# Visualización de datos

Práctica - Leroy Deniz Pedreira

# Práctica 1

## Sobre el juego de datos

El dataset elegido para esta práctica es *Data Science Salaries 2023*, accesible a través Kaggle desde este enlace, conteniendo información sobre las retribuciones salariales de profesionales en ciencia de datos de una diversidad de países.+

## Justificación

A la hora de elegir una disciplina en la cual especializarse, existen una gran variedad de motivos por los cuales decantarse en pro o en contra. Un factor nada despreciable es el salario, donde puede ser determinante a la hora de decidirse por esta especialidad en comparación con otras áreas.

Además, desembarcar en un área que no es la propia tiene ciertas incertidumbres de cara a aceptar una nueva posición. Conocer en términos generales los salarios percibidos facilita una de las grandes preguntas incómodas de una entrevista laboral. Naturalmente, se entiende que los datos aquí recogidos serán puramente orientativos.

Otro ítem que cabe destacar, es la evaluación de posibilidades de internacionalización, pudiendo ser un diferencial contar con información sistematizada sobre estos datos. Se puede enfocar a un mercado en detrimento de otros.

## Relevancia

El dataset, como su nombre lo indica, está actualizado a 2023. La última versión actualizada es de aproximadamente 20 días desde la fecha actual, sobre el 15 de abril.

Si bien no existe ninguna variable que determine el sexo en cada registro, se toman en cuenta diversos tipos de posiciones que apuntan específicamente al rol de cada posición, descartando los sesgos introducidos por el género al menos para el objetivo de este análisis.

El colectivo al que apunta este dataset y el público objetivo del análisis no son otros que todos los estudiantes y profesionales del área de la ciencia de datos. Como se ha comentado en el apartado anterior, presenta una utilidad de referencia a la hora de negociar una posición, así como también a la hora de decidir entre ofertas, puesto que es una disciplina en auge y existe una oferta laboral mucho mayor que la demanda de estos roles.

# Complejidad

El dataset se puede descargar en formato CSV. Este se compone de un total de 3755 registros, con datos en once columnas, siete de ellas son categóricas, las otras cuatro numéricas enteras. Es decir, tenemos presente en este dataset una serie de variables cuantitativas y cualitativas que dan una visión general del problema.

Las columnas presentes son las siguientes:

1. **work_year:** año en que se pagó el salario (cualitativa)
2. **experience_level:** nivel de experiencia en el trabajo durante el año (cualitativa)
3. **employment_type:** tipo de empleo para el puesto (cualitativa)
4. **job_title:** puesto en el que se trabajó durante el año (cualitativa)
5. **salary:** monto total del salario bruto pagado (cuantitativa)
6. **salary_currency:** moneda del salario pagado en código ISO 4217 (cualitativa)
7. **salary_in_usd:** salario en USD (cuantitativa)
8. **employee_residence:** país de residencia principal del empleado durante el año de trabajo en código ISO 3166 (cualitativa)
9. **remote_ratio:** cantidad general de trabajo realizado de forma remota (cuantitativa)
10. **company_location:** país de la oficina principal del empleador o sucursal contratante (cualitativa)
11. **company_size:** media del número de personas que trabajan en la empresa durante el año (cualitativa)

Se entiende que el conjunto de datos no es excesivamente simple, aunque tampoco es excesivamente complejo. Tiene un tamaño suficiente para poder identificar datos por canales alternativos y verificarlos sin perderse en millones de registros.

# Originalidad

El dataset tiene apenas unas pocas semanas, por lo cual la originalidad no es precisamente por la temática sino por la frescura de los datos. Pueden haber, y seguramente lo hayan, otras versiones más antiguas de datasets similares que muestren tendencias salariales para determinados perfiles profesionales, sin embargo el tiempo para construir un número de visualizaciones significativo con este dataset en particular es poco.

Cabe destacar, que al ser un dataset público en Kaggle y tratar una temática transversal a toda una profesión en auge, se tienen dudas suficientes de que no existan visualizaciones aunque no estén aún publicadas.

Tras una búsqueda en la web, no se han detectado visualizaciones sobre este dataset en particular. Este dataset ofrece la evolución en el tiempo de diversos perfiles profesionales asociados a la ciencia de datos, así como la localización geográfica de los registros. Esto facilita la utilización de mapas que permitan visualizar los perfiles mejor pagados por país, por ejemplo.

# Objetivos

El uso de este dataset para una visualización, tiene por objetivo responder, entre otras, a las siguientes preguntas:

a) ¿Cuál es el rol perfil mejor pagado y en qué país está?
b) ¿Cuál es la relación entre salario y trabajo remoto para esta disciplina?
c) ¿Cuál es la distribución de niveles de experiencia por país?

