INFO 526 - Assignment 5

Gayathri Renganathan

Reflection:

Though I had experience working with data and plots earlier. This class helped me understand my weakness and addressed the areas of improvement through various learning criteria. In this portfolio I have created 4 plots and 1 table to show how I have mastered the criteria listed in the syllabus. Throughout the semester your feedback has indicated I needed to work particularly hard on a few criteria that I was initially weak on. I missed to capture on a few criteria in the assignments

- 1. – How to keep the charts simple, how to effectively communicate with simple chart,
- 11. -using incorrect histogram bin width and how to avoid illusion effects in the chart.

I have worked on the feedback and addressed on my plot 1,2 and table 1 to keep the plots/table simple enough to address the questions (criteria: 1). Plot 1: also addresses the issue of using correct bin width to explain the data effectively(Criteria 11.d). In some of my assignments the color selections in the plot has led to illusion effects. I have worked hard to avoid the illusion effects in the final portfolio in all the plots.

In most assignments I had trouble to make the plot/table simple and effective. With practice, I have now successfully eliminated these problems (criteria 1) in all of the plots that follow. As a result, I think I deserve an **A** in the course, as my portfolio shows I have mastered almost all the criteria and have improved dramatically on those that I was struggling with before.

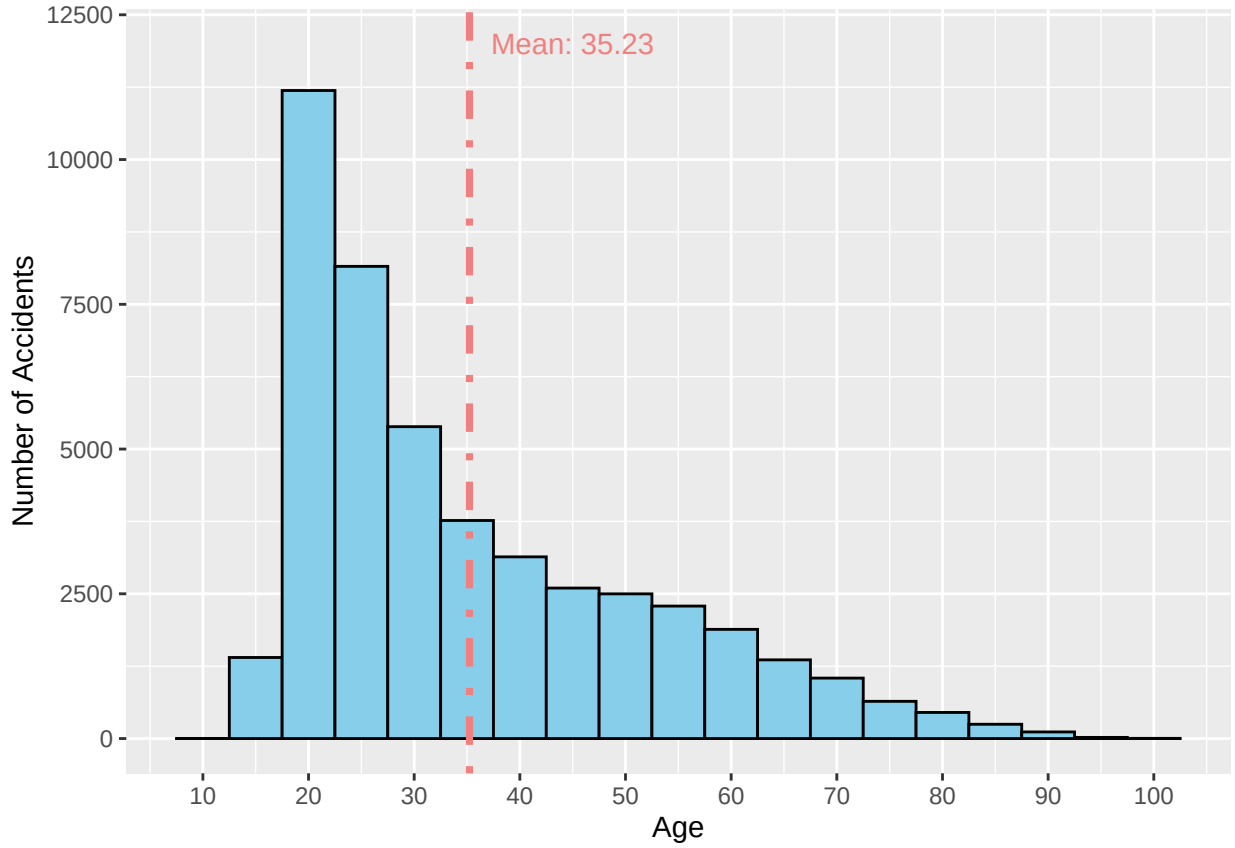
My portfolio plots are designed to specifically show mastery in the following criteria:

