Exploratory Data Analysis (EDA) and Visualization

Data Mining

Edgar Roman-Rangel.
edgar.roman@itam.mx

Department of Computer Science.
Instituto Tecnológico Autónomo de México, ITAM.

Outline

EDA

Exploratory data analysis (EDA).

- Consists in getting an overall picture of our data (exploring and investigating).
- It focuses on looking at the "position" and dispersion of each variable, and ultimately at its distribution.
- It also looks at the correlations between pairs of variables (univariate, bivariate, multivariate).
- It might as well include data cleaning.

The general goal of an EDA is: to get an overview of our data, and to identify useful and useless variables.

Centrality

Location - typical value of a variable.

- mean,
- median,
- mode,
- expected value,
- weighted mean,
- trimmed mean.

Dispersion

Variability of our data around its location.

- standard deviation,
- variance,
- mead absolute deviation (MAD).
- mead absolute deviation from the median.
- entropy,
- range,
- quantiles,
- interquartile range.

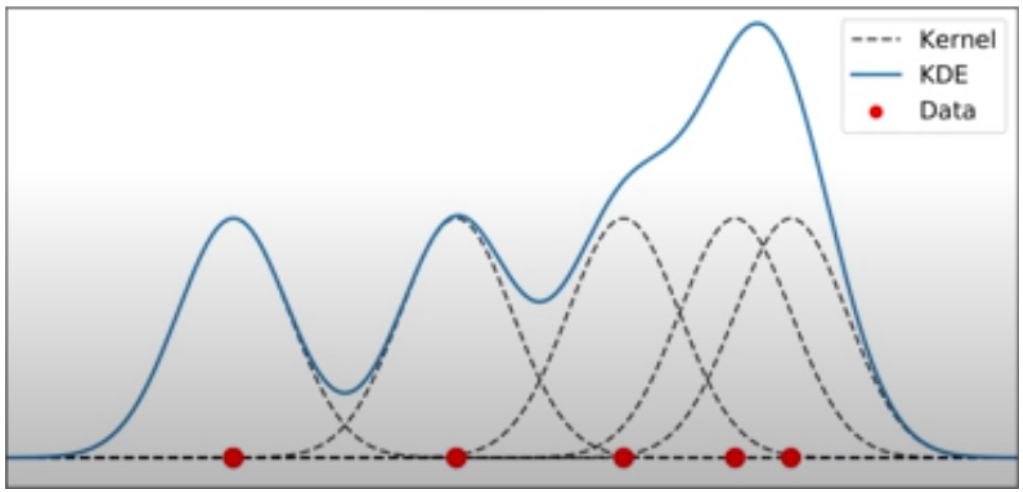
Distributions

More generally, we could be interested in being able to describe a variable using its distribution, either parametric or non-parametric.

- parametric: binomial, Gaussian, uniform, other.
- non-parametric: kernel density estimation and visualization tools.
- visualization tools: boxplots, frequency tables, histograms, density plots, scatter plots, etc.

Kernel Density Estimation

We can estimate (approximate) the probability density function of a dataset by the sum of **base density functions** (kernels), located at each sample point.



- which standard deviation?,
- which kernel function (Gaussian, triangle, square).

Other analyses

More generally, we could be interested in being able to describe a variable using its distribution, either parametric or non-parametric.

- minimum and maximum,
- biases and skewness,
- anomalies,
- temporal patterns,
- correlations and collinearities.

Binary, categorical, and ordinal variables

There are some tools that are preferred for categorical-like data. For instance,

- Frequency table for categorical or binned real-valued variables,
- histograms (including empty bins),
- bar or pie charts,
- mode and expected value.

Correlations

Often, we would like to explore more that one variable at a time,

- Correlation between pairs of independent variables,
- correlation between each independent variable and the target variable.

Depending on the type of data, correlation can be estimated by,

- Spearman, Pearson,
- mutual information,
- ANOVA test.

Visualization of correlations

Some visualization tools for this type of analysis are,

- Correlation matrix (with p-values),
- scatter plots,
- hexagonal binning chart for numeric data,
- contour plots for numeric data,
- contigency table (joint counting of possible outcome) for categorical data,
- boxplots with each box = possible outcome for categorical and numeric data.
- violin plot (extension of boxplot showing density).

