NYC Transport and Health Status

Objective

The objective is to build a database to assess the associations of individual health status and transportation in the New York City.

Background

Transportation is an important part of everyone’ s life, which is closely linked with another important component, public health. The connections between public health and transportation are varied and well documented in peer-reviewed journals in both the public health and transportation area. Mainly scholars consider that transportation has major impact on health and the development may have enhanced and health and increased health risks. For instance, A study in 2010, calculated that the costs of medical care and lost productivity associated with motor vehicle crashes exceeded $99 billion in 2005. Another study shows that the transportation did improve access for a person type of travel, since on average one walking briskly to a transit stop could be count as physical activity, which lower the risk of obesity, diabetes and heart disease. However, the U.S. Department of Health and Human Services’ 2008 Physical Activity Guidelines for Americans announced that the beneficial actually is as little as 60 minutes a week, research shows that at least 150 minutes a week will consistently reduce the risk of many chronic diseases and other adverse health outcomes.

Research Questions

We narrow down the objective into the following five questions. The relationship between patients’ hospital visiting frequency and his/her public transportation circumstances; The relationship between patients’ disease and traffic condition; the relationship between Hispanic male patients’ weights and his/her traffic condition; Whether patients at different ages have a preference for taking public transportation; The relationship between patients’ drug visiting frequency and traffic safety condition.

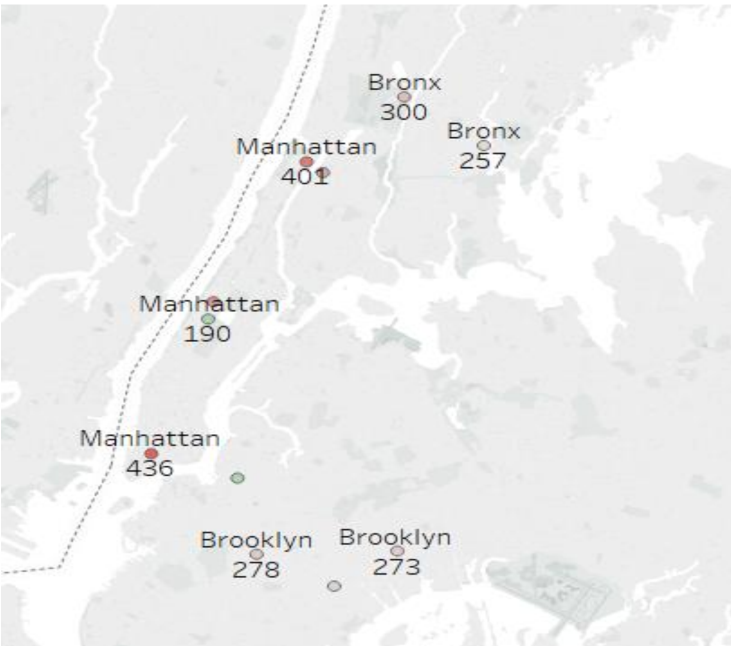
Dataset Description

In terms of external datasets, we searched many kinds of data and made some visualizations from websites, such as NYC Open Data and New York City Department of Tran sportation. In details, necessary variables for database building were extracted from a number of datasets and categorized into three main types. Specified tabled formed are listed below: 1). Transportation methods: bus stop shelter, bike parking shelter and subway station locations; 2). Traffic conditions: traffic speed, traffic volume counts, bridge rating, street pavement rating and vehicle classification counts and 3). Injuries. Besides, two more code tables were formed: ValueAsStringCode and VehicleTypeCode.

For the patients' data, we choose the EHR dataset as our main dataset here. The electronic health record(EHR) is the systematized collection of patient and population electronically-stored health information in a digital format. As we cannot get the true dataset, we use the OMOP which includes fake patient id with true column names. Variables of interests were selected from the OMOP dataset and were reorganized for the new dataset: 1)Patient: a table shows the basic information of patients like age, gender and the type of diseases, 2),DrugExposer: a table shows the details information about buying drugs from hospital, contains the column of exact date and DrugExposerID, 3)VisitOccurence: a table shows the details information about visiting occurence for each patients, contains the column like the exact date, and the total number of visit times can be calculated here, as well as 4)Observation: a table contains the details about patients' health status like the value of blood pressure and so on, which are parts of clinical notes.

