CITS4009_CDA_Project1

Amy HUNG (23702699)

1. Introduction

This dataset analysed is an extract from the US Accident Injury dataset, which can be obtained from Data.gov, published by US Department of Labour. The dataset contains data on all accidents, injuries and illnesses reported by US mine operators and contractors. The entire dataset spans across 15 years (2000 to 2015) and has a total of 202,814 observations, while this dataset analysed in this project covers only 2,000 observations of the entire dataset.

2. Data loading, overview and set up

2.1 Loading all the necessary libraries

```
library(ggplot2)
library(gridExtra)
library(ggthemes)
library(MASS)
library(dplyr)
library(usmap)
library(timeDate)
library(chron)
library(WVPlots)
library(scales)
library(treemapify)
library(lubridate)
library(tidyverse)
library(ggrepel)
```

2.2 Setting up plotting theme

Setting up a plotting theme to ensure all graphs can be look coherent and be aligned for the whole report.

```
cits4009_theme <- theme_minimal() +
                  theme(panel.grid.major.x = element_blank(), panel.grid.minor.x =
element_blank()) +
                  theme(plot.title = element_text(color = "darkred", face="bold"))
+
                  theme(axis.title = element_text(face="bold")) +
                  theme(legend.title = element_text(face="bold"))
```

2.3 Loading the dataset in R

Loading the dataset, which is in csv format, into R and assign the dataframe to “accident”.

```
accident <- read.csv("us_data_2000.csv", header = T, sep = ",", stringsAsFactors =
TRUE)
```

2.4 Quick overview on the dataset

2.4.1 head( )

Use head( ) to have a quick overview on the first 6 observations and all the column names in the dataset.

```
head(accident)
```

From the first glance of the dataset, we can see all the columns are in capital letters. Since R is case-sensitive, for easy processing and readability, we update all the column names from capital letters to lowercases.

```
names(accident) <- tolower(names(accident))
```

From the first 6 observations in the dataset, we can already notice there may be some invalid values or missing values need to further investigate in further analysis, such as:

- Unusual symbols (“?”):
  There are few columns, such as “ug_location_cd”, “ug_mining_method_cd” and “mining_equip_cd”, have unusual symbol (“?”) for data input. Referring to the data dictionary, these columns shall be some codes assigned to indicating different location, mining methods, mining equipment when accidents occurred.
- Invalid values (“NO VALUE FOUND”):
  This value is mostly found in the columns having written description to the coding columns as mentioned in the above, such as “ug_location”, “ug_mining_method” and “mining_equip”. When “?” is found in the coding columns, there will be “NO VALUE FOUND” in the description columns.
- Unknown values (NA):
  There are some NA data observed in several columns, such as “closed_doc_no” and “tot_exper”.
- Missing values (empty cells):
  There are some rows containing empty cells in the dataset. We have to consider if we have to remove these rows for our further data analysis or impute with other estimates to ensure the completeness of data input for detailed data analysis afterwards.

2.4.2 str( )

Use str( ) to further analyse the dataset, especially the data types of every column.

We can see there are 2,000 observations with 57 variables in total, comprised by 39 factors, 13 integer, 5 numeric variables. However, there are some of the columns are having inappropriate data type, causing their data do not seems to be reasonable or meaningful in the dataset, such as:

- Columns related to dates (“accident_dt” and “return_to_work_dt”):
  The dates input are treated as characters instead of dates.

- Columns related to time (“accident_time” and “shift_begin_time”):
  The time input are treated as integer instead of time.
- Columns related to document numbers (“document_no” and “closed_doc_no):
  The document numbers are treated as numbers without assigned as factors, so they are presented in scientific notation format, given that the document numbers are too long. This caused most of the document numbers are having the same values as shown, however,”document_no” is a unique key for every accident.
- Columns which are categorical numerical variables (“mine_id”, subunit_cd”, “cal_yr”, “cal_qtr”, “fiscal_yr”,“fiscal_qtr”, “fips_state_cd” and “coal_mine_ind”):
  There are no factors assigned to these categorical numberical variables.
- Column having too many levels (“narrative”):
  Since when we load the dataset into R, we have set stringasfactor = TRUE, so R will find all the strings data as factors. Since narrative on accidents are characters and different in every accident, there are 2000 levels (one per observation) have been noted from this column.

Besides, there are some columns are describing the same items but in either numerical (codes) or strings (words), such as “controller_id” and “controller_name”, “ug_location_cd” and “ug_location”, and “ug_mining_method_cd” and “ug_mining_method”. Given that they are representing the same items, their total number of levels should be the same. However, some columns are found to have mismatch in the number of levels, such as “controller_id” (612 levels) and “controller_name” (607 levels), “operator_id” (782 levels) and “operator_name” (874 levels), and “occupation_cd” (110 levels) and “occupation” (89 levels). This may indicate there might be missing levels in some of the columns or duplicate codes assigned to different items. Further investigations will be conducted on the difference level numbers in later stage.

2.4.3 summary( )

Use summary to analyse the data

Apart from the aforementioned invalid values (NA and “NO VALUE FOUND”), missing values and inappropriate data types leading to unreasonable or not meaningful statistics, there are others observations from the summary regarding to:

- Unreasonable input:

There are some input which seem to be unreasonable, such as the maximum values noted in “accident_time” and “shift_begin_time” are 9999. Since this column is representing time, the range of value shall be within 0001 and 2400. In order to further consider the treatment of these unreasonable time, we have to see the quantity of these value in the dataset.

i. accident_time

```
table_large_accident_time <- table(accident$accident_time[accident$accident_time>2
400])
names(dimnames(table_large_accident_time)) <- c("accident_time")
table_large_accident_time
```

```
## accident_time
## 9996 9997 9998 9999
##    1    1    3  134
```

For time values outside the normal time range in “accident_time”, we can see there are 1, 1, 3, and 134 observations having time values of 9996, 9997, 9998 and 9999 respectively. Quantity of time value 9996 to 9998 are only 5, which is considered to be insufficient to the whole dataset, and may be omitted for the time being. However, for the time values of 9999, they contribute for about 6.7% of the whole population. We cannot remove all these rows, we have to investigate if this is a sentinel value. However, by looking into the dataset, it seems there is no obvious pattern noted for the meaning of having time values 9999. We cannot conclude this is a sentinel value in this stage.

ii. shift_begin_time

```
table_large_shift_begin_time <- table(accident$shift_begin_time[accident$shift_beg
in_time>2400])
names(dimnames(table_large_shift_begin_time)) <- c("shift_begin_time")
table_large_shift_begin_time
```

```
## shift_begin_time
## 9999
##   85
```

There are only 85 unreasonable time values of 9999 found in “shift_begin_time”, which is fewer than what we notice in the “accident_time”. We may check if they are the same rows of “accident_time” having unreasonable time values.

```
accident[accident$shift_begin_time>2400, c("shift_begin_time", "accident_time")]
```

```
##      shift_begin_time accident_time
## 50               9999          9999
## 100              9999          9999
## NA                 NA            NA
## 203              9999          9999
## 238              9999          1315
## 240              9999          1400
## 267              9999          9999
## 270              9999          9999
## 285              9999          9999
## 303              9999          1100
## NA.1               NA            NA
## NA.2               NA            NA
## 412              9999          1000
## 421              9999           900
## 425              9999          9999
## 427              9999          1145
## 432              9999          1100
```

```
## 438              9999          1500
## 462              9999           800
## 465              9999          9999
## 486              9999          1500
## 496              9999           715
## 534              9999           915
## NA.3               NA            NA
## 585              9999          9999
## 667              9999          1130
## 668              9999          2100
## 670              9999          9999
## 673              9999          9999
## 682              9999          9999
## NA.4               NA            NA
## 741              9999          9999
## 780              9999          9999
## 783              9999          9999
## 833              9999          9999
## 864              9999          9999
## 877              9999          2000
## 887              9999          9999
## 891              9999          9999
## 901              9999           530
## 906              9999          9999
## 917              9999          1302
## 933              9999          2320
## 963              9999          9999
## 994              9999          9999
## 1014             9999          9999
## 1053             9999           700
## 1054             9999          9999
## 1105             9999          9999
## NA.5               NA            NA
## 1124             9999          9997
## 1133             9999          9999
## 1134             9999          2345
## 1153             9999          1000
## 1154             9999          1945
## 1156             9999          9999
## 1175             9999          1030
## 1236             9999          1740
## 1254             9999          9999
## NA.6               NA            NA
## 1361             9999          9999
## 1368             9999          9999
## 1371             9999          9999
## 1387             9999          1600
## 1447             9999          9999
## 1452             9999          9999
## 1453             9999          9999
```

```
## 1457            9999          9999
## 1466            9999          1315
## 1483            9999          2030
## 1508            9999          9999
## 1547            9999          1500
## 1624            9999           600
## 1626            9999          9999
## 1628            9999           710
## 1636            9999          9999
## 1642            9999          1700
## 1652            9999          1801
## 1672            9999           745
## 1720            9999          9999
## 1721            9999          9999
## 1722            9999          9999
## 1734            9999          9999
## 1765            9999           630
## 1775            9999          1708
## 1776            9999           715
## 1777            9999          1310
## NA.7              NA            NA
## 1822            9999          9999
## 1823            9999          9999
## NA.8              NA            NA
## 1896            9999          9999
## 1904            9999          9999
## 1949            9999          9999
```

From the above summary, they may not be in the same columns, and there maybe spread of unreasonable time values.

- Meaningless grouping:
  From the “equip_mfr_name”, we can see about 85% of the observation lie in the group of NO VALUE FOUND, Not reported, Not on this list and Not listed, which carry no valuable information to the data analysis. We may consider if this column is essential to the data analysis and more data has to be collected in later stage.
- Outliers:
  For “no_injuries”, “tot_exp”, “mine_exp”, “job_exp”, “schedule_charge”, “day_restrict” and “day_lost”, their maximum values are significantly greater than the value of Q3, median and mean. This may indicate these columns are right-skewed and the maximum values are outliers. We may plot boxplots later to identify outliers of these columns if they are essential to the data analysis.

2.4.4 is.na( )

Lastly, we have to analyse how many NAs are there in each column in order to consider the best way to handle NAs in the dataset.

```
apply(is.na(accident), 2, sum)
```

```
##             mine_id       controller_id     controller_name         operator_id
##                   0                   0                   0                   0
##       operator_name       contractor_id         document_no          subunit_cd
##                   0                   0                   0                   0
##             subunit         accident_dt              cal_yr             cal_qtr
##                   0                   0                   0                   0
##           fiscal_yr          fiscal_qtr       accident_time    degree_injury_cd
##                   0                   0                   0                   0
##       degree_injury       fips_state_cd      ug_location_cd         ug_location
##                   0                   0                   0                   0
## ug_mining_method_cd    ug_mining_method     mining_equip_cd        mining_equip
##                   0                   0                   0                   0
##        equip_mfr_cd      equip_mfr_name      equip_model_no    shift_begin_time
##                   0                   0                   1                   9
##   classification_cd      classification    accident_type_cd       accident_type
##                   0                   0                   0                   0
##         no_injuries           tot_exper          mine_exper           job_exper
##                   0                 360                 333                 328
##       occupation_cd          occupation         activity_cd            activity
##                   0                   0                   0                   0
##    injury_source_cd        injury_source    nature_injury_cd       nature_injury
##                   0                   0                   0                   0
##    inj_body_part_cd        inj_body_part     schedule_charge       days_restrict
##                   0                   0                 673                 543
##           days_lost          trans_term   return_to_work_dt     immed_notify_cd
##                 422                   0                   0                   0
##        immed_notify     invest_begin_dt           narrative       closed_doc_no
##                   0                   0                   0                1138
##      coal_metal_ind
##                   0
```

Most of the columns do not contain NA. However, there are some columns, especially for those related to experience, schedule charge, days and closed document have significant amount of NA. We will further investigate the reasons and consider treatments to those NA.

3. Initial transformations

3.1 Inappropriate data types:

According to the above findings, we will convert some columns to more appropriate data types.

```
accident_clean <- within(accident, {

  # Assign the numerical categorical variables to factors
  mine_id <- as.factor(mine_id)
  cal_yr <- as.factor(cal_yr)
  cal_qtr <- as.factor(cal_qtr)
  fiscal_yr <- as.factor(fiscal_yr)
  fiscal_qtr <- as.factor(fiscal_qtr)
  subunit_cd <- as.factor(subunit_cd)
  coal_metal_ind <- as.factor(coal_metal_ind)

  # Convert narrative to characters
  narrative <- as.character(narrative)

  # Remove scientific notation format from the Document numbers
  document_no <- format(document_no, scientific = FALSE)
  closed_doc_no <- format(closed_doc_no, scientific = FALSE)
})
```

3.2 Invalid values:

For the aforementioned invalid values (missing values, unreasonable input, NO VALUE FOUND, NA), we will assign them to NA at current stage. We may assign them to better estimates in Project 2.

```
accident_clean_v2 <- accident_clean %>% mutate_all(na_if,"?")
accident_clean_v2 <- accident_clean_v2 %>% mutate_all(na_if, "NO VALUE FOUND")
accident_clean_v2 <- accident_clean_v2 %>% mutate_all(na_if, "NOT MARKED")
accident_clean_v2 <- accident_clean_v2 %>% mutate_all(na_if, "")
```

3.3 Parsing and converting dates and times:

```
accident_clean_v2 <- within(accident_clean_v2, {

  # Change date columns to date format from characters
  accident_dt <- as.Date(accident_dt, format = "%d/%m/%Y")
  return_to_work_dt <- as.Date(return_to_work_dt, format = "%m/%d/%Y")
  invest_begin_dt <- as.Date(invest_begin_dt, format = "%m/%d/%Y")

  # Change time columns to time format from characters
  accident_time <- formatC(accident_time, width = 4, format = "d", flag = "0")
  accident_time[accident_time > 2400] <- NA           # Convert unreasonable time
values (9996 - 9999) to NA
  accident_time_format <- strptime(accident_time, format = "%H%M", tz = "GMT")
  accident_time_format <- as.POSIXlt(accident_time_format)

  shift_begin_time <- formatC(shift_begin_time, width = 4, format = "d", flag = "0
")
  shift_begin_time[shift_begin_time > 2400] <- NA     # Convert unreasonable time
values (9996 - 9999) to NA
  shift_begin_time_format <- strptime(shift_begin_time, format = "%H%M", tz = "GMT
")
  shift_begin_time_format <- as.POSIXlt(shift_begin_time_format)
})
```

3.4 Adding new columns and variables:

Adding new columns and variables that might by useful for further analysis in later stage.

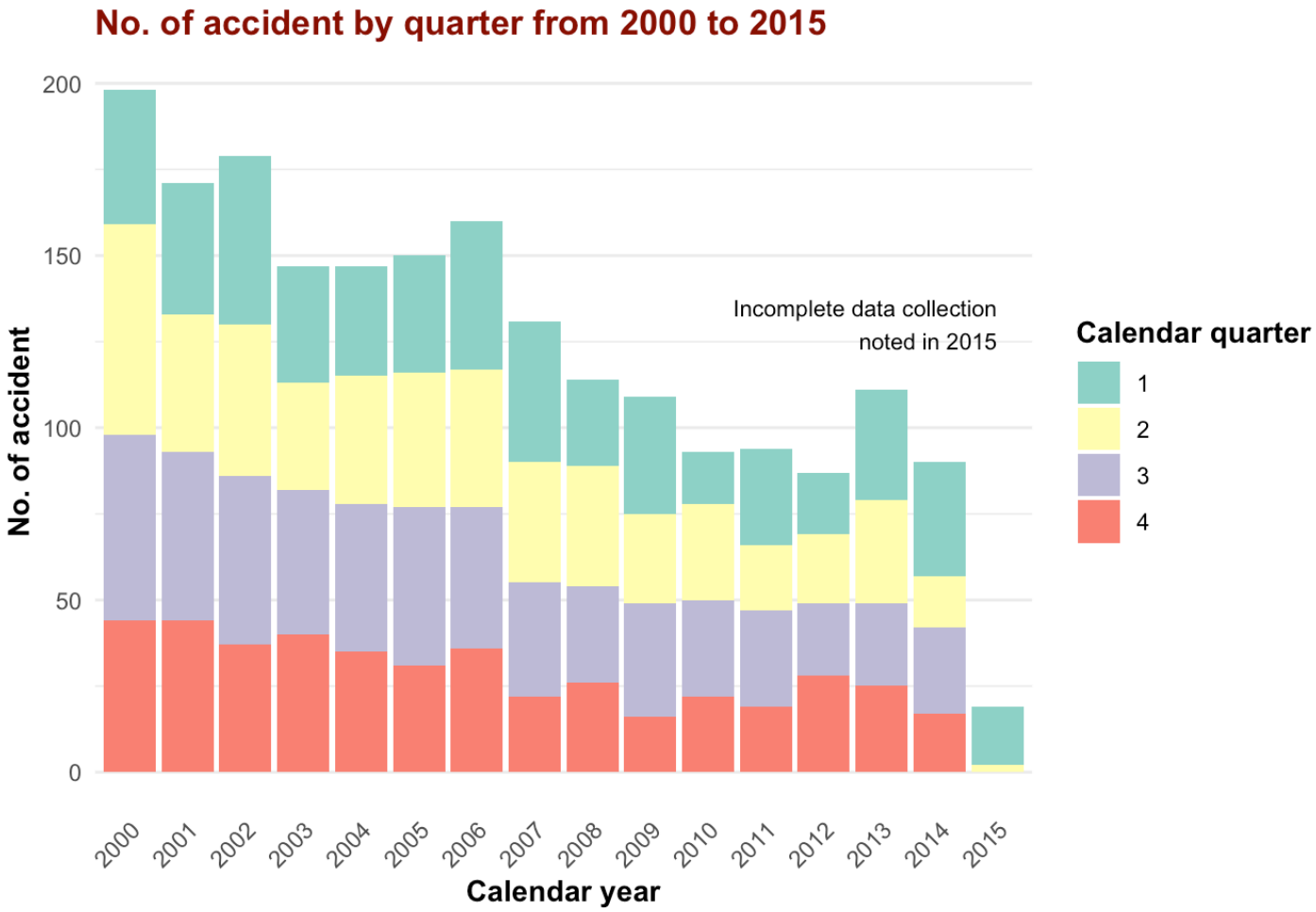
```
accident_clean_v2 <- within(accident_clean_v2, {

  # Adding new variables
  worktime_before_accident <- ifelse(
    as.numeric(difftime(accident_time_format, shift_begin_time_format, units = "ho
urs")) < 0,
    as.numeric(difftime(accident_time_format, shift_begin_time_format, units = "ho
urs")+24),
    as.numeric(difftime(accident_time_format, shift_begin_time_format, units = "ho
urs")))

  tot_exp_years <- NA
  tot_exp_years[tot_exper >= 0 & tot_exper < 10] <- "0 - 10 years"
  tot_exp_years[tot_exper >= 10 & tot_exper < 20] <- "10 - 20 years"
  tot_exp_years[tot_exper >= 20 & tot_exper < 30] <- "20 - 30 years"
  tot_exp_years[tot_exper >= 30 & tot_exper < 40] <- "30 - 40 years"
  tot_exp_years[tot_exper >= 40 & tot_exper < 50] <- "40 - 50 years"

  days_affected <- days_lost + days_restrict

  days_to_resume_work <- difftime(return_to_work_dt, accident_dt, tz , units = "da
ys")

  # Convert to factor from string type
  tot_exp_years <- factor(tot_exp_years)
})
```

4. Analysing the data

4.1 Analysing the timing of accident occurance

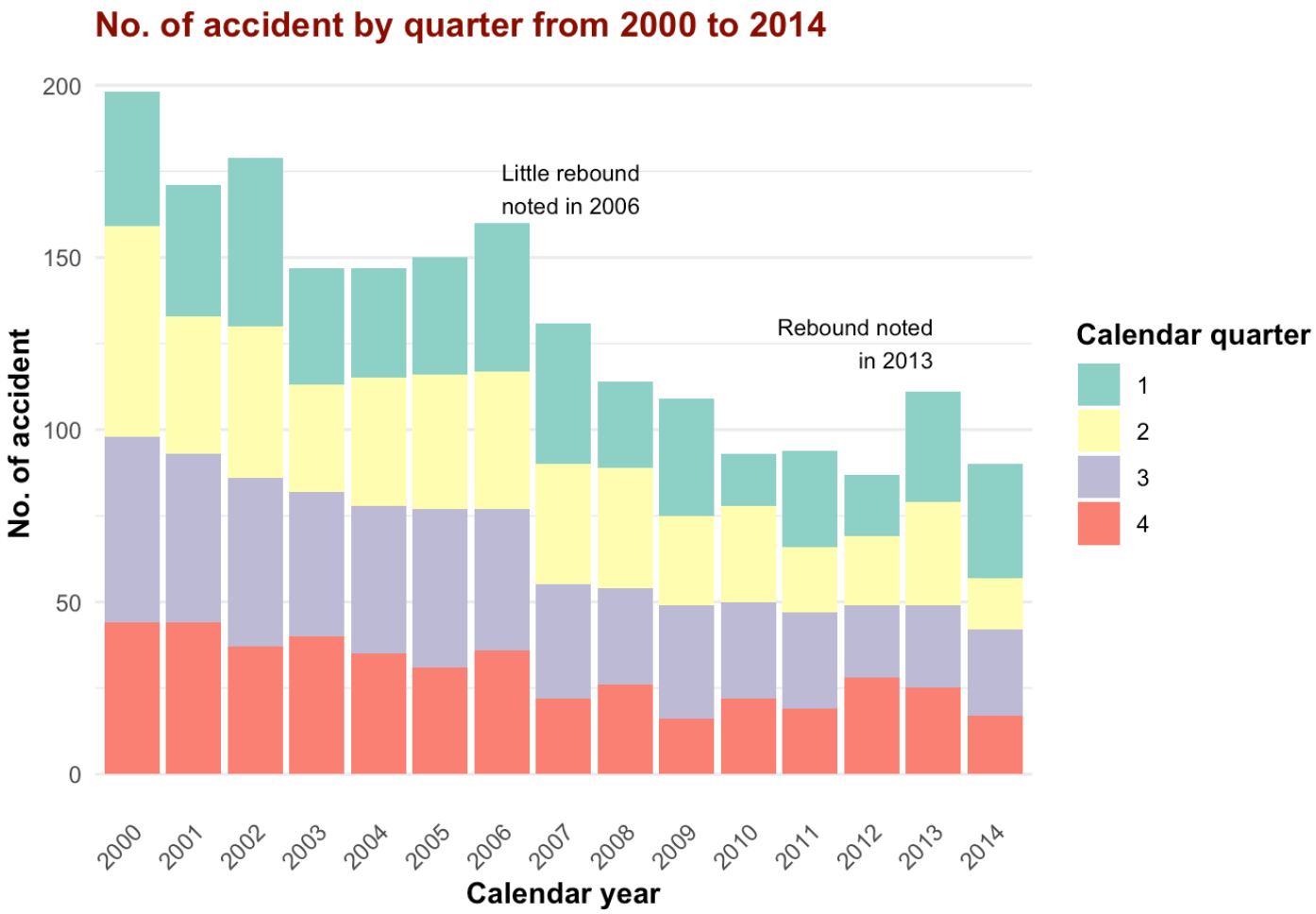
4.1.1 Analysis by year by quarter

```
ggplot(data = accident_clean_v2, aes(x=cal_yr, fill = cal_qtr)) +
  geom_bar() +
  labs(title = "No. of accident by quarter from 2000 to 2015", x = "Calendar year"
, y = "No. of accident", fill = "Calendar quarter") +
  cits4009_theme +
  theme(plot.title = element_text(size=13),
        axis.text.x = element_text(angle= 45, hjust=1),
        aspect.ratio = 0.8) +
  scale_fill_brewer(palette="Set3") +
  annotate("text", x = "2015", y =130, label = paste("Incomplete data collection",
"noted in 2015", sep = "\n"), size = 3, hjust = 1)
```



Though there was a significant drop accident cases in 2015, the data collected for 2015 is likely to be incomplete. There is no data collect for 2015Q3 and 2015Q4. Thus, for like-to-like comparison among the time frame (i.e. full year range), we will omit 2015 in our data analysis related to date.

```
accident_clean_v5 <- accident_clean_v2[!(accident_clean_v2$cal_yr == "2015"),]

ggplot(data = accident_clean_v5, aes(x=cal_yr, fill = cal_qtr)) +
  geom_bar() +
  labs(title = "No. of accident by quarter from 2000 to 2014", x = "Calendar year"
, y = "No. of accident", fill = "Calendar quarter") +
  cits4009_theme +
  theme(plot.title = element_text(size=13),
        axis.text.x = element_text(angle= 45, hjust=1),
        aspect.ratio = 0.8) +
  scale_fill_brewer(palette="Set3") +
  annotate("text", x = "2006", y =170, label = paste("Little rebound", "noted in 2
006", sep = "\n"), size = 3, hjust = 0) +
  annotate("text", x = "2013", y =125, label = paste("Rebound noted", "in 2013", s
ep = "\n"), size = 3, hjust = 1)
```



From the above chart, we can observe there is an overall decreasing trend in the number of accidents from 2000 (~200 cases) to 2014 (~100 cases). However, there was a little rebound noted during 2006 and 2013. While in terms of quarter, there is no significant seasonality pattern observed. The number of accidents seems distributed evenly in every quarter.

4.1.2 Analysis by accident time

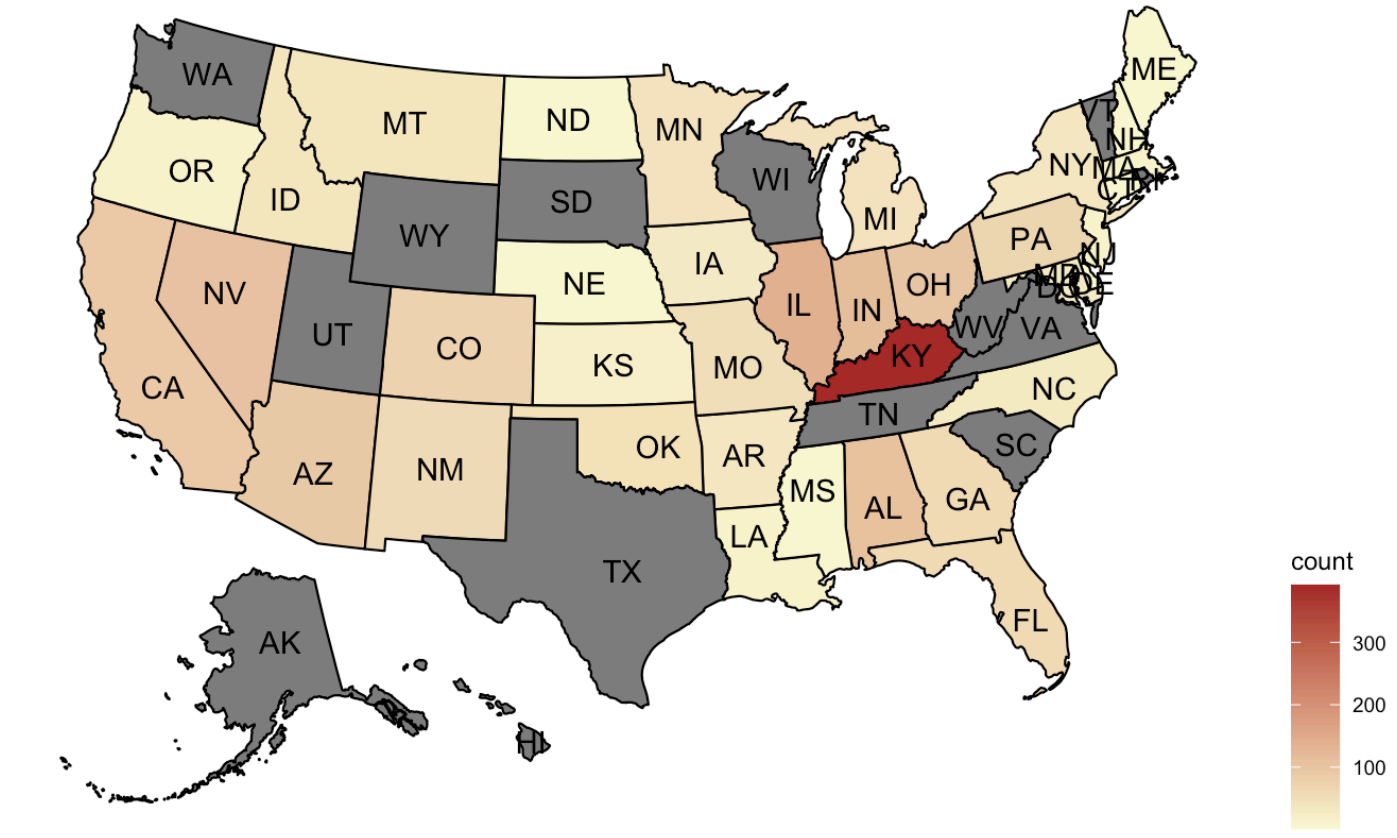
```
accident_time_table <- table(accident_clean_v2$accident_time)
accident_time_table.df <- as.data.frame(accident_time_table)
colnames(accident_time_table.df) <- c("accident_time", "count")

accident_time_table.df2 <- accident_time_table.df %>%
  mutate(accident_hour <- substr(accident_time, start = 1, stop = 2))
colnames(accident_time_table.df2) <- c("accident_time", "count", "accident_hour")
accident_time_table.df2$accident_hour <- as.factor(accident_time_table.df2$acciden
t_hour)

# Grouping the accident time by hours
accident_time_table.df3 <- aggregate.data.frame(accident_time_table.df2$count,list
(accident_time_table.df2$accident_hour), FUN = sum)
colnames(accident_time_table.df3) <- c("accident_hour", "count")
accident_time_table.df3$accident_hour <- paste(accident_time_table.df3$accident_ho
ur, ": 00")
accident_time_table.df3$accident_hour <- as.factor(accident_time_table.df3$acciden
t_hour)

# Plotting the line chart
ggplot(accident_time_table.df3, aes(x = accident_hour, y = count, group = 1)) +
  geom_rect(data=NULL, aes(xmin="07 : 00", xmax="17 : 00", ymin=-Inf, ymax=Inf), f
ill="lightyellow") +
  geom_point(color = "chocolate") +
  geom_line(color = "darkred")  +
  labs(title = "No. of accident by hours", x = "Accident time (in Hour)", y = "No.
of accident") +
  cits4009_theme +
  theme(plot.title = element_text(size=13),
        axis.text.x = element_text(angle= 45, hjust=1, size = 8),
        aspect.ratio = 0.8) +
  annotate("text", x = "07:00", y = 0, label = paste("Daytime"), size = 3, hjust =
0) +
  annotate("text", x = "10:00", y =190, label = paste("Peak noted", "at 10am", sep
= "\n"), size = 3, hjust = 0) +
  annotate("text", x = "11:00", y =55, label = paste("Significant drop noted", "at
12noon due to", "lunch hour", sep = "\n"), size = 3, hjust = 0.5) +
  annotate("text", x = "15:00", y =140, label = paste("2nd Peak noted", "at 2pm",
sep = "\n"), size = 3, hjust = 0) +
  annotate("text", x = "06:00", y =120, label = paste("Most of the cases", "were o
ccured in", "daytime", sep = "\n"), size = 3, hjust = 1)
```

From the above chart, we can observe that accidents were more likely occurred during daytime (highlighted in yellow) as compared to night time. Numbers of accident peaked at 10am and 2pm for nearly 200 cases and 140 cases respectively. There is a significant drop observed in between the two peaks at 12 noon, which is likely due to lunch hour and much fewer workers worked in that period of time. We will further analyse the data by looking in how long did the workers work before the accident happened in the next section.

4.2 Analysing the locations of accidents occurance

4.2.1 Analysis by US states

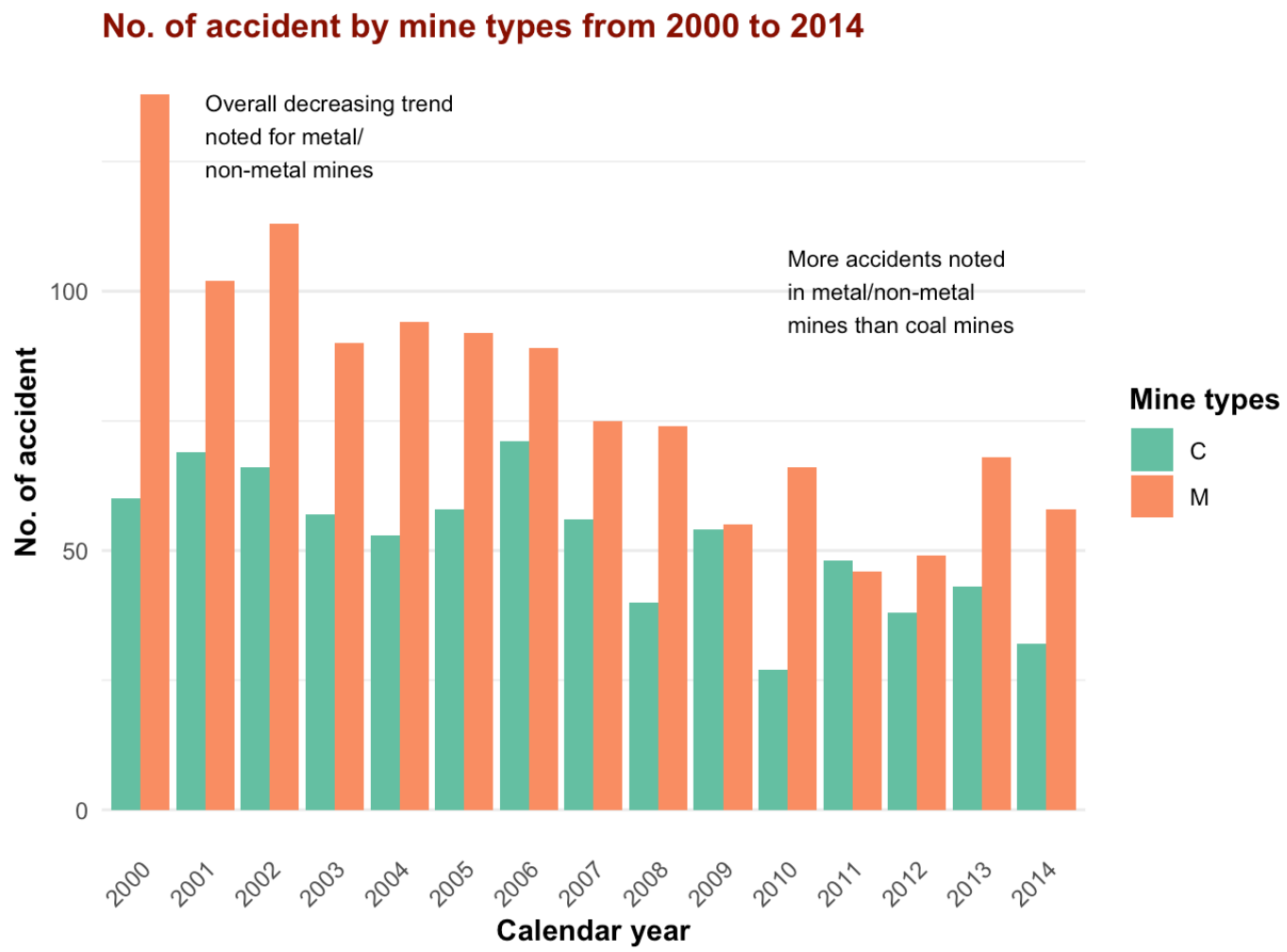
```
us_mine_map_table <- table(accident_clean_v2$fips_state_cd)
us_mine_map_df <- as.data.frame(us_mine_map_table)
colnames(us_mine_map_df) <- c("fips_code", "count")

# Convert fips_code to state names
us_mine_map_df$fips_code <- as.character(us_mine_map_df$fips_code)
fips_name <- fips_info(us_mine_map_df$fips_code)
us_mine_map_df2 <- cbind(us_mine_map_df, fips_name)

# Plot map
plot_usmap(data = us_mine_map_df2, value = "count", labels = TRUE) +
  ggtitle("No. of accident by US states") +
  theme(plot.title = element_text(size = 11,color = "darkred", face="bold"),
        legend.position = "right") +
  scale_fill_gradient(low = "lightgoldenrodyellow", high = "brown")
```



From the above map, we can see KY has the most accidents reported for over about 300 cases. We will further look down to specific mines.

