TEST 2 –PRACTICE

1. Data in R can be stored in **matrices** or **arrays** or other data structures. What are the two main data structures that we have used in class?

   vectors + data frames

2. Suppose you are running code in R, and you receive this error: **Error: could not find function "%>%"** What is the problem, and how can you fix it?

   library(dplyr) or library(tidyverse)

3. Consider these lines of code in R:

   ```
   round(53.24)   log(53.24)   read.csv(file.choose(), na.strings='')
   ```

   a. What is the technical term for **round, log**, and **read.csv**? function

   b. What is the technical term for the value **53.24** in round and log, and the **na.strings=''** in read.csv? arguments

   c. ggplot2 and dplyr are called packages

   d. BabyNames, NCHS and S are called data frames

4. Log transformations:

   a. Why/When do we use log transformations with data? When data is very skewed right.

   b. What is the value of $\log_{10}(100000)$? 5

   $10^{\square} = 100000$

5. What is the 'data-ink ratio' that Tufte discusses? (Define the measure)

   $$\frac{\text{Amount of ink devoted to data}}{\text{Total amount of ink}}$$

   (maximize it!)

6. Tufte has one opinion on 'chartjunk,' while the blog post you read describes another. Briefly describe each opinion:

   a. Tufte: Totally against

   b. Blog: Can make graph more memorable + attention-getting

7. The NCHS data frame (attached) is in your environment, and it is called **N**. (The individual vectors are not available by themselves.) It has over 30,000 lines; only a few are shown below.

   a. For the scatterplot on the last page, identify each variable and aesthetic. Then, **write the ggplot code** for making the graph.

Variable	Aesthetic
height	x
weight	y
smoker	shape
sex	color

```
ggplot(data = N,
  mapping = aes(x = height,
    y = weight, shape = smoker,
    color = sex)) +
  geom_point()
```

   b. Write code that would produce the mean of all of the ptfat values (percentage fat). You may use dplyr OR base package code.

```
N %>% summarise(mptf = mean(ptfat, na.rm = TRUE))
```

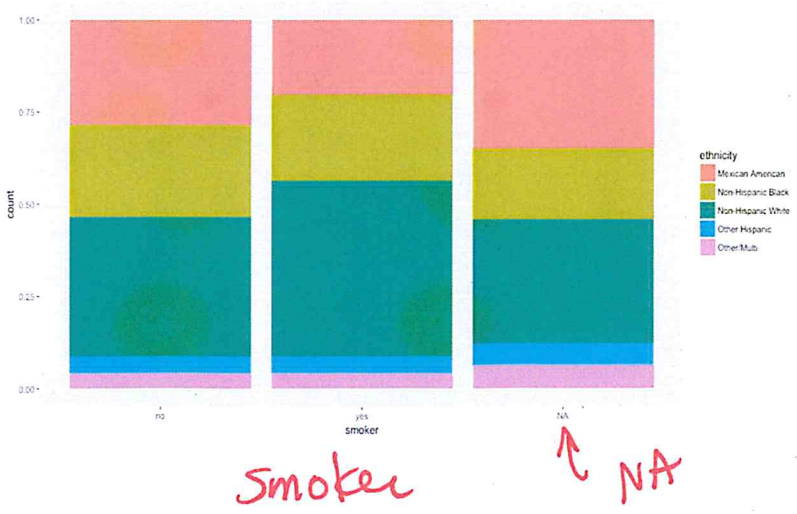
   c. Write **dplyr** code that will calculate the mean ptfat value for each different ethnicity. Use only 3 or so lines of code – don't create many data frames. (Note: there are no missing values for ethnicity)

```
N %>% group_by(ethnicity) %>%
  summarise(mptf = mean(ptfat, na.rm = TRUE))
```

   d. Write **dplyr** code to create a new data frame, called **NFem** that includes only females in the data frame.

```
NFem <- N %>% filter(sex == 'female')
```

   e. Write ggplot code that would use data frame **N** to produce **this** bar graph




```
ggplot(data = N,
  mapping = aes(x = smoker,
    fill = ethnicity)) +
  geom_bar(position = 'fill')
```

Smoker

NA

f. Suppose you wanted to create the above graph <u>without the third bar</u> (without the bar that has smoker equal to NA). Show how you could make a data frame, called **N2**, that could be used to make the graph above. (You do not need to make the graph – only create the data frame).

N2 ← N %>% filter(!is.na(smoker))

Use this data for question 7

	sex	age	pregnant	ethnicity	death	followup	smoker	diabetic	height	weight	waist	wci	bmi	ptfat
1	female	2	no	Non-Hispanic Black	NA	NA	no	0	0.916	12.50	0.457	0.07886587	14.89769	NA
2	male	77	no	Non-Hispanic White	alive	90	no	0	1.740	75.40	0.980	0.08711699	24.90421	14.338594
3	female	10	no	Non-Hispanic White	NA	NA	no	0	1.366	32.90	0.647	0.08171766	17.63171	NA
4	male	1	no	Non-Hispanic Black	NA	NA	no	0	NA	13.30	NA	NA	NA	NA
5	male	49	no	Non-Hispanic White	alive	74	yes	0	1.783	92.50	0.999	0.07908555	29.09639	16.450919
6	female	19	no	Other/Multi	alive	86	no	0	1.620	59.20	0.816	0.08030419	22.55754	19.648649
7	female	59	no	Non-Hispanic Black	alive	76	no	0	1.629	78.00	0.907	0.07461253	29.39358	17.339487
8	male	13	no	Non-Hispanic White	NA	NA	no	0	1.620	40.70	0.641	0.08098245	15.50831	6.668305
9	female	11	no	Non-Hispanic Black	NA	NA	no	0	1.569	45.50	0.646	0.07377525	18.48270	NA
10	male	43	no	Non-Hispanic Black	alive	79	no	0	1.901	111.80	1.080	0.07948423	30.93696	13.867352
11	male	15	no	Non-Hispanic White	NA	NA	no	0	1.719	65.00	0.765	0.07432172	21.99691	7.104769
12	male	37	no	Non-Hispanic White	alive	82	no	0	1.800	99.20	1.128	0.08590697	30.61728	15.345766
13	male	70	no	Mexican American	cardiovascular death	16	no	1	1.577	63.60	NA	NA	25.57371	23.646069
14	male	81	no	Non-Hispanic White	alive	85	yes	0	1.662	75.50	1.003	0.08574237	27.33285	16.626622
15	female	38	no	Non-Hispanic White	alive	92	yes	0	1.749	81.60	0.867	0.07343174	26.67538	16.575221
16	female	85	no	Non-Hispanic Black	other	62	no	0	1.442	41.50	0.744	0.08420643	19.95803	10.981205
17	male	2	no	Non-Hispanic Black	NA	NA	no	0	0.886	11.40	0.445	0.07942396	14.52237	NA
18	female	1	no	Non-Hispanic White	NA	NA	no	0	NA	11.10	NA	NA	NA	NA

Use this graph for question 7a: