Restaurant Visitor Forecast

Jay Cheung

jaycheung1096@gmail.com

ACCELERATE

BRIEF SELF INTRO

2011

CUHK BSC MATHEMATICS

Double Streams in Enrichment Math
and Applied Math

2013

POSTGRADUATE DIPLOMA IN EDUCATION

Major (Math Education), HKBU

2018

GM TEACHER

EMI secondary school
HKDSE Math, M2, Physics

NOW

ACCELERATE

Data Science and Machine Learning
Cohort 2

OBJECTIVES

Four Objectives

Insight

Looking into real data of catering industry

Analysis

possible features that affect the number of visitors

Prediction

Precise prediction of restaurant visitors of all kinds

Application

Apply similar strategies in HK restaurants

Data Source

AirREGI

kaggle™

DATA

AirREGI, an POS cash register app with reserve system for cafe and restaurants

2016-2017 visit and reserve data of 829 stores located in different locations of Japan

Dataset can be obtained in Kaggle.

Technical Requirement

python jupyter

01

Programming Langauge
Environment

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

NumPy

02

Data Manipulation

plotly

seaborn

matplotlib

plotnine

03

Graphical Visualisation

scikits learn

XGBoost

04

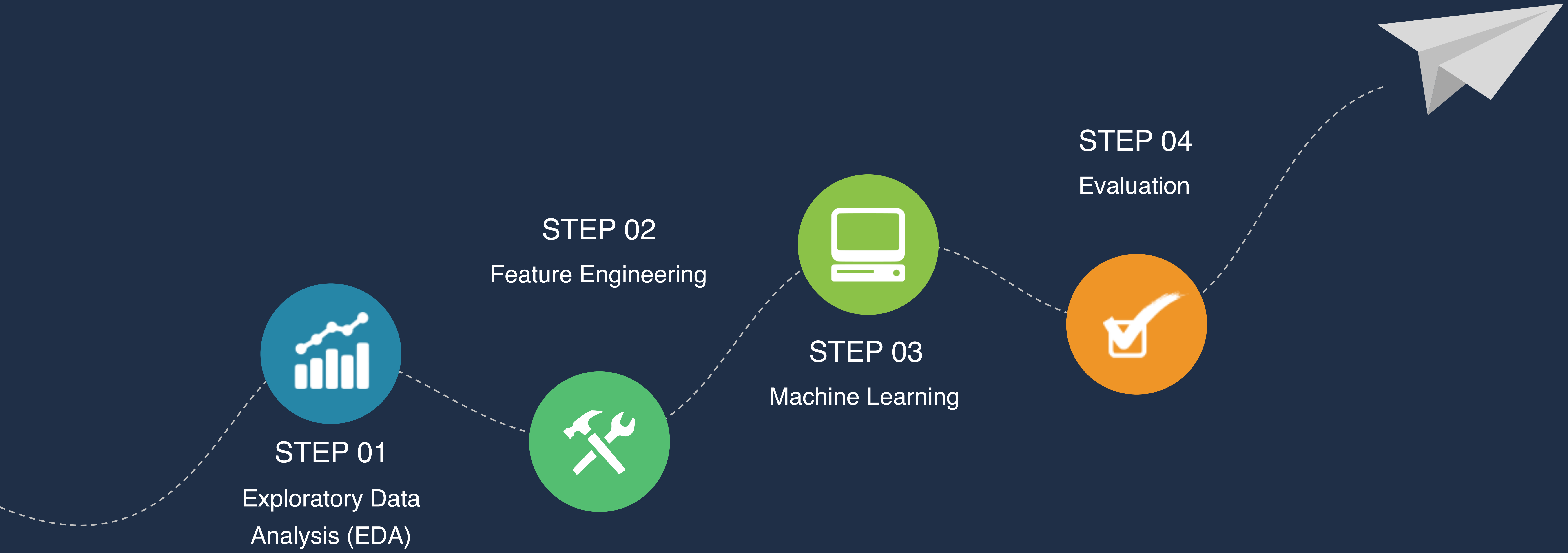
Statistics
Machine Learning

tableau

05

Software

Data Science Methodology



STEP 01

Exploratory Data Analysis (EDA)

