homework v

Ayush Kumar Shah

2020-10-02

Contents

Introduction

In this report, the two data sets introduced in the previous report are connected to the 311 data set, using `dplyr`. Before connecting them, several operations are performed on the datasets.

Also, few tables consisting of an extract of the data of each dataset as well as the final joined dataset are shown. Finally, a data dictionary for all the data in each dataset including the final joined dataset is also displayed. The connections between the columns will be shown in the next final report.

Reading the tidied nyc311 data

The nyc311 data was tidied and saved in the previous report. In this report, the saved tidied version of the nyc311 data set `tidied_nyc311.csv` is loaded.

A sample of 10,000 observations is also saved and used for initial computations but replaced later by the complete dataset.

```
nyc311_tidy <-fread("tidied_nyc311.csv",
              na.strings=c("","NA"))
# mini311<-nyc311[sample(nrow(nyc311),10000),]
# write_csv(mini311,"tidied_mini311.csv")
sample_tidy <- fread("tidied_mini311.csv", na.strings=c("","NA"))
```

Loading the additional datasets

```
nyc_popn <- read_csv('Projected_Population_2010-2040_-_Total_By_Age_Groups.csv')

## 2005-15 graduation
# nyc_grad <- read_csv('https://data.cityofnewyork.us/resource/qk7d-gecv.csv')
nyc_grad <- read_csv('2005_-_2011_Graduation_Outcomes_-_Borough_-_Ethnicity.csv')
names(nyc_grad)<-names(nyc_grad) %>%
stringr::str_replace_all("\\s", ".")
```

The two additional datasets introduced in the previous report are:

1. Projected Population 2010-2040 - Total By Age Groups: Source
2. 2005 - 2011 Graduation Outcomes - Borough - Ethnicity: Source

1. Projected Population 2010-2040 - Total By Age Groups

Projected total New York City population for five intervals from 2010 through 2040 by Borough, broken down by 18 age cohorts. (Age groups may not add up to the total due to rounding.)

This dataset is introduced so that the population information in the Borough can be known in correlation to the complaints in the nyc311 data.

2. 2005 - 2011 Graduation Outcomes - Borough - Ethnicity

Graduation results for all students by year; cohorts of 2001 through 2007 (Classes of 2005 through 2011). Graduation Outcomes as Calculated by the New York State Education Department. The New York State calculation method was first adopted for the Cohort of 2001 (Class of 2005).

Graduates are defined as those students earning either a Local or Regents diploma and exclude those earning either a special education (IEP) diploma or GED.

This dataset is introduced so that the educational status of the people in different Borough is available for further analysis in correlation with the nyc311 data.

Both of these datasets were obtained from the NYC OpenData. Also, they both contain the column `Borough` which makes them connectable to the `nyc311` data.

Connecting the datasets to nyc311

Converting the additional datasets to suitable form.

Before connecting the additional datasets to the nyc311, they are converted to suitable form.

Gathering the years in `nycpopn`

Table 1: Projected_Population 2010-2040 By AgeGroups

Borough	Age	2010	2015	2020	2025	2030	2035	2040
NYC Total	0-4	521990	535209	545778	547336	542426	540523	546426
NYC Total	15-19	539844	505783	492532	519298	535024	546062	546750
NYC Total	20-24	647483	646075	606203	591683	625253	643728	657403
NYC Total	25-29	736105	770396	763956	715824	698195	740437	762757
NYC Total	30-34	667657	707726	743916	740268	693684	675497	715486
NYC Total	35-39	592299	611239	649594	684249	682964	639237	621899

The years need to be gathered together in the `nyc_popn` data.

```
nyc_popn_tidy <-
  nyc_popn %>%
  gather('2010':'2040', key="Year", value="Population", convert = TRUE)

pander(head(nyc_popn_tidy, 10), caption = "Tidied Population data")
```

Table 2: Tidied Population data

Borough	Age	Year	Population
NYC Total	0-4	2010	521990
NYC Total	15-19	2010	539844
NYC Total	20-24	2010	647483
NYC Total	25-29	2010	736105
NYC Total	30-34	2010	667657
NYC Total	35-39	2010	592299

Borough	Age	Year	Population
NYC Total	40-44	2010	571825
NYC Total	45-49	2010	570273
NYC Total	50-54	2010	546204
NYC Total	55-59	2010	479661

