Topics:

- Box-plot; Bar-plot & scatter plot → prob.. (2000+ retail-data)
- Probability distributions
  - PMF, CDF
  - PDF
  - Bernouli & Binomial disb
  - Airticcels overbooking

Ops:

- Questions: Track → Question-tab → 
- chat: communication; Yes/no

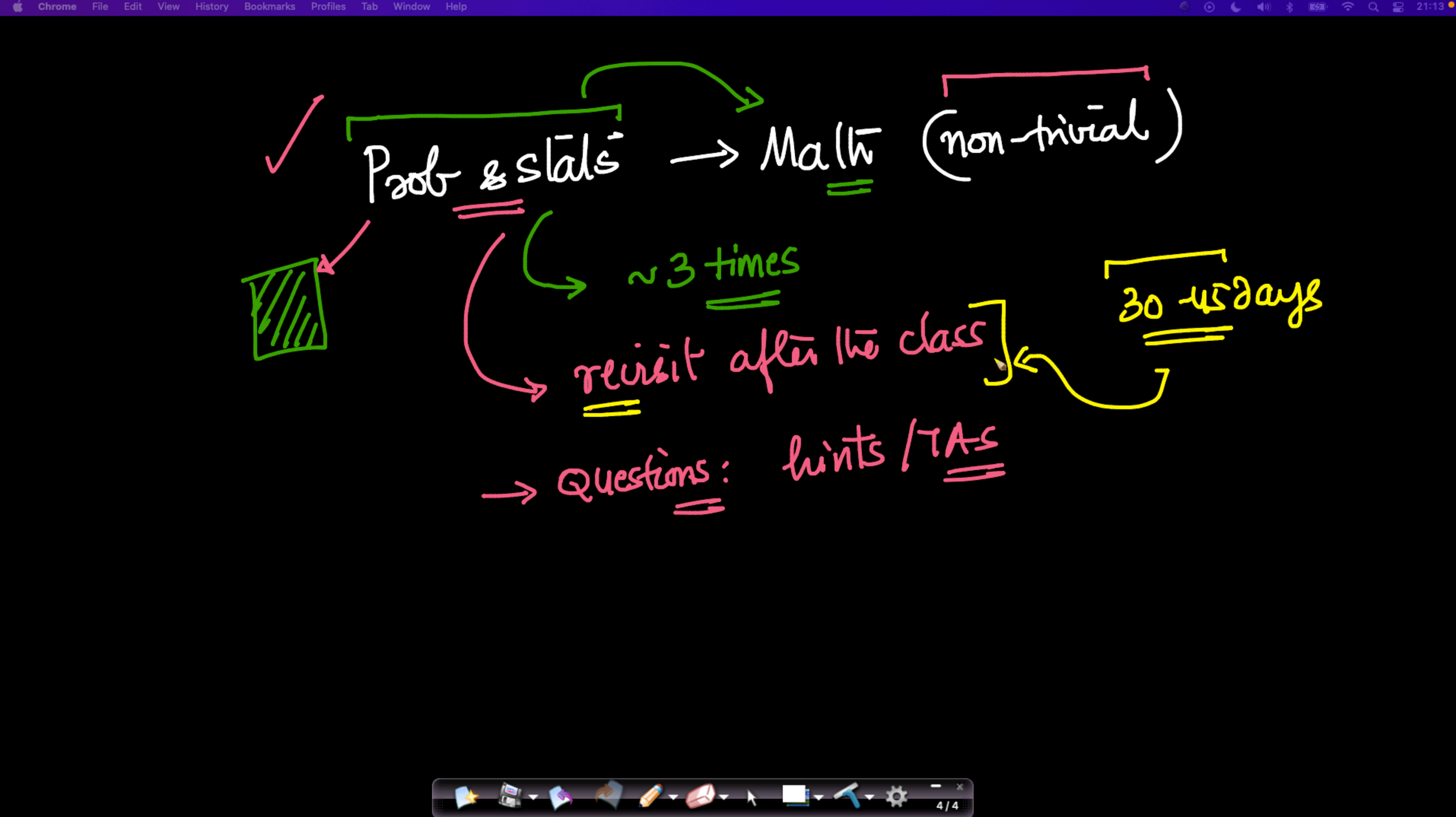
Bots: go-slow; pick-up-speed

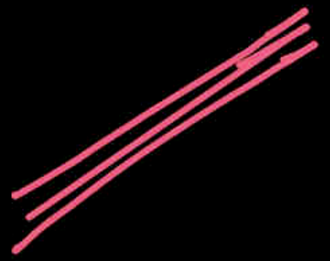
End: ----

→ scratch-pad: pdf of code

→ future: detailed text-notes

↳ 'TA' (or) whatsapp/slack

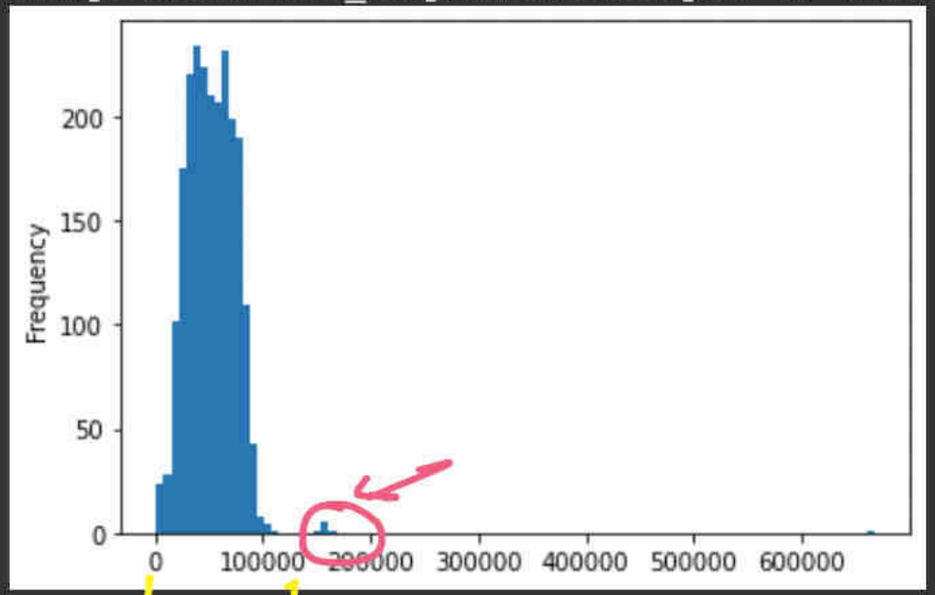

Prob & stats
Math
(non-trivial)
~3 times
revisit after the class
30-45 days
Questions: hints / TAs



Box plot

```
[ ] df[' Income '].plot.hist(bins=100)
    # few extreme values to the right.
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb674d2450>
```



histogram
KDE

```
#box plot

ax = df[' Income '].plot.box()
ax.set_ylabel("Income in Dollars")
```

```
Text(0, 0.5, 'Income in Dollars')
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Colal | New Tab

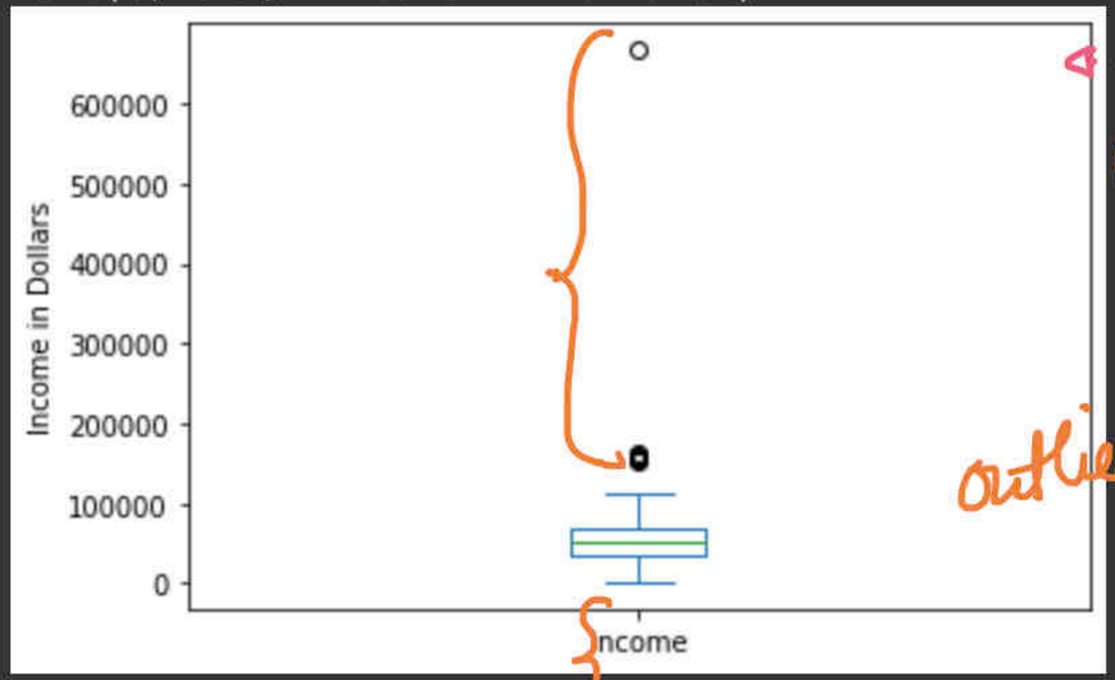
colab.research.google.com/drive/1l5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=SGqzqcJJbVA9

+ Code + Text

RAM Disk

```
[ ]  #box plot

ax = df[' Income '].plot.box()
ax.set_ylabel("Income in Dollars")
```

```
Text(0, 0.5, 'Income in Dollars')
```



outliers

$1.5\,IQR + Q_3$

whiskers

$Q_3 = P_{75}$

$Q_2 = P_{50}$

$Q_1 = P_{25}$

box

$Q_1 - 1.5\,IQR$

outliers

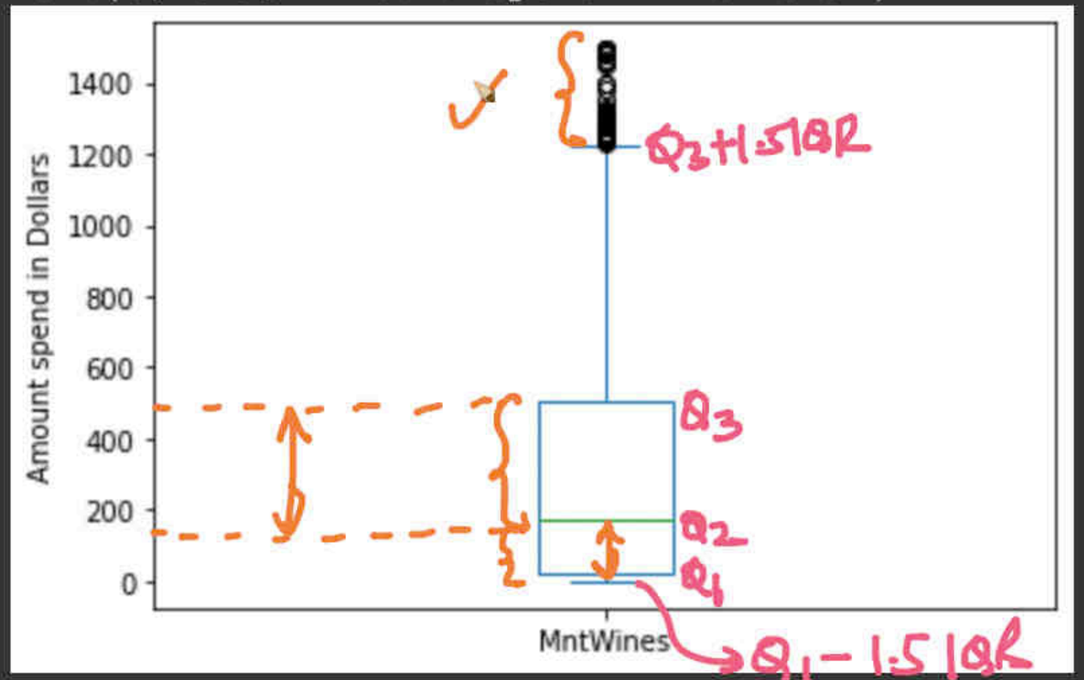
Income

```
[ ]  #box plot
```

7 / 8

```
[ ] #box plot
    ax = df['MntWines'].plot.box()
    ax.set_ylabel("Amount spend in Dollars")
```

```
Text(0, 0.5, 'Amount spend in Dollars')
```



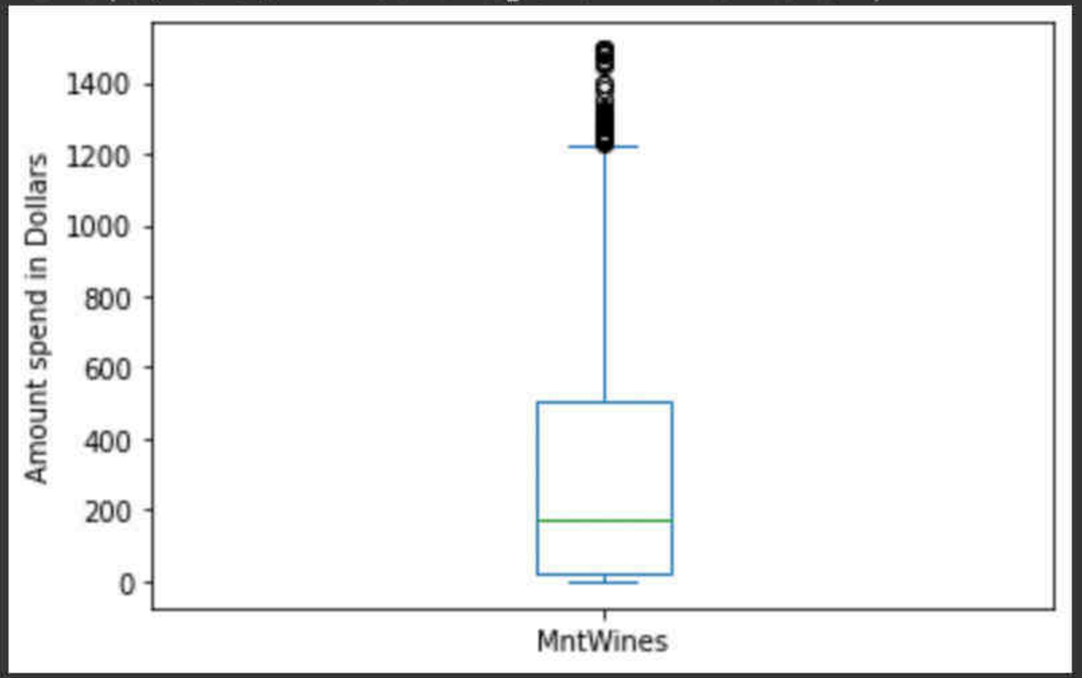

```
[ ] #bar plot
    df['Education'].value_counts().plot.bar()
```

+ Code + Text

RAM
Disk

```
[ ] #box plot
    ax = df['MntWines'].plot.box()
    ax.set_ylabel("Amount spend in Dollars")
```

```
Text(0, 0.5, 'Amount spend in Dollars')
```




```
[ ] #bar plot
    df['Education'].value_counts().plot.bar()
```

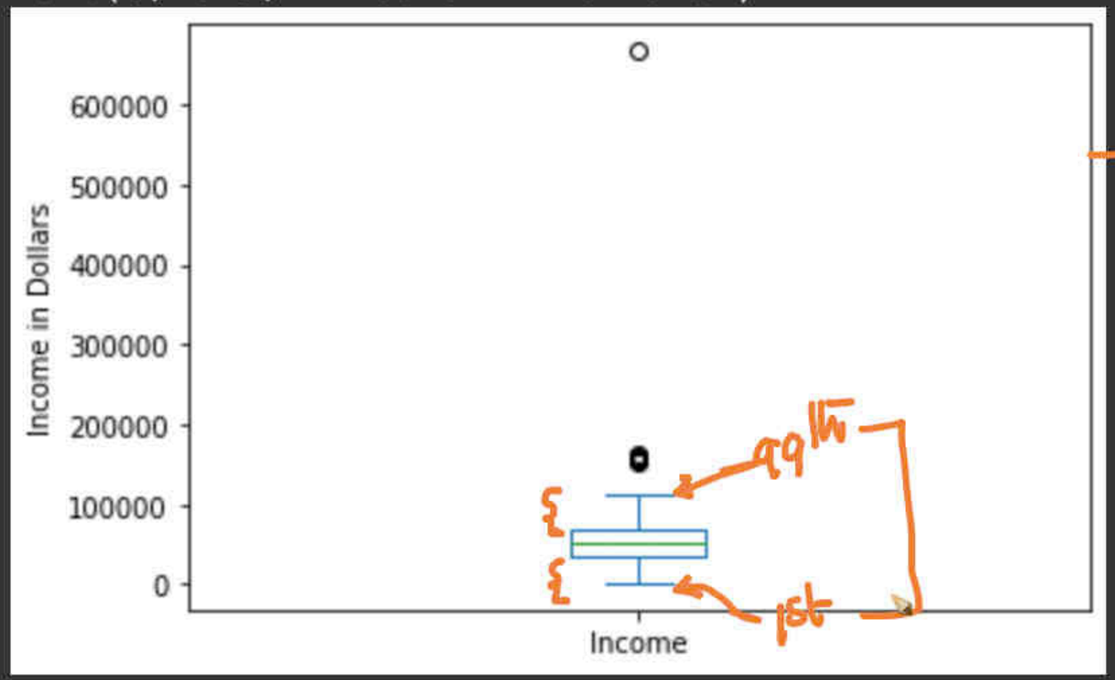
Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Colab | New Tab

colab.research.google.com/drive/1I5T7TVIAASw9TdI4JxqJxuDqRgFBGOW3#scrollTo=SGqzqcJJbVA9

+ Code + Text

RAM Disk

```
[ ]  Text(0, 0.5, 'Income in Dollars')
```



→ 1.5 IQR

→ whiskers:

```
#box plot
ax = df['MntWines'].plot.box()
ax.set_ylabel("Amount spend in Dollars")
```

```
Text(0, 0.5, 'Amount spend in Dollars')
```

10 / 11

pandas.DataFrame.plot.hexbin
pandas.DataFrame.plot.hist
pandas.DataFrame.plot.kde
pandas.DataFrame.plot.line
pandas.DataFrame.plot.pie
pandas.DataFrame.plot.scatter
pandas.DataFrame.boxplot
pandas.DataFrame.hist
pandas.DataFrame.sparse.density
pandas.DataFrame.sparse.from_spmatrix
pandas.DataFrame.sparse.to_coo
pandas.DataFrame.sparse.to_dense
pandas.DataFrame.from_dict
pandas.DataFrame.from_records
pandas.DataFrame.to_parquet
pandas.DataFrame.to_pickle
pandas.DataFrame.to_csv
pandas.DataFrame.to_hdf
pandas.DataFrame.to_sql
pandas.DataFrame.to_dict
pandas.DataFrame.to_excel
pandas.DataFrame.to_json
pandas.DataFrame.to_html
pandas.DataFrame.to_feather

pandas.DataFrame.boxplot

DataFrame.boxplot(*column=None, by=None, ax=None, fontsize=None, rot=0, grid=True, figsize=None, layout=None, return_type=None, backend=None, **kwargs*) [source]

Make a box plot from DataFrame columns.

Make a box-and-whisker plot from DataFrame columns, optionally grouped by some other columns. A box plot is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers extend from the edges of box to show the range of the data. By default, they extend no more than *1.5 * IQR (IQR = Q3 - Q1)* from the edges of the box, ending at the farthest data point within that interval. Outliers are plotted as separate dots.

For further details see Wikipedia's entry for boxplot.

Parameters: **column** : *str or list of str, optional*

Column name or list of names, or vector. Can be any valid input to `pandas.DataFrame.groupby()`.

by : *str or array-like, optional*

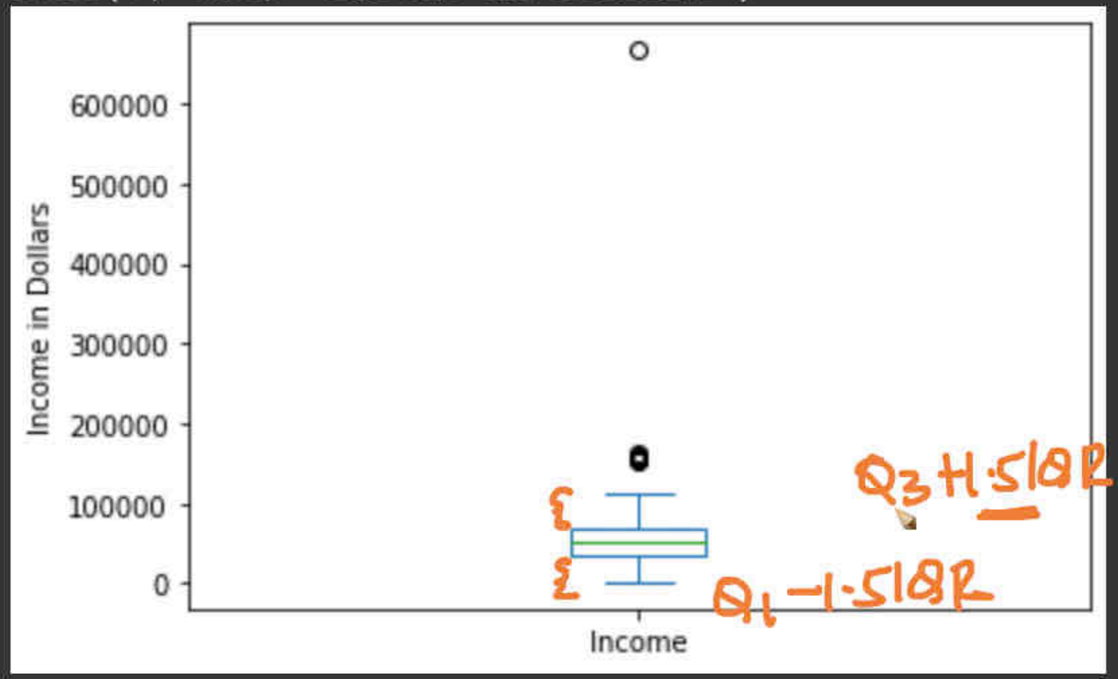
Column in the DataFrame to `pandas.DataFrame.groupby()`. One box-plot will be done per value of columns in *by*.

ax : *object of class matplotlib.axes.Axes, optional*

The matplotlib axes to be used by boxplot.

