Jianwen wu
Eco 20250
Prof. Foster

1)Group member: Crystal Hernandez, Keyi Long

2)Data:

```
> load("~/pums_NY.RData")
> attach(dat_pums_NY)
> summary(income_total[ (Hispanic == 1) & (Age>18) ])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -6400    5000   15000   25300   33000  746000
> x <-c(income_total[ (Hispanic == 1) & (Age>18) ])
> mean(x,na.rm=TRUE)
[1] 25298.76
> x1 <-mean(x,na.rm=TRUE)
> summary((Hispanic == 1) & (Age>18))
   Mode   FALSE    TRUE    NA's
logical  177653   18661       0
> n1 <- 18661
> sd(x)
[1] 38344.45
> sd1 <- sd(x)
> dat_NYC <- subset(dat_pums_NY, (dat_pums_NY$in_NYC == 1)&(dat_pums_NY$Age >= 18)&(dat_pums_NY
$Asian == 1))
> attach(dat_NYC)
> borough_f <- factor((in_Bronx + 2*in_Manhattan + 3*in_StatenI + 4*in_Brooklyn + 5*in_Queens), levels=c(1,2,
3,4,5),labels = c("Bronx","Manhattan","Staten Island","Brooklyn","Queens"))
> summary(borough_f)
       Bronx    Manhattan Staten Island      Brooklyn        Queens
         376          931          262         2268         4142
> xa <-c(income_total[borough_f])
> mean(xa,na.rm=TRUE)
[1] 30136.53
> x2 <-mean(xa,na.rm=TRUE)
> n2 <- 7979
> sd(xa)
[1] 37775.52
> sd2 <- sd(xa)
> (x1-x2)/sqrt((sd1^2/n1)+(sd2^2/n2))
[1] -9.53113
> z <- -9.53113
> pnorm(z)
[1] 7.779019e-22
> se <- sqrt(sd1*sd1/n1+sd2*sd2/n2)
> error <- qt(0.975,df=pmin(n1,n2)-1)*se
> left <- (x1-x2)-error
> right <- (x1-x2)+error
> left
[1] -5832.751
> right
[1] -3842.789
```

Explanation:
we compare Hispanic (older than 18) to Asian (older than 18) in NYC. We are
assuming that they have Asian has higher total income than Hispanic. Based on the data above, the average total income for Hispanic (older than 18) is 28628.04, the average total income for Asian (older than 18) is 30136.53. Therefore, the difference in average is -4837.77. The standard error of this difference is 507.5757. The 95% confidence interval is (-5832.751, -5832.751). Since the P-value is less than significant level in this case, they are statistically significantly different.

3)

```
> cor(rent_cost, income_total)
[1] 0.07373553
> mean(rent_cost)
[1] 624.7658
> mean(income_total)
[1] 37069.07
```

Explanation:
There is weak correlation between rent cost and income total. In other word, the increased of total income is not necessary affect the rent cost.

