Icahn School of Medicine at Mount Sinai

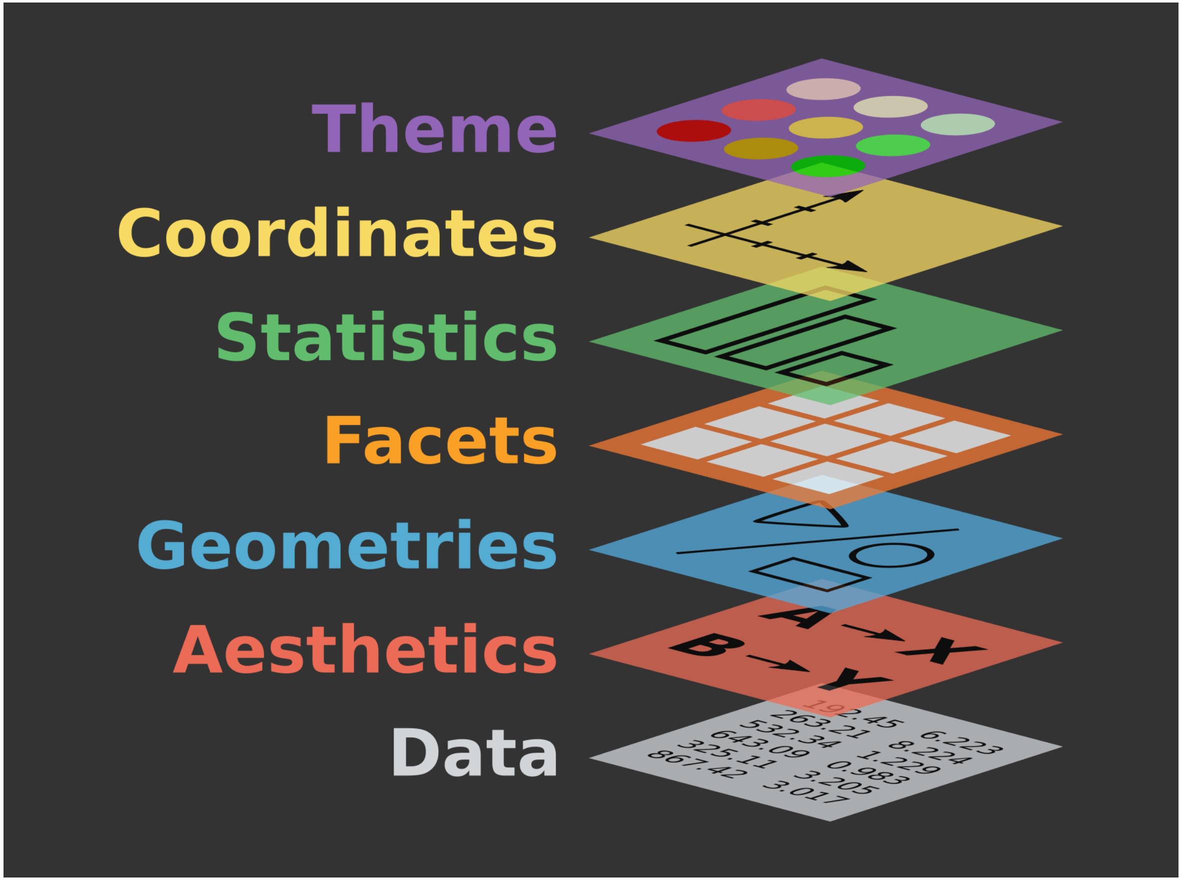
Data visualization

Part II

Joselyn C. Chávez Fuentes

02/20/2024

Let's recap

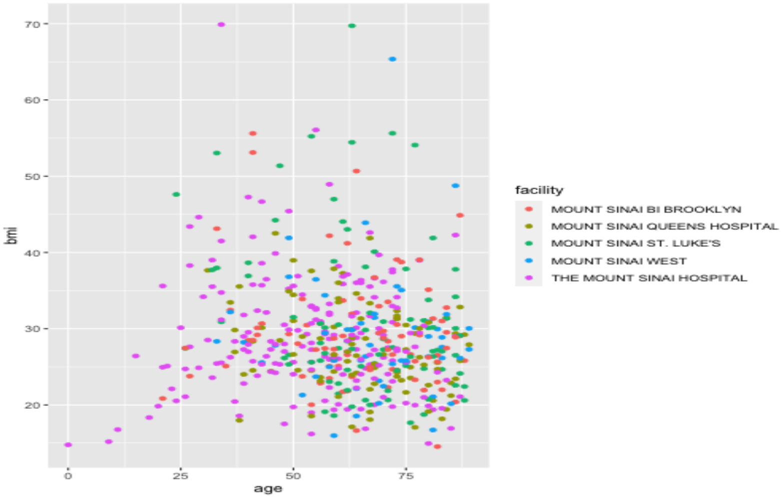

Theme
Coordinates
Statistics
Facets
Geometries
Aesthetics
A → X
B → Y
Data
192.45
263.21
532.34
643.09
325.11
867.42
6.223
8.224
1.229
0.983
3.205
3.017

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 would you include a third variable in the plot?

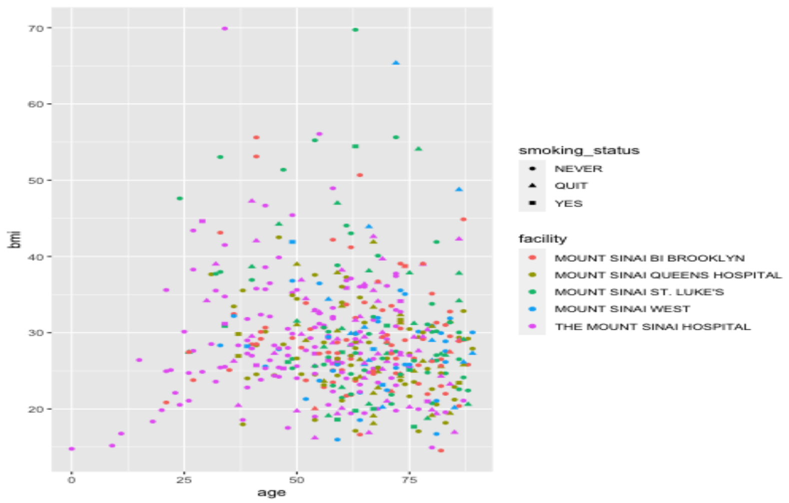
Plotting num vs num vs cat

```
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 = smoking_status)) +
  geom_point()
```



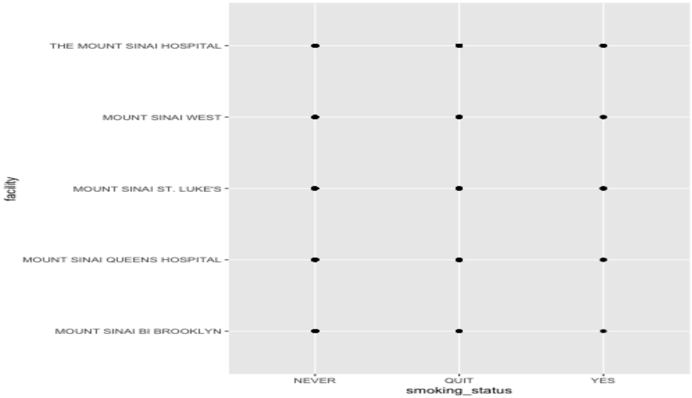
Your turn!

- Create a plot of bmi vs age
- Color by ethnicity
- Add shapes by sex

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

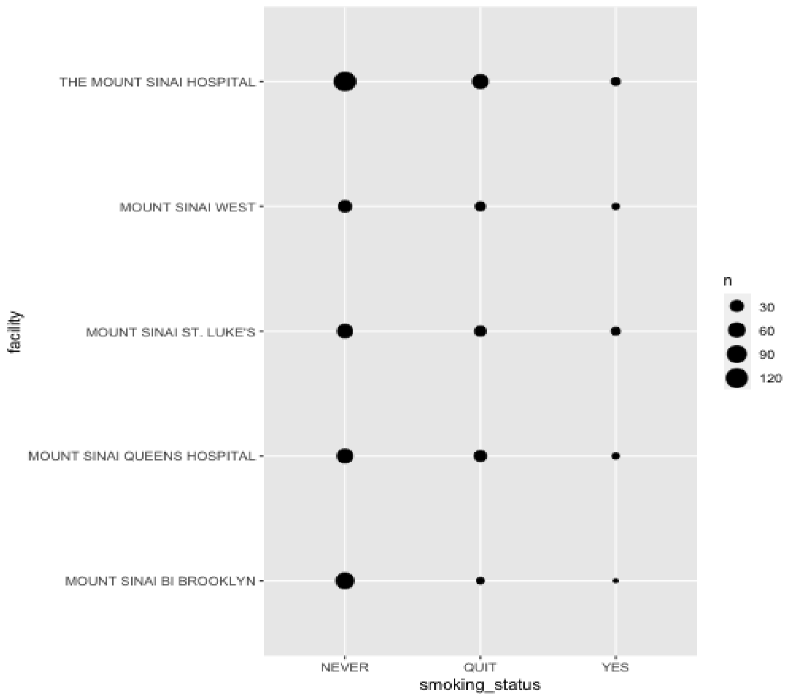
- First create a summarized tibble

```r
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
```

