EDA: DATA SCIENCE SALARIES

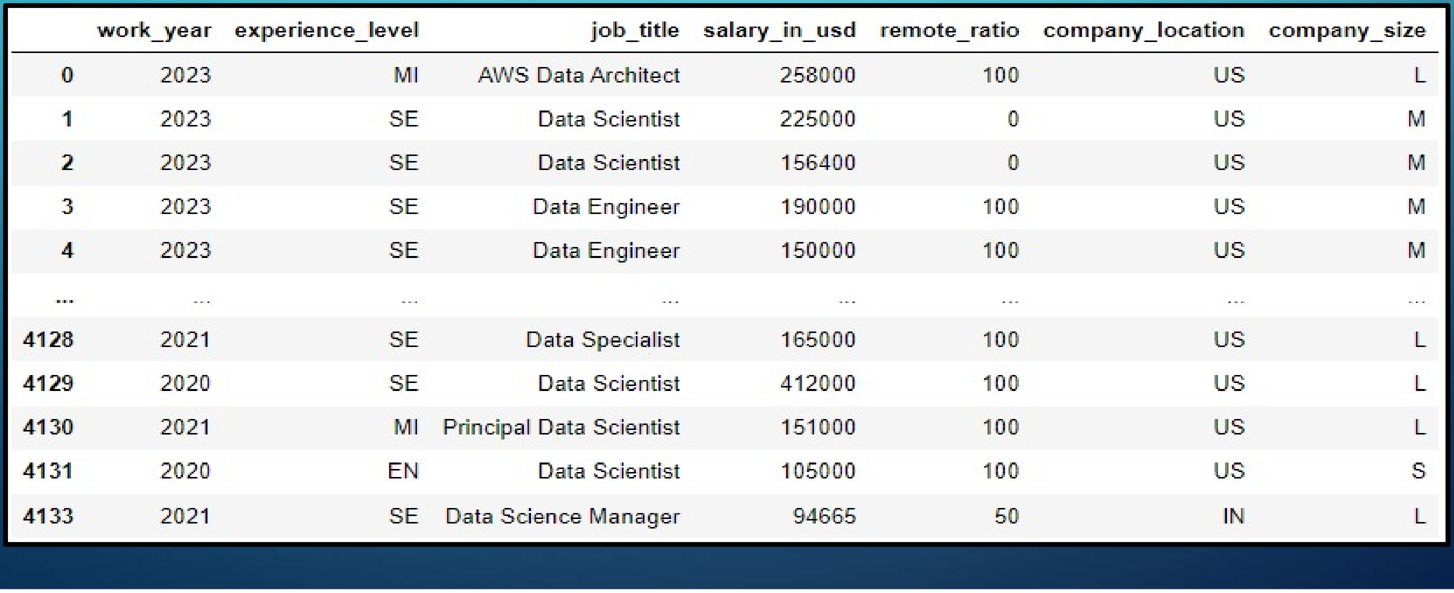
PATRICK REDINGTON

DATA SCIENCE SALARIES EDA/REGRESSION GOALS

- Overall goal is to discover what variables affect the salary of a position.
- Including how experience level, remote work ratio, company size, and the company's location affect the positions salary.
- Using these variables and others can a regression be built that can accurately predict salaries.

DATA SET – DATA SCIENCE SALARIES

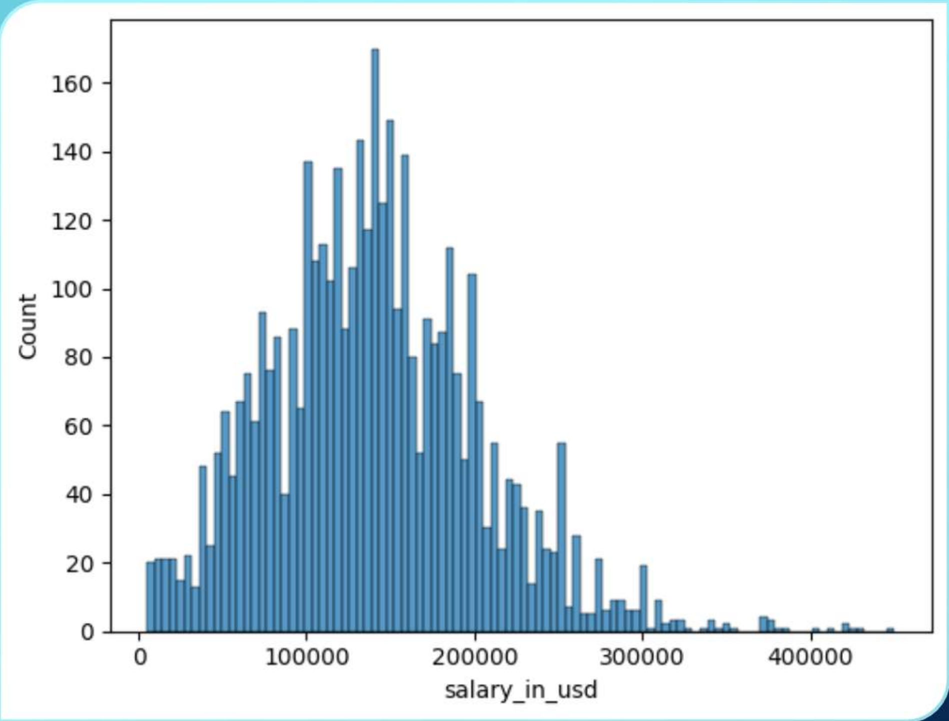
- The data set was obtained from Kaggle and is available for public domain use.
- Data was acquired from worldwide contributors and has been updated weekly since 2020.
- Summary of the data set is below.