Outline

Introduction

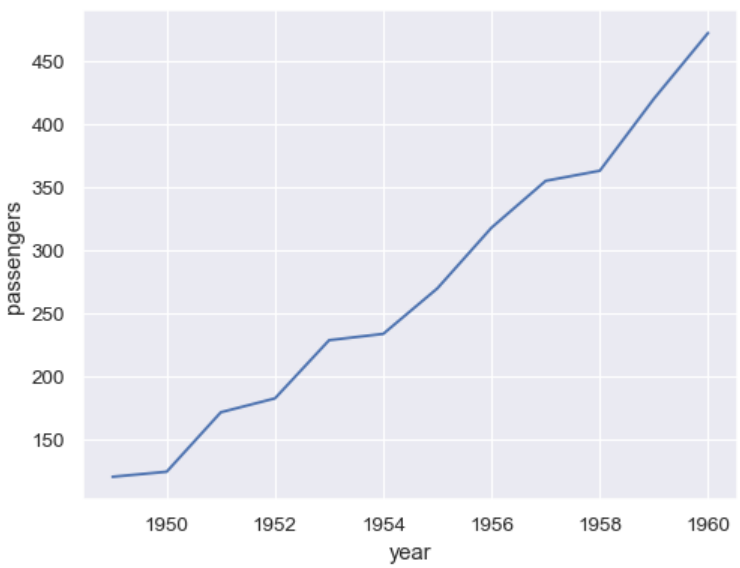
Plotting allows to present informative summaries of our data.

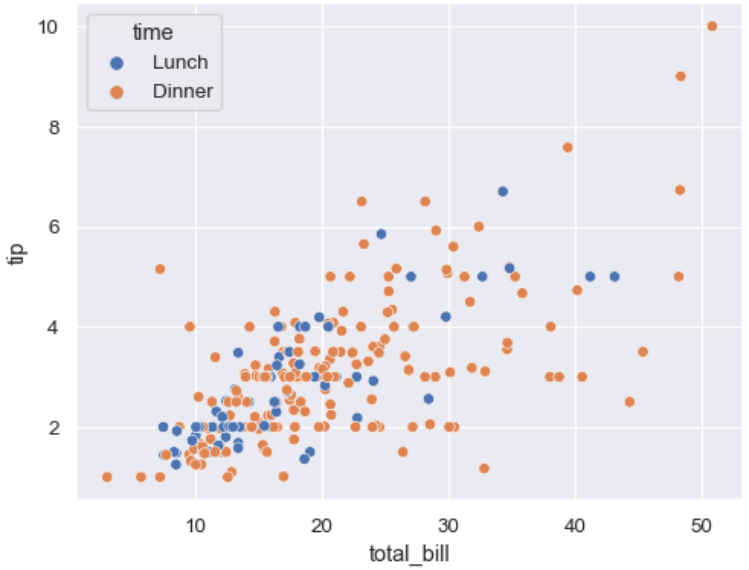
Plots and charts are a good tool to help the EDA.

Some common plot types include,

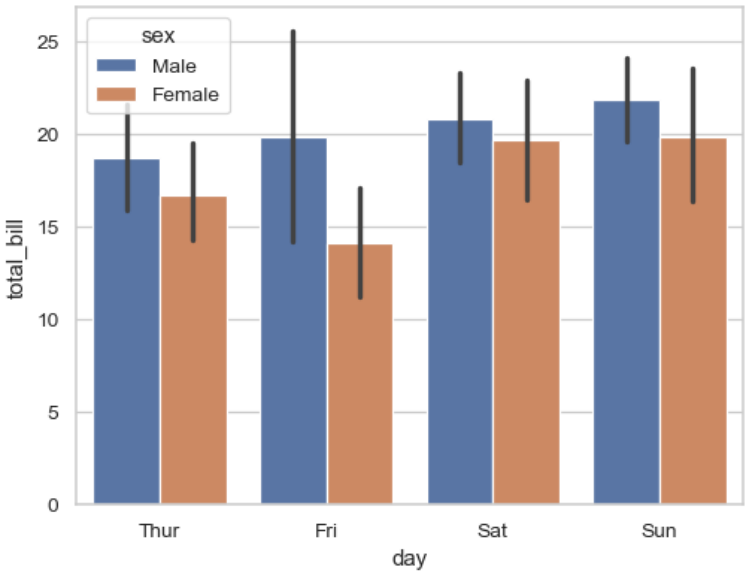
- line, scatter, bar, pie,
- histogram, boxplot, heatmaps,
- scatter matrices, correlation matrices,
- tree maps, network graphs.