- Whole Portfolio - 2,3,10,11,12,13,14
- 1, a plot - 1, 4, 5
- 2, a basic bar plot, 1, 7
- 3, a faceted plot - 1, 6, 7
- 4, a table - 1, 3, 7
- 5, sankey diagram - 4, 5, 7

Question

What is the distribution of ages of the driver that caused the accident?

Plot

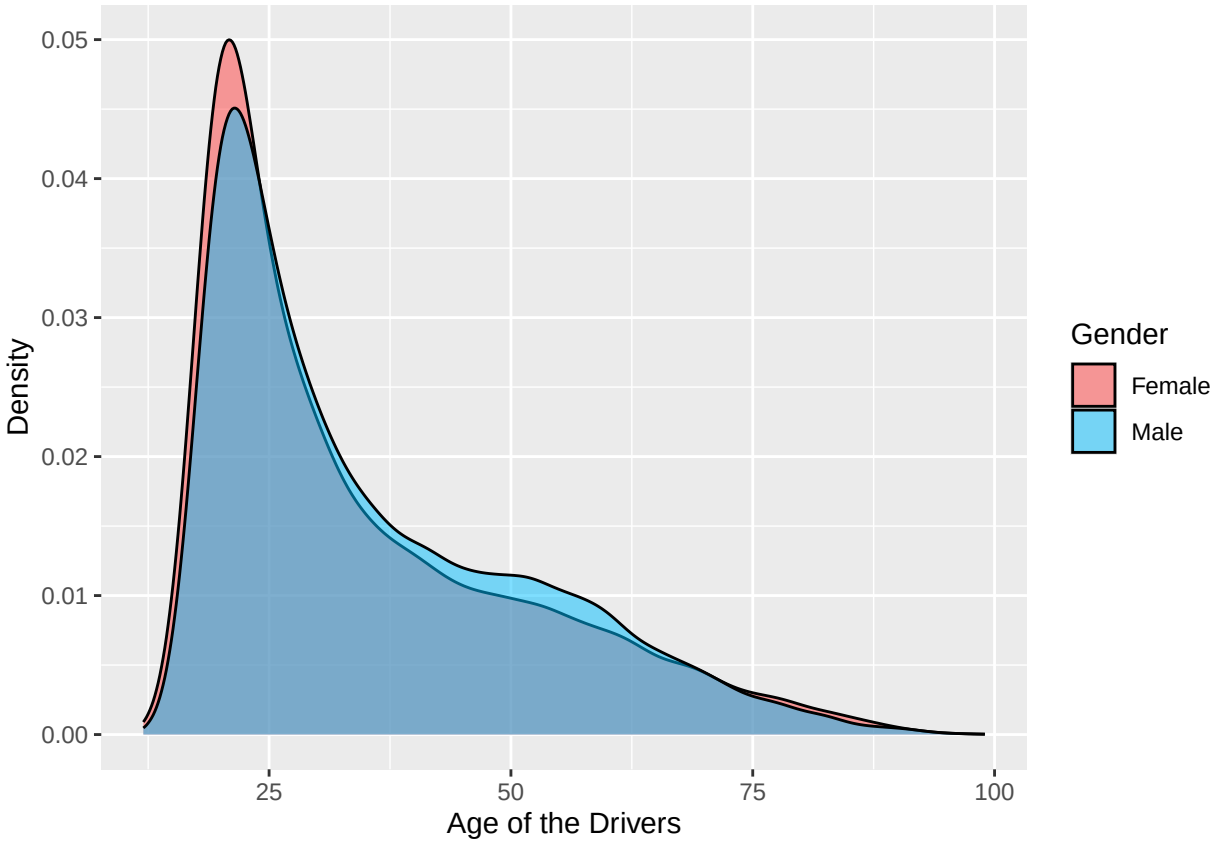


Plot 1: The Histogram plot shows the age distribution of the drivers that caused the accident. Accidents are more prominent in the young age drivers in the age of 18 to 28. The number of accidents decreases as the age of the driver increases. Very few accidents in the range of 13 -18 years