K-Nearest Neighbors

NAME: SUSHMA YADAV DUVVI
CWID: 50309074

Project Scope: In this assignment we will be using survey data of preferred programming languages in data science in a number of cities. We will use these results to predict the favorite programming languages for places that were NOT part of the survey.

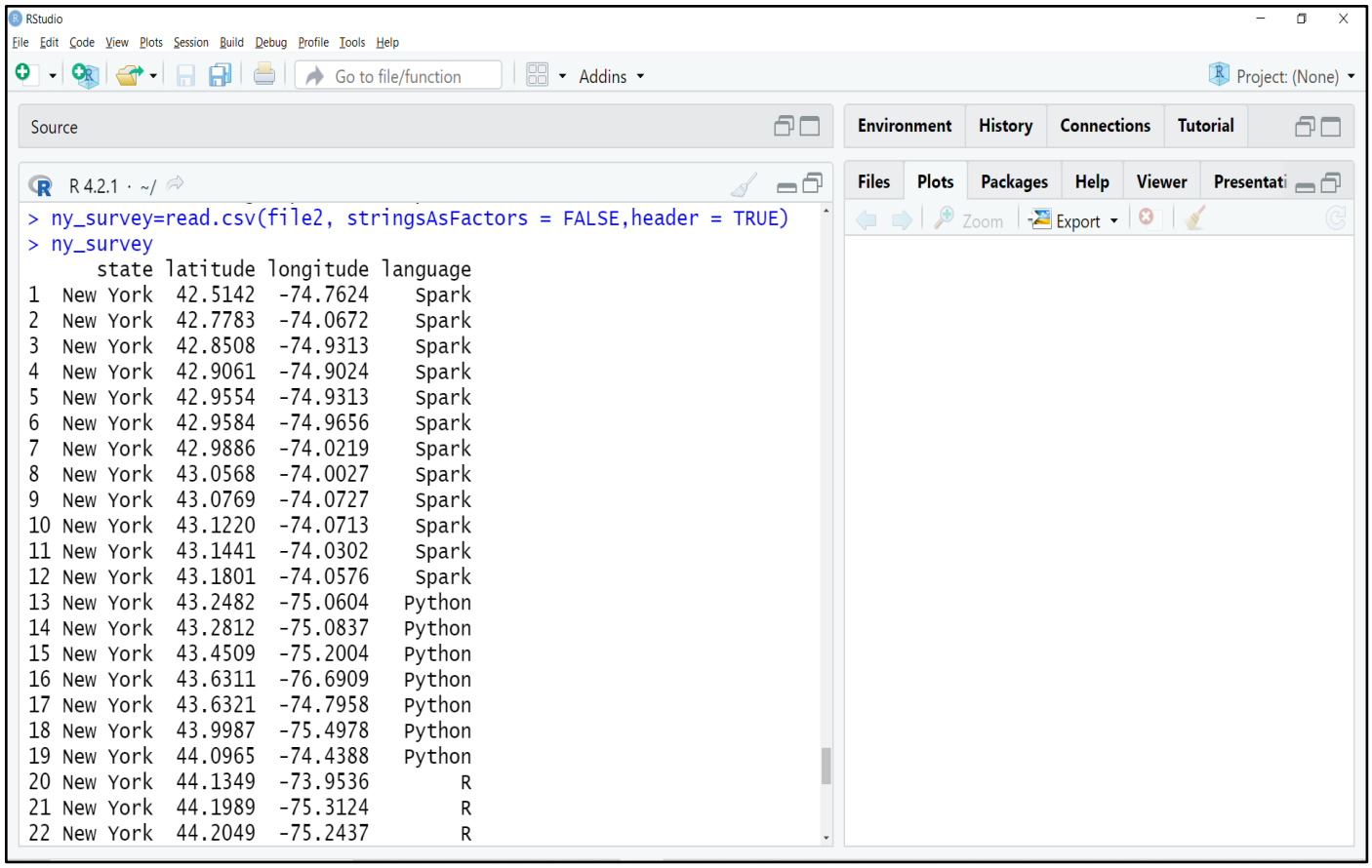
There are two data files for this assignment. The longitude and latitude of states and the survey data. You will be using longitude and latitude data to draw the maps of states and use survey data to display the survey results on the map. For this assignment we will be using the survey data of the New York state.**20**

