



ANOVA ANALYSIS

Analysis of Data Science Engineer's Salary

Team Leader:

110060012

Team Members:

109021226

109011214

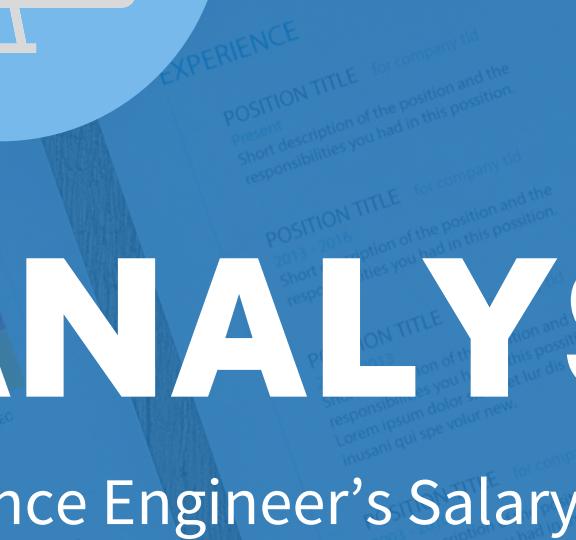
110071014

徐竣霆

呂柏緯

陳姿穎

蘇哲正



A detailed curriculum vitae (CV) template for Samantha Black, Sales Director. The CV includes sections for Personal Information, Experience, Education, Skills, and References. It features a professional layout with blue and white colors, matching the overall theme of the presentation.

Personal Information: Name: SAMANTHA BLACK, Position: sales director, Contact: PHONE 0028 01234 5678, EMAIL info@sambblack.com, WEBSITE www.mypage.com, SKYPE skype:sambblack.

Experience: Position Title: sales director, Company: for company ltd, Period: Present, Description: Short description of the position and the responsibilities you had in this position.

Education: WEB ADVERTISING SEMINAR, 2015, University of London, UK; GRAPHIC DESIGN CREW, 2013, London Art College, UK; HIGH SCHOOL UNIVERSITY, 2008 - 2014, Short description of the school and the responsibilities you had in this position. Lorem ipsum dolor sit amet, ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Skills: PHOTOGRAPHY, PHOTOSHOP, INDESIGN, WORDPRESS, TIME KEEPING, ORGANISATION.

References: ELIOT BROWN, 0028 01234 5678, info@eliottbrown.com; JAMES BROWN, 0028 01234 5678, info@jamesbrown.com; ROBERT BROWN, 0028 01234 5678, info@robertbrown.com.

Introduction

This dataset represents salary information for Machine Learning(Data Science) engineers.

For each entry, contains 11 columns, including details such as the level of experience, employment type, job title, salary amount, employee residence, remote work ratio, company location, and company size.

▲ experienc...	▲ employme...	▲ job_title	▲ # salary_in_...	▲ employee_...	▲ # remote_ra...	▲ company_l...	▲ company_...
MI	FT	Data Scientist	120000	AU	0	AU	S
MI	FT	Data Scientist	70000	AU	0	AU	S
MI	CT	Data Scientist	130000	US	0	US	M
MI	CT	Data Scientist	110000	US	0	US	M
MI	FT	Data Science Manager	240000	US	0	US	M
MI	FT	Data Science Manager	180000	US	0	US	M
SE	FT	Business Intelligence Engineer	202800	US	0	US	M
SE	FT	Business Intelligence Engineer	115000	US	0	US	M

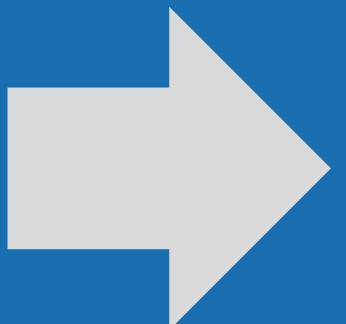
This information allows for analysis of salary trends, employment patterns, and other factors affecting machine learning engineer salaries in various locations and company settings.

Data Preprocessing

- We drop job_title(not important).
- We drop salary, salary_currency(USD as measurement).
- We drop those entries with company location not in US(US consists of 87.8%).

Total 16494 entries, 7 categorical features.

