MACHINE LEARNING

WITH PYTHON

EDA 1: EXPLORING DATA

Model workflow, feature selection

WARM-UP

A stranger barrels up and shows you this picture of some blueberry pastries, saying “Can you help me? I want to make these but I don’t know where to start??”.

You’re flattered… and slightly terrified.

You don’t have time to teach a full baking masterclass, but you decide to drop just enough info so they can figure it out on their own and leave you in peace.

What do you say to them?

COLAB WORKBOOK

Link: click for access

These are class notes. They're never due, but helpful to have when completing homework and studying for standards.

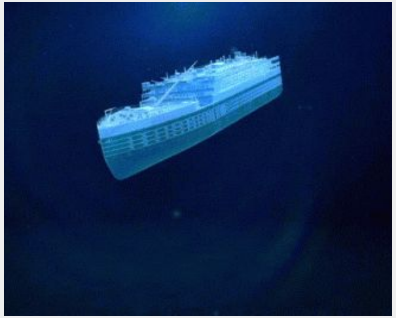
Datasets This Week

Lecture

Palmer Penguins

link: github

HW1 + Lab1



Titanic

link: github

Exercise

Imagine you've just been handed a random sample of a dataset

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
Chinstrap	Dream	50.7	19.7	203.0	4050.0	Male
Adelie	Dream	36.0	18.5	186.0	3100.0	Female
Adelie	Torgersen	41.5	18.3	195.0	4300.0	Male
Adelie	Dream	40.9	18.9	184.0	3900.0	Male
Gentoo	Biscoe	NaN	NaN	NaN	NaN	NaN
Gentoo	Biscoe	59.6	17.0	230.0	6050.0	Male

What does all of this even mean? How do you suppose we go about working with this?

feature / attribute / column

target

feature set

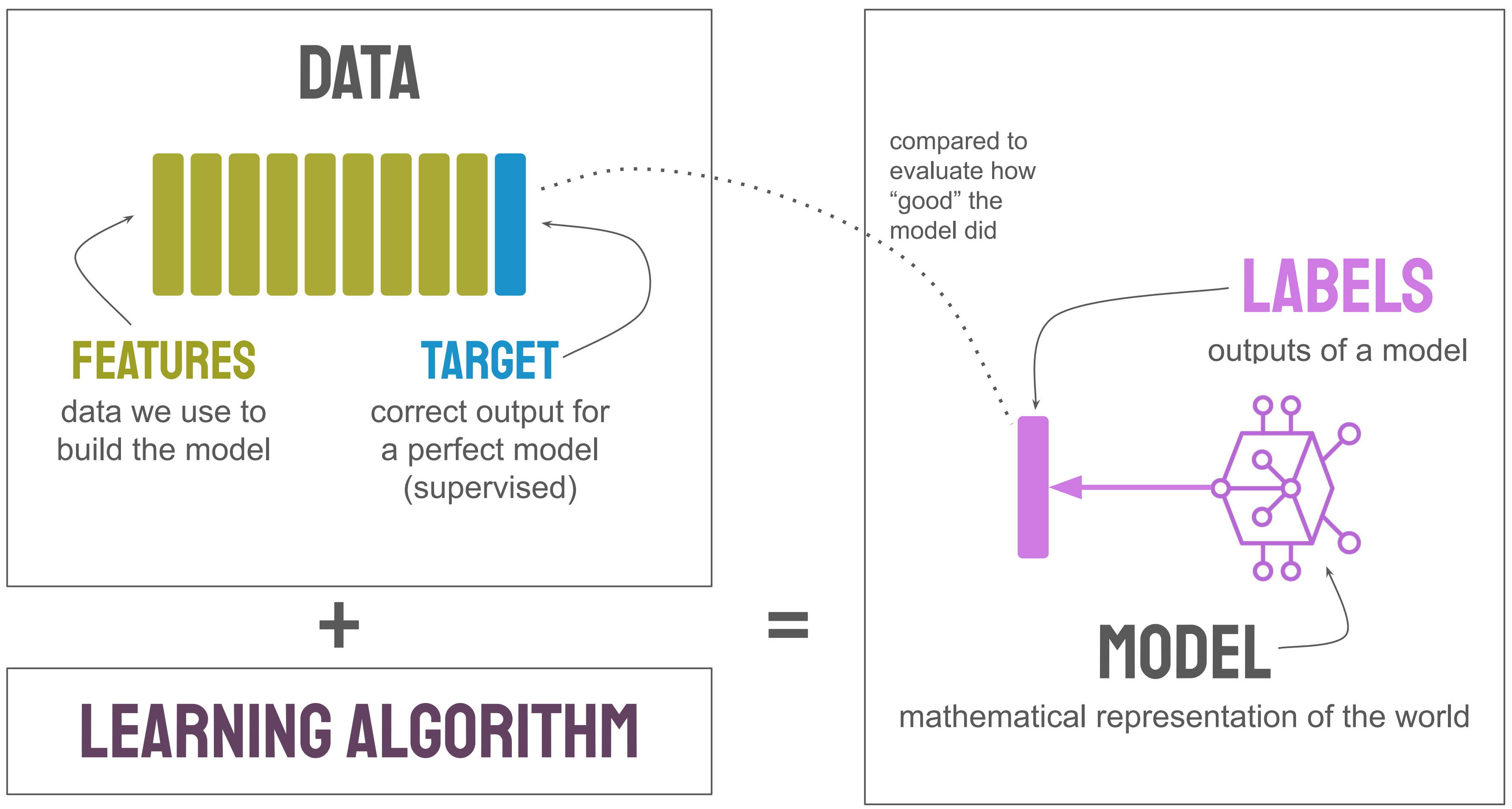
species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
Chinstrap	Dream	50.7	19.7	203.0	4050.0	Male
Adelie	Dream	36.0	18.5	186.0	3100.0	Female
Adelie	Torgersen	41.5	18.3	195.0	4300.0	Male
Adelie	Dream	40.9	18.9	184.0	3900.0	Male
Gentoo	Biscoe	NaN	NaN	NaN	NaN	NaN
Gentoo	Biscoe	59.6	17.0	230.0	6050.0	Male

class / label / ground truth

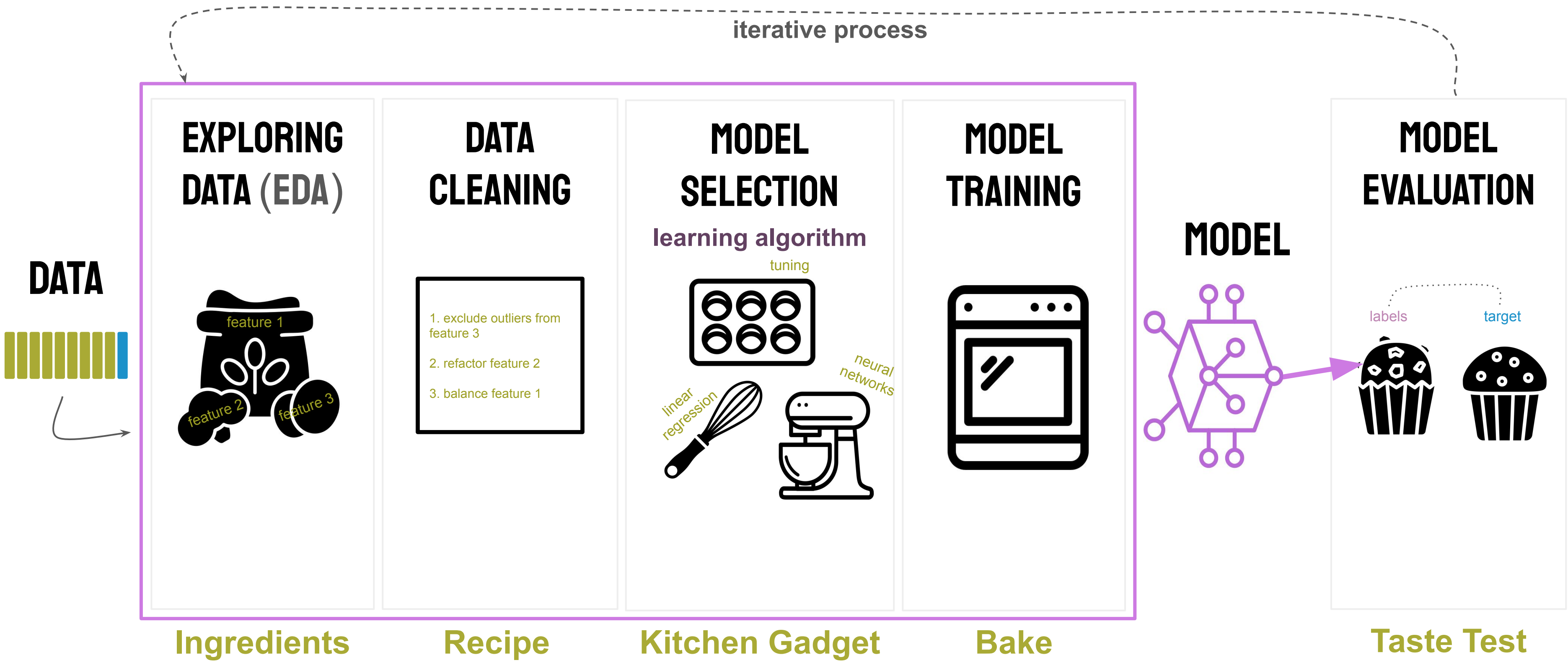
row / record / sample / instance

entry / value

DATASET TERMINOLOGY


DATA
FEATURES
data we use to build the model
TARGET
correct output for a perfect model (supervised)
+
LEARNING ALGORITHM
=
compared to evaluate how "good" the model did
LABELS
outputs of a model
MODEL
mathematical representation of the world

