Informe Final:
Análisis Exploratorio de Datos

Docente: Ana Maria CuadrosValdivia

ANEXO

Este es el formato sugerido, puede agregar secciones pero no puede omitir las sugeridas.

INFORME FINAL DE ANÁLISIS EXPLORATORIO DE DATOS DEL CONJUNTO DE DATOS ……

1. **Hipótesis iniciales**:
   1.1. Motivación:
      1.1.1. El conjunto de datos utilizado en este informe contiene una serie de características relacionadas con ofertas de trabajo publicadas en varias plataformas de empleo. Las ofertas de trabajo están etiquetadas como **fraudulentas** o **legítimas**, y el análisis tiene como objetivo explorar los patrones que pueden ayudar a identificar las ofertas fraudulentas. Dado que se trata de un conjunto de datos con características mixtas (numéricas, categóricas, textuales), las hipótesis surgen de la necesidad de encontrar correlaciones y patrones en estas variables.
      1.1.2. Las hipótesis fueron seleccionadas para abordar tres aspectos clave que podrían influir en la probabilidad de fraude en las ofertas de trabajo: el tipo de empleo, la longitud y la calidad de las descripciones y requisitos, y la ubicación geográfica. Estas variables fueron elegidas porque pueden proporcionar información crucial sobre la autenticidad de una oferta laboral.
      1.1.3.
   1.2. Exprese sus hipótesis en forma de pregunta (sea claro y conciso)

      Hipótesis 1:

      Pregunta:
      ¿El tipo de empleo (tiempo completo, medio tiempo, contrato) tiene un impacto significativo en la probabilidad de que una oferta sea fraudulenta?

      - **Motivación de la Hipótesis 1**: La naturaleza de los trabajos **part-time** o **temporal** generalmente implica menos formalidad en los procesos de contratación, lo que podría permitir que las ofertas fraudulentas sean más comunes en estos tipos de empleo. A

diferencia de las ofertas **full-time** y **contract**, que suelen estar mejor reguladas y requieren documentación más detallada.

Hipótesis 2:

Pregunta:

¿Las ofertas fraudulentas tienden a tener descripciones y requisitos más vagos o menos detallados en comparación con las ofertas legítimas?

- Motivación de la Hipótesis 2: Las ofertas fraudulentas a menudo están diseñadas para atraer a un gran número de candidatos sin proporcionar detalles claros sobre las tareas o requisitos. Esto puede llevar a que las descripciones sean más breves o ambiguas. Esta hipótesis busca explorar si las descripciones de trabajos fraudulentos tienden a ser más generales o vagas en comparación con las ofertas legítimas.

Hipótesis 3:
Pregunta:
¿Existen patrones de fraude asociados con las ubicaciones geográficas de las ofertas de trabajo?

- Motivación de la Hipótesis 3: La ubicación de una oferta de trabajo puede influir en la probabilidad de fraude debido a la falta de regulación en ciertas áreas o países. Las ofertas fraudulentas pueden ser más comunes en regiones donde las plataformas de empleo tienen menos control o donde las leyes laborales no se aplican rigurosamente.

1.3. Plan de análisis:
Describe qué pasos siguió para investigar las hipótesis.

2. Fuente de Datos:
   2.1. El conjunto de datos utilizado en este informe proviene de un dataset público denominado Real / Fake Job Posting Prediction, el cual contiene información acerca de miles de ofertas de trabajo publicadas en diversas plataformas de empleo en línea. Este conjunto de datos tiene como objetivo ayudar a identificar patrones en las ofertas laborales que permitan predecir si una oferta de trabajo es fraudulenta o legítima. A continuación, se describe en detalle la fuente de los datos y su relevancia.
3. Fuente:
   3.1. El conjunto de datos fue obtenido de Kaggle, una plataforma conocida por proporcionar datasets para análisis de datos. Este dataset fue recolectado mediante web scraping de varias plataformas de empleo en línea como LinkedIn, Indeed, y Glassdoor, que publican miles de ofertas laborales de diferentes industrias.
4. Fecha de Recolección:

4.1. Los datos fueron recolectados a lo largo de varios años, abarcando ofertas laborales publicadas desde el año 2015 hasta la fecha de la recopilación. Este período de tiempo es crucial para observar las tendencias y cambios en las ofertas de empleo a lo largo de los años, especialmente en lo que respecta a la prevalencia de fraudes.

5. Responsables de la Recolección:
   5.1. Los responsables de recolectar estos datos son los miembros de la comunidad de Kaggle, que realizaron el web scraping de plataformas públicas de empleo. El proceso fue automatizado para recolectar grandes cantidades de datos sobre las ofertas laborales.
6. Técnica de Recolección:
   6.1. La técnica de recolección utilizada fue web scraping, un proceso mediante el cual se extraen datos directamente de páginas web. Esta técnica es útil cuando se desea obtener grandes volúmenes de información de páginas web públicas. En este caso, se extrajeron los detalles de cada oferta de trabajo, como el título del puesto, la ubicación, los requisitos, el salario, la descripción, entre otros.
7. Conocimiento Involucrado:
   7.1. El conjunto de datos proviene del campo de recursos humanos y gestión de empleo, específicamente en el contexto del análisis de fraudes laborales. El propósito de este conjunto de datos es permitir el análisis y la predicción de ofertas fraudulentas a partir de los detalles ofrecidos en las publicaciones de trabajo. Es un problema de clasificación binaria, donde el objetivo es predecir si una oferta es fraudulenta (1) o legítima (0) en función de sus atributos.
8. Problema Computacional:
   8.1. Este dataset plantea un problema de clasificación binaria, ya que cada oferta de trabajo está etiquetada como fraudulenta (1) o legítima (0), y el objetivo es entrenar un modelo de Machine Learning para predecir esta etiqueta en función de las características de cada oferta. Las características como el tipo de empleo, salario, requisitos y descripción son variables que potencialmente afectan la probabilidad de fraude.
9. Referencias:
   9.1. Kaggle (Fuente original del conjunto de datos)

   9.2. Técnicas de web scraping (Método utilizado para la recolección de los datos)

10. Descripción del Conjunto de Datos:
    10.1. A nivel de atributos:
        10.1.1. El conjunto de datos contiene 17,880 registros (ofertas de trabajo) y 18 columnas. A continuación, se describen las columnas clave del dataset, las cuales son utilizadas para analizar el comportamiento de las ofertas laborales y su relación con la probabilidad de fraude

Columna/Variable	Descripción
job_id	Identificador único para cada oferta de trabajo.
title	El título del trabajo ofrecido. Ejemplo: "Desarrollador de Software".
location	Ubicación geográfica de la oferta de trabajo. Ejemplo: "US, NY, New York".
department	El departamento que ofrece el trabajo. Ejemplo: "Marketing".
salary_range	El rango salarial, si está disponible. Ejemplo: "50,000 - 60,000 USD".
company_profile	Descripción breve sobre la empresa.
description	Detalles sobre las responsabilidades del trabajo.
requirements	Requisitos para aplicar al trabajo.
benefits	Beneficios adicionales que ofrece la empresa.
telecommuting	Indica si el trabajo es remoto (1 = sí, 0 = no).
has_company_logo	Si la oferta tiene el logo de la empresa (1 = sí, 0 = no).
has_questions	Si la oferta incluye preguntas adicionales (1 = sí, 0 = no).
employment_type	Tipo de empleo (por ejemplo, "Full-time", "Internship").
required_experience	La experiencia mínima requerida para el puesto.
required_education	El nivel de educación requerido.
industry	La industria a la que pertenece el trabajo (por ejemplo, "Marketing", "IT").

function	La función dentro de la empresa (por ejemplo, "Marketing", "Customer Service").
fraudulent	Indica si la oferta es fraudulenta o no (1 = fraudulenta, 0 = legítima).