```
ax = df[' Income '].plot.box()
ax.set_ylabel("Income in Dollars")
```

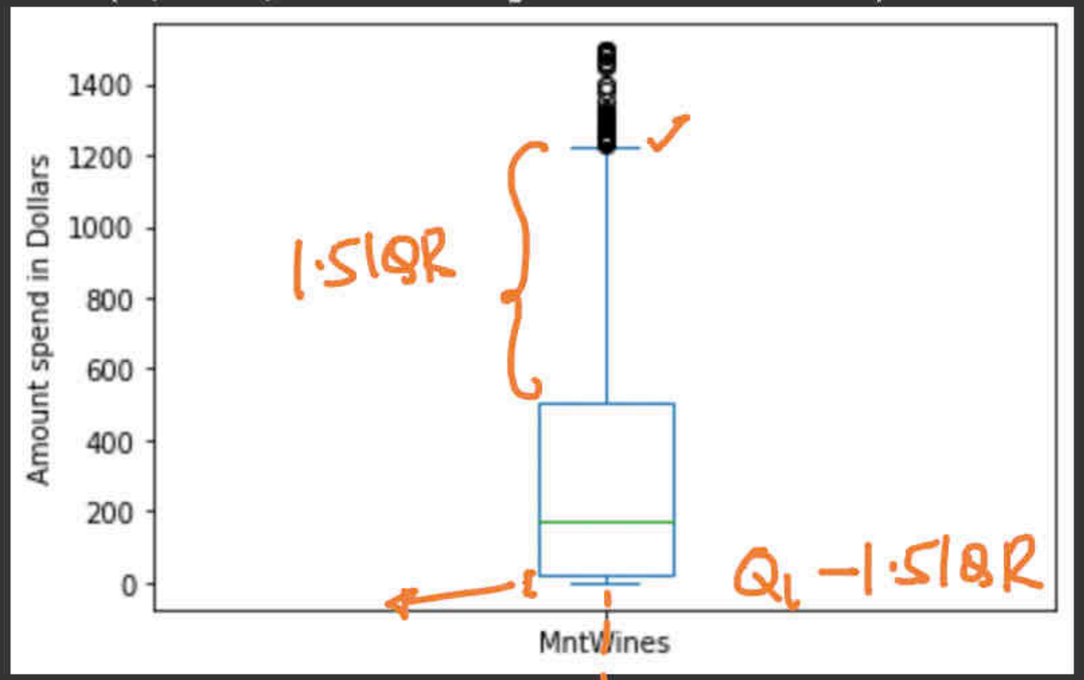
```
Text(0, 0.5, 'Income in Dollars')
```




```
[ ] #box plot
    ax = df['MntWines'].plot.box()
    ax.set_ylabel("Amount spend in Dollars")
```

```
[ ] #box plot
    ax = df['MntWines'].plot.box()
    ax.set_ylabel("Amount spend in Dollars")
```

```
Text(0, 0.5, 'Amount spend in Dollars')
```



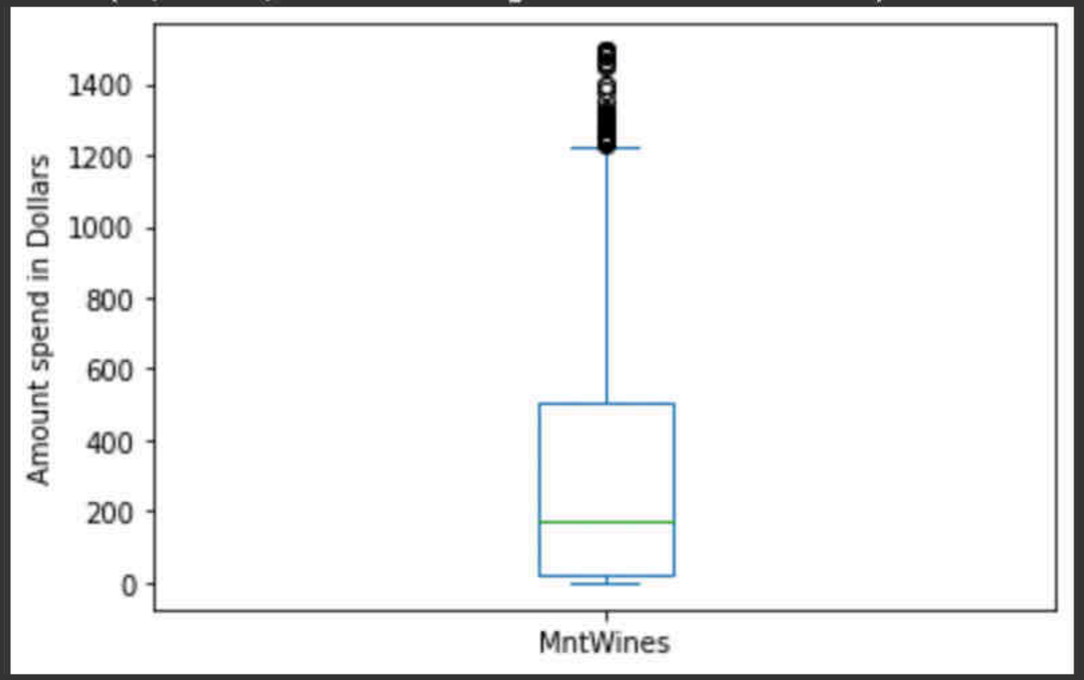
pandas: 1·5IQR (or) Max(min

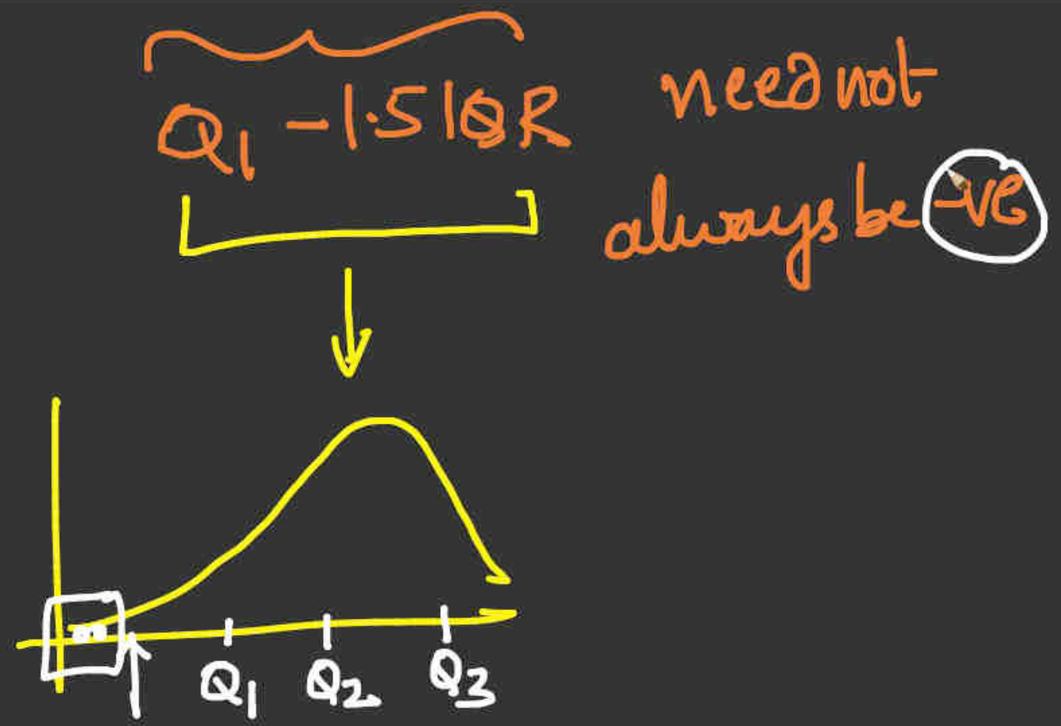
Q1 −1·5IQR < Min

```
[ ] #bar plot
    df['Education'].value_counts().plot.bar()
```

```
[ ] #box plot
    ax = df['MntWines'].plot.box()
    ax.set_ylabel("Amount spend in Dollars")
```

```
Text(0, 0.5, 'Amount spend in Dollars')
```






```
[ ] #bar plot
    df['Education'].value_counts().plot.bar()
```

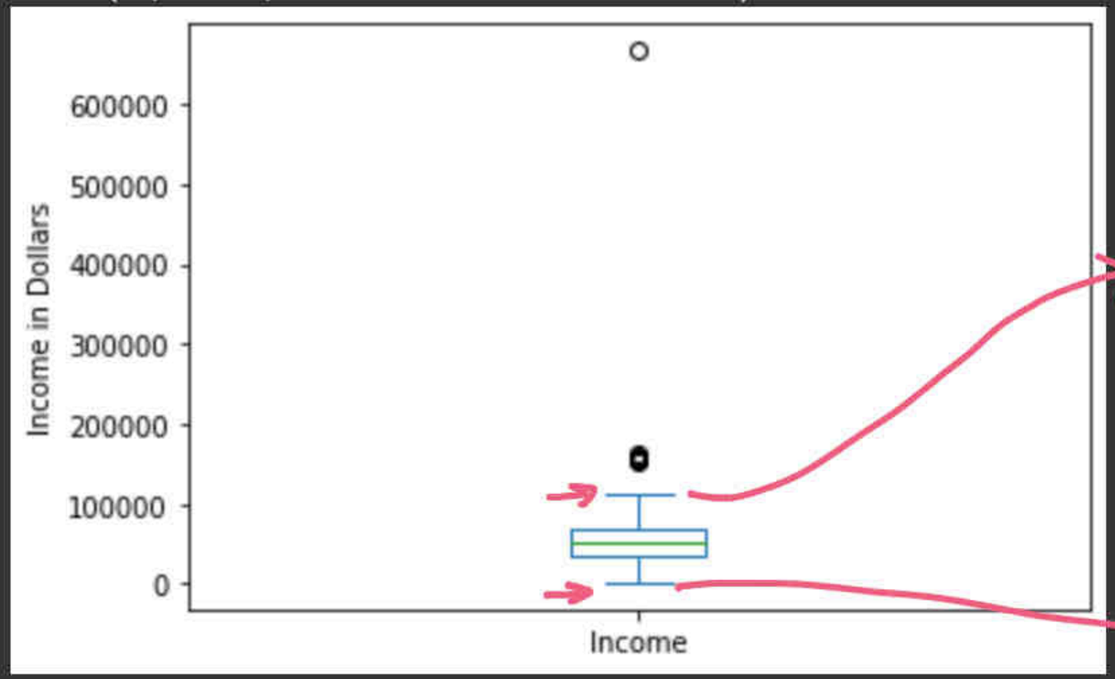
Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot –

colab.research.google.com/drive/1l5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=qhN6-BWVpojf

+ Code + Text

RAM Disk

```
Text(0, 0.5, 'Income in Dollars')
```



pandas:

min

UW = (Q3 + 1.5 IQR, Max r.v)

LW = max (Q1 - 1.5 IQR, Min r.v)

```
#box plot
ax = df['MntWines'].plot.box()
ax.set_ylabel("Amount spend in Dollars")
```

```
Text(0, 0.5, 'Amount spend in Dollars')
```

15 / 17

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot —

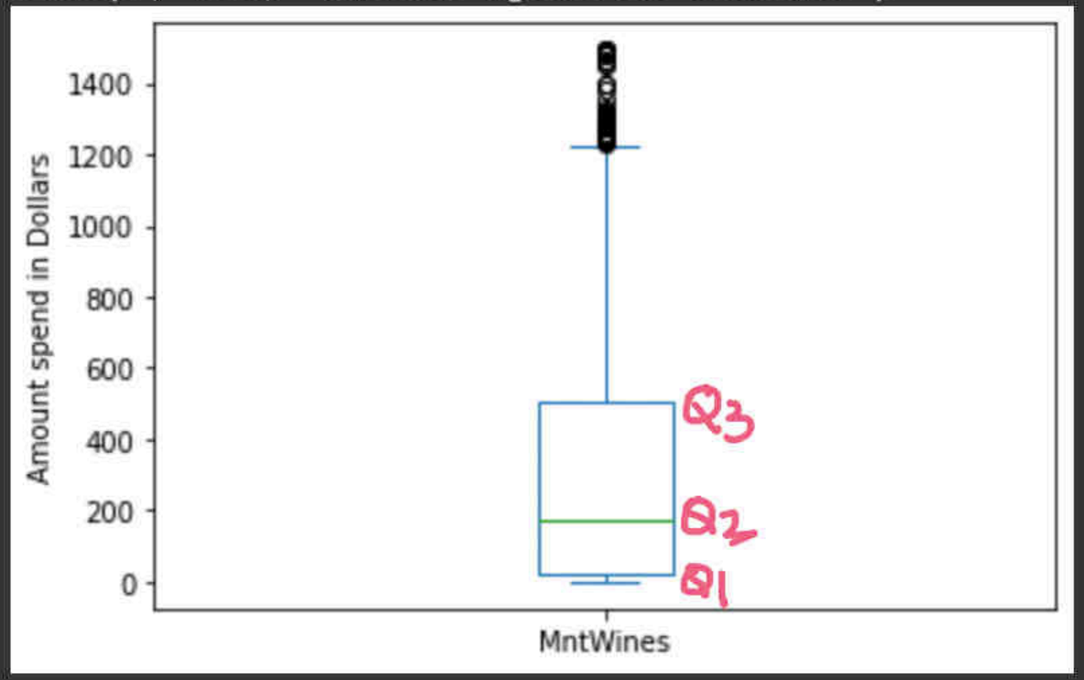
colab.research.google.com/drive/1I5T7TVIAASw9TdI4JxqJxuDqRgFBGOW3#scrollTo=qhN6-BWVpojf

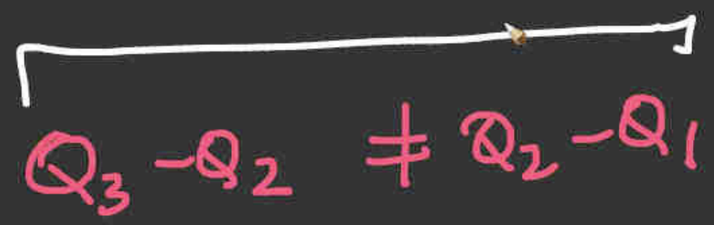
+ Code + Text

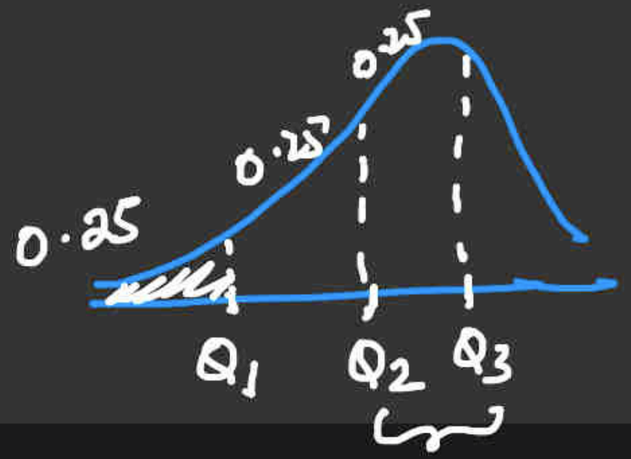
RAM
Disk

[ ]

```
Text(0, 0.5, 'Amount spend in Dollars')
```








```
#bar plot
df['Education'].value_counts().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb53e3f9d0>
```

16 / 17

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot –

colab.research.google.com/drive/1I5T7TVIAASw9TdI4JxqJxuDqRgFBGOW3#scrollTo=qhN6-BWVpojf

+ Code + Text

RAM
Disk

[ ]

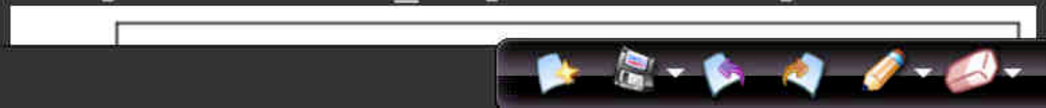
```
Text(0, 0.5, 'Amount spend in Dollars')
```

box-plot → ques ...

[ ]

```
#bar plot
df['Education'].value_counts().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb53e3f9d0>
```

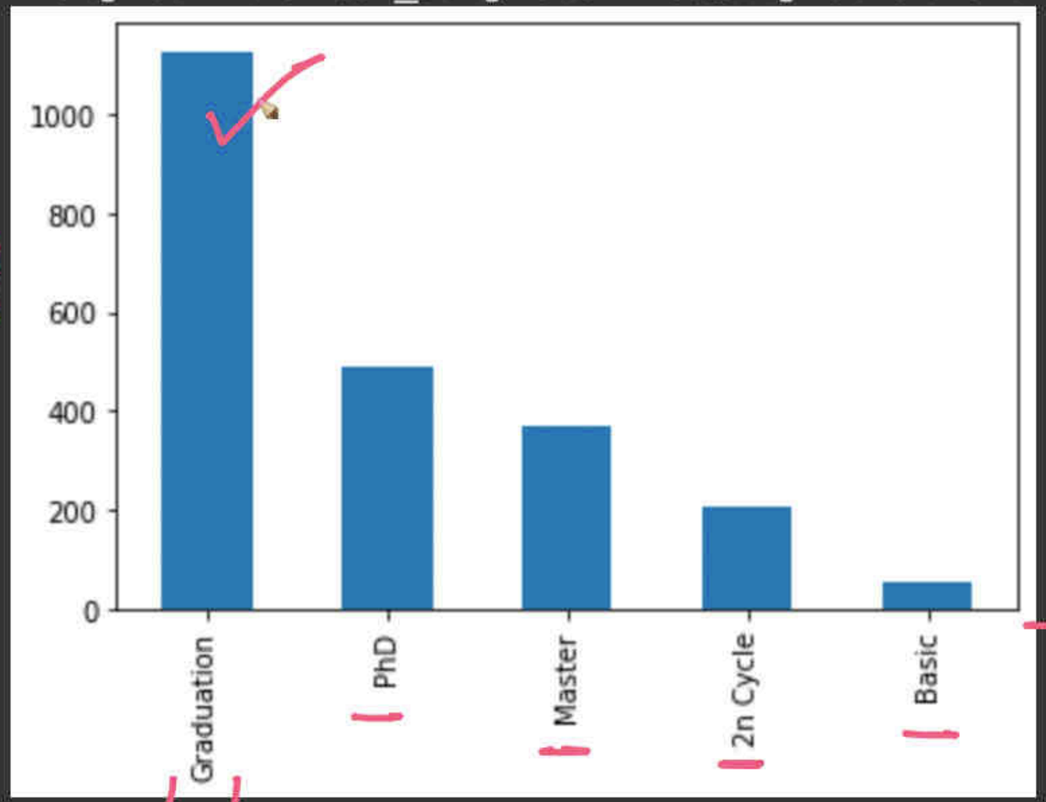


17 / 18

```
[ ]  #bar plot
     df['Education'].value_counts().plot.bar()
```

bar plot

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb53e3f9d0>
```

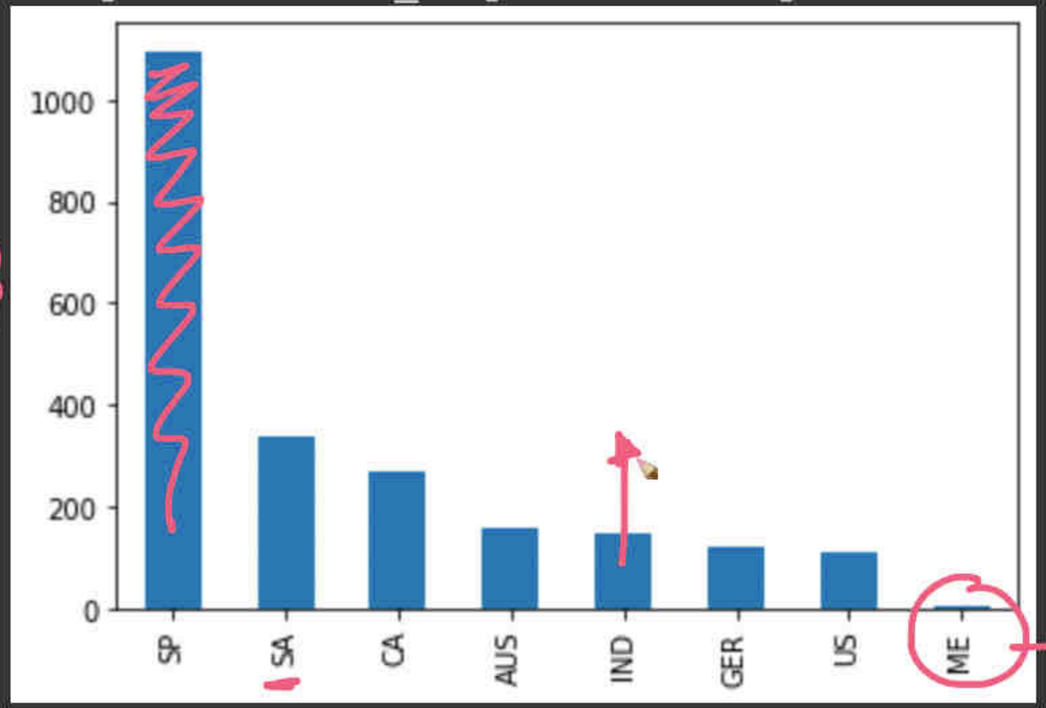


freq

→ values discrete r.v

```
df['Country'].value_counts().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb53db4f90>
```



freq

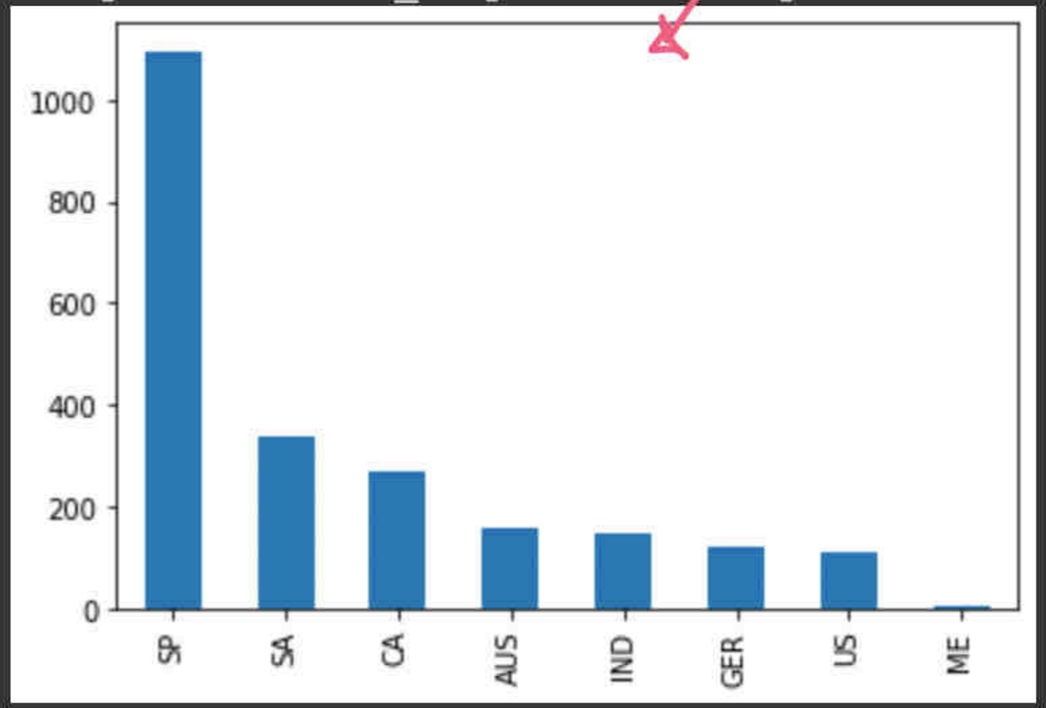
2000 +

country

```
[ ]  df['Marital_Status'].value_counts()
```

```
df['Country'].value_counts().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb53db4f90>
```



Categorical discrete r.v

```
[ ] df['Marital_Status'].value_counts()
```

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot —

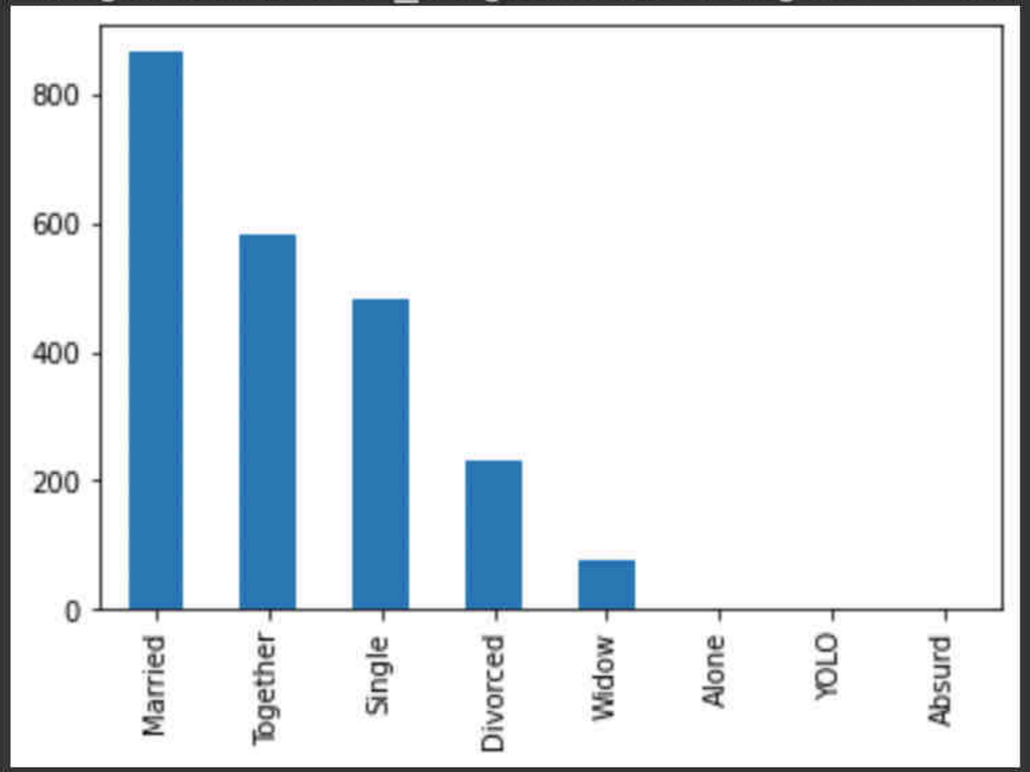
colab.research.google.com/drive/1l5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=A2JIMTNsnlMy

+ Code + Text

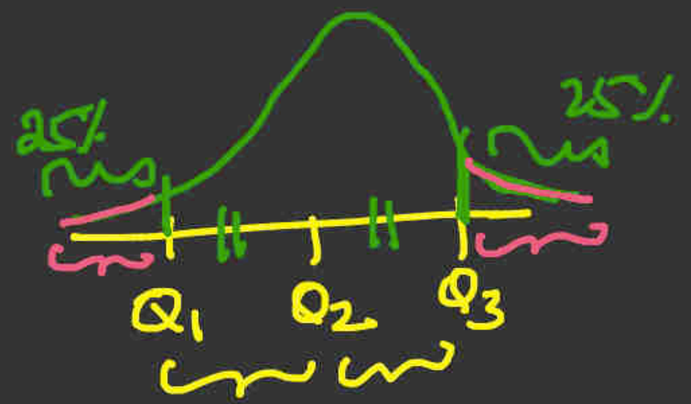
RAM
Disk

```
df['Marital_Status'].value_counts().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff8e90e8490>
```



