Assignment 6

Kyle Yeo

Due Friday, October 9, 11:59 PM

Problems

1

Transform and combine the necessary data sets so that you have two rows for each zip code (one row for each sex) and the columns of data listed below. Note that you will need to eliminate the data on obesity among children, and summarize the data across age cohorts within each zip code to accomplish this task. Display the first six rows of the transformed and combined data frame using the function `head()`.

- `zip` = zip code
- `sex` = sex (male or female)
- `adult_n` = estimated # of adults (of that sex)
- `obese_n` = estimated # of obese adults (of that sex)
- `obese_p` = estimated proportion of obese adults (of that sex)
- `pct_bach` = % adults (aged 25+, of the given sex) with at least a bachelors degree

```
obesity <- read_csv("C:/stat_240/data/obesity-hw.csv")
education <- read_csv("C:/stat_240/data/education.csv") %>%
  rename(male = pct_m_bach, female = pct_f_bach) %>%
  pivot_longer(c("female", "male"), names_to = "sex", values_to = "pct_bach")
#education

obesity1 <- obesity %>%
  filter(age != "05-17") %>%
  mutate(adult_n = pop) %>%
  mutate(obese_n = adult_n * (obese/bmi)) %>%
  select(-pop, zip, sex, adult_n, obese_n) %>%
  drop_na() %>%
  group_by(zip, sex) %>%
  summarise(adult_n = sum(adult_n), obese_n = sum(obese_n)) %>%
  mutate(obese_p = obese_n/adult_n)



final_result <- left_join(education, obesity1, by = c("zip", "sex")) %>%
  select(zip, sex, adult_n, obese_n, obese_p, pct_bach) %>%
  drop_na()


head(final_result)
```

```
## # A tibble: 6 x 6
##     zip sex    adult_n obese_n obese_p pct_bach
```

```
##   <dbl> <chr>    <dbl>   <dbl>   <dbl>    <dbl>
## 1 53001 female     671    259.   0.386     23
## 2 53001 male       682    330.   0.483     13
## 3 53002 female     977    369.   0.377     25.4
## 4 53002 male      1052    404.   0.384     16.2
## 5 53004 female    1244    553.   0.445     26.8
## 6 53004 male      1164    604.   0.519     23.3
```

2

Using the data from Question 1, we are going to investigate connections between obesity and education status (at least a bachelors degree or no bachelors degree) by sex. For this question, calculate the *estimated percentage of adults in Wisconsin who are obese* among those with at least a bachelors degree by sex. Similarly, calculate the *estimated percentage of adults in Wisconsin who are obese* among those without a bachelors degree by sex.
Display these values in a 2-by-2 table, i.e., a table with two rows - one for male and one for female, and two columns - one for each of the estimated percentages noted above (plus the first column sex). State any assumptions you need to make when carrying out these calculations. (Recall that you need to sum up totals of people before finding proportions.)

```
obe_edu <- final_result %>%
  mutate(bach = adult_n * (pct_bach/100), non_bach = adult_n - bach) %>%
  group_by(sex) %>%
  summarise(bach = sum(bach), non_bach = sum(non_bach), total_n = sum(adult_n)) %>%
  mutate(epao_with_bach = (bach/total_n)*100, epao_without_bach = (non_bach/total_n)*100) %>%
  select(sex, epao_with_bach, epao_without_bach)

obe_edu
```

