2021-1학기 응용머신러닝 Final Project

배달 데이터를 활용한 Regression/Clustering Modeling

데이터사이언스융합전공
2019312259 김유진

목 차

Introduction

Data 분석 및 처리

ML Method

Conclusion

Introduction

목 차

Introduction

Data 분석 및 처리

ML Method

Conclusion

Data 분석 및 처리

KT 빅데이터 플랫폼 - 업종-지역별 평균배달소요시간

kt 통신 빅데이터 플랫폼

검색 로그인 회원가입

통신 마켓플레이스 분석/시각화 커뮤니티 빅데이터 서비스 이용안내

Life Public Space

배달 상점 데이터

배달서비스를 제공하는 상점의 업종별 지역별(법정동,행정동,도로명주소) 정보

#배달상점 도로명주소코드 #배달상점 법정동코드, #배달상점 시군구, #배달상점

Data 분석 및 처리

EDA(Exploratory Data Analysis)

👉데이터: 업종-지역별 평균배달소요시간(2019-07-17 ~ 2020-09-30)

👉출처: 경기대학교 빅데이터 센터

	date	hour	menu	do	city	time
0	2019-07-17	0	도시락	경기도	의정부시	13.52
1	2019-07-17	0	돈까스/일식	경기도	의정부시	13.02
2	2019-07-17	0	돈까스/일식	충청북도	제천시	15.03
3	2019-07-17	0	배달전문업체	경기도	고양시 일산동구	19.49
4	2019-07-17	0	배달전문업체	경기도	의정부시	19.33
...	...	...	...	...	...	...
1940377	2020-09-30	23	회	경기도	화성시	30.40
1940378	2020-09-30	23	회	서울특별시	도봉구	30.97
1940379	2020-09-30	23	회	서울특별시	은평구	20.65
1940380	2020-09-30	23	회	전라북도	군산시	27.45
1940381	2020-09-30	23	회	충청남도	서산시	19.37

