5 PLANTILLAS DE ESTADÍSTICA PARA TRIUNFAR CON TUS PROYECTOS

ESTADÍSTICA PRÁCTICA MADE IN C2

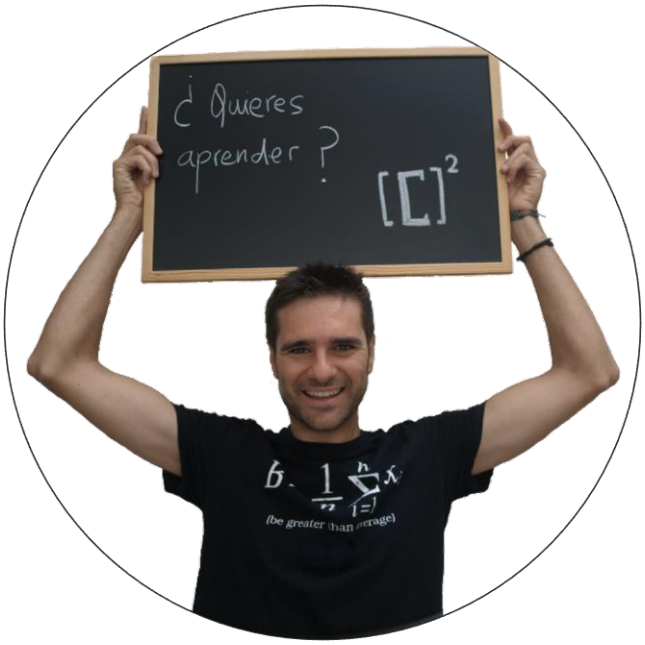
Bienvenid@ a Conceptos Claros

¡Hola!

Primero de todo quiero darte la bienvenida a esta guía y también a Conceptos Claros.

Así que antes de entrar en materia me voy a presentar un poco. Quiero que me conozcas un poco más. Te dejo con mi recorrido vital y por qué nació mi blog: conceptosclaros.com

Mi recorrido vital



Soy Jordi. Vivo en Barcelona con mi compañera de viaje Anna. Soy una mente inquieta y me encanta ser un sherpa de los datos.

Después de 7 años en el departamento de I+D de una multinacional, entendí que aplicar estadística y matemáticas en los datos es apasionante.

Entender con una visión práctica estas ciencias es especial. Me encanta formarme día tras día y buscar un sentido práctico.

Soy Ingeniero Industrial. Estudié el Máster Oficial en Ingeniería Biomédica en la especialidad de señales y Análisis de Datos.

Tengo una capacidad innata para resumir y hacer fácil lo difícil. No temas, estás en buenas manos.

¿Por qué nació Conceptos Claros?

Por otro lado descubrí que me encanta ser un guía, un mentor, un sherpa de otros que están un poco más atrás que yo. Quiero hacer crecer a los demás con mucho cariño y dedicación.

Comprobé este hecho trabajando con becarios codo con codo. Ellos quedaron encantados. Y me di cuenta de que puedo ayudar a muchas más personas. Quiero que tú aprendas como ellos lo hicieron.

¿Por qué es importante esta guía?

La era de los datos ha llegado, y la capacidad para analizarlos es una cualidad buscada y diferenciadora.

Por eso quiero ayudarte a aplicar estrategias de análisis de datos, para que puedes extraer conclusiones útiles para tus estudios e investigaciones.

De esta manera, serás un profesional mucho más completo y con una capacidad diferente al resto. Serás capaz de sumergirte en el mundo de los datos.

Me encantará ofrecerte el camino fácil y proporcionarte material entendible, práctico y paso a paso. Así vas a ganar tiempo y conocimiento.

Ahora ya sabes un poco de mí y por qué puede ser interesante lo que te explique.

Sigue leyendo y te explico más sobre la guía que al final es lo que quieres ☺

Jordi Ollé

¿En qué consiste esta guía?

A ver si te suena esta situación:

Has oído hablar de estadística alguna vez. O incluso has estudiado alguna asignatura en la universidad. Pero llega el momento de la verdad... y no sabes ni por dónde empezar.

No sabes qué técnica aplicar ni cómo utilizar la estadística como tu mejor aliada. Quizá, ahora mismo, es tu peor pesadilla.

Te sientes perdido y sin saber cómo enfocar el análisis de datos que te llevará al éxito de tu proyecto y como profesional investigador.

