Cayley : the 2nd project

코로나19 대시보드, 이렇게.

김가윤, 조용주

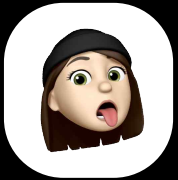
Team1

강영훈 김가윤 김가인 김현지 조용주 조현서

목차

오늘 이야기할 내용

- 코로나19 감염현황 데이터 EDA
- 코로나19 시도별 발생현황 데이터 EDA
- 코로나19 연령별, 성별 확진율 데이터 EDA
- 코로나19 선별진료소 현황 데이터 EDA
- 코로나19 해외 확진 현황 데이터 EDA
- 대시보드 레퍼런스



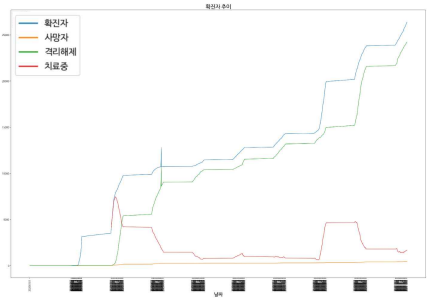
전처리

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Unnamed: 0     309 non-null    int64
 1   Unnamed: 0.1   309 non-null    int64
 2   누적 확진률         249 non-null    float64
 3   격리해제 수         309 non-null    int64
 4   등록일시분초          309 non-null    object
 5   사망자 수          309 non-null    int64
 6   확진자 수          309 non-null    int64
 7   검사진행 수         309 non-null    int64
 8   감염현황 고유값        309 non-null    int64
 9   기준일            309 non-null    int64
 10  기준시간           309 non-null    object
 11  수정일시분초          26 non-null     object
 12  누적 검사수         249 non-null    float64
 13  누적 검사 완료수      249 non-null    float64
 14  치료중 환자 수       249 non-null    float64
 15  결과 음성 수        249 non-null    float64
dtypes: float64(5), int64(8), object(3)
memory usage: 38.8+ KB
```

```
df['기준일'] = df['기준일'].astype(str)
year = df['기준일'].apply(lambda x: x[:4]).to_list()
month = df['기준일'].apply(lambda x: x[4:6]).to_list()
day = df['기준일'].apply(lambda x: x[6:]).to_list()
df["year"] = pd.DataFrame(year)
df["month"] = pd.DataFrame(month)
df["day"] = pd.DataFrame(day)
df['기준일'] = df[['year', 'month', 'day']].apply(lambda x: '-'.join(x), axis=1)
df['기준일'] = pd.to_datetime(df['기준일'])
df['기준일']
```

```
0     2020-01-01
1     2020-02-02
2     2020-02-03
3     2020-02-04
4     2020-02-05
         ...
304   2020-10-26
305   2020-10-27
306   2020-10-28
307   2020-10-29
308   2020-10-30
Name: 기준일, Length: 309, dtype: datetime64[ns]
```

시각화



	Unnamed: 0	누적 확진률	격리해제 수	등록일시분초	사망자 수	확진자 수	검사진행 수	감염현황 고유값	기준일	기준시간
249	249	NaN	30	2020-03-01 17:41:45.45	18	3736	33360	60	20200301	16:00
250	250	NaN	30	2020-03-01 10:06:32.32	17	3526	32422	59	20200301	09:00
251	251	NaN	28	2020-02-29 17:16:20.20	17	3150	35182	58	20200229	16:00
252	252	NaN	27	2020-02-29 10:14:50.50	16	2931	29154	57	20200229	09:00

시각화 방안

- 기준일이 같은 데이터를 1개로 합쳐주고,
  '(누적)확진자 수'를 활용하여, '일별 확진자 수'를 추가하고,
  '(누적)사망자 수'를 활용하여, '일별 사망자 수'를 추가한다면,
  다양한 시각화가 가능해집니다.

- 일별 확진자 수, 사망자 수, 격리 해제자 수 및
  치료 중 환자 수가 예시가 될 수 있죠.

문제점

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4510 entries, 0 to 4699
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   createdt       4510 non-null   datetime64[ns]
 1   deathcnt       4510 non-null   int64
 2   defcnt         4510 non-null   object
 3   gubun          4491 non-null   object
 4   incdec         4491 non-null   float64
 5   isolclearcnt   4510 non-null   float64
 6   isolingcnt     4510 non-null   object
 7   localocccnt    4492 non-null   object
 8   overflowcnt    4402 non-null   object
 9   qurrate        4014 non-null   object
 10  seq            4009 non-null   float64
 11  stdday         4009 non-null   object
 12  updatedt       24 non-null     object
dtypes: datetime64[ns](1), float64(3), int64(1), object(8)
memory usage: 493.3+ KB
```

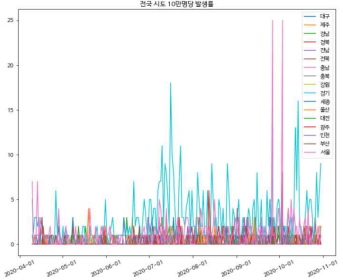
	createdt	deathcnt	defcnt	gubun	incdec	isolclearcnt	isolingcnt	localocccnt	overflowcnt
4180	2020-03-30 10:29:47.470	0	202	검역	13.0	0.0	-	725	2020년 03월 30일 00시
4181	2020-03-30 10:29:47.470	0	9	제주	1.0	4.0	1.34	724	2020년 03월 30일 00시
4182	2020-03-30 10:29:47.470	0	95	경남	1.0	65.0	2.83	723	2020년 03월 30일 00시
4183	2020-03-30 10:29:47.470	38	1298	경북	11.0	772.0	48.75	722	2020년 03월 30일 00시
4184	2020-03-30 10:29:47.470	0	9	전남	0.0	3.0	0.48	721	2020년 03월 30일 00시
...	...	...	...	...	...	...	...	...	...
4695	2020-03-02 19:27:57.570	0	인천	NaN	NaN	5.0	2020년 3월 1일 16시	NaN	NaN
4696	2020-03-02 19:27:57.570	9	대구	NaN	NaN	4.0	2020년 3월 1일 16시	NaN	NaN
4697	2020-03-02 19:27:57.570	0	부산	NaN	NaN	3.0	2020년 3월 1일 16시	NaN	NaN
4698	2020-03-02 19:27:57.570	0	서울	NaN	NaN	2.0	2020년 3월 1일 16시	NaN	NaN
4699	2020-03-02 19:27:57.570	18	합계	NaN	NaN	1.0	2020년 3월 1일 16시	NaN	NaN

코로나19 시도별 발생현황 데이터 EDA

문제점

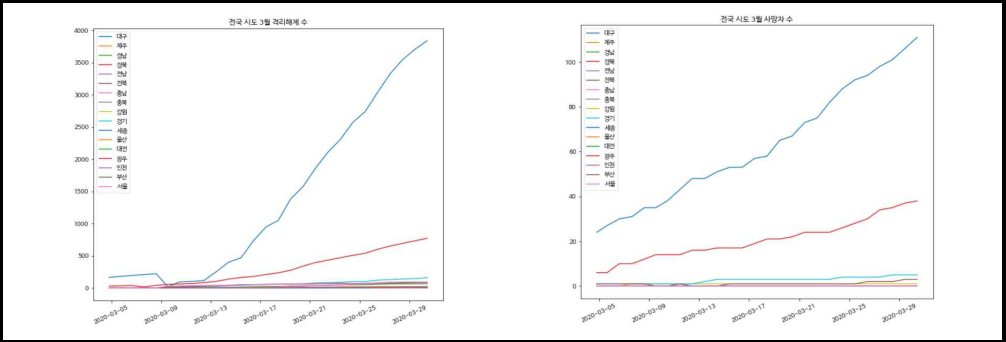
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4009 entries, 0 to 4008
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   createdt       4009 non-null   datetime64[ns]
 1   deathcnt       4009 non-null   int64
 2   defcnt         4009 non-null   int32
 3   gubun          4009 non-null   object
 4   incdec         4009 non-null   float64
 5   isolclearcnt   4009 non-null   float64
 6   isolingcnt     4009 non-null   int32
 7   localoccccnt   4009 non-null   int32
 8   overflowcnt    4009 non-null   int32
 9   qurrate        4009 non-null   object
 10  seq            4009 non-null   float64
 11  stdday         4009 non-null   object
 12  updatedt       24 non-null     object
dtypes: datetime64[ns](1), float64(3), int32(4), int64(1), object(4)
memory usage: 375.8+ KB
```



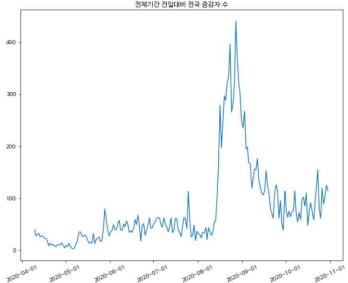
코로나19 시도별 발생현황 데이터 EDA

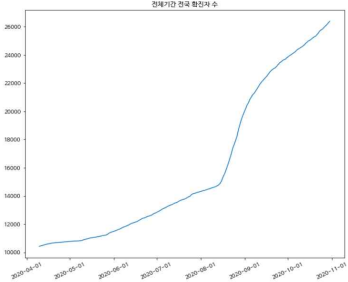
3월 칼럼 시각화 예시



코로나19 시도별 발생현황 데이터 EDA

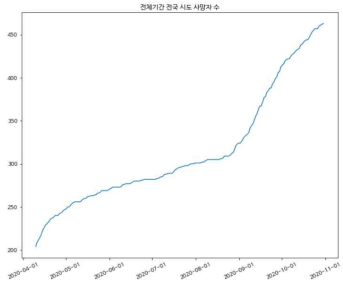
전국, 전체 기간 단위 예시

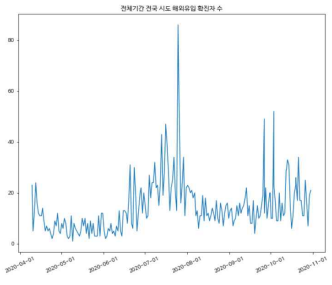




코로나19 시도별 발생현황 데이터 EDA

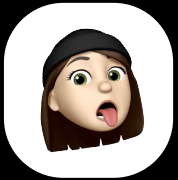
전국, 전체 기간 단위 예시





대시보드에는?

- 각 월마다 칼럼 별 변화 추이를 시각화합니다.
- 전국 단위 월별, 전체기간 추이를 시각화합니다.
- 전국 단위 시각화는 그래프도 좋지만 지도를 적극 활용해 한 눈에 변화를 알 수 있도록 합니다.
- 급격한 변화 양상이 보이는 구간에는 뉴스 링크 등을 걸어 원인을 설명하는 것도 좋을 것 같습니다.
- 자동 업데이트, 쉬운 시각화를 위한 함수를 구현합니다.



세션 준비자 : 김가윤 | 발표자 : 김가윤 | 디자인 : 조용주

코로나19 연령별, 성별 확진율 데이터 EDA

변수 슬라이싱 및 범주화

confcase	confcaserate	createdt	criticalrate	death	deathrate	gubun	seq	updatedt
685	2.60	2020-10-30 14:10:38.514	0.00	0	0.00	0-9	4336	null
1452	5.50	2020-10-30 14:10:38.514	0.00	0	0.00	10-19	4335	null
5147	19.51	2020-10-30 14:10:38.514	0.00	0	0.00	20-29	4334	null
3263	12.37	2020-10-30 14:10:38.513	0.06	2	0.43	30-39	4333	null
3545	13.44	2020-10-30 14:10:38.513	0.11	4	0.86	40-49	4332	null
4845	18.36	2020-10-30 14:10:38.513	0.43	21	4.54	50-59	4331	null
4202	15.93	2020-10-30 14:10:38.513	1.26	53	11.45	60-69	4330	null
2106	7.98	2020-10-30 14:10:38.513	7.22	152	32.83	70-79	4329	null
1140	4.32	2020-10-30 14:10:38.513	20.26	231	49.89	80 이상	4328	null
14096	53.42	2020-10-30 14:10:38.513	1.53	215	46.44	여성	4327	null
12289	46.58	2020-10-30 14:10:38.512	2.02	248	53.56	남성	4326	null

- 일자별 확진 수, 확진 비율, 치명률, 사망률 등의 자료를 연령대와 성별로 구분한 자료입니다.
- 2020.4.02 ~ 2020.10.30 기간의 자료입니다.

세션 준비자 : 강영훈 | 발표자 : 김가윤 | 디자인 : 조용주

변수 슬라이싱 및 범주화

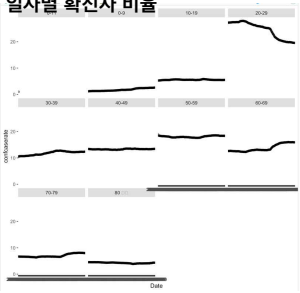
gubun	seq	updatedt	Date	Time
0-9	4336	null	2020-10-30	14:10:38.514
10-19	4335	null	2020-10-30	14:10:38.514
20-29	4334	null	2020-10-30	14:10:38.514
30-39	4333	null	2020-10-30	14:10:38.513
40-49	4332	null	2020-10-30	14:10:38.513
50-59	4331	null	2020-10-30	14:10:38.513
60-69	4330	null	2020-10-30	14:10:38.513
70-79	4329	null	2020-10-30	14:10:38.513
80 이상	4328	null	2020-10-30	14:10:38.513

- 날짜와 시간을 구분 짓고, 성별 분류는 다른 이름에 저장하였습니다.
- 추후 월별로 분류 계획입니다.

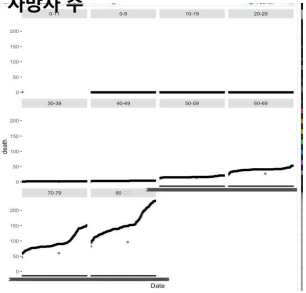
코로나19 연령별, 성별 확진율 데이터 EDA

데이터 시각화 (R)

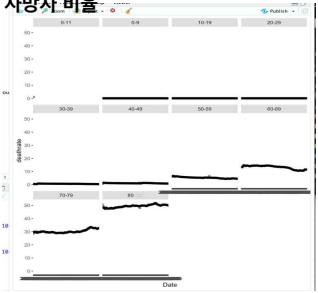
일자별 확진자 비율



사망자 수

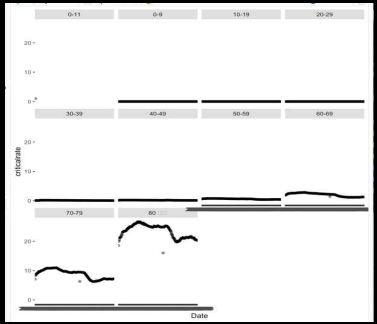


사망자 비율



코로나19 연령별, 성별 확진율 데이터 EDA

변수 슬라이싱 및 범주화



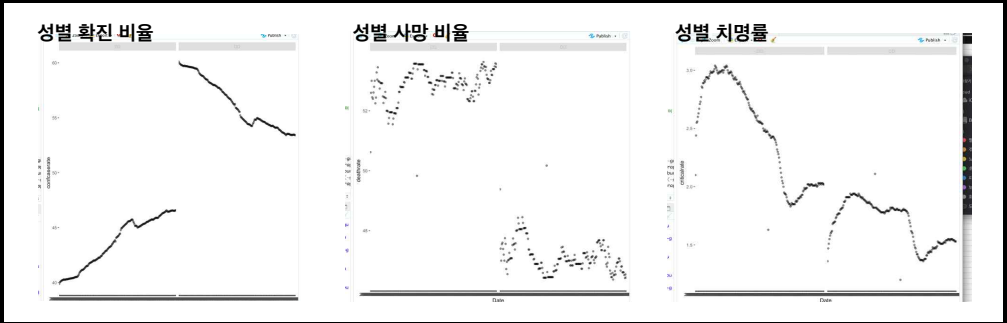
- 초반 확진자 수 비율은 20대에서 높았지만 점점 줄고 노년층 비중이 커지는 양상을 보입니다.
- 사망자 수와 비율은 70대 이상에서 계속 높은 비율을 유지하고요.
- 치명률 역시 70대 이후에서 높은 수준으로 유지되나 특정 시점 기준으로 다소 떨어집니다.
- 치료제 도입 혹은 면역력 증대를 원인으로 봅니다.

세션 준비자 : 강영훈 | 발표자 : 김가윤 | 디자인 : 조용주

코로나19 연령별, 성별 확진율 데이터 EDA

데이터 시각화 (R)

- 확진 비율은 여성이 적지만 사망 비율, 치명률 경우에는 여성이 남성보다 높은 것을 알 수 있습니다.



선별진료소 + 드라이브스루

- 시도/ 혹은 주소를 이용하여 각 지역별 선별 진료소 수와 밀집도를 확인합니다.

- 이를 통해 지역별 인구 밀집도를 예측합니다.
  선별진료소와 인구수가 비례하고 드라이브스루와 인구수가 반비례한다, 이런 식으로요.

- 라이브러리는 Matplotlib의 히스토그램(진료소 수)과
  판다스의 지도 시각화(밀집도)를 사용합니다.

세션 준비자 : 조현서 | 발표자 : 김가윤 | 디자인 : 조용주

코로나19 선별진료소 현황 데이터 EDA

확진자 수와 선별진료소 현황을 이용한 분석

- [지역별 인구 or 지역별 확진자 수]와 [지역별 선별진료소 수 or 지역별 드라이브 스루 수]를 비교하여 지역별 적절한 선별 진료소 수와 드라이브 스루 수를 책정합니다.

- 인구 대비 적정 선별 진료소 설치 수 기준 및 가장 확진자 억누르고 있는 지역을 기준 시, 도별 기준을 나눕니다.

- 라이브러리는 Matplotlib의 산점도와 버블차트를 사용합니다.

세션 준비자 : 조현서 | 발표자 : 김가윤 | 디자인 : 조용주

전처리

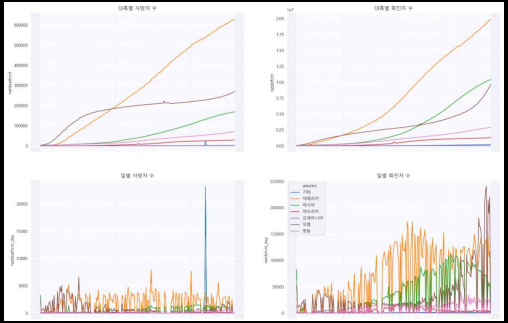
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43437 entries, 0 to 43436
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   areanm         43437 non-null  object
 1   areanmcn       42973 non-null  object
 2   areanmen       42973 non-null  object
 3   createdt       43437 non-null  object
 4   natdeathcnt    43437 non-null  int64
 5   natdeathrate   43437 non-null  float64
 6   natdefcnt      43437 non-null  int64
 7   nationnm       43437 non-null  object
 8   nationnmcn     42973 non-null  object
 9   nationnmen     42973 non-null  object
 10  seq            43437 non-null  int64
 11  stdday         43437 non-null  object
 12  updatedt       70 non-null     object
dtypes: float64(1), int64(3), object(9)
memory usage: 4.3+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 43435 entries, 0 to 43436
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   areanm         43435 non-null  object
 1   createdt       43435 non-null  object
 2   natdeathcnt    43435 non-null  int64
 3   natdeathrate   43435 non-null  float64
 4   natdefcnt      43435 non-null  int64
 5   nationnm       43435 non-null  object
 6   seq            43435 non-null  int64
dtypes: float64(1), int64(3), object(3)
memory usage: 3.9+ MB
```

