DA2I
Data Description and Preparation

Leandro Nunes de Castro, Ph.D.

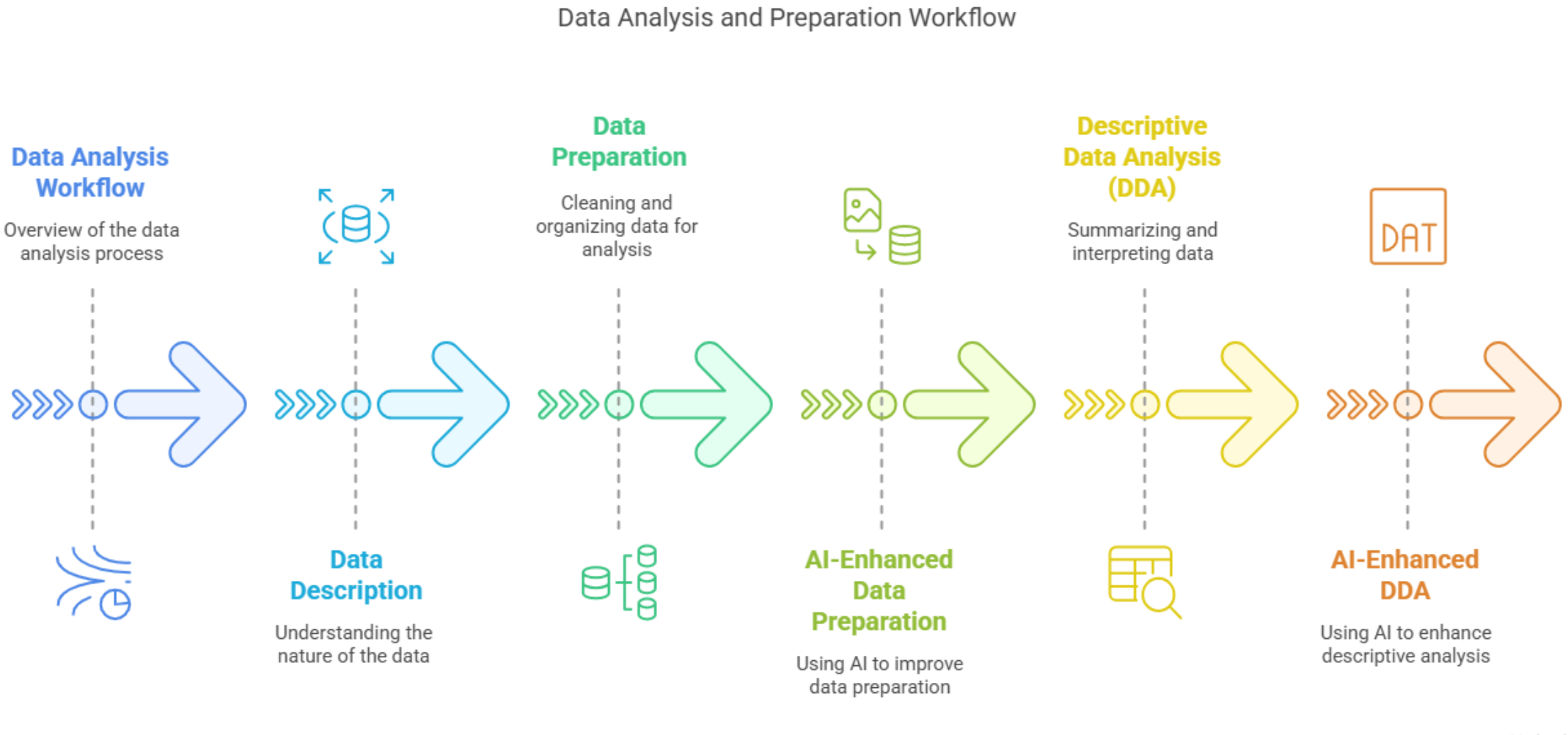
ldecastrosilva@fgcu.edu

DA2I

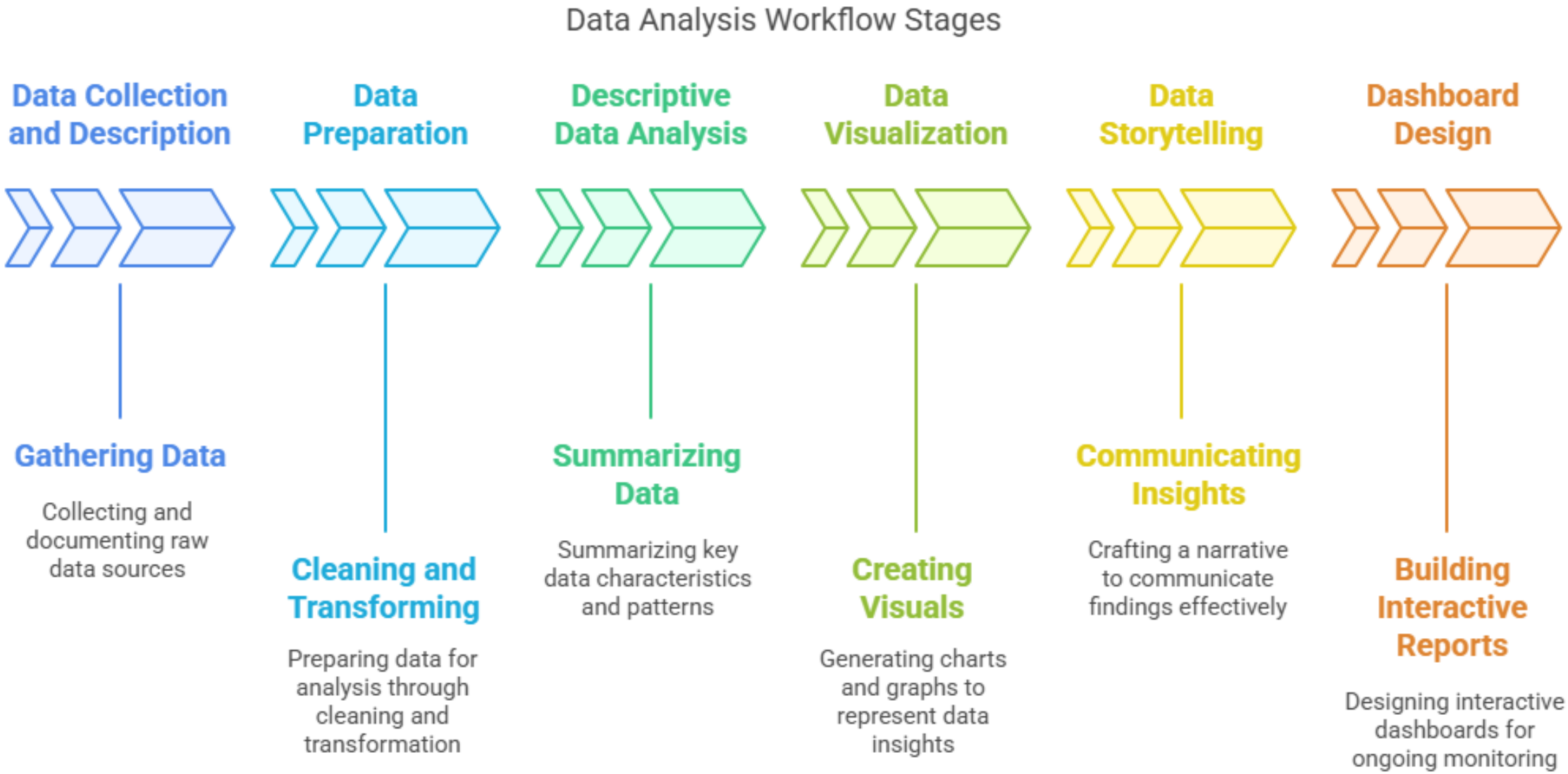
Data Description and Preparation

Summary

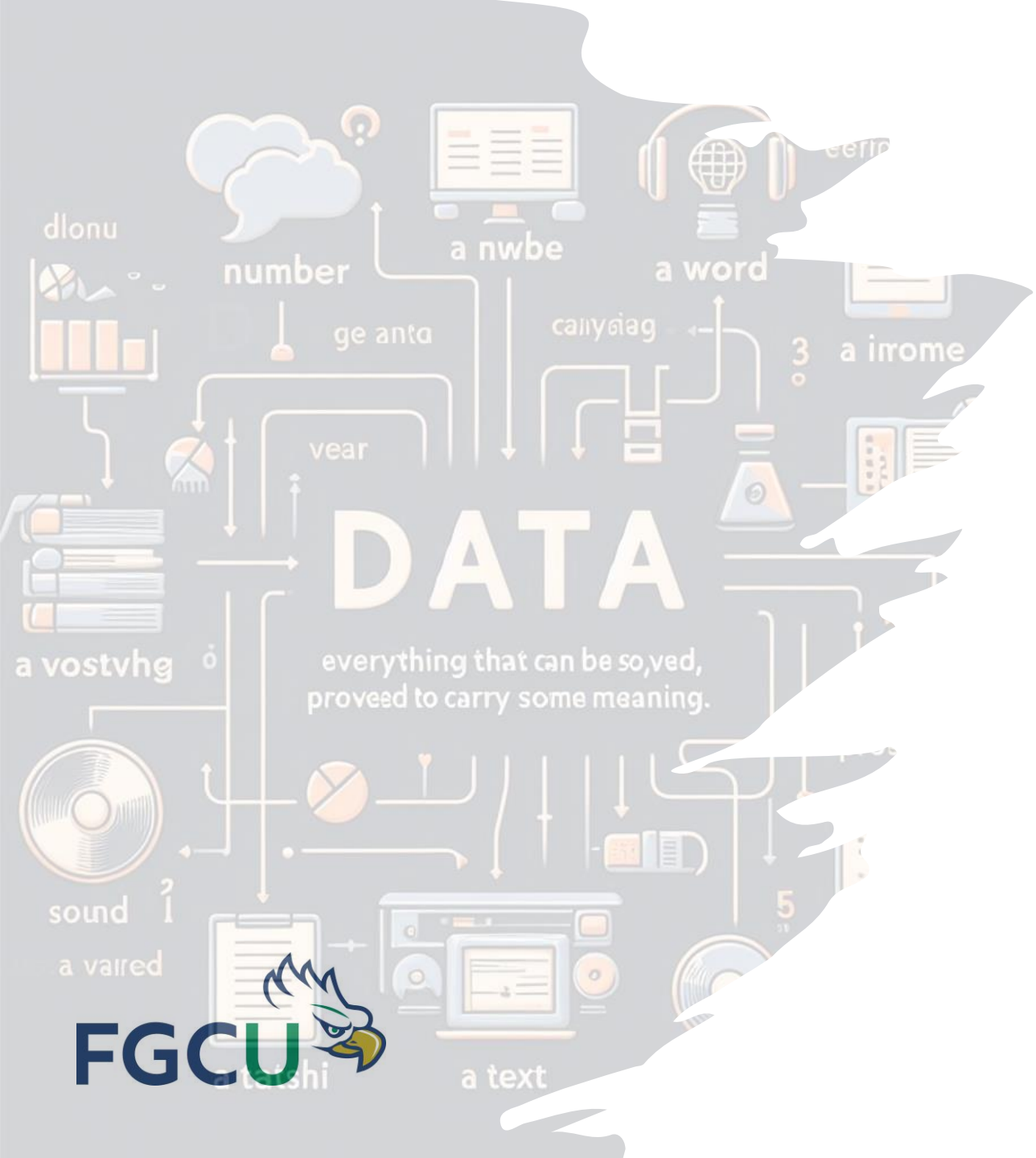
- The Data Science Workflow
- Data Description
- Data Preparation


Data Analysis and Preparation Workflow
Data Analysis Workflow
Overview of the data analysis process
Data Description
Understanding the nature of the data
Data Preparation
Cleaning and organizing data for analysis
AI-Enhanced Data Preparation
Using AI to improve data preparation
Descriptive Data Analysis (DDA)
Summarizing and interpreting data
DAT
AI-Enhanced DDA
Using AI to enhance descriptive analysis
Made with Napkin

The Data Analysis Workflow


Data Analysis Workflow Stages
Data Collection and Description
Data Preparation
Descriptive Data Analysis
Data Visualization
Data Storytelling
Dashboard Design
Gathering Data
Collecting and documenting raw data sources
Cleaning and Transforming
Preparing data for analysis through cleaning and transformation
Summarizing Data
Summarizing key data characteristics and patterns
Creating Visuals
Generating charts and graphs to represent data insights
Communicating Insights
Crafting a narrative to communicate findings effectively
Building Interactive Reports
Designing interactive dashboards for ongoing monitoring
Made with Napkin

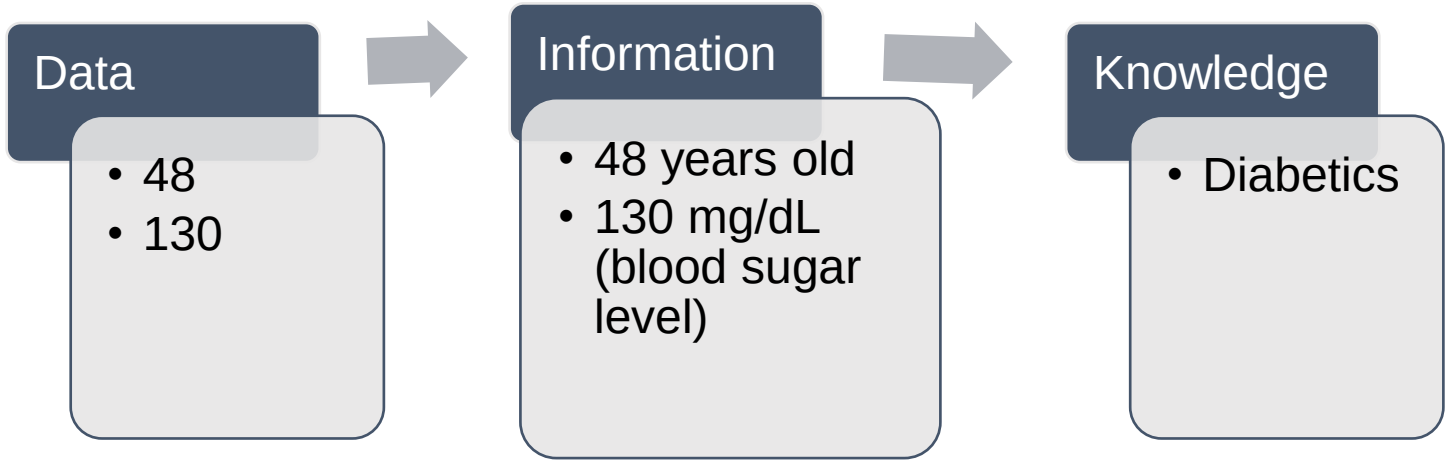
Data Description



Data

- **Data** is everything that can be used, moved, processed, or translated to carry some meaning. A number, a word, an image, a text, a graph, and a sound are all examples of data. In computational terms, anything that can be stored and/or processed is a kind of data.
- Data *vs* Datum.

