agenda : conclude plotting

random variable

```
import matplotlib.pyplot as plt
plt.style.use("seaborn-darkgrid")
```

```
%matplotlib inline ⟶ keep the plots in the notebook
x=[-3,5,7]
Y=[20,2,5]
plt.figure(figsize=(15,2))
plt.scatter(x,y,color="blue",s=200)   # s = size of points
plt.plot(x,y,color="red",linewidth=5)
plt.xlim(-4,8)
plt.ylim(0,11)
plt.xlabel('X axis',size=20)
plt.ylabel('Y axis',size=20)
plt.title('line plot',size=20)
plt.suptitle('Sales comparision',size=20,y=1.3)
plt.savefig('example.jpg',dpi=95)
```

ls ⇒ folders and files in the directory .

Matplotlib : Mtcars data

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

```
mtcars = pd.read_csv("mtcars.csv")

mtcars.head()

pd.crosstab(mtcars.gear, mtcars.cyl)
# for each categorical variable, it will give count for each variable

mtcars.gear.value_counts()
# it will give type and count of each entity

pd.crosstab(mtcars.gear, mtcars.cyl).plot(kind = "bar")
# we will get bar plot.

pd.crosstab(mtcars.gear, mtcars.cyl).plot(kind = "barh")
# we will get bar plot.
```

↓

it will give horizontal bar

create pie chart

```
plt.figure(figsize = (7,7))
mtcars["gear"].value_counts()
mtcars["cyl"].value_counts()
```

correlation plot →
- Positive correlation ⇒ speed ↑ distance covered ↑
- Negative correlation ⇒ age ↑ peace ↓
- No - correlation ⇒ rainfall in Bangalore vs US economy

Histogram

```
plt.hist(mtcars['mpg'], bins = 10)
```

Histogram

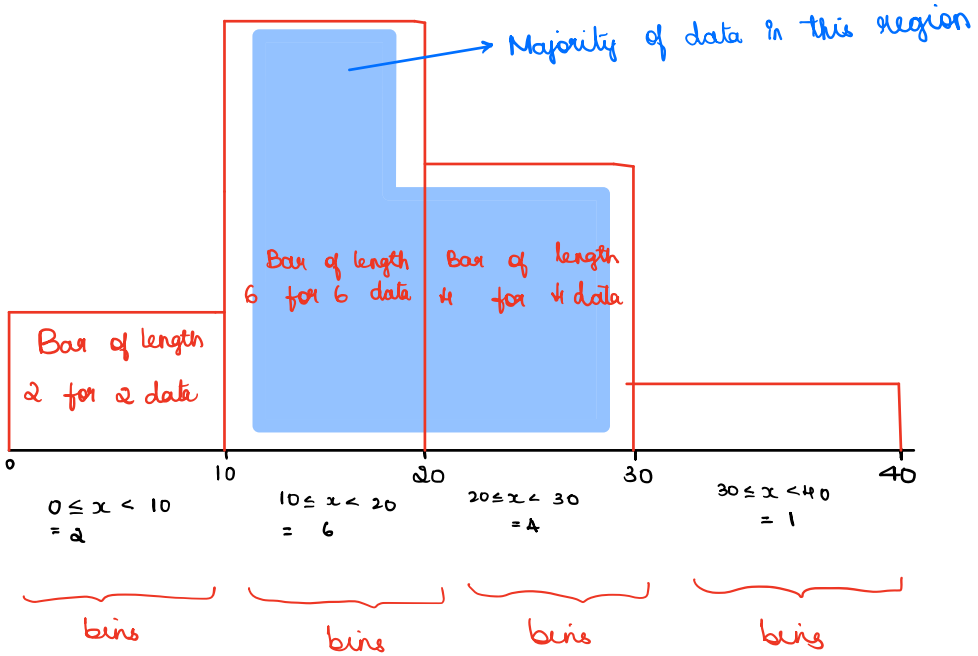
→ It will help us understand distribution of the data.

ex: data = 1, 3, 10, 13, 12, 10, 17, 15, 20, 25, 27, 23, 24, 37

ascending: 1, 3, 10, 10, 12, 13, 15, 17, 20, 23, 24, 25, 27, 37

→ histogram ⇒ Once you get the data divide them into small small bins.

minvalue = 1
maxvalue = 37
} range 1 to approx 40



Ex: BLR house prices → are their distribution same

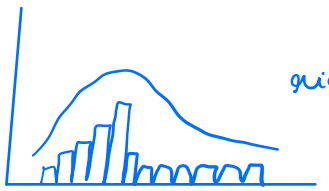
HYD house prices } → We can get this data by comparing their house prices.

→ check for mean

→ check for variance

Skewed data ⇒ Majority of the data is one sided

ex.



right skewed data

How to select bins = ? ——→ In large data 50 bins are good

Boxplot

plt.figure(figsize=(8,6), facecolor='green')

plt.box(mtcars['mpg'], vert=False)

Violin Plot

xlabel = 'mpg' data

We can see that spread is higher in some region and spread is lower in some region.

ie majority of data is in this region ⇒ region where spread is more.

Seaborn

import seaborn as sns

```
df = sns.load_dataset('tips')
```

EDA ⇒ Exploratory / Exhaustive data analysis.

Once you get the data what questions can you ask?

ex: Is there correlation between bill and tip

Can gender of people affect the amount of tip

Will a smoker leave more tip?