STEP 02

Feature Engineering

STEP 03

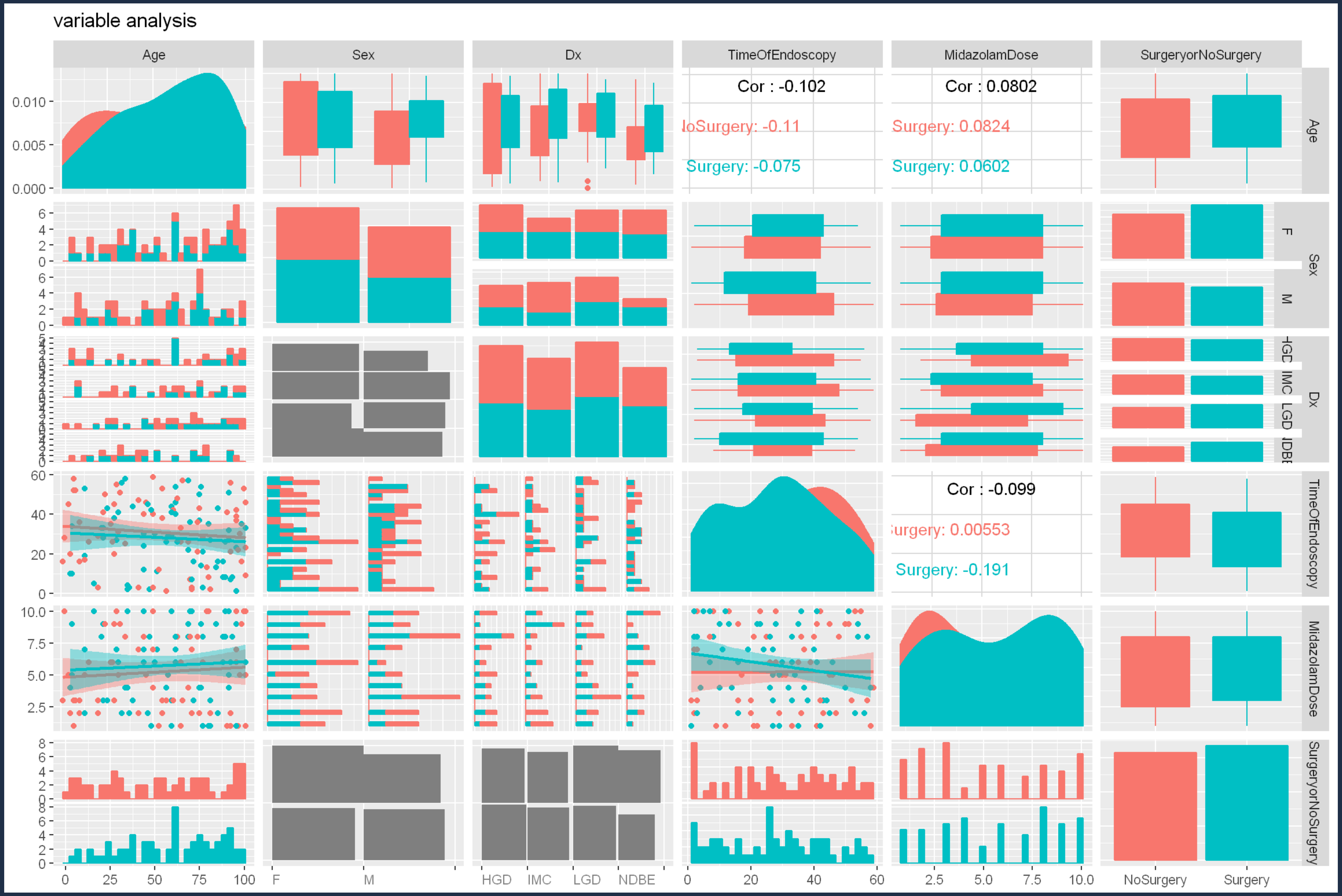
Machine Learning

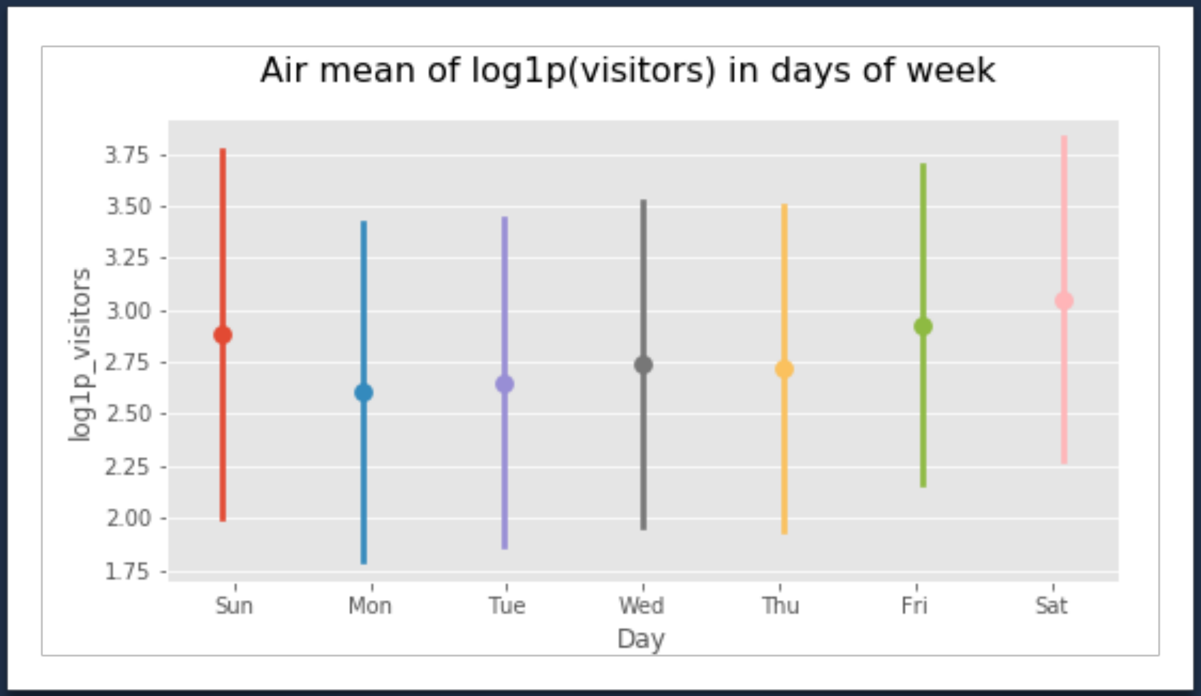
STEP 04

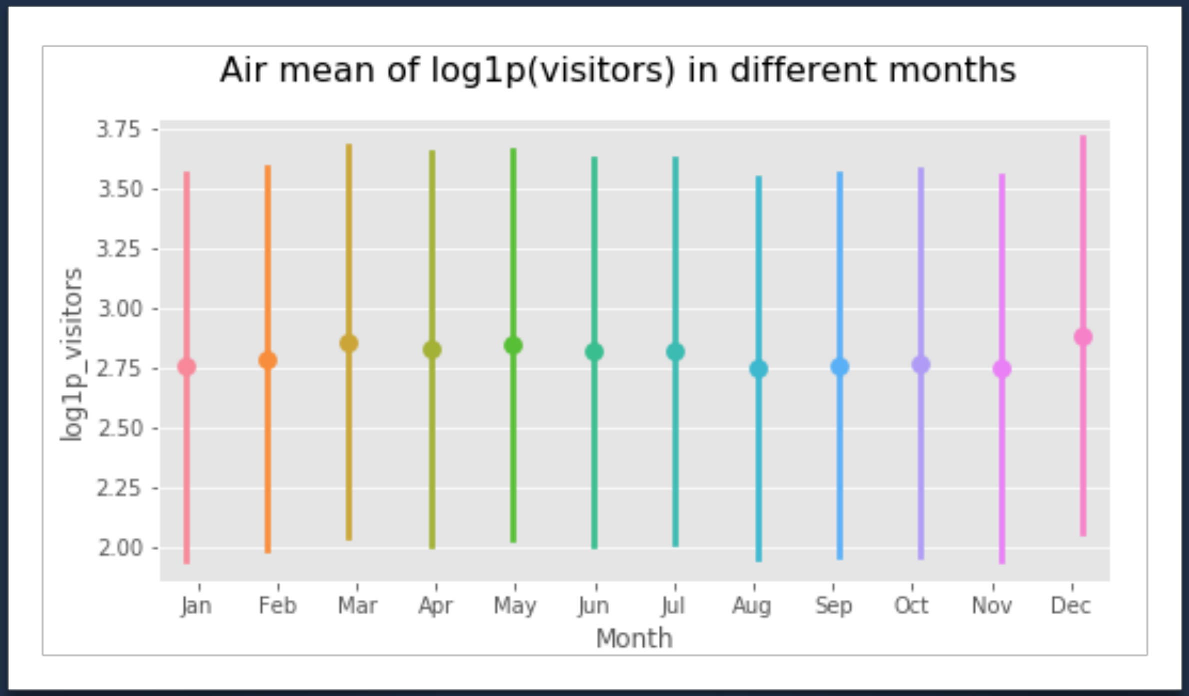
Evaluation

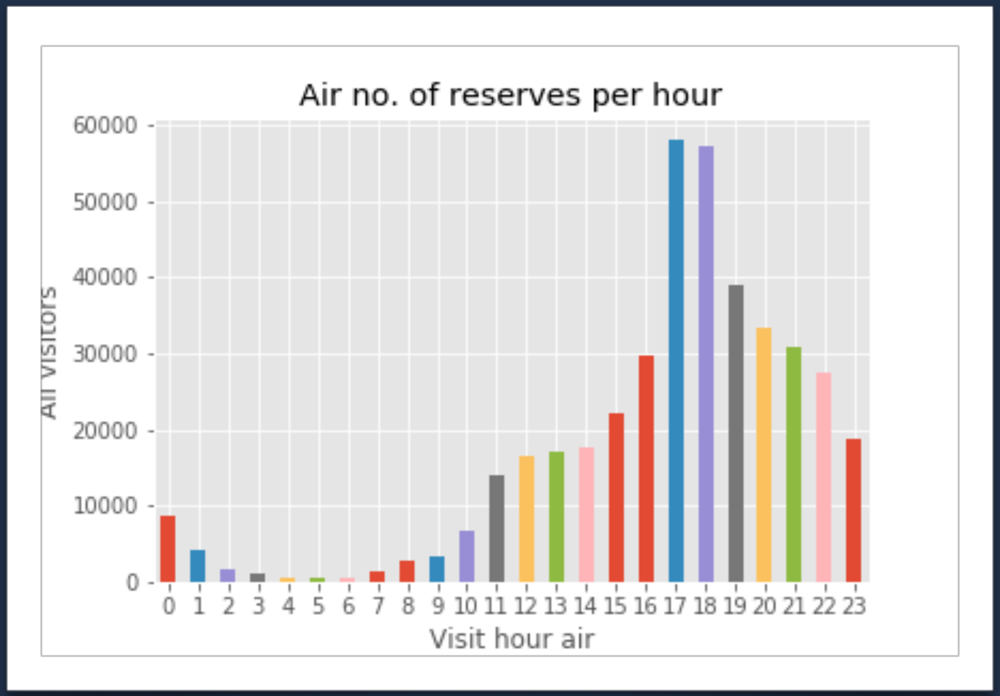
Exploratory Data Analysis

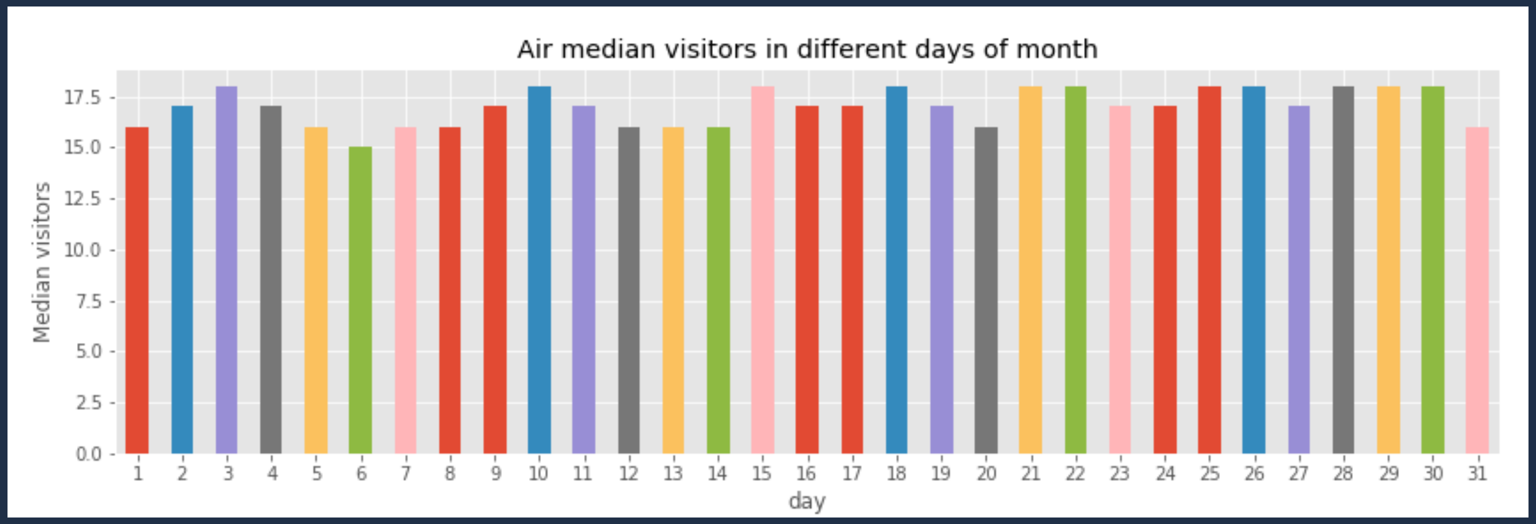
An approach to summarise the main characteristics of the dataset mainly by visualisation.

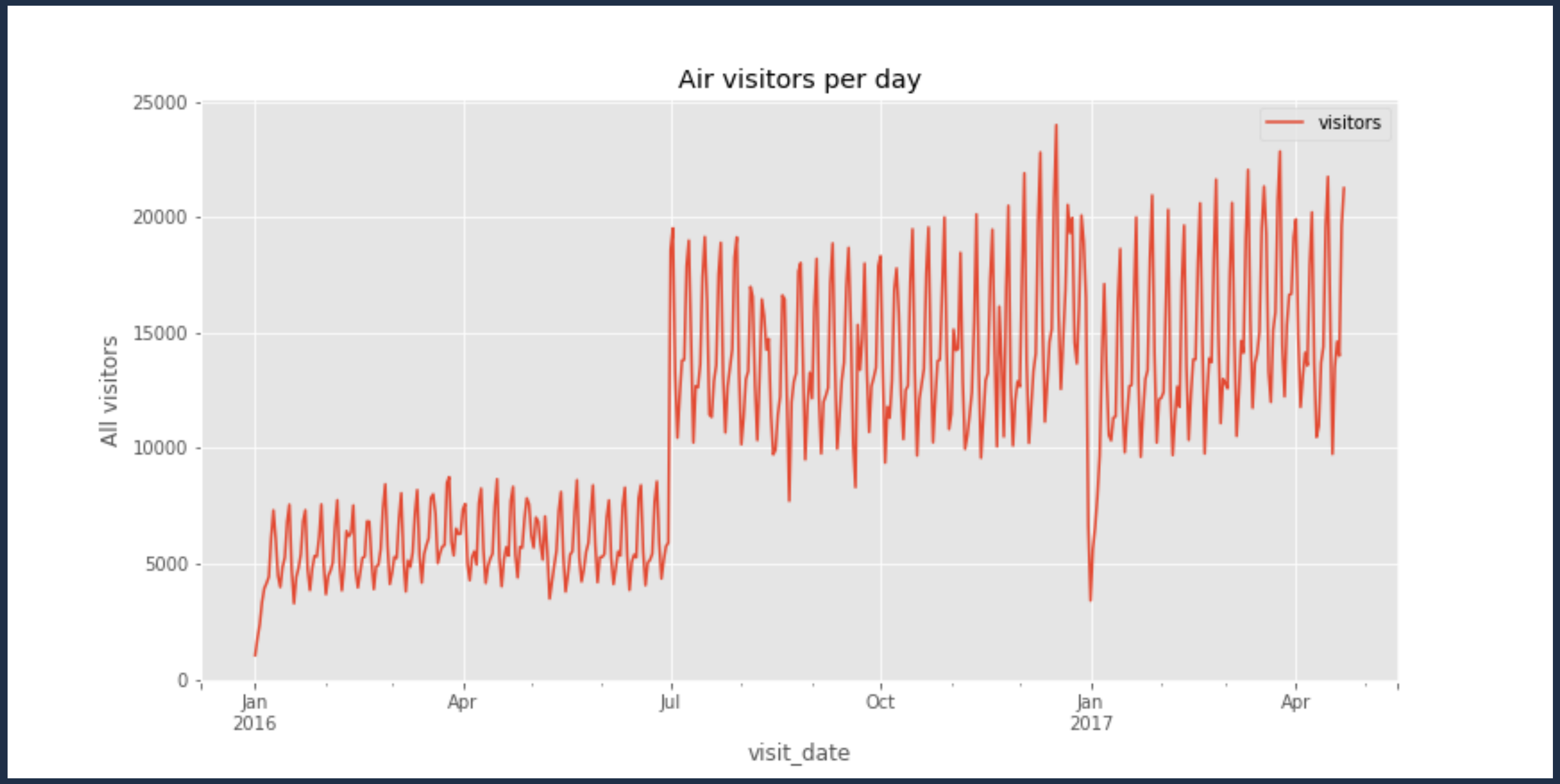


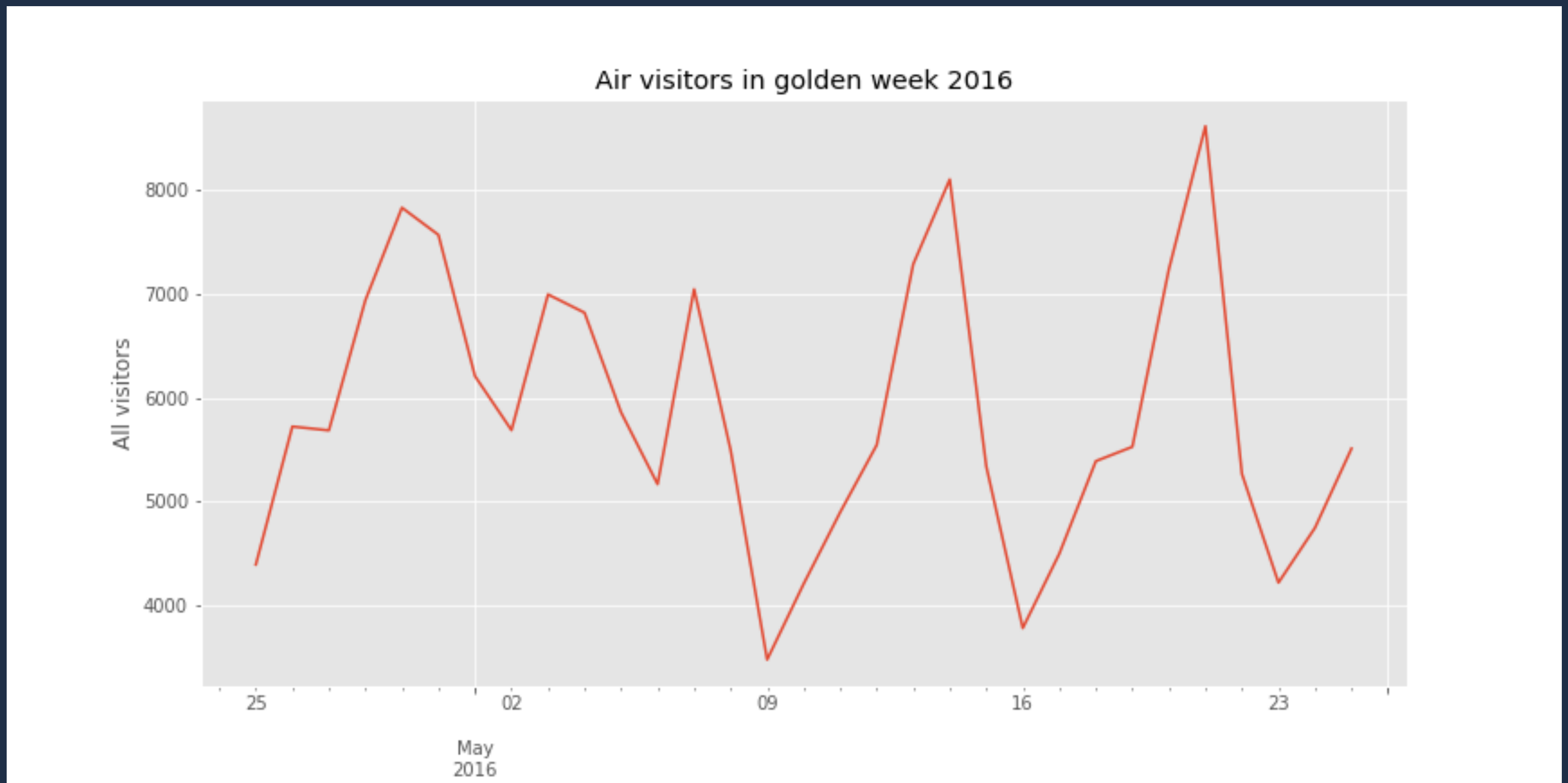

Air mean of log1p(visitors) in days of week
log1p_visitors
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Day

