Entrega del Proyecto Final

ENTREGA FINAL DEL PROYECTO

Formato: Tablero en archivo ejecutable de Power BI (.pbix). La documentación debe ser formato pdf.
Sugerencia: Adjuntar la fuente de datos del tablero, archivo plano xls, csv o txt. En caso de ser formato google slides, hacer público el documento.

>>Objetivos Generales:

1. Integrar los conocimientos adquiridos en la cursada.

>>Objetivos específicos:

1. Desarrollar un tablero de control.
2. Documentar la iniciativa de análisis de datos.

>>Aspectos a tener en cuenta:

Los archivos de la entrega final son un conglomerado de las entregas parciales previamente desarrolladas.

>>Se debe entregar:

Cada archivo debe llevar por título el nombre del proyecto y los nombres de los integrantes. En el caso del tablero, debe estar en la solapa de portada.

1. **En el archivo pdf:**

- **Debe incluir la documentación presentada en la primera entrega de proyecto final.**
- Portada: título principal del proyecto, nombre de los integrantes, institución y fecha de presentación.
- Tabla de contenido.
- Introducción.
- Tabla de versionado.
- Objetivo.
- Cuerpo del documento.
- Herramientas tecnológicas implementadas.
- Futuras líneas.

A. Sobre la base de datos

- Descripción de la base de datos usada en la que se explique la temática, los indicadores que se implementaron y la segmentación de estos.
- Imagen del diagrama de entidad-relación de la base de datos.
- Descripción de cada tabla usada, en la que se detalle la definición de cada columna.

B. Sobre la visualización:

- Objetivo.
- Alcance.
- Áreas de la organización que serán usuarios finales. (Solo si aplica al proyecto)
- Imagen de cada una de las solapas, con una breve descripción de qué información presenta y qué análisis permite hacer.
- Imagen del diagrama de entidad-relación perteneciente al desarrollo de la visualización con una breve descripción de qué información presenta cada tabla.

Futuras líneas: breve descripción de las posibles iniciativas que se pueden llevar a cabo para complementar el proyecto.

2. En el archivo .pbix:

- **Debe incluir los requerimientos de la segunda entrega de proyecto final.**

Aspectos generales del dashboard:
Desarrollarás y diseñarás un tablero en la herramienta Power BI con una extensión de entre tres y seis solapas.
El tablero debe tener una estructura eficiente en la que se integren todos los conocimientos vistos en el programa.

Recuerda que cuentas con un plazo de 20 días a partir de esta clase para realizar la entrega.

Portada

Titulo principal del proyecto: Dashboard informativo para posiciones relacionadas a Data Science

Nombre de los integrantes: Daniel Arbelaez, Daniel Ramirez

Institución: N/A

Fecha de presentación: 4 de julio de 2022

Tabla de contenido

- Entrega del Proyecto Final
  - Portada
  - Tabla de contenido
  - Introducción
  - Objetivo
  - Herramientas tecnológicas implementadas
  - Contenido del desarrollo
    - Base de datos
    - Visualización

- Futuras líneas

Introducción

En el presente proyecto se pretende entregar el desarrollo del proyecto final, el mismo que se utilizó durante toda la cursada.

Para cumplir con las consignas indicadas en la imagen anterior, se procederá a detallar en orden los ítems indicados que definen todos los requisitos del proyecto final.

Objetivo

En el presente trabajo se realizó inicialmente una busqueda de datasets que comprendían datos sobre salarios de los empleos en áreas de data science; seguido se realizó inicialmente el análisis de los datos que comprendían el dataset elegido, en este caso, "Data Science Jobs Salaries", el cual fue encontrado en el repositorio *Kaglee* al cual se puede acceder a través del siguiente link

Esquema en estrella: Según wikipedia, "*En las bases de datos usadas para data warehousing, **un esquema en estrella** es un modelo de datos que tiene una **tabla de hechos** (o tabla fact) que contiene los datos para el análisis, rodeada de las **tablas de dimensiones***". Dicho diagrama quedó de la siguiente forma:

