DATA SCIENCE INTERNSHIP

Project 1 – Data Science Job Salaries

UMID03072548194

Anjali Shibu

anjalishi1994@gmail.com

Contents

Introduction

The given dataset contains data science job salaries. I have downloaded it as a csv file 'Data Science Job Salaries' and uploaded in **Colab** environment .**Python** language is used.The dataset is loaded as pandas dataframe, analysed, cleaned, and EDA carried out to gain insights. This project also experiments with natural language processing (NLP) using titles to cluster job titles using unsupervised learning. At the end, we also use a predictive technique to predict data science job salaries in the upcoming year.

Import necessary libraries in Python

Import libraries

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import plotly.express as px
%pip install pycountry
from wordcloud import WordCloud
sns.set_style('whitegrid')
from scipy import stats
import statsmodels.api as sm
from statsmodels.formula.api import ols
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
%pip install prophet
from prophet import Prophet
```

We import the libraries needed to process our data as the first step.

Loading the dataset

```
[163] df = pd.read_csv('/content/Data Science Job Salaries.csv')
      df.head()
```

	Unnamed: 0	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	L
1	1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0	JP	S
2	2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	M
3	3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN	0	HN	S
4	4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US	50	US	L

Dataset is successfully loaded.We see that initial column has no name.

Data attributes

We use df.info() to find dataset attributes.We also make sure there no null values.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Unnamed: 0          607 non-null    int64 
 1   work_year           607 non-null    int64 
 2   experience_level    607 non-null    object
 3   employment_type     607 non-null    object
 4   job_title           607 non-null    object
 5   salary              607 non-null    int64 
 6   salary_currency     607 non-null    object
 7   salary_in_usd       607 non-null    int64 
 8   employee_residence  607 non-null    object
 9   remote_ratio        607 non-null    int64 
 10  company_location    607 non-null    object
 11  company_size        607 non-null    object
dtypes: int64(5), object(7)
memory usage: 57.0+ KB
```

Data cleaning and preprocessing

In this process, we check if there are duplicate rows.Sorted by columns to verify redundancy.

Check for duplicates

```
[165]
df[df.duplicated(keep=False)].sort_values(by=list(df.columns))
```

	Unnamed: 0	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size

```
[166] df = df.drop_duplicates()
```

```
[167] print("Duplicate rows:", df.duplicated().sum())
```

```
Duplicate rows: 0
```

This dataset did not have any duplicate rows.

Column removal :

Next step, we drop columns that do not contribute meaningful insights.We remove the ‘unnamed :0’ column as it is serial number,kind of an index value.

```
df = df.drop(columns=['Unnamed: 0']) # drop
```

```
df['work_year']= df['work_year'].astype('category')
```

```
columns_to_convert = ['experience_level', 'employment_type','job_title','salary_currency','employee_residence','remote_ratio','company_size','company_location']
df[columns_to_convert] = df[columns_to_convert].astype('category')
```

Data type optimization:

Some attributes with object type is converted into category type to be more efficient. Attributes that were converted are ‘experience_level’,’employment_type’,’job_title’,’salary_currency’,’employee_residence’,’remote_ratio’,’company_size’,’company_location’.Salary_in_usd will be used for further processing as numerical value.

Conversion reduced memory usage significantly. Improved computational efficiency for categorical operations. Maintained all categorical information while optimizing storage.

Unique values

To understand the distinct values present in categorical and key numerical columns, ensuring data consistency and identifying any anomalies.

Check unique values in the given data

```
68] unique_cols = ['work_year','job_title','experience_level','employment_type','company_location','company_size','employee_residence','remote_ratio','salary_currency']
    for col in unique_cols:
        unique_values = df[col].unique()
        print(f"Unique values in {col}:")
        print(unique_values)
        print('~'*50)
```

The output of the code gave distinct values.

- The work_year consists of years from 2020-2022.
- Job_title consists of 50 job titles.
- Experience level consists of 4 unique values-Entry,mid-level,senior level and executive level.
- Employment type contains part-time(PT),full-time(FT),contract(CT),freelance(FL)
- Company_location – 50 countries are available in the data.
- Company size – 3 unique values- Large ,small,medium-L,S,M
- Employment_residence-57 residence countries exist.
- Remote_ratio = 0(full on-site),50,100(full remote) 3 unique values
- Salary_currency = 17 values. We mainly use salary_in_usd for calculations.

Descriptive statistics

This gives a statistical summary of salary data

Dataset has 607 records.

Salary(USD) Distribution :

Mean $112298

Median(50th percentile):$ 101570

Standard deviation : $70957 (high variability)

Range:$2859(min) to $600000(max)

IQR = 150000 – 62726 = 87274.

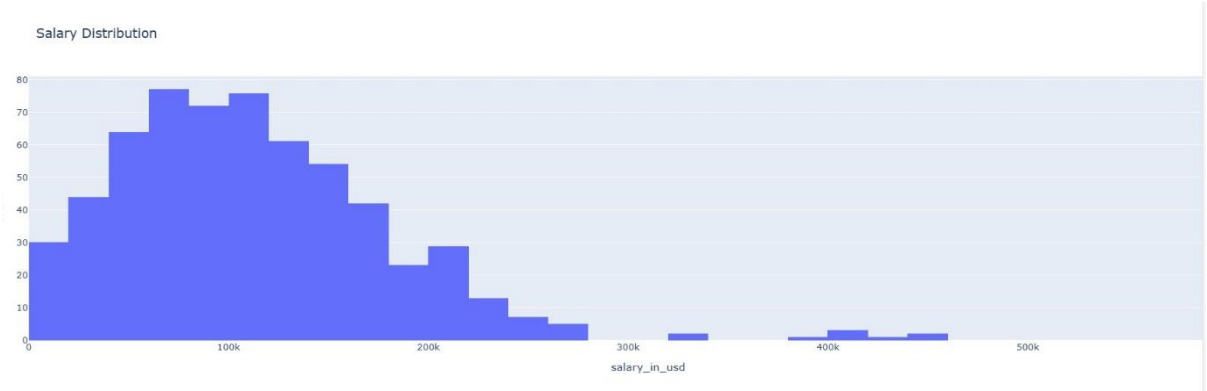
Here mean > median. Right-skewed distribution. 75% of salaries fall below $150000.

```
df.describe()
```

	salary	salary_in_usd
count	6.070000e+02	607.000000
mean	3.240001e+05	112297.869852
std	1.544357e+06	70957.259411
min	4.000000e+03	2859.000000
25%	7.000000e+04	62726.000000
50%	1.150000e+05	101570.000000
75%	1.650000e+05	150000.000000
max	3.040000e+07	600000.000000

EDA

Histogram



Histogram shows the frequency distribution of salary_in_usd (in thousands).

Peak concentration : Most salaries cluster in the lower range(left side of the graph).

A right-skewed distribution (long tail toward higher salaries).

Majority of roles fall under moderate salary brackets (likely early-career to mid-level positions).

Outliers Present :

A small number of high salaries extend beyond the main concentration.

