Adam Hendel
DS 710 Homework 8
R assignment

Code, Solutions, Comments Provided in BOLD

1. In this problem, you will create and apply a function that rates cities based on how appealing they are for you to live in.

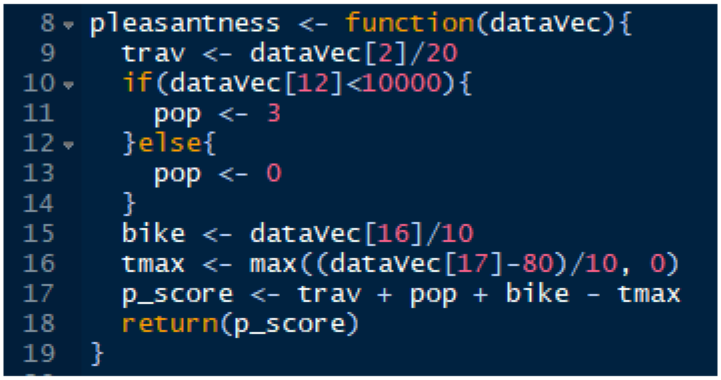
a. What factors are important to you in deciding where to live? The data set Best Cities.csv contains data on 10 U.S. cities, obtained from http://www.census.gov/quickfacts/, www.walkscore.com, and http://www.wunderground.com/. Develop your own formula to rate how pleasant a city would be to live in, in your opinion, based on the variables in this data set.

I wouldn't want to live in a super densely populated city, so if the population density is under 10,000 I will give the city 3 points, otherwise 0 points.

I also like use bicycle as transportation, so I will award a point for ever 10 values on the 'Bikeability' scale.

I don't mind the cold at all, but if it is a positive if the city have days above 80 degrees. Thus, I will award a point for every 10 degrees that the cities max temp is above 80.

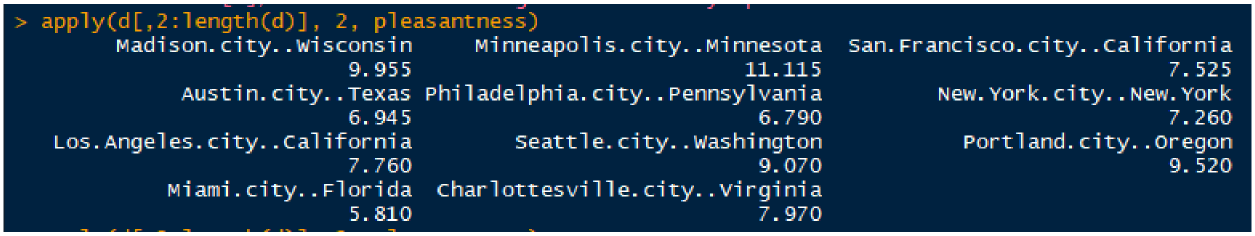
b. Create an R function that computes the pleasantness score of a city, based on a vector of data about it. You may assume that the vector contains data in the same order as it is listed in Best Cities.csv.



```
pleasantness <- function(datavec){
  trav <- datavec[2]/20
  if(datavec[12]<10000){
    pop <- 3
  }else{
    pop <- 0
  }
  bike <- datavec[16]/10
  tmax <- max((datavec[17]-80)/10, 0)
  p_score <- trav + pop + bike - tmax
  return(p_score)
}
```

c.Use apply() to apply your function to each city in the Best Cities.csv data set. Based on your criteria, which city is the best for you? Does this assessment seem accurate? If not, what would you want to change about your formula?

The assessment seems accurate; I enjoy living near Minneapolis but have also considered Seattle and almost moved to Madison a few years ago. If I were to improve it I would include more variables that might help differentiate cities like Portland and Seattle from one another.



```
> apply(d[,2:length(d)], 2, pleasantness)
        Madison.city..Wisconsin     Minneapolis.city..Minnesota  San.Francisco.city..California
                          9.955                          11.115                           7.525
             Austin.city..Texas Philadelphia.city..Pennsylvania         New.York.city..New.York
                          6.945                           6.790                           7.260
   Los.Angeles.city..California        Seattle.city..Washington           Portland.city..Oregon
                          7.760                           9.070                           9.520
            Miami.city..Florida  Charlottesville.city..Virginia
                          5.810                           7.970
```

2. Can we use statistical analysis of word lengths to identify the author of an anonymous essay? In Homework 7, you wrote a Python function that counted the lengths of words in the 1770 essay by “A Mourner”. Analysis of other articles published in *The Boston Gazette and Country Journal* in early 1770 finds that John Hancock wrote a 121-word article with a mean word length of 4.69 and standard deviation of 2.60.