Herramientas tecnológicas implementadas

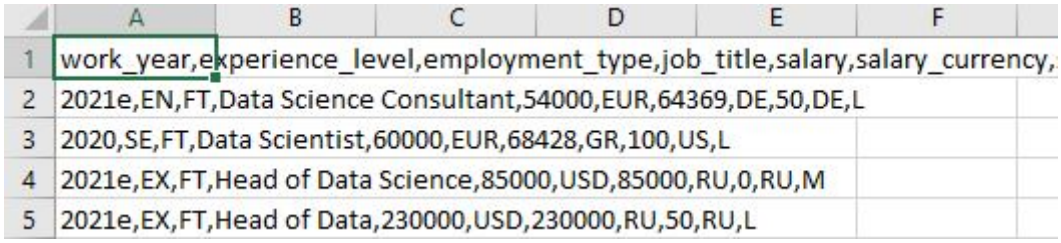
Para el presente proyecto se utilizaron las herramientas de Power BI, Excel y Visio, los cuales fueron necesarios, en el caso de Excel se utilizó en la definición de la base de datos y comprensión del dataset escogido, asimismo para el proceso de transformación al modelo estrella; para el caso de Visio fue usado para mayor comprensión de las entidades y la creación del modelo entidad relación; en el caso de PowerBI fue usado para la creación de los dashboard, medidas calculadas, transformaciones, tablas calculadas y medidas basadas en parámetros

Contenido del desarrollo

Base de datos

1. El dataset inicial se encontraba en formato `.csv` separado por comas, y fue seteado de forma tal que fue separado en 6 tablas principales, las cuales fueron usadas para el diseño del diagrama Entidad-Relación.

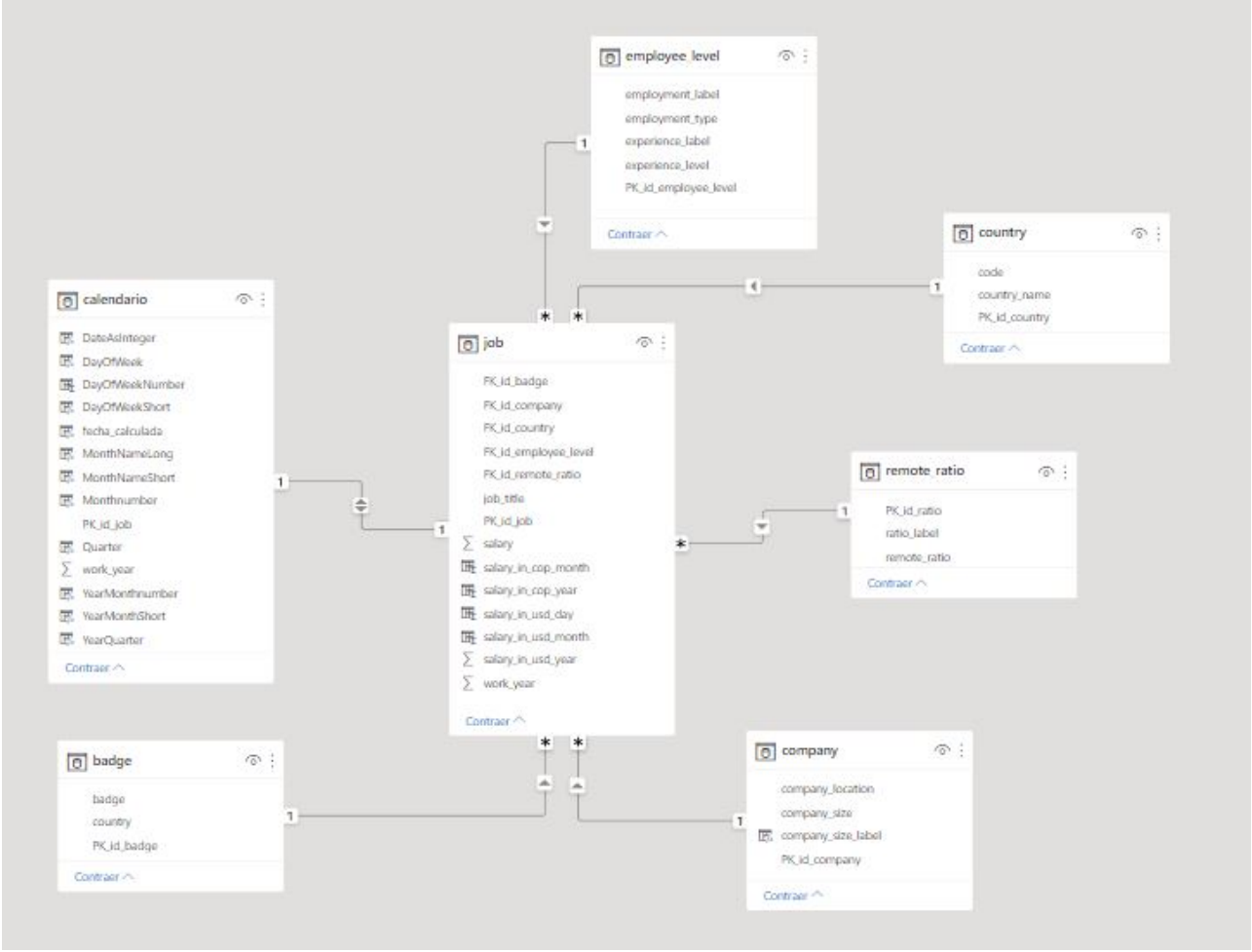
   Dataset original:



	A	B	C	D	E	F
1	work_year,experience_level,employment_type,job_title,salary,salary_currency,					
2	2021e,EN,FT,Data Science Consultant,54000,EUR,64369,DE,50,DE,L					
3	2020,SE,FT,Data Scientist,60000,EUR,68428,GR,100,US,L					
4	2021e,EX,FT,Head of Data Science,85000,USD,85000,RU,0,RU,M					
5	2021e,EX,FT,Head of Data,230000,USD,230000,RU,50,RU,L					

   El dataset final fue diseñado a partir de 6 tablas extraídas y editadas del dataset original, entre otras, job, remote_ratio, company, country, employee_level y badge. Por último, el modelo de datos se diseño con un esquema en estrella.

2. Imagen del diagrama de entidad-relación de la base de datos



3. **Listado de columnas por tablas con tipo de datos**:

Tabla	Columna	Tipo de dato
job	PK_id_job	Int
job	job_title	Varchar(40)
job	work_year	Varchar(5)
job	salary	Int
job	salary_in_usd_year	Int
job	FK_id_remote_ratio	Int
job	FK_id_company	Int
job	FK_id_badge	Varchar(5)
job	FK_id_employee_level	Int
-	-	-
remote_ratio	PK_id_remote_ratio	Int
remote_ratio	remote_ratio	Int

Tabla	Columna	Tipo de dato
remote_ratio	ratio_label	Varchar(20)
-	-	-
company	PK_id_company	Int
company	company_location	Varchar(5)
company	company_size	Varchar(5)
-	-	-
country	PK_id_country	Int
country	code	Varchar(5)
country	country_name	Varchar(20)
-	-	-
employee_level	PK_id_employee_level	Int
employee_level	experience_level	Varchar(5)
employee_level	experience_label	Varchar(40)
employee_level	employment_type	Varchar(5)
employee_level	employment_label	Varchar(20)
-	-	-
badge	PK_id_badge	Int
badge	badge	Varchar(5)
badge	country	Varchar(40)

Visualización

1. El objetivo del presente dashboard es presentar información de valor acerca de los salarios, ubicaciones, empresas entre otros, de los empleos en el área de Data Science.

2. Alcance: El presente proyecto pretende evidencias información de valor respecto de los salarios que obtuvieron empleos de Ciencia de Datos en los años 2020 y 2021. Lo anterior, tendrá relavancia para tomar decisiones respecto al área de enfoque según ingresos y demás apartados en el que se podría especializar un usuario que se interesa en el área de Ciencia de Datos. Adicionalmente podrá utilizarse para futuras proyecciones en el sentido de reutilizar dicho proyecto en aras de combinar datos y generar información de valor.