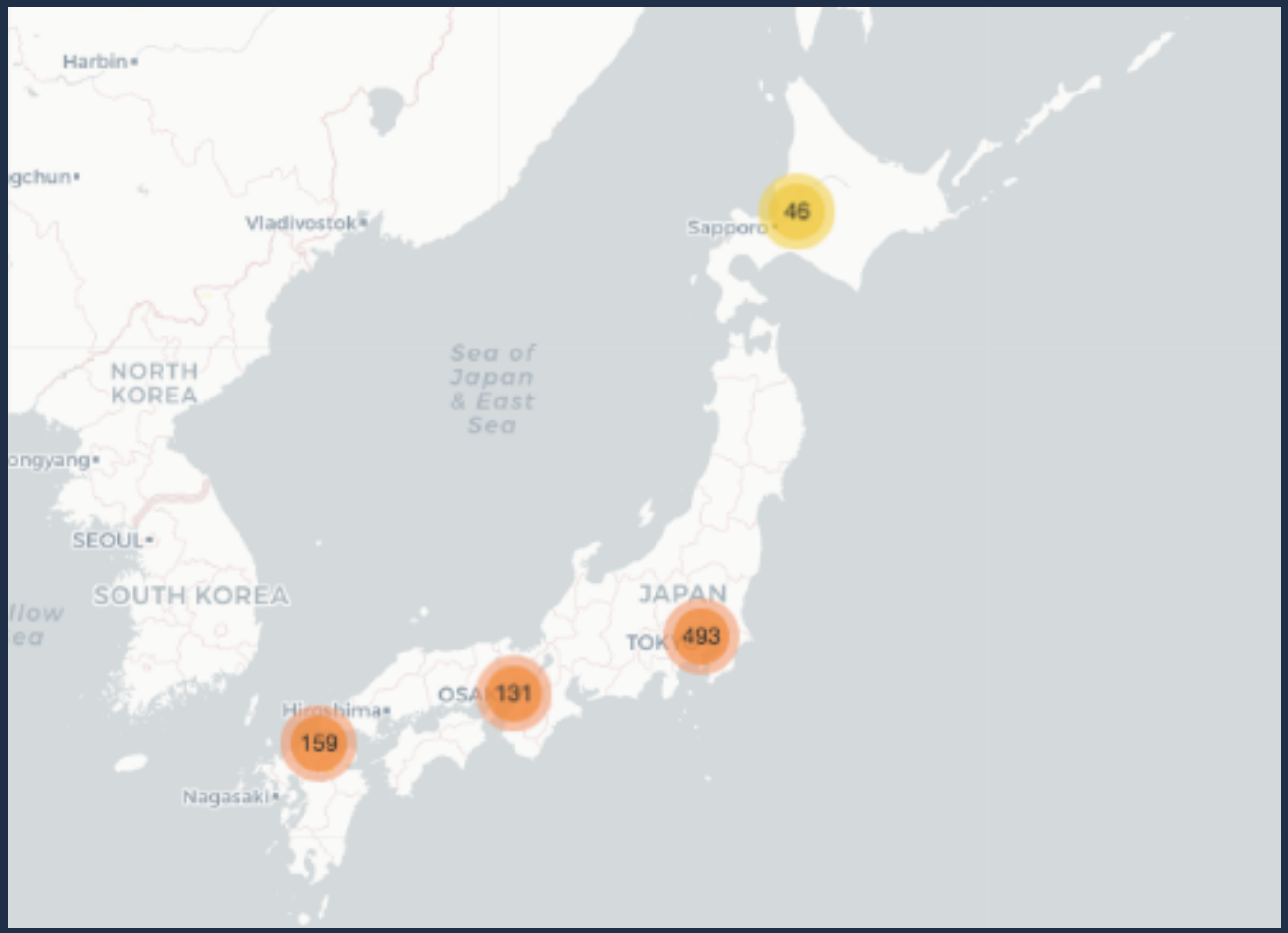

Air mean of log1p(visitors) in different months
log1p_visitors
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

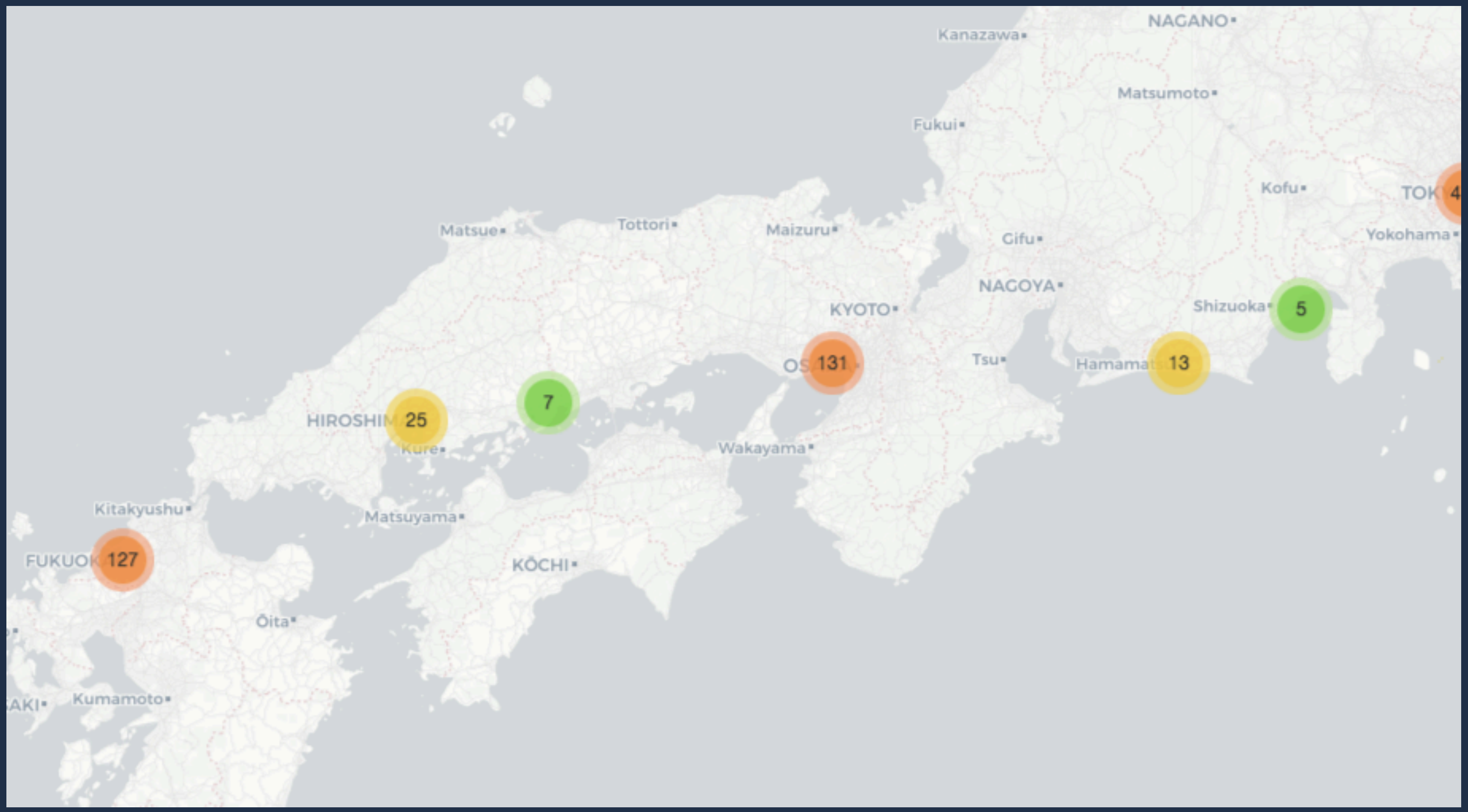

Air no. of reserves per hour
All visitors
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Visit hour air

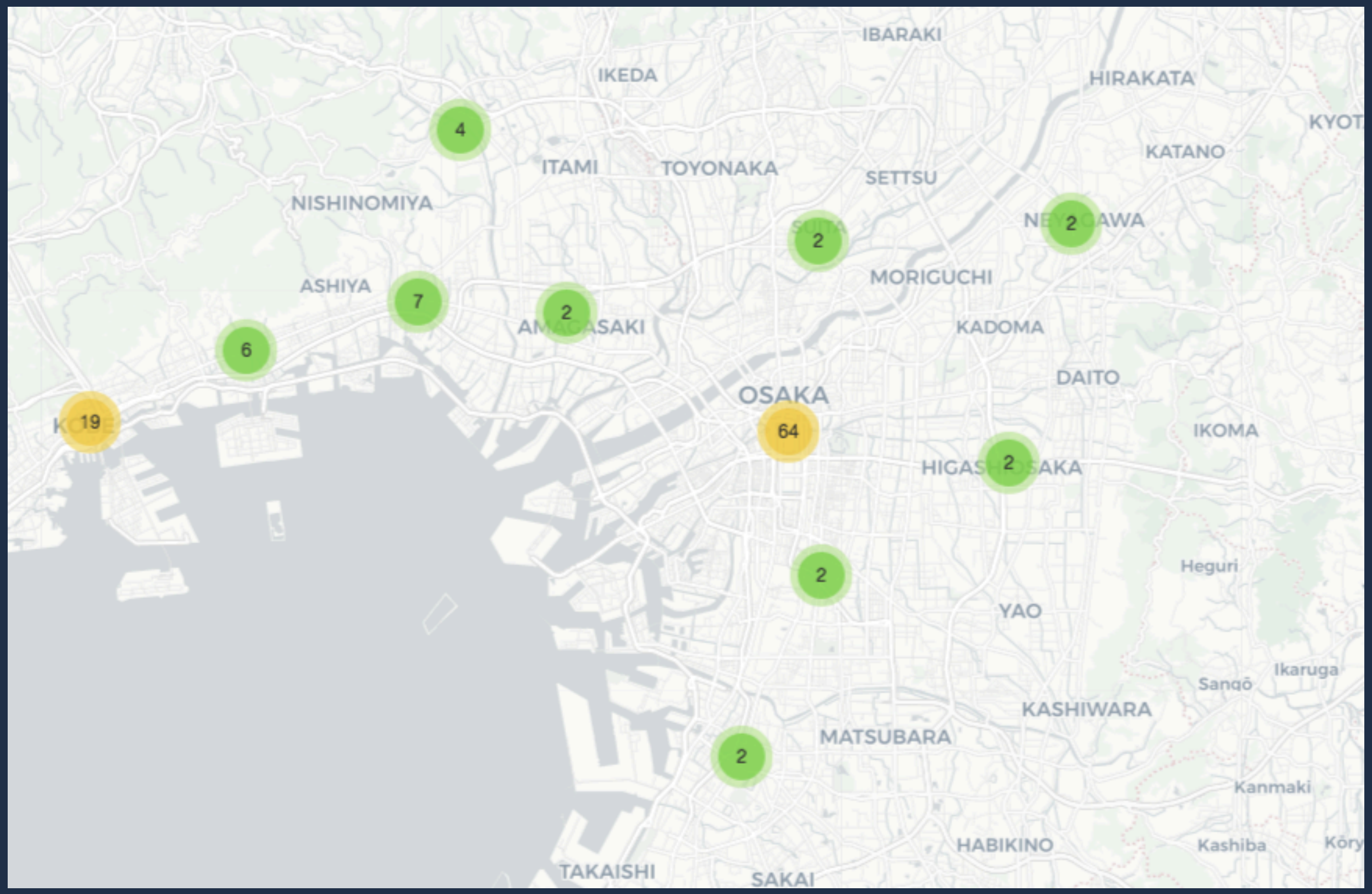

Air median visitors in different days of month
Median visitors
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
day

