

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 *latitude*.

```
> 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
```

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code to read two CSV files and display the first one.
- Environment/History/Connections/Tutorial:** A row of tabs for managing the session.
- Files/Plots/Packages/Help/Viewer/Presentation:** A second row of tabs for file management and output viewing.
- Console:** Displays the execution of the R code and the resulting data table.

R Code Executed:

```
> file1="c:/DATASETS/longitude_latitude.csv"
>
> file2="c:/DATASETS/ny_survey.csv"
>
> Longitude_Latitude=read.csv(file1, stringsAsFactors = FALSE,header
= TRUE)
> Longitude_Latitude
```

Output Data Table:

| | 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 |

The image shows the RStudio interface with the following components:

- Source Editor:** Contains the R code to read a CSV file and the resulting data frame output.
- Environment:** Shows the loaded data frame 'ny_survey'.
- Plots:** Empty.
- Packages:** Shows installed packages.
- Help:** Empty.
- Viewer:** Empty.
- Presentations:** Empty.

Source Editor Code:

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

Source Editor Output:

| | 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 |

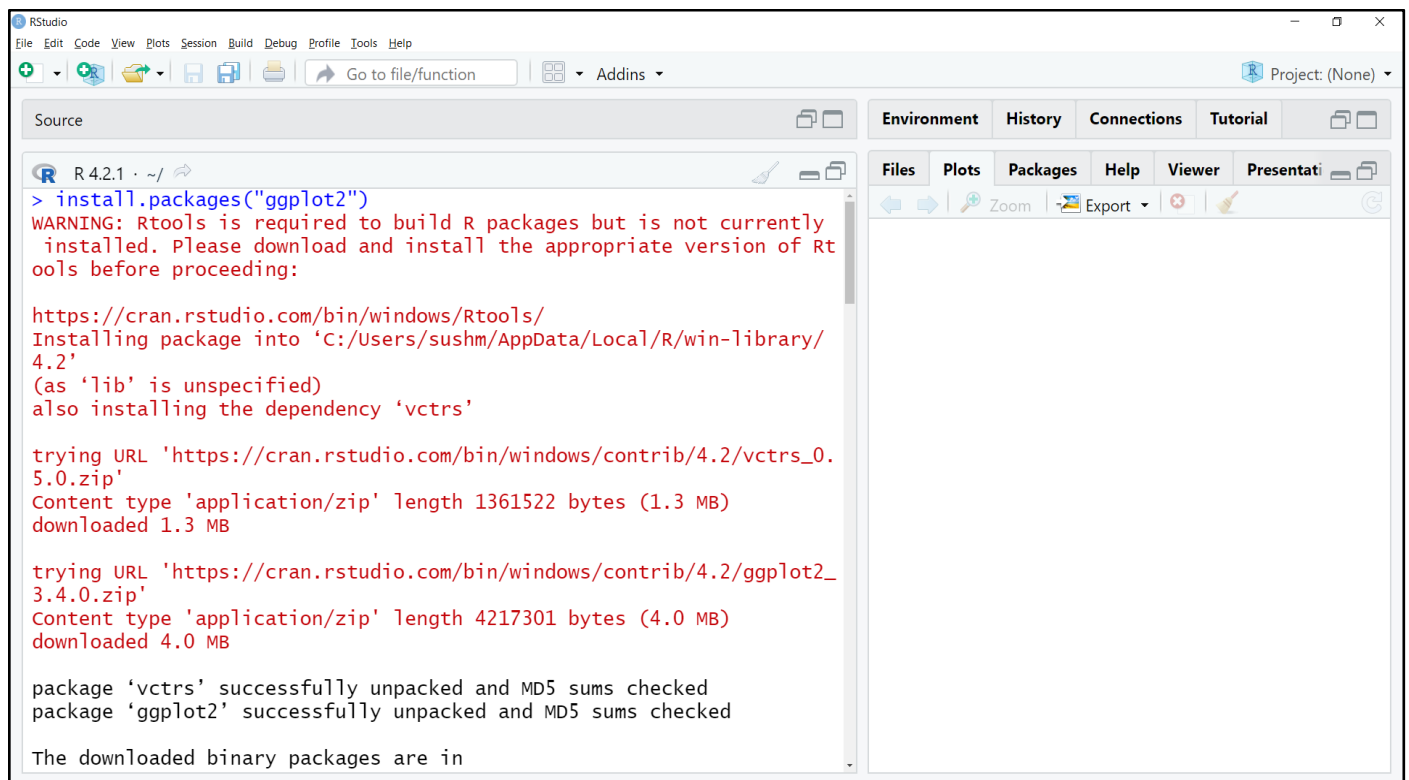
PART 2 : Using the *longitude* and *latitude* data in the file *longitude_latitude.csv* plot the map of the California state.

```
> 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" ))
```



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ Addins Go to file/function Project: (None)

Source
R 4.2.1 ~
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 'vctrs'

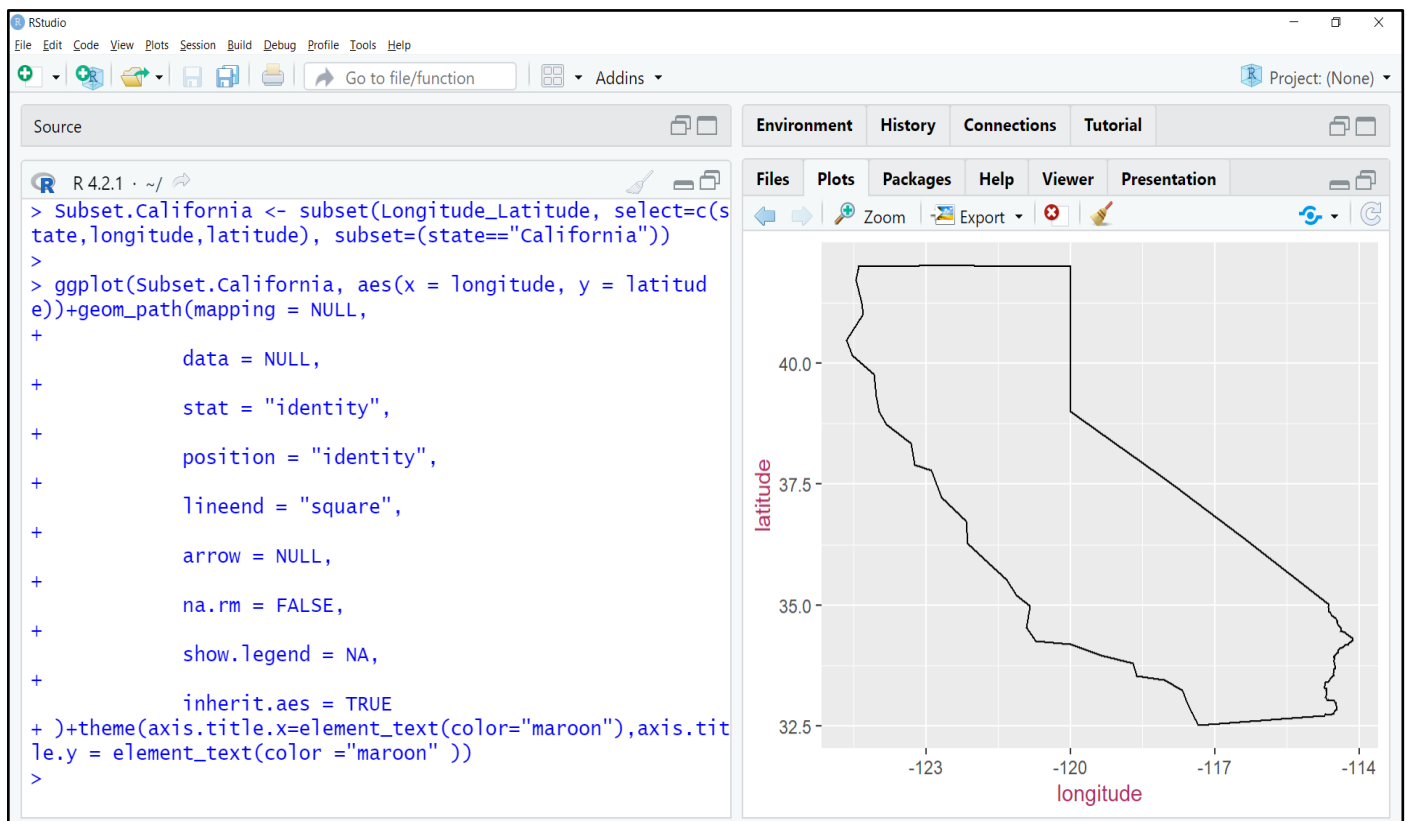
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/vctrs_0.
5.0.zip'
Content type 'application/zip' length 1361522 bytes (1.3 MB)
downloaded 1.3 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ggplot2_
3.4.0.zip'
Content type 'application/zip' length 4217301 bytes (4.0 MB)
downloaded 4.0 MB

