데이터 시각화

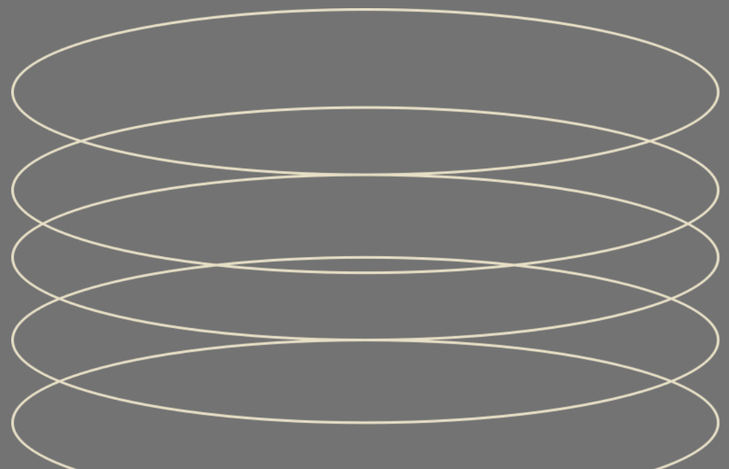
ch5.데이터수집,전처리,데이터 시각화

1조
고진영
감향임
하승아

python™

목차

1. 라이브러리



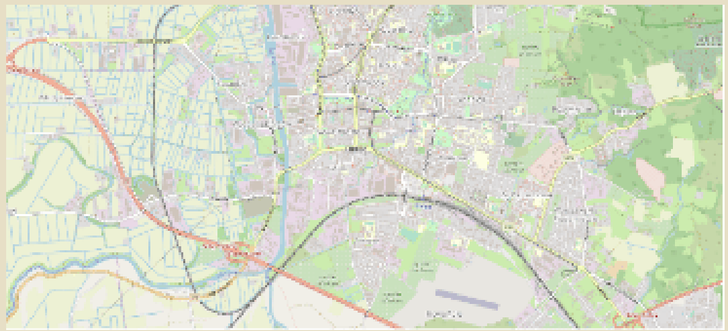
folium 이란?

지도 데이터에 위치 정보를 시각화하기 위한 라이브러리

- 파이썬의 강점인 데이터와 Leaflet.js 라이브러리의 매핑 강점을 토대로 제작

- 파이썬으로 데이터를 조작한다음 포리움을 통해 리플릿 맵에서 시각화

- 맵의 유형은 기본적으로 'Open Street Map'을 기반으로 동작

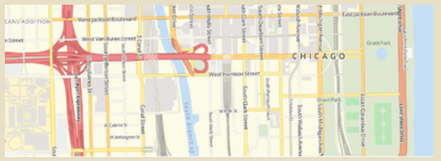


Open street Map

Stamen Terrain

stamen toner

Mapbox Bright



- 그 외에 'Stamen Terrain', 'Stamen Toner', 'Mapbox Bright', 와 'Mapbox Control room tiles' 형식을 내장

5.3.1 데이터 시각화

장점.

- 데이터를 python이라는 언어를 통해 쉽게 조작하고, folium 라이브러리를 호출해 leaflet.js 코드를 직접 짜지 않고 데이터를 시각화 할 수 있다.

단점.

- 생성된 leaflet 코드를 수정 하기 어렵다.
- 만약 웹 작업을 한다면, python 코드 실행 시 매번 새로 html코드를 만들기 때문에 잘못 하다 기존 html 작업본이 날아갈 수 있다.

- Folium은 방대한 양의 데이터를 시각화 하는 용도로만 사용 하는 것이 적합하다.

5.3.1 데이터 시각화

[포리움 함수를 통해 지도에 그림 그려보는 코드]

code: foliumTest.py

```
01  import folium
02
03  latitude = 37.566345
04  longitude = 126.977893
05
06  map_osm = folium.Map(location=[latitude, longitude])
07  map_osm.save('c:/imsi/map1.html')
08  print(type(map_osm)) # <class 'folium.folium.Map'> 객체
09
10  map_osm = folium.Map(location=[latitude, longitude], \
11                       zoom_start=16)
12  map_osm.save('c:/imsi/map2.html')
13
14  map_osm = folium.Map(location=[latitude, longitude], \
15              zoom_start=17, tiles='Stamen Terrain') #'Stamen Terrain'
```

위도 경도

06. map_osm = folium.Map(location = [latitude , longitude])

07. c드라이브의 imsi 폴더 map1.html 파일

10. 세부조건 - (zoon_start)초기 화면의 크기,숫자가 커질수록 줌인

matplotlib 이란?

