multicampus

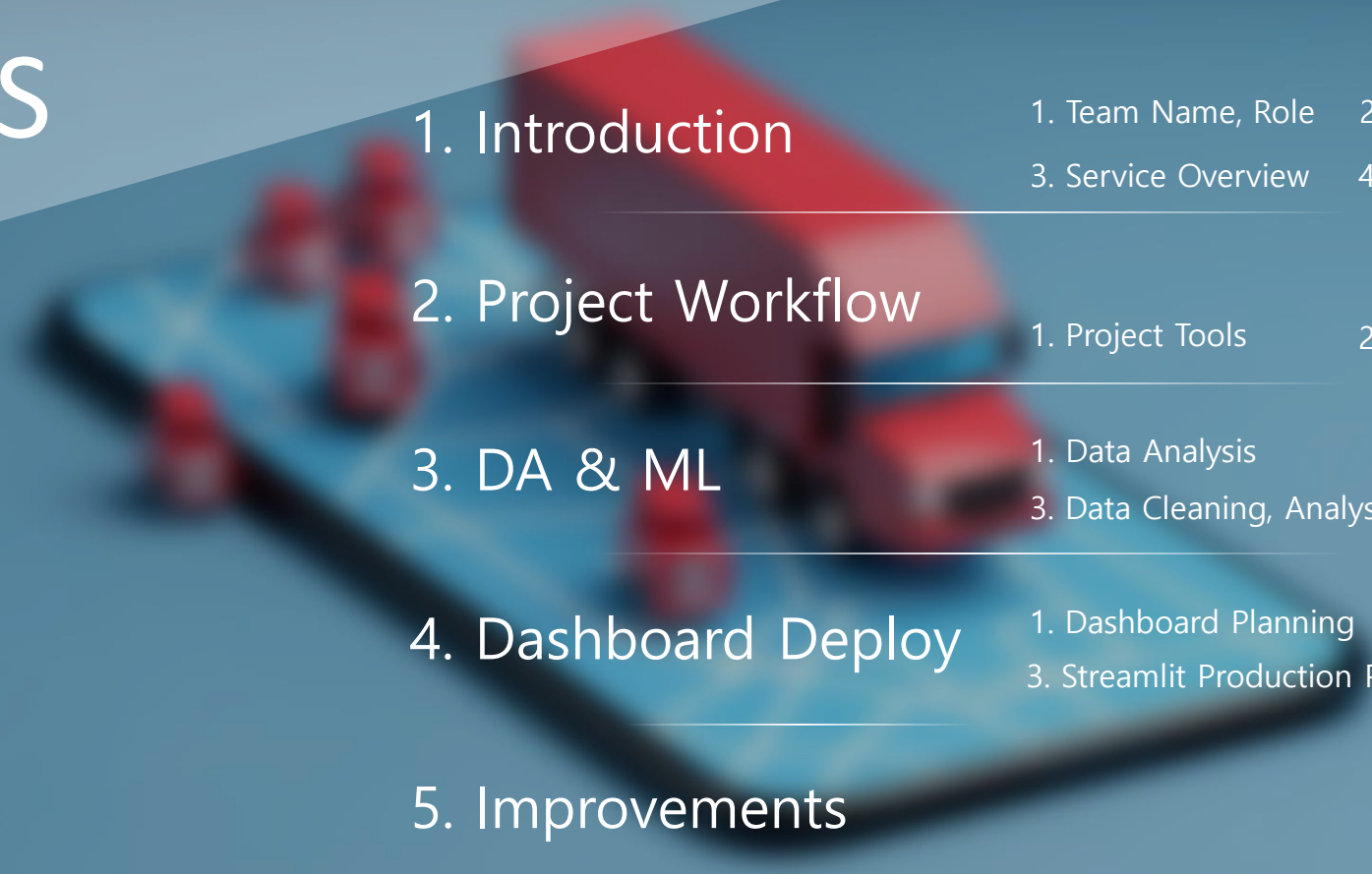
The 28th Data Analysis & Engineering (Python) Bootcamp

Final Project

Delivery Volume Forecasting & Analysis Service

Haeun Kim, YongEun Shim, Nayoon Na, Sanghyeok Lee, SeungJoon Lee, Yeonwoo Choi

CONTENTS



DeliverEase

Delivery + Ease

Project Objective: The primary goal of this initiative is to optimize and streamline the delivery process for enhanced convenience.

Data Cleaning

Haeun Kim YongEun Shim
Nayoon Lee Sanghyeok Lee

Project Management

SeungJoon Lee

Data Collection

Entire Team

Dashboard

Haeun Kim SeungJoon Lee
Nayoon Lee Yeonwoo Choi

Data Analysis / Statistical Analysis

Haeun Kim YongEun Shim
Nayoon Lee Sanghyeok Lee

Machine Learning

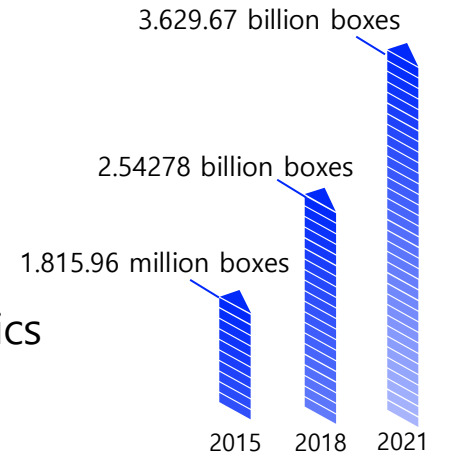
SeungJoon Lee

Topic Selection Background

Non-face-to-face becomes a trend after COVID-19, and volume surges

Courier volume fluctuates depending on regional and timing characteristics

Difficulty in effective staffing



Domestic parcel delivery market volume trends

E-commerce competition driving shorter delivery times

Serious issue of heavy workload for transport drivers

It is 'No Delivery Day' today, but Coupang continues operations, citing a high volume of orders.

As online shopping surges, couriers face a concerning risk of overwork-related stress.

Provided based on regional, temporal, and parcel-specific characteristics

Delivery Volume Forecasting & Analysis Service

Service Overview

FOR Site Manager for Courier Delivery Operations

IN Seoul

WHY

- Over 70% of the total parcel volume is handled in the Seoul metropolitan area
- In Seoul, the majority of delivery sub-terminals are independently operated by autonomous districts, facilitating clear distinctions
  (23 out of 25 autonomous district manage their respective operations)
- While there is a need for improvement in inventory management, the expansion of parcel workshop sites within the city is constrained
  - It is identified as the foremost area requiring delivery volume prediction services

DATA FROM CJ Logistics

- The preeminent transportation company boasting the largest domestic market share, consistently holding approximately 50% of the market share annually

Expectation Effectiveness

Enhance Inventory Management

Conducting demand forecasting using item volume data

Accurate prediction incorporating local and temporal factors
to prevent excess inventory in volume forecasting

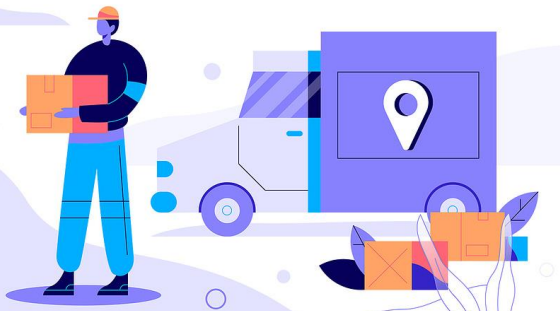
Optimizing subterminal locations

Utilizing analysis of regional characteristics and autonomous district-level data to facilitate the strategic selection of parcel delivery sub-terminal locations