$Q_3 - Q_2 \neq Q_2 - Q_1$



21 / 22

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot —

colab.research.google.com/drive/1l5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=A2JIMTNsnlMy

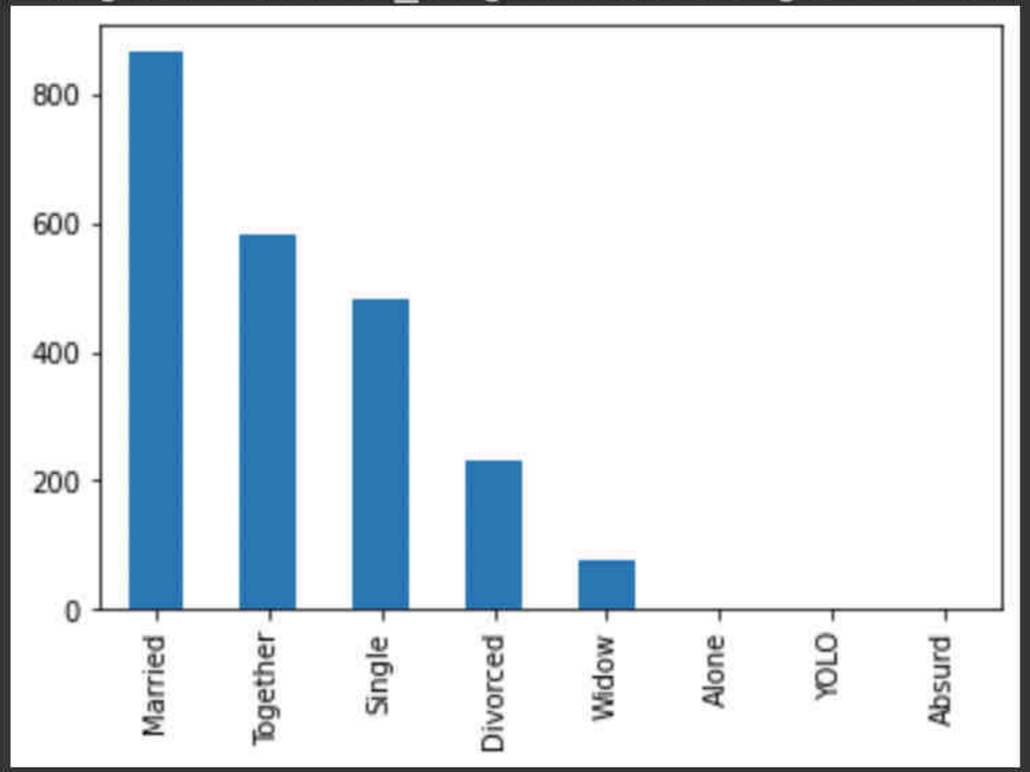
+ Code + Text

RAM
Disk

```
df['Marital_Status'].value_counts().plot.bar()
```

1s

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff8e90e8490>
```



numpy → { Seaborn, pandas } → Matplotlib

scipy.stats

22 / 23

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Colal | New Tab | pandas.DataFrame.boxplot —

colab.research.google.com/drive/1l5T7TVIAASw9TdI4JxqJxuDqRgFBGOW3#scrollTo=A2JIMTNsnlMy

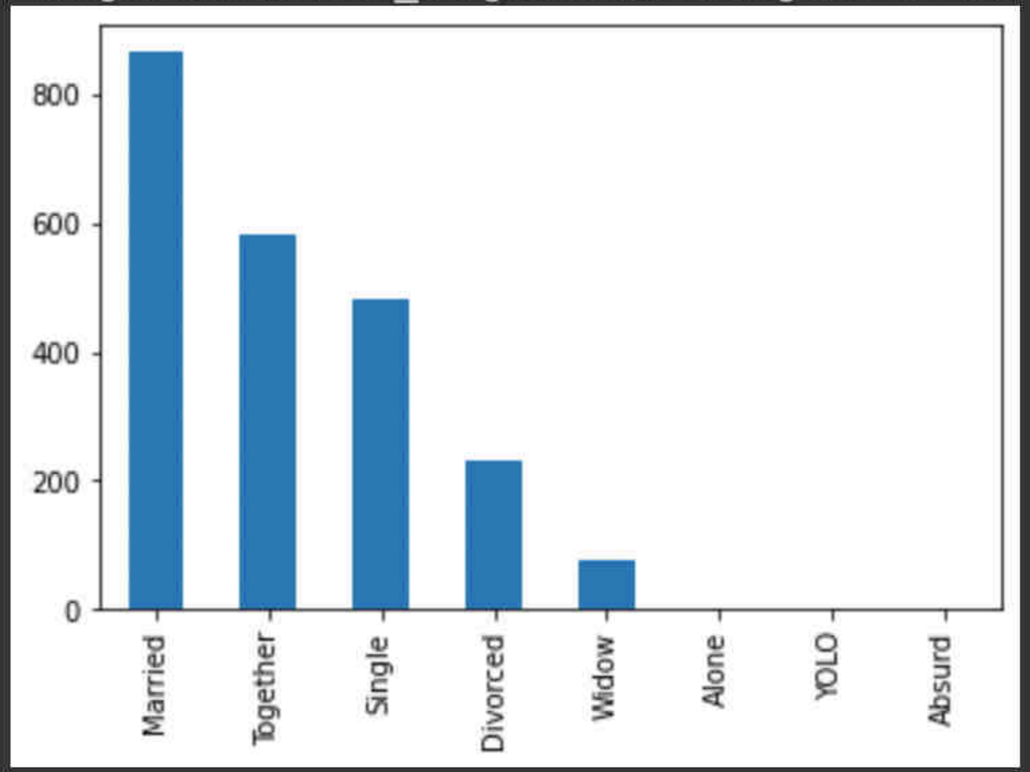
+ Code + Text

RAM
Disk

```
df['Marital_Status'].value_counts().plot.bar()
```

1s

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff8e90e8490>
```



integer → numerical → Continous (often)

age

→ categorical (ordinal) (nase)

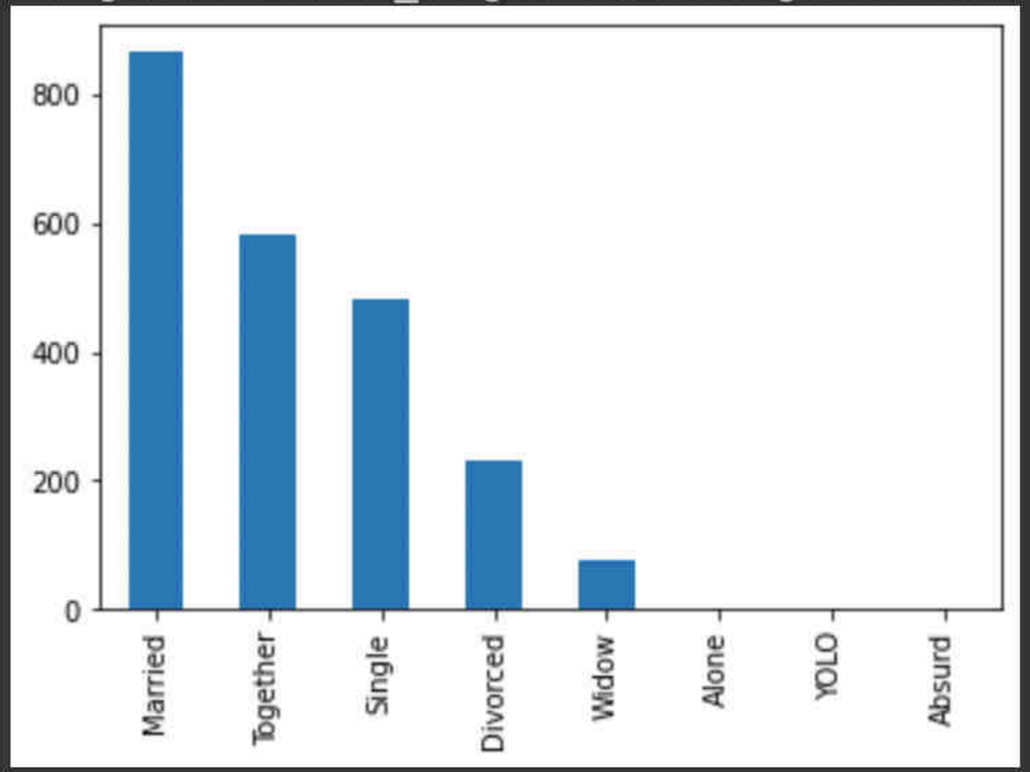
4, 5, 6, ... 90

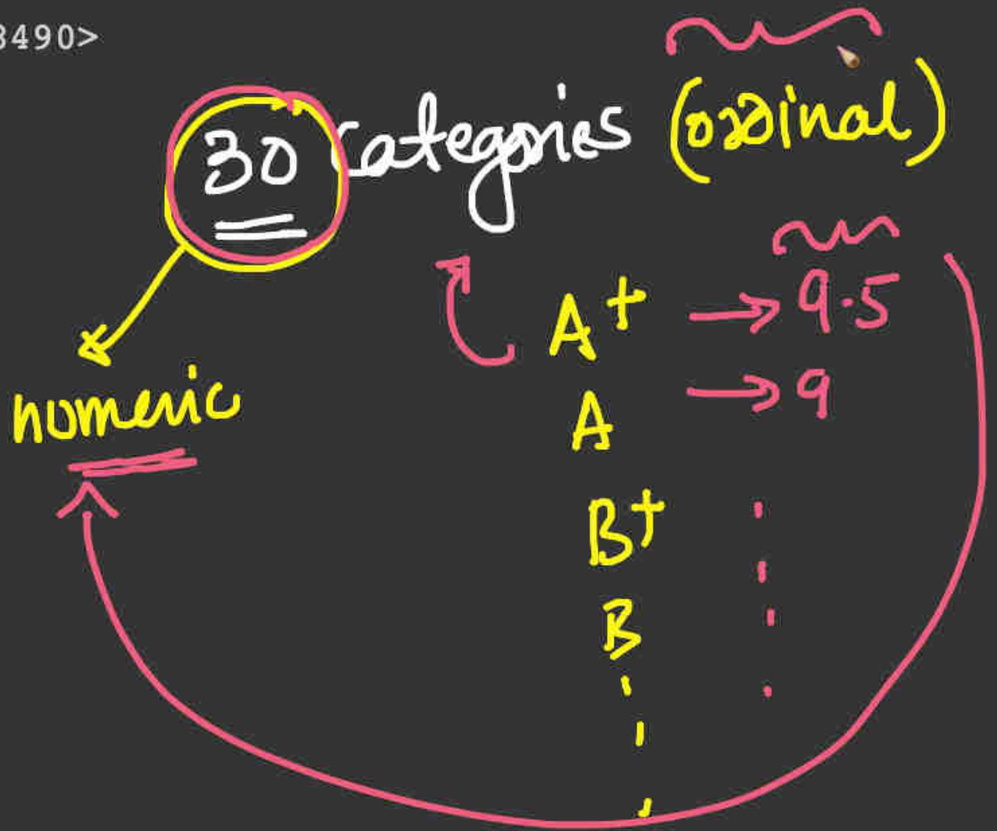
{ 10 | 20 | ...

23 / 24

```
df['Marital_Status'].value_counts().plot.bar()
```

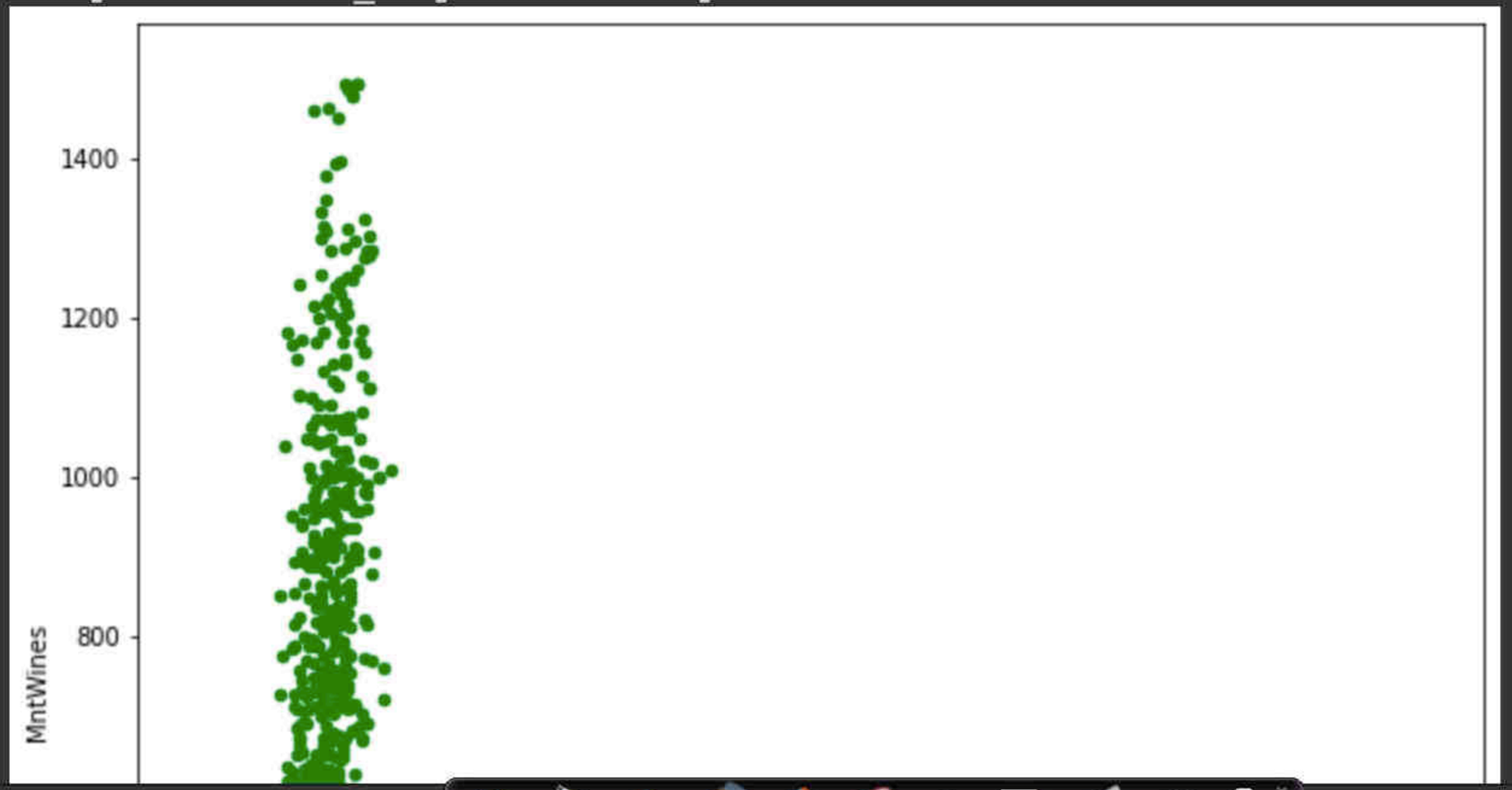
```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff8e90e8490>
```





```
#scatter plot
df.plot.scatter(x=' Income ', y='MntWines',  figsize = (10,10),c='green')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb53e3f310>
```



Chrome File Edit View History Bookmarks Profiles Tab Window Help 21:46

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot —

colab.research.google.com/drive/1I5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=A2JIMTNsnlMy

Update

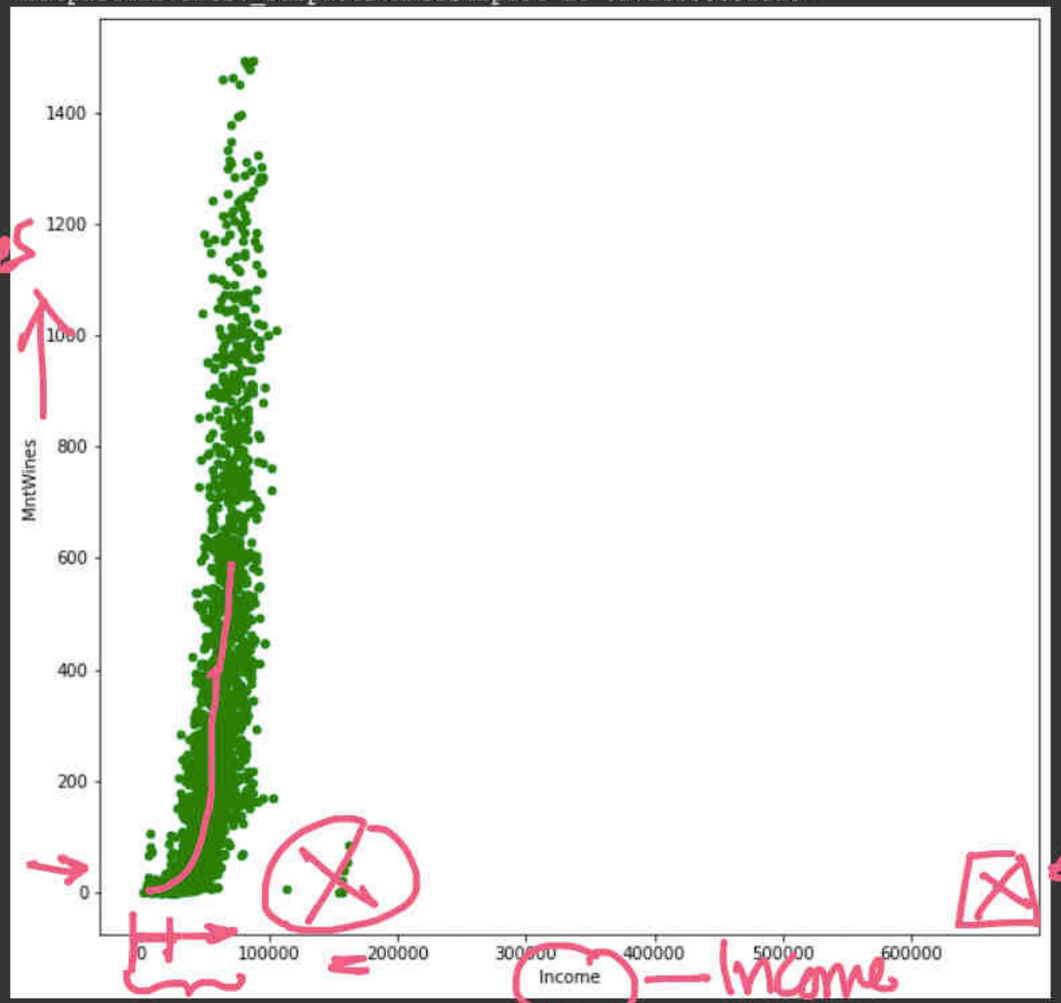
+ Code + Text

RAM Disk

```
ar.plot.scatter(x= 'Income' , y= 'MntWines' , figsize = (10,10), c= 'green' )
```

[ ]

<matplotlib.axes._subplots.AxesSubplot at 0x7feb53e3f310>



wines

Income

relationship

general trend:

Income ↑ ; Mnt wines ↑

[ ] plt.figure(figsize=(8,8))
plt.subplot(2,2,1)

26 / 26

Chrome File Edit View History Bookmarks Profiles Tab Window Help 21:46

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot –

colab.research.google.com/drive/1I5T7TVIAASw9TdI4JxqJxuDqRgFBGOW3#scrollTo=A2JiMTNsnlMy

Update

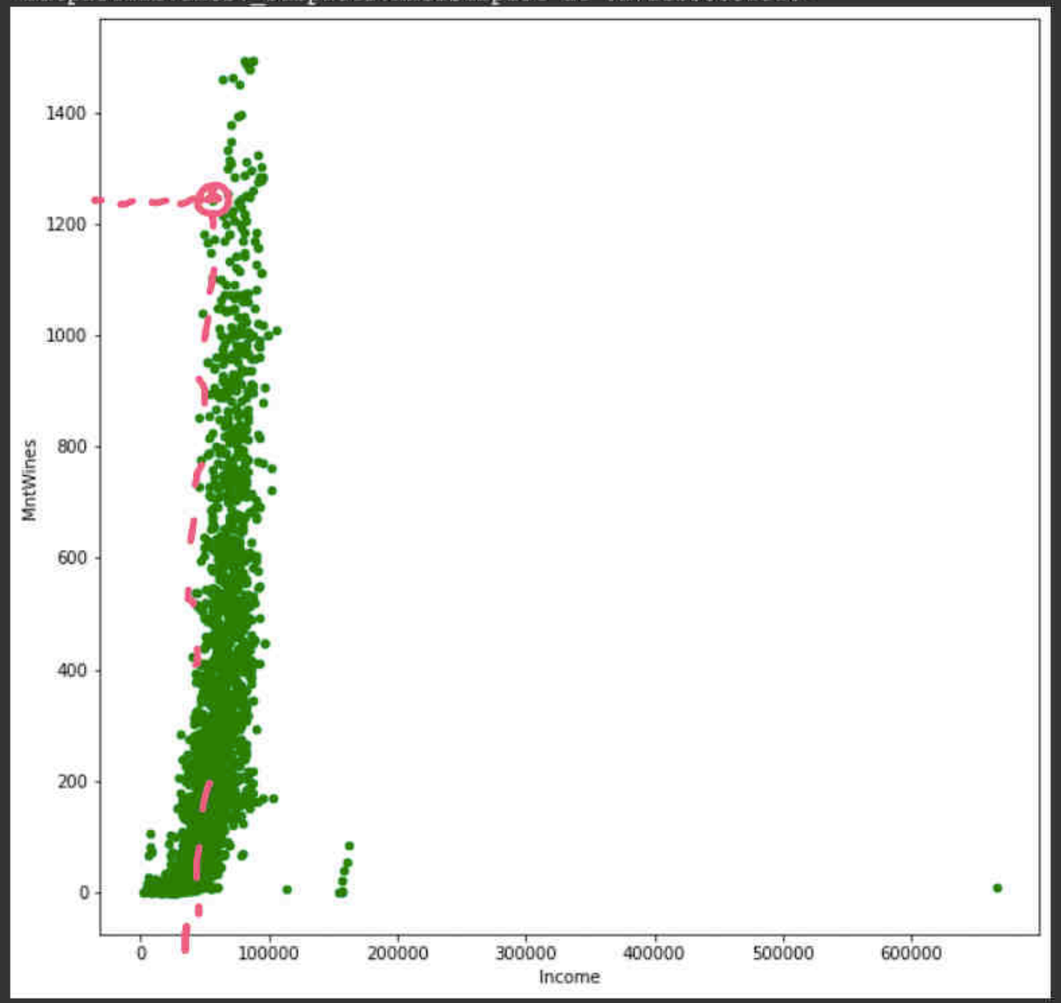
+ Code + Text

RAM Disk

```
ar.plot.scatter(x= 'Income' , y= 'MntWines' , figsize = (10,10), c= 'green' )
```

[ ]

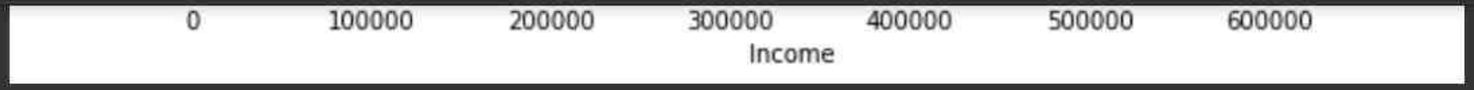
```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb53e3f310>
```



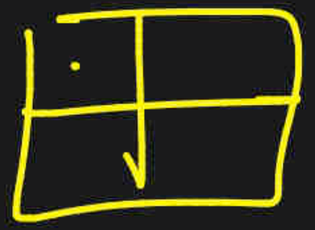
2000 +

```
[ ] plt.figure(figsize=(8,8))
    plt.subplot(2,2,1)
```

