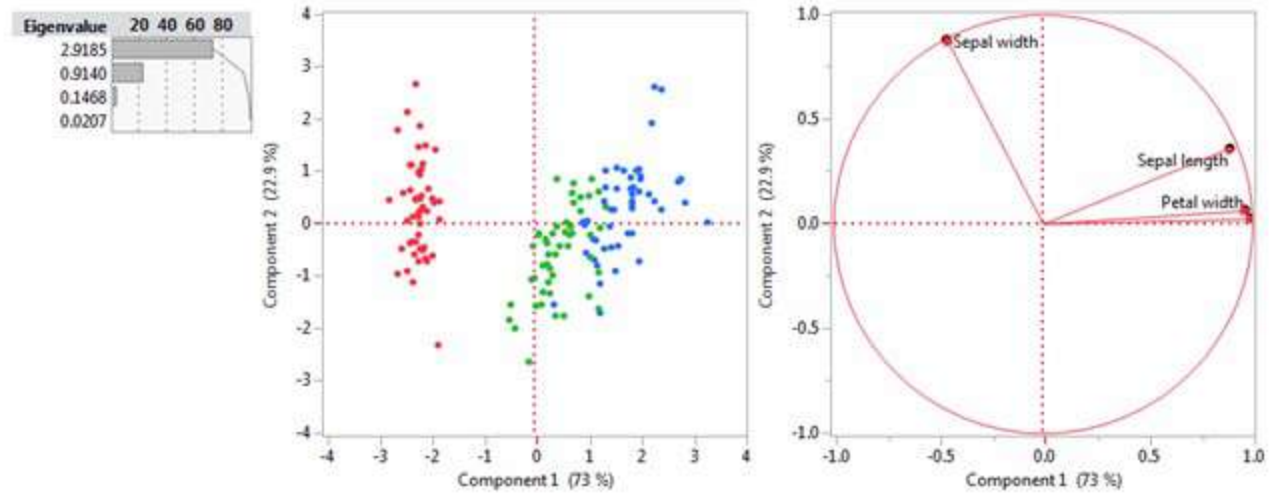
Master of Technology in Intelligent Systems

Problem Solving using Pattern Recognition

Module 3: Dimensionality Reduction

Charles Pang
Institute of Systems Science
National University of Singapore
Email: charlespang@nus.edu.sg



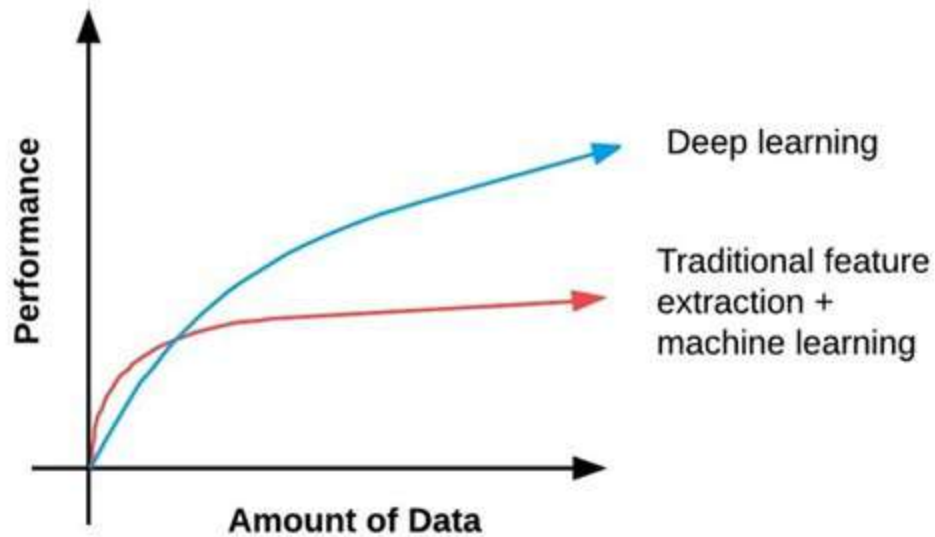
What is Dimensionality?

- The **dimensionality** of a dataset is the number of attributes or features/variables that the data possesses.
- For example:
  - Healthcare data: Height, Weight, Blood Pressure, Blood Glucose, Cholesterol
  - Image data: 8x8 image = 64 dimensions
  - Netflix rating: Customer ratings (480k) on every film (18k)
- Data with a *small* number of features are called **low dimensional data**
- Data with a *large* number of features are called **high dimensional data**

	Anne	Ben	Charlie	Doug	Eve	...
Star Wars	2	5	4	4	3	...
Harry Potter	3	4	5	3	?	...
Pretty Woman	4	?	2	?	5	...
Titanic	5	?	2	1	3	...
Lord of the Rings	?	5	5	4	4	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

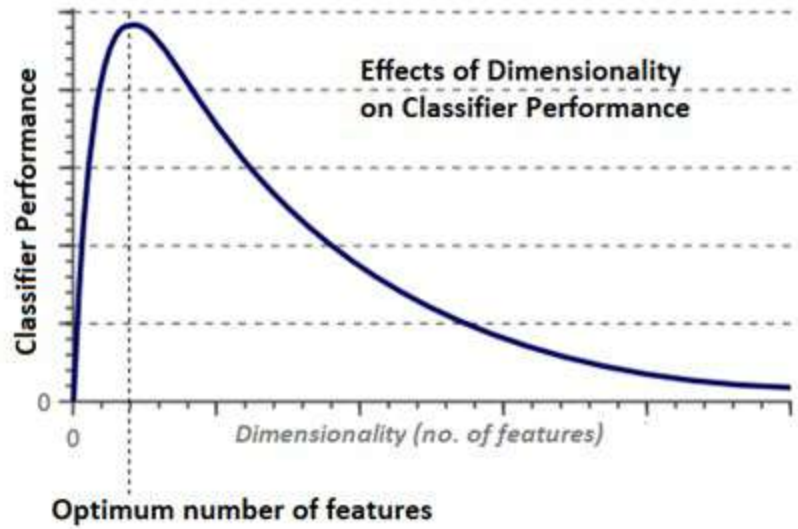
High Dimensionality and Computational Complexity

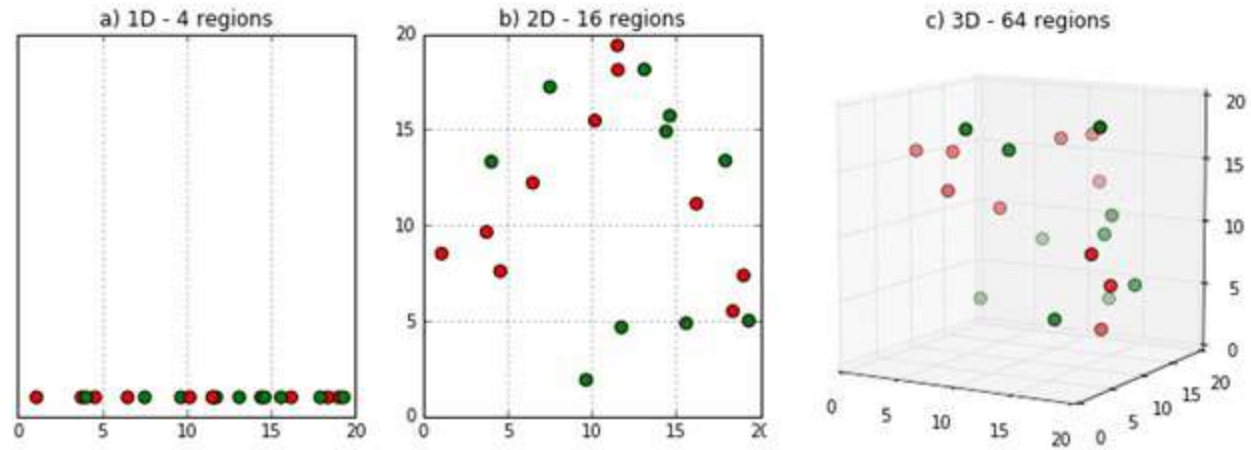
- High dimensional data tend to increase the **computational complexity** of machine learning algorithms.
- Large dataset occupy **bigger storage** spaces, incur high computational **loads**, build **complex analytical models**, and cause model **overfitting**.
- All these problems will result in slower learning cycles, and **poor prediction performance** on real data.



The Curse of Dimensionality

- The curse of dimensionality is a term introduced by Richard E. Bellman (1957) to describe the problem caused by the number of samples needed to estimate an arbitrary function with a given level of accuracy.
- As the number of features or dimensions grows, the amount of data you need to generalize accurately **grows exponentially**.
- Classifier performance will deteriorate after a threshold is reached.

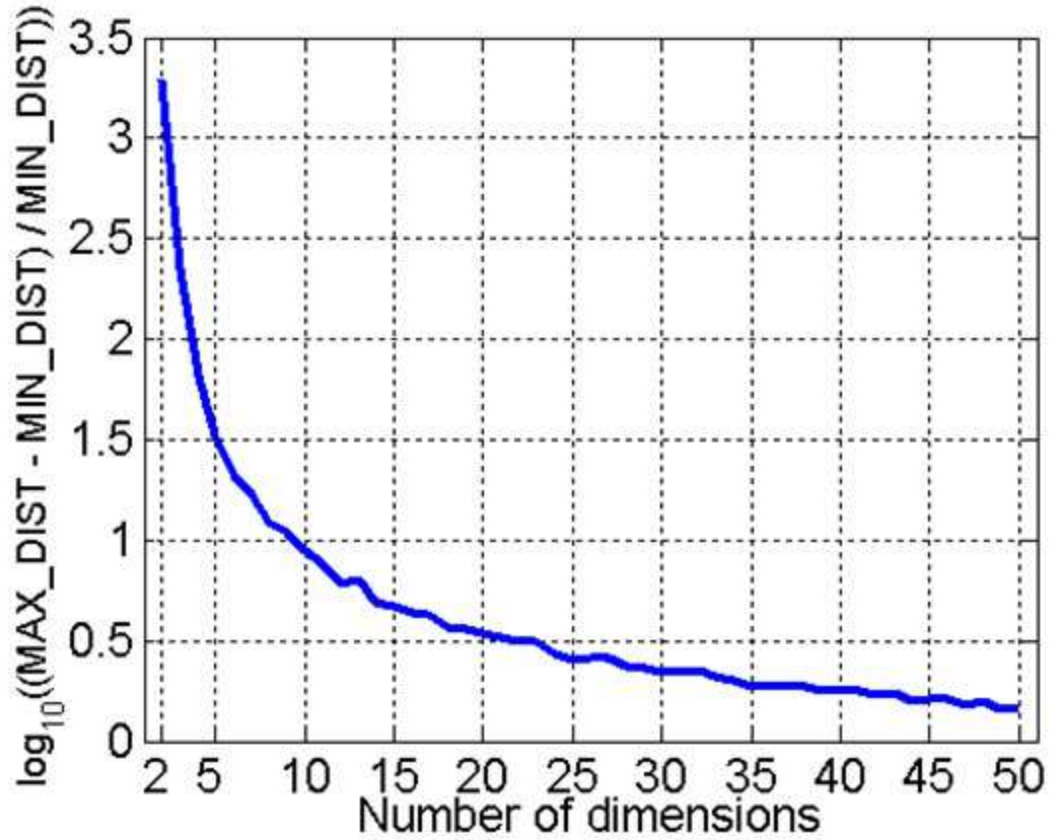




www.kdnuggets.com

Curse of Dimensionality (contd)

- The exponential growth in data causes **high sparsity** in the data space and unnecessarily increases storage space and processing time for the particular modelling algorithm.
- Moreover, the definitions of **density and distance** between points, which are critical for clustering and outlier detection, become less meaningful



Randomly generate 500 points
Compute difference between max and min distance between any pair of points

Intro to Datatmining, Tan, et.al.

Lower Dimensionality is Desirable

- Machine learning algorithms usually work better with a smaller set of features versus a larger set of features.
- When irrelevant or redundant features are eliminated, noise is reduced to yield more accurate and efficient models
- Fewer features will result in a more understandable model
- Visualisation is more possible with lower dimensional data
- The overall problem complexity is hopefully reduced, and hence time and effort spent by data scientists becomes more productive.

Methods to deal with high dimensionality

- Dimensionality reduction is useful in helping us to understand underlying **patterns and trends using simple visualisation**.
- Dimensionality reduction can also be a part of **Data Preparation** (i.e. the step before analysis).
- These are some ways to deal with high dimensionality:
  - Aggregation
  - Discretization
  - Feature Scaling
  - Sampling
  - Feature Subset Selection (feature elimination)
  - Feature extraction (Principal Component Analysis)

Feature Engineering?

