This is my portfolio, and it is my first solo project within data analysis, something I was quite excited to undertake. I had to find a dataset I had interest in and explore this dataset in any way I see fit, while applying the skills I learned during the level 4 Cambridge Spark bootcamp. During the 10-week bootcamp, I learned many skills including Python, Pandas (including processes within functions, data cleaning and aggregation and visualisation), SQL, maths and normal distribution within data analysis as well as participating in two mini group projects involving exploring and drawing conclusions from a dataset and presenting the findings to a group. I chose to upload my portfolio to GitHub rather than doing a PowerPoint presentation as I preferred the freedom I have with GitHub in terms of adding to my portfolio.

In my portfolio, I chose to find a dataset related to representation in gaming, whether it was gender or race. I am a fan of games and I am aware the disparity in the characters we see and many discriminatory controversies that arise when characters present in a way that certain gamers do not like. As I found a dataset on Kaggle relating to gender within gaming, I wanted to see the change in gender over the years, if any. I wanted to see if there are more female characters as they years go by as well as women in the developmental team. I also wanted the amount of sexualisation in female characters. I have been wanting to create a portfolio to show and use my newly acquired skills since I chose to pivot into data analysis and this bootcamp gave me the support and space to do so. I hope to one day be able to use these skills I’ve learned in this bootcamp in the data analysis field, hopefully within my areas of interest such the gaming industry, sports, psychology, or healthcare.

First, when looking for data, I searched within GOV.UK and ONS. However, despite being interested in health, gaming is something I wanted to look into more; and I was suggested to use Kaggle. I had in mind to use SQL queries to show what I’ve learned, but I realised I wouldn’t be getting my data from a database and chose to filter results by .csv files. After finding this dataset, I saw the information it provided and decided I would go ahead with working with this data. I wrote down initial things I wanted to find and do and followed that plan. I used Pandas as this lets me use functions and keywords that I’m familiar with due to this bootcamp; these included making plots, renaming columns and values, grouping values by a specific column and more. I also used masks to filter my data to find the insights I needed. I have included explanations in the Jupyter notebook explaining what I did and why during each set of codes. My hypotheses were:

1. There will be fewer female characters in games than male overall, and it will increase over the years.
2. There will be less women as part of the developmental team, and it will increase over the years.
3. There will be a difference in the relevant characters female characters play over the years.
4. There will be high amounts of sexualisation of the female characters.
5. There will be difference in which genres female characters show up more in.

I found most of the time, there are a low number of women that are on the developmental team of games and, almost every year the most occurring percentage (and therefore amount) of women that are on the team is zero. I’ve also found despite there usually being fewer female characters in games, it seems to be slowly increasing over the years. I also found much less sexualisation of female characters than I expected. As research has found diverse settings tend to fare better than non-diverse settings, what I found about less women and often no women being included in the development team of games and be applied to a real life setting by increasing the number of women in the setting for different perspectives. As around 40 - 50% of gamers are women (depending on the region), having female characters in games is something that is appreciated. During my dissertation a few years ago, I found some women who like to play games wouldn’t identity themselves as gamers due to stigma of what games a gamer should play; I believe this research as well as the insights I’ve drawn and could find in the future upon continuing to explore this data can be used in the making of future games.

In terms of what went well, I created a plan of what topic I wanted to study and what I wanted to find, and I followed that plan, while tweaking where I needed to. I also used a wide range of things I learned during the bootcamp and things I learned when searching how to complete an action; the knowledge I gained from the bootcamp helped me apply them to my work. I was also able to understand errors and how to fix them. I was also able to showcase the information I wanted to find. It was exciting for me as not only was I able to explore this dataset, being able to find and apply different processes within Pandas to simplify actions I needed to explore my data and find actionable insights was very exciting and further showed me this is something I want to do as a career.

I did struggle with whether I should remove outliers as I didn’t want my data to be skewed. I decided to keep any outliers and take out the calculations I did for said outliers as I felt any outliers were true outliers and they didn’t seem to be outliers due to error. Therefore, if I had the opportunity, I would further examine the data that has a column which states when the game began its development. This is as games can take anywhere from 2/3 years and often more, such as seven years, to finally get released. So even if two games were released in the same year, one may have begun development seven years prior and the landscape of society and other situations such as funding would have been different and would’ve affected the game in many ways such as the story and dialogue. So, if possible, I would love to find a database with a larger number of games and the initial development year which would make things such as outlier detection and an attempt to find causality much more reliable. With this database, I could also practice my use of SQL. I also want to do hypothesis testing and look at confidence intervals, something I need to brush up on. I also struggled with not knowing how to apply certain functions, such as missing a keyword, but searching on the web helped me to understand, tweak and apply to my own work. At first, I wasn’t making copies of the dataframe, which meant when I would change something and it was a mistake, I would have to re-run the initial cell in which I loaded the dataframe; I soon realised my mistake and made a copy so I would always see my unedited dataframe and not have issues when dealing with a mistake.

Next time, I would remove the columns that I don’t intend to use at all for the current exploration just to streamline what I want to see. I would also like to use Bokeh for visualisation due to the non-static plots and different types of visualisation tools. I would also do reliability and validity testing to discover if there is a link between sexualisation of female characters and if there is a genre (or genres) most associated with it.

I learned just how much time a project can take, and I will still enjoy it if it is about something I have a vested interest in. I also learned new functions and keywords to use when exploring data. It has showed me what I know is enough to gain insights into data and I can and will always continue to learn and I can apply this to current and future projects and hopefully in a working role in this field.