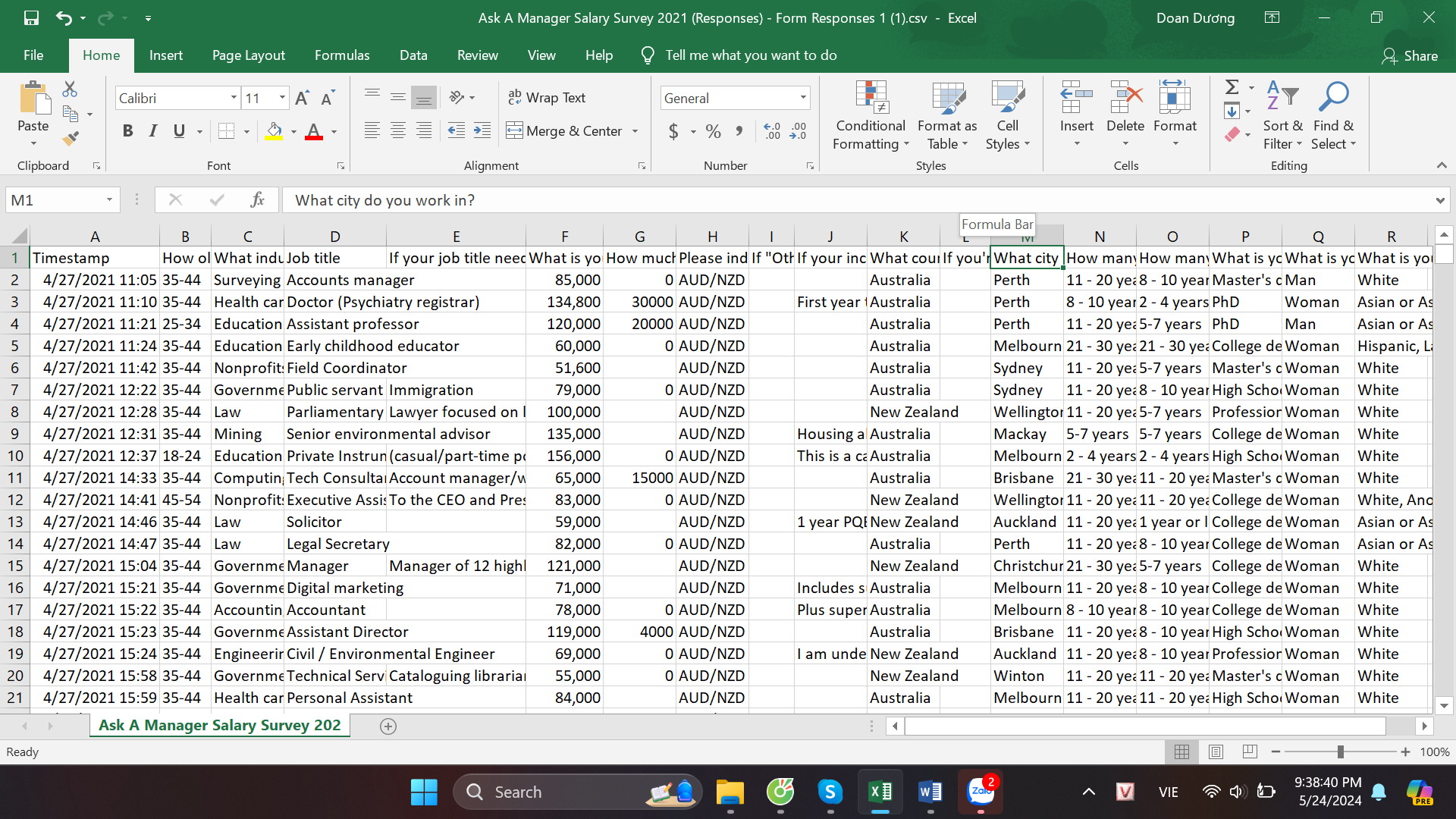
**REPORT**

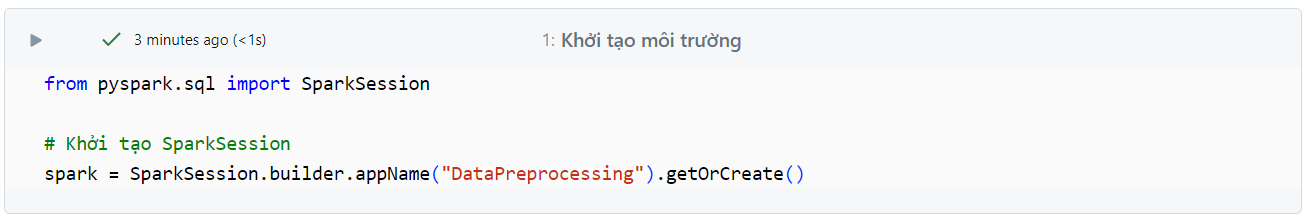
1. Overview of the dataset

The data is a Salary Survey from ‘AskAManager.org’. It’s US-centric-ish but does allow for a range of country inputs. Suppose a Business can use this data set to estimate the appropriate salary for its employees.

