Tidying Hyperwar Data

Ryan Heslin

2021-05-06

Today I'll tidy some data from Hyperwar, an archive of World War II primary source documents. The data are several tables of aircraft the Army Air Forces receieved each month, broken down by role and type. Challenges include nested headers, blank rows, and several nasty OCR errors.

```
url <- "https://www.ibiblio.org/hyperwar/AAF/StatDigest/aafsd-3.html"

sel <- "blockquote:nth-child(113) table , blockquote:nth-child(109) table, blockquote:nth-child(105) tal
```

Scraping is not too challenging.

```
raw <- read_html(url) %>%
    html_elements(css = sel) %>%
    html_table()
```

After some experimentation, I devise a function to clean the data.

```
clean_hyperwar <- function(tab) {
  tab <- tab[, colSums(is.na(tab)) != nrow(tab)]
  names(tab) <- c("Type", paste(tab[1, -1], names(tab)[-1]))
  tab %>% mutate(Category = str_extract(Type, "(^[^0-9\\-]+)(?=s--)"),
                 .before = Category) %>%
    fill(Category, .direction = "down") %>%
    mutate(
      Type = case_when(
        str_detect(Type, "2nd") ~ paste("2nd. Line", Category),

          str_detect(Type, "^Other") ~ paste("Other", Category),
          TRUE ~ Type
      )
      ) %>%
        mutate(
          across(
            -c(Type, Category),
            ~ str_replace_all(.x, c("-" = "0", "(\\d+,\\d{3})\\d"= "\\1", "," = "")) %>%
              as.numeric()
          )
        ) %>%
        filter(
          if_all(-c(Type, Category), ~ !is.na(.x)) &
            !str_detect(Type, "Total|1st|^Combat Airplanes$")
        ) %>%
```

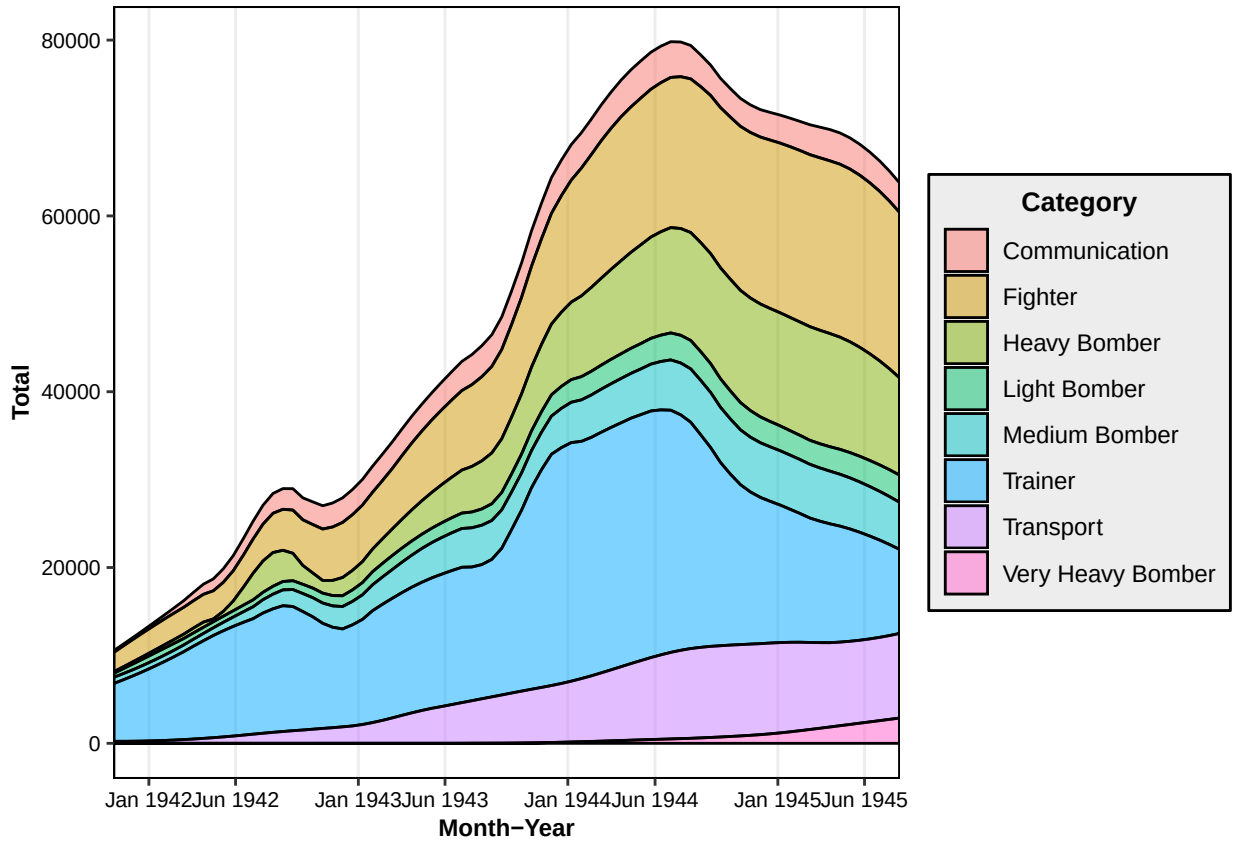
```
    pivot_longer(
      cols = -c(Type, Category),
      names_to = "Month-Year",
      values_to = "Number"
    ) %>%
    mutate(across(c(Type, Category), as.factor))
}
```

Data in hand, I fill in missing combinations.

```
cleaned <- map(raw, clean_hyperwar) %>%
    bind_rows() %>%
    mutate(`Month-Year` = my(paste(`Month-Year`))) %>%
    complete(nesting(Category, Type), `Month-Year`, fill = list(Number = 0))
```

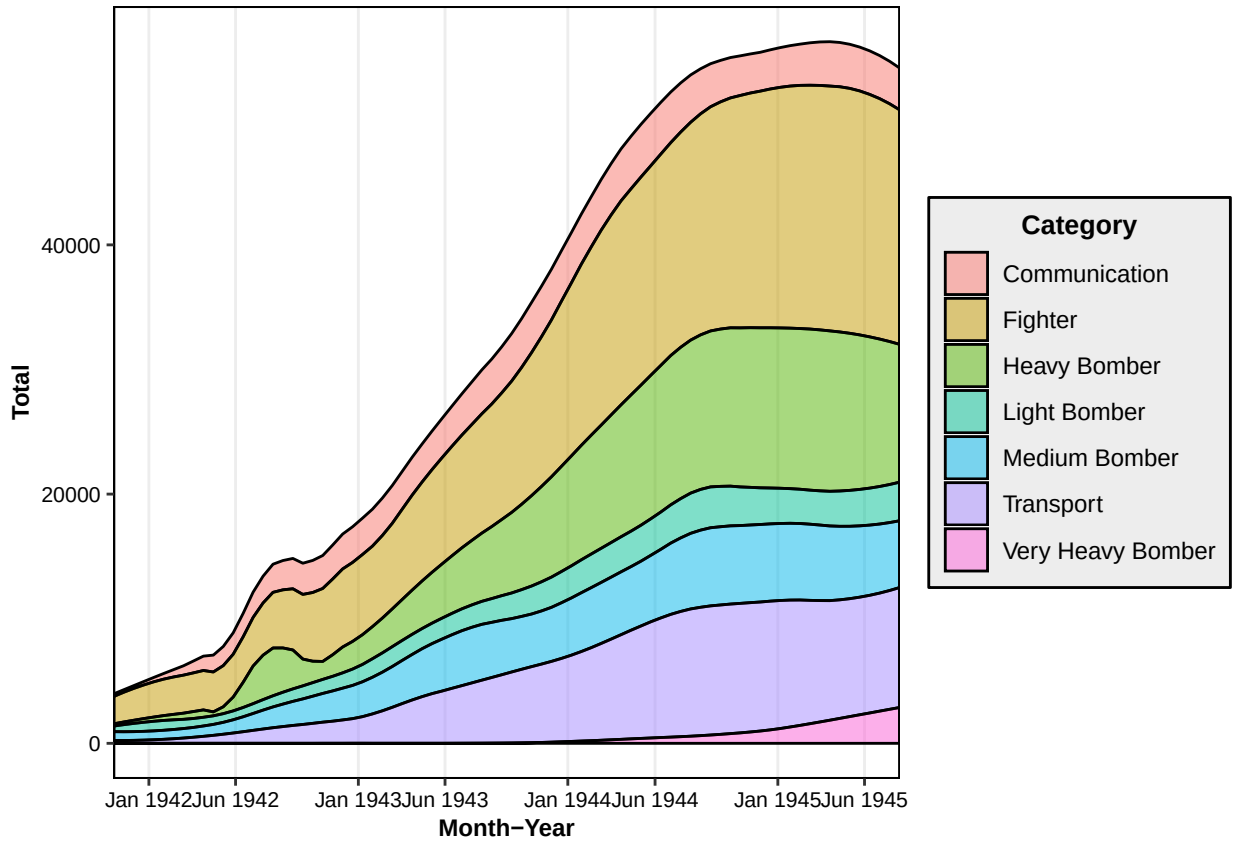
It seems trainers dominated acquisitions throughout the war, which makes sense.

```
my_theme()

breaks <- paste(rep(c("Jan", "Jun"), times = 4), rep(1941:1945, 2))

cleaned %>%
    group_by(`Month-Year`, Category) %>%
    summarize(Total = sum(Number)) %>%
    ggplot(aes(x = `Month-Year`, y = Total, fill = Category)) + stat_smooth(geom = "area", method = "lo
    position = position_stack(), outline.type = "full", color = "black", alpha = 0.5, bandwith = 0.2,
    span = 0.2) + scale_x_date(breaks = my(breaks), labels = breaks, expand = c(0, 0))
```



After filtering out trainers, it's clear fighters received surprisingly high emphasis even late in the war, when the Allies enjoyed air supremacy.

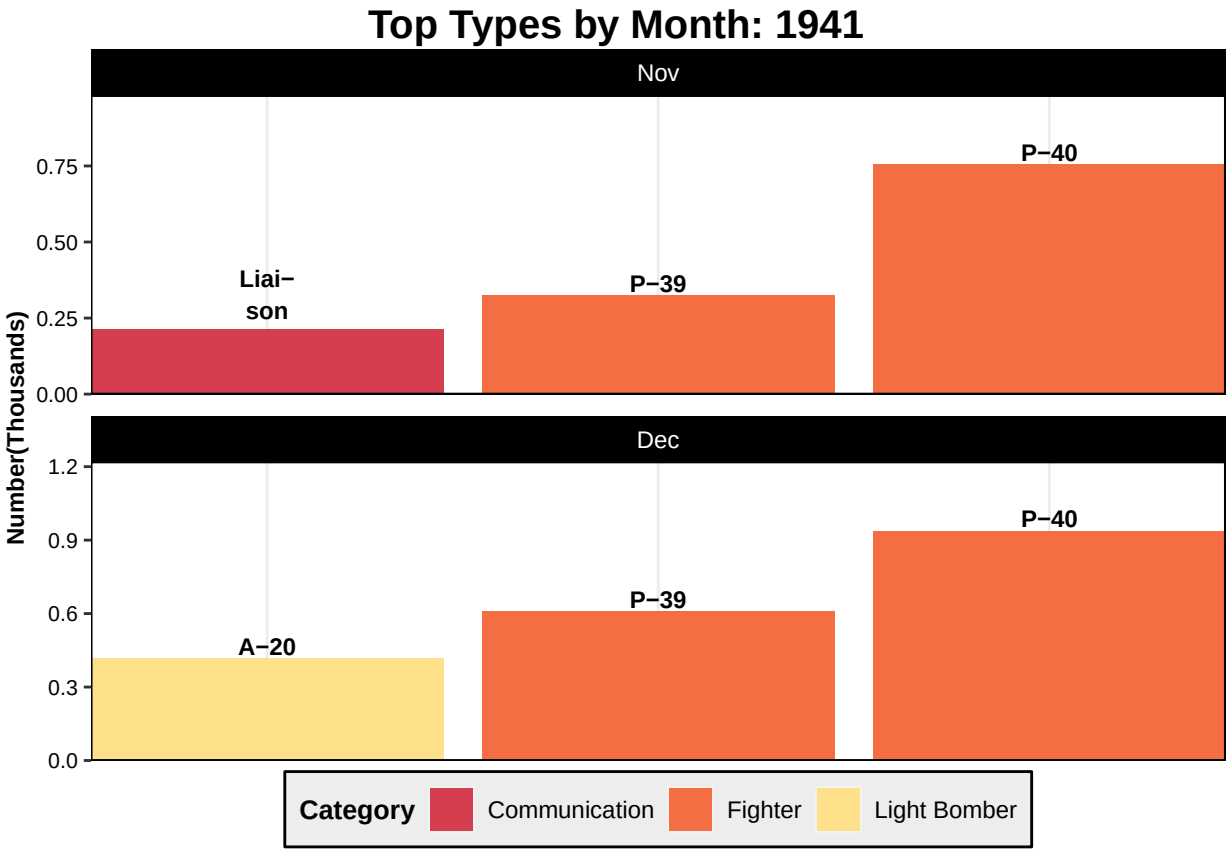
```
cleaned %>%
    filter(Category != "Trainer") %>%
    group_by(`Month-Year`, Category) %>%
    summarize(Total = sum(Number)) %>%
    ggplot(aes(x = `Month-Year`, y = Total, fill = Category)) + stat_smooth(geom = "area", method = "lo
    position = position_stack(), outline.type = "full", color = "black", alpha = 0.5, bandwith = 0.2,
    span = 0.2) + scale_x_date(breaks = my(breaks), labels = breaks, expand = c(0, 0))
```

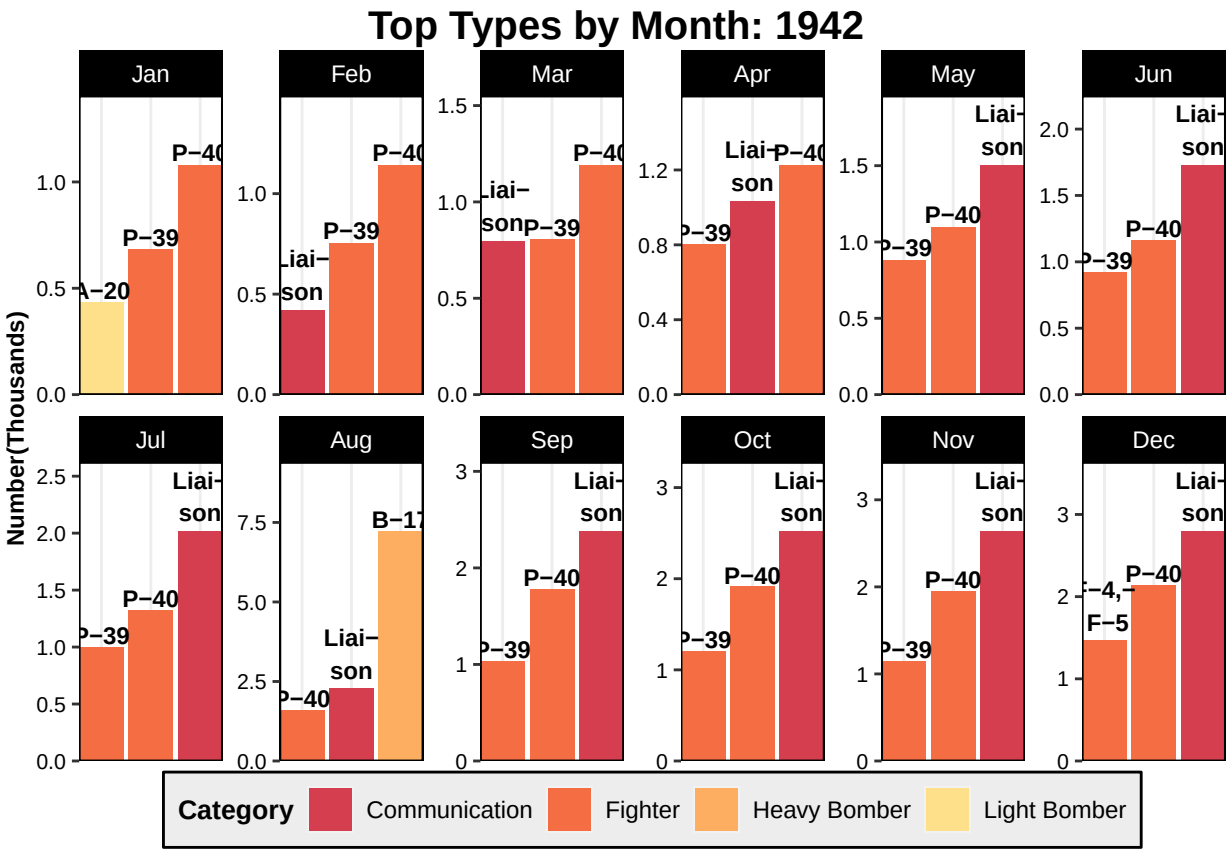


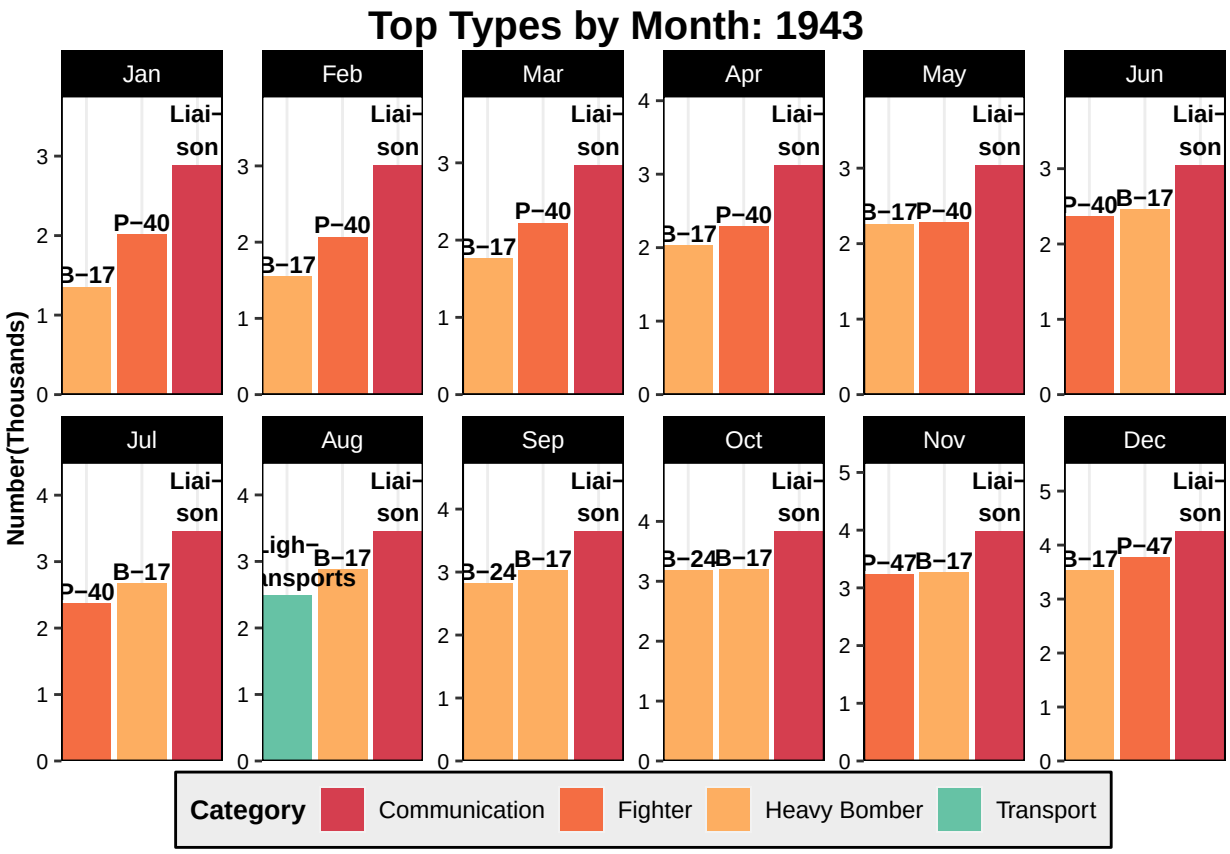
I plot the top three types produced each month. These plots reveal a shift from fighters to heavy bombers and transports midway through the war. There were initially some odd spikes I fixed by correcting OCR errors.

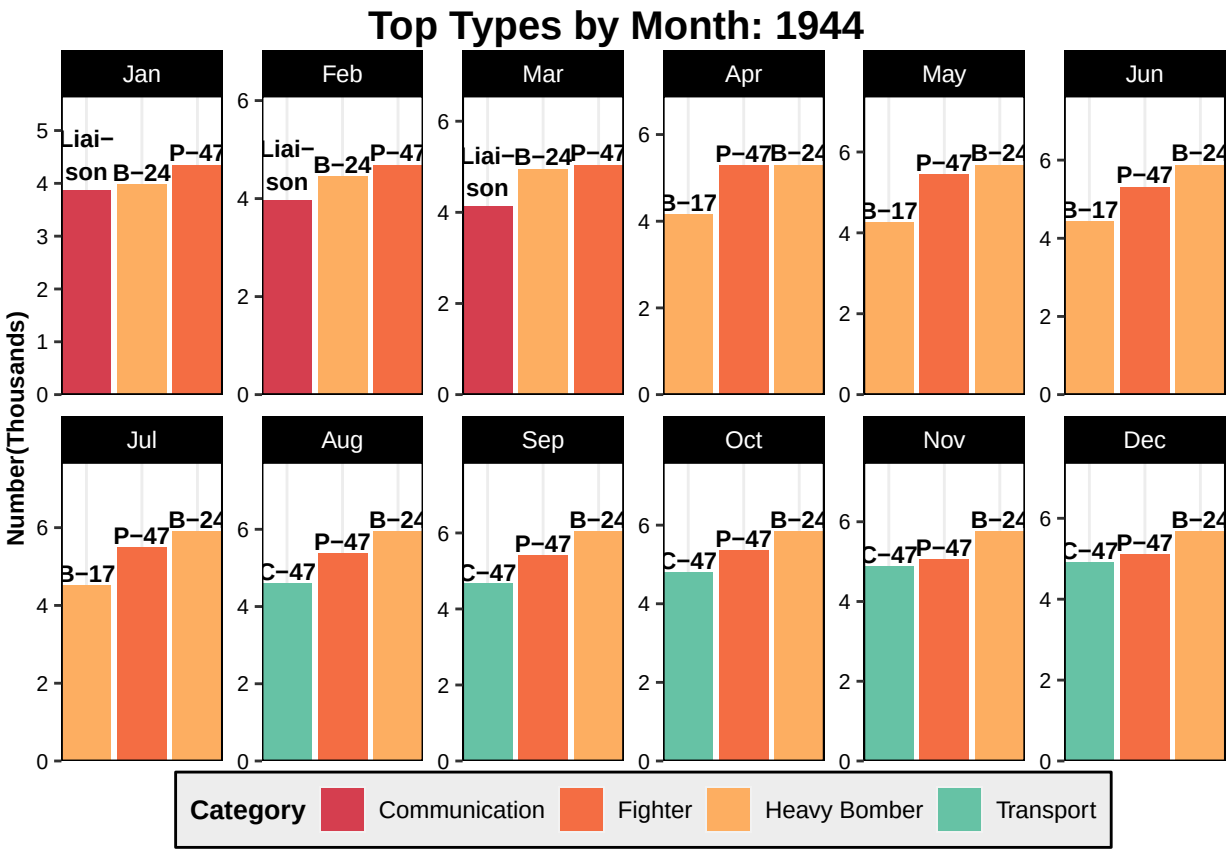
```
top_types <- cleaned %>%
    group_by(`Month-Year`) %>%
    filter(Category != "Trainer" & !str_detect(Type, "2nd|Other")) %>%
    slice_max(n = 3, order_by = Number) %>%
    mutate(Type_ordered = tidytext::reorder_within(Type, Number, within = `Month-Year`)) %>%
    ungroup() %>%
    arrange(`Month-Year`) %>%
    mutate(Year = year(`Month-Year`), Month = month(`Month-Year`, label = TRUE), Type = fct_recode(Type
        `C-47` = "C-47, C-53"))

pal <- RColorBrewer::brewer.pal(n = nlevels(top_types$Category), name = "Spectral") %>%
    setNames(levels(top_types$Category))
plots <- top_types %>%
    group_by(Year) %>%
    group_map(~ggplot(data = .x, aes(x = Type_ordered, label = str_replace(Type, "(.{4})(.+)", "\\1-\n\
        y = Number/1000, fill = Category)) + geom_col(position = "dodge") + geom_text(size = 3, vjust =
        fontface = "bold") + scale_fill_manual(values = pal) + tidytext::scale_x_reordered(expand = c(0
        0)) + facet_wrap(. ~ Month, nrow = 2, scales = "free") + scale_y_continuous(expand = expansion(m
        0.3))) + labs(title = paste("Top Types by Month:", .y), x = NULL, y = "Number(Thousands)") +
        theme(legend.position = "bottom", legend.box.spacing = unit(0.05, "cm"), axis.text.x = element_
            axis.ticks.x = element_blank(), ))
```

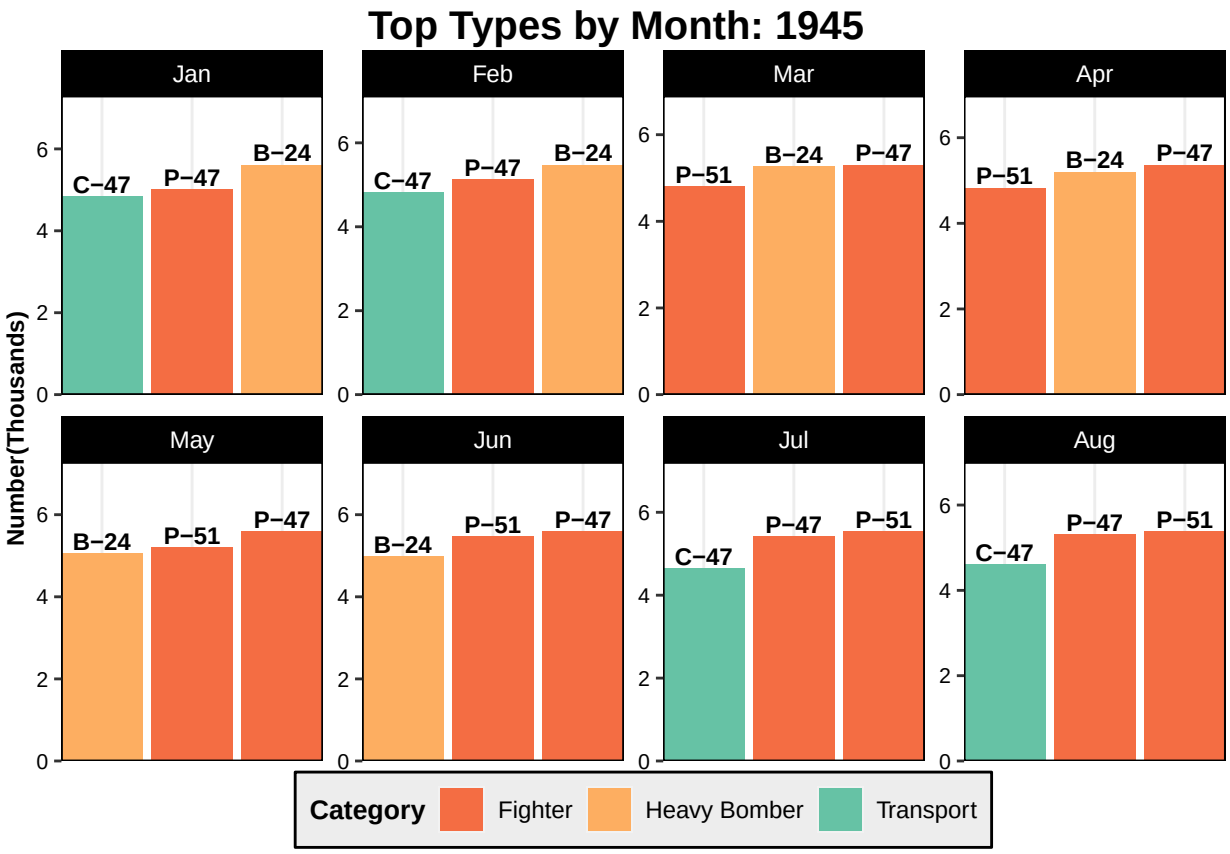
```
walk(plots, print)
```




Top Types by Month: 1942
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Number(Thousands)
P−39
P−40
Liai−
son
Category
Communication
Fighter
Heavy Bomber
Light Bomber


Top Types by Month: 1943
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Number(Thousands)
Liai-
son
P−40
B−17
B−24
P−47
Category
Communication
Fighter
Heavy Bomber
Transport


Top Types by Month: 1944
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Number(Thousands)
Liai−son
B−24
P−47
B−17
C−47
Category
Communication
Fighter
Heavy Bomber
Transport


Top Types by Month: 1945
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Number(Thousands)
0
2
4
6
C−47
P−47
B−24
P−51
Category
Fighter
Heavy Bomber
Transport