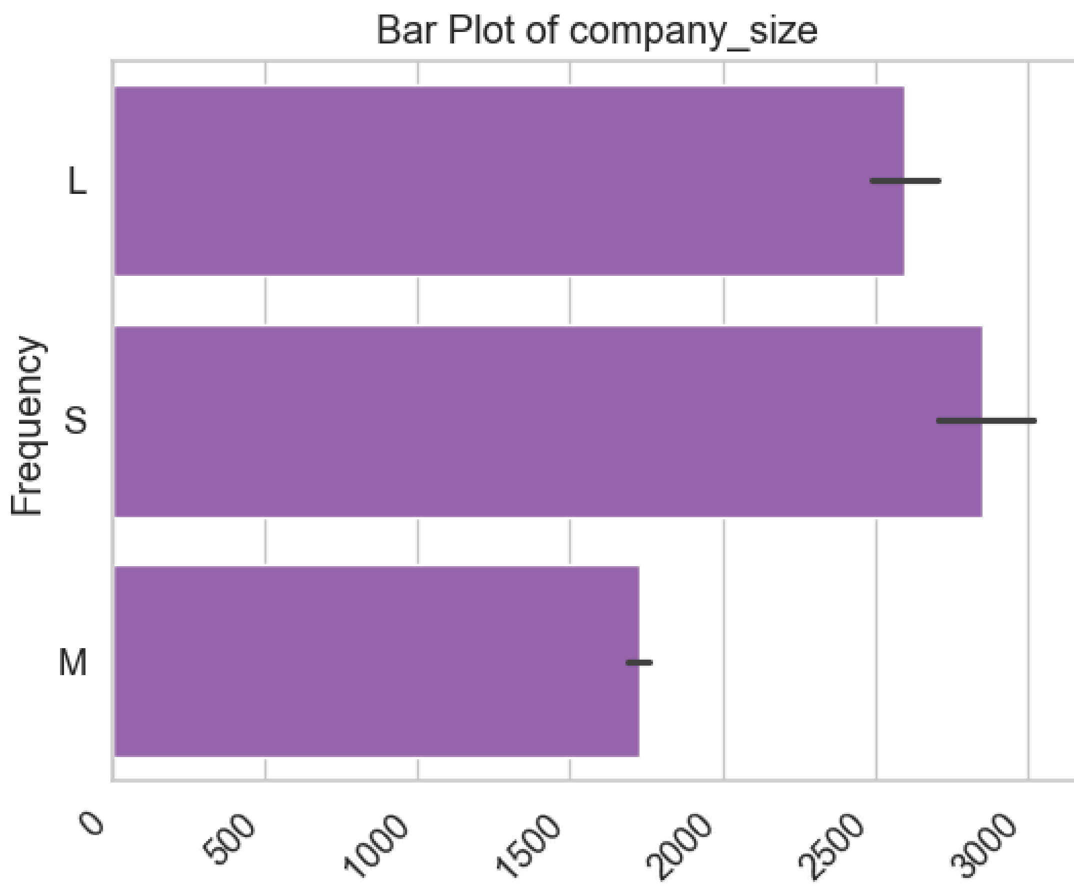
Team Name: Algo Warriors
Members: Rahul Basak, Aden Zhao, Arnav Kaul
PM: Samara Silverman
Oakley Reading Insights: Set boundaries early, be firm in our expectations, and do not enable dysfunctional behavior
PM Meeting insights: We had not started the project upon meeting her, but it was nice to formally introduce ourselves.
Summary Statistics for Dataset:


```
{'work_year': {'mean': np.float64(2022.3736351531293), 'median': np.float64(2022.0), 'min': np.int64(2020), 'max': np.int64(2023), 'std': np.float64(0.6914482342671989)}, '
experience_level': {'unique_values': ['SE', 'MI', 'EN', 'EX'], 'value_counts': {'SE': 2516, 'MI': 805, 'EN': 320, 'EX': 114}}, 'employment_type': {'unique_values': ['FT', '
CT', 'FL', 'PT'], 'value_counts': {'FT': 3718, 'PT': 17, 'CT': 10, 'FL': 10}}, 'job_title': {'unique_values': ['Principal Data Scientist', 'ML Engineer', 'Data Scientist',
'Applied Scientist', 'Data Analyst', 'Data Modeler', 'Research Engineer', 'Analytics Engineer', 'Business Intelligence Engineer', 'Machine Learning Engineer', 'Data Strateg
ist', 'Data Engineer', 'Computer Vision Engineer', 'Data Quality Analyst', 'Compliance Data Analyst', 'Data Architect', 'Applied Machine Learning Engineer', 'AI Developer',
 'Research Scientist', 'Data Analytics Manager', 'Business Data Analyst', 'Applied Data Scientist', 'Staff Data Analyst', 'ETL Engineer', 'Data DevOps Engineer', 'Head of D
ata', 'Data Science Manager', 'Data Manager', 'Machine Learning Researcher', 'Big Data Engineer', 'Data Specialist', 'Lead Data Analyst', 'BI Data Engineer', 'Director of D
ata Science', 'Machine Learning Scientist', 'MLOps Engineer', 'AI Scientist', 'Autonomous Vehicle Technician', 'Applied Machine Learning Scientist', 'Lead Data Scientist',
'Cloud Database Engineer', 'Financial Data Analyst', 'Data Infrastructure Engineer', 'Software Data Engineer', 'AI Programmer', 'Data Operations Engineer', 'BI Developer',
'Data Science Lead', 'Deep Learning Researcher', 'BI Analyst', 'Data Science Consultant', 'Data Analytics Specialist', 'Machine Learning Infrastructure Engineer', 'BI Data
Analyst', 'Head of Data Science', 'Insight Analyst', 'Deep Learning Engineer', 'Machine Learning Software Engineer', 'Big Data Architect', 'Product Data Analyst', 'Computer
 Vision Software Engineer', 'Azure Data Engineer', 'Marketing Data Engineer', 'Data Analytics Lead', 'Data Lead', 'Data Science Engineer', 'Machine Learning Research Engine
er', 'NLP Engineer', 'Manager Data Management', 'Machine Learning Developer', '3D Computer Vision Researcher', 'Principal Machine Learning Engineer', 'Data Analytics Engine
er', 'Data Analytics Consultant', 'Data Management Specialist', 'Data Science Tech Lead', 'Data Scientist Lead', 'Cloud Data Engineer', 'Data Operations Analyst', 'Marketin
g Data Analyst', 'Power BI Developer', 'Product Data Scientist', 'Principal Data Architect', 'Machine Learning Manager', 'Lead Machine Learning Engineer', 'ETL Developer',
'Cloud Data Architect', 'Lead Data Engineer', 'Head of Machine Learning', 'Principal Data Analyst', 'Principal Data Engineer', 'Staff Data Scientist', 'Finance Data Analyst
'], 'value_counts': {'Data Engineer': 1040, 'Data Scientist': 840, 'Data Analyst': 612, 'Machine Learning Engineer': 289, 'Analytics Engineer': 103, 'Data Architect': 101,
'Research Scientist': 82, 'Data Science Manager': 58, 'Applied Scientist': 58, 'Research Engineer': 37, 'ML Engineer': 34, 'Data Manager': 29, 'Machine Learning Scientist':
 26, 'Data Science Consultant': 24, 'Data Analytics Manager': 22, 'Computer Vision Engineer': 18, 'AI Scientist': 16, 'BI Data Analyst': 15, 'Business Data Analyst': 15, 'D
ata Specialist': 14, 'BI Developer': 13, 'Applied Machine Learning Scientist': 12, 'Machine Learning Infrastructure Engineer': 11, 'Big Data Engineer': 11, 'Director of Dat
a Science': 11, 'AI Developer': 11, 'Applied Data Scientist': 10, 'Head of Data': 10, 'Machine Learning Software Engineer': 10, 'Data Operations Engineer': 10, 'ETL Develop
er': 10, 'BI Analyst': 9, 'Head of Data Science': 9, 'Lead Data Scientist': 9, 'Data Science Lead': 8, 'Principal Data Scientist': 8, 'Data Quality Analyst': 7, 'NLP Engine
er': 7, 'Machine Learning Developer': 7, 'Data Infrastructure Engineer': 6, 'Lead Data Engineer': 6, 'Machine Learning Researcher': 6, 'Deep Learning Engineer': 6, 'Data An
alytics Engineer': 6, 'Lead Data Analyst': 5, 'Cloud Database Engineer': 5, 'Computer Vision Software Engineer': 5, 'Product Data Analyst': 5, 'Data Science Engineer': 5, '
