

2019.09.24

데이터 시각화 Data Visualization

5기 최보경

CONTENTS

1. 시각화란

2. 실전에선 어떤 툴을 쓰나

3. EDA 가이드라인

4. Cheat Sheet

5. Interactive EDA

6. 지도 시각화 (Folium)

PPT가 왜 세로예요?

- 주피터 노트북에 마크다운과 시각 자료를 적절히 섞었습니다.

오늘은 강의안과 실습 코드창을 함께 띄워주세요.

- 코드에서 쏟아지는 정보에 헤매지 않고 가이드라인을 잘 따라오면서 절차를 익혔으면 합니다.

[이렇게 준비됐나요?]

The left screen shows a presentation titled "Guidelines for Exploratory Data Analysis" with the following table of contents:

- 1. 데이터프레임 확인
 - 1. 데이터 읽어 오기
 - 2. 데이터셋 shape 파악
 - 3. 데이터의 통계량 파악
 - 4. 결측치 처리
 - 5. 변수 타입 파악과 변환
- 2. 변수에 대한 이해
 - 1. 변수 이름, 타입 파악
 - 2. 변수를 Segmentation
 - 3. 각 변수에 대해 궁금한 정보 생각
 - 4. 서로 영향을 줄 변수들에 대한 기대 가설
- 3. 단일 변수 분석
 - 1. 맞추고자 하는 타겟값 y부터 분석
 - 1. 통계량 파악과 해석
 - 2. 분포(강한 위주) 파악
 - 3. 이상치 제거
 - 4. 분포(강한 위주) 파악
 - 2. 범주형 변수의 빈도 파악
 - 1. 범주형 변수가 10개 이하일 때
 - 1. 빈도표
 - 2. 범주형 변수가 10개 이상일 때
 - 1. 데이터 상위 또는 하위로 자르기
 - 2. 빈도표
- 4. 이진 변수 분석
 - 1. 연속형 X 연속형
 - 2. 연속형 X 범주형

The right screen shows a Jupyter Notebook titled "how to taste folium visualization (autosaved)". It contains the following code and output:

```
In [13]: from folium.plugins import HeatMap
HeatMap(data=df_cut[['start_station_latitude', 'start_station_longitude']].groupby(
radius=8, max_zoom=11).add_to(base_map))

Out[13]: <folium.plugins.heat_map.HeatMap at 0x1c484775a20>

In [14]: base_map
Out[14]:
```

The output shows a map of New Jersey with a heatmap overlay. The heatmap shows a high concentration of data points in the central part of the state, specifically around the New York City area. The map is labeled with various cities and regions, including Newark, Jersey City, Hoboken, and Bayonne. The heatmap is titled "사용자 지정 SQL 쿼리 b2b vendor".

1. 시각화란

- 설득과 사실확인을 목적으로 한다 .
- 맥락과 상황에 맞는 데이터를 제공하되 과도한 정보를 제공해서는 안된다
- 데이터 애널리스트라면 커뮤니케이션과 함께 요구되는 능력이다
- 보는 대상이 누구인가?
 - 직접 그래프를 더 탐색해보고 싶은 목적이 있는 대상
: 동적인(interactive) 시각화 방식
 - 그렇지 않고, 보고 싶은 정보가 비교적 명확한 대상
: 정적인 그래프 기반으로 작성한 후, 추가적인 내용이 궁금하면 대시보드로 가도록 유도

정보 시각화 방법				
시간 시각화	분포 시각화	관계 시각화	비교 시각화	공간 시각화
막대그래프 (Bar graph) 누적 막대그래프 (Stacked Bar graph) 점그래프 (Point graph)	파이차트 (Pie chart) 도넛 차트 (Donut chart) 트리맵 (Tree map) 누적 연속 그래프 (Cumulative continuous graph)	스캐터 플롯 (Scatter plot) 버블 차트 (Bubble chart) 히스토그램 (Histogram)	히트맵 (heat map) 체르노프 페이스 (Chernoff face) 별그래프 (Star graph) 평행 좌표계 (Parallel coordinate system) 다차원 척도법 (Multi-dimensional scaling)	지도 매핑 (Data viz on map)

2. 실전에서는 어떤 툴을 쓰나

- 가장 많이 쓰는 것 ([대시보드](#) 생성용)