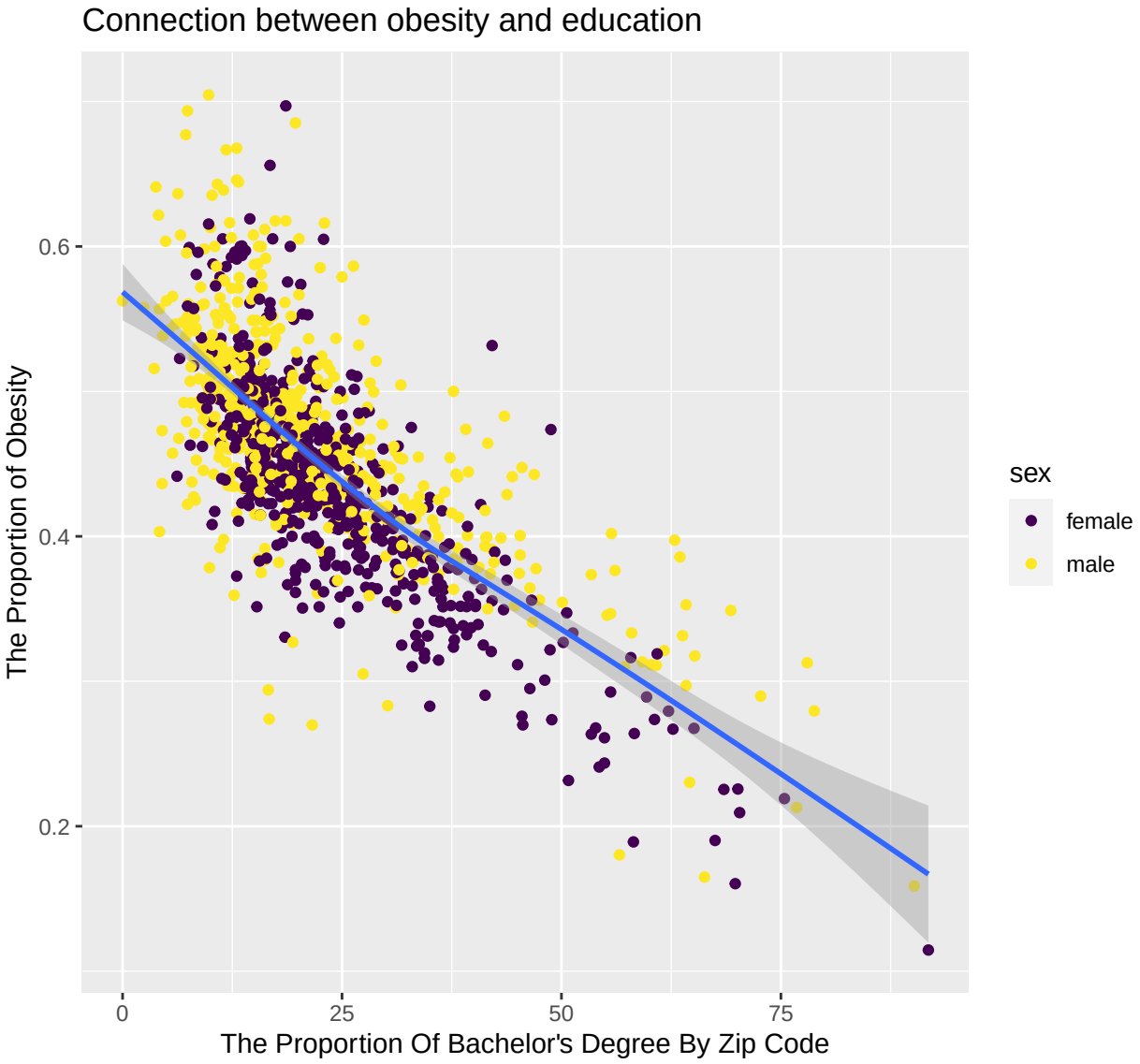
```
## # A tibble: 2 x 3
##   sex    epao_with_bach epao_without_bach
##   <chr>           <dbl>             <dbl>
## 1 female           30.9              69.1
## 2 male             29.0              71.0
```

```
#obe_edu
```

3

Make a scatter plot that displays the proportion of a zip code aged 25+ with a bachelor's degree on the x-axis and the proportion obese on the y axis. Use different colors for each sex and add a trend line or curve for each sex. Create appropriate labels and titles for the plot. Comment on any apparent patterns in the data.

```
ggplot(final_result, mapping = aes(x = pct_bach, y = obese_p)) +
  geom_point(aes(color = sex)) +
  geom_smooth() +
  xlab("The Proportion Of Bachelor's Degree By Zip Code") +
  ylab("The Proportion of Obesity") +
  ggtitle("Connection between obesity and education")
```



According to the above plot, we can figure out the pattern that people with a bachelor's degree tend to have less percent of obesity.

4

Transform and combine the necessary data sets so that you have one row for each zip code and the following columns of data. Note that you will need to eliminate the data on obesity among children and summarize the obesity data across age and sex cohorts within each zip code to accomplish this task. Display the first six rows of the transformed and combined data frame using the function `head()`.

- `zip` = zip code
- `adult_n` = estimated # of adults
- `obese_n` = estimated # of obese adults
- `non_obese_n` = estimated # of non-obese adults
- `obese_p` = estimated proportion of obese adults
- `households` = # of households
- `income` = median household income

- rural_n = # of residents in rural areas
- urban_n = # of residents in urban areas

```
obesity4 <- read_csv("C:/stat_240/data/obesity-hw.csv") %>%
  filter(age != "05-17") %>%
  drop_na() %>%
  mutate(adult_n = pop) %>%
  mutate(obese_n = adult_n * (obese/bmi)) %>%
  group_by(zip) %>%
  summarise(adult_n = sum(adult_n), obese_n = sum(obese_n)) %>%
  mutate(obese_p = obese_n/adult_n, non_obese_n = adult_n - obese_n)


income4 <- read_csv("C:/stat_240/data/income.csv") %>%
  drop_na()

obe_income <- left_join(obesity4, income4, by = "zip") %>%
  select(zip, adult_n, obese_n, non_obese_n, obese_p, households, income)
#obe_income

rural_ur <- read_csv("C:/stat_240/data/rural-urban.csv") %>%
  drop_na()

final_dataset <- left_join(obe_income, rural_ur, by = "zip") %>%
  mutate(urban_n = adult_n * (p_urban)) %>%
  mutate(rural_n = adult_n * (1-p_urban)) %>%
  select(-population, -p_urban, -rural, -urban)
#final_dataset
head(final_dataset)
```

```
## # A tibble: 6 x 9
##     zip adult_n obese_n non_obese_n obese_p households income urban_n rural_n
##   <dbl>   <dbl>   <dbl>       <dbl>   <dbl>      <dbl>  <dbl>   <dbl>   <dbl>
## 1 53001    1353    589.        764.   0.435        788  72206       0    1353
## 2 53002    2029    772.       1257.   0.381        869  85478       0    2029
## 3 53004    2408   1157.       1251.   0.481       1217  77989       0    2408
## 4 53005   15189   4632.      10557.   0.305       7556  97202   15189       0
## 5 53006     458    167.        291.   0.366        670  74107       0     458
## 6 53007    1437    607.        830.   0.422        895  41925    1437       0
```

5

Using the previous question's data frame, create a new variable `ru` that takes the value `rural` if 50% or more of the residents in the zip code live in rural areas, otherwise assign the value `urban`. Assume each adult in a zipcode has the median household income from that zip code. Under this assumption, calculate and display the average income for obese and non-obese adults for the state by `ru`. Your answer should have two rows and two columns.

```
final_dataset %>%
  mutate(ru = ifelse(rural_n > urban_n, "rural", "urban")) %>%
  group_by(ru) %>%
  summarise(obesity = weighted.mean(income, w=obese_n,na.rm=TRUE), non_obesity = weighted.mean(income,
```

```
## # A tibble: 2 x 3
##   ru    obesity non_obesity
##   <chr>   <dbl>       <dbl>
```

```
## 1 rural  56655.      57585.
## 2 urban  58473.      60960.
```

6

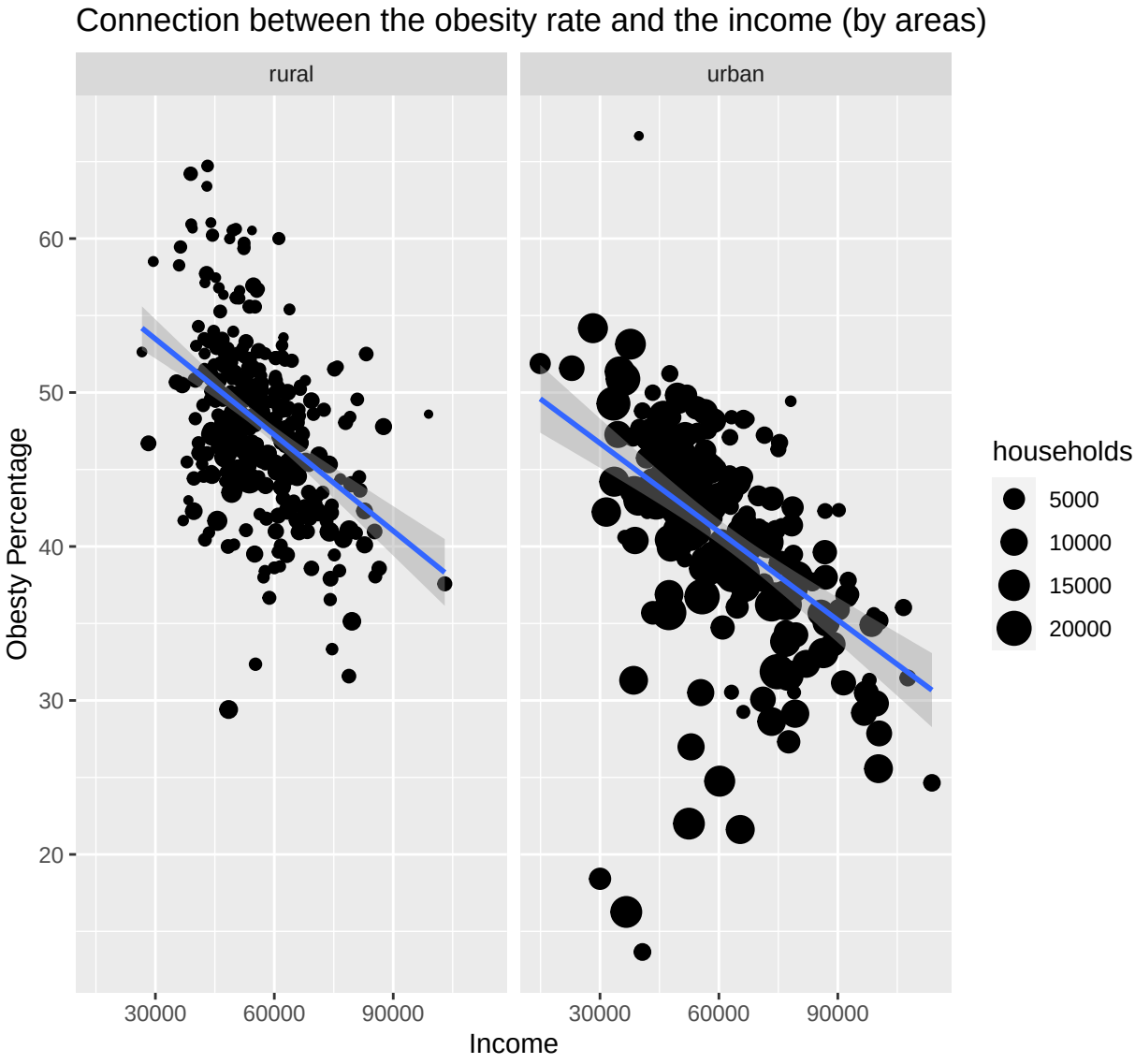
Make a scatter plot with one point for each zip code with the median household income on the x-axis and the percentage of obese adults on the y-axis. Make the area of the points proportional to the number of households represented (check out the `size` aesthetic). Create appropriate labels and titles for the plot, and facet by `ru`. Add a trend line/curve and comment on any apparent patterns.

```
data6 <- final_dataset %>%
  mutate(ru = ifelse(rural_n > urban_n, "rural", "urban"))
#data6

ggplot(data6, aes(x=income, y = obese_p*100)) +
  geom_point(aes(size = households)) +
  geom_smooth(method = "lm") +
  facet_wrap(~ru) +
  xlab("Income") +
  ylab("Obesty Percentage") +
  ggtitle("Connection between the obesity rate and the income (by areas)")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



People with more income tend to have less obesity rate and urban areas show us the pattern that has larger households and less obesity rate.

7

Transform and combine the necessary data sets so that you have four rows for each zip code (one row for the four age groups defined next) and the columns of data listed below. Define new age categories as "05-17", "18-34", "35-74", and "75-plus". Note that you will need to summarize the data across sex cohorts within each zip code to accomplish this task. Display the first six rows of the transformed and combined data frame using the function `head()`.

- `zip` = zip code
- `age_group` = "05-17", "18-34", "35-74", or "75-plus"
- `pop_n` = estimated # of individuals
- `obese_n` = estimated # of obese individuals
- `obese_p` = estimated proportion of obese individuals
- `rural_n` = estimated # of individuals who live in a rural household
- `urban_n` = estimated # of individuals who live in an urban household

```
obesity7 <- read_csv("C:/stat_240/data/obesity-hw.csv") %>%
  drop_na() %>%
  mutate(age_group = case_when(age == "05-17" ~ "05-17",
                           age == "18-34" ~ "18-34",
                           age == "35-54" ~ "35-74",
                           age == "55-74" ~ "35-74",
                           age == "75-plus" ~ "75-plus")) %>%
  rename(pop_n = pop) %>%
  mutate(obese_n = pop_n * (obese/bmi)) %>%
  group_by(zip, age_group) %>%
  summarise(pop_n = sum(pop_n), obese_n = sum(obese_n), obese_p = sum(obese_n/pop_n))
#obesity7

rural_ur1 <- read_csv("C:/stat_240/data/rural-urban.csv") %>%
  drop_na()
#read rural_urban.csv file and drop all the missing values


final_dataset7 <- left_join(obesity7, rural_ur1, by = "zip") %>%
  mutate(urban_n = pop_n * (p_urban)) %>%
  mutate(rural_n = pop_n * (1-p_urban)) %>%
  select(-urban, -rural, -population, -p_urban)
#final_dataset7

head(final_dataset7)
```

```
## # A tibble: 6 x 7
## # Groups:   zip [2]
##     zip age_group pop_n obese_n obese_p urban_n rural_n
##   <dbl> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 53001 18-34       304    79.2   0.260       0     304
## 2 53001 35-74      1049   509.    0.486       0    1049
## 3 53002 05-17       383    56.9   0.149       0     383
## 4 53002 18-34       565   142.    0.251       0     565
## 5 53002 35-74      1321   583.    0.441       0    1321
## 6 53002 75-plus     143    47.5   0.332       0     143
```

8

Using the previous question's data frame, calculate estimated percentages of obese individuals by age group and if they live in an urban or rural household. Display these values in a 4 by 2 table with one row for each age group range and separate columns for rural and urban.

```
epio_data <- final_dataset7 %>%
  mutate(ru = ifelse(rural_n > urban_n, "rural", "urban")) %>%
  group_by(age_group, ru) %>%
  summarise(epio = sum(obese_n) / sum(pop_n)*100)

epio_data %>%
  pivot_wider(names_from = c(ru), values_from = c(epio))
```

```
## # A tibble: 4 x 3
## # Groups:   age_group [4]
##   age_group rural urban
##   <chr>     <dbl> <dbl>
```

```
## 1 05-17      19.0  15.8
## 2 18-34      34.0  30.0
## 3 35-74      51.0  45.5
## 4 75-plus    38.6  31.6
```

9

Create a scatter plot with a point for each zip code and age_group to show percentage urban on the x-axis and percentage obese on the y-axis. Assign the color by age_group. Create appropriate labels and titles for the plot. Comment on any patterns in the plot.

```
obesity9 <- read_csv("C:/stat_240/data/obesity-hw.csv") %>%
  drop_na() %>%
  mutate(age_group = case_when(age == "05-17" ~ "05-17",
                           age == "18-34" ~ "18-34",
                           age == "35-54" ~ "35-74",
                           age == "55-74" ~ "35-74",
                           age == "75-plus" ~ "75-plus")) %>%
  rename(pop_n = pop) %>%
  mutate(obese_n = pop_n * (obese/bmi)) %>%
  group_by(zip, age_group) %>%
  summarise(pop_n = sum(pop_n), obese_n = sum(obese_n), obese_p = sum(obese_n/pop_n))
#obesity7

rural_ur2 <- read_csv("C:/stat_240/data/rural-urban.csv") %>%
  drop_na()
#read rural_urban.csv file and drop all the missing values


final_dataset9 <- left_join(obesity9, rural_ur2, by = "zip") %>%
  select(-urban, -rural, -population)
#we need p_urban variable for this question so that I did not remove it.


final_dataset9
```

```
## # A tibble: 1,808 x 6
## # Groups:   zip [581]
##      zip age_group pop_n obese_n obese_p p_urban
##    <dbl> <chr>     <dbl>   <dbl>   <dbl>   <dbl>
##  1 53001 18-34       304    79.2   0.260       0
##  2 53001 35-74      1049   509.    0.486       0
##  3 53002 05-17       383    56.9   0.149       0
##  4 53002 18-34       565   142.    0.251       0
##  5 53002 35-74      1321   583.    0.441       0
##  6 53002 75-plus     143    47.5   0.332       0
##  7 53004 18-34       671   240.    0.358       0
##  8 53004 35-74      1645   877.    0.533       0
##  9 53004 75-plus      92    39.8   0.432       0
## 10 53005 18-34      2755   614.    0.223       1
## # ... with 1,798 more rows
```

```
ggplot(final_dataset9, aes(x = p_urban*100, y = obese_p*100)) +
  geom_point(aes(color = age_group)) +
  xlab("Percentage Of Urban") +
```

```
  ylab("Percentage Of Obesity") +
  ggtitle("The connection between urban area and the obesity rate")
```