Si es así, ¡vamos bien! Quiero ayudarte a desbloquear tu mente y a utilizar la estadística como lo que es: una herramienta para brillar como investigador.

Voy abordar una preocupación muy recurrente y seguramente es la que te inquieta ahora mismo.

No sabes qué técnica/método estadístico aplicar para analizar tus datos ni cómo abordar un proyecto real de análisis de datos.

Para ayudarte a resolver esta inquietud voy a darte 5 plantillas (aunque el título de la guía son 4) para que puedas aclarar 5 aspectos que te permitirán solventar esta preocupación.

Te listo estos 5 puntos y así los verás más claro:

1- ¿Cuáles son las etapas de un proceso completo de análisis de datos?
2- ¿Qué es una tabla de datos y cómo está ordenada?
3- ¿Cómo puedo interpretar los datos? (La Exploración)
4- ¿Qué técnica estadística aplico en cada caso? (El Análisis)
5- ¿Qué software utilizo y cómo aprendo a manejarlo?

Las explicaciones de estos puntos las he resumido en formato plantillas. A continuación te muestro estas planillas que te comentaba.

PLANTILLA 1 – EL CAMINO A SEGUIR

¿Cuáles son las etapas de un proceso completo de Análisis de Datos?

Tener la visión global de un proceso completo de Análisis de Datos es fundamental. Y quiero que lo veas como una transformación de información.

Es decir, partes del estado inicial: plantear el problema y los objetivos.

Y pasas un estado final: listas conclusiones basadas en datos reales para poder resolver el problema planteado en la etapa inicial.

PLANTILLA 1 – EL CAMINO A SEGUIR

Te muestro las 6 etapas de un proceso global de análisis de datos des del punto de vista de la transformación de los datos.

Etapa 1 – El Problema

Todo nace de un problema, de una necesidad real. Tu estudio, tu proyecto, parte de esta premisa. Entender mejor la realidad y solucionar el problema que te preocupa. Este es tu objetivo como Analista de Datos.

En esta etapa **definirás el foco del estudio**.

Es decir: qué problema quieres abordar y definirás el objetivo del estudio

Etapa 2 – La Recolección

Es el diseño de un método de recolección de información. Más técnicamente, es un proceso de experimentación.

Puede ser una encuesta, pruebas en laboratorios, con pacientes, nutrirse de datos de marketing en redes sociales, etc.

En definitiva, es un plan de observación de la realidad para poder obtener DATOS.

Etapa 3 – La Limpieza

Los DATOS son observaciones de la realidad, y es un metal precioso en bruto. Es necesario pulirlo y encontrar lo más apreciado.

En esta etapa te encargarás de **homogeneizar los datos en cuanto a formato**, deshacer observaciones que no te interesan, y almacenar las más útiles.

PLANTILLA 1 – EL CAMINO A SEGUIR

Etapa 4 – La Exploración

Los DATOS se visualizan minuciosamente para intuir las pistas más relevantes que se esconden entre números y letras. Es la llamada exploración. En ella utilizarás **la estadística descriptiva** (ED).

Esta rama de la estadística se encarga de traducir los DATOS a gráficos y características sencillamente entendibles para nosotros. De esta forma puedes interpretarlos de manera eficaz y rápida.

Etapa 5 – El Análisis

Es el punto que quizá te esté preocupando. Es momento de responder a las preguntas como investigador con la ayuda de evidencias reales. Aquí entra en juego el conocimiento de técnicas estadísticas, y de tu propia creatividad para combinarlas y extraer las conclusiones que te interesan.

La famosa **estadística inferencial** (EI) es la rama por excelencia de esta etapa. Será tu mejor aliada. Se encarga de extraer conclusiones generales a partir de observaciones de un pequeño conjunto de la realidad, la muestra.

En otras palabras, proporciona herramientas para encontrar conclusiones de un conjunto grande (población) con la información de una pequeña parte de este conjunto (muestra). El contraste de hipótesis es la herramienta más famosa de esta etapa. Pero existen otras técnicas como: la predicción, la clasificación, o los métodos de causa-efecto, entre otros.

Etapa 6 – La Conclusión

Interpretarás los resultados del análisis y **listarás las conclusiones**. En definitiva, la información más valiosa de tus DATOS. Estarás mucho más cerca de solucionar el problema que habrás planteado en la etapa 1.

PLANTILLA 2 – LA MATERIA PRIMA

¿Qué es una tabla de datos y cómo está ordenada?

El Análisis de Datos se alimenta de Datos. Es de cajón. Entender que los Datos son, en realidad, tablas o matrices es una verdadera revelación.

Entender también que las variables son las características que mides de la realidad y se sitúan en columnas es otra clave.

En esta plantilla verás:

- Cómo es una tabla de datos
- Qué son las variables y las observaciones
- Qué tipo de variables son las más comunes en la práctica

PLANTILLA 2 – LA MATERIA PRIMA

Tu base de datos es la puerta que comunica con el mundo real. También puedes imaginarla como la materia prima.

Un proyecto de análisis de datos es un proceso de transformación de DATOS. Algo así como empezar con una simple tabla y llegar a obtener información útil. De DATOS a conclusiones cristalinas, que resolverán gran parte de tus inquietudes.

Los elementos de una tabla de datos

El punto de partida de un buen análisis son tus DATOS limpios en forma de tabla. Vale, pero ¿qué es una base de datos? De hecho, ya te lo he definido.

Es una tabla con filas y columnas. Como **una matriz**. Y cada celda contiene un código alfanumérico.

- Las **filas son observaciones** de la realidad
- Las **columnas son variables**

Fisher's *Iris* Data

Largo de sépalo	Ancho de sépalo	Largo de pétalo	Ancho de pétalo	Especies
5.0	2.0	3.5	1.0	*I. versicolor*
6.0	2.2	4.0	1.0	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.2	2.2	4.5	1.5	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.0	2.3	3.3	1.0	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
5.5	2.4	3.8	1.1	*I. versicolor*
4.9	2.5	4.5	1.7	*I. virginica*
5.1	2.5	3.0	1.1	*I. versicolor*
5.5	2.5	4.0	1.3	*I. versicolor*
5.6	2.5	3.9	1.1	*I. versicolor*
5.7	2.5	5.0	2.0	*I. virginica*
6.3	2.5	4.9	1.5	*I. versicolor*
6.3	2.5	5.0	1.9	*I. virginica*
6.7	2.5	5.8	1.8	*I. virginica*
5.5	2.6	4.4	1.2	*I. versicolor*
5.7	2.6	3.5	1.0	*I. versicolor*
5.8	2.6	4.0	1.2	*I. versicolor*
6.1	2.6	5.6	1.4	*I. virginica*

Las variables son características. Pueden ser longitud, temperatura, densidad, país de procedencia, tipo de tratamiento, calidad del tratamiento, peso, tipo de enfermedad, nivel de estrés, nota final de la asignatura, etc. Todo aquello que puedas medir y listar es una variable.

Las observaciones son las distintas mediciones de las variables. Pueden ser personas, animales, insectos, etc. En general, individuos. O también casos, situaciones, muestras. Al final, para resumirlo: observaciones de la realidad.

Cuantas más observaciones, más rica será tu base de datos. Más grande será. Y más información tendrás.

En la práctica, el número de observaciones o el número de individuos de una tabla de datos se simbolizan con la letra n pequeña.

Clasificar las variables es muy útil: numéricas, categóricas y ordinales.

Variables numéricas de escala

Expresan cantidad y tienen unidades: densidad (kg/m3), temperatura (°C), peso (kg), longitud (m), edad (años), etc.

A. Numéricas continuas: tienen decimales. Como el peso, la longitud, o la densidad
B. Numéricas discretas: no tienen decimales. Como la edad, o el sueldo en miles de euros

Variables categóricas

Son etiquetas nominales y expresan grupos o nombres. El país de procedencia, género, fumador, nombre del instituto.

A. Dicotómicas: identifican 2 grupos. Fumador o NO fumador, masculino y femenino, alto y bajo, grande, pequeño etc.
B. Politómicas: expresan muchos grupos. País de procedencia, nombres de universidades, carrera estudiada.

Variables Ordinales

Son un tipo de variables categóricas con un sentido de escala: calidad del servicio puede ser malo, regular, bueno, muy bueno. O la importancia de la enfermedad puede ser leve, sin riesgo, grave, muy grave.

PLANTILLA 3 – LA EXPLORACIÓN

¿Cómo puedo interpretar los datos?

El Análisis de Datos tiene dos herramientas muy claras. La exploración y el análisis.

Y para mí, la exploración tiene un sentido muy muy importante.

En pocas palabras explorar significa traducir tu tabla de datos en algo que se entienda.
En algo visual como gráficos o en características sencillas de entender.

