

Workshop#1



Santiago Gomez Castro

2226287

ETL

Docente: Javier Alejandro Vergara Zorrilla

Universidad Autónoma de Occidente

Facultad de Ingeniería

Santiago de Cali

2024

Metodología

Este trabajo busca explicar la solución a la actividad WorkShop#1, actividad que consiste en la limpieza y análisis de un csv que contenía datos sintéticos sobre empleados candidatos. El objetivo principal de la actividad es mostrar el conocimiento y habilidad al momento de usar herramientas como Python, bases de datos, entre otros.

1. Datos usados

First Name	Last Name	Email	Application Date	Country	YOE	Seniority
Bernadette	Langworth	leonard91@yahoo.com	2021-02-26	Norway	2	Intern
Camryn	Reynolds	zelda56@hotmail.com	2021-09-09	Panama	10	Intern
Larue	Spinka	ekey.schultz4@gmail.com	2020-04-14	Belarus	4	Mid-level
Arch	Spinka	elvera_kulas@yahoo.com	2020-10-01	Eritrea	25	Trainee
Larue	Altenwerth	minnie.gislason@gmail.com	2020-05-20	Myanmar	13	Mid-level
Alec	Abbott	juanita_hansen@gmail.com	2019-08-17	Zimbabwe	8	Junior
Allison	Jacobs	alba_rolfson2@yahoo.com	2018-05-18	Wallis and Futuna	19	Trainee
Nya	Col 2: [REDACTED]	madisen.zulauf@gmail.com	2021-12-09	Myanmar	1	Lead
Mose	Lakin	dale_murazik@hotmail.com	2018-03-13	Italy	18	Lead
Terrance	Zieme	dustin31@hotmail.com	2022-04-08	Timor-Leste	25	Lead
Aiyana	Goodwin	vallie.damore@yahoo.com	2019-09-22	Armenia	24	Intern
Emilia	Maelchi	peter.grady@gmail.com	2020-07-15	French Southern Territories	28	Lead
Terrell	Streich	meta92@yahoo.com	2021-12-27	Chad	3	Mid-level
Hilda	Rodriguez	jordan.hyatt@hotmail.com	2020-05-09	El Salvador	16	Junior
Hope	Hansen	clemmie.brueh@hotmail.com	2019-10-12	Mozambique	18	Architect
Arno	Altenwerth	cheyenne_rau2@gmail.com	2018-10-18	Brunei Darussalam	21	Mid-level
Betty	Crona	judd.wisozk55@gmail.com	2020-03-25	Morocco	28	Architect
Clint	Oberbrunner	dwight.jacobson@gmail.com	2021-05-23	Saint Helena	19	Senior

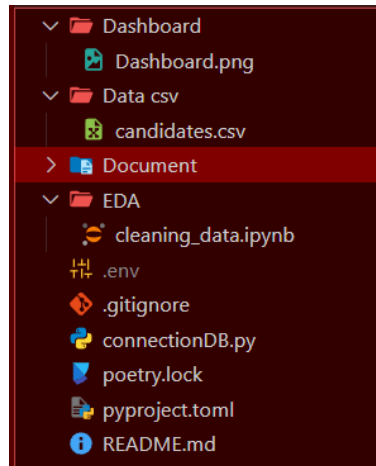
Los datos se encuentran en un csv, esto son datos relacionales organizados en filas y columnas, se encuentra 10 columnas, las cuales son:

- First Name
- Last Name
- Email
- Country
- Application
- Date
- Yoe (years of experience)
- Seniority
- Technology
- Code Challenge Score
- Technical Interview

Con 50.000 datos sintéticos de acuerdo al contexto de cada columna.

