Final Project ETL

CIS 9440-UWA
Final Project Milestone 3
Group Number - 19
Students: Hanqing Chen (hanqing.chen@baruchmail.cuny.edu)
Xinyue Chen (xinyue.chen@baruchmail.cuny.edu)
Candice Lee (candice.lee@baruchmail.cuny.edu)

Load public data

Dataset 1: New York State Statewide COVID-19 Testing_NY state
https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Testing/xdss-U53e

```
NYS_COV19.head()
```

	test_date	county	new_positives	cumulative_number_of_positives	total_number_of_tests	cumulative_number_of_tests
0	2020-03-01T00:00:00.000	Albany	0	0	0	0
1	2020-03-02T00:00:00.000	Albany	0	0	0	0
2	2020-03-03T00:00:00.000	Albany	0	0	0	0
3	2020-03-04T00:00:00.000	Albany	0	0	0	0
4	2020-03-05T00:00:00.000	Albany	0	0	3	3

Dataset 2: NYC Complaint Data__NYC Open Data
https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243

```
NYPD_COMPL.head()
```

	cmplnt_num	addr_pct_cd	cmplnt_fr_dt	cmplnt_fr_tm	crm_atpt_cptd_cd	juris_desc	ky_cd	law_cat_cd	loc_of_occur_desc	ofns_desc	...	susp_ag
0	885776788	66	2020-12-23T00:00:00.000	19:50:00	COMPLETED	N.Y. POLICE DEPT	101	FELONY	OUTSIDE	MURDER & NON-NEGL. MANSLAUGHTER	...	
1	350637195	77	2020-12-21T00:00:00.000	01:10:00	COMPLETED	N.Y. POLICE DEPT	101	FELONY	INSIDE	MURDER & NON-NEGL. MANSLAUGHTER	...	
2	347843168	43	2020-11-22T00:00:00.000	22:00:00	COMPLETED	N.Y. POLICE DEPT	104	FELONY	NaN	RAPE	...	UN
3	197941396	47	2020-11-22T00:00:00.000	09:50:00	COMPLETED	N.Y. POLICE DEPT	101	FELONY	INSIDE	MURDER & NON-NEGL. MANSLAUGHTER	...	
4	298404927	25	2020-11-21T00:00:00.000	15:38:00	COMPLETED	N.Y. HOUSING POLICE	101	FELONY	OUTSIDE	MURDER & NON-NEGL. MANSLAUGHTER	...	

Extract data needed for our project

NYPD_COMPL

```
NYPD_COMPL_df = NYPD_COMPL[["cmplnt_fr_dt","boro_nm","law_cat_cd","ofns_desc","susp_age_group",
                            "susp_race","susp_sex","vic_age_group","vic_race","vic_sex"]].copy()
NYPD_COMPL_df.head()
```

	cmplnt_fr_dt	boro_nm	law_cat_cd	ofns_desc	susp_age_group	susp_race	susp_sex	vic_age_group	vic_race	vic_sex
0	2020-12-23T00:00:00.000	NaN	FELONY	MURDER & NON-NEGL. MANSLAUGHTER	NaN	NaN	NaN	18-24	BLACK	M
1	2020-12-21T00:00:00.000	NaN	FELONY	MURDER & NON-NEGL. MANSLAUGHTER	NaN	NaN	NaN	25-44	BLACK	M
2	2020-11-22T00:00:00.000	BRONX	FELONY	RAPE	UNKNOWN	UNKNOWN	U	25-44	BLACK	F
3	2020-11-22T00:00:00.000	NaN	FELONY	MURDER & NON-NEGL. MANSLAUGHTER	25-44	BLACK	M	25-44	BLACK	F
4	2020-11-21T00:00:00.000	NaN	FELONY	MURDER & NON-NEGL. MANSLAUGHTER	NaN	NaN	NaN	18-24	BLACK HISPANIC	M

NYS_COVID

```
NYS_COV19_df = NYS_COV19[["test_date","county","new_positives","total_number_of_tests"]].copy()
NYS_COV19_df.head()
```

	test_date	county	new_positives	total_number_of_tests
0	2020-03-01T00:00:00.000	Albany	0	0
1	2020-03-02T00:00:00.000	Albany	0	0
2	2020-03-03T00:00:00.000	Albany	0	0
3	2020-03-04T00:00:00.000	Albany	0	0
4	2020-03-05T00:00:00.000	Albany	0	3

