

dm25s1

Topic 05 : Exploratory Data Analysis2

Part 02 : EDA Visualisation

Dr Bernard Butler

Department of Computing and Mathematics, WIT.
(bernard.butler@setu.ie)

Autumn Semester, 2025

Outline

- Selection of seaborn plots and advanced settings

Data Mining (Week 5)

Introduction

Motivating Example

Preparation

Data Handling

Exploring Data 1

Exploring Data 2

Building Models

Prediction

Clustering

Regression
1

Classification
1

Regression
2

Classification
2

Wrap up

EDA Visualisation — Summary

1. Introduction

2. Visualisation - selection of seaborn plots

3. Resources

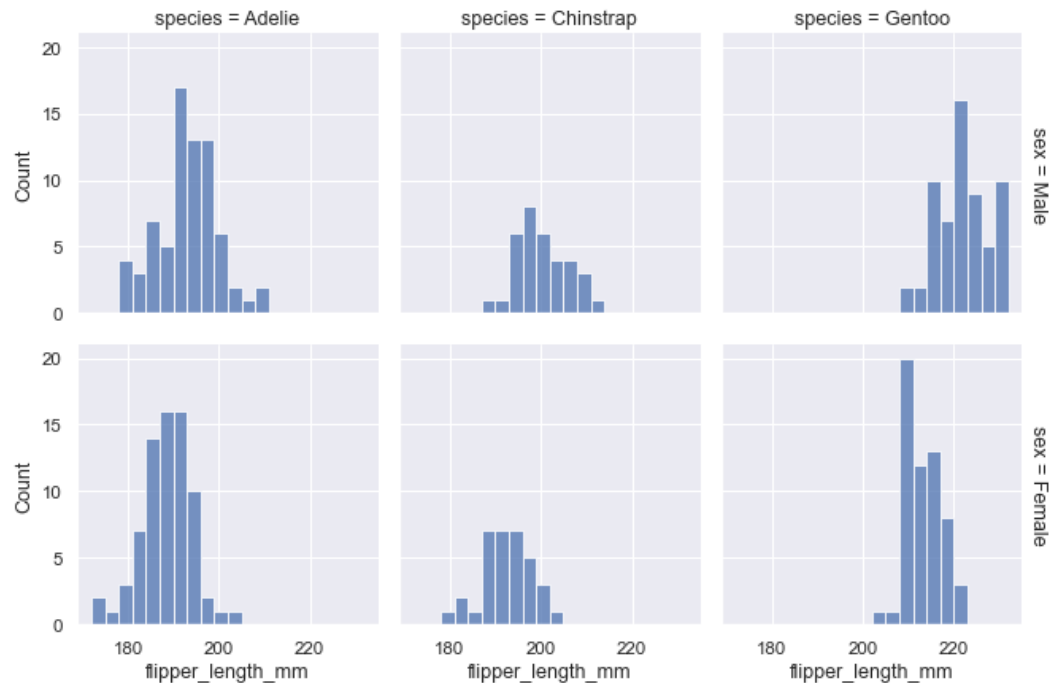
Need for More Advanced Plots

- The basic plot types allow us to visualise distributions and relationships
- But they have limitations if we wish to
 - Show a single plot with related elements rather than multiple hardcoded plots
 - Show that predictions from the model have uncertainty that varies over the range
 - Show the relationship between the *distributions* of 2 numerical variables
 - Generalise boxplots to show more distribution information, not just the quartiles
 - Plot a combination of 3 or more categorical and/or numerical variables in 2-D
 - Compare a selection of data instances over a selection of variables
 - Use attributes like colour, size or shape to convey information on categorical variables

Selected seaborn-based visualisations

- We could easily spend several weeks on EDA visualisation
- There is a long history of visualisation, from infographics to bubble plots
- Seaborn provides a gallery of data science-related visualisation examples
- We consider a selection today that are useful in practice
- Take a look at the [seaborn examples gallery](#) for more inspiration...

Histograms with facets



Source: https://seaborn.pydata.org/examples/faceted_histogram.html

What it does

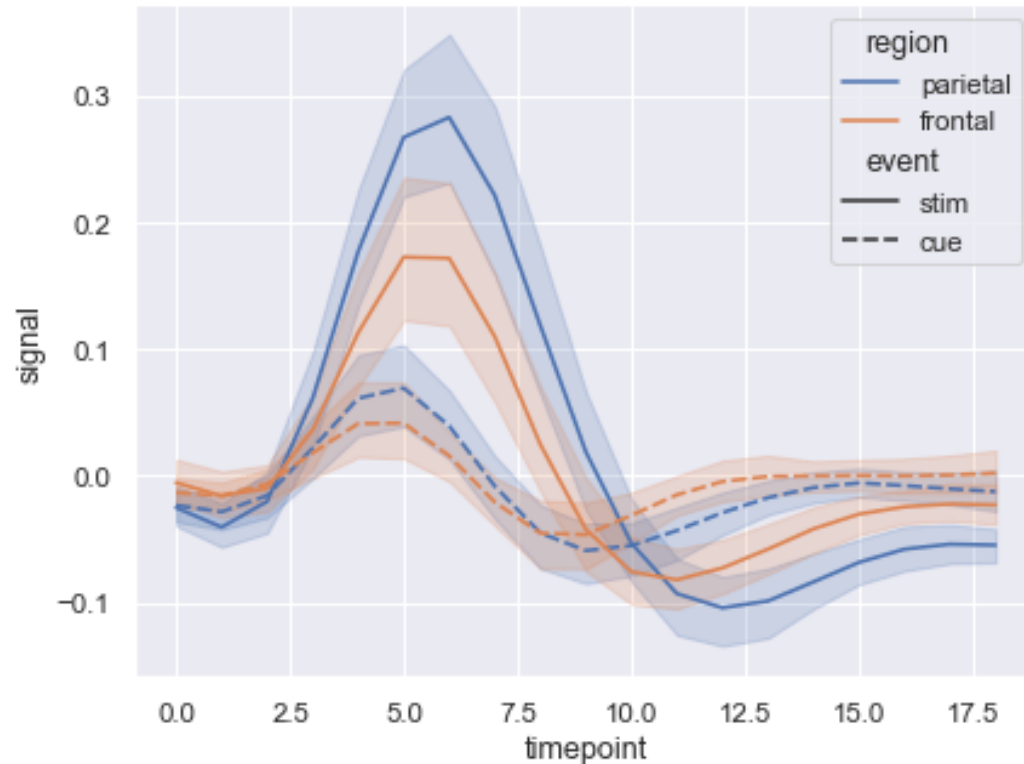
- Facets: show a grid of related plots
- Conditioned by 1 or 2 categorical variables
- Here: flipper length of penguins, by sex \times species.

