GHENT
UNIVERSITY

FACULTY OF SCIENCES

MASTER OF STATISTICAL DATA ANALYSIS
BAYRAK, CARTUYVELS, HEYVAERT, KOURFALI, O'SULLIVAN

DATABASES

SPARK PROJECT FOR NMBS / 14.12.2017

GHENT UNIVERSITY

BELGIAN PUBLIC TRANSPORTATION

Explore

Identify Problems

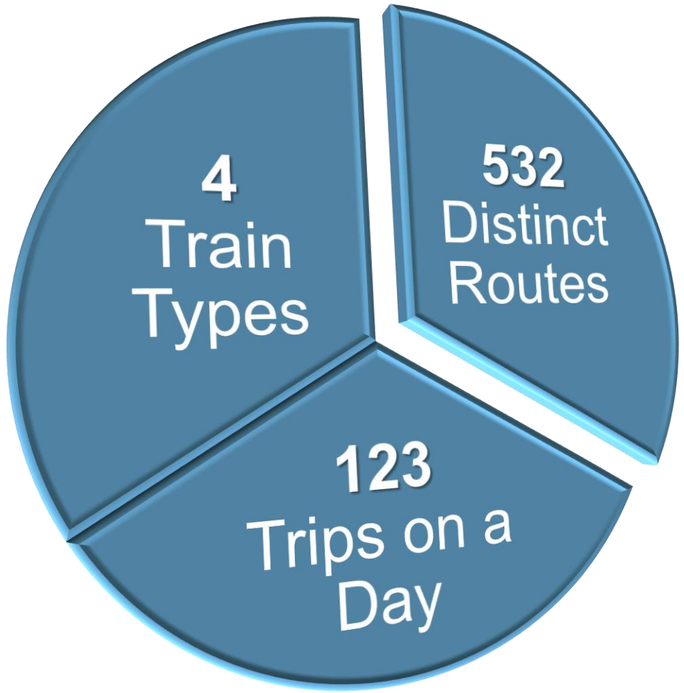
Suggest Solutions

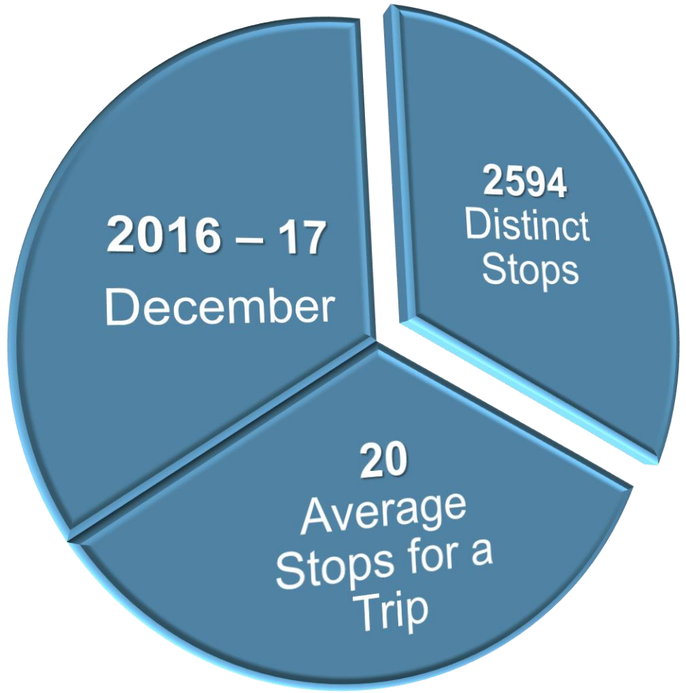
GHENT UNIVERSITY

