Final Project

AUTHOR
Madhur Thakur

A) Data import

```
library(readr)
library(tidyverse)
library(lubridate)
```

Import files

```
departments <- read_csv("departments.csv",
                        na = "NULL")
disease_types <- read_csv("disease_types.csv",
                          na = "NULL")
diseases <- read_csv("diseases.csv", na = "NULL")
encounters <- read_csv('encounters.csv', na = 'NULL')
medication_types <- read_csv('medication_types.csv', na = 'NULL')
medications = read_csv('medications.csv', na = 'NULL')
patients <- read_csv('patients.csv', na = 'NULL')
providers <- read_csv('providers.csv', na = 'NULL')
```

B) Data summary

1. Counts:

a. How many patients are in the data set?

```
nrow(patients)
```

```
[1] 500000
```

b. How many encounters?

```
nrow(encounters)
```

```
[1] 8673082
```

c. How many medication types (that is, different potential medications)?

```
nrow(medication_types)
```

```
[1] 2330
```

d. How many medications (that is, specific meds given to specific patients)?

```
nrow(medications)
```

```
[1] 2704000
```

e. How many disease types (that is, different potential diseases)?

```
nrow(disease_types)
```

```
[1] 9175
```

f. How many diseases (that is, specific diseases of specific patients)?

```
nrow(diseases)
```

```
[1] 9687836
```

g. How many departments?

```
nrow(departments)
```

```
[1] 74
```

h. How many providers?

```
nrow(providers)
```

```
[1] 6097
```

A table listing the number of patients stratified by: sex/gender, race/ethnicity, and marital status, simultaneously.

```
patient_count <- patients |>
  count(gender, race, marital_status)
patient_count
```

gender	race	marital_status	n
<chr>	<chr>	<chr>	<int>
F	AFRICAN AMERICAN	DIVORCED	392
F	AFRICAN AMERICAN	LIFE PARTNER	2
F	AFRICAN AMERICAN	MARRIED	1043
F	AFRICAN AMERICAN	SEPARATED	175
F	AFRICAN AMERICAN	SINGLE	3565
F	AFRICAN AMERICAN	UNKNOWN	56

gender <chr>	race <chr>	marital_status <chr>	n <int>
F	AFRICAN AMERICAN	WIDOWED	339
F	ASIAN	DIVORCED	46
F	ASIAN	MARRIED	1318
F	ASIAN	SEPARATED	14

1-10 of 126 rows Previous 1 2 3 4 5 6 ... 13 Next

2. Within the set of encounters ten most common:

a. Medications (i.e., the most common medications prescribed to patients)?

```
medication_10 <- medications |>
  count(medication_id) |>
  arrange(desc(n)) |>
  head(10) |>
  inner_join(medication_types, by = 'medication_id') |>
  select(medication_id, medication_name, n)
medication_10
```

medication_id <dbl>	medication_name <chr>	n <int>
1985	SODIUM CHLORIDE	113638
975	HEPARIN	70556
673	DOCUSATE SODIUM	69355
11	ACETAMINOPHEN	66279
410	CHLORHEXIDINE GLUCONATE	61478
799	FAT EMULSION	58110
541	0.9% SODIUM CHLORIDE	56427
1093	INSULIN REGULAR	52199
536	0.5 NORMAL SALINE WITH POTASSIUM CHLORIDE	51835
980	HEPARIN	44046

1-10 of 10 rows

b. Diseases (i.e., the most common diseases diagnosed for patients)?

```
disease_10 <- diseases |>
  count(disease_id) |>
  arrange(desc(n)) |>
  head(10) |>
  inner_join(disease_types, by = 'disease_id')
disease_10
```

disease_id	n	icd9cm
<dbl>	<int>	<chr>
3057	199128	401.9
1132	105793	250.00
8624	96188	V22.1
9081	77709	V74.1
5626	74145	729.5
5529	56882	724.2
1279	56880	272.4
3571	55351	493.90
1800	54355	311
8611	52834	V20.2

