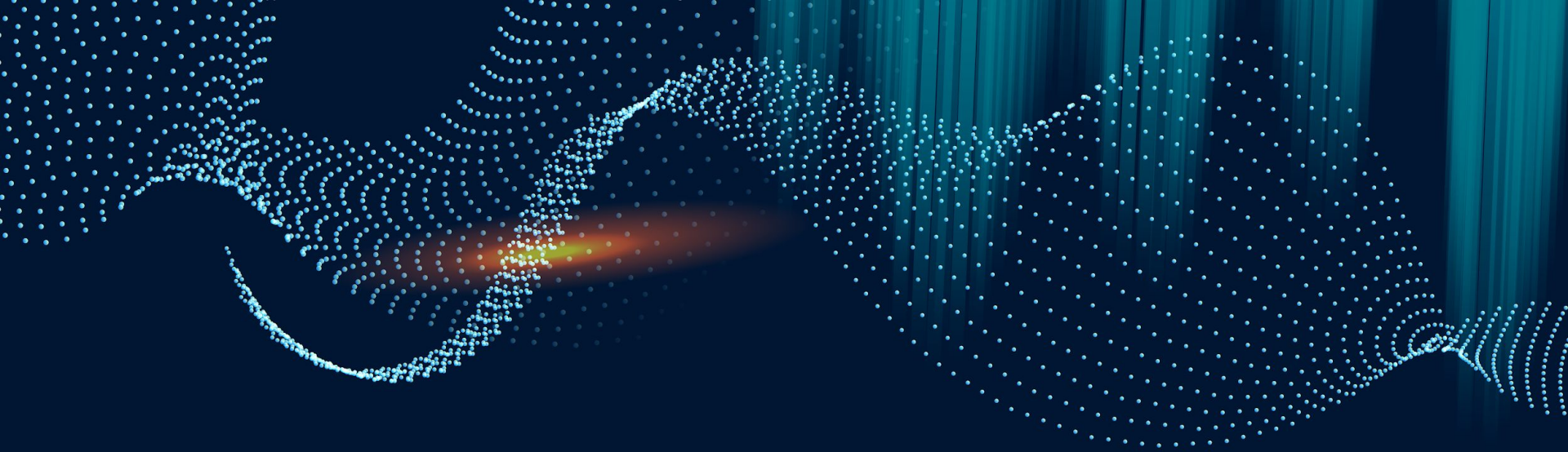
Risk Factor Analysis

Big Data Project

Gibran Cornejo

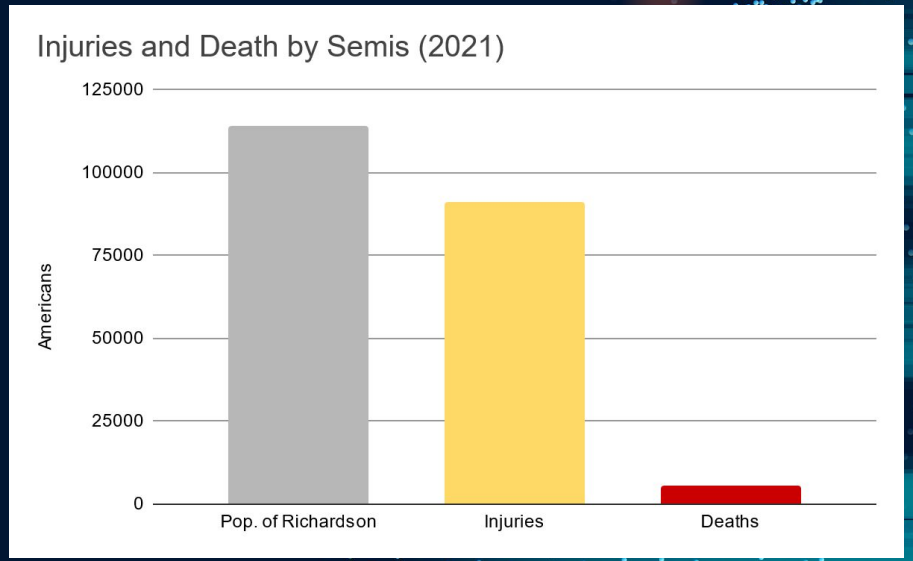


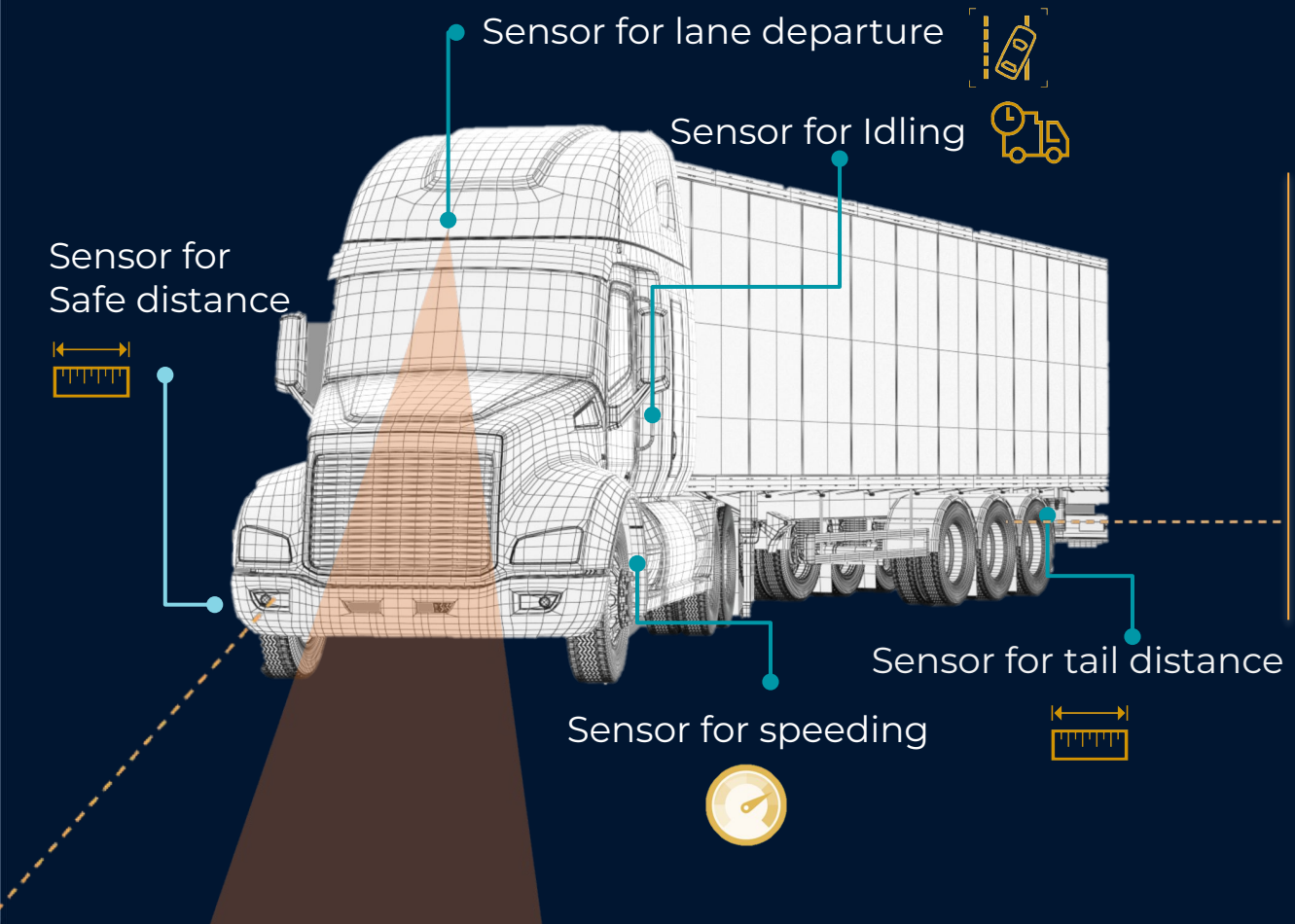
01 Background & Problem Statement

Background - Problem statement

Semi-trailers are much larger than normal trucks and cars. If driven recklessly, semi-trailers can cause car crashes and pile-ups, leading to serious injury and death.

Therefore it is essential for this industry to prioritize safety measures and ensure that their drivers are properly trained to operate these large vehicles. This includes implementing safety protocols and ensuring that the drivers are properly rested and not driving while fatigued.





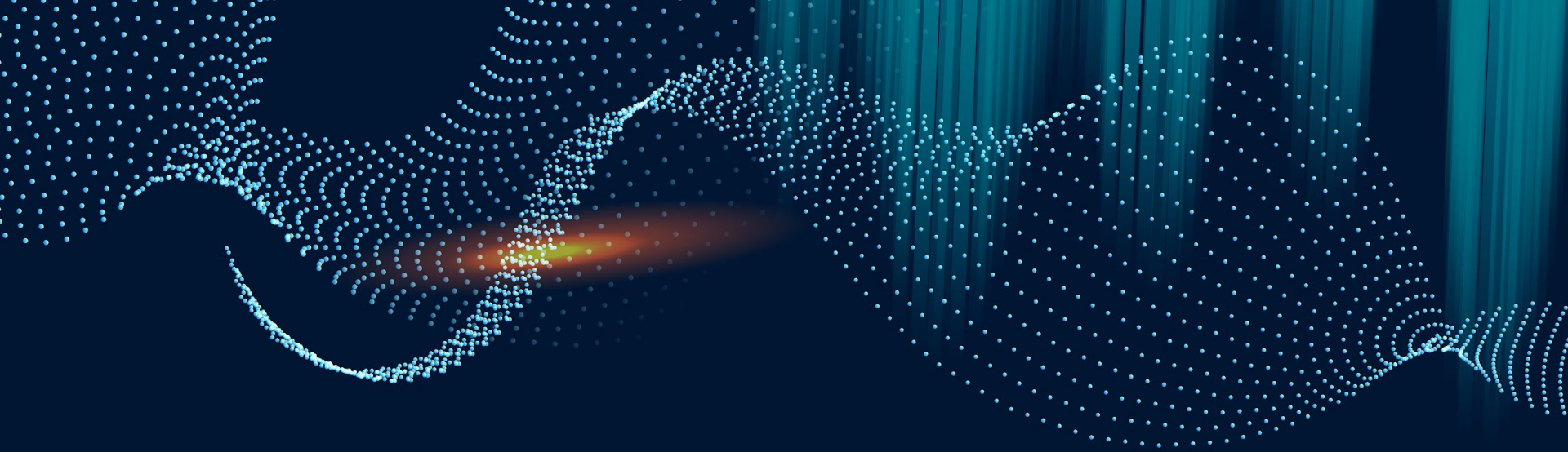
Sensor technology

Using sensor technology we can collect data from the trucks, such as the speed, tail and front distance among others, which leads to safer trips.

Our Objective

Given data of semi-trucks and their drivers, our goal is to inspect the data and answer some business question a trucking company would likely have, including:

- *Calculate and forecast a future "risk index" of the driver based on the model, event, velocity and event type.*
- *Analyze the most risky behaviours committed by these drivers.*
- *Identify risky locations for drivers.*
- *Identify risky drivers.*
- *Analyze the models driven by these drivers.*



02 Process flow

Architecture Diagram

Process Flow (Architecture Diagram)

DATA LAKE

1.Loaded data from CSV to HIVE using the HIVE CLI
2.Create “riskfactor” table using pig

HIVE DRIVER

-avg_miles
-drivermileage
-geolocation
-truck_mileage
-trucks_mg
-riskfactor
-trucks

HIVE CLI & UI

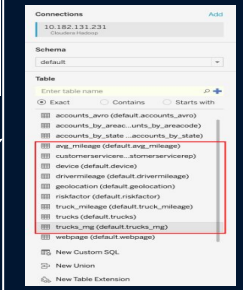
1.Load data from HDFS from HIVE tables created using “LOAD DATA INPATH”

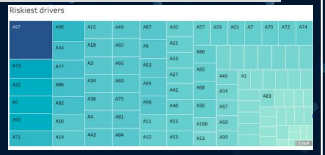
HDFS

ODBC Driver

Spark Connector (thrift-server)