PART 1 : Load data files longitude_latitude.csv **and** ny_survey.csv

The *language* column of the survey data contains the preferred programming language at the city located at *longitude* and *longitude*.

```
> file1="C:/DATASETS/longitude_latitude.csv"

> file2="C:/DATASETS/ny_survey.csv"

> Longitude_Latitude=read.csv(file1, stringsAsFactors = FALSE,header = TRUE)

> Longitude_Latitude

> ny_survey=read.csv(file2, stringsAsFactors = FALSE,header = TRUE)

> ny_survey
```

```
> file1="C:/DATASETS/longitude_latitude.csv"
>
> file2="C:/DATASETS/ny_survey.csv"
>
> Longitude_Latitude=read.csv(file1, stringsAsFactors = FALSE,header
 = TRUE)
> Longitude_Latitude
         state latitude longitude
1      Alabama  35.0041  -88.1955
2      Alabama  34.9918  -85.6068
3      Alabama  32.8404  -85.1756
4      Alabama  32.2593  -84.8927
5      Alabama  32.1535  -85.0342
6      Alabama  31.7947  -85.1358
7      Alabama  31.5200  -85.0438
8      Alabama  31.3384  -85.0836
9      Alabama  31.2093  -85.1070
10     Alabama  31.0023  -84.9944
11     Alabama  30.9953  -87.6009
12     Alabama  30.9423  -87.5926
13     Alabama  30.8539  -87.6256
14     Alabama  30.6745  -87.4072
15     Alabama  30.4404  -87.3688
16     Alabama  30.1463  -87.5240
17     Alabama  30.1546  -88.3864
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Environment History Connections Tutorial
R 4.2.1 · ~/
Files Plots Packages Help Viewer Presentati
Zoom
Export
> ny_survey=read.csv(file2, stringsAsFactors = FALSE,header = TRUE)
> ny_survey
state latitude longitude language
1 New York 42.5142 -74.7624 Spark
2 New York 42.7783 -74.0672 Spark