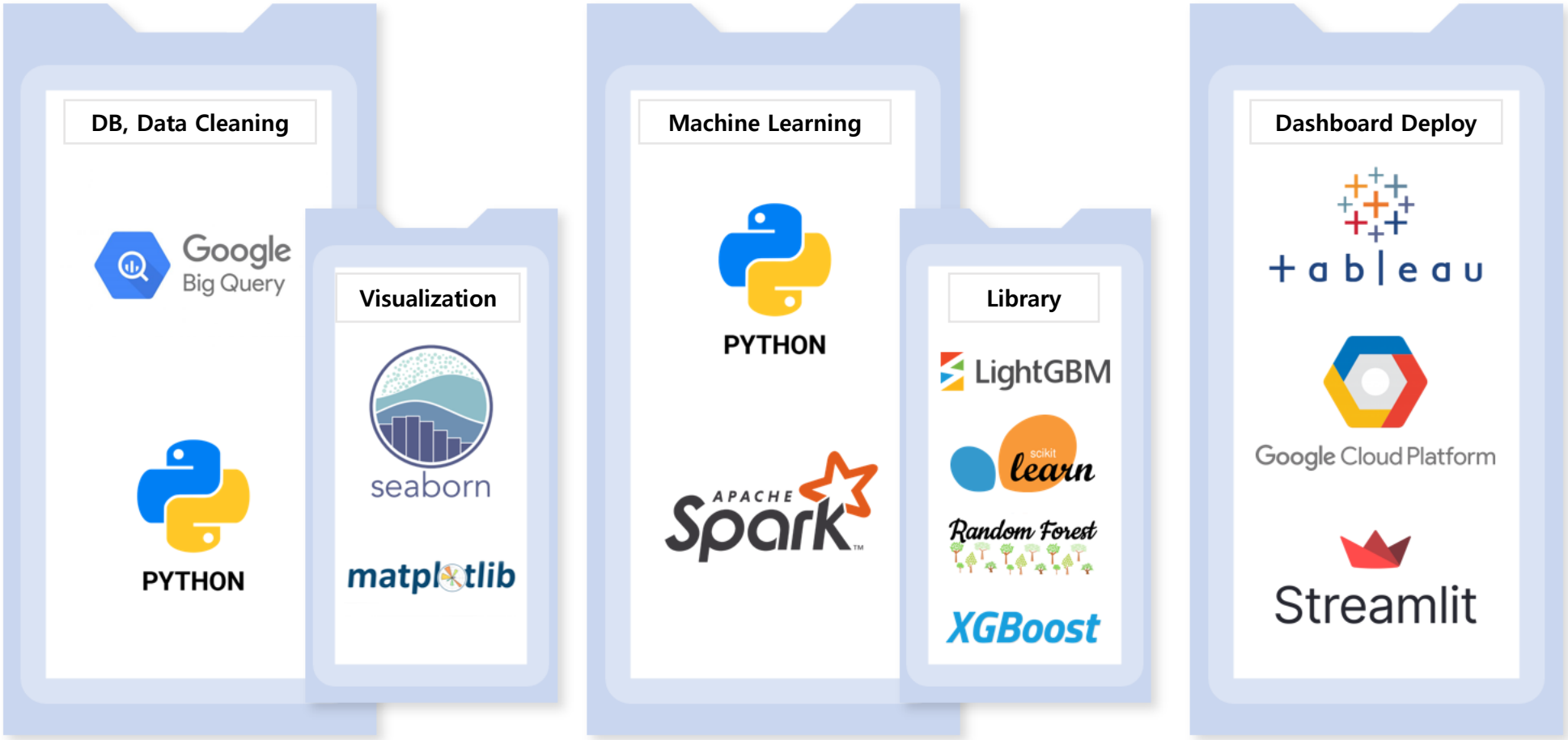
Enhance the working environment for courier personnel

Efficient workforce allocation through inventory management based on volume prediction

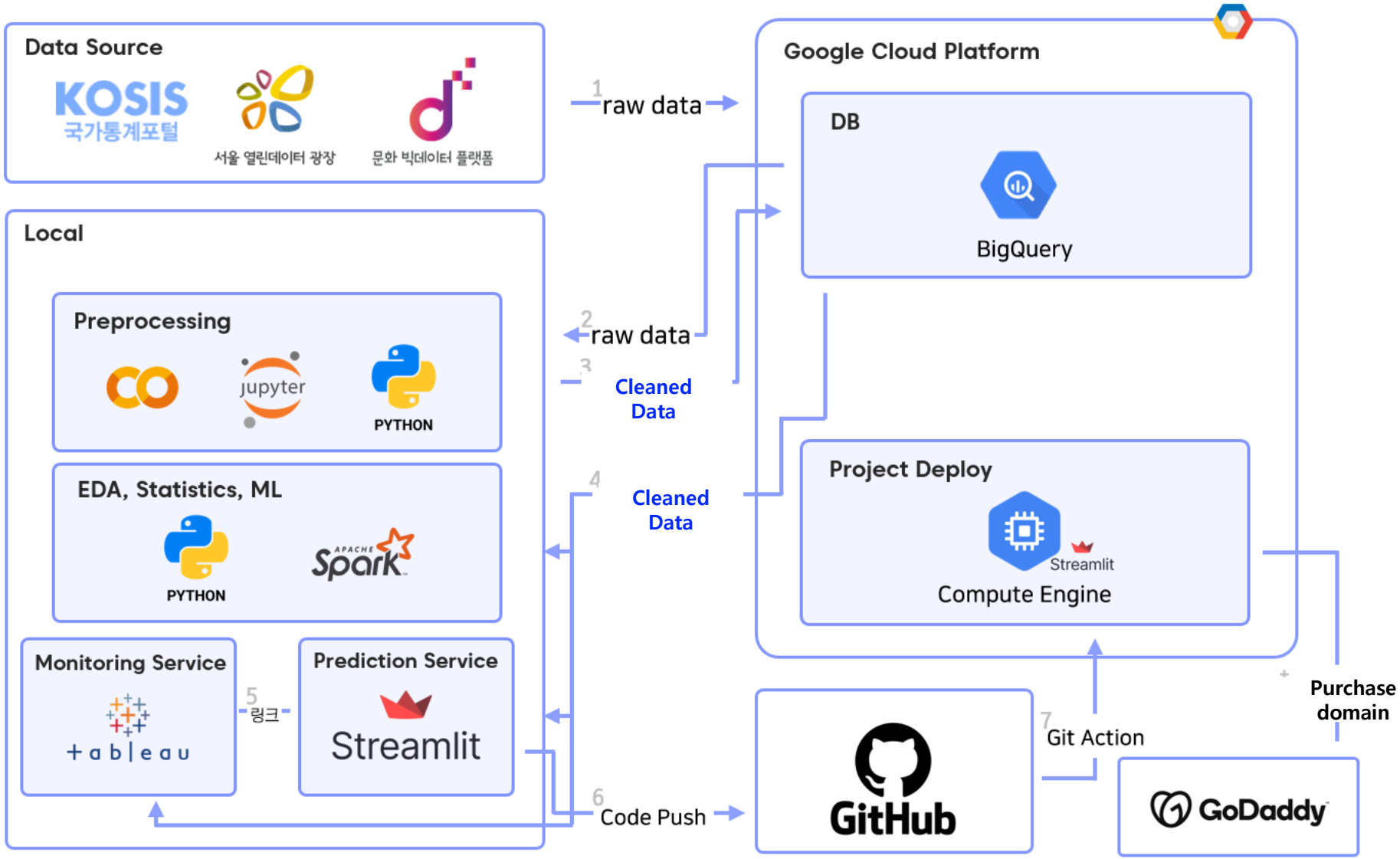
Enhancing the work environment through the reduction of workload.



Tools


DB, Data Cleaning
Google Big Query
PYTHON
Visualization
seaborn
matplotlib
Machine Learning
PYTHON
APACHE Spark
Library
LightGBM
scikit learn
Random Forest
XGBoost
Dashboard Deploy
tableau
Google Cloud Platform
Streamlit

System Architecture



Data Analysis Process

Understanding variables that vary the volume of delivery for delivery prediction
Data analysis for monitoring service

Hypothesis setting, Data Collection and cleaning

H0 : There is a significant relationship between the proportion of the elderly and the amount of living/health packages per capita.
H2 : There is no relationship between the proportion of the elderly and the amount of living/health packages per capita.

- Data secure & cleaning by ratio of distinguished elderly people

Conduct and test Correlation Analysis

Data Collection

Number of Data

Total

25

Datasets Collected

Contexts of Collected Data

Collection of Data by region
Logistics Volume

Population (age group statistics, birthrate, the number of household members)

House (number of apartment units, price)

Industry (pharmaceutical manufacturing and sales, distributor, food sanitation statistics)

Public Service (power usage, public library, oil usage)

Data Source

https://data.seoul.go.kr/

https://www.bigdata-culture.kr/

https://kdx.kr/

http://kostat.go.kr/

EDA / Analysis of the proportion of the elderly

Analysis Process

Compiled by total column of Seoul 2022 resident registration population (by age) Considering the difference in volume due to the difference in population, it is analyzed as the amount of living/health delivery per person

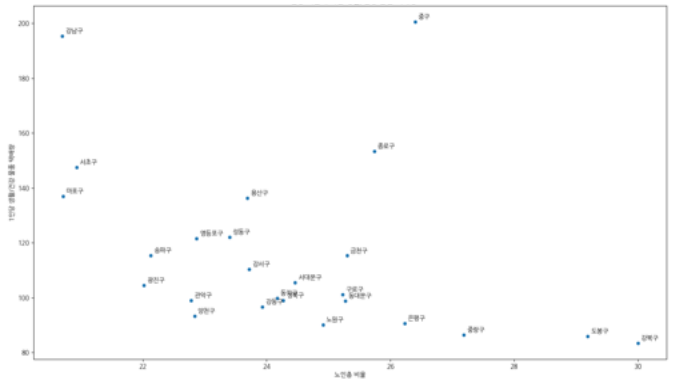
* *per capita daily/health delivery volume*

$$\frac{\text{Daily/health delivery volume}}{\text{Total population by region}}$$

Analysis Result

- Pearson correlation coefficient : -0.38 (negative correlation coefficient)
* p-value > 0.05 → **Not statistically significant**
* Although not statistically significant, the higher the percentage of elderly people, the weaker negative correlation between the amount of living health

A scatterplot of the proportion of Life/Health items and the elderly (2022)



Regression Analysis result of Life/Health items and the elderly (2022)

OLS Regression Results

Dep. Variable:	1인당 생활/건강 물품 택배량	R-squared:	0.146
Model:	OLS	Adj. R-squared:	0.109
Method:	Least Squares	F-statistic:	3.946
Date:	Fri, 10 Nov 2023	Prob (F-statistic):	0.0590
Time:	00:37:38	Log-Likelihood:	-119.03
No. Observations:	25	AIC:	242.1
Df Residuals:	23	BIC:	244.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	238.8497	62.375	3.829	0.001	109.818	367.882
노인층 비율	-5.0716	2.553	-1.987	0.059	-10.353	0.210

Omnibus:	21.750	Durbin-Watson:	1.077
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28.800
Skew:	1.961	Prob(JB):	5.57e-07
Kurtosis:	6.502	Cond. No.	259.

EDA / Elderly Rate Analysis

Reason Assumption

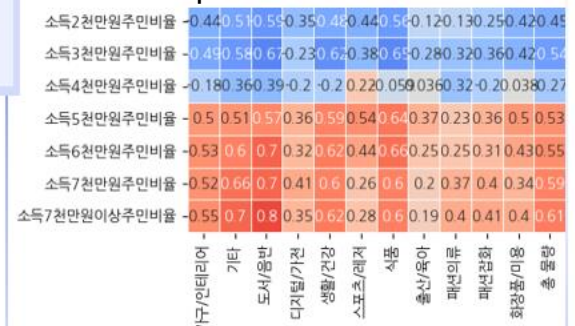
Korea's Poverty Rate for the Elderly **37.6 %**
(OECD Average 13.5%)

In the case of Gangnam-gu, where the income level is high, the elderly are very small and the number of living/health delivery per person is very high
the relationship between delivery volume and income level

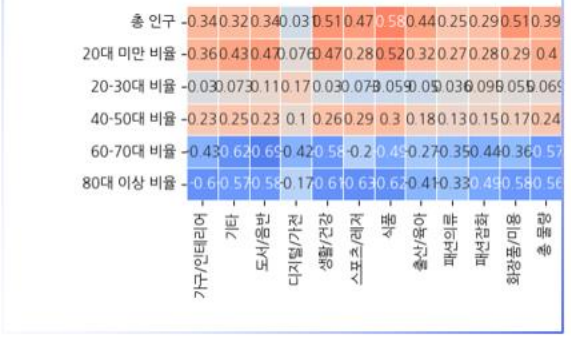
EDA / Income level, Age group Analysis

Pearson correlation coefficient analysis is conducted with differential population, age, and income level ratio data and life logistics data

Heat map between the percentage of delivery volume/per district income level



Heat Map between parcel volume / per-district age distribution



Analysis Result

Income level and **parcel delivery volume** generally have a **clear positive and negative correlation**

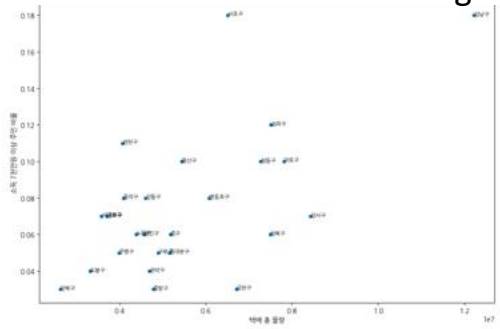
High-income areas tend to have a lot of parcel delivery
Areas with high low-income rates tend to have fewer packages

There is a clear correlation in furniture/interiors, others, books/recordings/living/health, and food items

사용 데이터
2022년 서울(보내는 곳)-서울(받는 곳) CJ 생활물류 데이터
변수1. 전국 시군구 단위 소득 구간대별 주민 비율 (2023년 8월)
변수2. 행정구역(시군구별) 주민등록세대수 (2022년)

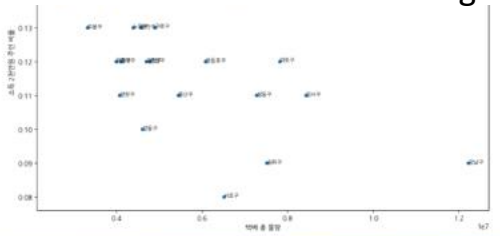
Conduct correlation analysis by adding income-related variables

Scatter plot of delivery volume/income of KRW 70 million or more among residents



택배 물량/소득 2천만원 주민비율간 산점도

Scatter plot of delivery volume/income of KRW 20 million or more among residents



EDA / Birth rate by district

Analysis Process

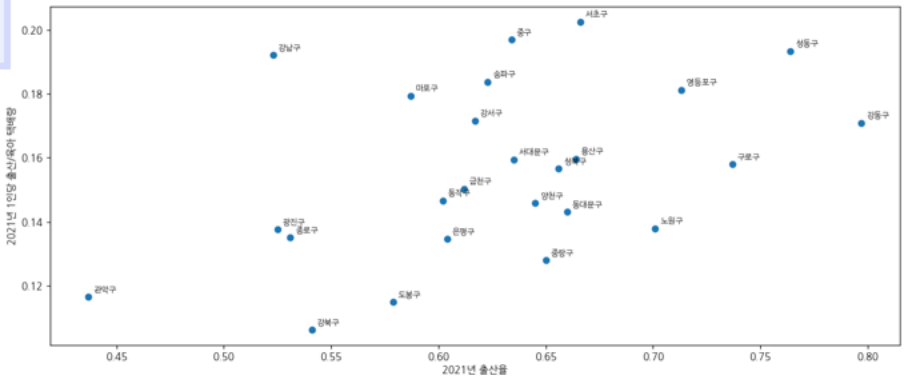
Considering the difference in volume due to the population difference, the analysis is based on the number of deliveries/childcare packages per person

* Birth/childcare delivery volume per person

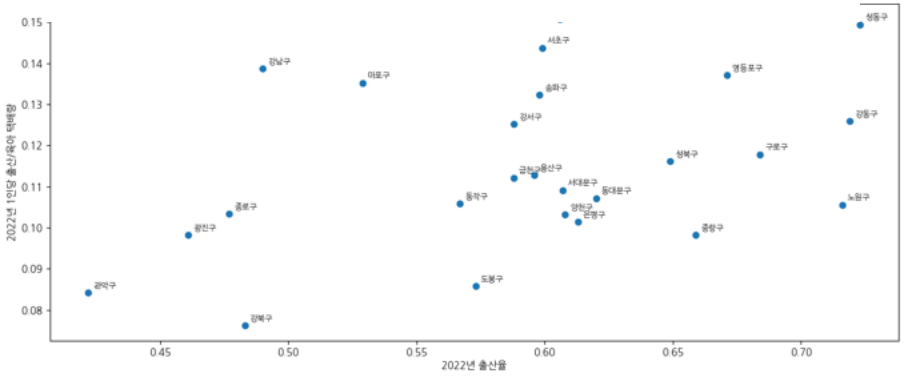
$$\frac{\text{Birth/childcare delivery volume}}{\text{Total population by district}}$$

Analysis of Pearson correlation coefficient with fertility data and life logistics data by autonomous district

A scatterplot of the number of childbirth/childcare items and the birth rate per person (2021)



A scatterplot of the number of childbirth/childcare items and the birth rate per person (2022)



Analysis Result

* **Overall rightward** scatter
* Pearson correlation coefficient is 0.44 in 2021 and 0.39 in 2022, **a weak positive correlation coefficient**
* However, a significant threshold for statistics comes out with a p-value of 0.05 for 2022

사용 데이터
2021/2022년 서울-서울 물류 데이터
데이터 1. 서울시 구별 출산율 (2021~2022)
데이터 2. 서울시 주민등록인구 (2021~2022)

EDA / Birth rate by district

Problem solving

Caused by a lack of data on the birth rate volume

Add National > Seoul data to logistics data

To explore variables, analyze all items

Analysis Result

+ Birth rate **correlation coefficient** rises (0.48 in 2021, 0.4 in 2022 / 0.44 in 2021 and 0.39 in 2022)

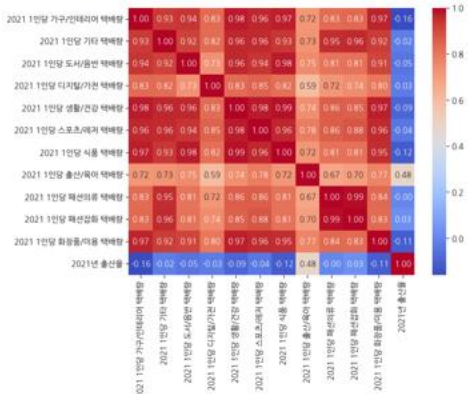
+ Statistical significance as p-value < 0.05

* Except for birth/childcare items, the absolute value of Pearson correlation coefficient is low, so there is no linear relationship

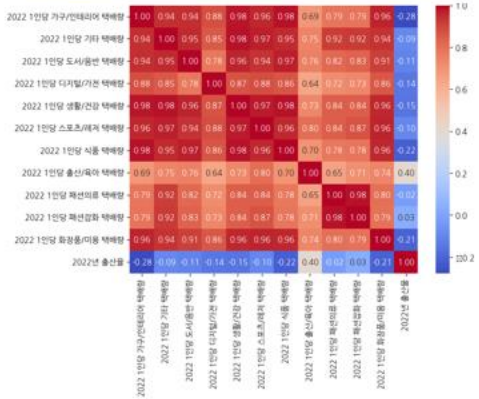
사용 데이터
2021/2022년 서울-서울 물류 데이터
데이터 1. 서울시 구별 출산율 (2021~2022)
데이터 2. 서울시 주민등록인구 (2021~2022)

Analysis of Pearson correlation coefficient with fertility data and household logistics data by district

Heatmap of per capita delivery volume and birth rate (2021)



Heatmap of per capita delivery volume and birth rate (2022)



EDA / Analysis Result

Additional analysis for machine learning data

After EDA Analysis

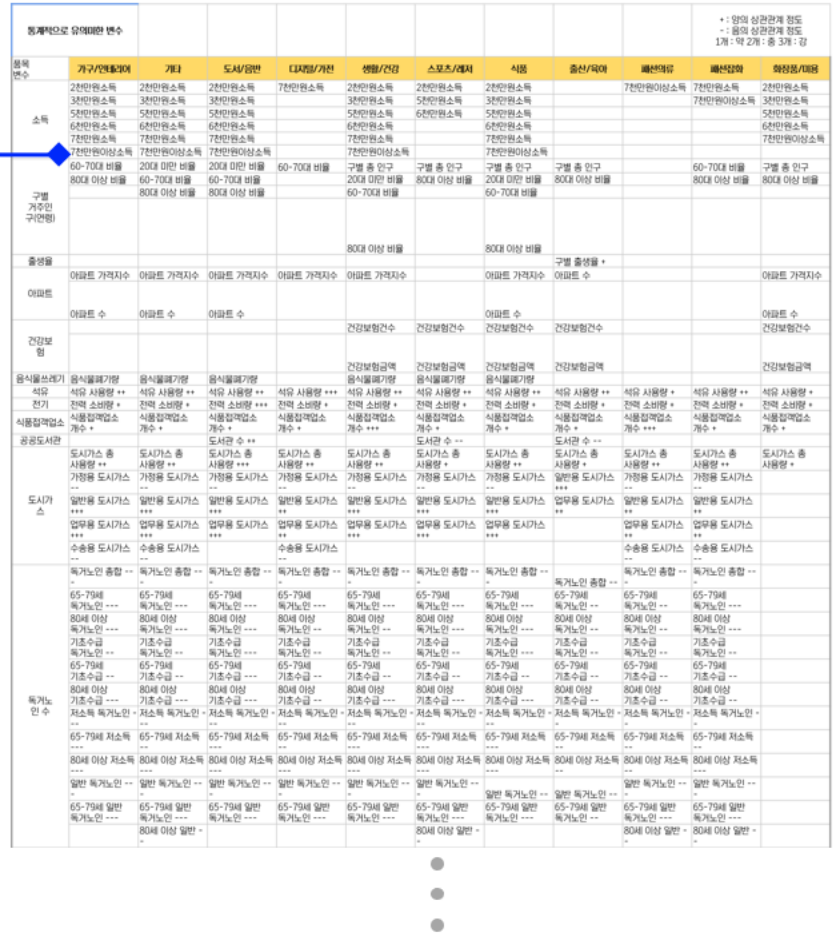
To produce item-by-item volume forecast data, correlate variables are organized by item

Statistically significant result

Significant correlations in 17 out of a total of 24 variable data sets were identified

Dataset of significant variables

Income, population by category, birth rate, number of apartments/price index, health insurance, food waste emissions, oil consumption, electricity consumption, number of food service establishments, number of public libraries, city gas usage, number of elderly people living alone, COVID-19 vaccination rate, number of households, number of registered hydrogen cars, population density, number of housing types



EDA / Visualization

Analysis about short term event

Analysis Background

Analysis of variables that could affect short term delivery volume, such as anniversaries and product launches

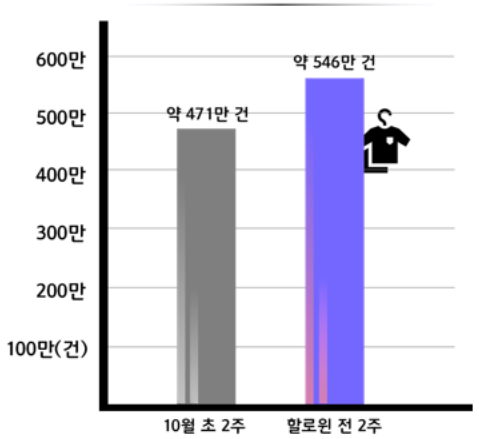
Analysis Process

Compare and visualize the amount of delivery before and after the variable period

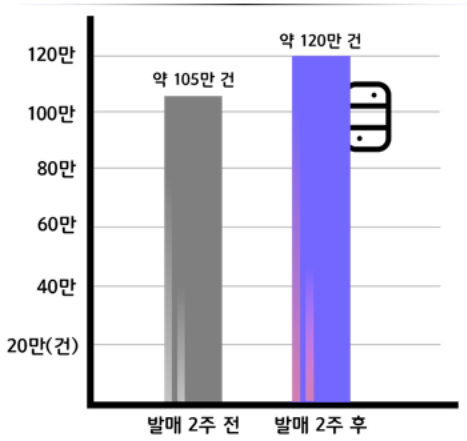
Visualized Variables

Holidays, idol singer albums released, new products released, new games released, Halloween, social distancing

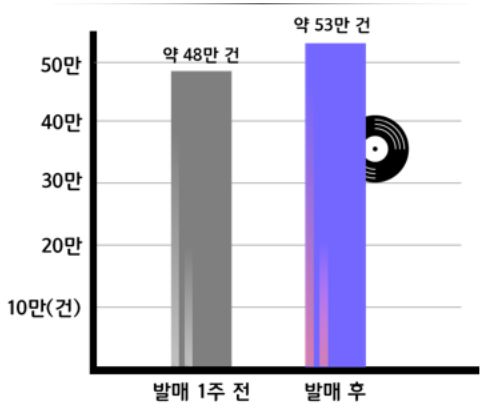
Halloween Fashion/Clothing Volume Changes in 2021



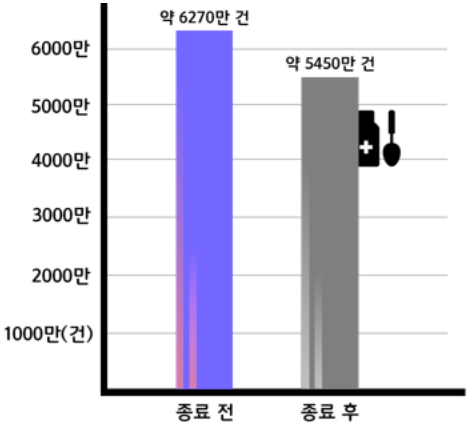
Zelda's Legend 'Tears of the Kingdom' Compared to Before and After the Release Date



Before and after the release date of BTS' "Proof" album

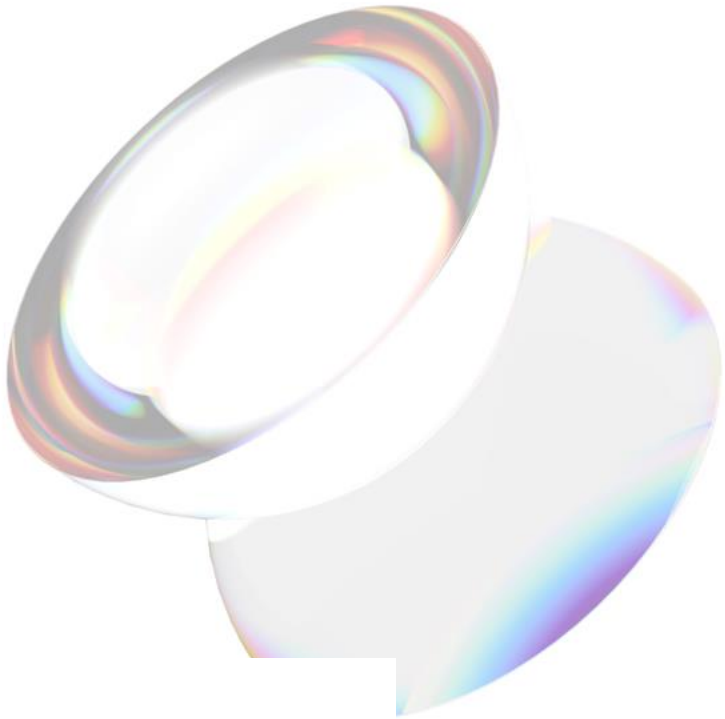


Changes in social distancing/health volumes



Data Analysis

FOR MACHINE LEARNING



Data Cleaning Summary

1. Pre-processing of destination/sending location/workplace-based data
2. Check the data - column name unification, data type change
3. Pre-processing and merging by each variable
4. Handle outliers
5. Elimination of strike period (Jan 2022)

Used Data

1. Logistic Data in Seoul (2021-2023 August)
2. Variables data

Machine Learning Data ERD

CJ Logistics Data

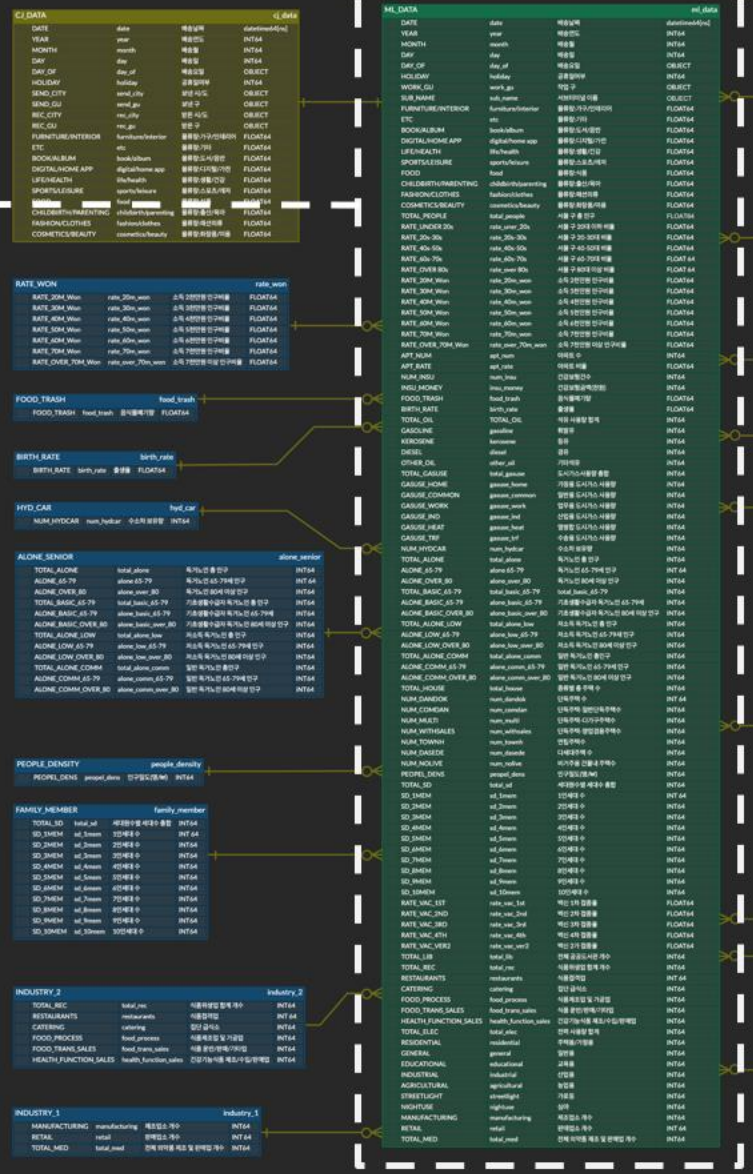
Nationwide > Seoul
Seoul > Nationwide
Seoul > Seoul

Sending
Receiving

Receiving in Seoul Sending from Seoul Work district

Create 3 datasets

Data for Machine Learning



Machine Learning

Independent variables

All variables correlated with delivery volume

Dependent variables

11 items of delivery volume
(Create 11 models)

Reasons to divide dependent variables by item

1. Provide information so that users can think of the type and number of vehicles they need, considering the characteristics of the logistics transport vehicle (size and refrigerated/frozen).
2. Differentiate items for each item because significant variables are different for each item, in case the model is trained later.

Additional Data cleaning

- Remove DATE column (use year/month/day column only)
- DAY_OF (Sunday) Change column to numeric (Monday=1, Tuesday=2...)
- Remove columns with duplicate information
- One-hot encoding variable simple expression (Seoul, Gangnam-gu, Seoul -> Gangnam...)