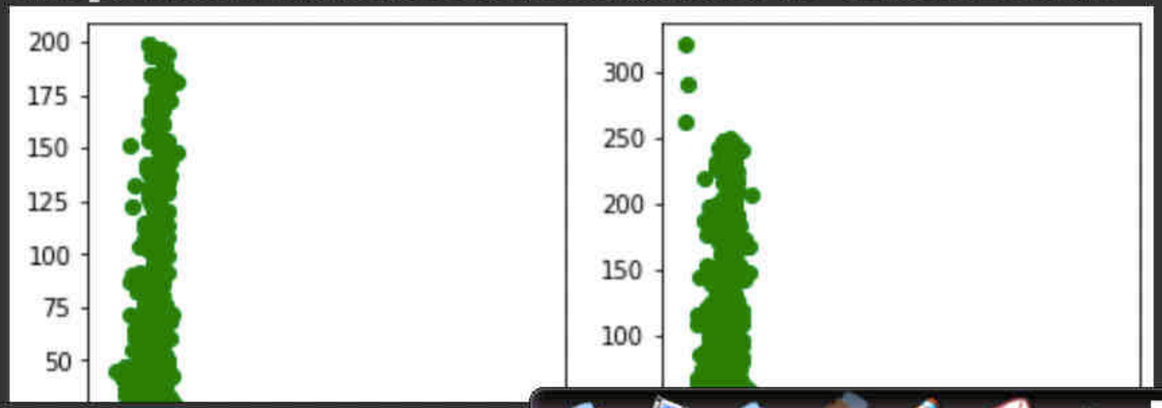
27 / 27



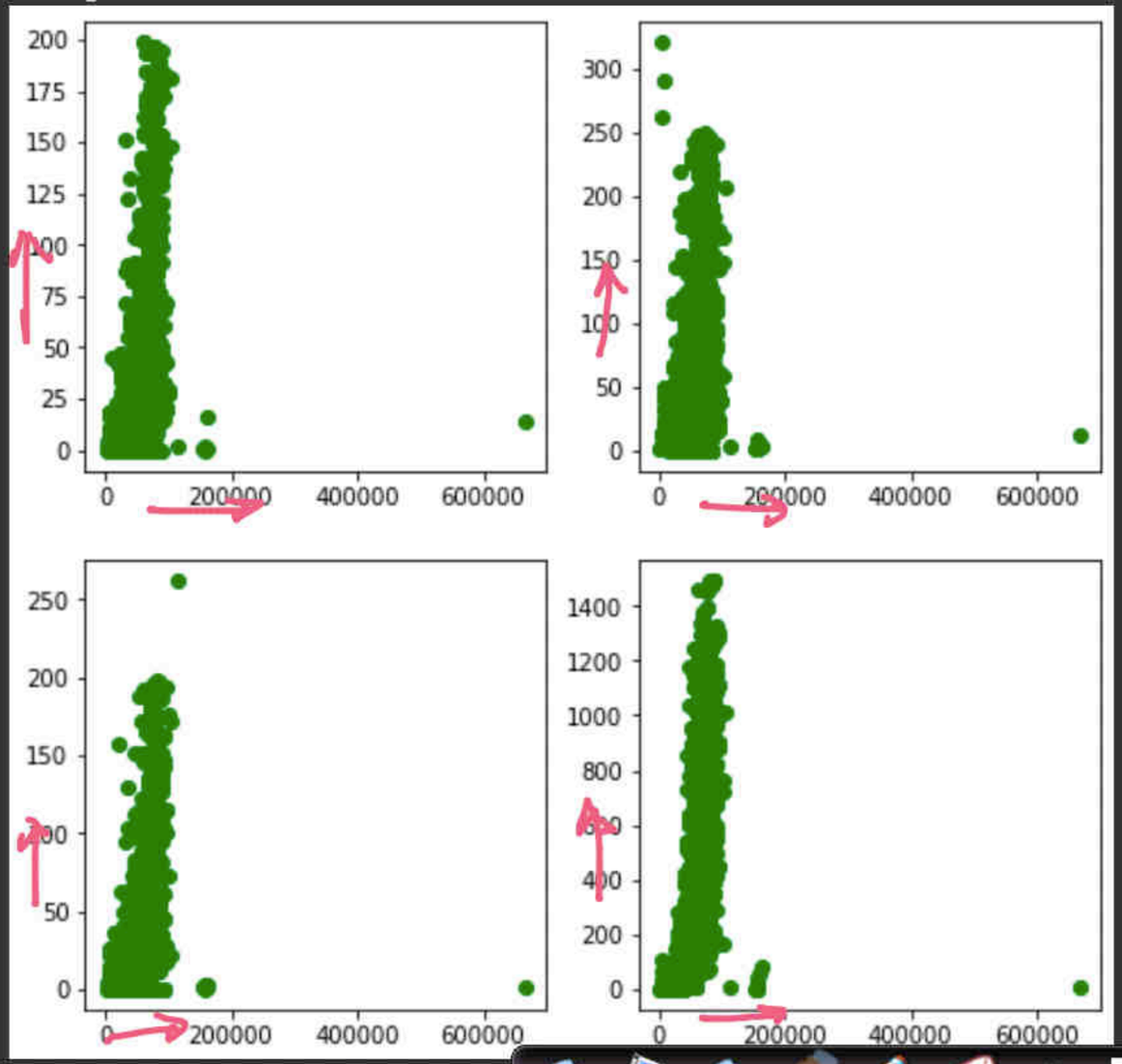

```
plt.figure(figsize=(8,8))
plt.subplot(2,2,1)
plt.scatter(x=df[' Income '], y=df['MntFruits'], c='green')

plt.subplot(2,2,2)
plt.scatter(x=df[' Income '], y=df['MntGoldProds'], c='green')

plt.subplot(2,2,3)
plt.scatter(x=df[' Income '], y=df['MntSweetProducts'], c='green')

plt.subplot(2,2,4)
plt.scatter(x=df[' Income '], y=df['MntWines'], c='green')
```



```
<matplotlib.collections.PathCollection at 0x7feb53a78d50>
```



<matplotlib.collections.PathCollection at 0x7feb53a78d50>



In general

Income ↑

Mnt- ... ↑

0 200000 400000 600000 0 200000 400000 600000

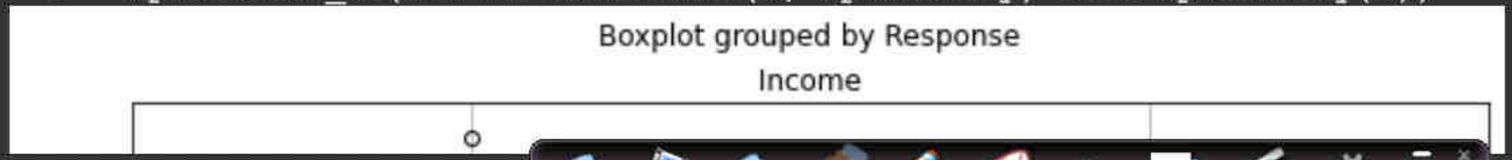
```
# Have they responded to the marketing survey by our company
df['Response']
```

Prime -Sub

```
0       1
1       1
2       0
3       0
4       1
       ..
2235    0
2236    0
2237    0
2238    0
2239    1
Name: Response, Length: 2240, dtype: int64
```

```
df.boxplot(by='Response', column=' Income ', figsize=(10,6))
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning: Creating an nda
  X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
```

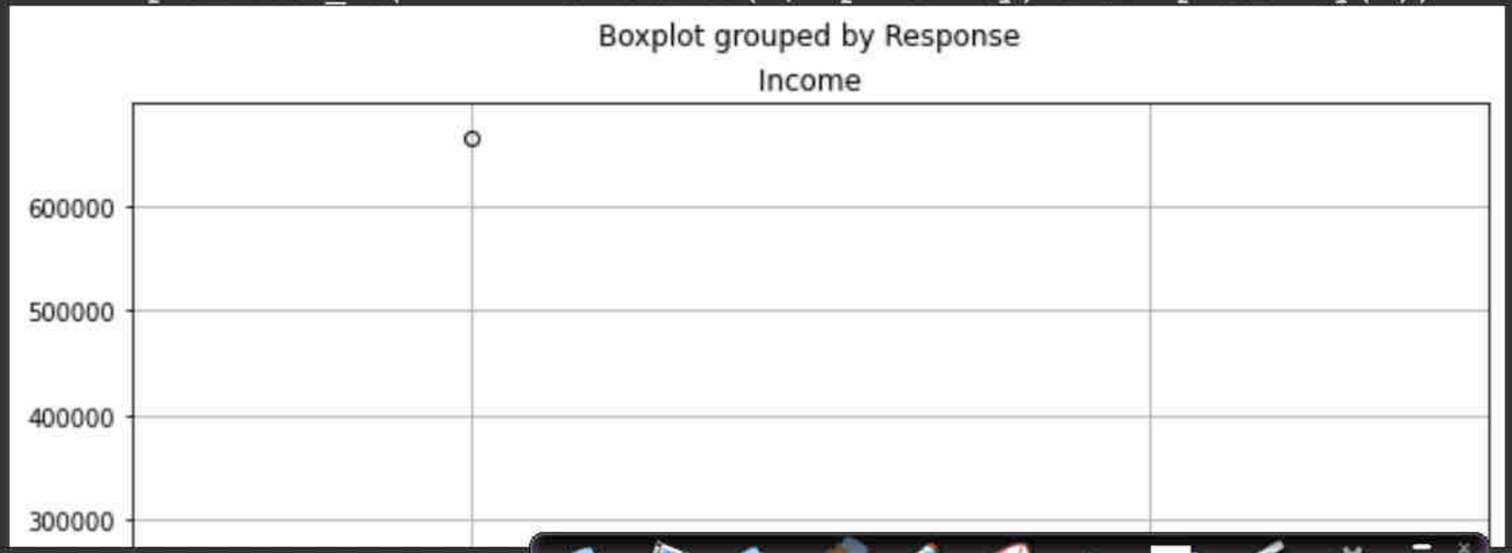


```
4       1
        ..
2235    0
2236    0
2237    0
2238    0
2239    1
Name: Response, Length: 2240, dtype: int64
```

```
[ ]  df.boxplot(by='Response', column=' Income ', figsize=(10,6))
     plt.show()
```

1 0

```
/usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning: Creating an nda
  X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
```



31 / 31

Chrome File Edit View History Bookmarks Profiles Tab Window Help 21:52

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot —

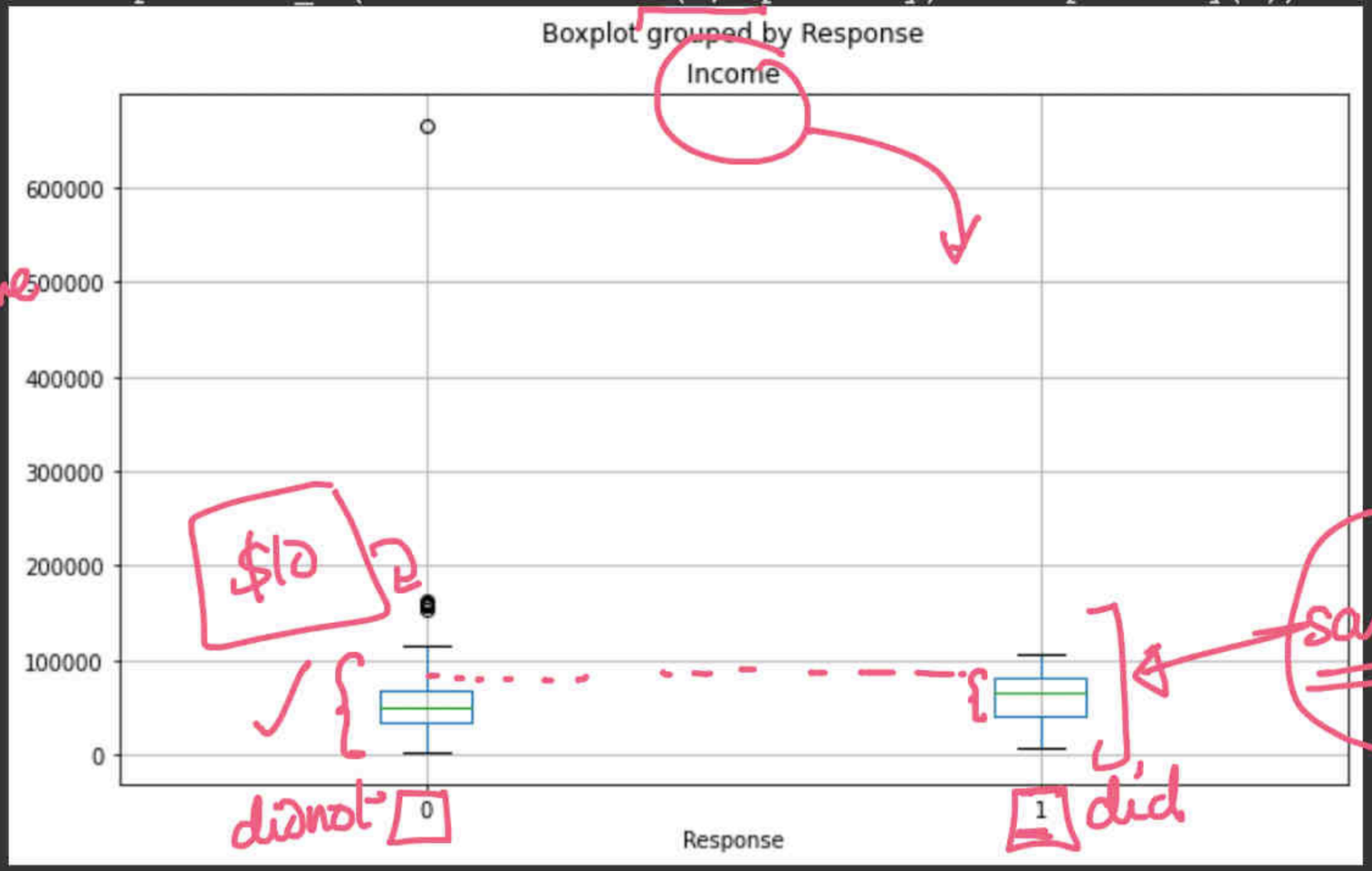
colab.research.google.com/drive/1l5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=pUstIlAPqkDB

Update

+ Code + Text

RAM Disk

```
/usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning: Creating an nda
[ ]   X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
```



Income

$10

diabetes

did

sample

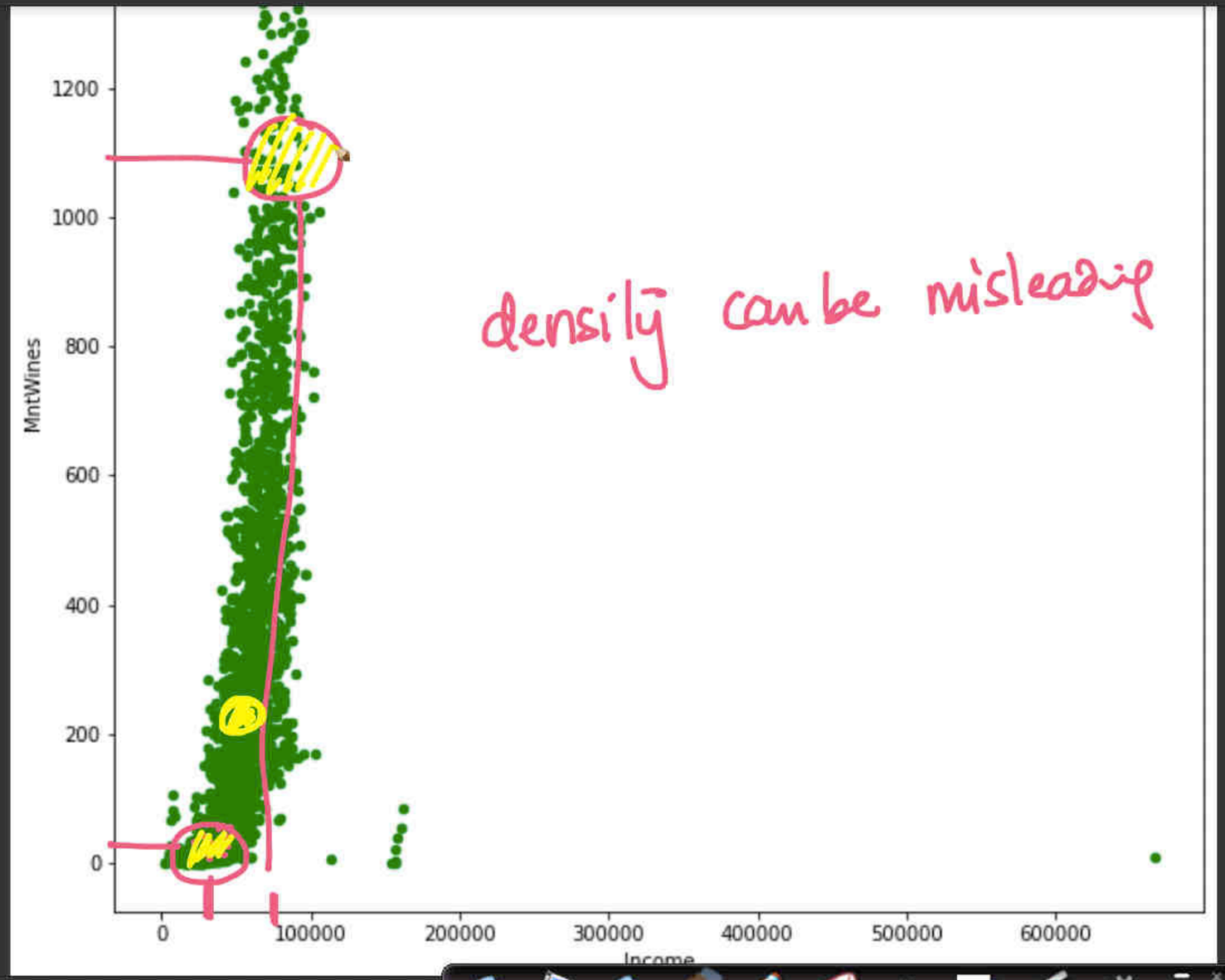
survey:

1

2240

0

32 / 32


density can be misleading
MntWines
1200
1000
800
600
400
200
0
0
100000
200000
300000
400000
500000
600000
Income

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot —

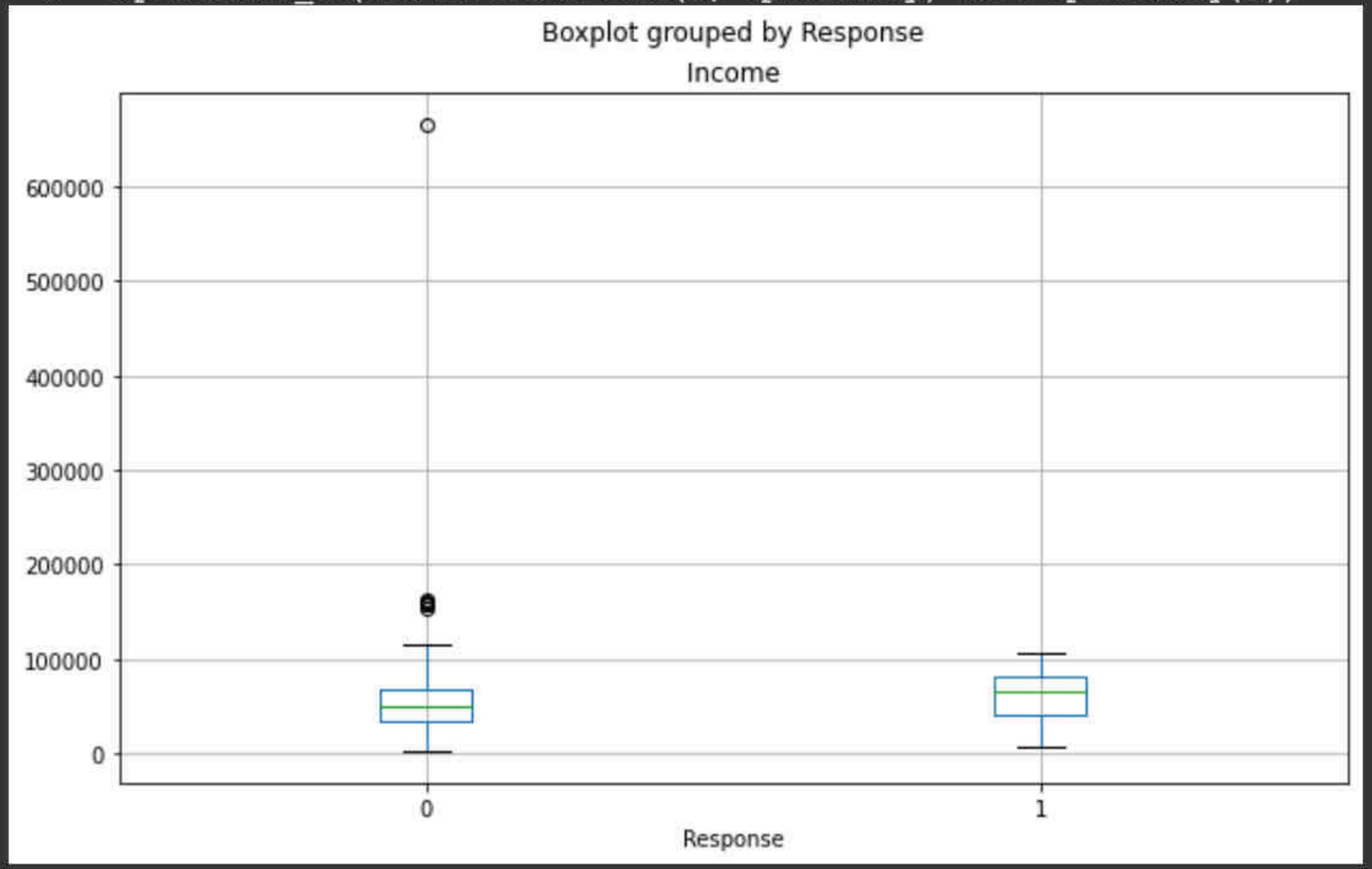
colab.research.google.com/drive/1l5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=Egg1hlOlqLJ7

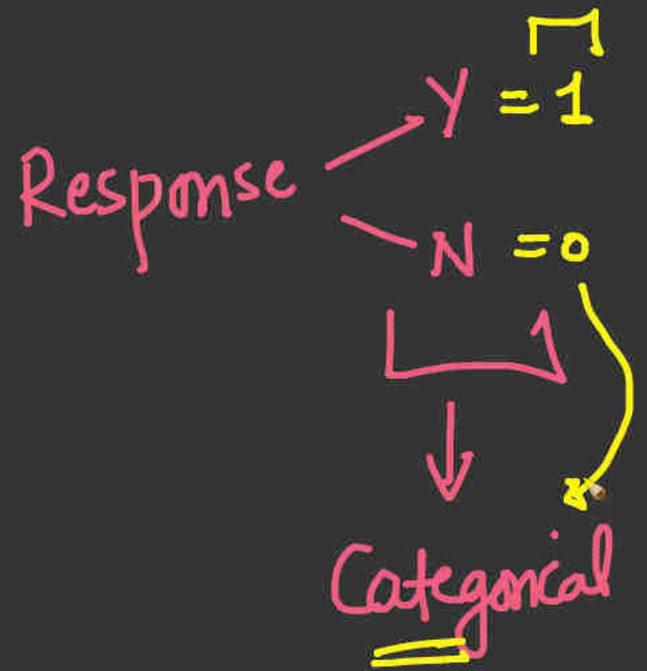
Update

+ Code + Text

RAM
Disk

```
[ ] /usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning: Creating an nda
    X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
```