1-10 of 10 rows | 1-3 of 4 columns

c. Departments?

```r
department_10 <- encounters |>
  count(department_id) |>
  arrange(desc(n)) |>
  head(10) |>
  inner_join(departments, by = 'department_id')
department_10
```

department_id	n	department_name
<dbl>	<int>	<chr>
22	6161008	INTERNAL MEDICINE
55	274655	RADIOLOGY - GENERAL
15	256618	FAMILY PRACTICE
30	205301	OPHTHALMOLOGY
32	152366	ORTHOPAEDIC
39	127959	PEDIATRICS - GENERAL
26	125574	OBG - GENERAL
6	99552	CANCER CENTER
52	98276	PSYCHIATRY
13	97810	EMERGENCY MEDICINE

1-10 of 10 rows

d. Providers?

```r
provider_10 <- encounters |>
  count(provider_id) |>
  arrange(desc(n)) |>
  head(10) |>
  inner_join(providers, by = 'provider_id') |>
  select( provider_id, n,first_name, middle_initial, last_name,gender)
provider_10
```

provider_id	n	first_name	middle_initial	last_name	gender
<dbl>	<int>	<chr>	<chr>	<chr>	<chr>
1	2072933	Zane	J	Fisk	M
3209	60372	Sarah	D	Nagle	F
3623	57313	Rodney	J	Bolin	M
5781	40602	Bruce	M	Powers	M
5763	37021	James	D	Fullerton	M
2610	36586	Thomas	R	Wilder-Neligan	M
5758	35262	John	M	Christmas	M
5663	32568	Jane	W	Hoff	F
5990	32403	Gerald	D	Vega	M
880	32093	Laurence	K	Smith	M

1-10 of 10 rows

C) Data manipulation

1. Using the height and weight data that are available, calculate BMIs and report their means in three ways: by sex/gender; by race/ethnicity; and by sex/gender and race/ethnicity simultaneously (i.e., three different tables).

```
encounters_bmi <- encounters |>
  filter(!is.na(height), !is.na(weight)) |>
  mutate(weight_kg = weight * 0.45359237,
         height_m = height * 0.0254) |>
  filter(height_m >= 1 & height_m <= 2.5,
         weight_kg >= 30 & weight_kg <= 200) |>
  mutate(bmi = weight_kg / height_m^2) |>
  inner_join(patients, by = 'patient_id')
```

Mean by gender

```
encounters_bmi |>
  group_by(gender) |>
  summarise(mean(bmi))
```

gender	mean(bmi)
<chr>	<dbl>
F	28.37837
M	27.90221

2 rows

```
encounters_bmi |>
  group_by(race) |>
  summarise(mean(bmi))
```

race <chr>	mean(bmi) <dbl>
AFRICAN AMERICAN	29.88142
ASIAN	23.74065
CAUCASIAN	28.39081
DECLINED	27.43509
HISPANIC	28.18227
MULTIRACIAL	26.19185
NATIVE AMERICAN	29.81011
OTHER	26.93206
PACIFIC ISLANDER	25.86115
UNKNOWN	26.50505

1-10 of 10 rows

```
encounters_bmi |>
  group_by(gender, race) |>
  summarise(mean(bmi))
```

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

gender <chr>	race <chr>	mean(bmi) <dbl>
F	AFRICAN AMERICAN	31.49467
F	ASIAN	23.35652
F	CAUCASIAN	28.62514
F	DECLINED	27.48967
F	HISPANIC	28.63072
F	MULTIRACIAL	27.29122
F	NATIVE AMERICAN	30.17885
F	OTHER	26.38891
F	PACIFIC ISLANDER	25.50838
F	UNKNOWN	26.18302

1-10 of 20 rows Previous **1** 2 Next

Filter out any encounters that have a stay of one day or less (i.e., keep encounters with LOSes longer than one day), and report the mean LOSes for the remainder in four ways: overall; by sex/gender; by department; and by sex/gender and department simultaneously.

```r
encounters_los <- encounters |>
  filter(!is.na(admit_date), !is.na(discharge_date)) |>
  mutate(los2=discharge_date-admit_date) |>
  mutate(los3 = as.numeric(los2)) |>
  filter(los3 > 1)

encounters_los_mean <- encounters_los |>
  inner_join(departments, by = 'department_id') |>
  inner_join(patients, by = 'patient_id') |>
  select(gender, race, department_name, los3)
```

```r
mean(encounters_los_mean$los3)
```

```
[1] 88578.97
```

```r
encounters_los_mean |>
  group_by(gender) |>
  summarise(mean(los3))
```

