# DATA SCIENCE INDUSTRY COMPENSATION ANALYSIS

*Project Report by Group 3*
Anushka Mondal
Ananyaa Shahi
Bhakti Kate
Vivid Liu

## INTRODUCTION

For our Data Science Industry Compensation Analysis, we delve into the “Data Science Salaries 2023” on Kaggle, exploring salary details across the data science and related professions from 2020 to 2023. The comprehensive dataset consists of 3755 rows and 11 columns with no missing values in each. It provides a wide range of variables including the work year, job titles, level of experience, type of employment, employee’s country of residence, the amount of work done remotely, salary amount paid, salary currency, company location, and size. We chose this dataset because it perfectly aligns with our career aspirations in data science and analytics, offering tailored and valuable insights relevant to our field. Sourced from aijobs.net, a reputable platform focusing on AI, Machine Learning, and Data Science jobs, this dataset ensures quality and reliability.

## DRIVING QUESTION

The driving questions for our analysis are: **What are the key elements that influence the variation in salaries for professionals in the data science and related fields in 2023? Also, can we predict or analyze salary trends based on these attributes for 2024?** The goal of our project is to extract insights and actionable recommendations, empowering us to make better career decisions. By analyzing the dataset, we aim to uncover the variables, patterns, and trends that drive the salary components in this fast-evolving Data field.

## DATA CLEANING

The following steps were performed for cleaning the dataset:

1. **Data Type Transformation:** Converted the data type of the column ‘work_year’ from int64 to string (object) to prevent arithmetic operations on this column.
2. **Duplicate Rows Removal:** Dropped 1171 duplicate rows, resulting in 2584 unique rows.
3. **Whitespace Removal:** Eliminated leading and trailing white spaces from the data to ensure consistency for analysis, sorting, and filtering purposes.
4. **Job Titles Standardization:** Identified similar job titles ('Machine Learning Engineer' and 'ML Engineer') and replaced the abbreviation ('ML Engineer') with its full form ('Machine Learning Engineer') using a map function, resulting in 92 unique job titles.

5. **Check for Newly Created Duplicates:** After modifying job titles, checked for new duplicate rows and found 8. Removed these duplicates, resulting in 2576 unique rows.
6. **Abbreviated Values Replacement:** Replaced abbreviated values in the 'experience_level' column ('EN', 'EX', 'MI', 'SE') with their expanded forms ('Entry - Level', 'Executive', 'Mid – Level', 'Senior').
7. **Abbreviated Employment Type Replacement:** Replaced abbreviated values in the 'employment_type' column ('FT', 'CT', 'FL', 'PT') with their expanded forms ('Full – Time', 'Contractor', 'Freelancer', 'Part – Time').

This is the screenshot of the final cleaned dataset:

```
ds_salaries_df.head(5)
```

| | work_year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio | company_location | comp |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2023 | Senior | Full-Time | Principal Data Scientist | 80000 | EUR | 85847 | ES | Fully Remote | ES | |
| **1** | 2023 | Mid-Level | Contractor | Machine Learning Engineer | 30000 | USD | 30000 | US | Fully Remote | US | |
| **2** | 2023 | Mid-Level | Contractor | Machine Learning Engineer | 25500 | USD | 25500 | US | Fully Remote | US | |
| **3** | 2023 | Senior | Full-Time | Data Scientist | 175000 | USD | 175000 | CA | Fully Remote | CA | |
| **4** | 2023 | Senior | Full-Time | Data Scientist | 120000 | USD | 120000 | CA | Fully Remote | CA | |

## EXPLORATORY DATA ANALYSIS

### Trendiest Job Title: Which is the trendiest job title by country?

The first step in finding out the trendiest job title was to create series that categorizes each country's job titles along with their corresponding salary lists. This was done by using groupby function.

```
# Group the dataframe by company location and job title, and aggregate the salaries into a list
salary_data_by_country = ds_salaries_df.groupby(['company_location', 'job_title'])['salary_in_usd'].apply(list)
print (salary_data_by_country)
```

```
company_location  job_title
AE                Lead Data Scientist                                                      [115000]
                  Machine Learning Engineer                                        [120000, 65000]
AL                3D Computer Vision Researcher                                             [10000]
AM                Machine Learning Engineer                                                 [50000]
AR                Data Analyst                                                              [50000]
                                                                          ...
US                Product Data Analyst                                            [100000, 140000]
                  Research Engineer                   [189110, 139000, 203000, 133000, 230000, 20000...
                  Research Scientist                  [220000, 130000, 110000, 210000, 136000, 21000...
                  Staff Data Scientist                                                     [105000]
VN                Data Engineer                                                             [12000]
Name: salary_in_usd, Length: 352, dtype: object
```

Following that we converted the series into a data frame called df_salary_data. To find the most common job title in each country we followed a step-by-step approach. First, apply function was used to go through all rows in order to count how many times each job title was reported. We then added a new column showing the maximum salary for each job for respective country. For a cleaner output we got rid of individual salary columns by using drop function. To locate the most common job title and its corresponding salary for each country, we used the idxmax function, which gives us the index value. Finally, we reset the index to ensure that the column names were correctly assigned to their respective headers.

The Output displays the trendiest job as per country and also, it's corresponding salaries. This analysis helps us make informed decision about the job market in our "Dream work location"

| | Country | Most Common Job Title | Most Common Job | Salary (USD) |
|---|---|---|---|---|
| **0** | AE | Machine Learning Engineer | 2 | 120000.0 |
| **1** | AL | 3D Computer Vision Researcher | 1 | 10000.0 |
| **2** | AM | Machine Learning Engineer | 1 | 50000.0 |
| **3** | AR | Data Analyst | 1 | 50000.0 |
| **4** | AS | 3D Computer Vision Researcher | 1 | 20000.0 |
| **...** | ... | ... | ... | ... |
| **67** | TH | Data Science Consultant | 1 | 29453.0 |
| **68** | TR | Data Scientist | 3 | 25000.0 |
| **69** | UA | AI Developer | 2 | 108000.0 |
| **70** | US | Data Engineer | 487 | 324000.0 |
| **71** | VN | Data Engineer | 1 | 12000.0 |

72 rows × 4 columns

## <u>Lucrative Job Title: Which is the most lucrative job title by country?</u>

Here, we group the data by company location in which we look at salary for job titles and find the index for highest paid job title. Create separate data frame using that index for each country. Then we rename the columns for better representation. This analysis empowers us with the knowledge necessary to target our "Dream Salary".

| | Country | Highest Paying Job Title | Salary (USD) |
|---|---|---|---|
| **0** | AE | Machine Learning Engineer | 120000.0 |
| **1** | AL | 3D Computer Vision Researcher | 10000.0 |
| **2** | AM | Machine Learning Engineer | 50000.0 |
| **3** | AR | Data Analyst | 50000.0 |
| **4** | AS | Business Data Analyst | 50000.0 |
| **...** | ... | ... | ... |
| **67** | TH | Data Science Consultant | 29453.0 |
| **68** | TR | Data Engineer | 28016.0 |
| **69** | UA | AI Developer | 108000.0 |
| **70** | US | Research Scientist | 450000.0 |
| **71** | VN | Data Engineer | 12000.0 |

72 rows × 3 columns

## By Country, Most Common Job V/S Highest paid Job

Here we merge the above two data frames and generate the output to visualize most common job and highest paying job for respective country. This analysis guides us to make optimal job choice with respect to salary.Best used during job switching.

| | Country | Most Common Job Title | Most Common Job Salary (USD) | Highest Paying Job Title | Highest Paid Job Salary (USD) |
|---|---|---|---|---|---|
| **0** | AE | Machine Learning Engineer | 120000.0 | Machine Learning Engineer | 120000.0 |
| **1** | AL | 3D Computer Vision Researcher | 10000.0 | 3D Computer Vision Researcher | 10000.0 |
| **2** | AM | Machine Learning Engineer | 50000.0 | Machine Learning Engineer | 50000.0 |
| **3** | AR | Data Analyst | 50000.0 | Data Analyst | 50000.0 |
| **4** | AS | 3D Computer Vision Researcher | 20000.0 | Business Data Analyst | 50000.0 |
| **...** | ... | ... | ... | ... | ... |
| **67** | TH | Data Science Consultant | 29453.0 | Data Science Consultant | 29453.0 |
| **68** | TR | Data Scientist | 25000.0 | Data Engineer | 28016.0 |
| **69** | UA | AI Developer | 108000.0 | AI Developer | 108000.0 |
| **70** | US | Data Engineer | 324000.0 | Research Scientist | 450000.0 |
| **71** | VN | Data Engineer | 12000.0 | Data Engineer | 12000.0 |

72 rows × 5 columns

## Demographic Pay Gaps: Are there any pay gaps related to most common job title in the world?

In order to identify the presence of demographic pay-gap, we choose to use the most common job title in the world to be our reference. To find that we used loc and idmax function. This helped locate the index of the job title. As per the data, Data Engineer was the most common job title.

```
Country                                       US
Worlds Most Common Job Title       Data Engineer
Most Common Job                              487
Salary (USD)                            324000.0
Name: 70, dtype: object
```

Further, we sort all countries with Data engineer roles, in a descending order of salary. Head and Tail provides us with the top 3 and bottom 3 countries. We use this information to plot graph.