2. Herramientas y estructura usada

Estructura:



Las herramientas usadas son:

- Python: Se empleo un script de Python para la subida de los datos a una base de datos relacional y almacenamiento de los datos.
- PostgreSQL: Siendo una base de datos relacional facilita el almacenamiento de los datos y gestión de los mismo para la realización de consultas o alguna modificación de ser necesario.
- Jupyter: Se creo un script de jupyter para la manipulación y modificación de los datos, también limpieza y poder obtener visualizaciones para entender mejor como se distribuyen de los datos.
- Poetry: El uso de un ambiente de trabaja es muy beneficioso para la manipulación y descarga de librerías de Python, en este caso se usó Poetry para la creación del ambiente.
- Git y GitHub: El sistema de versiones más popular para poder guardar las modificaciones y enviar el trabajo a terceros.
- Ubuntu: La mayor parte del trabajo se ha realizado en una máquina virtual Ubuntu para mayor facilidad en la creación y gestión de ambiente virtuales.
- Power BI: Se emplea de Power BI para la realización de las gráficas pedidas en el WorkShop.
- VScode: Siendo el editor de código más popular por su facilidad de uso y múltiples extensiones para un mejor trabajo.

- DBeaver: Es una herramienta de software que proporciona una interfaz gráfica para la visualización, gestión y manipulación de bases de datos.

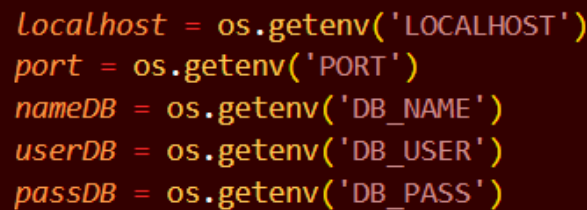
3. Manejo de credenciales

Se crea un archivo **.env** para guardar las credenciales de la base de datos para poder conectarse y realizar consultas, este se agrega en él **.gitignore** evitando ser guardado y compartido por git ya que esta información es delicada y no debe ser compartida a terceros.

A screenshot of a code editor showing a file named .env. The file contains five lines of environment variables: LOCALHOST=, PORT=, DB_NAME=, DB_USER=, and DB_PASS=. The sixth line is empty and highlighted. The editor has a dark theme with a yellow cursor at the end of the fifth line.

```
.env
1  LOCALHOST=
2  PORT=
3  DB_NAME=
4  DB_USER=
5  DB_PASS=
6
```

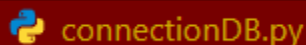
Usando la librería **dotenv** el código podrá acceder a las credenciales.


A code block containing Python code that uses the os module to retrieve environment variables. The variables are assigned to local variables: localhost, port, nameDB, userDB, and passDB.

```
localhost = os.getenv('LOCALHOST')
port = os.getenv('PORT')
nameDB = os.getenv('DB_NAME')
userDB = os.getenv('DB_USER')
passDB = os.getenv('DB_PASS')
```

