**Harnham Salary guides UK: Is they’re still a significant pay gap?**

**Overview:**

I was tasked with reporting some insights about Harnham’s salaries from 2018 – 2020. My main area of focus was to if there is still a pay gap between genders throughout the years in which Harnham has reported, their age range, likely salary ranges and if education is a factor.

For many decades, there has still been a differentiation between males and females in the workplace in terms of pay. This has become more prevalent than ever, seeing as how social media, the main marketing strategy for all industries nevermind Harnham UK.

If you’re active on social media (and even if you’re not) then its more than likely you have seen schemes to help get more women into tech, especially in the UK. There is still a demand for them. Let’s find out if we still have a gender pay gap.

**Contents:**

* **Data preparation + cleaning**
* **Data visualisations**
* **Overall analysis and next steps**

**Data preparation and cleansing:**

**Data used: Harnham salaries guides 2018 - 2020**

**Excel:**

1. **Exploratory analysis**

Each file was saved as a csv meaning it could be explored using excel at first. Straight away, there is far too many columns and there are even columns that should be values within records. Each dataset didn’t have the same number of columns and even had different names regarding data e.g. “What is your age range?” in UK 2018 was recorded as “How old are you?” in UK 2019 and onwards. With age ranges and age values respectively

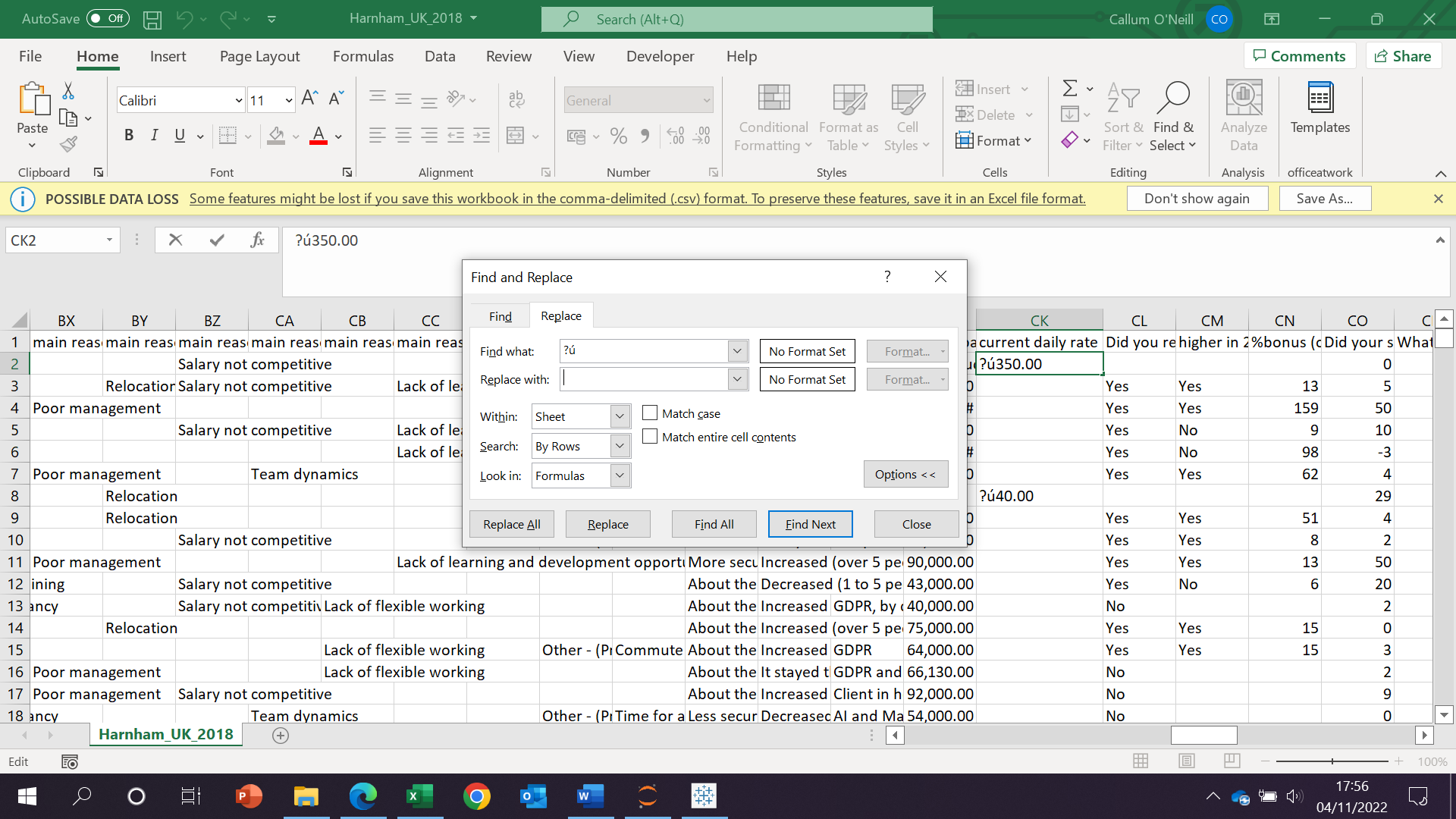
1. **Manipulating column names**

For the Column names that should have been rows, I simply changed the names of what they should be for my analysis e.g. ‘what are your reasons for seeking a new opportunity? relocation’ was simply changed to “what are your reasons for seeking a new opportunity?”. This would make it easier to explore and possibly merge in python.

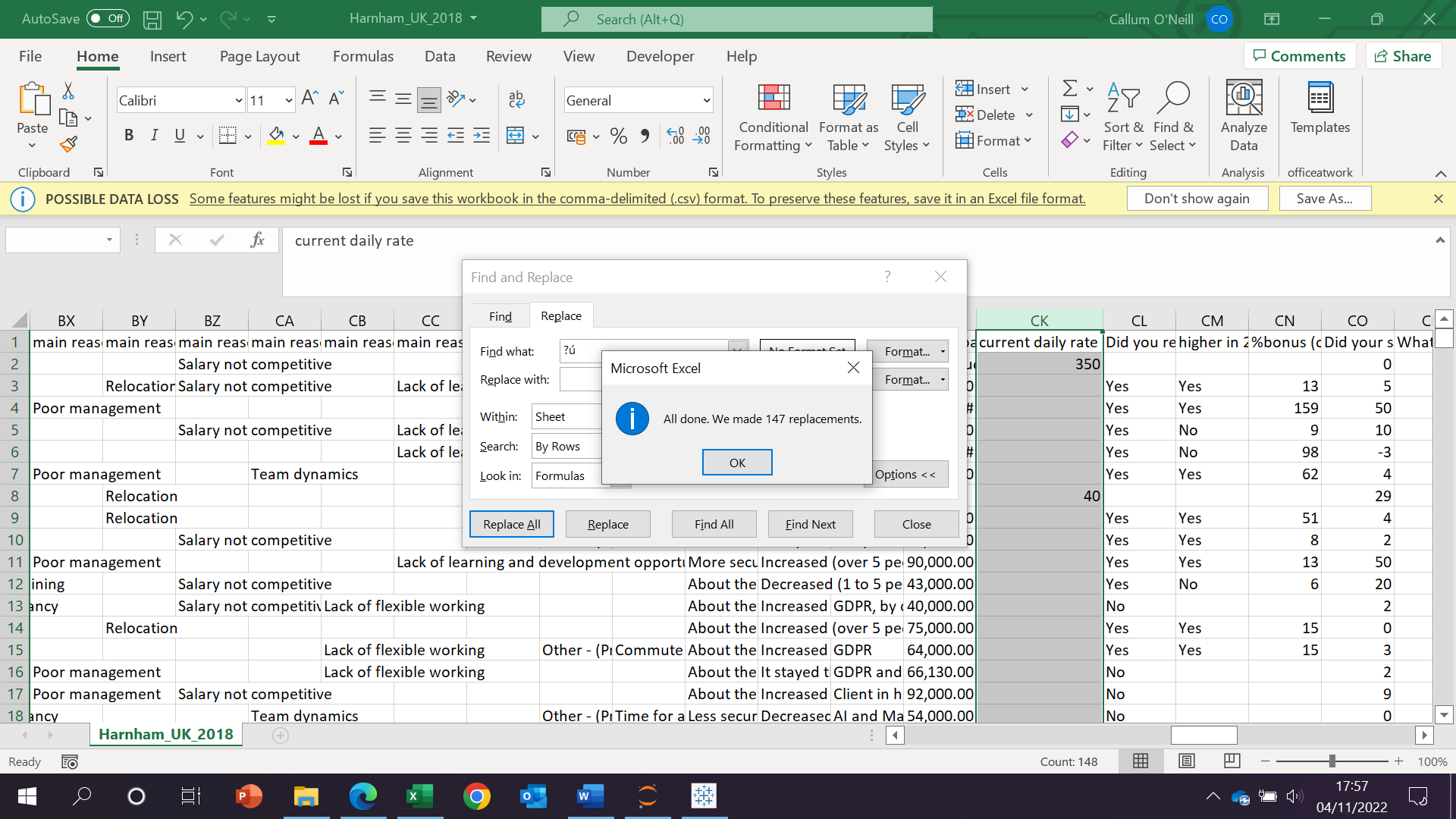
1. **Replacing special characters**

Some numerical values such as salary contained special characters in front of them. These were removed using find and replace:

**Before**



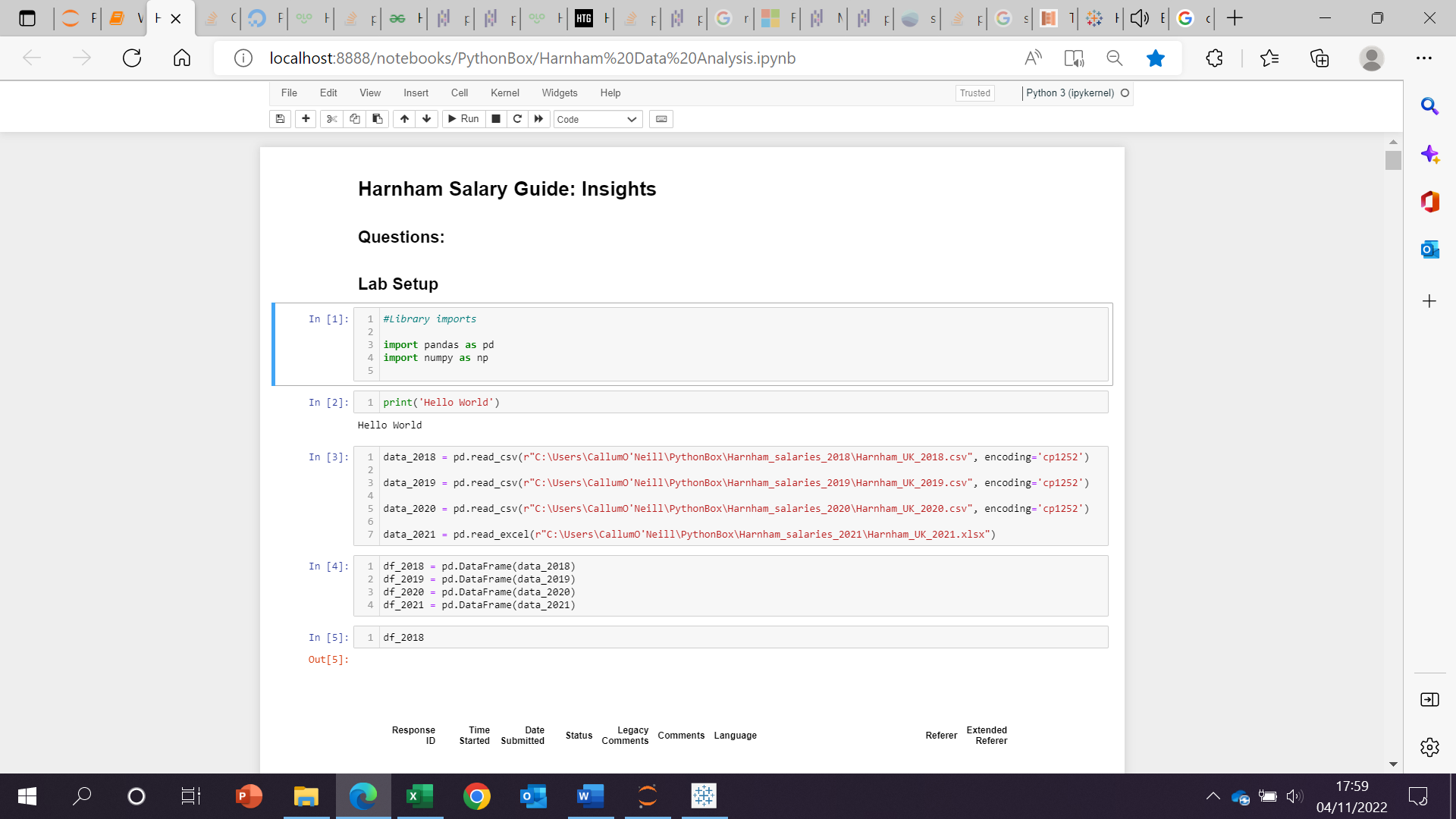
**After**

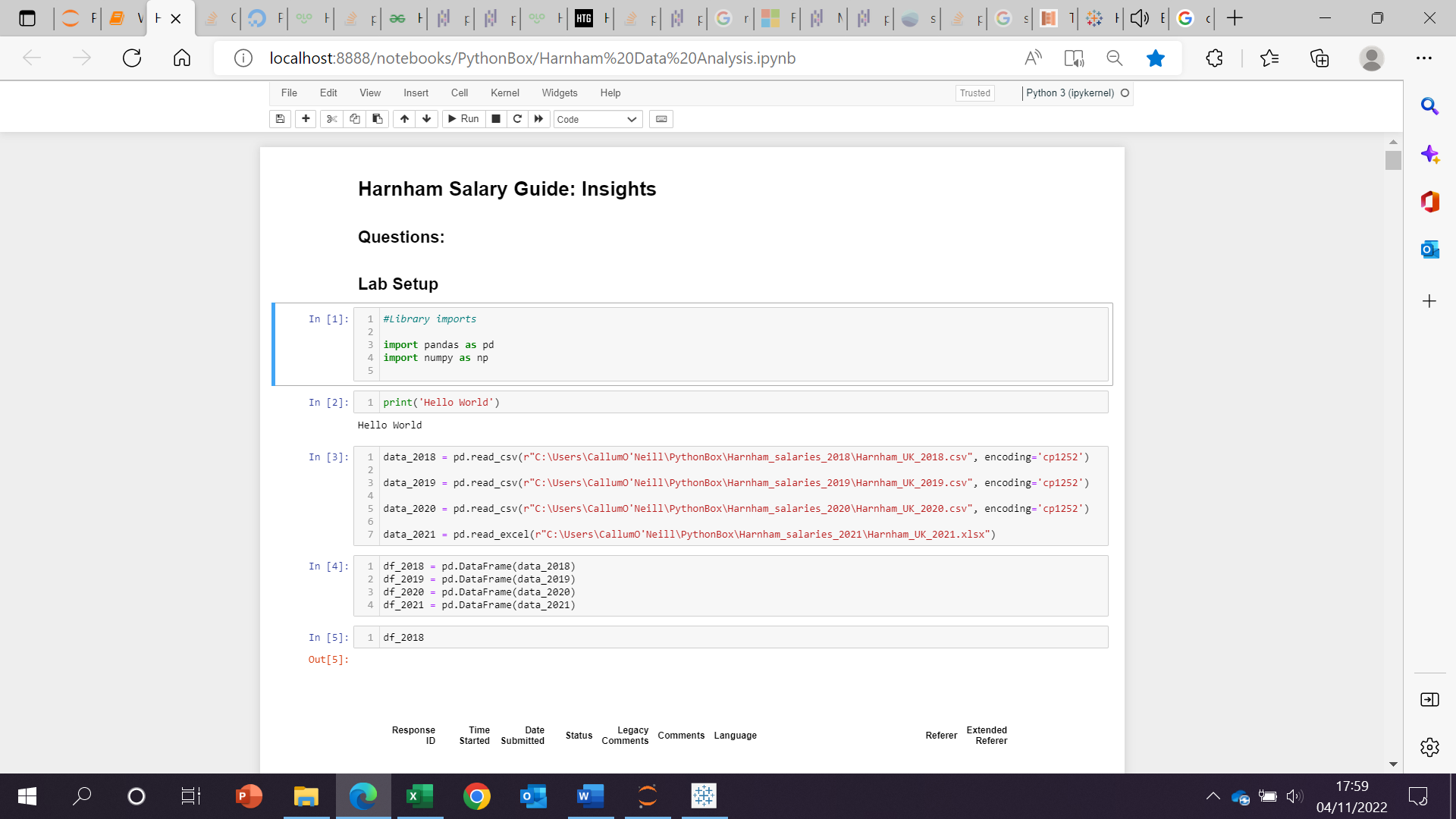


The same was applied to current salaries.

**Python:**

1. **Importing data, pandas, numpy and creating dataframes**



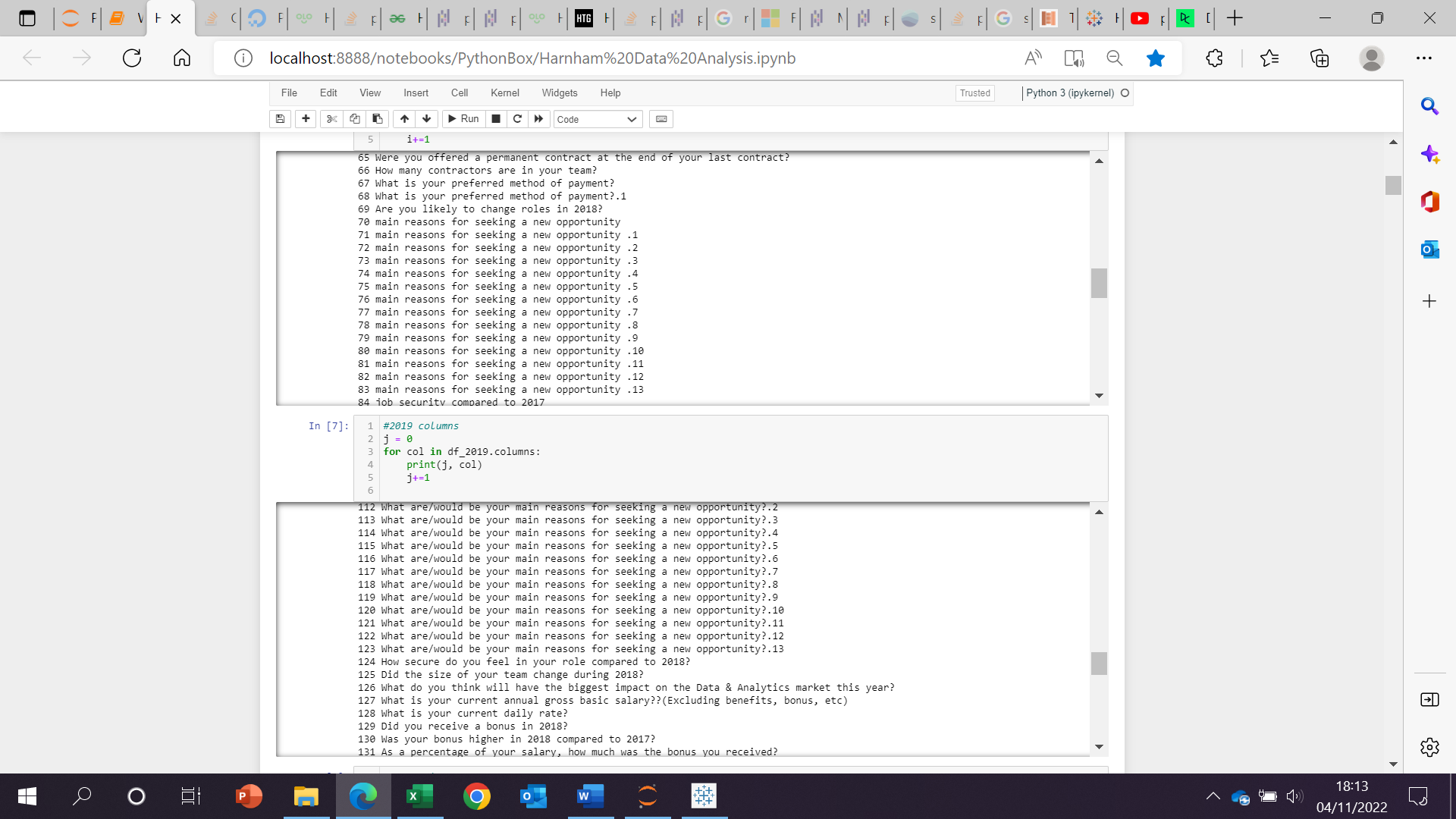


Pandas and numpy were imported as the main data manipulation libraries.

