

# Data Visualization



**Dr. José Ramón Iglesias**

DSP-ASIC BUILDER GROUP  
Director Semillero TRIAC  
Ingeniería Electronica  
Universidad Popular del Cesar

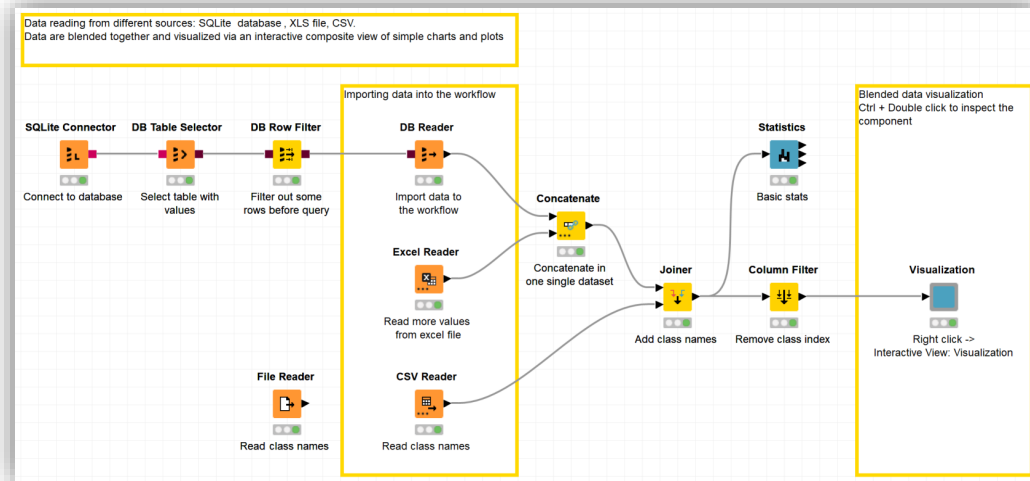
*„There is no excuse for failing to plot and look“*

What is the best way of plotting a dataset?

- Methods for One and Two Attributes
  - Barchart and Histogram
  - Boxplot
  - Scatter plot and density plot
- Methods for Higher-dimensional Data
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
  - Parallel Coordinates
  - Radar and Star Plots
  - Sunburst Chart
  - Correlation Analysis

# Datasets

- Datasets used : adult dataset and outliers dataset
- Example Workflows:
  - „Simple Visualizations“ <https://kni.me/w/dwugN1qYM2OOjzO4>
    - Read from CSV file, Excel file and SQLite.
    - bar chart and histogram
    - parallel coordinates
    - box plot
    - scatter plot
    - table view.



# Statistical Descriptors

Statistical measures can be used to describe a dataset:

- Range
- Min/max values
- Mean

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

- Variance

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2$$

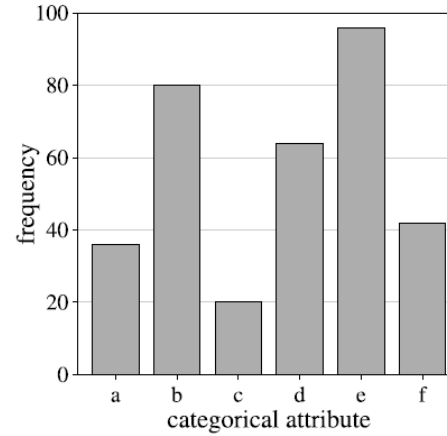
- Standard deviation

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2}$$

- **Median** (The middle number; found by ordering all data points and picking out the one in the middle - or if there are two middle numbers, taking the mean of those two numbers)
- **Mode** (Most frequently occurring value)
- **Percentiles (Quartiles)**
- **Number of missing values**
- ...

# Visualization Methods for One Attribute

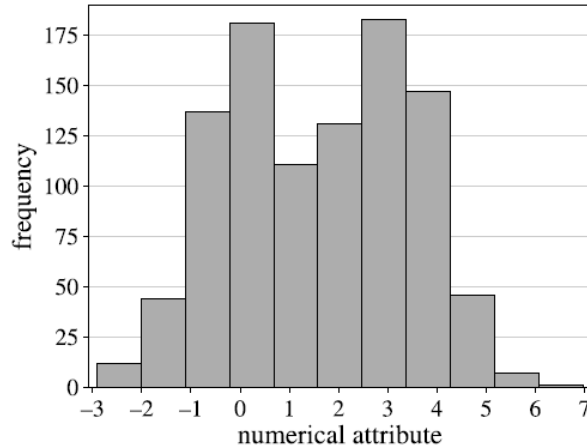
## Bar chart



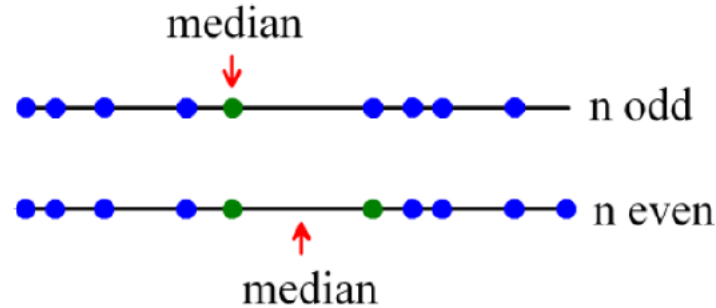
- A bar chart is a simple way to depict the frequencies of the values of a categorical attribute.



# Histogram

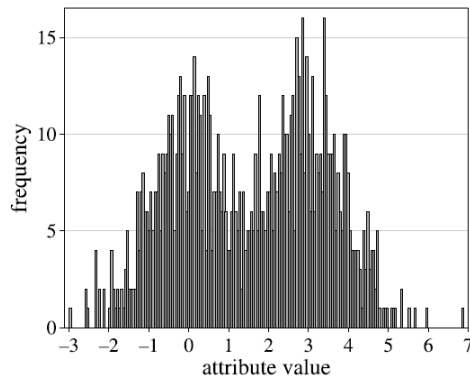
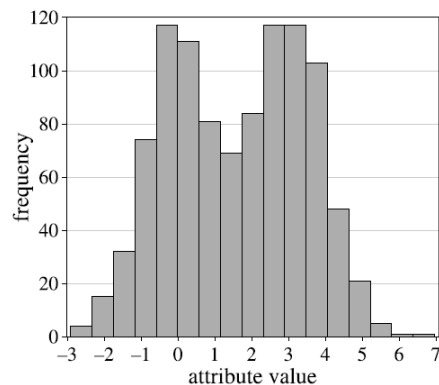


- A histogram shows the frequency distribution for a numerical attribute.
- The range of the numerical attribute is discretized into a fixed number of intervals (bins), usually of equal length.
- For each interval, the (absolute) frequency of values falling into it is indicated by the height of a bar.



- **Median:** The value in the middle (for values sorted in increasing order)
- **$q\%$ -quantile** ( $0 < q < 100$ ): The value for which  $q\%$  of the values are smaller and  $100-q\%$  are larger. The median is the 50%-quantile
- **Quartiles:** 25%-quantile (1st quartile), median (2nd quartile), 75%-quantile (3rd quartile)
- **Interquartile range (IQR):** 3rd quartile – 1st quartile

## Choice of number of bins



### – Best choice for number $k$ of bins in the histogram?

– Sturge's Rule  $k = \lceil \log_2(n) + 1 \rceil$

– Through fixed bin length  $h$

$$k = \left\lceil \frac{\max\{x_i\} - \min\{x_i\}}{h} \right\rceil \quad \text{with } h = \frac{3.5 \cdot s}{n^{\frac{1}{3}}} \quad \text{or} \quad h = \frac{2 \cdot IQR(x)}{n^{\frac{1}{3}}}$$

Where  $s$  is the standard deviation of input feature  $x$ ,  $x_i$  its value in the  $i$ -th sample, and  $n$  the number of samples in the dataset.

- **Skewness** is the 3rd standardized moment of  $X$ , that is:

$$\tilde{\mu}_3 = E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] = \frac{E[(X - \mu)^3]}{E[(X - \mu)^2]^{3/2}} = \frac{\mu_3}{\sigma^3}$$

- Skewness measures the asymmetry of the probability distribution of  $X$

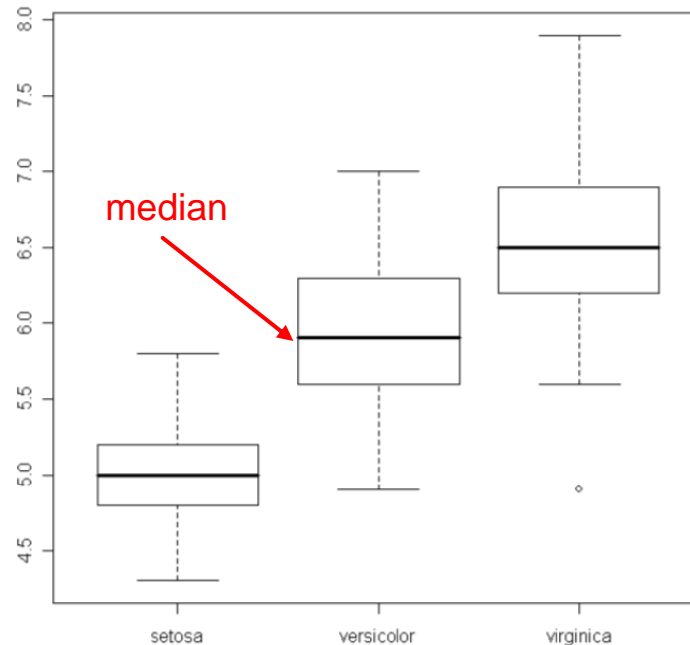
- **Kurtosis** is the 4th standardized moment of  $X$ , that is:

$$Kurt[X] = E \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = \frac{E[(X - \mu)^4]}{(E[(X - \mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

- Kurtosis measures the deviation from the peak in a Gaussian distribution: it measures the dispersion due to outliers
- Kurtosis of any univariate normal distribution is 3

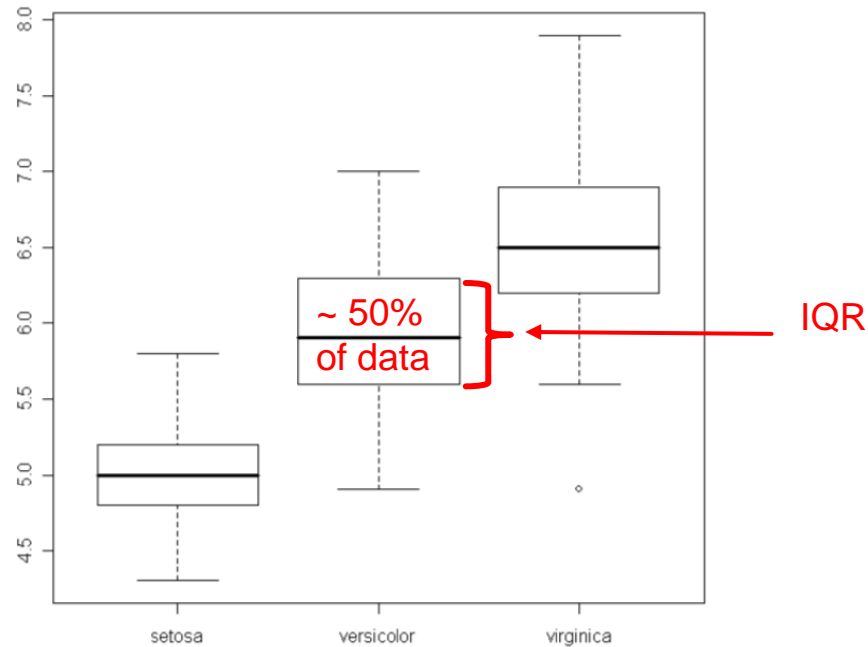
## Boxplots

- Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the ***median***, the IQR, and possible outliers



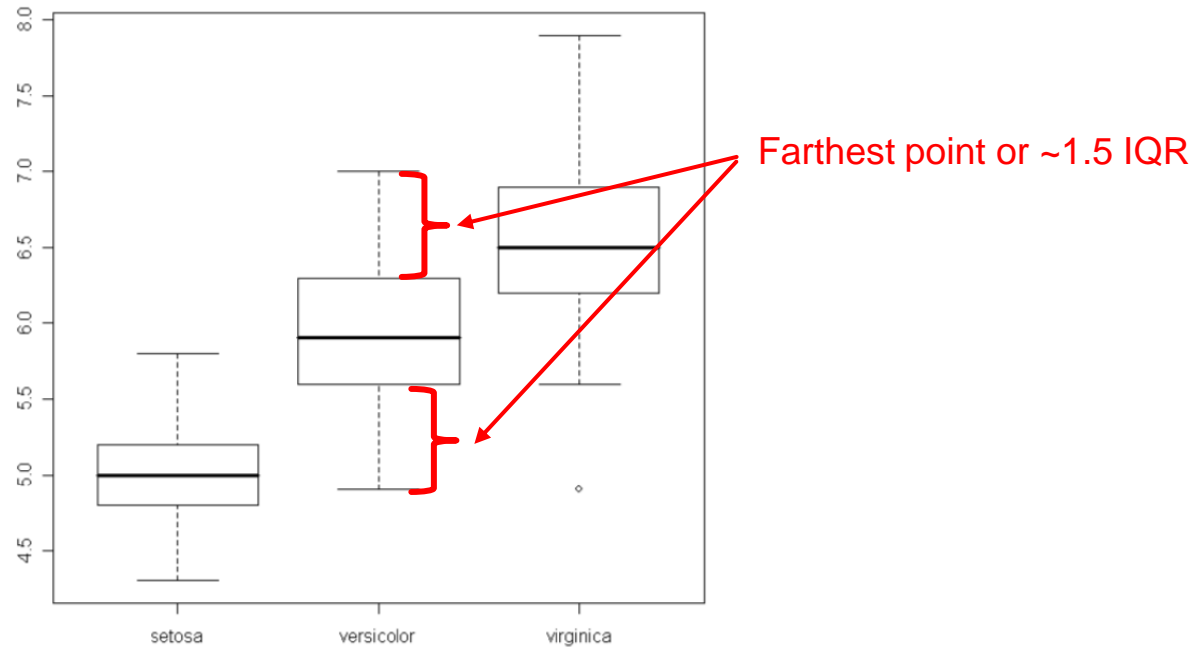
# Boxplots

- Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the ***IQR***, and possible outliers



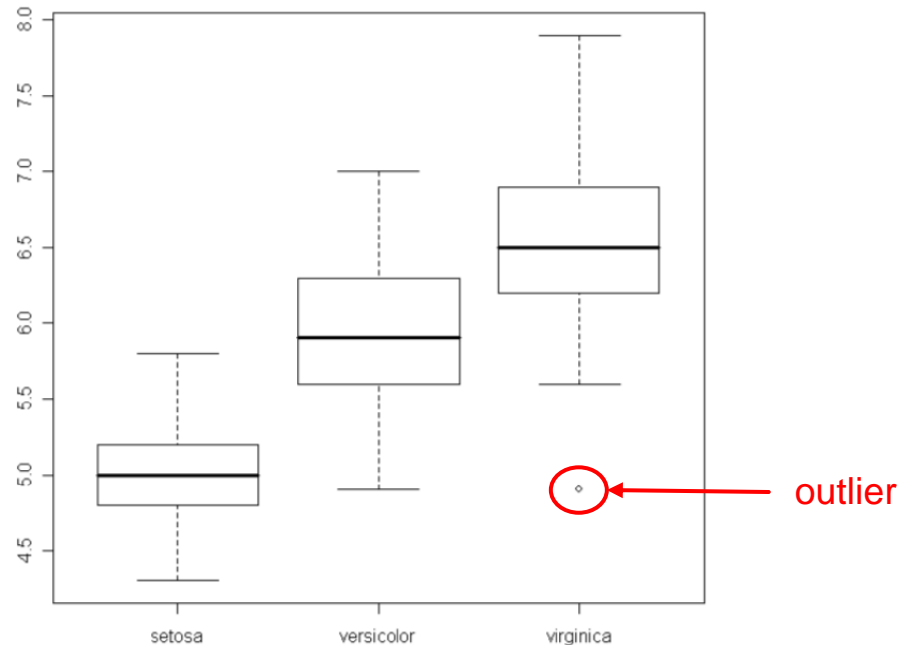
# Boxplots

- Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the **IQR**, and possible outliers



# Boxplots

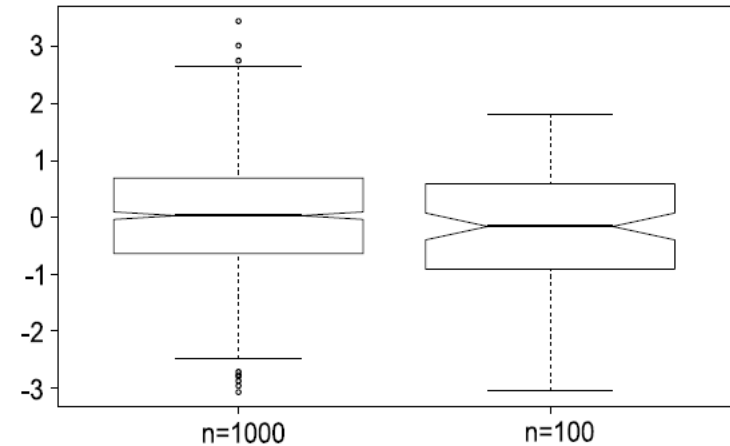
- Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the IQR, and possible ***outliers***





## Boxplots from normal distributions

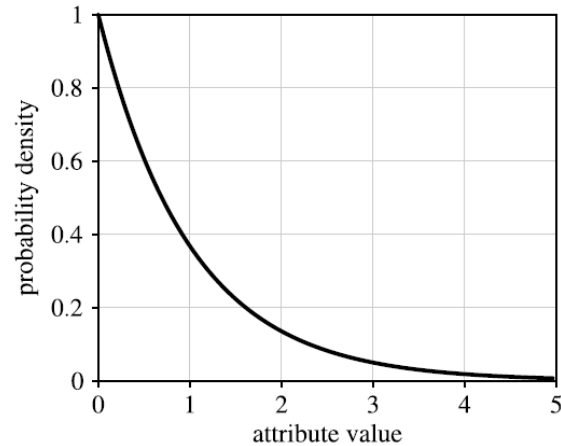
- The same distribution can result in different boxplots
- This depends on the sample size  $n$
- Two samples from normal distribution with different size  $n$
- For the small sample:
  - Whiskers have different length, even if it is the same symmetric distribution
  - No outliers



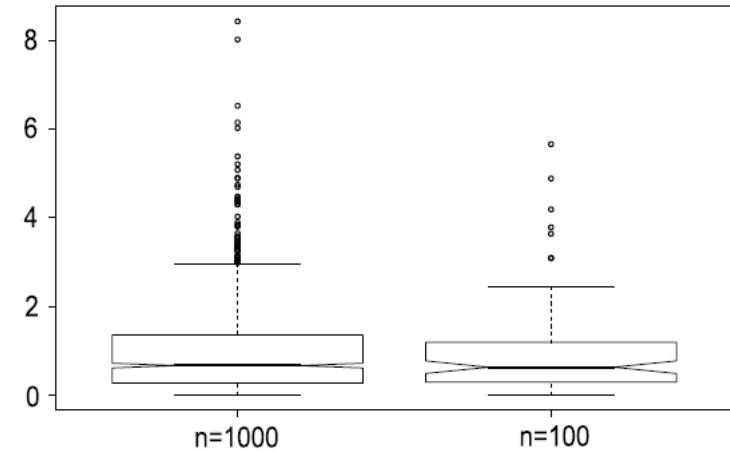
Boxplots from different samples from a standard normal distribution

## Boxplot of asymmetric distribution

- Boxplots of different samples from exponential distribution with  $\lambda = 1$



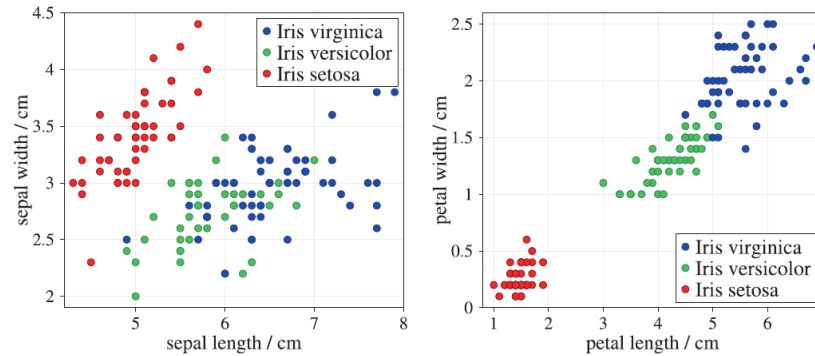
Exponential distribution with  $\lambda = 1$



Boxplots from different samples from exponential distribution with  $\lambda = 1$

# Visualization Methods for Two Attributes

# Scatter Plot

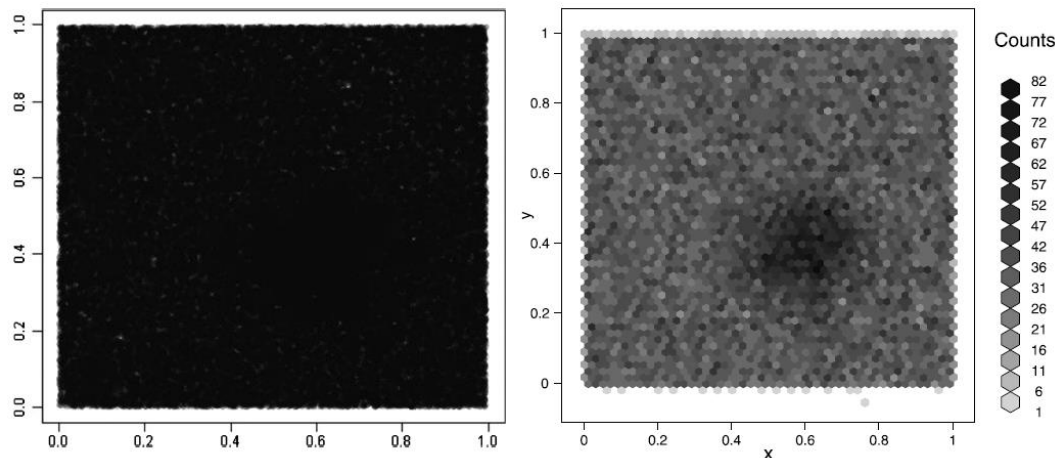


Petal length and width provide better class separation than sepal length and width

*Scatter plots of the Iris data set for sepal length vs. sepal width (left) and for petal length vs. petal width (right). All quantities are measured in centimetres*

- In scatter plots two attributes are plotted against each other
- Can be enriched with additional features (color, shape, size)
- Suitable for small number of points; not suitable for large datasets
- Points can hide each other -> add **Jitter** (a small random value to each point)

# Scatter Plot

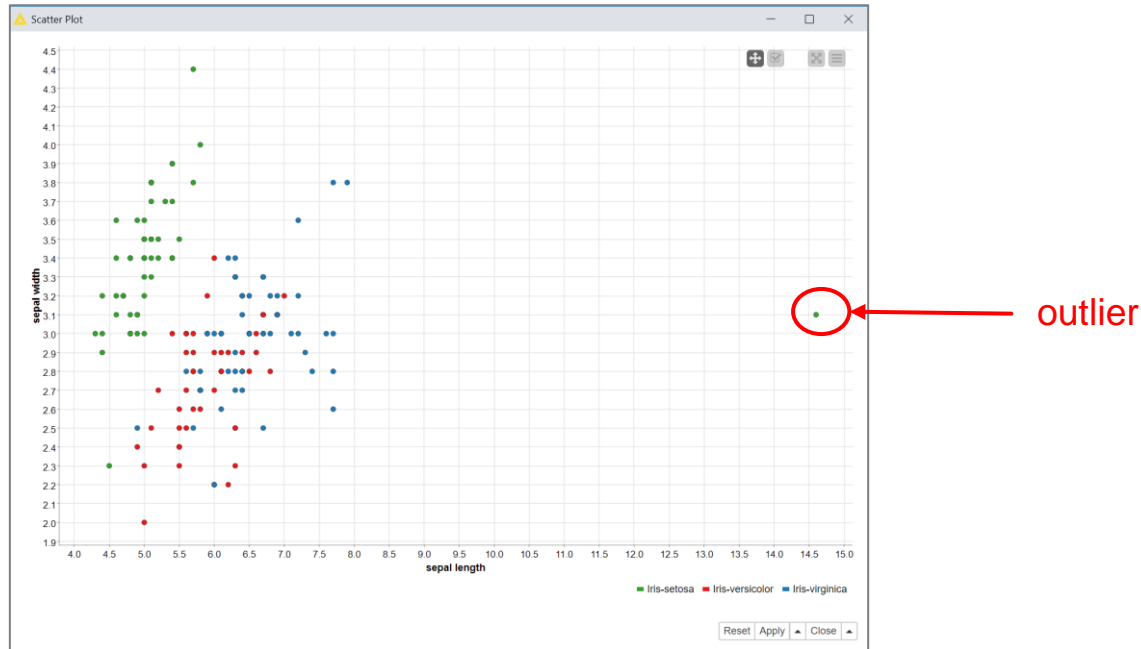


*Density plot (left) and a plot based on hexagonal binning (right) for a dataset with  $n = 100,000$  instances*

- Scatter plot is not suitable for large datasets
- Alternatives:
  - Density plot for example using semi-transparent points: the more points in the same place the less transparent
  - Binning points into rectangles or hexagons and heat scale color

# Scatter Plot to detect outliers

- Scatter plots can be used to detect outliers



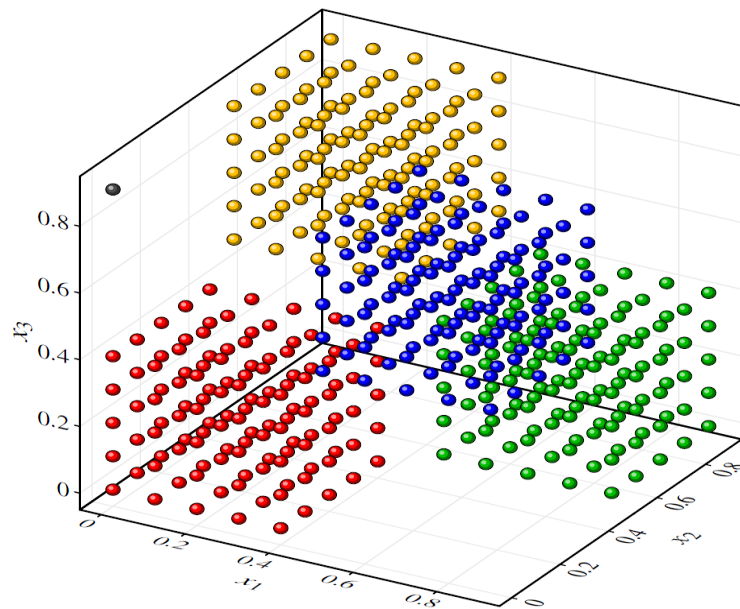
- Visualization can be used as a test
- **Good News**
  - Visualization reveals patterns or exceptions => there is something in the dataset
- **Bad News**
  - Visualization does not indicate anything specific => there might still be something in the dataset even if we do not see it
  - For example, if we do not see outliers for that combination of features, that does not mean that outliers do not exist in the dataset.

# Methods for Higher-Dimensional Data



A display or plot is **by definition two-dimensional**, so that only max. two axes (attributes) can be incorporated.

**3D** techniques can be used to incorporate three axes (attributes).



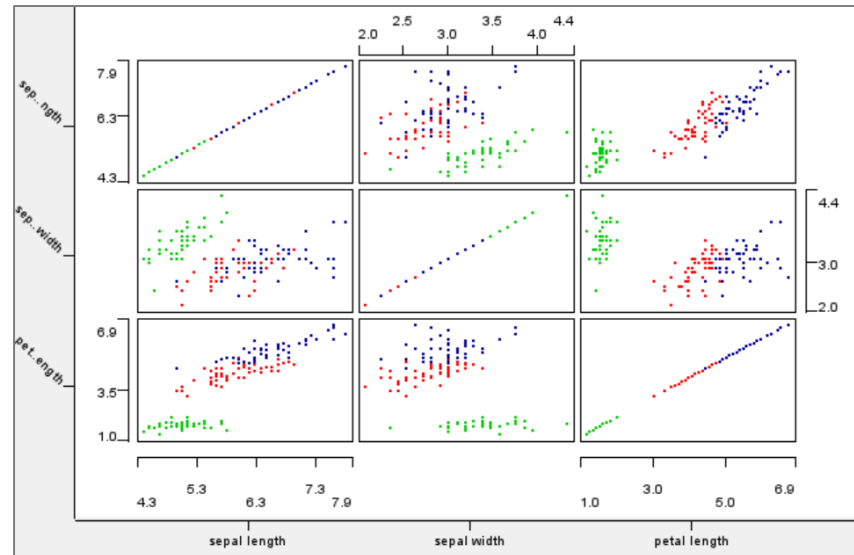
3D scatter plot

### Example

- A data set distributed over a cube in a **chessboard-like pattern**.
- The colors are only meant to make the different cubes more easily discernible. They do not indicate classes.
- Note the outlier in the upper left corner

## Visualization of three-dimensional data: Scatter Matrixes

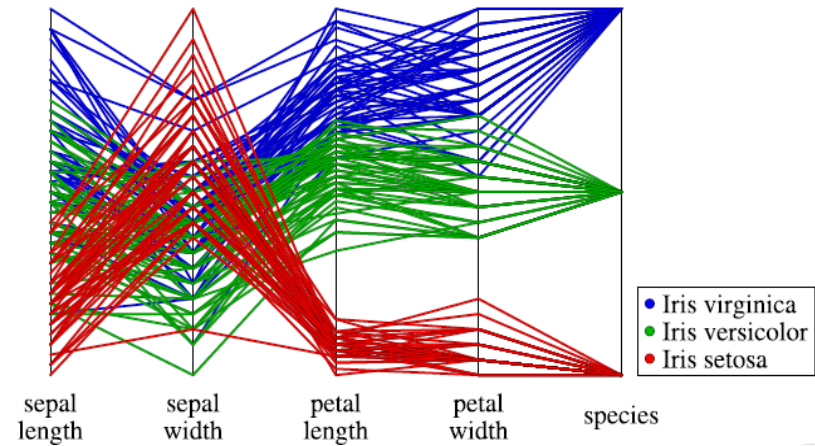
- A matrix of scatter plots  $m \times m$  where  $m$  is the number of attributes (data dimensionality)
- For  $m$  attributes there are  $\binom{m}{2} = m(m-1)/2$  possible scatter plots
- e.g. For 50 attributes there are 2450 scatter plots!



Scatter matrix

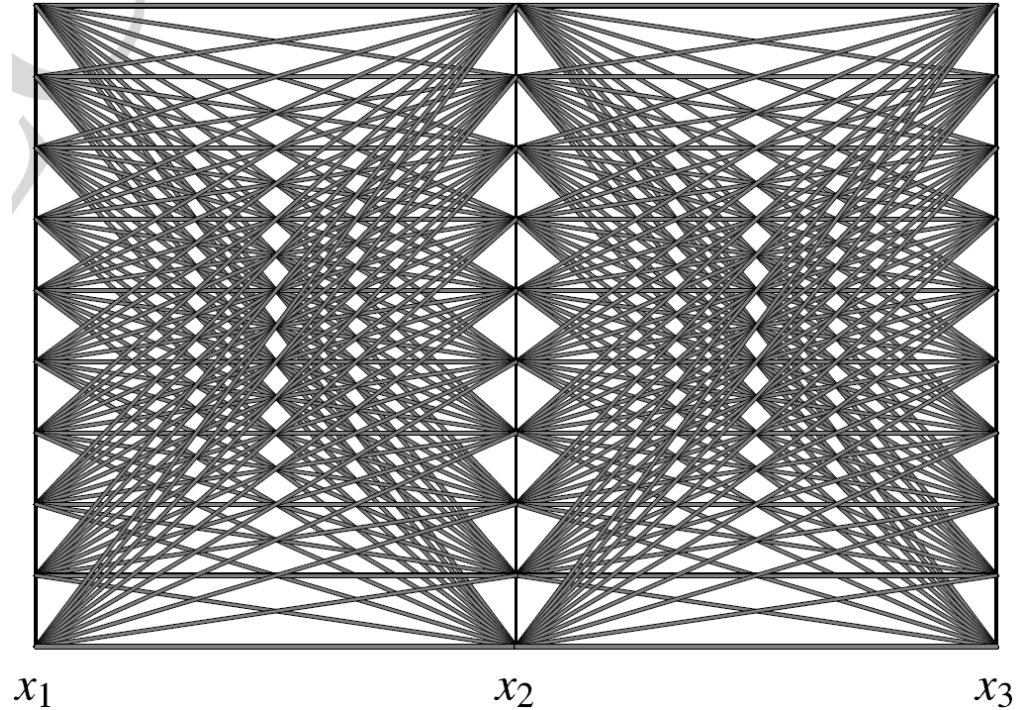
# Parallel Coordinates Plot

- Parallel coordinates draw the coordinate axes for each attribute parallel to each other, so that there is no limitation for the number of axes to be displayed.
- For each data object, a polyline is drawn connecting the values of the attributes on the corresponding axes.
- Maintains the original attributes
- Limited number of entries
- How do we spot correlation between features?