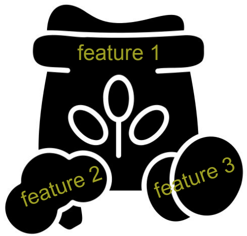
5 STEP MACHINE LEARNING WORKFLOW OR BAKING THE CAKE



5 STEP MACHINE LEARNING WORKFLOW OR BAKING THE CAKE

EXPLORING DATA (EDA)

statistical tests
distributions
missingness
outliers
correlations
visualizations



Ingredients

DATA CLEANING

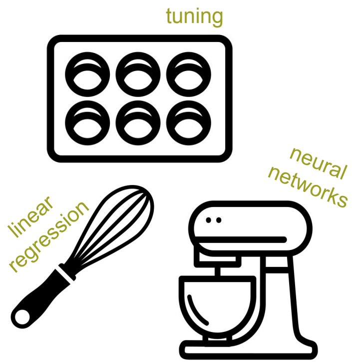
feature engineering
one-hot encoding
standardization
transformation
class balancing
hash/embeddings
deal with missingness

1. exclude outliers from feature 3
2. refactor feature 2
3. balance feature 1

Recipe

MODEL SELECTION

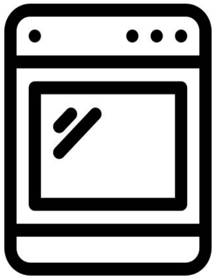
learning algorithm
hyperparameter tuning
selection
training set



Kitchen Gadget

MODEL TRAINING

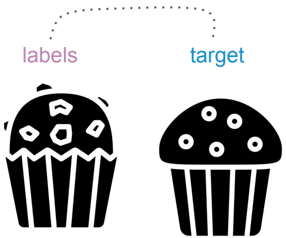
cross-fold validation
run models
dev set



Bake

MODEL EVALUATION

test set
accuracy
precision
recall
F1-score
confusion matrix
ROC or AUC
MSE or RMSE or MAE



Taste Test

Supervised

Teach model to give accurate **predictions** on new unseen data

Labeled data (has **target**)

- Classifying images
- Language translation
- Predicting housing prices
- Object detection
- Classifying mushrooms as poisonous/edible

Supervised:
Classification

Unsupervised

Teach model to discover **groups** and patterns

Unlabeled data

- Clustering
- Outliers
- Denoising signals
- Topic modeling
- Recommender systems
- Dimensionality reduction
- Grouping mushrooms by characteristics

Unsupervised:
Clustering

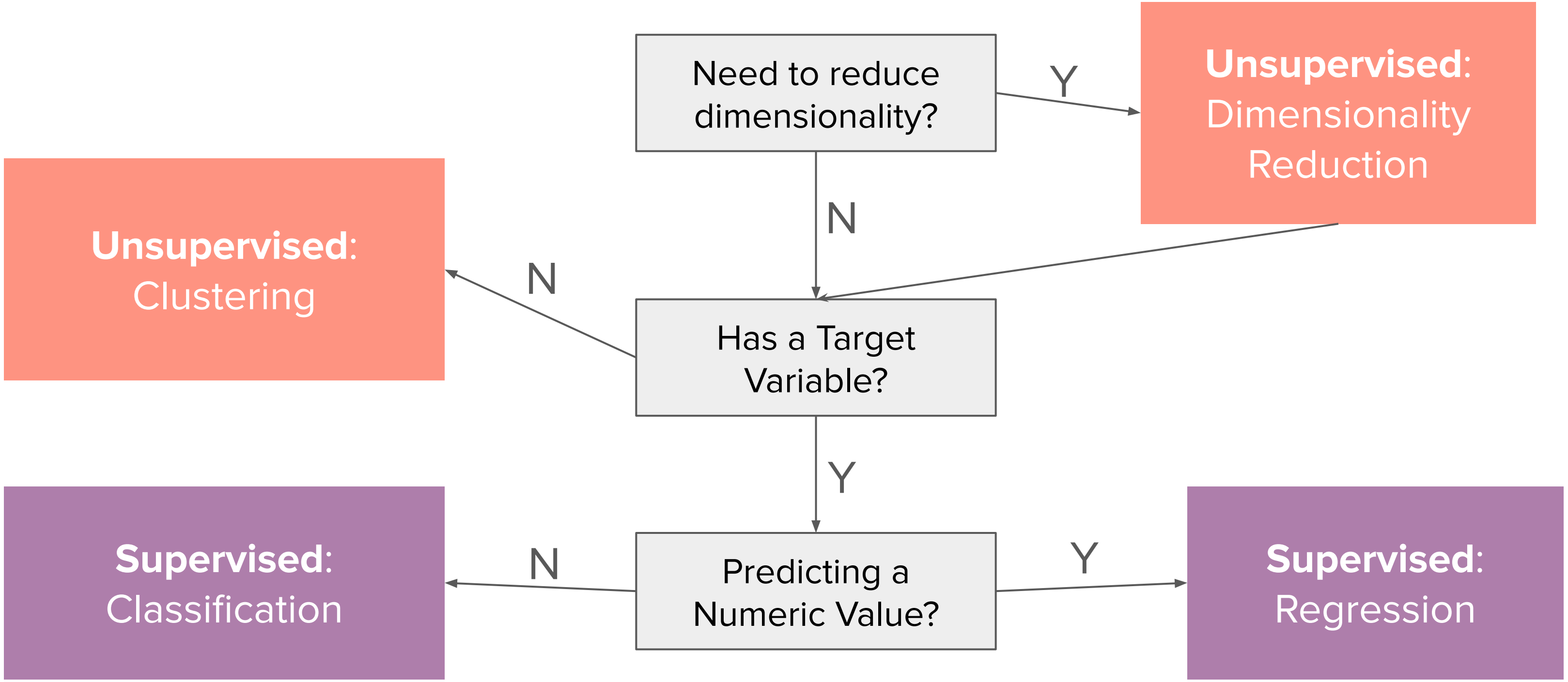
Reinforcement

Teach an agent about the world by rewarding good behavior

- go/chess
- generative design
- predator/prey

Reinforcement

Model Selection Flowchart



So what can you do with penguin data? Grab a partner, and spend five minutes creating an idea for each of the model types below. Once your group has this filled out, add your ideas to a board (each person does an approach)

Supervised Regression

Target variable:

Feature set:

Model rationale:

Unsupervised Clustering

Target variable:

Feature set:

Model rationale:

No free lunch

No single algorithm can solve all problems

Algorithm choice has consequences

Important to know pitfalls to help mitigate bias

Why learn all of this when genAI exists?

Need to know foundations in order to understand how model can generate new data from learned patterns

Understand complex architectures

Interpretability, bias, and ethics

Appropriate applications

feature types can be…

Numerical: numeric values

Ex: quiz scores: 7, 24, 35

Continuous: real numeric values (decimal)

