MINERIA DE CONTRASTES

JOSE MARIN GONZALEZ ZAVALA

MAY 2023

Descripción del proyecto

El proyecto consiste en implementar un flujo de minería de datos sobre un conjunto de datos con información salarial de una organización. El objetivo principal es encontrar patrones y relaciones significativas entre diferentes características de los colaboradores, como su género, edad, nivel educativo, departamento, seniority, rendimiento de evaluación y sueldo. Para ello, se utilizará la técnica de minería de datos llamada Minería de contrastes (Contrast Set Mining, CSM).

A lo largo del proyecto se resolverán las siguientes cuestiones:

- Cuáles son las características de los datos con las que trabajaremos.
- Cuáles son los departamentos con más colaboradores.
- La distribución de géneros, edades, sueldos y bonos.
- El contraste entre el sueldo y las demás características.
- Se analizará la relación que tengan las distintas variables.
- Se probarán hipótesis sobre las medias salariales en diferentes situaciones como hombres vs mujeres y managers vs el resto de los puestos.

Análisis exploratorio de datos.

Descripcion de los datos

Los datos que se utilizarán en este proyecto están descritos en formato CSV y contienen información sobre el puesto, género, edad, rendimiento de evaluación, nivel educativo, departamento, seniority, sueldo base y bonos de un conjunto de colaboradores de una organización.

Características

A continuación se describen las características principales de los datos con los que se trabajará:

- **JobTitle**: El puesto al cual pertenece el colaborador.
- **Gender**: El género del colaborador.
- **Age**: La edad del colaborador.
- **PerfEval**: El rendimiento de evaluación del colaborador, en una escala del 1 al 5 donde 1 es el mínimo y 5 es el máximo.
- **Edu**: El nivel educativo del colaborador.
- **Dept**: El área/departamento de trabajo del colaborador.
- **Seniority**: La antigüedad/maestría del puesto del colaborador.
- **BasePay**: El sueldo base del colaborador.
- **Bonus**: Los bonos de salario del colaborador.

Exploración inicial de datos

Antes de comenzar con el análisis de los datos, es importante realizar una exploración inicial para conocer las características de los mismos. Esta exploración nos ayuda a darnos una idea inicial de las columnas que tenemos, observar el tipo de información que contiene, el tipo de datos que contiene, si es que tenemos valores ausentes y detectar valores atípicos de manera rápida.

```
              jobTitle  gender  age  perfEval      edu            dept  \
0     Graphic Designer  Female   18         5  College      Operations
1    Software Engineer    Male   21         5  College      Management
2  Warehouse Associate  Female   19         4      PhD  Administration
3    Software Engineer    Male   20         5  Masters           Sales
4     Graphic Designer    Male   26         5  Masters     Engineering

   seniority  basePay  bonus
0          2    42363   9938
1          5   108476  11128
2          5    90208   9268
3          4   108080  10154
4          5    99464   9319

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   jobTitle   1000 non-null   object
 1   gender     1000 non-null   object
 2   age        1000 non-null   int64
 3   perfEval   1000 non-null   int64
 4   edu        1000 non-null   object
 5   dept       1000 non-null   object
 6   seniority  1000 non-null   int64
 7   basePay    1000 non-null   int64
 8   bonus      1000 non-null   int64
dtypes: int64(5), object(4)
memory usage: 70.4+ KB
None

               age     perfEval    seniority        basePay         bonus
count  1000.000000  1000.000000  1000.000000    1000.000000   1000.000000
mean     41.393000     3.037000     2.971000   94472.653000   6467.161000
std      14.294856     1.423959     1.395029   25337.493272   2004.377365
min      18.000000     1.000000     1.000000   34208.000000   1703.000000
25%      29.000000     2.000000     2.000000   76850.250000   4849.500000
50%      41.000000     3.000000     3.000000   93327.500000   6507.000000
75%      54.250000     4.000000     4.000000  111558.000000   8026.000000
max      65.000000     5.000000     5.000000  179726.000000  11293.000000
```

En este caso, se puede observar que no existen valores ausentes en los datos y que todos los tipos de datos hacen sentido al nombre de la columna y su contenido. Además, en cuanto a las siguientes columnas:

- Age: se tienen valores máximo y mínimos lógicos que van de 18 a 65.
- PerfEval y Seniority se mantienen en el rango de 1 a 5, ya que son las únicas calificaciones existentes.
- BasePay y Bonus hacen sentido a los sueldos de las profesiones mencionadas.

Feature Engineering

Para analizar los datos de una manera más efectiva, se crearán tres nuevas columnas:

- total_salary, que será la suma de basePay y bonus.
- total_salary_range, que es la agrupación de total_salary para volver categórica una variable continua.
- age_group, que es la agrupación de age para volver categórica una variable continua.

También se renombrarán las columnas para aplicar mejores prácticas de código, como separar las palabras con un "_" y convertir todas las letras a minúsculas.

Out[9]:

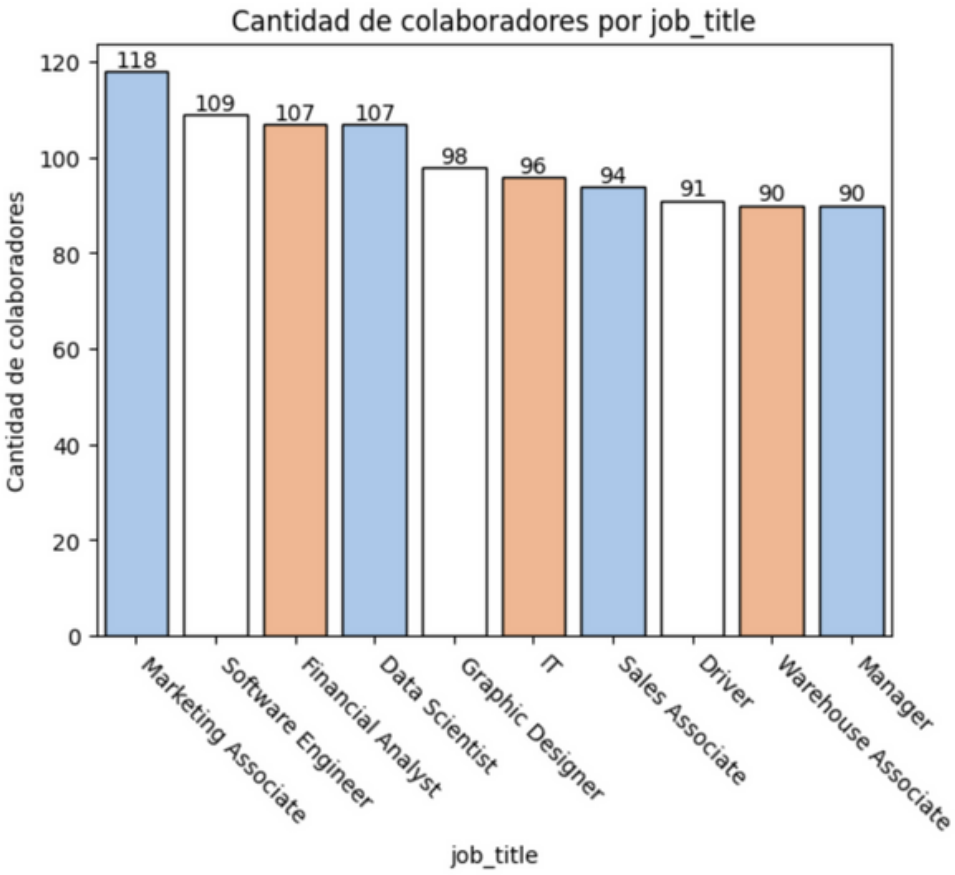
ob_title	gender	age	perf_eval	edu	dept	seniority	base_pay	bonus	total_salary	total_salary_range	age_group
Graphic Designer	Female	18	5	College	Operations	2	42363	9938	52301	40,000-80,000	0-19
Software Engineer	Male	21	5	College	Management	5	108476	11128	119604	80,001-120,000	20-39
ırehouse ssociate	Female	19	4	PhD	Administration	5	90208	9268	99476	80,001-120,000	0-19
Software Engineer	Male	20	5	Masters	Sales	4	108080	10154	118234	80,001-120,000	20-39
Graphic Designer	Male	26	5	Masters	Engineering	5	99464	9319	108783	80,001-120,000	20-39

Roles de trabajo

Se puede observar una uniformidad entre los roles de la empresa en un rango de 90 a 118 colaboradores donde. Así como se puede concluir que por las características de los puestos esta es una empresa que vende algún producto el cual debe ser almacenado y enviado, pero con un gran enfoque a lo tecnológico por la gran cantidad de puestos como ingeniero de software y data scientist, muy probablemente se trata de un E-commerce.

