Informe sobre el dataset Data Science Salaries 2023

Borgo Martin, Sandoval Jose, Molina Leandro

8 de mayo de 2023

Resumen

El informe analiza el conjunto de datos de salarios de ciencia de datos del año 2023 de Kaggle, proporcionando un análisis detallado de las variables incluidas en el conjunto de datos y comparación con otro conjunto de datos. Se utilizó Python y Excel para la realización de gráficos y tablas. Este informe proporciona información para los profesionales de la ciencia de datos y los interesados en comprender el mercado actual.

Índice

1. Introducción

Los trabajos para la ciencia de datos (data science) continúan en crecimiento, y las organizaciones están compitiendo para contratar a los expertos de la materia. Con el objetivo de entender este ambiente, se ha creado el siguiente informe. Este informe tiene como objetivo proporcionar un análisis a los datos obtenidos, que incluye información sobre la experiencia, el tipo de empleo, el cargo y la ubicación geográfica de los profesionales de la ciencia de datos, así como la moneda en la que se les paga, como también un análisis a la distribución de los salarios (usando coeficiente de Gini) y se termina comparando con otro dataset del mismo ámbito. Se espera ayudar a los profesionales o interesados del mercado data science a que comprendan las tendencias del mercado laboral.

2. Dataset Utilizado

Para este informe se utiliza el dataset proporcionado por kaggle[1]Es una plataforma que aloja competencias de aprendizaje automático, conjuntos de datos públicos y privados, y herramientas de exploración y análisis de datos. Este dataset contiene información sobre los salarios de data science y en diversos países del mundo. El conjunto de datos incluye varias variables, como el año de trabajo, el nivel de experiencia, el tipo de empleo, el título del trabajo, el salario bruto total pagado, la moneda en la que se pagó el salario y el salario en USD. También incluye información sobre el país de residencia del empleado, la cantidad de trabajo realizado de forma remota, la ubicación del país de la empresa y el tamaño de la empresa. Es un dataset confiable ya que ningún dato es "nulo".[2]

3. Análisis de variables y graficas

A continuación se analizaran la distribución de los salarios en dólares de acuerdo a la experiencia laboral[3] de cada uno de los encuestados (que participaron del dataset) que pertenecen al sector del Data Science, cada nivel de experiencia será evaluado mediante una gráfica de cajas resulta muy útil a la hora de determinar los sesgos y la dispersión de los datos, además de ofrecer información sobre la media y la mediana. Se calcularon las medias aritméticas y medianas para cada nivel de experiencia por el salario en dólares véase la Tabla 1.2. Como en la tabla 1.1 se calcula la cantidad de empleados por cada nivel de experiencia para todos los países.

[1] Véase: Salaries of Different Data Science Fields in the Data Science Domain

[2] Véase: en la sección basic exploring.

[3] En el dataset están referidos como:
EN, que se refiere a Entry-level / Junior
MI, que se refiere a Nivel medio / Pre-Senior.
SE, que corresponde a Senior-level / Experto.
EX, que se refiere a nivel ejecutivo / Director.

En la figura 1.1 se representa la distribución de los salarios de los analistas Junior (320) que forman parte de la muestra. A simple vista podemos notar que la distribución de los datos posee un gran sesgo positivo, eso quiere decir que los datos están más agrupados por debajo de la media. Realizando un análisis más riguroso con esos datos encontraremos que la mediana es 70000, esto significa que el 50 % (106) de los trabajadores Juniors se encuentran cobrando menos de ese monto y el 50 % restante más de ese monto. En este caso, como en los que veremos a continuación, la mediana resulta mucho más representativa, ya que la media se ve afectada por los valores altos como 300000 o 220000. Por lo tanto, la media no puede considerarse una buena medida para representar este conjunto de datos. Enfocándonos esta vez en la dispersión de los datos, si calculamos la desviación estándar, nos arrojará 52225.42 como resultado. Antes de seguir el desarrollo conviene hacer una aclaración sobre esto, debido a que estamos analizando salarios anuales en dólares y considerando los datos que se muestran en la figura 2.1, donde vemos que el 95 % de los encuestados actualmente residen en países donde las rentas son elevadas, consideramos que una desviación estándar de 10000 es buena en este caso. Con este punto aclarado podemos afirmar entonces que:

1. Los datos de la figura 1.1 están muy dispersos en esa distribución.

2. Debido a lo antes mencionado la media tampoco es muy representativa para ese conjunto de datos.

Pasa algo muy similar en las figuras 2.2 y la 3.1, donde se representan a los empleados con un nivel Pre-Senior (805) y Senior (2516) respectivamente. Si bien presentan algunas cualidades diferentes como el hecho de que en ambas figuras existe un sesgo positivo pero es casi mínimo en comparación con la primera gráfica analizada. Sucede lo mismo, la desviación estándar es extremadamente alta 54387.68 y 52225.42 respectivamente, lo que al igual que la distribución anterior deja a la media aritmética casi sin mucha importancia, sin contar que las gráficas que estamos observando ahora cuentan con mucho más valores atípicos, un segundo motivo para descartar a la media. Este mismo hecho se repite para la figura 4.1, donde están representados los analistas expertos, con la particularidad de que, al contrario que las distribuciones presentadas anteriormente, las cuales poseen un sesgo positivo; esta aparenta ser simétrica, porque si nos ponemos estrictos la media y la mediana tendrían que ser las mismas para ser considerada simétrica, cosa que no pasa aquí tampoco. Hay que tener en cuenta que de todas las muestras analizadas esta última es la que menor cantidad de datos tiene (114), en comparación con las demás que tienen muestras mucho mayores. Es probable que a medida que vayamos agregando más datos a este grupo la distribución de estos mismos irá cambiando. Algo a destacar es la dominancia de Estados Unidos en cantidad de empleados residiendo en su país, en la figura 5.1 donde se agrupan la cantidad de encuestados agrupados según los 5 países donde hay más residencia, discriminados a su vez según experiencia laboral, y tabla 1.3 donde se muestra los datos brutos con los que se construyó la gráfica. En esa tabla podemos ver la dominancia estadounidense en los 4 niveles de

experiencia en los 4 países con mayor cantidad de residentes (como también el total de empleados).

4. Desigualdad e igualdad en salarios en ciencia de datos.

Por último analizaremos qué tan desigual es la distribución de los salarios para cada uno de los niveles de experiencia laboral. Para esto utilizaremos el coeficiente de Gini.[4] Una vez calculado para cada uno obtenemos que el grupo con la desigualdad más alta en los salarios es el de los Juniors con un coeficiente de Gini de 0.3663, siguiéndole el grupo de los analistas Pre-Senior (nivel medio) con 0.2788; y por último los grupos que abarcan a los Seniors y Expertos en la cual el coeficiente de Gini es muy similares con 0.2058 por parte de Seniors y 0.2028 para Ejecutivos. Tomando todos estos datos podríamos llegar a concluir, que a medida que se tiene más experiencia laboral en el sector del Data Science, menor es la desigualdad salarial. Lo antes dicho se debe tomar con precaución porque:

1. Es probable que el dataset no esté hecho para que sea representativo siquiera de para una población nacional, muchos menos para que sea representativo a escala global.

2. Si las muestras de todos los niveles de experiencia fueran proporcionales se podría obtener más fiabilidad en las comparaciones, ya que por ejemplo en el caso de los analistas expertos tan solo hay 114 en todo el dataset, a diferencia de los seniors que son 2516.

3. El objetivo de este informe es estudiar este dataset a profundidad, no se pretende realizar un análisis representativo, pero se intenta hacer lo posible.

5. Comparativa de residentes y modalidad de trabajo entre dataset.

Para el siguiente análisis de datos hemos optado por comparar los datos actuales con los de otro dataset distinto[5]. Hemos encontrado un dataset, que al igual que el antes estudiado, presenta una cantidad muy similar de residentes en Estados Unidos, unos 2347 contra los 3004 del dataset analizado. Esto con el fin de contrastar los datos y ver que tan diferentes las muestras. Optamos por analizar cuál es la modalidad de trabajo más común de acuerdo a la experiencia laboral que posean, véase la figuras 6.1 y 6.2. Podemos observar que en ambas gráficas la preferencia es muy parecida, si bien varía la cantidad de datos de cada

