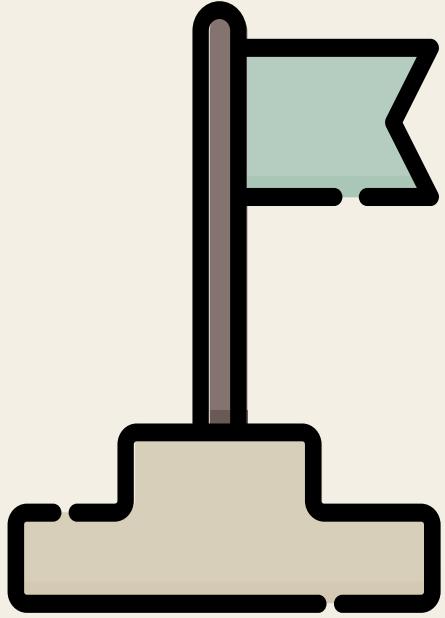


Data Science Practices Final Project



Data source : [Data Science Job Salaries](#)



1. EDA

Original Data

```
> head(df_original)
```

	X	work_year	experience_level	employment_type	job_title	salary
1	0	2020	MI	FT	Data Scientist	70000
2	1	2020	SE	FT	Machine Learning Scientist	260000
3	2	2020	SE	FT	Big Data Engineer	85000
4	3	2020	MI	FT	Product Data Analyst	20000
5	4	2020	SE	FT	Machine Learning Engineer	150000
6	5	2020	EN	FT	Data Analyst	72000

	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
1	EUR	79833	DE	0	DE	L
2	USD	260000	JP	0	JP	S
3	GBP	109024	GB	50	GB	M
4	USD	20000	HN	0	HN	S
5	USD	150000	US	50	US	L
6	USD	72000	US	100	US	L

> |

Data Structure

```
> str(df_original)
'data.frame': 607 obs. of 12 variables:
 $ X           : int 0 1 2 3 4 5 6 7 8 9 ...
 $ work_year   : int 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
 $ experience_level : chr "MI" "SE" "SE" "MI" ...
 $ employment_type   : chr "FT" "FT" "FT" "FT" ...
 $ job_title     : chr "Data Scientist" "Machine Learning Scientist" "Big Data Engineer" "Product
Data Analyst" ...
 $ salary        : int 70000 260000 85000 20000 150000 72000 190000 1100000 135000 125000 ...
 $ salary_currency : chr "EUR" "USD" "GBP" "USD" ...
 $ salary_in_usd    : int 79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...
 $ employee_residence: chr "DE" "JP" "GB" "HN" ...
 $ remote_ratio    : int 0 0 50 0 50 100 100 50 100 50 ...
 $ company_location : chr "DE" "JP" "GB" "HN" ...
 $ company_size     : chr "L" "S" "M" "S" ...
```

Distribution

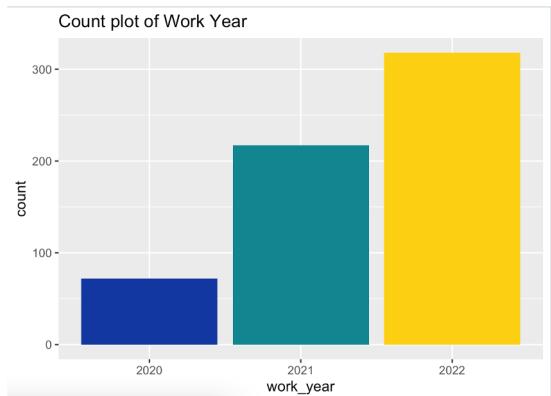
```
> # salary's range is large since contractors gets paid in hours while others get paid in month)
> summary(df_original)
```

X	work_year	experience_level	employment_type	job_title
Min. : 0.0	Min. :2020	Length:607	Length:607	Length:607
1st Qu.:151.5	1st Qu.:2021	Class :character	Class :character	Class :character
Median :303.0	Median :2022	Mode :character	Mode :character	Mode :character
Mean :303.0	Mean :2021			
3rd Qu.:454.5	3rd Qu.:2022			
Max. :606.0	Max. :2022			
salary	salary_currency	salary_in_usd	employee_residence	remote_ratio
Min. : 4000	Length:607	Min. : 2859	Length:607	Min. : 0.00
1st Qu.: 70000	Class :character	1st Qu.: 62726	Class :character	1st Qu.: 50.00
Median : 115000	Mode :character	Median :101570	Mode :character	Median :100.00
Mean : 324000		Mean :112298		Mean : 70.92
3rd Qu.: 165000		3rd Qu.:150000		3rd Qu.:100.00
Max. :30400000		Max. :600000		Max. :100.00
company_location	company_size			
Length:607	Length:607			
Class :character	Class :character			
Mode :character	Mode :character			

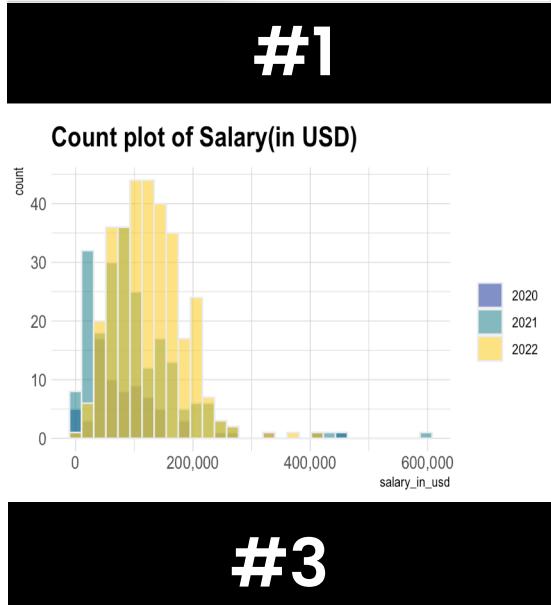
Categorical variable data

```
> table(df["experience_level"])
experience_level
  EN   EX   MI   SE
  88   26  213  280
> table(df["employment_type"])
employment_type
  CT   FL   FT   PT
    5    4  588   10
> table(df["remote_ratio"])
remote_ratio
  0  100  50
127 381  99
> table(df["company_size"])
company_size
  L    M    S
198 326   83
```

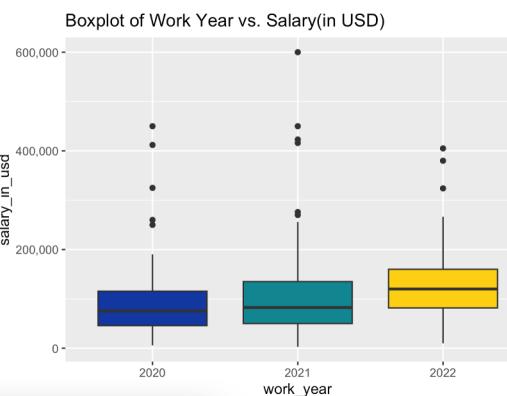
Visualization(Work Year)



