**4월 7일 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 7일 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. **서울통합인구수 CSV 수정**

서울에서 각 구 별 ‘총생활인구수’에 대한 데이터 컬럼들이 ‘시간대구분’으로 0~23시까지

나뉘어져 있음. 이를 0시를 기준으로 한다.

‘기준일ID’ 컬럼의 값은 Day를 기준으로 나뉘어져 있다. Month별 분석을 진행할 것이기 때문에, day로 되어있는 것들을 sum을통하여 Month로 합쳐주어 이 데이터의 Row개수를 줄일것이다.

이 csv 데이터 프레임은 총 32개의 Column과 219600개의 row를 가진다.

| csv\_test= csv\_test[['기준일ID','시간대구분','자치구코드','총생활인구수']]  csv\_test = csv\_test[csv\_test['시간대구분']==0]  csv\_test= csv\_test[['기준일ID','자치구코드','총생활인구수']]  csv\_test = csv\_test.astype({'기준일ID':'string'})  csv\_test = csv\_test.astype({'총생활인구수':'int'})  csv\_test["Year"] = csv\_test['기준일ID'].str[:4]  csv\_test["Month"] = csv\_test['기준일ID'].str[4:6]  csv\_test = csv\_test[['Year','Month','자치구코드','총생활인구수']]  grouped =csv\_test.groupby(["Year","Month","자치구코드"]).agg('mean')  grouped.to\_csv("zzz1.csv", mode='w') |
| --- |

**csv 변경 작업**

| 변경 전 | 변경 후 |
| --- | --- |
|  |  |

**csv 추가 append 작업**

해당 csv 파일에는 지역을 구분하는것이 ‘자치구코드’로 되어있다.

‘자치구코드’로는 다른 csv와 조인을 하는것이 어려워 보인다.

다른 파일에 ‘자치구코드’를 넣어주거나, 이 csv파일에 지역명칭을 넣어주는게 바람직하다. 따라서, 이 데이터프레임에 ‘자치구코드’에 따른 구 명칭을 추가해주는 작업을 진행하였다.

| code\_list=[11110,11140,11170,11200,11215,11230,11260,11290,11305,11320,11350,11380,11410,11440,11470,11500,11530,11545,11560  ,11590,11620,11650,11680,11710,11740]  name\_list=["종로구","중구","용산구","성동구","광진구","동대문구","중랑구","성북구","강북구","도봉구","노원구","은평구","서대문구",  "마포구","양천구","강서구","구로구","금천구","영등포구","동작구","관악구","서초구","강남구","송파구","강동구"]  for i in range(0,25):  opt = csv\_test['자치구코드'] == code\_list[i]  csv\_test.loc[(opt), 'Gu\_name'] = name\_list[i]  csv\_test.to\_csv("check123.csv", mode='w') |
| --- |

**CSV 변경 작업**

| 변경 전 | 변경 후 |
| --- | --- |
|  |  |

1. **버스 CSV 수정**

해당 버스 컬럼에서 ‘사용월’, ‘노선ID’, ‘노선번호’, ‘노선명’, ‘버스정류장ID’, ‘버스정류장명’, ‘승차승객수’, ‘하차승객수’, ‘작업일시’ 컬럼이 있다.

여기서, 사용월 컬럼에서 년도와 달력의 새로운 컬럼을 만들고

승차승객수, 하차 승객수 컬럼을 month와 year로 group을 한다음 1달치 승하차 인원수를 담는 데이터프레임으로 변경하고자 한다.

버리는 컬럼:’노선ID, 노선번호, 노선명, 버스정류장ID, 작업일시

새로운 컬럼을 만드는 컬럼: 사용월

사용되지만 변경되지 않는 컬럼:버스 정류장명

1달치 SUM을 통하여 변경되는 컬럼: 승,하차승객수

| import pandas as pd  csv\_test = pd.read\_csv('data/2021년 버스.csv', encoding="CP949")  csv\_test = csv\_test.drop(["작업일시"], axis=1)  csv\_test['월'] = csv\_test['사용월'].str.split('-').str[0]  csv\_test['년도'] = "20" + csv\_test['사용월'].str.split('-').str[1]  csv\_test=csv\_test.drop(["사용월"], axis=1)  month\_list= list(csv\_test['월'].unique())  month\_match=[1,2,3,4,5,6,7,8,9,10,11,12]  print(month\_list)  for i in range(0,12):  opt = csv\_test['월'] == month\_list[i]  csv\_test.loc[(opt), '월'] = month\_match[i]  csv\_test = csv\_test.astype({'승차승객수':'string','하차승객수':'string'})  csv\_test['승차승객수']=csv\_test['승차승객수'].str.replace(',','')  csv\_test['하차승객수']=csv\_test['하차승객수'].str.replace(',','')  csv\_test = csv\_test.astype({'월':'int', '년도': 'int', '승차승객수':'int','하차승객수':'int'})  csv\_test = csv\_test.drop(["노선ID","노선번호","노선명","버스정류장ID"], axis=1)  grouped =csv\_test.groupby(["년도","월","버스정류장명"]).agg({'승차승객수':sum, '하차승객수':sum})  grouped.to\_csv("zzz1.csv", mode='w') |
| --- |

**CSV 변경 작업**

| 변경 전 | 변경 후 |
| --- | --- |
|  |  |

469361\*3 의 row 개수 → 96330 \*3 개 row 개수로 축소 되었음.

1. 지하철 CSV 수정

해당 지하철데이터 컬럼에서 버스데이터와 마찬가지로

‘사용월’을 년도와 월의 컬럼으로 나누는 작업을 할것이다.

호선과 역 ID, 작업일시컬럼을 버리고 승,하차승객수 컬럼을 1달의 SUM값을 담도록 만들것이다.

**CSV 수정 코드**

| csv\_test = csv\_test.drop(["작업일시"], axis=1)  csv\_test['월'] = csv\_test['사용월'].str.split('-').str[0]  csv\_test['년도'] = "20" + csv\_test['사용월'].str.split('-').str[1]  csv\_test=csv\_test.drop(["사용월"], axis=1)  month\_list= list(csv\_test['월'].unique())  month\_match=[1,2,3,4,5,6,7,8,9,10,11,12]  print(month\_list)  for i in range(0,12):  opt = csv\_test['월'] == month\_list[i]  csv\_test.loc[(opt), '월'] = month\_match[i]  csv\_test = csv\_test.astype({'승차승객수':'string','하차승객수':'string'})  csv\_test['승차승객수']=csv\_test['승차승객수'].str.replace(',','')  csv\_test['하차승객수']=csv\_test['하차승객수'].str.replace(',','')  csv\_test = csv\_test.astype({'월':'int', '년도': 'int', '승차승객수':'int','하차승객수':'int'})  csv\_test = csv\_test.drop(["역ID","호선명"], axis=1)  grouped =csv\_test.groupby(["년도","월","지하철역"]).agg({'승차승객수':sum, '하차승객수':sum})  grouped.to\_csv("zzz1.csv", mode='w') |
| --- |

**CSV 변경 작업 전, 후**

| **변경 전** | **변경 후** |
| --- | --- |
|  |  |