4.2.2 Analysis by mine types

```
ggplot(data = accident_clean_v5, aes(x=cal_yr, fill = coal_metal_ind)) +
  geom_bar(position = "dodge") +
  labs(title = "No. of accident by mine types from 2000 to 2014", x = "Calendar ye
ar", y = "No. of accident", fill = "Mine types") +
  cits4009_theme +
  theme(plot.title = element_text(size=12.5),
        axis.text.x = element_text(angle= 45, hjust=1), aspect.ratio = 0.8) +
  scale_fill_brewer(palette="Set2") +
  annotate("text", x = "2001", y = 130, label = paste("Overall decreasing trend ",
"noted for metal/ ", "non-metal mines", sep = "\n"), size = 3, hjust = 0) +
  annotate("text", x = "2010", y = 100, label = paste("More accidents noted", "in
metal/non-metal", "mines than coal mines", sep = "\n"), size = 3, hjust = 0)
```

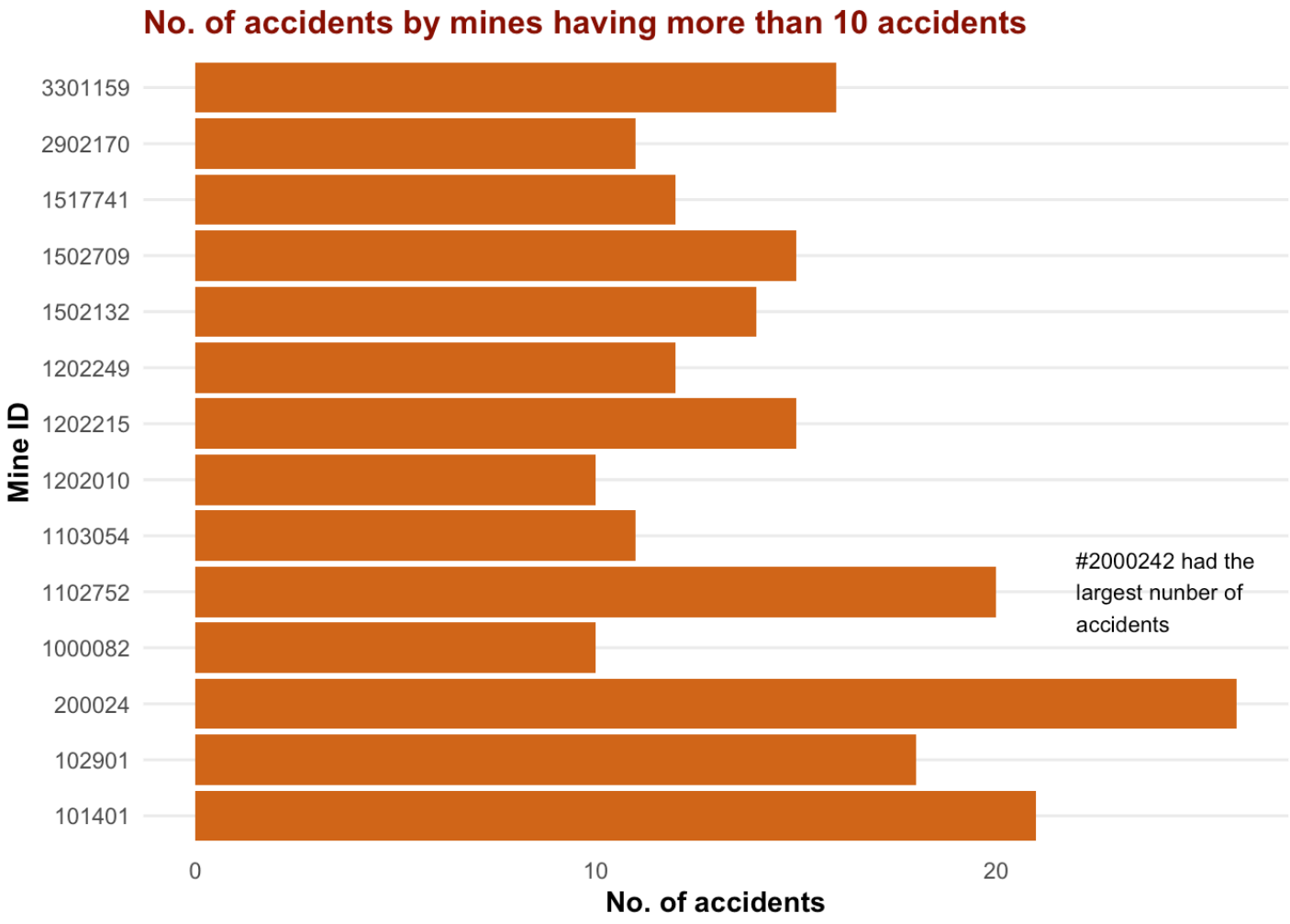


From the above chart, we can see there were more accidents occurred in metal/ non-metal mines as compared to coal mine almost all the years, except 2011. In term of number of accidents, there is a decreasing trend noted in metal/ non-metal mines throughout the years, from about 190 cases in 2000 decreased to about 60 cases in 2014. While for coal mines, there is no significant pattern can be observed for coal mines throughout the year, with the number of accidents fluctuated around 50 cases.

4.2.3 Analysis by mine ID

Since there were too many mines (over 1600 noted) had ever experienced accidents during the time frame, we have extracted those having accidents more frequently for our analysis.

```
mine_id_table <- table(accident_clean_v2$mine_id, accident_clean_v2$fips_state_cd)
mine_id_df <- as.data.frame(mine_id_table)
colnames(mine_id_df) <- c("mine_id", "fips_code"  ,"count")
mine_id_sort_df <- mine_id_df %>% arrange(desc(count))

# Extract the mines having more than 10 accidents
mine_id_over10 <- subset(mine_id_sort_df, mine_id_sort_df$count >= 10)
mine_id_over10_extract <- accident_clean_v2 %>% filter(mine_id %in% mine_id_over10
$mine_id)

# Plot the bar chart
ggplot(mine_id_over10_extract, aes(x = mine_id)) +
  geom_bar(fill = "chocolate") +
  labs(title = "No. of accidents by mines having more than 10 accidents", y = "No.
of accidents", x = "Mine ID") +
  coord_flip() +
  cits4009_theme +
  theme(plot.title = element_text(size=12.5)) +
  annotate("text", y = 22, x = "1102752", label = paste("#2000242 had the", "large
st nunber of", "accidents", sep = "\n"), size = 3, hjust = 0)
```



There are 14 mines had over 10 accidents from 2000 to 2014, while mine#200024 had the greatest number of accident, which is 26, followed by mine#101401 and mine#1102752 having 21 cases and 20 cases

respectively.

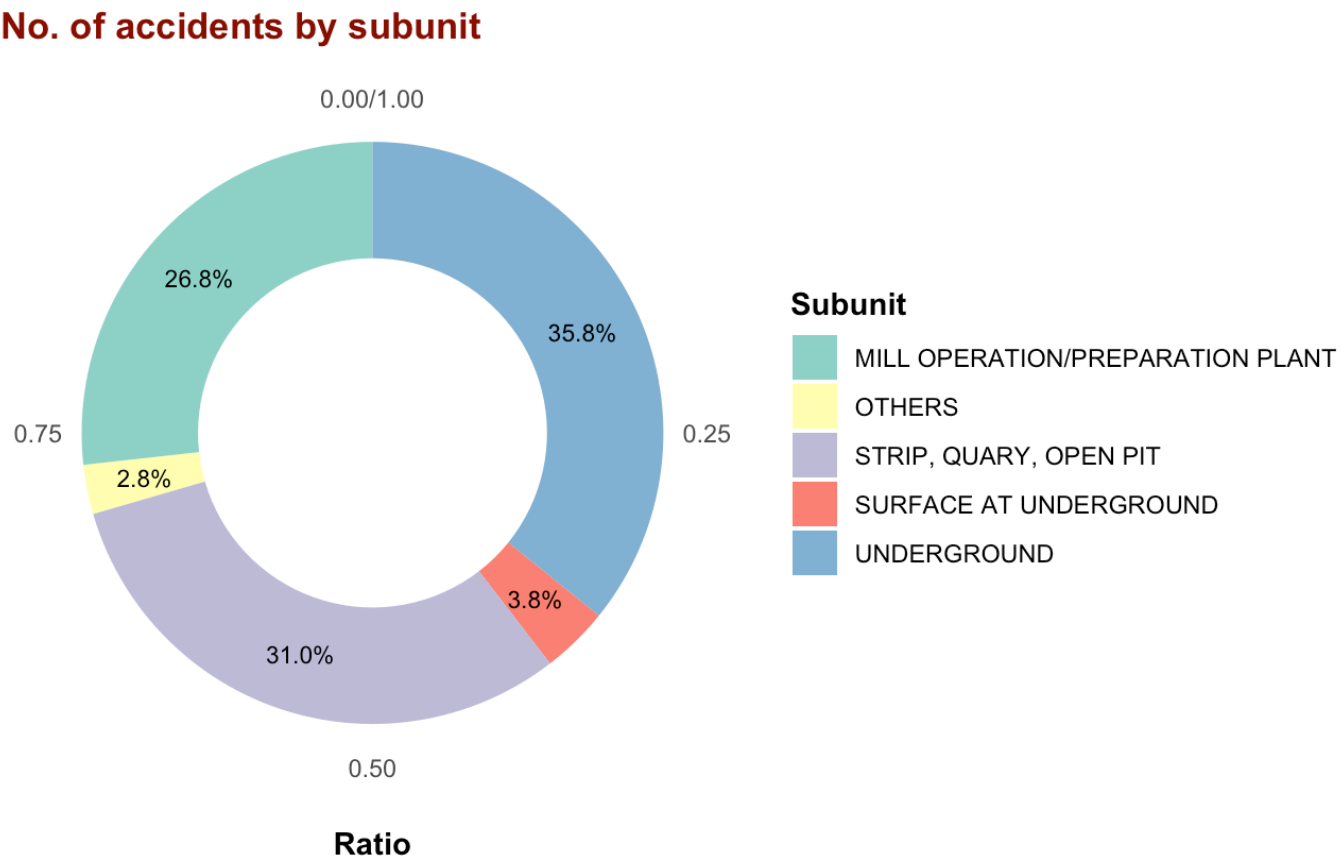
4.2.4 Analysis by locations within mines

```
subunit_df <- data.frame(accident_clean_v2$subunit)
colnames(subunit_df) <- "subunit"

# Calculation the % proportion of subunit
subunit_df2 <- subunit_df %>%
  group_by(subunit) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))
subunit_df2
```

```
## # A tibble: 9 × 4
##   subunit                            n   perc labels
##   <fct>                          <int>  <dbl> <chr>
## 1 AUGER                              1 0.0005 0.05%
## 2 CULM BANK/REFUSE PILE              1 0.0005 0.05%
## 3 INDEPENDENT SHOPS OR YARDS         5 0.0025 0.25%
## 4 OFFICE WORKERS AT MINE SITE        9 0.0045 0.45%
## 5 DREDGE                            39 0.0195 1.95%
## 6 SURFACE AT UNDERGROUND            75 0.0375 3.75%
## 7 MILL OPERATION/PREPARATION PLANT 535 0.268  26.75%
## 8 STRIP, QUARY, OPEN PIT           619 0.310  30.95%
## 9 UNDERGROUND                      716 0.358  35.80%
```

After calculating the proportion of subunit on total number of accidents, we can see there are few subunits having insignificant portions, such as auger, culm bank/ refuse pile, independent shops or yards, office workers at mine site and dredge, which aggregated portion is 2.8%. As a result, we will group these subunit to one category, “Others”.

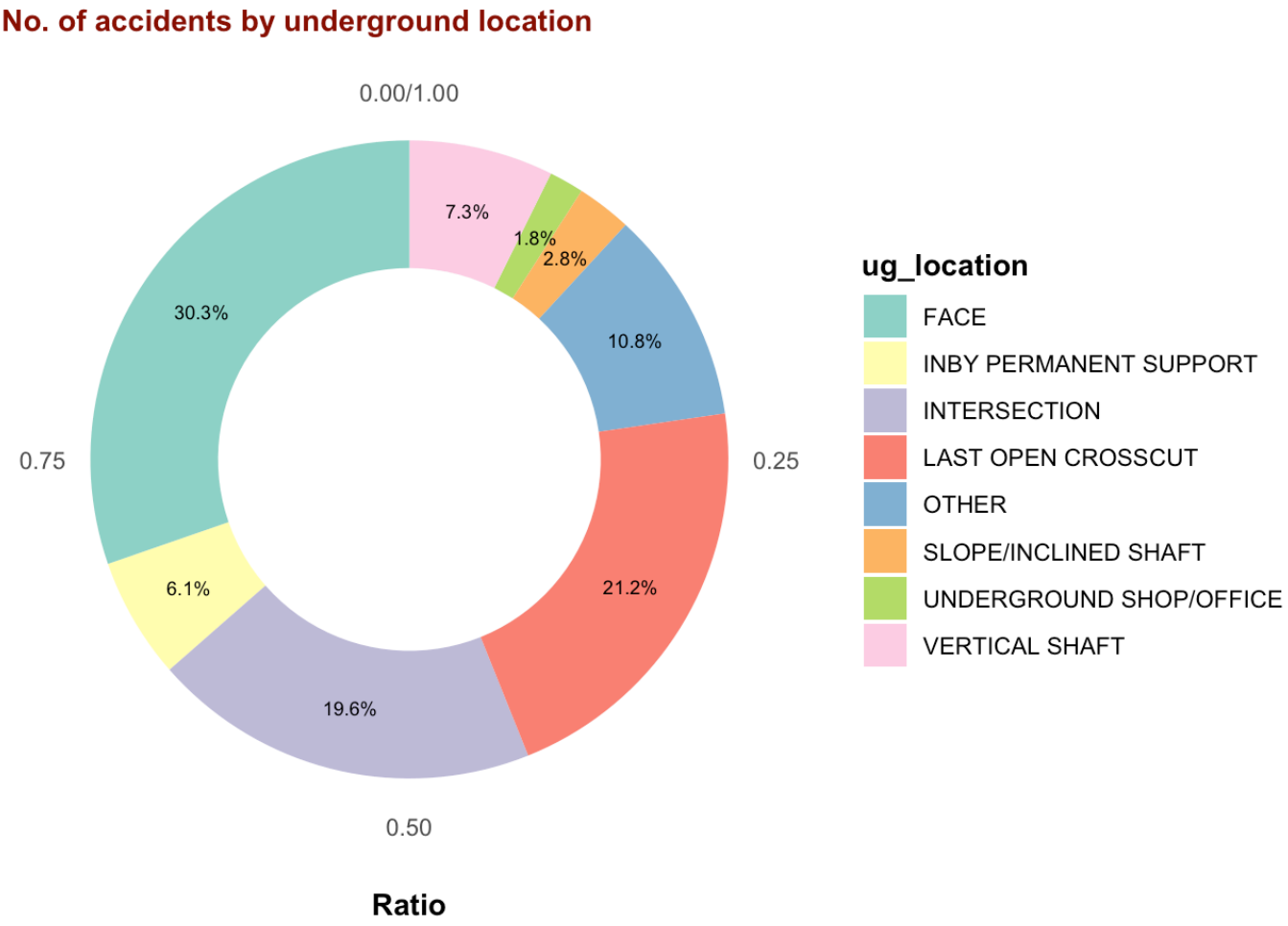
```
subunit_other <- c("AUGER", "CULM BANK/REFUSE PILE", "INDEPENDENT SHOPS OR YARDS",
"OFFICE WORKERS AT MINE SITE", "DREDGE")
subunit_df3 <- subunit_df
subunit_df3$subunit <- as.character(subunit_df3$subunit)
subunit_df3$subunit[subunit_df3$subunit %in% subunit_other] <- "OTHERS"
subunit_df3$subunit <- as.factor(subunit_df3$subunit)

# Recalculation the % proportion of subunit
subunit_df3 <- subunit_df3 %>%
  group_by(subunit) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

# Plot donut chart
hsize <- 2
subunit_df3 <- subunit_df3 %>% mutate(x = hsize)

ggplot(subunit_df3, aes(x= hsize, y = perc, fill=subunit)) +
  geom_col() +
  labs(title = "No. of accidents by subunit", fill = "Subunit", y = "Ratio", x = "
") +
  coord_polar("y") +
  xlim(c(0.2, hsize + 0.5)) +
  cits4009_theme +
  theme(plot.title = element_text(size = 13), legend.position = "right", axis.text
.y = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = eleme
nt_blank()) +
  scale_fill_brewer(palette="Set3") +
  geom_text(aes(label = labels), position = position_stack(vjust = 0.5), size = 3)
```



From the above chart, we can observe underground contributed the greatest portion of 36% in the total number of accident, followed by strip, quary, open pit of 31% and mill operation/ preparation plant of 27%. We will conduct an in-depth analysis on where the accidents happened in the underground in next stage.

4.2.5 Analysis by underground location

```
ug_location_omit_na <- na.omit(accident_clean_v2$ug_location)
ug_location_extract.df <- as.data.frame(ug_location_omit_na)
ug_location_extract_clean.df <- subset(ug_location_extract.df, ug_location_omit_na
!= "NOT MARKED")
colnames(ug_location_extract_clean.df) <- "ug_location"

# Calculation the % proportion of ug_location
ug_location_extract_clean.df2 <- ug_location_extract_clean.df %>%
  group_by(ug_location) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

# Plot donut chart
ug_location_extract_clean.df2 <- ug_location_extract_clean.df2 %>% mutate(x = hsiz
e)

ggplot(ug_location_extract_clean.df2, aes(x= hsize, y = perc, fill=ug_location)) +
  geom_col() +
  labs(title = "No. of accidents by underground location", legend = "Underground l
ocation", y = "Ratio", x = "") +
  coord_polar("y") +
  xlim(c(0.2, hsize + 0.5)) +
  scale_fill_brewer(palette="Set3") +
  cits4009_theme +
  theme(plot.title = element_text(size = 11), legend.position = "right", axis.text
.y = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = eleme
nt_blank()) +
  geom_text(aes(label = labels), position = position_stack(vjust = 0.5), size = 2.
5)
```

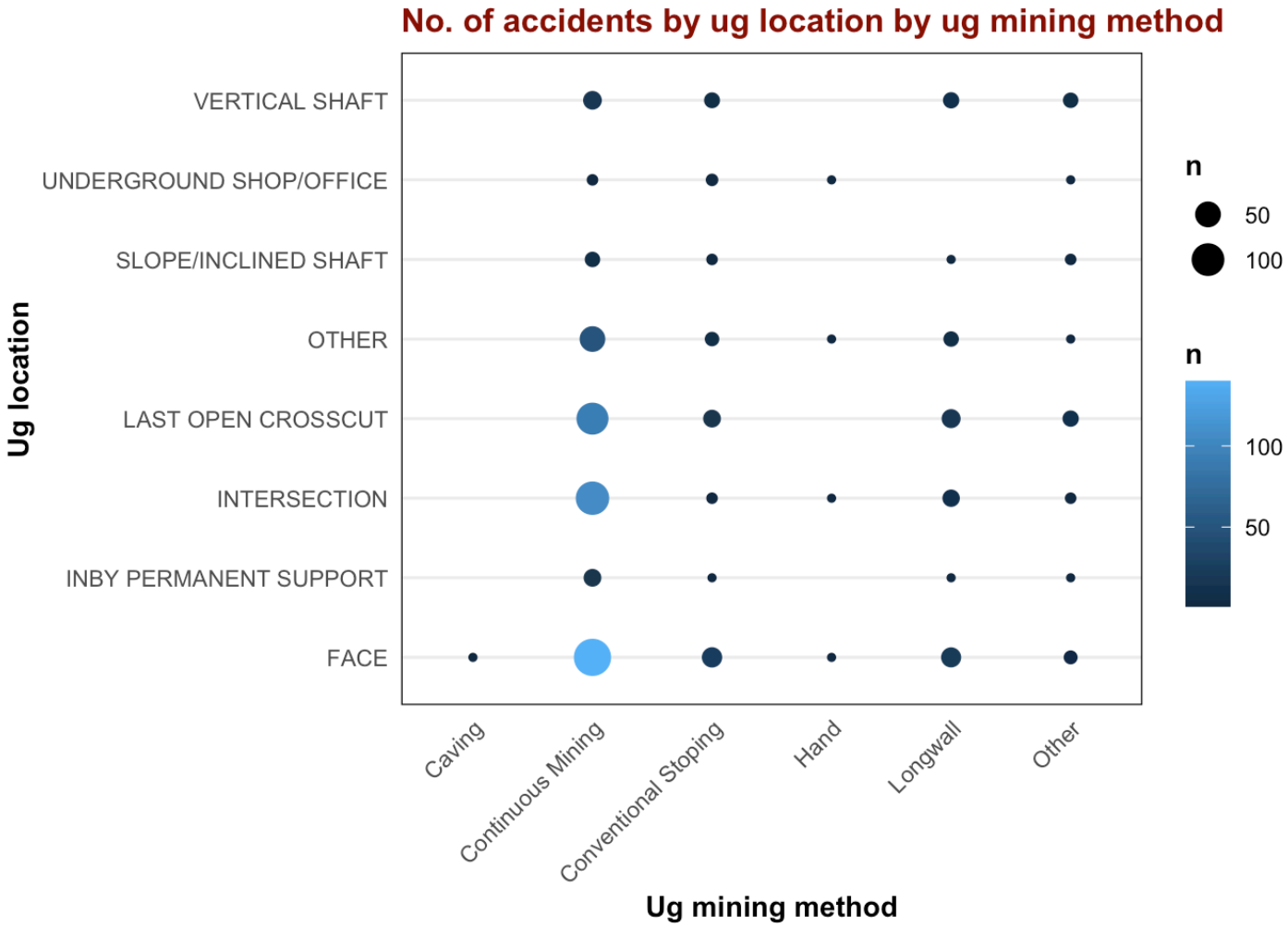


From the above chart, we can see that 30% of accidents were happened in the face of the mine, following by last open crosscut and intersection, which consist of 21% and 20%.

Further analysis by underground mining methods by underground location

Further analysis on which types of underground mining methods have caused the largest number of accidents in the underground.

```
ug_location_method_df <- data.frame(accident_clean_v2$ug_location, accident_clean_
v2$ug_mining_method)
colnames(ug_location_method_df) <- c("ug_location", "ug_mining_method")
ug_location_method_clean_df <- subset(ug_location_method_df, !is.na(ug_location) &
!is.na(ug_mining_method))

ggplot(ug_location_method_clean_df, aes(x = ug_mining_method, y = ug_location)) +
  geom_count(aes(color = ..n..)) +
  labs(title = "No. of accidents by ug location by ug mining method", x = "Ug mini
ng method", y = "Ug location") +
  cits4009_theme +
  theme(plot.title = element_text(size = 12.5), axis.text.x = element_text(angle=
45, hjust=1),  panel.border = element_rect(colour = "gray4", fill=NA, size=0.5))
```



From the above count plot, we can see Continuous Mining had caused the most accidents in all different underground locations, which may indicate this mining method is likely to be more risky as compared to other mining methods.

4.3 Analysing the types of accident

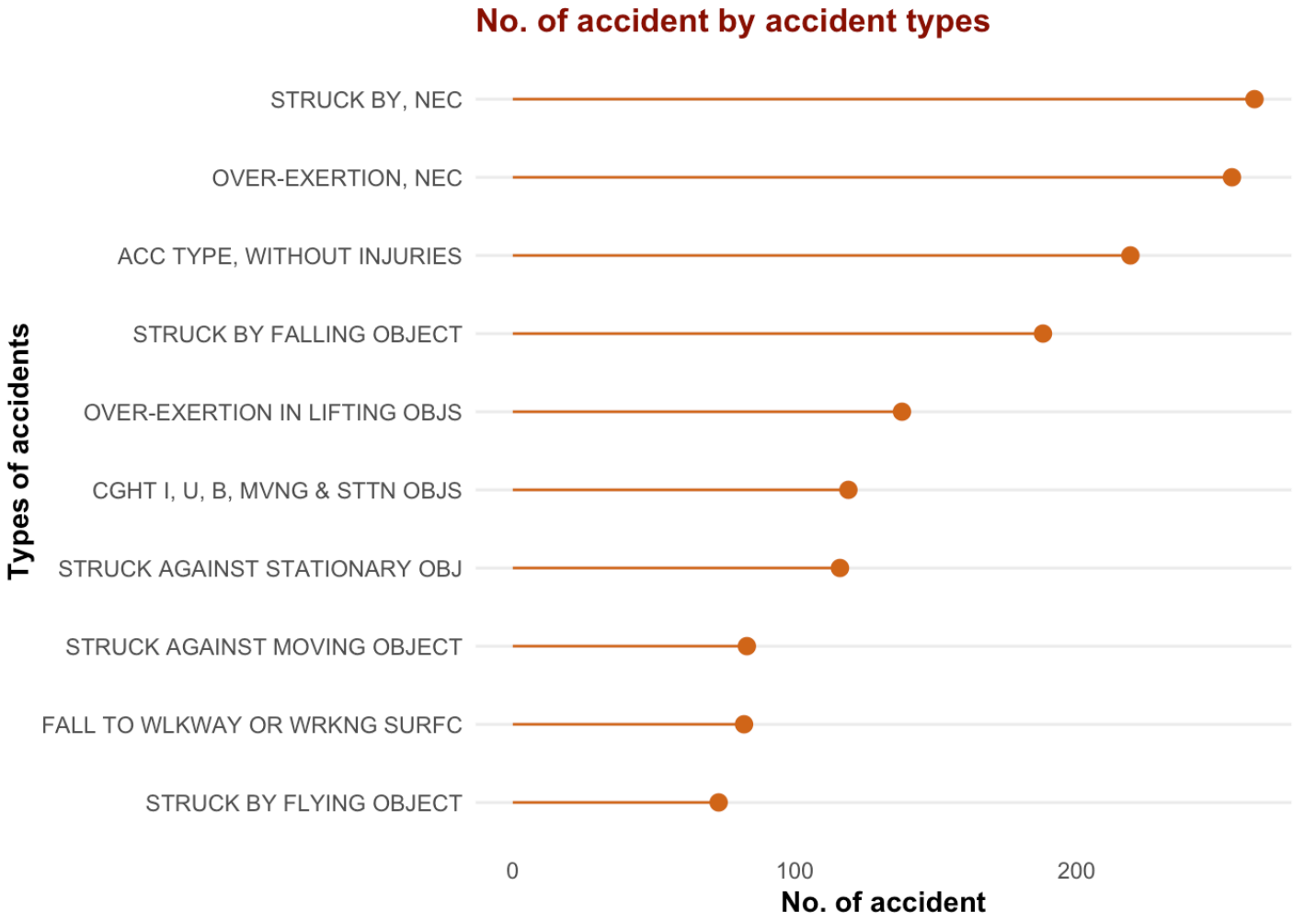
4.3.1 Analysis the frequency of each accident types

```
accident_type_table <- table(accident_clean_v2$accident_type)
accident_type_df <- data.frame(accident_type_table)
colnames(accident_type_df) <- c("accident_type", "count")
accident_type_df <- accident_type_df %>% arrange(desc(count))
nlevels(accident_clean_v2$accident_type)
```

```
## [1] 37
```

Since there are 37 levels of accident types in the dataset, while some of the levels only have insignificant contribution to the whole population, we will extract the top 10 accident types for analysis.

```
accident_type_top10_df <- head(accident_type_df, 10)
accident_type_top10_extract <- accident_clean_v2 %>% filter(accident_type %in% acc
ident_type_top10_df$accident_type)

ClevelandDotPlot(accident_type_top10_extract, "accident_type", sort = 1, title = "
No. of accident by accident types", color = "chocolate") +
  labs(x = "Types of accidents", y = "No. of accident") +
  coord_flip() +
  cits4009_theme +
  theme(plot.title = element_text(size=12.5))
```



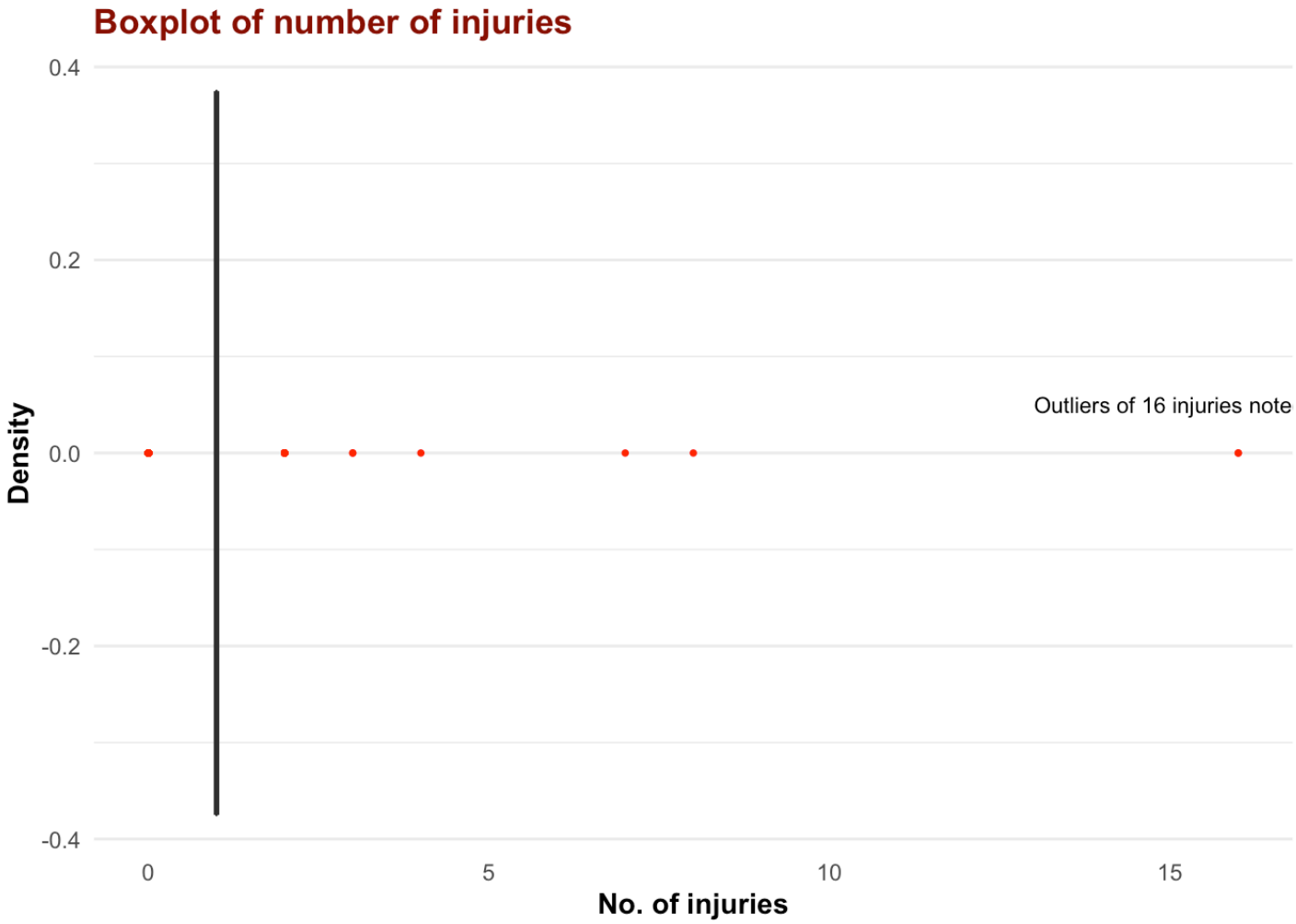
From the above chart, we can see the top 5 accident types contribute almost 50% of the total population. We will then analyse different accident types against the number of injuries to see if there is specific accident types causing more injuries.

