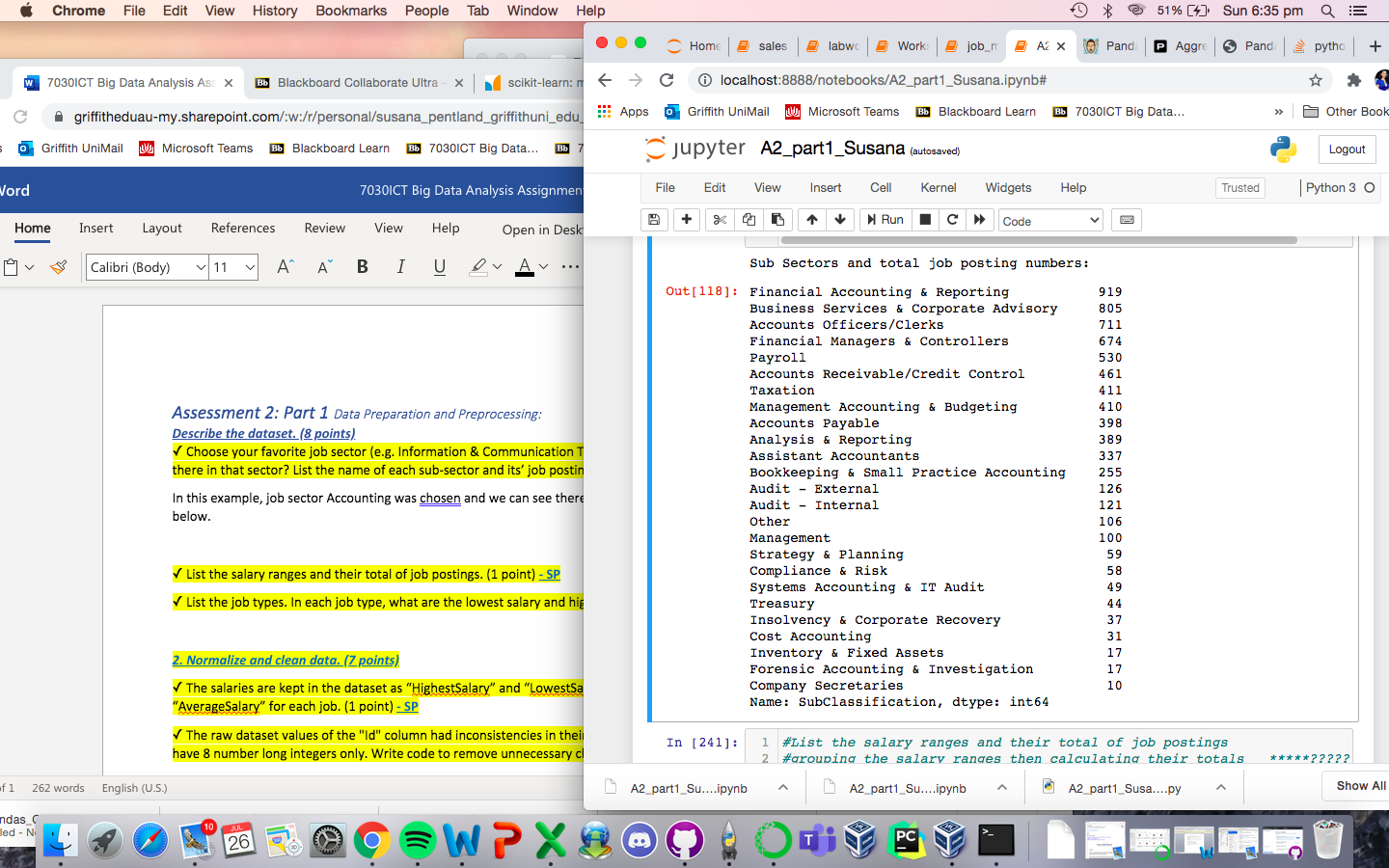
#### Assessment 2: Part 1 Data Preparation and Preprocessing: **Describe the dataset. (8 points)**

✓ Choose your favorite job sector (e.g. Information & Communication Technology), how many sub-sectors are there in that sector? List the name of each sub-sector and its’ job posting number. (1 point) **- SP**

