COMP3021

Information Visualization - Written Report

20317937

An introduction to the data and methods

New and cumulative total daily deaths suffered due to COVID-19 from February 29th 2020 to March 29th 2021 in England. This dataset provides the data split into the seven NHS regions within England: South West, East of England, South East, North East and Yorkshire, North West, Midlands and London. This dataset is provided by GOV.UK and does not include the totals presented by the government outside of the dataset, i.e. the dataset presents 91,544 total deaths on the date of 29/03/2021 but on 24/03/21 the government has presented 111K in their daily bulletin. This dataset was collected from https://coronavirus.data.gov.uk/.

All encodings for the visualization were carried out using Rstudio and the R language. A handful of libraries were used to clean and transform the data and also produce the visualizations. They include Rtools, ggplot2, dplyr, reshape2, lubridate and scales.

An analysis of the questions posed from the data, data cleaning and transformation, plotting and presentation

Question 1: How did the deaths differ by region?

This question was selected as it is understandably the most prevalent question for the dataset. Simple as it is, it could provide a valuable insight into managing future pandemics by comparison of the actions taken by each healthcare region.

Data cleaning and transformation:

Data cleaning and transformation began with an initial data frame (Q1) from the dataset to select the NHS Regions and the New Death Total (the number of new deaths each day; this is not cumulative) which was then grouped by the region. In order to get the total deaths over time for the arranged by regions, a mutate was performed on the Deaths to provide a sum and following this a rename of this new column from “sum(new_deaths_total)” to “Total_Deaths” to avoid difficulties in encoding the plot. Finally, deaths were then arranged by the default order for ease of comprehension within the data frame and a select(-c(new_deaths_total) as this column is not needed for the plot. Because there is an entry per day and the temporal factor isn’t

important the data were repeated (and identical) for every day that passed per region, producing hundreds, if not thousands, of unnecessary rows. To correct this, a secondary data frame was created using the results of the first one but only selecting for unique entries by the following code “Q1DF <- unique(Q1)”

Plotting the data:

The plot was created using ggplot and the second data frame (ggplot(Q1DF)). The ggplot aesthetic was then used to select the x and y axes variables as well as (during the declaration of the x variable in this statement) a reorder of the x axis to be “-total deaths” to plot the bars in descending order.

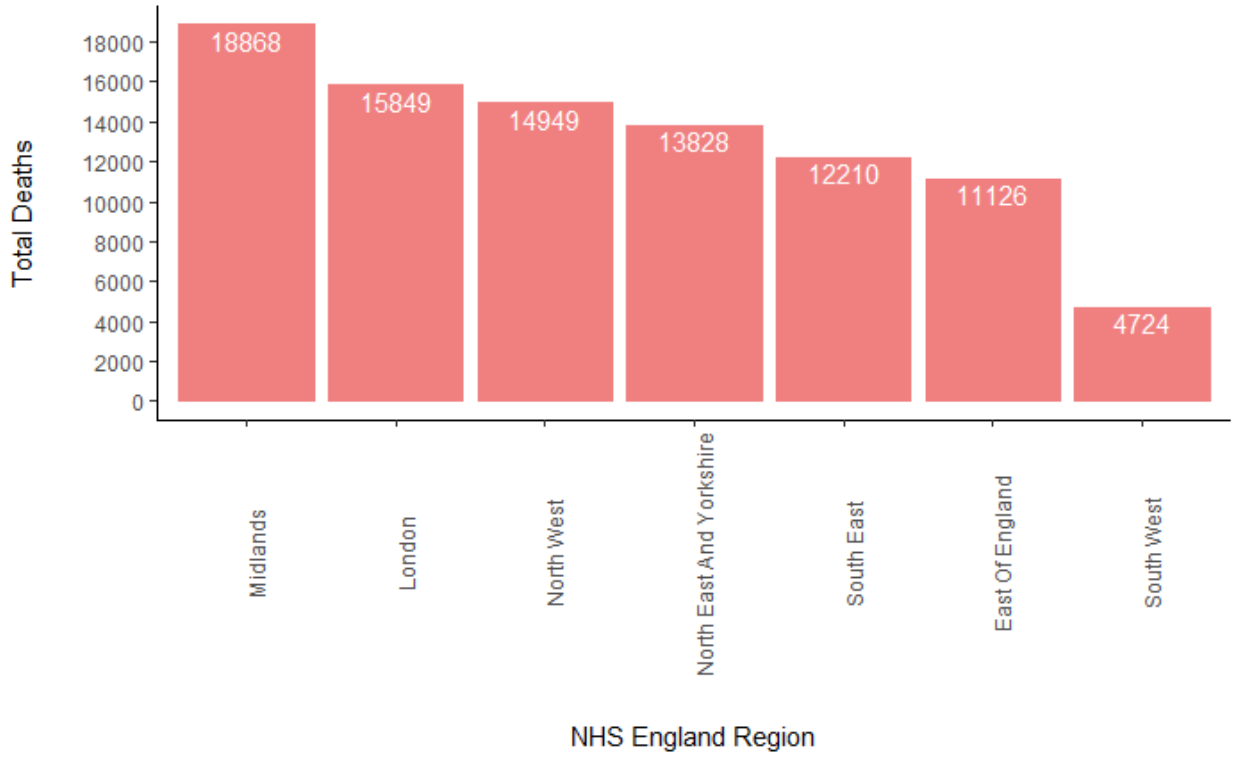
It was then turned into a bar plot using geom_bar where stat=’identity’ was used to assign the variables, and their data, to the axis along with fill and colour were used to make the graph more pleasant to look at.

