Los salarios: análisis estadístico

Inteligencia Artificial avanzada para la ciencia de datos - TC3006C

Grupo 101

Portafolio de Análisis - Módulo 1

Cristofer Becerra Sánchez - A01638659

***Resumen*—El vertiginoso crecimiento de campos como la Ciencia de Datos provoca preguntas interesantes sobre el mercado laboral de áreas emergentes como esta; en específico, preguntas como ¿el salario promedio de un Ingeniero en Datos es mayor al de un Científico de Datos?, ¿hay diferencia significativa en el salario promedio en función de la experiencia?, y ¿el salario promedio ofrecido por una empresa en Estados Unidos es mayor que en Reino Unido? Tales preguntas ayudan a esclarecer las diferentes condiciones en las cuales un profesionista de esta área puede conseguir un mejor salario. Para contestar las preguntas se implementó un modelo estadístico: la prueba de hipótesis. Esta se utilizó para realizar pruebas de normalidad de las muestras, pruebas de varianzas iguales, y, finalmente, pruebas de diferencias de medias. Se encontró que no hay diferencia significativa entre el salario promedio de un Ingeniero en Datos y un Científico de Datos con el mismo nivel de experiencia; además, se encontró que el salario promedio de un empleado intermedio en una empresa Estadounidense es mayor al de un empleado principiante; finalmente, se encontró que el salario promedio ofrecido por una empresa mediana en Estados Unidos es mayor que en Reino Unido.**

***Index Terms*—Salarios, Sueldos, Ciencia de Datos, Análisis de datos, Prueba de hipótesis**

I. Introducción

La Ciencia de Datos sigue siendo un campo relativamente nuevo que sin duda alguna ha crecido en las últimas décadas. Este vertiginoso crecimiento provoca preguntas interesantes sobre el mercado laboral de áreas emergentes como esta. A partir de una base de datos de Kaggle, es posible elucidar sobre posibles factores que influyen en un mayor o menor salario para un analista de datos o semejante. El presente trabajo establece las siguientes preguntas de investigación:

1. ¿El salario promedio en Estados Unidos de un Ingeniero en Datos es mayor al salario promedio de un Científico de Datos con el mismo nivel de experiencia?
2. ¿Hay diferencia significativa en el salario promedio en función de la experiencia en una empresa Estadounidense? En particular, ¿el salario promedio de un empleado intermedio es mayor que el salario promedio de un empleado principiante?
3. ¿El salario promedio ofrecido en una empresa mediana en Estados Unidos es mayor al que ofrece una empresa mediana en Reino Unido?

Cada pregunta pretende investigar tres diferentes factores puntuales de las distintas características del conjunto: el puesto (en este caso Científico de Datos y Ingeniero en Datos), el nivel de experiencia (intermedio y principiante), y ubicación de la empresa (Estados Unidos y Reino Unido). Además se realiza un intento de homogeneizar las muestras, es decir un c̈ontrolr̈udimentario: utilizar un mismo nivel de experiencia para comparar los puestos, usar sólo empresas Estadounidenses para comparar los niveles de experiencias, y usar sólo empresas medianas para comparar la ubicación de las empresas.

El objeto de investigación del presente reporte y similares es de suma importancia ya que da una idea sobre los potenciales cursos de acción de los futuros profesionistas de esta área para maximizar su desarrollo profesional y económico; además, este tipo de investigación puede ser de utilidad para diagnosticar las tendencias del mercado en el área de tecnología, y potencialmente orientar a inversionistas y emprendedores de las áreas de oportunidad en esta industria (y quizá incluso en otras).

II. Resultados y análisis

II-A. Salario en función del puesto

Para abordar la primera pregunta de investigación,

1. ¿El salario promedio en Estados Unidos de un Ingeniero en Datos es mayor al salario promedio de un Científico de Datos con el mismo nivel de experiencia?

se desarrollarán solo un caso de nivel de experiencia: el nivel nivel experto (Senior, SE). Dado que la pregunta implica la comparación de dos medias de muestras diferentes, es pertinente realizar el análisis con una prueba de hipótesis de diferencia de medias.