PLANTILLA 3 – LA EXPLORACIÓN

Has visto que la tabla de datos son números y letras ordenados en una tabla. Bien. Lo siguiente es interpretar la información escondida en esta tabla. Es momento de entenderte con los DATOS y hablar el mismo idioma.

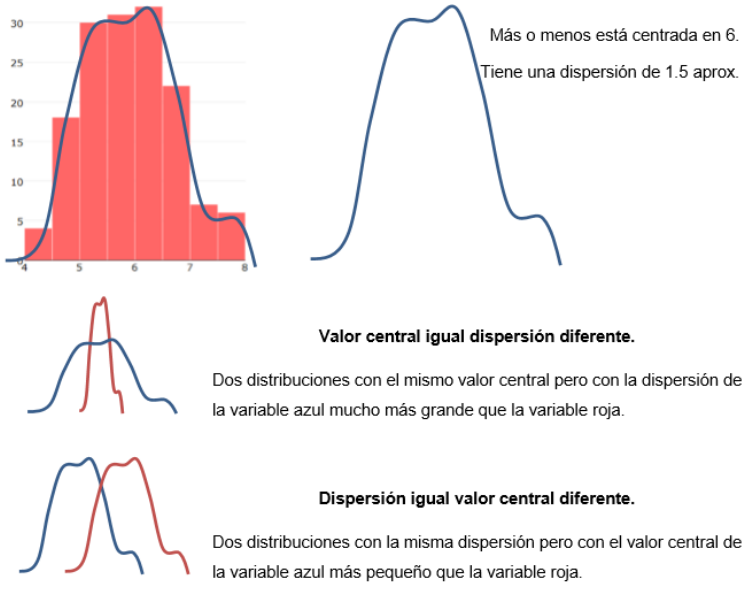
Gráficos y características

El **objetivo** principal de la **estadística descriptiva** (ED) es **utilizar gráficos y características numéricas sencillas** para comunicarte con el mismo idioma que tus datos. Es como un google translator. Las herramientas de ED te ayudan a transformar tu tabla de datos en:

- Gráficos para poder visualizar filas y filas de tu tabla de datos
- Características numéricas para evaluar la posición, centralidad, dispersión y frecuencias.
- Tablas de frecuencias para contar las observaciones de cada grupo o intervalo

La distribución

Las variables numéricas son filas infinitas de números. Pero podemos reordenar estas filas en forma de histograma y conseguir ver su distribución. La distribución es la forma cómo se ordena una variable numérica.

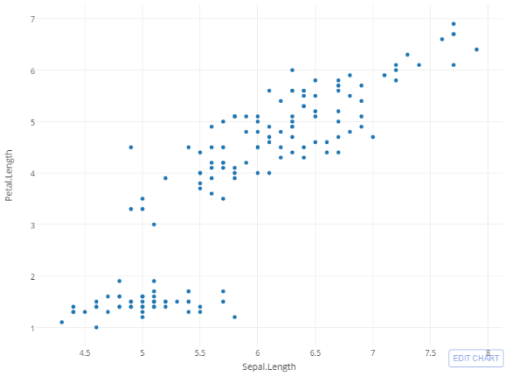


Las dos características de una distribución son:

- La Centralidad: es el valor más céntrico o dónde se concentran los valores. La media y la mediana miden esta característica.
- La dispersión: es el ancho de una distribución. La desviación estándar y la varianza cuantifican la dispersión.

PLANTILLA 3 – LA EXPLORACIÓN

Las relaciones entre variables



Relacionar variables numéricas es una buena praxis. El objetivo es ver a simple vista si dos variables numéricas se pueden relacionar entre sí.

Se utiliza el famoso scatterplot o diagrama de dispersión. Son los valores de 2 variables en el plano 2D en forma de puntos. Para ver posibles relaciones entre ellas.

Las tablas de contingencia

El histograma es, de lejos, la herramienta para resumir una variable numérica. Y en el caso de variables categóricas, utilizarás la tabla de contingencias.

Es una tabla resumen. Contarás las observaciones de cada grupo. La frecuencia es el número de observaciones de cada caso.

TABLA DE CONTINGENCIA DE DOS VARIABLES CATEGÓRICAS: FUMADOR, GÉNERO

Smoke	Gender	Frecuencia	Frecuencia Relativa
yes	male	33	4.552%
yes	female	44	6.069%
no	female	314	43.310%
no	male	334	46.069%
		725	

Smoke indica fumador (yes) o no fumador (no) y Gender, masculino (male) o femenino (female). Dos variables categóricas dicotómicas. Es una tabla de contingencia 2x2.