A theme was then applied to remove the background, minor and major grid elements and apply a clean line for the axes. One example of the removal of a feature is the “panel.grid.major = element_blank()” command. Theme was also used to provide a larger margin of white space between the axes labels and the axes data labels. It was then used a final time to rotate the x axis data labels 90^{O} as they were too long to fit in their standard orientation.

Geom_text was then used to add the aesthetic of data labels on the bars themselves that present the total numbers of deaths. Within this line, to justify the text to fit inside the bars, vjust = 1.2 was used and finally the selection of the colour (white) to be more pleasant to look at and legible.

Finally xlab and ylab were used to replace the default data frame axes titles to more appropriate end-product visualization titles.

The final plot:



I chose to produce a bar plot for this visualization as it is a simple plot of categorical/nominal data against ordinal data. I spent some time exploring different arrangements of the bars for the plot and came to the conclusion the descending order of the bars produced this quickest cognitive response from the user.

In addition, I did a short survey on potential colours for the graph and “lightcoral” was decidedly the most popular, despite the seriousness of the topic people did not feedback that it felt inappropriate but instead that it helped them comprehend the data more quickly. I chose to use data labels on the bars as, although I could have increased the number of y axis ticks, it is the most efficient and effective method to display the total deaths. This allows for a quicker comprehension of the visualization and the data it portrays and for a quicker analysis from the user of between-group differences.

No overall graph title was used for this deliberately, or for any graph. Following an undergrad in life sciences, for scientific writing it is not necessary or appropriate. Similarly, for visualizations like these, background elements, such as grid lines, offer zero value and are, more often than not, removed. As such, for this question and others, I chose to remove them. A cleaner visualization means the user is able to focus very specifically on what I am intending to show them and leads to a much faster and cleaner comprehension.

Question 2: How did the total deaths differ, by region, with and without a positive COVID test?

I selected this question as it piqued my interest to analyse the efficacy of testing carried out by the UK government and the NHS. I.e. if the number of deaths with a negative test were similar to, or disproportionately higher than, deaths with a positive test then a conclusion that the systems and/or tests put in place were failing could be formed.

Data cleaning and transformation:

The data frame Q2 was made by using a select for the data set and the NHS Region. Both deaths with a positive test and deaths with a negative test were grouped by the NHS region. Two independent mutates were carried out on both the deaths with a positive and negative COVID test to provide a sum of the total deaths; i.e. mutate (sum(new_deaths_with_positive_test)). They were then renamed to Positive_Test_DEATHS and Negative_Test_DEATHS respectively to make it easier when encoding the plot. The original, unmutated, columns for positive and negative deaths were removed using select(-c…) as they are not needed for the visualization. Finally, similar to Question 1, it was necessary to remove thousands of repeated rows so a secondary data frame was created using Q2DF <- unique(Q2).

Plotting the data:

This plot began more complex than the first plot as a gather(similar to melt) command was used on the Q2DF. This was necessary to be able to stack the two x variables on top of one another (or next to one another) and additionally produce a key for the plot. This produced a third and final dataframe “Q2DF_plotdata” that collected both deaths with a positive and negative test as one variable now named “Deaths”, described on the gather command line as the “value”.

