Project Diary

Monday, 5th December 2022

- Created Jupyter Notebook file:
  - `Forecasting the High-Risk Demographics for COVID-19 in London Using Deep Learning.ipynb`
- Sourced data from *UK Government coronavirus data hub*:
  - `cumCasesByPublishDate_2022-12-01.csv`
  - `femaleCases_2022-12-01.csv`
  - `maleCases_2022-12-01.csv`
  - `vaccinationsAgeDemographics_2022-12-01.csv`
- The UK Government data hub allows users to combine datasets; this is a major aid when preparing the datasets that I will be using.
- I encountered my first possible failure of the project; I can't seem to access the datastores on the NHS digital hub that I accessed previously.
  - I believe that the datasets I originally accessed may have been moved to the UK Government coronavirus data hub. I expected most of my data to be sourced from there, so currently this doesn't appear to be a setback for me.
- Sourced data from *London Datastore*:
  - `phe_cases_london_borough_2022-11-30.csv`
  - `phe_healthcare_admissions_age_2022-11-28.csv`
- The hardest challenge today has been deciding on the correct combination for the datasets that are to be used.
- I am considering contacting the relevant bodies at the UK Government to see if I can request specified data packages for my research.

Monday, 19th December 2022

- Made a start on cleaning the data that had been collected.

- I haven't had a response from Kamaran regarding the scope of the data; I shall proceed with the data that has been collected. I still believe there are relationships that I can discover within the data.
- For the male cases and female cases datasets, I have divided them by the age ranges. I feel that by having a more granular look at these datasets I may be able to find something more; the original data still remains intact regardless, so I can still analyse any inter-age relationships in the data.

Tuesday, 20th December 2022

- Collected another dataset, pertaining to the reinfections of coronavirus.

Wednesday, 21st December 2022

- Progress has been made on the cleaning of the data. I have now decided that I will also rename the columns of the initial DataFrames; if I leave the columns with the original names, a lost of context could be lost. If the data lacks context, it essentially becomes unusable because I simply wouldn't understand what I was looking at.
- I have added in markdown cells to give details to each dataset that is being implemented; while I might understand what is going on, someone who is looking at the notebook with fresh eyes might not understand background details about the data.
- I think I might be a little bit ahead of schedule. If that is the case, later this week I should aim to actually examine the data and find ways in which they could be integrated with each other; the aim of this project is to explore how regression models can be used to identify relationships. There will probably be relationships that can be found within single datasets but there could also be

other relationships that could be discovered by combining the datasets dynamically.
- Obsidian has been a useful tool for me, both for noting down my research and for keeping a diary during this project. There has been a new update that has introduced a module called 'Canvas'. This is essentially a dynamic mind map that can combine notes within itself. I have decided to implement this into my work scheme, as a way to have a clean and concise way to display an overview of the project.
- I think I will implement the use of RDDs to create some of the dynamic DataFrames. The sheer size of some of the datasets would make it cumbersome to manually create some of the datasets.

Thursday, 22nd December 2022

- As of this morning, the datasets that have been sufficiently prepared are:
  - Male Cases Dataset
  - Female Cases Dataset
  - Admissions Dataset
- The Cases Dataset still needs to broken down into the individual boroughs.
- The Vaccinations Dataset still needs to be broken down; I could potentially break this dataset down into the boroughs and also the areas within the boroughs. This would probably require the use of RDDs.
  - I have a feeling that the current form of the dataset still has use. I could use a string indexer on the borough and area columns prior to processing the data.
- The Vaccinations and Cases Dataset and the Reinfections Dataset need to be cleaned.
- The Vaccinations and Cases Dataset is going to be unusable; the data hasn't formatted correctly, giving undefined values for the columns related to vaccinations. This

shouldn't be much of an issue, as I should be able to reconstruct a facsimile of this dataset by combining the other datasets dynamically.
- I have received confirmation that I can 'redirect the focus of the project around technical issues'. Since this does pertain to big data, it is imperative that I discuss how there was an issue with the data, what the issue was, and how it could be solved.
- Along the same grain as the refocusing, I should reconsider whether I want to progress onto building a predictive model. I really think there is potential for it with the data that I have, as well as the ever-growing bank of data on the various sources.

