Data Visualization

Rimini – 25/10/2021

Alessia Angeli

Studente di dottorato in Data Science and Computation

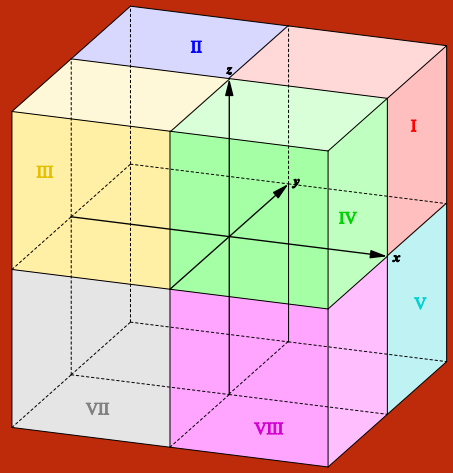
Dipartimento di Informatica – Scienza e Ingegneria

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

VARLAB: VIRTUAL AND AUGMENTED REALITY LAB

Contatti

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Data Visualization – Space & Attributes

Alessia Angeli

Studente di dottorato in Data Science and Computation

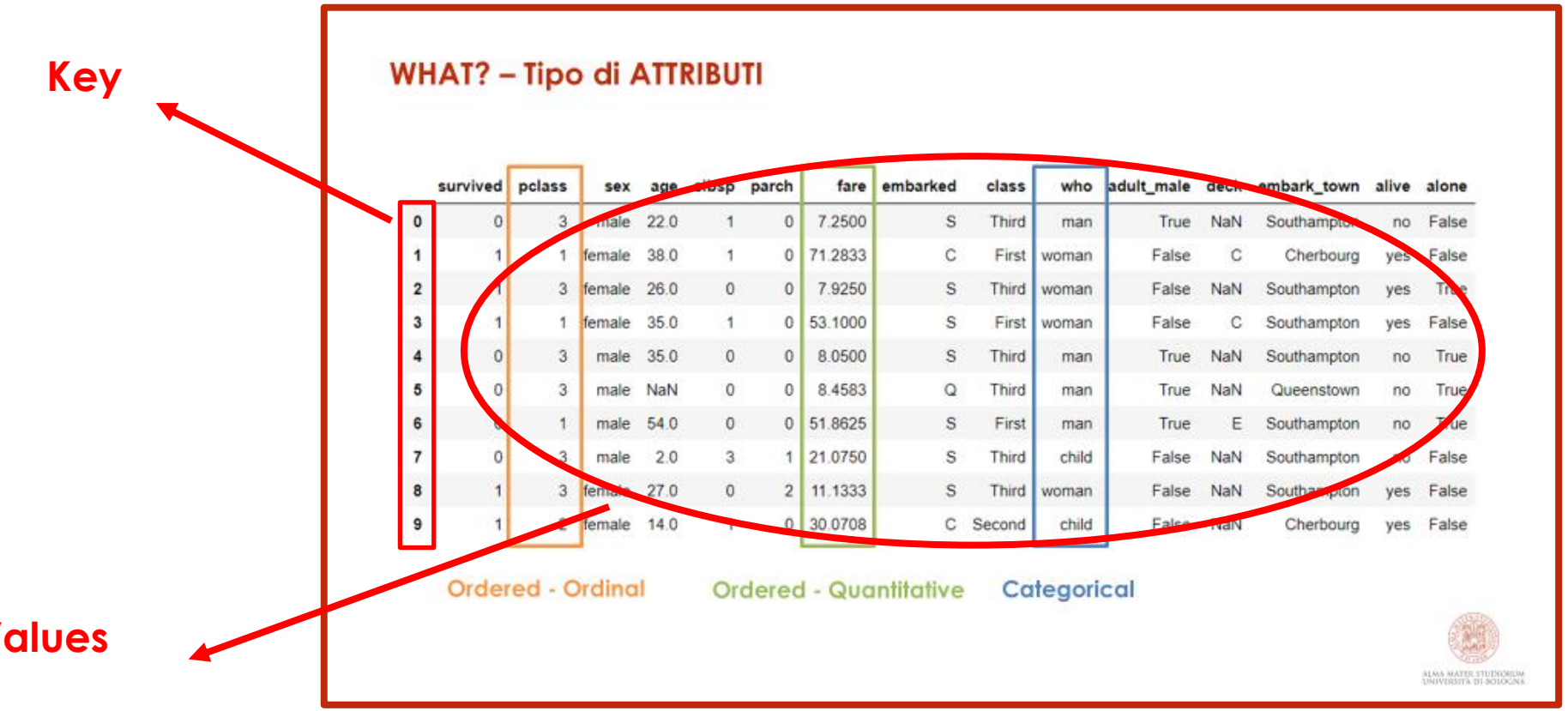
Dipartimento di Informatica – Scienza e Ingegneria

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Attributes: Keys and Values

- **Attribute – Key:** un attributo viene definito chiave quando identifica univocamente (in modo unico) un item (elemento);

- **Attribute – Value:** un attibuto se non è una chiave viene definito valore.

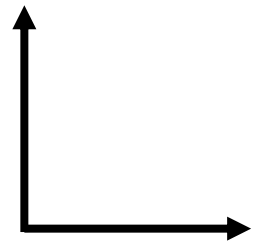


	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False

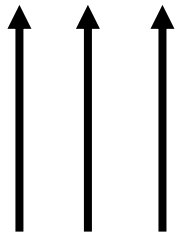
Space & Quantitative Attributes

- Esprimere il valore con la posizione;

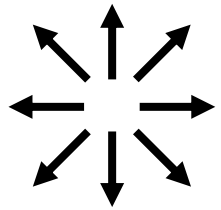
- Scegliere il modo più efficace per disegnare gli assi.



ORTOGONALI



PARALLELI

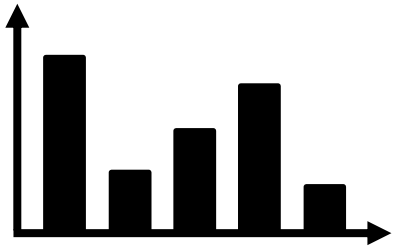


RADIALI

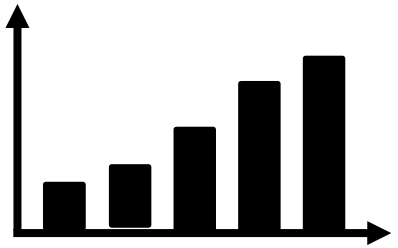
Space & Qualitative Attributes

- Codificare categorical key(s) in regioni separate;

- Scegliere un allineamento appropriato;

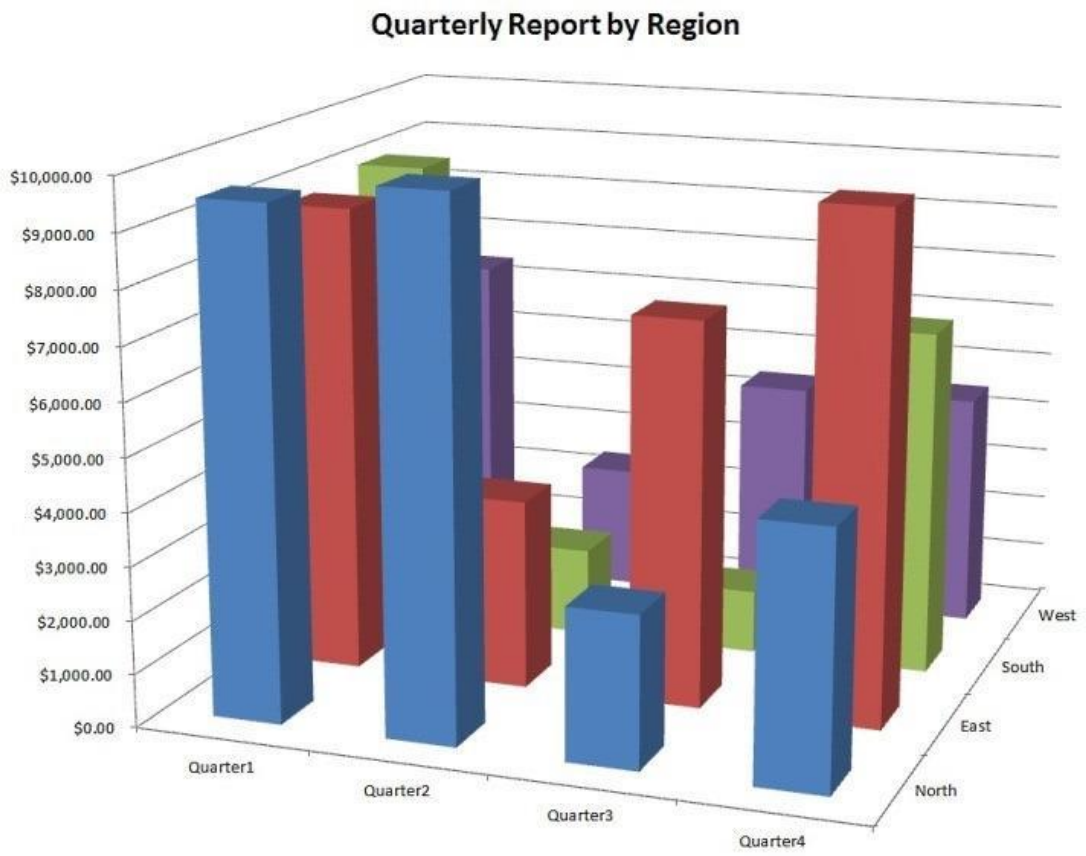


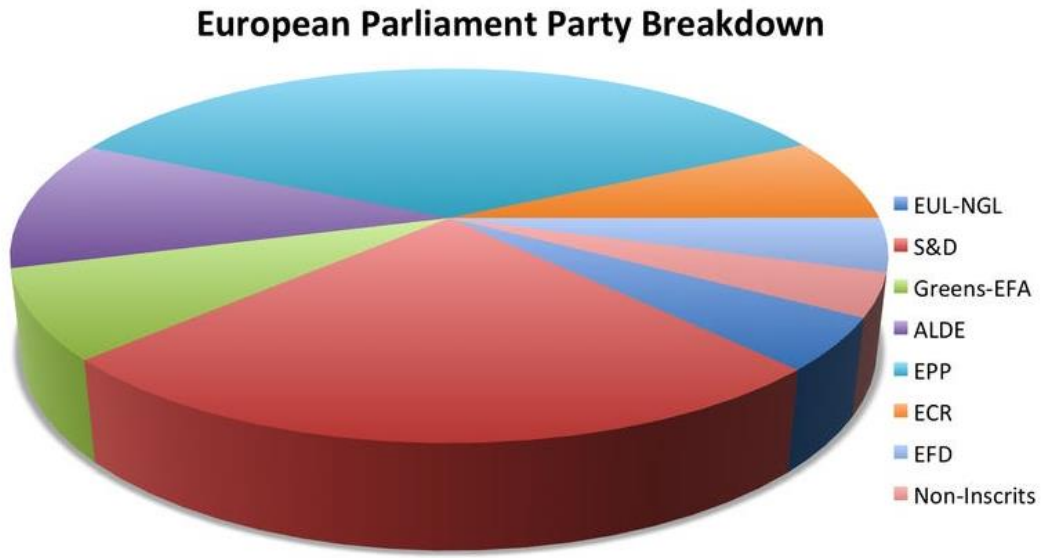
- Scegliere un ordine appropriato.



Attenzione alle visualizzazioni 3D!

3D… Senza valide ragioni? Anche no.





Data Visualization – Plots/Graphs

Alessia Angeli

Studente di dottorato in Data Science and Computation

Dipartimento di Informatica – Scienza e Ingegneria

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Quantitative Attributes

Scatter plot
Histogram
Scatter plot matrix
Box plot
Violin plot
Radar chart
…

Scatter plot

What?