DATA VISUALIZATION LAYER

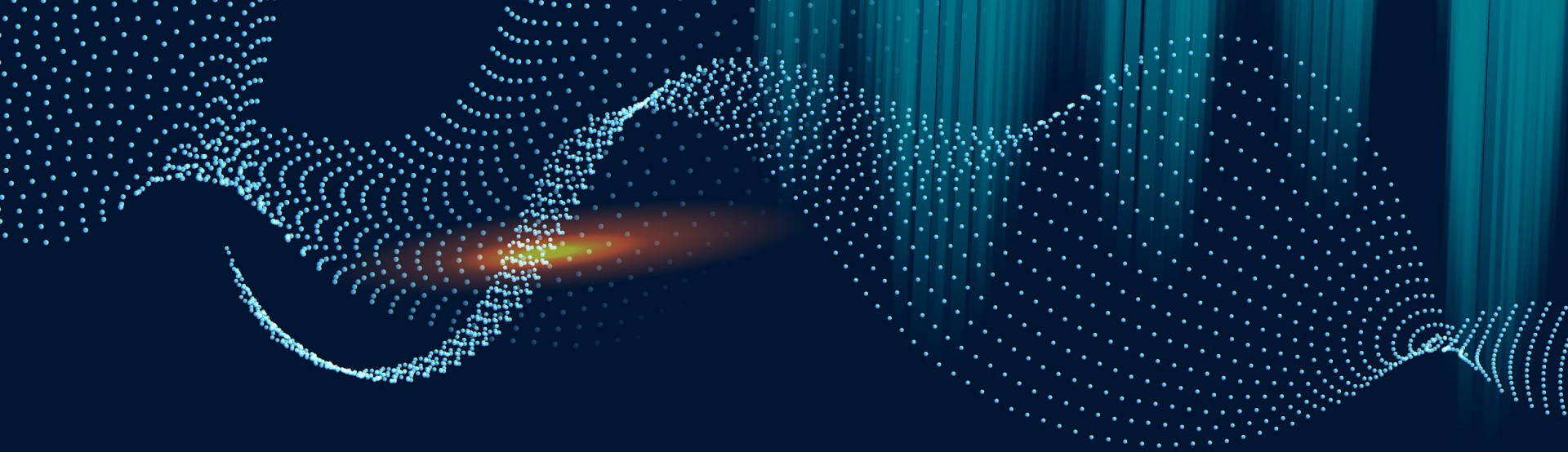




FORECASTING LAYER

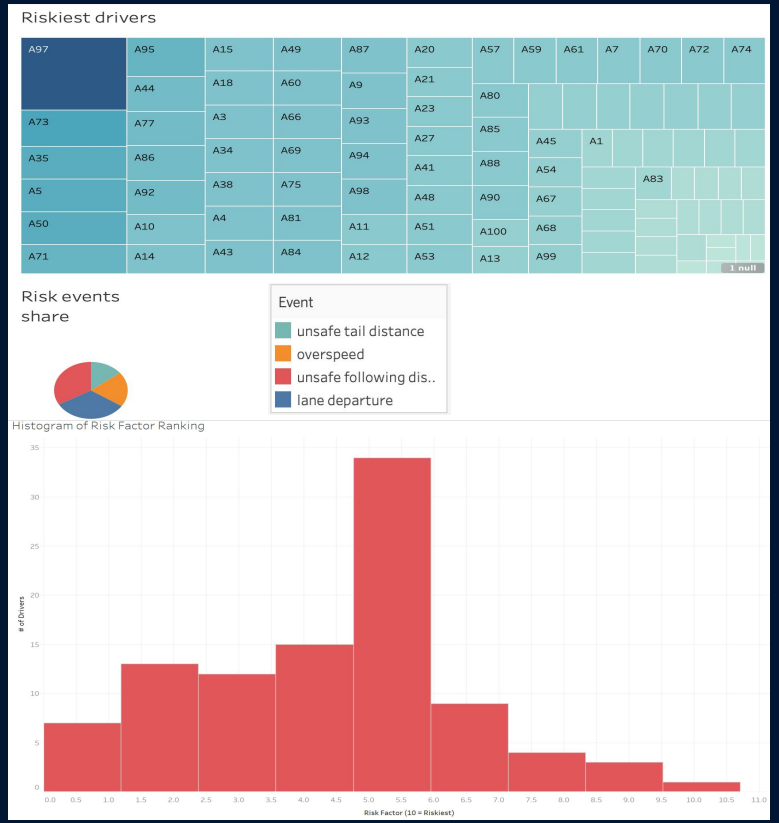
```
from pyspark import SparkContext, SparkConf
from pyspark.conf import SparkConf
from pyspark.sql import SparkSession, HiveContext
sparkSession = (SparkSession
                .builder
                .appName('example-pyspark-read-and-write-from-hive')
                .config("hive.metastore.uris", "thrift://http://10.182.131.21/:9083", conf=SparkConf())
                .enableHiveSupport()
                .getOrCreate()
                )
```

Decision Tree Acc - 95%



03 Graphics

Risk Factors

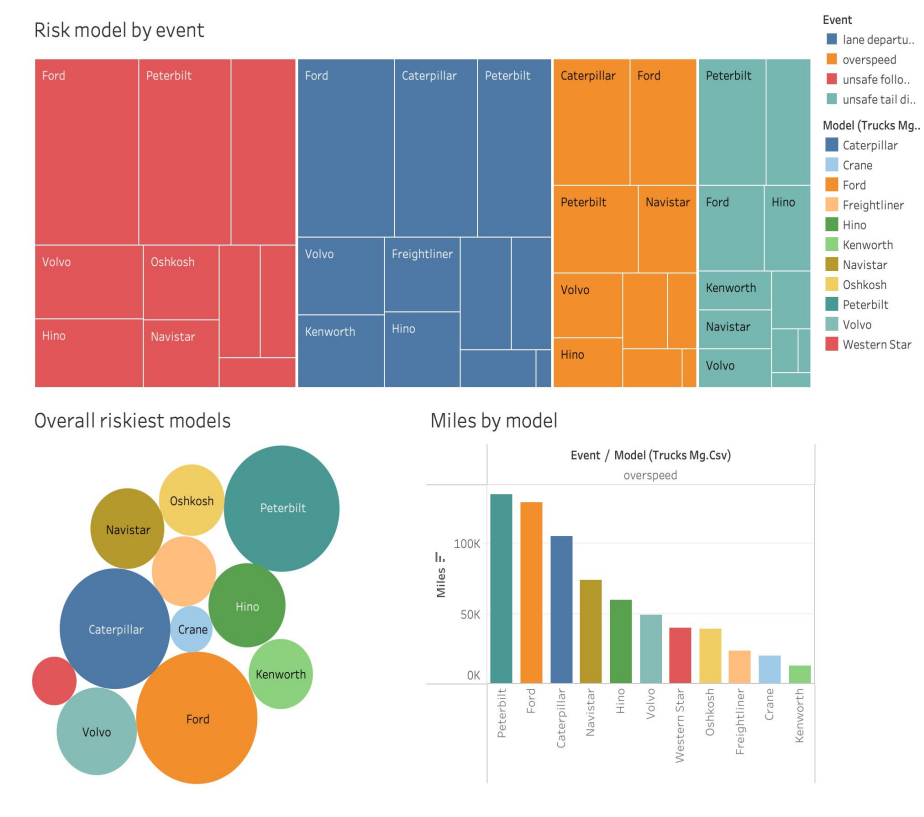


The graph depicted below illustrates the cohort's most high-risk drivers. It can be deduced that A97, A73, and A35 have made the greatest contributions to the overall risk factor.

Unsafe following distance is the most common risk factor, followed by lane departure

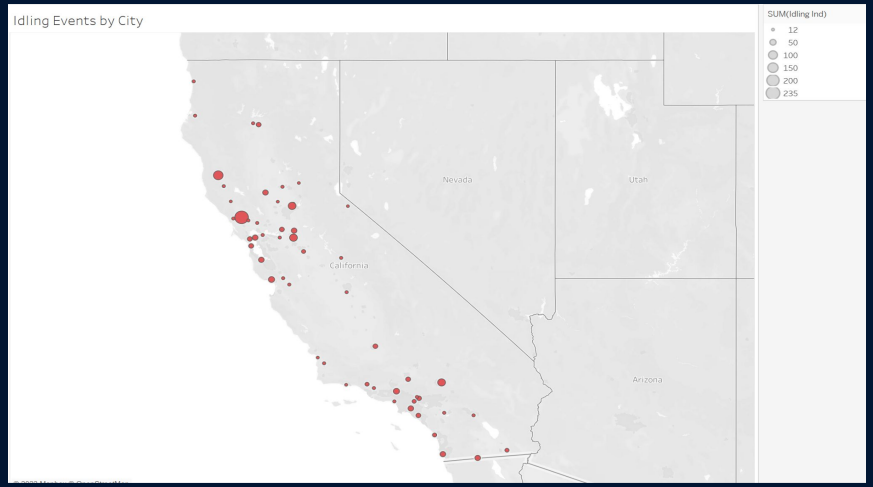
The histogram shows that most drivers responsible for risk events fall within the 5.0-6.0 range of the risk factor scale.

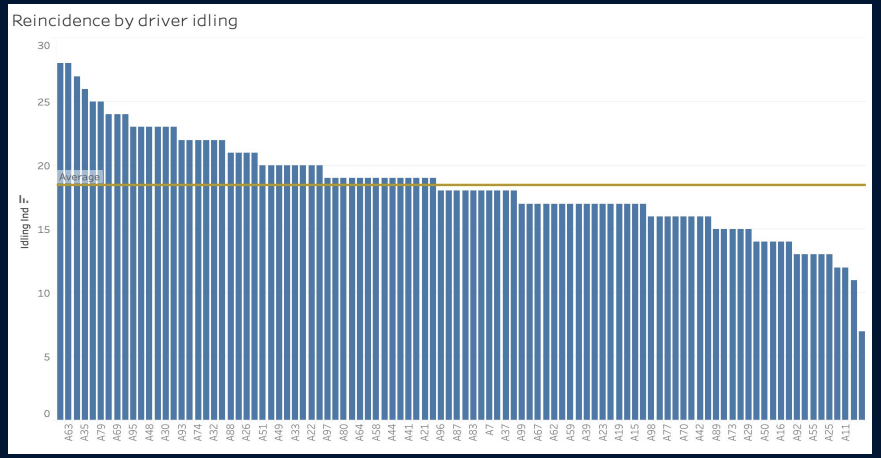
Risk by Model



The treemap shows the car companies with the highest incidence of risk events, with Ford leading in unsafe following distance and lane departure, Caterpillar in overspeeding, and Peterbilt in unsafe tail distance

Ford vehicles have the highest incidence of risk events overall. Peterbilt and Ford trucks have covered the greatest distance while overspeeding compared to other truck brands.



