Machine Learning Engineer Case Study

Data Sources

For this case study you will use the road safety data available from here:

http://data.gov.uk/dataset/road-accidents-safety-data

Please use the ALLSTATS19 data (accident, causalities and vehicles tables) for 2005 to 2014.

Description

These files provide detailed road safety data about the circumstances of personal injury road accidents in GB, the types (including Make and Model) of vehicles involved and the consequential casualties. The statistics relate only to personal injury accidents on public roads that are reported to the police, and subsequently recorded, using the STATS19 accident reporting form.

Task

The objective of this case study is to use Scala/Spark to:

- To implement a data ingestion pipeline that integrates the tables in a coherent way
- To profile the main characteristics of the data, and obtain interesting facts that are worth highlighting.
- To implement a clustering algorithm that provides insight into accident hotspots

Your data analysis should consist of the following components:

1) A short description and justification of the steps taken.
2) Performance and evaluation results of the clustering model processing.
3) Insights gained from the analysis.

Please return your completed outputs and code to HR within a week (7 days).