4.3.2 Analysis each accident types to number of injuries per accidents

```
summary(accident_clean_v2$no_injuries)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   0.923   1.000  16.000
```

```
ggplot(data = accident_clean_v2) +
  geom_boxplot(aes(x = no_injuries), outlier.colour = "red", outlier.shape = 16, o
utlier.size = 1, notch = FALSE) +
  labs(title = "Boxplot of number of injuries", x = "No. of injuries", y = "Densit
y") +
  cits4009_theme +
  annotate("text", x = 13, y = 0.05, label = paste("Outliers of 16 injuries noted"
), size = 3, hjust = 0)
```



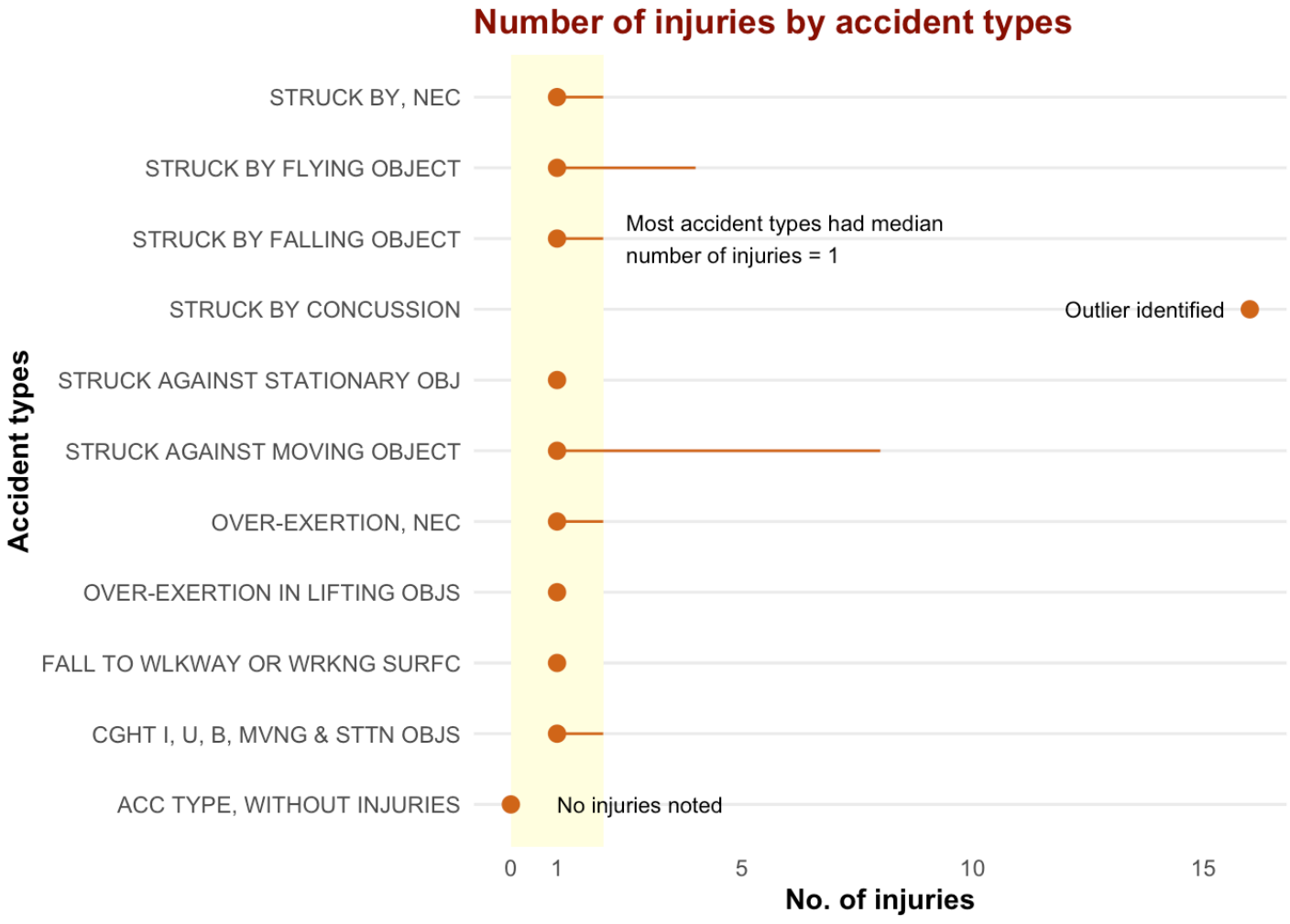
From the above analysis, we can see the number of injuries is right-skewed, having most of the observations are 1 and there are 6 outliers with large number of injuries. The highest number in injuries found in one single accident is 16. Accident type of this accident is worth to analyse as well. Since we are analysing the top 10 accident typesp previously, we have to ensure we have included this accident type (causing 16 injuries) in our anlysis.

```
# Identify the accident type of the outlier
accident_clean_v2$accident_type[accident_clean_v2$no_injuries == 16]
```

```
## [1] STRUCK BY CONCUSSION STRUCK BY CONCUSSION
## 37 Levels: ABSRTN RAD CAUST TXC & NOX SBS ... UNCLASSIFIED, INSUFFICENT DATA
```

The accident type causing 16 injuries is “STUCK BY CONCUSSION”. Since this accident type is not included in the top 10 accident type list, we have to include it in the dataframe for further analysis.

```
accident_type_extract <- accident_clean_v2 %>% filter(accident_type %in% accident_
type_top10_df$accident_type | accident_type == "STRUCK BY CONCUSSION")

ggplot(accident_type_extract) +
  geom_rect(data=NULL, aes(ymin= 0 , ymax= 2, xmin=-Inf, xmax=Inf), fill="lightyel
low") +
  stat_summary(aes(x = accident_type, y = no_injuries), fun.min = min, fun.max = m
ax, fun = median, color = "chocolate") +
  coord_flip() +
  cits4009_theme +
  labs(title = "Number of injuries by accident types", y = "No. of injuries", x =
"Accident types") +
  scale_y_continuous(breaks = c(0,1,5,10,15)) + annotate("text", x = "STRUCK BY FA
LLING OBJECT", y = 2.5, label = paste("Most accident types had median", "number of
injuries = 1", sep = "\n"), size = 3, hjust = 0) + annotate("text", x = "ACC TYPE,
WITHOUT INJURIES", y = 1, label = paste("No injuries noted"), size = 3, hjust = 0)
+ annotate("text", x = "STRUCK BY CONCUSSION", y = 12, label = paste("Outlier iden
tified"), size = 3, hjust = 0)
```



From the above chart, we can see most of the accident types having the median of number of injuries to be 1 only, except for “STRUCK BY CONCUSSION” and “ACC TYPE, WITHOUT INJURIES”. As noted in the summary above as well, number of injuries in “STRUCK BY CONCUSSION” is an outlier, consists of mean of 6 injuries in the accidents. Another outlier noted in this chart is “STRUCH AGAINST MOVING OBJECT”, the maximum number of injuries noted in accident is 8. As for accident type of “ACC TYPE, WITHOUT INJURIES”, there is no people injured, which match with the caption. Overall, these accidents seems not in large scales, except for the outliers we identified.

4.4 Analysis on the reasons of accident

4.4.1 Analysis on the workers

On this section, we will analysis on the affected workers, seriousness on accidents in terms of injuries and their recovery. The accidents reported which no injuries will not be applicable in these sections, so we will exclude all the row having no injuries.

```
accident_clean_v3 <- subset(accident_clean_v2, no_injuries != 0)
```

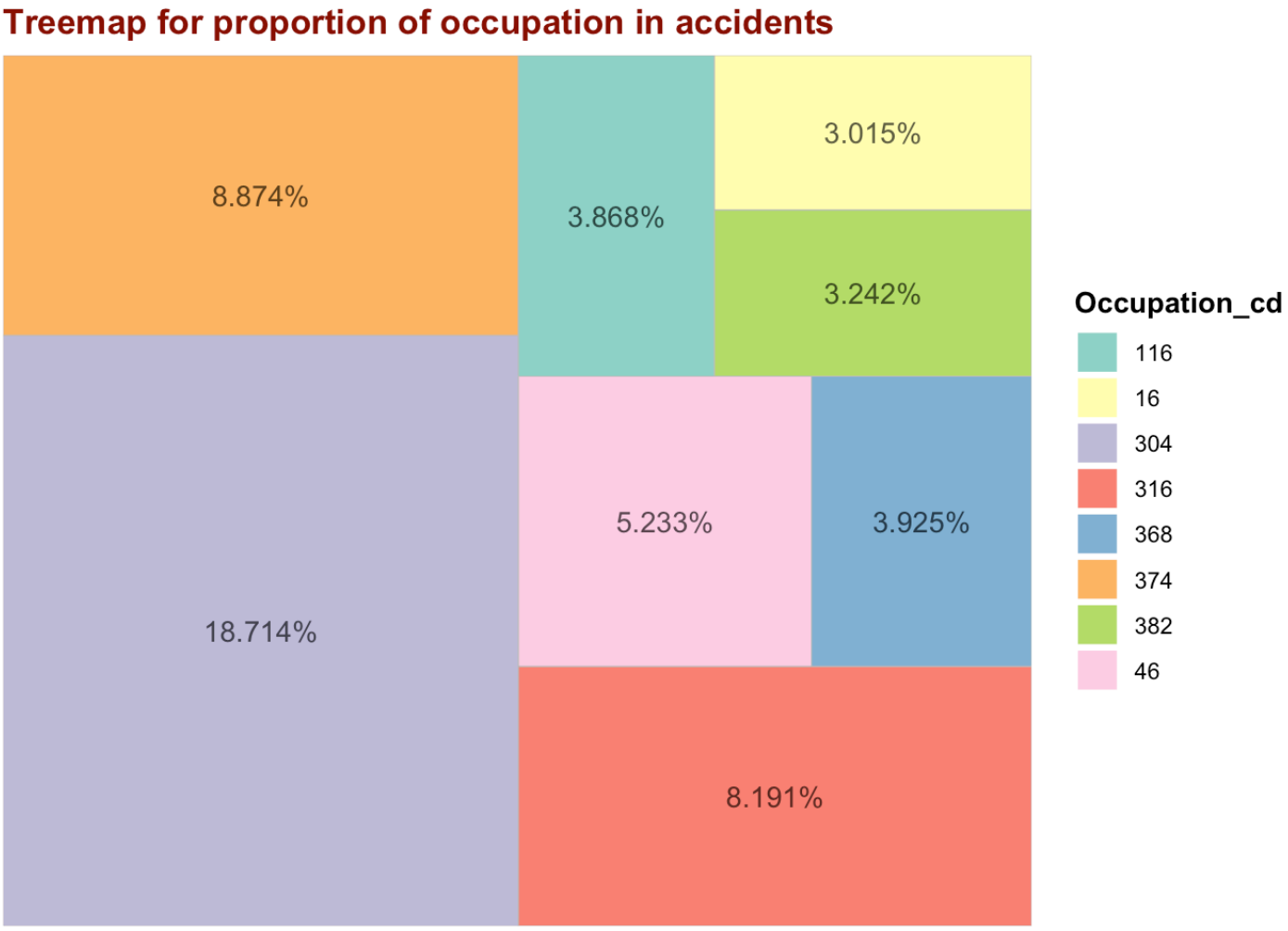
(i) Analysis on the occupation of workers

We will analysis if there are any relationship between occupation and occurrence of accident. Since the occupation descriptions are too lengthy, we will use occupation_cd in the analysis for readability.

```
occupation_df <- data.frame(accident_clean_v3$occupation_cd)
colnames(occupation_df) <- "occupation_cd"
occupation_df <- subset(occupation_df, occupation_cd != "?")

# Calculate the portion of occupation
occupation_df <- occupation_df %>%
  group_by(occupation_cd) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

# Extract occupation which contribute over 3%
occupation_extract_df <- subset(occupation_df, perc >= 0.03)
```

Since there are 108 occupations codes in the dataset, however most of them contribute less than 1% of total population. There are 7 levels contributed more than 3% of the total population, we will focus on them for the occupation analysis.

```
ggplot(occupation_extract_df, aes(fill = occupation_cd, area = n, label = labels))
+
  geom_treemap() +
  geom_treemap_text(place = "center", size = 11 , alpha = 0.7) +
  labs(title = "Treemap for proportion of occupation in accidents", fill = "Occupa
tion_cd") +
  cits4009_theme +
  scale_fill_brewer(palette="Set3")
```



From the above, we can see occupation_cd #304 has the highest portion of 18.7%, followed by occupation_cd #374 and #316 contributing 8.9% and 8.2% respectively. These occupations seems more risky to have accidents than the others.

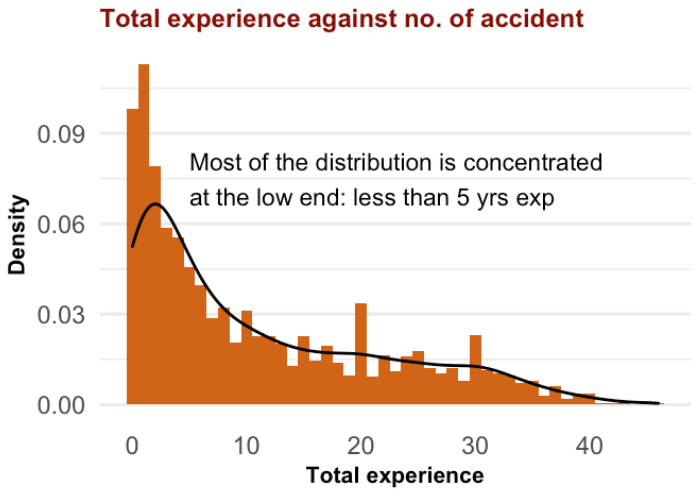
Matching back the occupation_cd to occupation description in the dataset, the occupations are: - occupation_cd #304 are Maintenance man, Mechanic, Repair/Serviceman, Boilermaker, Fueler, Tire tech, Field service tech; - occupation_cd #374 are Warehouseman, Bagger, Palletizer/Stacker, Store keeper, Packager, Fabricator, Cleaning plant operator; and - occupation_cd #316 are Laborer, Blacksmith, Bull gang, Parts runner, Roustabout, Pick-up man, Pitman

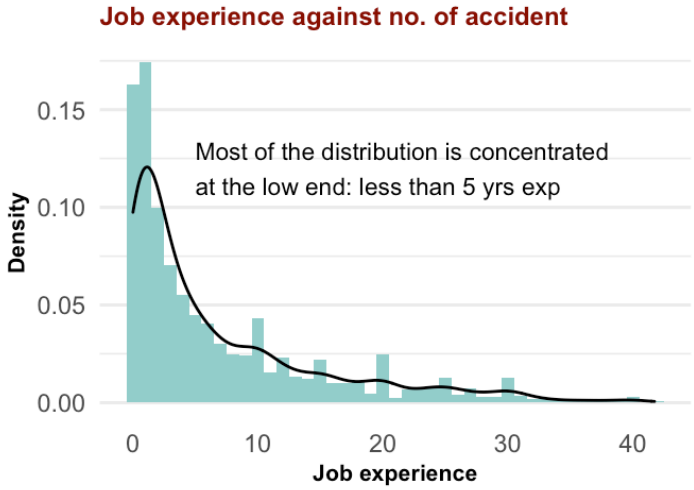
(ii) Analysis by working experience

```
pairs(~ tot_exper + mine_exper + job_exper, accident_clean_v3, main = "Working exp
erience against no. of accidents", col = "chocolate")
```

We can see from the above pair plot, there are strong positive linear relationship among total experience, mining experience and job experience of the affected workers. Thus, we may expect similar pattern from these 3 variables when analysing against the number of accidents and other variables.

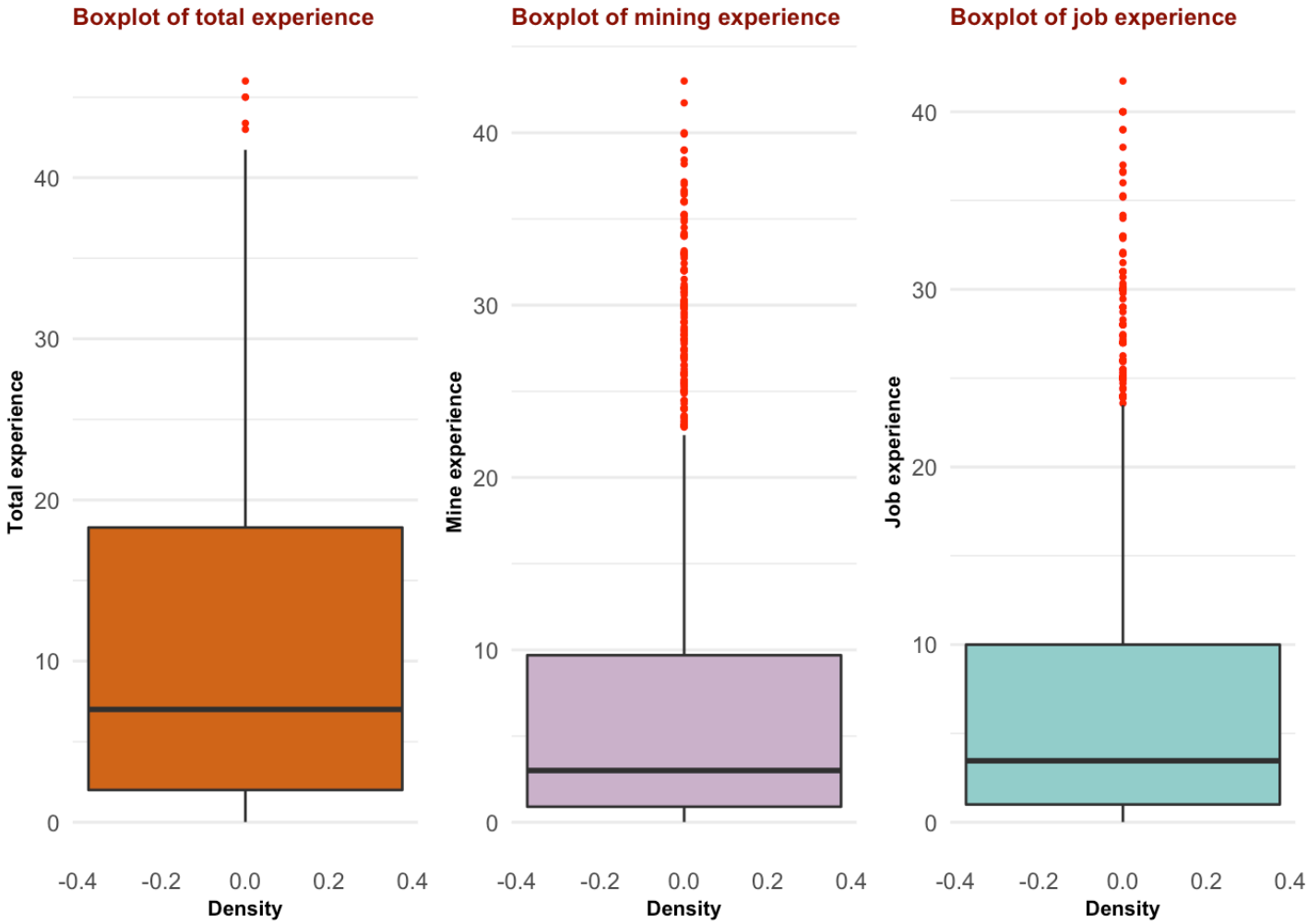
```
accident_tot_exp_hist <- ggplot(data=accident_clean_v3, aes(x = tot_exper, y=..den
sity..)) + geom_histogram(binwidth = 1, fill = "chocolate") + geom_density(color =
"black") + labs(title = "Total experience against no. of accident", x = "Total exp
erience", y = "Density") + cits4009_theme + theme(plot.title = element_text(size=9
), axis.title = element_text(size = 8)) + annotate("text", x = 5, y = 0.075, label
= paste("Most of the distribution is concentrated", "at the low end: less than 5 y
rs exp", sep="\n"), size = 3, hjust = 0)

accident_mine_exp_hist <- ggplot(data=accident_clean_v3, aes(x = mine_exper, y = .
.density..)) + geom_histogram(binwidth = 1, fill = "thistle3") + geom_density(colo
r = "black") + labs(title = "Mining experience against no. of accident", x = "Mini
ng experience", y = "Density") + cits4009_theme + theme(plot.title = element_text(
size=9), axis.title = element_text(size = 8)) + annotate("text", x = 5, y = 0.135,
label = paste("Most of the distribution is concentrated", "at the low end: less th
an 5 yrs exp", sep="\n"), size = 3, hjust = 0)

accident_job_exp_hist <- ggplot(data=accident_clean_v3, aes(x = job_exper, y = ..d
ensity..)) + geom_histogram(binwidth = 1, fill = "paleturquoise3") + geom_density(
color = "black") + labs(title = "Job experience against no. of accident", x = "Job
experience", y = "Density") + cits4009_theme + theme(plot.title = element_text(siz
e=9), axis.title = element_text(size = 8)) + annotate("text", x = 5, y = 0.12, lab
el = paste("Most of the distribution is concentrated", "at the low end: less than
5 yrs exp", sep="\n"), size = 3, hjust = 0)

grid.arrange(accident_tot_exp_hist, accident_mine_exp_hist, accident_job_exp_hist,
ncol=2)
```