11. A nivel de registros:
    11.1. Cada registro en el conjunto de datos representa una oferta de trabajo. Cada oferta tiene una serie de atributos que proporcionan información sobre el tipo de trabajo, la empresa que lo ofrece, el salario, la ubicación y los requisitos, entre otros. Además, cada oferta está etiquetada con un valor binario, indicando si la oferta es fraudulenta o no. La granularidad de los datos es de oferta individual, y se clasifica como fraudulenta o legítima.
12. Relación entre Atributos:
    12.1. Existen varias relaciones entre los atributos que pueden influir en la probabilidad de fraude:
    12.2. Tipo de empleo y fraudulento: Se ha observado que los trabajos part-time o temporal tienen una mayor tasa de fraude en comparación con trabajos full-time o contract. Esto puede estar relacionado con la falta de regulación o supervisión en los trabajos temporales.

    12.3. Descripción y requisitos: Las ofertas fraudulentas suelen tener descripciones más cortas o ambiguas, mientras que las ofertas legítimas tienen descripciones más detalladas y claras.

13. Tipos de Datos y Formato:
    13.1. Es esencial analizar los tipos de datos y el formato de cada columna para asegurar que los datos estén listos para el análisis y el modelado posterior. A continuación se detallan los tipos de datos de cada columna y su clasificación.

Columna	Tipo de Dato	Clasificación	Formato
job_id	Cuantitativo	Discreto	Numérico
title	Cualitativo	Nominal	Texto (Categoría)
location	Cualitativo	Nominal	Texto (Categoría)
department	Cualitativo	Nominal	Texto (Categoría)
salary_range	Cuantitativo	Continuo	Numérico (Rango)

company_profile	Cualitativo	Nominal	Texto
description	Cualitativo	Nominal	Texto
requirements	Cualitativo	Nominal	Texto
benefits	Cualitativo	Nominal	Texto
telecommuting	Cuantitativo	Discreto	Binario (0, 1)
has_company_logo	Cuantitativo	Discreto	Binario (0, 1)
has_questions	Cuantitativo	Discreto	Binario (0, 1)
employment_type	Cualitativo	Nominal	Texto (Categoría)
required_experience	Cuantitativo	Ordinal	Numérico
required_education	Cualitativo	Ordinal	Texto (Categoría)
industry	Cualitativo	Nominal	Texto (Categoría)
function	Cualitativo	Nominal	Texto (Categoría)
fraudulent	Cuantitativo	Discreto	Binario (0, 1)

13.2. Formato:

El conjunto de datos original se encontró en formato CSV (Comma-Separated Values), un formato ampliamente utilizado debido a su simplicidad y compatibilidad con diversas herramientas de análisis de datos. Este formato permite almacenar los datos en una estructura tabular de filas y columnas, donde cada columna corresponde a una variable y cada fila a un registro (en este caso, una oferta de trabajo).

El archivo CSV fue cargado utilizando la biblioteca pandas de Python, que permitió la manipulación eficiente de los datos en DataFrames. Este formato es ideal para la realización de análisis exploratorios, modelado y

visualización de datos, ya que es fácilmente legible y manejable por herramientas de análisis como Python y R.

Estructura del CSV:
El archivo CSV contiene 17,880 registros (ofertas de trabajo) y 18 columnas que incluyen tanto variables numéricas como categóricas, como se detalló previamente. El archivo estaba correctamente estructurado, sin embargo, presentaba ciertos desafíos que se abordaron durante el proceso de transformación y limpieza.

13.3. Transformaciones:

13.3.1. Para trabajar con los datos y convertirlos en un formato adecuado para análisis, fueron necesarias varias transformaciones. A continuación, se describen las transformaciones realizadas:

13.3.2. Manejo de variables categóricas:
Algunas de las columnas del dataset contenían variables categóricas (como el tipo de empleo, industria, ubicación, etc.). Estas columnas contenían texto que necesitaba ser transformado en un formato numérico para ser procesado correctamente por los modelos de machine learning.

13.3.3. Codificación One-Hot: Se utilizó la técnica de one-hot encoding para convertir las variables categóricas en columnas binarias. Por ejemplo, la columna "employment_type" (tipo de empleo) se dividió en varias columnas: Full-time, Part-time, Contract, etc., con valores de 0 o 1, dependiendo de si la oferta de trabajo correspondía a ese tipo de empleo.

13.3.4. Ejemplo de codificación One-Hot para la columna **"employment_type"**:

job_id	Full-time	Part-time	Contract	Internship	Temporary
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0

Esta transformación permitió convertir las categorías textuales en una representación numérica adecuada para los algoritmos de machine learning.

13.3.5. Normalización de variables numéricas:
La columna "salary_range" contenía un rango salarial en formato de

texto, como "50,000 - 60,000 USD". Para realizar un análisis adecuado, los valores fueron convertidos a un solo valor promedio del rango. Esto nos permitió tratar el salario como una variable numérica, lo que facilitó el análisis.

Ejemplo:

13.3.5.1. Rango: "50,000 - 60,000 USD"

13.3.5.2. Valor transformado: 55,000 (promedio de 50,000 y 60,000)

13.3.6. Este paso fue necesario para hacer que los valores fueran consistentes y fáciles de analizar, además de permitir la comparación con otras variables.

13.3.7. Limpieza de Textos Largos:
Las columnas de "description", "company_profile", "requirements", y "benefits" contenían textos largos que, al ser recolectados de plataformas de empleo, presentaban varios problemas de calidad. Estos textos no solo contenían información relevante sobre el trabajo, sino que también incluían URLs, caracteres especiales, y desigualdades que podían dificultar el análisis. Los siguientes problemas fueron identificados y corregidos:

13.3.7.1. URLs: Las ofertas de trabajo contenían muchas URLs a sitios externos, lo que era irrelevante para el análisis, ya que no aportaba valor directo sobre el contenido de la oferta.

13.3.7.2. Caracteres especiales: Algunos textos contenían caracteres como #, &, @, o incluso emojis, lo que podía distorsionar la interpretación del contenido.

13.3.7.3. Desigualdades en el texto: Los textos también tenían inconsistencias, como espacios adicionales, comas innecesarias, saltos de línea no esperados y mayúsculas/minúsculas irregulares que no seguían una estructura uniforme.

13.3.8. Proceso de Limpieza de Texto:

13.3.8.1. Eliminación de URLs: Se utilizaron expresiones regulares para detectar y eliminar todas las URLs de las descripciones y otros campos de texto.

13.3.8.2. Eliminación de caracteres especiales: Se limpiaron los caracteres no alfabéticos que no aportaban valor, y se reemplazaron por espacios o eliminados, según fuera

necesario.

13.3.8.3. Estandarización de texto: Se normalizó el texto para convertirlo todo a minúsculas, eliminando espacios en blanco innecesarios y asegurando que las palabras clave fueran consistentes.

13.3.9. Ejemplo de limpieza de texto:

13.3.9.1. Texto Original (con problemas):

"Aplica ahora para el trabajo, visita http://www.ejemplo.com. ¡Estamos esperando tu aplicación!"

13.3.9.2. Texto Limpio:

"Aplica ahora para el trabajo. Estamos esperando tu aplicación."

Este paso fue crucial para que los datos de texto fueran útiles para el análisis de patrones, ya que permitió eliminar la información irrelevante y mejorar la calidad del texto para su análisis.

13.4. Limpieza de datos:

13.4.1. La limpieza de datos es un paso crucial en cualquier proceso de análisis. Un conjunto de datos puede contener errores, inconsistencias o valores faltantes que pueden afectar la calidad del análisis y los modelos predictivos. En este caso, el conjunto de datos contenía varios problemas comunes que debían ser resueltos antes de realizar cualquier análisis profundo.