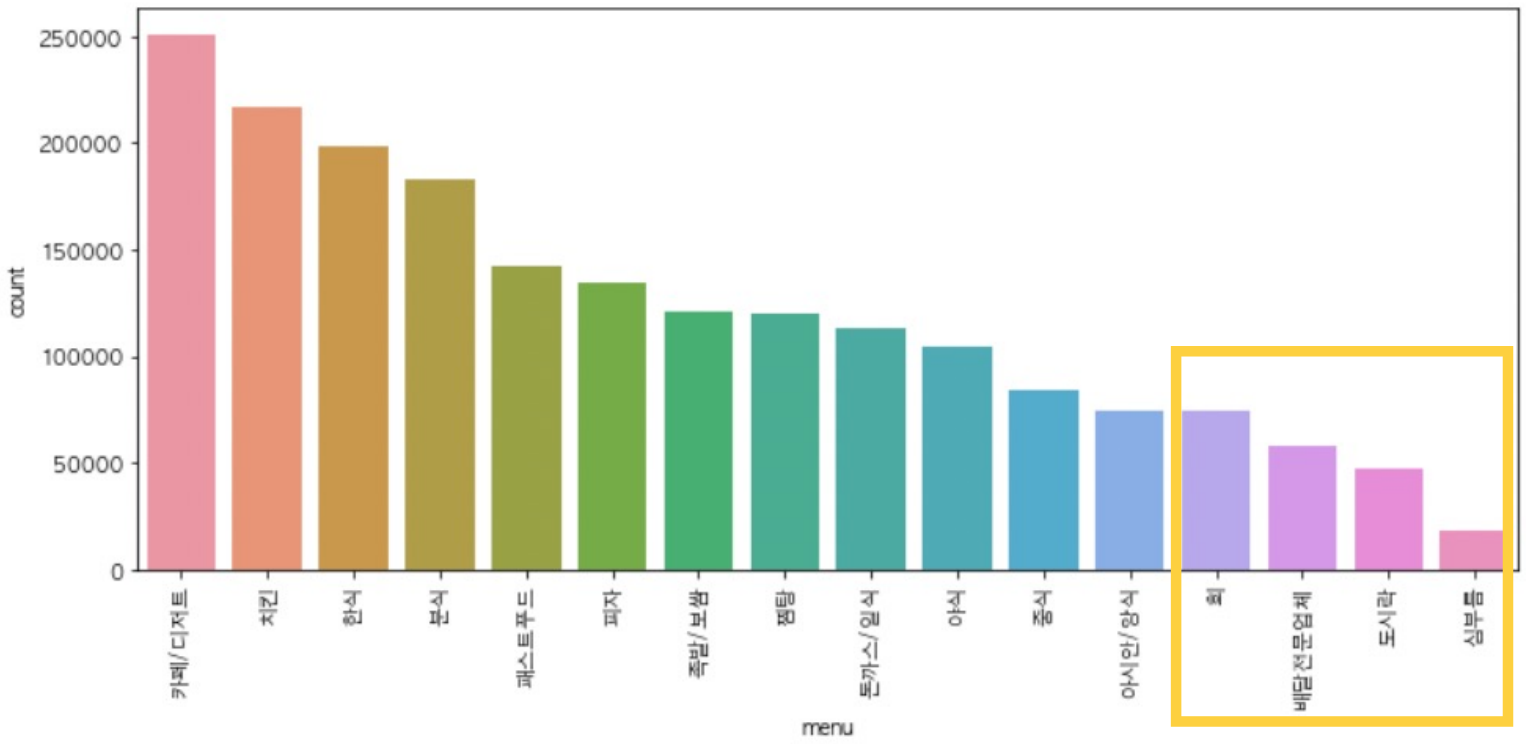
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1742717 entries, 0 to 1742716
Data columns (total 6 columns):
 #   Column  Dtype
---  ------  -----
 0   date    object
 1   hour    int64
 2   menu    object
 3   do      object
 4   city    object
 5   time    float64
dtypes: float64(1), int64(1), object(4)
memory usage: 79.8+ MB
```

Data 분석 및 처리

EDA(Exploratory Data Analysis)

1940382 rows x 6 columns → 데이터 row를 줄여보자!



👉하위 네 가지 메뉴 항목
'심부름, ' 도시락 ' , '배달전문업체', '회' 제거

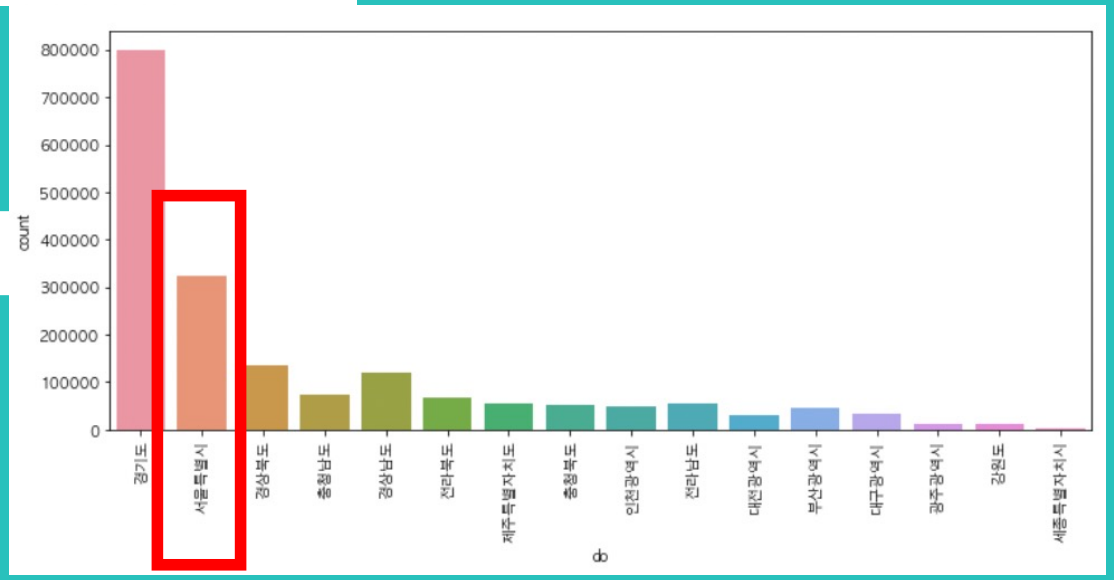
