SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

10.020 Data Driven World

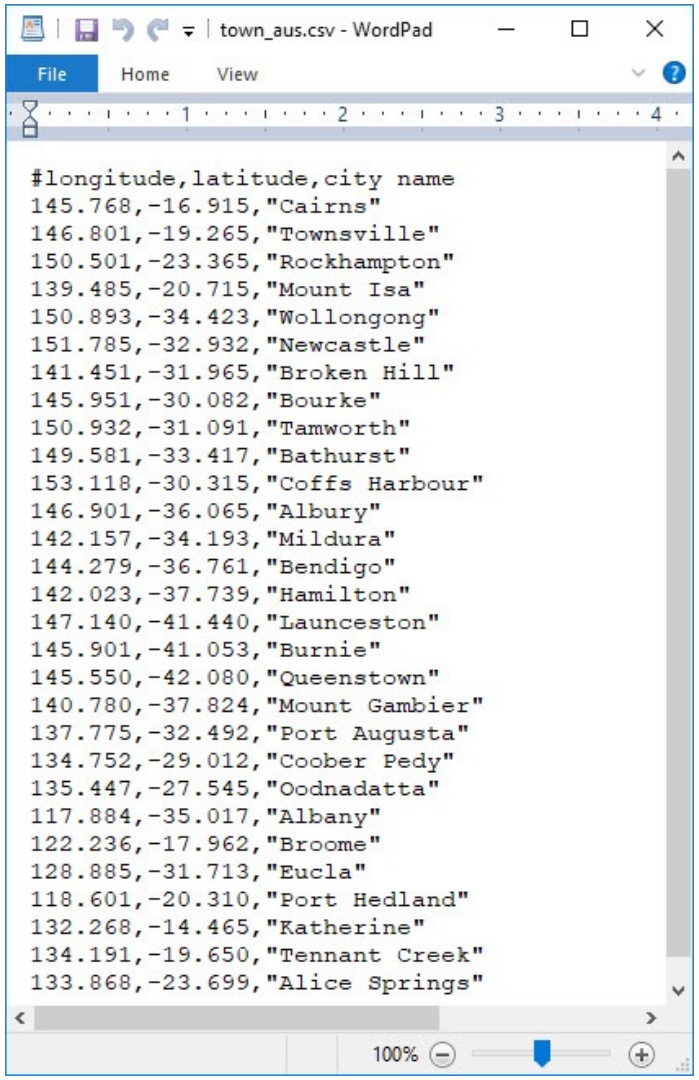
Visualizing Data

Peng Song, ISTD

Week 8, Lesson 2, 2021

Revision: Tabular Data

- A table is an arrangement of information or data, typically in rows and columns.
- A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values.



Revision: Panadas Library

- Data types
  - Series
  - DataFrame
- DataFrame operations
  - Get DataFrame information
  - Get a Column or Row as a Series
  - Get Rows and Columns as DataFrame
  - Select Data Using Conditions
  - Transpose Data Frame
  - Statistical Functions
  - Vector Operations

Revision: Data

- Data can have different forms

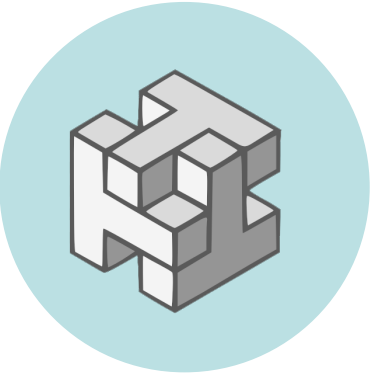
Text

Audio

Video

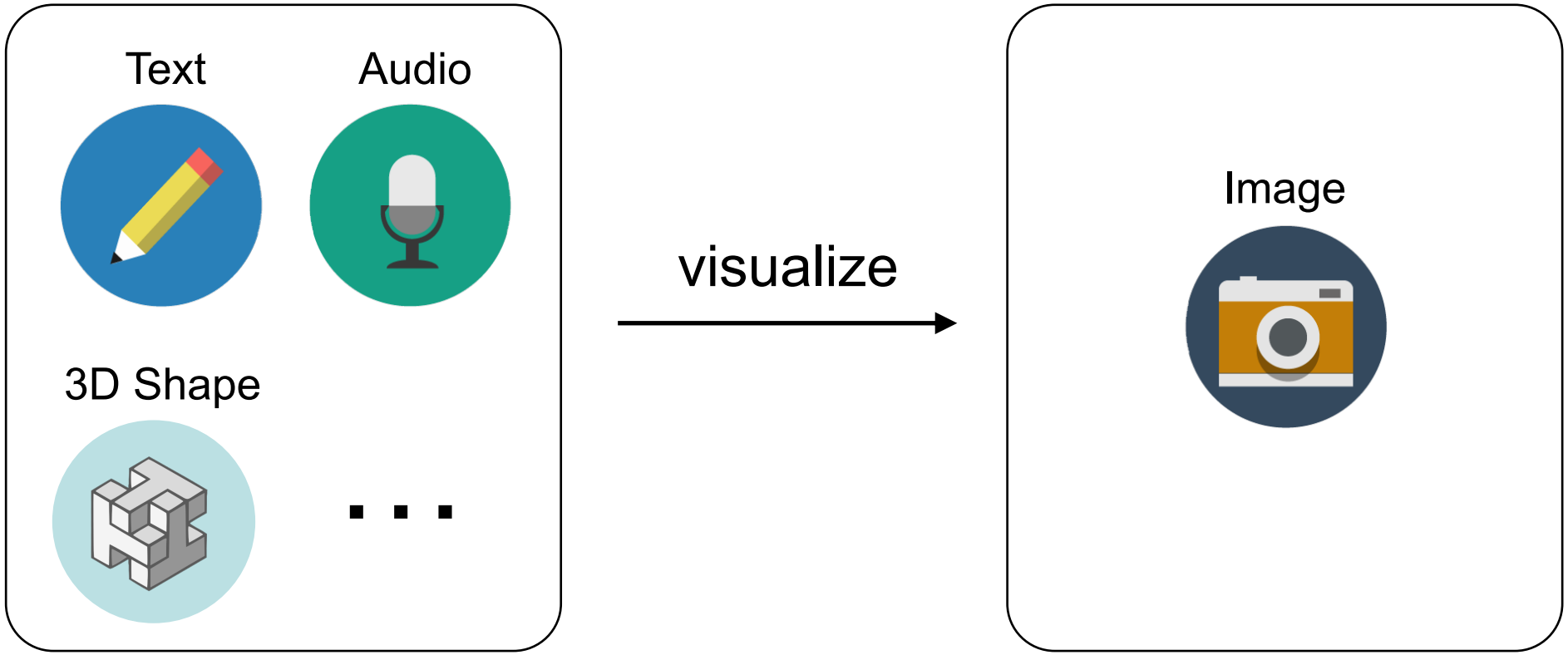
Image

3D Shape



What is Data Visualization?

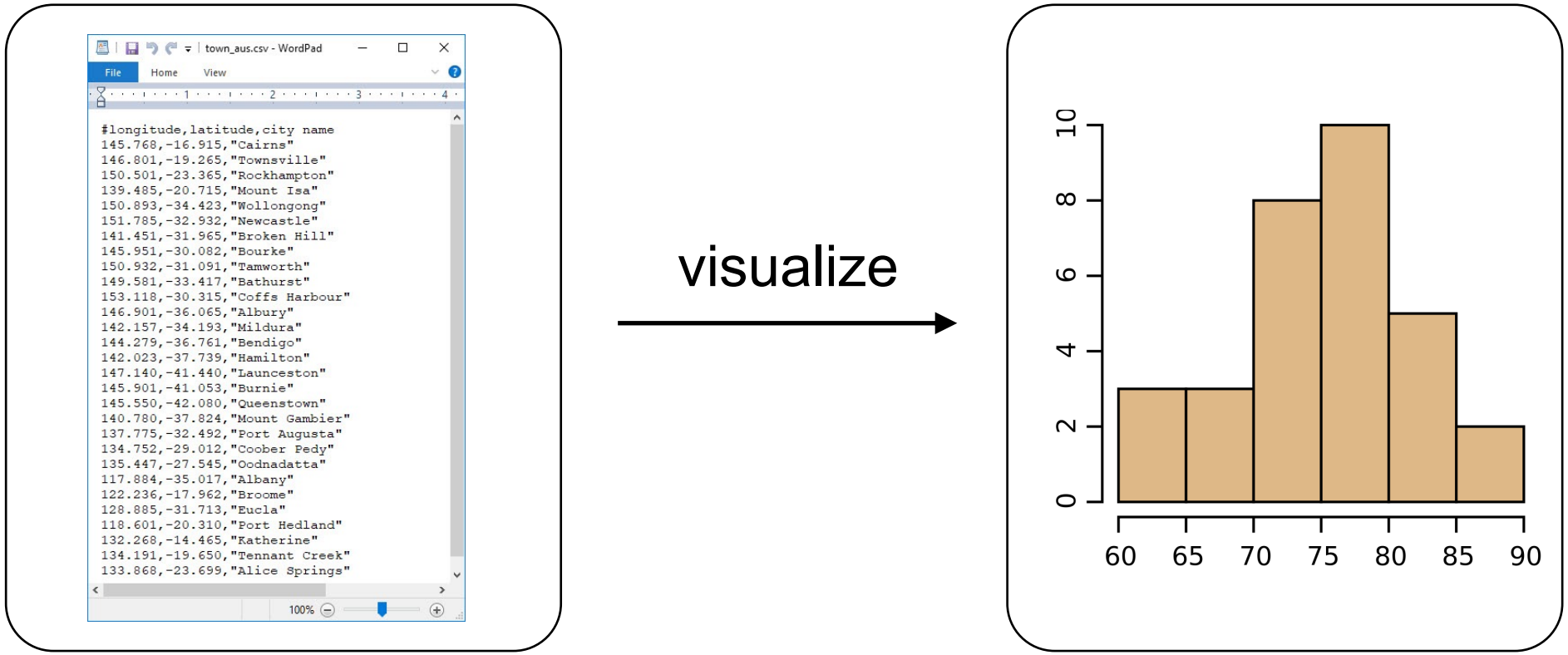
Data visualization is an interdisciplinary field that deals with the visual representation of data.