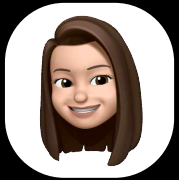
코로나19 해외 확진 현황 데이터 EDA

EDA 1. ‘20. 10. 30. 기준 대륙별 (단위 : 천)

EDA 2. 대륙별 사망자, 확진자 추이



- 아메리카의 경우 확진자, 사망자 모두 꾸준히 빠르게 증가하고 있습니다.
- 유럽은 5월까지 사망자가 급증하다가 유지되는 추세였으나 9월 들어 확진자가 급증하였죠.



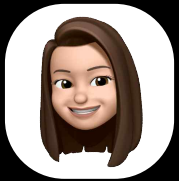
세션 준비자 : 김현지 | 발표자 : 김가윤 | 디자인 : 조용주

문제점

- 오기입이 있어서 일일이 찾아 바꿔야 했어요. 국가 명(nationnm) 통일이 제대로 안 되어 있었고, 아시아(areanm)에 속해 있던 러시아는 7월부터 유럽으로 집계되기도 하는 등, 기입 오류가 있었습니다.

- '기타' 대륙의 국가 대부분이 4월 이후로 집계가 멈추어 있어 이용이 불가능했습니다.

- 일별 확진자, 사망자 수를 파악하기 위해 '금일-전일 = 금일' 으로 계산했으나, 음수가 존재했습니다. 집계의 기준을 정확히 알아 봐야할 것 같습니다.



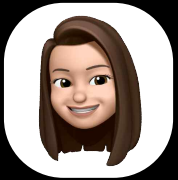
세션 준비자 : 김현지 | 발표자 : 김가윤 | 디자인 : 조용주

대시보드에는?

- 개별 국가가 223~186개라 대륙별로 구현했으나 해외 국가별 코로나19 확진자, 사망자 현황 시각화도 (지도&선 그래프 이용) 필요할 듯 합니다.

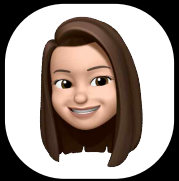
- 전세계 월&일 단위 추이 시각화 및 해외 국가별 인구수 대비 확진자, 사망자 분석 예정에 있습니다.

- 지도 시각화를 통해 보다 직관적으로 현황을 조명해야 합니다. 메소드로는 인터랙티브하게 자세한 수치를 보여줄 수 있는 bokeh, plotly 를 사용하면 좋을 것 같습니다.



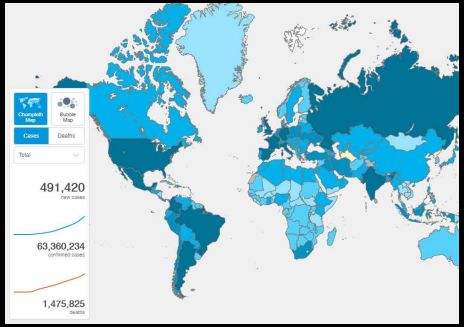
라이브러리 사용, 이렇게.

- Matplotlib : 기본 시각화 라이브러리로, 바, 파이, 히스토그램, 산점도 등 다양한 플롯을 제공합니다.
- Seaborn : Matplotlib에 색상, 통계 차트 등의 기능을 추가한 시각화 패키지로, Jointplot, pairchart 등 다차원 데이터의 비교 시각화가 가능합니다.
- Bokeh : 인터렉티브 그래프, 지도 시각화를 구현하며 화려한 디자인이 특징입니다.
- Plotly : D3.js를 이용해 인터렉티브 그래프를 구현하며, 지도 시각화도 가능합니다.
- Folium : 가벼운 지도 시각화 패키지로 pandas와 쉽게 연동할 수 있습니다.



대시보드 레퍼런스

WHO



- Choropleth 맵으로 각 주별 확진자 수에 따라 색을 달리하거나 Folium의 circle marker로 확진자 수에 따라 크기를 달리하였습니다.

- 지도 json 파일만 있으면 카피 가능합니다. Gmaps api도 가능하구요.

세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

대시보드 레퍼런스

Johns Hopkins Univ

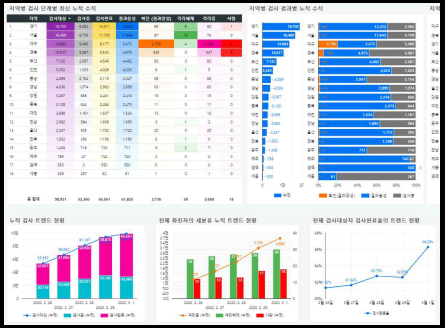
- 히트맵 스타일로 지역별 확진자 수를 나타냈습니다.

- 로딩 시간 매우 느려요. 한국 기준이라면 크게 지연 없이 구현 가능할 것 같기는 합니다.

세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

대시보드 레퍼런스

Google Data Studio



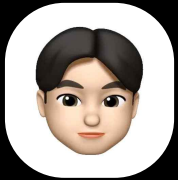
- 코로나 관련 검사누적부터 단계 누적 등 다양한 데이터를 바그래프 혹은 표로 정리하였습니다.
- 가독성이 너무 떨어집니다. 다양한 데이터를 다루려다보니 메인에 비주얼적 임팩트가 없어요.

세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

대시보드 레퍼런스

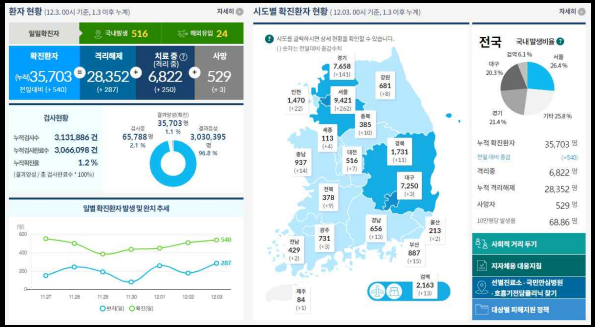
코로나 라이브

- 깔끔하고 예쁜 UI를 가졌습니다. 화면 전환 및 작동도 매우 부드러우나 모바일만 지원합니다.
- 데이터 업데이트 주기도 매우 짧습니다. 확진자 현황 등이 실시간으로 업데이트 됩니다.