여전히 많은 row...1742717 rows x 6 columns

	date	hour	menu	do	city	time
0	2020-05-01	0	돈까스/일식	경기도	고양시 덕양구	11.33
1	2020-05-01	0	돈까스/일식	경기도	광명시	19.74
2	2020-05-01	0	돈까스/일식	경기도	의정부시	21.24
3	2020-05-01	0	돈까스/일식	경기도	평택시	12.32
4	2020-05-01	0	돈까스/일식	전라북도	군산시	14.67
...	...	...	...	...	...	...
131401	2020-05-31	23	한식	인천광역시	부평구	42.29
131402	2020-05-31	23	한식	전라북도	군산시	35.55
131403	2020-05-31	23	한식	제주특별자치도	서귀포시	10.58
131404	2020-05-31	23	한식	충청남도	서산시	18.69
131405	2020-05-31	23	한식	충청북도	제천시	32.41

131406 rows × 6 columns

2020년 5월 데이터로 줄이기

'서울특별시'로 제한



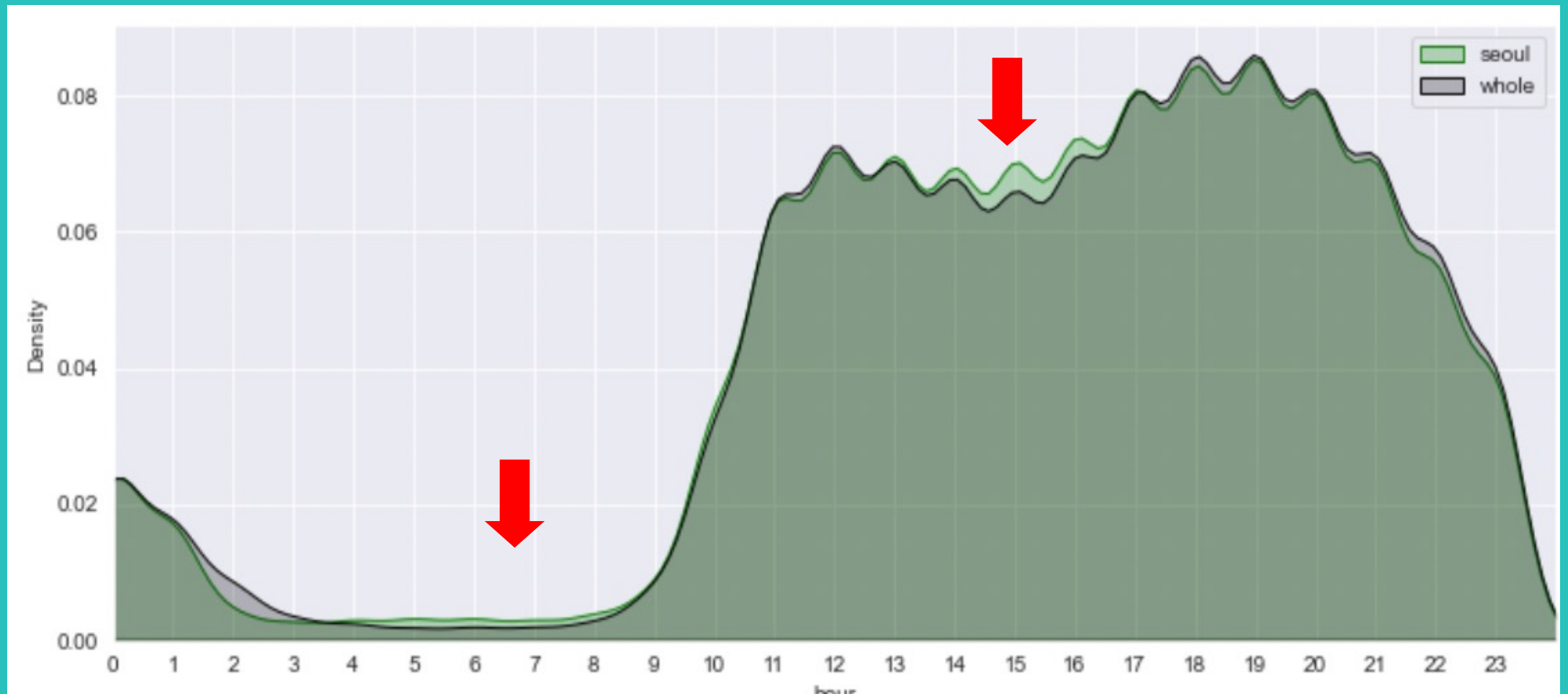
완성된 데이터프레임

	date	hour	menu	city	time
0	2020-05-01	0	분식	구로구	26.92
1	2020-05-01	0	분식	금천구	22.81
2	2020-05-01	0	야식	구로구	33.02
3	2020-05-01	0	야식	금천구	22.75
4	2020-05-01	0	야식	영등포구	18.93
...	...	...	...	...	...
20972	2020-05-31	23	피자	도봉구	19.76
20973	2020-05-31	23	한식	구로구	21.59
20974	2020-05-31	23	한식	금천구	16.05
20975	2020-05-31	23	한식	영등포구	33.09
20976	2020-05-31	23	한식	은평구	22.91

20977 rows × 5 columns

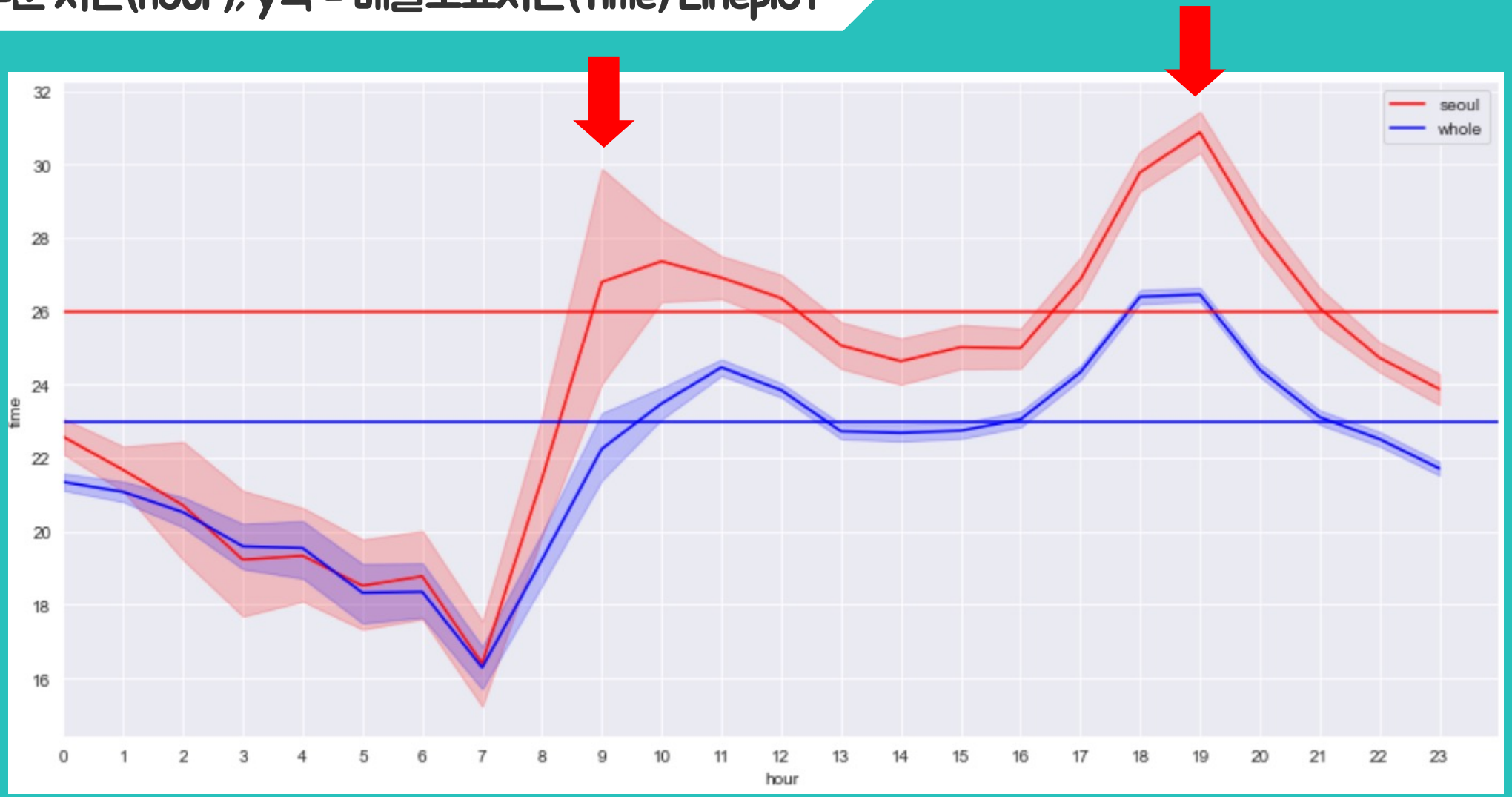
서울특별시/5월 배달 데이터 분석하기

Hour(주문시간) KDE plot



서울특별시/5월 배달 데이터 분석하기

X축 = 주문 시간(hour), y축 = 배달소요시간(time) Lineplot



Data 분석 및 처리

Data Preprocessing