Estas primeras preguntas pueden ser perfectamente contestadas mediante los datos del dataset. Se tiene presente toda la información relativa a las variables que es necesario cruzar, así como un número de registros considerable para poder inferir un parámetro. Los datos son históricos, por lo que pueden plantearse preguntas relacionadas a la evolución de los salarios por país o intentar medir la fortaleza de la relación experiencia/salario.

# Práctica 2

## Acceso a las visualizaciones

Visualización 1 - Cargos mejores pagados

Visualización 2 - Distribución de modalidad de perfil

Visualización 3 - Distribución de experiencia por país

## GitHub

https://github.com/leroydeniz/VD-PRA1

## Guión

Buenas noches, mi nombre es Leroy Deniz y a continuación presentaré la práctica 2 de Visualización de Datos del Máster en Ciencia de Datos de la UOC.

Para recordar, el juego de datos utilizado corresponde a los salarios del área de data science en los últimos tres años.

Esto permitiría, por ejemplo, identificar oportunidades de formación de cara a aspirar a puestos cuyo salario sea más elevado, y elegir el país donde estén las mejores oportunidades.

La **visualización 1** presenta la media de los salarios en los últimos tres años, organizados por posición y ordenados de forma ascendente.

A partir de esta información, visualizada en un gráfico de barras, puede verse que los perfiles peores pagados en su media son los de:

- Power BI Developer y Product Data Scientist.

En el lado opuesto, los perfiles mejores pagados son los de:

- Manager
- Lead
- Director
- Principal

Sin embargo, tanto los primeros como los últimos tienen unos pocos casos de muestra como se identifica por el grosor de las barras.

Así pues, el puesto que mejor paga en toda la muestra es de 450k dólares mientras que en la media, es superado por Data Science Tech Lead con 375k.

La mayor cantidad de registros están en los cargos de Data Engineer, Data Science y Data Analyst (en menor medida).

Respondiendo a la pregunta 1 sobre qué perfil es el mejor pagado y en qué país, este es el de Data Science Tech Lead en Estados Unidos.

Recordar en todo momento que hablamos de medias y no LA posición que mejor pagada está.

Por otro lado, con respecto a la **visualización 2**, aquí se busca ver el comportamiento entre salario y el tipo de posición. La visualización muestra los datos ordenados en tres grupos según el tipo de posición, donde:

- 0 significa trabajo 0% remoto o presencial.
- 50, posición híbrida repartiendo el tiempo en 50-50.
- 100, trabajo completamente remoto o full remote.

Los colores identifican el tamaño de la empresa y, el tamaño de cada punto se corresponde con la proporción de sueldo.

Tal como se puede ver, aquellos trabajos en calidad de híbrido, están presentes en los salarios más bien bajos, lo que puede corresponderse con posiciones en empresas que van migrando desde el paradigma convencional o, también, posiciones de ingreso.

Los colores indican también que en esta categoría son mayoría las empresas grandes, afirmando la convencionalidad comentada previamente.

Ahora bien, en aquellos trabajos totalmente remotos o presenciales, toman protagonismo las empresas medianas identificadas con una M. Estos cargos abarcan una concentración salarial más amplia desde los salarios más bajos, presentando muchos más casos de salarios altos frente a la modalidad híbrida.

Respondiendo a la pregunta 2 planteada en la práctica 1 sobre la relación entre el salario y modalidad de trabajo, basándome en los datos se podría decir que, a excepción de las posiciones híbridas, parecen estar bastante equiparadas tanto en remoto como en presencial.

Finalmente, en cuanto a la **visualización 3**, aquí se busca ver la relación entre la experiencia de los puestos y el país.

Se ha utilizado un gráfico de burbujas por cada país, donde cada una representa un nivel de experiencia.

- EN -> Entry level o inicial
- MI -> Mid-level o intermedio
- SE -> Senior o perfil de conocimiento consolidado
- EX -> Executive, experto o profesional

En el tamaño de cada burbuja se recoge el cardinal de cada experiencia.

Si tomamos como ejemplo el caso de España, se tiene una mayoría de niveles Senior e intermedios, y apenas unos pocos registros identifican a los perfiles iniciales y ejecutivos.

La misma tendencia se puede ver en el caso de Estados Unidos pero con un número significativamente mayor de datos.

Con esta explicación, se responde a la tercer pregunta de la práctica 1 en cuanto a relación entre experiencia y país.

De estos datos se puede tomar como síntesis:

- Primero, que el abanico de salarios percibidos en esta disciplina es ampliamente variado y, además, bastante elevado en posiciones con buena formación.

- Segundo, que de estos datos se desprende que las posiciones presenciales y remotas no están vinculadas a una cota salarial sino que, por el contrario, ambas están a la cabeza de aquellas retribuciones más altas.

-Tercero, que el nivel profesional más alto es un cargo que no sobra a nivel de contratación, pero que sin embargo tampoco es de los perfiles más contratados. Por el contrario, aquellos perfiles Senior o Middle son los que aseguran una posición cómoda de buena demanda.

Muchas gracias