Tecnológico de Monterrey

Actividad:

Reporte Final: Data Science Salaries Expectation

Módulo:

Módulo 1: Estadística e Inteligencia artificial avanzada para la ciencia de datos

Grupo:

TC3006C.101

Nombre:

Franco Quintanilla Fuentes - A00826953

Maestra:

Blanca R. Ruiz Hernández

Fecha:

14 de septiembre de 2022

Resumen

Como futuros Data Scientists debemos poder analizar cualquier tipo de datos, en especial si esos datos nos ayudan a poder determinar nuestros posibles salarios en un futuro en base a un estudio y un análisis estadístico. Dicho esto, identificamos las variables que hacen la diferencia tener un mejor salario que alguien más, aunque nuestro estudio tiene algunas limitaciones, ya que nos centramos en la cantidad de dinero en dólares, entre otras simplificaciones hechas. Uno de los resultados más importantes que obtuvimos fue que el rango de salarios de un Data Scientist puede oscilar entre los $116,315 y $128,059 dólares, en el año 2022, con un intervalo de confianza del 95%.

Introducción

Para nuestra investigación, fueron 3 las preguntas detonantes que optamos por utilizar en nuestro estudio.

1. ¿Cuál es el salario al que puede aspirar un analista de datos?
2. ¿Se han incrementado los salarios a lo largo del tiempo?
3. ¿Influye el tamaño de la compañía en el salario que ofrece en 2022?

Para poder contestar nuestras preguntas en base a los datos de Kaggle, tuvimos que realizar todo un análisis estadístico en donde nuestro objetivo fue poder responder esas preguntas basándonos en un análisis estadístico de trasfondo.

Análisis de resultados

Para hacer el análisis de los resultados, lo primero que tuvimos que hacer fue entender los datos que tenemos, que después de analizar e investigar un poco pudimos darnos cuenta que sería mucho más fácil enfocarnos en 4 variables principales: Titulo de empleo (job_title), Salario en dólares (salary_in_usd), Tamaño de la compañía (company_size), y Año Laboral (work_year).

Primero, de la base de datos original tuvimos que observar el comportamiento de esos datos, en donde nos dimos cuenta de lo sucios que estaban al visualizarlos.

Titulo de empleo

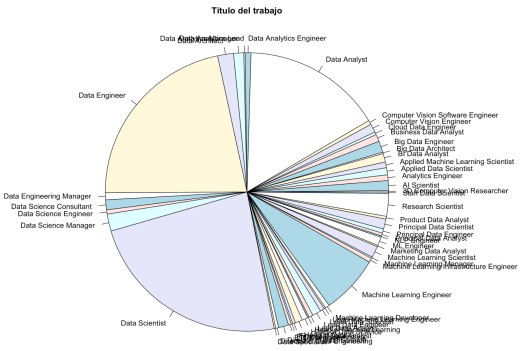

Figura 1: Pie Chart de la cantidad de “Títulos de empleo”.

Como podemos ver, son demasiados títulos, que ni el pie chart se ve bien, este fue un factor que después consideramos quitar para hacer nuestro análisis definitivo.

Salario en dólares

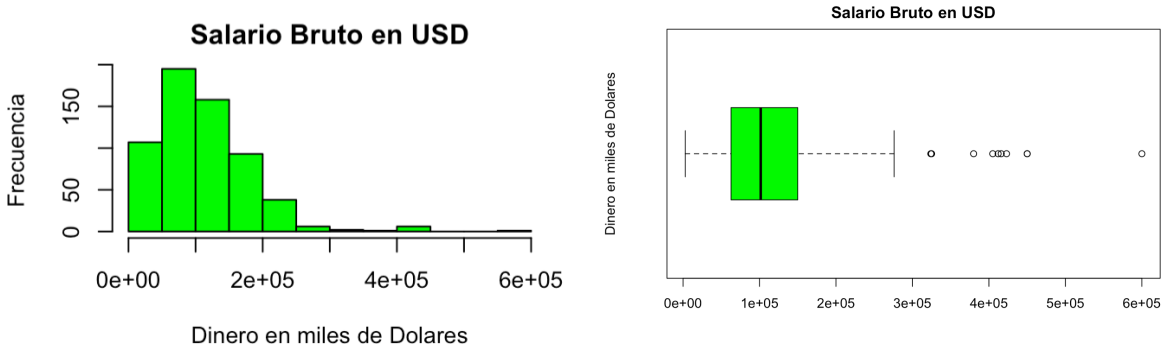

Figura 2: Distribución de los valores de Salario en USD, y Box plot del mismo.