Word cloud

Word cloud is a visually striking way to represent text data, where the size of each word indicates its frequency or importance. It helps to summarize-instantly see which words pop

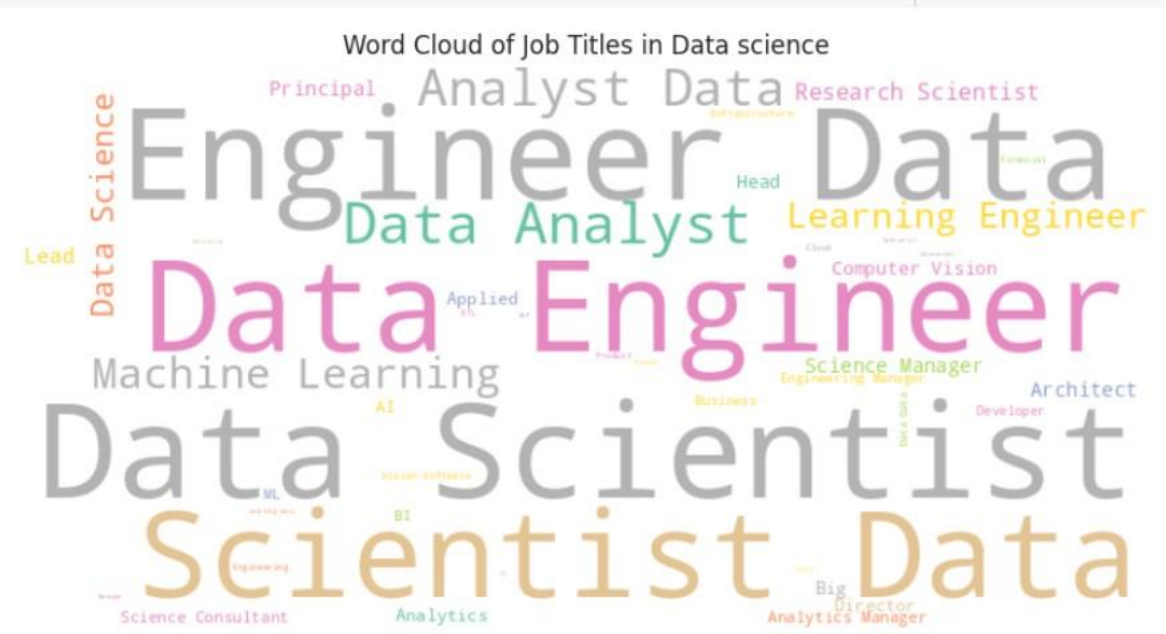
up most often, visualize-turn dry text into engaging graphics and explore-spot trends, themes or outliers in the language.

Job title analysis

1)Most common roles : Highlight ‘Data Scientist’,’Data Engineer’ and ‘Analyst’ appear prominently, showing these are core positions in the field.

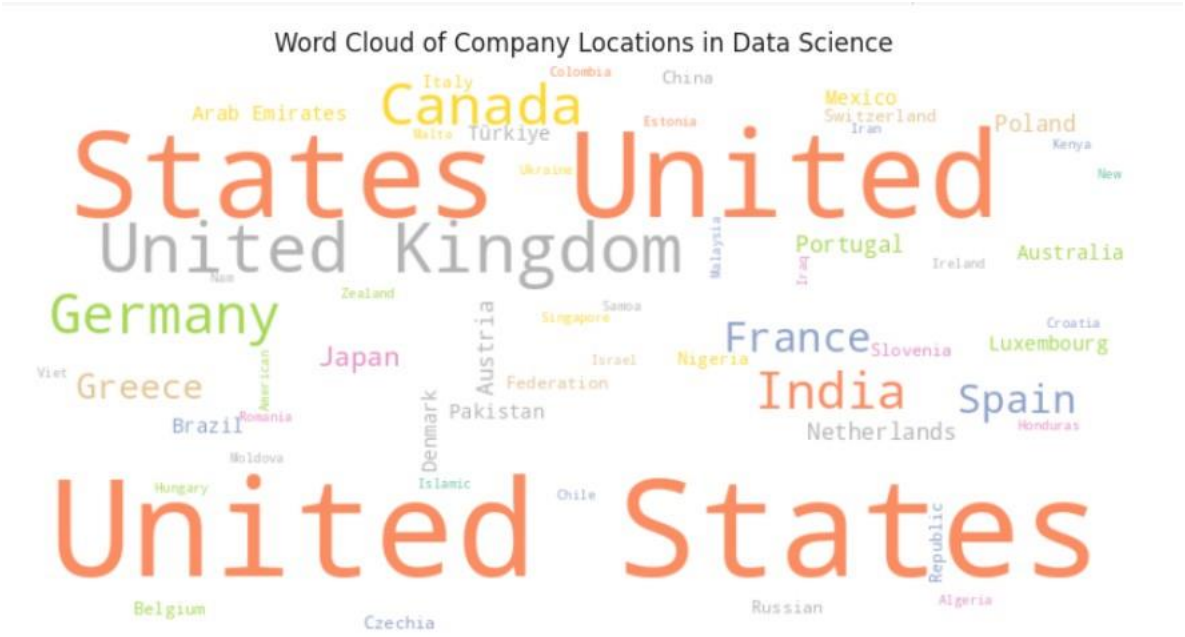
2) Specializations: Emergining specialities like Computer vision and analytics manager appear in the cloud.

3)Hierarchy : Point out senior roles like ‘Principal’,’ Manager’ and ‘Architect’ that demonstrate career progression paths in data science.

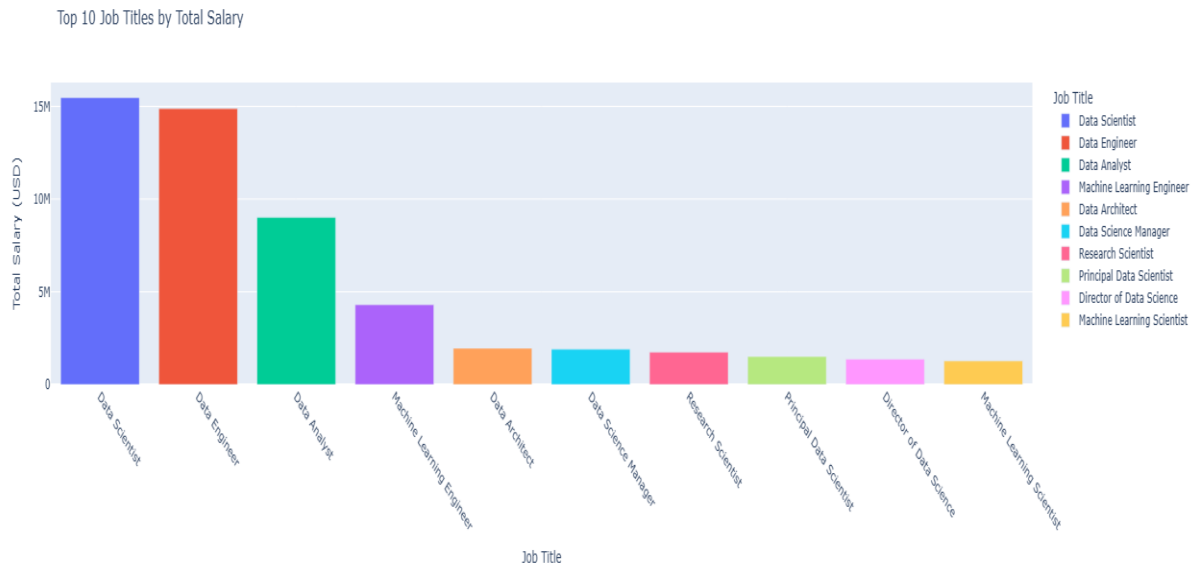


Company location analysis:

1) Geographic distribution: the US and UK appear multiple times, indicating they are major hubs for data science opportunities.
2) Global Presence: Mention the diversity of locations from Brazil to Australia to UAE, showing data science is a global field.
3) Emerging Markets: Note the presence of countries like Kenya and Iran, suggesting data science is growing in developing economies.



Top 10 titles by job salary



From the bar graph,we find:

1) Highest paying roles :

Director-level positions (Director of Data Science) and principal roles (Principal Data Scientist) command the highest salaries as expected.

Specialized engineering roles(Machine learning engineer,data architect) rank highly.

2) Core data roles dominance:

Data scientist,data engineer,data analyst, all appear in the top 10.

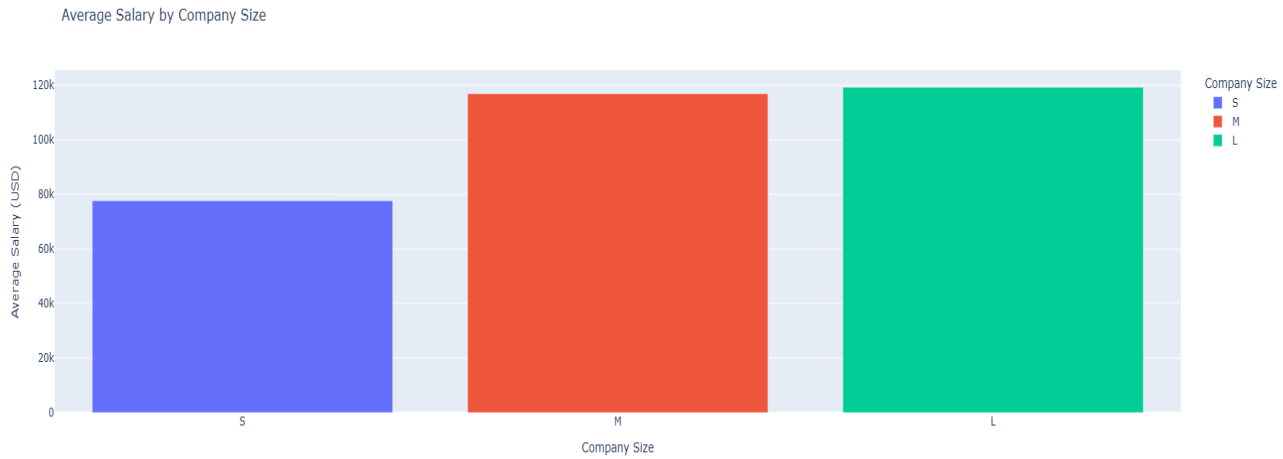
Data scientist appears first, reinforcing its position as the flagship role in the field.

3) Emerging specializations:

Machine learning appears twice (Engineer and scientist variants)

Research-focused roles (Research Scientist) make the list.

Average salary by company size



From the bar graph,it is seen that :

Small companies salary distribution is $60k-$80k average.Medium companies is around $80K-100k average.Large companies is around $100k-120k average. There exist clear positive correlation between company size and compensation.

Salary by experience level

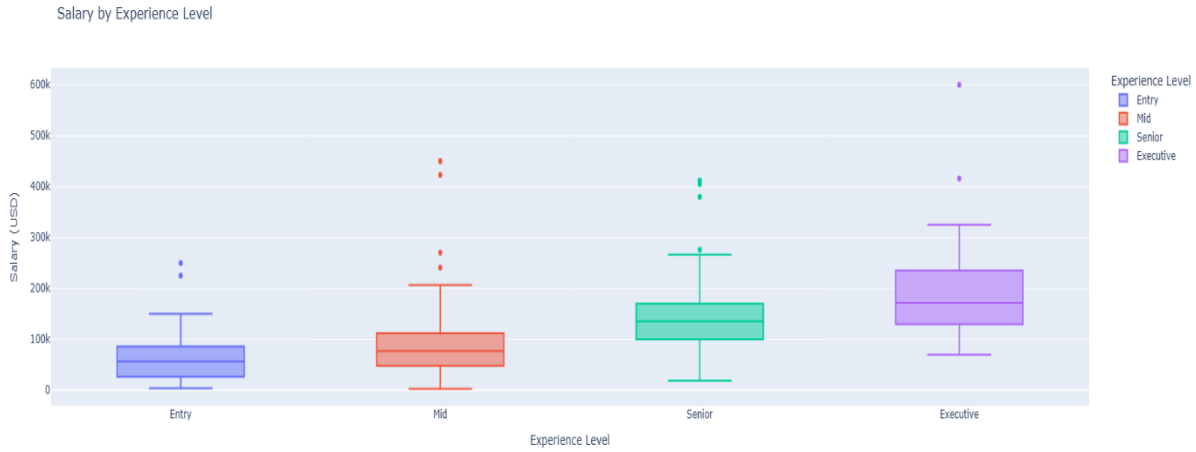
The data reveals a clear exponential growth pattern in compensation as professionals advance through their data science careers.

Median salary grows 3x from entry ($56.5k) to Executive ($171.4k)

90th percentile(upper fence) grows 5x from entry ($150k) to Executive ($325k)

Entry level – Tightest distribution (IQR : $26.6k-$85.8k)

Executive: Widest range($69.7k-$600k) ,reflecting performance based pay.



Career Planning:

Most significant jump occurs between mid to senior level roles(+76% median growth)

Executive roles offer 2.4x higher 75th percentile pay vs. senior roles.

Outlier interpretation:

Extreme high like $600k executives likely reflect:

Equity compensation in tech startups.

C-level positions at Fortune 500 companies .

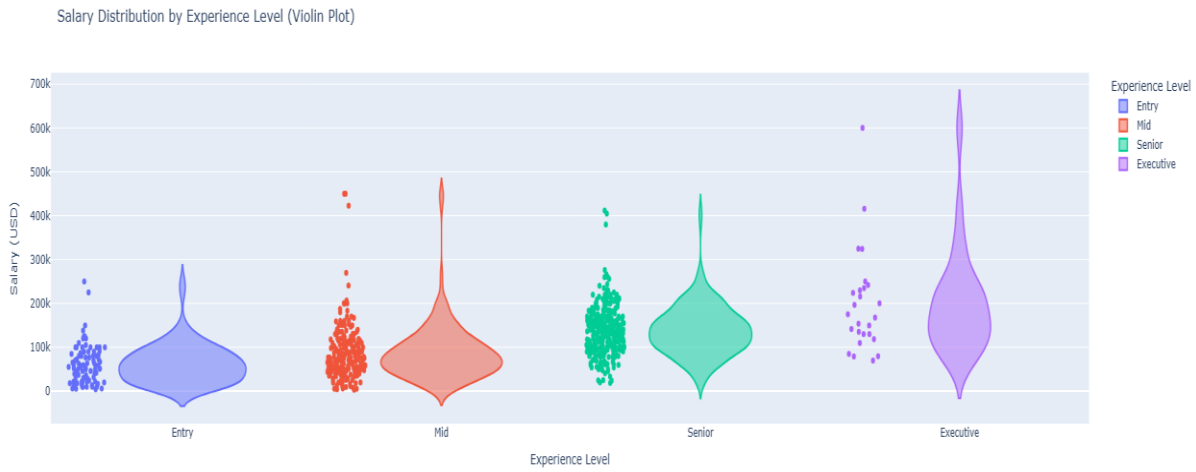
Average salary by experience level

The bar graph reveals a clear accelerating growth pattern in early-mid career(entry-senior) with a plateau effect at executive levels. This suggests the most lucrative skill investments happen before reaching seniority.