Aggregation

- Combining two or more features into a single feature
- Using **ratios**:

age	ed	employ	address	income	creddebt	othdebt	default
41	3	17	12	176	11.359392	5.008608	1
27	1	10	6	31	1.362202	4.000798	0
40	1	15	14	55	0.856075	2.168925	0

age	ed	employ	address	income	creddebt	othdebt	DebInc	default
41	3	17	12	176	11.359392	5.008608	0.093	1
27	1	10	6	31	1.362202	4.000798	0.173	0
40	1	15	14	55	0.856075	2.168925	0.055	0

- Using **change of scale**:
  - Cities aggregated into regions, states, countries, etc.
  - Days aggregated into weeks, months, or years

Discretization

- Converting a continuous attribute into an ordinal attribute- thereby diminishing data from a large domain of numeric values to a small subset of categorical values
- Discretization is one of the most influential data pre-processing task in machine learning because it can result in **remarkable improvements in learning speed and accuracy.**
- Many classification algorithms such as Decision Tree work best if both the independent and dependent variables have only a few values- producing shorter, more compact trees with accurate results.
- Two simple methods are (1) Equal Binning and (2) Equal Frequency

Feature Scaling

- Feature scaling is used to map an entire set of values to a new set of replacement values.
- Two methods are popularly used:
  - **Normalization** reduces the data to [0,1]. This is useful in cases where you need to have your parameters on the same positive scale. Note that outliers information will be lost

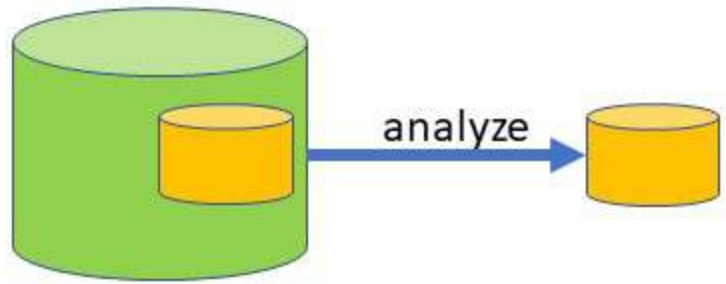
$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

  - **Standardization** rescales the $\frac{X - \mu}{\sigma}$ ave a mean of zero and a standard deviation of one.

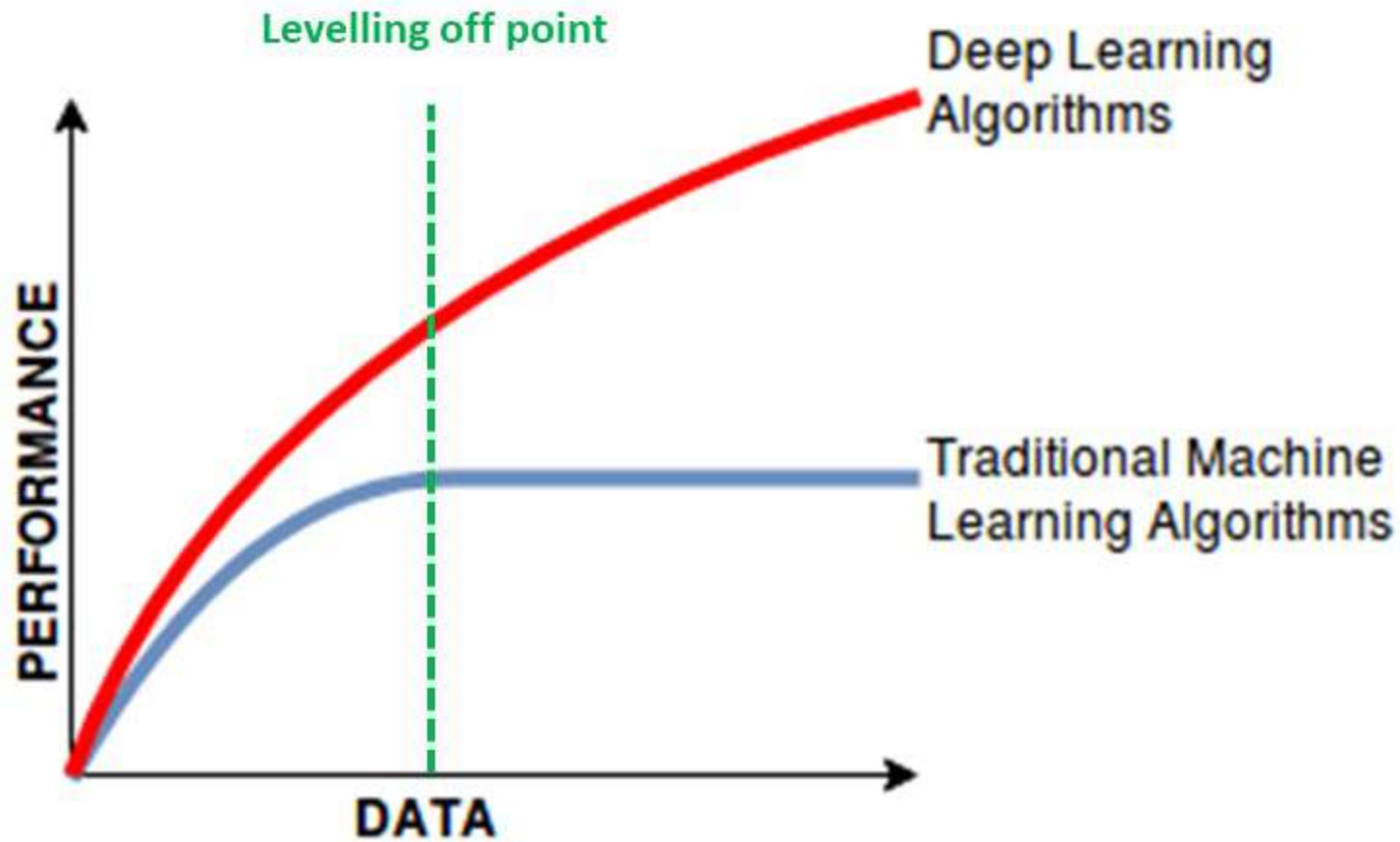
- While feature scaling and discretization are not specifically dimension reduction, it reduces complexity and helps improve learning algorithms.

Sampling

- Sampling is a technique used by statisticians because the cost of obtaining the entire set of data is too cost- and/or time- prohibitive
- For the Data Analysts, sampling is used because it is **too expensive or time consuming to process ALL the data**
- The key principle for effective sampling is as follows:
  - Sample must be **representative** of the entire data set
  - A sample is representative if it has approximately the same properties (of interest) as the original set of data



Determining the Proper Sample Size



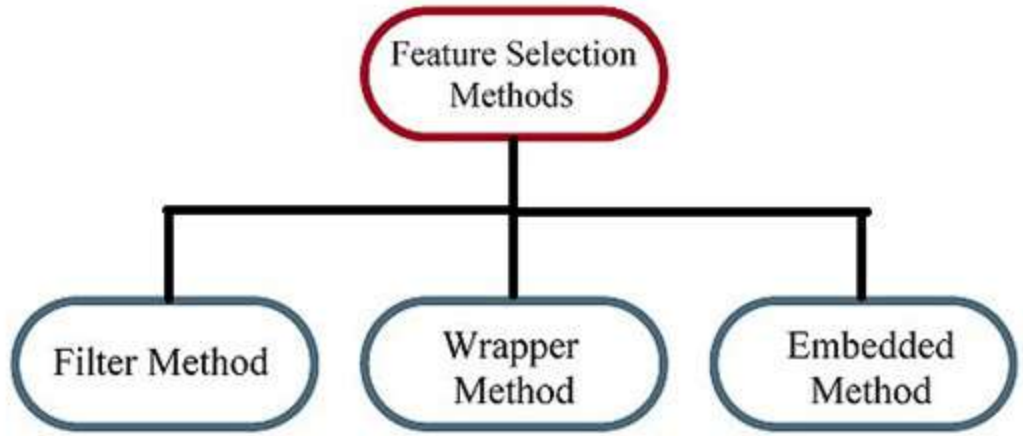
Feature Subset Selection

- We can reduce dimensionality by using only a subset of the features.
- On the surface, it might appear that we are throwing away useful information. But what if these are redundant and irrelevant features?
- Redundant features duplicate much or all of the information contained in one or more other attributes.
- For example, it is easy to see that *Total_Sales* being the product of *Price* and *Quantity*, can intuitively be eliminated without lost of information.
- However, features such as *Total Cholesterol, HDL, LDL, Triglyceride* may appear to make *Total Cholesterol* redundant- but in this case it can be a bit trickier.

Introduction to Data Mining , Tan, Steinbach, Kumar

Feature Subset Selection Methods

- Ideally, we want to try out all possible subsets of features and use them as input to the mining algorithm, and then take the subset that produces the best results.
- This method has the advantage of reflecting the objective and bias of the data mining algorithm that will eventually be used. In other words, you are customizing your data to a specific algorithm.



Feature Subset Selection methods

- **Embedded approaches** Feature selection occurs naturally as part of the data mining algorithm. Specifically, during its operation, the algorithm itself decides which attributes to use and which to ignore. E.g. decision tree classifiers.
- **Filter approaches** Features are selected before the data mining algorithm is run, using some approach that is independent of the analytical task. For example, we might select sets of attributes whose pairwise correlation is as low as possible. The disadvantage is that it ignores the inter-relationships of features that can affect prediction performance.
- **Wrapper approaches** tries to prevent the filter issues.
  Backward Feature Elimination: This method starts with all features present and removes one feature at the time.
  Forward Feature Selection: This method starts with no feature and adds one at a time.

Introduction to Data Mining , Tan, Steinbach, Kumar

Other Univariate techniques

Involves manual work by examining every feature and checking its importance with the target

- **Redundant/Irrelevant** – eliminating features that do not impact the accuracy or prediction – using domain knowledge
- **Missing Values** – features with too many missing values are unlikely to convey much useful information.
- **Low variance** – values in features that don't vary much behave like constants and will have no predictive power to the model
- **High correlation** – features with very similar trends are also likely to carry very similar information
- **Domain knowledge** – manually decide which features are not useful for the specific modelling purpose.

What is Variance?

- Variance is a measurement of the **spread between numbers** in a data set. It is calculated as the *average of the squared deviation from the mean.*

$$\text{variance } \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

- A variance of **zero** indicates that all the values within the feature are identical;
- A **large** variance indicates that values within the feature are far from the mean and each other, while a **small** variance indicates the opposite.
- We can analyse the variance by using either Variance (σ^2) or the Standard Deviation (σ)

Covariance Matrix

	Income	Education	Age	Residence	Employ	Savings	Debt	Credit cards
Income	3.424516e+08	21652.827586	44102.275862	25579.724138	22636.206897	3.324515e+07	-1.830487e+07	-1358.827586
Education	2.165283e+04	4.547126	2.262069	0.914943	0.379310	8.156736e+03	-4.909517e+03	-0.790805
Age	4.410228e+04	2.262069	21.406897	15.434483	14.379310	2.185434e+04	7.555862e+02	-0.751724
Residence	2.557972e+04	0.914943	15.434483	15.857471	13.896552	1.940961e+04	3.731172e+03	0.266667
Employ	2.263621e+04	0.379310	14.379310	13.896552	13.431034	1.688379e+04	4.558621e+03	0.103448
Savings	3.324515e+07	8156.735632	21854.344828	19409.609195	16883.793103	7.311108e+07	-1.692310e+07	-4394.574713
Debt	-1.830487e+07	-4909.517241	755.586207	3731.172414	4558.620690	-1.692310e+07	2.542695e+07	2994.758621
Credit cards	-1.358828e+03	-0.790805	-0.751724	0.266667	0.103448	-4.394575e+03	2.994759e+03	1.567816