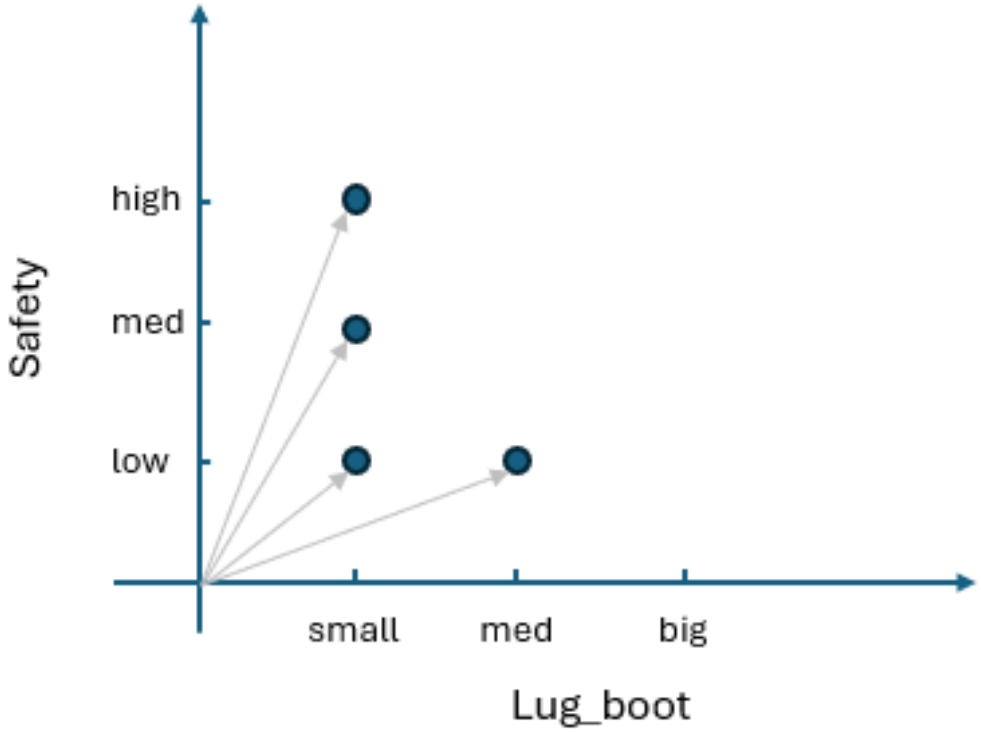
Data → Information → Knowledge



Tabular and Mathematical Representation of Data

Table 1: First four objects of the Car Evaluation Dataset available at UCI.

Car ID	Buying	Maintenance	Doors	Persons	Lug_boot	Safety	Class
1	vhigh	vhigh	2	2	small	low	unacc
2	vhigh	vhigh	2	2	small	med	unacc
3	vhigh	vhigh	2	2	small	high	unacc
4	vhigh	vhigh	2	2	med	low	unacc



Data Dictionary

Table 2: Example of a simple data dictionary for the Car Evaluation Dataset presented in Table 1.

Variable name	Definition (meaning)	Domain
Car ID	ID number of each car in the dataset	Integer number
Buying	Buying price	{v-high, high, med, low}
Maintenance	Level of maintenance required	{v-high, high, med, low}
Doors	Number of doors	{2, 3, 4, 5-more}
Persons	Number of persons accommodated	{2, 4, more}
Lug-boot	Trunk size	{small, med, big}
Safety	Level of safety	{low, med, high}
Class	Car acceptability	{unacc, acc, good, vgood}

Mammographic Dataset

Table 5: Mammographic dataset sample: first and last five objects in the dataset. Question marks, '?', indicate missing values.

Patient	BI-RADS	Age	Shape	Margin	Density	Severity
0	5	67	Lobular	Spiculated	Low	Malignant
1	4	43	Round	Circumscribed	?	Malignant
2	5	58	Irregular	Spiculated	Low	Malignant
3	4	28	Round	Circumscribed	Low	Benign
4	5	74	Round	Spiculated	?	Malignant
...	...	...	...	...	...	...
956	4	47	Oval	Circumscribed	Low	Benign
957	4	56	Irregular	Spiculated	Low	Malignant
958	4	64	Irregular	Spiculated	Low	Benign
959	5	66	Irregular	Spiculated	Low	Malignant
960	4	62	Lobular	Obscured	Low	Benign

FGCU

Data Dictionary

Table 14: Data dictionary for the Mammographic dataset.

Variable name	Definition (meaning)	Variable Type: Domain	Number of Missing Values
BI-RADS	Breast Imaging-Reporting and Data System. International system to evaluate, interpret and report breast imaging exams.	Ordinal: [1,5]	2
Age	Patient age in years	Integer	5
Shape	Mass shape	Nominal: {Round=1, Oval=2, Lobular=3, Irregular=4}	31
Margin	Mass margin	Nominal: {Circumscribed=1, Microlobulated=2, Obscured=3, Ill-defined=4, Spiculated=5}	48
Density	Mass density	Ordinal: {Mass density high=1, Iso=2, Low=3, Fat-containing=4}	76
Severity	Severity level	Binary: {Benign=0, Malignant=1}	0

Data Preparation

Introduction

- **Raw data**: source or primary data that has not been prepared or processed for being used; is the one originally input in a database by operators, sensors or any person or device.
- Main problems:
  - **Data overload**: excessive number of objects or variables.
  - **Incompleteness**: missing objects, values or variables.
  - **Inconsistency**: domain violations and discrepancies.
  - **Noise**: random variations or irregularities in the data that are not part of the underlying pattern or signal.

Dealing with Data Overload: Sampling

- Random Sampling with Replacement (RSWR)
- Random Sampling without Replacement (RSWoR)
- Systematic Sampling
- Group Sampling
- Stratified Sampling

Sampling Rate: 60%

	BI-RADS	Age	Shape	Margin	Density	Severity
1	5	67	Lobular	Spiculated	Low	Malignant
2	4	43	Round	Circumscribed	?	Malignant
3	5	58	Irregular	Spiculated	Low	Malignant
4	4	28	Round	Circumscribed	Low	Benign
5	5	74	Round	Spiculated	?	Malignant
6	4	47	Oval	Circumscribed	Low	Benign
7	4	56	Irregular	Spiculated	Low	Malignant
8	4	64	Irregular	Spiculated	Low	Benign
9	5	66	Irregular	Spiculated	Low	Malignant
10	4	62	Lobular	Obscured	Low	Benign

Sampling with Replacement

	BI-RADS	Age	Shape	Margin	Density	Severity
1	5	67	Lobular	Spiculated	Low	Malignant
1	5	67	Lobular	Spiculated	Low	Malignant
7	4	56	Irregular	Spiculated	Low	Malignant
10	4	62	Lobular	Obscured	Low	Benign
7	4	56	Irregular	Spiculated	Low	Malignant
3	5	58	Irregular	Spiculated	Low	Malignant

Sampling without Replacement

	BI-RADS	Age	Shape	Margin	Density	Severity
3	5	58	Irregular	Spiculated	Low	Malignant
4	4	28	Round	Circumscribed	Low	Benign
10	4	62	Lobular	Obscured	Low	Benign
6	4	47	Oval	Circumscribed	Low	Benign
1	5	67	Lobular	Spiculated	Low	Malignant
5	5	74	Round	Spiculated	?	Malignant

Systematic and Group Sampling Examples

	BI-RADS	Age	Shape	Margin	Density	Severity
1	5	67	Lobular	Spiculated	Low	Malignant
2	4	43	Round	Circumscribed	?	Malignant
3	5	58	Irregular	Spiculated	Low	Malignant
4	4	28	Round	Circumscribed	Low	Benign
5	5	74	Round	Spiculated	?	Malignant
6	4	47	Oval	Circumscribed	Low	Benign
7	4	56	Irregular	Spiculated	Low	Malignant
8	4	64	Irregular	Spiculated	Low	Benign
9	5	66	Irregular	Spiculated	Low	Malignant
10	4	62	Lobular	Obscured	Low	Benign

Systematic Sampling: Odd objects

	BI-RADS	Age	Shape	Margin	Density	Severity
1	5	67	Lobular	Spiculated	Low	Malignant
3	5	58	Irregular	Spiculated	Low	Malignant
5	5	74	Round	Spiculated	?	Malignant
7	4	56	Irregular	Spiculated	Low	Malignant
9	5	66	Irregular	Spiculated	Low	Malignant

Group Sampling: Class Severity = Benign

BI-RADS	Age	Shape	Margin	Density	Severity
4	28	Round	Circumscribed	Low	Benign
4	47	Oval	Circumscribed	Low	Benign
4	64	Irregular	Spiculated	Low	Benign
4	62	Lobular	Obscured	Low	Benign

Stratified Sampling Example: 50% Sample

BI-RADS	Age	Shape	Margin	Density	Severity
5	67	Lobular	Spiculated	Low	Malignant
4	43	Round	Circumscribed	?	Malignant
5	58	Irregular	Spiculated	Low	Malignant
4	28	Round	Circumscribed	Low	Benign
5	74	Round	Spiculated	?	Malignant
4	47	Oval	Circumscribed	Low	Benign
4	56	Irregular	Spiculated	Low	Malignant
4	64	Irregular	Spiculated	Low	Benign
5	66	Irregular	Spiculated	Low	Malignant
4	62	Lobular	Obscured	Low	Benign

Class Severity

BI-RADS	Age	Shape	Margin	Density	Severity
5	67	Lobular	Spiculated	Low	Malignant
5	74	Round	Spiculated	?	Malignant
4	56	Irregular	Spiculated	Low	Malignant
4	64	Irregular	Spiculated	Low	Benign
4	62	Lobular	Obscured	Low	Benign

Dealing with Incompleteness: Missing Values

- Ignore the object
- Manually input missing values
- Global constant imputation
- Hot-deck imputation
- Central tendency measure of the variable
- Central tendency measure of the variable class

Missing Values

	BI-RADS	Age	Shape	Margin	Density	Severity
1	5	67	Lobular	Spiculated	Low	Malignant
2	4	43	Round	Circumscribed	?	Malignant
3	5	58	Irregular	Spiculated	Low	Malignant
4	4	28	Round	Circumscribed	Low	Benign
5	5	74	Round	Spiculated	?	Malignant
6	4	47	Oval	Circumscribed	High	Benign
7	4	56	Irregular	Spiculated	Low	Malignant
8	4	64	Irregular	Spiculated	High	Benign
9	5	66	Irregular	Spiculated	Low	Malignant
10	4	62	Lobular	Obscured	High	Benign

Ignore the Object

	BI-RADS	Age	Shape	Margin	Density	Severity
1	5	67	Lobular	Spiculated	Low	Malignant
3	5	58	Irregular	Spiculated	Low	Malignant
4	4	28	Round	Circumscribed	Low	Benign
6	4	47	Oval	Circumscribed	High	Benign
7	4	56	Irregular	Spiculated	Low	Malignant
8	4	64	Irregular	Spiculated	High	Benign
9	5	66	Irregular	Spiculated	Low	Malignant
10	4	62	Lobular	Obscured	High	Benign

Manual Input

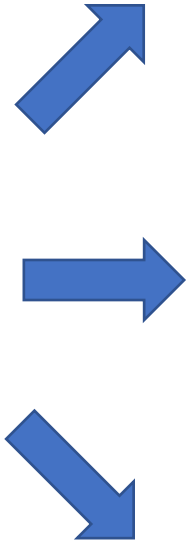
BI-RADS	Age	Shape	Margin	Density	Severity
5	67	Lobular	Spiculated	Low	Malignant
4	43	Round	Circumscribed	Low	Malignant
5	58	Irregular	Spiculated	Low	Malignant
4	28	Round	Circumscribed	Low	Benign
5	74	Round	Spiculated	High	Malignant
4	47	Oval	Circumscribed	High	Benign
4	56	Irregular	Spiculated	Low	Malignant
4	64	Irregular	Spiculated	High	Benign
5	66	Irregular	Spiculated	Low	Malignant
4	62	Lobular	Obscured	High	Benign

Global Constant

BI-RADS	Age	Shape	Margin	Density	Severity
5	67	Lobular	Spiculated	Low	Malignant
4	43	Round	Circumscribed	High	Malignant
5	58	Irregular	Spiculated	Low	Malignant
4	28	Round	Circumscribed	Low	Benign
5	74	Round	Spiculated	High	Malignant
4	47	Oval	Circumscribed	High	Benign
4	56	Irregular	Spiculated	Low	Malignant
4	64	Irregular	Spiculated	High	Benign
5	66	Irregular	Spiculated	Low	Malignant
4	62	Lobular	Obscured	High	Benign

Missing Values

	BI-RADS	Age	Shape	Margin	Density	Severity
1	5	67	Lobular	Spiculated	Low	Malignant
2	4	43	Round	Circumscribed	?	Malignant
3	5	58	Irregular	Spiculated	Low	Malignant
4	4	28	Round	Circumscribed	Low	Benign
5	5	74	Round	Spiculated	?	Malignant
6	4	47	Oval	Circumscribed	High	Benign
7	4	56	Irregular	Spiculated	Low	Malignant
8	4	64	Irregular	Spiculated	High	Benign
9	5	66	Irregular	Spiculated	Low	Malignant
10	4	62	Lobular	Obscured	High	Benign



Hot Deck

BI-RADS	Age	Shape	Margin	Density	Severity
5	67	Lobular	Spiculated	Low	Malignant
4	43	Round	Circumscribed	Low	Malignant
5	58	Irregular	Spiculated	Low	Malignant
4	28	Round	Circumscribed	Low	Benign
5	74	Round	Spiculated	Low	Malignant
4	47	Oval	Circumscribed	High	Benign
4	56	Irregular	Spiculated	Low	Malignant
4	64	Irregular	Spiculated	High	Benign
5	66	Irregular	Spiculated	Low	Malignant
4	62	Lobular	Obscured	High	Benign

Central Tendency of the Variable

BI-RADS	Age	Shape	Margin	Density	Severity
5	67	Lobular	Spiculated	Low	Malignant
4	43	Round	Circumscribed	Low	Malignant
5	58	Irregular	Spiculated	Low	Malignant
4	28	Round	Circumscribed	Low	Benign
5	74	Round	Spiculated	Low	Malignant
4	47	Oval	Circumscribed	High	Benign
4	56	Irregular	Spiculated	Low	Malignant
4	64	Irregular	Spiculated	High	Benign
5	66	Irregular	Spiculated	Low	Malignant
4	62	Lobular	Obscured	High	Benign

Central Tendency of the Class

BI-RADS	Age	Shape	Margin	Density	Severity
5	67	Lobular	Spiculated	Low	Malignant
4	43	Round	Circumscribed	Low	Malignant
5	58	Irregular	Spiculated	Low	Malignant
4	28	Round	Circumscribed	Low	Benign
5	74	Round	Spiculated	Low	Malignant
4	47	Oval	Circumscribed	High	Benign
4	56	Irregular	Spiculated	Low	Malignant
4	64	Irregular	Spiculated	High	Benign
5	66	Irregular	Spiculated	Low	Malignant
4	62	Lobular	Obscured	High	Benign

Normalization (Feature Scaling)

- Min-Max

$$x_i' = \frac{x_i - min}{max - min}(nmax - nmin) + nmin$$

- z-score

$$x_i' = \frac{x_i - mean}{\sigma}$$

where x_i is the i-th value of variable x, and x_i' is its value after normalization.