Machine Learning/1st Test

Used Data

- Seoul – Seoul Logistics data in 2022

Independent Variables (x)

- Year, month, day, holiday
  3 datasets
  20 컬럼
  Use **9094** raw data

Spark vs scikit-learn

- **Scikit-learn** is much better in **Performance and Wall Time**

RF vs LGBM vs XGBoost

- Chose the **highest** R2 value of **XGBoost** at 0.82 - 0.98 per item

Hyperparameter Tuning

- Use RandomizedSearchCV, conduct Cross-Validation

Performance after Tuning

- Improve performance of 8 items including book/record (-0.08), childbirth/parenting (-0.02), food (-0.02) and others
- R2 value: 80s-90s, RMSE value: 20% to 30% error with actual average volume in 10s-20s

Machine Learning/2nd Test

Used Data

- Seoul Logistics Data 2021-2023(August)

Model 33 models created in total

- Three types of models by item

1. By sub terminal
2. Seoul to nationwide
3. Nationwide to Seoul

17 Datasets
110 Columns(fields)
24306(work) **939171**(rec)
Used **924411**(sending) raw data

Work Location Model

Remove Outliers

제거 전

	Book/ Record	Life/ Health	Sports
R2	0.13	0.67	0.88

제거 후

	Book/ Record	Life/ Health	Sports
R2	0.87	0.94	0.90

나머지 품목 0.01~2 성능 상승

Assume the reason

Outliers in logistics between Seoul and the region, missing values in the case of books/records affecting publication and release

Sending and Receiving Model

Problem occurred

Approximately 40 times the total logistics model data capacity

Lack of local environmental memory due to overcapacity
-> Using Jupyterlab in a GCP environment

Hyper Parameter

parameter range, n_iter value adjustment
up to 0.02 performance improvement

Machine Learning/Final Performance Model

Total logistics model

R2
0.93-0.98
Average 0.96
Minimum:
Books/Record,
Maximum: Life/Health

RMSE
248-1688
Minimum: Furniture,
Maximum:
Fashion/Clothing

Sending logistics model

R2
0.87-0.97
Average 0.94
Minimum:
Books/Record,
Maximum:
Fashion/Clothing

RMSE
6.4-42.9
Minimum: Sports,
Maximum:
Fashion/Clothing

Receiving logistics model

R2
0.69-0.97
Average 0.83
Minimum:
Books/Record,
Maximum: Food

RMSE
6.4-54.1
Minimum: Sports,
Maximum:
Fashion/Clothing

Setting Scenario

Problem Situation

1. Situation to allocate workforce
2. Situation to allocate delivery volume because of partial strike, fire, accident and etc

Scenario to use dashboard

1-1 Selection of logistics volume in Gangseo district, the location in charge
1-2 Inquire only last New Years holiday
1-3 Inquire the logistics volume of feature items
2-1 Inquire the location and distance of a nearby terminal
2-2 Inquire the logistics volume of non-workable and nearby terminals

Action through dashboard

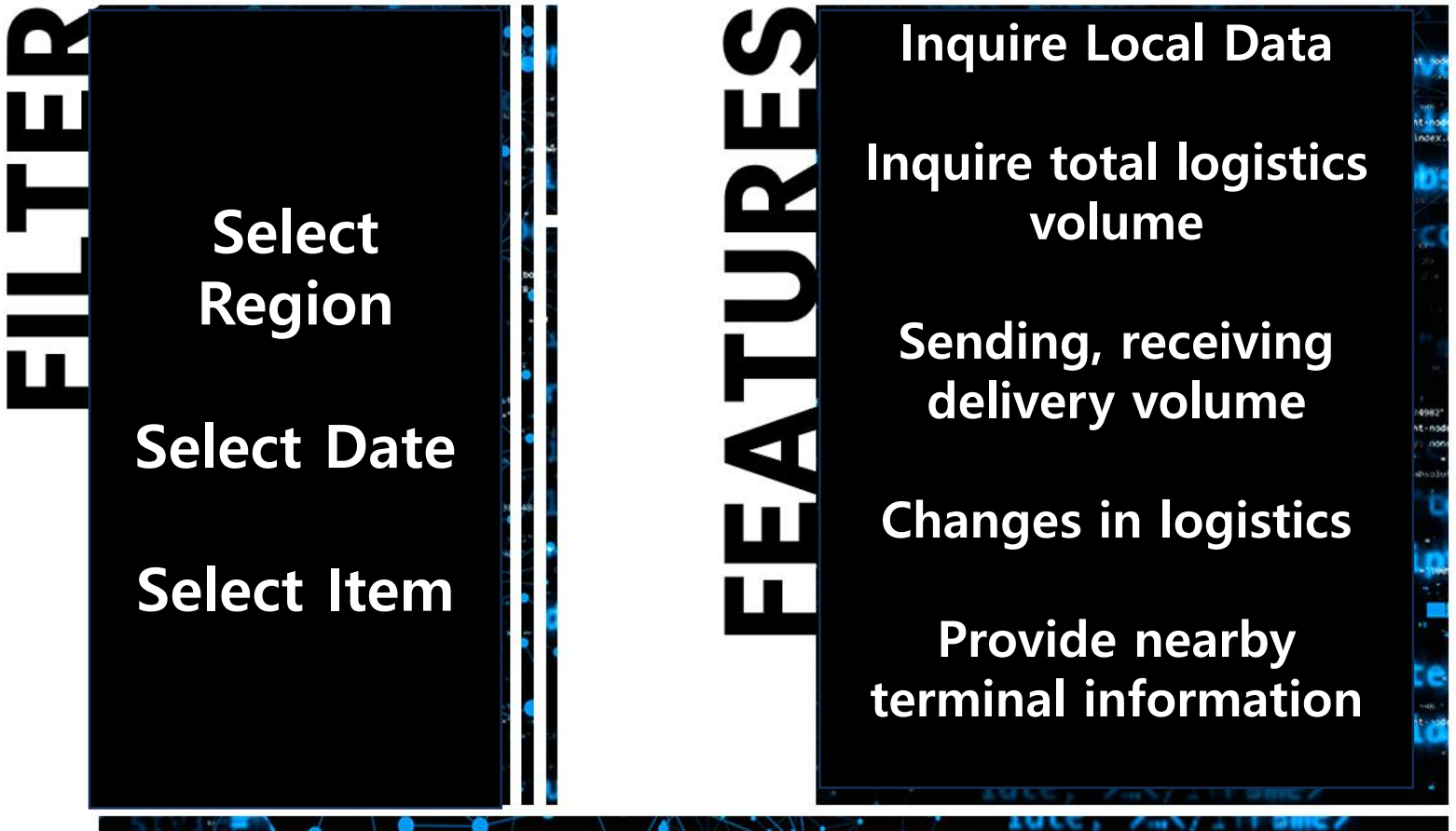
Inquire the total logistics volume of last year's facilities in the region and the logistics volume by item

Recruitment and allocation of manpower according to logistics volume

Expect the need for frozen trucks, large trucks, etc through the logistics volume of each item

