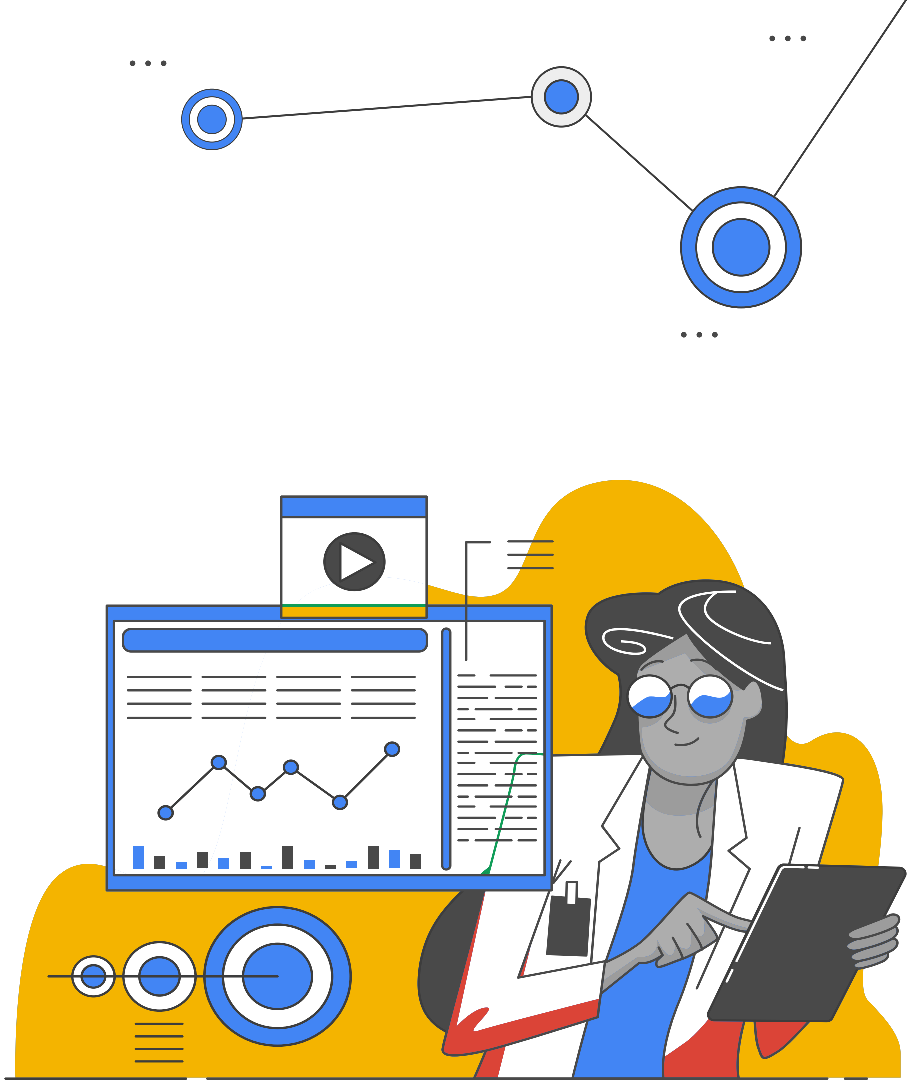
Hackabank

Standard Bank and Mobalyz Hackathon Solution



Team

Danel Adendorff
Lise Prinsloo
Minette Farrell
Rachel Rawraway

01

Problem Statement

Creating comprehensive behavioural profiles of taxi drivers with the aim of understanding the risk associated with driver characteristics.

02

Context

Taxis play a prominent role in the South African economy. Understanding what characterises more/less risky taxi behaviour gives very valuable insight into the market for potential investors.

03

Data

106 datasets, split according to each vehicle, with time stamped daily telematics data with odometer readings, ignitionStaete, speed, coordinates etc.

