Workshop 001 - ETL

Kevin Artunduaga - 2216155

1) Migración de los datos a la base de datos

Para la migración de los datos de candidates.csv estos fueron trasladados a la base de datos Postgresql, mediante la librería psycopg2, primero se realizó en el script la conexión a la base de datos llamada ETL, en conjunto de el host (use el local), el usuario y la contraseña, estos últimos dos puntos fueron guardados en un db_config.json para mantener estos datos ocultos. Luego de la conexión se creó un cursor para poder recorrer y manipular el conjunto de filas de la database.

Posteriormente se hizo la creación de la tabla el cual se llamó “candidatos” con las siguientes columnas:

Nombre	Tipo de dato
id	Serial Primary key
first_name	Varchar
last_name	Varchar
email	Varchar
application_date	Date
country	Varchar
YOE	Integer
seniority	Varchar
technology	Varchar
code_challenge_score	Integer
technical_interview_score	Integer
hired	Boolean

Se hizo la creación de la tabla con el siguiente comando `cursor.execute(create_table_query)` y se hizo el commit, para la inserción de los datos en las columnas creadas se usó el comando de PostgreSQL para cargar datos desde el archivo csv (candidates.csv) a la tabla creada, en donde se pone el nombre de las columnas que dispone el csv en el mismo orden que se

establecieron cada una de las columnas. Se indicó un FROM poniendo la dirección de donde se encuentra ubicado el csv junto al DELIMITER que es “;” es decir que las columnas están separadas por puntos y comas. Por último CSV header indicando que la primera línea del archivo contiene el encabezado de la columna y lo asocie con las columnas correspondientes, para hacer el cursor.execute(sql) de la función.

Ahora bien se añadió a la tabla como ya vimos una columna llamada hired (contratado), en las instrucciones se especificaba que para que un candidato quedará contratado tiene que tener el code challenge score y el technical interview score igual o mayor a 7, por lo tanto ejecute otro comando de PostgreSQL para actualizar los registros de la tabla con UPDATE, para hacer el SET a la columna hired a partir de las condiciones ya dichas, en donde al tener un tipo de dato booleano es TRUE o FALSE que en caso de que sea contratado será true y en caso de que no, será false para ya ahora si hacer el commit.

2) Análisis de los datos

Para el análisis de estos datos se puede ver que cuenta con columnas, y 50.000 filas, 3 de los tipos de las columnas son enteros y los demás tipo varchar

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 10 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   First Name               50000 non-null  object
 1   Last Name                50000 non-null  object
 2   Email                    50000 non-null  object
 3   Application Date         50000 non-null  object
 4   Country                  50000 non-null  object
 5   YOE                      50000 non-null  int64 
 6   Seniority                50000 non-null  object
 7   Technology               50000 non-null  object
 8   Code Challenge Score     50000 non-null  int64 
 9   Technical Interview Score  50000 non-null  int64 
dtypes: int64(3), object(7)
memory usage: 3.8+ MB
```

Para mirar si hay filas duplicadas se usó el comando `print(df.duplicated().sum())` y dio como resultado 0 es decir que no hay valores duplicados.

Quise hacer un conteo de los datos de seniority debido a que a lo hora de hacer las visualizaciones aparecian barras bastante parecidas que rondean entre 7000 cada una de las categorías que son:

```
Intern       7255
Mid-Level    7253
Trainee      7183
Junior       7100
Architect    7079
Lead         7071
Senior       7059
Name: Seniority, dtype: int64
```

Luego hice un df.describe() este comando hace un análisis rápido de las variables cuantitativas el cual como resultado tomó el YOE, code challenge score y el technical interview score.

	YOE	Code Challenge Score	Technical Interview Score
count	50000.000000	50000.000000	50000.000000
mean	15.286980	4.996400	5.003880
std	8.830652	3.166896	3.165082
min	0.000000	0.000000	0.000000
25%	8.000000	2.000000	2.000000
50%	15.000000	5.000000	5.000000
75%	23.000000	8.000000	8.000000
max	30.000000	10.000000	10.000000

En donde en cada columna hay 50.000 observaciones, se puede ver que el YOE(years of experience) cuenta con una media de 15.28 así que en promedio, los candidatos tienen alrededor de 15.29 años de experiencia laboral. Por otra parte la desviación estándar de YOE es la que contiene un valor más alto es decir que hay más variabilidad, el mínimo de las 3 tiene un 0 por lo tanto hay candidatos que no cuentan con experiencia laboral, y en los score sacaron la peor puntuación, pero también hay candidatos excepcionales que en como máximo obtuvieron los mejores score ósea 10.

En las visualizaciones se pide que se haga una de contratados por tecnologia asi que revise cuáles valores son únicos en Technology:

```
df['Technology'].unique()
```

```
array(['Data Engineer', 'Client Success', 'QA Manual',
       'Social Media Community Management', 'Adobe Experience Manager',
       'Sales', 'Mulesoft', 'DevOps', 'Development - CMS Backend',
       'Salesforce', 'System Administration', 'Security',
       'Game Development', 'Development - CMS Frontend',
       'Security Compliance', 'Development - Backend', 'Design',
       'Business Analytics / Project Management',
       'Development - Frontend', 'Development - FullStack',
       'Business Intelligence', 'Database Administration',
       'QA Automation', 'Technical Writing'], dtype=object)
```

La cual hay 24 tecnologías, ósea demasiadas para poder describir los datos en una gráfica tipo pastel, por lo tanto opte por agrupar las que tuvieran relación entre sí. Empecé con el primer grupo llamado “Software Development” incluyendo en él todas las que tuvieran que ver con el desarrollo (ya sea de backend, frontend o juegos); Development - CMS Backend, Development - CMS Frontend, Development - Backend, Development - Frontend, Development - FullStack, Game Development y DevOps.

Posteriormente el segundo grupo llamado “Administration and Security” estando en él todas las que hagan referencia a la administración y seguridad; System Administration, Security, Security Compliance y Database Administration. Luego el tercer grupo llamado “Sales and Customer Management” estando en este todas las carreras referentes a las ventas y el manejo de los clientes; Sales, Client Success, Social Media Community Management.

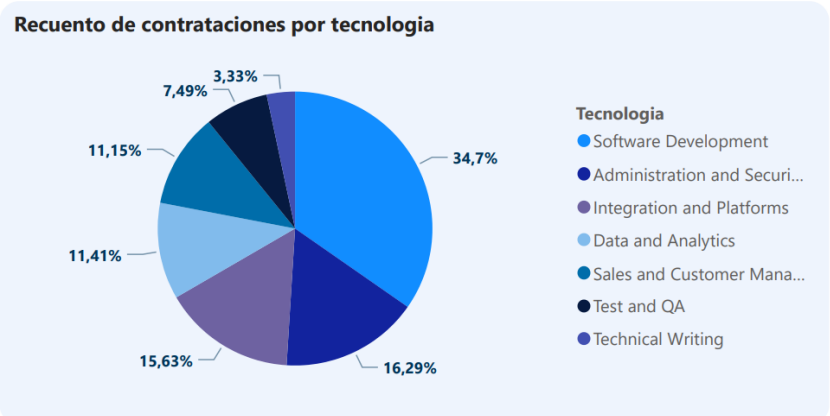
A continuación en el cuarto grupo llamado “Data and Analytics” están ubicadas las que hacen referencia a la administración de los datos y el análisis de ellos; Data Engineer, Business Analytics / Project Management y Business Intelligence. Por otro lugar el quinto grupo llamado “Test and QA” están las que se encargan de las pruebas tanto manuales como de automatización; QA Manual y QA Automation. En penúltimo lugar está el sexto grupo llamado “Integration and Platforms” incluyendo en él las que hacen referencia a la integración y diseño de las plataformas; Mulesoft, Salesforce, Adobe Experience Manager y Design.

Por último está el séptimo grupo llamado “Technical Writing” en donde está ubicada como su nombre lo dice un escritor técnico, no supe como agruparla con la sdemas debido a que es la única que no contaba con una relación en conjunto con las otras, así que se redujeron las tecnologías a 7 grupos que hacen parte a las tecnologías asignadas.