DESCRIPTIVE STATISTICS

GHENT UNIVERSITY

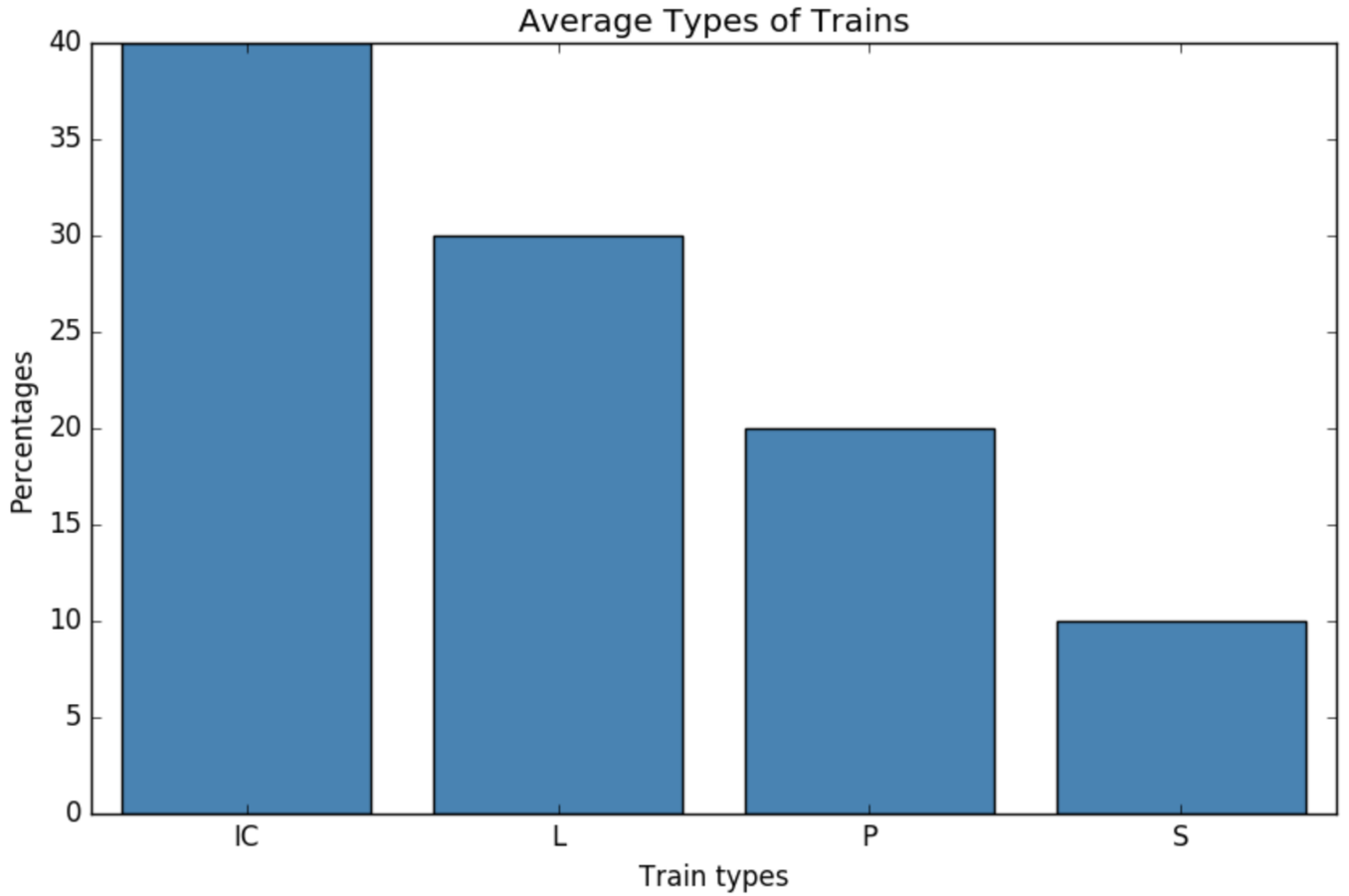
DESCRIPTIVE STATISTICS





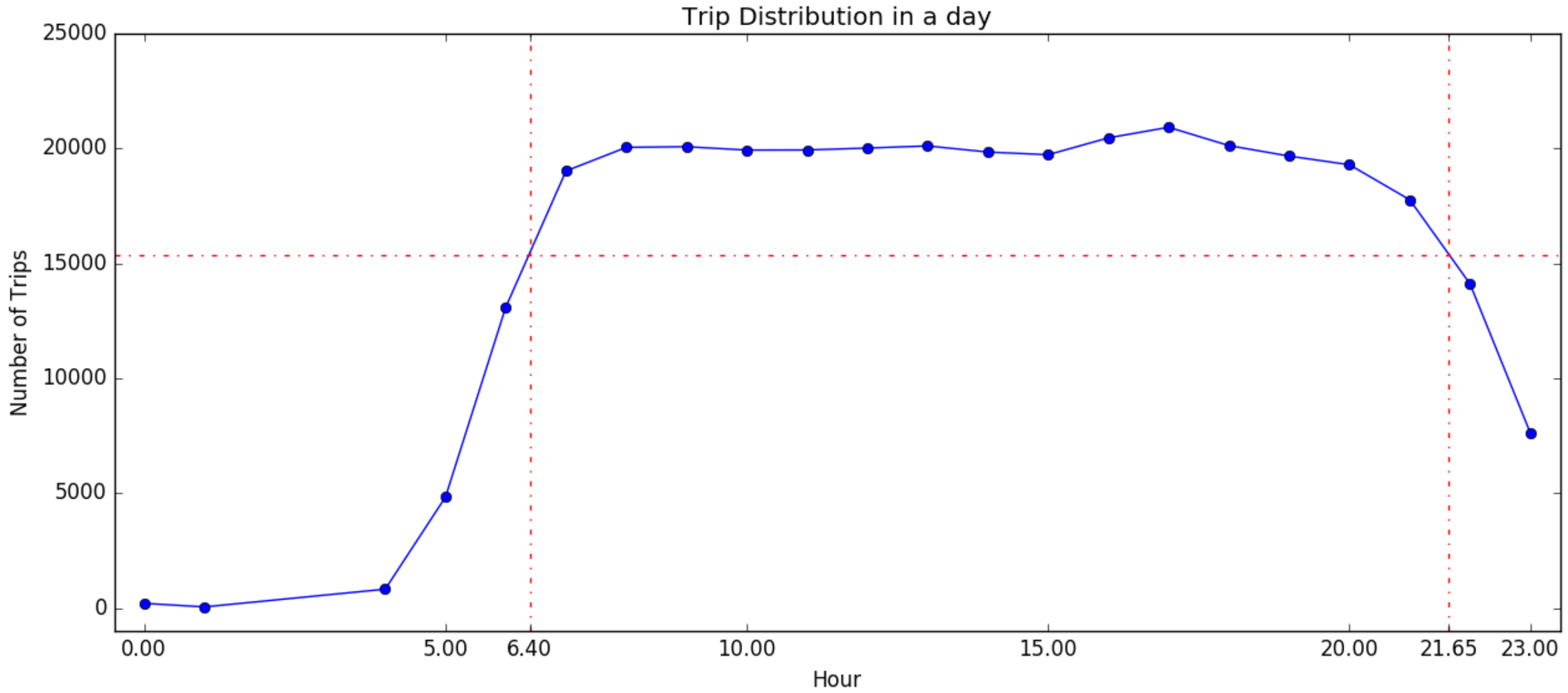
GHENT UNIVERSITY

DESCRIPTIVE STATISTICS



GHENT UNIVERSITY

DESCRIPTIVE STATISTICS

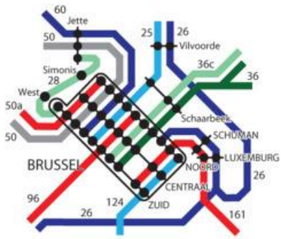


GHENT UNIVERSITY

INCOMING – OUTGOING TRAINS

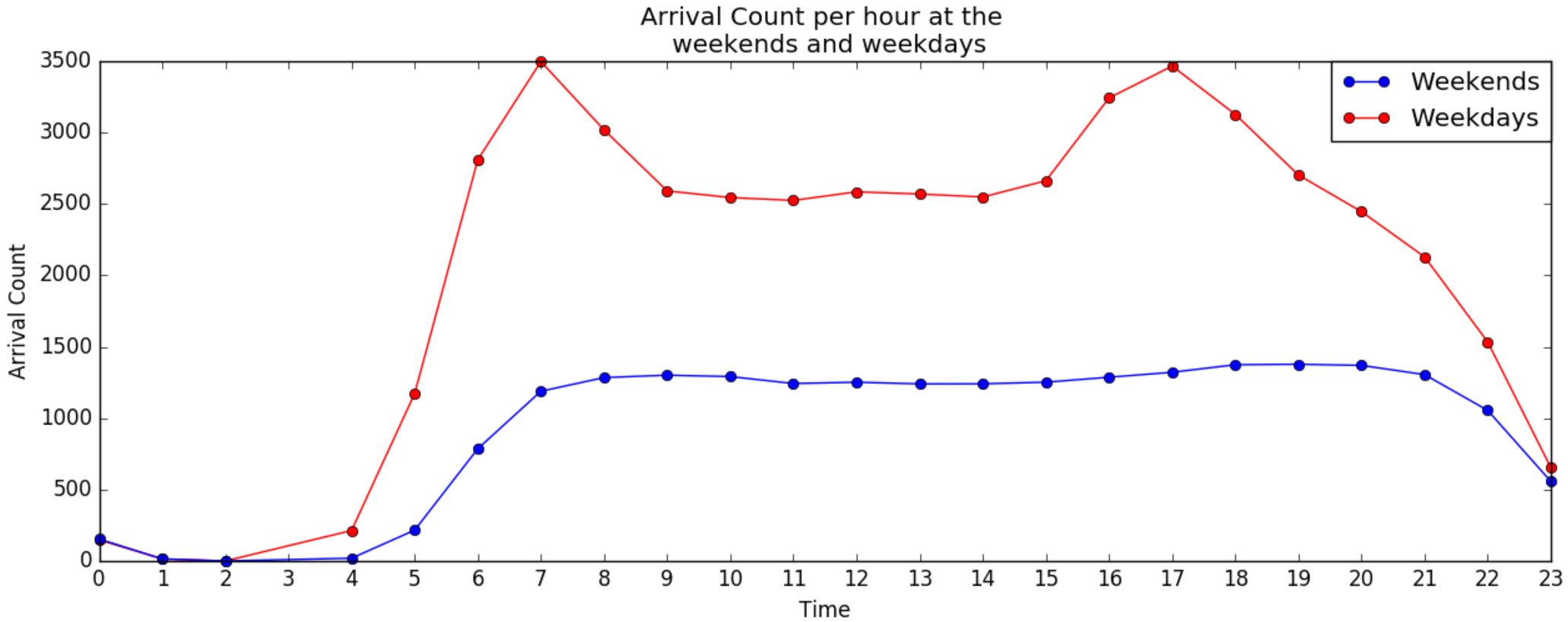