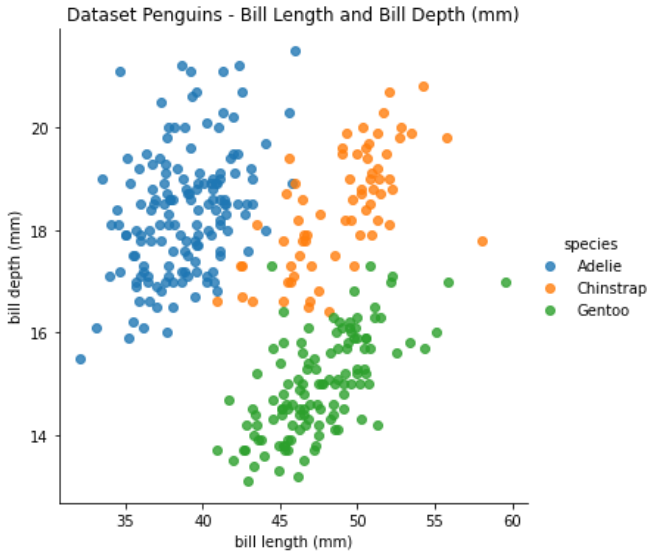
- 2 quantitative attributes;

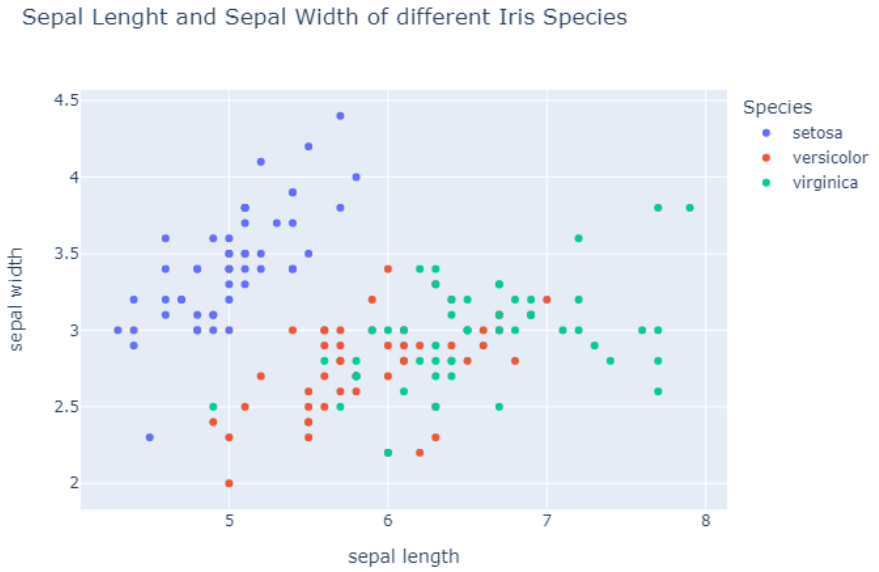
Why?

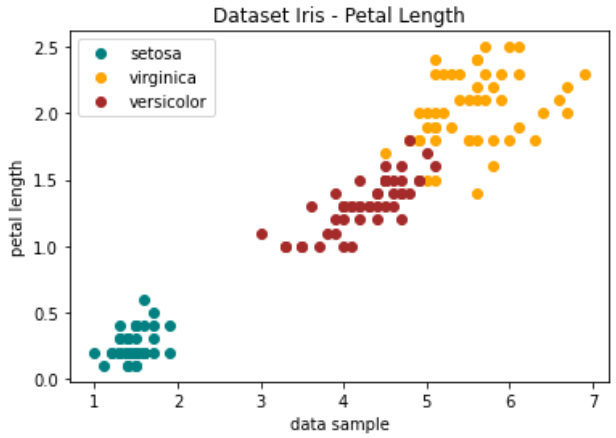
- Visualizzare correlazioni e distribuzioni;
- Identificare outliers, patterns e clusters;

Remarks

- Fino a ~100 items;
- Colore e dimesione possono essere usati per codificare categorical attributes aggiuntivi (bubble plot).





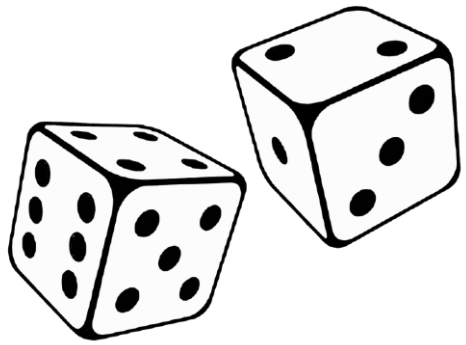


Definizione

DISTRIBUZIONE DI PROBABILITA': Una distribuzione di probabilità è un modello matematico che associa ai valori (possibili) di una variabile aleatoria (continua o discreta) le probabilità che tali valori possano essere assunti da tale variabile. Formalmente le distribuzioni vengono espresse da funzioni matematiche, **funzione densità di probabilità** e **funzione di probabilità**, rispettivamente per variabili aleatorie continue e discrete.

ESEMPIO

Si lanciano 2 dadi e si considera come variabile aleatoria la somma risultante.



Somma	# Combinazioni	Probabilità
2	1	0.03
3	2	0.06
4	3	0.08
5	4	0.11
6	5	0.14
7	6	0.17
8	5	0.14
9	4	0.11
10	3	0.08
11	2	0.06
12	1	0.03

Σ 36

Σ 1

Histogram

What?

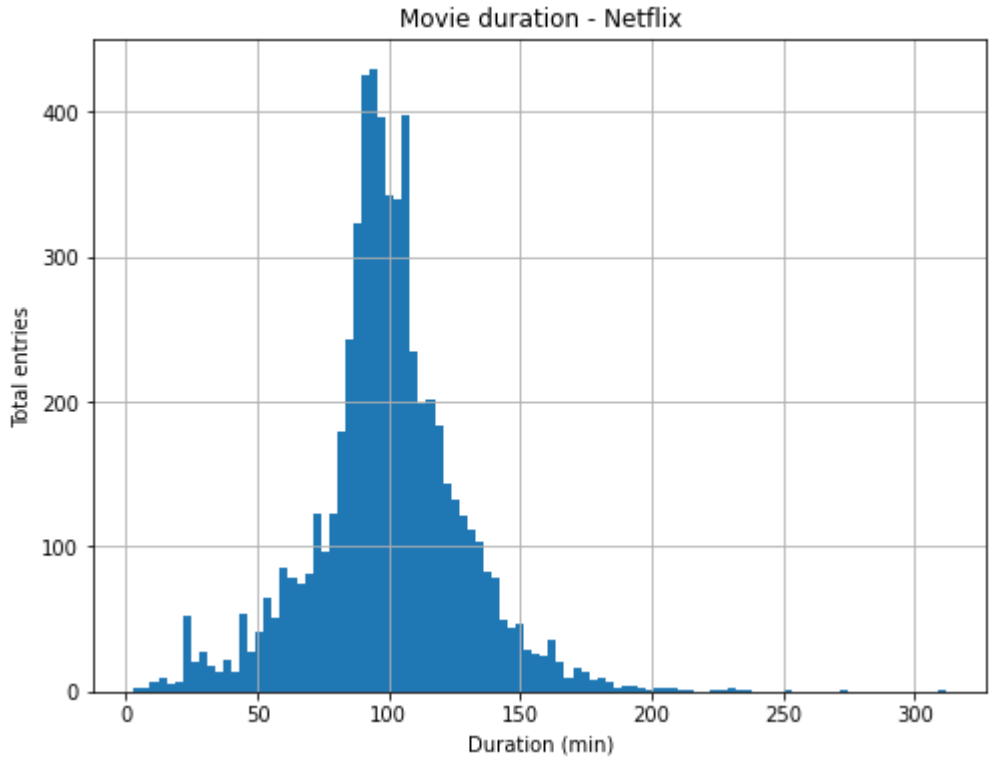
- 1 quantitative attribute;

Why?

- Visualizzare distribuzioni;
- Identificare patterns e range;

Remarks

- Una linea (o un'area) può essere visualizzata per mostrare la funzione di densità calcolata;
- Gli items possono essere visualizzati con dei punti.



Definizione

MATRICE: una matrice è una tabella ordinata. Le righe orizzontali vengono chiamate *righe* della matrice e le righe verticali *colonne* della matrice.

Generalmente una matrice si indica con una lettera maiuscola e viene scritta nel modo seguente:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

dove i pedici di ogni elemento della matrice indicano, rispettivamente, la riga e la colonna in cui l'elemento è posizionato.

Quindi a_{ij} è l'elemento della matrice A che si trova nella riga *i-esima* e nella colonna *j-esima*.

https://www.youmath.it/lezioni/algebra-lineare/matrici-e-vettori/1558-cos-e-una-matrice.html

Scatter plot matrix

What?

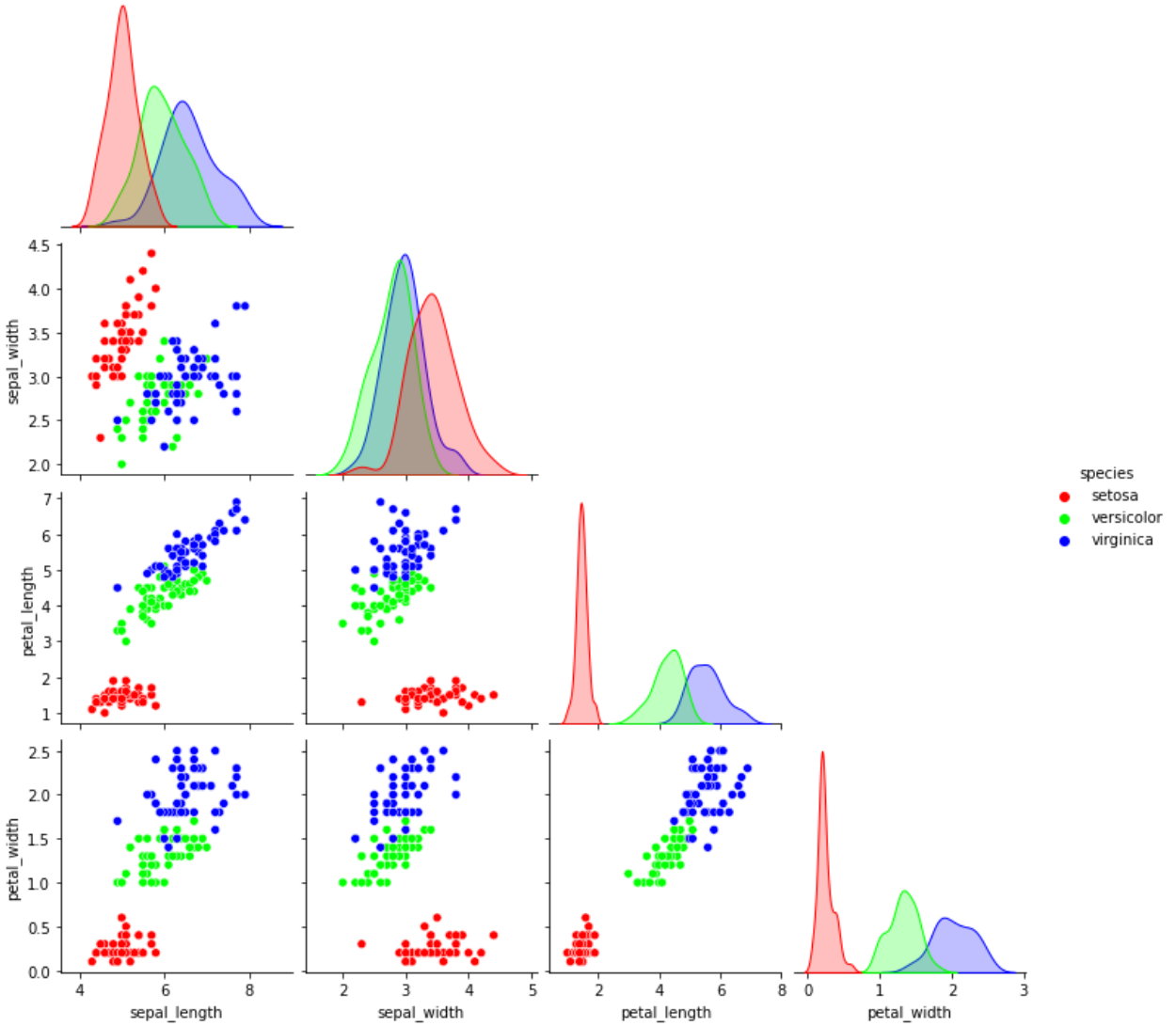
- *N* quantitative attributes;

Why?