Similarly to Question 1, ggplot was used on the data frame (Q2DF_plot2 <- ggplot(Q2DF_plotdata)) and aes was used to select the x and y variables, where y is “Deaths”, detailed above, rather than the variables from the original data set/frame. This same aes command also selected for the colours to be used, their fill colour and the act of grouping them together on the plot. Again, when declaring the x variable a reorder was used (aes(x=reorder(x, -y) to arrange the data in descending order.

Geom_bar was used next to define it as a bar plot. Stat=’identity’ was used once again to automatically assign the variables to their axes and position=’stack’ was used to plot the bars on top of one another.

In the exact same way as Question 1, theme was used to remove superfluous visualization elements, such as the background, add additional margin space and rotate the x axis data labels; for the same reason. Xlab and ylab were used in the same way once again to provide more appropriate axes labels.

Finally, additional axis breaks were added on the y axis to provide a better experience for the user in comprehension of what values were being displayed by the bars; in lieu of data labels on the bar for this visualization. This is achieved by overwriting the default y axis breaks assigned by ggplot and instead using a sequence(seq) from min to max value (0, 18000) to provide the scale and by = 2000 to provide the increments.

The final plot:

Once again, I chose a bar plot as it is the most effective way of displaying nominal data against ordinal data. I decided to “stack” the bars to conform with the scientific norm with this data type. Originally the bars were “dodge” so that they were side by side and data labels were assigned on each bar to give totals on the visualization. After long consideration I concluded this was not as effective in displaying both the y variables at the same time and also when it came to cognizance of their values and a comparison of between groups. By stacking them it results in a comparison of two *variables as one variable* and extrapolated to between group comparisons. By having them side by side, the user was having to do a much longer comparison of each individual variable (positive and negative test result deaths) per region and then trying to compare them between groups - a comparison of 14 bars versus 7. In general, then, the stacked version of this bar plot leads to significantly quicker user comprehension and a much more streamlined comparison of between group (region) data.

Data labels on the bars were removed as they became messy when combined with the stacked columns and their ultimate values are less significant in answering the question than the visual

elements presented. In order to compensate for this, the y axis ticks were increased from increments of 5,000 to 2,000 to allow the user, if they seek it, to look more closely into the approximate numbers for comparison. This number allows for a glance at the axis to draw approximate numbers for each of the three variables respectively.

Neutral colours were used to not inform the users cognition of a “positive” or “negative” bias to one or the other bar and vibrant colours were used to pique the interest of the user. A short survey again revealed the colours used were not considered inappropriate given the severity of the data being displayed. Once more I opted to use a deliberately clean visualization style so that nothing is detracted from the user’s comprehension of the visualization and it remains focused on only what I want the user to focus on.

The data were arranged in descending order to allow a quicker comprehension of the worst affected, and possibly worst managed, NHS regions and then compare efficiently the differences in negative test result deaths between the most and least affected regions. A comparison and some conclusions can be drawn here, however correlation does not mean causation and further studies would need to be carried out to quantify hypothetical conclusions - although these visualisations are effective in raising questions.

Question 3: A cumulative look at deaths over time

This question was selected to look at the steepest and shallowest regions of cumulative deaths to assess potential causes of increased death rates during the worst affected times.

Data cleaning and transformation:

The data cleaning was comparatively more simple for this question. The Q3 data frame was created by selecting the data set, date and cumulative total deaths. The regions were grouped together by date to remove superfluous data entries of each day per region and then summed together using a summarize(sum(cumulative_deaths_total)) command. Finally this transformed column was renamed to CTD from “sum(cumulative_deaths_total)” to remove difficulties during encoding the plot.

Plotting the data:

Using the Q3 data frame ggplot was used with the aes of “(date, CTD)” to select the variables. Geom_line was then used to define it as a line graph. At this point, size was set to 1 to dictate the thickness of the line and its colour to lightcoral to fit the theme of other visualizations. Geom_area was used next to fill the graph below the line. The fill colour was selected as lightcoral to match the line and the alpha was set as 0.1.