MLOps Engineer': 4, '3D Computer Vision Researcher': 4, 'Business Intelligence Engineer': 4, 'Data Operations Analyst': 4, 'Machine Learning Research Engineer': 4, 'Cloud D
ata Engineer': 3, 'Machine Learning Manager': 3, 'Lead Machine Learning Engineer': 3, 'Financial Data Analyst': 3, 'Data Scientist Lead': 2, 'Data Analytics Consultant': 2,
 'Marketing Data Analyst': 2, 'Data Modeler': 2, 'Principal Data Analyst': 2, 'Principal Data Engineer': 2, 'Data Lead': 2, 'Autonomous Vehicle Technician': 2, 'Insight Ana
lyst': 2, 'ETL Engineer': 2, 'Data Analytics Lead': 2, 'Applied Machine Learning Engineer': 2, 'Data Analytics Specialist': 2, 'AI Programmer': 2, 'Data Strategist': 2, 'Bi
g Data Architect': 2, 'Software Data Engineer': 2, 'Principal Data Architect': 1, 'Head of Machine Learning': 1, 'Cloud Data Architect': 1, 'Data DevOps Engineer': 1, 'BI D
ata Engineer': 1, 'Staff Data Scientist': 1, 'Deep Learning Researcher': 1, 'Staff Data Analyst': 1, 'Product Data Scientist': 1, 'Power BI Developer': 1, 'Compliance Data
Analyst': 1, 'Data Science Tech Lead': 1, 'Data Management Specialist': 1, 'Principal Machine Learning Engineer': 1, 'Azure Data Engineer': 1, 'Manager Data Management': 1,
 'Marketing Data Engineer': 1, 'Finance Data Analyst': 1}}, 'salary': {'mean': np.float64(190695.57177097205), 'median': np.float64(138000.0), 'min': np.int64(6000), 'max':
 np.int64(30400000), 'std': np.float64(671676.5005079068)}, 'salary_currency': {'unique_values': ['EUR', 'USD', 'INR', 'HKD', 'CHF', 'GBP', 'AUD', 'SGD', 'CAD', 'ILS', 'BRL
', 'THB', 'PLN', 'HUF', 'CZK', 'DKK', 'JPY', 'MXN', 'TRY', 'CLP'], 'value_counts': {'USD': 3224, 'EUR': 236, 'GBP': 161, 'INR': 60, 'CAD': 25, 'AUD': 9, 'SGD': 6, 'BRL': 6,
 'PLN': 5, 'CHF': 4, 'HUF': 3, 'DKK': 3, 'JPY': 3, 'TRY': 3, 'THB': 2, 'ILS': 1, 'HKD': 1, 'CZK': 1, 'MXN': 1, 'CLP': 1}}, 'salary_in_usd': {'mean': np.float64(137570.38988
015978), 'median': np.float64(135000.0), 'min': np.int64(5132), 'max': np.int64(450000), 'std': np.float64(63055.6252782241)}, 'employee_residence': {'unique_values': ['ES'
, 'US', 'CA', 'DE', 'GB', 'NG', 'IN', 'HK', 'PT', 'NL', 'CH', 'CF', 'FR', 'AU', 'FI', 'UA', 'IE', 'IL', 'GH', 'AT', 'CO', 'SG', 'SE', 'SI', 'MX', 'UZ', 'BR', 'TH', 'HR', 'P
L', 'KW', 'VN', 'CY', 'AR', 'AM', 'BA', 'KE', 'GR', 'MK', 'LV', 'RO', 'PK', 'IT', 'MA', 'LT', 'BE', 'AS', 'IR', 'HU', 'SK', 'CN', 'CZ', 'CR', 'TR', 'CL', 'PR', 'DK', 'BO',
'PH', 'DO', 'EG', 'ID', 'AE', 'MY', 'JP', 'EE', 'HN', 'TN', 'RU', 'DZ', 'IQ', 'BG', 'JE', 'RS', 'NZ', 'MD', 'LU', 'MT'], 'value_counts': {'US': 3004, 'GB': 167, 'CA': 85, '
ES': 80, 'IN': 71, 'DE': 48, 'FR': 38, 'PT': 18, 'BR': 18, 'GR': 16, 'NL': 15, 'AU': 11, 'MX': 10, 'IT': 8, 'PK': 8, 'JP': 7, 'IE': 7, 'NG': 7, 'AT': 6, 'AR': 6, 'PL': 6, '
PR': 5, 'TR': 5, 'BE': 5, 'SG': 5, 'RU': 4, 'LV': 4, 'UA': 4, 'CO': 4, 'CH': 4, 'SI': 4, 'BO': 3, 'DK': 3, 'HR': 3, 'HU': 3, 'RO': 3, 'TH': 3, 'AE': 3, 'VN': 3, 'HK': 2, 'U
Z': 2, 'PH': 2, 'CF': 2, 'CL': 2, 'FI': 2, 'CZ': 2, 'SE': 2, 'AS': 2, 'LT': 2, 'GH': 2, 'KE': 2, 'DZ': 1, 'NZ': 1, 'JE': 1, 'MY': 1, 'MD': 1, 'IQ': 1, 'BG': 1, 'LU': 1, 'RS
': 1, 'HN': 1, 'EE': 1, 'TN': 1, 'CR': 1, 'ID': 1, 'EG': 1, 'DO': 1, 'CN': 1, 'SK': 1, 'IR': 1, 'MA': 1, 'IL': 1, 'MK': 1, 'BA': 1, 'AM': 1, 'CY': 1, 'KW': 1, 'MT': 1}}, 'r
emote_ratio': {'mean': np.float64(46.271637816245004), 'median': np.float64(0.0), 'min': np.int64(0), 'max': np.int64(100), 'std': np.float64(48.589050470587566)}, 'company
_location': {'unique_values': ['ES', 'US', 'CA', 'DE', 'GB', 'NG', 'IN', 'HK', 'NL', 'CH', 'CF', 'FR', 'FI', 'UA', 'IE', 'IL', 'GH', 'CO', 'SG', 'AU', 'SE', 'SI', 'MX', 'BR
', 'PT', 'RU', 'TH', 'HR', 'VN', 'EE', 'AM', 'BA', 'KE', 'GR', 'MK', 'LV', 'RO', 'PK', 'IT', 'MA', 'PL', 'AL', 'AR', 'LT', 'AS', 'CR', 'IR', 'BS', 'HU', 'AT', 'SK', 'CZ', '
TR', 'PR', 'DK', 'BO', 'PH', 'BE', 'ID', 'EG', 'AE', 'LU', 'MY', 'HN', 'JP', 'DZ', 'IQ', 'CN', 'NZ', 'CL', 'MD', 'MT'], 'value_counts': {'US': 3040, 'GB': 172, 'CA': 87, 'E
S': 77, 'IN': 58, 'DE': 56, 'FR': 34, 'BR': 15, 'AU': 14, 'GR': 14, 'PT': 14, 'NL': 13, 'MX': 10, 'IE': 7, 'SG': 6, 'AT': 6, 'JP': 6, 'TR': 5, 'CH': 5, 'NG': 5, 'PL': 5, 'P
K': 4, 'LV': 4, 'DK': 4, 'IT': 4, 'PR': 4, 'SI': 4, 'BE': 4, 'CO': 4, 'UA': 4, 'HR': 3, 'TH': 3, 'RU': 3, 'AR': 3, 'CZ': 3, 'AE': 3, 'FI': 3, 'AS': 3, 'LU': 3, 'HU': 2, 'ID
': 2, 'LT': 2, 'RO': 2, 'SE': 2, 'KE': 2, 'EE': 2, 'CF': 2, 'IL': 2, 'GH': 2, 'EG': 1, 'MD': 1, 'CL': 1, 'NZ': 1, 'CN': 1, 'IQ': 1, 'DZ': 1, 'HK': 1, 'HN': 1, 'MY': 1, 'AL'
: 1, 'MA': 1, 'PH': 1, 'BO': 1, 'VN': 1, 'AM': 1, 'BA': 1, 'SK': 1, 'MK': 1, 'BS': 1, 'IR': 1, 'CR': 1, 'MT': 1}}, 'company_size': {'unique_values': ['L', 'S', 'M'], 'value
_counts': {'M': 3153, 'L': 454, 'S': 148}}}
```