- Visualizzare correlazioni e distribuzioni;
- Identificare outliers, patterns e clusters;

Remarks

- Fino a ~12 attributi e ~100 items;
- E' possibile visualizzare solo la parte triangolare inferiore della matrice.



Statistica descrittiva – alcune definizioni

Considerando un insieme di dati numerici si definiscono:

5 7 4 6 5

MEDIA (MEDIA ARITMETICA): rapporto tra la somma dei dati e il numero dei dati.

(5+7+4+6+5)/5=5.4

MODA: il valore del dato che si presenta con maggiore frequenza (possono essere presenti più valori di moda).

5 – dato con massima frequenza (2)

MEDIANA: è il valore centrale tra i dati ordinati in modo crescente o decrescente. Se l'insieme contiene un numero di dati dispari c'è un unico valore centrale e questo è la mediana. Se l'insieme contiene un numero di dati pari, invece, ci sono due valori centrali e di solito come mediana viene considerata la media aritmetica di questi.

5 – è il valore centrale in 4 5 5 6 7

https://www.youmath.it/lezioni/algebra-elementare/lezioni-di-algebra-e-aritmetica-per-scuole-medie/2064-media-moda-mediana.html

Statistica descrittiva – alcune definizioni

Oltre alla mediana, che divide a metà un insieme di dati ordinati, vengono usati anche altri indici che dividono tale insieme in determinate percentuali detti **quantili**, **quartili** e **percentili**.

PERCENTILI: sono un caso particolare dei quantili e, come si intuisce dal nome, dividono l'insieme di dati ordinati in 100 parti.

- il 1° percentile lascia alla sua sinistra un centesimo (1%) degli elementi dell'insieme ordinato;
- il 10° percentile lascia alla sua sinistra il 10% degli elementi;
- il 50° percentile (che coincide con la mediana) lascia alla sua sinistra il 50% degli elementi;
- …

QUARTILI: questi si ottengono dividendo l'insieme di dati ordinati in 4 parti uguali.

- il **primo quartile** (che coincide con il 25-esimo percentile) è il valore che lascia alla sua sinistra il 25% degli elementi;
- il **secondo quartile** (che coincide con la mediana e con il 50-esimo percentile) è il valore che lascia alla sua sinistra il 50% dei dati;
- il **terzo quartile** (che coincide con il 75-esimo percentile) è il valore che lascia il 75% degli elementi a sinistra e il 25% a destra.

https://www.webtutordimatematica.it/materie/statistica-e-probabilita/statistica-descrittiva/quartili-quantili-e-percentili

Box plot

What?

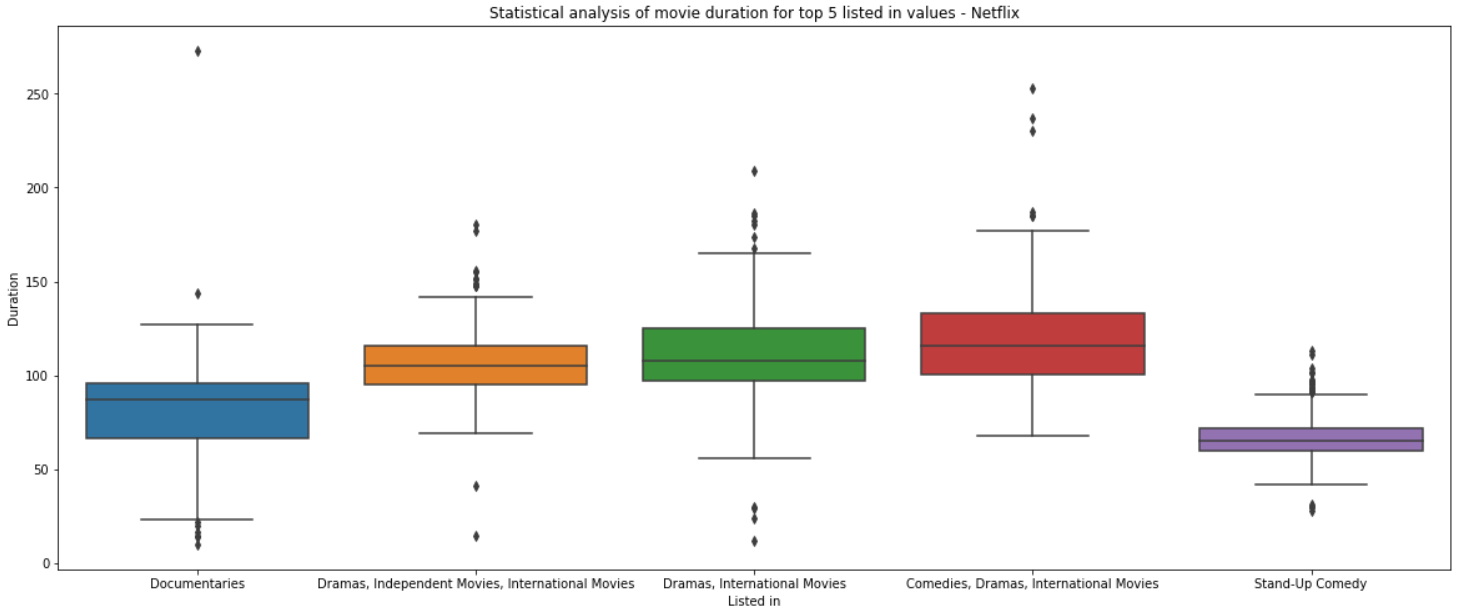
- *N* quantitative attributes (oppure 1 quantitative attribute ed 1 categorical key);

Why?

- Visualizzare distribuzioni;
- Identificare outliers, valori estremi, range etc.;

Remarks

- Il colore può codificare un categorical attribute aggiuntivo;
- Possibile effettuare raggruppamenti.



Violin plot

What?

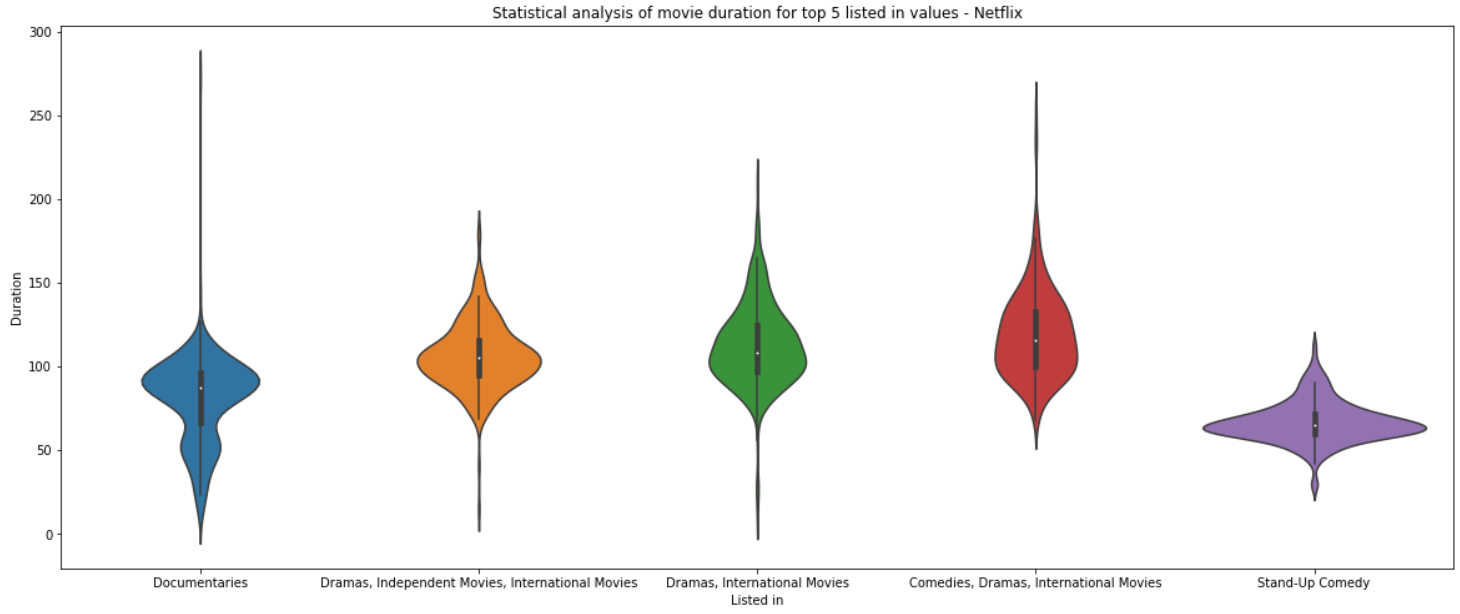
- *N* quantitative attributes (oppure 1 quantitative attribute ed 1 categorical key);

Why?

- Visualizzare distribuzioni;
- Identificare range;