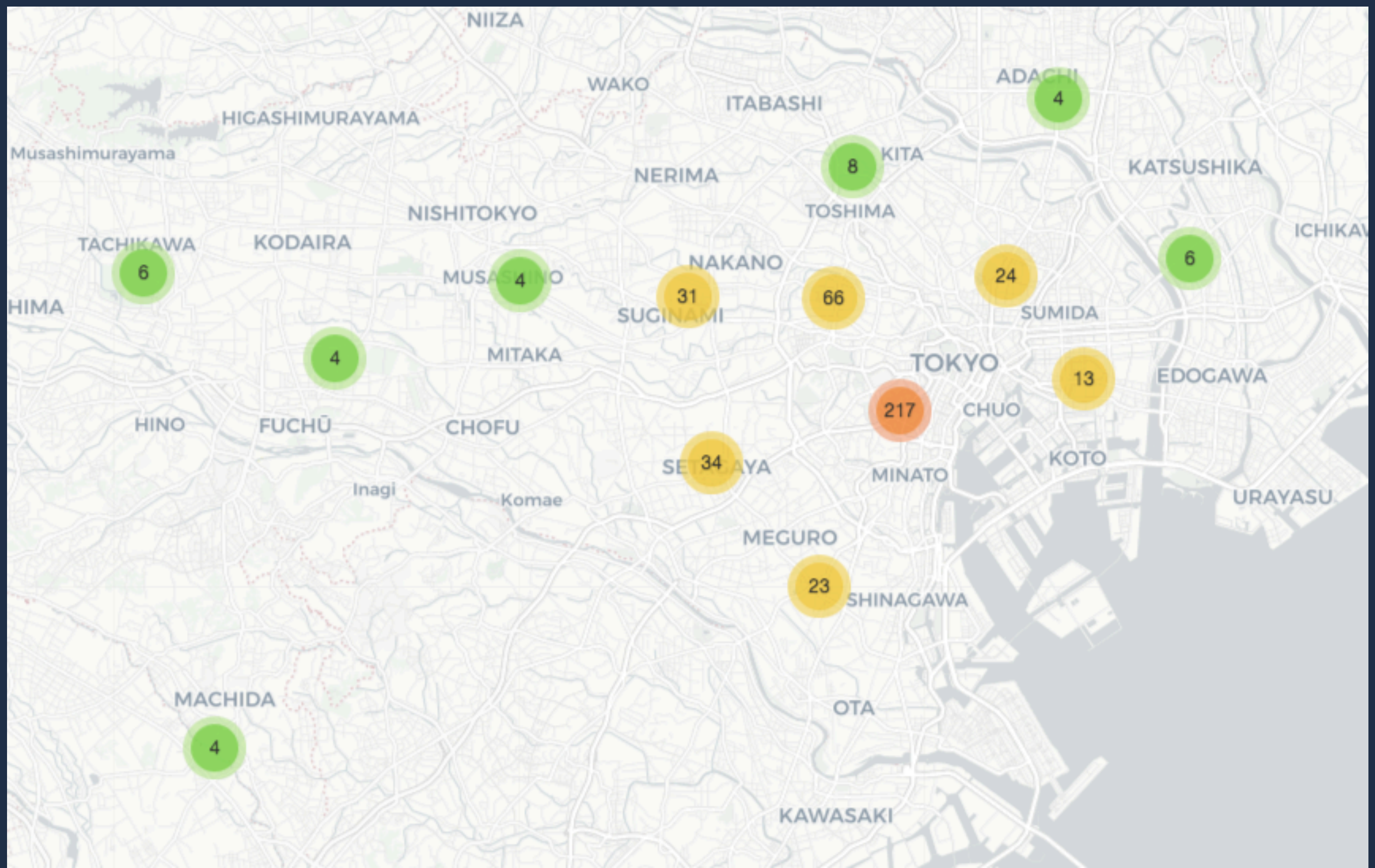

Air visitors per day
visitors
All visitors
25000
20000
15000
10000
5000
0
Jan
2016
Apr
Jul
Oct
Jan
2017
Apr
visit_date


Air visitors in golden week 2016
All visitors
8000
7000
6000
5000
4000
25
02
09
16
23
May
2016


Harbin
Vladivostok
NORTH KOREA
Sea of Japan & East Sea
SEOUL
SOUTH KOREA
Sapporo
46
JAPAN
493
131
159
Nagasaki

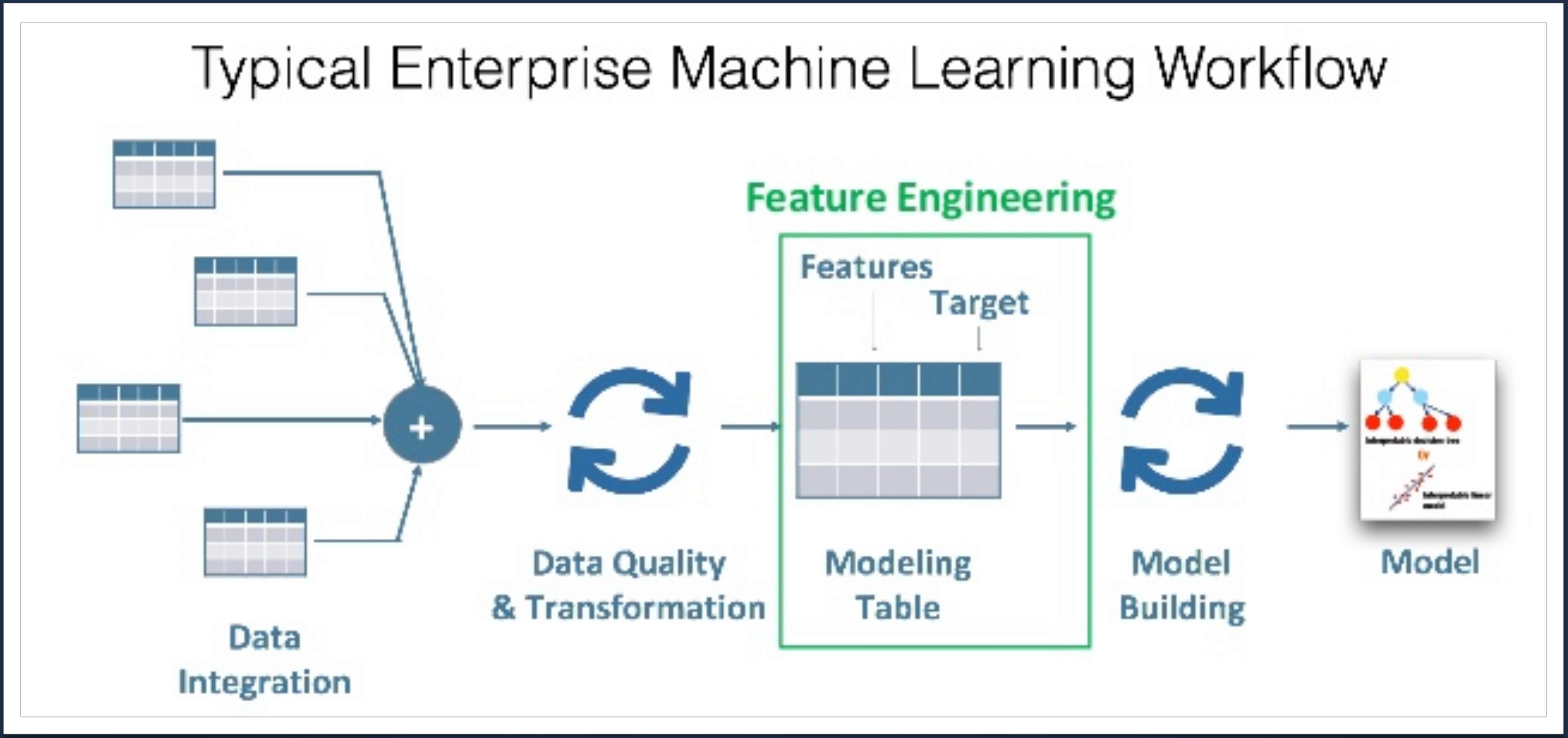

Kanazawa
NAGANO
Matsumoto
Fukui
Kofu
Matsue
Tottori
Maizuru
Gifu
Yokohama
KYOTO
NAGOYA
Shizuoka
5
131
Tsu
Hamamatsu
13
HIROSHIMA
25
7
Kure
Wakayama
Kitakyushu
Matsuyama
FUKUOKA
127
KOCHI
Oita
Kumamoto

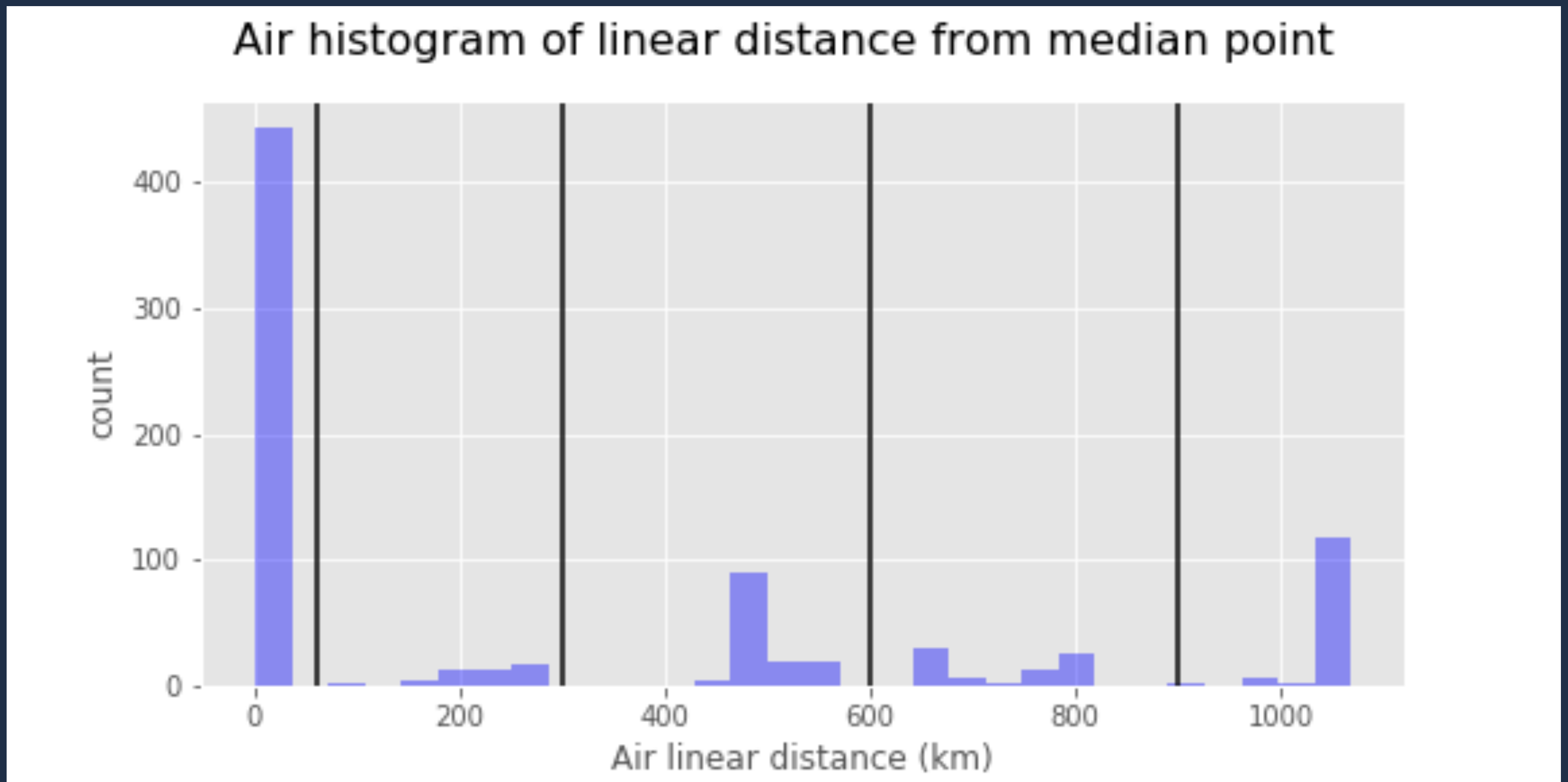

IBARAKI
IKEDA
HIRAKATA
KATANO
4
ITAMI
TOYONAKA
SETTSU
NISHINOMIYA
2
2
MORIGUCHI
ASHIYA
7
2
KADOMA
6
DAITO
19
OSAKA
64
IKOMA
2
2
YAO
Heguri
2
KASHIWARA
MATSUBARA
HABIKINO
TAKAISHI
SAKAI