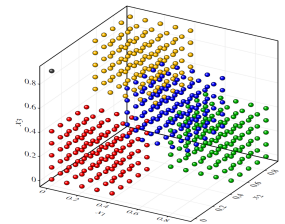


*Parallel coordinates plot for the Iris data set*

## Parallel Coordinates Plot: „Cube Data“

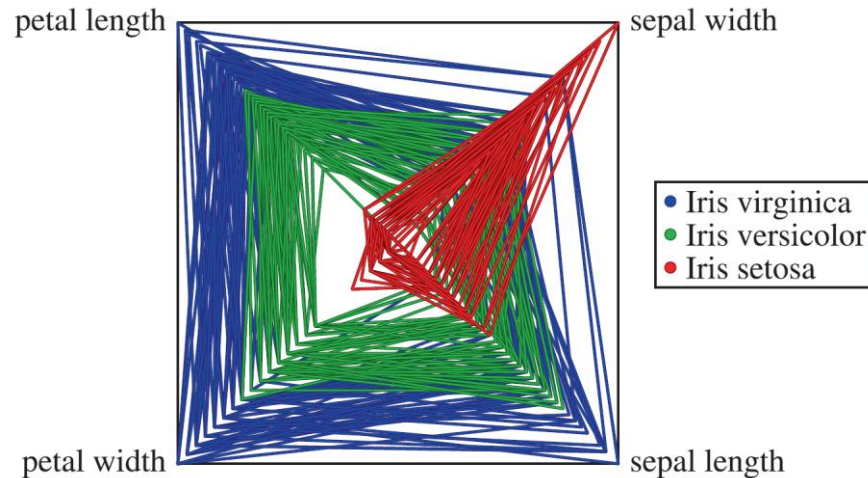


*Parallel coordinates plot for the Cube data*



## Radar Plot

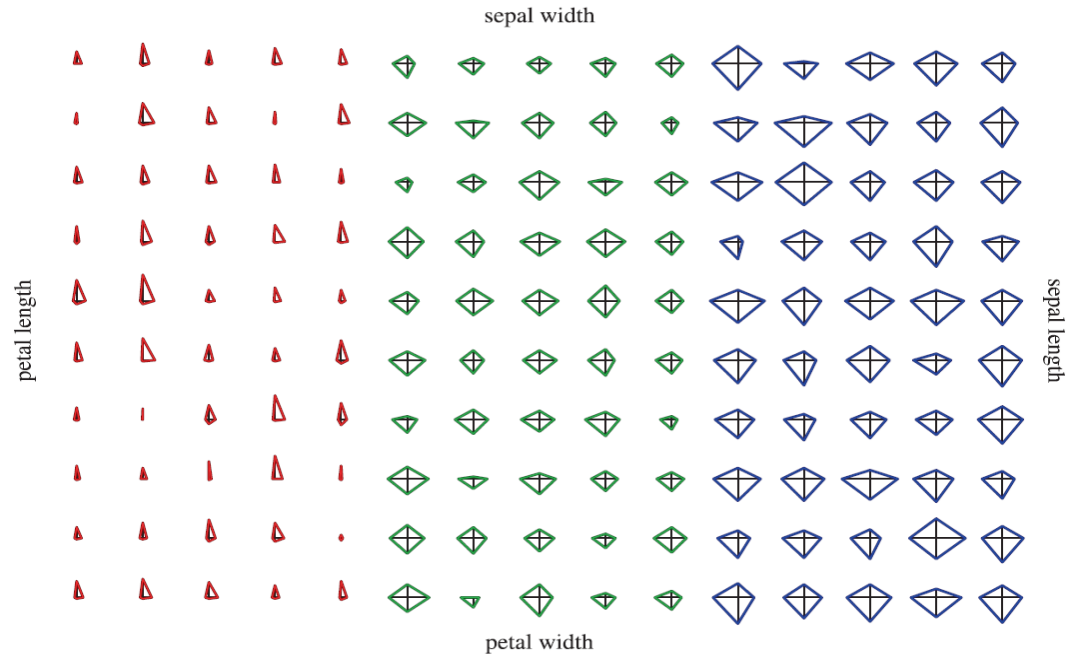
- Similar idea of the Parallel Coordinates plot
- Axes are drawn in a star-like fashion intersecting in one point
- Also called spider plots
- Suitable for small datasets



*Radar plot for the Iris data set*

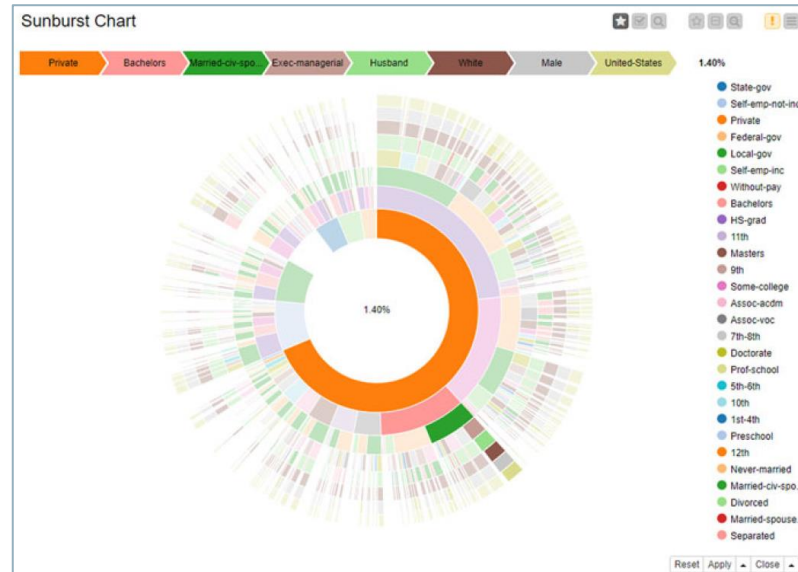
# Star Plots

- In a star plot each object is drawn separately
- In a radar plot fashion



*Star plot for the Iris data set*

# Sunburst Chart



- Display multidimensional hierarchical nominal data in a radial layout
- One section ⇔ one attribute
- Root attribute in the center, external sections are attributes located deeper in the hierarchy
- Area of a section represents the accumulated value of all descending sections

### **How can we transform a higher-dimensional data set to have two or three dimensions?**

- Preserve as much of the “structure” of the original data
- Define a measure to evaluate how well the original structure of the high-dimensional dataset is preserved after transformation
- Find the transformation that gives the best value for the given measure



# Correlation Analysis

How can we measure the similarity in behavior of two attributes?

- Pearson's correlation coefficient
- Spearman's rank correlation coefficient (Spearman's rho)

- **Pearson's correlation coefficient** is a measure for the **linear relationship** between two **numerical** attributes  $X$  and  $Y$  and is defined as:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{(n - 1) s_x s_y}$$

- where  $\bar{x}$  and  $\bar{y}$  are the (sample) mean values of the attributes  $X$  and  $Y$ , respectively, and  $s_x$  and  $s_y$  are the corresponding (sample) standard deviations
- $-1 \leq r_{xy} \leq 1$
- The larger the absolute value of the Pearson correlation coefficient, the stronger the linear relationship between the two attributes. For  $|r_{xy}| = 1$  the values of  $X$  and  $Y$  lie exactly on a line.
- Positive (negative) correlation indicates a linear relationship (a line) with positive (negative) slope.

