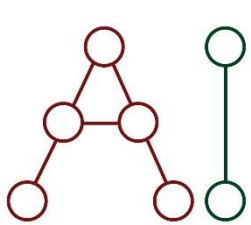


Exploratory Data Analysis

Assoc. Prof. Karl Ezra Pilario, Ph.D.
Process Systems Engineering Laboratory
Department of Chemical Engineering
University of the Philippines Diliman

Outline

- Exploratory Data Analysis
  - Introduction
  - Titanic Survival Data Set
  - Iris Flower Data Set
  - Taylor Swift Spotify Data Set

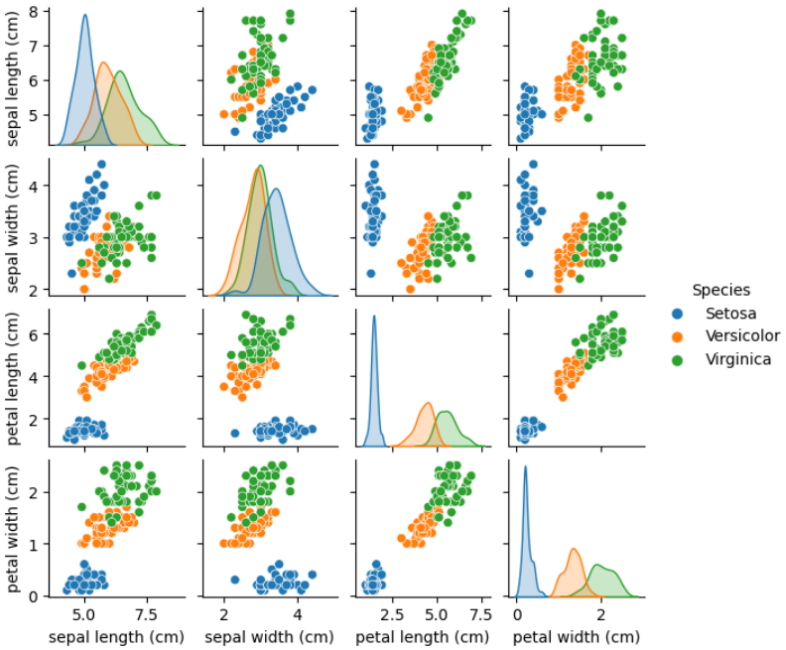
Exploratory Data Analysis

- Tools to analyze and investigate data sets and summarize their main characteristics.
- Often employs **data visualization** methods.
- Helps determine how best to manipulate data sources to get the answers you need, making it easier to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

Source: https://www.ibm.com/topics/exploratory-data-analysis

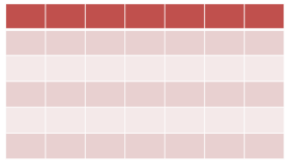
Example:

A correlogram visualization of the **Iris Flower Data Set**



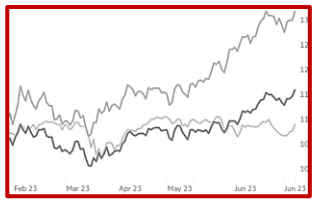
What are the different modalities of data?

Tabular Data

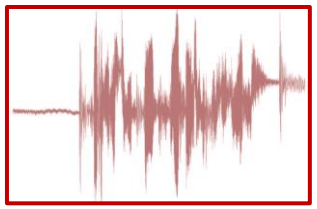


Text Data

Time Series Data

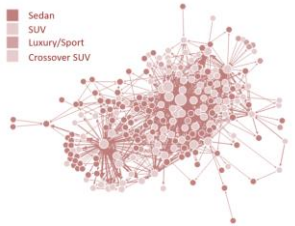


Audio / Speech Data



Images / Videos Data

Graph Data



- Data sets can be heterogeneous, i.e., different types of data are contained in one set.
- Multi-modal learning = learning from a combination of various modalities of data (speech, text, image, etc.).

Exploratory Data Analysis

For tabular data, different data types can exist in one table.

Example: Titanic Survival Data Set

Contains information on 1309 passengers aboard the Titanic and whether they survived or not. Goal: To predict the survival of passengers based on their attributes.

Attributes

Passenger ID	An identifier unique to a passenger
Survived	1 – survived, 0 – did not survive
Pclass	1, 2, 3 – travel class
Name	Passenger's name
Sex	Male / Female
Age	Passenger's age
SibSp	Number of siblings and spouses aboard
Parch	Number of parents and children aboard
Ticket	Ticket number
Fare	Amount paid for ticket
Cabin	Cabin of residence
Embarked	Q = Queenstown, C = Cherbourg, S = Southampton

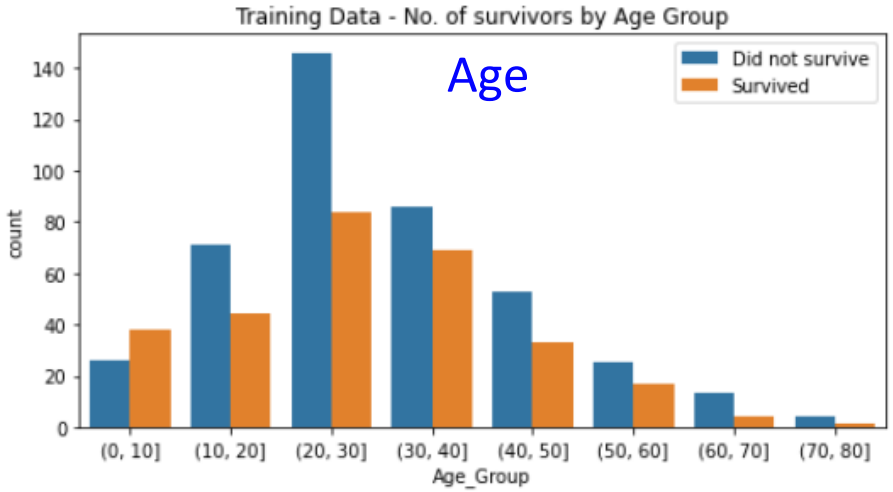
Integer, Ordinal
Integer, Binary
Integer, Categorical
String
String, Categorical
Continuous
Integer, Non-negative
String
Continuous
String
String, Categorical

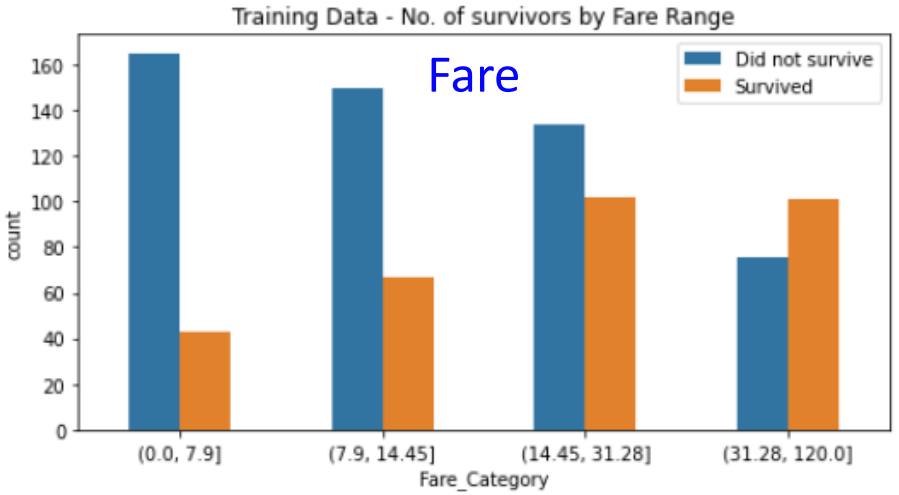
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

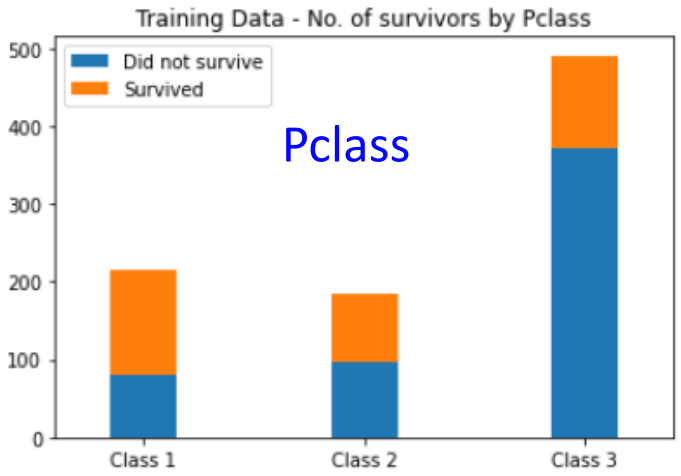
Example: Titanic Survival Data Set

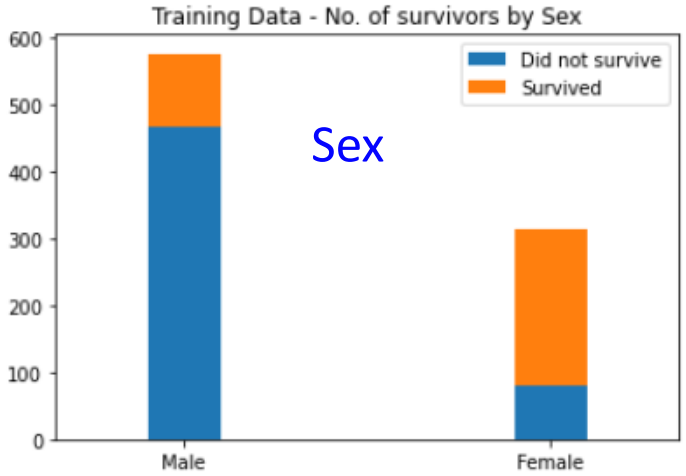
Contains information on 1309 passengers aboard the Titanic and whether they survived or not.
Goal: To predict the survival of passengers based on their attributes.