Ex: weights of hamsters: 4.23, 3.2

Categorical: finitely many values

Ex: animals: dog, cat, iguana

Ordinal: categorical, but with natural order

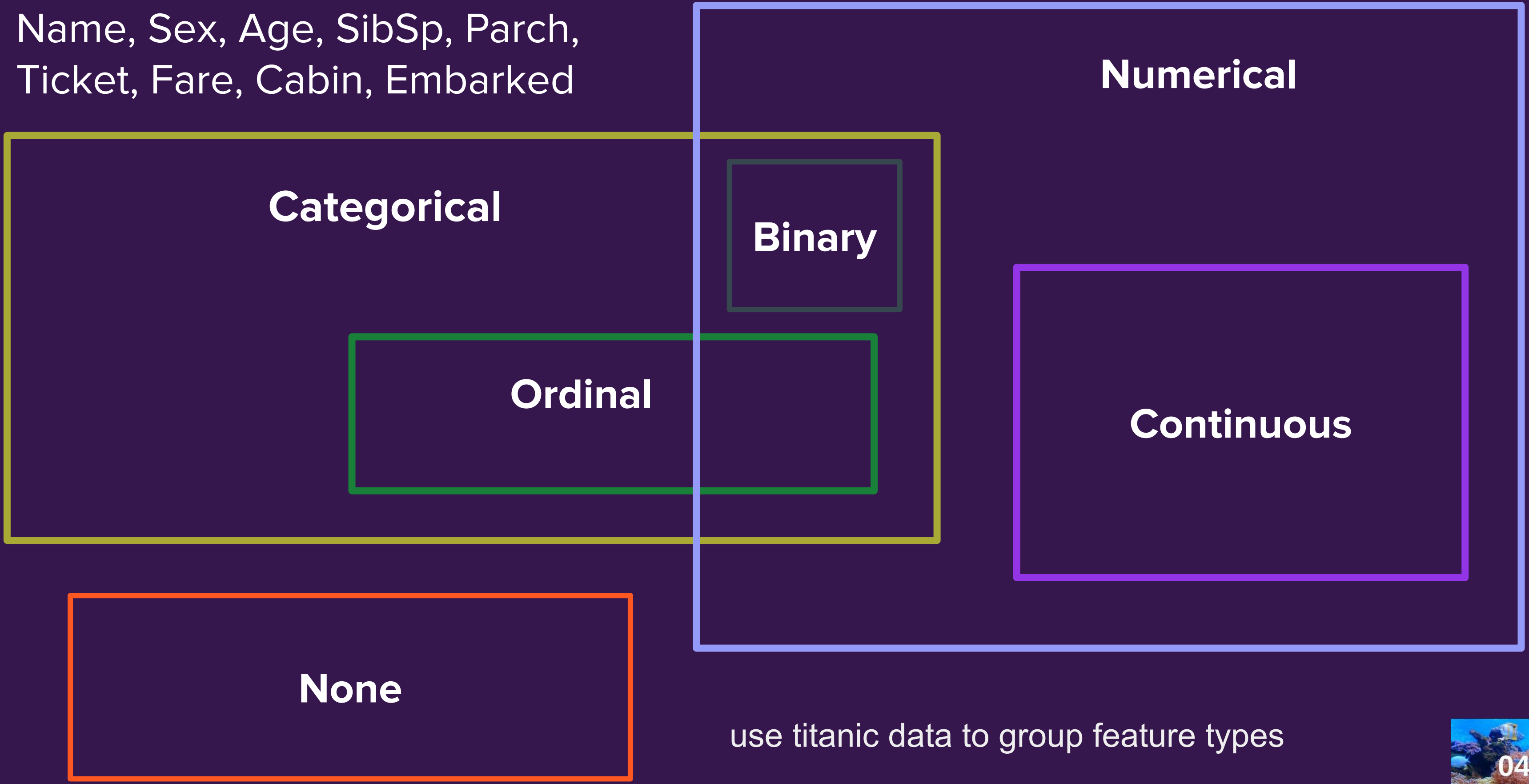
Ex: sizing: small, medium, large

Binary: two options (usually 0 and 1)

Ex: sex: Male, Female

categorical		numeric				categorical
		continuous		discrete		binary
species	**island**	**bill_length_mm**	**bill_depth_mm**	**flipper_length_mm**	**body_mass_g**	**sex**
Chinstrap	Dream	50.7	19.7	203.0	4050.0	Male
Adelie	Dream	36.0	18.5	186.0	3100.0	Female

PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked



use titanic data to group feature types

MACHINE LEARNING
WITH PYTHON

EDA2: EXPLORING DATA

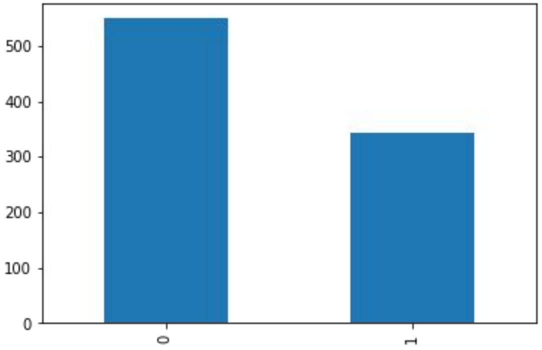
Statistics: descriptive, distributions, tests

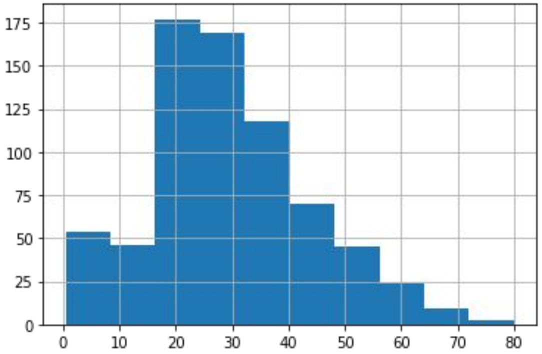
COLAB WORKBOOK

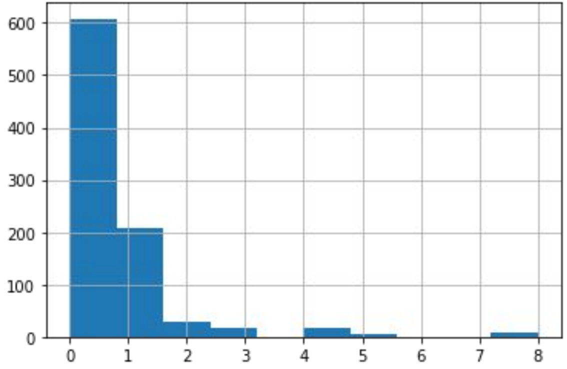
Link: click for access

Warm-up

1. For each of the following plots, describe what you see. Use relevant statistical terminology if you can remember it.







2. In your own words, describe mean, median, mode, and standard deviation.

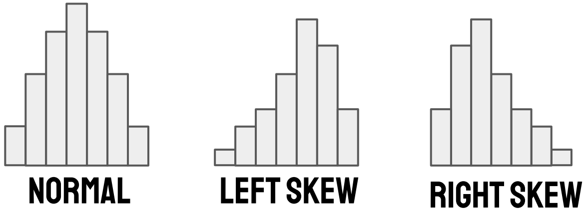
3. What are methods used to see whether there are differences among groups?

What should we be on the lookout for?

numerical

are distributions normal or skewed?

any features needing transformation?

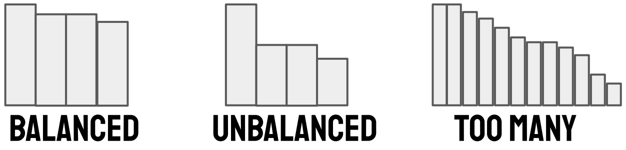


some algorithms do poorly on skewed data, so we usually transform these during data cleaning

categorical

balanced or unbalanced?

which categories could be combined?



which categories underrepresented?

should sampling occur?

