UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

GiAR
GRUPO DE INTELIGENCIA ARTIFICIAL Y ROBOTICA

Unidad 05:
Análisis exploratorio de un conjunto de datos

Aprendizaje Automático

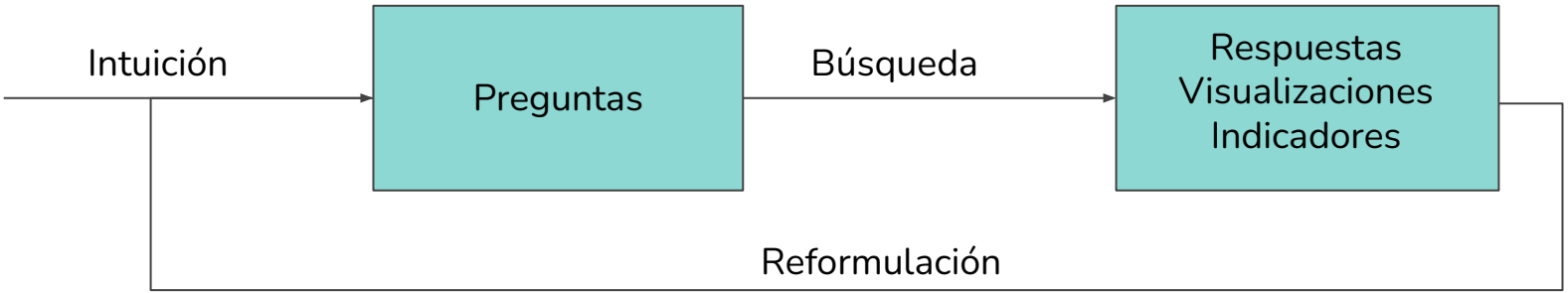
Docentes:
Diego P. Durante
ddurante@frba.utn.edu.ar

Ramiro Verrastro
ramiroverrastro@frba.utn.edu.ar

- Tema 1 - Introducción

Proceso de exploración

- Generar preguntas utilizando la intuición
- Buscar las respuestas mediante visualización u otros indicadores
- Reformular y generar nuevas preguntas utilizando lo aprendido



- Tema 1 - Introducción
- Tema 2 - Búsqueda de *insights*

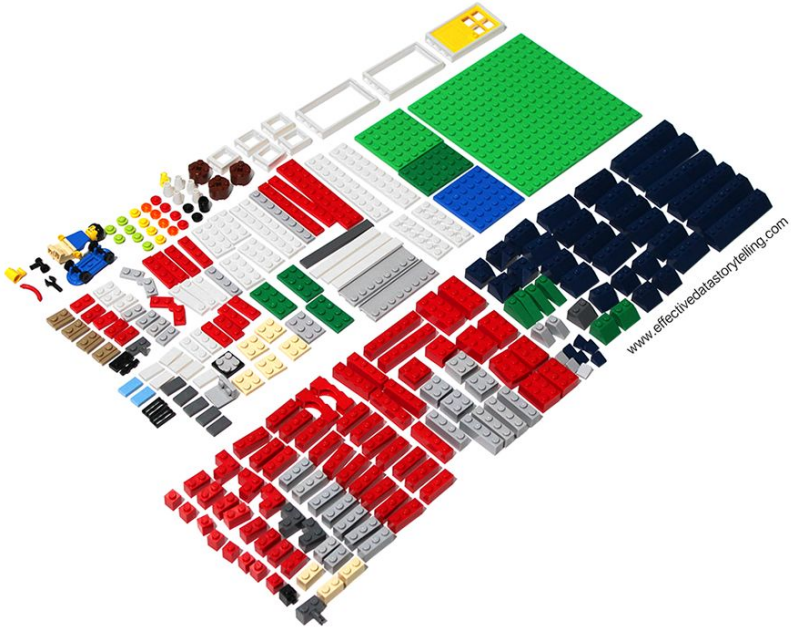
EDA

https://www.effectivedatastorytelling.com/post/a-deeper-dive-into-lego-bricks-and-data-stories

1 - Colección de datos

2 - Preparación de datos

3 - Visualización de datos



4 - Análisis de datos

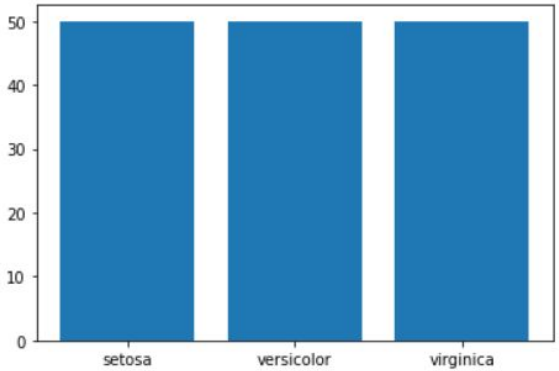
5 - Conclusiones desde datos

- Tema 1 - Introducción
- Tema 2 - Búsqueda de *insights*
- Tema 3 - Análisis de distribuciones

Distribución

Distribución de clases

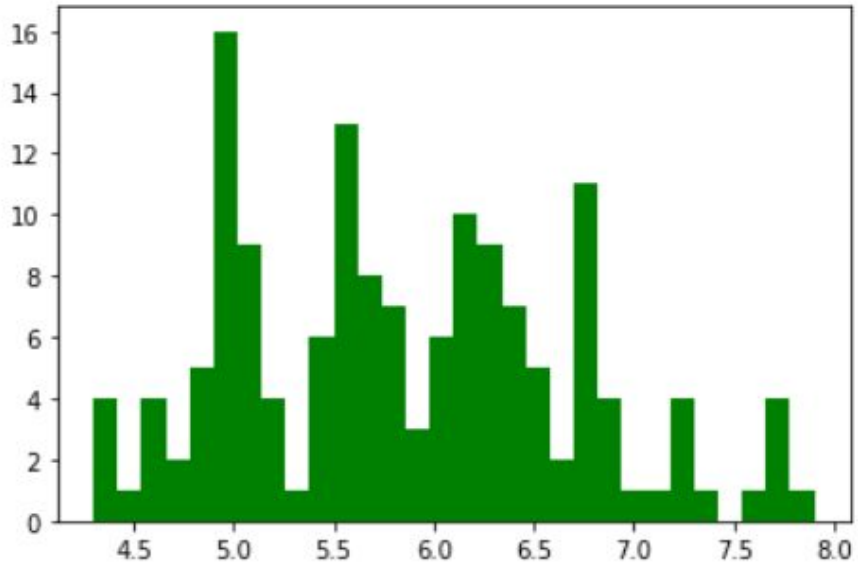
```
cantidad_de_ejemplos = df.groupby(['target_name']).size()
plt.bar(cantidad_de_ejemplos.index,cantidad_de_ejemplos.values)
plt.show()
```



Distribución

Distribución de un feature

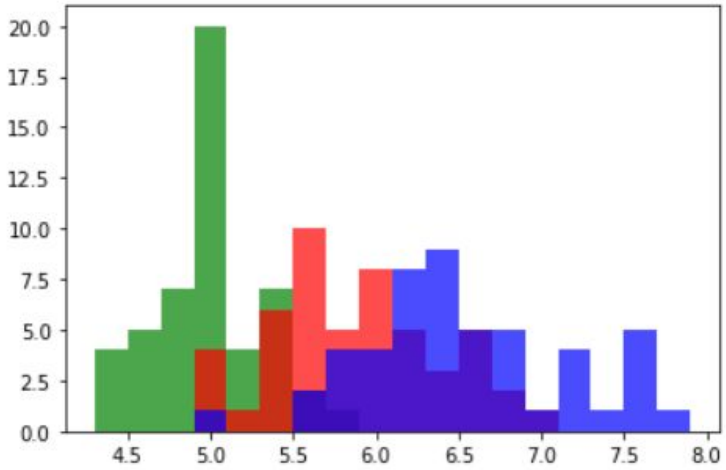
```
n, bins, patches = plt.hist(df['sepal length (cm)'], 30, facecolor='g')
```



Distribución

Distribución de features, separadas por clase

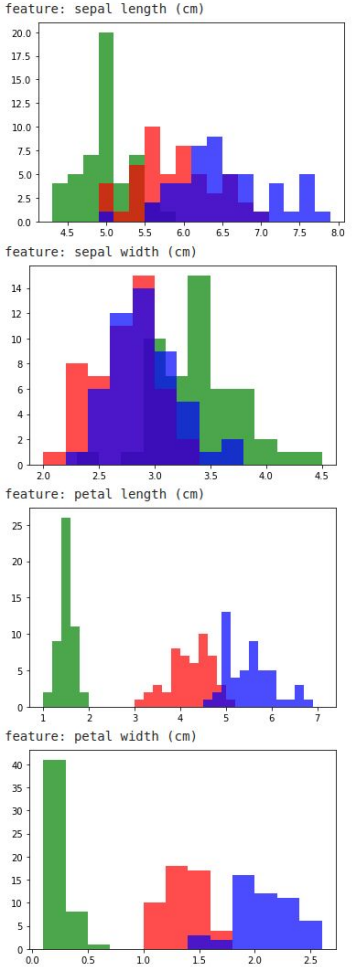
```
for clase,color in zip(['setosa', 'versicolor', 'virginica'],['g','r','b']):
  data = df[df['target_name']==clase]['sepal length (cm)']
  binwidth = 0.2
  plt.hist(data, bins=np.arange(min(data), max(data) + binwidth, binwidth), facecolor=color, alpha=0.7)
```



Varianza

```
for feature in ['sepal length (cm)','sepal width (cm)','petal length (cm)','petal width (cm)']:
  data = df[feature]
  var_feat = data.var()
  print("Varianza del feature {}: {:.3f}".format(feature,var_feat))
  for clase in ['setosa', 'versicolor', 'virginica']:
    data = df[df['target_name']==clase][feature]
    var_feat = data.var()
    print("Varianza del feature {}, para la clase {}: {:.3f}".format(feature,clase,var_feat))
```

```
Varianza del feature sepal length (cm): 0.686
Varianza del feature sepal length (cm), para la clase setosa: 0.124
Varianza del feature sepal length (cm), para la clase versicolor: 0.266
Varianza del feature sepal length (cm), para la clase virginica: 0.404
Varianza del feature sepal width (cm): 0.190
Varianza del feature sepal width (cm), para la clase setosa: 0.144
Varianza del feature sepal width (cm), para la clase versicolor: 0.098
Varianza del feature sepal width (cm), para la clase virginica: 0.104
Varianza del feature petal length (cm): 3.116
Varianza del feature petal length (cm), para la clase setosa: 0.030
Varianza del feature petal length (cm), para la clase versicolor: 0.221
Varianza del feature petal length (cm), para la clase virginica: 0.305
Varianza del feature petal width (cm): 0.581
Varianza del feature petal width (cm), para la clase setosa: 0.011
Varianza del feature petal width (cm), para la clase versicolor: 0.039
Varianza del feature petal width (cm), para la clase virginica: 0.075
```