Monday, 26th December 2022

- I have realised that something that has a lot of merit to talk about is the struggles of data collection direct from the source; the datastore offers a service that combines the datasets but it isn't always successful. Certain values can often be entered as null values and so the datasets become unusable and uncleanable.
- I might consider adding a week onto the data integration stage.
- I have added a dataset that details the tracking of different variants in London. I feel that I might be able to identify trends within this data as well.
- I am about to actually start combining the data into smaller DataFrames. I need to decide how they can be combined.
- The DataFrames have been converted. Now I just need to decide what DataFrames I want to combine. It would be amazing if they could all be combined into a single unity but I'm not currently confident as to whether that is possible.

Tuesday, 27th December 2022

- I have asked to see if I can steer towards this being a Big Data style project. I simply cannot see how I can do a full-blown machine learning project with the data available. I can **absolutely** see how I can do a Big Data style project; the data is rich with stats but none that will be able to show relationships. For that kind of data, I would probably need access to data that shows prevalence of symptoms, infection status, and mortality. For some reason, these are not available to the public. If everything goes as hoped, it is crucial that I speak about the struggles with the data, how certain data could be used for machine learning, and a little about the Five V's of the potential data.
- I did receive confirmation that I can change towards Big Data; at the least, that was the implication. Regardless, this is a massive boon for the project.
- I have also received confirmation that I can now change to using data from all over the UK. This is huge because I will be able to really showcase a 'big' scale of data. I shall collect data from pretty much anywhere. Hopefully, they share the same format. If they don't it is just another thing to discuss.
- As of right now, I don't think I am particularly set back. Will there be a restructure to my project timeline? Absolutely. Do I see it as a problem? Absolutely not. Why? I planned ahead for any possible time delays and decided to leave myself with, I think, a six week surplus. This is not a change that will take six extra weeks. In fact, I might use even less time. I am way more confident in the realm of Big Data, and frankly, more interested. A higher level of interest will keep me more interested in working at it, and with my nigh-unhealthy approach to work, I could end up expediting my work rate.
- I **must** make sure that I create a new Gantt chart. One thing that I have really got to understand in the short amount of time I have worked on this project is that Gantt charts and

similar time management diagrams are actually priceless; I know exactly what I should be doing at any one time, and I actually start to memorise this. This shouldn't surprise me given how this is the same effect my self-made gym schedule has caused, where I know pretty much every exercise is should be doing, down to the weights.

- I think I need to become a bit more organised with how I contact Kamaran. My problem is that I try to be too efficient with how I contact him; I realise something that needs clarifying, quickly write out an email, and send it. Sometimes, just minutes later, I will realise that there is something else, and I have to either try and recall the email or I just have to send another. Right now in life, it isn't the end of the world but bombarding someone with emails is rather bad etiquette for a future workplace. I must try to clean up this behaviour.
    - It could be said that having scheduled meetings could solve this but as I might have said before, if I cannot come to a meeting with a set of questions to ask Kamaran, I am simply insulting his time. **Never have a meeting without prior cause**. I have to be able to go in with questions, and leave with both answers **with** planned solutions. So, since I don't have any questions that warrant a meeting, I won't ask for a meeting just yet.
- Reading back over the diary, I see that early on I spoke about contacting the NHS datastore. This could still be a possibility for perhaps some free-time but it no longer feels anywhere near as necessary.
- I have inquired to see whether I should include this entire diary within my final report. I feel that it could be of value regarding a history of my personal learnings and struggles throughout but it is entirely possible that it would be fruitless ergo I have inquired.
    - In the event that I can, and thus someone is seeing this, I will have cleaned this up. Not a chance I will enter a crude and rudimentary format for this diary; purely professional here.

- Next week, I shall have to acquire some research papers to read through. I really hope there are some good ones. From a quick search, I think there are. The real task is about doing this in a way 'never done before'. I really hope I succeed in that, I don't want to be re-treading old steps.

Thursday, 29th December 2022

- I am going to collect all new datasets. I will collect datasets covering the whole of the UK and I will aim to make sure I log exactly where they came from, for the sake of replication. This should only take about an hour, most of the code that has already been written translates to these new datasets, due to the formatting. Perhaps I should speak about the formatting in my report.
- `generalCases_2022-12-22.csv`
- `maleCases_2022-12-22`
- `femaleCases_2022-12-22`
- `maleDeaths_2022-12-22`
- `femaleDeaths_2022-12-22`
- `variants_2022-12-22`
- `vaccinationDemographics_2022-12-22`

Monday, 2nd January 2023

- There are still some DataFrames that are in need of cleaning:
  - The `generalCases` DataFrame.
  - The `vaccinationDemographics` DataFrame.
- At the moment, I believe that once the cleaning is finished, I should be onto the next step of the project (reading through research).
- I must make sure that I do actually rebuild the Gantt chart. This can probably be done today, provided I keep focus.