Wikipedia

What is Data Visualization?

Data visualization is an interdisciplinary field that deals with the visual representation of data.



When is Data Visualization Useful?

1. **Too much data**:
   - do not have time to analyze it all (or read the analysis results)
   - show an overview, discover which questions are relevant
   - refine search either visually or analytically

2. **Qualitative / complex questions**:
   - cannot capture question compactly/exactly in a query
   - question/goal is inherently qualitative: understand what is going on
   - show an overview, answer the question by seeing relevant patterns

3. **Communication**:
   - transfer results to different (non technical) stakeholders
   - learn about a new domain or problem

Subfields of Data Visualization

Scientific Visualization: "The use of computers or techniques for comprehending data or to extract knowledge from the results of simulations, computations, or measurements"

[McCormick *et al.*, 1987]

Information Visualization: "Visualization applied to abstract quantities and relations in order to get insight in the data"

[Chi, 2000]

Software Visualization: "Software visualization is concerned with the static or animated 2D or 3D visual representation of information about software systems based on their structure, history, or behavior in order to help software engineering tasks"

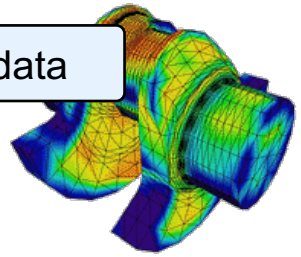
[Diehl, 2006]

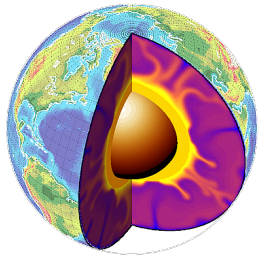
Subfields of Data Visualization

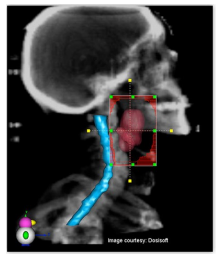
1985

Scientific Visualization:
- engineering
- geosciences
- medicine

spatial data





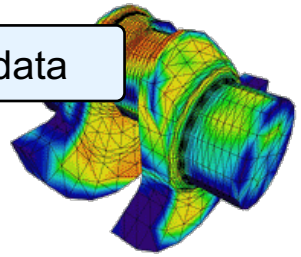


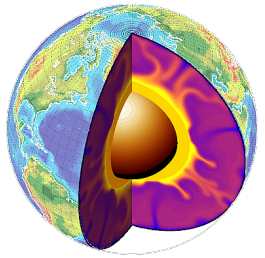
Subfields of Data Visualization

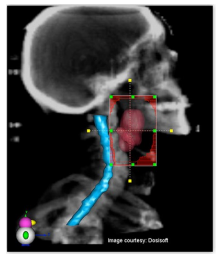
1985

Scientific Visualization:
- engineering
- geosciences
- medicine

spatial data



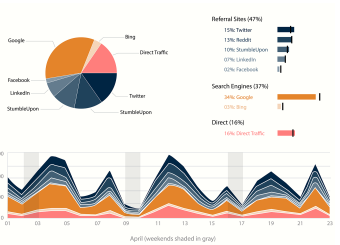




1995

Information Visualization:
- finance
- telecom
- business management

non-spatial data

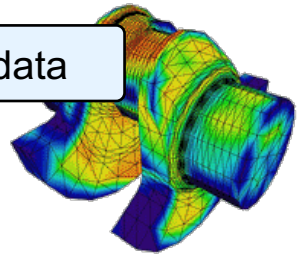


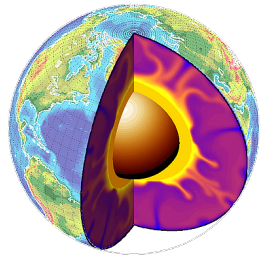
Subfields of Data Visualization

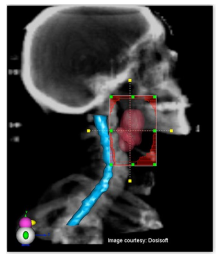
1985

Scientific Visualization:
- engineering
- geosciences
- medicine

spatial data



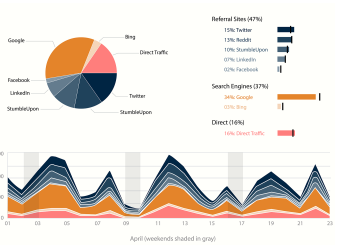




1995

Information Visualization:
- finance
- telecom
- business management

non-spatial data



2000

Software Visualization:
- the software industry

software data

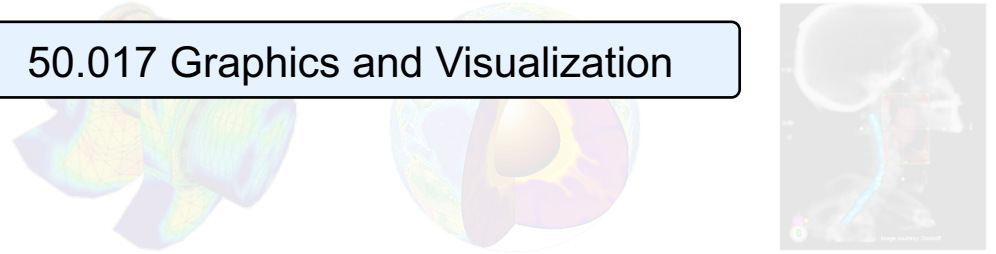
source: Alexandru Telea

Subfields of Data Visualization

1985

Scientific Visualization:
- engineering
- geosciences
- medicine

50.017 Graphics and Visualization



1995

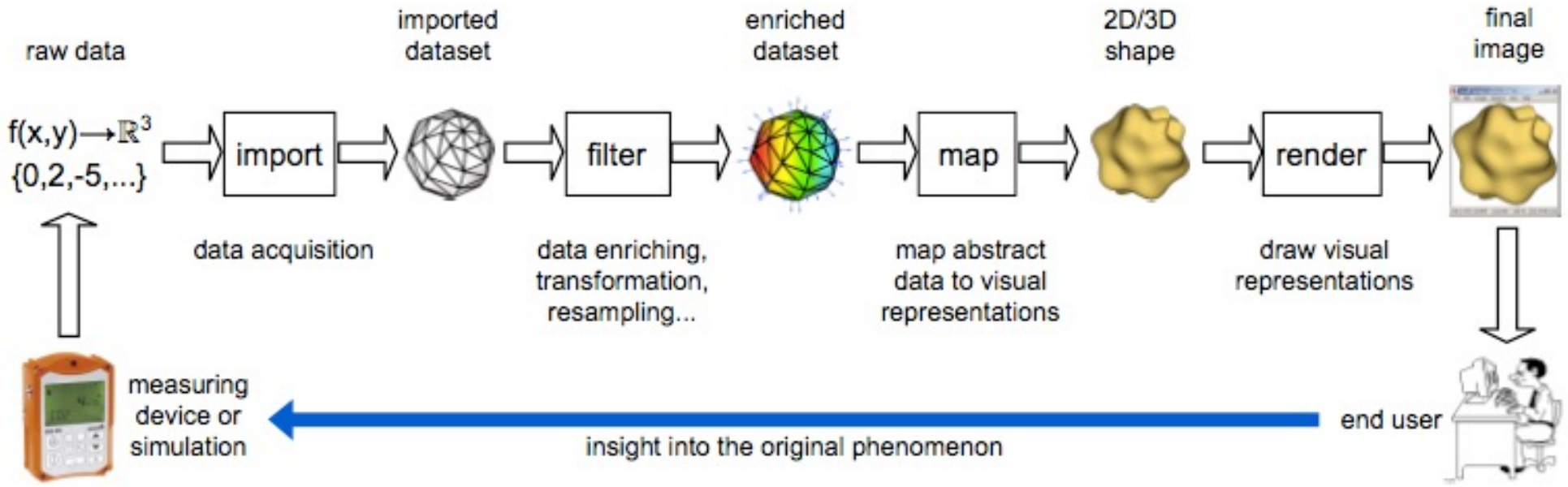
Information Visualization:
- finance
- telecom
- business management

Focus in 10.020 Data Driven World

2000

Software Visualization:
- the software industry

Data Visualization Pipeline

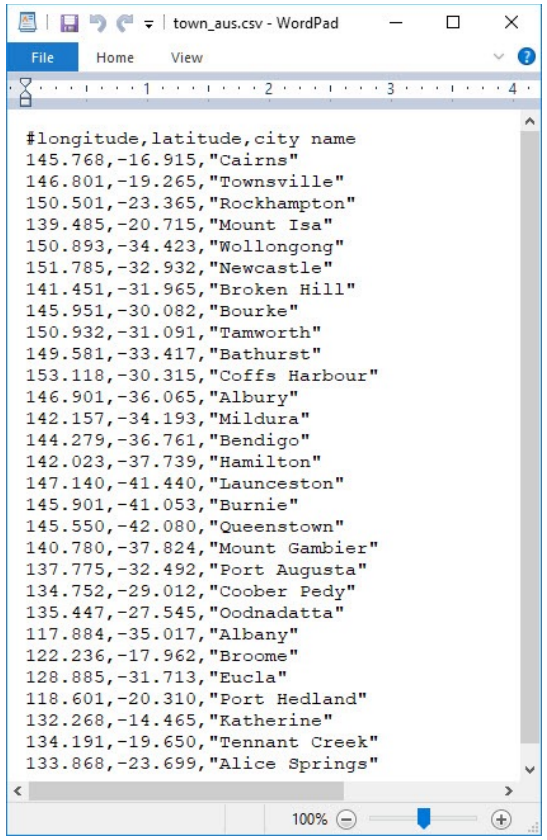


- transform raw data into insightful answers
- sequence of steps
  1. data acquisition (conversion, formatting, cleaning)
  2. data enrichment (transformation, resampling, filtering)
  3. **data mapping** (produce visible shapes from data)
  4. rendering (draw and interact with the shapes)