[4] La función utilizada para los cálculos se encuentra en la sección de referencias.
[5] Véase acá: AI/ML Salaries — Kaggle

nivel de experiencia laboral, se sigue notando que en la mayoría de los casos las empresas contratan a empleados para que realicen el trabajo de manera presencial o de forma remota. Incluso tomando las tablas 2.1[6] y 2.2 correspondientes a cada gráfica, podemos notar que la diferencia porcentual es casi mínima. Por ejemplo, el porcentaje de Seniors que trabajan remoto en el primer dataset es de 44,47 % mientras que en el otro es del 46,52 %, exactamente lo mismo ocurre en el nivel de Pre-Senior (el segundo grupo que más datos tiene) con un 59,45 % de presencialidad para el primer dataset y 46,09 % de presencialidad para los datos del segundo dataset, si bien la diferencia en este caso es más elevada, no es tan abrupta.

Referencias

[1] El dataset analizado y proporcionado por los docentes de la catedra de Estadistica y Probabilidad.

[2] Véase en la sección basic exploring: Fue utilizado para verificar si no habia datos nulos.

[3] El dataset comparado con el que nos proporciono los docentes.

[4] Se proporciona el código fuente de las graficas presentadas en este informe, a excepción de los diagramas de caja que fueron hechos con excel. Para ver la mayoría de graficas presentadas haga clic en la siguiente URL: **"salarios-DataScience.ipynb"**. Si quiere ver todas las graficas a recomiendo que se ejecute el archivo como también el csv que esta proporcionado con el código fuente. Haga clic aca para ver la carpeta

[5] Si quiere ver las graficas realizadas con el dataset que se comparo vea el archivo **"salarioComparar.ipynb"** igualmente se proporciona el dataset utilizado.

Figuras y tablas.

[6] Esta tabla tiene en la fila EX columna de los híbridos (50 %) tiene el símbolo "-" representa que no hay ningún ejecutivo trabajando en esa modalidad, por supuesto de USA.