4)

```
> rm(list = ls(all = TRUE)) # clear workspace
> setwd("~/Dropbox/CCNY/Statistics and Intro Econometrics/R Projects/PUMSdata-hw1")
> load("pums_NY.RData")
> head(dat_pums_NY)
  Age female PERNUM educ_nohs educ_hs educ_smcoll educ_as educ_bach educ_adv
1  43      1      1         0       0           0       0         0         1
2  45      0      2         0       0           1       0         0         0
3  33      0      1         0       1           0       0         0         0
4  57      0      1         0       1           0       0         0         0
5  52      1      2         1       0           0       0         0         0
6  26      0      3         0       0           0       1         0         0
  ANCESTR1D ANCESTR2D immig Hispanic Hisp_Mex Hisp_PR Hisp_Cuban Hisp_DomR
1      2610      9990     0        1        0       1          0         0
2       511      9990     0        0        0       0          0         0
3       880      9990     0        0        0       0          0         0
4      7060      9990     1        0        0       0          0         0
5      7060      9990     1        0        0       0          0         0
6      7060      9990     1        0        0       0          0         0
  white AfAm Amindian Asian race_oth Married divwidsep unmarried veteran
1     1    0        0     0        0       1         0         0       0
2     1    0        0     0        0       1         0         0       0
3     1    0        0     0        0       0         0         1       0
4     0    0        0     1        0       1         0         0       0
5     0    0        0     1        0       1         0         0       0
6     0    0        0     1        0       0         0         1       0
  has_AnyHealthIns has_PvtHealthIns Commute_car Commute_bus Commute_subway
1                1                1          1          0              0
2                1                1          1          0              0
3                1                1          1          0              0
4                0                0          0          0              0
5                0                0          0          0              0
6                0                0          0          0              0
```

```
  Commute_rail Commute_other below_povertyline below_150poverty
1            0             0                 0                 0
2            0             0                 0                 0
3            0             0                 0                 0
4            0             0                 1                 1
5            0             1                 1                 1
6            0             0                 1                 1
  below_200poverty foodstamps work_fullyr income_total income_wagesal
1                0          1          1       110000         110000
2                0          1          1        39000         39000
3                0          1          1        72000         72000
4                1          1          0            0             0
5                1          1          0         7000          7000
6                1          1          0            0             0
  HH_income owner_cost rent_cost occ_dum ind_dum in_NYC PUMA in_Bronx
1         0       2850         0    1820    7860      0 3106        0
2         0       2850         0    1550    3390      0 3106        0
3     72000          0       430    4210     770      0  100        0
4      7000          0       900       0       0      1 4103        0
5      7000          0       900    4520    8980      1 4103        0
6      7000          0       900       0       0      1 4103        0
  in_Manhattan in_StatenI in_Brooklyn in_Queens in_Westchester in_Nassau
1            0          0           0          0             1         0
2            0          0           0          0             1         0
3            0          0           0          0             0         0
4            0          0           0          1             0         0
5            0          0           0          1             0         0
6            0          0           0          1             0         0
  ROOMS BUILTYR2 UNITSSTR
1     8       10        4
2     8       10        4
3     2        9        3
4     3        5       10
5     3        5       10
6     3        5       10
> norm_varb <- function(X_in) {
+     (X_in - mean(X_in, na.rm = TRUE))/sd(X_in, na.rm = TRUE)
+ }
> dat_NYC <- subset(dat_pums_NY, (dat_pums_NY$in_NYC == 1)&(dat_pums_NY$Age >
20)&(dat_pums_NY$Age < 66))
> attach(dat_NYC)
> borough_f <- factor((in_Bronx + 2*in_Manhattan + 3*in_StatenI + 4*in_Brooklyn + 5*in_Queens),
levels=c(1,2,3,4,5),labels = c("Bronx","Manhattan","Staten Island","Brooklyn","Queens"))
> housing_cost <- owner_cost+rent_cost
> norm_inc_tot <- norm_varb(income_total)
> norm_housing_cost <- norm_varb(housing_cost)
>
> data_use <- data.frame(norm_inc_tot,norm_housing_cost)
> good_obs_data_use <- complete.cases(data_use,borough_f)
> dat_use <- subset(data_use,good_obs_data_use)
> y_use <- subset(borough_f,good_obs_data_use)
> detach(dat_NYC)
> set.seed(12345)
> NN_obs <- sum(good_obs_data_use == 1)
> select1 <- (runif(NN_obs) < 0.9)
> train_data <- subset(dat_use,select1)
```

```
> test_data <- subset(dat_use,(!select1))
> cl_data <- y_use[select1]
> true_data <- y_use[!select1]
> summary(cl_data)
       Bronx     Manhattan Staten Island      Brooklyn        Queens
        5568          5546          2095         13443         11915
> prop.table(summary(cl_data))
       Bronx     Manhattan Staten Island      Brooklyn        Queens
  0.14437213    0.14380170    0.05432105    0.34856224    0.30894288
> summary(train_data)
 norm_inc_tot       norm_housing_cost
 Min.   :-0.749438   Min.   :-1.338548
 1st Qu.:-0.535068   1st Qu.:-0.623880
 Median :-0.270258   Median :-0.232514
 Mean   :-0.002506   Mean   :-0.002493
 3rd Qu.: 0.186855   3rd Qu.: 0.439614
 Max.   :13.900225   Max.   : 8.552795
> require(class)
Loading required package: class
> for (indx in seq(44, 88, by= 4)) {
+     pred_borough <- knn(train_data, test_data, cl_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)
+
+     num_correct_labels <- sum(pred_borough == true_data)
+     correct_rate <- num_correct_labels/length(true_data)
+     print(c(indx,correct_rate))
+
+ }
[1] 44.0000000  0.3585084
[1] 48.0000000  0.3589804
[1] 52.0000000  0.3580363
[1] 56.0000000  0.3644088
[1] 60.0000000  0.3639367
[1] 64.0000000  0.3639367
[1] 68.0000000  0.3618126
[1] 72.0000000  0.3596885
[1] 76.0000000  0.3580363
[1] 80.0000000  0.3596885
[1] 84.0000000  0.3625207
[1] 88.0000000  0.3618126
> for (indx in seq(44, 55, by= 1)) {
+     pred_borough <- knn(train_data, test_data, cl_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)
+
+     num_correct_labels <- sum(pred_borough == true_data)
+     correct_rate <- num_correct_labels/length(true_data)
+     print(c(indx,correct_rate))
+
+ }
[1] 44.0000000  0.3596885
[1] 45.0000000  0.3603965
[1] 46.0000000  0.3587444
[1] 47.0000000  0.3582724
[1] 48.0000000  0.3582724
[1] 49.0000000  0.3568563
[1] 50.0000000  0.3589804
[1] 51.0000000  0.3596885
[1] 52.0000000  0.3570923
```

[1] 53.0000000 0.3578003
[1] 54.0000000 0.3627567
[1] 55.0000000 0.3618126
>

Explanation:

From the prob. Of every class, we assume most data is should be “Brooklyn”. And we want to predict the class after 44, and we get the output and match our prediction.