3 New York 42.8508 -74.9313 Spark
4 New York 42.9061 -74.9024 Spark
5 New York 42.9554 -74.9313 Spark
6 New York 42.9584 -74.9656 Spark
7 New York 42.9886 -74.0219 Spark
8 New York 43.0568 -74.0027 Spark
9 New York 43.0769 -74.0727 Spark
10 New York 43.1220 -74.0713 Spark
11 New York 43.1441 -74.0302 Spark
12 New York 43.1801 -74.0576 Spark
13 New York 43.2482 -75.0604 Python
14 New York 43.2812 -75.0837 Python
15 New York 43.4509 -75.2004 Python
16 New York 43.6311 -76.6909 Python
17 New York 43.6321 -74.7958 Python
18 New York 43.9987 -75.4978 Python
19 New York 44.0965 -74.4388 Python
20 New York 44.1349 -73.9536 R
21 New York 44.1989 -75.3124 R
22 New York 44.2049 -75.2437 R

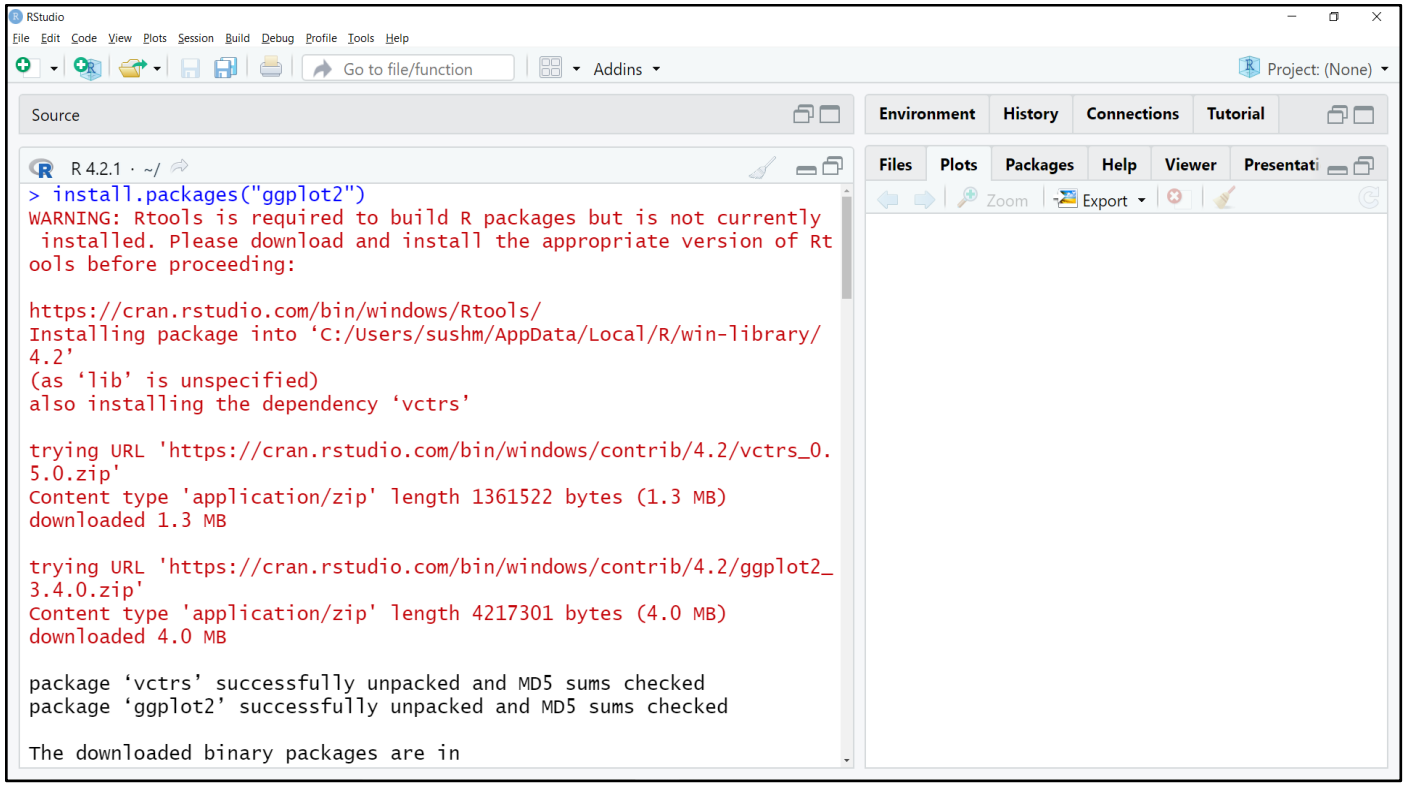
<u>PART 2 : Using the *longitude* and *latitude* data in the file *longitude_latitude.csv* plot the map of the California state.</u>

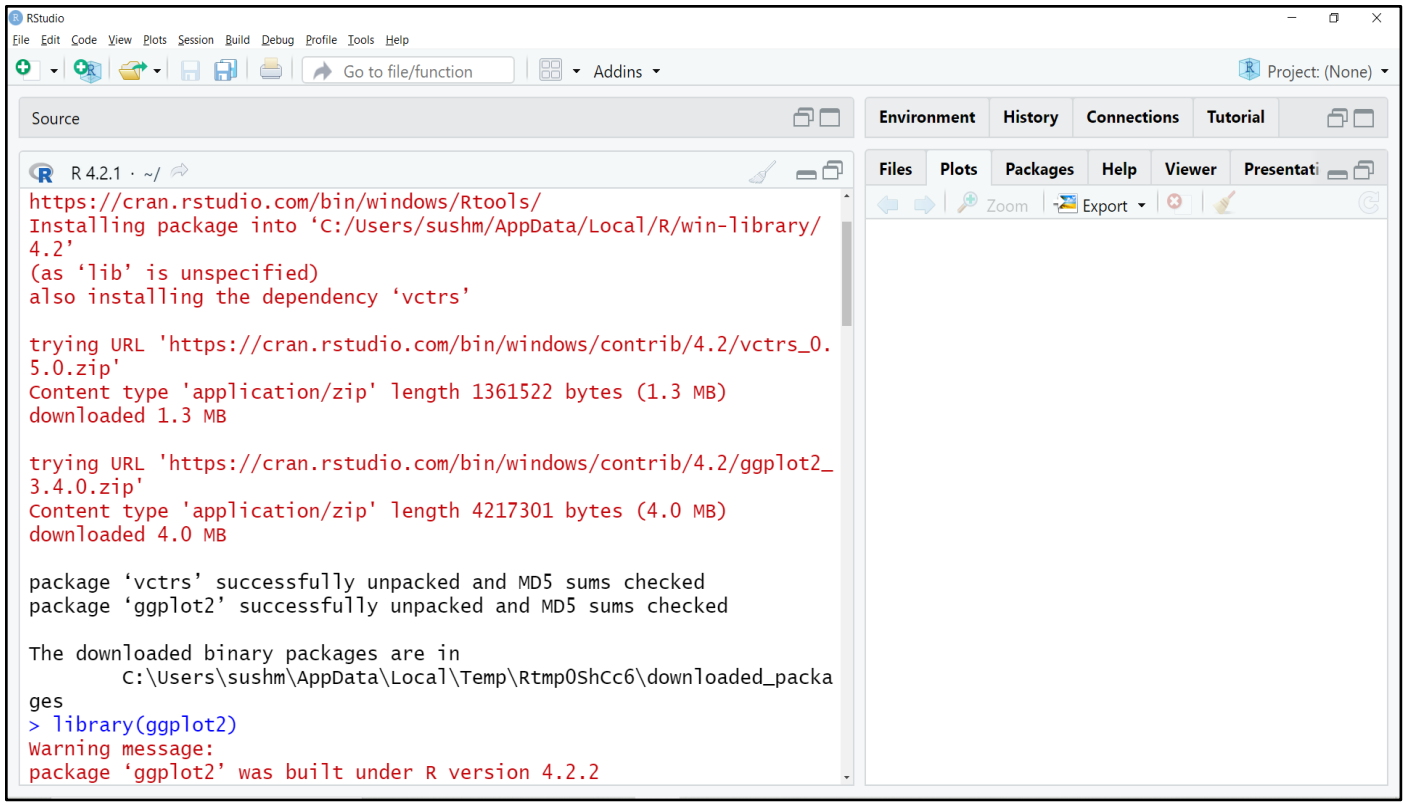
> install.packages("ggplot2")

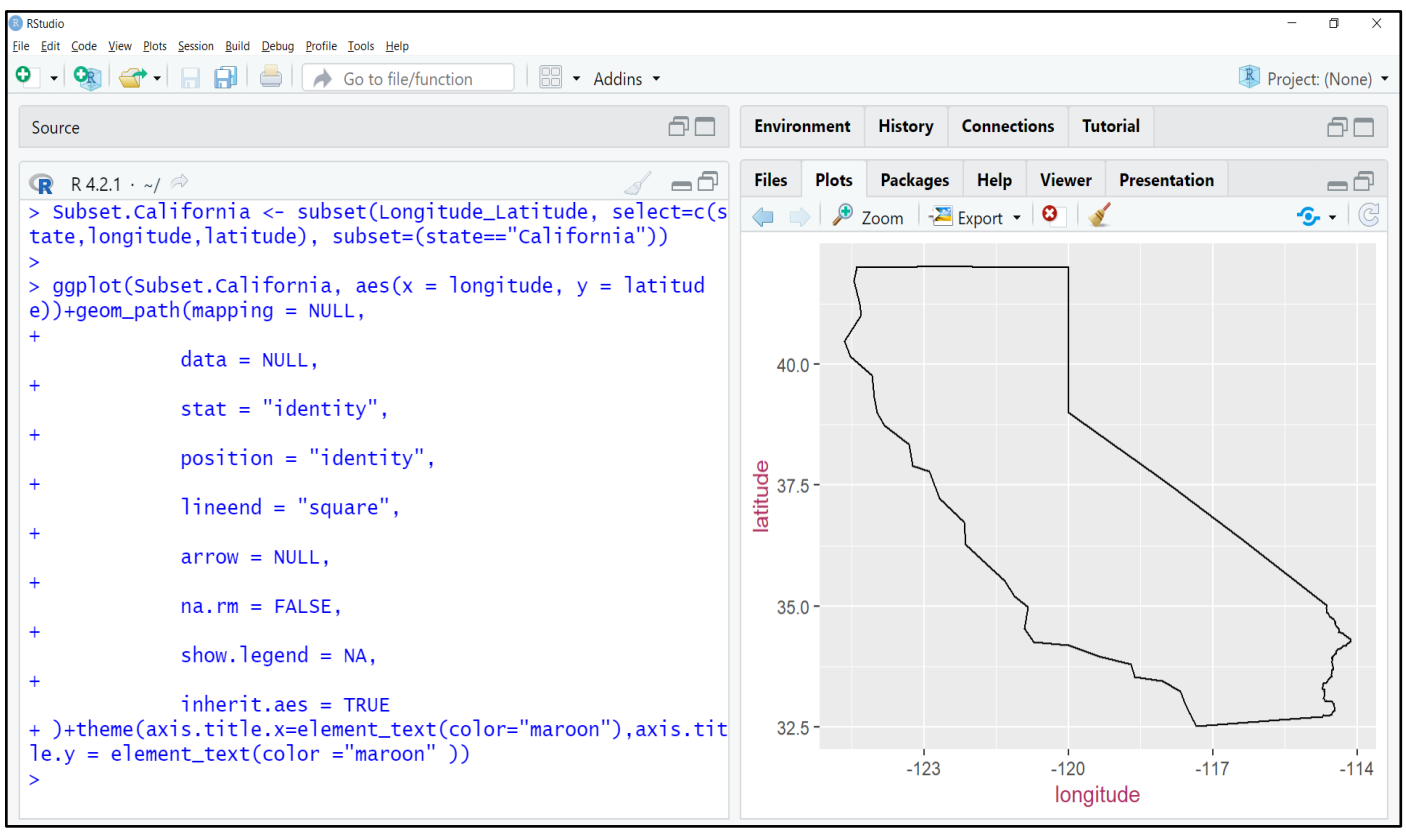
> library(ggplot2)

> Subset.California <- subset(Longitude_Latitude, select=c(state,longitude,latitude), subset=(state=="California"))

> ggplot(Subset.California, aes(x = longitude, y = latitude)) + geom_path(mapping = NULL, data = NULL, stat = "identity", position = "identity", lineend = "square", arrow = NULL, na.rm = FALSE, show.legend = NA, inherit.aes = TRUE) + theme(axis.title.x=element_text(color="maroon"),axis.title.y = element_text(color ="maroon"))





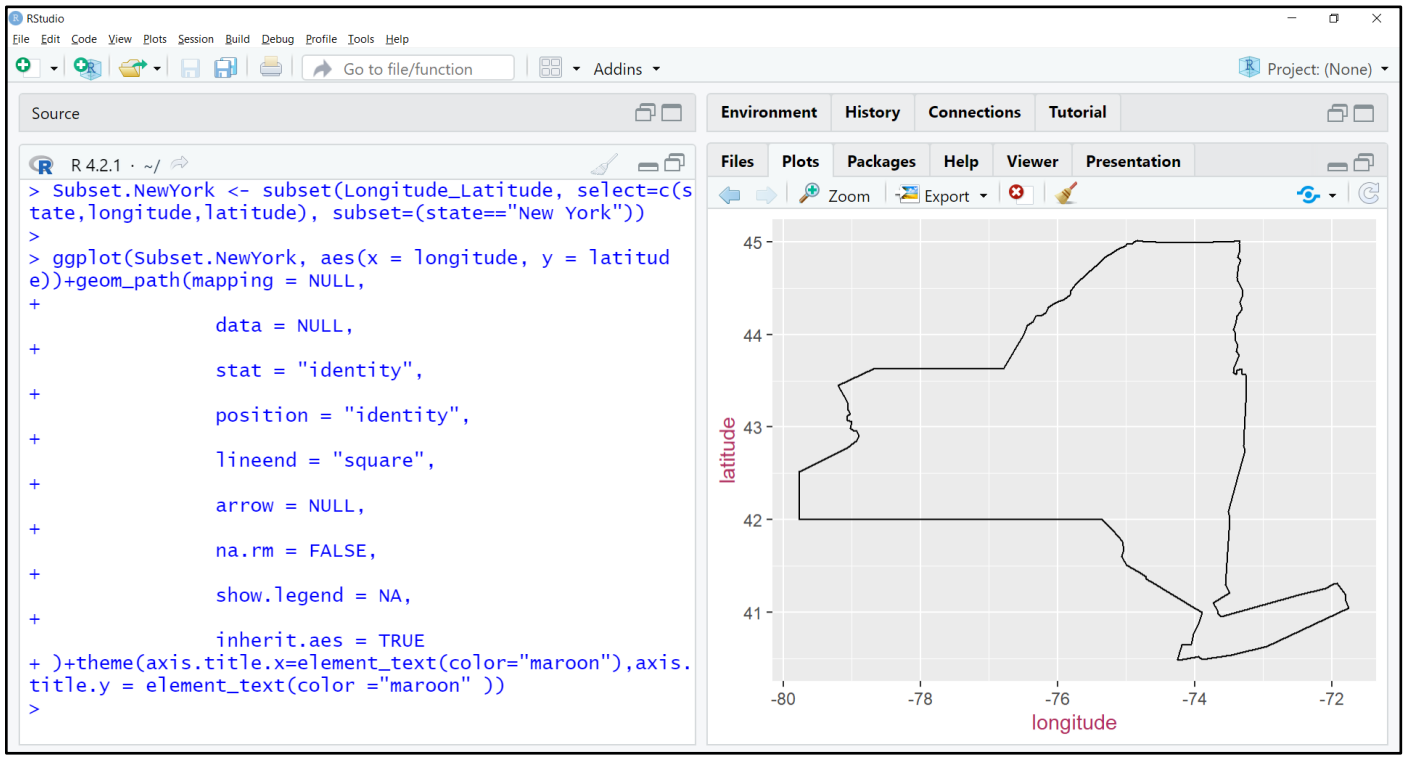


PART 3: Draw map and display data

<u>Part A</u>: Using the *longitude* and *latitude* data in the file *longitude_latitude.csv* plot the map of the New York state.

