

Curso de

Análisis Exploratorio de Datos

Maria Cruz
 @chelimsky

EDA

KDD

Knowledge Discovery
in Databases

SEMMA

Sample, Explore, Modify,
Model, and Assess

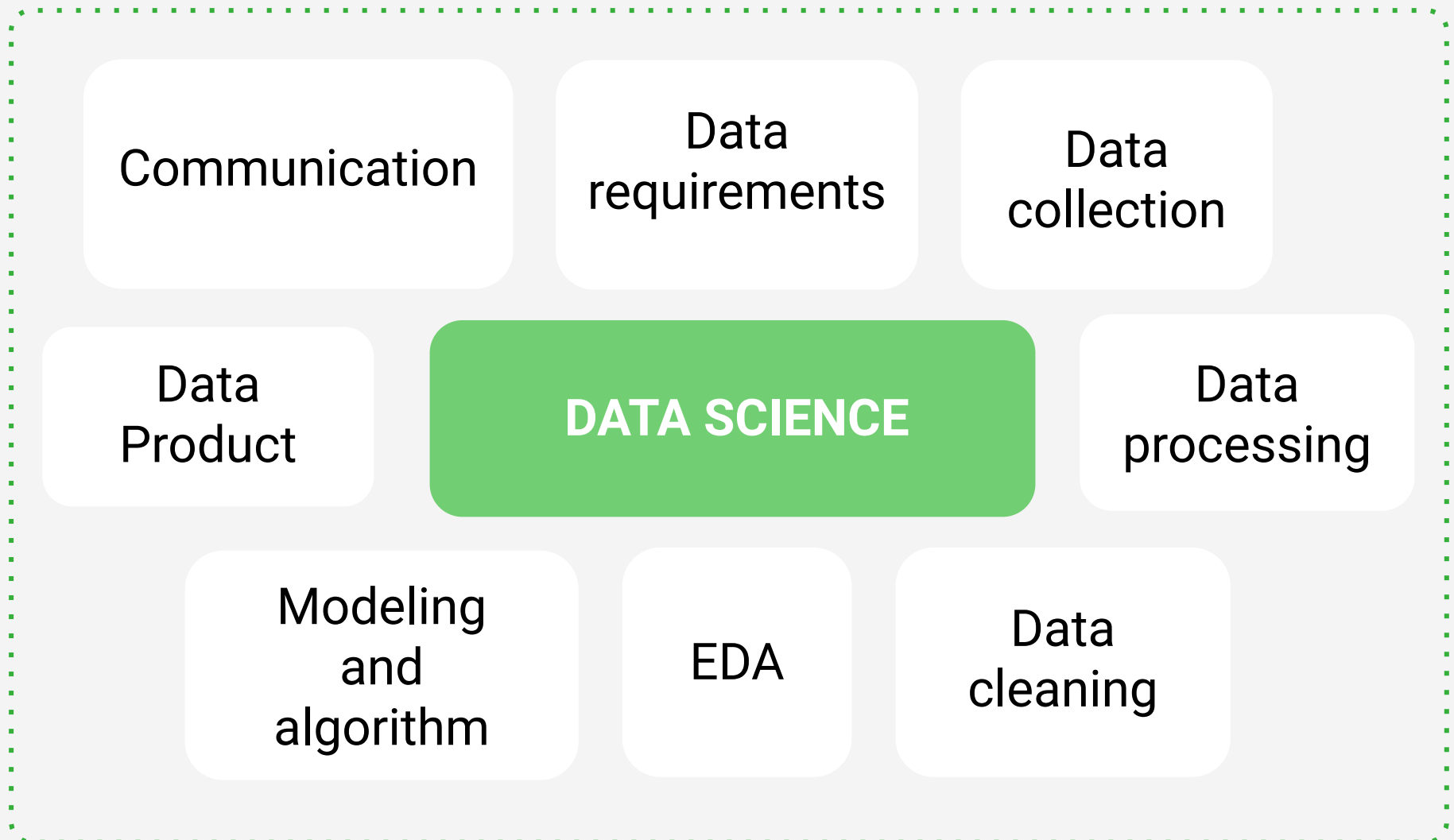
CRISP-DM

Cross Industry Standard
Process for Data Mining



Metodología para poner orden

CRISP-DM



Etapas del EDA

Problem
definition

Stage 1

Data
analysis

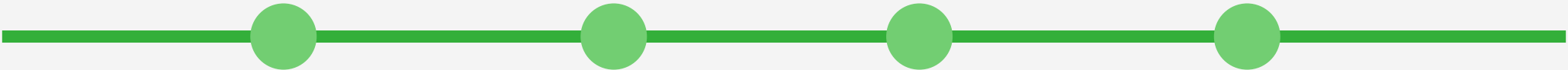
Stage 3

Stage 2

Data
preparation

Stage 4

Development and
representation
of the results



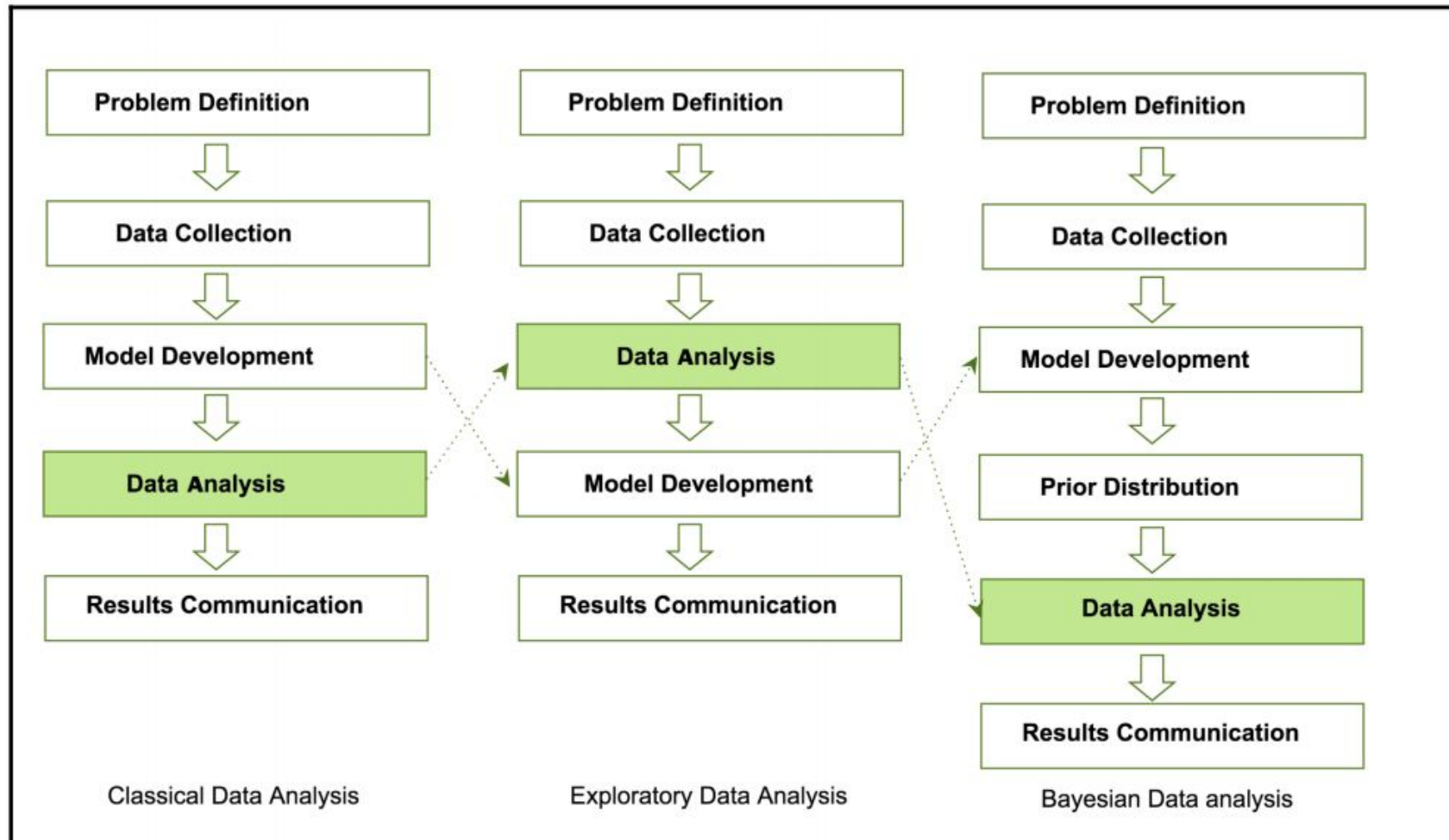
EDA

Comparación con análisis
estadístico y análisis bayesiano

Análisis típicos de datos

| | | |
|--------------------------------------|--------------------|-----------------|
| Análisis clásico | Análisis bayesiano | EDA |
| Resultados directo a la comunicación | Prior probability | Cambio dinámico |

