2019.09.24

데이터 시각화
Data Visualization

5기 최보경

Growth Hackers

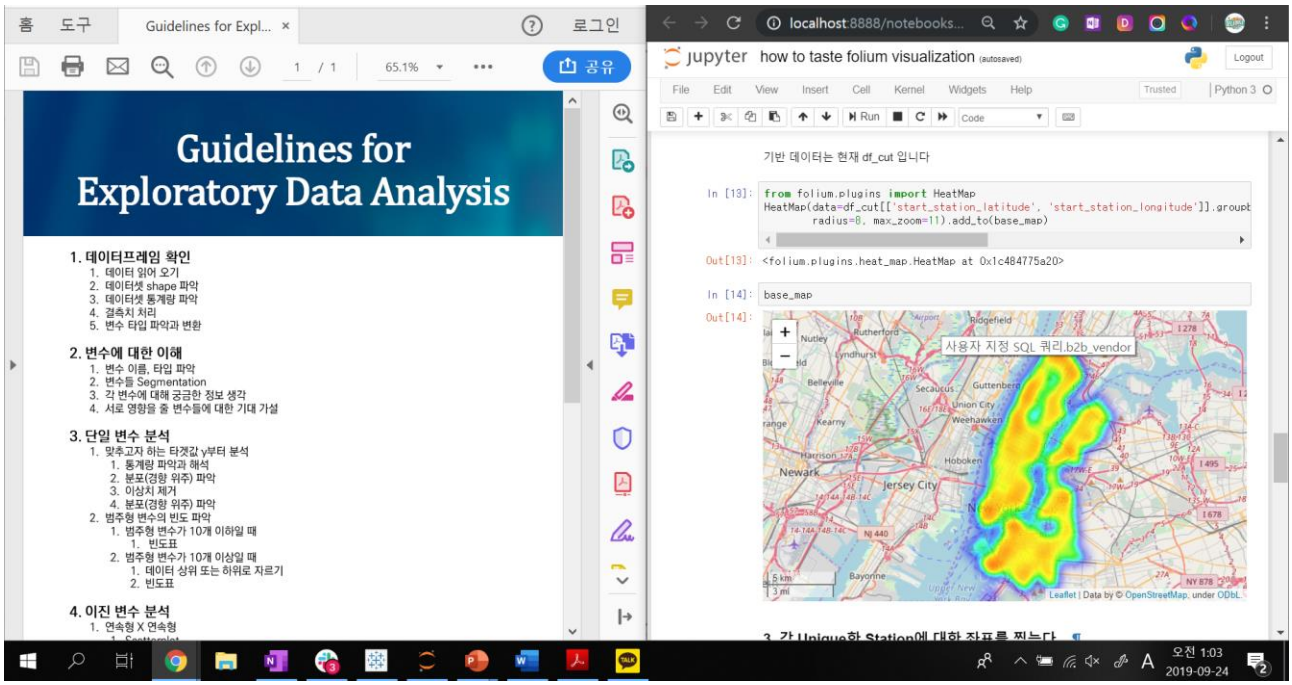
CONTENTS

PPT가 왜 세로에요?

- 쥬피터 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?]



Growth Hackers

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 (Dataviz on map)

Growth Hackers

2. 실전에서는 어떤 툴을 쓰나

- 가장 많이 쓰는 것 (대시보드 생성용)
  Tableau
  Google spreadsheet
  Zeppelin

- 심화 데이터 전처리 & 모델링과 섞을 때
  Python
  - Matplotlib
  - Seaborn
  - Folium
  - Pyecharts
  - Plotly

  R (ggplot) 또는 SQL

- 요즘 뜨는 신박한 것
  Interactive EDA
  - Ipywidget
  - Widget (Colabotory notebook의 경우)

Growth Hackers

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 건까지 지난 일 수

Growth Hackers

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.