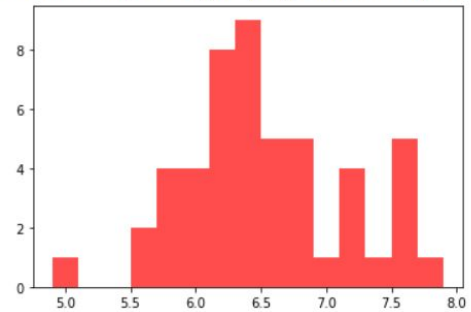


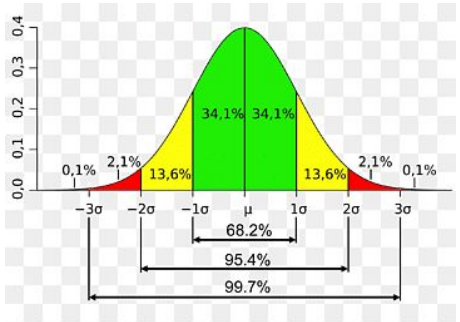
Valores Típicos, inusuales, faltantes

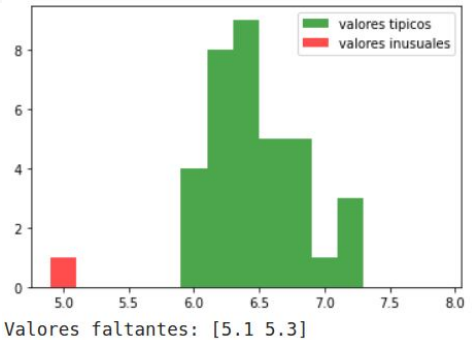
```
feature='sepal length (cm)'
clase = 'virginica'
print('feature: {}, clase: {}'.format(feature,clase))
data = df[df['target_name']==clase][feature]
data_typ = data[(data > (data.mean()-data.std())) & (data < (data.mean()+data.std()))]
data_out = data[(data < (data.mean()-2*data.std())) & (data < (data.mean()+2*data.std()))]
binwidth = 0.2
n, bins, patches = plt.hist(data, bins=np.arange(min(data), max(data) + binwidth, binwidth), facecolor='r', alpha=0.7)
plt.show()
plt.hist(data_typ, bins=np.arange(min(data), max(data) + binwidth, binwidth), facecolor='g', alpha=0.7,label='valores tipicos')
plt.hist(data_out, bins=np.arange(min(data), max(data) + binwidth, binwidth), facecolor='r', alpha=0.7,label='valores inusuales')
plt.show()

print("Valores faltantes: {}".format(bins[np.where(n == 0)]))
```

feature: sepal length (cm), clase: virginica

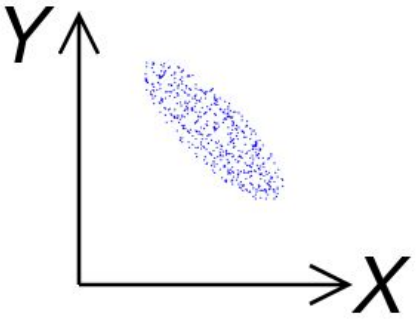




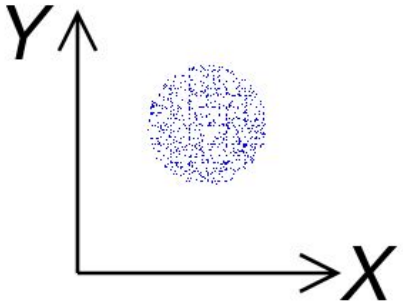


Valores faltantes: [5.1 5.3]

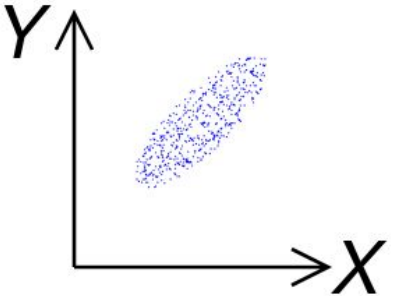
Covarianza



$\text{cov}(X,Y) < 0$



$\text{cov}(X,Y) \approx 0$



$\text{cov}(X,Y) > 0$

Matriz de covarianza

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable a and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values

```
df.drop(columns=['target']).cov()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
sepal length (cm)	0.685694	-0.042434	1.274315	0.516271
sepal width (cm)	-0.042434	0.189979	-0.329656	-0.121639
petal length (cm)	1.274315	-0.329656	3.116278	1.295609
petal width (cm)	0.516271	-0.121639	1.295609	0.581006

Matriz de correlación

$$\frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

```
df.drop(columns=['target']).corr()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
sepal length (cm)	1.000000	-0.117570	0.871754	0.817941
sepal width (cm)	-0.117570	1.000000	-0.428440	-0.366126
petal length (cm)	0.871754	-0.428440	1.000000	0.962865
petal width (cm)	0.817941	-0.366126	0.962865	1.000000

- Tema 1 - Introducción
- Tema 2 - Búsqueda de *insights*
- Tema 3 - Análisis de distribuciones
- Tema 4 - Visualización