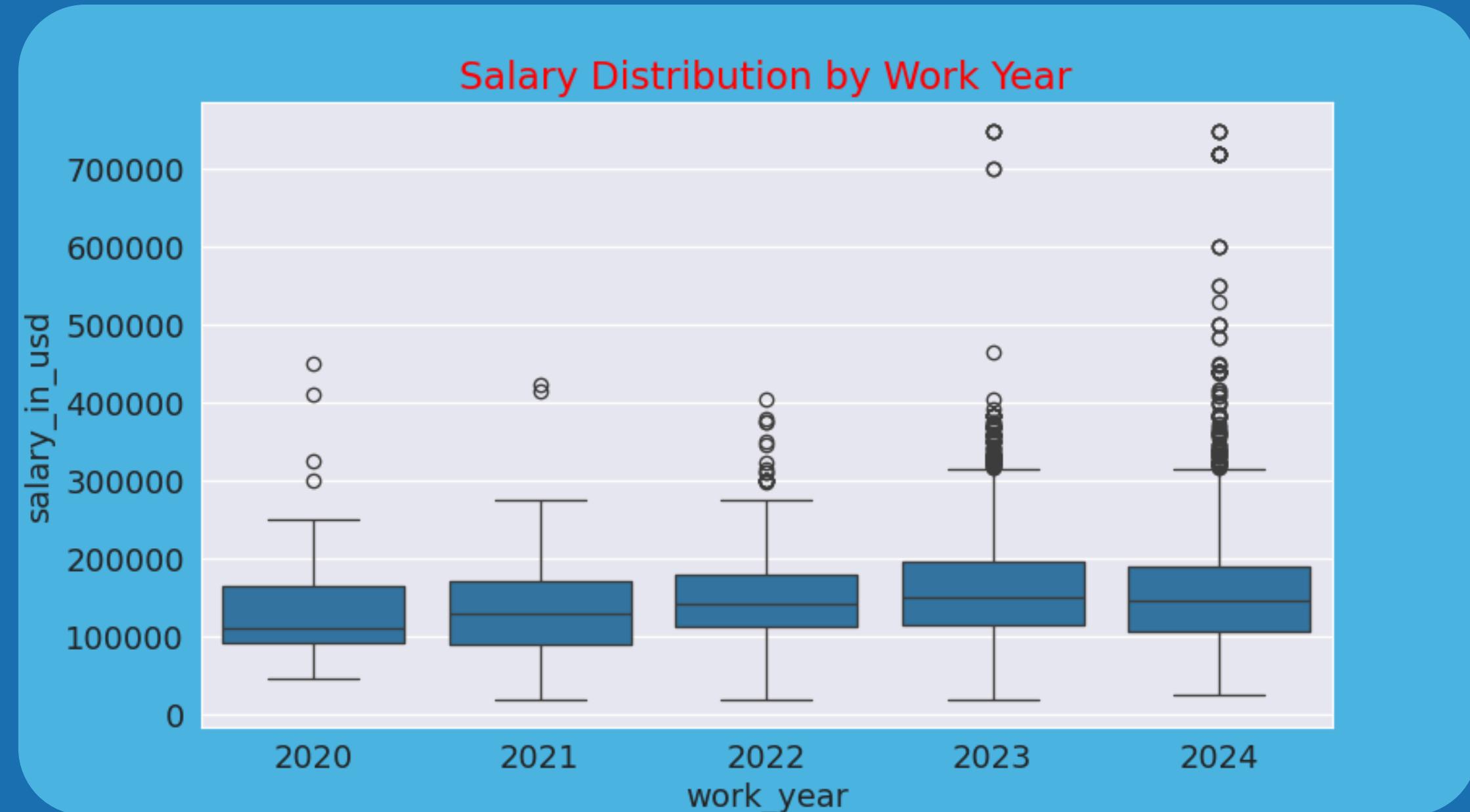
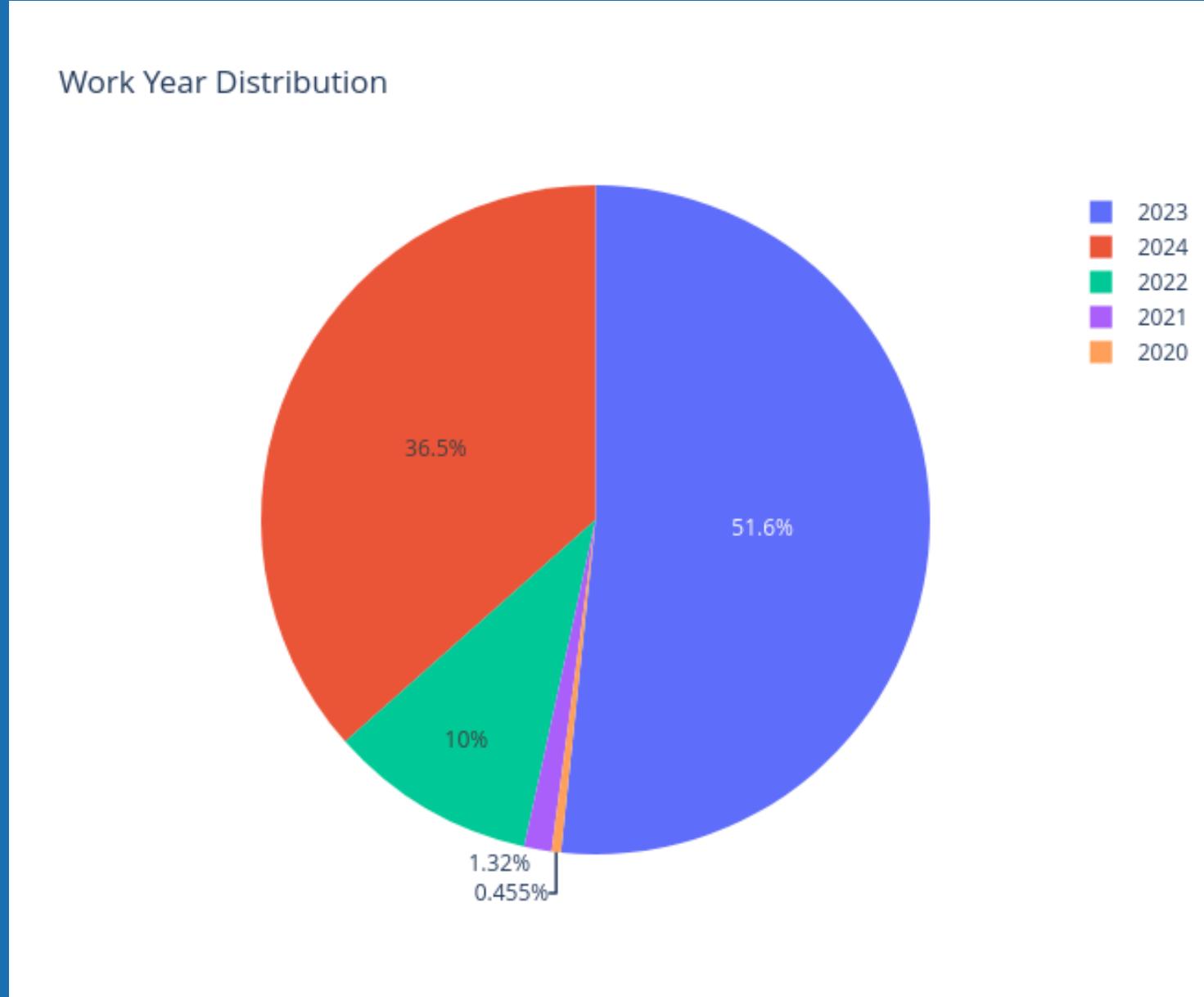
```
RangeIndex: 16494 entries, 0 to 16493
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
 --- 
 0   work_year        16494 non-null   int64  
 1   experience_level 16494 non-null   category
 2   employment_type  16494 non-null   category
 3   job_title         16494 non-null   category
 4   salary            16494 non-null   int64  
 5   salary_currency   16494 non-null   category
 6   salary_in_usd    16494 non-null   int64  
 7   employee_residence 16494 non-null   category
 8   remote_ratio      16494 non-null   int64  
 9   company_location  16494 non-null   category
 10  company_size     16494 non-null   category
dtypes: category(7), int64(4)
memory usage: 656.4 KB
None
```



14478 entries, 4 categorical features.