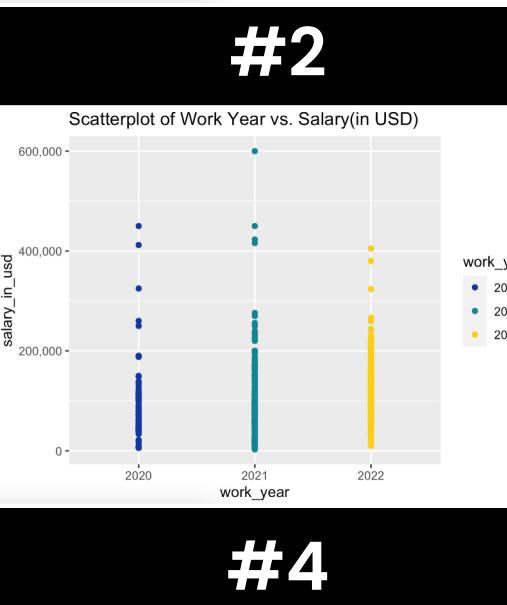
#1



#3



#2



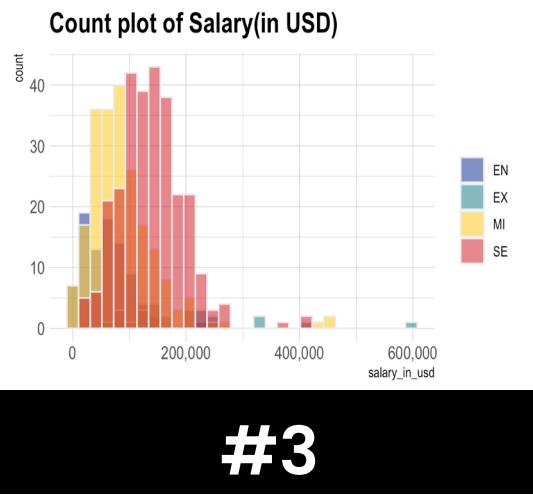
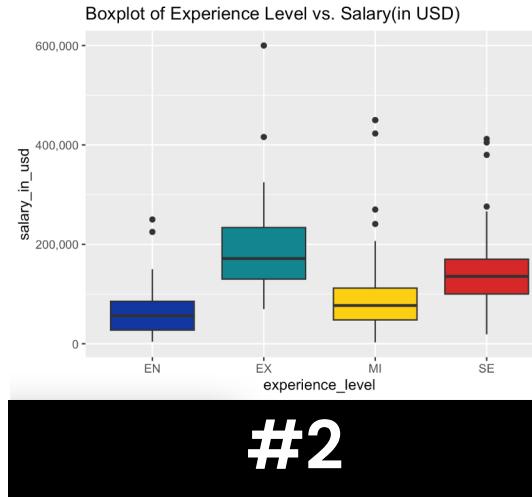
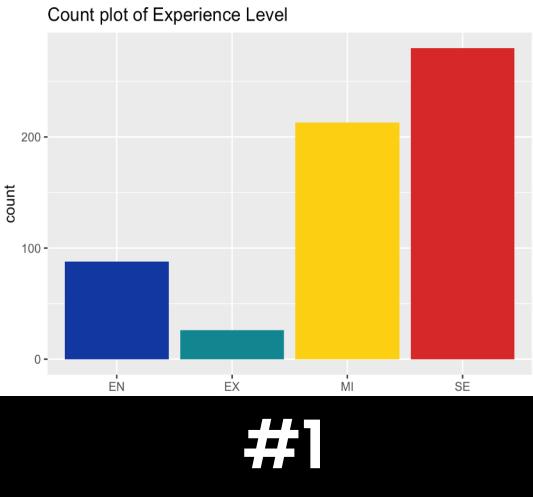
#4

- **Countplot**
Work Year
- **Countplot(Salary)**
Salary(In USD)

- **Boxplot**
Work vs. Salary(In USD)

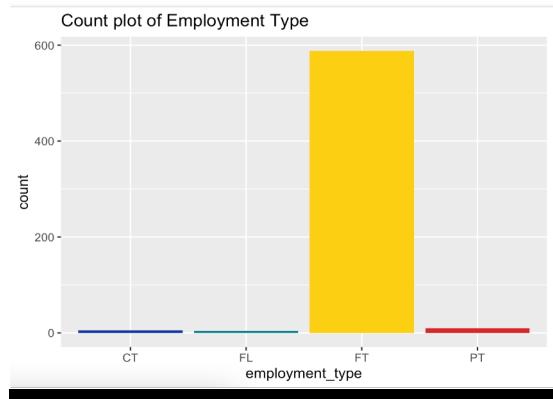
- **Scatterplot**
Work vs. Salary(In USD)

Visualization(Experience Level)

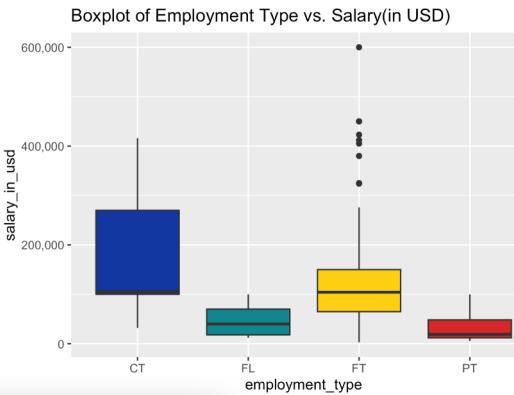


- **Countplot**
Experience Level
- **Countplot(Salary)**
Experience Level vs. Salary(In USD)
- **Boxplot**
Experience Level vs. Salary(In USD)
- **Scatterplot**
Experience Level vs. Salary(In USD)

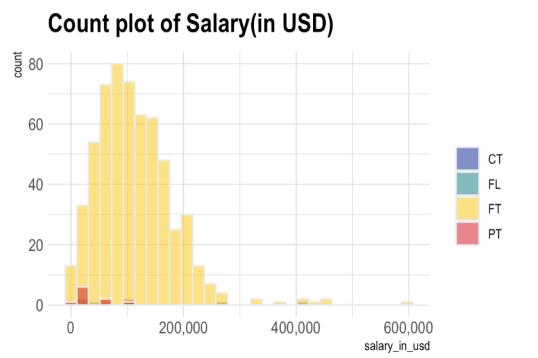
Visualization(Employment Type)



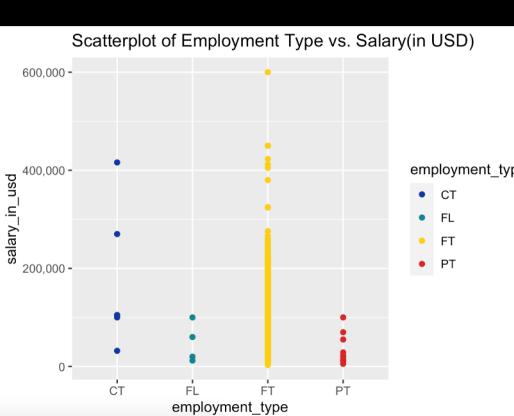
#1



#2



#3



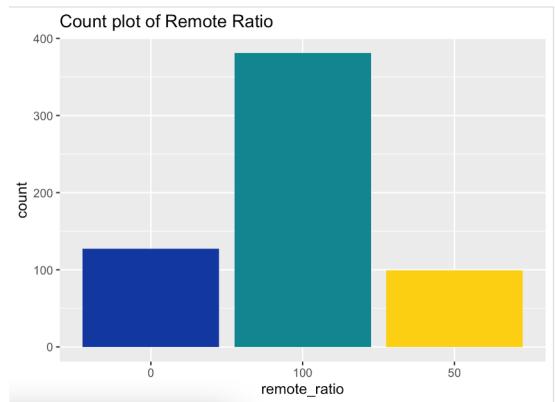
#4

- **Countplot**
Employment Type
- **Countplot(Salary)**
Employment Type vs. Salary(In USD)

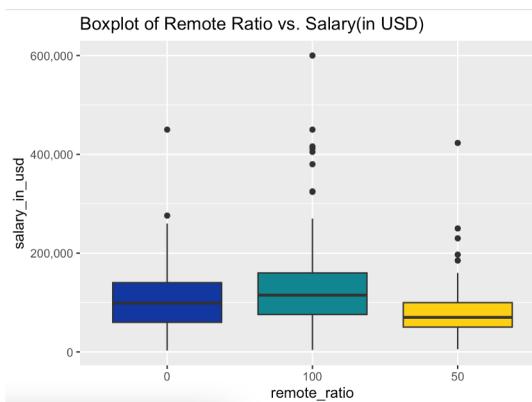
- **Boxplot**
Employment Type vs. Salary(In USD)

- **Scatterplot**
Employment Type vs. Salary(In USD)

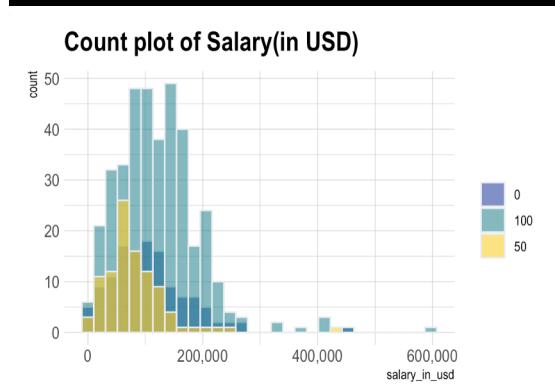
Visualization(Remote Rate)



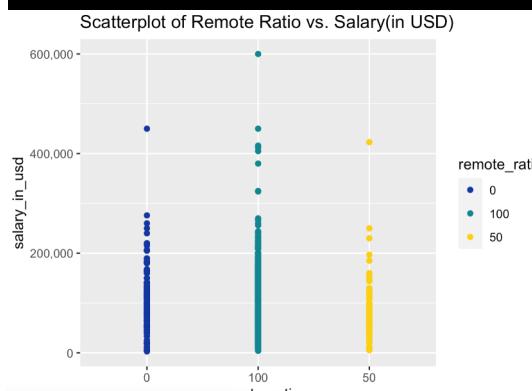
#1



#2



#3



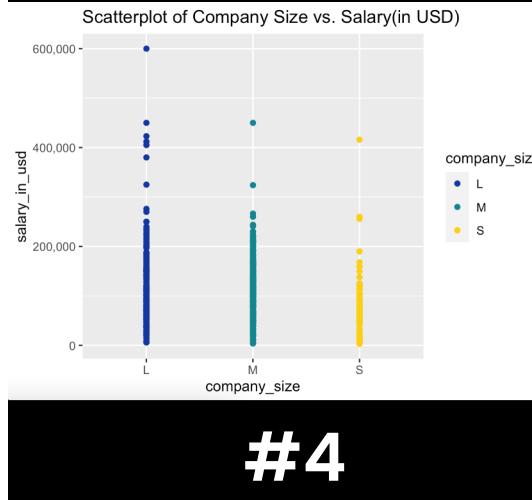
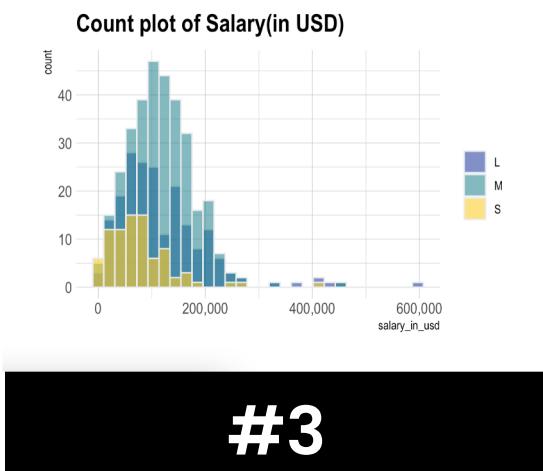
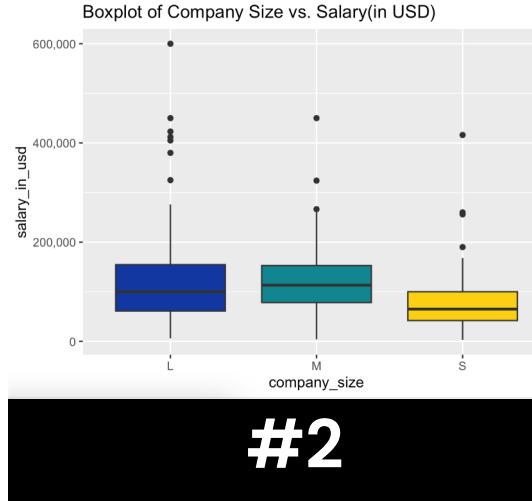
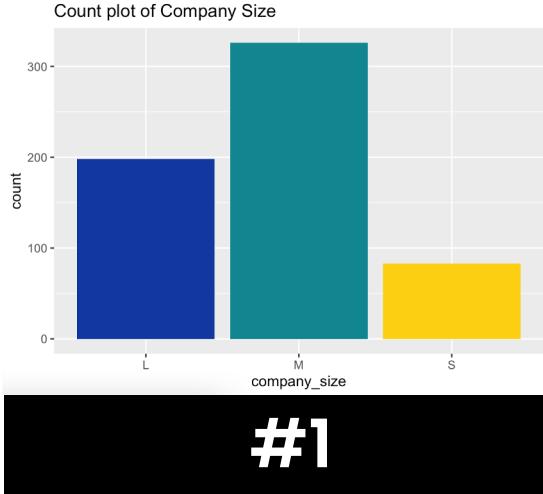
#4

- **Countplot**
Remote Ratio
- **Countplot(In USD)**
Remote Ratio vs. Salary(In USD)

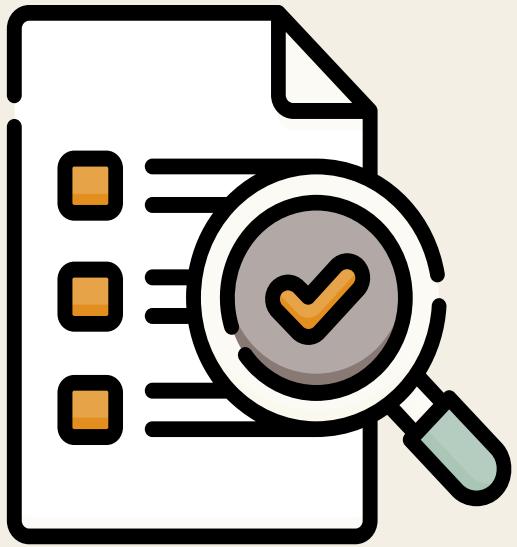
- **Boxplot**
Remote Ratio vs. Salary(In USD)

- **Scatterplot**
Remote Ratio vs. Salary(In USD)

Visualization(Company Size)



- **Countplot**
Company Size
- **Countplot(Salary)**
Company Size vs. Salary(In USD)
- **Boxplot**
Company Size vs. Salary(In USD)
- **Scatterplot**
Company Size vs. Salary(In USD)



2. 資料處理

Missing Value

```
> # check if there are missing values in data
> apply(df_original, 2, function(x) any(is.na(x))) # No missing value
      X      work_year   experience_level   employment_type      job_title
    FALSE        FALSE          FALSE          FALSE        FALSE
  salary  salary_currency  salary_in_usd employee_residence  remote_ratio
  FALSE        FALSE          FALSE          FALSE        FALSE
company_location      company_size
  FALSE        FALSE
```

Outliers

```
df_drop_salary_outliers <- df[!(df$salary_in_usd >= 400000 | df$salary_in_usd <= 10000),]

> df_FT[(df_FT$employee_residence != df_FT$company_location & df_FT$remote_ratio == 0),]
  work_year experience_level      job_title salary_in_usd employee_residence remote_ratio
183     2021              MI Data Engineer       26005                 RO            0
282     2021              EN Research Scientist 1000000                 JE            0
  company_location company_size
183           US             L
282           CN             L
```

Country

```
developed <- c('DE', 'JP', 'GB', 'US', 'HU', 'NZ', 'FR', 'PL', 'PT', 'GR', 'AE', 'NL', 'CA', 'AT',
             'ES', 'DK', 'HR', 'IT', 'SG', 'BE', 'RU', 'RO', 'SI', 'HK', 'TR', 'LU', 'CZ', 'MY',
             'EE', 'AU', 'IE', 'CH', 'IL') #33
developing <- c('HN', 'IN', 'PK', 'CN', 'MX', 'NG', 'PH', 'BG', 'IQ', 'VN', 'BR', 'UA', 'MT', 'CL',
               'IR', 'CO', 'MD', 'KE', 'RS', 'PR', 'JE', 'AR', 'BO', 'AS') #24
```

```
> pairwise.t.test(df_FT$salary_in_usd, df_FT$employee_residence, p.adj='bonferroni')
```

Pairwise comparisons using t tests with pooled SD

data: df_FT\$salary_in_usd and df_FT\$employee_residence

developed

developing 6.3e-16

P value adjustment method: bonferroni

```
> pairwise.t.test(df_FT$salary_in_usd, df_FT$company_location, p.adj='bonferroni')
```

Pairwise comparisons using t tests with pooled SD

data: df_FT\$salary_in_usd and df_FT\$company_location

developed

developing <2e-16

P value adjustment method: bonferroni

Job titles

```
> sort(table(df$job_title), decreasing = TRUE)
```

Data Scientist	143	Data Engineer	132	Data Analyst	97
Machine Learning Engineer	41	Research Scientist	16	Data Science Manager	12
Data Architect	11	Big Data Engineer	8	Machine Learning Scientist	8
AI Scientist	7	Data Analytics Manager	7	Data Science Consultant	7
Director of Data Science	7	Principal Data Scientist	7	BI Data Analyst	6
Computer Vision Engineer	6	Lead Data Engineer	6	ML Engineer	6
Applied Data Scientist	5	Business Data Analyst	5	Data Engineering Manager	5
Head of Data	5	Analytics Engineer	4	Applied Machine Learning Scientist	4
Data Analytics Engineer	4	Head of Data Science	4	Computer Vision Software Engineer	3
Data Science Engineer	3	Lead Data Analyst	3	Lead Data Scientist	3
Machine Learning Developer	3	Machine Learning Infrastructure Engineer	3	Principal Data Engineer	3
Cloud Data Engineer	2	Director of Data Engineering	2	ETL Developer	2
Financial Data Analyst	2	Principal Data Analyst	2	Product Data Analyst	2
3D Computer Vision Researcher	1	Big Data Architect	1	Data Analytics Lead	1
Data Specialist	1	Finance Data Analyst	1	Head of Machine Learning	1
Lead Machine Learning Engineer	1	Machine Learning Manager	1	Marketing Data Analyst	1
NLP Engineer	1	Staff Data Scientist	1		

```
Analyst <- c("Product Data Analyst", "Data Analyst", "Business Data Analyst",
           "Lead Data Analyst", "BI Data Analyst", "Marketing Data Analyst",
           "Data Analytics Manager", "Finance Data Analyst", "Principal Data Analyst",
           "Financial Data Analyst")
Scientist <- c("Data Scientist", "Lead Data Scientist", "Director of Data Science",
             "Research Scientist", "Data Science Consultant", "AI Scientist",
             "Principal Data Scientist", "Data Science Manager", "Head of Data",
             "Applied Data Scientist", "Head of Data Science", "Data Specialist")
Engineer <- c("Big Data Engineer", "Lead Data Engineer", "Data Engineer",
             "Data Engineering Manager", "Data Analytics Engineer", "Cloud Data Engineer",
             "Computer Vision Software Engineer", "Director of Data Engineering",
             "Data Science Engineer", "Principal Data Engineer", "Computer Vision Engineer",
             "Data Architect", "Big Data Architect", "Analytics Engineer", "ETL Developer",
             "NLP Engineer")
Machine_Learning <- c("Machine Learning Scientist", "Machine Learning Engineer",
                      "Machine Learning Manager", "Machine Learning Infrastructure Engineer",
                      "Machine Learning Developer", "Applied Machine Learning Scientist",
                      "ML Engineer", "Head of Machine Learning", "Lead Machine Learning Engineer")
```

```
> pairwise.t.test(df_FT$salary_in_usd, df_FT$job_title, p.adj='bonferroni')
```

Pairwise comparisons using t tests with pooled SD

data: df_FT\$salary_in_usd and df_FT\$job_title

	Analyst	Engineer	Machine_Learning
Engineer	0.0181	-	-
Machine_Learning	0.8062	1.0000	-
Scientist	0.0066	1.0000	1.0000

P value adjustment method: bonferroni

```
> pairwise.t.test(df_FT$salary_in_usd, df_FT$job_title, p.adj='bonferroni')
```

Pairwise comparisons using t tests with pooled SD

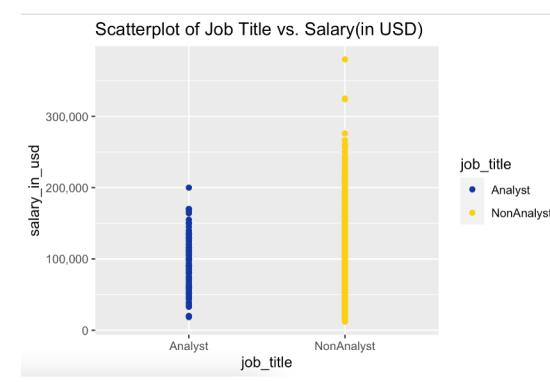
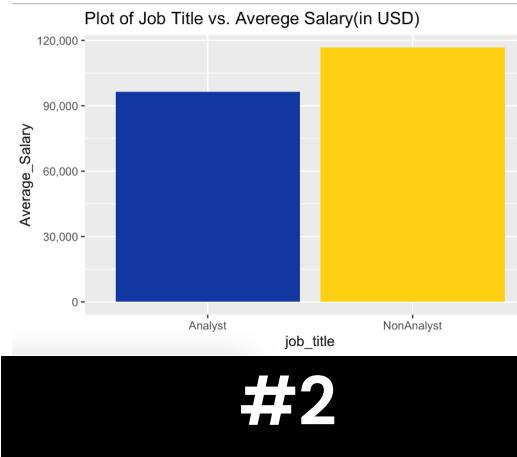
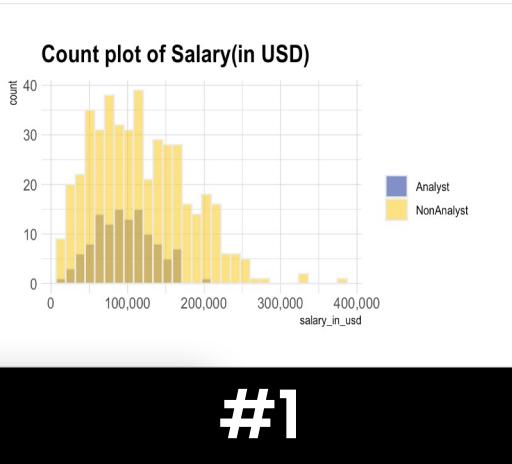
```
data: df_FT$salary_in_usd and df_FT$job_title
```

Analyst

NonAnalyst 0.00081

P value adjustment method: bonferroni

Visualization(Job Titles)



#3

Analyst

Company Size

NonAnalyst

Company Size vs. Salary(In USD)

Work Year

```
> aov(salary_in_usd~work_year, data = df_FT) %>% summary()
   Df    Sum Sq  Mean Sq F value    Pr(>F)
work_year     2 1.038e+11 5.190e+10   15.91 1.9e-07 ***
Residuals   564 1.840e+12 3.262e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # pairwise t test(bonferroni method) in work_year to see which pairs are different
> pairwise.t.test(df_FT$salary_in_usd, df_FT$work_year, p.adj='bonferroni')

  Pairwise comparisons using t tests with pooled SD

data: df_FT$salary_in_usd and df_FT$work_year

  2020   2021
2021 1.00000 -
2022 0.00073 1.8e-06

P value adjustment method: bonferroni
> # set work_year into two categories: "before 2022" and "2022"
> df_FT$work_year[df_FT$work_year != "2022"] <- "before 2022"
```

Remote ratio

```
> # remote_ratio
> # there exist a significant difference between 3 levels
> aov(salary_in_usd~remote_ratio, data = df_FT) %>% summary()
      Df   Sum Sq  Mean Sq F value    Pr(>F)
remote_ratio  2 1.079e+11 5.396e+10   16.58 1.01e-07 ***
Residuals    564 1.836e+12 3.255e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # pairwise t test(bonferroni method) in remote_ratio to see which pairs are different
> pairwise.t.test(df_FT$salary_in_usd, df_FT$remote_ratio, p.adj='bonferroni')

  Pairwise comparisons using t tests with pooled SD

data: df_FT$salary_in_usd and df_FT$remote_ratio

  0      100
100 0.1550 -
50  0.0026 5.3e-08

P value adjustment method: bonferroni
> # set remote_ratio into two categories: "partial" and "non-partial"
> df_FT$remote_ratio <- ifelse(df_FT$remote_ratio == "50", "partial", "non-partial")
```

Company Size

```
> pairwise.t.test(df_FT$salary_in_usd, df_FT$company_size, p.adj='bonferroni')

  Pairwise comparisons using t tests with pooled SD

data: df_FT$salary_in_usd and df_FT$company_size

L      M
M 0.75275 -
S 0.00048 4.7e-06

P value adjustment method: bonferroni
> # set company_size into two categories: "S"(small) and "L"(large) (Combining original M and L)
> df_FT$company_size[df_FT$company_size == "M"] <- "L"
|
```

