# Introducción al Análisis de Datos: Programación Estadística con Python

Alberto Sanz, Ph.D. www.linkedin.com/in/alberto-sanz-4b6bb5106
asanz@edem.es

Máster en *Data Analytics* para la Empresa
Octubre de 2020

OBJETIVO

El objetivo central de este curso consiste en conseguir que los alumnos se sientan cómodos sean competentes en la analítica de datos mediante Python. El curso cubrirá desde la recolección y depuración de datos hasta su análisis mediante técnicas estadísticas bivariadas e introducción al modelo de regresión lineal.

Todos los análisis del curso llevarán siempre asociado un respaldo gráfico orientado a la comunicación eficaz de resultados en entornos empresariales.

COMPETENCIAS A ADQUIRIR

Analítica práctica de datos y representación gráfica univariada y bivariada.
Regresión lineal básica.
Fundamentos de programación en Python.
Uso del entorno de desarrollo Spyder
Use de Google Colab Notebooks

CONTENIDO:

## Sesiones 1 , 2 y 3. Introducción a Python y Spyder

(Presencial 5 horas)

Visión general del entorno de programación

- Python como lenguaje orientado a objetos
- Tus scripts no son (sólo) tuyos: buenas prácticas en la escritura de algoritmos. ¡Documenta siempre tu código!
- Aprendiendo a aprender Python:
    - Un poco de orden en la galaxia de Internet:
        - Tutoriales valiosos
        - Blogs confiables y valiosos
- Cargando librerías: os, pandas, numpy y matplotlib
- Leyendo los primeros datos en Python
- Expandiendo y enriqueciendo nuestros datasets: merge
- Nuestros primeros gráficos con Python

*Temas clave:* Python, spyder, pandas, numpy, matplotlib, pd.read_excel, pd.read_csv.

## Sesión 4. Describiendo nuestros datos con representaciones gráficas y numéricas

(Videoconferencia 2 horas)

Instrumentos para medir el mundo: nominales *versus* cuantitativos.

- Bargraphs and percentage tables
- Histograms and descriptives
- Refining our plots
- [Sub-setting]

*Temas clave:* code, python objects, object, category, int64, float64, (& other variable types), subsetting.

## Sesión 5. Sub-setting

(Videoconferencia 2 horas)

- Seleccionando variables relevantes
- Seleccionando casos válidos o de interés
- Gestión de las distintas versiones de nuestro dataset.

*Temas clave:* pandas, sub-setting.

## Sesión 6 . Mejorando nuestros análisis descriptivos: Gestionando lo inesperado: Limpieza y formato en nuestros datos

(Presencial 2h)

- Evitando artefactos: ¡Representa siempre tus datos!
- Outliers, wild codes y otros especímenes no es esperados: Refinando nuestra limpieza de los datos.
- Reagrupando, recodificando y recalculando variables: ¡Cuidado! ¡Nunca machaques tus datos!
- Gráficos bivariados: Explorando relaciones en los datos: una cuestión de perspectiva.
- Excursus: Google Colab notebooks

*Temas clave:* na, 99's, special codes in quantitative variables, wild codes, duplicates, general trends vs. specific cases, type coercion, boxplots, scatterplots.



Alberto Sanz, Ph.D. asanz@edem.es

## Sesión 7. Buscando patrones de asociación en nuestros datos. Análisis bivariado (I):

(Presencial 2h)

- Comparación de medias: t tests y ANOVA.
- Informes profesionales con apoyo gráfico.

*Temas clave*: t test, ANOVA, p.value, error bars, confidence interval plots.

## Sesión 8. Buscando patrones de asociación en nuestros datos. Análisis bivariado (II):

(Videoconferencia 2 horas)

- Comparación de porcentajes. Tablas cruzadas y gráficos de barras múltiples
- Informes profesionales con apoyo gráfico.

*Temas clave*: Column vs. row percentages, Chi 2, adjusted residuals, p.value.

## Sesión 9. Buscando patrones de asociación en nuestros datos. Análisis bivariado (III):

(Presencial 2 horas)

- Correlación. Diagramas de dispersión / *Scatter plots*.
- Visualizando lo cuantitativo discreto: *jitter* y tamaño de los marcadores.
- Trucos del oficio: Manejando la no linealidad en un mundo lineal.
- Reflexiones sobre la correlación y la (no) causalidad.