**Observation:**
• We see that there is definite pay-gap for the most common job title in the world.
• Data Engineer: US highest paying country ($325k), Vietnam lowest paying country ($12k)

For most common job title in each country we find the corresponding highest paying country. Here we use nested for loop to go thorugh df_salary_data dataframe to find the corresponding highest paying country. Further we create a data frame to store our values and then merge it with df_most_common_job via index matching.

| | Country | Most Common Job Title | Salary (USD) | Highest Paying Country | Highest Paying Country Salary |
|---|---|---|---|---|---|
| 0 | AE | Machine Learning Engineer | 120000.0 | US | 342300.0 |
| 1 | AL | 3D Computer Vision Researcher | 10000.0 | CR | 50000.0 |
| 2 | AM | Machine Learning Engineer | 50000.0 | US | 342300.0 |
| 3 | AR | Data Analyst | 50000.0 | GB | 430967.0 |
| 4 | AS | 3D Computer Vision Researcher | 20000.0 | CR | 50000.0 |
| ... | ... | ... | ... | ... | ... |
| 67 | TH | Data Science Consultant | 29453.0 | US | 145000.0 |
| 68 | TR | Data Scientist | 25000.0 | US | 412000.0 |
| 69 | UA | AI Developer | 108000.0 | IN | 300000.0 |
| 70 | US | Data Engineer | 324000.0 | US | 324000.0 |
| 71 | VN | Data Engineer | 12000.0 | US | 324000.0 |

72 rows × 5 columns

## Impact of remote ratio column: Does work modality have any significant impact on salaries?

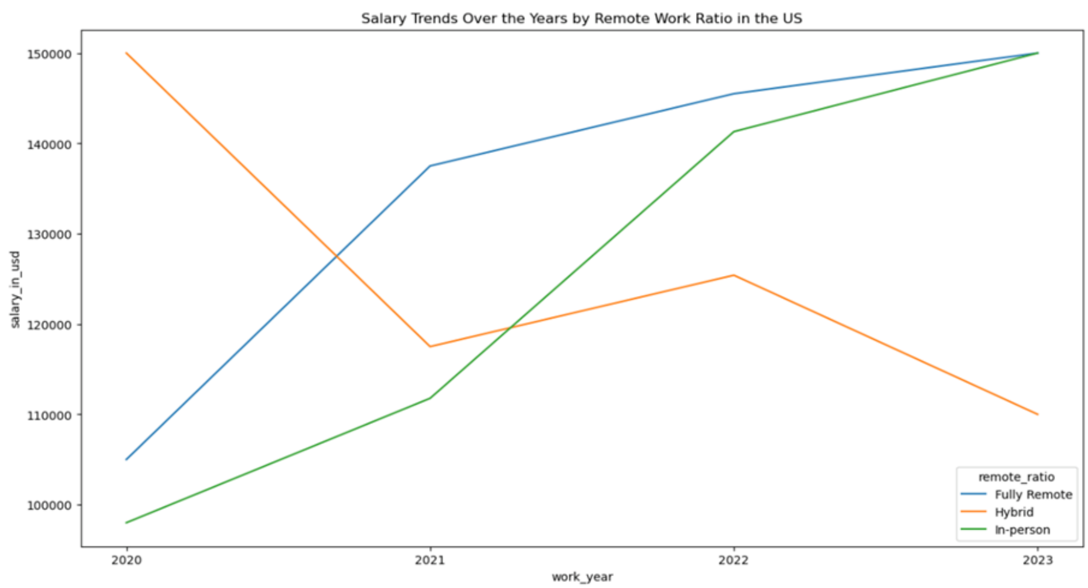
We opted to reclassify the 'remote_ratio' column values for better categorization. The numerical values were substituted with descriptive terms to enhance clarity:

- 100: Replaced by 'Fully Remote'
- 50: Updated to 'Hybrid'
- 0: Transformed into 'In-person'

| | work_year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio | company_location | c |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2023 | Senior | Full-Time | Principal Data Scientist | 80000 | EUR | 85847 | ES | Fully Remote | ES | |
| **1** | 2023 | Mid-Level | Contractor | Machine Learning Engineer | 30000 | USD | 30000 | US | Fully Remote | US | |
| **2** | 2023 | Mid-Level | Contractor | Machine Learning Engineer | 25500 | USD | 25500 | US | Fully Remote | US | |
| **3** | 2023 | Senior | Full-Time | Data Scientist | 175000 | USD | 175000 | CA | Fully Remote | CA | |
| **4** | 2023 | Senior | Full-Time | Data Scientist | 120000 | USD | 120000 | CA | Fully Remote | CA | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **3750** | 2020 | Senior | Full-Time | Data Scientist | 412000 | USD | 412000 | US | Fully Remote | US | |

Firstly, the dataset is filtered to include only US locations. Then, the data is grouped by 'remote_ratio' and 'work_year', calculating the median salary for each group. Finally, a line plot is generated using Seaborn and Matplotlib, showcasing salary trends across different remote work ratios over the years, allowing for visual comparison and analysis of salary trends based on remote work practices.

In analyzing Salary Trends Over the Years based on Remote Work Ratio in the US, we chose to employ the median salary over the mean. This decision was influenced by the median's resilience to outliers, making it a more robust measure for skewed distributions and scenarios where extreme values can significantly impact the average. The median's interpretability and simplicity enhance its effectiveness in representing a central tendency, especially in salary comparisons, ensuring a more representative measure of typical salary values across different remote work ratios.



**Observation:**

•During the peak pandemic years of 2020 and 2021, remote workers' median salaries surged. Surprisingly, in 2022 and 2023, despite the return to in-person work, remote job compensation continued to rise.

•By 2023, median salaries for in-person and remote work equalized, showing no difference. However, hybrid work experienced a sharp decline in median salary, becoming the least lucrative option.

**Highest paid remote_ratio entry level positions: Which are the highest paid entry level jobs in US in 2023 as per work modality?**

**NOTE:** We have chosen the US company location because we believe that the majority of our classmates would want to work in the US after graduation.

The analysis first involved grouping the dataset by 'remote_ratio' and 'job_title', computing median salaries, and pinpointing the highest and lowest-paid job titles for each remote work ratio in 2023. Following this, a specific focus on Entry-Level positions entailed data filtering, median salary calculation based on 'remote_ratio' and 'job_title', identification, and display of the highest-paid Entry-Level job titles in 2023. Additional job information was merged and duplicate rows were dropped to streamline the results for clarity and precision in understanding the highest-paid roles within different remote work scenarios for Entry-Level positions.

| | remote_ratio | job_title | salary_in_usd | company_size | experience_level | employment_type |
|---|---|---|---|---|---|---|
| 0 | Fully Remote | Machine Learning Scientist | 225000.0 | L | Entry-Level | Full-Time |
| 1 | Hybrid | Research Scientist | 220000.0 | L | Entry-Level | Full-Time |
| 2 | Hybrid | Research Scientist | 220000.0 | M | Entry-Level | Full-Time |
| 5 | In-person | Computer Vision Engineer | 172500.0 | M | Entry-Level | Full-Time |

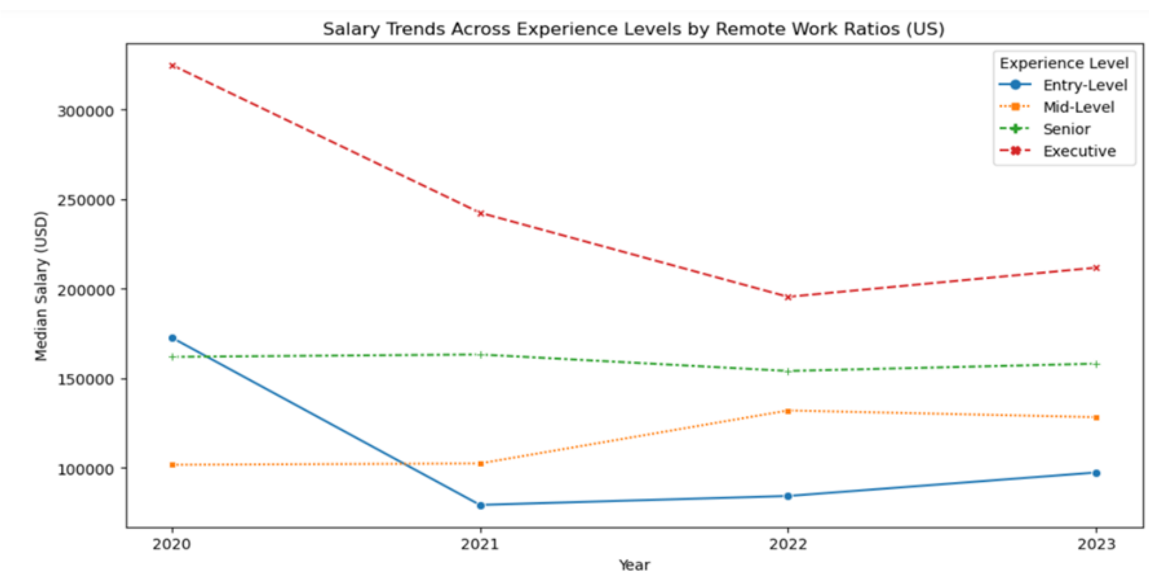
**Lowest paid remote_ratio entry level positions: Which are the lowest paid entry level jobs in US in 2023 as per work modality?**

The process is initiated by selecting essential columns from the 'lowest_paid_jobs_entry_level' DataFrame, including 'remote_ratio', 'job_title', and 'salary_in_usd'. Subsequently, this data was merged with 'us_data_entry_level' to incorporate supplementary job information like 'company_size', 'experience_level', and 'employment_type', ensuring a comprehensive view. To ensure clarity and uniqueness, duplicate rows were dropped, resulting in a refined dataset. Finally, the resulting lowest-paid Entry-Level job titles across various remote work ratios were displayed, offering a concise overview of these roles within distinct remote work scenarios.

| | remote_ratio | job_title | salary_in_usd | company_size | experience_level | employment_type |
|---|---|---|---|---|---|---|
| 0 | Fully Remote | AI Scientist | 12000.0 | M | Entry-Level | Full-Time |
| 1 | Fully Remote | AI Scientist | 12000.0 | M | Entry-Level | Part-Time |
| 2 | Fully Remote | AI Scientist | 12000.0 | S | Entry-Level | Part-Time |
| 3 | Fully Remote | BI Analyst | 12000.0 | L | Entry-Level | Part-Time |
| 4 | Fully Remote | BI Analyst | 12000.0 | L | Entry-Level | Full-Time |
| 5 | Hybrid | Business Data Analyst | 48000.0 | L | Entry-Level | Full-Time |
| 6 | Hybrid | Business Data Analyst | 48000.0 | L | Entry-Level | Contractor |
| 7 | In-person | Data Analyst | 62500.0 | M | Entry-Level | Full-Time |
| 16 | In-person | Data Analyst | 62500.0 | L | Entry-Level | Full-Time |
| 19 | In-person | Data Analyst | 62500.0 | S | Entry-Level | Part-Time |
| 25 | In-person | Data Analyst | 62500.0 | S | Entry-Level | Full-Time |
| 26 | In-person | Data Analyst | 62500.0 | L | Entry-Level | Part-Time |

**Impact of experience and remote_ratio: Did experience levels play a role in salary trends as per different work modes?**

The process involved analyzing the impact of experience levels and remote work ratios on salary trends in the US. Initially, the dataset was filtered for US locations, and a specific order for experience levels ('Entry-Level', 'Mid-Level', 'Senior', 'Executive') was defined. The data was then grouped by 'work_year', 'experience_level', and 'remote_ratio', calculating median salaries for each group. Subsequently, a line plot was generated using Seaborn and Matplotlib, displaying salary trends across different remote work ratios and experience levels.



**Observation:**

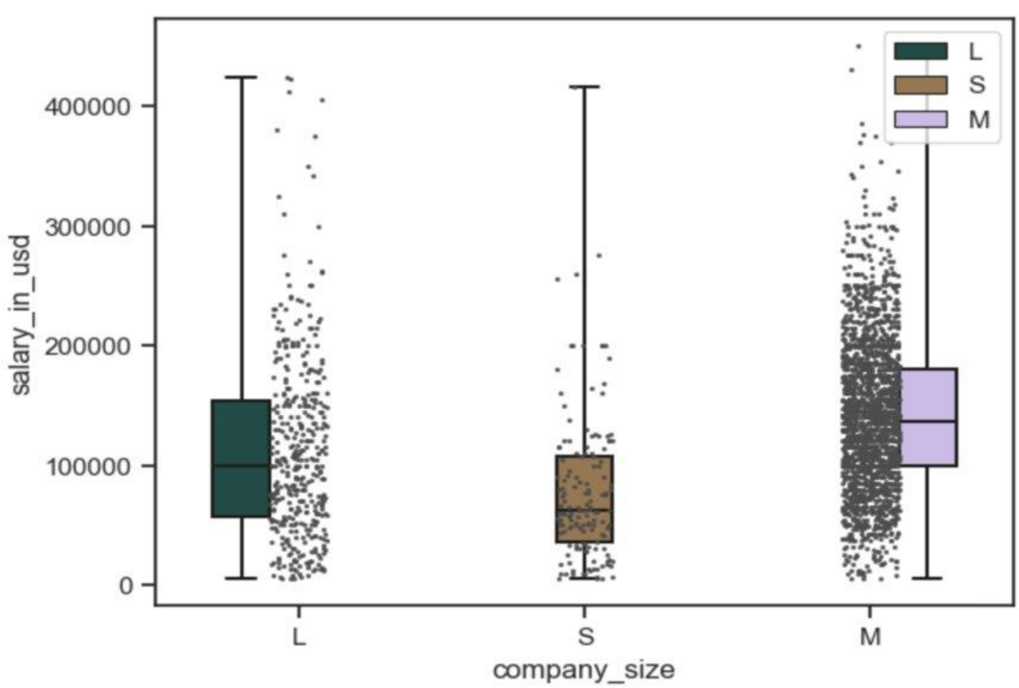
Entry and executive-level median salaries declined significantly during the peak pandemic years from 2020 to 2021 and showed minimal growth from 2022 to 2023. Meanwhile, senior-level salaries consistently increased and mid-level salaries grew from 2021 to 2022 as the pandemic eased.

**<u>Salary variation based on company size: Do salaries vary by company size (small, medium, and large)?</u>**

Our objective of doing this analysis was to explore the distribution of salaries in the field of Data Science across different company sizes. Our aim was to identify any patterns or trends that may emerge, particularly in relation to the size of the companies.