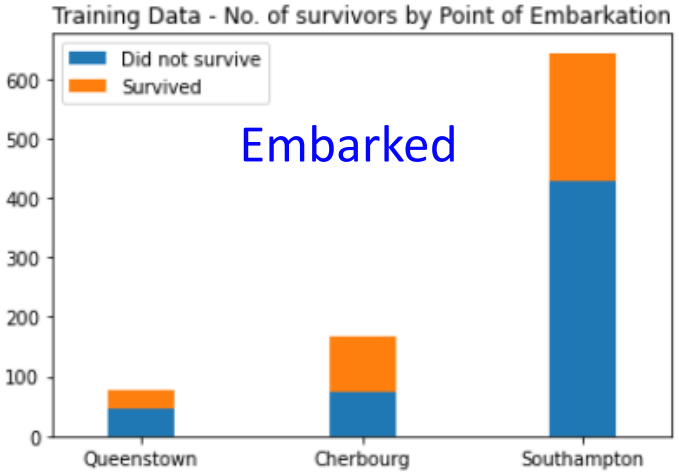
Bar plots and **histograms** are useful for visualizing the "count" of values in the data set.











Example: Titanic Survival Data Set

Before training any classifier, only the relevant numerical (**Num**) and categorical (**Cat**) columns should be retained. Other columns can be dropped.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
		Label	**Cat**		**Cat**	**Num**	**Num**	**Num**		**Num**		**Cat**

- This is called feature selection.

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

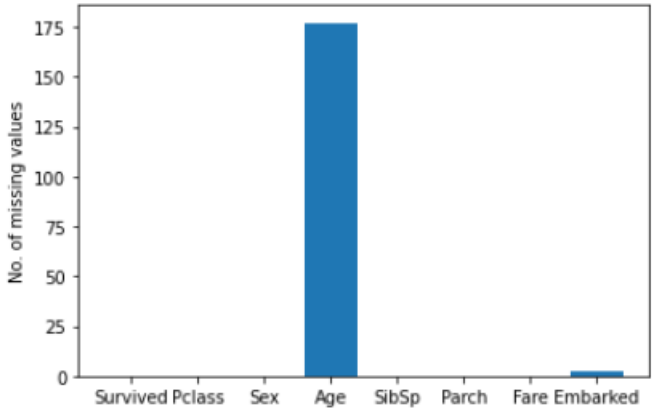
Example: Titanic Survival Data Set

Here is our current data set:

	Label	Cat	Cat	Num	Num	Num	Num	Cat
	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

Here is the current count of missing values:

- "Age" has **177** missing values.
- "Embarked" has **2** missing values.

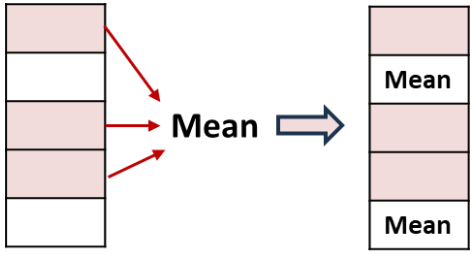


How to deal with missing values?

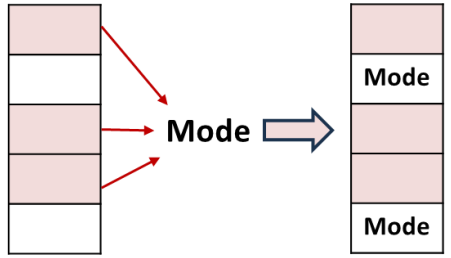
Two options:

- Remove rows with missing values; or,
- **Imputation** = the act of filling in missing values by estimating them.

Mean imputation



Most-frequent imputation



Other imputers: median imputer, iterative imputer, kNN imputer, etc.

Example: Titanic Survival Data Set

We set up one last preprocessing task before ML training: **Column Transformation**

Before

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
857	1	male	51.000000	0	0	26.5500	S
52	1	female	49.000000	1	0	76.7292	C
386	3	male	1.000000	5	2	46.9000	S
124	1	male	54.000000	0	1	77.2875	S
578	3	female	29.699118	1	0	14.4583	C
...	...	...	...	...	...	...	...
835	1	female	39.000000	1	1	83.1583	C
192	3	female	19.000000	1	0	7.8542	S
629	3	male	29.699118	0	0	7.7333	Q
559	3	female	36.000000	1	0	17.4000	S
684	2	male	60.000000	1	1	39.0000	S

After

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0.0	1.0	1.623937	0.0	0.0	-0.122530	2.0
1	0.0	0.0	1.470203	1.0	0.0	0.918124	0.0
2	2.0	1.0	-2.219399	5.0	2.0	0.299503	2.0
3	0.0	1.0	1.854537	0.0	1.0	0.929702	2.0
4	2.0	0.0	-0.013392	1.0	0.0	-0.373297	0.0
...	...	...	...	...	...	...	...
618	0.0	0.0	0.701536	1.0	1.0	1.051455	0.0
619	2.0	0.0	-0.835798	1.0	0.0	-0.510258	2.0
620	2.0	1.0	-0.013392	0.0	0.0	-0.512765	1.0
621	2.0	0.0	0.470936	1.0	0.0	-0.312290	2.0
622	1.0	1.0	2.315737	1.0	1.0	0.135667	2.0

For Numerical Data:

- "Age", "Fare"
- *Standard Scaling*

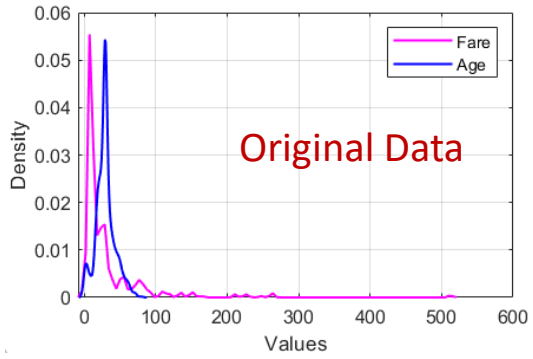
For Categorical Data:

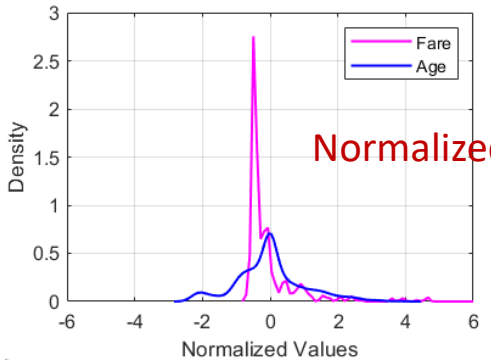
- "Pclass", "Sex", "Embarked"
- *Ordinal Encoding*

- Other features are just retained.

Data Normalization

- Normalization removes the effect of differing scales and biases.
- All data are centered to *zero-mean* and scaled to *unit-variance*.





Mean $\boldsymbol{\mu} = \frac{1}{N}\sum \boldsymbol{x}_i$

Population standard deviation $\boldsymbol{\sigma}_P = \sqrt{\frac{\sum(\boldsymbol{x}_i - \boldsymbol{\mu})^2}{N}}$

Sample standard deviation $\boldsymbol{\sigma}_S = \sqrt{\frac{\sum(\boldsymbol{x}_i - \boldsymbol{\mu})^2}{N-1}}$

Data Standardization (Standard scaler)

$$\boldsymbol{x}_i' = \frac{\boldsymbol{x}_i - \boldsymbol{\mu}}{\boldsymbol{\sigma}_P}$$

Min-max scaler

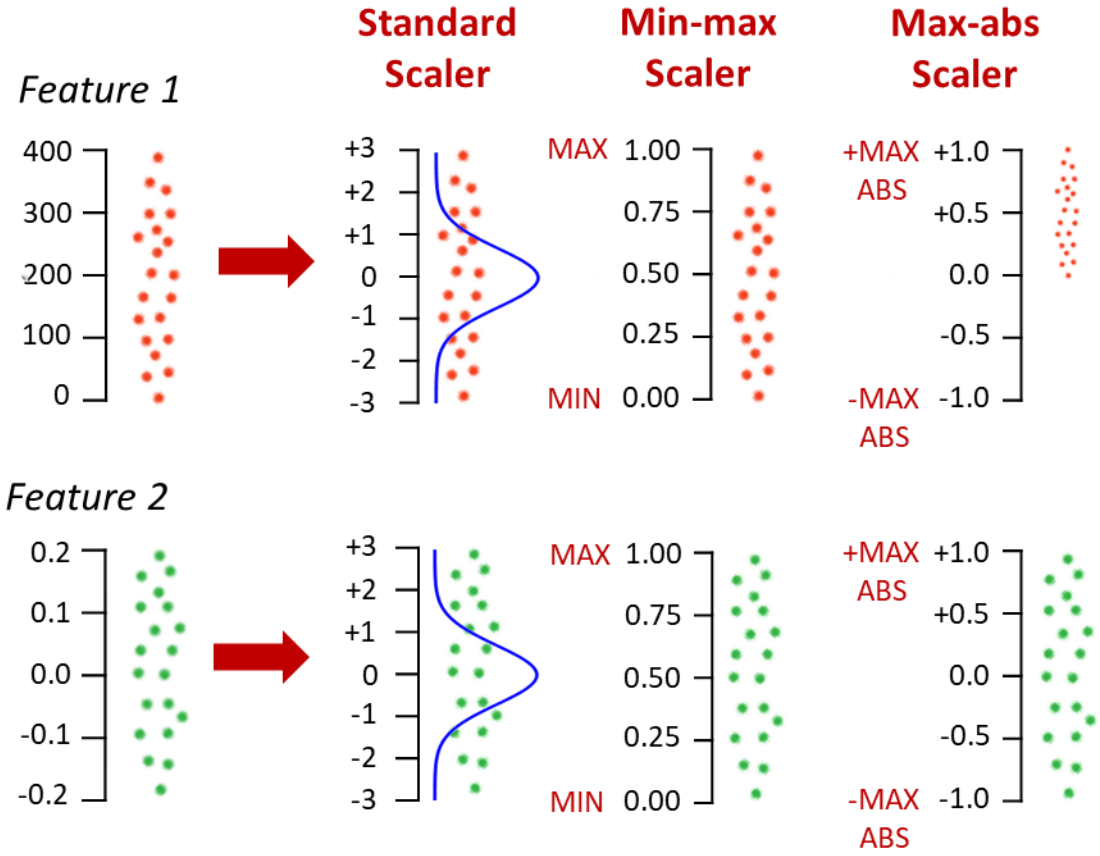
$$\boldsymbol{x}_i' = \frac{\boldsymbol{x}_i - \min \boldsymbol{x}_i}{\max \boldsymbol{x}_i - \min \boldsymbol{x}_i}$$

Max-abs scaler

$$\boldsymbol{x}_i' = \frac{\boldsymbol{x}_i}{\max |\boldsymbol{x}_i|}$$

What happens to the data points after normalization?

