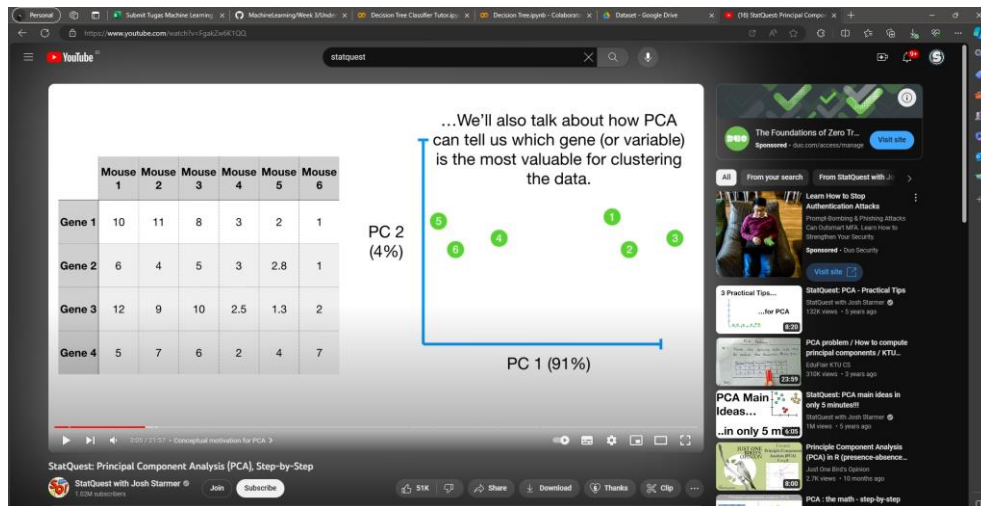


Nama : Sabilly Artowibowo

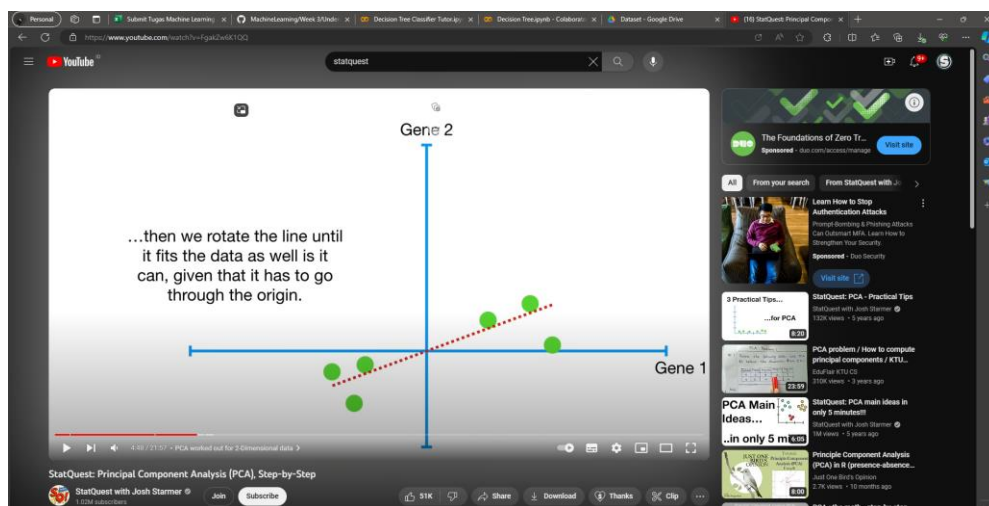
NIM : 1103204057

3 Link Statquest.

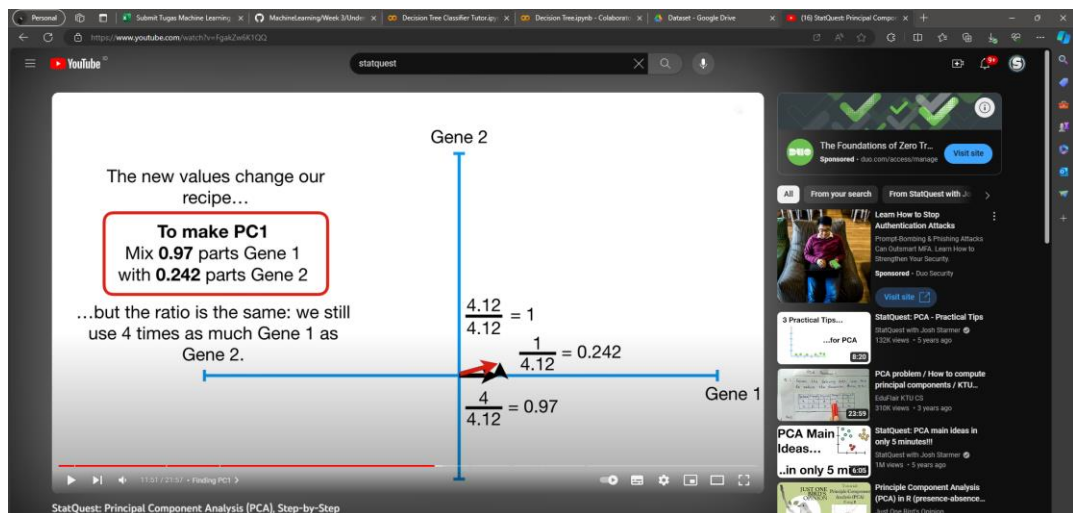
### 1. StatQuest: Principal Component Analysis (PCA), Step-by-Step