package 'vctrs' successfully unpacked and MD5 sums checked
package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\sushm\AppData\Local\Temp\Rtmp0ShCc6\downloaded_packa
ges
> library(ggplot2)
Warning message:
package 'ggplot2' was built under R version 4.2.2

```

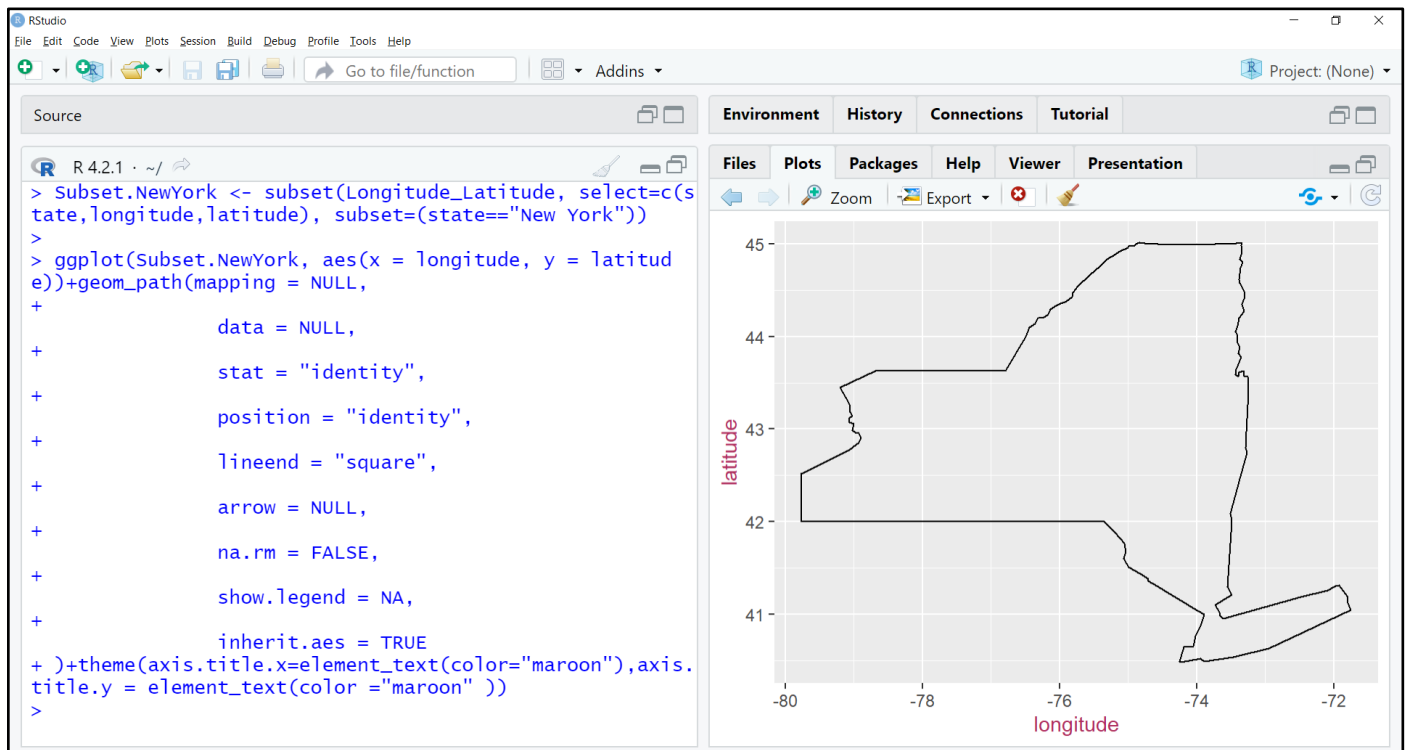


PART 3: Draw map and display data

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

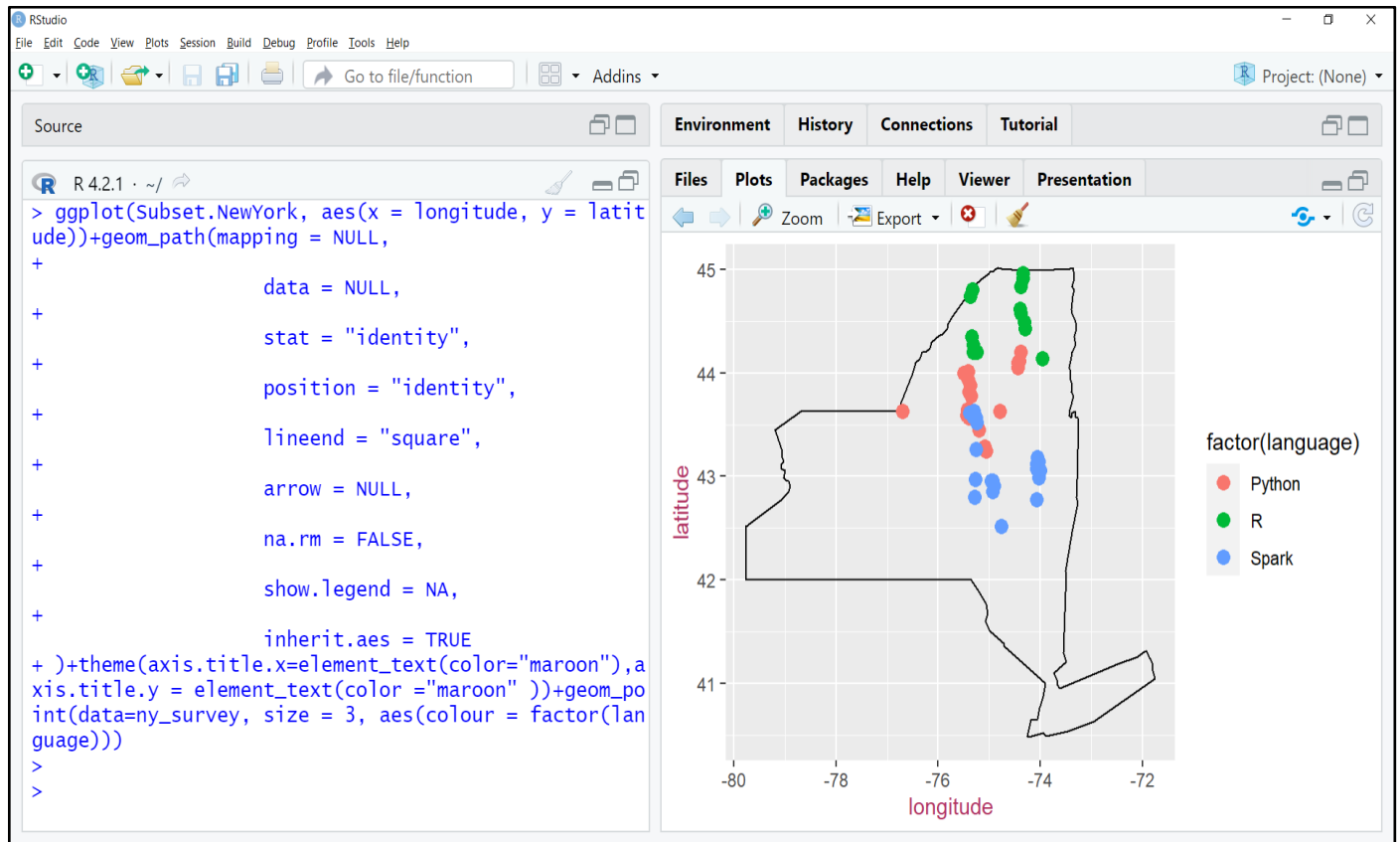
```
> Subset.NewYork <- subset(Longitude_Latitude, select=c(state,longitude,latitude),
subset=(state=="New York"))
```

