* NLP applied to Zoopla listings and then linked to Land Reg PPD (would need to start work *before* the summer as application to UBDC required: <https://www.ubdc.ac.uk/data-services/data-catalogue/housing-data/zoopla-property-data/>) to examine opportunities for enhanced hedonic modelling using non-spatial/non-categorical features. See: <https://www.youtube.com/watch?v=SNF_TnKE-GI&t=1s> for context. The idea here would be to try to fill in the huge gaps caused by the paucity of detail in the Land Registry data set when it’s used for house price modelling (details like garages or swimming pools) so that we have a better understanding of underlying prices and also how regional preferences in property impact price dynamics. Care would need to be given to duplicate listings and re-listings at new prices, something that [Pewtils](https://pewresearch.github.io/pewanalytics/text.html) and general NLP might help with.
* Impact on COVID-19 on STL market in (ideally) London, Manchester, and Edinburgh. See <https://www.wired.co.uk/article/airbnb-london-short-term-rentals> and <https://www.wired.co.uk/article/airbnb-coronavirus-london> for context. Objective is to link listings across time periods (and to produce a means of doing so that is robust and replicable—ideally even to changes in management such as a private landlord turning the listing over to a professional management company for day-to-day management) and understand recent trends in up to 3 cities. There are a number of angles to this, but you’d undoubtedly be looking at the minimum at both pricing and availability: how have prices changed, how many listings have left the market, what are their characteristics (e.g. more multi-hosts?), etc.
* Scaling of ML neighbourhood change modelling approach (building on [my work](https://journals.sagepub.com/doi/abs/10.1177/0042098018789054) & [Bin & Adam’s work](https://www.ucl.ac.uk/bartlett/casa/publications/2020/jan/casa-working-paper-215) as well as [this recent publication](https://www.tandfonline.com/doi/full/10.1080/10106049.2019.1595177) and [this one](https://onlinelibrary.wiley.com/doi/10.1111/gean.12273)) to country-scale (MSOA-level?) and analysis of resulting clusters: are there areas of Cornwall or the Highlands that appear to have experienced similar changes to areas of Manchester or London? So there are a series of technical improvements that need to be reconciled, as well as two types of scaling issue (performance challenges ahead *and* major questions about how best to compare areas with radically different compositions: are the problems caused by 2nd [or 18th] homes in South Kensington remotely similar to those caused by 2nd homes in Devon and Cornwall?)
* A fully open gridded population assignment process: build on exploratory work undertaken in R to grid the entire UK and attempt to use Iterative Proportional Fitting to then probabilistically assign households to housing based on local, regional, and national trends so as to improve population modelling, particularly in very low-density and very high-density areas. Adam and I had discussed this ages ago but didn’t have time or the opportunity to write it up as research. What we are aiming to achieve is the ability to probabilitistically assign households at a very fine scale. This would enable us to create custom-geographies that persist over time because the population can be constructed at each Census interval.
* Supporting, but separate from, preceding task: use of ML to identify building typologies from a mix of satellite/MasterMap/parcel data and then linkage to LR and related data. National scale (but less precise) version of: <https://colouringlondon.org/>. This would help us to fill in gaps where there are no transactions in the Land Registry (because no sales) and where all we have is a count of terraces, semi-detached, etc. At the OA or above level. So we’re looking to know that the South side of Street X is composed of 15 terraced houses, etc.
* Extending work around identifying Ghost Hotels on Airbnb using a mix of spatial and textual similarity analysis from InsideAirbnb data. If you tackled this in your final assessment for Intro to Programming (CASA0013) then you will need to be *very careful* not to plagiarise your own work on this topic, though you can obviously build on any code that you’ve already written. This would be ideal for someone who saw from the work that they did for the I2P module that it could be taken further and would like an opportunity to do so: you would expected to conduct a much more rigorous literature review, as well as to go much further with NLP as part of the analytical process, and it would likely be a good idea to try this out on several cities where we know that such problems exist (e.g. London, Edinburgh, San Francisco, and New York City). I recently came across the [Pewtils Text Analytics library](https://pewresearch.github.io/pewanalytics/text.html) and that looks like it might be relevant.