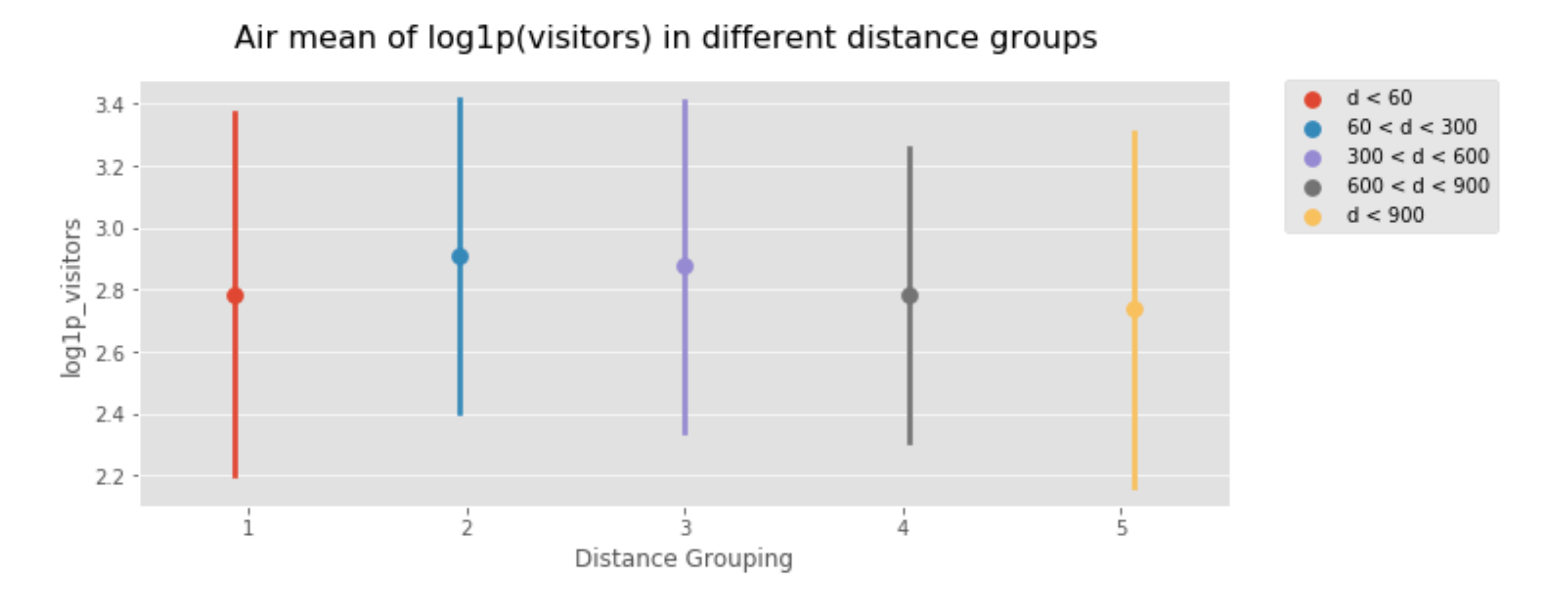

NIIZA
WAKO
ITABASHI
4
HIGASHIMURAYAMA
Musashimurayama
NERIMA
8
KITA
KATSUSHIKA
TOSHIMA
NISHITOKYO
TACHIKAWA
KODAIRA
6
NAKANO
6
4
31
66
24
SUMIDA
4
MITAKA
TOKYO
13
EDOGAWA
HINO
FUCHU
CHOFU
217
CHUO
Inagi
Komae
34
MINATO
KOTO
URAYASU
MEGURO
23
SHINAGAWA
MACHIDA
4
OTA
KAWASAKI

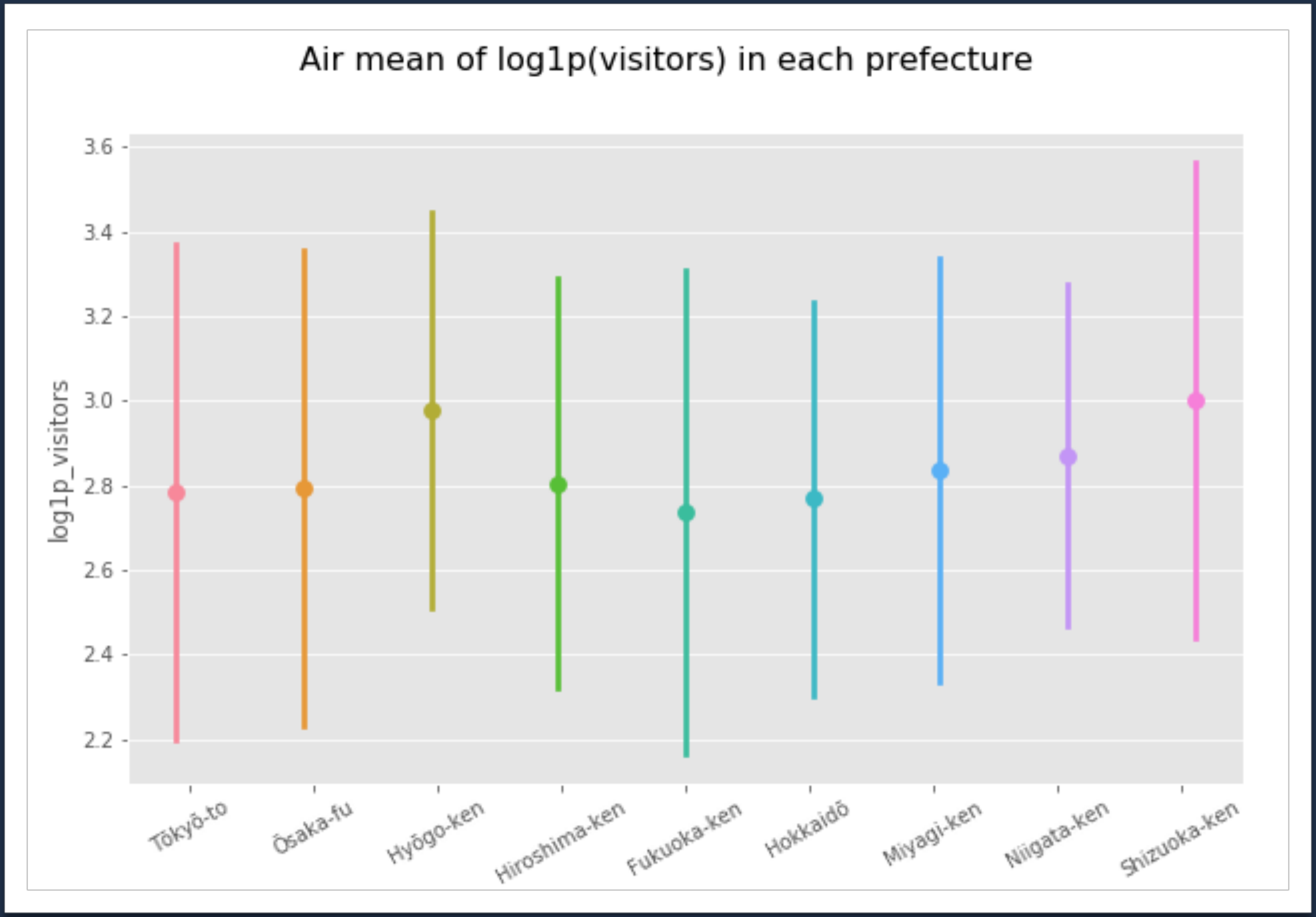
Feature Engineering

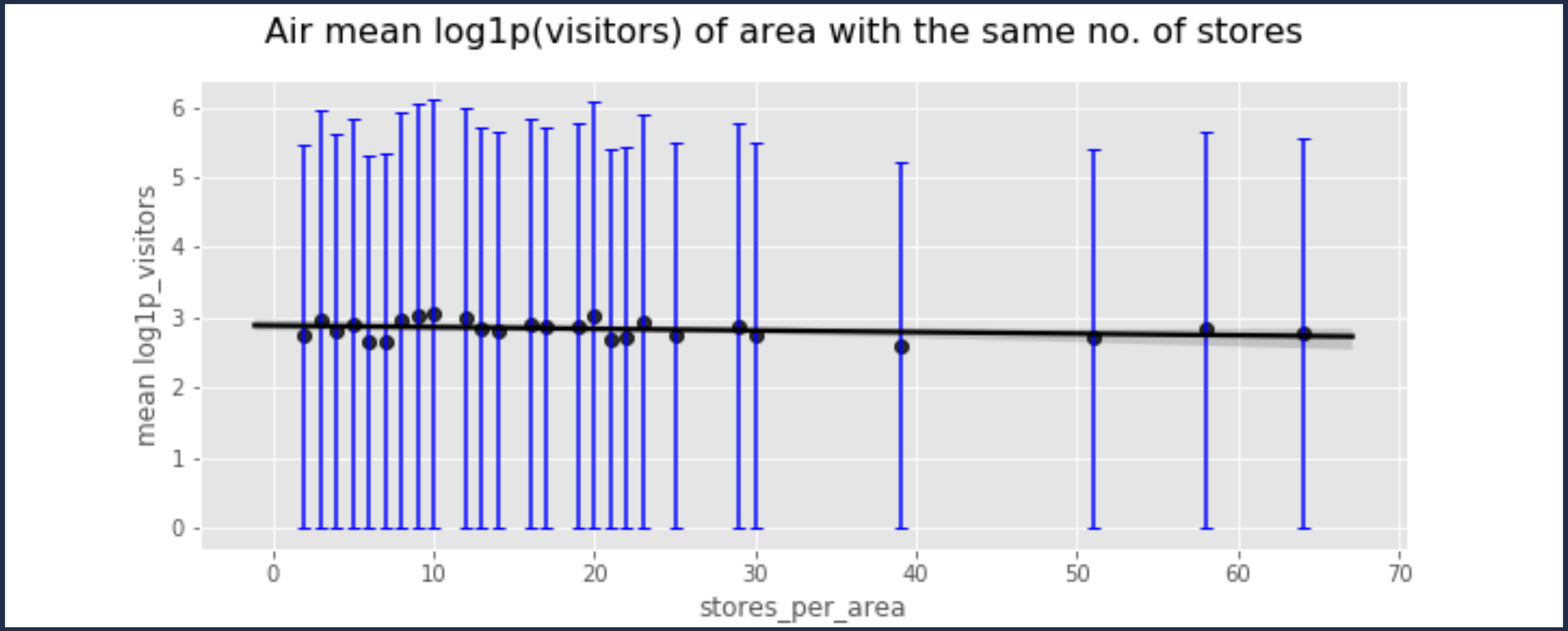
Create new features from the original ones



