

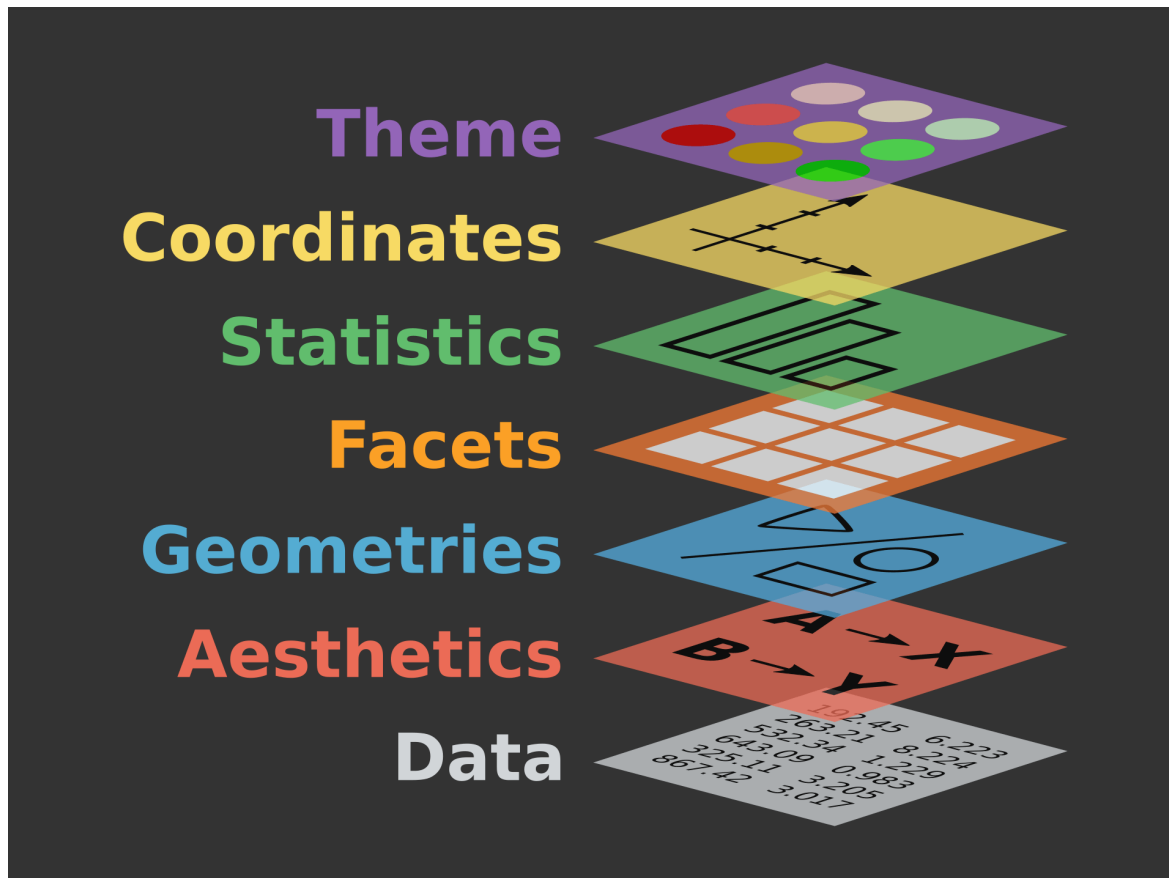
Data visualization

Part II

Joselyn Chávez

02/21/2023

Let's recap



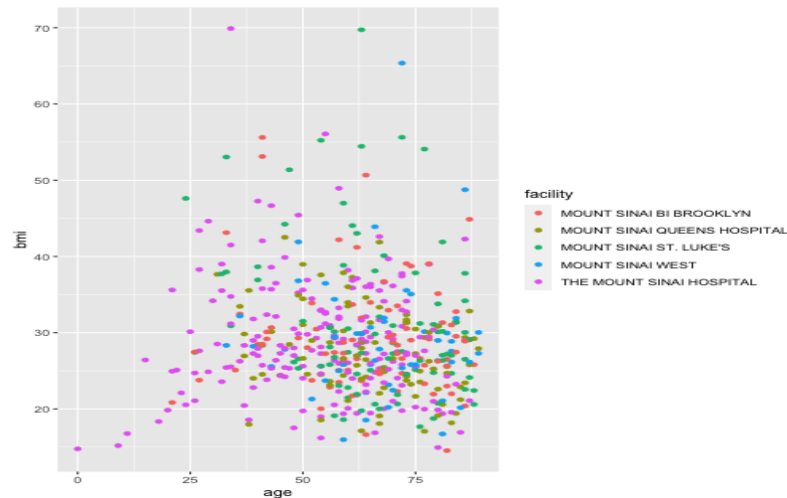
Let's recap

- What geometry would you use for plotting two numerical variables?
- What geometry would you use for plotting categorical vs continuous variables?
- How can you include a third variable in the plot?