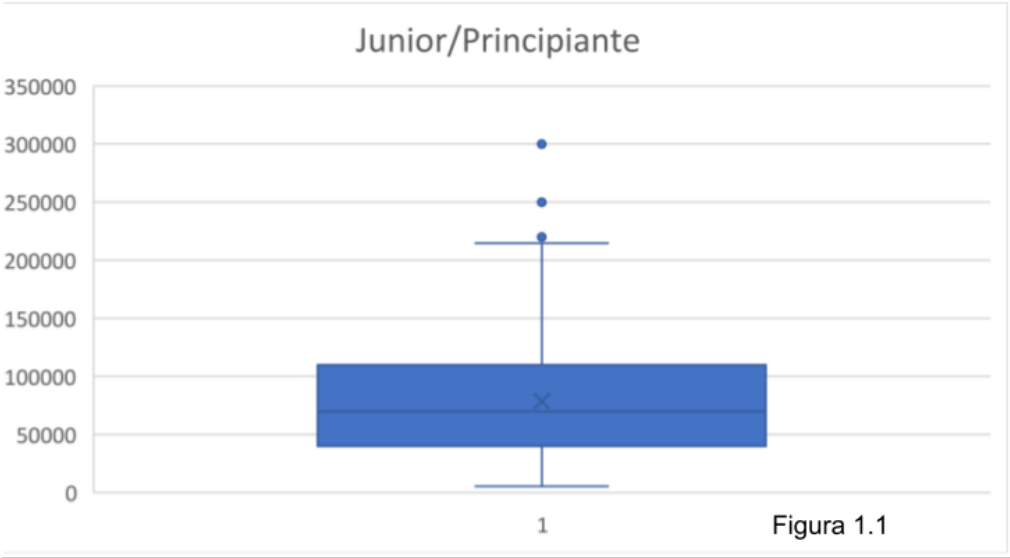

Figura 1.1

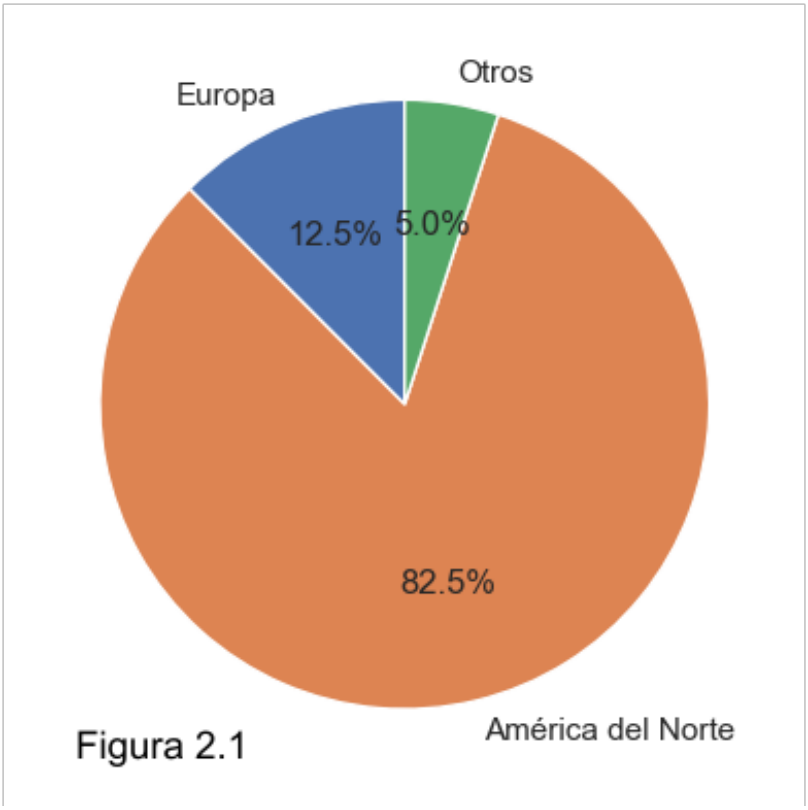

Figura 2.1

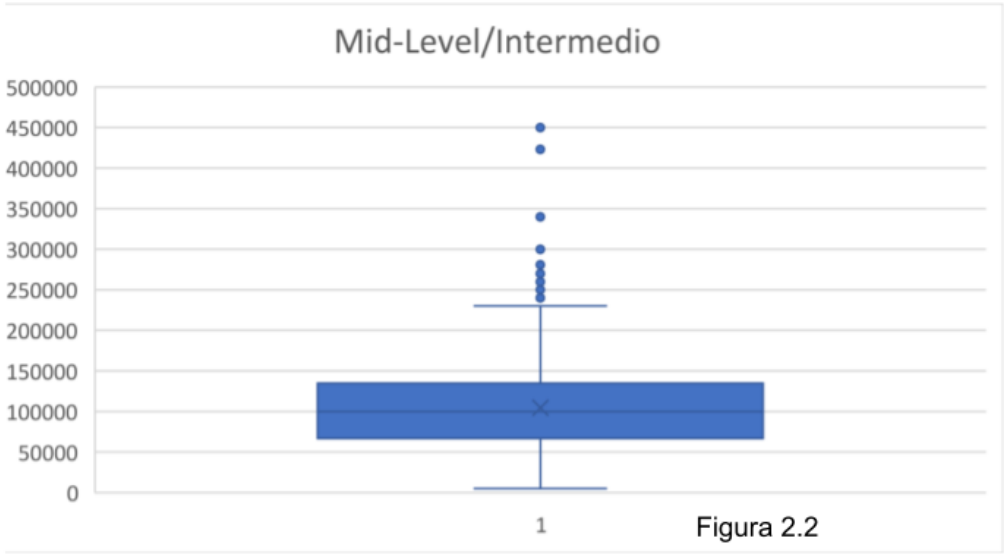
Mid-Level/Intermedio
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
1