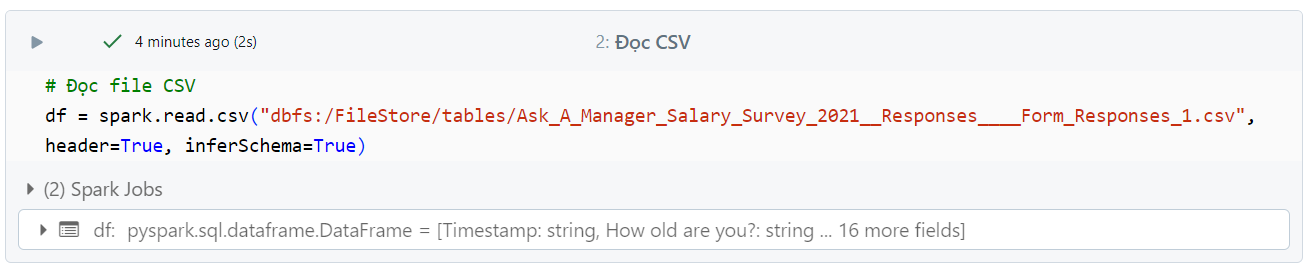




1. Data preprocessing
2. Initialize the environment

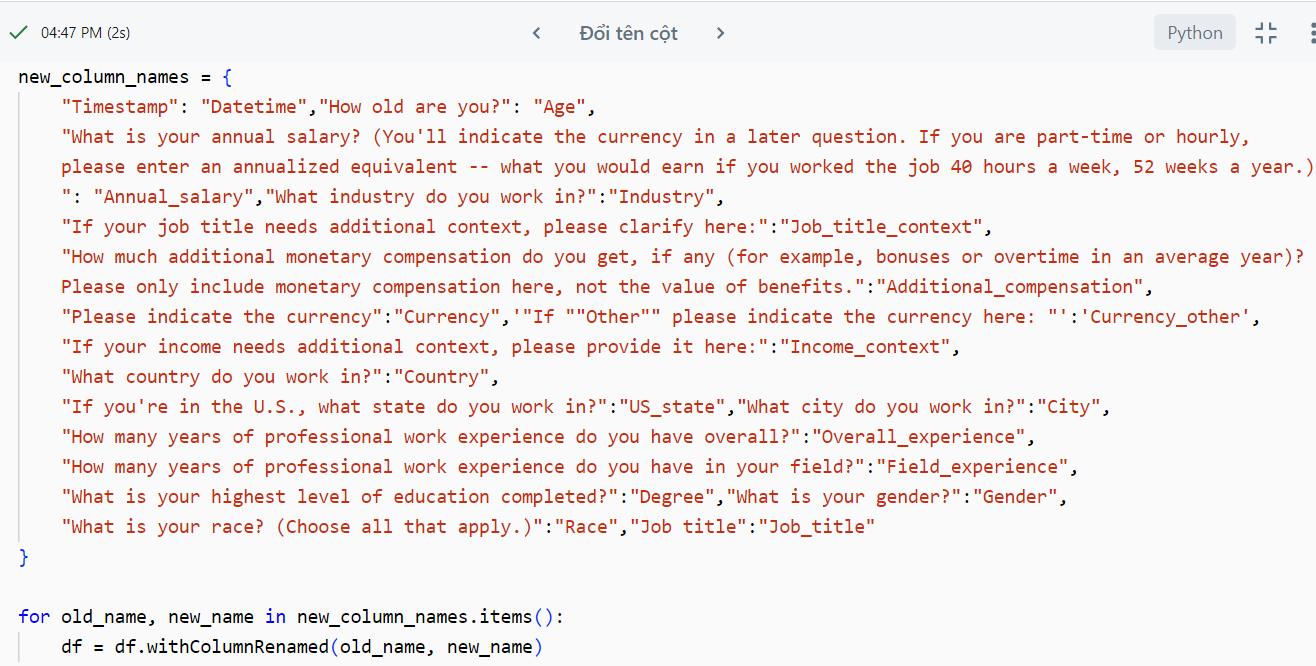


1. Read the file into the dataframe

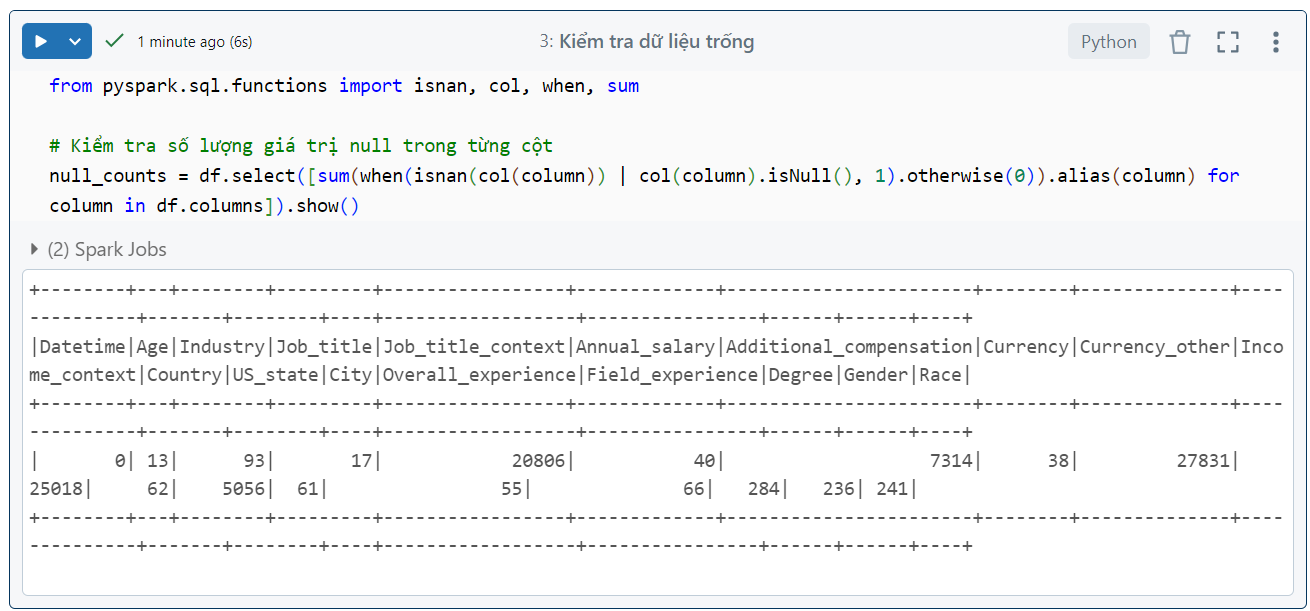


1. Rename columns

Because the columns in the current dataset are questions in the survey, I renamed the columns to be more consistent with the meaning and more concise for analysis.

