Analysis of the transport network and Passenger load in the Valle de Aburrá

El futuro digital es de todos | MinTIC

correlation·one
DATA SCIENCE FOR ALL™
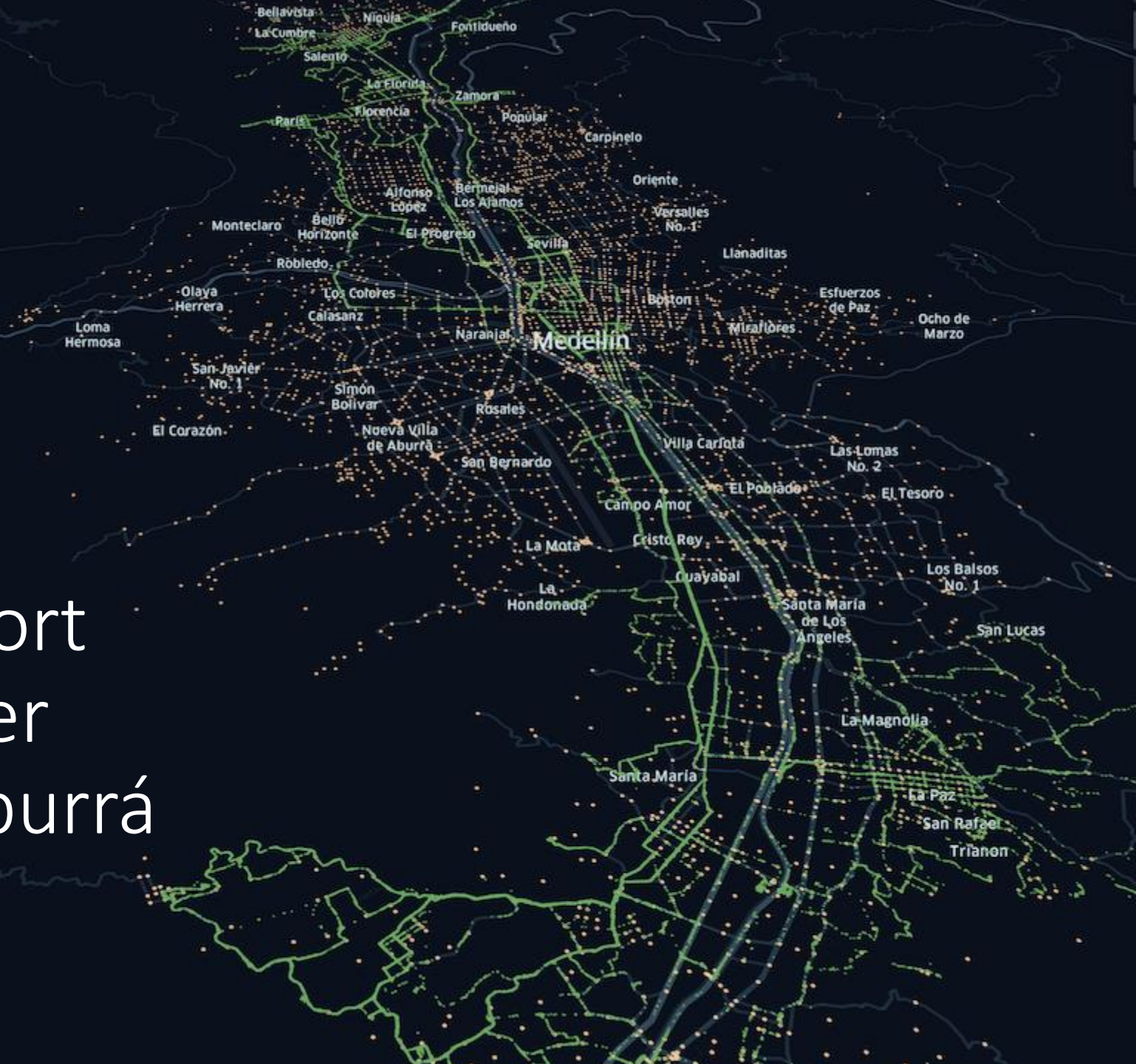
Meet Our Team