Se aborda entonces la comparación del salario promedio entre un Científico de Datos y un Ingeniero en Datos para un mismo nivel de experiencia, el nivel experto. Una vez identificada la herramienta estadística a utilizar, se plantea la prueba de hipótesis a probar,

$$H_0 : \ \mu_1 = \mu_2, \tag{1}$$

$$H_a : \ \mu_1 > \mu_2, \tag{2}$$

donde μ_1 es el salario promedio poblacional de un Ingeniero en Datos experto, y μ_2 es el salario promedio poblacional de un Científico de Datos experto. Todas las pruebas de hipótesis del presente reporte utilizarán una regla de decisión de un nivel de significancia de $\alpha = 0{,}05$.

No obstante, antes de realizar dicha prueba de hipótesis se debe realizar una prueba de bondad de ajuste a una distribución normal con la finalidad de implementar la prueba de hipótesis

Cristofer Becerra Sánchez, pertence al Tecnológico de Monterrey Campus Monterrey, Monterrey, Nuevo León, Mexico, C.P. 64849

de las medias con la distribución correcta. Las hipótesis de tal prueba de bondad de ajuste son de la forma

$$H_0 : f(x) = f_n(x) \tag{3}$$
$$H_a : f(x) \neq f_n(x) \tag{4}$$

donde $f(x)$ es la distribución de los salarios de cada puesto, y $f_n(x)$ es una distribución normal. Se implementará una prueba de normalidad de Shapiro-Wilk para todas las pruebas de bondad de ajuste posteriores. Comenzando con la muestra de los salarios de los Científicos de Datos expertos, se obtuvo un estadístico $W = 0{,}9825$, con su respectivo p-value de $p = 0{,}5407$. La prueba retorna un p-value mayor que el nivel de significancia establecido, por lo cual no es posible rechazar H_0. Por lo tanto, se dice que los salarios de los Científicos de Datos siguen una distribución normal. La figura 1 contiene un QQ-plot y un histograma que cimentan mejor las conclusiones de la prueba.

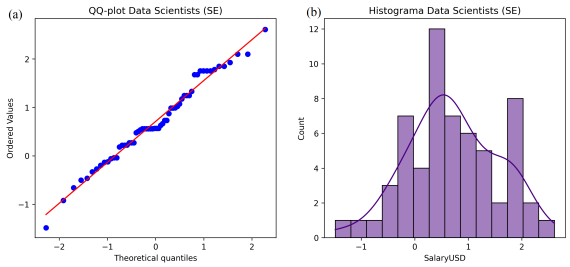

Figura 1. (a) QQ-plot de la muestra de los Científicos de Datos; el eje x representa los valores de los cuantiles teóricos que debería seguir una distribución normal, mientras que el eje y representa los cuantiles observados a partir de a muestra. (b) Histograma de la muestra de Científicos de Datos en Estados unidos; se grafica la curva de densidad sobre las barras que representan la frecuencia de los datos en el intervalo.

Se realiza lo mismo para la muestra de los Ingenieros en Datos expertos, en donde se obtuvo un estadístico de prueba $W = 0{,}9888$, con un p-value $p = 0{,}8384$, por lo que puede decirse lo mismo que la muestra anterior: es imposible rechazar H_0 y por lo tanto los salarios de los Ingenieros en Datos expertos se distribuyen normalmente. De igual manera, el respectivo QQ-plot y el histograma se despliegan en la figura 2.

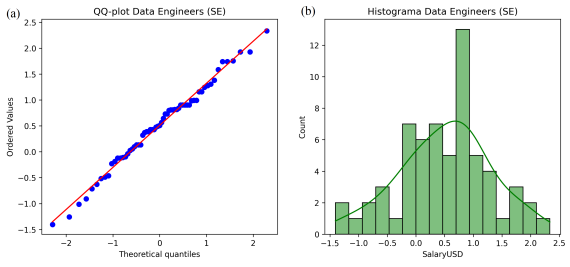

Figura 2. (a) QQ-plot de la muestra de los Ingenieros en Datos que trabajan en empresas de Estados Unidos. (b) Histograma de la misma muestra de Ingenieros en Datos; también se grafica la curva de densidad sobre las barras que, de igual manera, representan la frecuencia de los datos en el intervalo dado.