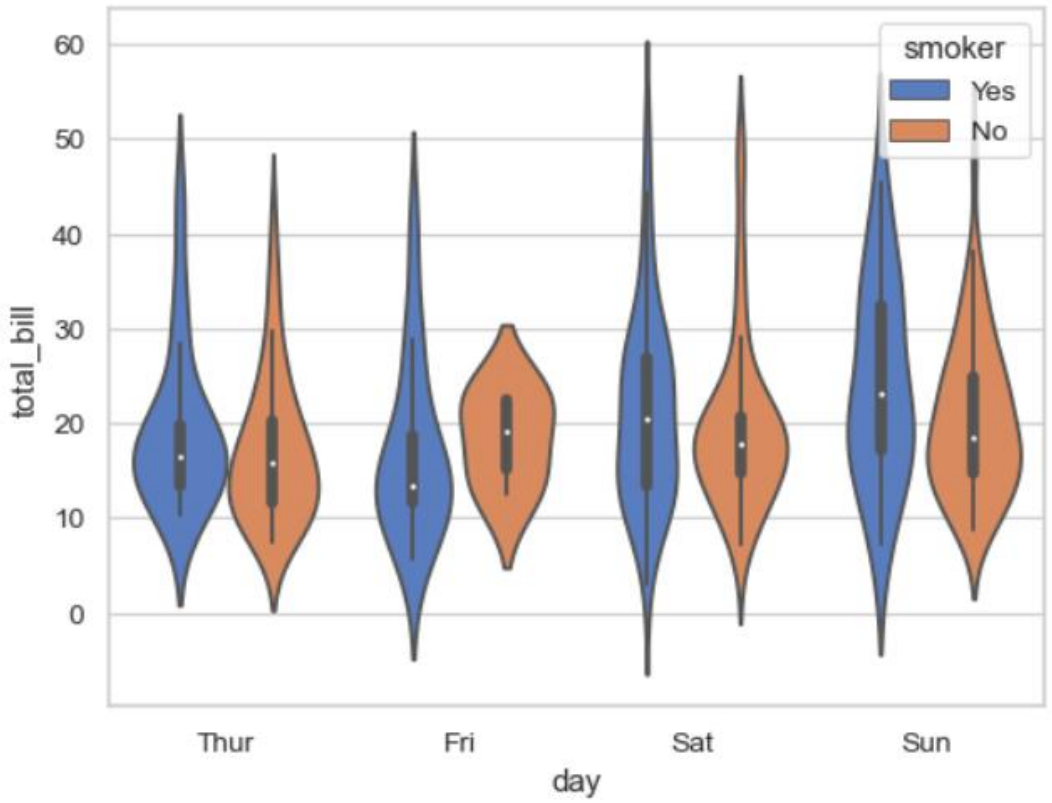
Remarks

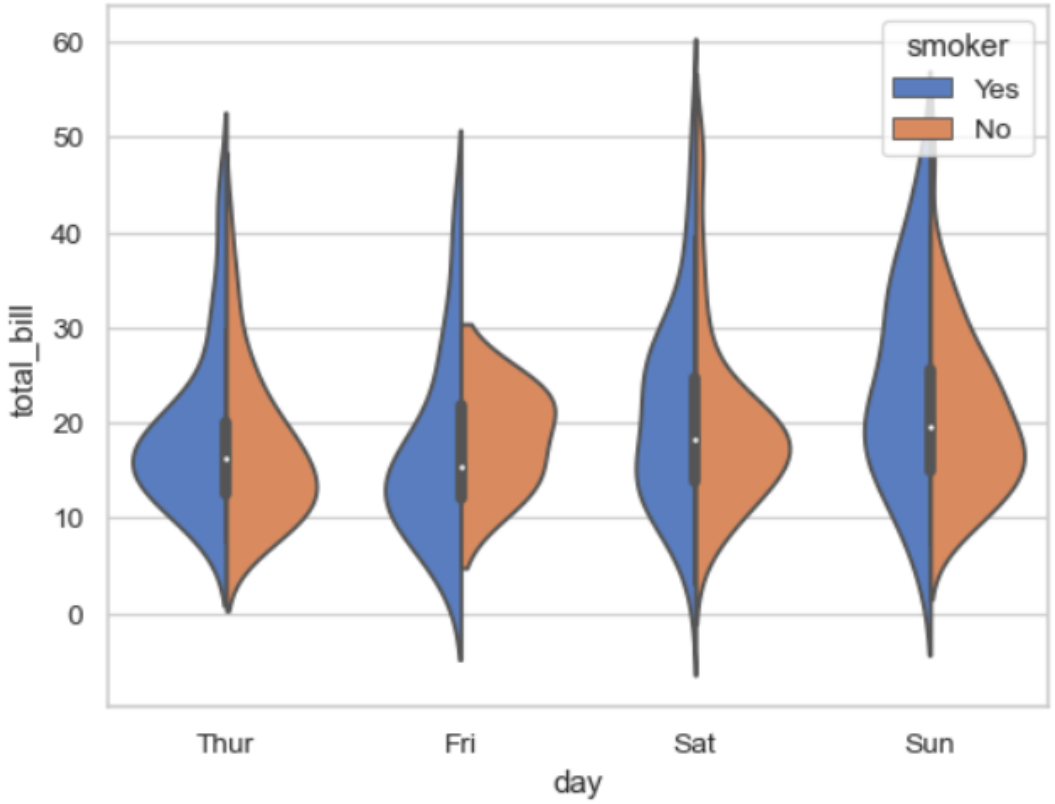
- Il colore può codificare un categorical attribute aggiuntivo (è possibile effettuare anche uno split se i dati lo consentono).



Esempi – SOLO immagine

Violin plot





https://seaborn.pydata.org/generated/seaborn.violinplot.html

Radar chart

What?

- *N* quantitative attributes;

Why?

- Identificare patterns;
- Confrontare valori;

Remarks

- Fino a ~12 attributi;
- Il colore può codificare una categorical key aggiuntiva (fino a 3-4 valori).

Qualitative Attributes

Bar plot
Multi-set bar plot
Pie chart
Word Cloud
…

Bar plot

What?

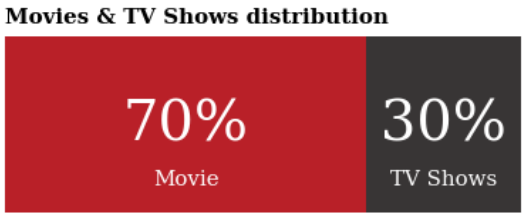
- 1 quantitative attribute;
- 1 categorical key;

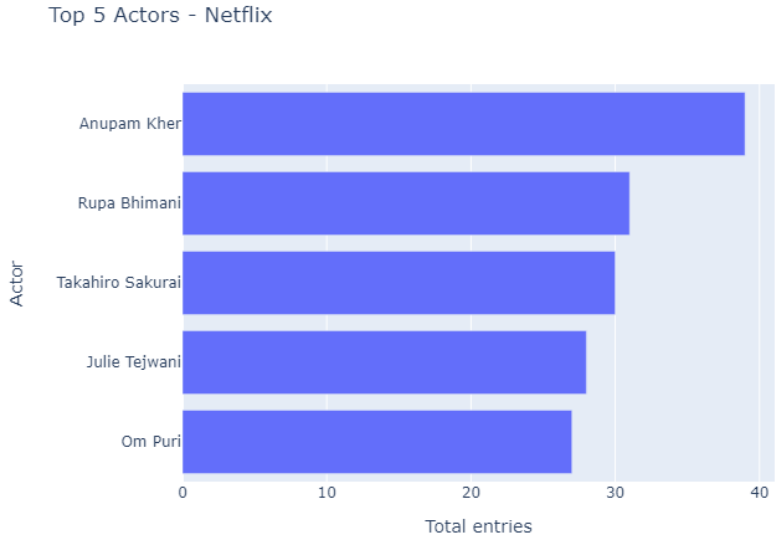
Why?

- Confrontare/evidenziare valori;
- Identificare valori estremi;

Remarks

- Fino a ~100 barre;
- Keys vs valori ordinati;
- Non adatto per visualizzare trends.





Multi-set bar plot

What?

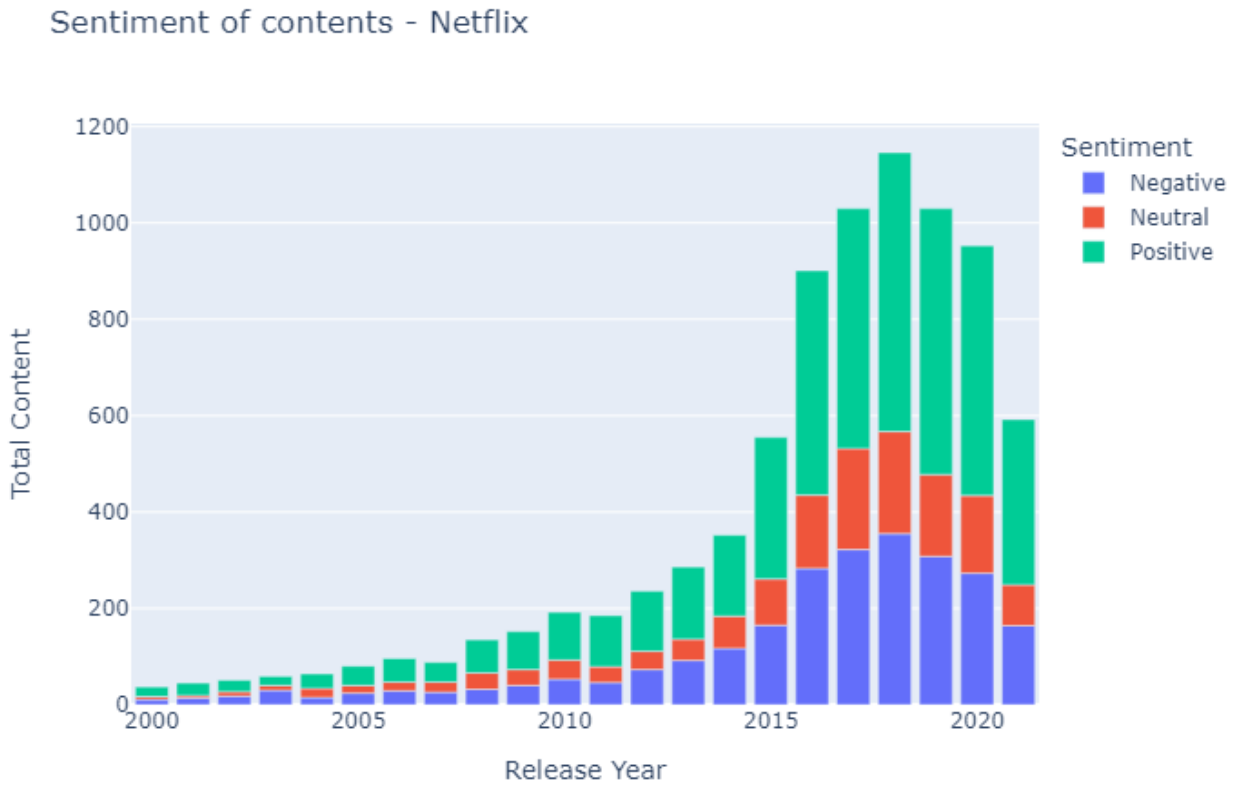
- 1 quantitative attribute;
- 2 categorical keys;

Why?

- Confrontare valori;
- Identificare patterns;

Remarks

- Visualizzare fino a ~100 barre;
- Riuscire a raggruppare/confrontare items, patterns.



Pie chart

What?

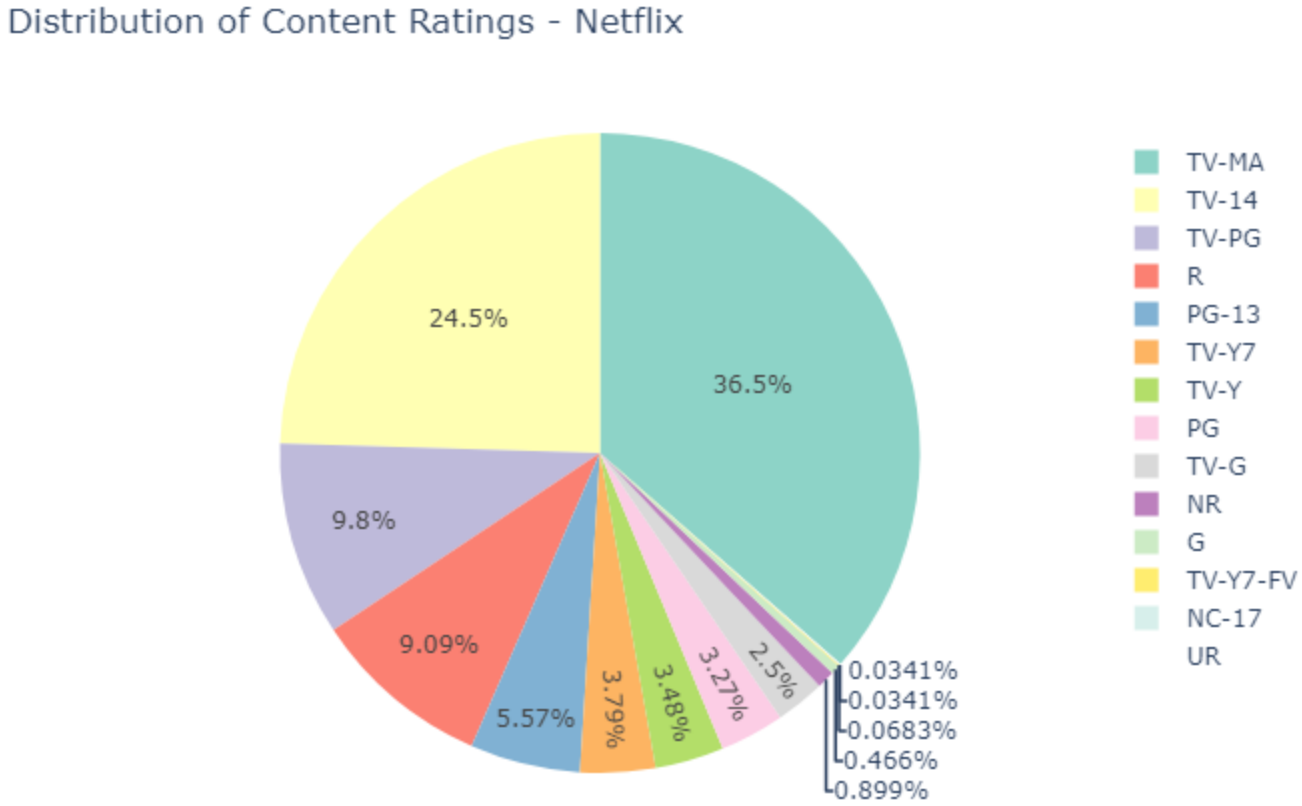
- 1 quantitative attribute;
- 1 categorical key;