Como podemos ver, hay demasiados valores atípicos en nuestros datos, por lo que también nuestra distribución se va a comportar de esa manera tan sesgada, por lo que estos son los datos que vamos a limpiar en base a los cuartiles.

Tamaño de la compañía

Para el tamaño de la empresa usamos un barplot, ya que solo tenemos 3 variables categóricas, Chica, Mediana y Grande.

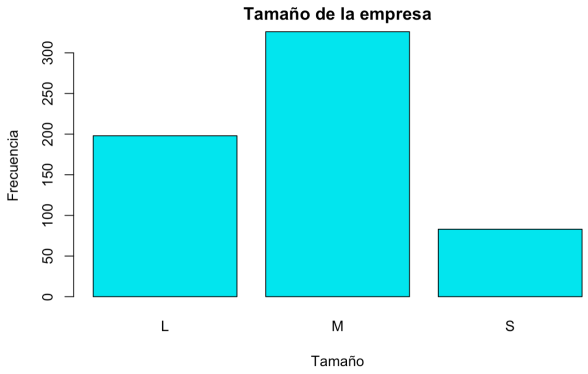

Figura 3: Bar plot del conteo de personas en base al tamaño de la empresa.

En esta figura podemos ver que la mayoría de los trabajadores se encuentran laborando en una empresa de tamaño mediano.

Año Laboral

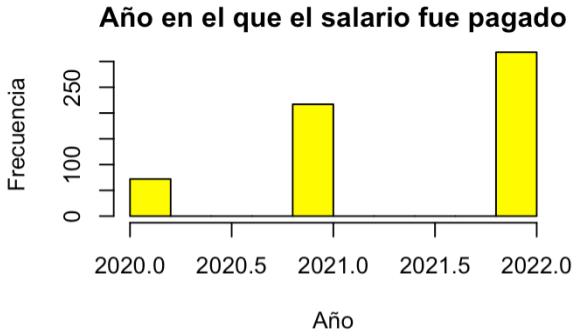

Figura 4: Visualización de distribución de año laboral.

Para el caso de año laboral, podemos ver que la mayoría de las personas en este estudio se encuentran en el año 2022.

Después de analizar estos datos, nos damos cuenta de que tenemos que limpiar nuestra variable de Salario en USD para poder visualizar mejor su comportamiento y su distribución. Para esto, utilizamos la limpieza de datos en base a los cuartiles, el cual es un método muy comúnmente usado en estadística y data science.

Después de limpiar los datos de Salary in USD, vamos a visualizar si quedaron algunos datos atípicos, y su nuevo comportamiento con el histograma y su frecuencia.

