Hickey Lab!

Mariah Culpepper

```
library(RColorBrewer)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.0     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(dplyr)
library(ggplot2)
library(tidytext)
```

```
data = read.csv("/Users/riahcul/Downloads/locofspatialdata.csv")
data[data == ""] <- NA
```

```
x_freqs <- data |>
  group_by(x_col_name) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

y_freqs <- data |>
  group_by(y_col_name) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

celltype_freqs <- data |>
  group_by(celltype_col_name) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

uniqueregion_freqs <- data |>
  group_by(uniqueregion_col_name) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

modality_freqs <- data |>
  group_by(modality) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

disease_freqs <- data |>
  group_by(disease) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

tissue_freqs <- data |>
  group_by(tissue) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))
```

```
print(x_freqs)
```

```
## # A tibble: 25 × 2
##    x_col_name        Frequency
##    <chr>                 <int>
##  1 x                        15
##  2 X                        14
##  3 centroid-0                6
##  4 Location_Center_X         3
##  5 X_centroid                3
##  6 centroid_x                3
##  7 X:X                       2
##  8 X_cent                    2
##  9 Center_X                  1
## 10 NominalPostion_X          1
## # ℹ 15 more rows
```

```
print(y_freqs)
```

```
## # A tibble: 25 × 2
##    y_col_name        Frequency
##    <chr>                 <int>
##  1 y                        15
##  2 Y                        14
##  3 centroid-1                6
##  4 Location_Center_Y         3
##  5 Y_centroid                3
##  6 centroid_y                3
##  7 Y:Y                       2
##  8 Y_cent                    2
##  9 Center_Y                  1
## 10 NominalPosition_Y         1
## # ℹ 15 more rows
```

```
print(celltype_freqs)
```

```
## # A tibble: 33 × 2
##    celltype_col_name Frequency
##    <chr>                 <int>
##  1 <NA>                     10
##  2 cell_type                 6
##  3 phenotype                 5
##  4 cellType                  4
##  5 Cell Type                 3
##  6 Cluster                   3
##  7 ClusterName               3
##  8 celltype                  3
##  9 Annotation                2
## 10 CellType                  2
## # ℹ 23 more rows
```

```
print(uniqueregion_freqs)
```

```
## # A tibble: 31 × 2
##    uniqueregion_col_name Frequency
##    <chr>                     <int>
##  1 unique_region                11
##  2 fov                           8
##  3 region                        7
##  4 separated by File             4
##  5 Region                        3
##  6 roi                           3
##  7 <NA>                          3
##  8 frame                         2
##  9 imageID                       2
## 10 FileName                      1
## # ℹ 21 more rows
```

```
print(modality_freqs)
```

```
## # A tibble: 18 × 2
##    modality                       Frequency
##    <chr>                              <int>
##  1 CODEX                                 21
##  2 IMC                                   14
##  3 MIBI                                   8
##  4 CyCIF                                  4
##  5 CODEX,MIBI                             3
##  6 CosMx SMI                              3
##  7 3D Cell DIVE                           1
##  8 CODEX, multiplexed ISH                 1
##  9 CODEX,CyCIF,mIHC,MxIF,IMC,MIBI         1
## 10 IBEX                                   1
## 11 MACSima                                1
## 12 MALDI                                  1
## 13 MIF                                    1
## 14 MIF (using CODEX)                      1
## 15 Orion                                  1
## 16 Visium                                 1
## 17 mIHC                                   1
## 18 <NA>                                   1
```

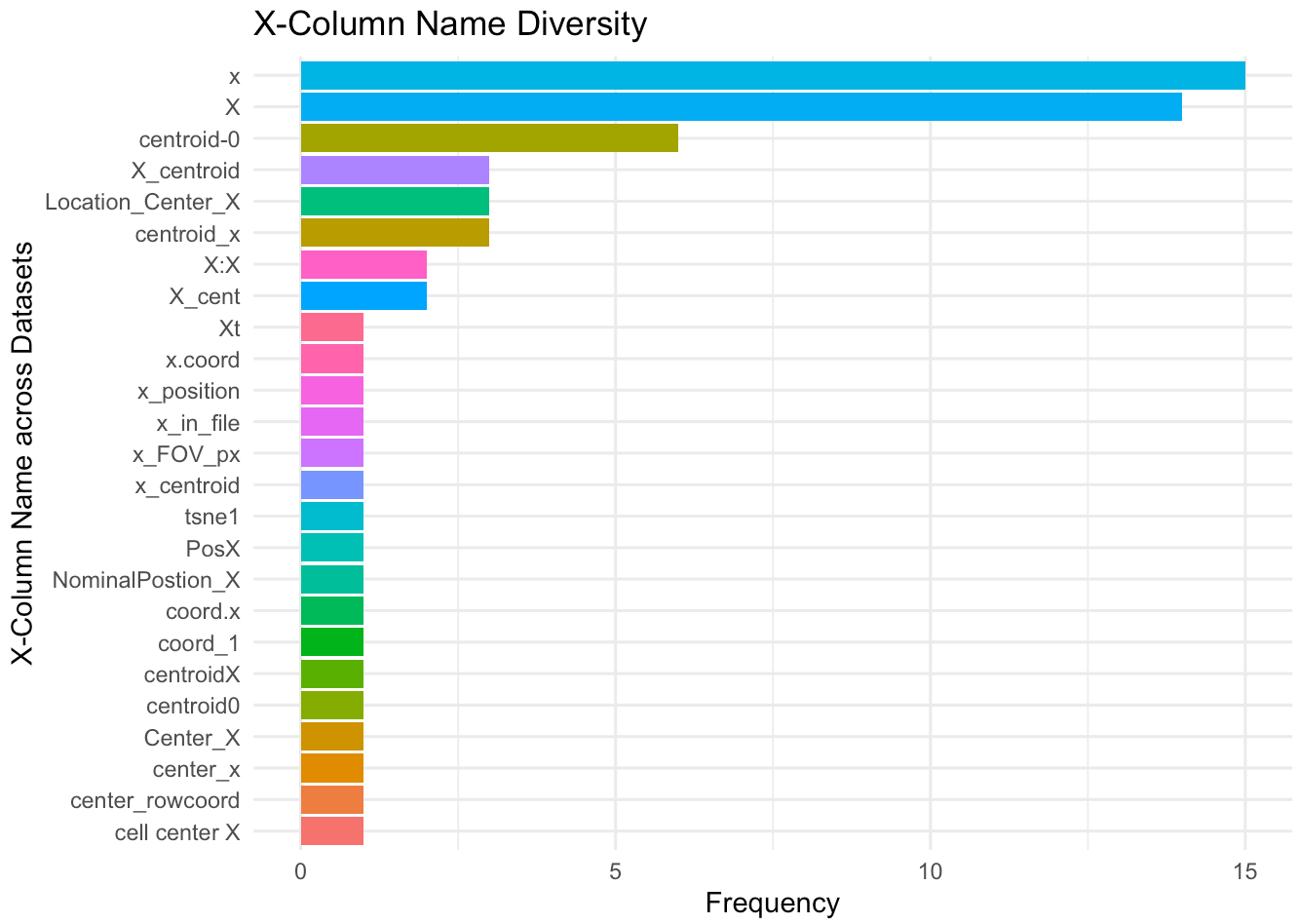
```
print(disease_freqs)
```

```
## # A tibble: 9 × 2
##   disease             Frequency
##   <chr>                   <int>
## 1 Cancer                     39
## 2 Normal                     11
## 3 Normal,Cancer               4
## 4 Infection                   3
## 5 Inflammation                3
## 6 <NA>                        2
## 7 Normal,Inflammation         1
## 8 Osteoarthritis              1
## 9 Type 1 Diabetes             1
```

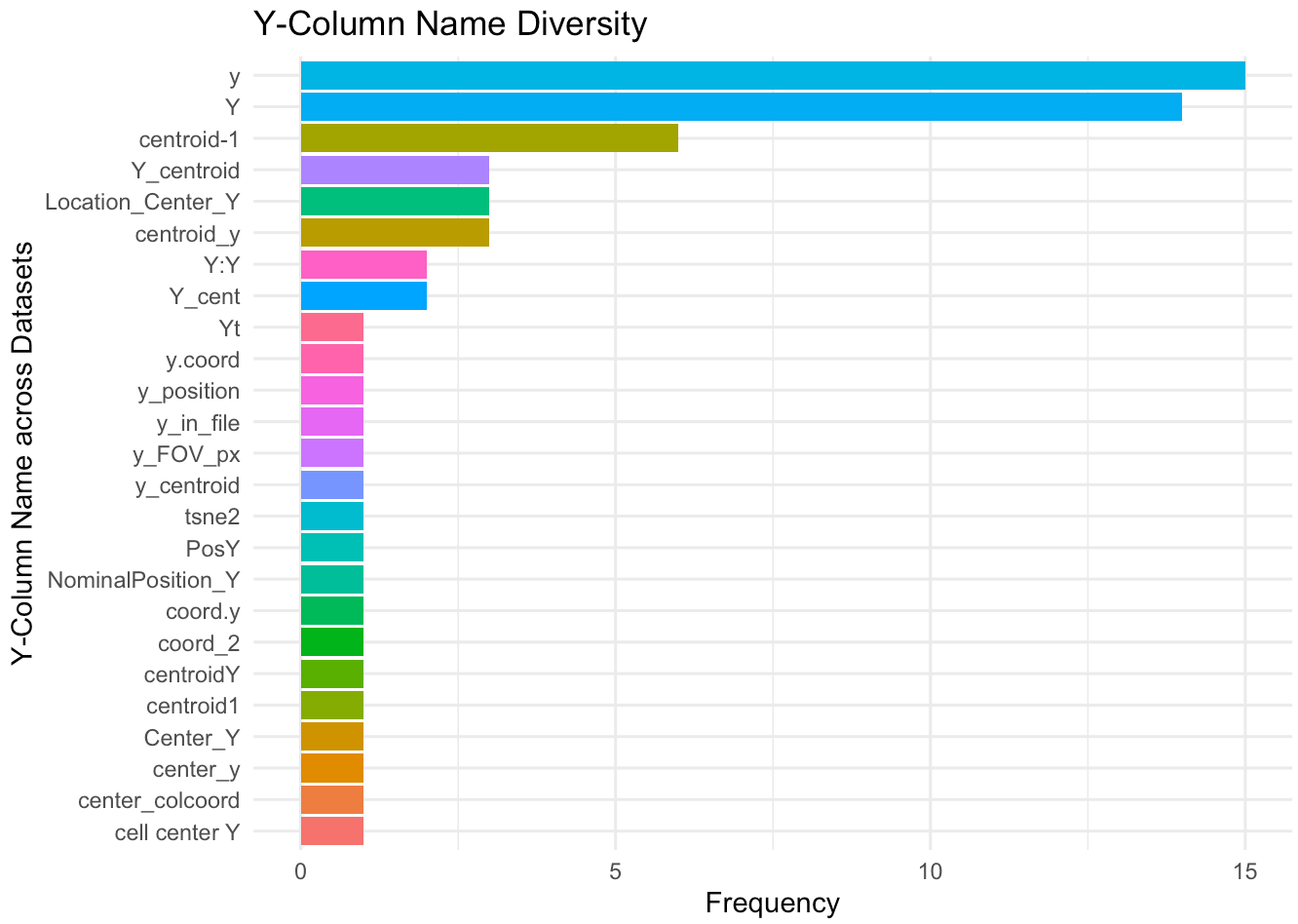
```
print(tissue_freqs)
```

```
## # A tibble: 28 × 2
##    tissue     Frequency
##    <chr>          <int>
##  1 breast             9
##  2 colon              7
##  3 lung               7
##  4 lymph node         7
##  5 esophagus          4
##  6 brain              3
##  7 kidney             3
##  8 head,neck          2
##  9 pancreas           2
## 10 skin               2
## # ℹ 18 more rows
```

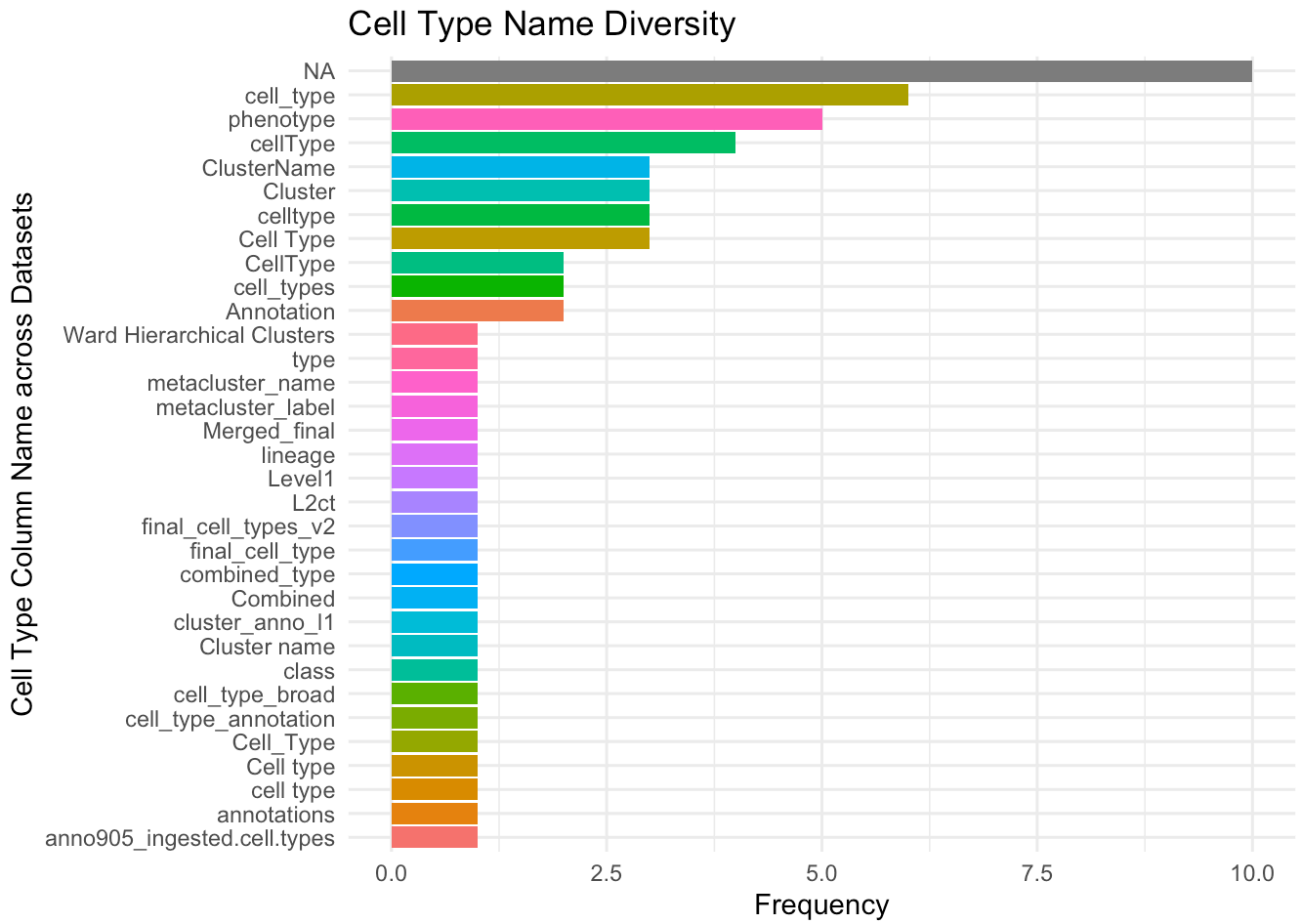
```
x_freqs |>