Ya que se determinó mediante dos pruebas de Shapiro-Wilk que ambas distribuciones muestrales siguen aquella a la de una normal, se debe comprobar otra consideración importante: la homogeneidad de varianzas. Para esto se utiliza una prueba de Levene, en la cual la hipótesis nula establece que las muestras provienen de poblaciones cuyas varianzas son idénticas. Dicha prueba retornó un estadístico $W = 0{,}038$, el cual implica un p-value de $p = 0{,}8458$, que es sustancialmente mayor que el nivel de significancia $\alpha = 0{,}05$. De esto se aduce que no se rechaza H_0, y por tanto puede afirmarse con un nivel de significancia del 95 % que las varianzas poblacionales son iguales.

Se continúa ahora con la prueba de hipótesis de diferencia de medias con una distribución Z. A pesar de que se desconocen las desviaciones estándar poblacionales de ambas muestras, esta decisión se fundamenta en que el tamaño de muestra es $n > 30$ en los dos casos, por lo cual la desviación estándar muestral representa una excelente aproximación a la desviación estándar poblacional.

La función utilizada entregó un estadístico $z_0 = 1{,}3483$, con el correspondiente p-value de $p = 0{,}1776$. Del estadístico de prueba z_0 y su respectivo p-value se aduce que no se rechaza la hipótesis nula, H_0. No hay evidencia suficiente para afirmar que el sueldo promedio de un Ingeniero en Datos en Estados Unidos es mayor al de un Científico de Datos (sin controlar para otros posibles factores). Se puede obtener una intuición al graficar sus distribuciones sobrepuestas mediante un histograma y un boxplot, tal como se observa en la figura 3.

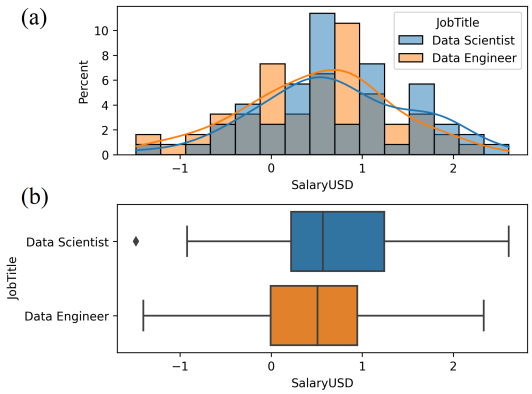

Figura 3. (a) Histograma de las distribuciones de ambas muestras; el eje y está normalizado tal que la suma de la altura de todas las barras es igual a 100, es decir, un porcentaje; se grafica la estimación de la densidad de la distribución a partir de la altura de las barras. (b) Diagrama de caja y bigotes de ambas muestras; se tiene un solo dato atípico para la muestra de los Científicos de Datos; se observa que las distribuciones abarcan casi los mismos valores pero los bigotes de la muestra de los Científicos de Datos está recorrida ligeramente a la derecha.

Al observar la figura se vuelve evidente la conclusión, ya que los datos se concentran prácticamente en la misma región, con algunas desviaciones insignificantes. A pesar de que la distribución de los Científicos de Datos tiene una deformación importante en la cola derecha, no se afecta la conclusión ya que es probable que esta misma recorra la media de la muestra más cerca a la media de los Ingenieros en Datos.

II-B. Salario en función de la experiencia

Continuando con la segunda pregunta de investigación,

2. ¿Hay diferencia significativa en el salario promedio en función de la experiencia en una empresa Estadounidense? En particular, ¿el salario promedio de un empleado intermedio es mayor que el salario promedio de un empleado principiante?

se nota inmediatamente que también se trata de la comparación de dos medias muestrales, por lo cual el modelo estadístico de una prueba de hipótesis de diferencia de medias también resulta adecuado. Por tal razón la prueba de hipótesis en este caso toma la misma forma,

$$H_0 : \ \mu_1 = \mu_2, \tag{5}$$
$$H_a : \ \mu_1 > \mu_2, \tag{6}$$

sin embargo, ahora μ_1 representa el salario promedio de un empleado de nivel intermedio, y μ_2 es el salario promedio de un empleado de nivel principiante. De nueva cuenta es necesario implementar una prueba de bondad de ajuste de Shapiro-Wilk para proseguir adecuadamente con la prueba de hipótesis.

Se obtuvo un estadístico de prueba $W_0 = 0{,}8474$, con un p-value de $p = 0{,}0004$. Se obtiene un p-value mucho menor que α por lo cual se debe rechazar H_0. Por lo tanto, se dice con un nivel de significancia del 95 % que los salarios empleados principiantes de empresas Estadounidenses no se distribuyen como una normal. En la figura 4 se puede apreciar el QQ-plot y la distribución de la muestra mediante un histograma.

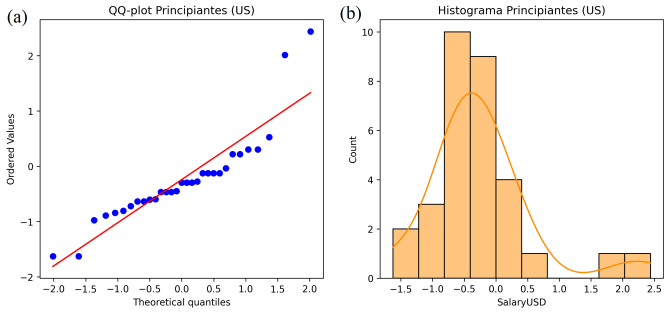

Figura 4. (a) QQ-plot de la muestra de empleados principiantes en empresas de Estados Unidos; nótese como la cola izquierda distorsiona la pendiente de tal manera que la tendencia del resto de la distribución (que en otras condiciones tendría un buen ajuste) parece tener una dirección diferente; además se tienen los dos valores atípicos de la cola derecha. (b) Se resalta la misma distribución del QQ-plot pero en forma de histograma; se logra apreciar una buena forma de campana de la distribución principal, con excepción de las colas mencionadas anteriormente.

Se obstiene un ajuste muy poco razonable a la recta de cuantiles teóricos, debido a las colas. La cola derecha se nota particularmente en el histograma, donde se tienen dos valores atípicos alrededor de las 2 desviaciones estándar de la media. Se realiza lo mismo para los empleados de nivel intermedio, y se obtuvo un estadístico $W_0 = 0{,}9844$, cuyo p-value es de $p = 0{,}3808$. Se tiene el caso opuesto al anterior: el p-value es mayor que el nivel de significancia, por lo cual no se rechaza la hipótesis nula y se dice que esta muestra sí se distribuye como una distribución normal. La gráfica complementaria a la prueba de Shapiro-Wilk, es decir, el QQ-plot y el histograma, se encuentran en la figura 5.

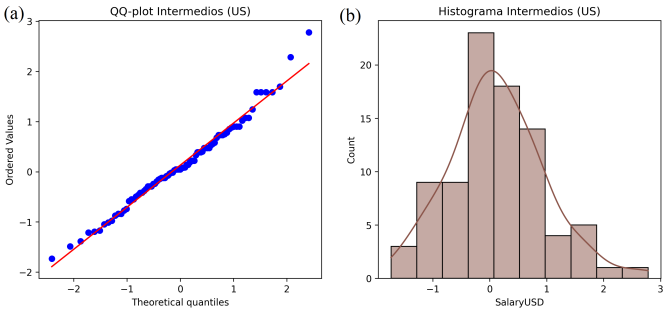

Figura 5. (a) QQ-plot de los empleados de nivel intermedio de empresas en Estados Unidos; se distingue un buen ajuste a la línea recta, con la excepción de la cola derecha. (b) Histograma de la misma muestra que en (a); se nota un ligero sesgo a la derecha, sin embargo, la forma es parecida a la de una distribución normal.