3. Solapas del dashboard (con una breve descripción de qué información presenta y que análisis permite hacer)

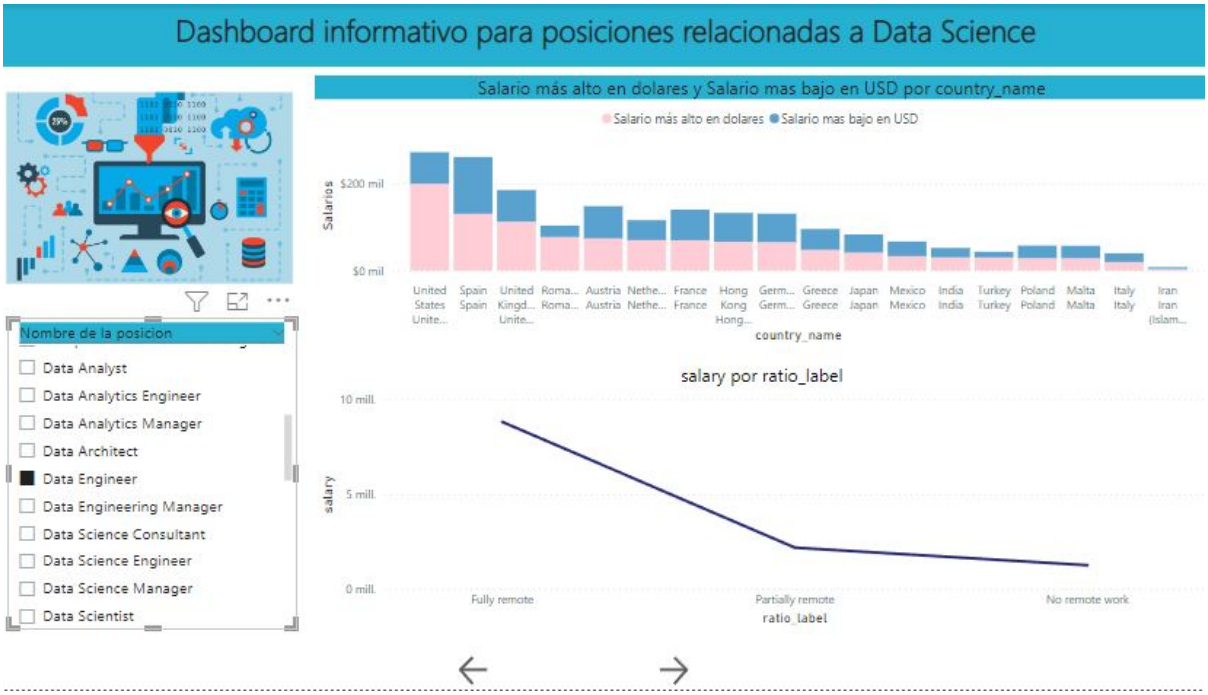
    1. Portada: En esta solapa se da una presentación visual que incorpora los botones interactivos para acceder a las diferentes solapas del dashboard. En esta sección es posible identificar el objeto de

análisis que comprende información relacionada con Data Science.

2. Principal: En esta solapa se presenta en la parte de la izquiera filtros relacionados con país, posición, tipo de empleo, año y rango de salario; seguido se muestran medidas generales relacionadas con cantidad de países, salario mayor en usd, salario más bajo en usd, número de posiciones por país y top 5 de empleos con mayor salario al promedio. En esta primer solapa es posible analizar aspectos claves que fueron obtenidas de la Base de Datos como lo es el salario mayor y que puede tomarse como un punto relevante a la hora de hacer un análisis y que de acuerdo a la cantidad de posiciones por país se interpreta que USA es el país con mayores registros de empleos en el dataset.

3. Estadísticos: Se presenta una segmentación de datos en la parte de la izquierda donde se incluyen todas la posiciones (empleos) que registran en la base de datos y que dicho filtro modifica los valores del gráfico de columnas apiladas y el gráfico de líneas. En primer lugar en esta solapa es posible distinguir entre los diferentes posiciones (empleos) y que para el caso de Data Engineer se pueden aislar los datos de salario (mayor y menor) y ratio de remoto (full remote, partial remote y no remote), permitiendo así evidenciar que la mayoría de salarios

superiores corresponden a empleos que son full remotos según el dataset obtenido.

