EHR Data Quality Control and Analysis

Fereshteh Izadi

19 July 2025

```
# Set path to data directory
data_path <- "C:/Users/User/Downloads/test_data/data/dest/"

# Load main datasets
```

```
patients <- read_csv(paste0(data_path, "patients.csv"),show_col_types = FALSE)
conditions <- read_csv(paste0(data_path, "conditions.csv"),show_col_types = FALSE)
observations <- read_csv(paste0(data_path, "observations.csv"),show_col_types = FALSE)
medications <- read_csv(paste0(data_path, "medications.csv"),show_col_types = FALSE)
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

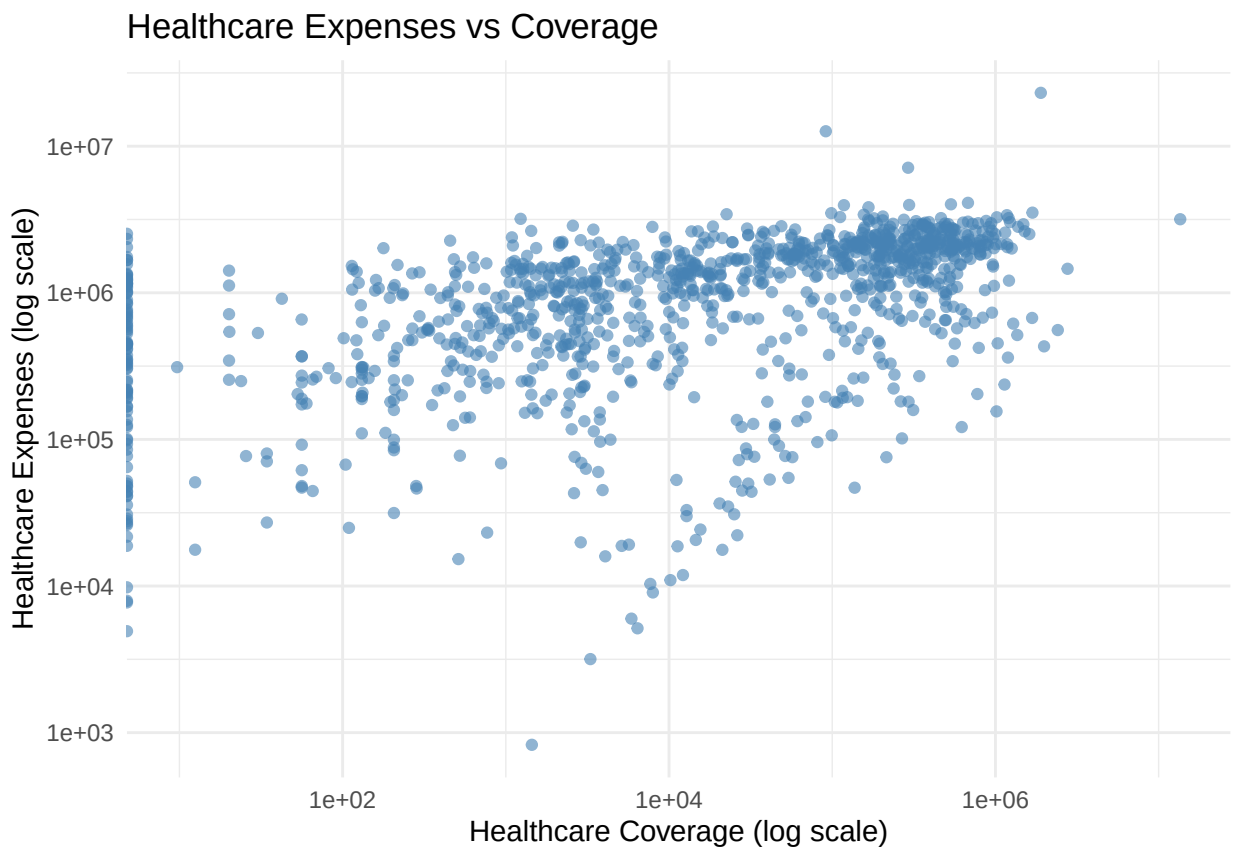
```
encounters <- read_csv(paste0(data_path, "encounters.csv"),show_col_types = FALSE)

# Load dictionaries
dict_snomed <- read_csv(paste0(data_path, "dictionary_snomed.csv"),show_col_types = FALSE)
dict_rxnorm <- read_csv(paste0(data_path, "dictionary_rxnorm.csv"),show_col_types = FALSE)
dict_loinc  <- read_csv(paste0(data_path, "dictionary_loinc.csv"),show_col_types = FALSE)
```

Patients Quality Control (QC)

```
# Convert date columns
patients$BIRTHDATE <- as.Date(patients$BIRTHDATE)
patients$DEATHDATE <- as.Date(patients$DEATHDATE)

# Calculate age
patients$AGE <- as.numeric(floor(interval(patients$BIRTHDATE, Sys.Date()) / years(1)))

# HEALTHCARE_EXPENSES vs HEALTHCARE_COVERAGE Scatter plot with log scales
ggplot(patients, aes(x = HEALTHCARE_COVERAGE, y = HEALTHCARE_EXPENSES)) +
  geom_point(alpha = 0.6, color = "steelblue") +
  scale_x_log10() +
  scale_y_log10() +
  labs(
    title = "Healthcare Expenses vs Coverage",
```

```
    x = "Healthcare Coverage (log scale)",
    y = "Healthcare Expenses (log scale)"
  ) +
  theme_minimal()
```

```
## Warning in scale_x_log10(): log-10 transformation introduced infinite values.
```




```
# Filter out invalid data
patients <- patients %>%
  mutate(
    # Logical flags for invalid healthcare expense/coverage
    FLAG_INVALID_EXPENSE = HEALTHCARE_EXPENSES < 500 | HEALTHCARE_EXPENSES > 2e7,
    FLAG_INVALID_COVERAGE = HEALTHCARE_COVERAGE > HEALTHCARE_EXPENSES
  ) %>%
  filter(
    AGE >= 0 & AGE <= 100,
    is.na(DEATHDATE) | BIRTHDATE <= DEATHDATE,
    !(LAT == 0 & LON == 0),
    LAT >= -90 & LAT <= 90,
    LON >= -180 & LON <= 180,
    !(HEALTHCARE_EXPENSES == 0 & HEALTHCARE_COVERAGE > 1000),
    GENDER %in% c("M", "F", "", NA),
    !((PREFIX == "Mr." & GENDER == "F") |
      (PREFIX %in% c("Mrs.", "Ms.", "Miss") & GENDER == "M"))
```

```
  )

# Remove US state + UK city mismatches
us_states <- state.abb
uk_cities <- c("London", "Manchester", "Leeds", "Birmingham", "Glasgow", "Liverpool")
patients <- patients[!(patients$CITY %in% uk_cities & patients$STATE %in% us_states), ]

# Fix race coding
patients$RACE[patients$RACE == "XJniDSe"] <- "other (possibly miscoded)"

# ZIP formatting
patients$ZIP <- sprintf("%05d", patients$ZIP)

# The number of patients and unique patients after QC
print(dim(patients))
```

```
## [1] 930  28
```

```
print(n_distinct(patients$Id))
```

```
## [1] 930
```

Encounters QC

```
# Convert date columns
encounters$START <- as.POSIXct(encounters$START)
encounters$STOP <- as.POSIXct(encounters$STOP)
# Filter out invalid data
encounters <- encounters %>%
  filter(is.na(STOP) | STOP >= START) %>%
  filter(BASE_ENCOUNTER_COST >= 0, TOTAL_CLAIM_COST >= 0, PAYER_COVERAGE >= 0) %>%
  filter(PATIENT %in% patients$Id)
```

Conditions QC

```
# Convert date columns
conditions$START <- as.Date(conditions$START)
conditions$STOP <- as.Date(conditions$STOP)