GHENT UNIVERSITY

INCOMING – OUTGOING TRAINS



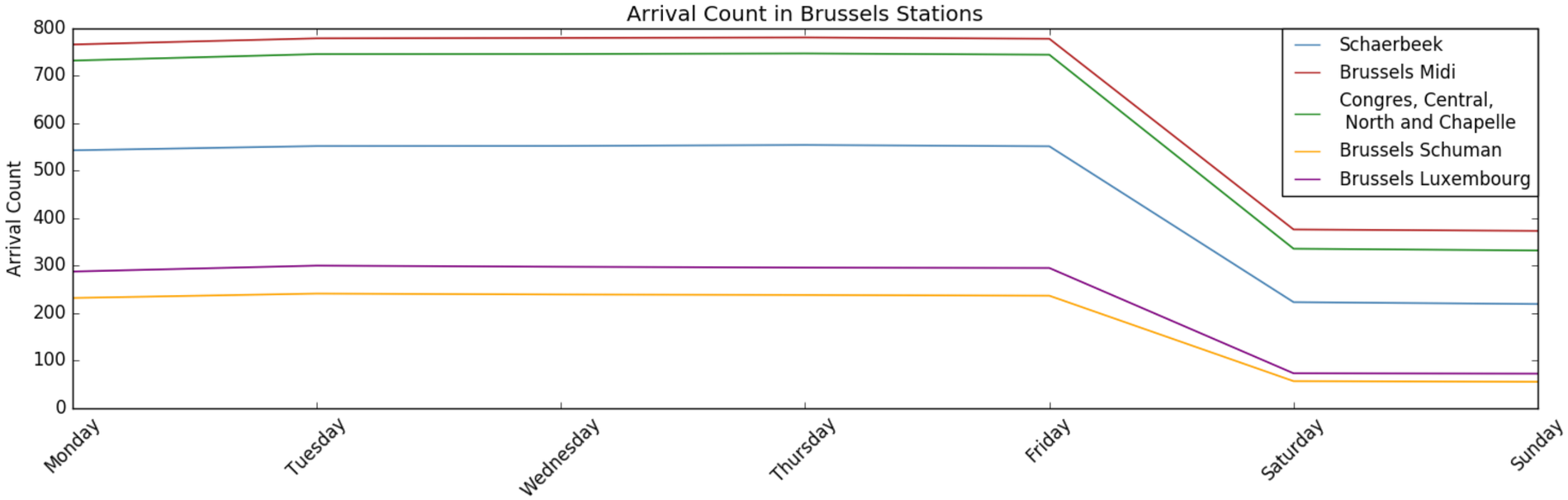
GHENT UNIVERSITY

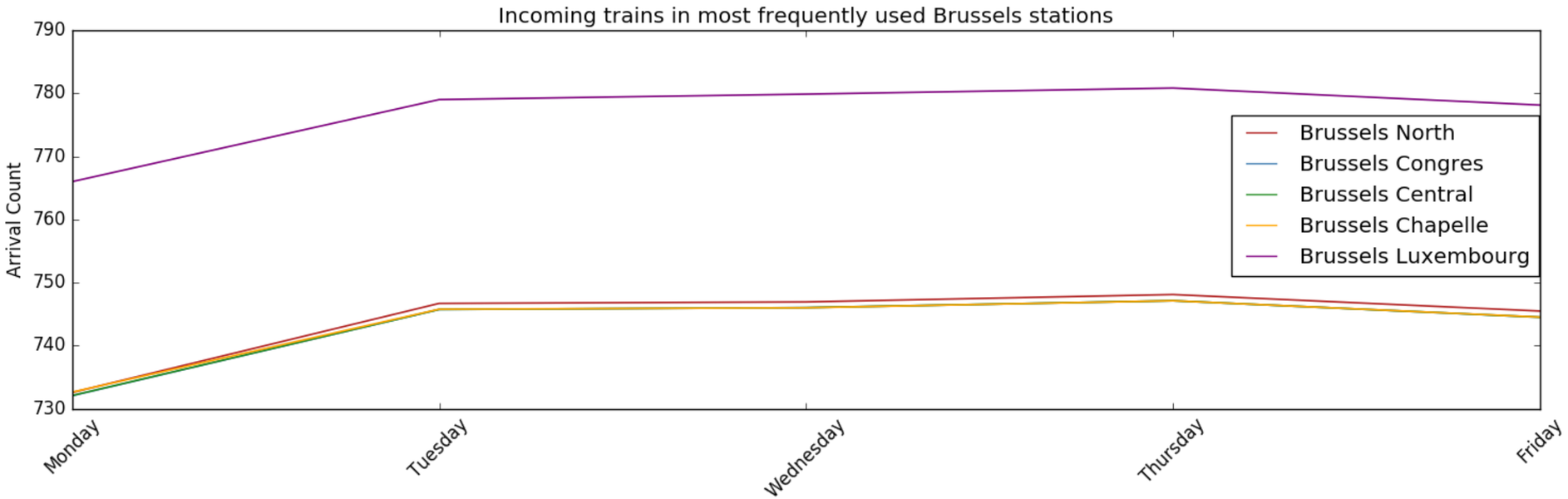
INCOMING – OUTGOING TRAINS



GHENT
UNIVERSITY

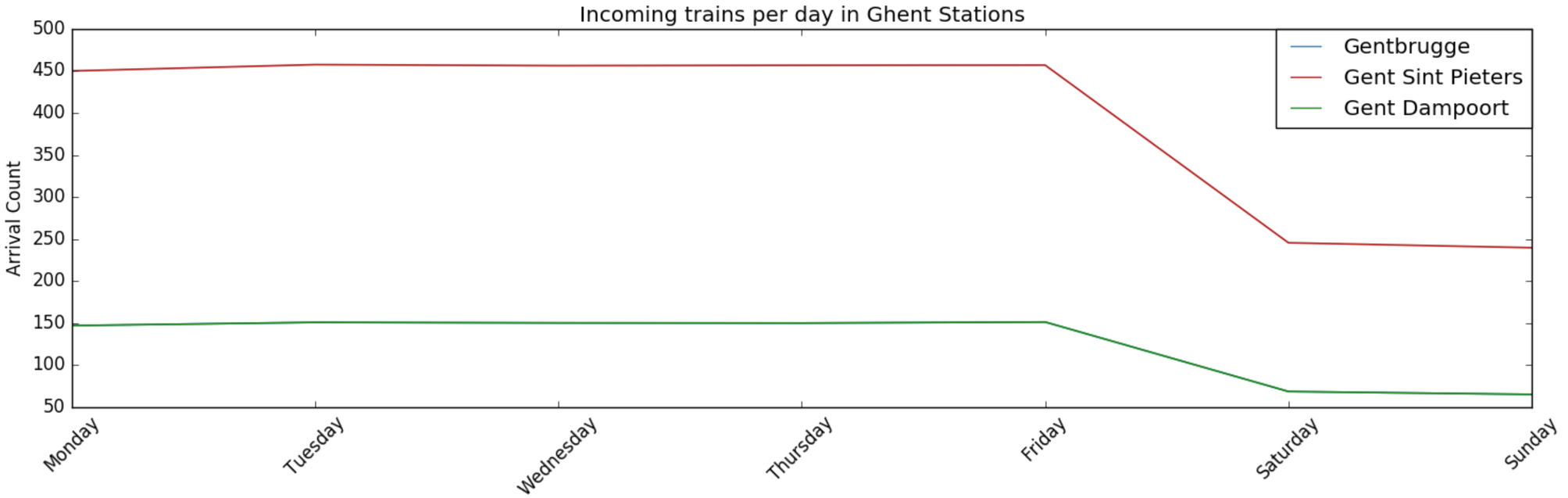
INCOMING – OUTGOING TRAINS

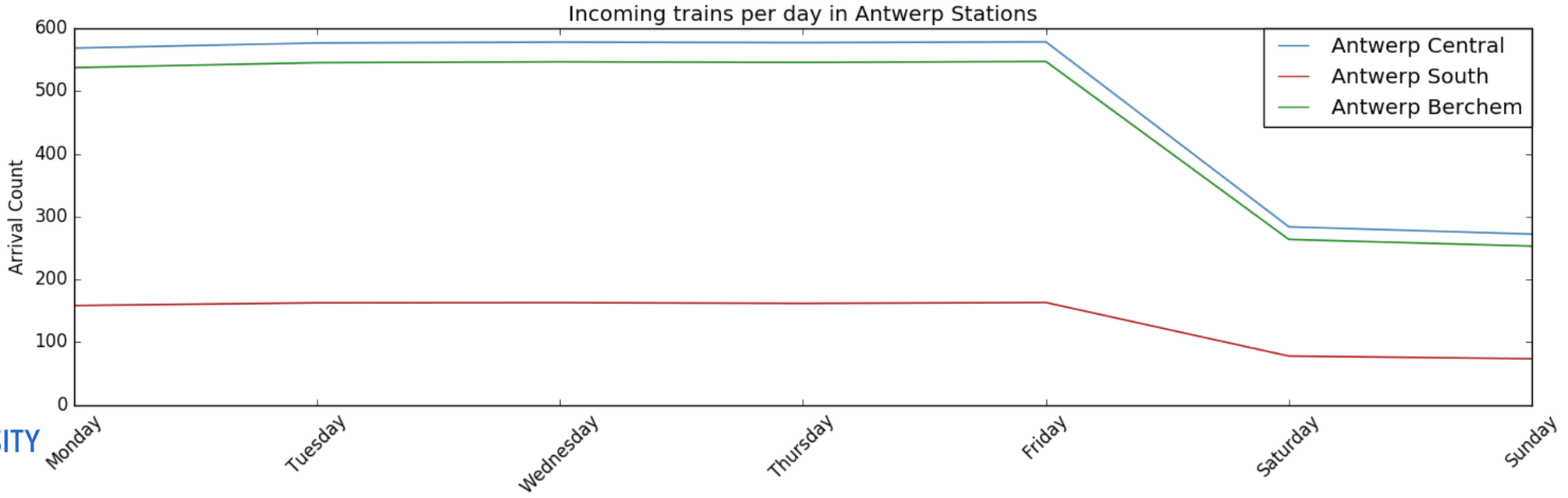




GHENT UNIVERSITY

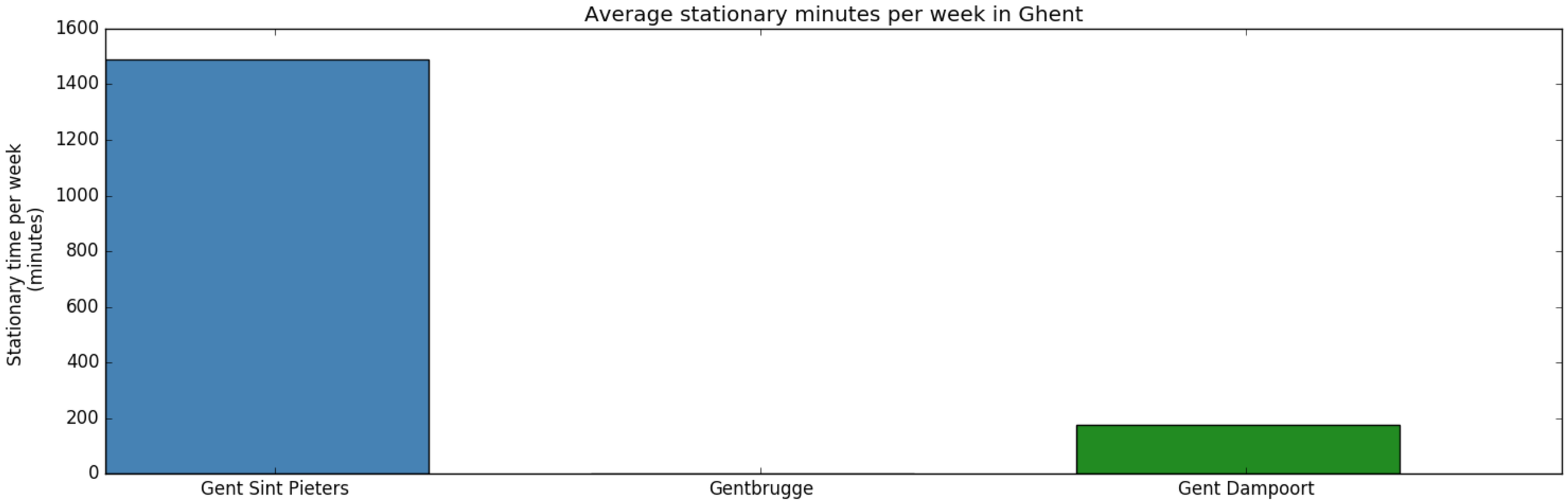