Visualización de datos

Gráfico de líneas

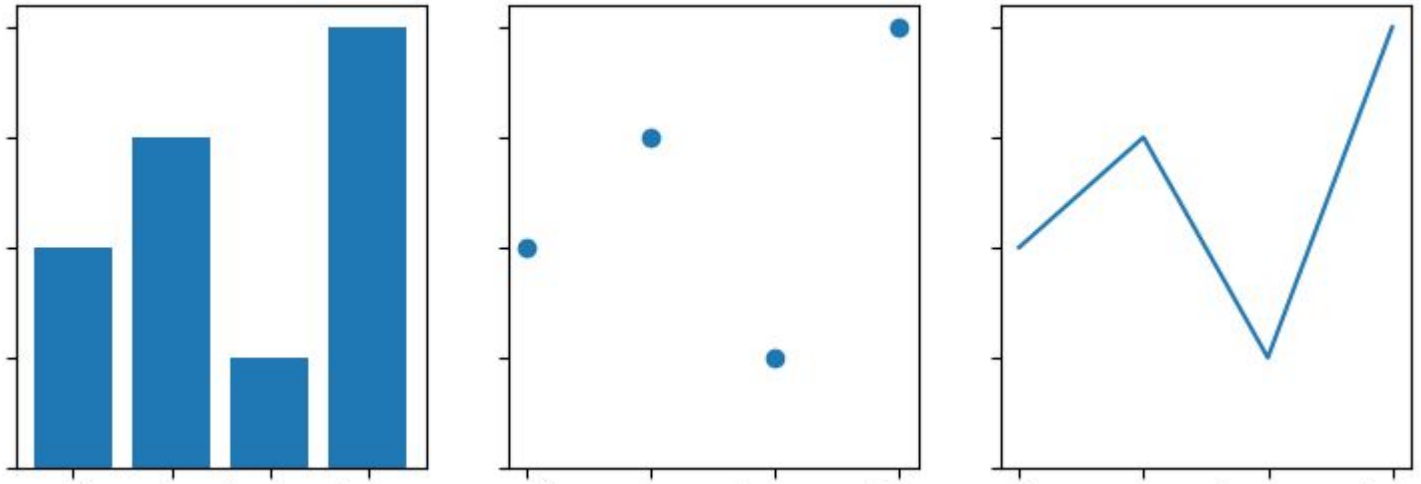


Gráfico de líneas

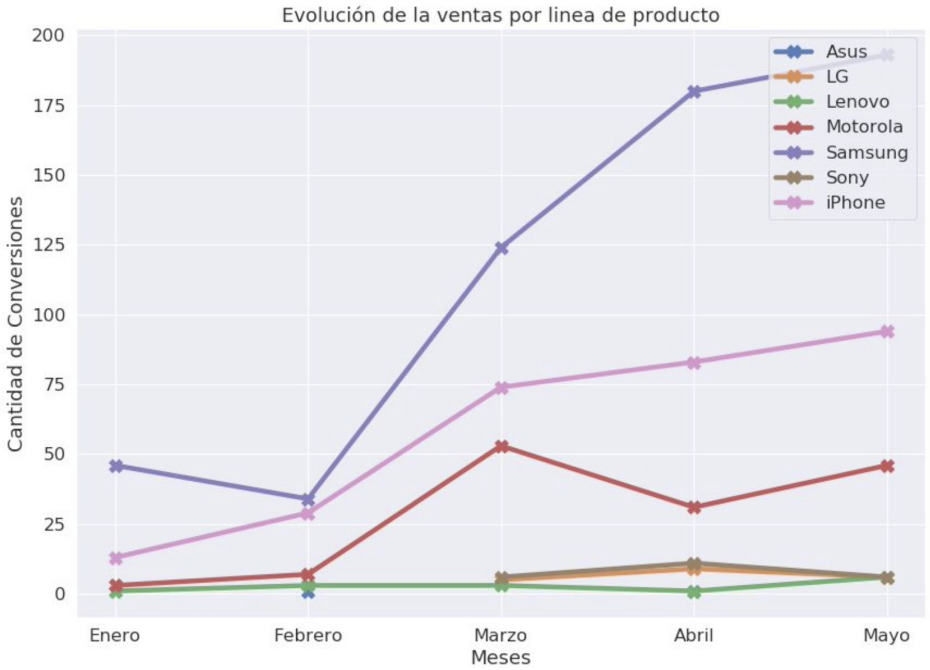


Gráfico de barras

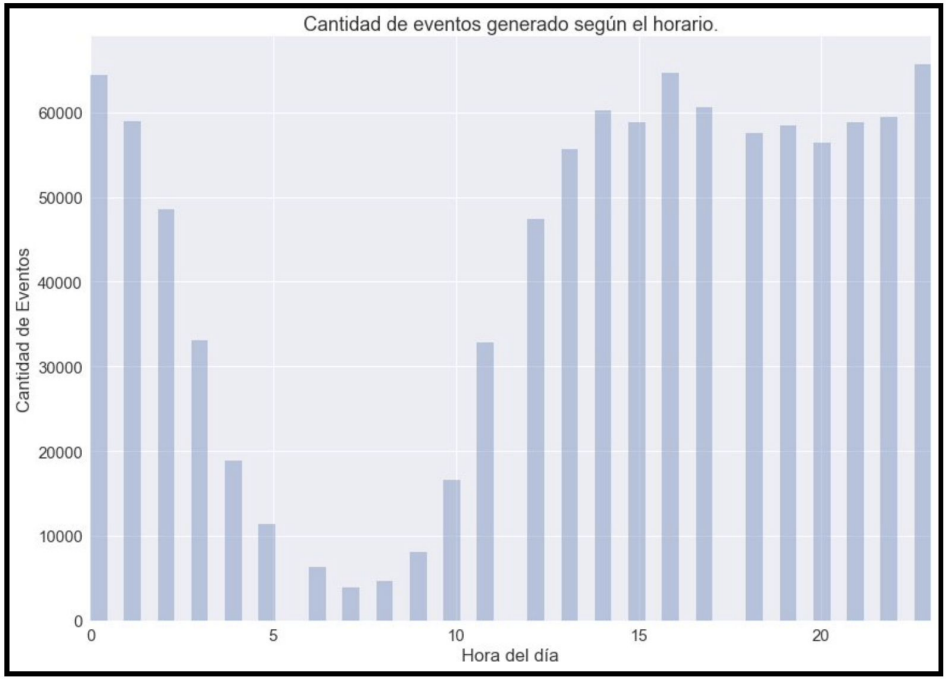


Gráfico de torta

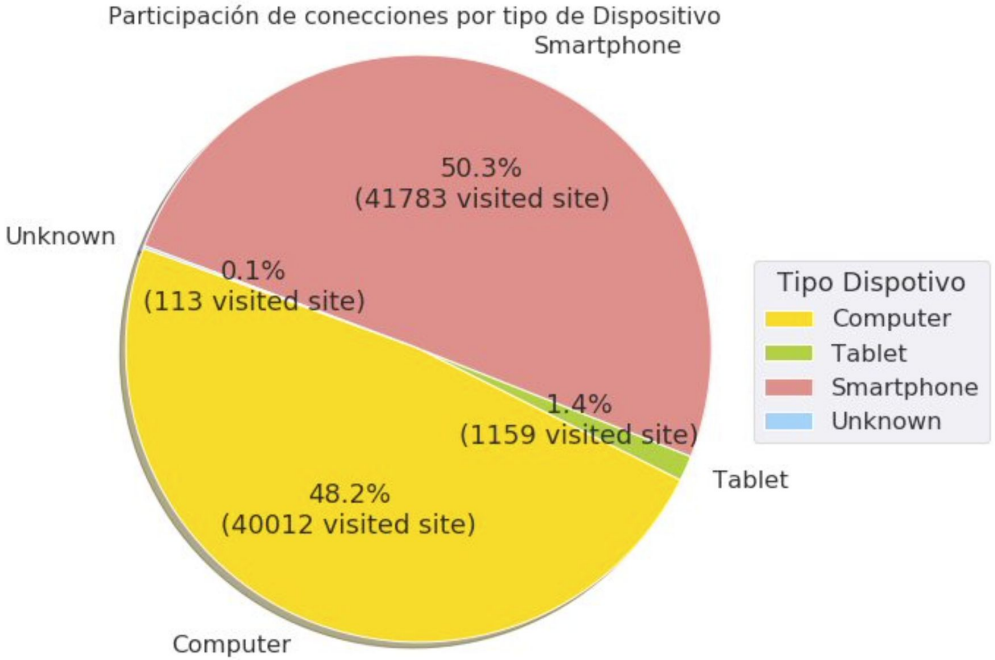
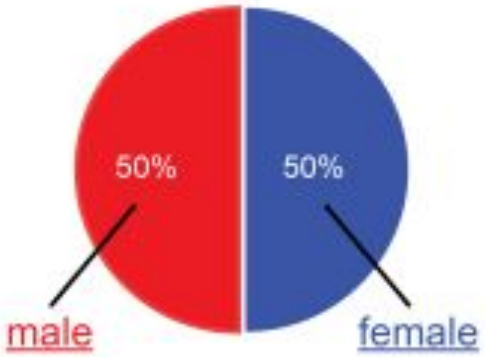