Data Cleansing

NYPD_COMPL

- Drop null value

```
NYPD_COMPL_df.dropna(inplace=True)
NYPD_COMPL_df.shape
```

```
(318265, 10)
```

- Drop wrong value of age_group

```
NYPD_COMPL_df.susp_age_group.value_counts()
```

```
UNKNOWN    144584
25-44      100238
45-64       33880
18-24       29613
<18          6617
65+          3312
2020           10
2019            2
-977            1
1020            1
-962            1
-942            1
-12             1
-71             1
-928            1
-965            1
1925            1
Name: susp_age_group, dtype: int64
```

```
age_group=["UNKNOWN","25-44","18-24","<18","65+","45-64"]
NYPD_COMPL_df=NYPD_COMPL_df[NYPD_COMPL_df['susp_age_group'].isin(age_group)]
NYPD_COMPL_df=NYPD_COMPL_df[NYPD_COMPL_df['vic_age_group'].isin(age_group)]
NYPD_COMPL_df.reset_index(inplace=True,drop=True)
```

- Limit date

```
NYPD_COMPL_df=NYPD_COMPL_df[NYPD_COMPL_df["cmplnt_fr_dt"]>="2020-01-01T00:00:00.000"]
```

- Drop duplicates

```
NYPD_COMPL_df=NYPD_COMPL_df.drop_duplicates()
```

NYS_COVID

- Drop null value

```
NYS_COV19_df.dropna(inplace=True)
NYS_COV19_df.shape
```

```
(25854, 4)
```

- Drop unnecessary county value

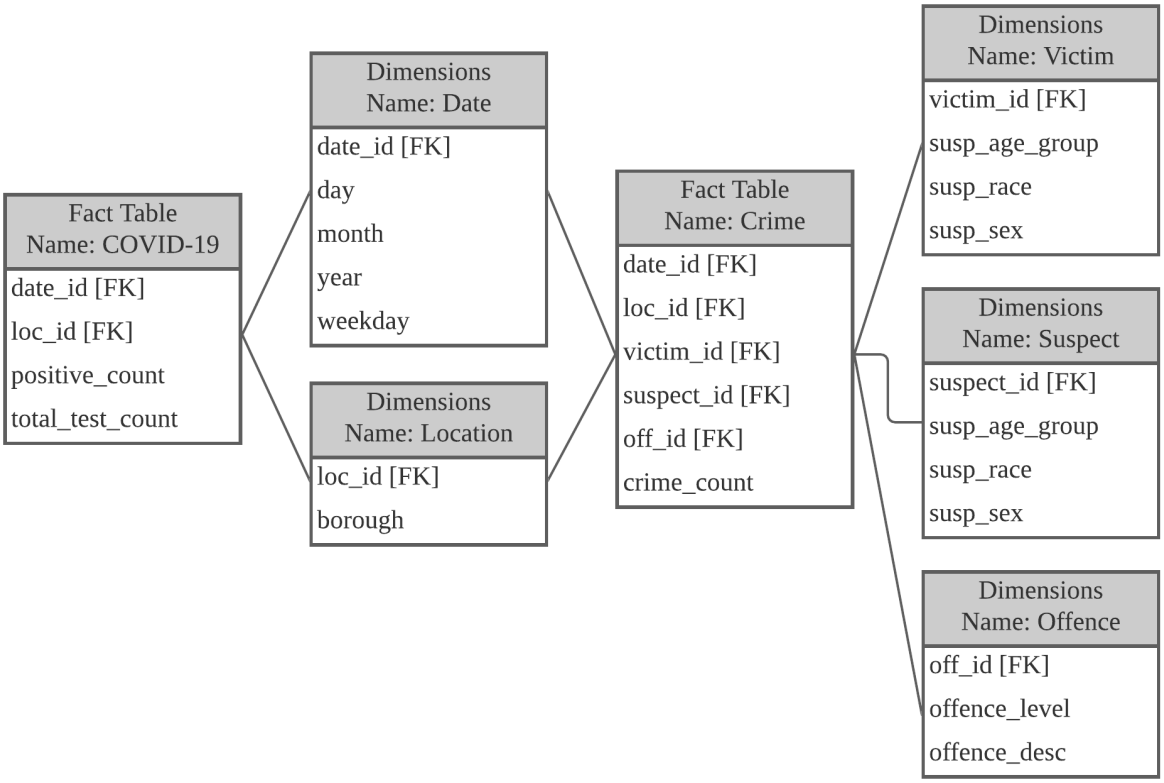
```
NYC_list=["New York","Bronx","Richmond","Kings","Queens"]
NYC_COV19_df=NYS_COV19_df[NYS_COV19_df.county.isin(NYC_list)]
NYC_COV19_df.reset_index(inplace=True,drop=True)
NYC_COV19_df
```