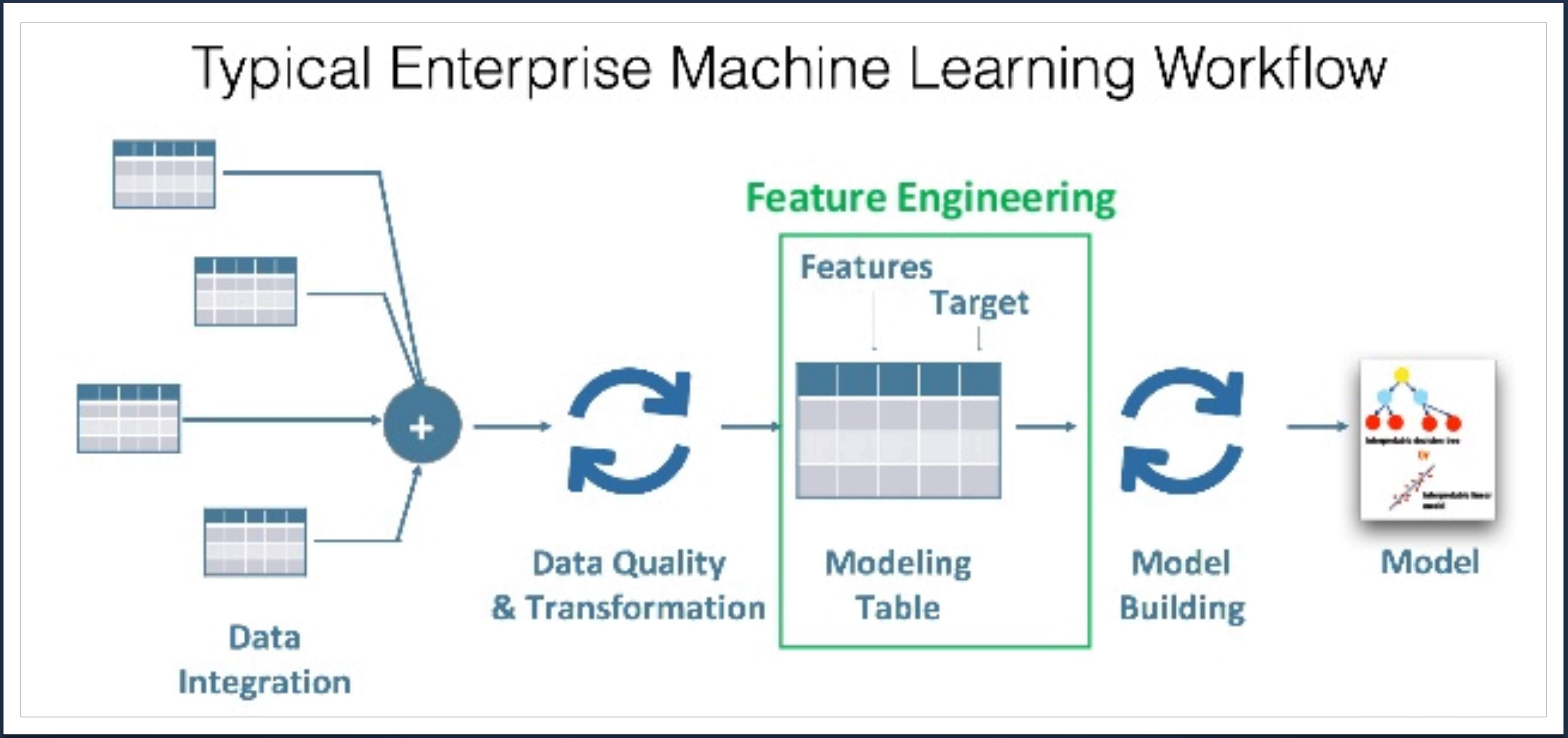

Air histogram of linear distance from median point
count
400
300
200
100
0
0
200
400
600
800
1000
Air linear distance (km)


Air mean of log1p(visitors) in different distance groups
log1p_visitors
3.4
3.2
3.0
2.8
2.6
2.4
2.2
1
2
3
4
5
Distance Grouping
d < 60
60 < d < 300
300 < d < 600
600 < d < 900
d < 900


Air mean of log1p(visitors) in each prefecture
log1p_visitors
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
Tōkyō-to
Ōsaka-fu
Hyōgo-ken
Hiroshima-ken
Fukuoka-ken
Hokkaidō
Miyagi-ken
Niigata-ken
Shizuoka-ken


Air mean log1p(visitors) of area with the same no. of stores
mean log1p_visitors
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
stores_per_area

Modeling

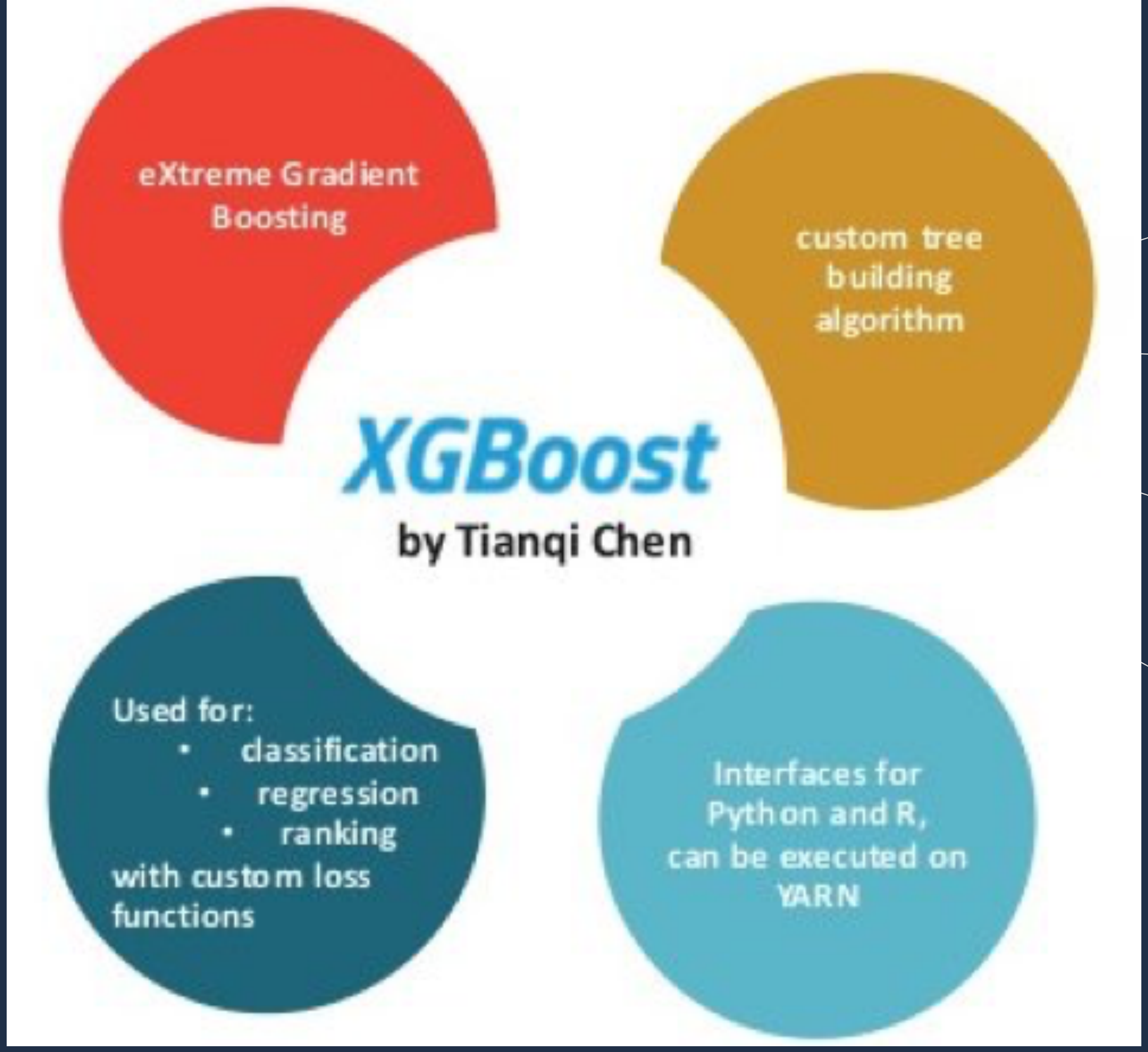


Typical Enterprise Machine Learning Workflow

Facebook’s Prophet ML Package

- **Open-source** forecasting model by facebook research team
- Numerous applications across **Facebook**
- Based on **additive model** with non-linear trends
- Capture **seasonalities** of data plus **holiday** effects
- Handles **missing data** and **outliers** automatically

XGBoost ML Algorithm


eXtreme Gradient Boosting
custom tree building algorithm
XGBoost
by Tianqi Chen
Used for:
• classification
• regression
• ranking
with custom loss functions
Interfaces for Python and R, can be executed on YARN

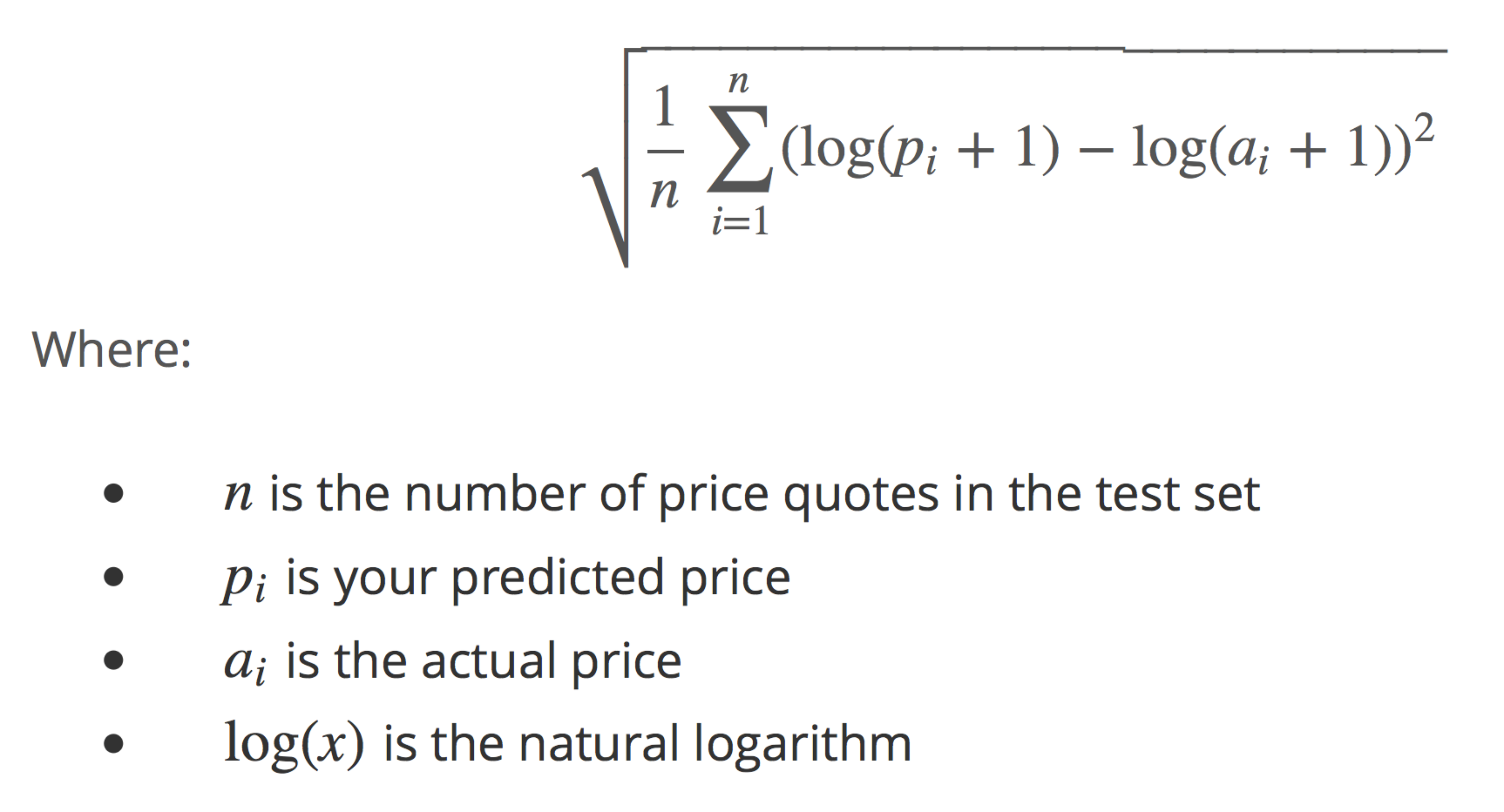
eXtreme Gradient Boosting

Open-source gradient boosted **decision tree algorithm**

Speedy and high performance

Dominates **Kaggle** competitions

RMSLE (Root Mean Squared Logarithmic Error)