파이썬에서 그래프를 그려주는 라이브러리

- Pip 명령어 (pip install matplotlib)을 이용하여 설치

- Line plor, bar chart, histogram, box Ploat, ccatter plot 등의 다양한 차트와 플롯 스타일을 지원.

<맷플롯립 사용용도>

데이터 분석 이전에 데이터 이해를 위한 시각화나
데이터 분석 후에 결과를 시각화 하기 위해서 사용

5.3.1 데이터 시각화

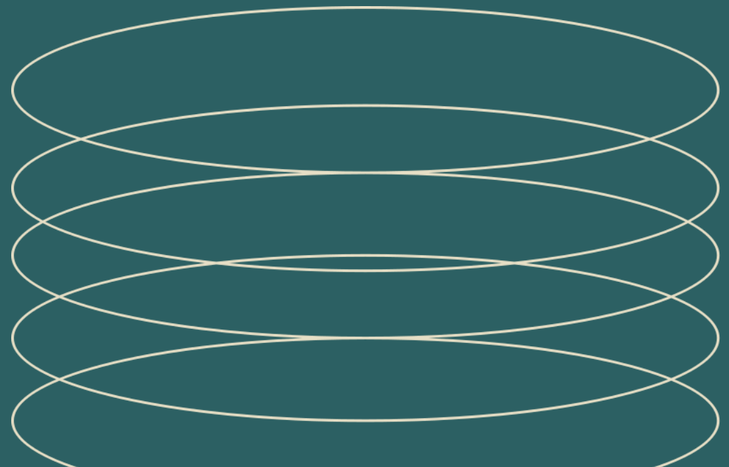
장점.

- 그래프의 세부적인 부분까지 일일이 세부 조정이 가능하기 때문에, 그래프를 한번에 다 그리지 않고 필요한 영역을 수정해 가면서 그릴 수 있다.

단점.

- 현재에는 맷플롯립을 사용하는 것보다 seaborn(분석에 더 적합함)이나 plotly(D3 기반의 대화형 시각화 그래프를 그릴 수 있음-> interactive함) 같은 패키지를 이용하면 분석에 더 적합하거나 혹은 인터랙티브가 가능한 그래프를 그릴 수 있다.

2. 지오코딩



3. 지오 코딩

- (geocoding) 주소나 산, 호수의 이름 등 고유 명칭을 가지고 위도와 경도의 좌푯값을 얻는 것
- 반대로, Reverse Geocording이라고 하면, 위도와 경도 값으로부터 주소를 얻는 것을 말함

<지오코딩 사용방법>

1. 구글 스프레드 시트 부가기능 이용
2. 파이썬 Geopy 패키지 이용 -> 파이썬 Geopy라이브러리는 pip 설치 후 사용
   개발자가 직접 구현하기보다는 네이버, 카카오 API 등을 사용하면 쉽게 구할 수 있음

3. 구글맵 API, 카카오맵 API 등 지도 서비스 이용

1)카카오 맵 장/단점 : 장점 – 갱신이 빠르고 국내 교통이 잘 시각화 되어 있음. / 단점 – 아시아권을 제외하고 해외가 잘 보이지 않는다.