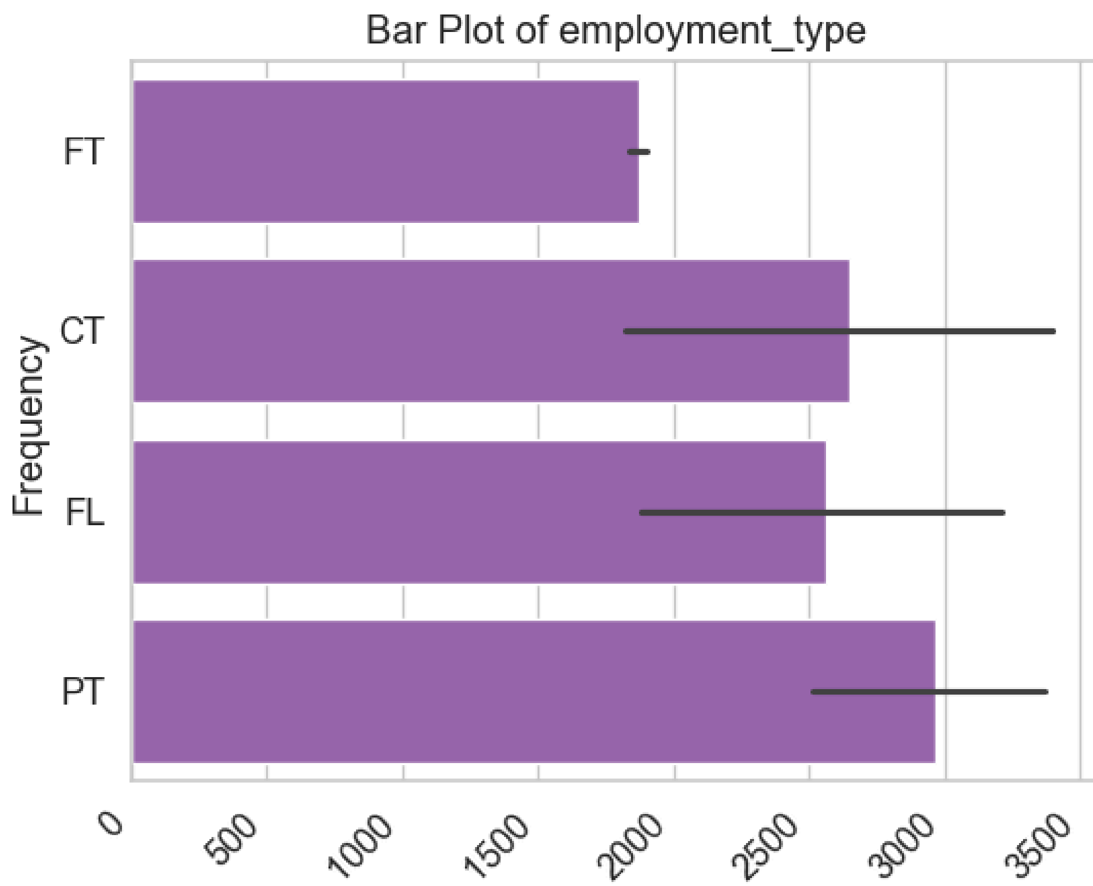

Plots and descriptions:

The plot below tells us the relative frequencies of company sizes in our data set. There are less medium companies.

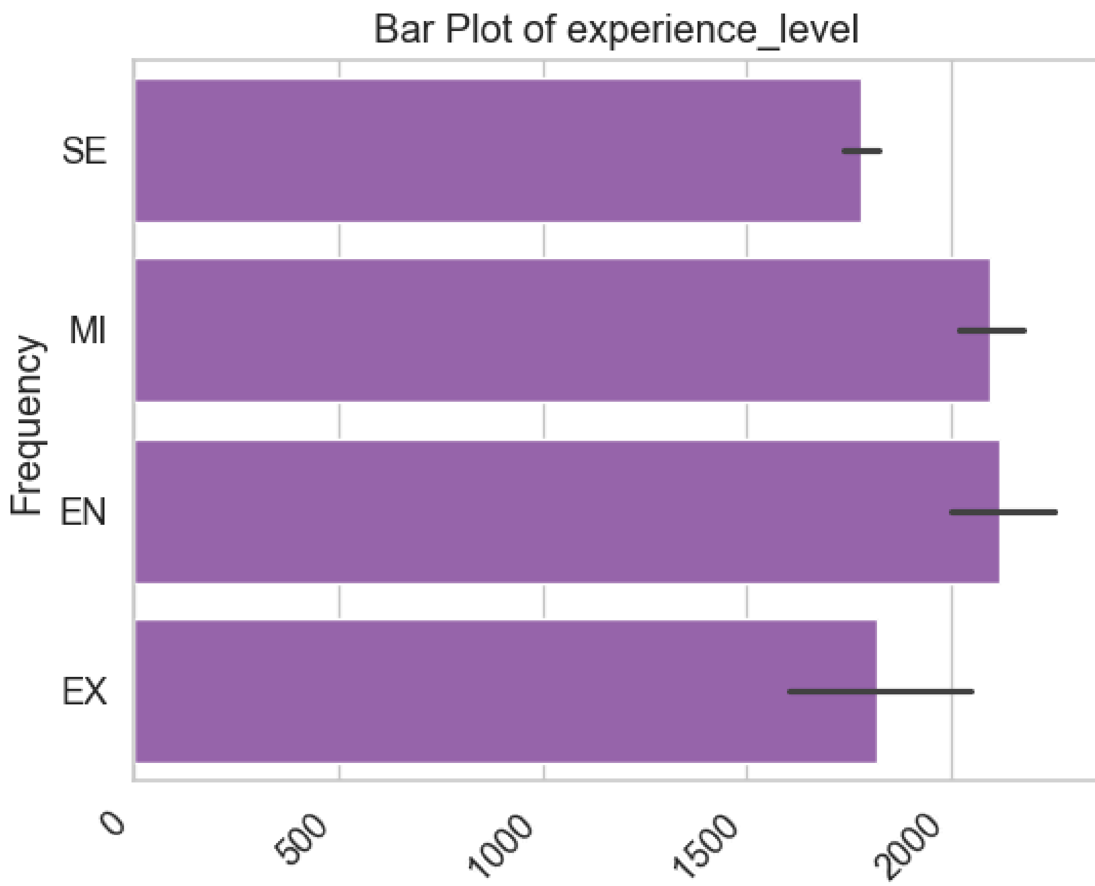


The plot below tells us the relative frequencies of employee residences. It’s not very useful since the plot is so cluttered. Maybe a different plot would be better suited for this particular column.

The plot below tells us about the relative frequencies of employment_type. There are a lot of part time workers.

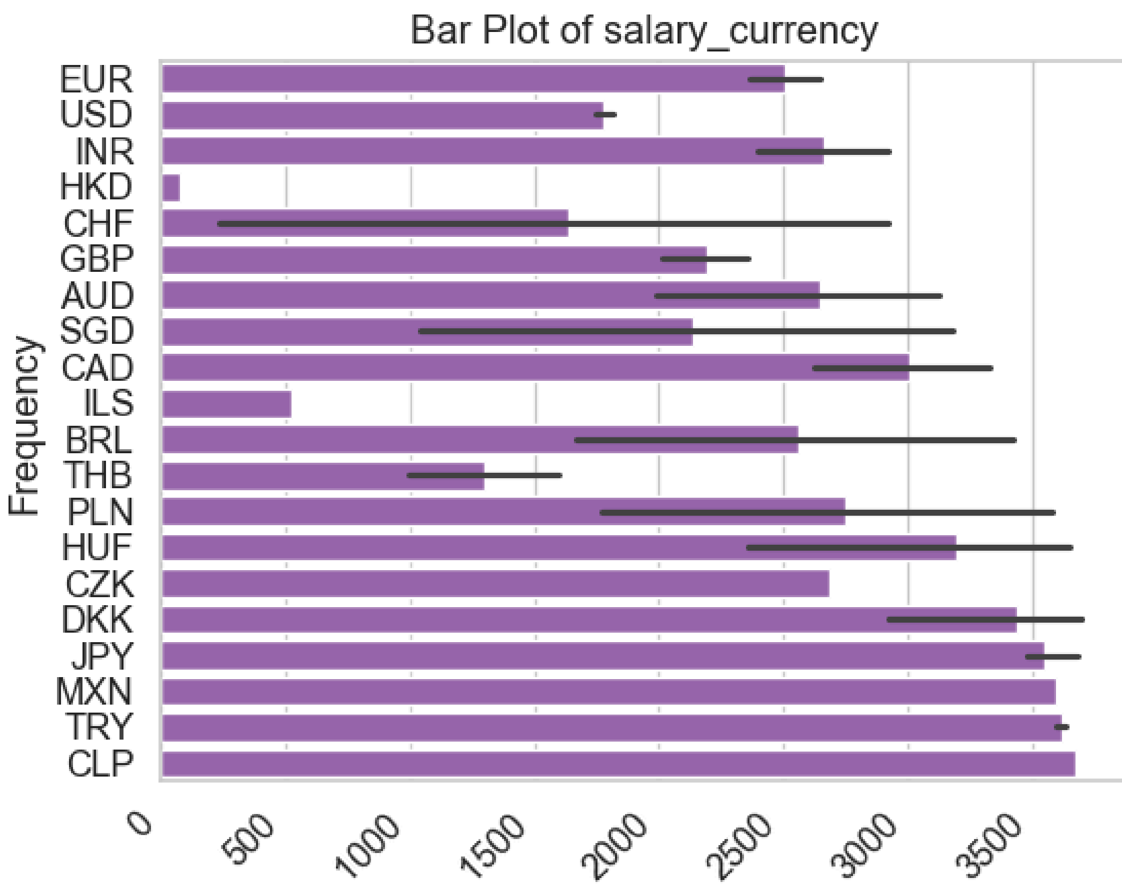


The plot below tells us about the relative frequencies of experience level. It’s pretty even.



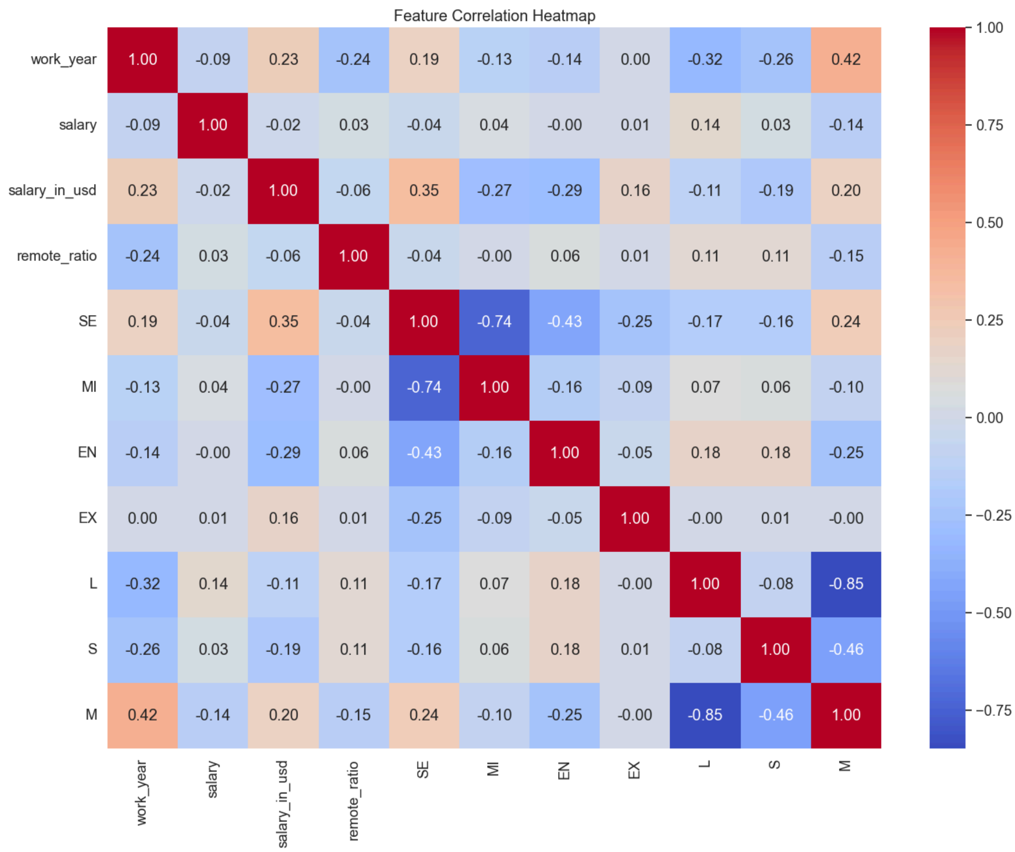
The plot below tells us about the relative frequencies of job title. Again, it's not very useful because of how cluttered it is, which tells us that a different plot would perhaps be better suited.

The plot below tells us about the relative frequencies of salary currency. Again, it's fairly even with outliers on the lower end like HKD.

