Data Wrangling and Visualization Homework

QBS 103: Foundations of Data Science

Due: August 2, 2024

Assignment Overview

Unless otherwise specified, you may use any combination of base R, *reshape2*, and/or *tidyverse* to complete the data wrangling portion of this assignment. For the data visualization portion, all plots must be generated using *ggplot2* using the basic guidelines laid out in class for high quality figures (note: you must define a theme for all plots). Any figures created using *ggpubr* or other packages will receive at most half credit as the primary goal here is to get you comfortable working in *ggplot2*. The final assignment should be submitted as both an R markdown file and a knitted PDF file.

We're going to create a toy dataset of 10,000 pregnant women at high risk of gestational diabetes. Gestational diabetes (diabetes diagnosed during pregnancy) is diagnosed using an oral glucose tolerance test in which a fasting glucose is measured, a sugary beverage is consumed, and glucose is measured again an hour later. Here, we're going to create data in which carrying a male baby puts you at higher risk of being diagnosed with gestational diabetes.

1. Generate and Summarize Random Data (15 pts)

Generate a random data set (**remember to set a random seed**) of 10,000 pregnant women with the following characteristics:

1. Age is uniformly distributed between 18 and 35 years old (Variable Name: Age).
2. There is a probability of 0.5 that each mother is carrying a female infant (Variable Name: InfantSex). This variable should be formatted as a factor variable with levels "Female" and "Male".
3. Define fasting glucose measures (Variable Name: Glucose1) as normally distributed. Mothers carrying a male infant have a mean score of 85 and a standard deviation of 6 mg/dL. Mothers carrying a female infant have a mean score of 80 and a standard deviation of 6 mg/dL.
4. Define 1 hour glucose measures (Variable Name: Glucose2) as normally distributed. Mothers carrying a male infant have a mean score of 165 and a standard deviation of 9 mg/dL. Mothers carrying a female infant have a mean score of 155 and a standard deviation of 9 mg/dL.
5. Define a summary variable for gestational diabetes (Variable Name: Diagnosis) which is "Gestational Diabetes" if either Glucose1 is higher than 95 or Glucose2 is higher than 180 and "Healthy" otherwise.

Use the *summary()* function to **print summaries of your data** for male infants and female infants separately (hint: you will need to subset your data to do this) diagnosis should appear summarized as a factor variable. Do not print the entire data frame.

Generate a boxplot showing the distribution of Glucose1 (Fasting Glucose) on the y axis, Diagnosis on the x axis, and color the plot by InfantSex. Define a unique color palette for your boxplot with the colors of your choosing.

2. Plot The Distribution of Glucose Measures by Timepoint and Diagnosis (10 pts)

Convert your dataset into a **long format data frame** with a single Glucose variable and a separate variable designating Timepoint (Fasting or One Hour). **Print the entries for Subject 1**. Do not print the entire data frame.

Generate a boxplot of the distribution of Glucose (y axis) for all subjects where Timepoint is on the x-axis and the plot is colored by Diagnosis. Set a color palette that is unique from the color scheme used in Part 1. Define axis titles, x axis labels, and legend title using *ggplot2* (i.e. do not simply change the values in your data frame).

x axis title: "Timepoint"

x axis labels: "Baseline" and "One Hour"

y axis title: "Glucose (mg/dL)"

legend title: "GDM Diagnosis"

3. Plot the Distribution of Glucose Values by Maternal Age (10 pts)

Generate a scatter plot with maternal age plotted on the x axis (labeled "Maternal Age (yrs)") and glucose on the y axis (labeled "Glucose (mg/dL)". Color points by Time point (Baseline or One Hour) and define a new color palette than those used in parts 1 and 2. Separate into two plots by infant sex. For each plot, add text with the mean (standard deviation) fasting and 1 hr glucose in the same colors used for the color palette. Define the location of the text such that it is legible does not overlap with any points (note: this should be plotted as an annotation, not a plot title). Figures should have titles identifying "Mothers of Female Infants" and "Mothers of Male Infants" respectively. Finally, use the *ggarrange()* function in *ggpubr* or a similar function for arranging multiple plots of your choosing to produce a single figure with both plots adjacent, labeled as figures A and B, respectively.

4. Generate Wide Table of Summary Statistics (5 pts)

Generate a wide format table summarizing the age, fasting glucose, and one hour glucose of all subjects by both disease status and infant sex. Your table should have 4 rows in the following order: Healthy & Female, Gestational Diabetes & Female, Healthy & Male, Gestational Diabetes & Male and should summarize mean age, mean and sd fasting glucose, and mean and sd one hour glucose.