Basic Analysis of Salaries of CS University Graduates in Singapore

Wesley Tan Yan Long

INTRODUCTION

This report is for me to practice what I have learned in coding and showcase how I have used **Python**, **Pandas**, and **Numpy** to analyze the salaries of computer science students who have graduated from computer science universities in Singapore.

This report contains:

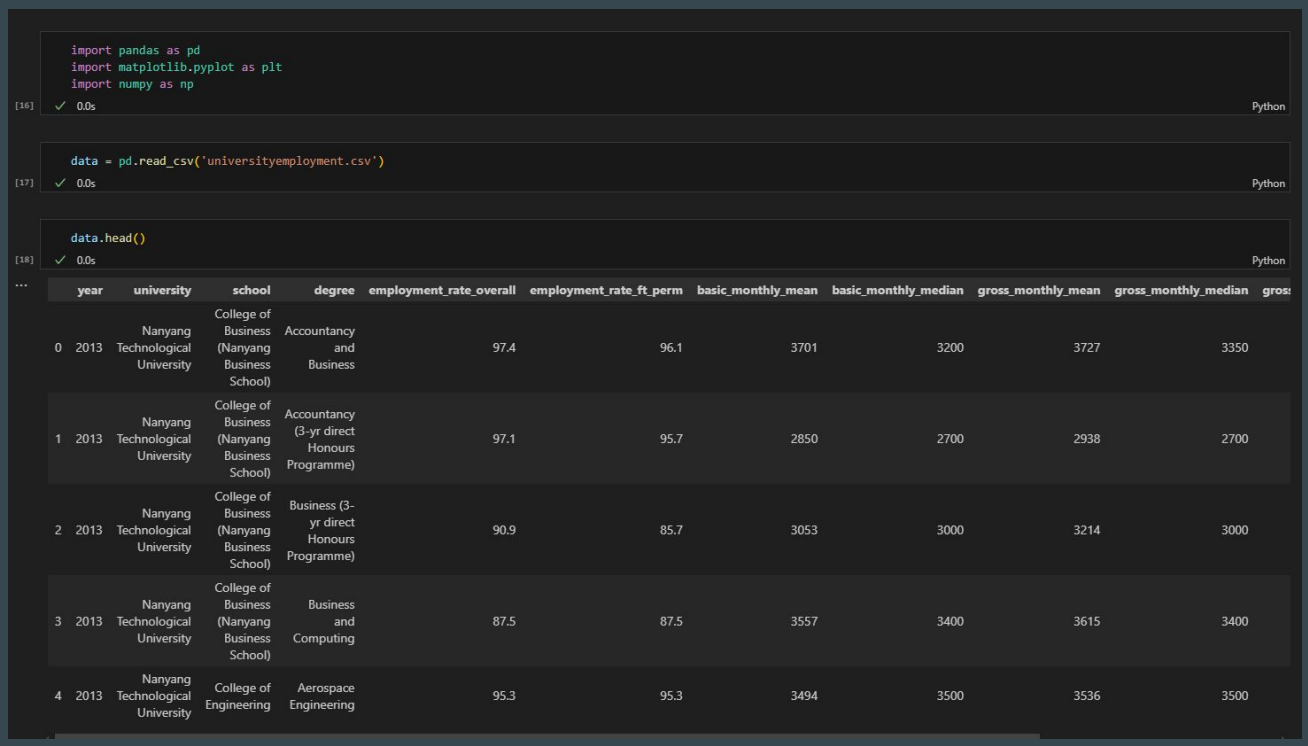
1. Obtaining the dataset
2. Cleaning of the dataset
3. Using the dataset for analysis
4. Prediction of variables
5. Conclusion

All of the coding will be done using Visual Studio Code.

1. Obtaining of dataset

After searching the web to find a suitable dataset for analysis, I found a official government website ('www.data.gov.sg'), that provides open source data that allow users to do research studies.

I will be navigating to the education tab to obtain the dataset for analysis.



```
[16] ✓ 0.0s
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
Python

[17] ✓ 0.0s
data = pd.read_csv('universityemployment.csv')
Python

[18] ✓ 0.0s
data.head()
Python
```

	year	university	school	degree	employment_rate_overall	employment_rate_ft_perm	basic_monthly_mean	basic_monthly_median	gross_monthly_mean	gross_monthly_median	gros
0	2013	Nanyang Technological University	College of Business (Nanyang Business School)	Accountancy and Business	97.4	96.1	3701	3200	3727	3350	
1	2013	Nanyang Technological University	College of Business (Nanyang Business School)	Accountancy (3-yr direct Honours Programme)	97.1	95.7	2850	2700	2938	2700	
2	2013	Nanyang Technological University	College of Business (Nanyang Business School)	Business (3-yr direct Honours Programme)	90.9	85.7	3053	3000	3214	3000	
3	2013	Nanyang Technological University	College of Business (Nanyang Business School)	Business and Computing	87.5	87.5	3557	3400	3615	3400	
4	2013	Nanyang Technological University	College of Engineering	Aerospace Engineering	95.3	95.3	3494	3500	3536	3500	

Fig 1.1

After downloading the dataset, I headed to Visual Studio Code to import the dataset. (shown in fig 1.1)

Python will be the main language used for this project as well as I have imported all the necessary python libraries (Pandas, Matplotlib and Numpy) to assist me.

2. Cleaning of dataset

For personal taste, I do not need the full spelling of the universities so I will be renaming all the universities to their acronyms for easier viewing.

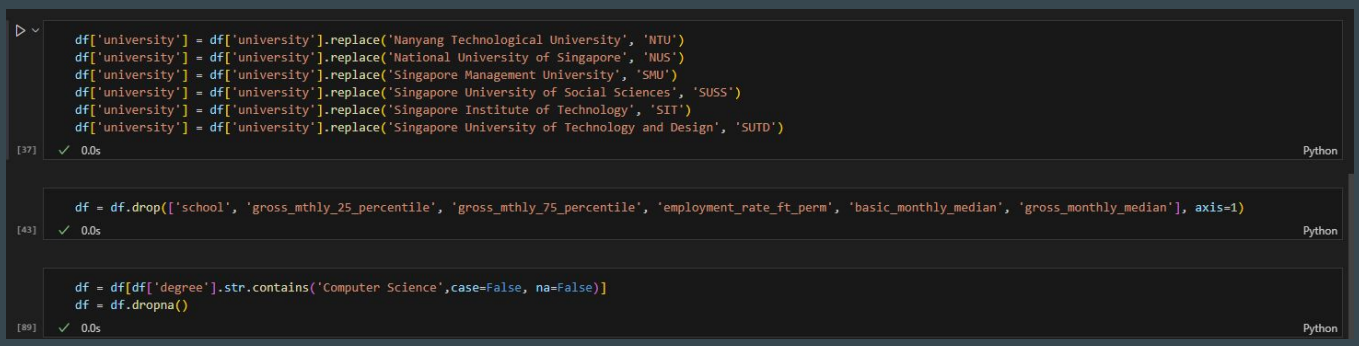
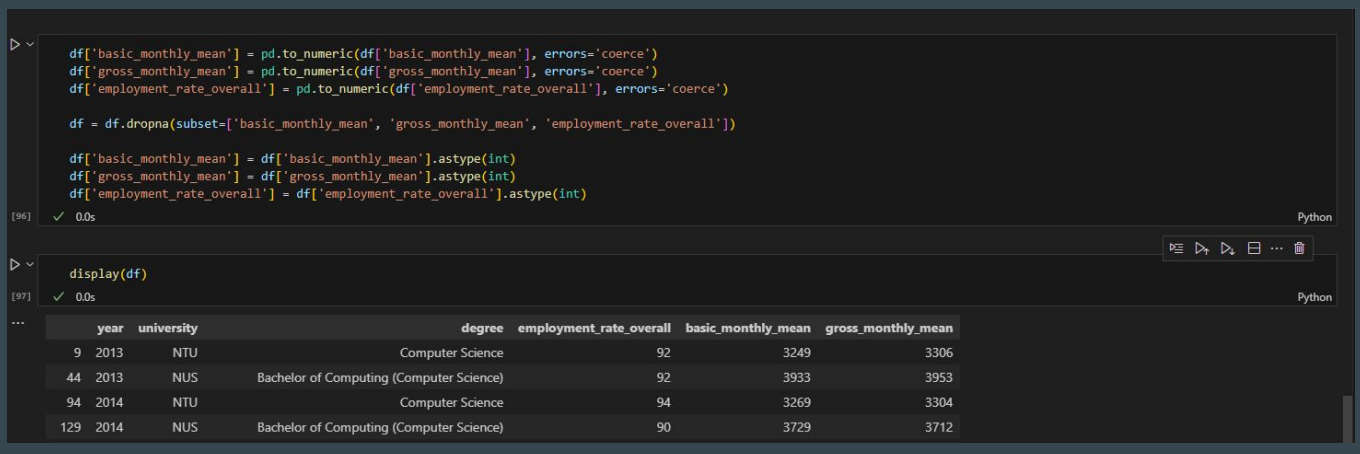


Fig 2.1

As the dataset contain many informations, I will be dropping column that is not needed using the .drop() syntax.