Normalization Example: Iris dataset

Dataset Sample

Object	Sepal length	Sepal width	Petal length	Petal width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

Dataset Statistics

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
mean	5.84	3.06	3.76	1.20
std	0.83	0.44	1.77	0.76
min	4.30	2.00	1.00	0.10
25%	5.10	2.80	1.60	0.30
50%	5.80	3.00	4.35	1.30
75%	6.40	3.30	5.10	1.80
max	7.90	4.40	6.90	2.50

	Original	Min-Max Normalized	Z-Score Normalized
0	5.1	0.22	-0.90
1	4.9	0.17	-1.14
2	4.7	0.11	-1.39
3	4.6	0.08	-1.51
4	5.0	0.19	-1.02
5	5.4	0.31	-0.54
6	4.6	0.08	-1.51
7	5.0	0.19	-1.02
8	4.4	0.03	-1.75
9	4.9	0.17	-1.14
10	5.4	0.31	-0.54
50	7.0	0.75	1.40
51	6.4	0.58	0.67
52	6.9	0.72	1.28
53	5.5	0.33	-0.42
54	6.5	0.61	0.80
55	5.7	0.39	-0.17
56	6.3	0.56	0.55
57	4.9	0.17	-1.14
58	6.6	0.64	0.92
59	5.2	0.25	-0.78
60	5.0	0.19	-1.02
100	6.3	0.56	0.55
101	5.8	0.42	-0.05
102	7.1	0.78	1.52
103	6.3	0.56	0.55
104	6.5	0.61	0.80
105	7.6	0.92	2.13
106	4.9	0.17	-1.14
107	7.3	0.83	1.77
108	6.7	0.67	1.04
109	7.2	0.81	1.64
110	6.5	0.61	0.80

AI-Enhanced Data Preparation

Sampling with AI

- For the mammographic dataset, prompt:

"*Using a stratified sampling approach, sample 20% of the dataset. Assume variable 'severity' as the target variable*."

"*Using a stratified sampling approach, sample 20% of the dataset. Assume variable 'shape' as the target variable*."

- Analyze the results.

Finding and Replacing Missing Values with AI

- Open the mammographic_masses_nominal dataset using Excel.
- Observe the missing values represented with ‘?’.
- Apply filters in all variables to observe the missing values.
- Prompt the tools to find missing values:

“*Find the missing values of the mammographic data (they are represented by the question mark)*”

- Prompt the tools to replace missing:

“*Replace these missing values by a central tendency measure of the variable and save the dataset with the name mammographic_data_wo_missing_values.*”

Data Normalization with AI

- For the Iris dataset of Fisher, prompt the tools to:

“*Normalize the iris dataset attached using a min-max method and the z-score*”

- Analyze the results.

Leandro Nunes de Castro

ldecastrosilva@fgcu.edu

https://www.linkedin.com/in/lndecastro/

FGCU

DA2I
Descriptive Data Analysis

Leandro Nunes de Castro, Ph.D.

ldecastrosilva@fgcu.edu

DA2I

Descriptive Analysis

Summary

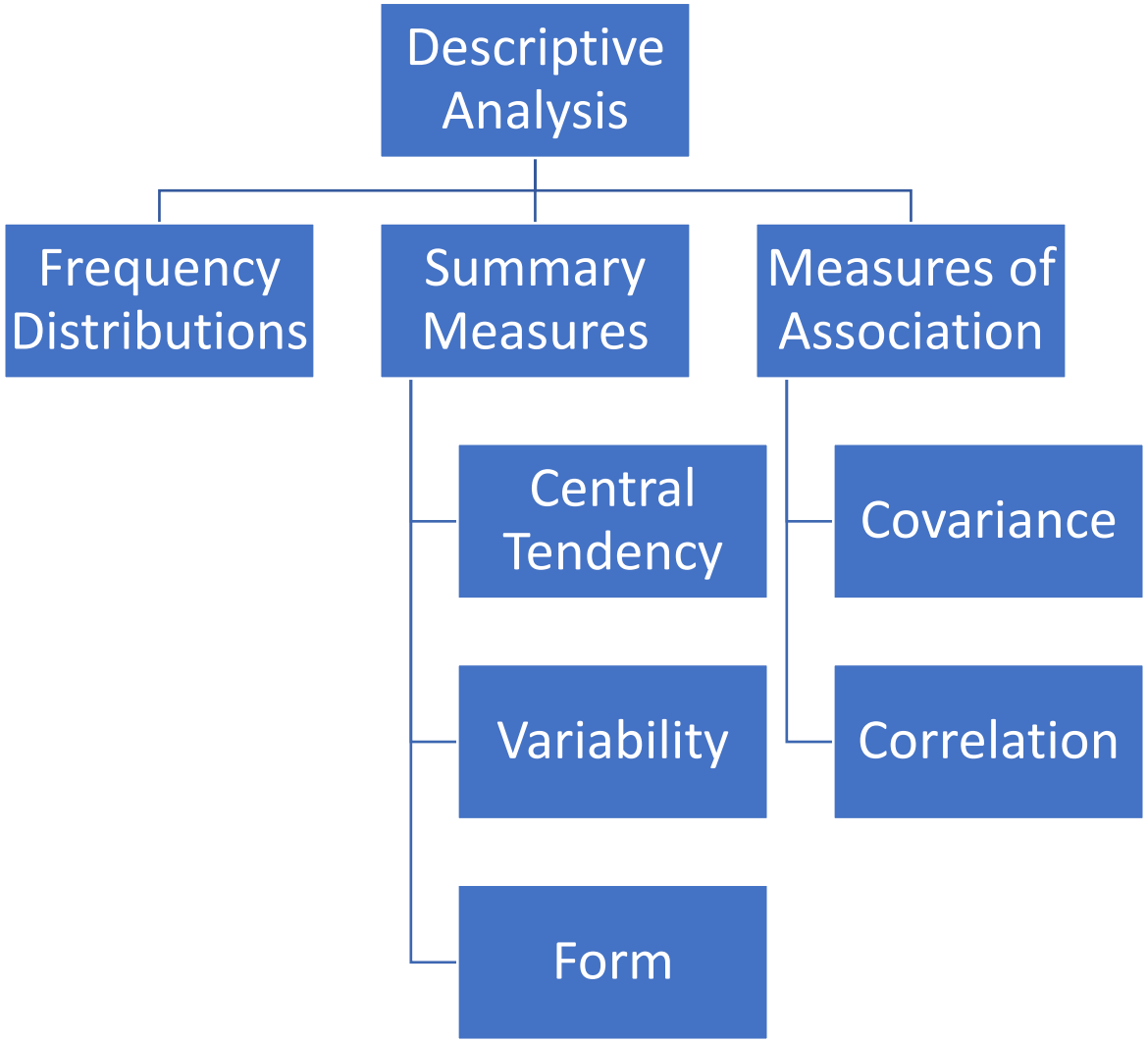
- What is Descriptive Data Analysis
- Distributions
- Summary Measures
  - Central Tendency
  - Variability
  - Relative Position
  - Measures of Shape
- The Normal Distribution
- Association Measures
- Linear Regression

What is Descriptive Data Analysis (DDA)?

- DDA involves a range of methods and techniques capable of summarizing, organizing, characterizing, and describing data in numerical terms.
- DDA differs from *data analytics* in the sense that it does not involve generalizing beyond the data available.
- Generalization is the capability of responding appropriately to unknown data, what usually requires building a model or a solution that is capable of extrapolating what it learnt from a given set of data to these new, unknown data samples.

Some questions to be answered:

1. How are the variables distributed?
2. What are the typical values of each variable?
3. What is the dispersion (variability) of each variable?
4. What is the shape of the variable's distribution?
5. What is the type and level of association among variables?


Descriptive Analysis
Frequency Distributions
Summary Measures
Measures of Association
Central Tendency
Variability
Form
Covariance
Correlation

Forest Fires Dataset

Table 6: Forest Fires dataset sample: first and last five objects in the dataset.

Obj	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.00
1	7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.00
2	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.00
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.00
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...
512	4	3	aug	sun	81.6	56.7	665.6	1.9	27.8	32	2.7	0.0	6.44
513	2	4	aug	sun	81.6	56.7	665.6	1.9	21.9	71	5.8	0.0	54.29
514	7	4	aug	sun	81.6	56.7	665.6	1.9	21.2	70	6.7	0.0	11.16
515	1	4	aug	sat	94.4	146.0	614.7	11.3	25.6	42	4.0	0.0	0.00
516	6	3	nov	tue	79.5	3.0	106.7	1.1	11.8	31	4.5	0.0	0.00

517 objects
13 variables

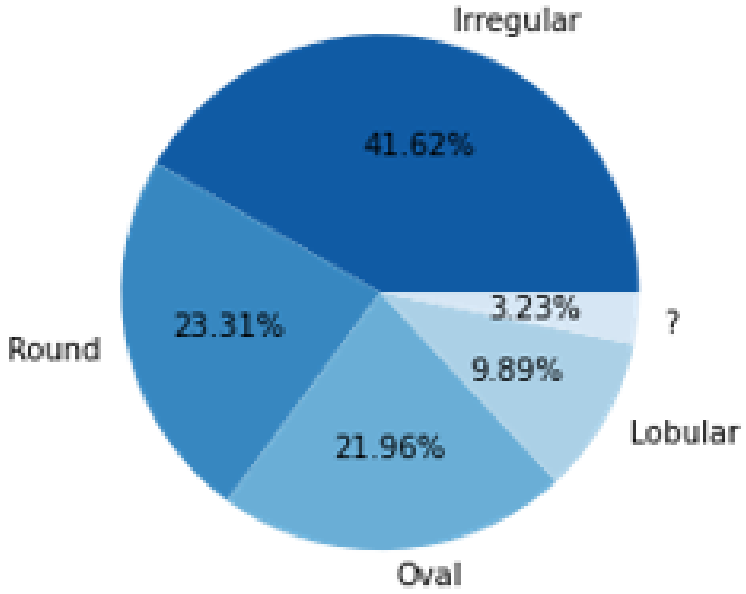
FFMC: Fine Fuel Moisture Code, which is a numeric rating of the moisture content of litter and other cured fine fuels.
DMC: Duff Moisture Code, which is a numeric rating of the average moisture content of loosely compacted organic layers.
DC: Drought Code, which is a numeric rating of the drying potential of deep organic layers.

Distributions

Frequency Distributions

Table 13: Frequency table and pie chart of variable 'Shape' in the mammographic dataset.

Shape	Absolute Frequency	Relative Frequency (%)	Cumulative Frequency (%)
Irregular	400	41.62	41.62
Round	224	23.31	64.93
Oval	211	21.96	86.89
Lobular	95	9.89	96.77
?	31	3.23	100.00



Frequency Table and Pie Chart with AI

- For the normalized or unnormalized mammographic dataset, prompt:

“*Print the frequency table with the absolute, relative, and cumulative frequency, then plot the pie chart of variable 'Shape' in the mammographic dataset*”

- Analyze the results.

Frequency Table and Histogram

Table 13: Frequency table for variable 'temp' of the Forest Fires dataset.