Figura 2.2

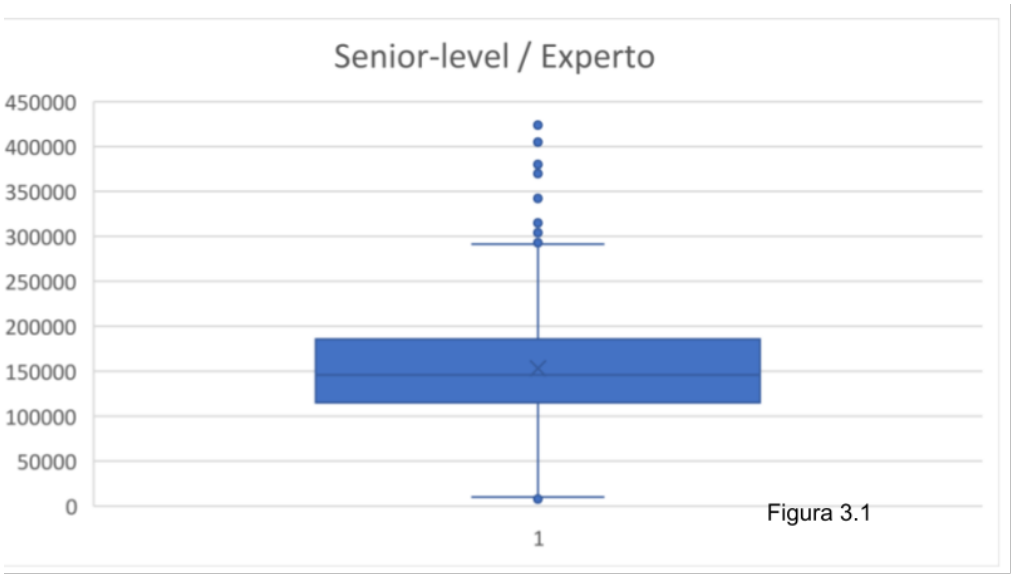

Figura 3.1

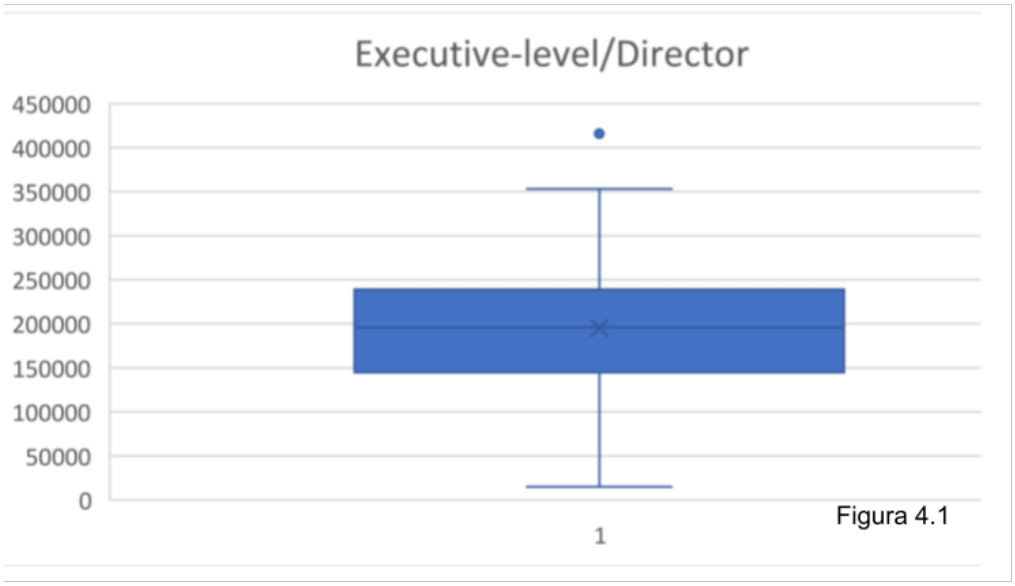

Figura 4.1

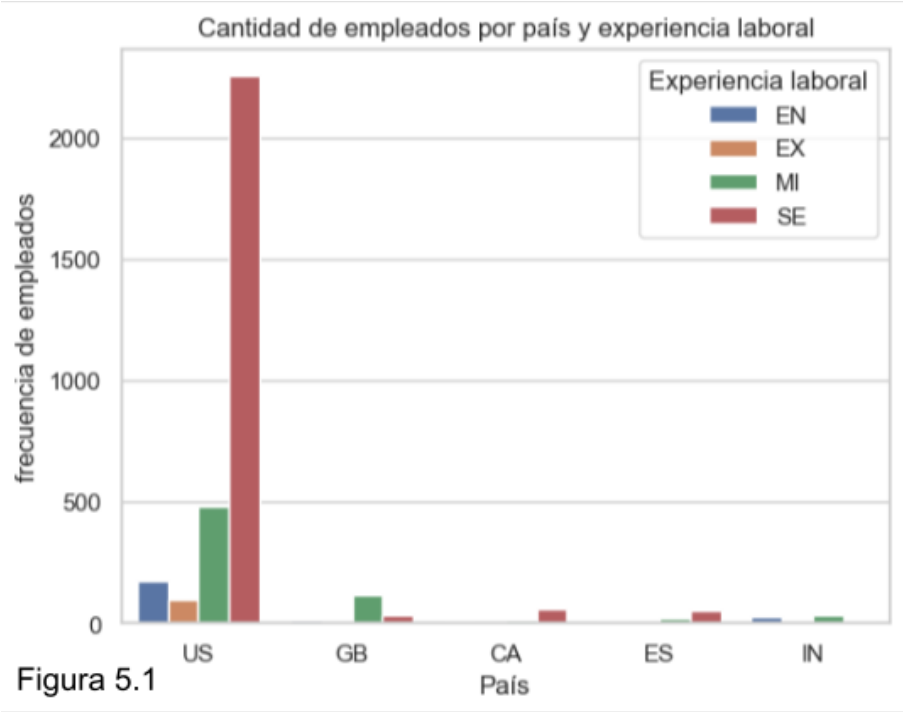

Figura 5.1

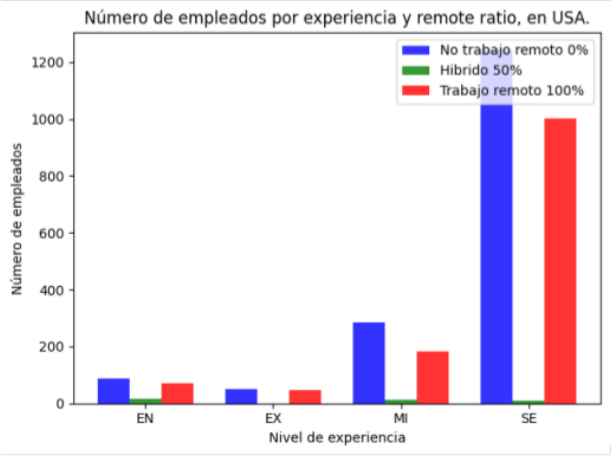

Figura 6.1
Dataset original

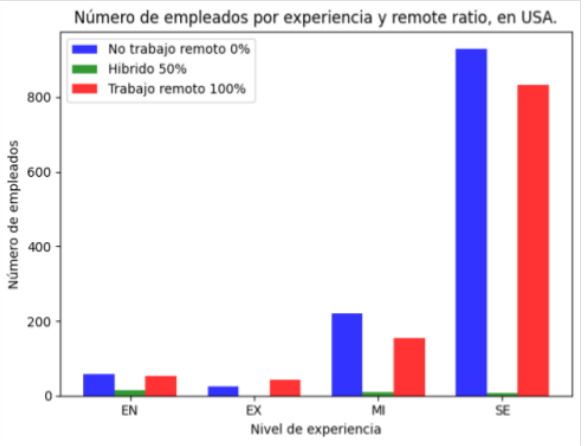

Figura 6.2
Dataset encontrado

Nivel de experiencia	Frecuencia
EN	320
EX	114
MI	805
SE	2516

Tabla 1.1

Tabla 1.2

Nivel de experiencia	Media Salario en USD	Mediana Salario en USD
EN	78546.284375	70000
MI	104525.939130	100000
SE	153051.071542	146000
EX	194930.929825	196000

Nivel de experiencia	EN	EX	MI	SE	Total
Paises donde viven los empleados					
US	173	97	481	2253	3004
GB	13	2	117	35	167
CA	7	3	15	60	85
ES	5	1	22	52	80
IN	26	2	32	11	71

Tabla 1.3

Trabajo remoto en USA.	0	50	100
Experiencia de nivel			
EN	89.0	15.0	69.0
EX	49.0	-	48.0
MI	286.0	12.0	183.0
SE	1243.0	8.0	1002.0

Tabla 2.1
Dataset original

Trabajo remoto en USA.	0	50	100
Experiencia de nivel			
EN	75	60	121
EX	31	6	46
MI	313	71	295
SE	1021	42	925

Tabla 2.2
Dataset encontrado