Overview

rows = 607

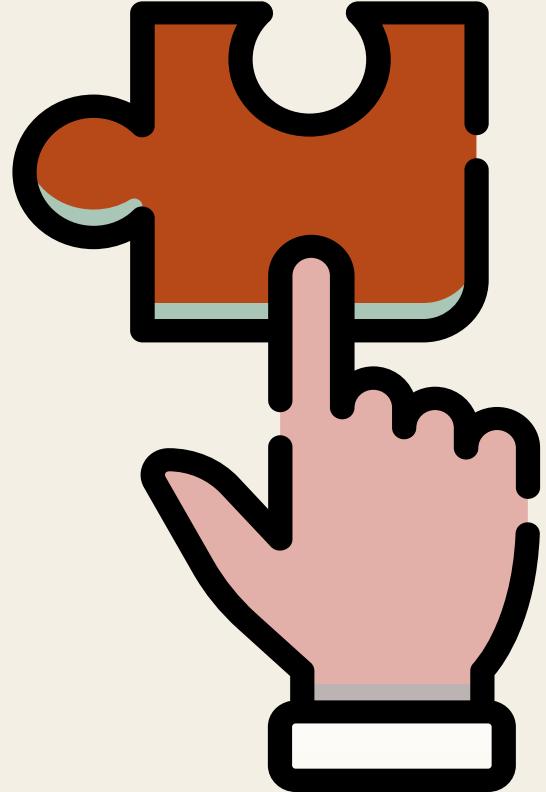
work_year	experience_level	employment_type	job_title	salary_in_usd	employee_reside
2020	MI	FT	Data Scientist	79833	DE
2020	SE	FT	Machine Learning Scientist	260000	JP
2020	SE	FT	Big Data Engineer	109024	GB
2020	MI	FT	Product Data Analyst	20000	HN
2020	SE	FT	Machine Learning Engineer	150000	US
2020	EN	FT	Data Analyst	72000	US



rows = 567

work_year	experience_level	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	comp
before 2022	MI	NonAnalyst	79833	developed	non-partial	developed	L
before 2022	SE	NonAnalyst	260000	developed	non-partial	developed	S
before 2022	SE	NonAnalyst	109024	developed	partial	developed	L
before 2022	MI	Analyst	20000	developing	non-partial	developing	S
before 2022	SE	NonAnalyst	150000	developed	partial	developed	L
before 2022	EN	Analyst	72000	developed	non-partial	developed	L

Dummy base : work_year_2022, experience_level_EN, job_title_Analyst, employee_residence_developing, remote_ratio_partial, company_location_developing, company_size_S



3. Linear Regression (training:testing = 8:2)

```
> summary(linear_all_model)
```

Call:

```
lm(formula = salary_in_usd ~ ., data = train_df_FT_dummy)
```

Residuals:

Min	1Q	Median	3Q	Max
-102535	-28910	-4264	22562	173775

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-39798.3	14447.6	-2.755	0.00612 **
`work_year_before 2022`	178.0	4959.0	0.036	0.97139
experience_level_EX	83180.1	12105.1	6.872	2.16e-11 ***
experience_level_MI	179.9	7123.2	0.025	0.97986
experience_level_SE	44851.8	7111.0	6.307	6.88e-10 ***
job_title_NonAnalyst	28646.4	5272.9	5.433	9.17e-08 ***
employee_residence_developed	27766.6	13357.5	2.079	0.03822 *
`remote_ratio_non-partial`	28011.5	6208.9	4.512	8.26e-06 ***
company_location_developed	44291.4	15799.9	2.803	0.00528 **
company_size_L	15141.1	6602.1	2.293	0.02229 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44950 on 443 degrees of freedom

Multiple R-squared: 0.3971, Adjusted R-squared: 0.3849

F-statistic: 32.42 on 9 and 443 DF, p-value: < 2.2e-16

```
> anova(linear_all_model)
Analysis of Variance Table
```

Response: salary_in_usd

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
`work_year_before 2022`	1	7.3776e+10	7.3776e+10	36.5187	3.210e-09 ***
experience_level_EX	1	7.6085e+10	7.6085e+10	37.6619	1.869e-09 ***
experience_level_MI	1	1.0345e+11	1.0345e+11	51.2053	3.473e-12 ***
experience_level_SE	1	1.2157e+11	1.2157e+11	60.1753	6.030e-14 ***
job_title_NonAnalyst	1	4.3370e+10	4.3370e+10	21.4680	4.738e-06 ***
employee_residence_developed	1	1.0768e+11	1.0768e+11	53.3009	1.337e-12 ***
`remote_ratio_non-partial`	1	3.9283e+10	3.9283e+10	19.4449	1.301e-05 ***
company_location_developed	1	1.3699e+10	1.3699e+10	6.7809	0.009523 **
company_size_L	1	1.0625e+10	1.0625e+10	5.2595	0.022294 *
Residuals	443	8.9496e+11	2.0202e+09		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
> vif(linear_all_model)
```

`work_year_before 2022`	experience_level_EX	experience_level_MI
1.372920	1.320321	2.627794
experience_level_SE	job_title_NonAnalyst	employee_residence_developed
2.826042	1.041232	3.074574
`remote_ratio_non-partial`	company_location_developed	company_size_L
1.181401	3.028380	1.123086

```
> linear_all_model_comparison
```

Best Subsets Regression

Model Index Predictors

1	experience_level_SE
2	experience_level_EX experience_level_SE
3	experience_level_EX experience_level_SE company_location_developed
4	experience_level_EX experience_level_SE job_title_NonAnalyst company_location_developed
5	experience_level_EX experience_level_SE job_title_NonAnalyst `remote_ratio_non-partial` company_location_developed
6	experience_level_EX experience_level_SE job_title_NonAnalyst `remote_ratio non-partial` company_location_developed company_size_L
7	experience_level_EX experience_level_SE job_title_NonAnalyst employee_residence_developed `remote_ratio_non-partial` company_location_developed company_size_L
8	`work_year_before_2022` experience_level_EX experience_level_SE job_title_NonAnalyst employee_residence_developed `remote_ratio_non-partial` company_location_developed company_size_L
9	`work_year_before_2022` experience_level_EX experience_level_MI experience_level_SE job_title_NonAnalyst employee_residence_developed `remote_ratio_non-partial` company_location_developed company_size_L

Subsets Regression Summary

Model	R-Square	Adj. R-Square		Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
		R-Square	R-Square									
1	0.1407	0.1388	0.1331	182.4306	11148.1903	9861.3420	11160.5380	1.281287e+12	2840936556.1735	6285442.0510	0.8669	
2	0.2375	0.2341	0.2224	113.2853	11096.0378	9809.2573	11112.5014	1.139453e+12	2531993754.4245	5602084.4115	0.7727	
3	0.3181	0.3136	0.3032	56.0589	11047.4276	9761.0835	11068.0070	1.021279e+12	2274362681.1041	5032267.3811	0.6940	
4	0.3486	0.3428	0.3325	35.6319	11028.6839	9742.5843	11053.3793	977749559906.4053	2182178243.6986	4828535.4298	0.6659	
5	0.3814	0.3745	0.3631	13.5745	11007.3227	9721.7228	11036.1339	930687127478.8279	2081667364.2321	4606403.6460	0.6352	
6	0.3912	0.3830	0.3704	8.3366	11002.0517	9716.6704	11034.9789	917926398218.9318	2057588144.5508	4553431.5070	0.6279	
7	0.3971	0.3876	0.3716	6.0018	10999.6407	9714.4425	11036.6838	9.11079e+11	2046668600.4813	4529620.6641	0.6246	
8	0.3971	0.3863	0.3682	8.0006	11001.6395	9716.4865	11042.7984	913133123971.1831	2055722238.5901	4550058.0169	0.6273	
9	0.3971	0.3849	0.3657	10.0000	11003.6389	9718.5310	11048.9137	915197717575.2903	2064819380.0287	4570639.9302	0.6301	

AIC: Akaike Information Criteria

SBIC: Sawa's Bayesian Information Criteria

SBC: Schwarz Bayesian Criteria

MSEP: Estimated error of prediction, assuming multivariate normality

FPE: Final Prediction Error

HSP: Hocking's Sp

APC: Amemiya Prediction Criteria

```
> # best model subset
> linear_all_model_comparison$predictors[7]
[1] "experience_level_EX experience_level_SE job_title_NonAnalyst employee_residence_developed `remote_ratio_non-partial` company_location_developed company_size_L"
```

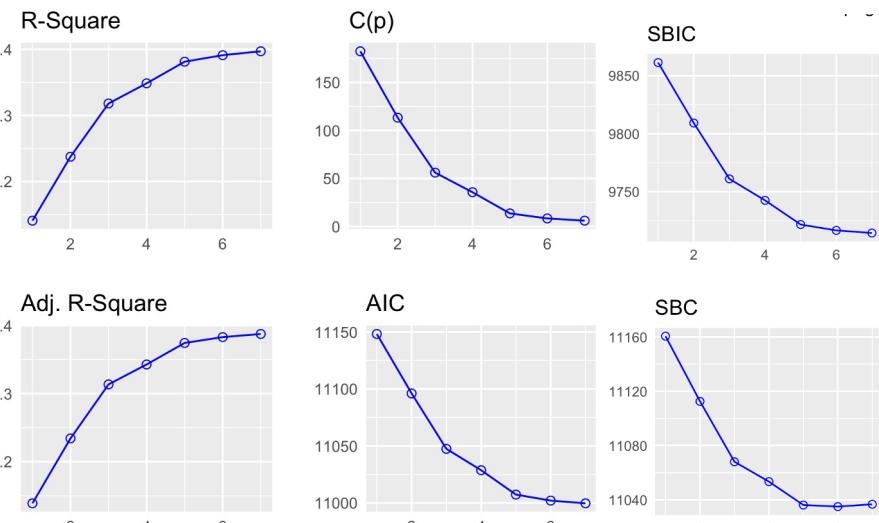
Discarded variables : work_year