Plotting num vs num vs cat

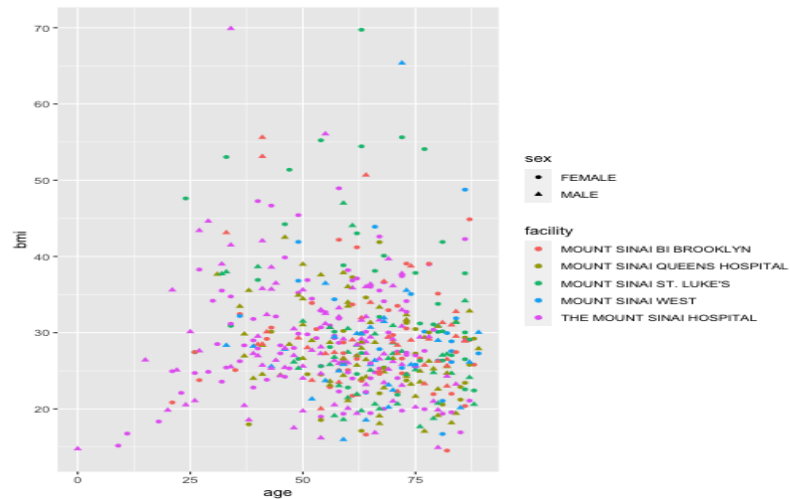
```
library(tidyverse)
sinai_covid <- read_csv("Sinai_covid.csv")

ggplot(sinai_covid,
       aes(x = age,
           y = bmi,
           color = facility)) +
  geom_point()
```



Plotting num vs num vs cat vs cat

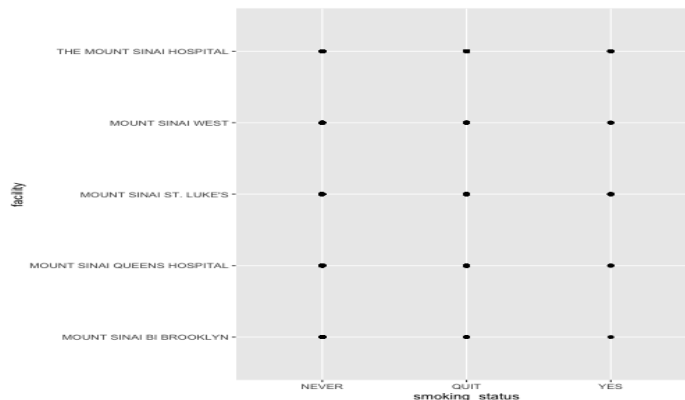
```
ggplot(sinai_covid, aes(x = age,  
                        y = bmi,  
                        color = facility,  
                        shape = sex)) +  
  
geom_point()
```



Plotting cat vs cat?

- How would you compare smoking_status vs facility?

```
ggplot(sinai_covid,  
      aes(x = smoking_status,  
          y = facility)) +  
  geom_point()
```



Plotting cat vs cat vs num

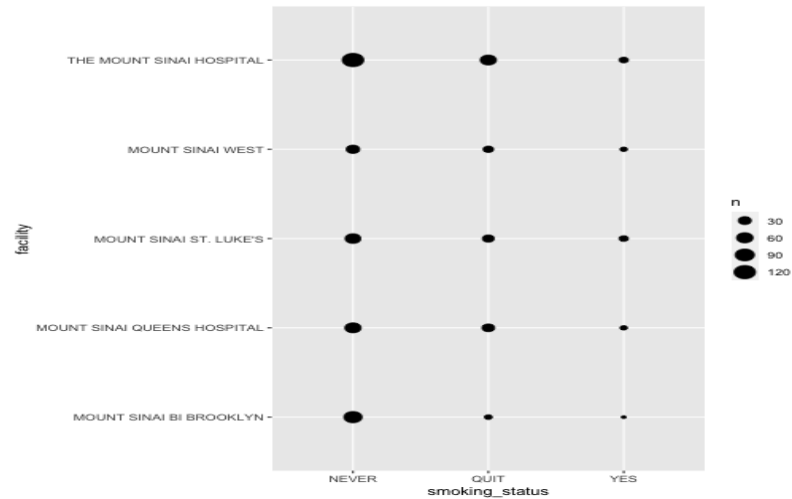
- First create the summarized tibble

```
count_smoking <- sinai_covid %>%  
  group_by(facility) %>%  
  count(smoking_status)  
  
count_smoking[1:3,]
```

```
## # A tibble: 3 × 3  
## # Groups:   facility [1]  
##   facility          smoking_status      n  
##   <chr>          <chr>          <int>  
## 1 MOUNT SINAI BI BROOKLYN NEVER      78  
## 2 MOUNT SINAI BI BROOKLYN QUIT       5  
## 3 MOUNT SINAI BI BROOKLYN YES        1
```

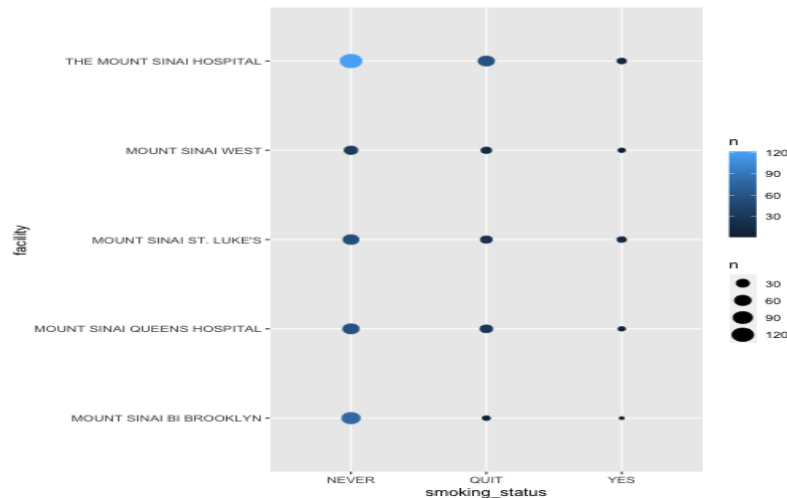
- Plot

```
ggplot(count_smoking,  
       aes(x = smoking_status,  
           y = facility,  
           size = n)) +  
  geom_point()
```



Plotting 3 variables (cat-cat-num)

```
ggplot(count_smoking,  
  aes(x = smoking_status,  
      y = facility,  
      size = n, color = n)) +  
  geom_point()
```