When to use it

- Have a key variable, represented by a suitable plot
- Wish to view dependence on 1 or 2 categorical variables in same plot group

Line plots with error bands

Source: https://seaborn.pydata.org/examples/errorband_lineplots.html



What it does

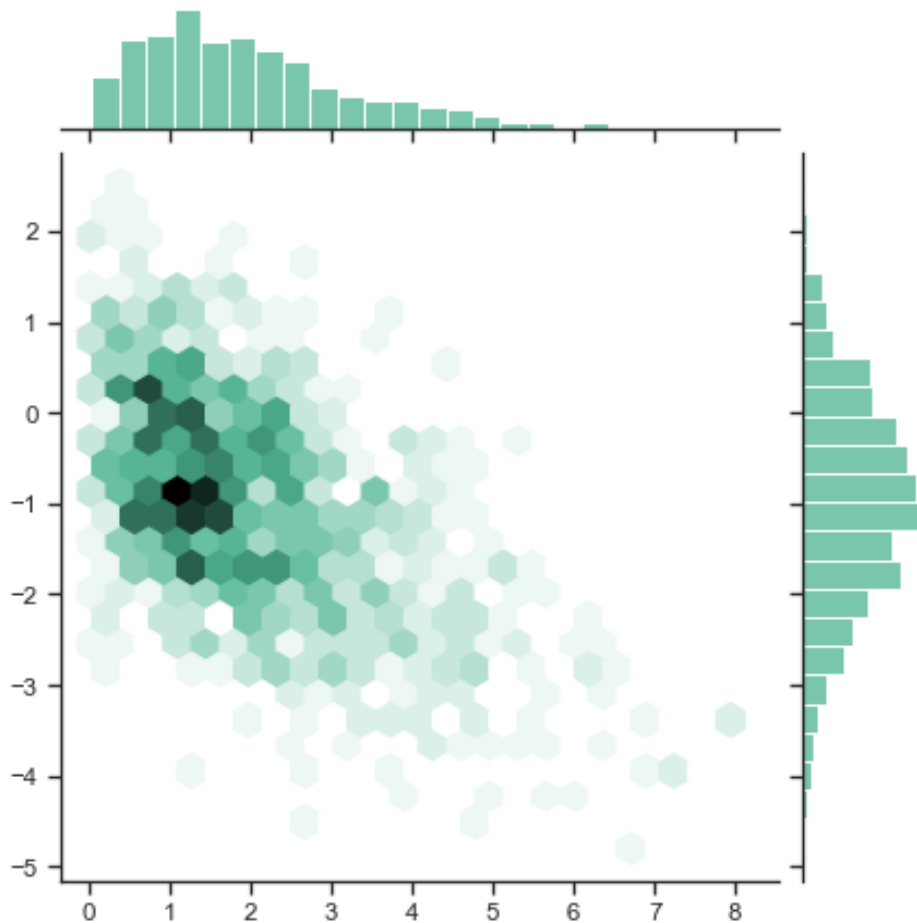
- Multiple numeric variables as lineplots
- Use of colour and linetype
- Overlaid on error bands

When to use it

- Multiple numeric variables on same scale
- Highlight uncertainties

Binning with distribution plots

Source: https://seaborn.pydata.org/examples/hexbin_marginals.html



What it does

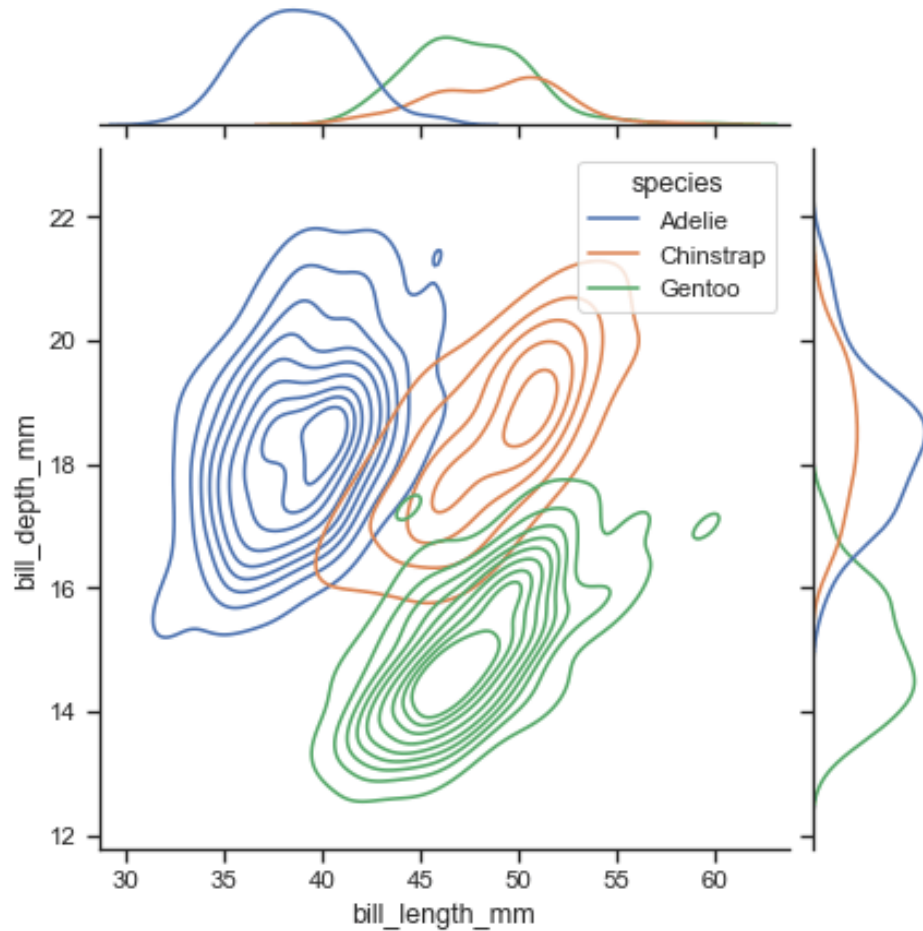
- Compare histograms of 2 numeric columns
- Binning provides a heatmap

