MEMORIA

Análisis Exploratorio de Datos (EDA) sobre

Salarios en el campo Data Science

**Realización de un análisis exploratorio de datos sobre salarios dentro del mundo de los datos.**

**Se va realizar utilizando el lenguaje de programación Python y diferentes librerías como numpy, pandas y matplotlib, entre otras.**

**Los datos provienen de la plataforma Kaggle, e incluyen información relevante para el estudio que llevaremos a cabo.**

Objetivo:

**Conocer la distribución de salarios en el campo del data para la toma de decisiones futuras.**

Hipótesis planteadas:

**1. Análisis variable principal-Salary**

**2. ¿Qué relación existe entre el salario y el nivel de estudios dentro del campo de Data?**

**3. ¿Cual es la proporción de hombres y mujeres en cada tipo de empleo?**

**4. ¿Qué relación tiene los años de experiencia y el salario?**

**5. ¿Dentro del campo de data, hay diferencias entre los salarios de hombre y mujeres?**

**6. Dentro del campo de data, ¿quién tiene mayor salario de entrada?**

Librerías utilizadas:

* **numpy**
* **pandas**
* **matplotlib.pyplot**
* **seaborn**
* **scipy.stats**
* **plotly.express**

Análisis exploratorio:

**El análisis ha seguido la siguiente estructura:**

1. **Planteamiento de hipótesis.**
2. **Carga de datos .**
3. **Exploración inicial.**
4. **Limpieza de datos.**
5. **Análisis univariante de la variable principal - Salary.**
6. **Estudio variables y comprobación de hipótesis.**

### Exploración inicial:

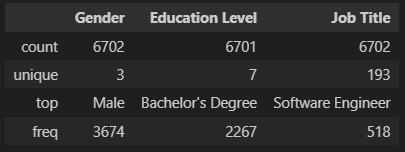
**Descripción de los datos:**

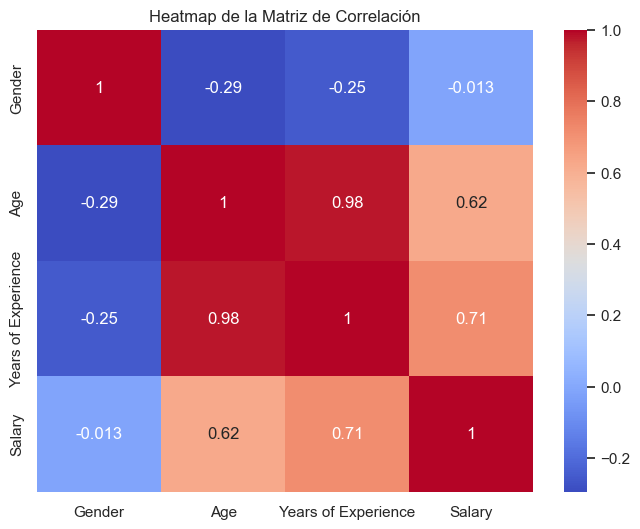
**Se dispone de una base de datos compuesta por 6 columnas y 6704 filas, con datos tanto numéricos como cuantitativos.**

| **NOMBRE COLUMNAS** | **DESCRIPCIÓN** | **TIPO** |
| --- | --- | --- |
| **Age** | **Edad del usuario.** | **Cuantitativa discreta** |
| **Gender** | **Género del usuario.** | **Cualitativa nominal** |
| **Education Level** | **Nivel máximo de estudios finalizados por el usuario.** | **Cualitativa ordinal** |
| **Job Title** | **Nombre del puesto de trabajo que desempeña el usuario.** | **Cualitativa nominal** |
| **Years of Experience** | **Años de experiencia en el puesto de trabajo.** | **Cuantitativa discreta** |
| **Salary** | **Salario anual.** | **Cuantitativa continua** |

**Realizamos una primera aproximación estadística de la distribución de los datos.**







### Limpieza de datos:

**Se realiza la limpieza y ordenación de los datos a través de diferentes cálculos y métodos de imputación.**

**Se han eliminado dos filas completas de datos nulos.**

**Se han imputado las medias correspondientes en la columna salary, filtrando por tipo de trabajo, nivel de estudios y género.**