Bins	Absolute Frequency	Relative Frequency	Cumulative Frequency
(0.0, 3.33]	1	0.19	0.19
(3.33, 6.66]	20	3.87	4.06
(6.66, 9.99]	15	2.90	6.96
(9.99, 13.32]	47	9.09	16.05
(13.32, 16.65]	75	14.51	55.32
(16.65, 19.98]	128	24.76	40.81
(19.98, 23.31]	119	23.02	78.34
(23.31, 26.64]	69	13.35	91.68
(26.64, 29.97]	30	5.80	97.49
(29.97, 33.3]	13	2.51	100.00

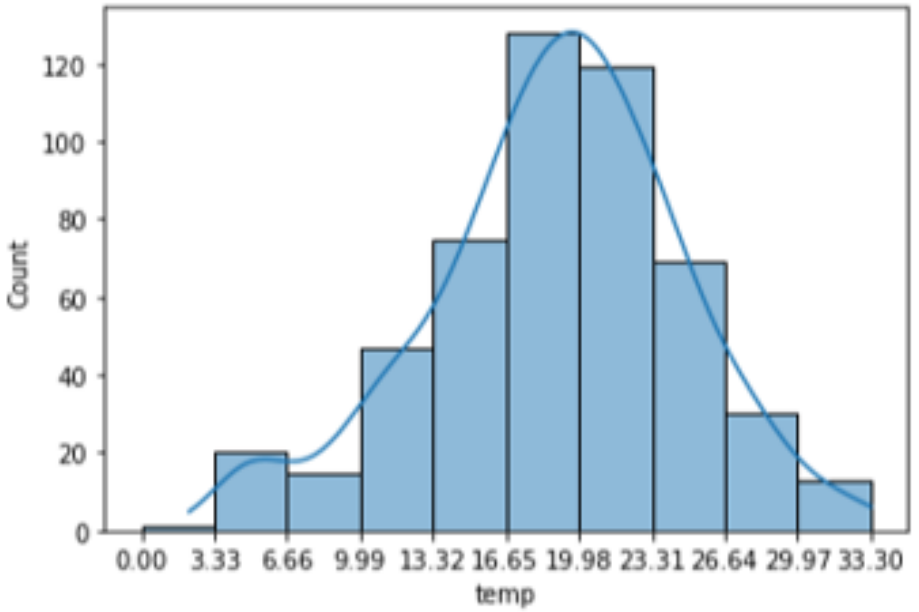
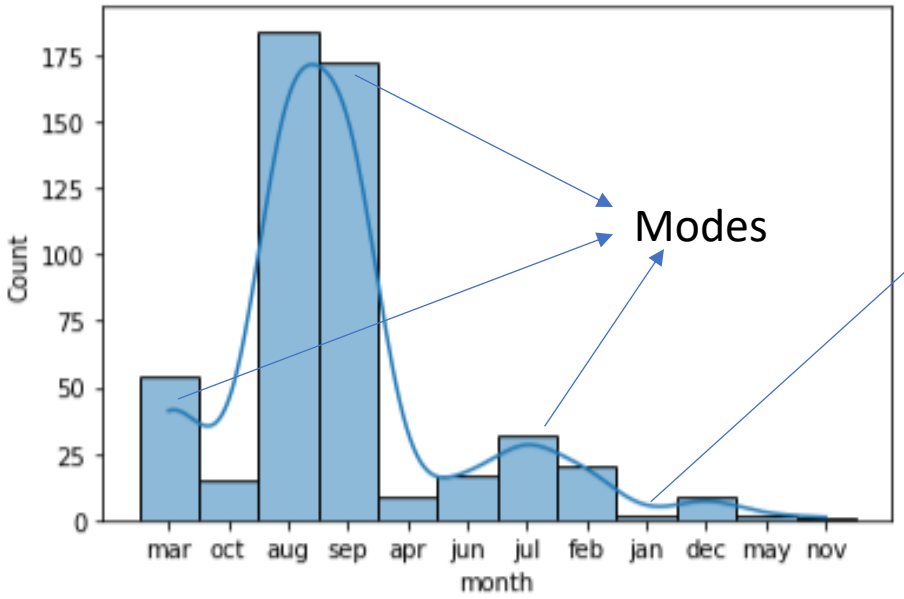
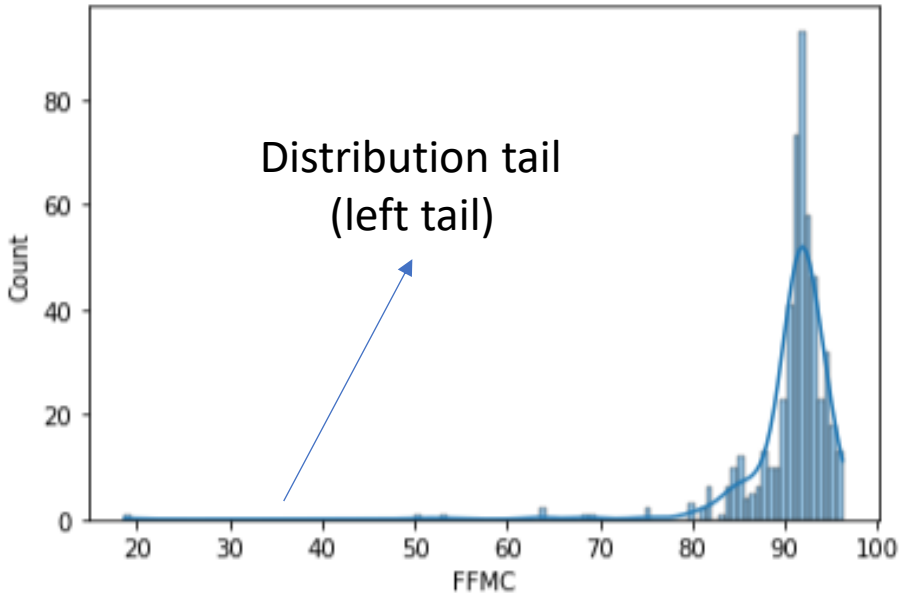


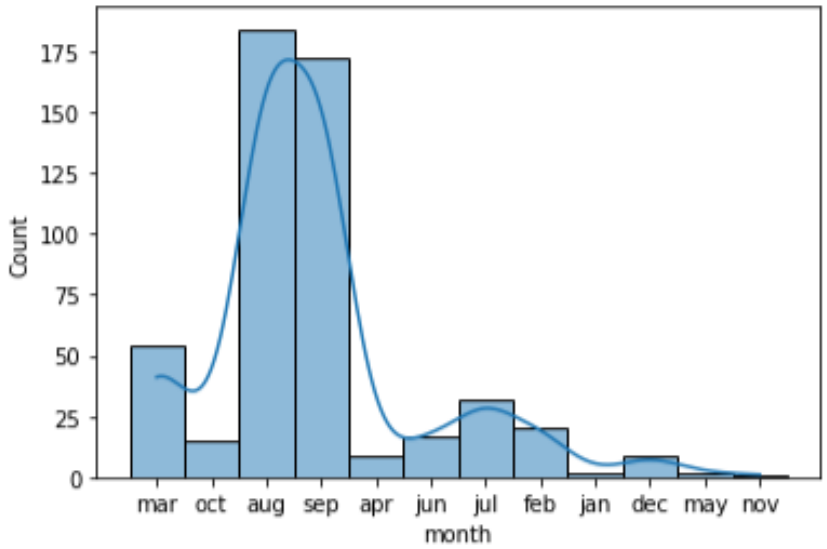
Figure 10: Histogram for the variable 'temp' of the Forest Fires dataset.

Distributions and Histograms

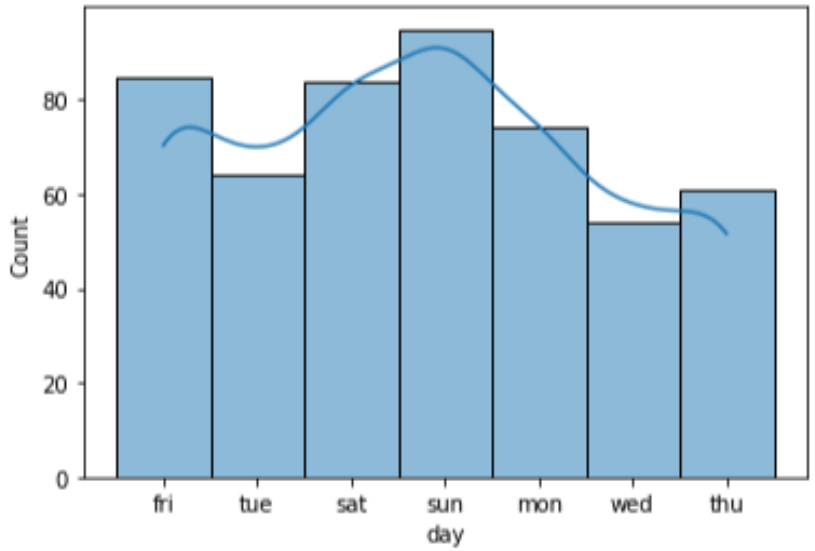




Forest Fires Dataset: Frequency Distributions

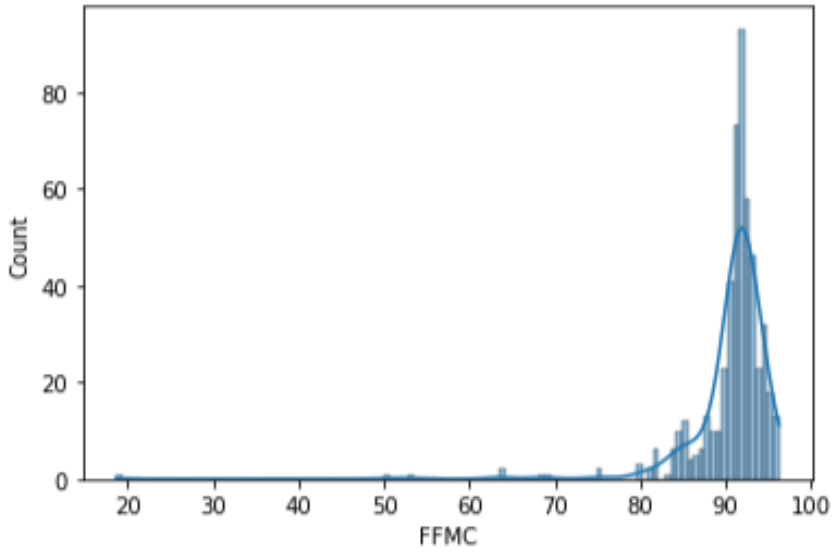


Multimodal
Right-skewed (positive skew)

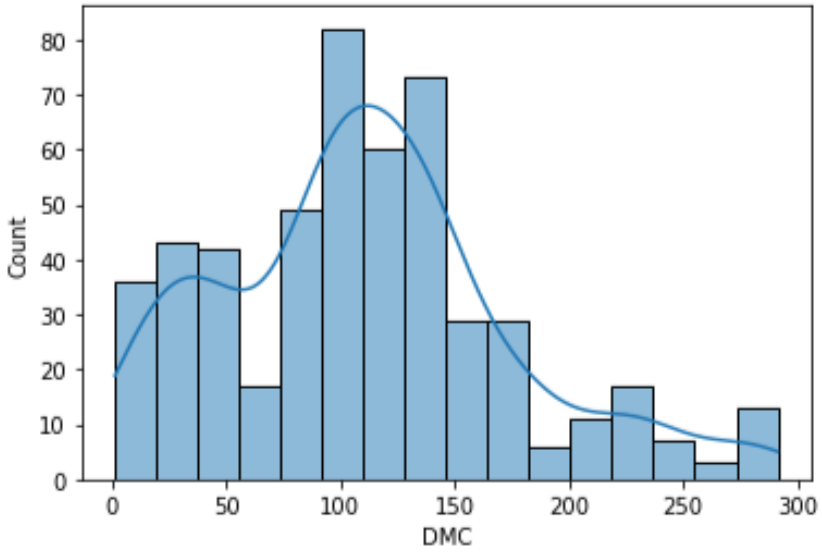


Multimodal
Almost uniform

Forest Fires Dataset: Frequency Distributions

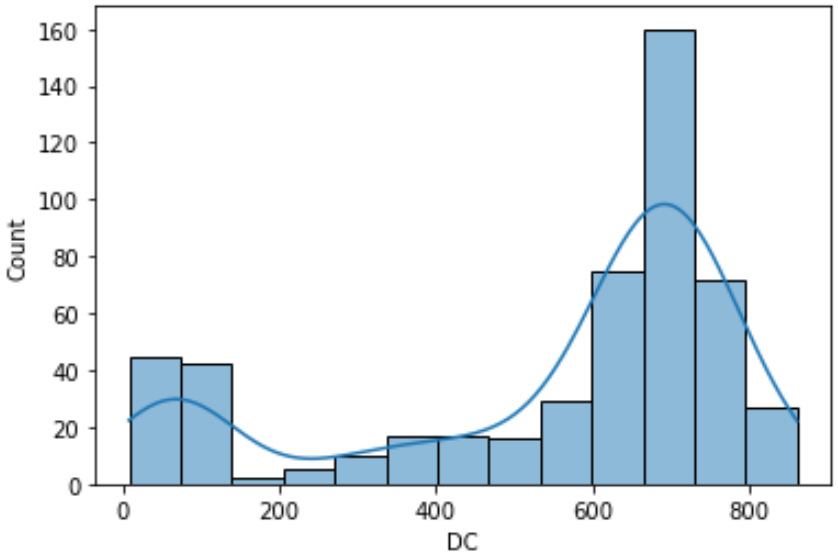


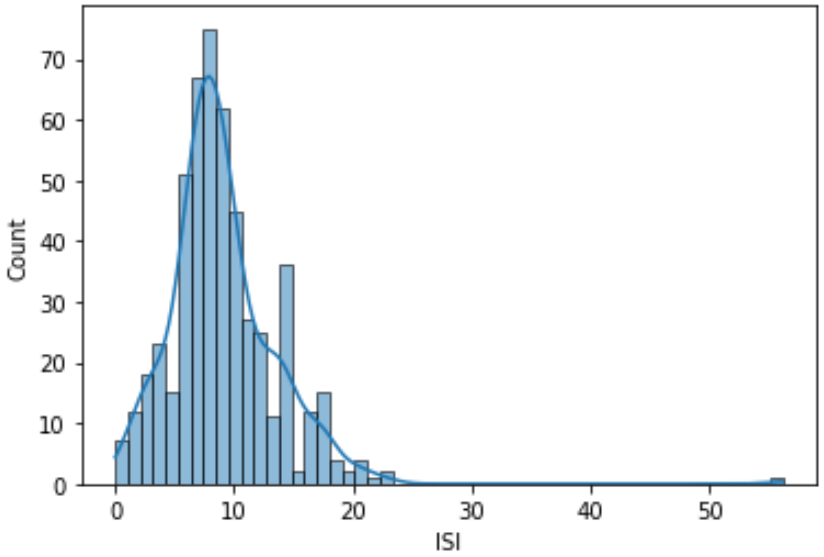
Unimodal
Left-skewed (negative skew)