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$$

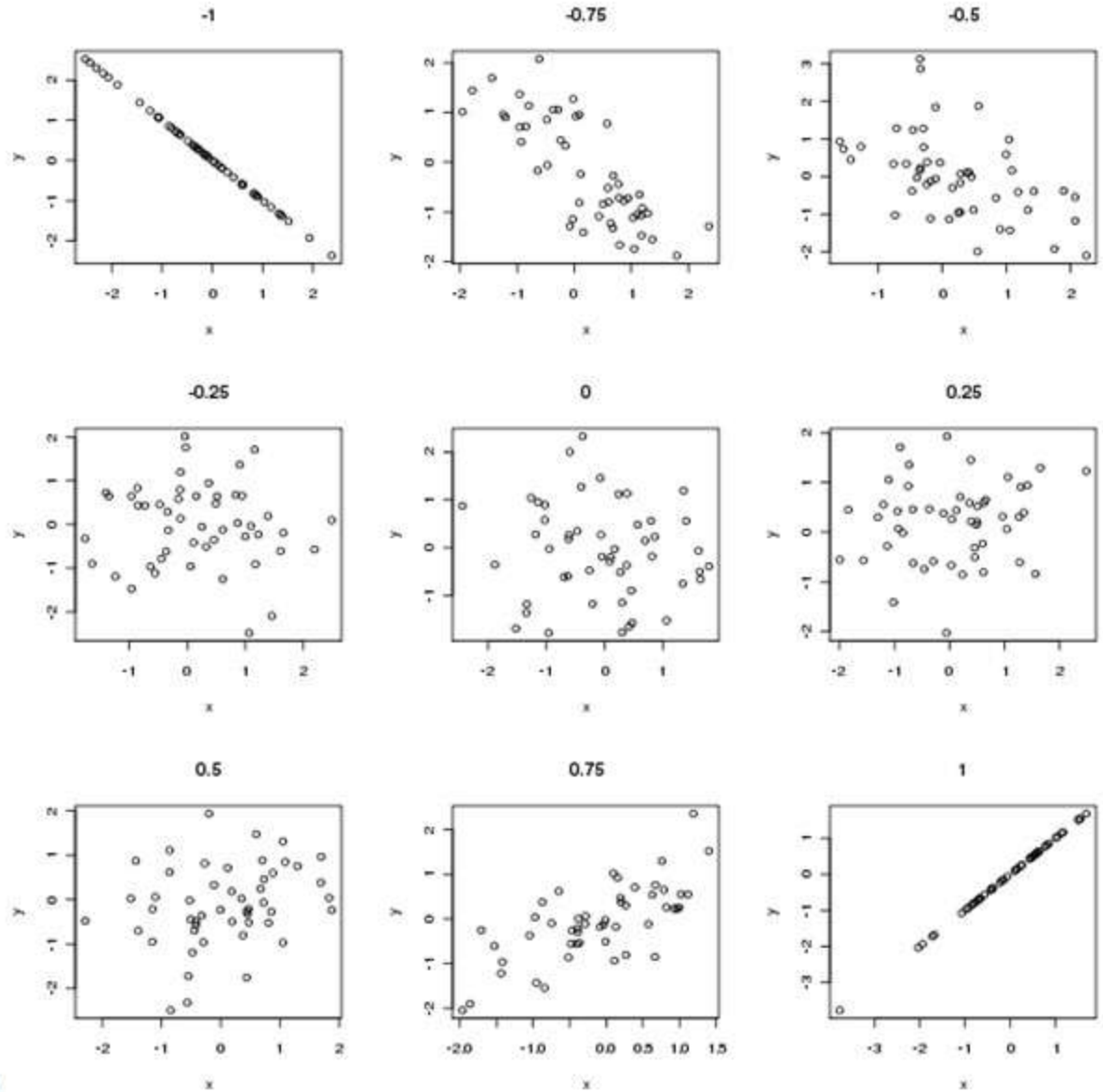
```
data.cov()
```

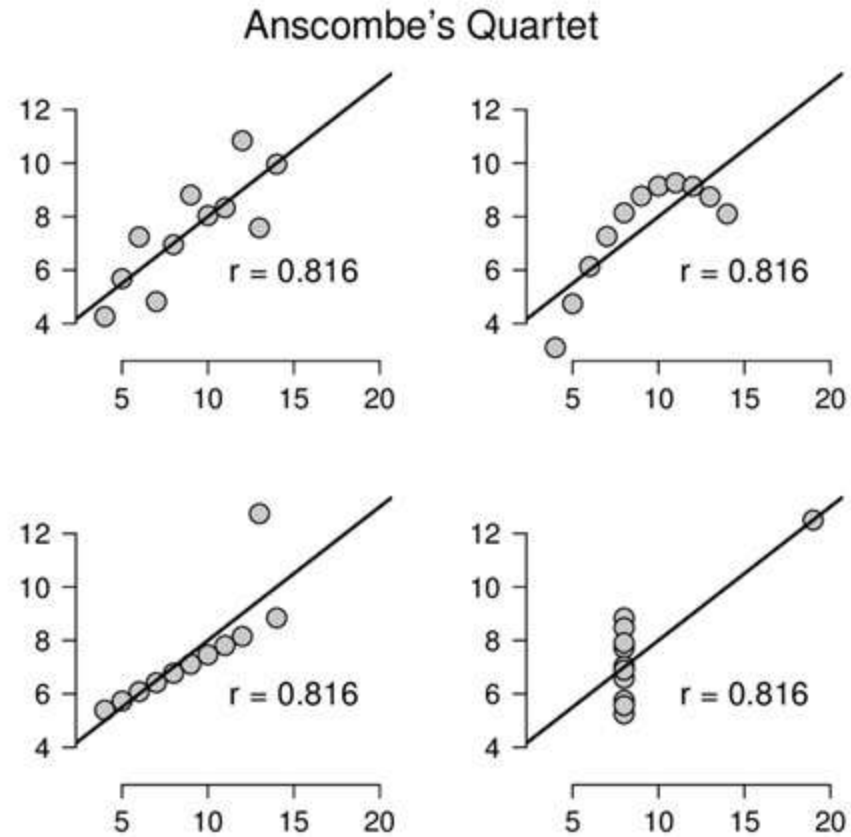
What is Correlation?

Correlation is a measure of the extent to which **two variables are related** or associated.

- There are **three possible results** of a correlation analysis:
  - **Positive correlation:** both variables either increase or decease at the same time. Taller people tend to be heavier.
  - A **negative correlation:** increase in one variable is associated with a decrease in the other. Time you spent playing games versus your test scores .
  - **Zero correlation:** no relationship between two variables. Number of cockroaches in your home and your salary amount
- When a pair of variables are highly correlated it maybe possible to remove one to reduce dimensionality without much loss of information.
- Which one should we keep? Usually the one with a higher correlation to the target.

Examples of Correlations





High correlation doesn't necessarily mean that the underlying relationship is linear

Correlation Matrix

	Income	Education	Age	Residence	Employ	Savings	Debt	Credit cards
Income	1.000000	0.548715	0.515092	0.347120	0.333772	0.210105	-0.196164	-0.058643
Education	0.548715	1.000000	0.229277	0.107748	0.048537	0.447359	-0.456587	-0.296178
Age	0.515092	0.229277	1.000000	0.837719	0.848021	0.552420	0.032386	-0.129758
Residence	0.347120	0.107748	0.837719	1.000000	0.952216	0.570044	0.185815	0.053482
Employ	0.333772	0.048537	0.848021	0.952216	1.000000	0.538795	0.246679	0.022543
Savings	0.210105	0.447359	0.552420	0.570044	0.538795	1.000000	-0.392501	-0.410466
Debt	-0.196164	-0.456587	0.032386	0.185815	0.246679	-0.392501	1.000000	0.474315
Credit cards	-0.058643	-0.296178	-0.129758	0.053482	0.022543	-0.410466	0.474315	1.000000

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})}$$

```
data.corr()
```

Difference between Covariance and Correlation Matrices

- Both measure the relationship and the dependency between two variables.
- Covariance indicates the **direction** of the linear relationship between variables.
- Correlation on the other hand measures both the **strength and direction** of the linear relationship between two variables.
- What sets them apart is the fact that **correlation values are standardized** whereas, covariance values are not.

Feature Subset Selection Summary

- The advantages of feature elimination methods include simplicity and maintaining interpretability of your variables.
- The use of domain knowledge means that you are keeping only the features that are intuitively relevant to solving your problem.
- Algorithms such as Decision Trees or Cluster Analysis can help detect unimportant features that can be ignored.
- On the flipped side, you may loose the opportunity of gaining some beneficial information from those features that you have eliminated.
- You will have to experiment.

Feature Extraction Techniques

- Dimensionality reduction is a **feature extraction** method as opposed to feature elimination that we have seen earlier.
- In feature extraction, all original variables are extracted to form **new independent variables**.
- Suppose you have 20 features in your dataset. Feature extraction will create (as many as 20) new independent variables where each new variable is a **combination of each of the original features**.
- You can choose to **keep only the more important** (new) variables and ignore the least important ones.
- Principal Component Analysis is a very popular feature extraction method.

PRINCIPAL COMPONENT ANALYSIS

Background of Principal Component Analysis

- PCA was invented in 1901 by Karl Pearson, as an analogue of the principal axis theorem in mechanics; it was later independently developed and named by Harold Hotelling in the 1930s.
- It is known by many names depending on the field of application: discrete Karhunen–Loève transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, proper orthogonal decomposition (POD) in mechanical engineering, etc.
- PCA concepts can be challenging even for the mathematical oriented students, since the statistical definitions need to be connected to matrix algebra.
- This course does not go into the math but will concentrate on its applications using available packaged libraries. There are many resources on the internet for those who are interested in the math.

Source: https://en.wikipedia.org/wiki/Principal_component_analysis

What is Principal Component Analysis?

- PCA uses a vector space transform to project the data from a high-dimensional space into a lower-dimensional space. Its main purpose is to reduce a **correlated multidimensional** data set (x_i) to an **uncorrelated lower dimensional** space (y_j) with **maximum variance** called principal components
- The principal components are linear combinations of the x_s :

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\
y_2 &= a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\
&\vdots \\
y_p &= a_{1p}x_1 + a_{2p}x_2 + \cdots + a_{pp}x_p.
\end{aligned}
$$

- Each component is a weighted sum of the x_i where the a_{ij} are the weights, or coefficients.

Properties of Principal Components

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\
y_2 &= a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\
&\vdots \\
y_p &= a_{1p}x_1 + a_{2p}x_2 + \cdots + a_{pp}x_p.
\end{aligned}
$$

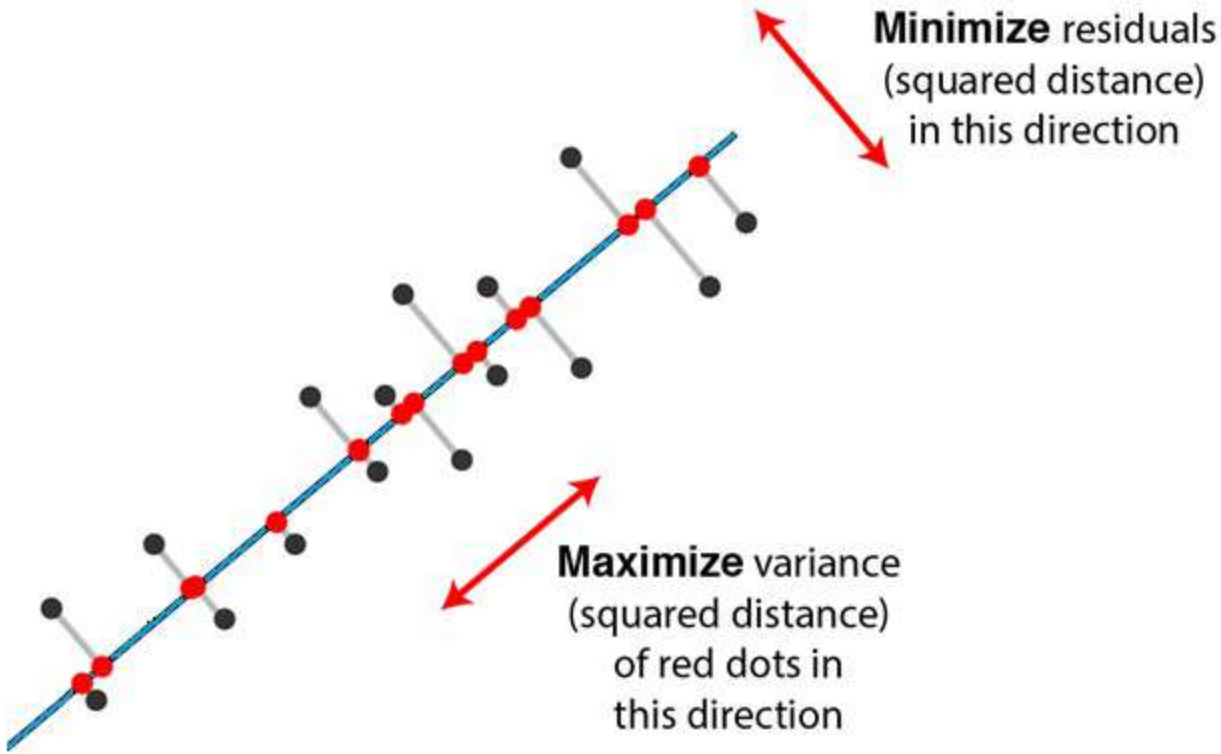
y_s principal component
y_s original Feature
a_{ij} coefficients

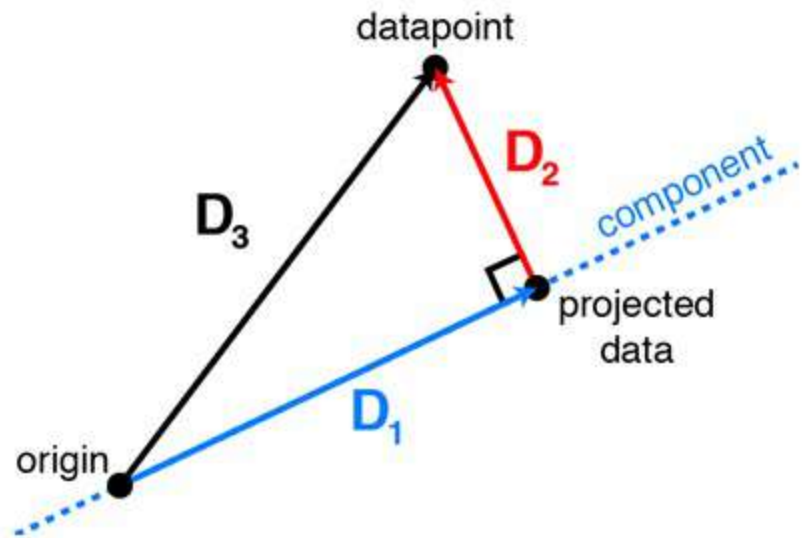
The properties of Principal Components are:

- are linear combinations of the original features,
- are orthogonal (perpendicular) to each other, and
- capture the maximum amount of variance in the data.

Principal Component Analysis Method

- PCA tries to find a new axis that captures the maximum variance within the data once it is projected onto the new axis:

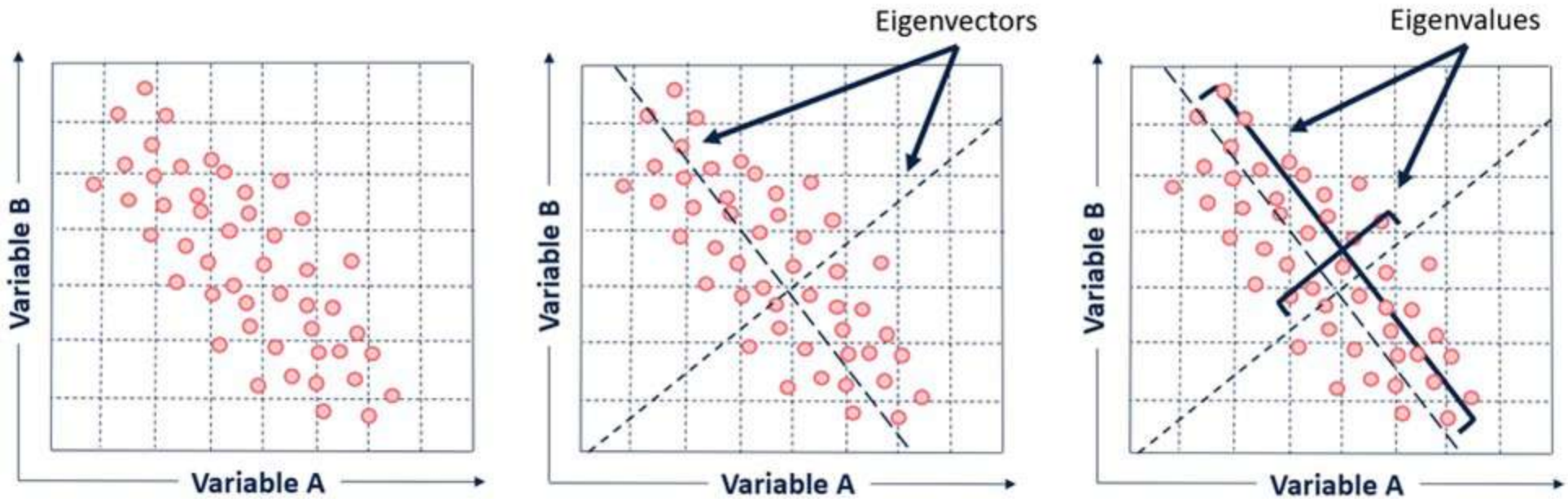




$$D_3^2 = D_1^2 + D_2^2$$

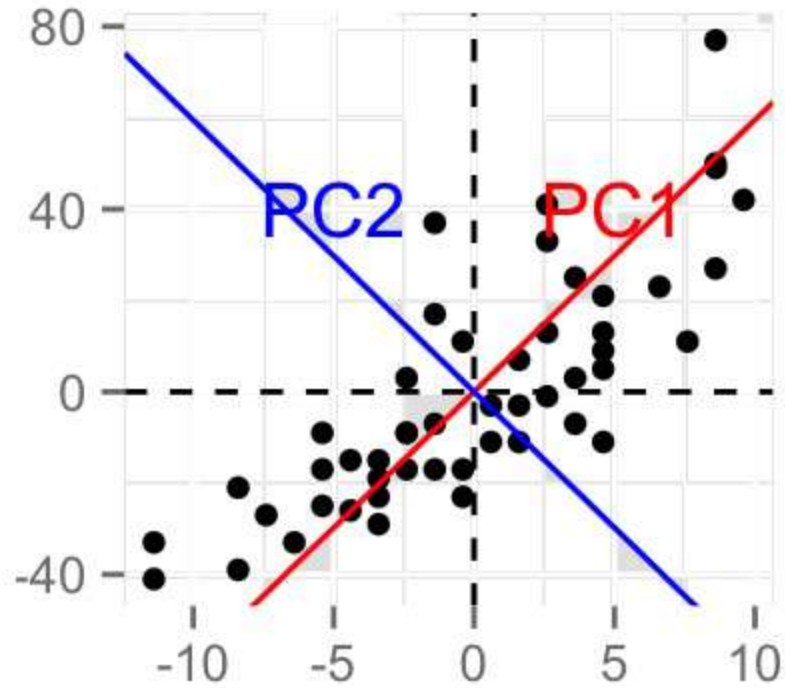
this is constant | maximize this | **or** | minimize this

Concept of eigenvectors and eigenvalues



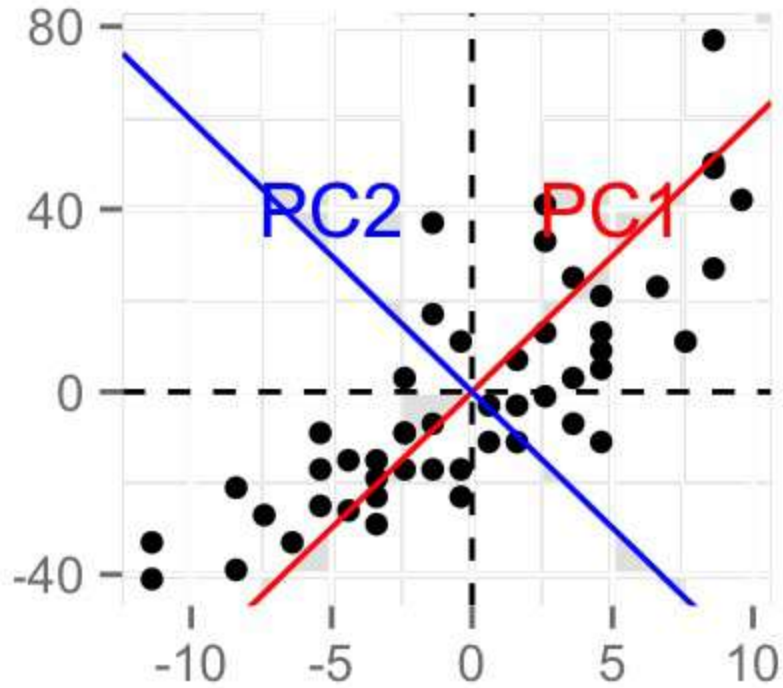
Principal Component Analysis Method

- The new axes are called **principal components** and the values of the new variables are called principal components **score**
- Each principal component is a **linear combination** of the original variables
- The **1st PC** accounts for the maximum variance in the data
- The **2nd PC** accounts for the maximum variance that has not been accounted by the 1st variable
- The **pth PC** accounts for the maximum variance that has not been accounted by the p-1 variables



Principal Component Analysis Method

- The Principal Components themselves are considered as **new variables**, and they are all **uncorrelated**
- If a substantial amount of the total variance in the data is accounted for by a few principal components then one can use these fewer no. of variables (subset) for further investigation instead of original features.
- This subset of variables can be identified by looking at the **eigen** values



PCA EXAMPLE

PCA Example: Loan Application

Income	Education	Age	Residence	Employ	Savings	Debt	Credit cards
50000	16	28	2	2	5000	1200	2
72000	18	35	10	8	12000	5400	4
61000	18	36	6	5	15000	1000	2
88000	20	35	4	4	980	1100	4
91100	18	38	8	9	20000	0	1
45100	14	41	15	14	3900	22000	4
36200	14	29	6	5	100	7000	5
41000	12	34	9	8	5000	200	3
40000	16	32	8	7	19000	1760	2
32000	16	30	2	2	16000	550	1
29000	16	28	1	4	2100	4600	2
21240	12	26	2	2	100	10010	3
58700	12	38	9	9	4500	7800	5
41000	14	29	5	4	300	10000	6
38720	16	36	11	11	24500	540	2
88240	16	38	13	12	13600	8100	2
40000	18	39	7	6	16000	1300	2

Income: Yearly Income
Education: no. of years
Age: years
Residence: years at current addr
Employ: years at current coy
Savings: current saving
Debt: current debts
Credit cards: #credit cards owned

Approach for this Example

1. **Basic Data understanding**
   - Summary statistics
   - Frequency Distributions
   - Correlation Matrix/Heatmap
2. **Check for appropriateness of PCA Method**
   - Bartlett's Sphericity Test
   - Kaiser-Meyer-Olkin Test
3. **Run PCA and examine the outputs**
   - Component Loading matrix
   - Interpretation and labelling
   - Loading Plot
4. **How Many Components to retain?**
   - Eigenvalues (Kaiser Rule)
   - Total variance explained
   - Scree Plot
   - Communalities Matrix
5. **Proceeding to the next step**
   - Eigenvectors
   - Component Scores
   - Score Plot

1.Basic Data Understanding

Summary statistics

Frequency Distributions

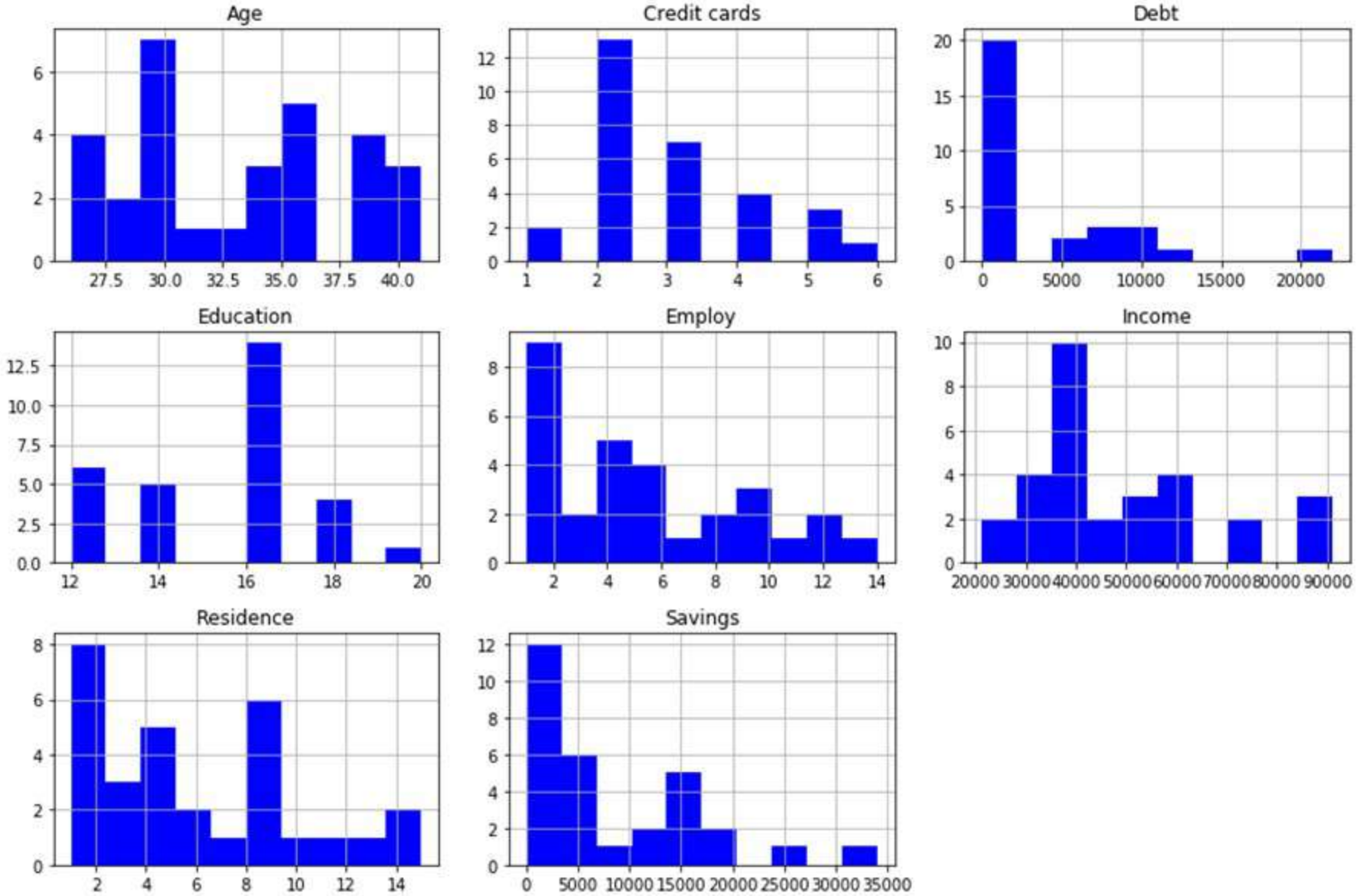
Correlation Matrix/Heatmap

Summary Statistics

```
Data Dimensions:   (30, 8)
Missing Values found: 0
Summary Statistics:
```

	miss	count	mean	std	min	25%	50%	75%	max
Income	0	30.0	48646.00	18505.45	21240.0	37055.00	41100.0	58125.00	91100.0
Education	0	30.0	15.27	2.13	12.0	14.00	16.0	16.00	20.0
Age	0	30.0	32.80	4.63	26.0	29.00	33.0	36.00	41.0
Residence	0	30.0	5.93	3.98	1.0	2.25	5.0	8.00	15.0
Employ	0	30.0	5.50	3.66	1.0	2.00	4.0	8.00	14.0
Savings	0	30.0	8319.33	8550.50	0.0	1950.00	4750.0	14050.00	34000.0
Debt	0	30.0	3812.00	5042.51	0.0	800.00	1200.0	6600.00	22000.0
Credit cards	0	30.0	2.87	1.25	1.0	2.00	2.5	3.75	6.0