# Filter out invalid data
conditions <- conditions %>%
  filter(is.na(STOP) | START <= STOP) %>%
  filter(PATIENT %in% patients$Id) %>%
  filter(ENCOUNTER %in% encounters$Id) %>%
  filter(CODE %in% dict_snomed$CODE)

conditions <- left_join(conditions, dict_snomed, by = "CODE")
```

```
## Warning in left_join(conditions, dict_snomed, by = "CODE"): Detected an unexpected many-to-many rela
## i Row 304 of `x` matches multiple rows in `y`.
## i Row 15 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
# Add 0/1 flags for socioeconomic factors based on DESCRIPTION content
conditions <- conditions %>%
  mutate(
    factor_education = as.integer(grepl("education|high school|primary school|higher education", DESCRI
    factor_employment = as.integer(grepl("employment|unemployed|labor force", DESCRIPTION, ignore.case =
    factor_stress = as.integer(grepl("stress|anxiety|panic", DESCRIPTION, ignore.case = TRUE)),
    factor_refugee = as.integer(grepl("refugee", DESCRIPTION, ignore.case = TRUE)),
    factor_criminal = as.integer(grepl("criminal", DESCRIPTION, ignore.case = TRUE)),
    factor_violence = as.integer(grepl("violence|abuse|victim", DESCRIPTION, ignore.case = TRUE)),
    factor_transport = as.integer(grepl("transport", DESCRIPTION, ignore.case = TRUE)),
    factor_drugs_alcohol = as.integer(grepl("alcohol|misuses drugs", DESCRIPTION, ignore.case = TRUE)),
    factor_social_isolation = as.integer(grepl("social|isolation|limited contact", DESCRIPTION, ignore.
    factor_military = as.integer(grepl("armed forces|military", DESCRIPTION, ignore.case = TRUE)),
    factor_housing = as.integer(grepl("housing", DESCRIPTION, ignore.case = TRUE))
  )
```

Observations QC

```
# Convert date columns
observations$DATE <- as.POSIXct(observations$DATE)
observations$CODE <- as.character(observations$CODE)

dict_loinc$CODE <- as.character(dict_loinc$CODE)

# Filter out invalid data
observations <- observations %>%
  left_join(dict_loinc, by = "CODE") %>%
  filter(!is.na(DESCRIPTION), !is.na(VALUE), VALUE != "", !is.na(TYPE), TYPE != "")
```

```
## Warning in left_join(., dict_loinc, by = "CODE"): Detected an unexpected many-to-many relationship be
## i Row 10 of `x` matches multiple rows in `y`.
## i Row 1 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
observations$VALUE_NUMERIC <- suppressWarnings(as.numeric(observations$VALUE))
observations <- observations[observations$TYPE != "numeric" | !is.na(observations$VALUE_NUMERIC), ]

# Weight-kg mismatch
mismatch <- grepl("weight", tolower(observations$DESCRIPTION)) & !grepl("kg", tolower(observations$UNIT
observations <- observations[!mismatch, ]

observations <- observations %>%
  filter(PATIENT %in% patients$Id, ENCOUNTER %in% encounters$Id)
```

```r
# Join observations with patients to get DEATHDATE
observations <- observations %>%
  left_join(patients %>% select(Id, DEATHDATE), by = c("PATIENT" = "Id"))

# Keep only observations that occurred before or on the DEATHDATE, or if DEATHDATE is missing
observations <- observations %>%
  filter(is.na(DEATHDATE) | DATE <= DEATHDATE)
```

Medications QC

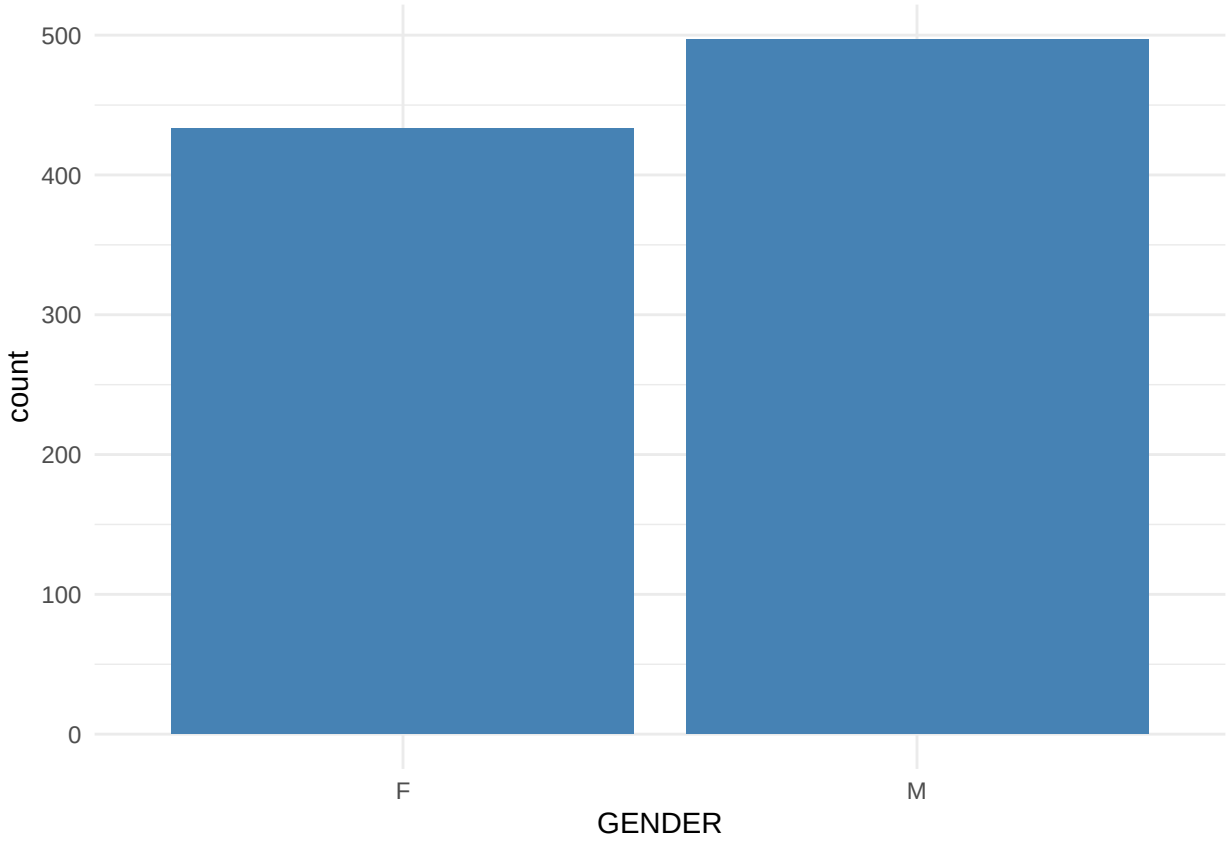
```r
# Convert date columns
medications$START <- as.POSIXct(medications$START)
medications$STOP <- as.POSIXct(medications$STOP)

medications <- medications %>%
  filter(is.na(STOP) | START <= STOP, PATIENT %in% patients$Id, ENCOUNTER %in% encounters$Id) %>%
  filter(BASE_COST >= 0, PAYER_COVERAGE >= 0, TOTALCOST >= 0)

medications <- medications %>%
  left_join(dict_rxnorm, by = "CODE") %>%
  filter(!is.na(DESCRIPTION)) %>%
  filter(BASE_COST <= TOTALCOST) %>%
  filter(abs((BASE_COST + PAYER_COVERAGE) - TOTALCOST) <= 1)
```

```
## Warning in left_join(., dict_rxnorm, by = "CODE"): Detected an unexpected many-to-many relationship b
## i Row 51 of `x` matches multiple rows in `y`.
## i Row 9 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