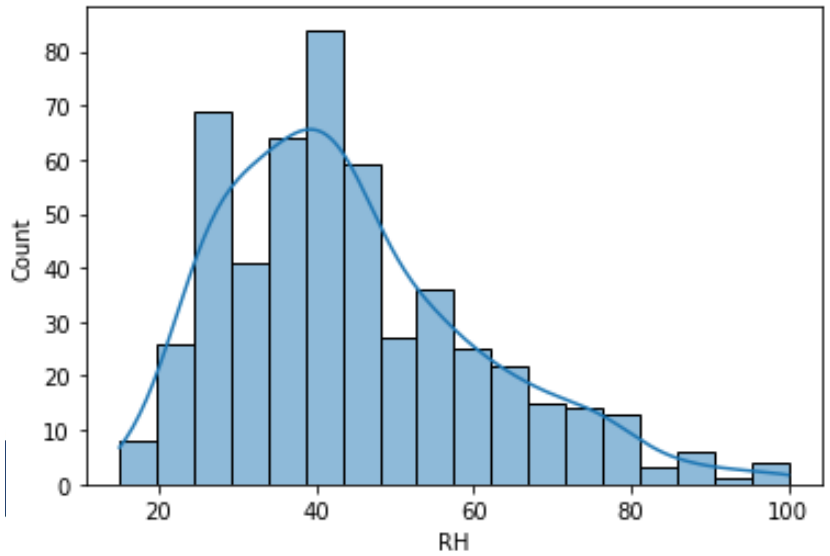


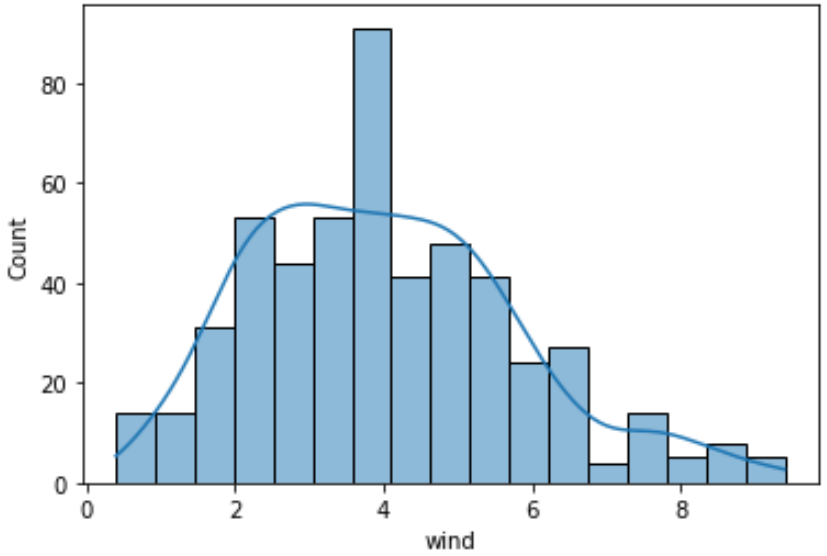
Bimodal
Right-skewed (positive skew)

Forest Fires Dataset: Frequency Distributions









Contingency Tables

Table 16: Contingency table for the pair 'Shape' x 'Severity' of the mammographic dataset

	Severity		
Shape	**Benign = 0**	**Malignant = 1**	**Total**
Round = 1	158	32	190
Oval = 2	149	31	180
Lobular = 3	39	42	81
Irregular = 4	81	298	379
Total	427	403	830

- There are 158 women with the mass shape 'Round' that present a 'Benign' diagnosis, and 32 women with the mass shape 'Round' that present a 'Malignant' diagnosis…

Contingency Table and Frequency Distribution with AI

"*Draw the contingency table for variables 'Shape' vs 'Severity' of the mammographic dataset*"

- For the forestfires dataset, prompt:

"*Plot the frequency distribution for all variables of the attached forestfires dataset*"

In **Claude.ai** you may try: "*I want you to show the visuals. You can create an interactive artifact to show them.*"

Summary Measures

Central Tendency Measures

- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Median: central value.
- Mode: most frequent value.
- Midpoint: $(max + min)/2$

- Trimmed mean: $\bar{x} = \frac{1}{n_t}\sum_{i=1}^{n_t} x_i$
- Mean of a frequency distribution:

$$\bar{x} = \frac{\sum_{i=1}^{n}(f_i \cdot x_i)}{\sum_{i=1}^{n} f_i}$$

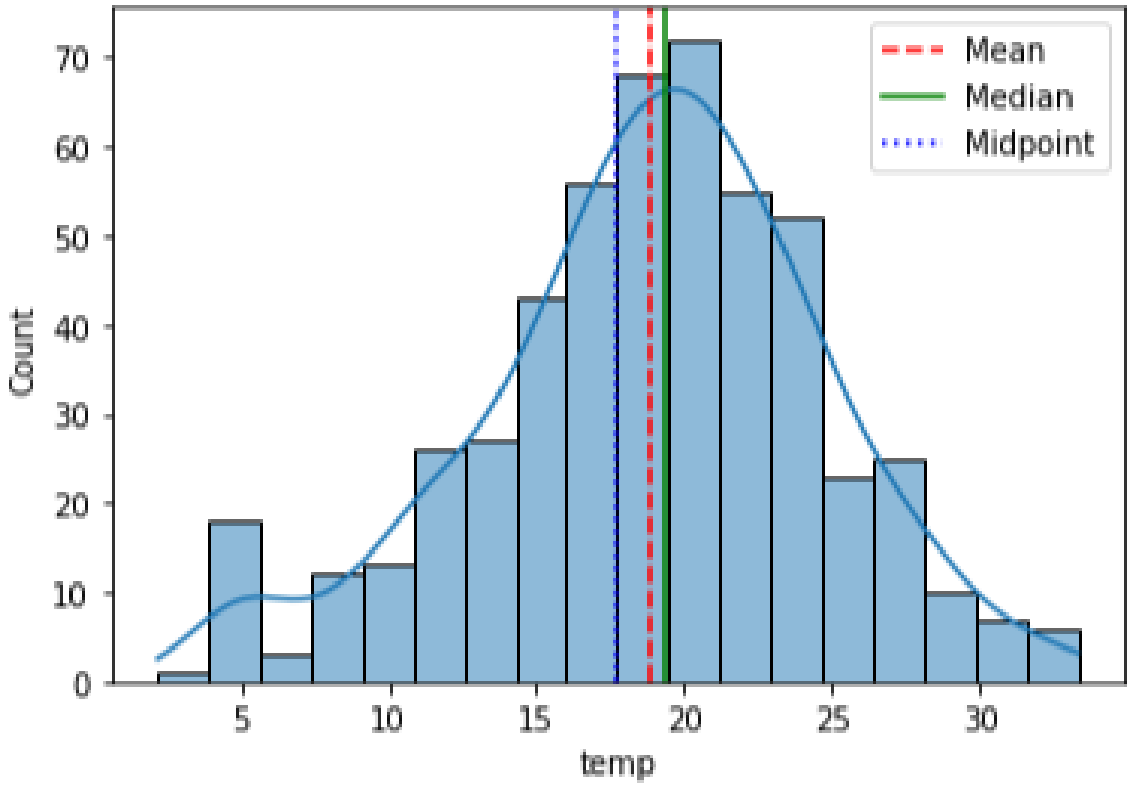
- Weighted average: $\bar{x} = \frac{\sum_{i=1}^{n}(w_i \cdot x_i)}{\sum_{i=1}^{n} w_i}$

Central Tendency Measures

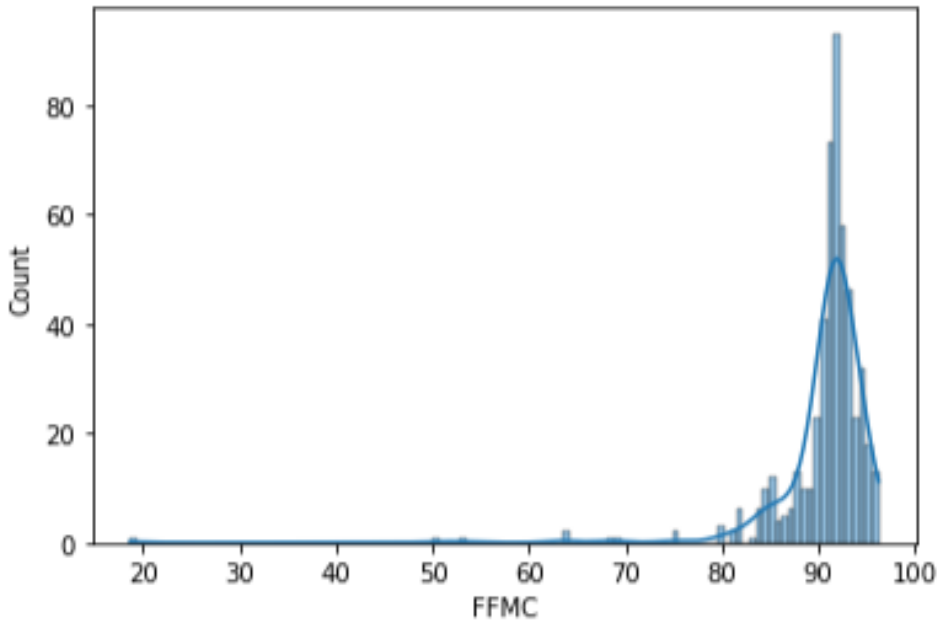
Table 17: Comparison of the different central tendency measures.

Measure	Sensitivity to Outliers	Computation	Existence	When to Use
Mean	High	All values	Always	Normal distributions; variation measures are needed; there are no extreme values
Median	No	All values	Always	Skewed distributions; there are extreme values
Mode	No	Some values	Not always or Multiple	Categorical data; when need to find the most frequent value
Midpoint	High	Extreme values	Always	Middle point is desired
Weighted Mean	High	All values	Always	Different values have different importance in the average
Trimmed Mean	No	Some values	Always	There are extreme values; skewed distributions

Central Tendency Measures



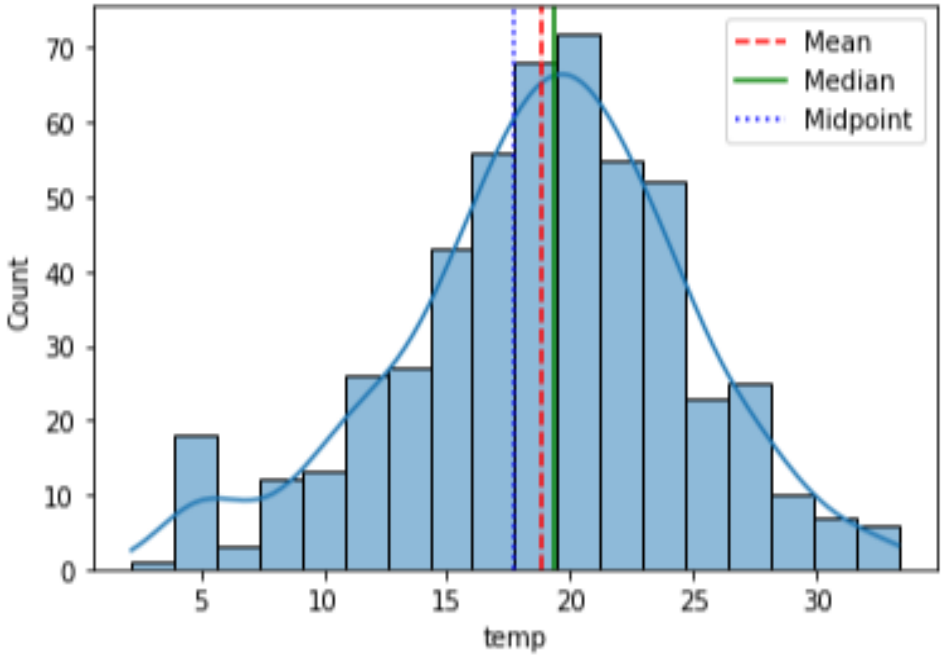
Central Tendency Measures: Interpretation



FFMC
Mean: 90.64
Median: 91.60
Midpoint: 57.45
Weighted Mean: 91.74
Trimmed Mean: 91.45

Interpretation:
The mean and median are very close, indicating a symmetric distribution. However, the midpoint is significantly lower than the other measures, suggesting that there might be a left skew and some outliers in the dataset. The trimmed mean is close to the mean and median, indicating that after removing some extreme values the distribution becomes more symmetric and with shorter tails.

Central Tendency Measures: Interpretation



temp
Mean: 18.89
Median: 19.30
Midpoint: 17.75
Weighted Mean: 21.70
Trimmed Mean: 19.09

Interpretation:
The mean, median and midpoint are close, indicating a symmetric distribution. The midpoint slightly smaller than the mean indicate a small left skew. The trimmed mean is close to the mean and median, indicating that after removing some extreme values the distribution becomes more symmetric and with shorter tails.

Variability Measures

- **Variability measures**, also called **dispersion measures**, provide numeric indices about the spread of the data, that is, the extent to which the values are spread out from the average.
- The most common variability measures are the **range**, **interquartile range**, **semi-interquartile range**, **variance**, **standard deviation**, and the **variation coefficient**.

Variability Measures

- Range: $R = x^L - x^l.$
- Interquartile range: $IQR = Q_3 - Q_1.$
- Semi-interquartile range: $sIQR = (Q_3 - Q_1)/2.$
- Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Standard deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Coefficient of variation: $CV = \frac{s}{\bar{x}}.100\%$

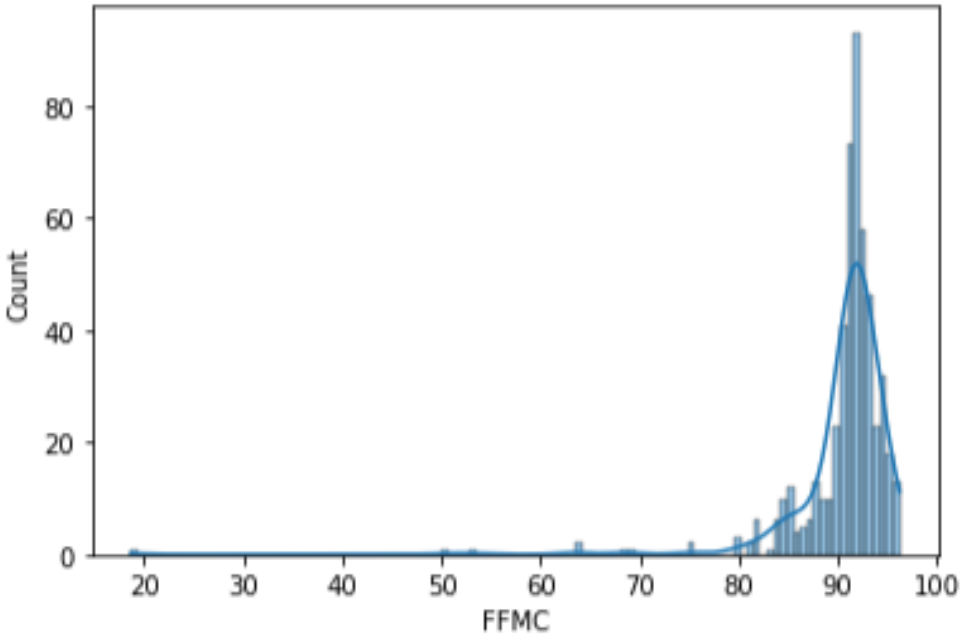
Variability Measures

Table 17: Comparison of the different variability measures.