indicates that people are driving without proper license or possible data entry issue.

Question

What is the distribution of ages of the driver across gender that caused the accident?

Plot



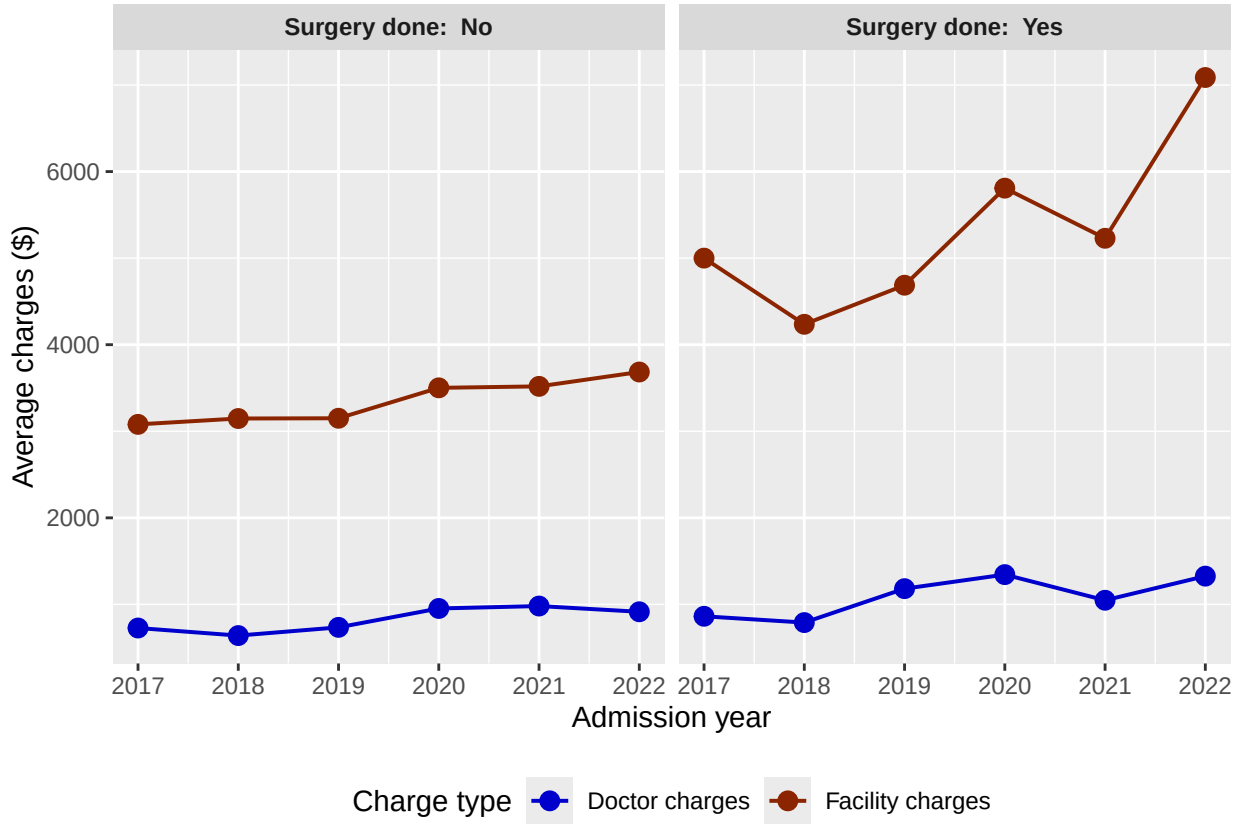
Plot2: The Density plot shows the age distribution of the Drivers that caused the accident across Gender. Age distribution is similar for Female and Male with minor differences. Female drivers has slightly higher accidents than male drivers until the age of 25. Male drivers has slightly higher accidents than female drivers in the age group of 30-60 years.