```
#Mapping ordinal features
city_mapping = {
    '구로구': 0,
    '영등포구': 1,
    '동작구': 2,
    '구로구': 3,
    '은평구': 4,
    '금천구': 5,
    '도봉구': 6,
    '양천구':7,
    '관악구': 8,
    '노원구': 9,
    '강남구': 10,
    '강동구': 11,
    '서초구': 12,
    '서대문구': 13
    '강서구': 14
}
```

```
#월요일(1), 화(2), 수(3), 목(4), 금(5), 토(6), 일(7) mapping 시키기

date_mapping = {
    '2020-05-04':1, '2020-05-11':1,'2020-05-25': 1,
    '2020-05-05':2,'2020-05-19':2,'2020-05-26':2,'2020-05-12': 2,
    '2020-05-20':3,'2020-05-13':3,'2020-05-06':3,'2020-05-27': 3,
    '2020-05-21':4,'2020-05-07':4,'2020-05-28':4,'2020-05-14': 4,
    '2020-05-15':5,'2020-05-01':5,'2020-05-22':5,'2020-05-08': 5,
    '2020-05-09':6,'2020-05-23':6,'2020-05-02':6,'2020-05-16':6,'2020-05-30':6,'2020-05-29': 6
    '2020-05-03':7,'2020-05-10':7,'2020-05-17':7,'2020-05-24':7,'2020-05-31': 7
}
encoding['date'] = encoding['date'].map(date_mapping)
```

👉Categorial value → 숫자로 encoding

목 차

Introduction

Data 분석 및 처리

ML Method

Conclusion

Machine Learning Method

배달 소요 시간 예측

Regression 적용

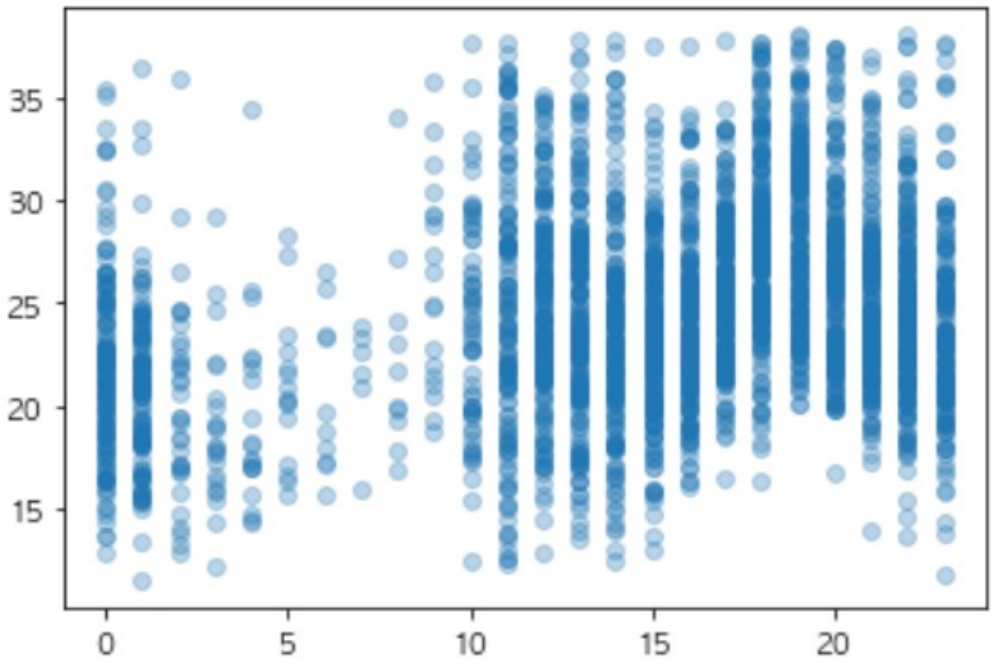
주문 분포 확인

Clustering 적용

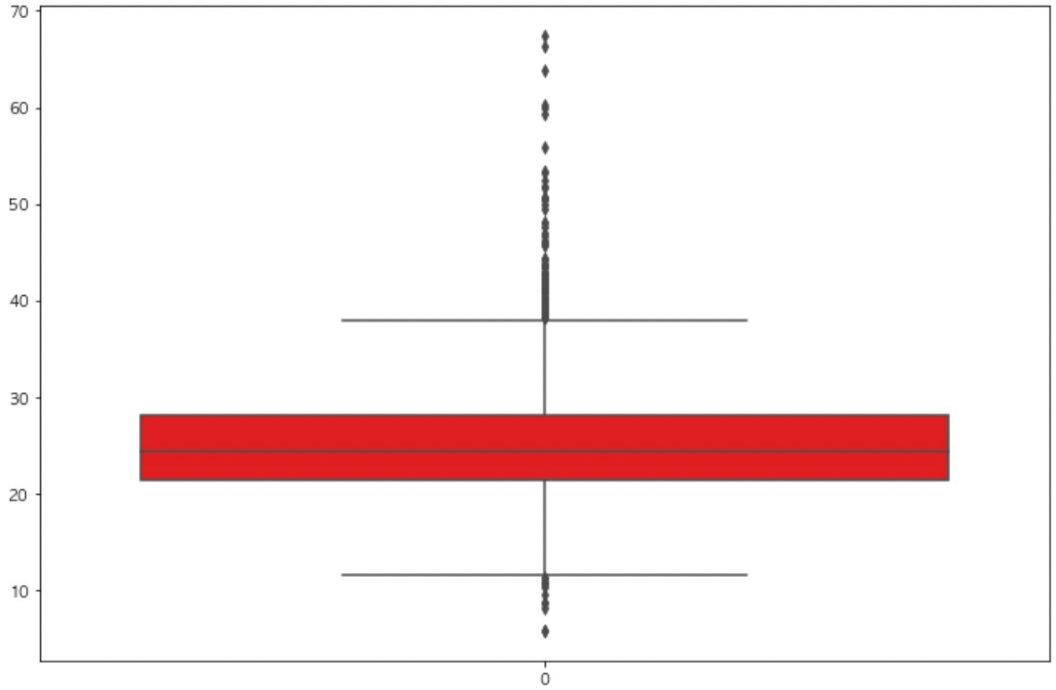
Machine Learning Method

Regression 적용

X= 주문 시간(hour), y=배달소요시간(time) scatterplot



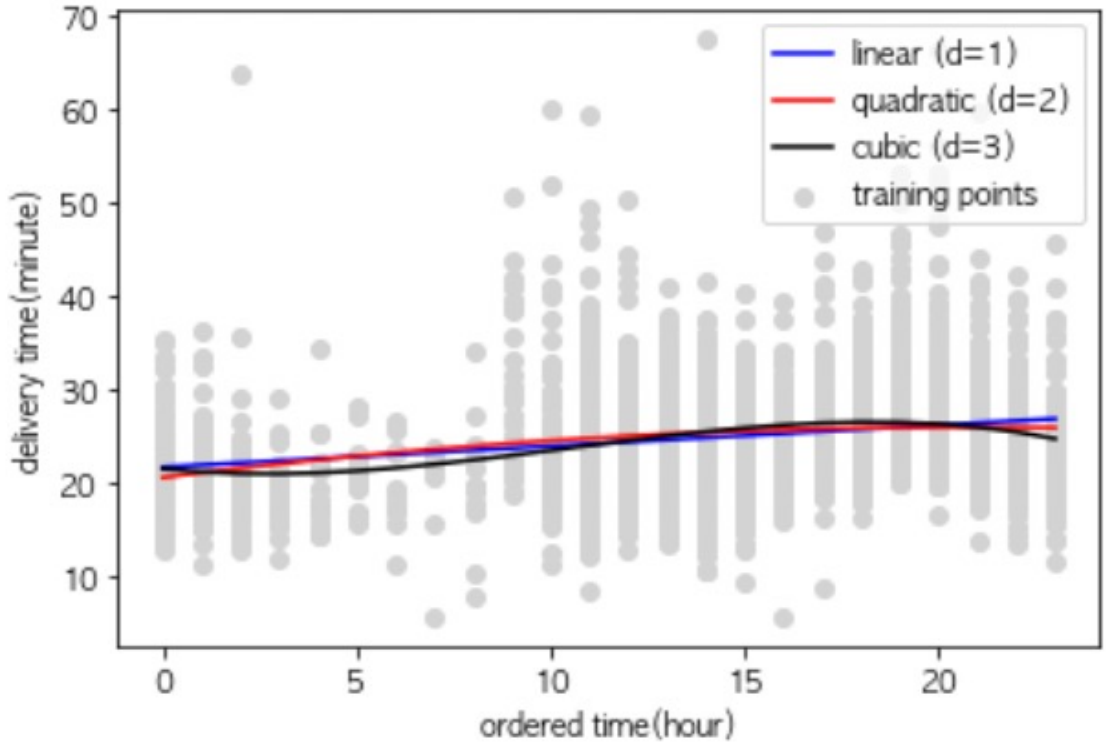
Boxplot을 이용해 확인한 Time Outlier



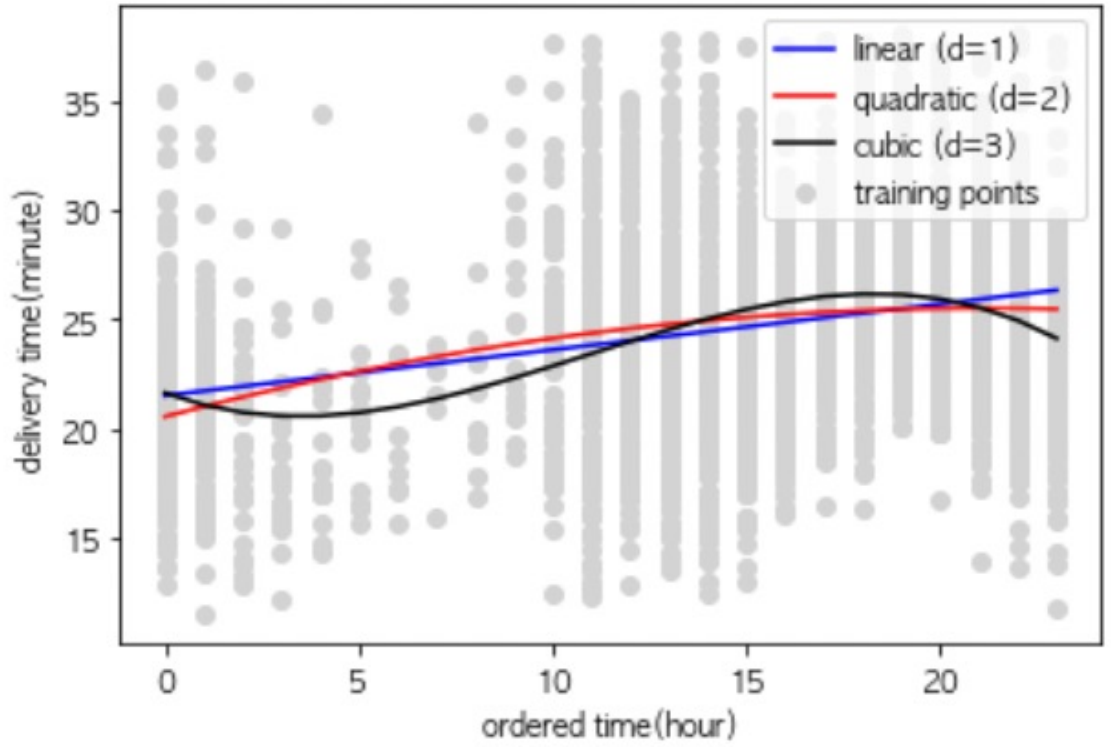
Machine Learning Method

Outlier 제거 전/후 Regression 결과 비교 ex) menu - chicken 데이터