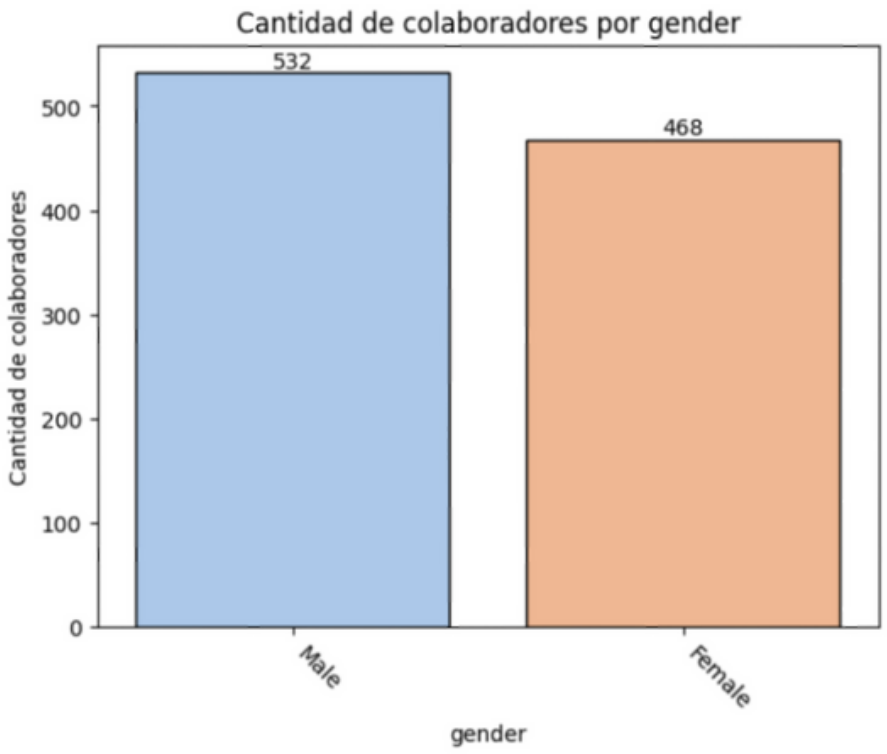
Otros datos a destacar son:

- El título con más colaboradores es Marketing Associate.
- El título con menos colaboradores es Manager.



Géneros en la empresa

Se puede observar una uniformidad entre los géneros de la empresa con una variación de 64 colaboradores entre grupos, por lo cual se consideraría balanceado.

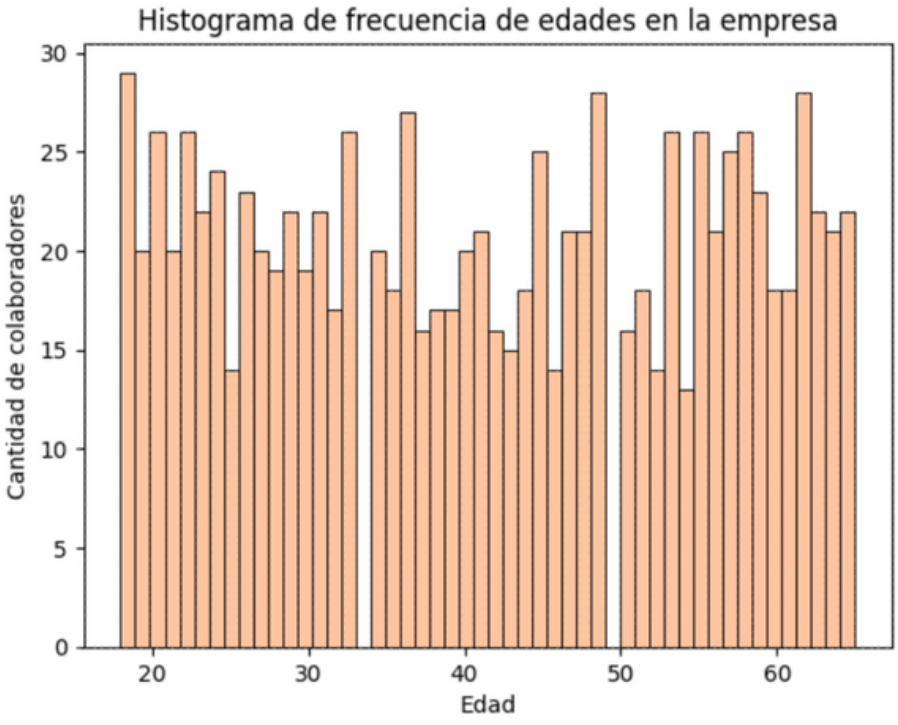


Principales frecuencias de las edades

Se puede observar una uniformidad entre las edades en la empresa con picos alrededor de 18, 33, 37, 48 y 62 años. También se pueden destacar valles en 26, 35 y 49 años.

Otros datos a destacar son:

- La edad mínima es de 18 años.
- La mediana de edad es de 41 años.
- La edad máxima es de 65 años.



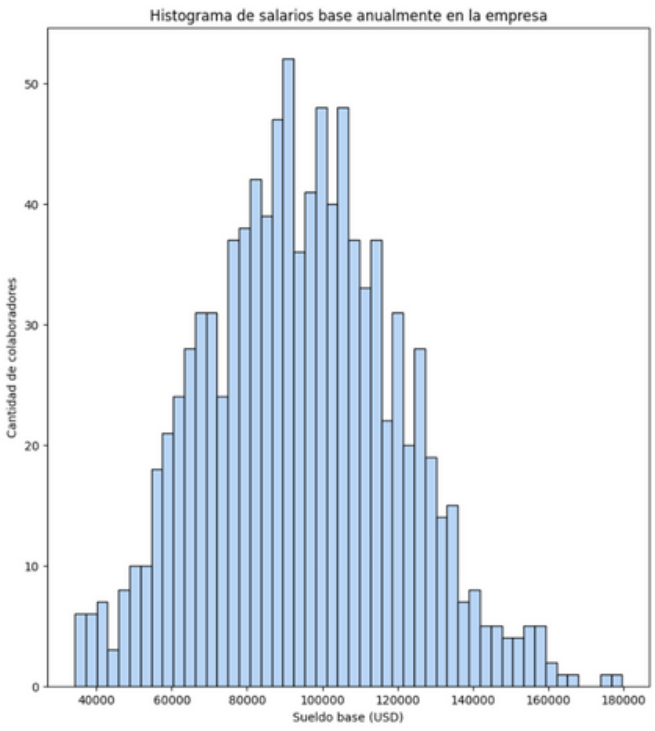
Sueldos base y bonos

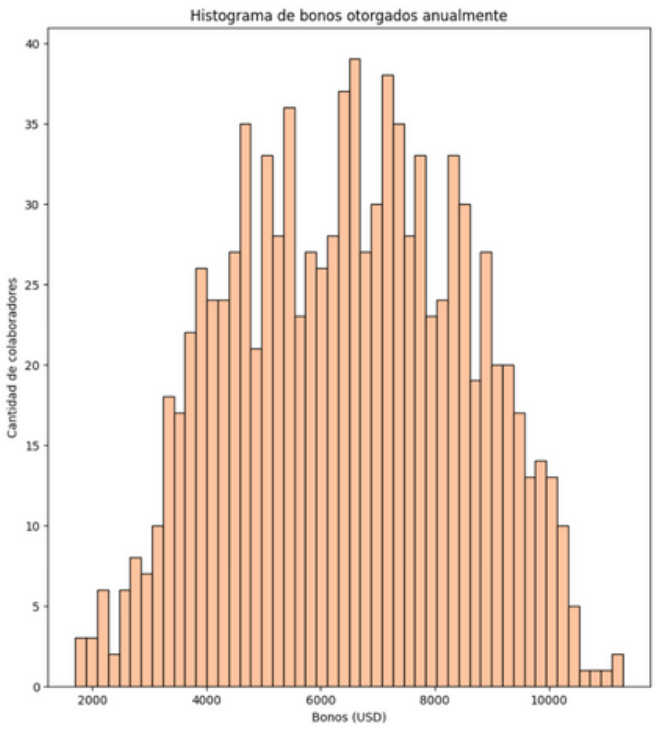
Se puede observar una distribución normal para ambos conjuntos de datos, aunque ligeramente más distribuidos en los bonos.

Los rangos en los que se encuentran los sueldos base es de $34,208 a $17,9726 y los bonos se encontrarían en un rango de $1,703 a $11,293.

Otros datos a destacar son:

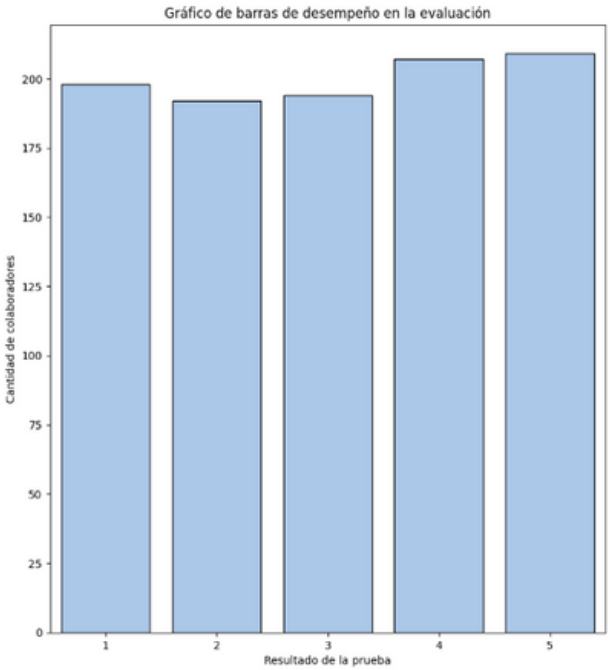
- La mayoría de salarios base se encuentra distribuido alrededor de $80,000 y $100,000 USD.
- La mayoría de bonos se encuentran distribuidos alrededor de $4,000 y $9,000 USD.

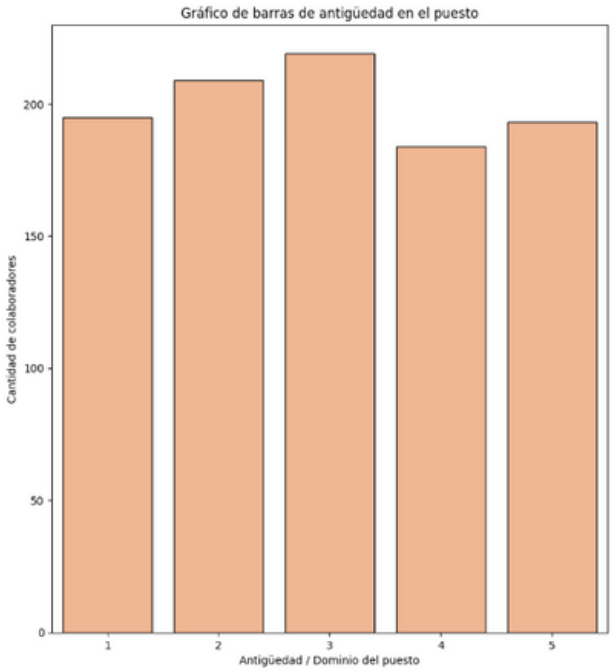




Performance y Seniority

Se puede observar una uniformidad para ambos conjuntos de datos.
Para la nota de evaluación se puede destacar que la mayoría de resultados se muestran en las calificaciones 4 y 5, lo que nos dice que la empresa tiene un buen desempeño en general por parte de los colaboradores.
Para la etiqueta de seniority se puede destacar que la mayoría de colaboradores se encuentran en un nivel mid/medio, mientras que la menor parte de colaboradores se encuentran en un nivel semi senior y hay un poco más en un nivel senior.



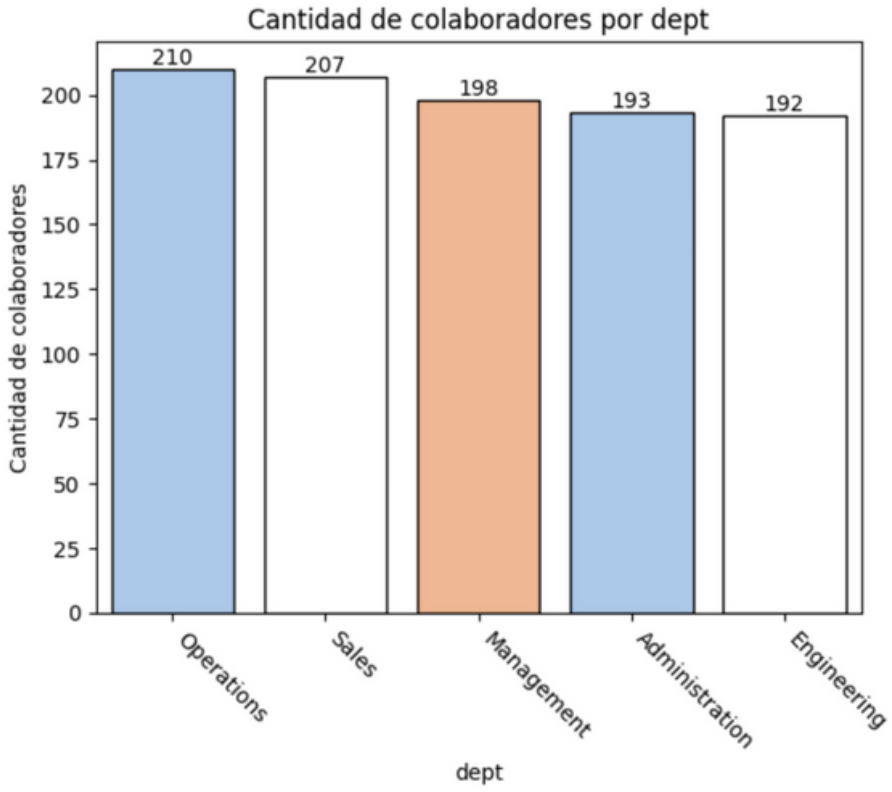


Departamentos

Se puede observar una uniformidad entre los roles de la empresa en un rango de 192 a 210 colaboradores donde.

Datos a destacar son:

- El departamento con más colaboradores es Operations.
- El departamento con menos colaboradores es Engineering.

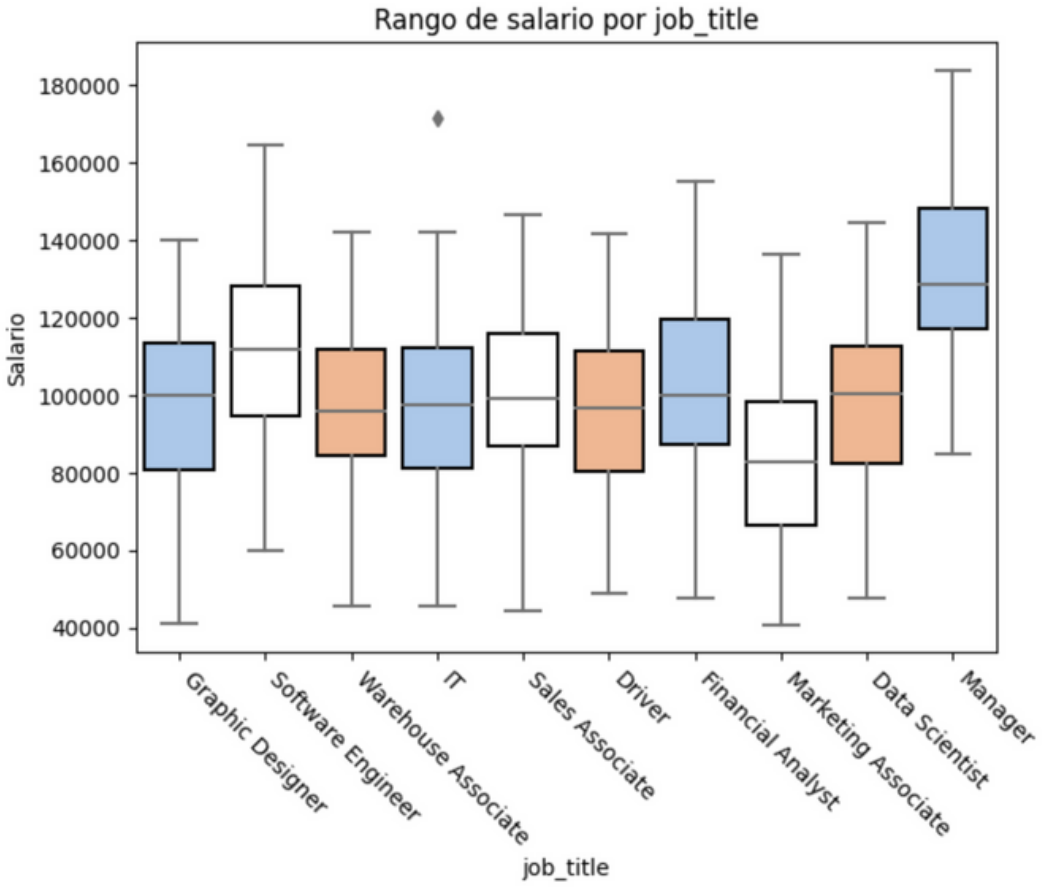


Sueldos por rol de trabajo

Se puede destacar que los managers se llevan la delantera en cuanto a sueldos considerablemente, así como que este es el puesto que menos se encuentra en la empresa, los demás puestos se encuentran en un rango similar.