Measure	Sensitivity to Outliers	Computation	Comments
Range	Yes	Extreme values	Sensitive to extreme values and does not take into account the data distribution
IQR	No	Half of the values	Not sensitive to extreme values and is suitable for skewed data
sIQR	No	Half of the values	Not sensitive to extreme values and is suitable for skewed data
Variance	Yes	All values	Sensitive to extreme values and with unit measured as the square of the x unit
Standard Deviation	Yes	All values	Sensitive to extreme values and measured in the same unit as x
Coefficient of Variation	Yes	All values	Sensitive to extreme values and suitable for data with a mean close to zero

Variability Measures: Forest Fires Data

```
*Variability Measures*
Range of variable FFMC: 77.50
IQR of variable FFMC: 2.70
sIQR of variable FFMC: 1.35
Variance of variable FFMC: 30.41
Standard deviation of variable FFMC: 5.51
Variation coefficient of variable FFMC: 6.08
```



- The difference between the highest and lowest 'FFMC' values is 77.50, which is its range.
- IQR = 2.70 means that 50% of the objects are within a 2.70 range of values.
- The sIQR represents the spread of the middle 50% of the 'FFMC' values around the median, and is equal to 1.35, which is half of the IQR value found.
- By knowing that 'FFMC' $\in$ [18.7, 96.20], a variance of 30.41 indicates that the 'FFMC' values are spread out over a wide range.
- A coefficient of variation (CV) equals to 6.08 is a relatively high value. A CV greater than 1 usually indicates high variability in the data.

Relative Position Measures

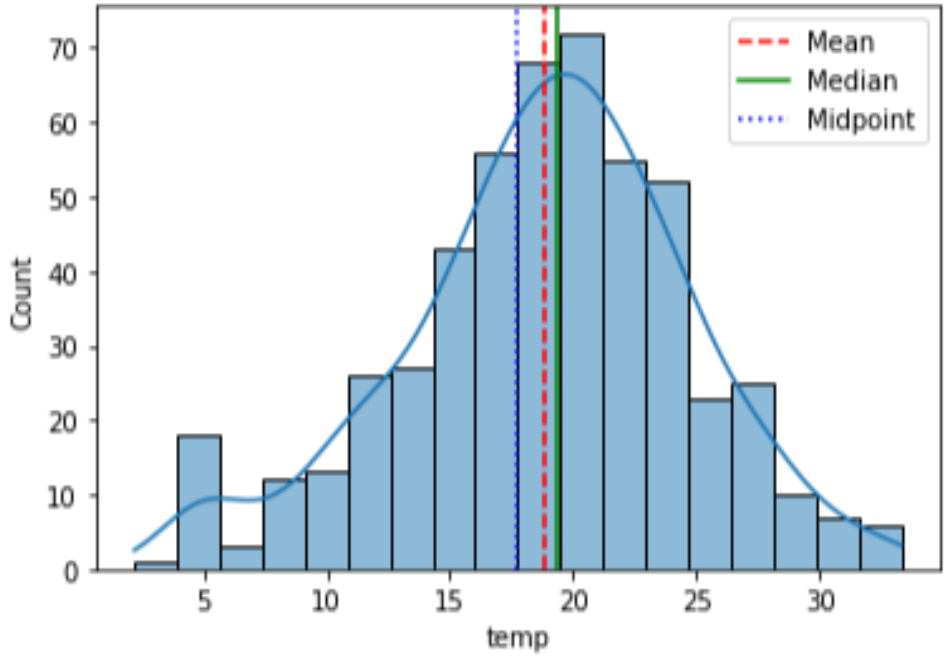
- There are situations in which you may want to know how a given value, e.g., a score, compares with others.
- For example, if you scored 6.3 in an exam, but the average score was 8.2, then you performed relatively poorly in relation to the group that took the exam.
- Measures that can be used to compare the relative performance of a value, that is, how it compares in relation to others, are called **relative position measures**.

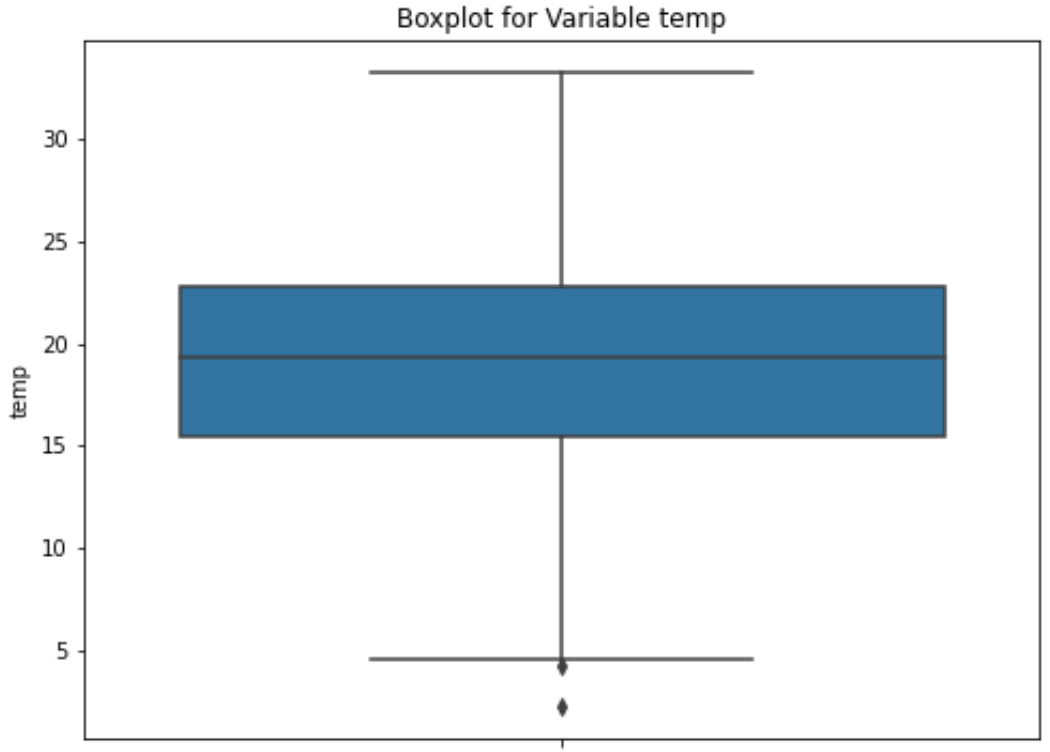
Relative Position Measures

- z-score $z = \frac{x - \bar{x}}{s}$

- Quantiles and Quartiles
  - First quartile (Q1): divides the 25% lowest ordered values from the remaining 75%.
  - Second quartile (Q2): divides the 50% lowest ordered values from the remaining 50%, that is, it is the same as the median.
  - Third quartile (Q3): divides the 75% lowest ordered values from the remaining 25%.

Relative Position Measures: Interpretation





Z-score for temp value 5: -2.39
Z-score for temp value 10: -1.53
Z-score for temp value 15: -0.67
Z-score for temp value 20: 0.19
Z-score for temp value 25: 1.05
Z-score for temp value 30: 1.91

Measures of Shape

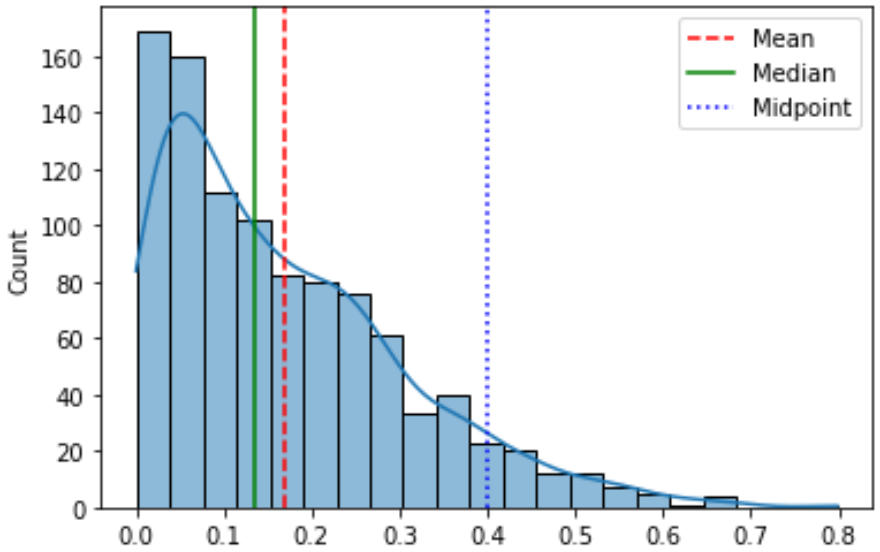
- The **shape** of a distribution brings important information about the underlying data, such as where the data are accumulated, the presence or absence of outliers, and if the distribution is more or less **skewed**, and more or less **spread**.

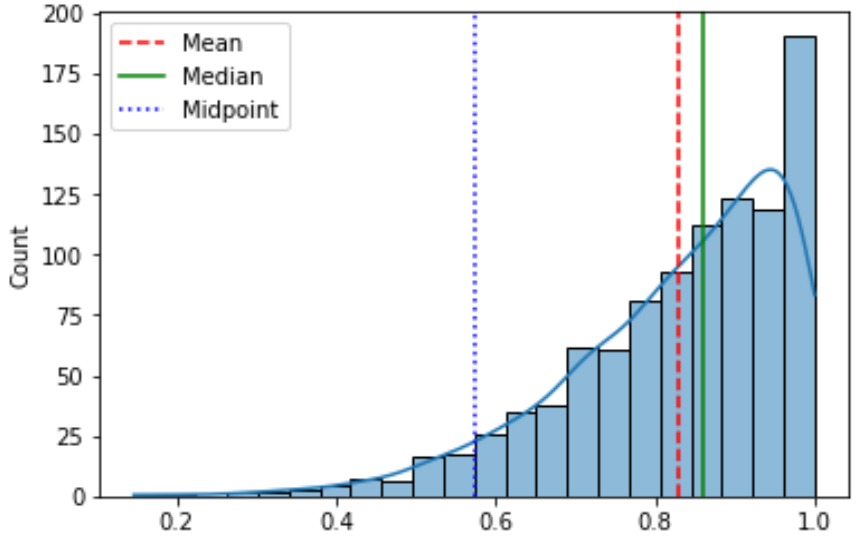
Measures of Shape: Skewness

- **Skewness** is a measure of the asymmetry (lack of symmetry) of a distribution, allowing us to quantify the shape, in terms of direction and length, of its tail.
- A **positively skewed**, or **right-skewed**, distribution has a long tail to its right, whilst a negatively skewed, or left-skewed, distribution has a long tail to its left.
- Therefore, **in positively skewed distributions the mean is usually greater than the median, which is greater than the mode**.
- By contrast, **in negatively skewed distributions, the mean is usually smaller than the median, which is smaller than the mode**.

Measures of Shape: Skewness

- Fischer-Pearson Skewness Coefficient $\gamma = \frac{m_3}{m_2^{3/2}}, \quad m_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^i\,(x_i - \bar{x})^i$





Right-skewed distribution: Mean, median and midpoint: 0.17 0.13 0.40
Skewness (Fischer-Pearson Coefficient): 1.06

Left-skewed distribution: Mean, median and midpoint: 0.83 0.86 0.57
Skewness (Fischer-Pearson Coefficient): -1.14

Measures of Shape: Kurtosis

- **Kurtosis** is a measure of the taildness of the distribution, and is useful to analyze the peak, the tails of the curve, and the presence of outliers.
- The distribution can have a steeper or flatter peak, and a longer or shorter tail.

Measures of Shape: Kurtosis

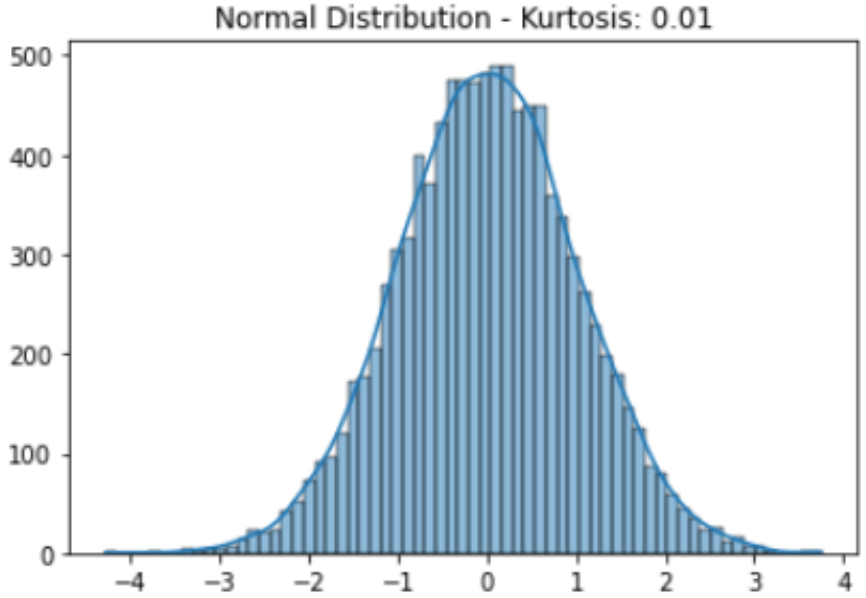
- Excess Kurtosis $\beta = \frac{m_4}{(s^2)^2} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{(s^2)^2} - 3$

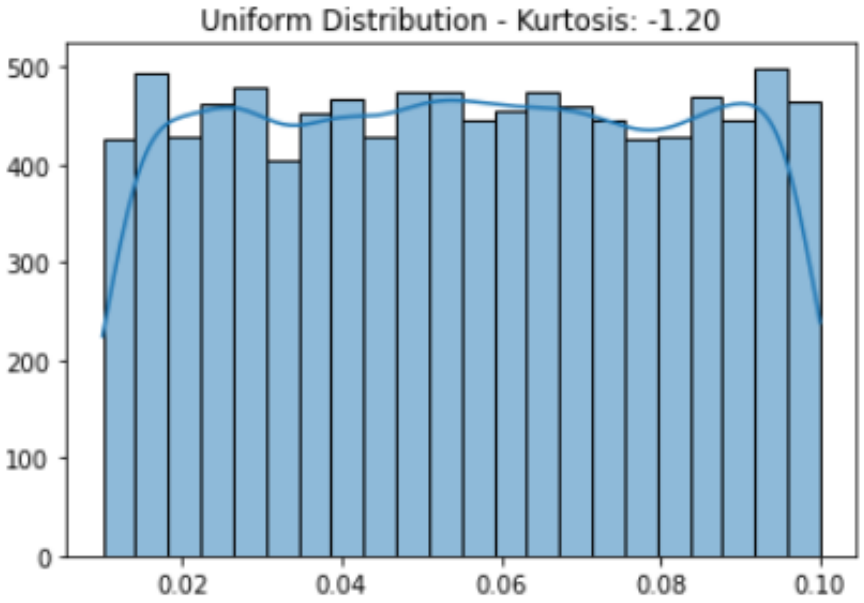
Platykurtic distribution: *negative kurtosis* (usually has a flatter peak and shorter tail, meaning that it produces less extreme values than normal distributions).