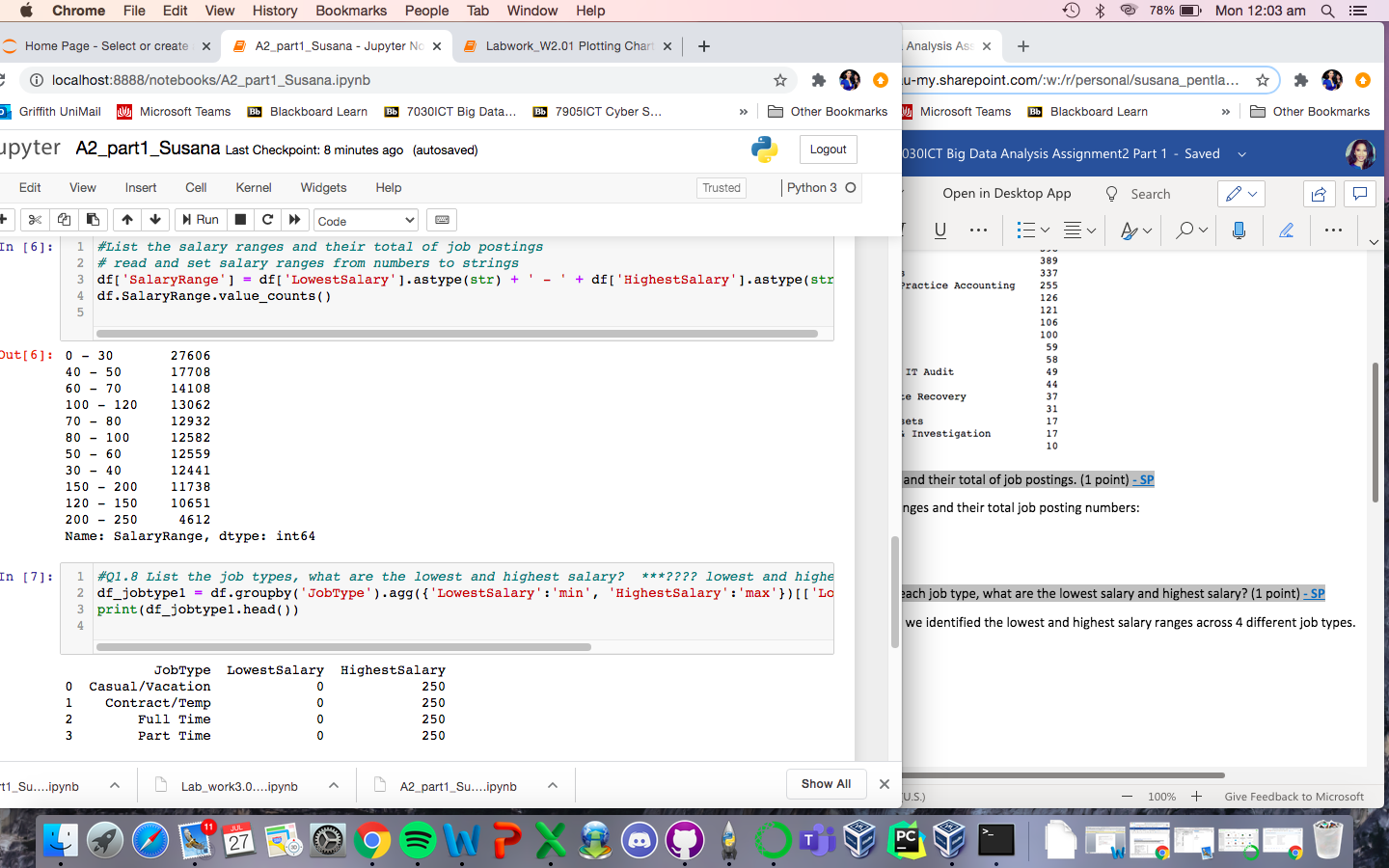
In this example, job sector Accounting was chosen, and we can see there are a total of 25 sub sectors within Accounting sector, as listed below.

This was determined by aggregating the total number of job postings within each sub sectors of the Accounting sector.



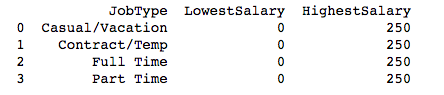
✓ List the salary ranges and their total of job postings. (1 point) **- SP**

In this list below, we have the salary range indicated on the left with their corresponding total job posting numbers on the right. This was calculated by creating a dataset from the data already provided of the Accounting sector’s salaries. We then identified the salary range by using the minimum and maximum salary values within this sector and aggregating the totals within each range group.



✓ List the job types. In each job type, what are the lowest salary and highest salary? (1 point) **- SP**

From the dataset given, we identified that there are 4 different job types that had the following lowest and highest salary, as listed below. This was calculated by grouping the different job types, and aggregating the minimum and maximum salaries of each job type.



***2. Normalize and clean data. (7 points)***

✓ The salaries are kept in the dataset as “HighestSalary” and “LowestSalary”. You should calculate the “AverageSalary” for each job. (1 point) **- SP**

With the minimum and maximum salaries provided in the dataset, we were able to calculate the average salary for each job by finding the mean value.

*TBC – not sure if this is entirely correct now\*\*\*\**

*Shouldn’t it be highestSalary – lowestSalary / 2 ?*

✓ The raw dataset values of the "Id" column had inconsistencies in their representation. The Id values should have 8 number long integers only. Write code to remove unnecessary characters. (1 point) **- SP**

During the data cleaning process, some types of challenges we would expect to see in the dataset include formatting, missing data, erroneous data, irrelevant and especially inconsistent data such as in the ID column. The dataset given is already clean with consistent format in ID column, so no further processing was required.