To extract the data is used read\_csvs and included some encoding due to compatibility issues. A data frame for each was created.

1. **Columns and indexing**

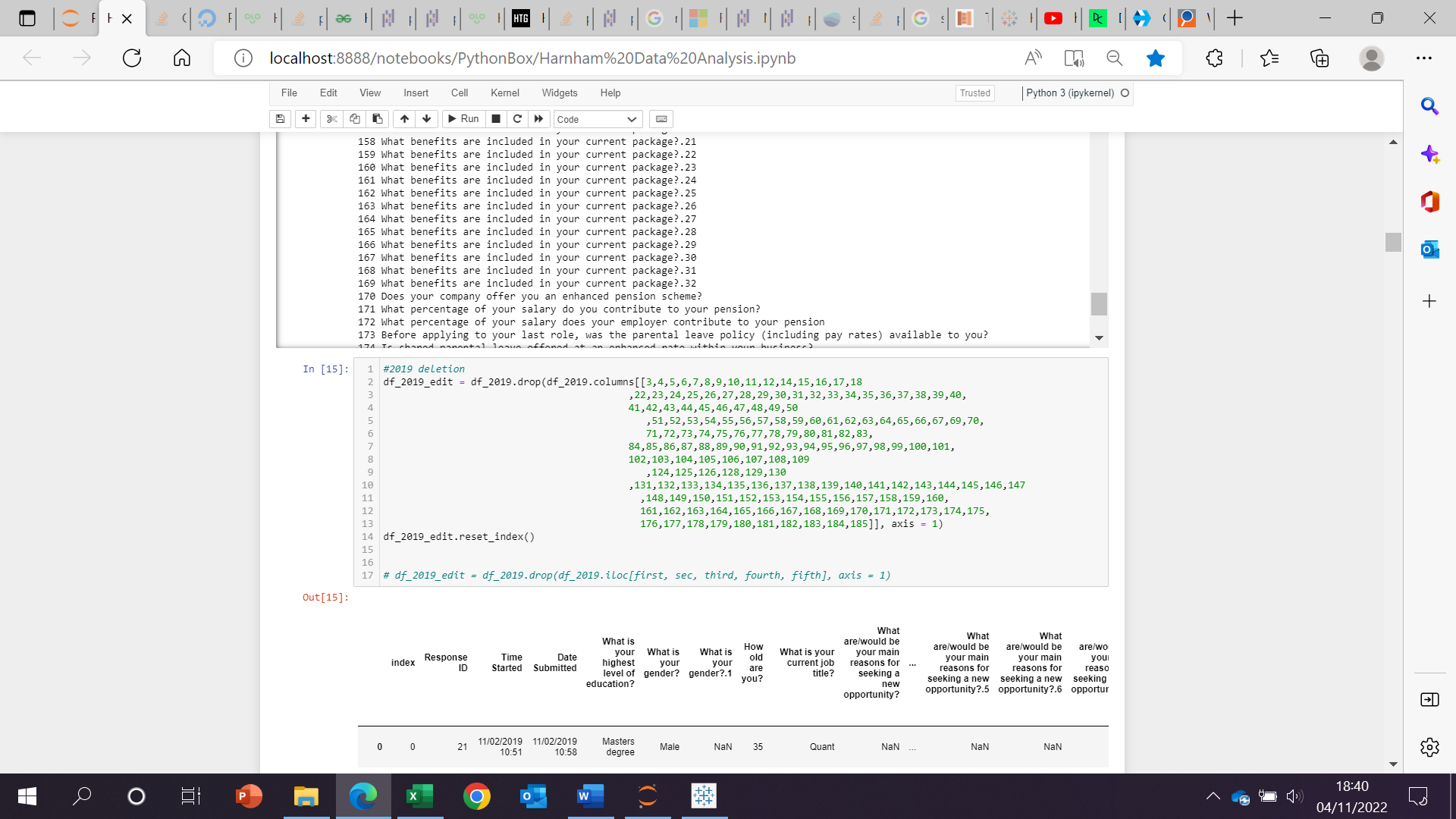
Columns of each data frame were explored to see what could be merged and what to keep for analysis. This was done using a for loop:



All the columns and their respective indexes were returned and explored. The same was done for 2018 and 2020 data.

1. **Column deletion**

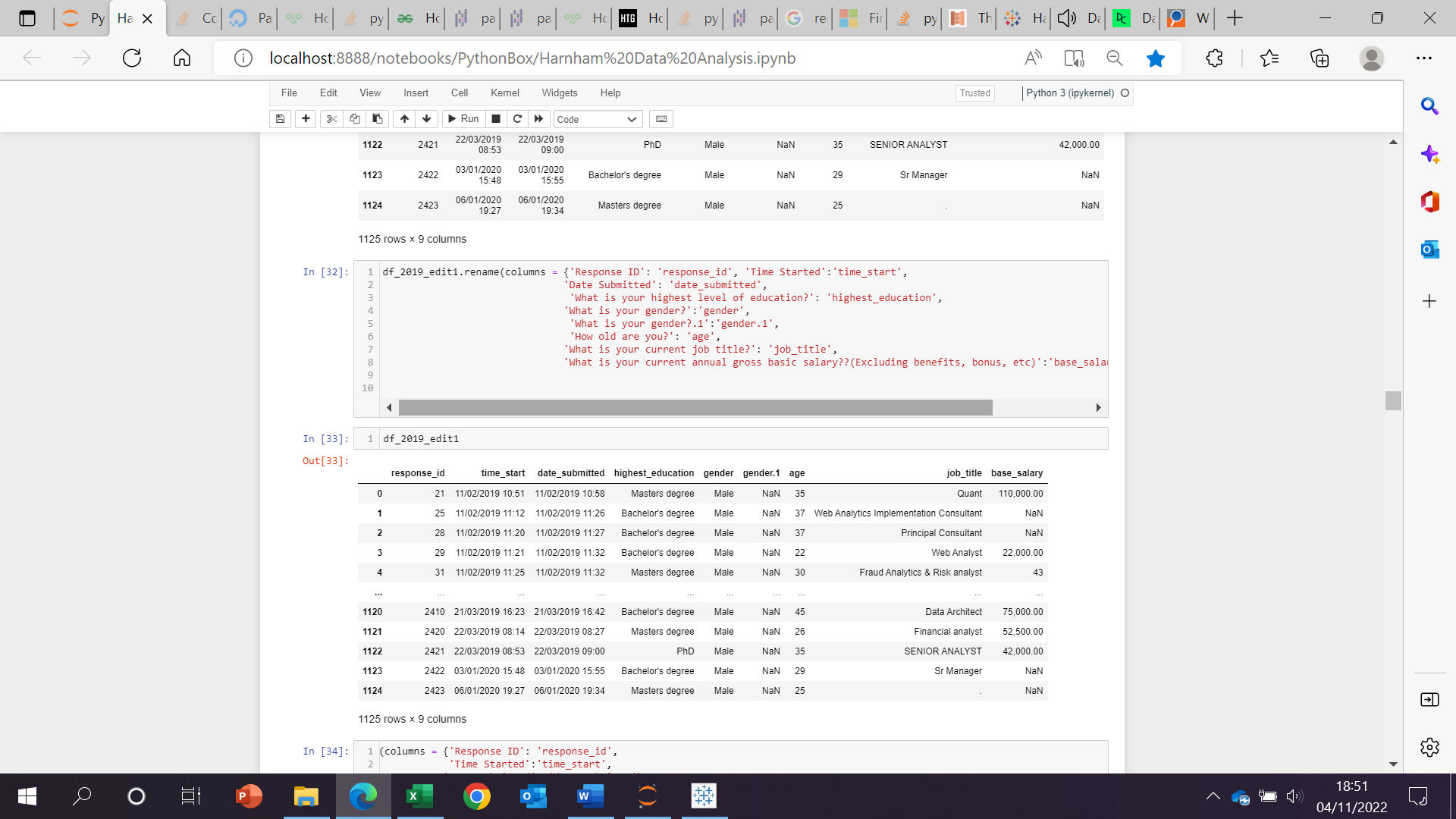
Upon deciding what columns to include for analysis, I then deleted the rest of the columns. This was done using the drop function and stored in a variable so that it doesn’t reset. Rather than writing out all the column names, their index numbers were used.



They had to be written out like this because multiple ranges were required to be deleted and too many index ranges weren’t accepted. The index was rest after the transformation. The same was done for others.

