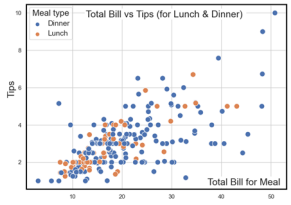
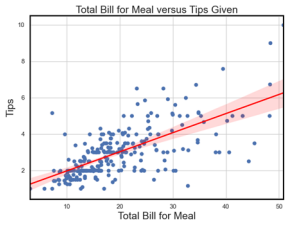
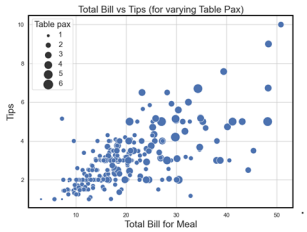
Chapter 5.3 – Visualising Relationships in Data

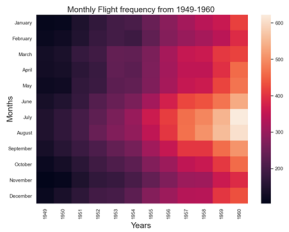
Contents

- Basic Relation Plots
- Scatter Pots
- Bubble Plots
- Heatmaps





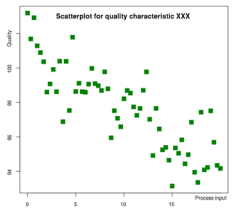




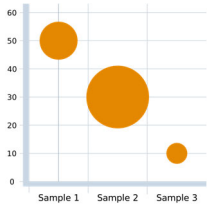
Basic Relation Plots

Seeing Relationship Among Variables

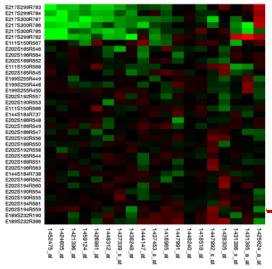
- Relation plots help show relationships or patterns among variables.
- **Scatter and Bubble plots** – Scatter plots are useful to visualise correlation between two variables for one or more groups. The value of an additional third variable can be encoded by the dot size in a bubble plot.
- **Heatmaps** – When one or more of these variables are non-continuous or non-numeric, a heatmap can be used. Relationships between categories are visualised through a grid of values encoded in different colours.



Scatter Plot



Bubble Plot

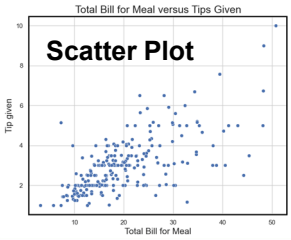


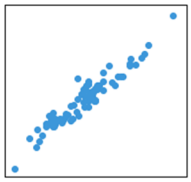
Heatmap

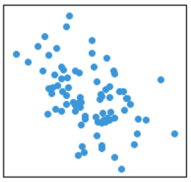
Scatter Plots

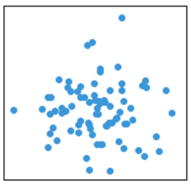
Seeing Relationship in the Dots

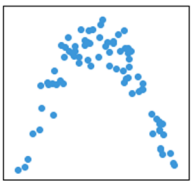
- Scatter plots use the 2-D position of dots to represent values of two different numeric variables. The collection of these scattered dots can reveal **correlational relationships** or **patterns** between these two variables.
- Given a horizontal value (independent variable), a highly **correlated** relationship can readily predict the vertical value (dependent variable) and such pairwise relationships can be described in many ways (e.g. positive/negative, strong/weak, linear/non-linear)[1]




strong, positive (linear)


moderate, negative (linear)


no relationship (uncorrelated)

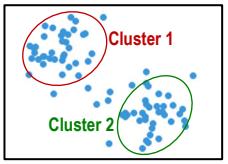

strong (non-linear)

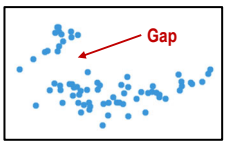
[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/

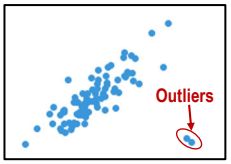
Scatter Plots

Seeing Groupings in the Dots

- Scatter plots can show interesting grouping patterns of dots[1].
- **Clusters** – are distinctive groupings in the dots forming various discernible shapes. Some clusters are circular, while others may be elongated.
- **Empty regions or gaps** – are areas that have a distinctively lack of dots. These unexpected gaps can be informative as they reveal incompatible value pairs between the two variables of interest.
- **Outliers** – are dots that are far off and do not appear to belong to any cluster. These data points could be erroneously recorded or they could be rarely occurring events of special interest.