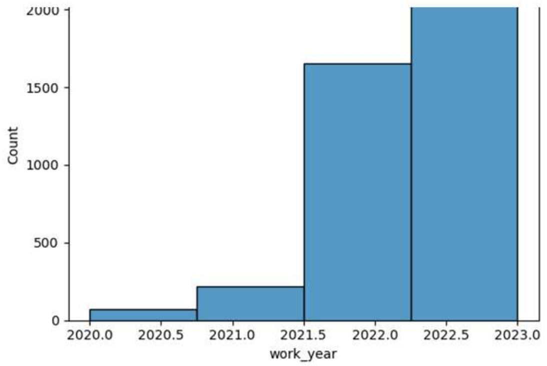


	work_year	experience_level	job_title	salary_in_usd	remote_ratio	company_location	company_size
0	2023	MI	AWS Data Architect	258000	100	US	L
1	2023	SE	Data Scientist	225000	0	US	M
2	2023	SE	Data Scientist	156400	0	US	M
3	2023	SE	Data Engineer	190000	100	US	M
4	2023	SE	Data Engineer	150000	100	US	M
...	...	...	...	...	...	...	...
4128	2021	SE	Data Specialist	165000	100	US	L
4129	2020	SE	Data Scientist	412000	100	US	L
4130	2021	MI	Principal Data Scientist	151000	100	US	L
4131	2020	EN	Data Scientist	105000	100	US	S
4133	2021	SE	Data Science Manager	94665	50	IN	L

SALARY IN USD

- States the salary reported converted to USD, if not reported as USD already.
- Statistics:
  - count   4093.000000
  - mean   140116.351332
  - std   62983.078569
  - variance 3966868186
  - min   5132.000000
  - 25%   99050.000000
  - 50%   136000.000000
  - 75%   180000.000000
  - max   450000.000000
  - Name: salary_in_usd
  - dtype: float64