this includes both categorical and numeric features

To see a table of common statistics for all features we can use describe!

```
penguins.describe(include= 'all')
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
count	344	344	342.000000	342.000000	342.000000	342.000000	333
unique	3	3	NaN	NaN	NaN	NaN	2
top	Adelie	Biscoe	NaN	NaN	NaN	NaN	Male
freq	152	168	NaN	NaN	NaN	NaN	168
mean	NaN	NaN	43.921930	17.151170	200.915205	4201.754386	NaN
std	NaN	NaN	5.459584	1.974793	14.061714	801.954536	NaN
min	NaN	NaN	32.100000	13.100000	172.000000	2700.000000	NaN
25%	NaN	NaN	39.225000	15.600000	190.000000	3550.000000	NaN
50%	NaN	NaN	44.450000	17.300000	197.000000	4050.000000	NaN
75%	NaN	NaN	48.500000	18.700000	213.000000	4750.000000	NaN
max	NaN	NaN	59.600000	21.500000	231.000000	6300.000000	NaN

NUMERICAL FEATURES

		species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
	count	344	344	342.000000	342.000000	342.000000	342.000000	333
	unique	3	3	NaN	NaN	NaN	NaN	2
	top	Adelie	Biscoe	NaN	NaN	NaN	NaN	Male
	freq	152	168	NaN	NaN	NaN	NaN	168
balancing point of distribution	**mean**	NaN	NaN	43.921930	17.151170	200.915205	4201.754386	NaN
spread	**std**	NaN	NaN	5.459584	1.974793	14.061714	801.954536	NaN
smallest value of data	**min**	NaN	NaN	32.100000	13.100000	172.000000	2700.000000	NaN
	25%	NaN	NaN	39.225000	15.600000	190.000000	3550.000000	NaN
middle value of data	**50%**	NaN	NaN	44.450000	17.300000	197.000000	4050.000000	NaN
	75%	NaN	NaN	48.500000	18.700000	213.000000	4750.000000	NaN
largest value of data	**max**	NaN	NaN	59.600000	21.500000	231.000000	6300.000000	NaN

DESCRIPTIVE STATS: NUMERICAL FEATURES

add your answers to your workbook

Mean: the balancing point of the data

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{\underbrace{n}_{\text{sample size}}} = \frac{\text{sum of all records}}{\text{number of records}}$$

Median: middle of ordered data, also called the **50th percentile**

Example 1: Five employee salaries at a small start-up:

\$33k, \$35k, $\boxed{\$35k,}$ \$37k, \$210k

Find **mean**: $\frac{\$33k+\$35k+...+\$210k}{5}$ = **\$70k**

Find **median**: **\$35k**

If mean and median are quite **different** from each other, then the data is **skewed**. When **median < mean** the distribution is **right** skewed. If **median > mean**, then **left**. If significantly << or >>, then add a very in front. e.g. 35k<<70k ➔ very right skewed.

DESCRIPTIVE STATS: NUMERICAL FEATURES

Standard deviation: (approximately) the average **distance** of the data **from the mean**, also called spread.

$$s = \sqrt{\frac{\sum_{i=1}^{N} \boxed{(x_i - \overline{x})}^2}{N-1}}$$

when **s = 0**, then **no variability**

add your answers to your workbook

Example2: For the following two scenarios, give a list of six numbers chosen from the set: **0, 1, 2, 3, 4, 5** (repeats are ok)

Scenario1: Six numbers with **smallest** possible standard deviation

All numbers are the same → s=0
Set: 4, 4, 4, 4, 4, 4

Scenario2: Six numbers with **largest** possible standard deviation

Numbers equally on far ends
Set: 0, 0, 0, 5, 5, 5

DESCRIPTIVE STATS: NUMERICAL FEATURES

Kth percentile: data value s.t. k% of the data is less than or equal to that value

Ex: If you scored in 34% percentile in reading in the 3rd grade, then your score was better than 34% of the other 3rd grade students

Quartiles: special percentiles that split the data into quarters

First quartile **Q1**: 25th percentile

Second quartile **M**edian: 50th percentile

Third quartile **Q3**: 75th percentile

Here is how we could order the following columns from:

most left skewed (mean<median) to most right skewed

bill_length
bill_depth
flipper_length
body_mass

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
count	344	344	342.000000	342.000000	342.000000	342.000000	333
unique	3	3	NaN	NaN	NaN	NaN	2
top	Adelie	Biscoe	NaN	NaN	NaN	NaN	Male
freq	152	168	NaN	NaN	NaN	NaN	168
mean	NaN	NaN	43.921930	17.151170	200.915205	4201.754386	NaN
std	NaN	NaN	5.459584	1.974793	14.061714	801.954536	NaN
min	NaN	NaN	32.100000	13.100000	172.000000	2700.000000	NaN
25%	NaN	NaN	39.225000	15.600000	190.000000	3550.000000	NaN
50%	NaN	NaN	44.450000	17.300000	197.000000	4050.000000	NaN
75%	NaN	NaN	48.500000	18.700000	213.000000	4750.000000	NaN
max	NaN	NaN	59.600000	21.500000	231.000000	6300.000000	NaN

very left skewed — balanced — very right skewed

body_mass bill_length bill_depth flipper_length

CATEGORICAL FEATURES

		species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
number of records	**count**	344	344	342.000000	342.000000	342.000000	342.000000	333
number of categories	**unique**	3	3	NaN	NaN	NaN	NaN	2
top category	**top**	Adelie	Biscoe	NaN	NaN	NaN	NaN	Male
count in top category	**freq**	152	168	NaN	NaN	NaN	NaN	168
	mean	NaN	NaN	43.921930	17.151170	200.915205	4201.754386	NaN
	std	NaN	NaN	5.459584	1.974793	14.061714	801.954536	NaN
	min	NaN	NaN	32.100000	13.100000	172.000000	2700.000000	NaN
	25%	NaN	NaN	39.225000	15.600000	190.000000	3550.000000	NaN
	50%	NaN	NaN	44.450000	17.300000	197.000000	4050.000000	NaN
	75%	NaN	NaN	48.500000	18.700000	213.000000	4750.000000	NaN
	max	NaN	NaN	59.600000	21.500000	231.000000	6300.000000	NaN

DESCRIPTIVE STATS: CATEGORICAL FEATURES

Let’s talk about some things to keep an eye out for when looking at these features

If unique > 5-8. We usually convert categorical columns to binary columns, each column being a unique category.

Too many categories => lots of columns => dimension increases.

Solution: consolidate to 5 or less categories.

If top frequency >> average count. Want each category to have similar counts, this is called balanced.

Too many records of single category (unbalanced) means low diversity in training models => overfitting/underfitting problem

Solution: up/down sample accordingly.

Discuss the categorical features. Which are balanced vs unbalanced.

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
count	344	344	342.000000	342.000000	342.000000	342.000000	333
unique	3	3	NaN	NaN	NaN	NaN	2
top	Adelie	Biscoe	NaN	NaN	NaN	NaN	Male
freq	152	168	NaN	NaN	NaN	NaN	168
mean	NaN	NaN	43.921930	17.151170	200.915205	4201.754386	NaN
std	NaN	NaN	5.459584	1.974793	14.061714	801.954536	NaN