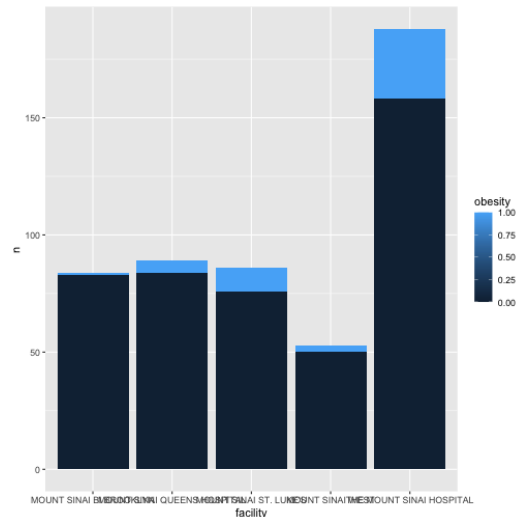


Let's practice

- Count the number of patients with obesity (0 and 1) per facility
- Plot the data mapping the size of the dots to the obesity counts

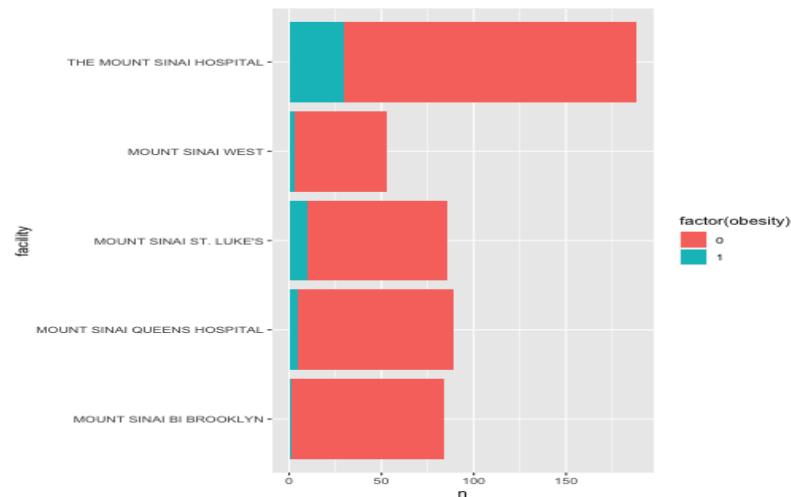
Another way

```
ob_count <- sinai_covid %>%  
  group_by(facility) %>%  
  count(obesity)  
  
ggplot(ob_count, aes(x = facility, y = n,  
                     fill = obesity)) +  
  geom_col()
```



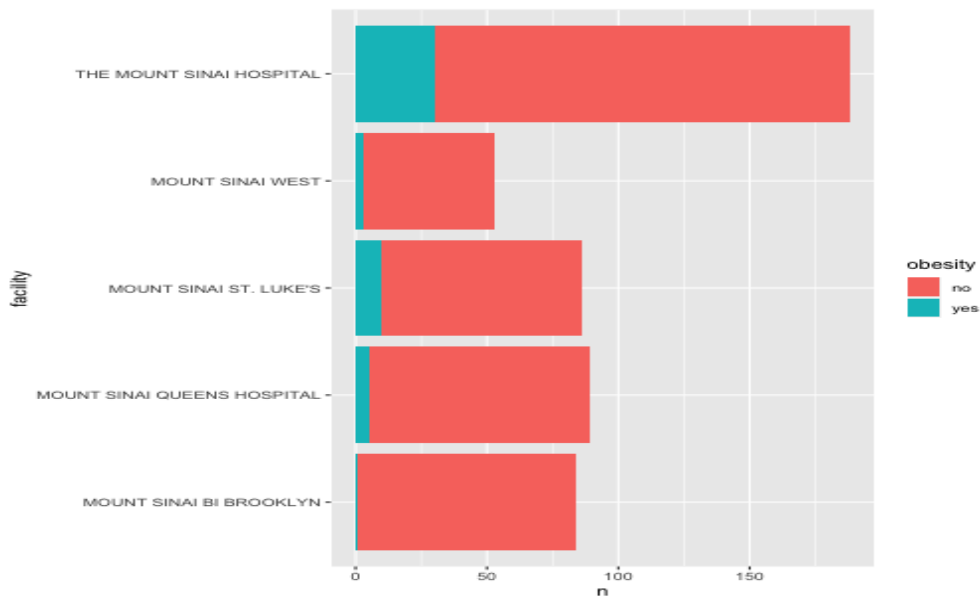
- Obesity must be a "categorical" variable
- Let's treat the variable as factor

```
ggplot(ob_count,  
       aes(x = n,  
           y = facility,  
           fill = factor(obesity))) +  
geom_col()
```



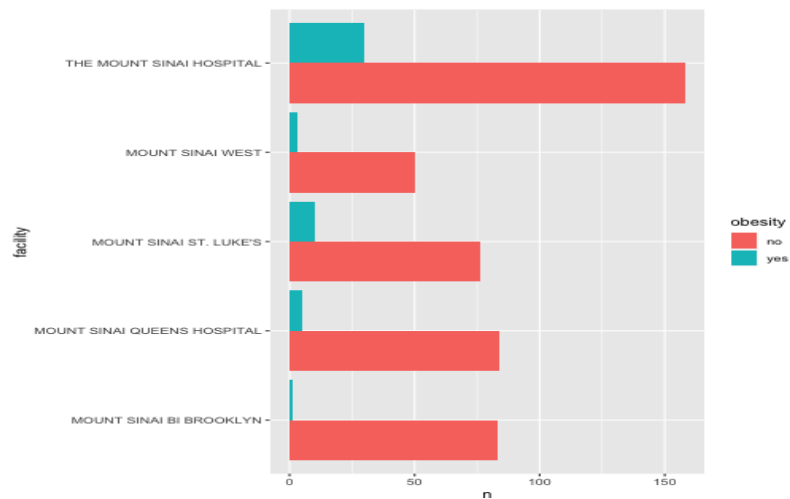
Or replace values

```
ob_count <- ob_count %>%  
  mutate(obesity = case_when(obesity == 0 ~ "no",  
                              obesity == 1 ~ "yes"))  
  
ggplot(ob_count, aes(x = n, y = facility,  
                     fill = obesity)) +  
  geom_col()
```