Otros datos a destacar son:

- Solo las personas de TI tienen sueldos con valores atípicos.
- El título con menor salario es Marketing Associate y recordemos que es el puesto que más se presenta en la empresa.

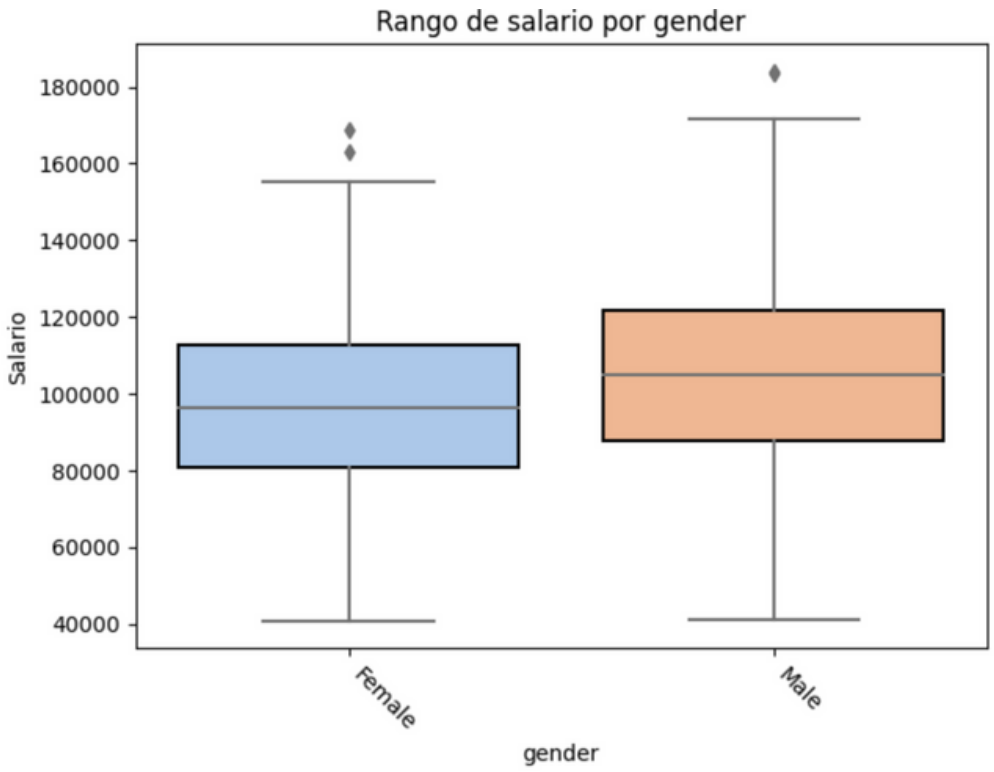


Sueldos por role de genero

Se puede destacar que los hombres se llevan la delantera en cuanto a sueldos aunque muy poco.

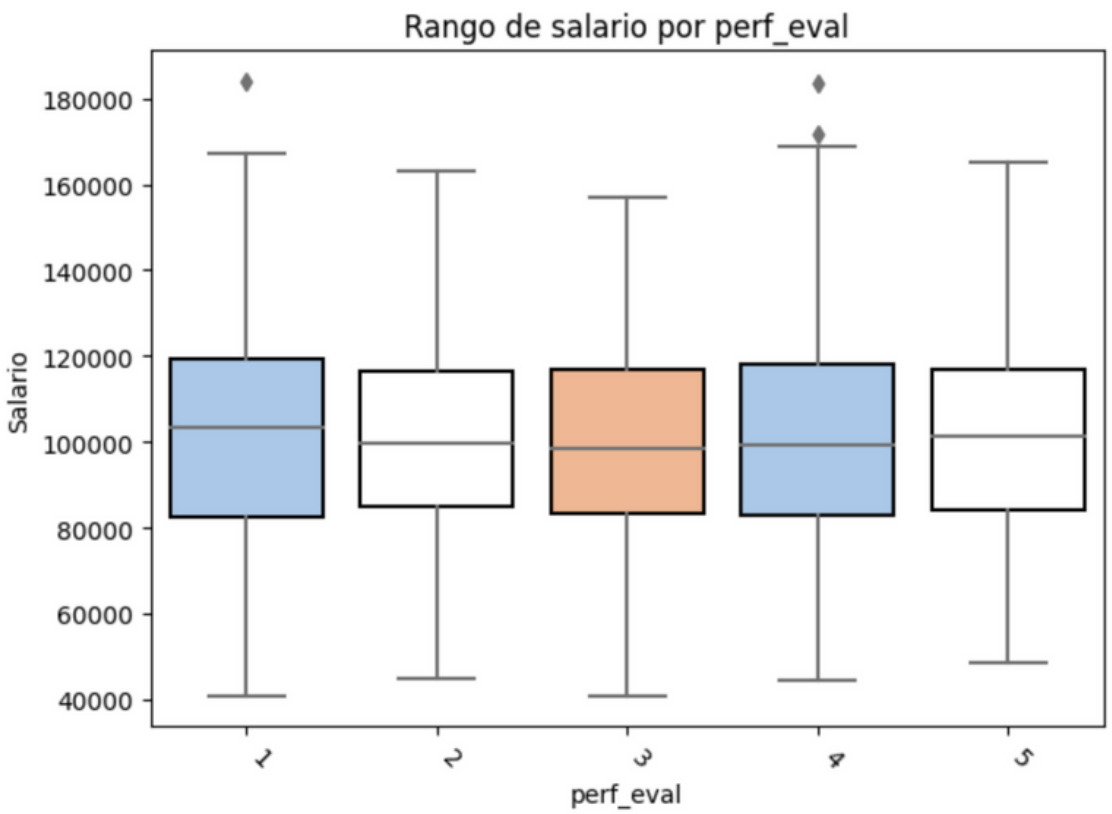
Otros datos a destacar son:

- El mayor sueldo es de un hombre.
- Las mujeres tienen más valores atípicos.



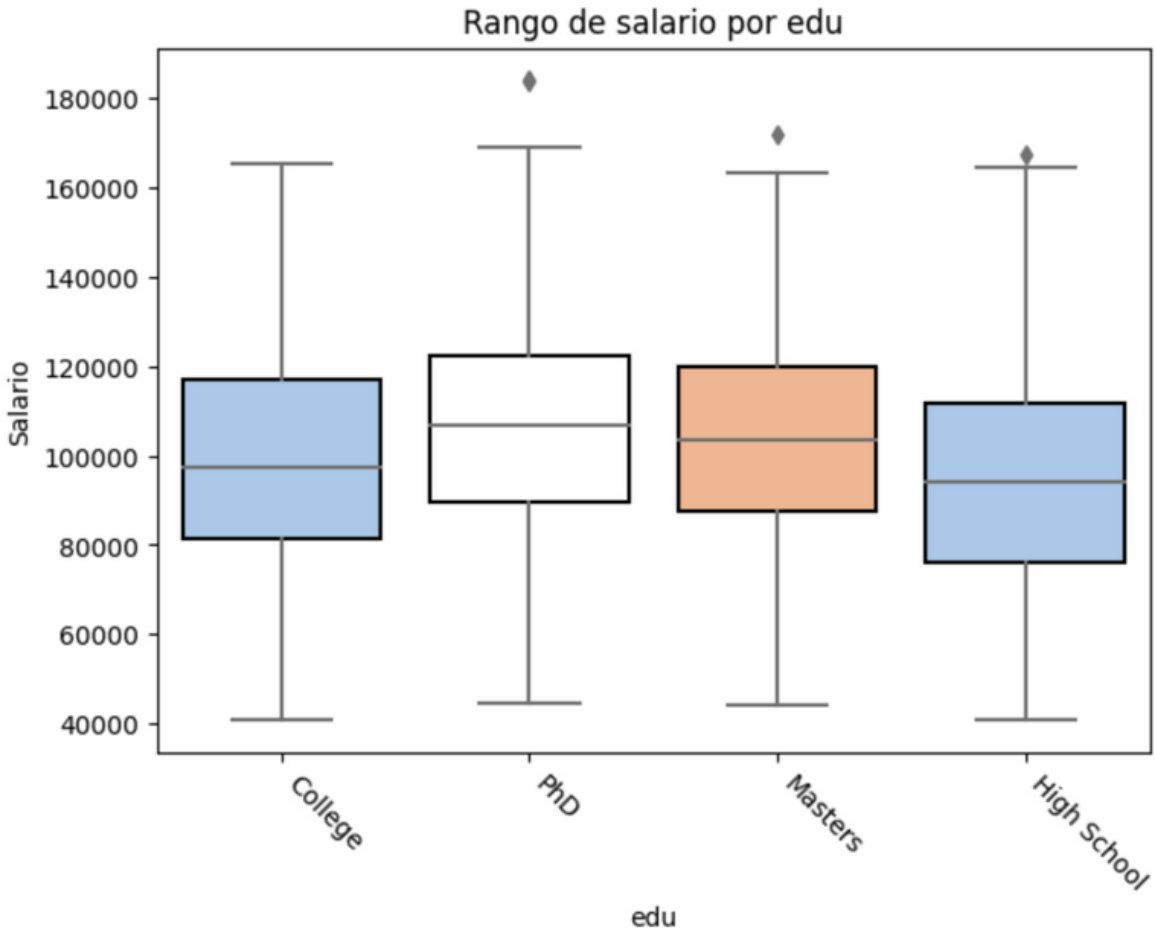
Sueldos por rendimiento de evaluación

Se puede observar cómo todos los desempeños están en el mismo rango por lo cual no atribuiría algún cambio en el sueldo al desempeño.



