BS730 – Spring 2021
Project 1

Details

This project must be completed individually – no collaboration at all is allowed between students on any aspect of the project. The project is worth 15% of your final grade. We encourage you to work as independently as possible on the project although you may ask the teaching team for assistance with programming issues when absolutely necessary. Your project submission should include completed tables, write-up and any graphs for each part. Your R code needs to be easily readable (neat and tidy), clearly labelled and included as an Appendix. Please submit on Blackboard by the beginning of Class 9. Late submissions are penalized 5 points per day unless prior arrangements have been made with your course instructor. No late projects will be accepted after three days past the due date and time.

Background

The Jackson Heart Study (JHS), which initiated in 1998, is a single-site, prospective, epidemiologic investigation of cardiovascular disease among African Americans. The JHS is funded by the National Heart, Lung and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). It is a population-based longitudinal study. The JHS recruited 5306 African American residents living in the Jackson, Mississippi, metropolitan area of Hinds, Madison, and Rankin Counties. Recruitment was limited to non-institutionalized adult African American men and women, 35-84 years old, except in a nested family cohort where those 21 to 34 years of age were also eligible. The final cohort of participants enrolled during the baseline exam included 6.6% of all African American men and women residents of the Jackson Mississippi Metropolitan Statistical Area aged 35-84 (N=76,426, US Census 2000). Because there is a greater prevalence of cardiovascular disease among African Americans compared to other ethnic groups, the purpose of the Jackson Heart Study is to explore the reasons for this disparity and to uncover new approaches to reduce it. (1)

The JHS TRANS-Data Package consists of a de-identified set of JHS data that has been modified by steps such as: creating a set of anonymized IDs, including a ~50% random sample of participants consenting to all study data usage, truncating low frequency classifications for categorical variables (<5), and setting all dates to the 15th day of the month. The JHS TRANS-Data Package is intended to be TRANSformative for research and research education. (2)

Diet is a potentially modifiable factor that contributes to cardiovascular disease (CVD). Prior studies have shown that African Americans, particularly those residing in the southeastern US, have dietary patterns that differ from the general US population. Study of individual level and neighborhood level variables may help in understanding factors, including socioeconomic measures and access to healthy food, that may influence dietary patterns in African Americans. (3)

The current project will use a subset of the JHS TRANS-Data package. The aims of the project are:

1. To characterize consumption of dark green vegetables among JHS participants included in the TRANS-Data package at baseline.

2. To examine the association between various factors and consumption of dark green vegetables among JHS participants included in the TRANS-Data package.

A description of the variables in the dataset "**jhst_proj1sp21.csv**" can be found on page 3.

References

(1) *Wyatt SB, Diekelmann N, Henderson F, Andrew ME, Billingsley G, Felder SH et al. A community-driven model of research participation: the Jackson Heart Study Participant Recruitment and Retention Study. Ethn Dis 2003; 13(4):438-455.*

(2) https://www.jacksonheartstudy.org/Research/Data-Science

(3) Gao Y, Hickson DMA, Talegawkar S, et al. Influence of individual life course and neighbourhood socioeconomic position on dietary intake in African Americans: the Jackson Heart Study. BMJ Open 2019;9:e025237. doi:10.1136/bmjopen-2018-025237

Data Dictionary

Variable Name	**Variable Type in R**	**Description**	**Coding**
Subjid	integer	Participant ID	Integer value between 1 and 2750
Age	Numeric	Participant age	Range 22.9 to 83.2 years
Sex	Character	Participant sex	Female Male
Darkgrnveg	Numeric	Daily servings dark-green vegetables consumed	Range 0.01 to 4.27
Fish	Numeric	Fish consumed	Range 0.0 to 2.99
Eggs	Numeric	Eggs consumed	Range 0.0 to 3.36
VitaminD2	Numeric	25(OH) Vitamin D2 (ng/mL)	Range 0.0 to 17.8
VitaminD3	Numeric	25(OH) Vitamin D3 (ng/mL)	Range 3.3 to 33.3
BMI	Numeric	Body mass index (kg/m^2) at visit 2	Range 18.4 to 58.80 kg/m^2
HTN	Character	Hypertension	Yes No
Diabetes	integer	Diabetes	1=Yes 0=No 9= Don't know
Currentsmoker	integer	Indication of participant's current cigarette smoking status	1=Yes 0=No 9= Don't know
Activeindex	Numeric	Active Living Index	Range 1.0 to 4.5
Dailydiscr	Numeric	everyday experiences: score	Range 1.0 to 5.67
Perceivedstress	integer	total global stress score	Range 0 to 19
Depression	integer	Total Depressive Symptoms Score	Range 0 to 41
Income	integer	Income category at visit 1	1=Affluent 2=Upper-middle 3=Lower-middle 4=Poor 999= Don't know
nbmedHHincome	integer	Median Household Income in Census Tract	Range $13,110 to $104,707
nbpctPoverty	Numeric	% Below Poverty in Census Tract	Range 0.01 to 0.53
nbSESanascore	Numeric	Census Tract SES Score (Diez-Roux 1990)	Range -8.91 to 8.42
nbK3FavorFoodstore	Numeric	Density of Favorable Food Stores (3 Mile Kernel)	Range 0.0 to 0.72
nbK3paFacilities	Numeric	Density of Physical Activity Facilities (3 Mile Kernel)	Range 0.0 to 1.74

Part 1. Data Manipulation (30 points total)

A. Read the dataset **jhst_proj1SP21.csv** into R as a dataframe called "*proj1*". (3 points)

B. Within *proj1* make the following changes:
   i. For all variables, recode any values of "Don't know" to missing (NA). (3 points)
   ii. Using the coding in the table below, create the following five new variables: *Agecat*, *darkgrnVegQ2*, *darkgrnVegQ3*, *income2cat, Desert*. Note that *Darkgrnveg*, *Income*, and *nbK3FavorFoodstore* columns have missing values. (20 points)
   iii. Exclude participants who have missing values for any of the following variables: *Darkgrnveg*, *Income*, *nbmedHHincome*, *nbpctPoverty*, *nbSESanascore*, *nbK3FavorFoodstore* , *nbK3paFacilities* , *Activeindex*, *Dailydiscr* ,and *Perceivedstress* (4 points)

Original Variable	**New Variable**	**Coding for New Variable**
Age	Agecat	1 is for <= 30 years 2 is for >30 to <=50 years 3 is for >50 to <=70 years 4 is for >70 years
Darkgrnveg	darkgrnvegQ2	1: >= the median value of darkgrnveg 0: < the median value of darkgrnveg (Assume the median value of Darkgrnveg = .5)
Darkgrnveg	darkgrnvegQ3	1: >= the third quartile of darkgrnveg 0: < the third quartile of darkgrnveg (Assume the third quartile value of Darkgrnveg = 1.03)
Income	income2cat	1: Higher income (Affluent or Upper-middle) 0: Lower income (Lower-middle or Poor)
nbK3FavorFoodstore	Desert	1: nbK3FavorFoodstore = 0 0: nbK3FavorFoodstore >0

Part 2. Statistical Analysis (65 points total)

A. Fill in the table below with the appropriate summary statistics and p-values. (35 points)

Table 1. Distribution of risk factors stratified by dark green vegetable consumption (based on darkgrnvegQ2) among Jackson Heart Study participants measured at baseline in the TRANS data set.

	Total Participants (N=2121)	**Dark Green Vegetable Consumption (darkgrnQ2)**		**P-value**
		Low (N=1076 50.7%)	**High (N= 1045 49.2%)**	
Demographics				
Age (years) = mean(sd)	54.2(11.6)	53.8(11.8)	54.4(11.3)	0.2118
Age category = n(%)				0.6079
< 30 years	43(2%)	25(2.3%)	18(1.7%)	
30-50 years	745(35%)	387(35.9%)	358(34.2%)	
50-70 years	1166(54%)	579(53.8%)	587(56.1%)	
>70 years	167(78%)	85(7.89%)	82(7.8%)	
Male sex = n(%)	780(36%)	459(36.2%)	483(37.3%)	
Individual Measures				
Body mass index (kg/m^2) =mean(sd)	31.89(7.04)	31.30(6.6)	32.50(7.43)	8.106e-05
Activity Index = mean(sd)	2.1(0.79)	2.03(0.77)	2.18(0.808)	3.011e-05
Daily Discrimination = mean(sd)	2.12(1)	2.09(0.98)	2.15(1.02)	0.1706
Perceived Stress = mean(sd)	5.32(4.3)	5.2(4.37)	5.44(4.32)	0.1979
High income (based on income2cat) = n(%)	1437(67.7%)	686(47.7%)	751(52.2%)	6.455e-05
Neighborhood Measures				
Median Household Income in Census Tract = mean(sd)	34298.43(15803.7)	33371.95(15321.47)	35252.41(16237.48)	0.006166
% Below Poverty in Census Tract = mean(sd)	0.23(0.12)	0.23(0.12)	0.22(0.12)	0.007955
Census Tract SES Score (Diez-Roux 1990)mean(sd)	-2.38(3.77)	-2.63(3.72)	-2.12(3.8)	0.00176
Density of Favorable Food Stores (3 Mile Kernel) mean(sd)	0.26(0.2)	0.26(0.21)	0.26(0.2)	0.7189
Density of Physical Activity Facilities (3 Mile Kernel) nmean(sd)	0.44(0.35)	0.43(0.35)	0.45(0.35)	0.2999

B. Based on your results in Table 1, which factors showed a statistically significant association with dark green vegetable consumption category? Specify the direction for each significant association. (6 points)
Ans-
The statistically significant values observed were:

1. BMI: Mean BMI is less for low dark green vegetable consumption.
2. Activity Index: Mean is less for low level dark green vegetable consumption
3. High income: Count of High income(Affluent and upper middle) is less as compared to low income(poor and lower middle)
4. Median Household income in census tract: Again mean is low for low level dark green vegetable consumption
5. % below poverty in census tract: The mean is greater for low level dark green vegetable consumption
6. Census tract SES score: Mean is low for low level dark green vegetable consumption

C. You would like to further explore the relationship between the density of favorable food stores within a three mile kernel and the consumption of dark green vegetables. You are specifically interested in whether the density of favorable food stores is associated with a particularly high level of consumption. Test if the mean density of favorable food stores differs between those with the highest quartile of consumption compared to others (use the variable darkgrnvegQ3). Report the null and alternative hypotheses, name of the test, test statistic, p-value, the appropriate measure of effect (+ 95% confidence interval), and your conclusion. But sure to describe all steps taken to determine the appropriate test statistic and p-value. (12 points)

Ans –
The overall mean (sd) density of favorable food stores within three mile kernel is 0.26(0.2). The calculated mean (sd) with respect to dark green vegetables consumption for higher group was approximately same for both higher and lower groups. The variance for both the groups were equal as the evaluated ratio of variances came out to be 0.99 with p value 0.96 (>0.05).
Null Hypothesis (Ho) = the mean of both groups are equal. Ho: µ1 = µ2
Alternate Hypothesis (H1) = the mean of both groups are not equal. H1: µ1 ≠ µ2.
The level of significance was kept at 5%.
The test used to compute was two sample t test with test statistics as -0.62936 and p value as 0.5292. The 95% confidence interval was calculated as from 0.027 to 0.01.
As the p value was > 0.05, we fail to reject null hypothesis and can thus conclude that the mean of both the high and low dark green vegetables are equal with comparison to density of favorable food stores within three mile kernel.

D. Another nutrient of interest is Vitamin D. Vitamin D is actually a family of nutrients, with vitamins D2 and D3 being the most common forms in humans. Vitamin D3 comes from animal sources and vitamin D2 from plant sources. Your skin can also produce D3 in response to sunlight or UV light. In our diet D2 typically comes from fortified foods, although mushrooms grown in UV light are also a source. Both D2 and D3 were measured in Jackson Heart Study participants and we are interested in knowing if these levels differ. Test the hypothesis that the mean difference in vitamin D3 versus D2 is zero at the baseline visit for Jackson Heart Study participants. Report the null and alternative hypotheses, name of test, test statistic, p-value, the appropriate measure of effect (+ 95% confidence interval), and your conclusion. (12 points)

Ans –

The mean for Vitamin D2 is 3.77 and for Vitamin D3 is 13.10. For statistical analysis, I performed paired t test between both the variables.

Null hypothesis (Ho): mean difference between Vitamin D2 and D3 is 0. Ho: µd = 0

Alternate hypothesis (h1): mean difference Vitamin D2 and D3 is not equal to 0. H1: µd ≠ 0.

The level of significance was 5%. The sample mean of D2 Vitamin was less by 10.5nf/mL as compared to D3. Paired t test was performed. The test statistics = -44.912 with 1031 degree of freedom and p value = < 2.2e-16. As the p value is less than 0.05, we reject the null hypothesis and conclude that there is significant evidence of mean difference for both the D2 and D3 vitamin groups. The 95% confidence interval was from -10.9ng/mL to -10.4ng/mL.

PASTE YOUR CODE FOR PARTS 1-2 HERE. (5 points)

```
setwd("C:/Users/Neha/Desktop/BU/bs730/project1")

## Part 1
## Ques A
proj1<- read.csv('jhst_proj1sp21(1) (1).csv', header = TRUE)


## Ques B
proj1$agecat <- cut(x = proj1$Age, breaks = c(0, 30,50, 70, Inf), labels = c(1,2,3,4))
table(proj1$agecat) #for category of age


proj1$darkgrnvegQ2 <- ifelse(proj1$Darkgrnveg >= 0.5, 1, 0)
table(proj1$darkgrnvegQ2) #for dark green vegetable category

proj1$darkgrnvegQ3 <- ifelse(proj1$Darkgrnveg >= 1.03, 1,0)
table(proj1$darkgrnvegQ3)


proj1$income2cat[proj1$Income == 1 | proj1$Income ==2] <- 1
proj1$income2cat[proj1$Income == 3 | proj1$Income ==4] <- 0
table(proj1$income2cat) #for income category

#Dessert
proj1$desert <- ifelse(proj1$nbK3FavorFoodstore >0, 0, 1)
table(proj1$desert) #for desert category

#the don't know values as NA
proj1$Diabetes[proj1$Diabetes == 9] <- NA
```

```
proj1$Currentsmoker[proj1$Currentsmoker == 9] <- NA
proj1$Income[proj1$Income == 999] <- NA

#excluding the missing values from the table
proj1 <- proj1[(!is.na(proj1$Darkgrnveg)) & (!is.na(proj1$Income)) &
(!is.na(proj1$nbmedHHincome)) &
             (!is.na(proj1$nbpctPoverty)) & (!is.na(proj1$nbSESanascore)) &
             (!is.na(proj1$nbK3FavorFoodstore)) & (!is.na(proj1$nbK3paFacilities)) &
             (!is.na(proj1$Activeindex)) & (!is.na(proj1$Dailydiscr)) &
             (!is.na(proj1$Perceivedstress)), ]


dim(proj1)##2121,27
number = table(proj1$darkgrnvegQ2) #total number
prop.table(number) #low and high level dark green vegetable number

## Part 2
#Ques A
#1st row age(years)
mean(proj1$Age)
sd(proj1$Age)
tapply(proj1$Age, proj1$darkgrnvegQ2, FUN = mean)
tapply(proj1$Age, proj1$darkgrnvegQ2, FUN = sd)
t.test(proj1$Age ~ proj1$darkgrnvegQ2, var.equal = TRUE)#aasume variance is true for all in this
part


##2nd row
#table age cat wise
agecat_table <- table(proj1$agecat)
agecat_table
prop.table(agecat_table)

#table consumption wise
agecat_consumption <- table(proj1$agecat, proj1$darkgrnvegQ2)
agecat_consumption
prop.table(agecat_consumption, 2)
chisq.test(proj1$agecat , proj1$darkgrnvegQ2, correct = FALSE)


##male sex

male <- table(proj1$Sex == 'Male')
male
prop.table(male)
male_consumption <- table(proj1$Sex=='Male', proj1$darkgrnvegQ2)
male_consumption
prop.table(male_consumption,2)
```

```
t.test(proj1$Sex=='Male' ~ proj1$darkgrnvegQ2, var.equal = TRUE)


##BMI
mean(proj1$BMI)
sd(proj1$BMI)
tapply(proj1$BMI, proj1$darkgrnvegQ2, FUN = mean)
tapply(proj1$BMI, proj1$darkgrnvegQ2, FUN = sd)
t.test(proj1$BMI ~ proj1$darkgrnvegQ2, var.equal = TRUE)

##activity index
mean(proj1$Activeindex)
sd(proj1$Activeindex)
tapply(proj1$Activeindex, proj1$darkgrnvegQ2, FUN = mean)
tapply(proj1$Activeindex, proj1$darkgrnvegQ2, FUN = sd)
t.test(proj1$Activeindex ~ proj1$darkgrnvegQ2, var.equal = TRUE)

##Daily discrimination
mean(proj1$Dailydiscr)
sd(proj1$Dailydiscr)
tapply(proj1$Dailydiscr, proj1$darkgrnvegQ2, FUN = mean)
tapply(proj1$Dailydiscr, proj1$darkgrnvegQ2, FUN = sd)
t.test(proj1$Dailydiscr ~ proj1$darkgrnvegQ2, var.equal = TRUE)


##percieved stree
mean(proj1$Perceivedstress)
sd(proj1$Perceivedstress)
tapply(proj1$Perceivedstress, proj1$darkgrnvegQ2, FUN = mean)
tapply(proj1$Perceivedstress, proj1$darkgrnvegQ2, FUN = sd)
t.test(proj1$Perceivedstress ~ proj1$darkgrnvegQ2, var.equal = TRUE)


##high income
high_income <- table(proj1$income2cat == 1)
high_income
prop.table(high_income)
highincome_cons <- table(proj1$income2cat == 1, proj1$darkgrnvegQ2)
highincome_cons
prop.table(highincome_cons, 1)
chisq.test(proj1$income2cat, proj1$darkgrnvegQ2, correct = FALSE)

##household income
mean(proj1$nbmedHHincome)
sd(proj1$nbmedHHincome)
tapply(proj1$nbmedHHincome, proj1$darkgrnvegQ2, FUN= mean)
tapply(proj1$nbmedHHincome, proj1$darkgrnvegQ2, FUN= sd)
t.test(proj1$nbmedHHincome ~ proj1$darkgrnvegQ2, var.equal = TRUE)
```

```
##%below poverty
mean(proj1$nbpctPoverty)
sd(proj1$nbpctPoverty)
tapply(proj1$nbpctPoverty, proj1$darkgrnvegQ2, FUN= mean)
tapply(proj1$nbpctPoverty, proj1$darkgrnvegQ2, FUN= sd)
t.test(proj1$nbpctPoverty ~ proj1$darkgrnvegQ2, var.equal = TRUE)


##SES score
mean(proj1$nbSESanascore)
sd(proj1$nbSESanascore)
tapply(proj1$nbSESanascore, proj1$darkgrnvegQ2, FUN= mean)
tapply(proj1$nbSESanascore, proj1$darkgrnvegQ2, FUN= sd)
t.test(proj1$nbSESanascore ~ proj1$darkgrnvegQ2, var.equal = TRUE)



##  Favourable food
mean(proj1$nbK3FavorFoodstore)
sd(proj1$nbK3FavorFoodstore)
tapply(proj1$nbK3FavorFoodstore, proj1$darkgrnvegQ2, FUN= mean)
tapply(proj1$nbK3FavorFoodstore, proj1$darkgrnvegQ2, FUN= sd)
t.test(proj1$nbK3FavorFoodstore ~ proj1$darkgrnvegQ2, var.equal = TRUE)

##Physical activity
mean(proj1$nbK3paFacilities)
sd(proj1$nbK3paFacilities)
tapply(proj1$nbK3paFacilities, proj1$darkgrnvegQ2, FUN= mean)
tapply(proj1$nbK3paFacilities, proj1$darkgrnvegQ2, FUN= sd)
t.test(proj1$nbK3paFacilities ~ proj1$darkgrnvegQ2, var.equal = TRUE)



## Ques C
mean(proj1$nbK3FavorFoodstore)
sd(proj1$nbK3FavorFoodstore)
tapply(proj1$nbK3FavorFoodstore, proj1$darkgrnvegQ3, FUN= mean)
tapply(proj1$nbK3FavorFoodstore, proj1$darkgrnvegQ3, FUN= sd)
var.test(proj1$nbK3FavorFoodstore ~ proj1$darkgrnvegQ3, conf.level = 0.95) # variance is equal
t.test(proj1$nbK3FavorFoodstore ~ proj1$darkgrnvegQ3, var.equal = TRUE)


## Ques D
mean(proj1$VitaminD2, na.rm = TRUE)
sd(proj1$VitaminD2, na.rm = TRUE)
mean(proj1$VitaminD3, na.rm = TRUE)
```

```
sd(proj1$VitaminD3, na.rm = TRUE)
t.test(proj1$VitaminD2, proj1$VitaminD3, paired = TRUE, conf.level = 0.95)
```