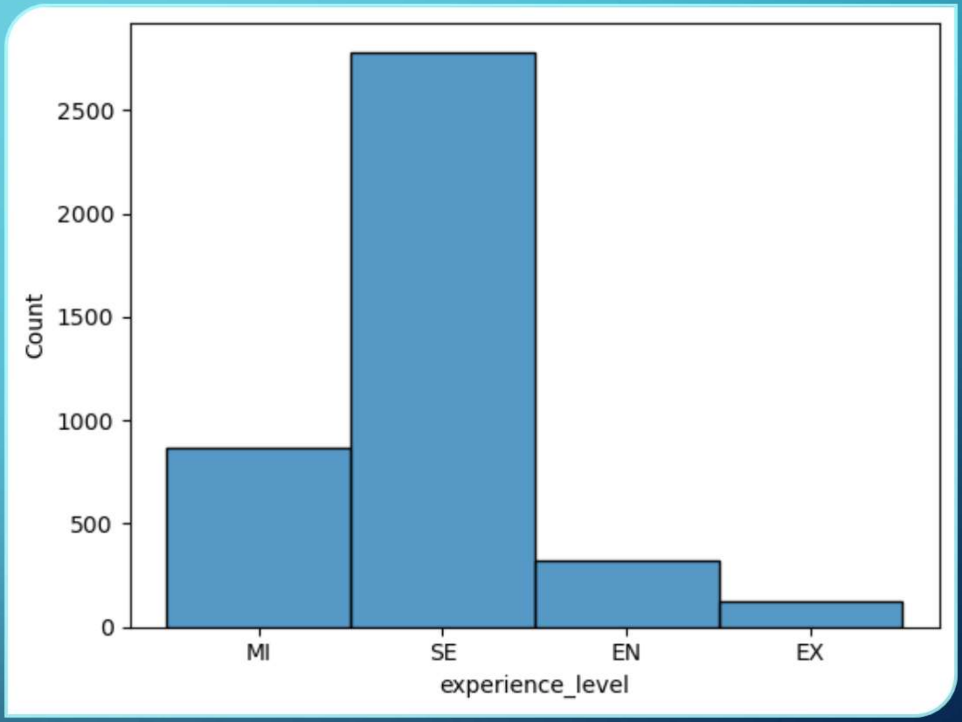


WORK YEAR

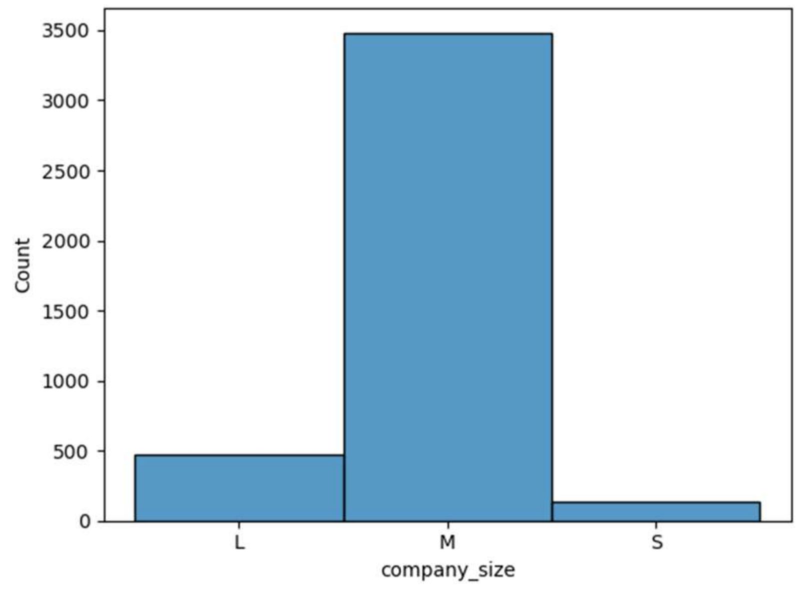
- States the year the salary was paid to the employee.
- Statistics:
  - count 4093.000000
  - mean 2022.437332
  - std 0.676584
  - variance 0.457765
  - min 2020.000000
  - 25% 2022.000000
  - 50% 2023.000000
  - 75% 2023.000000
  - max 2023.000000
  - Name: work_year
  - dtype: float64

EXPERIENCE LEVEL

- States the experience level of the position ; EN (Entry-level), MI (Mid-level), SE (Senior-level), and EX (Executive-level).
- Statistics:
    - count 4093
    - unique 4
    - variance 0.440626
    - top SE
    - freq 2784
    - Name: experience_level
    - dtype: object



