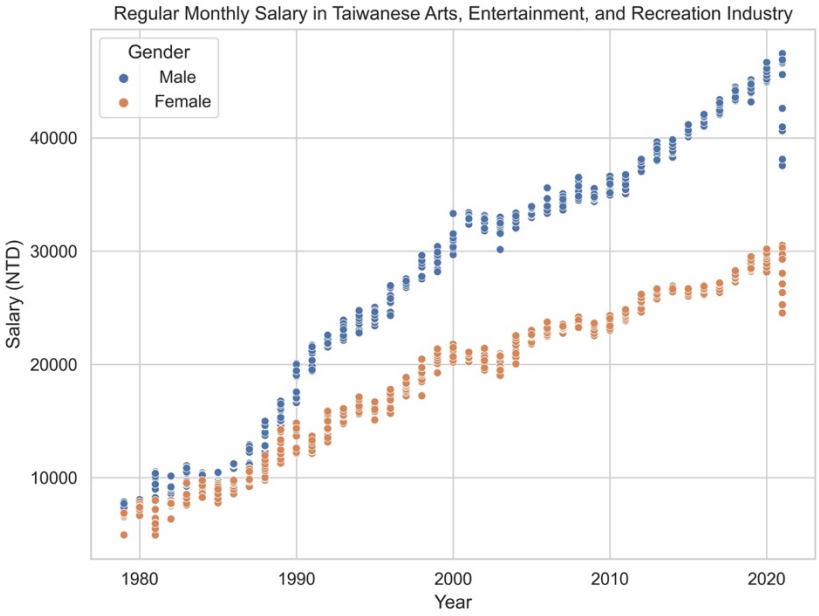
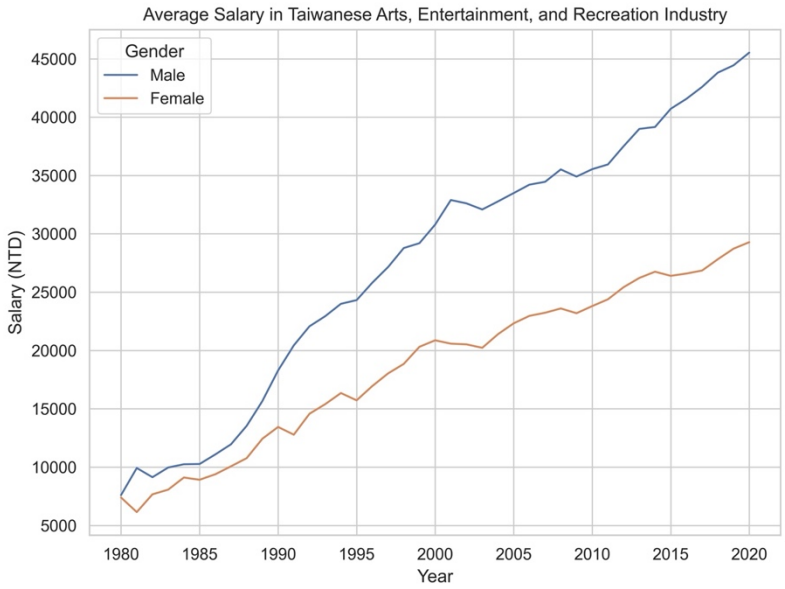
Intro to Data Science - Assignment 2

The analysis aims to explore possible causes of the gender pay gap in the Taiwanese Arts, Entertainment, and Recreation Industry*. The method is to apply data analysis techniques to explore the relationship on variables such as the time, number of employees, and monthly salary.