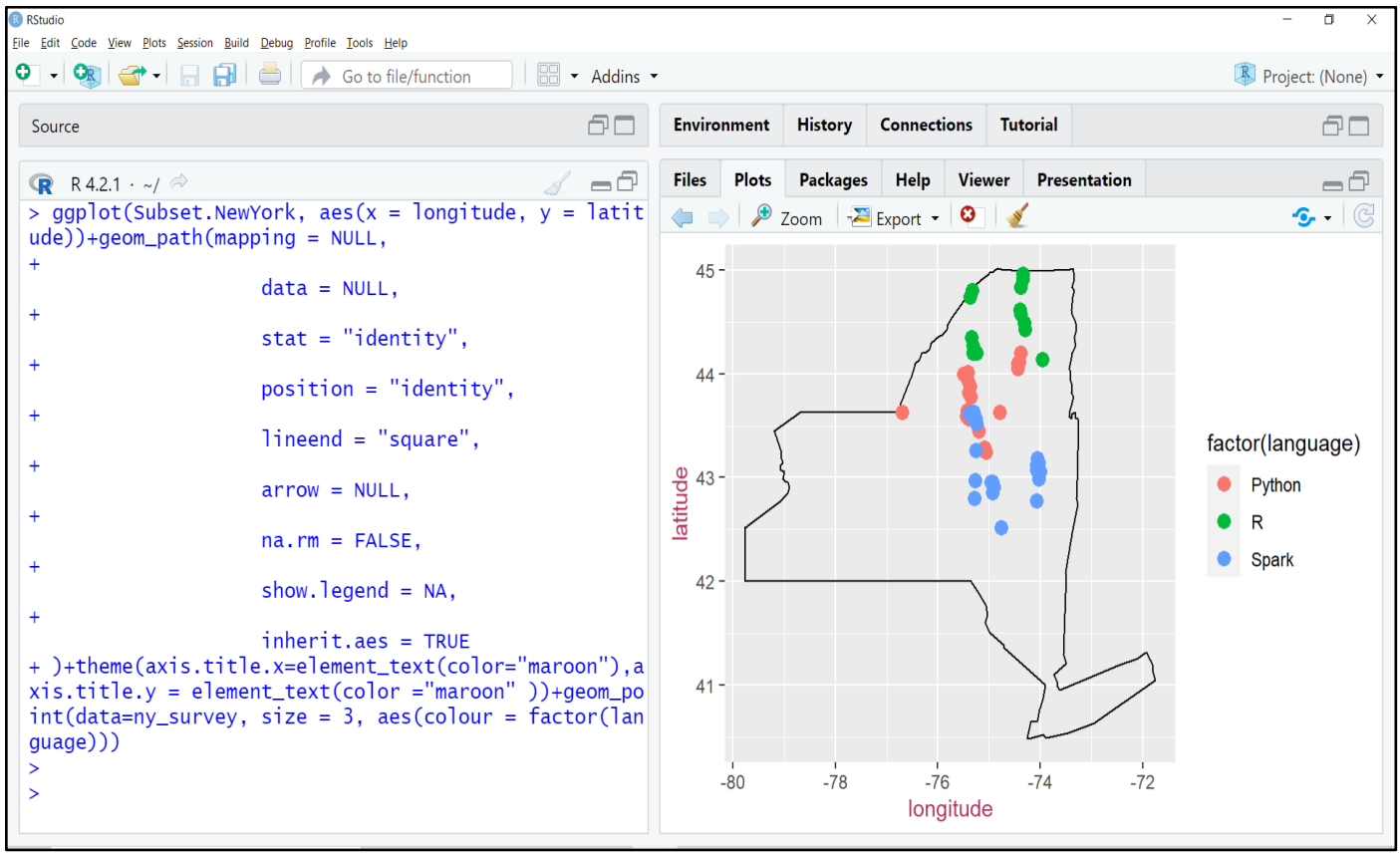
> Subset.NewYork <- subset(Longitude_Latitude, select=c(state,longitude,latitude), subset=(state=="New York"))

> ggplot(Subset.NewYork, aes(x = longitude, y = latitude)) + geom_path(mapping = NULL, data = NULL, stat = "identity", position = "identity", lineend = "square", arrow = NULL, na.rm = FALSE, show.legend = NA, inherit.aes = TRUE) + theme(axis.title.x=element_text(color="maroon"),axis.title.y = element_text(color ="maroon"))



<u>Part B</u> : Using the data in the file *ny_survey.csv* plot the survey responses on the map you created in Part A.

> ggplot(Subset.NewYork, aes(x = longitude, y = latitude)) + geom_path(mapping = NULL, data = NULL, stat = "identity", position = "identity", lineend = "square", arrow = NULL, na.rm = FALSE, show.legend = NA, inherit.aes = TRUE) + theme(axis.title.x=element_text(color="maroon"),axis.title.y = element_text(color ="maroon")) + geom_point(data=ny_survey, size = 3, aes(color = factor(language)))



Part 4: Apply KNN

In this part we want to use the R implementation of K-Nearest Neighbors on the survey data for the New York state to predict the favorite programming languages for places that weren't part of the survey.

Use knn() function to answer the following questions.

Solution :

STEP 1 : Installing the required packages

install.packages("class")

library(class)

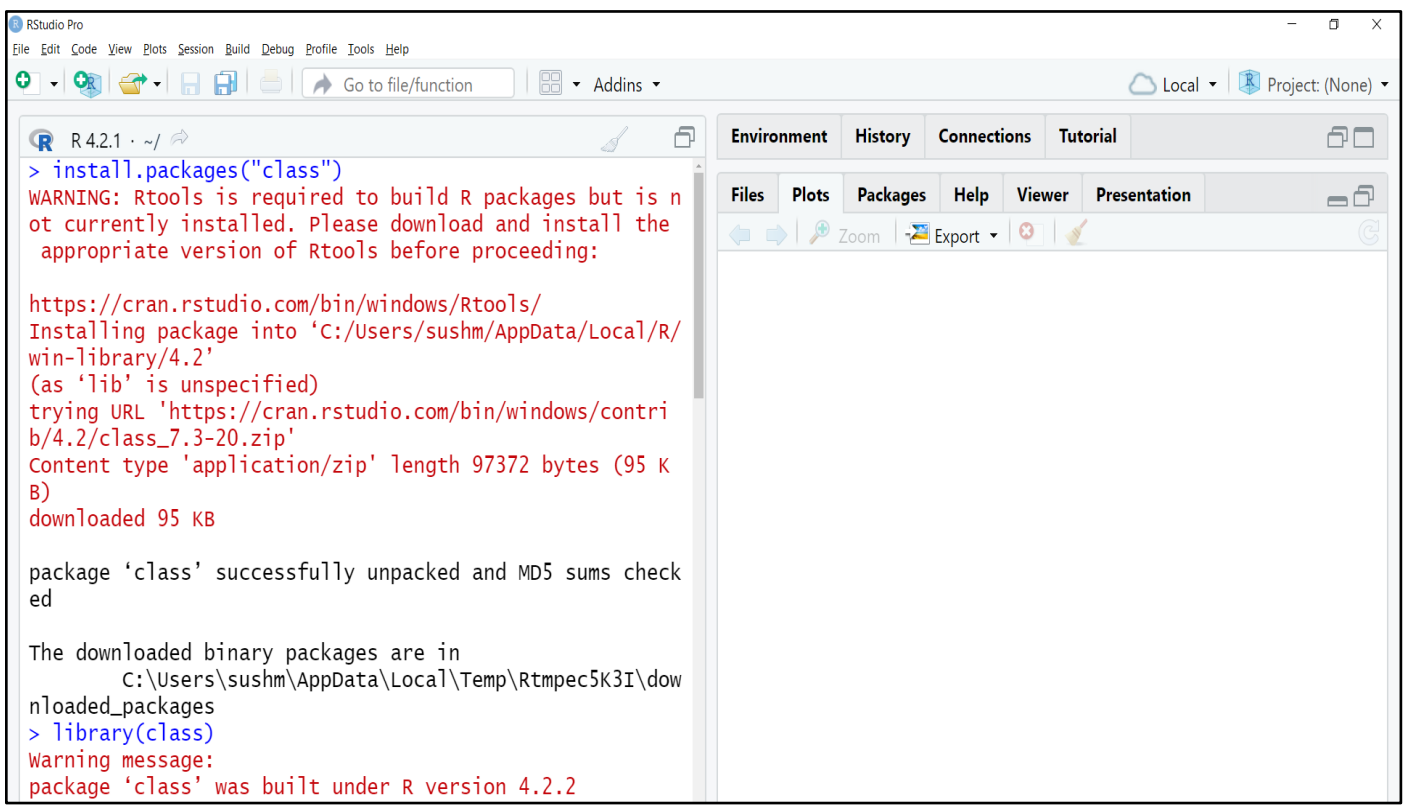
library("class")

install.packages("e1071")

library("e1071")

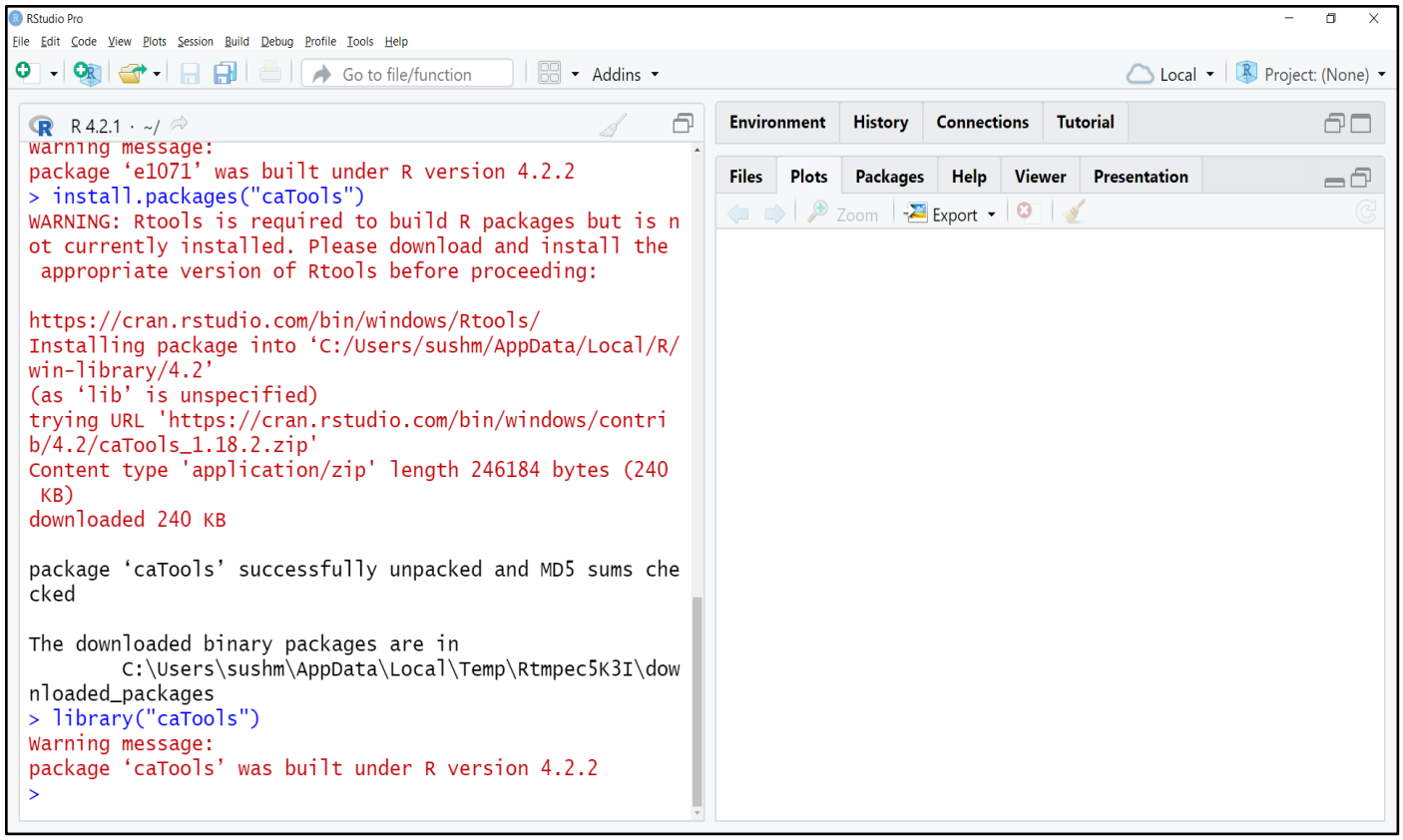
install.packages("caTools")

library("caTools")



```
> install.packages("e1071")
WARNING: Rtools is required to build R packages but is n
ot currently installed. Please download and install the
 appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/sushm/AppData/Local/R/
win-library/4.2'
(as 'lib' is unspecified)
also installing the dependency 'proxy'

trying URL 'https://cran.rstudio.com/bin/windows/contri
b/4.2/proxy_0.4-27.zip'
Content type 'application/zip' length 180353 bytes (176
 KB)
downloaded 176 KB

trying URL 'https://cran.rstudio.com/bin/windows/contri
b/4.2/e1071_1.7-12.zip'
Content type 'application/zip' length 663936 bytes (648
 KB)
downloaded 648 KB

package 'proxy' successfully unpacked and MD5 sums check
ed
package 'e1071' successfully unpacked and MD5 sums check
ed
```

```
win-library/4.2'
(as 'lib' is unspecified)
also installing the dependency 'proxy'

trying URL 'https://cran.rstudio.com/bin/windows/contri
b/4.2/proxy_0.4-27.zip'
Content type 'application/zip' length 180353 bytes (176
 KB)
downloaded 176 KB

trying URL 'https://cran.rstudio.com/bin/windows/contri
b/4.2/e1071_1.7-12.zip'
Content type 'application/zip' length 663936 bytes (648
 KB)
downloaded 648 KB

package 'proxy' successfully unpacked and MD5 sums check
ed
package 'e1071' successfully unpacked and MD5 sums check
ed

The downloaded binary packages are in
        C:\Users\sushm\AppData\Local\Temp\Rtmpec5K3I\dow
nloaded_packages
> library("e1071")
Warning message:
package 'e1071' was built under R version 4.2.2
```


RStudio Pro
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Local
Project: (None)
R 4.2.1 · ~/
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Zoom
Export
warning message:
package 'e1071' was built under R version 4.2.2
> install.packages("caTools")
WARNING: Rtools is required to build R packages but is n
ot currently installed. Please download and install the
appropriate version of Rtools before proceeding:
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/sushm/AppData/Local/R/
win-library/4.2'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contri
b/4.2/caTools_1.18.2.zip'
Content type 'application/zip' length 246184 bytes (240
KB)
downloaded 240 KB
package 'caTools' successfully unpacked and MD5 sums che
cked
The downloaded binary packages are in
C:\Users\sushm\AppData\Local\Temp\Rtmpec5K3I\dow
nloaded_packages
> library("caTools")
Warning message:
package 'caTools' was built under R version 4.2.2
>

STEP 2 : Extracting the longitude and latitude

> ny_survey$longitude=factor(ny_survey$longitude)

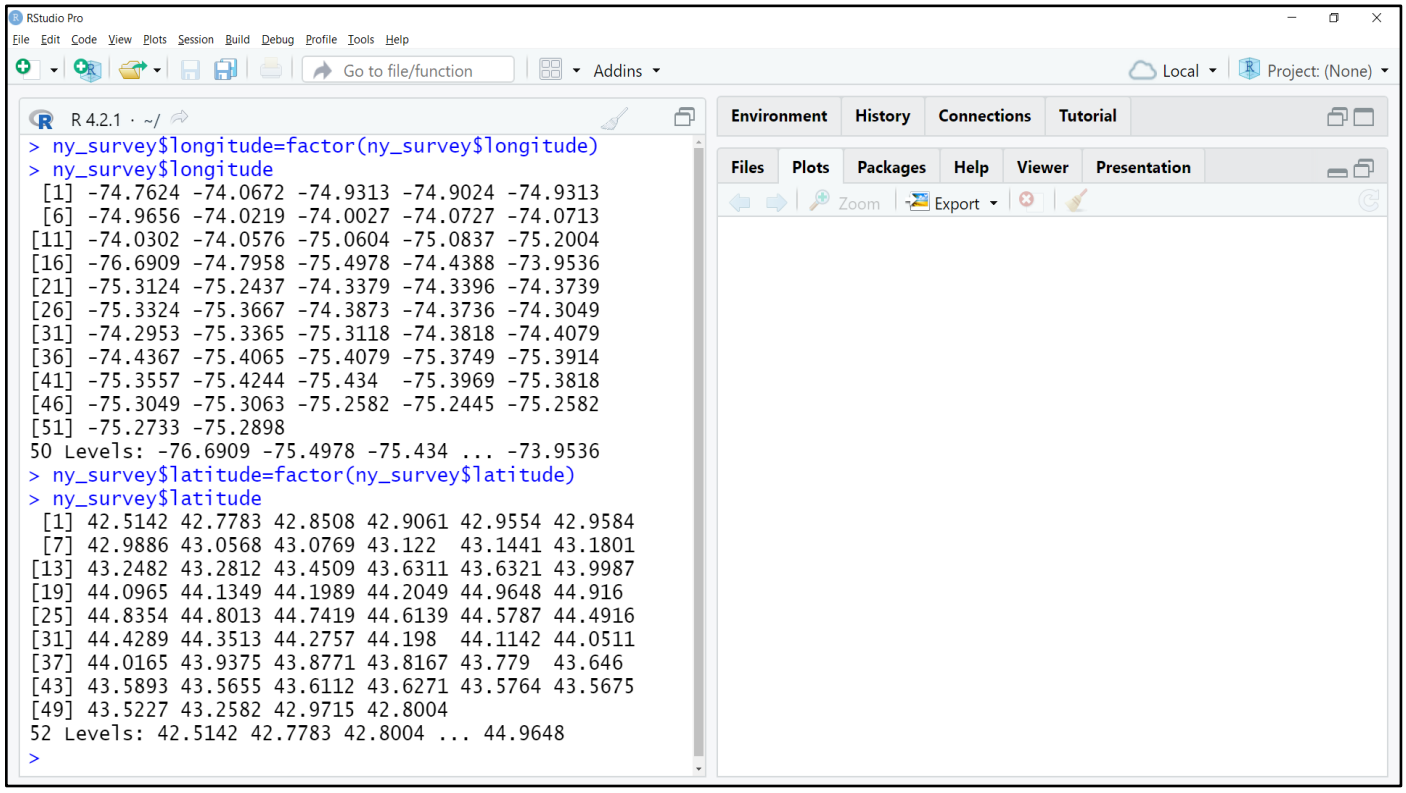
> ny_survey$longitude

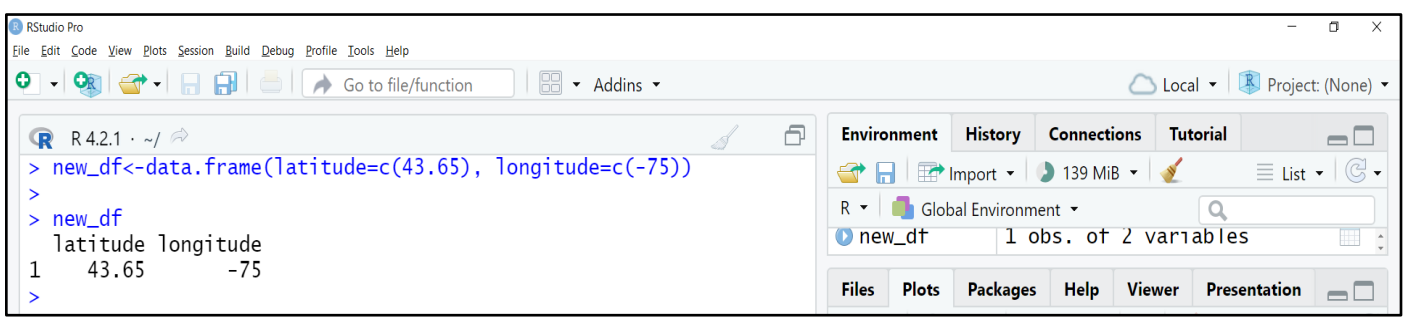
> ny_survey$latitude=factor(ny_survey$latitude)

> ny_survey$latitude

> new_df<-data.frame(latitude=c(43.65), longitude=c(-75))

> new_df





STEP 3 : Train Dataset and Test Dataset

SurveyData<-sample(1:nrow(ny_survey),size=nrow(ny_survey)*0.7, replace=FALSE)

train dataset

train.ny_survey<-ny_survey[SurveyData,]

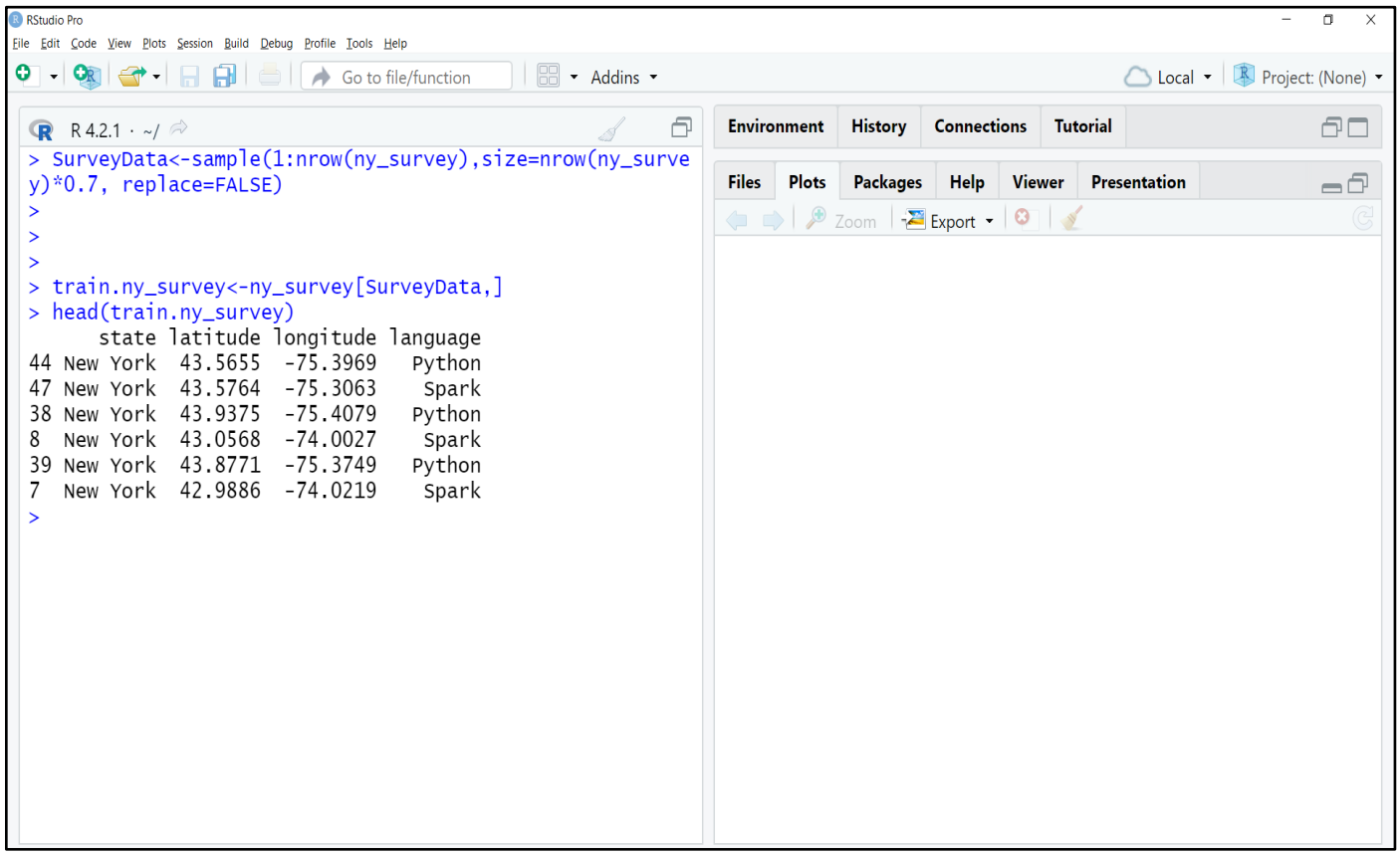
head(train.ny_survey)

test Dataset

test.ny_survey<-ny_survey[-SurveyData,]

test.ny_survey

head(test.ny_survey)

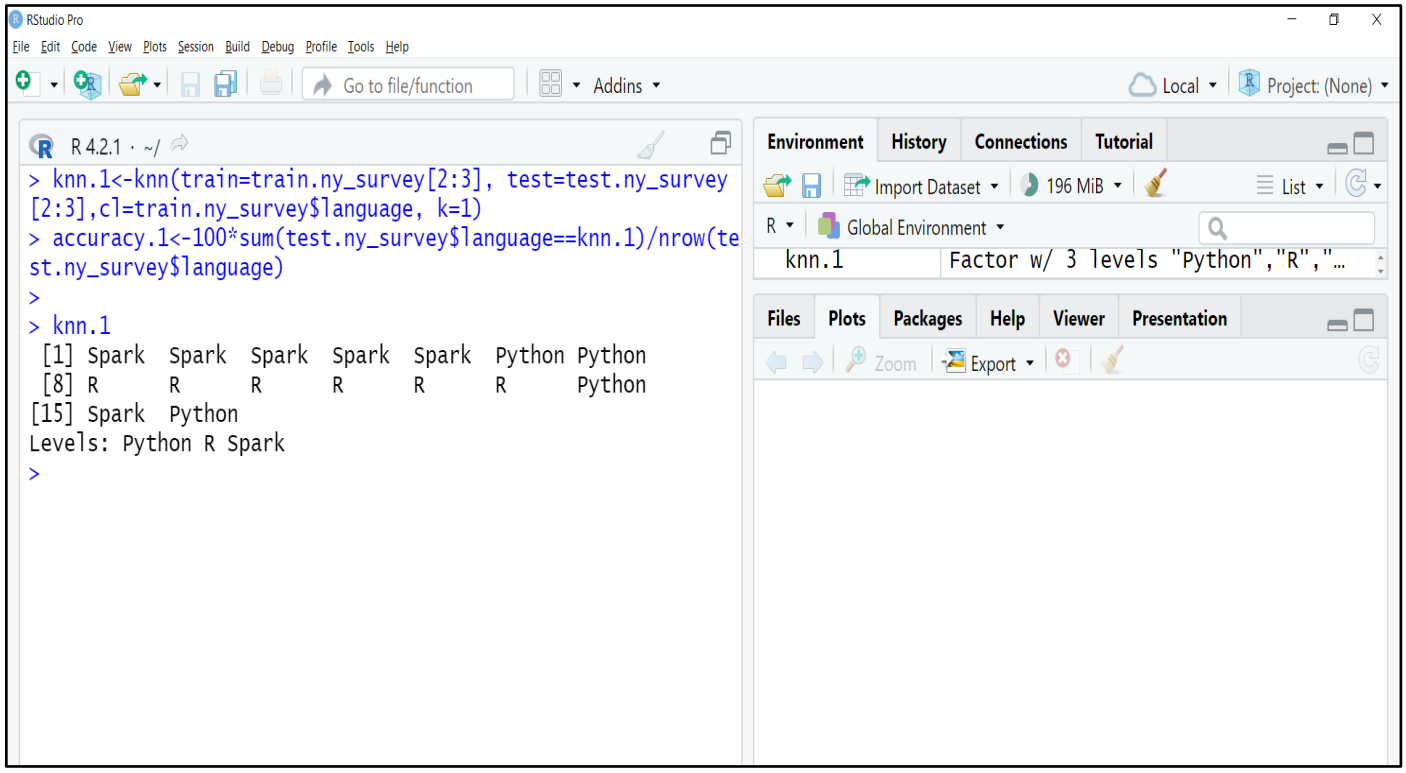


```
> test.ny_survey<-ny_survey[-SurveyData,]
> test.ny_survey
      state latitude longitude language
4  New York  42.9061  -74.9024    Spark
9  New York  43.0769  -74.0727    Spark
11 New York  43.1441  -74.0302    Spark
16 New York  43.6311  -76.6909   Python
17 New York  43.6321  -74.7958   Python
20 New York  44.1349  -73.9536        R
21 New York  44.1989  -75.3124        R
22 New York  44.2049  -75.2437        R
25 New York  44.8354  -74.3739        R
33 New York  44.2757  -75.3118        R
34 New York   44.198  -74.3818   Python
35 New York  44.1142  -74.4079   Python
43 New York  43.5893   -75.434   Python
45 New York  43.6112  -75.3818    Spark
46 New York  43.6271  -75.3049    Spark
51 New York  42.9715  -75.2733    Spark
>
```

```
> head(test.ny_survey)
      state latitude longitude language
4  New York  42.9061  -74.9024    Spark
9  New York  43.0769  -74.0727    Spark
11 New York  43.1441  -74.0302    Spark
16 New York  43.6311  -76.6909   Python
17 New York  43.6321  -74.7958   Python
20 New York  44.1349  -73.9536        R
>
```

Question 1: For k=1, what is the preferred language for the location at *longitude=-75* and *latitude=43.65* ?

```
> knn.1<-knn(train=train.ny_survey[2:3], test=test.ny_survey[2:3],cl=train.ny_survey$language, k=1)

> accuracy.1<-100*sum(test.ny_survey$language==knn.1)/nrow(test.ny_survey$language)

> knn.1

> Value_k1<-knn(train=train.ny_survey[2:3],test=new_df,cl=train.ny_survey$language, k=1)