min	NaN	NaN	32.100000	13.100000	172.000000	2700.000000	NaN
25%	NaN	NaN	39.225000	15.600000	190.000000	3550.000000	NaN
50%	NaN	NaN	44.450000	17.300000	197.000000	4050.000000	NaN
75%	NaN	NaN	48.500000	18.700000	213.000000	4750.000000	NaN
max	NaN	NaN	59.600000	21.500000	231.000000	6300.000000	NaN

unbalanced

balanced

island species

sex

STATISTICAL TESTS

t-test categorical numerical

Determines **difference in means** between two categories

H0: there is no difference in average **body mass** among Chinstrap and Gentoo **species**

chi-sq categorical categorical

Determines **association** between two features

H0: there is no association between **penguin species** and **islands**

ANOVA categorical numerical

Determines **difference in means** between two+ categories

H0: there is no difference in average **body mass** among all three penguin **species**

correlation numerical numerical

Determines **linear relationship** between two features

H0: there is no linear relationship between **bill length** and **bill depth**

HYPOTHESIS TESTING

Trying to find enough **evidence** to disprove a **statement**

H_1: there is a significant difference between average body mass among two penguin species

The statement we're trying to disprove is called our **null hypothesis**, H_0. The alternative is H_1.

H_0: there is no difference between average body mass among two penguin species

The first bit of evidence we're collecting is a test statistic. The test we use is determined by the type of features being compared.

We have one numeric feature and one categorical feature, so we use a **t-test**.

Note: missing values will need to be removed for testing purposes

HYPOTHESIS TESTING

Trying to find enough **evidence** to disprove a **statement**

What's considered small? Typically between **0.01 and 0.005**, sometimes 0.05. Depends on the application.

The second bit of evidence we're collecting is a **p-value**, which tells us how likely it is to see a test statistic more extreme than the one we calculated. If small, then statistically significant.

If we run the test and get a small p-value, then we reject the null and conclude there is a significant difference in the body mass of penguin species. If p-value isn't small, cannot reject the null (inconclusive)

If significant, then need to further explore to unpack what's happening as p-value isn't enough evidence.

If significant, create some visualizations to see what is going on.

goal of this next step is to collect more evidence around the claim that something is indeed significant.

The misconception about p-values

What they actually measure. It's evidence that we see a given value assuming the null is true. If small chance to see that value, we reject the null. We're not saying it's not true, just very unlikely.

What to do next. p-values are just one piece of evidence. Gather more by doing other stats (confidence intervals, sample size) and visualizations (boxplots, histograms, heatmaps)

Note: To do HW1 #4 you'll need to know which viz to do (we'll learn this next time!)

Why the definition of 'small' changes. Depending on the context, it may be catastrophic if you reject something based on an arbitrary threshold. On the other hand, some folks pick a certain threshold so their results become significant. This is called p-hacking and is problematic as it undermines valid research practices.

More info pls. ASA has great advice on how to interpret p-values (link here).

Interpret the statistical tests in your colab workbook by filling in the blanks. Some of the notes here may be helpful.

H_1: there is a significant difference between average body mass among penguin species

H_0: there is no difference between average body mass among penguin species

If we run the test and get a small p-value, then we reject the null and conclude there is a significant difference in the body mass of penguin species. If p-value isn't small, cannot reject the null (inconclusive)

t-test: determines **difference in means** between two categories

ANOVA: determines **difference in means** between two+ categories

chi-sq: determines **association** between two categorical features

correlation: determines **linear relationship** between two numeric features

Quiz INTRO

We’ll continue content at 1:50. Spend ~5 minutes on each page, adding explanation where needed.

HOMEWORK WORK TIME

HW1

Let's get a start on some more homework questions. Grab a partner to get started. If you have any questions on finding the dataset please let me know!

[View on our website]

MACHINE LEARNING
WITH PYTHON

EDA3: EXPLORING DATA

Visualizations

COLAB WORKBOOK

Link: click for access

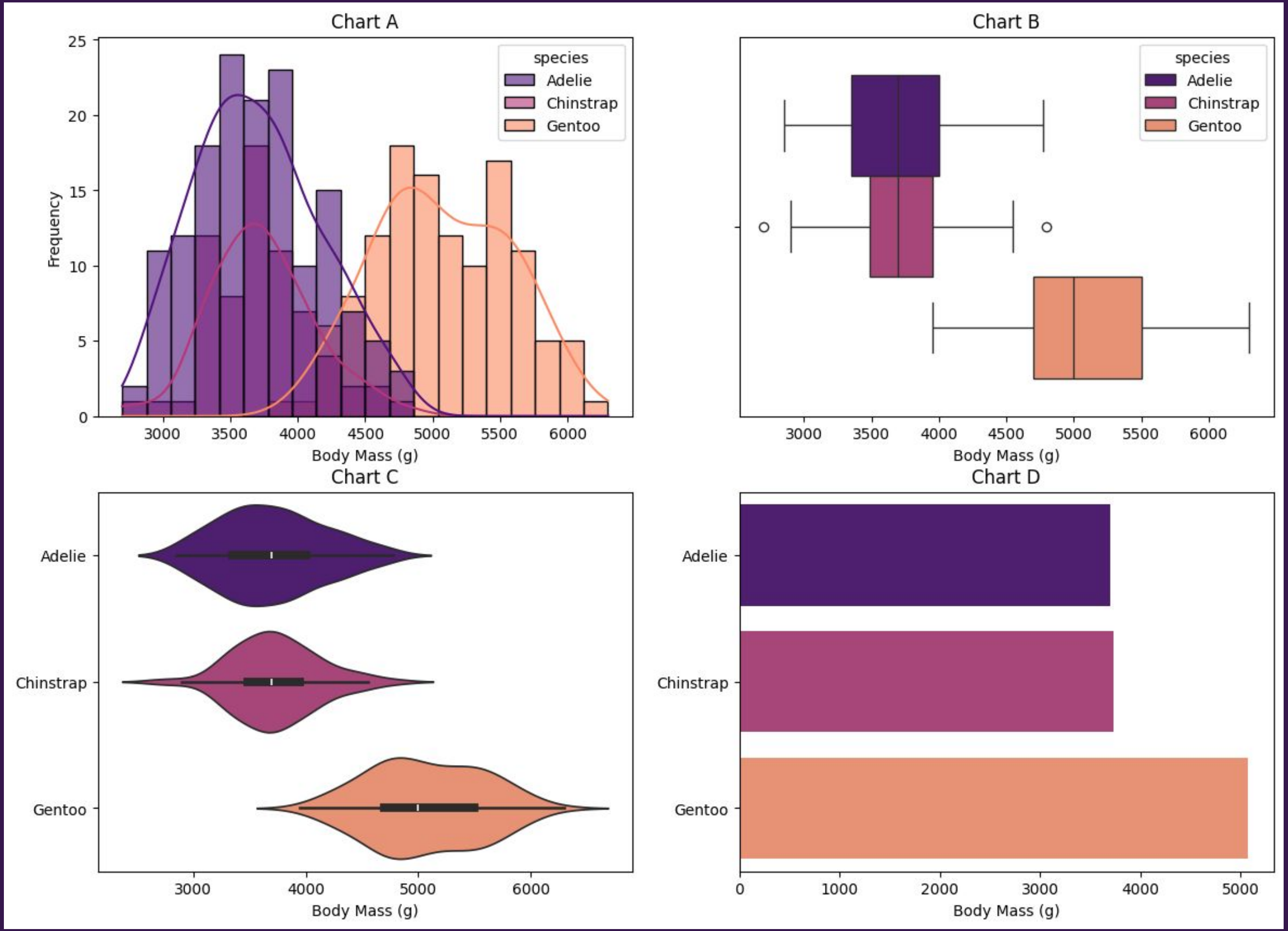
Warm-up

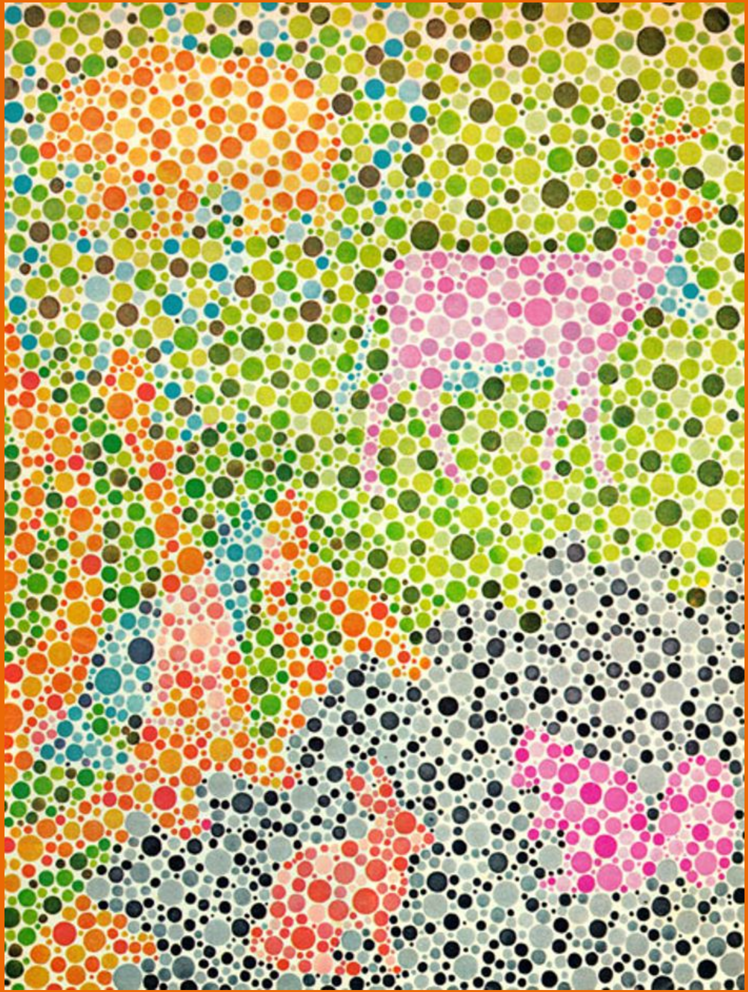
Each of the following visualizations compares body mass across penguin species.

For each chart, describe what it does well, and what it could improve upon.

Consider things like statistical measures, visual appeal, informativeness, etc.

Lastly, come up with an informative and appealing title for each chart.





sim daltonism link

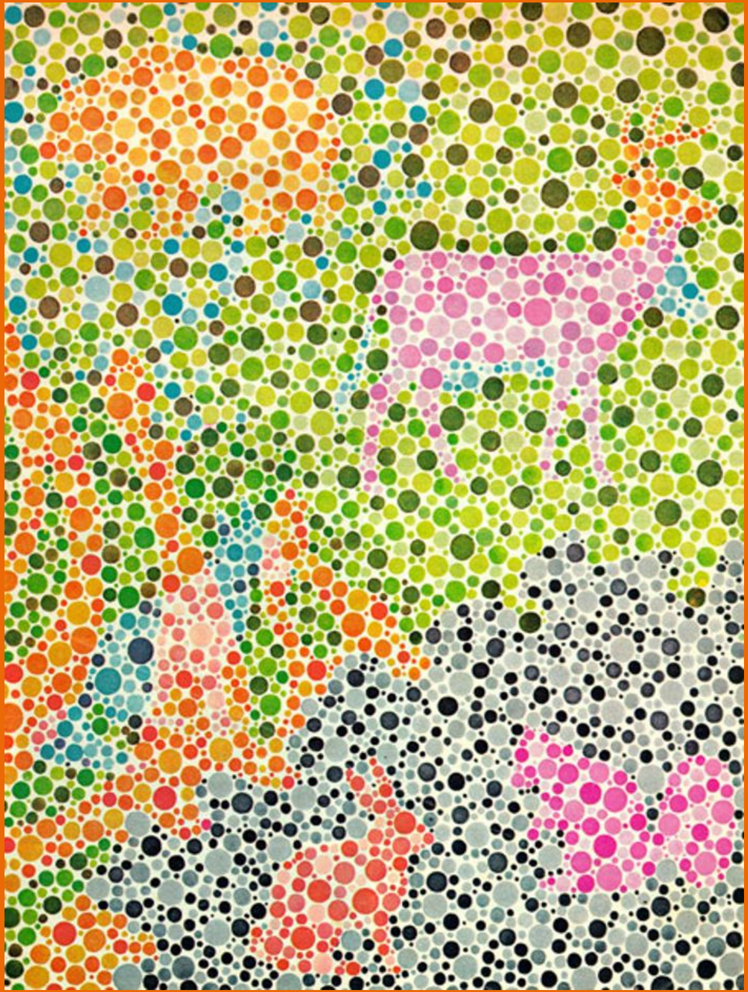


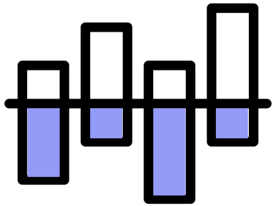
photo by Field and Stream magazine

Impactful visuals leads to better insights

Better insights leads to more informed decisions

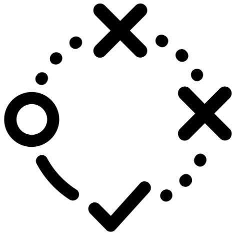
…so how do we create impactful visualizations?

tell a story

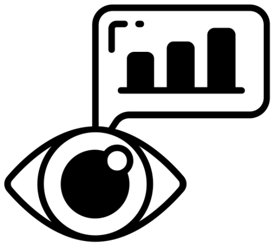


color matters

avoid cognitive overload



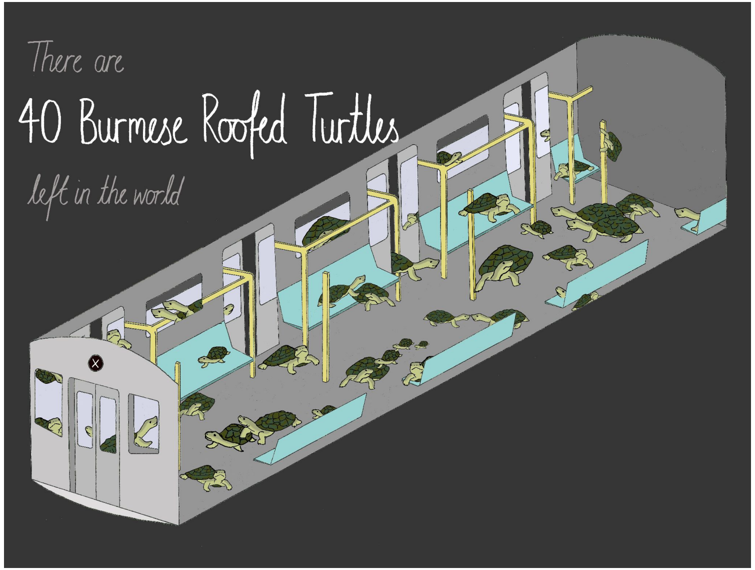
experiment!

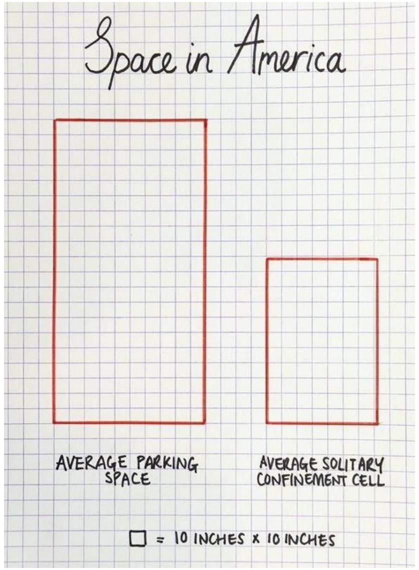


simple but rich

Visualizations

top 10 dos and don’ts reading: link
how to use colors in visualizations video (link)





Visualizations

Sketching with Data with Mona Chalabi ([link]())

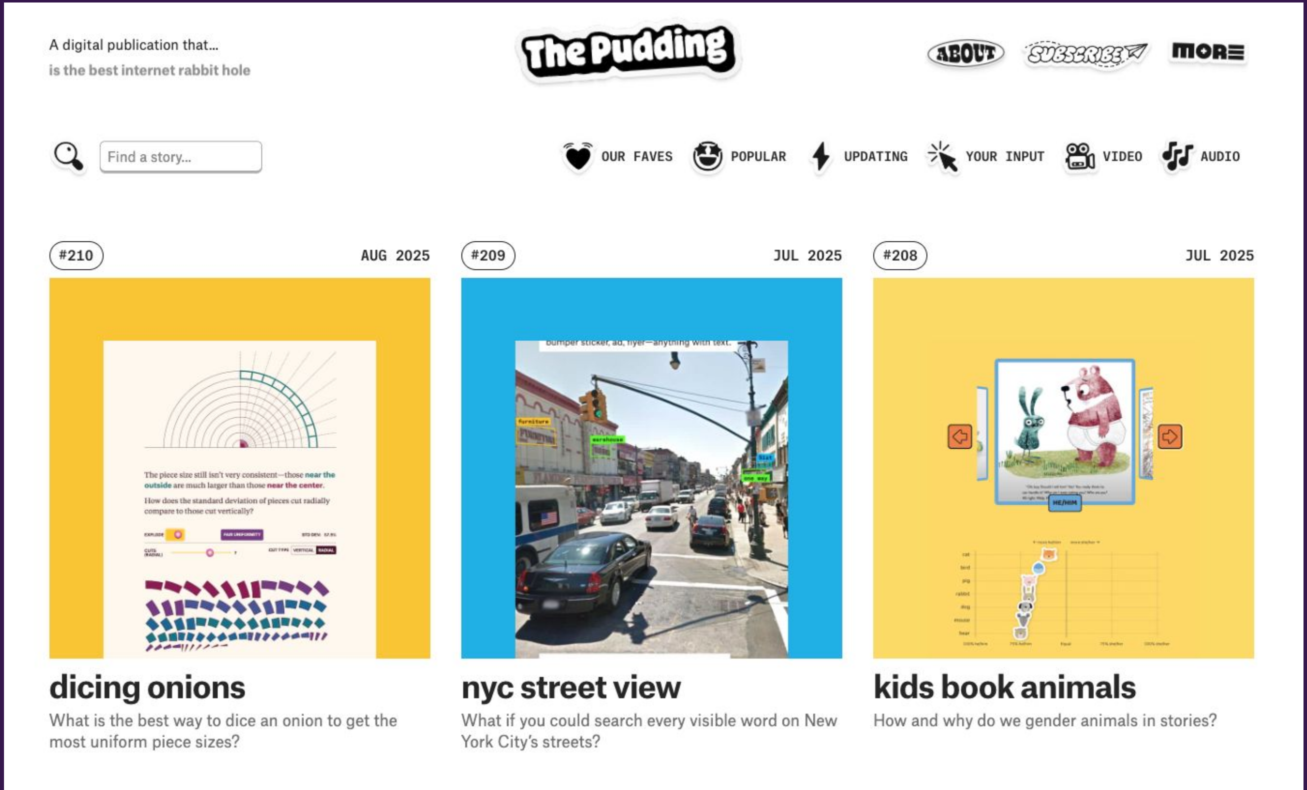
Exercise

Visit The Pudding (pudding.cool) and explore some stories.

Take note of the visual cues they use.

What makes the visualizations or charts nice to look at?

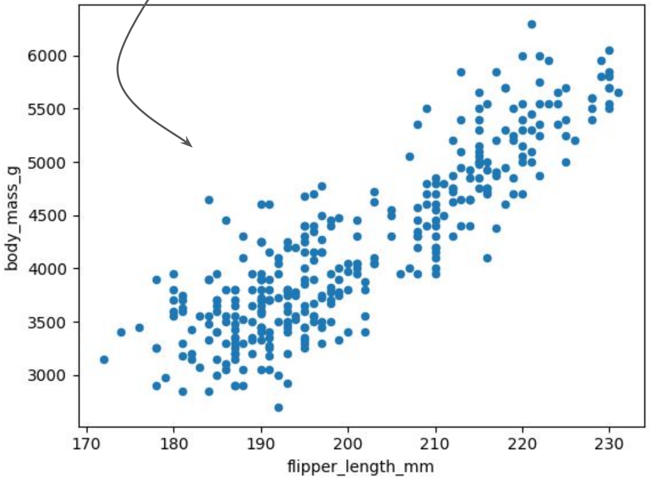
What narrative do they use to help you understand the story being told?

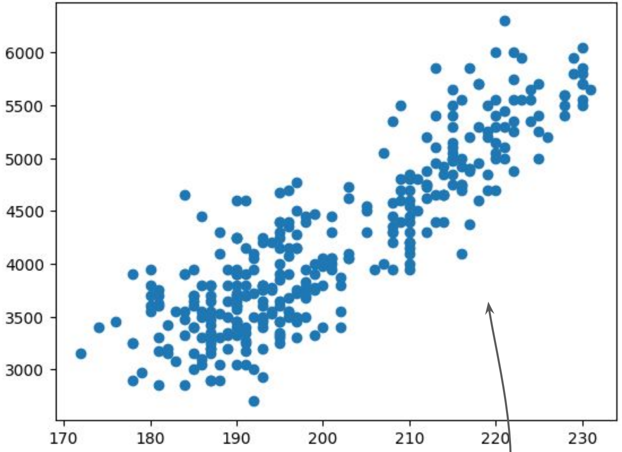


plot

documentation