2)구글맵 장/단점 : 장점 – 전세계를 전부 볼 수 있다. / 단점- 국내 갱신이 상대적으로 느리며 세부적으로 보여지고 있진 않다

-> API는 구글 클라우드 플랫폼과 같이 지도와 관련된 서비스를 이용하여 key 등록 후 라이브러리를 설치하고 import해서 사용

3. 지오 코딩

code: getGeocoderApi03.py

```
01  import folium, requests
02
03  address = '서울 마포구 신수동 451번지 세왕청마루아파트 상가 101호'
04  url = 'https://dapi.kakao.com/v2/local/search/address.json?query=' +
    address
05
06  api_key = '인증키 입력'
07  header = {'Authorization': 'KakaoAK ' + api_key}
08
09  def getGoocoder(address):
```

03. 찾고 싶은 주소지 정보 입력

06. 카카오 개발자 사이트에 접속하여서 인증키를 받아서 입력

```
10      result = ""
11      r = requests.get(url, headers=header)
12
13      if r.status_code == 200:
14          try:
15              result_address = r.json()["documents"][0]["address"]
16              result = result_address["y"], result_address["x"]
17          except Exception as err:
18              return None
19      else:
20          result = "ERROR[" + str(r.status_code) + "]"
21
22      return result
23
24  address_latlng = getGoocoder(address)
25  latitude = address_latlng[0]
26  longitude = address_latlng[1]
27
28  print('주소지 :', address)
29  print('위도 :', latitude)
30  print('경도 :', longitude)
31
32  shopinfo = '고촌 신수점'
33  foli_map = folium.Map(location=[latitude, longitude], zoom_start=17)
34  myicon = folium.Icon(color='red', icon='info-sign')
35  folium.Marker([latitude, longitude], popup=shopinfo, \
36                icon=myicon).add_to(foli_map)
37
38  folium.CircleMarker([latitude, longitude], radius=300, color='blue', \
39              fill_color='red', fill=False, popup=shopinfo).add_to(foli_
    map)
40
41  foli_map.save('c:/imsi/my_map_graph.html')
42  print('파일 저장 완료')
```

[주소지를 통해 위도 경도로 바꾸고 지도에 그림]

3. 지오 코딩 실행 결과

실행결과

주소지 : 서울 마포구 신수동 451번지 세양청마루아파트 상가 101호
위도 : 37.5477899311394
경도 : 126.93671586157
파일 저장 완료

' C:/imsi/my_map_graph.html'에 저장된 HTML문서

3. 데이터 그룹화

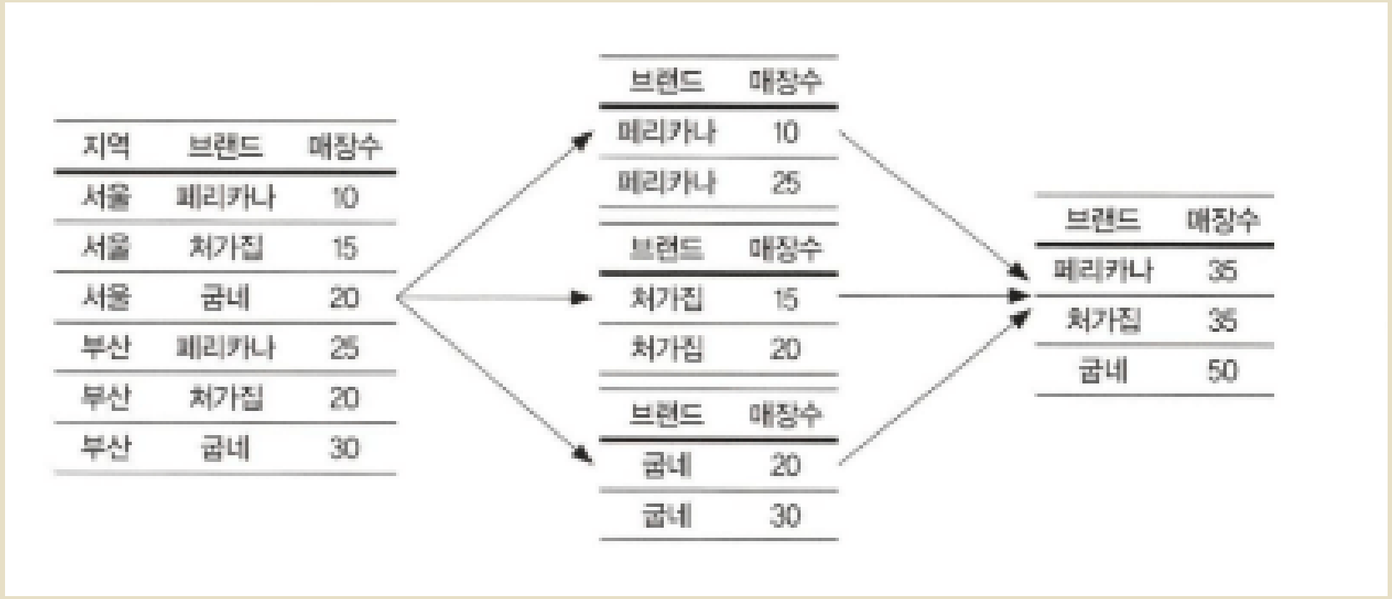
데이터 그룹화란?

객체를 분리하고, 함수를 적용하여 최종 결과를 조합하는 과정
= “끼리끼리 모아 틀에 놓는 것”

파이썬에서는 데이터를 그룹화하기 위해서 주로 groupby()함수를 사용함
groupby함수를 사용하면 카테고리에 따른 데이터를 그룹화하여 반환함.

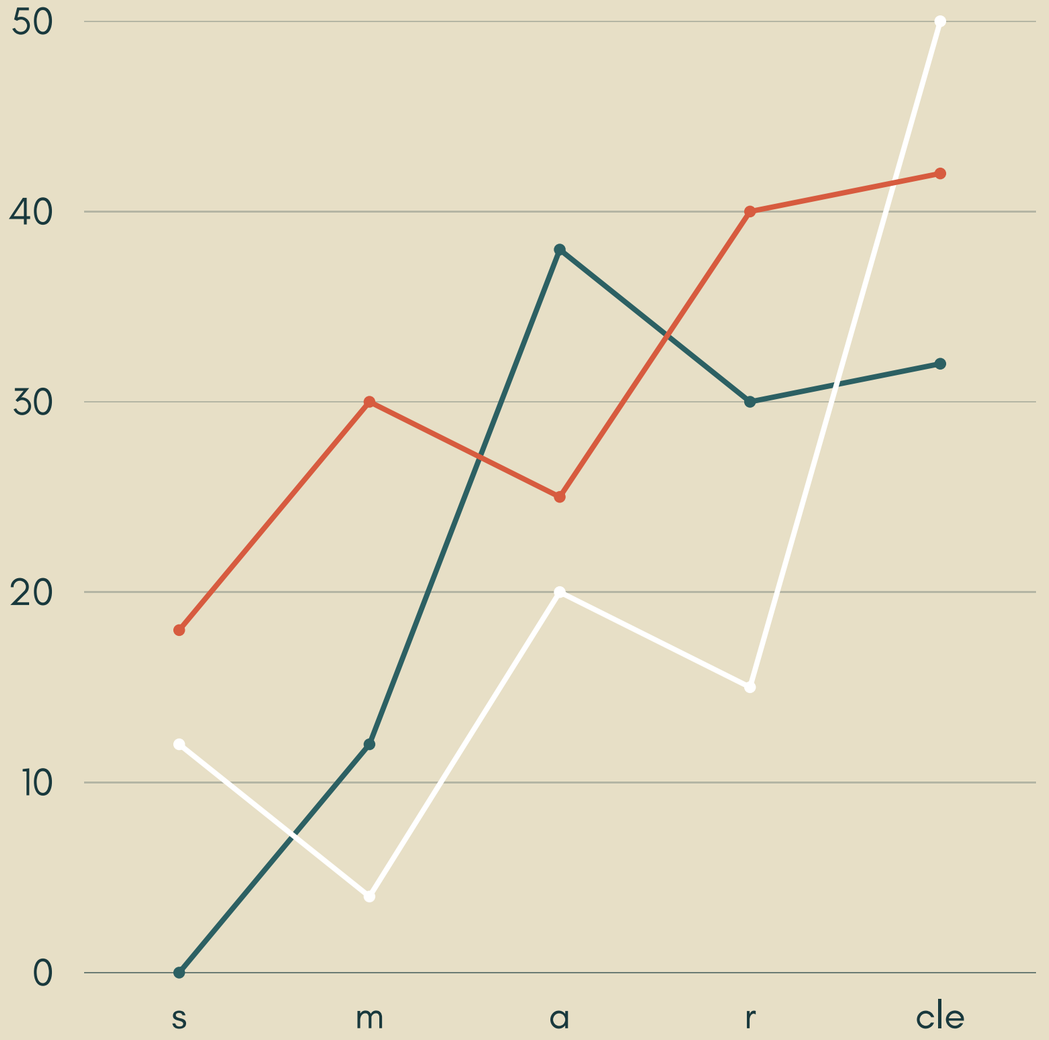
그룹화를 하는 이유

간단한 집계가 아닌 조건부 집계가 하고싶은 경우 사용'하며 Key값을 기준으로 그룹을 효율적으로 묶을 수 있음.



-> 다음 사진과 같이, 브랜드 별로 묶는 집계가 필요할 때 사용

4. 그래프 그리기



데이터 그룹화, 그래프 그리기

code: makeChickenGraph.py

```
01  import matplotlib.pyplot as plt
02  import pandas as pd
03
04  plt.rcParams['font.family'] = 'Malgun Gothic'
05
06  csv_file = 'allStoreModified.csv'
07  myframe = pd.read_csv(csv_file, index_col=0, encoding='utf-8')
```

01. matplotlib 선언

```
12  mycolor = ['r', 'g', 'b', 'm']
13  brand_dict = {'cheogajip':'처가집', 'goobne':'굽네', 'kyochon':'교촌',
    'pelicana':'페리카나', 'nene':'네네'}
```

12. 그래프에 사용할 색상 설정

데이터 그룹화, 그래프 그리기

```
15  mygrouping=myframe.groupby(['brand'])['brand']
16  chartData=mygrouping.count()
17
18  newindex=[brand_dict[idx] for idx in chartData.index]
19  chartData.index=newindex
20  print(chartData)
```

(아래 사진에 brand 확인할 수 있음)

	brand	store	sido	gungu	address	phone
0	cheogajip	장성점	전라남도	장성군	전라남도 장성군 장성읍 영천로 133-2	061-393-9289
1	cheogajip	신사점	서울특별시	은평구	서울특별시 은평구 신사동 40-6	02-304-7770