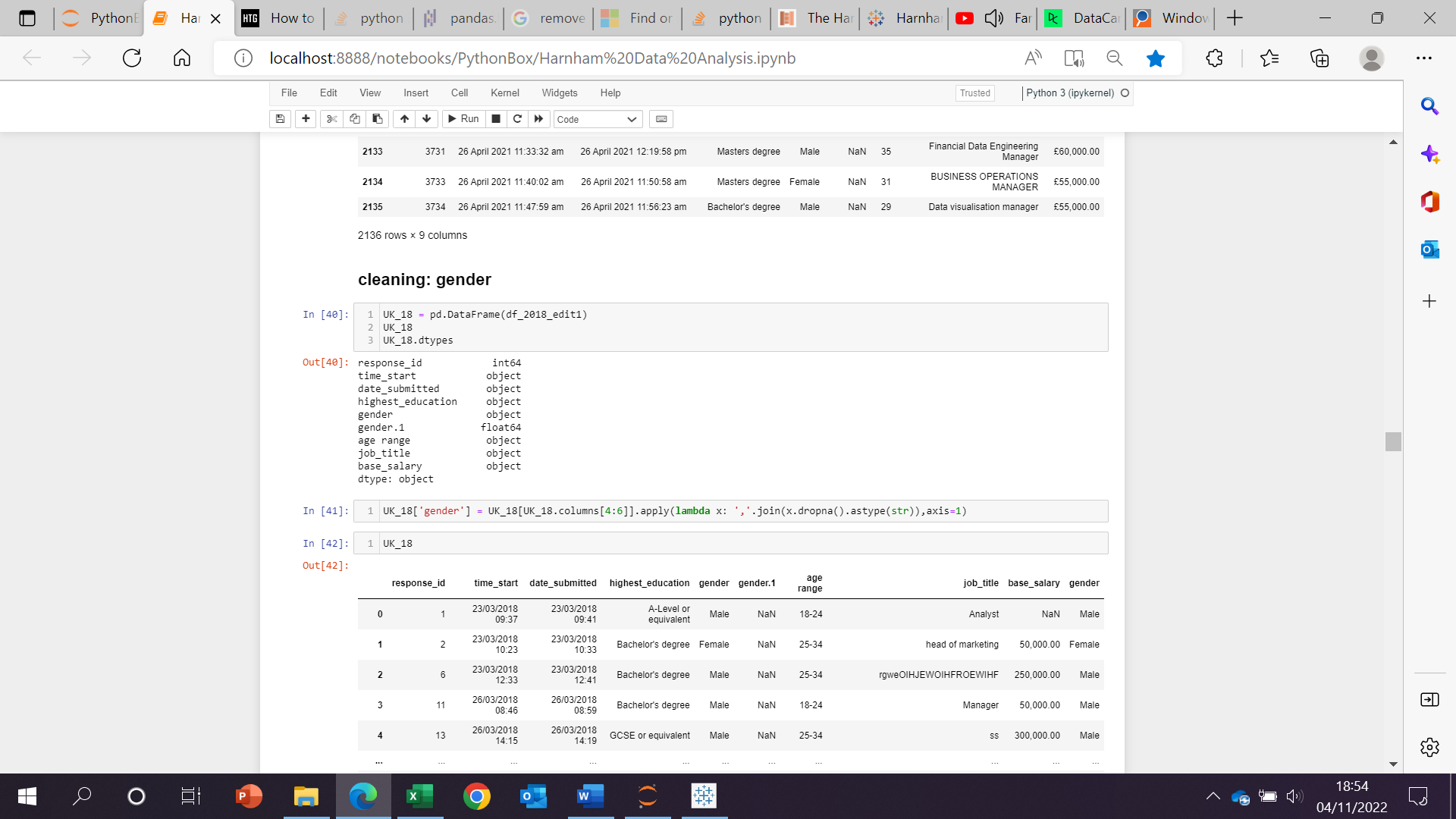
1. **Renaming columns**

Once all the columns of interest we’re extracted, they were all renamed:



1. **Merging common columns:**

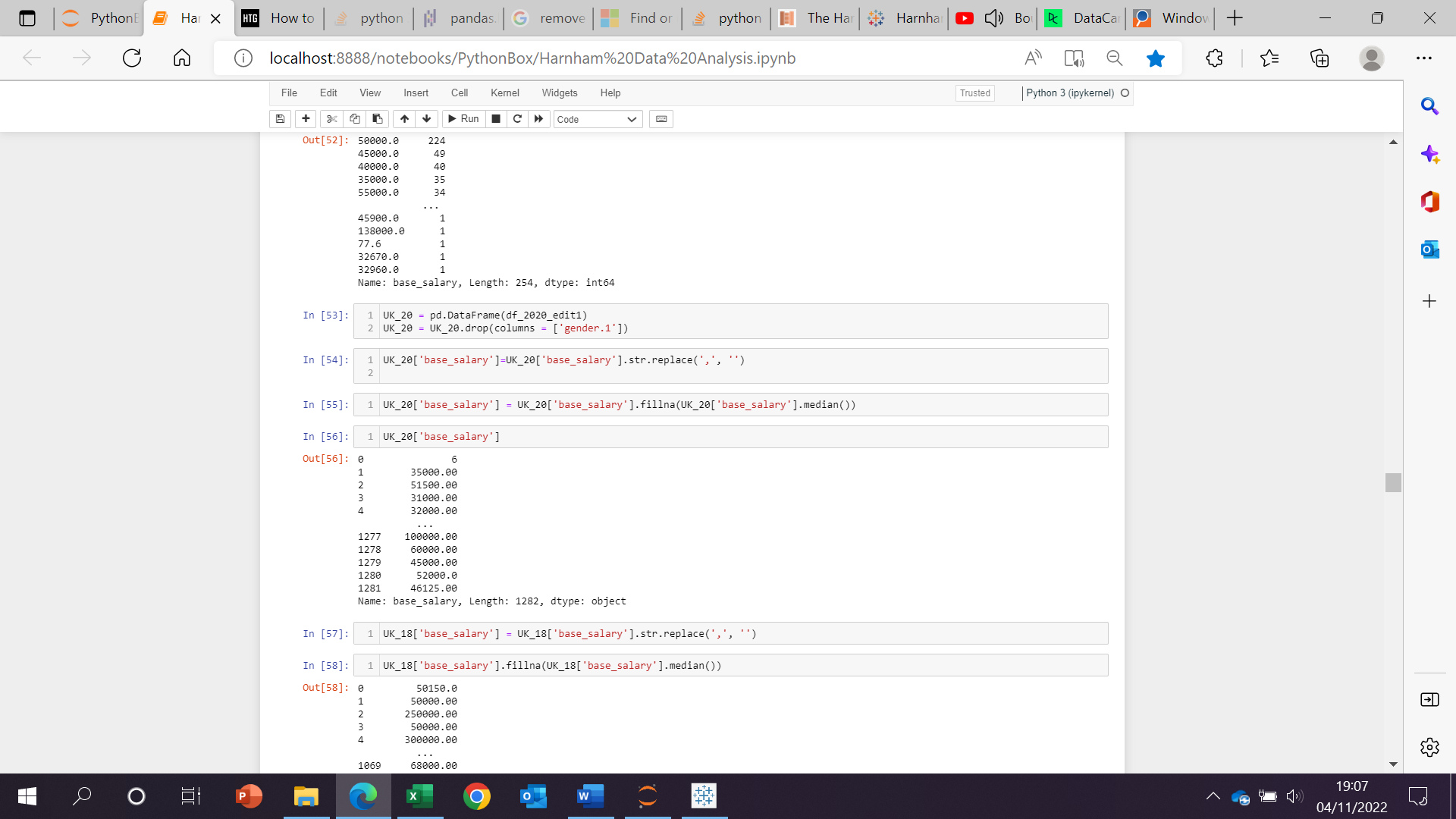
In 2018, there were two gender columns. Both had a significant number of values in, so they were combined using the following function:

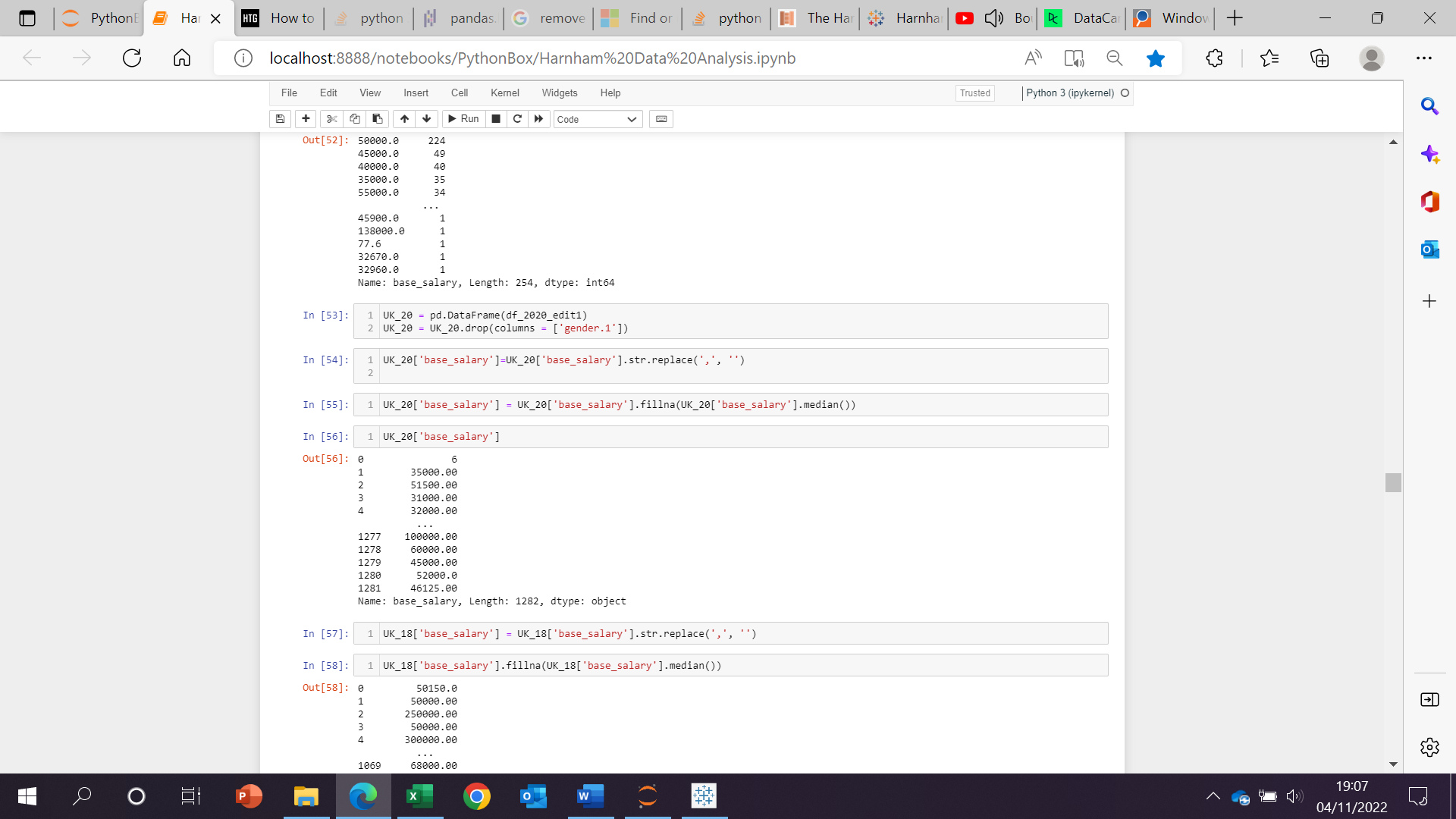


This created a new column with the combined values from both gender columns. This was not necessary for 2019 and 2020 as the secondary column contained only one value so they were dropped.