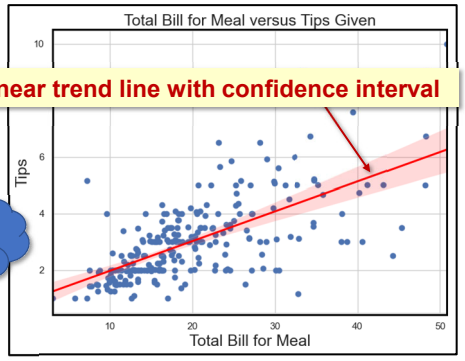


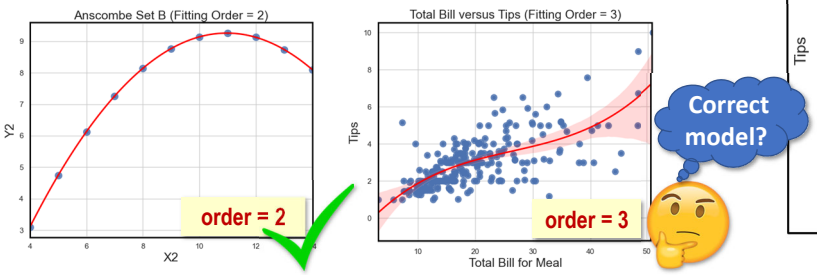
[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/

Scatter Plots

Seeing Trends in the Dots

- Trend lines can be drawn to show the mathematically **best fit** to the data.
- Seaborn's `regplot()` [2] method is used to fit a **linear regression model** to the data. The shaded area about the line is the **confidence interval**.





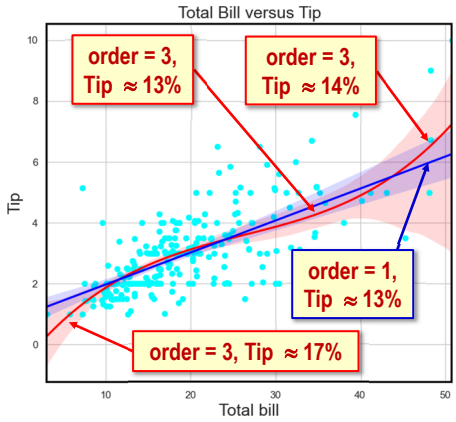
- Lines of higher polynominal order can be used if such fit makes sense to the data.

[2] Seaborn, Plot data and a linear regression model fit , https://seaborn.pydata.org/generated/seaborn.regplot.html

Scatter Plots

When a Polynomial May Not Fit

- The polynomial regression model with order=1 provides a constant estimation of a 13% tip irrespective of the cost of the meal.
- A polynomial model of order=3 predicts a varying tip, which about 17% with meals < $10, drops to about 13% when meal cost starts increasing and then goes up again to about 14% as meals gets even more expensive.
- It seems unlikely that a single polynomial can **model all the data** (irrespective of increasing total bill), especially if the data is noisy and there may exist outliers at the two extremities.

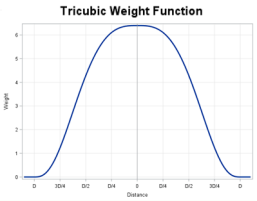


Data from "*tips.csv*"

Scatter Plots

Using LOWESS

- LOcally Weighted Scatterplot Smoothing (LOWESS) was developed by William Cleveland in 1979 at AT&T Bell Labs.
- Lowess is a **non-parametric** technique that makes no assumption about the model that should fit the relationship between the independent and dependent variable[3].
- Lowess is computed **piecewise** (i.e. point by point) as a series of **least squares linear regression** about the point using a nearest-neighbour subset of the data.
- This subset of neighbouring points used at each point is determined by a parameter (e.g. `frac` in statsmodels's version of `lowess`) that creates a window size that is a faction of all points in the dataset.
- In each regression, points within the nearest-neighbours window are **weighted** by their proximity, usually using the tri-cubic weight function given by $w(x)=(1-|d|^3)^3$.



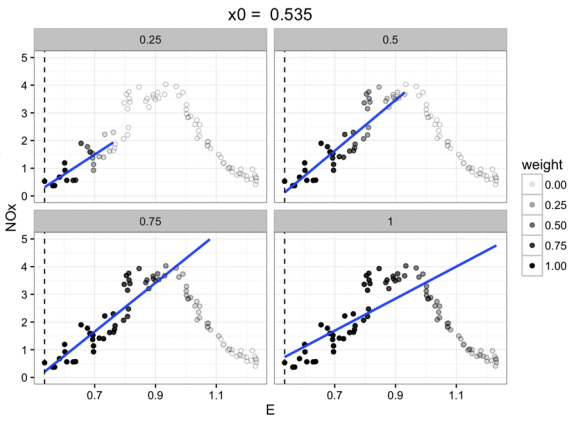
[3] J.B. Mailman, Data Smoothing in Data Science Visualization (2021), https://towardsdatascience.com/data-smoothing-for-data-science-visualization-the-goldilocks-trio-part-1-867765050615

Scatter Plots

How Does LOWESS Work?

- The `lowess` algorithms moves from **left to right** doing **piecewise** concatenation of individual least-squares linear regression using a **weighted** fraction of all nearest data points to the current point being computed.
- The Statsmodels Python module provides a useful implementation of `lowess`[4]. The size of the neighbourhood is given by `frac`, a float value between 0 to 1.
- It can be imported in Spyder with the following:

```
from statsmodels.nonparametric.smoothers_lowess import lowess
```



Animation by David Robinson from http://varianceexplained.org/files/loess.html

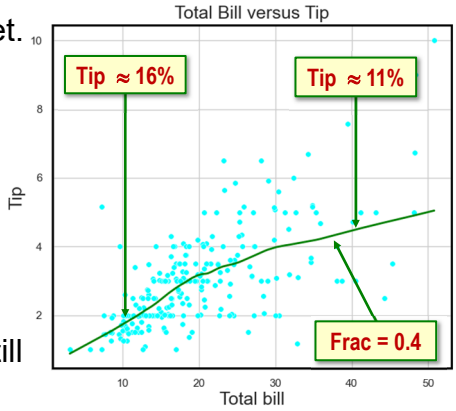
[4] Statsmodels Python library reference for its lowess method, https://www.statsmodels.org/devel/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html

Scatter Plots

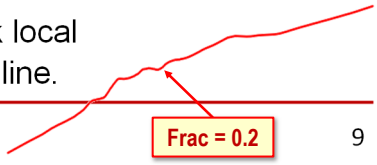
Appyling LOWESS

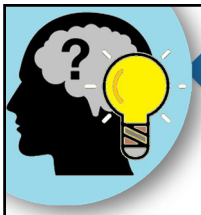
- Lowess with **`frac`**=0.4 was applied to the tips dataset.

```
LS = lowess(Tips, TotalBill, frac = 0.4)
ax = sns.lineplot(x=LS[:,0], y=LS[:,1])
```



- The results show that a tip of about 16% is given for meals < $20 but as the cost of the meal increases, the tip starts to drop to about 11%.
- By using a sizable portion (but not all) of the dataset, lowess reduces the **influence of outliers** but can still track the **changing trend** over the entire total bill.
- Reducing the fraction (**`frac`**) of data points used will track local variations better but reduces the smoothness of the trend line.

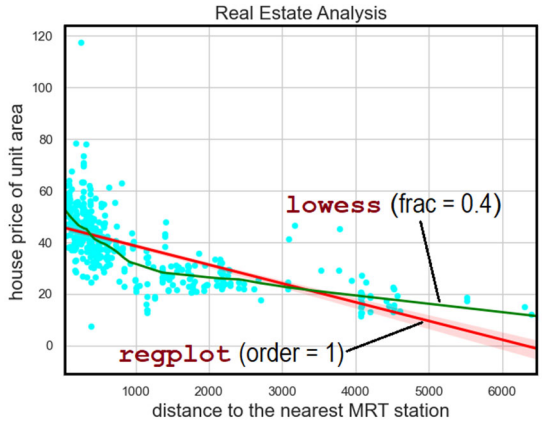




Think and Apply

What Affects Property Prices?

- A dataset of 414 property transactions that list various property-related data regarding each transaction.