```

> # stepwise selection
> bothFit.p <- ols_step_both_p(linear_all_model, prent = 0.15, prem = 0.15)
> bothFit.p

```

Stepwise Selection Summary							
Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	experience_level_SE	addition	0.141	0.139	182.4310	11148.1903	53183.1639
2	experience_level_EX	addition	0.238	0.234	113.2850	11096.0378	50153.1245
3	company_location_developed	addition	0.318	0.314	56.0590	11047.4276	47481.1098
4	job_title_NonAnalyst	addition	0.349	0.343	35.6320	11028.6839	46458.1030
5	`remote_ratio_non-partial`	addition	0.381	0.374	13.5750	11007.3227	45326.1076
6	company_size_L	addition	0.391	0.383	8.3370	11002.0517	45014.1869
7	employee_residence_developed	addition	0.397	0.388	6.0020	10999.6407	44845.8646



Best model variables :

- 1.experience_level
- 2.job_title
- 3.employee_residence
- 4.remote_ratio
- 5.company_location
- 6.company_size

Discarded variables

- 1.work_year

```
> summary(linear_model)
```

Call:
lm(formula = salary_in_usd ~ ., data = selected_train_df_FT_dummy)

Residuals:

Min	1Q	Median	3Q	Max
-102466	-28958	-4294	22539	173787

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-39547	12624	-3.133	0.00185 **
experience_level_EX	83148	12058	6.896	1.85e-11 ***
experience_level_MI	168	7107	0.024	0.98115
experience_level_SE	44811	7011	6.391	4.17e-10 ***
job_title_NonAnalyst	28665	5243	5.468	7.62e-08 ***
employee_residence_developed	27731	13305	2.084	0.03772 *
`remote_ratio_non-partial`	27934	5814	4.804	2.13e-06 ***
company_location_developed	44277	15777	2.806	0.00523 **
company_size_L	15088	6426	2.348	0.01931 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44900 on 444 degrees of freedom

Multiple R-squared: 0.3971, Adjusted R-squared: 0.3863

F-statistic: 36.56 on 8 and 444 DF, p-value: < 2.2e-16

```
> anova(linear_model)
```

Analysis of Variance Table

Response: salary_in_usd

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
experience_level_EX	1	7.9033e+10	7.9033e+10	39.2092	8.994e-10 ***
experience_level_MI	1	1.2343e+11	1.2343e+11	61.2371	3.737e-14 ***
experience_level_SE	1	1.5640e+11	1.5640e+11	77.5939	< 2.2e-16 ***
job_title_NonAnalyst	1	3.4962e+10	3.4962e+10	17.3450	3.746e-05 ***
employee_residence_developed	1	1.2273e+11	1.2273e+11	60.8859	4.371e-14 ***
`remote_ratio_non-partial`	1	4.8132e+10	4.8132e+10	23.8790	1.435e-06 ***
company_location_developed	1	1.3724e+10	1.3724e+10	6.8088	0.009377 **
company_size_L	1	1.1113e+10	1.1113e+10	5.5131	0.019312 *
Residuals	444	8.9496e+11	2.0157e+09		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:

```
lm(formula = salary_in_usd ~ .^2, data = selected_train_df_FT_dummy)
```

Residuals:

Min	1Q	Median	3Q	Max
-125959	-25128	-3128	20549	175887

Coefficients: (5 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	80235	50547	1.587	0.11319
experience_level_EX	-154756	75443	-2.051	0.04086 *
experience_level_MI	-5273	37881	-0.139	0.88935
experience_level_SE	-35328	37629	-0.939	0.34836
job_title_NonAnalyst	3915	43084	0.091	0.92765
employee_residence_developed	-40313	86207	-0.468	0.64030
`remote_ratio_non-partial`	-62548	33573	-1.863	0.06315 .
company_location_developed	39342	95380	0.412	0.68021
company_size_L	-43320	33154	-1.307	0.19205
experience_level_EX:experience_level_MI	NA	NA	NA	NA
experience_level_EX:experience_level_SE	NA	NA	NA	NA
experience_level_EX:job_title_NonAnalyst	69903	38210	1.829	0.06804 .
experience_level_EX:employee_residence_developed	35511	59268	0.599	0.54938
experience_level_EX:`remote_ratio_non-partial`	54020	36857	1.466	0.14348
experience_level_EX:company_location_developed	NA	NA	NA	NA
experience_level_EX:company_size_L	113635	36792	3.089	0.00214 **
experience_level_MI:experience_level_SE	NA	NA	NA	NA
experience_level_MI:job_title_NonAnalyst	-2164	19185	-0.113	0.91023
experience_level_MI:employee_residence_developed	-12992	42799	-0.304	0.76161
experience_level_MI:`remote_ratio_non-partial`	2714	16855	0.161	0.87213
experience_level_MI:company_location_developed	-6673	47698	-0.140	0.88880
experience_level_MI:company_size_L	33648	17000	1.979	0.04843 *
experience_level_SE:job_title_NonAnalyst	17660	19245	0.918	0.35933
experience_level_SE:employee_residence_developed	21130	42896	0.493	0.62257
experience_level_SE:`remote_ratio_non-partial`	36702	16882	2.174	0.03026 *
experience_level_SE:company_location_developed	7881	47849	0.165	0.86925
experience_level_SE:company_size_L	14436	17857	0.808	0.41930
job_title_NonAnalyst:employee_residence_developed	-5228	59793	-0.087	0.93037
job_title_NonAnalyst:`remote_ratio_non-partial`	2284	18471	0.124	0.90163
job_title_NonAnalyst:company_location_developed	10658	68048	0.157	0.87561
job_title_NonAnalyst:company_size_L	9427	21479	0.439	0.66097
employee_residence_developed:`remote_ratio_non-partial`	47225	42135	1.121	0.26301
employee_residence_developed:company_location_developed	NA	NA	NA	NA
employee_residence_developed:company_size_L	34573	31085	1.112	0.26668
`remote_ratio_non-partial`:company_location_developed	13715	48605	0.282	0.77794
`remote_ratio_non-partial`:company_size_L	17857	15907	1.123	0.26227
company_location_developed:company_size_L	-15177	37163	-0.408	0.68320

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44040 on 421 degrees of freedom

Multiple R-squared: 0.4499, Adjusted R-squared: 0.4094

F-statistic: 11.11 on 31 and 421 DF, p-value: < 2.2e-16

```

> summary(linear_model3)

Call:
lm(formula = salary_in_usd ~ . + experience_level_EX:company_size_L +
    experience_level_MI:company_size_L + experience_level_SE:`remote_ratio_non-partial`,
    data = selected_train_df_FT_dummy)

Residuals:
    Min      1Q  Median      3Q     Max 
-109697 -27426 -4179  23454 170656 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         -23744.5   13704.6  -1.733  0.08387 .  
experience_level_EX                  216.3     32507.6   0.007  0.99469  
experience_level_MI                 -14383.7   12685.2  -1.134  0.25745  
experience_level_SE                  24538.7   11872.7   2.067  0.03933 *  
job_title_NonAnalyst                28819.9    5188.7   5.554 4.82e-08 *** 
employee_residence_developed       25157.8   13211.3   1.904  0.05753 .  
`remote_ratio_non-partial`          18082.7    7215.5   2.506  0.01257 *  
company_location_developed         45431.7   15600.1   2.912  0.00377 ** 
company_size_L                      3678.7    8175.5   0.450  0.65295  
experience_level_EX:company_size_L  97053.8   34138.4   2.843  0.00468 ** 
experience_level_MI:company_size_L 20585.7   13240.7   1.555  0.12073  
experience_level_SE:`remote_ratio_non-partial` 27962.6  11971.4   2.336  0.01995 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44320 on 441 degrees of freedom
Multiple R-squared:  0.4165,    Adjusted R-squared:  0.4019 
F-statistic: 28.62 on 11 and 441 DF,  p-value: < 2.2e-16

```

Prediction validation (train/test set)

> validation

	MAPE	RMSE	MAE
linear_model3	0.4262145	45478.7	33008.41