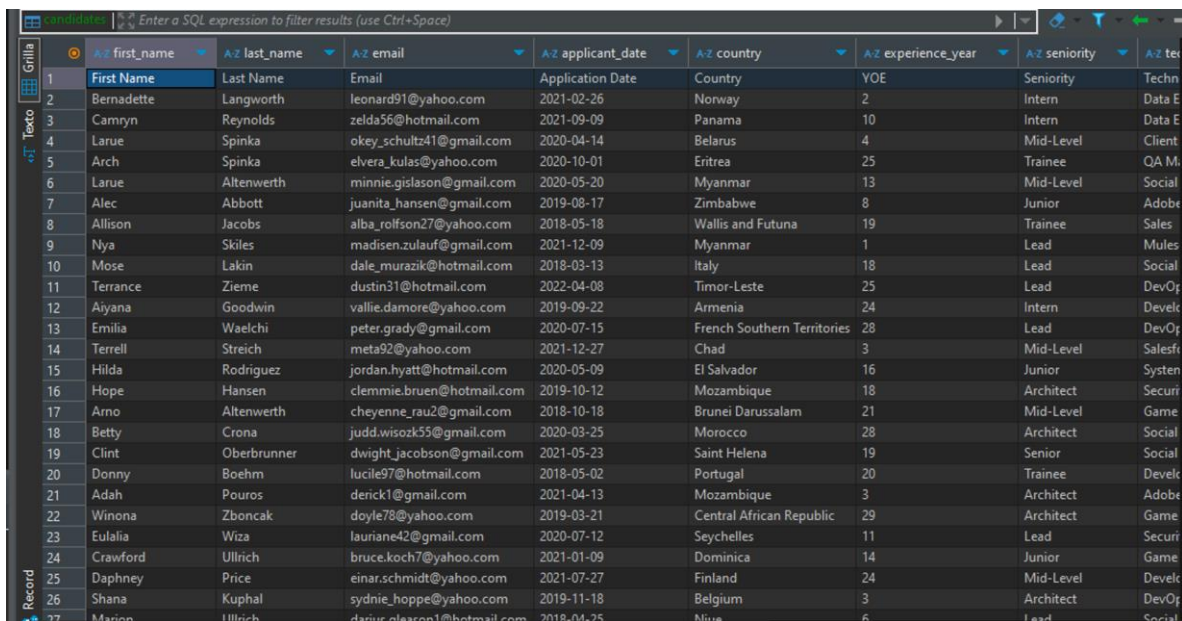
4. Migración de los datos a PostgreSQL

Se creo un sript de Python llamado **connectionDB.py** para la migración inicial de los datos desde el csv a PostgreSQL para poder empezar con la manipulación de la información.

A small image showing the Python logo (two interlocking snakes, one blue and one yellow) followed by the filename connectionDB.py in a monospaced font.

```
 connectionDB.py
```

Aquí revisamos la subida de los datos dentro de la base de datos.



	first_name	last_name	email	applicant_date	country	experience_year	seniority	title
1	First Name	Last Name	Email	Application Date	Country	YOE	Seniority	Techn
2	Bernadette	Langworth	leonard91@yahoo.com	2021-02-26	Norway	2	Intern	Data E
3	Camryn	Reynolds	zelda56@hotmail.com	2021-09-09	Panama	10	Intern	Data E
4	Larue	Spinka	okey_schultz41@gmail.com	2020-04-14	Belarus	4	Mid-Level	Client
5	Arch	Spinka	elvera_kulas@yahoo.com	2020-10-01	Eritrea	25	Trainee	QA M
6	Larue	Altenwerth	minnie.gislason@gmail.com	2020-05-20	Myanmar	13	Mid-Level	Social
7	Alec	Abbott	juanita_hansen@gmail.com	2019-08-17	Zimbabwe	8	Junior	Adobe
8	Allison	Jacobs	alba_rolfson27@yahoo.com	2018-05-18	Wallis and Futuna	19	Trainee	Sales
9	Nya	Skiles	madisen.zulauf@gmail.com	2021-12-09	Myanmar	1	Lead	Mules
10	Mose	Lakin	dale_murazik@hotmail.com	2018-03-13	Italy	18	Lead	Social
11	Terrance	Zieme	dustin31@hotmail.com	2022-04-08	Timor-Leste	25	Lead	DevOp
12	Aiyana	Goodwin	vallie.damore@yahoo.com	2019-09-22	Armenia	24	Intern	Develo
13	Emilia	Waelchi	peter.grady@gmail.com	2020-07-15	French Southern Territories	28	Lead	DevOp
14	Terrell	Streich	meta92@yahoo.com	2021-12-27	Chad	3	Mid-Level	Salesfr
15	Hilda	Rodriguez	jordan.hyatt@hotmail.com	2020-05-09	El Salvador	16	Junior	System
16	Hope	Hansen	clemmie.bruen@hotmail.com	2019-10-12	Mozambique	18	Architect	Securi
17	Arno	Altenwerth	cheyenne_rau2@gmail.com	2018-10-18	Brunei Darussalam	21	Mid-Level	Game
18	Betty	Crona	judd.wisozk55@gmail.com	2020-03-25	Morocco	28	Architect	Social
19	Clint	Oberbrunner	dwight_jacobson@gmail.com	2021-05-23	Saint Helena	19	Senior	Social
20	Donny	Boehm	lucile97@hotmail.com	2018-05-02	Portugal	20	Trainee	Develo
21	Adah	Pouros	derick1@gmail.com	2021-04-13	Mozambique	3	Architect	Adobe
22	Winona	Zboncak	doyle78@yahoo.com	2019-03-21	Central African Republic	29	Architect	Game
23	Eulalia	Wiza	lauriane42@gmail.com	2020-07-12	Seychelles	11	Lead	Securi
24	Crawford	Ullrich	bruce.koch7@yahoo.com	2021-01-09	Dominica	14	Junior	Game
25	Daphney	Price	einar.schmidt@yahoo.com	2021-07-27	Finland	24	Mid-Level	Develo
26	Shana	Kuphal	sydney_hoppe@yahoo.com	2019-11-18	Belgium	3	Architect	DevOp
27	Mason	Ullrich	darius_nolan1@hotmail.com	2018-04-25	Nicar	6	Lead	Social

5. EDA

- Limpieza y análisis

Se usa de Jupyter para la manipulación y modificación de los datos para el cumplimiento de las tareas del Workshop.

