

Reports of the Interactive Cares Task

Created : By Mehedi Azad

InteractiveCares-Task

Data Scientist
Machine Learning Engineer
Research Scientist
Data Engineer
AI Scientist
Data Analyst
NLP Engineer
Computer Vision Engineer
Data Architect Machine Learning Scientist

Reports of Data Science Domain job Salary

Dataset link:

<https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries>

Project Title: EDA & Prediction of Data Domain job Salaries

Programming Language used: Python

Environment: Google Colab / Jupyter Notebook

Machine Learning Library used: Scikit Learn

**Data Manipulation, Computing Library used
Library used: Pandas & Numpy**

Data Visualization Library used:

Matplotlib

Seaborn

Plotly

Used Machine Learning Models:

1. Random Forest
2. GradientBoosting
3. XGB
4. ExtraTree
5. DecisionTree
6. LGBM
7. KNeighbors

Best Model: Random Forest

Best Accuracy: 0.948997

Data Manipulation Tools used::

1. Pandas
2. Numpy

Visualization Tools used::

1. Matplotlib
2. Seaborn
3. plotly

This project is about Exploratory Data Analysis and Predictions on Salaries of Data Domain field.

Explaining the dataset

This Dataset contains:

1. Rows: 607
2. Columns: 9
3. There are 7 Categorical Columns
4. There are 2 Numerical Columns

Dataset Analysis:

work_year: The year the salary was paid.

experience_level: The experience level in the job during the year with the following possible values

- EN : Entry-level

- MI : Junior Mid-level
- SE : Intermediate Senior-level
- EX Expert Executive-level / Director

employment_type: The type of employment for the role

- PT : Part-time
- FT : Full-time
- CT : Contract
- FL : Freelance

job_title: The role worked in during the year.

salary: The total gross salary amount paid.

salary_currency: The currency of the salary paid as an ISO 4217 currency code.

salary_in_usd: The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com).

employee_residence: Employee's primary country of residence in during the work year as an ISO 3166 country code.

remote_ratio: The overall amount of work done remotely, possible values are as follows

- 0 : No remote work (less than 20%)
- 50 : Partially remote or Hybrid
- 100 : Fully remote (more than 80%)

company_location: The country of the employer's main office or contracting branch as an ISO 3166 country code.

company_size: The average number of people that worked for the company during the year

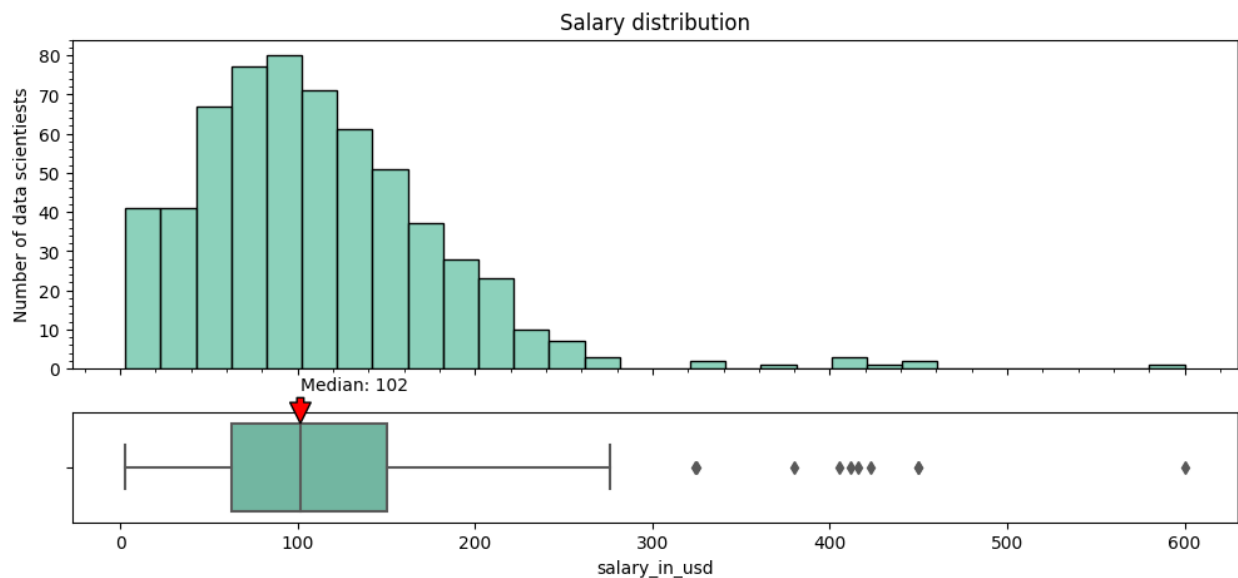
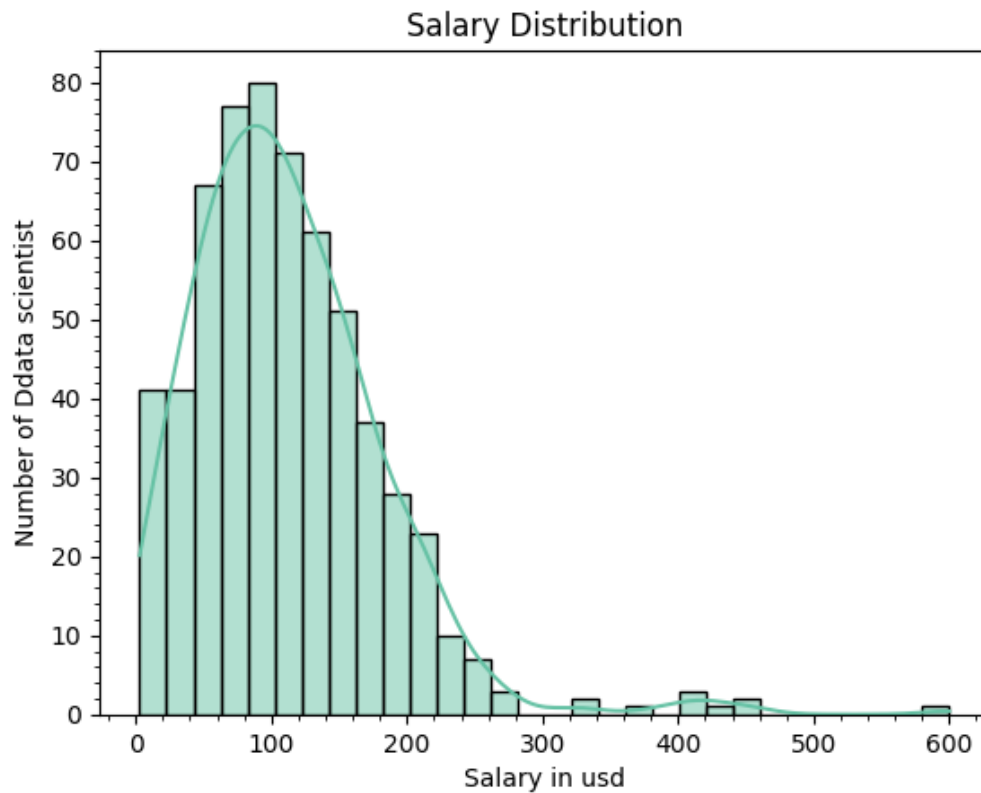
- S : less than 50 employees (small)

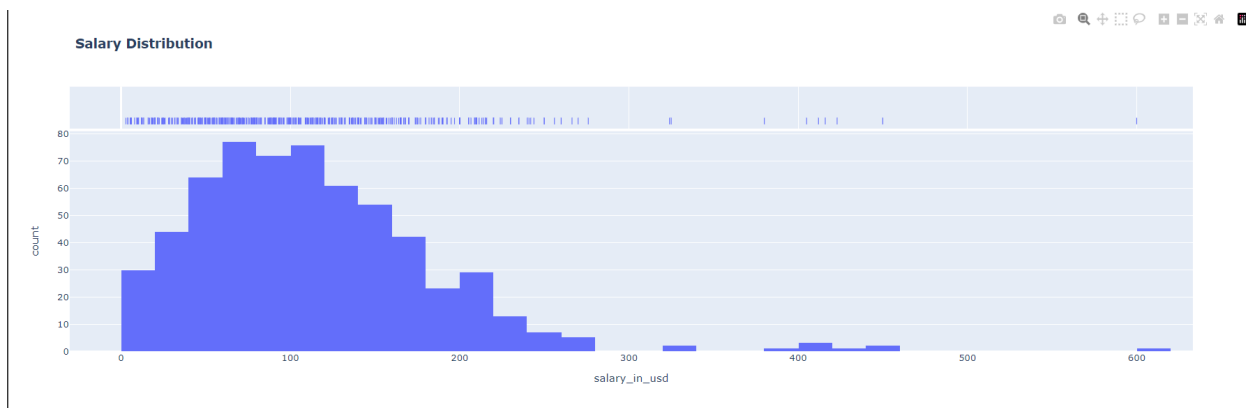
- M : 50 to 250 employees (medium)
- L : more than 250 employees (large)