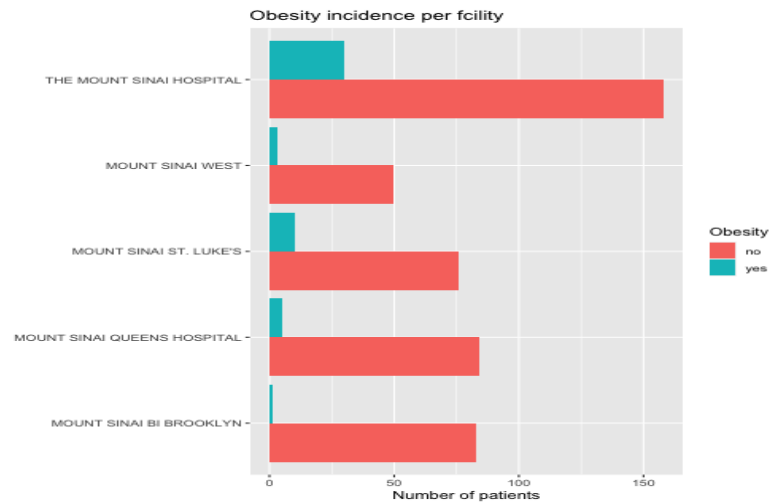
Splitting the bars

```
ggplot(ob_count,  
       aes(x = n,  
           y = facility,  
           fill = obesity)) +  
  geom_col(position = position_dodge())
```



Adding titles

```
ggplot(ob_count,  
       aes(x = n, y = facility,  
           fill = obesity)) +  
  geom_col(position = position_dodge()) +  
  labs(title = "Obesity incidence per fcility",  
       x = "Number of patients",  
       y = "",  
       fill = "Obesity")
```



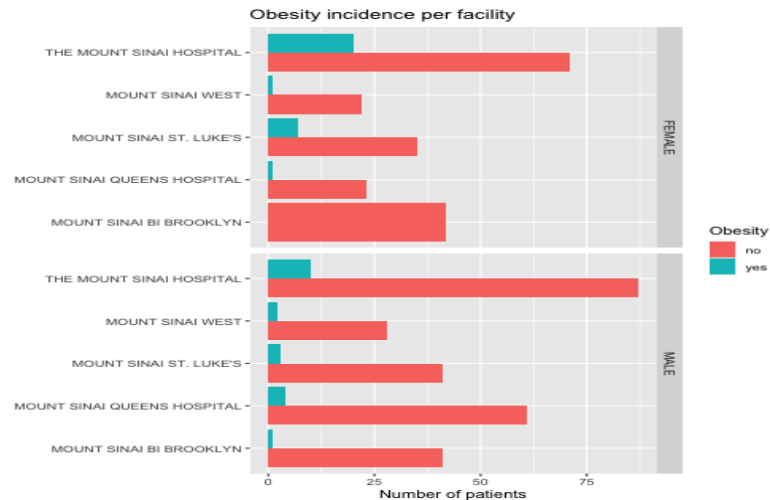
Facets

```
ob_count <- sinai_covid %>%  
  group_by(facility, sex) %>%  
  count(obesity) %>%  
  mutate(obesity = case_when(obesity == 0 ~ "no",  
                             obesity == 1 ~ "yes"))  
  
ob_count[1:3,]
```

```
## # A tibble: 3 × 4  
## # Groups:   facility, sex [2]  
##   facility          sex  obesity      n  
##   <chr>            <chr> <chr>    <int>  
## 1 MOUNT SINAI BI BROOKLYN FEMALE no      42  
## 2 MOUNT SINAI BI BROOKLYN MALE  no      41  
## 3 MOUNT SINAI BI BROOKLYN MALE  yes       1
```

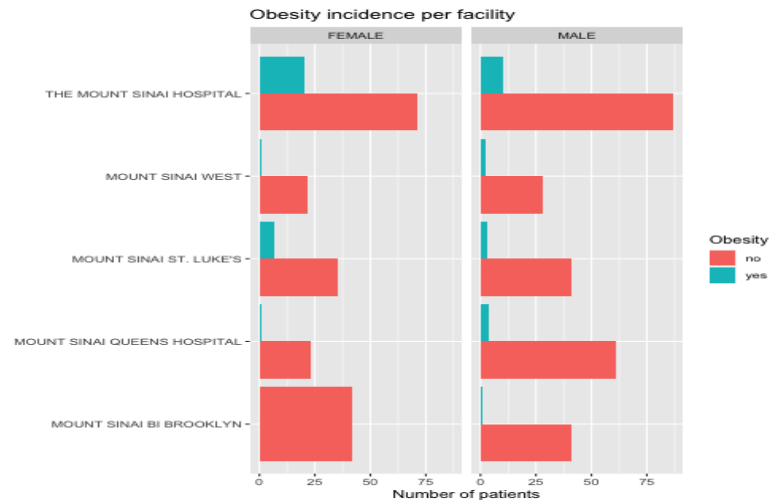

Facet grid

```
ggplot(ob_count,  
      aes(x = n, y = facility,  
          fill = obesity)) +  
geom_col(position = position_dodge()) +  
labs(title = "Obesity incidence per facility",  
      x = "Number of patients",  
      y = "",  
      fill = "Obesity") +  
facet_grid(rows = vars(sex))
```