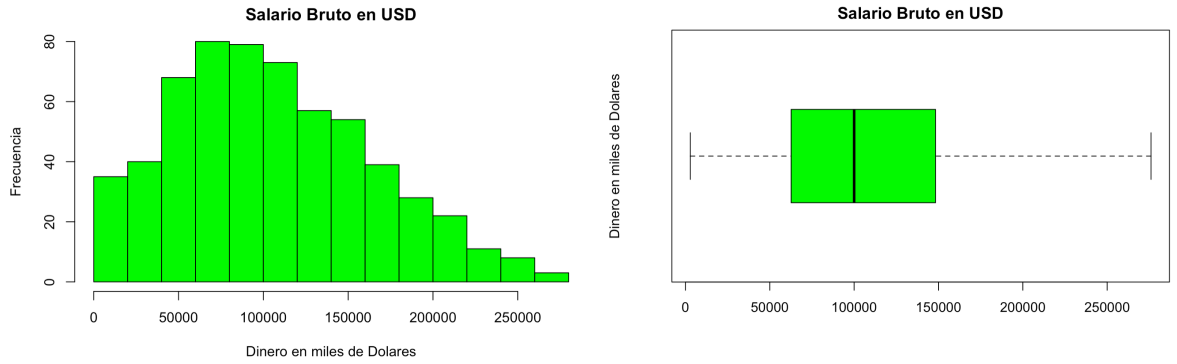

Figura 5: Distribución y Box plot de Salario en USD limpia.

Como podemos ver, cambia muchísimo el comportamiento de la distribución y la frecuencia de los valores, una vez la variable se encuentra libre de valores atípicos.

Con nuestra variable de salario limpia, optamos por hacer una visualización más avanzada del salario en USD, por cada Título Del Empleo, y que a su vez la separe por el tamaño de la empresa. Para lograr esto, tuvimos que dividir los datos en cada uno de sus respectivos años (2020, 2021, 2022), en donde sus gráficos son los siguientes.

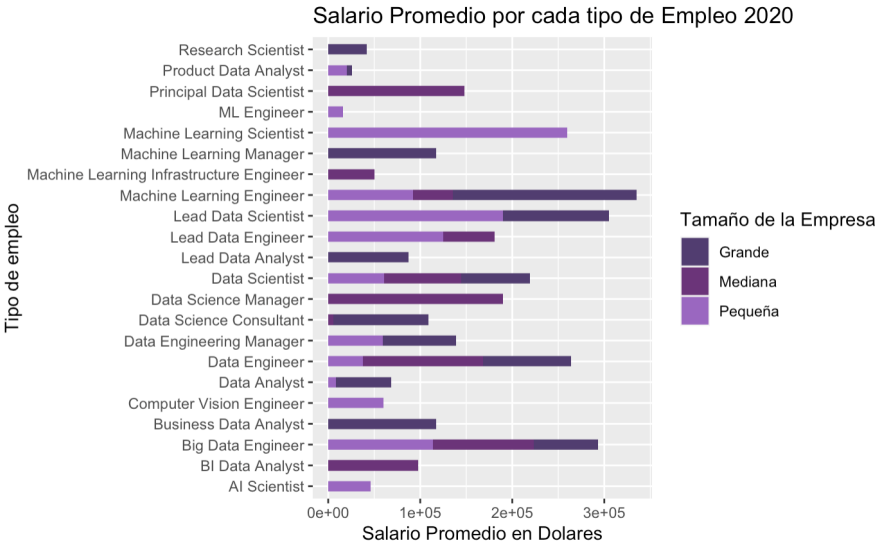

Figura 6: Salario Promedio por cada tipo de Empleo 2020.

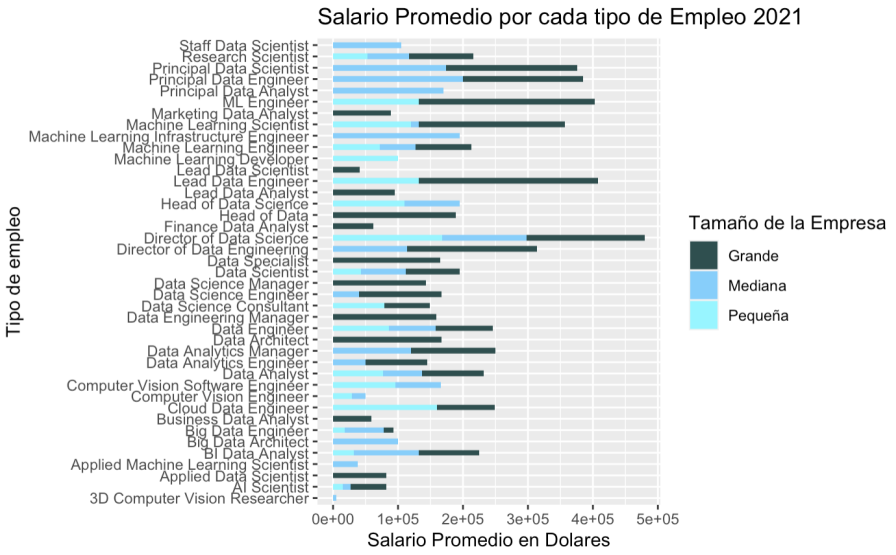

Figura 7: Salario Promedio por cada tipo de Empleo 2021.