We utilized the Seaborn library in Python for data visualization, specifically focusing on box plots to visualize the distribution of salaries across various company sizes. We plotted ‘company_size’ on x-axis, ‘salary_in_usd’ on y-axis, and ‘company_size’ as hue. Additionally, we incorporated strip plots to display individual data points, providing a more comprehensive view of the salary distribution.

```
<Axes: xlabel='company_size', ylabel='salary_in_usd'>
```



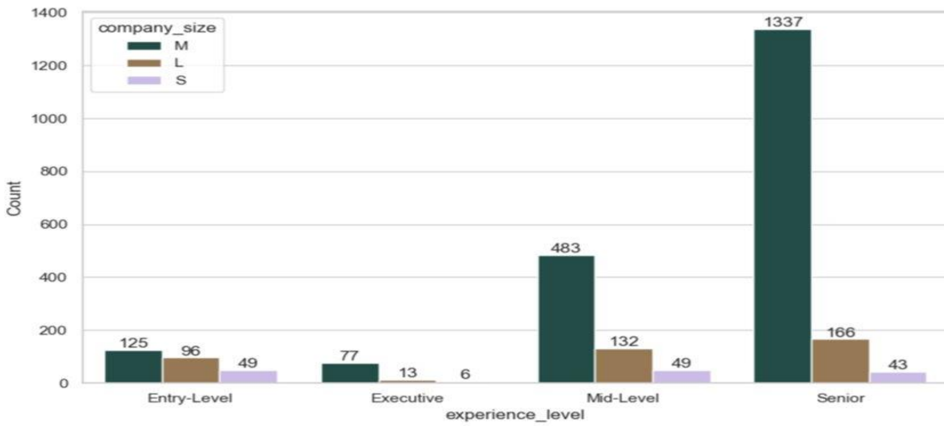
**Observations:**
1) Company Size Impact on Salaries: By comparing the box plots for Small, Medium, and Large sized companies, we were able to identify a noticeable impact of company size on Data Science salaries. Medium sized companies demonstrated a leading position in the employment rate within the field of data science, accompanied by higher median salary ranges when contrasted with both large and small firms.
2) Individual Data Points: The strip plot complements the box plot by displaying individual data points. This allows for a more granular examination of salary distribution within each company size category.

**<u>Salary variation based on company size: How is salary distributed across companies based on their sizes?</u>**

Our objective of this analysis is to examine the distribution of experience levels within different company sizes in the field of Data Science. We aim to understand the count of individuals at various experience levels across diverse company size categories.

We decided that grouping the data based on two categorical variables: 'experience_level' and 'company_size' would come in handy for this section. The value_counts() function is then applied to determine the count of occurrences for each combination of these variables. The resulting data is structured into a Data Frame named ‘exp_size_df’.

Furthermore, we have utilized the Seaborn library in Python to create a grouped bar chart that visualizes this. Our plot consists of x-axis as 'experience_level', y-axis as 'Count', and hue as 'company_size' The bar chart allows for a clear comparison of the count of individuals in each category.



**Observation:**
The grouped bar chart represents the count of individuals at different experience levels within each company size category.

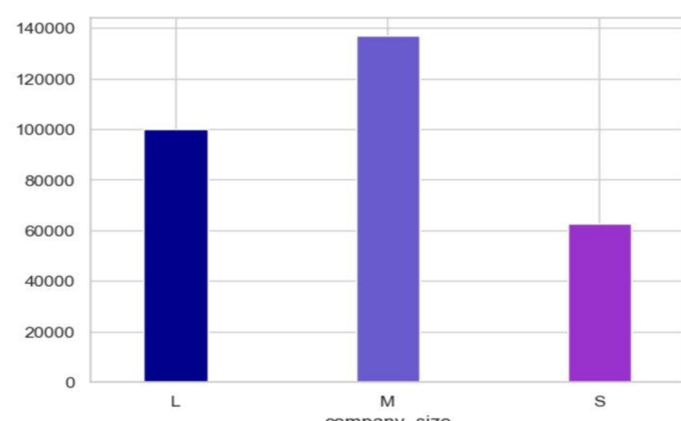
1. In mid-sized companies, the grouped bar chart suggests a structure where the number of seniors is highest, followed by mid-level professionals, entry-level employees, and executives.
2. Similarly, large-sized companies exhibit a comparable pattern with seniors with a higher number, followed by mid-level and entry-level roles, and finally executives.
3. Conversely, small-sized companies present a distinctive structure, emphasizing a balance between entry-level and mid-level positions, followed by seniors and executives.

**<u>Salary variation based on company size: How is salary distributed across companies based on their sizes?</u>**

The goal of this analysis is to explore the relationship between median salaries and company sizes in the field of Data Science.

We decided to do the analysis by grouping the data based on the variable 'company_size.' We are calculating the median salary for each company size category and putting it in a series named ‘salary_vs_company_size’.

We utilized the Matplotlib library in Python to create a bar chart. The x-axis represents different company sizes, and the y-axis represents median salaries. Three distinct bar colors are employed to distinguish between company size categories: 'small,' 'medium,' and 'large.'



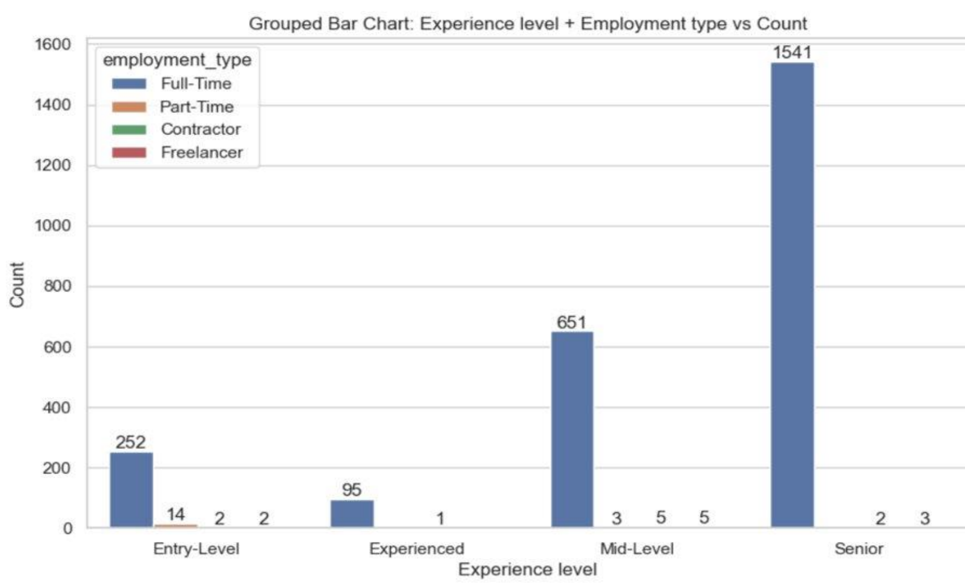
**Observation:**
We are clearly able to see median salary distribution across company sizes. The bar chart provides a clear visual representation of the median salaries within each company size category. We can see that medium-sized companies surpass both large and small companies in providing higher median salaries.

**Employment types and experience: Do employment types have a dependency on experience?**

The objective of this analysis is to explore the distribution of employment types across different experience levels in the Data Science field. The count of individuals for each combination of experience level and employment type is examined to provide insights into workforce composition.

We used the ‘groupby’ function to group the data based on two categorical variables: 'experience_level' and 'employment_type.' The value_counts() function is then applied to calculate the count of occurrences for each combination of these variables. The resulting data is structured into a DataFrame named employment_vs_experience_dt.

Additionally, we utilized seaborn and matplotlib libraries in Python to create a grouped bar chart. The x-axis represents different experience levels, the y-axis represents the count of individuals, and different colors represent distinct employment types.



**Observation:**
From the grouped bar chart, we can see the count of individuals at different experience levels, categorized by employment types. Colors distinguish between employment types, providing a clear comparison. Most of the companies are offering full time roles for data science jobs. The number of full-time roles is highest among senior-level positions, followed sequentially by mid-level, entry-level, and executive roles. Job roles for part-time, contractors and freelancers are low.

**Correlation between Employment and Experience types**

The objective of this analysis is to assess the association between experience levels and employment types in the Data Science field using a Chi-Square test.