- I will probably add a branch to the project canvas, where I will jot down any analysis ideas that come to mind; over the past couple of weeks, I have had ideas that have been lost due to not noting them down.
- For the `generalCases` DataFrame, I had to remove some null values that were appearing. I knew how to do it with `df.dropna()` but I thought I might have to do it another way to not mess up the data; I didn't. I have decided to implement this for **every** DataFrame, as it will add to the insurance of cleanliness in all of the data.
- I might have forgot to note this down at the time of doing so but I had to manually delete columns from the `vaccinationDemographics` dataset `csv` file because of the sheer size of it. The columns that are still being used are still being cleaned within the notebook.
- While cleaning the `vaccinationDemographics` DataFrame, I have realised that the columns that pertain to people completely vaccinated mirrors the column for people who have received their second vaccine; they are one and the same. For the sake of data capacity, I am removing the former mentioned columns.
- All of the DataFrames have now been cleaned. I shall work on making the new Gantt chart and it will naturally include some changes. A notable one is that I will start combining the DataFrames after I have researched the methods that I will use. There is no use in me combining DataFrames to find out that some data shouldn't be combined with others and should be with another set, etc. This will also allow me some time to work on my other outstanding assignment for another module (Graphics).
- I noticed I don't have a DataFrame for general deaths, so I have implemented one.
  - *`generalDeaths`*
- The current volumes of each of the DataFrames are:
  - `maleCases` - 176054 rows × 6 columns = 1056324 cells
  - `femaleCases` - 176054 rows × 6 columns = 1056324 cells
  - `maleDeaths` - 171 rows × 6 columns = 1026 cells

> - `femaleDeaths` - 171 rows × 6 columns = 1026 cells
> - `generalCases` - 8775 rows × 5 columns = 43875 cells
> - `generalDeaths` - 9070 rows × 5 columns = 45350 cells
> - `vaccinationDemographics` - 133920 rows × 11 columns = 1473120 cells
> - `reinfections` - 196455 rows × 7 columns = 1375185 cells
> - `variants` - 8460 rows × 6 columns = 50760 cells

Sunday, 22nd January 2023

- External assignments for the first semester have been completed. It is now the exam week but there are no exams. I still need to remake a Gantt chart. That can be accomplished tomorrow, if necessary; looking back at the current Gantt chart, I might not need to update but I will still reassess.
- The week beginning tomorrow will be mostly dominated by completing my research. I have lagged behind on research, in so far as I haven't spent four weeks doing it, but I might actually be able to complete it within the week. The best part of my time plan is that I was quite generous with certain tasks, so I might be fine.
- I really want to get a big grasp on just how I am going to be examining this data. It is all well and good talking about "I'm going to analyse it" but **how**?
- I have made a start on writing the interim report since it better fits the style of my writing; I prefer to write sections of reports **as I am doing them**.
- I have written the introduction, which was quite beneficial as it needed a rewrite due to the refocus of the project. I found it even more helpful because it allowed me to actually sit down and properly think about where I want the project to go from here; past this report, I am locked in for the long road.
- I am including in the literature review a set of questions that will lead the research. I am doing this as I feel it

will keep me on track. I can quite easily rack up a ton of ideas on how I could do things; that simply isn't helpful when trying to complete an aimed project. By having a set of parameters to keep me on track, I shouldn't wander and I will actually have a solid reference as to what I am looking for.

- I am done writing for the night, finished before the *Previous Studies* section. Tomorrow is where the proper research begins. I don't foresee this being too difficult a task now, due to my research objectives. I have narrowed it down to cover a juicy area without being bombastic.

Monday, 23rd January 2023

- I have started my research, and I have found some pretty good papers that are very heavily detailed in their methodology. This is a major boon as I was worried that I would be left with a piddly amount.
- I might do the Gantt chart today, depending on how I feel, and I also need to update my canvas spreadsheet; not only will the canvas aid in my tracking but it could be a really nice addition to the progress update.
- I've had a little bit of a panic, concerning the focus of my project. The toughest part so far has been confidence in my idea; I truly enjoy Big Data and also love discussions surrounding COVID-19, yet I keep having these feelings of how there isn't enough data or the idea is weak. I'm fine now but it's a weird anxiety.
- I've started to put together some ideas based on early research:

  > - Examine the data types.
  > - Examine the collection of the data.
  > - Examine the 5 V's of the data.
  > - Investigate gender disparities.
  > - Examine efficacy of vaccinations on the different variants.