Tableau

Google spreadsheet

[Zeppelin](#)

- 심화 데이터 전처리 & 모델링과 석을 때

Python

- Matplotlib

- Seaborn

- Folium

- Pyecharts

- Plotly

R (ggplot) 또는 SQL

- 요즘 뜨는 신박한 것

Interactive EDA

- lpywidget

- Widget (Colabotory notebook의 경우)

3. EDA 실습 전

[실습 데이터 SCHEMA]

nyc_citibike.csv

2018년 5월 1일 하루 일자의 뉴욕 bike 대여 기록

- Start_date: 대여 시작 일자
- End_date: 대여 종료 일자
- trip_duration: 주행 시간 (초 단위)
- Start_hour: 대여 시작 시간대
- End_hour: 대여 종료 시간대
- Start/End station_id: 대여 시작/종료 역 ID
- Start/End station_name: 대여 시작/종료 역 이름
- Start/End latitude: 대여 시작/종료 역 위도
- Start/End longitude: 대여 시작/종료 역 경도
- Bike_id: 바이크 기기별 고유 ID
- User_type:
 - Customer = 24 hour pass or 7 day pass user
 - Subscriber = Annual member
- Birth_year: 출생연도
- Gender:
 - Unknown
 - Male
 - Female
- Day_since_register: 가입한 후로 해당 바이크 예약 건
까지 지난 일 수

3. EDA 실습 전

What is EDA?

▪ Exploratory Data Analysis (탐색적 데이터 분석)

- 기본 도구는 도표(plot), 그래프(graph), 요약 통계(summary statistics)
 - 모든 변수의 분포를 도표화하고, 시계열 데이터를 도표화하며, 변수를 변환하고, 산점도 행렬을 이용하여 변수들의 대응 관계를 파악하며, 모든 변수의 요약 통계를 생성하는 등
 - 데이터를 체계적으로 둘러보는 하나의 방법
- EDA와 Data Visualization은 사실 다르다
EDA는 연구의 초기 단계에서 이루어지고, 데이터 시각화는 분석 결과를 커뮤니케이션 하기위해 연구의 마지막 단계에서 행해진다.
- EDA에서 얻은 이해는 알고리즘의 발전을 알려 주고 향상 시키는 데에 사용할 수 있다.

3. EDA 가이드라인 [전체]

[참고]

주어진 데이터셋이 있을 때, 이를 파헤쳐야 하는데 그 방식이 막막할 때 이 순서를 따를 수 있다. 하지만 데이터의 도메인과 시각화 목적을 고려해서 필요한 부분만 사용하거나 순서를 조정할 것을 권장. 더 좋은 시각화 방식과 개인에게 맞는 툴을 여러 구글링과 경험을 통해 각자만의 가이드라인을 세웠으면 좋겠다.

1. 데이터프레임 확인

1. 데이터 읽어 오기
2. 데이터셋 shape 파악
3. 데이터셋 통계량 파악
4. 결측치 처리
5. 변수 타입 파악과 변환

2. 데이터 도메인과 변수 이해

1. 변수 이름, 타입 파악
2. 변수들 Segmentation
3. 각 변수에 대해 궁금한 정보 생각
4. 서로 영향을 줄 변수들에 대한 기대 가설

3. 단일 변수 분석

1. (맞추고자 하는 타겟값 y부터 분석) 연속형 변수의 분포 파악
 1. 통계량 파악
 2. 분포(경향 위주) 파악
 3. 이상치 제거
 4. 분포(경향 위주) 파악
2. 범주형 변수의 빈도 파악
 1. 범주형 변수가 30개 이하일 때
 1. 빈도표
 2. 범주형 변수가 30개 이상일 때
 1. 데이터 상위 또는 하위로 자르기
 2. 빈도표

4. 이진 변수 분석

1. 연속형 X 연속형
 1. Scatterplot
 1. Pandas Visualization
 2. Scatterplot with Regression Fit
 1. Seaborn
2. 범주형 X 범주형
 1. Vertical Countplot
 2. Horizontal Countplot
3. 범주형 X 연속형
 1. 범주형 변수가 10개 이하일 때
 1. Seaborn의 Boxplot
 2. Seaborn의 Catplot - Ex. Boxen
 2. 범주형 변수가 10개 이상일 때
 1. 데이터 상위 또는 하위로 자르기
 2. Horizontal Boxplot

