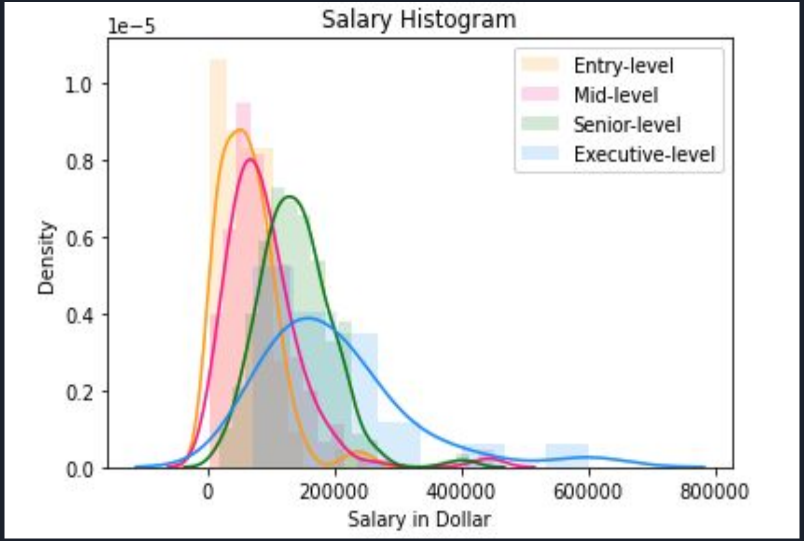
Data Science Career Analysis

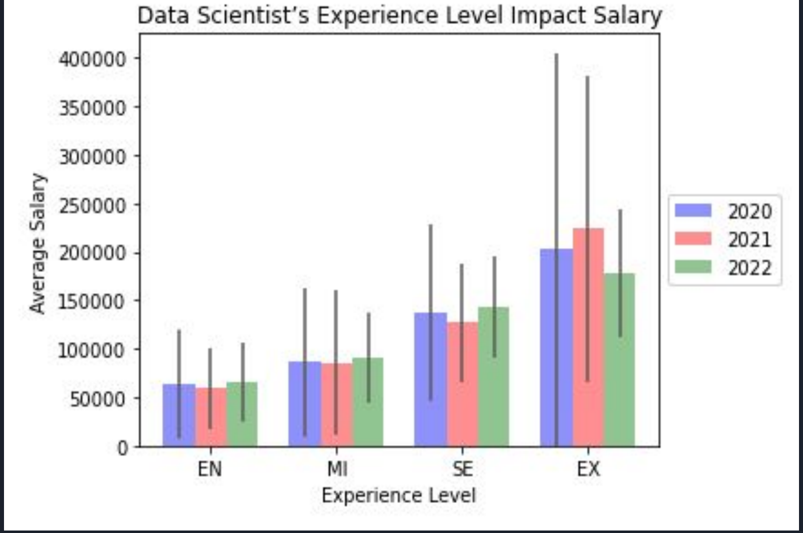
Team 3: Kyle Admire, Mai Dang, Travis Frocione, Xenia Liu