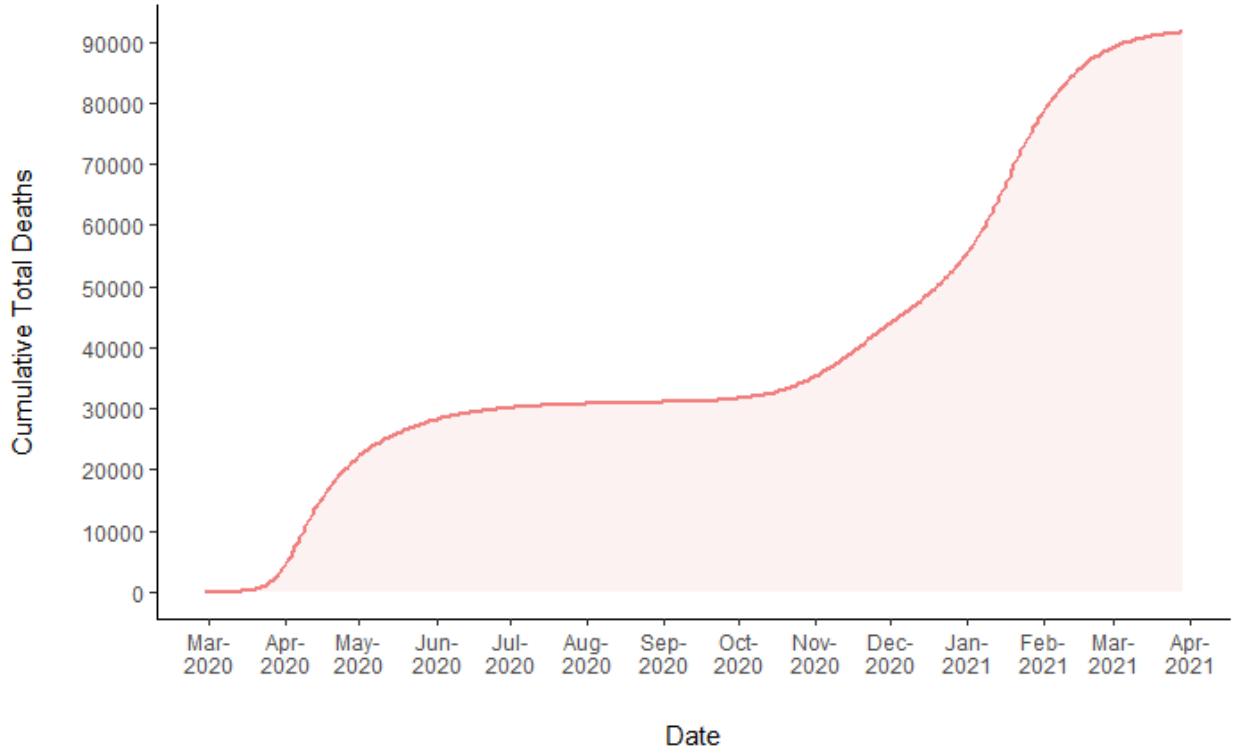
Once more, the theme was used to remove superfluous elements of the visualization by assigning elements such as panel.grid.major and panel.background as element_blank(). In the same way as before, theme was also used to apply a larger margin between axis title labels and

the data break labels. Xlab and ylab were used once more to assign appropriate x and y axes titles.

Scale_x_date was used during this encoding to define the date break amounts and also what data to display on the x axis as the tick labels. %b-\n%Y produces a label with the month and then the year on a new line; the day of the month was not relevant or fitting for this scale. Simply, date_breaks = “1 month” allowed for easy breaks of one month at a time.

scale _y_continuous was used in a similar way to Question 2 to replace the default y scaling of break labels. A sequence within the data range was selected for the scale and was sorted by intervals of 10,000.

The final plot:



A line graph was selected as it is the most efficient way to plot nominal data (cumulative deaths) against a temporal variable (date). *This is the reasoning behind the choices for line graphs for all of the following questions*. The choice to fill below the line was twofold. Firstly, it fits the norm of this scientific data type. Secondly, by removing superfluous design elements such as grid lines and a background, not only was the line dull to look at but interestingly led to a slower comprehension time upon feedback.

Grouping the data on the x axis by month, with the assistance of the lubridate library, was the most sensible because it provided a deep insight into deaths from the overall time scale without being either overloaded or uninformative.

