Práctica 2: Limpieza y Análisis de Datos.

1. Objetivo

El objetivo de este documento es describir la práctica 2 “Limpieza y Análisis de Datos” de la asignatura *Tipología y ciclo de vida de los datos*, asignatura del Máster en Ciencia de Datos de la Universitat Oberta de Catalunya.

2. Miembros del equipo

La práctica ha sido realizada por los estudiantes del Máster en Ciencia de Datos **Francisco Jesús Cárdenas Ruiz** y **Jesús Manuel Montero Garrido.**

3. Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en
- función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

4. Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

5. Resolución

5.1 Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El objetivo de la práctica ha sido el tratamiento de una dataset que fue creado en la práctica 1 por los mismos alumnos del máster. El dataset se denomina *Valores Climatológicos Diarios de España, VCDE*. Para esta práctica nos hemos quedado con los valores climatológicos diarios de España procedentes del sistema AEMET OpenData.

El VCDE se encuentra publicado en la siguiente dirección electrónica:

https://zenodo.org/record/6621823

La descripción de los campos del dataset se encuentra en la tabla 1

ID de campo	**Descripción**	**Requerido**	**Tipo de dato**	**Formato**	**Unidad de medida**
fecha	fecha del dia	S	string	AAAA-MM-DD	N/A
indicativo	indicativo climatológico	S	string		N/A
nombre	nombre (ubicación) de la estación	S	string		N/A

provincia	provincia de la estación	S	string		N/A
altitud	altitud de la estación en m sobre el nivel del mar	S	float		m
tmed	Temperatura media diaria	N	float	Valor decimal con dos dígitos enteros y dos dígitos decimales	grados celsius
prec	Precipitación diaria de 07 a 07	N	float	Valor decimal con dos dígitos enteros y dos dígitos decimales	mm (Ip = inferior a 0,1 mm)
tmin	Temperatura Mínima del día	N	float	Valor decimal con dos dígitos enteros y dos dígitos decimales	grados celsius
horatmin	Hora y minuto de la temperatura mínima	N	string	HH:mm	Hora UTC
tmax	Temperatura Máxima del día	N	float	Valor decimal con dos dígitos enteros y dos dígitos decimales	grados celsius
horatmax	Hora y minuto de la temperatura máxima	N	string	HH:mm	Hora UTC
dir	Dirección de la racha máxima	N	float		decenas de grado sexagesiman codificado
velmedia	Velocidad media del viento	N	float	Valor decimal con dos dígitos	m/s

				enteros y dos dígitos decimales	
racha	Racha máxima del viento	N	float	Valor decimal con dos dígitos enteros y dos dígitos decimales	m/s
horaracha	Hora y minuto de la racha máxima	N	string	Valor decimal con dos dígitos enteros y dos dígitos decimales	Hora UTC
sol	Insolación	N	float	Valor decimal con dos dígitos enteros y dos dígitos decimales	horas
presmax	Presión máxima al nivel de referencia de la estación	N	float	Valor decimal con dos dígitos enteros y dos dígitos decimales	hPa
horapresmax	Hora de la presión máxima (redondeada a la hora entera más próxima)	N	string		Hora UTC
presmin	Presión mínima al nivel de referencia de la estación	N	float	Valor decimal con dos dígitos enteros y dos dígitos decimales	hPa
horapresmin	Hora de la presión mínima (redondeada a la hora entera más próxima)	N	string	Número entero que representa las octas de cielo cubiero por nubes bajas	Hora UTC

Tabla 1: Campos del dataset VCDE

Los campos del dataset VCDE están descritos en la Tabla 1. Los registros del VCDE están formados por los valores climatológicos diarios registrados por las estaciones de las redes sinóptica y climatológica (estaciones con indicativo sinóptico) y de las red mesoscalar (estaciones sin indicativo sinóptico). **El período de tiempo es el comprendido entre el 1 de enero de 1970 hasta el 24 de mayo de 2022**.

La metodología de obtención de los datos es la explicada en el documento de entrega de la primera práctica 1 por los mismos alumnos del máster.

La enorme importancia de las series de datos del conjunto de datos se debe a varios factores: la enorme calidad de las observaciones, a que son observaciones ininterrumpidas durante ya más de 30 años (1 de enero de 1970 - 24 de mayo de 2022), salvo muy cortos periodos de tiempo debido a problema logísticos y de infraestructura de los diferentes estaciones meteorológicas y observatorios.

Se pretende utilizar las series de datos del conjunto de datos VCDE para conocer relaciones entre las diferentes variables, poder hacer regresiones entre los diferentes parámetros y también podría utilizarse para para conocer aspectos de las variaciones interdecadales en la atmósfera, así como para valorar el calentamiento global que se hace especialmente patente a partir de la década de los 80.

5.2 Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

Al conjunto de datos original del dataset de la primera práctica se le han añadido los siguientes campos que inicialmente no estaban incluidos:

- **Nombre:** nombre de la estación. Este campo se ha incluido porque era necesario tener referencias que aportaran más significado que el indicativo.
- **Provincia:** provincia a la que pertenece la estación meteorológica. Este parámetro se incluye para podernos referir a ella en el tratamiento de datos que realizamos
- **Altitud:** altitud sobre el nivel del mar a la que se encuentra instalada la estación meteorológica. Este parámetro se incluye para poder analizar el impacto que pueda tener este parámetro sobre otras variables meteorológicas.

