EDA

COGS 108 Fall 2025
Jason Chen
Week 5

xic007@ucsd.edu

OH: Tue 3-5 pm

Due dates

- Project Proposal: TODAY (10/29)
- D4: Friday (10/31)

Project Proposal

- Due: TODAY
- Just make sure you've pushed your completed Project Proposal to your github group repo by 11:59 pm
  - Nothing else to submit

Project Proposal

- Work with your group to make a strong proposal
- Practice your git/github commands and strategies
- Use ReviewNB to look at changes between jupyter notebooks in Git
- Follow the instructions fully!

D4: Descriptive AND Exploratory Data Analysis

Part I : Data & Wrangling

df['height'] = df['height'].apply(function)

The `apply()` method allows you to apply a function along one of the axis of the DataFrame, default 0, which is the index (row) axis.

Syntax

```
dataframe.apply(func, axis, raw, result_type, args, kwds)
```

Parameters

The `axis`, `raw`, `result_type`, and `args` parameters are keyword arguments.

Parameter	Value	Description
func		Required. A function to apply to the DataFrame.

axis	`0`
	`1`
	`'index'`
	`'columns'`
		Optional, Which axis to apply the function to. default 0.

raw	`True` `False`	Optional, default False. Set to true if the row/column should be passed as an ndarray object

result_type	`'expand'`
	`'reduce'`
	`'broadcast'`
	`None`
		Optional, default None. Specifies how the result will be returned

args	*a tuple*	Optional, arguments to send into the function
kwds	*keyword arguments*	Optional, keyword arguments to send into the function

Part II : EDA

fig =
sns.pairplot(df[['column1',
'column2']]);

A pairplot is useful because it allows you to visualize the relationship between multiple variables in a dataset at once.

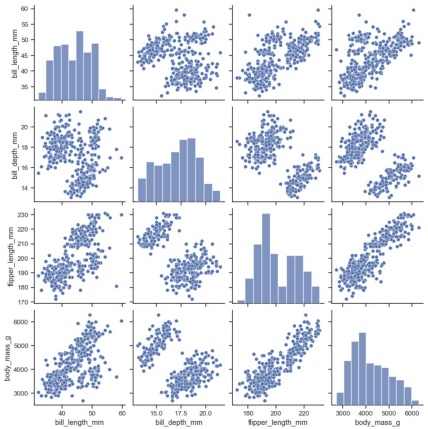
```
seaborn.pairplot(data, *, hue=None, hue_order=None, palette=None, vars=None,
x_vars=None, y_vars=None, kind='scatter', diag_kind='auto', markers=None,
height=2.5, aspect=1, corner=False, dropna=False, plot_kws=None, diag_kws=None,
grid_kws=None, size=None)
```

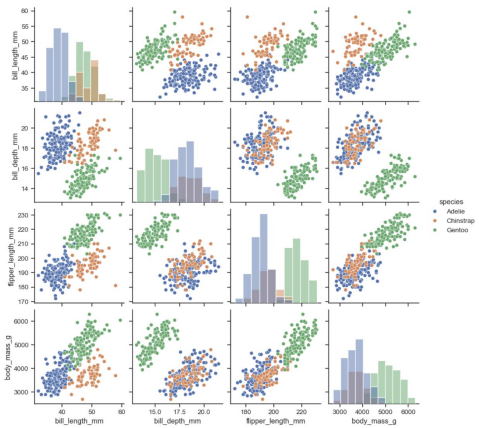
Plot pairwise relationships in a dataset.

By default, this function will create a grid of Axes such that each numeric variable in `data` will by shared across the y-axes across a single row and the x-axes across a single column. The diagonal plots are treated differently: a univariate distribution plot is drawn to show the marginal distribution of the data in each column.

It is also possible to show a subset of variables or plot different variables on the rows and columns.

This is a high-level interface for `PairGrid` that is intended to make it easy to draw a few common styles. You should use `PairGrid` directly if you need more flexibility.





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

Part III : Pairwise analysis

T-tests are used to check whether the unknown population means of given pair of groups are equal.

Null hypothesis: the means of two groups are equal

t_val, p_val = ttest_ind(df_1, df_2)

scipy.stats.ttest_ind

```
scipy.stats.ttest_ind(a, b, axis=0, equal_var=True, nan_policy='propagate',
permutations=None, random_state=None, alternative='two-sided', trim=0, *, keepdims=False)
```

[source]

Calculate the T-test for the means of *two independent* samples of scores.

This is a test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances by default.

Parameters: **a, b** : *array_like*

The arrays must have the same shape, except in the dimension corresponding to *axis* (the first, by default).

axis : *int or None, default: 0*

If an int, the axis of the input along which to compute the statistic. The statistic of each axis-slice (e.g. row) of the input will appear in a corresponding element of the output. If `None`, the input will be raveled before computing the statistic.

equal_var : *bool, optional*

If True (default), perform a standard independent 2 sample test that assumes equal population variances [1]. If False, perform Welch's t-test, which does not assume equal population variance [2].

New in version 0.11.0.

nan_policy : *{'propagate', 'omit', 'raise'}*

Defines how to handle input NaNs.

- `propagate`: if a NaN is present in the axis slice (e.g. row) along which the statistic is computed, the corresponding entry of the output will be NaN.
- `omit`: NaNs will be omitted when performing the calculation. If insufficient data remains in the axis slice along which the statistic is computed, the corresponding entry of the output will be NaN.
- `raise`: if a NaN is present, a `ValueError` will be raised.

Ref: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

Part III : Pairwise analysis

Cohen's d goes beyond "significant or not" and measure how big the difference is.

Cohen's d = $(\text{mean}_1 - \text{mean}_2)$ / pooled SD

Describes the magnitude of difference in standard-deviation units.

Part III : Pairwise analysis

One-way Anova tests whether more than two means differ.

ANOVA = Analysis of Variance

Tests the null hypothesis that all group means are equal.

F-statistic = between-group variance / within-group variance.

Large F → Between-group variance is much bigger than within-group variance,
meaning the group means are likely different.

statsmodels.stats.oneway.anova_oneway

```
statsmodels.stats.oneway.anova_oneway(
    data,
    groups=None,
    use_var='unequal',
    welch_correction=True,
    trim_frac=0
)                                                    [source]
```

Part III : Pairwise analysis

Tukey's HSD is a post-hoc analyses conducted after a significant ANOVA, and finds which groups differ.

HSD = Honestly Significant Difference

Performs pairwise comparisons between all group means.

Adjusts p-values to control family-wise error rate (avoids false positives).

statsmodels.stats.multicomp.pairwise_tukeyhsd

```
statsmodels.stats.multicomp.pairwise_tukeyhsd(
    endog,
    groups,
    alpha=0.05,
    use_var='equal'
)                                                    [source]
```

Section Materials

Section materials can be accessed at:
https://github.com/JasonC1217/COGS108_FA25_B07-B08

THANKS!

Questions on EdStem or office hours

Office hours: Tue, 3-5 PM