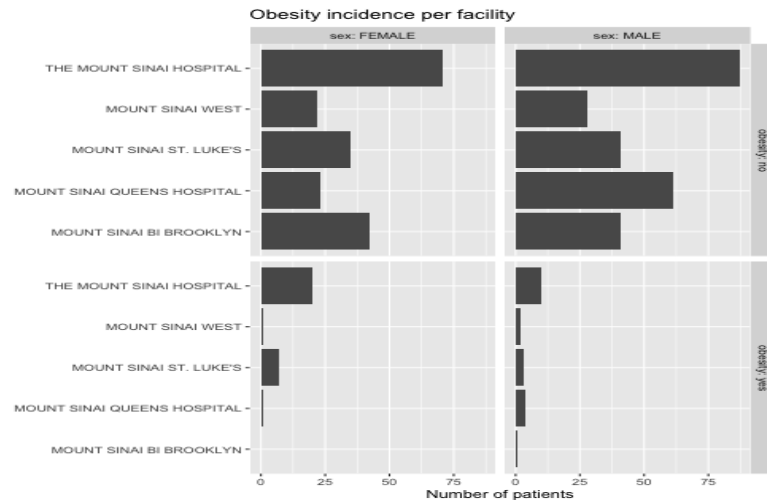
Facet grid

```
ggplot(ob_count,  
       aes(x = n, y = facility,  
           fill = obesity)) +  
  geom_col(position = position_dodge()) +  
  labs(title = "Obesity incidence per facility",  
       x = "Number of patients",  
       y = "",  
       fill = "Obesity") +  
  facet_grid(cols = vars(sex))
```



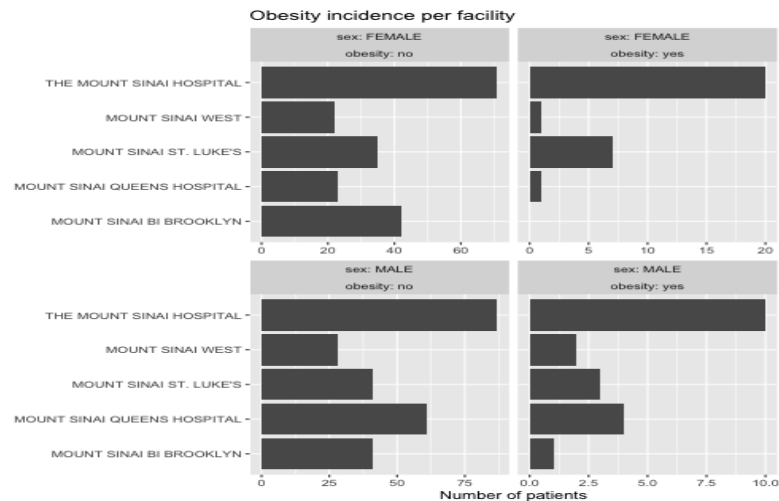
Facet grid

```
ggplot(ob_count,  
       aes(x = n, y = facility)) +  
  geom_col() +  
  labs(title = "Obesity incidence per facility",  
       x = "Number of patients",  
       y = "") +  
  facet_grid(rows = vars(obesity),  
            cols = vars(sex),  
            labeller = label_both)
```



Facet wrap

```
ggplot(ob_count,
       aes(x = n, y = facility)) +
  geom_col() +
  labs(title = "Obesity incidence per facility",
       x = "Number of patients",
       y = "") +
  facet_wrap(vars(sex, obesity), scales = "free_x",
            labeller = label_both)
```

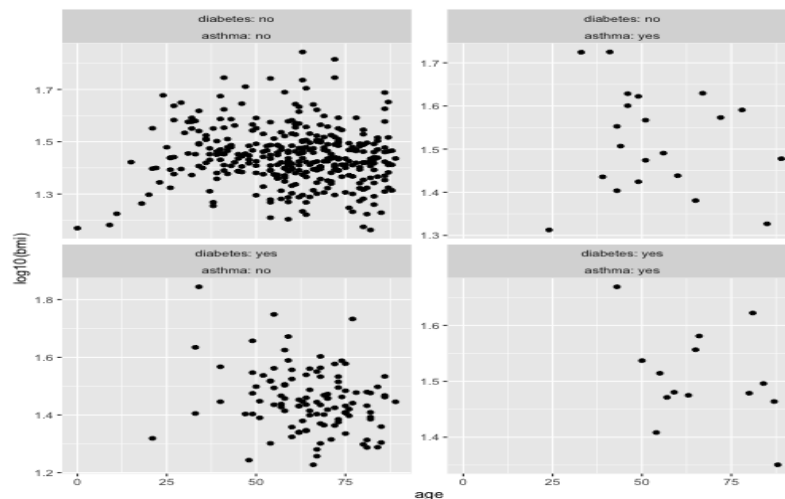


Let's practice!