1. **Salary cleaning:**

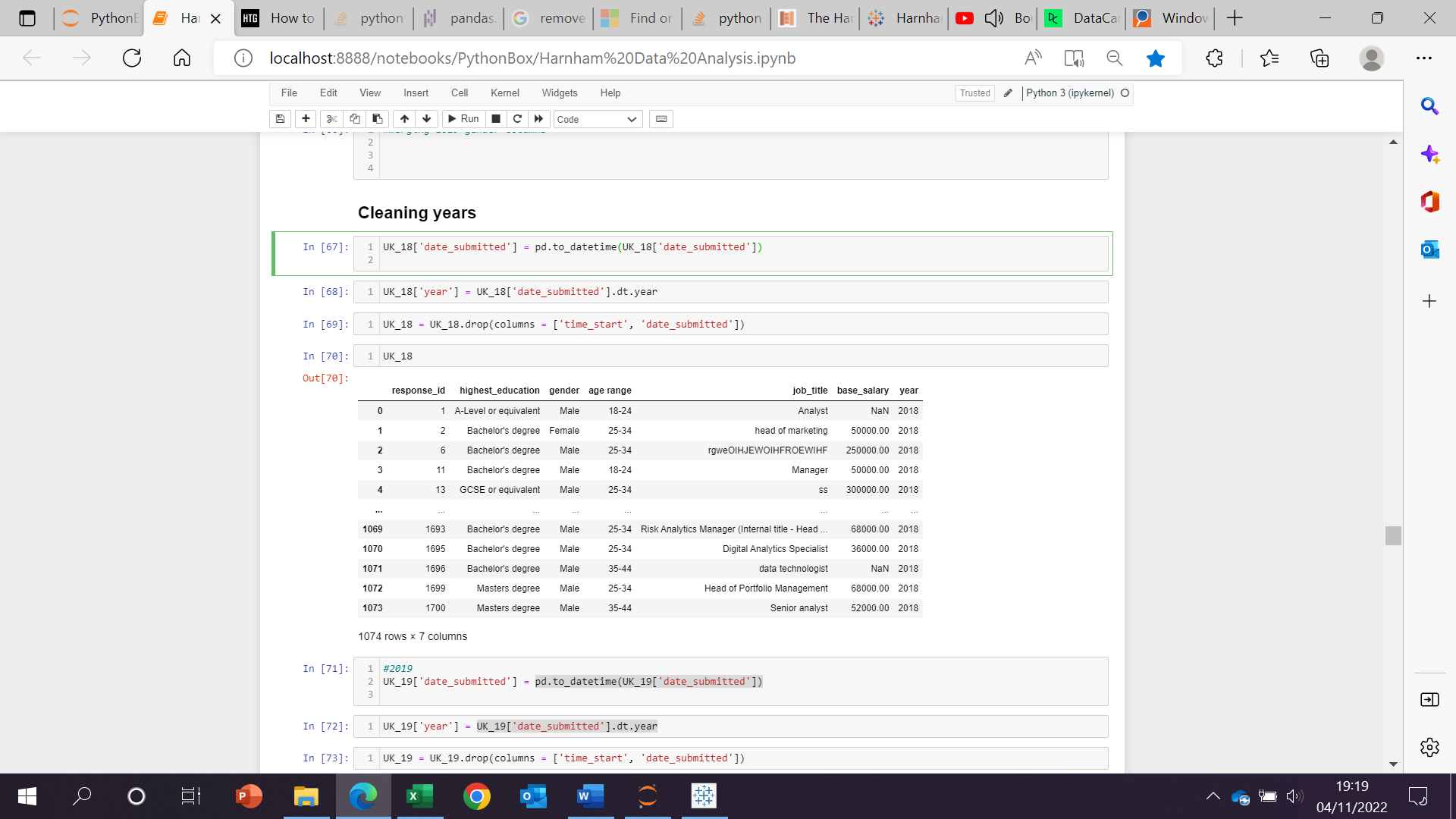
First, commas were removed: all applied to datasets



Secondly, null values were replaced with the median salary:

1. **Extracting years:**

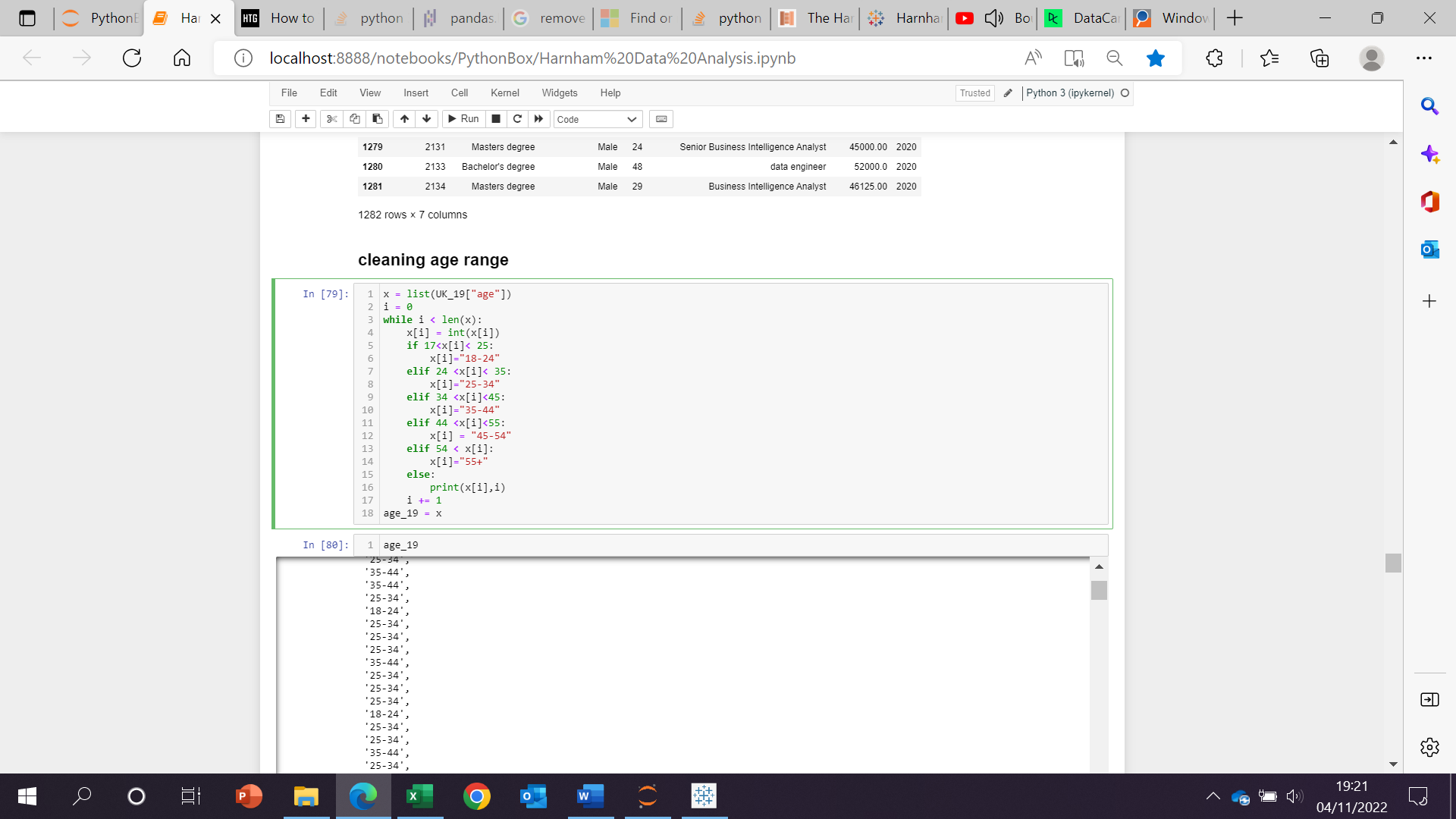
Years were extracted from the date submitted column in each dataset:



The original columns were then deleted once the years column was combined to the original dataframe.

1. **Creating age ranges:**

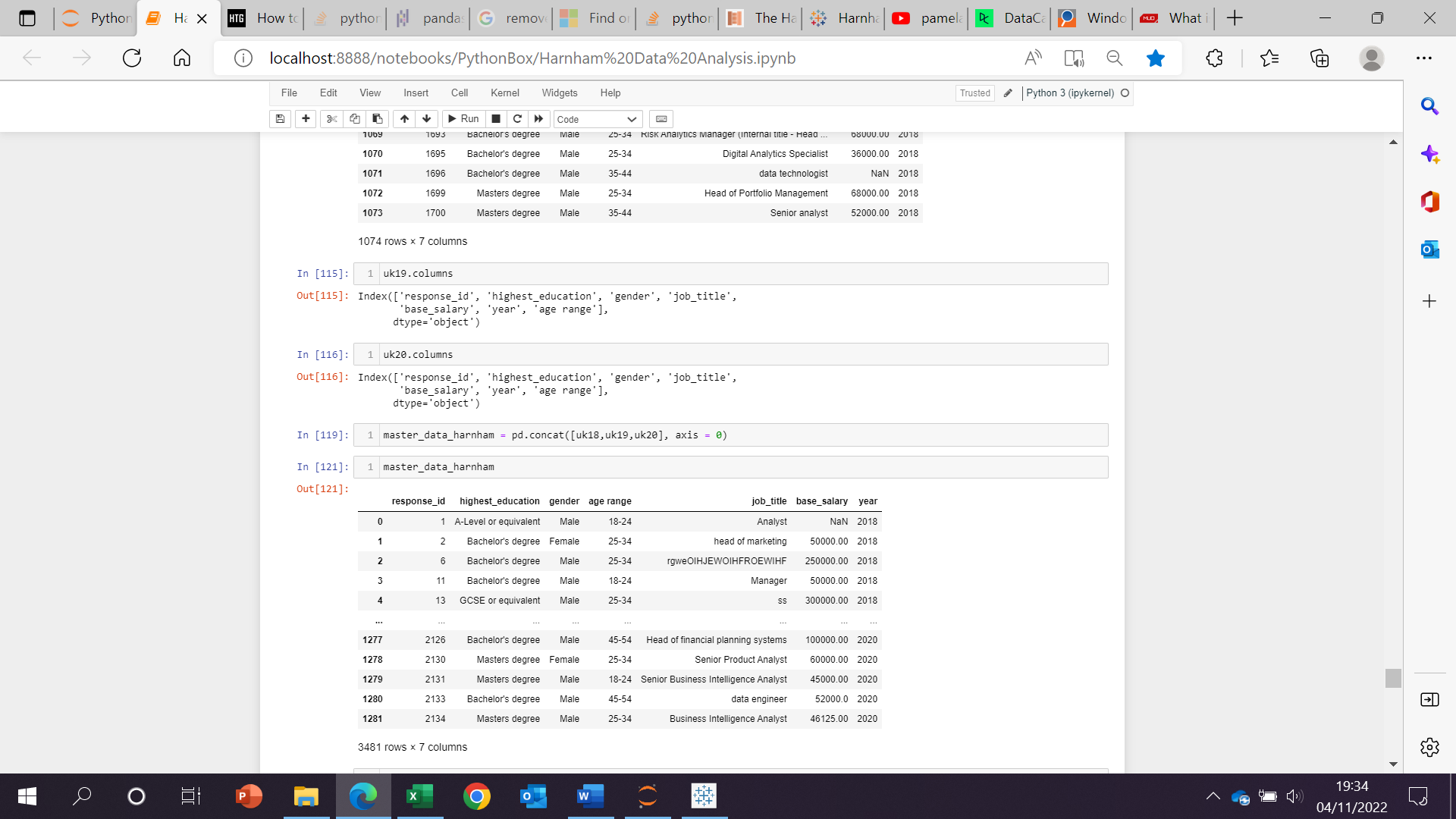
In 2019 and 2020, the age column was given an age rather than a range. This was cleaned using a for loop:



The new column was combined back to the original dataframe and original age list was removed.