2	cheogajip	중곡역점	서울특별시	광진구	서울특별시 광진구 긴고랑로 5, 1층(중곡동)	02-3409-8292
3	cheogajip	응암점	서울특별시	은평구	서울특별시 은평구 백련산로 36, 상가동 106호(응암...	02-303-8295
4	cheogajip	돈암점	서울특별시	성북구	서울특별시 성북구 아리랑로6길 4, 1층(동선동5가)	02-6489-0101
5	cheogajip	거여점	서울특별시	송파구	서울특별시 송파구 거마로7길 3, 1층 101호(거여동)	02-3402-1511
6	cheogajip	밀양내이점	경상남도	밀양시	경상남도 밀양시 묵성로4길 16	055-356-9989
7	cheogajip	남지점	경상남도	창녕군	경상남도 창녕군 남지읍 남지중앙로 90	055-536-7333
8	cheogajip	신호동점	부산광역시	강서구	부산광역시 강서구 신호산단3로 66 1층 101호	051-831-0318
9	cheogajip	율하2지구점	경상남도	김해시	경상남도 김해시 율하로 479-1, 331동 1층 104호(장...	055-327-1811
10	cheogajip	남부동점	경상남도	양산시	경상남도 양산시 남부동 605-1번지 106호	055-383-9133
11	cheogajip	세곡점	서울특별시	강남구	서울특별시 강남구 헌릉로571길 7, 1층 103호(세곡...	02-451-8989
12	cheogajip	오금점	서울특별시	송파구	서울특별시 송파구 마천로 165, 1층(오금동)	02-407-1257
13	cheogajip	구산점	경상남도	김해시	경상남도 김해시 구산동 1083-7번지 101호	055-336-4366
14	cheogajip	예산점	충청남도	예산군	충청남도 예산군 예산읍 벚꽃로 155번길 43	041-335-7277
15	cheogajip	신창점	광주광역시	광산구	광주광역시 광산구 신창로36번길 19, 1층(신창동)	062-961-9282
16	cheogajip	장덕점	광주광역시	광산구	광주광역시 광산구 장덕로39번길 14, 1층(장덕동)	062-961-9281
17	cheogajip	학동방림점	광주광역시	동구	광주광역시 동구 천변우로 603, 102-101(학동)	062-233-9280
18	cheogajip	은남점	광주광역시	광산구	광주광역시 광산구 목련로 273번안길 43, 2층(은남동)	062-716-9282
19	cheogajip	전대점	광주광역시	북구	광주광역시 북구 우치로 100번길 3-2, 1층(용봉동)	062-262-9280