그래프

분석 방법

연속형 X 연속형

- (추세선이 있는) Scatter plot

- Correlation 분석
(두 변수 간 상관관계 여부)

범주형 X 범주형

- 누적막대그래프
- 100%기준 누적 막대 그래프

- Chi-Square분석
(두 변수가 독립적인지 여부)

범주형 X 연속형

- 누적막대그래프
- 범주 별 Histogram

- 범주의 종류에 따라
 - 2개: T-test/Z-test
 - 3개 이상: ANOVA
(집단 별 평균 차가 유의한지 여부)

5. 3개 이상의 변수 분석

1. Bubble Scatterplot (버블차트)
2. Heatmap(히트맵)
 1. Groupby로 데이터 정렬
 2. Seaborn의 Heatmap

6. 상관관계 분석

1. Heatmap(히트맵)
 1. 전체 데이터셋 히트맵
 2. 각 변수별 상위 N개 히트맵

3. EDA 가이드라인 [1~3단계]

Session 03. EDA Guideline.ipynb

1. 데이터프레임 확인

1. 데이터 읽어 오기
2. 데이터셋 shape 파악
3. 데이터셋 통계량 파악
4. 결측치 처리
5. 변수 타입 파악과 변환

2. 데이터 도메인과 변수 이해

1. 변수 이름, 타입 파악
2. 변수들 Segmentation
3. 각 변수에 대해 궁금한 정보 생각
4. 서로 영향을 줄 변수들에 대한 기대 가설

3. 단일 변수 분석

1. (맞추고자 하는 타겟값 y 부터 분석) 연속형 변수의 분포 파악
 1. 통계량 파악
 2. 분포(경향 위주) 파악
 3. 이상치 제거
 4. 분포(경향 위주) 파악
2. 범주형 변수의 빈도 파악
 1. 범주형 변수가 30개 이하일 때
 1. 빈도표
 2. 범주형 변수가 30개 이상일 때
 1. 데이터 상위 또는 하위로 자르기
 2. 빈도표

3. EDA 가이드라인 [4~6단계]

4. 이진 변수 분석

1. 연속형 X 연속형

1. Scatterplot

1. Pandas Visualization

2. Scatterplot with Regression Fit

1. Seaborn

2. 범주형 X 범주형

1. Vertical Countplot

2. Horizontal Countplot

3. 범주형 X 연속형

1. 범주형 변수가 10개 이하일 때

1. Seaborn의 Boxplot

2. Seaborn의 Catplot - Ex. Boxen

2. 범주형 변수가 10개 이상일 때

1. 데이터 상위 또는 하위로 자르기

2. Horizontal Boxplot

그래프

분석 방법

연속형 X 연속형

- (주세선이 있는) Scatter plot

- Correlation 분석
(두 변수 간 상관관계 여부)

범주형 X 범주형

- 누적막대그래프
- 100%기준 누적 막대 그래프

- Chi-Square 분석
(두 변수가 독립적인지 여부)

범주형 X 연속형

- 누적막대그래프
- 범주 별 Histogram

- 범주의 종류에 따라
 - 2개: T-test/Z-test
 - 3개 이상: ANOVA
(집단 별 평균 차가 유의한지 여부)

5. 3개 이상의 변수 분석

1. Bubble Scatterplot (버블차트)

2. Heatmap(히트맵)

1. Groupby로 데이터 정렬

2. Seaborn의 Heatmap

6. 상관관계 분석

1. Heatmap(히트맵)

1. 전체 데이터셋 히트맵

2. 각 변수별 상위 N개 히트맵

방금 했는데 기억이 안나요

- 밥 먹고 파이썬으로 EDA만 하는 사람이 아니라면 당연히 파라미터를 까먹습니다.
- 기본만 익혀 두세요.

차트

Line Chart: `plt.plot(x, y, ...)` 색: color, 점 모양: marker, 선 모양: linestyle, 라벨: label ...

Bar Chart: `plt.bar(x, y, ...)` 색: color, 너비: width, 라벨: label ...