Job seekers should aim for promotions every 2-3 years early career to maximize the 50%+ jumps.

For employers,retention efforts should focus on mid-level talent.

Density of salary in experience level



The violin plot reveals what the bar chart hides :

Entry-lives salaries are rigid, but executive pay follows a power law.

The real salary growth happens from mid to senior transition where distribution splits.

For **job seekers** in early careers= focus on skill building rather than salary jumps.

Mid-career- specialise or transition to leadership for higher pay bands.

Senior – Target equity/performance bonuses, base salary plateaus.

For **employers**:

Retention risk : mid-level bulge suggests high competition for skilled individual contributors.

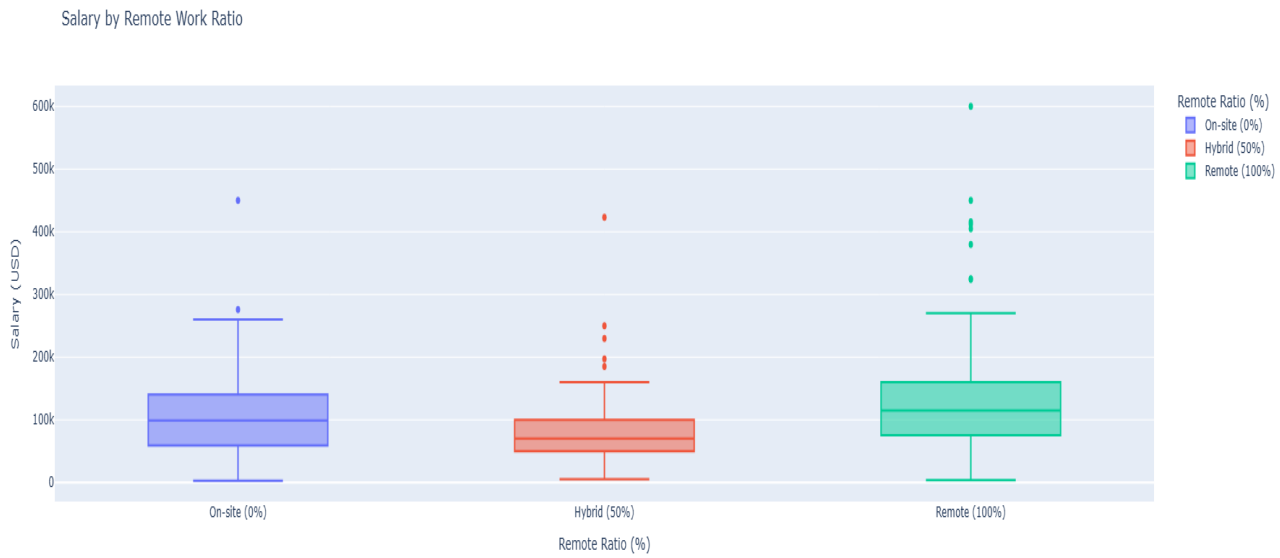
Executive hiring: Wide range means top talent demands outlier packages.

Remote work v.s salary

The chart illustrates the relationship between salary levels and the degree of remote work, categorized into three work arrangements: On-site (0% remote), Hybrid (50% remote) and

fully remote (100% remote). The vertical axis represents salary (in USD), while the horizontal axis shows the remote work ratio.

Average Salary by Remote Ratio:
remote_ratio
On-site (0%) 106354.622047
Hybrid (50%) 80823.030303
Remote (100%) 122457.454068



1. Fully Remote (100% remote):

Highest average salary ($122457), suggesting a premium for remote work.This may reflect demand for specialized skills or cost saving for employers (eg, no geographic constraints).

2. On-site Roles (0% remote)

Second highest average salary ($106355), potentially tied to industries that require physical presence (e.g manufacturing, healthcare) or location based pay adjustments (e.g : high-cost urban areas).

3. Hybrid roles (50% remote)

Lowest average salary($ 80823), which could indicate roles with less specialization, mid-career positions, or employers transitioning to remote policies without full salary parity.

The given dataset challenges the assumption that on-site roles always pay more, highlighting the growing value of remote work in certain sectors.

The hybrid salary dip may warrant further investigation (e.g are these roles concentrated in lower- paying industries or junior levels?)

Geographic and seniority breakdowns could clarify discrepancies(e.g remote salaries may skew higher if they include senior tech workers)

Salary trend over time

The line graph depicts a steady upward trajectory in average salaries for data science roles over a three year period.It reflects growing demand and valuation of data expertise in the job market.

1)2020

Average salary : $ 95813. Baseline year, likely impacted by initial pandemic uncertainty, with hiring slowdowns in some sectors.

2)2021

Average salary : $ 99854(+4.2% YoY growth). Gradual recovery and increased reliance on data-driven decision making as businesses adapted to remote work.

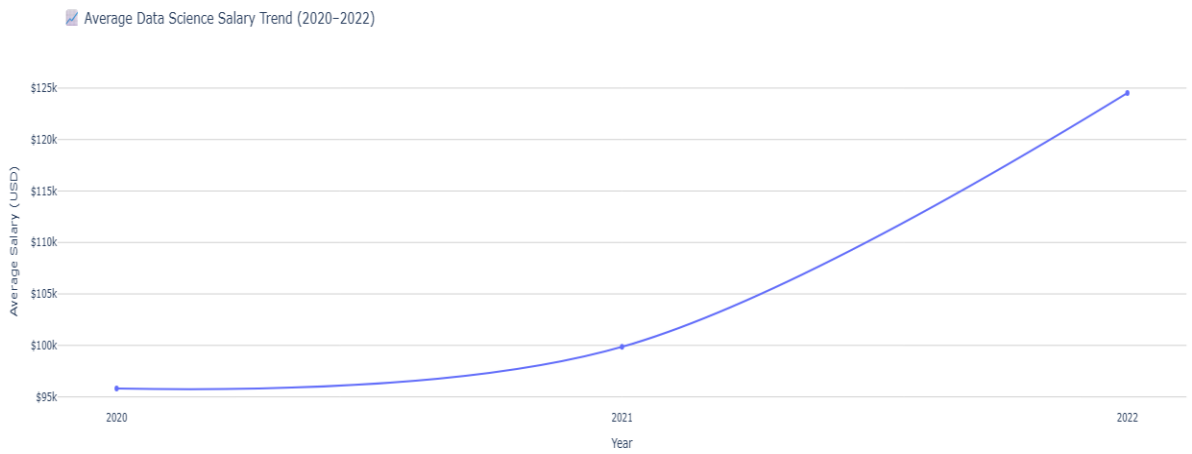
3)2022

Average salary: $124522(+24.7% YoY growth). Sharp surge, likely driven by:

High demand for AI/ML, analytics, automation skills. Talent shortages and competitive hiring in tech and adjacent industries. Inflation-adjusted salary adjustments or premium for specialized roles.

Implications:

The 22% cumulative growth from 2020 to 2022 underscores data science's escalating strategic importance. The 2022 spike may correlate with broader tech salary trends or a post-pandemic war for talent.



Salary distribution by remote work and experience level-Box plot and ANOVA

The box plot reveals how salaries vary across experience levels (Entry,Mid,Senior,Executive) and remote work categories (On-site, Hybrid,Remote). Key observations:

- **Remote work premium:**

Fully remote roles (100%) consistently show higher median salaries across all experience levels compared to hybrid or on-site roles, with the gap widening at senior/executive levels.

Executive remote roles dominate the upper quartile (up to $500k+), suggesting remote work is increasingly tied to high-impact, leadership positions.

- **Experience-level progression:**

Entry-level : Tight salary ranges (~$50k-$150k), with minimal variation by remote ratio.

Senior/ Executive : Wider dispersion , especially for remote roles, indicating greater pay negotiation leverage or specialization premiums.

- **Hybrid anomaly:**

Hybrid roles (1-50% remote) often fall between on-site and remote salaries but show less variability, possibly reflecting standardized policies in transitioning organizations.

```
ANOVA Results:
                                        sum_sq     df          F  \
C(experience_level)                 6.578716e+11    3.0  59.080451
C(remote_category)                  5.698862e+10    2.0   7.676833
C(experience_level):C(remote_category)  4.293222e+10    6.0   1.927773
Residual                            2.208478e+12  595.0        NaN

                                        PR(>F)
C(experience_level)                 1.926800e-33
C(remote_category)                  5.108389e-04
C(experience_level):C(remote_category)  7.425576e-02
Residual                                  NaN
```

Statistical significance (ANOVA)

The ANOVA tests whether experience level, remote category, and their interaction significantly explain salary variance.

- Main effects
  1. Experience level (p=1.926e-33) : Highly significant, confirming seniority drives salary differences.
  2. Remote category (p = 5.11e-4) : Statistically significant , validating that remote work arrangements independently affect pay.
- Interaction effect(p = 0.074) : Marginally insignificant but suggests trends like remote senior roles earn disproportionately more may warrant deeper analysis (e.g post-hoc tests).
- Effect-size:
  Remote category's sum of squares (6.58e+11) is substantial, though dwarfed by experience level (2.21e+12), emphasizing seniority primacy.

  Implications:
  Remote work as a salary accelerator : Fully remote roles command higher pay , especially for senior talent , likely due :
  Access to global talent pools (employers competing for top-tier candidates).
  Cost savings from relocation/office space reinvested in salaries.

  Hybrid's middle ground : Hybrid policies may standardize pay, potentially limiting upside for top performers.

  Actionable insights :
  For employers : Remote flexibility can attract elite talent but requires competitive compensation.
  For employees : Negotiating remote options may yield higher returns at senior levels.

Geographic analysis –Average salary by country

The bubble chart visualizes the average salary distribution across countries, where bubble size represents the salary magnitude in USD. The general trends can be interpreted as follows:

1. Salary range:

- The vertical axis (Salary_in_usd) shows a spread from $20k to $140k, indicating significant global disparities in compensation.
- Larger bubbles (higher salaries) are concentrated at the top, while smaller bubbles (lower salaries) cluster toward the bottom.

2. Regional patterns :

- High-Income countries (eg. US,UK,Germany,Switzerland,Russia): Likely occupy the upper tier ($100k+), reflecting strong demand for tech/executive talent and higher living costs.
- Emerging markets (eg. India, Brazil,Poland) : Likely appear in the mid-lower range ($40k-$80k),aligning with localized pay scales and cost of labor.
- Outliers : Some bubbles may deviate due to niche industries (eg. Oil/gulf countries) or currency conversion artifacts.

3. Bubble size variation

Disproportionately large bubbles at higher salary levels suggest non-linear pay gaps (eg. A $140k salary may be 3-5x the median in certain regions).

Implications:

Talent Migration : Countries with smaller bubbles may experience brain drain to higher-paying regions.

Remote work equity : Global companies could leverage salary disparities to attract remote talent cost-effectively (eg. Hiring mid-level roles from lower-cost countries).

Data limitations : Cross referencing with GDP or industry hotspots would strengthen insights.

Job title clustering using Silhouette score,K-means,PCA

Optimal number of clusters: 6 (silhouette score: 0.15)

Cluster 0:
['Machine Learning Scientist', 'Machine Learning Engineer', 'Machine Learning Manager', 'Machine Learning Infrastructure Engineer', 'Machine Learning Developer', 'Applied Machine Learning Scientist', 'Head of Machine Learning', 'Lead Machine Learning Engineer']
Categories (50, object): ['3D Computer Vision Researcher', 'AI Scientist', 'Analytics Engineer', 'Applied Data Scientist', ..., 'Principal Data Scientist', 'Product Data Analyst', 'Research Scientist', 'Staff Data Scientist']

Cluster 1:
['Data Scientist', 'Lead Data Scientist', 'Research Scientist', 'AI Scientist', 'Principal Data Scientist', 'Applied Data Scientist', 'Staff Data Scientist']
Categories (50, object): ['3D Computer Vision Researcher', 'AI Scientist', 'Analytics Engineer', 'Applied Data Scientist', ..., 'Principal Data Scientist', 'Product Data Analyst', 'Research Scientist', 'Staff Data Scientist']

