

Assignment 3

HİLAL SELEN BÜTÜN

2025-10-20

```
install.packages("ggplot2")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)

install.packages("ggmosaic")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)

install.packages("treemapify")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)

library(ggplot2)
library(ggmosaic)
library(treemapify)
# Installing necessary package(s)
```

Dataset

The **salary** dataset contains information about the average annual salaries of data professionals across different roles and experience levels. It includes numerical data on salaries (**mean_salary**) in Rupees as well as categorical variables describing the job role (**position**) and experience level (**experience**). The experience levels are ordered from Junior → Intermediate → Senior → Executive, reflecting increasing levels of expertise and responsibility. This dataset can be used to explore which positions and experience levels contribute most to overall salary distribution in the data science field.

Source: Kaggle – Data Science Fields Salary Categorization (accessed October 2025).

To load the dataset into your environment, just run the code below:

```
salary <- data.frame(
  position = c("Data Analyst", "Data Analyst", "Data Analyst", "Data Analyst",
               "Data Engineer", "Data Engineer", "Data Engineer", "Data Engineer",
               "Data Scientist", "Data Scientist", "Data Scientist"),
  experience = c("Junior", "Executive", "Intermediate", "Senior",
                 "Junior", "Executive", "Intermediate", "Senior",
                 "Junior", "Intermediate", "Senior"),
  mean_salary = c(4293623, 9548340, 5705070, 8905628,
                  4689309, 19534312, 6841836, 10903873,
                  4402653, 6527813, 12171827))
```

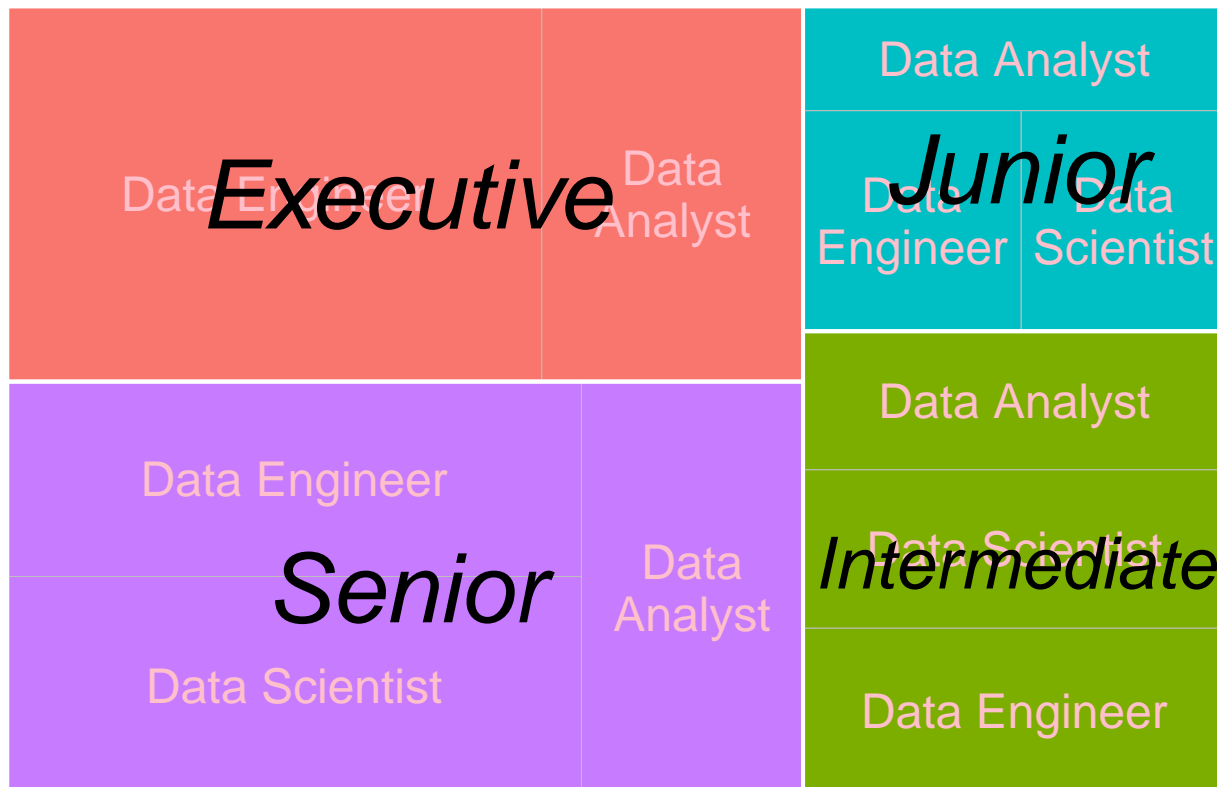
Drawing a plot for proportion

1. Draw a plot showing the proportion of total mean salaries by experience and position. The plot should clearly display the hierarchical contribution of each job title within its experience level.

```
salary$prop <- salary$mean_salary / sum(salary$mean_salary)

ggplot(salary, aes(area = prop,
                    fill = experience,
                    label = position,
                    subgroup = experience)) +
  geom_treemap() +
  geom_treemap_text(colour = "pink", place = "centre", reflow = TRUE) +
  geom_treemap_subgroup_border(colour = "white", size = 2) +
  geom_treemap_subgroup_text(place = "centre", grow = FALSE,
                              alpha = 1, colour = "black", fontface = "italic") +
  labs(title = "Proportion of Total Mean Salaries by Experience and Position") +
  theme(legend.position = "none")
```

Proportion of Total Mean Salaries by Experience and Position



2. Interpret the plot (30 pts).

This treemap chart shows the percentage of average salary by experience level in different job positions. The size of the boxes indicates the percentage of total salary for that group. Executive and Senior positions receive the highest share, followed by Intermediate and Junior positions. In the Executive position, the highest salary is earned by the Data Engineer, followed by the Data Analyst. In the Senior position, the highest salary is earned by the Data Engineer and Data Scientist, followed by the Data Analyst. In the Intermediate position, the percentages for all positions are similar. In the Junior position, the Data Engineer and Data Scientist are close, while the Data Analyst has the lowest. In summary, the largest share of the salary distribution is seen

at the Executive and Senior levels, while the smallest share is seen at the Junior level.