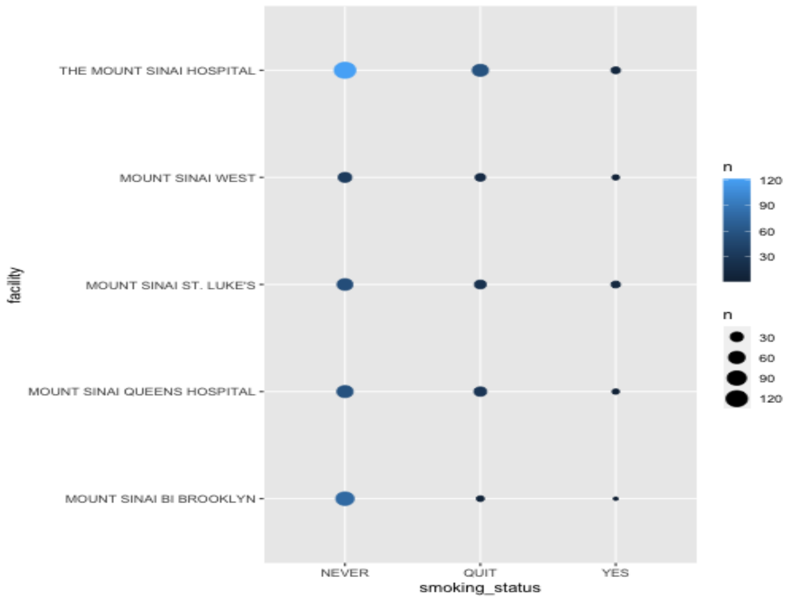
- Plot

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



Plotting cat vs cat vs num

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

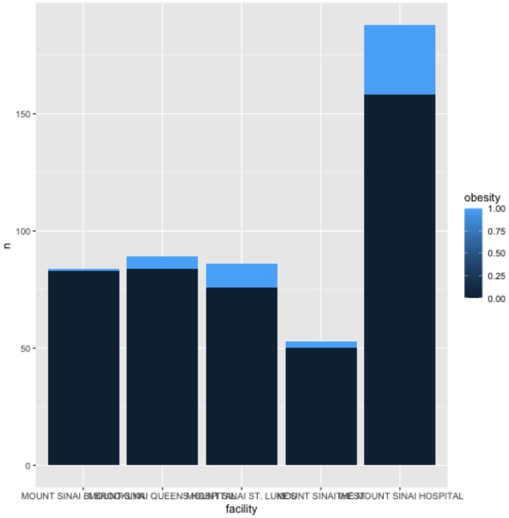


Let's practice

- Count the number of patients with obesity (0 and 1) per facility.
- Create a dots plot, mapping the size of the dots to the obesity counts.

An alternative plot

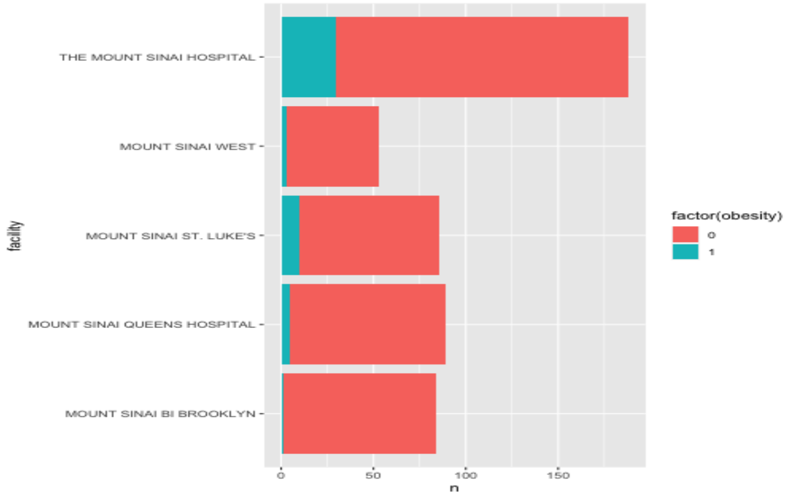
```
ob_count <- sinai_covid %>%
  group_by(facility) %>%
  count(obesity)

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



How do we fix the label?

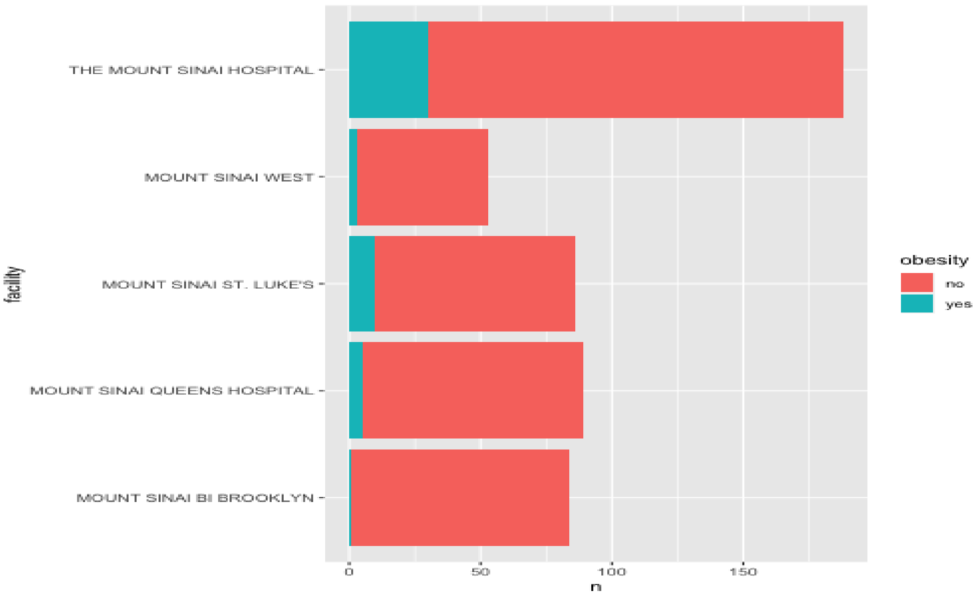
- Let's treat the variable Obesity as factor

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



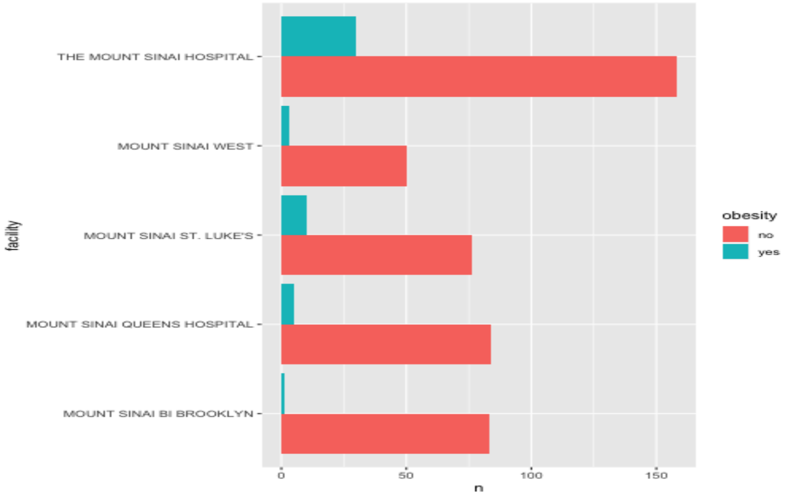
Alternatively, replace the values

```
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()
```



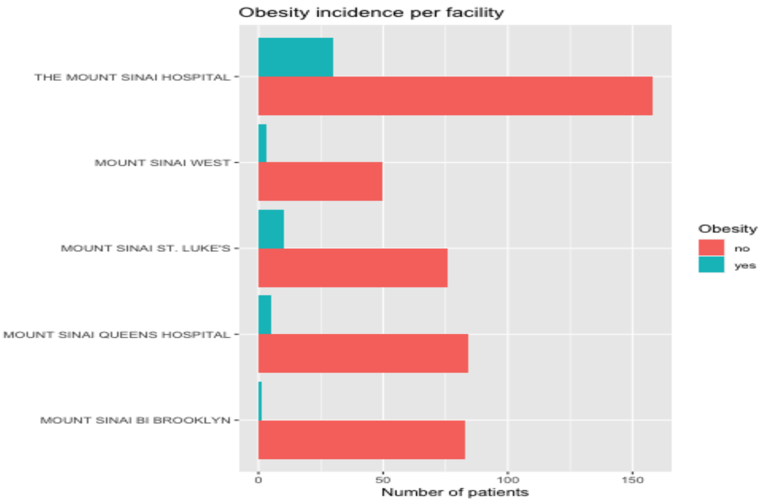
Splitting the bars

```
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 facility",
       x = "Number of patients",
       y = "",
       fill = "Obesity")
```



Your turn

- Create a summarizing tibble with the number of patients per ethnicity and asthma status.

- Use the summarized tibble to create a bar plot, using a position dodge and coloring by asthma status.

- Add a title to the plot, modify the axis titles, and change the legend title.

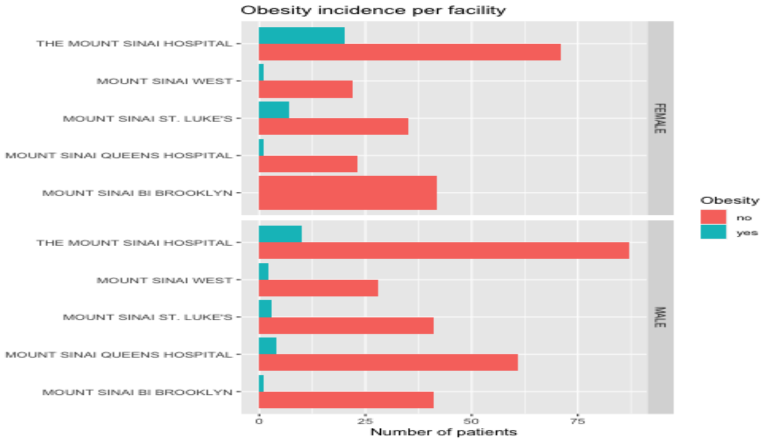
Facets

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

ob_count[1:3,]
```

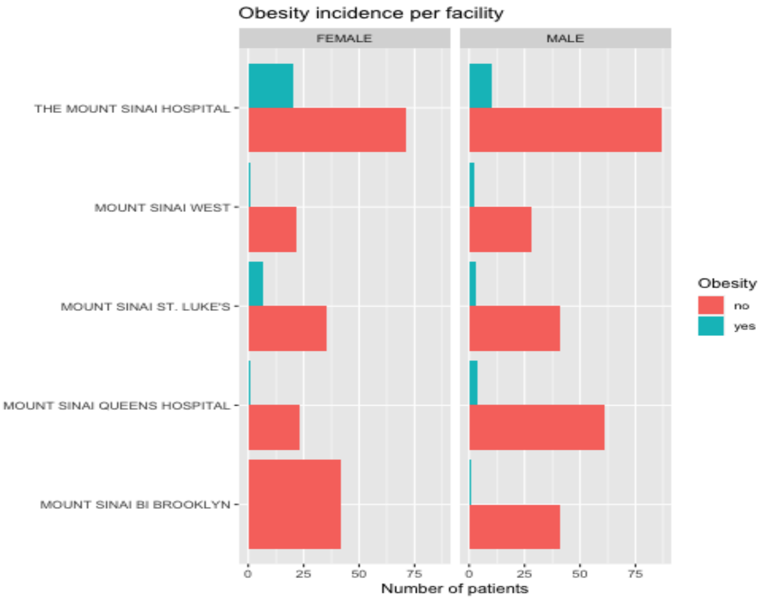
```
## # A tibble: 3 × 4
## # Groups:   facility, sex [2]
##   facility                sex    obesity     n
##   <chr>                   <chr>  <chr>   <int>
## 1 MOUNT SINAI BI BROOKLYN FEMALE no         42
## 2 MOUNT SINAI BI BROOKLYN MALE   no         41
## 3 MOUNT SINAI BI BROOKLYN MALE   yes   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

Facet grid

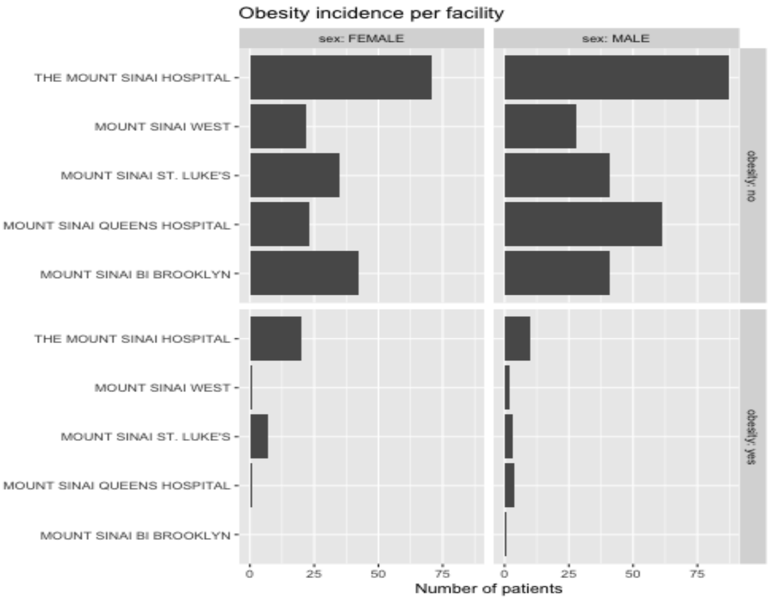
```
ggplot(ob_count,
       aes(x = n, y = facility,
           fill = obesity)) +
  geom_col(position = position_dodge()) +
  labs(title = "Obesity incidence per facility",
       x = "Number of patients",
       y = "",
       fill = "Obesity") +
  facet_grid(rows = vars(sex))
```