	test_date	county	new_positives	total_number_of_tests
0	2020-03-01T00:00:00.000	Bronx	0	0
1	2020-03-02T00:00:00.000	Bronx	0	0
2	2020-03-03T00:00:00.000	Bronx	0	1
3	2020-03-04T00:00:00.000	Bronx	0	0
4	2020-03-05T00:00:00.000	Bronx	0	5
...	...	...	...	...
2080	2021-04-17T00:00:00.000	Richmond	250	7083
2081	2021-04-18T00:00:00.000	Richmond	206	4749
2082	2021-04-19T00:00:00.000	Richmond	187	3561
2083	2021-04-20T00:00:00.000	Richmond	161	5209
2084	2021-04-21T00:00:00.000	Richmond	197	7130

- Unify county value with NYPD_COMPL data

```
boro_list = {"Bronx":'BRONX',"Queens":'QUEENS',"New York":'MANHATTAN',"Kings":'BROOKLYN',"Richmond":'STATEN ISLAND'}
NYC_COV19_df['county'] = [boro_list[item] for item in NYC_COV19_df['county']]
```

Data Transformation (Dimension and Fact Tables)



Suspect Dimension

```
suspect_df=NYPD_COMPL_df[["susp_age_group","susp_race","susp_sex"]]
suspect_df=suspect_df.drop_duplicates()
Suspect_df=suspect_df.reset_index(drop=True)
```

```
suspect_df["suspect_id"]=list(range(100,100+len(suspect_df)))
```

```
suspect_dim=suspect_df[['suspect_id','susp_age_group', 'susp_race', 'susp_sex']]
```

Victim Dimension

```
victim_df=NYPD_COMPL_df[["vic_age_group","vic_race","vic_sex"]]
victim_df=victim_df.drop_duplicates()
victim_df=victim_df.reset_index(drop=True)
```

```
victim_df["victim_id"]=list(range(1000,1000+len(victim_df)))
```

```
victim_dim=victim_df[["victim_id","vic_age_group","vic_race","vic_sex"]]
```

Offence Dimension

```
offence_df=NYPD_COMPL_df[["law_cat_cd","ofns_desc"]]
offence_df=offence_df.drop_duplicates()
offence_df=offence_df.reset_index(drop=True)
```

```
offence_df["off_id"]=list(range(5000,5000+len(offence_df)))
```

```
offence_dim=offence_df[["off_id","law_cat_cd","ofns_desc"]]
```

Location Dimension

```
loc_list = NYPD_COMPL_df.boro_nm.unique().tolist()
    # create blank list of dimension rows
dimension_rows = []

    # create author dimension with a surrogate key
for loc_id, boro_nm in enumerate(loc_list, start = 50):
    temp_list = [loc_id, boro_nm]
    dimension_rows.append(temp_list)

loc_dim = pd.DataFrame(data=dimension_rows,
                       columns = ["loc_id","borough"])
```

Date Dimension

```
date_df=NYS_COV19_df[['test_date']]
date_df=date_df.rename(columns={"test_date":"full_date"})
date_df=date_df.drop_duplicates()
date_df2=NYPD_COMPL_df[['cmplnt_fr_dt']]
date_df2= date_df2.rename(columns={"cmplnt_fr_dt":"full_date"})
date_df2=date_df2.drop_duplicates()
date_df=date_df.append(date_df2)
date_df=date_df.drop_duplicates()
```

```
date_df['full_date']=date_df['full_date'].str[0:10]
date_df['date_id']=date_df['full_date'].str[0:4]+date_df['full_date'].str[5:7]+date_df['full_date'].str[8:10]
date_df['full_date']=pd.to_datetime(date_df['full_date'])
date_df["year"]=date_df['full_date'].dt.year
date_df["month"]=date_df['full_date'].dt.month
date_df["day"]=date_df['full_date'].dt.day
date_df["weekday"]=date_df['full_date'].dt.day_name()
```

```
date_df=date_df[["date_id","day","month","year","weekday"]]
date_df=date_df.sort_values("date_id")
date_dim=date_df.reset_index(drop=True)
```

COVID-19 Fact Table

```
NYC_COV19_df1= NYC_COV19_df1.merge(date_dim, left_on='date_id', right_on='date_id',
                  how='inner')
NYC_COV19_df1= NYC_COV19_df1.merge(loc_dim, left_on='county', right_on='borough',
                  how='inner')
NYC_COV19_df1
```

```
NYC_COV19_fact=NYC_COV19_df1.drop(["test_date","county","day","month","year","weekday","borough"], axis=1)
```

```
NYC_COV19_fact
```

	new_positives	total_number_of_tests	date_id	loc_id
0	0	0	20200301	50
1	0	0	20200302	50
2	0	1	20200303	50
3	0	0	20200304	50
4	0	5	20200305	50
...	...	...	...	...
2080	250	7083	20210417	54
2081	206	4749	20210418	54
2082	187	3561	20210419	54
2083	161	5209	20210420	54
2084	197	7130	20210421	54

Crime Fact Table

```
NYPD_COMPL_df1= NYPD_COMPL_df1.merge(date_dim, left_on='date_id', right_on='date_id',
                  how='inner')
NYPD_COMPL_df1= NYPD_COMPL_df1.merge(loc_dim, left_on='boro_nm', right_on='borough',
                  how='inner')
NYPD_COMPL_df1= NYPD_COMPL_df1.merge(offence_dim, left_on=["law_cat_cd","ofns_desc"],
                                     right_on=["law_cat_cd","ofns_desc"],
                  how='inner')
NYPD_COMPL_df1= NYPD_COMPL_df1.merge(victim_dim, left_on=["vic_age_group","vic_race","vic_sex"],
                                     right_on=["vic_age_group","vic_race","vic_sex"],
                  how='inner')
NYPD_COMPL_df1= NYPD_COMPL_df1.merge(suspect_dim, left_on=["susp_age_group","susp_race","susp_sex"],
                                     right_on=["susp_age_group","susp_race","susp_sex"],
                  how='inner')
NYPD_COMPL_df1
```

```
NYPD_COMPL_fact=NYPD_COMPL_df1.drop(["cmplnt_fr_dt","boro_nm","law_cat_cd","ofns_desc","susp_age_group",
                                     "susp_race","susp_sex","vic_age_group","vic_race","vic_sex",
                                     "day","month","year","weekday","borough"],axis=1)
```

```
NYPD_COMPL_fact
```

	date_id	loc_id	off_id	victim_id	suspect_id
0	20201122	50	5000	1000	100
1	20201106	50	5000	1000	100
2	20200222	50	5000	1000	100
3	20200430	50	5000	1000	100
4	20200804	51	5000	1000	100
...	...	...	...	...	...
275125	20200806	50	5011	1005	200
275126	20200801	54	5011	1012	200
275127	20201102	53	5004	1033	200
275128	20200626	51	5016	1074	200
275129	20200704	53	5011	1076	221

Load Dimension and Fact Tables into Google BigQuery

```python
key_path='crime-covid-178c3ff212a1.json'
# fill in file path to your key here
credentials = service_account.Credentials.from_service_account_file(
    key_path, scopes=["https://www.googleapis.com/auth/cloud-platform"],)
client = bigquery.Client(credentials=credentials, project=credentials.project_id)
```