Histogram: `plt.hist(x, ...)` 구간: range, 구간 개수: bins, 스타일: histtype, 색: color ...

Scatter Plot: `plt.scatter(x, y, ...)` 색: color, 점 크기: size, 점 모양: marker, alpha: 투명도 ...

주요 함수

`plt.show()`: 설정된 차트를 display하는 함수

`plt.figure()`: 차트에 대한 각종 설정을 넣는 함수 차트 크기: `figsize`

`plt.subplot(n, m, x)`: $n*m$ 등분한 칸 중 x 번 째 칸에 차트를 그리겠다는 선언

`plt.title(x)`: x 를 타이틀로 넣음 / `plt.xlabel(x)`: x 를 횡축 라벨로 넣음 / `plt.ylabel(x)`: x 를 종축 라벨로 넣음

`plt.grid(True)`: 그리드를 넣음 / `plt.legend()`: 범례를 넣음

`plt.xticks(xs, labels)`: xs 가 나타내는 x 축 위치들에 `labels`의 원소들을 라벨로 넣음

`plt.annotate(label, (x, y))`: (x, y) 가 나타내는 위치에 `label`을 라벨로 넣음

- 가장 효율적인 키워드로 구글링해서 Documentation 또는 Reference 웹사이트들을 찾아서 바로 이해하고 써주는 것이 중요해요

Matplotlib

Line color

color / c

'C0' 'C1' 'C2' 'C3' 'C4' 'C5' 'C6' 'C7' 'C8' 'C9'

Line width

linewidth / lw

1 2 3 4 5

Cap style

'butt' 'round' 'projecting'

Dash cap style

'butt' 'round' 'projecting'

Line style

linestyle / ls

'--' '---' '---' '---' '---' (0, (0.01, 2))

Antialias

False True

Antialiased

Marker edge color

ec / ec

'C0' 'C1' 'C2' 'C3' 'C4' 'C5' 'C6' 'C7' 'C8' 'C9'

Marker face color

fc / fc

'C0' 'C1' 'C2' 'C3' 'C4' 'C5' 'C6' 'C7' 'C8' 'C9'

Marker edge width

ew / lw

1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5

Marker size

ms / s

25 50 75 100 125 150 175 200 225 250

Filled markers

marker

'o' 's' 'x' 'x' 'x' 'x' 'x' 'x' 'x' 'x'

Unfilled markers

marker

'o' 's' 'x' 'x' 'x' 'x' 'x' 'x' 'x' 'x'

Unicode markers

marker

's' 's' 's' 's' 's' 's' 's' 's' 's' 's'

Marker spacing

markervery

10 [0, -1] (25, 5) [0.25, -1]

Line collection

LineCollection

Circle collection

CircleCollection

Ellipse collection

EllipseCollection

Polygon collection

PolyCollection

Path collection

PathCollection

Regular polygon collection

RegularPolyCollection

Star collection

StarPolygonCollection

Asterisk collection

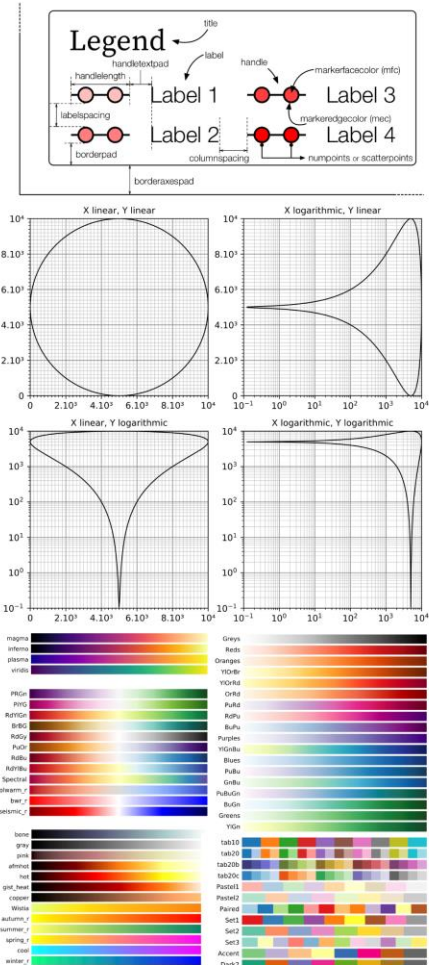
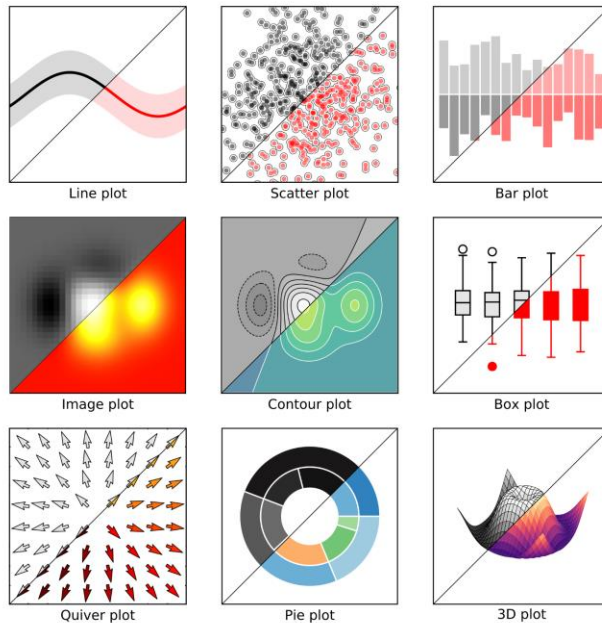
AsteriskPolygonCollection

Text styles

bold italic font size

10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92 94 96 98 100 102 104 106 108 110 112 114 116 118 120 122 124 126 128 130 132 134 136 138 140 142 144 146 148 150 152 154 156 158 160 162 164 166 168 170 172 174 176 178 180 182 184 186 188 190 192 194 196 198 200 202 204 206 208 210 212 214 216 218 220 222 224 226 228 230 232 234 236 238 240 242 244 246 248 250 252 254 256 258 260 262 264 266 268 270 272 274 276 278 280 282 284 286 288 290 292 294 296 298 300 302 304 306 308 310 312 314 316 318 320 322 324 326 328 330 332 334 336 338 340 342 344 346 348 350 352 354 356 358 360 362 364 366 368 370 372 374 376 378 380 382 384 386 388 390 392 394 396 398 400 402 404 406 408 410 412 414 416 418 420 422 424 426 428 430 432 434 436 438 440 442 444 446 448 450 452 454 456 458 460 462 464 466 468 470 472 474 476 478 480 482 484 486 488 490 492 494 496 498 500 502 504 506 508 510 512 514 516 518 520 522 524 526 528 530 532 534 536 538 540 542 544 546 548 550 552 554 556 558 560 562 564 566 568 570 572 574 576 578 580 582 584 586 588 590 592 594 596 598 600 602 604 606 608 610 612 614 616 618 620 622 624 626 628 630 632 634 636 638 640 642 644 646 648 650 652 654 656 658 660 662 664 666 668 670 672 674 676 678 680 682 684 686 688 690 692 694 696 698 700 702 704 706 708 710 712 714 716 718 720 722 724 726 728 730 732 734 736 738 740 742 744 746 748 750 752 754 756 758 760 762 764 766 768 770 772 774 776 778 780 782 784 786 788 790 792 794 796 798 800 802 804 806 808 810 812 814 816 818 820 822 824 826 828 830 832 834 836 838 840 842 844 846 848 850 852 854 856 858 860 862 864 866 868 870 872 874 876 878 880 882 884 886 888 890 892 894 896 898 900 902 904 906 908 910 912 914 916 918 920 922 924 926 928 930 932 934 936 938 940 942 944 946 948 950 952 954 956 958 960 962 964 966 968 970 972 974 976 978 980 982 984 986 988 990 992 994 996 998 1000

<https://github.com/rougier/matplotlib-cheatsheet>



Growth Hackers

4. Cheat Sheet – Seaborn

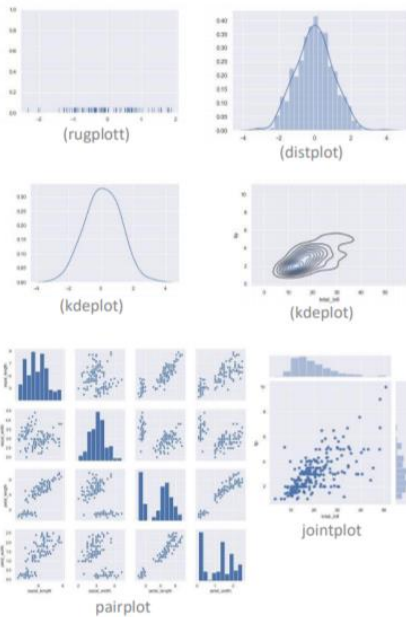


<http://www.interactivechaos.com>

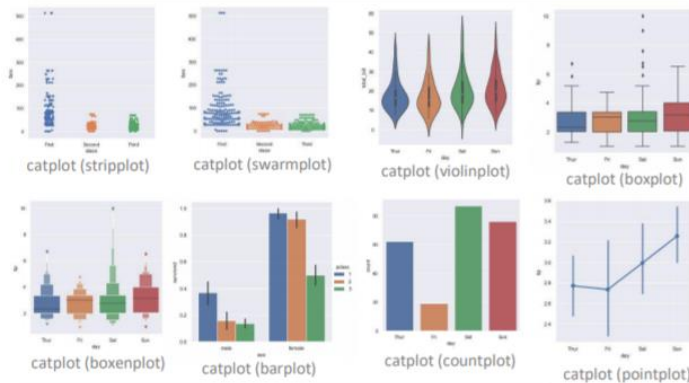
SEABORN CHEAT SHEET

Figure-level function (equivalent axes-level function)

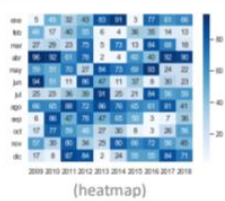
DISTRIBUTIONS



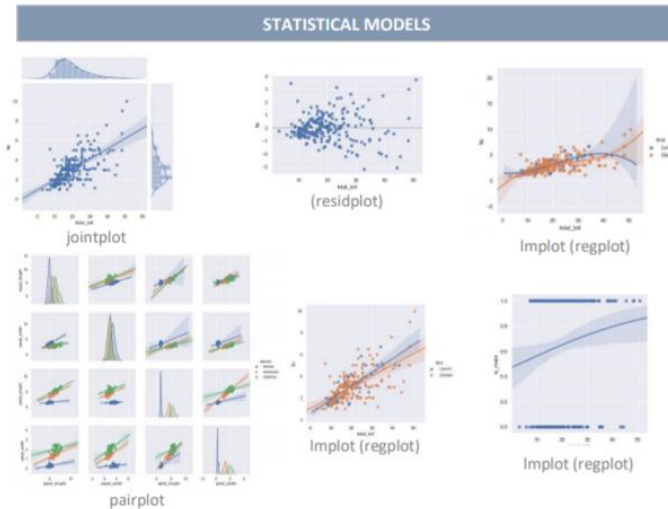
RELATIONSHIPS BETWEEN QUANTITATIVE AND QUALITATIVE VARIABLES



HEAT MAPS



RELATIONSHIPS BETWEEN QUANTITATIVE VARIABLES



https://www.interactivechaos.com/sites/default/files/data/seaborn_cheat_sheet.pdf

5. Interactive EDA 맛보기

함수와 함께 짜서 쓰는 Interactive (동적인) EDA 방식이 있습니다.

- 직접 그래프를 더 탐색해보고 싶은 목적이 있는 분들에게 보내는 그래프라면 추천
- 여러가지 y변수를 두고 이진 변수 관계 분석을 해야 할 때 추천
- 다만, 무겁고 느리다.
 - ✓ Jupyter Notebook에서는 Ipywidget, Plotly
 - ✓ Colab Notebook 에서는 Widget

[관련 링크]

<https://towardsdatascience.com/interactive-controls-for-jupyter-notebooks-f5c94829aee6>

[sample codes]

[https://nbviewer.jupyter.org/github/zzsza/TIL/blob/master/python/visualization\(cufflinks\).ipynb](https://nbviewer.jupyter.org/github/zzsza/TIL/blob/master/python/visualization(cufflinks).ipynb)

6. 지도 시각화 (Folium)

