# Generative AI Tools for Quantitative Research

## Video 4 transcript

Video: https://youtu.be/-gI6dA-pR0o

Liam Wright: Hello, I’m Liam Wright, and today I’m going to be using LLMs within the Positron IDE to perform an analysis, so I’m going to take crime data at the community safety partnership level, and convert that into a map or display it as a map.

I’m not a geographer, so this is something relatively new to me and I wanted to do this as inspiration to show that with LLMs you can do some interesting analyses very quickly.

Now, I’ve already downloaded the data on crime rates from the ONS website. There’s a sheet within it called Table C5 that contains the per capita crime rates, so we have individual rows for the different community safety partnerships, and then in the I column, we have the per capita total of recorded crime rate.

That is data that I’ve downloaded and I’ve got it in this folder that I have open within Positron. I’ve also downloaded some data from the Ordnance Survey website that gives shape files which are boundaries for the different community safety partnership.

To open up, we’re going to be using Claude, which has a really useful integration within Positron. We’ll have information at the end of the blog post showing you how to set it up and a link to the Positron website, talking more about that. But here I’m just going to click this robot icon which brings up a sidebar, which has the general sort of chat interface that you might expect from an LLM.

Now, the Claude Positron assistant has three different modes: we have agent, which we’re not going to use, and we also have ask and edit. Ask is the usual sort of interaction way where you ask a question and you get an answer, and I’ll type the question into this box and then we’ll get the answer on the left-hand side.

You’ll see at the minute I’ve just opened a new session of R, so nothing’s loaded into memory, there’s been no outputs so far or anything like that. Let me just resize my screen slightly.

So I’ve already written some prompts just to make this a bit simpler. So first we’re just going to ask, “I have data on crime rates in England and Wales at the community safety partnership level in a file called crimerates.xls,” so that’s an Excel spreadsheet, “Within the sheet Table C5. I want to put these data as a map in R, the boundaries are in a shape file called CSPboundaries.shp. First give me code to import the spreadsheet into R.” I’ve given it the variable names, which start in row 8. So just to show, we have variables names in row 8. And then the CSP codes, because we have human readable names in column D, but the codes that we’ll use for merging are in column C. And the crime rates are in column I, which we discussed already. I’m also saying that some rows have missing data and need to be dropped, and the variable names should be renamed to something sensible, because we don’t want to read in the data sent to R and then have a variable name called ‘community safety partnership code’ because that’s not going to be easy to work with later.

So I’ll just send that now. So it gives me back a response relatively quickly. If we scroll, if we hover over a box of code, it’s got various different buttons. We can run the code using the play button. We can add it into R editor, so into this open file using this button, we can insert it at cursor with this copy and so on. I’m just going to add it in here using the third button, and then we’ll just load this data. So I’m not going to read it in detail but basically we’ll be able to see whether it worked or not in a second. So that’s just running.

Okay, so we’ve got an error, it said, “Sheet 9 has 32 columns, but call types has length 4.” So I’m just going to send that error back into a Positron assistant, to Claude, to get it to change the code. So I’ll just delete that code that we have already. Okay, let’s try again. So there, that’s worked now, so we’ve got CSP code, CSP name and the crime rate.

Okay, so that’s what we wanted. So the next prompt I’m going to give, I’m just going to ask, “Now give me the code to read in the shape file,” which is the file that contains the boundaries, “And draw a map so I can see the boundaries are as expected.” So we’re going to get a plot that just has an outline of England and Wales hopefully, if this has worked correctly. So let’s just scroll down the file here.

Okay, it gives me some information on the code, I’m not going to read that, I just want to insert the code here and get things running. So it’s going to load two packages, GG Plot 2, which creates maps or creates plots, rather, and SF, which is what you use for working with geographic data. So let’s run that code. Okay. So I’m just going to open this in a new window. So you can see that it’s given us the outline of England and Wales, so we know that it’s probably read in the data correctly.