```

> summary(cv_model)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-126675 -28065 -3414  23975 228325 

Coefficients:
                                         Estimate Std. Error t value
(Intercept)                         -29324    12201   -2.403
experience_level_EX                  7359     32429    0.227
experience_level_MI                 -7371     11481   -0.642
experience_level_SE                  28934     11191   2.586
job_title_NonAnalyst                29712     4673    6.358
employee_residence_developed       23924     12218   1.958
`\\`remote_ratio_non-partial\\` `    17727     6456    2.746
company_location_developed         44566     14214   3.135
company_size_L                      7614      7638   0.997
`experience_level_EX:company_size_L` 91625     33786   2.712
`experience_level_MI:company_size_L` 16855     12138   1.389
`experience_level_SE:\\`remote_ratio_non-partial\\` ` 28522     11172   2.553
                                         Pr(>|t|)    
(Intercept)                         0.01657 *  
experience_level_EX                  0.82056  
experience_level_MI                  0.52111  
experience_level_SE                  0.00998 ** 
job_title_NonAnalyst                4.28e-10 *** 
employee_residence_developed       0.05071 .  
`\\`remote_ratio_non-partial\\` `      0.00623 ** 
company_location_developed         0.00181 ** 
company_size_L                      0.31930  
`experience_level_EX:company_size_L` 0.00690 ** 
`experience_level_MI:company_size_L` 0.16551  
`experience_level_SE:\\`remote_ratio_non-partial\\` ` 0.01095 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 44510 on 555 degrees of freedom
 Multiple R-squared: 0.4343, Adjusted R-squared: 0.4231
 F-statistic: 38.73 on 11 and 555 DF, p-value: < 2.2e-16

Prediction validation(5 folds CV)

```
> print(cv_model)
Linear Regression

567 samples
 8 predictor

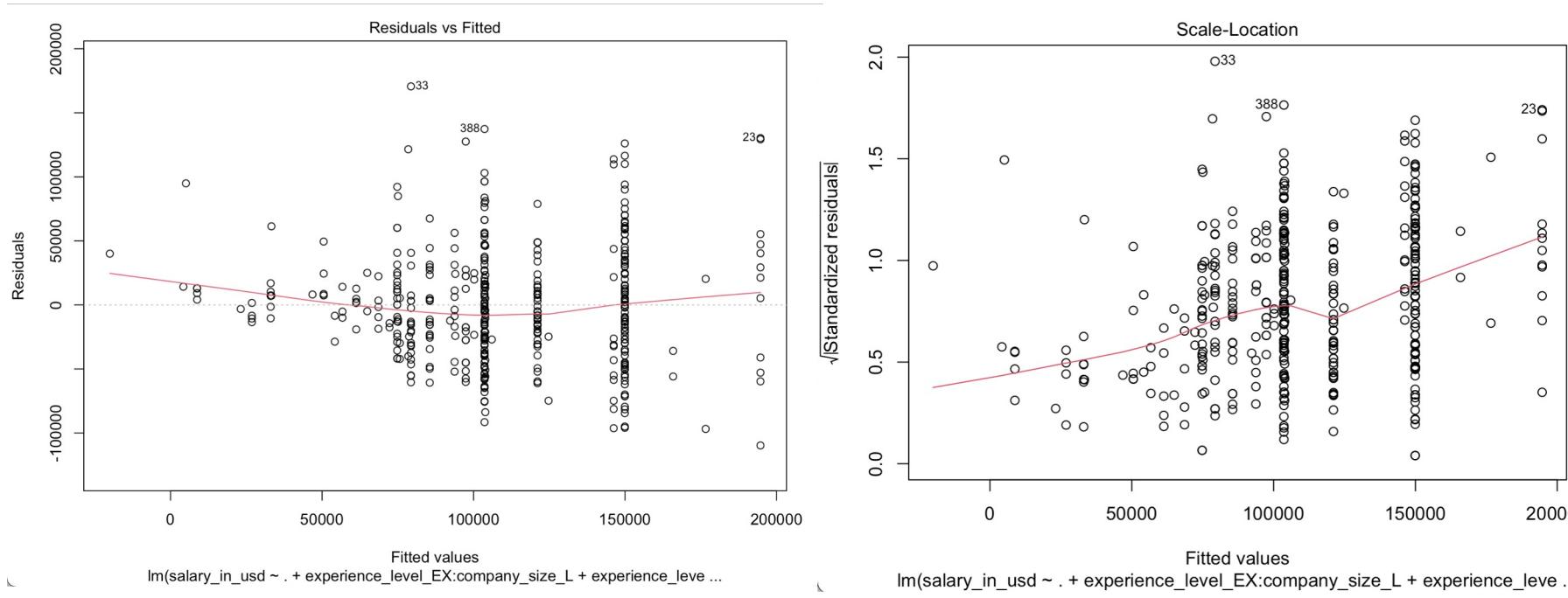
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 454, 454, 454, 453, 453
Resampling results:

  RMSE     Rsquared     MAE
44802.06  0.4194596  34122.3

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Prediction validation(5 folds CV)

Residual Analysis



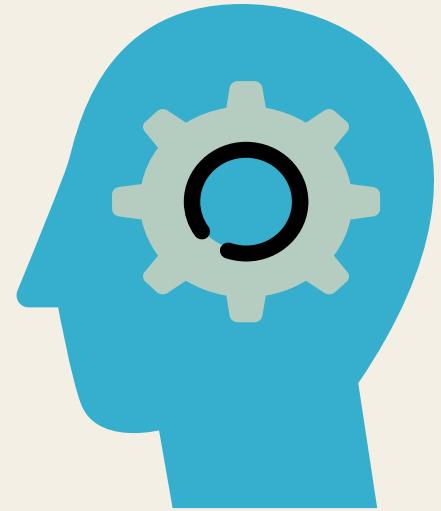
```

> # normality test
> shapiro.test(linear_model3$residuals) # alpha = 0.05
Shapiro-Wilk normality test

data: linear_model3$residuals
W = 0.9803, p-value = 8.052e-06

> # independence test
> durbinWatsonTest(linear_model3) # alpha = 0.05
lag Autocorrelation D-W Statistic p-value
1      0.02267397      1.953977  0.608
Alternative hypothesis: rho != 0
>
> # homogeneity test
> ncvTest(linear_model3) # alpha = 0.05
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 23.04468, Df = 1, p = 1.5828e-06

```



4. Logistic Regression (training:testing = 8:2)

```
> summary(df_FT_dummy$salary_in_usd)
   Min. 1st Qu. Median      Mean 3rd Qu.      Max.
12103    66144 105000    112462 150000    380000
```

salary_in_usd
79833
260000
109024
20000
150000
72000
190000
35735
135000
125000
51321
40481
39916
87000
85000
41689
114047
56000



salary_in_usd
0
1
1
0
1
0
1
0
1
1
0
0
0
0
0
1
0

Threshold = median

```
> summary(logi_all_model)
```

Call:

```
glm(formula = salary_in_usd ~ ., family = binomial, data = train_df_FT_dummy_classifi)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.03519	-0.90835	-0.00013	0.67768	2.10202

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-21.0741	678.5908	-0.031	0.975225
`work_year_before_2022`	-0.1484	0.2607	-0.569	0.569099
experience_level_EX	2.7567	0.7528	3.662	0.000250 ***
experience_level_MI	0.4333	0.4282	1.012	0.311565
experience_level_SE	2.1748	0.4249	5.118	3.09e-07 ***
job_title_NonAnalyst	0.9812	0.2819	3.481	0.000500 ***
employee_residence_developed	1.1062	0.7401	1.495	0.134997
`remote_ratio_non-partial`	1.2725	0.3643	3.493	0.000478 ***
company_location_developed	16.3148	678.5907	0.024	0.980819
company_size_L	0.7274	0.3815	1.907	0.056554 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 627.62 on 452 degrees of freedom

Residual deviance: 469.80 on 443 degrees of freedom

AIC: 489.8

Number of Fisher Scoring iterations: 16

```
Call: glm(formula = salary_in_usd ~ experience_level_SE + company_location_developed +
experience_level_EX + `remote_ratio_non-partial` + job_title_NonAnalyst +
company_size_L + employee_residence_developed + company_size_L:employee_residence_developed +
job_title_NonAnalyst:company_size_L + `remote_ratio_non-partial`:job_title_NonAnalyst,
family = "binomial", data = train_df_FT_dummy_classifi)
```

