**4월 6일 계획서**

| **오전** | **데이터 수집** | **09:00~11:40** |
| --- | --- | --- |
|  | **점심시간** | **11:45~13:15** |
| **오후** | **데이터 정제** | **13:15~16:00** |
| **데이터 분석** | **16:00~18:00** |

**4월 6일 진행 내용**

| **오전** | 1. **코로나 데이터 수집** | **09:00~09:30** |
| --- | --- | --- |
| 1. **킥보드 데이터 수집** | **09:30~11:40** |
| **점심** | **점심시간** | **11:45~13:15** |
| **오후** | 1. **킥보드 데이터 수정** | **13:15~14:30** |
| 1. **코로나 데이터 수정** | **14:30~15:30** |
| 1. **서울시 인구 데이터 수정** | **15:30~** |

**1) 코로나 데이터 수집**

https://ncov.mohw.go.kr/

코로나 확진자 수 데이터셋

설명: 서울시의 코로나 확진자 데이터셋 구함.

필요한 컬럼과 필요없는 컬럼을 나누고, 필요한 부분만 담고 있는 csv파일로 만듬.

**2) 킥보드 데이터 수집**

<https://stophyun.tistory.com/207>

킥보드 위치에 대한 스크래핑 방법. 필요한 데이터는 사용량, 이용객수 이므로 필요가 없음.

<https://www.findatamall.or.kr/fsec/myPage/purchaseInfo.do?cmnx=67&sCateType=1&speriod=&sdayf=&sdayt=&stype=&skey=>

쉐어카. 킥보드 데이터. 필요한 데이터. but, 불필요한 컬럼과 성별 ,나이별로 사용자수가 나눠져 있음. 이를 통합해야 할것 같음.

**3) 킥보드 데이터 수정**

킥보드에 대한 정보를 사용하기 쉽게 csv파일을 데이터프레임에서 불필요한 컬럼과,

groupby를 이용하여 묶어서 데이터를 알아보기 쉽게 정리함.

| csv\_test = csv\_test.drop(["성별","연령대\_10","연령대\_05","매출금액(천원)"], axis=1)  grouped =csv\_test.groupby(["기준년도","기준월","카테고리"])  grouped = grouped.first()  grouped.to\_csv("check123.csv", mode='w') |
| --- |

| **수정 전** | **수정 후** |
| --- | --- |
|  |  |

**4) 코로나 데이터 수정**

서울의 코로나 확진자 데이터셋이 일별로 되어있는데, 이를 년도, 월로 나누고,

groupby로 년도-월 로 묶어 확진자를 묶음.

| **수정전** | **수정후** |
| --- | --- |
|  |  |

**5) 서울시 인구 데이터 수정**

서울의 인구 데이터가 너무 큼. 그리고, 월별로 안되어 있으며, 필요없는 컬럼들이 많음

해당 컬럼들을 축소화 하고, 필요한 컬럼을 만드는 작업을 진행함.

66만개에 이르는 row를 600개로 간소화 시킴.

| **수정전** | **수정후** |
| --- | --- |
|  |  |