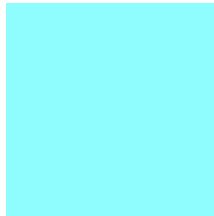FELIPE GARCIA
/felipe-garcia-lizalda-ba39651a6

SANDRA RUIZ
/sandra-milena-ruiz-reyes

CHRISTIAN VELEZ
/cvelez

CARLOS TAIMAL
/taimalyepes

SEBASTIAN MARIN
/

BORIS MARTINEZ
/boris-martinez-0265011b2

JORGE SAAVEDRA
/jorgesaavedragomez
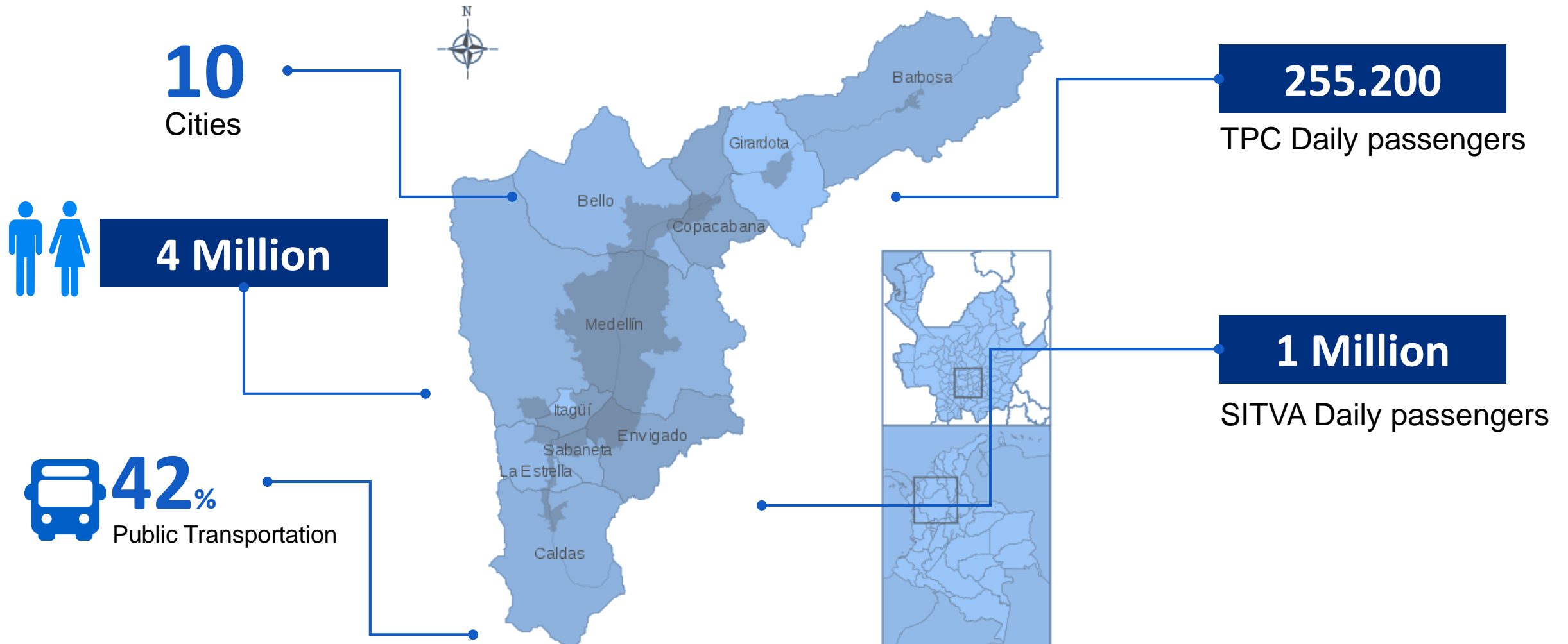
TEAM #21