  ggplot(aes(x = reorder(x_col_name, Frequency), y=Frequency, fill = x_col_name)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "X-Column Name Diversity",
    x = "X-Column Name across Datasets",
  ) +
  coord_flip()
```



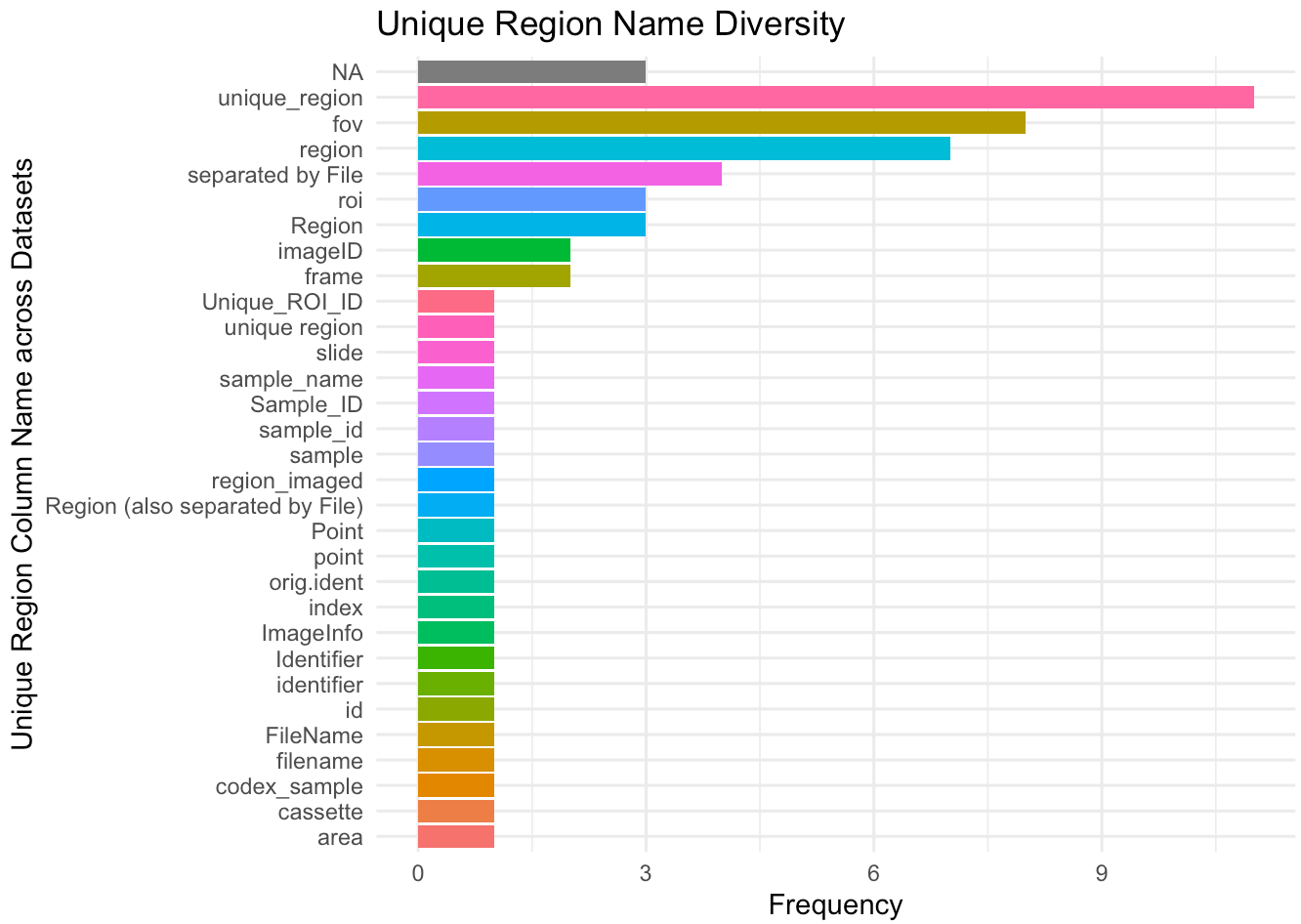

```
y_freqs |>
  ggplot(aes(x = reorder(y_col_name, Frequency), y=Frequency, fill = y_col_name)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Y-Column Name Diversity",
    x = "Y-Column Name across Datasets",
  ) +
  coord_flip()
```



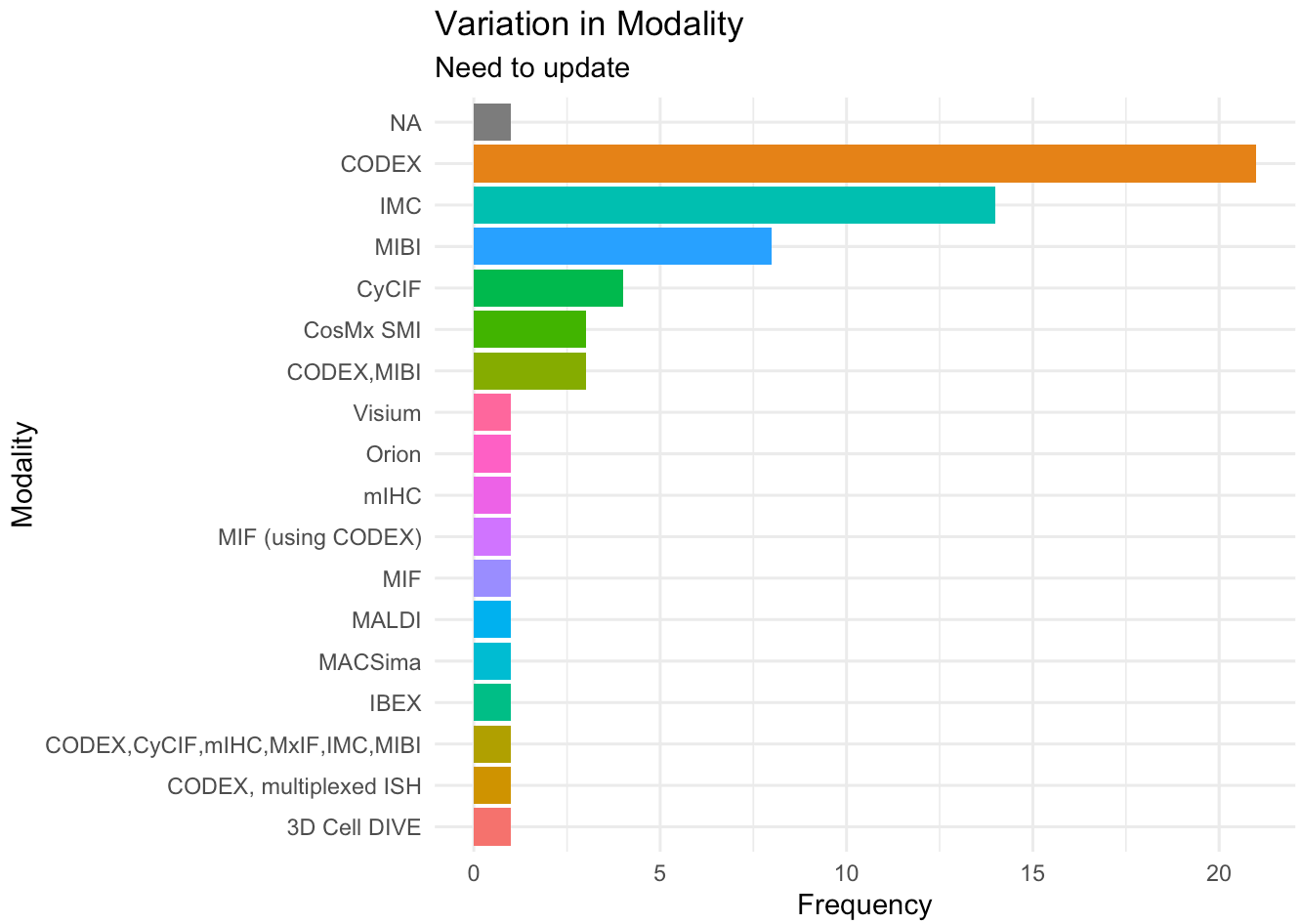

```
celltype_freqs |>
  ggplot(aes(x = reorder(celltype_col_name, Frequency), y=Frequency, fill = celltype_col
_name)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Cell Type Name Diversity",
    x = "Cell Type Column Name across Datasets",
  ) +
  coord_flip()
```



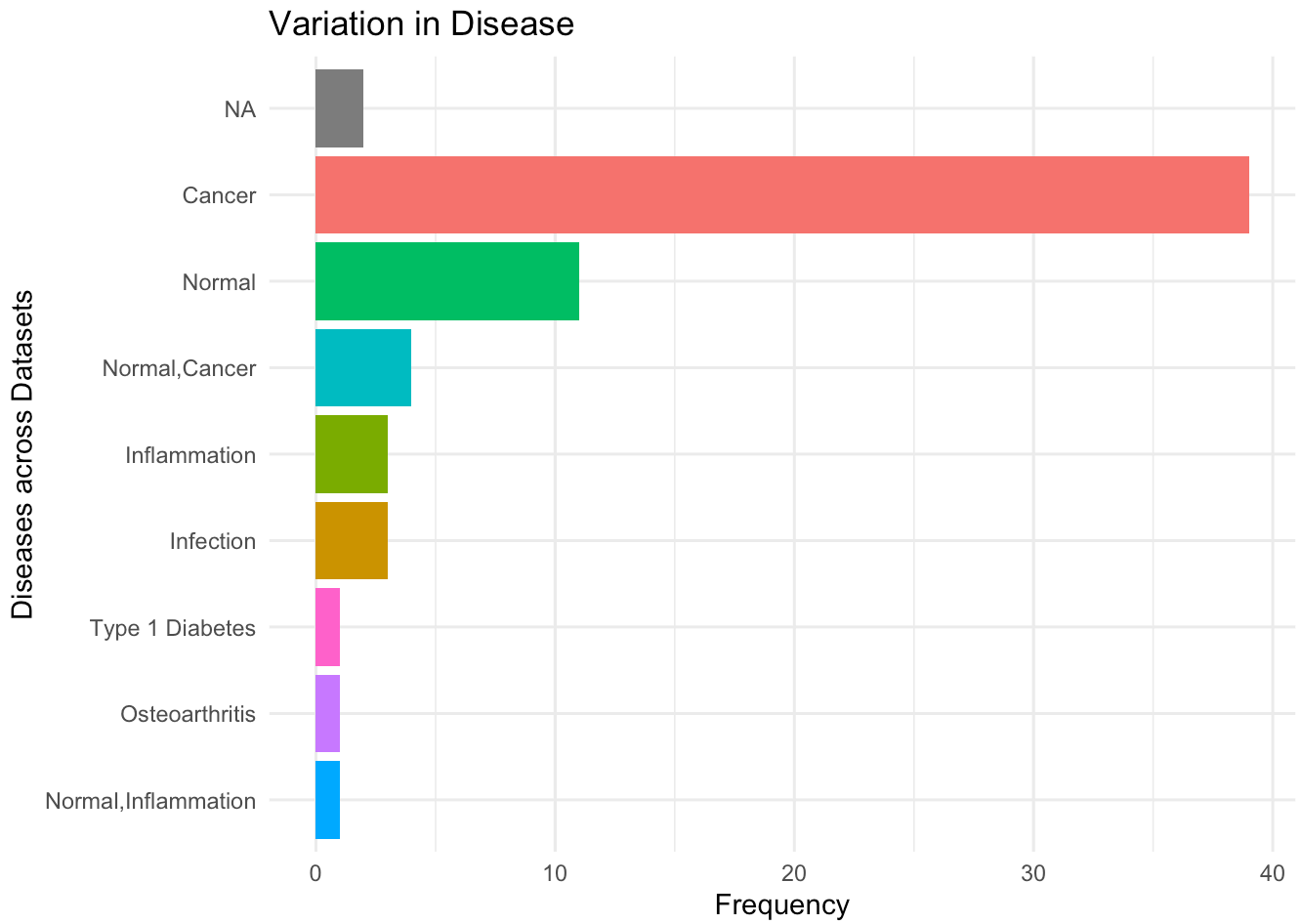