Empezaremos con la obtención de los datos desde la base de datos usando la librería **sqlalchemy** para la conexión a la base de datos y **dotenv** para traer las credenciales.

```
load_dotenv()

localhost = os.getenv('LOCALHOST')
port = os.getenv('PORT')
nameDB = os.getenv('DB_NAME')
userDB = os.getenv('DB_USER')
passDB = os.getenv('DB_PASS')

engine = create_engine(f'postgresql+psycopg2://{userDB}:{passDB}@{localhost}:{port}/{nameDB}')

inspector = inspect(engine)
```

Primero realizamos una visualización de los datos para entender los datos que este posee y si hay valores nulos (null).

```

...   first_name last_name email applicant_date country \
0 First Name Last Name Email Application Date Country
1 Bernadette Langworth leonard91@yahoo.com 2021-02-26 Norway
2 Camryn Reynolds zelda56@hotmail.com 2021-09-09 Panama
3 Larue Spinka okey_schultz41@gmail.com 2020-04-14 Belarus
4 Arch Spinka elvera_kulas@yahoo.com 2020-10-01 Eritrea

experience_year seniority technology code_challenge_score \
0 YOE Seniority Technology Code Challenge Score
1 2 Intern Data Engineer 3
2 10 Intern Data Engineer 2
3 4 Mid-Level Client Success 10
4 25 Trainee QA Manual 7

technical_interview_score
0 Technical Interview Score
1 3
2 10
3 9
4 1
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50001 entries, 0 to 50000
Data columns (total 10 columns):
# Column Non-Null Count Dtype
---
...
9 technical_interview_score 50001 non-null object
dtypes: object(10)
memory usage: 3.8+ MB
None

```

Observamos los datos y columnas que posee la base de datos, vemos un poco la variedad de los datos y como ninguna columna no posee algún valor nulo.

En nuestro caso la primera fila se guardó los nombres las columnas, estos datos son innecesarios para nuestros análisis, entonces los borramos.

```

first_name last_name email applicant_date country \
0 Bernadette Langworth leonard91@yahoo.com 2021-02-26 Norway
1 Camryn Reynolds zelda56@hotmail.com 2021-09-09 Panama
2 Larue Spinka okey_schultz41@gmail.com 2020-04-14 Belarus
3 Arch Spinka elvera_kulas@yahoo.com 2020-10-01 Eritrea
4 Larue Altenwerth minnie.gislason@gmail.com 2020-05-20 Myanmar

experience_year seniority technology \
0 2 Intern Data Engineer
1 10 Intern Data Engineer
2 4 Mid-Level Client Success
3 25 Trainee QA Manual
4 13 Mid-Level Social Media Community Management

code_challenge_score technical_interview_score
0 3 3
1 2 10
2 10 9
3 7 1
4 9 7

```

Revisamos en la columna email si hay valores duplicados, encontramos como algunos correos se repiten, pero esto es normal por lo que son generados de forma artificial.

```
Frecuencia de valores en la columna 'email':
email
fern70@gmail.com      3
marianne31@yahoo.com  3
charley51@gmail.com   2
brooks60@hotmail.com  2
rogers12@gmail.com    2
..
diana70@gmail.com     1
marion91@hotmail.com  1
sister51@hotmail.com  1
marvin_parker@gmail.com 1
abigayle.crooks@yahoo.com 1
Name: count, Length: 49833, dtype: int64
```

Revisamos la columna **experience_year** si existen datos outliers.