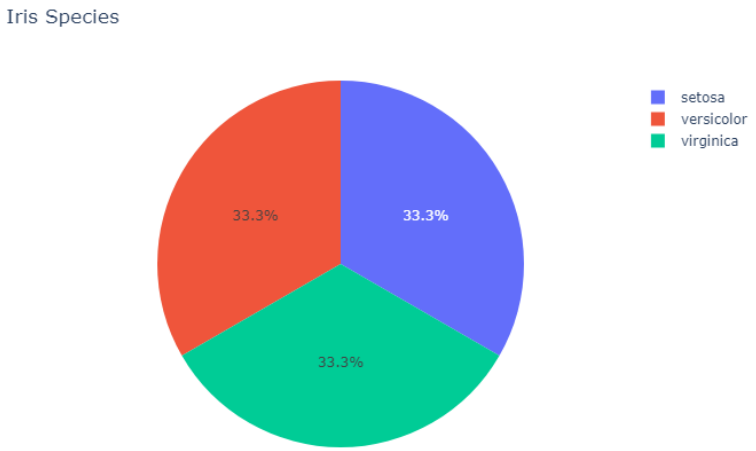
Why?

- Evidenziare una parte rispetto al tutto;

Remarks

- Meno dettagliato ed accurato rispetto al layout lineare;
- L'area centrale può essere rimossa (donut chart).





Word cloud

What?

- 1 categorical attribute -> testo;
- 1 quantitative attribute -> frequenza;

Why?

- Visualizzare distribuzione di parole;
- Visualizzare un sommario;

Remarks

- Scarsa accuratezza;
- Bias dovuto alla lunghezza e alla struttura delle parole.

2 Categorical Keys

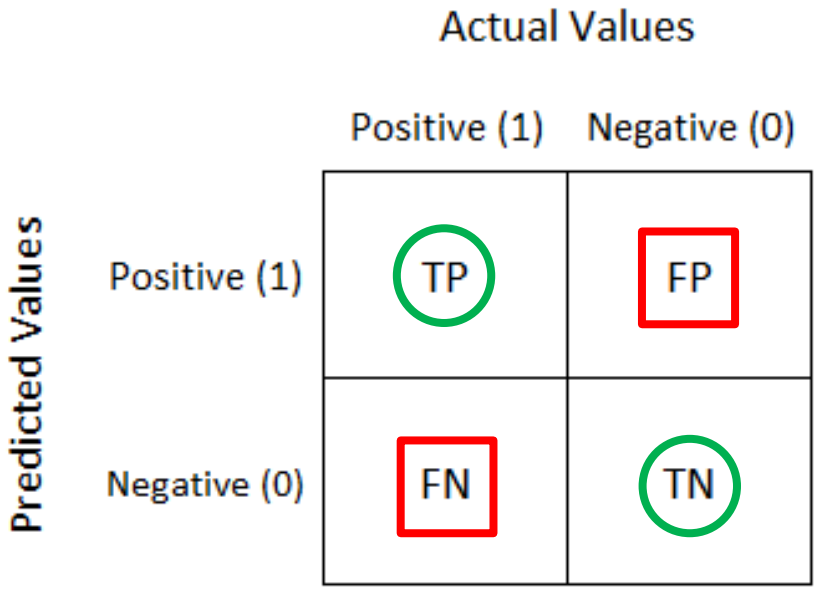
Heatmap

…

Definizione

MATRICE DI CONFUSIONE: è un metodo per visualizzare le performance di un algoritmo rispetto ad un problema di classificazione dove gli outputs possono essere due o più classi.

Nel caso di problema di classificazione binario (due classi in outputs) la matrice di confusione sarà composta da quattro elementi: **True Positive (TP)**, **False Positive (FP)**, **False Negative (FN)**, **True Negative (TN)**.



Inoltre, la matrice di confusione è estremamente comoda per calcolare *Precision*, *Recall*, *Accuratezza*, … (se ne parlerà nelle prossime lezioni).

https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Heatmap

What?

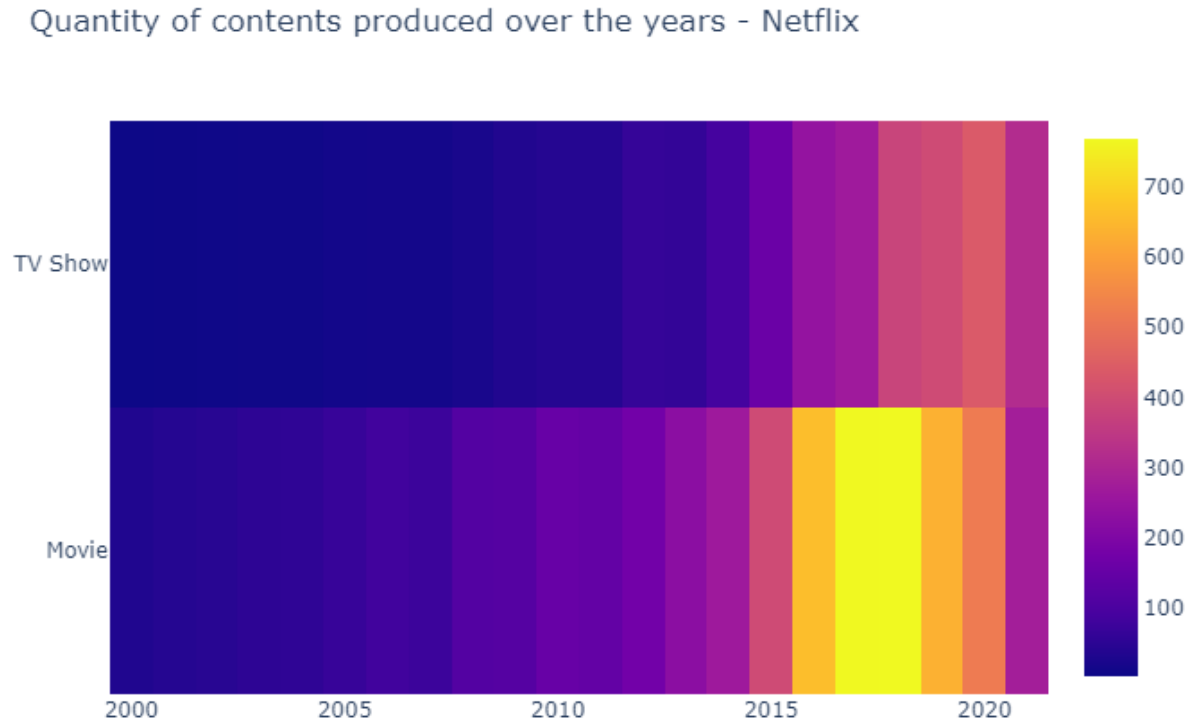
- 2 categorical key;
- 1 quantitative attribute;

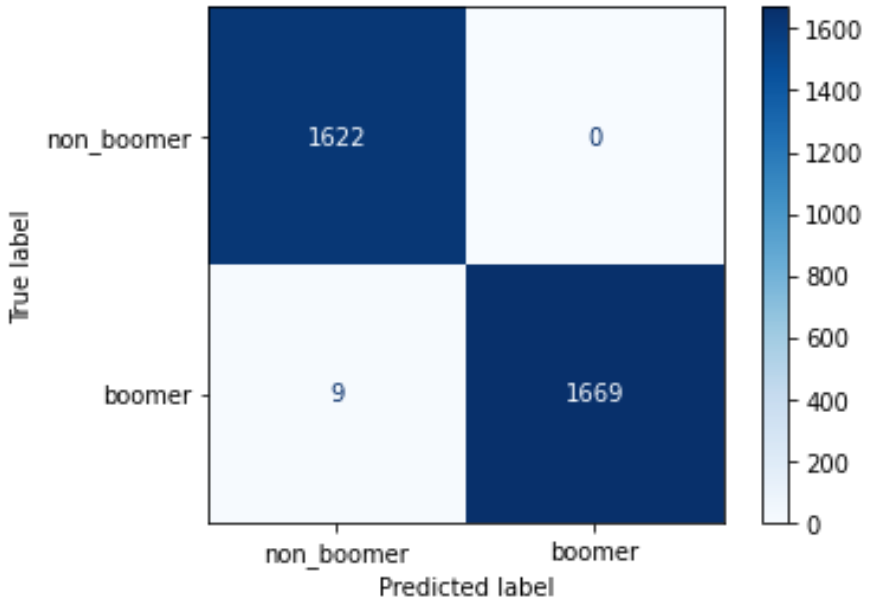
Why?

- Visualizzare correlazioni;
- Identificare patterns, outliers;
- Confusion matrix for classification result visualization;

Remarks

- Fino a ~1M di items;
- L'ordine delle keys influisce la visibilità dei patterns.





For dealing with time

Line graph
Stacked area graph
…

Line graph

What?

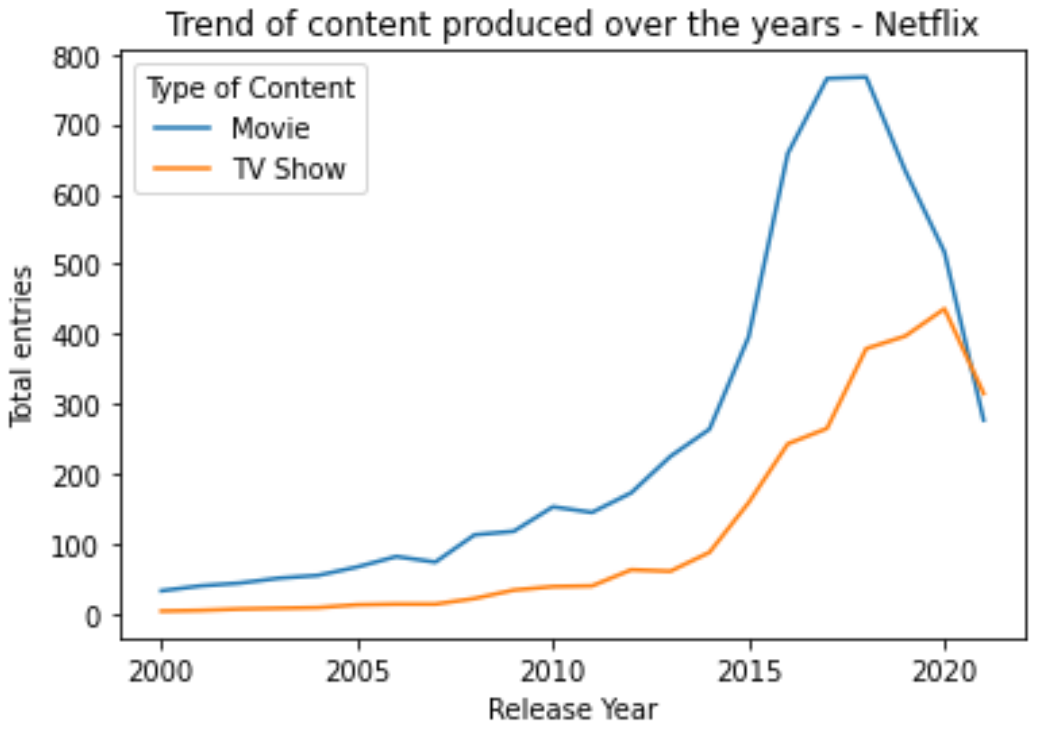
- 1 ordered key -> time;
- 1 quantitative attribute;

Why?