Comparativo entre análisis clásico, bayesiano y EDA



Software tools para desarrollar EDA

EDA -Software

Python-Jupyter

localhost:8888/notebooks/Clustering%20-Copy1.ipynb

jupyter Clustering -Copy1 Last Checkpoint: 09/30/2020 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run

```
In [3]: #ama
comprehend = boto3.client(service_name='comprehend', region_name='us-east-1')
```

```
In [4]: text = "It is raining today in Seattle"
print('Calling DetectKeyPhrases')
print(json.dumps(comprehend.detect_key_phrases(Text=text, LanguageCode='en'), sort_keys=True, indent=4))
print('End of DetectKeyPhrases\n')
```

```
Calling DetectKeyPhrases
{
  "KeyPhrases": [
    {
      "BeginOffset": 14,
      "EndOffset": 19,
      "Score": 1.0,
      "Text": "today"
    }
  ],
  "ResponseMetadata": {
    "HTTPHeaders": {
      "content-length": "77",
      "content-type": "application/x-amz-json-1.1",
      "date": "Thu, 01 Oct 2020 01:27:16 GMT",
      "x-amzn-requestid": "0ffc20b8-581a-4823-9bf1-04ce149bff75"
    },
    "HTTPStatusCode": 200,
    "RequestId": "0ffc20b8-581a-4823-9bf1-04ce149bff75",
    "RetryAttempts": 0
  }
}
```

EDA -Software

AWS -SAGEMAKER

Amazon SageMaker Studio File Edit View Run Kernel Git Tabs Settings Help

xgboost_customer_churn.ipynb

- Have the predictor variable in the first column
- Not have a header row

But first, let's convert our categorical features into numeric features.

```
[ ]: model_data = pd.get_dummies(churn)
      model_data = pd.concat([model_data['Churn?_True'], model_data.drop(['Churn?_True'], axis=1)], axis=1)
```

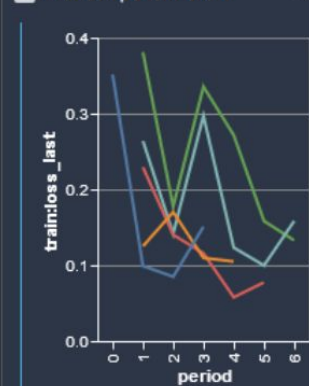
And now let's split the data into training, validation, and test sets. This will help prevent us from overfitting the model, and allow us to test the models accuracy on data it hasn't already seen.

```
[ ]: train_data, validation_data, test_data = np.split(model_data.sample(frac=1, random_state=42), [int(0.33 * len(model_data)), int(0.33 * len(model_data))])
      train_data.to_csv('train.csv', header=False, index=False)
      validation_data.to_csv('validation.csv', header=False, index=False)
```

Now we'll upload these files to S3.

```
[ ]: boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'train.csv')).upload_file(train_data.to_csv(index=False).get_value())
      boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'validation.csv')).upload_file(validation_data.to_csv(index=False).get_value())
```

Trial Component Chart



Trial Component List

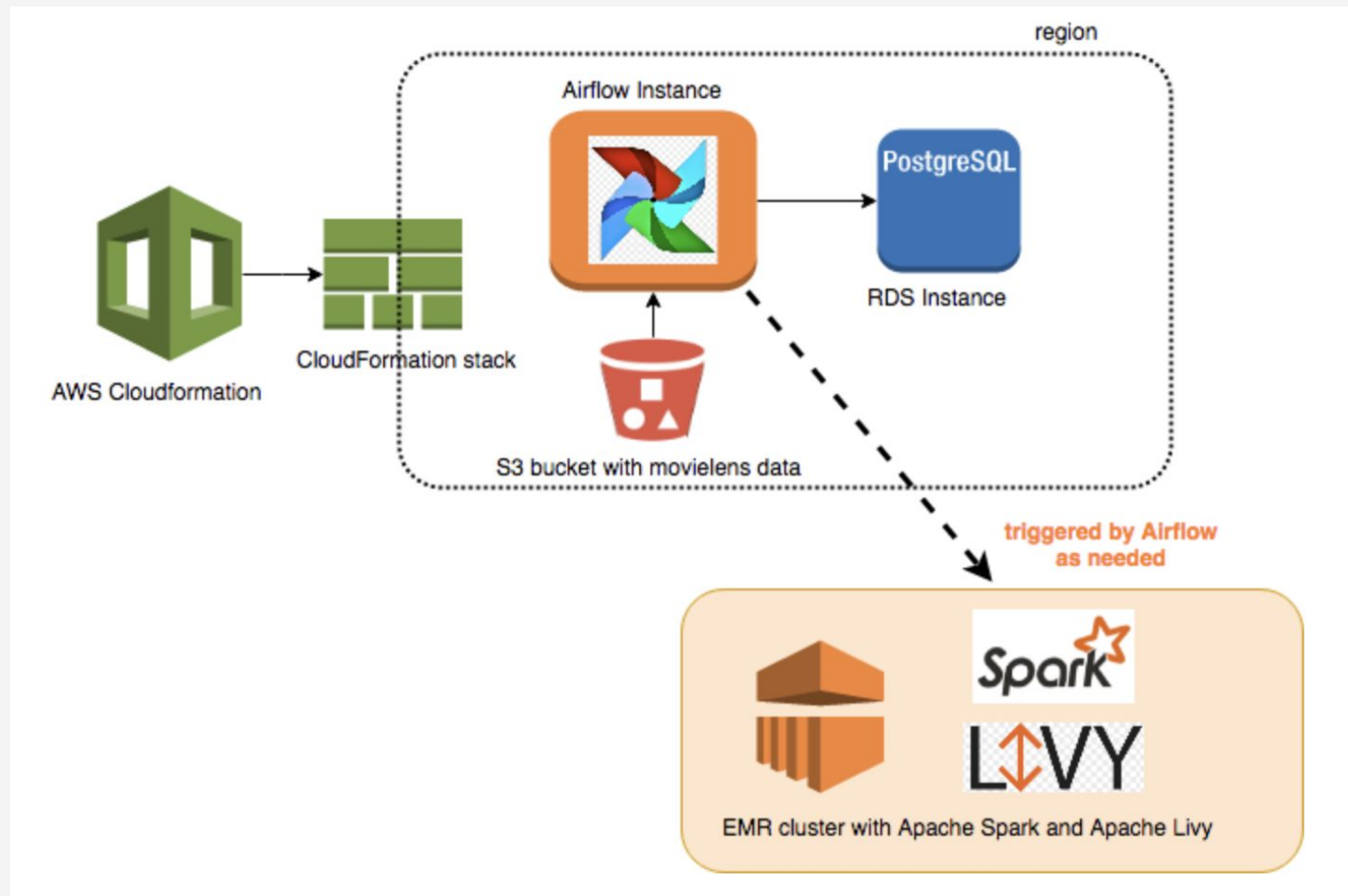
10 rows selected

| Status | Experiment | Type | Trial | Trial c |
|-------------|-------------------------|--------------|---------|---------|
| ✓ Completed | customer-churn-predi... | Training job | Trial-3 | Tra |
| ✓ Completed | customer-churn-predi... | Training job | Trial-2 | Tra |
| ✓ Completed | customer-churn-predi... | Training job | Trial-1 | Tra |
| ✓ Completed | customer-churn-predi... | Training job | Trial-0 | Tra |