Te voy a dar acceso a un mini curso de 8 días y te voy a compartir la guía de la exploración. Donde te resumiré los gráficos más utilizados y útiles y cuando utilizarlos.

Atento al mini curso porque recibirás también la guía de la exploración gratuita.

PLANTILLA 4 – EL ANÁLISIS

¿Qué técnica estadística aplico en cada caso?

Como te decía en la plantilla anterior tienes dos herramientas muy importantes: la exploración y el análisis.

Si la exploración te ayuda a entender tu tabla de datos. El análisis te ayuda a sacar conclusiones con evidencias estadísticas.

Con la ayuda de métodos y cálculos estadísticos vas a poder sacar información útil de tus datos. Que es de lo que se trata al final ☺

PLANTILLA 4 – EL ANÁLISIS

Como te decía al principio de esta guía, la preocupación que tienes ahora mismo es no saber qué técnica estadística aplicar para aprovechar tus datos. O dicho de otro modo, qué test estadístico utilizar en cada caso.

Bien. Espero, que con las 3 primeras plantillas te haya situado y tengas un enfoque mucho más práctico. Ahora intentaré responder a tu inquietud número 1.

Estadística Inferencial

¡Si! El análisis estadístico de siempre se basa en **la estadística inferencial**. ¿En qué consiste? En **obtener conclusiones generales** (de una población) **a partir de una pequeña parte** de esta población (muestra). El verbo inferir significa extraer una conclusión general a partir de datos obtenidos de una muestra.

La muestra es una parte pequeña de una población. ¡Y claro! Las conclusiones que saques dependerán de los datos que tengas. O lo que es lo mismo, de cómo hayas escogido tu muestra.

Contraste de hipótesis

Una de las técnicas por excelencia de la estadística inferencial (EI) es el **contraste de hipótesis** (CH). Como es un concepto complicado te he preparado una ficha para explicártelo un poco más en detalle. Espero que esta ficha tengas claro qué es y para qué sirve el contraste de hipótesis.

DESCARGA LA FICHA DEL CONTRASTE DE HIPÓTESIS

Si quieres más también puedes echarle un ojo a este ejemplo.

PLANTILLA 4 – EL ANÁLISIS

2 Tipos de test estadísticos

En la ficha del contraste de hipótesis te explico qué es un test estadístico. Existen dos tipos en la práctica. Los más precisos, pero con más restricciones: test o pruebas paramétricas.

O los menos precisos, pero con menos restricciones: test o pruebas NO paramétricas. Utiliza, siempre que puedas, pruebas paramétricas.

Pruebas Paramétricas: basadas en distribuciones de probabilidad conocidas: como la distribución normal, la t-student, etc. Utilizan parámetros como la media, la desviación estándar, etc. como comparadores. Las restricciones que tienes que cumplir son la normalidad y la igualdad de varianzas (a veces hay más).

Pruebas NO Paramétricas: basadas en rangos y frecuencias. No utilizan las fórmulas de distribuciones, sino que se basan en el rango. En el orden de los datos. Son menos precisas que las paramétricas, pero te pueden servir en muchas ocasiones.

El Mapa Mental del Análisis

Me he dado cuenta las técnicas de análisis de datos se pueden dividir en 6 tipos de problemas. Te he puesto el tipo de problema y un pequeño ejemplo de cada uno:

1. Distinguir si un grupo es diferente a otro. *"Comparación de medias"*
2. Distinguir si las proporciones son diferentes de un grupo al otro.
3. Ver si los grupos tienen relación en la tabla de contingencias. *"Dependencia test Chi-cuadrado"*
4. Analizar si hay relación entre variables numéricas. *"Análisis de Correlación"*
5. Calcular un modelo matemático que permita predecir una variable en función de otras. Por ejemplo *"Regresión Lineal Simple o Logística"*
6. Técnicas de predicción y reconocimiento de patrones. Por ejemplo "Clustering", "Algoritmos de Clasificación"

Para solucionar estos problemas tipo hay un montón de técnicas y escoger la que más te convenga no es tarea fácil. Entendiendo estos problemas tipos puedes escoger la técnica apropiada sin agobios.

Por eso en los próximos días te voy a dar acceso a un mini curso gratuito dónde te podrás descargar un resumen de las técnicas más comunes para solucionar estos 6 problemas tipo.

