DATA SCIENCE SALARIES

BY:
Aurora Corzas
Alejandro Chávez
Victor Olguín
Fernando Bueno

Pitch Deck

PROBLEM

DESCRIPTION

The objective is to make an interactive visualisation tool that gives information regarding salary, position, location, and other related factors in order to better inform career decisions related to Data Science.

Pitch Deck

QUESTIONS

This are the most insightful questions we would like to answer with our project.

- Where are the data scientist located?
- What role has the highest income?
- Is there a gender pay gap in data science?
- Top paying companies
- Does it make a difference to hold a higher degree?
- How much do years of experience affect the total compensation?
- Men/women rate per company
- More inclusive companies (according to ethnicity)

DATASET CARACTERISTICS AND EXPLORATION

Our Data set has a large variety of columns but it doesnt have a standard, in some columns

Exploring

First we needed to explore the data

Reading Dataset:

```
df = pd.read_csv("../Databases/project3_salaries_initial.csv")
df
```

Exploring General Info in the Dataset:

```
df.info()
```

Counting NaN values per column:

```
df.isnull().sum()
```

Companies distribution

```
df["company"].value_counts()
```

Position distributions

```
df["title"].value_counts()
```

Cleaning:

In order to clean we used:

Renaming Columns:

```
df.rename(columns={'rowNumber': 'rownumber', 'Masters_Degree' : 'mastersdegree',
```

Droping Columns:

```
df.drop(columns=["race", "education", "otherdetails", "highschool", "somecollege", "basesalary",
```

Estandardizing Time Stamps:

```
df["timestamp"] = pd.to_datetime(df["timestamp"]).dt.date
df
```

Keeping only the top 20 companys:

```
top_20_companies = df["company"].value_counts().nlargest(20).index
top_20_companies

top_20_df = df[df["company"].isin(top_20_companies)].copy()
top_20_df["totalyearlycompensation"].describe()
```

Adding ID's

```
top_20_df.insert(0, "id", range(1, len(top_20_df)+1))
top_20_df
```

UPLOADING OUR DATASET TO FLASK

```python
import csv
import sqlite3


INPUT_CSV_PATH = 'top20.csv'
OUTPUT_SQLITE_PATH = 'top20_pk.sqlite'


try:

    # Import csv and extract data
    with open(INPUT_CSV_PATH, 'r') as fin:
        dr = csv.DictReader(fin)
        all_info = [(i['id'], i['timestamp'], i['company'], i['level'], i['title'],
        print(all_info)

    # Connect to SQLite
    sqliteConnection = sqlite3.connect(OUTPUT_SQLITE_PATH)
    cursor = sqliteConnection.cursor()
```

We start by taking the CSV and turning it into a sqlite file, using the sqlite3 library

Then we use SQLAlchemy to create a variable that references our table in the sqlite file

```python
from flask import Flask, jsonify
from flask_cors import CORS

DB_PATH = "sqlite:///top20_pk.sqlite"

print(os.getcwd())
# python -m http.server [PORT]

#################################################
# Database Setup
#################################################
engine = create_engine(DB_PATH)


# reflect an existing database into a new model
Base = automap_base()
# reflect the tables
Base.prepare(autoload_with = engine)
      You, yesterday • Test de Flask …

top_20_companies = Base.classes.top20
```

```python
@app.route("/api/v1.0/id")
def id():
    # Create our session (link) from Python to the DB
    session = Session(engine)

    # Query all companies
    results = session.query(top_20_companies.id).all()

    session.close()

    # Convert list of tuples into normal list
    all_ids = [int(x) for x in np.ravel(results)]

    return jsonify(all_ids)
```

Finally we use flask to create our api endpoints, referencing our table variable

WHERE ARE THE DATASCIENTIST LOCATED?

The data provides a broad overview of the average yearly compensations across different cities globally. As expected, regions with a strong tech presence and higher living costs tend to have more people. Meanwhile, cities in developing countries, despite having robust tech sectors, tend to have lower cocentrations of DS due to economic differences and cost of living.

HOW IS GENDER DISTRIBUTED

- Overall, the data suggests that men, on average, earn more than women, especially as years of experience increase.
- There are, however, certain anomalies or specific years where women's compensation surpasses that of men, but these seem to be exceptions rather than a consistent trend.