Filtering by total population in each Borough

Since the `nyc311` has no information of Age group, so only the observations with total population of each Borough is filtered.

```
nyc_popn_tidy <- nyc_popn_tidy %>%
  filter(Age == "Total") %>%
  select(-Age)

pander(head(nyc_popn_tidy))
```

Borough	Year	Population
NYC Total	2010	8242624
Bronx	2010	1385108
Brooklyn	2010	2552911
Manhattan	2010	1585873
Queens	2010	2250002
Staten Island	2010	468730

Converting graduation dataset to suitable form

Table 4: 2005-2011 Graduation Outcomes - Borough

Borough	Cohort.Year	Cohort.Category	Demographic	Total.Cohort.Num	Total.Grads.Num
Bronx	2001	4 Year June	Asian	638	479
Bronx	2001	5 Year	Asian	638	528
Bronx	2001	6 Year	Asian	638	534
Bronx	2002	4 Year June	Asian	681	497
Bronx	2002	5 Year	Asian	681	564
Bronx	2002	6 Year	Asian	681	574

The `Cohort.Ctaegory` is parsed as number and added with the `Cohort.Year` to calculate the `Graduation.Year`. Then, only the relevant columns are selected. Finally, duplicates are removed by grouping the data according to the columns, `Graduation.Year` and `Borough`.

```
nyc_grad_tidy <- nyc_grad %>%
  mutate(Duration = parse_number(Cohort.Category),
         Graduation.Year = Cohort.Year + Duration) %>%
  select(Graduation.Year, Borough, c(5:7)) %>%
  group_by(Graduation.Year, Borough) %>%
  summarize_all(max)
pander(head(nyc_grad_tidy), split.table = "Inf")
```

Graduation.Year	Borough	Total.Cohort.Num	Total.Grads.Num	Total.Grads.Pct.of.cohort
2005	Bronx	6150	2261	75.10%
2005	Brooklyn	9382	4232	63.80%
2005	Manhattan	5739	2924	81.90%
2005	Queens	5335	2836	67.20%
2005	Staten Island	2304	1700	80.50%
2006	Bronx	6623	2922	82.80%

Check Borough's values

Since the tables will be connected by `Borough`, it is checked if there are any unspecified values in the column or mismatch in all the tables.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Borough	n
BRONX	1277987
BROOKLYN	2288036
MANHATTAN	1547810
QUEENS	1863259
STATEN ISLAND	434449

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Borough	n
Bronx	7
Brooklyn	7
Manhattan	7
Queens	7
Staten Island	7

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Borough	n
Bronx	7
Brooklyn	7
Manhattan	7
NYC Total	7
Queens	7
Staten Island	7

Since there are no Unspecified or null values, we are good to go. However, the values of Borough in the 2