- Session 03. Folium Visualization.ipynb
- 간단하게 두 가지 기능만 사용해보자.
 1. Folium 설치
 2. 빈 캔버스 역할을 하는 지도 그리기
 3. 히트맵 그리기
 4. 상위 10개 지역 Marker 찍기
 5. 로컬에 html로 저장하기

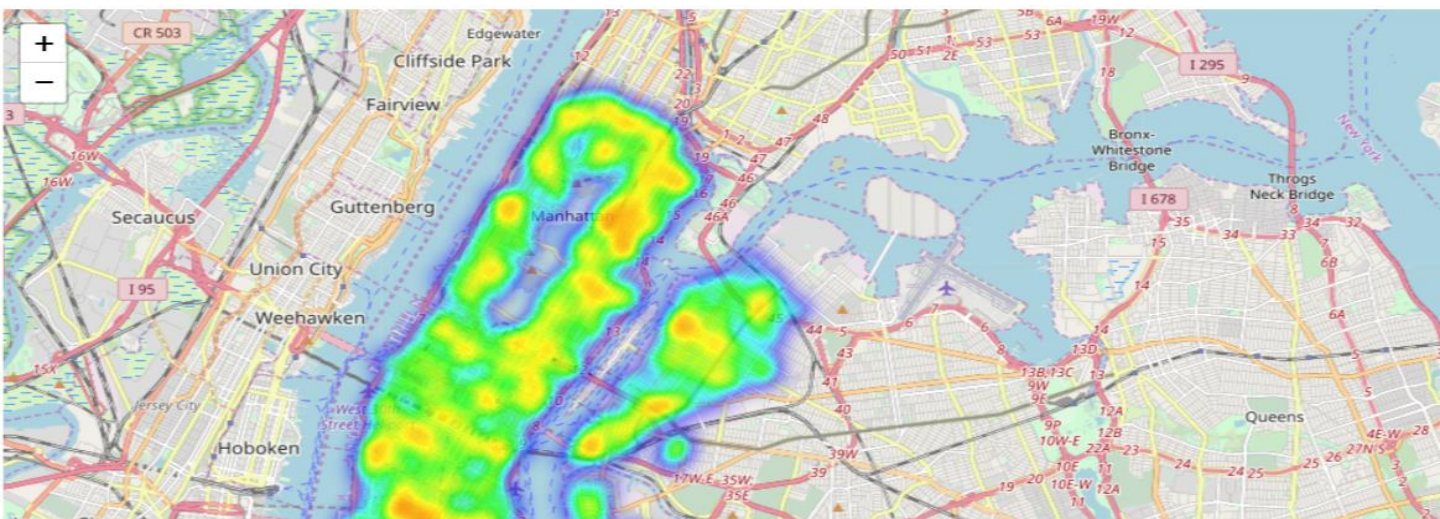
- 레퍼런스

folium 공식 깃허브

<https://github.com/python-visualization/folium>

folium 공식 documentation

<https://python-visualization.github.io/folium/>



퀘스트 정보

- Session03.Quest.ipynb 파일에 각 자리를 만들어 두었어요.
 - 실습에서 다룬 nyc_citibike.csv에서
 1. 연령대(ex.10,20,30,40대) 변수와 함께 어떤 변수와의 관계를 보면 좋을지 기대 가설을 세우고, 적절한 방식으로 시각화 후 해석 보태주세요.
 - 연령은 25세, 26세~ 아닌 20대의 BIN 형태 연령대 (범주형 변수)로 묶어주세요.
 2. Bike_id (바이크 하나하나에 붙어있는 고유 아이디) 에 따른 trip_duration을 시각화해주세요. 평균이어도 좋고, 누적이어도 좋습니다. 해석 보태 주세요.
 - Groupby 함수 사용해주시면 편합니다.
 3. Bike_id, Trip_duration, + 한 가지 변수 더 추가해서 3개 이상 변수 시각화 방식으로 시각화 해주세요. 해석은 안 보태 주셔도 됩니다.

레퍼런스

- <https://python-graph-gallery.com/272-map-a-color-to-bubble-plot/>
- https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html
- <https://seaborn.pydata.org/tutorial/categorical.html>
- <https://datascienceschool.net/view-notebook/d0b1637803754bb083b5722c9f2209d0/>
- <https://brunch.co.kr/@jjason68/12>
- <https://pinkwink.kr/984>