... Max 9
Min 0

Ya después de revisar que los datos estén bien, pasamos a la creación de la columna donde muestre si el candidato fue aceptado o rechazado.

first_name	last_name	email	applicant_date	country	
0	Bernadette	Langworth	leonard91@yahoo.com	2021-02-26	Norway
1	Camryn	Reynolds	zelda56@hotmail.com	2021-09-09	Panama
2	Larue	Spinka	okey_schultz41@gmail.com	2020-04-14	Belarus
3	Arch	Spinka	elvera_kulas@yahoo.com	2020-10-01	Eritrea
4	Larue	Altenwerth	minnie.gislason@gmail.com	2020-05-20	Myanmar

experience_year	seniority	technology	
0	2	Intern	Data Engineer
1	10	Intern	Data Engineer
2	4	Mid-Level	Client Success
3	25	Trainee	QA Manual
4	13	Mid-Level	Social Media Community Management

code_challenge_score	technical_interview_score	hired	
0	3	3	0
1	2	10	0
2	10	9	1
3	7	1	0
4	9	7	1

La columna lleva el nombre de hired, donde solo dominan 2 valores, 1 significando que el candidato fue aprobado y 0 que el candidato fue rechazado.

Se suben los nuevos datos a la base de datos.

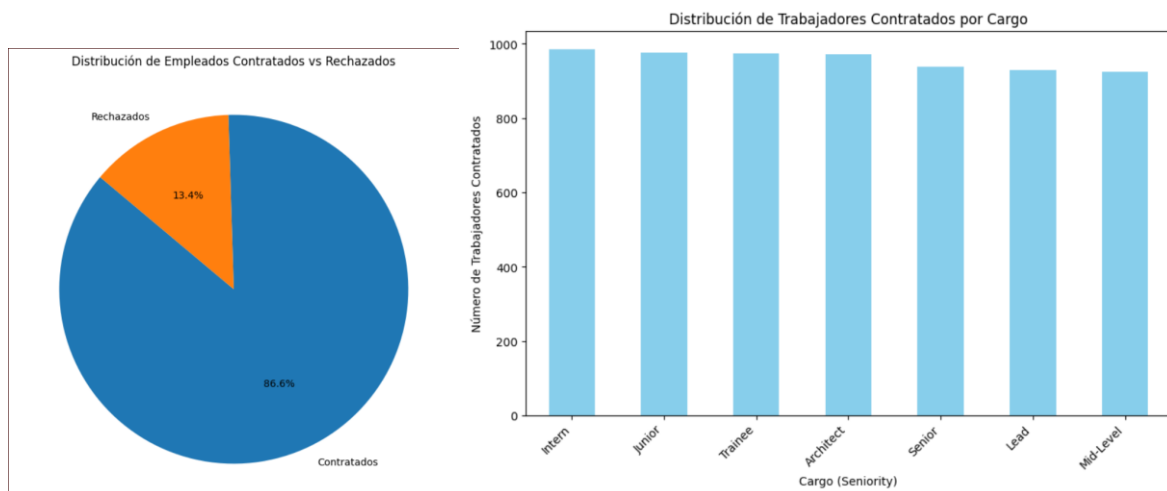
```
try:
    df.to_sql('candidates', engine, if_exists='replace', index=False)
    print(f"Tabla 'candidates' se actualizó.")

except Exception as e:
    print(f"Error al subir los datos: {e}")

finally:
    engine.dispose()
```

Tabla 'candidates' se actualizó.