Exploratory Data Analysis

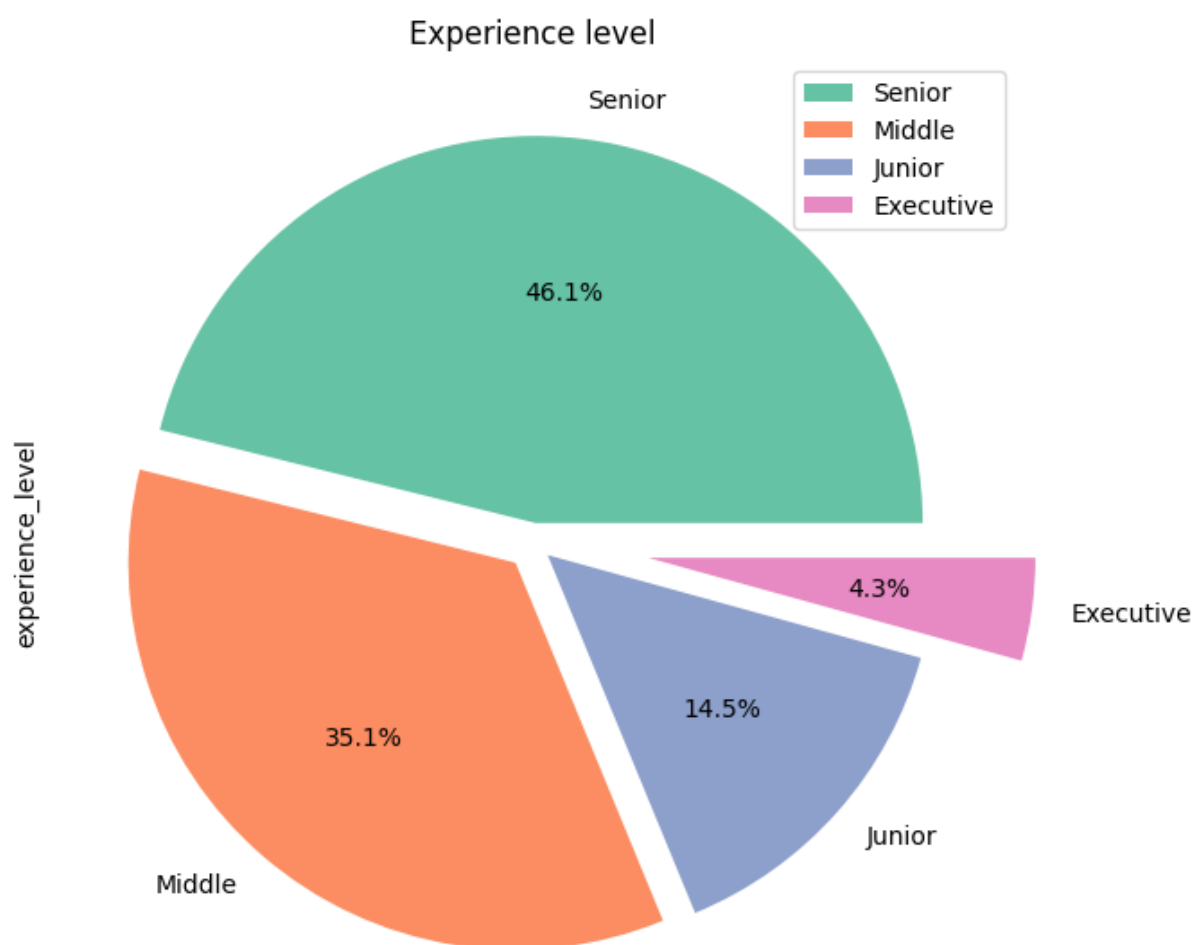
Let's See Univariate Analysis

Distribution of salary





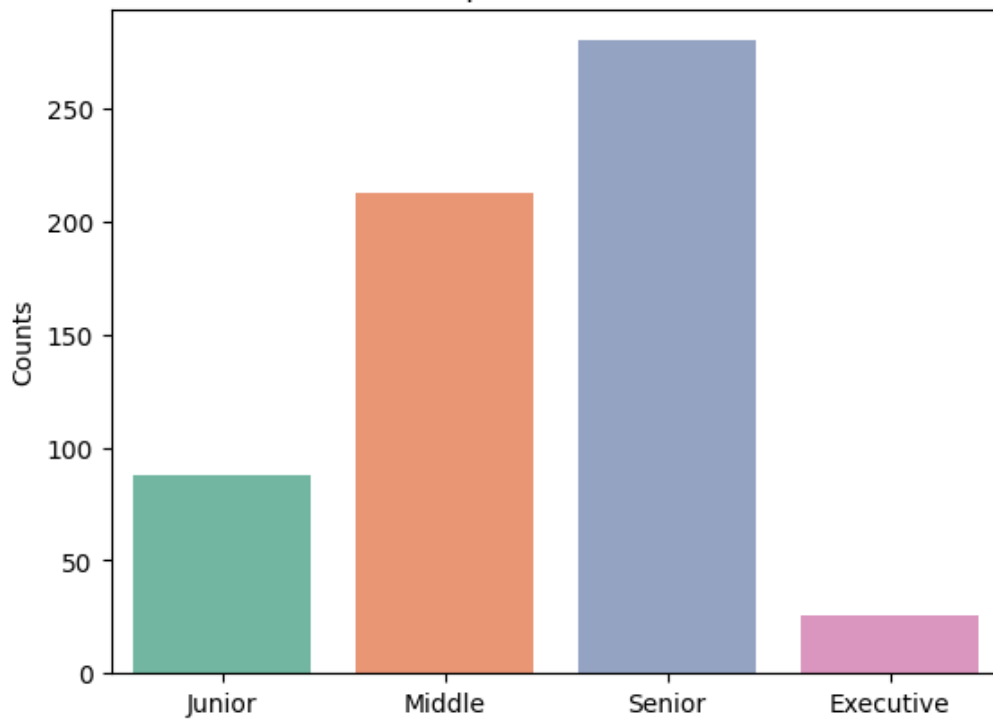
experience_level



Total Jobs based on Experience Level



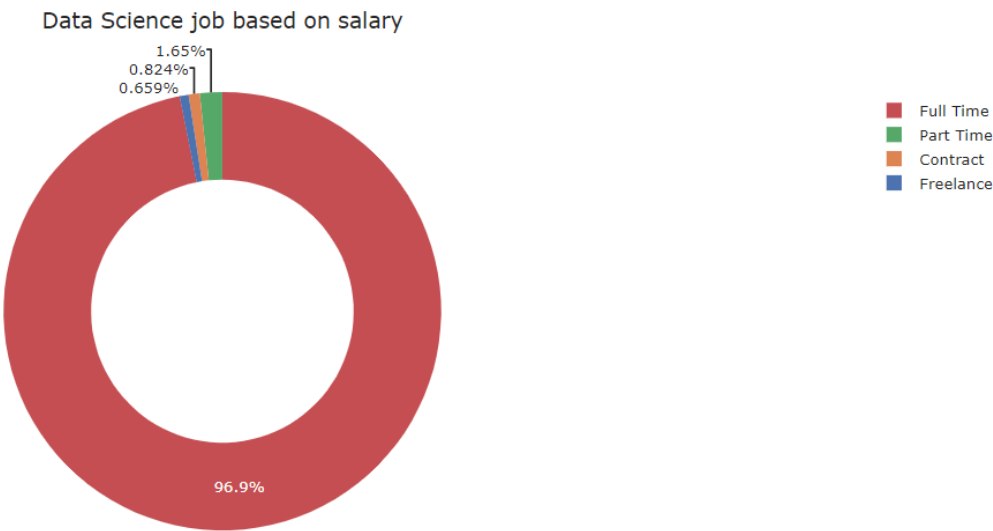
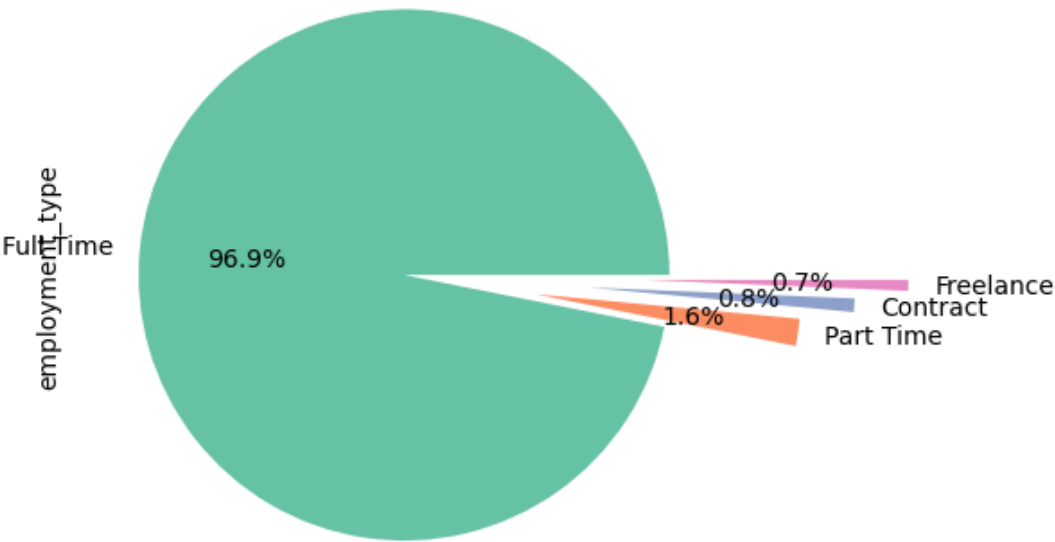
Experience Level

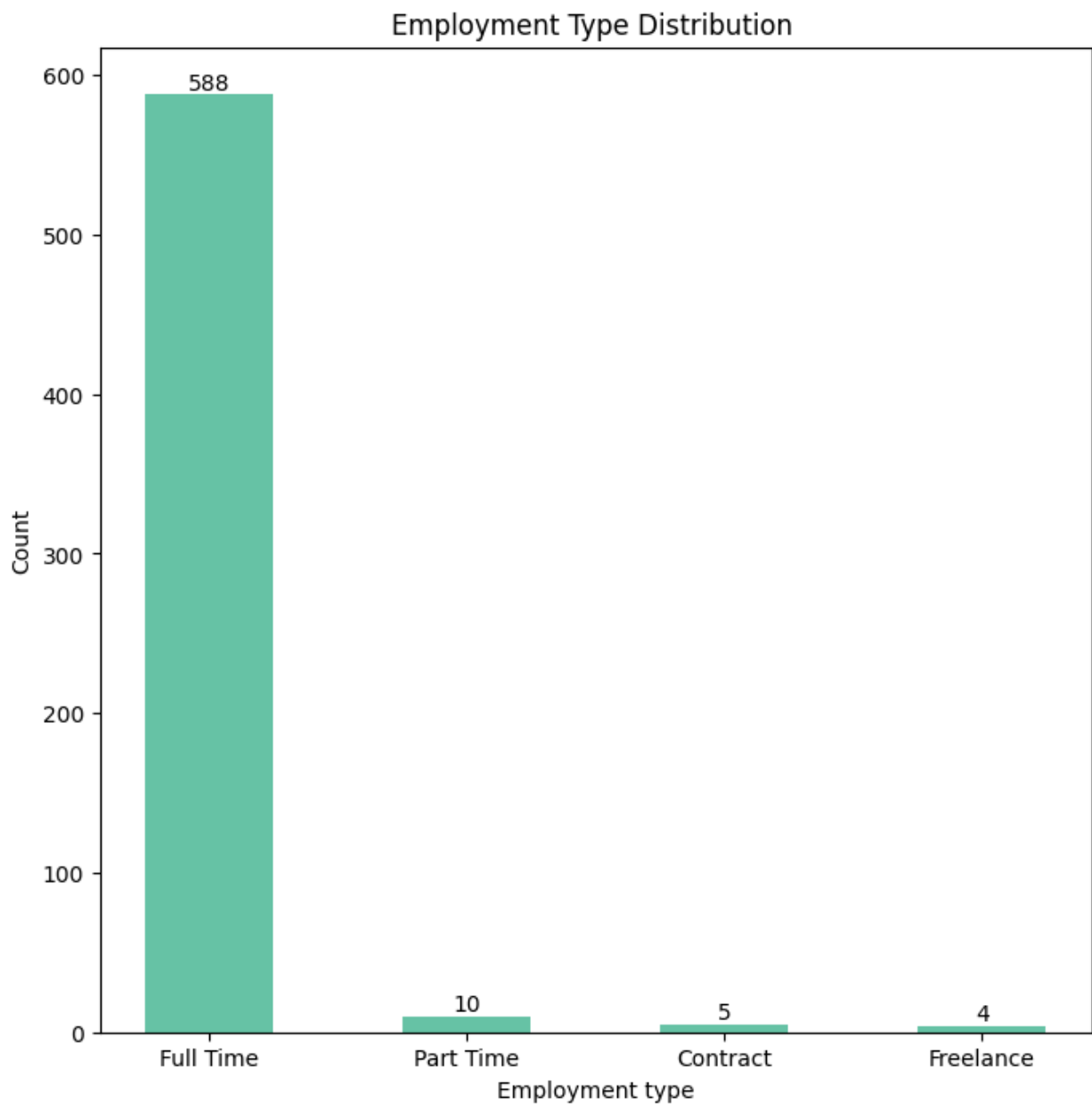


Observation:

We can see from the above pie chart that senior level jobs mostly requires experience.