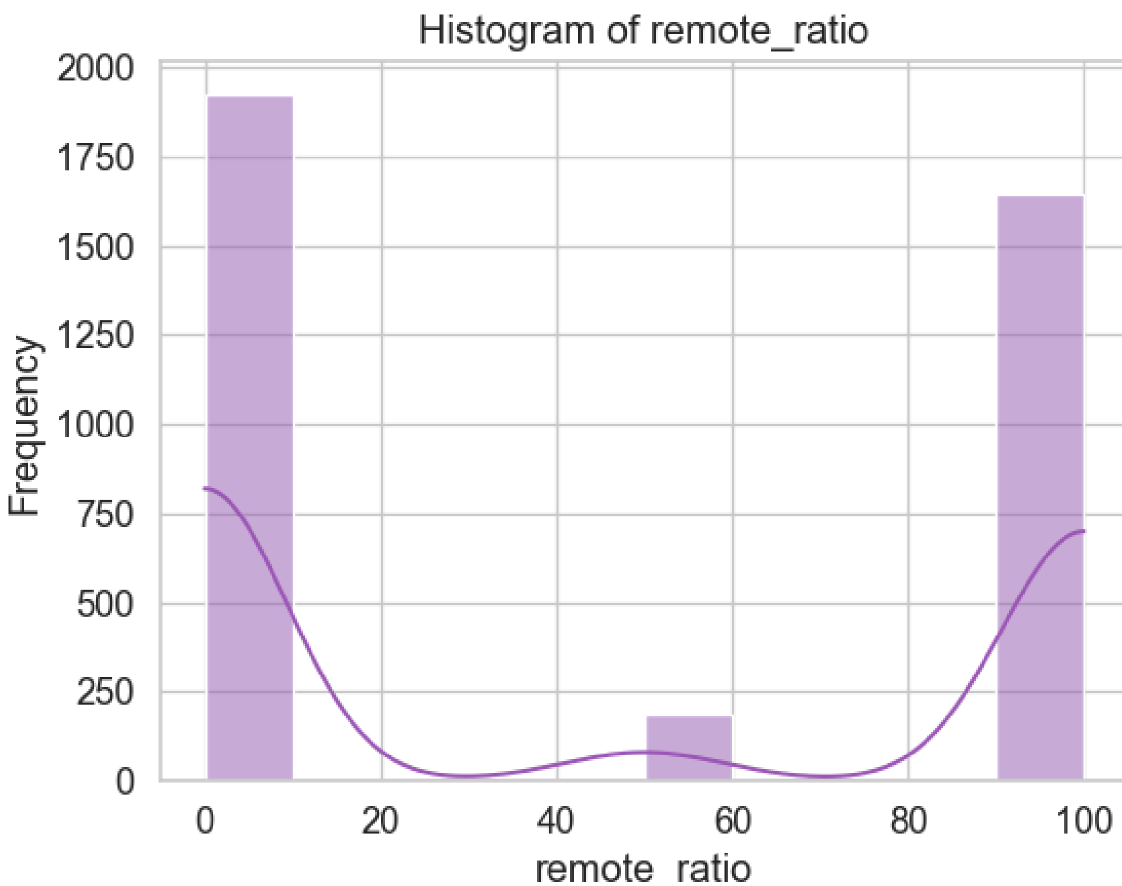


The plot below tells us about the distribution of salaries by remote work ratio. (the title is a typo). Strangely, 50% remote work jobs earn noticeably less money.

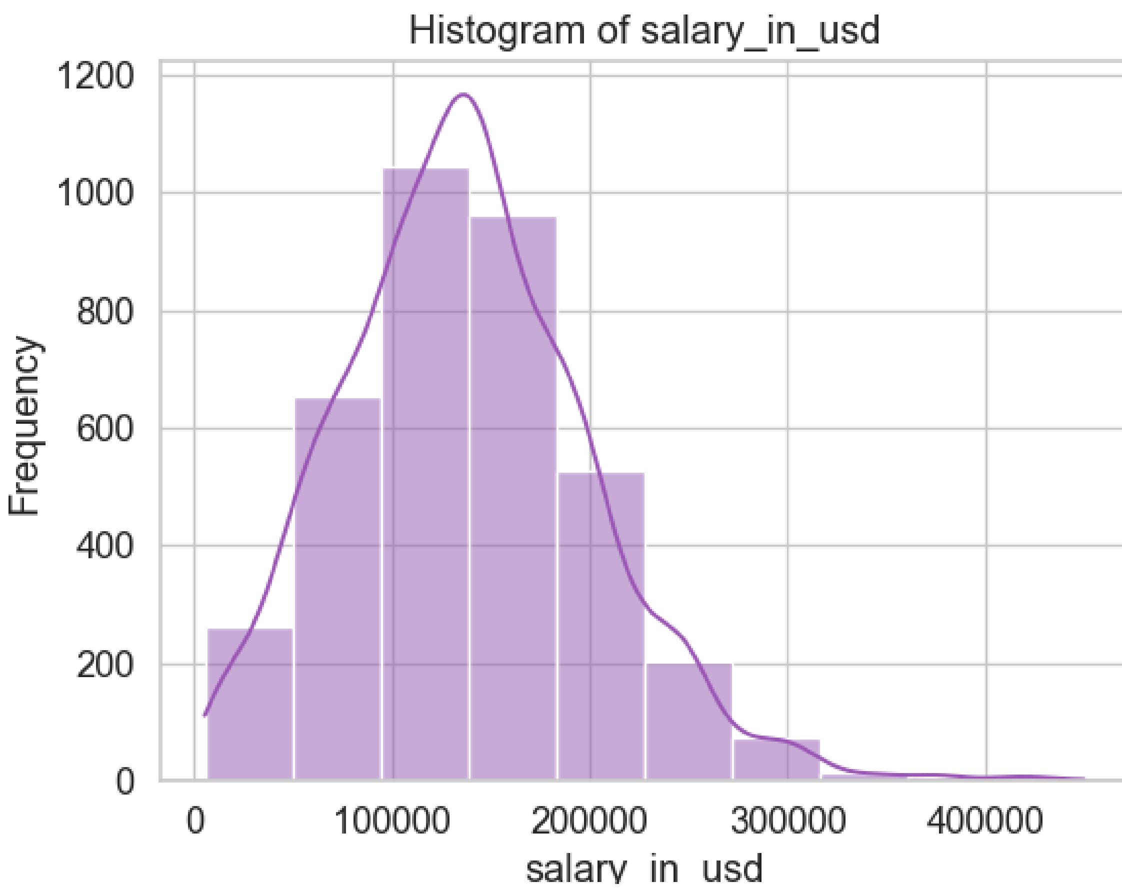
The plot below is a feature correlation heat map of the features: work year, salary, salary_in_usd, remote_ratio, job title, and company size. The plot tells us a lot of things, like remote ratio and work year are negatively correlated, suggesting that as the number of years someone has been working increases, the proportion of remote work decreases. Strange.,