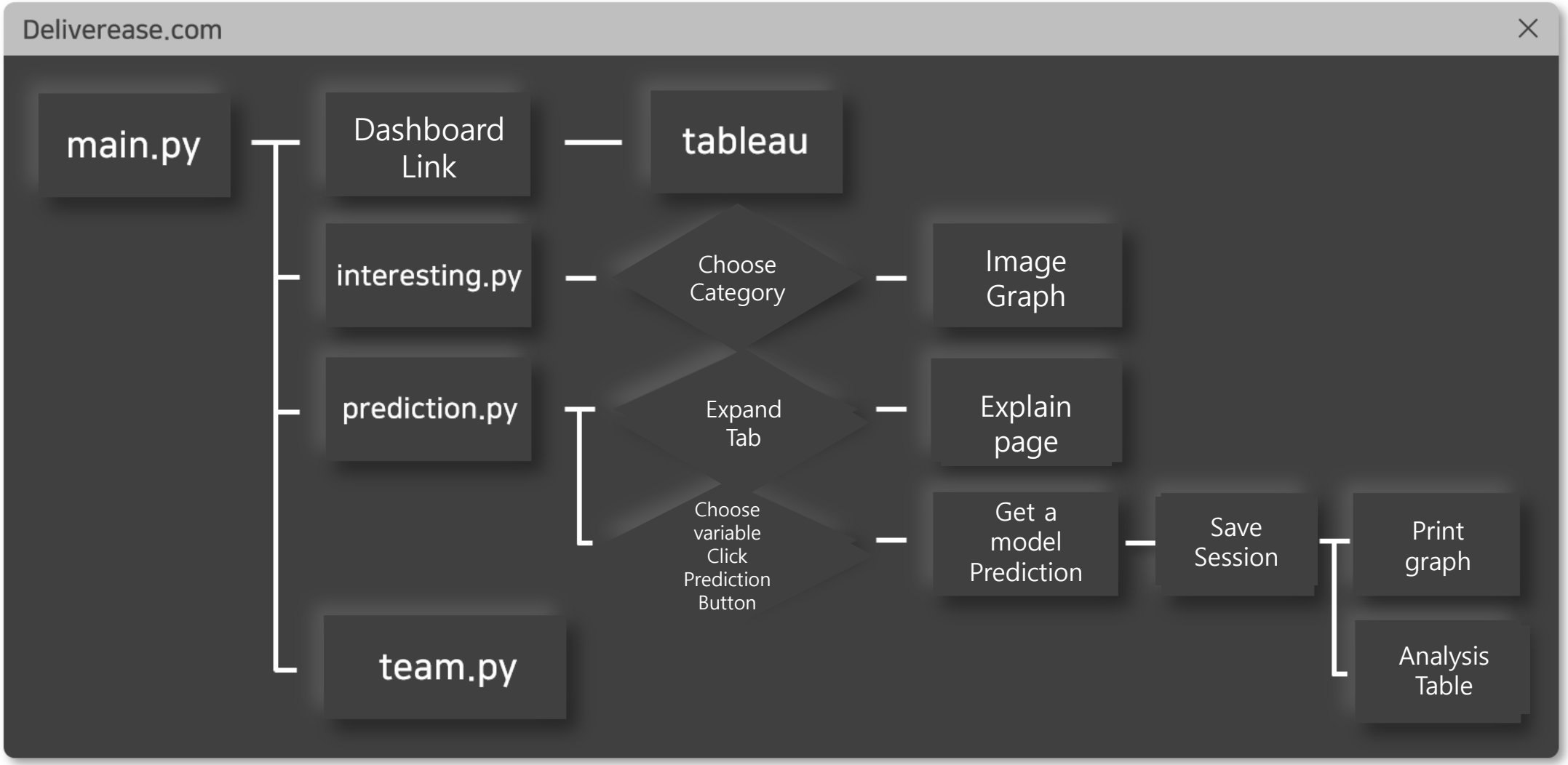
Dashboard Function Plan

Identify needs through dashboard scenarios

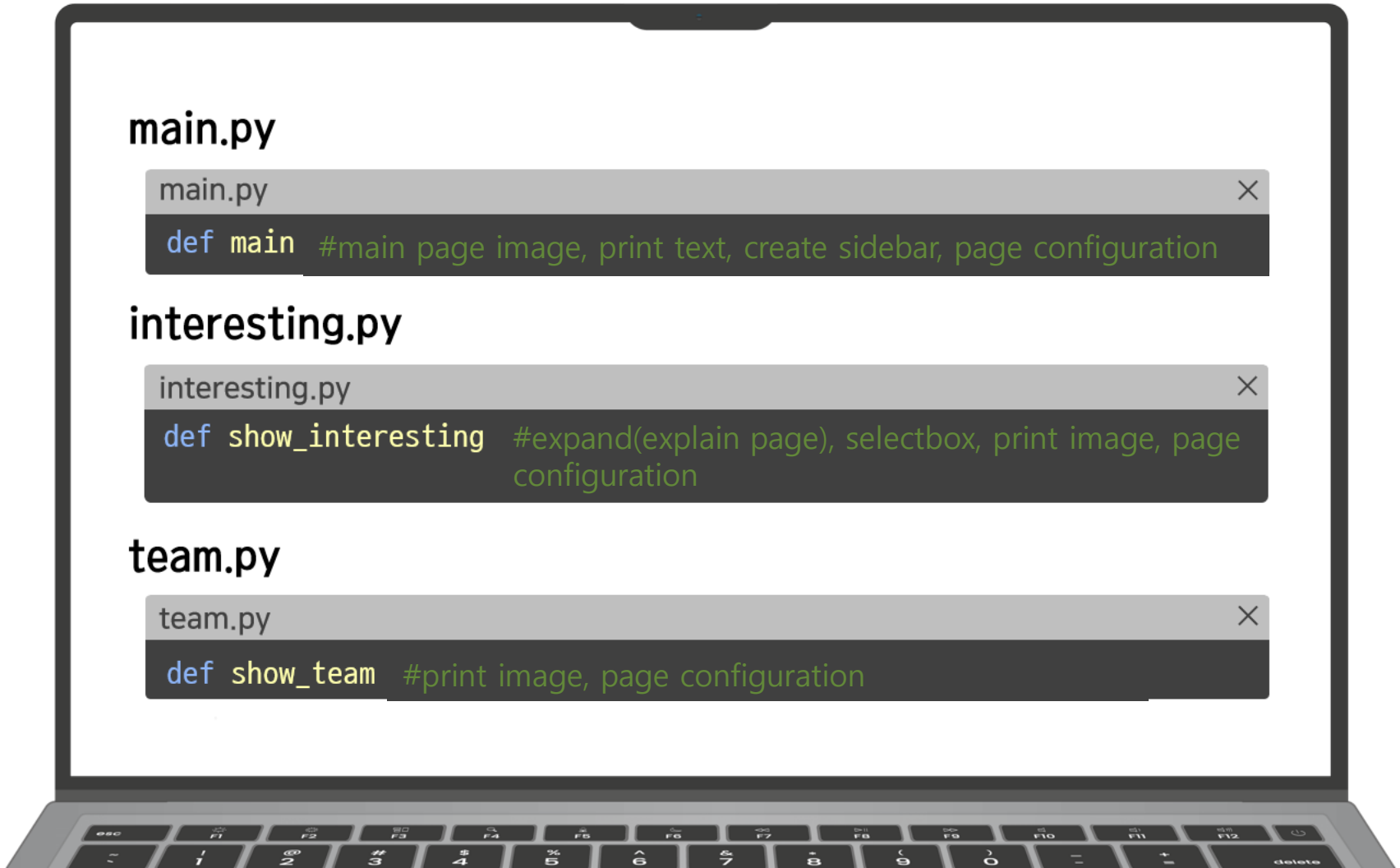



NOW LET'S MAKE
DELIVERY EASE
http://deliverease-multifpp.com

Streamlit Structure



Function for each page



prediction.py

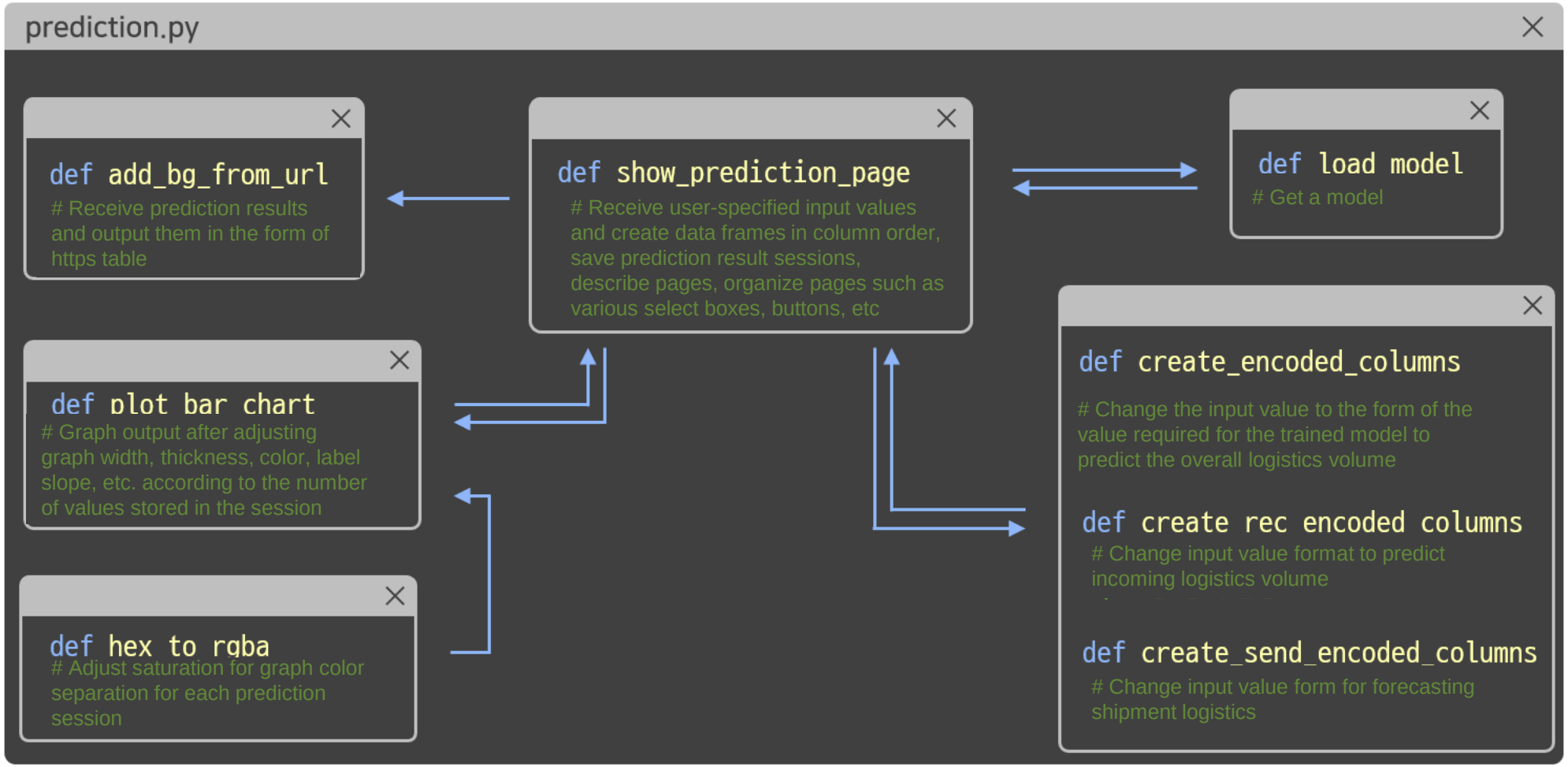
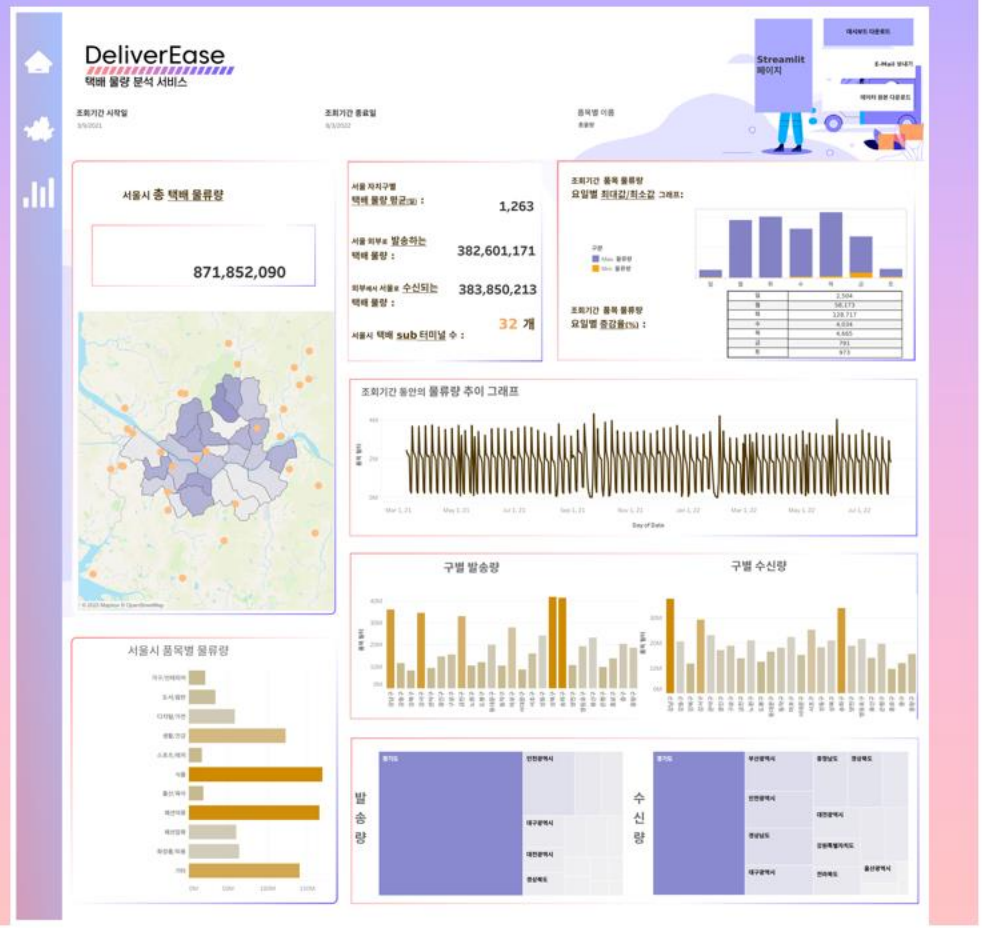


Tableau Dashboard

Seoul Dashboard

https://public.tableau.com/app/profile/.13586515/viz/df3_1_re/sheet14_1?publish=yes



Dashboard by ddistrict

https://public.tableau.com/app/profile/.13586515/viz/df3_2_re/sheet20?publish=yes

Improvement

DATA

Shortage in logistics data

Lack of data accumulated for a long time
(2021-2023.08)
Data of CJ Logistics only exists

Requires subcategories in data

PLANNING ANALYSIS

Lack of domain knowledge

Whether the actual site provides the necessary services

Requires additional variables

Need to explore more variables that affect logistics changes

FEATURES

Tableau

Implement a trend graph in which the total value and average value of each item are drawn simultaneously

Streamlit

Implement to represent only unique values for each prediction

Resources