세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

질병관리청

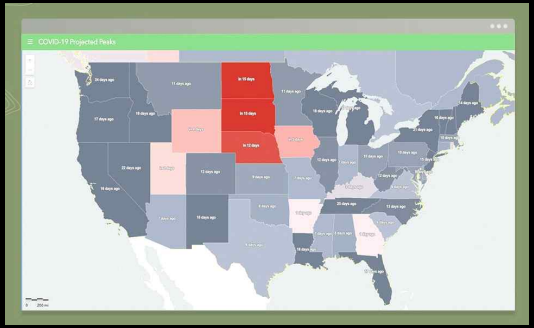


- 구글 데이터 스튜디오와 같이 다양한 데이터를 다루지만 가독성이 좋습니다.
- 메인 지도 컬러링과 숫자 표기 방식이 시각화와 정보전달 모두에 가장 적합하다고 느꼈어요. 네이버나 카카오도 같은 방식의 시각화를 코로나 배너에서 메인으로 보여줍니다.

세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

대시보드 레퍼런스

Esri - IHME 예상 피크 대시보드

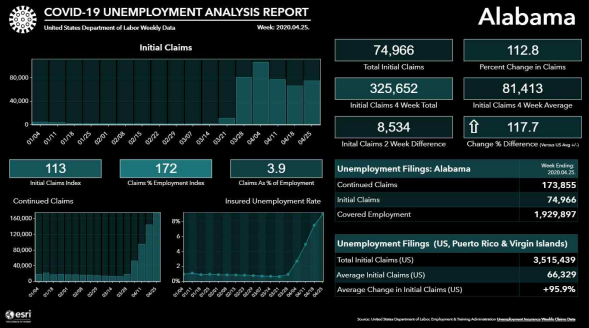


- 기존 확진자 데이터와 머신러닝을 통해 향후 확진자 수를 예측하는 모델입니다.

세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

대시보드 레퍼런스

Esri - 미국 내 실업자 현황

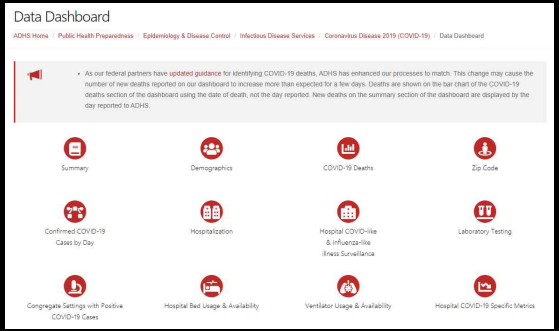


- 코로나19 이후 미국 내 실업자 현황을 보여주는 대시보드입니다.
- 단순 실업자 데이터가 아니라 사회학적 데이터들을 가져와도 유의미할 것 같아요.

세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

대시보드 레퍼런스

Arizona Department of Health Services



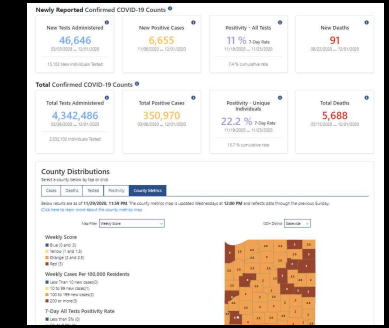
- 한 페이지에 모든 대시보드를 나타내지 않고 아이콘을 통해 구분 후 클릭시 해당 대시보드로 이동하는 방식으로 구성되어 있습니다.

- 한 페이지 내에 가독성 있게 내용들을 담을 수 없다면 이 대시보드처럼 구분하는 것이 가장 이상적일 것 같아요.

세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

대시보드 레퍼런스

Indiana COVID-19 Dashboard



- 관련 숫자들을 메인에 보여주고 하단에 지도 등으로 데이터를 시각화하여 보여줍니다.
- 데이터 허브 및 대중적 시각화 사이트로 기능하기라는 본 프로젝트 목표에 부합하는 타협안으로 보입니다.

세션 준비자 : 조용주 | 발표자 : 조용주 | 디자인 : 조용주

코로나19 대시보드, 이렇게.

발표를 들어주셔서 감사합니다.
Presented by Team1. Designed by Peniel Cho.