gender	mean(los3)
<chr>	<dbl>
F	76504.59
M	104042.65

2 rows

By gender

```r
encounters_los_mean |>
  group_by(gender) |>
  summarise(mean(los3))
```

gender	mean(los3)
<chr>	<dbl>
F	76504.59
M	104042.65

2 rows

by department

```r
encounters_los_mean |>
  group_by(department_name) |>
  summarise(mean(los3))
```

department_name	mean(los3)
<chr>	<dbl>
ALLERGY AND IMMUNOLOGY	22010.30
ANESTHESIOLOGY	33388.03
AUDIOLOGY	51388.24
BEHAVIORAL HEALTH	530487.99
BURN UNIT	1007139.84
CANCER CENTER	44471.05
CARDIAC REHABILITATION	45629.46
CARDIOLOGY	363909.06
CLINICAL RESEARCH	45803.80
DENTISTRY	44015.54

1-10 of 74 rows Previous **1** 2 3 4 5 6 ... 8 Next

by gender and department

```
encounters_los_mean |>
  group_by(gender, department_name) |>
  summarise(mean(los3))
```

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

gender	department_name	mean(los3)
<chr>	<chr>	<dbl>
F	ALLERGY AND IMMUNOLOGY	22456.95
F	ANESTHESIOLOGY	33278.41
F	AUDIOLOGY	46682.67
F	BEHAVIORAL HEALTH	454105.66
F	BURN UNIT	1169096.53
F	CANCER CENTER	44415.76
F	CARDIAC REHABILITATION	45510.39
F	CARDIOLOGY	367666.37
F	CLINICAL RESEARCH	45173.29
F	DENTISTRY	43644.62

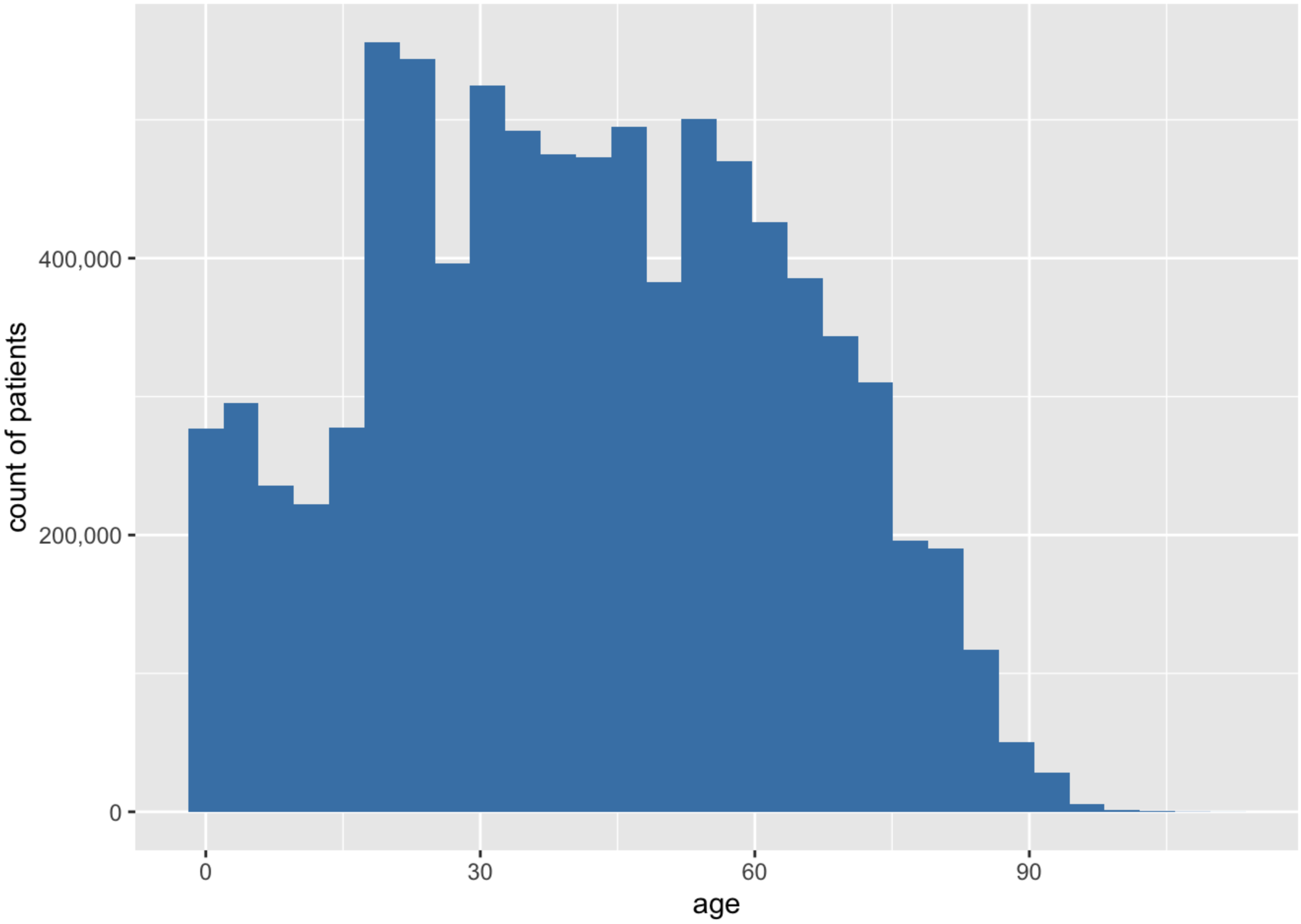
1-10 of 147 rows Previous **1** 2 3 4 5 6 ... 15 Next