> - Predictive model on cases, deaths, and vaccine demographics.

- I have decided to break down the literature review into three overarching segments: data handling, data analysis, machine learning integration. These are the three main steps that play into how my project will play out so it makes sense to draw from papers concerning how I should approach them.

Tuesday, 24th January 2023

- A preliminary draft of the literature review has been completed and sent off to Kamaran for review. I'm actually glad we had to do this as part of the project; I feel as though it has helped give me a bit more direction in where I want to go, but it has also given me a bit more reassurance. I no longer feel as though I need to be producing an equivalent to Hawking or von Neumann. It has also reminded me of the fact that failure to produce isn't necessarily a failure to study; even if I can't produce relationships in the data, I have still made a finding.
- I just need to complete the progress update, which shouldn't be difficult due to this very diary, and then I should be finished with the interim report. I feel that regardless of whether or not I need to change my Gantt chart, I will still recreate it from this point in time. Keeping things fresh can be really helpful, like a reminder.

Wednesday, 25th January 2023

- Almost complete with my interim report and I find out that not only did I not document the origin of the `reinfections` dataset but it isn't actually what I need. Regardless, the link to it is *this*.

Sunday, 29th January 2023

- The datasets for the deaths are complete rubbish. Total incompetence on my end. Only track to the past 28 days. Honestly, I knew this at the time but I was being lazy. Swapping them out right now, then I will finish the alterations.

Monday, 30th January 2023

- I have collected the new dataset for `generalDeaths`. The other datasets for male and female deaths will remain. I forgot to note that they are the only ones available. While it isn't as large a volume as I hoped for, it is what I will have to deal with.
- I have actually found the page that lists the proper documentation for all the datasets. This is massive, since I can now update the interim report with proper descriptions for the harvested datasets.
- I have discovered the original `reinfections` dataset. Super pleased, I didn't want to lose it because it was so large and had really juicy data. I am definitely learning the hard way the importance of proper documentation.
- Reading through the documentation of the `reinfections` dataset has made something clear; there are individual, 5-year age brackets but there are also two separate brackets for 0-59 and 60+. I knew that they wouldn't have that in there for no reason but now I think those two brackets could be more useful than previously understood.
- Discovered, at such a late time, that I was actually over-extending myself with the way I was breaking down the DataFrames. The pandas utility offers the simple dynamic via grouping but I thought that was limited to RDDs. Notably, I spoke about how RDDs could be used. Perhaps I should discuss, later on in the final report, how I might not have actually needed to use them.

- Despite the annoying feeling that I have wasted time, I am now on track to blaze through this current stage of the task. By tomorrow's end, I am certain to have finished the statistical modelling for four of the nine datasets, and who knows how much further I could go. If I finish this stage within the week, I will be amazingly pleased. The regression analysis is going to be the real beast, not only to program but also to discover if it even works. Having an extra week to work on it according to the plan would be huge.
- I also need to make sure that after tomorrow, on Wednesday, I drop some images of the statistical modelling into the interim report. No way am I making this sort of progress and not getting marks for it.

Tuesday, 31st January 2023

- Not currently at my station but I have completed the statistical modelling for region and age distribution. This means I am essentially half way to completing the first four dataset models.