Mode: Command Ln 1, Col 1 xgboost_customer_churn.ipynb

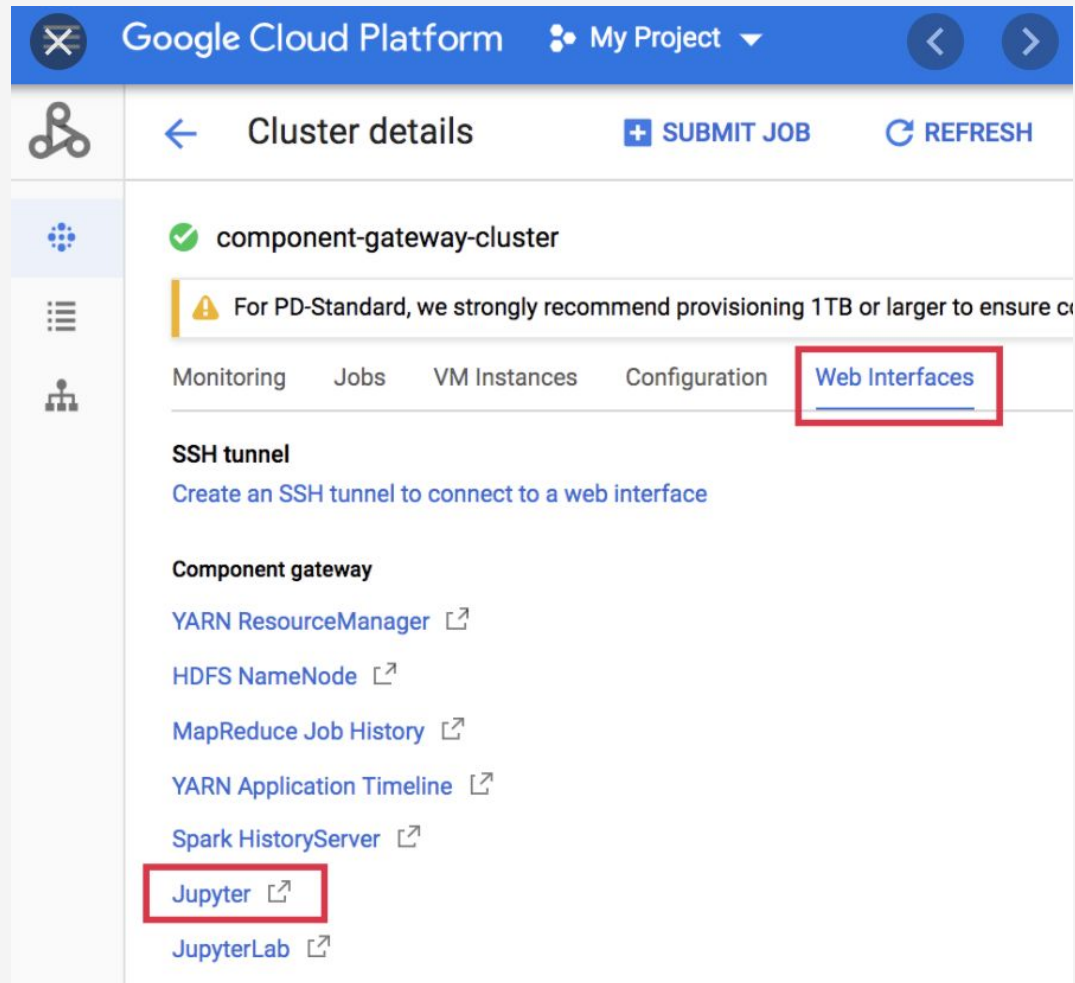
EDA -Software

AWS- EMR



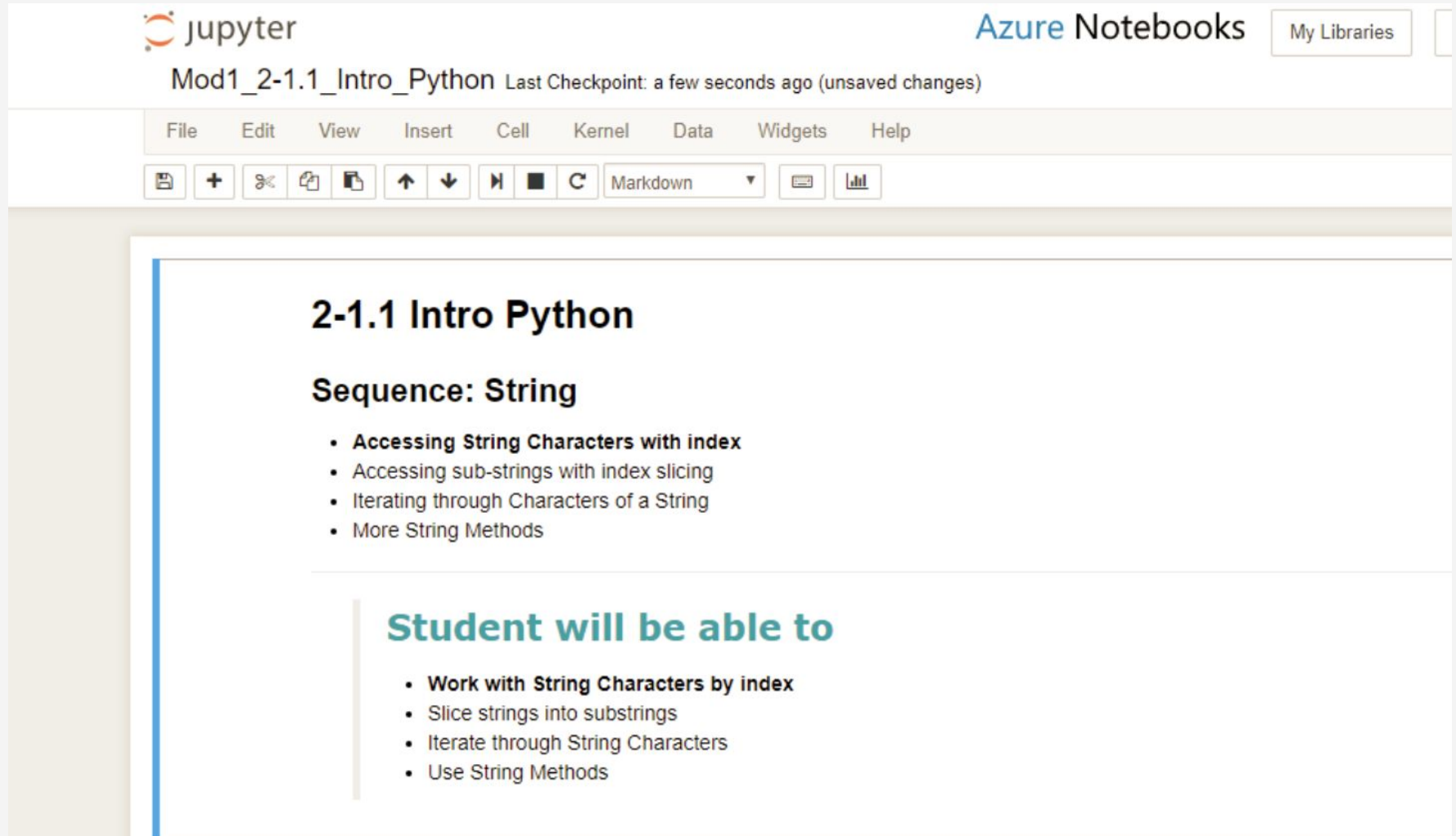
Fuente: <https://aws.amazon.com/blogs/big-data/build-a-concurrent-data-orchestration-pipeline-using-amazon-emr-and-apache-livy/>

Google - Jupyter notebook cloud



Fuente: <https://cloud.google.com/dataproc/docs/tutorials/jupyter-notebook?hl=es-419>

Azure - Notebooks



The screenshot displays the Azure Jupyter Notebook interface. At the top, the Jupyter logo is on the left, and 'Azure Notebooks' is on the right next to a 'My Libraries' button. Below this, the notebook title 'Mod1_2-1.1_Intro_Python' is shown, followed by the status 'Last Checkpoint: a few seconds ago (unsaved changes)'. A menu bar contains 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Data', 'Widgets', and 'Help'. Below the menu is a toolbar with icons for saving, adding cells, undo, redo, copy, paste, and navigation, along with a cell type dropdown set to 'Markdown'. The main content area has a blue vertical bar on the left and contains the following text:

2-1.1 Intro Python

Sequence: String

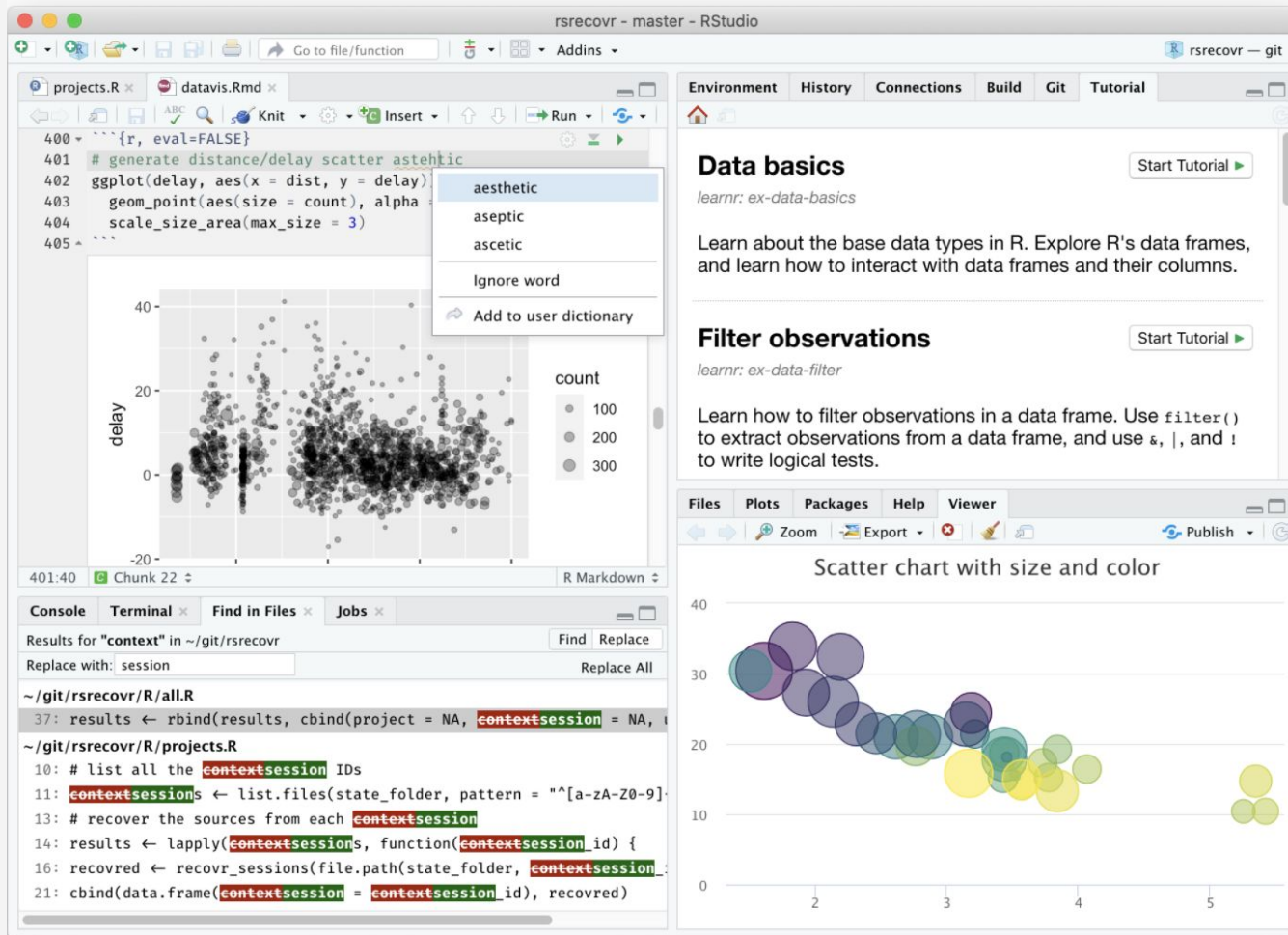
- **Accessing String Characters with index**
- Accessing sub-strings with index slicing
- Iterating through Characters of a String
- More String Methods

Student will be able to

- **Work with String Characters by index**
- Slice strings into substrings
- Iterate through String Characters
- Use String Methods

Fuente: <https://notebooks.azure.com/help/jupyter-notebooks>

R - Rstudio



Fuente: <https://blog.rstudio.com/2020/05/27/rstudio-1-3-release/>

KNIME



[Hub](#) [Blog](#) [Forum](#) [Events](#) [Use Cases](#) [Careers](#) [Contact](#)

[Download](#)



[SOFTWARE](#) / [PRICING](#) / [COMMUNITY](#) / [LEARNING](#) / [PARTNERS](#) / [ABOUT](#)

End to End Data Science

At KNIME, we build software to create and productionize data science using one easy and intuitive environment, enabling every stakeholder in the data science process to focus on what they do best.



Visualizaciones de EDA

Transformación de los datos



Estadística descriptiva

Distribución de los datos

Discreta o Continua

Continua

Cuando puede tomar cualquier valor dentro de un intervalo.

Discreta

Cuando no puede tomar ningún valor entre dos consecutivos.

Ejemplos

Continua

Temperaturas registradas en un observatorio; tiempo en recorrer una distancia en una carrera.

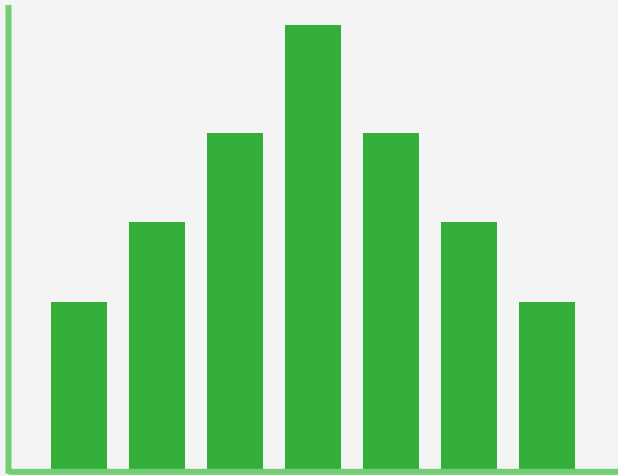
Discreta

Número de empleados de una fábrica; número de hijos; número de cuentas ocultas en Suiza.

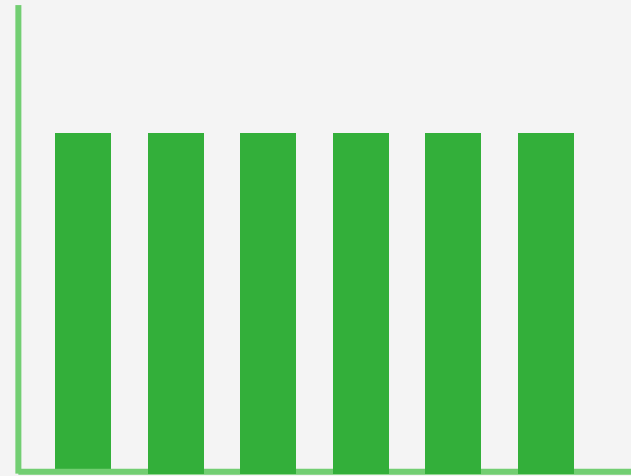
Tipos de distribuciones

Conoce los diferentes tipos de distribución de datos: uniforme discreta, Bernoulli, binomio, binomio negativo, Poisson, geométrica, uniforme continua, normal (curva de campana), exponencial, gamma y beta.