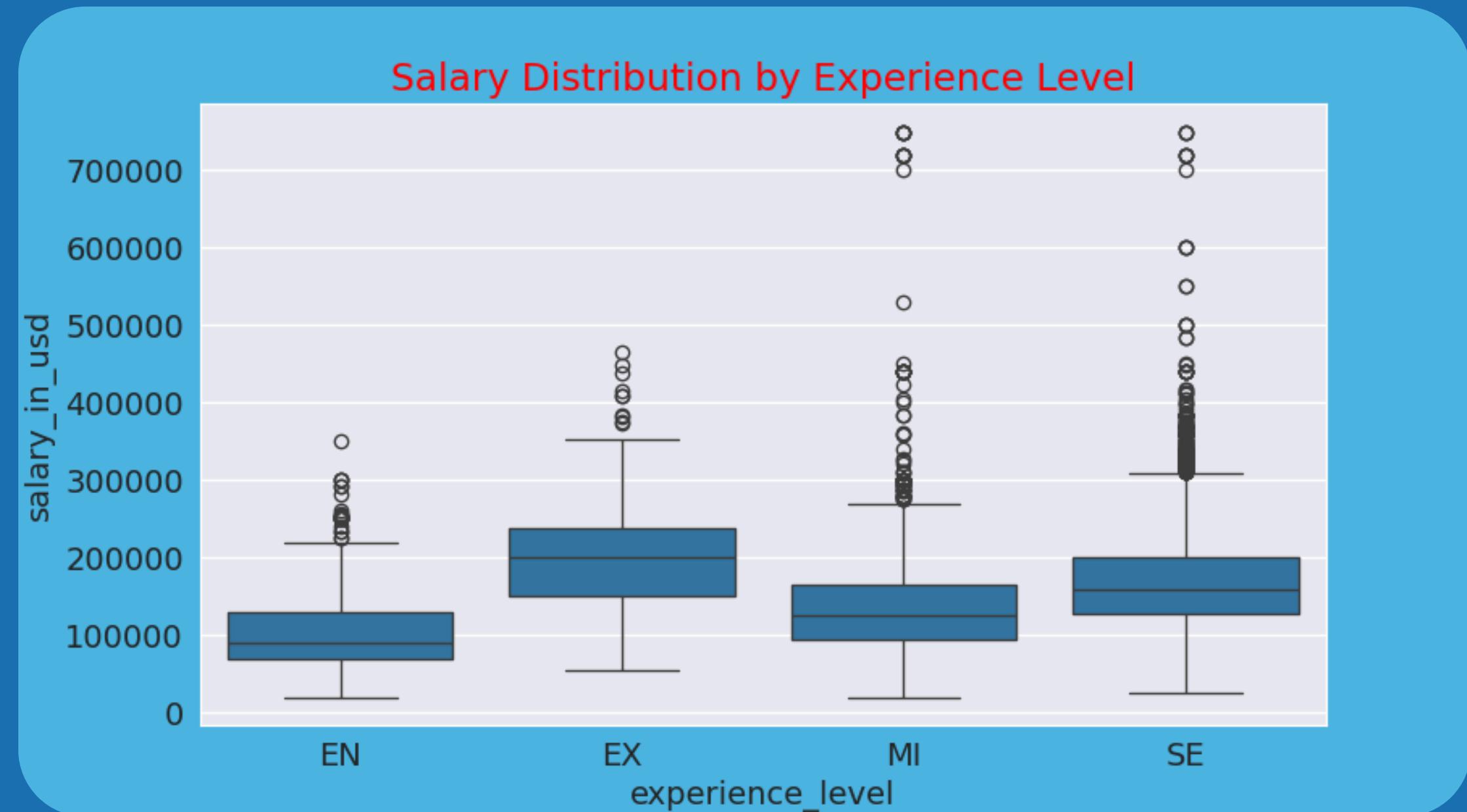
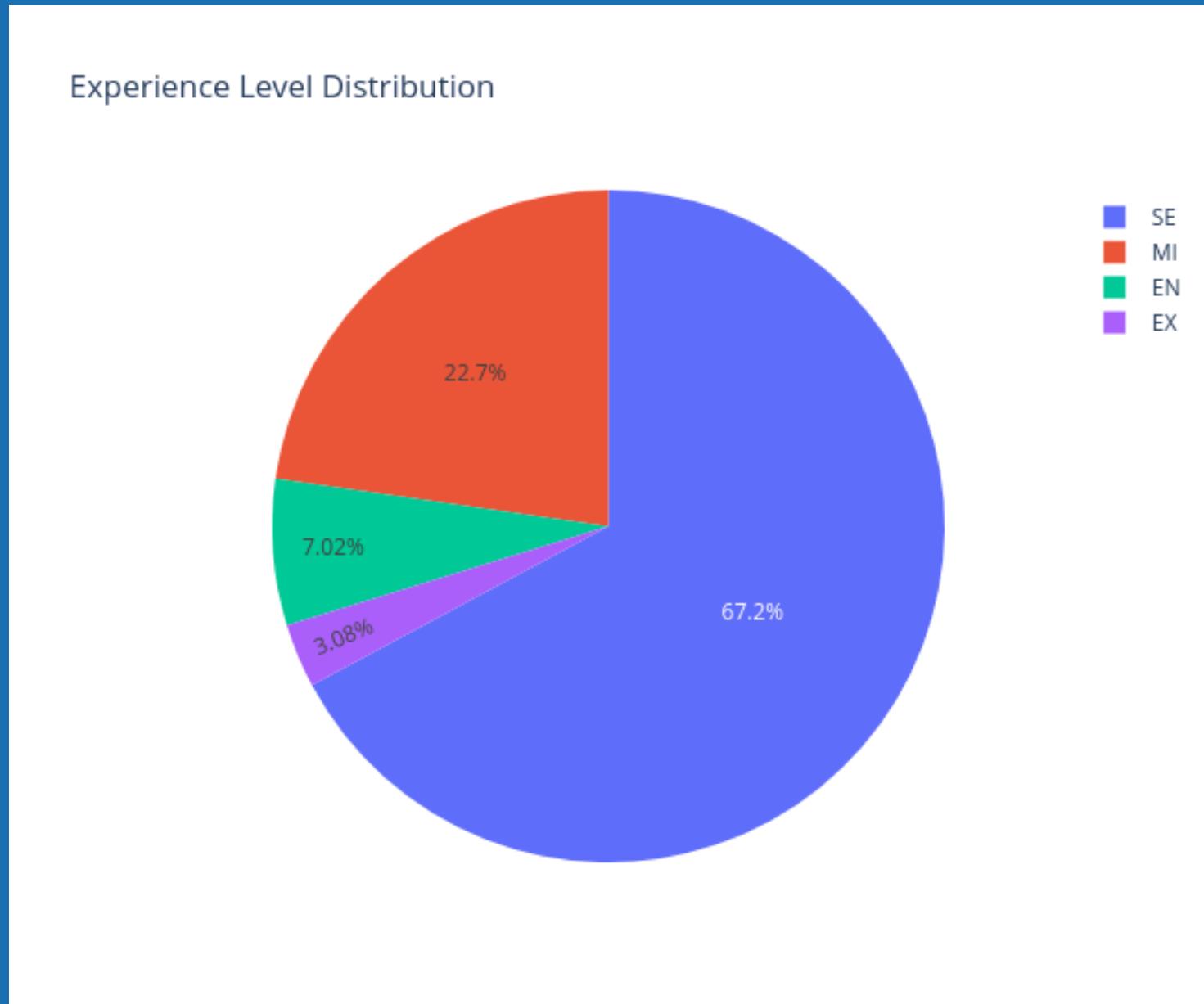
```
Index: 14478 entries, 2 to 16492
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
 --- 
 0   work_year        14478 non-null   int64  
 1   experience_level 14478 non-null   category
 2   employment_type  14478 non-null   category
 3   salary_in_usd    14478 non-null   int64  
 4   employee_residence 14478 non-null   category
 5   remote_ratio      14478 non-null   int64  
 6   company_size     14478 non-null   category
dtypes: category(4), int64(3)
memory usage: 512.3 KB
None
```