Sueldos por educacion

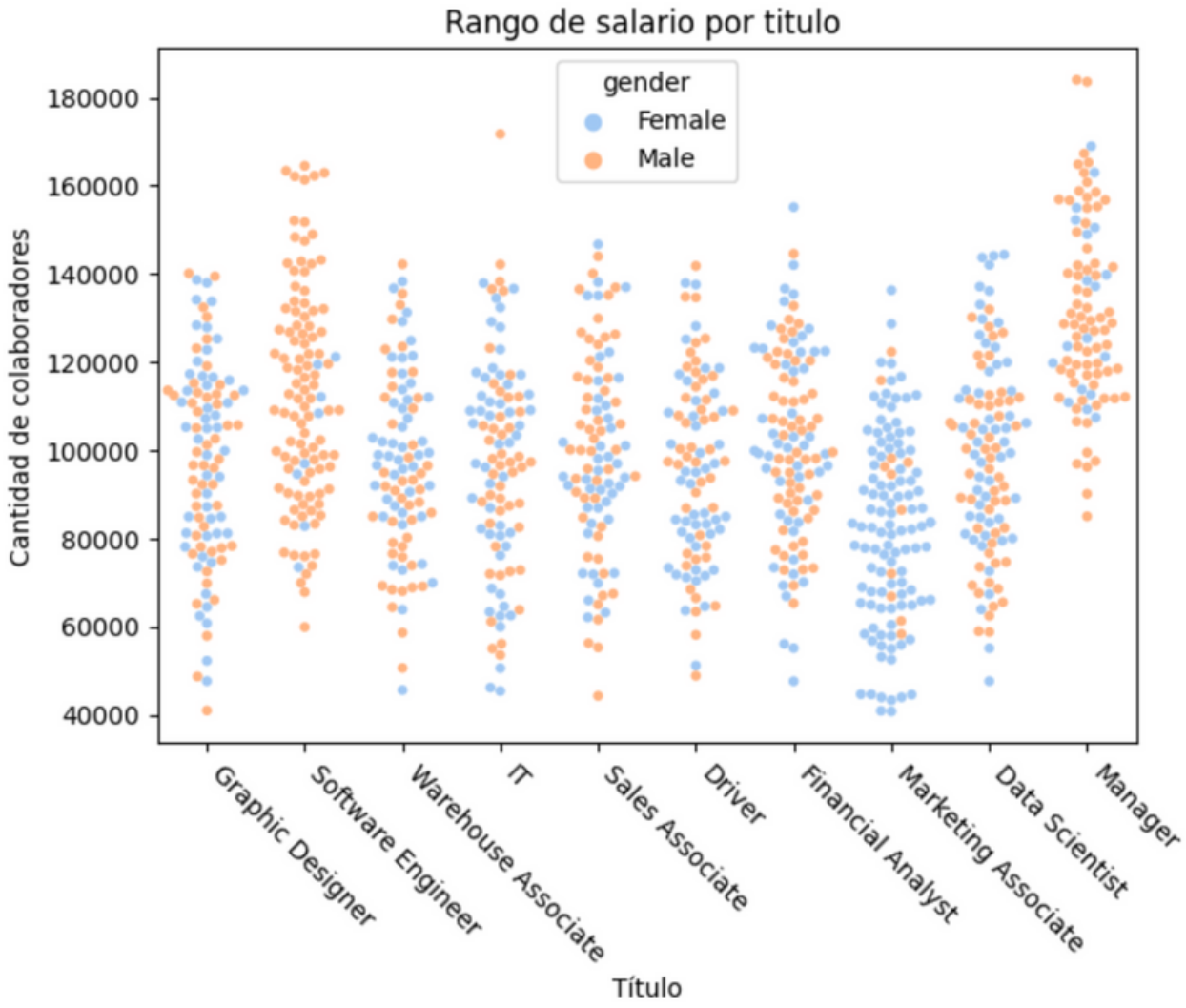
Se observa una correlación positiva respecto al nivel de educación, es decir, existe una relación directa entre estudios y sueldo (entre más estudies más ganarás).



Sueldos por título y por género

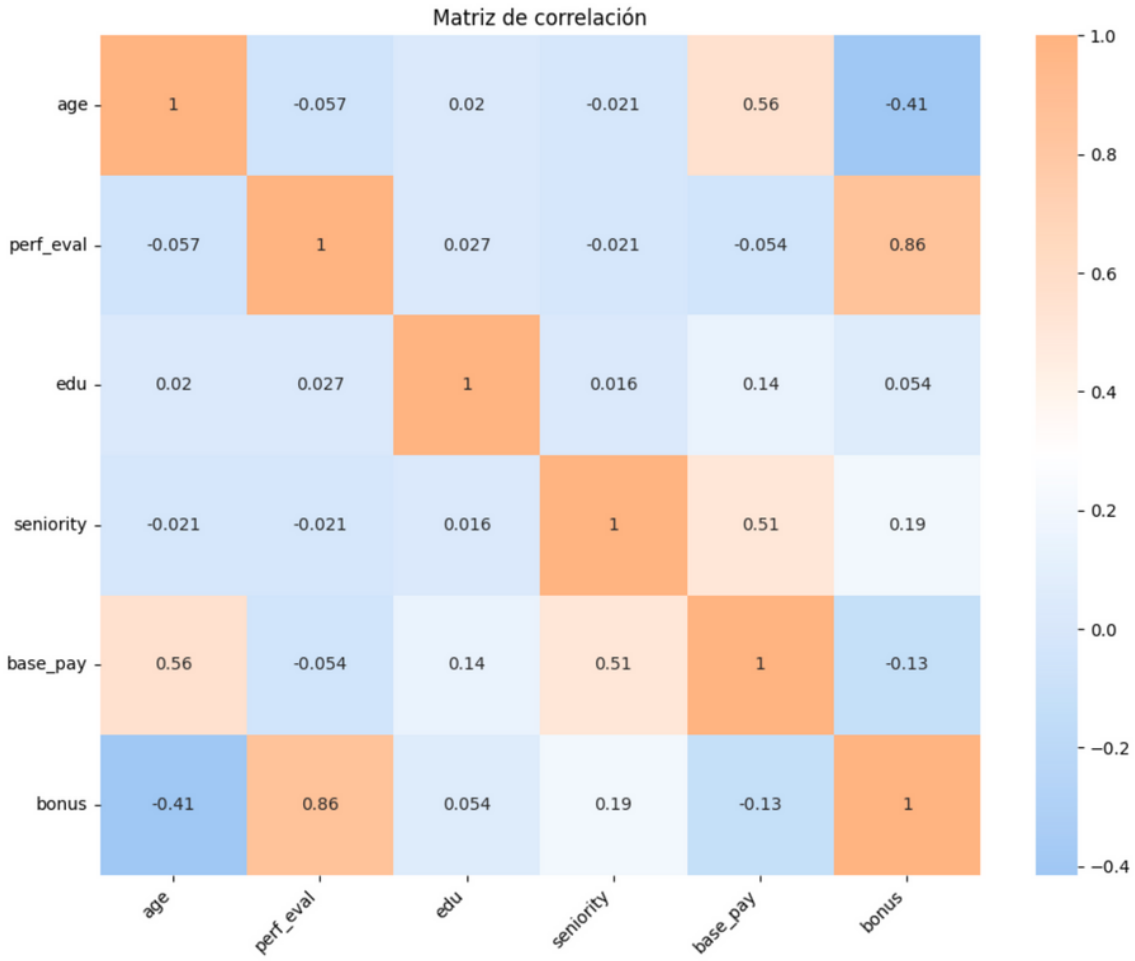
Se puede destacar que:

- Los managers hombres se llevan la delantera en cuanto a sueldos considerablemente y las mujeres del área de marketing son las que cuentan con los salarios más bajos.
- Los ingenieros de software engineer son principalmente hombres.
- El puesto de marketing associate está compuesto principalmente por mujeres.