D) Data visualization

1. A histogram of patient age in years at time of encounter.

```
ggplot(encounters, aes(x = age)) +
  geom_histogram(fill = 'steelblue') +
  labs(x = 'age', y = 'count of patients') +
  scale_y_continuous(labels = scales::comma_format())
```

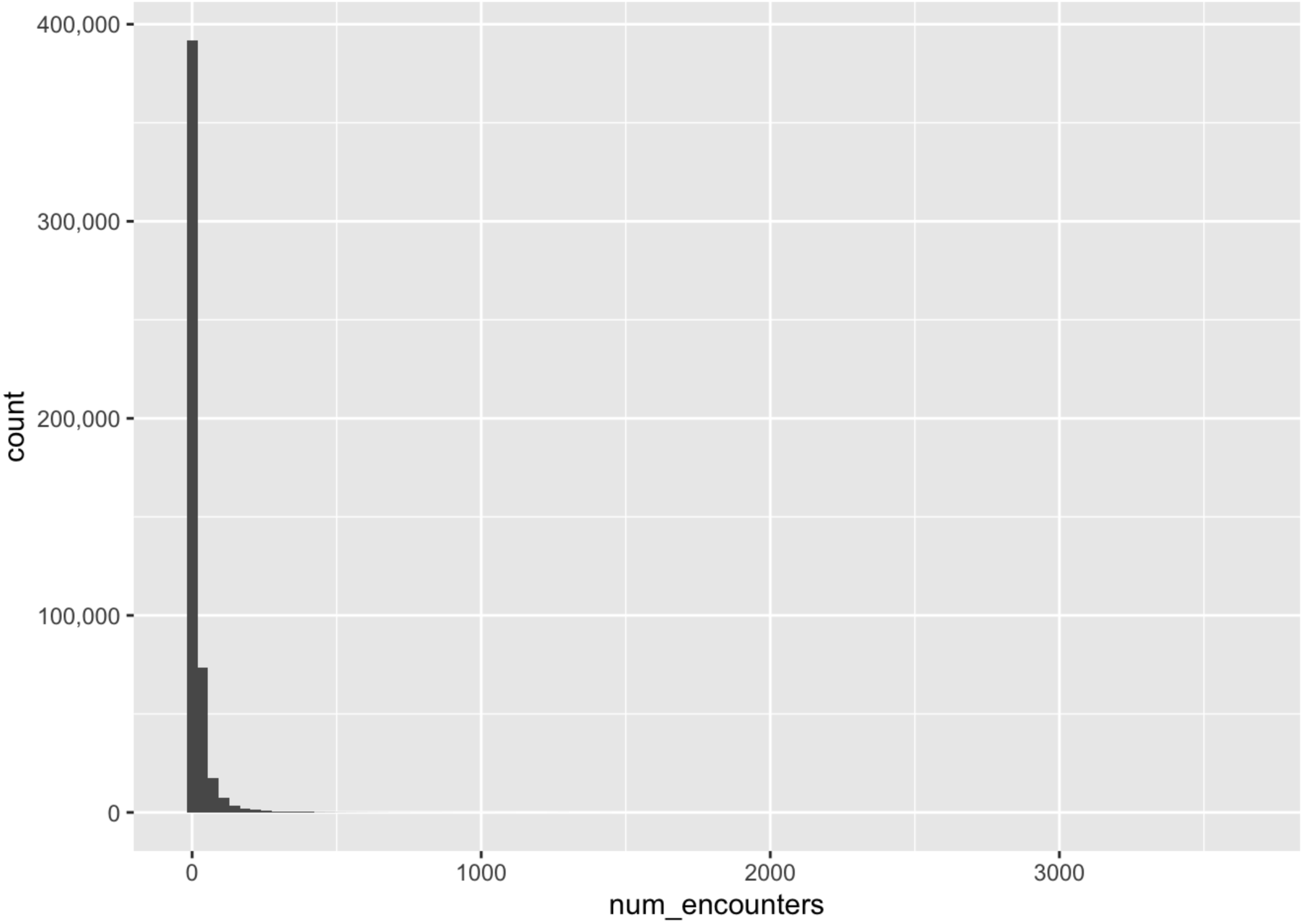
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2. A histogram of number of encounters

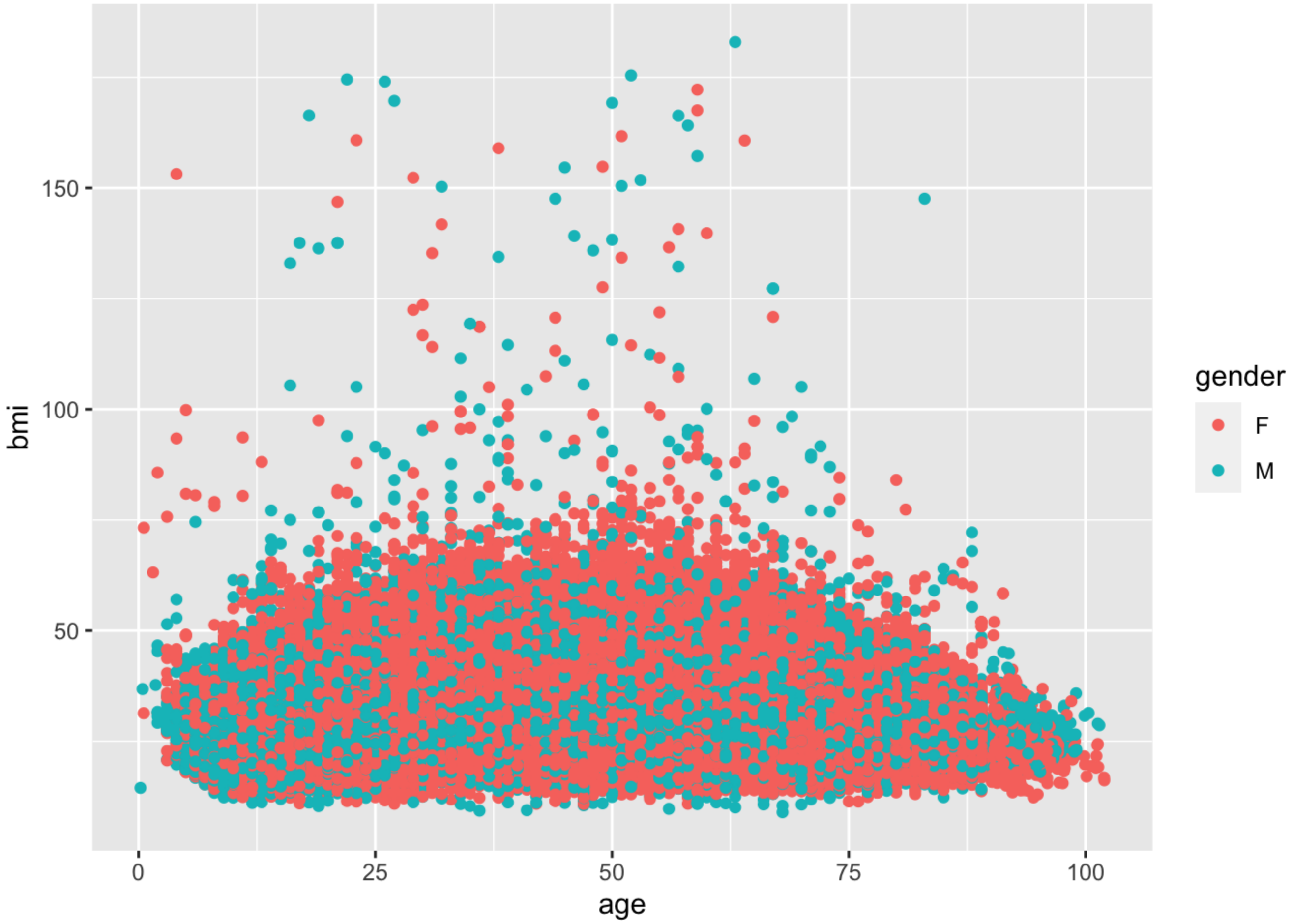
```
patient_enc <- encounters |>
  select(encounter_id, patient_id) |>
  group_by(patient_id) |>
  summarise(num_encounters = n())

ggplot(patient_enc, aes(x = num_encounters)) +
  geom_histogram(bins = 100)+
  scale_y_continuous(labels = scales::comma_format())
```