Question

Question1: How has the relationship between doctor versus facility charges changed over time?

Question2: If surgery was performed as part of that visit, how does it affect the amount charged for the doctor?

Plot



Plot 3: The line plot shows the trend of how the Average Doctor charges and Average Facility charges varied over the years. Average Facility charges are very high compared to the Average Doctor charges across all the years. The average facility charges are elevated when surgery was performed. The average doctor charges are slightly higher when surgery was performed.

Question

Question: How has the total amount charged for an ER visit fluctuated over time?

ER visit total charges by year				
Year	**Median**	**Q1**	**Q3**	**No. of visits**
2017	2,500	1,200	5,000	6,500
2018	2,400	1,100	5,100	6,400
2019	2,500	1,100	5,200	6,100
2020	2,900	1,200	5,900	4,600
2021	2,800	1,200	6,200	5,100
2022	2,900	1,300	6,100	4,300

Table 1: Summary statistics of emergency room(ER) visit total charges from 2017 to 2022 based on the MEPS Survey data in America. The increasing median total charges over the years highlights the rising uncertainity in ER costs.

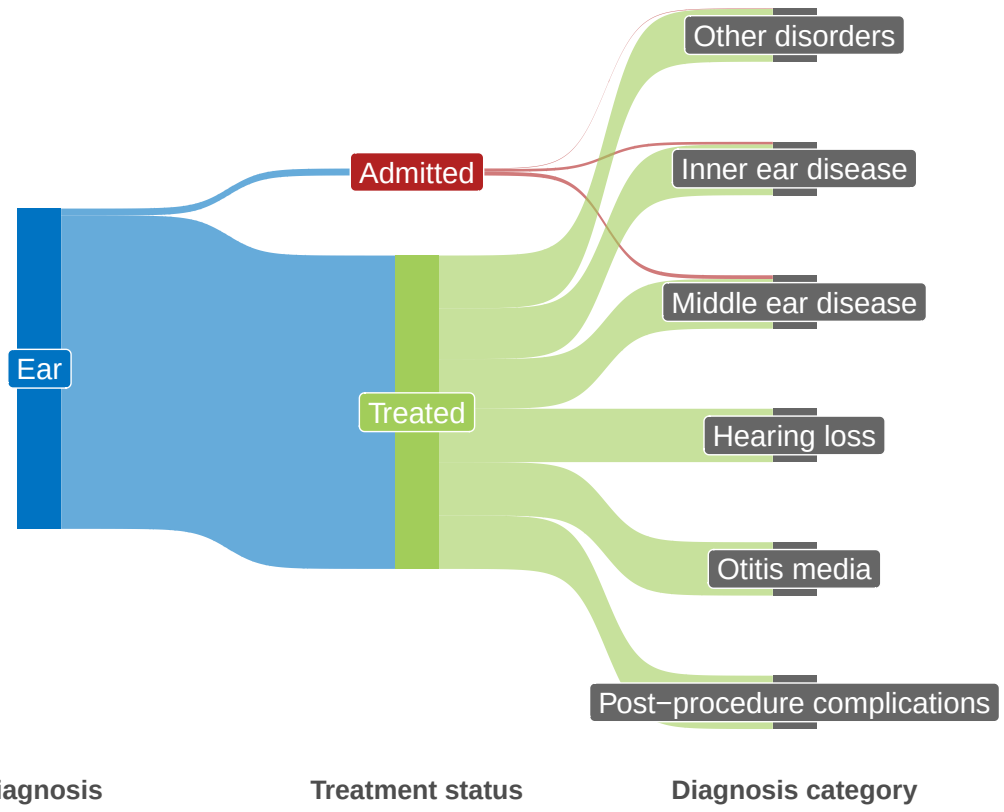
Table Notes:

- All numbers are rounded off to the nearest hundreds.
- Median, Q1, Q3 are in USD($).
- Q1 refers to the first quartile (25th percentile), while Q3 refers to the third quartile (75th percentile) of the data.

Question

Question: What is the distribution of Admitted and Treated Patients under "EAR" diagnosis?

Plot



Plot 4: The Sankey diagram shows the distribution of Admitted and Treated patients across Ear-related disease category. The diagram illustrates that most of the patients diagnosed under Ear-related diseases were treated. Only people diagnosed with Middle ear disease / Inner ear disease / Other disorders were admitted.

My Code

```
# Loading libraries
library(dplyr)
library(ggplot2)
library(readxl)
library(tidyr)
library(ggsankey)
library(reshape2)
library(stringr)
suppressWarnings(library(gt))
```

```
# Reading the Crash report Dataset
crashReport <- read.csv("Tempe_Crash_Data_Report_2024.csv")

# To get the drivers who caused the accident , we can check if the driver has
# done any violation / alcohol use / drug use
#
# For Driver1, Remove the records age > 100 and filter for 'Driver' type
# Filter drivers without any violations

crashReport_Drv1_cleaned <- crashReport %>%
  filter(Age_Drv1 < 100 & Unittype_One == "Driver"
         & (Violation1_Drv1 != "No Improper Action" |
              AlcoholUse_Drv1 != "No Apparent Influence"
            | DrugUse_Drv1 != "No Apparent Influence" )) %>%
  select(Incidentid, Year, Age_Drv1, Unittype_One, Gender_Drv1 ) %>%
  rename(Age_Drv = Age_Drv1, Unit_type = Unittype_One,
         Gender = Gender_Drv1) %>%
  mutate(group = 'Driver1')

# For Driver2, Remove the records age > 100 and filter for 'Driver' type
# Filter drivers without any violations

crashReport_Drv2_cleaned <- crashReport %>%
  filter(Age_Drv2 < 100 & Unittype_Two == "Driver"
         & (Violation1_Drv2 != "No Improper Action" |
              AlcoholUse_Drv2 != "No Apparent Influence"
            | DrugUse_Drv2 != "No Apparent Influence" )) %>%
  select(Incidentid, Year,Age_Drv2, Unittype_Two, Gender_Drv2 ) %>%
  rename(Age_Drv = Age_Drv2, Unit_type = Unittype_Two,
         Gender = Gender_Drv2) %>%
  mutate(group = 'Driver2')

#Combine Driver1 and Driver2 dataset
crashReport_cleaned = rbind(crashReport_Drv1_cleaned,
                            crashReport_Drv2_cleaned)

# Calculate the average age of the Drivers
mean_age <- round(mean(crashReport_cleaned$Age_Drv), 2)

# Annotation data for Geom_text in Histogram
annotations <- data.frame(x = 45, y = 12000,
```

```
                          label = paste("Mean:", mean_age ))

# Histogram plot to show the age distribution of the drivers
ggplot(data= crashReport_cleaned, aes(x = Age_Drv))+
  geom_histogram(fill= 'skyblue', col = 'black', binwidth = 5)+
  geom_vline(aes(xintercept = mean_age), color = 'lightcoral',
             linewidth = 1.25, linetype = 'dotdash') +
  geom_text(data = annotations,
            aes(x = x, y = y, label = label),
            size = 4,
            col= 'lightcoral') +
  scale_x_continuous(
    breaks = c(seq(10, 110, 10))
  )+
  labs(x = "Age", y = "Number of Accidents")
```

```
# Color coding for different gender
cols <- c("brown1", "deepskyblue")

# Density plot to show the density distribution of the Driver's age
ggplot(data= crashReport_cleaned, aes(x = Age_Drv, fill = Gender))+