- **Visualización**

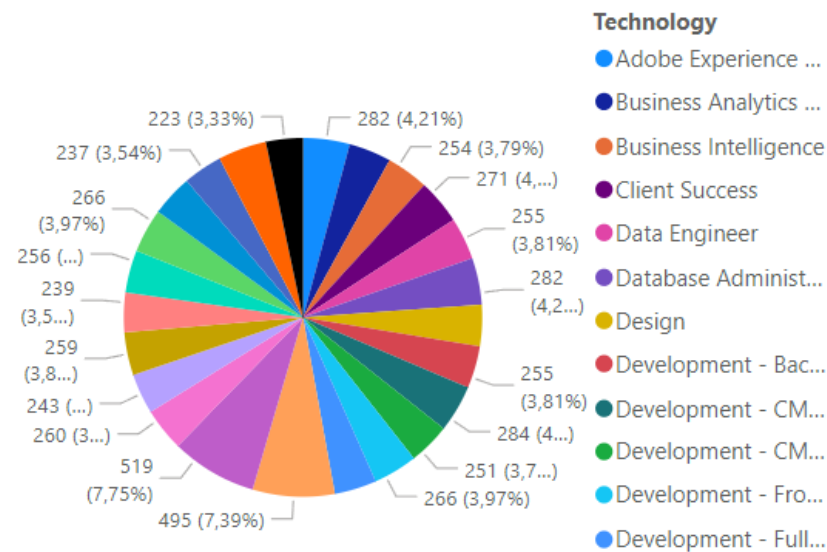


Estas visualizaciones preliminares nos muestran, por el lado del grafico de paste; El porcentaje de los candidatos que fueron aceptas(azul) con los que fueron rechazados(naranja), dando a entender que un gran porcentaje los candidatos lograron ser aceptados, y el grafico de barras; Muestra los cargos que fueron aceptados, mostrando la necesidad de una gran variedad de cargos requeridos sin dejar ninguno como el cargo más rechazada, claramente hay cargos que fueron más aceptados a diferencia de otros, pero la gráfica muestra lo parejo que han esta todos estos.

6. Graficas Power BI

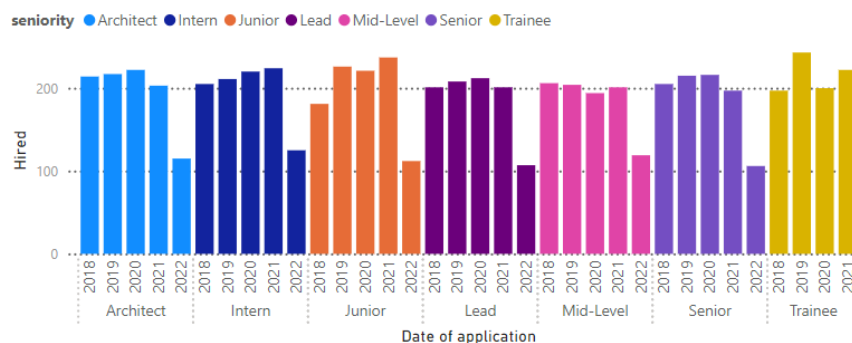
- La grafica de pastel muestra los porcentajes de empleados contratados y la tecnología que manejan, siendo Adobe Experience Manager la tecnología con mayor contratación seguida por Business analytics/Proyect Management y Business Intelligence, estas son las tecnologías con más contratación, podemos concluir que sabiendo alguna de estas tecnologías el candidato tiene una mayor oportunidad para pasar y ser aceptado.

Hired by technology

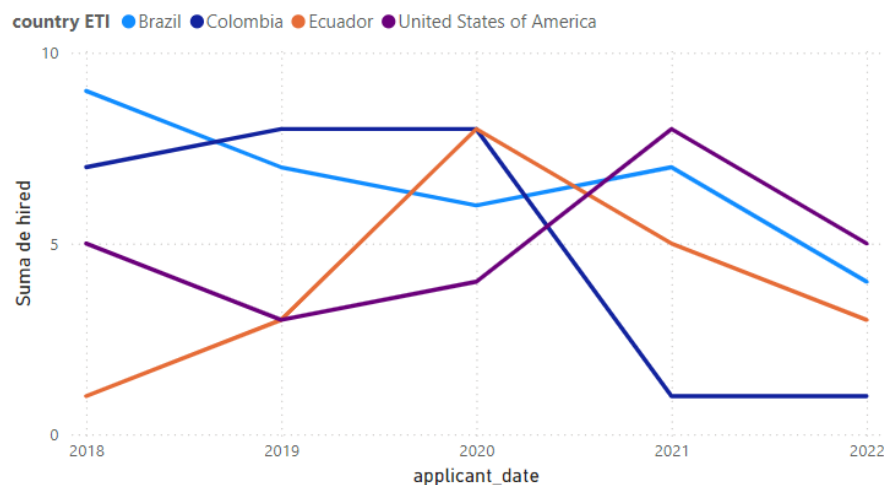


- La grafica de barra vertical muestra las fechas de aplicación de los candidatos aceptados y el cargo, observamos como el año 2019 fue el año con mayor aplicación de candidatos con el cargo de **Trainee** y fueron aceptados, se ve una tendencia en el año 2022, año el que menos candidatos aplicaron exceptuando los candidatos con el cargo de **Trainee**, de resto de los años muestra como es parejo el número de candidatos que aplicaron y fueron aceptados.

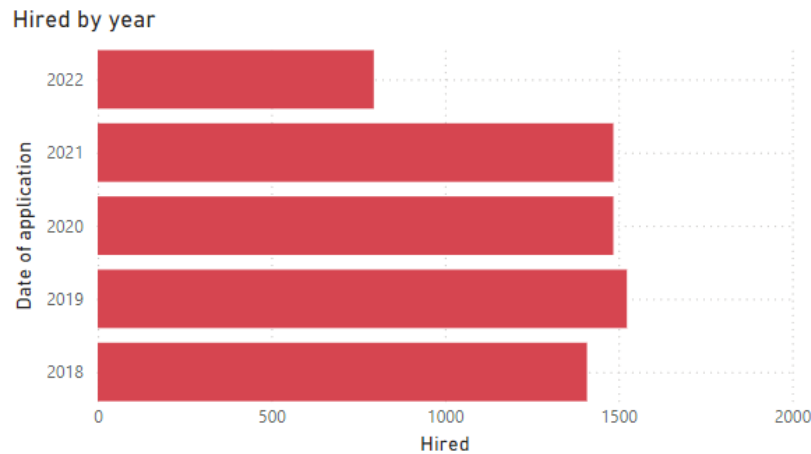
Hires by seniority



- El grafico múltiple muestra como se distribuyen los candidatos que fueron contratados, año en el que aplicaron y la nacionalidad. Vemos la tendencia descendente de Brasil, siendo el país con mayores candidatos contratado en el año 2018, en el año 2021 muestra un aumento de candidatos aceptados, pero vuelve a tomar la misma tendencia descendente. Ecuador empieza en el año 2018 con el menor número de candidatos aceptados, pero tiene una tendencia ascendente hasta 2020 siendo el pico más alto de candidatos aprobados y después de este año empieza con una tendencia descendente. Colombia en el año 2018 empieza como el segundo país con mayor número de candidatos aceptados, Tiene una leve tendencia ascendente hasta el año 2020, a partir de este año observamos el desplome de los candidatos aceptados hasta 2021 donde posee una tendencia estable, haciendo que Colombia pase de ser el segundo país con mayores candidatos aceptados a ser el último. El último país por analizar es Estados Unidos (United States of America), empezando como el tercer país con mayores candidatos aceptados y una tendencia descendente hasta el año 2019, desde este año empieza una tendencia ascendente hasta el 2021 siendo pico máximo de candidatos aceptados y terminando como el país con mayores candidatos aceptados en el año 2022.



La grafica de barra horizontal muestre la fecha de aplicación de los candidatos y los que fueron aceptados, observamos que el año con mayor numero de candidatos aceptados es el 2019 con más de 1500 candidatos aceptados y el año 2022 siendo el año con menor candidatos aceptados, por ultimo los años 2020 y 2021 son los años con un numero parejo de candidatos aceptados.



Conclusiones generales

Los datos son muy completos y variados en cuestión de país de origen, cargo de los candidatos, tecnologías entre otras. Estos análisis permiten la indagación dentro de la base de datos y la naturaleza de los datos, con propósito de análisis o visualización de estos para entender ya de forma visual como se distribuyen los datos.

Hablando ahora de las gráficas observamos la variación de las tecnologías y cargos de los candidatos aceptados, mostrando que ninguna es superior o mejor sobre las demás, sino que estas obtienen un nivel de aceptación parecido, por supuesto hay con más números de aceptados pero esta diferencia no es extensa. La fecha de aplicación de los candidatos que fueron aceptados tiene una tendencia similar en general a excepción del año 2022, año donde la cantidad de candidatos aceptados fue menor a comparación de los años anteriores.