Gráfico de torta



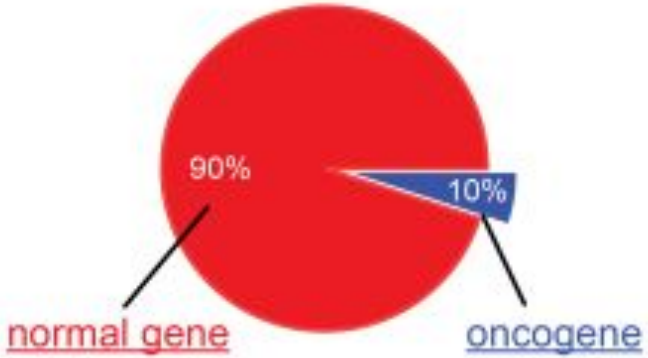
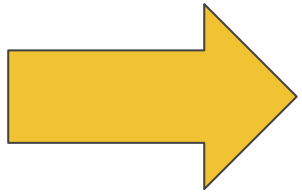
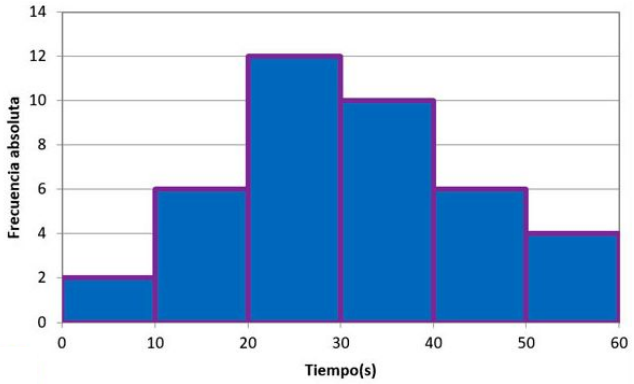


Gráfico en mosaico

Histograma

[0 - 10)	5
[10 - 20)	15
[20 - 30)	25
[30 - 40)	35
[40 - 50)	45
[50 - 60]	55
Total	





Histograma

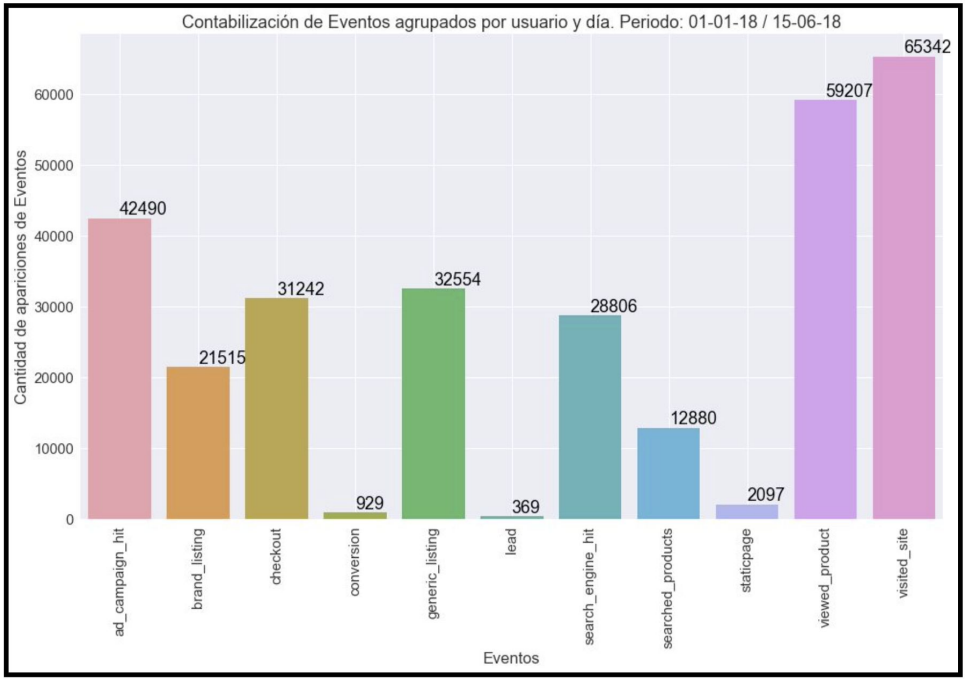

Contabilización de Eventos agrupados por usuario y día. Periodo: 01-01-18 / 15-06-18
Cantidad de apariciones de Eventos
Eventos
0
10000
20000
30000
40000
50000
60000
42490
21515
31242
929
32554
369
28806
12880
2097
59207
65342
ad_campaign_hit
brand_listing
checkout
conversion
generic_listing
lead
search_engine_hit
searched_products
staticpage
viewed_product
visited_site