employment_type

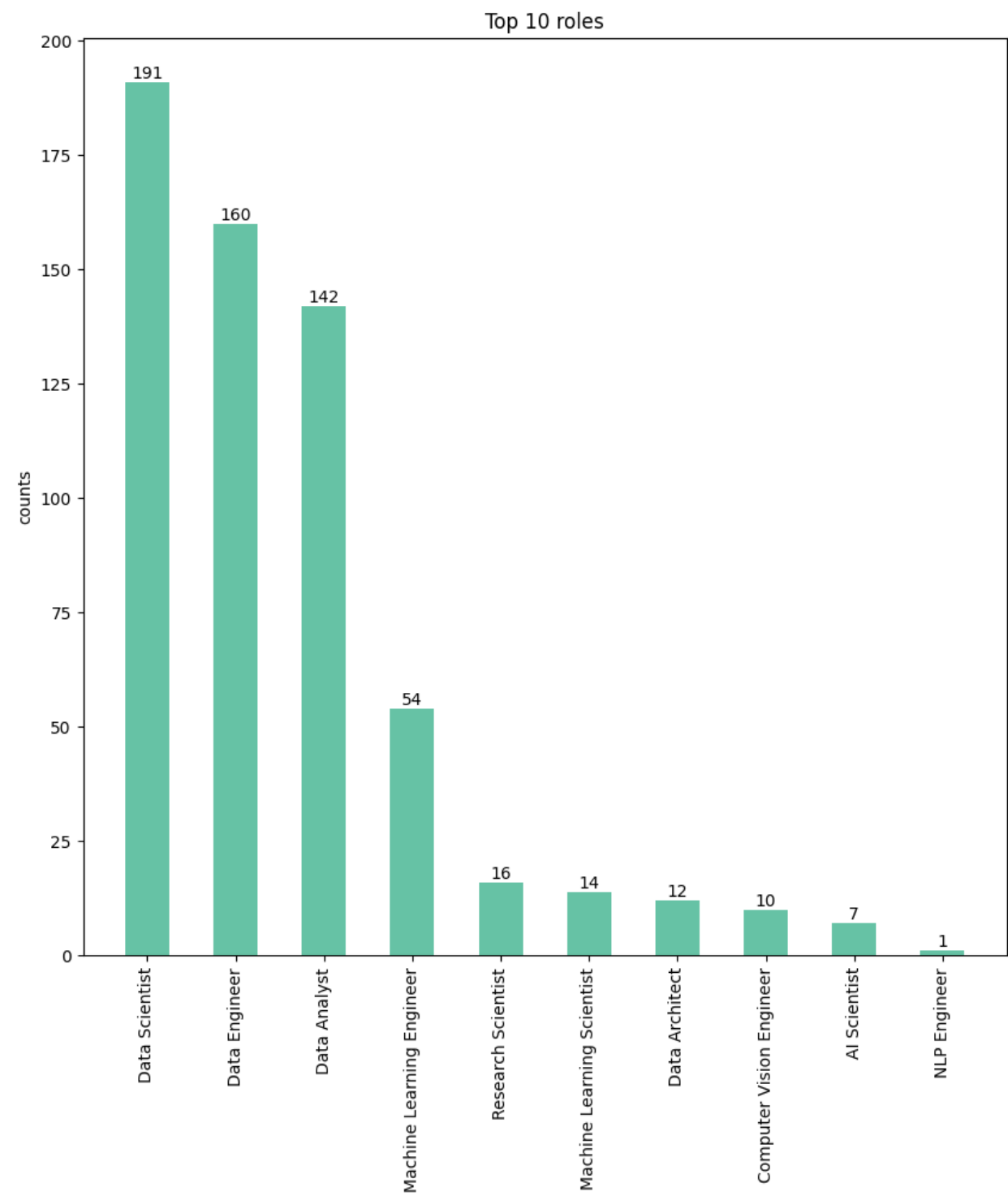


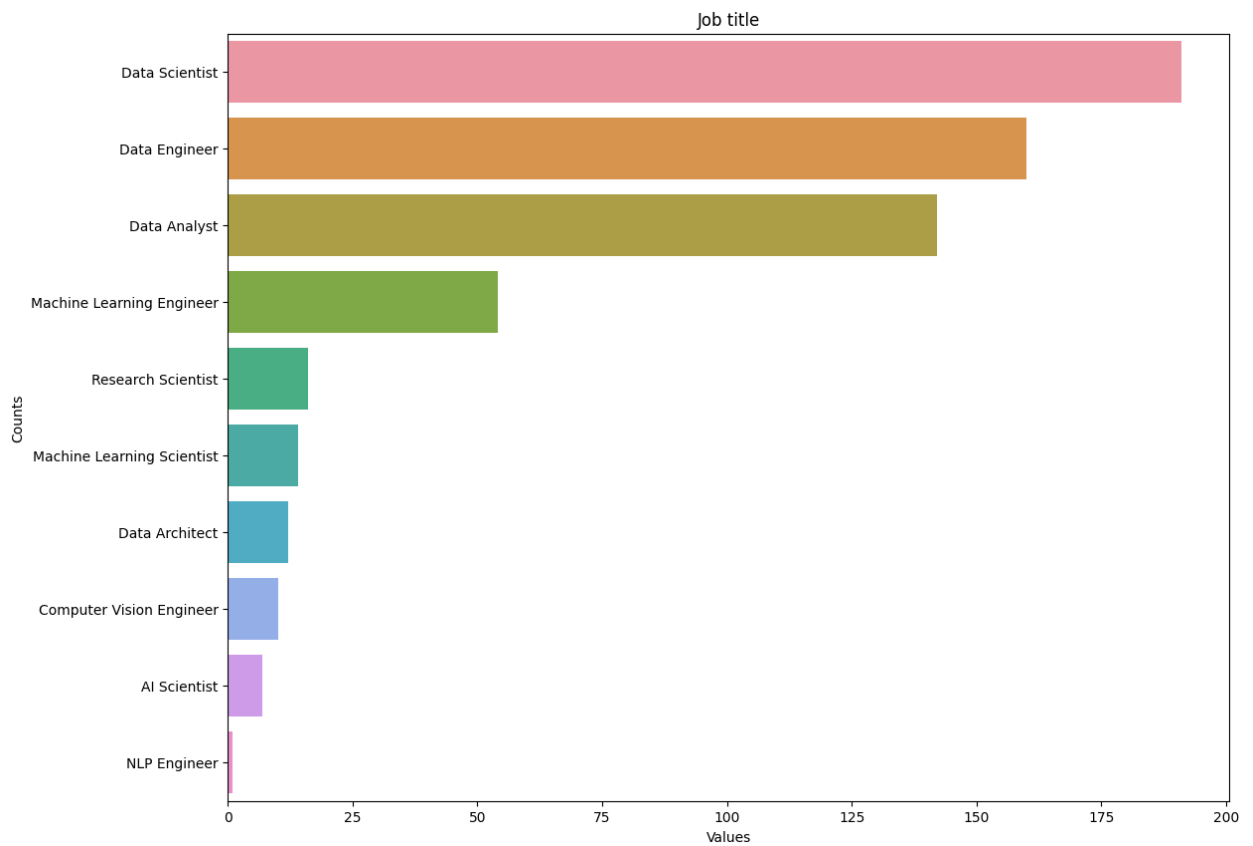
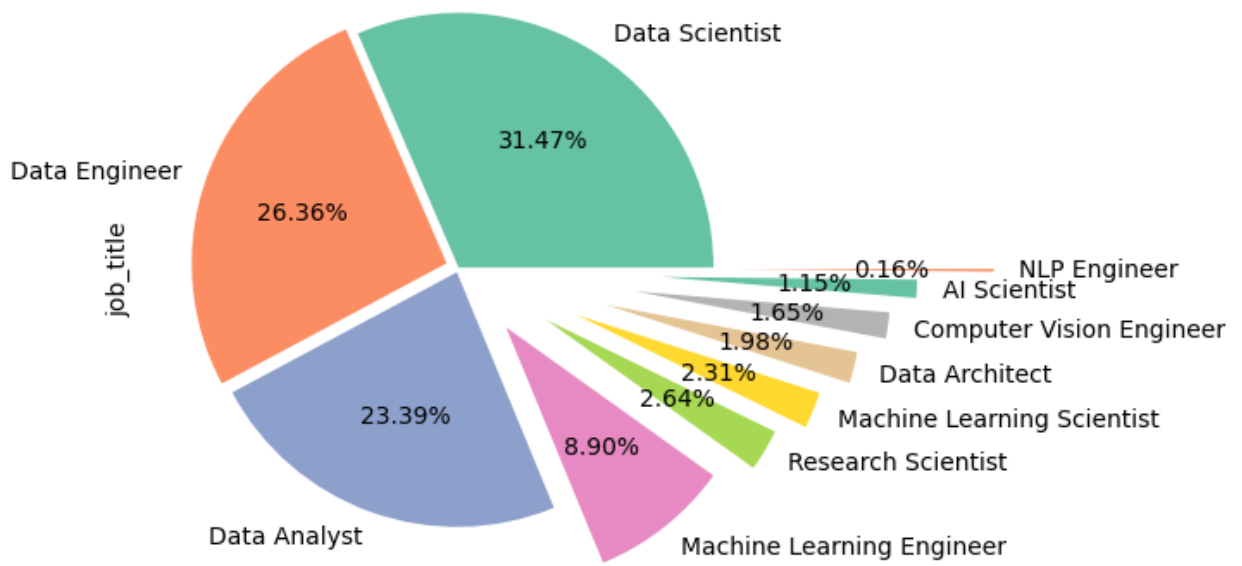


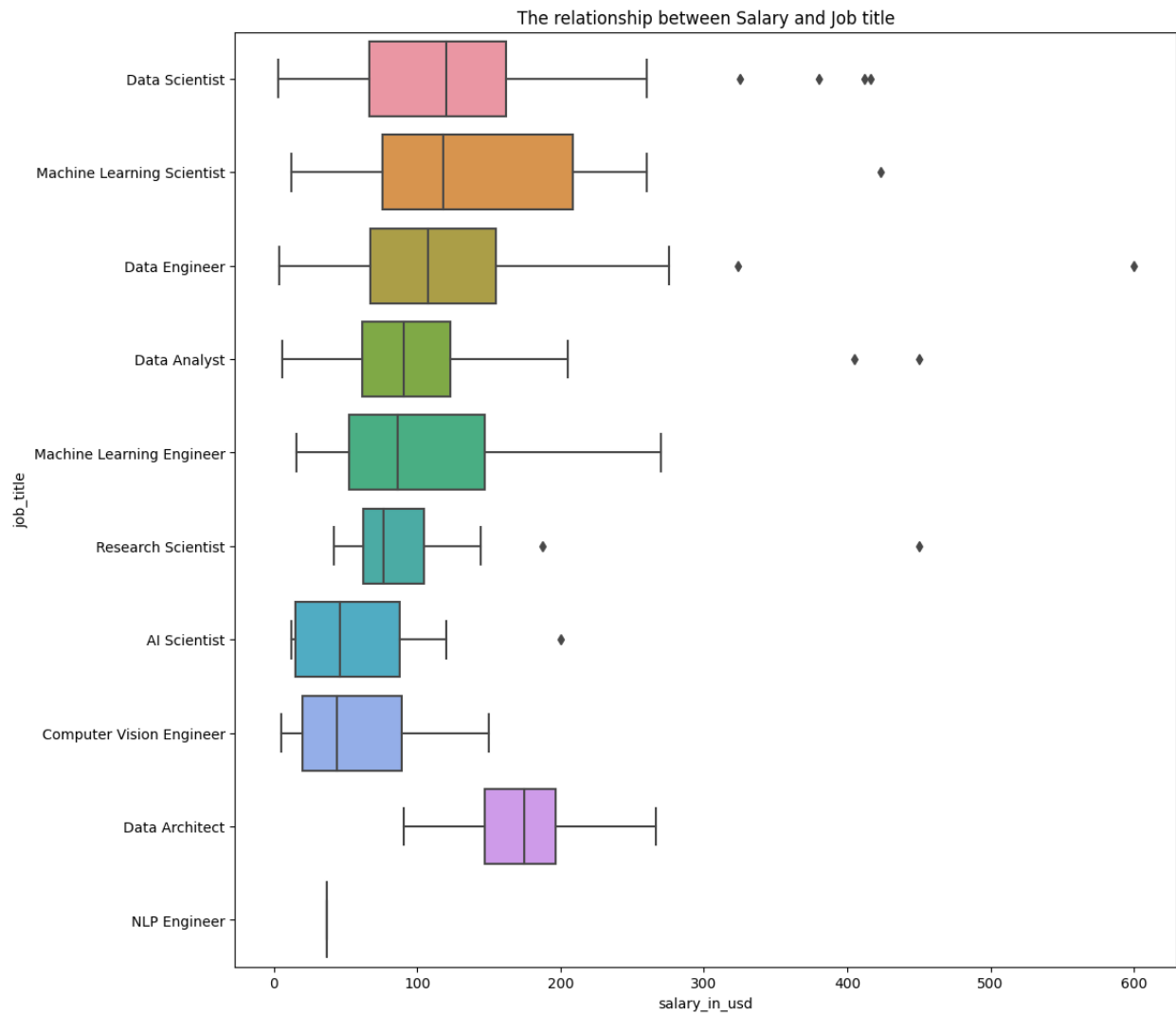
Observation:

We can see that 96.9% of the jobs are Full-time jobs.
Contract and freelancing jobs are not given that much importance in Data Science.
Part Time jobs are also rare.

job_title







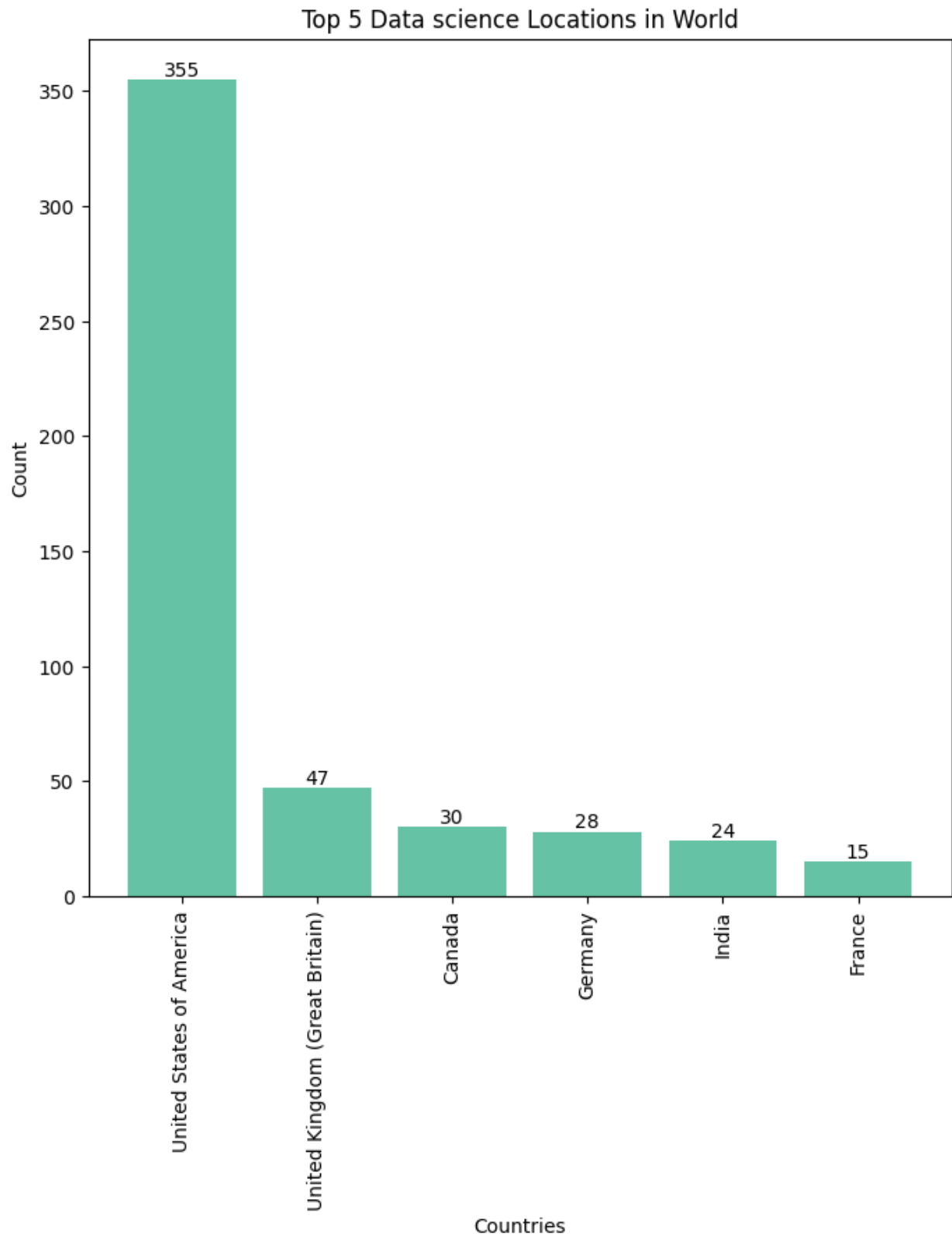
Observations:

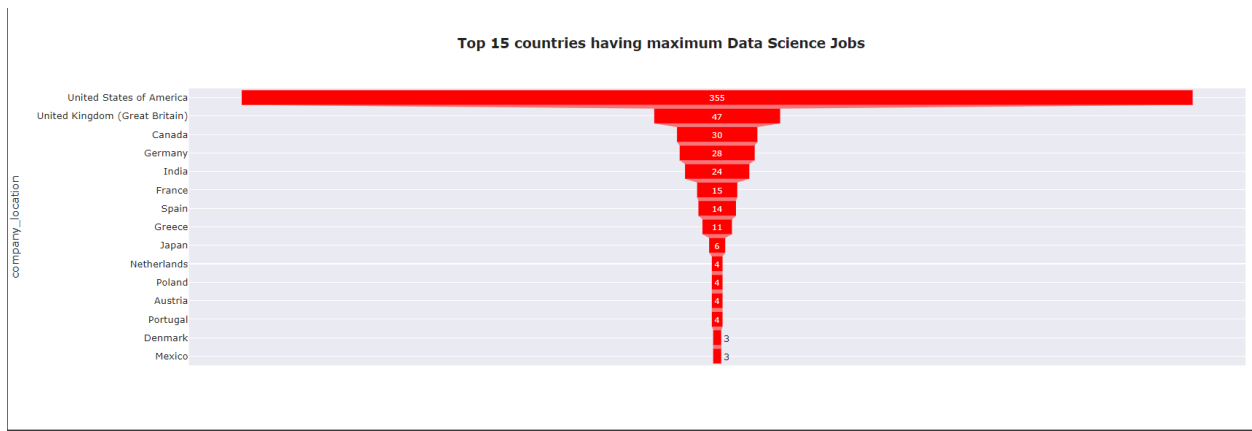
After Centralize we got 10 main job category.

Data Scientist;s Average Salary gain top position.

Machine learning Engineer's Salary got Second Position.

Employee Residence and Company location

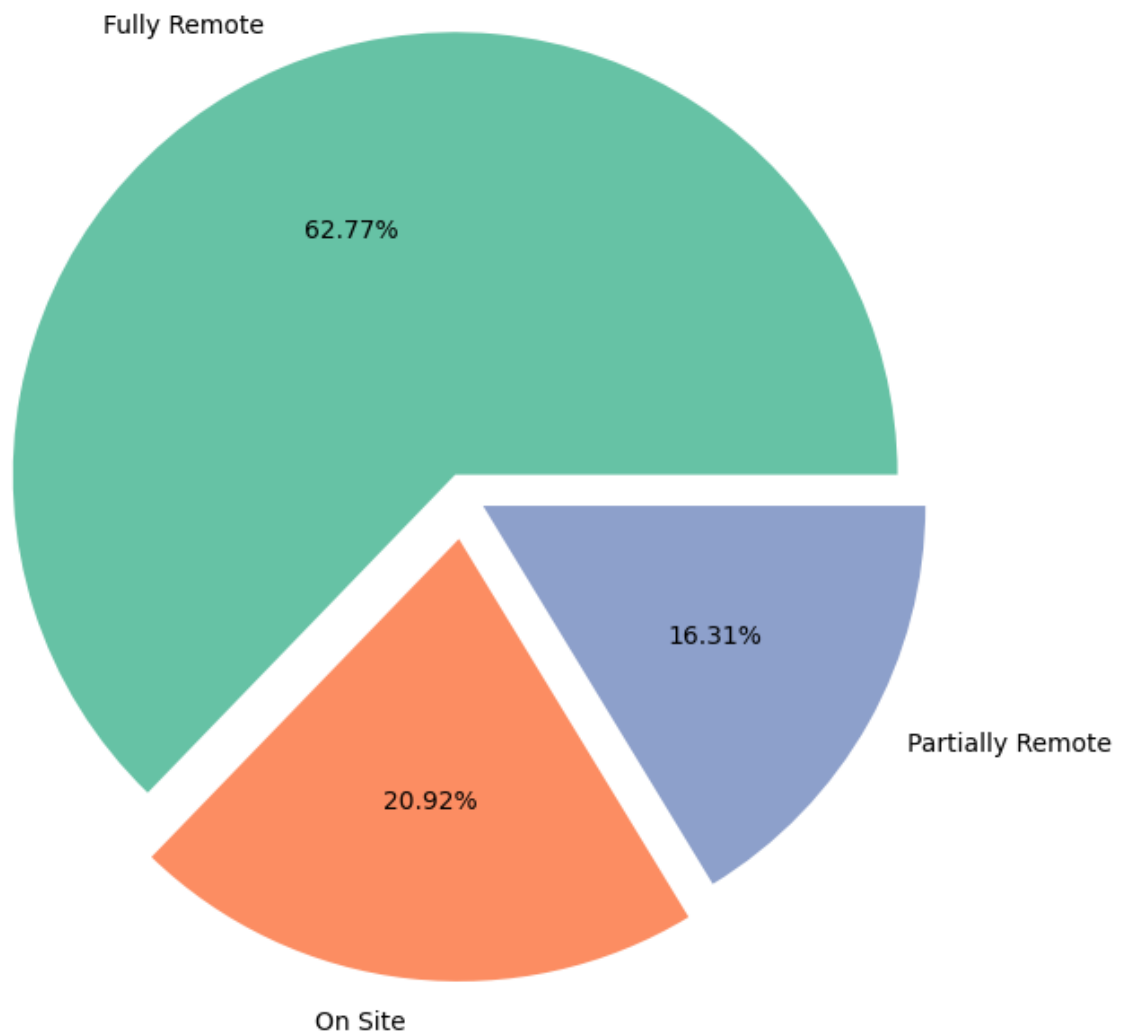


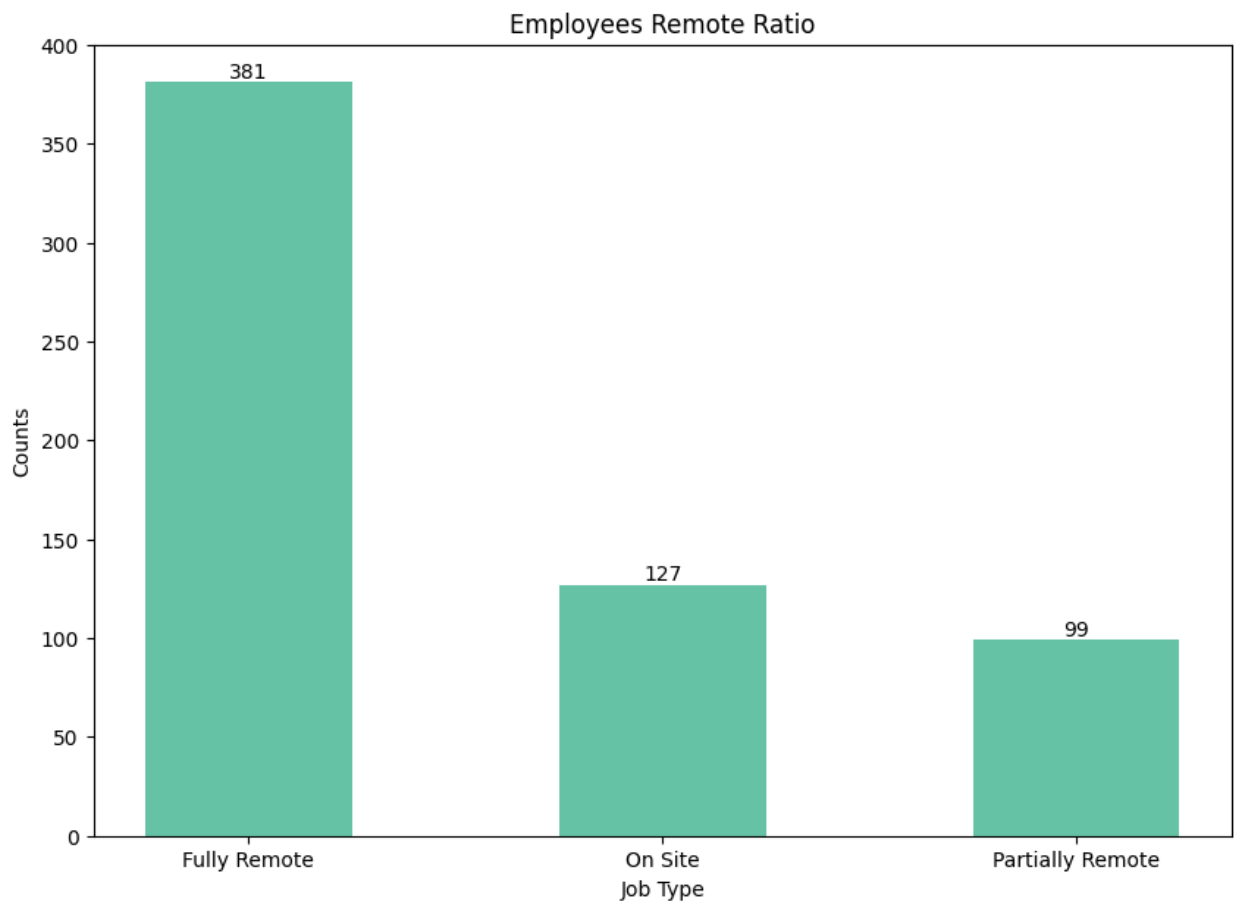


Observations:

Maximum Data Domain Jobs seen in the USA.
Then UK, Canada, Germany, India, France so on

Remote ratio



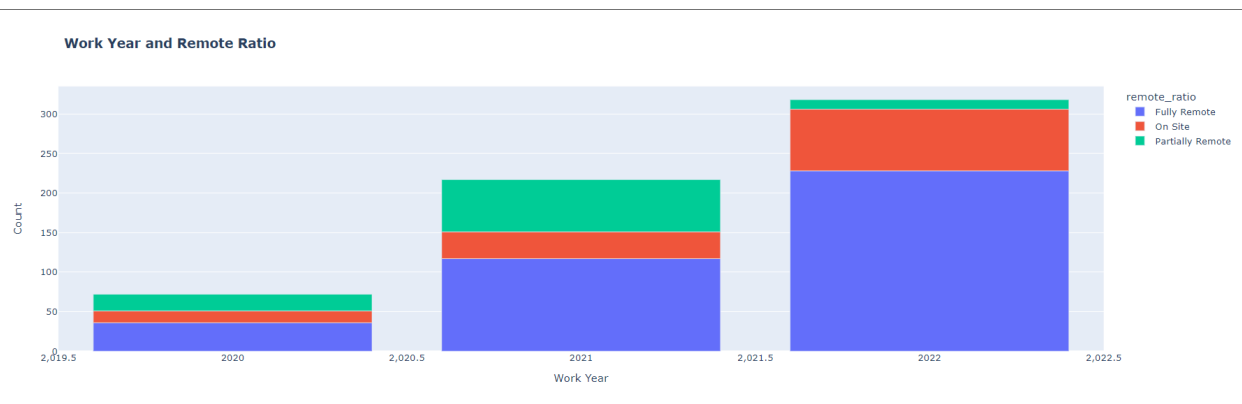


Observations:

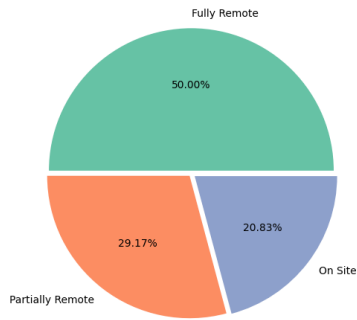
Fully Remote job type ratio is way more than other types.
Then On-Site job type
Lastly, Partially or Hybrid job type.

Let's See Bivariate Analysis

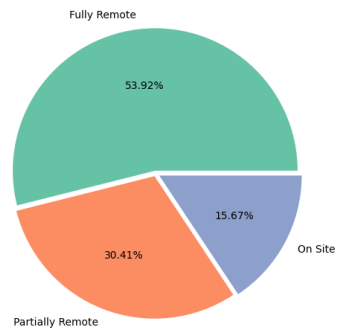
Work Year and Remote Ratio



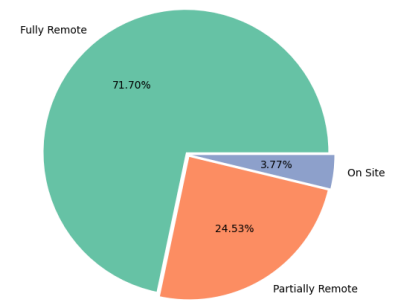
Remote ratio in the year 2020



Remote ratio in the year 2021



Remote ratio in the year 2022



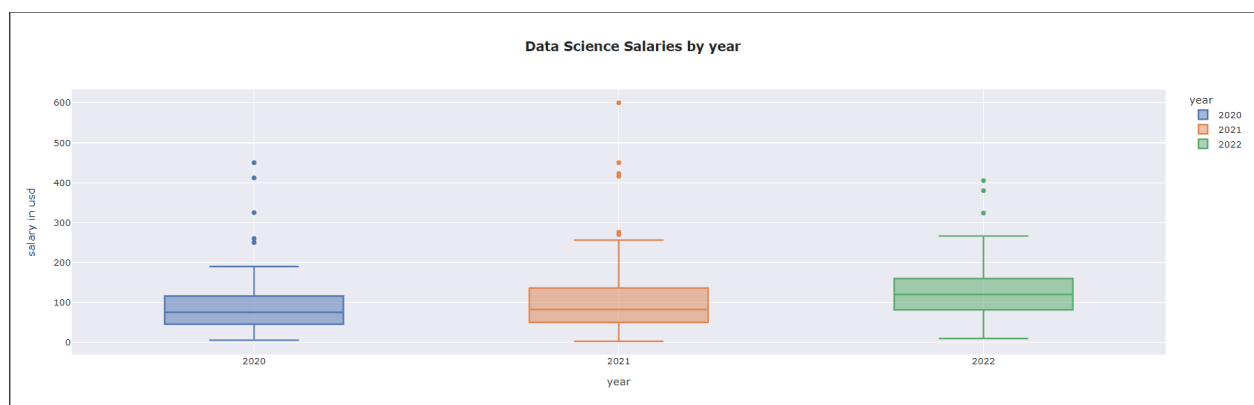
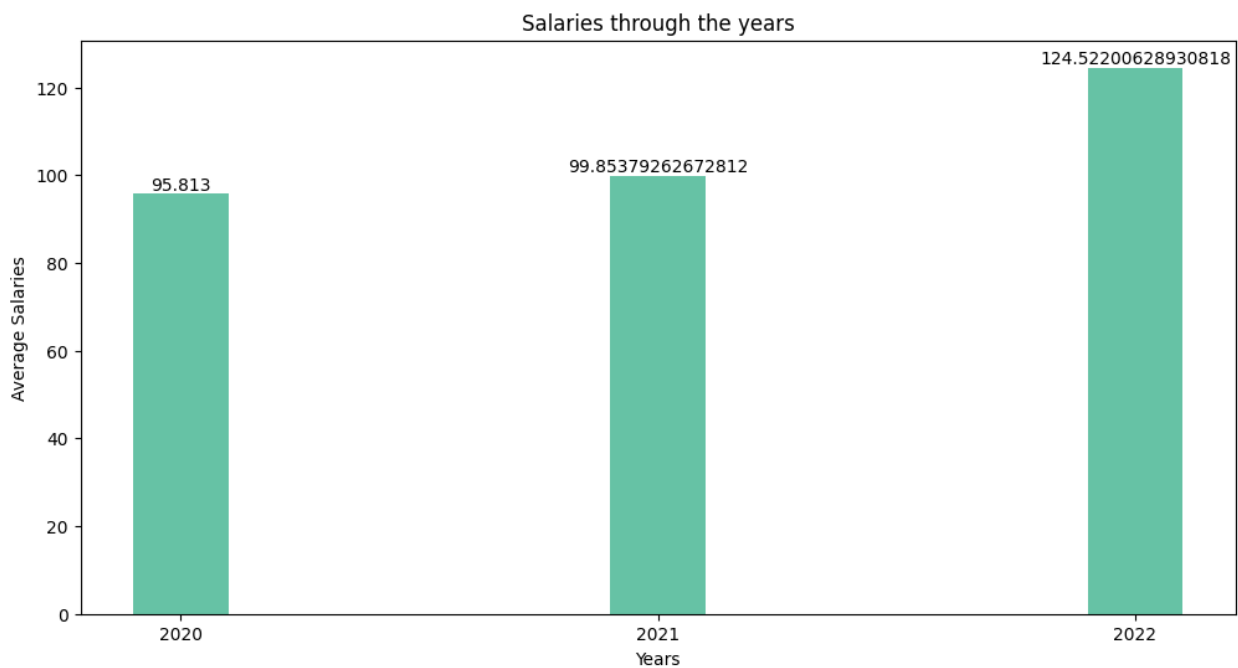
Observations:

Fully Remote ratio Increased each year.

Partially or Hybrid ratio Increased in 2021 year but decreased 2020.

On Site ratio decreased each year.

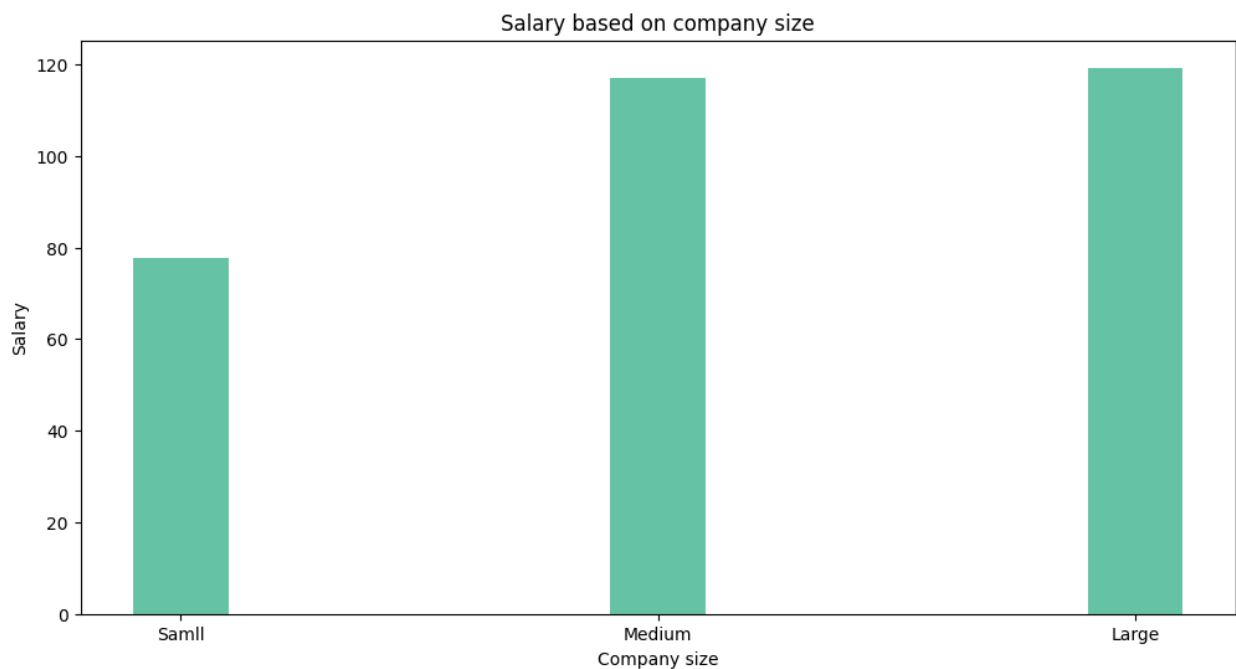
Relationship between salary and work year



Observations:

Average salaries increased each year.