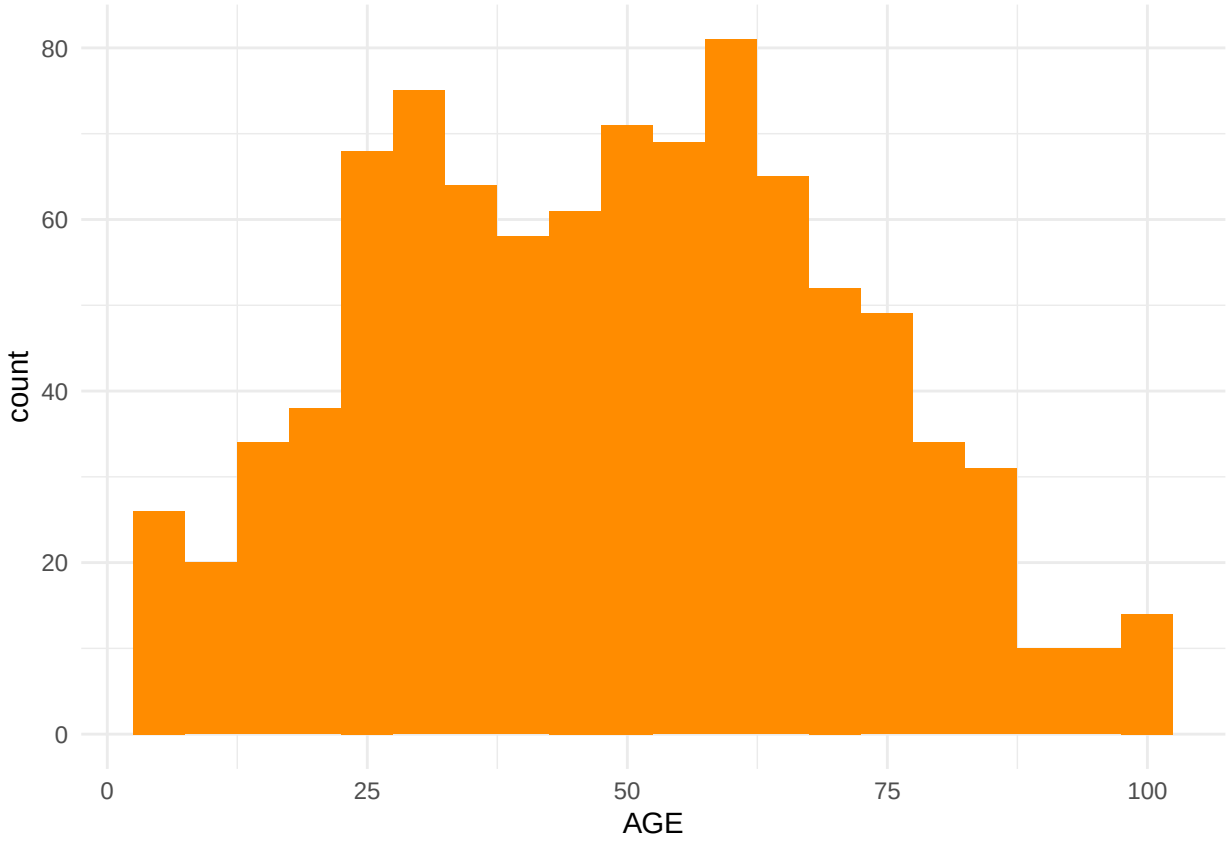
```r
# Join medications with patients to get DEATHDATE
medications <- medications %>%
  left_join(patients %>% select(Id, DEATHDATE), by = c("PATIENT" = "Id"))

# Filter out medications that start after the patient's death
medications <- medications %>%
  filter(is.na(DEATHDATE) | START <= DEATHDATE)
```

Descriptive Summaries and Visualizations

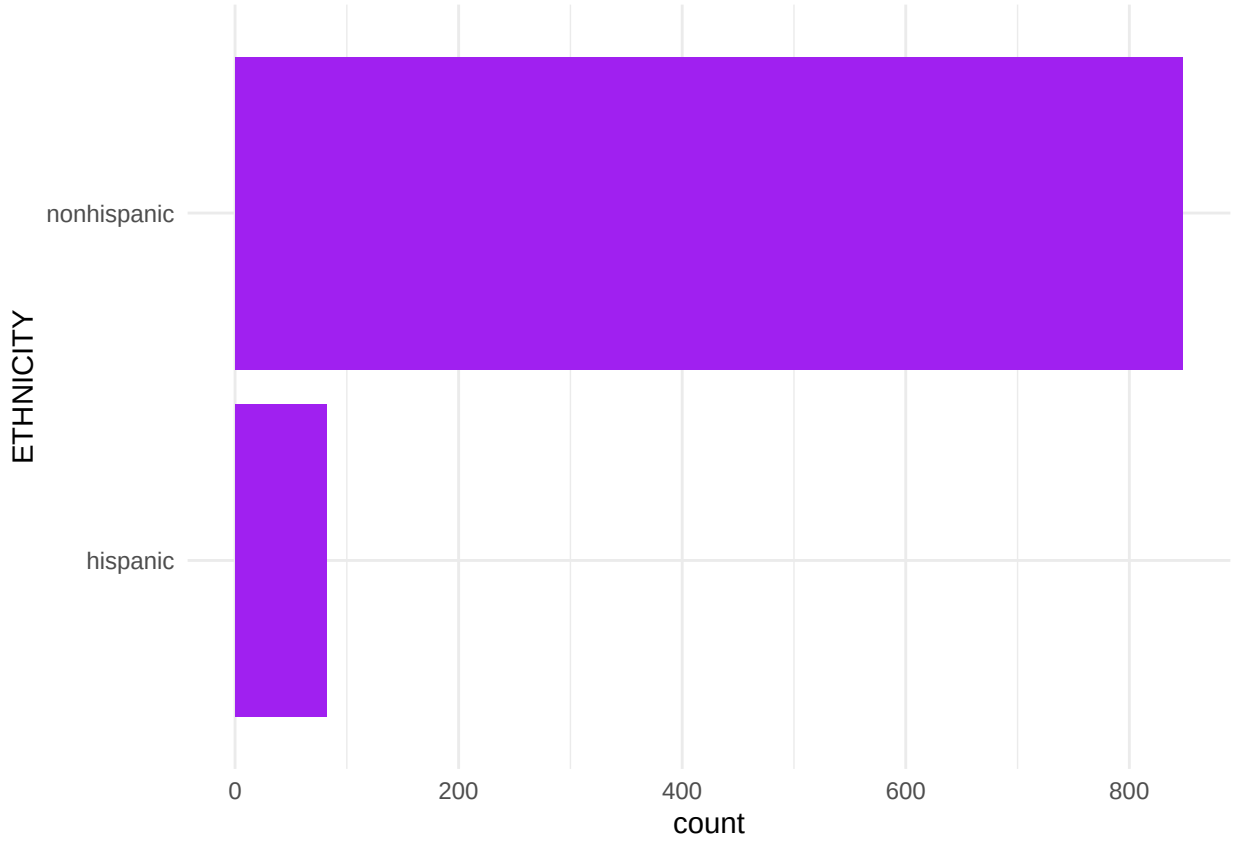
```r
# Gender distribution
ggplot(patients, aes(x = GENDER)) + geom_bar(fill = "steelblue") + theme_minimal()
```