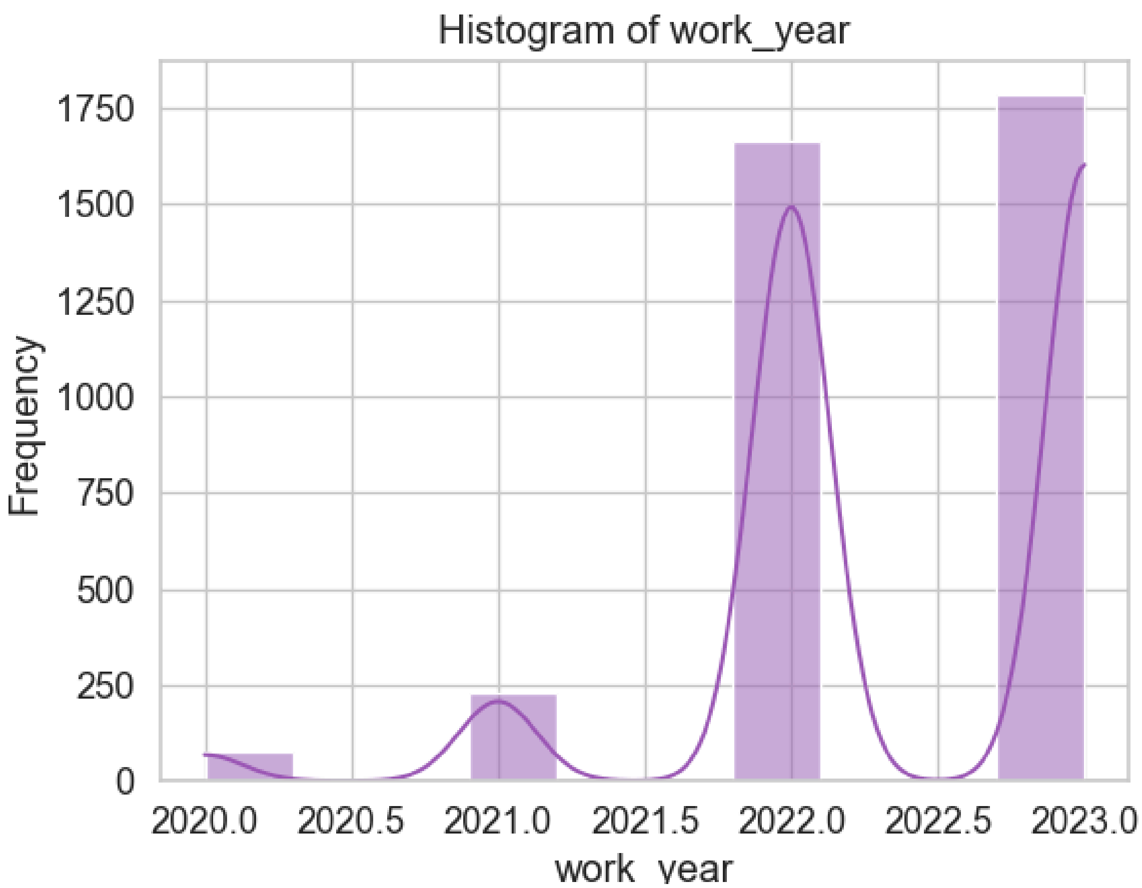
The histogram below tells us the frequencies of remote ratios. In person and remote are pretty equal.



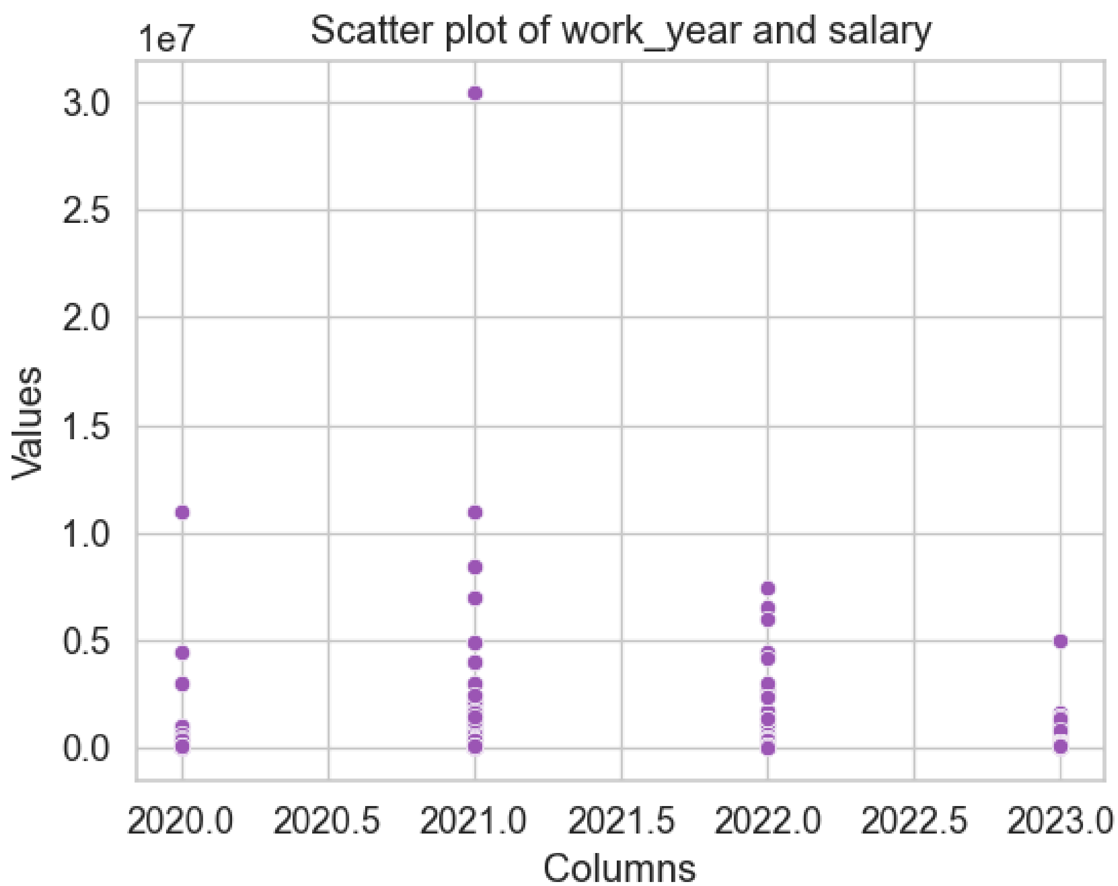
The histogram below tells us about the frequency of salaries. The most common salary seems to be $150k.