34 / 34

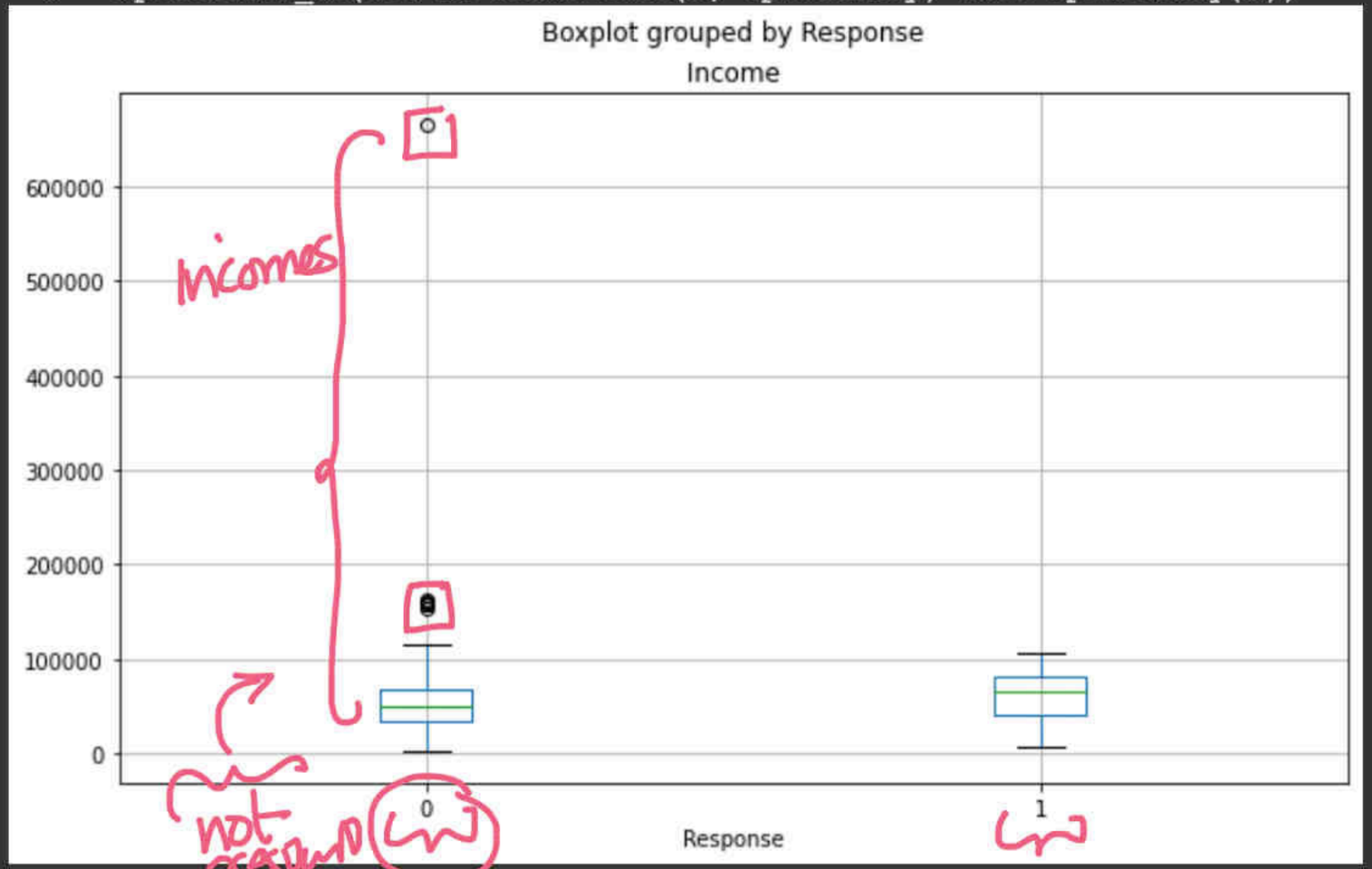
Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot —

colab.research.google.com/drive/1l5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=Egg1hlOlqLJ7

+ Code + Text

RAM
Disk

```
[ ] /usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning: Creating an nda
    X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
```



35 / 35

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Cola | New Tab | pandas.DataFrame.boxplot —

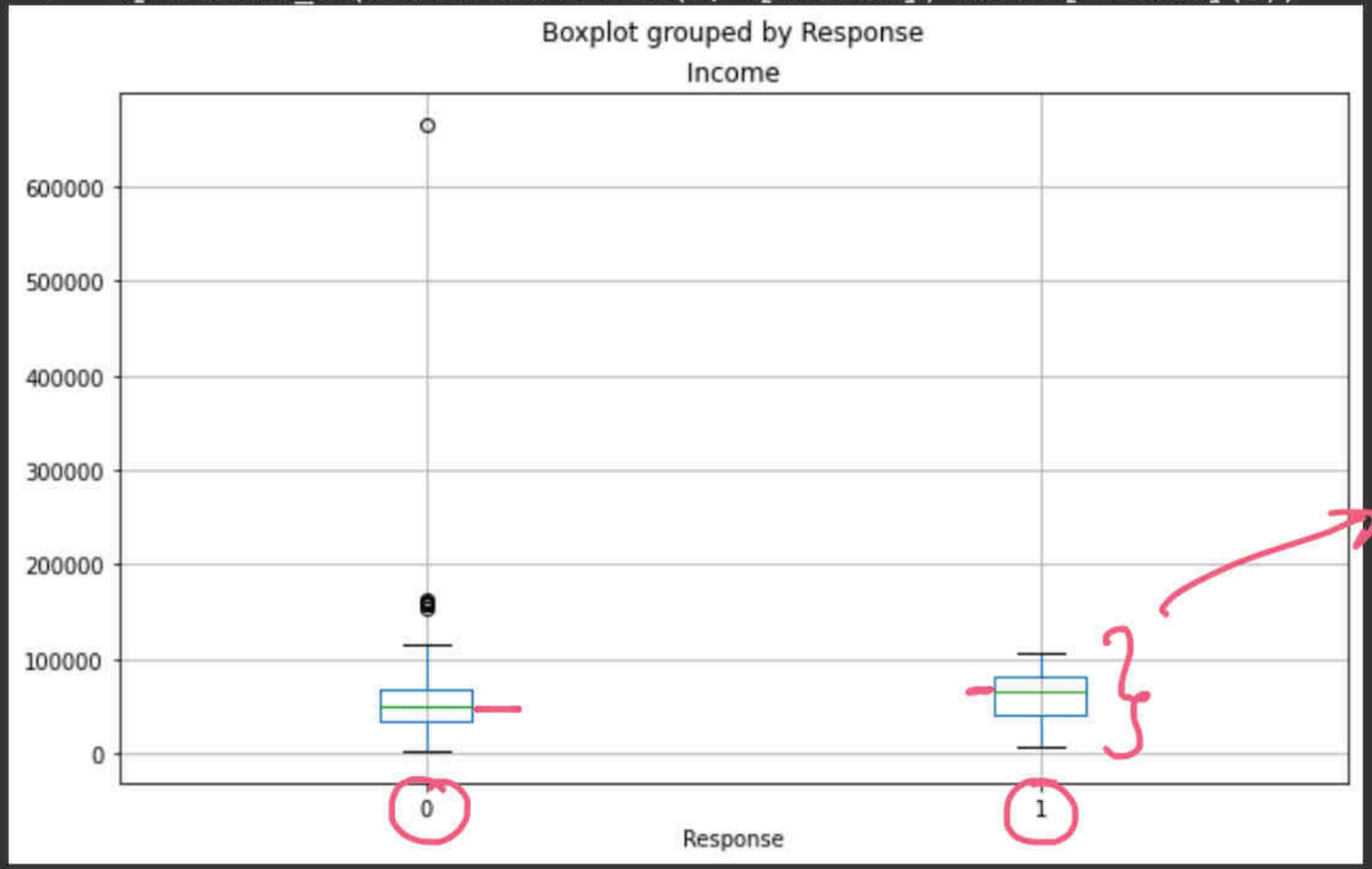
colab.research.google.com/drive/1l5T7TVIAASw9Tdl4JxqJxuDqRgFBGOW3#scrollTo=Egg1hlOlqLJ7

Update

+ Code + Text

RAM
Disk

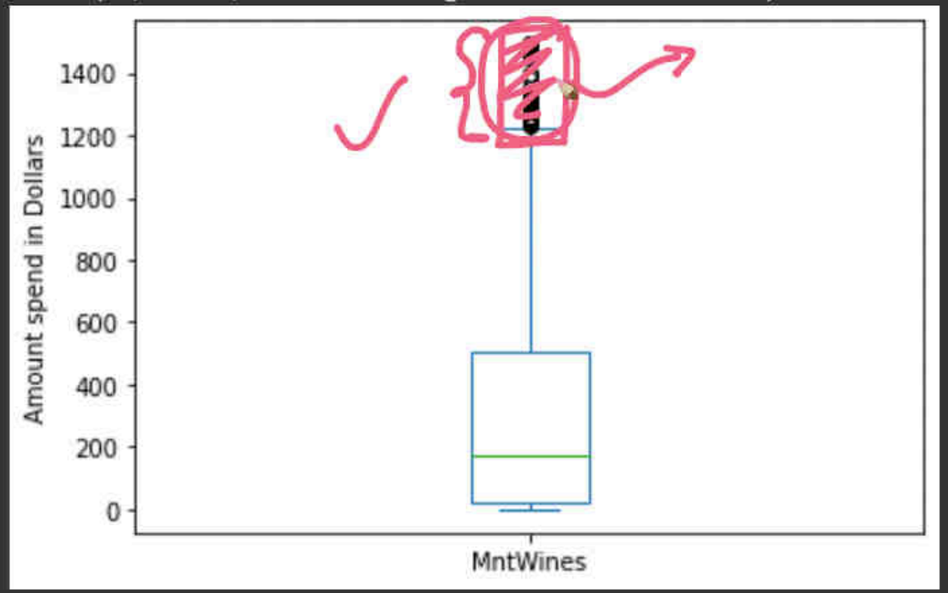
```
[ ] /usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning: Creating an nda
    X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
```



36 / 36

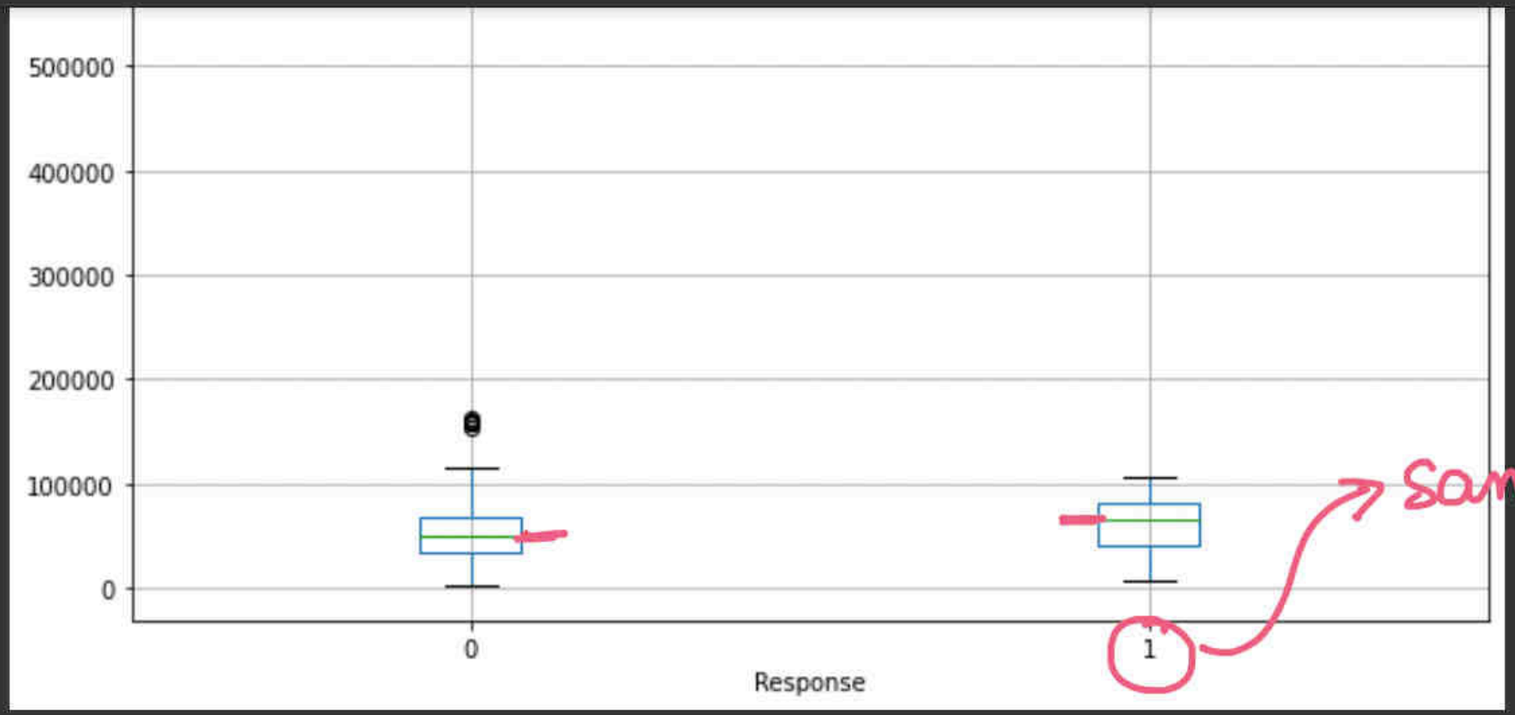

```
[ ]  #box plot
     ax = df['MntWines'].plot.box()
     ax.set_ylabel("Amount spend in Dollars")
```

```
Text(0, 0.5, 'Amount spend in Dollars')
```



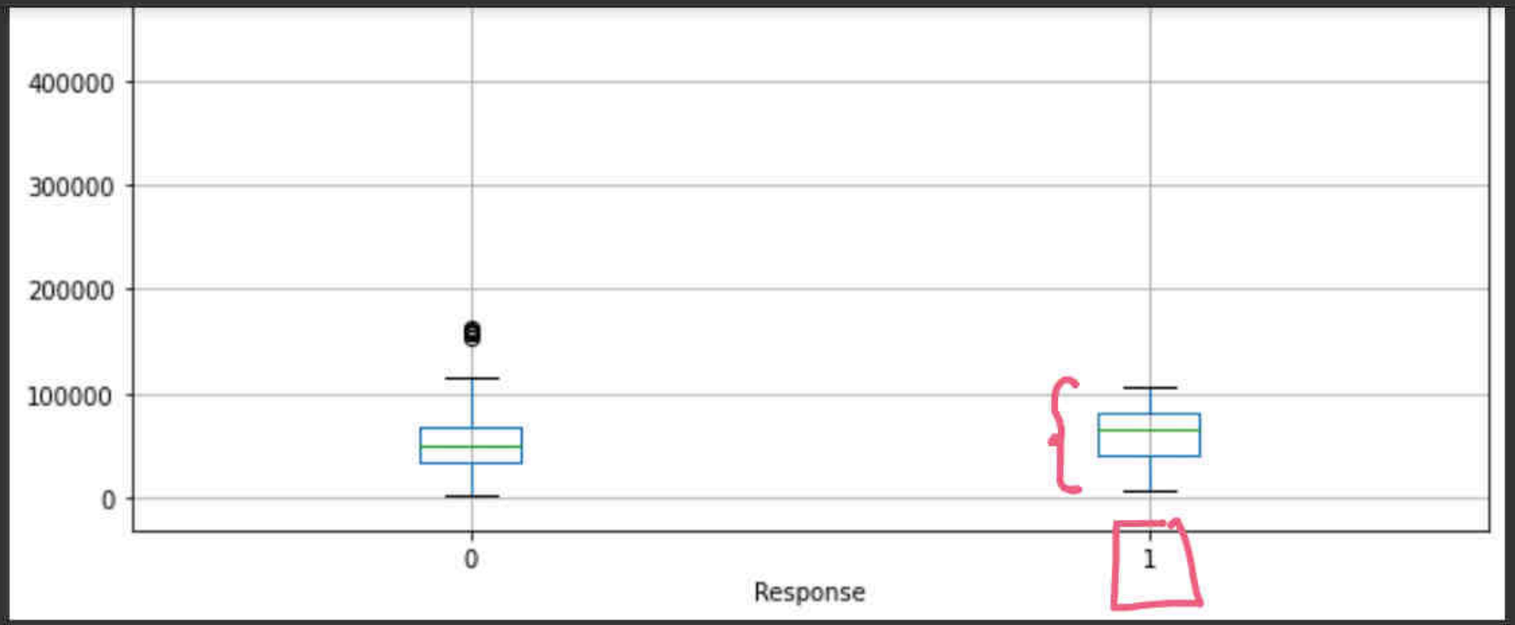

```
[ ]  #bar plot
     df['Education'].value_counts().plot.bar()
```

[ ]



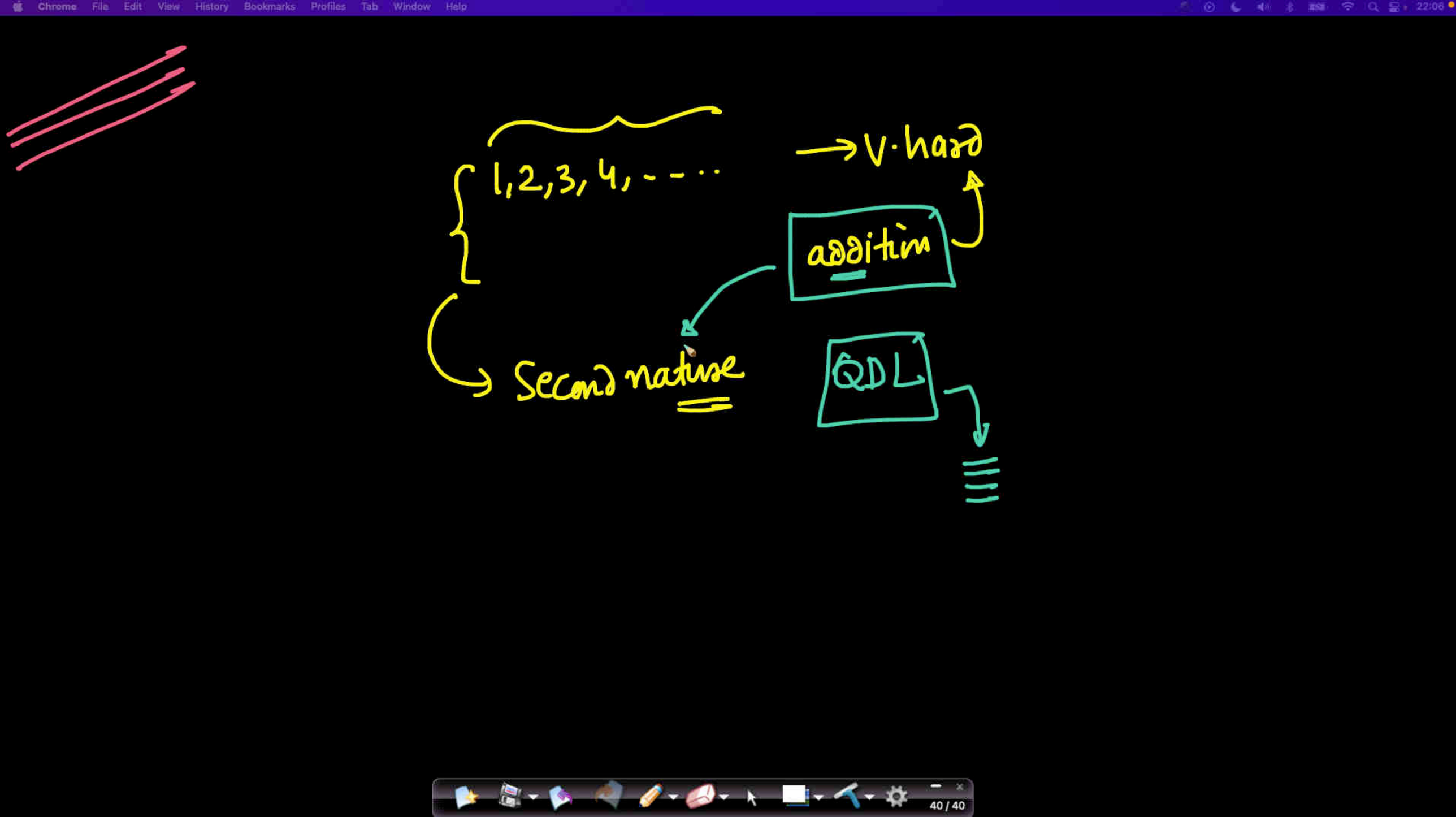
[ ]

[ ]



200 M
↓
2240
R
1
0

[ ]


1, 2, 3, 4, - - - ..
→ V·hard
addition
Second nature
QDL

which plot to use where

→ Continous r.v: work-exp

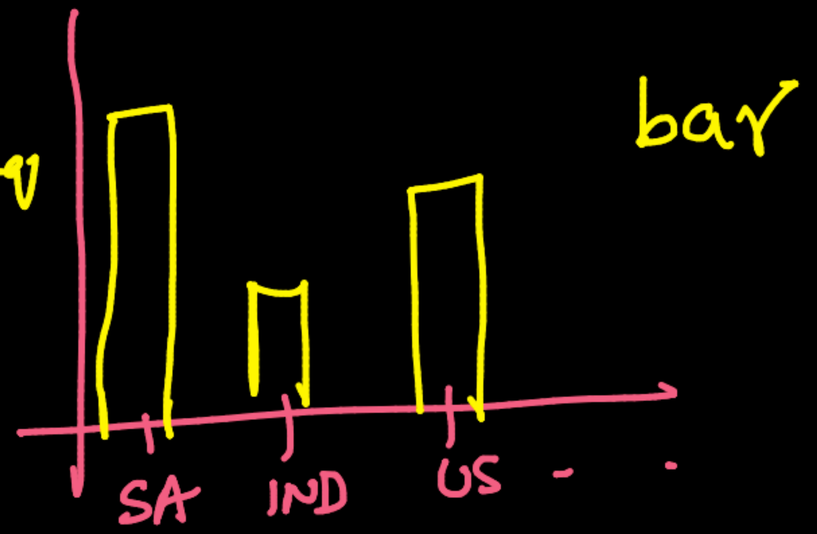
→ histogram: KDE: density plots

bin-width

→ bar plot: 0-2 2-4 4-6 6-8 8+

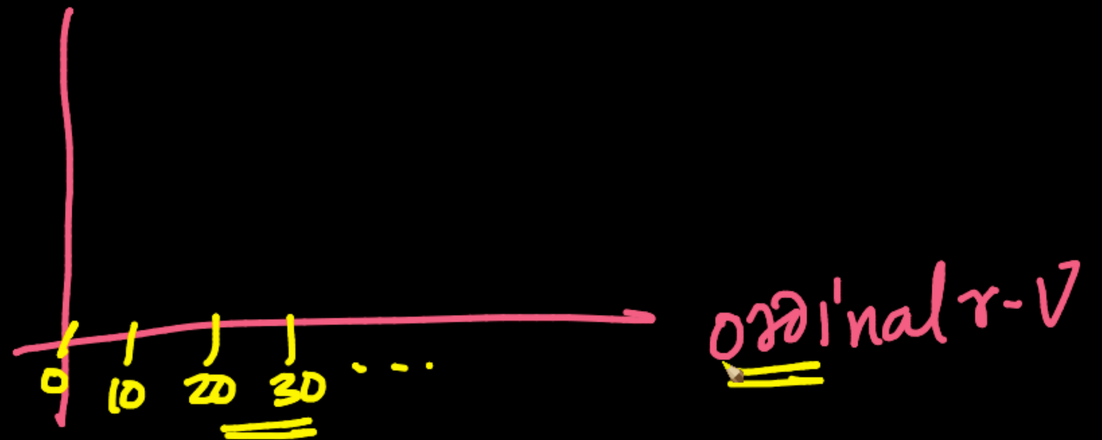
binned

Categorical
Country
(not-ordinal)



