MAKE SCHOOL

Mean, Median and Mode Variance and Standard Deviation

QL 1.1

Learning Outcomes

By the end of today, you should be able to...

1. Use mean, mode, median, range, and standard deviation to describe data
2. Compare and contrast the above concepts and know when to use which
3. Implement these concepts algorithmically using various libraries

Warm up

5 mins

1. Open the terminal.
2. Make a folder called "**ql-titanic**" and inside boot up an instance of jupyter notebook.
3. Make a new notebook called "**mean-median-mode**".
4. In the notebook, run:

   > **print(“Hi descriptive statistics!”)**

What’s the shape of our data?

- Whenever you first get a dataset, you want to get a sense of it:
    - Where are its ends?
    - Where is its middle?
    - How big of a spread are there?
- Consider this data: age of titanic passengers
    - Was it a boat full of children or only adults?
    - What was the youngest and oldest passengers?
    - Were they very young or very old or people from any age were onboard?

Blackboard: Mean, Median, Mode

summary

Average ➜ (arithmetic mean): sum of all values divided by the number of items in the data set.

Median: ➜ The value in the middle of the data set. If there is no middle, then find the two most middle item and take their mean. That would be the median.

Mode ➜ the most frequent item or number in the data set.

Calculate by hand

5 mins

Calculate the mean, median and mode of the following numbers:

a. Salaries: $70k, $90k, $90k, $80k, $130k, $500k, $60K

b. -5, -4, -1, 2, 4, 6, 6, 7, 0

For each series above, which parameter (mean, median or mode) is the better representative of the series?

Teacher Demo: Mean, Mode, Median with Numpy, scipy and pandas (10 mins)

-

5 mins

Titanic - Average Passenger Age

# Starter code:
```
import pandas as pd

# read in the CSV
df = pd.read_csv('titanic.csv')
# create a list of Age values,
# not including N/A values
ls_age = df['Age'].dropna()
```

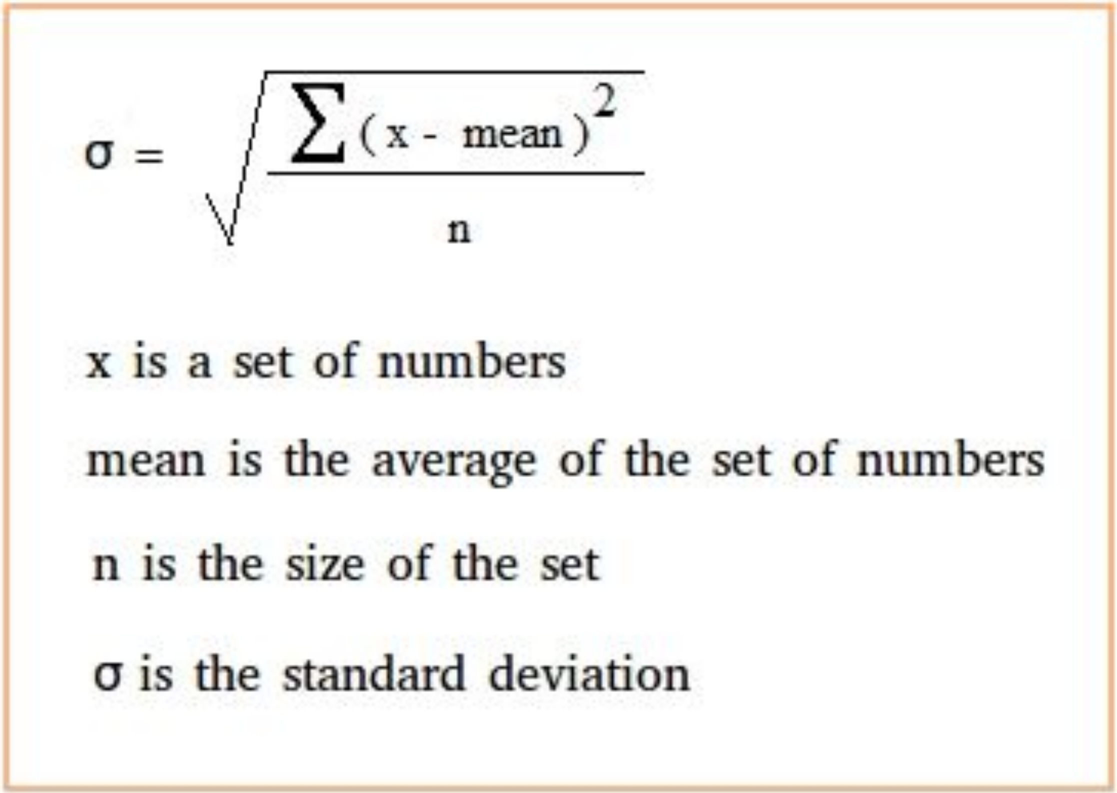
Using Titanic CSV:

1. What is the mean, median and mode of age of passengers on the Titanic?
2. Who was the oldest passenger aboard the ship?
3. How much did the cheapest ticket cost?
4. What was the range of ticket prices? (range = max - min)

10 mins break

Variance and Standard Deviation

Blackboard: Variance and Standard Deviation



$$\sigma = \sqrt{\frac{\sum (x - \text{mean})^2}{n}}$$

x is a set of numbers

mean is the average of the set of numbers

n is the size of the set

σ is the standard deviation

Variance & Standard deviation (manually)

5 mins

Calculate variance and standard deviation of the following two salary series :

a. Software developer salaries: $70k, $90k, $90k, $80k, $500k, $60K
b. High School Teacher salaries: $70k, $47k, $55k, $55k, $62k

Which job has more salary variation?

Teacher Demo: Calculating Variance & SD programmatically (10 mins)

- Let’s solve the previous question (the previous slide) programmatically. We have:
    - Software developer salaries: $70k, $90k, $90k, $80k, $500k, $60k
    - Teacher salaries: $70k, $47k, $55k, $55k, $62k

Variance & Standard deviation (programmatically)

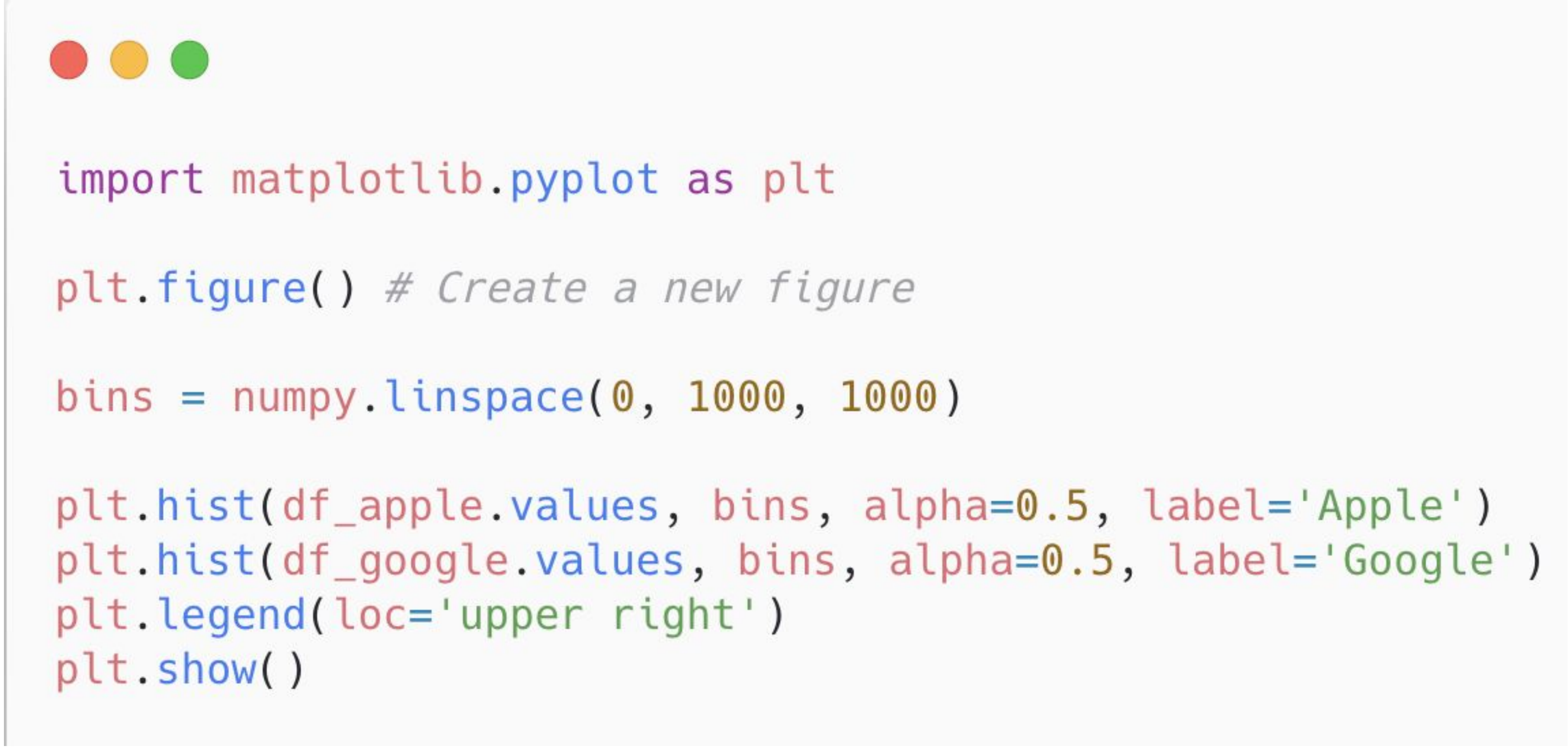
10 mins

We are going to compare the Standard deviation of Apple stock price with google's.