→ do smoker & gender, correlate with tip

→ connect tip with the rest of data

→ Once we ask questions, try to answer them using analysis

Stripplot

two items at a time

```
sns.stripplot(x='time', y='tip', data=df, jitter=True)
plt.ylabel('Tip ($)')
plt.show()
```

```
df.groupby[['smoker'],['tips']].describe()
```

three types of data

```
sns.stripplot(x='day', y='tip', hue='sex', data=df, jitter=True)
plt.ylabel('Tip ($)')
plt.show()
```

Swarmplot → This is similar to violin plot

```
sns.swarmplot(y='tip', data=df)
plt.ylabel('tip($)')
plt.show()
```

Box and violin plot

Subplot

→ Create multiple plots in same layout.

```
plt.subplot(1, 2, 1) ——→ 1st plot
```

1 → row is 1

2 → column is 2

Random Variable

1) Discreet random variable
2) Continuous random variable

1) **Discreet random variable :**

probability distribution function

Total probability should always be 1

Probability is always positive and its inbetween 0 to 1

ex $P_1 + P_2 = 1$

$0 \leq P_1 \leq 1$ & $0 \leq P_2 \leq 1$

Problem: The daily sales of a large flat panel TV's at a store

0 = 0.4

1 = 0.25

2 = 0.2

3 = 0.05

4 = 0.10

a) What is the probability of a sale? 0.6

b) probability of selling atleast 3 TVs?

atleast 3 TV ⇒ 3 or 4

$P(x=3) + P(x=4) = 0.05 + 0.1 = 0.15$

Bernoulli Random Variable: Random Variable that has only two outcomes / two possibilities

Bernoulli's distribution tries to find the probability of first and probability of second distribution.

Binomial distribution

⇒ Outcomes are more than 2 & within a range.

→ they are related

↓ Binomial distribution can be broken down in Bernoulli's

ex : 5 exams and I have to fail in all the exams

$= P((X_1 = 0) \cap (X_2 = 0) \cap (X_3 = 0) \cap (X_4 = 0) \cap (X_5 = 0))$ ⇒ 15 independent events

$= P(X_1 = 0) * P(X_2 = 0) * P(X_3 = 0) * P(X_4 = 0) * P(X_5 = 0)$

$= (0.5)^5 = (1/2)^5 = 1/32$

Probability = $\frac{\text{favourable cases}}{\text{total number of case}}$

2) Continuous Random Variables

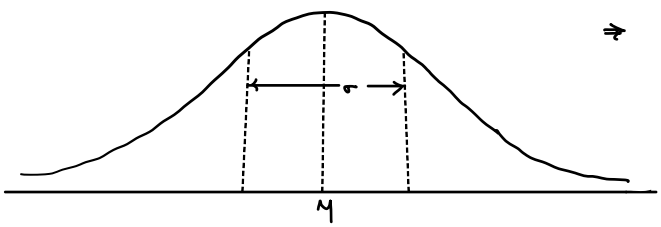
ex : Temperature can take infinite values

Therefore one particular temperature will be $1/\infty$ value $\simeq 0$

probability of single value will be 0

probability of a range will be value

⇒ Here we will consider density.

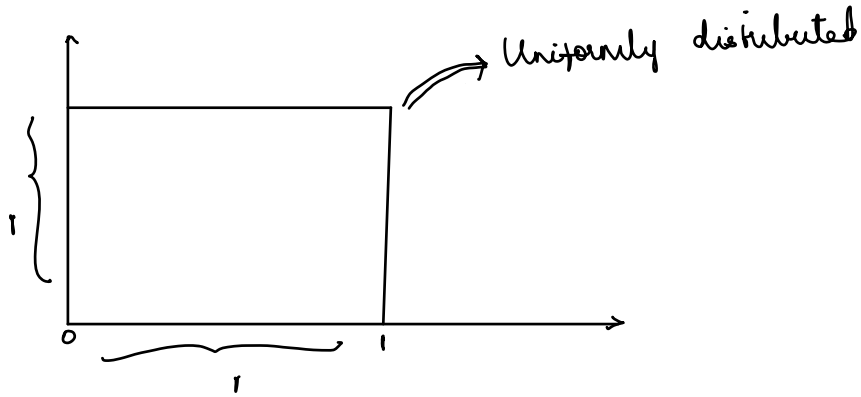


⇒ for each continuous distribution we have density

→ area under the curve (for a particular range) will give

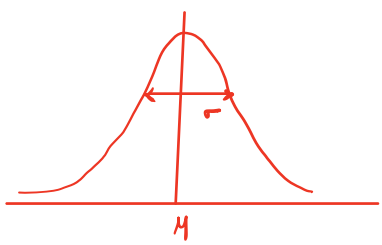
the probability in that region.

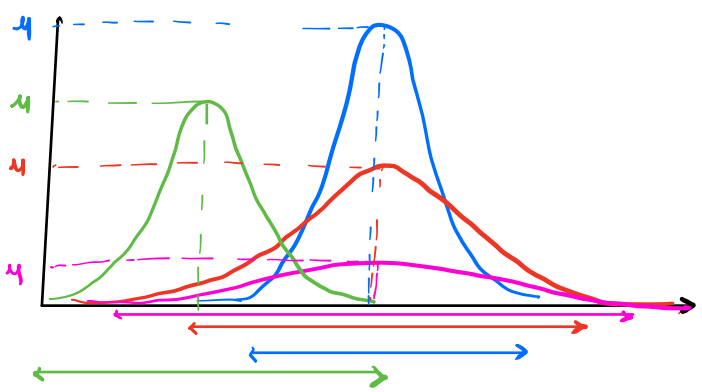
Uniformly distributed data



⇒ If we have density graph, the total area under the graph must be 1

Normal distribution





if you add a constant, only mean will change but distribution Variance will be same

if we multiply a constant, mean and variance both will change.