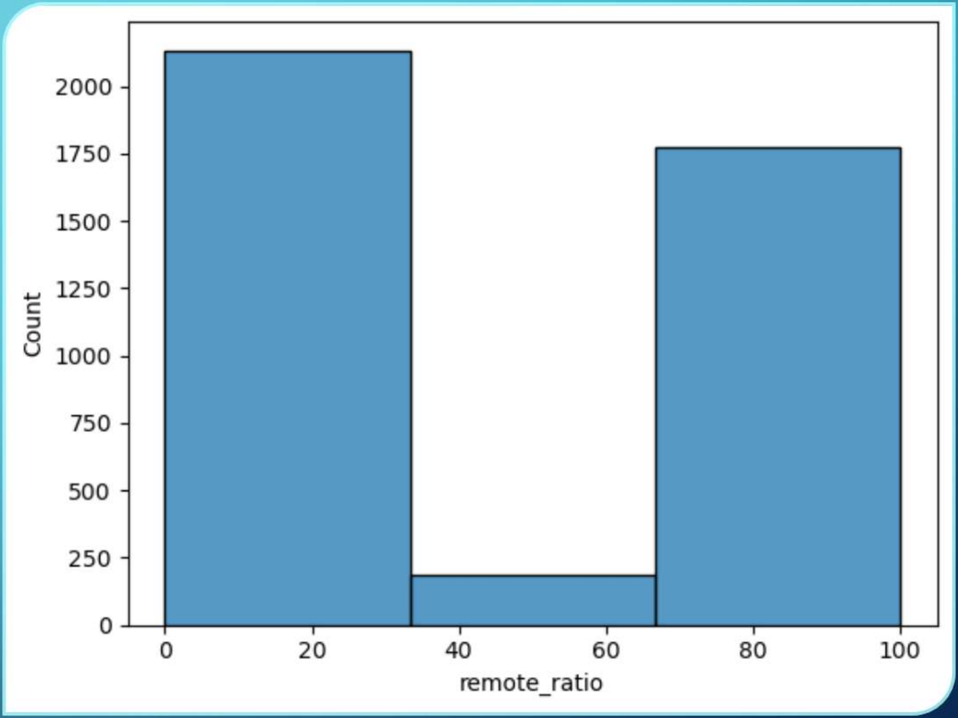
COMPANY SIZE

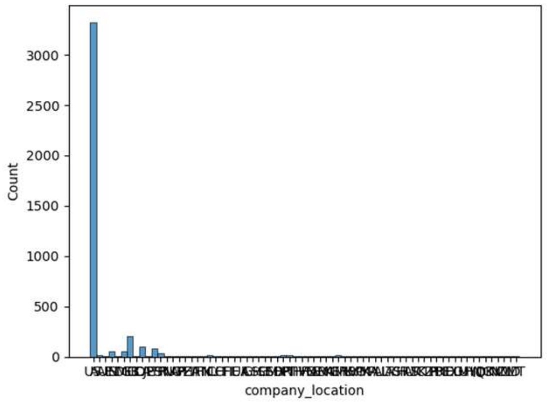


- The average number of people that worked for the company during the year: S (less than 50) M (50 to 250) L (more than 250).
- Statistics:
  - count 4093
  - unique 3
  - variance 0.14212
  - top M
  - freq 3484
  - Name: company_size
  - dtype: object

REMOTE RATIO

- States the amount of remote work; 0 (less than 20%) 50 (20%-80%) 100 (more than 80%).
- Statistics:
  - count 4093.000000
  - mean 45.614464
  - std 48.671951
  - variance 0.236895
  - min 0.000000
  - 25% 0.000000
  - 50% 0.000000
  - 75% 100.000000
  - max 100.000000
  - Name: remote_ratio
  - dtype: float64



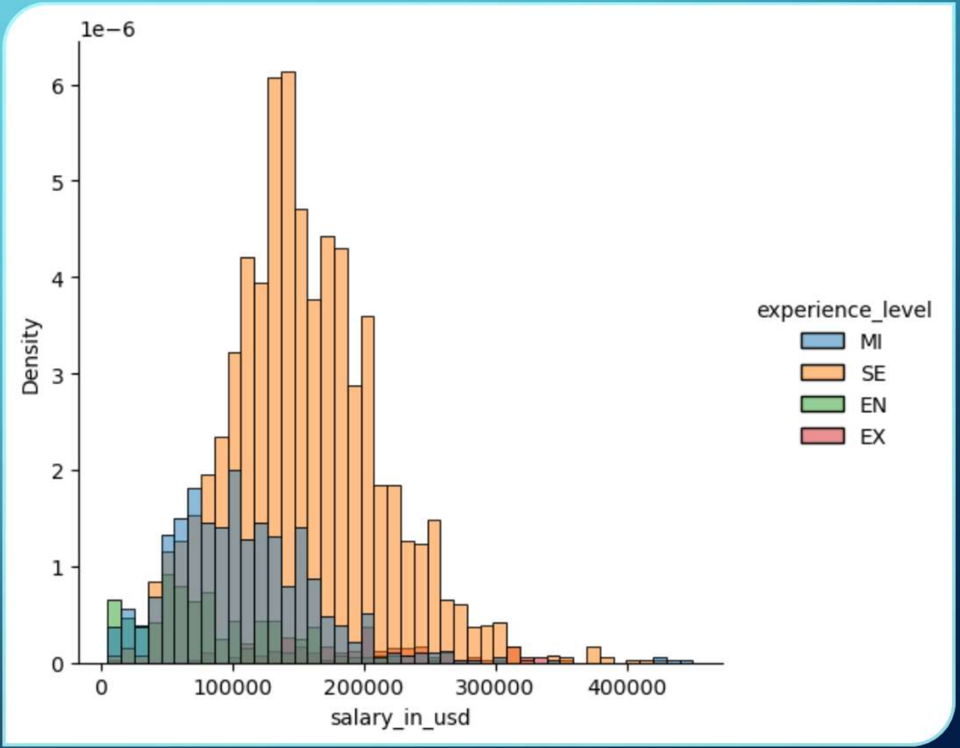


COMPANY LOCATION

- Sates the country of the employer's main office or contracting branch as an ISO 3166 country code.
- Statistics:
  - count 4093
  - unique 70
  - variance 297.09829
  - top US
  - freq 3326
  - Name: company_location
  - dtype: object