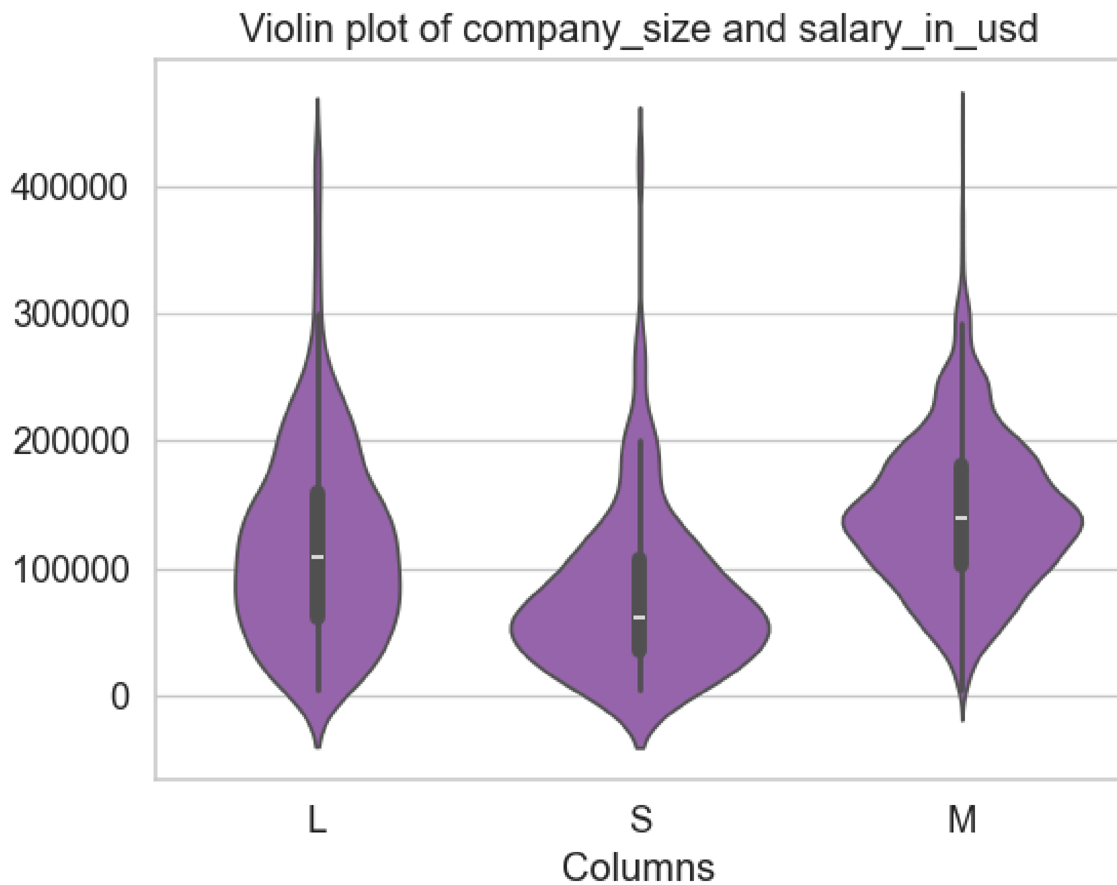


The histogram below tells us about the frequency of work years. 2022 and 2023 are very common.



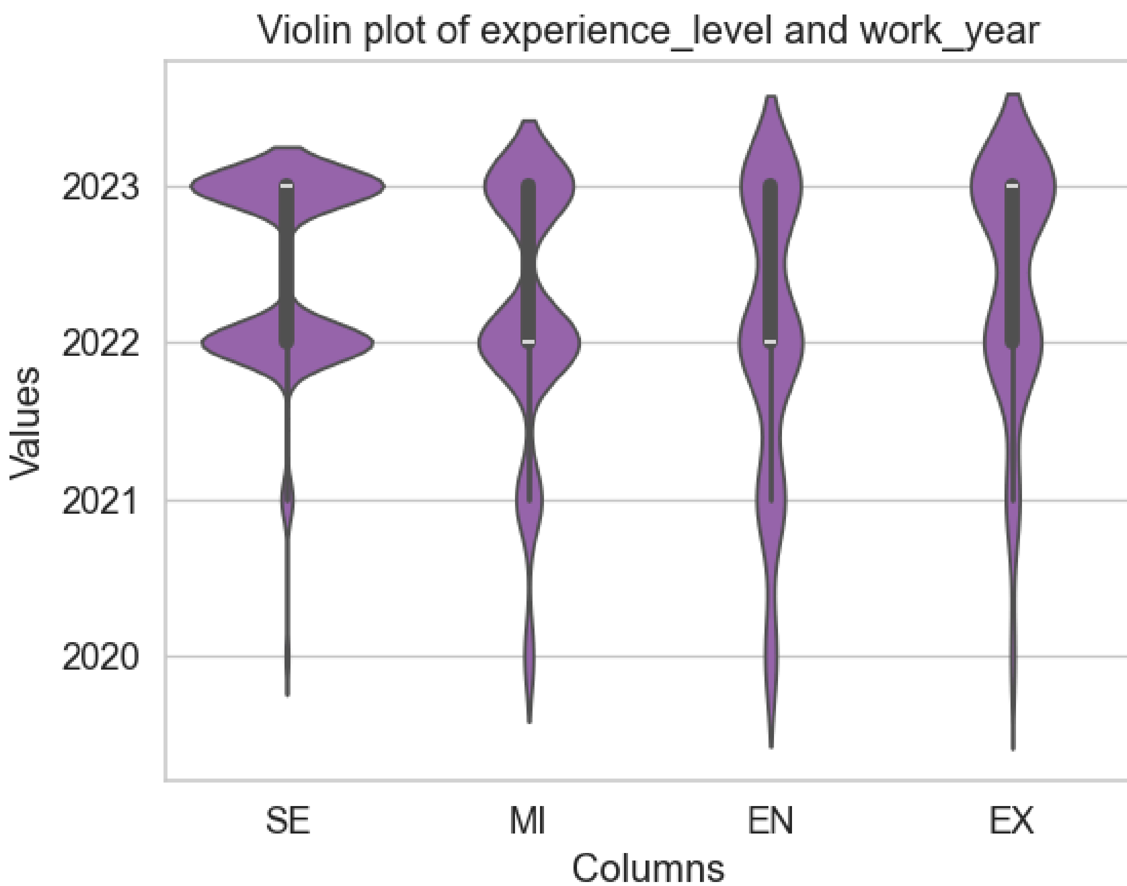
This scatter plot tells us how correlated wrk_year and salary are. There doesn't appear to be much of a correlation though