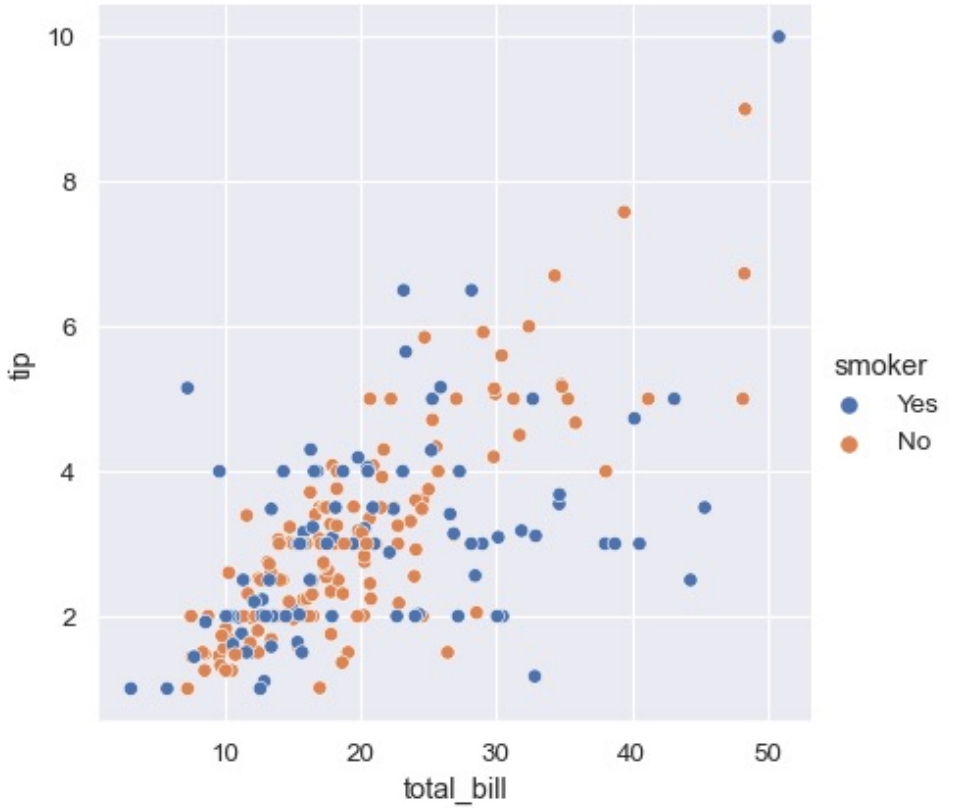
source: Alexandru Telea

Tabular Data Visualization – Mapping

Mapping is not ‘neutral’ or natural, but reflects the problem/question to be solved



mapping



Tabular Data Visualization: Tasks

1. **Distribution** of each category of data
2. **Relation** between different categories of data

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
0	2017-01	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44.0	Improved	1979	61 years 04 months	232000.0
1	2017-01	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67.0	New Generation	1978	60 years 07 months	250000.0
2	2017-01	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67.0	New Generation	1980	62 years 05 months	262000.0
3	2017-01	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68.0	New Generation	1980	62 years 01 month	265000.0
4	2017-01	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67.0	New Generation	1980	62 years 05 months	265000.0
...	...	...	...	...	...	...	...	...	...	...	...
95853	2021-04	YISHUN	EXECUTIVE	326	YISHUN RING RD	10 TO 12	146.0	Maisonette	1988	66 years 04 months	650000.0
95854	2021-04	YISHUN	EXECUTIVE	360	YISHUN RING RD	04 TO 06	146.0	Maisonette	1988	66 years 04 months	645000.0
95855	2021-04	YISHUN	EXECUTIVE	326	YISHUN RING RD	10 TO 12	146.0	Maisonette	1988	66 years 04 months	585000.0
95856	2021-04	YISHUN	EXECUTIVE	355	YISHUN RING RD	10 TO 12	146.0	Maisonette	1988	66 years 08 months	675000.0
95857	2021-04	YISHUN	EXECUTIVE	277	YISHUN ST 22	04 TO 06	146.0	Maisonette	1985	63 years 05 months	625000.0

95858 rows × 11 columns

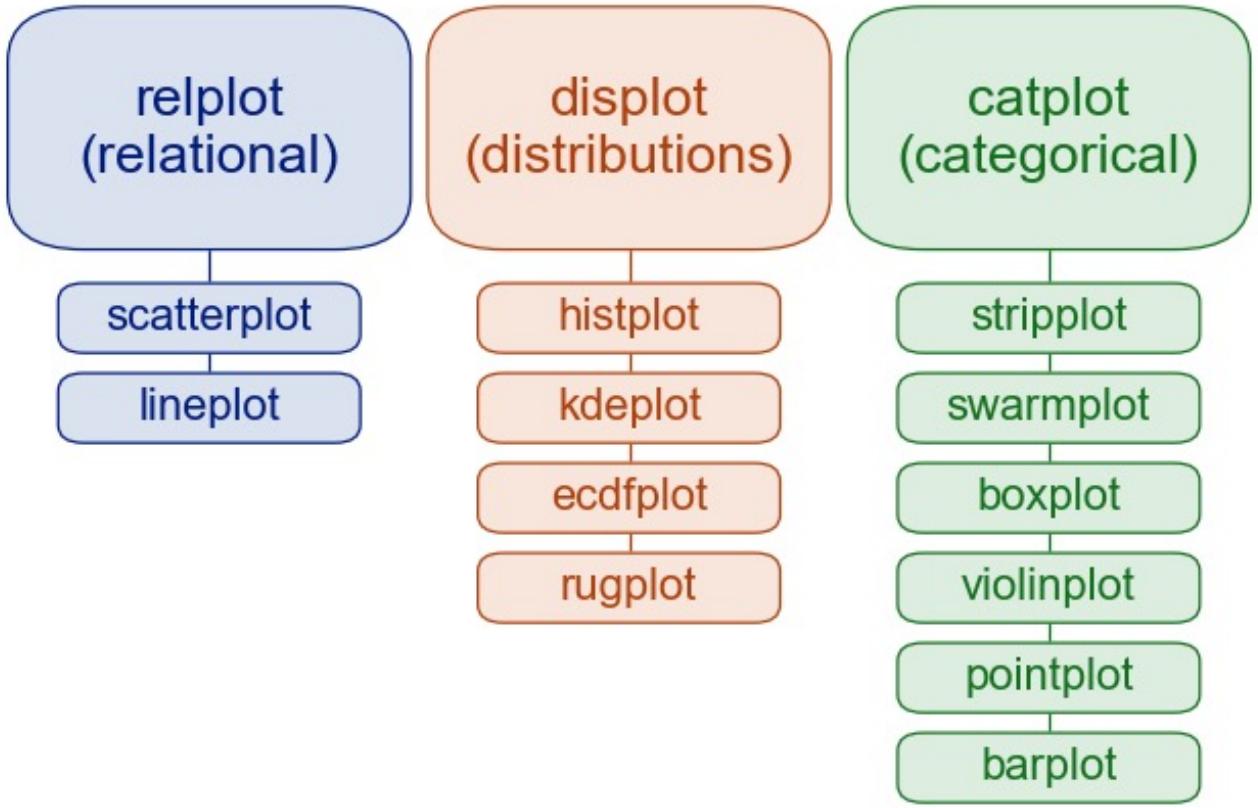
Data Visualization in Python

- Python draw common plots to visualize data using **Matplotlib** and **Seaborn**.
- Seaborn works on top of Matplotlib and you will need to import both packages in most of the cases.

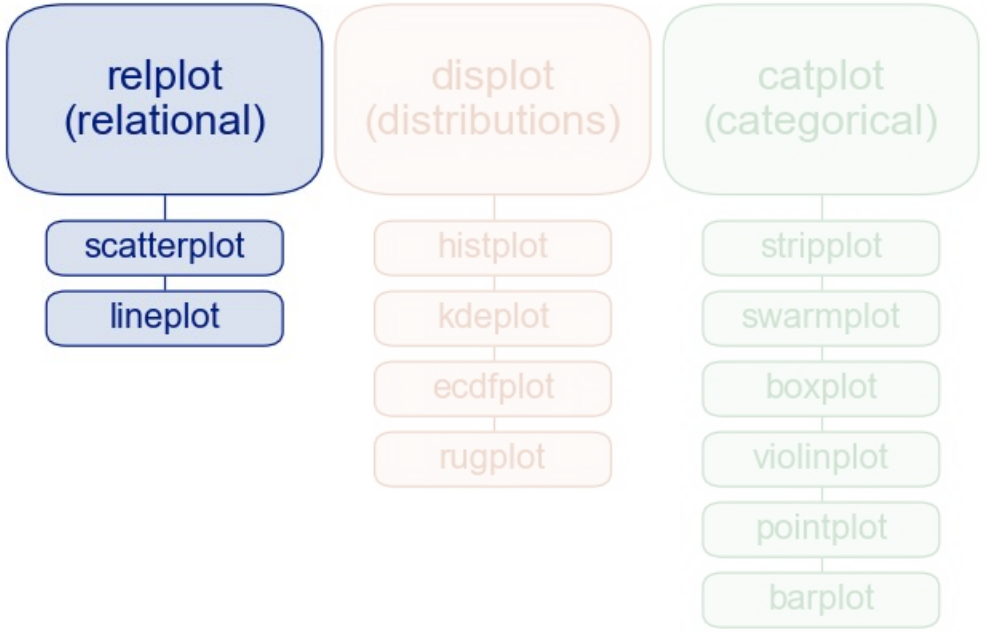
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

Categories of Plots

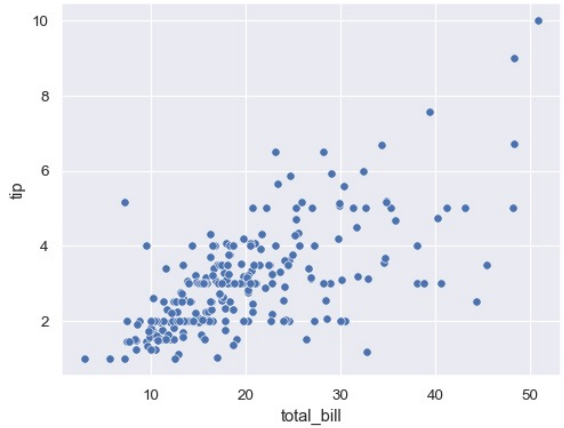
- There are different categories of plot in Seaborn packages as shown in Seaborn documentation.