Visualization - Work Year



Visualization - Experience Level

EN: Entry Level SE: Senior Level
MI: Mid Level EX: Executive Level



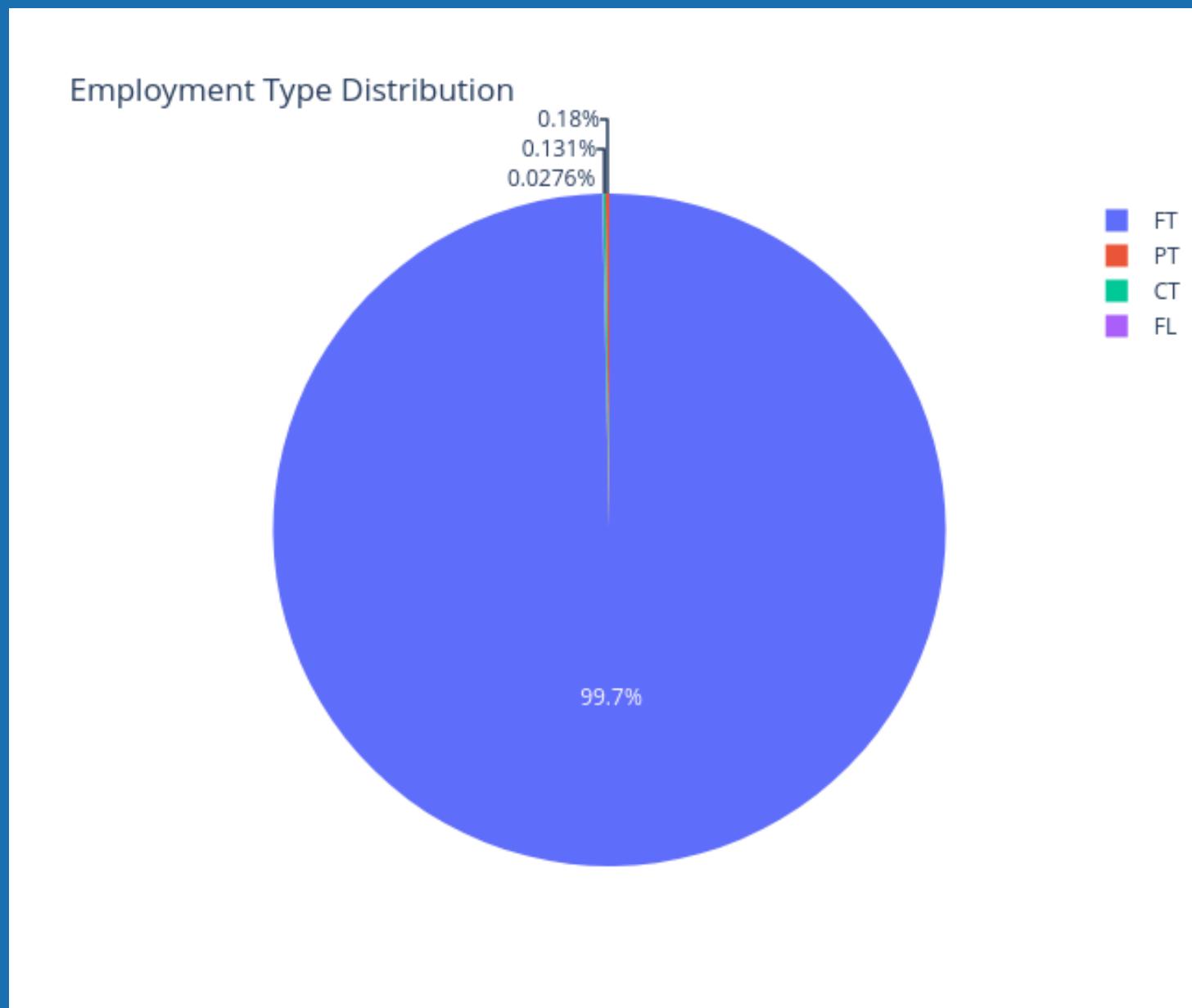
Visualization - Employment Type

FL: Freelancer

PT: Part Time

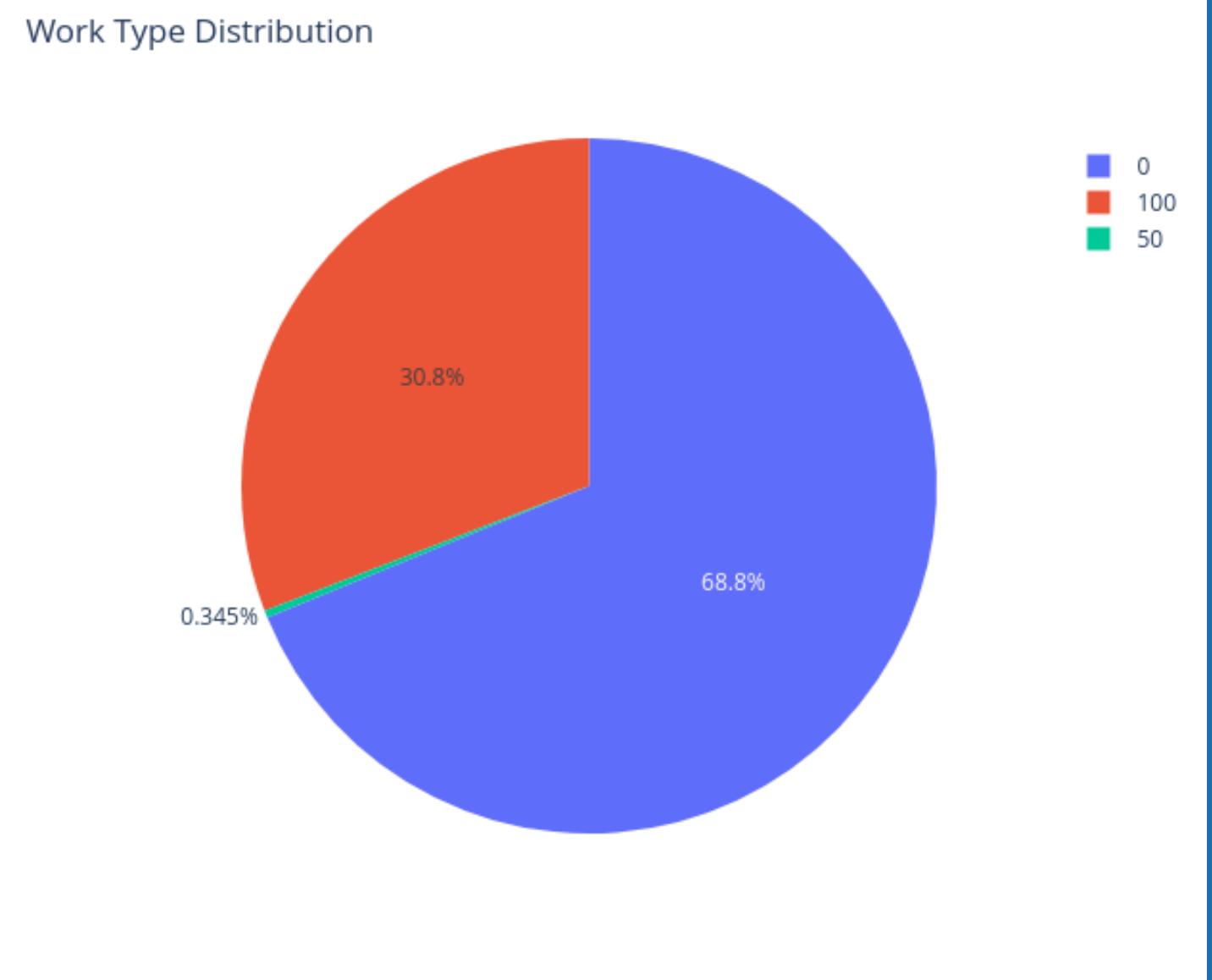
CT: Contract basis

FT: Full Time



Visualization - Work Type

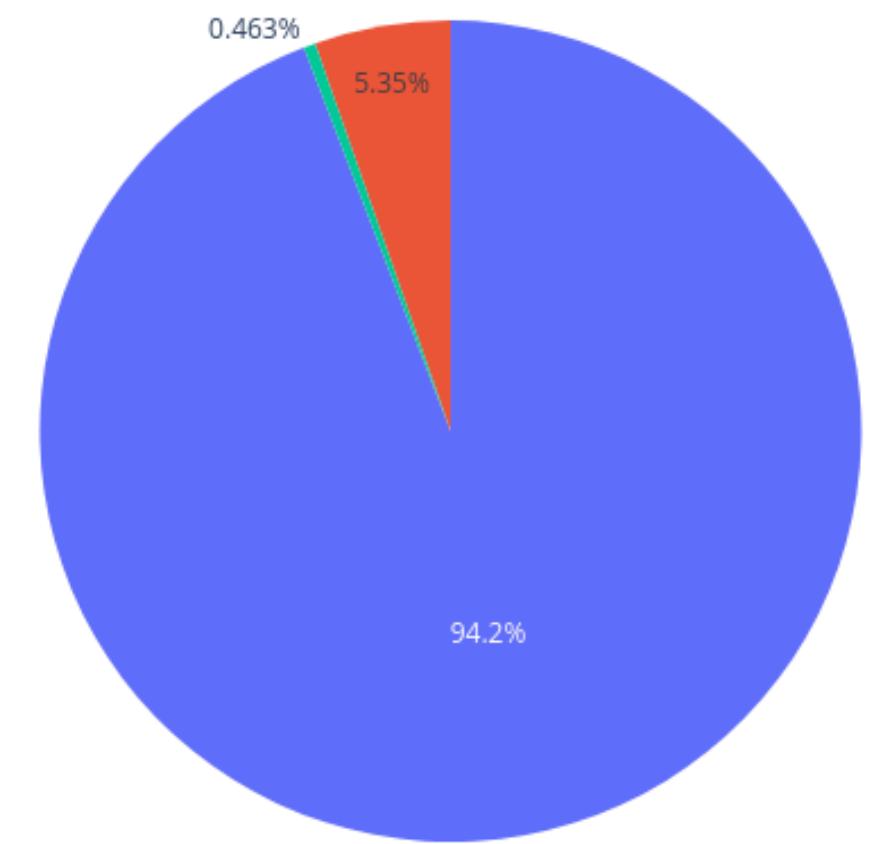
0: Office Work
50: Hybrid
100: Work from Home