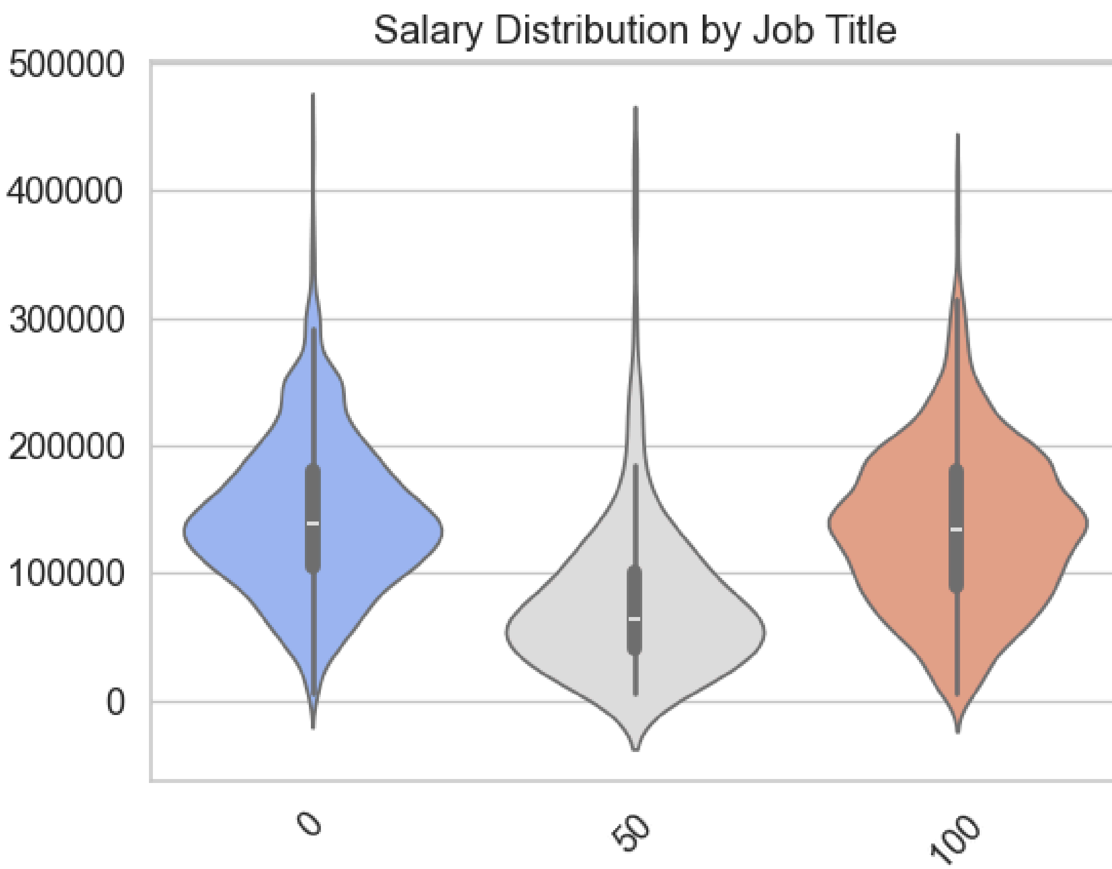


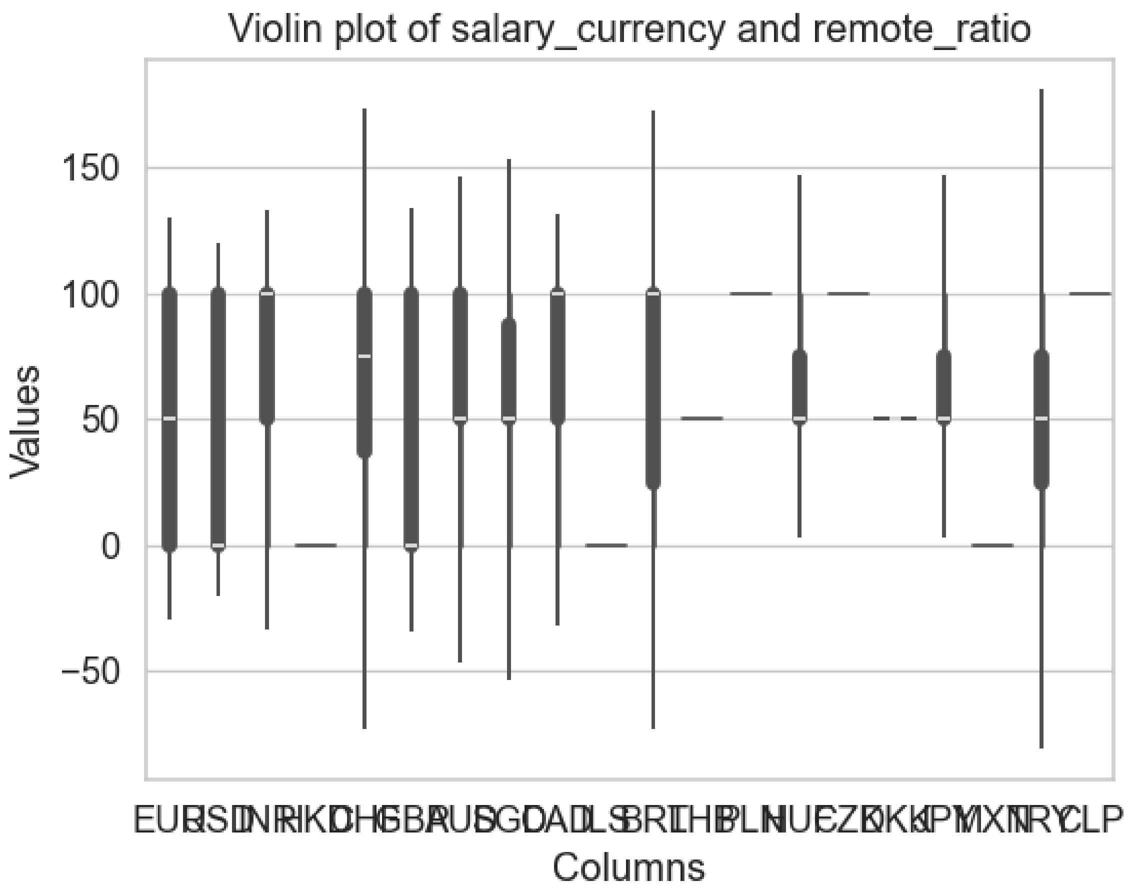
This violin plot illustrates how company size and salary are correlated. Small companies pay less, which makes sense.

The violin plot illustrates how experience level and work year are related. The results make sense.



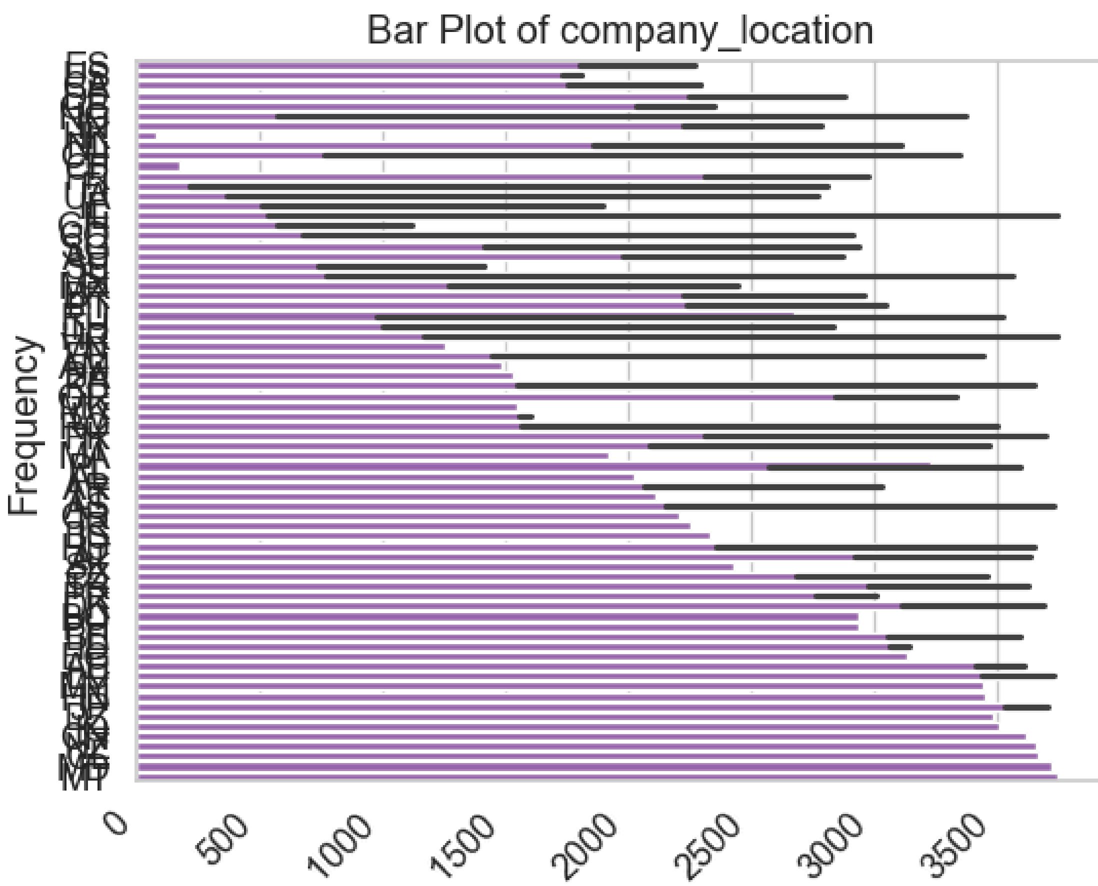
This violin plot illustrates how remote work ratio and salary distribution are related (the title is a typo). Remote and in person work do not have much of a difference.





This violin plot is too crowded, which suggests that a different plot would represent this data better.

This bar plot is too crowded to be useful, unfortunately. Maybe rotating the axis labels would help.



Interesting patterns:
Remote work did not seem to have an effect on salary compared to full-time, but half remote saw a significant salary drop.

There were data science jobs all around the world, which was cool to see visualized.

Insights:
It’s very easy for categorical column plots to become too crowded. I wonder what methods there are to solve this problem.

Questions: none
Signatures: Rahul Basak, Arnav Kaul, Aden Zhao