histogram

Income

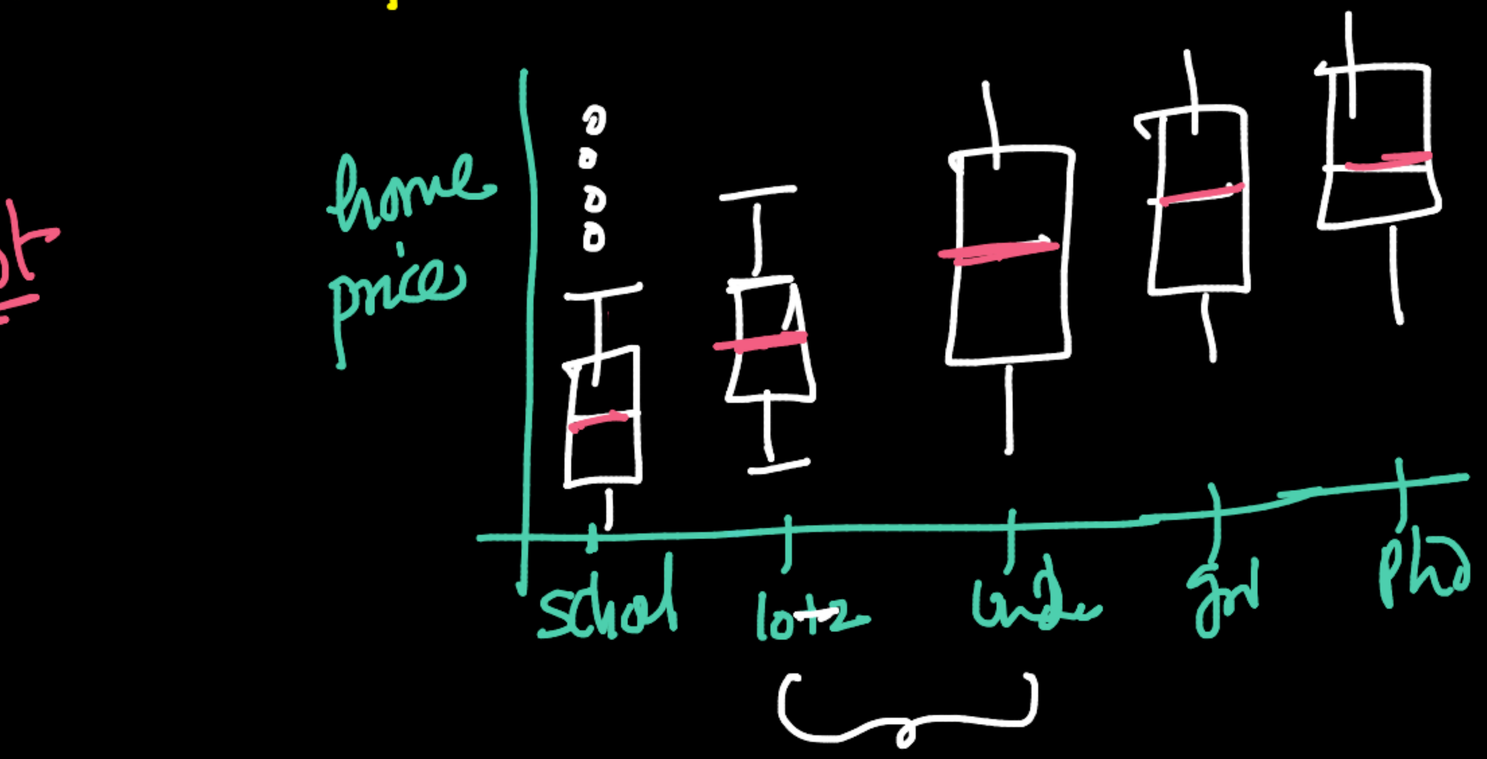


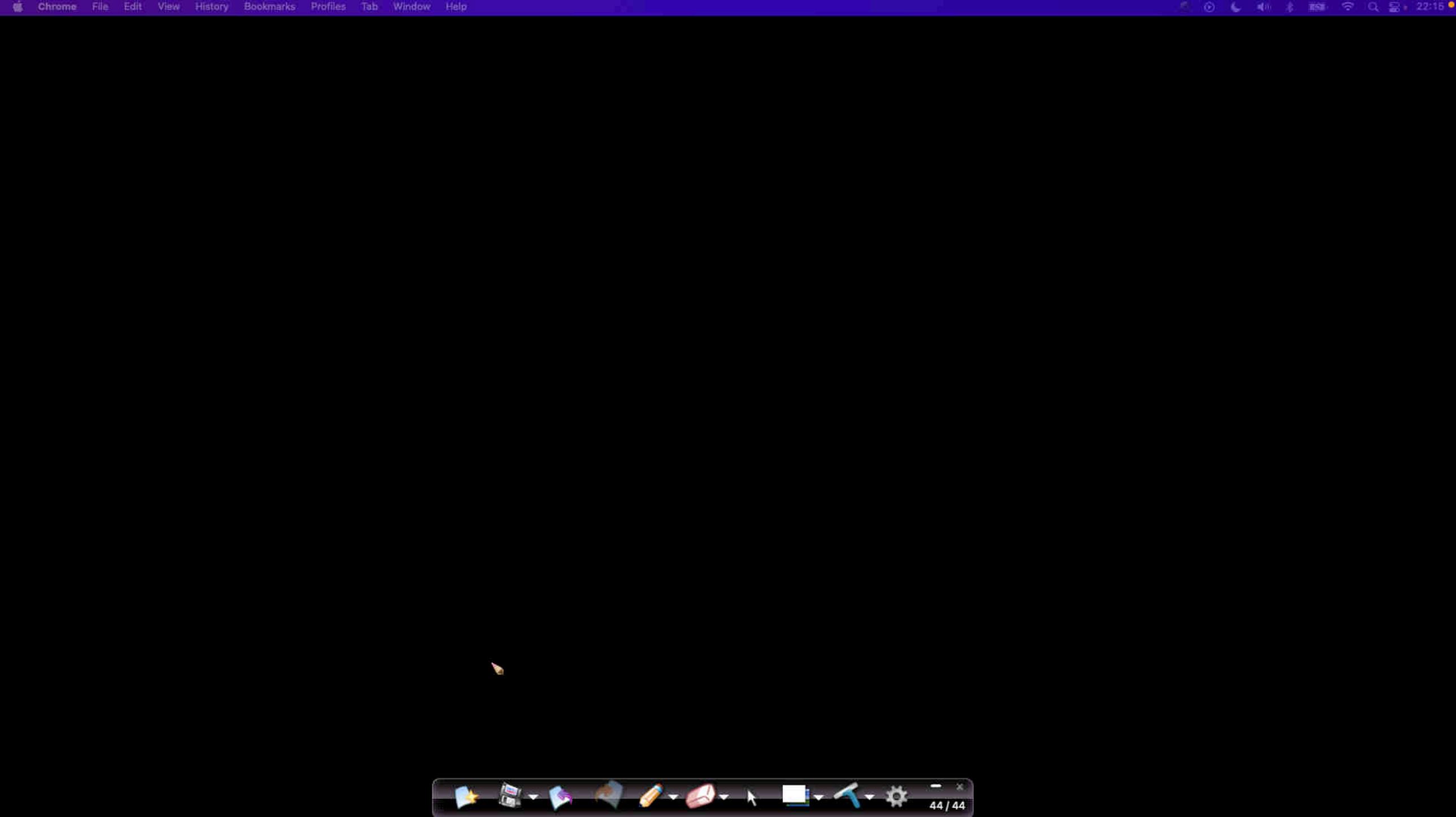
→ home-prices : outliers → boxplot

→ home-price & education-level (ordinal) relationship

scatter

bar plot

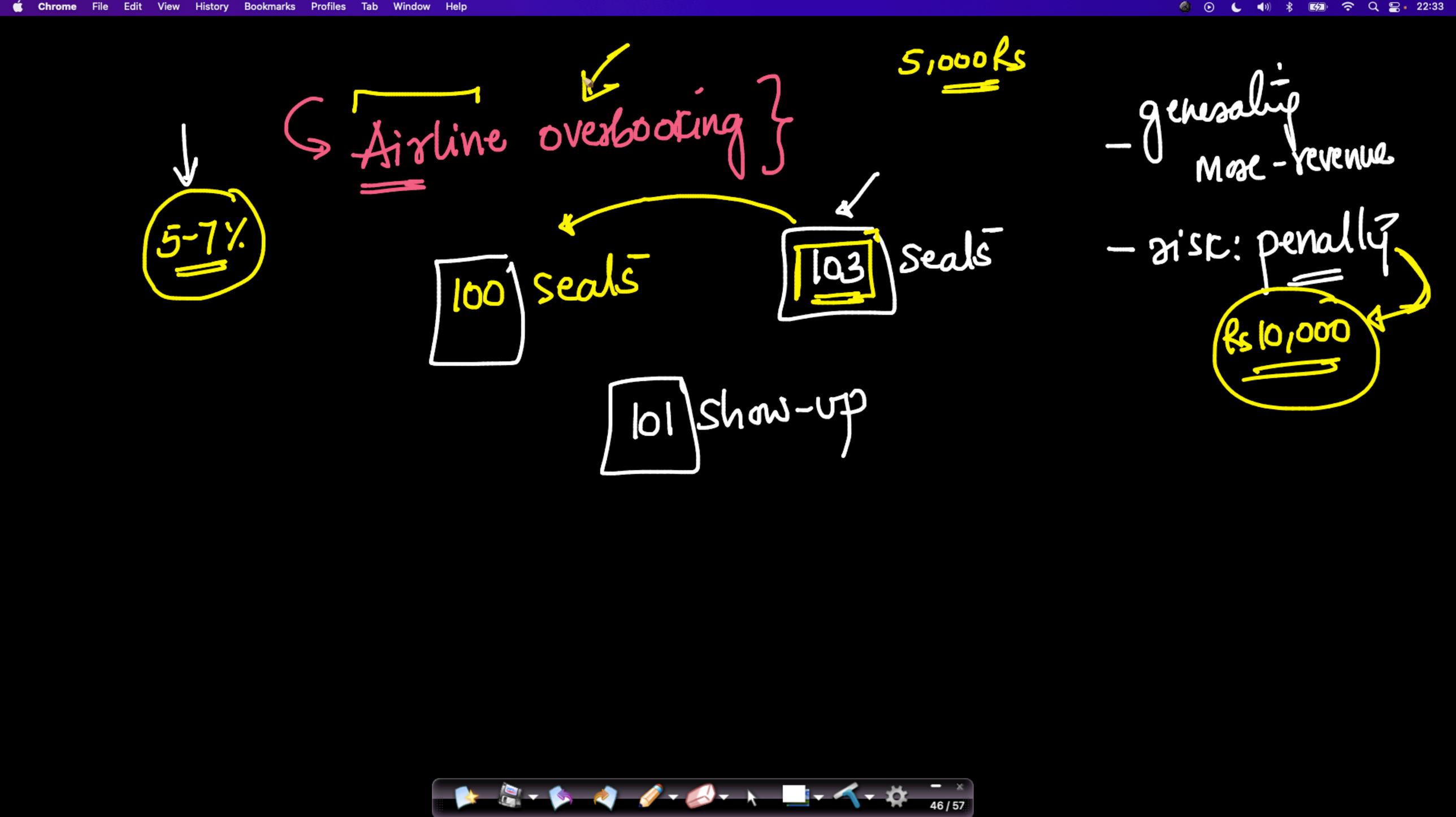



Chrome File Edit View History Bookmarks Profiles Tab Window Help
22:15
44 / 44

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Probability - basics + 2½

- what ; where
- why
- PMF ; PDF ; CDF ..
- real-world ex
- intuition


Airline overbooking
5,000 Rs
5-7%
100 seats
103 seats
101 show-up
- generating more revenue
- risk: penalty
Rs 10,000

Q

Airline
- data [showing-up]

100-seats

- max net rev by overbooking

↳ revenue - penalty

↓ more tickets

↓ >100 show-up

prob model

```
[3] flight.csv          100%[===================>]  11.89K  --.-KB/s    in 0s

2022-04-26 16:58:15 (57.7 MB/s) - 'flight.csv' saved [12175/12175]
```

```
flights = pd.read_csv('./flight.csv')
flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Passenger_ID  1000 non-null   int64
 1   Flight_ID     1000 non-null   object
 2   Arrived       1000 non-null   int64
dtypes: int64(2), object(1)
memory usage: 23.6+ KB
```

```
flights.head()
```

	Passenger_ID	Flight_ID	Arrived
0	1811	A320	1

```
[1]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt

     import seaborn as sns
     from scipy import stats
```

```
     id = "1PazlhissU63pozkOjJckyjIJuM_u-0JF"
     print("https://drive.google.com/uc?export=download&id=" + id)
```

```
https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF
```

```
[3]  !wget "https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF" -O flight.csv
```

```
--2022-04-26 16:58:14--  https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF
Resolving drive.google.com (drive.google.com)... 74.125.26.101, 74.125.26.138, 74.125.26.100, ...
Connecting to drive.google.com (drive.google.com)|74.125.26.101|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://doc-14-14-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc717deffksulhg5h7mbp1/4n3oik3dp7qkgo
Warning: wildcards not supported in HTTP.
--2022-04-26 16:58:15--  https://doc-14-14-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc717deffksulhg5h7mbp1
Resolving doc-14-14-docs.googleusercontent.com (doc-14-14-docs.googleusercontent.com)... 173.194.218.132, 2607:f8b0:
Connecting to doc-14-14-docs.googleusercontent.com (doc-14-14-docs.googleusercontent.com)|173.194.218.132|:443... co
HTTP request sent, awaiting response... 200 OK
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Probability2.ipynb - Colaborat | ProbabilityDisb_1.ipynb - Col | New Tab | pandas.DataFrame.boxplot —

colab.research.google.com/drive/1JZ-TbmRIY8mmD018C_grjyQHluBGNQxZ#scrollTo=46VInWzHWhRZ

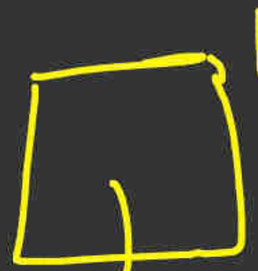
+ Code + Text

RAM Disk

https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF

[3] !wget "https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF" -O flight.csv

```
--2022-04-26 16:58:14--  https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF
Resolving drive.google.com (drive.google.com)... 74.125.26.101, 74.125.26.138, 74.125.26.100, ...
Connecting to drive.google.com (drive.google.com)|74.125.26.101|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://doc-14-14-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc717deffksulhg5h7mbp1/4n3oik3dp7qkgo
Warning: wildcards not supported in HTTP.
--2022-04-26 16:58:15--  https://doc-14-14-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc717deffksulhg5h7mbp1
Resolving doc-14-14-docs.googleusercontent.com (doc-14-14-docs.googleusercontent.com)... 173.194.218.132, 2607:f8b0:
Connecting to doc-14-14-docs.googleusercontent.com (doc-14-14-docs.googleusercontent.com)|173.194.218.132|:443... co
HTTP request sent, awaiting response... 200 OK
Length: 12175 (12K) [text/csv]
Saving to: 'flight.csv'

flight.csv          100%[===================>]  11.89K  --.-KB/s    in 0s

2022-04-26 16:58:15 (57.7 MB/s) - 'flight.csv' saved [12175/12175]
```



Terminal/Shell/cmd/mt

[4]
```
flights = pd.read_csv('./flight.csv')
flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

50 / 50

https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF

[3] !wget "https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF" -O flight.csv

```
--2022-04-26 16:58:14--  https://drive.google.com/uc?export=download&id=1PazlhissU63pozkOjJckyjIJuM_u-0JF
Resolving drive.google.com (drive.google.com)... 74.125.26.101, 74.125.26.138, 74.125.26.100, ...
Connecting to drive.google.com (drive.google.com)|74.125.26.101|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://doc-14-14-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc717deffksulhg5h7mbp1/4n3oik3dp7qkgo
Warning: wildcards not supported in HTTP.
--2022-04-26 16:58:15--  https://doc-14-14-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc717deffksulhg5h7mbp1
Resolving doc-14-14-docs.googleusercontent.com (doc-14-14-docs.googleusercontent.com)... 173.194.218.132, 2607:f8b0:
Connecting to doc-14-14-docs.googleusercontent.com (doc-14-14-docs.googleusercontent.com)|173.194.218.132|:443... co
HTTP request sent, awaiting response... 200 OK
Length: 12175 (12K) [text/csv]
Saving to: 'flight.csv'

flight.csv          100%[===================>]  11.89K  --.-KB/s    in 0s

2022-04-26 16:58:15 (57.7 MB/s) - 'flight.csv' saved [12175/12175]
```

[4]
```
flights = pd.read_csv('./flight.csv')
flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
[4] flights = pd.read_csv('./flight.csv')
    flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Passenger_ID  1000 non-null   int64
 1   Flight_ID     1000 non-null   object
 2   Arrived       1000 non-null   int64
dtypes: int64(2), object(1)
memory usage: 23.6+ KB
```

```
flights.head()
```

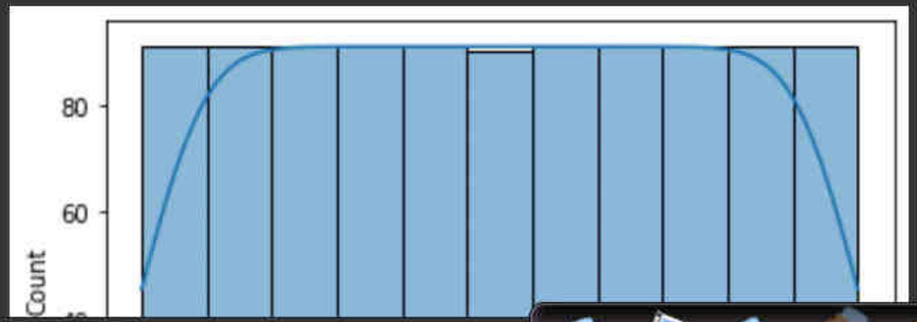
	Passenger_ID	Flight_ID	Arrived
0	1811	A320	1
1	1812	A320	1
2	1813	B777	1
3	1814	B737	1
4	1815	B737	1

```
memory usage: 23.6+ KB
```

```
[ ] flights.head()
```

	Passenger_ID	Flight_ID	Arrived
0	1811	A320	1
1	1812	A320	1
2	1813	B777	1
3	1814	B737	1
4	1815	B737	1

```
[ ] sns.histplot(flights["Passenger_ID"], kde=True)
    plt.show()
```



```
[4] flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Passenger_ID  1000 non-null   int64
 1   Flight_ID     1000 non-null   object
 2   Arrived       1000 non-null   int64
dtypes: int64(2), object(1)
memory usage: 23.6+ KB
```

```
[ ] flights.head()
```

	Passenger_ID	Flight_ID	Arrived
0	1811	A320	1
1	1812	A320	1
2	1813	B777	1
3	1814	B737	1
4	1815	B737	1

```
[4] flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Passenger_ID  1000 non-null   int64
 1   Flight_ID     1000 non-null   object
 2   Arrived       1000 non-null   int64
dtypes: int64(2), object(1)
memory usage: 23.6+ KB
```

```
[ ] flights.head()
```

	Passenger_ID	Flight_ID	Arrived
0	1811	A320	1
1	1812	A320	1
2	1813	B777	1
3	1814	B737	1
4	1815	B737	1

```
[4] flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Passenger_ID  1000 non-null   int64
 1   Flight_ID     1000 non-null   object
 2   Arrived       1000 non-null   int64
dtypes: int64(2), object(1)
memory usage: 23.6+ KB
```

```
[ ] flights.head()
```

	Passenger_ID	Flight_ID	Arrived
0	1811	A320	1
1	1812	A320	1
2	1813	B777	1
3	1814	B737	1
4	1815	B737	1

```
[4]  Data columns (total 3 columns):
      #   Column        Non-Null Count  Dtype
     ---  ------        --------------  -----
      0   Passenger_ID  1000 non-null   int64
      1   Flight_ID     1000 non-null   object
      2   Arrived       1000 non-null   int64
     dtypes: int64(2), object(1)
     memory usage: 23.6+ KB
```

random-var

A = 1 or 0

↳ categorical / binary

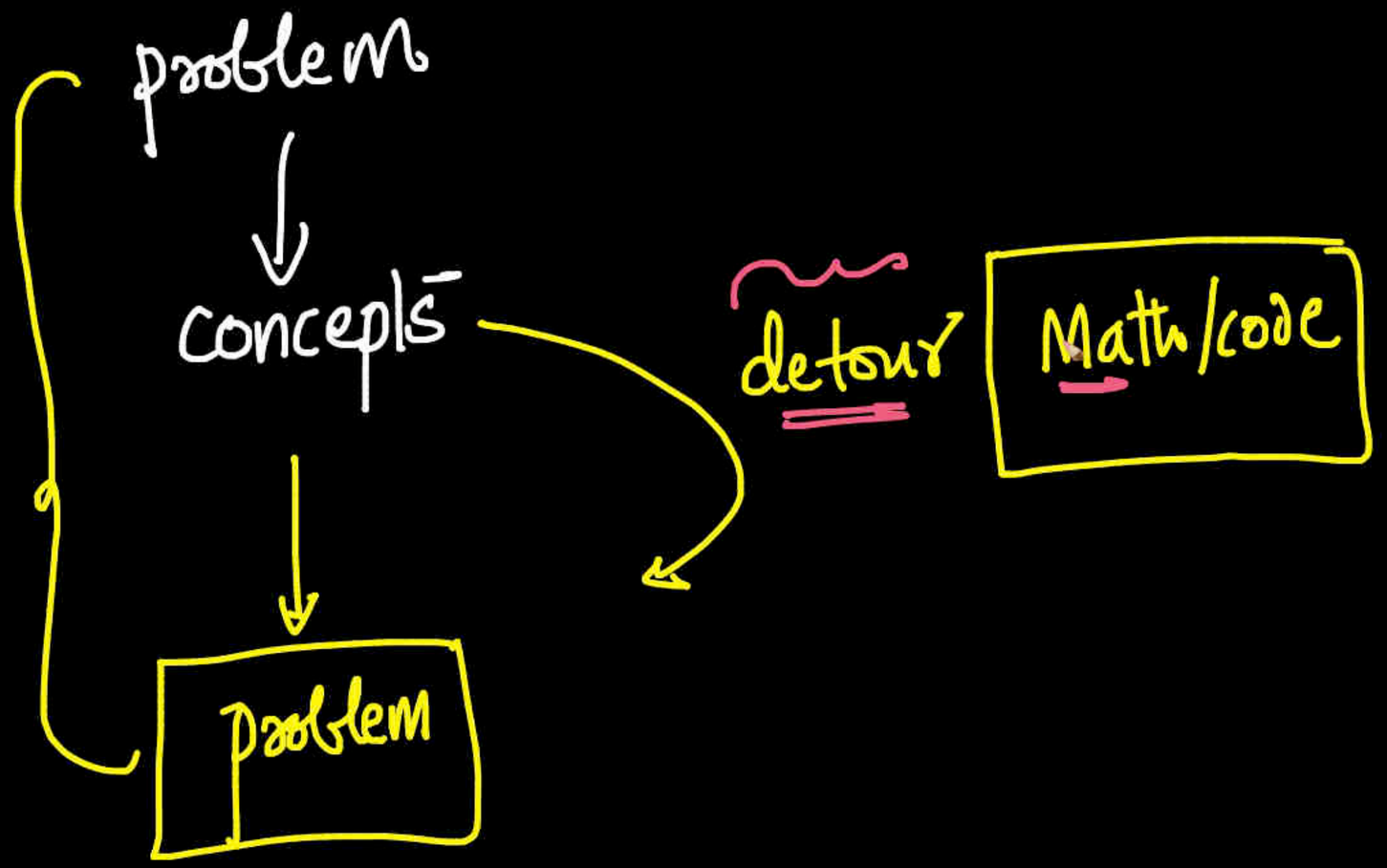
```
[ ]  flights.head()
```

	Passenger_ID	Flight_ID	Arrived
0	1811	A320	1
1	1812	A320	1
2	1813	B777	1
3	1814	B737	1
4	1815	B737	1

pID: r.v numeric → continous

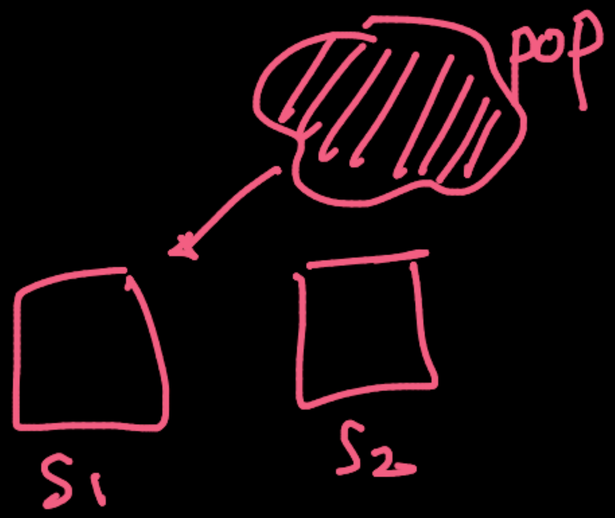
1000 – 9999

```
sns.histplot(flights["Passenger_ID"], kde=True)
plt.show()
```