```r
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))
```



```
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)
```



Let's practice!

- Count the number of patients depending on their smoking status per ethnic group and asthma status.

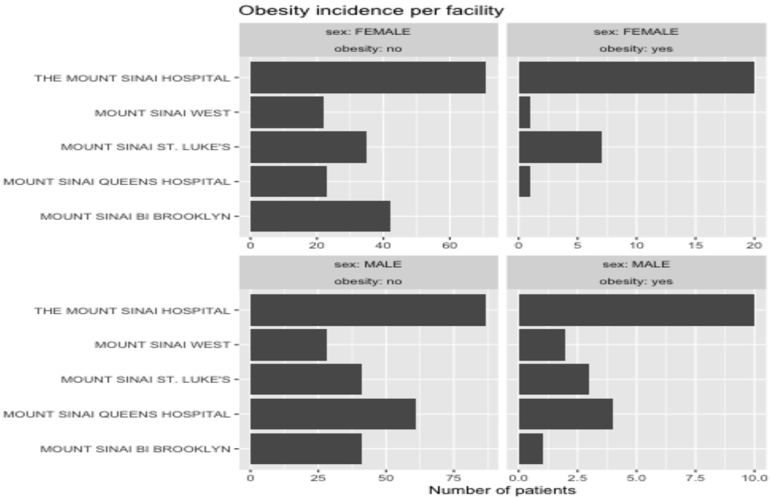
- Represent the data on a column plot, color them by asthma status.

- 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.

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)
```



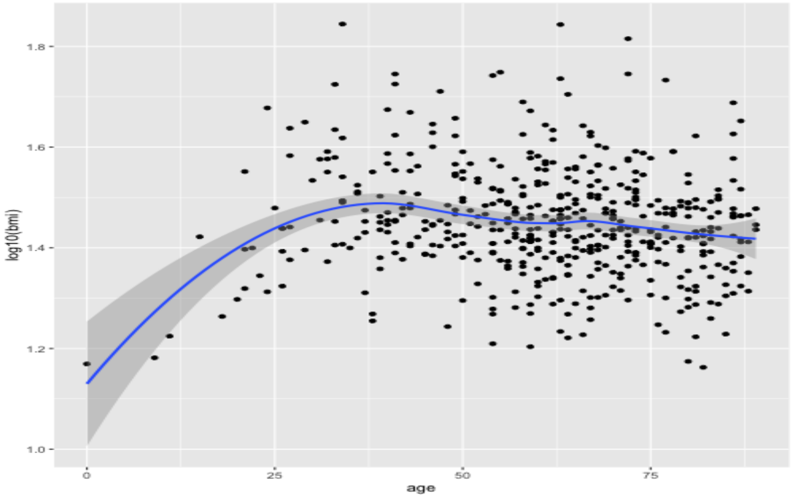
Let's practice!

- Create the previous plot again, but use facet_wrap instead of facet_grid.

Statistical transformations

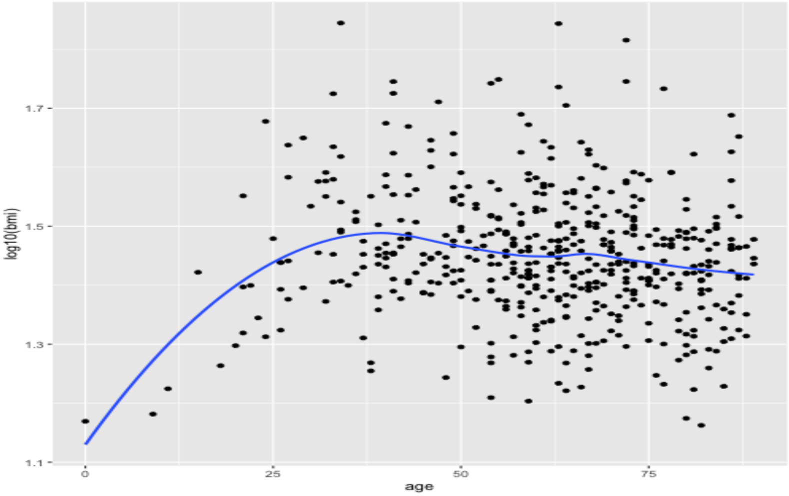
```
sinai_covid %>%
  ggplot(aes(x = age, y = log10(bmi))) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



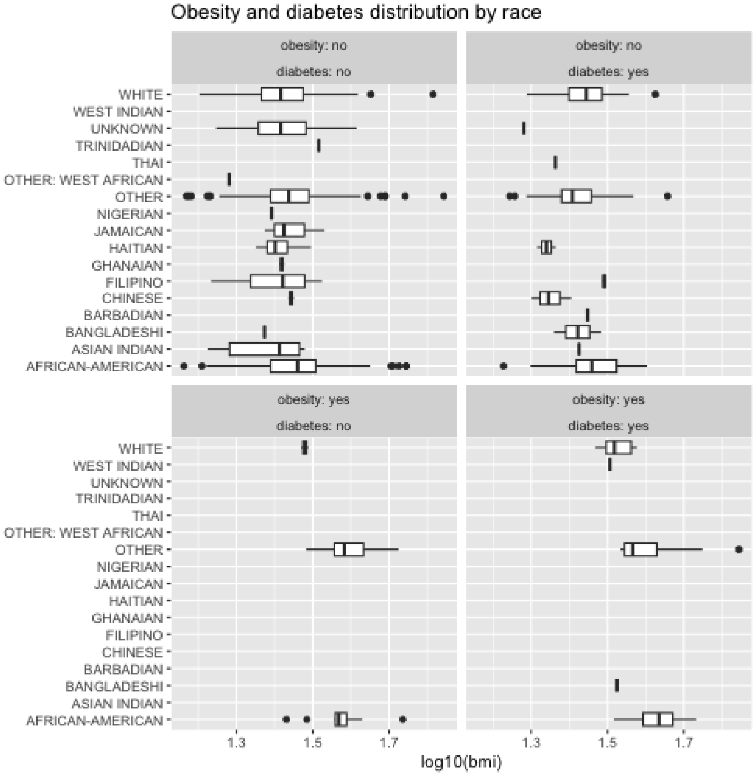
```
sinai_covid %>%
  ggplot(aes(x = age, y = log10(bmi))) +
  geom_point() +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Your turn!