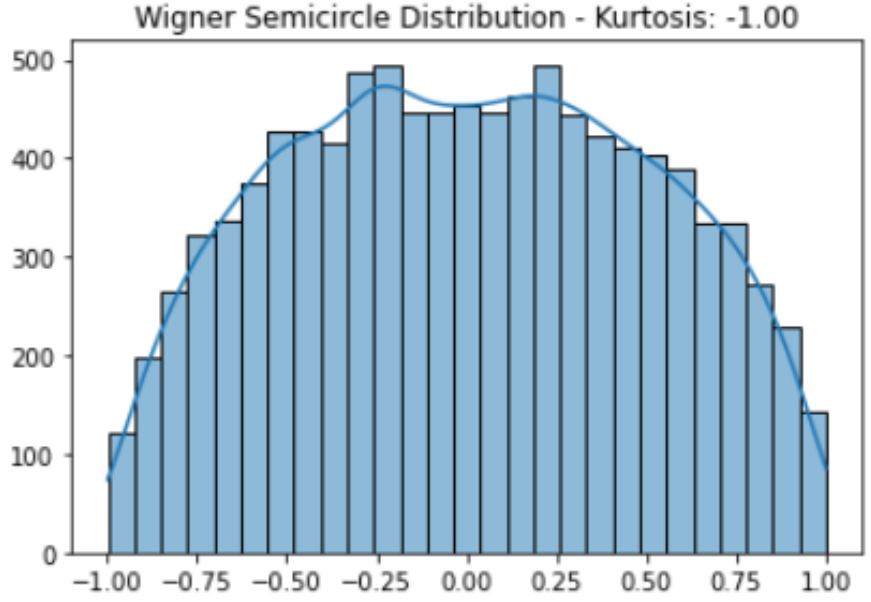
Leptokurtic distribution: *positive kurtosis* (usually has a longer tail and steeper peak, producing more outliers than a normal distribution).

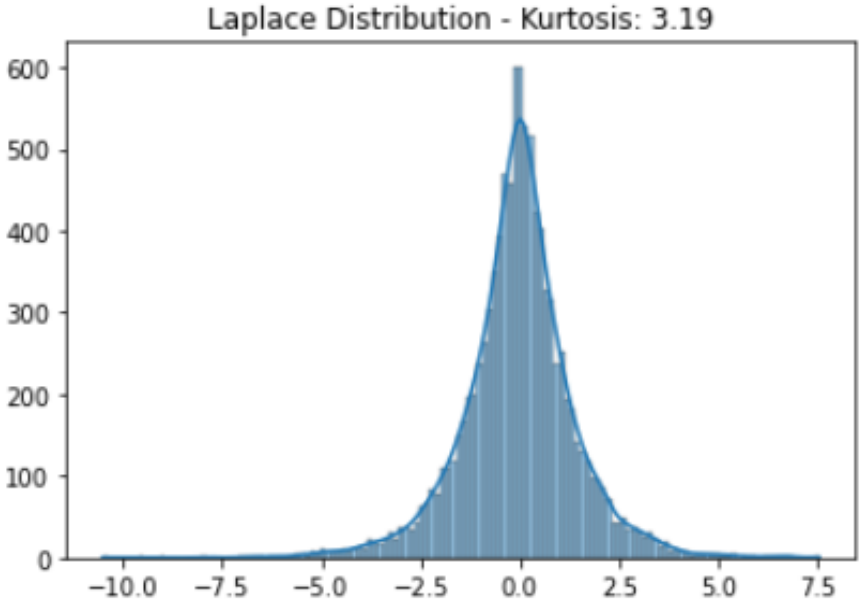
Mesokurtic distribution: close to zero or zero kurtosis, like the normal distribution.

Measures of Shape: Kurtosis







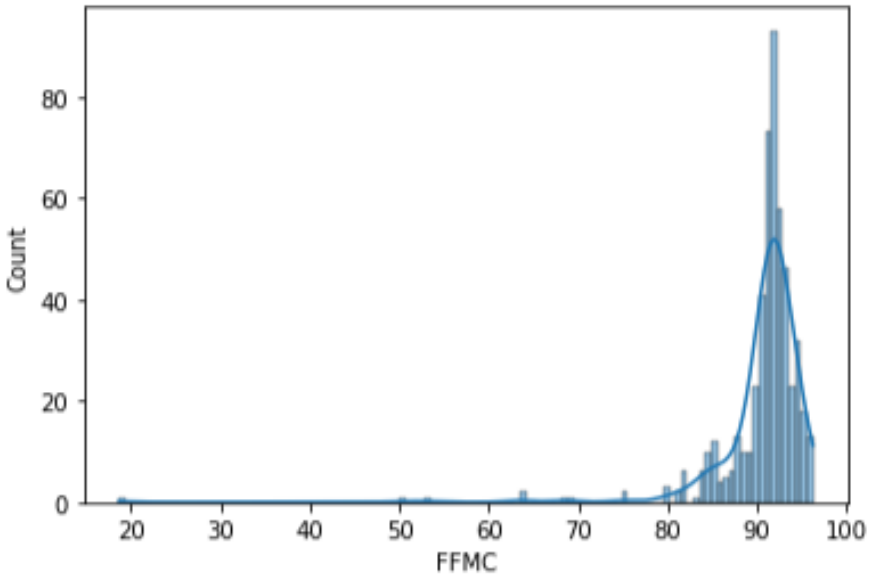


Measures of Shape: Forest Fires Data

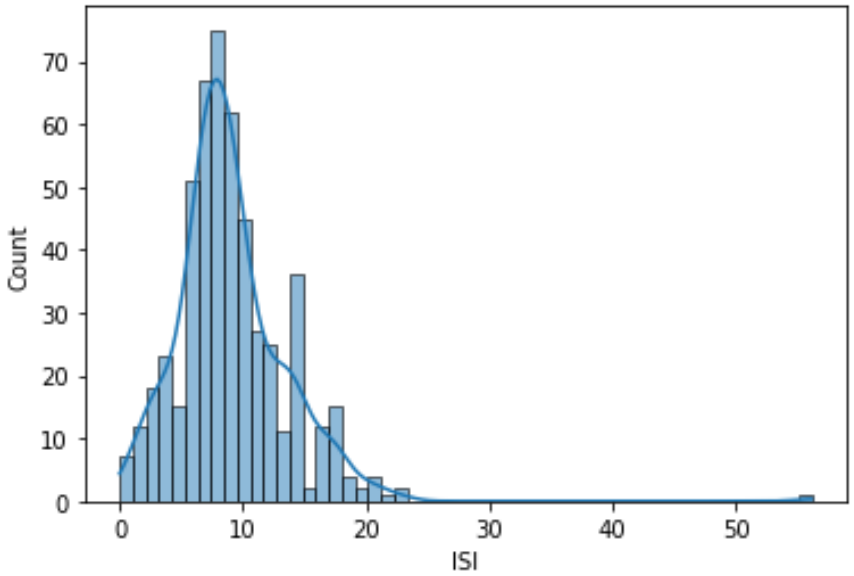
FFMC Skewness: -6.58
FFMC Kurtosis: 67.07 (Leptokurtic)
DMC Skewness: 0.55
DMC Kurtosis: 0.20 (Leptokurtic)**
DC Skewness: -1.10
DC Kurtosis: -0.25 (Platykurtic)**
ISI Skewness: 2.54
ISI Kurtosis: 21.46 (Leptokurtic)

temp Skewness: -0.33
temp Kurtosis: 0.14 (Leptokurtic)
RH Skewness: 0.86
RH Kurtosis: 0.44 (Leptokurtic)
wind Skewness: 0.57
wind Kurtosis: 0.05 (Leptokurtic)****
rain Skewness: 19.82
rain Kurtosis: 421.30 (Leptokurtic)

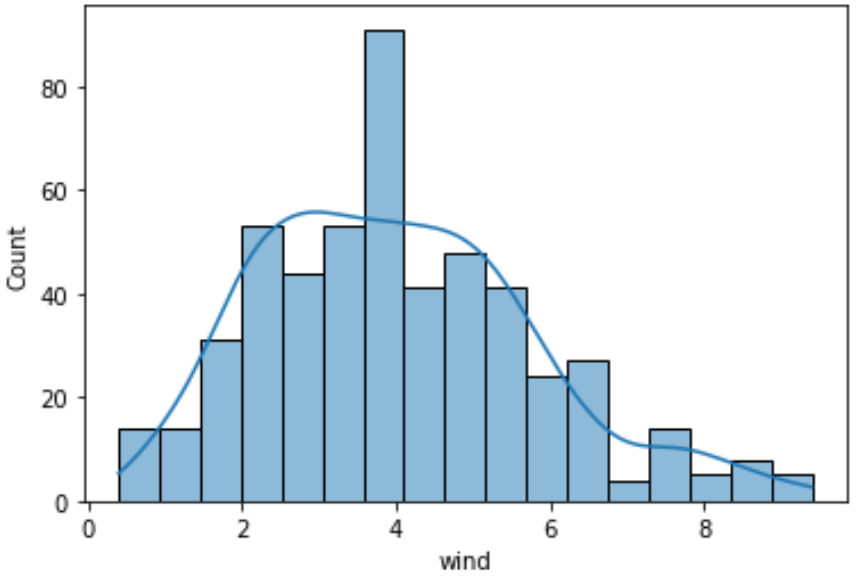
Measures of Shape: Interpretation



FFMC Skewness: -6.58 (left skew)
FFMC Kurtosis: 67.07 (Leptokurtic)

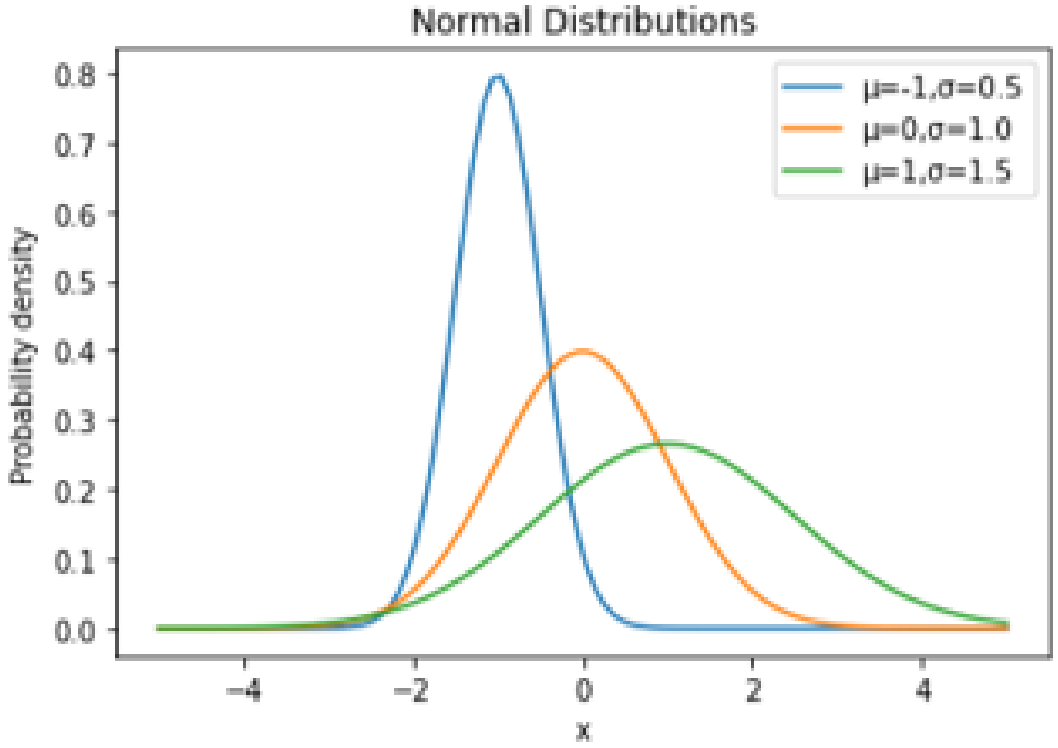


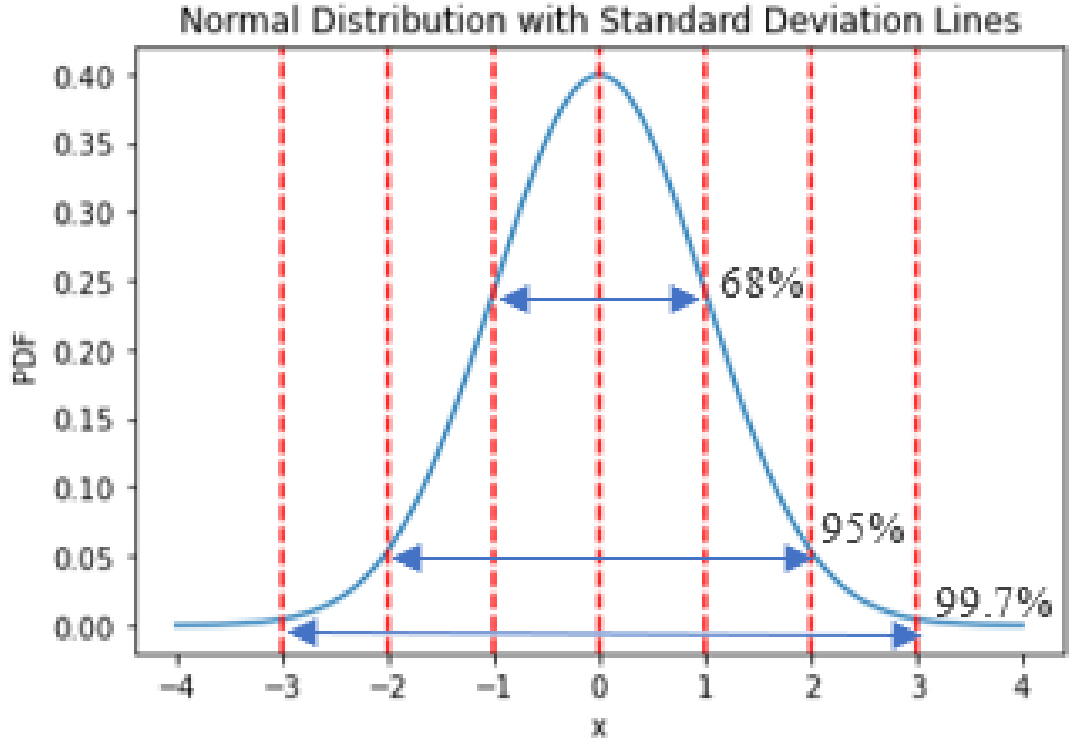
ISI Skewness: 2.54 (right skew)
ISI Kurtosis: 21.46 (Leptokurtic)



wind Skewness: 0.57 (right skew)
wind Kurtosis: 0.05 (Mesokurtic)

The Normal Distribution





The *empirical rule* 68-95-99.7 is known, as follows:

- Approximately 68% of all values are within one standard deviation from the mean;
- Approximately 95% of all values are within two standard deviations from the mean;
- Approximately 99.7% of all values are within three standard deviations from the mean.

