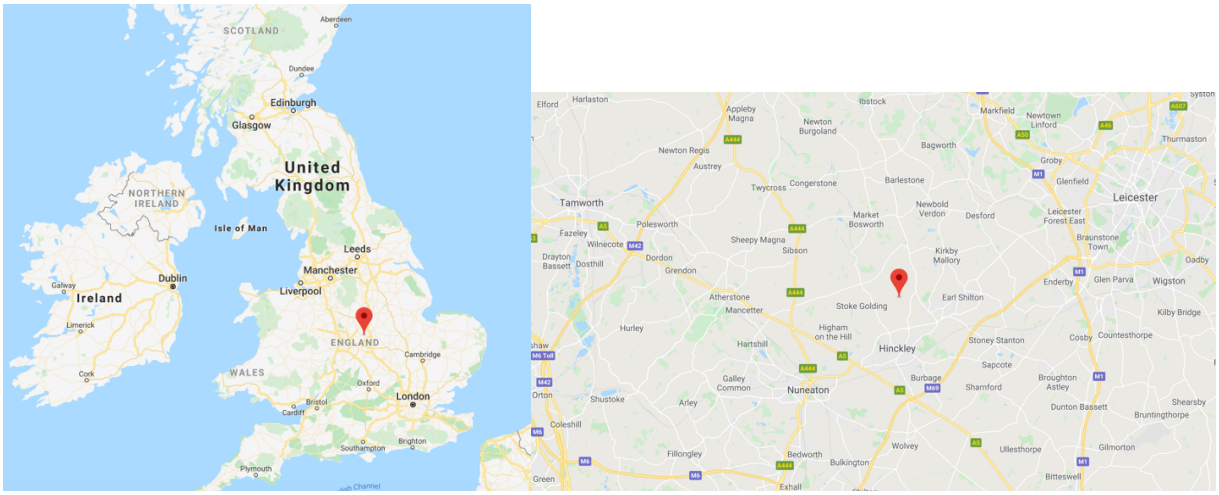
Bradley Sundsbarm
CSCI551 assignment 5

The Data:
The UK government amassed traffic data from 2000 and 2016, recording over 1.6 million accidents in the process and making this one of the most comprehensive traffic data sets out there. The part I used was from 2012 to 2014, which had 464,698 records of traffic accidents in the UK.
https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales/data

Among all the data in these reports, I chose to run spark to look at these specific fields: Latitude, longitude, day of the week, road type, time, number of casualties, junction control and speed limit.

Generalizations from my data:
I was curious as to where most of these accidents were occurring, so I ran standard deviation and means on the lat/long coordinates of each incident:
mean of longitude = -1.375156
standard deviation of longitude = 1.382137
mean of latitude = 52.575498
standard deviation of latitude = 1.436370



Before running my pyspark code I figured the coordinate would be much close to London. But after looking at the map with my results zoomed out its easy to see why not. London is a big city yes of course but Manchester, Liverpool, Leeds, Birmingham (closest city to marker) are also highly populated, so they have an impact on the data significantly. Since there are thousands of accidents around these locations, they had a significant impact on the final result, and since the standard deviation is somewhat low, considering these are coordinates. It is reasonable to conclude that most accidents occur in those main cities around the coordinate.

Other Results I wanted to see out of curiosity:
Road_type: highest one by far was Single carriageway 351,268. Dual and roundabout had ~65k and ~32k respectively.
Definition of Single Carriageway - A single carriageway or undivided highway is a road with one, two or more lanes arranged within a single carriageway with no central reservation to separate opposing flows of traffic.

I also wanted to get an idea of what day and time would it be most dangerous to drive. According to my time/day data around 330-5pm on Friday, Saturday, or Sunday was the most dangerous time to drive. Found this interesting because 330-5pm is typically associated with people getting off work, but only the workday in the top 3 was Friday.

Speed limit was also interesting but makes sense first place was 30km/h and second was 60, with 304,842 and 64465 respectively. The results show speed limits 10 – 70. Not terribly sure as to why 30km/h is so much higher than the rest. But it's reasonable to hypothesize that because of my coordinate above, 30km is likely the top speed limit in the cities.

Results I did not use:
Number of Casualties - because it did not make sense the CSV description states that it should only be values 1 – 3 depending on severity. But when I calculated the count for it, I got values anywhere from 1 to 24+ which does not make sense I have yet to year of an accident causing more than 5-10 deaths at the absolute most. According to this between 2012-2014 there was 10 deaths for 43 incidents haha. I would have to do more research, but as of now I will ignore this data.
Junction Control - seemed pretty obvious.

Contrast of Spark to MPI

I don't think you can really compare them in terms of easier or harder, both are trying to do completely different things. It's like comparing apples to oranges. Obviously, spark was easier to write for me since the only MPI coding I've done was in c++, which with getting your head around send/receive is no easy task. Not to mention the amount of code you need for spark, or should I say lack-there-of haha.
Another reason you can't compare them is that their goals are not the same. Yes, they utilize parallel computing, but they are solving different things. Spark is looking over data and filtering/gathering info. MPI is running user written algorithms on data.
Spark is just clearly easier when it comes down to processing massive data and extremely useful for SQL applications.
MPI is more geared towards you want to run this crazy big algorithm on this giant amount of data and you want results faster than one computer can give you.