Se tiene una descripción más completa en 001_inspeccion_VCDE.ipynb

5.3 Limpieza de datos

5.3.1 ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Indicado en notebook 002_datos_problematicos.ipynb

5.3.2 Identifica y gestiona los valores extremos.

El conjunto de datos está muy tratado por las unidades de climatología de la Agencia Estatal de Meteorología y tal como se indica en el apartado "*Valores infinitos o muy grandes"* del notebook 002_datos_problematicos.ipynb no se encuentran valores extremos a considerar.

5.4 Análisis de datos

5.4.1 Selección de los grupos de datos que se quieren analizar/comparar (p. e., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)

La selección de grupos de datos que se van a comparar/analizar está descrito en el notebook 004_analisis_exploratorio.ipynb

5.4.2 Comprobación de la normalidad y homogeneidad de la varianza.

La única variable que es normal es la temperatura

5.4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Se ha realizado una cálculo de los estadísticos básicos para las variables numéricas del dataset (media, mediana, desviación típica, mínimo, máximo).

Además, se ha realizado un análisis de correlación de las variables numéricas del dataset.

Finalmente, se ha realizado una visualización de una serie temporal de un ejemplo concreto (Grazalema) para un año completo.

5.5 Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

Se han generado los histogramas de todas las variables numéricas y se observa que la temperaturas (media, máxima y mínima) siguen una distribución normal en cambio el resto de parámetros (viento, rachas y precipitación) no siguen una distribución normal.

Se realizan representaciones gráficas indicadas en el archivo 004_analisis_datos.ipynb

5.6 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

A partir de los datos representados de la estación seleccionada (Grazalema) se observa que las temperaturas tienen una tendencia ascendente. Los valores de finales del año son superiores en 10 grados a los del comienzo del año. Se puede inferir que las temperaturas han aumentado durante el año.

5.7 Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Disponible en el repositorio que se indica en el documento de entrega

6. Agradecimientos

Los datos del dataset VCDE han sido obtenidos de los siguientes orígenes:

- Agencia Estatal de Meteorología de España

Los datos meteorológicos procedentes de OGIMET son copyright del organismo o institución de cada país. Los datos meteorológicos procedentes de AEMET a través del API AEMET OpenData son copyright de AEMET y se reproducen citando a AEMET como autora de los mismos de conformidad con las condiciones de uso y reutilización, la

autorización de reutilización y cesión no exclusiva de derechos de propiedad intelectual, la condiciones generales para la reutilización y resto de condiciones establecidas en Nota Legal de AEMET y Resolución 40 de la OMM

7. Inspiración

La web y el API de OpenData permiten obtener datos meteorológicos libres en virtual de la resolución 40 de la Organización Meteorológica Mundial. Pueden ser reutilizados para cualquier tipo de análisis o exploración.

El conjunto de datos es interesante por las siguientes razones:

- Permite realizar estudios climatológicos a partir de datos obtenidos de las estaciones.
- Evolución del cambio climático.
- La fiabilidad del dato está respaldada por entes públicos y mundiales como los diferentes servicios climatológicos nacionales.

8. Licencia

Se ha optado por “Creative Commons Zero v1.0 Universal” por ser una de las licencias más utilizadas que permiten compartir, adaptar y modificar a partir del material publicado, siempre y cuando se dé crédito al creador original y no se apliquen restricciones adicionales ya sean tecnológicas o legales a los posteriores proyectos derivados del original.

9. Código

El código referido en el apartado 5.7 se encuentra disponible en GitHub en el siguiente enlace:

Disponible en el repositorio https://github.com/fjcardenasuoc/analisis_vcde

10. Dataset

El dataset actualizado ha sido publicado en

https://zenodo.org/record/6621823#.Yp-4CqhByUk

Vídeo

El vídeo de la práctica está disponible en el siguiente enlace de Google Drive de la UOC:

https://drive.google.com/file/d/1qpYSSnvnB1iDbWtlz9IPthOCdHPyrlt-/view?usp=sharing

Bibliografía

- **Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.**
- **Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2 Scraping the Data.**
- **Sene, K. (2016). Meteorological Observations. In: Hydrometeorology. Springer, Cham. https://doi.org/10.1007/978-3-319-23546-2_2**
- **Redes de observación de superficie y en altura: http://www.aemet.es/es/idi/observacion/observacion_convencional**
- **OGIMET: https://ogimet.com/**
- **AEMET OpenData: https://opendata.aemet.es**
- **Resolución de 30 de diciembre de 2015, de la Agencia Estatal de Meteorología, por la que se establecen los precios públicos que han de regir la prestación de servicios meteorológicos y climatológicos.**
- **WMO Global Observing System: https://public.wmo.int/en/programmes/global-observing-system**
- **Resolución 40 de la OMM: https://community.wmo.int/resolution-40**
- **Nota Legal de AEMET: http://www.aemet.es/es/nota_legal**