  geom_density(alpha = 0.5)+
  scale_fill_manual(values = cols)+
  labs(x = "Age of the Drivers", y = "Density")
```

```
# Reading the Household emergency room visit cost dataset
er_cost <- read_excel("Household_Emergency_Room_Visit_Costs_2022.xlsx")

# SURGPROC = 1 , indicates surgery being performed in the visit
# Grouping data by year, Surgery_done
# calculate the average doctor & facility charge
er_grouped_charge <- er_cost %>%
  mutate("Surgery_Done" = ifelse(SURGPROC == 1, "Yes", "No")) %>%
  group_by(ERDATEYR,Surgery_Done) %>%
  summarise("Avg Doctor Charge" = mean(ERDTC),
            "Avg Facility Charge" = mean(ERFTC)) %>%
  melt(id.vars = c("ERDATEYR","Surgery_Done"), variable.name = "Charge_Type")


# Line plot to show the average doctor charge and average facility charge
# vary over the years
# Whether surgery was performed or not is indicated by the facets.
ggplot(data = er_grouped_charge,
       aes(x = ERDATEYR, y = value, color = Charge_Type) )+
  geom_line(linewidth = 0.7)+
  geom_point(size = 3)+
  facet_wrap(~Surgery_Done,
             nrow = 1,
             labeller = labeller(Surgery_Done =
                c("Yes" = "Surgery done:  Yes",
                  "No" = "Surgery done:  No")))+
  scale_color_manual(name= "Charge type",
                     values=c('mediumblue','orangered4'),
```

```
                    labels = c('Doctor charges','Facility charges'))+
  scale_x_continuous(breaks = c(seq(2017, 2022)))+
  labs(x = "Admission year", y = "Average charges ($)")+
  theme(legend.position = "bottom",
        strip.text = element_text(face="bold", size=9))
```

```
# There are no missing values in the year column
# There are 1670 records with 0 in ERTC - no charges in doctor and facility
# EVNTIDX is the unique event identifier and there are no duplicated record.
er_cost_filtered <- er_cost %>%
  filter(ERTC > 0) %>%
  distinct(EVNTIDX, .keep_all=TRUE)

# Grouping data by year and calculate the average doctor & facility charge
er_grouped_charge <- er_cost_filtered %>%
  group_by(ERDATEYR) %>%
  summarise(
            "Median" = prettyNum(round(median(ERTC),
                                       digits = -2),
                                 big.mark = ",", scientific = FALSE),
            "Q1" = prettyNum(round(quantile(ERTC, probs = 0.25),
                                   digits = -2),
                             big.mark = ",", scientific = FALSE),
            "Q3" = prettyNum(round(quantile(ERTC, probs = 0.75),
                                   digits = -2),
                             big.mark = ",", scientific = FALSE),
            "No. of visits" = prettyNum(round(n(),digits = -2),
                              big.mark = ",", scientific = FALSE))

# Create table using gt and update formatting
er_gt <- er_grouped_charge %>%
  gt() %>%
  cols_label(ERDATEYR = "Year")%>%
  opt_table_font(
    font = list(system_fonts(name = "monospace-code"))
  ) %>%
  tab_options(column_labels.font.weight = "bold",
              table.font.size = 14) %>%
  tab_header(title = "ER visit total charges by year") %>%
  cols_width(everything() ~ px(70))


gtsave(er_gt, "er_cost.png", expand = 10) #to export the table
knitr::include_graphics("er_cost.png") #to pull it back in
```

```
