6103 Data Mining
Final Project

How much salary will we earn in the future?

Team 2 | Brooklyn Chen, Nayaeun Kwon, Manojkumar Yerraguntla

Outline

- SMART Questions
- About Dataset
- Features
- Exploratory Data Analysis
- Modeling
- Conclusion

SMART Questions

1. Over time, salary by job title is changed?
2. How large is the wage disparity between gender?
3. Is there a significant difference between the yearly salary in top 10 companies and the yearly salary in not-top-10 companies?
4. People working for more years in the same company have higher salary?
5. In the recent four years, the average salary keeps the same?
6. How much does the employee’s experience affect his/her salary ?

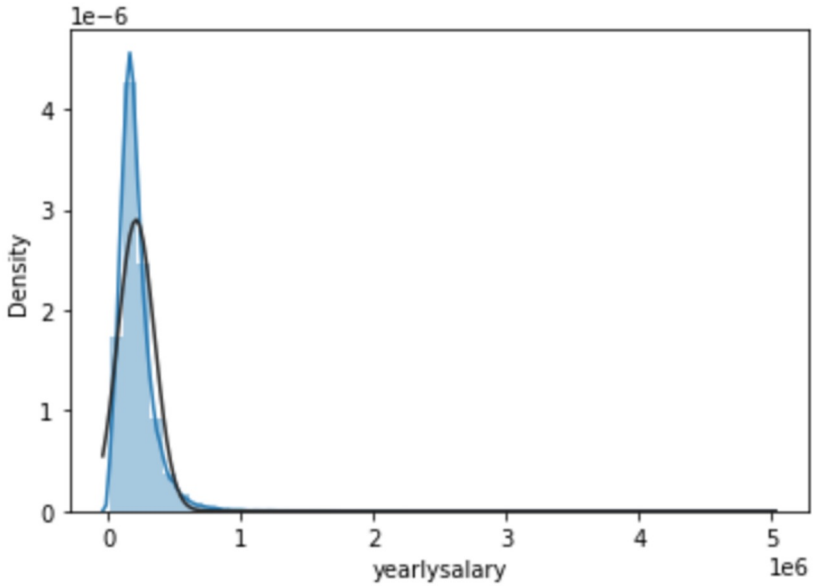
About Dataset

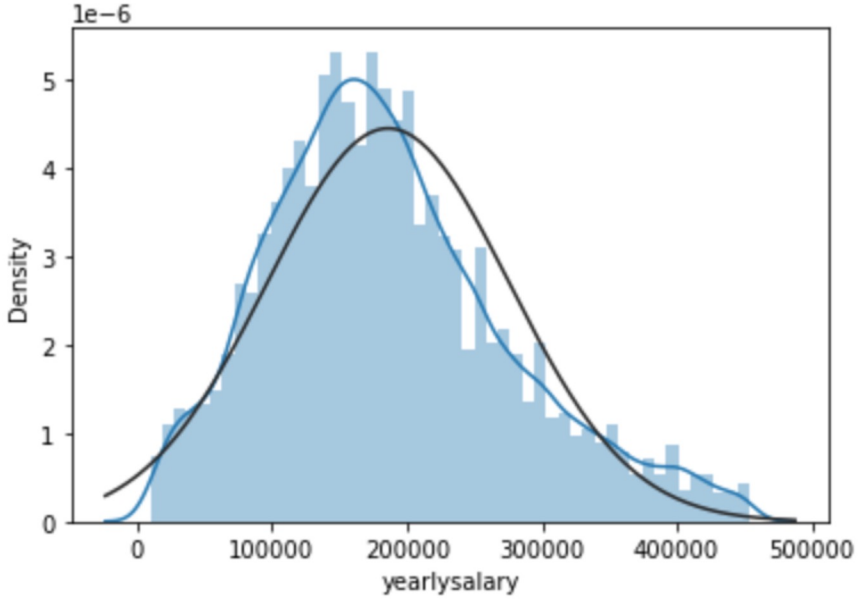
- From Kaggle
- Approximately 62,000 observations, after removing NA values and outliers, it was remained about 35,000 observations
- 29 variables, but left 8 variables:

  Timestamp, company, title, yearsofexperience, yearsatcompany, gender, yearlysalary.