INCOMING - OUTGOING TRAINS





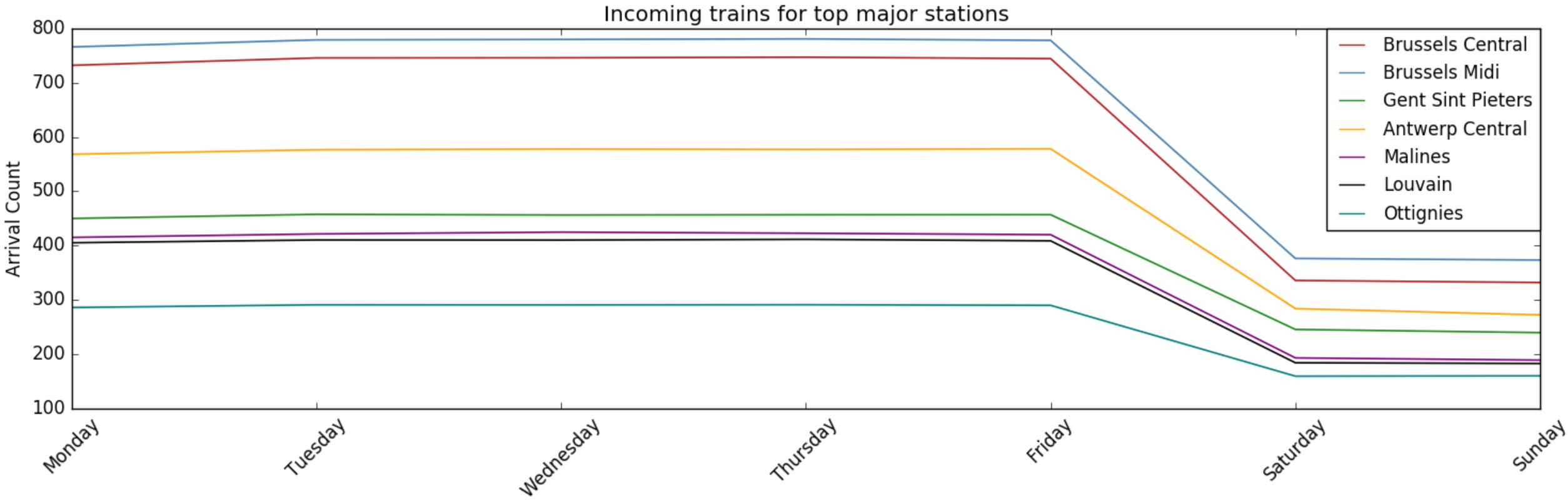
GHENT UNIVERSITY

INCOMING – OUTGOING TRAINS



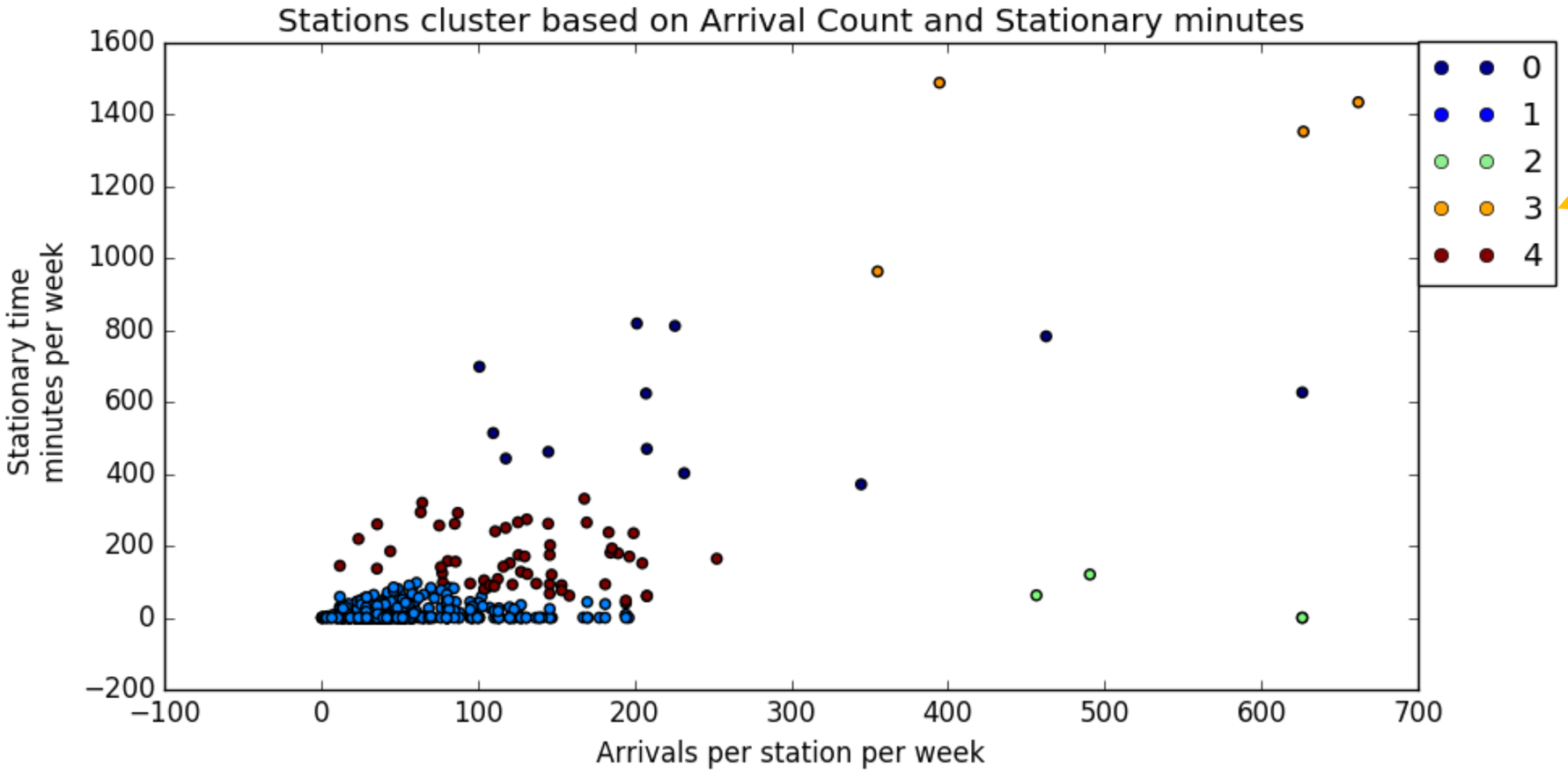
GHENT UNIVERSITY

INCOMING - OUTGOING TRAINS



GHENT UNIVERSITY

INCOMING – OUTGOING TRAINS



- Brussels Nord
- Brussels Midi
- Malines
- Gent Sint Pieters

GHENT UNIVERSITY

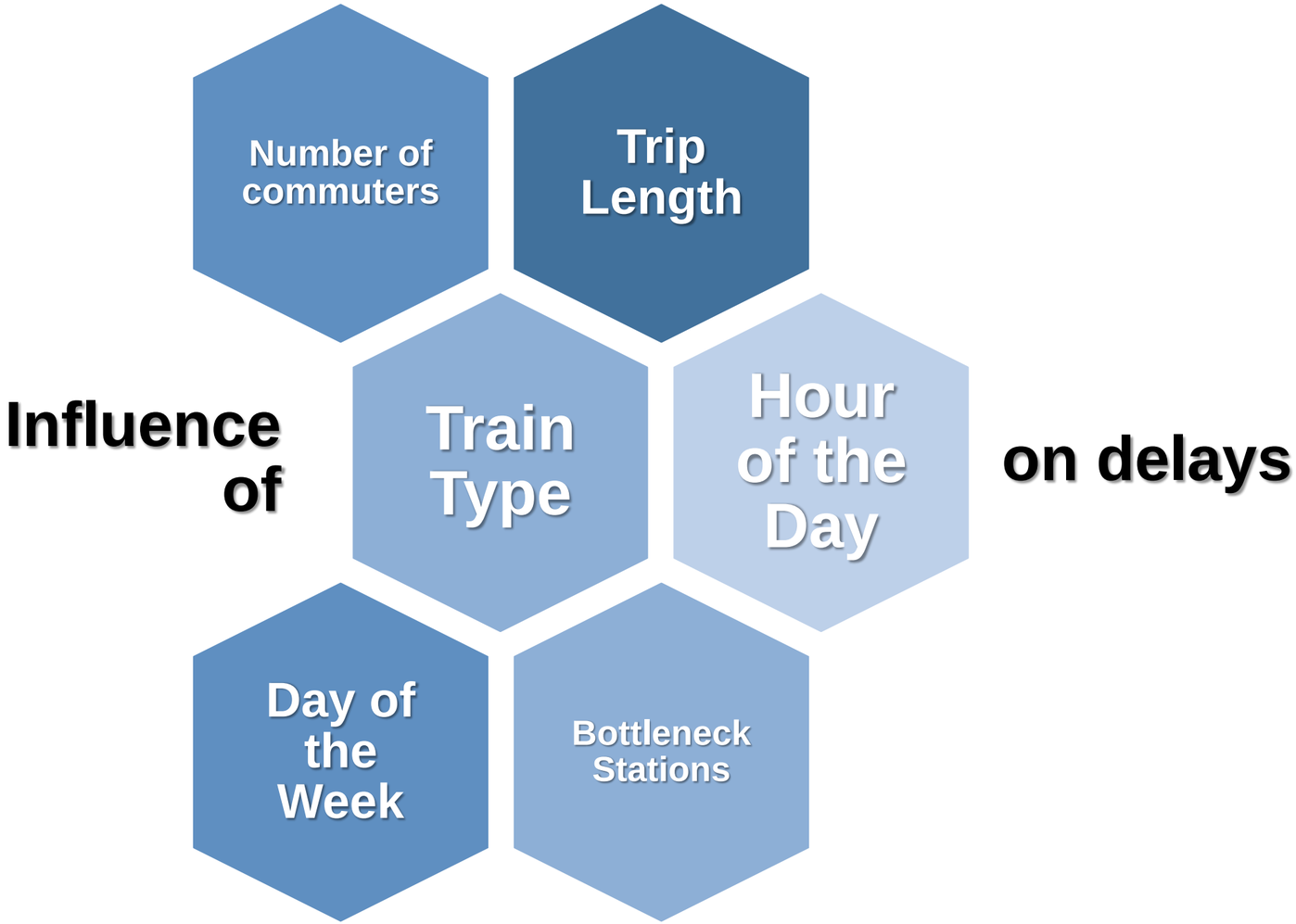