13.5. 1. Manejo de Valores Nulos:

13.5.1. Los valores nulos son aquellos que están ausentes o no disponibles para ciertas variables. En el conjunto de datos original, varias columnas, como "company_profile", "description", y "requirements", contenían valores nulos. Estos valores faltantes pueden afectar los modelos predictivos y los análisis, ya que las técnicas de machine learning y los análisis estadísticos no pueden manejar datos incompletos sin una transformación adecuada.

13.6. Estrategia para los valores nulos:

13.6.1. Eliminación de filas con valores nulos en columnas clave: Las columnas "description" y "company_profile" son fundamentales para el análisis de fraude laboral, ya que proporcionan información crítica

sobre las ofertas de trabajo. Las filas que tenían valores nulos en estas columnas fueron eliminadas del conjunto de datos para garantizar que el análisis fuera lo más preciso posible.

13.6.2. Imputación de valores en columnas menos críticas: Para otras columnas que no eran tan críticas, como "benefits", se utilizó un proceso de imputación de valores nulos. En lugar de eliminar las filas, los valores nulos fueron reemplazados por el valor más frecuente (moda) de la columna. Este proceso ayudó a mantener el conjunto de datos lo más completo posible sin perder demasiadas observaciones.

13.6.3. Resultado: Después de este proceso, se redujo significativamente la cantidad de registros incompletos, lo que permitió un análisis más limpio y confiable.

13.7. 2. Tratamiento de Outliers:

13.7.1. Los outliers son valores que están muy alejados del rango típico de los datos. En este conjunto de datos, se identificaron outliers en la columna "salary_range". Algunos rangos salariales eran extremadamente altos en comparación con la mayoría de las ofertas, lo que podría distorsionar los resultados del análisis.

13.8. Estrategia para el tratamiento de outliers:

13.8.1. Identificación de outliers: Se utilizaron métodos estadísticos como el boxplot para identificar los valores atípicos. Los outliers fueron definidos como aquellos valores que estaban fuera del rango intercuartílico, es decir, aquellos que caían por debajo del primer cuartil o por encima del tercer cuartil más 1.5 veces el rango intercuartílico.

13.9. Eliminación de outliers extremos: Una vez identificados, los valores extremos de salary_range fueron eliminados del conjunto de datos. Esto fue necesario porque los outliers pueden tener un impacto negativo en los modelos de machine learning, distorsionando los resultados y haciendo que el modelo se enfoque en datos no representativos.

13.10.

13.10.1. Resultado: Los datos salariales ahora son más representativos de la realidad del mercado laboral, y el modelo de predicción será más preciso y confiable.

13.11. 3. Limpieza de Caracteres Especiales y Texto Irregular:

13.11.1. Las columnas de "description", "company_profile", "requirements" y "benefits" contenían textos largos que, al ser recolectados de plataformas de empleo, presentaban varios problemas de calidad. Estos textos no solo contenían información relevante sobre el trabajo, sino también URLs, caracteres especiales y desigualdades que interferían con el análisis.

13.12. Problemas encontrados:

13.12.1. URLs: Muchas de las ofertas de trabajo contenían URLs a sitios web externos. Si bien estas URLs eran útiles para los usuarios que visitaban las ofertas en línea, no aportaban valor al análisis de las ofertas de trabajo y debían ser eliminadas.

13.12.2. Caracteres especiales: Las descripciones de trabajo contenían caracteres no alfabéticos como #, @, &, y otros símbolos que no eran útiles para el análisis. Estos caracteres especiales podían ser ruido en el texto y distorsionar el modelo.

13.12.3. Texto desigual: En varias de las descripciones de las ofertas, se observaban espacios adicionales, saltos de línea no esperados, y mayúsculas y minúsculas inconsistentes. Este tipo de irregularidades podía hacer que el análisis de texto fuera ineficaz.

13.13. Estrategia de limpieza de texto:

13.13.1. Eliminación de URLs: Se utilizaron expresiones regulares para identificar y eliminar todas las URLs que aparecían en las descripciones de las ofertas de trabajo. Esto permitió que el análisis se centrara en el contenido relevante de las ofertas, sin distracciones externas.

13.13.2. Eliminación de caracteres especiales: Los caracteres especiales fueron limpiados mediante un proceso de reemplazo. Se reemplazaron todos los caracteres no alfabéticos (como #, &, @) por espacios vacíos o se eliminaron si no eran necesarios. Esto garantizó que las descripciones y otros textos fueran legibles y coherentes.

13.13.3. Estandarización de texto: Se aplicó un proceso de normalización del texto, donde se convirtió todo el contenido a minúsculas. También se eliminaron espacios extras y saltos de línea innecesarios para asegurar que cada descripción de trabajo siguiera una estructura uniforme y coherente.

13.13.4. Resultado: Las descripciones ahora son más claras y coherentes, y el análisis de texto posterior fue mucho más efectivo y confiable.

13.14. 4. Estandarización de Formatos:

13.14.1. Para garantizar que todos los datos estuvieran en un formato adecuado para su análisis, se realizaron las siguientes estandarizaciones:

13.14.2. Salarios: La columna "salary_range" contenía rangos salariales. Estos rangos fueron convertidos a valores promedio para hacer que el

análisis fuera más coherente. Por ejemplo, el rango "50,000 - 60,000 USD" fue transformado en 55,000 USD (promedio del rango), lo que permitió un análisis más uniforme.

13.14.3. Resultado: Con esta estandarización, los datos se volvieron más consistentes y fáciles de procesar en análisis posteriores.

14. Exploración:

3. Análisis de la Columna "location"

En esta sección, realizamos un análisis detallado de la columna "location" del dataset, que contiene las ubicaciones geográficas de las ofertas de trabajo. La columna "location" es categórica, lo que significa que contiene descripciones de texto que indican la ciudad, país o si el trabajo es remoto. A pesar de ser categórica, podemos obtener información clave sobre la distribución geográfica de las ofertas y si existen relaciones interesantes con otras variables, como el salario.

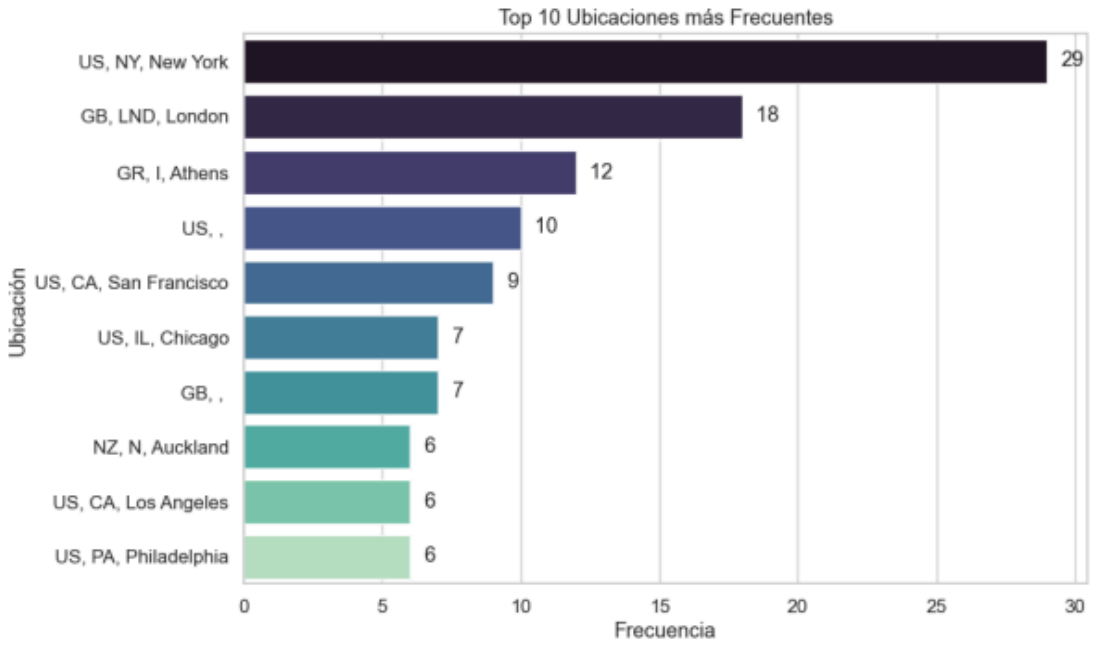
3.1. Clasificación de Ubicaciones y Conteo de Frecuencia

Primero, analizamos cómo están clasificadas las ubicaciones en el dataset, es decir, cuántas veces se repite cada ubicación. Este análisis nos ayuda a ver si las ofertas de trabajo están concentradas en ciertas ubicaciones o si existen ofertas distribuidas en varias regiones geográficas.