Y poder utilizar, por fin, la estadística como una herramienta práctica ☺

PLANTILLA 5 – EL SOFTWARE

¿Qué software utilizo y cómo aprendo a manejarlo?

Tener claras las etapas, qué es una tabla de datos, la exploración y el análisis es muy muy importante. Pero la ejecución lo es mucho más.

Hoy quiero hablarte del software que yo utilizo y cómo empezar a manejarlo.

PLANTILLA 5 – EL SOFTWARE

Si has llegado hasta aquí ya has ganado mucho. Habrás visto que para llevar a la práctica todas estas enseñanzas necesitas un PC y un software.

Es momento de practicar con tus datos y empezar a utilizar la estadística en tu realidad como profesional. Y, ¿cuál es el siguiente paso? Básicamente son dos:

1. ¿Qué software utilizo?
2. ¿Cómo utilizo el software?

¿Qué software utilizo?

Una preocupación muy normal derivada de la plantilla 3 y 4, es qué herramienta o, dicho de otro modo, qué software es el mejor para ti. Te lo voy a poner fácil. Existen dos caminos:

CAMINO 1 – El Investigador Científico

No quiero programar y quiero utilizar un software sencillo para poder afirmar mis hipótesis como investigador sin complicarme la vida:

- R + RCommander – software libre con capacidad de calcular análisis estadísticos sin necesidad de programar. Pero puede utilizar funcionalidades de R completas. Porque también te permite añadir sentencias de código.
- SPSS – software de pago y comercial con capacidad muy buena para calcular análisis estadísticos sin necesidad de programar. No puedes crear rutinas repetitivas y tienes que pagar para usarlo.

CAMINO 2 – El Científico de Datos

Quiero convertirme en un Científico de Datos. Aprender una herramienta que me permita crecer como profesional y llegar a ser un técnico e investigador adaptado a la era de los datos y con mayor capacidad técnica.

- R + RStudio – software libre con capacidad para crear análisis ad hoc según lo que necesitas. Es muy fácil de implementar la repetibilidad de tus análisis ya que se utilizan códigos programables y de fácil adaptación. La robustez es la principal característica de este software.

Yo soy partidario de utilizar el software libre. Y me decanto por R. De esta manera no dependes de licencias comerciales.

Fíjate que R cubre los dos caminos que te he planteado. Es igual en que situación estés.

Para que me entiendas un poquito más:

R es el motor de cálculo.

RStudio y RCommander son interfaces de usuario del motor de cálculo R.

Puedes utilizar la que quieras o combinarlas si lo prefieres.

- RStudio necesita que tu entres los comandos a mano
- RCommander funciona a base de clicks.

Los cálculos son los mismos. Aunque para utilizar R al máximo potencial es mejor RStudio.

Si quieres avanzar y ser un buen profesional en análisis de datos, con capacidades más avanzadas te recomiendo utilizar R+RStudio. Utilizarás R a toda máquina.

¿Cómo utilizo el software?

Es cuestión de práctica. Pero mejor empezar paso a paso. Te he preparado una guía de R para que vayas paso a paso y te explico lo esencial para empezar con éxito.

So no te la has descargado aún te dejo con una guía para empezar con R sin morir en el intento.

DESCARGA LA GUÍA PARA EMPEZAR CON R SIN MORIR EN EL INTENTO

¿Y AHORA QUÉ?

¿Cómo aplicar todo esto en tu base de datos?

Te puedo ayudar. Conmigo aprenderás a utilizar todo lo que has visto en estas plantillas y superar lo más difícil: la curva de aprendizaje inicial. Y te ayudaré tanto a aprender la estadística más esencial como utilizarla en la práctica con el software de aplicación.

Si quieres realmente adaptarte a la era de los datos y ser único en tu sector puedes acceder al máster Analiza tus Datos y transformarte en un científico de datos.

¿ME AYUDAS A DIFUNDIR ESTA GUÍA?

Comparte la guía con tus colegas pinchando en el icono de tu red favorita

Ayudo a investigadores y profesionales técnicos a aprender herramientas de análisis de datos para mejorar sus capacidades técnicas y sentirse mejores y más valorados

Jordi

PD: por favor dime qué te ha parecido esta guía contándome tu experiencia a jordi@conceptosclaros.com . Así podré mejorar el contenido y ser más efectivo con mis mensajes.