At the same time, I will also be filtering out the degree column to only show computer science related degree. (shown in fig 2.1)



	year	university	degree	employment_rate_overall	basic_monthly_mean	gross_monthly_mean
9	2013	NTU	Computer Science	92	3249	3306
44	2013	NUS	Bachelor of Computing (Computer Science)	92	3933	3953
94	2014	NTU	Computer Science	94	3269	3304
129	2014	NUS	Bachelor of Computing (Computer Science)	90	3729	3712

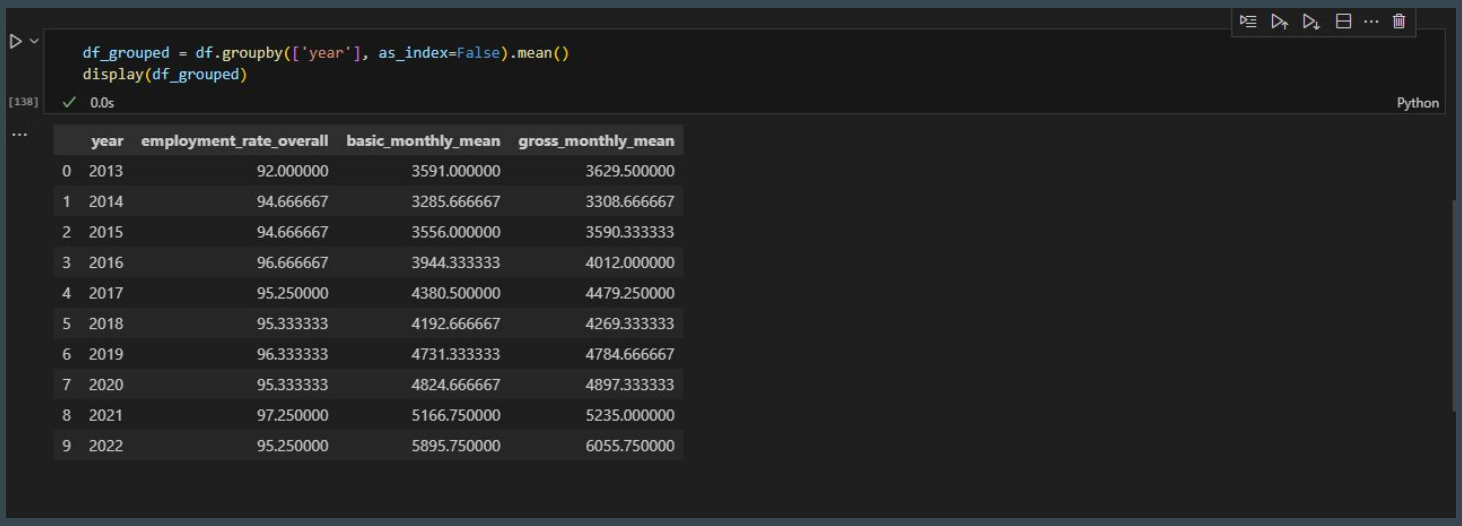
Fig 2.2

Now, the dataset I imported is saved as a string even though what we see is integer. I will need to convert the necessary columns to integer for calculation later on. (shown in fig 2.2)

3. Using the dataset for analysis

My next step is to group all the universities by year, this would mean the values in the other column will need to combine as well.

To tackle this i will be grouping the years together using .groupby() syntax as well as combining the values and finding the mean of it using .mean() syntax. (shown in fig 3.1)



```
df_grouped = df.groupby(['year'], as_index=False).mean()
display(df_grouped)
```

[138] ✓ 0.0s Python

	year	employment_rate_overall	basic_monthly_mean	gross_monthly_mean
0	2013	92.000000	3591.000000	3629.500000
1	2014	94.666667	3285.666667	3308.666667
2	2015	94.666667	3556.000000	3590.333333
3	2016	96.666667	3944.333333	4012.000000
4	2017	95.250000	4380.500000	4479.250000
5	2018	95.333333	4192.666667	4269.333333
6	2019	96.333333	4731.333333	4784.666667
7	2020	95.333333	4824.666667	4897.333333
8	2021	97.250000	5166.750000	5235.000000
9	2022	95.250000	5895.750000	6055.750000

Fig 3.1

Do note that instead of grouping the universities together, I could have filter out the specific university I want to target instead and analyse the salaries of computer science graduates from specific university.