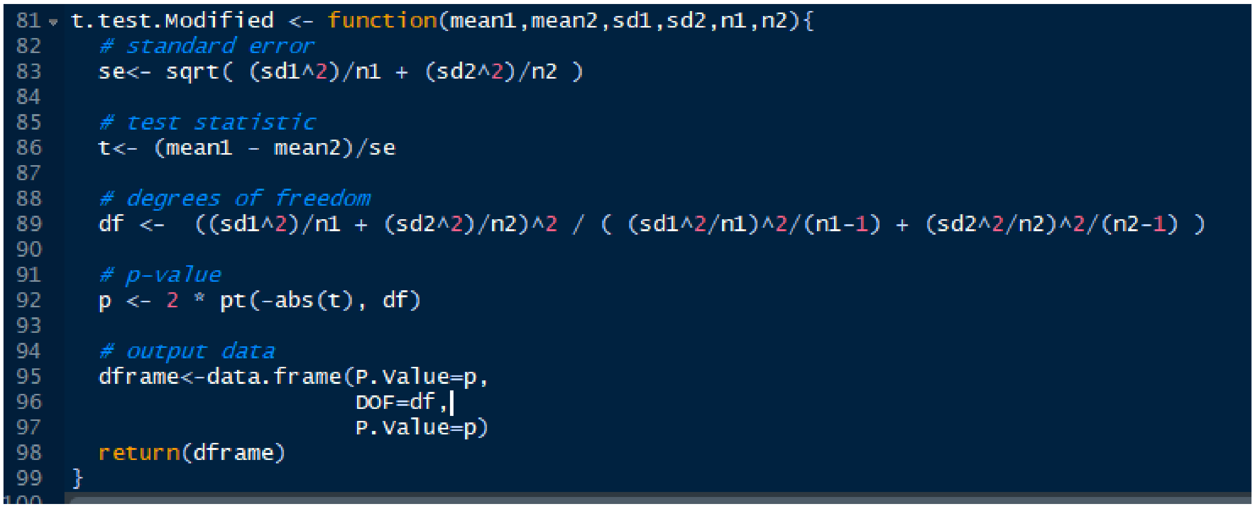
a. We want to use R to assess whether it is plausible that John Hancock was “A Mourner”, based on his mean word length. Explain why a 2-sided, 2-sample t-test is appropriate for this.

We want to test whether the mean between “A Mourner” is different from the other article John Hancock wrote. That is, don’t want to simply know if it is greater than or less than, rather greater than or less than. Thus, we will use the two-sided t test.

b. Explain why the t.test() function is *not* appropriate for the data we have available.

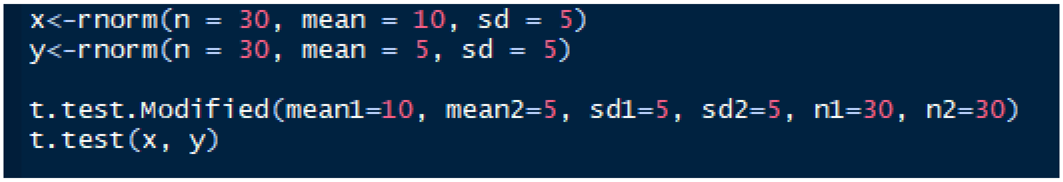
The two sample sizes and variances are not equal.

c. Write your own function for performing a 2-sided, 2-sample t-test for equality of means when the raw data are not available. Use the following information. (If the formulas do not show up correctly, please view the pdf version of the assignment.)