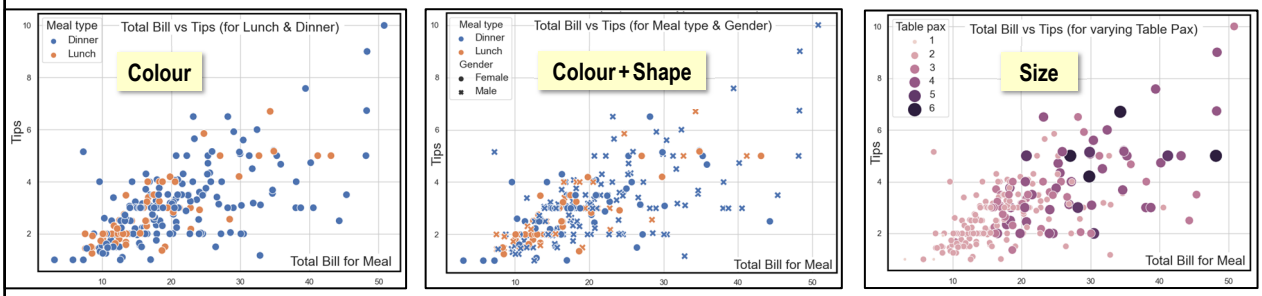
- Unit area price of the property
- Age of the property
- Distance from a MRT station
- Number of convenient stores nearby

- Better property price nearer MRT station?
- Is older cheaper?
- Does more convenience have a price?
- What really makes the difference?

Bubble Plots

Handling Additional Data Dimensions

- Additional **categorical** variables (e.g. countries, gender) can be added to the scatter plot by using points with **different nominal visual attributes** (e.g. colour or shapes)[1].
- However, if the relationship of a **third** additional variable to be depicted is **ordinal** (i.e. ordered) or **quantitative** (i.e. numeric), then the **size the dots** can be used to encode this quantitative value in a **Bubble Plot**[5].

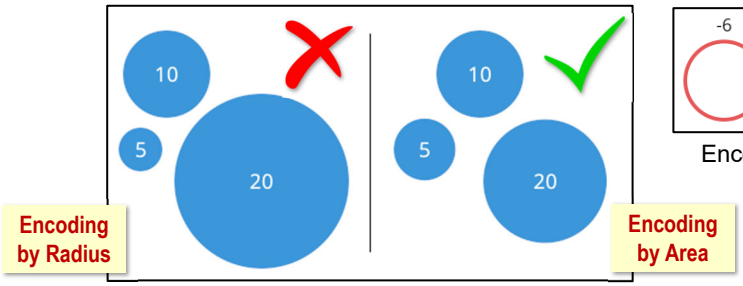


[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/
[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/

Bubble Plots

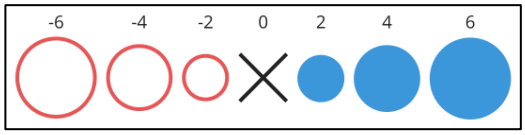
Encoding Quantitative Values to Bubble Size

- For **proportional** visual comparison, the bubble's **area** (not radius) is scaled to the value[5]. A dot with value *2n*, is only $(\sqrt{2})=1.41$ times the radius of a dot of value *n*.
- **Negative** values can be incorporated using **unfilled bubbles** with proportional areas. However, if variables with the negative ranges can be encoded in **one of the two axes** instead, that should be preferred[5].



Encoding by Radius

Encoding by Area



Encoding negative values in bubble plots[5]

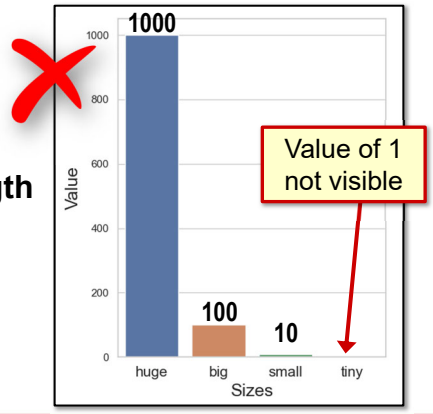
[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/

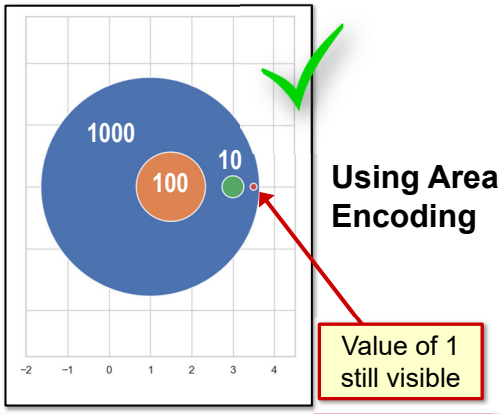
Bubble Plots

Length versus Area

- When the **differences** in magnitudes between categories are **large**, **area** encoding make better comparative visuals then value encoding using length[6].

Using Length Encoding





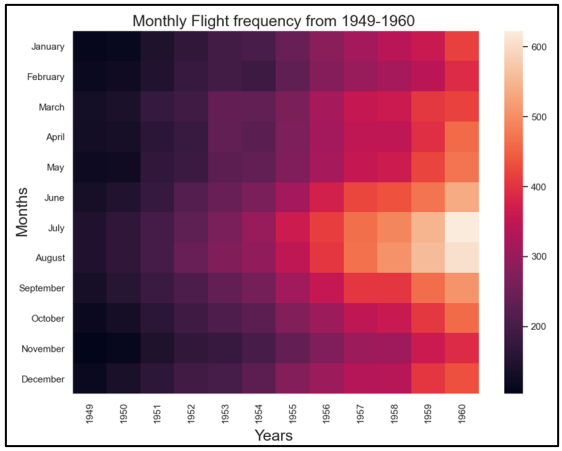
Using Area Encoding

[6] Nathan Yau, Data Points, Visualisation that Means Something, Wiley (2013)

Heatmaps

Seeing Relationships using Grids

- Heatmap depict values for a main variable of interest in **coloured rectangles** across two axis variables[7].
- Any patterns in the values for one or both of these axes can be observed by how the cell **colour changes** across each axis.
- These two axis variables are either **categorical** or if quantitative, they are binned into **discrete ranges** to form the desired number of cells.



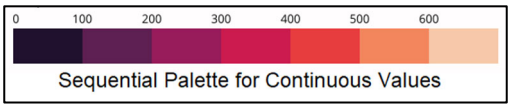
Heatmap plot using Seaborn's **heatmap()**[8]

[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/
[8] Seaborn, Plot rectangular data as a color-encoded matrix, https://seaborn.pydata.org/generated/seaborn.heatmap.html

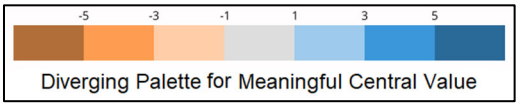
Heatmaps

Choosing the Colours

- Since colour is a key feature of heatmaps, an **appropriate colour palette** should be chosen to match the nature of the primary variable of interest[7].
- A **sequential** colour ramp between value and colour is often used. Lighter colours map to larger values & darker colours to smaller values (or vice versa).



- If the values have a meaningful zero point, a **diverging** colour palette would add better visual clarity.



- The colour distribution **legend** must be included in the heatmap to allow associated values of the colours to be interpreted.

[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/

Summary

Relation Plots

- Relation plots help us visualise **relationships** or **patterns** in the data.
- **Scatter plots** uses 2D position of dots to visualise **spatial relationships** (e.g. clustering, strong correlation, outliers, etc.) among data with **two** interval or ratio scale variables.
- **Linear regression** and **LOWESS** allow us to extract best fit trend lines within these scattering of dots.
- **Bubble plots** extend scatter plots ability to handle one more **quantitative dimension**, but discrimination of area variations is limited.
- **Heatmaps** allow us to visualised relationships over a **coloured grid**.

References for Relation Plots

[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/

[2] Seaborn, Plot data and a linear regression model fit , https://seaborn.pydata.org/generated/seaborn.regplot.html

[3] J.B. Mailman, Data Smoothing in Data Science Visualization (2021), https://towardsdatascience.com/data-smoothing-for-data-science-visualization-the-goldilocks-trio-part-1-867765050615

[4] Statsmodels Python library reference for its lowess method, https://www.statsmodels.org/devel/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html

[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/

[6] Nathan Yau, Data Points, Visualisation that Means Something, Wiley (2013)

[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/

[8] Seaborn, Plot rectangular data as a color-encoded matrix, https://seaborn.pydata.org/generated/seaborn.heatmap.html

Note: All online articles were accessed between May to June 2021