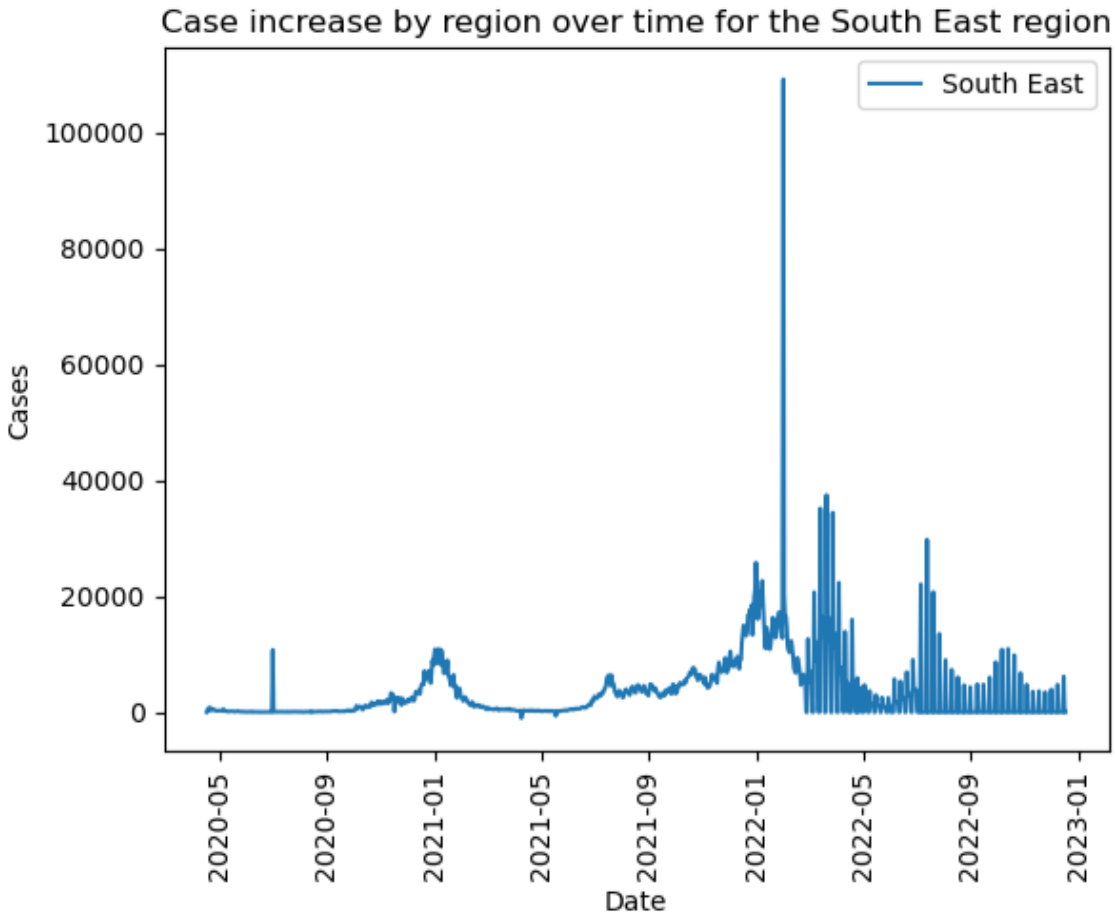
Thursday, 2nd February 2023

- Ignore the date on this entry, it is 1:21 in the morning. I have spent all day finishing off the modelling for the male cases, which will be copied and pasted onto other datasets. I have an idea to add an accompanying table for the pie chart. Super pleased, and it gets better; the models are giving me ideas on possible regression analyses.
- The male and female deaths datasets are rubbish. I completely misunderstood what they were when I collected them; they are a 28 day span, they are the values for the total of the past 28 days. Totally and utterly unusable since there is no time series. This is a complete failure on the collection end, and I should bring that up in my

final report. Seems I will have to do away with them but no worry yet.