El futuro digital es de todos   MinTIC   correlation·one
DATA SCIENCE FOR ALL™

PROBLEM BACKGROUND

10
Cities

4 Million

42%
Public Transportation

255.200
TPC Daily passengers

1 Million
SITVA Daily passengers

Barbosa
Girardota
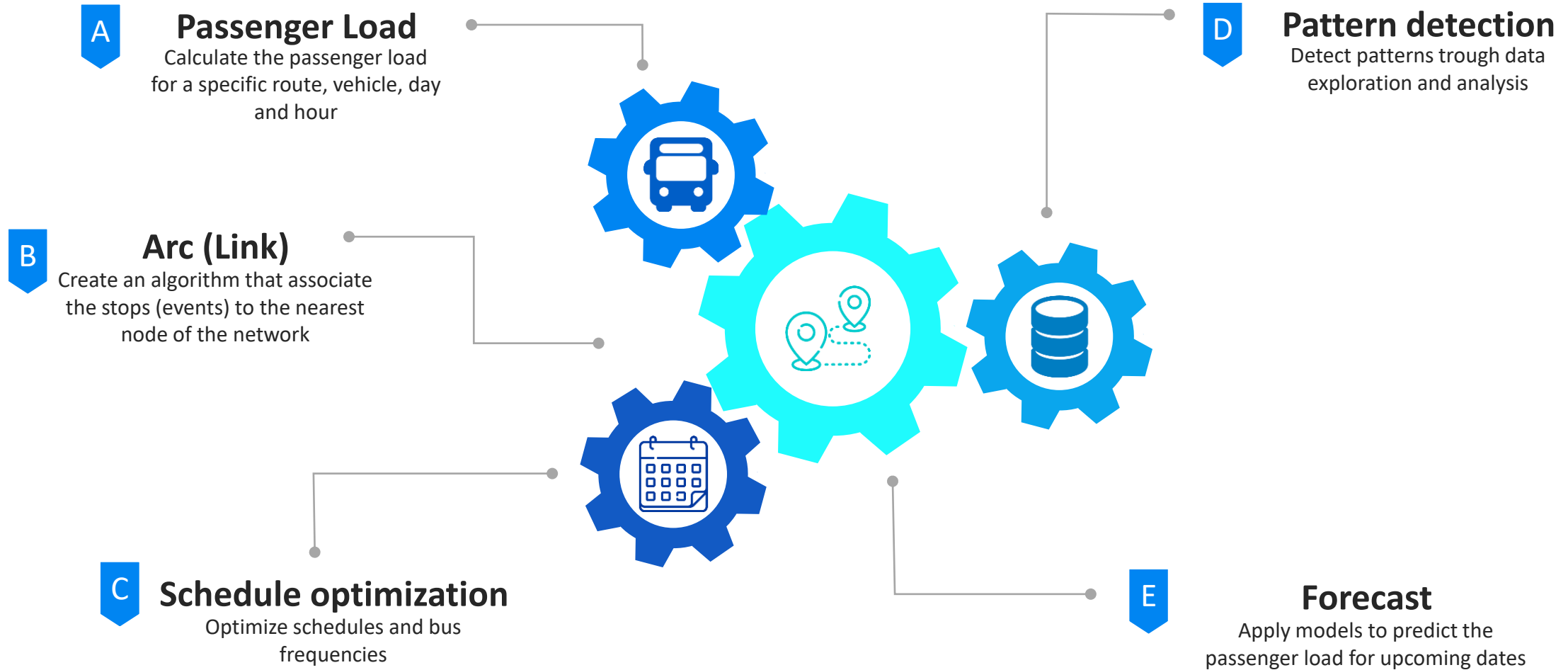Bello
Copacabana
Medellín
Itagüí
Envigado
Sabaneta
La Estrella
Caldas

