BI Engineer - Challenge

Introduction

At Stuart, being a last-mile delivery platform that connects business to fleets of geo-localised couriers, we depend on the quality of our recommendation engine that defines the most optimised route for drivers.
As BI engineers, we've been approached by our Product Analytics team, asking us to provide them with an insight into the most common road accidents in the UK. They'd like to get an overview in a visual manner to help them understand the possible risks for our couriers. They're also looking to reproduce/reuse the pipeline themselves.

Challenge

You have been tasked with providing insights for the Product Analytics team into the road traffic incidents that happened in the UK over the past few years.

- (Mandatory) They'd like us to provide them with raw data and explain the insights, i.e. what are they seeing?
- (Mandatory) They need to be able to run the pipeline locally. In the future they'd like to re-use the pipeline to extract additional datasets, therefore making the pipeline modular would be most beneficial to them.
- (Optional) Having implemented a modular approach, they'd be keen to have an ability to orchestrate it.

1. **Have a look at the following two datasets:**

- *AccidentInformation.csv*: every line in the file represents a unique traffic accident (identified by the AccidentIndex column), featuring various properties related to the accident as columns. Date range: 2005-2017
- *Vehicle_Information.csv*: every line in the file represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns. Date range: 2004-2016

2. **Using any language/framework/tool, do the following:**

- Leveraging the Kaggle API, extract the two datasets, focus on making the extraction process reusable if possible.
- Load the most recent accidents we have vehicle data for into a Postgres database along with the dates these occurred, day of the week, severity index, driver home area, age bands, vehicle age and journey purpose.
- Visualise the percentage comparing the different age bands for the incidents and the age of their vehicles. Also, feel free to be creative and draw any other insights/visualizations you can think of.
- Make the process reproducible and modular, ensure data correctness and anticipate any errors.

3. **Explain in the README file:**

- How to run it and how to query the data, explain what we are seeing.
- Why you've chosen each tool/language/framework for the task.
- Any data quality practices you would enforce as well as error handling.
- How you would test for correctness, i.e. reconcile.

Guidance

Take your time and when you're done send us a link to your public Github repository. Any questions, give us a shout!