About Dataset

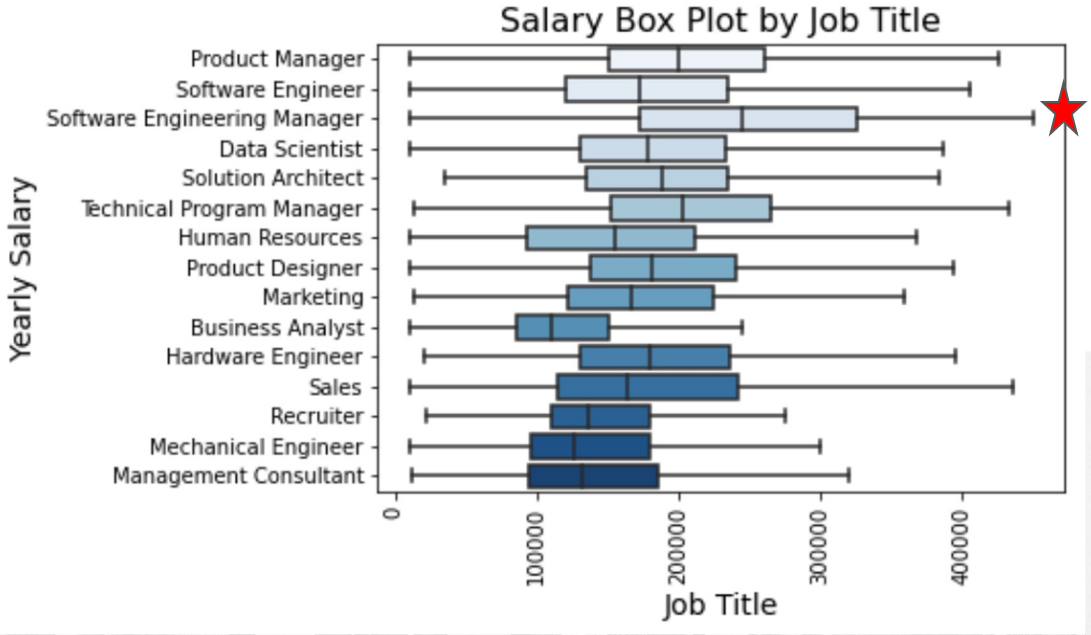
- Target Feature: **yearlysalary**





Question 1

Over time, salary by job title is changed?



Question 1

Over time, salary by job title is changed?

H_0: Year by year salary means of each job position are equal.
H_1: Year by year salary means of each job position are not equal.

Software Engineer

```
The p-value approach to hypothesis testing in the decision rule
F-score is: 155.70539235423766  and p value is: 1.1102230246251565e-16
Null Hypothesis is rejected.
```

Product Manager

```
The p-value approach to hypothesis testing in the decision rule
F-score is: 5.165934529555198  and p value is: 0.0014683665224930476
Null Hypothesis is rejected.
```

Software Engineering Manager

```
The p-value approach to hypothesis testing in the decision rule
F-score is: 5.192236014126059  and p value is: 0.0014497189297676405
Null Hypothesis is rejected.
```

Hardware Engineer

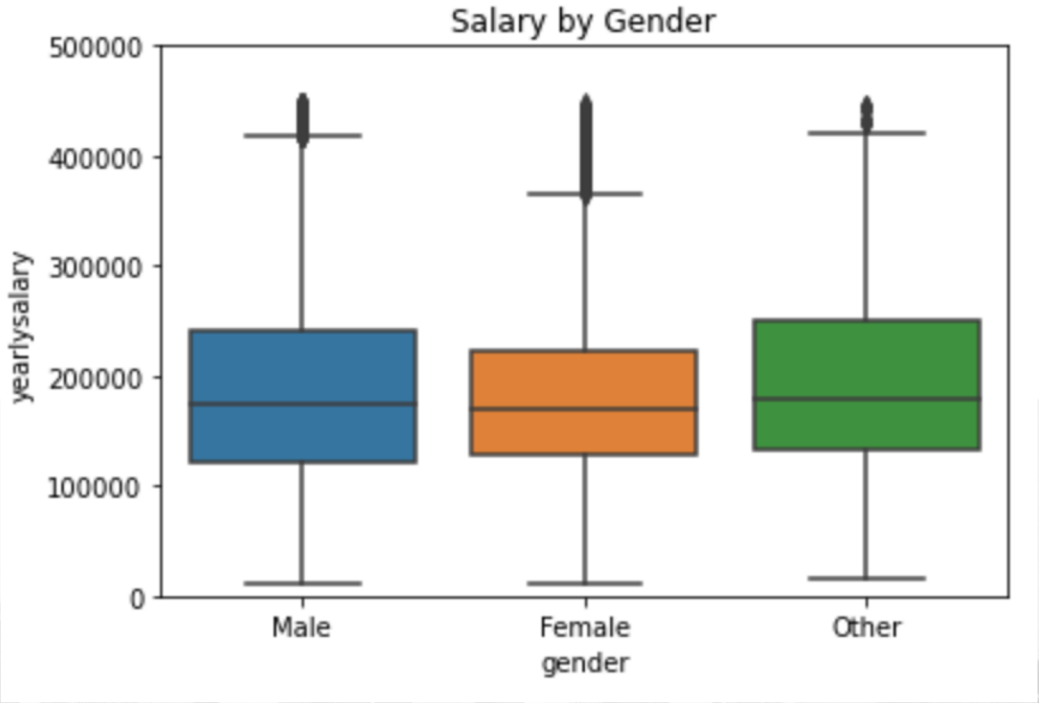
```
The p-value approach to hypothesis testing in the decision rule
F-score is: 7.983152783826113  and p value is: 2.8408647506306117e-05
Null Hypothesis is rejected.
```