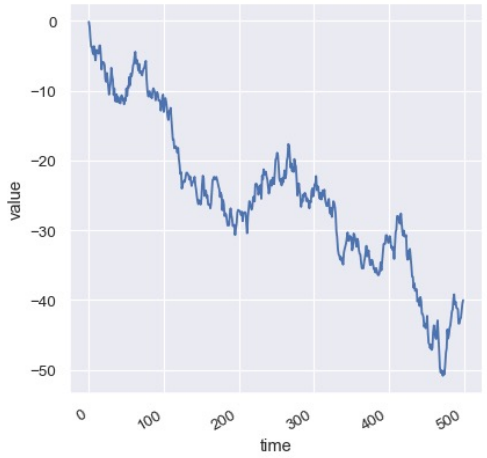
Categories of Plots



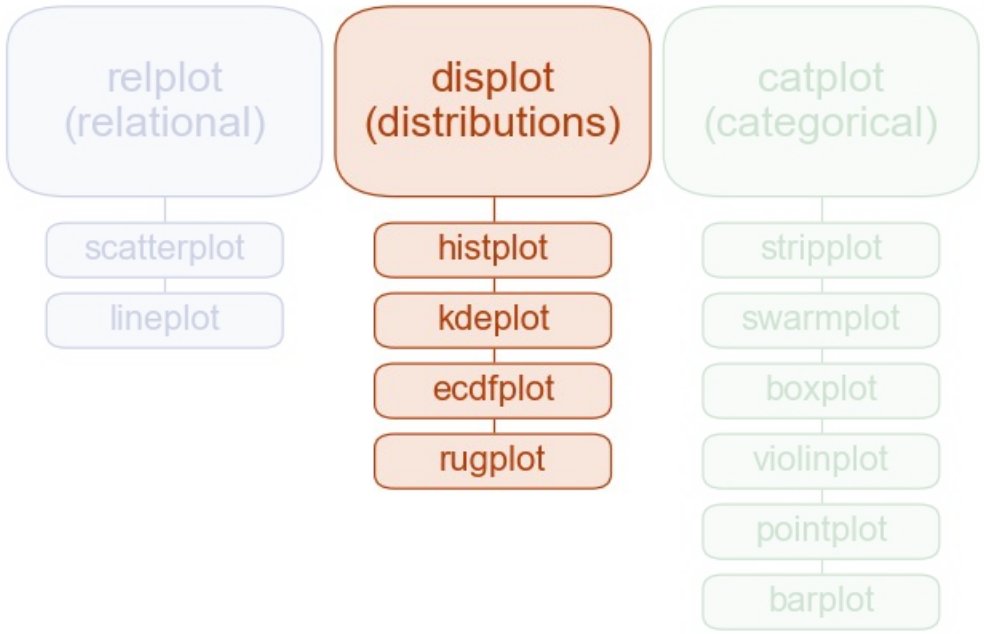
scatter plot



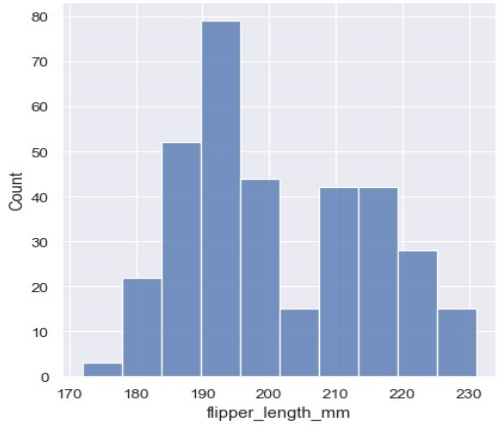
line plot



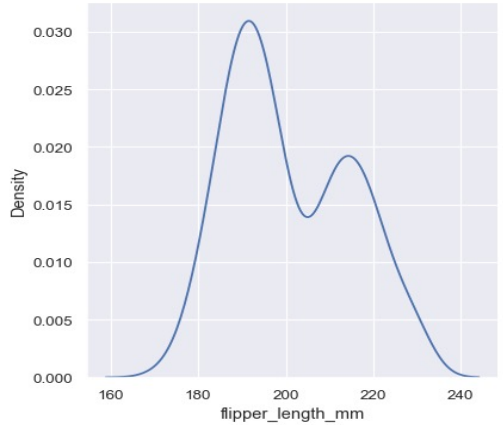
Categories of Plots



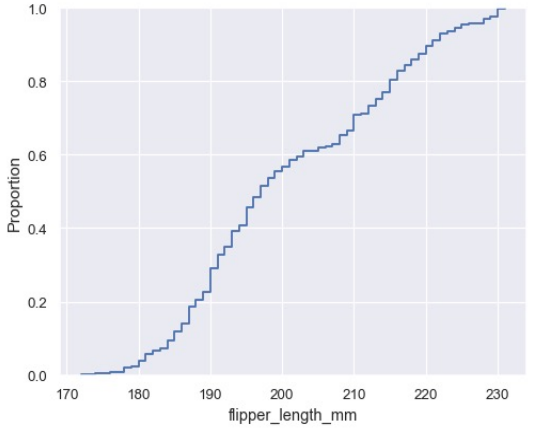
histogram (hist)



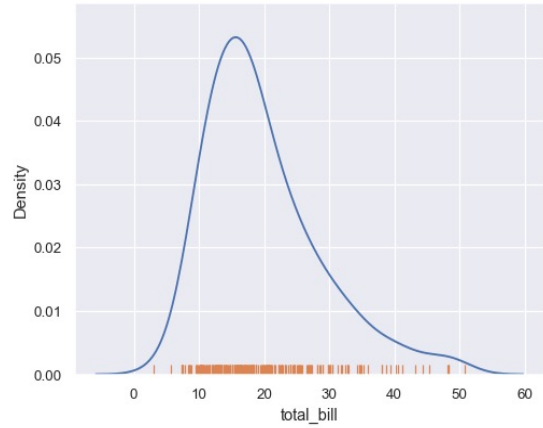
kernel density estimation (kde)



empirical cumulative distribution function (ecdf)



rugplot



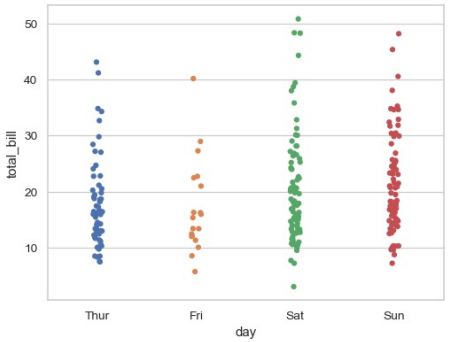
Categories of Plots

relplot (relational)
- scatterplot
- lineplot

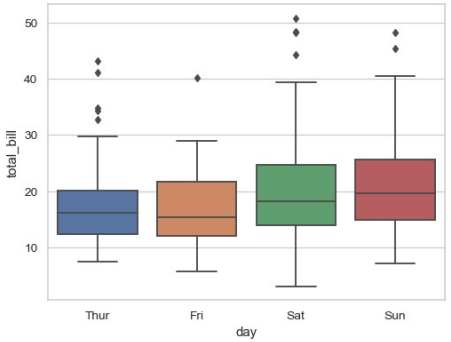
displot (distributions)
- histplot
- kdeplot
- ecdfplot
- rugplot

catplot (categorical)
- stripplot
- swarmplot
- boxplot
- violinplot
- pointplot
- barplot