Visualization - Company Size

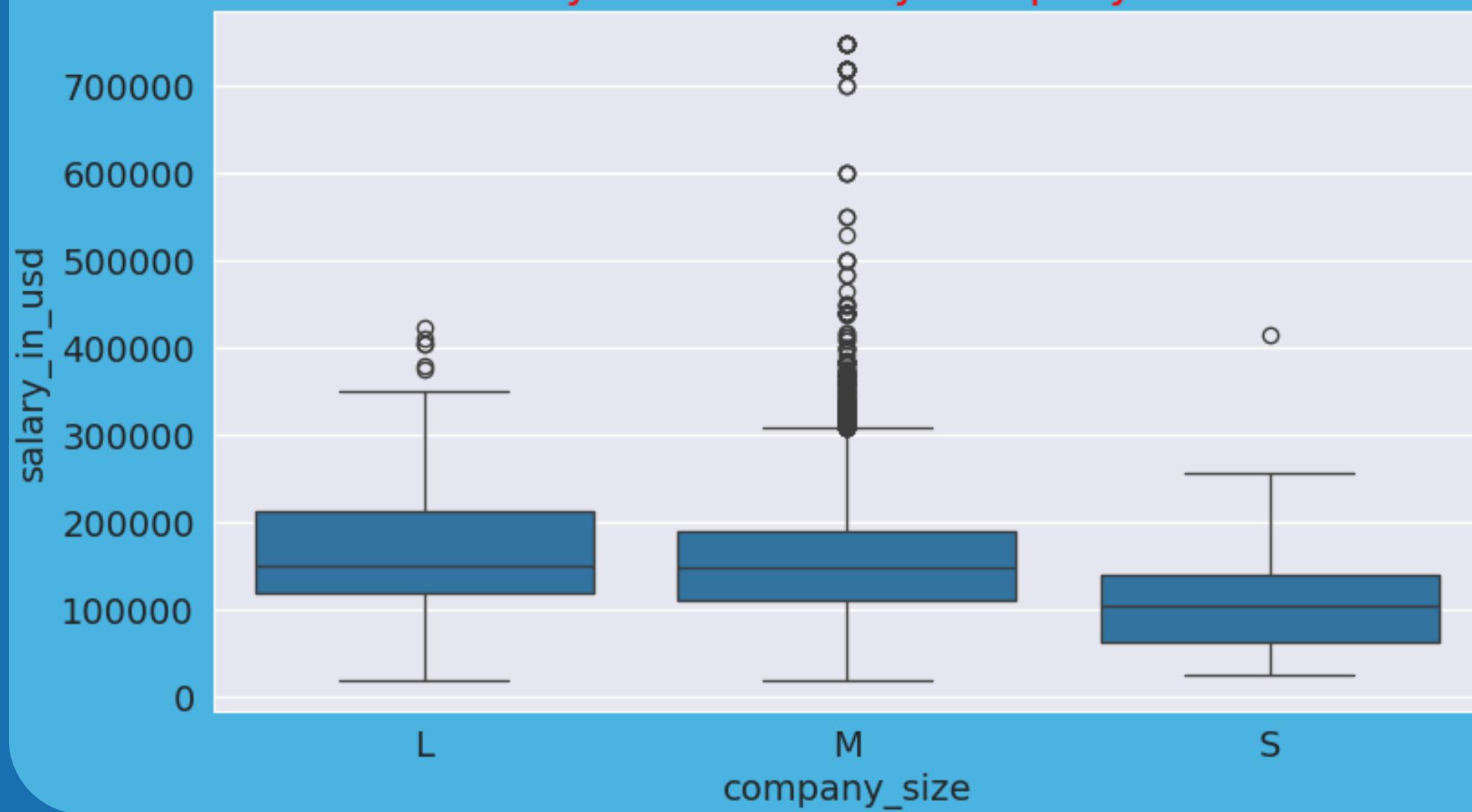
S: Small
M: Medium
L: Large

Company Size Distribution



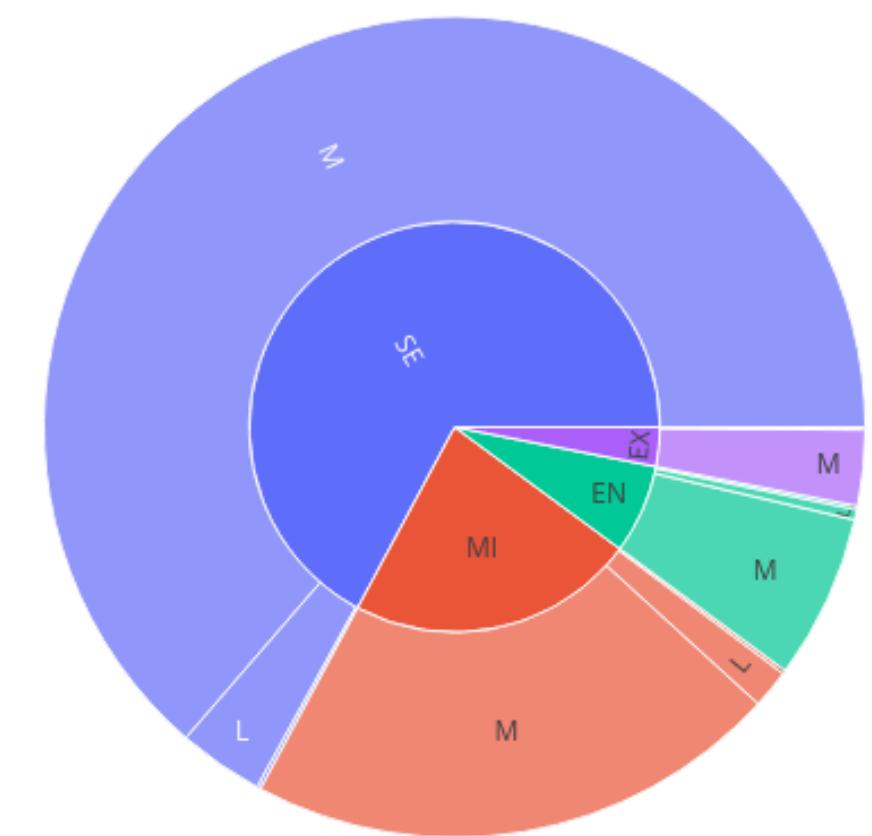
M
L
S

Salary Distribution by Company Size

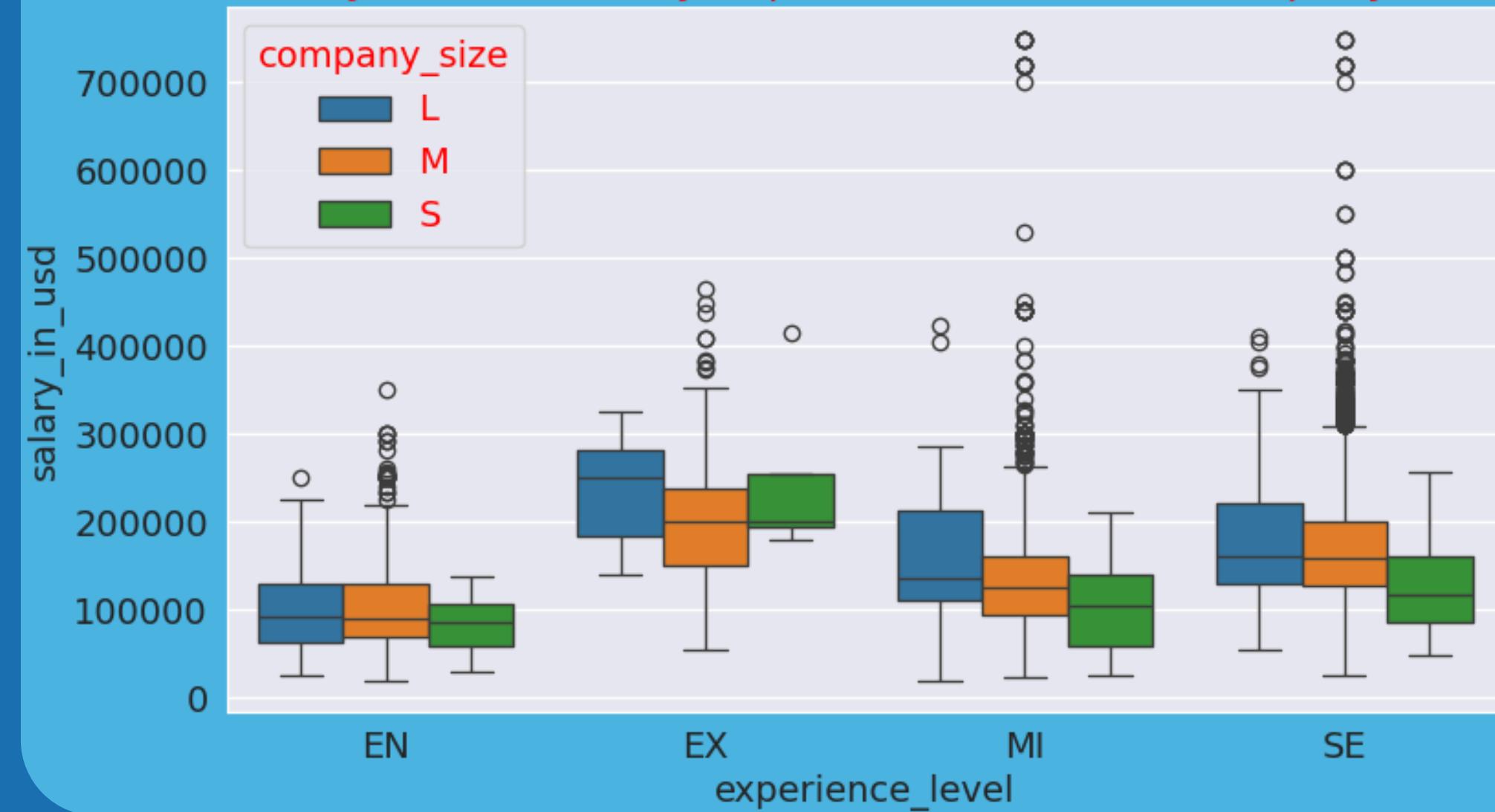


Visualization - Company Size ~ Experience Level

Experience Level and Company Size Distribution



Salary Distribution by Experience Level and Company Size



Hypothesis Setting

- Want to know that if there is an impact on salary by
 - company size
 - experience level
- Company size
 - Set $H_0: \mu_1 = \mu_2 = \mu_3$ against $H_1: H_0$ is NOT true
- Experience level
 - Set $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_1: H_0$ is NOT true

Introduction - Summarization

Company Size

company_size <fctr>	group_size <int>	mean_salary <dbl>	var_salary <dbl>
L	775	163876.5	4180752190
M	13636	156768.4	4318185977
S	67	114857.2	4259542012

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	163877	2358	69.486	< 2e-16
data\$company_sizeM	-7108	2424	-2.932	0.00338
data\$company_sizes	-49019	8361	-5.863	4.64e-09

company_size <fctr>	group_size <int>	mean_salary <dbl>	var_salary <dbl>
L	769	162034.2	3772159152
M	13343	151866.3	3049765625
S	66	110294.4	2908749533