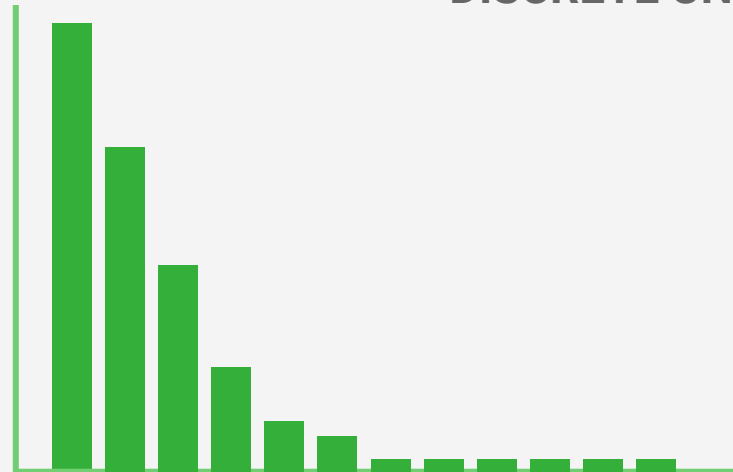
Tipos de distribuciones



BINOMIAL



DISCRETE UNIFORM



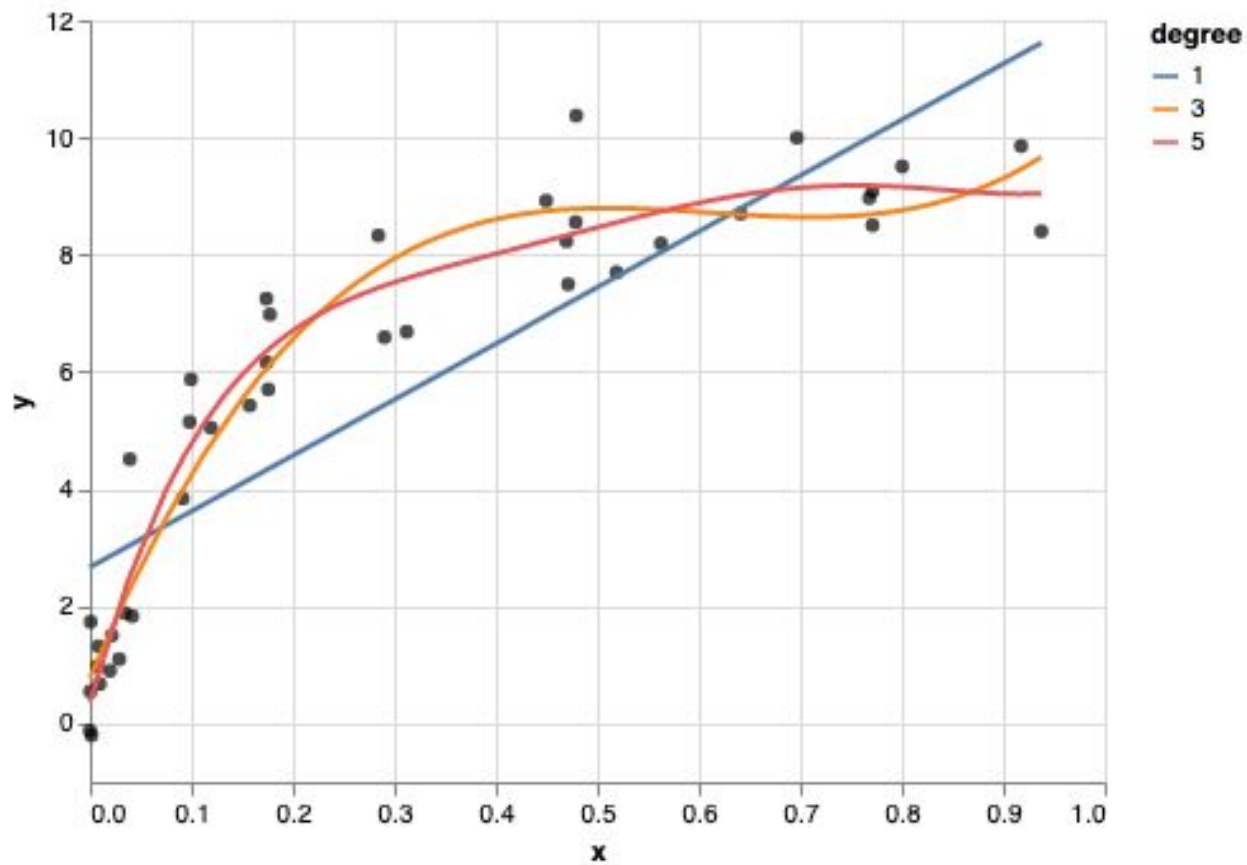
WARING YULE

Cómo determinar la distribución

1. Realiza una representación gráfica de tus datos.
2. Descarta primero lo que no puede ser.
 - a. Si hay algún pico en el conjunto de datos, no puede ser una distribución uniforme discreta.
 - b. Si los datos tienen más de un pico, no es Poisson o binomial.
 - c. Si tiene una sola curva, no hay picos secundarios, y tiene una pequeña pendiente en cada lado, podría ser una distribución Poisson o gamma. Pero no podrá ser una distribución uniforme discreta.

Ajuste de la curva

$$R^2 = 0.99$$



Medida de tendencia central

“

No se puede predecir el
comportamiento individual, pero
sí el comportamiento promedio.

”

Alejandro Quintela del Río

La ley de los grandes números



SIMÉON-DENIS
POISSON

Dice que (bajo ciertas condiciones generales) la media de n variables aleatorias X_1, X_2, \dots, X_n se aproxima a la media de las n medias $\mu_1, \mu_2, \dots, \mu_n$ (donde $\mu_i = E(X_i)$)

$$\frac{X_1 + X_2 + \dots + X_n}{n} \longrightarrow \frac{\mu_1 + \mu_2 + \dots + \mu_n}{n}$$

El teorema del límite central

Cuando el tamaño de la muestra es lo suficientemente grande, la distribución de las medias sigue aproximadamente una distribución normal.

Medidas estadísticas

Media

Mediana

Moda

Min

Max

Producto de valores

Suma acumulada

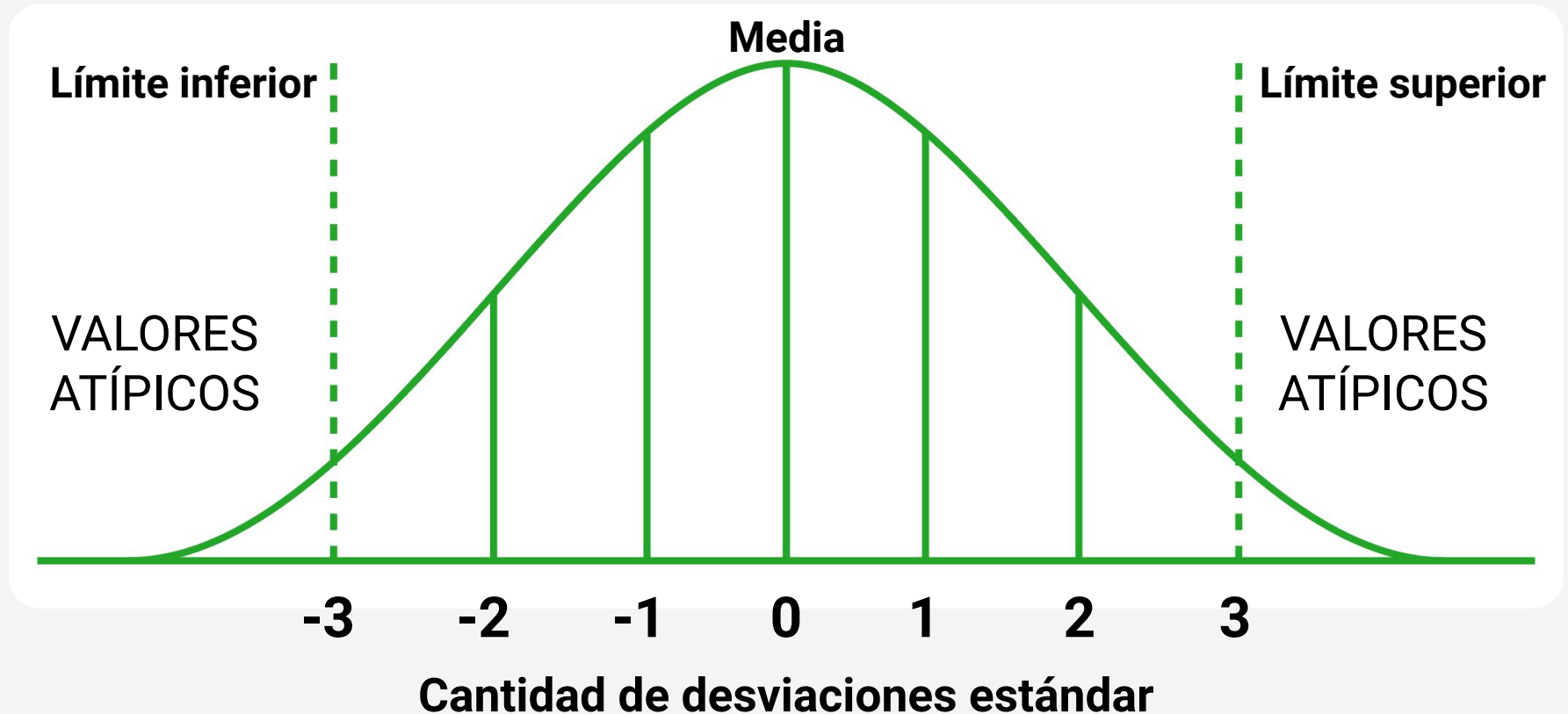
Medidas de dispersión

Desviación estándar

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

La desviación estándar de una población repasa la cantidad de dispersión de los datos de una población entera.