But in this case, I chose the grouping method for my analysis.

3. Using the dataset for analysis

Once I have combine the universities by year, we will be plotting my graph is X as my independent variable (year) and Y as my dependent variable (gross monthly mean). This will allow me to see the correlation between the two variable for the analysis. I will be using Matplotlib to assist me with the graph.

I will also be adding linear regression line for me to predict the future variable. I will be using Numpy for helping me with calculating of the regression line.

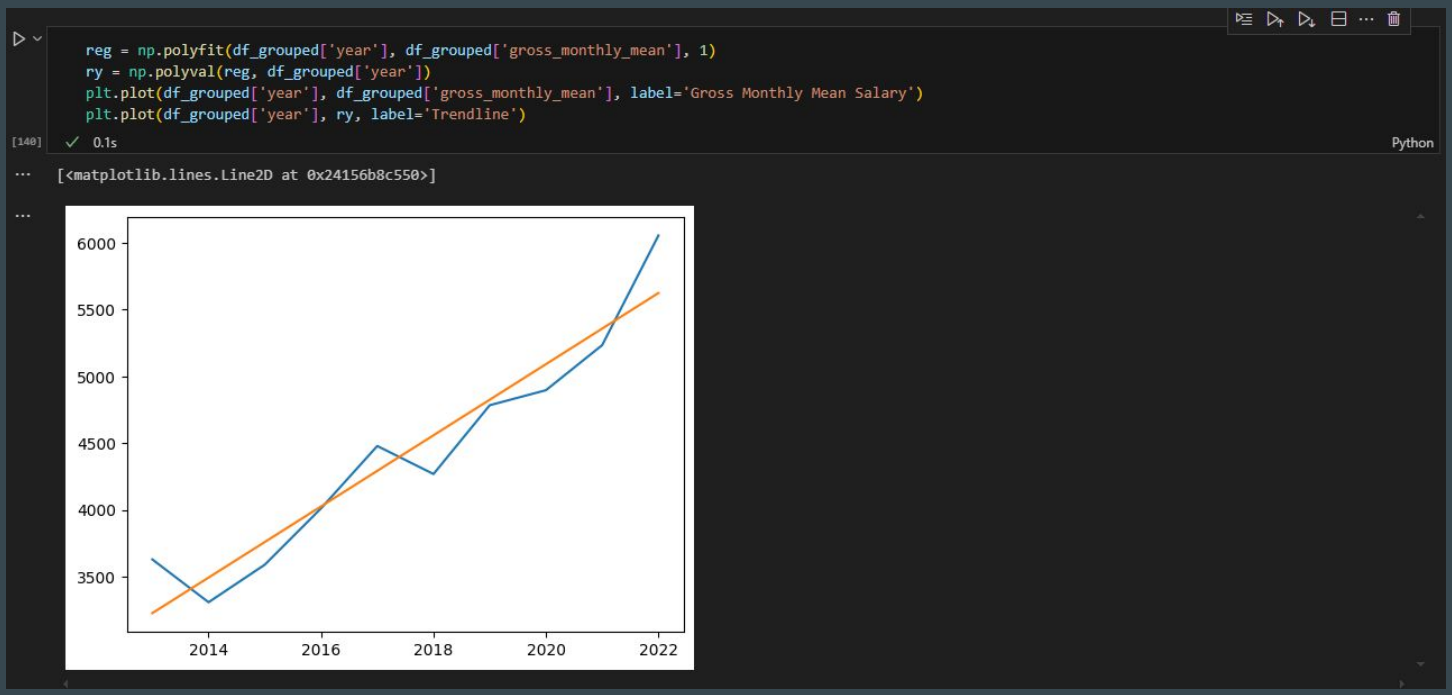


Fig 3.2

This is the output I received, for the blue line to be the graph of year against gross monthly mean, and orange line being the linear regression line. (refer to fig 3.2)