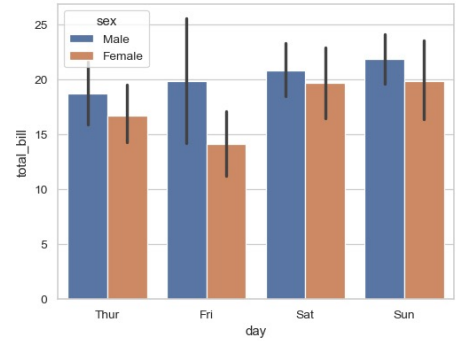
strip plot



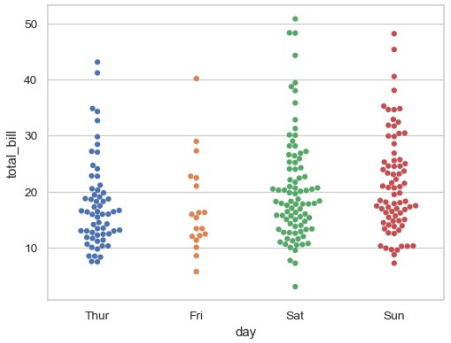
box plot



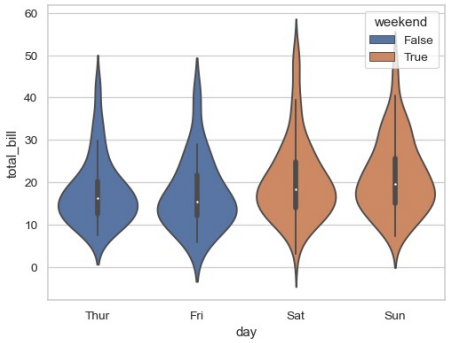
bar plot



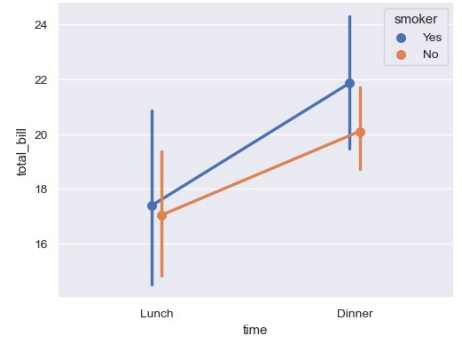
swarm plot



violin plot

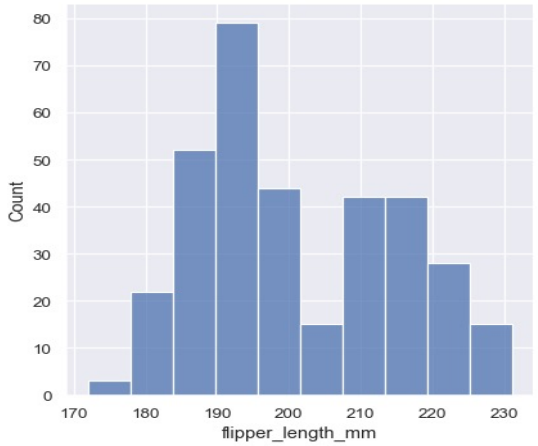


point plot

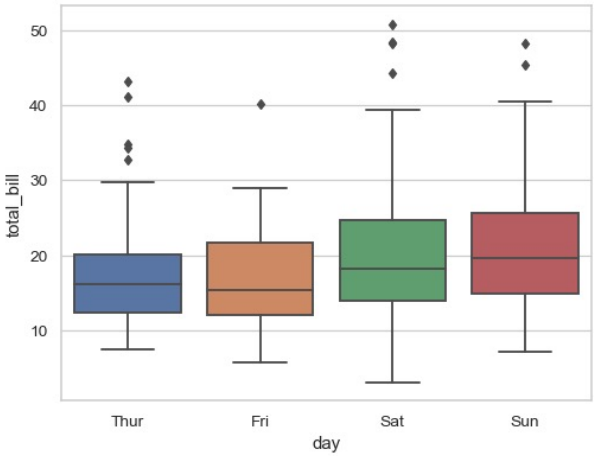


Four Plots

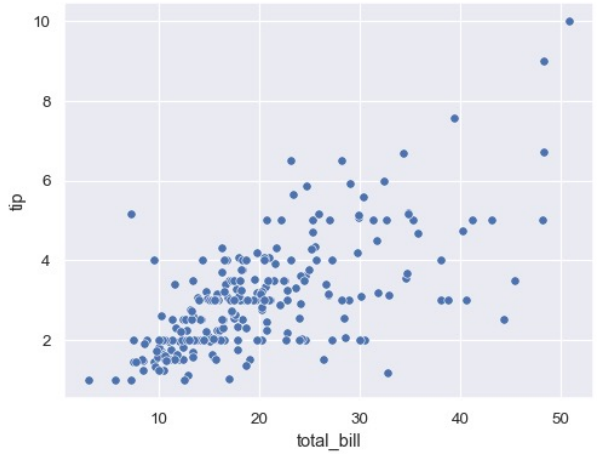
#1 hist plot



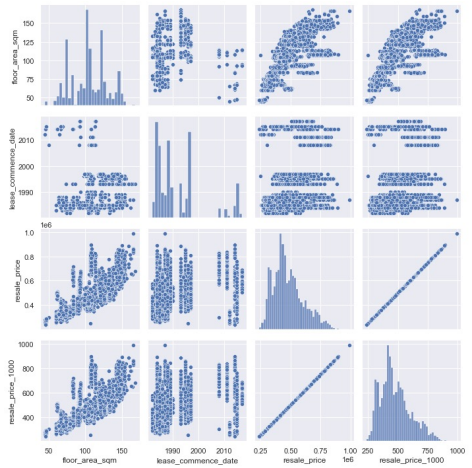
#2 box plot



#3 scatter plot



#4 pair plot



#1 Histogram Plot

- Get the data for resale in Tampines only.

```
df_tampines = df.loc[df['town'] == 'TAMPINES',:]
df_tampines
```

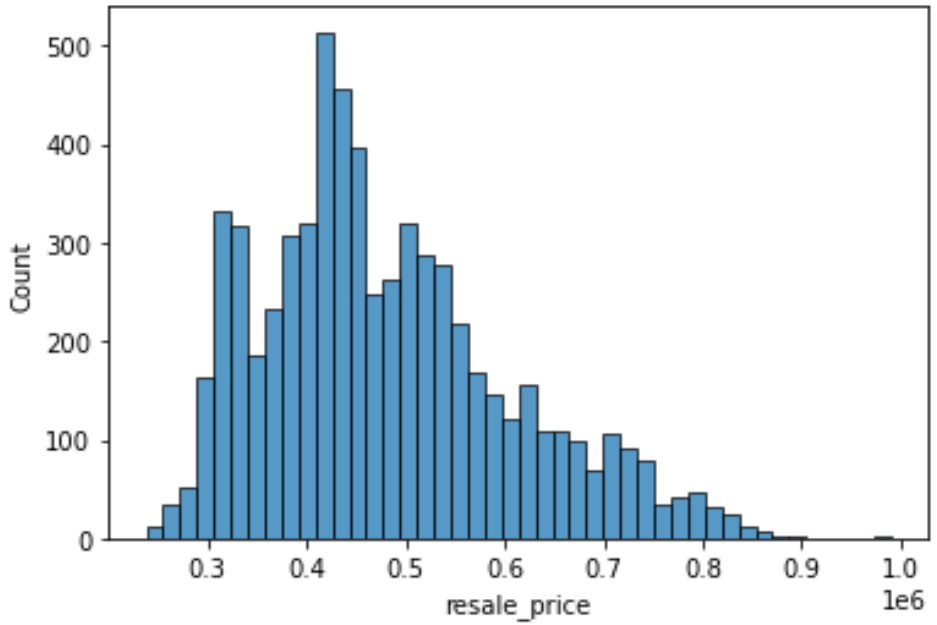
	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
917	2017-01	TAMPINES	2 ROOM	299A	TAMPINES ST 22	01 TO 03	45.0	Model A	2012	94 years 02 months	250000.0
918	2017-01	TAMPINES	3 ROOM	403	TAMPINES ST 41	01 TO 03	60.0	Improved	1985	67 years 09 months	270000.0
919	2017-01	TAMPINES	3 ROOM	802	TAMPINES AVE 4	04 TO 06	68.0	New Generation	1984	66 years 05 months	295000.0
920	2017-01	TAMPINES	3 ROOM	410	TAMPINES ST 41	01 TO 03	69.0	Improved	1985	67 years 08 months	300000.0
921	2017-01	TAMPINES	3 ROOM	462	TAMPINES ST 44	07 TO 09	64.0	Simplified	1987	69 years 06 months	305000.0
...	...	...	...	...	...	...	...	...	...	...	...
95671	2021-04	TAMPINES	EXECUTIVE	495E	TAMPINES ST 43	04 TO 06	147.0	Apartment	1994	71 years 10 months	630000.0
95672	2021-04	TAMPINES	EXECUTIVE	477	TAMPINES ST 43	04 TO 06	153.0	Apartment	1993	71 years 04 months	780000.0
95673	2021-04	TAMPINES	EXECUTIVE	497J	TAMPINES ST 45	10 TO 12	139.0	Premium Apartment	1996	74 years 03 months	695000.0
95674	2021-04	TAMPINES	EXECUTIVE	857	TAMPINES ST 83	01 TO 03	154.0	Maisonette	1988	66 years	735000.0
95675	2021-04	TAMPINES	MULTI-GENERATION	454	TAMPINES ST 42	01 TO 03	132.0	Multi Generation	1987	65 years 04 months	600000.0

6392 rows × 11 columns

#1 Histogram Plot

- Plot the resale price distribution using histplot.

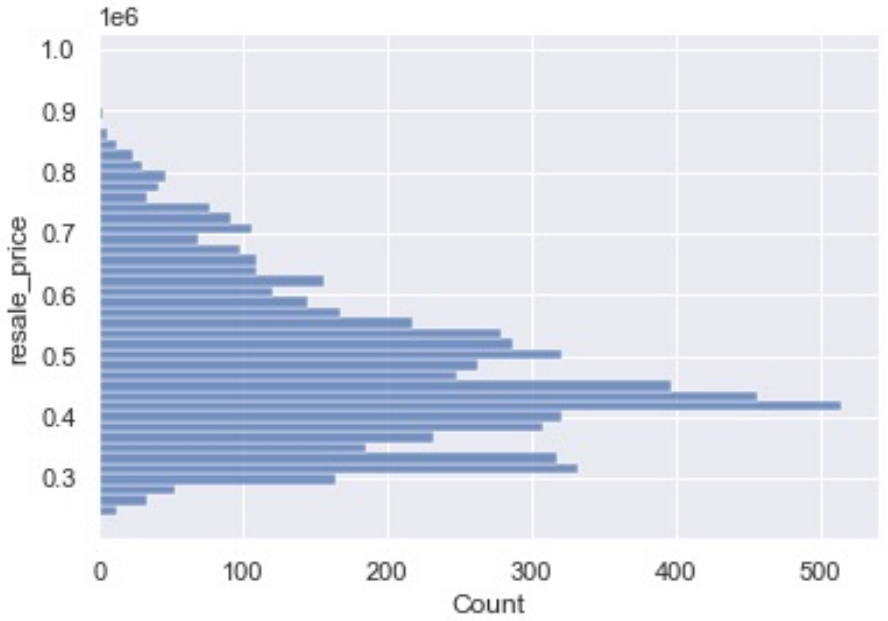
```
sns.histplot(x='resale_price', data=df_tampines)
```



#1 Histogram Plot

- Change the plot to show it vertically.