Conceptual motivation for PCA



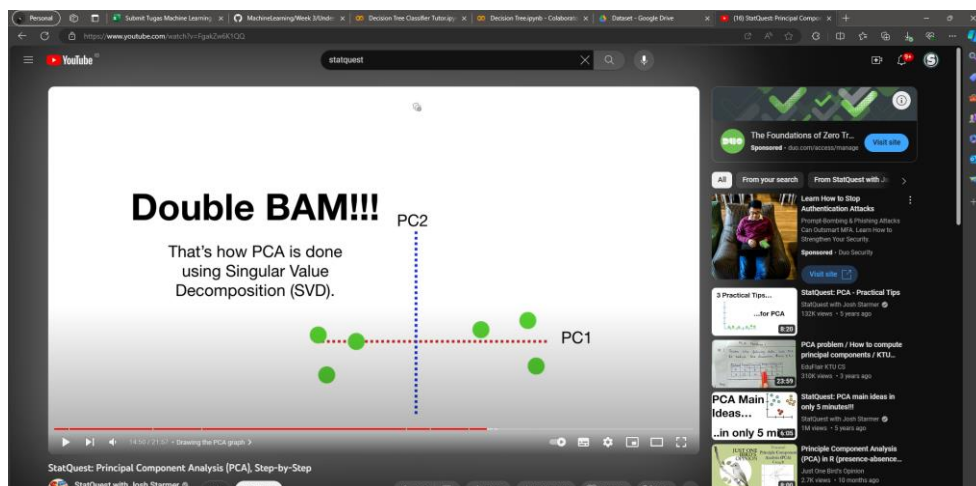
PCA Worked out for 2D data



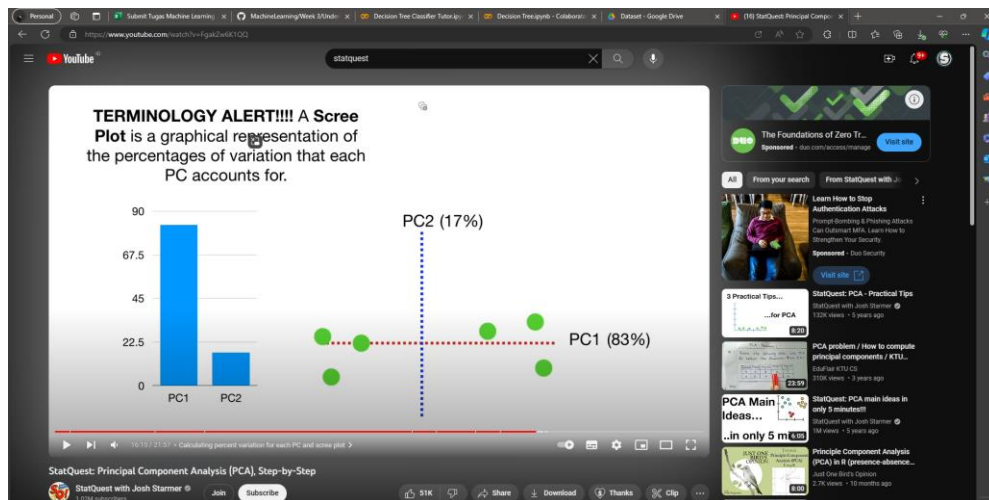
Finding PC1



Finding PC2



## Drawing the PCA Graph



## Calculating percent variation

### 2. StatQuest: PCA in Python

```
import pandas as pd
import numpy as np
import random as rd
from sklearn.decomposition import PCA
from sklearn import preprocessing
import matplotlib.pyplot as plt

genes = ['gene' + str(i) for i in range(1,101)]
wt = ['wt' + str(i) for i in range(1,6)]
ko = ['ko' + str(i) for i in range(1,6)]

data = pd.DataFrame(columns=['wt', 'ko'], index=genes)

for gene in data.index:
    data.loc[gene, 'wt'] = np.random.poisson(lam=rd.randrange(10,1000), size=5)