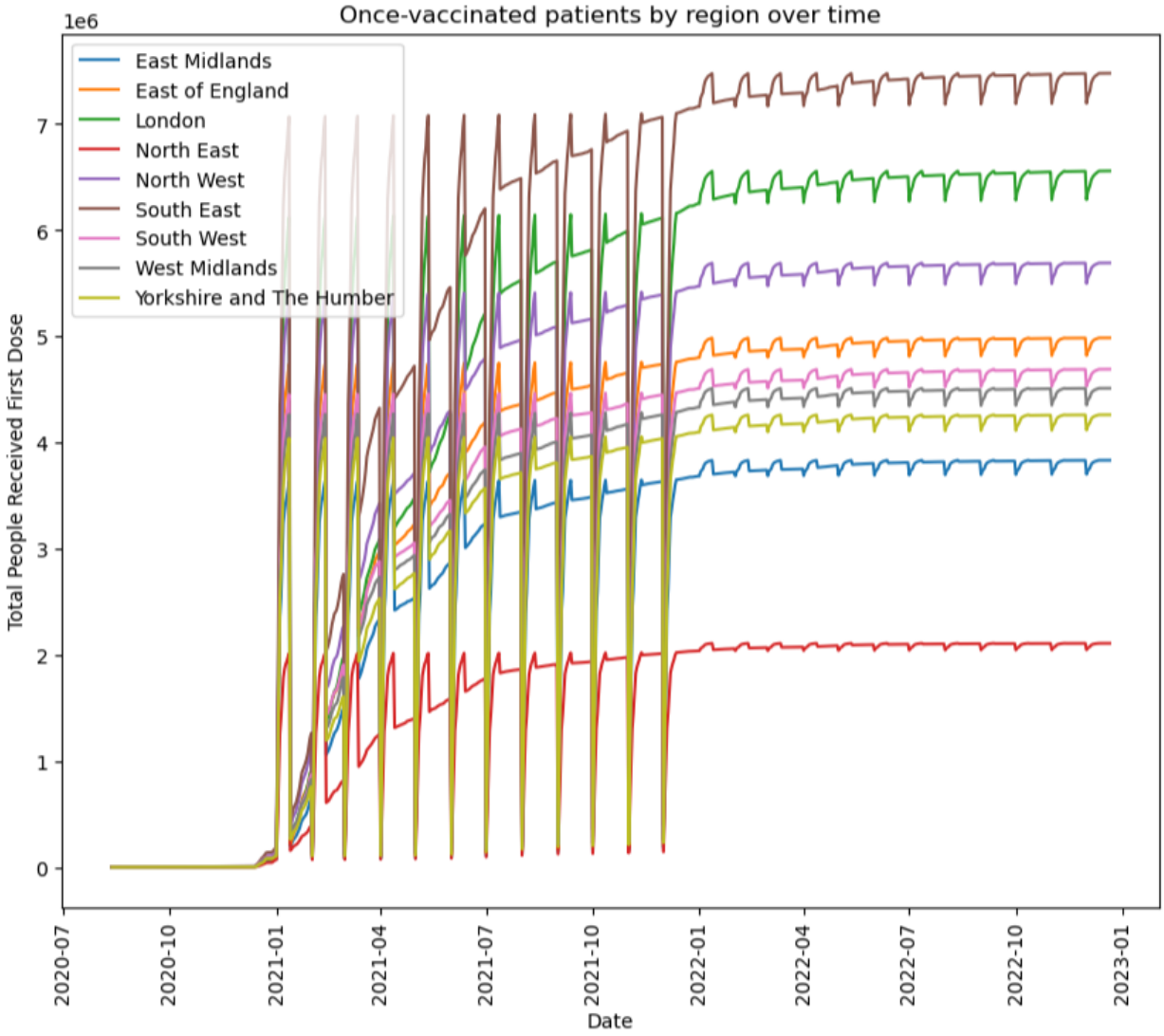
Monday, 6th February 2023

- I am going through the general cases dataset and I am discovering some unusual outputs in the modelling, see this image. Looking at the dataset again, I think this is because of how the dataset is a composite set, and I have had to remove some null values. Since I'm not even using the rate values that I have collected, I should actually get a clean set with just the cases; this will actually allow me to show off more of the usage of PySpark for data calculation.



- The new dataset for general cases was sourced from *here*. I tested it to see if there was any change and there actually wasn't. I'm not sure what it is that causes the strange data patterns.
- I am going through the vaccination data; it is complete trash. The graph was coming out really weird, see this

image. I couldn't understand how values could be jumping around like that so I took a segment of the data (a single region) and went through it. For some reason, at random intervals, the values will plummet and then count up again. The fact that they somewhat increase is what creates this strange trend line through the noise.



- I have an idea as to how I will solve this problem; if I remove all of the rows where the value of the cumulative column drops, I will eradicate the problem. Will I lose data? Yes. Will it be a massive improvement to the readability of the data? Yes. I really don't to lose this data, and it's better to lose a finger than a hand.
- The idea just didn't work. The dataset is completely bunk. I will try *this data* instead.
- The other dataset doesn't have any harvestable data. I need to report on this in the final report. I'm genuinely amazed that the Government of all institutions would collect such a shoddy dataset.

- Stress avoided. The data is clearly not false therefore I shall simply remake the total by remaking the cumulative columns. I feel like I keep saying this but the way in which I averted this catastrophe must be documented.
- Moving onto finish the last two datasets. They shouldn't pose anywhere near as much of a problem.
- I think I will finish the statistical modelling tonight if I push on but I don't think I will jump straight onto the regression analysis tomorrow. I have another assignment that I want to make a start on; a classmate has said it took them three days to complete a full draft and since I have four days remaining this week, I want to get it out of the way. By getting it out of the way, I can be more firmly on track with the project.
- I added a function that formats the age ranges so that they are all the same across all datasets.

Tuesday, 14th February 2023

- I am back to working on the project. I have built the linear regression model to measure the impact of sex, region, and age. I also attempted to build a logistic regression model, however it told me that I lacked the sufficient memory capacity; I have 32GB of RAM. I should talk about this as a limitation in the report.
- I will create unique documentation pages for each prototype I make to keep good track of how they went.
- The first model I made took the Sex, Region, and Age as independent variables, and the Cases as the dependent variables. One Hot Encoding has been used to account for the categorical variables.
- I have gone through the combined male and female cases dataset with different variations of a linear regression model; none of them perform very well, suggesting their is little relationship between any of the dependent variables and the independent variables.

Wednesday, 15th February 2023

- I have been doing some mini-research and it turns out that a low R^2 value isn't necessarily a bad thing, considering the models are being used exploratively and not predictively. I shall consult with Kamaran to see what he thinks about this in Thursday's meeting.
- I think I might have figured out why the models are having crazy results; they're from time series data but I'm not actually incorporating that into the model. I need to factor in the passing of time into the modelling, otherwise the model just gets values that just have a range without context.
- I am working on a method that will allow me to input the time series into the regression as a simple value of 'days since'. It makes sense now why the model wasn't finding any hard relations because it was just getting input that were identical except for a difference in the cases. The sex, region, and age would never actually change.
- The latest model has a much better R^2 value than all of the other models made so far; it is ~0.5 better than the next best model. Based upon the models made so far, I think that there is actually little connection between sex and cases. The introduction of the time series made the massive increase but that should be expected really. I think now I will move onto my next model.
- I am now going to run through different model types on the same premise.

Tuesday, 21st March 2023

- I am back working on the project, after working through other assignments to get a clear schedule; I still have a few minor tweaks to make to one assignment but that can be finished this evening.
- I am back under way with building models, starting with a model to measure the impact of the variants on the cases.

Building this model could be pretty expeditious, given the structure that I can already refer to from my previous model. I also think that this model will produce more interesting results. The sex impact model was never really expected to produce anything ground-breaking, since a virus that has bias for sexes would be wild but I feel it was a success because of how the model results actually reflect that. On the topic of this model though, contextual knowledge of the outbreak suggests that there is a relationship between certain variants and outcomes, it's just a matter of whether the regression modelling used can identify any relationship with the data it is given.

Tuesday, 28th March 2023

- I have discovered that in my regression model implementations, I have been encoding the 'Days' column that I use. This is actually unnecessary and just increases the complexity of the model without any benefit. Luckily, I believe that the model didn't suffer in the results.

Thursday, 30th March 2023

- I have been able to get some pretty interesting model results from the modified variant impact data. I am trying to see if I can pass through both cases and deaths to see if there is a combined effect but I am struggling to get it to work.

Friday, 31st March 2023

- I have finished the modelling for the variant data. Now I am doing modelling with data about vaccines and their impact. This will be the last set of models I do with the data, as it will have been completely harvested by that

point. I might consider, during the final stages of the results to take the 'most successful' models, get some predictions on them with real data that came after December, and see how they fare in prediction.
- I think it's worth discussing why support vector regression seems to take so much longer to train and how that might affect similar scenarios.

Monday, 3rd April 2023

- I have started writing the dissertation but during some minor research for the writing, I have realised that there are some regression models that I have not tried yet. I shall create a few models with these different formats, for all three topics, to see if there is any other potential yield.

Tuesday, 4th April 2023

- I have just learned that it's only linear-based models that can produce beta coefficients. Non-linear models like kNN, decision trees, and neural networks cannot produce beta coefficients. I must, absolutely must, speak about this in the report. I must be able to clearly justify why some forms of regression modelling were chosen but others were not.
- I think I will not pursue any non-linear models due to the complexity of using them.

Thursday, 6th April 2023

- I think I might replace the pie charts with tree maps. They can show the same idea but when you have as many variables as I do, they are so much cleaner. Part of this project is

about optimising user experience with the visualisation, so this is almost a must.

- I have implemented the tree maps and they are so much cleaner. Distribution is so much more clear.