There are also two codes table here to show the exact type of vehicle in VehicleClassification table.

Based on the data of traffic volume, one map is generated as below which shows the total traffic volume during 24 hours in different area of NYC.



API

Because the address of the patient in EHR dataset only show the special address which includes the street name, we need to transform the special address into FIPS code to help us do further analysis. One of the APIs we used here is Google Maps Geocoding API, it can convert one special address into longitude and latitude, like for “W 96 ST”, the result will be “(40.79, -73.97)”. Another API used here is the block API from Federal Communication Commision, which can convert longitude and latitude into FIPS code. Based on both APIs, we can convert all addresses of patient into FIPS code to do further analysis.

E-R Diagram

Based on the purpose of our project which is analyzing the relationship between health status and transportation in New York City, the patient table is put in the center of the E-R diagram and connected with other three related table: DrugExposure, Observation and VisitOccurence. For this part, we suppose that each patient have many times to buy drug from the hospital,each patient needs to visit hospital at least one time, as well as each patient has many clinic notes. This relationship is shown as “one to many”.

Another part of E-R diagram is the transportation part, which is based on the location table. There are four tables connected to location table, which are PavementRating, Injury, Station, Road tables. We also suppose that each location has many stations (bus, bike and subway), injuries, roads and pavements. This is assumption is the reason why the relationship among all those tables and location is “one to many”.

In order to connect the patient and transportation part, we connect the location table and patient table. The suppose here is one patient has one exact FIPS code, and each FIPS has many patients. So the relationship here is one location to many patients.

The all tables and the relationships are shown as the E-R diagram.

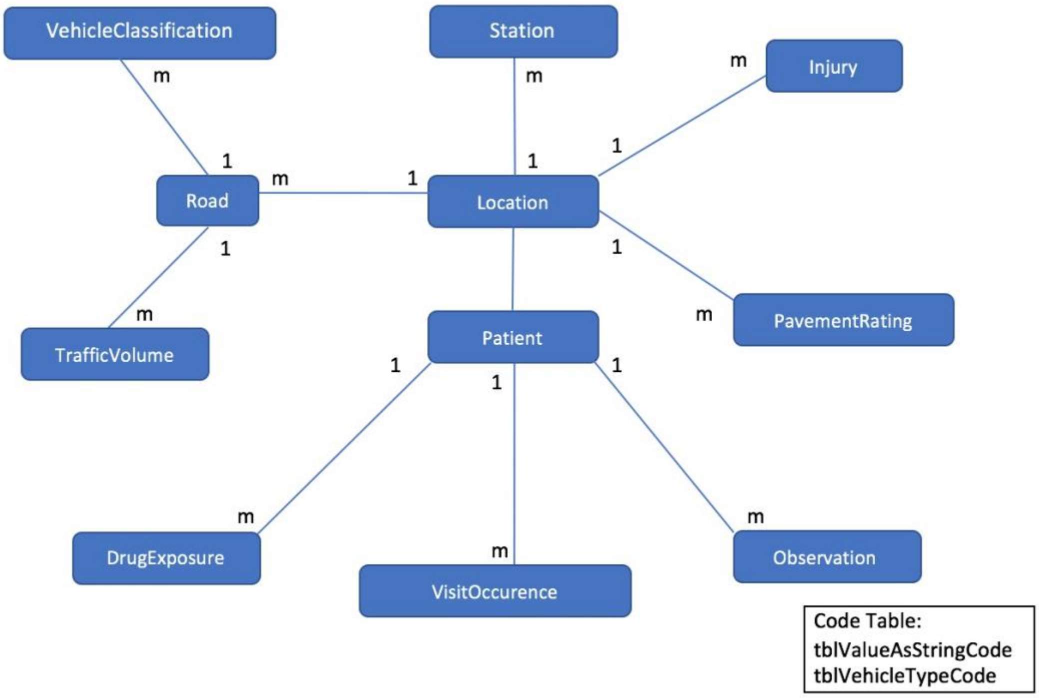


Table Schema

tblDrugExposure (DrugExposureID, PatientID, DrugExposureStartDate)