**Se ha imputado la moda en el dato faltante de la columna Education level, filtrada por Job title y Gender.**

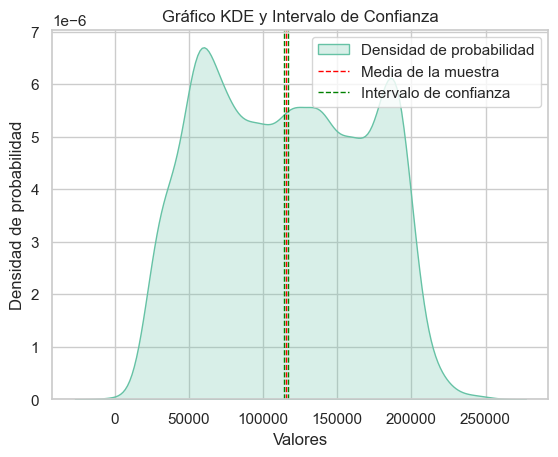
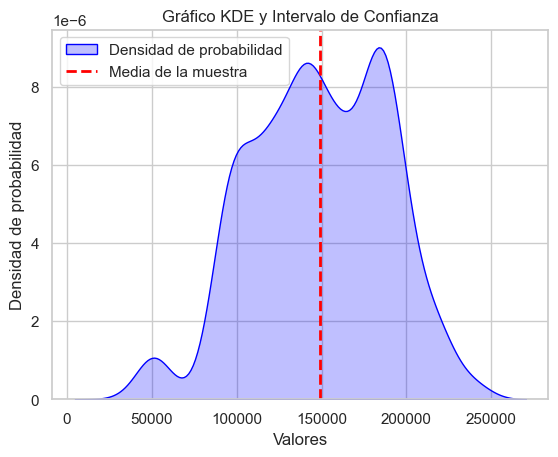
**Se han reducido el número de valores diferentes de la columna Education level, unificando valores con diferente nombre pero que contenían el mismo tipo de información.**

### Análisis univariante

### **En este apartado, analizamos la variable principal del proyecto, el salario.**

### **Se estudia la distribución del salario medio para todos los grupos de trabajo y para el grupo de empleos relacionados con los datos.**

### **Salarios datos: Salarios del conjunto general:**



**La asimetría: indica la simetría de la distribución de una variable respecto a la media aritmética.**

**La curtosis (o apuntamiento) mide cómo de escarpada o achatada está una curva o distribución. Indica la cantidad de datos que hay cercanos a la media, de manera que a mayor grado de curtosis, más escarpada (o apuntada) será la forma de la curva.**

**Curtosis: -0.46**

**Simetría: -0.21**

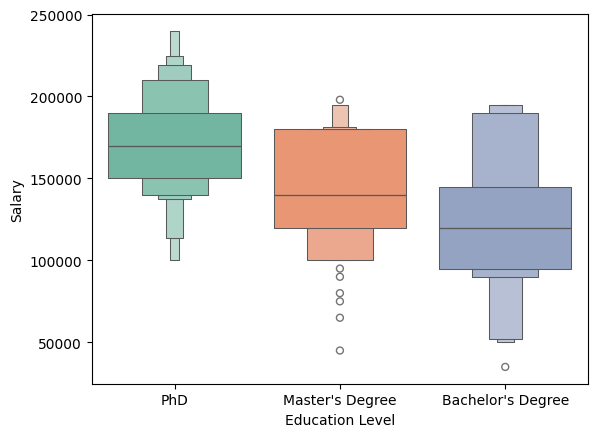
**La distribución es platicúrtica,**

**La distribución es asimétrica negativa (sesgo hacia la izquierda).**

**Según la prueba estadística de Shapiro-Wilk, la distribución de los datos de salarios dentro del campo de data, no sigue una distribución normal.**

Estudio de variables y comprobación de hipótesis.