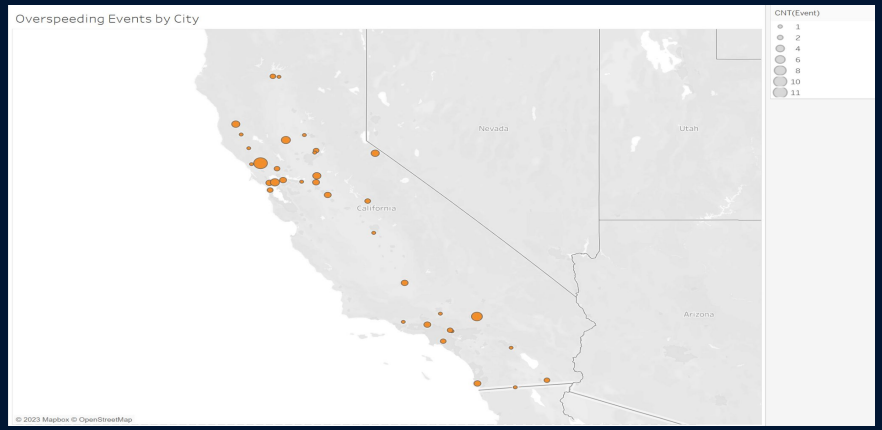


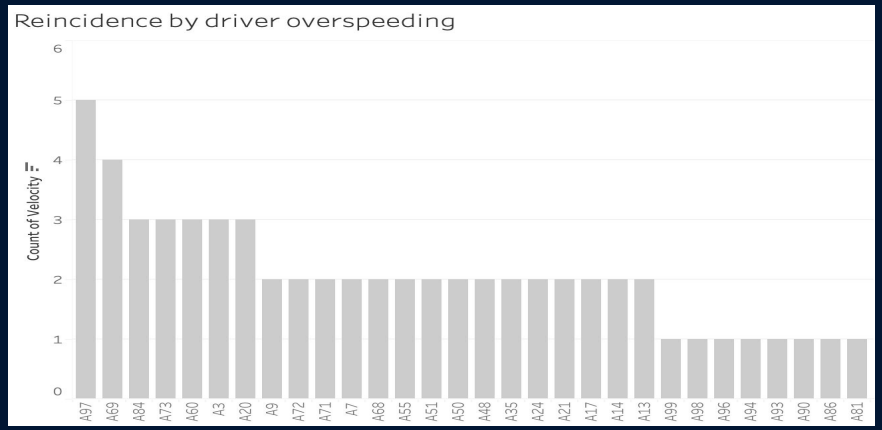
Idling events

Idling is a significant risk factor, with 47 drivers performing it more frequently than average, and Santa Rosa being the location with the highest incidence of idling events.

The average driver contributes to this event around 18.4 times during their tenure, with drivers A63 and A35 having the highest number of idling events.

The histogram is slightly right-skewed, indicating the presence of outliers who engage in excessive idling behavior.





Overspeeding

Speeding is less frequent than idling among drivers but still accounts for a significant number of risk events. Overspeeding is mainly observed in Santa Rosa, requiring the manager to implement measures to reduce losses.

Only seven drivers have more than three instances of speeding according to the histogram.Among them A97 has the most number of speeding incidents.

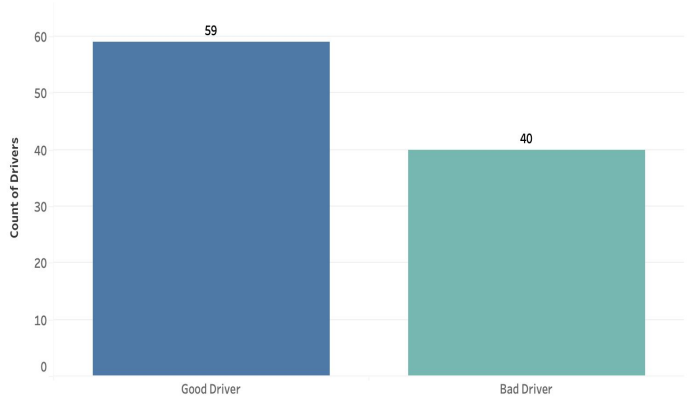
Dynamic Sets

Performace Based on Risk factor and Avg Mileage

	Average_Mileage > 5	Average_Mileage < 5
Risk_factor < 5	31	28
Risk_factor > 5	22	19