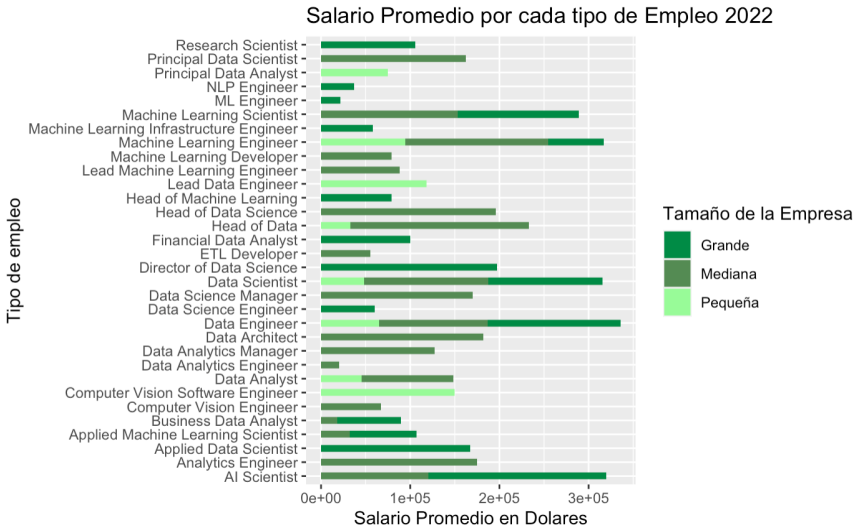

Figura 8: Salario Promedio por cada tipo de Empleo 2022.

Con estas gráficas, nosotros podemos ver que del 2020 al 2022 se han tanto creado como eliminado distintos tipos de empleos, lo cual nos ayuda a tomar una decisión más informada sobre la expectativa de salario como un científico de datos.

A partir de este resultado, vamos a tomar la decisión de irnos con los datos más nuevos, es decir, con los datos del 2022, esto más que nada por la creación y eliminación de nuevos y viejos empleos, el cual personalmente, es un factor que influye en la toma de decisiones.

Con esto, podemos sacar los valores como la media y la desviación estándar de nuestros nuevos datos enfocados en el año 2022. Calculando estos resultados, obtuvimos que:

- La media de 'Salary in USD' en el 2022 es de: 122,187.3
- La desviación estándar de 'Salary in USD' en el 2022 es de: 53,170.39

Pero como estos valores son crudos, necesitamos una visualización para comprenderlos mejor. Por lo que vamos a graficar su histograma para visualizar su distribución.

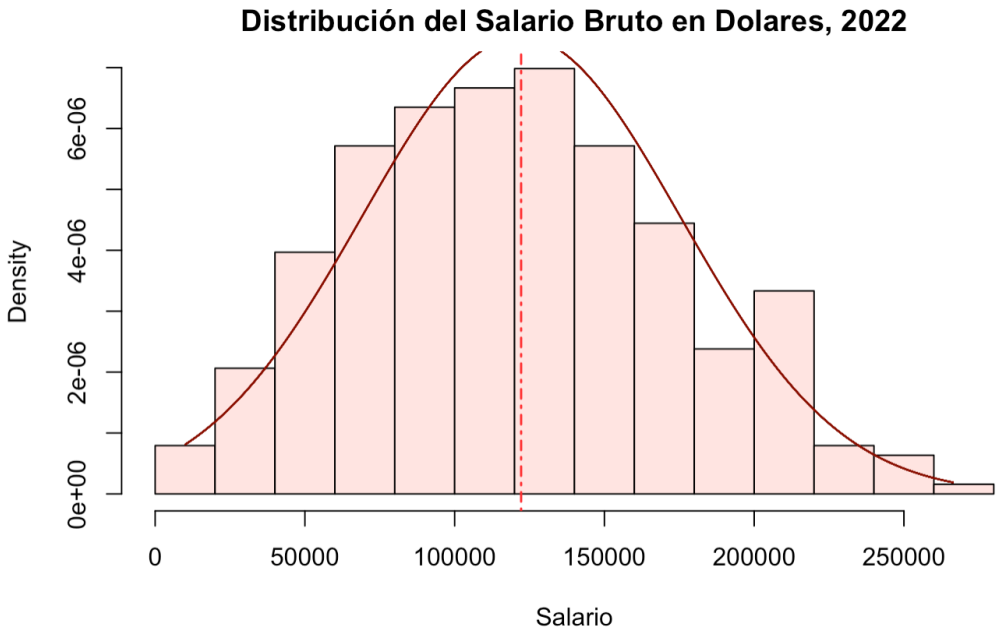

Figura 9: Distribución del Salario Bruto en Dólares en el año 2022.

Vemos que los datos siguen la tendencia de una distribución normal, aunque con la ayuda de QQ plot podremos inferir mejor por que no se adapta de la mejor manera, además de el histograma, observamos que la línea punteada es el valor de la media y la línea roja oscura es la representación de la distribución de los datos del salario en USD en el 2022.

Dicho esto, nuestro siguiente paso es graficar el QQ-Plot.

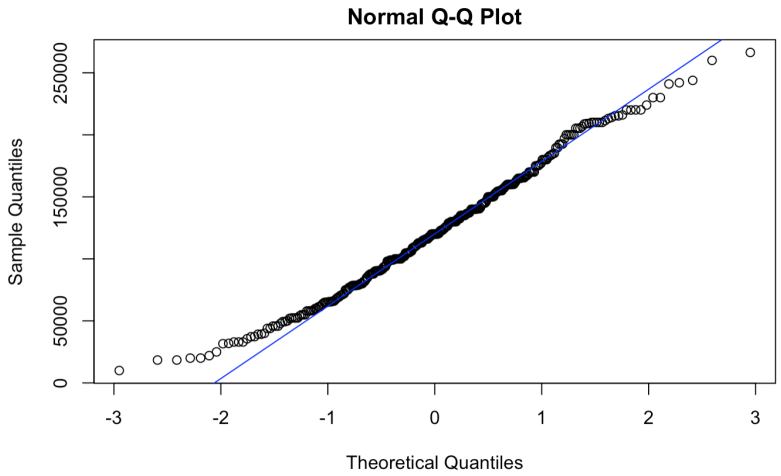

Figura 10: Prueba de normalidad con QQ plot.

Gracias a la gráfica de Q-Q Plot podemos observar que nuestros datos del Salario en USD en el 2022 se comporta como una distribución con colas delgadas es decir, alta curtosis y una distribución leptocúrtica.

Hecho lo anterior, ahora vamos a sacar los intervalos de confianza de nuestro modelo, esto con ayuda del z.test. Los resultados que obtuvimos fueron que en Lower fue de $116,315 y que en Upper fue de $128,059. Lo que quiere decir que el rango de salario de un "científico de datos" ronda entre los ***$116,315 y $128,059 dólares***, con un intervalo de confianza del 95%.

Con este análisis inicial de los intervalos de confianza, podemos agarrar un poquito más de rango y aplicarlo en distintos enfoques de las preguntas base de nuestro análisis, por lo que nos enfocaremos en:

- Salario en los distintos tamaños de empresa en el 2022

	Grande	***Mediana***	***Pequeña***
Media de los salarios en los distintos tamaños de empresa	$119,249.1	$124,962.9	$77,046.54
Desviación estándar de los salarios en los distintos tamaños de empresa	$59,162.4	$51,835.73	$38,510.28

En donde los intervalos de confianza para el salario en los distintos tamaños de empresa en el 2022 nos queda de la siguiente manera, visto ya gráficamente.