04

Model

First employed an autoencoder to reduce complexity, then clustered vehicles together based on similar behaviour, using k-means clustering.

Data Pipeline

Data pre-processing

Understand data

Exploratory Data Analysis:

- Schema
- Data types
- Missing values

»

Feature selection

Understand driver's behaviour

A few features that are considered as indicative are

- acceleration
- speed
- time of day

»

Feature Engineering

Model driver behaviour

A new table is created that summarises each vehicle's behaviour in terms of

- ratios
- penalties

Feature Engineering

Ratios

These features represent the proportion of instances that a specific condition was met

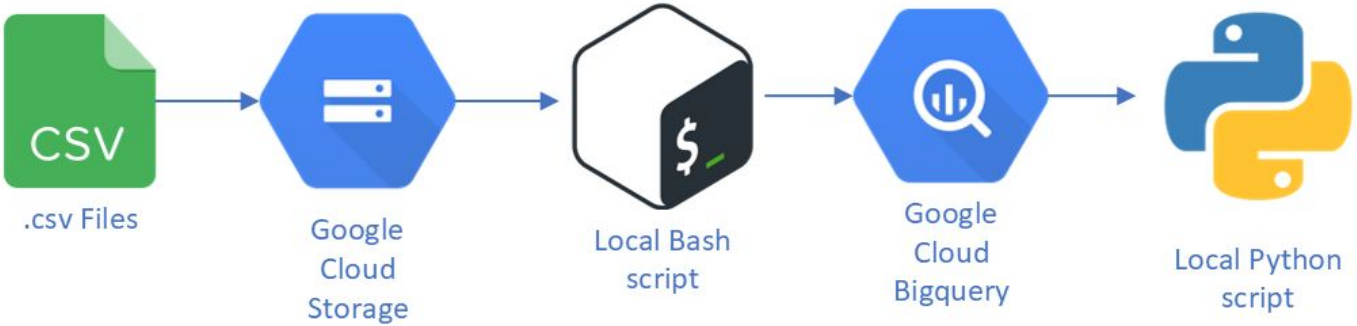
- Speeding
- Dangerous area
- Dangerous times
- Corner speeding
- Idle ratio

Penalties

These features add penalties to a driver each time they perform a specific risky behavior

- Speeding
- Corner speeding
- Braking
- Acceleration
- Net Acceleration

Initial Local Pipeline



Insecure, inconvenient and unscalable

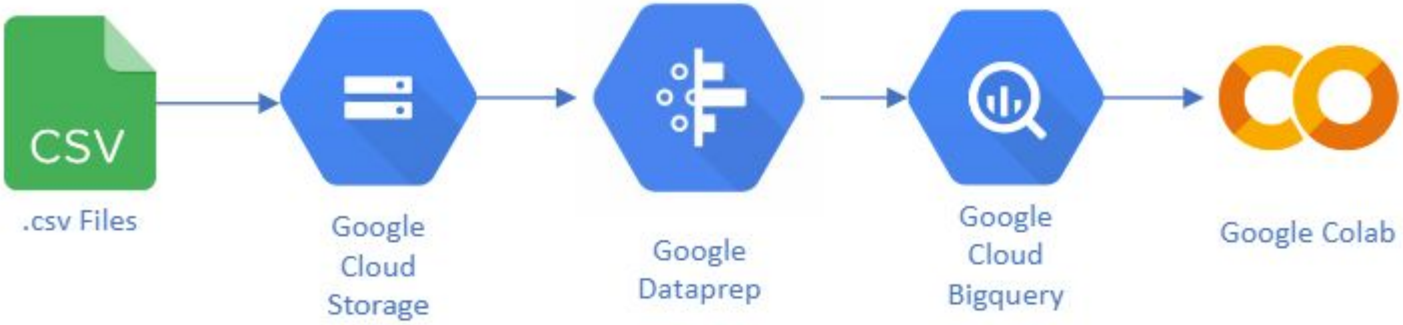
$ Bash might not be supported on all operating systems.
$ Python needs a virtual environment to install libraries.
$ Requires a service key to connect to BigQuery.
$ Slow data transfer and updates.
$ Manual data cleaning and pre-processing.

Cloud Based Solution

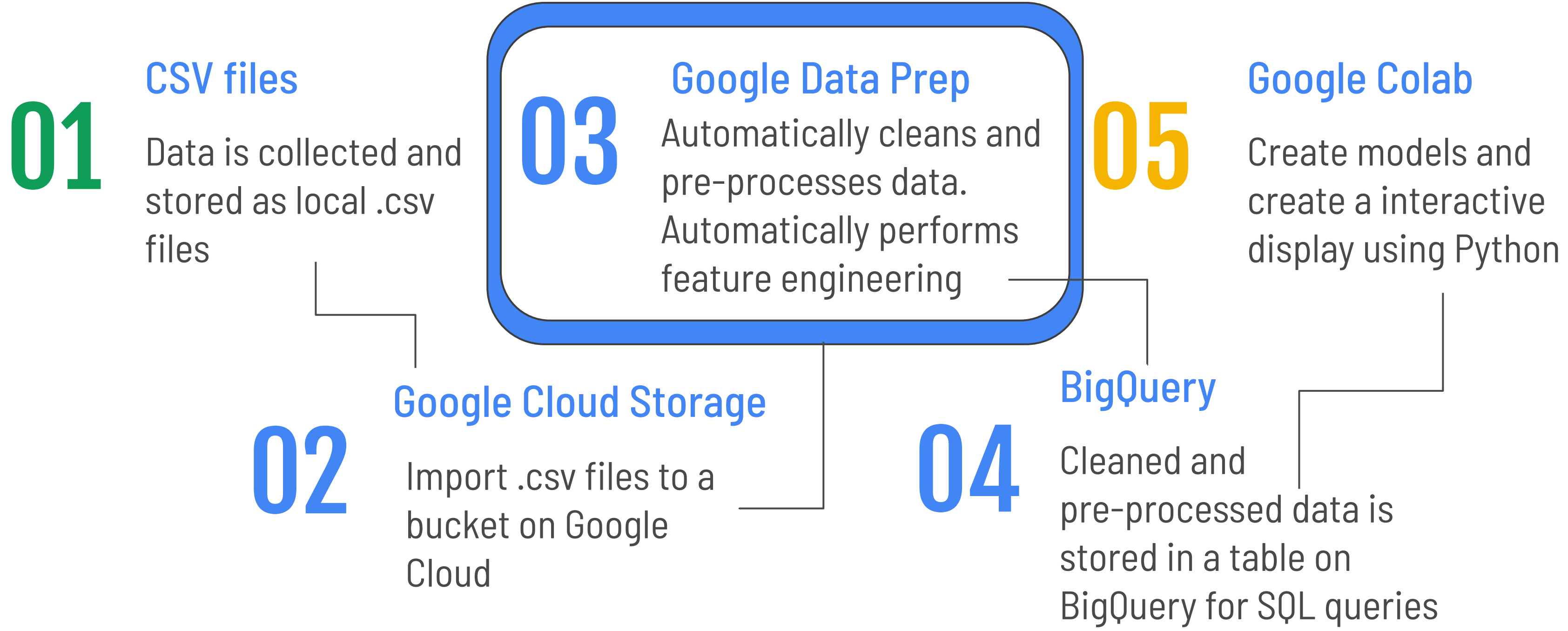
Google Cloud offers premium low-cost cloud-based, scalable software for creating a data pipeline.