```
> 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" ))
```



Part B : 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

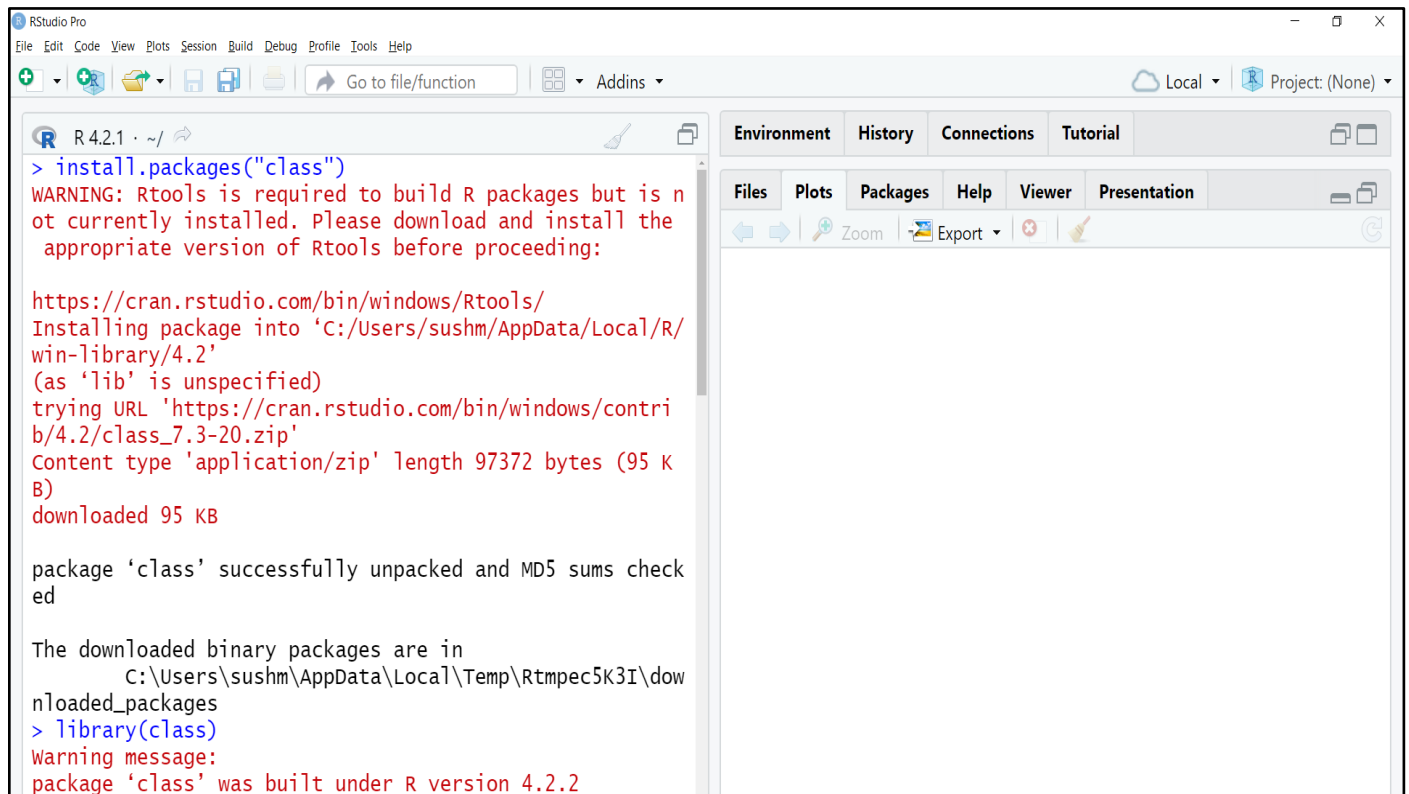
In this part we want to use the R implementation of K-Nearest Neighbors on the survey data for the New York state to predict the favorite programming languages for places that weren't part of the survey.

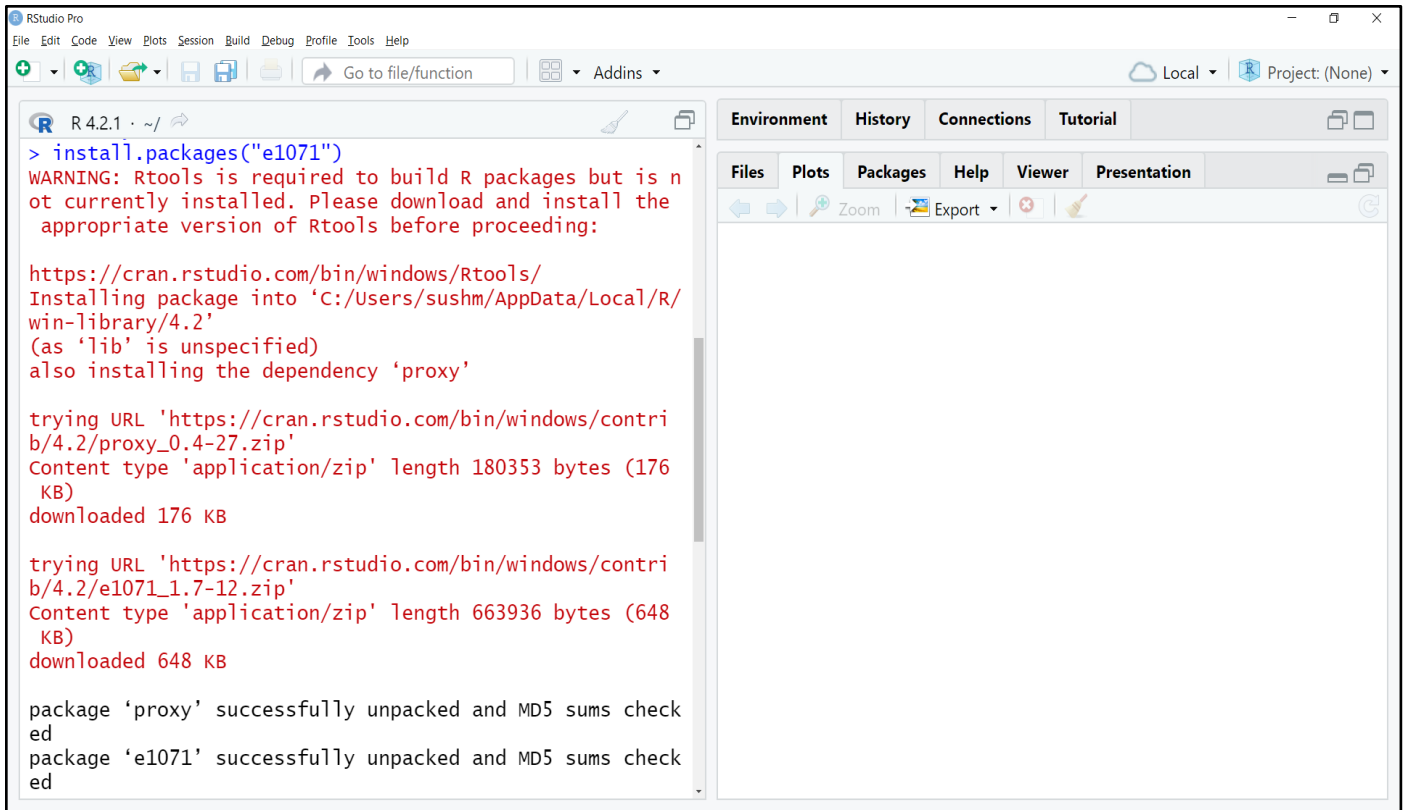
Use `knn()` function to answer the following questions.

Solution :

STEP 1 : Installing the required packages