- The scatter of data is preserved. *Information is intact!*
- Normalization improves machine learning by treating *all features equally.*



Example: Titanic Survival Data Set

We now set up the **ML pipeline**:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0.0	1.0	1.623937	0.0	0.0	-0.122530	2.0
1	0.0	0.0	1.470203	1.0	0.0	0.918124	0.0
2	2.0	1.0	-2.219399	5.0	2.0	0.299503	2.0
3	0.0	1.0	1.854537	0.0	1.0	0.929702	2.0
4	2.0	0.0	-0.013392	1.0	0.0	-0.373297	0.0
...	...	...	...	...	...	...	...
618	0.0	0.0	0.701536	1.0	1.0	1.051455	0.0
619	2.0	0.0	-0.835798	1.0	0.0	-0.510258	2.0
620	2.0	1.0	-0.013392	0.0	0.0	-0.512765	1.0
621	2.0	0.0	0.470936	1.0	0.0	-0.312290	2.0
622	1.0	1.0	2.315737	1.0	1.0	0.135667	2.0

Randomly split the instances into a Training Data Set and a Testing Data Set.

Perform data preprocessing

Perform model training and testing

70%

Training Data

Fit and **Transform** the data using the Column Transform

Train the ML classifier

Fitted Column Transform

Trained model

30%

Test Data

Transform the data using the Column Transform

Test the trained ML classifier

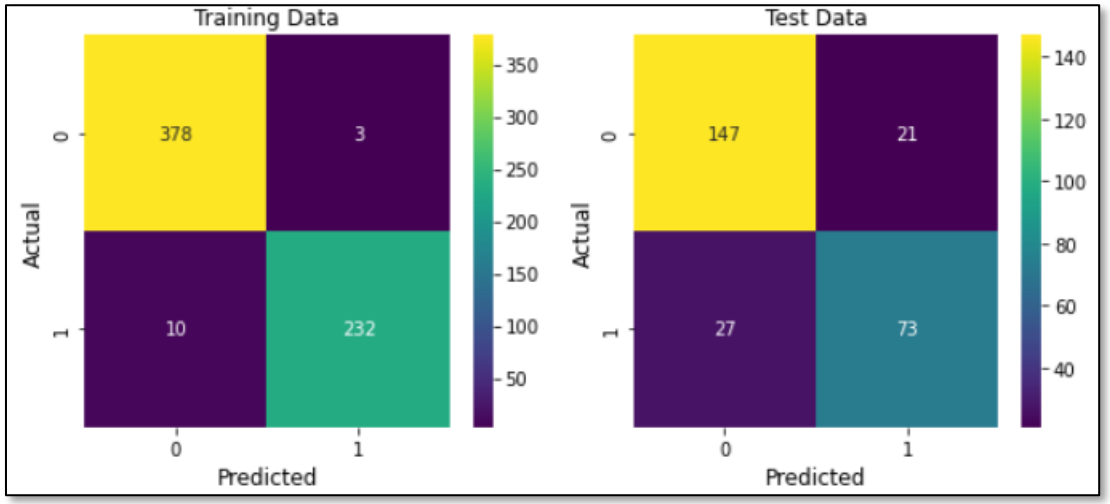
Report performance metrics

Training Accuracy

97.9%

Test Accuracy

82.1%



Outline

- Exploratory Data Analysis
  - Introduction
  - Titanic Survival Data Set
  - Iris Flower Data Set
  - Taylor Swift Spotify Data Set

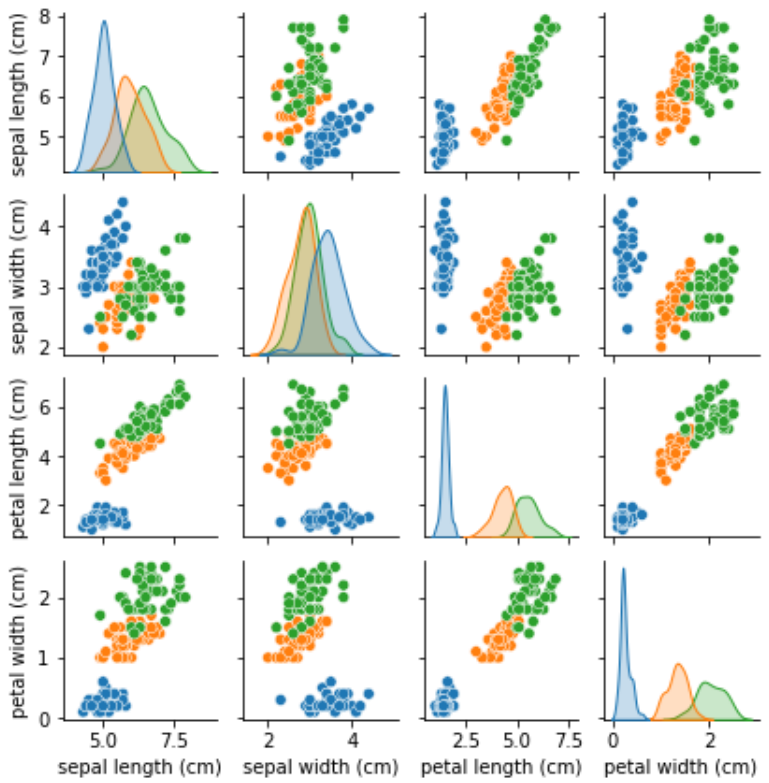
Example: Iris Flower Data Set

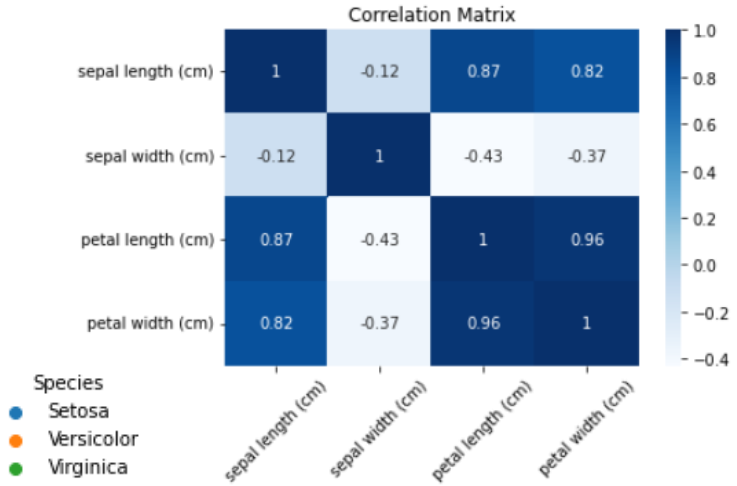
The data set contains measurements of 150 iris flowers in terms of their sepal length, sepal width, petal length, and petal width. There are 3 species of flowers, Setosa, Versicolor, and Virginica, with 50 samples each.

Versicolor Setosa Virginica

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	Species
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Virginica
146	6.3	2.5	5.0	1.9	Virginica
147	6.5	3.0	5.2	2.0	Virginica
148	6.2	3.4	5.4	2.3	Virginica
149	5.9	3.0	5.1	1.8	Virginica

Pair plots (or **correlograms / correlation matrices**) are useful for finding correlated features.





If two features are positively correlated:

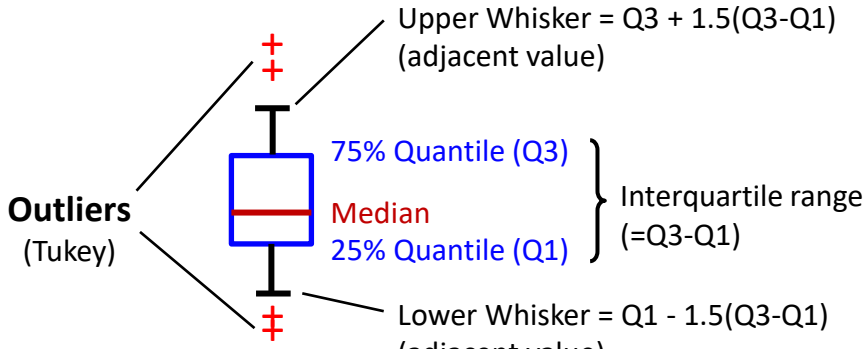
- When one is high, the other is also high
- When one is low, the other is also low

Example: Iris Flower Data Set

The data set contains measurements of 150 iris flowers in terms of their sepal length, sepal width, petal length, and petal width. There are 3 species of flowers, Setosa, Versicolor, and Virginica, with 50 samples each.

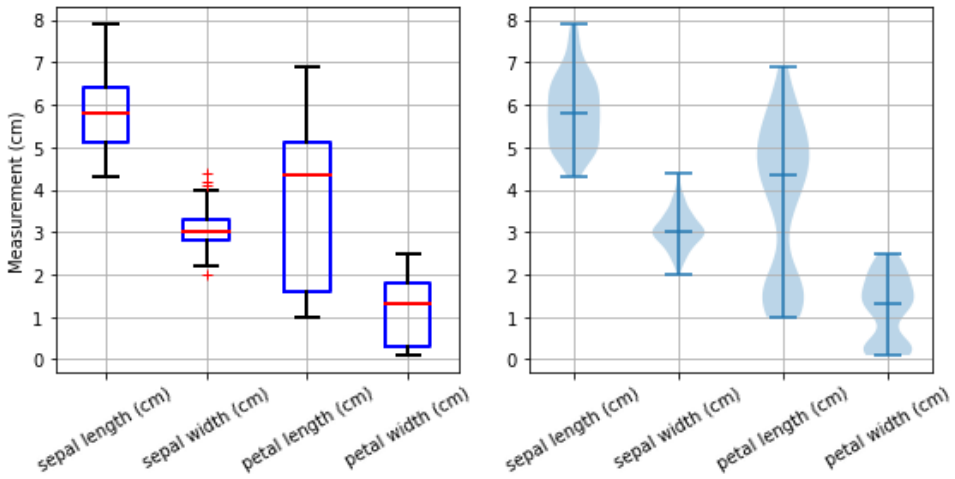
Box-and-whisker plots (or **box plots**) and **Violin plots** are useful for visualizing the distributions of values.