**2. ¿Qué relación existe entre el salario y el nivel de estudios dentro del campo de Data?**

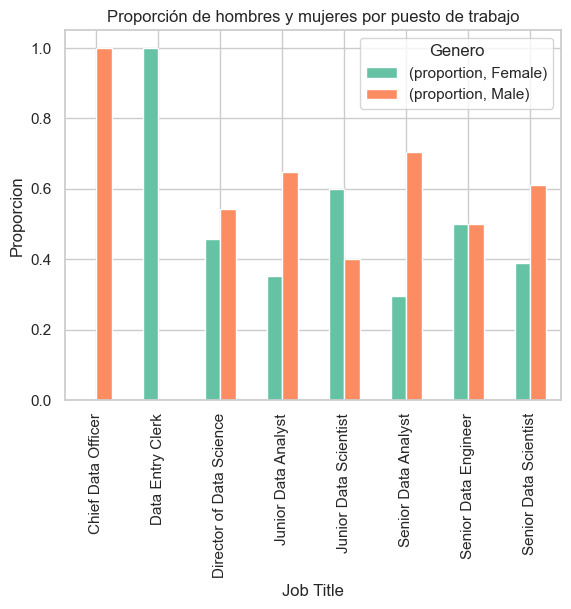
****

**Realización de la prueba de Kruskal-Wallis para más de dos muestras independientes: stat = 269.9976845259498, p\_value = 2.3482684427143934e-59**

**Rechazamos la hipótesis nula. Hay evidencia de al menos una diferencia significativa en los salarios según el nivel alcanzado de estudios.**

**3. ¿Cuál es la proporción de hombres y mujeres en cada tipo de empleo?**

**Se representa en un gráfico de barras la distribución de hombres y mujeres por tipo de puesto de trabajo. A simple vista, se aprecian diferencias entre las cantidades totales, pero comprobamos estadísticamente si es así en el puesto de trabajo Director of Data Science.**

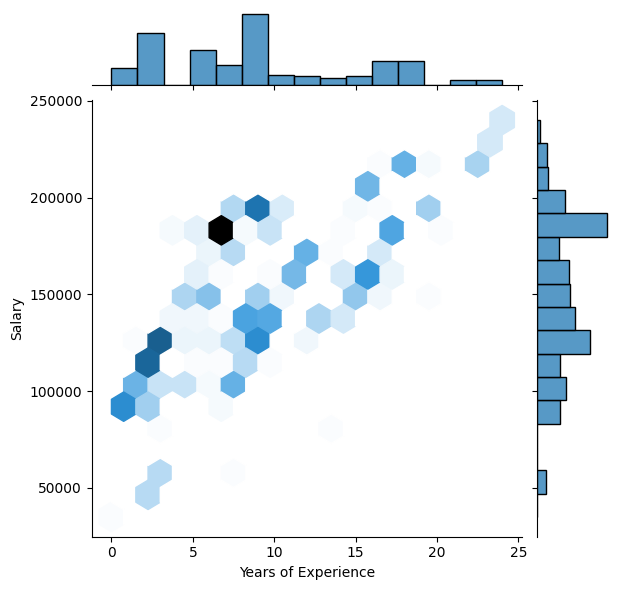
****

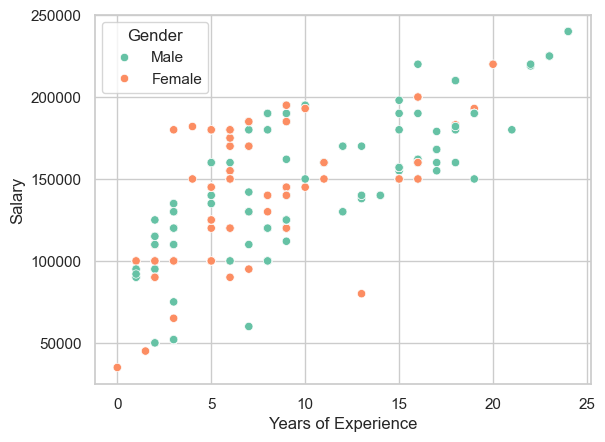
**Utilizo la prueba de proporciones Z para rechazar o no la hipótesis nula. El resultado es el siguiente:**

**No hay suficiente evidencia para rechazar la hipótesis nula. No hay diferencia significativa en la proporción de hombres y mujeres dentro del puesto de Director of Data Science**

