Hi everyone, so for the midterm, I thankfully got this Oscars Demographics dataset, and I had a lot of fun tinkering with it. In thik talk, I'm gonna give you a tour of this data, walk you through how I approached it and share with you some of my thoughts, findings, and takeaways over the analysis.

First, let's talk a little bit about the data, it has a fairly small size, about 90 kilobytes in total. So I was quite ambitious at first to try to dig some insights directly from the csv, and it looks something like this. From an overview, it has 441 observations and 27 columns. It was a bit messy, so I turned to JupyterLab to have a closer look at things.

I loaded the data, and printed the names of the 27 columns out. Though it sounds like a lot, there weren't actually many usable data in these columns, in that a couple columns were either missing data or just providing duplicated information in different ways. But basically, what I'm getting at is, this dataset gives a good chunk of information about the Oscars Award winners, the name of the movies, their awards by category, their race and religion, their sexual orientation, their date of birth, their birth places, and the year they got the award, and a column indicates whether they've also got the golden globe award, a url to their bios on the Notable Names Database,. As you've probably noticed, there aren't many numerical data, most of the them is categorical, more specifically, nominal data, and the reason why I still put so many variables here as continuous data, is that many columns were about the confidence of a specific column.

After the first sight, some questions that came into my mind are:
How are these winners distributed by their race & ethnicity?
…
With these questions in mind, I started to look into the data, and did some preliminary data preparation and cleaning work. I removed all these extraneous columns and only left those that were of my interest to answer these questions. And I found that there were actually some duplicates in the data, as shown here, so I used the drop_duplicates method to clean that up. After I did this, there were 336 observations left compared to the original 441.

Then I was ready to dig deeper into the data to answer the question about its race/ ethnicity diversity. I used the value_counts method to get this information, and after I plotted it with a pie chart, I was sadly to find the result was super skewed, more than 90% of these winners were white. As the Oscars increasingly become an international event among the world's film industry, it is very disappointing to find this lack of racial diversity in 2019.
Beyond this frustration, I was wondering how was the gender diversity though? I did pretty much the same thing with the award column, this time I plotted the data with a bar chart. However, this chart doesn't tell that much information, because there would be awards for actor and actresses anyway, and theoretically they should generally look the same. This guess was confirmed after I combined the best actor and supporting actor, and best actress and supporting actress, as shown in this chart. However, what would really be interesting to look into and is not there in this dataset is, what's the gender diversity for the Oscars best directors award? As curious as I was,

I searched this on Google, I again, sadly found that, this has been a heavily male-dominated field, There has only ever been one female winner r for best director in the history of the Academy Awards – and that is not going to change in 2019. (Kathryn Bigelow)

Well, so far the data tells a sad and biased story, let's take a look at the other questions that are on my list. Given the birth place and birthdates information, I was curious about where were most of these winners from and which month were most of them born in? I imported the wordcloud module, and generated the top 30 words from the birthplace column. There we go,not very surprisingly, as we can see, the most salient ones are …., metropolitan and traditionally diverse and artistic cities. On a related note, I was also curious if there was any pattern in terms of the month in which most of these winners were born in. So I extracted the month from the birth date column with the help of the DatetimeIndex method, I made it a separate column, and plotted it with seaborn's catplot function.Generally this distribution is fairly balanced, April and December seem to be the most creative months though.

The last question I was interested is What is the chance to win both the Oscars and the Golden Globe Award. To answer that, I looked into the column called golden, which indicates whether a movie also won the golden globe award with boolean values true and false, since the possibility of a movie winning the golden globe and the Oscars award are independent, I divided the total number of golden globe winners by the whole , and the possibility of winning both awards is about 1.49%.

Finally, I wanna share with you some of my takeaways out of this exploratory data analysis to close this presentation.
As someone who works on data visualization, I feel like the way you approach a dataset is somewhat similar to this mantra for visual information-seeking. You start with the overview, take a look at your data, get a sense of its size, data types, variables, etc. Then you zoom into the dataset, and have closer look at things that are of interest, filtering through different information.

My second takeaway here is, high quality data is the key, i learned this one the hard way, I was at first really interested in looking at the age that most winners won their first Oscars, and from there, maybe I will be able to identify a certain age that people tend to be most successful or something. However, after spending hours working with the date-of-birth column, I figured this dataset cannot be used to do that, because there were mixed formats for their year of birth and were not able to be parsed by python.

My last takeway is to use visualization aross data analyses, and more importantly, to map the visualization types to the tasks accordingly. For example, in this case, we know pie chart is good at conveying the part of whole relationship, and bar chart works well for categorical data, when plotting the racial and ethnicity information with a pie, we can intuitively see how biased the data is.