

The coolest thing about data analysis...

...is that it makes you think very hard about the world. I'm having a short break between jobs. After 2 weeks of cycling, kayaking, camping, tidying, diy, meditation, niggling chores and piano playing, I'm finally ready to start thinking about data again.

Kaggle is a lovely thing. Not only does it encourage learning and cooperation between data scientists and create better algorithms for worthy causes, it also provides nice clean data to play with (believe me, raw data is way messier than this). I'm currently geeking out on the San Francisco crime dataset - basically a huge CSV containing the date, category (and details), location (with police district) and outcomes of SF crimes from 2003 to 2015.

I could just throw a bunch of learning algorithms at this dataset (and, in truth, am, because that's what Kaggle's about) but I'm finding it interesting to think first about what I know about crime patterns from my time in the UK, whether those transfer to the US, and what else might be interesting to look for if I know more about the area. For instance, are car crimes grouped near easy escape routes (thank you SFdata.gov for the major roads map); are there more thefts on area 'boundaries' (a UK study a while back showed more crimes near areas that perpetrators felt comfortable in); are there more juvenile and petty crimes [near schools](#) (also in SFdata.gov)? I know that this is consciously testing out my own biases, but I'm on holiday and it's an interesting thought experiment.

There's a similar style of dataset ([Taarifa's water point challenge](#)) on DrivenData. Some similar thinking might be interesting there too.