- Count the number of patients depending on their smoking status per ethnic group and asthma status.
- Represent the data on a column plot, color them by asthma status
- Add a plot title, axis title and change the legend title to remove the underscores
- Split the plots in columns and rows by smoking and asthma status

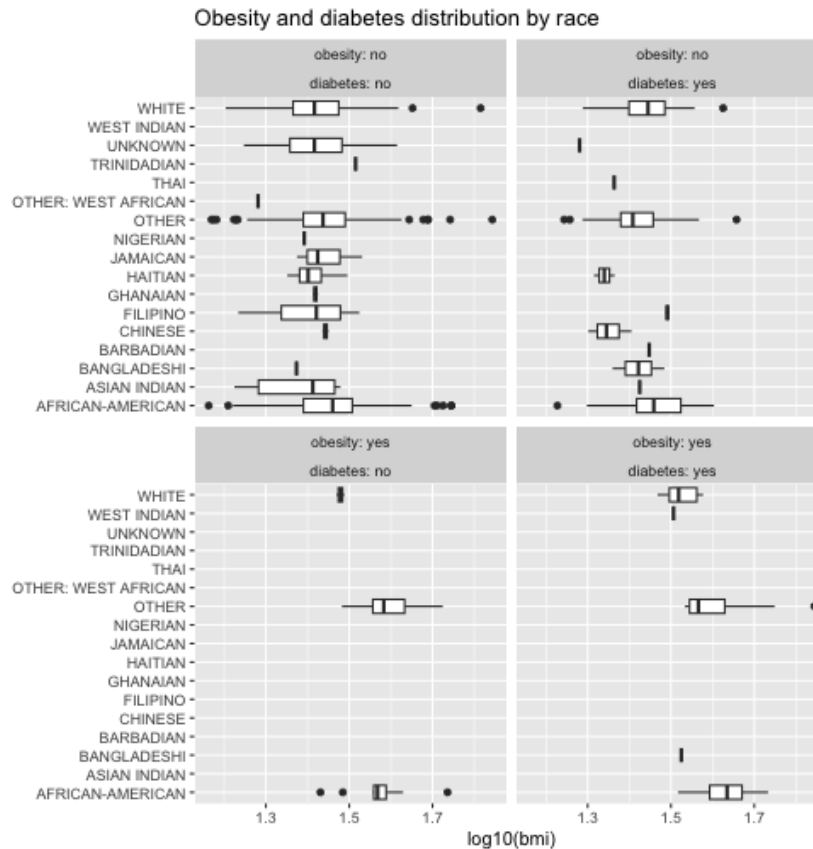
Statistical transformations

```
sinai_covid %>%  
  mutate(asthma = case_when(asthma == 0 ~ "no",  
                             asthma == 1 ~ "yes"),  
         diabetes = case_when(diabetes == 0 ~ "no",  
                              diabetes == 1 ~ "yes")) %>%  
  ggplot(aes(x = age, y = log10(bmi))) +  
  geom_point() +  
  facet_wrap(vars(diabetes, asthma), scales = "free_y",  
            labeller = label_both)
```



Let's practice!

- Write the code needed to get the following plot:

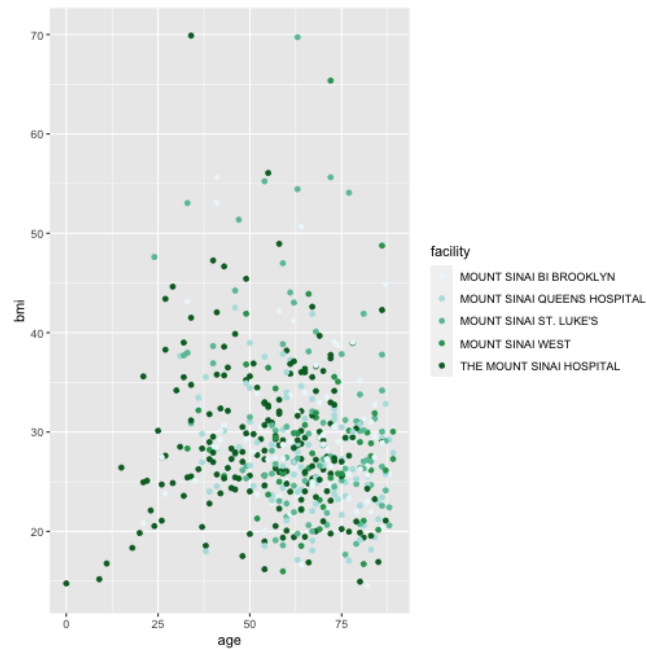


- Possible Answer

```
sinai_covid %>%  
  mutate(obesity = case_when(obesity == 0 ~ "no",  
                              obesity == 1 ~ "yes"),  
         diabetes = case_when(diabetes == 0 ~ "no",  
                              diabetes == 1 ~ "yes")) %>%  
  ggplot(aes(x = log10(bmi), y = race)) +  
  geom_boxplot() +  
  facet_wrap(vars(obesity, diabetes),  
            labeller = label_both) +  
  labs(title = "Obesity and diabetes distribution by race",  
       y = "")
```

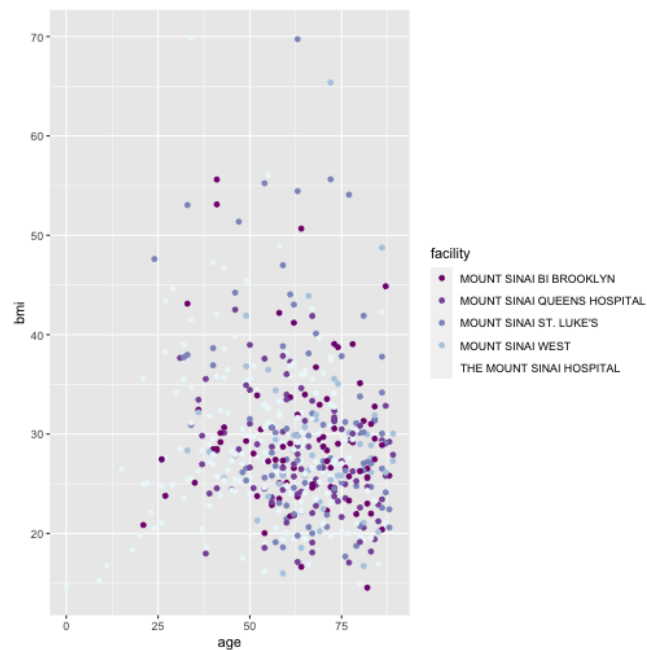

Scales

```
ggplot(sinai_covid, aes(x = age, y = bmi,  
                        color = facility)) +  
  geom_point() +  
  scale_color_brewer(palette = 2)
```