Agenda

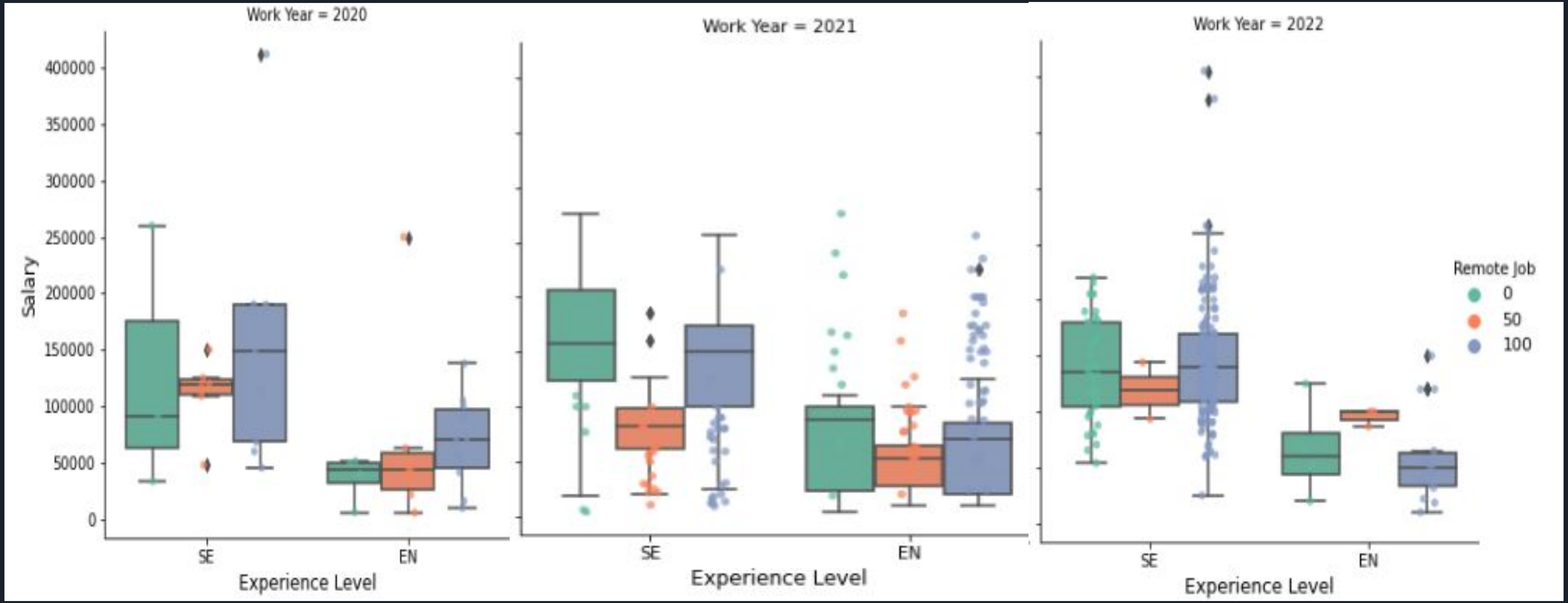
1. How does a data scientist's experience level impact their salary?
2. Is remote work more rewarding for entry level or senior level data scientists?
3. What is the most commonly used programming language for data science?
4. Which cities and states have the most demand for data scientists?
5. Which areas have the highest salaries for data scientists?
6. What is the percentage of males and females with careers in data science?
7. How does education impact the salary of a data scientist?
8. What are the highest paying positions in data science?

Q1: How does a data scientist's experience level impact their salary?

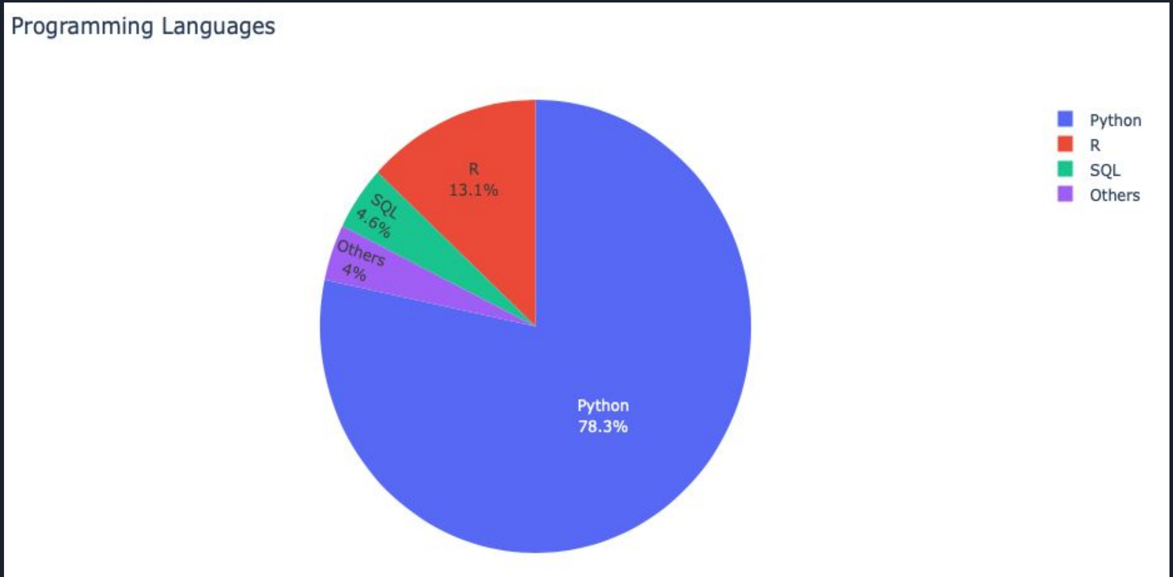


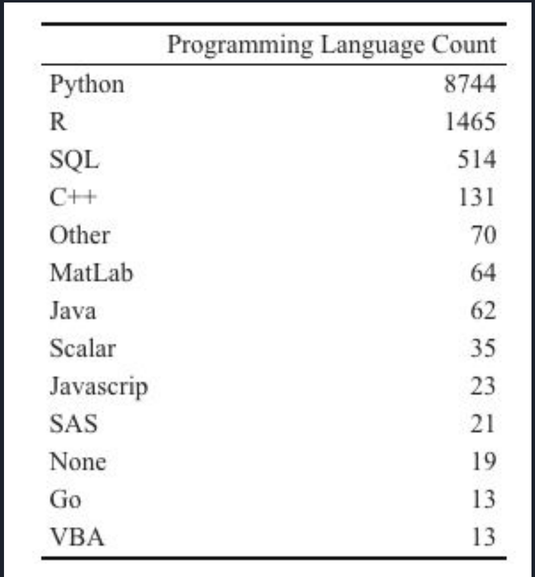


Q2: Is remote work more rewarding for entry level or senior level data scientists?



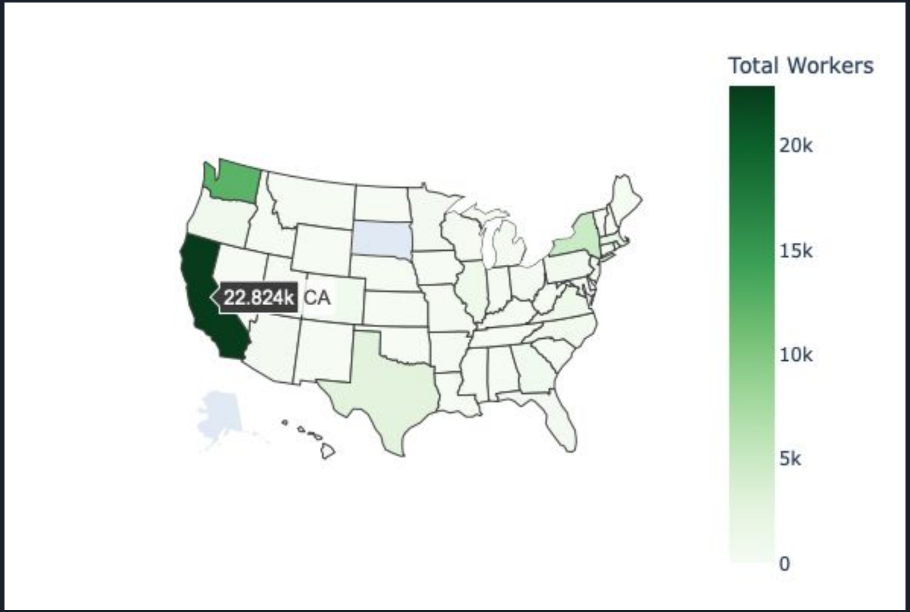
Q3: What is the most commonly used programming language for data science?





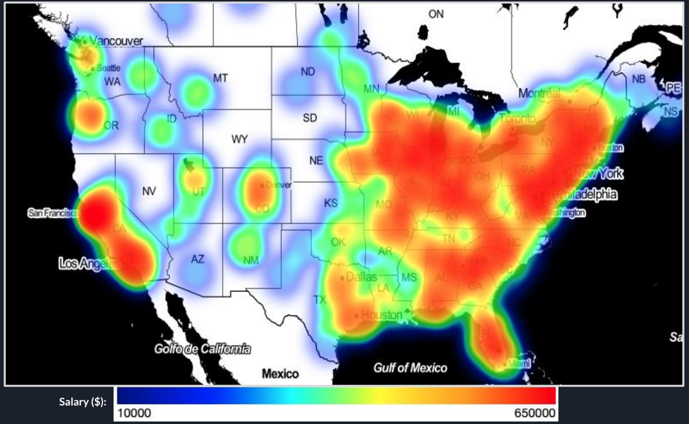
	Programming Language Count
Python	8744
R	1465
SQL	514
C++	131
Other	70
MatLab	64
Java	62
Scalar	35
Javascrip	23
SAS	21
None	19
Go	13
VBA	13