What is a box plot?

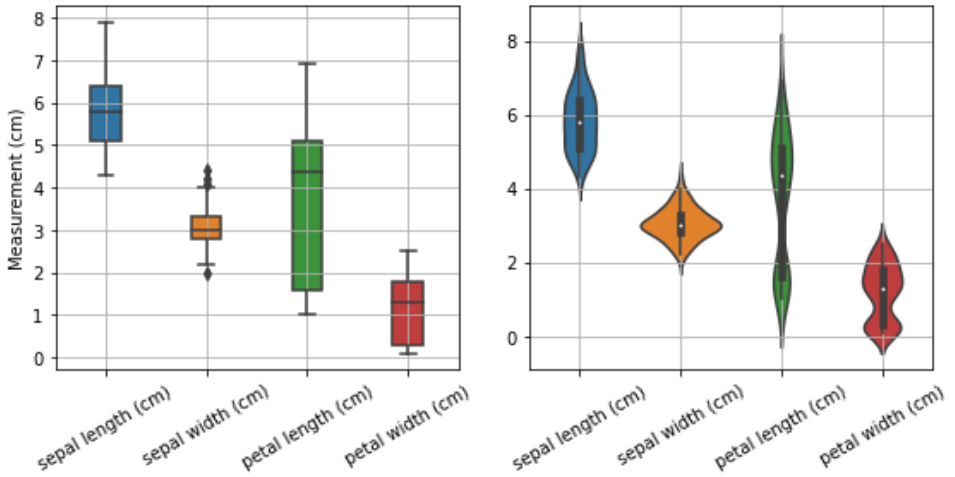


Box plot **Violin plot**

Matplotlib



Seaborn

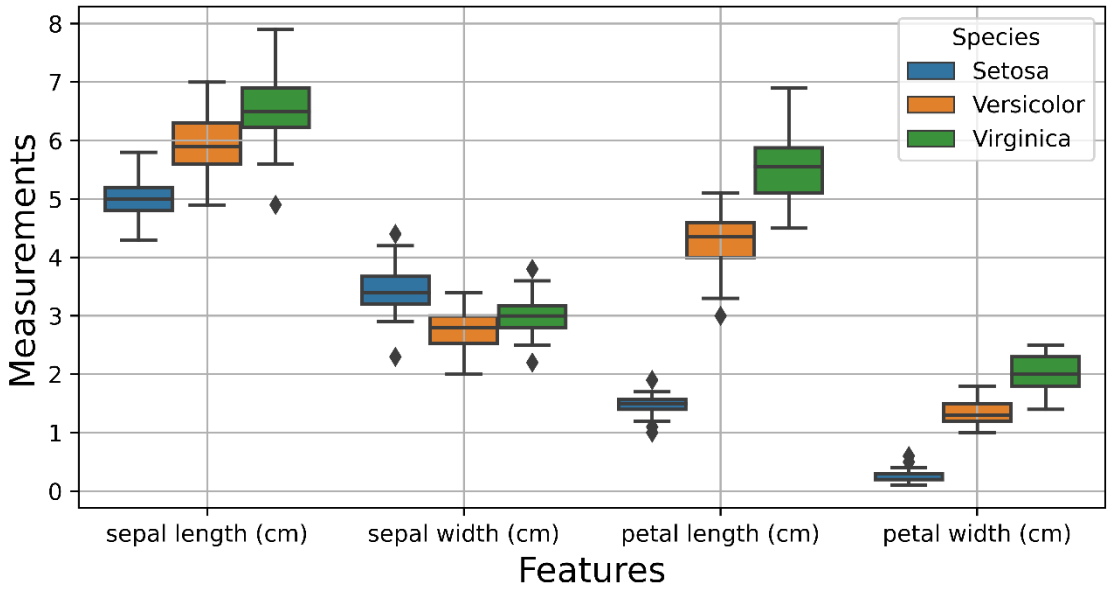


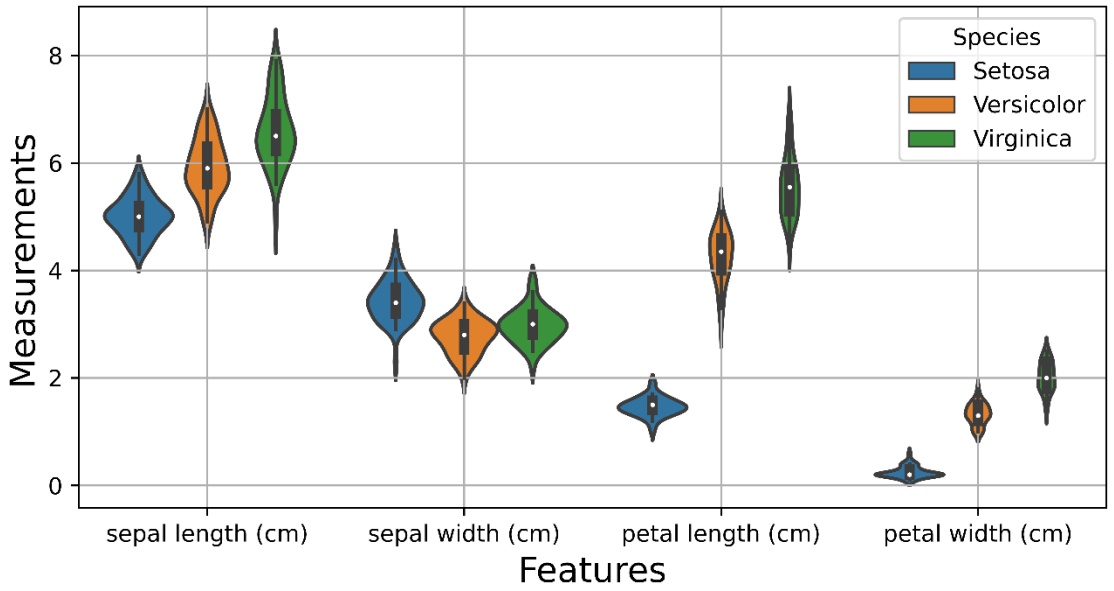
Example: Iris Flower Data Set

The data set contains measurements of 150 iris flowers in terms of their sepal length, sepal width, petal length, and petal width. There are 3 species of flowers, Setosa, Versicolor, and Virginica, with 50 samples each.

Box plots can be **grouped** according to each feature, then visualized per class (or any categorical variable).

Distributions of measurements per feature per species

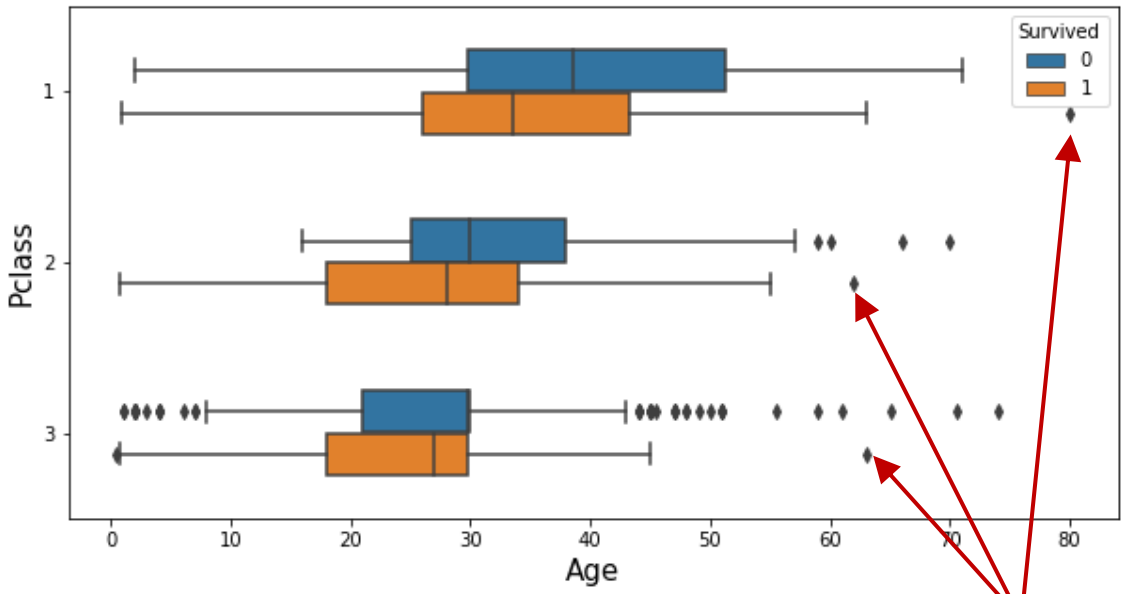




Example: Titanic Survival Data Set

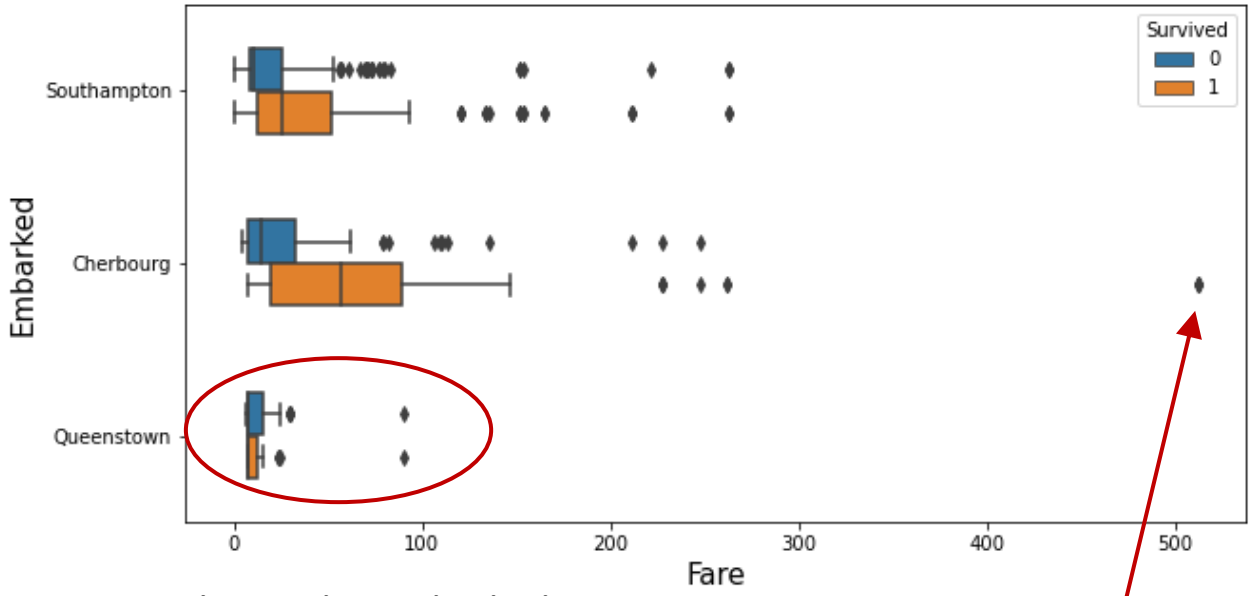
Grouped boxplots can also be applied to the Titanic Data Set.