ANALYSIS OF DELAYS

GHENT
UNIVERSITY

ANALYSIS OF DELAYS

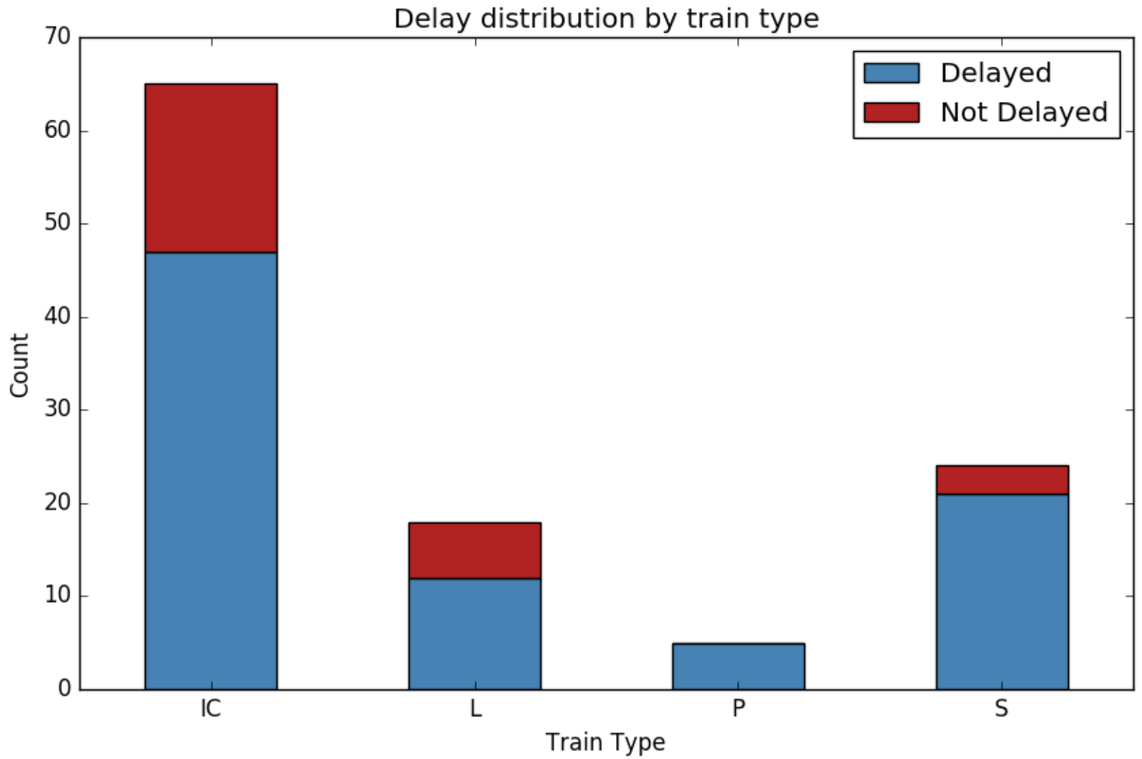
GHENT UNIVERSITY

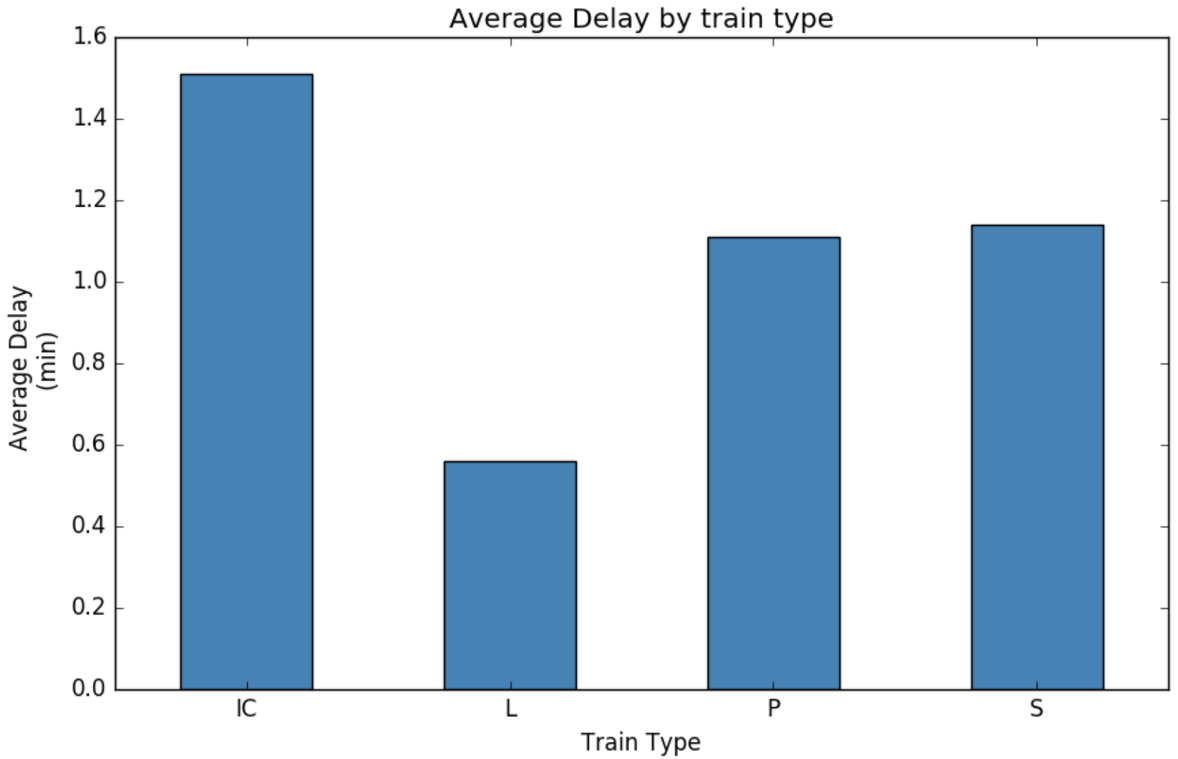
ANALYSIS OF DELAYS



GHENT UNIVERSITY

ANALYSIS OF DELAYS

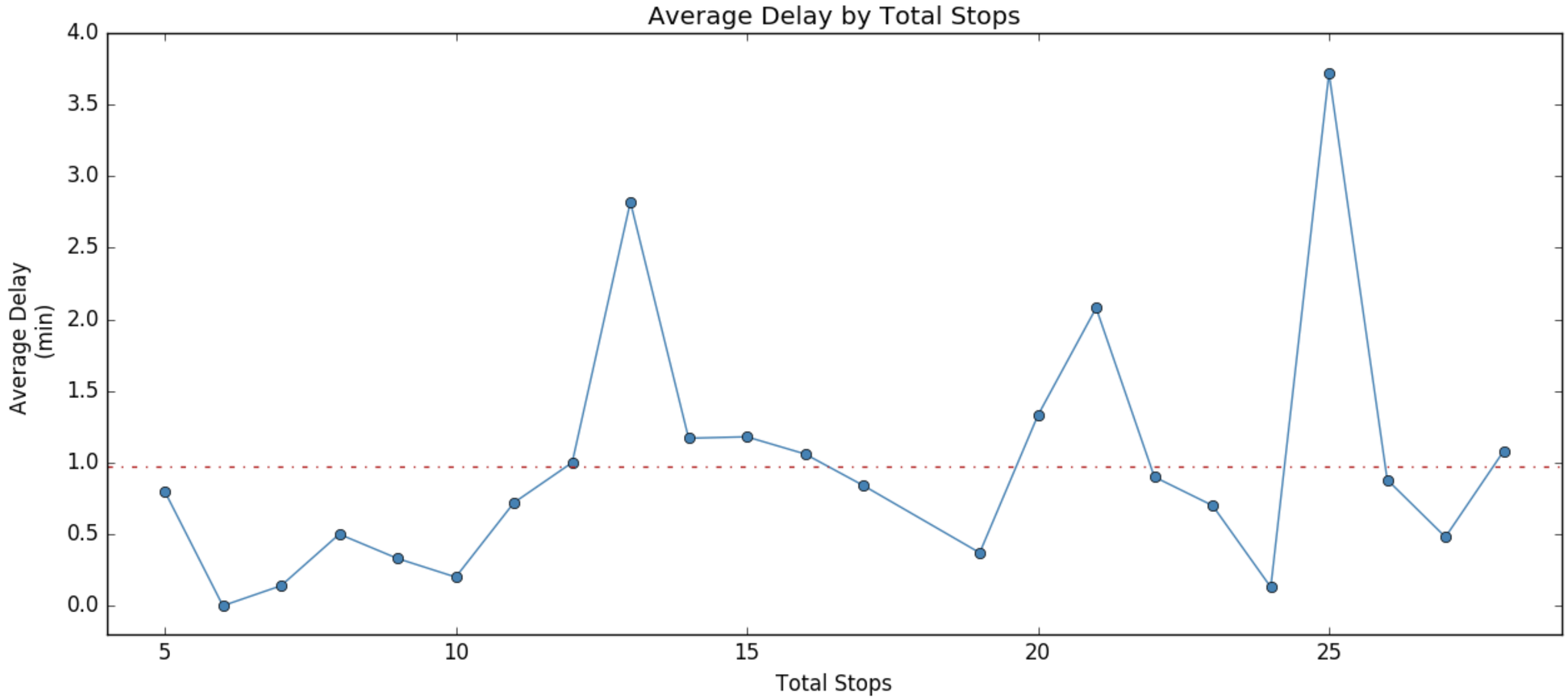




* Not delayed: average delay time <= 0.25 min

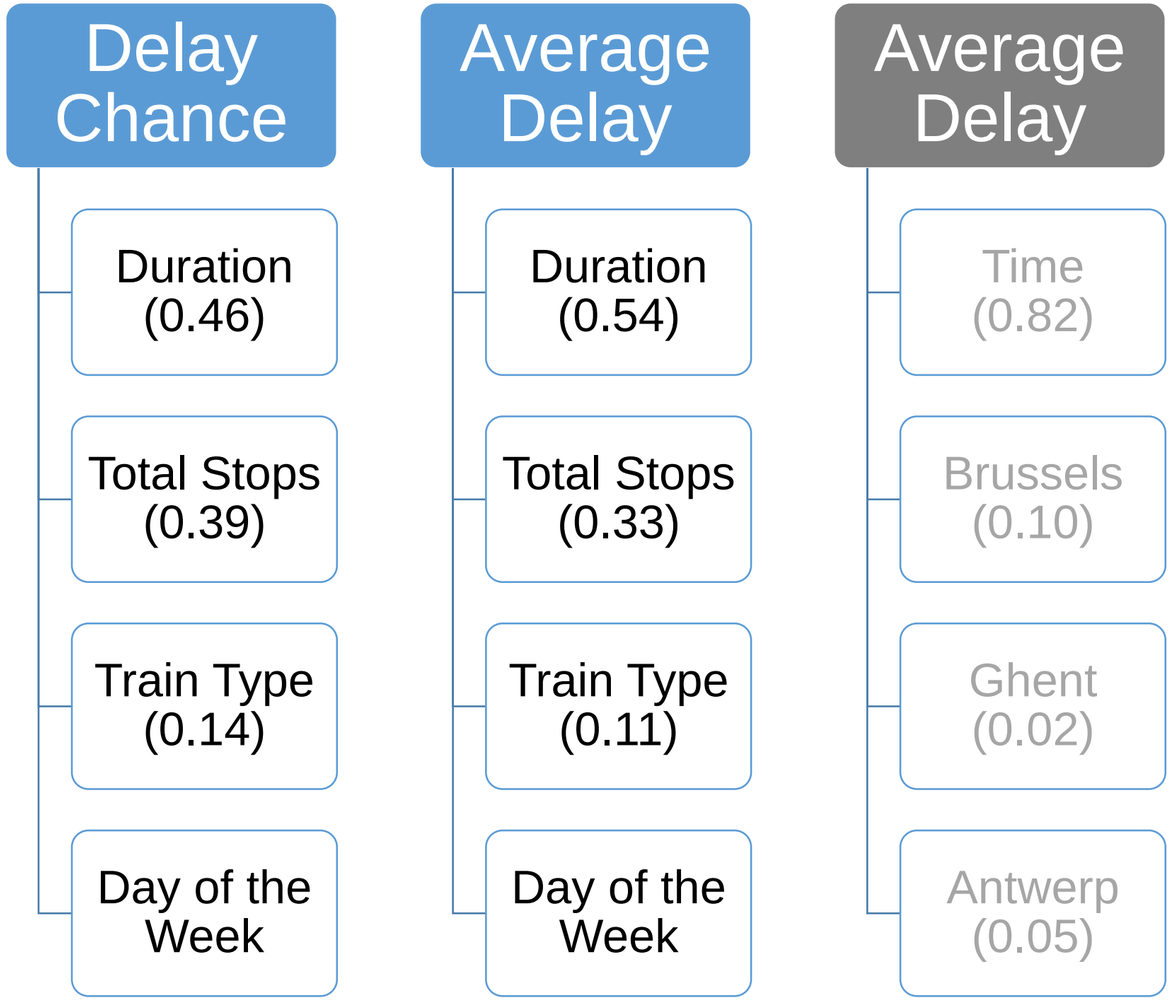