TOP PAYING COMPANYS

META

LINKEDIN

GOOGLE

UBER

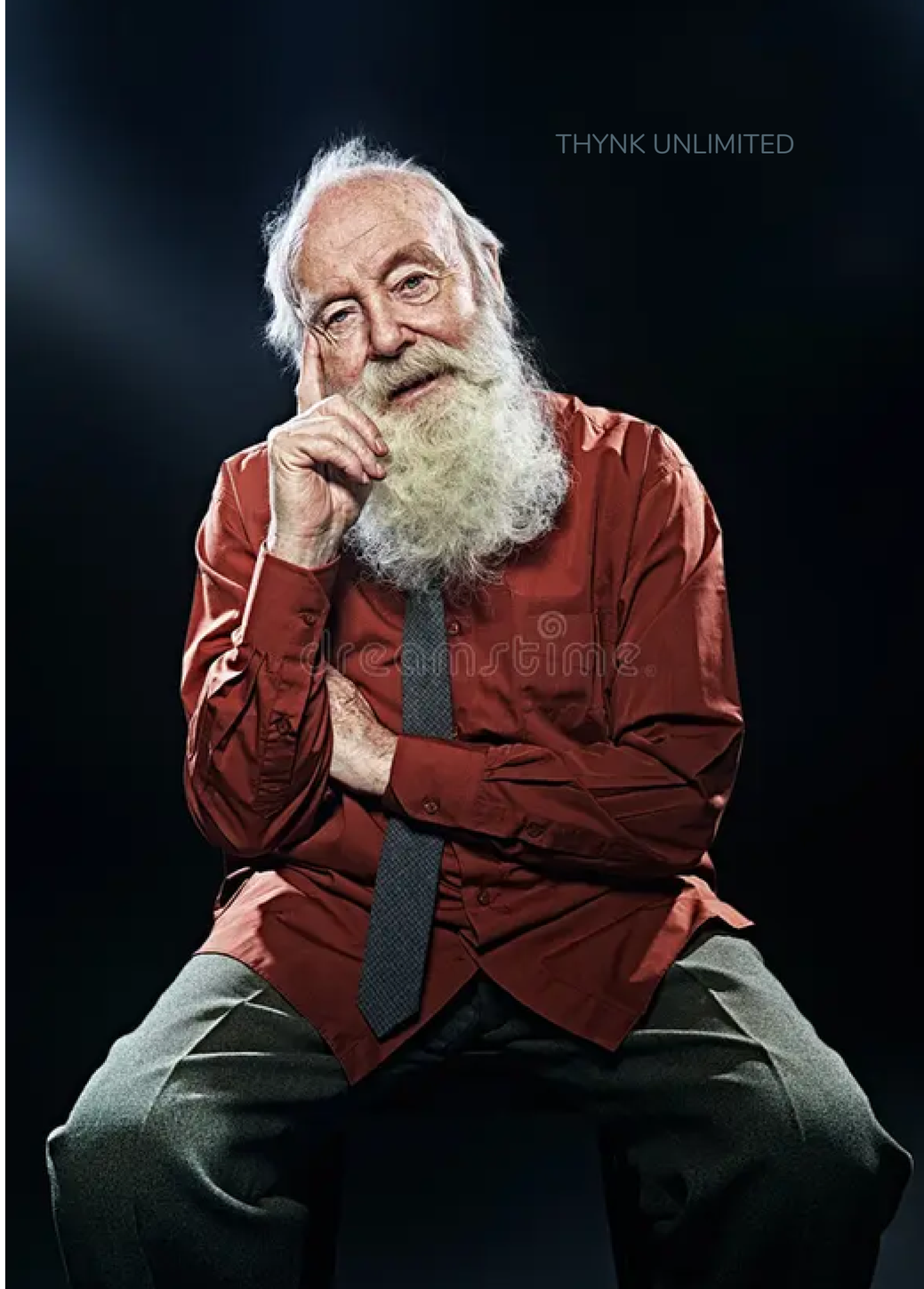
HOW MUCH DO YEARS OF EXPERIENCE AFFECT THE TOTAL COMPENSATION?

In 1-10 YEARS There's the largest growth.
154,233.64- $279,237.55

37 years of experience at $678,666.67.

Lowest Compensation $154,233.64.

Through the years the growth decreases or tends to be more variable



MOST COMMON JOB TITLES OR POSITIONS

1. Software Engineer 
2. Data scientist 
3. Product Manager 

MORE INCLUSIVE COMPANIES

Distribution:
Asian employees are the majority (6,212).
White employees come second (3,614).
Hispanic (617), Black (373), and Two or More races (386) follow.

Top Performers:
Amazon, Microsoft, and Google have the highest numbers across all racial categories.
Underrepresented Groups

Smallest groups
Black and Hispanic employees are underrepresented in most companies.

Key Observations
Many tech companies have higher Asian representation.
Black representation is notably low in companies like Apple and Bloomberg.
Recommendation

Pitch Deck

DOES IT MAKE A DIFFERENCE TO HOLD A HIGHER DEGREE?

Degree	Income
Bachelors	$ 234,723.22
Masters	$ 242,799.31
Doctorate	$ 309,015.74

Summary:

- The percentage increase from Bachelor's to Master's is approximately 3.44%.
- The percentage increase from Master's to Doctorate is approximately 27.27%.

CONCLUSION

In conclusion, for an optimal career path: pursue what you're passionate about, target US or EU companies that resonate with you, continuously upgrade your skills through courses and certifications, and remember – patience is key. Experience not only brings enhanced salaries but also greater career prospects!