- Write the code to create the following plot:

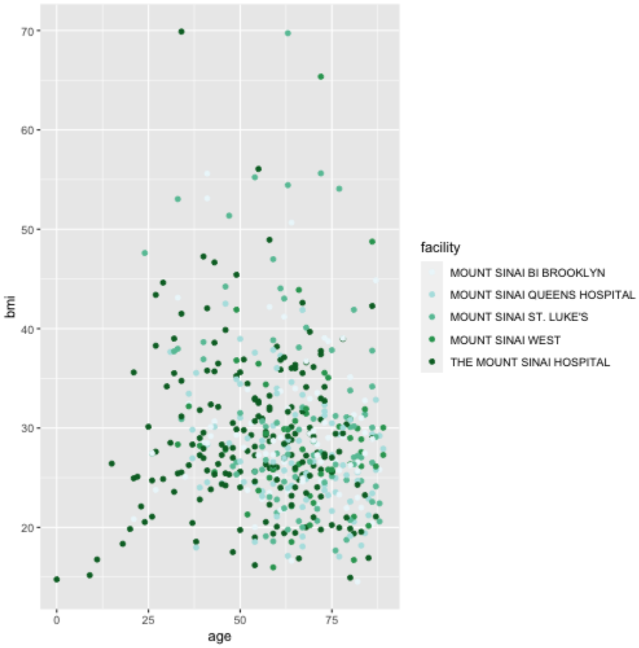


- Possible Answer

```
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 = "")
```