Diagrama de Pareto

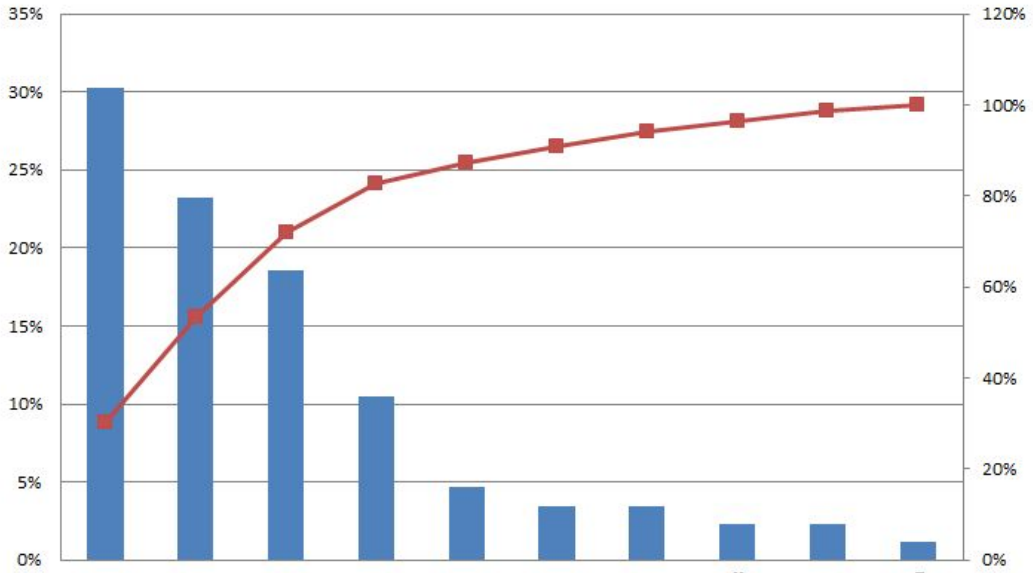
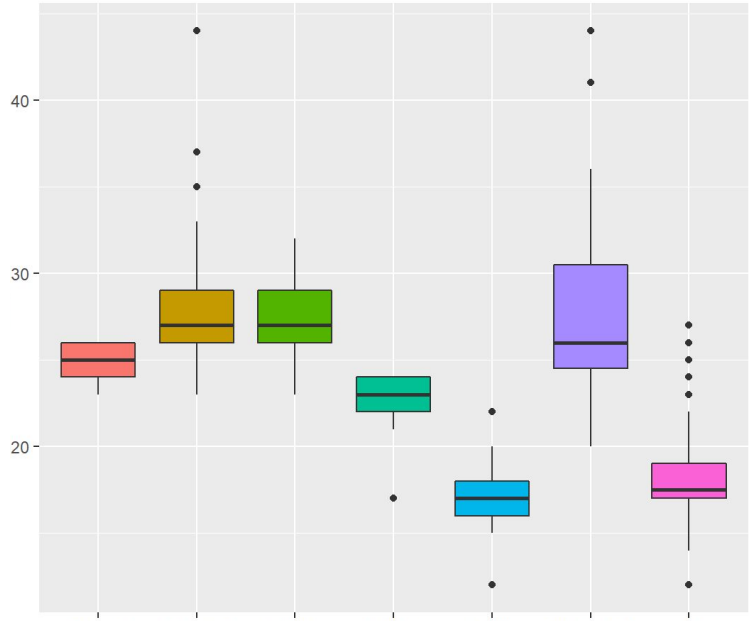
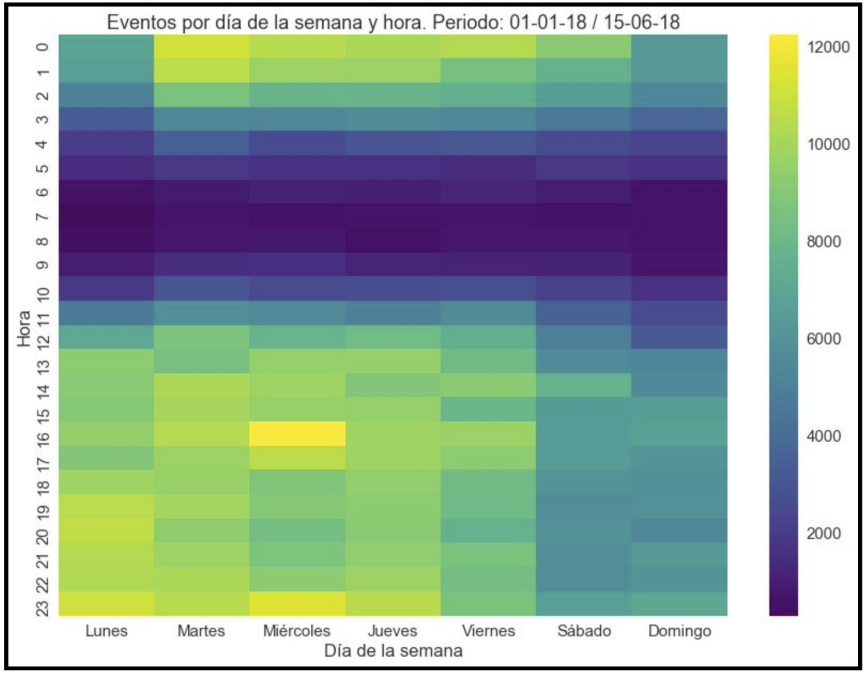


Gráfico de caja



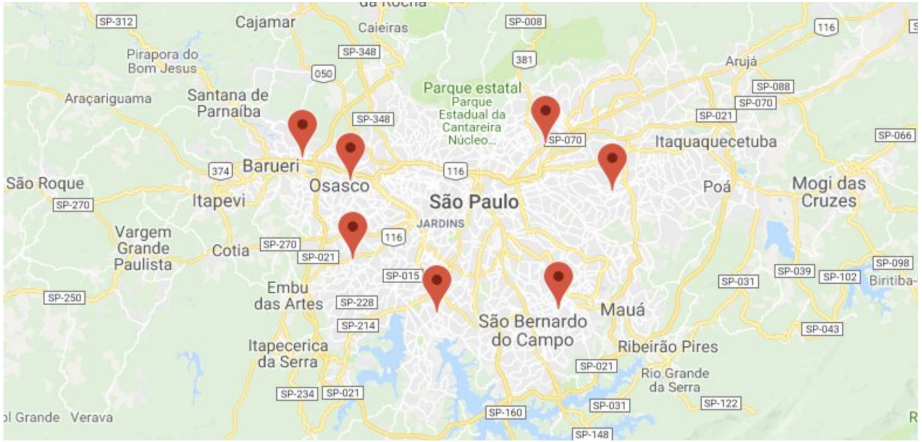
Mapa de calor



Mapa de calor

Mapa de calor

Mapa con los Quiosues



Mapa de calor con Zoom sobre la zona de San Pablo. Eventos generados.