Distributions of Age by Pclass and Survived



Oldest survivors at the time

Distributions of Fare by Embarked and Survived



Those who embarked in Queenstown generally paid less.

The one who paid the most for a ticket survived.

Outline

- Exploratory Data Analysis
  - Introduction
  - Titanic Survival Data Set
  - Iris Flower Data Set
  - Taylor Swift Spotify Data Set

Example: Taylor Swift Spotify Data Set

- Contains information on Taylor Swift's songs from the Spotify API.
- Data set is updated **monthly**: https://www.kaggle.com/datasets/jarredpriester/taylor-swift-spotify-dataset
- Shown here is the data downloaded on 18 September 2023. Total no. of instances: 487
- Typical goal: Predict **Popularity**.

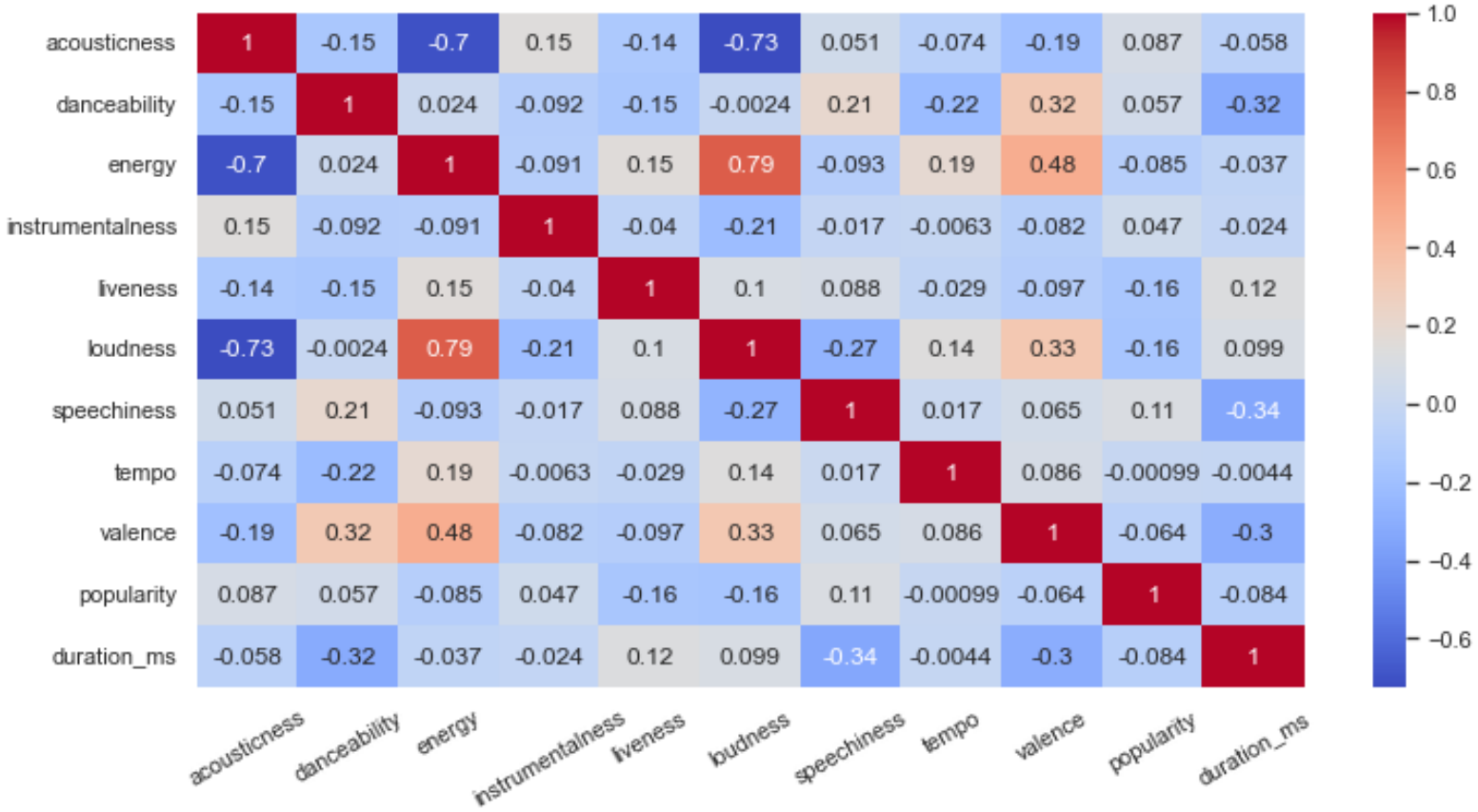
Name	Song name
Album	Album name
Release Date	YYYY-MM-DD
Track Number	Order of the song in the album where it appeared
Id	Spotify id for the song
Uri	Spotify uri for the song
Acousticness	[0, 1] where 1 = most acoustic
Danceability	[0, 1] where 1 = most danceable
Energy	[0, 1] where 1 = most energetic
Instrumentalness	[0, 1] where 1 = no vocals, mostly instrumental
Liveness	[0, 1] where 1 = most likely has a live audience
Loudness	Average loudness in decibels
Speechiness	[0, 1] where 1 = mostly speech, like audiobooks
Tempo	Average speed in beats per minute (BPM)
Valence	[0, 1] where 1 = most positive, cheerful, or happy
Popularity	[0, 100] No. of recent streams relative to other artists
Duration_ms	Duration of the track in milliseconds

name	album	release_date	track_number	id	uri	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity	duration_ms
Mine (Taylor's Version)	Speak Now (Taylor's Version)	2023-07-07	1	7G0gBu6nLdhFDPRLc0HdDG	spotify:track:7G0gBu6nLdhFDPRLc0HdDG	0.00444	0.646	0.783	0.000001	0.1710	-2.846	0.0356	121.080	0.490	87	231706
Sparks Fly (Taylor's Version)	Speak Now (Taylor's Version)	2023-07-07	2	3MytWN8L7shNYzGl4tAKRp	spotify:track:3MytWN8L7shNYzGl4tAKRp	0.02510	0.588	0.758	0.000000	0.1310	-2.347	0.0305	114.991	0.387	84	261230
Back To December (Taylor's Version)	Speak Now (Taylor's Version)	2023-07-07						0.635	0.000000	0.1170	-3.927	0.0273	142.063	0.203	88	294189
Speak Now (Taylor's Version)	Speak Now (Taylor's Version)	2023-07-07						0.677	0.000000	0.1430	-2.871	0.0325	118.995	0.639	84	242473
Dear John (Taylor's Version)	Speak Now (Taylor's Version)	2023-07-07						0.470	0.000000	0.1630	-5.016	0.0296	119.294	0.133	84	405906

Example: Taylor Swift Spotify Data Set

Correlation matrix

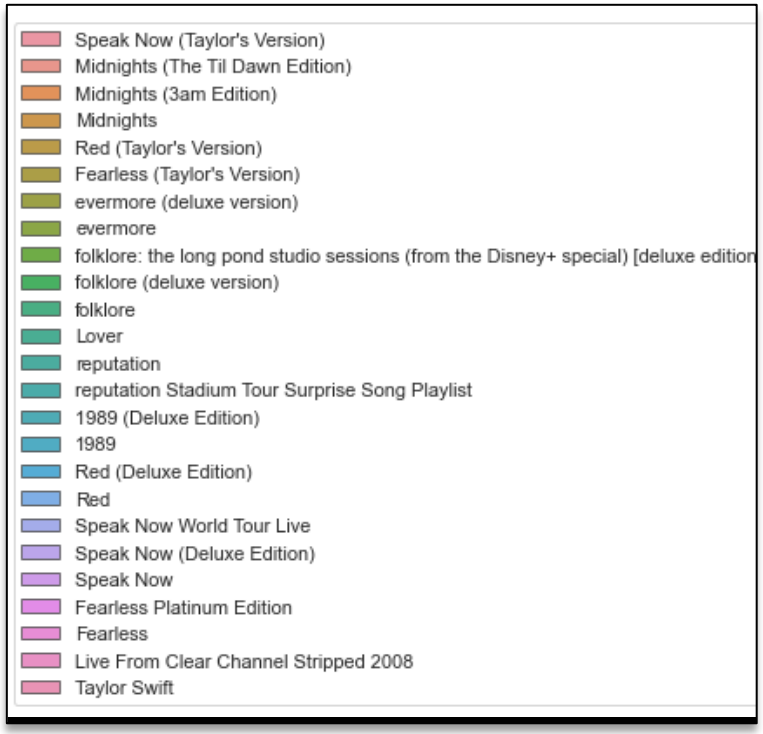
- **Popularity** has little linear correlation with other numerical features.
- **Energy** and **loudness** are the two most positively correlated features (0.79).
- **Acousticness** and **loudness** are the two most negatively correlated features (-0.71).
- **Valence** is positively correlated with **energy** (0.47), which means that happy Taylor Swift songs are also energetic.