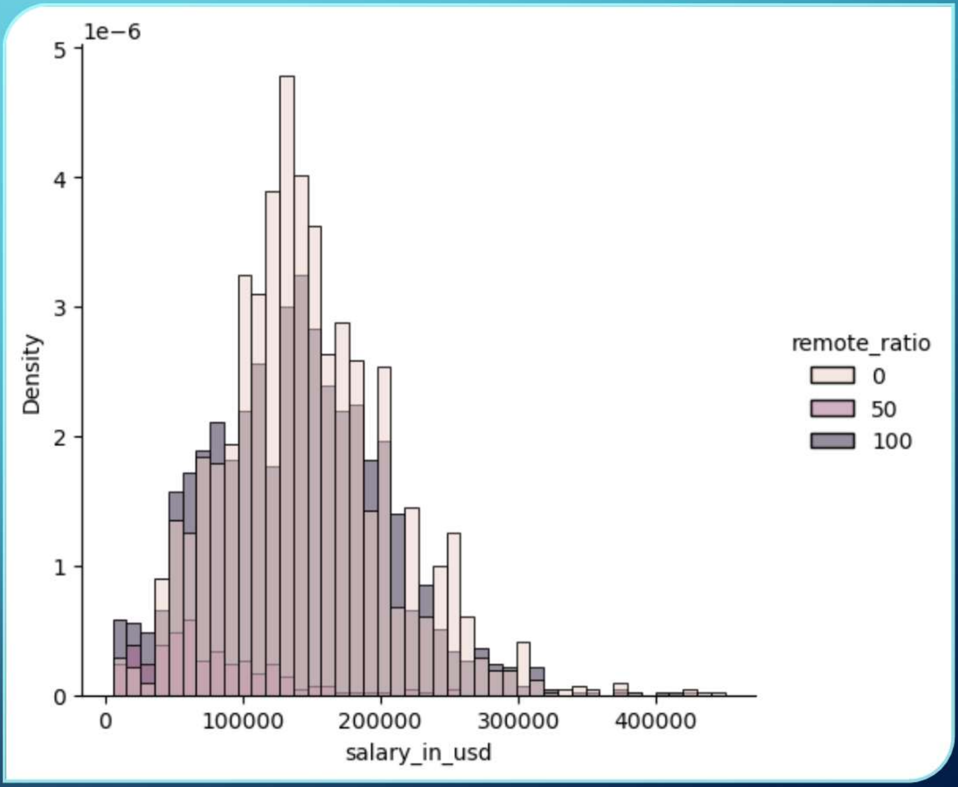
PMF –SALARY COMPARED BY EXPERIENCE LEVELS

- Majority of data points are from senior level positions.
- As expected, the higher the level of experience of the position the father salary probability is pushed to the right on the pay scale.
- Means:
  - EN - $80,192.33
  - MI - $107,652.77
  - SE - $154,698.15
  - EX - $193,833.15

