Week 12: Activity

Mitchell Levy

Data Visualization

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to b
```

```
library(dplyr)
library(reshape2)
```

```

Attaching package: 'reshape2'

The following object is masked from 'package:tidyr':

    smiths
```

```
library(RColorBrewer)
```

1. You will need to clean the data set for your analysis by recoding, correcting variable types, etc.

I am just going to do number 1 as I go.

Load the file.

```
stroke <- read.csv("stroke_data.csv")
stroke
```

id	gender	age	hypertension	heart_disease	ever_married
<int>	<chr>	<dbl>	<int>	<int>	<chr>
9046	Male	67.00	0	1	Yes
51676	Female	61.00	0	0	Yes
31112	Male	80.00	0	1	Yes
60182	Female	49.00	0	0	Yes
1665	Female	79.00	1	0	Yes
56669	Male	81.00	0	0	Yes
53882	Male	74.00	1	1	Yes
10434	Female	69.00	0	0	No
27419	Female	59.00	0	0	Yes
60491	Female	78.00	0	0	Yes

1–10 of 5,110 rows | 1–6 of 12 columns Previous 1 2 3 4 5 6 511 Next

2. Add a column that shows the BMI category for each subject.

```
stroke |>
  mutate(bmiCat = case_when(
    (bmi < 18.5) ~ "Underweight",
    (bmi >= 18.5 & bmi <= 24.9) ~ "Normal",
    (bmi >= 25 & bmi < 30) ~ "Overweight",
    (bmi >= 30) ~ "Obese",
    TRUE ~ NA_character_)) -> stroke
stroke
```

id	gender	age	hypertension	heart_disease	ever_married
<int>	<chr>	<dbl>	<int>	<int>	<chr>
9046	Male	67.00	0	1	Yes
51676	Female	61.00	0	0	Yes
31112	Male	80.00	0	1	Yes
60182	Female	49.00	0	0	Yes
1665	Female	79.00	1	0	Yes
56669	Male	81.00	0	0	Yes
53882	Male	74.00	1	1	Yes
10434	Female	69.00	0	0	No
27419	Female	59.00	0	0	Yes
60491	Female	78.00	0	0	Yes

1–10 of 5,110 rows | 1–6 of 13 columns Previous 1 2 3 4 5 6 511 Next

3. For one of the figures you must sample the data and show errorbars for the Body Mass Index OR the average glucose level for one of the categorical variables.

I am going to do the average glucose levels compared to if they were ever married.

Since we did not learn how to do this with multiple samples, I had to look up how to do this, so yeah this is not all my code, but I tried for 3 and a half hours and could not figure out how to do this.

```
# Define the number of samples and the number of repetitions
n_samples <- 50 # sample size
n_repeats <- 100 # number of repeated samples

# Perform the sampling multiple times
stroke_samples <- bind_rows( # bind rows just puts each repeated sample on top of each p
  lapply(1:n_repeats, function(i) { # this will run everything an n_repeats amount of ti
    stroke |>
      group_by(ever_married) |>
      slice_sample(n = n_samples) |> # randomly choose 25 glucose levels
      mutate(repeat_id = i) # Add a variable so we know which repeated sample we are on
  })
)
stroke_samples
```

id	gender	age	hypertension	heart_disease	ever_married
<int>	<chr>	<dbl>	<int>	<int>	<chr>
54117	Male	7.00	0	0	No
23604	Male	4.00	0	0	No
1737	Female	16.00	0	0	No
57979	Male	8.00	0	0	No
64128	Male	10.00	0	0	No
30734	Male	15.00	0	0	No
51579	Male	27.00	0	0	No
64393	Male	56.00	0	0	No
32826	Male	0.08	0	0	No
66229	Female	17.00	0	0	No

1–10 of 10,000 rows | 1–6 of 14 columns Previous 1 2 3 4 5 6 1000 Next

Back to my own code! Let's get the data now that we want to graph!

```
stroke_samples |>
group_by(repeat_id, ever_married) |>
summarise(glucose_bar = mean(avg_glucose_level),
          glucose_sd = sd(avg_glucose_level)) |>
mutate(glucose_se = glucose_sd/sqrt(2))-> stroke_means
```

```
`summarise()` has grouped output by 'repeat_id'. You can override using the
`.groups` argument.
```

```
stroke_means
```

repeat_id	ever_married	glucose_bar	glucose_sd	glucose_se
<int>	<chr>	<dbl>	<dbl>	<dbl>
1	No	90.0190	21.67890	4.375780
1	Yes	102.0892	44.12899	8.825799
2	No	99.2512	31.40205	6.280411
2	Yes	109.1780	51.71403	10.2238055
3	No	93.5834	27.15647	5.431294
3	Yes	121.4886	59.58903	11.917205
4	No	95.4584	35.69966	7.139933
4	Yes	96.3526	45.19565	9.039129
5	No	92.8948	24.55668	4.911336
5	Yes	126.9502	60.08815	12.017631

1–10 of 200 rows 1 2 3 4 5 6 20 Next

```
stroke_means |>
ggplot(aes(x = glucose_bar, y = repeat_id, color = ever_married)) + # Color mapped to
  alpha(70, 100)) + # Set x-axis limits
  geom_point(size = 3, position = position_jitter(width = 0, height = 0.3)) + # Jitter p
  geom_errorbar(aes(xmin = glucose_bar - 1.96 * glucose_se,
                    xmax = glucose_bar + 1.96 * glucose_se),
                width = 0.2) + # Add error bars with width for visibility
  labs(
    title = "Average Glucose Levels by Marriage Status", x = "Average Glucose Level", y
  scale_color_manual(values = c("Yes" = "#a1dab4ff", "No" = "#41b6c4ff")) +
  theme_minimal() + # Use a clean theme
  theme(
    legend.position = "right",
    axis.title = element_text(size = 20),
    axis.text = element_text(size = 20),
    title = element_text(size = 20),
    legend.text = element_text(size = 20)
  )
```

Comment on your findings...focus on uncertainty.

When looking at this graph, you can immediately see that people who have been married have higher average glucose levels. However, what is also interesting is the levels of uncertainty for each category. While unfortunately it can be hard to see the error bars for never married, it is clear that the error bars are very long for each category. This means that the data could really fall anywhere, and that being married or not married does not make you definitively higher or lower average glucose levels. It just means that married people in general tend to have higher levels.

One of your figures must show a scatterplot, some sort of model, and the uncertainty bands separated by one of the categorical variables.

```
stroke
```

id	gender	age	hypertension	heart_disease	ever_married
<int>	<chr>	<dbl>	<int>	<int>	<chr>
9046	Male	67.00	0	1	Yes
51676	Female	61.00	0	0	Yes
31112	Male	80.00	0	1	Yes
60182	Female	49.00	0	0	Yes
1665	Female	79.00	1	0	Yes
56669	Male	81.00	0	0	Yes
53882	Male	74.00	1	1	Yes
10434	Female	69.00	0	0	No
27419	Female	59.00	0	0	Yes
60491	Female	78.00	0	0	Yes

1–10 of 5,110 rows | 1–6 of 13 columns Previous 1 2 3 4 5 6 511 Next

I am going to look at glucose level, age, and gender.

```
stroke |>
  filter(!is.na(avg_glucose_level), !is.na(gender), !is.na(age)) |>
  ggplot(aes(x = age, y = avg_glucose_level, color = gender)) +
  geom_point(alpha = 0.15) +
  geom_smooth(
    method = "lm",
    formula = y ~ splines::bs(x, 6),
    se = TRUE,
    size = 0.5 # Increase size of the smoothed line
  ) +
  scale_color_manual(values = c("Male" = "deepskyblue", "Female" = "tomato3")) +
  ggtitle("How Age Relates to Glucose Levels, by Gender") +
  labs(x = "Age", y = "Average Glucose Level")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```


6. Comment on your findings, especially any uncertainty in the models and what they mean.

This graph shows the average glucose levels of male and females as they age. As you can see, the average levels slowly rise for both genders over time. However, the males are consistently just barely higher than the females. The standard error is very small, which means that there is a high likelyhood of that being where the true average is for this data.

7. Construct any other visualization and comment on what it shows.

```
stroke
```

id	gender	age	hypertension	heart_disease	ever_married
<int>	<chr>	<dbl>	<int>	<int>	<chr>
9046	Male	67.00	0	1	Yes
51676	Female	61.00	0	0	Yes
31112	Male	80.00	0	1	Yes
60182	Female	49.00	0	0	Yes
1665	Female	79.00	1	0	Yes
56669	Male	81.00	0	0	Yes
53882	Male	74.00	1	1	Yes
10434	Female	69.00	0	0	No
27419	Female	59.00	0	0	Yes
60491	Female	78.00	0	0	Yes

1–10 of 5,110 rows | 1–6 of 13 columns Previous 1 2 3 4 5 6 511 Next

```
# Convert bmi to numeric (if it's a character or factor type)
stroke <- stroke |>
  mutate(bmi = as.numeric(bmi))
```

```
Warning: There was 1 warning in `mutate()`.
ℹ In argument: `bmi = as.numeric(bmi)`.
Caused by warning:
! NAs introduced by coercion
```

```
# Reshape the data to long format
stroke_long <- stroke |>
  select(age, bmi, avg_glucose_level) |>
  pivot_longer(cols = c(bmi, avg_glucose_level), names_to = "variable", values_to = "val

# Create the plot with two lines (one for bmi and one for avg_glucose_level)
ggplot(stroke_long, aes(x = age, y = value, color = variable)) +
  geom_smooth(size = 1, se = FALSE) +  # Draw lines for each variable
  labs(title = "BMI and Average Glucose Level vs. Age",
       x = "Age",
       y = "Value") +                  # Shared y-axis label
  scale_color_manual(values = c("bmi" = "#e9a3c9", "avg_glucose_level" = "cyan")) + # Cu
  theme_minimal()
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

Warning: Removed 201 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

This graph shows a smoothed line for the average glucose level and bmi of the data set. As you can see, the as people age, their average glucose level appears to increase. Initially, it appears that the bmi will also increase with age, but around 40, bmi starts to remain consistent, and then decreases slightly. I made this graph because I was curious what the trends would be for these two variables over time. It is a little too simple, but I did not know what else to do.