From the above histogram, we can see that all tot_exper, mine_exper and job_exper are right-skewed variables (similar distributions as we expected. We will look at the boxplot to identify the outliers.

```
tot_exp_outliers <- ggplot(data = accident_clean_v3) + geom_boxplot(aes(x = tot_ex
per), outlier.colour = "red", outlier.shape = 16, outlier.size = 1, notch = FALSE,
fill = "chocolate") + labs(title = "Boxplot of total experience", x = "Total exper
ience", y = "Density") + cits4009_theme + coord_flip() + theme(plot.title = elemen
t_text(size=9), axis.title = element_text(size = 8))

mine_exp_outliers <- ggplot(data = accident_clean_v3) + geom_boxplot(aes(x = mine_
exper), outlier.colour = "red", outlier.shape = 16, outlier.size = 1, notch = FALS
E, fill = "thistle3") + labs(title = "Boxplot of mining experience", x = "Mine exp
erience", y = "Density") + cits4009_theme + coord_flip() + theme(plot.title = elem
ent_text(size=9), axis.title = element_text(size = 8))

job_exp_outliers <- ggplot(data = accident_clean_v3) + geom_boxplot(aes(x = job_ex
per), outlier.colour = "red", outlier.shape = 16, outlier.size = 1, notch = FALSE,
fill = "paleturquoise3") + labs(title = "Boxplot of job experience", x = "Job expe
rience", y = "Density") + cits4009_theme + coord_flip() + theme(plot.title = eleme
nt_text(size=9), axis.title = element_text(size = 8))

grid.arrange(tot_exp_outliers, mine_exp_outliers, job_exp_outliers, ncol=3)
```

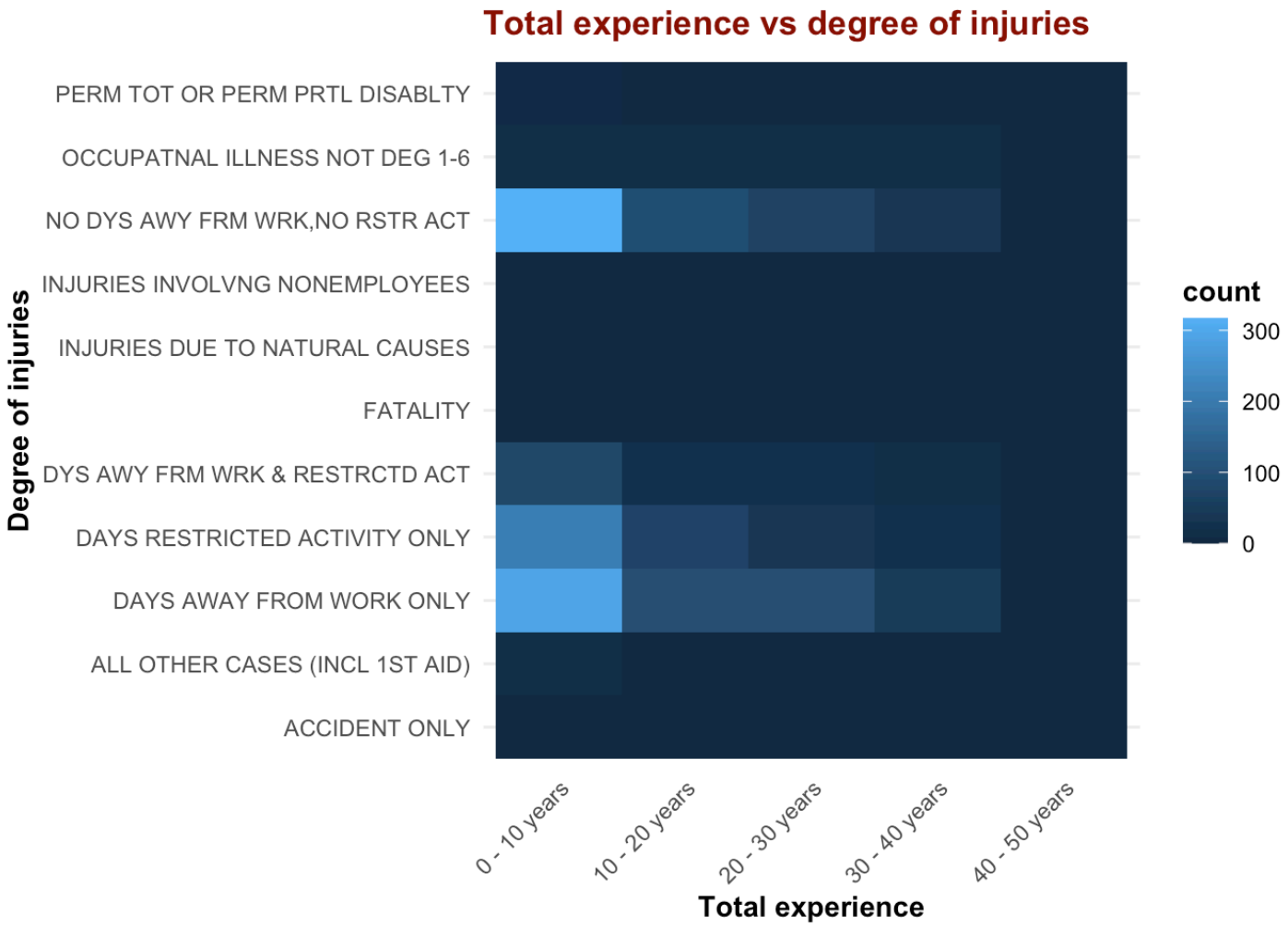


To sum up, we can observe that it was more likely for workers having less experience to have accidents, however, there were still very experienced workers (over 40 years experience) got accidents in their work.

Further analysis of experience against seriousness of injury

Since all tot_exper, mine_exper and job_exper are having similar distributions, we will use total_exper for further analysis.

```
exp_vs_injury_table <- table(accident_clean_v3$tot_exp_years, accident_clean_v3$de
gree_injury)
exp_vs_injury_df <- data.frame(exp_vs_injury_table)
colnames(exp_vs_injury_df) <- c("tot_exp_years", "degree_injury", "count")
exp_vs_injury_df <- subset(exp_vs_injury_df, exp_vs_injury_df$degree_injury != "NO
VALUE FOUND")

ggplot(exp_vs_injury_df, aes(x = tot_exp_years, y = degree_injury)) + geom_tile(ae
s(fill = count)) + labs(title = "Total experience vs degree of injuries", x = "Tot
al experience", y = "Degree of injuries") + cits4009_theme + theme(plot.title = el
ement_text(size=13), axis.text.x = element_text(angle= 45, hjust=1))
```



From the above, we can see the more experienced workers have fewer accidents and most of the accidents happened in the cohort of 0 - 10 years experience. Moreover, most of the affected workers can work again after certain day lost or restrict activity, instead of having permanent disability in their lives.

(iii) Analysis by working time

In order to analysis if there is relationship between working hours and probability of having accidents, a new variable “worktime_before_accident” has been created to show the number of working hours of the affected worker before the accident happened. “worktime_before_accident” has been calculated from the difference between the shift_begin_time and accident_time, assuming the accident was happened after the workers’ shift begin.

```
summary(accident_clean_v2$worktime_before_accident)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.500   4.500   4.873   7.000  23.917     185
```

From the summary, we can see the distribution of worktime_before_accident is right-skewed since the difference between Q3 and maximum observation is so significant, which indicates the existence of outliers. As for the maximum observation, the accident time was earlier than the shift begin time. By looking to the accident type, the accident was happened when workers was travelling to work.

```
which(accident_clean_v2$worktime_before_accident > 13)
```

```
##  [1]   51  176  247  262  466  478  788  810  835  858  985 1101 1242 1280 1460
## [16] 1477 1554 1581 1779 1921
```

```
ggplot(accident_clean_v2, aes(x = worktime_before_accident)) + geom_histogram(aes(
y=..density..), fill = "chocolate") + geom_density(color = "black") + labs(title =
"No. of accident by worktime before accident", x = "Working time before accident (
in hours)", y = "Density") + cits4009_theme + annotate("text", x = 17, y = 0.01, l
abel = paste("Outliers noted beyond 15 hours"), size = 3, hjust = 0) + annotate("t
ext", x = 8, y = 0.1, label = paste("Most of the distribution concentrate", "at th
e low end: fewer then 10 hours", sep = "\n"), size = 3, hjust = 0)
```

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

However, it is unreasonable for a worker to work for so long since normal working hours is 8 hours. From the above, we have try to identify the accidents having workers worked over 13 hours before the accident. In total, we have identified there are 18 observations (1% of total population). Since the portion is insignificant, we will exclude these rows as outliers.

```
worktime_before_accident_exl_outlier <- subset(accident_clean_v2, accident_clean_v
2$worktime_before_accident <= 13)

ggplot(worktime_before_accident_exl_outlier, aes(x = worktime_before_accident)) +
geom_histogram(aes(y=..density..), binwidth = 1, fill = "chocolate") + geom_densit
y(color = "black") + labs(title = "No. of accident by worktime before accident (ex
cl. outliers)", x = "Working time before accident (in hours)", y = "Density") + ci
ts4009_theme
```

From the above chart, we can see most of accidents were occurred within the workers' normal working hours and distributed quite evenly on every hour. Thus, the occurrence of accident is less likely due to the tiredness of workers or working overtime.

(iv) Further anaysis on the seriousness of injuries of the affected workers

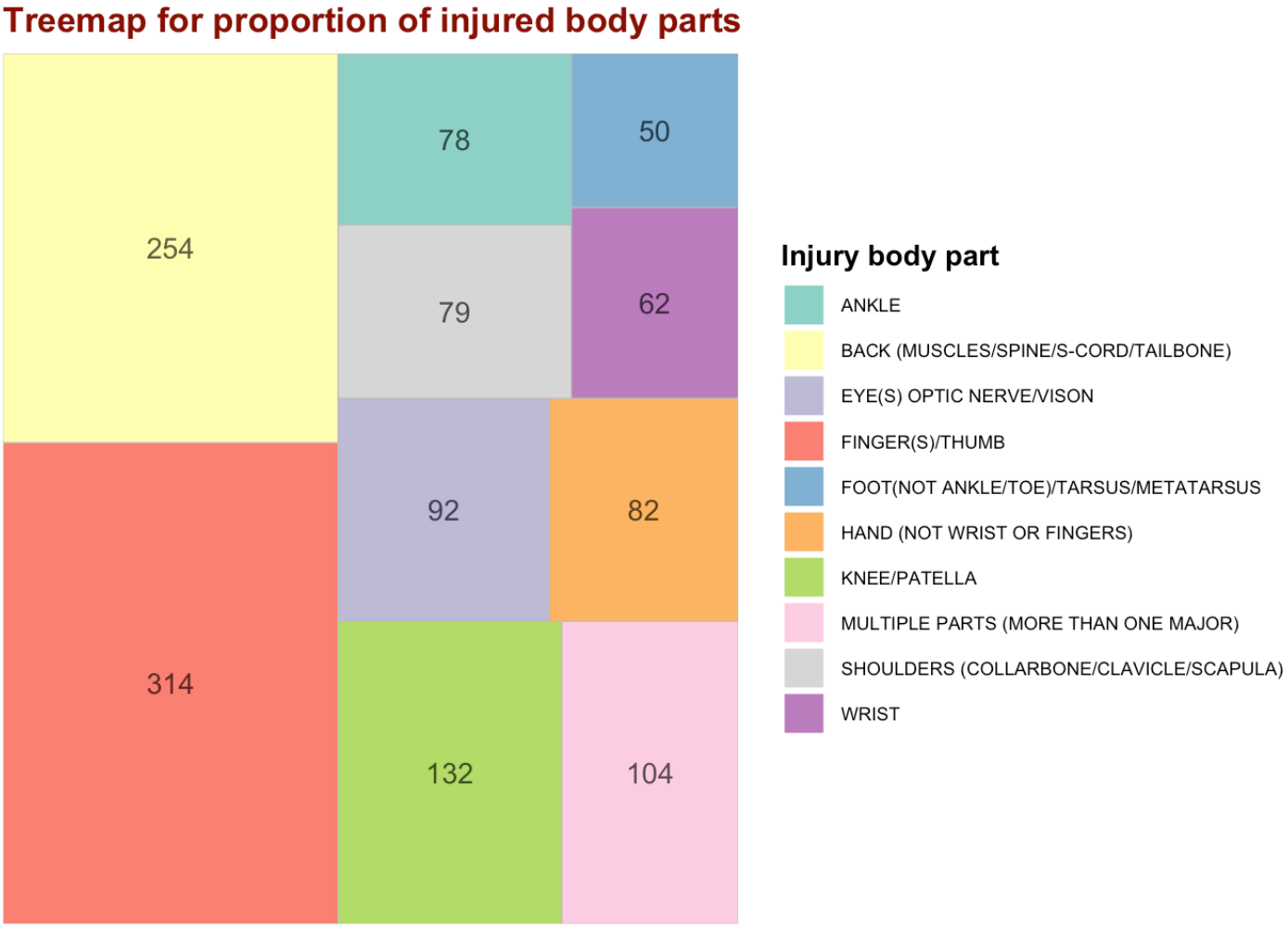
I. Analysis on the injured body parts

```
nlevels(accident_clean_v3$inj_body_part)
```

```
## [1] 45
```

Since there are too many levels of factors noted in inj_body_part and some levels have only occurred less than 10 times in the dataset, we will focus on the top 10 most frequent inj_body_part.

```
inj_body_part_counting <- count(accident_clean_v3, inj_body_part)
inj_body_part_counting <- inj_body_part_counting[order(-inj_body_part_counting$n,
inj_body_part_counting$inj_body_part),]

inj_body_part_counting_top10 <- head(inj_body_part_counting, 10)

ggplot(inj_body_part_counting_top10, aes(fill = inj_body_part, area = n, label = n
)) + geom_treemap() + geom_treemap_text(place = "center", size = 11 , alpha = 0.7)
+ labs(title = "Treemap for proportion of injured body parts", fill = "Injury body
part") + cits4009_theme + scale_fill_brewer(palette="Set3") + theme(legend.text=el
ement_text(size=7))
```