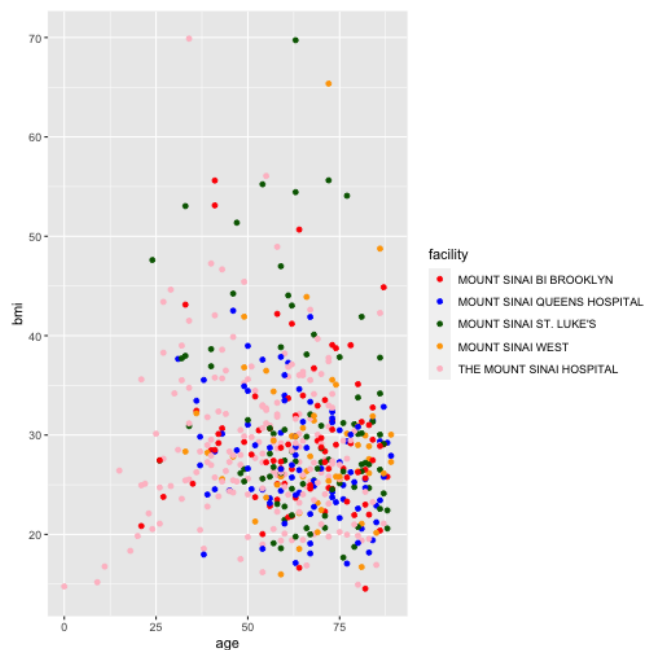
- Invert scale direction

```
ggplot(sinai_covid,  
      aes(x = age, y = bmi,  
          color = facility)) +  
  geom_point() +  
  scale_color_brewer(palette = 3, direction = -1)
```



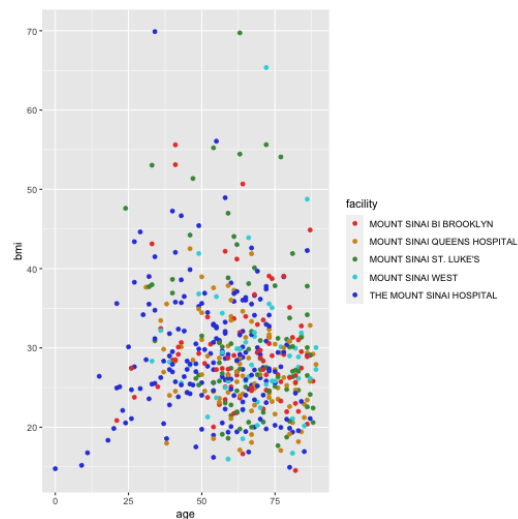
- Using manual scale

```
ggplot(sinai_covid,  
      aes(x = age, y = bmi,  
          color = facility)) +  
  geom_point() +  
  scale_color_manual(values = c("red", "blue", "darkgreen",  
                                "orange", "pink"))
```



- Using manual scale. Look for "html color picker" on Google browser

```
ggplot(sinai_covid,  
      aes(x = age, y = bmi,  
          color = facility)) +  
  geom_point() +  
  scale_color_manual(values = c("#eb4034", "#d19617",  
                                "#429642",  
                                "#31d5de", "#3148de"))
```



Customized position

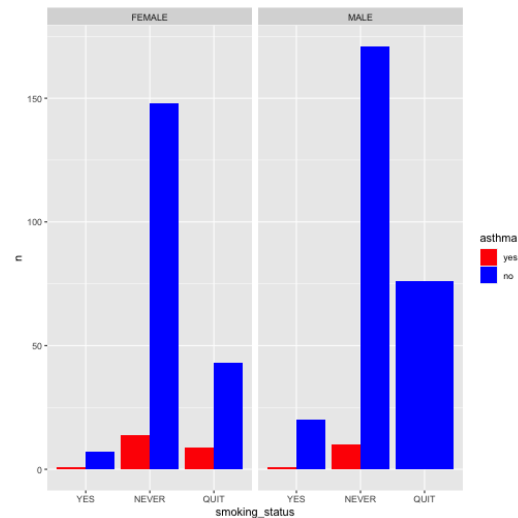
```
count_smoking <- sinai_covid %>%  
  group_by(asthma, sex) %>%  
  count(smoking_status) %>%  
  mutate(asthma = case_when(asthma == 0 ~ "no",  
                             asthma == 1 ~ "yes"),  
         asthma = factor(asthma, levels = c("yes", "no"))  
  )  
  
count_smoking[1:3,]
```

```
## # A tibble: 3 × 4  
## # Groups:   asthma, sex [1]  
##   asthma sex      smoking_status      n  
##   <fct> <chr>    <chr>                <int>  
## 1 no    FEMALE NEVER             148  
## 2 no    FEMALE QUIT             43  
## 3 no    FEMALE YES              7
```

```

ggplot(count_smoking,
       aes(x = smoking_status,
           y = n,
           fill = asthma)) +
  geom_col(position = position_dodge()) +
  scale_x_discrete(limits = c("YES", "NEVER", "QUIT")) +
  scale_fill_manual(values = c("yes" = "red",
                                "no" = "blue")) +
  facet_wrap(vars(sex))

```



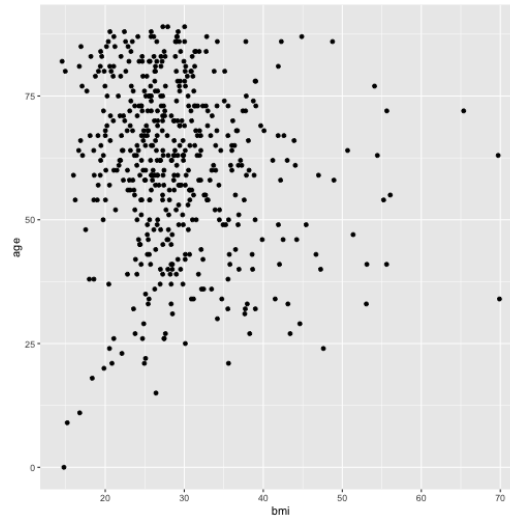
Your turn!