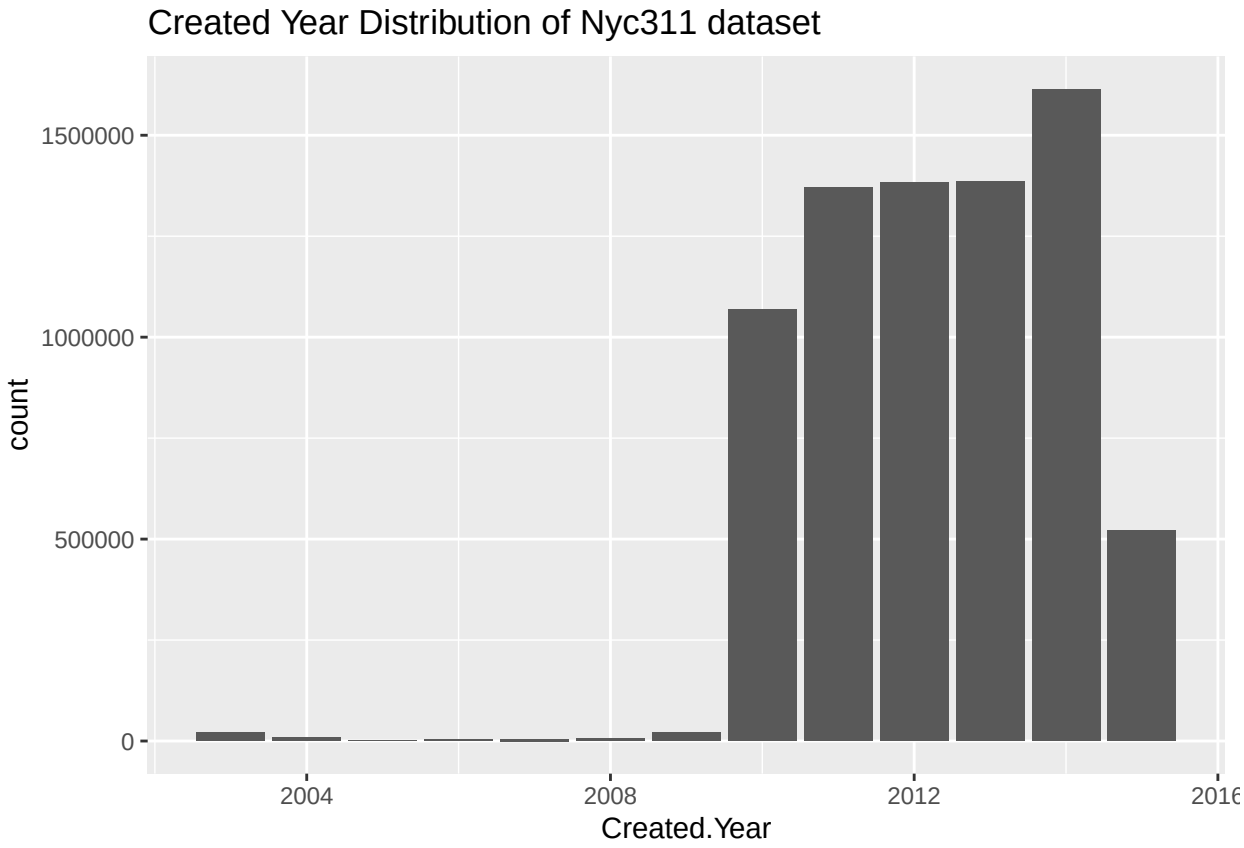
tables table must be converted to uppercase.

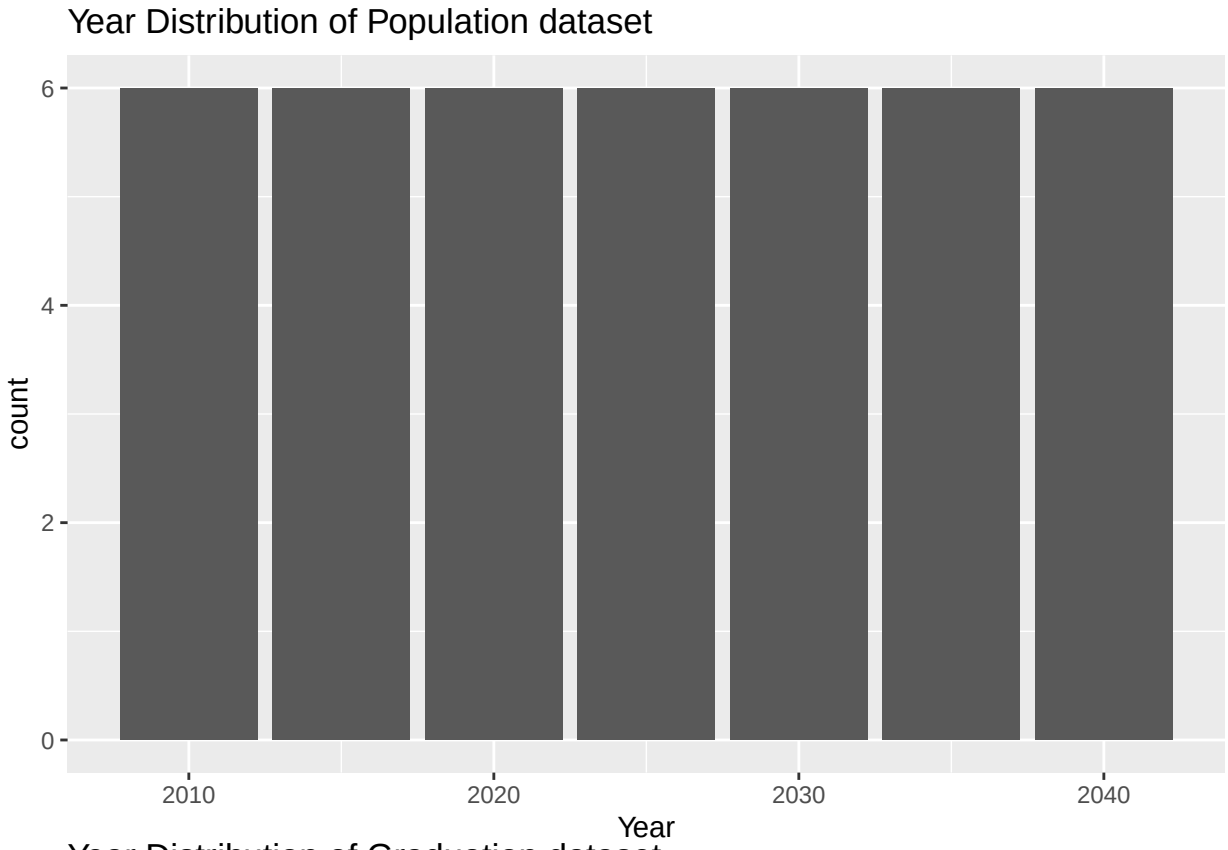
Converting to uppercase

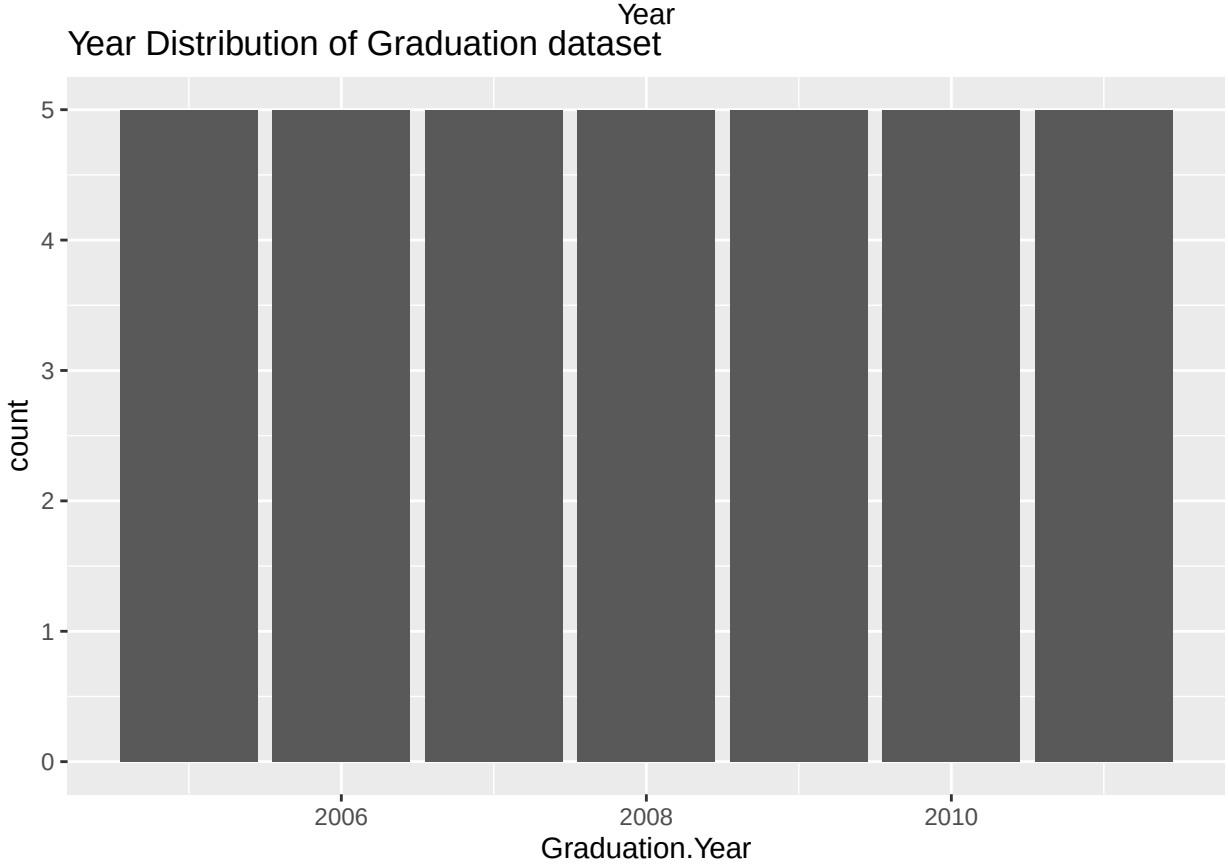
```
nyc_grad_tidy$Borough <- nyc_grad_tidy$Borough %>%
                            str_to_upper()
nyc_popn_tidy$Borough <- nyc_popn_tidy$Borough %>%
                            str_to_upper()
```

Viewing the year distribution of the tables

Since these tables contain different year values, we need to sample it based on a particular year. So, the year distributions of all three datasets are checked.







The results show that 2010 would be a good year to perform the analysis since it is common in all the tables.

Joining the datasets

Now that all the datasets are processed, they are ready to be connected by the columns `Year` and `Borough`. Note that the column name for `Year` differs in the three tables.

Also, 2010 is selected as the year for the analysis, hence all the tables are filtered accordingly.

```
nyc_combined <- nyc311_tidy %>%
  left_join(nyc_popn_tidy, by = c("Created.Year" = "Year", "Borough" = "Borough")) %>%
  left_join(nyc_grad_tidy, by = c("Created.Year" = "Graduation.Year", "Borough" = "Borough")) %>%
  filter(Created.Year == 2010)
```

View extracts of the data

Before joining

Table 9: Tidy Nyc311 data (continued below)

Created.Date	Created.Month	Created.Day	Created.Year	Created.Time
04/14/2015 02:14:40 AM	4	14	2015	02:14:40 AM
04/14/2015 02:10:12 AM	4	14	2015	02:10:12 AM
04/14/2015 02:03:01 AM	4	14	2015	02:03:01 AM
04/14/2015 02:02:40 AM	4	14	2015	02:02:40 AM
04/14/2015 02:00:04 AM	4	14	2015	02:00:04 AM
04/14/2015 01:52:15 AM	4	14	2015	01:52:15 AM

Table 10: Table continues below

Closed.Date	Agency	Agency.Name	Complaint.Type	Descriptor
04/14/2015 03:03:22 AM	NYPD	New York City Police Department	Vending	In Prohibited Area
NA	NYPD	New York City Police Department	Blocked Driveway	No Access
NA	NYPD	New York City Police Department	Noise - Street/Sidewalk	Loud Music/Party
NA	NYPD	New York City Police Department	Noise - Street/Sidewalk	Loud Talking
04/14/2015 02:47:33 AM	NYPD	New York City Police Department	Noise - Street/Sidewalk	Loud Talking

Closed.Date	Agency	Agency.Name	Complaint.Type	Descriptor
04/14/2015 02:11:10 AM	NYPD	New York City Police Department	Noise - Street/Sidewalk	Loud Talking

Table 11: Table continues below

Location.Type	Incident.Zip	Incident.Address	Cross.Street.1
Street/Sidewalk	10465	3775 EAST TREMONT AVENUE	RANDALL AVENUE
Street/Sidewalk	11234	1524 RYDER STREET	FLATLANDS AVENUE
Street/Sidewalk	11204	NA	NA
Street/Sidewalk	11211	361 METROPOLITAN AVENUE	HAVEMEYER STREET
Street/Sidewalk	10025	NA	NA
Street/Sidewalk	11205	NA	NA

Table 12: Table continues below

Cross.Street.2	Address.Type	City	Status	Due.Date	Borough
ROOSEVELT AVENUE	ADDRESS	BRONX	Closed	04/14/2015 10:14:40 AM	BRONX
AVENUE P	ADDRESS	BROOKLYN	Open	04/14/2015 10:10:12 AM	BROOKLYN
NA	INTERSECTION	BROOKLYN	Open	04/14/2015 10:03:01 AM	BROOKLYN
HAVEMEYER STREET	ADDRESS	BROOKLYN	Assigned	04/14/2015 10:02:40 AM	BROOKLYN
NA	INTERSECTION	NEW YORK	Closed	04/14/2015 10:00:04 AM	MANHATTAN
NA	INTERSECTION	BROOKLYN	Closed	04/14/2015 09:52:15 AM	BROOKLYN