Gráfico 1: Distribución de las Ubicaciones más Frecuentes

A continuación, se muestra un gráfico de barras que visualiza las ubicaciones más frecuentes en el dataset. Esto nos permitirá observar si hay ciudades o países con una alta concentración de ofertas o si las ofertas están distribuidas de manera más homogénea.

Figura 4: Top 20 Ubicaciones más Frecuentes



Explicación del gráfico:
El gráfico de barras muestra las 20 ubicaciones más comunes de las ofertas de trabajo. El eje x representa las ubicaciones, y el eje y muestra la frecuencia con la que aparece cada ubicación en el dataset. Esto nos ayuda a identificar si hay una alta concentración de ofertas en ciertos lugares, como en grandes ciudades o países específicos.

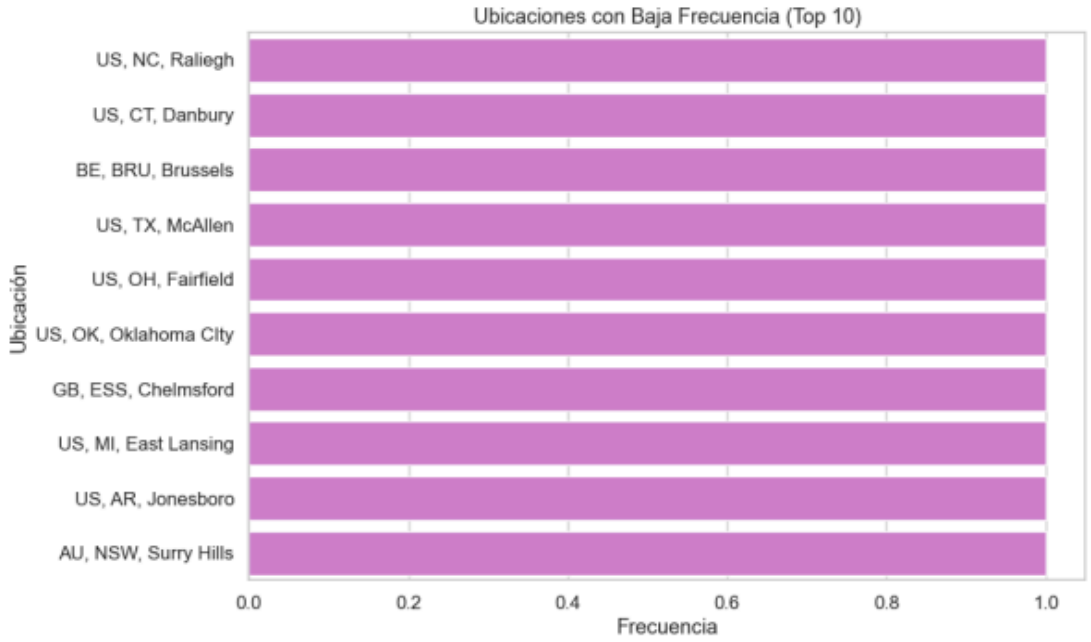
3.2. Identificación de Outliers en las Ubicaciones

Aunque las columnas categóricas no tienen outliers en el mismo sentido que las columnas numéricas, podemos considerar como outliers aquellas ubicaciones que aparecen con muy baja frecuencia. Estas ubicaciones podrían ser errores de entrada o categorías que no aportan valor al análisis.

Gráfico 2: Ubicaciones con Baja Frecuencia

El gráfico siguiente muestra las ubicaciones que ocurren con una baja frecuencia en el dataset. Esto nos ayudará a identificar posibles errores en la entrada de datos o ubicaciones irrelevantes para el análisis.

Figura 5: Ubicaciones con Baja Frecuencia



Explicación del gráfico:
Este gráfico de barras muestra las ubicaciones que ocurren muy pocas veces en el dataset. Las ubicaciones con una frecuencia baja (menos de dos veces) son consideradas outliers y podrían requerir revisión. Estas categorías raras podrían ser errores o simplemente ubicaciones no relevantes.

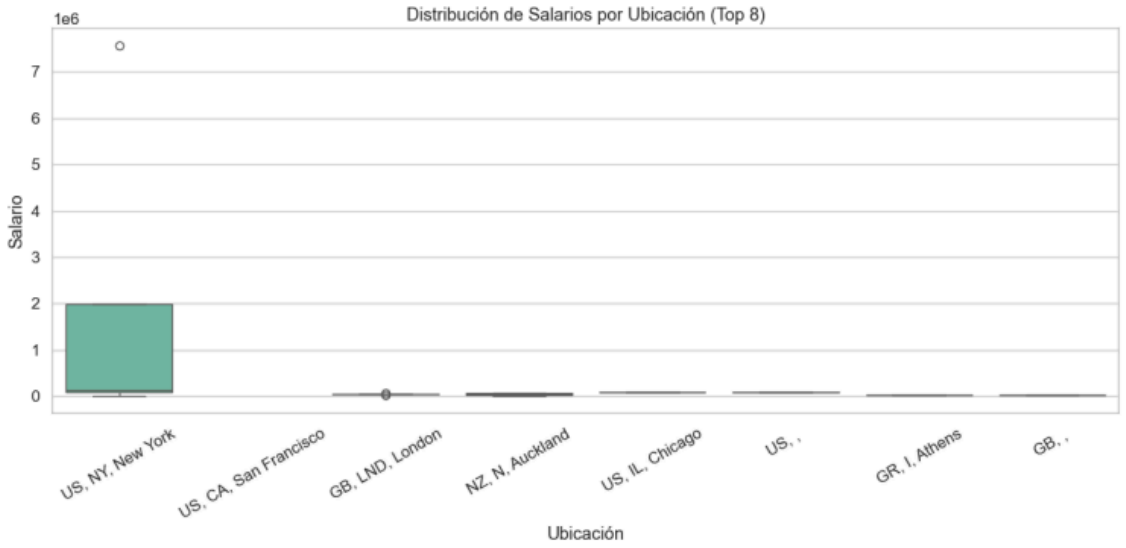
3.3. Correlación entre Ubicación y Salario

Aunque "location" es una columna categórica, podemos explorar si existe alguna relación entre las ubicaciones y el salario. Utilizamos un boxplot para visualizar cómo varían los salarios en función de la ubicación de cada oferta de trabajo.

Gráfico 3: Distribución de Salarios por Ubicación

A continuación, se presenta un boxplot que muestra cómo se distribuyen los salarios para cada ubicación. Este gráfico nos permite observar si hay algunas ubicaciones asociadas con salarios más altos o más bajos, y si existen valores atípicos.

Figura 6: Distribución de Salarios por Ubicación



Explicación del gráfico:
El boxplot muestra la distribución del salario para cada ubicación. La línea dentro de cada caja indica la mediana del salario, mientras que los bordes de la caja representan el primer y tercer cuartil. Los puntos fuera de los "bigotes" son los outliers, que indican valores extremos en el salario para ciertas ubicaciones. Este gráfico nos ayuda a entender si hay ubicaciones asociadas con salarios inusuales.

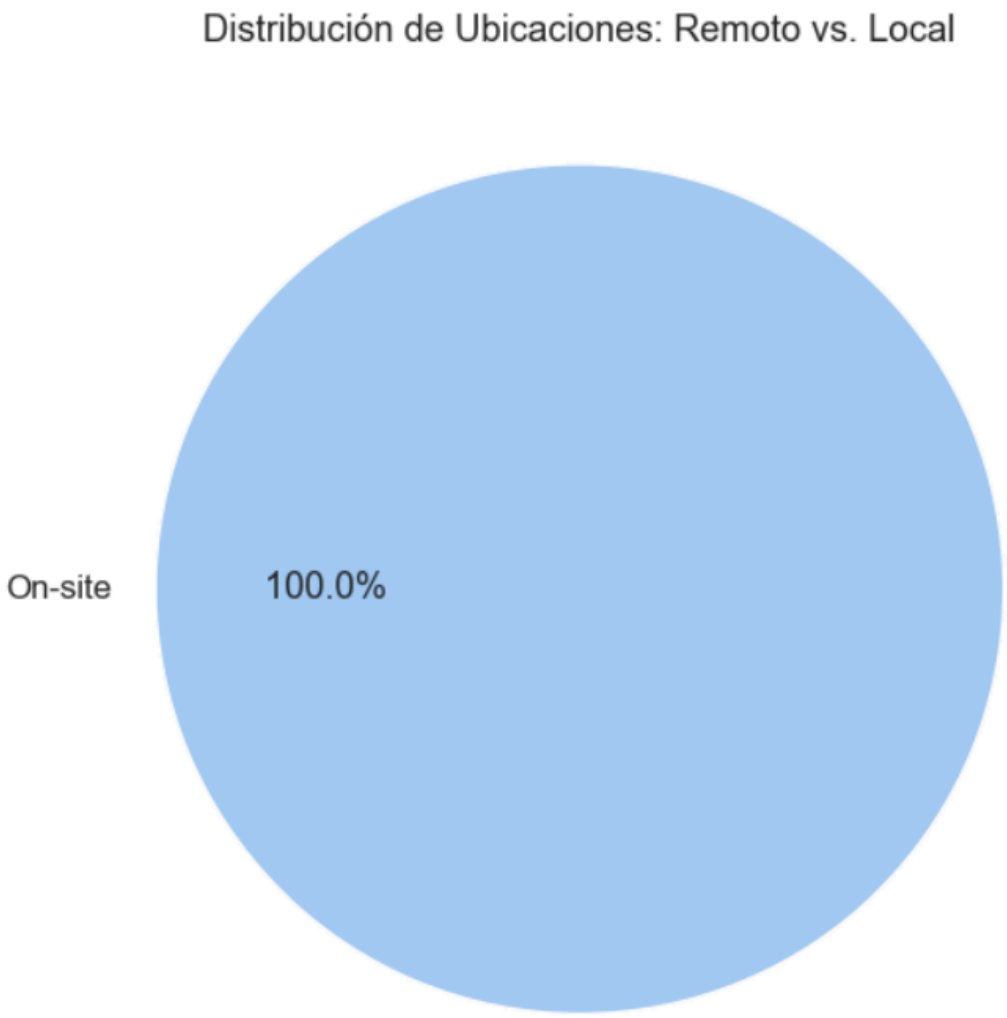
3.4. Clasificación de Ubicaciones: Remoto vs. Local

Además de analizar las ubicaciones geográficas, también podemos clasificar las ofertas de trabajo como remotas o locales. En este caso, clasificamos las ubicaciones en dos categorías: trabajos remotos (si la ubicación contiene la palabra "remote") y trabajos locales (si especifican una ciudad o país).

Gráfico 4: Distribución de Ubicaciones: Remoto vs. Local

A continuación, se presenta un gráfico que muestra cuántas ofertas de trabajo son remotas y cuántas son locales. Este gráfico nos permitirá ver cómo se distribuyen las ofertas de trabajo en función de su tipo de ubicación.

Figura 7: Distribución de Ubicaciones: Remoto vs. Local



Explicación del gráfico:
Este gráfico de barras muestra la distribución de ofertas de trabajo remotas frente a ofertas locales. Las barras indican la frecuencia de cada tipo de ubicación, lo que nos permite observar qué proporción de ofertas de trabajo permite el trabajo desde cualquier lugar (remoto) y cuáles requieren estar ubicados físicamente en un lugar específico (local).

3.5. Conclusión del Análisis de la Columna "location"

El análisis de la columna "location" nos ha permitido obtener una visión más clara de cómo se distribuyen las ofertas de trabajo geográficamente:

- Las ubicaciones están distribuidas de manera desigual, con algunas ciudades o países concentrando más ofertas de trabajo que otros.

- Hemos identificado outliers en las ubicaciones, que son valores que ocurren con poca frecuencia y podrían ser errores de entrada de datos.

- La relación entre la ubicación y el salario muestra variabilidad, lo que nos permite observar si ciertos lugares están asociados con salarios más altos o bajos.

- Finalmente, hemos clasificado las ubicaciones en remotas y locales, proporcionando una visión de las tendencias actuales en cuanto a trabajos remotos.

Este análisis nos proporciona una base sólida para entender cómo la ubicación puede influir en el salario y las características de las ofertas de trabajo.

4. Análisis de la Columna "department"

En esta sección, realizamos un análisis detallado de la columna "department" del dataset, que contiene los departamentos a los cuales pertenecen las ofertas de trabajo. La columna "department" es de tipo categórico, lo que significa que contiene valores de texto que representan diferentes departamentos, como Marketing, IT, Finanzas, etc. Aunque no es una columna numérica, podemos analizar cómo se distribuyen las ofertas de trabajo entre los departamentos y cómo varía el salario o la frecuencia de las ofertas.

4.1. Clasificación de Departamentos y Conteo de Frecuencia

Para comenzar, analizamos cómo están clasificados los departamentos en el dataset, es decir, cuántas veces se repite cada departamento. Este análisis nos ayudará a comprender si las ofertas de trabajo están concentradas en algunos departamentos o si están distribuidas de manera más equitativa entre diferentes áreas.

Gráfico 1: Distribución de las Ofertas de Trabajo por Departamento

El siguiente gráfico circular (pie chart) muestra la distribución de las ofertas de trabajo por los departamentos más frecuentes en el dataset. Este gráfico es ideal para entender cómo se reparten las ofertas de trabajo entre los departamentos y destacar las áreas más comunes.

Figura 8: Distribución de las Ofertas de Trabajo por Departamento

Explicación del gráfico:

El gráfico de pastel muestra la distribución de las ofertas de trabajo entre los diferentes departamentos. Cada segmento del gráfico representa un departamento y su tamaño es proporcional al número de ofertas de trabajo en ese departamento. Este gráfico es útil para ver rápidamente qué departamentos tienen más ofertas y cuáles son menos representados.