When to use it

- Interested in the co-occurrence of 2 numeric columns
- Columns are correlated, wish to understand this

Contour plots of distributions

Source: https://seaborn.pydata.org/examples/joint_kde.html



What it does

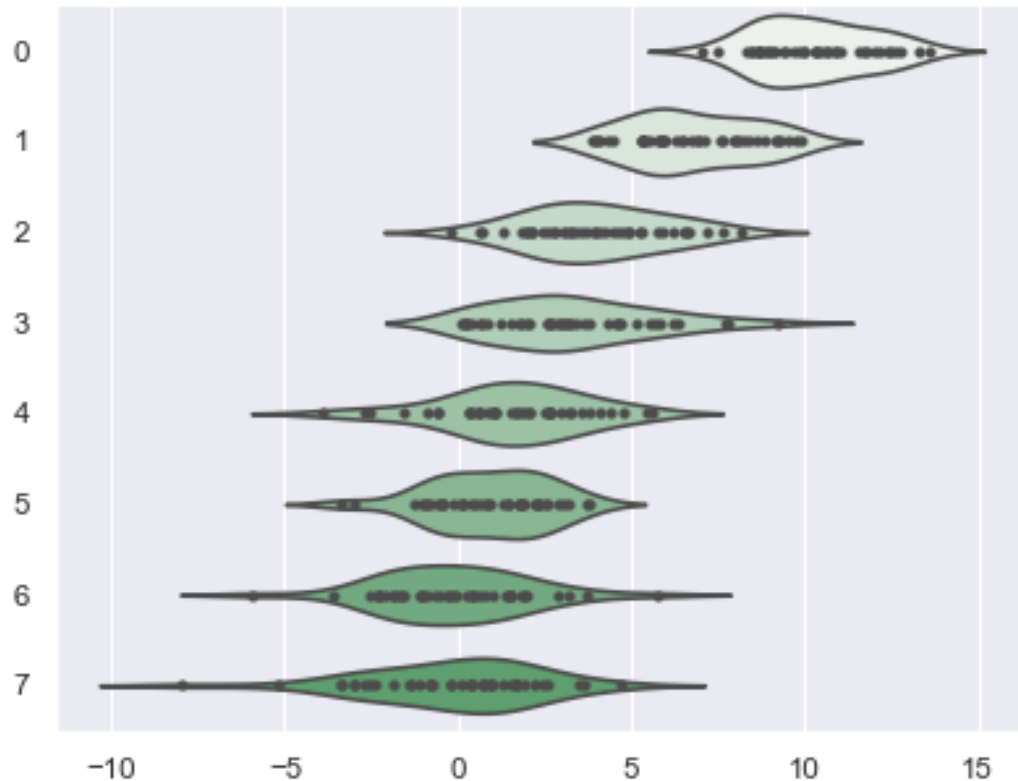
- Penguin bill length \times bill width per species
- Two ways of showing distributions

When to use it

- 2 numeric features, split by 1 categorical feature

Violin plots

Source: https://seaborn.pydata.org/examples/simple_violinplots.html



What it does

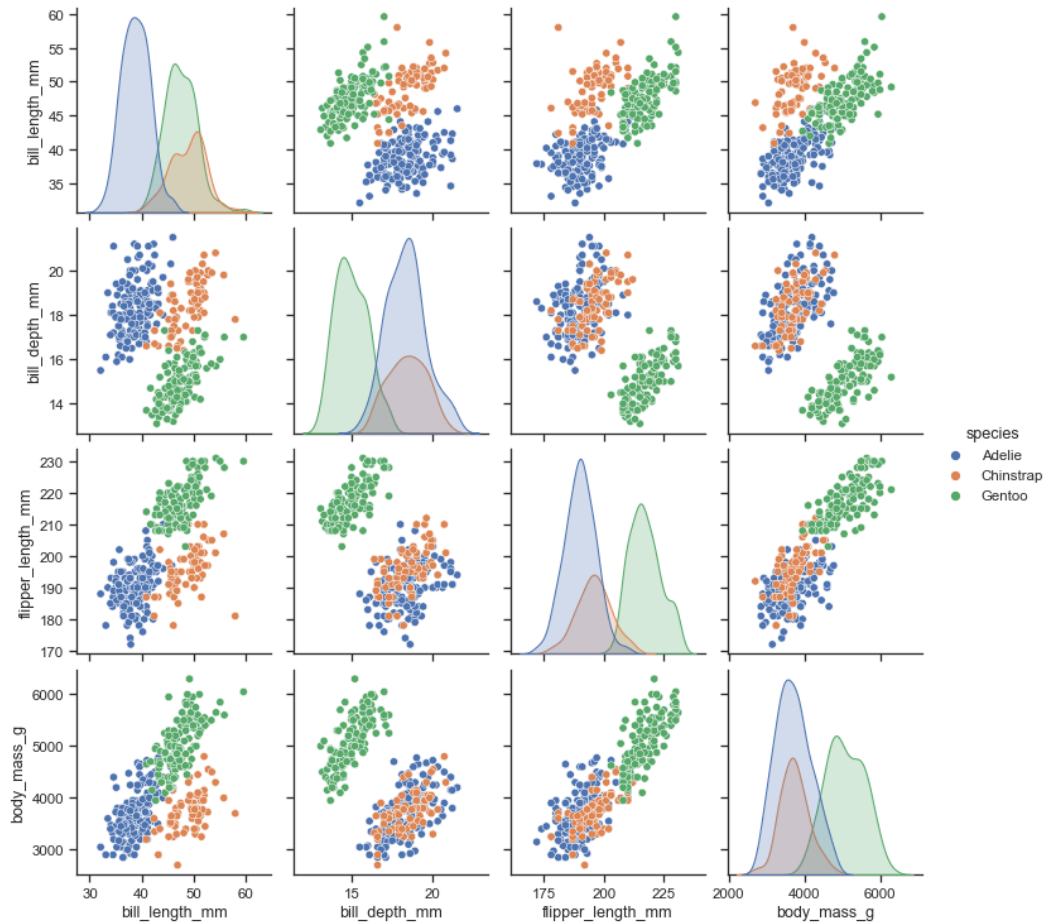
- Numeric variable, split by category
- Alternative to boxplot
- data points shown here