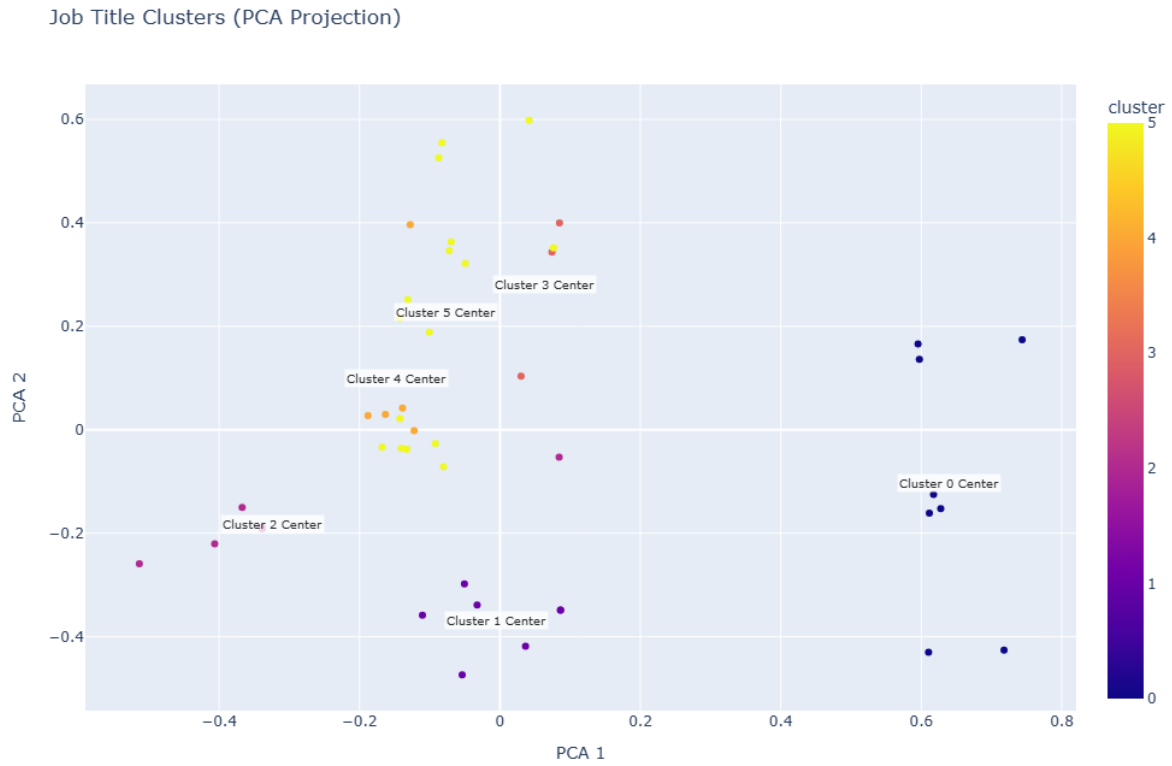
Cluster 2:
['Product Data Analyst', 'Data Analyst', 'Business Data Analyst', 'Lead Data Analyst', 'BI Data Analyst', 'Marketing Data Analyst', 'Financial Data Analyst', 'Finance Data Analyst', 'Principal Data Analyst', 'ETL Developer']
Categories (50, object): ['3D Computer Vision Researcher', 'AI Scientist', 'Analytics Engineer', 'Applied Data Scientist', ..., 'Principal Data Scientist', 'Product Data Analyst', 'Research Scientist', 'Staff Data Scientist']

Cluster 3:

['Computer Vision Engineer', '3D Computer Vision Researcher', 'Computer Vision Software Engineer']
Categories (50, object): ['3D Computer Vision Researcher', 'AI Scientist', 'Analytics Engineer', 'Applied Data Scientist', ..., 'Principal Data Scientist', 'Product Data Analyst', 'Research Scientist', 'Staff Data Scientist']

Cluster 4:
['Data Science Consultant', 'Director of Data Science', 'Data Science Manager', 'Data Science Engineer', 'Head of Data Science']
Categories (50, object): ['3D Computer Vision Researcher', 'AI Scientist', 'Analytics Engineer', 'Applied Data Scientist', ..., 'Principal Data Scientist', 'Product Data Analyst', 'Research Scientist', 'Staff Data Scientist']

Cluster 5:
['Big Data Engineer', 'Lead Data Engineer', 'Data Engineer', 'Data Engineering Manager', 'ML Engineer', ..., 'Data Architect', 'Big Data Architect', 'Analytics Engineer', 'NLP Engineer', 'Data Analytics Lead']
Length: 17
Categories (50, object): ['3D Computer Vision Researcher', 'AI Scientist', 'Analytics Engineer', 'Applied Data Scientist', ..., 'Principal Data Scientist', 'Product Data Analyst', 'Research Scientist', 'Staff Data Scientist']



1. Methodology

- Algorithm : K-means clustering with PCA (Principal component analysis) for dimensionality reduction.

- Optimal clusters 6 determined by Silhouette score=0.15, indicating weak but discernible structure.
- Data : Job titles from data/tech roles , grouped by semantic and functional similarity.

2. Insights and implications

- Technical vs. Business Alignment:

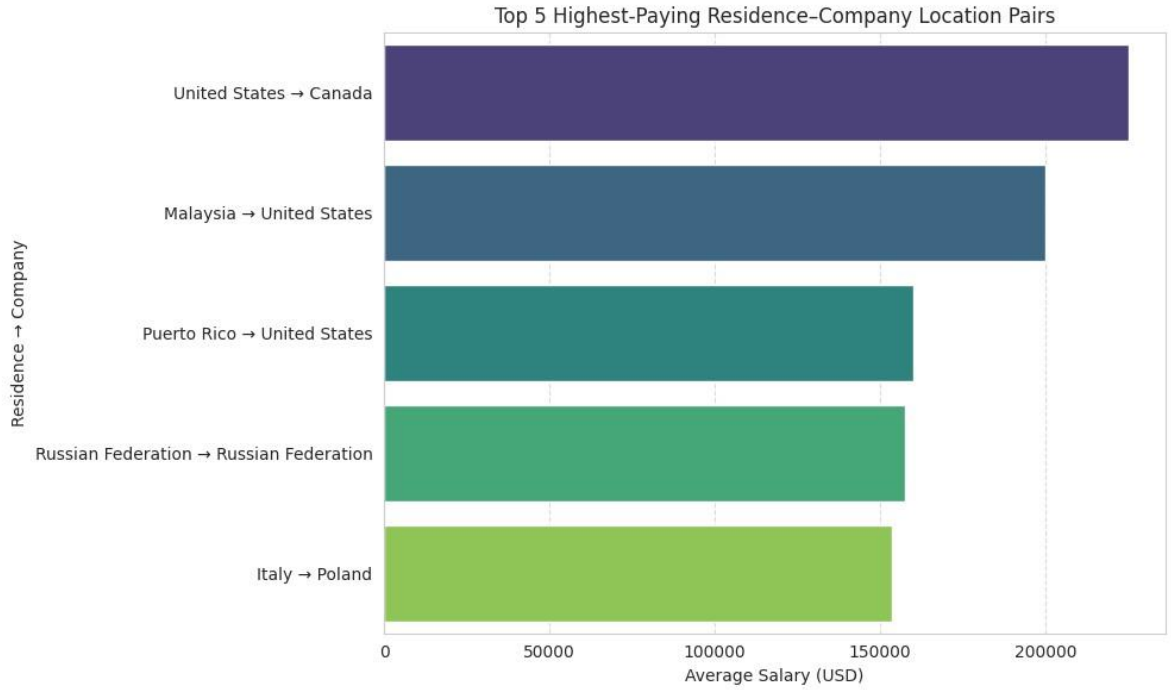
  Cluster 0,3 and 5(ML,Computer vision, data engineering) emphasize technical skills (e.g algorithms, distributed systems).
  Cluster 2 and 4 (Analytics, Leadership) align with business outcomes and team management.

- Salary Potential:

  Highest-paying clusters likely correlate with technical specialization (e.g. ML, Computer Vision) and leadership roles .
  Analytics roles (cluster 2) may show narrower salary bands due to standardized business demands.