Frequency Distributions

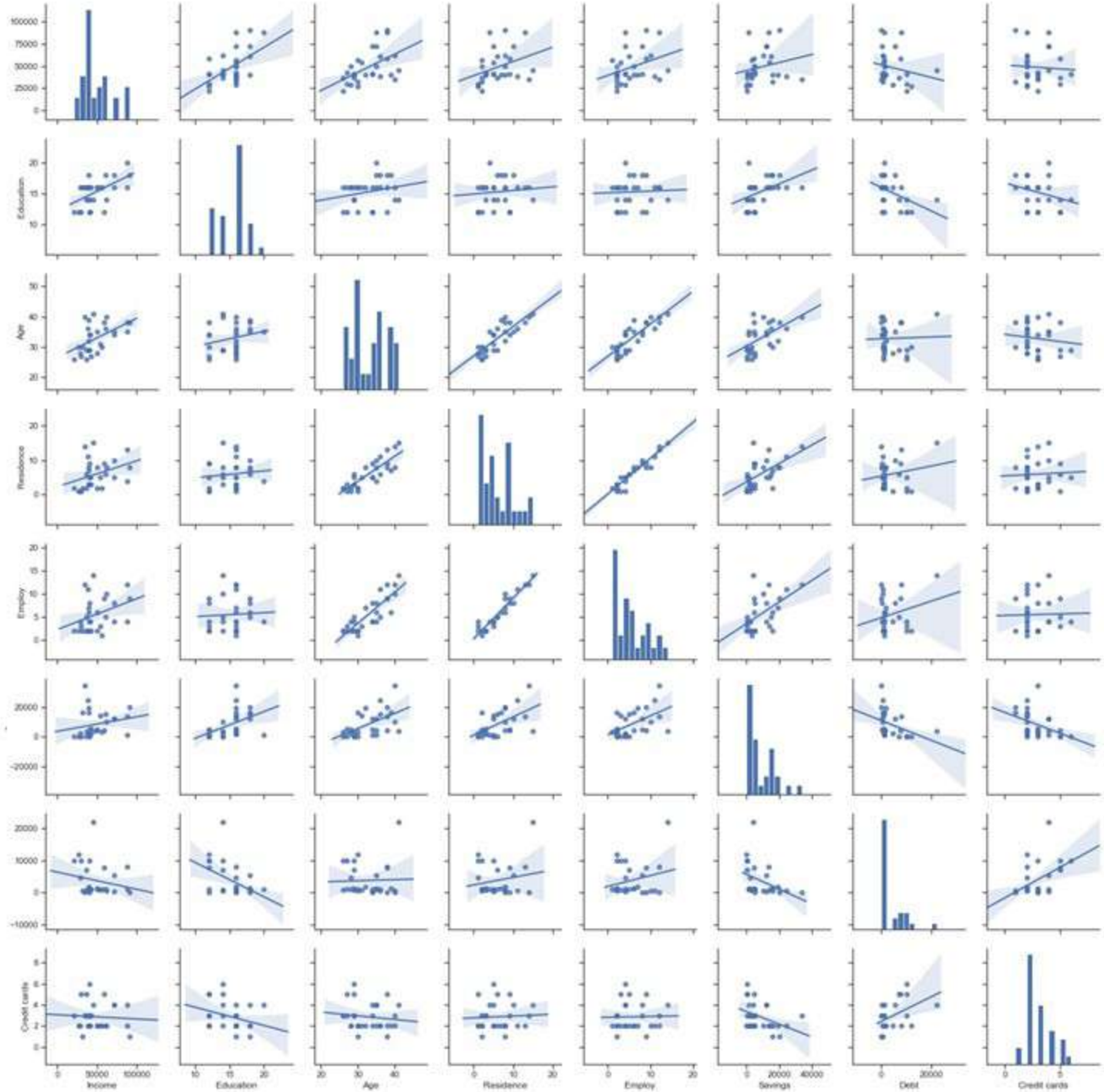


Correlation Matrix

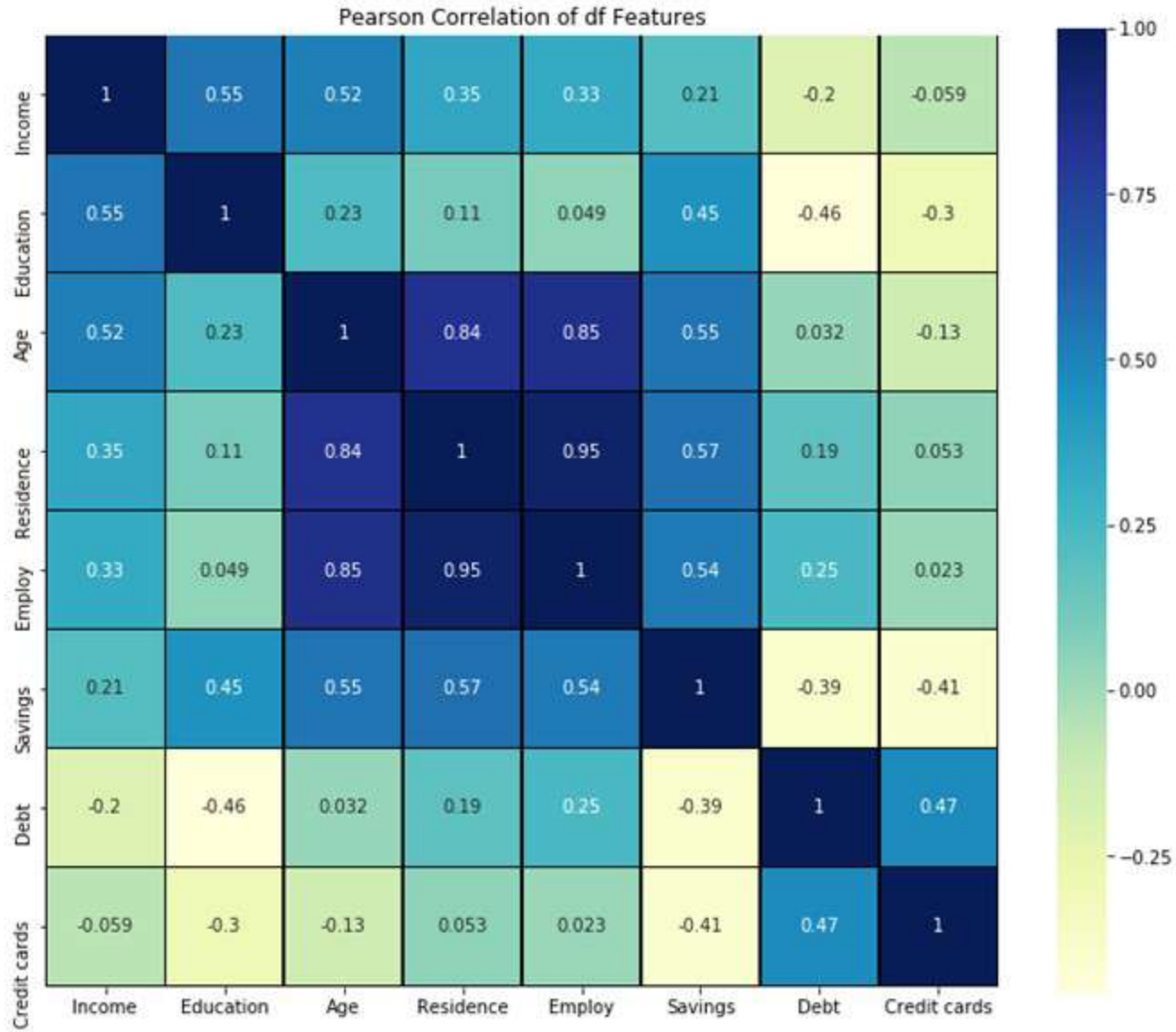
Corelation Matrix:

	Income	Education	Age	Residence	Employ	Savings	Debt	Credit cards
Income	1.00	0.55	0.52	0.35	0.33	0.21	-0.20	-0.06
Education	0.55	1.00	0.23	0.11	0.05	0.45	-0.46	-0.30
Age	0.52	0.23	1.00	0.84	0.85	0.55	0.03	-0.13
Residence	0.35	0.11	0.84	1.00	0.95	0.57	0.19	0.05
Employ	0.33	0.05	0.85	0.95	1.00	0.54	0.25	0.02
Savings	0.21	0.45	0.55	0.57	0.54	1.00	-0.39	-0.41
Debt	-0.20	-0.46	0.03	0.19	0.25	-0.39	1.00	0.47
Credit cards	-0.06	-0.30	-0.13	0.05	0.02	-0.41	0.47	1.00

Scatter Plots



Correlation Heatmap



2.Check for appropriateness of PCA

Bartlett's Sphericity Test

Kaiser-Meyer-Olkin Test

Kaiser-Meyer-Olkin Test

- The **Kaiser-Meyer-Olkin (KMO)** test is a Measure of Sampling Adequacy – i.e. how suited is your data for factor analysis. The KMO is a statistic that indicates the proportion of variance in your variables that might be caused by underlying factors. KMO returns values between 0 and 1.
- A **rule of thumb** for interpreting the statistic:
  - 1 > KMO > 0.8 indicate the sampling is good.
  - 0.8 > KMO > 0.5 indicates the sampling is adequate
  - KMO < 0.5 indicate the sampling is NOT adequate
  - KMO Values close to zero means that there are large partial correlations compared to the sum of correlations. In other words, there are widespread correlations which are a large problem for factor analysis.

Bartlett's Test

- **Bartlett's Test of Sphericity** compares an observed correlation matrix to the identity matrix. Essentially it checks to see if there is a certain redundancy between the variables that we can summarize with a few number of factors.
- The **null hypothesis** of the test is that the **variables are orthogonal**, i.e. not correlated.
- The **alternative hypothesis** is that the **variables are not orthogonal**, i.e. they are correlated enough to where the correlation matrix diverges significantly from the identity matrix.
- This test is often performed before we use a data reduction technique such as principal component analysis or factor analysis to verify that a data reduction technique can actually compress the data in a meaningful way. We seek to REJECT the NULL.

Running the KMO and Bartlett's Test

- These 2 tests provide a minimum standard which should be passed before a Principal Component analysis can be conducted.

```
chi_square_value: 173.01622444105612
P-value: 3.2842239591607713e-23
CONCLUSION: REJECT THE NULL - Correlations EXISTS

Kaiser-Meyer-Olkin: 0.685653942535218
CONCLUSION: Sampling is Adequate
```

3.Run the Principal Component Analysis

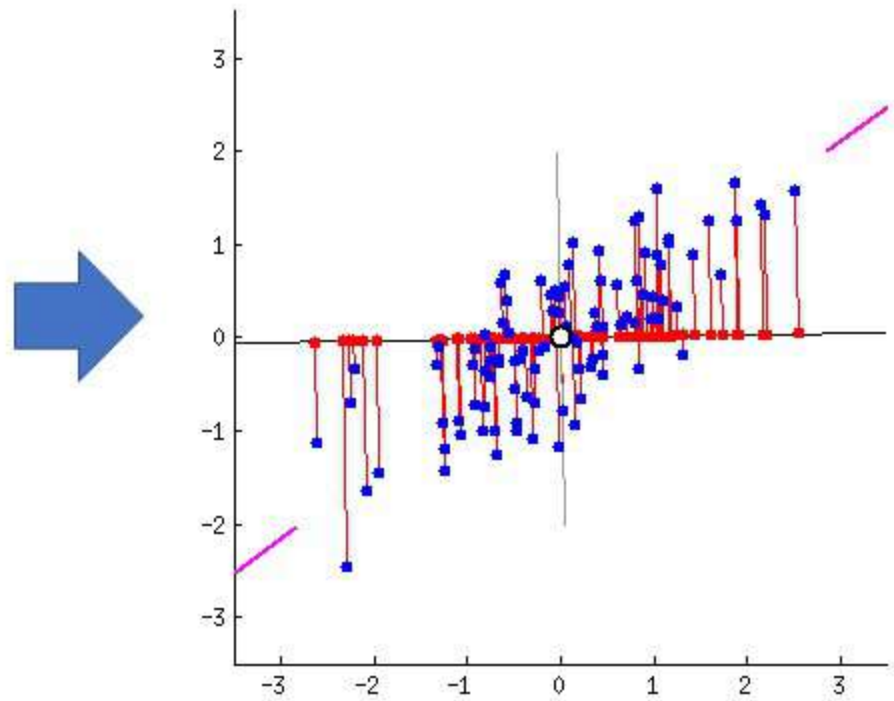
Component Loading matrix

Interpretation and labelling

Loading Plot

PCA Transformation

Income	Education	Age	Residence	Employ	Savings	Debt	Credit cards
50000	16	28	2	2	5000	1200	2
72000	18	35	10	8	12000	5400	4
61000	18	36	6	5	15000	1000	2
88000	20	35	4	4	980	1100	4
91100	18	38	8	9	20000	0	1
45100	14	41	15	14	3900	22000	4
36200	14	29	6	5	100	7000	5
41000	12	34	9	8	5000	200	3
40000	16	32	8	7	19000	1760	2
32000	16	30	2	2	16000	550	1
29000	16	28	1	4	2100	4600	2
21240	12	26	2	2	100	10010	3
58700	12	38	9	9	4500	7800	5
41000	14	29	5	4	300	10000	6
38720	16	36	11	11	24500	540	2
88240	16	38	13	12	13600	8100	2
40000	18	39	7	6	16000	1300	2



Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Income	0.60	-0.21	0.70	-0.26	0.16	-0.20	-0.01	-0.01
Education	0.45	-0.66	0.42	0.18	-0.41	0.15	-0.04	0.01
Age	0.93	0.20	0.00	-0.16	0.11	0.20	0.24	-0.01
Residence	0.89	0.41	-0.09	0.09	0.02	0.04	-0.18	-0.14
Employ	0.88	0.45	-0.13	-0.01	0.01	0.01	-0.13	0.15
Savings	0.77	-0.33	-0.38	0.32	-0.09	-0.24	0.13	-0.00
Debt	-0.13	0.87	0.08	-0.21	-0.44	-0.10	0.07	-0.02
Credit cards	-0.24	0.67	0.49	0.52	0.13	0.01	0.06	0.01

Component Loading Matrix

Loading Matrix:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Income	0.5912	0.2112	0.6908	0.2529	-0.1548	0.0061	0.2015	0.0063
Education	0.4463	0.6484	0.4097	-0.1749	0.3990	-0.0116	-0.1456	0.0365
Age	0.9115	-0.1973	0.0042	0.1542	-0.1122	0.0105	-0.1988	-0.2327
Residence	0.8784	-0.4039	-0.0927	-0.0844	-0.0226	0.1343	-0.0348	0.1725
Employ	0.8644	-0.4445	-0.1244	0.0124	-0.0092	-0.1498	-0.0094	0.1302
Savings	0.7611	0.3197	-0.3746	-0.3177	0.0918	0.0035	0.2318	-0.1232
Debt	-0.1266	-0.8543	0.0797	0.2048	0.4369	0.0151	0.1001	-0.0693
Credit cards	-0.2320	-0.6598	0.4784	-0.5129	-0.1250	-0.0117	-0.0089	-0.0559

- Component ***loadings*** are the ***correlation*** coefficients between the variables (rows) and components (columns). It is analogous to *Pearson's r*, and the squared *loading* is the percent of variance in that variable explained by the component.
- Loadings are typically used for the interpretation and labelling of the derived components because they show the correlations between the original fields (rows), and the derived components (columns).