Data Scientist

```
The p-value approach to hypothesis testing in the decision r
F-score is: 4.618955571726775  and p value is: 0.003188886046185724
Null Hypothesis is rejected.
```

Question 2

Is there the wage disparity between gender?



Question 2

Is there the wage disparity between gender?

ANOVA to figure out salary difference by gender difference
H_0: Average salary by gender are equal.
H_1: Average salary by gender are not equal.
α = 0.05

	SS	df	MS	F	P-value	F crit
Source of Variation						
Between Groups	163899537557.229	2	81949768778.615	10.577	0.0	3.689
Within Groups	278691537628696.812	35969	7748103578.879			
Total	278855437166254.031	35972	7752013709.726			

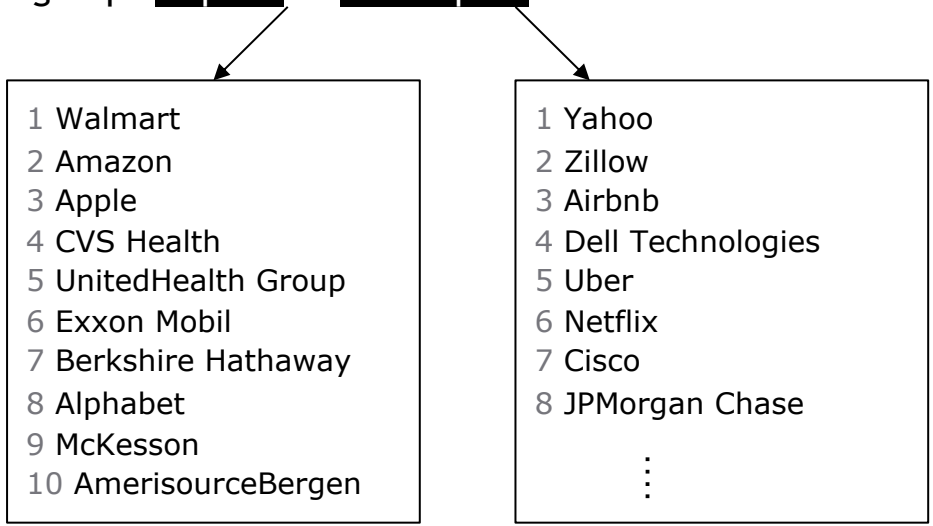
```
The p-value approach to hypothesis testing in the decision rule
F-score is: 10.576751839250923  and p value is: 2.5581453431011703e-05
Null Hypothesis is rejected.
```

Question 3

Is there a significant difference between the yearly salary in top 10 companies and the yearly salary in not-top-10 companies?

Question 3

- There are 1435 different companies in this dataset
- Based on the rank on Fortune 500, we divided the companies into two groups: top 10 vs not-top-10



Question 3


Boxplot of Yearly Salary
Grouped by Topcompany
Yearly Salary
400000
300000
200000
100000
0
Not-Top-10
Top 10
topcompany

Question 3

Independent-Sample T Test (2 groups)

H0: The means for the two groups are equal.

H1: The means for the two groups are not equal.

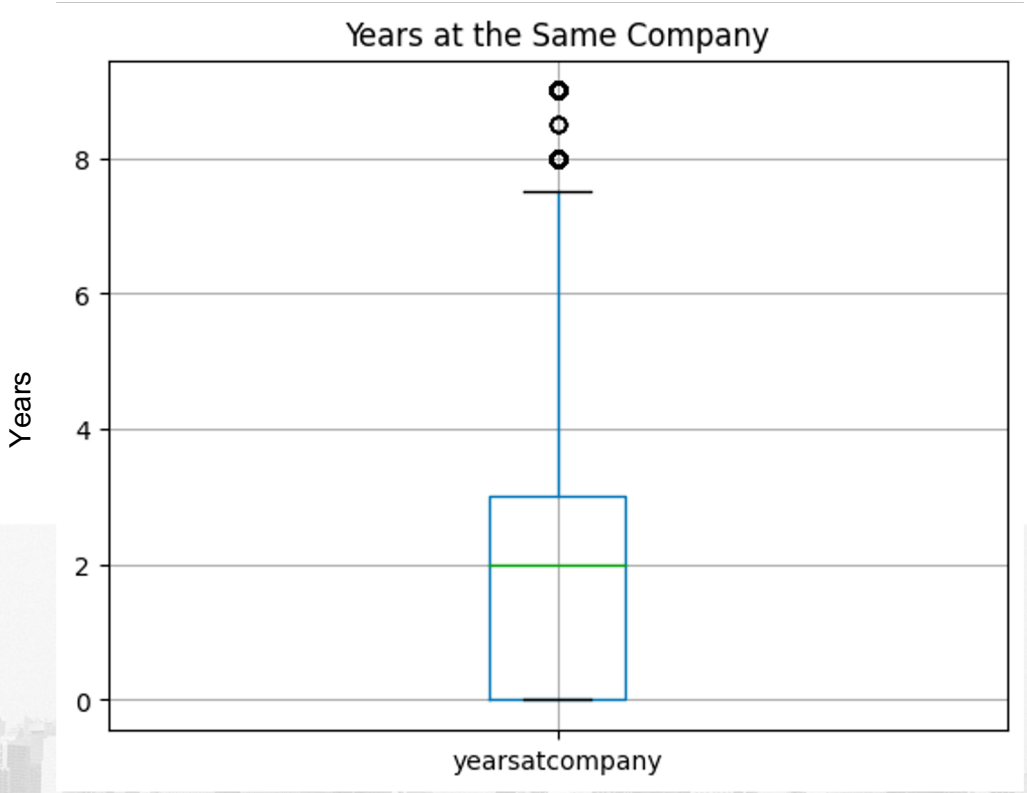
	Variable	N	Mean	SD	SE	95% Conf.	Interval
0	top10	6063.0	212370.938	80553.228	1034.521	210342.910	214398.967
1	not10	29910.0	179779.338	88483.127	511.626	178776.530	180782.147
2	combined	35973.0	185272.427	88045.521	464.215	184362.552	186182.302

	Independent t-test	results
0	Difference (top10 - not10) =	32591.600
1	Degrees of freedom =	35971.000
2	t =	26.538
3	Two side test p value =	0.000
4	Difference < 0 p value =	1.000
5	Difference > 0 p value =	0.000
6	Cohen's d =	0.374
7	Hedge's g =	0.374
8	Glass's delta1 =	0.405
9	Point-Biserial r =	0.139

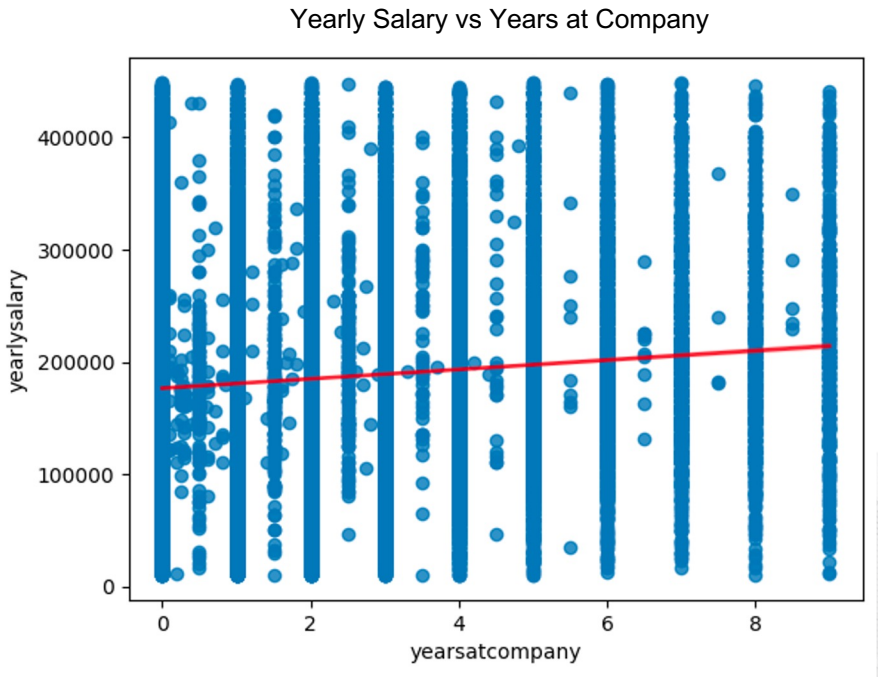
Question 4

People working for more years in the same company have higher salary?