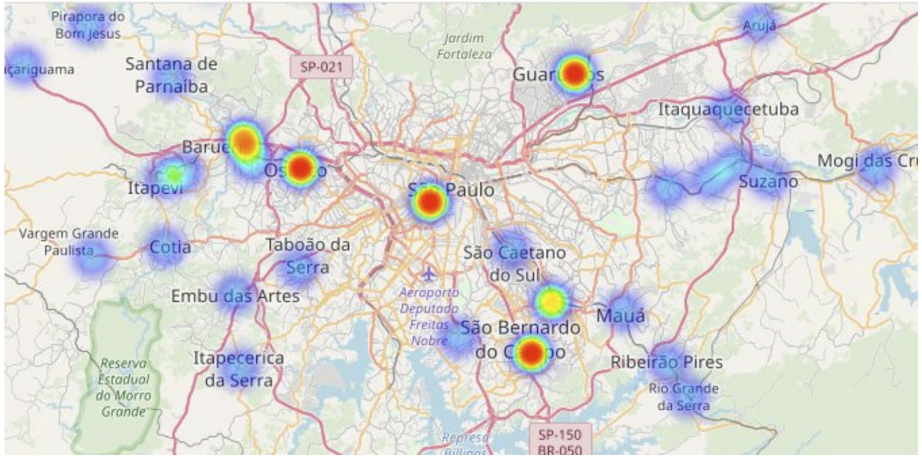


Diagrama de árbol

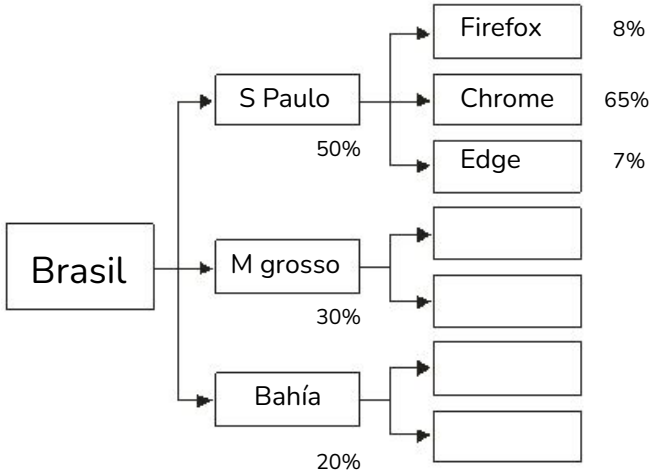
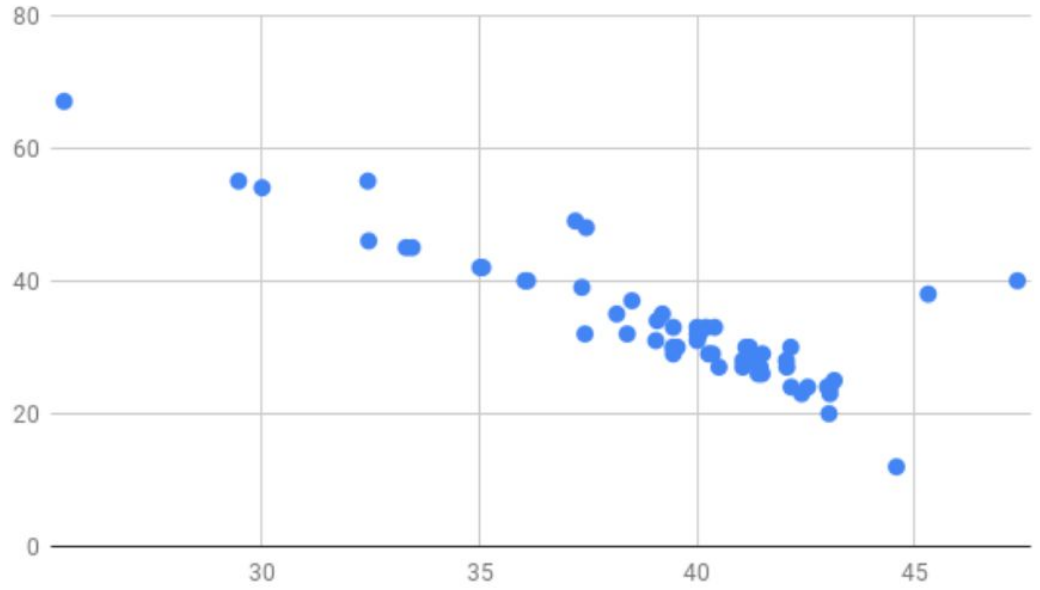
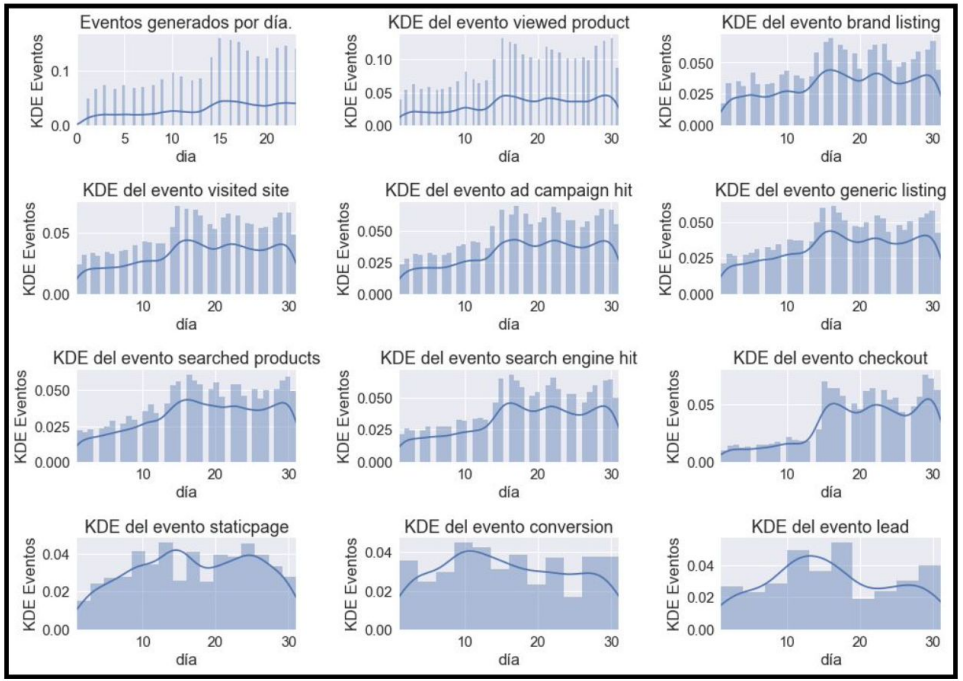