$n = 1000$

Classical Stat: $\gtrsim 30$ randomly ✓



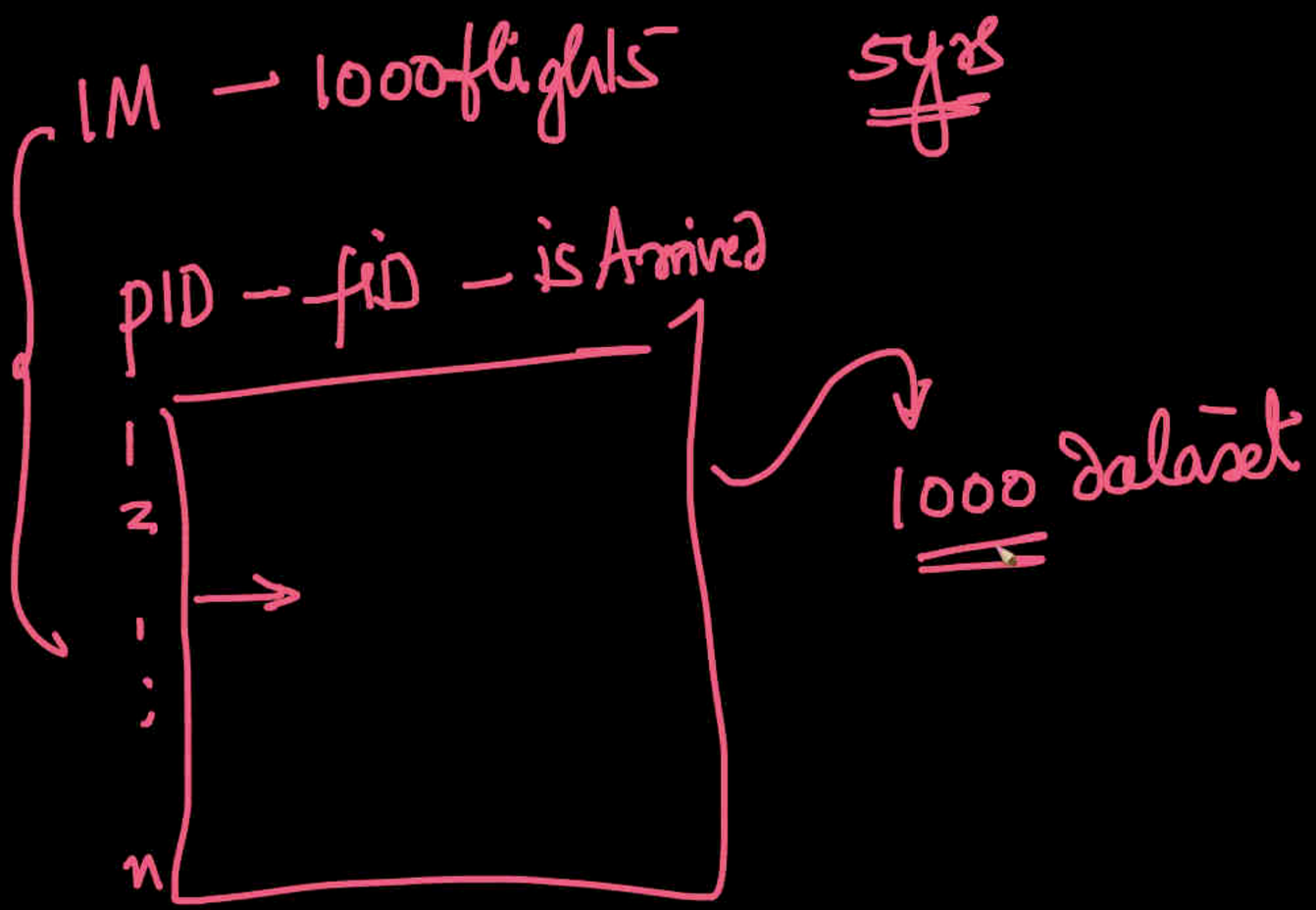
~ 100 ✓

ML: few thousands — few millions

airline

1M – 1000 flights

sys



Prob - dist: f()

outcomes

1

discrete r.v

r.v

A: arrived

0

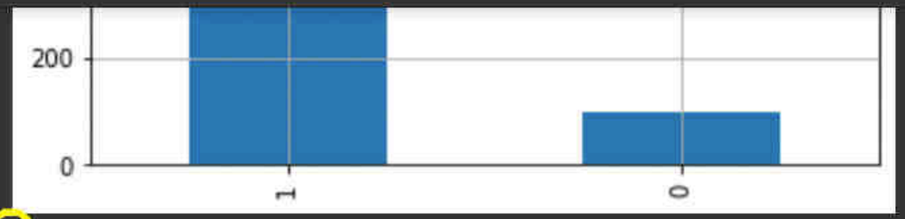
Measure Thy

1000

$p(A=0) = 0.098$

$p(A=1) = 0.902$ ✓

[5]



0.902 → 902
0.098 → 98

```
showsup_probability = flights['Arrived'].value_counts(normalize=True)[1]
print(showsup_probability)
```

```
0.902
```

```
[ ] flights['Arrived'].value_counts()
```

```
1    902
0     98
Name: Arrived, dtype: int64
```

$\frac{902}{1000} = P(A=1) = 0.902$

```
[ ] import math

    PENALTY = 10000
    def comb(n, r):

      num1 = math.factorial(n)
      num2 = math.factorial(r)
      num3 = math.factorial(n-r)
```

$\frac{98}{1000} = P(A=0)$

Q

Fair dice — 6 sides

1, 2, 3, 4, 5

4.5

r.v X: represent the outcome of a throw

{1, 2, 3, 4, 5, 6} ← Categorical

$P(X=1) = \frac{1}{6} = P(X=2) \cdots = P(X=6)$

Prob. dist

discrete r.v

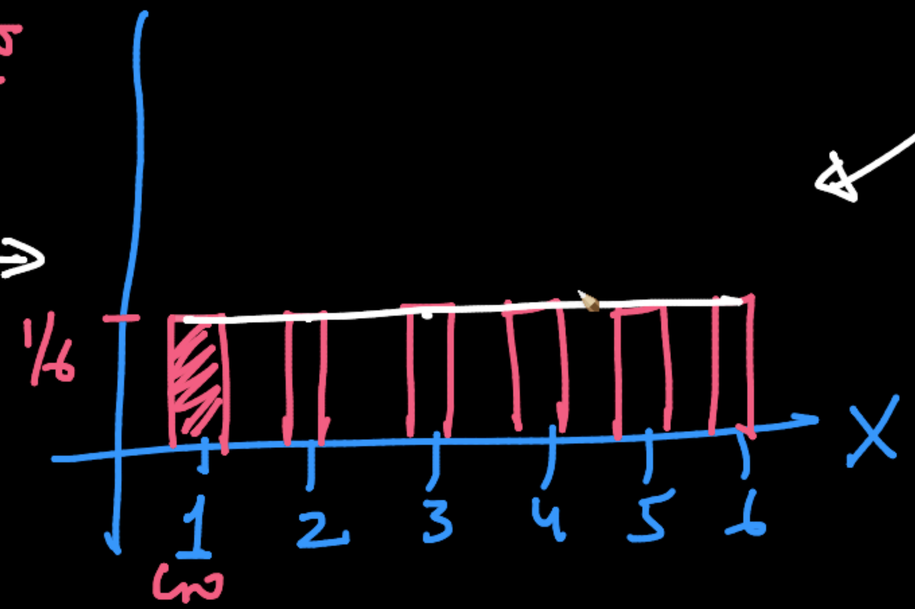
X: dice throw outcomes

prob. mass fn

PMF

{ Uniformly distributed discrete r.v

prob



$P(X=i) = 1/6$

$\forall\, i = 1,2,3,4,5,6$

$\exists$: there exists

Q

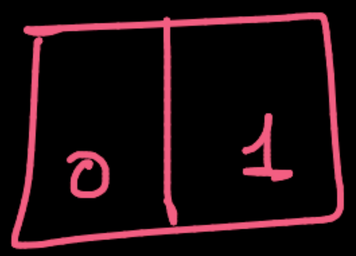
discrete

A = 0 or 1

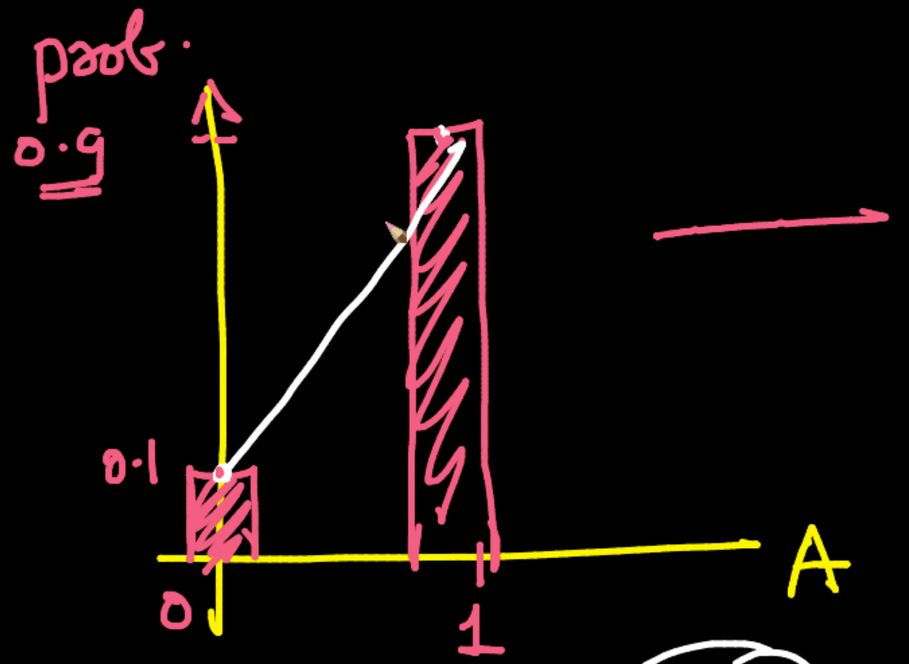
✓ $p(A=0) = 0.1$

✓ $p(A=1) = 0.9$

1.0



u



viz ✓

Extremely

intuition

Terminology → overwhelming
→ contextualize
] — revision

{ 80 mul-prob

PMF : discrete r.v

PDF (later): Continous r.v

Chrome File Edit View History Bookmarks Profiles Tab Window Help

- → notes ✓
- → work-out the concepts in your own [test]

Feynman

freq:
- week → thorough
- month
- 3-months

code

Q

A: 0 or 1

$P(A=0) = 0.1$ $P(A=1) = 0.9$

Uniform r.v ? → NO

Q

Fair Coin

X: T → 0, H → 1 or

$P(X=T) = P(X=H) = 1/2$

Uniform r.v

CDF: Cumulative disb. fn

discrete r.v

$P(X \le i) \quad \forall i = 1,2,3,4,5,6$

fair dice $X \in \{1,2,3,4,5,6\}$

$P(X \le 1) = 1/6$

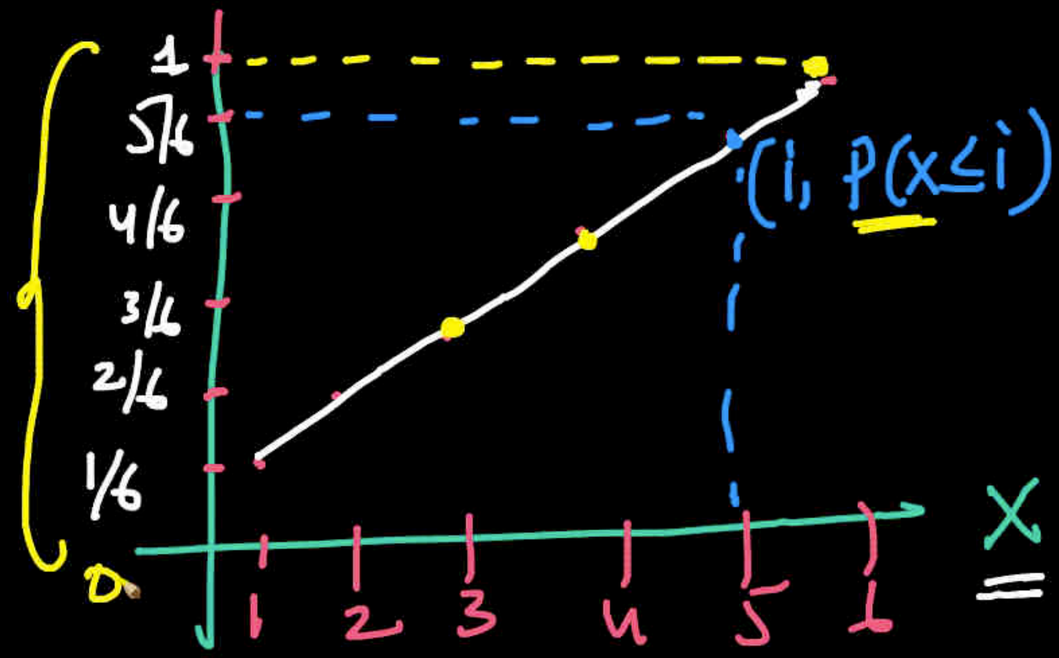
$P(X \le 2) = 2/6 = P(X=1) + P(X=2)$

$\vdots$

$P(X \le 6) = 6/6$

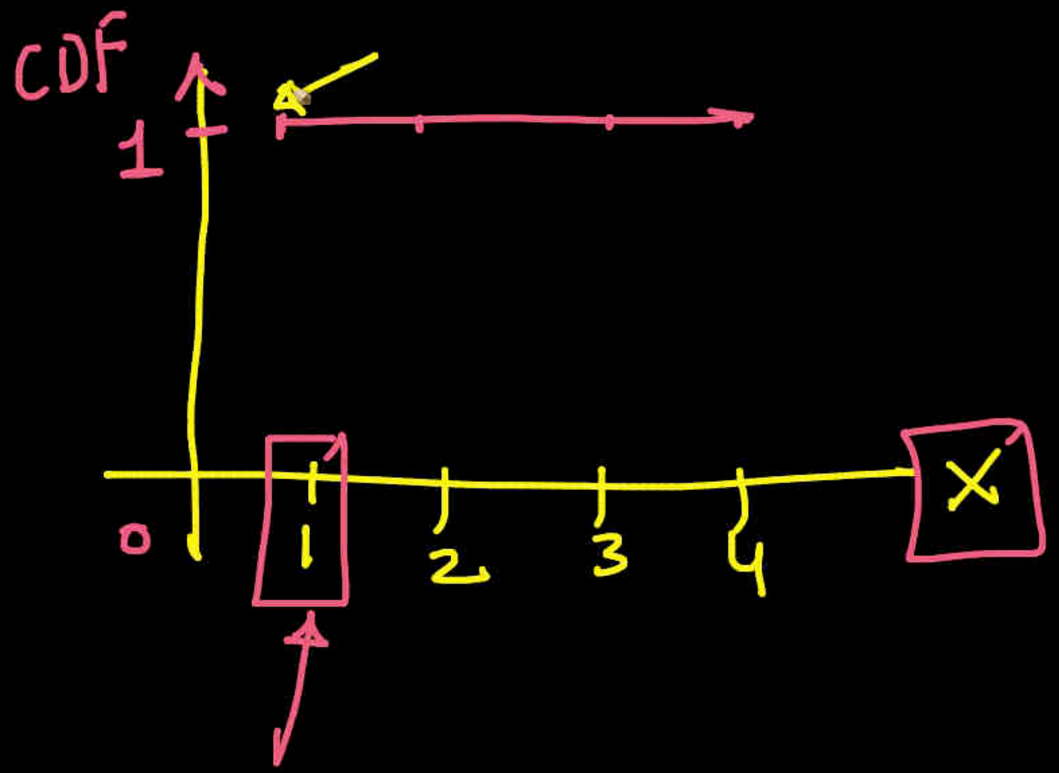
PMF: $P(X=i) \quad \forall i$

CDF



- non-decr graph
- max-val = 1

$$P(X \leq 4) = P(X \leq 3) + \boxed{\quad} \geq 0$$



$P(X \leq 1) = 1$

$P(X \leq 2) = 1$

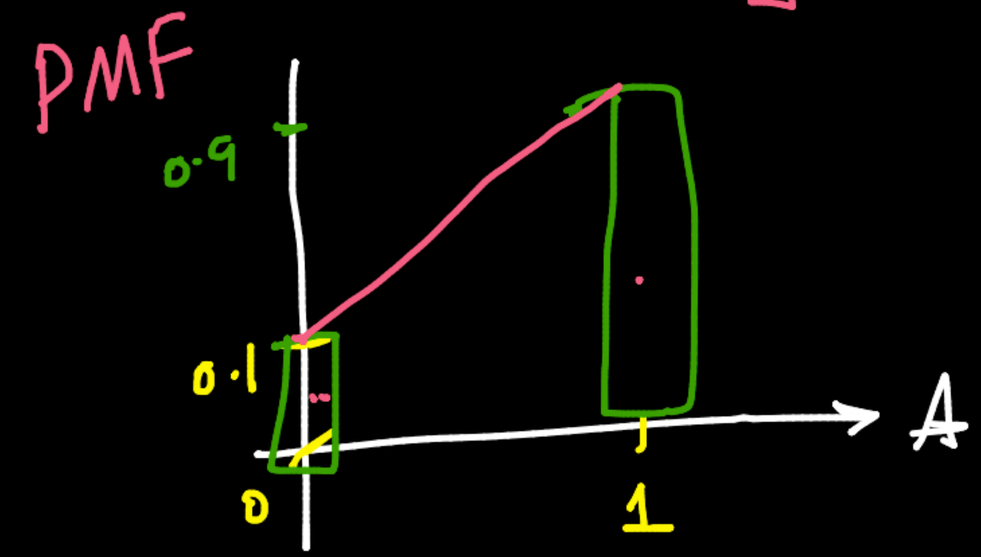
$P(X \leq 3) = 1$

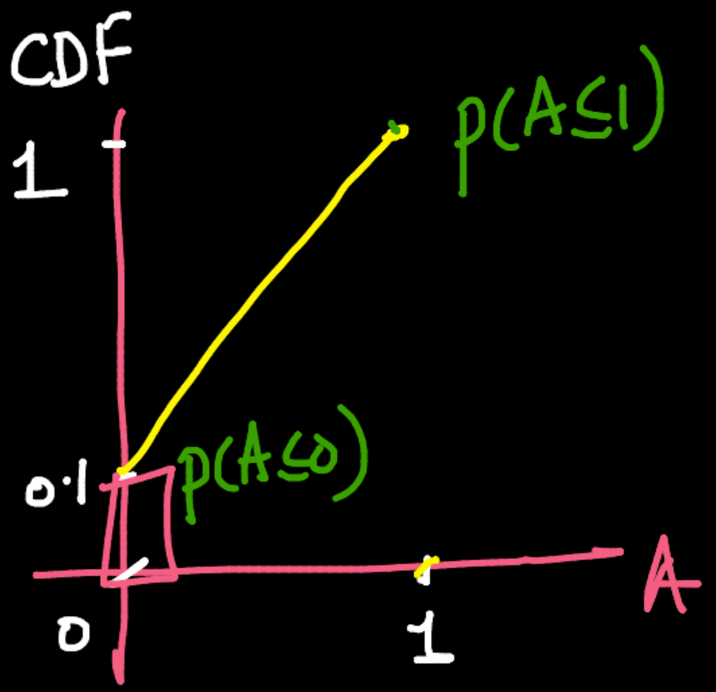
$P(X \leq 4) = 1$

$X \in \{1, 2, 3, 4\}$

discrete r.v

CDF ⟷ PMF





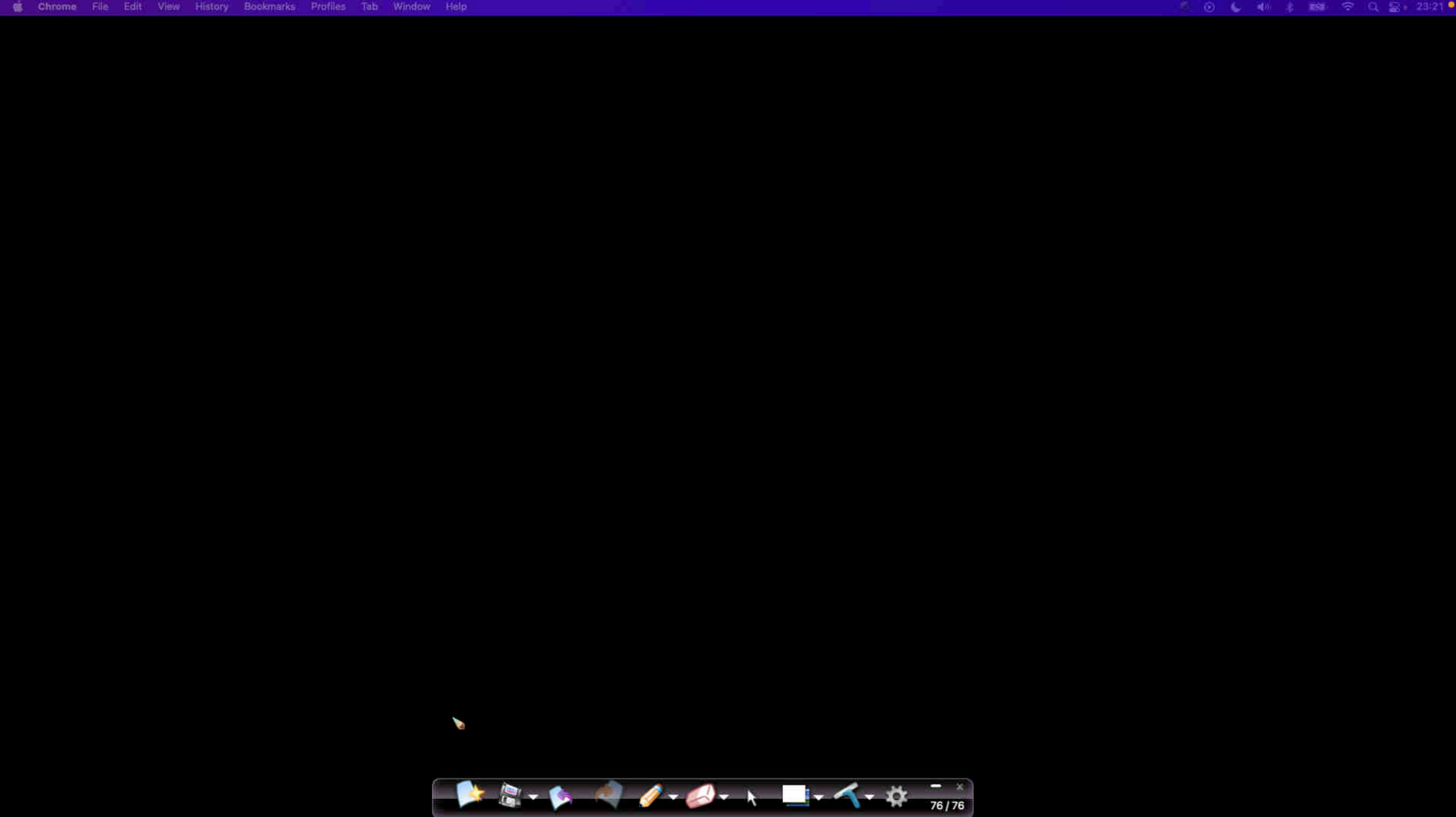
$p(A \leq 0) = 0.1$

$p(A \leq 1) = 1.0$

$A \in \{0, 1\}$

$p(A = 1)$

- CDF ⟶ PMF (difference of consecutive values) — proof-
- PMF → CDF (sum of all previous values)


Chrome File Edit View History Bookmarks Profiles Tab Window Help
23:21
76 / 76

Q

r.v

X discrete

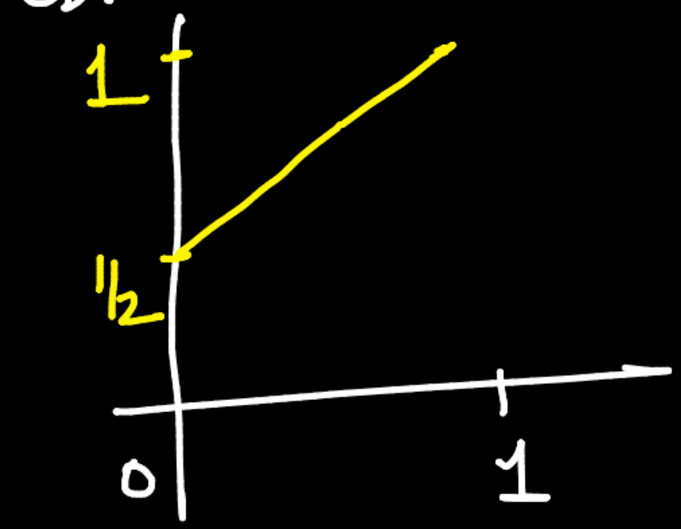
H & T

1 0

create mapping

CDF ✓

CDF

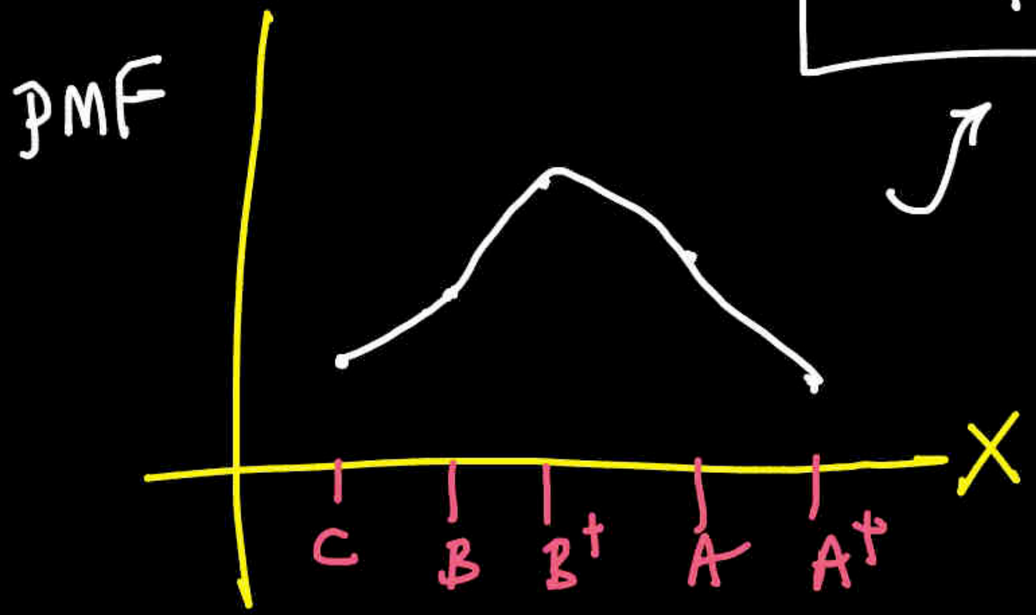


$P(X \leq i)$

PMF $p(X=i)$

$p(X=T) = 1/2 = p(X=H)$

entropy (label) → Decision Trees



A+
A
B+
B
C

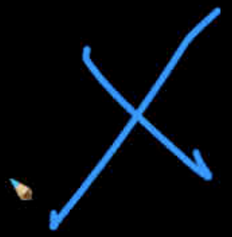
9 - 10:30 - class

10:30 - 10:40 break

10:40 - 12 - class

Q & A

Slow



✓ {
- 9:00 – 10:30
- 10:30 – 10:35
- 10:35 – 11:30
- 11:30

} — stick to this

reuse → speed up

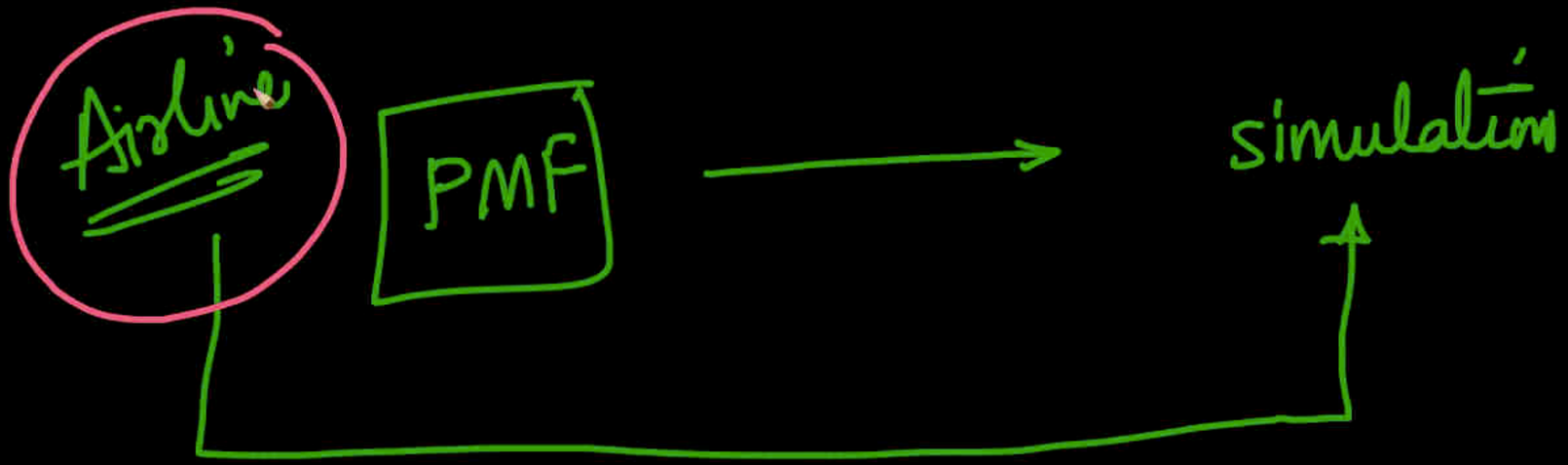
[5 classes later]

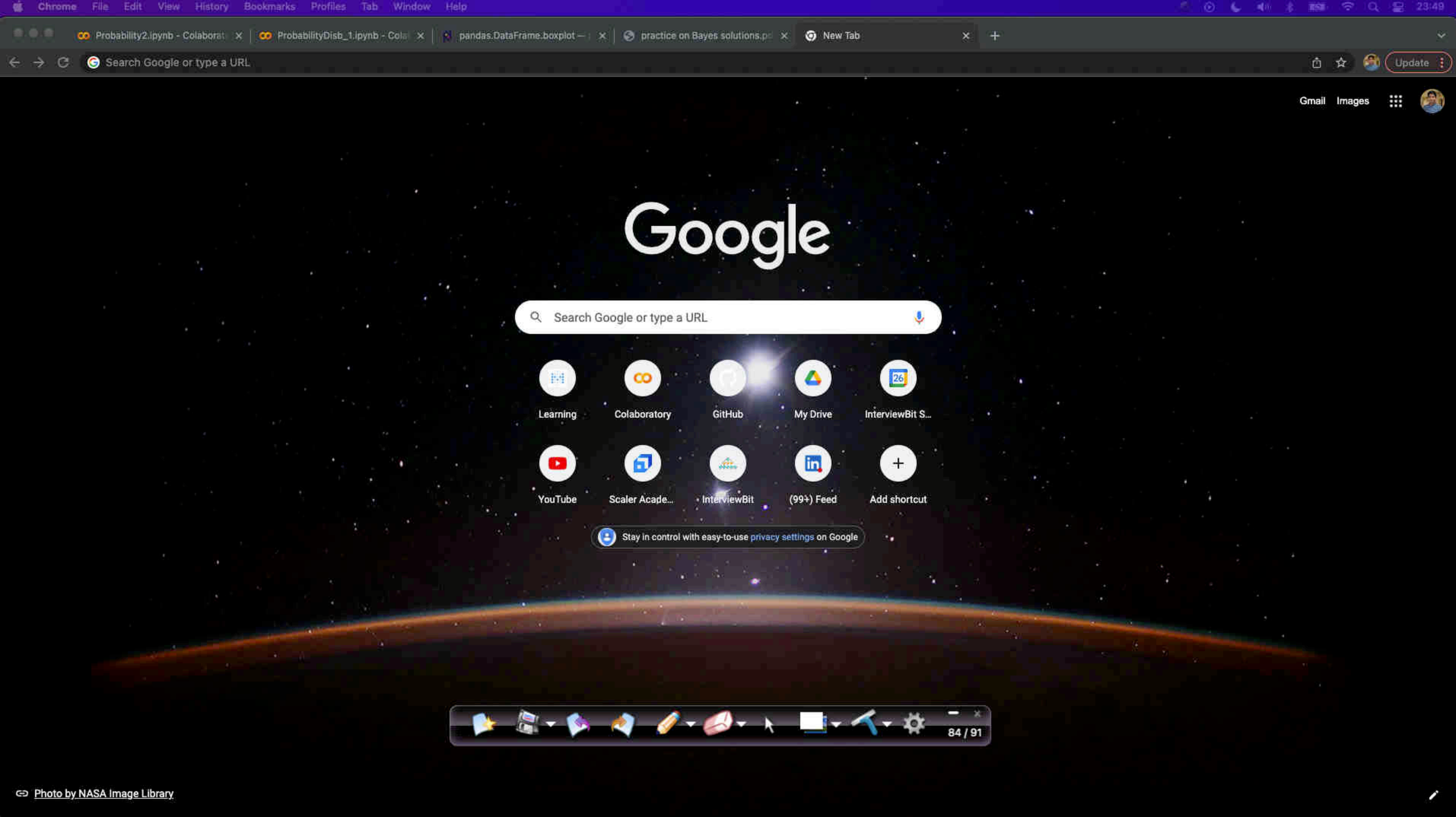
CDF, PDF of Continous r.V

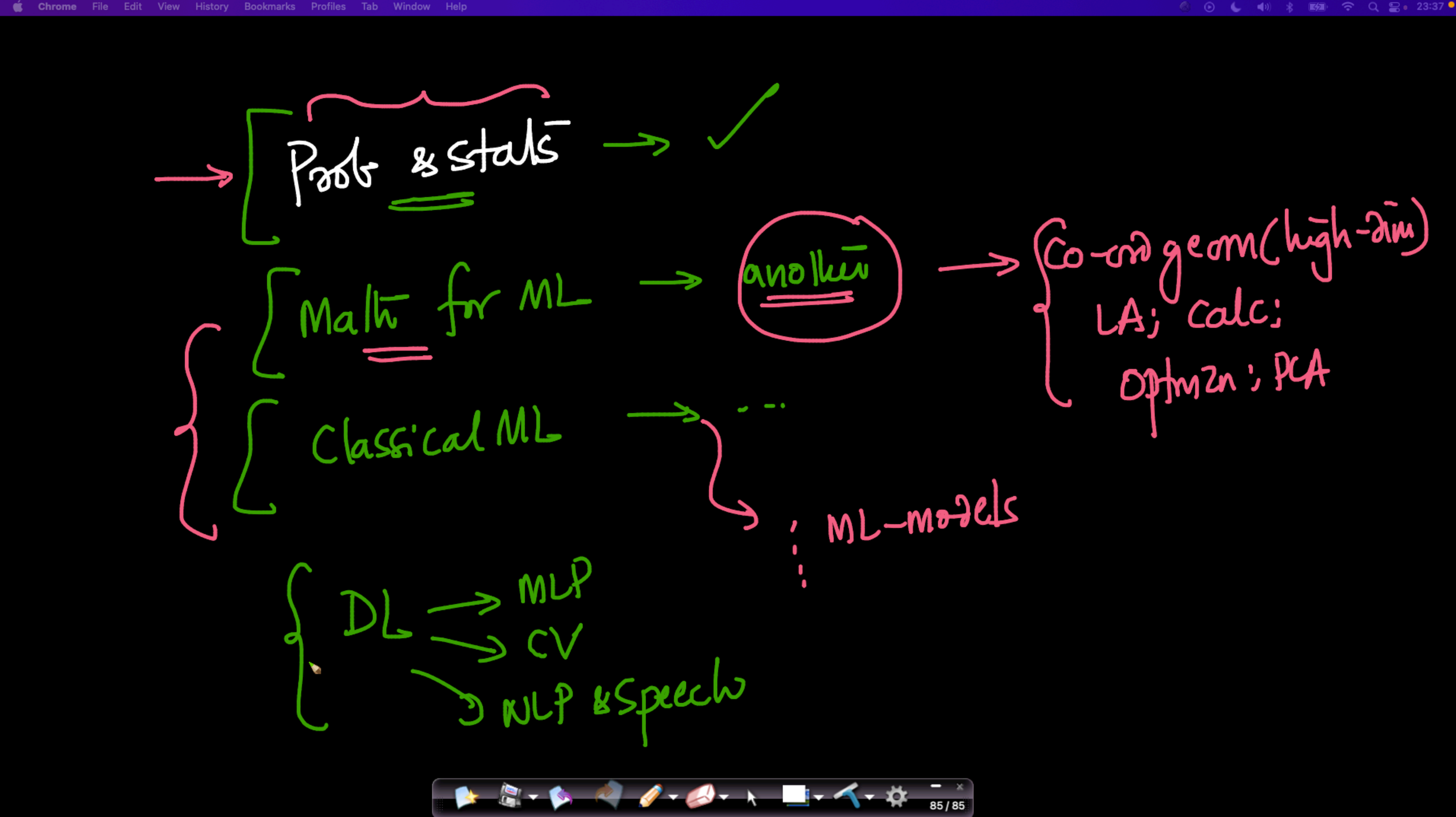
decent - pace

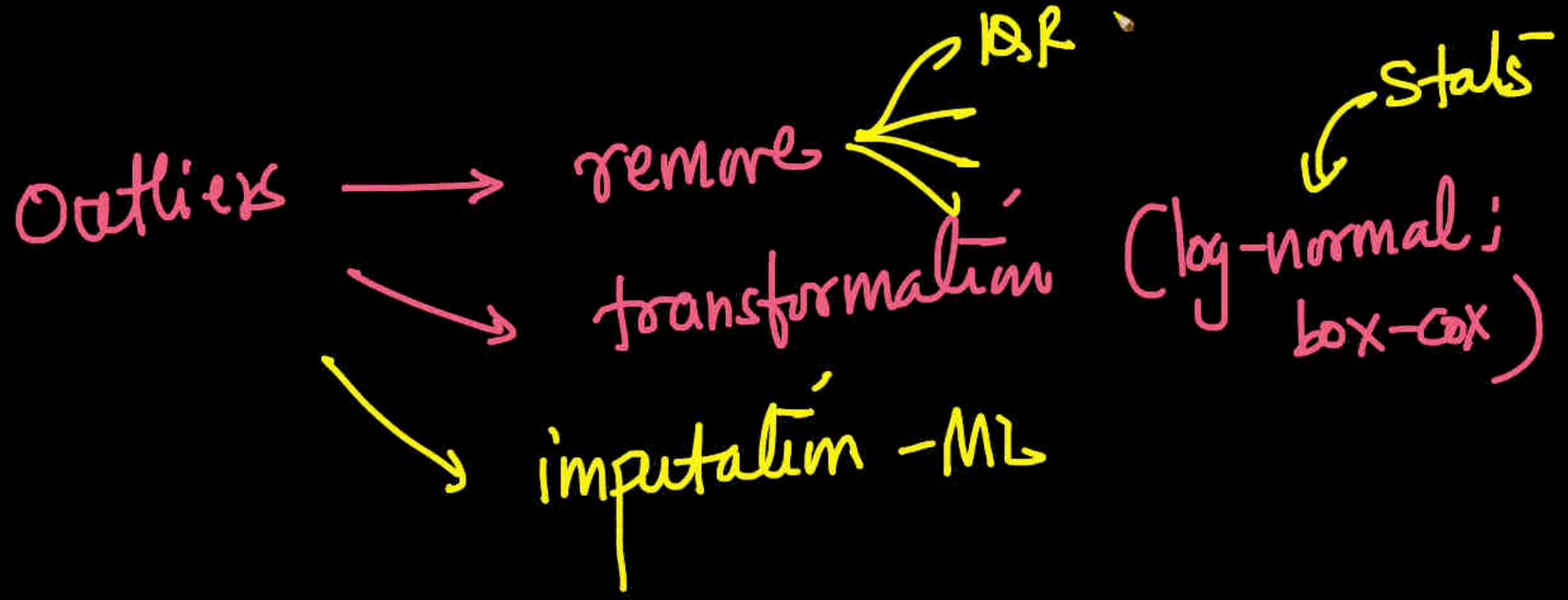
Bernoulli r.V

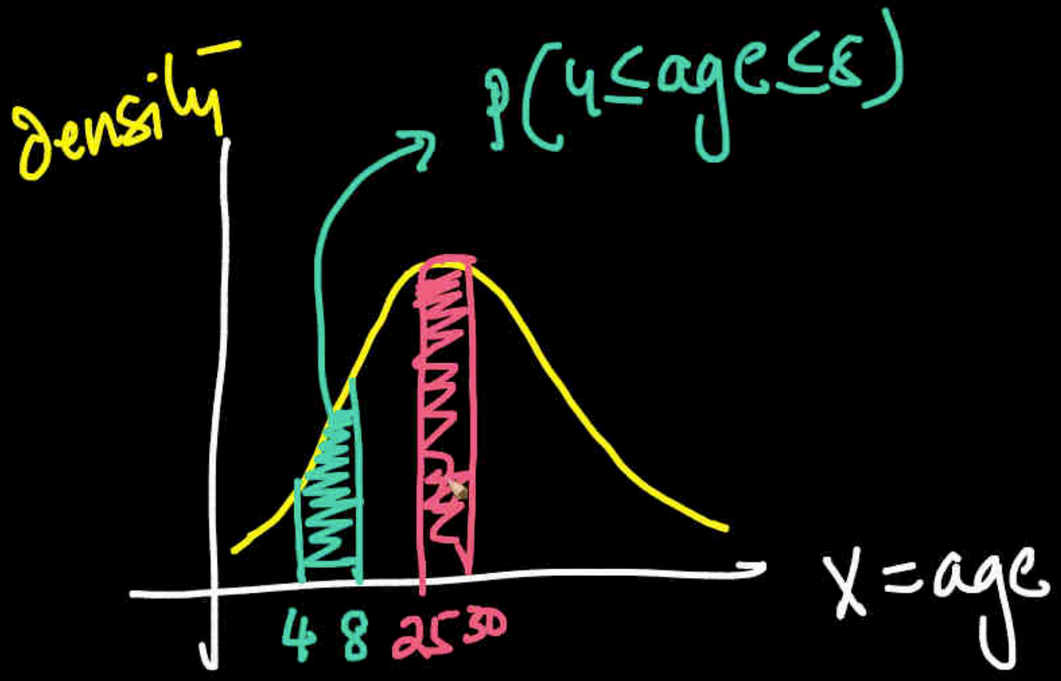
Binomial r.V ⟹ Simulation - code




Chrome File Edit View History Bookmarks Profiles Tab Window Help
Probability2.ipynb - Colaborat
ProbabilityDisb_1.ipynb - Cola
pandas.DataFrame.boxplot —
practice on Bayes solutions.p
New Tab
Search Google or type a URL
Update
Gmail Images
Google
Search Google or type a URL
Learning
Colaboratory
GitHub
My Drive
InterviewBit S...
YouTube
Scaler Acade...
InterviewBit
(99+) Feed
Add shortcut
Stay in control with easy-to-use privacy settings on Google
84 / 91
Photo by NASA Image Library


Probs & stats
Math for ML
analysis
Co-ord geom (high-dim)
LA; calc;
Optmzn; PCA
Classical ML
ML-models
DL
MLP
CV
NLP & Speech
85 / 85


Outliers
remove
IQR
transformation
Stats
(log-normal; box-cox)
imputation - ML

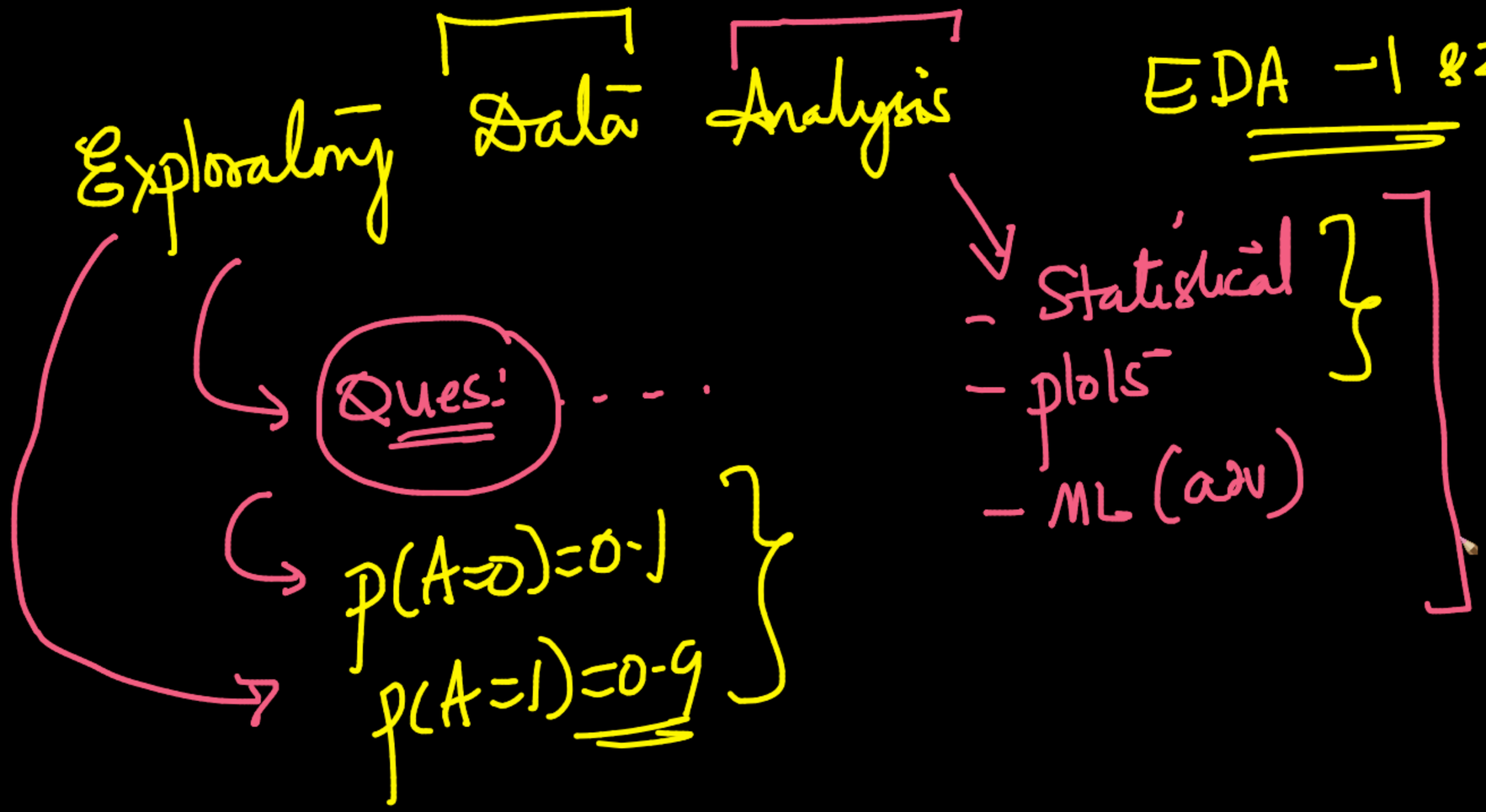


PDF & CDF (next-class)

EDA:

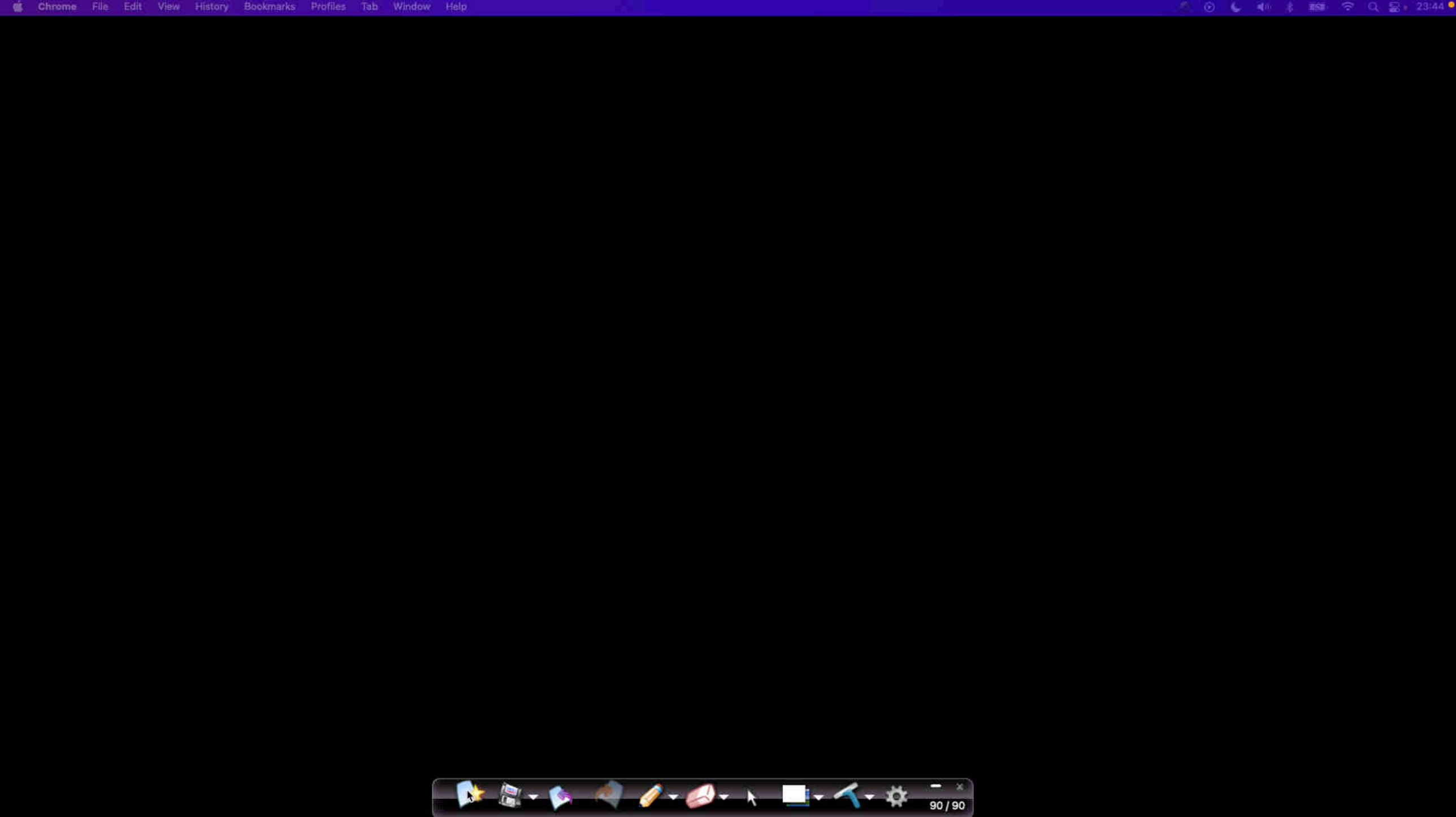
Exploratory Data Analysis

EDA -1 &2



Outlier

- IQR →
- ML
- DL
-


Chrome File Edit View History Bookmarks Profiles Tab Window Help
90 / 90