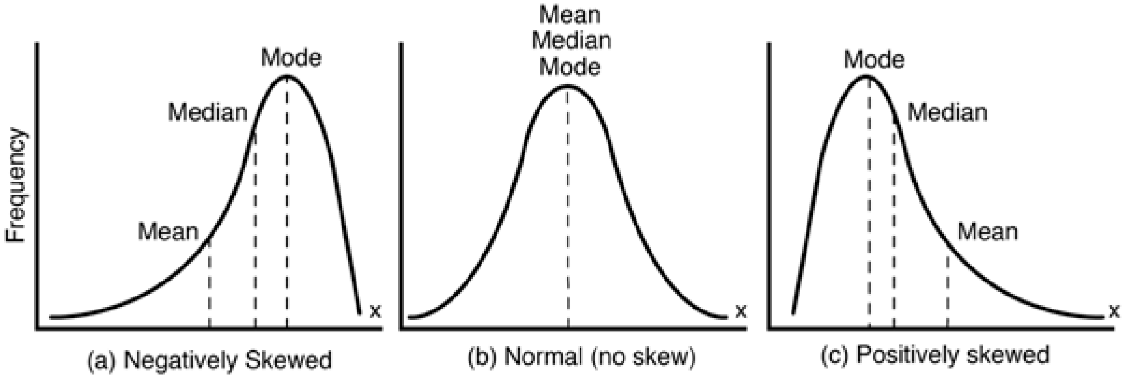
- **symetryczny**, lub
- **asymetryczny**
  - asymetria dodatnia (prawostronna),
  - asymetria ujemna (lewostronna).

Charakterystyki kształtu — współczynnik asymetrii, skośność

Współczynnikiem asymetrii nazywamy wielkość charakteryzującą stopień i kierunek asymetrii rozkładu empirycznego badanej cechy i wyznaczamy go ze wzoru

$$A = \frac{n\sum_{i=1}^{n}(X_i - \overline{X})^3}{(n-1)(n-2)S^3}.$$

- $A = 0$ oznacza, że obserwacje są symetrycznie rozłożone względem średniej,
- $A > 0$ określa dodatnią asymetrię,
- $A < 0$ określa ujemną asymetrię.



Źródło 5: Źródło: https://www.researchgate.net

Miary korelacji — współczynnik korelacji Pearsona

W przypadku jednoczesnego badania dwóch cech pewnej populacji naszą próbą jest ciąg par

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

gdzie X_i oraz Y_i oznaczają, odpowiednio, wartości pierwszej i drugiej cechy przyjmowane przez i-ty element próby.

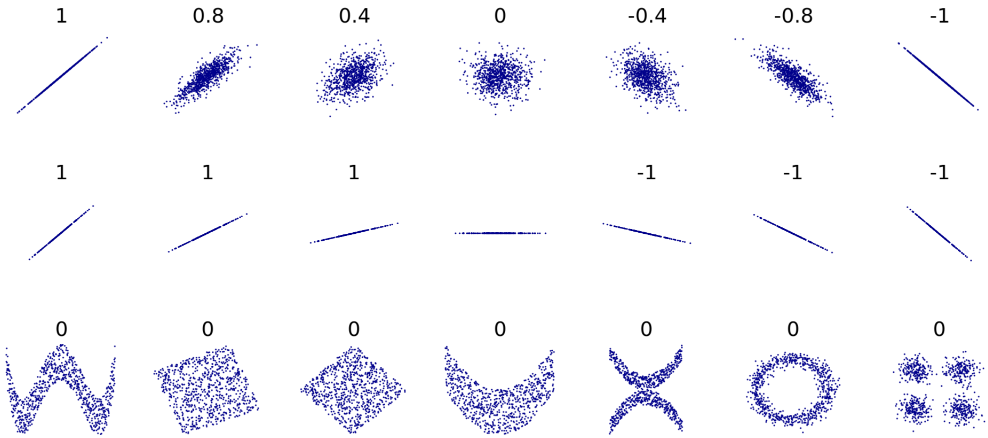
Wówczas **współczynnik korelacji Pearsona** jest dany wzorem

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

Miary korelacji — współczynnik korelacji Pearsona

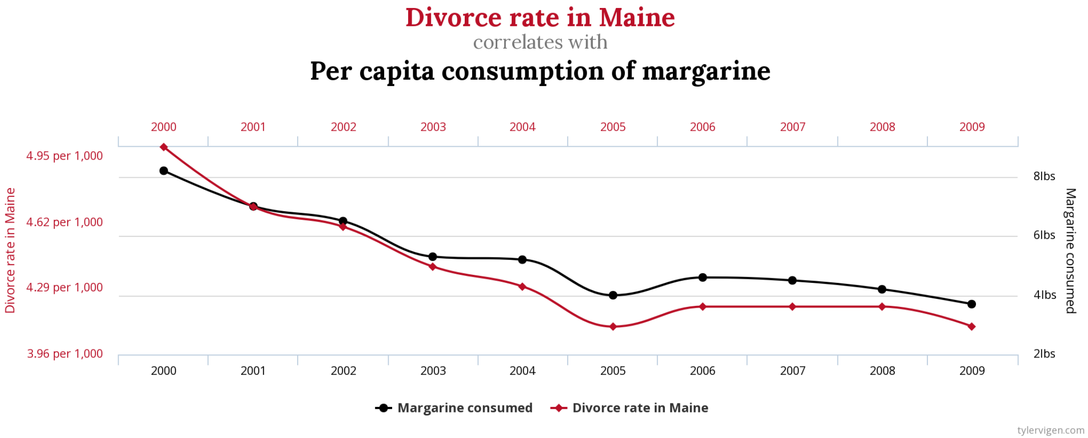
$r \in [-1, 1]$:

- $r = 1$, oznacza, że punkty $(X_1, Y_1), \ldots, (X_n, Y_n)$, leżą na prostej rosnącej i wskazują silną korelację liniową dodatnią;
- $r = -1$, oznacza, że punkty $(X_1, Y_1), \ldots, (X_n, Y_n)$, leżą na prostej malejącej i wskazują silną korelację liniową ujemną;
- $r = 0$, oznacza to brak skorelowania liniowego, co nie jest jednoznaczne z brakiem istnienia jakiegokolwiek związku między badanymi cechami.



Źródło 6: Źródło: https://en.wikipedia.org/wiki/Correlation$_a$nd$_d$ependence

Fałszywe korelacje



Źródło 7: Źródło: http://tylervigen.com/spurious-correlations

Materiały na podstawie

- Grzegorzewski P., Bobecka K., Dembińska A., Pusz J., Rachunek prawdopodobieństwa i statystyka, WSISiZ, Warszawa, wyd. V - 2008.
- http://www.biecek.pl/Eseje/
- http://smarterpoland.pl/index.php/category/zly-wykres/
- https://python-graph-gallery.com/
- https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/