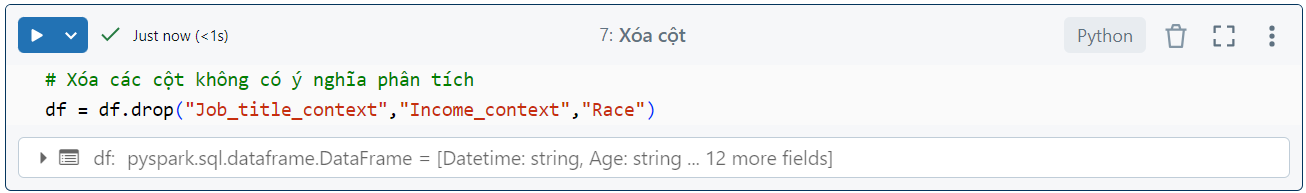


1. Check for blank rows of each column



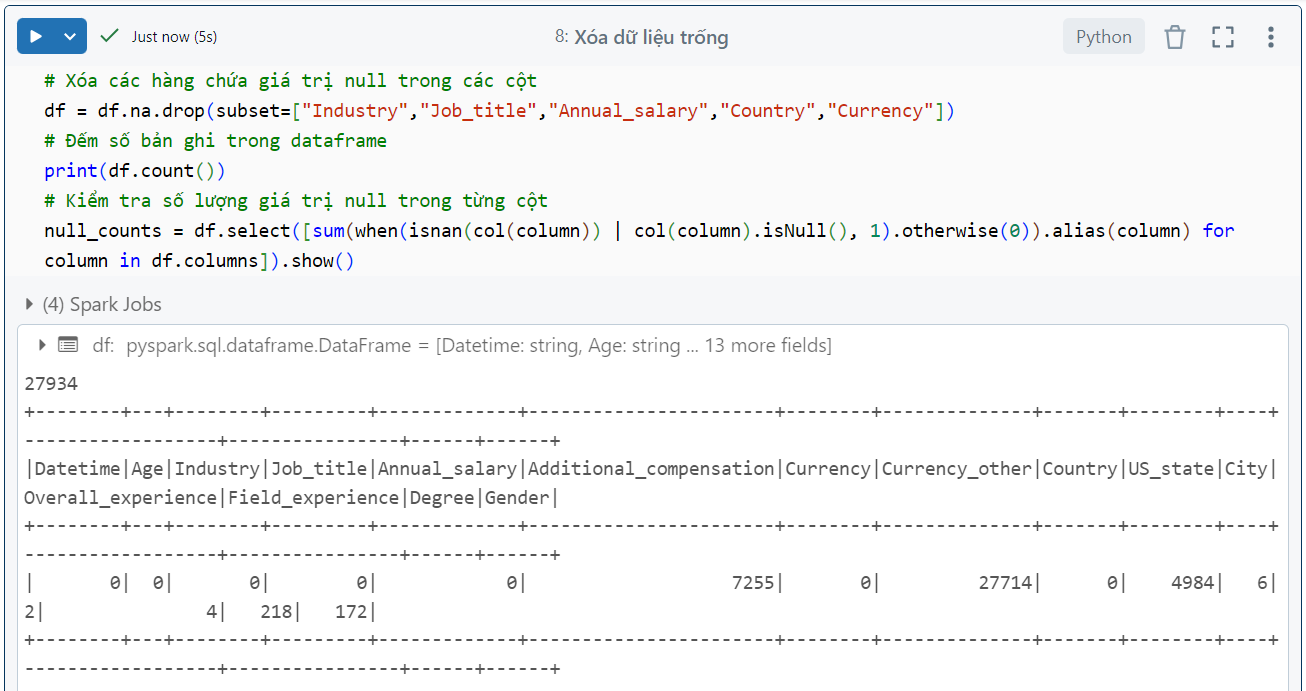
1. Delete columns

Delete columns that have no analytical significance. Because columns like job context, income context are not necessary for the criteria for determining salary for a business's employees. Race is even more unnecessary because race should not be used to differentiate between employees.