También se realiza la prueba de hipótesis de Levene para comprobar la igualdad de varianzas en este caso. La función retorna un p-value de $p = 0{,}4850$ para un estadístico de prueba $W_0 = 0{,}7428$, y por tal razón se puede decir que las distribuciones provienen de poblaciones con varianzas idénticas con un nivel de confianza del 95 %. Se procede entonces con la prueba de hipótesis de diferencia de medias.

Con el precedente de que una de las distribuciones muestrales no se ajusta a la de una normal, se opta por utilizar una prueba de hipótesis de diferencia de medias con una distribución t de Student para varianzas iguales. De tal suerte que la prueba regresa un estadístico de prueba $t_0 = 2{,}1424$ que corresponde a un p-value $p = 0{,}0171$. El p-value es menor que α, por lo cual, debe rechazarse la hipótesis nula, H_0; hay evidencia suficiente para afirmar que el salario promedio de un empleado intermedio de una empresa Estadounidense es mayor al de un empleado principiante con un nivel de significancia del 95 %. Se grafican ambas distribuciones para visualizar dicho resultado en la figura 6.

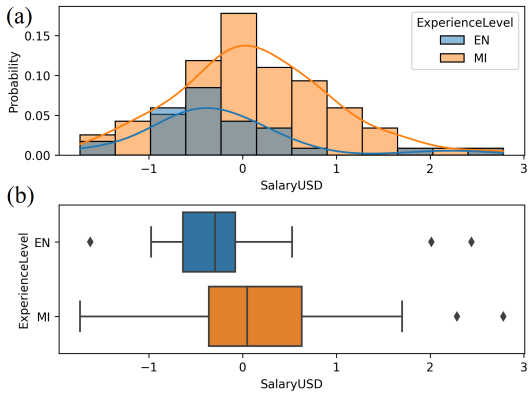

Figura 6. (a) Histograma de las distribuciones de ambas muestras: los principiantes (EN) y los intermedios (MI); el eje y está normalizado tal que representa la densidad de probabilidad, es decir, que el área total del histograma es igual a 1. (b) Diagrama de caja y bigotes de las muestras de los salarios de empleados de nivel principiante (EN) y los salarios de empleados de nivel intermedio (MI); se osberva claramente el menor rango intercuartílico de los principiantes, y el corrimiento hacia la derecha de la mediana de los intermedios con respecto a los principiantes.

En este caso las distribuciones también están relativamente traslapadas, pero es más evidente que la media de los

principiantes (EN) es considerablemente menor que la de los intermedios (MI); a pesar de que los bigotes del diagrama de caja de los intermedios abarca un mayor intervalo que el de los principiantes, es evidente que tanto la media como la mediana son mayores que en el caso de los principiantes.

II-C. Salario en función de la ubicación de la empresa

Finalmente se aborda la tercera y última pregunta de investigación,

3. ¿El salario promedio ofrecido en una empresa mediana en Estados Unidos es mayor al que ofrece una empresa mediana en Reino Unido?

De igual suerte que en los dos casos anteriores, es conveniente implementar una prueba de hipótesis de diferencia de medias ya que se está contrastando la media de los salarios ofrecidos por empresas medianas en Estados Unidos con la media de salarios de empresas del mismo tamaño en Reino Unido. Entonces, las hipótesis se enuncian simbólicamente como,

$$H_0 : \ \mu_1 = \mu_2, \tag{7}$$
$$H_a : \ \mu_1 > \mu_2, \tag{8}$$

donde μ_1 es el salario promedio que ofrece una empresa mediana Estadounidense, y μ_2 es el salario promedio que ofrece una empresa mediana Británica.