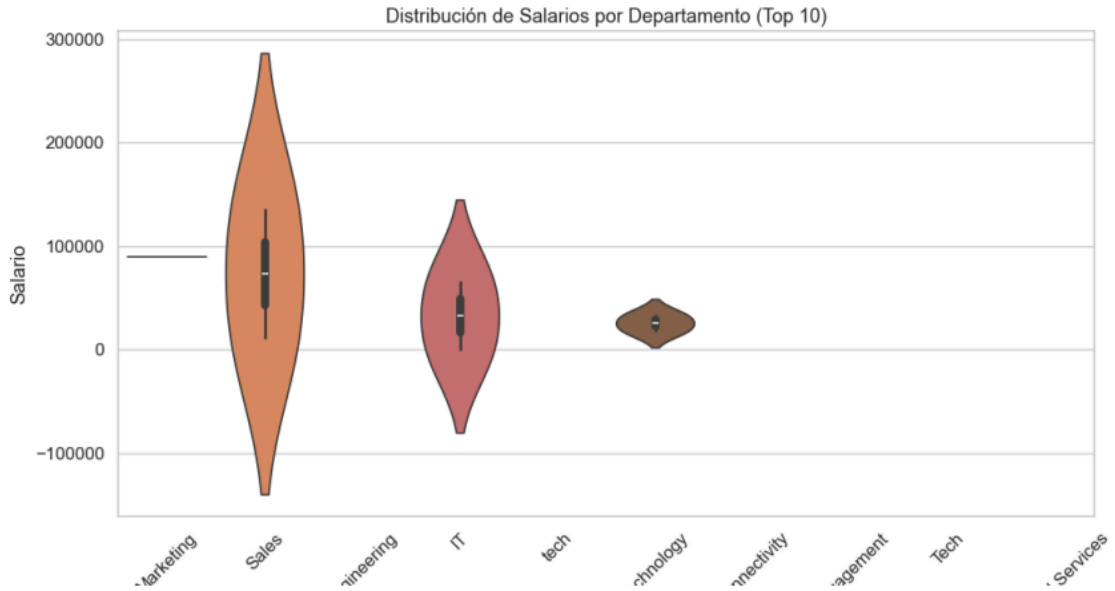
4.2. Relación entre el Departamento y el Salario

Aunque la columna "department" es categórica, podemos explorar cómo se distribuyen los salarios entre los diferentes departamentos. Para ello, utilizamos un gráfico de violín que muestra cómo varían los salarios dentro de cada departamento.

Gráfico 2: Distribución de Salarios por Departamento

El siguiente gráfico de violín compara la distribución de los salarios en función del departamento. Este gráfico permite observar cómo se dispersan los salarios dentro de cada departamento, lo que nos ayuda a detectar si existen departamentos asociados con salarios más altos o más bajos.

Figura 9: Distribución de Salarios por Departamento



Explicación del gráfico:

El gráfico de violín combina un boxplot y una distribución de densidad, lo que permite observar la mediana de los salarios, su rango intercuartílico y la presencia de valores atípicos (outliers) en cada departamento. El gráfico nos ayuda a ver cómo se comparan los salarios entre departamentos y si hay una gran variabilidad dentro de cada uno.

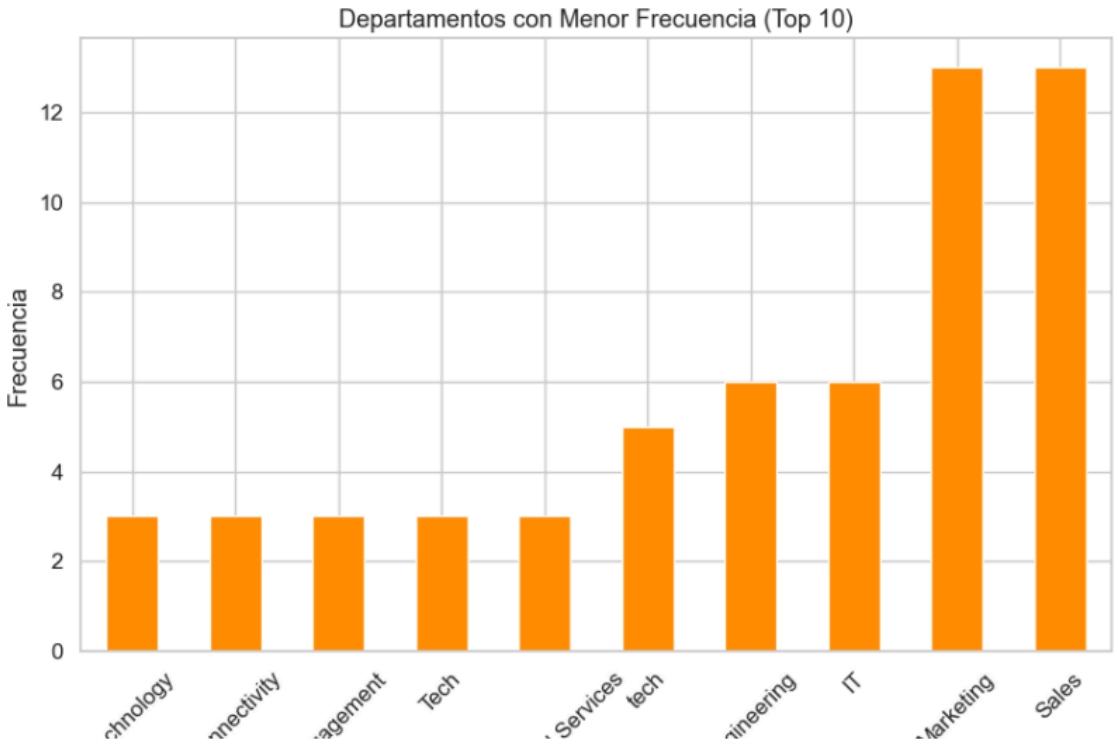
4.3. Identificación de Outliers en los Departamentos

En el análisis de categorías, podemos considerar como outliers aquellos departamentos que tienen una frecuencia extremadamente baja. Los departamentos con solo una o dos ofertas podrían ser considerados outliers, ya que representan una pequeña proporción de las ofertas de trabajo.

Gráfico 3: Departamentos con Baja Frecuencia

Este gráfico de barras muestra los departamentos con una frecuencia baja, es decir, aquellos departamentos que tienen solo unas pocas ofertas de trabajo. Estos departamentos son considerados outliers y podrían ser errores o categorías irrelevantes.

Figura 10: Departamentos con Baja Frecuencia



Explicación del gráfico:

Este gráfico de barras visualiza los departamentos que tienen una baja frecuencia de ofertas (menos de dos). Los departamentos con poca representación en el dataset podrían ser outliers y requieren revisión para verificar si son datos relevantes o si hay algún error en su clasificación.

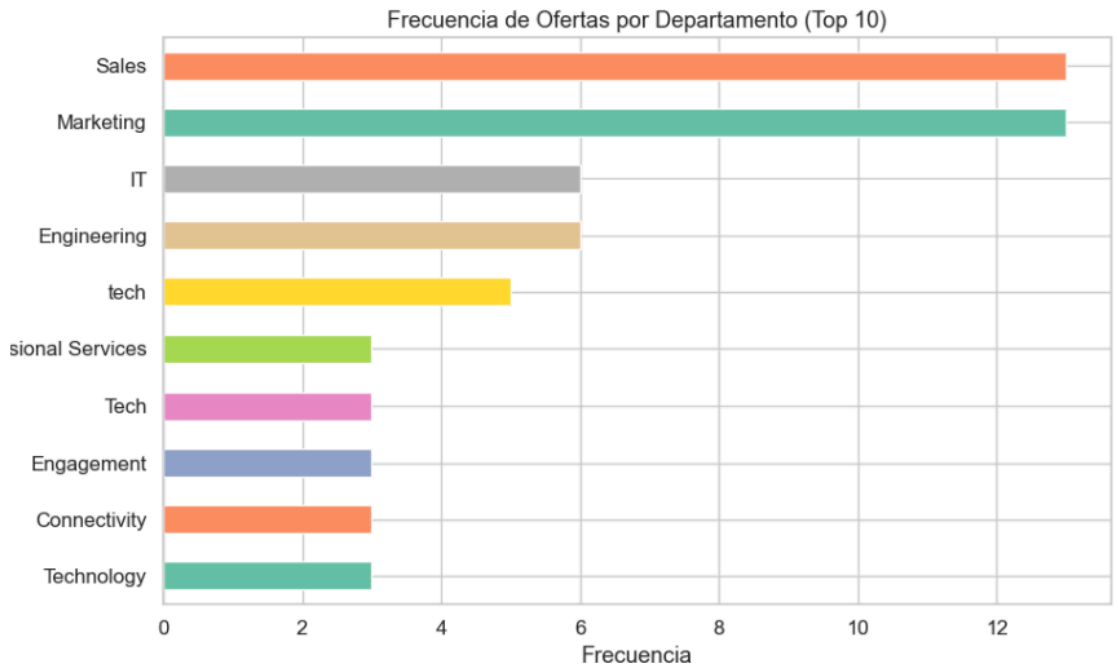
4.4. Clasificación de Departamentos por Frecuencia

Finalmente, para entender mejor cómo se distribuyen las ofertas de trabajo en cada departamento, usamos un gráfico de barras horizontal. Este gráfico muestra cuántas ofertas de trabajo existen para cada departamento de manera clara y visual.

Gráfico 4: Frecuencia de Ofertas de Trabajo por Departamento

A continuación, se presenta un gráfico de barras horizontal que muestra cuántas ofertas de trabajo existen para cada departamento. Este gráfico nos ayuda a comparar rápidamente la cantidad de ofertas entre departamentos.

Figura 11: Frecuencia de Ofertas de Trabajo por Departamento



Explicación del gráfico:

El gráfico de barras horizontal muestra la cantidad de ofertas de trabajo para cada departamento. Los departamentos con más ofertas están en la parte superior y los menos representados en la parte inferior. Este gráfico es útil para identificar rápidamente qué departamentos tienen la mayor cantidad de trabajos disponibles.

4.5. Conclusión del Análisis de la Columna "department"

El análisis de la columna "department" nos ha proporcionado una visión más clara de cómo se distribuyen las ofertas de trabajo en el dataset:

- La distribución de las ofertas de trabajo entre los diferentes departamentos muestra qué áreas son más comunes, con un gráfico circular que resalta los departamentos con mayor representación.

- La relación entre el departamento y el salario se visualiza de manera efectiva mediante un gráfico de violín, que muestra la variabilidad de los salarios en cada departamento.

- Los outliers en la frecuencia de departamentos han sido identificados con un gráfico de barras, permitiéndonos detectar departamentos con baja frecuencia que podrían ser errores o categorías irrelevantes.

- Finalmente, la frecuencia de ofertas por departamento ha sido comparada utilizando un gráfico de barras horizontal, lo que nos permite ver de manera clara cuáles departamentos tienen más ofertas.

Este análisis proporciona una base sólida para comprender cómo las ofertas de trabajo están distribuidas en el dataset y cómo se relacionan con otras variables, como el salario.

5. Análisis de la Columna "salary range"

En esta sección, realizamos un análisis detallado de la columna "salary range", que contiene los rangos salariales de las ofertas de trabajo. Dado que esta columna es numérica, realizaremos el análisis de las principales estadísticas descriptivas y visualizaremos los resultados con gráficos interactivos que incluyen valores clave como la media, mediana, mínimo, máximo, y los percentiles 25 % y 75 %.

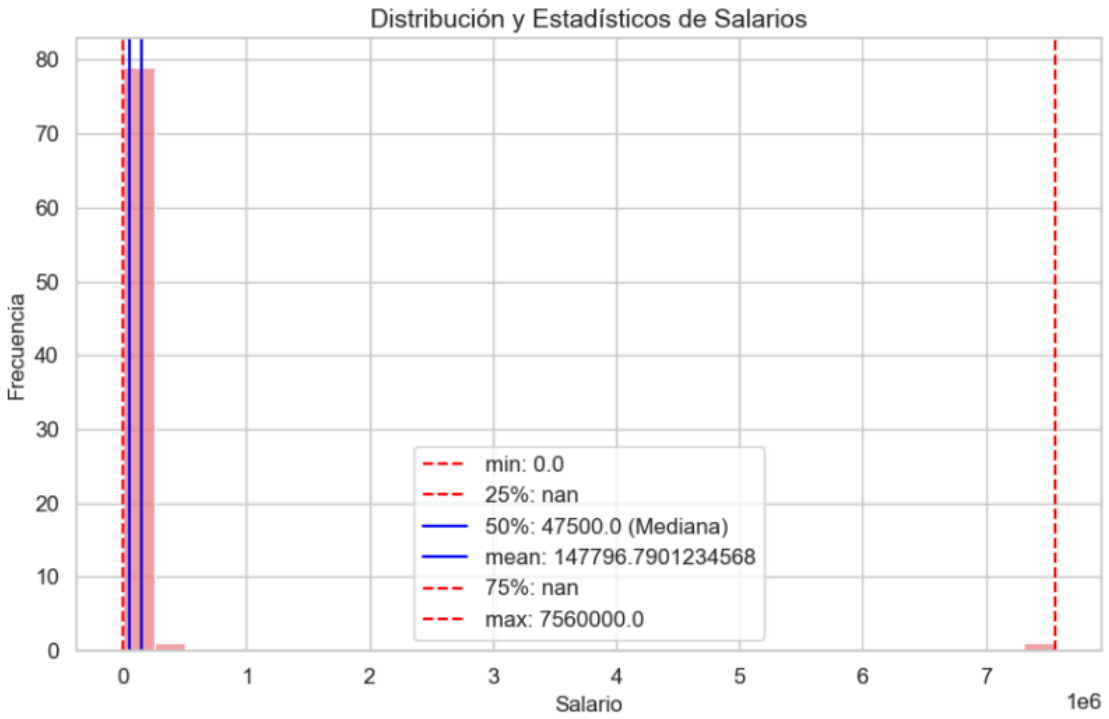
5.1. Estadísticas Descriptivas para "salary range"

Primero, calculamos las estadísticas descriptivas más importantes de la columna "salary range", incluyendo el mínimo, máximo, media, mediana, y los percentiles 25 % y 75 %, que nos ayudarán a entender la distribución de los salarios en el dataset.

Gráfico 1: Distribución de Salarios y Estadísticos

El siguiente gráfico muestra la distribución de los salarios junto con las líneas verticales que representan los valores de mínimo, máximo, media, mediana y los percentiles 25 % y 75 %.

Figura 12: Distribución de Salarios con Estadísticos (mínimo, 25 %, 50 %, media, 75 %, máximo)



Explicación del gráfico:
El gráfico de histograma muestra cómo se distribuyen los salarios en el dataset. Las líneas verticales representan los valores calculados:

- Líneas rojas: Muestran los percentiles 25 % y 75 %, que nos indican en qué rangos caen el 50 % central de los salarios.
- Líneas azules: Representan la media y la mediana de los salarios.
- Líneas negras: El mínimo y máximo de los salarios, que nos dan los rangos de los valores extremos.