- Plot facility vs bmi
- Use diabetes to color the boxplots
- Change the order of facilities to put The Mount Sinai Hospital at the beginning
- Choose your favorite colors to modify the diabetes coloring

Coordinate systems

- `coord_flip`

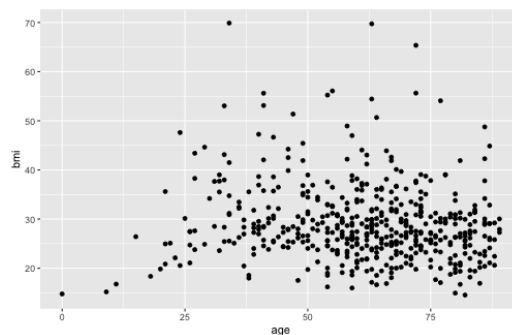
```
ggplot(sinai_covid,  
       aes(x = age, y = bmi)) +  
  geom_point() +  
  coord_flip()
```



- coord_fixed

Makes x and y axis proportional to each other

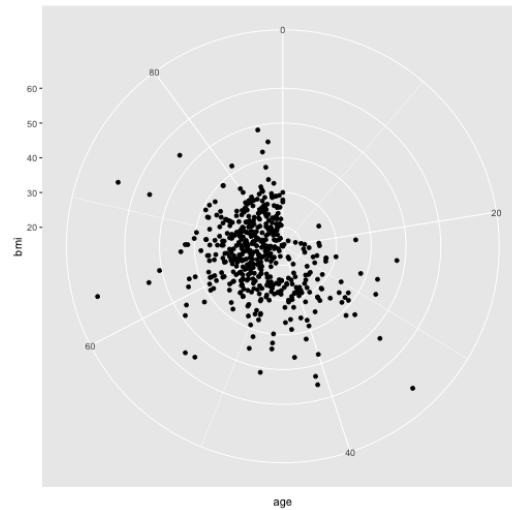
```
ggplot(sinai_covid,  
      aes(x = age, y = bmi)) +  
  geom_point() +  
  coord_fixed()
```



- `coord_polar`

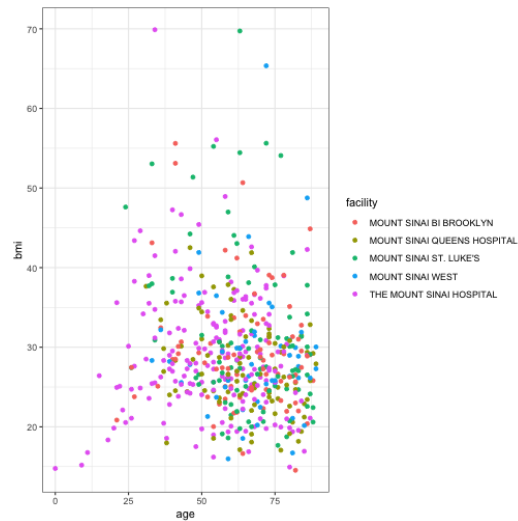
Used for navigation, considers direction and angle

```
ggplot(sinai_covid,  
       aes(x = age, y = bmi)) +  
  geom_point() +  
  coord_polar()
```

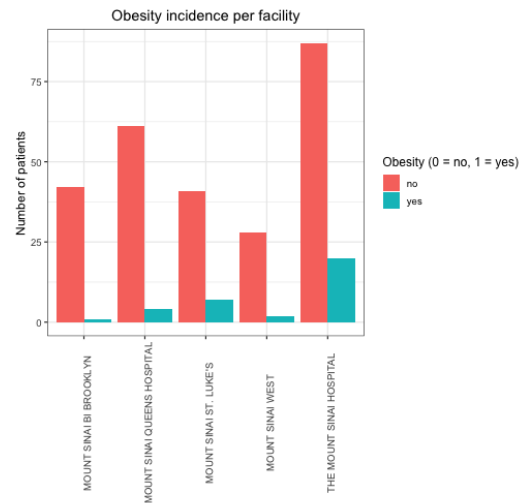


Themes

```
ggplot(sinai_covid,  
  aes(x = age, y = bmi,  
      color = facility)) +  
  geom_point() +  
  theme_bw()
```



```
ggplot(ob_count,
       aes(x = facility, y = n, fill = factor(obesity))) +
  geom_col(position = position_dodge()) +
  labs(title = "Obesity incidence per facility",
       x = "",
       y = "Number of patients",
       fill = "Obesity (0 = no, 1 = yes)") +
  theme_bw(base_size = 12) +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90))
```



Your turn!

- Count the number of patients grouping by smoking status, ethnic group and asthma status
- Plot the number of patients by ethnic group using vertical bars
- Change bars color by asthma status
- Add a plot title, axis titles and modify the legend title if needed
- Explore the available themes and use one of them
- Modify the angle and size of the axis text
- Split in several plots by smoking status

Thanks!



Illustration
by Allison
Horst