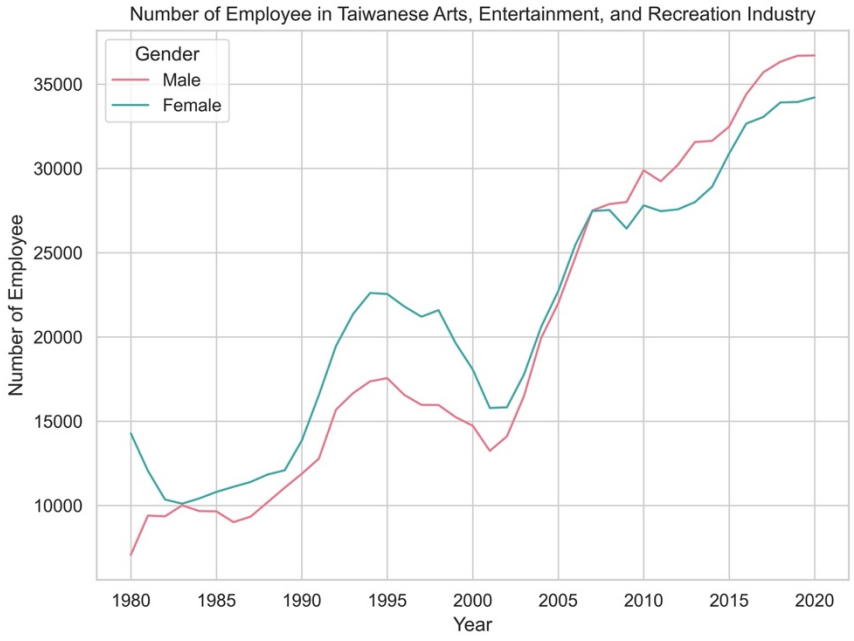
The original data were collected from Taiwan Executive Yuan's which is an open data source from the Taiwanese government. The database provides monthly income and the number of employees from 1980 to 2021.



Overall, the salary for both genders has an upward trend as the year progresses. There's no significant outlier in this dataset, except for a noticeable downfall during 2020 to 2021, which is most likely to be affected by the pandemic. Therefore, not much cleaning is needed for this dataset. To fit into the regression model, I decided to gain the means out of months as the representation of the average salary for each year. A new chart is shown below:

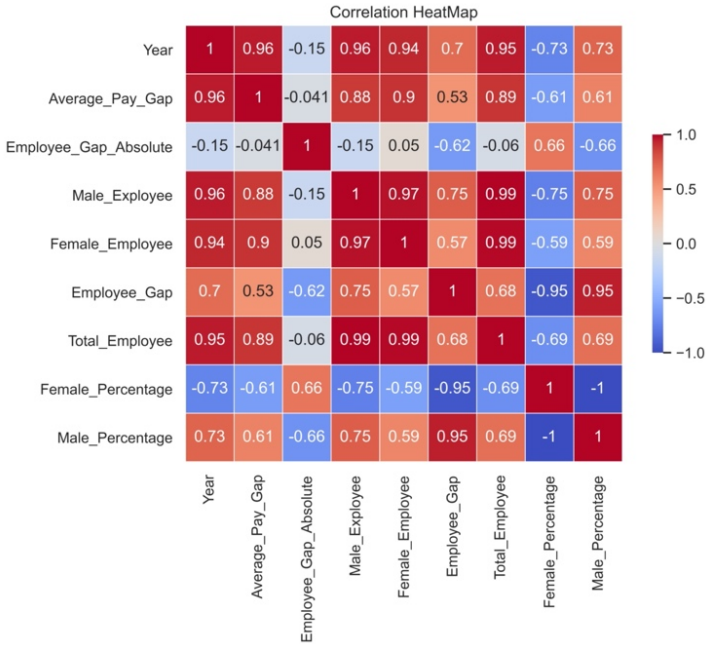


To add in variables, the number of employees in the creative industry were collected as well. The line chart indicates that from around 2007, male workers exceeded female workers in the industry, which could be an interesting point of exploring its historical background and its correlation with the pay gap.



2

From the data collected, multiple variables are created: Year, Average Salary, Average Pay Gap, Number of Male Employees, Number of Female Employees…etc. From the image of the heatmap below, I found the correlation between the number of female employees and the average pay gap is 0.89, indicating a positive correlation between the two. However, the reasons behind this correlation would



require further investigation of the historical background and demographic information such as age, marital status and job title.

3

I therefore applied linear regression with *the sklearn* package and *statsmodel's* OLS to test if the number of female employees in the industry would significantly predict the pay gap between genders. I would say the overall regression was statistically significant (R2 = 0.8) (P = 0.006). That means, from the diagram, I could guess that the two variables are related, but through simple linear regression, the claim is supported by numbers.

The prediction is that it's likely that when the female employee increases to 60,000 people, the predicted monthly pay gap could widen up to 30,626 NTD.

```
In [275]: #make prediction
          y_pred = regressor.predict([[60000]])
          print('predicted response:', y_pred, sep='\n')
```

```
predicted response:
[[30626.47125569]]
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.808
Model:	OLS	Adj. R-squared:	0.803
Method:	Least Squares	F-statistic:	164.1
Date:	Sat, 05 Mar 2022	Prob (F-statistic):	1.49e-15
Time:	17:45:04	Log-Likelihood:	-371.02
No. Observations:	41	AIC:	746.0
Df Residuals:	39	BIC:	749.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2875.7345	982.965	-2.926	0.006	-4863.970	-887.499
x1	0.5584	0.044	12.811	0.000	0.470	0.647

Omnibus:	18.898	Durbin-Watson:	0.390
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.115
Skew:	1.361	Prob(JB):	1.29e-06
Kurtosis:	5.909	Cond. No.	6.72e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.72e+04. This might indicate that there are strong multicollinearity or other numerical problems.

4

Multiple linear regression was used to test if the number of female employees and the labour gap between two genders taken significantly of the monthly pay gap. According to the overall regression was statistically significant (R2 = 0.809, p = 0.049). It was found that the gap of employees taken did not significantly predict the pay gap (p = 0.751). The result also implied that there's a strong multicollinearity of the variables which means the variables are correlated, thus is less reliable as inputs of a regression model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Average_Pay_Gap   R-squared:                       0.809
Model:                            OLS   Adj. R-squared:                  0.798
Method:                 Least Squares   F-statistic:                     80.22
Date:                Sat, 05 Mar 2022   Prob (F-statistic):           2.29e-14
Time:                        18:37:46   Log-Likelihood:                -370.96
No. Observations:                  41   AIC:                             747.9
Df Residuals:                      38   BIC:                             753.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -2616.2037   1283.833     -2.038      0.049   -5215.187     -17.221
Female_Employee     0.5487      0.054     10.243      0.000       0.440       0.657
Employee_Gap        0.0460      0.144      0.320      0.751      -0.245       0.337
==============================================================================
Omnibus:                       19.056   Durbin-Watson:                   0.375
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.673
Skew:                           1.399   Prob(JB):                     1.61e-06
Kurtosis:                       5.790   Cond. No.                     8.68e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.68e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
coefficient of determination: 0.8085145950389272
intercept: -2616.2037027179776
slope: [0.54865133 0.04596768]
```

```
predicted_pay_gap = regressor.predict([[45000, 30000]])
print(predicted_pay_gap)
```

```
[23452.13651265]
```

5

Still, the simple regression model could help predict the trend of pay gap which might help the researchers to monitor and further investigate the possible cause of the phenomena.

The data collected by the government are imperfect: gender is divided in binary, therefore caused lack of representation. It’s difficult to interpret the data because there is no further information of why the pay gap widen as women joined the workforce. Other variable, for instance, age, education level and job title would help analyse create a better model that would also be more interpretable for the result of the predictions.

Data Reference : Directorate-General of Budget, Accounting and Statistics, Executive Yuan, R.O.C.(Taiwan) https://eng.dgbas.gov.tw/mp.asp?mp=2 (last access: 21/03/2022)