When to use it

- Numeric attribute by categorical feature
- Interested in the shape of the distribution

Scatterplot matrix

Source: https://seaborn.pydata.org/examples/scatterplot_matrix.html



What it does

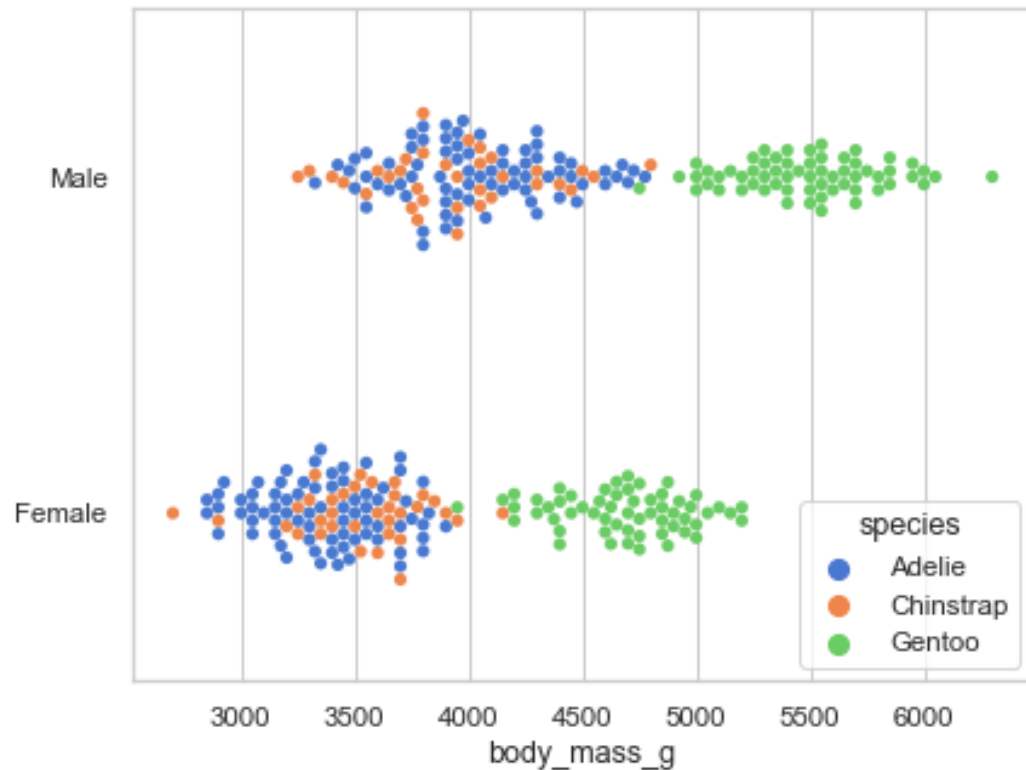
- Penguin data - 4 numeric features (bill length, bill depth, flipper length, body mass), 1 categorical feature (species) with 3 levels
- All combinations shown

When to use it

- Look at many numeric variables together
- Can use colour or other indicator to show categorical variable

Scatterplot with categorical variables

Source: https://seaborn.pydata.org/examples/scatterplot_categorical.html



What it does

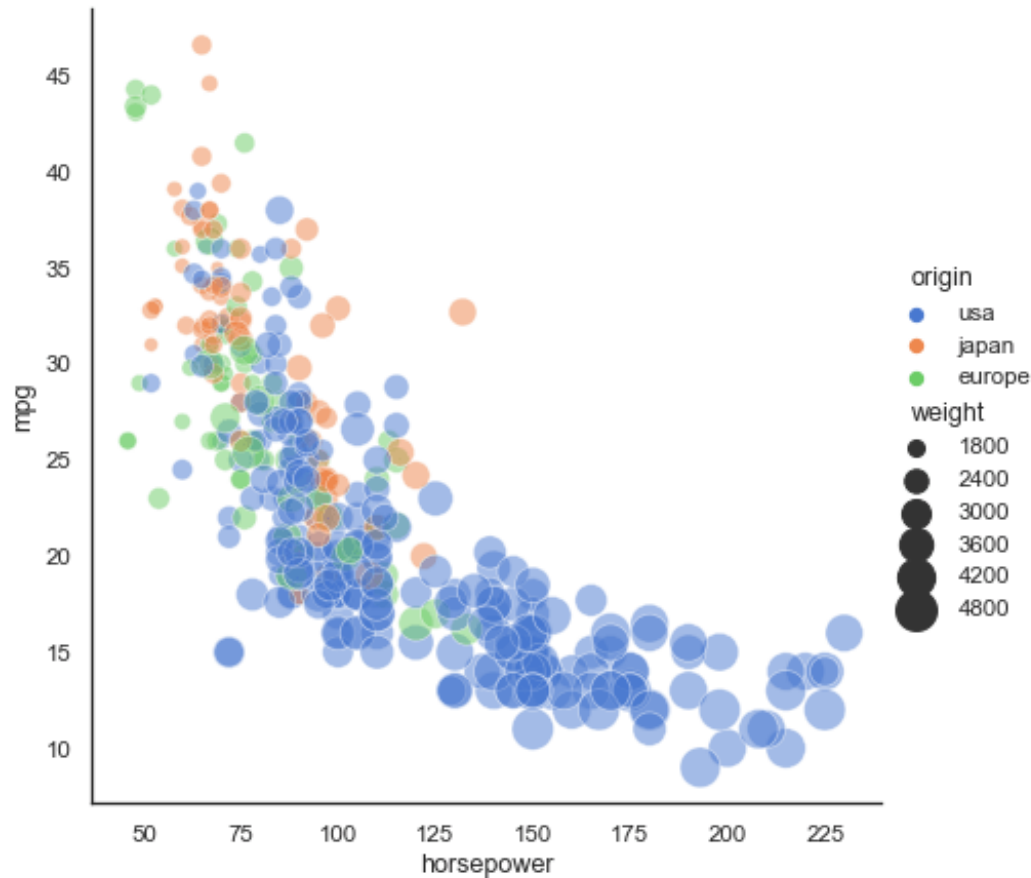
- Show numerical variable in terms of 1 or more categorical variables