```
install.packages("class")
library(class)
library("class")
install.packages("e1071")
library("e1071")
install.packages("caTools")
library("caTools")
```





```

R 4.2.1 · ~/
> install.packages("e1071")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

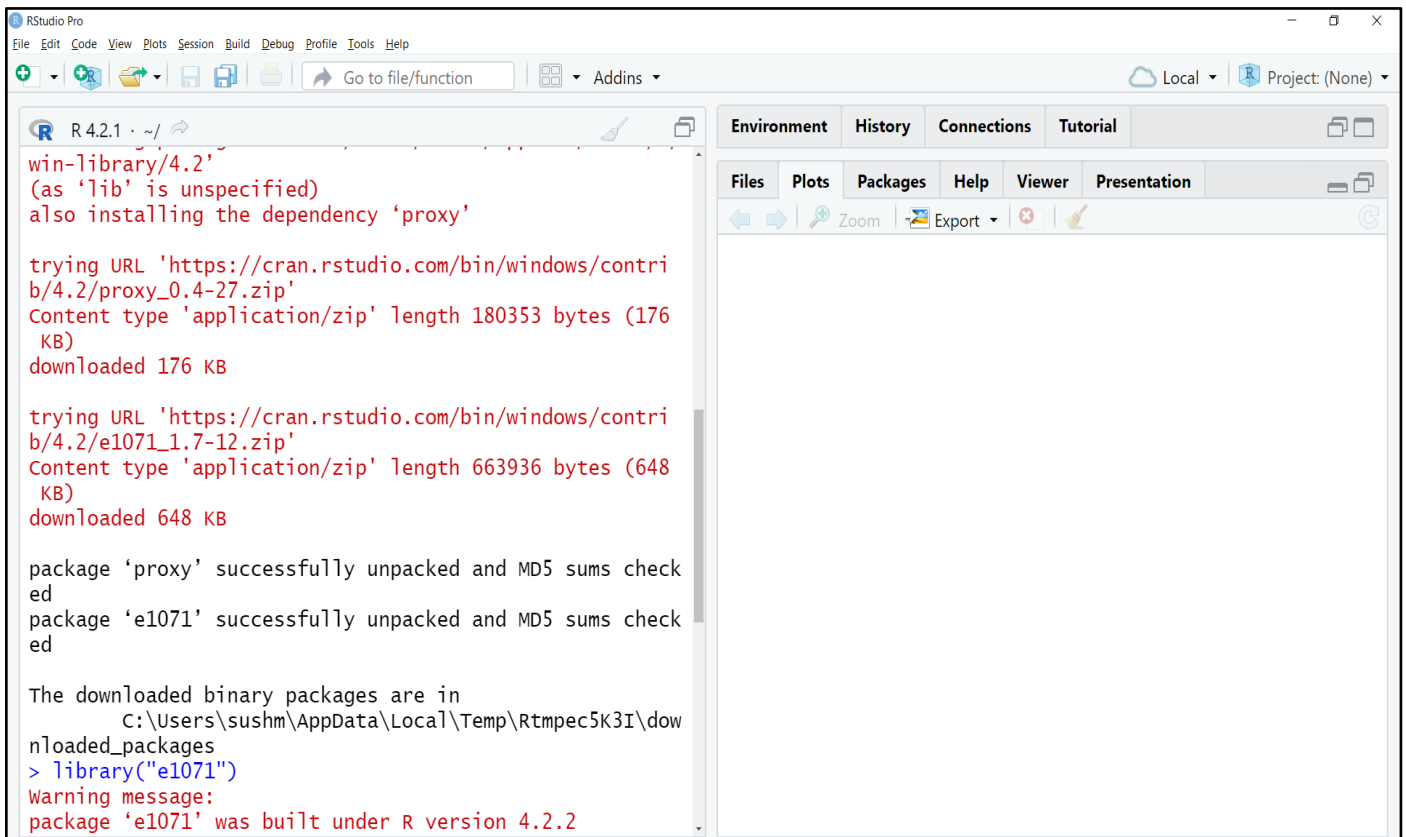
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/sushm/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
also installing the dependency 'proxy'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/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/contrib/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 checked
package 'e1071' successfully unpacked and MD5 sums checked

```



```

R 4.2.1 · ~/
win-library/4.2'
(as 'lib' is unspecified)
also installing the dependency 'proxy'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/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/contrib/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 checked
package 'e1071' successfully unpacked and MD5 sums checked

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

```

The screenshot shows the RStudio Pro interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for creating a new file, opening a file, saving, and other standard functions. The main window is divided into two panes. The left pane, titled 'R 4.2.1 · ~/', contains the R console output. The right pane has tabs for Environment, History, Connections, and Tutorial, and a sub-panel with tabs for Files, Plots, Packages, Help, Viewer, and Presentation. The console output shows the following text:

```
R 4.2.1 · ~/
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 not
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)
```

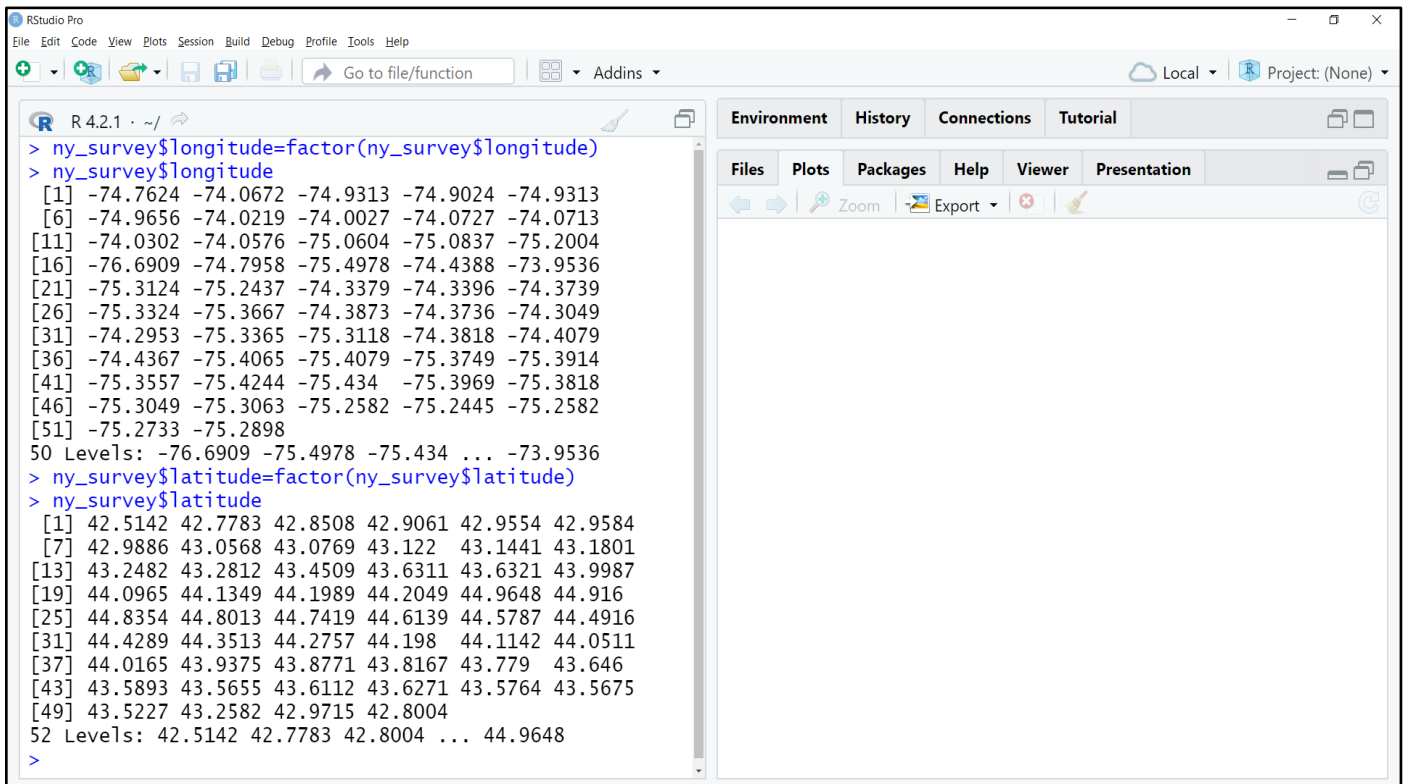
```
> ny_survey$longitude
```

```
> ny_survey$latitude=factor(ny_survey$latitude)
```

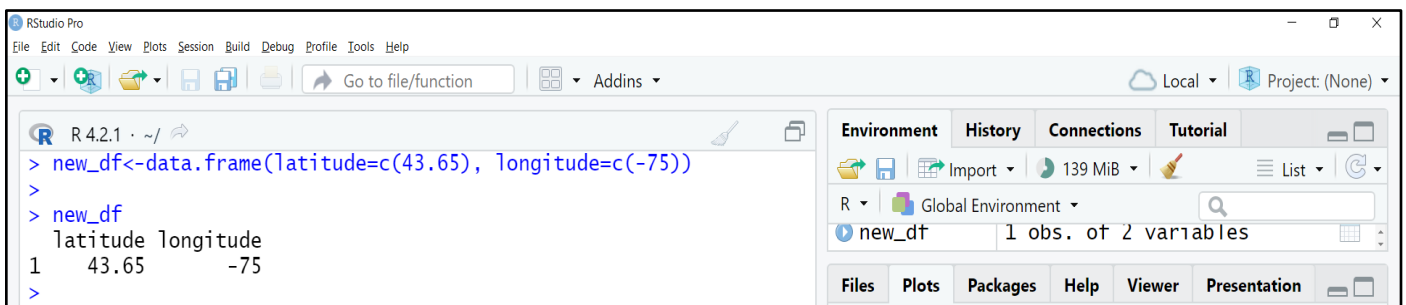
```
> ny_survey$latitude
```

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

```
> new_df
```



```
R 4.2.1 ~/  
File Edit Code View Plots Session Build Debug Profile Tools Help  
Go to file/function Addins Local Project: (None)  
Environment History Connections Tutorial  
Files Plots Packages Help Viewer Presentation  
Zoom Export  
> ny_survey$longitude=factor(ny_survey$longitude)  
> ny_survey$longitude  
[1] -74.7624 -74.0672 -74.9313 -74.9024 -74.9313  
[6] -74.9656 -74.0219 -74.0027 -74.0727 -74.0713  
[11] -74.0302 -74.0576 -75.0604 -75.0837 -75.2004  
[16] -76.6909 -74.7958 -75.4978 -74.4388 -73.9536  
[21] -75.3124 -75.2437 -74.3379 -74.3396 -74.3739  
[26] -75.3324 -75.3667 -74.3873 -74.3736 -74.3049  
[31] -74.2953 -75.3365 -75.3118 -74.3818 -74.4079  
[36] -74.4367 -75.4065 -75.4079 -75.3749 -75.3914  
[41] -75.3557 -75.4244 -75.434 -75.3969 -75.3818  
[46] -75.3049 -75.3063 -75.2582 -75.2445 -75.2582  
[51] -75.2733 -75.2898  
50 Levels: -76.6909 -75.4978 -75.434 ... -73.9536  
> ny_survey$latitude=factor(ny_survey$latitude)  
> ny_survey$latitude  
[1] 42.5142 42.7783 42.8508 42.9061 42.9554 42.9584  
[7] 42.9886 43.0568 43.0769 43.122 43.1441 43.1801  
[13] 43.2482 43.2812 43.4509 43.6311 43.6321 43.9987  
[19] 44.0965 44.1349 44.1989 44.2049 44.9648 44.916  
[25] 44.8354 44.8013 44.7419 44.6139 44.5787 44.4916  
[31] 44.4289 44.3513 44.2757 44.198 44.1142 44.0511  
[37] 44.0165 43.9375 43.8771 43.8167 43.779 43.646  
[43] 43.5893 43.5655 43.6112 43.6271 43.5764 43.5675  
[49] 43.5227 43.2582 42.9715 42.8004  
52 Levels: 42.5142 42.7783 42.8004 ... 44.9648  
>
```