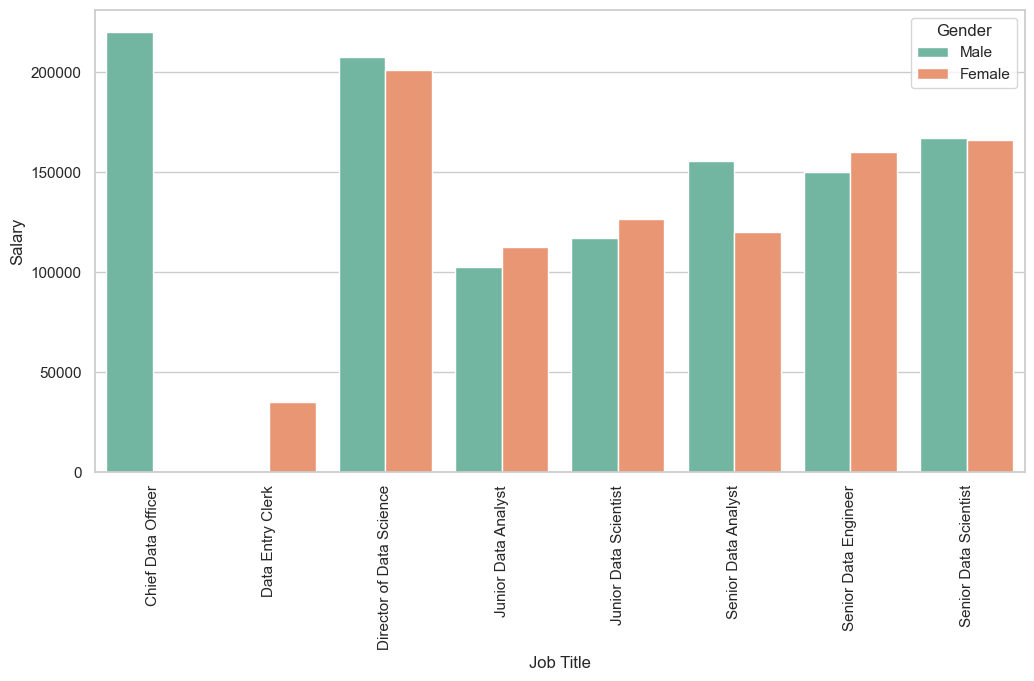
**4. ¿Qué relación existe entre los años de experiencia y el salario?**

**Se estudia la correlación entre ambas variables, dando como resultado una correlación positiva de 0.71282707.**

****

****

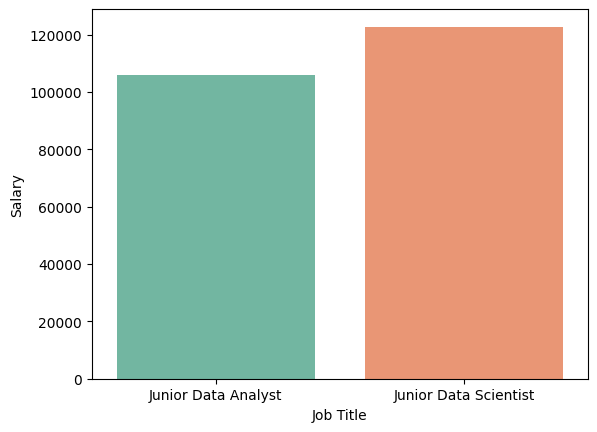
**5. ¿Dentro del campo de data, hay diferencias entre los salarios de hombre y mujeres?**

****

**En este caso, utilizo la prueba de Mann-Witney estudiar dos muestras independientes. El resultado es el siguiente:**

**No hay suficiente evidencia para rechazar la hipótesis nula. No hay diferencia significativa entre los salarios de las mujeres y los salarios de los hombres en el campo de data**

**6. Dentro del campo de data, ¿quién tiene mayor salario de entrada?**

****

**De nuevo utilizo Prueba de Mann-Whitney para dos muestras independientes:**

**Rechazamos la hipótesis nula. Hay evidencia de una diferencia significativa entre los salarios de los junior data scientist y data analyst.**

FUENTES DE DATOS

Notebooks proporcionados por The Bridge.

<https://www.kaggle.com/>

<https://www.python.org/>

<https://bookdown.org/>

<https://numpy.org/>

<https://matplotlib.org/>

<https://pandas.pydata.org/>

<https://seaborn.pydata.org/>