tblInjury (InjuryID, LocationID, PedInjurie, BikeInjuri, MVOInjurie, LocationName)

tblLocation (LocationID, FIPS, Borough)

tblObservation (ObservationID, PatientID, ObservationDate, ValueAsStringCode, ValueAsNumber)

tblPatient (PatientID, LocationID, YearOfBirth, GenderSourceValue, RaceSourceValue, EthnicitySourceValue, CconceptName)

tblPavementRating (PavementRatingID, LocationID, PavementID, Length, Rating, Borough)

tblRoad (RoadID, LocationID, LOCATION, Speed)

tblstation (ID, LocationID, TypeCode, Count)

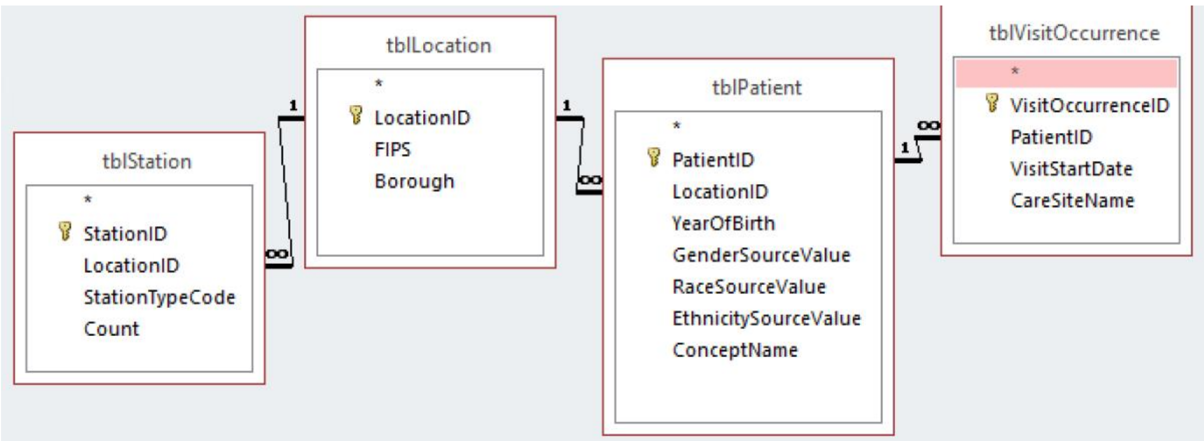
tblTrafficVolume (TrafficVolumeID, RoadID, TrafficVolumeCounts, TimeTypeCode, Address, Borough) tblVehicleClassification (<u>VehicleClassificationID</u>, RoadID, RoadNumber, Count, VeichelTypeCode)

tblVisitOccurrence (<u>VisitOccurrenceID</u>, PatientID, VisitStartDate, CareSiteName)

Query

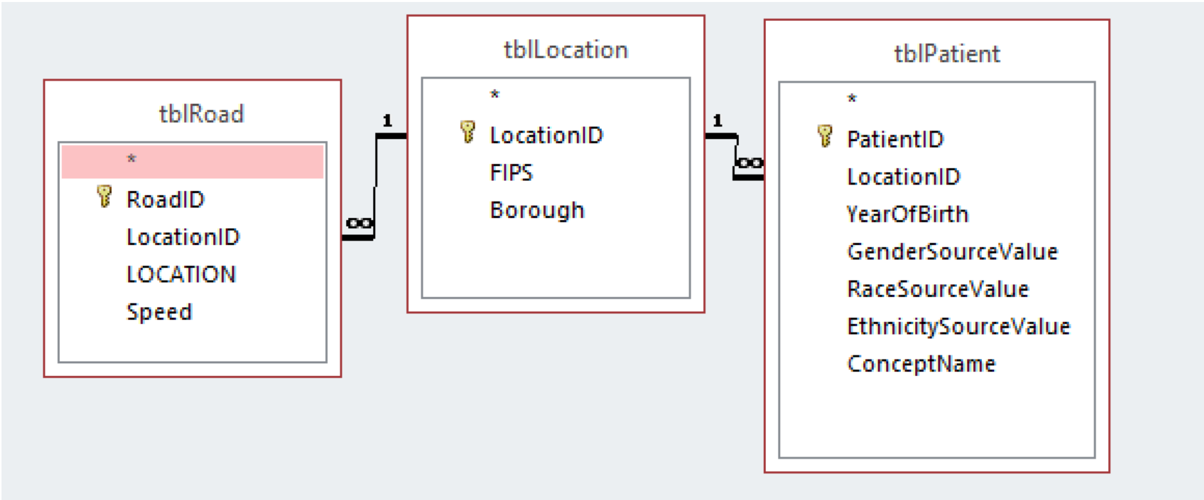
1. Provide a list of the frequency (count) of patient visiting the NYP, Weill Cornell,MountSaina and MSK during 2005 and 2008 and the total amount of bike shelters, bus stops and subway stations. This query can help to analyze the relationship between patients' hospital visiting frequency and his/her public transportation circumstances.