Okay, now I’m going to say another prompt, which is, “Give me the code to merge the crime rates data with the shape file data and plot the crime rates on the map, check the merges have been performed correctly.” So I’m just going to enter that now. Okay, you can see this is running very fast. Notice also that it says, “The shape file has a column – CSP23CD that should match with your CSP code column.” So I didn’t tell it anything about the shape file, the name of the columns or anything like that, but it’s worked out that because behind the scenes, Positron is sending information from my current session to the API calls to Claude, so it’s given us information about…it’s sending information about the metadata to the various objects, so it’s been able to work out what the relevant columns should be without me explicitly declaring them.

Okay, so let me run that now. Here we go. So let’s see that in a new window again. So we now have this nice plot which has got the crime rates on it, you can see that they’re coloured by the level of crime. A light yellow is a crime rate of 400, a dark purple is a crime rate towards zero, I guess. Notice here that there’s very few points that actually have these lighter colours. Presumably this is because there’s a very skewed distribution, so we might want to use Claude to change this plot so that we can see, better distinguish between community safety partnership areas based on their crime rates.

So I’m just going to add another prompt, just asking, saying that, “The distribution of crime rates is skewed, so make appropriate adjustments to the visualisation.” Okay, so it’s just going to send me, give me a histogram to check the distribution and then it’s also going to, in this case, I think just plot quantile breaks instead. So again, we’re just going to put this code into the file, so we can use it later or store it for later if we want to. So let’s just make this first plot. So you see that we have this distribution of crime rates which is skewed, I’ll just go back to that. But there’s also, it appears, this huge outlier here. So it might make sense to remake the plot without that outlier included. And now we’ve got this different plot here that is accounting for the skewedness, but again, that’s not worked very well. I think what it’s done is just used the different scale, maybe a long scale or something like that.

So let’s give it one other prompt, what I think we’ll…yeah, this will be our penultimate prompt, “There appears to be outliers. Remove crime rate outliers and create new visualisation.” Just make that a bit smaller. So, “Looking at your data, I can see from the summary… “ So let’s wait for that to finish. Okay, this might take a few seconds, because it seems to be quite a long bit of code. Okay, so at the top of the response, it says, “Looking at your data, I can see the summary that there are potential outliers with a maximum of 446.41 compared to a mean of 81.” So it’s going to remove outliers using the interquartile method to create a cleaner visualisation. So I notice here that we ha this plot that gave us this outlier, I think maybe there was another bit of…yeah, there was also this output as well, that said that the maximum is 446.41. So again, this information, what’s in the console, what’s in our metadata from the objects is being sent to Claude and it’s really allowed us to integrate the LLMs into our workflow, so we can work really fast.

So now let me just try and run this code instead. I’m just going to, again, add it to the end of the file and let’s… So firstly it’s just telling me some stuff about how many outliers are removed, so I think this is the upper bound without the outliers, it seems like there’s five outliers that have been removed; that includes Westminster, which is this one with this huge high crime rate, but then there’s also Camden, Manchester, Kensington and Chelsea, and Middlesbrough. So it’s just going to create a data frame that removes the outliers, I think, and then it’s going to plot them again. So let’s do this.

Okay. There we go. So let’s look at this now. So now I can see that the outliers have been removed. The top pf the crime rate on this graph is about 150. There’s much more distinction between the different areas, and it also says at the bottom, “Grey areas indicate missing data or outliers.” So now with a better visualisation, it’s able to give us a sense of differences between regions or community safety partnership areas, rather, by removing the outliers that were making the plot not very good.

Okay, so there’s one other prompt I could do, which is to, “Suggest some analyses that I can perform with this data and give me the code to do so.” I’m just going to put that into the chatbot now, so you can see it’s giving us the code for spatial autocorrelation analysis. I imagine it’s also going to give us different types of analysis as well, some simple descriptives, I’m not going to run those right now because I think that will take too long, but you can see how quickly we’re able to make a very usable plot using data that you might not have been…had much experience with. I certainly don’t have much experience with geographic data, so it’s really…LLMs are really incredible in enhancing our abilities, very fast, but obviously we need to make sure that we are critical and know that what we’re doing as we use them. Thank you.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK