Health_Records_Project_Nicholas_Pinero

Nicholas Pinero

2025-02-04

Introduction

This project analyzes simulated hospital health records using R and tidyverse.
The goal is to demonstrate a full data science workflow including: - Data cleaning and preprocessing - Feature engineering - Dataset merging - Exploratory data analysis - Healthcare insights and recommendations

Datasets:

```
library(tidyverse)
```

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

```
medications <- read_csv("medications.csv")
```

```
## Rows: 5 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): condition, recommended_medication
## dbl (1): dosage_mg
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
messy_health_records <- read_csv("messy_health_records.csv")
```

```
## Rows: 8670 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): condition, blood_pressure, medication_given, hospital_location
```

```
## dbl  (4): record_id, patient_id, cholesterol_level, patient_age
## date (1): visit_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
patients <- read_csv("patients.csv")
```

```
## Rows: 3513 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): patient_name, gender, insurance_status
## dbl  (1): patient_id
## date (1): admission_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Cleaning and Preprocessing

1. Check for missing values:

```
messy_health_records %>%
  summarise(
    record_id = sum(is.na(record_id)),
    patient_id = sum(is.na(patient_id)),
    visit_date = sum(is.na(visit_date)),
    condition = sum(is.na(condition)),
    blood_pressure = sum(is.na(blood_pressure)),
    cholesterol_level = sum(is.na(cholesterol_level)),
    medication_given = sum(is.na(medication_given)),
    hospital_location = sum(is.na(hospital_location)),
    patient_age = sum(is.na(patient_age))
    )
```

```
## # A tibble: 1 x 9
##   record_id patient_id visit_date condition blood_pressure cholesterol_level
##       <int>      <int>      <int>     <int>          <int>             <int>
## 1         0          0          0         0            437               439
## # i 3 more variables: medication_given <int>, hospital_location <int>,
## #   patient_age <int>
```

The above code checks if there are missing values in each column and outputs the number of missing values in each column.

2. Handle missing values:

```
cleaned_health_records <- messy_health_records %>%
  drop_na()

cleaned_health_records
```

```
## # A tibble: 7,431 x 9
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                      <dbl>
##  1      7908       4911 2020-11-25 Obesity      Elevated                    128.
##  2      2115       1530 2020-03-29 Hypertension Normal                      155.
##  3      1545       3840 2020-03-05 Obesity      Stage 1 Hyper~              237.
##  4      7028       1178 2020-10-19 Diabetes     Elevated                    178
##  5      6493       1893 2020-09-27 Hypertension Normal                      247.
##  6      5216       4292 2020-08-05 Diabetes     Stage 1 Hyper~              151.
##  7      6388       4426 2020-09-23 Heart Disea~ Stage 2 Hyper~              193.
##  8       412       3595 2020-01-18 Obesity      Elevated                    220.
##  9      4202       1062 2020-06-24 Asthma       Elevated                    220.
## 10      7500       1336 2020-11-08 Diabetes     Normal                      126.
## # i 7,421 more rows
## # i 3 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>
```

To handle missing values, I removed rows containing NAs. While imputation could have been considered, dropping missing values was the most straightforward choice given the dataset's size and structure.

3. Convert incorrect data types:

```
cleaned_health_records <- cleaned_health_records %>%
  mutate(visit_date = ymd(visit_date))

cleaned_health_records
```

```
## # A tibble: 7,431 x 9
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                      <dbl>
##  1      7908       4911 2020-11-25 Obesity      Elevated                    128.
##  2      2115       1530 2020-03-29 Hypertension Normal                      155.
##  3      1545       3840 2020-03-05 Obesity      Stage 1 Hyper~              237.
##  4      7028       1178 2020-10-19 Diabetes     Elevated                    178
##  5      6493       1893 2020-09-27 Hypertension Normal                      247.
##  6      5216       4292 2020-08-05 Diabetes     Stage 1 Hyper~              151.
##  7      6388       4426 2020-09-23 Heart Disea~ Stage 2 Hyper~              193.
##  8       412       3595 2020-01-18 Obesity      Elevated                    220.
##  9      4202       1062 2020-06-24 Asthma       Elevated                    220.
## 10      7500       1336 2020-11-08 Diabetes     Normal                      126.
## # i 7,421 more rows
## # i 3 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>
```

I mutated the visit_date column of my cleaned_health_records into a date type by using the ymd function which was used in this case because the data was already in a year-month-day format.

4. Remove duplicate records

```
removed_duplicates <- nrow(cleaned_health_records) - nrow(cleaned_health_records %>% unique())
removed_duplicates
```

```
## [1] 140
```

```
cleaned_health_records <- cleaned_health_records %>% unique()
cleaned_health_records
```

```
## # A tibble: 7,291 x 9
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                      <dbl>
##  1      7908       4911 2020-11-25 Obesity      Elevated                    128.
##  2      2115       1530 2020-03-29 Hypertension Normal                      155.
##  3      1545       3840 2020-03-05 Obesity      Stage 1 Hyper~              237.
##  4      7028       1178 2020-10-19 Diabetes     Elevated                    178
##  5      6493       1893 2020-09-27 Hypertension Normal                      247.
##  6      5216       4292 2020-08-05 Diabetes     Stage 1 Hyper~              151.
##  7      6388       4426 2020-09-23 Heart Disea~ Stage 2 Hyper~              193.
##  8       412       3595 2020-01-18 Obesity      Elevated                    220.
##  9      4202       1062 2020-06-24 Asthma       Elevated                    220.
## 10      7500       1336 2020-11-08 Diabetes     Normal                      126.
## # i 7,281 more rows
## # i 3 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>
```

To remove duplicate rows I used the unique() function in R. I calculated the number of removed duplicate rows by subtracting the numbers of rows in the original dataset from the number of rows in the non-duplicate dataset. Ultimately there were 140 dupliacte rows that were removed.

5. Detect and handle outliers:

```
IQR_cholesterol <- IQR(cleaned_health_records$cholesterol_level, na.rm = TRUE)

Q1_cholesterol <- quantile(cleaned_health_records$cholesterol_level, probs = 0.25, na.rm = TRUE)

Q3_cholesterol <- quantile(cleaned_health_records$cholesterol_level, probs = 0.75, na.rm = TRUE)

lower_bound_cholesterol <- Q1_cholesterol - 1.5 * IQR_cholesterol

upper_bound_cholesterol <- Q3_cholesterol + 1.5 * IQR_cholesterol


IQR_patient_age <- IQR(cleaned_health_records$patient_age, na.rm = TRUE)

Q1_patient_age <- quantile(cleaned_health_records$patient_age, probs = 0.25, na.rm = TRUE)

Q3_patient_age <- quantile(cleaned_health_records$patient_age, probs = 0.75, na.rm = TRUE)
```

```r
lower_bound_patient_age <- Q1_patient_age - 1.5 * IQR_patient_age

upper_bound_patient_age <- Q3_patient_age + 1.5 * IQR_patient_age

cleaned_health_records <- cleaned_health_records %>%
  filter(cholesterol_level >= lower_bound_cholesterol & cholesterol_level <= upper_bound_cholesterol) %
  filter(patient_age >= lower_bound_patient_age & patient_age <= upper_bound_patient_age)

cleaned_health_records
```

```
## # A tibble: 7,291 x 9
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                      <dbl>
##  1      7908       4911 2020-11-25 Obesity      Elevated                    128.
##  2      2115       1530 2020-03-29 Hypertension Normal                      155.
##  3      1545       3840 2020-03-05 Obesity      Stage 1 Hyper~              237.
##  4      7028       1178 2020-10-19 Diabetes     Elevated                    178
##  5      6493       1893 2020-09-27 Hypertension Normal                      247.
##  6      5216       4292 2020-08-05 Diabetes     Stage 1 Hyper~              151.
##  7      6388       4426 2020-09-23 Heart Disea~ Stage 2 Hyper~              193.
##  8       412       3595 2020-01-18 Obesity      Elevated                    220.
##  9      4202       1062 2020-06-24 Asthma       Elevated                    220.
## 10      7500       1336 2020-11-08 Diabetes     Normal                      126.
## # i 7,281 more rows
## # i 3 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>
```

I used the IQR method to remove outliers. This was done by finding the IQR, Q1, and Q3 values from both the cholesterol_level and patient_age column of the messy_health_data dataset. Upon calculating these values I found the upper and lower bounds of each column using the formulas lower bound = Q1 - 1.5 * IQR and upper bound = Q3 + 1.5 * IQR. I finally removed all rows from these two columns that contained values outside of these bounds.

Data Transformation and Feature Engineering

1. Create a new column visit_month:

```r
cleaned_health_records <- cleaned_health_records %>% mutate(visit_month = month(visit_date))

cleaned_health_records
```

```
## # A tibble: 7,291 x 10
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                      <dbl>
##  1      7908       4911 2020-11-25 Obesity      Elevated                    128.
##  2      2115       1530 2020-03-29 Hypertension Normal                      155.
##  3      1545       3840 2020-03-05 Obesity      Stage 1 Hyper~              237.
##  4      7028       1178 2020-10-19 Diabetes     Elevated                    178
##  5      6493       1893 2020-09-27 Hypertension Normal                      247.
##  6      5216       4292 2020-08-05 Diabetes     Stage 1 Hyper~              151.
```

```
##  7      6388       4426 2020-09-23 Heart Disea~ Stage 2 Hyper~              193.
##  8       412       3595 2020-01-18 Obesity      Elevated                    220.
##  9      4202       1062 2020-06-24 Asthma       Elevated                    220.
## 10      7500       1336 2020-11-08 Diabetes     Normal                      126.
## # i 7,281 more rows
## # i 4 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>, visit_month <dbl>
```

I added a new column, visit_month, to the cleaned_health_records data set that extracts the month variable from each row of the visit_date column.

2. Create a categorical column patient_age_group:

```
cleaned_health_records <- cleaned_health_records %>%
  mutate(patient_age_group = case_when(
    patient_age >= 18 & patient_age <= 35 ~ "Young",
    patient_age >= 36 & patient_age <= 60 ~ "Middle-aged",
    patient_age >= 61 ~ "Senior"
    ))

cleaned_health_records
```

```
## # A tibble: 7,291 x 11
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                      <dbl>
##  1      7908       4911 2020-11-25 Obesity      Elevated                    128.
##  2      2115       1530 2020-03-29 Hypertension Normal                      155.
##  3      1545       3840 2020-03-05 Obesity      Stage 1 Hyper~              237.
##  4      7028       1178 2020-10-19 Diabetes     Elevated                    178
##  5      6493       1893 2020-09-27 Hypertension Normal                      247.
##  6      5216       4292 2020-08-05 Diabetes     Stage 1 Hyper~              151.
##  7      6388       4426 2020-09-23 Heart Disea~ Stage 2 Hyper~              193.
##  8       412       3595 2020-01-18 Obesity      Elevated                    220.
##  9      4202       1062 2020-06-24 Asthma       Elevated                    220.
## 10      7500       1336 2020-11-08 Diabetes     Normal                      126.
## # i 7,281 more rows
## # i 5 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>, visit_month <dbl>, patient_age_group <chr>
```

I added a new column with the mutate() and case_when() functions that checks the age of the patient and assigns the patient either “Young”, “Middle-aged”, or “Senior” in the new patient_age_group column.

3. Create a new column cholesterol_risk:

```
cleaned_health_records <- cleaned_health_records %>%
  mutate(cholesterol_risk = case_when(
    cholesterol_level < 160 ~ "Low",
    cholesterol_level >= 160 & cholesterol_level <= 200 ~ "Moderate",
    cholesterol_level > 200 ~ "High"
```

```
  ))

cleaned_health_records
```

```
## # A tibble: 7,291 x 12
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                       <dbl>
##  1      7908       4911 2020-11-25 Obesity      Elevated                     128.
##  2      2115       1530 2020-03-29 Hypertension Normal                       155.
##  3      1545       3840 2020-03-05 Obesity      Stage 1 Hyper~               237.
##  4      7028       1178 2020-10-19 Diabetes     Elevated                     178
##  5      6493       1893 2020-09-27 Hypertension Normal                       247.
##  6      5216       4292 2020-08-05 Diabetes     Stage 1 Hyper~               151.
##  7      6388       4426 2020-09-23 Heart Disea~ Stage 2 Hyper~               193.
##  8       412       3595 2020-01-18 Obesity      Elevated                     220.
##  9      4202       1062 2020-06-24 Asthma       Elevated                     220.
## 10      7500       1336 2020-11-08 Diabetes     Normal                       126.
## # i 7,281 more rows
## # i 6 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>, visit_month <dbl>, patient_age_group <chr>,
## #   cholesterol_risk <chr>
```

A new column called "cholesterol_risk" was added with the mutate() function that checks the cholesterol_level column and returns a value using the case_when() in the new column of either "Low", "Moderate", or "High".

4. Summarize the number of visits per hospital location and condition:

```
cleaned_health_records %>%
  group_by(hospital_location, condition) %>%
  summarise(visits = n()) %>%
  pivot_wider(names_from = condition, values_from = visits)
```

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

```
## # A tibble: 5 x 6
## # Groups:   hospital_location [5]
##   hospital_location Asthma Diabetes `Heart Disease` Hypertension Obesity
##   <chr>              <int>    <int>           <int>        <int>   <int>
## 1 Chicago              297      320             290          327     308
## 2 Houston              295      283             304          256     305
## 3 LA                   284      292             271          303     321
## 4 Miami                297      282             274          288     266
## 5 NYC                  263      318             279          287     281
```

The cleaned_health_records were grouped by the hospital_location and condition columns of the dataset. I then used the pivot_wider() function to widen the the condition column.

Data Joining and Merging

1. Merge the patient dataset:

```
merged_health_records <- cleaned_health_records %>%
  inner_join(patients, by = "patient_id")

glimpse(merged_health_records)
```

```
## Rows: 7,291
## Columns: 16
## $ record_id         <dbl> 7908, 2115, 1545, 7028, 6493, 5216, 6388, 412, 4202,~
## $ patient_id        <dbl> 4911, 1530, 3840, 1178, 1893, 4292, 4426, 3595, 1062~
## $ visit_date        <date> 2020-11-25, 2020-03-29, 2020-03-05, 2020-10-19, 202~
## $ condition         <chr> "Obesity", "Hypertension", "Obesity", "Diabetes", "H~
## $ blood_pressure    <chr> "Elevated", "Normal", "Stage 1 Hypertension", "Eleva~
## $ cholesterol_level <dbl> 128.5, 155.2, 237.2, 178.0, 246.8, 151.3, 192.7, 220~
## $ medication_given  <chr> "No", "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes",~
## $ hospital_location <chr> "NYC", "LA", "Miami", "Chicago", "Houston", "NYC", "~
## $ patient_age       <dbl> 34, 20, 37, 65, 55, 25, 39, 21, 32, 40, 62, 55, 88, ~
## $ visit_month       <dbl> 11, 3, 3, 10, 9, 8, 9, 1, 6, 11, 4, 1, 10, 12, 1, 7,~
## $ patient_age_group <chr> "Young", "Young", "Middle-aged", "Senior", "Middle-a~
## $ cholesterol_risk  <chr> "Low", "Low", "High", "Moderate", "High", "Low", "Mo~
## $ patient_name      <chr> "Patient_1", "Patient_2", "Patient_3", "Patient_4", ~
## $ gender            <chr> "Female", "Male", "Other", "Female", "Male", "Other"~
## $ admission_date    <date> 2015-01-02, 2015-01-03, 2015-01-04, 2015-01-05, 201~
## $ insurance_status  <chr> "Private", "Medicare", "Medicaid", "Private", "Priva~
```

The inner_join() function was used to merge the messy_health_records dataset with the data in the patients dataset that corresponds with the patient_id.

2. Merge the medication dataset:

```
merged_health_records <- merged_health_records %>%
  inner_join(medications, by = "condition")

glimpse(merged_health_records)
```

```
## Rows: 7,291
## Columns: 18
## $ record_id             <dbl> 7908, 2115, 1545, 7028, 6493, 5216, 6388, 412, ~
## $ patient_id            <dbl> 4911, 1530, 3840, 1178, 1893, 4292, 4426, 3595,~
## $ visit_date            <date> 2020-11-25, 2020-03-29, 2020-03-05, 2020-10-19~
## $ condition             <chr> "Obesity", "Hypertension", "Obesity", "Diabetes~
## $ blood_pressure        <chr> "Elevated", "Normal", "Stage 1 Hypertension", "~
## $ cholesterol_level     <dbl> 128.5, 155.2, 237.2, 178.0, 246.8, 151.3, 192.7~
## $ medication_given      <chr> "No", "Yes", "Yes", "Yes", "No", "Yes", "No", "~
## $ hospital_location     <chr> "NYC", "LA", "Miami", "Chicago", "Houston", "NY~
## $ patient_age           <dbl> 34, 20, 37, 65, 55, 25, 39, 21, 32, 40, 62, 55,~
## $ visit_month           <dbl> 11, 3, 3, 10, 9, 8, 9, 1, 6, 11, 4, 1, 10, 12, ~
```