Example: Taylor Swift Spotify Data Set

Box plots

- All features that have a range of values within [0, 1] are plotted as box plots **per album**.
- Legend:



Acousticness

Danceability

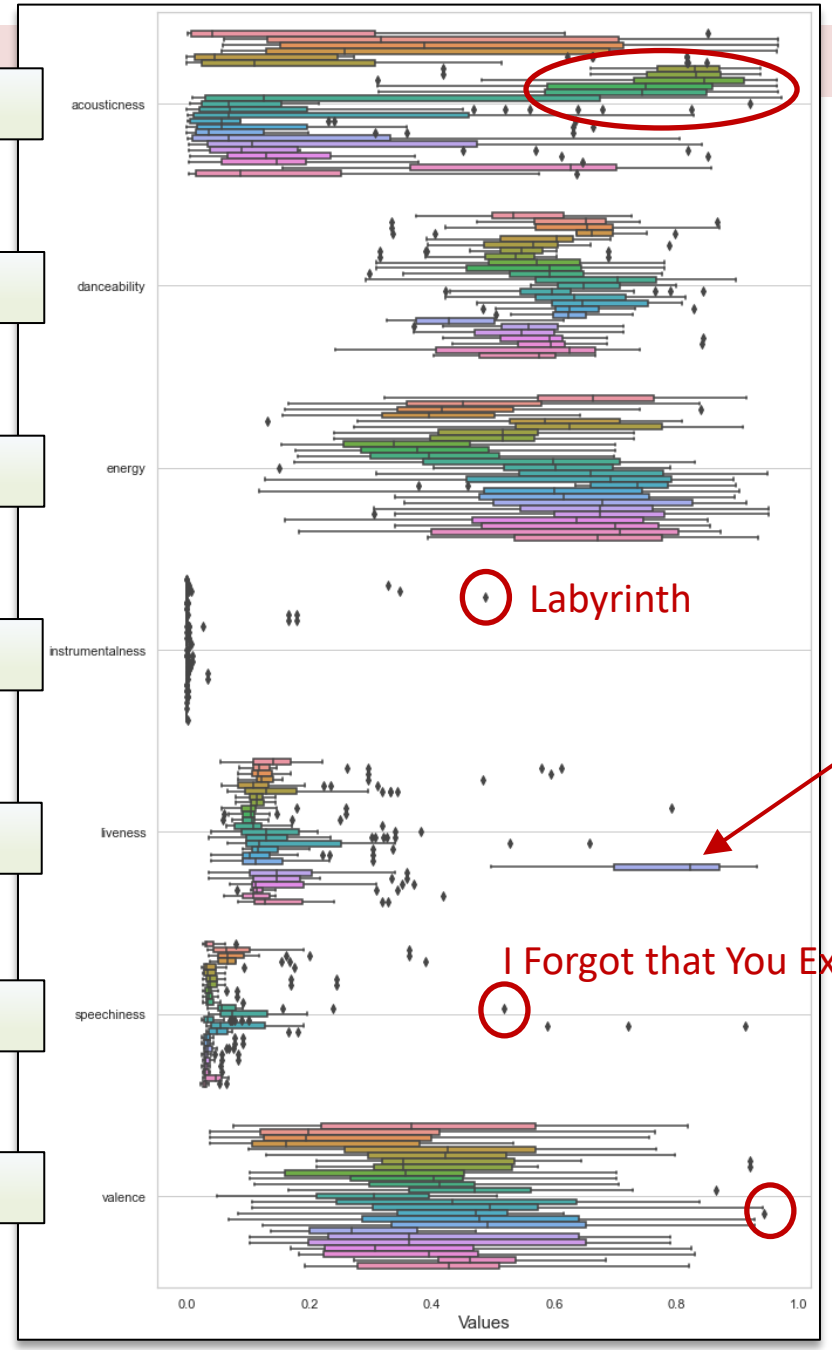
Energy

Instrumentalness

Liveness

Speechiness

Valence



Evermore (album)
Folklore (album)

Labyrinth

Speak Now
(World Tour Live)

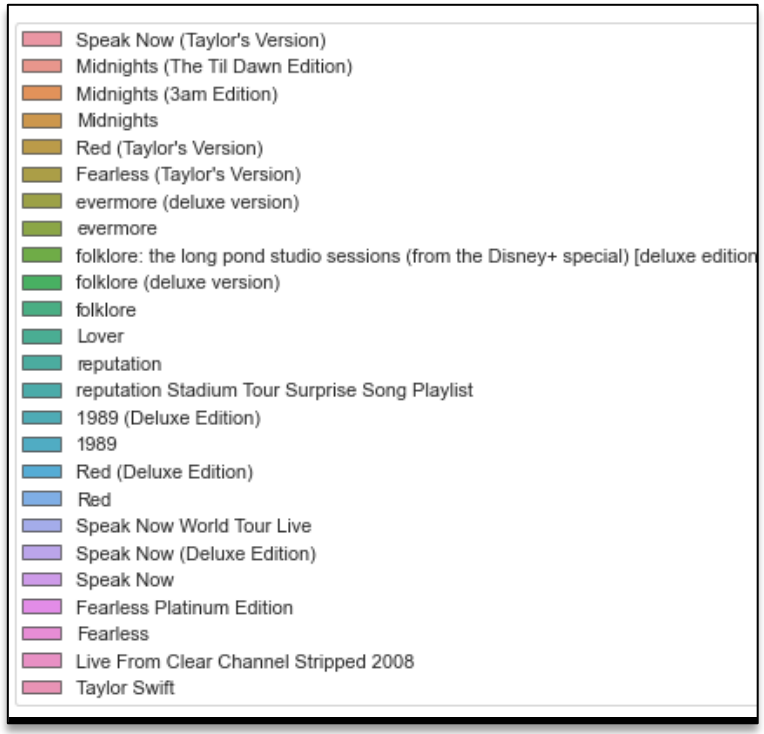
I Forgot that You Existed

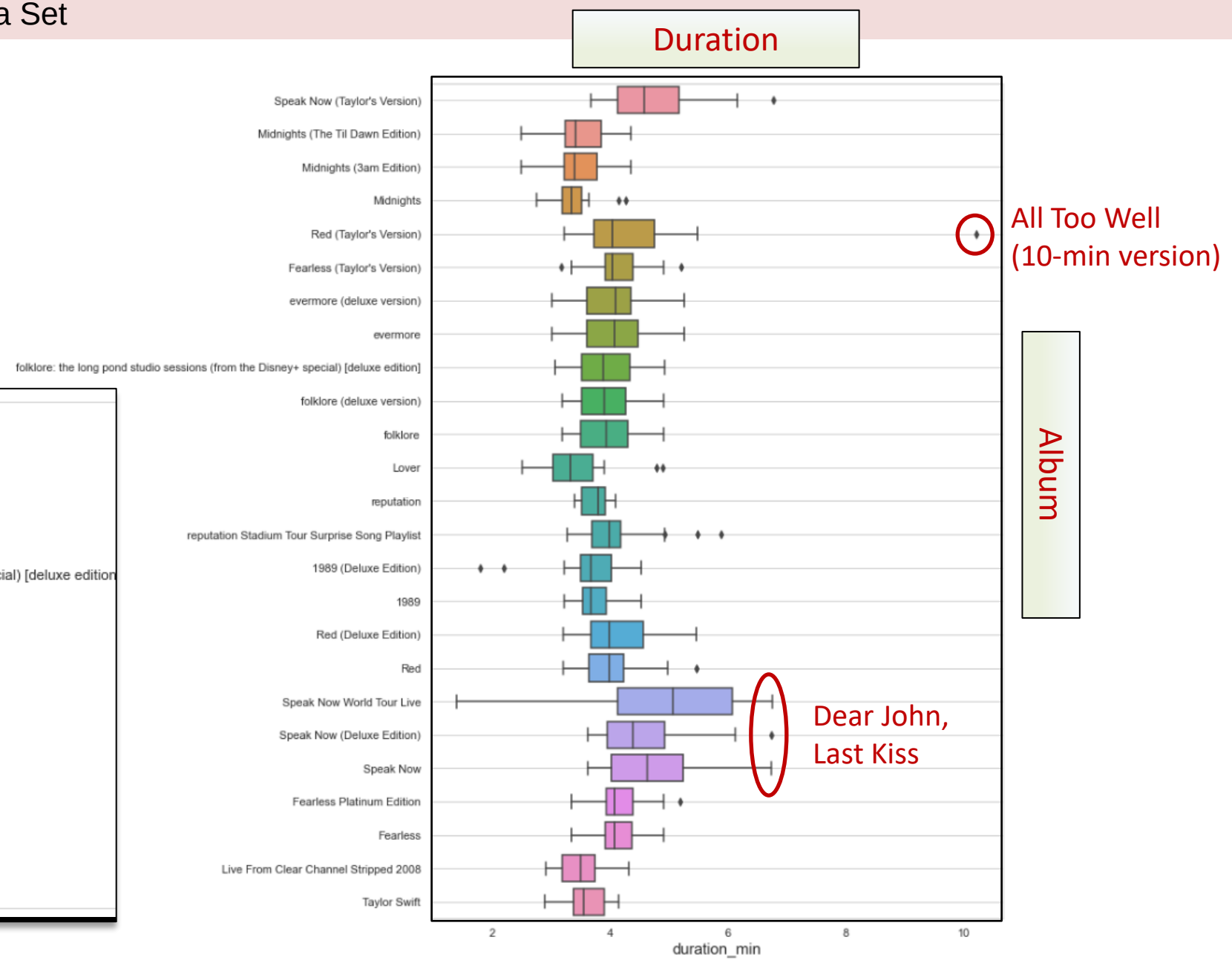
Shake It off

Example: Taylor Swift Spotify Data Set

Box plots

- Distributions of track durations plotted as box plots **per album**.
- Legend:

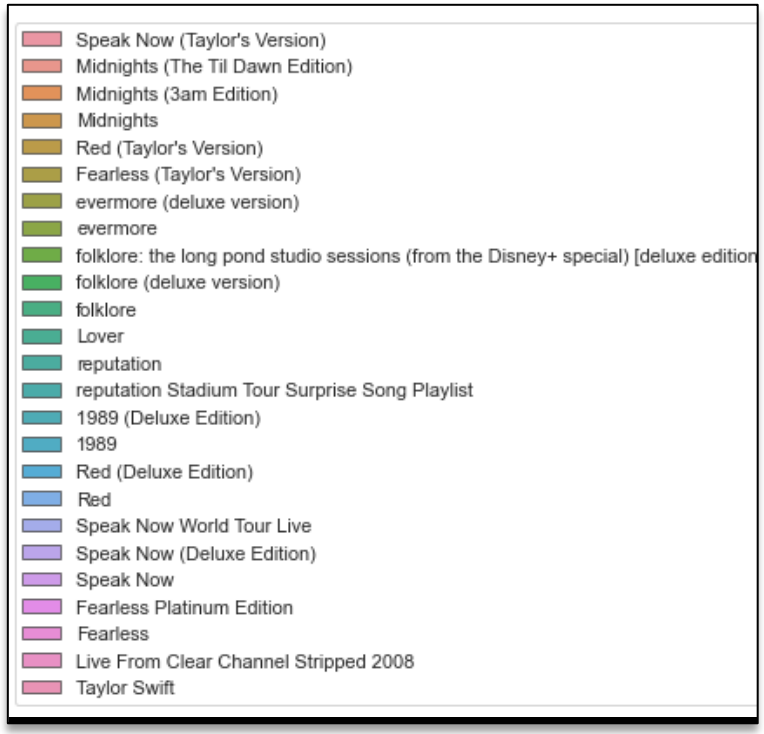


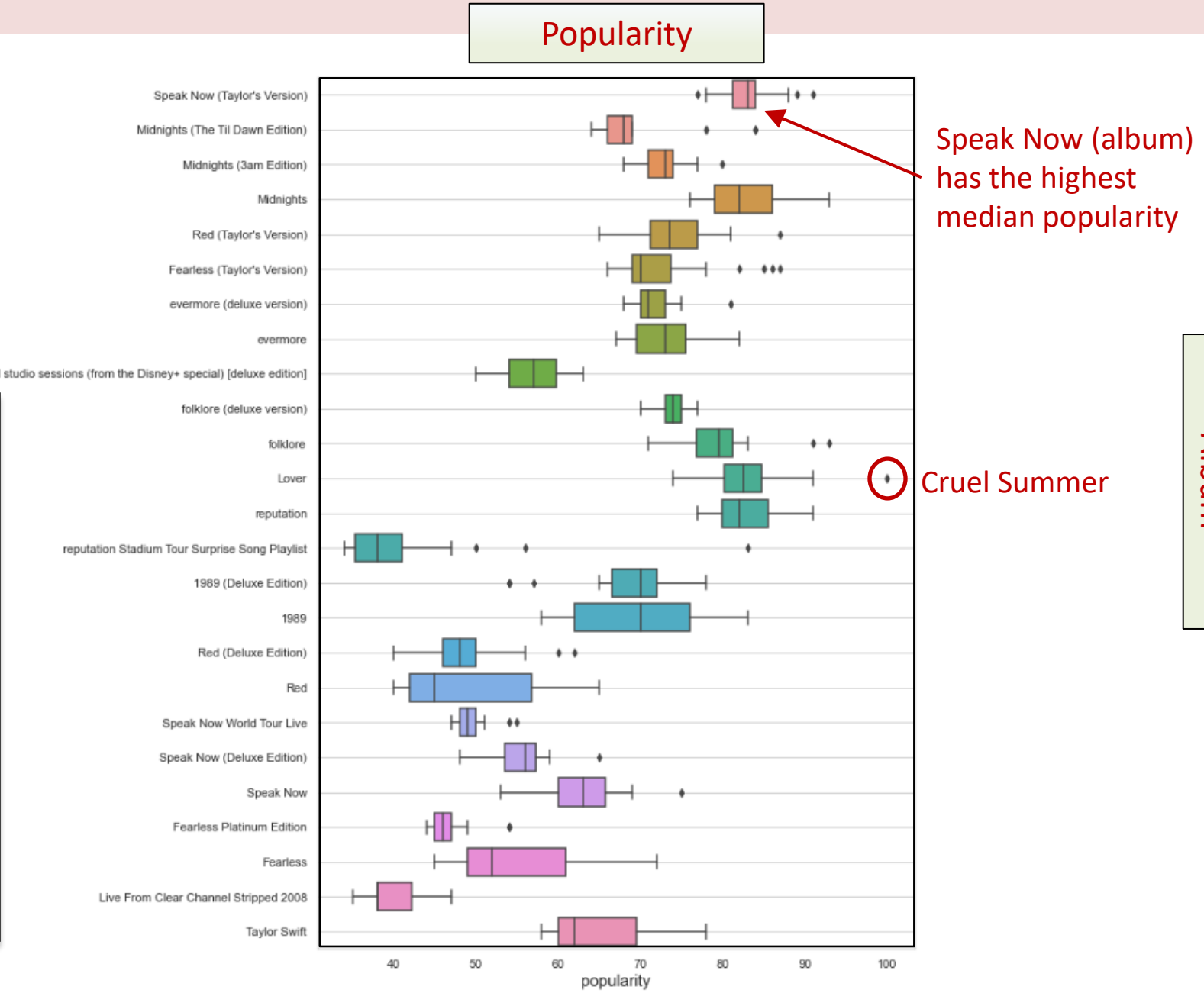


Example: Taylor Swift Spotify Data Set

Box plots

- Distributions of popularity plotted as box plots **per album**.
- Legend:





Example: Taylor Swift Spotify Data Set

Another way to explore the data is to reveal **rankings**.

Most popular tracks

	name	popularity
234	Cruel Summer	100
67	Anti-Hero	93
224	august	93
235	Lover	91
218	cardigan	91
...	...	...
285	Treacherous	35
297	Invisible	35
290	Change	34
300	Jump Then Fall	34
305	Hey Stephen	34

Longest tracks

	name	duration_ms
107	All Too Well (10 Minute Version) (Taylor's Ver...	613026
4	Dear John (Taylor's Version)	405906
394	Dear John - Live/2011	404680
422	Dear John	403920
402	Dear John	403920
...	...	...
62	Glitch	148781
39	Glitch	148781
330	Blank Space - Voice Memo	131186
329	I Wish You Would - Voice Memo	107133
393	I Want You Back - Live/2011	83253

Most acoustic tracks

	name	acousticness
249	It's Nice To Have A Friend	0.971000
33	Sweet Nothing	0.967000
232	hoax	0.966000
198	hoax - the long pond studio sessions	0.962000
187	my tears ricochet - the long pond studio sessions	0.946000
...	...	...
384	The Story Of Us - Live	0.000480
83	22 (Taylor's Version)	0.000443
78	State Of Grace (Taylor's Version)	0.000328
344	State Of Grace	0.000197
120	Change (Taylor's Version)	0.000191

Most danceable tracks

	name	danceability
238	I Think He Knows	0.897
52	Vigilante Shit	0.870
305	Hey Stephen	0.843
363	Treacherous - Original Demo Recording	0.828
241	Cornelia Street	0.824
...	...	...
155	tolerate it	0.316
182	the lakes - bonus track	0.313
169	exile (feat. Bon Iver)	0.310
237	The Archer	0.292
471	Change - Live From Clear Channel Stripped 2008	0.243

Most energetic tracks

	name	energy
409	Haunted	0.950
483	I'm Only Me When I'm With You	0.934
407	Better Than Revenge	0.917
11	Haunted (Taylor's Version)	0.915
385	Mean - Live/2011	0.915
...	...	...
198	hoax - the long pond studio sessions	0.155
265	New Year's Day	0.151
97	State Of Grace (Acoustic Version) (Taylor's Ve...	0.131
328	I Know Places - Voice Memo	0.128
365	State Of Grace - Acoustic	0.118

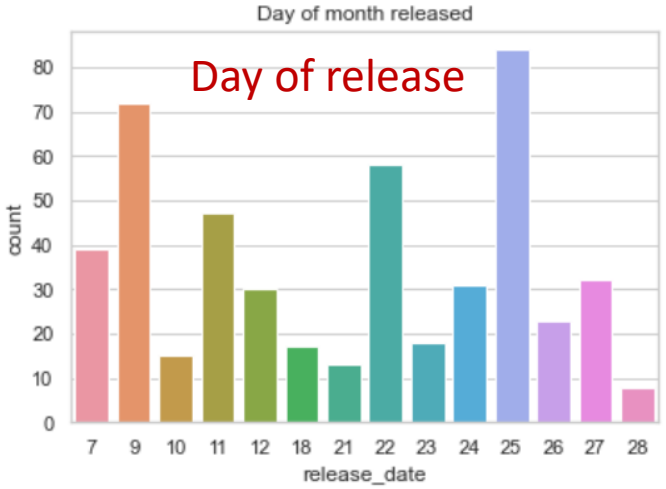
Happiest tracks

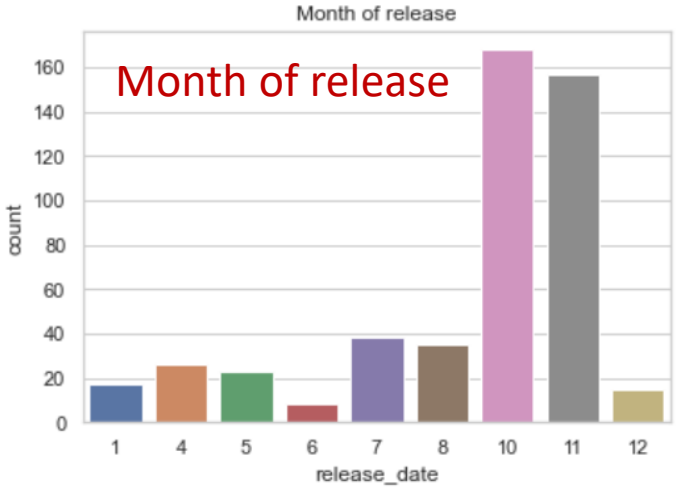
	name	valence
336	Shake It Off	0.9430
374	Stay Stay Stay	0.9280
164	closure	0.9200
240	Paper Rings	0.8650
305	Hey Stephen	0.8380
...	...	...
44	Karma (feat. Ice Spice)	0.0734
365	State Of Grace - Acoustic	0.0682
59	Bigger Than The Whole Sky	0.0680
255	Delicate	0.0499
23	Maroon	0.0382