# Reading the Household emergency room visit cost dataset
hcup <- read.csv("HCUP_Edited_Arizona.csv")

# Filtering the dataset to remove NA  values
hcup_filtered <- hcup[!with(hcup, is.na(Treat.and.Release) &
                                  is.na(Patient.Admitted)), ]

# Data restructuring
```

```r
hcup_grouped <- hcup_filtered %>%
 filter(DiagnosisCode == "EAR") %>%
  group_by(DiagnosisCode, Diagnoses)  %>%
  summarise(
            Patients_sum = sum(Patient.Admitted,Treat.and.Release,
                               na.rm = TRUE),
            Patients_admitted_sum = sum(Patient.Admitted, na.rm = TRUE),
            Treated_sum = sum(Treat.and.Release, na.rm = TRUE)) %>%
  mutate(admitted_perc =
           round(Patients_admitted_sum / Patients_sum * 100),
         treated_perc =
           round(Treated_sum / Patients_sum * 100),
         `Diagnoses` =
           recode(`Diagnoses`,
                "EAR002: Diseases of middle ear and mastoid" =
                  "Middle ear disease",
                "EAR003: Diseases of inner ear and related conditions" =
                  "Inner ear disease",
                "EAR001: Otitis media" = "Otitis media",
                "EAR006: Oth disorders of the ear" = "Other disorders",
                "EAR004: Hearing loss" = "Hearing loss",
                "EAR005: Postproc ear/mastoid process complication" =
                  "Post-procedure complications"),
         DiagnosisCode = recode(DiagnosisCode,
                "EAR" = "Ear")) %>%
  select(DiagnosisCode, Diagnoses, admitted_perc,treated_perc) %>%
  gather(key="category", value="value", 3:4) %>%
  mutate(`category` = recode(`category`,
                             "admitted_perc" = "Admitted",
                             "treated_perc" = "Treated")) %>%
  uncount(value)

hcup_sankey1 <- hcup_grouped  %>%
  make_long(DiagnosisCode,category, Diagnoses)

# Reordering the nodes
hcup_sankey1$node <- factor(
  hcup_sankey1$node,
  levels =
    c("Ear", "Treated","Admitted", "Post-procedure complications", "Otitis media",
      "Hearing loss", "Middle ear disease", "Inner ear disease", "Other disorders")
)
hcup_sankey1$next_node <- factor(
  hcup_sankey1$next_node,
  levels =
    c("Ear", "Treated","Admitted", "Post-procedure complications", "Otitis media",
      "Hearing loss", "Middle ear disease", "Inner ear disease", "Other disorders")
)

# Named colors for nodes
named.col2 <-
  c(Ear = "#0073C2FF", Treated = "darkolivegreen3" ,Admitted = "firebrick",
    "Post-procedure complications" = "gray40" ,"Otitis media" = "gray40" ,
```

```
    "Hearing loss" = "gray40","Other disorders" = "gray40",
    "Inner ear disease" = "gray40","Middle ear disease" = "gray40" )

# Sankey Plot
ggplot(hcup_sankey1, aes(x = x,
                      next_x = next_x,
                      node = node,
                      next_node = next_node,
                      fill = factor(node))) +
  geom_sankey(flow.alpha = 0.6, smooth = 10,width = 0.115) +
  geom_sankey_label(aes(label = node), color= "white") +
  scale_fill_manual(values = named.col2, name = "")+
  theme_sankey(base_size = 12, ) +
  labs(x = NULL) +
  theme(legend.position = "none")+
  theme(axis.text.x = element_text(face = "bold"))+
  scale_x_discrete(labels = c("Diagnosis", "Treatment status",
                              "Diagnosis category"))
```