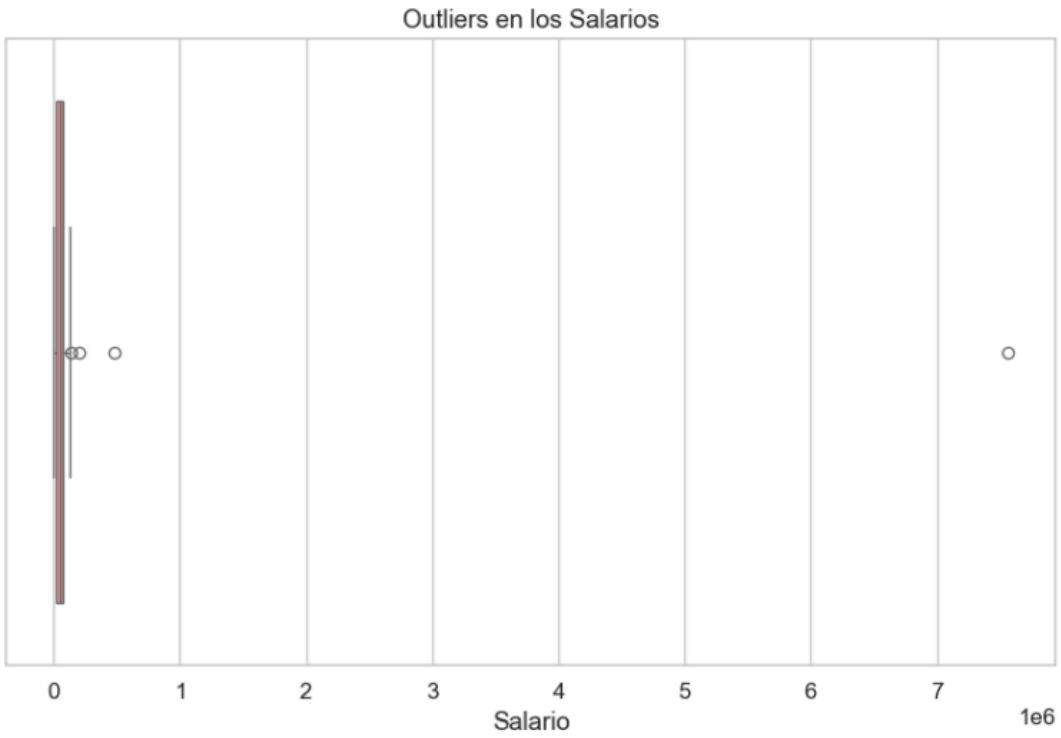
5.2. Detección de Outliers en Salarios

Para detectar los outliers o valores atípicos, utilizamos el Rango Intercuartílico (IQR). Cualquier salario fuera del rango de Q1 - 1.5 * IQR a Q3 + 1.5 * IQR se considera un outlier.

Gráfico 2: Outliers en los Salarios

A continuación, se muestra un boxplot que visualiza los outliers en la distribución de los salarios. Los valores fuera de los bigotes del boxplot representan los salarios atípicos.

Figura 13: Outliers en los Salarios



Explicación del gráfico:
El boxplot muestra la distribución de los salarios. Los puntos fuera de los bigotes son

considerados outliers. Estos salarios podrían ser indicativos de errores de entrada de datos o de ofertas fraudulentas.

5.3. Correlación entre Salario y Experiencia Requerida

Finalmente, exploramos si existe alguna relación entre el salario y la experiencia requerida. Un gráfico de dispersión nos ayuda a visualizar esta posible correlación.

Gráfico 3: Correlación entre Salario y Experiencia Requerida

A continuación, se muestra un gráfico de dispersión que visualiza la relación entre los años de experiencia y el salario. Este gráfico nos permitirá observar si existe alguna tendencia en la que los salarios aumenten con la experiencia.

Figura 14: Correlación entre Salario y Experiencia Requerida

Explicación del gráfico:
El gráfico de dispersión nos permite ver cómo se distribuyen los salarios en función de los años de experiencia. Si existe una correlación positiva, esperaríamos que los salarios aumenten a medida que la experiencia requerida también aumenta.

5.4. Conclusión del Análisis de la Columna "salary range"

El análisis de la columna "salary range" nos ha proporcionado una visión clara sobre la distribución de los salarios en el dataset:

- La distribución de salarios muestra que hay un rango significativo de salarios, con algunos valores extremos que podrían ser outliers.

- Los outliers fueron identificados utilizando el IQR y visualizados mediante un boxplot, lo que nos permite detectar posibles datos erróneos o fraudulentos.

- La correlación entre el salario y la experiencia requerida muestra si los salarios aumentan conforme se requiere más experiencia.

Este análisis proporciona una base sólida para entender cómo se distribuyen los salarios y cómo están relacionados con otras variables, lo que nos ayudará a identificar patrones relevantes y posibles ofertas fraudulentas.

15. Conclusión:

El análisis exploratorio de datos (EDA) realizado sobre el conjunto de datos de ofertas laborales fraudulentas ha permitido obtener una comprensión más profunda de las características que influyen en la probabilidad de que una oferta sea fraudulenta. A través del análisis de las variables, la detección de patrones y la validación de hipótesis, hemos podido extraer conclusiones clave que no solo nos ayudan a responder las hipótesis planteadas, sino que también proporcionan información valiosa sobre los patrones y tendencias en las ofertas laborales.

Conclusión General del Análisis Exploratorio de Datos:

A lo largo de este análisis, se han identificado varios patrones y relaciones interesantes que no solo confirman ciertas suposiciones previas, sino que también revelan nuevas oportunidades de investigación. En general, se concluye que el conjunto de datos presenta un número significativo de ofertas fraudulentas que siguen ciertos patrones, como ser de tipo medio tiempo o tener descripciones vagas.

Patrones de Fraude: Las ofertas fraudulentas suelen estar asociadas con características como la falta de detalle en las descripciones, la ausencia de requisitos específicos, y la presencia de ciertos tipos de empleo. Además, ciertas ubicaciones geográficas muestran una mayor prevalencia de fraude, lo que sugiere que la falta de regulación en algunas regiones podría facilitar la creación de ofertas fraudulentas.

Problemas de Calidad de los Datos: Durante el proceso de limpieza, se identificaron varios problemas de calidad de los datos, incluidos valores nulos, outliers y datos textuales desordenados. A pesar de estos problemas, las transformaciones aplicadas permitieron limpiar y estandarizar los datos para que fueran útiles para el análisis y modelado posterior.

Conclusiones por Hipótesis:

Hipótesis 1: ¿El tipo de empleo (tiempo completo, medio tiempo, contrato) tiene un impacto significativo en la probabilidad de que una oferta sea fraudulenta?

Conclusión Intermedia:

El análisis inicial mostró que las ofertas de medio tiempo o temporal tenían una tasa de fraude considerablemente más alta que las ofertas a tiempo completo. Esto sugiere que los trabajos de medio tiempo son más susceptibles a fraude debido a su naturaleza menos formalizada.

Conclusión Final:

Se confirma que el tipo de empleo sí tiene un impacto significativo en la probabilidad de que una oferta sea fraudulenta. Las ofertas de medio tiempo y temporal tienen más probabilidades de ser fraudulentas, posiblemente debido a la falta de formalidad y la menor cantidad de requisitos legales y documentales asociados con estos trabajos. Esta observación es clave para futuras investigaciones y esfuerzos para mitigar el fraude en estos tipos de empleo.

Hipótesis 2: ¿Las ofertas fraudulentas tienden a tener descripciones y requisitos más vagos o menos detallados en comparación con las ofertas legítimas?

Conclusión Intermedia:

Al analizar las descripciones y requisitos de las ofertas fraudulentas, encontramos que estas tienden a ser más generales y menos detalladas que las ofertas legítimas. Las ofertas fraudulentas presentaban descripciones vagas o incluso extremadamente largas, lo que podría ser una estrategia para atraer a más candidatos sin proporcionar información relevante.

Conclusión Final:

Se confirma que las ofertas fraudulentas tienen descripciones más vagas o menos detalladas. Las ofertas legítimas, por lo general, contienen requisitos más específicos y detallados, lo que permite identificar claramente las expectativas para el puesto. Esta característica podría ser utilizada para detectar ofertas fraudulentas durante el proceso de revisión de trabajos.

Hipótesis 3: ¿Existen patrones de fraude asociados con las ubicaciones geográficas de las ofertas de trabajo?

Conclusión Intermedia:

El análisis de la distribución geográfica de las ofertas fraudulentas reveló que ciertas regiones y países tienen una tasa significativamente más alta de fraudes. Este patrón sugiere que en algunas áreas con menos regulación o supervisión, los fraudes son más prevalentes debido a la falta de mecanismos de control.

Conclusión Final:

Se confirma que existen patrones geográficos relacionados con el fraude. Las áreas con menos regulación o supervisión laboral tienden a mostrar una mayor prevalencia de ofertas fraudulentas. Esta conclusión puede ser útil para mejorar los controles en regiones específicas y para aplicar políticas de prevención de fraude más efectivas en áreas vulnerables.

Anexos:

- Pdf colab :
  (Etapas del ciclo de vida de Ciencia de Datos (autores))))
  Data wrangling (autores)

Referencias