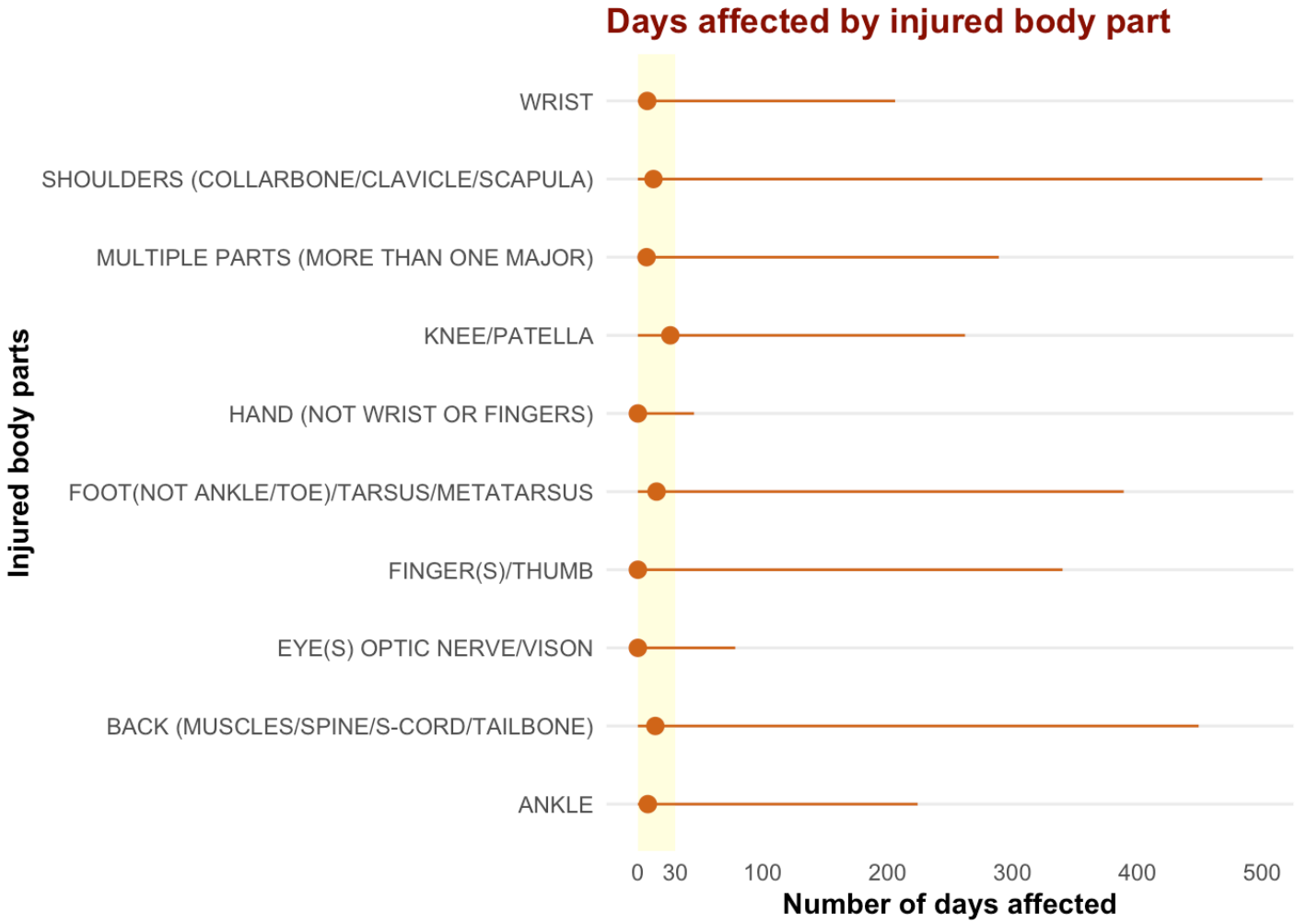


From the above, we can see there are 3 most common injury body parts for the workers, which are FINGER/ THUMB, BANK and KNEE/ PATELLA, contribute 314 cases, 254 cases and 132 cases of accident respectively.

II. Analysis on number of days affected for normal working on different injury body part

New variable of “days_affected” represents the number of days affected the workers to perform normal work duties after the accident, which is calculated from the sum of days_lost and days_restrict, is added to the dataset.

```
days_affected_df <- data.frame(accident_clean_v3$inj_body_part, accident_clean_v3$
days_affected)
colnames(days_affected_df) <- c("inj_body_part", "days_affected")
days_affected_df2 <- days_affected_df %>% filter(inj_body_part %in% inj_body_part_
counting_top10$inj_body_part)
days_affected_df2 <- na.omit(days_affected_df2)

ggplot(days_affected_df2) + geom_rect(data=NULL, aes(ymin= 0 , ymax= 30, xmin=-Inf
, xmax=Inf), fill="lightyellow") + stat_summary(aes(y = days_affected, x = inj_bod
y_part), fun.min = min, fun.max =max, fun = median, color = "chocolate") + scale_y
_continuous(breaks = c(0,30,100,200,300,400,500)) + labs(title = "Days affected by
injured body part", x = "Injured body parts", y = "Number of days affected") + coo
rd_flip() + cits4009_theme
```



From the above, we can see the median of number of days affected for all injured body part lie between 0 to 30 days, with knee or patella may need longer time to recover. Besides, we can also see there were accidents which workers require much longer time to recover. The highest number of days affected is

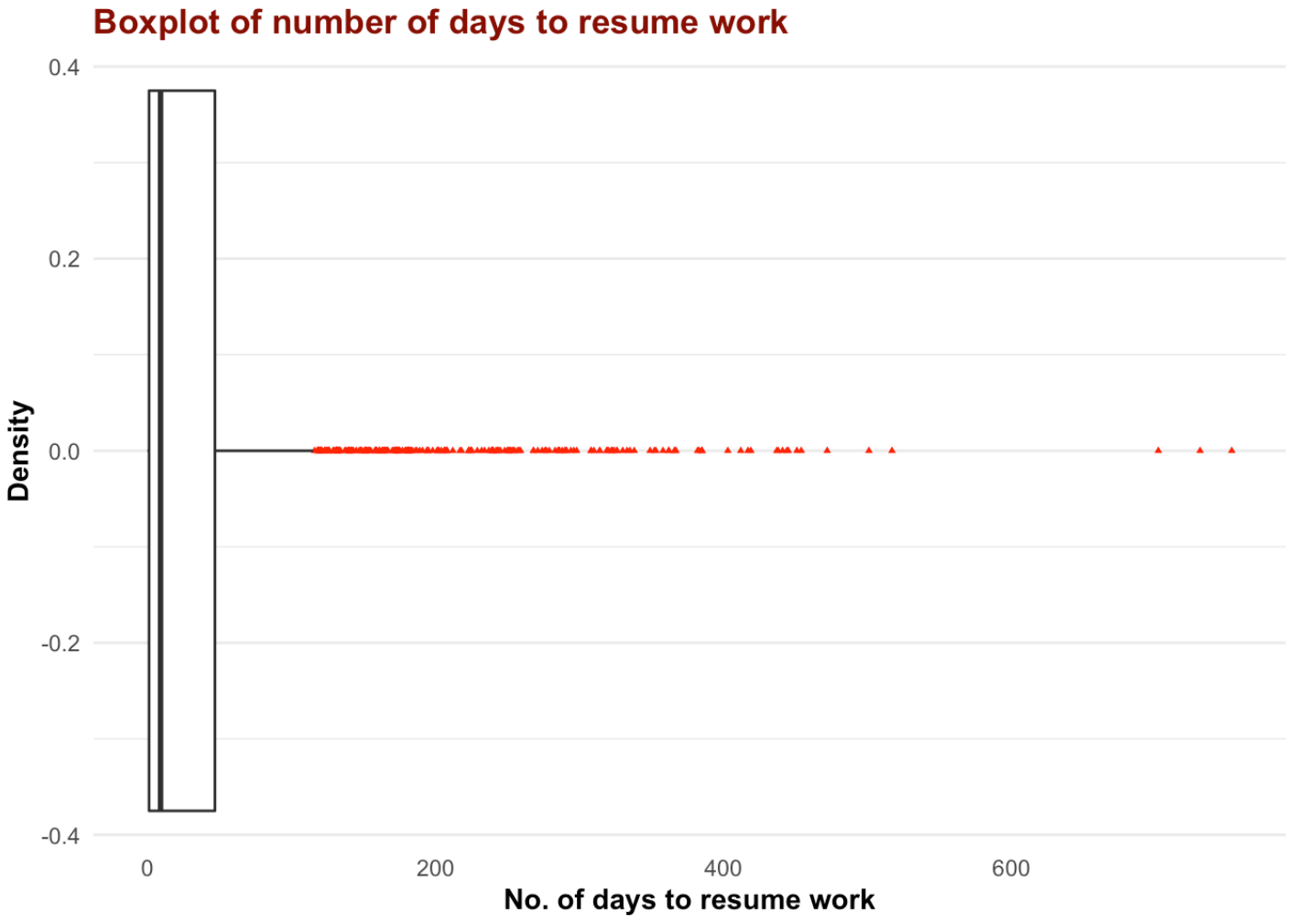
about 500 days on injured shoulders. As for hands and eyes, they have the shortest period in affecting normal work as compared to other body parts.

III. Analysis on number of days to resume work

New variable of “days_to_resume_work” represents the number of days taken by affected workers to resume work after the accident, which is calculated from the difference between return_to_work_dt and accident_dt, is added to the dataset.

```
days_to_resume_work_df <- data.frame(accident_clean_v3$days_to_resume_work)
colnames(days_to_resume_work_df) <- "days_to_resume_work"
days_to_resume_work_df <- subset(days_to_resume_work_df, !is.na(days_to_resume_wor
k))

ggplot(data = days_to_resume_work_df) + geom_boxplot(aes(x = days_to_resume_work),
outlier.colour = "red", outlier.shape = 17, outlier.size = 0.7, notch = FALSE) + l
abs(title = "Boxplot of number of days to resume work", x = "No. of days to resume
work", y = "Density") + cits4009_theme
```

```
## Don't know how to automatically pick scale for object of type difftime. Default
ing to continuous.
```

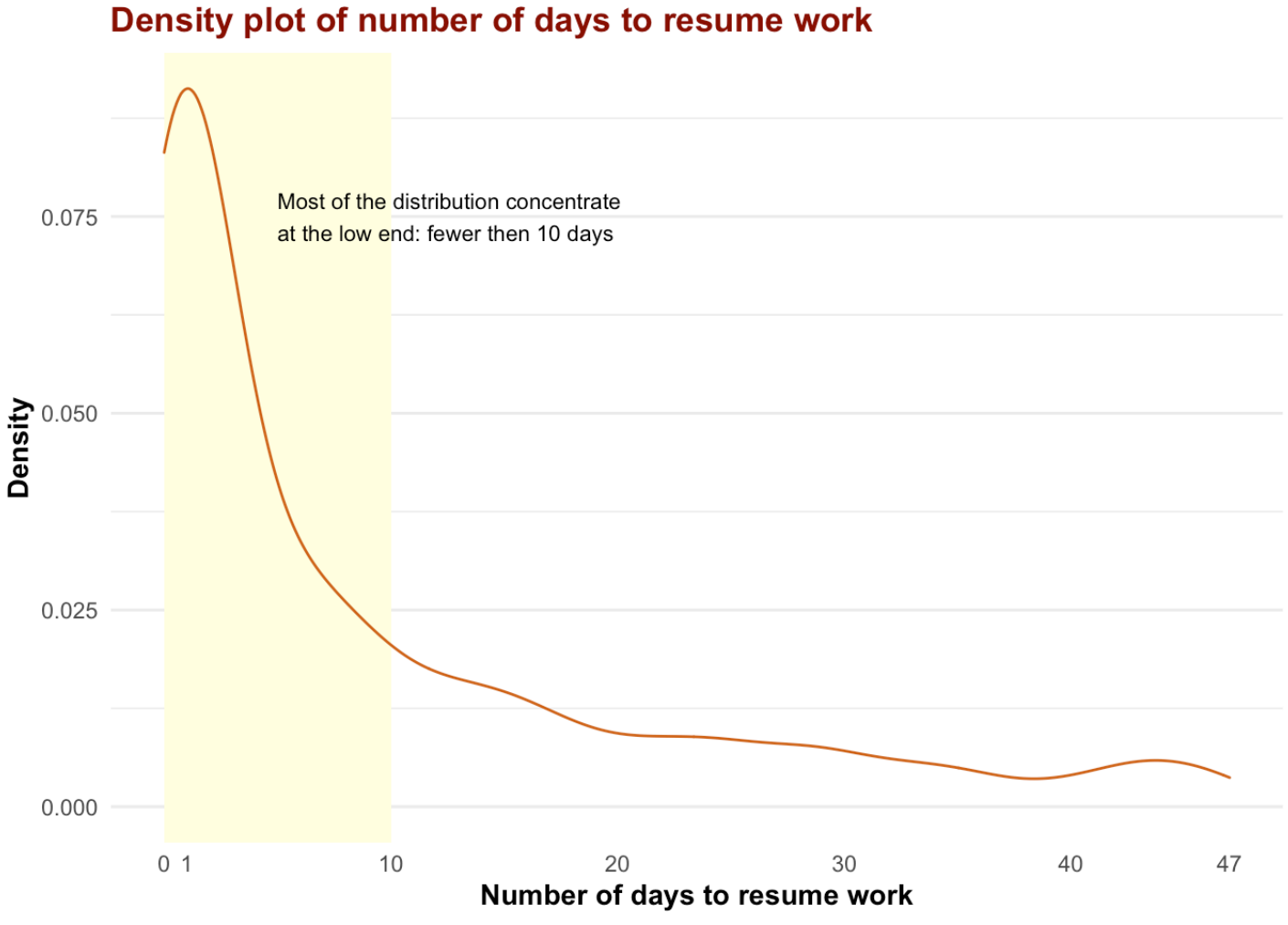


From the above, we can see the distribution of days_to_resume_work is extremely right-skewed owing to the outliers noted on the right, having significant gap between the accident date and resume to work date. In order to do analysis on the majority of affected workers, we will exclude the outliers (those greater than Q3, i.e. 47 days).

```
quantile(days_to_resume_work_df$days_to_resume_work)
```

```
## Time differences in days
##     0%    25%    50%    75%   100%
##   0.00   1.00   9.00  46.75 753.00
```

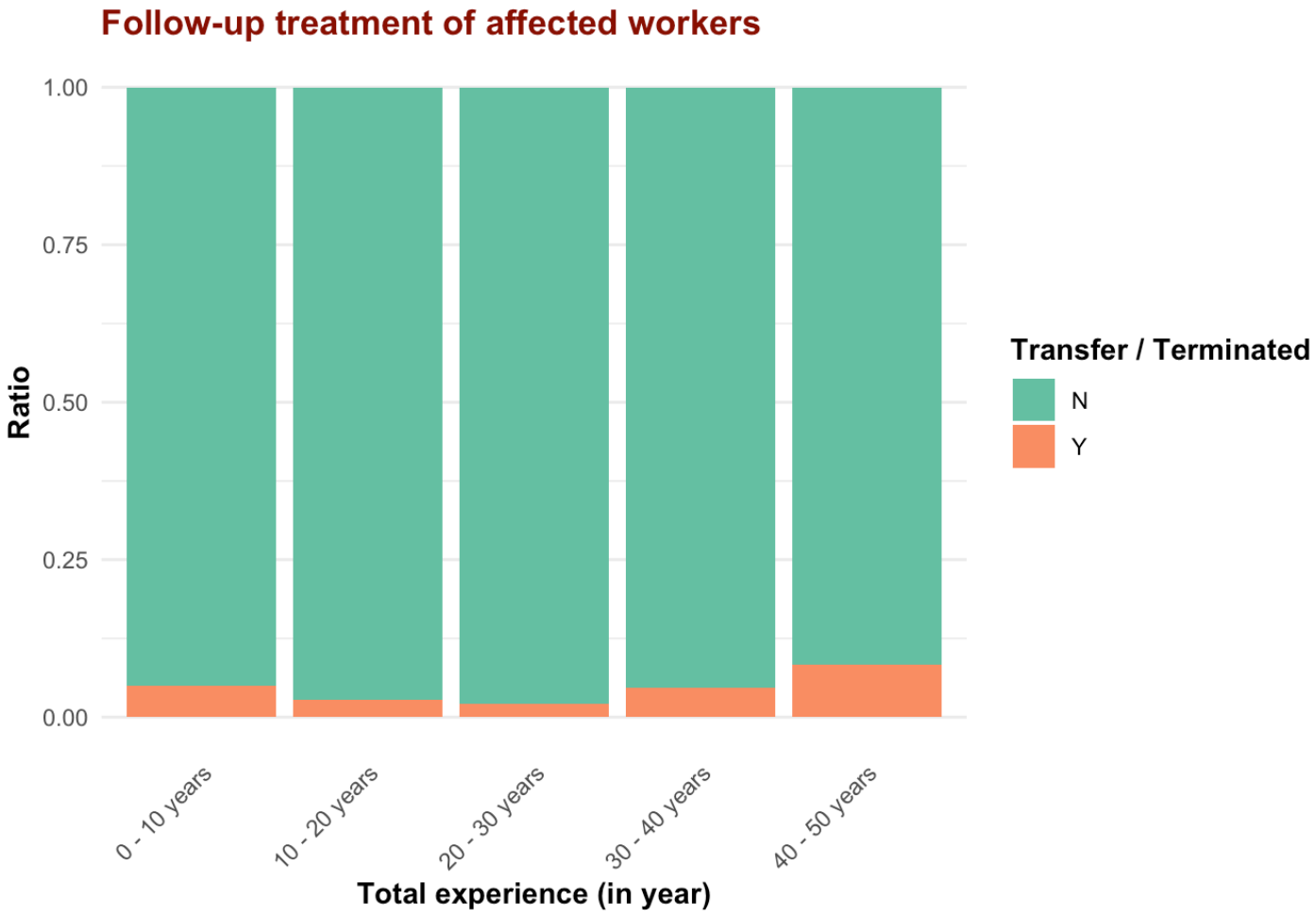
```
days_to_resume_work_majority_df <- subset(days_to_resume_work_df, days_to_resume_w
ork_df$days_to_resume_work <= 47)

ggplot(days_to_resume_work_majority_df, aes(x = days_to_resume_work)) + geom_rect(
data=NULL, aes(xmin= 0 , xmax= 10, ymin=-Inf, ymax=Inf), fill="lightyellow")+ geom
_density(color = "chocolate")  + scale_x_continuous(breaks = c(0,1,10,20,30,40,47)
) + labs(title = "Density plot of number of days to resume work", x = "Number of d
ays to resume work", y = "Density") + cits4009_theme + annotate("text", x = 5, y =
0.075, label = paste("Most of the distribution concentrate", "at the low end: fewe
r then 10 days", sep = "\n"), size = 3, hjust = 0)
```



By excluding the outliers where days to resume work greater than the 3rd quantile, we can see the majority of workers chose to resume their work within 10 days.