Interpretation and labelling

- The interpretation process involves the examination of the loading values and their signs and identification of significant correlations
- The goal is to understand the **information that they convey** and **name** them accordingly
- Typically correlations above 0.4 (sometimes subjective) in absolute value are considered to be a practical significance
- The interpretation process ends with the labelling of the derived components with names that appropriately summarize their meaning
- Sometimes you will find data scientists applying another Rotation (e.g. varimax) to the matrix in order to make it easier to interpret the loadings.

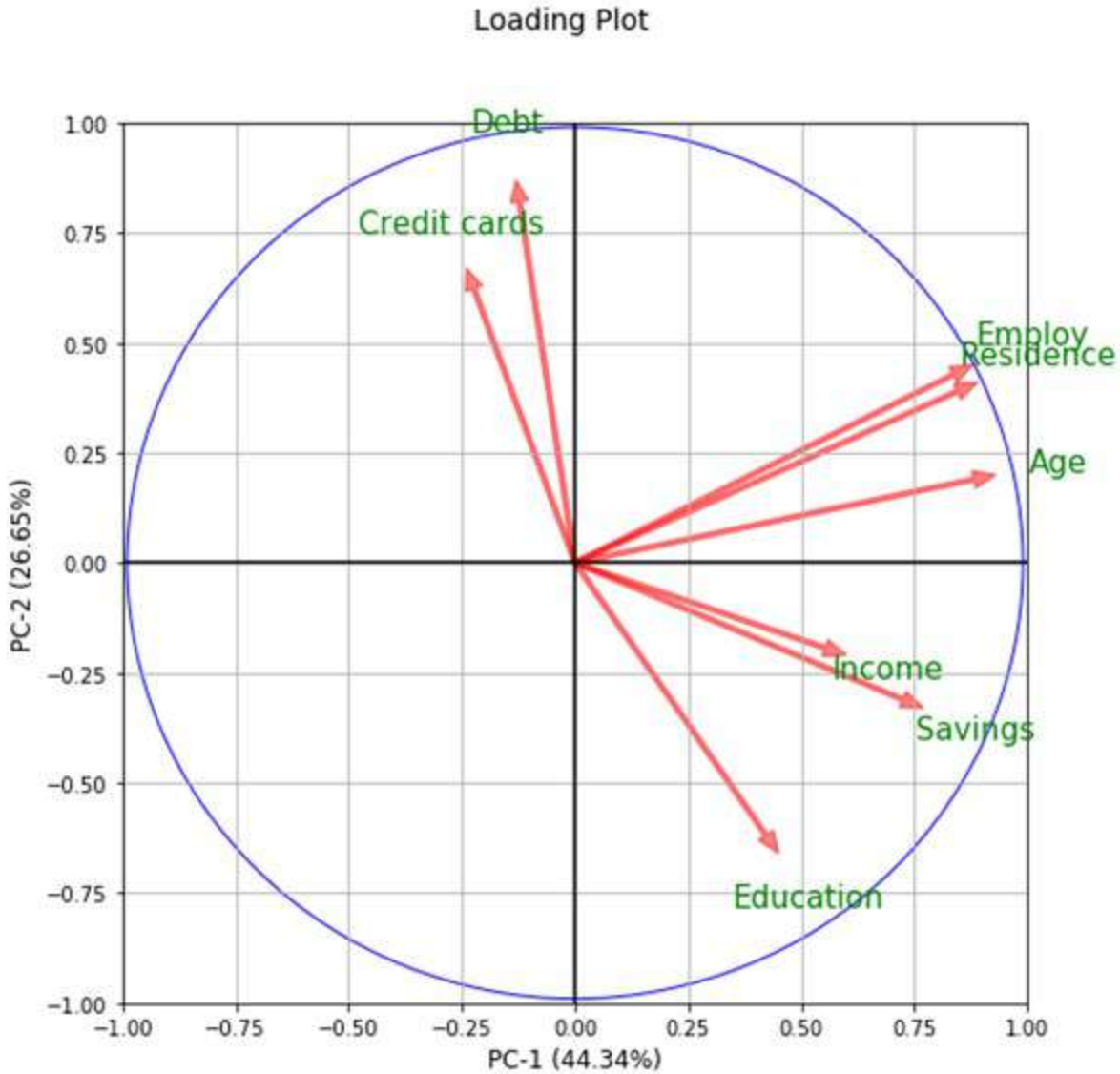
Component Loading Matrix - Interpretation

Loading Matrix:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Income	0.5912	0.2112	0.6908	0.2529	-0.1548	0.0061	0.2015	0.0063
Education	0.4463	0.6484	0.4097	-0.1749	0.3990	-0.0116	-0.1456	0.0365
Age	0.9115	-0.1973	0.0042	0.1542	-0.1122	0.0105	-0.1988	-0.2327
Residence	0.8784	-0.4039	-0.0927	-0.0844	-0.0226	0.1343	-0.0348	0.1725
Employ	0.8644	-0.4445	-0.1244	0.0124	-0.0092	-0.1498	-0.0094	0.1302
Savings	0.7611	0.3197	-0.3746	-0.3177	0.0918	0.0035	0.2318	-0.1232
Debt	-0.1266	-0.8543	0.0797	0.2048	0.4369	0.0151	0.1001	-0.0693
Credit cards	-0.2320	-0.6598	0.4784	-0.5129	-0.1250	-0.0117	-0.0089	-0.0559

- Look at the loadings and determine the meaning of the derived components with respect to the original fields.
- The goal is to understand the information that they convey and name them accordingly.

Loading Plot



- The Loading Plot graphs the **loadings** of Principal components.
- Loading Plots allow you to identify which variables have the **largest effect** on each component.
- Loadings close to -1 or 1 indicate that the variable **strongly influences** the component.
- Loadings close to 0 indicate that the variable has a **weak influence** on the component.
- Evaluating the loadings can also help you **characterize each component** in terms of the variables.

3.How Many Components to retain?

Eigenvalues (Kaiser Rule)

Total variance explained

Scree Plot

Communalities Matrix

What are eigenvalues?

- **Eigenvalues** represent the total amount of variance that can be explained by a given principal component. The **variance** can be considered as a measure of the field's information. A standardized field has a standard deviation and a variance value of 1 – hence it carries one unit of information.

Loading Matrix:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Income	0.5912	0.2112	0.6908	0.2529	-0.1548	0.0061	0.2015	0.0063
Education	0.4463	0.6484	0.4097	-0.1749	0.3990	-0.0116	-0.1456	0.0365
Age	0.9115	-0.1973	0.0042	0.1542	-0.1122	0.0105	-0.1988	-0.2327
Residence	0.8784	-0.4039	-0.0927	-0.0844	-0.0226	0.1343	-0.0348	0.1725
Employ	0.8644	-0.4445	-0.1244	0.0124	-0.0092	-0.1498	-0.0094	0.1302
Savings	0.7611	0.3197	-0.3746	-0.3177	0.0918	0.0035	0.2318	-0.1232
Debt	-0.1266	-0.8543	0.0797	0.2048	0.4369	0.0151	0.1001	-0.0693
Credit cards	-0.2320	-0.6598	0.4784	-0.5129	-0.1250	-0.0117	-0.0089	-0.0559

```
Eigenvalues                          : [3.548 2.132 1.045 0.532 0.411 0.166]
Eigenvalues (%Explained_Variance): [44.34 26.65 13.06  6.64  5.14  2.08]
Eigenvalues (%Cumulative)          : [44.34 70.99 84.05 90.69 95.83 97.91]
```

Eigenvalues are **the sum of squared** component loadings across all items.

The Kaiser Rule

- The **Kaiser Rule**- is perhaps the most widely used criterion for selecting which components to keep. It is based on the idea that a component should be considered insignificant if it does worse than a single field.
- Each single field contains one unit of standardized variance, thus components with eigenvalues below 1 are not extracted.

```
                PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
Eigenvalues   : [3.67  2.206  1.081  0.55   0.425  0.172  0.13   0.043]
```

How many components should be retained?

Total Variance Explained

Eigenvalues can also be expressed in terms of a **percentage of the total variance** of the original fields. The second row of the table denotes the proportion of the variance attributable to each component, and the next row denotes the proportion of the variance jointly explained by all components up to that point.

The first 4 components explains 91% of the variability with 8% loss of information.

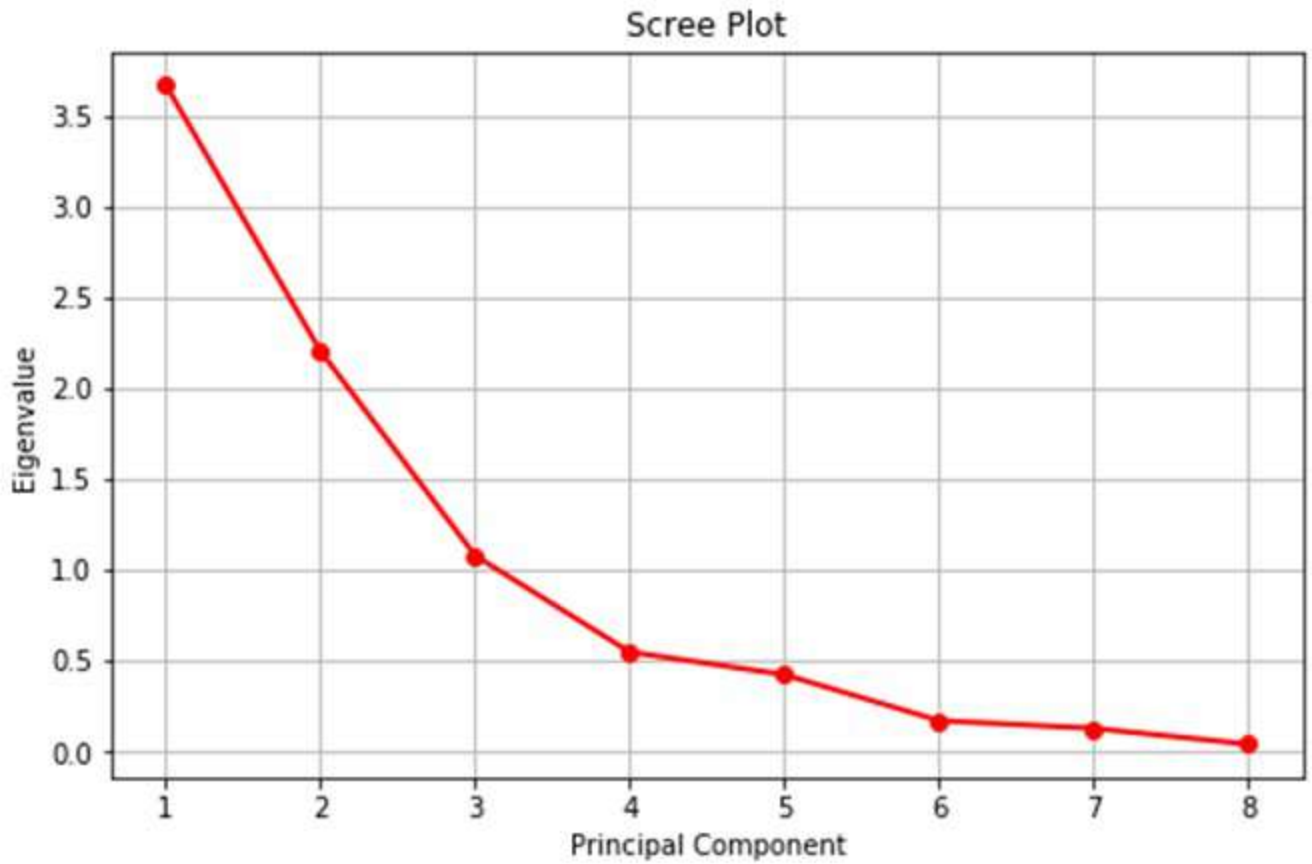
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Eigenvalues	3.67	2.206	1.081	0.55	0.425	0.172	0.13	0.043
% Variance Explained	44.34	26.65	13.06	6.64	5.14	2.08	1.57	0.51
% Cumulative)	44.34	70.99	84.05	90.69	95.83	97.91	99.48	99.99

How many components should be retained?

Scree Plot

The Scree Plot shows the eigenvalues vs the components in decreasing order:

```
Eigenvalues          : [3.67  2.206 1.081 0.55  0.425 0.172 0.13  0.043]
% Variance Explained: [44.34 26.65 13.06  6.64  5.14  2.08  1.57  0.51]
% Cumulative)        : [44.34 70.99 84.05 90.69 95.83 97.91 99.48 99.99]
```



Scree Plot test

Look for a large drop, followed by a "plateau" in the eigenvalues – i.e. a bend or elbow. This indicates a transition from the large to smaller values. Beyond the elbow point, the variances explained tapers off.

If the scree is "not ideal" – i.e. you don't see any clear bends, then fall back on previous 2 criteria.

Communalities

	2 PCs	3 PCs
Income	0.4041	0.8941
Education	0.6381	0.8145
Age	0.9049	0.9049
Residence	0.9602	0.9683
Employ	0.9769	0.9938
Savings	0.7018	0.8462
Debt	0.7738	0.7802
Credit cards	0.5065	0.7466

Does the solution account for all the original variables?