```
## $ patient_age_group      <chr> "Young", "Young", "Middle-aged", "Senior", "Mid~
## $ cholesterol_risk       <chr> "Low", "Low", "High", "Moderate", "High", "Low"~
## $ patient_name           <chr> "Patient_1", "Patient_2", "Patient_3", "Patient~
## $ gender                 <chr> "Female", "Male", "Other", "Female", "Male", "O~
## $ admission_date         <date> 2015-01-02, 2015-01-03, 2015-01-04, 2015-01-05~
## $ insurance_status       <chr> "Private", "Medicare", "Medicaid", "Private", "~
## $ recommended_medication <chr> "Orlistat", "Lisinopril", "Orlistat", "Metformi~
## $ dosage_mg              <dbl> 120, 10, 120, 500, 10, 500, 20, 120, 90, 500, 9~
```

The previously created merged_health_data and the medications data are merged by condition column using the inner_join() function similar to the previous problem.

3. Filter out records:

```
merged_health_records %>%
  filter(dosage_mg <= 200)
```

```
## # A tibble: 5,796 x 18
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                      <dbl>
##  1      7908       4911 2020-11-25 Obesity      Elevated                    128.
##  2      2115       1530 2020-03-29 Hypertension Normal                      155.
##  3      1545       3840 2020-03-05 Obesity      Stage 1 Hyper~              237.
##  4      6493       1893 2020-09-27 Hypertension Normal                      247.
##  5      6388       4426 2020-09-23 Heart Disea~ Stage 2 Hyper~              193.
##  6       412       3595 2020-01-18 Obesity      Elevated                    220.
##  7      4202       1062 2020-06-24 Asthma       Elevated                    220.
##  8      2361       4672 2020-04-08 Asthma       Elevated                    150.
##  9       208       3156 2020-01-09 Hypertension Stage 2 Hyper~              157.
## 10      6670       4573 2020-10-04 Hypertension Elevated                    178.
## # i 5,786 more rows
## # i 12 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>, visit_month <dbl>, patient_age_group <chr>,
## #   cholesterol_risk <chr>, patient_name <chr>, gender <chr>,
## #   admission_date <date>, insurance_status <chr>,
## #   recommended_medication <chr>, dosage_mg <dbl>
```

I filtered out all rows in merged_health_data that had a dosage_mg greater than 200.

4. Identify the top 5 most frequently treated conditions:

```
merged_health_records %>%
  group_by(condition) %>%
  summarise(total_treated = n()) %>%
  arrange(desc(total_treated)) %>%
  head(5)
```

```
## # A tibble: 5 x 2
##   condition     total_treated
```

```
##   <chr>                <int>
## 1 Diabetes              1495
## 2 Obesity               1481
## 3 Hypertension          1461
## 4 Asthma                1436
## 5 Heart Disease         1418
```

The 5 most frequently treated conditions were found by taking the merged_health_data and grouping the data by condition. I then summarized the total amount of patients that were treated for each condition with n(). Finally, I arranged the data in descending order based on total treated patients and then took the top 5 conditions with head(5).

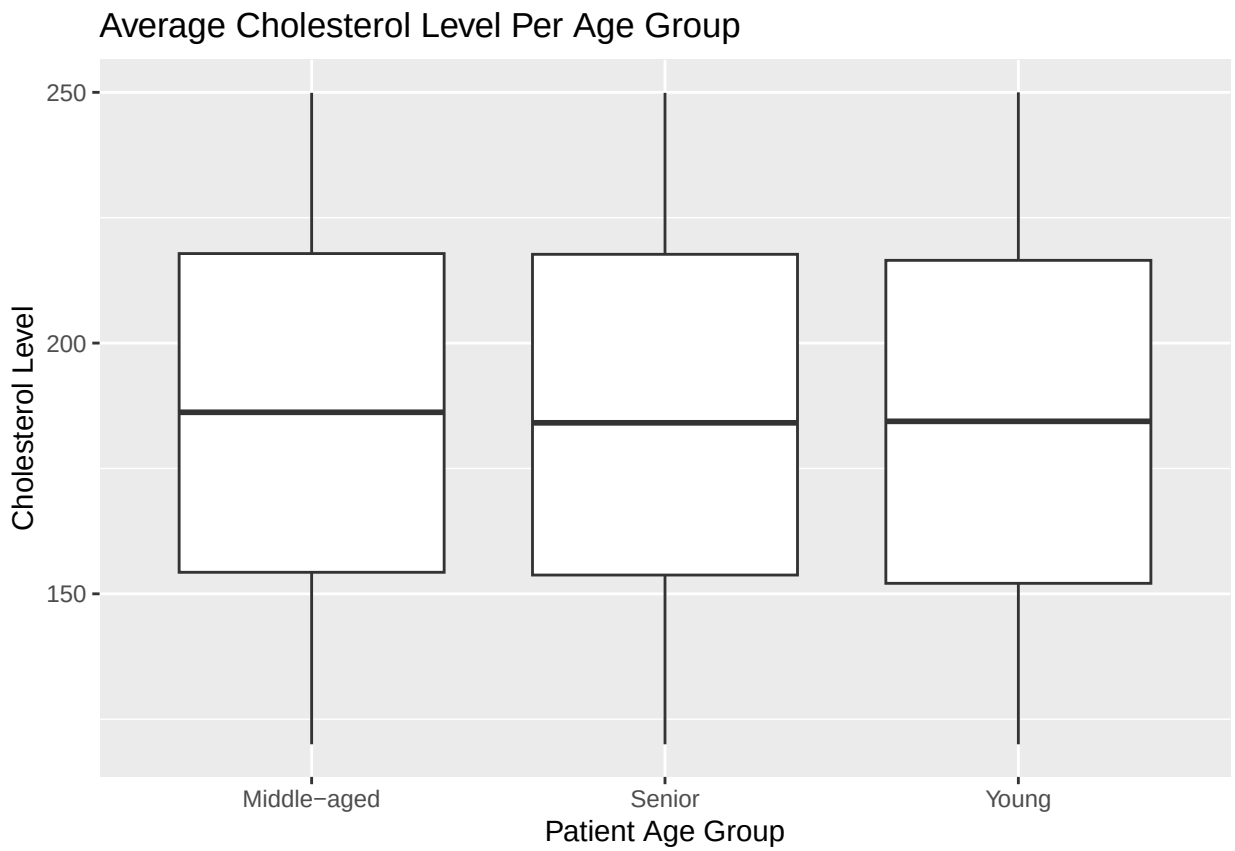
Exploratory Data Analysis

1. What is the average cholesterol level?:

```
summary_average_cholesterol<- merged_health_records %>%
  group_by(patient_age_group) %>%
  summarise(average_cholesterol_level = mean(cholesterol_level))
summary_average_cholesterol
```

```
## # A tibble: 3 x 2
##   patient_age_group average_cholesterol_level
##   <chr>                                <dbl>
## 1 Middle-aged                           186.
## 2 Senior                                185.
## 3 Young                                 185.
```

```
merged_health_records %>% ggplot() +
  geom_boxplot(aes(x = patient_age_group, y = cholesterol_level)) +
  labs(title = "Average Cholesterol Level Per Age Group",
       x = "Patient Age Group",
       y = "Cholesterol Level")
```

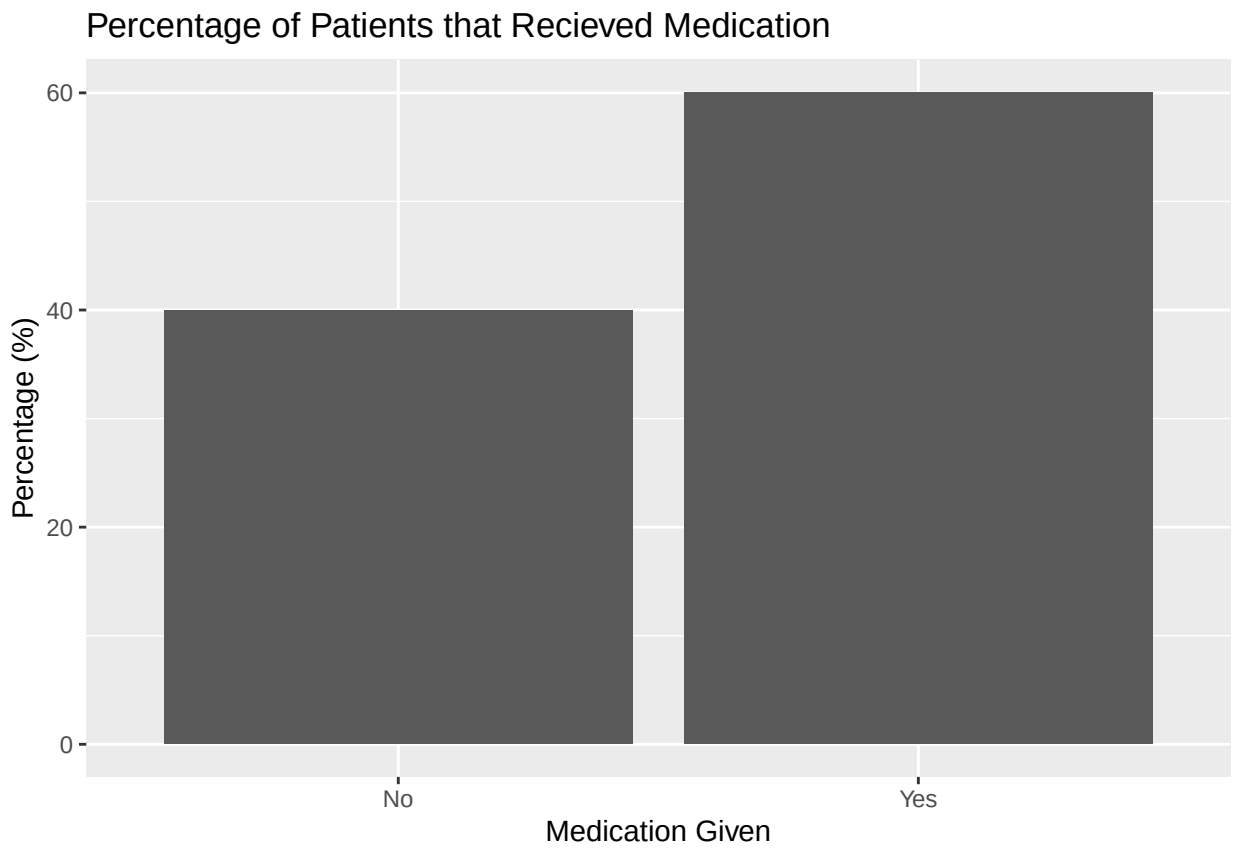


Every age group has a similar average cholesterol level with Middle-aged people having a slightly higher level at 185.7108 than Seniors at 185.2054 and Young at 184.6385.

2. What percentage of patients recieved medication?:

```
summary_medication_percent <- merged_health_records %>%
  group_by(medication_given) %>%
  summarise(percentage = n()) %>%
  mutate(percentage = percentage/sum(percentage)*100)

summary_medication_percent %>%
  ggplot() +
  geom_col(aes(x = medication_given, y = percentage)) +
  labs(title = "Percentage of Patients that Recieved Medication",
       x = "Medication Given",
       y = "Percentage (%)")
```



Around 60% of patients recieved medication during there hospital visit whereas around 40% did not receive medication.

3. Which hospital location treated the highest number of patients?:

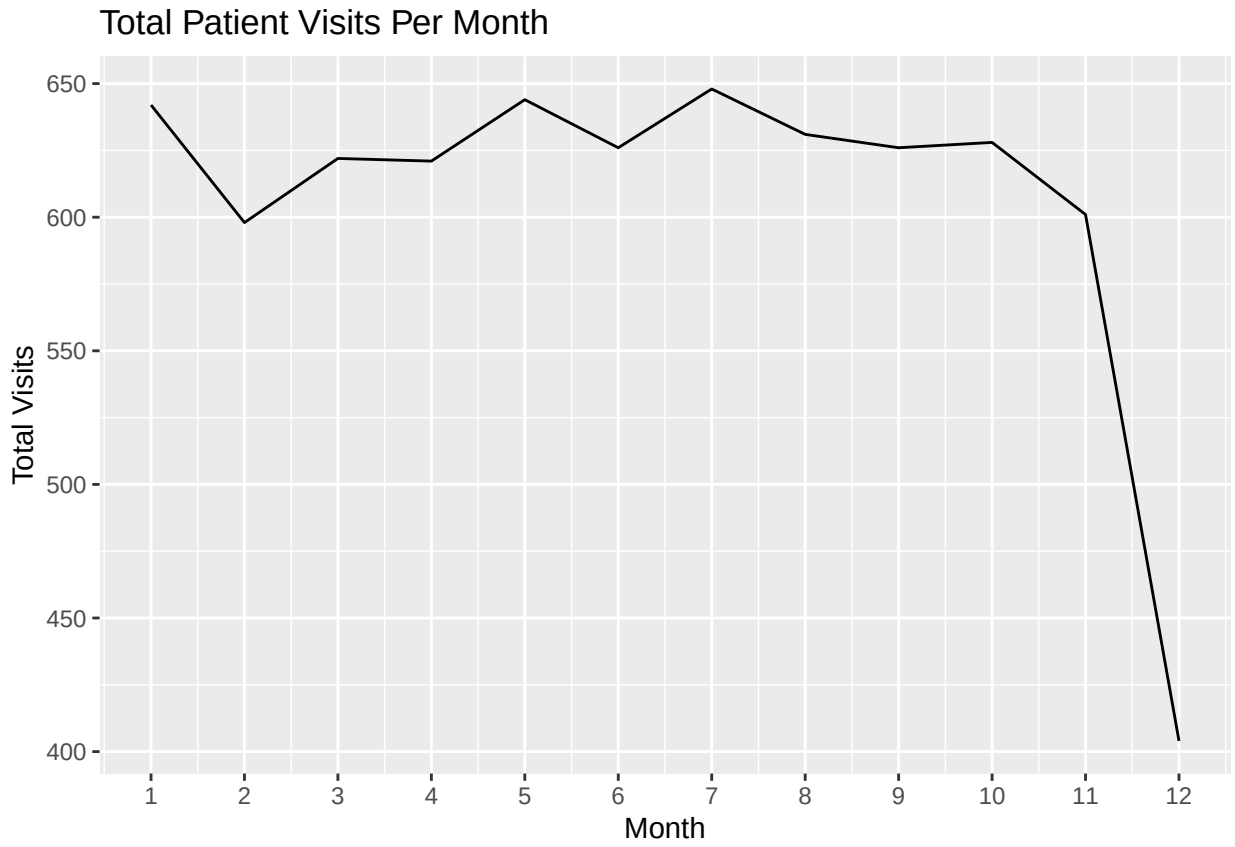
```
merged_health_records %>%
  group_by(hospital_location) %>%
  summarise(num_patients = n()) %>%
  arrange(desc(num_patients))
```

```
## # A tibble: 5 x 2
##   hospital_location num_patients
##   <chr>                    <int>
## 1 Chicago                   1542
## 2 LA                        1471
## 3 Houston                   1443
## 4 NYC                       1428
## 5 Miami                     1407
```

The merged_health_records data is grouped my hospital location and summarized by the number of patients treated at each hospital with the n() function. I then arranged the data in descending order. According to the data above, Chicago treated the highest number of patients at 1542.

4. Create a time series plot

```
merged_health_records %>%
  group_by(visit_month) %>%
  summarise(total_visits = n()) %>%
  ggplot(aes(x = visit_month, y = total_visits)) +
  geom_line() +
  labs(title = "Total Patient Visits Per Month",
       x = "Month",
       y = "Total Visits") +
  scale_x_continuous(breaks=seq(1,12,1))
```



A time series plot that plots total patient visits to the hospital by monthwas created. This was accomplished by grouping the data by visit_month and summarizing the total visits that occurred in each month. I plotted this summarized data into a line plot with visit_month on the x-axis and total_visits on the y-axis to create an easily understandable plot that shows the number of patients that these hospitals receive in each month. According to the data above, total patient visits tends to hover between 600 and 650 every month with the sole exception of December which sees a large decline in total visits dropping close to 400.

Healthcare Insights and Recomendations

1. Which age has the highest cholesterol levels?:

```
merged_health_records %>%
  group_by(patient_age_group) %>%
  summarise(cholesterol_level = mean(cholesterol_level)) %>%
  arrange(desc(cholesterol_level))
```

```
## # A tibble: 3 x 2
##   patient_age_group cholesterol_level
##   <chr>                         <dbl>
## 1 Middle-aged                    186.
## 2 Senior                         185.
## 3 Young                          185.
```

According to the above data, Middle-aged people have the highest cholesterol levels at 185.7108 which is slightly higher than Seniors at 185.2054 and young people at 184.6385. I found this information by grouping my data based on age group and summarizing the mean of the cholesterol levels of each age group and then arrange the new data in descending order.

2. Which condition is most frequently treated across all hospitals?:

```
merged_health_records %>%
  group_by(hospital_location, condition) %>%
  summarise(num_condition = n()) %>%
  arrange(hospital_location, desc(num_condition)) %>%
  slice(1)
```

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

```
## # A tibble: 5 x 3
## # Groups:   hospital_location [5]
##   hospital_location condition    num_condition
##   <chr>             <chr>                <int>
## 1 Chicago           Hypertension           327
## 2 Houston           Obesity                305
## 3 LA                Obesity                321
## 4 Miami             Asthma                 297
## 5 NYC               Diabetes               318
```

```
merged_health_records %>%
  group_by(condition) %>%
  summarise(num_condition = n()) %>%
  arrange(desc(num_condition))
```

```
## # A tibble: 5 x 2
##   condition     num_condition
```

```
##   <chr>                <int>
## 1 Diabetes              1495
## 2 Obesity               1481
## 3 Hypertension          1461
## 4 Asthma                1436
## 5 Heart Disease         1418
```

According to the data above, diabetes is the most common condition across the entire data set with it also being the most common condtion in NYC. In the other hospital locations Hyptertension is the most common condition in Chicago, asthma is the most common condition in Miami, and obesity is the most common condition in Houstin and LA.

3. Recommendations:

Based on our findings, I would encourage patients to listen to their doctor and take medication if it is prescribed. Also, the most common conditions overall were Diabetes and Obesity which can both be cause by unhealthy eating patterns. So one major suggestion I'd have for these patients is to consider exercising regularly and to reach out to a dietitian to create a dietary plan to hopefully either lose weight or reduce the negative effects of these conditions.

Identifying High-Risk Patients

1. I will define high risk as patients with High cholesterol (>200 mg/dL) AND Stage 2 Hypertension:

```
high_risk_patients <- merged_health_records %>%
  filter(cholesterol_level > 200 & blood_pressure == "Stage 2 Hypertension")
```

I found the highest risk patients by filtering for patients with cholesterol level greater than 200 and blood pressure equal to Stage 2 Hypertension.

2. Rank the top 10 most at-risk patients based on cholesterol and blood pressure data:

```
high_risk_patients %>%
  arrange(desc(cholesterol_level)) %>%
  head(10)
```

```
## # A tibble: 10 x 18
##    record_id patient_id visit_date condition    blood_pressure cholesterol_level
##        <dbl>      <dbl> <date>     <chr>        <chr>                       <dbl>
##  1      3023       3931 2020-05-05 Asthma       Stage 2 Hyper~              250.
##  2      8080       4320 2020-12-02 Asthma       Stage 2 Hyper~              250.
##  3      3947       2981 2020-06-13 Asthma       Stage 2 Hyper~              250.
##  4      1826       1572 2020-03-17 Diabetes     Stage 2 Hyper~              250.
##  5      3588       1945 2020-05-29 Diabetes     Stage 2 Hyper~              250.
##  6      4646       3043 2020-07-12 Asthma       Stage 2 Hyper~              250.
##  7      3621       1172 2020-05-30 Asthma       Stage 2 Hyper~              250.
##  8      8338       1785 2020-12-13 Hypertension Stage 2 Hyper~              250.
```

```
##  9      7051       1864 2020-10-20 Hypertension Stage 2 Hyper~              250.
## 10      1990       3279 2020-03-23 Heart Disea~ Stage 2 Hyper~              249.
## # i 12 more variables: medication_given <chr>, hospital_location <chr>,
## #   patient_age <dbl>, visit_month <dbl>, patient_age_group <chr>,
## #   cholesterol_risk <chr>, patient_name <chr>, gender <chr>,
## #   admission_date <date>, insurance_status <chr>,
## #   recommended_medication <chr>, dosage_mg <dbl>
```

I found the top 10 most at-risk patients by taking the previously created high_risk_patients data set and arranging it in descending order based on cholesterol level and then took the top 10 highest risk patients with the head(10) function.

3. Provide a scatter plot:

```
merged_health_records %>%
  ggplot(aes(x = patient_age, y = cholesterol_level)) +
  geom_point() +
  labs(title = "Scatter Plot for Cholesterol Level vs Patient Age",
       x = "Patient Age",
       y = "Cholesterol Level")
```

This scatter plot was created from the merged_health_data previously created showing cholesterol level based on patient age. The scatter plot is very dense and it is clear that there is not a strong correlation between patient age and cholesterol level.

R Appendix

```r
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)

medications <- read_csv("medications.csv")

messy_health_records <- read_csv("messy_health_records.csv")

patients <- read_csv("patients.csv")
messy_health_records %>%
  summarise(
    record_id = sum(is.na(record_id)),
    patient_id = sum(is.na(patient_id)),
    visit_date = sum(is.na(visit_date)),
    condition = sum(is.na(condition)),
    blood_pressure = sum(is.na(blood_pressure)),
    cholesterol_level = sum(is.na(cholesterol_level)),
    medication_given = sum(is.na(medication_given)),
    hospital_location = sum(is.na(hospital_location)),
    patient_age = sum(is.na(patient_age))
    )
cleaned_health_records <- messy_health_records %>%
  drop_na()

cleaned_health_records
cleaned_health_records <- cleaned_health_records %>%
  mutate(visit_date = ymd(visit_date))

cleaned_health_records
removed_duplicates <- nrow(cleaned_health_records) - nrow(cleaned_health_records %>% unique())
removed_duplicates

cleaned_health_records <- cleaned_health_records %>% unique()
cleaned_health_records
IQR_cholesterol <- IQR(cleaned_health_records$cholesterol_level, na.rm = TRUE)