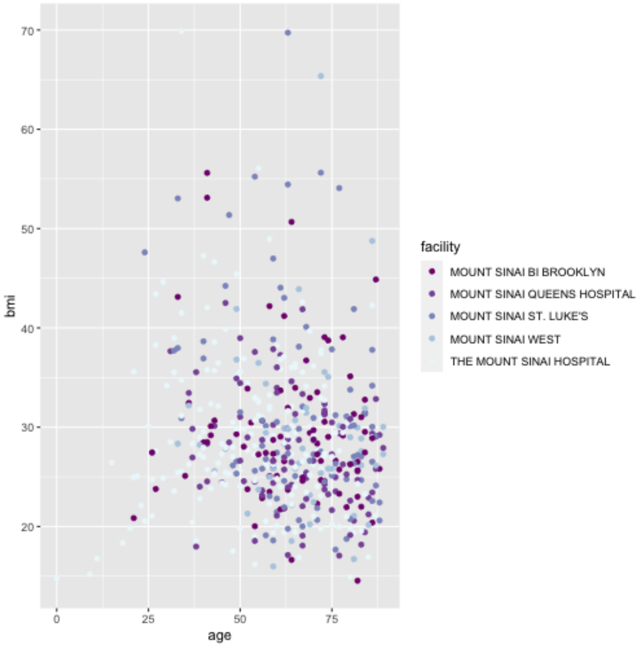
Scales

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



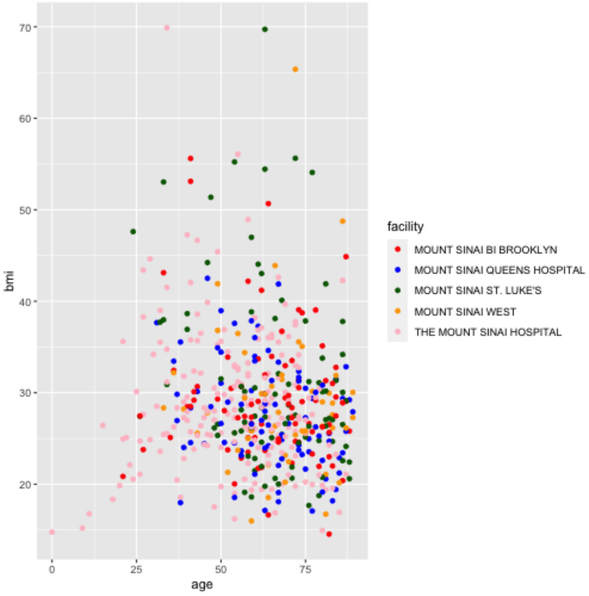
- Invert scale direction

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



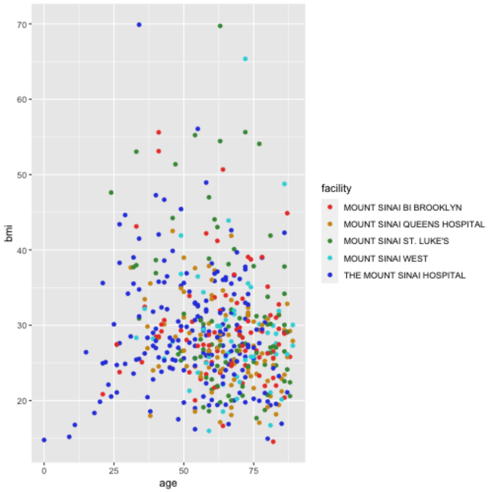
- Using the manual scale

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



- Using the 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"))
```

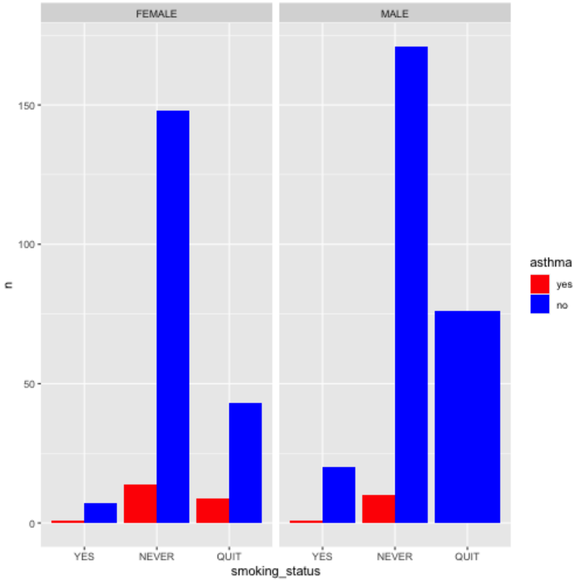


Customized position

```
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))
```

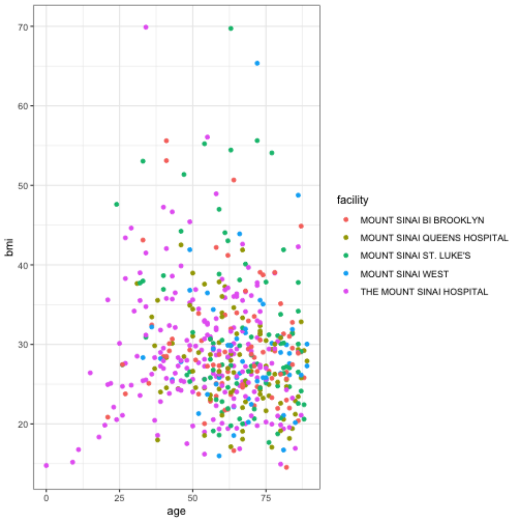


Your turn!

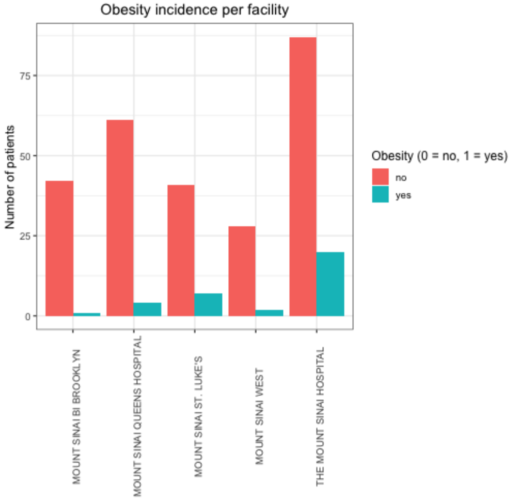
- Plot facility vs bmi.
- Color by diabetes.
- Change the order of facilities to place The Mount Sinai Hospital at the beginning of the x axis.
- Choose your favorite colors to modify the diabetes coloring.

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. Color the bars by asthma status.
- Add a plot title, axis titles and modify the legend title. Explore the available themes and use one.
- Modify the angle and size of the text of the axis. Split in several plots by smoking status.

Thanks!