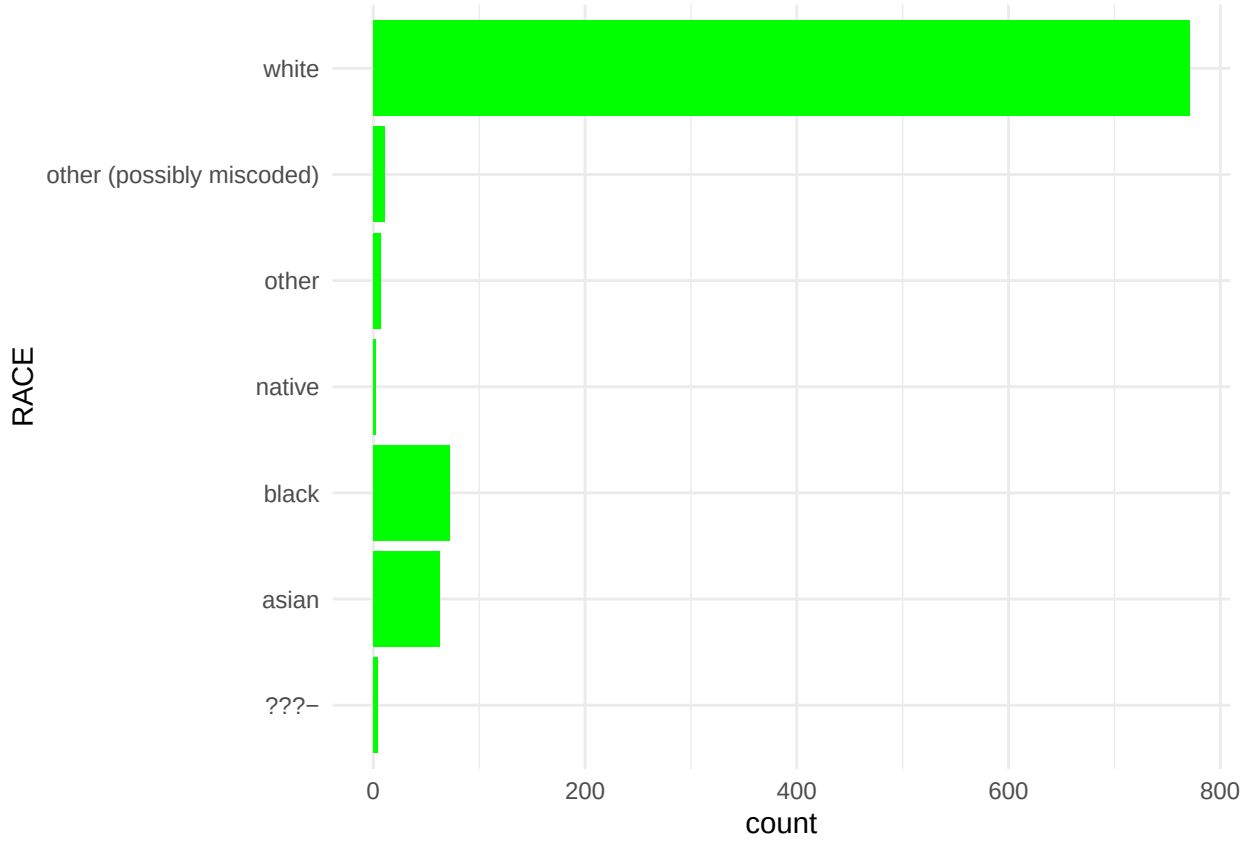

```
# Age distribution
ggplot(patients, aes(x = AGE)) + geom_histogram(binwidth = 5, fill = "darkorange") + theme_minimal()
```




```r
# Ethnicity and Race
ggplot(patients, aes(x = ETHNICITY)) + geom_bar(fill = "purple") + coord_flip() + theme_minimal()
```




```
ggplot(patients, aes(x = RACE)) + geom_bar(fill = "green") + coord_flip() + theme_minimal()
```



Top Medications, Conditions, and Labs

```
top_meds <- medications %>%
  group_by(PATIENT, DESCRIPTION) %>% summarise(n = n()) %>%
  group_by(DESCRIPTION) %>% summarise(total = sum(n)) %>% arrange(desc(total)) %>% head(10)
```

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

```
print(top_meds)
```

```
## # A tibble: 10 x 2
##    DESCRIPTION                                                      total
##    <chr>                                                            <int>
##  1 Hydrochlorothiazide 25 MG Oral Tablet                             5944
##  2 Simvastatin 10 MG Oral Tablet                                     1639
##  3 1 ML Epoetin Alfa 4000 UNT/ML Injection [Epogen]                  1015
##  4 insulin human  isophane 70 UNT/ML / Regular Insulin  Human 30 UNT/ML I~   992
##  5 lisinopril 10 MG Oral Tablet                                       915
##  6 24 HR Metformin hydrochloride 500 MG Extended Release Oral Tablet  510
##  7 Acetaminophen 325 MG Oral Tablet                                   480
##  8 Simvastatin 20 MG Oral Tablet                                      455
##  9 Amlodipine 5 MG Oral Tablet                                        415
## 10 amLODIPine 2.5 MG Oral Tablet                                      406
```

```r
conditions_summary <- conditions %>%
  group_by(DESCRIPTION) %>% summarise(n = n()) %>% arrange(desc(n))
top_conditions <- head(conditions_summary, 10)
least_conditions <- tail(conditions_summary, 10)

print(top_conditions)
```

```
## # A tibble: 10 x 2
##    DESCRIPTION                                  n
##    <chr>                                    <int>
##  1 Full-time employment (finding)           11051
##  2 Stress (finding)                          4130
##  3 Part-time employment (finding)            2047
##  4 Social isolation (finding)                1067
##  5 Limited social contact (finding)          1030
##  6 Viral sinusitis (disorder)                1025
##  7 Not in labor force (finding)               900
##  8 Victim of intimate partner abuse (finding)  695
##  9 Acute viral pharyngitis (disorder)          554
## 10 Normal pregnancy                            482
```

```r
print(least_conditions)
```

```
## # A tibble: 10 x 2
##    DESCRIPTION                                                          n
##    <chr>                                                            <int>
##  1 Infection caused by Pseudomonas aeruginosa                           1
##  2 Injury of heart (disorder)                                           1
##  3 Injury of kidney (disorder)                                          1
##  4 Macular edema and retinopathy due to type 2 diabetes mellitus (disorde~     1
##  5 Major depression disorder                                            1
##  6 Male Infertility                                                     1
##  7 Microalbuminuria due to type 2 diabetes mellitus (disorder)          1
##  8 Pyelonephritis                                                       1
##  9 Spina bifida occulta (disorder)                                      1
## 10 Tear of meniscus of knee                                             1
```

```r
top_labs <- observations %>%
  group_by(DESCRIPTION) %>% summarise(n = n()) %>% arrange(desc(n)) %>% head(10)
print(top_labs)
```

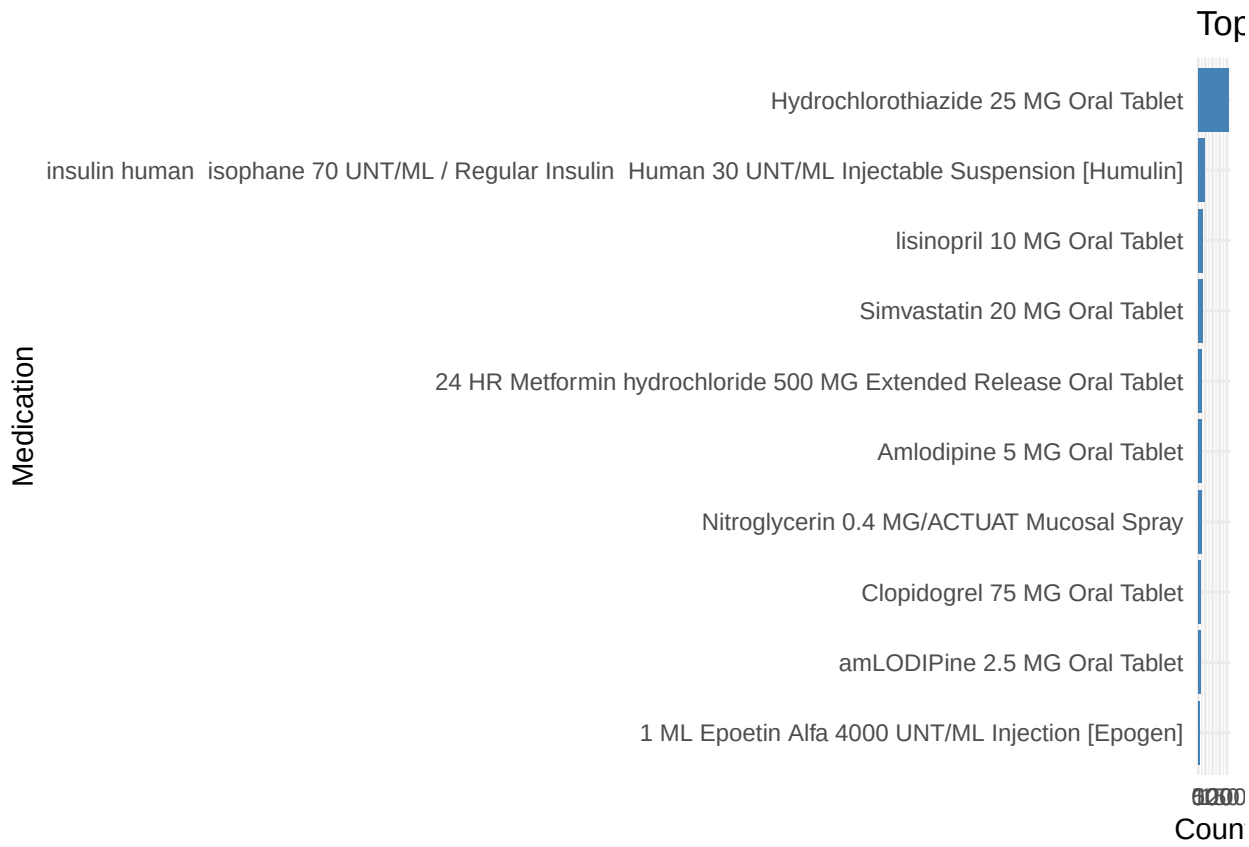
```
## # A tibble: 10 x 2
##    DESCRIPTION                                                    n
##    <chr>                                                      <int>
##  1 Diastolic Blood Pressure                                   11386
##  2 Systolic Blood Pressure                                    11386
##  3 Pain severity - 0-10 verbal numeric rating [Score] - Reported 11372
##  4 Body Weight                                                10114
##  5 Heart rate                                                  9948
##  6 Respiratory rate                                            9948
##  7 Body Height                                                 9648
##  8 Tobacco smoking status NHIS                                 9616
```

```
##  9 Body Mass Index                                                9109
## 10 Housing status                                                 7584
```

Medications by Socioeconomic Group

Medications for Unemployed Patients

```r
# Step 1: Find encounters where unemployment is recorded
unemployed_encounters <- conditions %>%
  filter(grepl("unemployed|not in labor force", DESCRIPTION, ignore.case = TRUE)) %>%
  distinct(PATIENT, ENCOUNTER)

# Step 2: Join to medications
medications_unemployed <- medications %>%
  semi_join(unemployed_encounters, by = c("PATIENT", "ENCOUNTER"))

# Step 3: Plot top medications
medications_unemployed %>%
  count(DESCRIPTION, sort = TRUE) %>%
  slice_max(n, n = 10) %>%
  ggplot(aes(x = reorder(DESCRIPTION, n), y = n)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Top Medications Among Unemployed Patients",
    x = "Medication", y = "Count"
  ) +
  theme_minimal()
```



Conditions with Highest Out-of-Pocket Costs

```
cond_cost <- merge(conditions, encounters[, c("Id", "TOTAL_CLAIM_COST", "PAYER_COVERAGE")],
                   by.x = "ENCOUNTER", by.y = "Id") %>%
  group_by(DESCRIPTION) %>%
  summarise(
    avg_claim_cost = mean(TOTAL_CLAIM_COST, na.rm = TRUE),
    avg_patient_pay = mean(TOTAL_CLAIM_COST - PAYER_COVERAGE, na.rm = TRUE)) %>%
  arrange(desc(avg_patient_pay))
```

Reasons for Visit

```
reasons <- encounters %>%
  filter(!is.na(REASONCODE)) %>%
  count(REASONCODE, sort = TRUE) %>%
  left_join(dict_snomed, by = c("REASONCODE" = "CODE")) %>%
  rename(ReasonDescription = DESCRIPTION)
print(head(reasons, 20))
```

```
## # A tibble: 20 x 3
##    REASONCODE     n ReasonDescription
```

```
##         <dbl> <int> <chr>
##  1   72892002  3585 Normal pregnancy
##  2   55822004  1640 Hyperlipidemia
##  3   88805009  1536 Chronic congestive heart failure (disorder)
##  4  444814009  1210 Viral sinusitis (disorder)
##  5   10509002   633 Acute bronchitis (disorder)
##  6  195662009   632 Acute viral pharyngitis (disorder)
##  7  254837009   380 Malignant neoplasm of breast (disorder)
##  8  271737000   284 Anemia (disorder)
##  9  192127007   280 Child attention deficit disorder
## 10   59621000   242 Hypertension
## 11   75498004   213 Acute bacterial sinusitis (disorder)
## 12  195967001   193 Asthma
## 13   36971009   176 Sinusitis (disorder)
## 14   55680006   145 Drug overdose
## 15   43878008   131 Streptococcal sore throat (disorder)
## 16   65363002   122 Otitis media
## 17  233678006   115 Childhood asthma
## 18   74400008   106 Appendicitis
## 19   82423001    99 Chronic pain
## 20  196416002    89 Impacted molars
```