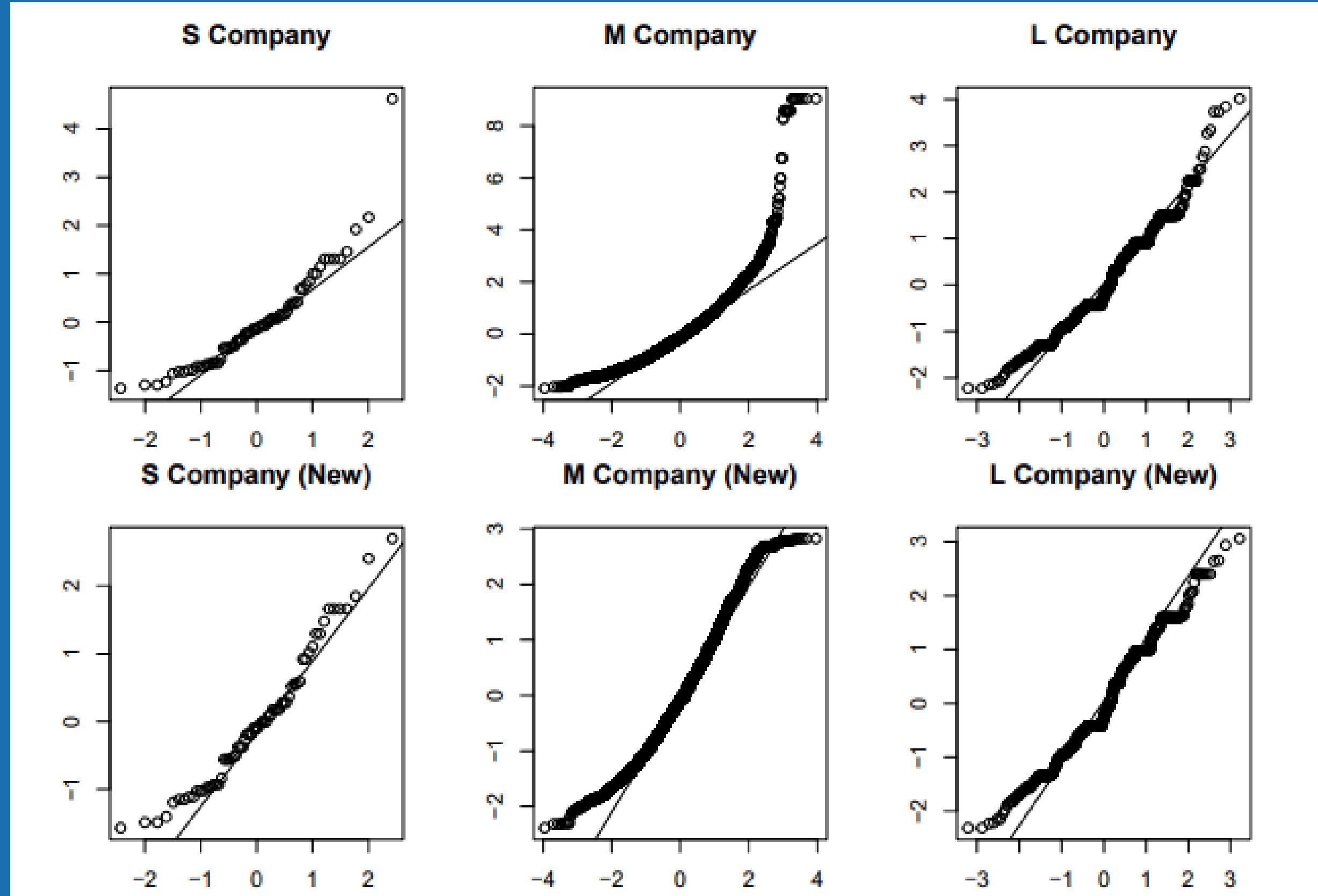
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	162034	2004	80.856	< 2e-16
processed_data\$company_sizeM	-10168	2061	-4.934	8.16e-07
processed_data\$company_sizes	-51740	7128	-7.259	4.11e-13

<--- Raw data

<--- Processed data
remove the outlier that outside
[Q1 - 1.5 IQR , Q3+1.5 IQR]

Model Diagnostics - Normality

Company Size



<--- Raw data

<--- Processed data
remove the outlier that outside
[Q1 - 1.5 IQR , Q3+1.5 IQR]

Model Diagnostics - Homogeneity

Company Size

```
Levene's Test for Homogeneity of Variance (center = mean)
  Df F value Pr(>F)
group  2 4.7596 0.008582 **
      14475
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<--- Raw data

```
Levene's Test for Homogeneity of Variance (center = mean)
  Df F value Pr(>F)
group  2 20.242 1.666e-09 ***
      14175
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<--- Processed data
remove the outlier that outside
[Q1 - 1.5 IQR , Q3 + 1.5 IQR]

ANOVA

Company Size

```
          Df  Sum Sq  Mean Sq F value Pr(>F)
company_size     2 1.563e+11 7.817e+10   18.14 1.36e-08 ***
Residuals    14475 6.240e+13 4.311e+09
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<---- Raw data

```
          Df  Sum Sq  Mean Sq F value Pr(>F)
company_size     2 1.917e+11 9.587e+10   31.05 3.52e-14 ***
Residuals    14175 4.378e+13 3.088e+09
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<---- Processed data
remove the outlier that outside
[Q1 - 1.5 IQR , Q3 + 1.5 IQR]

Introduction - Summarization

Experience Level

experience_level <fctr>	group_size <int>	mean_salary <dbl>	var_salary <dbl>
EN	1017	102587.1	2216775262
EX	446	200829.7	4753351772
MI	3291	135429.4	4232276146
SE	9724	167913.8	3877693626

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	102587	1950	52.60	<2e-16
data\$experience_level EX	98243	3532	27.81	<2e-16
data\$experience_level MI	32842	2232	14.72	<2e-16
data\$experience_level SE	65327	2050	31.87	<2e-16

<--- Raw data

experience_level <fctr>	group_size <int>	mean_salary <dbl>	var_salary <dbl>
EN	992	98559.14	1591651109
EX	436	196034.46	3813384376
MI	3224	129911.32	2304660037
SE	9501	163198.42	2824313693