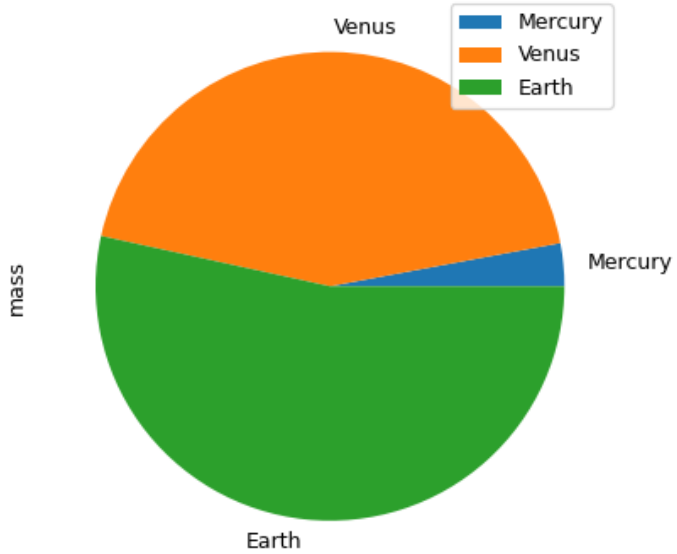
Line and scatter





Bar and pie


sex
Male
Female
total_bill
0
5
10
15
20
25
Thur
Fri
Sat
Sun
day

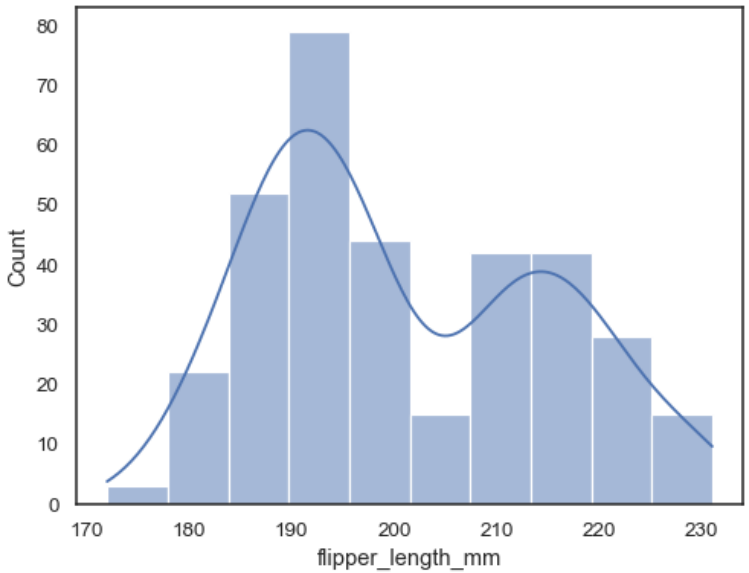

Mercury
Venus
Earth
Venus
Mercury
Earth
mass

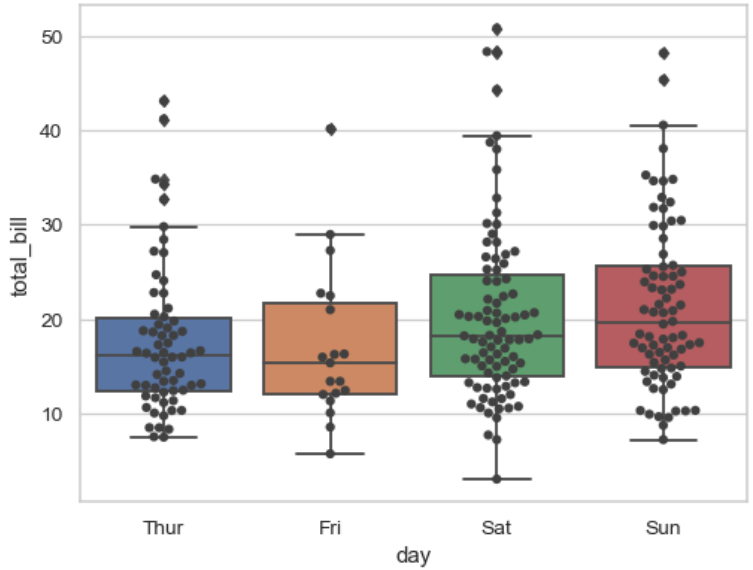
Frequency table

Table 1-5. A frequency table of population by state

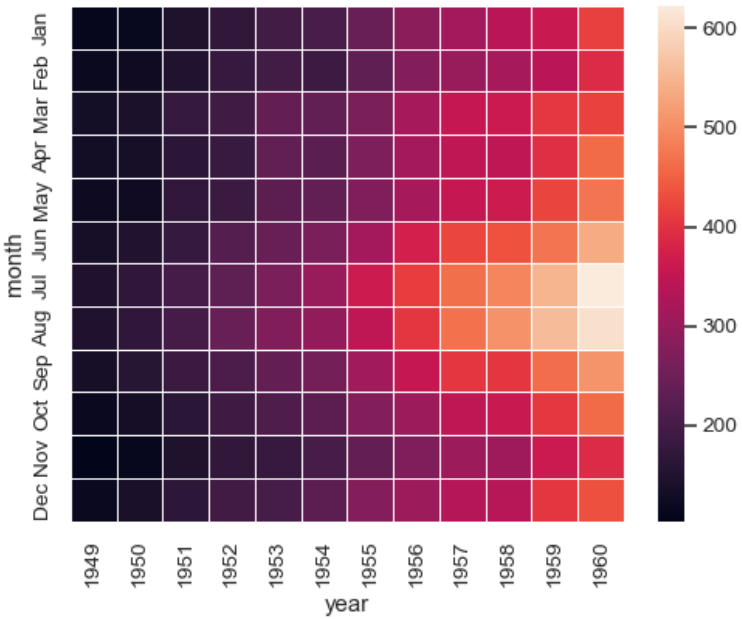
BinNumber	BinRange	Count	States
1	563,626–4,232,658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV,NM,NV,UT,KS,AR,MS,IA,CT,OK,OR
2	4,232,659–7,901,691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA,WA
3	7,901,692–11,570,724	6	VA,NJ,NC,GA,MI,OH
4	11,570,725–15,239,757	2	PA,IL
5	15,239,758–18,908,790	1	FL
6	18,908,791–22,577,823	1	NY
7	22,577,824–26,246,856	1	TX
8	26,246,857–29,915,889	0	
9	29,915,890–33,584,922	0	
10	33,584,923–37,253,956	1	CA