Scalable (vertical and horizontal)
Secure (IAM)
Serverless (eliminates management)
Services and APIs (seamless integration)
Speed (fast processing)

Cloud Based Pipeline



Cloud Based Pipeline



01

CSV files

Data is collected and stored as local .csv files

02

Google Cloud Storage

Import .csv files to a bucket on Google Cloud

03

Google Data Prep

Automatically cleans and pre-processes data. Automatically performs feature engineering

04

BigQuery

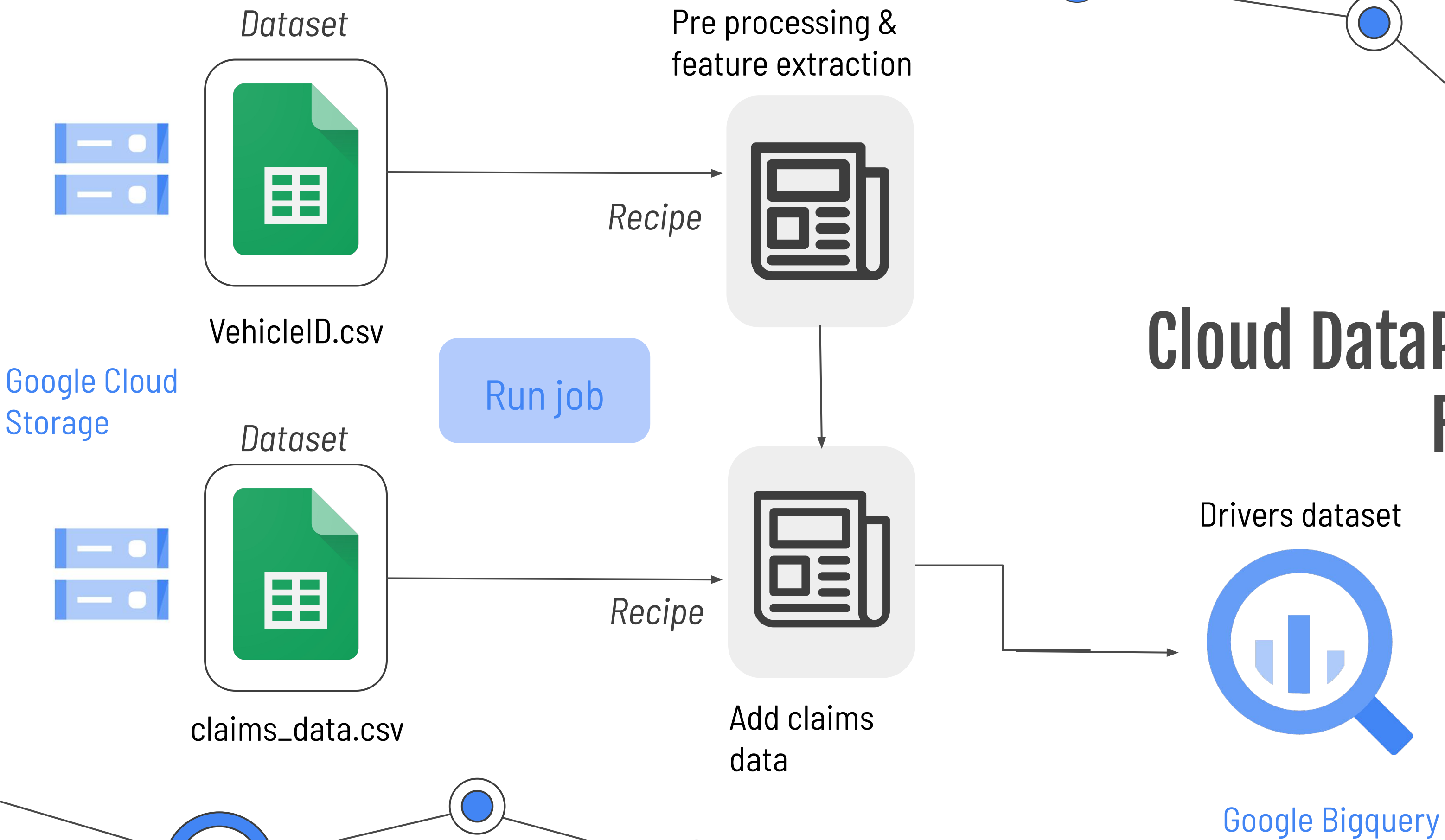
Cleaned and pre-processed data is stored in a table on BigQuery for SQL queries

05

Google Colab

Create models and create a interactive display using Python

Cloud DataPrep Flow



Model Approach

Data prep

Standardise and normalise the data

Dimensionality reduction

Autoencoder neural network

Clustering

K-Means clustering and class assignment

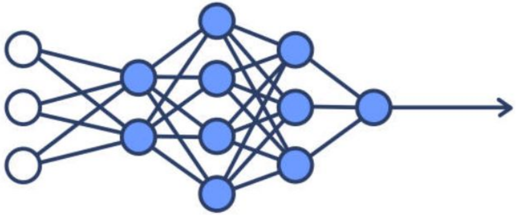
Dimensionality reduction

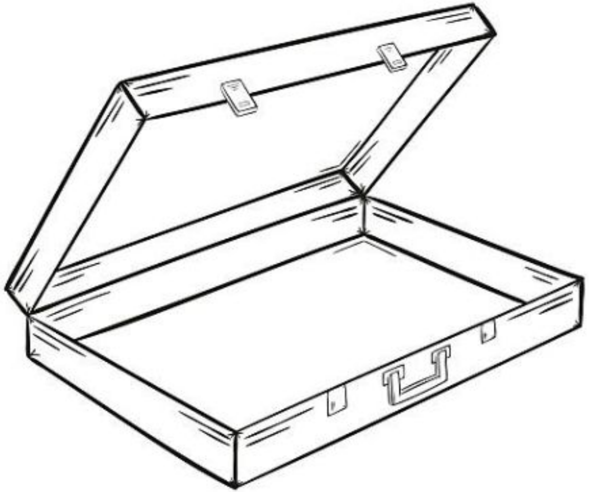
Auto-encoder

Neural network based AI technique

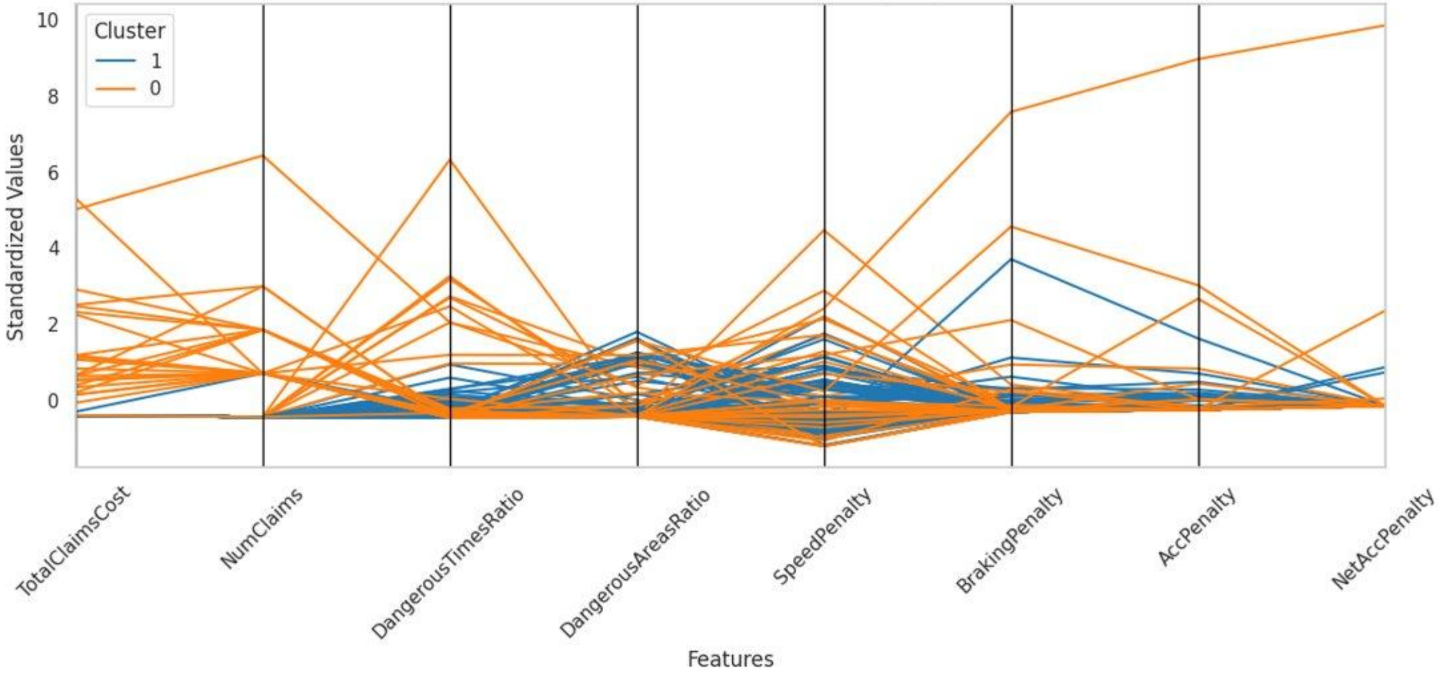
Learn & represent most essential features

Data compression, noise reduction & uncovers hidden patterns



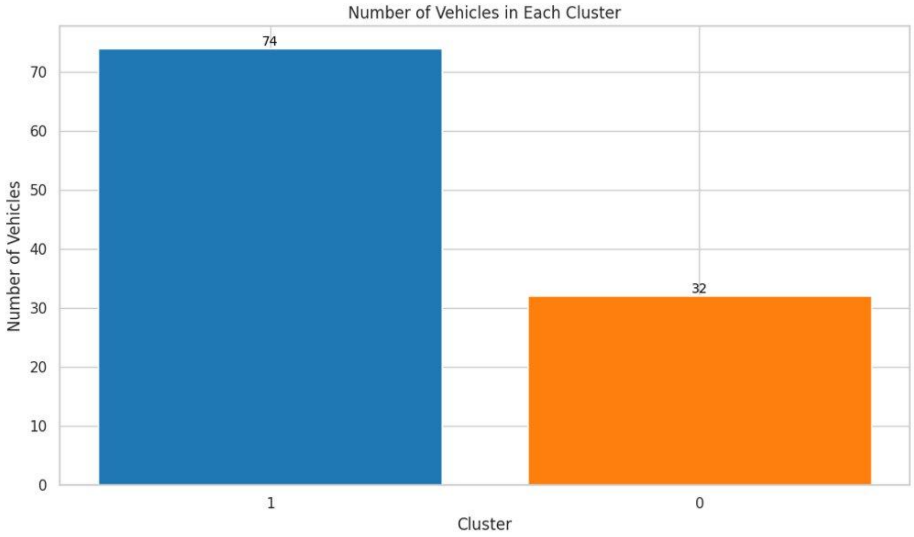


Parallel Coordinate Plot of Clusters from K-means



Our model clusters the vehicles into two distinct clusters, cluster 0 (bad) and cluster 1 (good).

Analysis

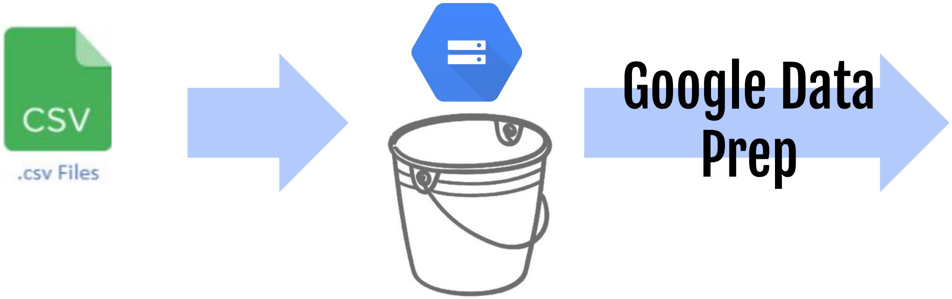


From the 106 vehicles in the dataset, we classify 74 vehicles with good behaviour and 32 vehicles with bad behaviour.