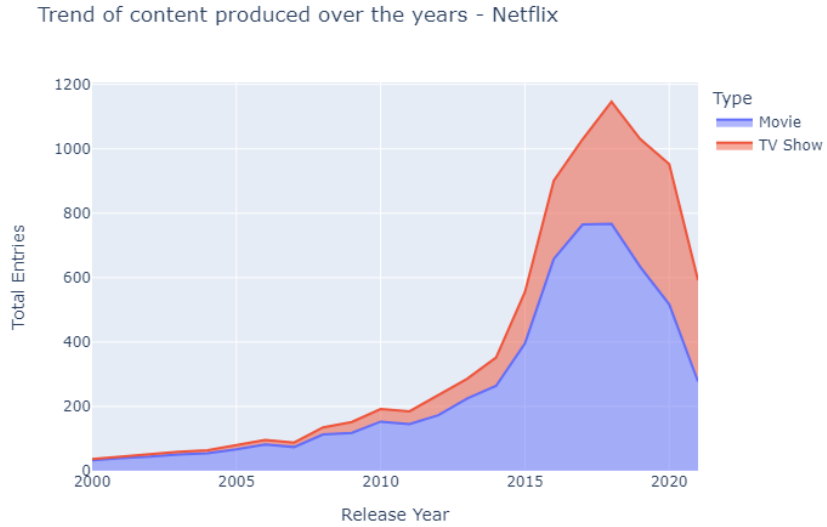
- Identificare e confrontare trends;

Remarks

- Fino a 10-20 linee;
- Il colore può codificare un categorical attribute additivo.



Stacked area graph

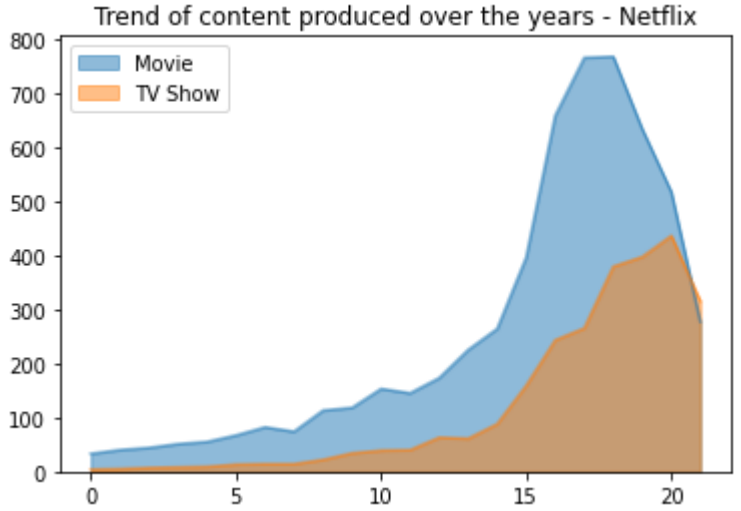


What?

- 1 ordered key -> time;
- 1 categorical attribute;

Why?

- Visualizzare trends;
- Evidenziare una parte rispetto al tutto;
- Confrontare valori;



Remarks

- Fino a pochi valori.

ATTENZIONE ALLA BASE DI RIFERIMENTO DELLE AREE COLORATE

Hands-On – Caso Studio – Visualizzazione Dati NETFLIX

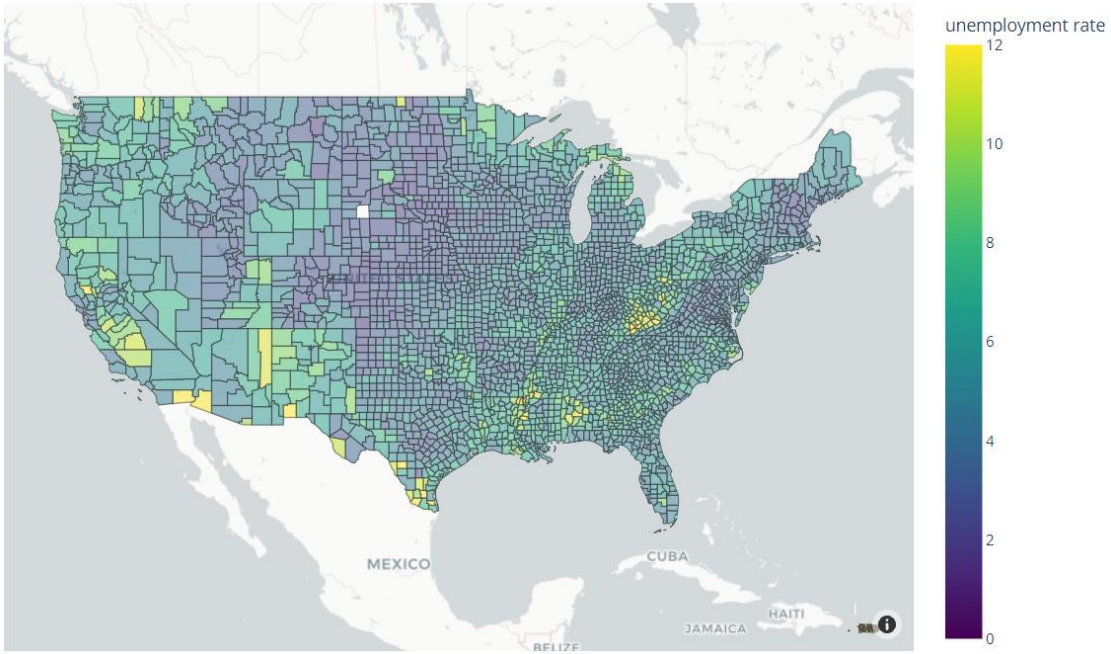
- **Notebook 6 – Caso Studio Netflix**

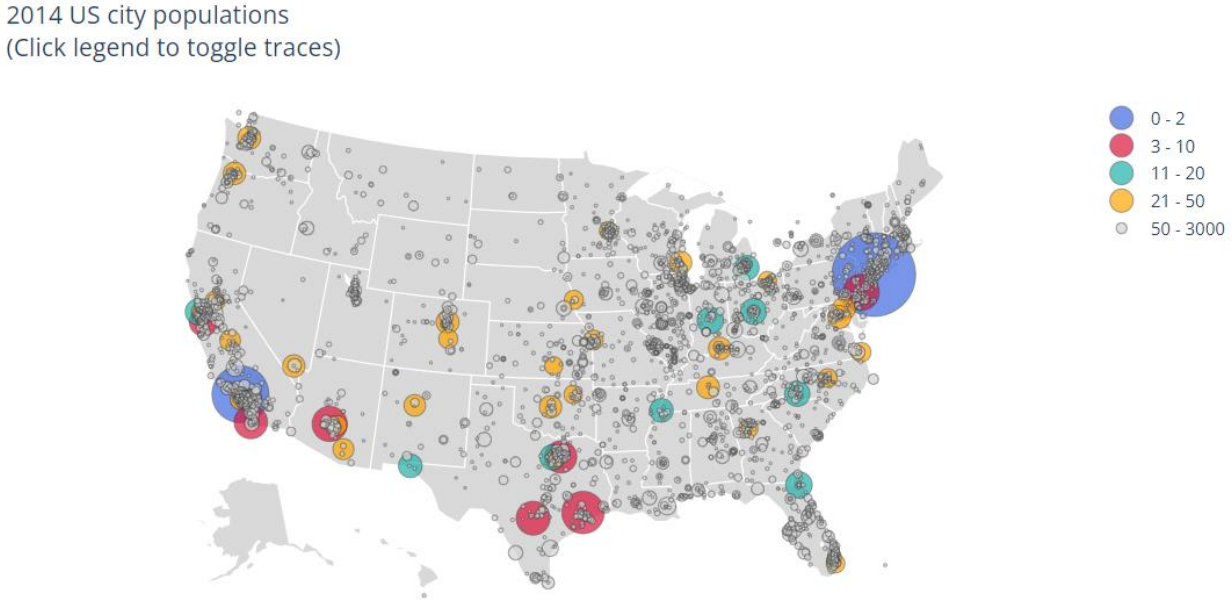
https://www.netflix.com/it/

https://www.kaggle.com/shivamb/netflix-shows

Esempi – SOLO immagine

Cloropeth map e Bubble map

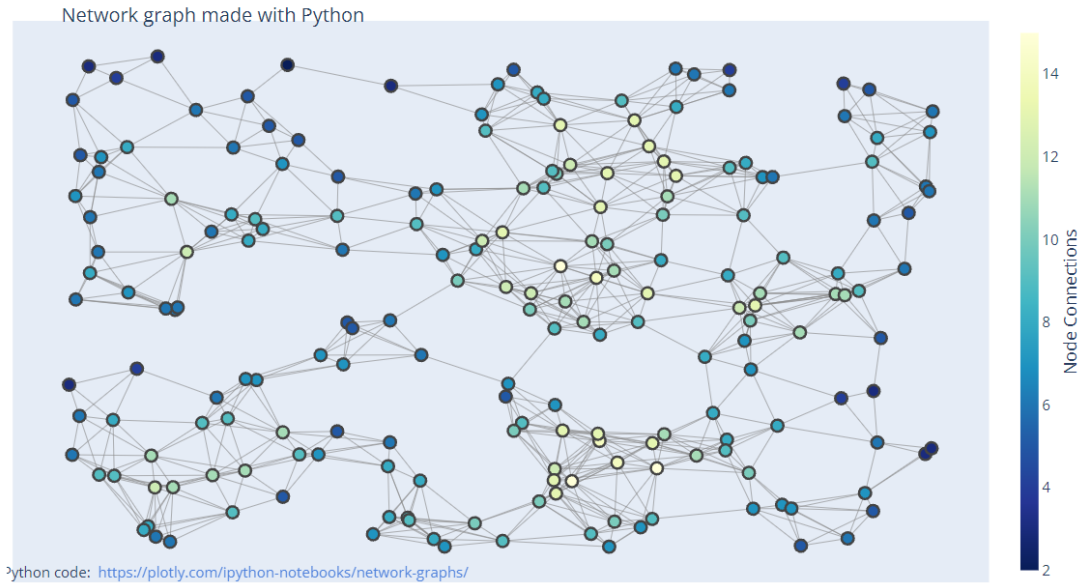


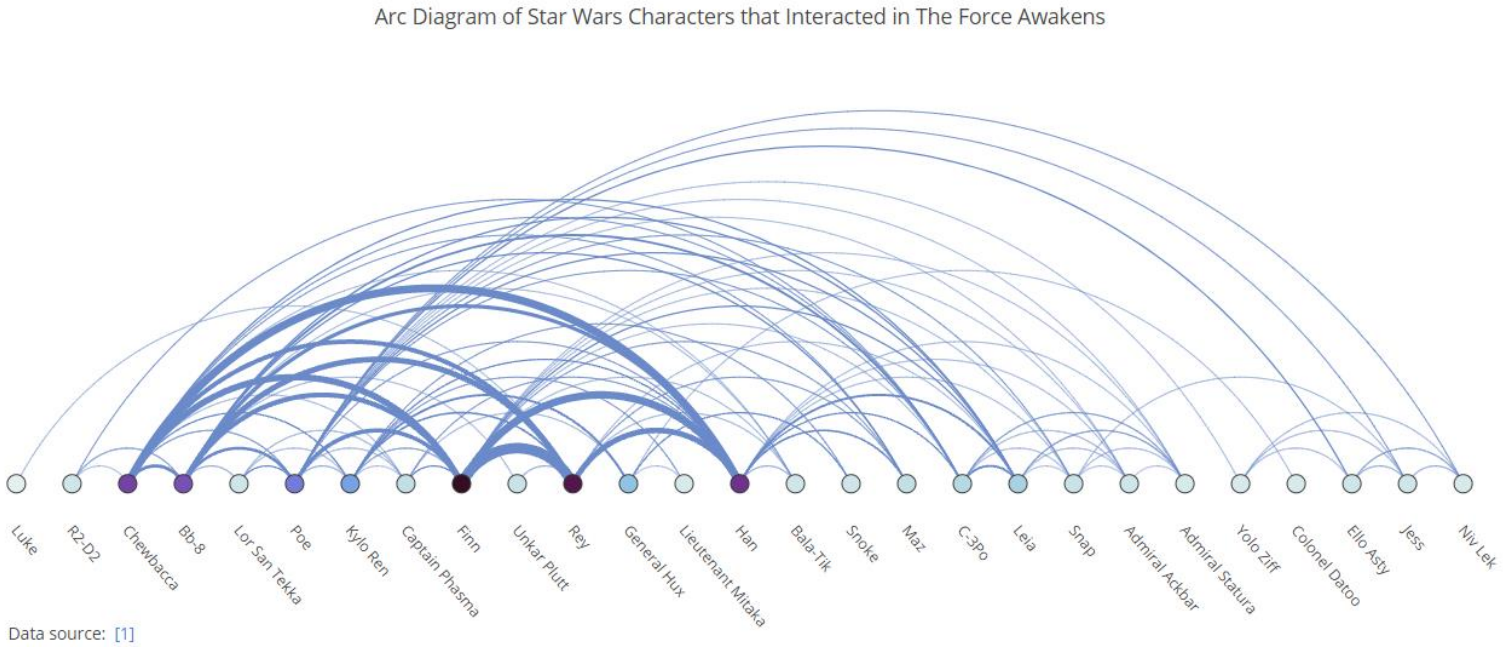


https://plotly.com/python/maps/

Esempi – SOLO immagine

Node-link diagram e Arc diagram



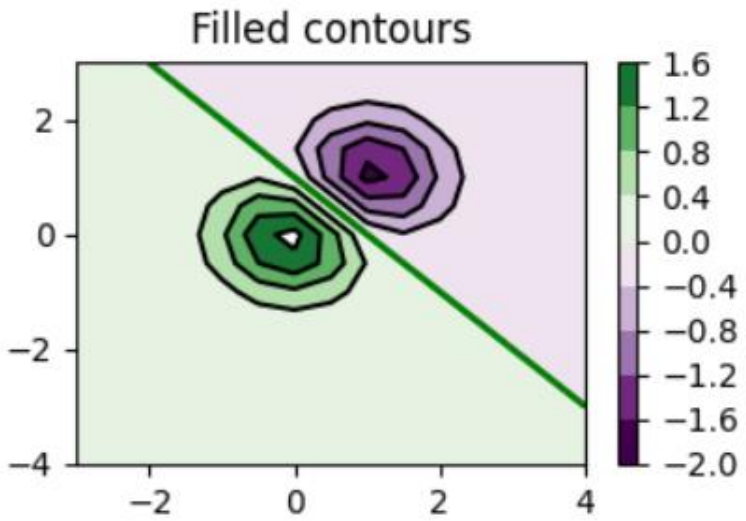


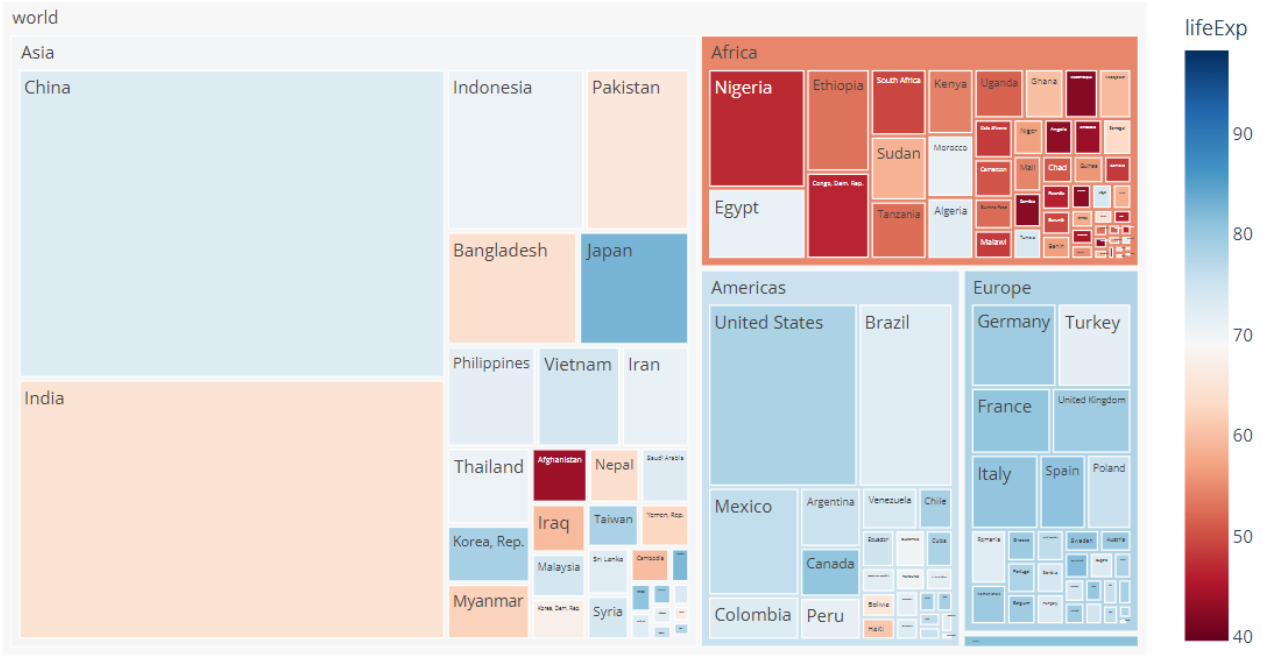
https://plotly.com/python/network-graphs/

https://chart-studio.plotly.com/~empet/13574/arc-diagram-of-star-wars-characters-that-interacted-in-the-force-awakens/#plot

Esempi – SOLO immagine

Isocontour plot e Tree map





https://matplotlib.org/stable/gallery/images_contours_and_fields/contour_image.html#sphx-glr-gallery-images-contours-and-fields-contour-image-py

https://plotly.com/python/treemaps/

Visualizzazione Interattiva e Visualizzazione AR

Alessia Angeli

Studente di dottorato in Data Science and Computation

Dipartimento di Informatica – Scienza e Ingegneria

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Visualizzazione interattiva

L'utente ha la possibilità di:

Cambiare

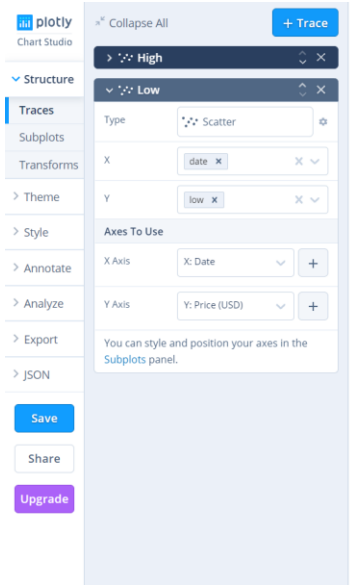
- Codifica (tipo di grafico)
- Parametri (colore, dimensioni, aggiungere elementi al grafico)
- Organizzazione (allineamento colonne/righe)
- Ordine
- Tipo di aggregazione

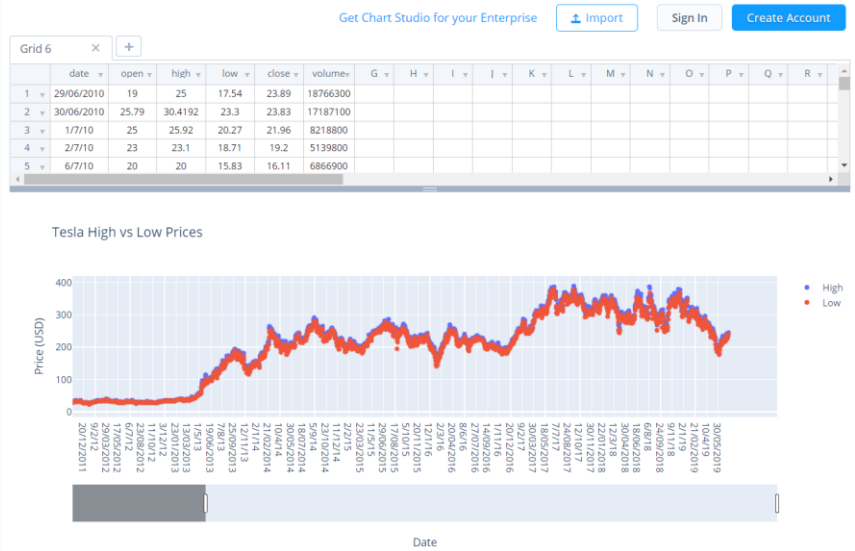
Selezionare

- Items
- Tooltips per ulteriori informazioni

Navigare

- Tra items
- Tra attributi





https://towardsdatascience.com/interactive-data-visualization-167ae26016e8

https://www.tableau.com/learn/articles/interactive-map-and-data-visualization-examples

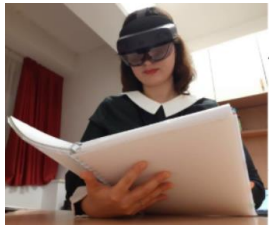
Visualizzazione interattiva - VANTAGGI

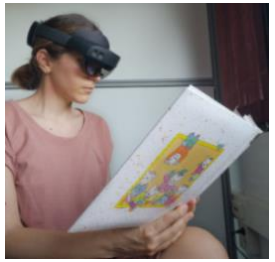
- Tipo di visualizzazione flessibile, potente, intuitivo.
- L'analisi esplorativa dei dati può cambiare durante lo stesso processo di analisi.
- Possibile cambio di attività fluido attraverso codifiche visive diverse a supporto di attività diverse.
- Le transizioni animate possono fornire un supporto eccellente alla visualizzazione.
- C'è un'evidenza empirica che le transizioni animate aiutino le persone a rimanere concentrate.

Visualizzazione interattiva - SVANTAGGI

- L'interazione ha un costo in termini di tempo (variabile, a volte significativo)
- L'interazione impone un carico cognitivo.
- I controlli per l'interazione potrebbero richiedere molto spazio sullo schermo.
- Gli utenti potrebbero non interagire come pianificato dal designer (e.g., i registri del NYTimes mostrano che circa il 90% degli utenti non interagisce oltre lo scrollytelling (Aisch, 2016)).

