Job Openings in NYC (KNN Classification)

```
require(tidyverse)

## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse
1.2.1 --

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

require(class)

## Loading required package: class

nyc_data = read_csv('NYC_Jobs_2.csv')

## Parsed with column specification:
## cols(
##   `Work Location` = col_character(),
##   IT_Salary_From = col_double(),
##   IT_Salary_To = col_double(),
##   NonIT_Salary_from = col_double(),
##   NonIT_Salary_To = col_double(),
##   Annual_salary_from = col_integer(),
##   Annual_Salary_to = col_double(),
##   Daily_Salary_from = col_integer(),
##   Daily_Salary_to = col_integer(),
##   Hourly_Salary_from = col_integer(),
##   Hourly_Salary_to = col_integer(),
##   Annual_Salary_freq = col_integer(),
##   Daily_salary_freq = col_integer(),
##   Hourly_salary_freq = col_integer(),
##   Total_Opening = col_integer(),
##   Non_IT = col_integer(),
##   IT = col_integer(),
##   Full_Time = col_integer(),
##   Part_Time = col_integer()
## )

nyc_data
```

```
## # A tibble: 217 x 19
##    `Work Location` IT_Salary_From IT_Salary_To NonIT_Salary_fr~
##    <chr>                    <dbl>        <dbl>            <dbl>
##  1 1 Bay St., S.I~              0            0            54141
##  2 1 Centre St., ~         73148.       90358.           58897.
##  3 1 Centre Stree~              0            0           36546.
##  4 1 Court Square~              0            0           43246.
##  5 1 Fordham Plaz~              0            0           26156.
##  6 1 Metro Tech, ~              0            0           33875
##  7 1 Murray Hulbe~              0            0           48492.
##  8 1 Police Plaza~         39841        52045            50322.
##  9 10 Walker Rd, ~              0            0           27276.
## 10 100 Church St.~         56646.       85387.           46892.
## # ... with 207 more rows, and 15 more variables: NonIT_Salary_To <dbl>,
## #   Annual_salary_from <int>, Annual_Salary_to <dbl>,
## #   Daily_Salary_from <int>, Daily_Salary_to <int>,
## #   Hourly_Salary_from <int>, Hourly_Salary_to <int>,
## #   Annual_Salary_freq <int>, Daily_salary_freq <int>,
## #   Hourly_salary_freq <int>, Total_Opening <int>, Non_IT <int>, IT <int>,
## #   Full_Time <int>, Part_Time <int>
```

```
select_data = select(nyc_data, c(`Work Location`, IT, Non_IT))

salary = round((nyc_data$Annual_salary_from +  nyc_data$Annual_Salary_to) /2)


new_data = mutate(select_data, salary = salary)

location_openings =gather(new_data, IT_cat, Openings, 2:3)

category =as.numeric(as.factor(location_openings$IT_cat))

final_data_0 = mutate(location_openings, IT=category )
final_data = filter(final_data_0,Openings != 0)
final_data
```

```
## # A tibble: 255 x 5
##    `Work Location`                salary IT_cat Openings    IT
##    <chr>                           <dbl> <chr>     <int> <dbl>
##  1 1 Centre St., N.Y.              73626 IT           28     1
##  2 1 Police Plaza, N.Y.            68158 IT            2     1
##  3 100 Church St., N.Y.            67014 IT           17     1
##  4 100 Gold Street                 80828 IT           22     1
##  5 11 Metrotech Center Brooklyn N  66495 IT            2     1
##  6 110 William St. N Y             66416 IT            2     1
##  7 120-55 Queens Blvd, Queens Ny   55852 IT            2     1
##  8 120 Broadway, New York, NY      78828 IT            4     1
##  9 125 Worth Street, Nyc           61279 IT            2     1
## 10 130 Stuyvesant Place, S.I.      74115 IT            2     1
## # ... with 245 more rows
```

```
input = subset(final_data, select = c(salary, Openings))
label = final_data$IT_cat

input_n = sapply(input, function(x){(x-min(x))/(max(x)-min(x))})

location_dummies = model.matrix(~`Work Location`-1,data=final_data)

input_n_new = data.frame(input_n, location_dummies)
#input_n_new

set.seed(1234)
indices = sample(1:2, size=nrow(input_n_new), replace = T, prob = c(.8,.2))


data = data.frame(indices==1, input_n_new)

training_input = input_n_new[indices == 1,]
testing_input = input_n_new[indices == 2,]

training_label = label[indices==1]
testing_label = label[indices==2]

set.seed(1234)

#sqrt(nrow(training_input))

predications = knn(train = training_input,
test=testing_input,cl=training_label, k=13)

sum(predications==testing_label)/length(testing_label)

## [1] 0.8297872

table(predications,testing_label)

##             testing_label
## predications IT Non_IT
##       IT      1      1
##       Non_IT  7     38
```