Una vez más, se realiza la prueba de normalidad de Shapiro-Wilk para determinar la distribución a utilizar en la prueba de hipótesis. Se comienza con el subconjunto de empresas medianas de Estados Unidos. La prueba indica un estadístico de prueba $W_0 = 0{,}9896$, con un p-value de $p = 0{,}1207$, que es mayor al nivel de significancia establecido, α. Dicha razón obliga a aceptar la hipótesis nula con un nivel de significancia del 95 %, es decir, que esta muestra sigue una distribución normal. El QQ-plot y el histograma de la figura 7 apoyan la conclusión.

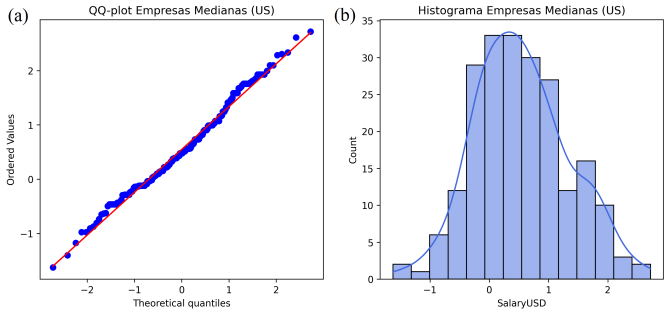

Figura 7. (a) QQ-plot de la muestra de salarios de empleados de empresas medianas en Estados Unidos; el ajuste es considerablemente bueno para los cuantiles centrales, pero comienza a adoptar la forma de una especie de w alrededor de las colas. (b) Histograma de la distribución de los salarios de empleados de empresas medianas en Estados Unidos.

Se observa un buen ajuste la recta teórica, aunque es notable una especie de oscilación alrededor de la línea recta, tanto por encima como por debajo; esto parece indicar regiones donde se acumulan los datos, perdiendo ligeramente la forma de una normal. Esta forma se observa en el histograma, que tiene un pico relativamente ancho, y la cola derecha tiene una acumulación importante de datos.

Lo siguiente es la prueba del subconjunto británico. Se obtiene un estadístico $W_0 = 0{,}9182$, con un p-value con valor de $p = 0{,}0211$; con un nivel de significancia de $\alpha = 0{,}05$ se sigue necesariamente el rechazo de la hipótesis nula; por lo tanto, se dice que el subconjunto de empresas medianas de Reino Unido no sigue una distribución normal. De igual manera en la figura 8 se aprecia el QQ-plot acompañado por un histograma que apoyan la conclusión a la que se llegó.

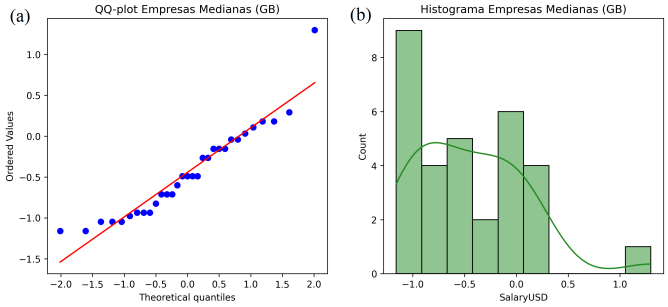

Figura 8. (a) QQ-plot de la muestra de salarios de empleados de empresas medianas en Reino Unido. (b) Histograma de la distribución de los salarios de empleados de empresas medianas en Reino Unido.

En este caso se tiene una forma del histograma con un sesgo importante, tanto, que difícilmente parece una distribución conocida; este sesgo repercute también en las colas, por lo cual los cuantiles observados están muy lejos de cuantiles teóricos, afectando el ajuste.