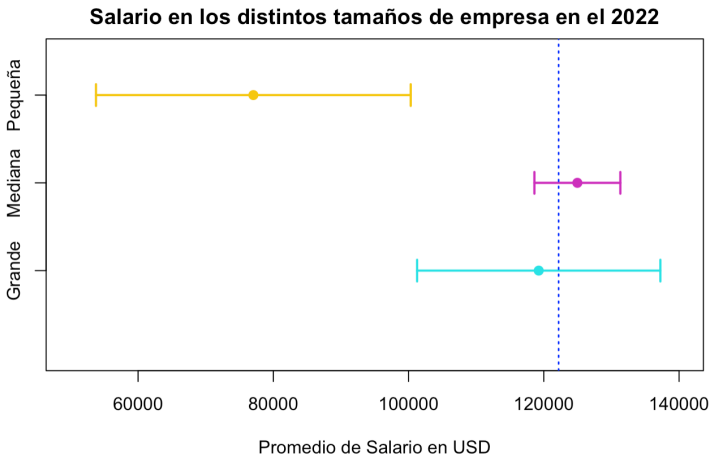

Figura 11: Intervalos de confianza del salario en los distintos tamaños de empresa en el 2022.

- Salario en los distintos años laborales

	2020	***2021***	***2022***
Media de los salarios en los distintos años laborales	$82,775.88	$92,860.44	$122,187.3
Desviación estándar de los salarios en los distintos años laborales	$53,887.35	$61,531.28	$53,170.39

En donde los intervalos de confianza para el salario en los distintos años laborales nos queda de la siguiente manera, visto ya gráficamente.

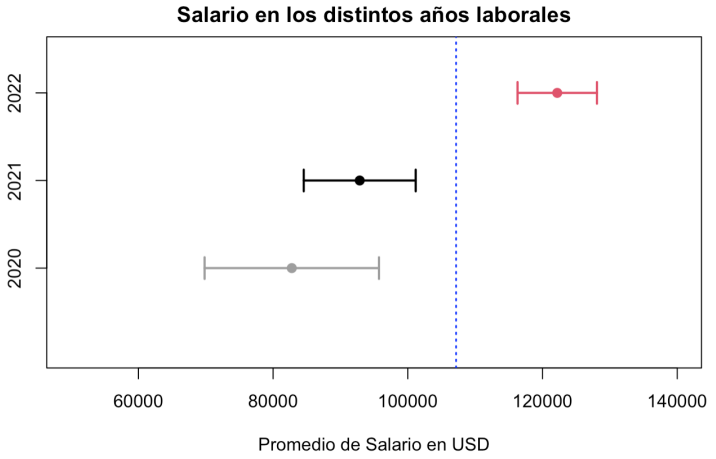

Figura 11: Intervalos de confianza del salario en los distintos años laborales.

Conclusión

Gracias a todo este análisis estadístico, podemos responder nuestras preguntas base.

1. El rango de salarios de un científico de datos puede oscilar entre los $116,315 y $128,059 dólares, en el año 2022, en base a nuestras estadísticas con un intervalo de confianza del 95%, además se puede observar visualmente en la gráfica de intervalos de confianza de "Salario en los distintos años laborales".
2. En el gráfico de los intervalos de confianza de "Salario en los distintos años laborales", podemos observar que efectivamente los salarios han incrementado a lo largo del tiempo, ya que en el año 2022, podemos observar que su intervalo de confianza es mayor a la media poblacional de los salarios en dólares de los años anteriores, donde además se ve que está alejado de los demás intervalos de confianza.
3. En base al análisis podemos afirmar que el tamaño de una empresa pequeña en 2022 si influye en el salario que ofrece, no obstante, cuando hablamos de una empresa mediana y una empresa grande, podemos afirmar con el 95% de confianza que no varía, y podríamos decir que son la misma, por los intervalos de confianza.

Anexos

Repositorio de Github

https://github.com/francoquintanilla0/Data-Science-Salaries