$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

Where:

- n is the number of price quotes in the test set
- p_i is your predicted price
- a_i is the actual price
- $\log(x)$ is the natural logarithm

The **lower** the better

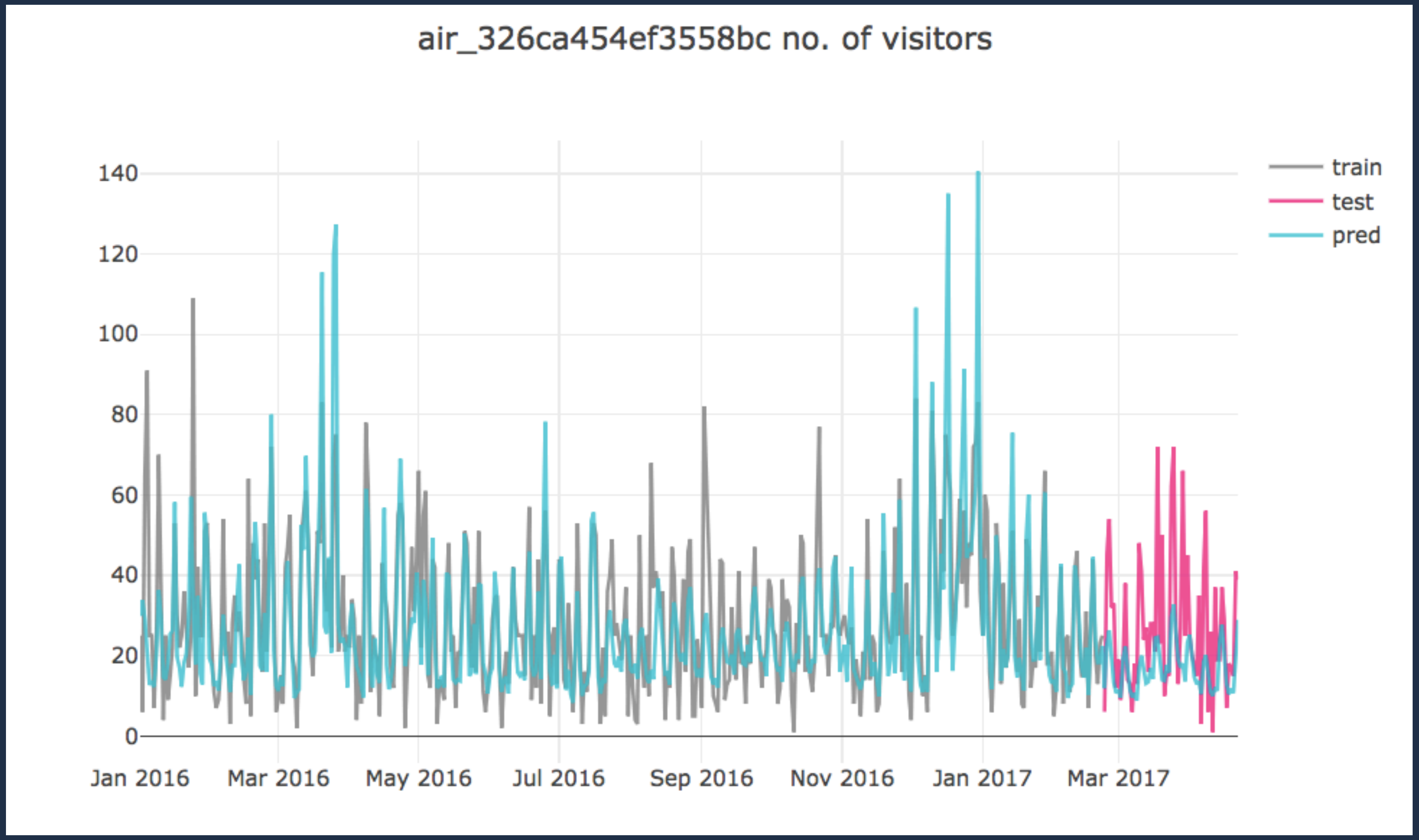
The **difference** between the predicted values and the actual values

Penalises an **under-predicted** estimate more

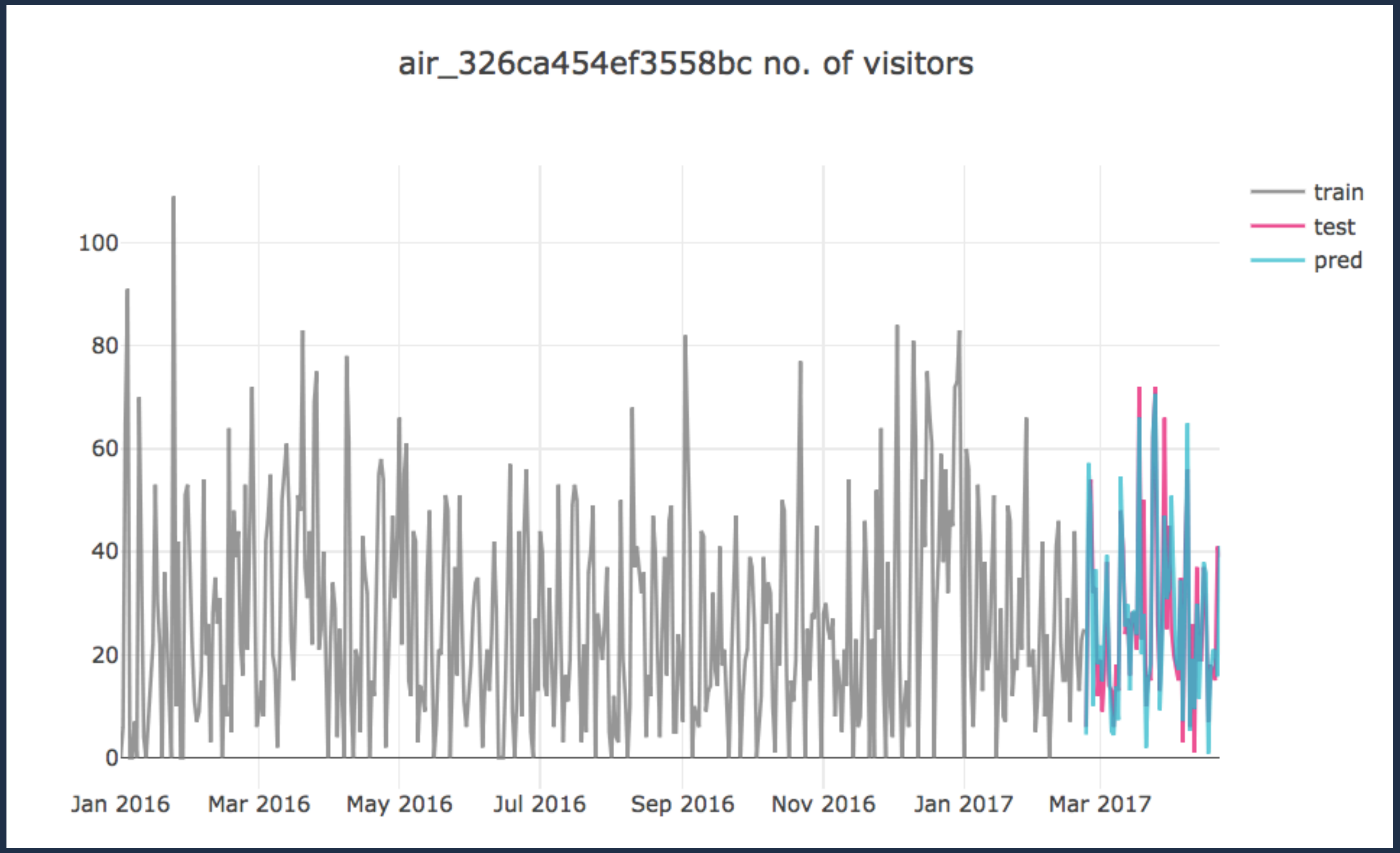
ETL process comparison between Python and Tableau

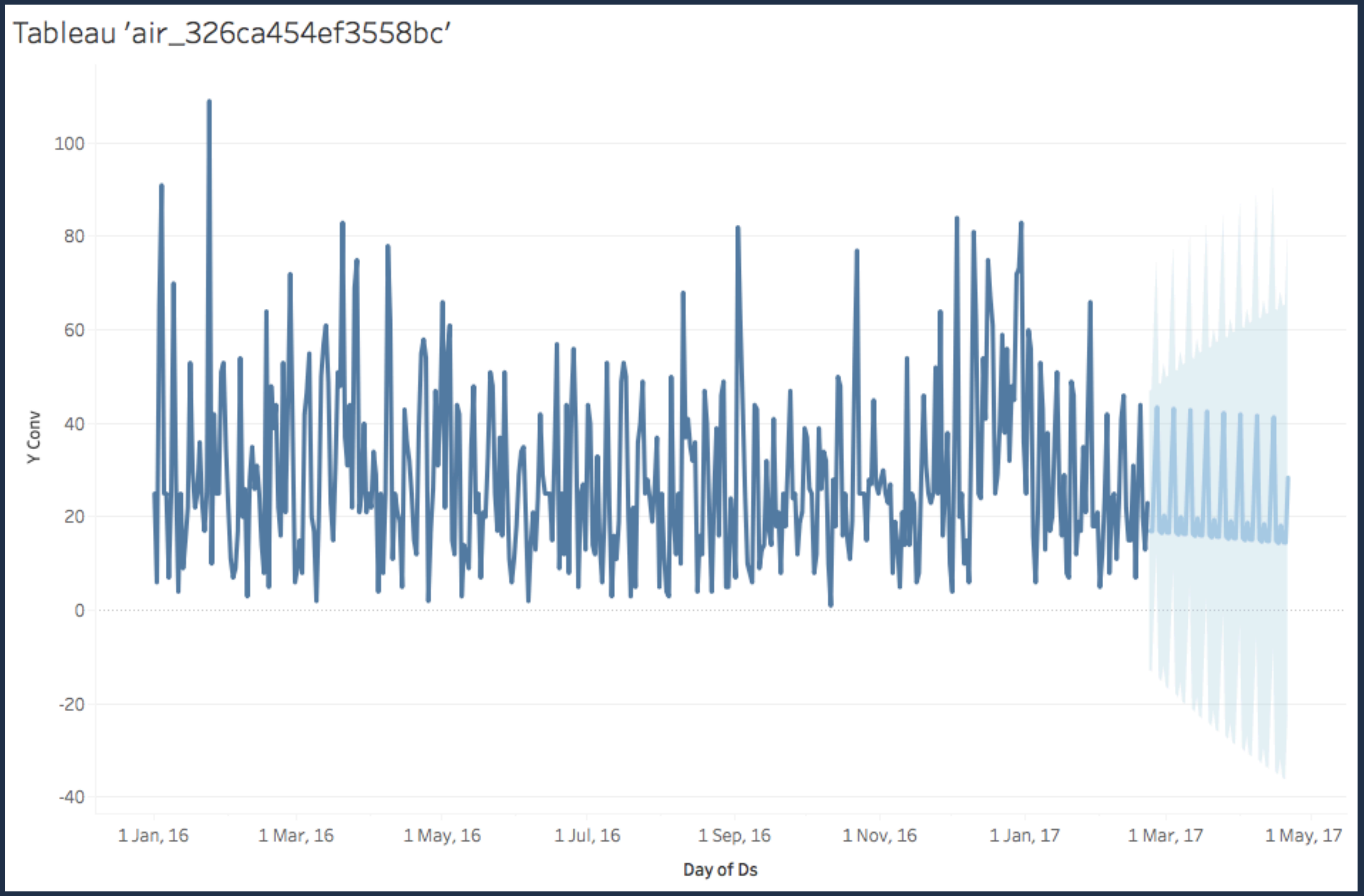
A popular data visualising software for business intelligence