IV. Analysis on the follow-up treatment of affected workers

```
consequence_exp_df <- data.frame(accident_clean_v3$tot_exp_years, accident_clean_v
3$trans_term)
colnames(consequence_exp_df) <- c("tot_exp_years", "trans_term")
consequence_exp_df_clean <- subset(consequence_exp_df, !is.na(tot_exp_years) & tra
ns_term != "")

ggplot(consequence_exp_df_clean, aes(x = tot_exp_years, fill = trans_term)) + geom
_bar(position = "fill") + labs(title = "Follow-up treatment of affected workers",
x = "Total experience (in year)", y = "Ratio", fill = "Transfer / Terminated") + c
its4009_theme + theme(plot.title = element_text(size=13), axis.text.x = element_te
xt(angle= 45, hjust=1), aspect.ratio = 0.8) + scale_fill_brewer(palette="Set2")
```

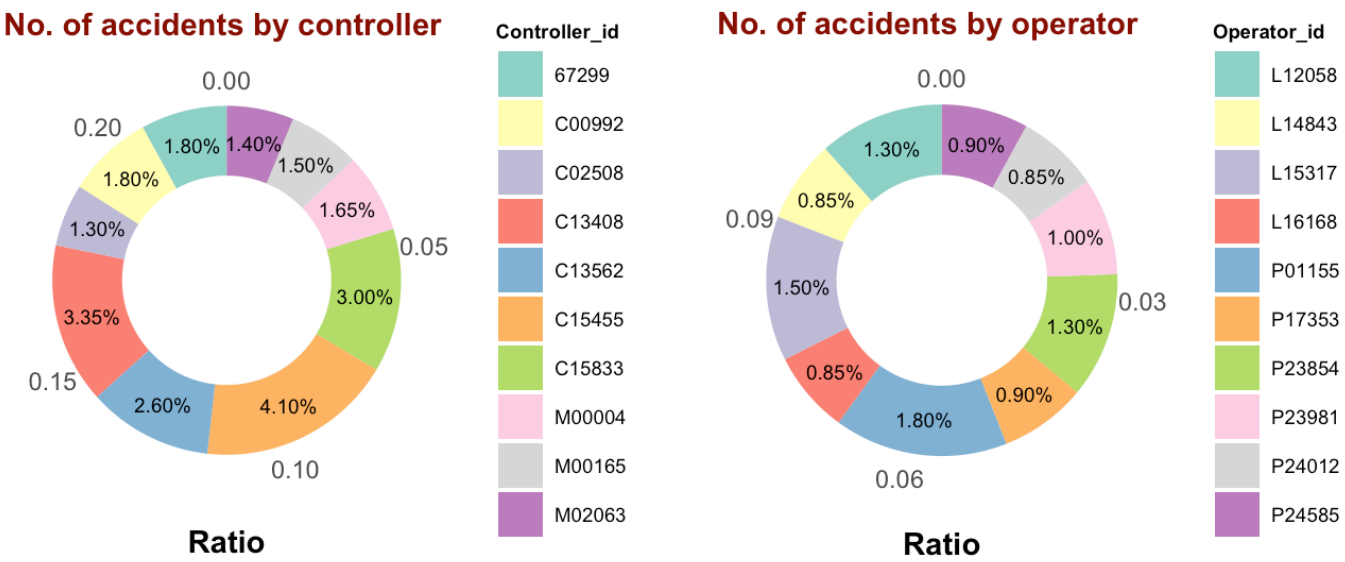


From the above analysis, we can see over 90% of affected workers for all experience groups did not transfer or terminated after the accidents. However, for more experienced workers (with 40 - 50 years experience), they are more likely to be transferred or terminated from the roles as compared to other experience groups.

4.5 Analysis on controller and operator

```
controller_extract_df <- as.data.frame(accident_clean_v2$controller_id)
colnames(controller_extract_df) <- "controller_id"

controller_extract_df <- controller_extract_df %>%
  group_by(controller_id) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

operator_extract_df <- as.data.frame(accident_clean_v2$operator_id)
colnames(operator_extract_df) <- "operator_id"

operator_extract_df <- operator_extract_df %>%
  group_by(operator_id) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

# Sorting the dataframe by the number of occurrence of controllers and operators
controller_extract_df2 <- controller_extract_df[order(controller_extract_df$n, dec
reasing = TRUE),]
operator_extract_df2 <- operator_extract_df[order(operator_extract_df$n, decreasin
g = TRUE),]

# Extracting the top 10 controllers and operators
controller_extract_top10_df <- head(controller_extract_df2, 10)
operator_extract_top10_df <- head(operator_extract_df2, 10)

# Plotting the donut chart
controller_extract_top10_df <- controller_extract_top10_df %>% mutate(x = hsize)
operator_extract_top10_df <- operator_extract_top10_df %>% mutate(x = hsize)

controller_donut <- ggplot(controller_extract_top10_df, aes(x= hsize, y = perc, fi
ll=controller_id)) + geom_col() + labs(title = "No. of accidents by controller", f
ill = "Controller_id", y = "Ratio", x = "") + coord_polar("y") + xlim(c(0.2, hsize
+ 0.5)) + scale_fill_brewer(palette="Set3") + cits4009_theme + theme(plot.title =
element_text(size = 11), legend.position = "right", legend.text = element_text(siz
e = 7), legend.title = element_text(size = 7), axis.text.y = element_blank(), pane
l.grid.major = element_blank(), panel.grid.minor = element_blank()) + geom_text(ae
s(label = labels), position = position_stack(vjust = 0.5), size = 2.5)

operator_donut <- ggplot(operator_extract_top10_df, aes(x= hsize, y = perc, fill=o
perator_id)) + geom_col() + labs(title = "No. of accidents by operator", fill = "O
perator_id", y = "Ratio", x = "") + coord_polar("y") + xlim(c(0.2, hsize + 0.5)) +
scale_fill_brewer(palette="Set3") + cits4009_theme + theme(plot.title = element_te
xt(size = 11), legend.position = "right", legend.text = element_text(size = 7), le
```

```
gend.title = element_text(size = 7), axis.text.y = element_blank(), panel.grid.maj
or = element_blank(), panel.grid.minor = element_blank()) + geom_text(aes(label =
labels), position = position_stack(vjust = 0.5), size = 2.5)

grid.arrange(controller_donut, operator_donut, ncol = 2)
```

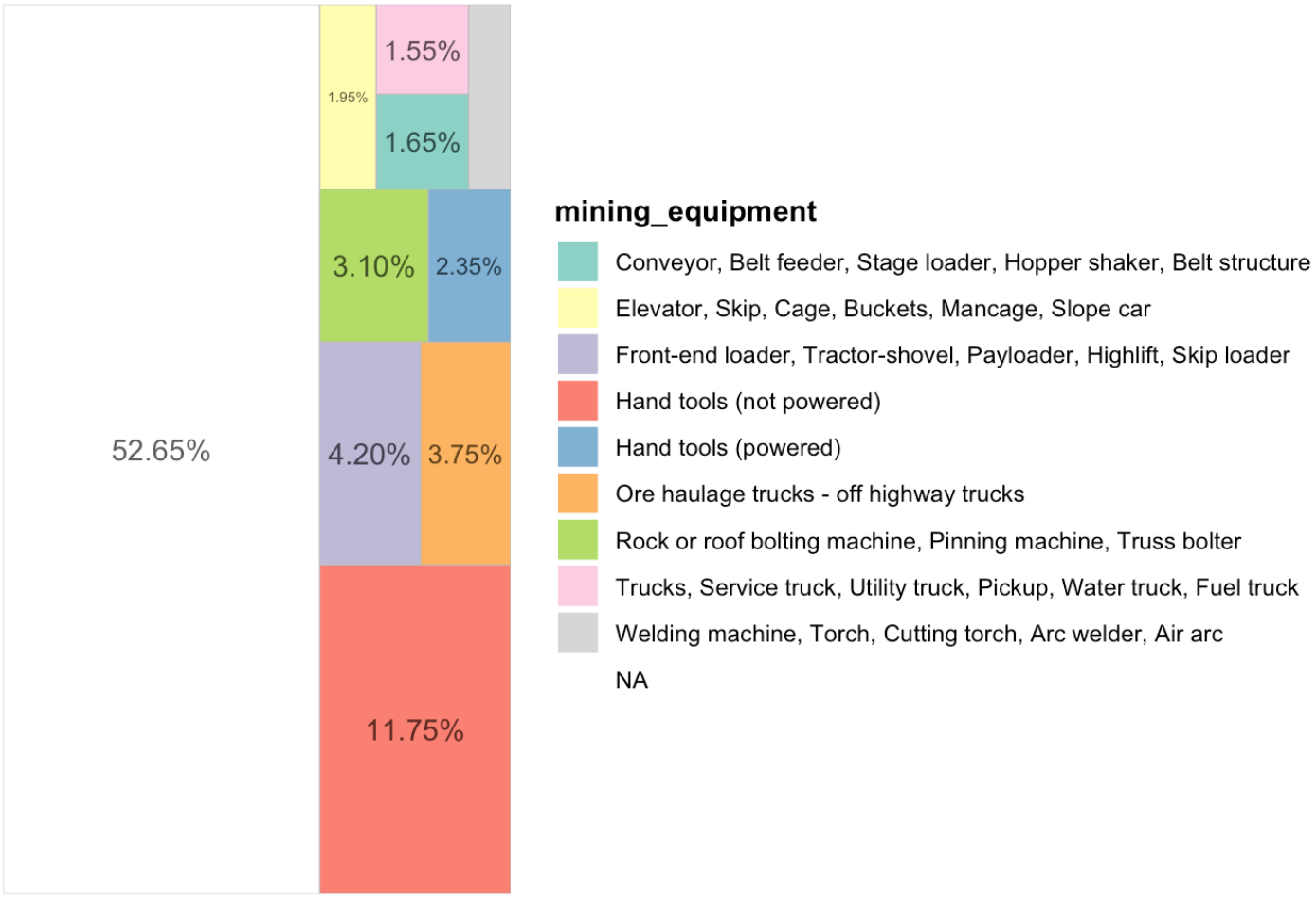


For controllers and operators, it seems that the distribution of them are quite diverse, there is no dominating controller and operators can be noted in terms of number of accidents. For controller, there may be 4 controllers had more accidents than the others, which are C15455 (Alliance Resource Partners LP) of 4.1%, C13408 (Robert E Murray) of 3.4%, C15833 (Peabody Energy) of 3.0% and C13562 (James River Coal Company) of 2.6%. Nevertheless, they only contributed to less than 5% of the whole population, which can be treat as insignificant.

4.6 Analysis on mining equipment

```
mining_equip_extract_df <- as.data.frame(accident_clean_v2$mining_equip)
colnames(mining_equip_extract_df) <- "mining_equip"

mining_equip_extract_df <- mining_equip_extract_df %>%
  group_by(mining_equip) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

# Sorting the dataframe by the number of occurrence of mining equipment
mining_equip_extract_df2 <- mining_equip_extract_df[order(mining_equip_extract_df$
n, decreasing = TRUE),]

# Extracting the top 10 mining equipment
mining_equip_extract_df2 <- head(mining_equip_extract_df2, 10)

# Plot the heatmap of mining equipment
ggplot(mining_equip_extract_df2, aes(fill = mining_equip, area = n, label = labels
)) + geom_treemap() + geom_treemap_text(place = "center", size = 11 , alpha = 0.7)
+ labs(title = "Treemap for proportion of mining equipment", fill = "mining_equipm
ent") + cits4009_theme + scale_fill_brewer(palette="Set3")
```

Treemap for proportion of mining equipment



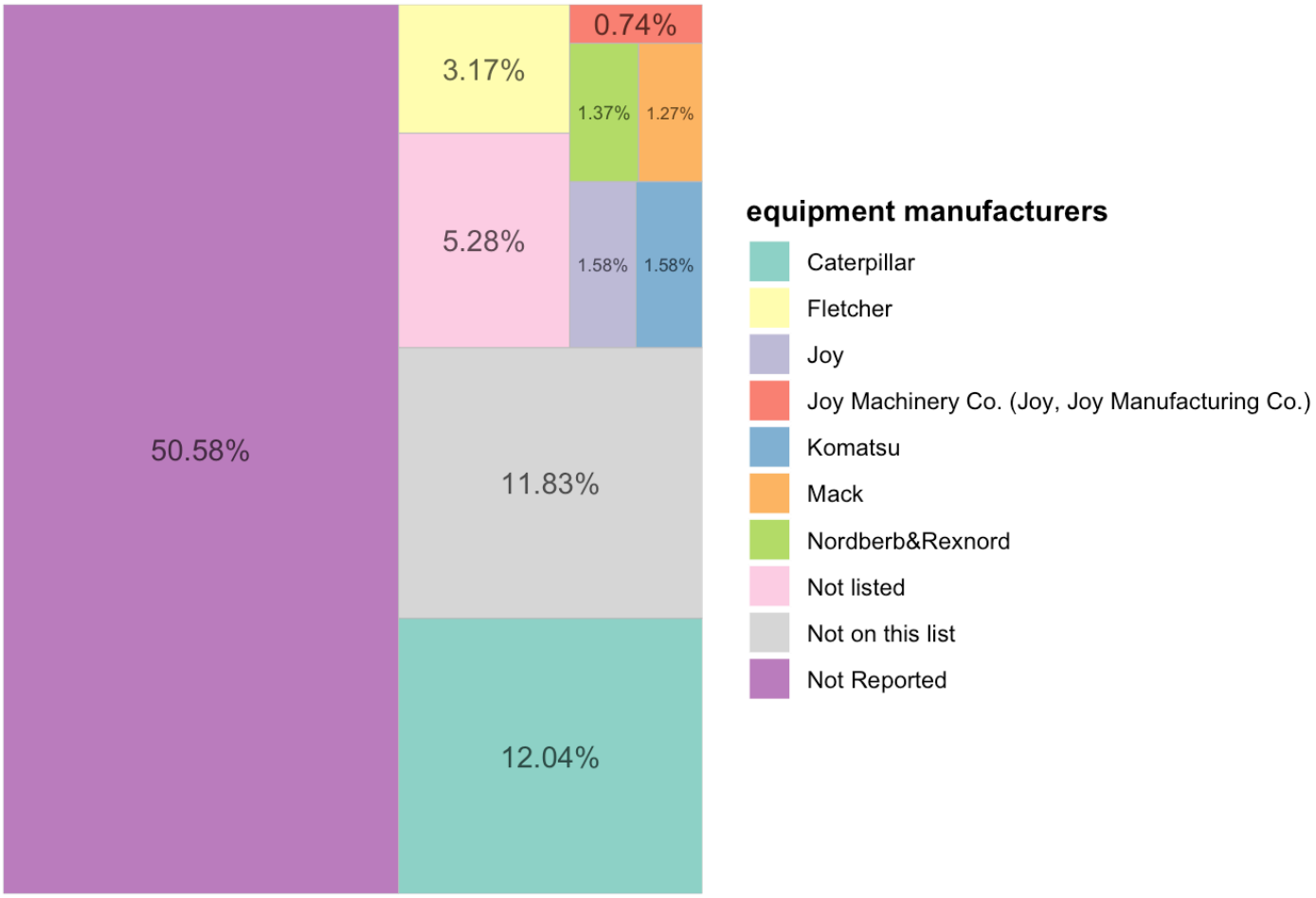
From the above, we can see over 50% are NA. In this case, NA may mean there was no mining equipment involved in the accidents. As for the accidents involved mining equipment, not powered hand tools contributed the highest portion of 11.8%.

Further analysis on manufactures

Looking further to see any specific equipment manufacturers produce equipment with higher risk of having accidents.

```
# Remove the accidents did not involve equipment
accident_clean_v4 <- subset(accident_clean_v2, !is.na(mining_equip))

equip_mfr_extract_df <- as.data.frame(accident_clean_v4$equip_mfr_name)
colnames(equip_mfr_extract_df) <- "equip_mfr"

equip_mfr_extract_df <- equip_mfr_extract_df %>%
  group_by(equip_mfr) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

# Sorting the dataframe by the number of occurrence of equipment manufacturers
equip_mfr_extract_df2 <- equip_mfr_extract_df[order(equip_mfr_extract_df$n, decrea
sing = TRUE),]

# Extracting the top 10 equipment manufacturers
equip_mfr_extract_df2 <- head(equip_mfr_extract_df2, 10)

# Plot the heatmap of equipment manufacturers
ggplot(equip_mfr_extract_df2, aes(fill = equip_mfr, area = n, label = labels)) + g
eom_treemap() + geom_treemap_text(place = "center", size = 11 , alpha = 0.7) + lab
s(title = "Treemap for proportion of equipment manufacturers", fill = "equipment m
anufacturers") + cits4009_theme + scale_fill_brewer(palette="Set3")
```

Treemap for proportion of equipment manufacturers



Apart from the accidents did not involve mining equipment, we can see from the above chart, there were over 65% of the equipment manufacturers were not reported or not listed. In order to conduct detailed analysis on safety standard of mining equipment by each manufacturers, more information shall be obtained.

Apart from that, we can also see there is one manufacturer, Caterpillar, contributed about 12%. Caterpiller is one of the world's leading manufacturer of construction and mining equipment. Thus, highest portion noted may due to it is the sole manufacturer for some of the mining equipment. However, safety standard on their products may worth to further analysis.