15. 그룹화 하기

groupby 라는 코드를 사용하면 같은 값을 하나로 묶어
통계 또는 집계 결과를 나타냄
해당 코드에서는 brand 별로 값을 묶었음

20. 그룹화 한 결과 출력

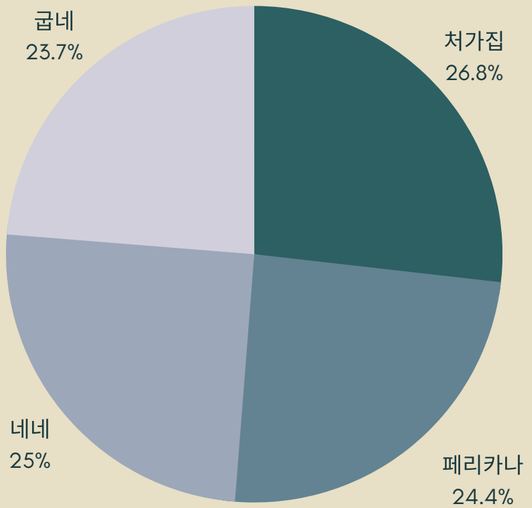
브랜드별 매장 개수

브랜드	개수
처가집	1204
굽네	1066
네네	1125
페리카나	1098

파이 그래프 그리기

```
22  plt.figure()
23  chartData.plot(kind='pie', legend=False, autopct='%1.2f%%',
    colors=mycolor)
24  filename='makeChickenGraph01.png'
25  plt.savefig(filename, dpi=400, bbox_inches='tight')
26  print(filename + ' 파일이 저장되었습니다.')
```

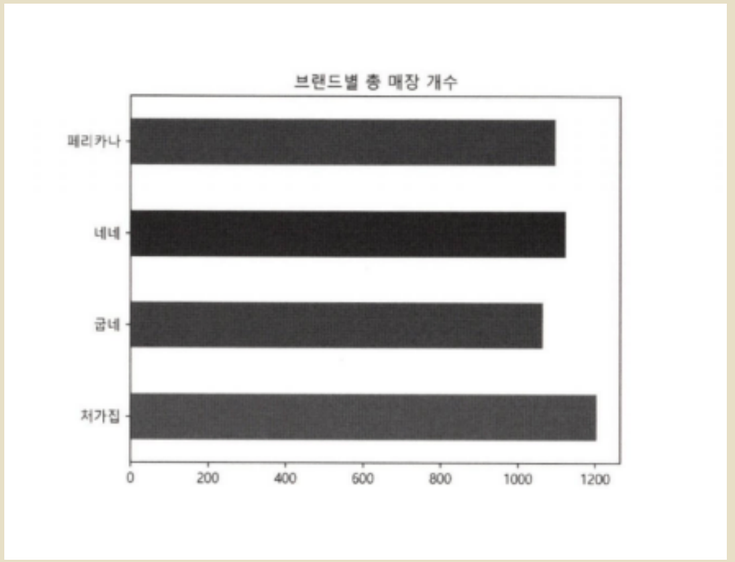
kind : 파이 그래프 선택
legend : 범례를 추가하는 옵션
autopct : 파이 조각의 전체 대비 백분율 표시
colors : 파이 조각의 색상 표현