> print(Value_k1)
```



Question 2: For k=2, what is the preferred language for the location at *longitude=-75* and *latitude=43.65* ?

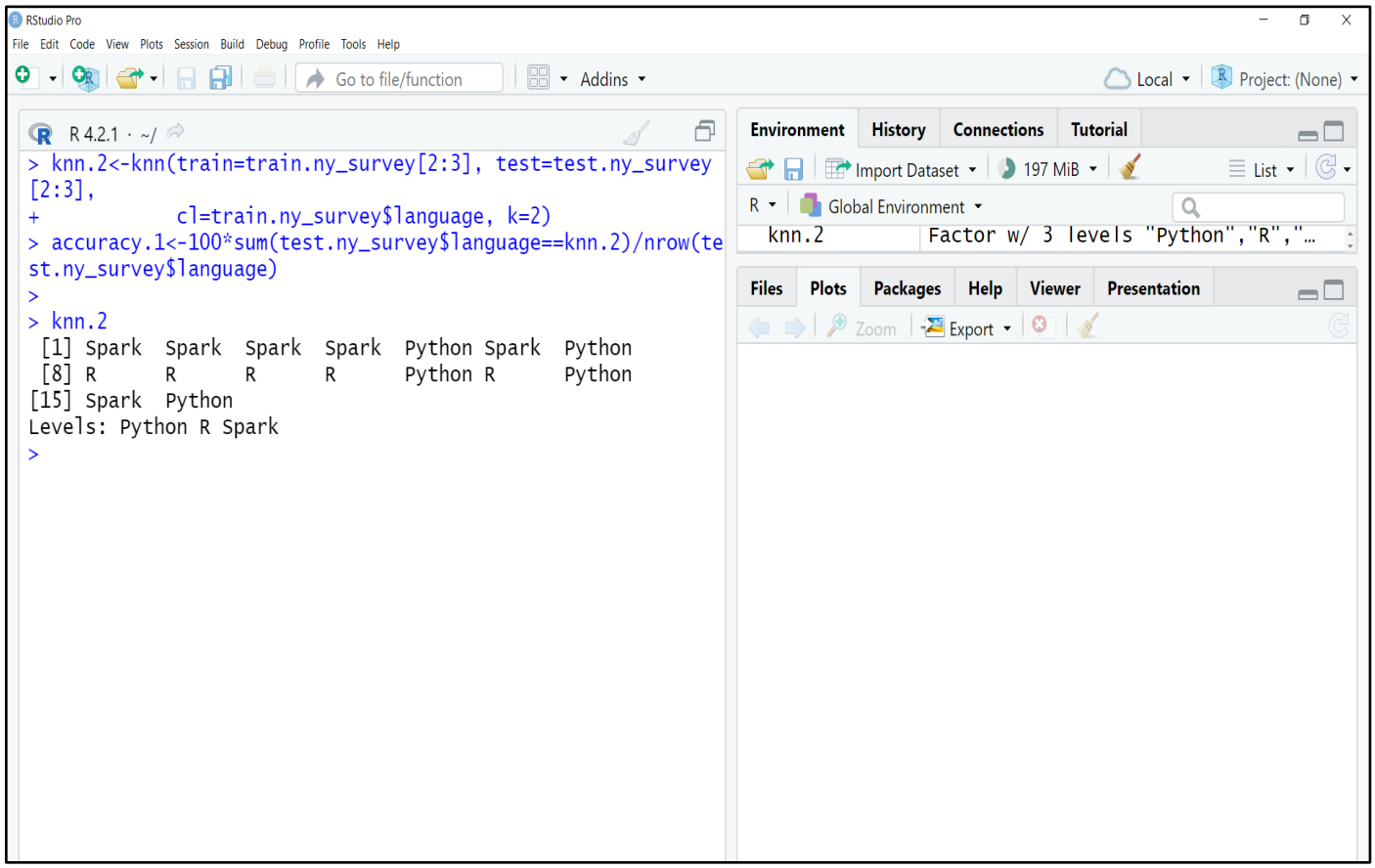
> knn.2<-knn(train=train.ny_survey[2:3], test=test.ny_survey[2:3],cl=train.ny_survey$language, k=2)

> accuracy.1<-100*sum(test.ny_survey$language==knn.2)/nrow(test.ny_survey$language)

> knn.2

> Value_k2<-knn(train=train.ny_survey[2:3],test=new_df,cl=train.ny_survey$language, k=2)

> print(Value_k2)



Question 3: For k=3, what is the preferred language for the location at *longitude=-75* and *latitude=43.65* ?

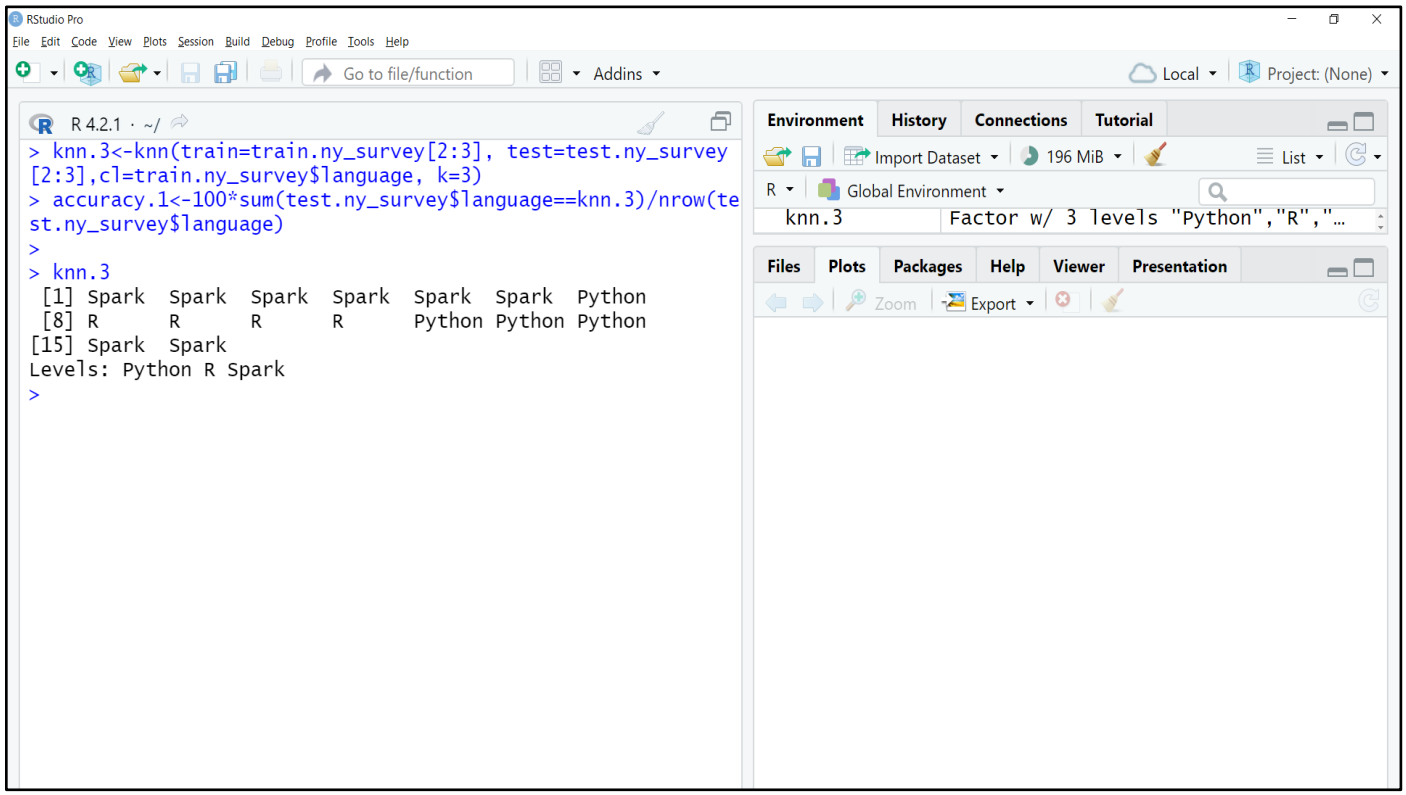
> knn.3<-knn(train=train.ny_survey[2:3], test=test.ny_survey[2:3],cl=train.ny_survey$language, k=3)

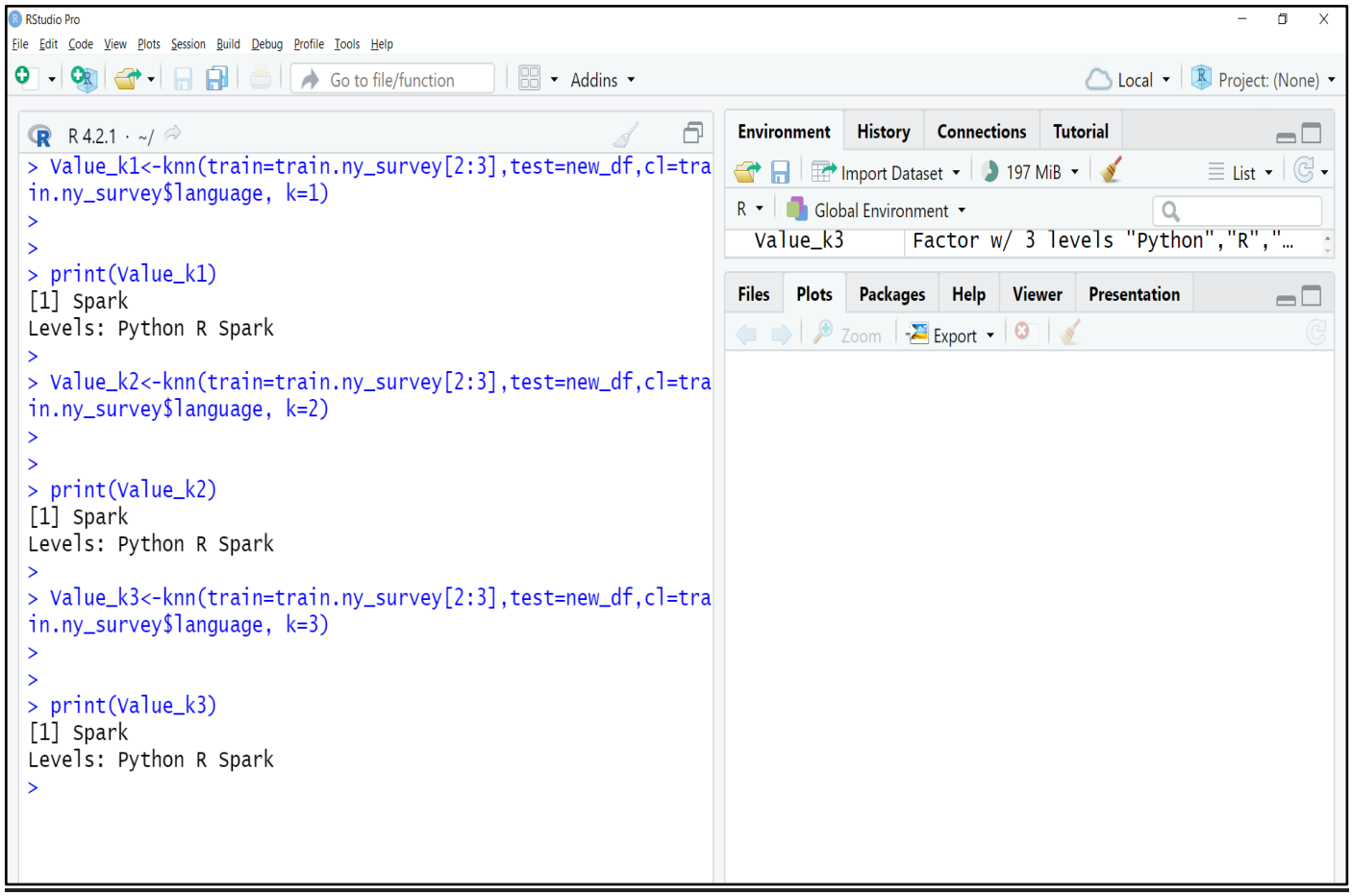
> accuracy.1<-100*sum(test.ny_survey$language==knn.3)/nrow(test.ny_survey$language)

> knn.3

> Value_k3<-knn(train=train.ny_survey[2:3],test=new_df,cl=train.ny_survey$language, k=3)

> print(Value_k3)




RStudio Pro
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Local
Project: (None)
R 4.2.1 · ~/
> Value_k1<-knn(train=train.ny_survey[2:3],test=new_df,cl=tra
in.ny_survey$language, k=1)
>
>
> print(Value_k1)
[1] Spark
Levels: Python R Spark
>
> Value_k2<-knn(train=train.ny_survey[2:3],test=new_df,cl=tra
in.ny_survey$language, k=2)
>
>
> print(Value_k2)
[1] Spark
Levels: Python R Spark
>
> Value_k3<-knn(train=train.ny_survey[2:3],test=new_df,cl=tra
in.ny_survey$language, k=3)
>
>
> print(Value_k3)
[1] Spark
Levels: Python R Spark
>
Environment History Connections Tutorial
Import Dataset
197 MiB
List
R
Global Environment
Value_k3 Factor w/ 3 levels "Python","R","...
Files Plots Packages Help Viewer Presentation
Zoom
Export

Part 5: Display the new point on the state map.

Use the map from Part B in Step 3 to display the data point at longitude=-75 and latitude=43.65.

Hint : You can use the same technique as before to add the new point to the existing map. Simply add the following statement to the code generated the map in Part B in Step 3 :

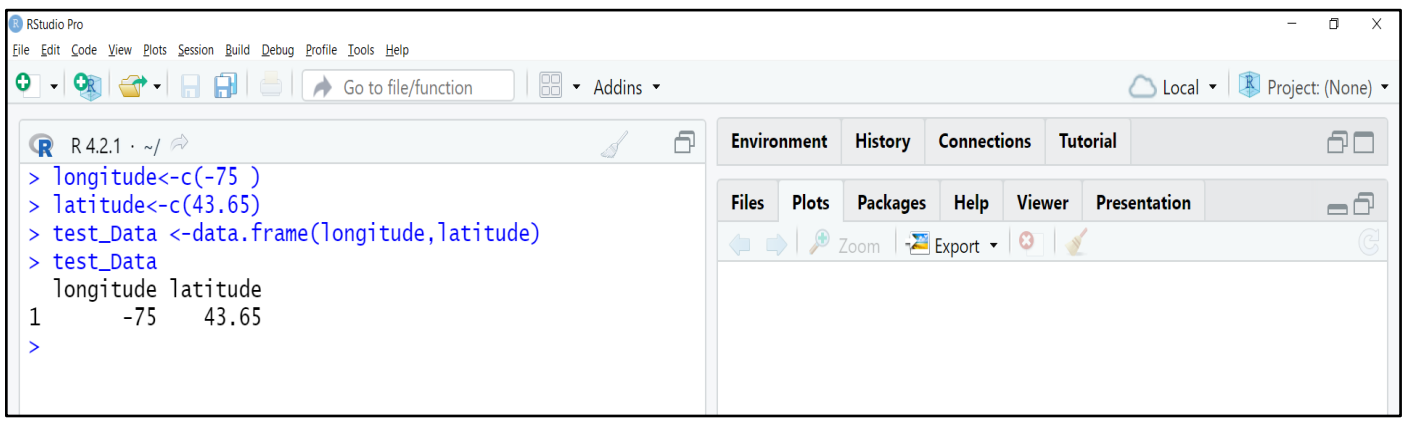
geom_point(data=testingDataFrame, shape=25, fill="blue", color="darkred", size=5)

where testing DataFrame is the data frame for longitude = -75 and latitude = 43.65

Solution :

Step 1 : Adding the longitude and latitude point to Test data frame "test_Data"

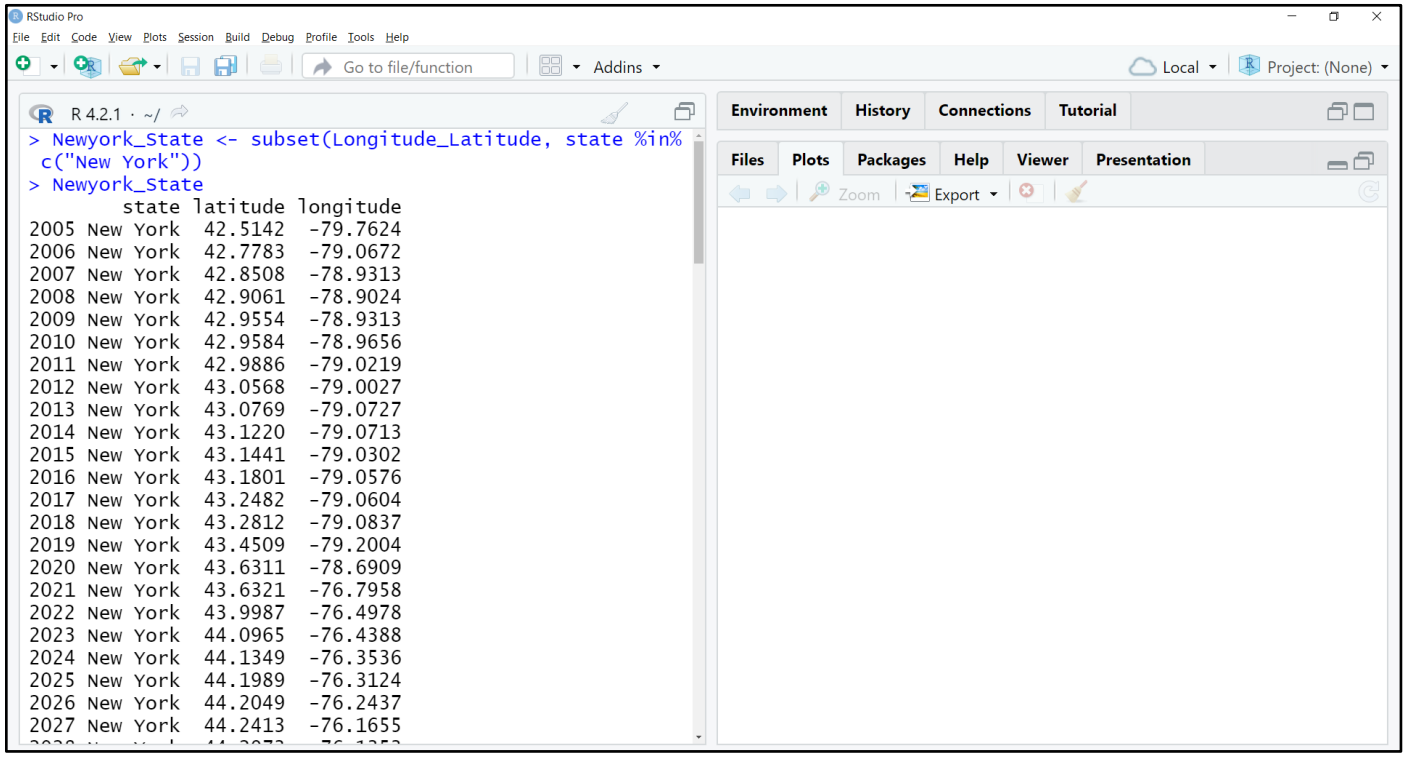
```
> longitude<-c(-75 )
> latitude<-c(43.65)
> test_Data <-data.frame(longitude,latitude)
> test_Data
```



Step 2 : Creating a subset “Newyork_State”

> Newyork_State <- subset(Longitude_Latitude, state %in% c("New York"))

> Newyork_State



Step 3 : Drawing the required ggplot

> ggplot(Subset.NewYork, aes(x = longitude, y = latitude)) + geom_path(mapping = NULL, data = NULL, stat = "identity", position = "identity", lineend = "square", arrow = NULL, na.rm = FALSE, show.legend = NA, inherit.aes = TRUE) + theme(axis.title.x=element_text(color="maroon"),axis.title.y = element_text(color ="maroon")) + geom_point(data=ny_survey, size = 3, aes(color = factor(language)))+geom_point(data=test_Data, shape=25, fill="blue", color="darkred", size=5)