- **Spearman's rank correlation coefficient (Spearman's rho)** is defined as:

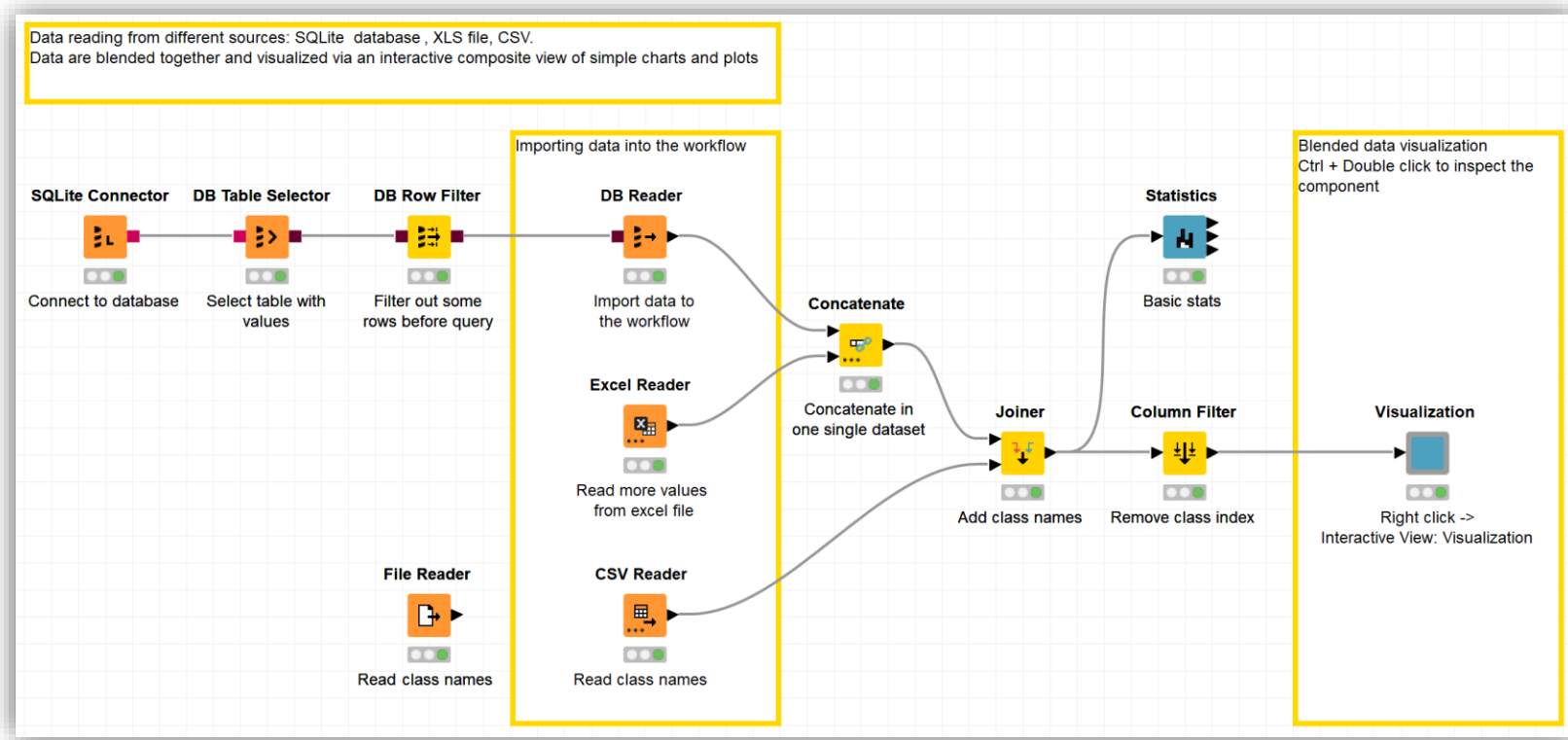
$$\rho = 1 - 6 \frac{\sum_{i=1}^n (r(x_i) - r(y_i))^2}{n (n^2 - 1)}$$

where  $r(x_i)$  is the rank of value  $x_i$  when we sort the list  $(x_1, x_2, \dots, x_n)$  in increasing order.  $r(y_i)$  is defined analogously.

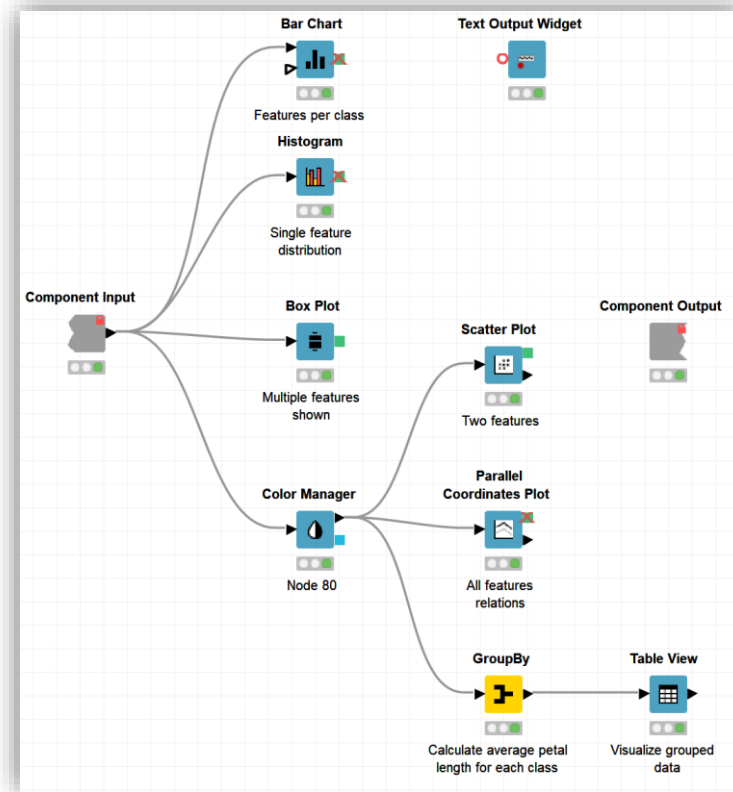
- When the rankings of the  $x$ - and  $y$ -values are exactly in the same order, Spearman's rho will yield value 1.
- If they are in reverse order, Spearman's rho will yield value  $-1$ .

# Practical Examples with KNIME Analytics Platform

## – Simple visualization



## – Inner workings of the visualization component



## – Interactive view





- Methods for One and Two Attributes
  - Barchart and Histogram
  - Boxplot
  - Scatter plot and density plot
- Methods for Higher-dimensional Data
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
  - Parallel Coordinates
  - Radar and Star Plots
  - Sunburst Chart
  - Correlation Analysis

**Thank you**