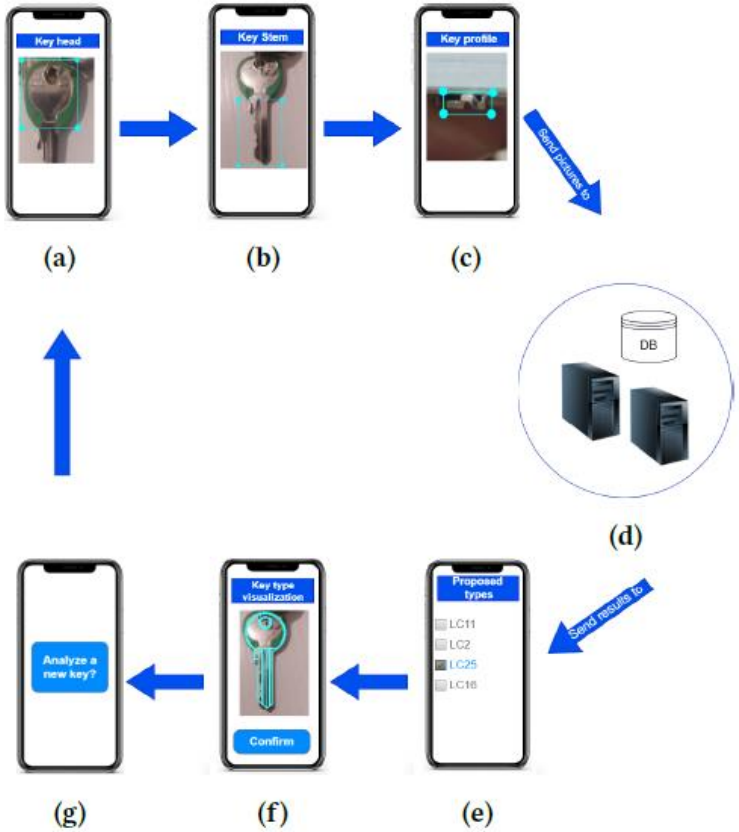
Visualizzazione e Augmented Reality (AR)

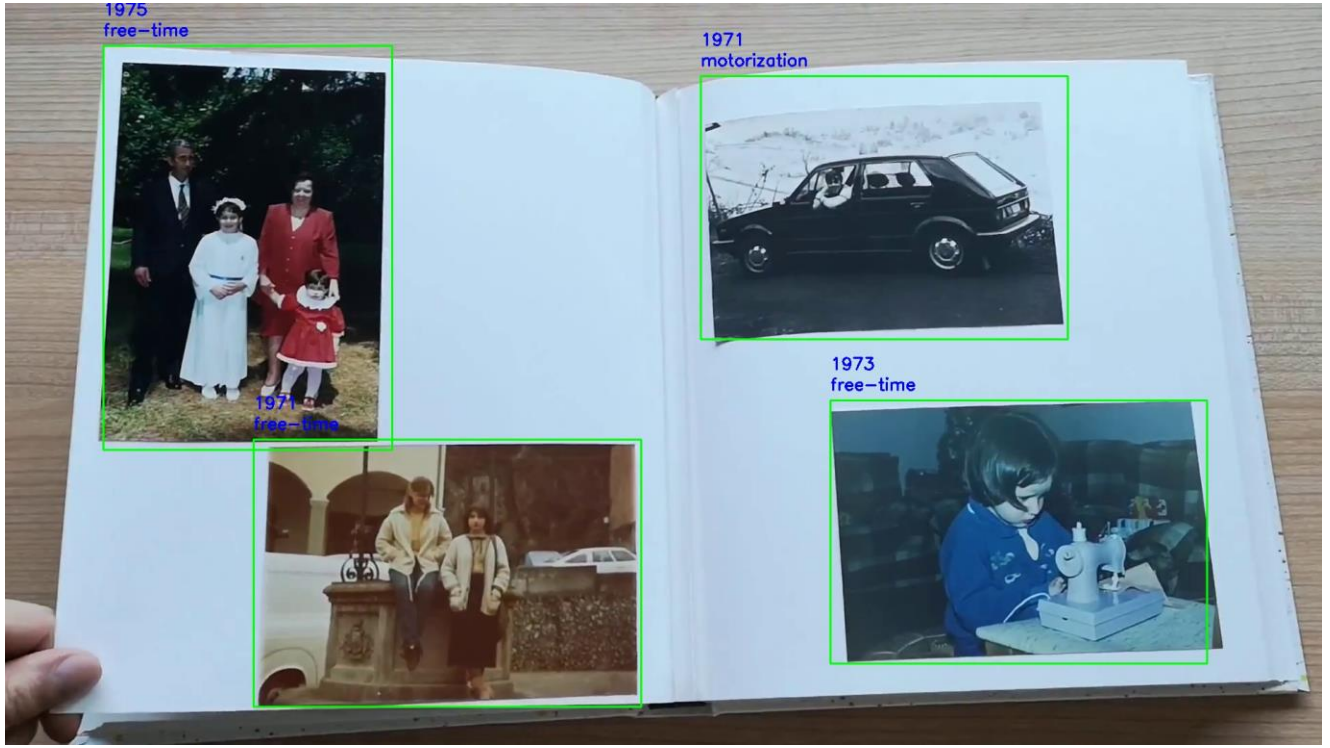




https://www.microsoft.com/it-it/hololens

Visualizzazione e Augmenter Reality (AR) – alcuni nostri progetti





Stacchio, L., Angeli, A., Hajahmadi, S., Marfia, G. (2021). Revive Family Photo Albums through a Collaborative Environment Exploiting the HoloLens 2. In Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), to appear.

Stacchio, L., Hajahmadi, S., & Marfia, G. (2021, March). Preserving Family Album Photos with the HoloLens 2. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 643-644). IEEE.

Stacchio, L., Angeli, A., & Marfia, G. (2021, September). Empowering Locksmith Crafts via Mobile Augmented Reality. In *Proceedings of the Conference on Information Technology for Social Good* (pp. 305-308).

Solo una parte…

Alessia Angeli

Studente di dottorato in Data Science and Computation

Dipartimento di Informatica – Scienza e Ingegneria

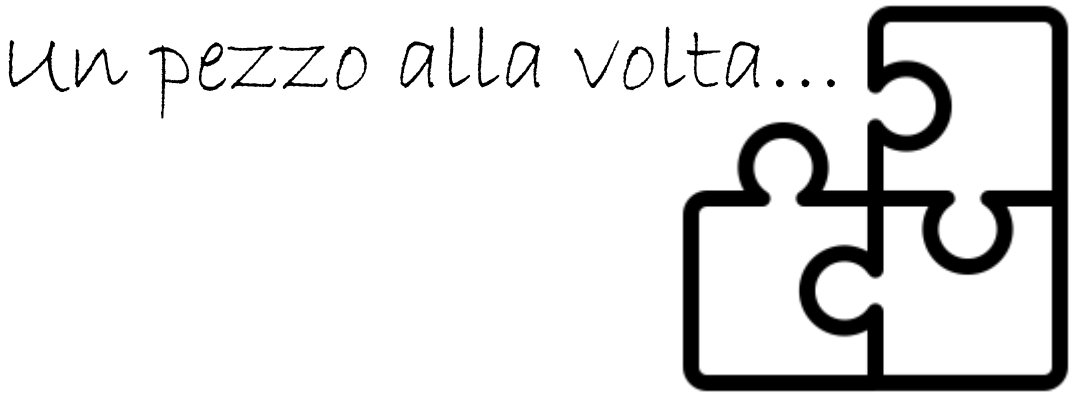
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Solo una parte…

La visualizzazione dati non è solo quello visto in queste lezioni, è molto di più…

Una parte fondamentale è analizzare/selezionare i dati/risultati da mostrare. Spesso questa parte è una delle più complesse sia per analizzare i dati con l'obiettivo di costruire un modello (**feature selection** e **feature extraction**) sia con l'obiettivo di visualizzare dati e/o risultati (**analisi esplorativa**).

Ricordando poi, come già detto in precedenza, che come prima operazione i dati vanno controllati: dati mancanti, dati oggettivamente non corretti (e.g., età di uomini con valori negativi, prezzi di articoli con valori negativi), …



Data Visualization – Quindi…

Alessia Angeli

Studente di dottorato in Data Science and Computation

Dipartimento di Informatica – Scienza e Ingegneria

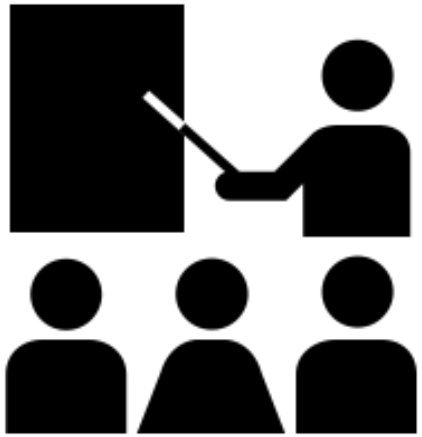
Visualizzazione efficace?

Come abbiamo visto, ci sono molti modi per definire l'efficacia ma… quello che conta di più è il fatto che i vostri utenti o lettori siano in grado di **estrarre informazioni dai dati visualizzati**:

- In modo accurato;
- Con uno sforzo ragionevole;
- Con grande sicurezza.

Il test è l'**acquisizione di conoscenza**:

- Se i vostri utenti non imparano niente di nuovo… c'è qualcosa che non va!
- L'acquisizione di conoscenza deve essere il vostro metro di valutazione.



GUAI IN VISTA… Quando?

Il problema

- Non capire di che cosa hanno bisogno gli utenti.

Dal problema alla visualizzazione

- Mostrare agli utenti le cose sbagliate.

Il tipo di visualizzazione

- La strada che si ha intrapreso per visualizzare i dati non funziona.

L'algoritmo

- Il codice è lento.

Quindi… PER CERCARE DI NON FINIRE NEI GUAI!

Il problema

- Capire chi sono i vostri utenti (e.g., studenti scuole superiori, operai, impiegati);
- Capire quale problema devono risolvere con (i loro) dati;
- Raccogliere una serie di obiettivi e/o azioni che gli utenti vorrebbero raggiungere/eseguire con (i loro) dati.

Dal problema alla visualizzazione

- Dopo avere compreso il problema, capire che cosa gli utenti devono visualizzare e per quali attività;
- Se si fallisce qui significa che non si sta mostrando le cose giuste.

Il tipo di visualizzazione

- Pianificare visualizzazioni efficaci per i dati e per le attività identificate;
- È qui che si deve sapere come scartare alternative scadenti e ideare buoni progetti.

L'algoritmo

- Assicurarsi che l'algoritmo sia veloce, accurato ed efficace.

Un consiglio

«Sbagliando si impara...» (?) – in Data Visualization sì!

La maggior parte delle codifiche non è subito ottimale e si continua ad imparare a crearne di migliori fallendo. Più fallisci, più la tua codifica migliora. Non smettere di ripetere. Non accontentarti della prima visualizzazione che ti viene in mente! Genera e confronta alternative.

Riferimenti

Corso «Data and Results Visualization», Daniele Loiacono, Politecnico di Milano (2019).

Corso Coursera (online) "Visualization for Data Journalism", Margaret Yee Man Ng, Univeristy of Illinois at Urbana-Champaign (2021).

Sitografia presente slide per slide.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Alessia Angeli

Dipartimento di Informatica – Scienza e Ingegneria

alessia.angeli2@unibo.it