Question 4



Question 4

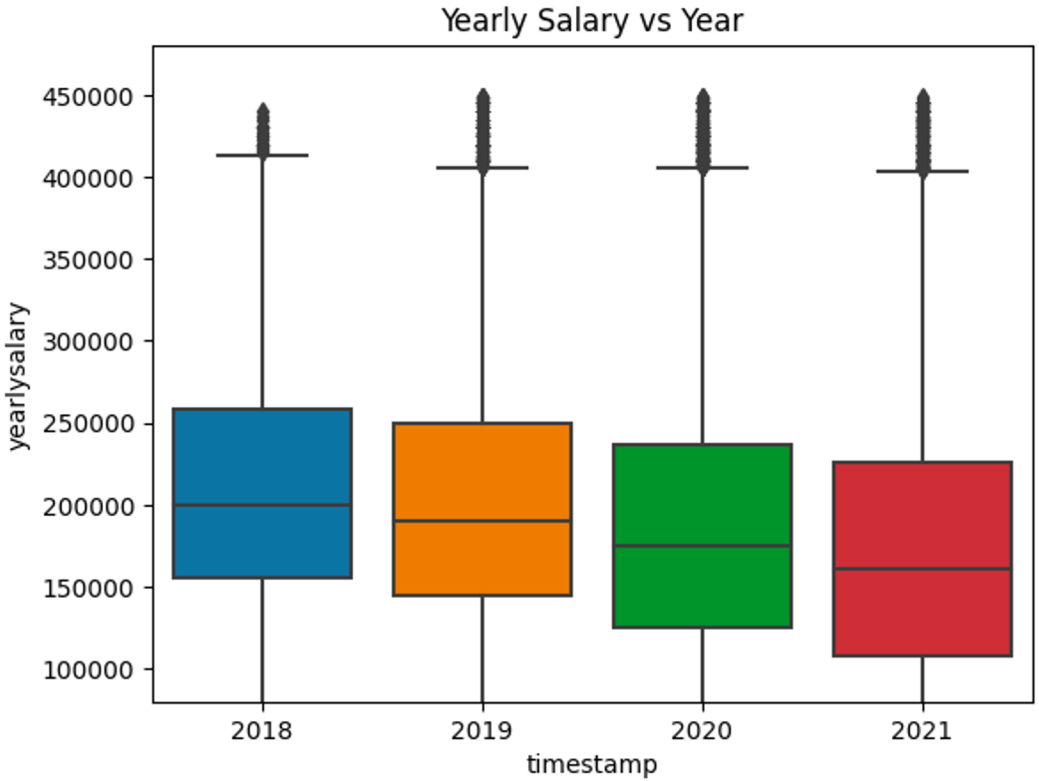



```
Pearsons correlation: 0.099
```

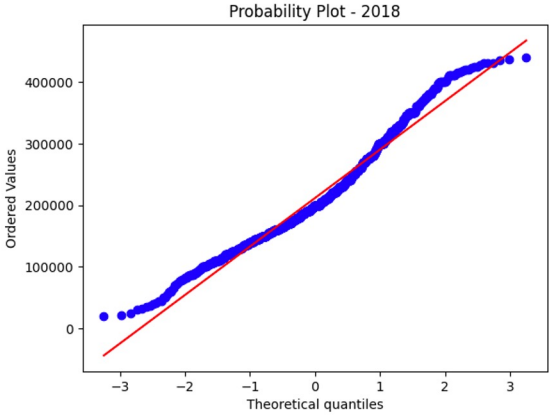
Question 5

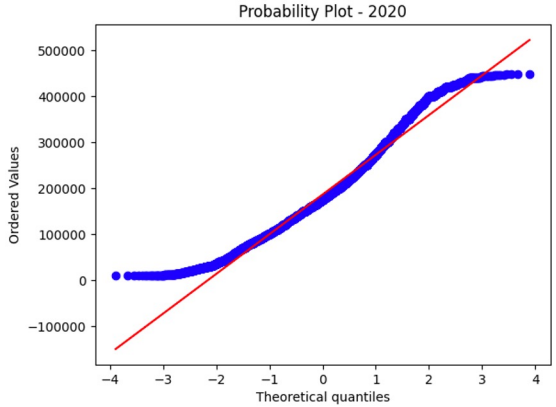
In the recent four years, the average salary keeps the same?