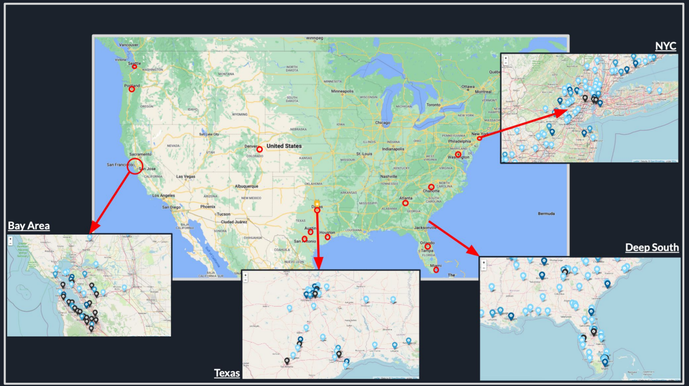
Q4: What areas have the most demand for data scientists?



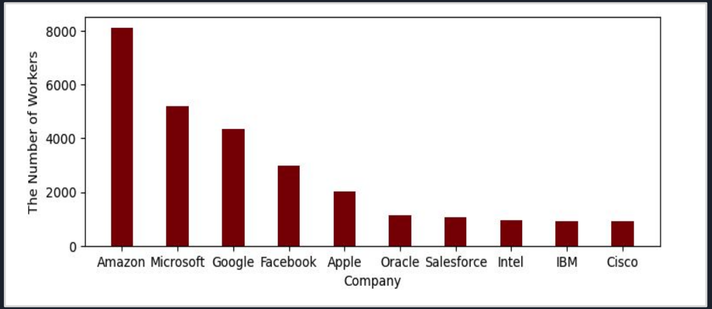
Q5: Which areas have the highest salaries for data scientists?