Summaries

```
summary(patients)
```

```
##       Id              BIRTHDATE            DEATHDATE
##  Length:930         Min.   :1924-12-04   Min.   :1945-10-20
##  Class :character   1st Qu.:1960-03-30   1st Qu.:1997-01-05
##  Mode  :character   Median :1975-11-20   Median :2008-07-02
##                     Mean   :1976-06-17   Mean   :2005-04-22
##                     3rd Qu.:1994-07-28   3rd Qu.:2015-05-22
##                     Max.   :2021-09-05   Max.   :2021-09-29
##                                          NA's   :833
##      SSN              DRIVERS            PASSPORT           PREFIX
##  Length:930         Length:930         Length:930         Length:930
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     FIRST               LAST              SUFFIX             MAIDEN
##  Length:930         Length:930         Length:930         Length:930
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    MARITAL              RACE            ETHNICITY           GENDER
##  Length:930         Length:930         Length:930         Length:930
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   BIRTHPLACE          ADDRESS              CITY              STATE
##  Length:930         Length:930         Length:930         Length:930
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     COUNTY              ZIP                 LAT             LON
##  Length:930         Length:930         Min.   :41.50   Min.   :-73.45
##  Class :character   Class :character   1st Qu.:42.10   1st Qu.:-71.43
##  Mode  :character   Mode  :character   Median :42.31   Median :-71.13
##                                        Mean   :42.25   Mean   :-71.30
##                                        3rd Qu.:42.45   3rd Qu.:-70.99
##                                        Max.   :42.89   Max.   :-69.98
##
##  HEALTHCARE_EXPENSES HEALTHCARE_COVERAGE      AGE         FLAG_INVALID_EXPENSE
##  Min.   :     826    Min.   :      0     Min.   :  3.00   Mode :logical
##  1st Qu.:  589426    1st Qu.:   1675     1st Qu.: 30.25   FALSE:930
##  Median : 1248438    Median :  22745     Median : 49.00
##  Mean   : 1305014    Mean   : 158385     Mean   : 48.58
##  3rd Qu.: 1914155    3rd Qu.: 214946     3rd Qu.: 65.00
##  Max.   :12641789    Max.   :2411214     Max.   :100.00
##
##  FLAG_INVALID_COVERAGE
##  Mode :logical
##  FALSE:900
##  TRUE :30
##
##
##
##
```

```
summary(encounters)
```

```
##       Id                 START                         STOP
##  Length:49339       Min.   :1925-10-01 21:07:04   Min.   :1925-10-01 21:22:04
##  Class :character   1st Qu.:2001-10-31 20:33:41   1st Qu.:2001-11-01 06:28:47
##  Mode  :character   Median :2013-07-16 01:52:59   Median :2013-07-16 02:35:10
##                     Mean   :2007-08-27 23:09:14   Mean   :2007-08-28 04:39:02
##                     3rd Qu.:2017-09-04 05:48:47   3rd Qu.:2017-09-04 06:03:47
##                     Max.   :2021-11-19 16:50:22   Max.   :2021-11-19 17:05:22
##
##    PATIENT           ORGANIZATION         PROVIDER            PAYER
##  Length:49339       Length:49339       Length:49339       Length:49339
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##  ENCOUNTERCLASS          CODE           BASE_ENCOUNTER_COST TOTAL_CLAIM_COST
##  Length:49339       Min.   :  1505002   Min.   : 77.49      Min.   :     0.0
##  Class :character   1st Qu.:162673000   1st Qu.: 77.49      1st Qu.:   129.2
##  Mode  :character   Median :185347001   Median :129.16      Median :   786.3
##                     Mean   :260883600   Mean   :113.68      Mean   :  3976.6
##                     3rd Qu.:390906007   3rd Qu.:129.16      3rd Qu.:  1615.7
##                     Max.   :702927004   Max.   :129.16      Max.   :873646.2
##
##  PAYER_COVERAGE        REASONCODE
##  Min.   :     0.00   Min.   :1.734e+06
##  1st Qu.:     0.00   1st Qu.:6.256e+07
##  Median :     0.00   Median :7.289e+07
##  Mean   :   866.70   Mean   :5.840e+12
##  3rd Qu.:    20.14   3rd Qu.:1.957e+08
##  Max.   :227851.81   Max.   :1.094e+16
##                      NA's   :36206
```

```
summary(medications)
```

```
##      START                         STOP                          PATIENT
##  Min.   :1931-03-09 03:45:19   Min.   :1931-03-27 03:45:19   Length:16967
##  1st Qu.:1997-11-03 03:45:08   1st Qu.:1998-02-07 07:37:02   Class :character
##  Median :2008-12-06 23:41:21   Median :2009-01-20 11:58:49   Mode  :character
##  Mean   :2005-06-23 23:21:23   Mean   :2005-07-30 17:20:11
##  3rd Qu.:2016-06-08 03:55:18   3rd Qu.:2016-06-19 21:56:31
##  Max.   :2021-11-18 14:01:22   Max.   :2021-11-18 14:01:22
##                                NA's   :494
##     PAYER             ENCOUNTER             CODE           BASE_COST
##  Length:16967       Length:16967       Min.   : 106258   Min.   :   0.01
##  Class :character   Class :character   1st Qu.: 310798   1st Qu.:   0.01
##  Mode  :character   Mode  :character   Median : 310798   Median :   0.02
##                                        Mean   : 427546   Mean   : 383.79
##                                        3rd Qu.: 314231   3rd Qu.:  60.56
##                                        Max.   :2123111   Max.   :6994.54
##
##  PAYER_COVERAGE        DISPENSES          TOTALCOST         REASONCODE
##  Min.   :0.0000000   Min.   :    1.00   Min.   :   0.01   Min.   : 10509002
##  1st Qu.:0.0000000   1st Qu.:    1.00   1st Qu.:   0.12   1st Qu.: 55822004
##  Median :0.0000000   Median :    1.00   Median :   0.56   Median : 59621000
##  Mean   :0.0001532   Mean   :   12.09   Mean   : 383.88   Mean   :111906137
##  3rd Qu.:0.0000000   3rd Qu.:   12.00   3rd Qu.:  60.56   3rd Qu.: 59621000
##  Max.   :0.6200000   Max.   :45000.00   Max.   :6994.54   Max.   :706870000
##                                                           NA's   :2852
##  DESCRIPTION          DEATHDATE
##  Length:16967       Min.   :1945-10-20
##  Class :character   1st Qu.:2008-11-15
##  Mode  :character   Median :2008-11-15
##                     Mean   :2009-07-03
##                     3rd Qu.:2012-10-12
##                     Max.   :2021-09-29
##                     NA's   :12753
```

```r
summary(conditions)
```

```
##      START                STOP              PATIENT
##  Min.   :1930-08-05   Min.   :1930-10-21   Length:31334
##  1st Qu.:1996-04-30   1st Qu.:2000-05-09   Class :character
##  Median :2011-07-23   Median :2013-06-20   Mode  :character
##  Mean   :2005-01-29   Mean   :2007-07-07
##  3rd Qu.:2017-02-08   3rd Qu.:2018-03-07
##  Max.   :2021-11-15   Max.   :2021-11-18
##                       NA's   :7014
##   ENCOUNTER              CODE           DESCRIPTION        factor_education
##  Length:31334       Min.   :1.734e+06   Length:31334       Min.   :0.00000
##  Class :character   1st Qu.:9.130e+07   Class :character   1st Qu.:0.00000
##  Mode  :character   Median :1.609e+08   Mode  :character   Median :0.00000
##                     Mean   :5.463e+13                      Mean   :0.02582
##                     3rd Qu.:2.376e+08                      3rd Qu.:0.00000
##                     Max.   :1.094e+16                      Max.   :1.00000
##
##  factor_employment factor_stress    factor_refugee   factor_criminal
##  Min.   :0.0000    Min.   :0.0000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.0000    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.0000    Median :0.0000   Median :0.0000   Median :0.000000
##  Mean   :0.4488    Mean   :0.1356   Mean   :0.0015   Mean   :0.005234
##  3rd Qu.:1.0000    3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.000000
##  Max.   :1.0000    Max.   :1.0000   Max.   :1.0000   Max.   :1.000000
##
##  factor_violence   factor_transport   factor_drugs_alcohol
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.00000   Median :0.000000   Median :0.000000
##  Mean   :0.03763   Mean   :0.006159   Mean   :0.007245
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.000000
##
##  factor_social_isolation factor_military    factor_housing
##  Min.   :0.00000         Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.00000         1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.00000         Median :0.000000   Median :0.000000
##  Mean   :0.06692         Mean   :0.001213   Mean   :0.004564
##  3rd Qu.:0.00000         3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.00000         Max.   :1.000000   Max.   :1.000000
##
```

```r
summary(observations)
```

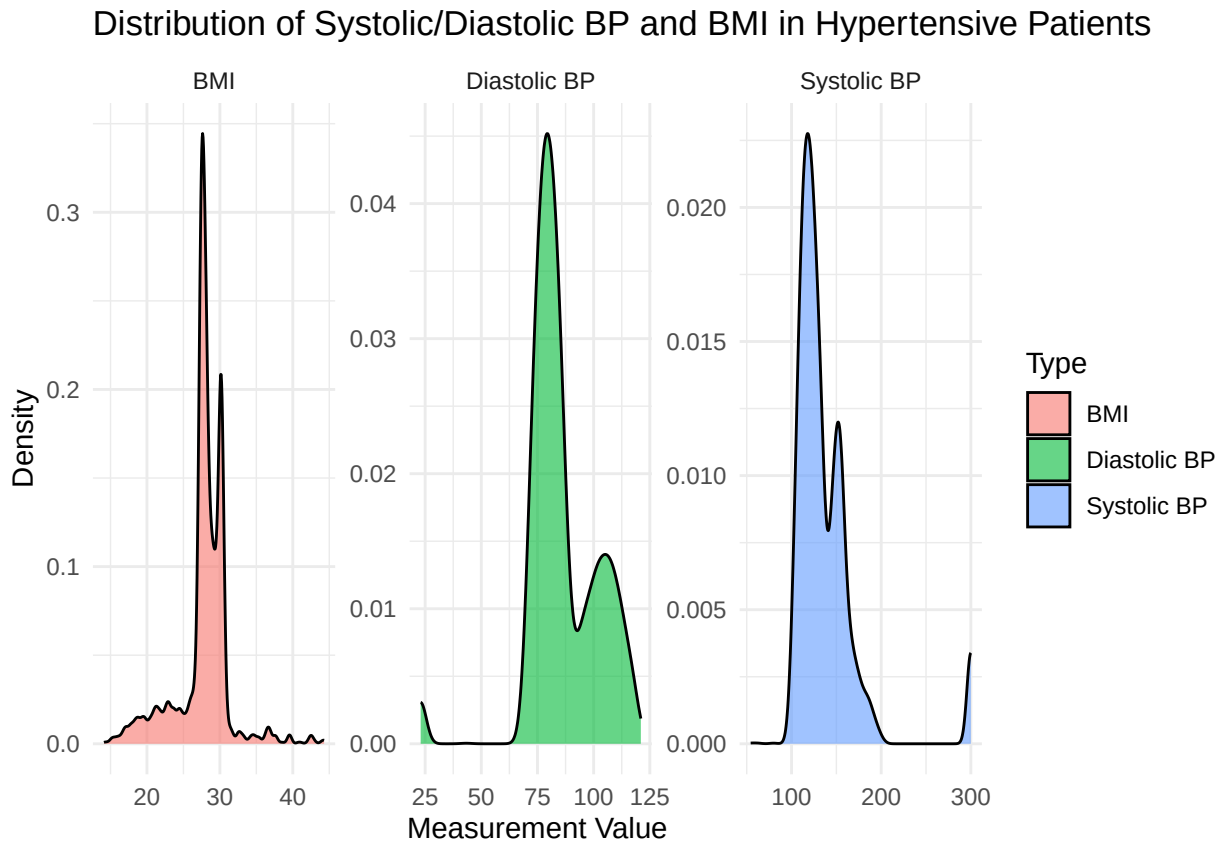
```
##       DATE                        PATIENT           ENCOUNTER
##  Min.   :1935-11-30 10:04:43   Length:438591      Length:438591
##  1st Qu.:2013-07-18 12:07:29   Class :character   Class :character
##  Median :2016-07-31 18:42:32   Mode  :character   Mode  :character
##  Mean   :2015-01-30 02:48:47
##  3rd Qu.:2019-07-23 08:20:50
##  Max.   :2021-11-18 16:26:22
```

```
##
##    CATEGORY             CODE              VALUE              UNITS
##  Length:438591      Length:438591      Length:438591      Length:438591
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      TYPE            DESCRIPTION        VALUE_NUMERIC        DEATHDATE
##  Length:438591      Length:438591      Min.   :    -3.7   Min.   :1945-10-20
##  Class :character   Class :character   1st Qu.:     5.3   1st Qu.:2005-11-09
##  Mode  :character   Mode  :character   Median :    29.2   Median :2011-07-10
##                                        Mean   :  3086.7   Mean   :2009-04-11
##                                        3rd Qu.:    92.0   3rd Qu.:2017-04-06
##                                        Max.   :957744.0   Max.   :2021-09-29
##                                        NA's   :168780     NA's   :356492
```

Exploring and comparing the distribution of: systolic and diastolic blood pressure and BMI measurements in patients with diagnosed hypertension

```r
# 1. Get hypertensive patients
hypertensive_patients <- conditions %>%
  filter(grepl("hypertension", DESCRIPTION, ignore.case = TRUE)) %>%
  distinct(PATIENT)

# 2. Extract their blood pressure and BMI observations
bp_bmi_data <- observations %>%
  semi_join(hypertensive_patients, by = "PATIENT") %>%
  filter(CODE %in% c("55284-4", "8480-6", "8462-4", "39156-5")) %>%  # systolic, diastolic, BMI
  mutate(
    Type = case_when(
      CODE == "8480-6" ~ "Systolic BP",
      CODE == "8462-4" ~ "Diastolic BP",
      CODE == "39156-5" ~ "BMI",
      CODE == "55284-4" ~ "Blood Pressure Panel",
      TRUE ~ "Other"
    ),
    VALUE = as.numeric(VALUE)
  ) %>%
  filter(!is.na(VALUE))  # remove any non-numeric or NA values

# 3. Plot distributions
bp_bmi_data %>%
  filter(Type %in% c("Systolic BP", "Diastolic BP", "BMI")) %>%
  ggplot(aes(x = VALUE, fill = Type)) +
  geom_density(alpha = 0.6) +
  facet_wrap(~Type, scales = "free") +
  labs(
    title = "Distribution of Systolic/Diastolic BP and BMI in Hypertensive Patients",
    x = "Measurement Value",
    y = "Density"
```

```
  ) +
  theme_minimal()
```



The crude, and adjusted (to the UK population as much as possible) prevalence of hypertension

```
# Step 1: Identify hypertension codes from SNOMED dictionary
htn_codes <- dict_snomed %>%
  filter(grepl("hypertension", tolower(DESCRIPTION))) %>%
  pull(CODE)

# Step 2: Mark hypertensive patients
htn_patients <- conditions %>%
  filter(CODE %in% htn_codes) %>%
  distinct(PATIENT)

# Step 3: Mark all patients as hypertensive or not
patients <- patients %>%
  mutate(HYPERTENSION = ifelse(Id %in% htn_patients$PATIENT, 1, 0))

# Step 4: Crude prevalence
crude_prev <- mean(patients$HYPERTENSION) * 100
cat("Crude Prevalence of Hypertension: ", round(crude_prev, 2), "%\n")
```

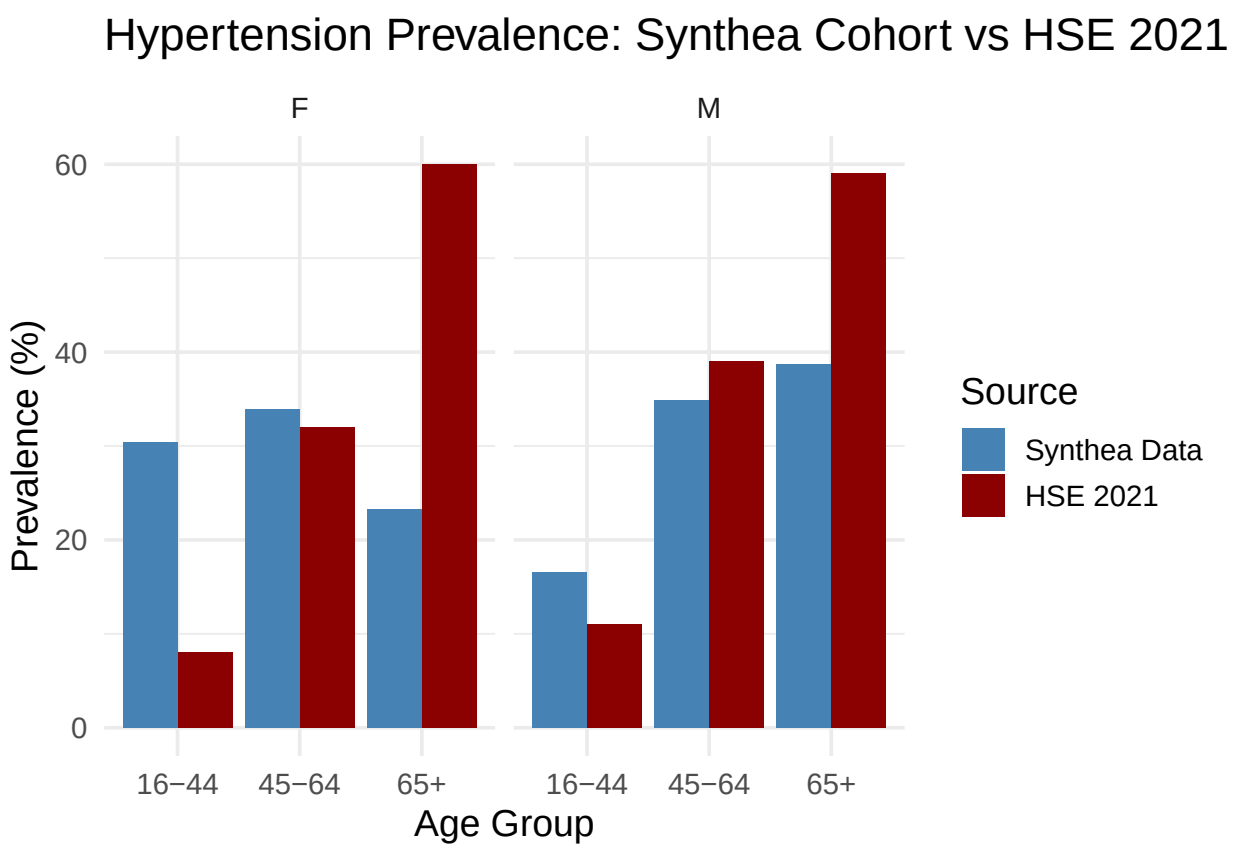
```
## Crude Prevalence of Hypertension:  28.17 %
```

```
library(dplyr)
library(lubridate)
library(ggplot2)
library(tidyr)

# Step 1: Age and age group
patients <- patients %>%
  mutate(BIRTHDATE = as.Date(BIRTHDATE),
         AGE = as.numeric(floor(interval(BIRTHDATE, Sys.Date()) / years(1))),
         age_group = case_when(
           AGE < 45 ~ "16-44",
           AGE < 65 ~ "45-64",
           TRUE ~ "65+"
         ))

# Step 2: UK HSE 2021 reference prevalence (from Table 12)
hse_prevalence <- tribble(
  ~age_group, ~GENDER, ~hse_prev,
  "16-44", "M", 0.11,
  "45-64", "M", 0.39,
  "65+",   "M", 0.59,
  "16-44", "F", 0.08,
  "45-64", "F", 0.32,
  "65+",   "F", 0.60
)

# Step 3: Synthea cohort's prevalence by age group and gender
cohort_prev <- patients %>%
  filter(GENDER %in% c("M", "F")) %>%
  group_by(age_group, GENDER) %>%
  summarise(cohort_prev = mean(HYPERTENSION), n = n(), .groups = "drop")

# Step 4: Join for comparison
comparison <- left_join(cohort_prev, hse_prevalence, by = c("age_group", "GENDER"))

# View comparison table
print(comparison)
```

```
## # A tibble: 6 x 5
##   age_group GENDER cohort_prev     n hse_prev
##   <chr>     <chr>        <dbl> <int>    <dbl>
## 1 16-44     F            0.304   158     0.08
## 2 16-44     M            0.166   247     0.11
## 3 45-64     F            0.340   159     0.32
## 4 45-64     M            0.349   126     0.39
## 5 65+       F            0.233   116     0.6
## 6 65+       M            0.387   124     0.59
```

```
# Step 5: Reshape for plotting
plot_data <- comparison %>%
  pivot_longer(cols = c("cohort_prev", "hse_prev"),
```

```
              names_to = "source", values_to = "prevalence")

# Step 6: Plot
ggplot(plot_data, aes(x = age_group, y = prevalence * 100, fill = source)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~GENDER) +
  scale_fill_manual(values = c("cohort_prev" = "steelblue", "hse_prev" = "darkred"),
                    labels = c("Synthea Data", "HSE 2021")) +
  labs(title = "Hypertension Prevalence: Synthea Cohort vs HSE 2021",
       x = "Age Group", y = "Prevalence (%)", fill = "Source") +
  theme_minimal(base_size = 14)
```