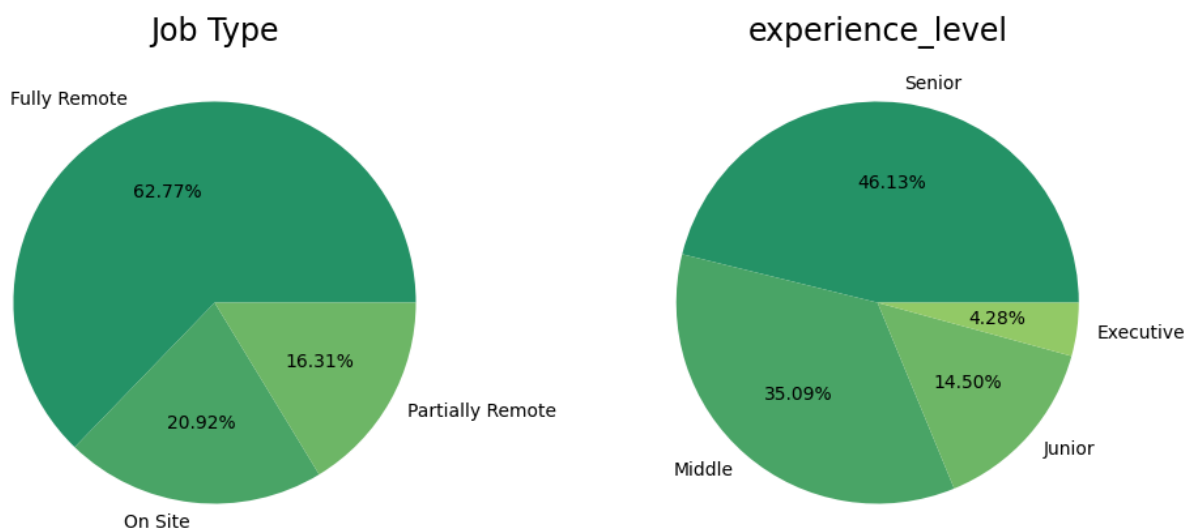
Relationship between Salary and company size



Observations:

comparatively Larger company pays more salary.

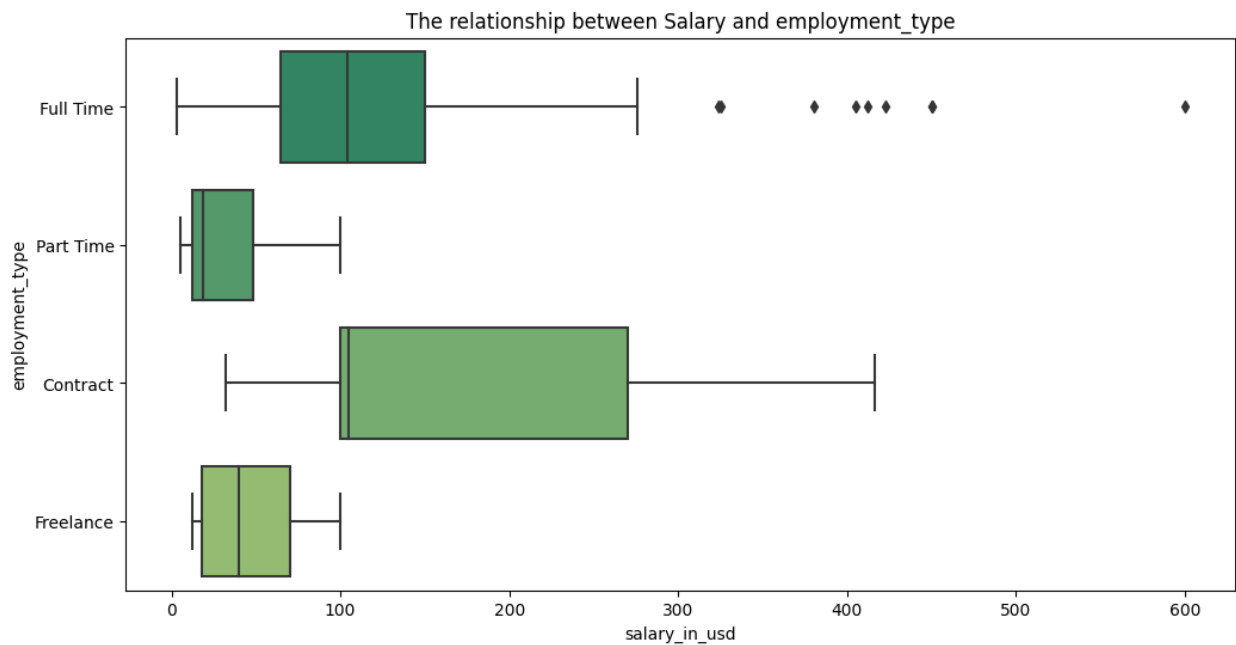
Job Types and Experience Level distributions



Observations:

Fully Remote job has the highest number of job opening.
Senior level job has the highest number of job opening.

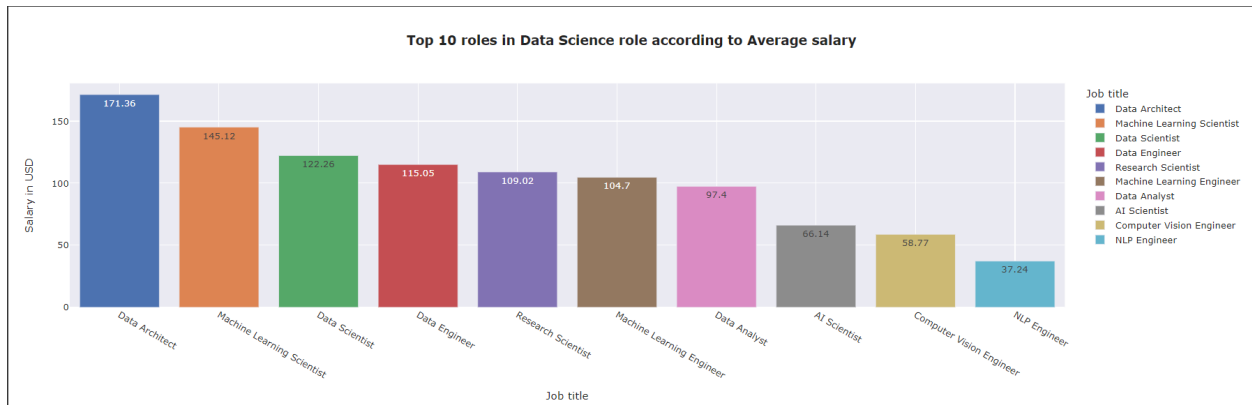
relationship between Salary and employment_type



Observations:

Contract based employee get higher amount Average salary. Then Full time based employee.
Then Freelance and part time based employee.
Some full time based employee gets higher amount than others.

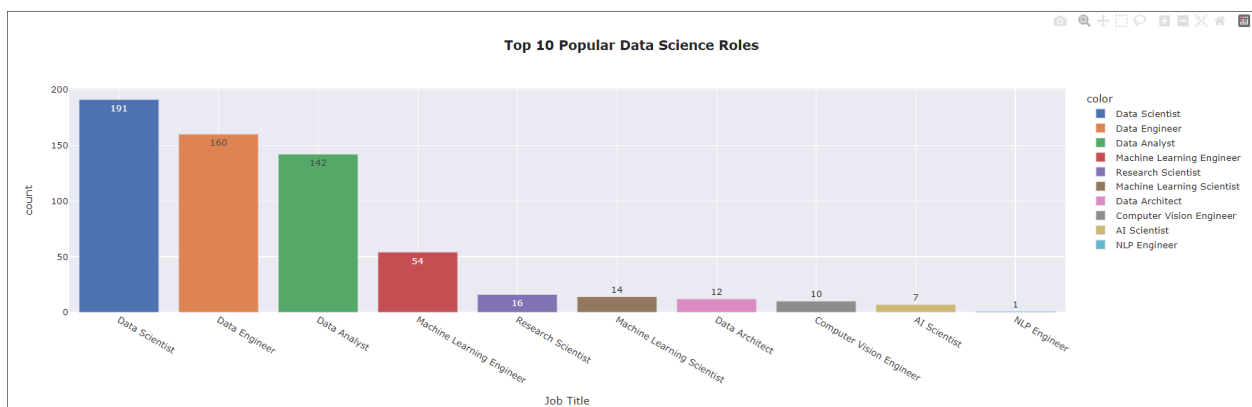
Top 10 Roles in Data Science based on Average Pay



Observations:

Average Salary of Data Architech is Higher than others.

top 10 popular Data science role

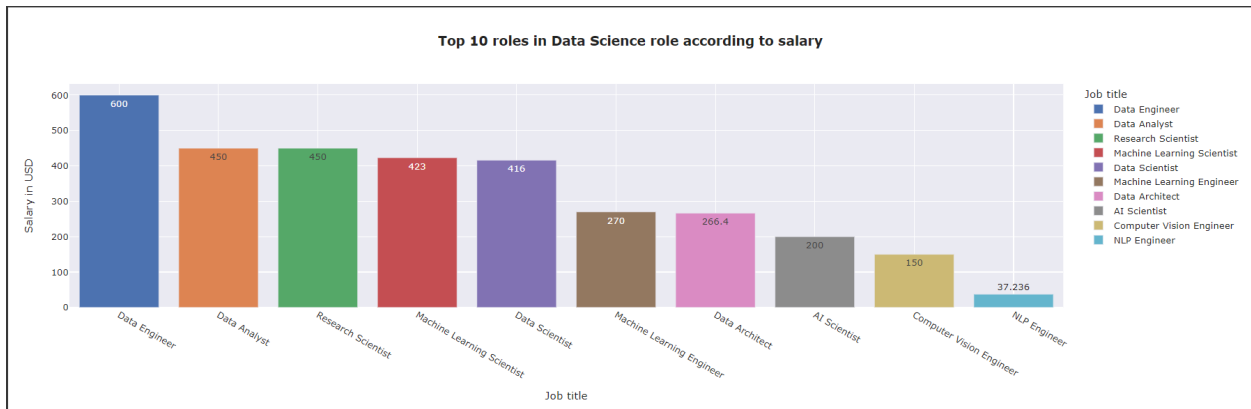


Observations:

There are more jobs in Data Science Field.