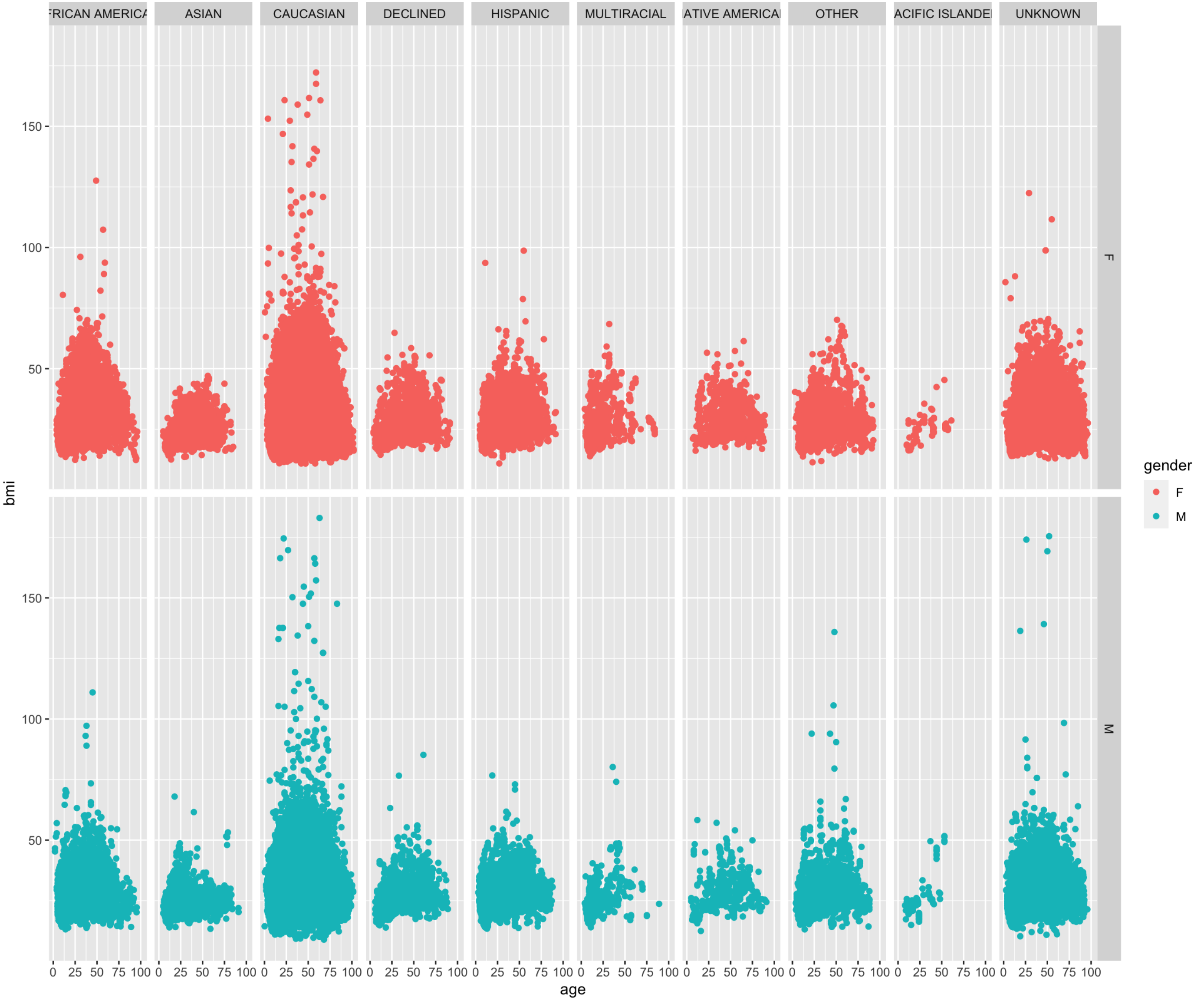
3. A scatterplot of BMIs by age, giving different colors to different sexes/genders.

```
encounters_bmi |>
  ggplot(aes(x = age, y = bmi, color = gender))+
  geom_point()
```