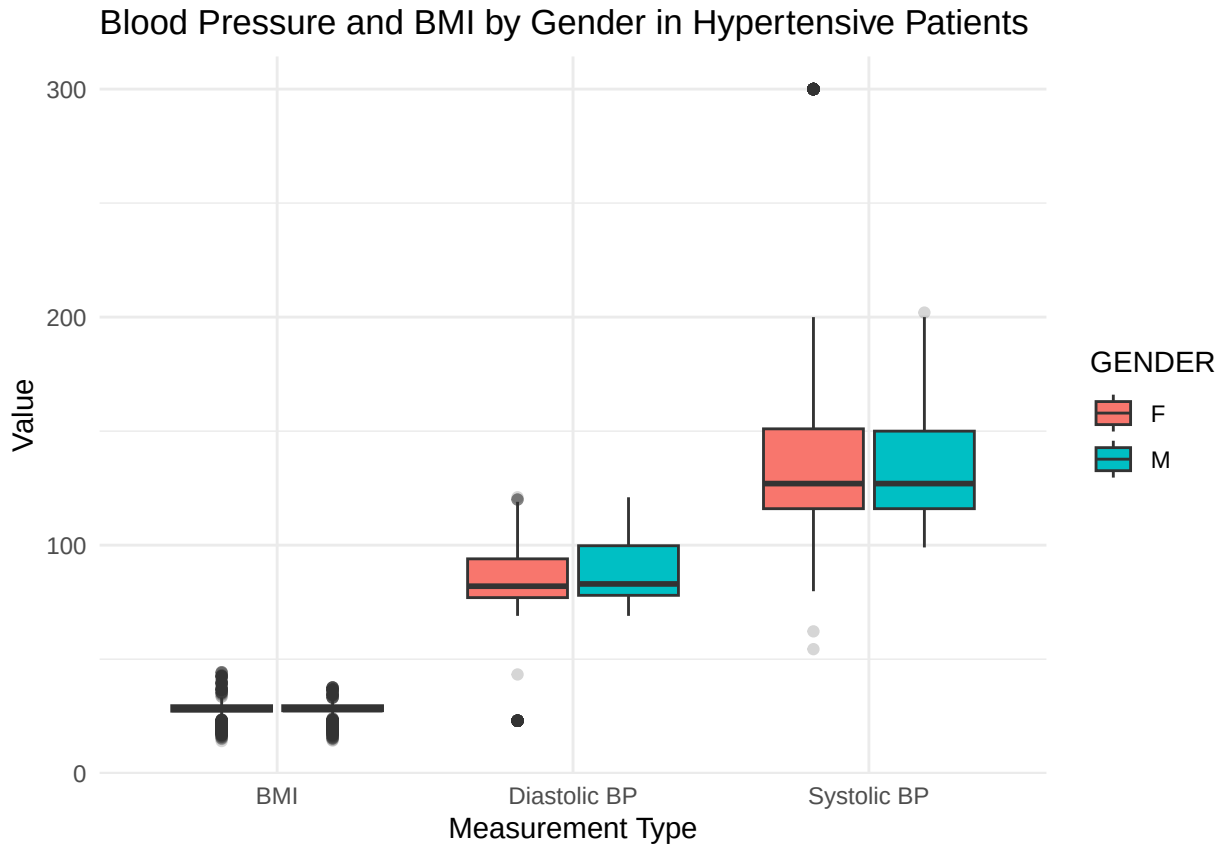
Further Analysis: Stratified Summary and Plots

```
bp_bmi_summary <- bp_bmi_data %>%
  group_by(Type) %>%
  summarise(
    count = n(),
    mean = round(mean(VALUE_NUMERIC, na.rm = TRUE), 1),
    median = round(median(VALUE_NUMERIC, na.rm = TRUE), 1),
    sd = round(sd(VALUE_NUMERIC, na.rm = TRUE), 1),
    .groups = 'drop'
  )
```

```
bp_bmi_summary
```

```
## # A tibble: 3 x 5
##   Type         count  mean median    sd
##   <chr>        <int> <dbl>  <dbl> <dbl>
## 1 BMI           3355  27.7   27.9   3.6
## 2 Diastolic BP  4026  86.1   83    15.7
## 3 Systolic BP   4026 138.   127    37.8
```

```
bp_bmi_data <- bp_bmi_data %>%
  left_join(patients %>% select(Id, GENDER), by = c("PATIENT" = "Id"))

ggplot(bp_bmi_data, aes(x = Type, y = VALUE_NUMERIC, fill = GENDER)) +
  geom_boxplot(outlier.alpha = 0.2) +
  labs(
    title = "Blood Pressure and BMI by Gender in Hypertensive Patients",
    x = "Measurement Type", y = "Value"
  ) +
  theme_minimal()
```




```
bp_bmi_data <- bp_bmi_data %>%
  left_join(patients %>% select(Id, AGE), by = c("PATIENT" = "Id")) %>%
  mutate(AgeGroup = cut(
    AGE,
```

```
    breaks = c(0, 30, 50, 70, 120),
    labels = c(" 30", "31-50", "51-70", "70+")
  ))

ggplot(bp_bmi_data, aes(x = Type, y = VALUE_NUMERIC, fill = AgeGroup)) +
  geom_boxplot(outlier.alpha = 0.2) +
  labs(
    title = "BP and BMI by Age Group in Hypertensive Patients",
    x = "Measurement Type", y = "Value"
  ) +
  theme_minimal()
```