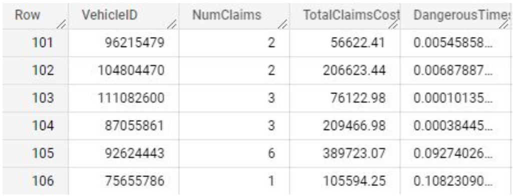
Cluster 0 follows a trend of more claims, more dangerous activity and more penalties. It is for that reason we classify cluster 0 as bad vehicles

Cluster 1 on the other hand tends to have less claims, less dangerous activity and less penalties. Hence, we classify cluster 1 as good vehicles

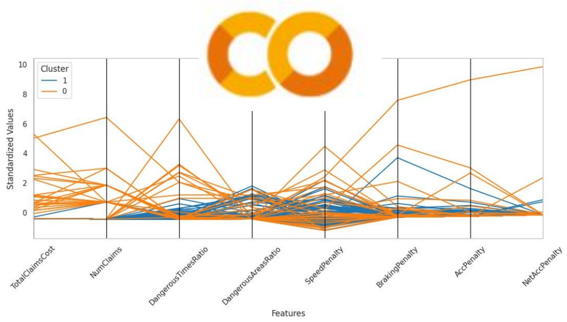
FULL SOLUTION

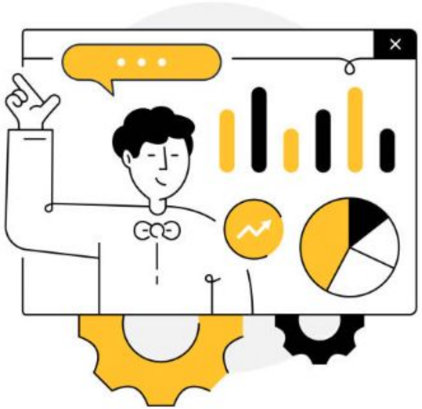


Google Data Prep



Row	VehicleID	NumClaims	TotalClaimsCost	DangerousTime
101	96215479	2	56622.41	0.00545858...
102	104804470	2	206623.44	0.00687887...
103	111082600	3	76122.98	0.00010135...
104	87055861	3	209466.98	0.00038445...
105	92624443	6	389723.07	0.09274026...
106	75655786	1	105594.25	0.10823090...

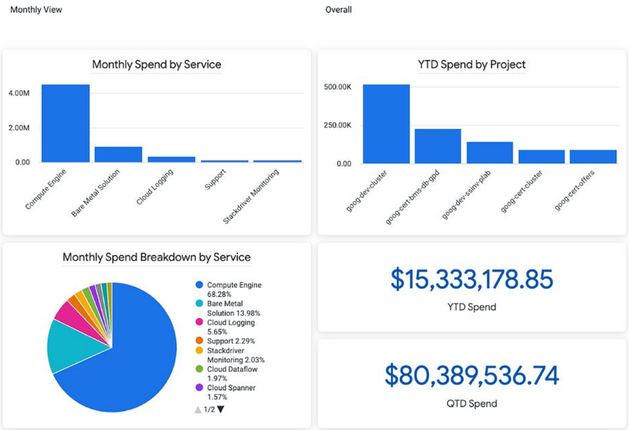




Next steps

Build individual driver profiles

Dashboard



$15,333,178.85

YTD Spend

$80,389,536.74

QTD Spend

Thanks

Team Hackabank