GHENT UNIVERSITY

ANALYSIS OF DELAYS



GHENT UNIVERSITY

RANDOM FOREST



Random Forest Classification

AUC = 0.55

Accuracy = 69%

Random Forest Regression

Random Forest Regression

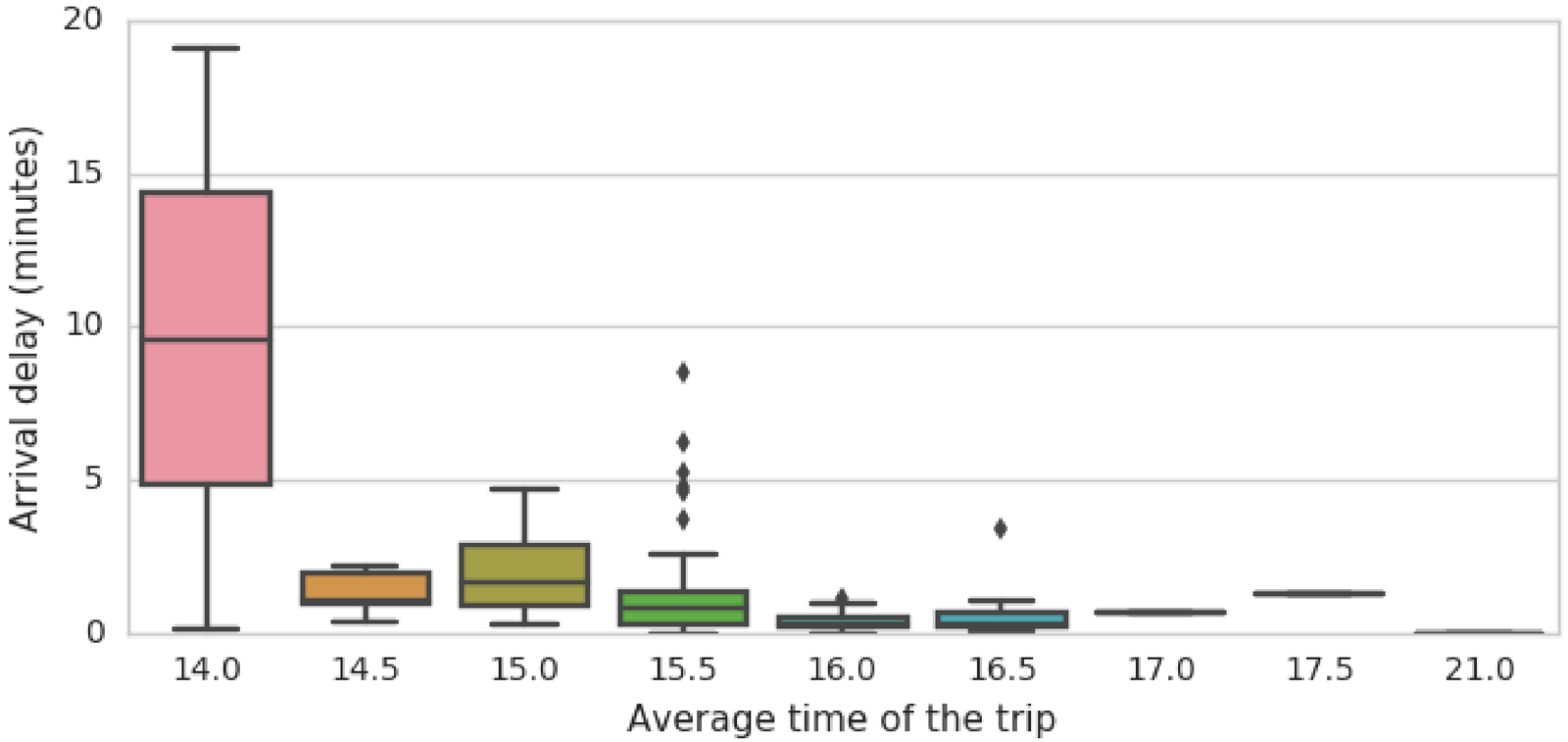
GHENT UNIVERSITY

INFLUENCE OF TIME OF DAY

- Is the delay time longer at certain moments?
- Include possible bottleneck cities (Brussels, Ghent, …)

