Resumen

Este analisis de datos tiene como objetivo identificar las condiciones que hacen que una persona especialista en analizar datos tenga un mejor sueldo. Para encontrar respuesta las principales preguntas planteadas, se utilizo una base de datos de Kaggle con información de personas especialistas en analisis de datos alrededor del mundo.

Introducción

Este analisis se enfoca en buscar respuesta por medio de vizualización de datos a las siguientes preguntas: ¿En qué países se ofrecen mejores salarios?, ¿Se han incrementado los salarios a lo largo del tiempo? y ¿Influye el nivel de experiencia en el salario?

1) EXPLORACIÓN DE LA BASE DE DATOS

A) Accede a la base de datos de Data Science Jobs Salaries: Aquí Descargar Aquí.

```
import pandas as pd
import plotly.express as px
import plotly.graph_objects as go

ds = pd.read_csv('ds_salaries.csv')
```

B) Explora las variables y familiarizate con su significado. La página de Data Science Jobs Salaries (Enlaces a un sitio externo.) de Kaggle te puede ser de utilidad.

```
#Identifica la cantidad de datos y variables presentes.
print("--------Columnas en data frame---------")
print(ds.columns)
```

```
--------Columnas en data frame---------
Index(['Unnamed: 0', 'work_year', 'experience_level', 'employment_type',
       'job_title', 'salary', 'salary_currency', 'salary_in_usd',
       'employee_residence', 'remote_ratio', 'company_location',
       'company_size'],
      dtype='object')
```

```
#Clasifica las variables de acuerdo a su tipo y escala de medición.
print("---------Tipos de variables---------")
print(ds.dtypes)

v_cuali = ['experience_level', 'employment_type', 'job_title', 'salary_currency', 'employe
v_cuant = ['salary', 'salary_in_usd']

```

```
---------Tipos de variables---------
Unnamed: 0             int64
work_year              int64
experience_level      object
employment_type       object
job_title             object
salary                 int64
salary_currency       object
salary_in_usd          int64
employee_residence    object
remote_ratio           int64
company_location      object
company_size          object
dtype: object
```

```
df_cuali = ds[['experience_level', 'employment_type', 'job_title', 'salary_currency', 'emp
df_cuanti = ds[['salary','salary_in_usd']]
```

C) Exploración de la base de datos

1. Calcula medidas estadísticas

* Variables cuantitativas
    - Medidas de tendencia central: promedio, media, mediana y moda de los datos.
    - Medidas de dispersión: rango: máximo - mínimo, varianza, desviación estándar.

```
# Medidas de tendencia central: promedio, media, mediana y moda de los datos.

mean = df_cuanti.mean()
median = df_cuanti.median()
mode = df_cuanti.mode()

print("---MEDIA---")
print(mean)
print("------------")
print("---MEDIANA---")
print(median)
print("------------")
```

```
print("---MODA---")
print(mode)
print("------------")

```

```
---MEDIA---
salary           324000.062603
salary_in_usd    112297.869852
dtype: float64
------------
---MEDIANA---
salary           115000.0
salary_in_usd    101570.0
dtype: float64
------------
---MODA---
   salary  salary_in_usd
0   80000       100000.0
1  100000            NaN
------------
```

```
#Medidas de dispersión: rango: máximo - mínimo, varianza, desviación estándar.

rango = df_cuanti.max() - df_cuanti.min()
varianza = df_cuanti.var()
des_est = df_cuanti.std()

print("---RANGO---")
print(rango)
print("------------")
print("---VARIANZA---")
print(varianza)
print("------------")
print("---DESVIACIÓN ESTANDAR---")
print(des_est)
print("------------")

```

```
---RANGO---
salary           30396000
salary_in_usd      597141
dtype: int64
------------
---VARIANZA---
salary           2.385040e+12
salary_in_usd    5.034933e+09
dtype: float64
------------
---DESVIACIÓN ESTANDAR---
salary           1.544357e+06
salary_in_usd    7.095726e+04
dtype: float64
------------
```

* Variables cualitativas
    - Tabla de distribución de frecuencia
    - Moda

```
# Tabla de distribución de frecuencia
cols = df_cuali.columns
for i in range(1, cols.size):
    print(df_cuali.groupby(cols[i]).agg(frequency=(cols[i], "count")))

```

```
                   frequency
employment_type
CT                         5
FL                         4
FT                       588
PT                        10
                                          frequency
job_title
3D Computer Vision Researcher                     1
AI Scientist                                      7
Analytics Engineer                                4
Applied Data Scientist                            5
Applied Machine Learning Scientist                4
BI Data Analyst                                   6
Big Data Architect                                1
Big Data Engineer                                 8
Business Data Analyst                             5
Cloud Data Engineer                               2
Computer Vision Engineer                          6
Computer Vision Software Engineer                 3
Data Analyst                                     97
Data Analytics Engineer                           4
Data Analytics Lead                               1
Data Analytics Manager                            7
Data Architect                                   11
Data Engineer                                   132
Data Engineering Manager                          5
Data Science Consultant                           7
Data Science Engineer                             3
Data Science Manager                             12
Data Scientist                                  143
Data Specialist                                   1
Director of Data Engineering                      2
Director of Data Science                          7
ETL Developer                                     2
Finance Data Analyst                              1
Financial Data Analyst                            2
Head of Data                                      5
Head of Data Science                              4
Head of Machine Learning                          1
```

```
Lead Data Analyst                          3
Lead Data Engineer                         6
Lead Data Scientist                        3
Lead Machine Learning Engineer             1
ML Engineer                                6
Machine Learning Developer                 3
Machine Learning Engineer                 41
Machine Learning Infrastructure Engineer   3
Machine Learning Manager                   1
Machine Learning Scientist                 8
Marketing Data Analyst                     1
NLP Engineer                               1
Principal Data Analyst                     2
Principal Data Engineer                    3
Principal Data Scientist                   7
Product Data Analyst                       2
Research Scientist                        16
Staff Data Scientist                       1
```

```
#MODA
df_cuali.mode()
```

	experience_level	employment_type	job_title	salary_currency	employee_residence	c
0	SE	FT	Data	USD	US	

2) Explora los datos usando herramientas de visualización

* Variables cuantitativas:
  - Medidas de posición: cuartiles, outlier (valores atípicos), boxplots
  - Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asi

```
#Medidas de posición: cuartiles, outlier (valores atípicos), boxplots
import plotly.express as px

print("---CUARTILES---")
print(df_cuanti.quantile([.25, .50, .75]))
print("---OUTLIER---")
q1 = df_cuanti.quantile(.25)
q3 = df_cuanti.quantile(.75)
iqr = q3 - q1
print(q3)
print("---BOXPLOTS---")
fig = px.box(df_cuanti, y= v_cuant)
fig.show()

```

```
---CUARTILES---
        salary  salary_in_usd
0.25   70000.0        62726.0
0.50  115000.0       101570.0
0.75  165000.0       150000.0
---OUTLIER---
salary           165000.0
salary_in_usd    150000.0
Name: 0.75, dtype: float64
---BOXPLOTS---
```

* Variables categóricas
    - Distribución de los datos (diagramas de barras, diagramas de pastel)

```
cols = df_cuali.columns
for i in range(1, cols.size):
    aux = df_cuali.groupby(cols[i]).agg(
    frequency=(cols[i], "count"))

    fig = px.bar(aux, x='frequency')
```

```
    fig.show()

```

3) Identifica problemas de calidad de datos (registros duplicados, valores faltantes, outliers, etc).

```
from scipy.stats import zscore
#Valores faltantes?
print(ds.isnull().values.any())
#Valores nulos?
print(ds.isna().values.any())

```

```
    False
    False
```

```
# Delete outliers
df_cuanti = df_cuanti[(zscore(df_cuanti) < 3).all(axis=1)]
print("---BOXPLOTS---")
fig = px.box(df_cuanti, y=v_cuant)
fig.show()

```

```
    ---BOXPLOTS---
```

D) Preparación de los datos: 1. Selecciona el conjunto de datos a utilizar.

```
    * Decide qué conjunto de datos se utilizará. Identifica variables objetivo. En caso necesario, ex

    * Maneja datos categóricos: transforma a datos numéricos si es necesario.

    * En caso de necesidad de recorte de datos (atípicos, faltantes, duplicados, etc), explica el mot

    * Maneja apropiadamente datos atípicos.
```

```
v_selec = ds[['experience_level','company_location', 'work_year', 'salary_in_usd']]
```

```
2. Transforma los datos en caso necesario.
    * Revisa si es necesario discretizar los datos
    * Revisa si es necesario escalar y normalizar los datos
    * Construye atributos si es conveniente
```

2) ANALIZA LOS DATOS Y CONTESTA TUS PREGUNTAS GUÍA

```

fig = px.histogram(v_selec, x='company_location', y='salary_in_usd', color="company_locati
fig.show()
```

```
fig = px.histogram(v_selec, x='work_year', y='salary_in_usd', color="work_year",
                   title='¿Se han incrementado los salarios a lo largo del tiempo?')
fig.show()

```

```
fig = px.histogram(v_selec, x='experience_level', y='salary_in_usd',
                   color="experience_level", title='¿Influye el nivel de experiencia en el
fig.show()

```

Productos pagados de Colab - Cancela los contratos aquí