*Temas clave*: Pearson's r, scatterplot, non linearity, jitter.

## Sesión 10. La regresión implica control y predicción sobre nuestros datos. Regresión (I)

(Presencial 2h)

- Modelizando con una línea: Introducción a la regresión lineal simple.
- Modelo estadístico de la línea de regresión. Ajuste del modelo a la realidad de tus datos.
- Visualizando la regresión simple.
- Predicción: Mantengamos el rango bajo control. ¿Hay alguien ahí?

*Temas clave*: Model fit, R2, residuals, predicted values, confidence interval plot, range, rugs.

## Sesiones 11 y 12. Mejorando el control de factores en nuestros modelos: Modelos de regresión múltiple (Regresión II).

(Videconferencia 2 horas + Presencial 1,5 horas)

- Introducción a los modelos de regresión múltiple. Noción de control.
- Variables independientes dicotómicas en los modelos de regresión.
- Variables independientes nominales de más de dos categorías: Dummies
- Estimando relaciones no lineales en un mundo lineal.



Alberto Sanz, Ph.D. asanz@edem.es

## Sesión 13 Adaptando la regresión a todos los terrenos. Explicar lo cualitativo en un mundo cuantitativo?: Modelos de **regresión logística.** (Regresión III).

( Presencial 1,5 horas)

- Introducción a los modelos de regresión múltiple. Noción de control.
- Variables independientes dicotómicas en los modelos de regresión.
- Variables independientes nominales de más de dos categorías: Dummies
- Estimando relaciones no lineales en un mundo lineal.

*Temas clave*: Model fit, Pseudo R2, residuals, predicted values, dummy variables, **logistic regression**.

## Sesiones 14 y 15 Fundamentos de Programación para la puesta en producción: Estructuras de datos. Control de flujo y expansión de Python con nuestras propias funciones y clases.

(presencial 2 horas + presencial 2 horas)

- Control de flujo
  - Ejecución condicional
  - Bucles
- Estructuras de datos vistas en el curso:
  - Constantes y variables, vectores y tuplas, matrices, diccionarios.
- Funciones
- Todo en uno: Programación orientada a objetos en Python.
  - Objetos: Clases, atributos y métodos.

*Temas clave*: Flow control, if , loops, variable, list, tuple, array, functions, object, class, attibute, method

## Sesión 16 Presentación prueba grupal final. Hasta el infinito y más allá.

(Videoconferencia 2h)

- Presentación de los trabajos finales
- Sumario del curso
- Trucos del oficio.

## EVALUACIÓN

| **Trabajo Grupal** | **Prueba final grupal** |
|---|---|
| Descripción: Problema práctico de análisis e informe | Descripción: Problema práctico de análisis e informe |
| Formato de entrega: Google Colab Notebook | Formato de entrega: Google Colab Notebook |
| Fecha de presentación: Antes de la sesión 7 | Fecha de presentación: Sesión 16 |
| Peso sobre la nota del curso: 30 Por ciento | Peso sobre la nota del curso: 70 Por ciento |



Alberto Sanz, Ph.D. asanz@edem.es

## BIBLIOGRAFÍA Y RECURSOS

Manual de referencia para el curso:
Haslwanter, Thomas (2016) An introduction to Statistics with Python. Ed. Springer

Recursos en la web.
La lista de recursos web es infinita, pero algunos tutoriales de utilidad pueden ser:

- Buenas prácticas (manual de estilo) escribiendo código en Python
  https://pep8.org/

- Sobre tipos de datos en Pandas
  https://pbpython.com/pandas_dtypes.html

- Una extensión específica sobre datos categóricos en Pandas:
  https://pbpython.com/pandas_dtypes_cat.html

- Sobre gráficos en Matplotlib:
  https://matplotlib.org/tutorials/index.html

Cursos on-line.
Para aquellos que quieran avanzar en paralelo sobre los fundamentos de programación que veremos a lo largo del curso.

- https://www.edx.org/es/course/programacion-para-todos-empezando-con-python.



Alberto Sanz, Ph.D. asanz@edem.es