```
MSE      Linear: 35.20,          Quadratic: 34.88,        Cubic: 34.49
R2       Linear: 0.05,  Quadratic: 0.06,         Cubic: 0.07
```




```
MSE      Linear: 22.81,          Quadratic: 22.56,        Cubic: 22.01
R2       Linear: 0.07,  Quadratic: 0.08,         Cubic: 0.10
```

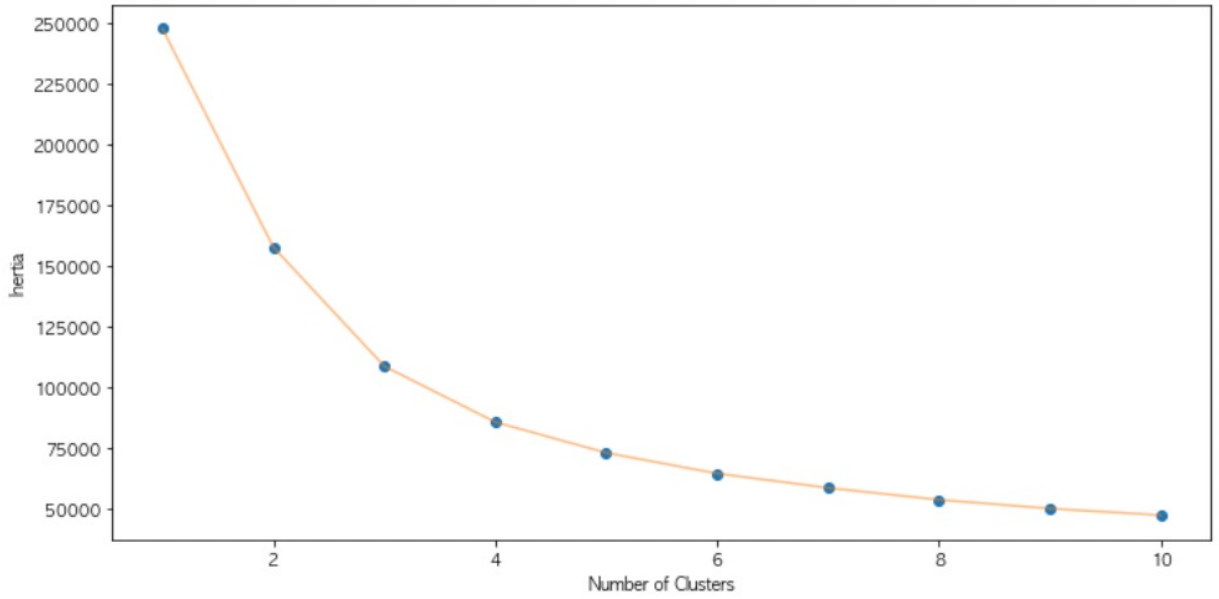


Machine Learning Method

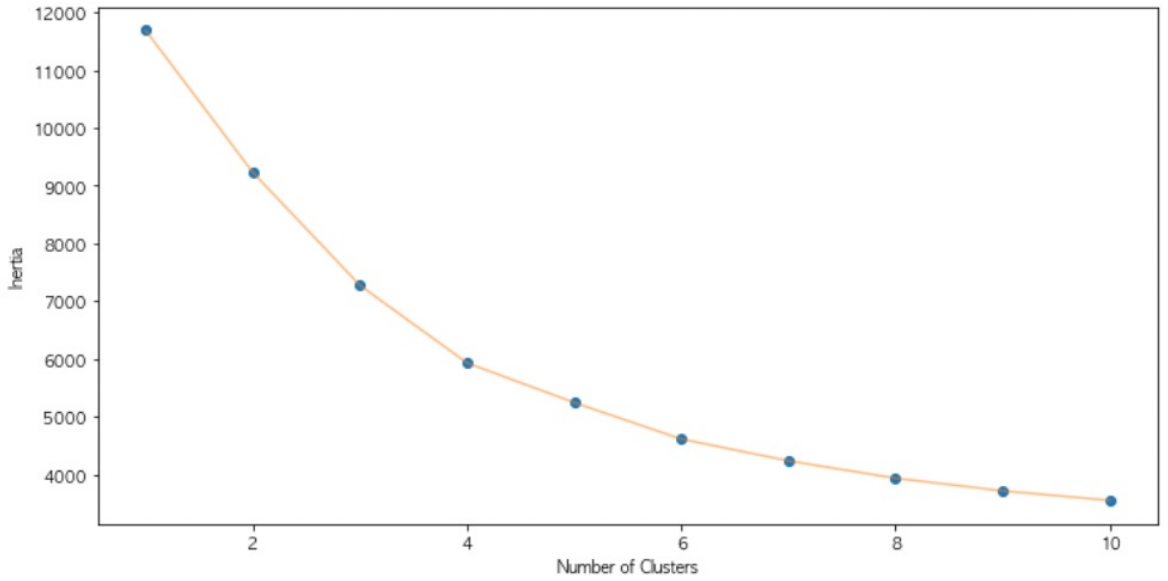
Clustering 적용 - K-means Clustering ex) menu - chicken 데이터

K값에 따른 군집 내 거리 확인

```
#스케일링 전
change_n_clusters(k_range, encoding)
```

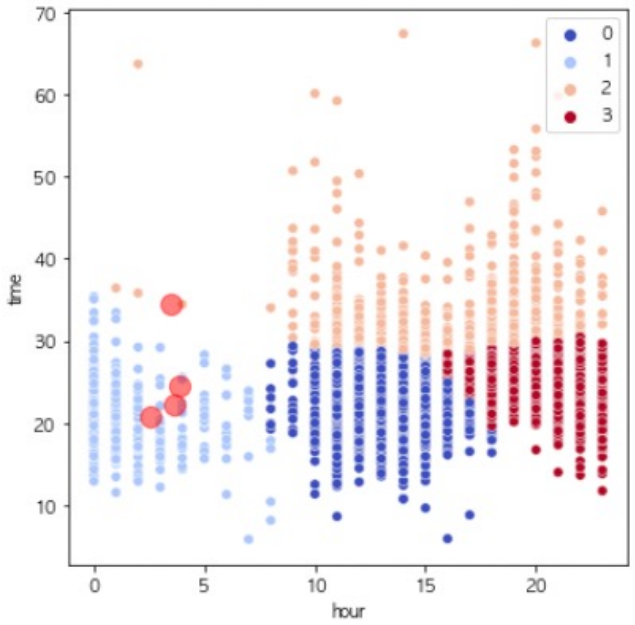


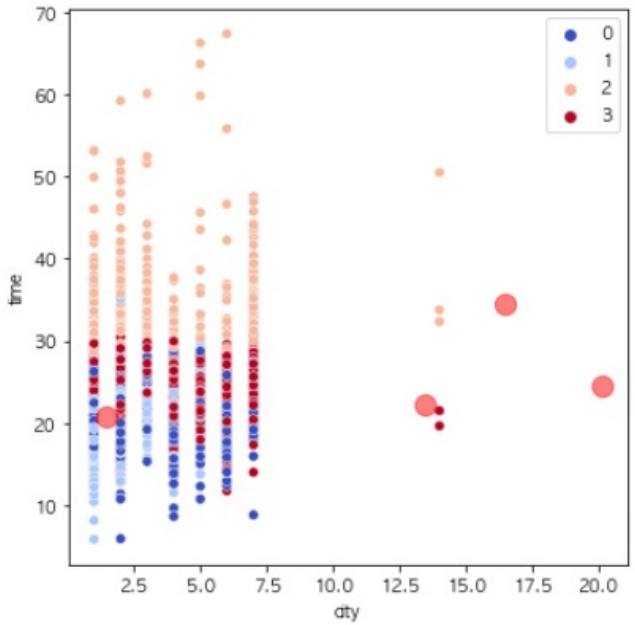