- Hybrid roles:
  Cluster 5 includes 'ML Engineer' and 'Analytics Engineer' reflecting blurred lines between data science and engineering.
  Actionable use cases:
  HR/Talent acquisition : Tailor recruitment strategies by cluster (eg, niche skills for ML v.s broad analytics for Cluster 2)
  Career pathing : Identify transition opportunities (eg. Data analyst-data scientist-ML Engineer)

Highest paying top 5 Employee-Location pair



The visualization highlights the highest-paying geographic pairings between employee residence and company location, revealing unique cross-border and domestic compensation trends.

- Highest average salary $225k USD is for **US-Canada**. Likely reflect tech or finance roles where US companies hire. Canadian talent remotely (or vice versa) ,leveraging cost arbitrage or specialized skills.
- Second highest average salary $200k USD for **Malaysia-US**. Suggests US companies sourcing high-value roles (eg. Software engineering,data science) from Malaysia, possibly due to competitive talent pools or regional hubs.
- **Puerto Rico- US** has notable salary range of average $160k USD. May indicate tax incentives (eg. Puerto Rico's Act 60) or niche industries (eg. Biotech, crypto) driving premium pay.
- **Russian Federation- Russian Federation.** Domestic pairing with salaries average $ 157500 USD.Could represent local tech/energy sectors or senior roles in multinationals Russian offices.
- **Italy-Poland** has an average of $153667 USD. Might reflect Eastern Europe tech hubs (eg. Poland's growing IT sector) attracting Italian talent or companies.

Implications:

Remote work globalization : High-paying cross-border pairs (eg. US- Canada) underscore the rise of remote talent arbitrage , where companies tap into cost-effective or specialized labour markets.

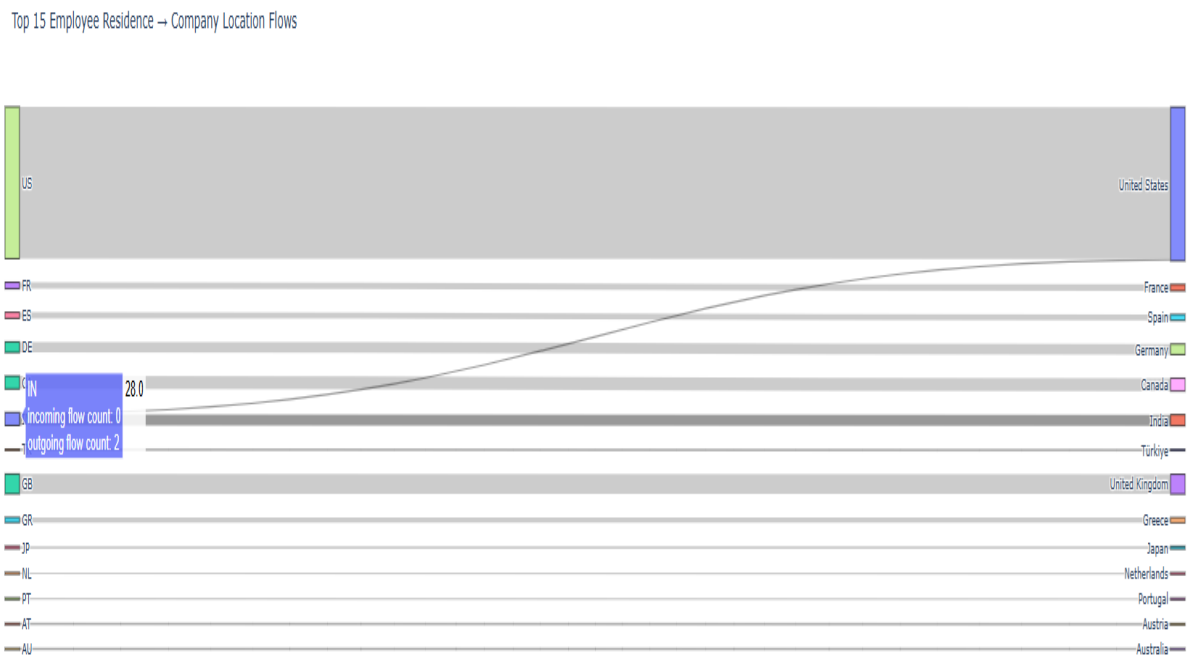
Tax and policy influence : Pairings like Puerto Rico- US highlight how tax policies can shape compensation strategies.

Domestic Premiums : Russia's domestic high pay may signal localized demand for scarce skills (eg. Cybersecurity,oil/gas tech).

Recommendations:

- For Employers : Use these insights to benchmark salaries for remote or international hires.
- For Professionals : Target companies in high-paying residence-company pairs(eg. US firms hiring remotely from Canada)

Top 15 employee residence flow - Sankey diagram



This visualization maps the most common international work arrangements, showing where the employees reside versus where their companies are located. The flows highlight globalization trends in remote work, outsourcing and talent migration.

<u>Observations:</u>

Dominant hubs:

US is the most frequent company location, attracting employees from Canada,India,UK,France,Germany (likely due to tech/finance roles).

European crossflows : Strong connections between Germany(DE)-France(FR), Netherlands(NL)-Portugal(PT).

Emerging Patterns:

India-US : Reflects outsourcing or remote hiring in tech/engineering.

Turkey-EU countries : May indicate regional talent mobility in manufacturing or IT.

Implications :

Remote work expansion : Companies increasingly hire across borders, with the US and EU as central nodes.

Talent Competition : Countries like Canada and India feed into high-demand markets (US/EU), potentially creating salary pressure.

Policy impact : Visa programs (eg. H-1B in the US, Blue Card in EU) likely influence these flows.