- Communality represents the total amount of variance of a specific field that is jointly accounted for by all the components. It is calculated as the sum of squared loadings of the field across all components.
- **High communality** values indicate that the original field is **sufficiently explained** by the reduced PCA solution.
- **Low communality** implies an **insignificant** contribution to the formation of the PCA Solution
- communalities > 0.5 are desired

5.Proceeding to the next step

Eigenvectors

Eigenvectors (Linear Coefficients):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Income	0.314	-0.145	0.676	-0.347	0.241	-0.494	-0.018	-0.030
Education	0.237	-0.444	0.401	0.240	-0.622	0.357	-0.103	0.057
Age	0.484	0.135	0.004	-0.212	0.175	0.487	0.657	-0.052
Residence	0.466	0.277	-0.091	0.116	0.035	0.085	-0.487	-0.662
Employ	0.459	0.304	-0.122	-0.017	0.014	0.023	-0.368	0.739
Savings	0.404	-0.219	-0.366	0.436	-0.143	-0.568	0.348	-0.017
Debt	-0.067	0.585	0.078	-0.281	-0.681	-0.245	0.196	-0.075
Credit cards	-0.123	0.452	0.468	0.703	0.195	0.022	0.158	0.058

Eigenvectors represent the coefficients or weights for each variable in the component.
The **principal components** are the linear combinations of the original variables that account for the variance in the data.
The **eigenvectors**, which are comprised of **coefficients** corresponding to each variable, are used to calculate the principal **component scores**.
The coefficients indicate the relative weight of each variable in the component.

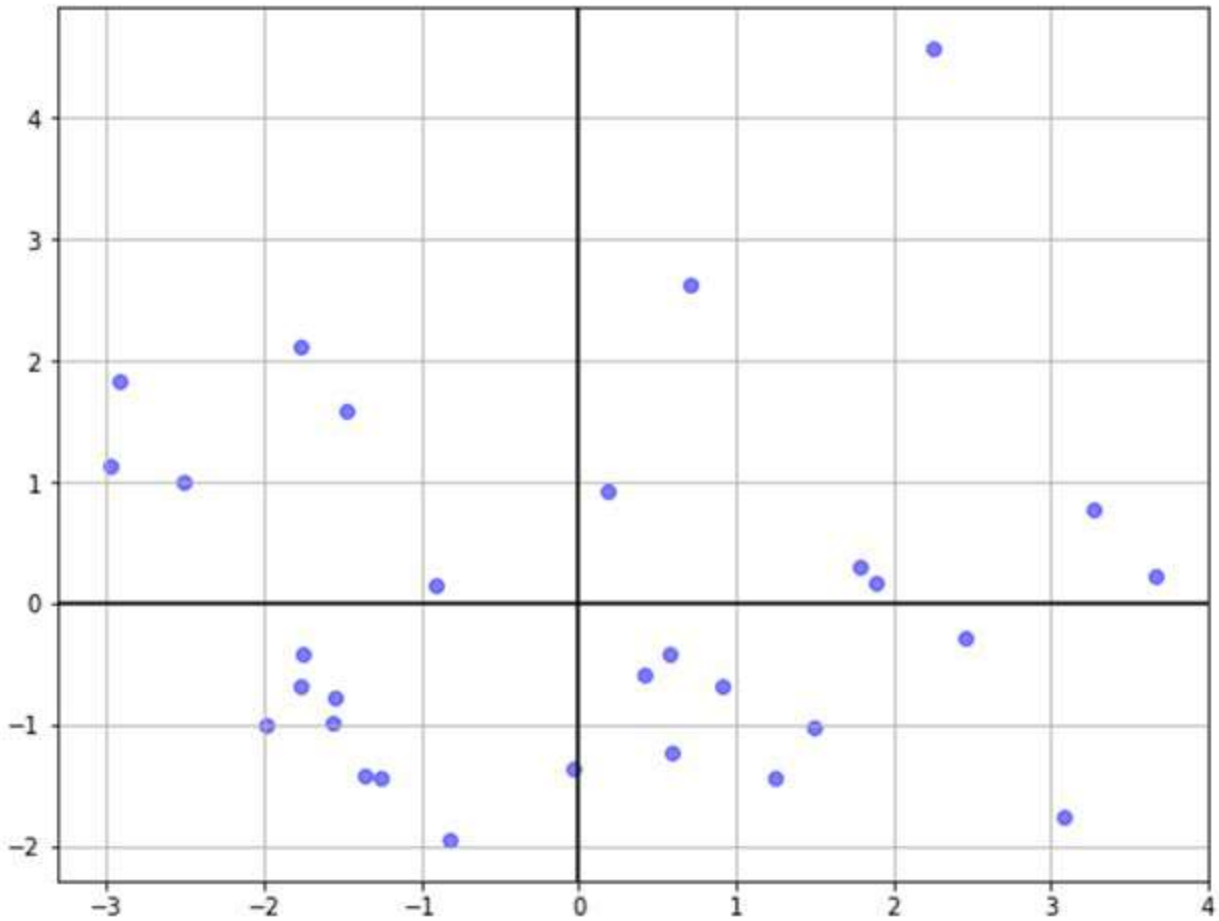
```
PC1 = 0.314*Income + 0.237*Education + 0.484*Age +
      0.466*Residence + 0.459*Employ + 0.404*Savings +
      -0.67*Debt + -0.123*Credit cards
```

Component Scores

	PC1	PC2	PC3
0	-1.36	-1.42	0.17
1	1.79	0.31	1.51
2	1.25	-1.44	0.33
3	0.60	-1.23	3.17
4	3.09	-1.76	0.66
5	2.26	4.58	0.05
6	-1.47	1.59	0.53
7	0.19	0.93	-0.92
8	0.91	-0.68	-1.11
9	-0.82	-1.95	-1.37
10	-1.77	-0.68	-0.47
11	-2.97	1.13	-0.93
12	0.72	2.63	0.60
13	-1.76	2.11	1.18
14	2.46	-0.28	-1.62
15	3.27	0.78	0.74
16	1.50	-1.02	-0.54
17	3.67	0.23	-1.91
18	-2.91	1.84	0.14
19	-1.26	-1.45	0.53
20	-1.75	-0.42	-0.13

- The PCA algorithm derives new composite fields, named **component scores**, that denote the values of each record in the revealed components.
- Component scores are produced through linear transformations of the original fields, by using coefficients that correspond to the loading values.
- They can be used as any other fields in subsequent tasks.
- Scores are standardized values (mean=0; SD=1)
- The scores represent the number of standard deviations above or below the overall mean where each customer lies.

Score Plot



- The score plot graphs the scores of the First principal component versus the scores of the Second principal component.
- If the first two components account for most of the variance in the data, you can use the score plot to assess the data structure and detect clusters, outliers, and trends.
- Groupings of data on the plot may indicate two or more separate distributions in the data.
- If the data follow a normal distribution and no outliers are present, the points are randomly distributed around zero.

Some Assumptions for PCA

When you choose to analyse your data using PCA, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using PCA. You need to do this because it is only appropriate to use PCA if your data "passes" four assumptions that are required for PCA to give you a valid result.

- **Sampling adequacy**: ideally, there should be 150+ cases and there should be ratio of at least five cases for each variable (Pallant, 2010)
- **Suitability (Correlations exist):** there should be some **linear** correlation among the factors to be considered for PCA
- **Continuous variables:** variables should be interval or ratio.
- **Outliers:** PCA is sensitive to outliers; disproportionate influence on your results. Remove data that are > 3 standard deviations

What is meant by data loss?

- The goal of PCA is to extract the smallest number of components which **account for as much as possible of the information** of the original features.
- Hence, there will be some amount of information that will be lost.
- To illustrate information lost, we will reconstruct the original data using only the saved principal components.

```
[[ 19.   4.  90. 150.]
 [ 43.  12.  30.  35.]
 [ 13.   3.  10.  20.]
 [ 60.  14. 100.  80.]
 [  5.   1.  30.  55.]
 [ 56.  11.  25.  35.]
 [ 25.   7.  30.  28.]
 [  3.   1.  65.  82.]
 [ 40.   9.  15.  30.]
 [ 65.  15.  20.  40.]]
```

Original Data

```
[[ 23.   5. 106. 138.]
 [ 35.   8.  27.  36.]
 [ 37.   9.  14.  20.]
 [ 28.   6.  73.  97.]
 [ 33.   8.  39.  52.]
 [ 36.   8.  24.  33.]
 [ 35.   8.  25.  34.]
 [ 29.   7.  65.  86.]
 [ 36.   9.  19.  26.]
 [ 36.   8.  24.  33.]]
```

1 PC (55.098%)

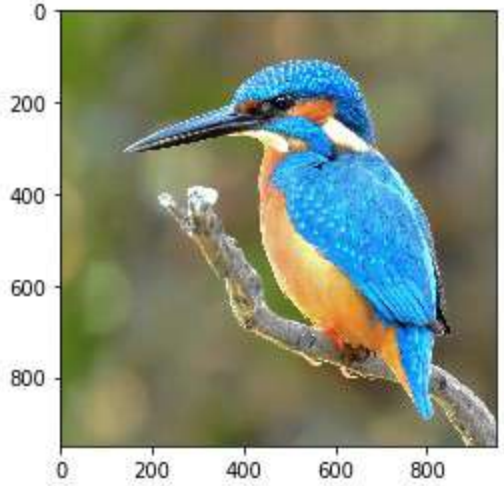
```
[[ 14.   3. 103. 139.]
 [ 44.  10.  29.  36.]
 [ 14.   4.   7.  22.]
 [ 66.  15.  85.  92.]
 [  5.   1.  30.  55.]
 [ 54.  12.  30.  31.]
 [ 28.   7.  22.  34.]
 [  6.   1.  58.  88.]
 [ 38.   9.  20.  26.]
 [ 61.  14.  32.  30.]]
```

2 PCs (96.564%)

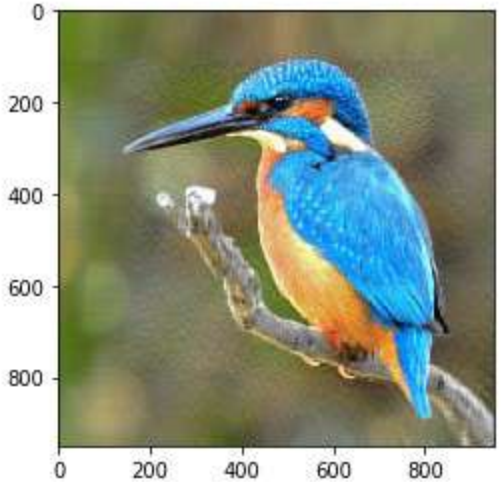
```
[[ 19.   3.  90. 149.]
 [ 43.  10.  30.  34.]
 [ 12.   3.   9.  20.]
 [ 59.  14.  99.  80.]
 [  4.   1.  29.  55.]
 [ 55.  12.  24.  35.]
 [ 25.   6.  30.  27.]
 [  2.   1.  64.  82.]
 [ 39.   9.  14.  30.]
 [ 65.  14.  20.  39.]]
```