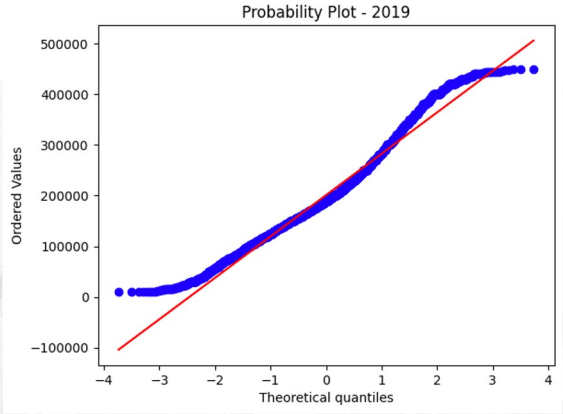
Question 5

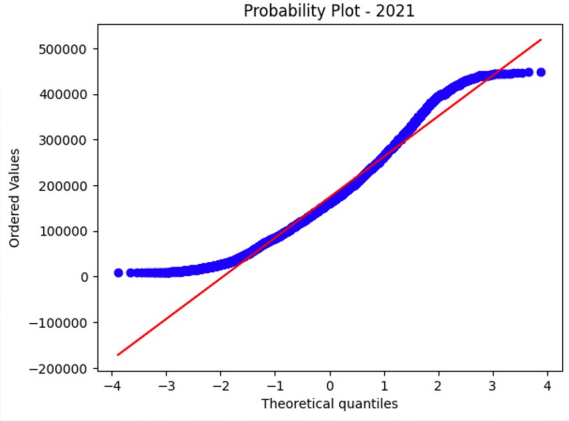

Yearly Salary vs Year
yearlysalary
450000
400000
350000
300000
250000
200000
150000
100000
2018
2019
2020
2021
timestamp

Question 5


Probability Plot - 2018
Ordered Values
Theoretical quantiles
400000
300000
200000
100000
0
−3
−2
−1
0
1
2
3


Probability Plot - 2020
Ordered Values
Theoretical quantiles
500000
400000
300000
200000
100000
0
−100000
−4
−3
−2
−1
0
1
2
3
4


Probability Plot - 2019
Ordered Values
Theoretical quantiles
500000
400000
300000
200000
100000
0
−100000
−4
−3
−2
−1
0
1
2
3
4


Probability Plot - 2021
Ordered Values
Theoretical quantiles
500000
400000
300000
200000
100000
0
−100000
−200000
−4
−3
−2
−1
0
1
2
3
4

Question 5

ANOVA Test

H_0: μ1= μ2 = μ3 = μ4 = μ5

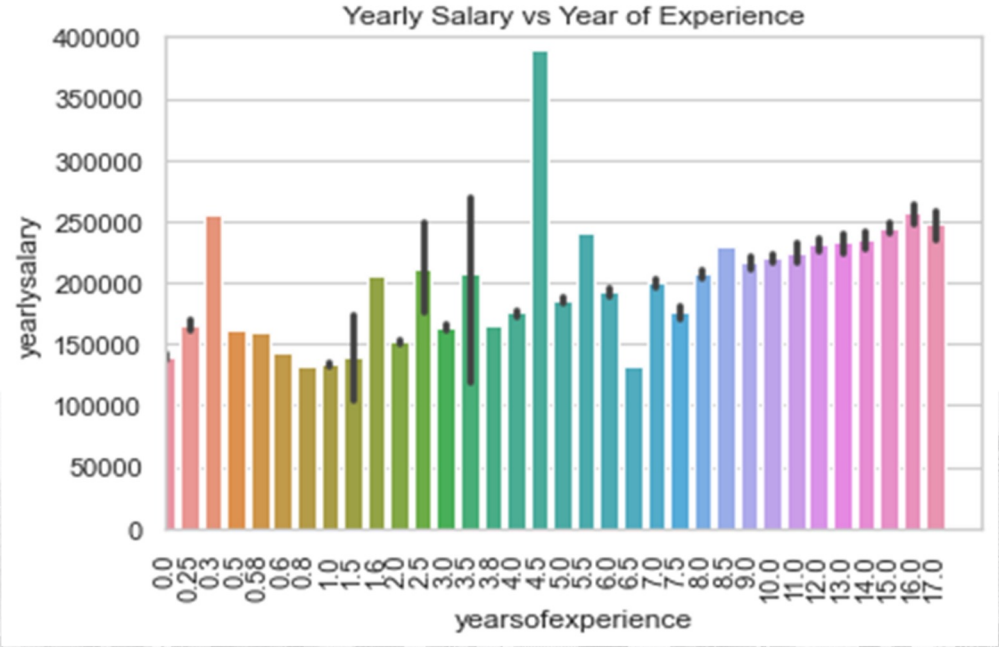
H_1: Not all yearly salary means are equal