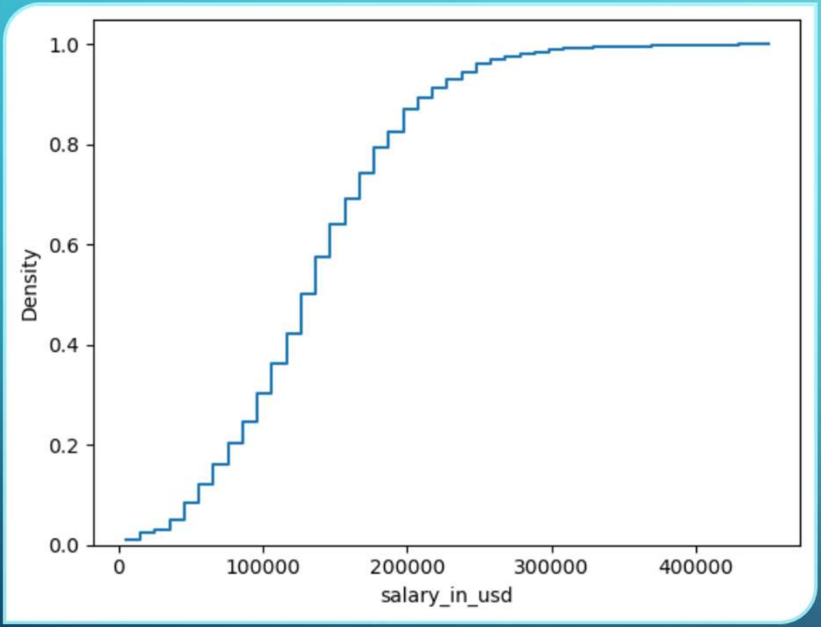


PMF- SALARIES COMPARED BY REMOTE WORK RATIO

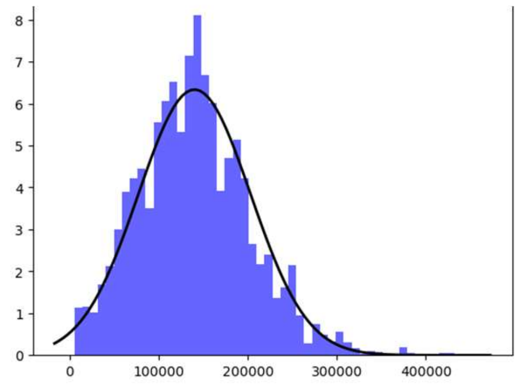
- Majority of data points come either 0% remote work or 100%.
- Hybrid work seems to be on the lower end of the pay range.
- 0% vs 100% seems to be normally distributed across the pay range.

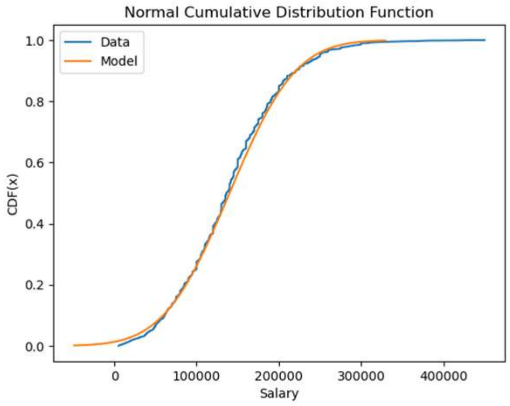


CDF- REPORTED SALARY



- Around 60% of the salaries are between 100K and 200K. With a somewhat linear distribution in this range.
- After 200K the distribution begin to level out.
- Less than 5-7% make below 50K, which may be due to bad data or third world countries.



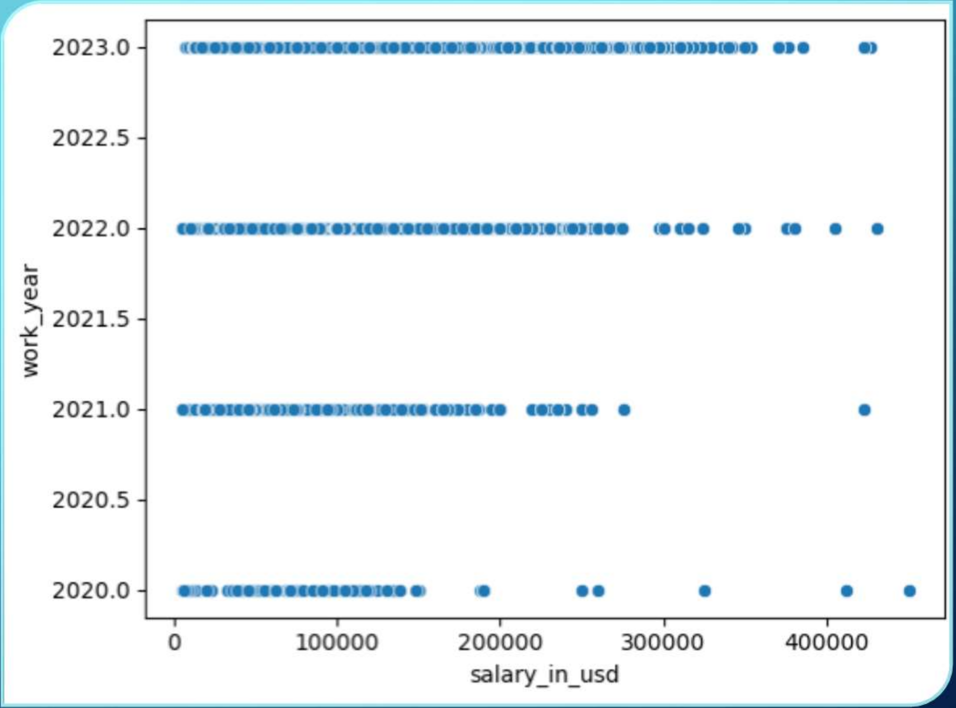


DISTRIBUTION

- The salaries are mostly normal distributed.
- As can be seen in the distribution plot and the CDF plot comparing it against the model.
- There is deviation on the lower end, as expected from the original CDF. Which may be cause by bad data or a few outliers.
- The data also increases quicker in the middle that may be an affect caused by the lower end data.