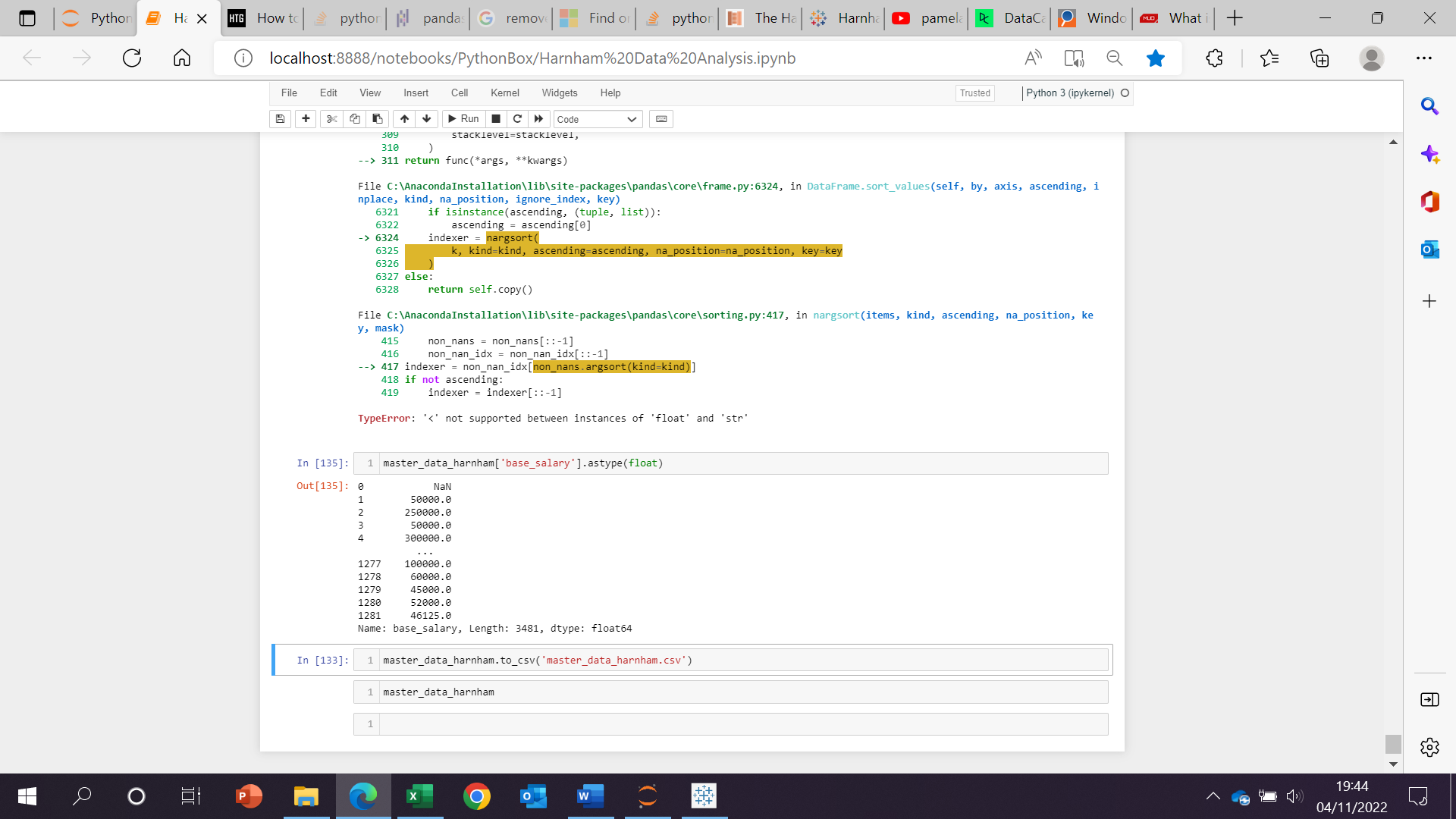
1. **Concatenation:**

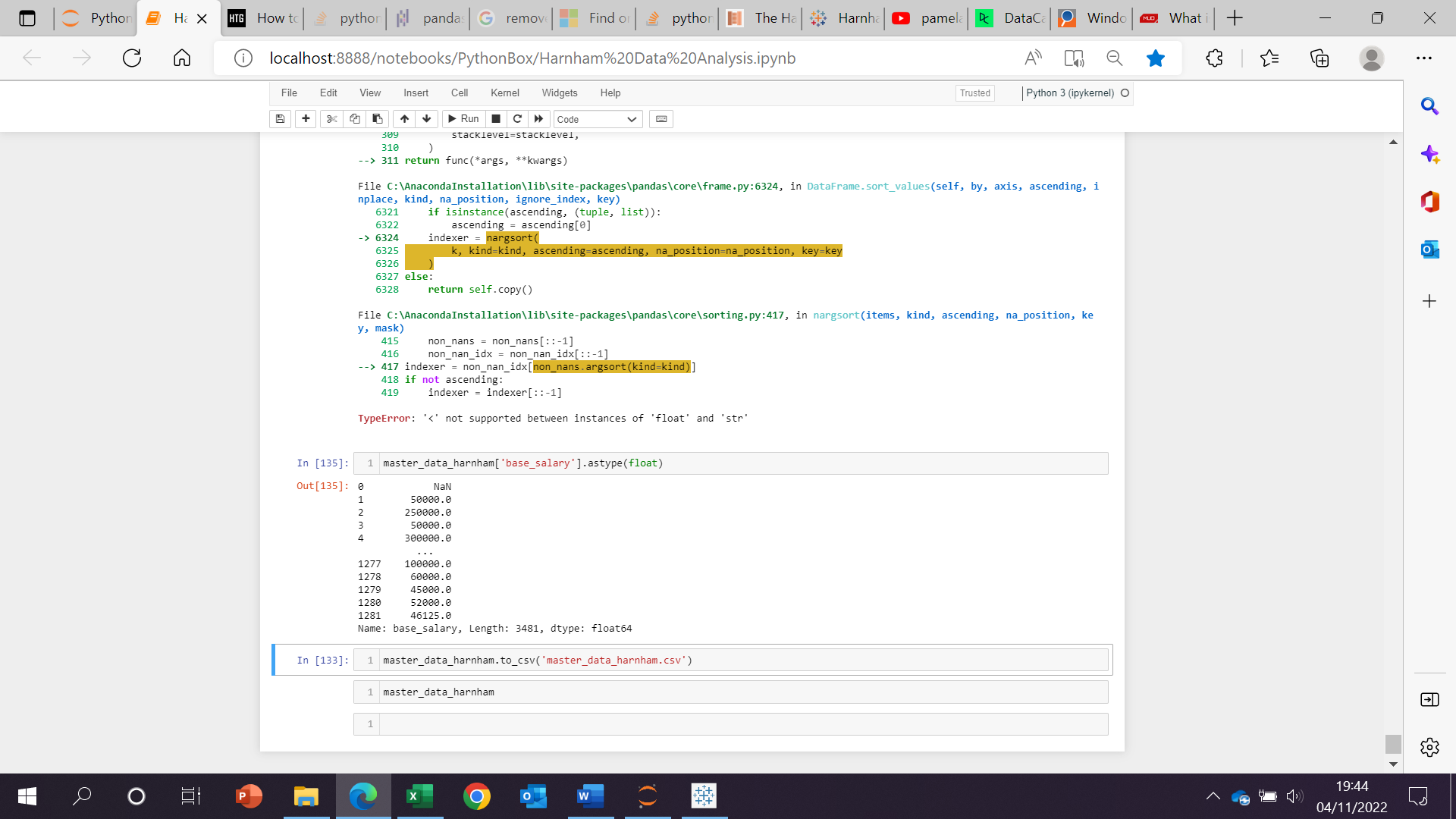
All data, separated by years, were concatenated (combined) to create a master data set:



1. **Converted numerical objects into floats and Data Frame saved as a csv**

Salary was converted to a float rather than an object in the master data set then finally was saved as a csv:





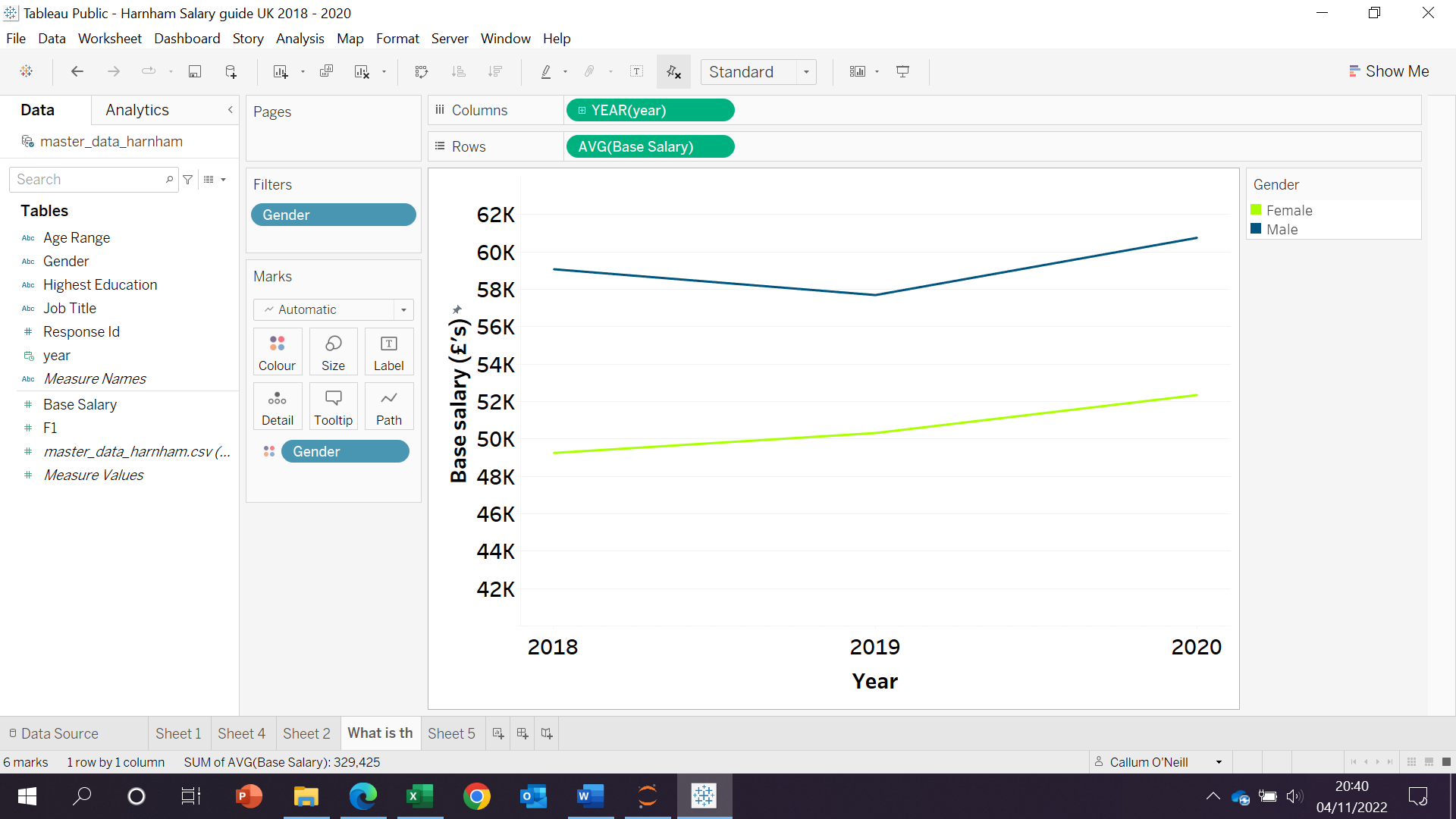
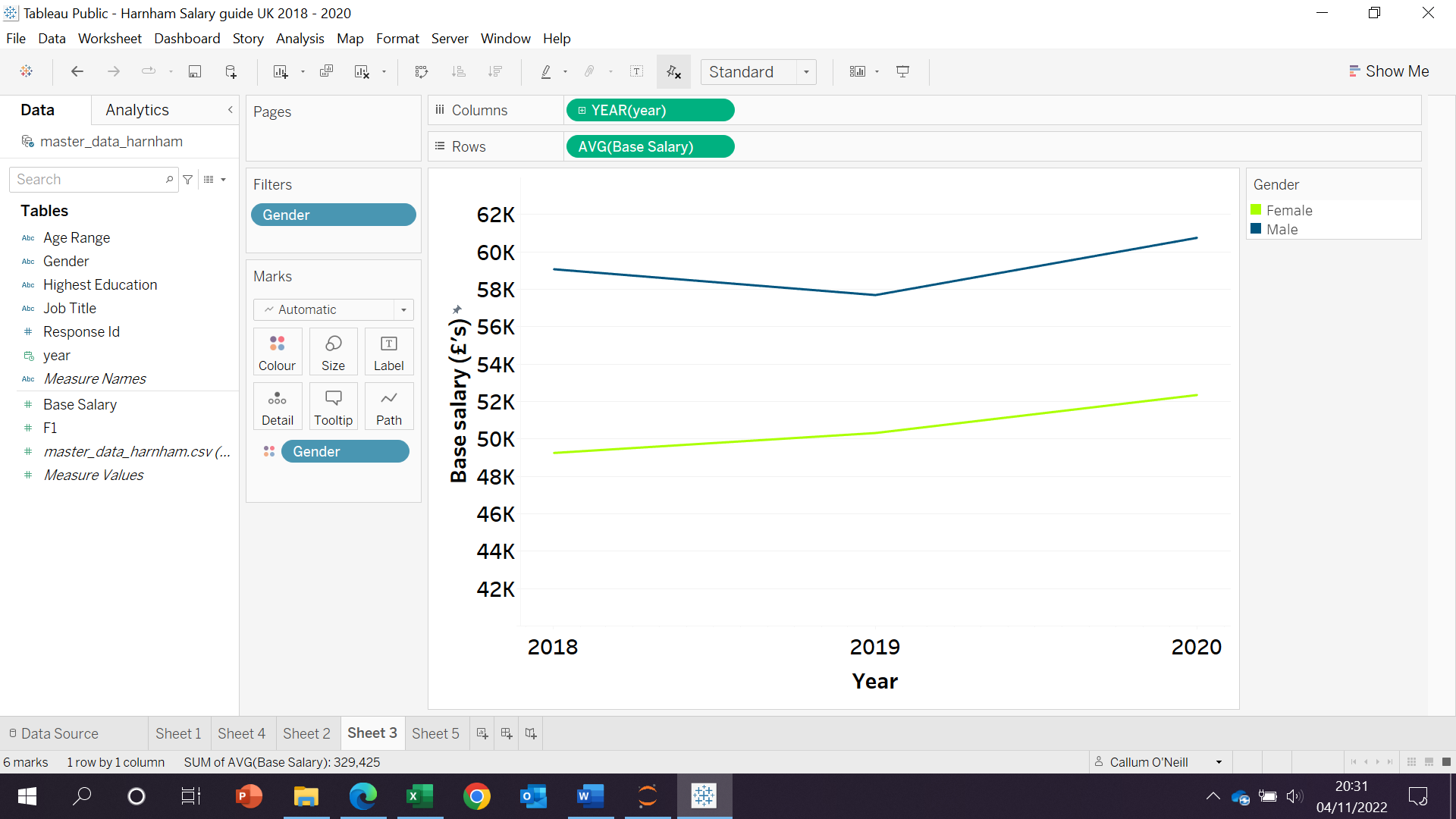
**Tableau:**

The saved csv was imported into tableau to conduct visualisations on the relevant insights. Small number of null values were removed alongside some incorrectly filled gender data and distinct gender assignments.

**Visualisations**

**What is the moving average of salaries for both males and females?**

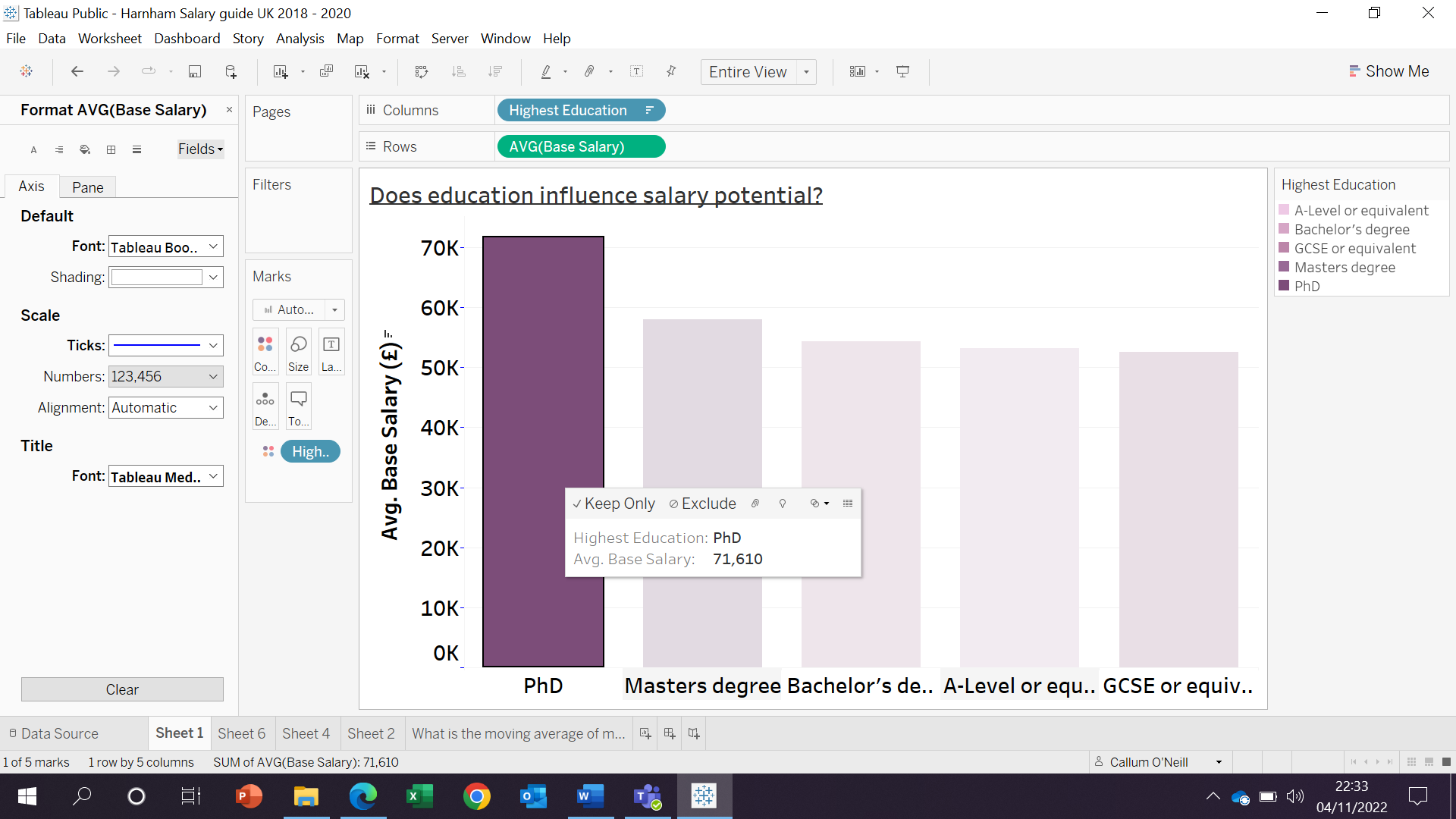
Despite the push for more women into data and tech within the 21st century, it’s clear there is still quite a large gap in the moving average of salaries year on year. The general trend is on the increase for both male and female. In 2020, height of the covid-19 pandemic, data become more in demand due to more people workingfrom home and government pushing for covid-19 case monitoring and vaccine rollout:



* **2020: average** male salary on was approx. **14% higher**
* **Female salaries: approx 6%** increase from 2018 to 2020

**Does education influence salary potential?**

When looking at what links to a high salary or even a stepping-stone, nothing has more polarising opinions than education. Some think it’s necessary for success, others not so much with prime examples of mark Zuckerberg and bill gates being college dropouts. It’s still clear that education has a part to play as PhD (almost at top tier for education) will grant, on average, a high salary in data and analytics.