Desviación estándar



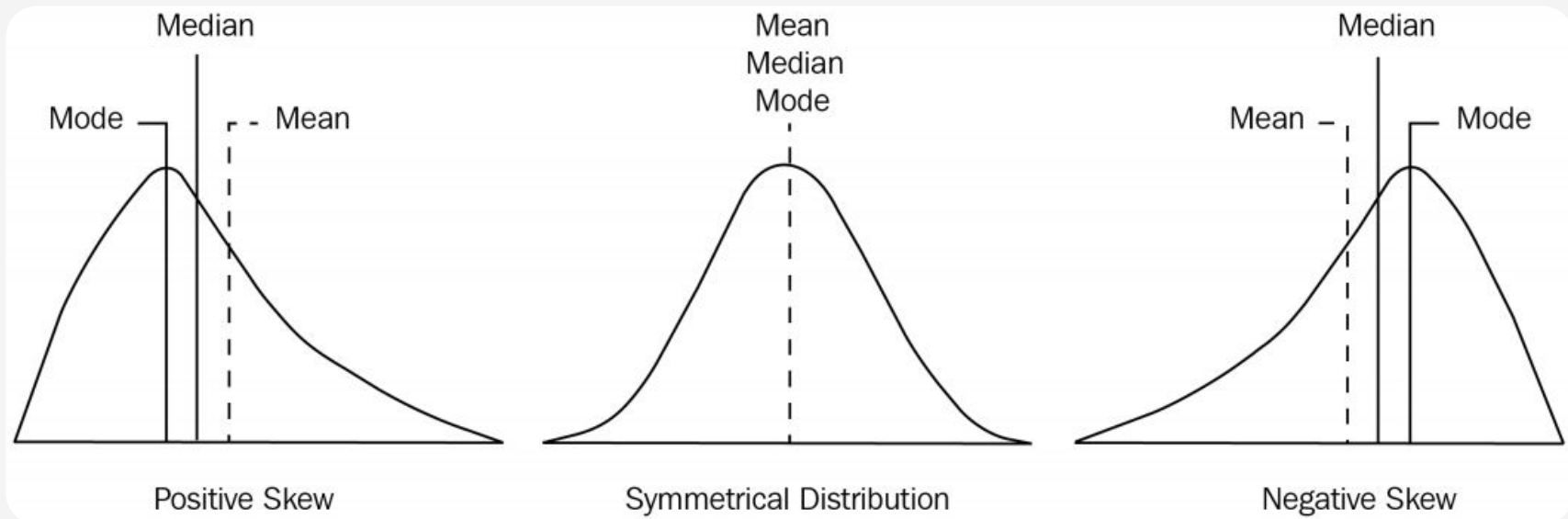
Un valor bajo de la desviación típica indica que los números del conjunto están relativamente concentrados alrededor de la media.

Varianza

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{X}^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j>i} (x_i - x_j)^2$$

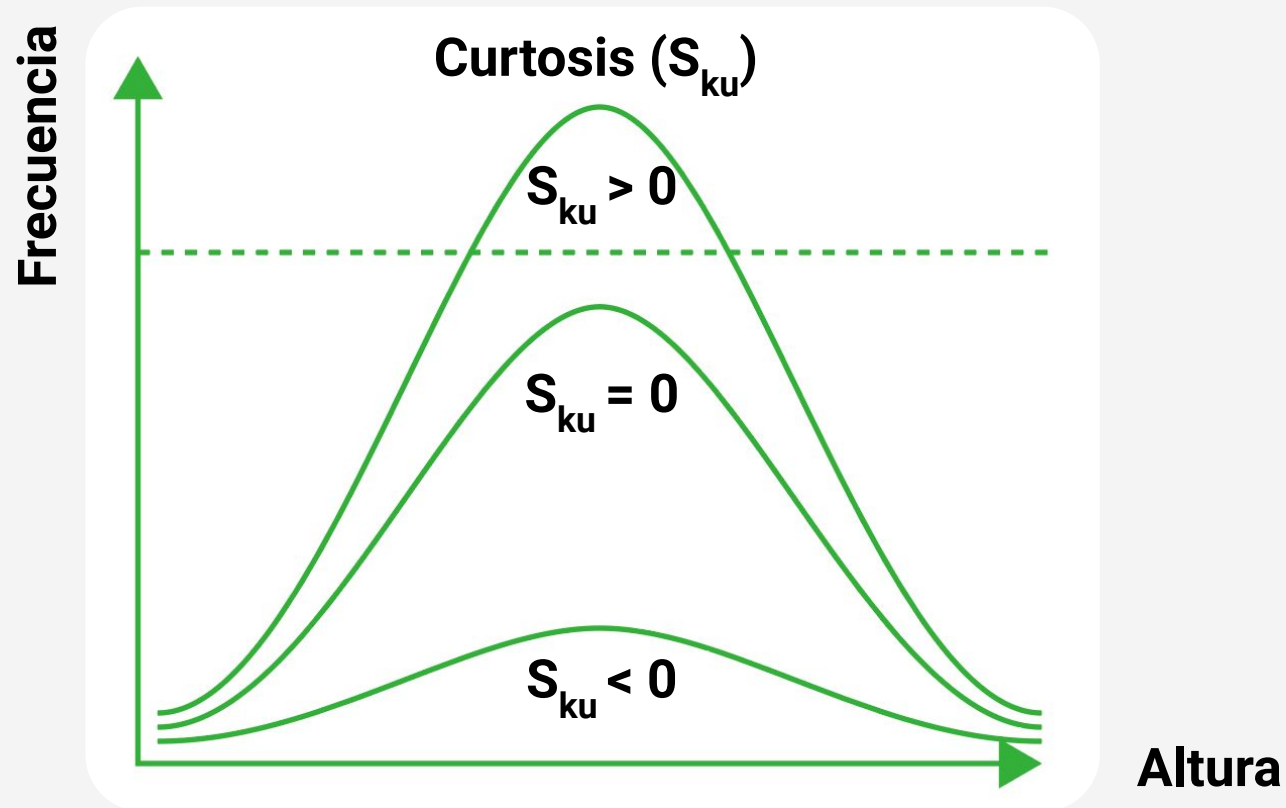
La varianza es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media.

Asimetría estadística (Skewness)



Las medidas de asimetría son indicadores que permiten establecer el grado de simetría (o asimetría) que presenta una distribución de probabilidad de una variable aleatoria sin tener que hacer su representación gráfica.

Curtosis



La curtosis de una variable estadística/aleatoria es una característica de forma de su distribución de frecuencias/probabilidad.