Then Data Engineer and Data Analyst

Top 10 roles in Data Science role according to salary

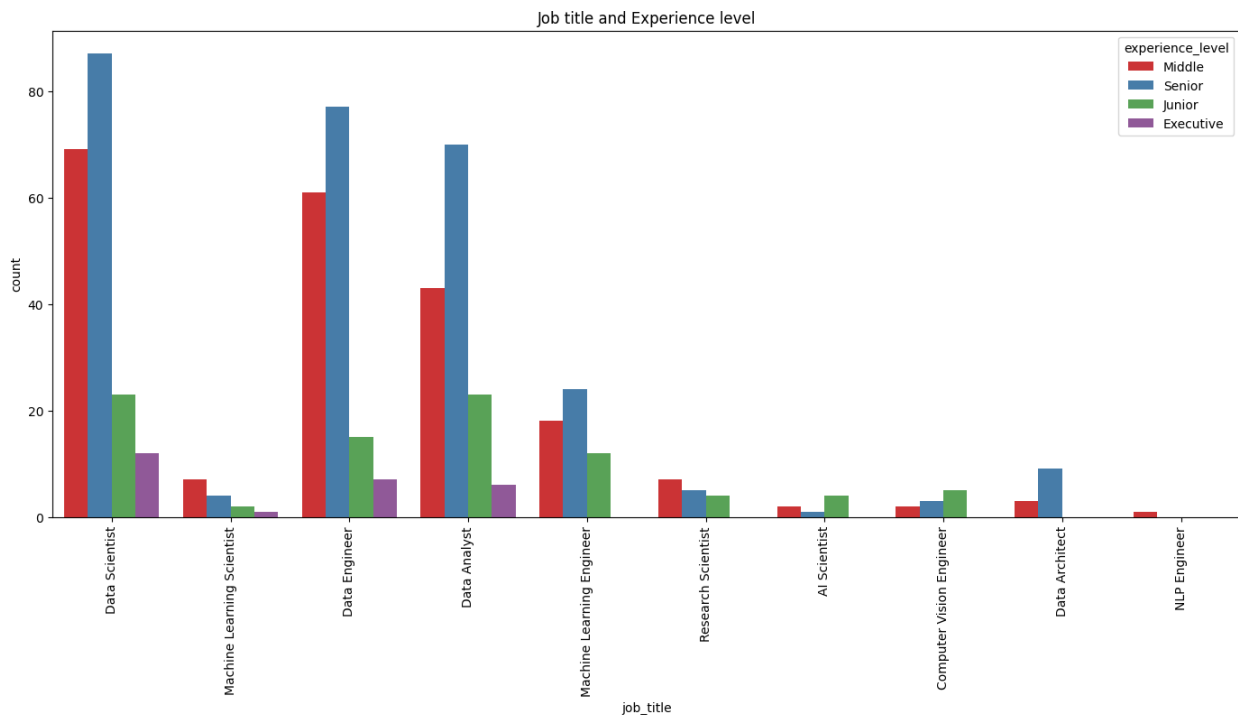


Observations:

According to Salary, Data Engineer earn more than others.

Data Analyst, Research Scientist, Machine learning Engineer and Data Scientist earn almost same.

Job title and Experience level



Observations:

we can see that these positions, Experience Level: Senior are higher than others.

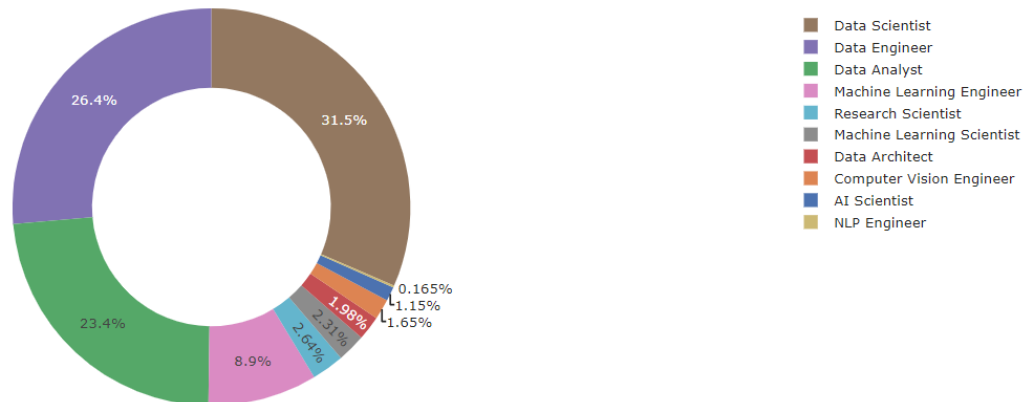
1. Data Scientist,
2. Data Scientist,
3. Data Engineer,

4. Data Anlyst,
5. Machine Learning Engineer, &
6. Data Architech

that means company are hiring more Senior levels employee in this positions.
Middle experience level employee comes second at these positions.

Relationship of Job Title and Salary

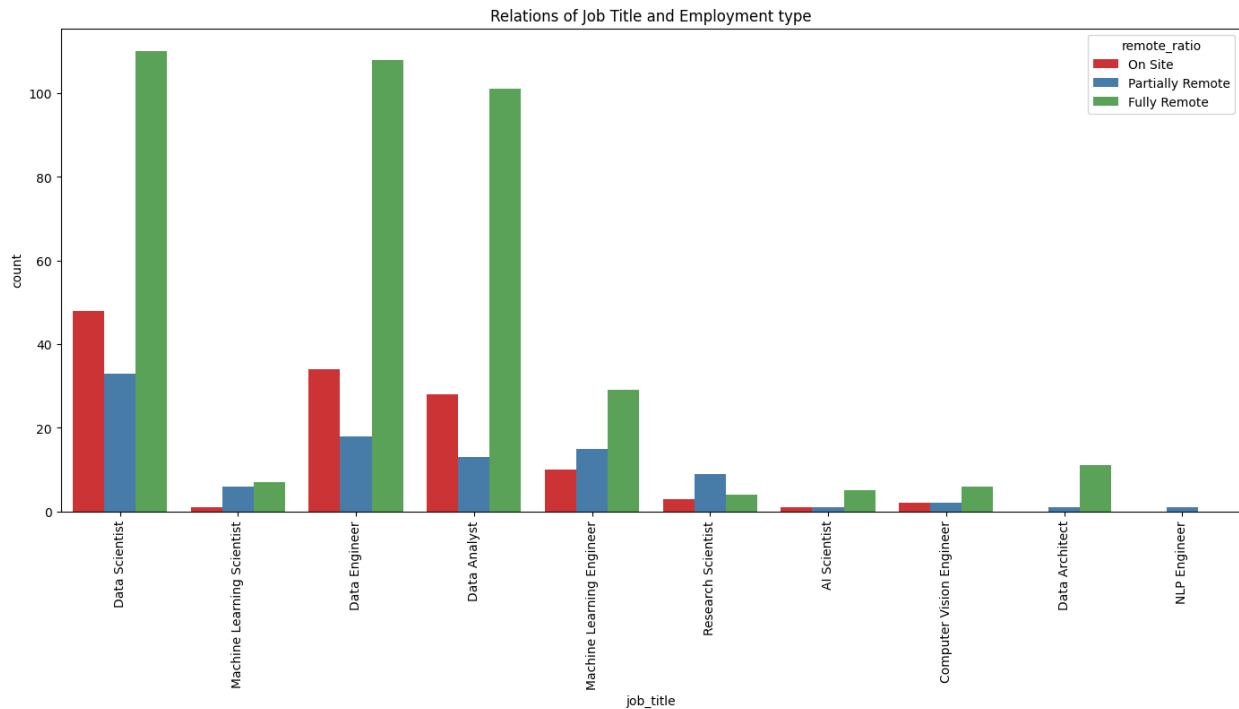
Top 10 roles according to Experience Level



Observations:

Senior Employee is more in Almost every Job role.

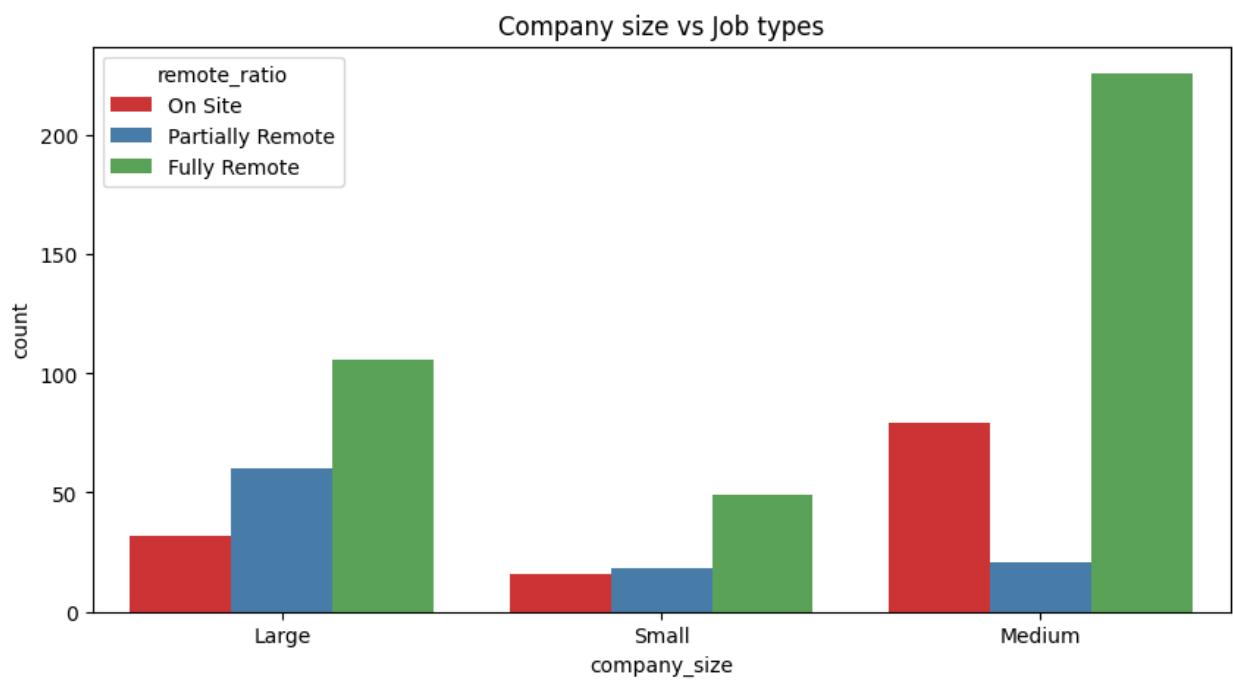
Job title and Employment Type



Observations:

Count of Fully Remote is in Top position in most job title.

Company Size VS Job Type Counts



Observations:

Small-sized companies provide minimum onsite opportunities,
whereas Medium-sized companies provide maximum remote work opportunities.

Relationship between Salary and Expreience

Mean Salaries according to Experience Level



Maximum Salaries according to Experience Level

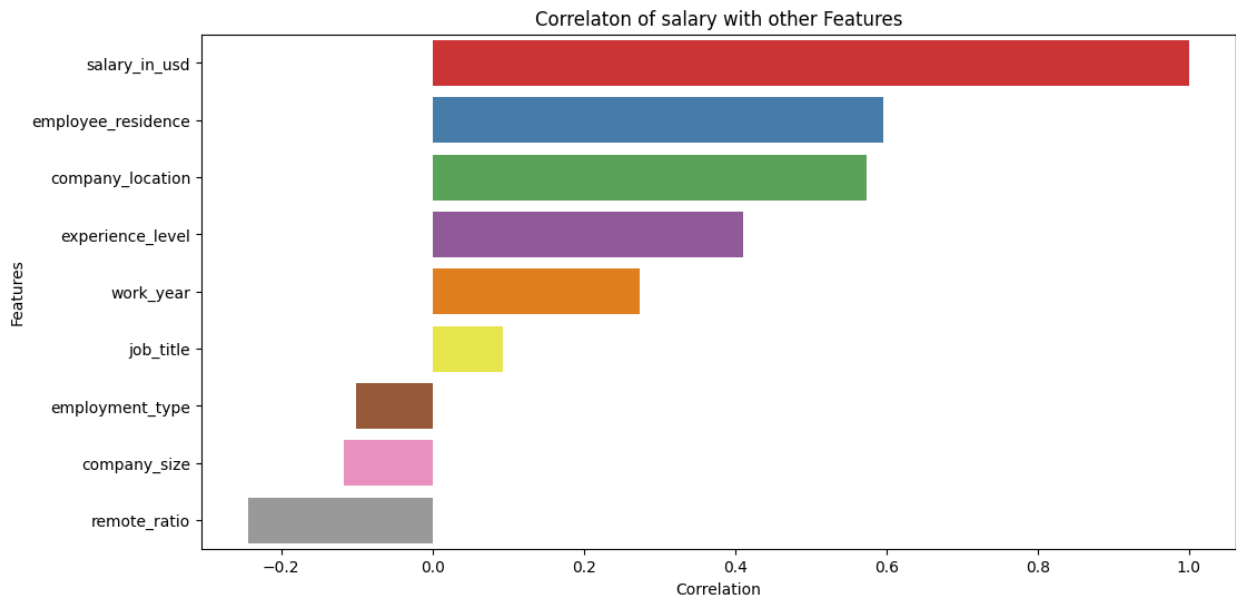


Observations:

According to Experience level, Mean & Maximum Salary of Executive levels got first position.
'Senior Level' in Mean salary section got second position.
'Middle Level' in Maximum salary section got second position.

Multivariate Analysis

Corelation between Features



Observations:

The Highest correlation of salary is with:

1. Employee Residence
2. Company Location
3. Experience Level

Employment type & Company size has a Negative correlation

job type and company size VS salary

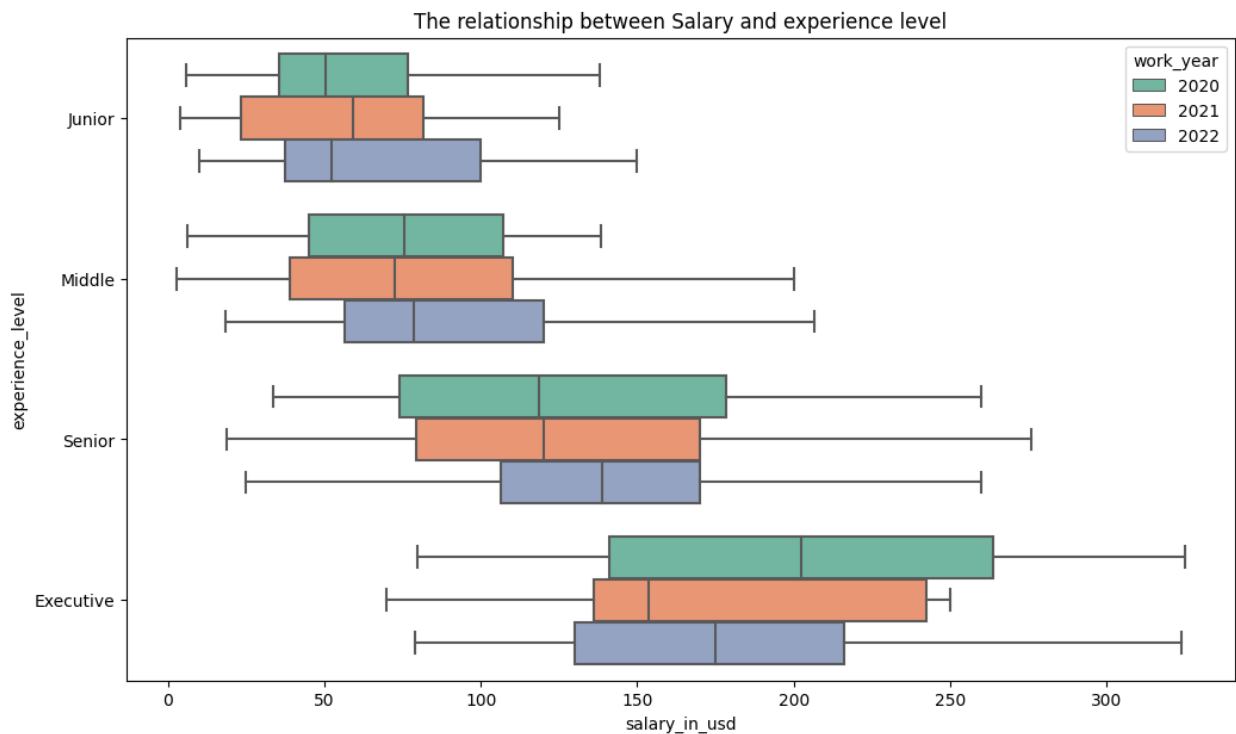
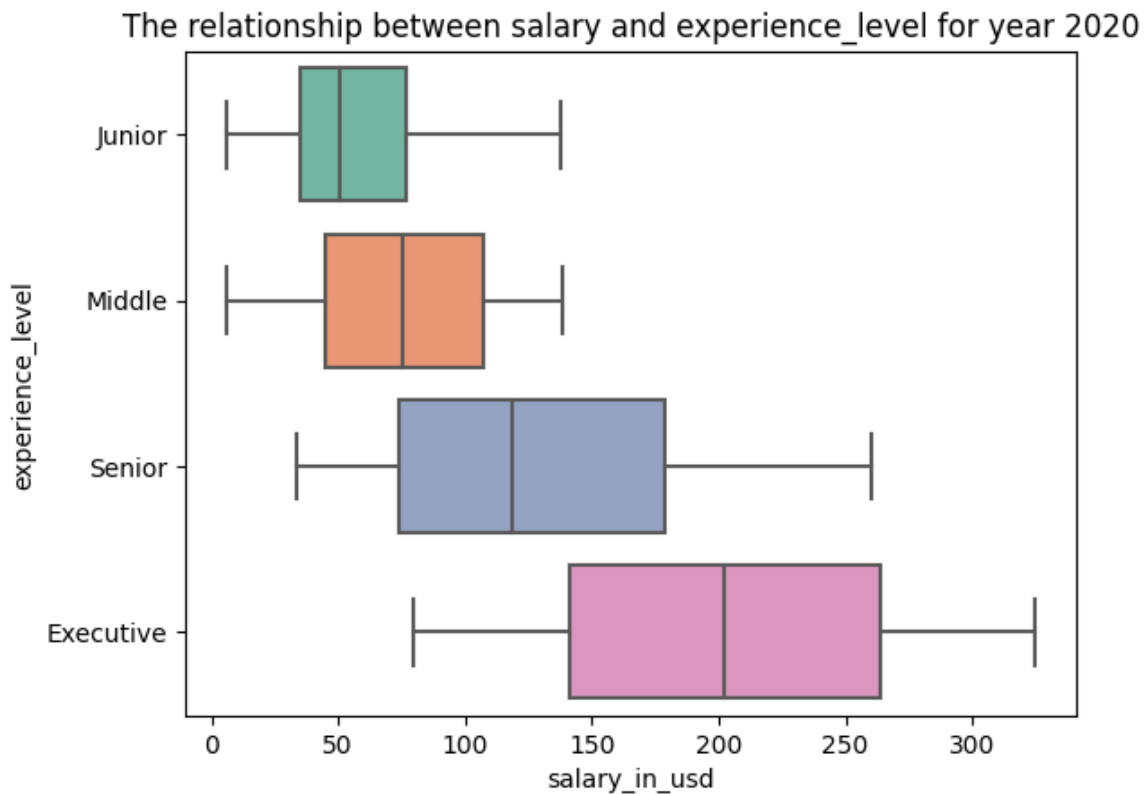
![[image]](<https://github.com/AzadMehedi/Projects/assets/49702660/b8f3b8b8-199f-43fe-ae82-e1fe58ef352a>)

Observations:

Lowest salary is seen in onsite job for a small company.

Higher salary is seen in Fully remote job for a laege company.

The relationship between salary_in_usd and experience_level and work_year



1. in 2020, 2021, 2023 : Executive levels employee got more average salary than others.
2. in 2020, 2021: Executive levels employee got max salary than others.
3. in 2021: Middle levels employee got max salary than others.

4. in 2020, 2022: Executive levels employee got max salary than others.

Regression Analysis

1. The factorize function in Pandas assigns a unique numerical value to each distinct category in a categorical column. It returns two values: a new column with the numerical codes for each category, and an array that contains the unique categories themselves.
2. In the given code, the line `data[column], _ = pd.factorize(data[column])` assigns the numerical codes to the column in the data DataFrame, replacing the original categorical values. The underscore `_` is used to discard the array of unique categories returned by factorize because it is not needed in this code.
3. By applying factorize to each categorical column, the original categorical data is transformed into numerical data, allowing it to be used as input in the subsequent regression analysis.

Applying Random Forest Regressor

OLS Regression Results						
Dep. Variable:	salary_in_usd	R-squared (uncentered):	0.954			
Model:	OLS	Adj. R-squared (uncentered):	0.953			
Method:	Least Squares	F-statistic:	1236.			
Date:	Sun, 13 Aug 2023	Prob (F-statistic):	0.00			
Time:	15:12:25	Log-Likelihood:	-2894.8			
No. Observations:	607	AIC:	5810.			
Df Residuals:	597	BIC:	5854.			
Df Model:	10					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
new_mean_salary	0.9907	0.134	7.416	0.000	0.728	1.253
new_minimum_salary	0.0065	0.077	0.085	0.933	-0.145	0.158
new_maximum_salary	0.0019	0.058	0.033	0.974	-0.112	0.115
job_title_Data Engineer	0.1090	3.088	0.035	0.972	-5.956	6.174
experience_level	0.1838	1.834	0.100	0.920	-3.419	3.787
remote_ratio	-0.1047	1.479	-0.071	0.944	-3.009	2.800
company_location_Germany	-3.1171	5.661	-0.551	0.582	-14.236	8.002
job_title_Data Scientist	-0.3553	2.835	-0.125	0.900	-5.923	5.212
work_year	4.387e-05	0.003	0.017	0.987	-0.005	0.005
job_title_Data Architect	0.0127	8.615	0.001	0.999	-16.907	16.932
Omnibus:	320.522	Durbin-Watson:	2.109			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7632.787			
Skew:	1.817	Prob(JB):	0.00			
Kurtosis:	19.988	Cond. No.	1.51e+04			

Observations:

Top 10 feature importance are:

[index----feature name-----score]

1. new_mean_salary 0.853894
2. new_minimum_salary 0.075712
3. new_maximum_salary 0.060554
4. job_title_Data Engineer 0.002372
5. experience_level 0.001205
6. remote_ratio 0.001189
7. company_location_Germany 0.001139
8. job_title_Data Scientist 0.000671
9. work_year 0.000573
10. job_title_Data Architect 0.000489
11. squared value of Accuracy is : 0.91%. That means our model/ regression line can fit or touch 91% feature variable. squared value of Accuracy is : 0.91%. That means our model/ regression line can fit or touch 91% feature variable.
12. Adj. R-squared (uncentered): 0.953 means if we increase the feature variables then our accuracy will increase. but if we increase the feature variables & R-squared increase but Adj. R-square doesn't then adding more feature doesn't increase the model. here R-squared = 0.954 & Adj. R-square = 0.953 which is says R-squared ~ Adj. R-square. so no need to add more feature variables.
13. Prob (F-statistic): the value of Prob (F-statistic) should below (0.05). if Prob (F-statistic) > 0.05 then our model is not good for regression. its very important Prob (F-statistic) < 0.05
14. $P > |t|$ - Those which coefficients $p > |t|$ value is more than 0.05 ($p > |t| > 0.05$) we can remove them because they are not important for our model.
15. Df Model- Number of Independent features model consumed. Dep. Variable- Number of dependent feature (target/ label)
16. Mean Absolute Error - (MAS) : 9.83875394333003
17. Mean Squared Error -(MSE) : 331.0808322445805
18. Root Mean Squared Error -(RMSE) : 18.195626734041905

19. Accuracy-(R2 Score): 0.9136138665760188 ~ 91%

20. Now comparing the Accuracy and error of different models


21. Observations:

Model Name-----Accuracy

Random Forest: 0.945127
GradientBoosting: 0.941213
XGB: 0.940079
DecisionTree: 0.924845
ExtraTree: 0.917075
LGBM: 0.893387

KNeighbors: 0.389352

Comparing Model Accuracy

	R_square	MSE	MAE	RMSE	
Random Forest	0.945127	0.196458	0.116022	0.340620	
GradientBoosting	0.941213	0.203342	0.117459	0.342724	
XGB	0.940079	0.205294	0.122234	0.349620	
DecisionTree	0.924845	0.229915	0.124327	0.352601	
ExtraTree	0.917075	0.241508	0.138793	0.372550	
LGBM	0.893387	0.273838	0.158087	0.397601	
KNeighbors	0.389352	0.655365	0.385903	0.621211	