Q1_cholesterol <- quantile(cleaned_health_records$cholesterol_level, probs = 0.25, na.rm = TRUE)

Q3_cholesterol <- quantile(cleaned_health_records$cholesterol_level, probs = 0.75, na.rm = TRUE)

lower_bound_cholesterol <- Q1_cholesterol - 1.5 * IQR_cholesterol

upper_bound_cholesterol <- Q3_cholesterol + 1.5 * IQR_cholesterol


IQR_patient_age <- IQR(cleaned_health_records$patient_age, na.rm = TRUE)

Q1_patient_age <- quantile(cleaned_health_records$patient_age, probs = 0.25, na.rm = TRUE)

Q3_patient_age <- quantile(cleaned_health_records$patient_age, probs = 0.75, na.rm = TRUE)

lower_bound_patient_age <- Q1_patient_age - 1.5 * IQR_patient_age
```

```r
upper_bound_patient_age <- Q3_patient_age + 1.5 * IQR_patient_age

cleaned_health_records <- cleaned_health_records %>%
  filter(cholesterol_level >= lower_bound_cholesterol & cholesterol_level <= upper_bound_cholesterol) %
  filter(patient_age >= lower_bound_patient_age & patient_age <= upper_bound_patient_age)

cleaned_health_records
cleaned_health_records <- cleaned_health_records %>% mutate(visit_month = month(visit_date))

cleaned_health_records
cleaned_health_records <- cleaned_health_records %>%
  mutate(patient_age_group = case_when(
    patient_age >= 18 & patient_age <= 35 ~ "Young",
    patient_age >= 36 & patient_age <= 60 ~ "Middle-aged",
    patient_age >= 61 ~ "Senior"
    ))

cleaned_health_records
cleaned_health_records <- cleaned_health_records %>%
  mutate(cholesterol_risk = case_when(
    cholesterol_level < 160 ~ "Low",
    cholesterol_level >= 160 & cholesterol_level <= 200 ~ "Moderate",
    cholesterol_level > 200 ~ "High"
  ))

cleaned_health_records
cleaned_health_records %>%
  group_by(hospital_location, condition) %>%
  summarise(visits = n()) %>%
  pivot_wider(names_from = condition, values_from = visits)
merged_health_records <- cleaned_health_records %>%
  inner_join(patients, by = "patient_id")

glimpse(merged_health_records)
merged_health_records <- merged_health_records %>%
  inner_join(medications, by = "condition")

glimpse(merged_health_records)
merged_health_records %>%
  filter(dosage_mg <= 200)
merged_health_records %>%
  group_by(condition) %>%
  summarise(total_treated = n()) %>%
  arrange(desc(total_treated)) %>%
  head(5)
summary_average_cholesterol<- merged_health_records %>%
  group_by(patient_age_group) %>%
  summarise(average_cholesterol_level = mean(cholesterol_level))
summary_average_cholesterol

merged_health_records %>% ggplot() +
  geom_boxplot(aes(x = patient_age_group, y = cholesterol_level)) +
  labs(title = "Average Cholesterol Level Per Age Group",
```

```
       x = "Patient Age Group",
       y = "Cholesterol Level")
summary_medication_percent <- merged_health_records %>%
  group_by(medication_given) %>%
  summarise(percentage = n()) %>%
  mutate(percentage = percentage/sum(percentage)*100)

summary_medication_percent %>%
  ggplot() +
  geom_col(aes(x = medication_given, y = percentage)) +
  labs(title = "Percentage of Patients that Recieved Medication",
       x = "Medication Given",
       y = "Percentage (%)")
merged_health_records %>%
  group_by(hospital_location) %>%
  summarise(num_patients = n()) %>%
  arrange(desc(num_patients))
merged_health_records %>%
  group_by(visit_month) %>%
  summarise(total_visits = n()) %>%
  ggplot(aes(x = visit_month, y = total_visits)) +
  geom_line() +
  labs(title = "Total Patient Visits Per Month",
       x = "Month",
       y = "Total Visits") +
  scale_x_continuous(breaks=seq(1,12,1))
merged_health_records %>%
  group_by(patient_age_group) %>%
  summarise(cholesterol_level = mean(cholesterol_level)) %>%
  arrange(desc(cholesterol_level))
merged_health_records %>%
  group_by(hospital_location, condition) %>%
  summarise(num_condition = n()) %>%
  arrange(hospital_location, desc(num_condition)) %>%
  slice(1)

merged_health_records %>%
  group_by(condition) %>%
  summarise(num_condition = n()) %>%
  arrange(desc(num_condition))
high_risk_patients <- merged_health_records %>%
  filter(cholesterol_level > 200 & blood_pressure == "Stage 2 Hypertension")
high_risk_patients %>%
  arrange(desc(cholesterol_level)) %>%
  head(10)
merged_health_records %>%
  ggplot(aes(x = patient_age, y = cholesterol_level)) +
  geom_point() +
  labs(title = "Scatter Plot for Cholesterol Level vs Patient Age",
       x = "Patient Age",
       y = "Cholesterol Level")
knitr::purl(input = "Report.Rmd", output = "Report.R", documentation = 0)
```