The analysis employs the chi2_contingency function from the scipy.stats module in Python. A contingency table is created using the pd.crosstab function, with experience levels as rows and employment types as columns. The Chi-Square test is then performed on this contingency table to assess the independence of the two categorical variables.

```
from scipy.stats import chi2_contingency
```

```
cross_tab = pd.crosstab(index=ds_salaries_df['experience_level'],columns=ds_salaries_df['employment_type'])
chi_square_result = chi2_contingency(cross_tab,)
chi_square_result
```

```
Chi2ContingencyResult(statistic=107.89611391840003, pvalue=3.9399062330006367e-19, dof=9, expected_freq=array([[1.0
4813665e+00, 1.04813665e+00, 2.66121894e+02, 1.78183230e+00],
       [3.72670807e-01, 3.72670807e-01, 9.46211180e+01, 6.33540373e-01],
       [2.57763975e+00, 2.57763975e+00, 6.54462733e+02, 4.38198758e+00],
       [6.00155280e+00, 6.00155280e+00, 1.52379425e+03, 1.02026398e+01]]))
```

**Observation:**
H0 - The two columns are not correlated.
H1 - The two columns are correlated.
Result - The probability of H0 being true.
In this case, since p-value is very less (in comparison of alpha, say 0.05), we will reject the H0 and conclude that the two columns are correlated.

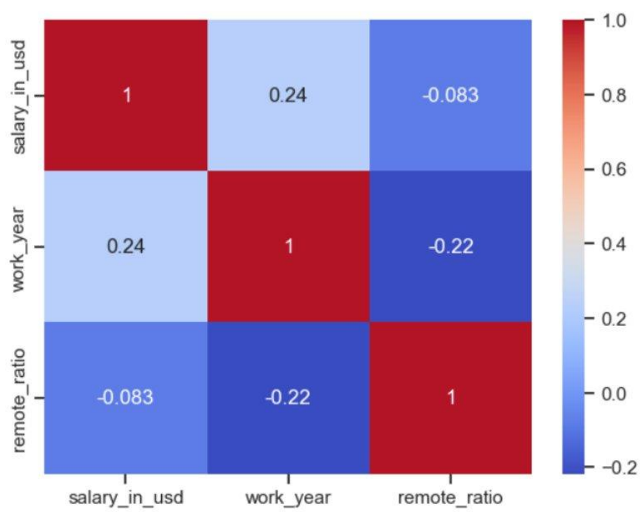
## Finding correlation between numerical values of the dataset

The objective of this analysis is to explore the correlation between salary, work year, and remote work ratio in the Data Science field. A correlation matrix is computed and visualized using a heatmap to identify potential relationships between these continuous variables.

The analysis focuses on a subset of the dataset, including the variables 'salary_in_usd,' 'work_year,' and 'remote_ratio.' The 'work_year' variable is converted to integers to facilitate correlation calculations. The correlation matrix is computed using the corr() function, and a heatmap is generated using the Seaborn library to visualize the strength and direction of correlations.

```
sns.heatmap(corr, annot=True, cmap='coolwarm')
ax[0].set_title('Continuous-Continuous Correlation')
```

```
Text(0.5, 1.0, 'Continuous-Continuous Correlation')
```

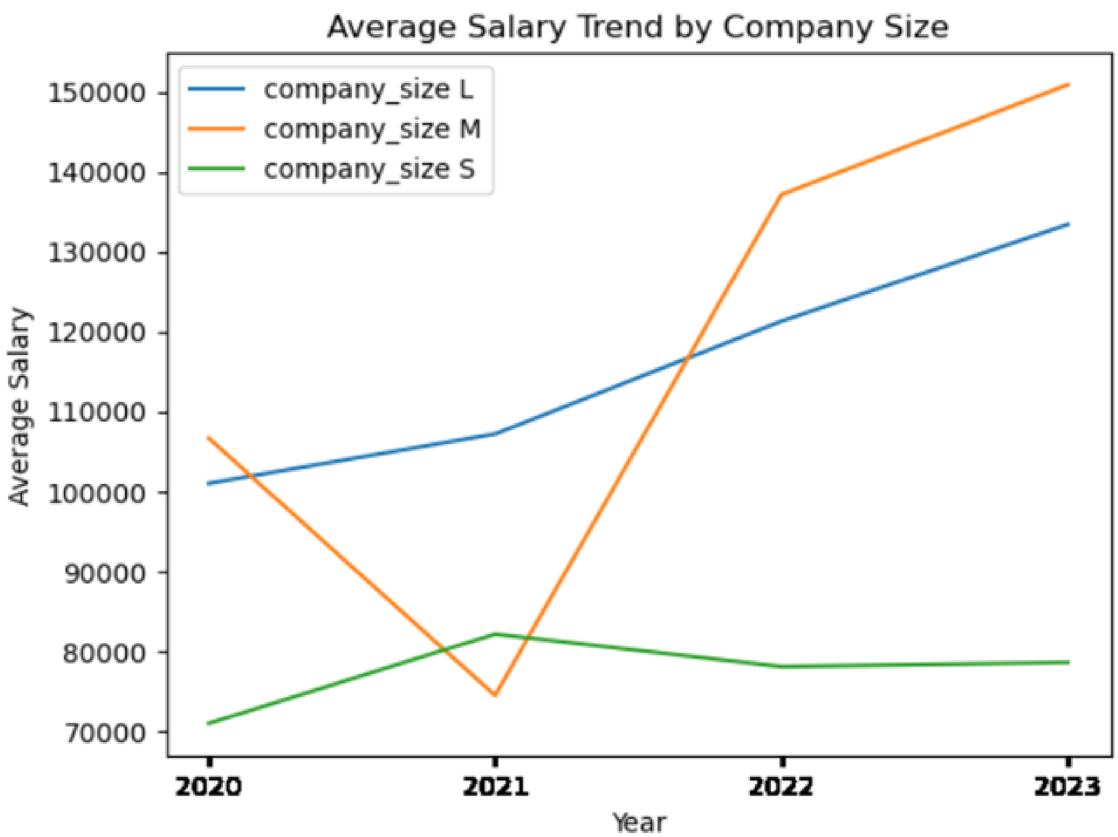



**Observation:**

The analysis reveals that both remote work ratio and work year exhibit weak correlations with salary in USD within the Data Science sector. The correlation coefficients of 0.24 for remote ratio and -0.083 for work year suggest limited predictive strength, indicating that these variables have relatively low influence on salary variations.

**Compensation Trends: Are organizations staying up-to-date with the compensation trends in the industry? What kind of talent are companies looking for based on the highest paying job title?**

We developed two key questions centering around compensation trends. Firstly, we wanted to understand whether companies have been staying up-to-date with the compensation trends in the industry. To tackle this, we grouped the data by company size (L, M, S) and year and calculated average salaries using the mean function. Visualizing these averages on a line plot showcased a gradual increase in overall average salaries over the years. Notably, companies with small and medium sizes saw a spike and a decline in 2021 respectively, distinguishing them within this pattern.




The second question looked into the type of talent sought by companies based on the highest paying job title. To accomplish this, we followed a two-step process. Initially, we grouped data by job title to determine each title's average salary, arranging them in descending order to identify the highest paying role. The Data Science Tech Lead appeared as the highest paying job title. Filtering out data by the job title "Data Science Tech Lead" revealed it as a senior position at a large company, working as a full-time employee based in the US. This output gave us insight on the kind of talent holding the highest compensation in the Data Science industry.