    data.loc[gene, 'ko'] = np.random.poisson(lam=rd.randrange(10,1000), size=5)

print(data.head())
```

The head() method returns the first 5 rows of data.

## Load Modules and generate data

```
scaled_data = preprocessing.scale(data.T)
```

One last note about scaling with sklearn vs scale() or prcomp() in R:  
In sklearn, variation is calculated as:

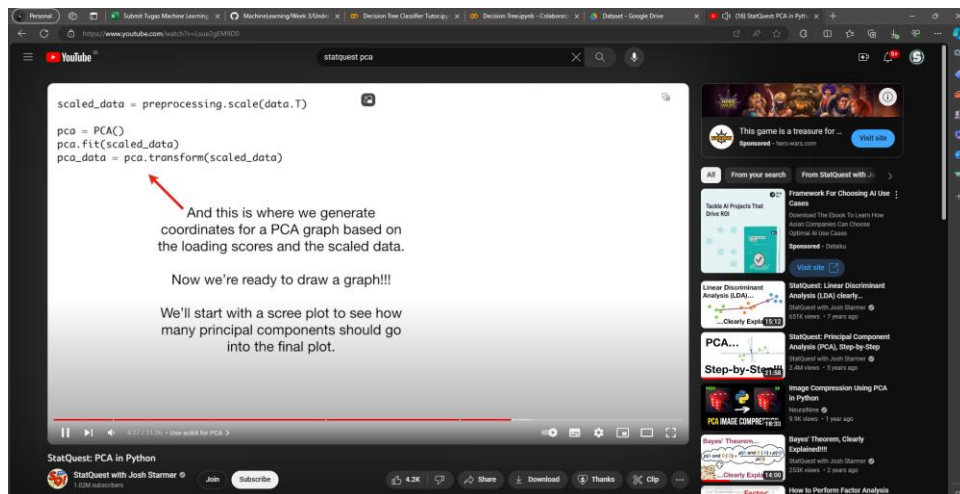
$$\frac{(\text{measurements} - \text{mean})^2}{\text{the number of measurements}}$$

In R using scale() or prcomp(), variation is calculated as:

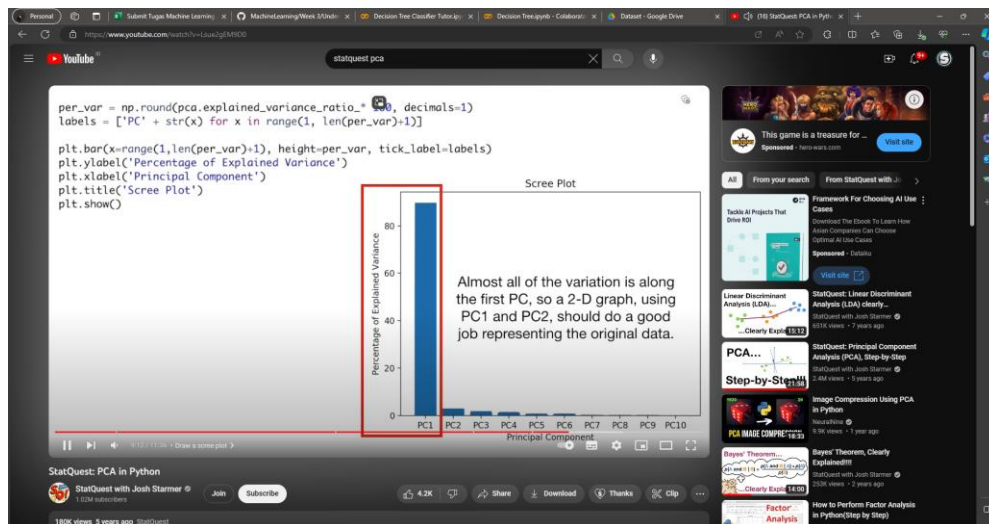
$$\frac{(\text{measurements} - \text{mean})^2}{\text{the number of measurements} - 1}$$