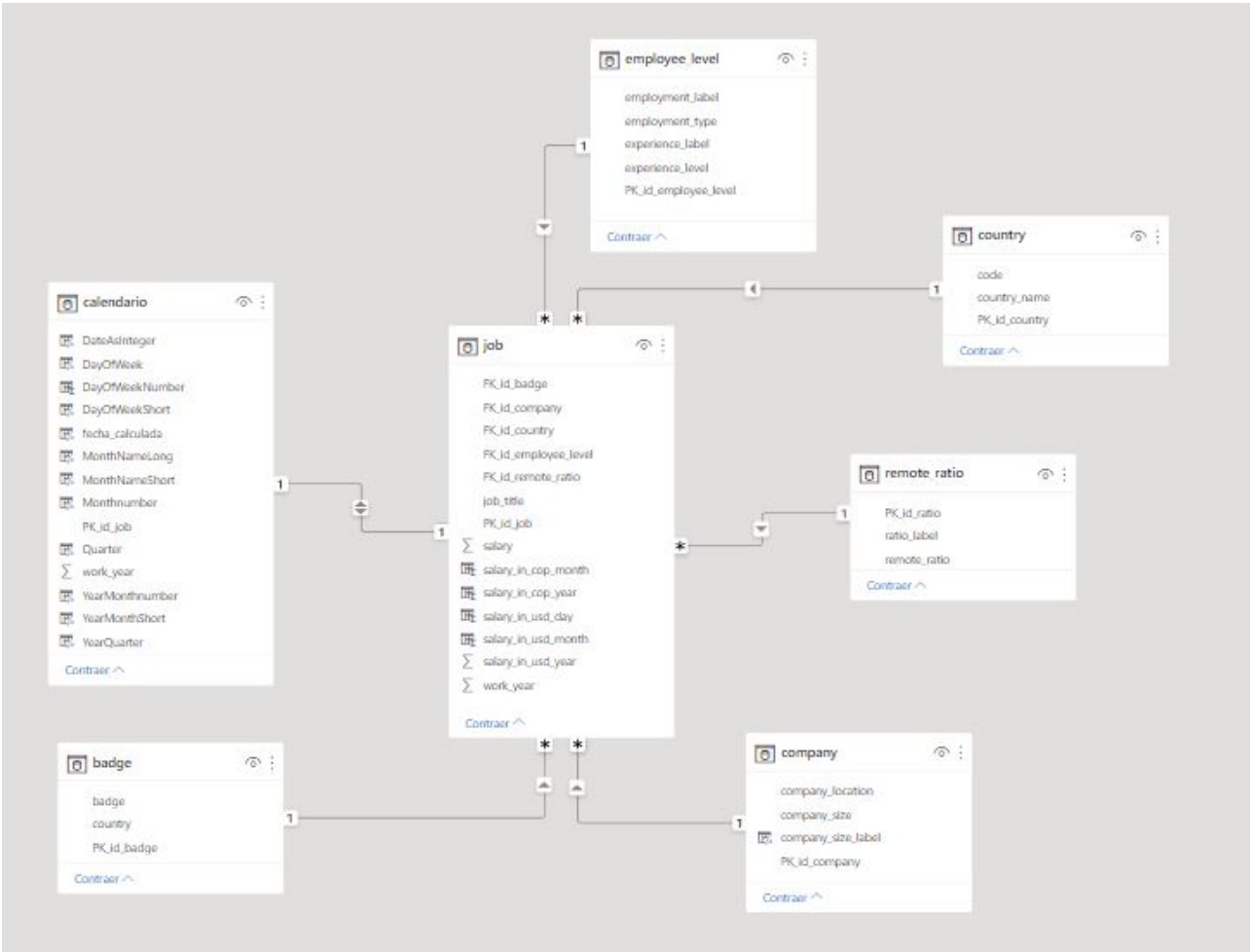


4. Mapa: En esta sección se muestra una jerarquía del Top 15 de los menores salarios de Data Science relacionado con el nivel de empleo en una visualización de mapa de árbol.

5. Medidas: En esta solapa podemos encontrar variedad de medidas, medidas avanzadas, una tabla que contiene nombre de empleo y su correspondiente salario que puede ser modificado mediante el parámetro de precio de dolar en COP. En la presente sección es posible examinar

diferentes medidas que pueden aportar valor a las anteriores secciones.