```
penguins.plot(kind = "scatter",
              x= "flipper_length_mm",
              y = "body_mass_g")
```





matplotlib

documentation

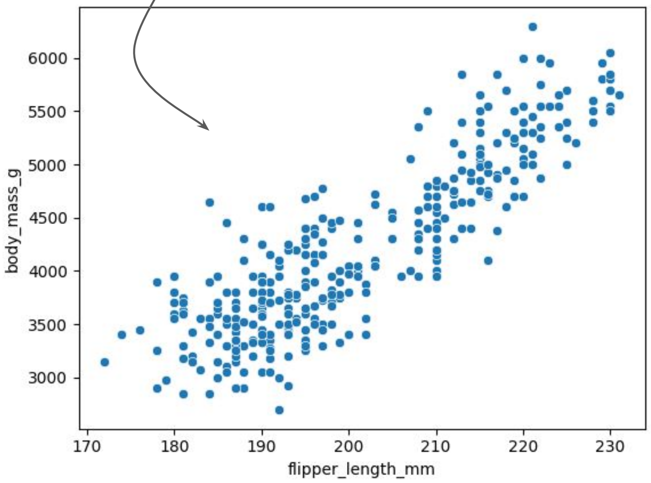
color names

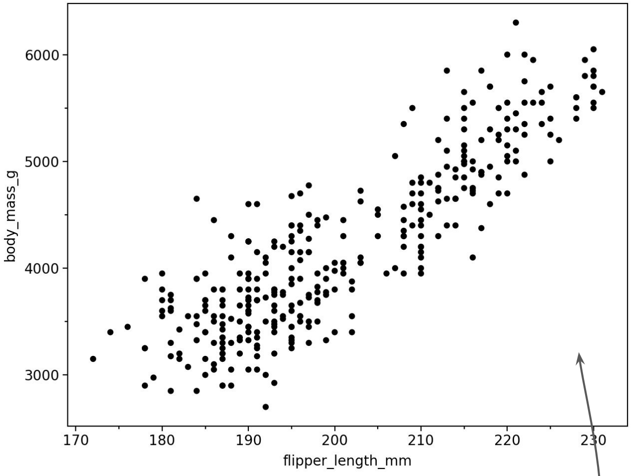
```
plt.scatter(penguins["flipper_length_mm"],
            penguins["body_mass_g"])
```

seaborn

documentation

```
sns.scatterplot(data = penguins,
                x = "flipper_length_mm",
                y = "body_mass_g")
```





plotnine

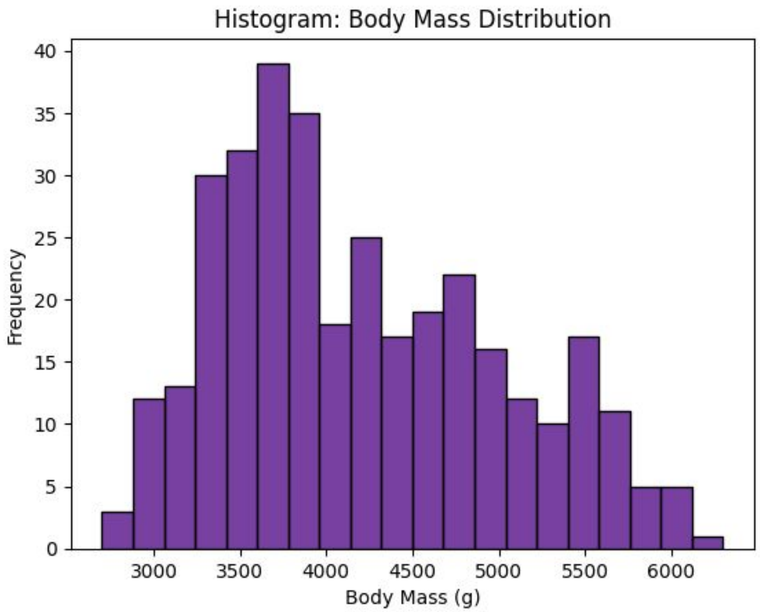
documentation