SITVA

The Integrated transportation system of the Valle de Aburrá (SITVA) is conformoed by the Metro, Cable Transport, Tram-Train, exclusive buses and the Subsystem of collective public transport of passengers (TPC)
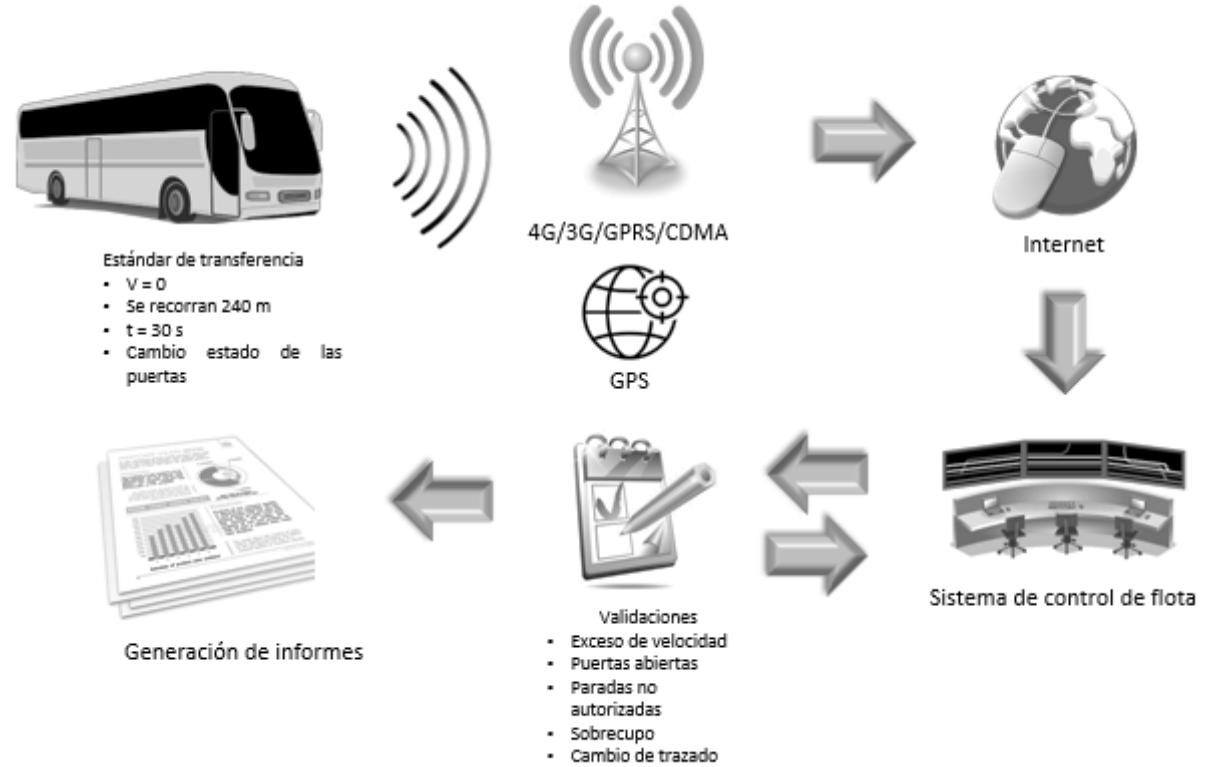
GOALS

A Passenger Load
Calculate the passenger load for a specific route, vehicle, day and hour

B Arc (Link)
Create an algorithm that associate the stops (events) to the nearest node of the network

C Schedule optimization
Optimize schedules and bus frequencies

D Pattern detection
Detect patterns trough data exploration and analysis

E Forecast
Apply models to predict the passenger load for upcoming dates

GATHERING DATA

75% Of the routes

59 Companies

3850 Vehicles

Estándar de transferencia
- V = 0
- Se recorran 240 m
- t = 30 s
- Cambio estado de las puertas

4G/3G/GPRS/CDMA

GPS

Internet

Sistema de control de flota

Validaciones
- Exceso de velocidad
- Puertas abiertas
- Paradas no autorizadas
- Sobrecupo
- Cambio de trazado

Generación de informes

Vehicles with Sensors

Capture time, latitude, longitude, # of passengers boarding and alighting.

Transmitted

Data Transmitted online every 3 minutes or less or also in batch to the GTPC
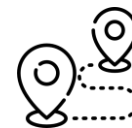
KML Files

XML file for route visualization within 2D and 3D maps

DATA

2019-11 to 2020-05
6 Million records per day

TECHNICAL ARCHITECTURE

Data Transformation Flow

| Data Source | Centralized DB | Analysis & Report | Dashboard |
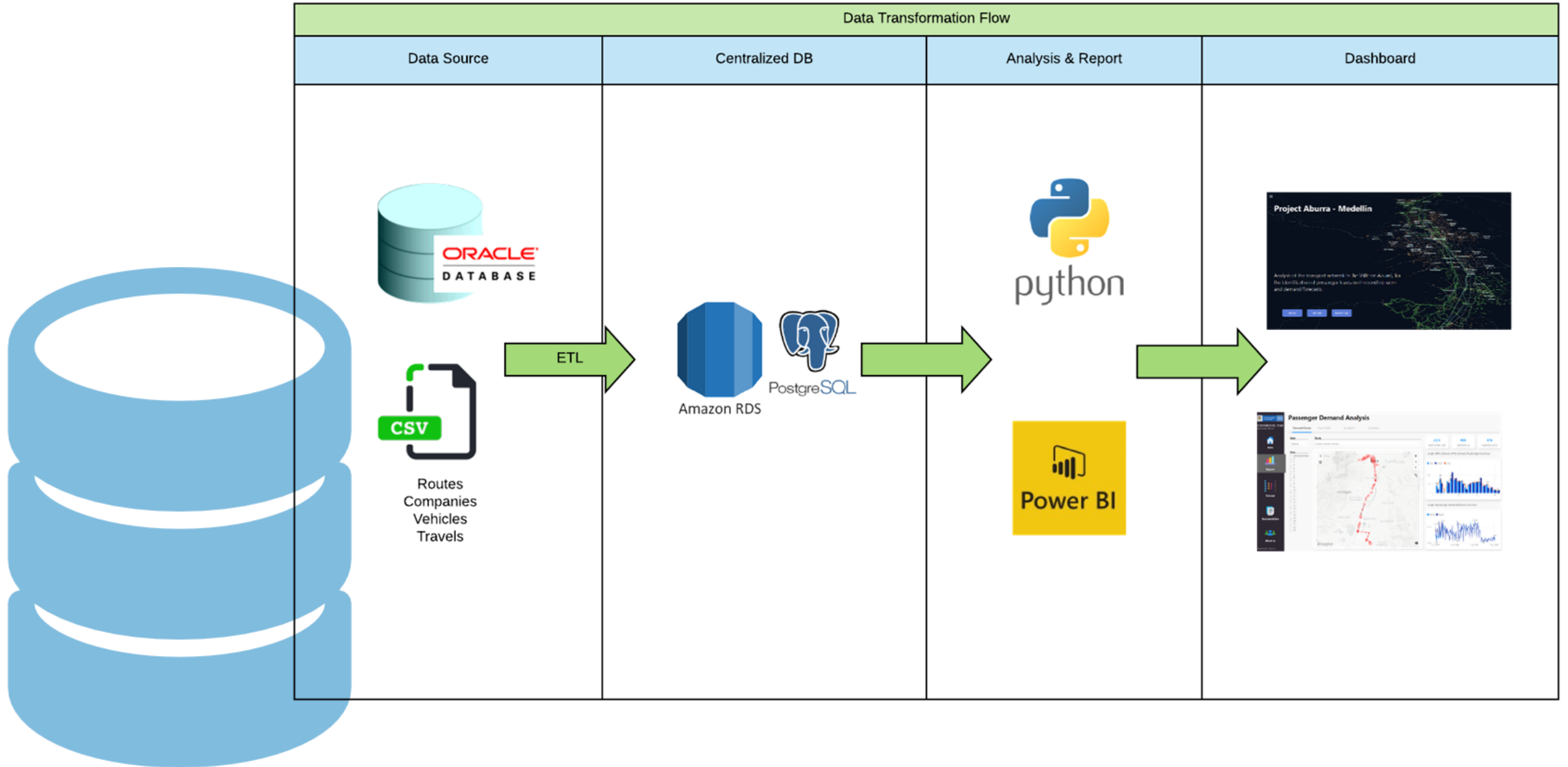
DATA CLEANSING & FINDINGS

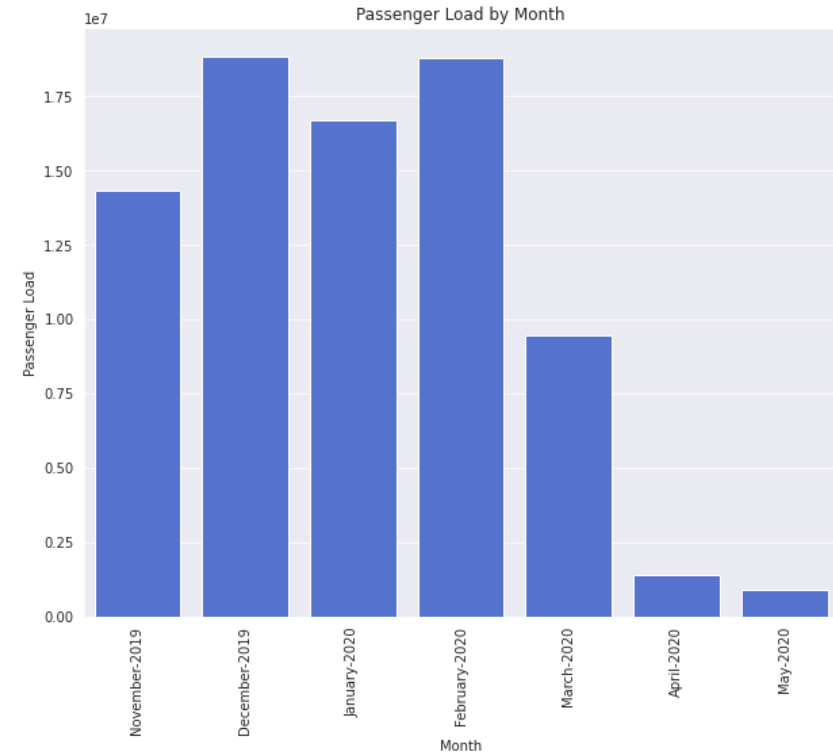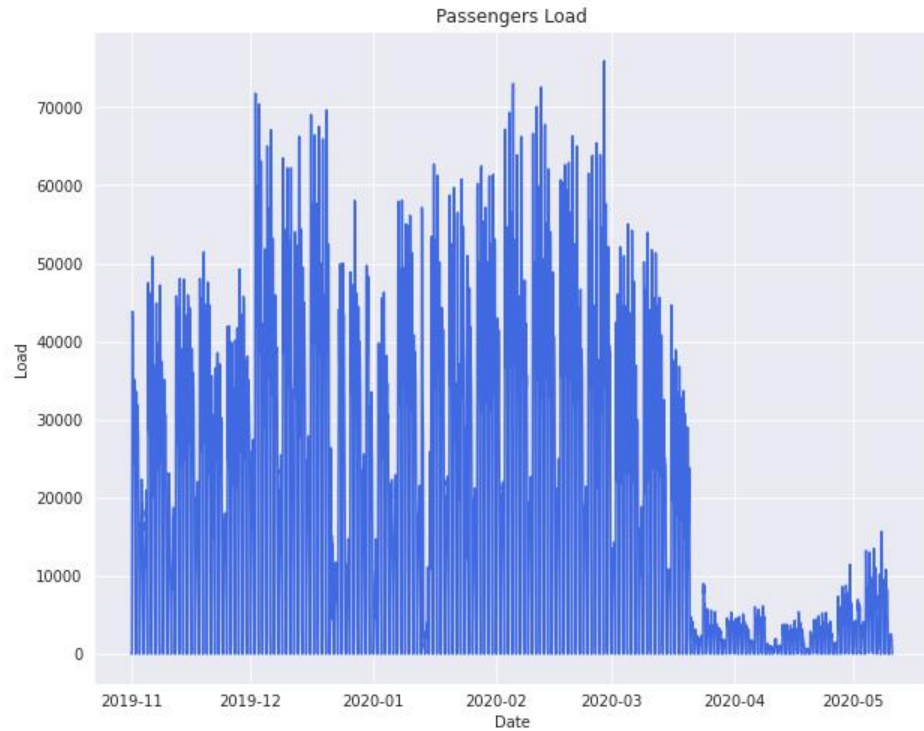|	
1	Values between 0 and 99 for the number of Passengers boarding and alighting.
2	Latitude and longitude that don't belong to the Valle de Aburrá Metropolitan Area.
3	Records in unconventional hours.
4	Trips that start with people alighting

TIME SERIES

Passenger Load
There is a stational pattern that repeats as weeks goes by.

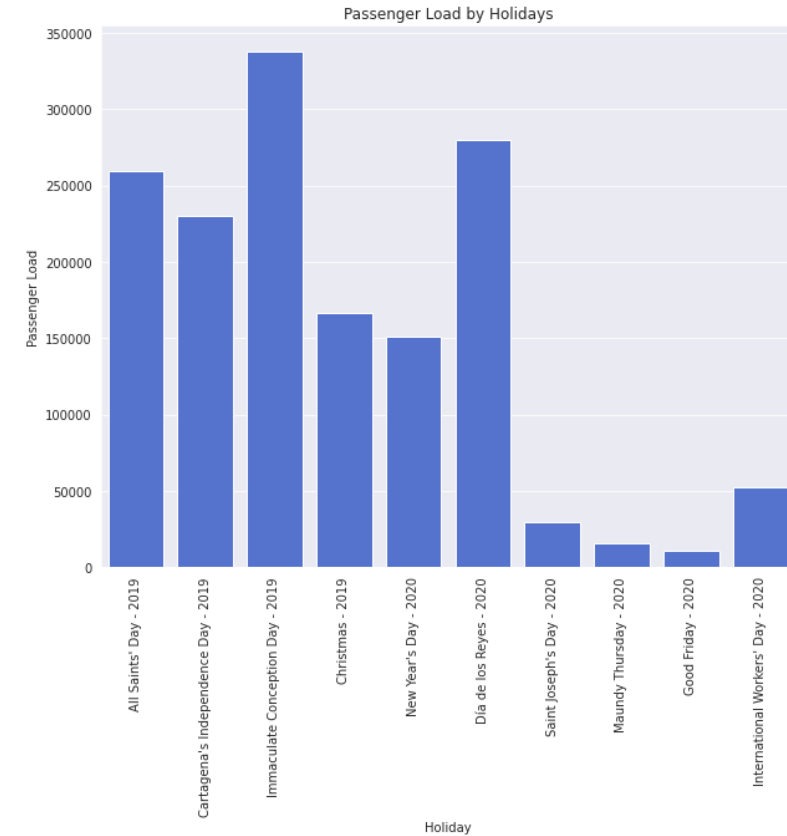**Passengers Load by Month**
The maximum demand was reached around February and the minimums are reached after the lockdown as a result of the COVID-19 pandemic

TIME SERIES
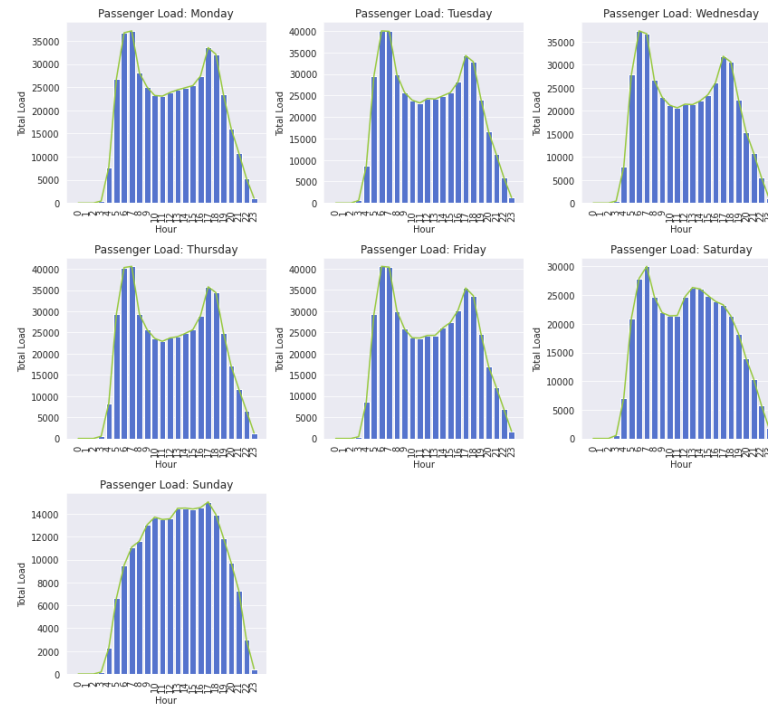
Passengers by day

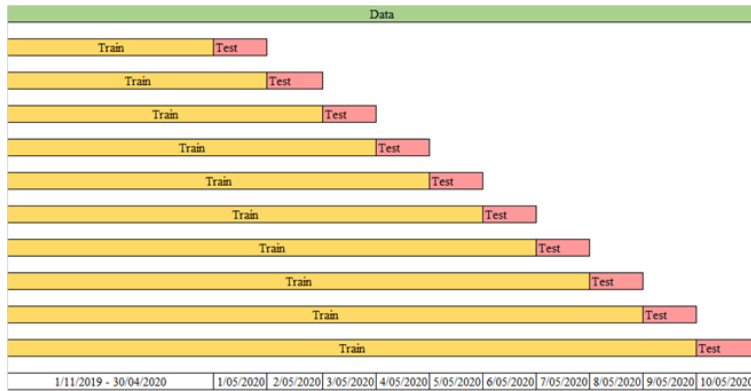There is a peak between 6 and 7 am which corresponds to the beginning of the workday.

Passengers by Holidays

The Immaculate Conception Day is the one for which there's a higher public transportation demand;

PREDICTIVE MODEL

Random Forest Algorithm

ARIMA/SARIMA Models
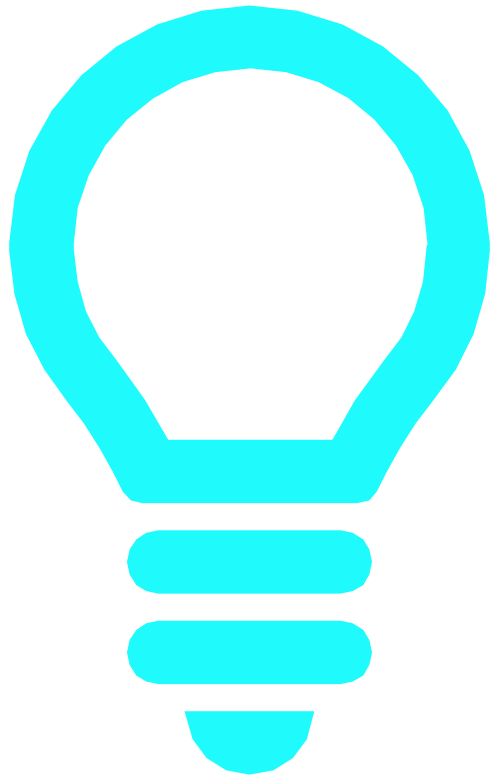
DASHBOARD

https://tinyurl.com/y5ao296g

Analysis of the transport network in the Valle de Aburrá, for the identification of passenger loads on the road network and demand forecasts.

INTRO   REPORT   DATAFOLIO

CONCLUSIONS AND SUGGESTIONS

1

Exploratory data analysis showed that despite the huge volume of information gathered daily from the vehicle's sensors, only around 20% of the data has values about boarding and alighting greater than zero and it was valuable for analysis

2

Inconsistencies in data like the recording of events at unconventional hours, passenger alighting at the beginning of the route, more passengers getting off than boarding, or extreme values in both the number of passengers who get on and those who get off, distort the calculation of the load of passengers in each link

1

It is important to adopt techniques that guarantee the quality of the information, specifically for the geolocation coordinates and the boarding and alighting values capture by the sensors

2

For future analysis, the analytical component must be expanded. As a first phase this work has used the analysis for forecasting the demand of passengers but can be improved with a better data cleaning and an optimization of the variables in the models

THANK YOU