3) Gráficos

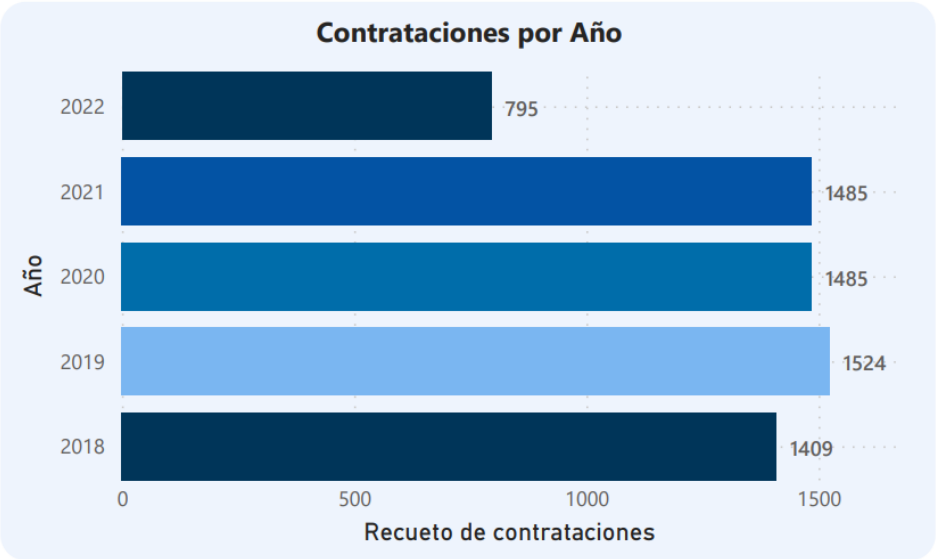
Para la visualizaciones del taller se esperaban estos 5 gráficos;

- **Contrataciones por tecnología (gráfico circular)**



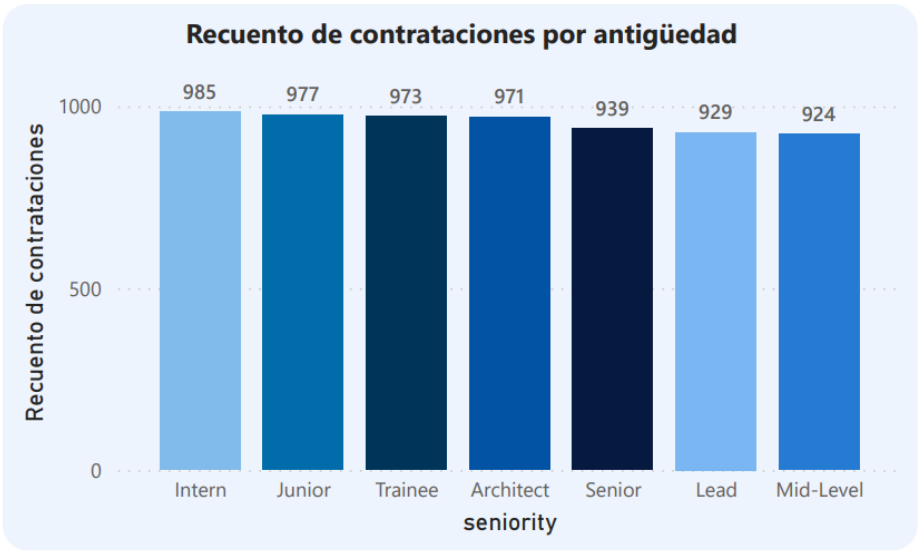
Para este gráfico como ya lo mencione anteriormente, agrupe las tecnologías en 6 grupos debido a que contaban con bastantes, así que lo que podemos concluir con esta visualización es que hay más personal contratado del desarrollo de software que las demás.

- **Contrataciones por año (gráfico de barras horizontales)**



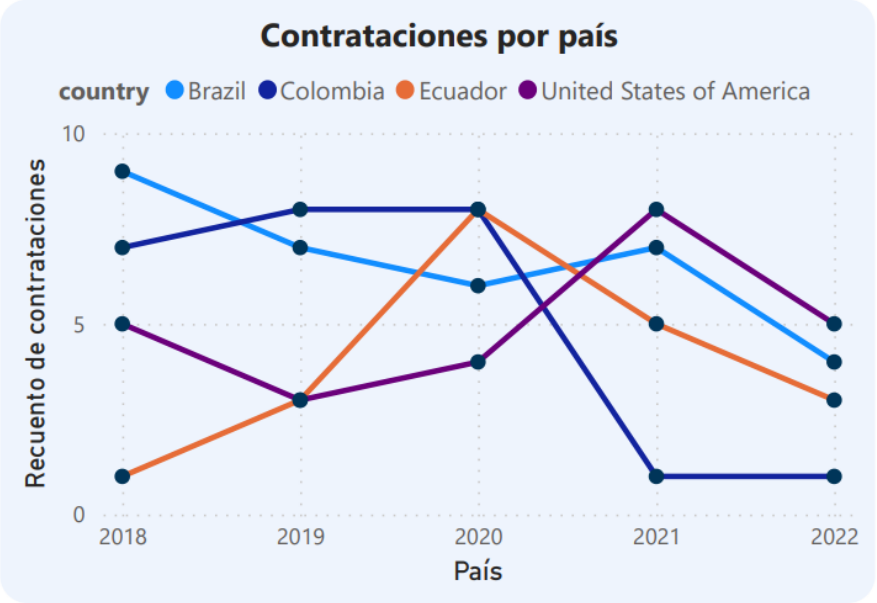
Para esta gráfica tipo barras horizontales, se puede ver que en el año donde más contrataciones hubieron fue en el 2019 con 1524, mientras que en los demás años se manejo estable respecto a los demás años el 2022 solo cuenta con 795, lo más probable es que sea porque aún no habían completado todos los datos respecto a este año.

- **Contrataciones por antigüedad (gráfico de barras)**



En este gráfico cada una de las contrataciones cuenta con 7 niveles de experiencia o antigüedad, en donde la que cuenta con una mayor cantidad es el “Intern” tiene sentido debido a que son las personas que estan en pasantías o en prácticas laborales y tienen más posibilidad de ser contratados al querer ganar experiencia, y de alguna manera los que cuentan con un número menor de contratados son los “mid-level” el cual son los que llevan de 5 a 10 años de experiencia laboral.

- **Contrataciones por país durante años (solo EE. UU., Brasil, Colombia y Ecuador) (multilínea)**



En esta gráfica las contrataciones por país es bastante variable pero a medida que vamos avanzando en los años USA aumenta su número de contrataciones, mientras que a pesar de que Colombia arranca con un buen número de contrataciones para el 2022 solo cuenta con 1, demostrando un decaimiento en el mundo laboral para las contrataciones en Colombia.

Aunque al hacer el análisis exploratorio de datos, se pudo encontrar que el año 2022 está incompleto y va hasta el mes de Julio nomas, asi que las gráfica de contratados por año y esta, estan mal porque no hace el análisis de todo el año completo.

```
df['Application Date'] = pd.to_datetime(df['Application Date'])

# Filtrar las fechas al año 2022
df_2022 = df[df['Application Date'].dt.year == 2022]
cuenta_meses = df_2022['Application Date'].dt.month.nunique()

print("Cantidad de meses en el año 2022:", cuenta_meses)
```

```
Cantidad de meses en el año 2022: 7
```