Gráfico de dispersión



Subplots

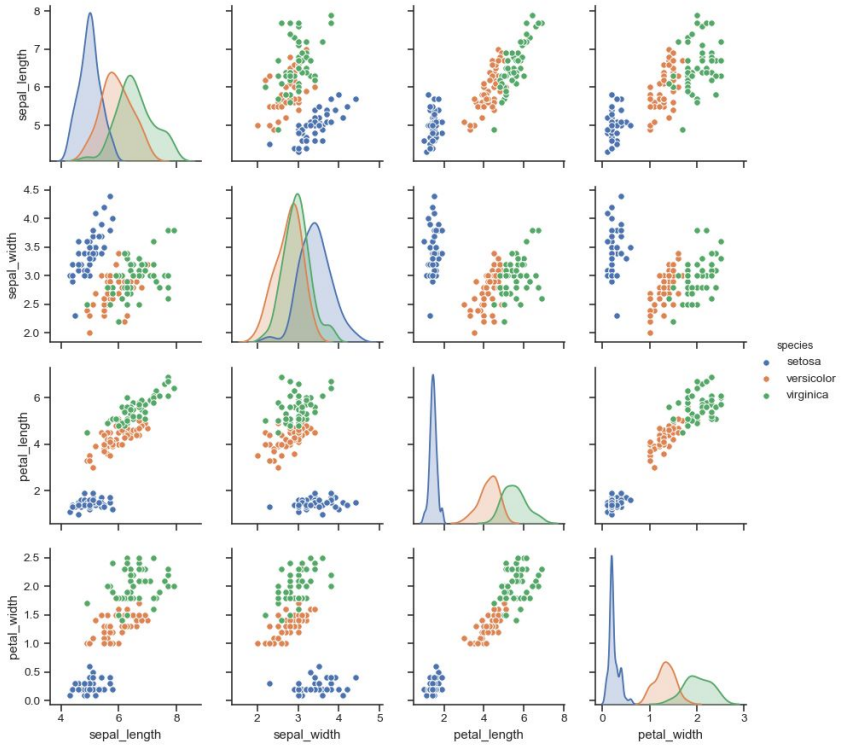

Eventos generados por día.
KDE del evento viewed product
KDE del evento brand listing
KDE del evento visited site
KDE del evento ad campaign hit
KDE del evento generic listing
KDE del evento searched products
KDE del evento search engine hit
KDE del evento checkout
KDE del evento staticpage
KDE del evento conversion
KDE del evento lead
KDE Eventos
día
dia

Nube de palabras

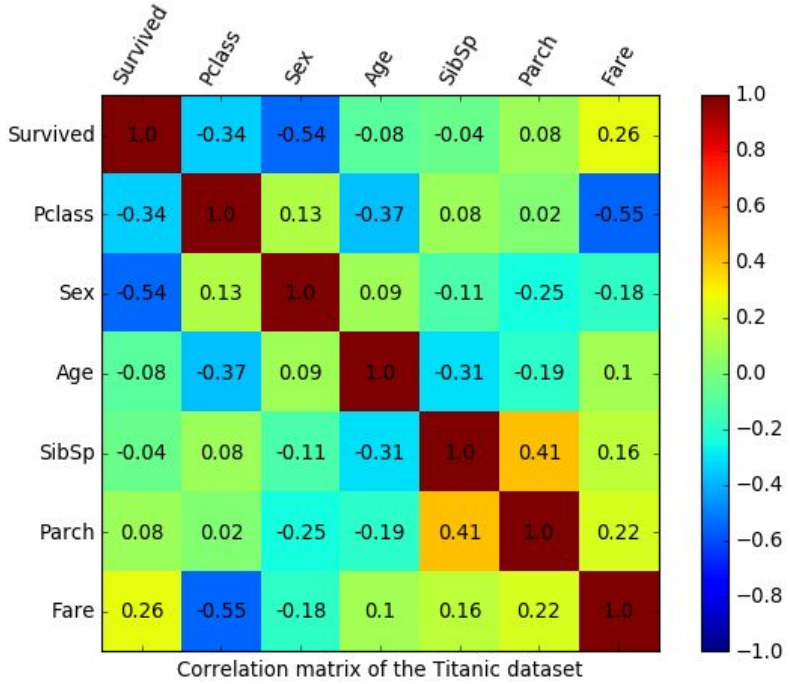
shutterstock.com · 53816176

Matriz de gráficos de dispersión

- setosa
- versicolor
- virginica

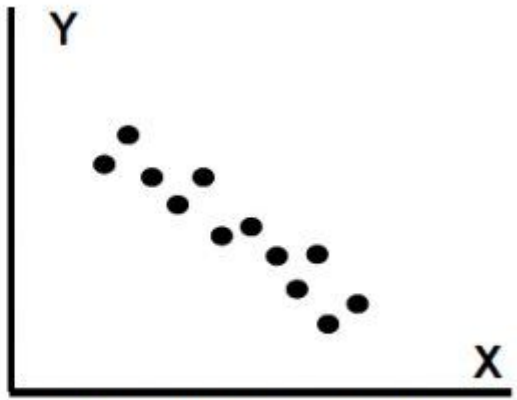


Matriz de correlación



Matriz de correlación

Correlación positiva



Correlación negativa

Preguntas?

Muchas gracias