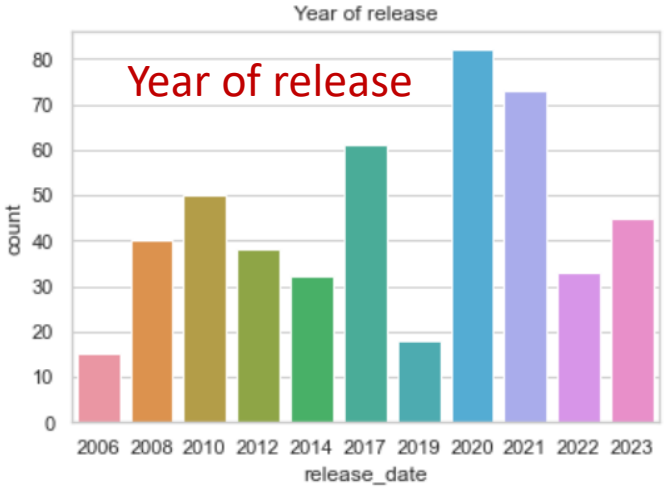
Example: Taylor Swift Spotify Data Set

Using the **date-and-time** columns, we can:

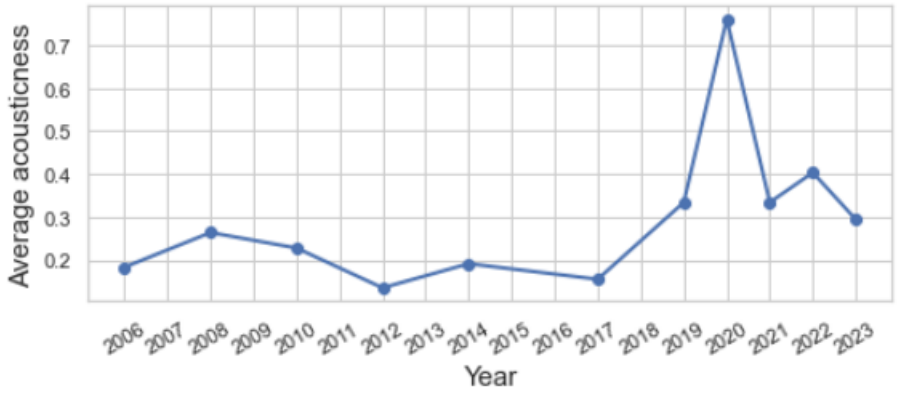
- Investigate the frequency of track release per day, month, year
- Track the evolution of feature values with time



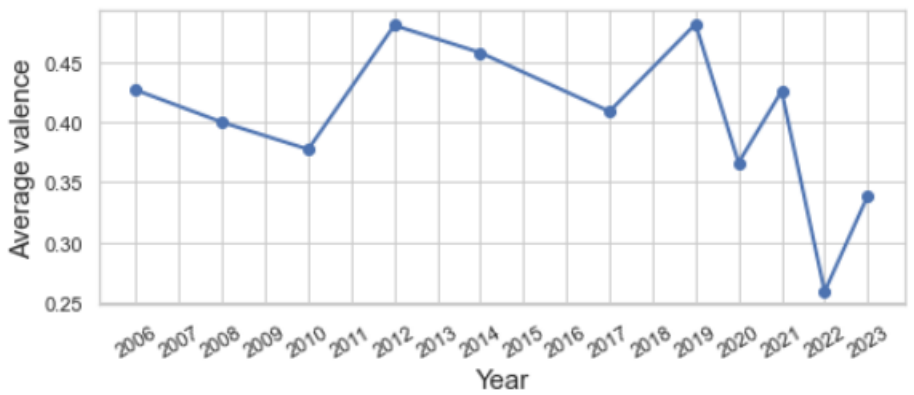




Average Acousticness vs. Release Year



Average Valence vs. Release Year



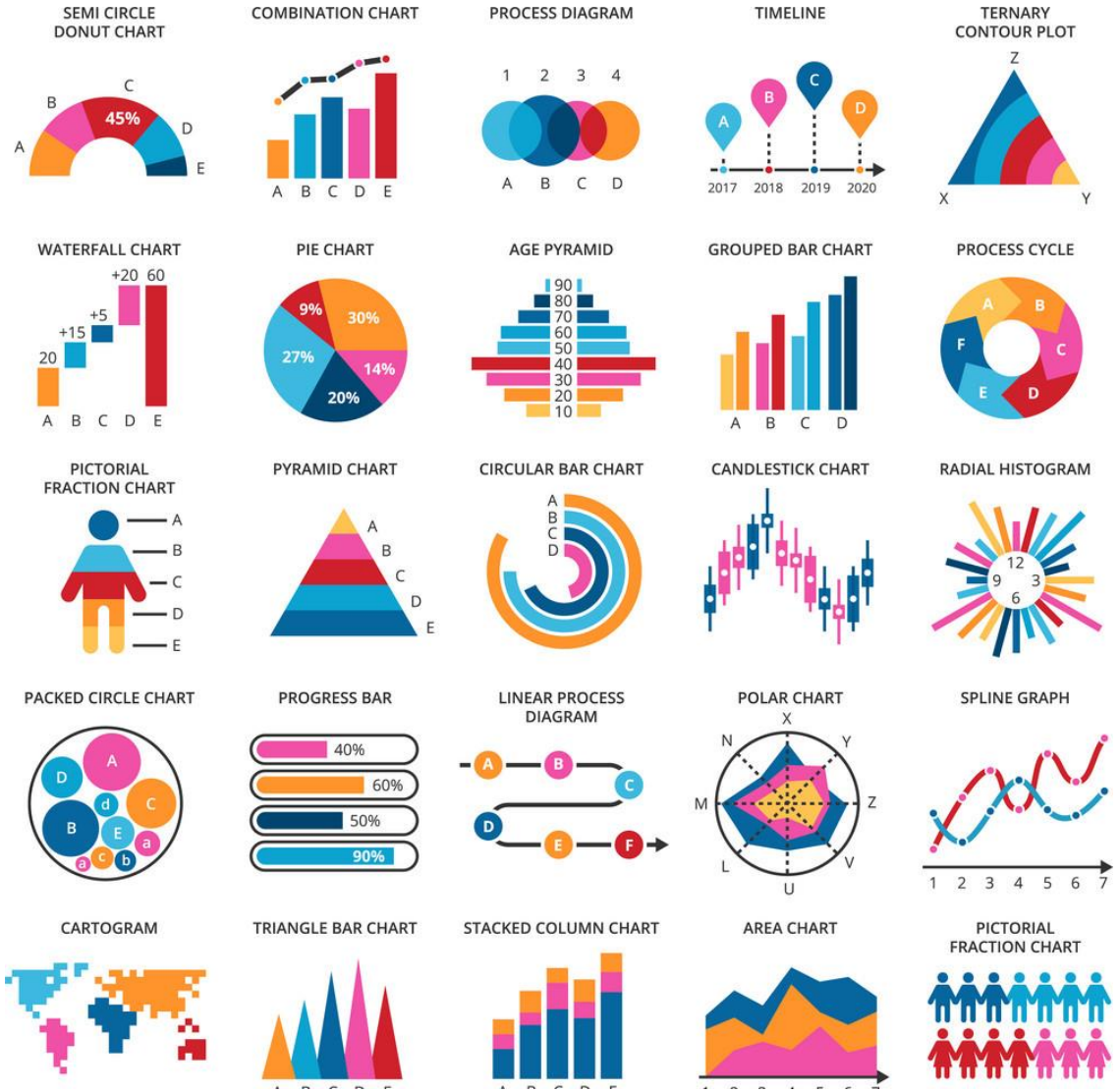
Summary

- Exploratory Data Analysis
  - Introduction
  - Titanic Survival Data Set
    - Tabular Data can come with mixed data types
    - Bar plots / histograms are best for counting data points
    - Missing values can be removed or imputed
    - Data Normalization is important!
    - Ordinal Encoding, Label Encoding, One-hot Encoding
  - Iris Flower Data Set
    - Box plots / violin plots are great for visualizing data distributions
    - Pair plots / correlation matrices can find related features
    - Grouped / nested box plots can be used with categorical data
  - Taylor Swift Spotify Data Set
    - Sorting / ranking the data points by feature values
    - Dates and times can be extracted from tabular data
    - Visualizing the evolution of features with time

EDA tools that we didn't cover:

- Dimensionality reduction (t-SNE, PCA) – **Week 6 and 7**
- Clustering and Density Estimation – **Week 8**
- Hand-crafting new features
- Autocorrelation in Time Series
- EDA on other data modalities

Other plots you may be interested in…



- Know which plots are best to use when.
- Effective data storytelling is as important as the analysis itself.
- There are plenty of code examples on the internet to learn from.
- Practice on your own!

Further Reading

- Hastie et al. (2008). *The Elements of Statistical Learning*. 2nd Ed. Springer.
- Bishop (2006). *Pattern Recognition and Machine Learning*. Springer.
- https://towardsdatascience.com/a-beginners-guide-to-kaggle-s-titanic-problem-3193cb56f6ca
- https://towardsdatascience.com/handling-missing-data-like-a-pro-part-2-imputation-methods-eabbf10b9ce4
- https://scikit-learn.org/stable/modules/impute.html
- https://www.kdnuggets.com/2022/07/scikitlearn-imputer.html
- https://seaborn.pydata.org/tutorial/categorical.html
- Recent news on multi-modal AI: https://science.org/doi/full/10.1126/science.adk6139
- https://www.kaggle.com/code/aaronjones32/predicting-song-popularity-from-spotify-dataset