‘택배 없는 날’ 끝나자 하루 택배물량 900건 폭증…택배기사 “야근 불가피한 업무강도” (택배 없는 날 이후 물량 증가)

신촌人의 택배 상자에는 OOO이 많다? (신촌 택배 물량 분석, 기타 지역도 존재)

추석 택배물량 폭주 집하 마감… 선물 못 보낸 시민들 발만 동동 (추석 택배 물량 증가로 택배사들이 집하를 마감하면서 미처 선물을 보내지 못한 시민들 발생)

1인가구 증가에 온라인 거래↑..택배 물동량 70% 수도권 몫 - 데일리팝 (1인가구 증가에 온라인 거래↑..택배 물동량 70% 수도권 몫)

(서울시 생활물류서비스 시설 확충과 지원방안 - 택배서비스 중심으로 - | 서울연구원 : [연구보고서] 서울시 생활물류서비스 시설 확충과 지원방안 - 택배서비스 중심으로)

"2030년 경기도 생활물류시설 185만㎡ 부족" ("2030년 경기도 생활물류시설 185만㎡ 부족" - [연구보고서a] 수도권 생활물류 1,000만 개 시대 새로운 물류시스템 구축이 필요)

쿠팡, 오늘 ‘택배 없는 날’인데 운영 강행…“물량도 몰려” (쿠팡로지틱스서비스(쿠팡CLS)는 올해도 택배 없는 날에 동참하지 않은 채 운영을 강행)

[스페셜 리포트] 택배업 (국내 택배시장 주요 지표)

추석 전 택배물량 평소보다 17% 증가 예상…4주간 특별관리 | YTN (추석 전 택배물량 평소보다 17% 증가 예상…4주간 특별관리)

‘택배없는 날’ 물량 폭탄 맞은 기사들…“후유증 무서워 쉬겠나” | 동아일보 (‘택배 없는 날 이후 물량 폭탄)

코로나19로 급증한 물류창고…스마트 물류 주도권은 누가? - 머니투데이 (코로나 이후 증가한 물류창고업 등록 건수 18년 253개 ,22년 586건)

[D리포트] "미리 택배 보냅시다"…'미주단' 추석 산더미 택배 방지 캠페인 ([D리포트] "미리 택배 보냅시다"…'미주단' 추석 산더미 택배 방지 캠페인- 작년대비 추석 물량 많았다는 뉴스)

"6년간 20여명 사망"…명절 특수에도, 택배기사 파업 '현재진행형' – 아주경제

온라인 쇼핑 2배 늘 동안 택배기사는… 이유 있는 '과로사' - 한국일보

쿠팡, 오늘 ‘택배 없는 날’인데 운영 강행…“물량도 몰려” - 한겨레

무료 이모티콘 출처

multicampus

The 28th Data Analysis & Engineering (Python) Bootcamp

Final Project

DeliverEase