When to use it

- More detailed alternative to violinplot

Scatterplot with bubbles

Source: https://seaborn.pydata.org/examples/scatter_bubbles.html



What it does

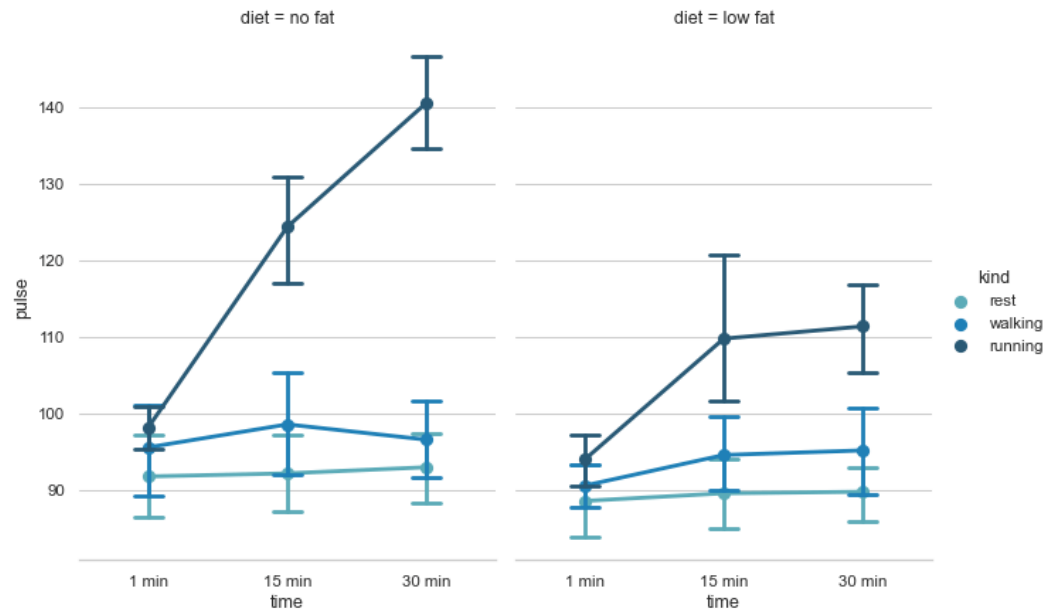
- Auto mpg data, mpg \times horsepower
- Plot features represent categorical features
- Note grouping of numeric variable to create categories

When to use it

- Represent multiple categorical variables in terms of 2 numerical variables

Pointplot for Analysis of Variance

Source: https://seaborn.pydata.org/examples/pointplot_anova.html



What it does

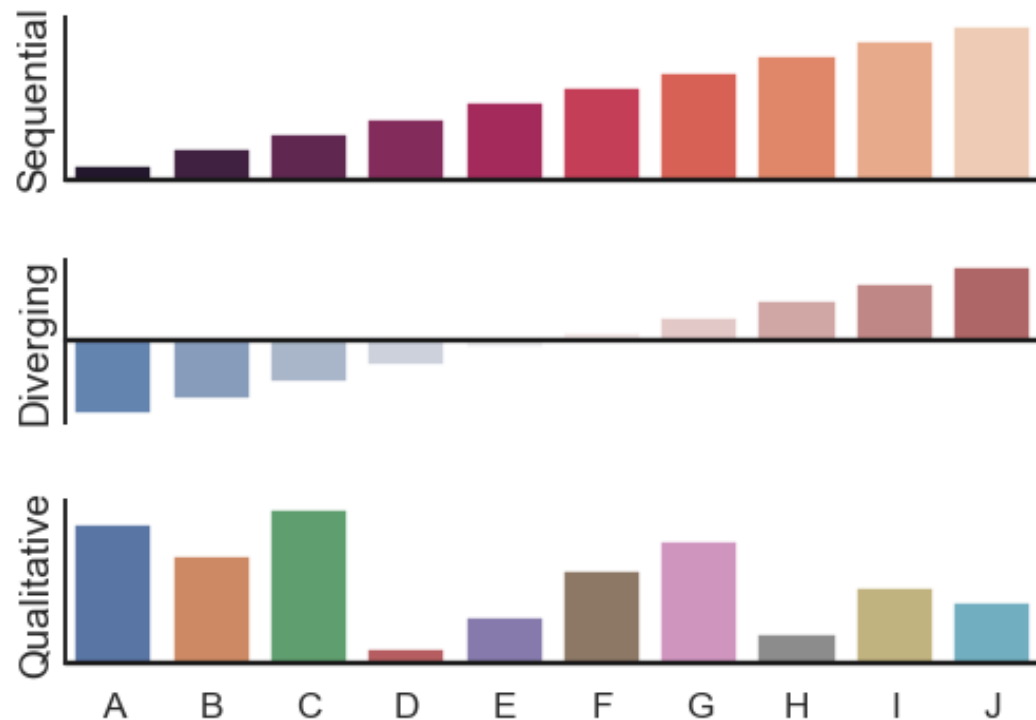
- Trend in pulse rates, by time \times activity (ordered categories)
- Rich plot, with drill down capability