Here is the activity.

Plotting data

Let’s plot the data so we see their distribution:



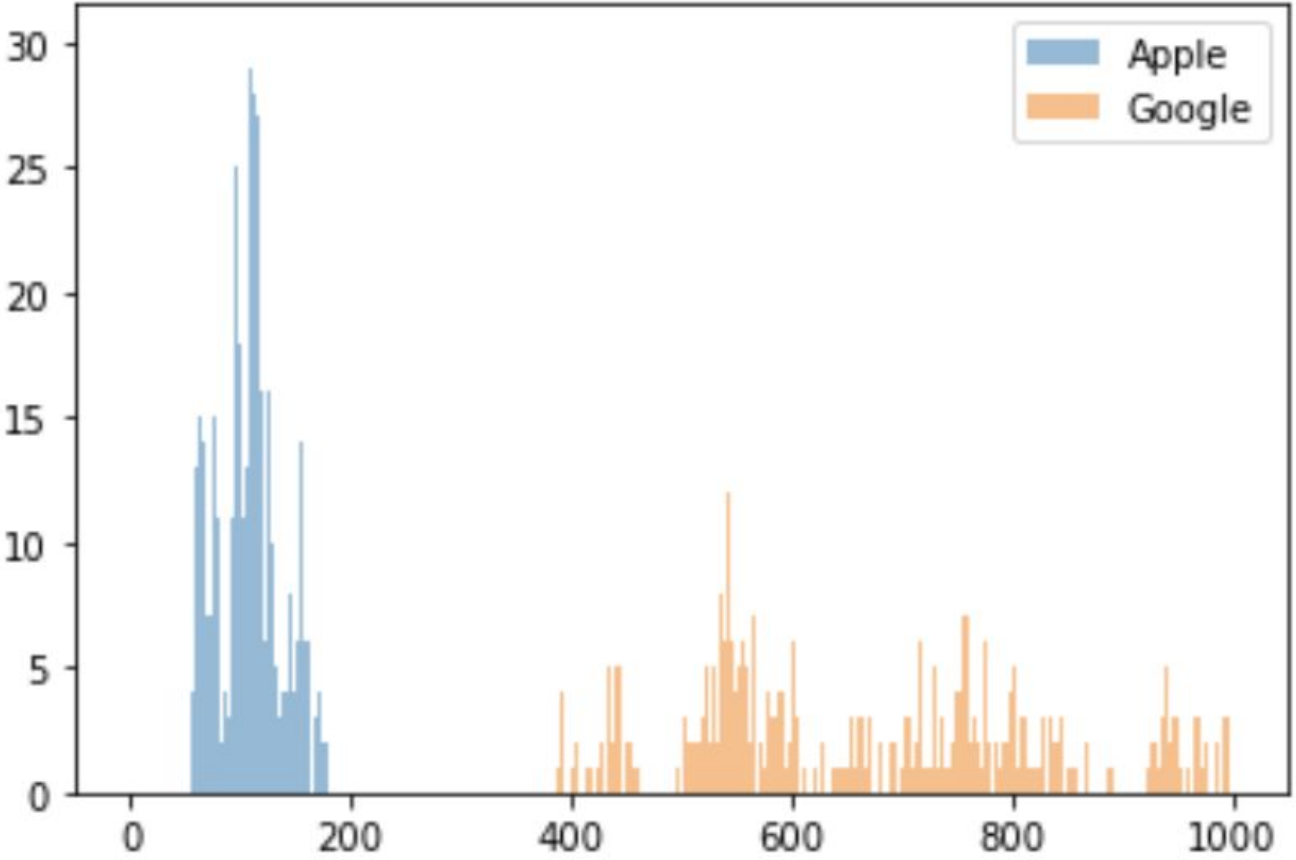
```python
import matplotlib.pyplot as plt

plt.figure() # Create a new figure

bins = numpy.linspace(0, 1000, 1000)

plt.hist(df_apple.values, bins, alpha=0.5, label='Apple')
plt.hist(df_google.values, bins, alpha=0.5, label='Google')
plt.legend(loc='upper right')
plt.show()
```

Standard Deviation Figure



Summary

- Variance and standard deviation are measure of how much the data is spread.
- Standard deviation is more used because it has the same unit as the data.