The bad news is that these differences will have a minor effect the final graph. This because the coordinates on the final graph come from multiplying the loading scores by the scaled values.

## Scaling and centering



Use scikit for PCA

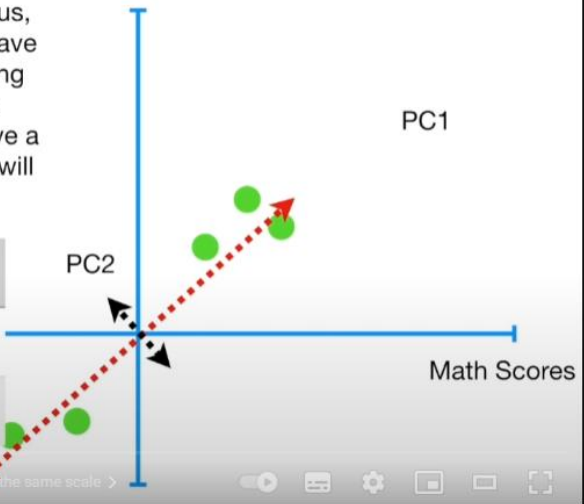


Draw a scree Plot

### 3. StatQuest: PCA - Practical Tips

The standard practice is to divide each variable by its standard deviation. Thus, if a variable has a wide range, it will have a large standard deviation and dividing by it will scale the values a lot. If a variable has a narrow range, it will have a small standard deviation and scaling will be minimal.

	Student 1	Student 2	Student 3	Student 4	...
Math	9.5	8.8	9.3	7.5	...
Reading	9	8	10	7	...

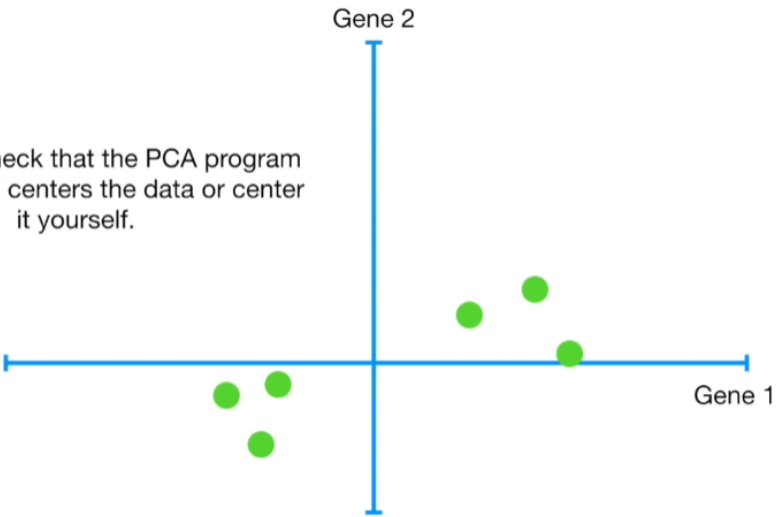


StatQuest: PCA - Practical Tips

StatQuest with Josh ...  
1.02M subscribers

Join Subscribe 3K Share Download ...

Make sure the data are on the same scale



So double check that the PCA program you are using centers the data or center it yourself.

StatQuest: PCA - Practical Tips

StatQuest with Josh ...  
1.02M subscribers

Join Subscribe 3K Share Download ...

Make sure data centered