When to use it

- Numeric target as function of multiple categorical features

Colour palettes

Source: https://seaborn.pydata.org/examples/palette_choices.html



What it does

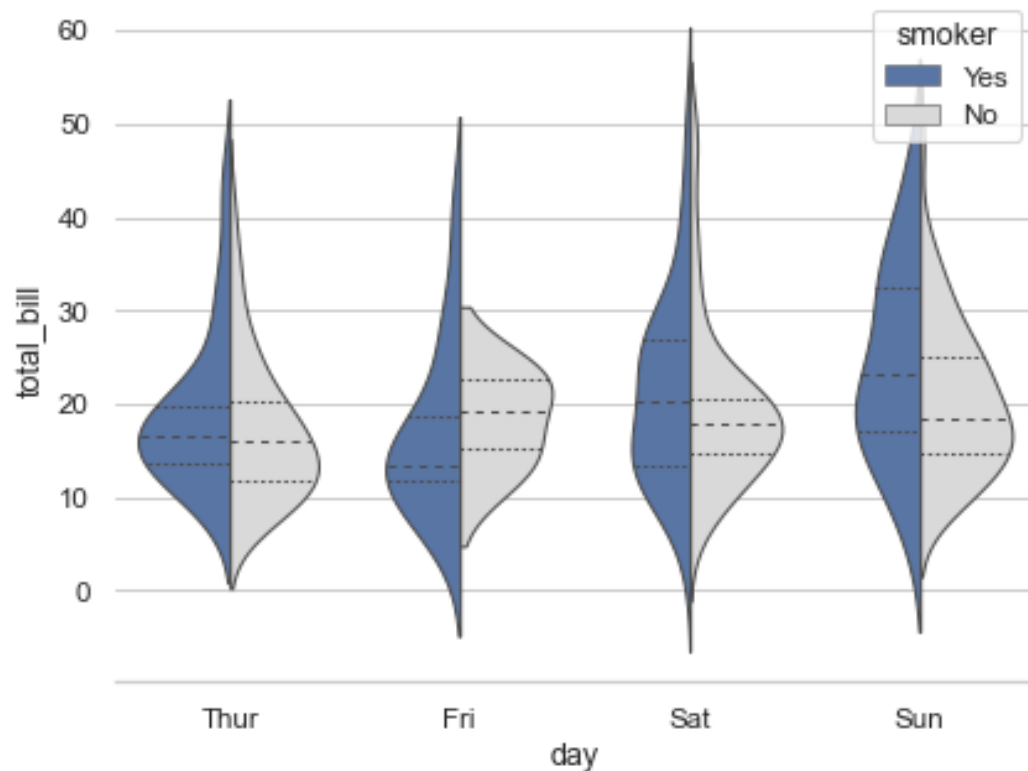
- Options for choosing palettes
- Qualitative, Sequential, Diverging

When to use it

- Qualitative: unordered categorical variable
- Sequential: ordered categorical variable
- Diverging: ordered sequential variable

Grouped Violinplots

Source: https://seaborn.pydata.org/examples/grouped_violinplots.html



What it does

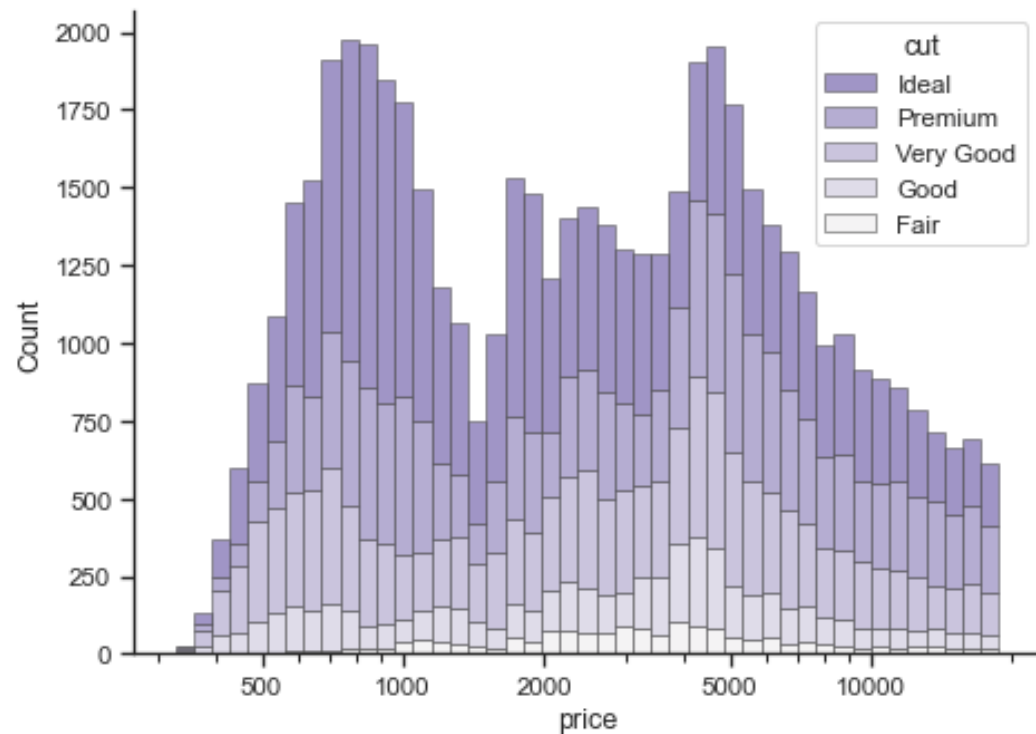
- Tips data: total_bill by day × smoker
- Note splits in “violins” to accomodate a category