To explore the relationship between patients' hospital visiting frequency and their transportation conditions, four tables (tblPatient, tblLocation, tblStation and tblVisitOcurrence) were selected and joined. By LocationID, we can link tblStation, tblLocation and tblPatient, and output the number of bus stops, subway stations and bike shelters for a location. Meanwhile, we can count the patient's hospital visiting frequency by counting VisitOcurrenceID and the visiting frequency for each location can be calculated. In this way, the relationship of interest could be analyzed.



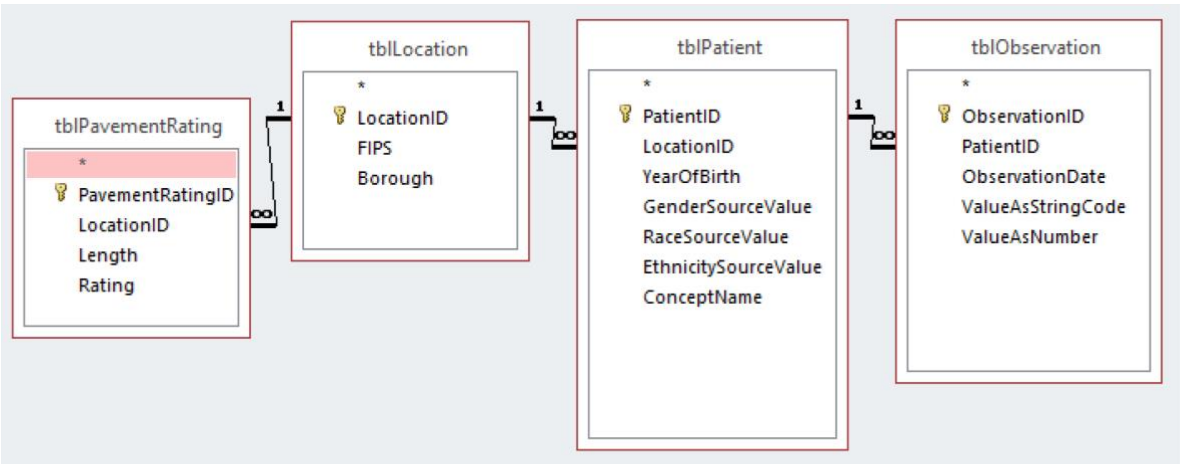
2. Provide a list of whether patients have mental disease and the average traffic speed. This query is to analyze the relationship between patients' mental health and traffic condition.

This query is to build up the association of patients' mental health with traffic condition. Three tables tblRoad, tblLocation and tblPatient are selected and mental health was selected by specifying ConceptName. Linking these three tables by LocationID, the condition of mental health could be assessed by evaluating the speed of this particular address.