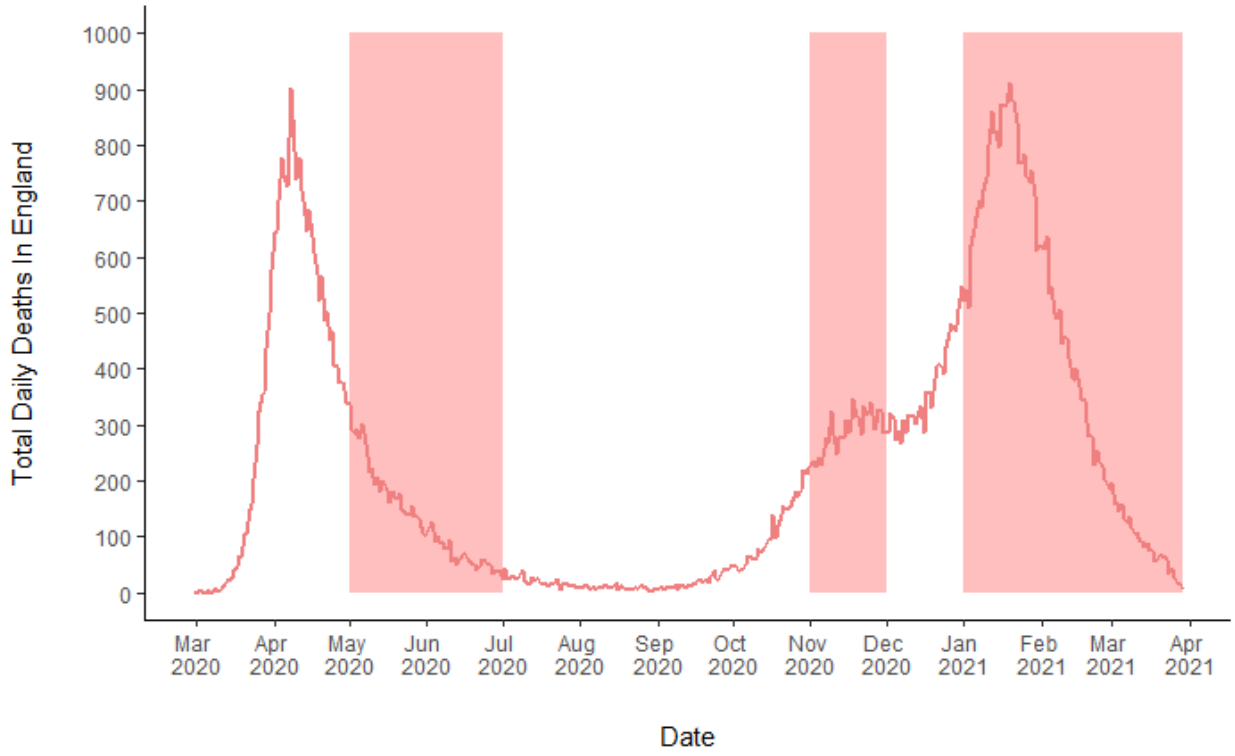
Again, the colour was chosen as lightcoral to be neutral informing bias and was found to be visually engaging without being insensitive giving the severity of the data set.

Refining the questions with the methods to achieve these plots

Question 4: How did the national lockdowns in England affect the cumulative total deaths?

This question was derived from Question 3. We can see rise and fall in the rates of deaths but the plot doesn't infer any causation. By applying some refined visuals to the cumulative graph, we can start to assess the efficacy of lockdowns and if they were a driver for reduced death rates.

Once I produced this plot, I realized that the dates of lockdowns against the cumulative data didn't show their efficacy effectively as it required the viewer to extrapolate death rates from the steepness of the line which is a more qualitative conclusion to draw than intended. In light of this, the data were cleaned in the exact same way as for Question 3 but daily deaths were summarized and grouped by date instead of cumulative deaths. The plot, which used simple geom_rect commands to add the highlighted areas based off of the data frame date values, follows below:



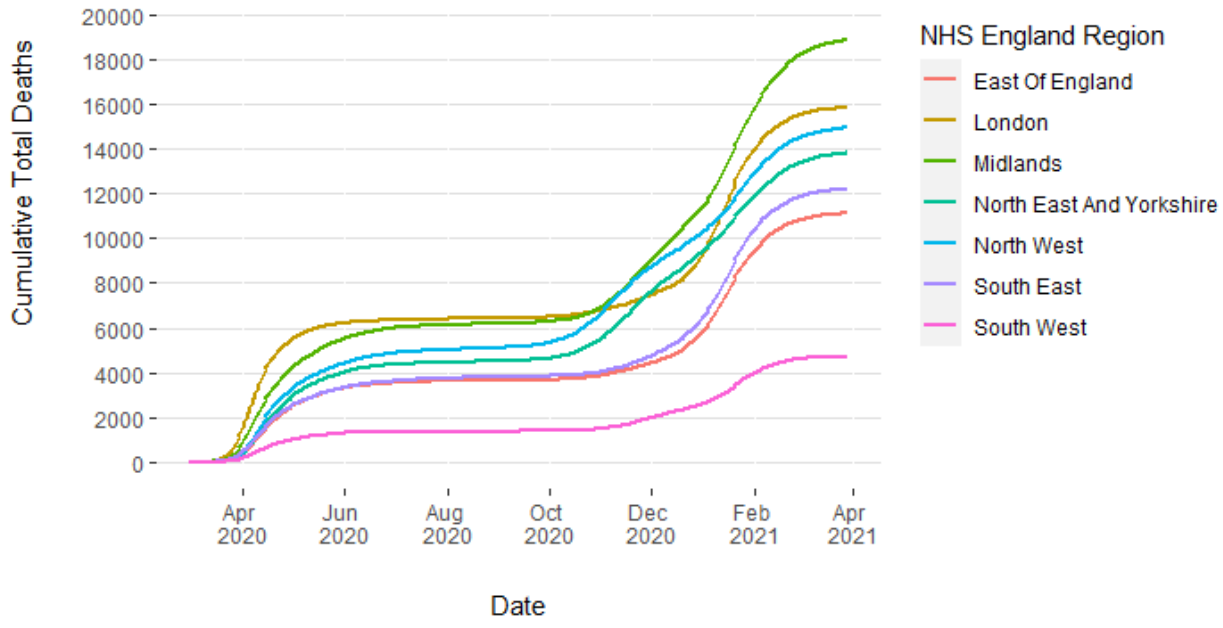
The highlighted areas now show the dates at which England was in a national lockdown and, although further analysis is necessary to form quantifiable and conclusive results, hypotheses can be drawn to the efficacy of a lockdown. Thanks to this refined visualization it can be seen that the first lockdown, though deaths were in steady decline by this point, continued to decline and remain low. For the second lockdown it appears it brought a prompt halt to what looked to be an exponential increase to daily deaths before it; and you can see a sharp rise shortly after the lockdown ended. Finally the third and final lockdown began during a sharp incline of daily deaths and brought about a swift and effective decrease in daily deaths - though this final conclusion is to be taken with a pinch of salt as vaccines also began their rollout.

Question 5: How did deaths vary over time by region

This question follows up both Question 1 and 3. This allows a deeper insight into times at which each region suffered their worst, and succeeded with their best, responses to COVID for further analysis - using this for future preventative measures for pandemics.

These data were cleaned and transformed in the same way described for Question 3, a select, group by and summarize sum were performed and finished with a rename of the sum column for easier plot encoding. The key difference was that three variables were selected for, not two.