Exploración de correlación

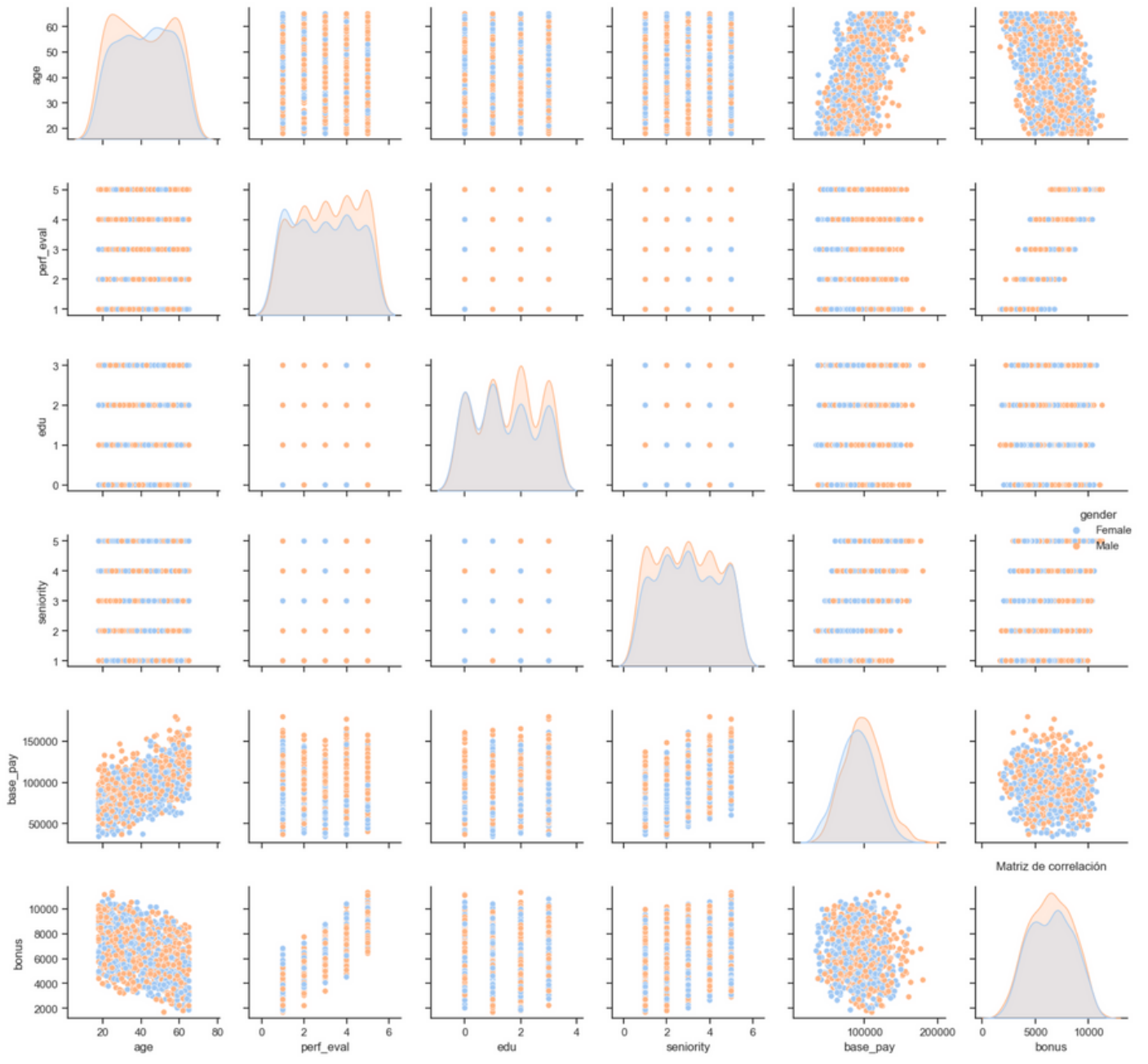
En la exploración de correlación se buscarán relaciones entre las diferentes variables numéricas.
Se discriminarán las columnas del sueldo que repiten información para tener datos claros y solamente se dejará total_salary, ya que contiene toda la información que necesitamos. Además, se realizará un etiquetado cardinal de características con la columna.



En conclusión, se han explorado los datos de una organización en busca de patrones y relaciones significativas entre diferentes características de los colaboradores, como su género, edad, nivel educativo, departamento, seniority, rendimiento de evaluación y sueldo. Se han encontrado relaciones interesantes, como la correlación positiva entre el nivel educativo y el sueldo, y se han destacado datos interesantes, como que los managers hombres se llevan la delantera en cuanto a sueldos.

Estos hallazgos pueden ser útiles para la organización al momento de tomar decisiones importantes, como la asignación de salarios o la contratación de nuevos colaboradores.

Exploración de correlación respecto al genero



Conclusiones EDA

Conclusiones generales del análisis univariado y bivariado de características

- Esta empresa probablemente sea un e-commerce debido a que cuenta con envíos y almacenes, pero tiene una gran cantidad de desarrolladores de software y científicos de datos.
- La empresa está constituida uniformemente entre todos los departamentos, pero podemos destacar personal en marketing y desarrollo de software.
- Los géneros se encuentran equilibrados aunque ligeramente hay más hombres.
- Encontramos todo rango de edades en la empresa desde 18 años hasta 65 años.
- Los rangos en los que se encuentran los sueldos base es de $34,208 a $17,9726 y los bonos se encontrarían en un rango de $1,703 a $11293.
- La mayoría de colaboradores se encuentran en un nivel mid/medio.
- Departamento con más colaboradores: Operations y que menos es Engineering.
- Podemos destacar que los managers se llevan la delantera en cuanto a sueldos considerablemente, así como que este es el puesto que menos se encuentra en la empresa, los demás puestos se encuentran en un rango similar.
- Podemos destacar que los hombres se llevan la delantera en cuanto a sueldos aunque muy poco.
- Todas las pruebas de desempeño están en el mismo rango por lo cual no atribuiría algún cambio en el sueldo al desempeño.
- Existe una correlación positiva respecto al nivel de educación, en otras palabras, existe una relación directa entre estudios y sueldo (entre más estudies más ganarás).
- Los ingenieros de Software Engineer son principalmente hombres.
- El puesto de Marketing Associate está compuesto principalmente por mujeres

Conclusiones EDA

Conclusiones generales de la correlación entre variables
En este apartado se trabajará con un mapa de calor de correlación de variables del cual podemos destacar:

- Existe una correlación positiva respecto a la edad y el sueldo base, lo que significa que una mayor edad generalmente representa un mayor sueldo.
- Existe una correlación negativa respecto a la edad y el bono base, lo que significa que una mayor edad generalmente representa un menor bono.
- La edad no está relacionada con el nivel de experiencia, el nivel de educación o el desempeño.
- El bonus está positivamente relacionado con el desempeño, entre más bonus mejor desempeño.
- El nivel de educación está ligeramente relacionado positivamente con el sueldo base y entre mejor educación tengas, tendrás un mejor sueldo, pero no hay un impacto muy grande.
- Las personas con mayor dominio del trabajo ganan más en bonos, pero ganan aún más de salario base.
- No hay relación entre el dominio del puesto y las otras variables.

Conclusiones generales de los test de hipótesis

- El promedio de ingresos entre géneros difiere.
- El promedio de ingresos entre puesto difiere respecto a los managers

Pre-procesamiento de datos.

Buenas practicas del pre-procesamiento de datos

1. **Limpieza de datos**: Los datos pueden venir en formatos diferentes, y a menudo están llenos de errores, valores faltantes, o ruido. Debes limpiar tus datos removiendo o corrigiendo los errores, y tratando los valores faltantes de una manera que tenga sentido para tu análisis. Los métodos comunes para tratar los valores faltantes incluyen la imputación (rellenar con valores promedio o medianos), la eliminación de las filas o columnas afectadas, o el uso de algoritmos que puedan manejar los valores faltantes.
2. **Normalización**: Algunos algoritmos de minería de datos son sensibles a la escala de los datos. Por ejemplo, una característica con un rango de valores mucho más grande que otra puede dominar en el proceso de aprendizaje. La normalización ayuda a resolver este problema escalando todos los datos a un rango común, como 0-1.
3. **Transformación de datos**: Esto puede incluir la codificación de variables categóricas (como convertir una variable de género en '0' para hombres y '1' para mujeres), la creación de nuevas variables a partir de las existentes (ingeniería de características), la discretización de variables continuas, entre otras.
4. **Reducción de datos**: Los conjuntos de datos pueden ser muy grandes y difíciles de manejar. La reducción de datos intenta encontrar una representación más pequeña de los datos que preserva la utilidad de los mismos para el problema en cuestión. Esto puede implicar técnicas como la selección de características (escoger un subconjunto de características que son más relevantes para el problema), la selección de instancias (escoger un subconjunto de instancias que son más representativas del conjunto completo), o la aplicación de técnicas de reducción de dimensionalidad (como el Análisis de Componentes Principales).
5. **Balanceo de datos**: En problemas de clasificación, a menudo se tiene un número desigual de instancias para cada clase. Algunos algoritmos de minería de datos pueden sesgarse hacia la clase mayoritaria, lo que puede llevar a resultados pobres para la clase minoritaria. El balanceo de datos busca resolver este problema, ya sea recopilando más datos para la clase minoritaria, submuestreando la clase mayoritaria, o generando ejemplos artificiales de la clase minoritaria.