```
(ggplot(penguins,
          aes("flipper_length_mm", "body_mass_g"))
    + geom_point()
)
```

Univariate → single feature

more bins = smaller binwidth
less bins miss details, but
more bins prone to noise

Histograms show distribution and skewness of a **continuous** feature




```
sns.histplot(data=penguins,
             x='body_mass_g',
             color='indigo',
             bins=20)

plt.title('Body Mass Distribution')
plt.xlabel('Body Mass (g)')
plt.ylabel('Frequency')
```

take note symmetry vs skewness; modality; clumps or gaps

```
penguins['body_mass_g'].plot(kind='hist', bins=20,color='indigo', edgecolor='black')
```

Sometimes it's best to see a table of common statistics of a **numeric** features rather than creating a visualization

	body_mass_g
count	342.000000
mean	4201.754386
std	801.954536
min	2700.000000
25%	3550.000000
50%	4050.000000
75%	4750.000000
max	6300.000000

```
penguins['body_mass_g'].describe()
```

Notice 50% percentile is slightly less than the mean.. This means it is skewed to the __**right**__

Sometimes it's best to see a table of the distribution of a **categorical** feature rather than a visualization

	count
species	
Adelie	152
Gentoo	124
Chinstrap	68

```
penguins['species'].value_counts()
```

When value counts are roughly the same for each category this is called **balanced**. Highly unbalanced data is less ideal for ML outcomes.

Would you consider the species feature to be balanced or unbalanced?

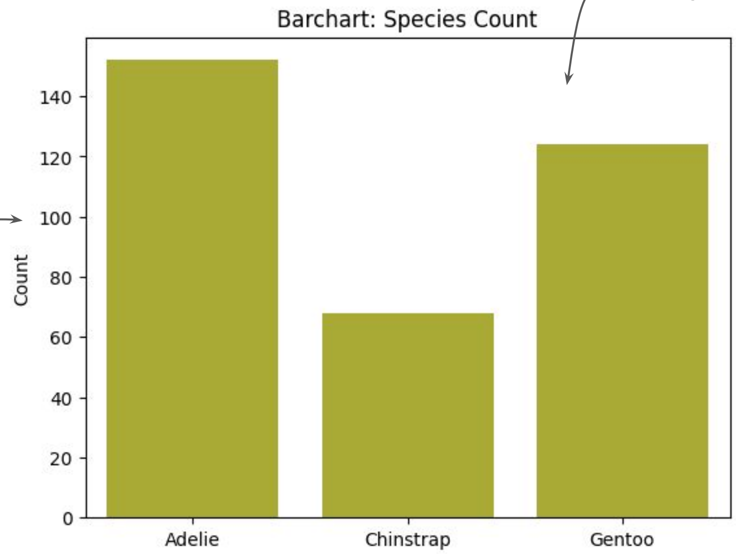
Mode: category/ies with highest value count

Univariate → single feature

Barcharts show counts within a categorical feature

axes should be scaled correctly

not cluttered: 3-5 categories good, anymore may require you to combine smaller categories together OR use color only on important categories




```
sns.countplot(data=penguins,
              x='species',
              color='tab:olive')

plt.title('Barchart: Species Count')
plt.xlabel(None)
plt.ylabel('Count')
```

take note

balanced vs unbalanced; appropriate axes

```
penguins['species'].value_counts().plot(kind='bar', x='species', y='count', color='tab:olive')
```

Bivariate → **two features**

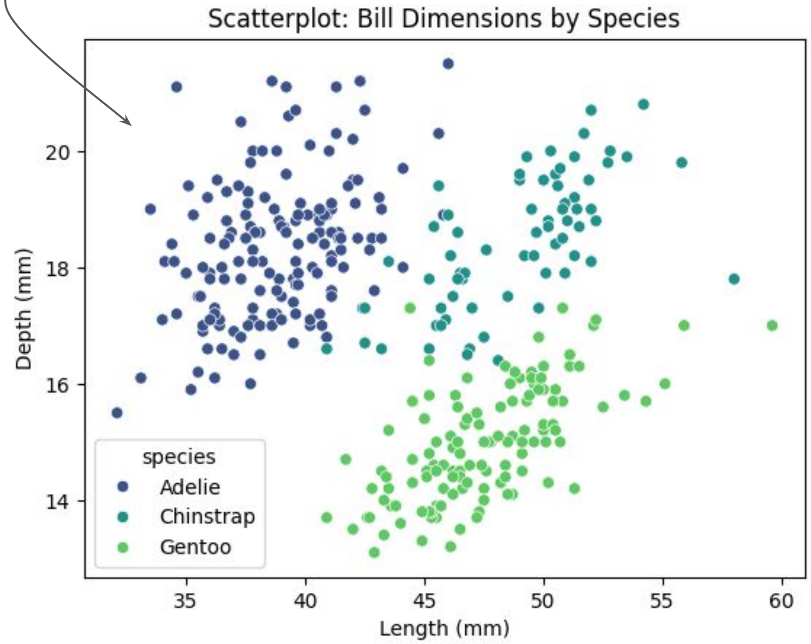
Scatterplots show correlations between two **continuous** features

relatively small data, somewhat linear

y: **depth**

x: **length**

color: **species**




```
sns.scatterplot(data=penguins,
                x='bill_length_mm',
                y='bill_depth_mm',
                hue = 'species',
                palette='viridis')
plt.title('Bill Dimensions by Species')
plt.xlabel('Length (mm)')
plt.ylabel('Depth (mm)')
```

take note

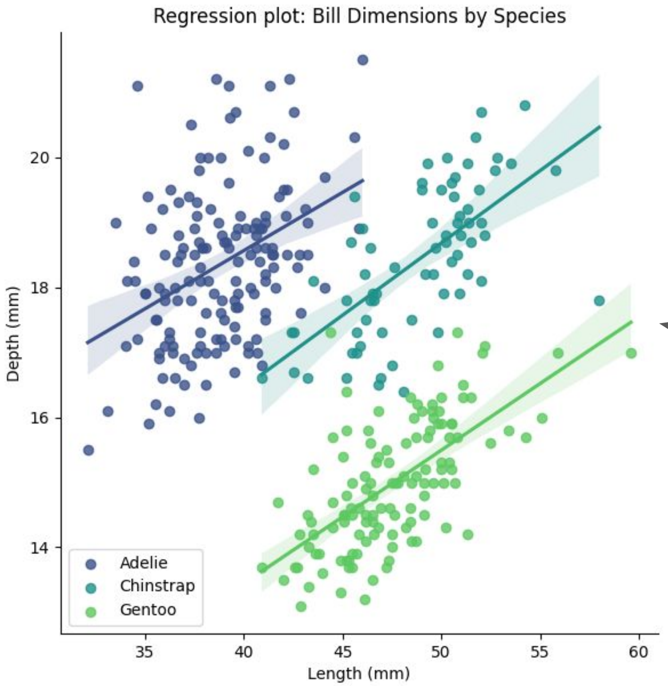
correlation != causation

```
penguins.plot(kind='scatter', x='bill_length_mm', y='bill_depth_mm', color='indigo')
```

Multivariate → **3+** features

Regression plots show correlations between two **continuous** features and fit a line of best fit

y: depth | x: length | color: species