4. Prediction of variables

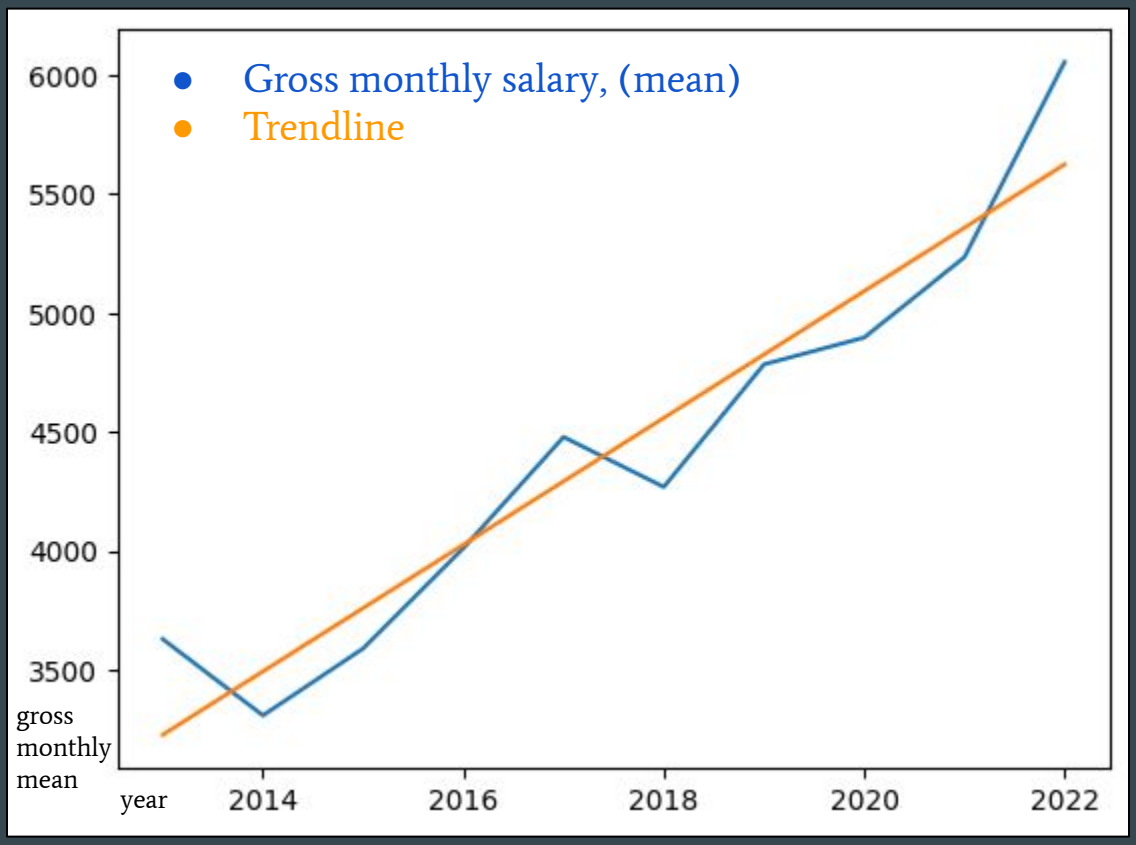


Fig 4.1

With my regression line completed, I can use this value to predict the possible future variables. Currently, the dataset provide information up to 2022. I will try to calculate prediction for 2023, 2024, and 2025. (refer to fig 4.1)

```
predict=np.poly1d(reg)
print(f'2023 prediction: {int(predict(2023))}')
print(f'2024 prediction: {int(predict(2024))}')
print(f'2025 prediction: {int(predict(2025))}')
```

[146] ✓ 0.0s Python

```
2023 prediction: 5891
2024 prediction: 6158
2025 prediction: 6424
```

Fig 4.2

With more Numpy syntax, I can calculate the prediction value using .poly1d() syntax. With this I can find out what is the possible salary of graduates in 2023, 2024, and 2025. (refer to fig 4.2)

5. Conclusion

Based on the trendline on the graph, I can see that the salaries of graduates has been increasing over the years. This is a good sign that increment of salaries will help the graduates tackle the current economy around the world.

With the predicted value, I can expect to see the mean graduates earning:

- 2023: Around S$5891
- 2024: Around S$6158
- 2025: Around S$6424

This is based on the dataset extracted from the internet. There is no way to confirm that the information provided is verified as the dataset is through interviews with the graduates. Graduates can possibly give a false value which will affect the analysis.

With more time, I would do more analysis on each of the university graduates salaries, as well as the employment rate between them.

THANK YOU!

I would like to thank you for taking your time to read my report. This report is to showcase how I apply coding skills to do analysis on a dataset.

Although it is on a basic level, I am very proud of this analysis work as it has open the world of data analytics. I had so much regarding producing from start to finish!

Citation:

Ministry of Education. (2022). Graduate Employment Survey - NTU, NUS, SIT, SMU, SUSS & SUTD (2024) [Dataset]. data.gov.sg. Retrieved March 11, 2025 from https://data.gov.sg/datasets/d_3c55210de27fcccda2ed0c63fdd2b352/view