3 PCs (96.985%)

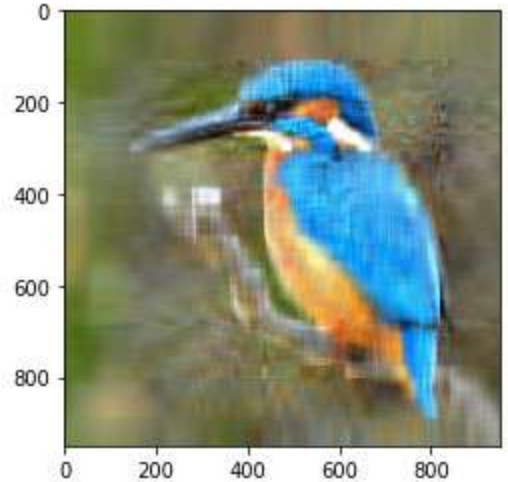
Data loss in Image Compression



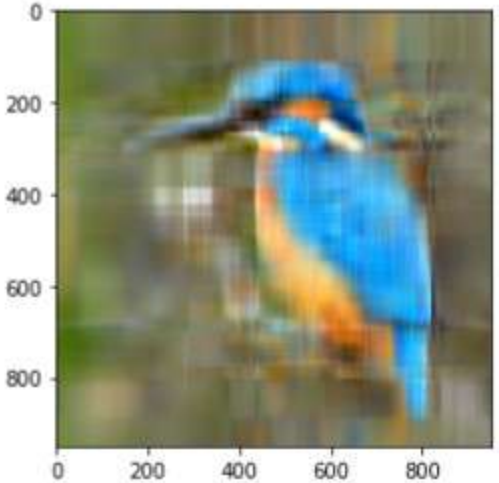
100%, 3800



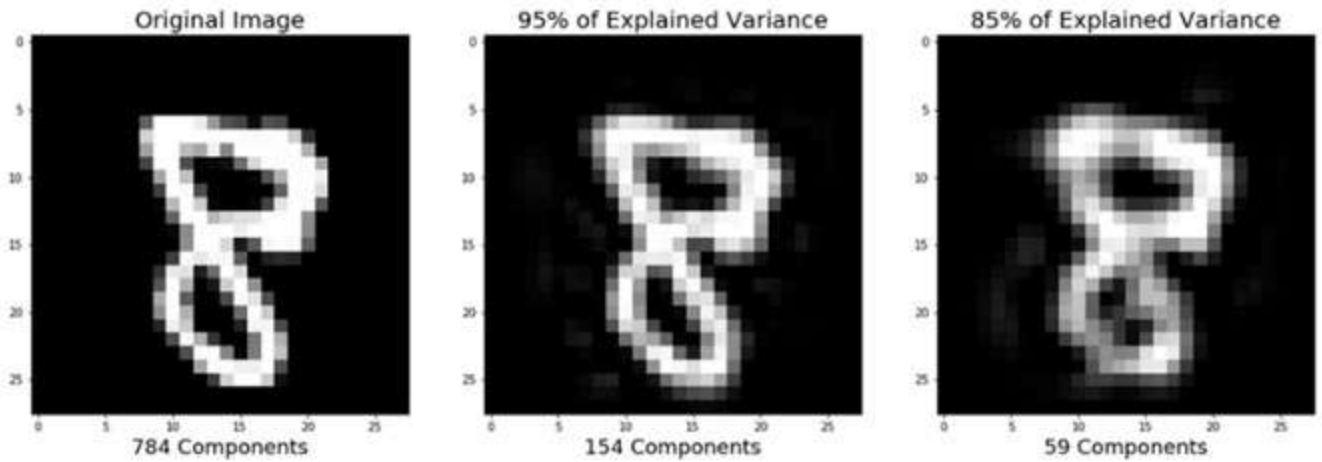
95%, 60 PC



85%, 18 PC



75%, 9 PC



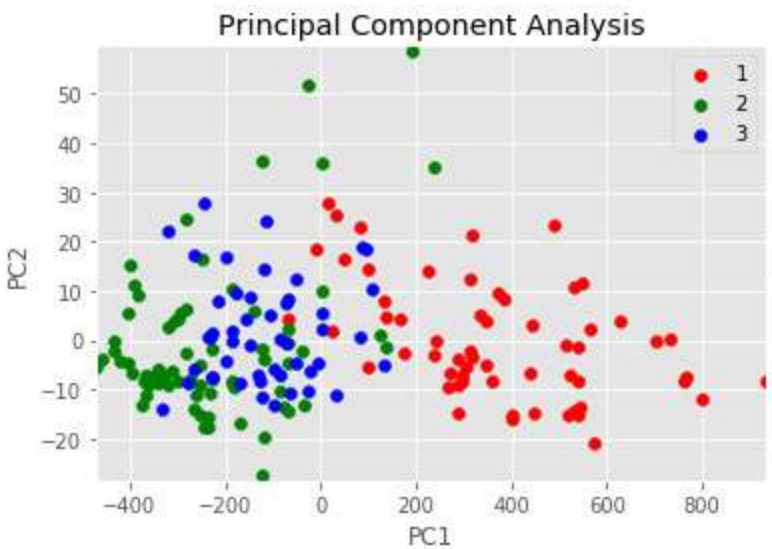
Principal Component Analysis Using Python

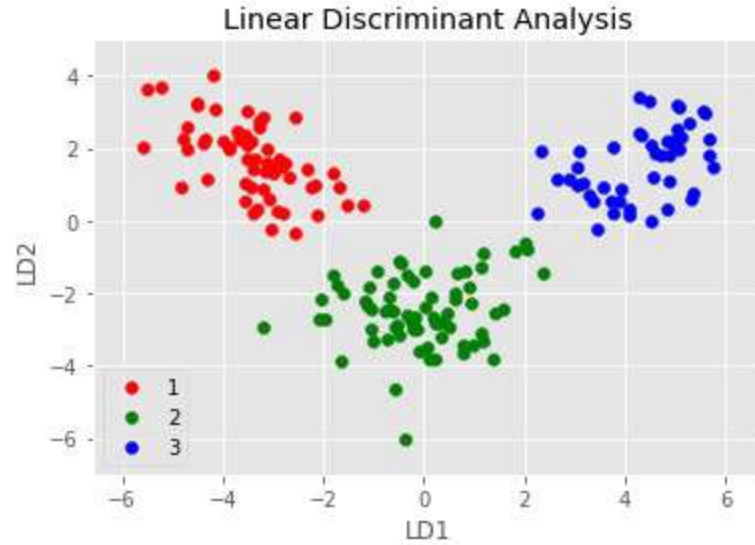
Since we are not doing PCA from scratch, the process of performing it is as follows:-

1. Read in the dataset
2. Basic Data Understanding and Preparation
3. Standardizing the data
4. Call the PCA algorithm
5. Plot the results
6. Decide on the number of components to retain
7. Proceeding to the next step

LINEAR DISCRIMINANT ANALYSIS (LDA)

What if you had access to class labels?





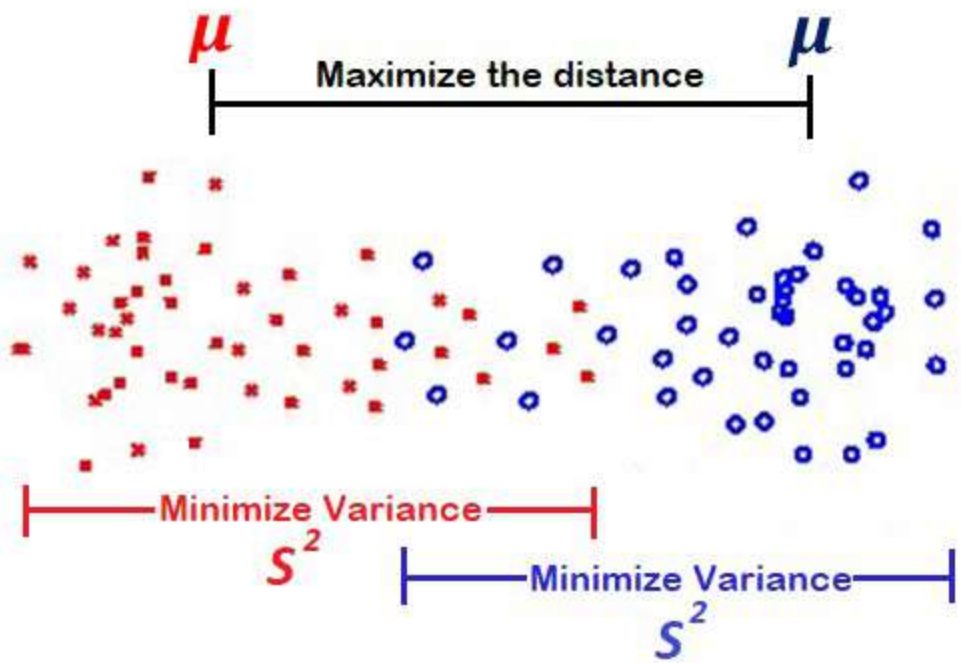
We not only reduce the number of features BUT at the same time create class separability.
This is most useful for multi-variate classification-type problems.

Linear Discriminant Analysis

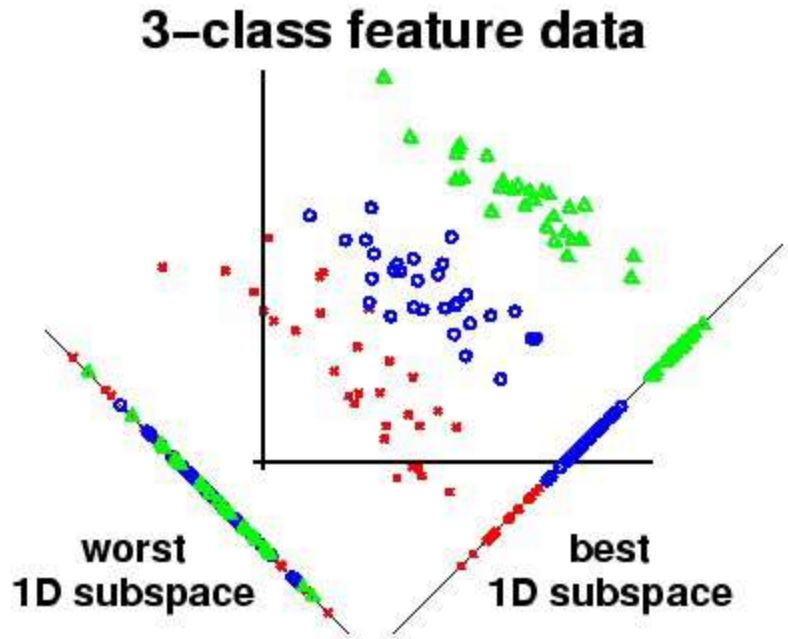
- Originally developed in 1936 by R.A. Fisher, Discriminant Analysis is a classic method of classification.
- There are lots of similarities between LDA and PCA.
- Whereas PCA yields the directions (principal components) that maximize the variance of the data, LDA also aims to find the directions that **maximize the separation** (or discrimination) between different classes, which can be useful in pattern classification problem.
- LDA takes the **class label** into consideration whereas PCA ignores them.
- Since it uses the class information, LDA is a "**supervised learning**" method while PCA is an unsupervised learning method.

The LDA Method

- LDA finds a new dimension that yields
  - Maximum separation between the class means
  - Minimum variance within class



$$\frac{(\mu - \mu)}{s^2 + s^2}$$

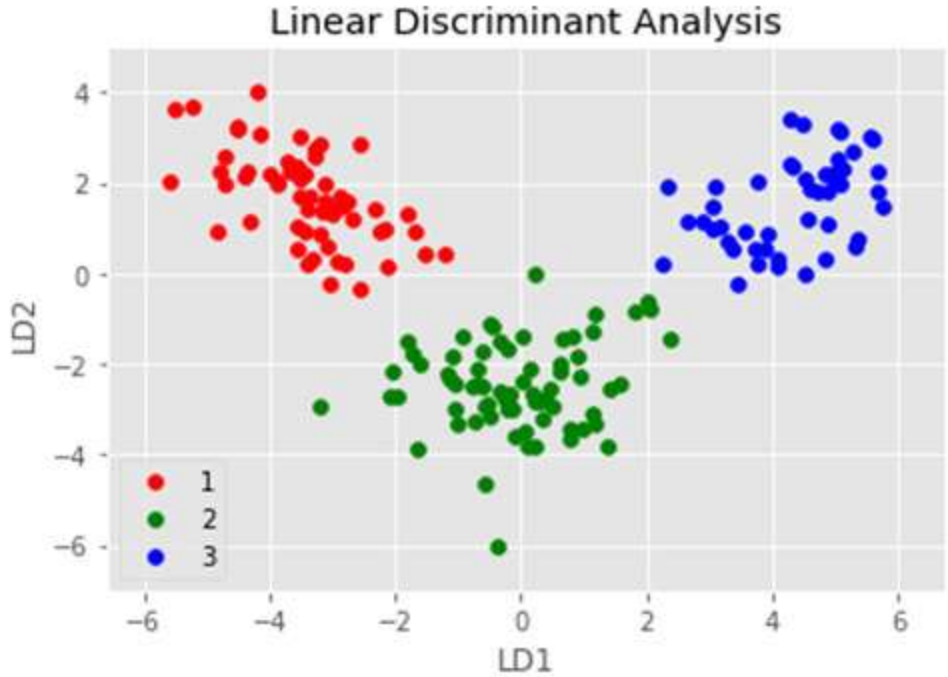


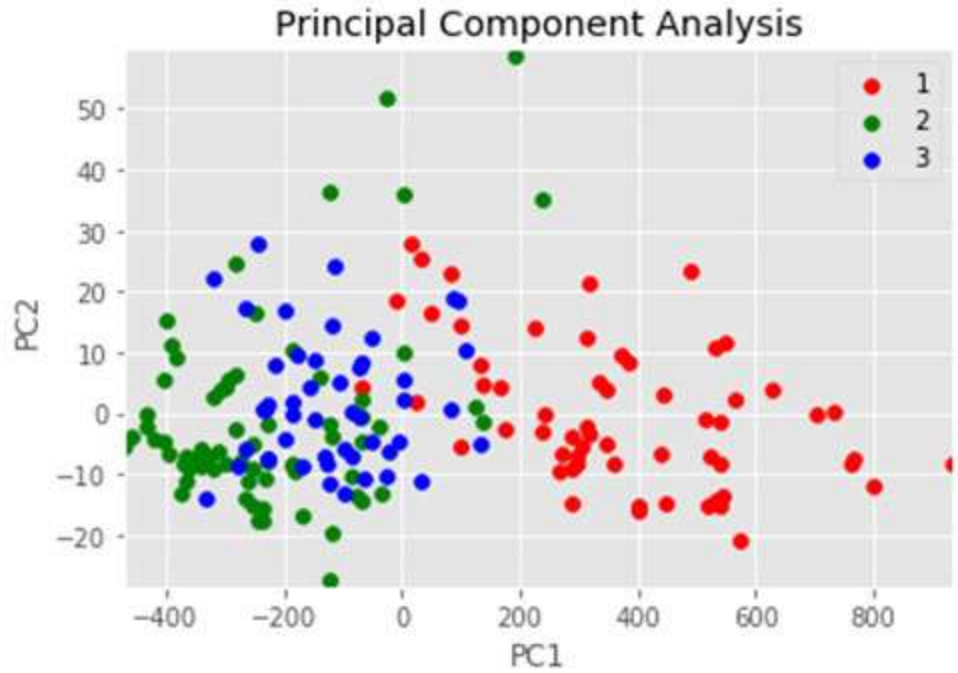
Example

	alcohol	malic_acid	ash	ty_of_ash	agnesium	l_phenols	flavanoids	d_phenols	hocyanins	_intensity	hue	ted_wines	proline
0	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
2	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
3	14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480
4	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
5	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
6	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
7	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295
8	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045
9	13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045
10	14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510
11	14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280
12	13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320
13	14.75	1.73	2.39	11.4	91	3.1	3.69	0.43	2.81	5.4	1.25	2.73	1150
14	14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	7.5	1.2	3	1547
15	13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310
16	14.3	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280
17	13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	1.13	2.57	1130

This data set has 178 examples from three classes. Each example consists of 13 real-valued features.

Results of LDA vs PCA






```
LDA Explained Variance [0.6875 0.3125]
PCA Exlained Variance: [0.9981 0.0017]
```

Demo, Exercises and Workshop

Thank You!