```
R 4.2.1 ~/  
File Edit Code View Plots Session Build Debug Profile Tools Help  
Go to file/function Addins Local Project: (None)  
Environment History Connections Tutorial  
Files Plots Packages Help Viewer Presentation  
Import 139 MiB List  
R Global Environment  
new_df 1 obs. of 2 variables  
latitude longitude  
1 43.65 -75  
>
```

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,]
```

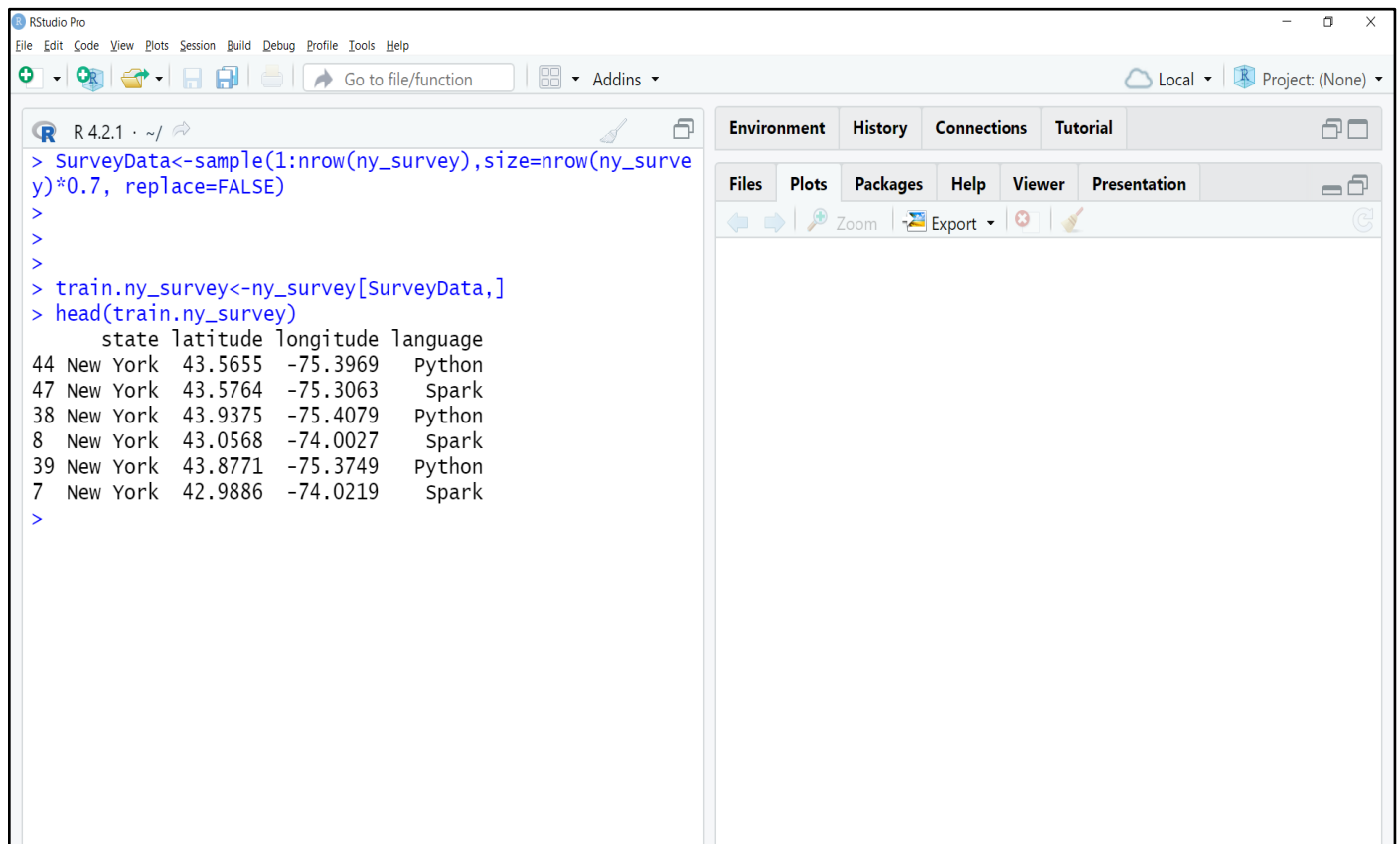
```
head(train.ny_survey)
```

test Dataset

```
test.ny_survey<-ny_survey[-SurveyData,]
```

```
test.ny_survey
```

```
head(test.ny_survey)
```



RStudio Pro interface showing the execution of a command in the console. The command `test.ny_survey<-ny_survey[-SurveyData,]` has been executed, resulting in a data frame with 16 rows and 5 columns: `state`, `latitude`, `longitude`, `language`, and an unnamed column. The data is displayed in the console output.

| | 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 | |

RStudio Pro interface showing the execution of a command in the console. The command `head(test.ny_survey)` has been executed, resulting in a data frame with 6 rows and 5 columns: `state`, `latitude`, `longitude`, `language`, and an unnamed column. The data is displayed in the console output.

| | 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 ?

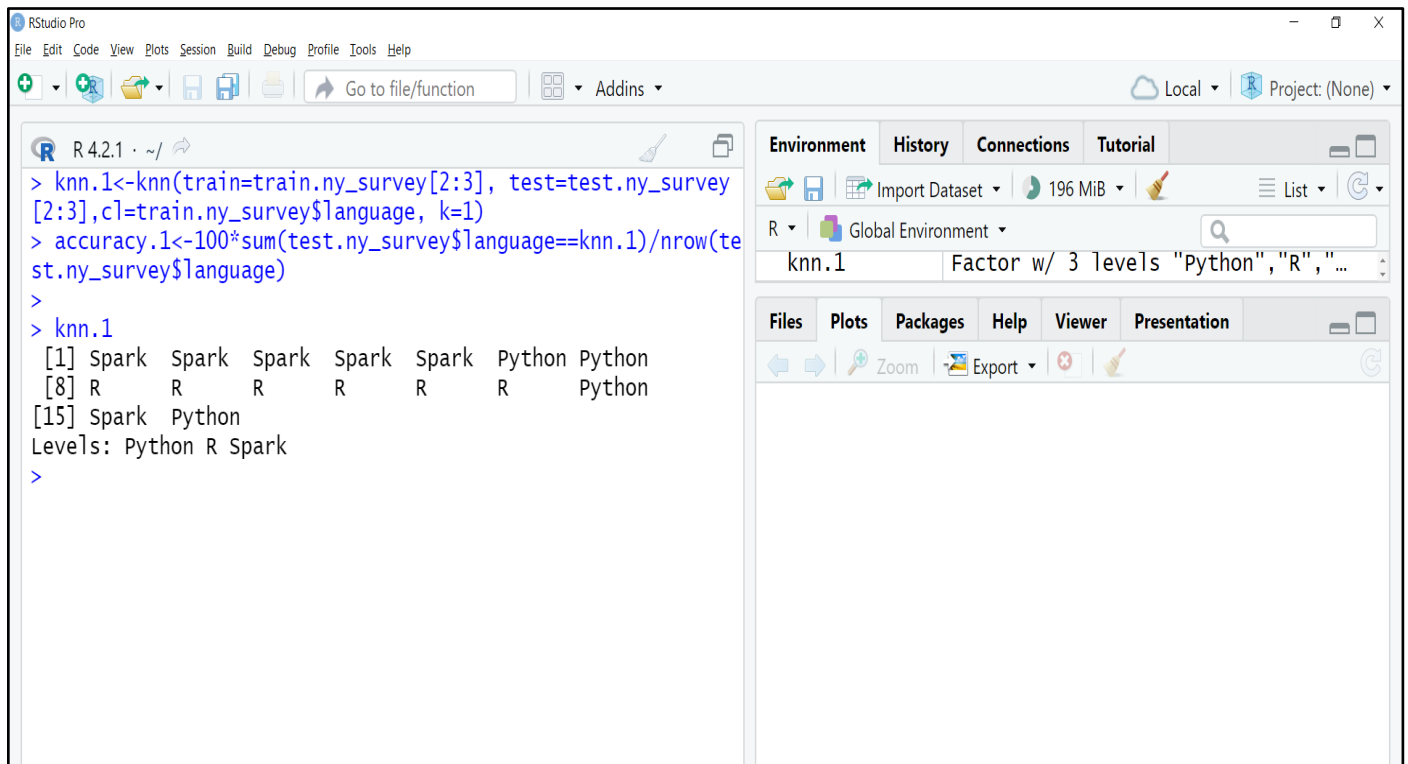
```
> knn.1<-knn(train=train.ny_survey[2:3], test=test.ny_survey[2:3],cl=train.ny_survey$language,
k=1)
```

```
> 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 ?

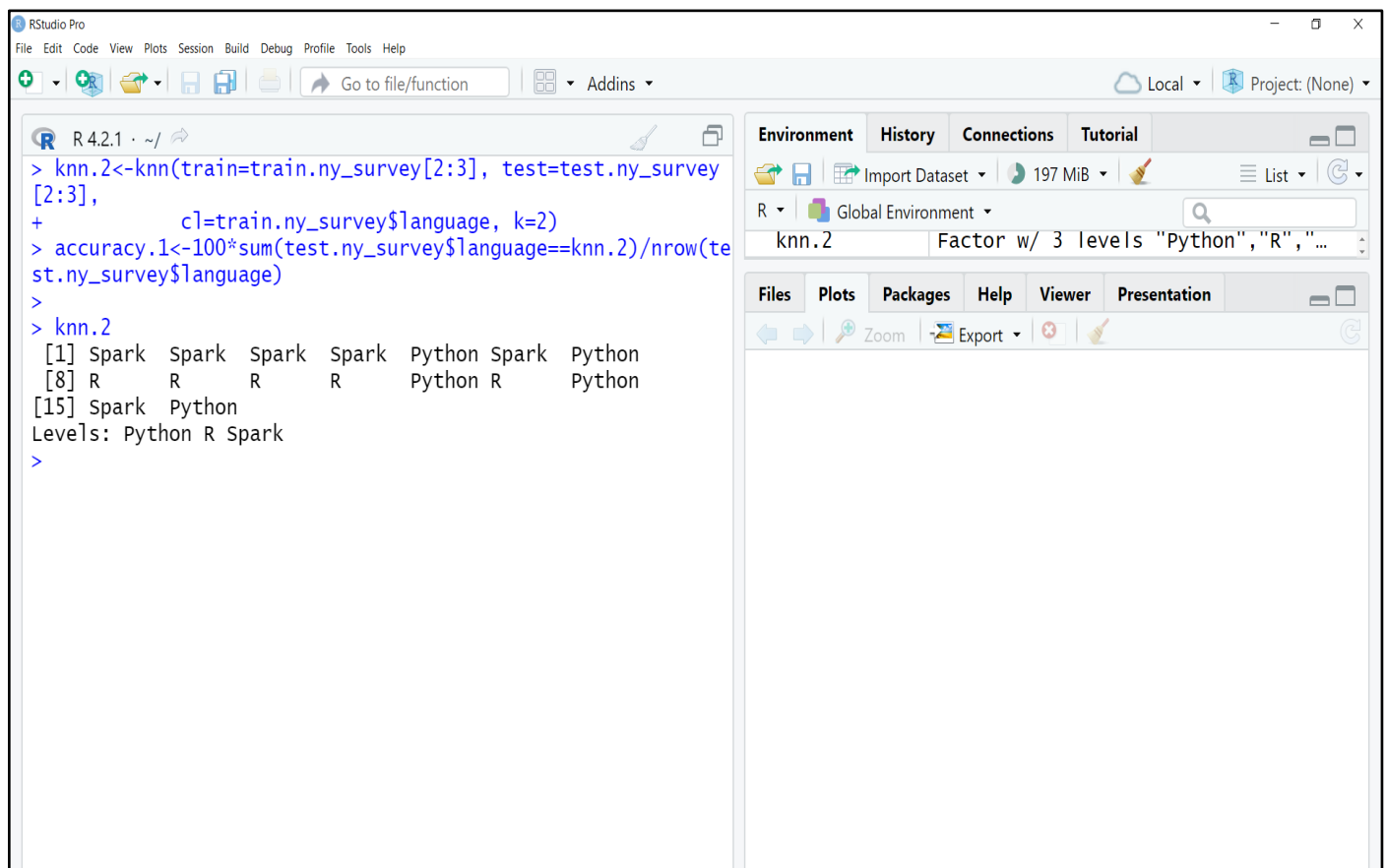
```
> knn.2<-knn(train=train.ny_survey[2:3], test=test.ny_survey[2:3],cl=train.ny_survey$language,
k=2)
```

```
> 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 ?