When to use it

- Adding a second categorical variable to violinplot
- Alternative to faceting

Stacked histograms

Source: https://seaborn.pydata.org/examples/histogram_stacked.html



What it does

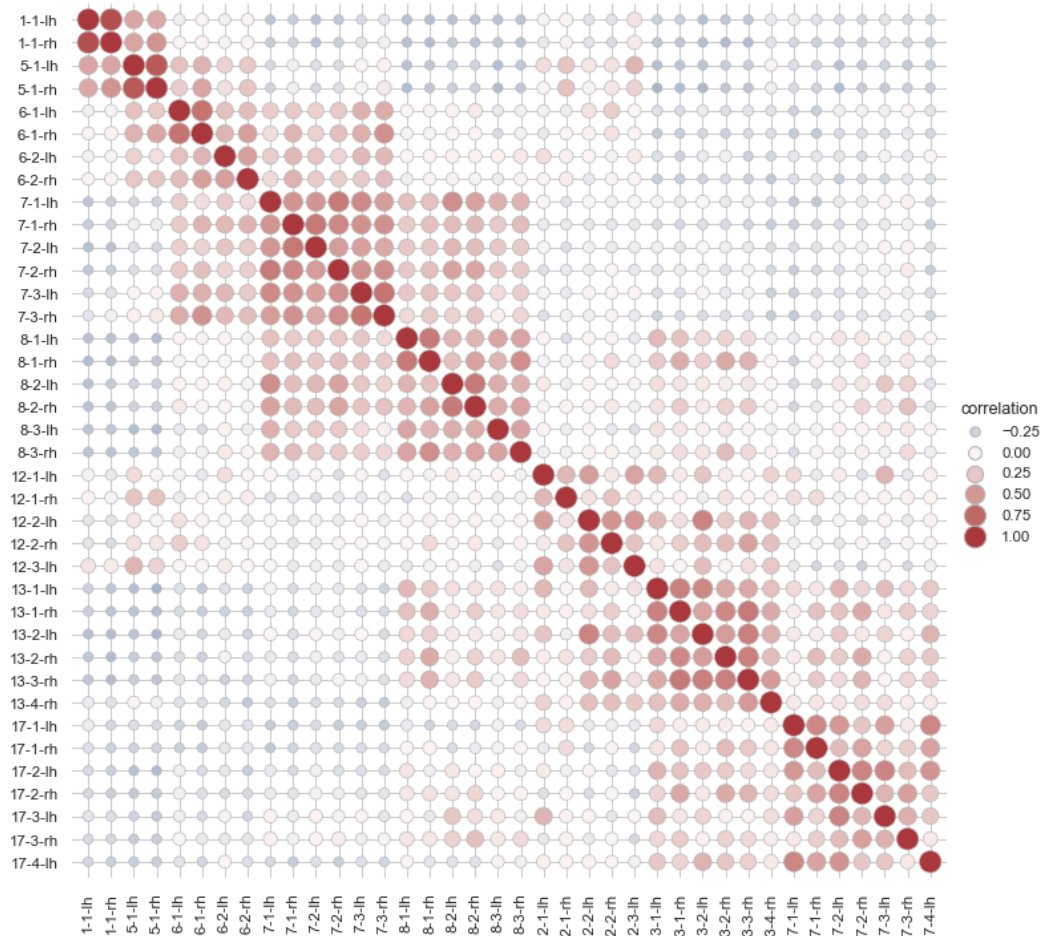
- Diamond valuation data
- Show distribution of price by cut
- Be careful: stacked not overlaid!

When to use it

- Can compare histograms by category variable
- Alternative to faceting

Heatmap with scatterplot

Source: https://seaborn.pydata.org/examples/heat_scatter.html



What it does

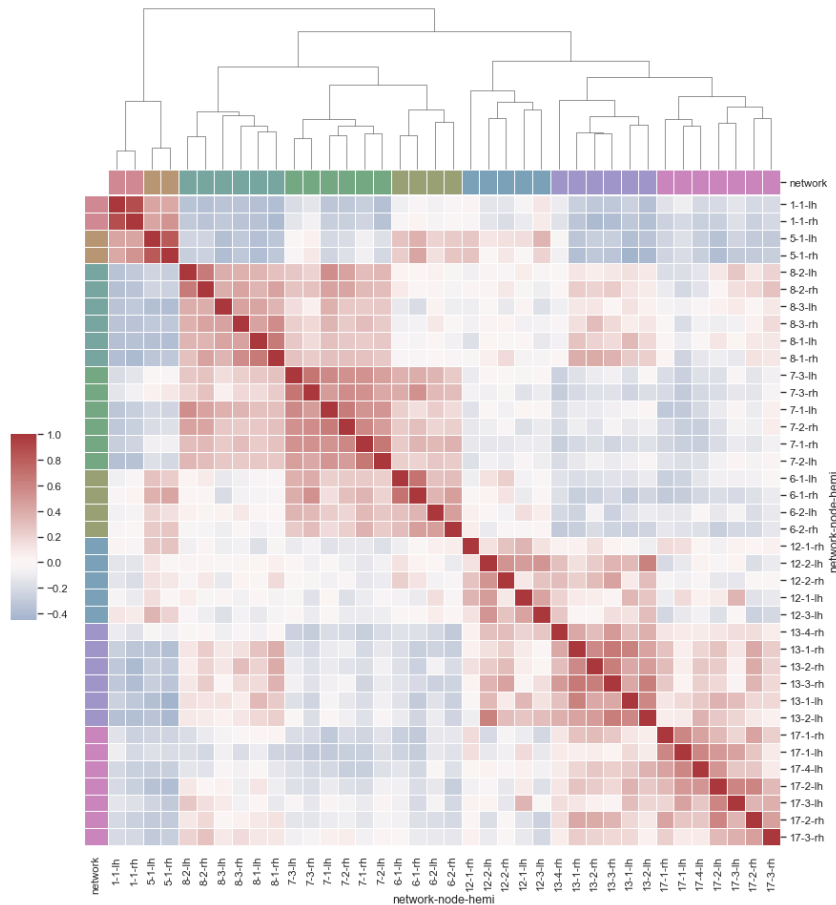
- Network data
- Highlighting correlated flows
- Use of colour and size of bubbles

When to use it

- Emphasise sign and magnitude of correlations

Heatmap with dendrogram

Source: https://seaborn.pydata.org/examples/structured_heatmap.html



What it does

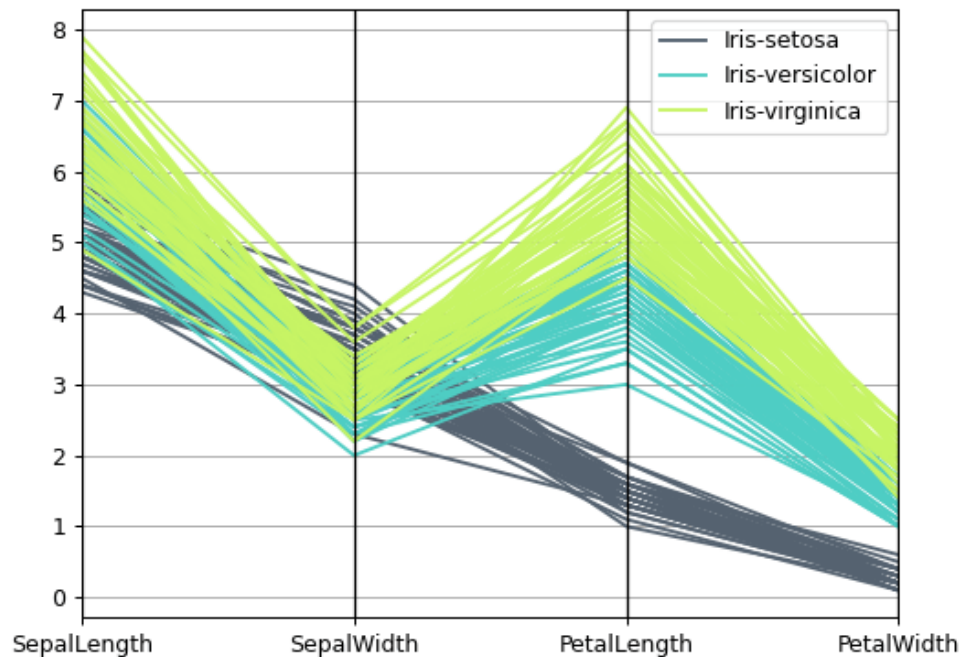
- Heatmap of correlations
- Dendrogram clusters them to highlight similar values

When to use it

- Need to identify groups of correlated numerical variables

Parallel coordinate plots

Source: https://pandas.pydata.org/docs/reference/api/pandas.plotting.parallel_coordinates.html



What it does

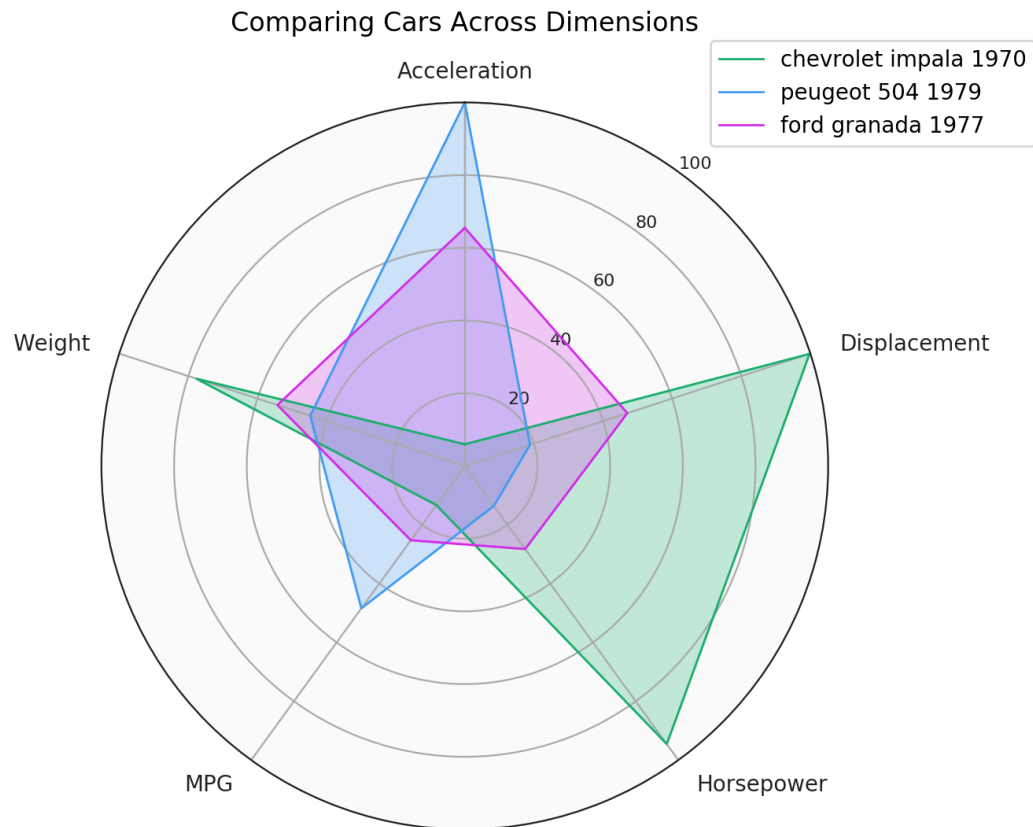
- Each piecewise linear “line” represents an instance
- Each vertical split (context) line represents a numerical variable (feature or target)
- Instance lines pass through values they take on the context variables

When to use it

- Need to compare subsets of instances (note use of colour to distinguish)
- Compare instances based on a (subset) of their numerical values

Radar charts

Source: <https://www.pythoncharts.com/matplotlib/radar-charts/>



What it does

- Each polygon represents an instance, colour legend identifies the instance
- Each radial line represents a numerical variable
- Vertices of the polygon indicate the value an instance takes on that variable

When to use it

- Have a small number of instances to compare over selected numerical variables
- Visualise correlations between numerical variables, for selected instances

Plot aesthetics

- A well-designed plot looks attractive and can provide much more information, with less viewer effort
- Here are some general tips
 - Use a suitable choice of colour palette - see [this guidance from the seaborn project](#) and related references.
 - Many seaborn plots offer a setting `hue =`, which is particularly helpful when set to a categorical feature or target column.
 - Lines and marker symbols can all be configured, so some lines could be dashed or symbols can have different sizes
 - Legends can be used to explain what the line and symbol variations mean.
 - Often two or more plot types side-by-side can be more informative, e.g., a qq-plot beside a histogram
 - Side-by-side plots can also be used to compare behaviour across settings
 - 2-D plot groupings (arranged as, say, 2×2 or 2×3 can help to compare 2 different settings
 - Every plot should have a title and axes should be labeled

Resources

Resources

Resources

Guides

- 1 hour, Youtube on generating seaborn plots — excellent (but wrong on interpretation of box plot)
www.youtube.com/watch?v=6GUZXDef2U0&t=1363s

Articles on Exploratory Data Analysis

- Exploratory Data Analysis (EDA) and Data Visualization with Python
www.kite.com/blog/python/data-analysis-visualization-python/
- When Should You Delete Outliers from a Data Set?
humansofdata.atlan.com/2018/03/when-delete-outliers-dataset

Visualisation

- (Seaborn) Example Gallery
<https://seaborn.pydata.org/examples/index.html>