Top States and Cities





Brand Awareness

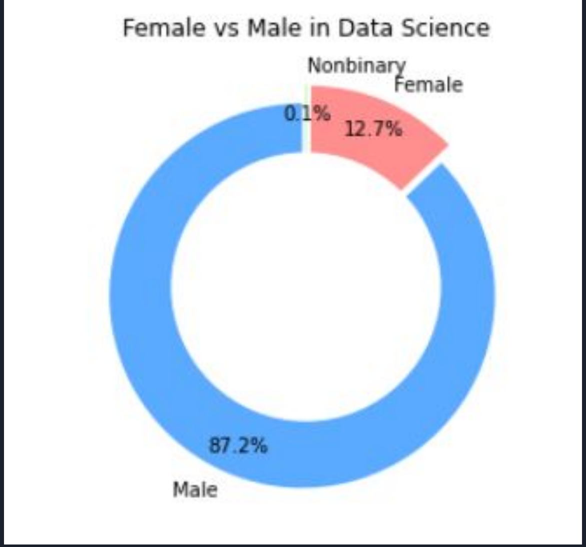


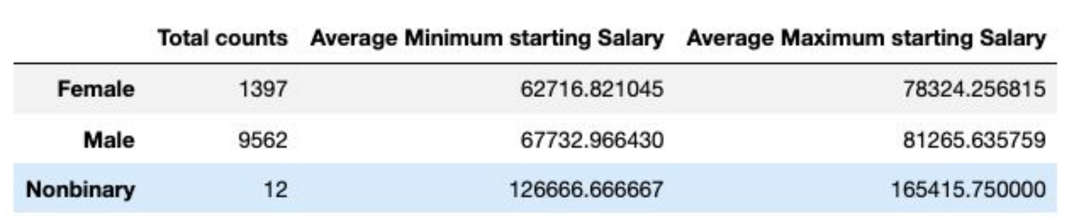
*****Pro Tip*****

Find interesting companies in areas you want to reside

Identify the centers of influence

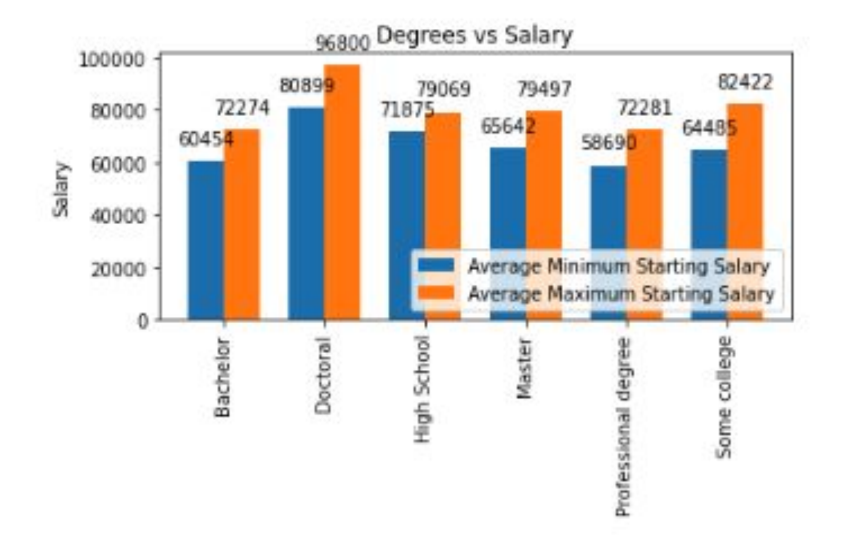
Q6: What is the percentage of males and females with careers in data science?

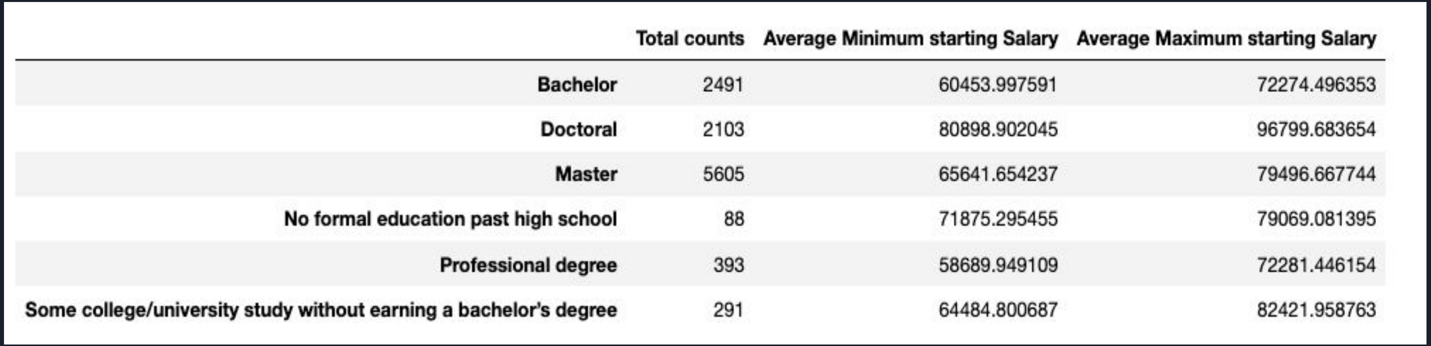




	Total counts	Average Minimum starting Salary	Average Maximum starting Salary
Female	1397	62716.821045	78324.256815
Male	9562	67732.966430	81265.635759
Nonbinary	12	126666.666667	165415.750000

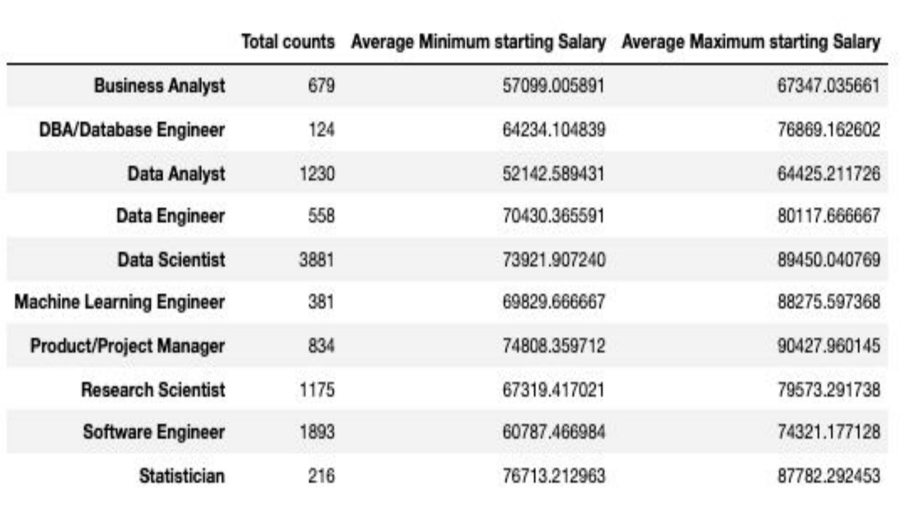
Q7: How does education impact the salary of a data scientist?