Growth Hackers

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

Growth Hackers

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

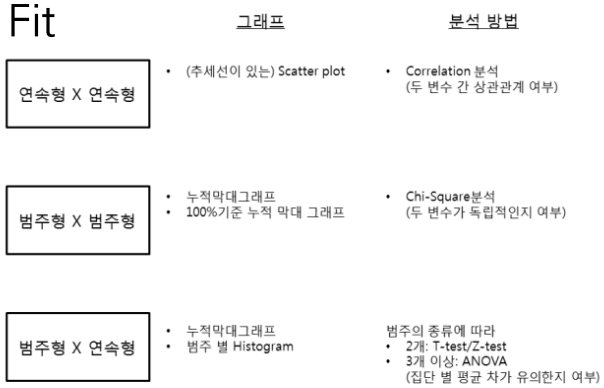
1.(맞추고자 하는 타겟값 y부터 분석) 연속형 변수의 분포 파악
    1.통계량 파악
    2.분포(경향 위주) 파악
    3.이상치 제거
    4.분포(경향 위주) 파악
2.범주형 변수의 빈도 파악
    1.범주형 변수가 30개 이하일 때
        1. 빈도표
    2.범주형 변수가 30개 이상일 때
        1.데이터 상위 또는 하위로 자르기
        2.빈도표

Growth Hackers

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



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

Growth Hackers

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 ...