Driving Rating
- Good Driver
- Bad Driver

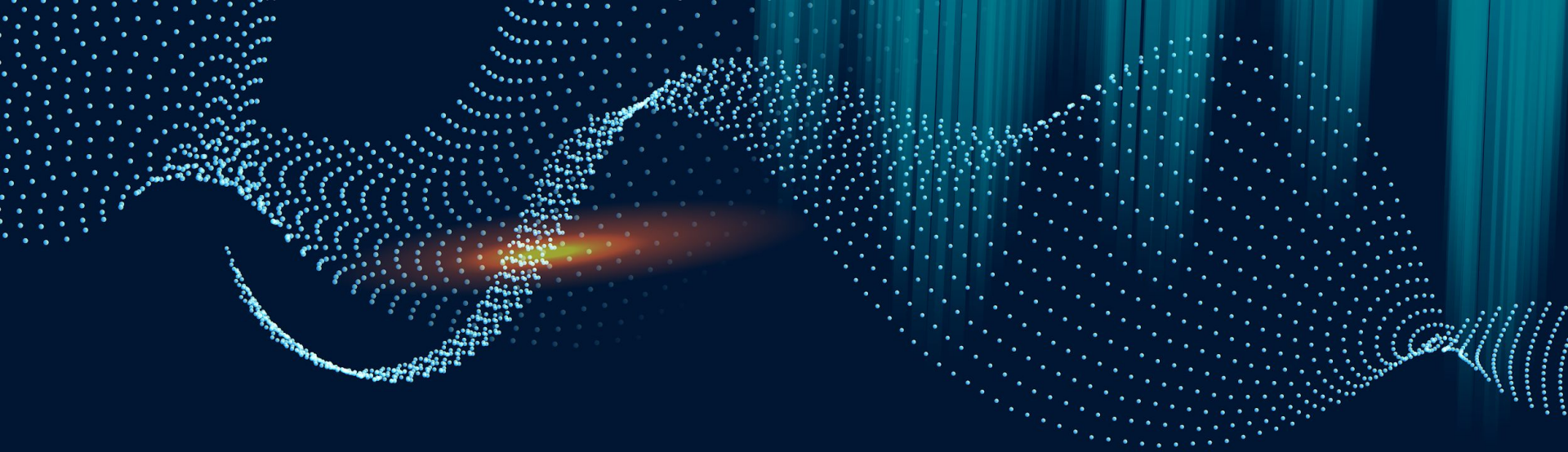
Count of Drivers based on Risk Factors



The drivers have been classified into two groups, namely, good drivers and bad drivers, based on their risk factors and the average mileage of their trucks.

Drivers are classified as good or bad based on their risk factors and truck mileage. 31 drivers have a risk factor of less than 5 and an average mileage greater than 5, indicating that they may be among the best drivers.

If only the risk factor is considered, 59 drivers have a risk factor of less than 5 and are labeled as 'good drivers.'



04 FORECASTING, INSIGHTS & RESULTS

Forecasting

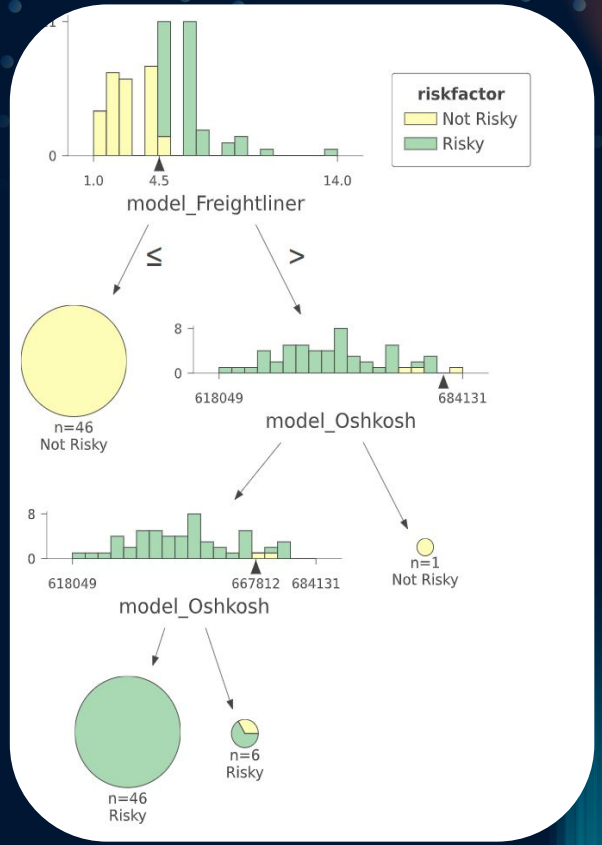
To avoid overfitting we decided to reduce the depth of the decision tree or limit the number of leaf nodes.

Hyperparameters:

- *max_depth : 3*

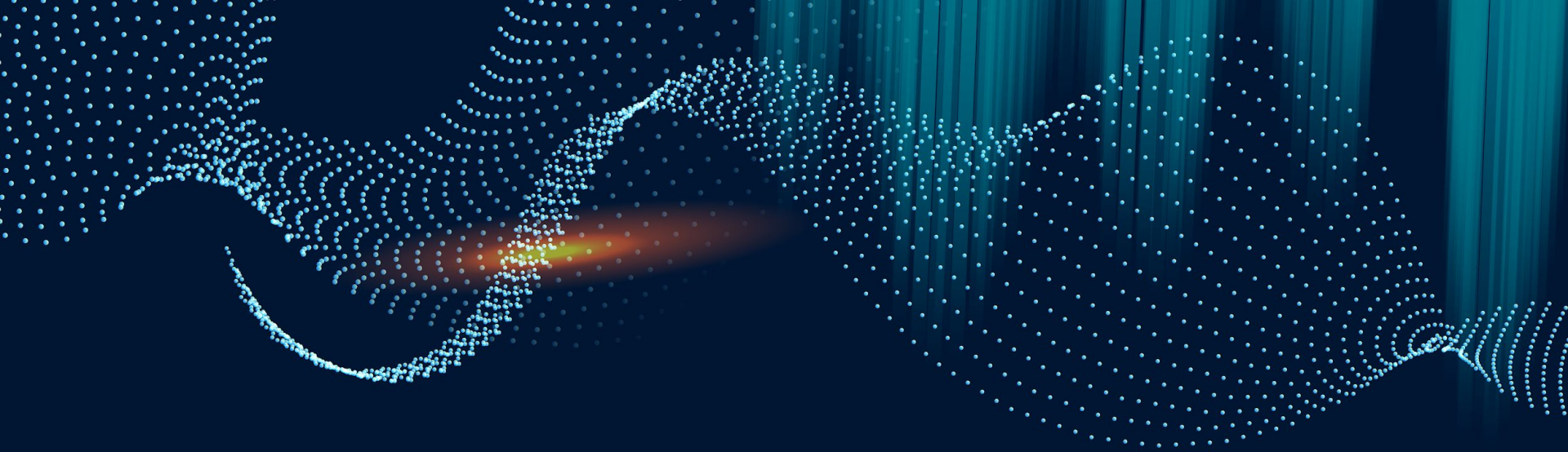
```
# train decision tree model
dtc = DecisionTreeClassifier(max_depth=3)
dtc.fit(X_train, y_train)
y_pred = dtc.predict(X_test)
dtc_accuracy = accuracy_score(y_test, y_pred)
print("Decision Tree Accuracy: ", dtc_accuracy)
```

```
Decision Tree Accuracy:  0.95
```



Insights: Practical Suggestions for Improved Results

- *Adding a layer of background check to identify potentially dangerous drivers before they are hired.*
- *Increasing the frequency of safety inspections to weed out potential risky drivers.*
- *Embedding electronic logging devices to record a driver's KPI's.*
- *Coming up with strategies to improve Driver Training by introducing workshops and distributing regularly updated manuals.*



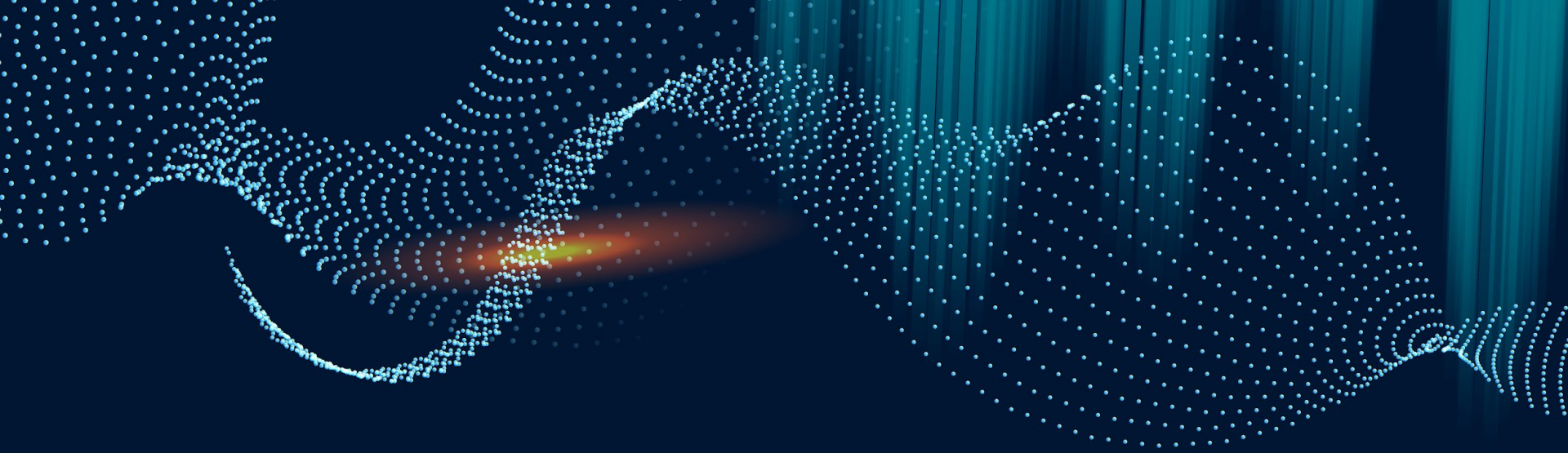
05 CHALLENGES FACED & RESOURCES

Challenges faced

- Connecting HIVE with Spark.
- Connecting Tableau to the server causes the workflow to be slower, since it must process data from the server rather than a local resource.
- Learning how to model data in Tableau and joining the different tables:
  - Working on Dynamic Sets and implementing different visualizations was challenge and finally collaborating all the sheets in Dashboard.

Sources

- *“A&I Online - Motor Carrier Analysis and Information Resources Online.” Ai.fmcsa.dot.gov, ai.fmcsa.dot.gov/CrashStatistics/. Accessed 28 Apr. 2023*.



THANK YOU !