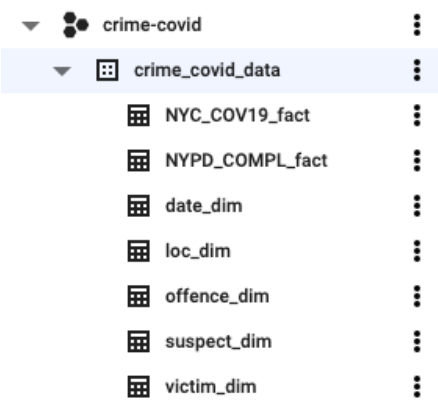
```python
def load_df_to_bigquery(df, table_name):
    dataset_id =  'crime-covid:crime_covid_data'
    dataset_ref = client.dataset(dataset_id)
    job_config = bigquery.LoadJobConfig()
    job_config.autodetect = True
    job_config.write_disposition = "WRITE_TRUNCATE"

    upload_table_name = 'crime_covid_data.'+str(table_name)

    load_job = client.load_table_from_dataframe(df, upload_table_name,
                                                job_config=job_config)

    print("Starting job {}".format(load_job))
```

```python
load_df_to_bigquery(df=date_dim, table_name='date_dim')
load_df_to_bigquery(df=loc_dim, table_name='loc_dim')
load_df_to_bigquery(df=offence_dim, table_name='offence_dim')
load_df_to_bigquery(df=victim_dim, table_name='victim_dim')
load_df_to_bigquery(df=suspect_dim, table_name='suspect_dim')
load_df_to_bigquery(df=NYC_COV19_fact, table_name='NYC_COV19_fact')
load_df_to_bigquery(df=NYPD_COMPL_fact, table_name='NYPD_COMPL_fact')
```



Schema Details **Preview**

Row	new_positives	total_number_of_tests	date_id	loc_id
1	4	31	20200309	50
2	2	31	20200311	50
3	3	36	20200312	50
4	10	99	20200313	50
5	8	80	20200314	50

NYPD_COMPL_fact

SHARE TABLE COPY TABLE DELETE TABLE EXPORT

Schema Details Preview

Row	date_id	loc_id	off_id	victim_id	suspect_id
1	20201122	50	5000	1000	100
2	20201106	50	5000	1000	100
3	20200222	50	5000	1000	100
4	20200430	50	5000	1000	100
5	20200925	50	5000	1009	100

date_dim

SHARE TABLE COPY TABLE DELETE TABLE EXPORT

Schema Details Preview

Row	date_id	day	month	year	weekday
1	20200103	3	1	2020	Friday
2	20200110	10	1	2020	Friday
3	20200117	17	1	2020	Friday
4	20200124	24	1	2020	Friday
5	20200131	31	1	2020	Friday

loc_dim

QUERY TABLE SHARE TABLE COPY TABLE DELETE TABLE EXPORT

Schema Details Preview

Row	loc_id	borough
1	50	BRONX
2	51	QUEENS
3	52	MANHATTAN
4	53	BROOKLYN
5	54	STATEN ISLAND

offence_dim

SHARE TABLE COPY TABLE DELETE TABLE EXPORT

Schema Details Preview

Row	off_id	law_cat_cd	ofns_desc
1	5000	FELONY	RAPE
2	5006	FELONY	CRIMINAL MISCHIEF & RELATED OF
3	5007	FELONY	ROBBERY
4	5008	FELONY	GRAND LARCENY OF MOTOR VEHICLE
5	5010	FELONY	GRAND LARCENY

suspect_dim

SHARE TABLE COPY TABLE DELETE TABLE EXPORT

Schema Details Preview

Row	suspect_id	susp_age_group	susp_race	susp_sex
1	163	65+	BLACK	F
2	168	65+	WHITE HISPANIC	F
3	172	65+	UNKNOWN	F
4	180	65+	ASIAN / PACIFIC ISLANDER	F
5	182	65+	BLACK HISPANIC	F

victim_dim

SHARE TABLE COPY TABLE DELETE TABLE EXPORT

Schema Details Preview

Row	victim_id	vic_age_group	vic_race	vic_sex
1	1078	UNKNOWN	BLACK	D
2	1104	18-24	BLACK	D
3	1105	25-44	BLACK	D
4	1114	45-64	BLACK	D
5	1082	UNKNOWN	WHITE	D