GHENT UNIVERSITY

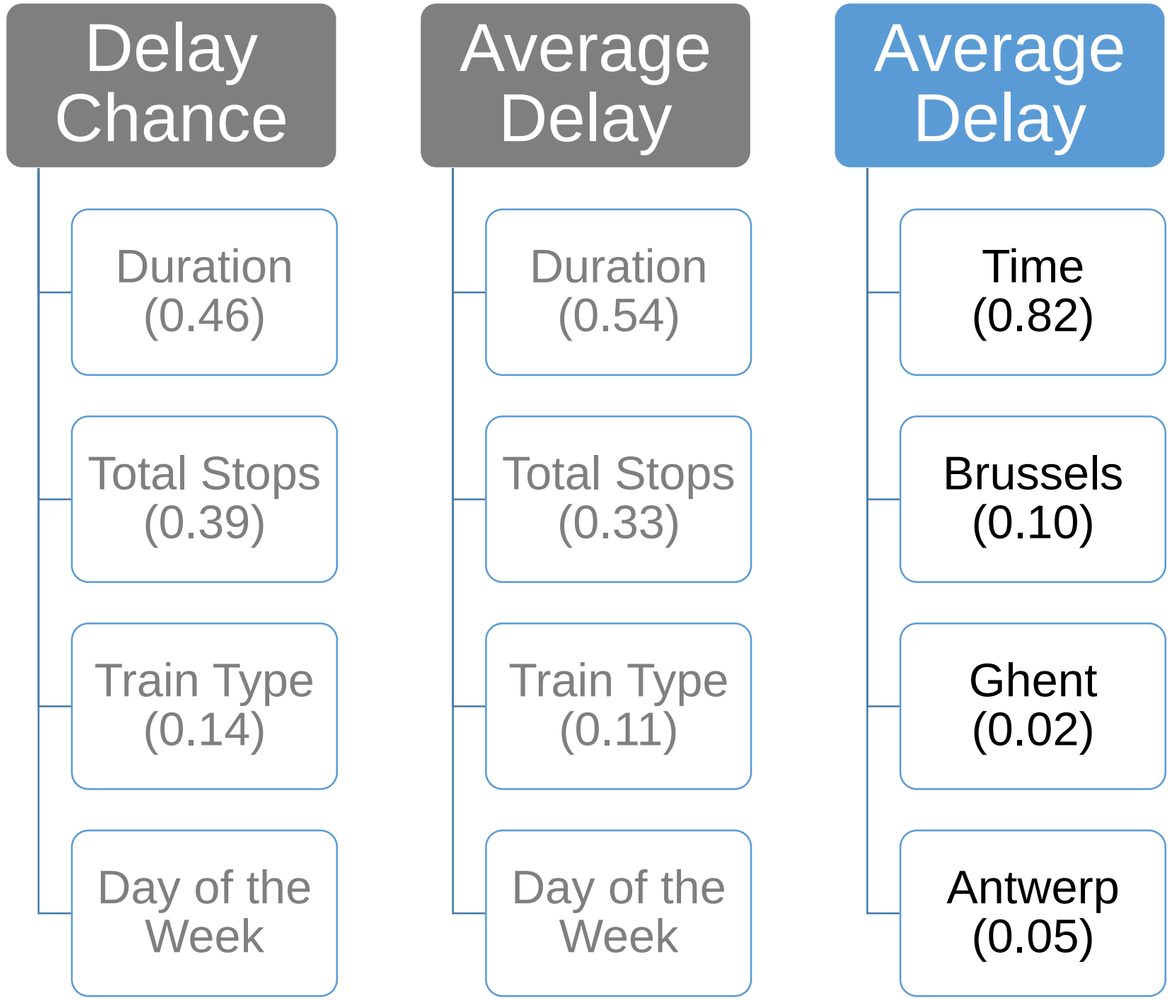
INFLUENCE OF TIME OF DAY



- Divide 24h day in categories of half an hour each
- Dummies for each train: whether it passed in Brussels, Ghent or Antwerp
- Perform random forest regression

GHENT UNIVERSITY

RANDOM FOREST



Random Forest Classification

AUC = 0.55

Accuracy = 69%

Random Forest Regression

Random Forest Regression

GHENT UNIVERSITY

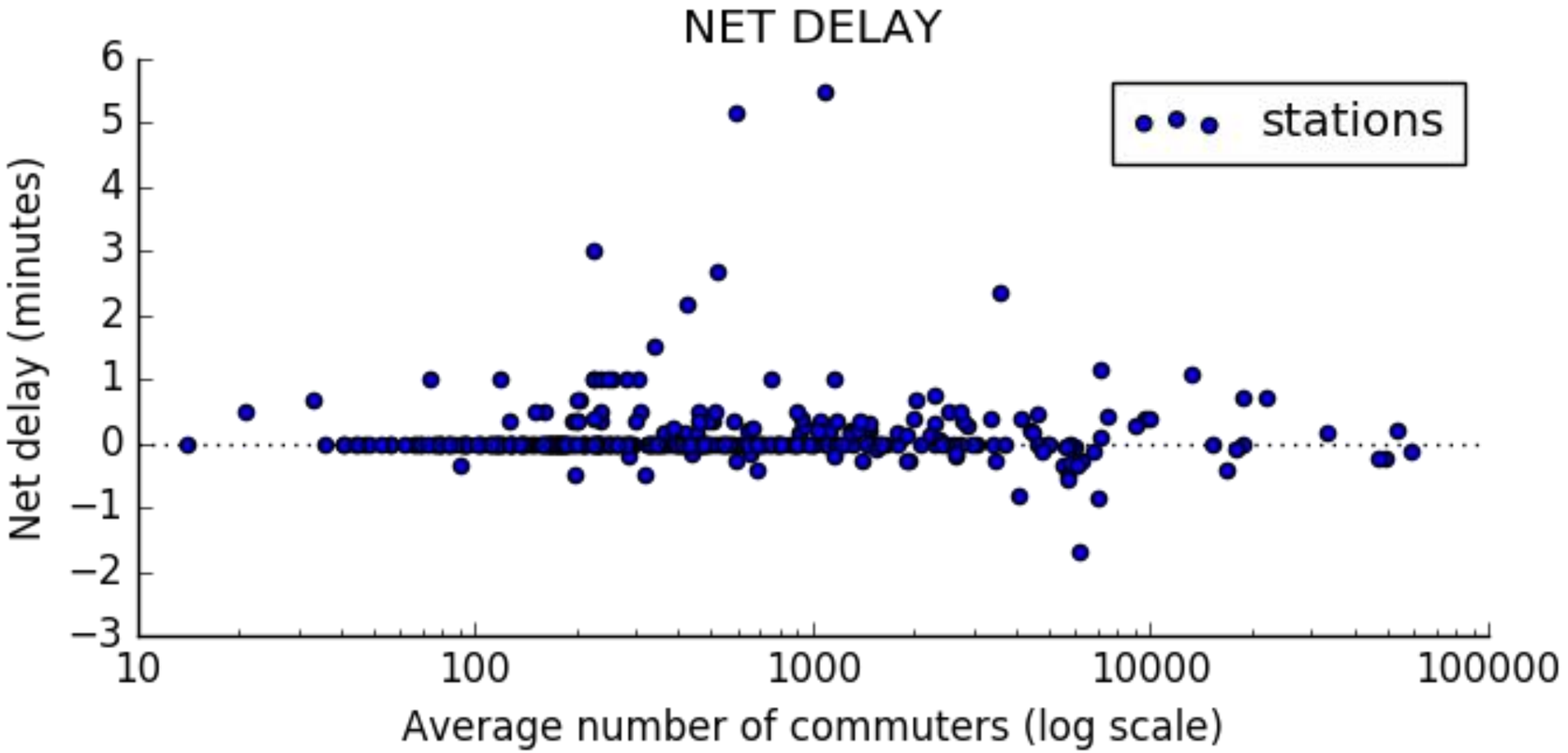
INFLUENCE OF COMMUTERS

- Are trains more easily delayed in busier stations?

**Extra dataset: average number of commuters per station on weekday*

GHENT UNIVERSITY

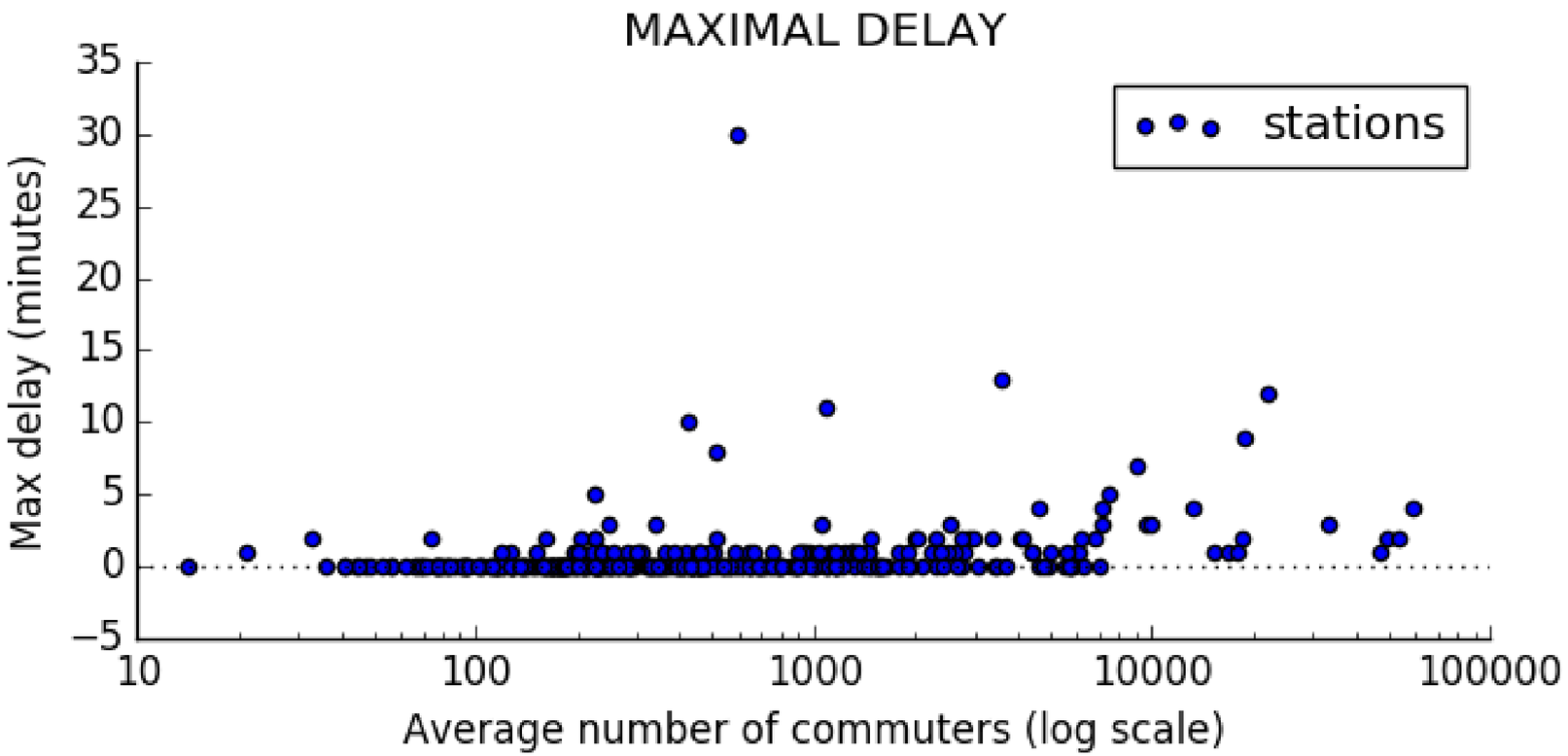
INFLUENCE OF COMMUTERS



Net delay: difference of departure delay and arrival delay

GHENT UNIVERSITY

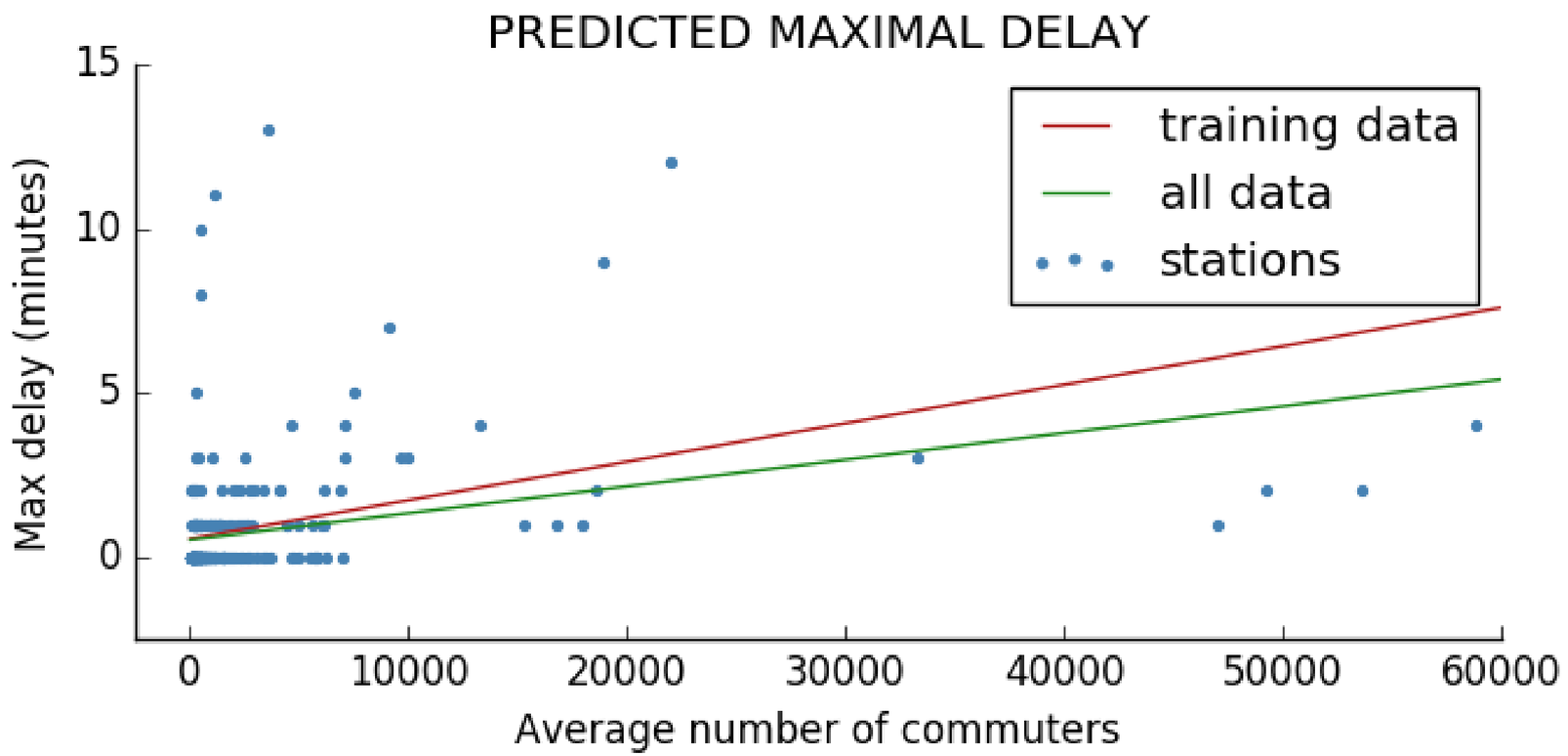
INFLUENCE OF COMMUTERS



Max delay: the maximal netto delay per station in the data set

GHENT UNIVERSITY

INFLUENCE OF COMMUTERS

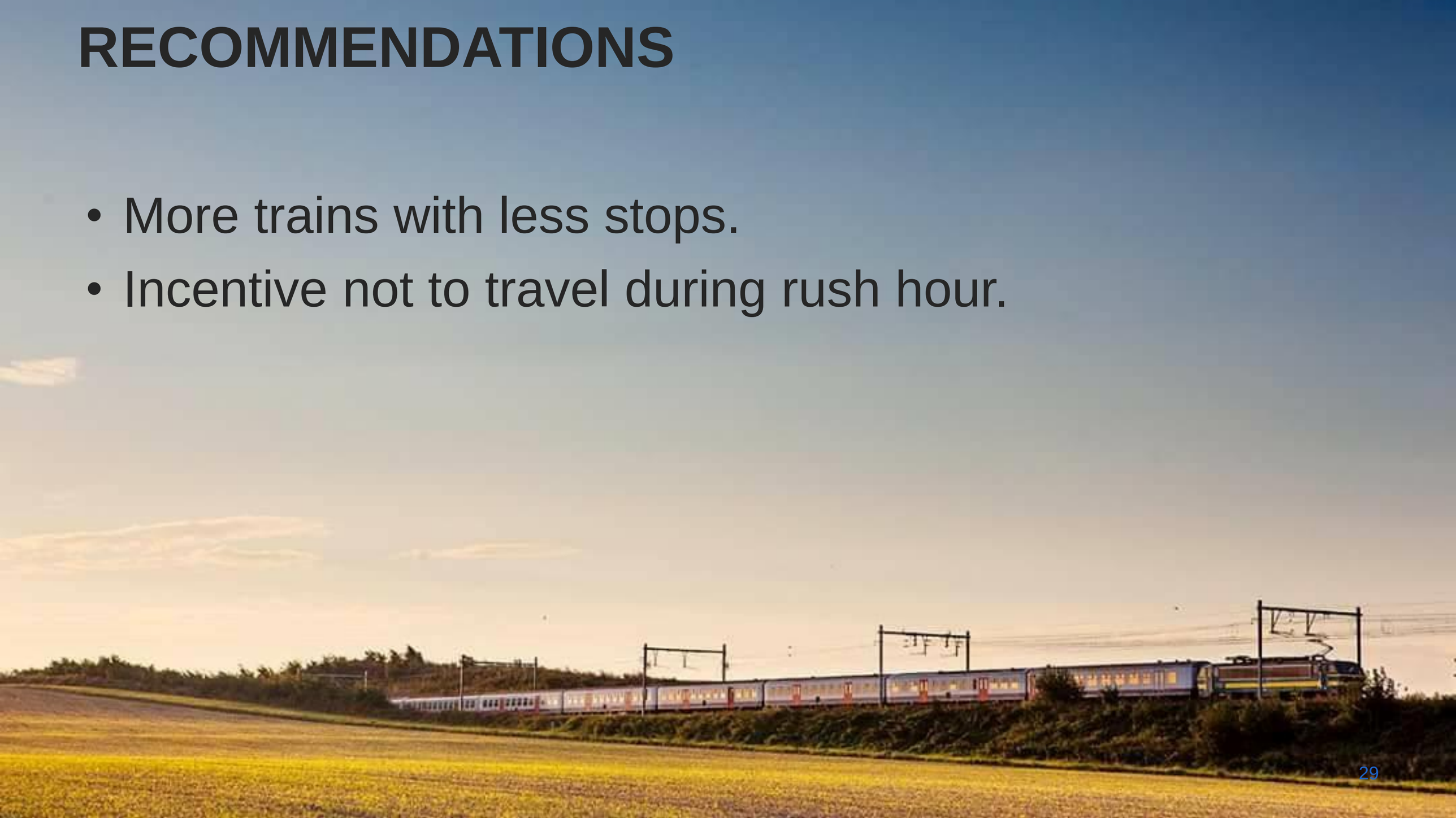


- Simple linear regression
- Slope has significant *p*-value!

GHENT UNIVERSITY

RECOMMENDATIONS

- More trains with less stops.
- Incentive not to travel during rush hour.



FACULTY OF SCIENCES

MASTER OF STATISTICAL DATA ANALYSIS
BAYRAK, CARTUYVELS, HEYVAERT, KOURFALI, O'SULLIVAN

THANK YOU

SPARK PROJECT FOR NMBS / 14.12.2017

GHENT UNIVERSITY