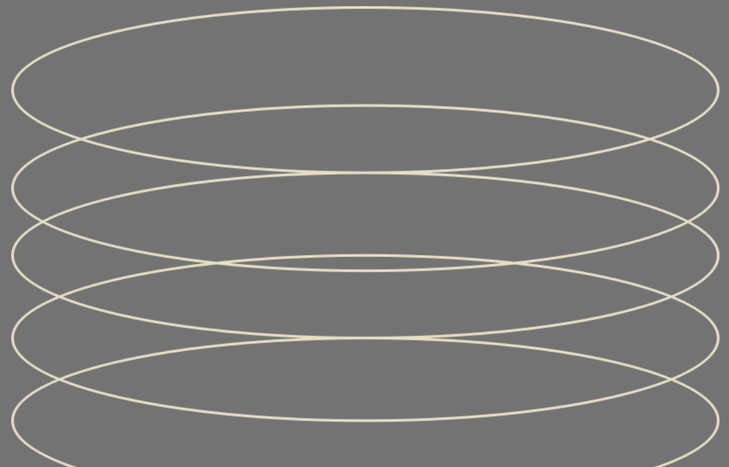
막대 그래프 그리기

```
28  plt.figure()
29
30  chartData.plot(kind='barh', rot=0, title='브랜드별 총 매장 개수',
    legend=False, color=mycolor)
31  filename='makeChickenGraph02.png'
32  plt.savefig(filename, dpi=400, bbox_inches='tight')
33  print(filename + ' 파일이 저장되었습니다.')
```

kind : 차트의 유형(barh – 가로 막대)
rot : 눈금 이름을 회전시킬 각도를 지정(0~360)
title : 그래프 제목을 문자열로 지정



+ 데이터 시각화



데이터 시각화

데이터 시각화란?

차트, 그래프 또는 맵과 같은 시각적 요소를
사용해 데이터를 표시하는 프로세스

데이터 시각화의 이점

1. 전략적 의사 결정가능
(빠르고 효과적인 의사결정을 내리기 가능)

패턴을 식별하고 추세를 찾고 인사이트를 얻어서
경쟁에서 앞서 나갈 수 있음

데이터 시각화가 중요한 이유

기업이 처리해야하는 데이터 소스
1. 내부 및 외부 웹 사이트
2. 스마트 디바이스
3. 내부 데이터 수집 시스템
4. 소셜 미디어

이런 원시 데이터는 이해하고 사용하기 어려울 수 있음.
따라서 데이터 간의 관계를 식별하고 숨겨진 패턴이나 추세를 감지
할 수 있도록 데이터에 시각적 형태를 부여하는 기능을 함.

<파이썬을 이용하여 데이터분석을 하는 이유>

1. 파이썬이란 언어의 직관성

C언어와 파이썬이 모두 하이레벨 언어에 속함에도 불구하고 그 중에서 가장 배우기 쉬운 언어를 고르자면 파이썬이다.

2. 대량의 데이터를 처리하는 속도

웹사이트의 방문 로그처럼 지속적으로 발생하는 대량의 데이터를 분석하거나 웹에서 크롤링해오는 대량의 데이터를 처리하기에는 파이썬이 제격이다.

수식 계산과 함수 같이 기본적으로 필요한 도구들을 제공함과 동시에 대량의 데이터를 빠르게 처리할 수 있고 다양한 분석 및 예측 관련 라이브러리가 지원된다.

3. 다양한 라이브러리 사용가능

데이터를 분석하기위해 Pandas 등의 라이브러리를 제공해 줌

<데이터 분석 스터디를 하는 이유>

1. 분석 실무 역량 제고

딱딱한 프로그래밍 언어로써의 파이썬이 아니라 분석 실무에서 접하는 다양한 문제를 이해하고 직접 해결해보는 경험을 통해 실무에서 파이썬 사용 역량을 제고하는 기회를 만들어준다.

2. 프로세스 이해

실제 분석을 할 경우에 대부분의 시간을 전처리와 데이터 이해/탐색에 대부분의 시간을 사용한다는 것을 체감하고 실제 실무에서 분석 프로세스를 이해할 수 있다.

3. 소프트 스킬의 중요성 이해

분석 스터디를 통해서 혼자 공부하는 것에서 그치지 않고 함께 소통하며 분석자로써의 능력을 기를 뿐만 아니라 마인드셋이나 태도, 원칙등도 자연스럽게 취득할 수 있다.

감사합니다!