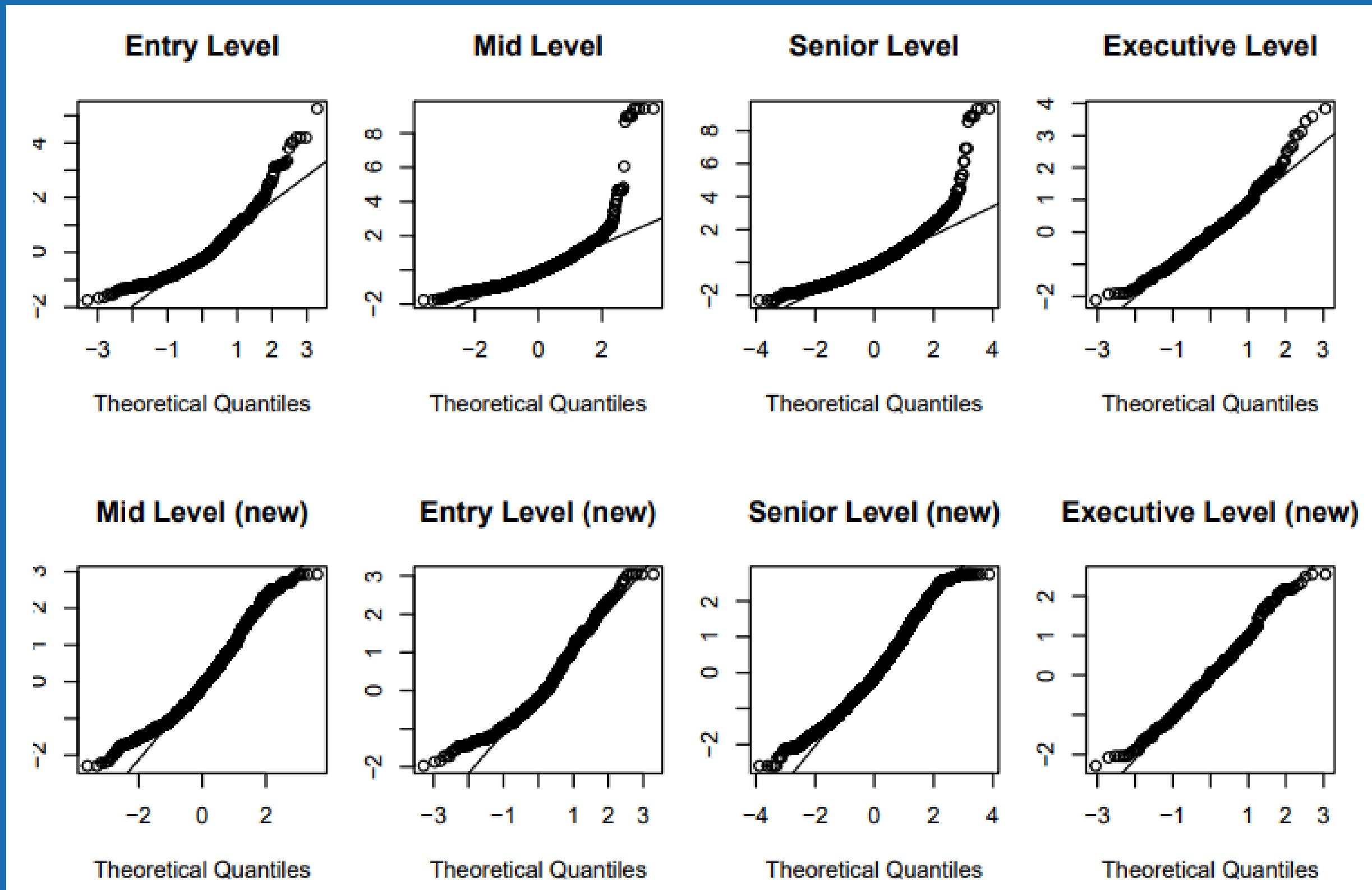
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	98559	1634	60.30	<2e-16
processed_data\$experience_level EX	97475	2958	32.95	<2e-16
processed_data\$experience_level MI	31352	1869	16.77	<2e-16
processed_data\$experience_level SE	64639	1718	37.63	<2e-16

<--- Processed data

remove the outlier that outside
[Q1 - 1.5 IQR , Q3+1.5 IQR]

Model Diagnostics - Normality

Experience Level



<--- Raw data

<--- Processed data
remove the outlier that outside
[Q1 - 1.5 IQR , Q3+1.5 IQR]

Model Diagnostics - Homogeneity

Experience Level

```
Levene's Test for Homogeneity of Variance (center = mean)
  Df F value    Pr(>F)
group      3 26.702 < 2.2e-16 ***
               14474
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<--- Raw data

```
Levene's Test for Homogeneity of Variance (center = mean)
  Df F value    Pr(>F)
group      3 55.746 < 2.2e-16 ***
               14149
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<--- Processed data
remove the outlier that outside
[Q1 - 1.5 IQR, Q3 + 1.5 IQR]

ANOVA

Experience Level

```
Df      Sum Sq   Mean Sq F value Pr(>F)
experience_level    3 6.557e+12 2.186e+12     565 <2e-16 ***
Residuals          14474 5.599e+13 3.869e+09
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<--- Raw data

```
Df      Sum Sq   Mean Sq F value Pr(>F)
experience_level    3 6.443e+12 2.148e+12     810.4 <2e-16 ***
Residuals          14149 3.750e+13 2.650e+09
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<--- Processed data
remove the outlier that outside
[Q1 - 1.5 IQR , Q3 + 1.5 IQR]

Conclusion

- The size of a company has an impact on salaries.
 - The difference between the size “M” and “L” is smaller than the difference between “S” and the others.
- The experience level has a significant impact on salaries.
- Kick out the outliers may lead to
 - increase the normality.
 - worsen the homogeneity of variance.
 - make the results of ANOVA more significant.

Reference

- Dataset: [Machine Learning Engineer Salary in 2024](#)
- Article reference: [Data Scientist Job Salaries Analysis](#)

THANKS FOR LISTENING



SAMANTHA BLACK
sales director

EXPERIENCE

POSITION TITLE: Present
Short description of the position and the responsibilities you had in this position.

POSITION TITLE: Previous
Short description of the position and the responsibilities you had in this position.

POSITION TITLE: Previous
Short description of the position and the responsibilities you had in this position.

POSITION TITLE: Previous
Short description of the position and the responsibilities you had in this position.

POSITION TITLE: Previous
Short description of the position and the responsibilities you had in this position.

EDUCATION

WEB ADVERTISING SEMINAR
2015
University of London, UK

GRAPHIC DESIGN CREW
2013
London Art College, UK
Leader of the group, lorem ipsum

HIGH SCHOOL UNIVERSITY
2008 - 2014
Short description of the school and the responsibilities you had in this position.
Lorem ipsum dolor sit amet, ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

SCHOOL TITLE LOREM
2004 - 2008
Short description of the position and the responsibilities you had in this position.

PROFESSIONAL STATEMENT
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse suscipit efficitur lectus, Fusce laculis, leo nec vulputate efficitur lorem interdum elit, ut vestibulum nisl metus, non mi.

Aliquam dictum porta erat nec commodo. Maecenas vestibulum massa in justo pellentesque, non eleifend dolor ornare. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse suscipit efficitur lectus, Fusce laculis, leo nec vulputate efficitur lorem interdum elit, ut vestibulum nisl metus, non mi.

Aliquam dictum porta erat nec commodo. Maecenas vestibulum massa in justo pellentesque, non eleifend dolor ornare. Lorem ipsum dolor sit amet, consectetur nisl.

SKILLS

PHOTOGRAPHY
PHOTOSHOP
INDESIGN
WORDPRESS
TIME KEEPING
ORGANISATION

REFERENCES

ELIOT BROWN
0028 01234 5678
eliot@mypage.com

ELIOT BROWN
0028 01234 5678
eliot@mypage.com

ELIOT BROWN
0028 01234 5678
eliot@mypage.com