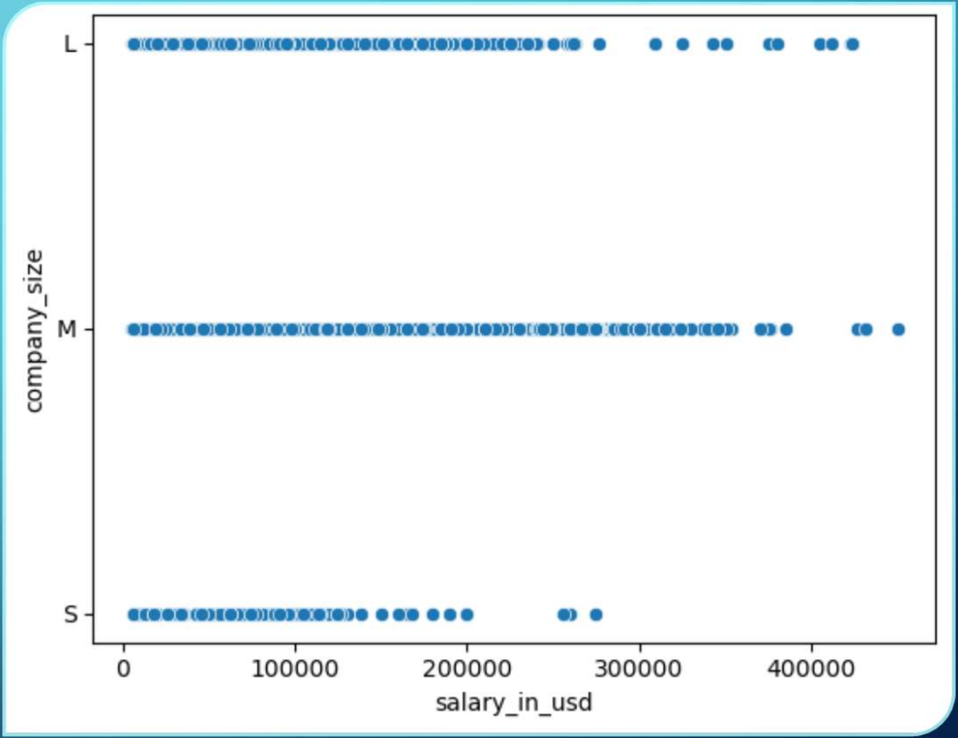
SCATTER PLOTS: SALARY VS. WORK YEAR

- The data points increase over the years, but there is a clear trend to the right as the years progress.
- There is a correlation of 0.230 as well which shows there is some correlation even if minor.
- This is expected due to the overall wage increases over the years and the increase in demand for data scientist.



SCATTER PLOTS: SALARY VS. COMPANY SIZE

- The majority of the data points are from medium and large companies.
- There initially looks to be a correlation between small and medium/large companies on pay ranges.
- The actual correlation is -0.005 though, so this just looks like there is due to the distribution of the data.