```
#스케일링 이후

change_n_clusters(k_range, scaled_encoding)
```

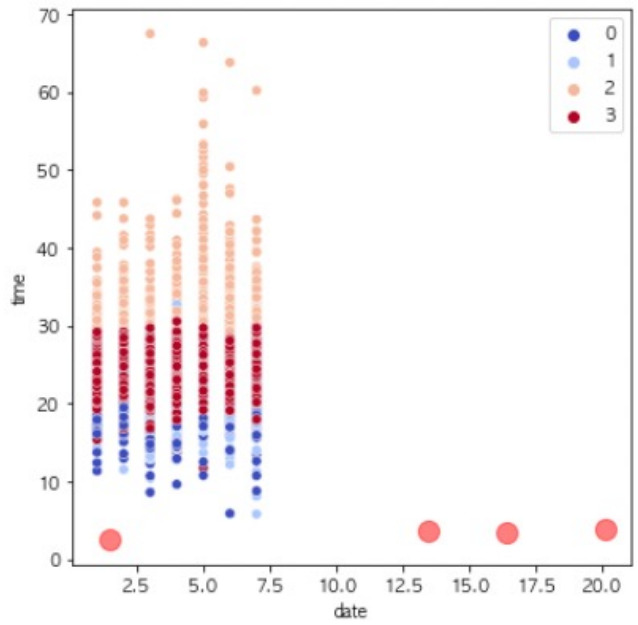


K=4 clustering 결과.

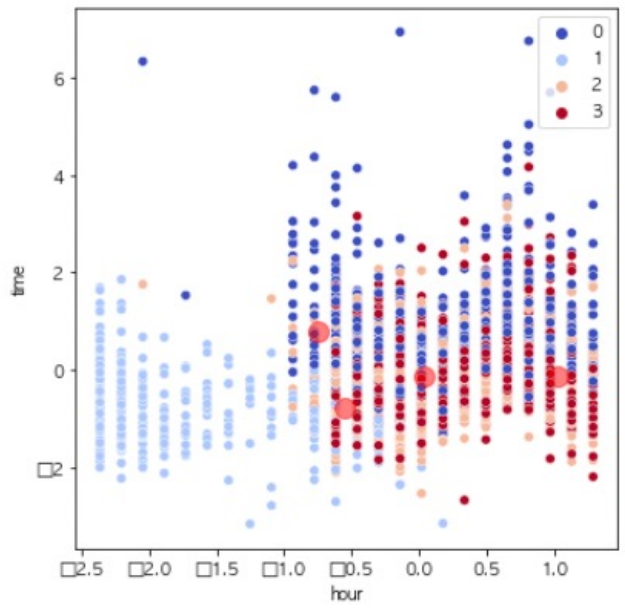
Scaling 전

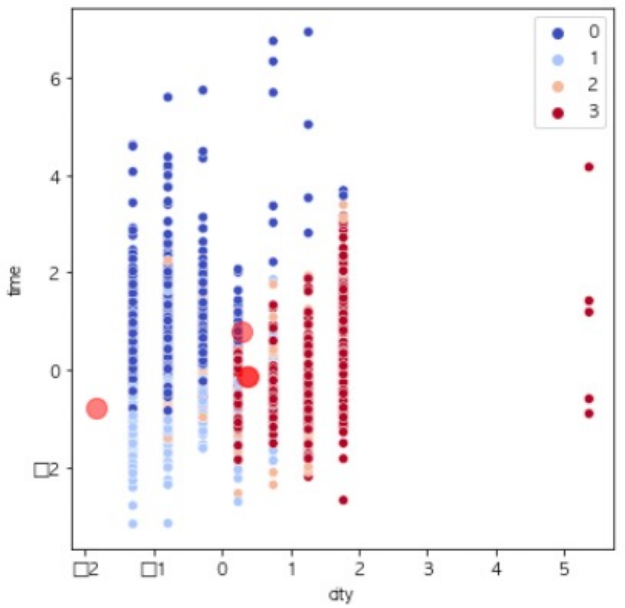


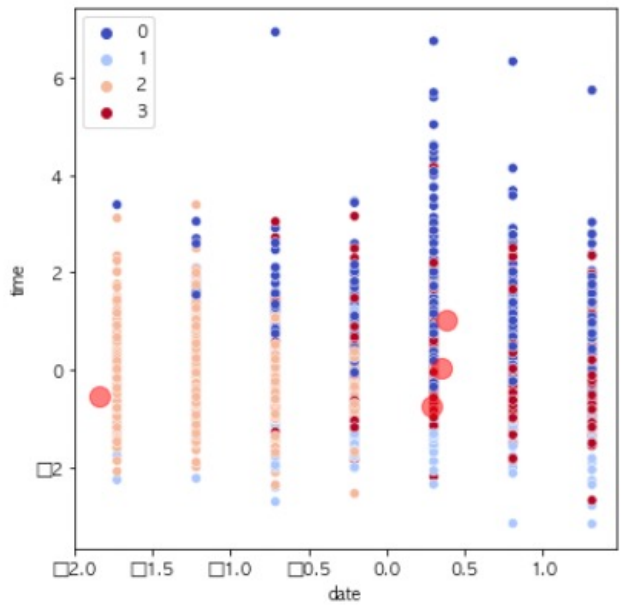




Scaling 후







목 차

Introduction

Data 분석 및 처리

ML Method

Conclusion

감사합니다!