De manera reiterada, se debe poner a prueba si las varianzas iguales; la intención de tal procedimiento es utilizar los argumentos y funciones indicadas para la prueba de hipótesis de dos muestras. Esto se realiza con la prueba de Levene. Esta prueba retornó un valor de estadístico $W = 4{,}5367$, con el correspondiente p-value de $p = 0{,}0342$, quien es menor que el nivel de significancia α. Por tanto, debe rechazarse H_0 en este caso. Dicho resultado implica que las muestras no provienen de poblaciones con varianzas idénticas con un nivel de confianza del 95 %. En consecuencia, la prueba de hipótesis de diferencia de medias se realizará con una distribución t de Student para varianzas diferentes.

La prueba resulta en un estadístico $t_0 = 8{,}8209$ muy grande y, de forma consecuente, un p-value extremadamente pequeño, $p \to 0$ (Python lo imprime como 0,0, lo cual indica que es tan cercano a cero como la precisión de la máquina lo permite); por esta razón se rechaza H_0. En consecuencia, puede afirmarse con un nivel de significancia del 95 % que el salario que ofrece una empresa mediana de Estados Unidos, en promedio, es mayor al que ofrece una empresa de Reino Unido, en promedio. Dicha conclusión se aprecia en la visualización del histograma y el diagrama de caja (ver figura 9).

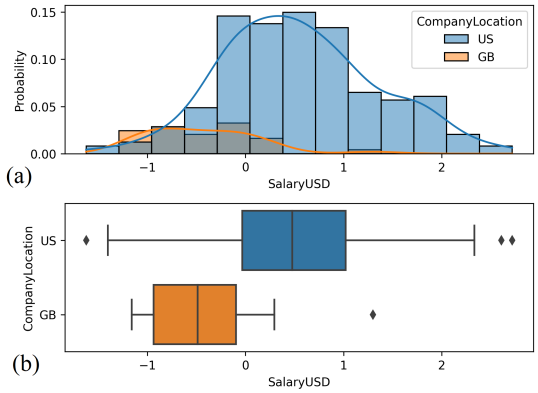

Figura 9. (a) Histograma de las distribuciones de ambas muestras: los salarios de empresas medianas en Estados Unidos (US) y los salarios de empresas medianas en Reino Unido (GB); el eje y está normalizado tal que representa la densidad de probabilidad, es decir, que el área total del histograma es igual a 1. (b) Diagrama de caja y bigotes de las muestras de los salarios de empresas medianas en Estados Unidos y los salarios de empresas medianas en Reino Unido.

Al observar la figura parece obvia la conclusión: la frecuencia de salarios mayores de empresas estadounidenses es aplastante en comparación a los de Reino Unido; se tiene una muestra más grande, y además es evidente la gran diferencia entre medias: la media de los registros de salarios de empresas medianas en Estados Unidos presenta un corrimiento importante a la derecha en comparación a la media de Reino Unido.

III. Conclusión

El análisis anterior permite hacer varias conclusiones. Primero, puede afirmarse que no hay diferencia significativa entre el salario promedio de un Data Engineer y un Data Scientist para el mismo nivel de experiencia: nivel intermedio. No obstante, esta conclusión lleva en realidad varios asteriscos; debido a que no se realizó un ANOVA con las variables categóricas (puesto, nivel de experiencia, año, ubicación de la empresa, etc.) es imposible realizar esta misma afirmación con seguridad si se desagrega para el resto de los factores. Para investigaciones posteriores sería pertinente realizar este análisis para determinar con certeza los factores principales, y si, en efecto, la presente conclusión se mantiene sólida.

Se encontró también que el salario promedio de un empleado intermedio en una empresa Estadounidense es mayor al de un empleado principiante. Los mismos asteriscos de la conclusión anterior aplican aquí.

Finalmente, se encontró que el salario promedio ofrecido por una empresa mediana en Estados Unidos es mayor que el de una empresa mediana en Reino Unido. Vale la pena tomar las mismas precauciones que en los dos casos anteriores.

Anexo

Enlace al repositorio de GitHub, *Data-Science-Salaries*, con todos los archivos del proyecto (Jupyter Notebook, Notebook en formato .py, la base de datos utilizada, y el presente documento PDF):

https://github.com/crisb-7/Data-Science-Salaries