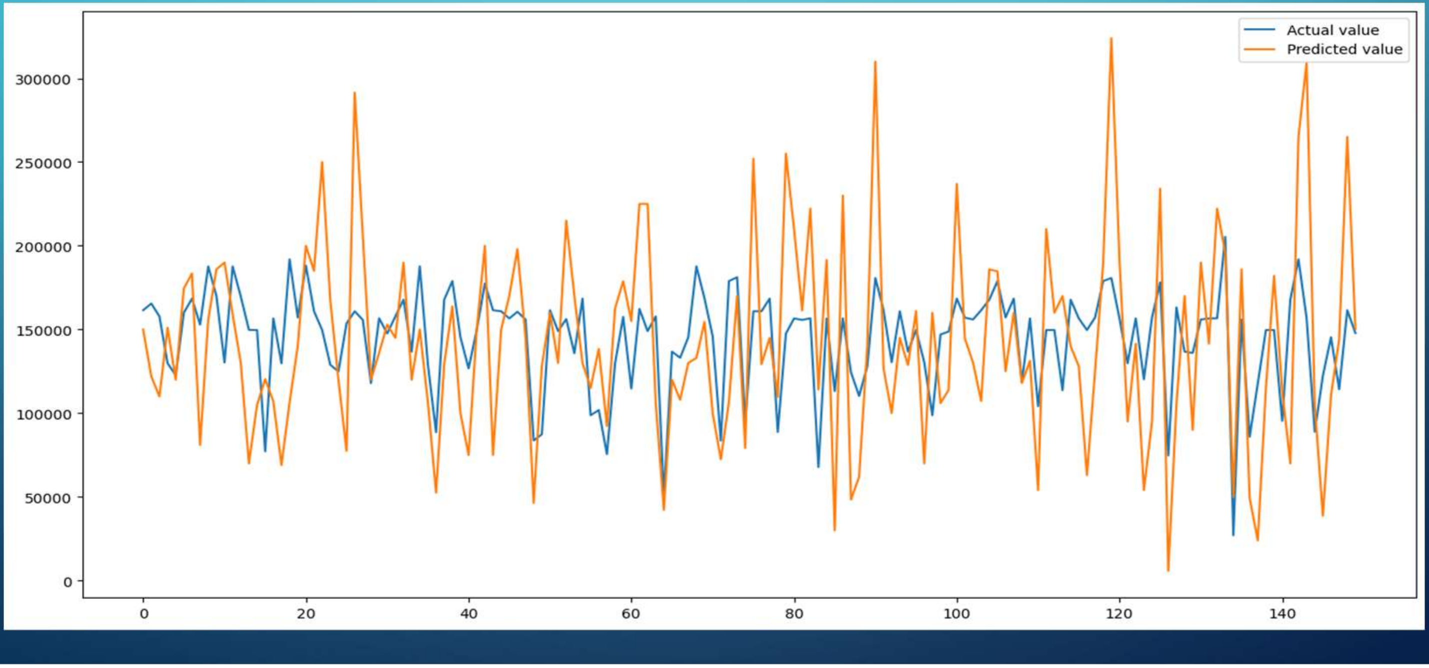


HYPOTHESIS TEST – PEARSON CORRELATION

- Hypothesis 1 – The amount of remote work affects the employee's salary.
  - After 1000 iterations the largest correlation was -0.121 and the largest P-value was 0.690.
  - Due to the low correlation and rather large P-Value the correlation coefficient is not statistically significant.
- Hypothesis 2 – The employee's experience level affects their salary.
  - After 1000 iterations the largest correlation was 0.470 and the largest P-value was 5.25e-72
  - The correlation is rather large, and the P-value is extremely low. So, the correlation coefficient is statistically significant.

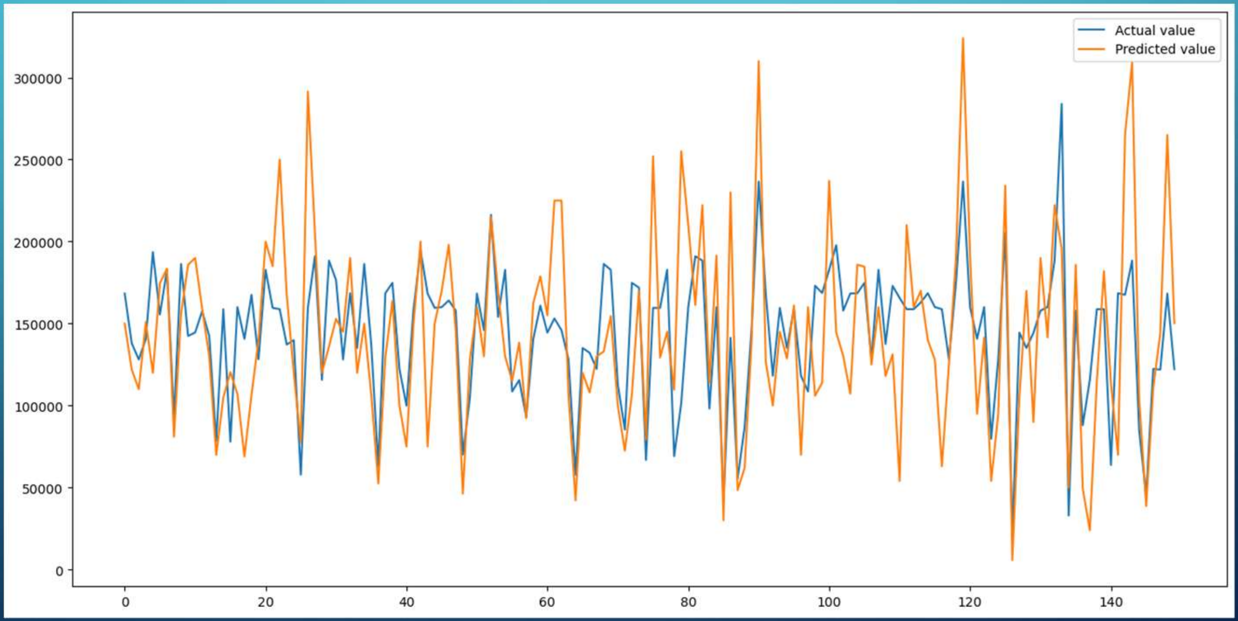
LINEAR REGRESSION MODEL

- R-Squared = 0.297
- MAE = 40962.34
- Overall model is not a good fit and only explains 29% of the variances.



RANDOM FOREST REGRESSION MODEL

- R-Squared = 0.560
- MAE = 36727.56
- Random forest model still isn’t acceptable but drastically better than the linear regression with 56% of the variances accounted for.


Actual value
Predicted value