4. Diagrama Entidad - Relación: Para el diseño del presente diagrama se tuvo en cuenta que la tabla jobs se consideró como la tabla hechos por tratarse del tema principal en cuanto a los salarios de empleos en áreas de Data Science, las demás fueron clasificadas como tablas de dimensiones.



Listado de tablas con PK y FK:

- Tabla "job":

PK_id_job	job_title	work_year	salary	salary_in_usd_year	FK_id_remote_ratio	FK_id_company	FK_id_badge	FK_id_employee_level	FK_id_country
5	Machine Learning Engineer	2021	125000	125000	3	5	USD	1	237
6	Data Analytics Manager	2021	120000	120000	3	6	USD	2	237
7	Research Scientist	2020	450000	450000	1	6	USD	4	237
11	Manager Data Science	2021	144000	144000	3	2	USD	2	237
14	Data Scientist	2021	150000	150000	3	6	USD	4	237
15	Data Science Consultant	2020	103000	103000	3	2	USD	4	237

- Tabla "remote_ratio":

PK_id_remote_ratio	remote_ratio	ratio_label
1	0	No remote work
2	50	Partially remote
3	100	Fully remote

- Tabla "company"

PK_id_company	company_location	company_size
1	DE	L
2	US	L
3	RU	M
4	RU	L
5	US	S
6	US	M
7	FR	L
8	AT	L
9	CA	L
10	UA	L
11	NG	L

- Tabla "country"

PK_id_country	code	country_name
1	AF	Islamic Republic of Afghanistan Afghanistan
2	AX	Åland Åland Islands
3	AL	Albania Albania
4	DZ	Algeria Algeria
5	AS	American Samoa American Samoa
6	AD	Andorra Andorra
7	AO	Angola Angola
8	AI	Anguilla Anguilla
9	AQ	Antarctica Antarctica
10	AG	Antigua and Barbuda Antigua and Barbuda
11	AR	Argentina Argentina
12	AM	Armenia Armenia

- Tabla "employee_level"

PK_id_employee_level	experience_level	experience_label	employment_type	employment_label
1	EN	Entry-level / Junior	FT	Full-time
2	SE	Senior-level / Expert	FT	Full-time
3	EX	Executive-level / Director	FT	Full-time
4	MI	Mid-level / Intermediate	FT	Full-time
5	EN	Entry-level / Junior	PT	Part-time
6	MI	Mid-level / Intermediate	PT	Part-time
7	MI	Mid-level / Intermediate	CT	Contract
8	SE	Senior-level / Expert	CT	Contract
9	EX	Executive-level / Director	CT	Contract
10	MI	Mid-level / Intermediate	FL	Freelance
11	SE	Senior-level / Expert	FL	Freelance
12	EN	Entry-level / Junior	CT	Contract

- Tabla "badge"

PK_id_badge	badge	country
EUR	Euro	Bandera de Unión Europea Eurozona en la Unión Europea
USD	Dólar estadounidense	Bandera de Estados Unidos Estados Unidos, incluyendo lo
CAD	Dólar canadiense	Bandera de Canadá Canadá
INR	Rupia india	Bandera de Bután Bután, Bandera de la India India
PLN	Złoty	Flag of Poland.svg Polonia
GBP	Libra esterlina	Bandera de Reino Unido Reino Unido, incluyendo las Depe
HUF	Forinto	Flag of Hungary.svg Hungría
SGD	Dólar de Singapur	Bandera de Singapur Singapur
MXN	Peso mexicano	Flag of Mexico.svg México
TRY	Lira turca	Bandera de Turquía Turquía
CLP	Peso chileno	Bandera de Chile Chile

Futuras líneas

Con este proyecto es posible impulsar un plan orientado a reconocer la evolución de los salarios para los empleos relacionados con el área de Data Science ya que el presente se basó en los datos recopilados en los años 2020 y 2021. Adicional, para el futuro proyecto es posible unificar la información y generar mayor valor a la información que se obtenga teniendo en cuenta su evolución cronológica y también mientras se adaptan dichas posiciones a la realidad de la empresa con adopción y la evolución tecnológica.