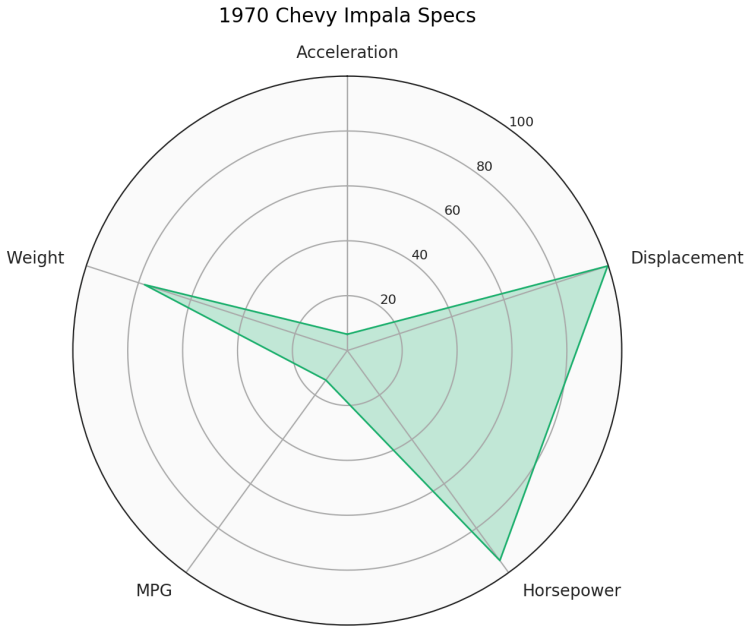
Histogram and boxplot



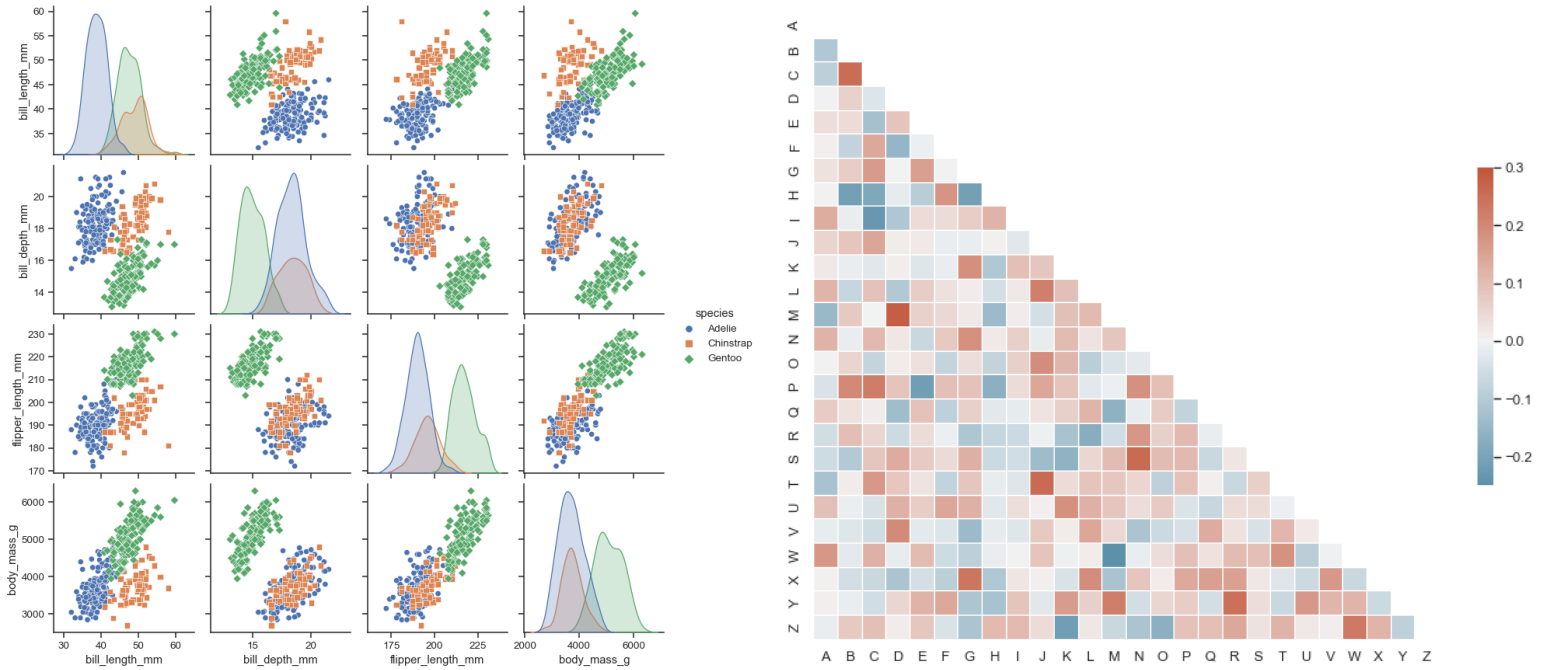


Heatmap and radar

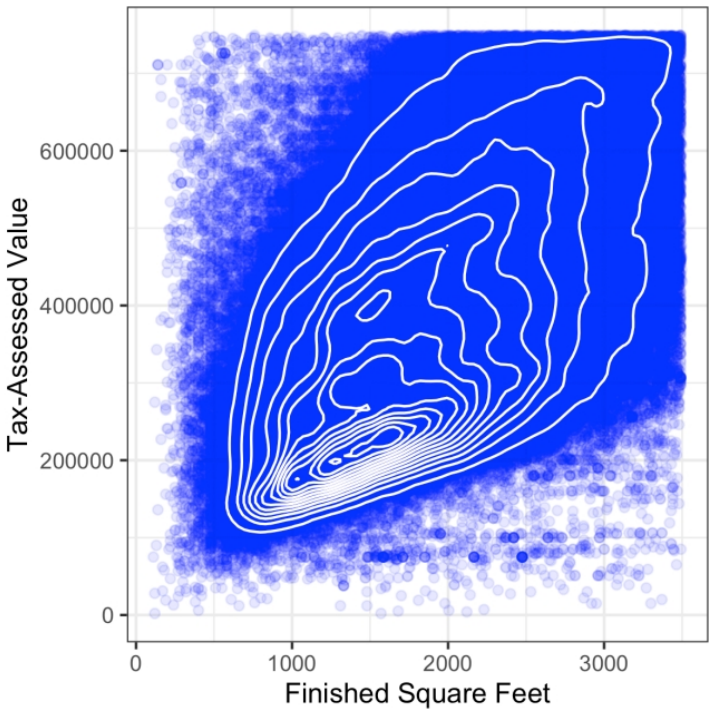


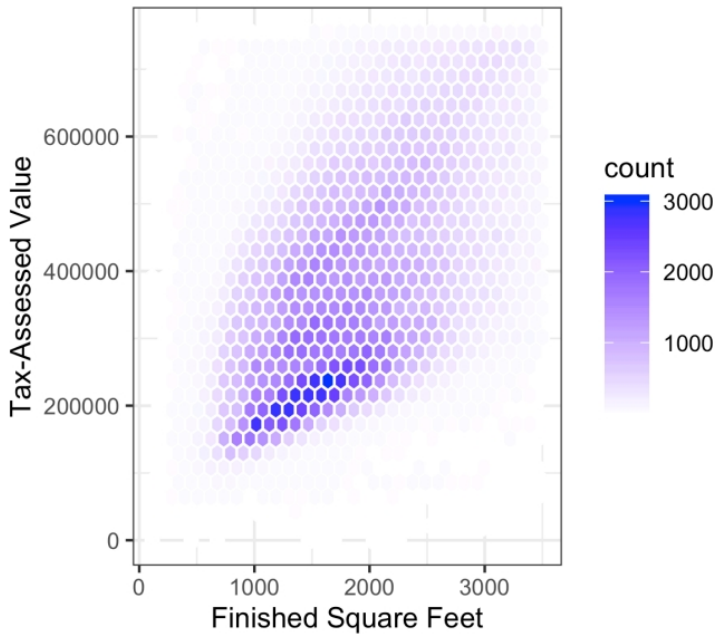


Correlation

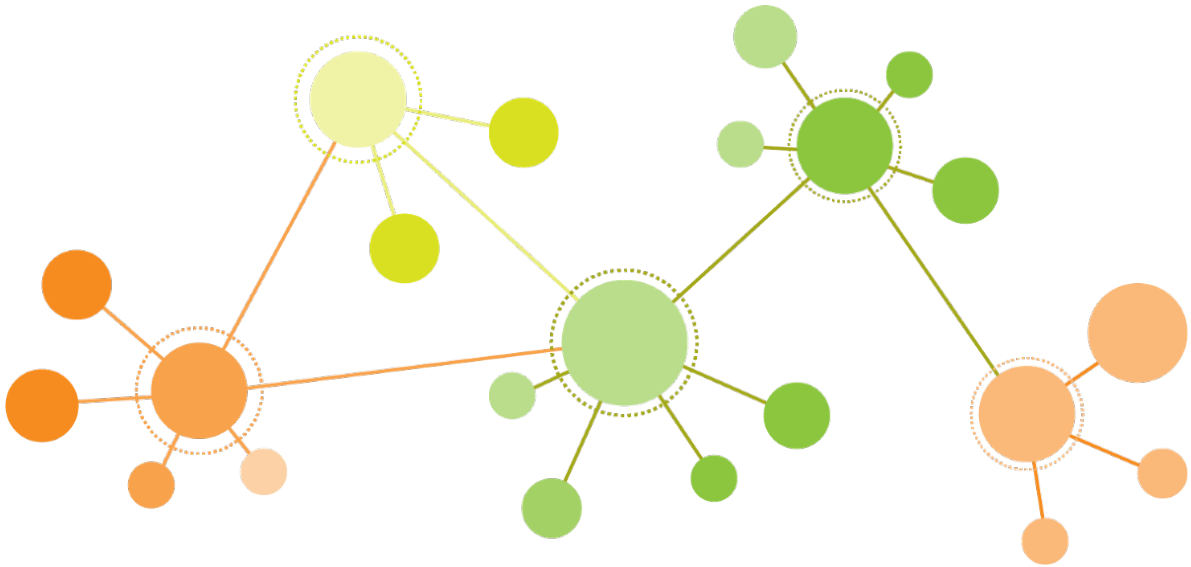


Contour plots and Hexagonal binning

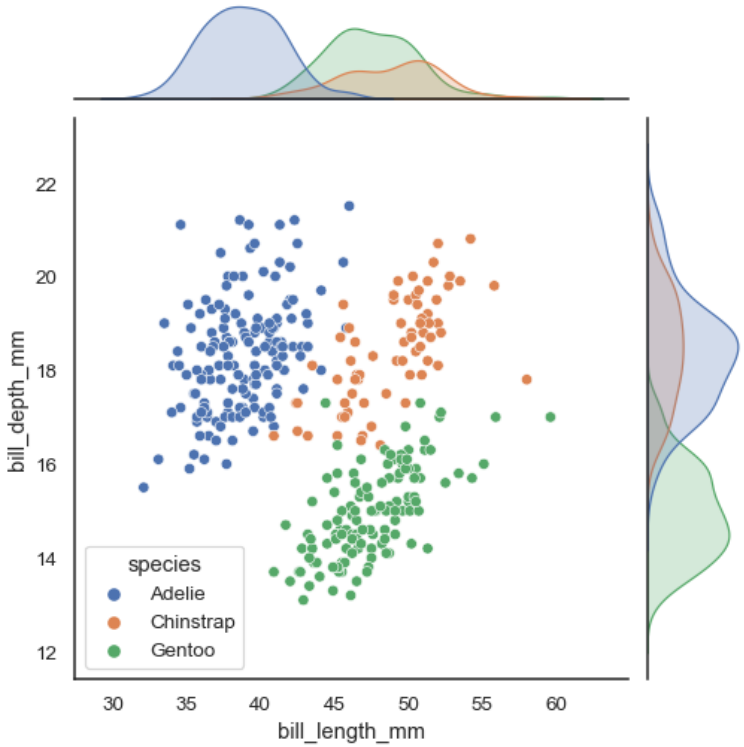


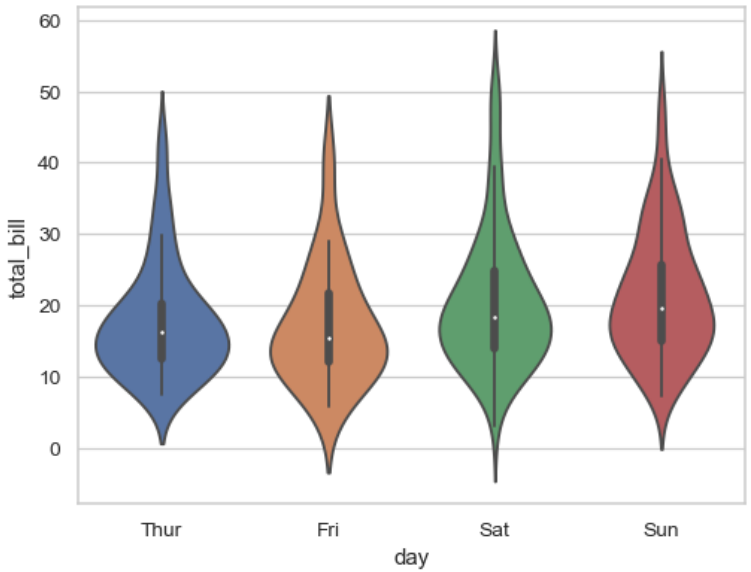


Network graphs



Others


bill_depth_mm
bill_length_mm
species
Adelie
Chinstrap
Gentoo
22
20
18
16
14
12
30
35
40
45
50
55
60


total_bill
day
Thur
Fri
Sat
Sun
60
50
40
30
20
10
0

References

Jiawei Han, Micheline Kamber, Jian Pei.
"Data Mining Concepts and Techniques". Ch 2.2, and Ch 3.2.
Elsevier. 2012.

EDA example.
https://medium.datadriveninvestor.com/step-by-step-exploratory-data-analysis-of-titanic-dataset-2d0fb09b0e86

Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, Kenneth C.Lichtendahl, Jr.
"Data Mining for Business Analytics: Concepts, Techniques, and Applications in R". Ch 3.
Wiley. 2018.

Matplotlib.
https://matplotlib.org/

Seaborn.
https://seaborn.pydata.org/index.html

Q&A

Thank you!

edgar.roman@itam.mx