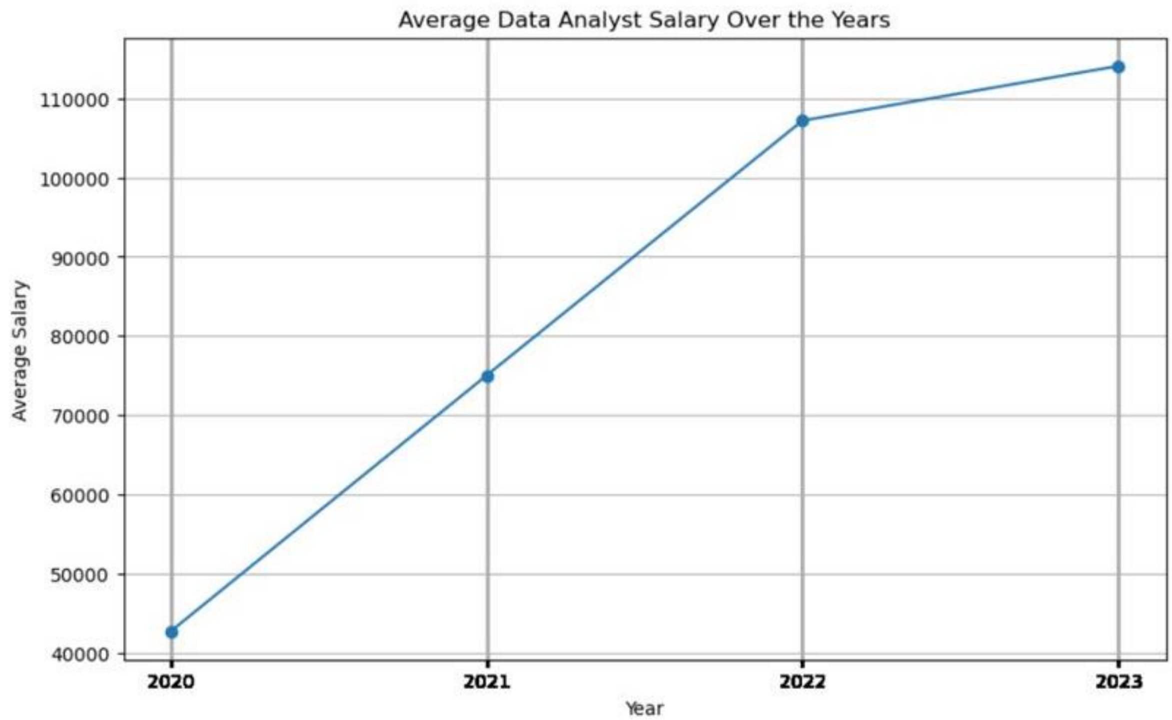
The Highest Paying Job Title

| _year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio | company_location | company_size |
|---|---|---|---|---|---|---|---|---|---|---|
| 2022 | SE | FT | Data Science Tech Lead | 375000 | USD | 375000 | US | 50 | US | L |

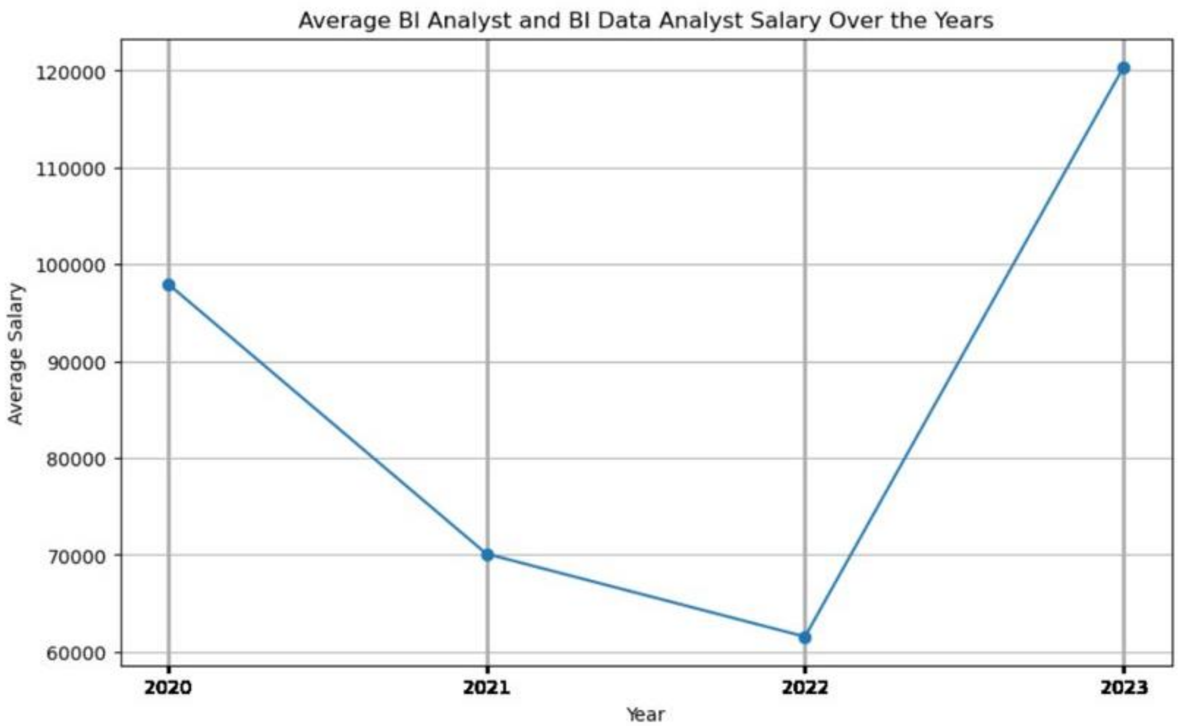
**<u>Salary trends over the years: Can we identify trends in salaries over time, such as year-over-year growth or fluctuations? What does the salary trend for Data Analysts and BI (Data) Analysts look like over the years?</u>**

To dig deeper into compensation trends in a different perspective, we decided to look at salary trends over time, focusing on identifying patterns of year-over-year growth or fluctuation. This involved grouping data by work year using the .groupby() function and calculating the average yearly salaries with the .mean() method. We visualized the shifts of the average yearly salaries via a line graph. Notably, we saw a substantial increase in salaries in 2022, indicating significant growth within the data science domain.

Next, we narrowed our focus to look at the salary trends in specific job titles that are more relevant to us and our peers: Data Analysts and BI Analysts/BI Data Analysts. For Data Analysts, we filtered data specific to this role, calculated average salaries by year, and illustrated the numbers on a line graph. The visualization showed a remarkable doubling of Data Analyst salaries from 2020 to 2022.




On the other hand, analyzing salary trends for BI Analysts/BI Data Analysts required more steps when filtering out the data due to the different job titles. We stored the BI Analyst and BI Data Analyst job titles in a list as one single variable. Then, the .isin() method was implemented to find rows of data that contain “BI Analysts” and “BI Data Analysts” in the “job_title” column. After these initial steps, the following steps were the same as finding out the average salary trend for Data Analysts. Therefore, we applied the same procedure, resulting in a line graph of the year-over-year average salary trend of BI Analysts and BI Data Analysis. Interestingly, we noted a decline from 2020 to 2022 followed by an upward trend from 2022 to 2023 for BI Analysts and BI Data Analysts, offering a unique perspective on these roles across time.




## DATA MODELING

**Chosen Model:** Multiple Linear Regression

**Purpose:** The purpose of utilizing the Multiple Linear Regression model is to predict average salaries associated with various job titles in the United States for the year 2024. This predictive analysis aims to uncover the relationships between categorical factors, such as job titles, and their influence on salary outcomes.

**Benefits:** By employing the Multiple Linear Regression model, we aim to gain a comprehensive understanding of how different categorical factors impact average salaries. This analysis will provide valuable insights into the key predictors influencing salary outcomes, enabling a deeper comprehension of the factors driving compensation variations across various job titles.

**Insights:** The primary goal is to offer clear insights into how categorical factors, including job titles, experience levels, employment types, and other variables, influence the average salary. Through this analysis, we seek to highlight the significance and relative impact of each predictor on salary determination.

**Aim:** This report aims to support classmates and individuals seeking employment opportunities in the United States during the year 2024. By considering a comprehensive array of factors and their influence on average salaries, the aim is to provide informed guidance to assist in making well-rounded career decisions. This analysis intends to empower individuals by presenting a holistic view of factors affecting salaries, thereby aiding them in making informed choices during their job hunt.

## DATA MODELING STEPS

### 1) <u>Prepare Data</u>

- **Filtering Data for Specific Criteria:** The dataset (**ds_salaries_df**) was filtered to include only records related to companies located in the United States during the years 2020 to 2023.
- **Grouping and Calculating Average Salary:** The filtered data was grouped by various columns such as 'job_title', 'work_year', 'experience_level', 'employment_type', 'employee_residence', 'remote_ratio', and 'company_location'. Then, the average salary (**salary_in_usd**) for each group was calculated.
- **Creating a Feature for 2024:** A new feature for the year 2024 was created by making a copy of the processed data for the prediction year ('work_year' column was updated to 2024).
- **Encoding Categorical Variables:** Categorical variables ('job_title', 'work_year', 'experience_level', 'employment_type', 'employee_residence', 'remote_ratio', 'company_location') were encoded using one-hot encoding via **pd.get_dummies()**, with **drop_first=True** to avoid multicollinearity issues. This process is essential for transforming categorical variables into a format suitable for machine learning models.

| | job_title | work_year | experience_level | employment_type | employee_residence | remote_ratio | company_location | salary_in_usd |
|---|---|---|---|---|---|---|---|---|
| **0** | AI Developer | 2023 | Mid-Level | Full-Time | US | Fully Remote | US | 200000.000000 |
| **1** | AI Scientist | 2021 | Entry-Level | Part-Time | BR | Fully Remote | US | 12000.000000 |
| **2** | AI Scientist | 2021 | Entry-Level | Part-Time | PK | Fully Remote | US | 12000.000000 |
| **3** | AI Scientist | 2022 | Entry-Level | Full-Time | US | Fully Remote | US | 50000.000000 |
| **4** | AI Scientist | 2022 | Experienced | Full-Time | US | Fully Remote | US | 200000.000000 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **320** | Research Scientist | 2023 | Mid-Level | Full-Time | US | Fully Remote | US | 193633.333333 |
| **321** | Research Scientist | 2023 | Mid-Level | Full-Time | US | In-person | US | 116250.000000 |

## 2) Fitting the Multiple Linear Regression model

- ❖ **Importing Necessary Libraries:** The required libraries from the scikit-learn (**sklearn**) package were imported. Specifically, the **linear_model** module was imported to access the **LinearRegression** class.
- ❖ **Initializing the Linear Regression Model:** An instance of the **LinearRegression** class was created and assigned to the variable **lr**. This model serves as the chosen algorithm for fitting a linear regression to the data.
- ❖ **Fitting the Model:** The **fit()** method from the linear regression model (**lr**) was utilized to train the model. The training data (**X=avg_salary_encoded**) and the target variable (**y=avg_predict_salary['salary_in_usd']**) were passed as arguments to the **fit()** function. The **X** parameter represents the independent variables (features), while **y** represents the dependent variable (target variable) that the model aims to predict.

## 3) Predict average salaries for 2024

- ❖ **Data Preparation for 2024 Prediction:**
- ➢ Similar to the previous encoding step during the training phase, categorical variables ('job_title', 'work_year', 'experience_level', 'employment_type', 'employee_residence', 'remote_ratio', 'company_location') from the avg_salary_2024 dataset were encoded using one-hot encoding via pd.get_dummies().
- ➢ drop_first=True was utilized to ensure consistency with the encoding structure applied during the training phase.
- ❖ **Making Predictions for 2024:**
- ➢ The pre-trained linear regression model (lr) was employed to predict salaries for the year 2024 based on the prepared encoded dataset (avg_salary_2024_encoded).
- ➢ The predict() method from the linear regression model was utilized to generate salary predictions for the year 2024 using the encoded features of job titles, work year, experience level, employment type, employee residence, remote ratio, and company location.
- ❖ **Displaying Predictions:**
- ➢ The predicted salaries for the year 2024 were printed or displayed using print(predictions_2024).
- ❖ **Ensuring Consistency:**
- ➢ It's essential to maintain consistency in data preprocessing steps, including encoding, between the training and prediction phases. Therefore, one-hot encoding was applied again during the prediction phase (avg_salary_2024_encoded) to

ensure that the model interprets and predicts accurately based on the encoding structure it learned during the training phase. This consistency helps in correctly transforming new data ('avg_salary_2024') into a format suitable for the model's predictions.

| | job_title | work_year | avg_salary |
|---|---|---|---|
| **0** | AI Developer | 2024 | 193372.605697 |
| **1** | AI Scientist | 2024 | 2035.750801 |
| **2** | AI Scientist | 2024 | -7882.182908 |
| **3** | AI Scientist | 2024 | 118344.392849 |
| **4** | AI Scientist | 2024 | 209293.998827 |
| **...** | ... | ... | ... |
| **320** | Research Scientist | 2024 | 170997.905079 |
| **321** | Research Scientist | 2024 | 174823.005377 |
| **322** | Research Scientist | 2024 | 193281.524666 |
| **323** | Research Scientist | 2024 | 197106.624964 |
| **324** | Staff Data Scientist | 2024 | 85117.817092 |

325 rows × 3 columns

***This data frame displays the average salaries for 2024 for data science job titles based in the US.***

## Analyzing Salary Trends: 2023 vs. 2024 for US-based Data Science Job Titles

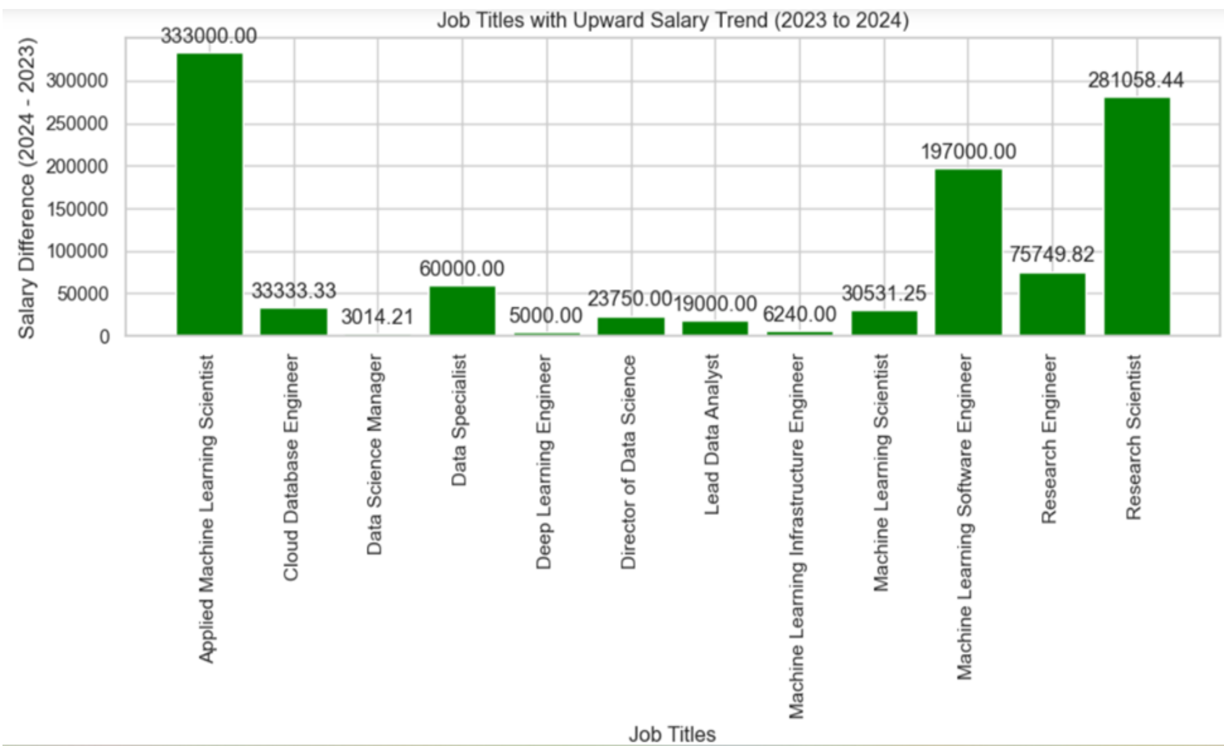
Following the development of a predictive model for forecasting average salaries in 2024 for data science job titles situated in the US, the subsequent step involved a comparative analysis with average salaries from 2023. The aim was to discern distinct trends among job titles, identifying those exhibiting upward, stable, or declining salary trajectories.

To figure out the trends, we first prepared the data in the following ways:

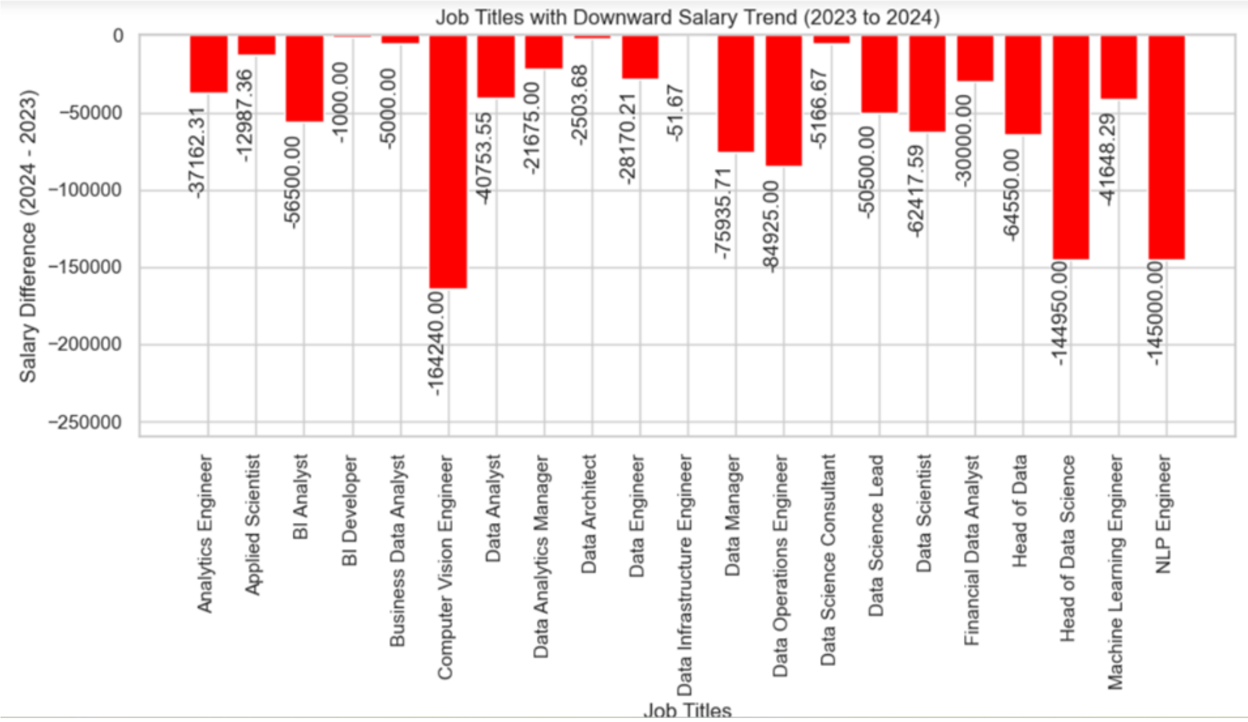
1. **Merging Datasets:** Combined average salaries for 2023 and cleaned average salaries for 2024 based on job titles using an inner merge operation via **pd.merge()** on the 'job_title' column.
2. **Comparison and Calculation:** Calculated the differences ('salary_diff') between 2023 and 2024 salaries by subtracting 2023 salaries ('avg_salary_2023') from 2024 salaries ('salary_in_usd') within the merged DataFrame ('comparison_df').

| | job_title | avg_salary_2023 | work_year | experience_level | employment_type | employee_residence | remote_ratio | company_location | salary_in_usd | salary_diff |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | AI Developer | 200000.000000 | 2024 | Mid-Level | Full-Time | US | Fully Remote | US | 200000.000000 | 0.000000 |
| 1 | Analytics Engineer | 167162.307692 | 2024 | Entry-Level | Full-Time | US | Hybrid | US | 130000.000000 | -37162.307692 |
| 2 | Applied Machine Learning Engineer | 130000.000000 | 2024 | Mid-Level | Full-Time | US | In-person | US | 130000.000000 | 0.000000 |
| 3 | Applied Machine Learning Scientist | 90000.000000 | 2024 | Mid-Level | Full-Time | US | Hybrid | US | 423000.000000 | 333000.000000 |
| 4 | Applied Scientist | 194951.000000 | 2024 | Senior | Full-Time | US | Fully Remote | US | 181963.636364 | -12987.363636 |
| 5 | BI Analyst | 132500.000000 | 2024 | Entry-Level | Full-Time | US | Hybrid | US | 76000.000000 | -56500.000000 |

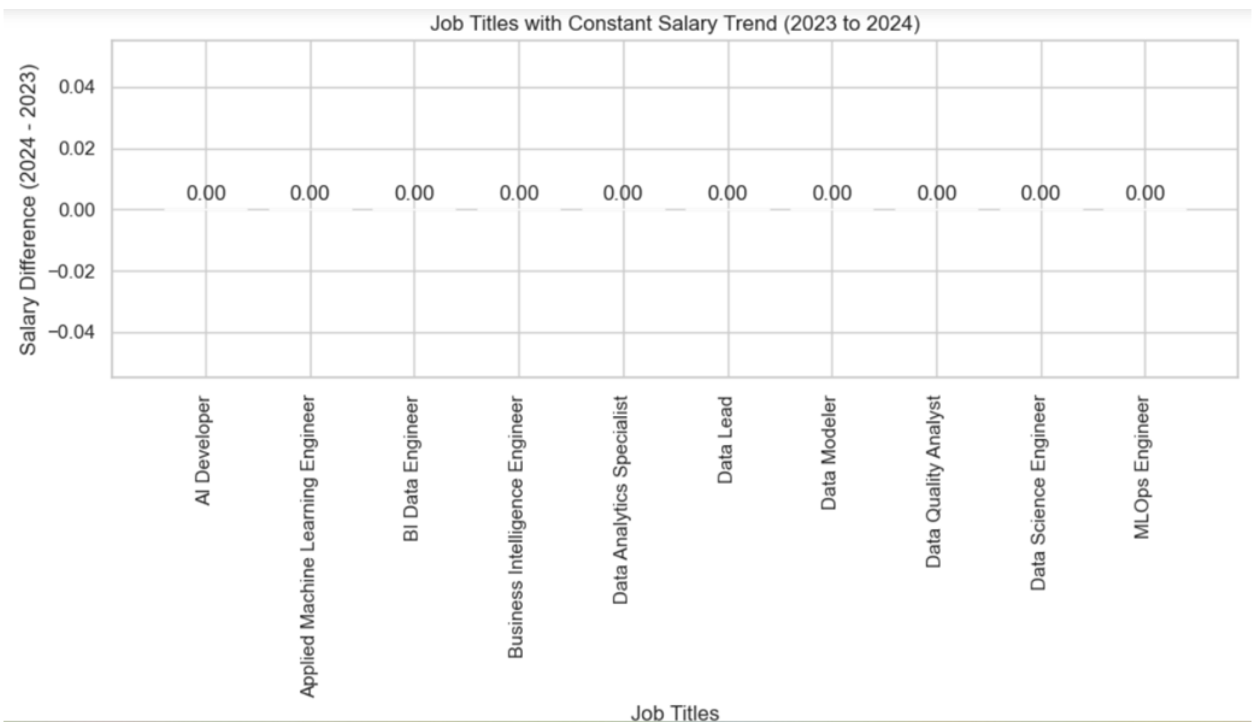
The following job titles all have an upward salary trend in 2024, of which **Applied Machine Learning Scientist** being the highest.



The following job titles all have a downward salary trend in 2024, of which **Computer Vision Engineer** being the lowest.




The following job titles have a constant salary trend:



## CONCLUSION

During this project, we were able to gain a comprehensive understanding of Python as a programming language, particularly in its application to various aspects of data science. The practical utilization of Python for tasks such as Exploratory Data Analysis (EDA), data modeling, cleaning, visualization, and descriptive analysis has been an important learning aspect for us. Moreover, despite coming from humble technical backgrounds, the course of this project has taught us to successfully navigate and write code. The experience has not only equipped us with practical coding ability but has also provided us with invaluable insights into the data analytics landscape. Additionally, our sense of adaptability and teamwork was enhanced via this project, as it has taught us the importance of effective collaboration and accommodation of diverse skill sets and needs within a team environment.

Throughout the project, we encountered various challenges that enriched our learning experience. One significant obstacle was the presence of incomplete data in the dataset, which hindered our chance to conduct a year-over-year analysis and visualization. The vast scope of the project made us look carefully into prioritization and establishing boundaries to effectively incorporate desired features into our code. The introduction of a new programming language, for most of us, posed a formidable challenge, given our non-technical background. Furthermore, materializing our thoughts into code required an active effort, often involving seeking additional resources online to overcome coding hurdles. The diverse schedules and varying levels of technical understanding among team members made us harmonize our pace, facilitating effective collaboration and ensuring everyone's contribution aligned with each other. Despite these challenges, the experience has been instrumental in honing our problem-solving skills, adaptability, and teamwork, essential attributes in the dynamic landscape of data science projects.