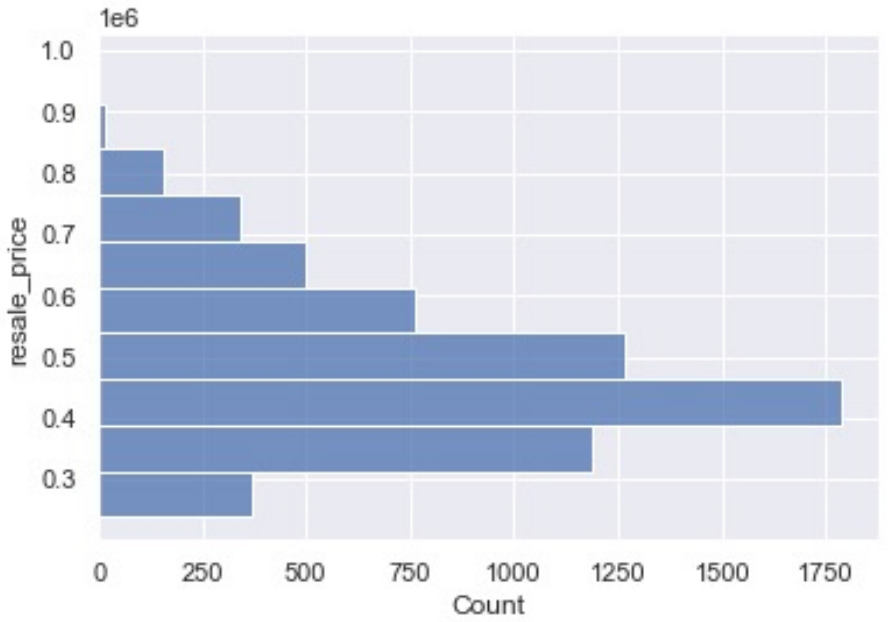
```
sns.set()
sns.histplot(y='resale_price', data=df_tampines)
```



#1 Histogram Plot

- Specify the number of bins.

```
sns.histplot(y='resale_price', data=df_tampines, bins=10)
```

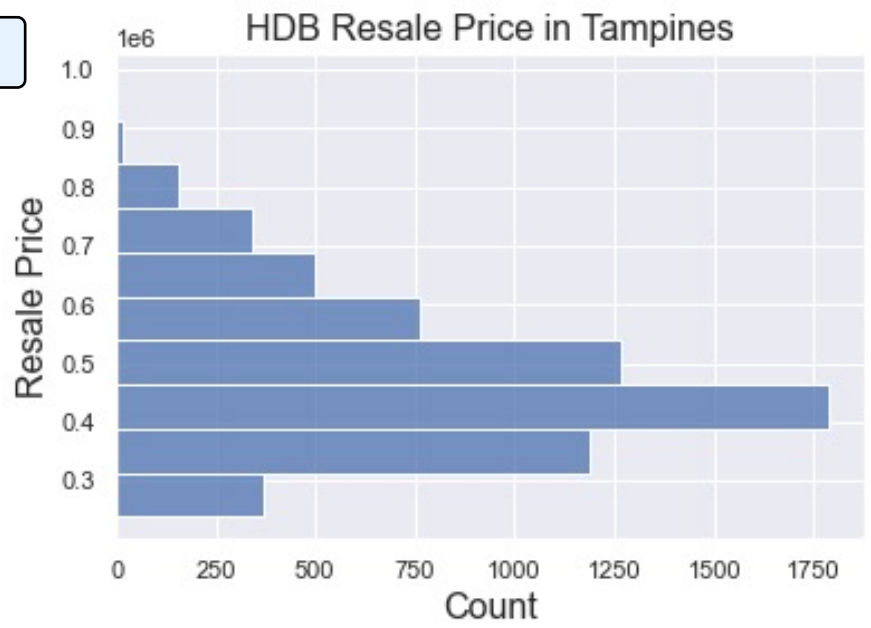


#1 Histogram Plot

- Use some of Matplotlib functions to change the figure's labels and title.

```
myplot = sns.histplot(y='resale_price', data=df_tampines, bins=10)
myplot.set_xlabel('Count', fontsize=16)
myplot.set_ylabel('Resale Price', fontsize=16)
myplot.set_title('HDB Resale Price in Tampines', fontsize=16)
```

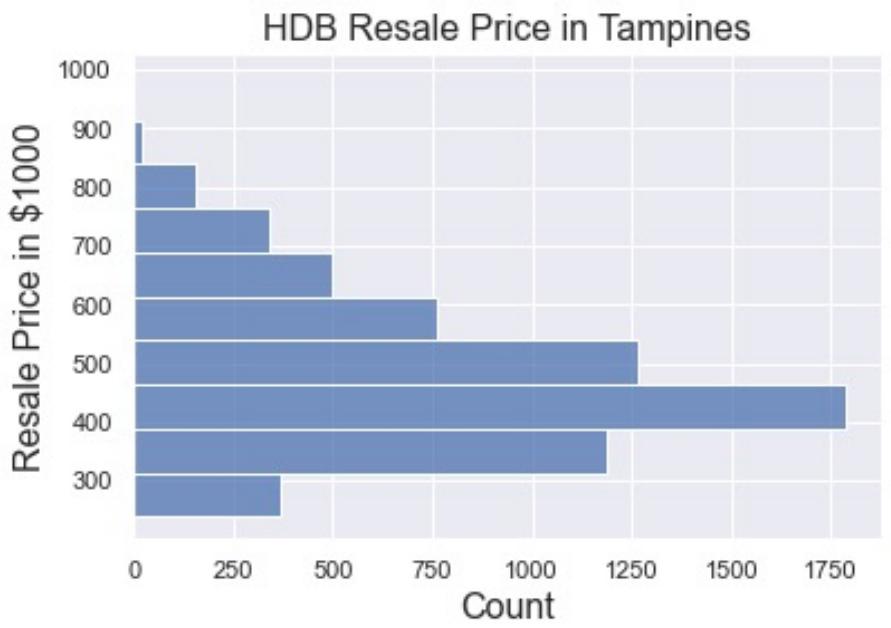
Too large number



#1 Histogram Plot

- Use some of Matplotlib functions to change the figure's labels and title.

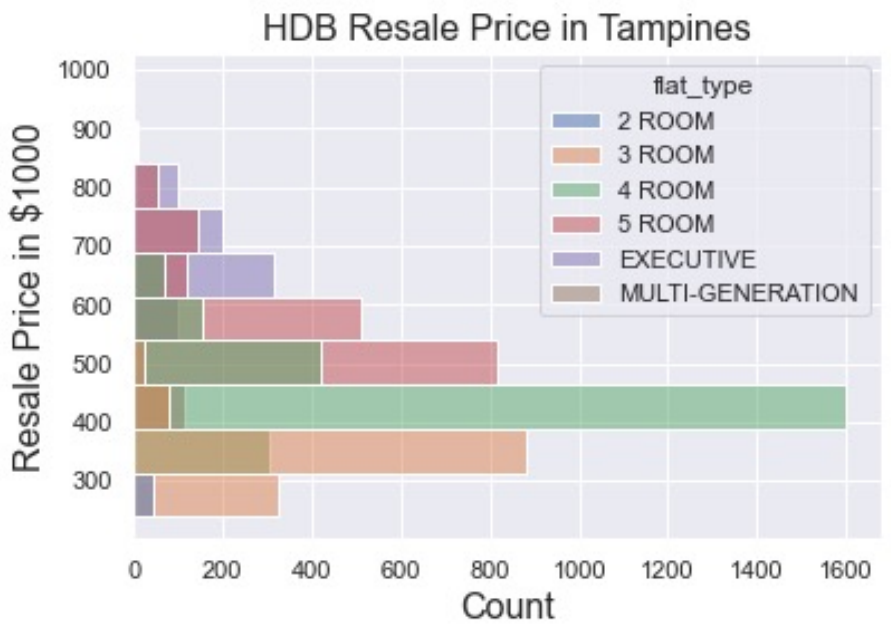
```
df_tampines['resale_price_1000'] =
df_tampines['resale_price'].apply(lambda price: price/1000)
myplot = sns.histplot(y='resale_price_1000', data=df_tampines, bins=10)
myplot.set_xlabel('Count', fontsize=16)
myplot.set_ylabel('Resale Price in $1000', fontsize=16)
myplot.set_title('HDB Resale Price in Tampines', fontsize=16)
```



#1 Histogram Plot

- See the distribution of the resale price according to the flat type

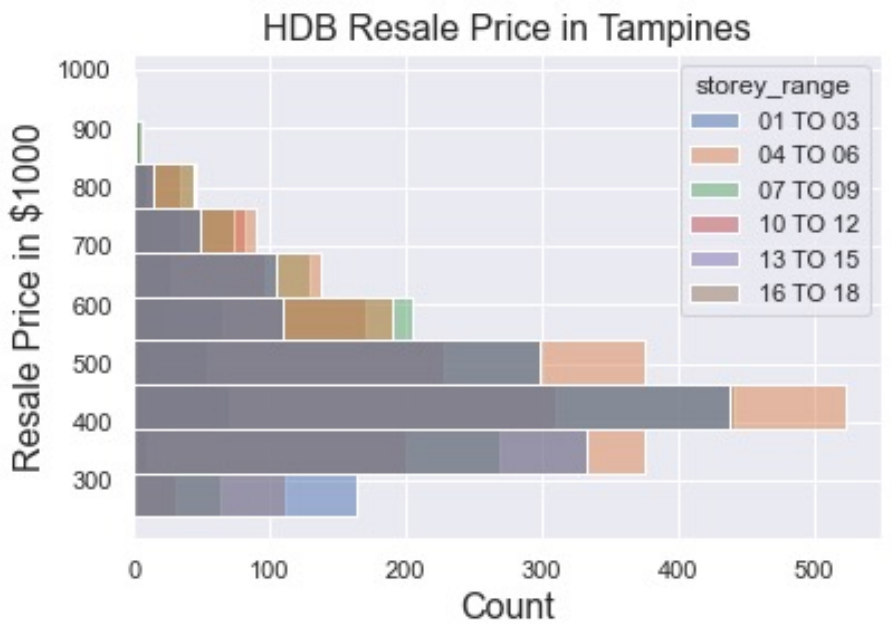
```
myplot = sns.histplot(y='resale_price_1000', hue='flat_type',
data=df_tampines, bins=10)
myplot.set_xlabel('Count', fontsize=16)
myplot.set_ylabel('Resale Price in $1000', fontsize=16)
myplot.set_title('HDB Resale Price in Tampines', fontsize=16)
```



#1 Histogram Plot

- See the distribution of the resale price according to the story range

```
myplot = sns.histplot(y='resale_price_1000', hue='storey_range',
data=df_tampines, bins=10)
myplot.set_xlabel('Count', fontsize=16)
myplot.set_ylabel('Resale Price in $1000', fontsize=16)
myplot.set_title('HDB Resale Price in Tampines', fontsize=16)
```



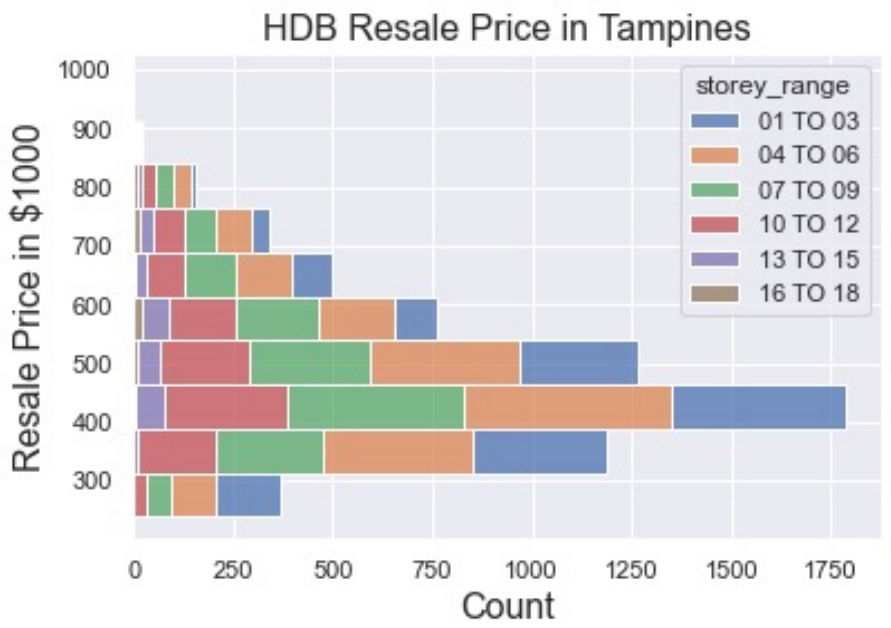
Why do the bars have more than 6 colors?

Bars with different colors **layered** on top of each other

#1 Histogram Plot

- Set the multiple argument when there are multiple data in the same area.

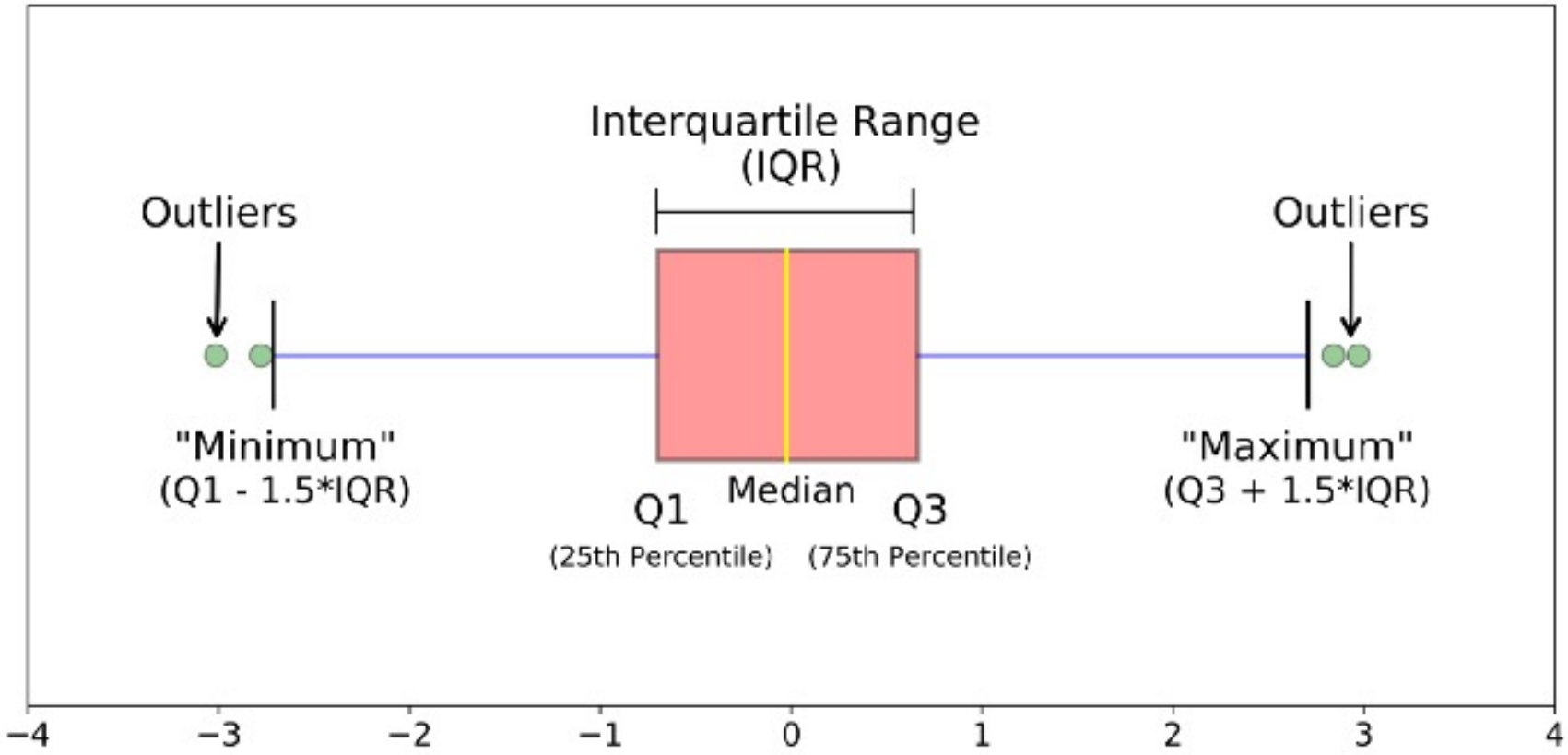
```
myplot = sns.histplot(y='resale_price_1000', hue='storey_range',
multiple='stack', data=df_tampines, bins=10)
myplot.set_xlabel('Count', fontsize=16)
myplot.set_ylabel('Resale Price in $1000', fontsize=16)
myplot.set_title('HDB Resale Price in Tampines', fontsize=16)
```



Bars with different colors **stacked** on top of each other

#2 Box Plot

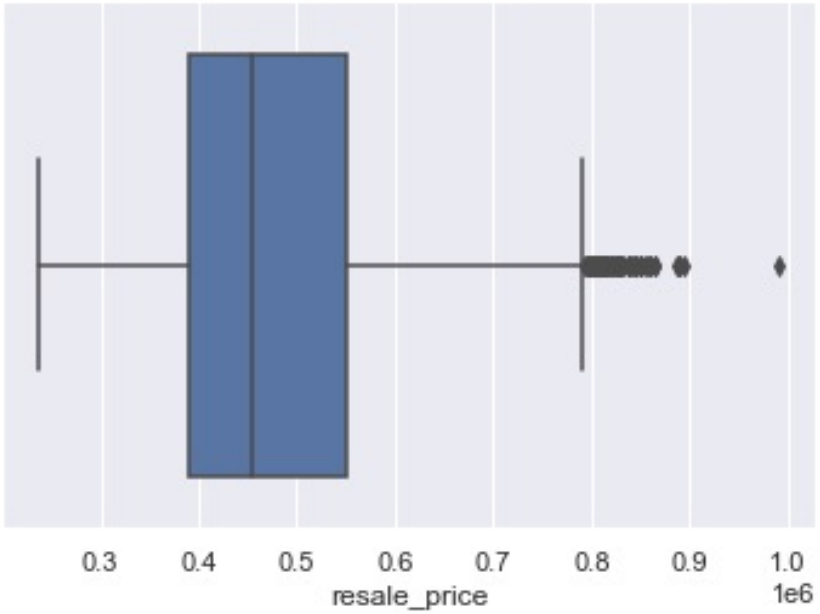
- Understand the boxplot



#2 Box Plot

- Use the boxplot to see some descriptive statistics of the data.