Agrupamiento de datasets

Integración de datos

Pivot tables y cross-tabulations

Correlación

Análisis multivariable empleando el dataset Titanic

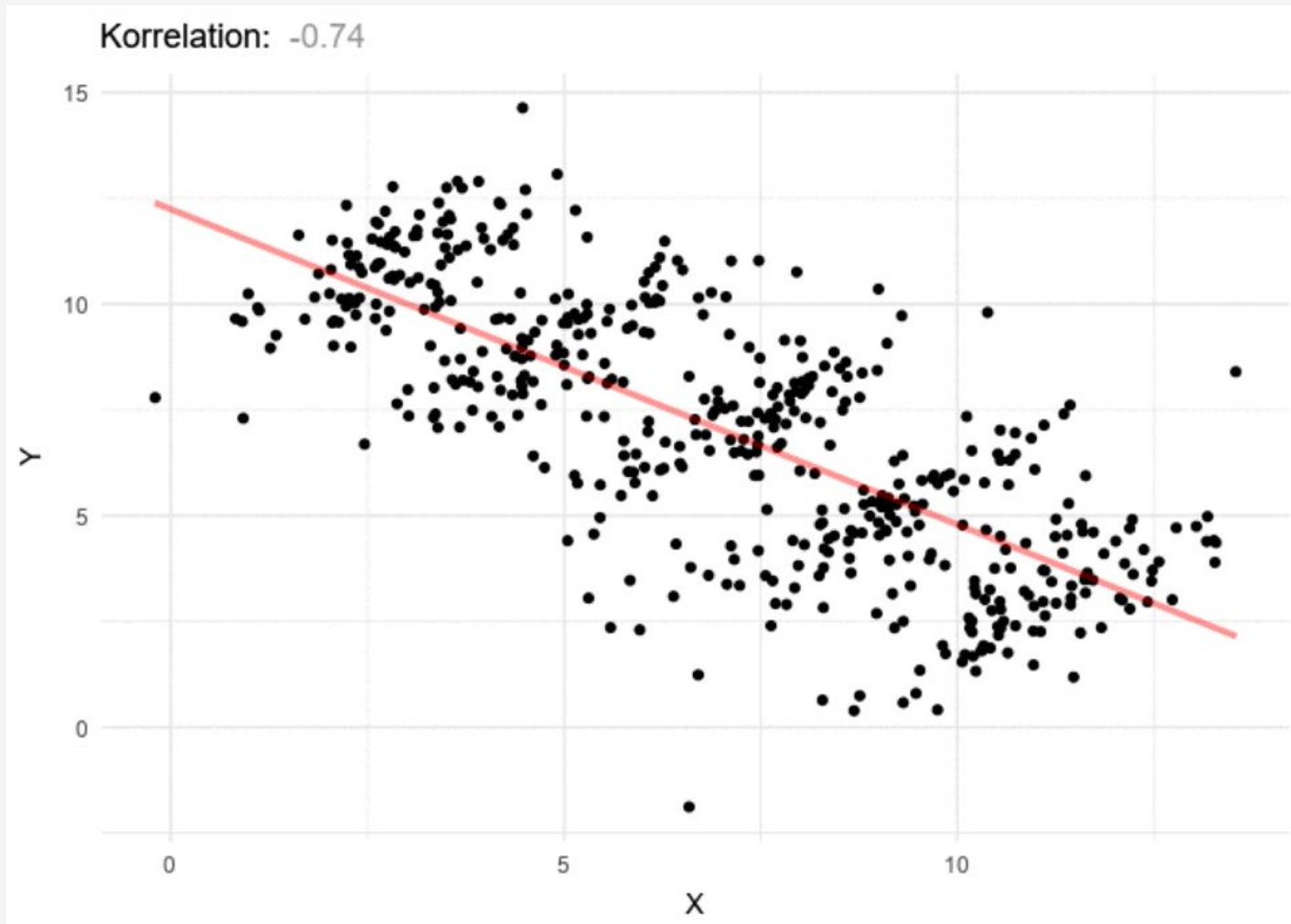
Paradoja de Simpson

Definición

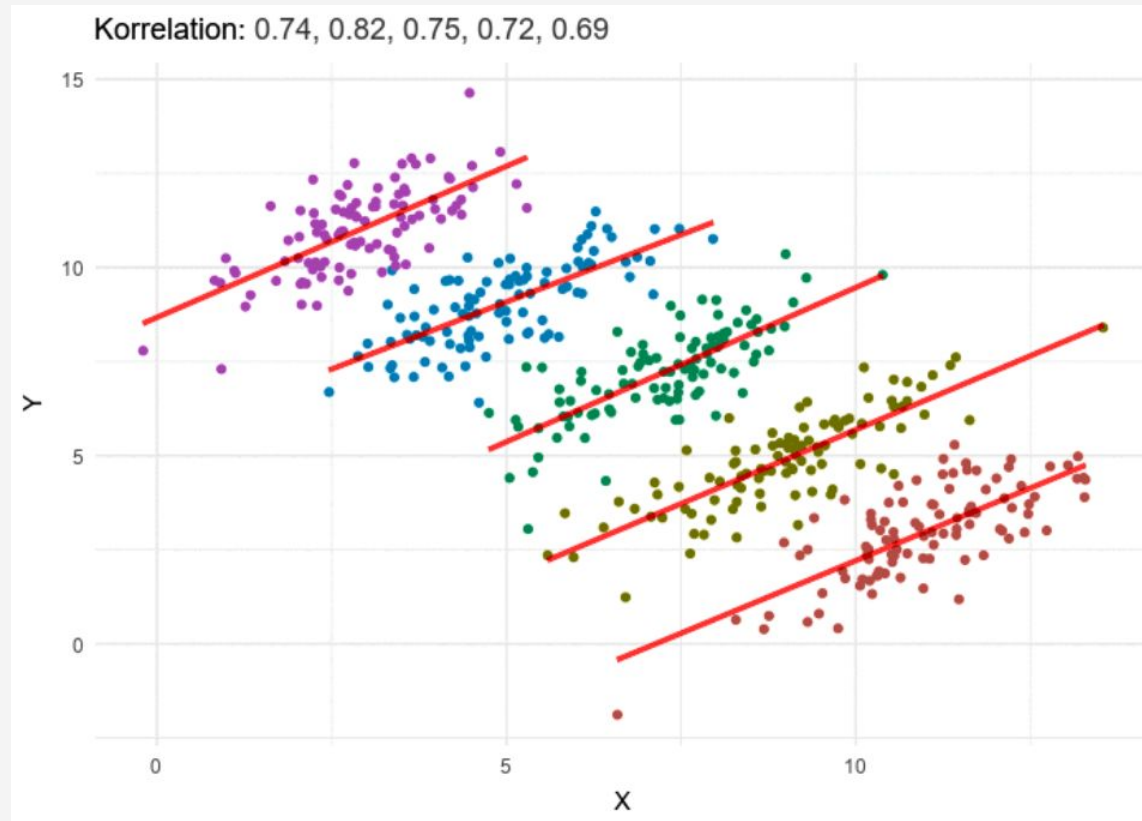
Aparece en varios grupos de datos, desaparece cuando estos grupos se combinan y en su lugar aparece la tendencia contraria para los datos agregados.

La idea básica es que incluso una correlación elevada encontrada entre dos variables puede ser interpretada erróneamente.

Tenemos un conjunto de datos con una correlación negativa de 0.74



Fuente: https://commons.wikimedia.org/wiki/File:Simpsons_paradox_-_animation.gif



Pero si se consideran los grupos determinados por una tercera variable, se puede observar que, para cada grupo, la correlación obtenida para cada uno de ellos tiene una magnitud parecida, pero de signo contrario.