```
sns.lmplot(data=penguins,
           x='bill_length_mm',
           y='bill_depth_mm',
           hue="species",
           palette='viridis',
           legend=None, height=6)
plt.legend(loc='lower left', ncol=1)
```

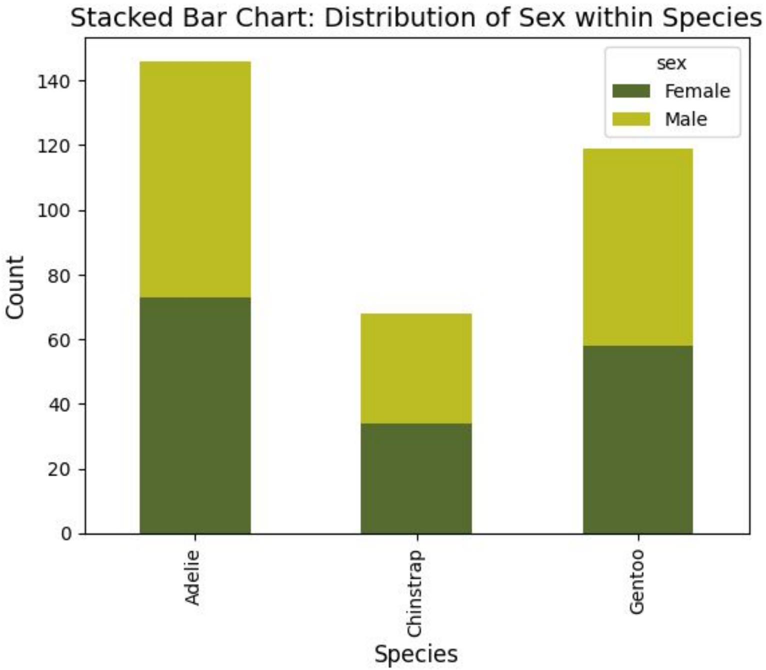
take note correlation != causation

here, bill depth and length are positively correlated as we can see from the positive line

Bivariate → **two** features

Stacked barcharts compare two **categorical** features

y: **count** | x: **species** | color: **sex**

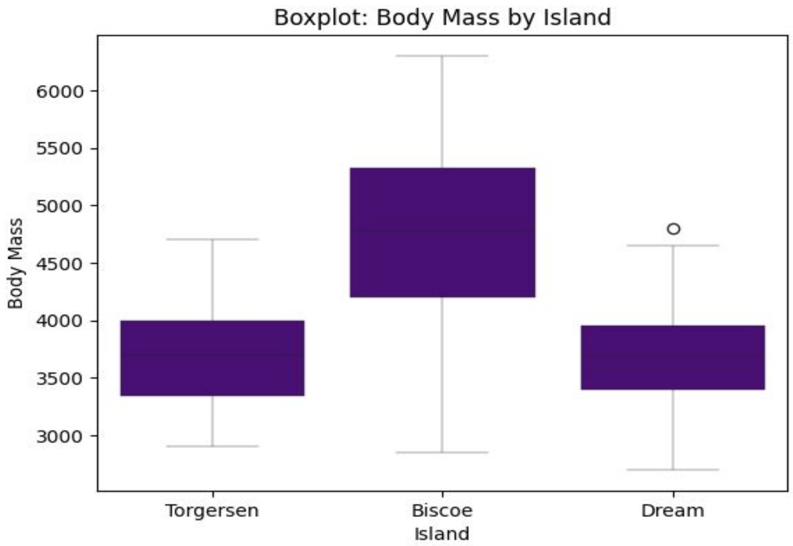


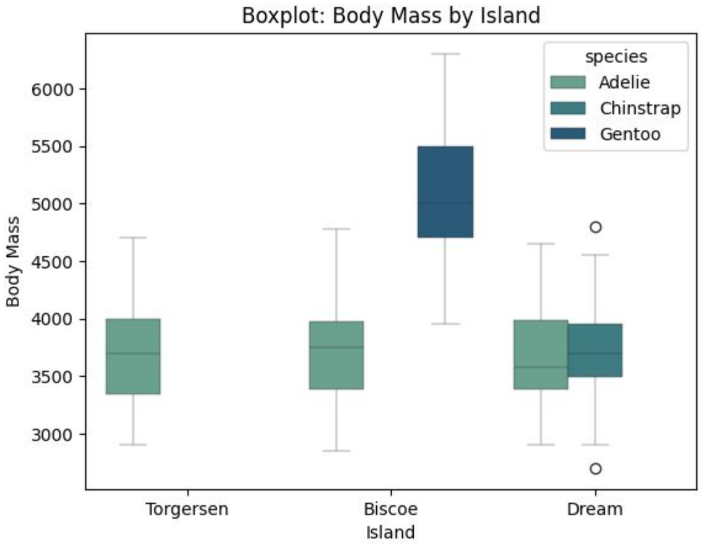

```
pivot_data = penguins.pivot_table(index='species',
                   columns='sex',
                   aggfunc='size',
                   fill_value=0)

pivot_data.plot(kind='bar', stacked=True,
                color= ['green', 'tab:olive'])
```

take note difficult to compare sizes (not exact)

Boxplots compare the distribution, skewness, and/or outliers of a single **numerical** feature to 0+ **categorical** features

y **body mass** x **island** color **species**






```
sns.boxplot(data = penguins, x = 'island',y ='body_mass_g',
            hue = 'species', palette = 'crest', linewidth=0.3)
```

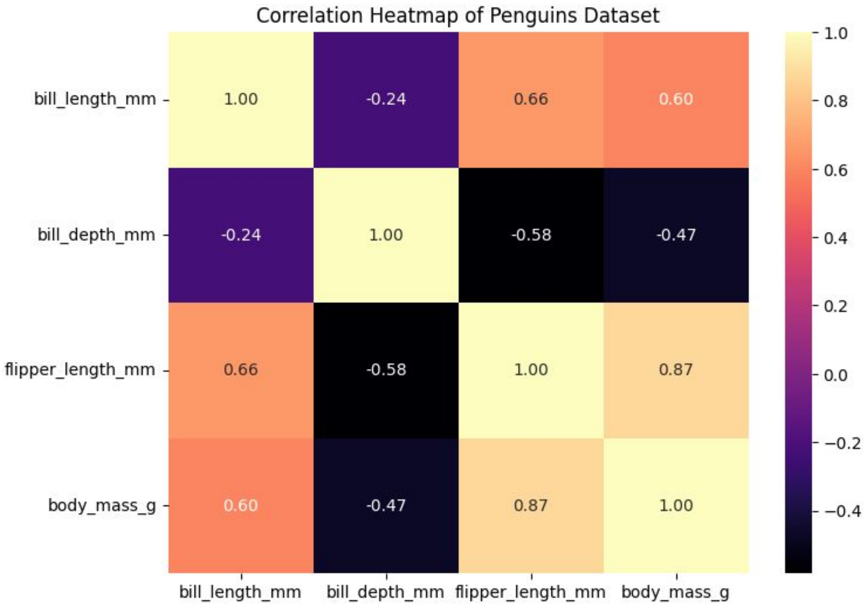
take note requires interpretation and needs categorical feature to compare with

Multiivariate → many features

Heatmaps show correlation between all **numerical** features

numerical | **numerical** | **numerical**

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
bill_length_mm	1.000000	-0.235053	0.656181	0.595110
bill_depth_mm	-0.235053	1.000000	-0.583851	-0.471916
flipper_length_mm	0.656181	-0.583851	1.000000	0.871202
body_mass_g	0.595110	-0.471916	0.871202	1.000000



take note requires interpretation and not great for small and/or sparse datasets

```
num_penguins = penguins.select_dtypes(include = ['float64', 'int64'])
corr_matrix = num_penguins.corr()
sns.heatmap(corr_matrix, annot=True, cmap='magma', fmt='.2f', cbar=True)
```

Multii<u>variate</u> → **many** <u>features</u>

Pairplots show pairwise relationships between all **numerical** features

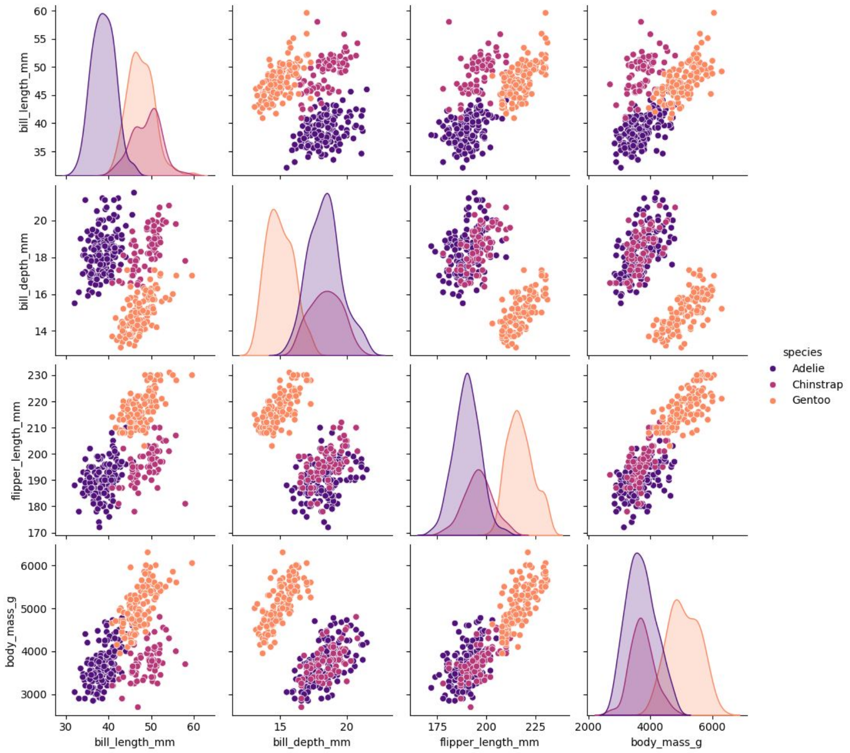
numerical | **numerical** | **categorical**

color by categorical feature

Typical to start here, then do some statistical tests or further visualizations

take note

requires interpretation and not great for small and/or sparse datasets




```
sns.pairplot(penguins, hue='species',palette='magma', dropna=True)
```

Your turn!

Below are some pairings of features. Decide which chart type and/or summary would be best.

Once you have selected your charts, pick two to create.

Opening the slide deck in a new tab may be helpful.

set A

island

set B

flipper_length_mm

set C

flipper_length_mm

bill_length_mm

set D

bill_length_mm

island

set E

island

flipper_length_mm

bill_length_mm

MACHINE LEARNING

WITH PYTHON

PC 1

Three Datasets

PC 1

Identify three data sets that you're interested in. Prioritize using the datasets listed in the project overview (see course site).

Post on the Forum under "Three datasets of interest".

Include the following for each dataset selected: a link to a repo, a two sentence description, and the reasoning behind why you find that dataset interesting.

Keep this informal as I'm curious to see what types of datasets folks are interested in.

LAB WORK TIME

Lab1

View on our website

LAB WORK TIME

Lab1

[View on our website]