Park.Facility.Name	School.Name	Latitude	Longitude	Created.Hour
Unspecified	Unspecified	40.83	-73.82	2
Unspecified	Unspecified	40.62	-73.94	2
Unspecified	Unspecified	40.62	-74	2
Unspecified	Unspecified	40.71	-73.96	2
Unspecified	Unspecified	40.8	-73.96	2
Unspecified	Unspecified	40.69	-73.96	1

Table 14: Projected Population 2010-2040 data (Tidy)

Borough	Year	Population
NYC TOTAL	2010	8242624
BRONX	2010	1385108
BROOKLYN	2010	2552911

Borough	Year	Population
MANHATTAN	2010	1585873
QUEENS	2010	2250002
STATEN ISLAND	2010	468730

Table 15: 2005-2011 Graduation Outcomes - Borough (Tidy)

Graduation.Year	Borough	Total.Cohort.Num	Total.Grads.Num	Total.Grads.Pct.of.cohort
2005	BRONX	6150	2261	75.10%
2005	BROOKLYN	9382	4232	63.80%
2005	MANHATTAN	5739	2924	81.90%
2005	QUEENS	5335	2836	67.20%
2005	STATEN ISLAND	2304	1700	80.50%
2006	BRONX	6623	2922	82.80%

After joining

Table 16: Final combined dataset (continued below)

Created.Date	Created.Month	Created.Day	Created.Year	Created.Time
12/31/2010 11:59:12 PM	12	31	2010	11:59:12 PM
12/31/2010 11:57:30 PM	12	31	2010	11:57:30 PM
12/31/2010 11:54:59 PM	12	31	2010	11:54:59 PM
12/31/2010 11:54:00 PM	12	31	2010	11:54:00 PM
12/31/2010 11:52:00 PM	12	31	2010	11:52:00 PM
12/31/2010 11:50:45 PM	12	31	2010	11:50:45 PM

Table 17: Table continues below

Closed.Date	Agency	Agency.Name	Complaint.Type	Descriptor
01/01/2011 02:14:44 AM	NYPD	New York City Police Department	Blocked Driveway	No Access
01/01/2011 12:46:41 AM	NYPD	New York City Police Department	Blocked Driveway	No Access
01/01/2011 08:03:26 AM	NYPD	New York City Police Department	Blocked Driveway	No Access
03/16/2011 09:36:00 AM	DOT	Department of Transportation	Street Light Condition	Lamppost Leaning
02/01/2011 12:45:00 PM	DEP	Department of Environmental Protection	Noise	Noise, Barking Dog (NR5)

Closed.Date	Agency	Agency.Name	Complaint.Type	Descriptor
01/01/2011 02:10:59 AM	NYPD	New York City Police Department	Blocked Driveway	Partial Access

Table 18: Table continues below

Location.Type	Incident.Zip	Incident.Address	Cross.Street.1
Street/Sidewalk	11212	632 THOMAS BOYLAND STREET	SUTTER AVENUE
Street/Sidewalk	11365	71-01 SUTTON PLACE	71 AVENUE
Street/Sidewalk	11236	8714 AVENUE A	EAST 87 STREET
NA	10006	86 TRINITY PLACE	RECTOR STREET
NA	11209	555 OVINGTON AVE	5 AVE
Street/Sidewalk	11368	34-22 101 STREET	34 AVENUE

Table 19: Table continues below

Cross.Street.2	Address.Type	City	Status	Due.Date	Borough
BLAKE AVENUE	ADDRESS	BROOKLYN	Closed	01/01/2011 07:59:12 AM	BROOKLYN
72 AVENUE	ADDRESS	FRESH MEADOWS	Closed	01/01/2011 07:57:30 AM	QUEENS
EAST 88 STREET	ADDRESS	BROOKLYN	Closed	01/01/2011 07:54:59 AM	BROOKLYN
TRINITY CEMETERY BOUNDARY	ADDRESS	NEW YORK	Closed	NA	MANHATTAN
6 AVE	ADDRESS	BROOKLYN	Closed	NA	BROOKLYN
35 AVENUE	ADDRESS	CORONA	Closed	01/01/2011 07:50:45 AM	QUEENS

Table 20: Table continues below

Park.Facility.Name	School.Name	Latitude	Longitude	Created.Hour
Unspecified	Unspecified	40.67	-73.91	23
Unspecified	Unspecified	40.73	-73.81	23
Unspecified	Unspecified	40.65	-73.92	23
Unspecified	Unspecified	40.71	-74.01	23
Unspecified	Unspecified	40.63	-74.02	23
Unspecified	Unspecified	40.76	-73.87	23

Population	Total.Cohort.Num	Total.Grads.Num	Total.Grads.Pct.of.cohort
2552911	11130	7322	83.30%
2250002	6867	4289	83.40%
2552911	11130	7322	83.30%
1585873	7766	5050	89.80%
2552911	11130	7322	83.30%

Population	Total.Cohort.Num	Total.Grads.Num	Total.Grads.Pct.of.cohort
2250002	6867	4289	83.40%

Data Dictionary

Table 22: Data dictionary for Projected Population 2010-2040 dataset

Column.Name	Description	DataType
`Borough`	Name of the New York City Borough	Text
`Age`	One of 18 Age cohorts like '0-4', '15-19', 'Total', and so on	Text
`Year`	Year in which the population is projected	Number
`Population`	The projected population value	Number

Table 23: Data dictionary for 2005-2011 Graduation Outcomes dataset

Column.Name	Description	DataType
`Borough`	Name of the New York City Borough	Text
`Graduation.Year`	The cohort's year of graduation	Number
`Total.Cohort.Num`	Number of students in the cohort	Number
`Total.Grads.Num`	Number of students who graduated in the cohort	Number
`Total.Grads.Pct.of.cohort`	Percentage of students who graduated in the cohort	Number

Table 24: Data dictionary for final Nyc combined dataset

Column.Name	Description	DataType
Created.Date	Date Service Request (SR) was created	Floating Timestamp
Created.Month	Month SR was created (1-12)	Number
Created.Day	Day of month SR was created (1-31)	Number
Created.Year	Year SR was created	Number
Created.Time	Time SR was created	Floating Timestamp
Created.Hour	Hour SR was created (0-24)	Number
Closed Date	Date SR was closed by responding agency	Floating Timestamp
Population	Total population of the Borough	Number
Total.Cohort.Num	Number of students in the cohort	Number
Total.Grads.Num	Number of students who graduated in the cohort	Number
Total.Grads.Pct.of.cohort	Percentage of students who graduated in the cohort	Number
Agency	Acronym of responding City Government Agency	Text
Agency Name	Full Agency name of responding City Government Agency	Text
Complaint Type	This is the first level of a hierarchy identifying the topic of the incident or condition.Complaint Type may have a corresponding Descriptor (below) or may stand alone.	Text
Descriptor	This is associated to the Complaint Type, and provides further detail on the incident or condition.Descriptor values are dependent on the Complaint Type, and are not always required in SR.	Text
Location.Type	Describes the type of location used in the address information	Text
Incident.Zip	Incident location zip code, provided by geo validation.	Text
Incident.Address	House number of incident address provided by submitter.	Text
Cross.Street.1	First Cross street based on the geo validated incident location	Text
Cross.Street.2	Second Cross Street based on the geo validated incident location	Text
Address.Type	Type of incident location information available.	Text
City	City of the incident location provided by geovalidation.	Text
Status	Status of SR submitted	Text
Due.Date	Date when responding agency is expected to update the SR. This is based on the Complaint Type and internal Service Level Agreements (SLAs).	Floating Timestamp
Resolution.Description	Describes the last action taken on the SR by the responding agency. May describe next or future steps.	Text
Borough	Provided by the submitter and confirmed by geovalidation.	Text
Park.Facility.Name	If the incident location is a Parks Dept facility, the Name of the facility will appear here	Text
Latitude	Geo based Lat of the incident location	Number
Longitude	Geo based Long of the incident location	Number

Conclusion

Hence, the two additional datasets were joined with the `nyc311` data by converting the datasets into appropriate form (tidying, grouping and removal of redundant columns), performing a left join by columns `Year` and `Borough` and finally filtering it by the `Year==2010` constraint.

The data extracts of tables before and after joining were were also displayed. Finally, the data dictionary for of the two additional datasets and the final combined dataset was displayed.