Measures of Association

- Covariance: $cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- Covariance matrix: $\Sigma_{ij} = cov(x_i, x_j),\ \forall i,j$

	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain
X	5.35	1.54	-0.27	-7.17	-49.31	0.07	-0.69	3.22	0.08	0.04
Y	1.54	1.51	-0.31	0.61	-30.87	-0.14	-0.17	1.25	-0.04	0.01
FFMC	-0.27	-0.31	30.47	135.27	452.59	13.38	13.83	-27.11	-0.28	0.09
DMC	-7.17	0.61	135.27	4101.95	10838.50	89.10	174.64	77.12	-12.09	1.42
DC	-49.31	-30.87	452.59	10838.50	61536.84	259.19	714.75	-158.64	-90.43	2.63
ISI	0.07	-0.14	13.38	89.10	259.19	20.79	10.44	-9.86	0.87	0.09
temp	-0.69	-0.17	13.83	174.64	714.75	10.44	33.72	-49.97	-2.36	0.12
RH	3.22	1.25	-27.11	77.12	-158.64	-9.86	-49.97	266.26	2.03	0.48
wind	0.08	-0.04	-0.28	-12.09	-90.43	0.87	-2.36	2.03	3.21	0.03
rain	0.04	0.01	0.09	1.42	2.63	0.09	0.12	0.48	0.03	0.09

Measures of Association

- It is now time to measure the association between two variables and there are different measures of association (MAs) to calculate the strength and direction of the relationship between two variables.
- Most MAs are within the [–1, 1] range, where –1 indicates a perfect negative correlation, 0 indicates no correlation, and +1 indicates a perfect positive correlation.
- Two variables are negatively correlated when one variable increases and the other decreases, and vice-versa; they are positively correlated if they either increase or decrease simultaneously; and when there is no correlation it means that the change of one variable does not exert any influence in the change of the other.

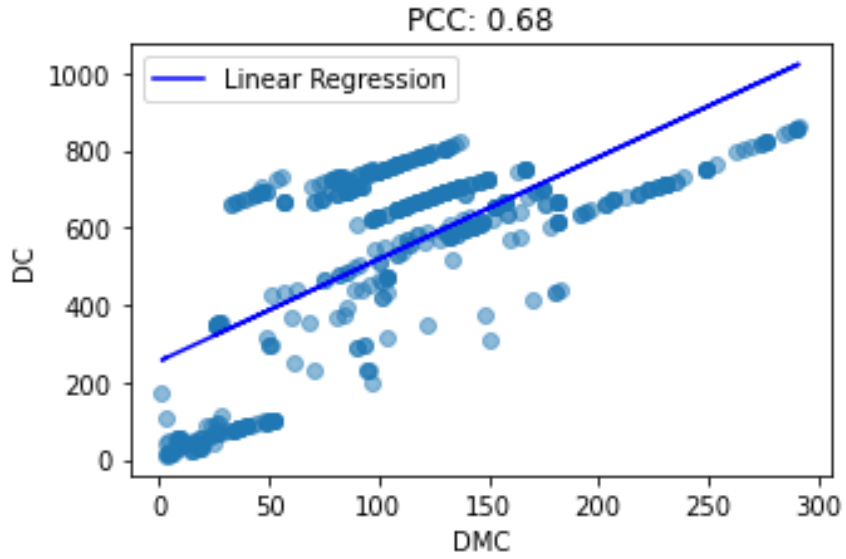
Measures of Association

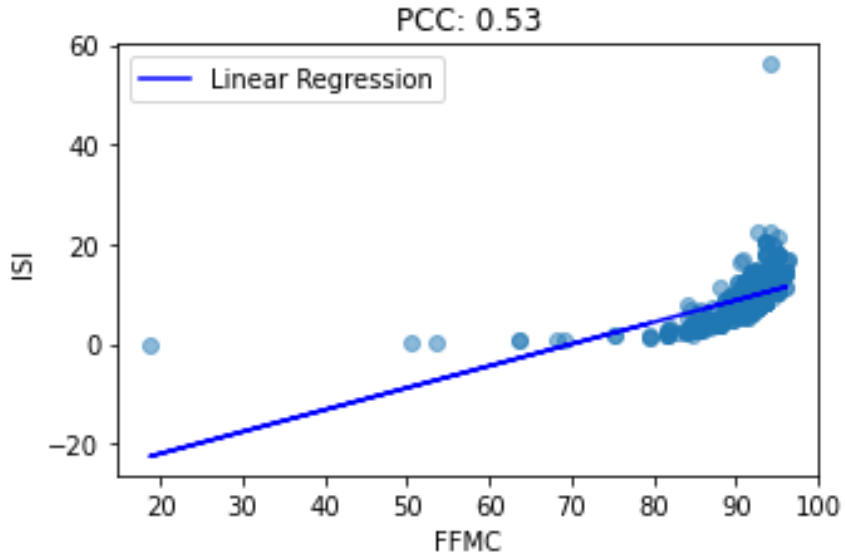
- Correlation: $\rho(x,y) = \frac{cov(x,y)}{\sigma(x).\sigma(y)}$ $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2} \cdot \sqrt{\sum_{j=1}^{n}(y_j - \bar{y})^2}}$

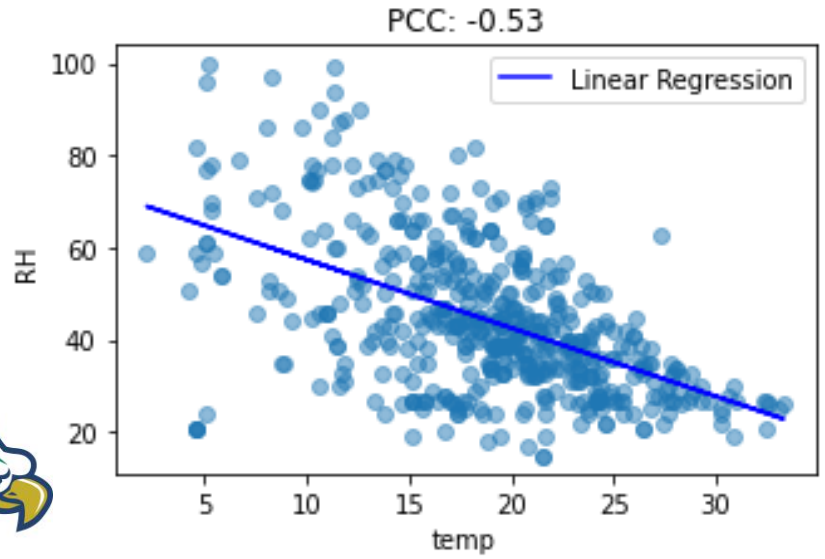
**Forest Fires Dataset: PCC
Pearson Correlation Coefficient (PCC)

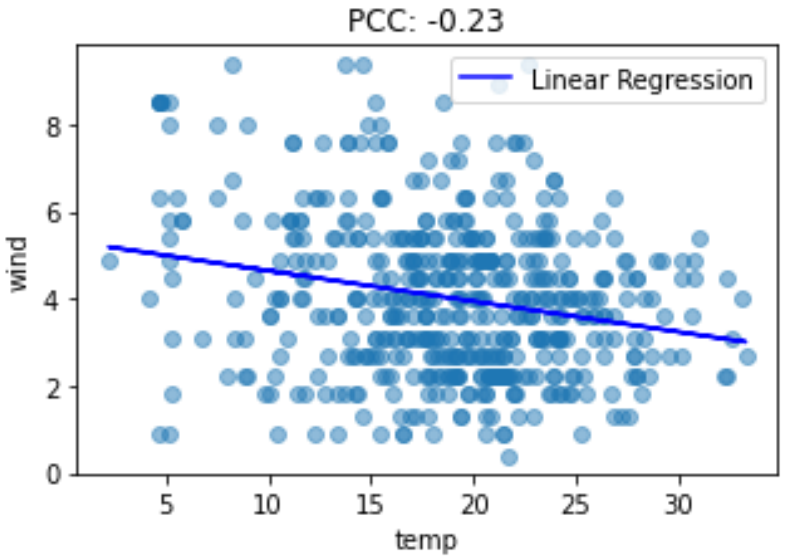
	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	1.00	0.54	-0.02	-0.05	-0.09	0.01	-0.05	0.09	0.02	0.07	0.06
Y	0.54	1.00	-0.05	0.01	-0.10	-0.02	-0.02	0.06	-0.02	0.03	0.04
FFMC	-0.02	-0.05	1.00	0.38	0.33	0.53	0.43	-0.30	-0.03	0.06	0.04
DMC	-0.05	0.01	0.38	1.00	0.68	0.31	0.47	0.07	-0.11	0.07	0.07
DC	-0.09	-0.10	0.33	0.68	1.00	0.23	0.50	-0.04	-0.20	0.04	0.05
ISI	0.01	-0.02	0.53	0.31	0.23	1.00	0.39	-0.13	0.11	0.07	0.01
temp	-0.05	-0.02	0.43	0.47	0.50	0.39	1.00	-0.53	-0.23	0.07	0.10
RH	0.09	0.06	-0.30	0.07	-0.04	-0.13	-0.53	1.00	0.07	0.10	-0.08
wind	0.02	-0.02	-0.03	-0.11	-0.20	0.11	-0.23	0.07	1.00	0.06	0.01
rain	0.07	0.03	0.06	0.07	0.04	0.07	0.07	0.10	0.06	1.00	-0.01
area	0.06	0.04	0.04	0.07	0.05	0.01	0.10	-0.08	0.01	-0.01	1.00

Linear Regression + Correlation

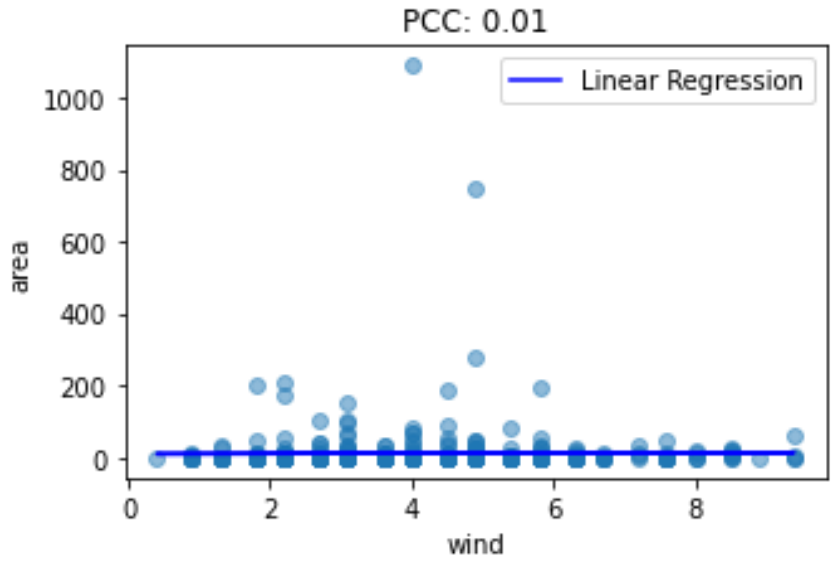


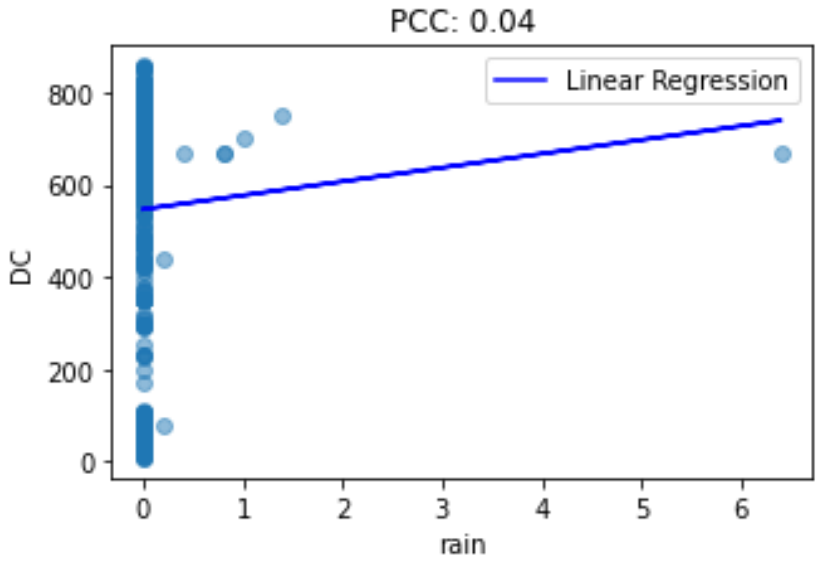






Linear Regression + Correlation





Linear regression: $y = a + b.x$

Descriptive Analysis with AI

- Prompt:

“For the forestfires dataset, do:

1) Create a table with the central tendency, variability measures, and the measures of shape of the variables in the forest fires dataset.

2) Plot the correlation matrix of all numeric variables.

3) Plot the dispersion graphs of some pairs of numeric variables involving ‘temp’ and ‘ffmc’ with the best linear regressors showing their correlation trend.”

Leandro Nunes de Castro

ldecastrosilva@fgcu.edu