Prophet

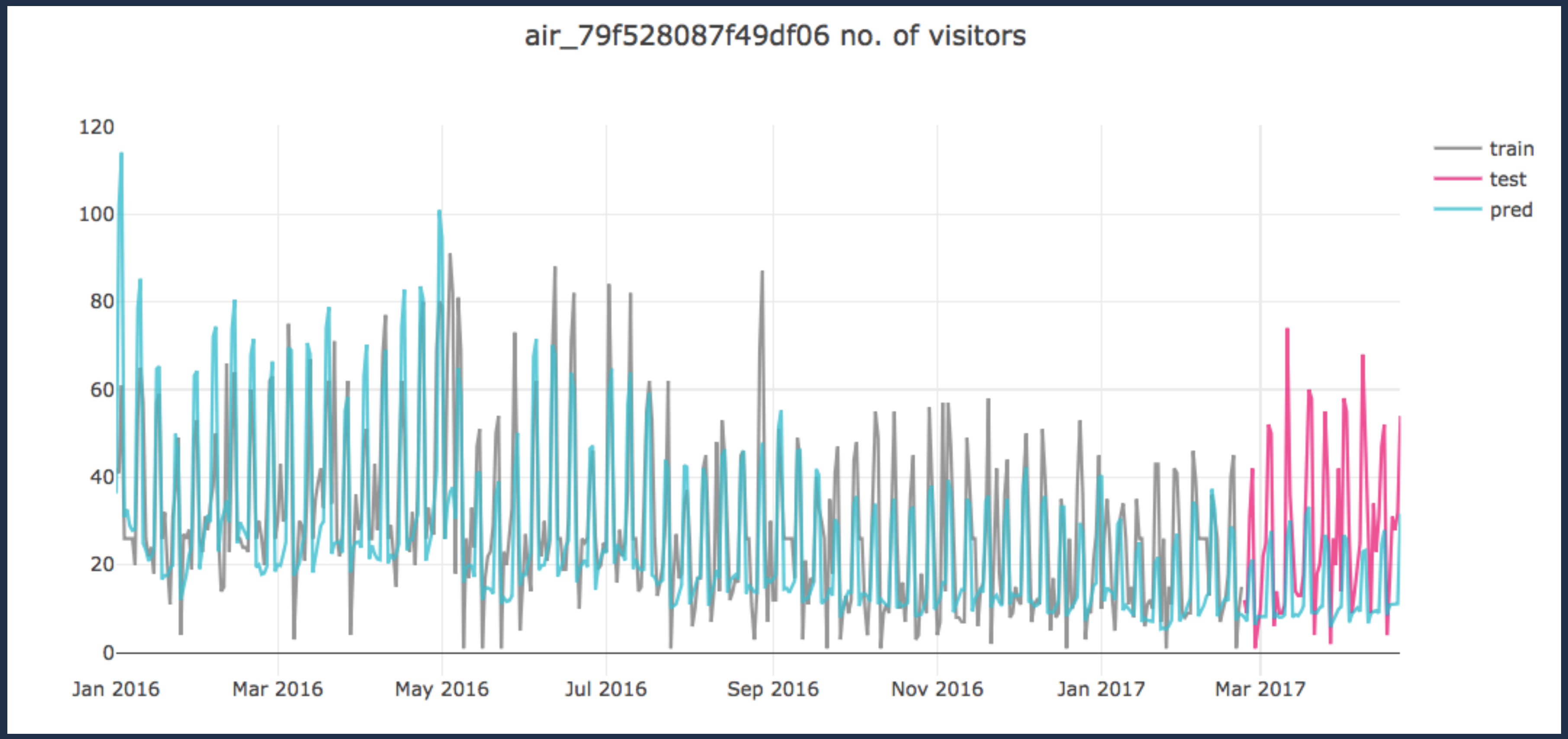


XGBoost

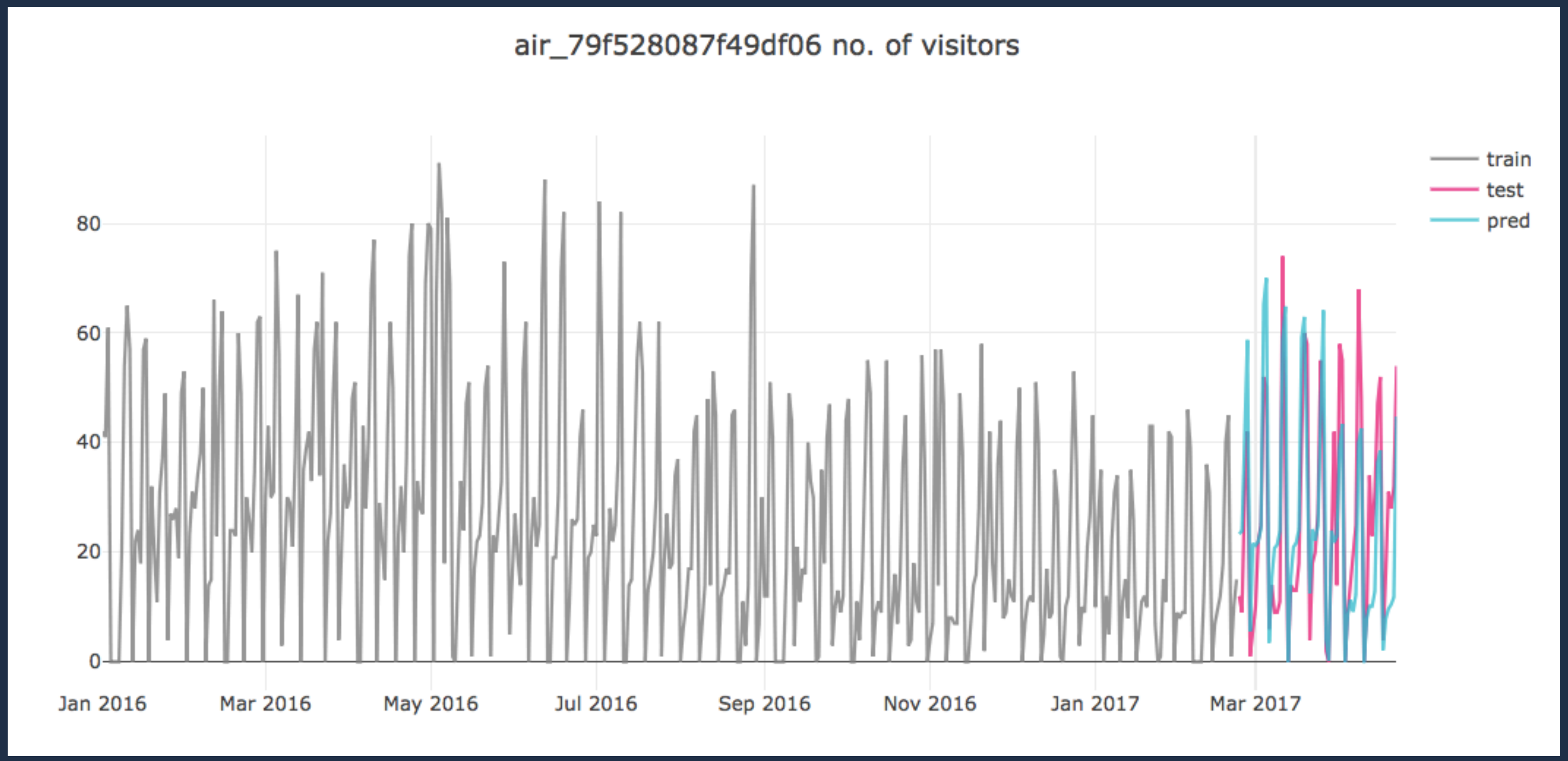


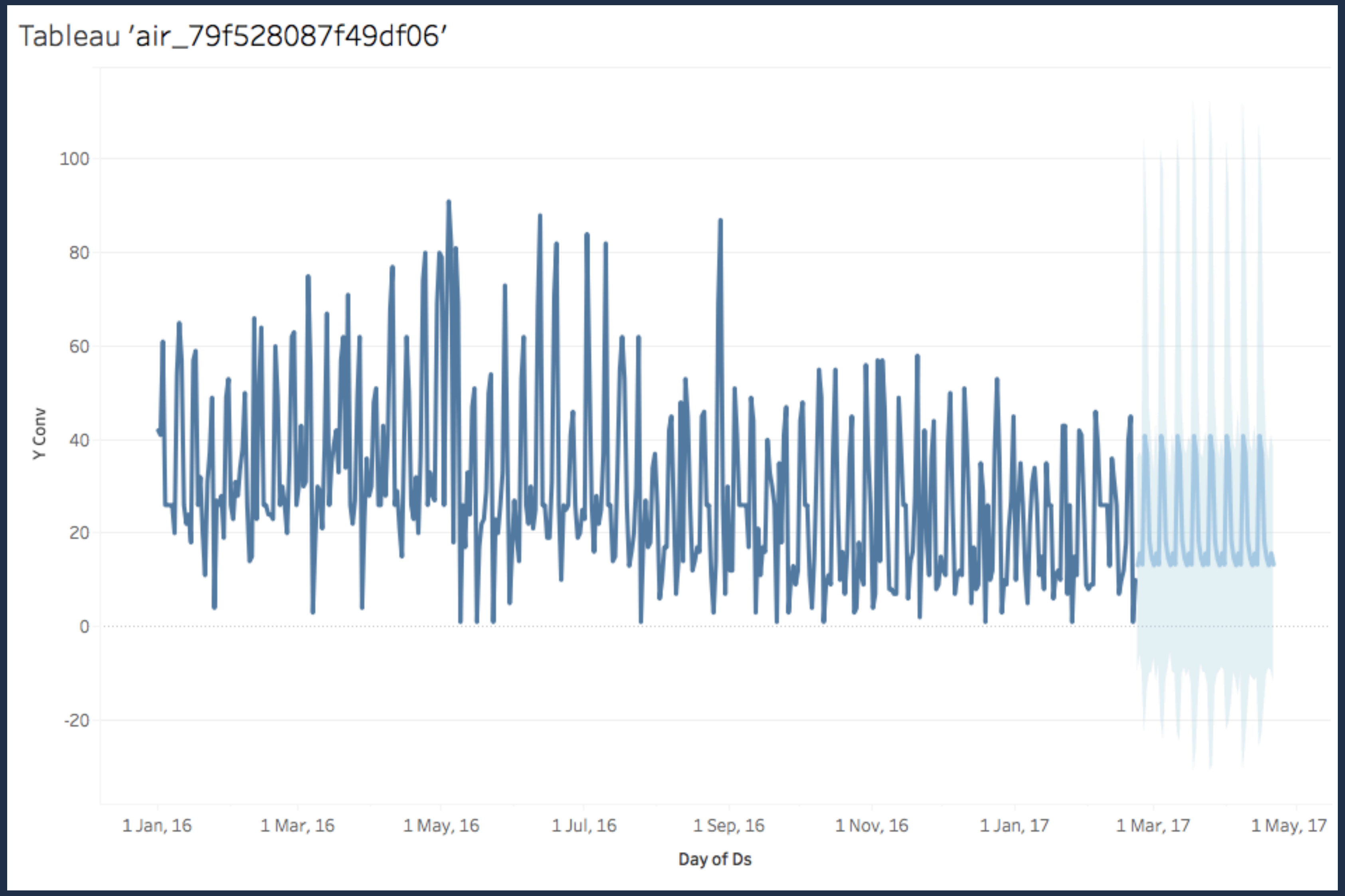


Prophet



XGBoost





thank you

To review the slide

Visit: https://https://github.com/jaycheung1096