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	4549359324285.643	3	1516453108095.214	198.848	0.0	3.116
Within Groups	274306077841968.0	35969	7626180261.947			
Total	278855437166253.656	35972	7752013709.726			

Null Hypothesis is rejected.

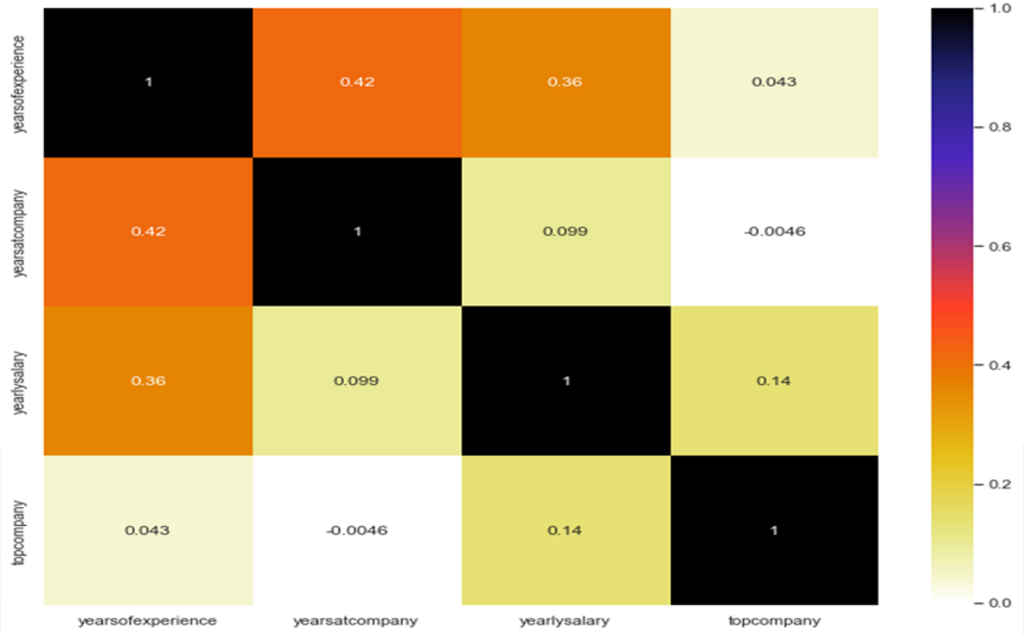
Question 6

How much does the employee’s experience affect his/her salary ?



Question 6

Pearson's Correlation:



Correlation of yearsatcompany is 0.099; Correlation of yearsofexpericence is 0.36

Additional Preprocessing for Modeling

Unique Values in each

```
timestamp : 4
company : 1435
title : 15
location : 881
yearsofexperience : 34
yearsatcompany : 52
gender : 3
yearlysalary : 451
topcompany : 2
```

Converting Multi Categorical Features into Dummy variables:

1. get_dummies() Function
2. One Hot Encoding

Shape of DataFrame after Creating Dummies:

```
✓ df_dummies2.shape ···

(35973, 68)
```

Train Test Split for Modeling

```
#Defining X and y Variables for Train and Test Split
X = df_dummies2.loc[:,['yearsofexperience','company_Amazon','company_Microsoft','company_Google',
'company_Facebook','company_Apple','company_Oracle','company_Salesforce','company_IBM','company_Intel',
'company_Cisco','company_CapitalOne','company_Uber','company_VMware','company_LinkedIn',
'company_JPMorganChase','company_GoldmanSachs','company_Qualcomm','company_Intuit','company_Bloomberg',
'company_PayPal','location_Seattle_WA','location_SanFrancisco_CA','location_NewYork_NY','location_Redmond_WA',
'location_Sunnyvale_CA','location_MountainView_CA','location_SanJose_CA','location_Austin_TX',
'location_Bangalore_KA_India','location_Cupertino_CA','location_MenloPark_CA','location_Boston_MA',
'location_London_EN_UnitedKingdom','location_SantaClara_CA','location_PaloAlto_CA','location_Chicago_IL',
'location_SanDiego_CA','location_Toronto_ON_Canada','location_Bellevue_WA','location_Bengaluru_KA_India',
'gender_Female','gender_Male','gender_Other','timestamp_2018','timestamp_2019','timestamp_2020',
'timestamp_2021','title_BusinessAnalyst','title_DataScientist','title_HardwareEngineer',
'title_HumanResources','title_ManagementConsultant','title_Marketing','title_MechanicalEngineer',
'title_ProductDesigner','title_ProductManager','title_Recruiter','title_Sales','title_SoftwareEngineer',
'title_SoftwareEngineeringManager','title_SolutionArchitect','title_TechnicalProgramManager',]]
y = df_dummies2[['yearlysalary']]
```