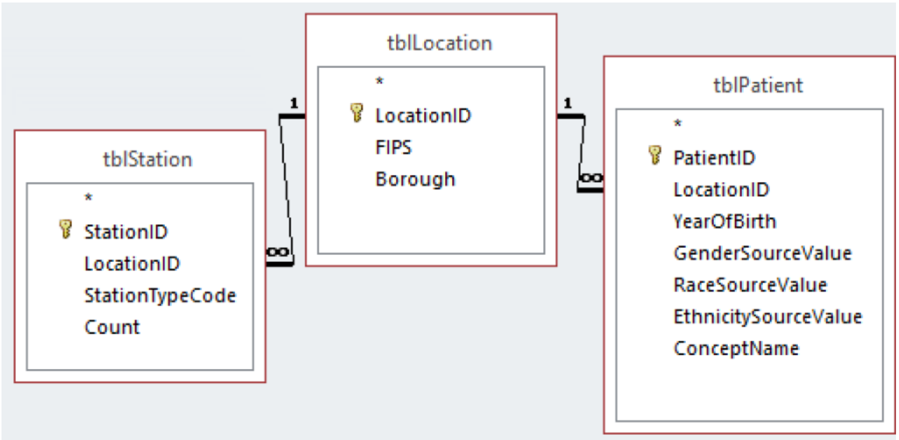
3. Provide a list of Hispanic male patients' weights and pavement rating. This query can help to analyze the relationship between Hispanic male patients' weights and his/her traffic condition.

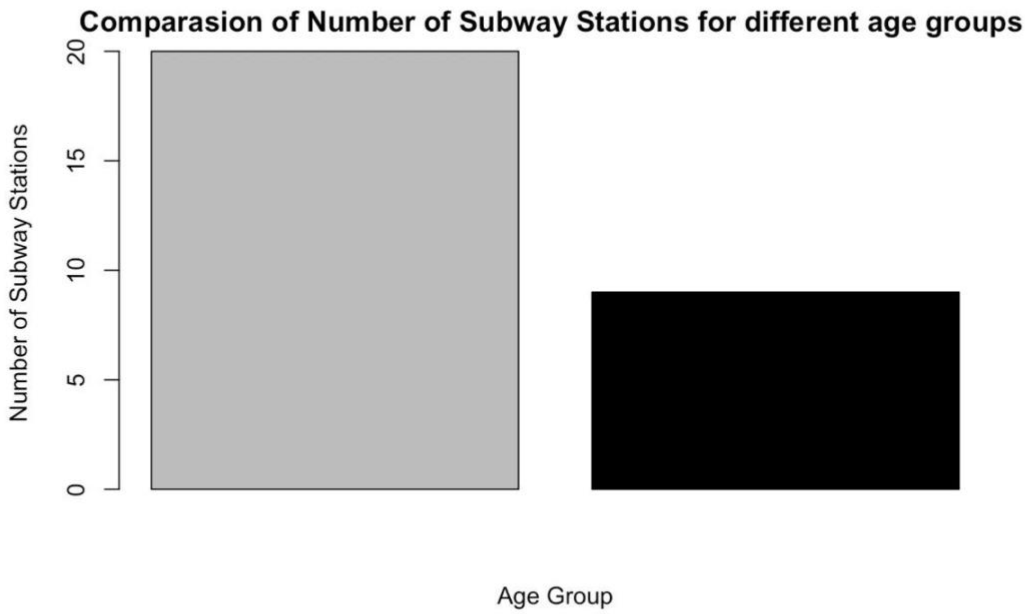
By specifying ValueStringCode = "Weight" in tblObservation, the query output weights of all patients. Linking tblObservation and tblPatient by PatientID, we could select patients whose GenderSourceValue = "Male" and RaceSourceValue = "Hispanic" and calculate the average weight for each location. More things to do is linking other two tables by LocationID, and thus form a relationship between Hispanic male patients' weights and his/her traffic condition.



4. Provide a list of subway and patients in different age group. This query is to analyze whether patients at different ages have a preference for taking public transportation.