Coefficients:

	(Intercept)
	-50.3418
experience_level_SE	1.8329
company_location_developed	16.1378
experience_level_EX	2.4076
`remote_ratio_non-partial`	16.8529
job_title_NonAnalyst	32.3606
company_size_L	13.8527
employee_residence_developed	-0.7329
company_size_L:employee_residence_developed	2.9906
job_title_NonAnalyst:company_size_L	-15.9165
`remote_ratio_non-partial`:job_title_NonAnalyst	-15.5784

Degrees of Freedom: 452 Total (i.e. Null); 442 Residual
Null Deviance: 627.6
Residual Deviance: 463.6 AIC: 485.6

Stepwise (consider interaction)

Best model variables :

experience_level
job_title
employee_residence
remote_ratio, company_location
company_size

company_size_L : employee_residence_developed
job_title_NonAnalyst : company_size_L
remote_ratio_nonpartial : job_title_NonAnalyst

Discarded variables : work_year

```
> confusionMatrix(pred_tab)
Confusion Matrix and Statistics
```

Predicted		
Actual	0	1
0	40	13
1	11	50

Accuracy : 0.7895
95% CI : (0.7031, 0.8602)

No Information Rate : 0.5526
P-Value [Acc > NIR] : 1.102e-07

Kappa : 0.5758

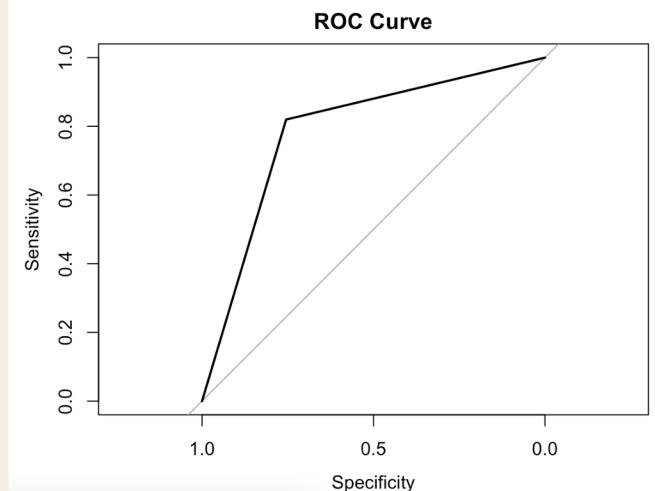
McNemar's Test P-Value : 0.8383

Sensitivity : 0.7843
Specificity : 0.7937
Pos Pred Value : 0.7547
Neg Pred Value : 0.8197
Prevalence : 0.4474
Detection Rate : 0.3509
Detection Prevalence : 0.4649
Balanced Accuracy : 0.7890

'Positive' Class : 0

```
> precision
[1] 0.7936508
> F1_Score
[1] 0.7889477
> # calculate area under curve
> auc(roc_object)
Area under the curve: 0.7872
```

Prediction validation (train/test set)



```
> summary(cv_logi_model)
```

Call:
NULL

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.30524	-0.80756	-0.00004	0.64130	2.14229

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-53.7756	2836.0655	-0.019	0.9849
experience_level_SE	2.4302	0.4063	5.981	2.22e-09 ***
experience_level_MI	0.6811	0.4147	1.642	0.1005
company_location_developed	17.2156	950.1917	0.018	0.9855
experience_level_EX	3.5375	0.7557	4.681	2.86e-06 ***
`\\`remote_ratio_non-partial\\`	17.7278	1906.4203	0.009	0.9926
job_title_NonAnalyst	34.1036	2672.1534	0.013	0.9898
company_size_L	15.2166	1872.4225	0.008	0.9935
employee_residence_developed	-0.7949	1.1228	-0.708	0.4790
`company_size_L:employee_residence_developed`	2.5472	1.4342	1.776	0.0757 .
`job_title_NonAnalyst:company_size_L`	-16.7009	1872.4224	-0.009	0.9929
`\\`remote_ratio_non-partial\\`:job_title_NonAnalyst`	-16.4925	1906.4203	-0.009	0.9931

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 785.98 on 566 degrees of freedom
Residual deviance: 561.92 on 555 degrees of freedom
AIC: 585.92

Number of Fisher Scoring iterations: 17

Prediction validation (5 folds CV)

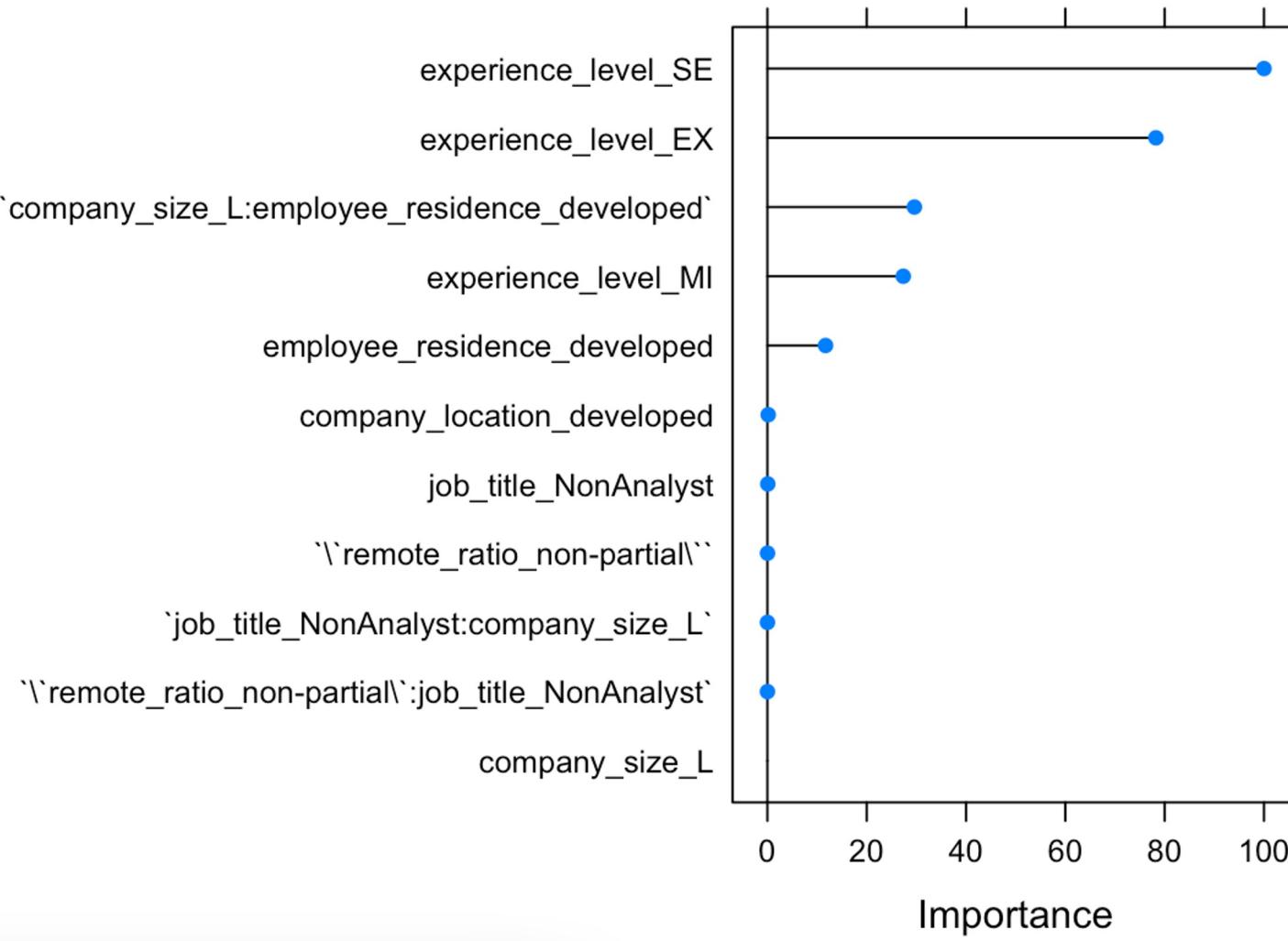
```
> print(cv_logi_model)
```

Generalized Linear Model

567 samples
8 predictor
2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 454, 453, 454, 454, 453
Resampling results:

Accuracy	Kappa
0.7458624	0.4918138



Important variables