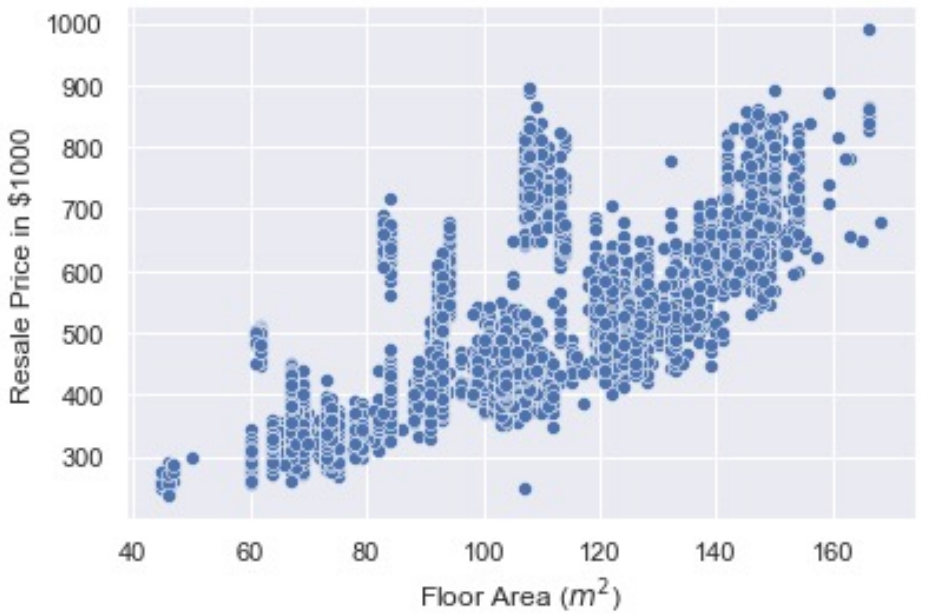
```
sns.boxplot(x='resale_price', data=df_tampines)
```



#3 Scatter Plot

- Plot the floor area and resale price to see if there is any relationship.

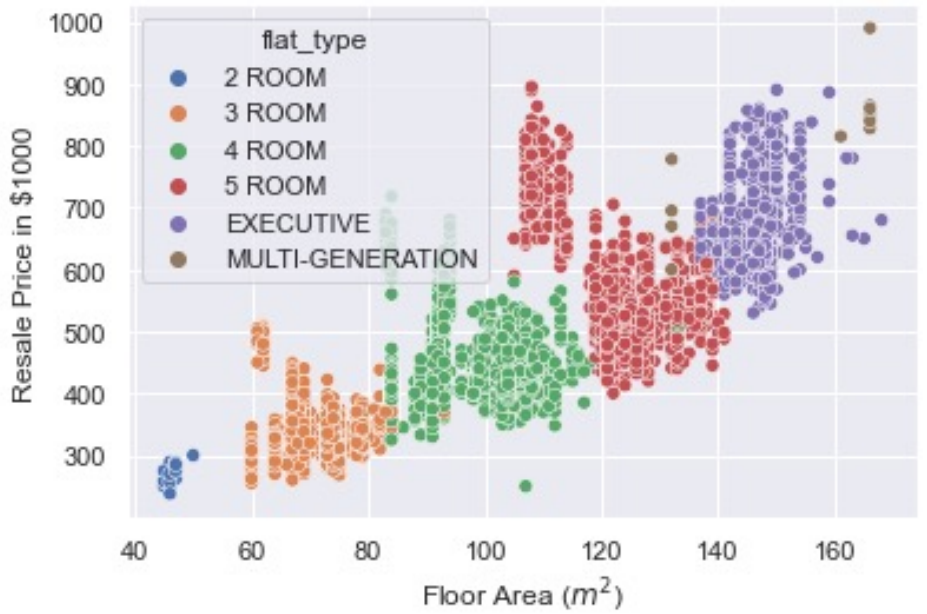
```
myplot = sns.scatterplot(x='floor_area_sqm', y='resale_price_1000',
data=df_tampines)
myplot.set_xlabel('Floor Area ($m^2$)’)
myplot.set_ylabel('Resale Price in $1000')
```



#3 Scatter Plot

- We can again use the hue argument to see any category in the plot.

```
myplot = sns.scatterplot(x='floor_area_sqm', y='resale_price_1000',
hue='flat_type', data=df_tampines)
myplot.set_xlabel('Floor Area ($m^2$)’)
myplot.set_ylabel('Resale Price in $1000')
```



#4 Pair Plot

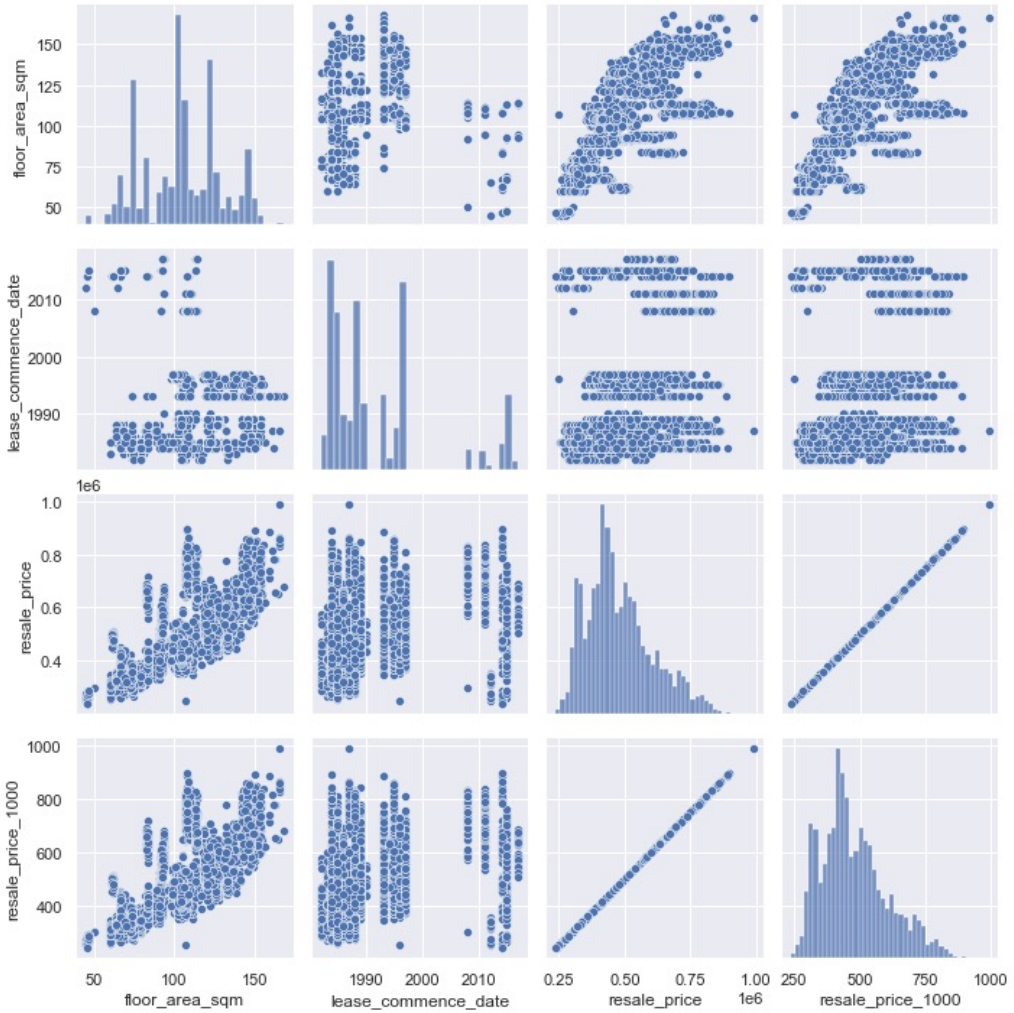
- Plots the relationship on multiple data columns.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
917	2017-01	TAMPINES	2 ROOM	299A	TAMPINES ST 22	01 TO 03	45.0	Model A	2012	94 years 02 months	250000.0
918	2017-01	TAMPINES	3 ROOM	403	TAMPINES ST 41	01 TO 03	60.0	Improved	1985	67 years 09 months	270000.0
919	2017-01	TAMPINES	3 ROOM	802	TAMPINES AVE 4	04 TO 06	68.0	New Generation	1984	66 years 05 months	295000.0
920	2017-01	TAMPINES	3 ROOM	410	TAMPINES ST 41	01 TO 03	69.0	Improved	1985	67 years 08 months	300000.0
921	2017-01	TAMPINES	3 ROOM	462	TAMPINES ST 44	07 TO 09	64.0	Simplified	1987	69 years 06 months	305000.0
...	...	...	...	...	...	...	...	...	...	...	...
95671	2021-04	TAMPINES	EXECUTIVE	495E	TAMPINES ST 43	04 TO 06	147.0	Apartment	1994	71 years 10 months	630000.0
95672	2021-04	TAMPINES	EXECUTIVE	477	TAMPINES ST 43	04 TO 06	153.0	Apartment	1993	71 years 04 months	780000.0
95673	2021-04	TAMPINES	EXECUTIVE	497J	TAMPINES ST 45	10 TO 12	139.0	Premium Apartment	1996	74 years 03 months	695000.0
95674	2021-04	TAMPINES	EXECUTIVE	857	TAMPINES ST 83	01 TO 03	154.0	Maisonette	1988	66 years	735000.0
95675	2021-04	TAMPINES	MULTI-GENERATION	454	TAMPINES ST 42	01 TO 03	132.0	Multi Generation	1987	65 years 04 months	600000.0

6392 rows × 11 columns

#4 Pair Plot

```
myplot = sns.pairplot(data=df_tampines)
```



diagonal: histograms
others: scatter plots

Cohort Problem CS3

CS3. *Histogram and Box plot:* Plot the histogram for the median value in \$1000 for the Boston's housing price.

Task 1: Plot the histogram with default bin values.

Task 2: Plot the histogram with 5 bins only.

Task 3: Plot the histogram with the following bin edges 0, 10, 20, 30, 40, 50.

Task 4: Plot the same data using a box plot in a horizontal manner.

Cohort Problem CS4

CS4. *Scatter plot:* Do the following plots.

Task 1: Display scatter plot of "RM" versus "MEDV".

Task 2: Display a scatter plot "weighted distances to five Boston employment centers" versus "Median value of owner-occupied homes in \$1000s".

Task 3: Display a scatter plot "proportion of non-retail business acres per town" versus "Median value of owner-occupied homes in \$1000s".

Thank You!