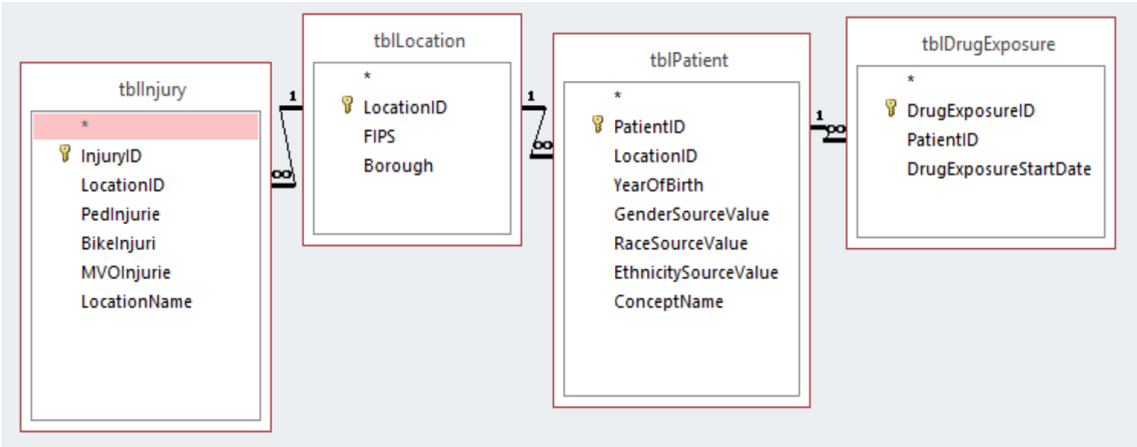
By specifying StationTypeCode = "Subway" and counting its number grouping by LocationID, we can link the table tblStation and tblPatient. In tblPatient, we define patients whose YearOfBirth < 1979 are "Old" group, and the rest are "Young" group. Comparing the number of subway stations by different groups of age, the consideration query could be analyzed.





5. Provide a list of the frequency (count) of patient visiting the drug service providers and Motor Vehicle injuries. This query is to analyze the relationship between patients' drug visiting frequency and traffic safety condition.

This query is designed to link Injury table and DrugExpouse table in order to provide information for connection between injury type and the drug service visiting times. To do so, we use LocationID and PatientID to link these tables. In the injury table, we count the motor vehicle injuries grouping b y LocationID. In the DrugExpouse table, we count the variable “DrugExposureStratDate” to record the times of visit for each patient. Total drug service visiting times for each location can be calculated to assess their relationship.



Linear Regression

Based on the built-up dataset and designed queries, clients can analysis many associations. For instance, qryPublicTransportationVisiting provide the information of the total amount and each amount of bike shelters, bus stops and subway stations, stations’ location, Patient information and the time of patient visit the hospital. Therefore, client can run a sample linear regression to check if there is an association between the visit hospital frequency of a patient and the his/her public transportation circumstances. The following is a sample:

```
Call:
lm(formula = n_visit ~ Bus + Subway, data = qdat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.17308 -0.09615 -0.05769 -0.05769  0.90385 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  1.07692    0.42106   2.558   0.0285 *
Bus         -0.05769    0.32024  -0.180   0.8606  
Subway       0.03846    0.10212   0.377   0.7143  
```

Since our data set is not adequate, the result didn’ t the significant. However, among with the increasing of the scale of the dataset. The client can seek the true association between patients’ hospital visiting frequency and his/her public transportation circumstances.

Conclusion

For this project, we built a database that linked traffic burden, traffic safety and transportation tools to the EHR data that contains information of patients' health condition, drug service visiting and hospital visiting. And according to this database, we can extract data to analyze the relationship between NYC transportation and residents' health status.

Discussion

For building up the tables of EHR data, we just assumed that each patient had one disease. However, it is not always the case. It can be the case that one patient has multiple diseases or one patient has no disease. To include this kind of information, it is better for us to add one more table to the relationship with the tblPatient.

In our project, one patient matches to one disease. In real life, the relationship between patients and disease is one to many. Therefore, A additional type disease code table may could be built. Another thing is the scale of dataset could be enhanced. Since including more data, the higher accuracy of analysis will be.

Reference

OHDSI/ CommonDataModel, by clairblacker, 2017. https://github.com/OHDSI/CommonDataModel/wiki/DRUG_ERA. Accessed 20 Jun, 2017.

Public Road. Publication Numver : FHWA-HRT-13-004
https://www.fhwa.dot.gov/publications/publicroads/13mayjun/05.cfm. Issue No: Vol. 76 No. 6. May/June 2013

Trandsport and public health. https://www.eea.europa.eu/signals/signals-2016/articles/transport-and-public-health
29 Jun 2016

Appendix

The code below shows how the block API from Federal Communication Commision works, and based on exact address, we can get the longitude and latitude in the results

```
> gGeoCode <- function(address,verbose=FALSE) {
+   if(verbose) cat(address,"\n")
+   u <- construct.geocode.url(address)
+   doc <- getURL(u)
+   x <- fromJSON(doc,simplify = FALSE)
+   if(x$status=="OK") {
+     lat <- x$results[[1]]$geometry$location$lat
+     lng <- x$results[[1]]$geometry$location$lng
+     return(c(lat, lng))
+   } else {
+     return(c(NA,NA))
+   }
+ }
> construct.geocode.url("W 96 ST")
[1] "http://maps.google.com/maps/api/geocode/json?address=W%2096%20ST&sensor=false"
>
> x <- gGeoCode("W 96 ST")
> x
[1]  40.79406 -73.97036
```

.

The code below shows how the google maps geocoding API works in R which can conver t the longitude and latitude into the exact FIPS code.

```
> for( i in 1:21){
+   x<- mapply(latlong2fips,latitude=a$lat[i], longitude=a$long[i])
+   return(x)
+ }
> for( i in 1:21){
+   x<- mapply(latlong2fips,latitude=a$lat[i], longitude=a$long[i])
+   print(x)
+ }
[1] "360050281003001"
[1] "360050319001000"
[1] "360050449022000"
[1] "360470331003001"
```

The output of Query1:

qryPublicTransportationVisiting					
PatientID	LocationID	Bus	Bike	Subway	n_visit
1	1	1	1	2	1
2	2	1	1	1	1
3	3	1	1	2	2
4	4	1	1	2	1
6	6	1	1	1	1
9	9	1	1	4	1
10	10	1	1	1	1
11	11	1	1	2	1

12	12	1	1	2	1
13	13	1	1	1	1
14	14	1	1	1	1
15	15	1	1	1	1
16	16	2	1	1	1
19	19	1	1	1	2
21	20	1	1	1	1

The output of Query2:

qryDrugInjuryRelationship			
PatientID	LocationID	n_injuries	n_drug
1	1	1	2
2	2	3	2
3	3	1	3
4	4	5	3
5	5	3	3
6	6	5	3
7	7	1	3
8	8	3	3
9	9	1	3
10	10	3	3
11	11	3	2
12	12	2	2
13	13	5	2
14	14	6	2
15	15	4	2
16	16	2	2
17	17	5	2
18	18	6	2
19	19	5	2
20	20	8	2
21	20	8	2

The output of Query3:

qryPavementWeight

LocationID	PatientID	Rating	ValueAsNumber
1	1	0	114
2	2	6	115
2	2	6	120
3	3	8	116
4	4	7	117
5	5	8	117
5	5	8	118
6	6	3	119
7	7	7	120
8	8	9	145
8	8	9	121
9	9	6	122
10	10	9	123
11	11	8	124
11	11	8	133
12	12	8	125
13	13	5	126
14	14	8	123
14	14	8	127
15	15	9	128
16	16	8	129
16	16	9	129
17	17	6	130
18	18	8	131
19	19	8	132
20	20	6	133
20	21	6	127

The output of Query4:

qryPublicTransportAgeGroup			
LocationID	PatientID	Subway	Age_Group
1	1	2	Old
2	2	1	Old
3	3	2	Young
4	4	2	Old
5	5	1	Young

6	6	1	Old
7	7	1	Old
8	8	1	Old
9	9	4	Young
10	10	1	Old
11	11	2	Old
12	12	2	Old
13	13	1	Old
14	14	1	Old
15	15	1	Old
16	16	1	Young
17	17	1	Old
18	18	1	Old
19	19	1	Young
20	20	1	Old
20	21	1	Old

The output of Query5:

qrySpeedMental			
LocationID	PatientID	avg_speed	Mental_disease
1	1	49.08	No
1	1	51	No
2	2	55.3	No
3	3	28.58	No
4	4	37.28	No
5	5	26.1	No
6	6	30.9	No
7	7	43.3	Yes
8	8	17.39	Yes
9	9	15.53	Yes
10	10	22.99	No
11	11	20.5	No
12	12	27.3	No
13	13	18.2	No
14	14	19.4	No
15	15	27.9	No
16	16	39.4	No
17	17	28.8	No
18	18	42.25	No

19	19	52.82	Yes
20	20	37.9	Yes
20	21	37.9	Yes