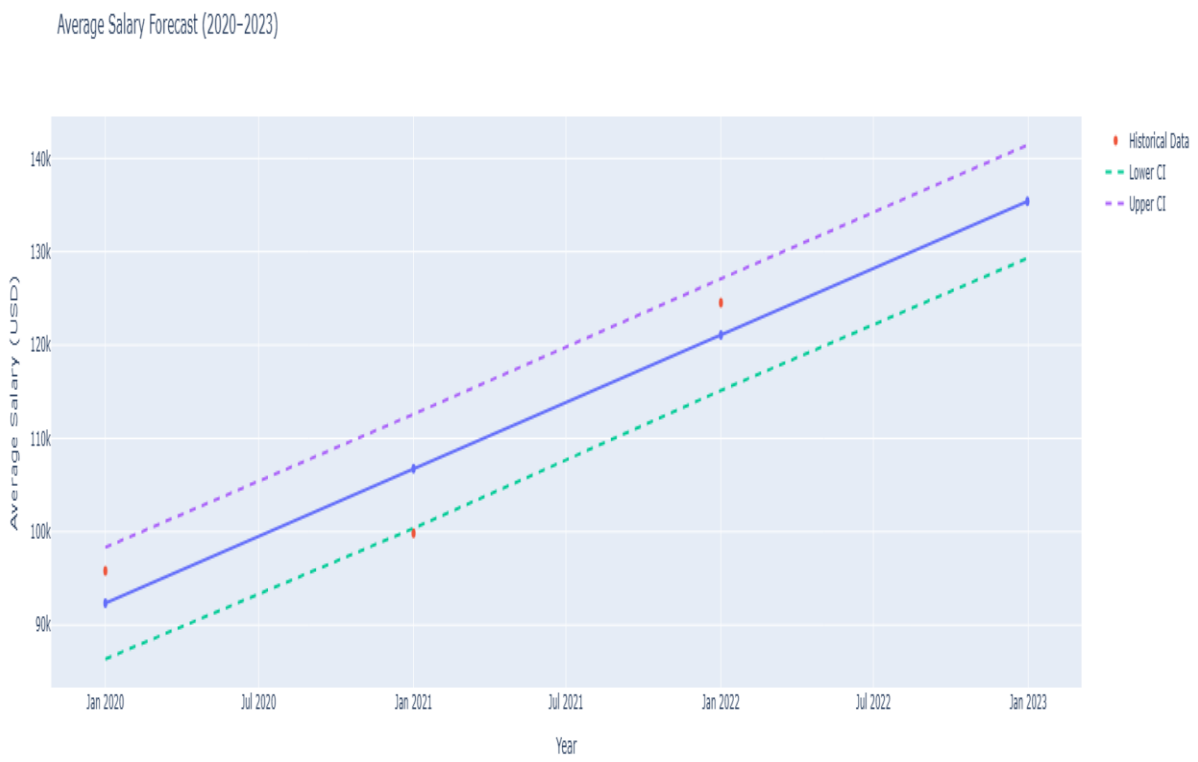
Recommendations :

For Employers: Prioritize visa-friendly policies or remote hubs to access global talent(eg. Portugal for EU time zones).

For Professionals: Target companies in high-flow corridors (eg. US firms hiring remotely from Canada).

Average salary forecast for 2023

Using Prophet



The line graph shows a clear upward convex curve, indicating accelerating growth through 2022 that moderates in 2023.

Confidence bands widen in 2023 forecast, reflecting increased economic uncertainty.

Actionable recommendations:

1. For employers:

Budget for 6-9% salary increases to remain competitive in 2023. Consider location-based adjustments (per prior geographic analysis). Monitor Q3-Q4 2023 trends for potential stabilization.

2. For Professionals

Negotiation Benchmark: Use the $135K average for senior roles.

Skill Premiums: Specialized roles (ML, cloud architecture) may command +20-30% over average.

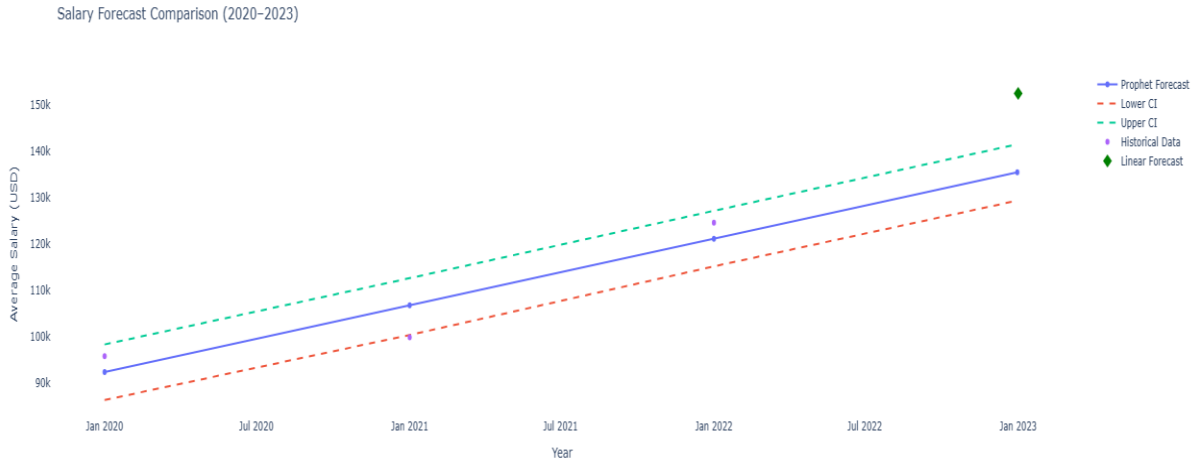
Timing: Early 2023 hires may have captured peak salaries before moderation.

Methodological Notes: Used annual aggregates to smooth monthly volatility

Limitations: Doesn't account for sudden economic shocks (eg. Banking crisis March 2023).Industry mix changes could skew averages.

Linear regression v.s prophet forecast comparison

Linear forecast for 2023: $152,377



Prophet's strengths

Better handles non-linear trends (eg.2022 hypergrowth)

Accounts for potential saturation effects

Narrower CI bands suggest higher confidence

Linear Model Risks

Overly optimistic given 2023 market conditions

Does not account for :

- Tech layoffs (Q1 2023)
- Reduced VC funding
- Geographic salary normalization

Actionable insights

Scenario Planning:

- Base Case- Prophet's $135k
- Upside Case - $145k if demand rebounds unexpectedly
- Downside risk -$125k if recession deepens.

Hiring strategy

- Use Prophet's upper bound as salary cap for critical roles.
- For budgeting, average the models $144k.

Professional benchmarking

- Senior roles : $135-$150k range reflects current uncertainty
- Negotiate based on specialisation (AI/ML roles may still command linear model premiums)

Conclusion

This project analysed a dataset of data science job salaries to uncover key trends and insights. The dataset, containing 607 records from 2020 to 2022, was cleaned, pre-processed and explored using various statistical and visualization techniques. Key findings include:

1. Salary trends :

- Salaries showed a steady increase , rising from $95813 in 2020 to $124522 in 2022, reflecting growing demand for data science roles.
- The highest-paying roles were director-level and specialized positions (e.g. Principal data Scientist, Machine Learning Engineer).

2. Experience level :

- Median salaries grew exponentially with experience : Entry-level ($56.5k), Mid-level ($101k), Senior ($138k), Executive ($171k)

3. Company Size:

- Larger companies paid higher average salaries (Large: $100K–120K, Medium: $80K–100K, Small: $60K–80K).

4. Remote Work:

- Fully remote roles commanded the highest average salaries ($122K), followed by on-site ($106K) and hybrid ($81K).

5. Geographic Analysis:

- High-income countries (e.g., US, UK) offered the highest salaries, while emerging markets (e.g., India, Brazil) had lower averages.
- Cross-border employment pairs (e.g., US-Canada, Malaysia-US) revealed unique compensation trends.

6. Job Title Clustering:

- K-means clustering grouped job titles into 6 clusters, with technical roles (e.g., ML, Computer Vision) and leadership roles commanding higher salaries.

7. Salary Forecast for 2023:

- Using Prophet, the average salary was projected to reach $135K, with a potential range of $125K–$145K depending on economic conditions.

Actionable Insights:

For Employers: Budget for competitive salaries, especially for remote and senior roles.

For Professionals: Target high-paying clusters (e.g. ML, leadership) and negotiate based on specialization and location.

Trends: Remote work and experience level significantly impact salaries, with technical and leadership roles offering the highest compensation.

This analysis provides actionable insights for both employers and job seekers in the data science field.