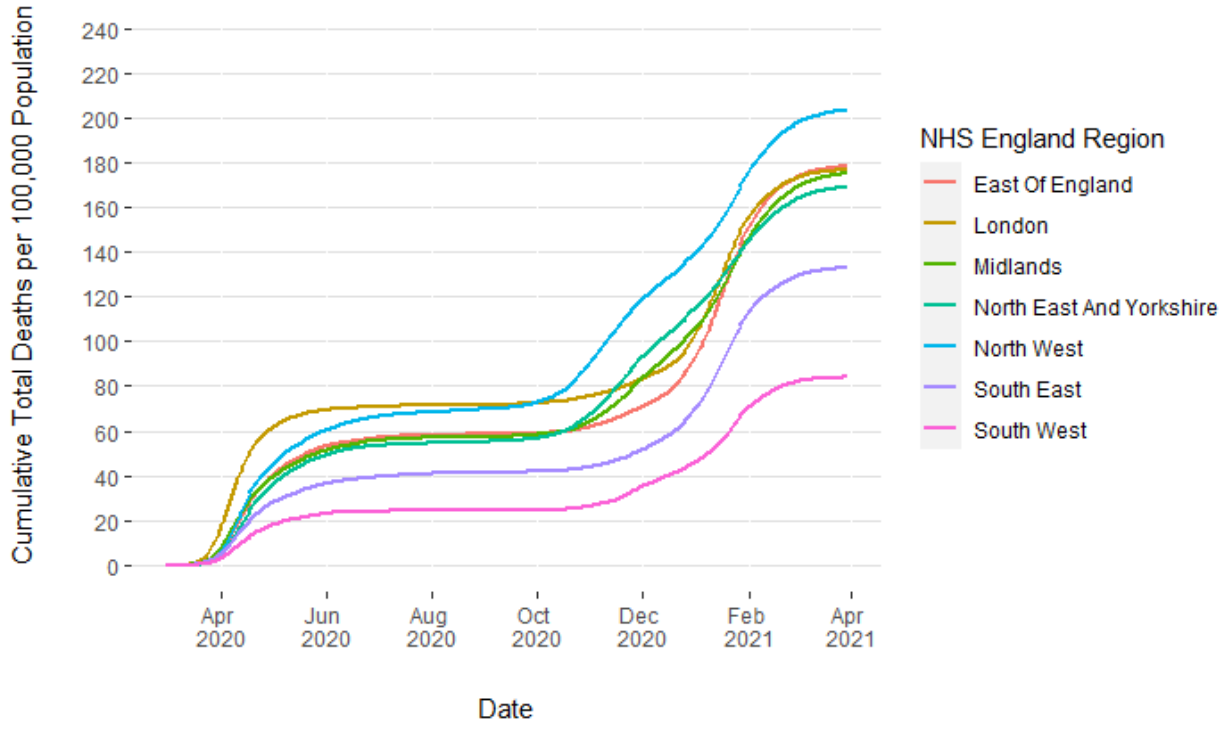
The plot was encoded in an almost identical way as Question 3, also. The key difference is within the geom_line command. As this plot required multiple lines and were of a variable not displayed on the x or y axes, geom_line(aes(colour=nhs_england_region) was used to select the variables to be displayed and to be differentiated by colour. Scale of x and y commands were used to change the data label intervals on both axes and also labs(col=) was used to change the key label pertaining to the colour coded variable. Unlike the previous plots, a graphical element was kept in order to assist the user's cognition. The horizontal lines were kept using panel.grid.major.y=element_line(), at the same time other elements were removed, so that when the user looks to the final point of the graph then can follow a line back across to the y axis and form an estimate on what the total figure was. No labels were included to provide this total death figures as this was a comparison between regions and this superfluous information would detract the users focus away from the question's specificity. The plot follows below:



Question 6: How did deaths vary over time by region, per 100,000 population.

Question 6 is a follow up to the revised Question 5 and is potentially a more conclusive answer to the question of whether the NHS region's total population is responsible for increased daily death rates.

In order to encode for this plot it was necessary to add to the original data set. Both new total deaths and cumulative deaths per day per 100,000 population were calculated after researching the total *estimated* populations within each of the seven NHS regions in England. Otherwise, the data were cleaned and transformed in the same way as Question 5, but selected for, and summarized by, the new cumulative deaths per 100,000 variable. The encoding for this plot is identical to Question 5, save for the variable selected to be shown by colour. The plot follows below:



Typically, this graph would be followed by a statistical analysis to test for significant differences overall and between groups. Upon reflection, a table would be added to assist the user's comprehension of this visualization that would contain the populations per region such as the below table.

NHS England Region	Total Population
Midlands	10,769,965
South East	9,180,135
London	8,961,989
North East and Yorkshire	8,172,908
North West	7,341,196
East of England	6,236,072
South West	5,624,696