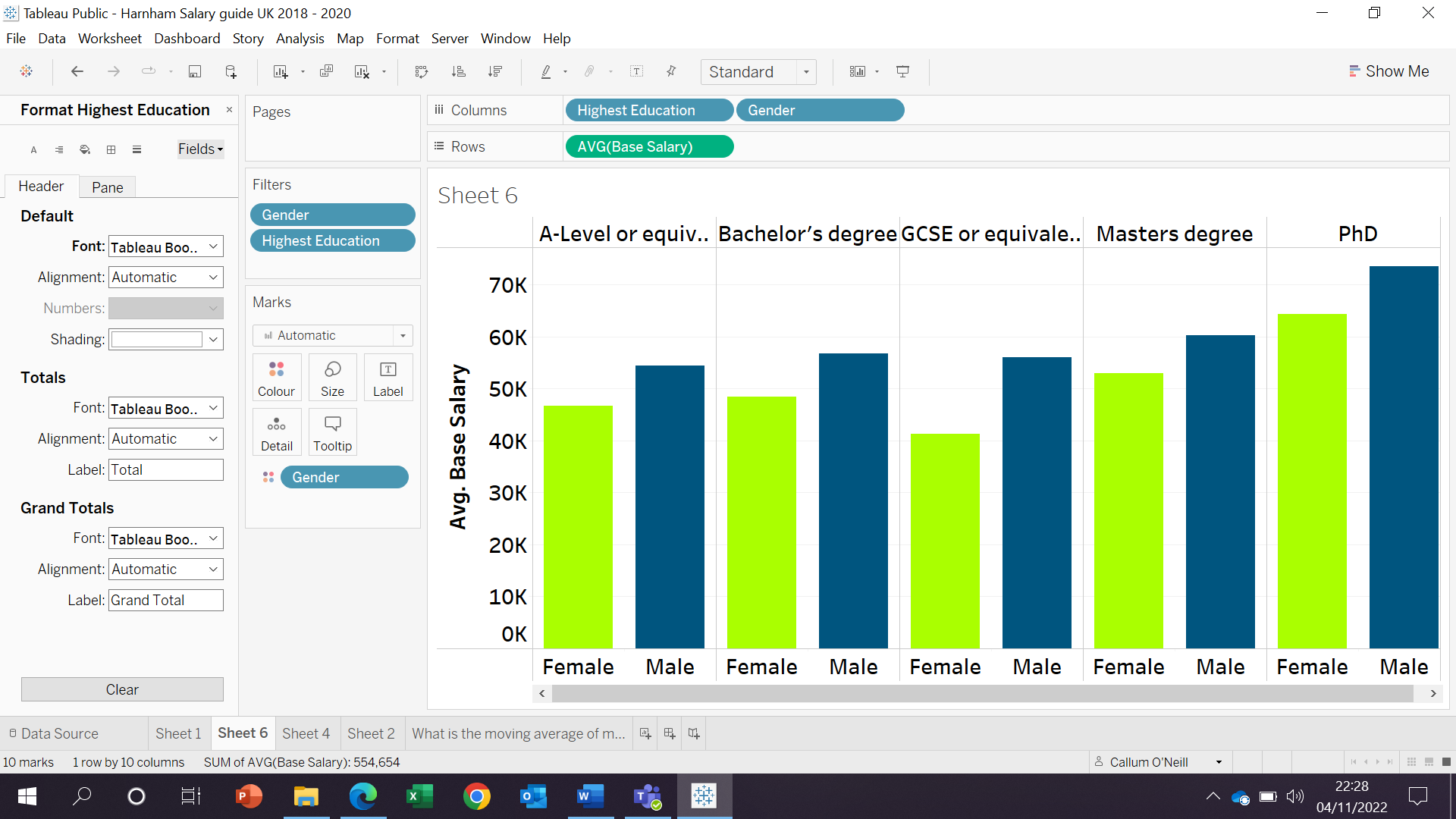


* **A-level, GCSE, bachelor’s degree: not significantly different** from each other
* **phD** has a an approx. **19% increase** compared to master’s degrees

**How does education look for male and female salary potential?**

There is still minimal difference, and it seems that men are still earning more regardless of their level of education. At the GCSE and equivalent level, the gap is most significant.

* Do males have the most potential even if their highest qualifications are GCSE’s or equivalent compared to females?
* Is it accurate to judge earning potential for both genders on GCSEs only?

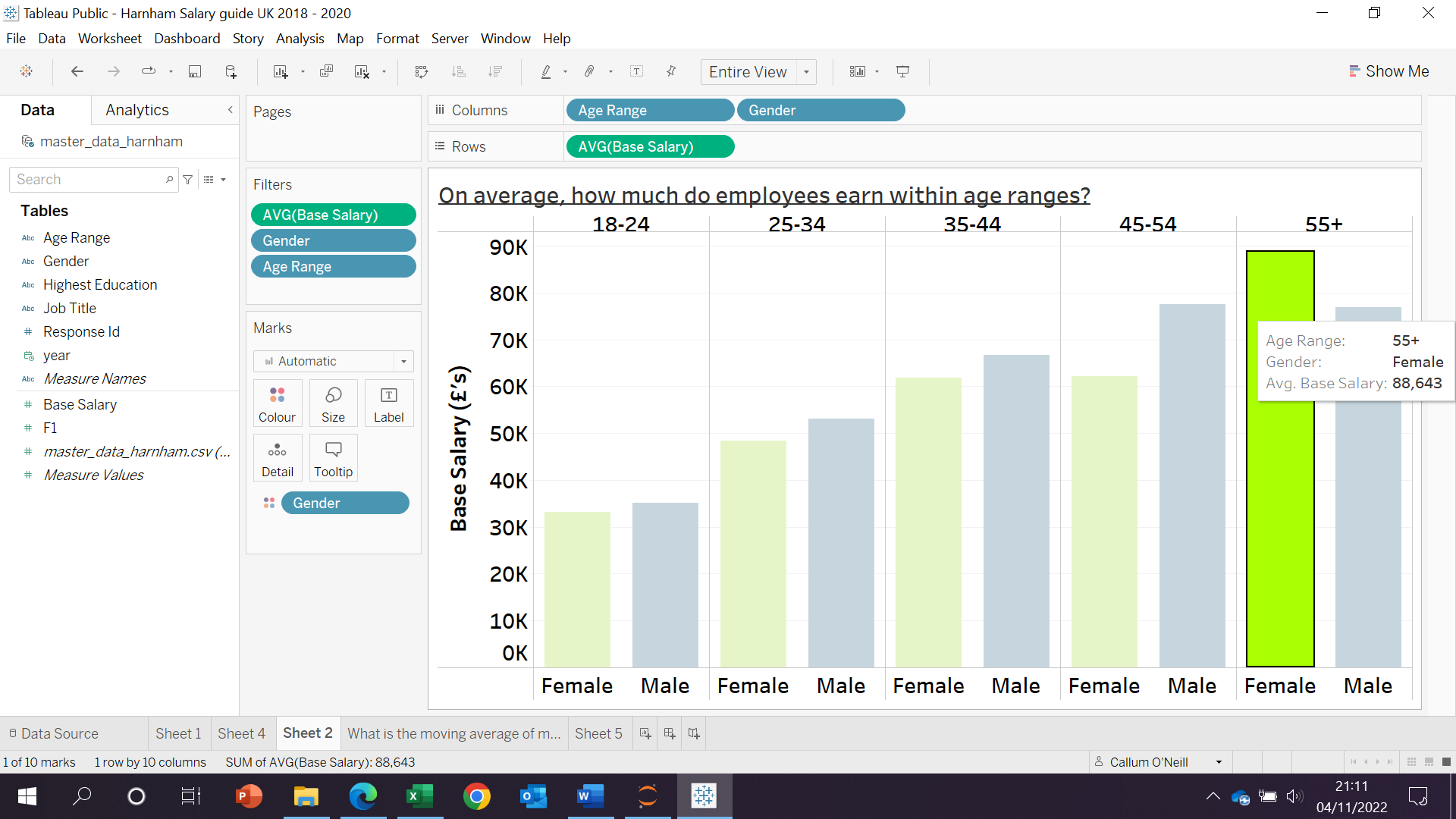


* **GCSE’s: Approx. 26% difference** in average salaries for males
* **phD:** males earn **approx. 12%** more than females

**On average, how much do employees earn within age ranges?**

With confidence, it would be correct to assume that as employees become older, they become more skilled and likely take on greater responsibilities. In age 55+, we can see that females are earning more compared to males.

* Do women become better leaders as they get older?



* **55+:** females approx.. **13%** increase
* This is the only group that has provided this insight
* Are women better leaders?
* perhaps there is a greater demand for them?

**Overall analysis and next steps**

So its clear that males, overall are still earning more than women. it would be right to assume that as males and females get older in the data and analytics industry, never mind harnham UK, their salaries increase. Education for salary potential seems to be most significant if you’re qualified at the phD level. This may be due to the high-level applicable nature to industry when studying for one.

For next steps:

* Look at number of hours worked per week
* Amount of experience within industry
* 2021 results