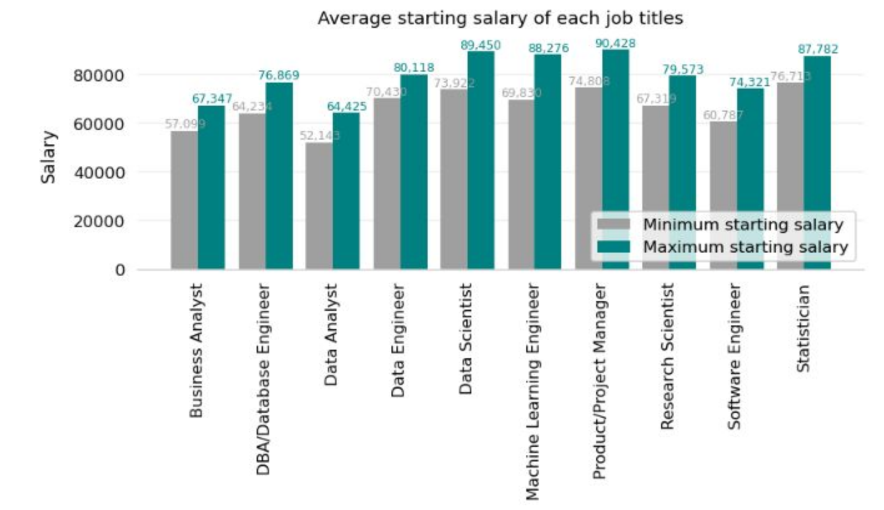


	Total counts	Average Minimum starting Salary	Average Maximum starting Salary
Bachelor	2491	60453.997591	72274.496353
Doctoral	2103	80898.902045	96799.683654
Master	5605	65641.654237	79496.667744
No formal education past high school	88	71875.295455	79069.081395
Professional degree	393	58689.949109	72281.446154
Some college/university study without earning a bachelor's degree	291	64484.800687	82421.958763

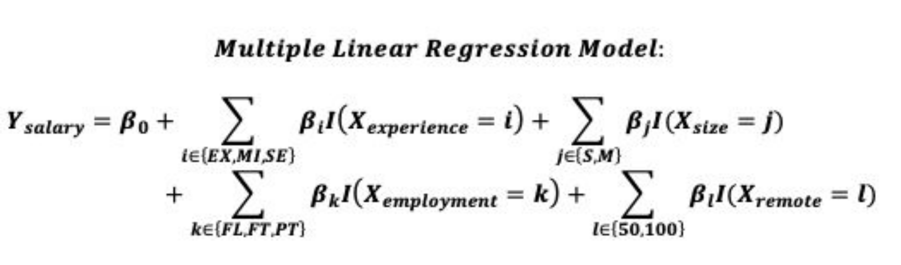
Q8: What are the highest paying positions in data science?



	Total counts	Average Minimum starting Salary	Average Maximum starting Salary
Business Analyst	679	57099.005891	67347.035661
DBA/Database Engineer	124	64234.104839	76869.162602
Data Analyst	1230	52142.589431	64425.211726
Data Engineer	558	70430.365591	80117.666667
Data Scientist	3881	73921.907240	89450.040769
Machine Learning Engineer	381	69829.666667	88275.597368
Product/Project Manager	834	74808.359712	90427.960145
Research Scientist	1175	67319.417021	79573.291738
Software Engineer	1893	60787.466984	74321.177128
Statistician	216	76713.212963	87782.292453