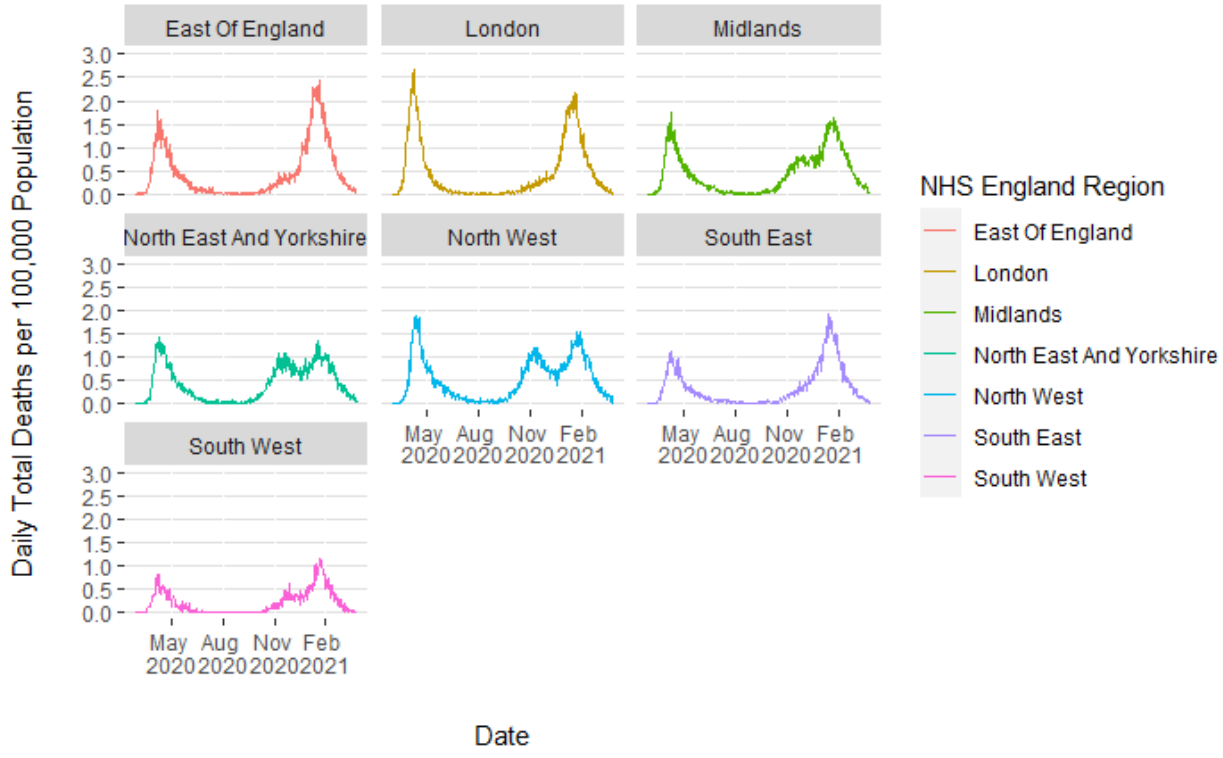
With the addition of this table for the visualization it becomes much quicker for the user to comprehend whether, or not, and albeit without scientific analysis to declare significance, the total population was one of the biggest drivers for increased death rates. To put this into context,

we can see very quickly the South East region has one of the highest populations but one of the lowest death rates per 100,000 population on the visualization - a vital comprehension not possible without the table and refinement of the question provided by the graph.

Question 7: How did the new, not cumulative, daily deaths differ per 100,000 population?

Cumulative totals provided interesting visualizations that provoke more rigorous investigation. In turn, I decided to look at the new daily deaths per 100,000 population and look for patterns and differences between groups. Because of the vast amounts of data points, at first it did not provide easy comprehension from the user when all plotted together. However, by applying facetting, the visualizations became much easier and efficient to analyse when the graphs were broken down individually.

Data cleaning, transformation and plotting was the same across the board as it was for Question 6, with the addition of facet.wrap() to create the nested mini-series of graphs - as well as selecting for daily and not cumulative deaths. The plot follows below:



This graph, contextually, details some of the most important information this data set provides. Patterns are readily apparent and the differences between the regions are very interesting. This graph, along with future deeper scientific analysis, could provide some invaluable and life-saving insight into preventative measures for future pandemics. Specifically pertaining to the measures implemented in different NHS regions and the approach to enforcing or policing these implementations.

This graph presents the question of why do these regions differ so dramatically? Wealth, lifestyle and social class could all play their role as an effect on deaths caused by COVID. The patterns for southern regions all loosely match and similarly the northern regions match each other, too. The midlands, acting as a buffer between the two, geographically, is the most unique pattern and somewhat reflects an intermediate between the common southern and northern patterns.

Future work

This data set, with the addition of these visualizations, provides valuable insight into the effects and management of COVID-19 within the NHS England regions. The visualizations yield answers to the key questions asked, but a more rigorous and scientific exploration, as an expansion of these topics, would be necessary to test for significance and provide empirical evidence to the findings. Statistical analysis would be exigent as the first stepping-off point following this report. Once significance was established, a more hands-on investigation into the NHS regions at a government handling and NHS trust level is required to answer the questions from this report beyond reproach, identify weak-points in the NHS as well as their relationship/handling with the government, and also prepare the country against future pandemics.