Reduccion de datos.

Dado que el modelo que desarrollaremos se basa en la busca de relacion entre caracteristicas categoricas es nuestro deber trabajar con las columnas lineales como serian la edad, el sueldo base y los bonos para agrupar eso en rangos y de esta manera volverlas variables categoricas.

Esto lo realizaremos por medio de funciones las cuales nos ayudaran a agrupar estos datos, para posteriormente eliminar las columnas lineales y solamente trabajar con las categoricas.

Modelado del argotimo

Mining for Contrasting Sets (STUCCO)

STUCCO es un algoritmo de minería de datos que se utiliza para descubrir conjuntos de contraste significativos entre diferentes grupos de datos. El algoritmo se basa en la idea de encontrar conjunciones de atributos y valores que difieren de manera significativa en su distribución entre los grupos.

El proceso implica la construcción de un árbol de búsqueda de contrastes, donde se generan conjuntos de contraste especializando conjuntos vacíos mediante la adición de términos adicionales. Luego, se calcula el soporte de cada conjunto de contraste en cada grupo y se realizan pruebas de significancia para determinar si las diferencias son estadísticamente significativas.

Utiliza diferentes estrategias de poda para reducir la complejidad computacional y descartar conjuntos de contraste que no cumplen con los criterios de significancia. Además, se implementa un control de errores tipo I para evitar falsos positivos al realizar múltiples pruebas de hipótesis.

Definamos nuestros grupos de contraste

De esta manera podremos enfocar nuestro caso de estudio y empezar a calcular medidas imporntantes como sera el soporte, lo que en palabras simpres en la frencuencia con la que se repite una clase

< 40k

Puesto	Genero	Rango_edad	Educacion
DS	H	20-39	Collegue
SE	M	< 19	Collegue
MK	H	20-39	Collegue

SUP (Puesto = DS ∧ Genero = H | < 40k) = 1 / 3 = 33%

40k - 80k

Puesto	Genero	Rango_edad	Educacion
SE	M	40-59	Collegue
GD	M	20-39	Collegue
DS	H	40-59	Master

SUP (Educacion = Collegue ∧ Genero = H | 40k - 80k) = 2 / 3 = 66%

80k - 120k

Puesto	Genero	Rango_edad	Educacion
FA	H	20-39	Master
MG	H	40-59	PhD
MG	H	> 60	PhD

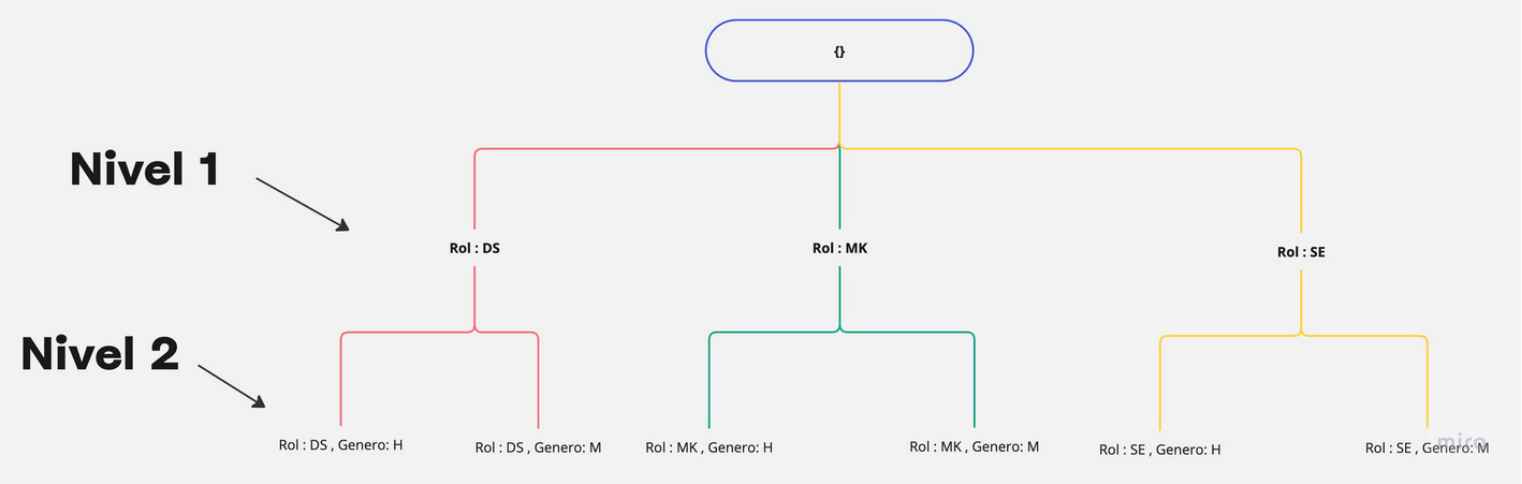
SUP (Puesto = MG ∧ Genero = H | < 40k) = 0 / 3 = 0%

Es importante medir esto ya que mas adelante nos ayudara a hacer los calculos necesarios para filtrar con las siguientes condiciones:

Definición de **desviaciones**: • Una desviación es un conjunto de contrastes que es **significativo** y **grande**. • Un conjunto de contrastes en el cual al menos dos grupos difieren en su soporte se denomina **significativo**. • Un conjunto de contrastes en el cual la diferencia máxima entre los soportes es mayor que un parámetro mindev se denomina grande.

Niveles del arbol de busqueda

De esta manera podremos enfocar nuestro caso de estudio y empezar a calcular medidas imporntantes como sera el soporte, lo que en palabras simpres en la frencuencia con la que se repite una clase



- **Nivel 1**: En el nivel 1, sólo se considera una característica a la vez. En el contexto de tus datos, se generarán conjuntos de contraste para cada característica individualmente. Por ejemplo, "job_title" = "Software Engineer", "gender" = "Male", etc.
- **Nivel 2**: En el nivel 2, se consideran dos características a la vez. Por ejemplo, se podrían generar conjuntos de contraste para combinaciones de "job_title" y "gender", como ("job_title" = "Software Engineer", "gender" = "Male").
- **Nivel 3**: En el nivel 3, se consideran tres características a la vez, y así sucesivamente.

Podado de nodos

Para determinar si la diferencia en el soporte es estadísticamente significativa, realizamos una prueba estadística, en particular, una prueba de chi-cuadrado. La hipótesis nula de esta prueba es que "El soporte para el conjunto de contraste es el mismo en todos los grupos". En otras palabras, suponemos inicialmente que no hay diferencia en el soporte entre los grupos.

Calculamos la estadística de chi-cuadrado, que mide cuánto difieren las distribuciones observadas de lo que esperaríamos si la hipótesis nula fuera cierta. A continuación, comparamos este valor con la distribución chi-cuadrado para determinar el valor p de la prueba, que es la probabilidad de obtener un resultado al menos tan extremo como el observado si la hipótesis nula fuera cierta.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$O \rightarrow$ *Observed values*

$$E_{ij} = \frac{\sum_{i=1}^{2} O_{ij} \sum_{j=1}^{c} Oij}{N}$$

$E \rightarrow$ *Expected values*

$N \rightarrow$ *total number of observations*

Podado de nodos

Estos nodos son importante ser podados nos ayuda a determinar que grupos de contraste son mas importantes y ayuda a reducir el coste computacional filtrar de este modo.

De este modo podremos obtener un mejor resumen de las

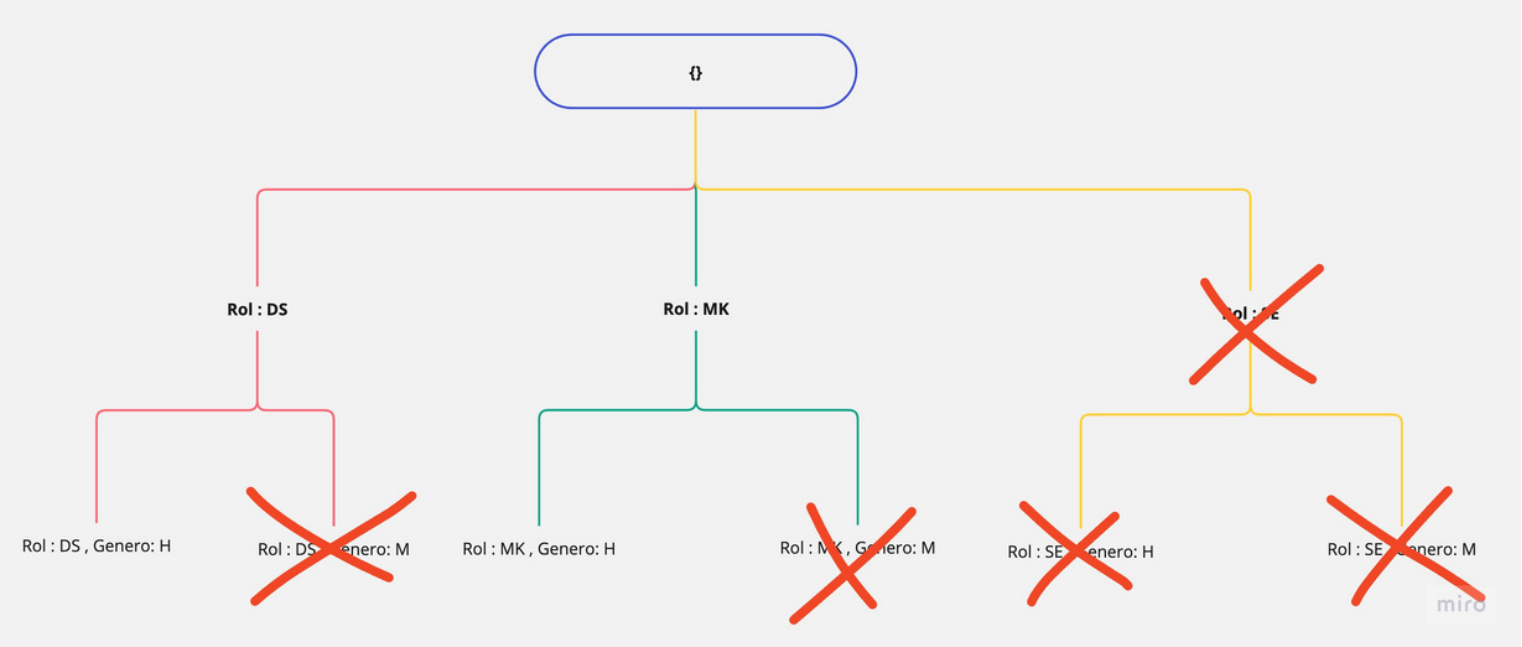
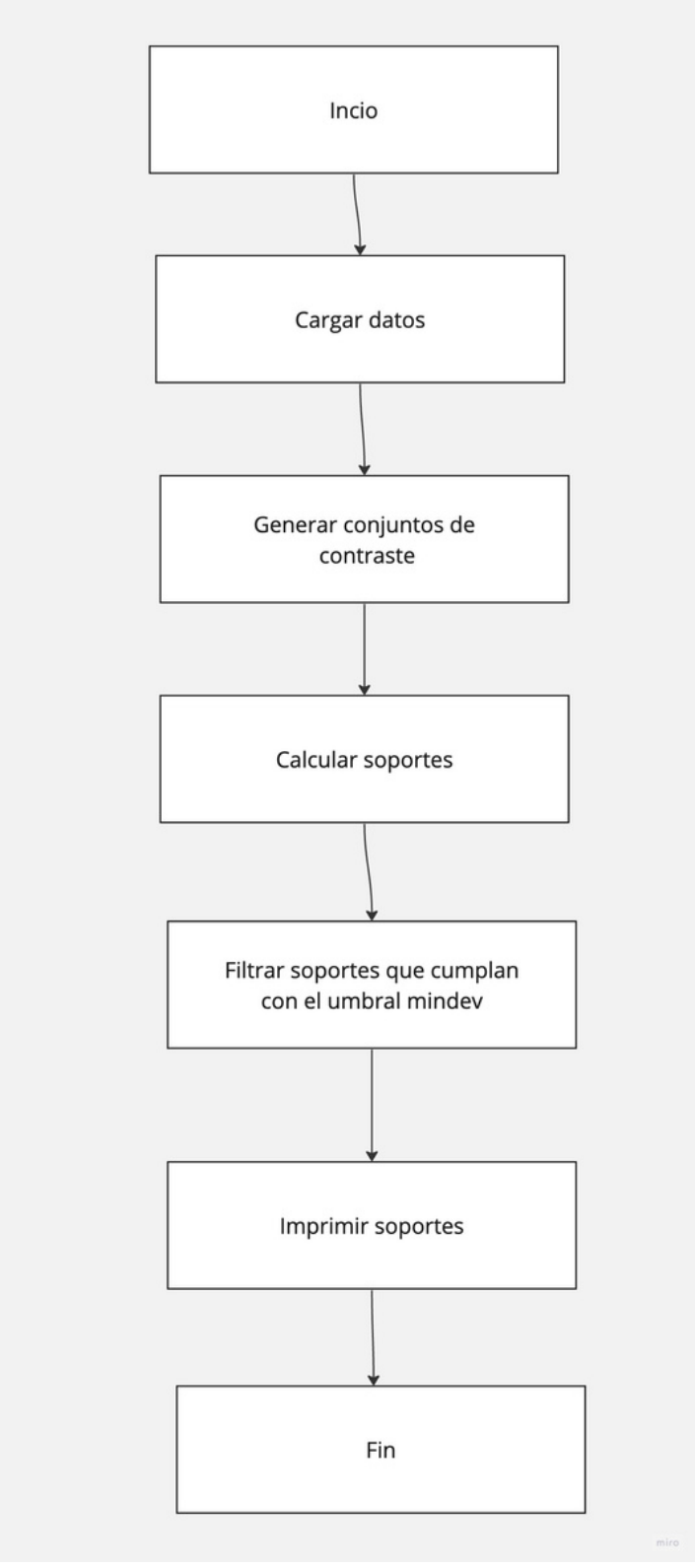


Tabla de resultados

Contrast-Set	Observed % Ph.D.	Observed % Bach.	Expected % Ph.D.	Expected % Bach.	χ^2	p
workclass = State-gov	21.0	5.4			225.1	6.9e-51
occupation = sales	2.7	15.8			74.9	4.8e-18
hour per week > 60	8.4	3.2			43.4	4.4e-11
native country = U.S.	80.5	89.5			45.9	1.3e-11
native country = Canada	1.9	0.5			18.6	1.6e-5
native country = India	1.6	0.5			15.2	9.5e-5
salary > 50K	72.6	41.3			220.2	8.3e-50
sex = male ∧ salary > 50K	61.8	34.8	58.8	28.5	173.6	1.2e-39
occupation = prof-specialty ∧ sex = female ∧ salary > 50K	7.6	2.6	10.7	3.5	48.2	3.8e-12

Diagrama de caja



Resumen de ejecuciones

Se logro ejecutar el codigo al punto de generar los conjuntos de contraste y se calculo el soporte para cada clase de la siguiente manera:

```
{1: {'Manager': 0.03,
  'Data Scientist': 0.14,
  'Sales Associate': 0.07,
  'Software Engineer': 0.19,
  'Graphic Designer': 0.12,
  'Driver': 0.05,
  'IT': 0.07,
  'Warehouse Associate': 0.17},
 2: {('Manager', 'Male'): 0.03,
  ('Manager', 'Female'): 0.0,
  ('Financial Analyst', 'Male'): 0.03,
  ('Financial Analyst', 'Female'): 0.05,
  ('Data Scientist', 'Male'): 0.05,
  ('Sales Associate', 'Male'): 0.04,
  ('Sales Associate', 'Female'): 0.03,
  ('Software Engineer', 'Male'): 0.18,
  ('Software Engineer', 'Female'): 0.01,
  ('Graphic Designer', 'Male'): 0.06,
  ('Graphic Designer', 'Female'): 0.06,
  ('Driver', 'Male'): 0.01,
  ('Driver', 'Female'): 0.04,
  ('Marketing Associate', 'Male'): 0.01,
  ('Marketing Associate', 'Female'): 0.07,
  ('IT', 'Male'): 0.02,
  ('IT', 'Female'): 0.05,
  ('Warehouse Associate', 'Male'): 0.07,
  ('Warehouse Associate', 'Female'): 0.1},
 3: {('Software Engineer', 'Male', 5): 0.17}}
```

	Level	Cand.	Dev.	Surp.
0	1.0	8.0	0.04	0.40
1	2.0	19.0	-0.99	1.15
2	3.0	1.0	0.07	0.07
total	total	28.0	-0.88	1.62