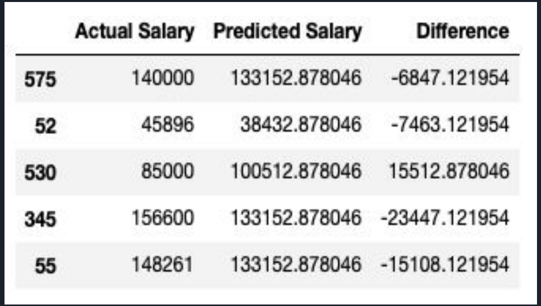


Statistical Modeling (Prediction)



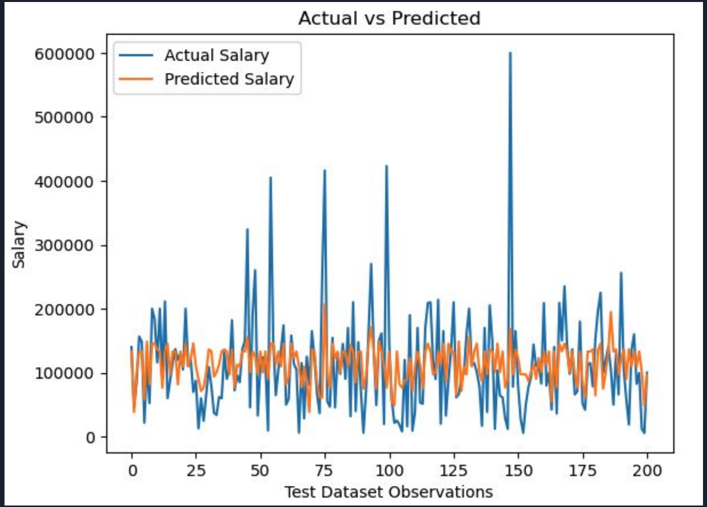
Multiple Linear Regression Model:

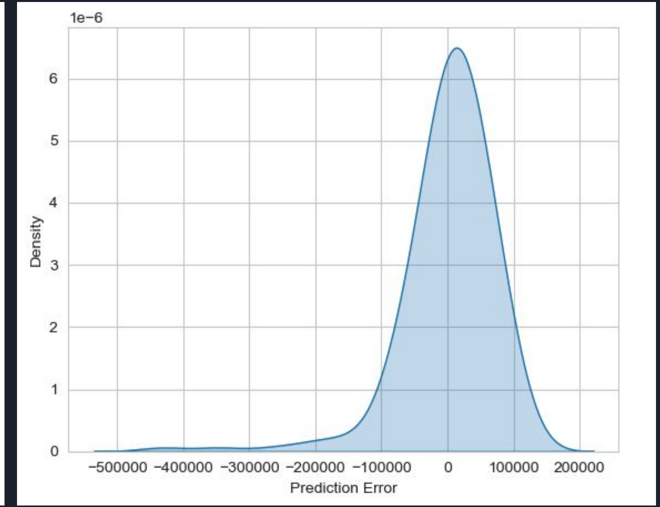
$$Y_{salary} = \beta_0 + \sum_{i \in \{EX,MI,SE\}} \beta_i I(X_{experience} = i) + \sum_{j \in \{S,M\}} \beta_j I(X_{size} = j) + \sum_{k \in \{FL,FT,PT\}} \beta_k I(X_{employment} = k) + \sum_{l \in \{50,100\}} \beta_l I(X_{remote} = l)$$



	Actual Salary	Predicted Salary	Difference
575	140000	133152.878046	-6847.121954
52	45896	38432.878046	-7463.121954
530	85000	100512.878046	15512.878046
345	156600	133152.878046	-23447.121954
55	148261	133152.878046	-15108.121954

Difference = Predicted Salary - Actual Salary

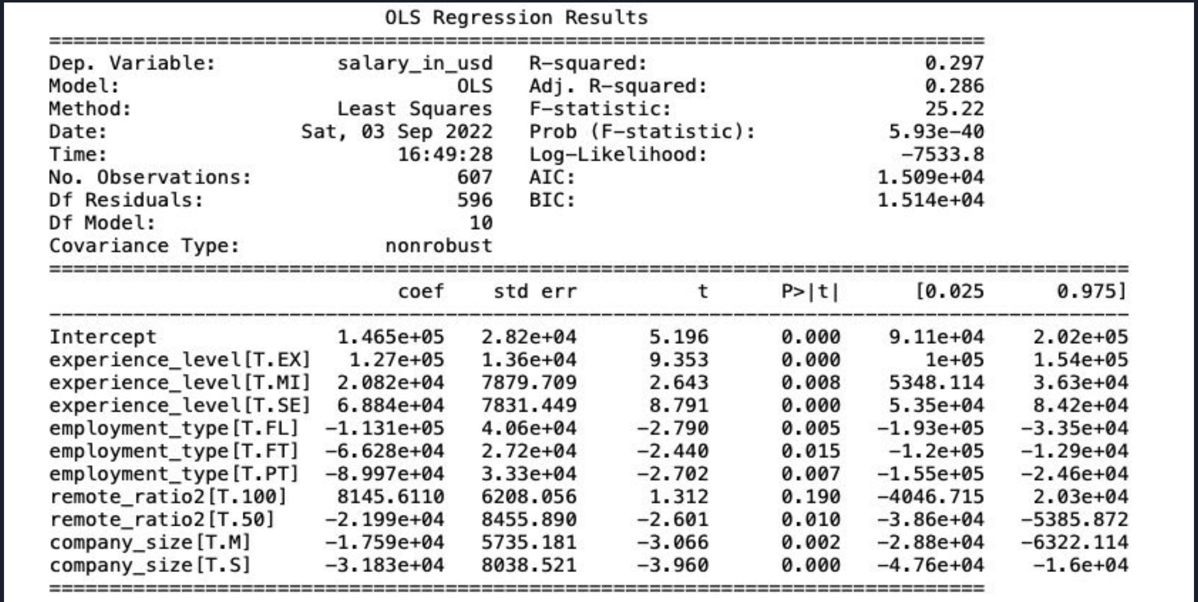




Statistical Modeling (Inference)

Ho: not significantly different from the baseline

Ha: significantly different from the baseline



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          salary_in_usd   R-squared:                       0.297
Model:                            OLS   Adj. R-squared:                  0.286
Method:                 Least Squares   F-statistic:                     25.22
Date:                Sat, 03 Sep 2022   Prob (F-statistic):           5.93e-40
Time:                        16:49:28   Log-Likelihood:                -7533.8
No. Observations:                 607   AIC:                         1.509e+04
Df Residuals:                     596   BIC:                         1.514e+04
Df Model:                          10
Covariance Type:            nonrobust
==========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                1.465e+05   2.82e+04      5.196      0.000    9.11e+04    2.02e+05
experience_level[T.EX]    1.27e+05   1.36e+04      9.353      0.000       1e+05    1.54e+05
experience_level[T.MI]   2.082e+04   7879.709      2.643      0.008    5348.114    3.63e+04
experience_level[T.SE]   6.884e+04   7831.449      8.791      0.000    5.35e+04    8.42e+04
employment_type[T.FL]   -1.131e+05   4.06e+04     -2.790      0.005   -1.93e+05   -3.35e+04
employment_type[T.FT]   -6.628e+04   2.72e+04     -2.440      0.015    -1.2e+05   -1.29e+04
employment_type[T.PT]   -8.997e+04   3.33e+04     -2.702      0.007   -1.55e+05   -2.46e+04
remote_ratio2[T.100]     8145.6110   6208.056      1.312      0.190   -4046.715    2.03e+04
remote_ratio2[T.50]     -2.199e+04   8455.890     -2.601      0.010   -3.86e+04   -5385.872
company_size[T.M]       -1.759e+04   5735.181     -3.066      0.002   -2.88e+04   -6322.114
company_size[T.S]       -3.183e+04   8038.521     -3.960      0.000   -4.76e+04     -1.6e+04
==============================================================================
```

Baseline:

Experience Level – Entry
Employment Type – Contract
Remote Ratio – 0 %
Company Size – Large

Interpretation:

If p-value < 0.05, we reject the Ho, which means it is significantly different from the baseline. Check the coefficient to see if the salary is bigger or smaller than the baseline

Otherwise, no significant difference.

Summary

- There is currently a high demand for data scientists, which is creating many job opportunities across the United States.
- Remote and hybrid work environments are becoming more common for data scientists.
- As data scientists gain more experience and expertise, there is a high ceiling for potential earnings.

Thank you!