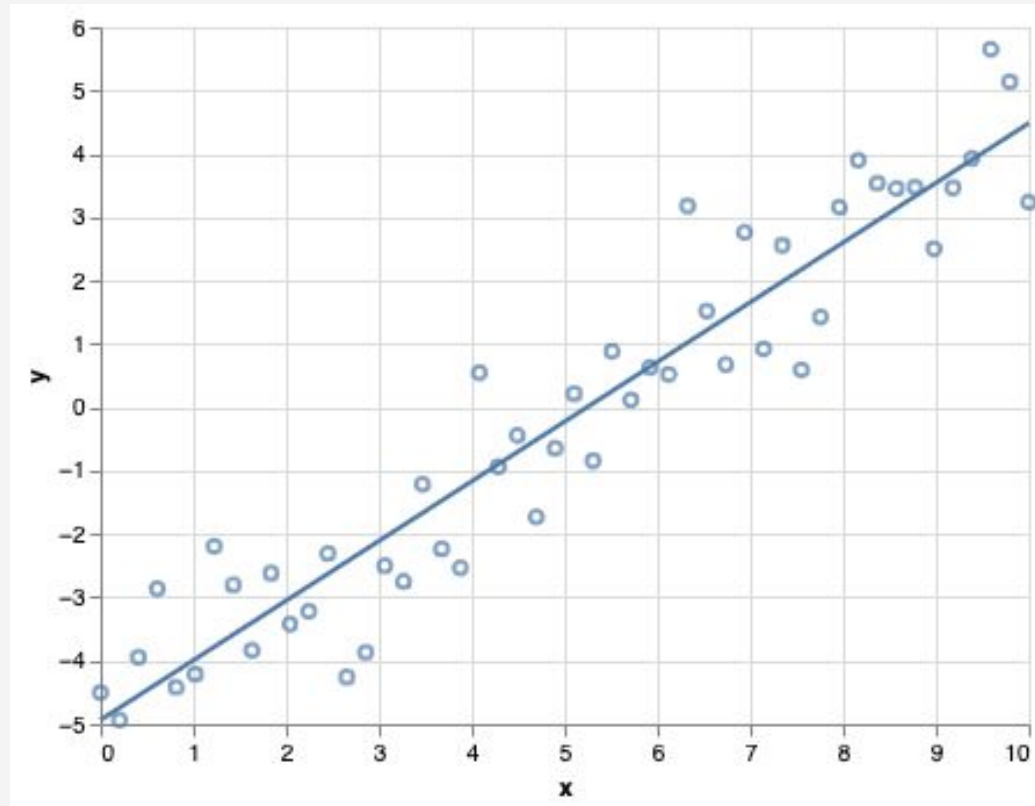
**Correlación no
implica causalidad**



Cum hoc ergo propter hoc

con esto, por tanto a causa de esto

Correlación



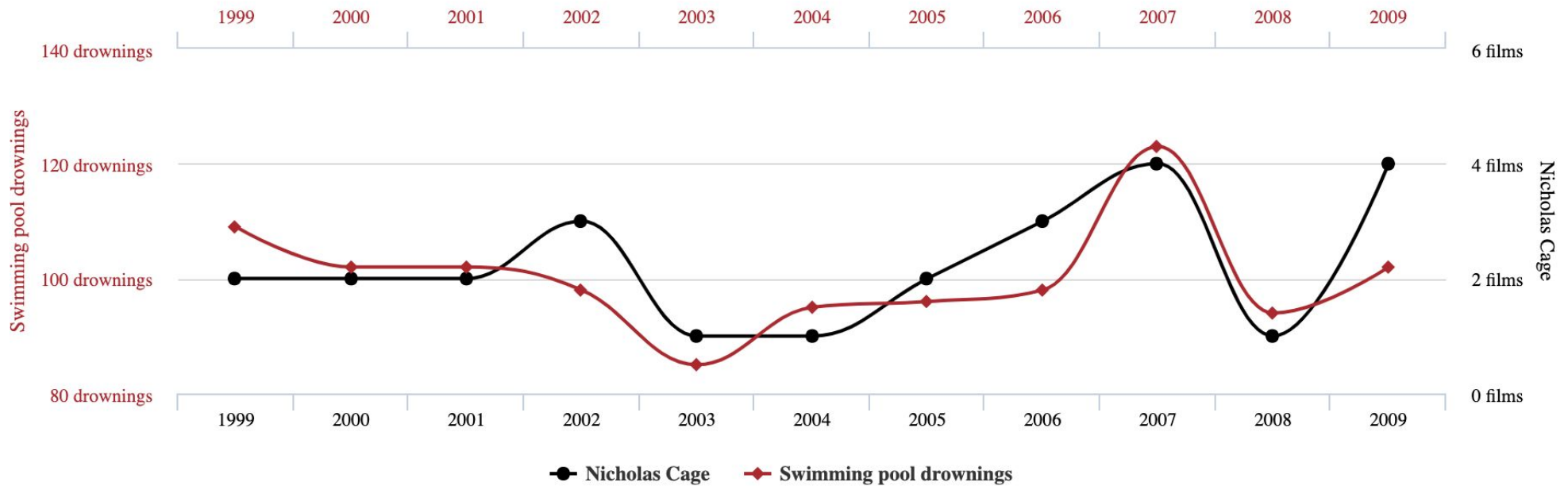
A partir de ciertos datos obtenidos de cada una de esas variables uno estima si hay alguna relación entre ellas.

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

Correlation: 66.6% ($r=0.666004$)



tylervigen.com

Fuente:
<https://www.tylervigen.com/spurious-correlations>

Análisis de Series de Tiempo (TSA)

TSA con Open Power System Data

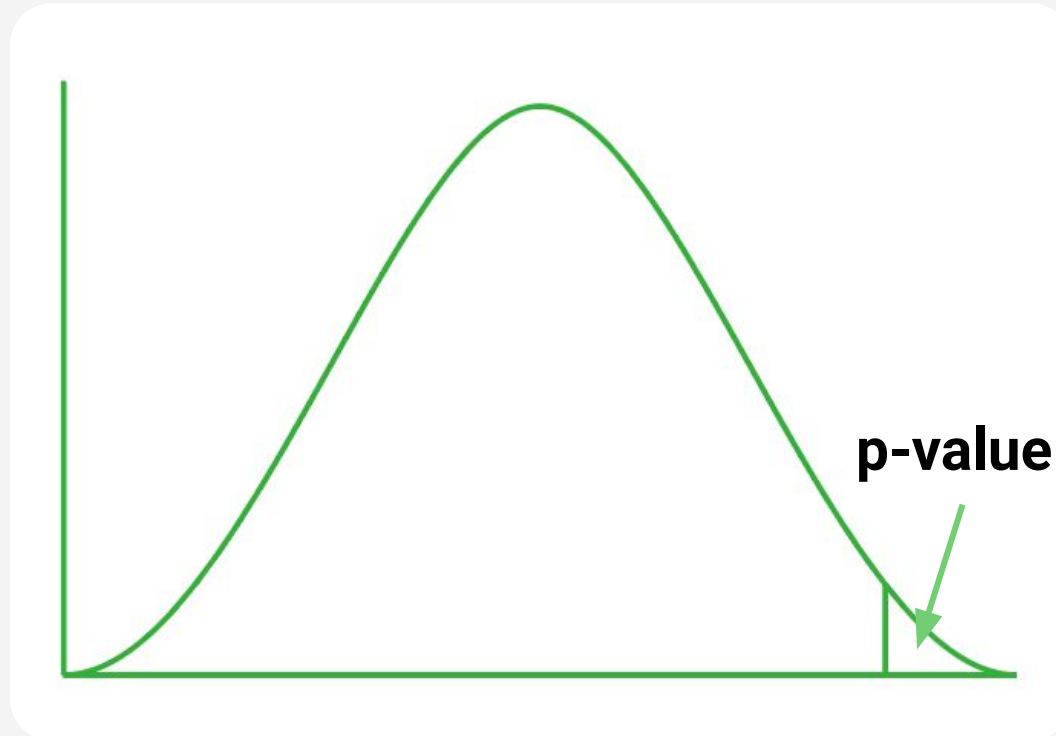
Desarrollo y evaluación de modelos

Etapas de evaluación de un modelo

Pruebas de hipótesis

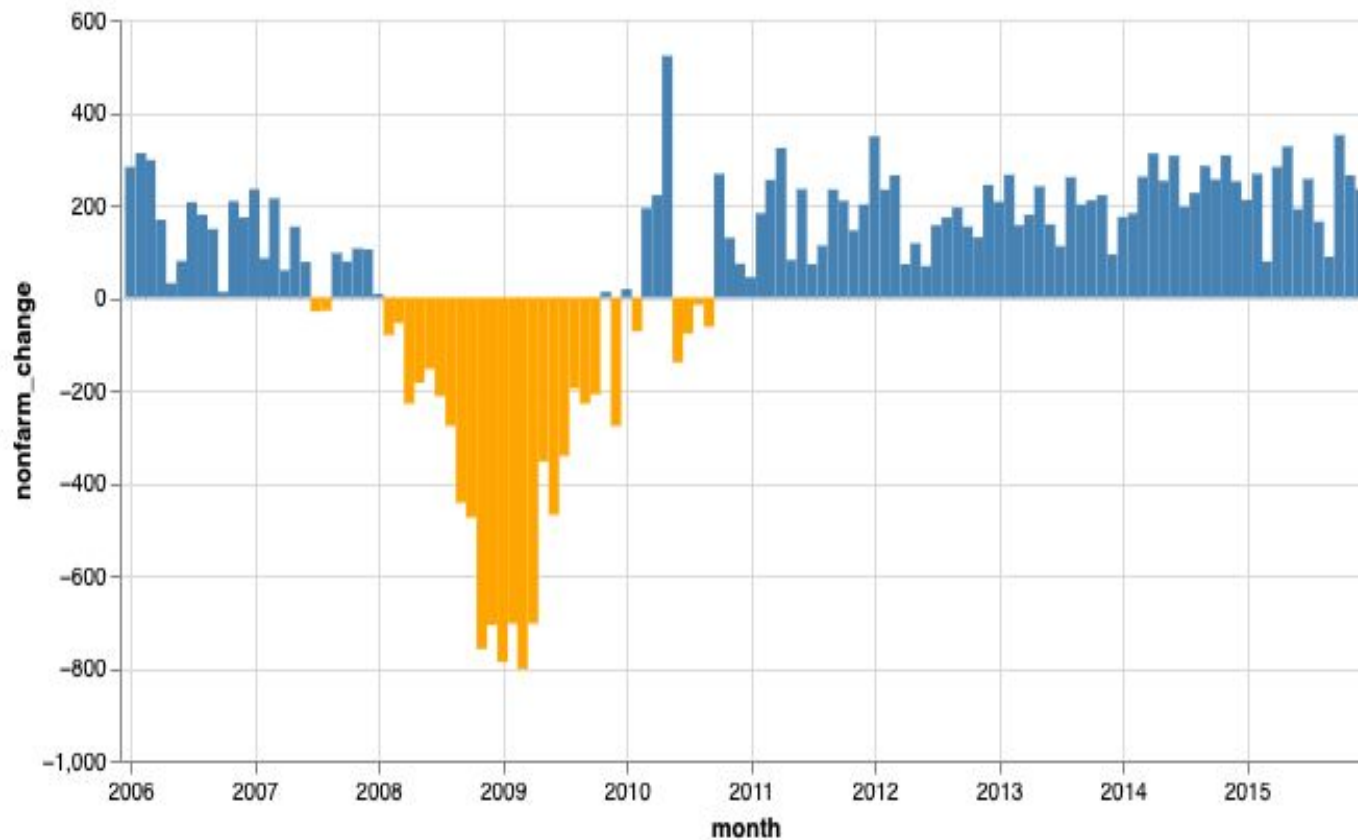
- Anova
- Z-test
- T-test
- Chi-squared test

p-value



El valor p ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativos.

Dividir el conjunto de datos en dos:
entrenamiento y de testing



Ejemplo

1. Entrenas el modelo
2. Evaluación del modelo

$$\text{precisión} = \frac{VP}{VP + FP}$$

$$\text{exactitud} = \frac{VP + VN}{VP + FP + FN + VN}$$

Ejemplo



Ejemplo preciso y exacto



Ejemplo preciso y no exacto

Regresión y evaluación de hipótesis

Métricas de evaluación y regresión

Análisis exploratorio completo