Splitting into Train and Test Subsets: 80:20

```
# Splitling the data into Train and test Data
X_train,X_test,y_train,y_test = train_test_split(X,y,train_size=0.8,random_state=100)
```

Linear Regression Model Results

OLS Regression Results

Dep. Variable:	yearlysalary	R-squared:	0.522
Model:	OLS	Adj. R-squared:	0.521
Method:	Least Squares	F-statistic:	653.9
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	0.00
Time:	20:05:43	Log-Likelihood:	-4.4734e+05
No. Observations:	35973	AIC:	8.948e+05
Df Residuals:	35912	BIC:	8.953e+05
Df Model:	60		
Covariance Type:	nonrobust		

```
Mean Squared Error of Linear Regression : 3720129792.96678
```

```
The Training R Square value is: 0.523208
The Testing R Square value is: 0.516748
```

Linear Regression Model Results

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.556e+04	964.190	47.257	0.000	4.37e+04	4.75e+04
yearsofexperience	7794.3770	82.331	94.671	0.000	7633.005	7955.749
company_Amazon	2.098e+04	1139.233	18.416	0.000	1.87e+04	2.32e+04
company_Microsoft	531.2986	1675.102	0.317	0.751	-2751.952	3814.549
company_Google	5.597e+04	1474.350	37.961	0.000	5.31e+04	5.89e+04
company_Facebook	7.835e+04	2094.699	37.404	0.000	7.42e+04	8.25e+04
company_Apple	4.209e+04	3070.410	13.710	0.000	3.61e+04	4.81e+04
company_Oracle	9606.8961	2565.426	3.745	0.000	4578.584	1.46e+04
company_Salesforce	1.67e+04	2665.495	6.267	0.000	1.15e+04	2.19e+04
company_IBM	-2.664e+04	2674.330	-9.962	0.000	-3.19e+04	-2.14e+04
company_Intel	-4936.1787	2890.221	-1.708	0.088	-1.06e+04	728.742
company_Cisco	-1.251e+04	2934.978	-4.263	0.000	-1.83e+04	-6758.412
company_CapitalOne	-3132.3086	2788.390	-1.123	0.261	-8597.636	2333.019
company_Uber	4.779e+04	2862.536	16.695	0.000	4.22e+04	5.34e+04
company_VMware	-3614.3919	3444.831	-1.049	0.294	-1.04e+04	3137.580
company_LinkedIn	5.267e+04	3387.475	15.550	0.000	4.6e+04	5.93e+04

Random Forest Regression Model Results

- Can be used for both Regression and classification problems
- Implemented using assemble method
- N_estimators Used: 100

R Square Values of Model, Training and Testing Datasets:

```
Mean Squared Error of Random Forest Regression:  2299991475.35531
R Square value of Random Forest Regression: 0.703296
The Training R Square value is: 0.471604
The Testing R Square value is: 0.462184
```

Comparison of Linear and Random Forest Models

Model Name	MSE Value	R Square Value
Linear Regression	3720129792.96678	0.51675
Random Forest Regression	2299991457.35531	0.70329

Random forest is the better model with 0.70329 R Square value.

Summary

1. The means salary of each job position is different from each other.
2. The average salary of gender is not equal
3. The average yearly salary in the US top 10 companies was higher than not-top-10 companies' average yearly salary
4. Working in the same company for more years does not mean those people have higher salaries
5. The means of yearly salary from 2018 to 2021 are not equal
6. After a period of time, the salary won't get affected much by yearsofexperience.
7. Random Forest Regression is the better model with high R2 Square when compared to the Linear Regression.

Thank You

Team 2 | Brooklyn Chen, Nayaeun Kwon, Manojkumar Yerraguntla