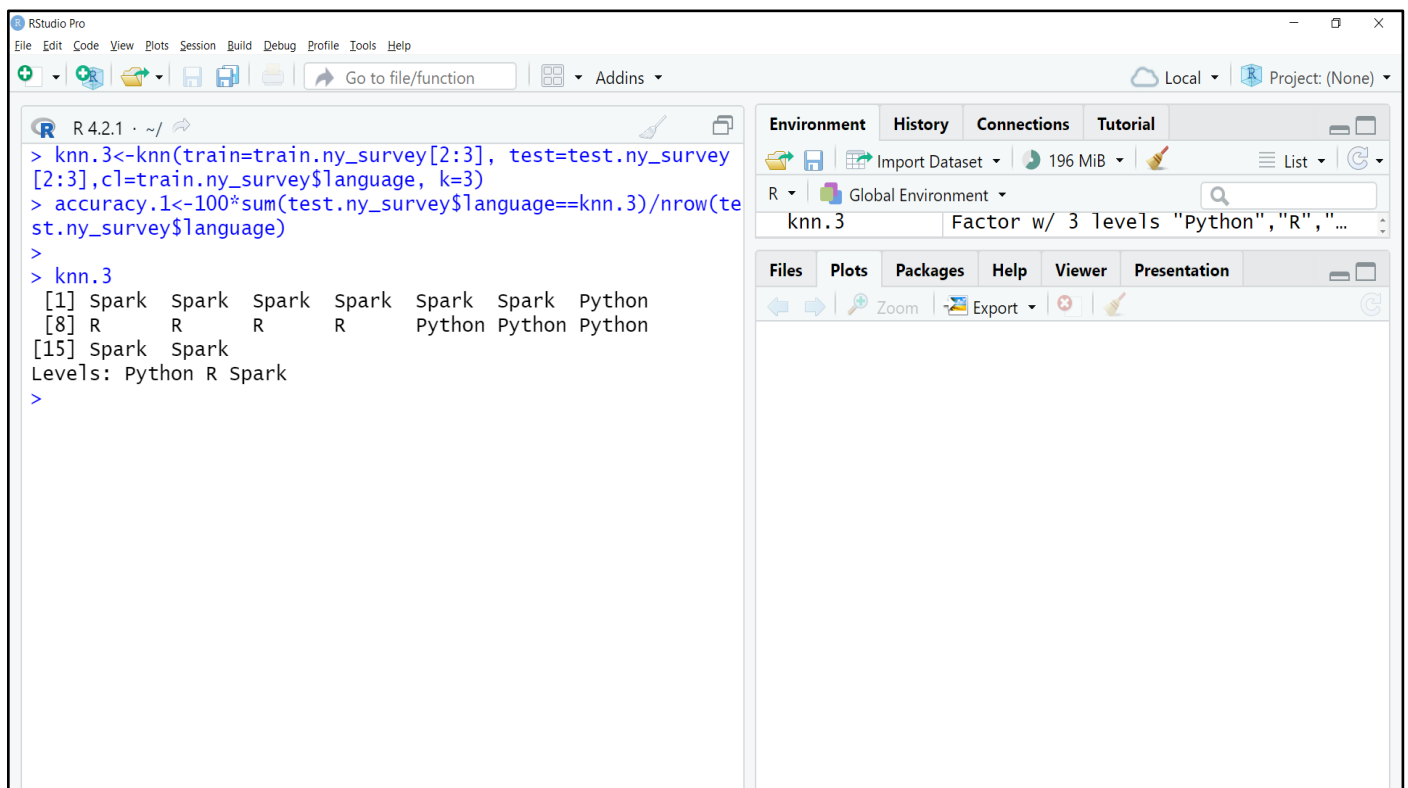
```
> knn.3<-knn(train=train.ny_survey[2:3], test=test.ny_survey[2:3],cl=train.ny_survey$language,
k=3)
```

```
> 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)
```



The screenshot displays the RStudio Pro interface. The main editor window on the left contains the following R code and its output:

```
> Value_k1<-knn(train=train.ny_survey[2:3],test=new_df,c1=train.ny_survey$language, k=1)
>
>
> print(Value_k1)
[1] Spark
Levels: Python R Spark
>
> Value_k2<-knn(train=train.ny_survey[2:3],test=new_df,c1=train.ny_survey$language, k=2)
>
>
> print(Value_k2)
[1] Spark
Levels: Python R Spark
>
> Value_k3<-knn(train=train.ny_survey[2:3],test=new_df,c1=train.ny_survey$language, k=3)
>
>
> print(Value_k3)
[1] Spark
Levels: Python R Spark
>
```

The right-hand pane shows the 'Environment' tab. It indicates the current environment is 'Global Environment' with a memory usage of 197 MiB. A single variable, 'Value_k3', is listed as a 'Factor w/ 3 levels "Python","R","..."'. Below this, there are tabs for 'Files', 'Plots', 'Packages', 'Help', 'Viewer', and 'Presentation', which are currently empty.

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 :

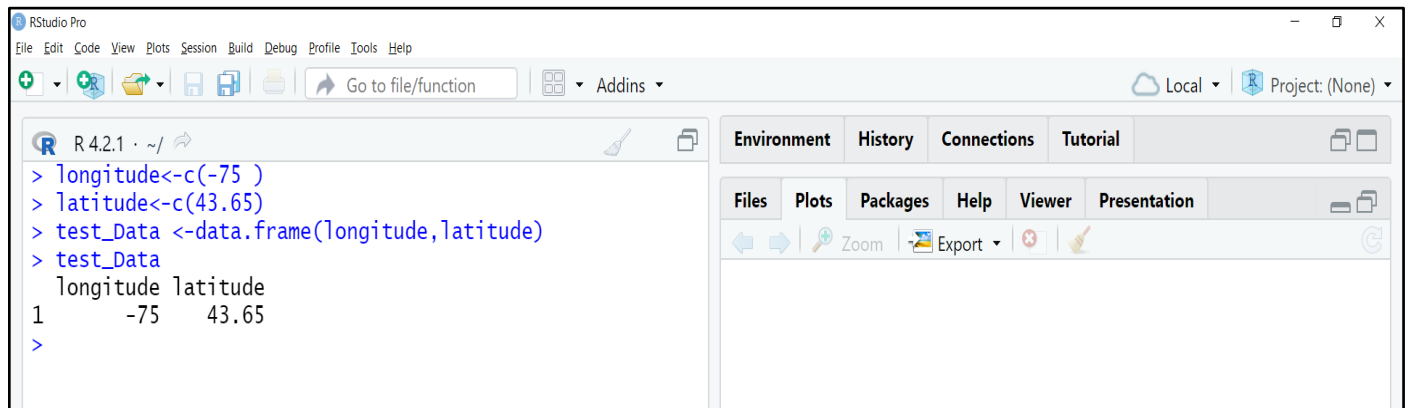
```
geom_point(data=testingDataFrame, shape=25, fill="blue", color="darkred", size=5)
```

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
```

The screenshot shows the RStudio Pro interface. The console window displays the following R code and its output:

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

The output is a data frame with three columns: `state`, `latitude`, and `longitude`. It contains 23 rows of data for New York, indexed from 2005 to 2027. The `state` column is constant at "New York".

| | state | latitude | longitude |
|------|----------|----------|-----------|
| 2005 | New York | 42.5142 | -79.7624 |
| 2006 | New York | 42.7783 | -79.0672 |
| 2007 | New York | 42.8508 | -78.9313 |
| 2008 | New York | 42.9061 | -78.9024 |
| 2009 | New York | 42.9554 | -78.9313 |
| 2010 | New York | 42.9584 | -78.9656 |
| 2011 | New York | 42.9886 | -79.0219 |
| 2012 | New York | 43.0568 | -79.0027 |
| 2013 | New York | 43.0769 | -79.0727 |
| 2014 | New York | 43.1220 | -79.0713 |
| 2015 | New York | 43.1441 | -79.0302 |
| 2016 | New York | 43.1801 | -79.0576 |
| 2017 | New York | 43.2482 | -79.0604 |
| 2018 | New York | 43.2812 | -79.0837 |
| 2019 | New York | 43.4509 | -79.2004 |
| 2020 | New York | 43.6311 | -78.6909 |
| 2021 | New York | 43.6321 | -76.7958 |
| 2022 | New York | 43.9987 | -76.4978 |
| 2023 | New York | 44.0965 | -76.4388 |
| 2024 | New York | 44.1349 | -76.3536 |
| 2025 | New York | 44.1989 | -76.3124 |
| 2026 | New York | 44.2049 | -76.2437 |
| 2027 | New York | 44.2413 | -76.1655 |

The RStudio interface also shows the Environment pane on the right, which is currently empty. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The bottom status bar shows the R version as 4.2.1 and the project as (None).

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)
```