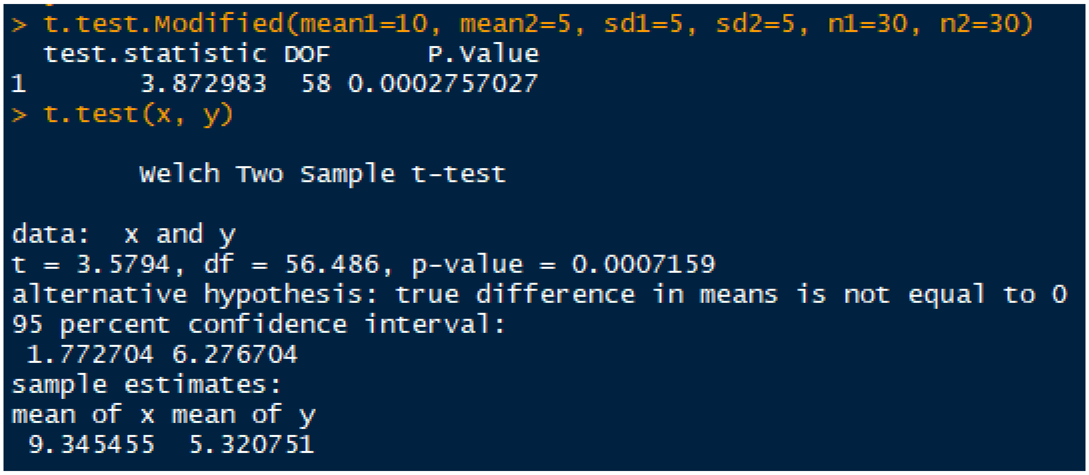


```r
t.test.Modified <- function(mean1,mean2,sd1,sd2,n1,n2){
  # standard error
  se<- sqrt( (sd1^2)/n1 + (sd2^2)/n2 )

  # test statistic
  t<- (mean1 - mean2)/se

  # degrees of freedom
  df <-   ((sd1^2)/n1 + (sd2^2)/n2)^2 / ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-1) )

  # p-value
  p <- 2 * pt(-abs(t), df)

  # output data
  dframe<-data.frame(P.Value=p,
                     DOF=df,
                     P.Value=p)
  return(dframe)
}
```

d. Test your function by comparing it to t.test() on a pair of samples. You may wish to use rnorm() to generate random data from a normal distribution. If the p-value from your function doesn’t match the p-value from t.test(), then revise your code from part c.



```
x<-rnorm(n = 30, mean = 10, sd = 5)
y<-rnorm(n = 30, mean = 5, sd = 5)

t.test.Modified(mean1=10, mean2=5, sd1=5, sd2=5, n1=30, n2=30)
t.test(x, y)
```



```
> t.test.Modified(mean1=10, mean2=5, sd1=5, sd2=5, n1=30, n2=30)
  test.statistic DOF      P.Value
1       3.872983  58 0.0002757027
> t.test(x, y)

        Welch Two Sample t-test

data:  x and y
t = 3.5794, df = 56.486, p-value = 0.0007159
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.772704 6.276704
sample estimates:
mean of x mean of y
 9.345455  5.320751
```

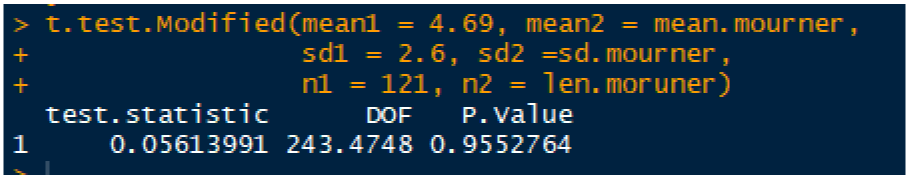
The two p-values are very close and the code seems to be correct. Further testing indicates correct form.

e. Apply your function to assess whether it is plausible that Hancock was A Mourner. Write your conclusion as a sentence.

Note: The null hypothesis for a 2-sample t-test of this question is

$$H_0\colon \mu_{Mourner} = \mu_{Hancock}$$

i.e., that A Mourner and Hancock have the *same* mean word length. In other words, the null hypothesis is that it *is* plausible that Hancock was "A Mourner."



```
> t.test.Modified(mean1 = 4.69, mean2 = mean.mourner,
+                 sd1 = 2.6, sd2 =sd.mourner,
+                 n1 = 121, n2 = len.moruner)
  test.statistic      DOF   P.Value
1     0.05613991 243.4748 0.9552764
```

With a P value of 0.95, there is not enough evidence to reject the null hypothesis that the two means are equal. Therefore, it is not plausible that John Hancock is "A Mourner".