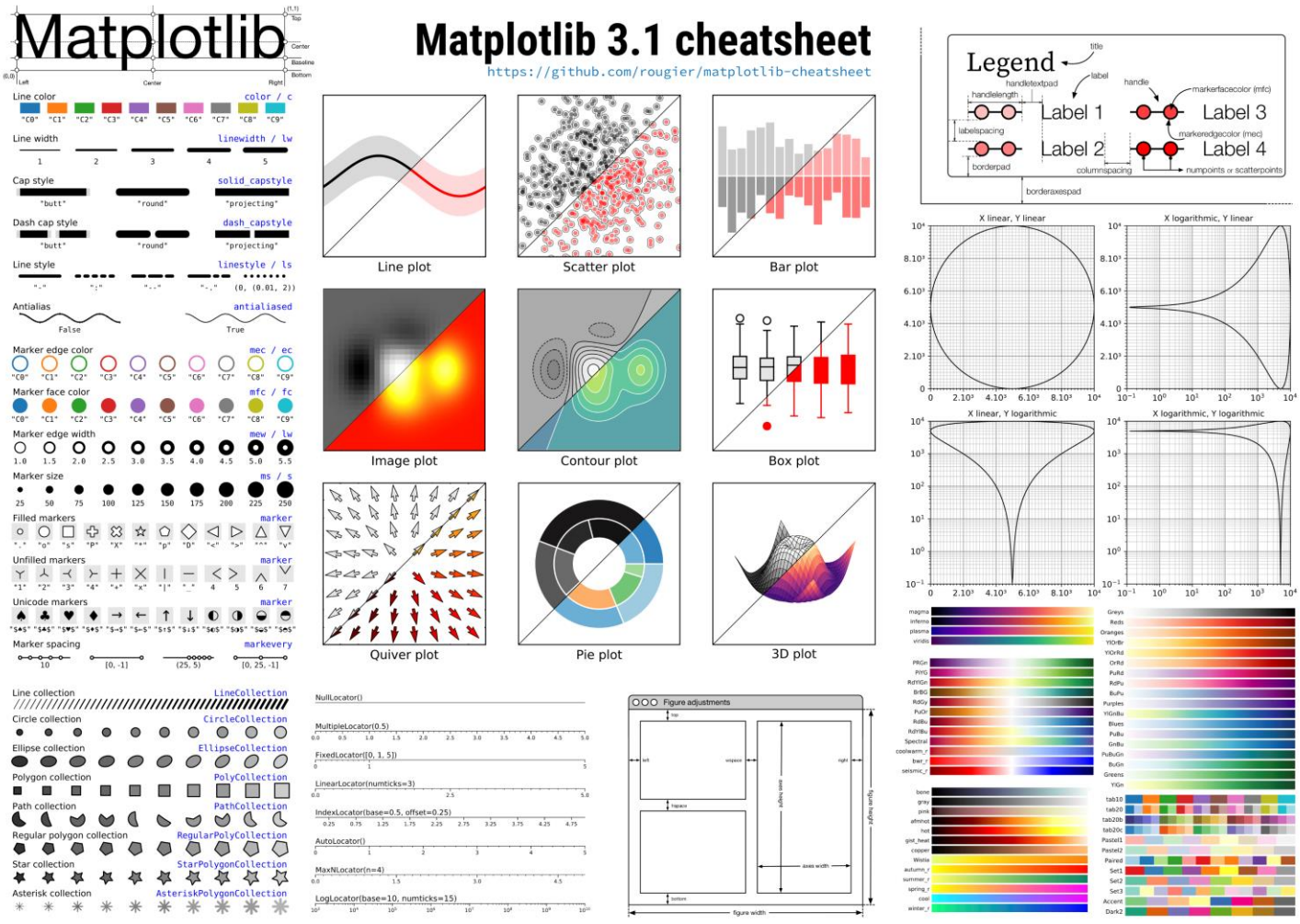
주요 함수

plt.show(): 설정된 차트를 display하는 함수
plt.figure(): 차트에 대한 각종 설정을 넣는 함수 차트 크기: figsize
plt.subplot(n, m, x): n*m등분한 칸 중 x번 째 칸에 차트를 그리겠다는 선언
plt.title(x): x를 타이틀로 넣음 / plt.xlabel(x): x를 횡축 라벨로 넣음 / plt.ylabel(x): x를 종축 라벨로 넣음
plt.grid(True): 그리드를 넣음 / plt.legend(): 범례를 넣음
plt.xtics(xs, labels): xs가 나타내는 x축 위치들에 labels의 원소들을 라벨로 넣음
plt.annotate(label, (x, y)): (x, y)가 나타내는 위치에 label을 라벨로 넣음

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

Growth Hackers

4. Cheat Sheet – Matplotlib



https://github.com/rougier/matplotlib-cheatsheet/blob/master/README.md?fbclid=IwAR2B3mqWisCzsCQbkWq13Kv5gZqXDDWnn5Yk1RnmdbUWOhIJy8xgxI7JL28

Growth Hackers

4. Cheat Sheet – Seaborn

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

Growth Hackers

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

Growth Hackers

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

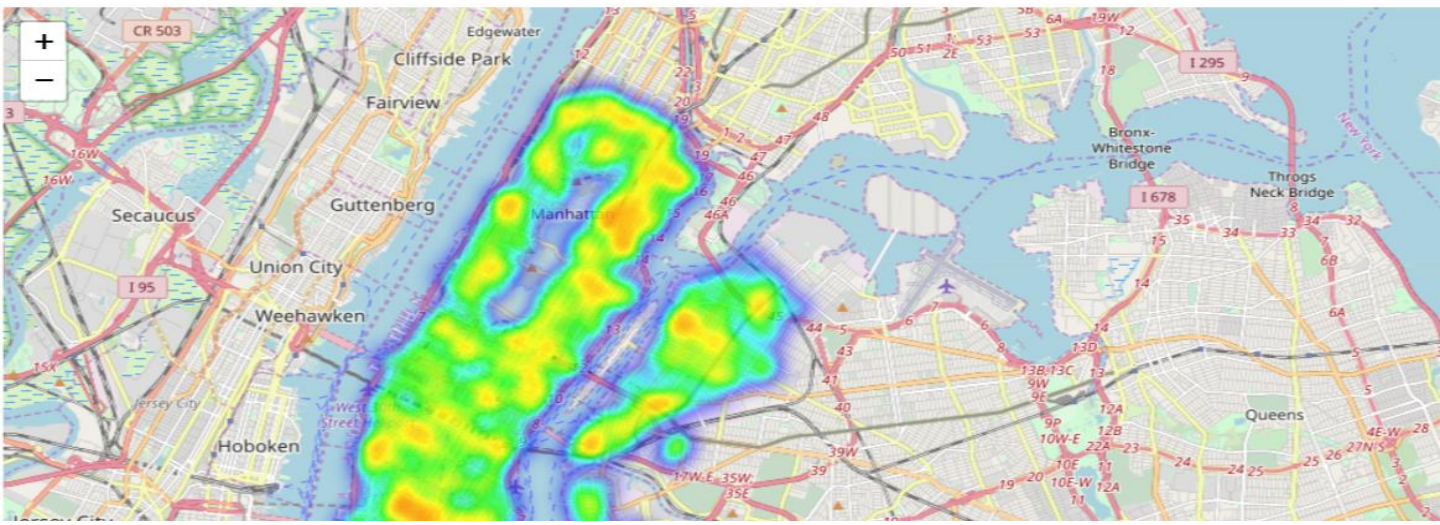
- 레퍼런스

  folium 공식 깃허브

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

  folium 공식 documentation

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



Growth Hackers

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 보태 주세요.
       - 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.

Growth Hackers

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

Growth Hackers