A set of panels (facets) of scatterplots of BMI by age, with different plots for each combination of sex/gender and race/ethnicity.

```
ggplot(encounters_bmi, aes(x = age, y = bmi)) +
  geom_point(aes(color = gender)) +
  facet_grid(gender~race)
```



E) Missing values.

Some variables may have missing data. What are the approximate rates of missing data? Would any of these have an impact on data analysis? You may keep this discussion mostly qualitative.

```
colMeans(is.na(encounters))
```

```
   encounter_id      patient_id      admit_date discharge_date   department_id 
      0.0000000       0.0000000       0.0000000      0.1846650       0.0000000 
    provider_id             age     bp_systolic    bp_diastolic     temperature 
      0.0000000       0.0000000       0.9220743      0.9220799       0.9288419 
          pulse          height          weight 
      0.9215037       0.9371708       0.8838089 
```

```
range(encounters$bp_systolic, na.rm = TRUE)
```

```
[1]   0 313
```

```
range(encounters$age, na.rm = TRUE)
```

```
[1]   0.0000 111.6995
```

```
range(encounters$temperature, na.rm = TRUE)
```

```
[1]  32.90 107.96
```

```
range(encounters$weight, na.rm = TRUE)
```

```
[1]    0.01 1260.00
```

```
range(encounters$height, na.rm = TRUE)
```

```
[1]   0.06 131.88
```

```
range(encounters$pulse, na.rm = TRUE)
```

```
[1]   0 250
```

PART 2

1. Latest BMI

```
latest_bmi <- encounters |>
  filter(!is.na(height), !is.na(weight)) |>
  arrange(admit_date) |>
  tail(1000) |>
  mutate(weight_kg = weight * 0.45359237,
         height_m = height * 0.0254) |>
  filter(height_m >= 1 & height_m <= 2.5,
         weight_kg >= 30 & weight_kg <= 200) |>
  mutate(bmi = weight_kg / height_m^2) |>
  inner_join(patients, by = 'patient_id')

summary(latest_bmi$bmi)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.67   23.03   27.02   28.71   32.95   66.94 
```

```
latest_bmi |>
  group_by(gender) |>
  summarise(mean(bmi))
```

gender	mean(bmi)
<chr>	<dbl>
F	29.40415
M	27.88112

2 rows

```r
latest_bmi |>
  group_by(race) |>
  summarise(mean(bmi))
```

race	mean(bmi)
<chr>	<dbl>
AFRICAN AMERICAN	30.00441
ASIAN	24.54578
CAUCASIAN	28.65035
DECLINED	32.70522
HISPANIC	29.79523
MULTIRACIAL	30.83771
NATIVE AMERICAN	20.24764
UNKNOWN	24.89087

8 rows

```r
latest_bmi |>
  group_by(gender, race) |>
  summarise(mean(bmi))
```

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

gender	race	mean(bmi)
<chr>	<chr>	<dbl>
F	AFRICAN AMERICAN	29.64465
F	ASIAN	24.97388
F	CAUCASIAN	29.48918
F	DECLINED	34.12172
F	HISPANIC	30.19329
F	MULTIRACIAL	30.90600
F	UNKNOWN	21.46317
M	AFRICAN AMERICAN	30.27851
M	ASIAN	24.03207
M	CAUCASIAN	27.63251

1-10 of 15 rows Previous **1** 2 Next