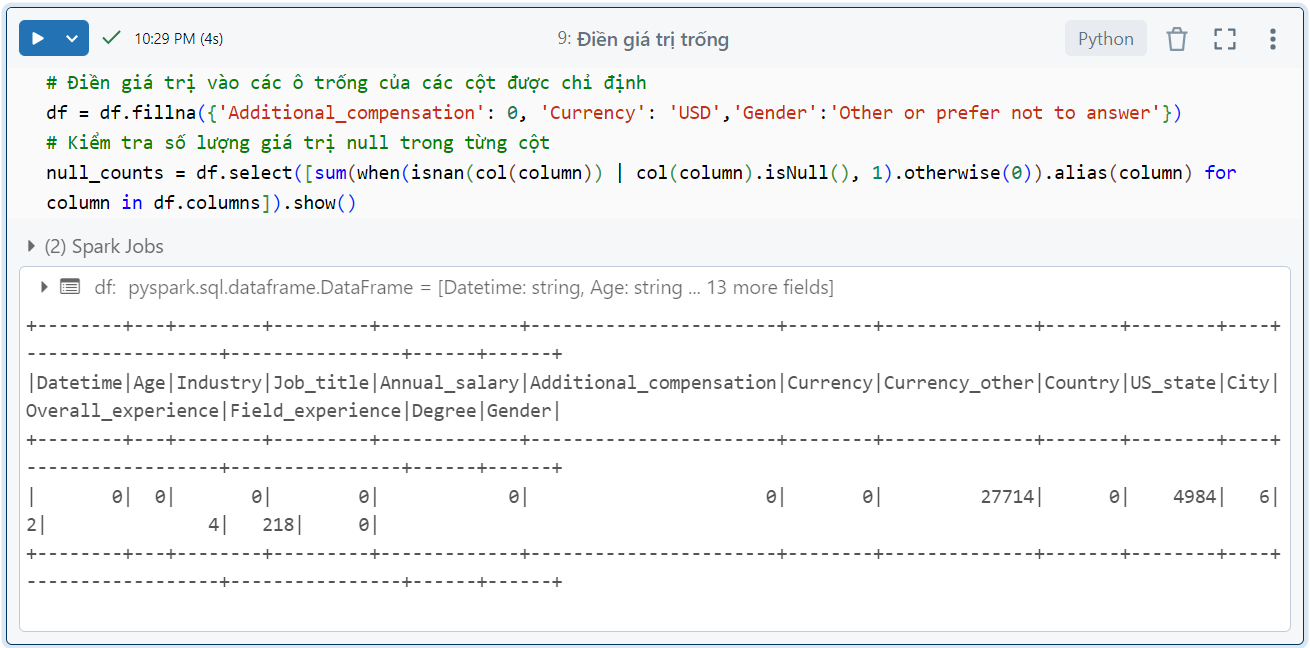
1. Delete null values

These columns are important criteria for businesses to consider salary levels. If filled out arbitrarily, it will lead to many deviations in experimental results. Therefore, deleting would be more suitable in cases where there are few null records in each column like this.



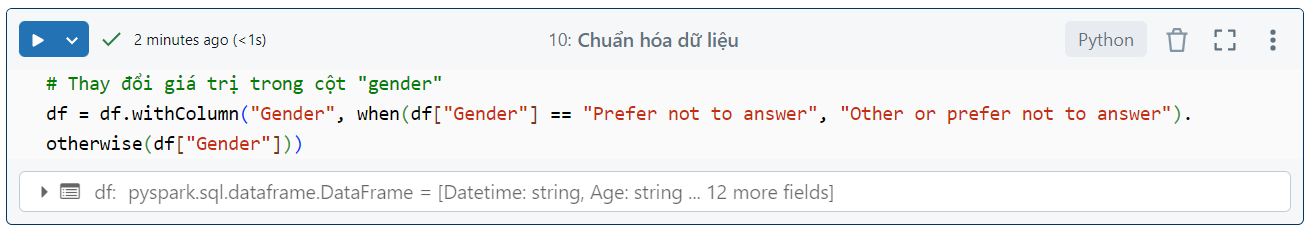
1. Fill in blank values

* If the “Additional\_compensation” column is null, I will fill it with a value of “0” for easy calculation and less impact on the actual statistical value.
* In the "Currency" column, I will fill in the blank cells with “USD” because this survey was taken mainly in the US, so filling in USD will be less likely to be misleading.
* In the “Gender” column, I will fill in the value " Other or prefer not to answer" because that option is already available in the gender field and it does not affect the model too much when filled in.



1. Normalize "Gender" column data

Normalize gender column data because their two values have the same meaning.



1. Normalize “Country” column data

Normalize country names to capital letters and remove extra spaces so they become a common convention in this dataframe. for ease of analysis in the following steps.



1. Handle data with missing column "US\_states"

Normalize the data if the person is not in the US then fill in the data as "None" and if in the US but do not fill in then the blank data is "unknown".



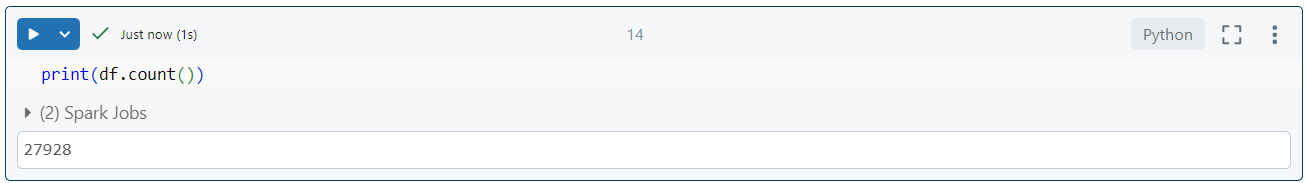
1. Merged into one currency column

* Update the other currency in the “Currency” column with the primary currency of the country he or she is working on to further detail and optimize query performance later.
* Remove the “Currency\_other” column to avoid redundant information and wasting memory.





At this point, our empty data is almost done processing. And our data is still quite a lot, so I will delete the empty records to reduce the data size.



1. Process leftover data

We have completed the section on handling missing data