```
uniqueregion_freqs |>
  ggplot(aes(x = reorder(uniqueregion_col_name, Frequency), y=Frequency, fill = uniquere
gion_col_name)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Unique Region Name Diversity",
    x = "Unique Region Column Name across Datasets",
  ) +
  coord_flip()
```



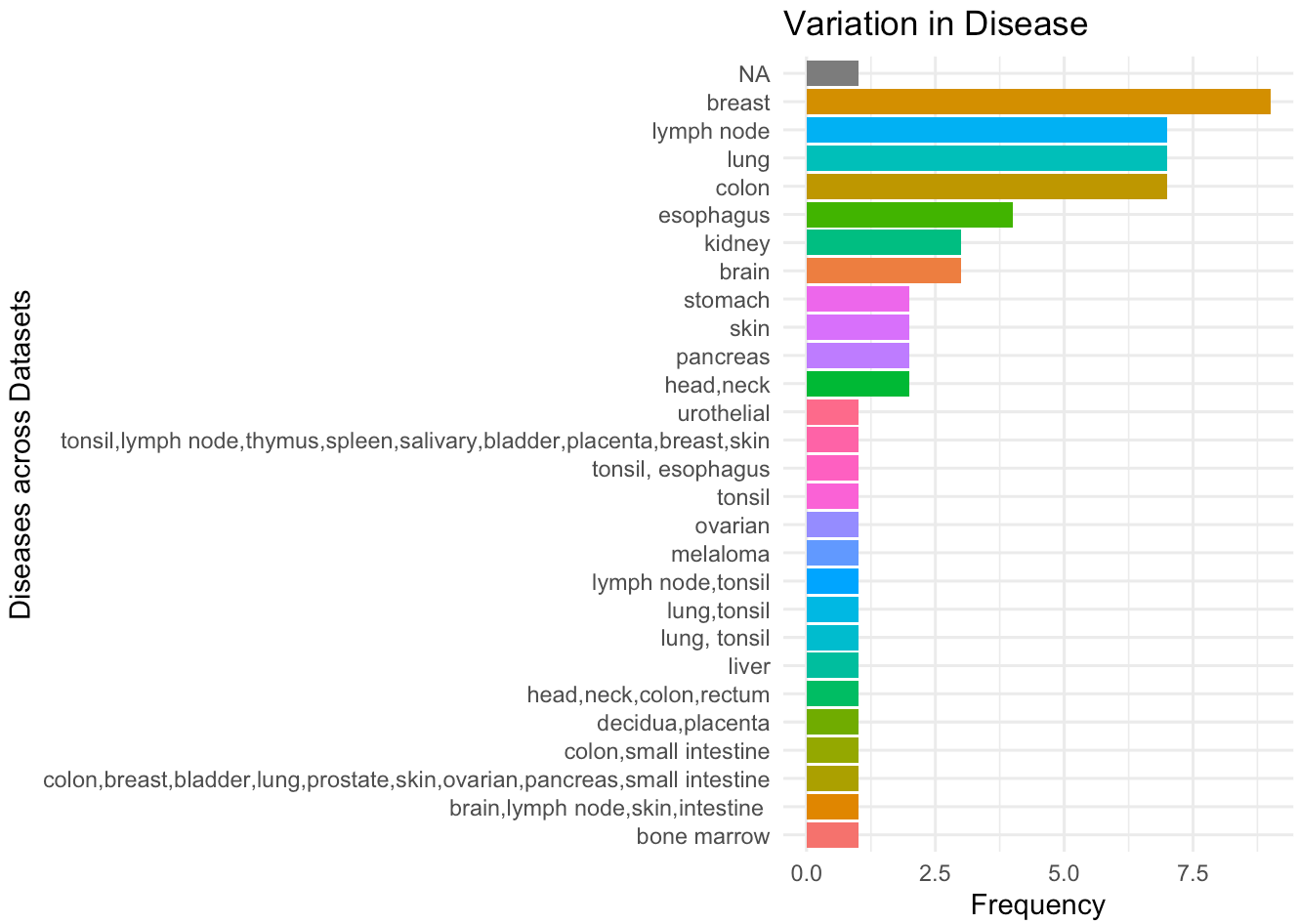

```
modality_freqs |>
  ggplot(aes(x = reorder(modality, Frequency), y=Frequency, fill = modality)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Variation in Modality",
    x = "Modality",
    subtitle = "Need to update"
  ) +
  coord_flip()
```




```
disease_freqs |>
  ggplot(aes(x = reorder(disease, Frequency), y=Frequency, fill = disease)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Variation in Disease",
    x = "Diseases across Datasets",
  ) +
  coord_flip()
```

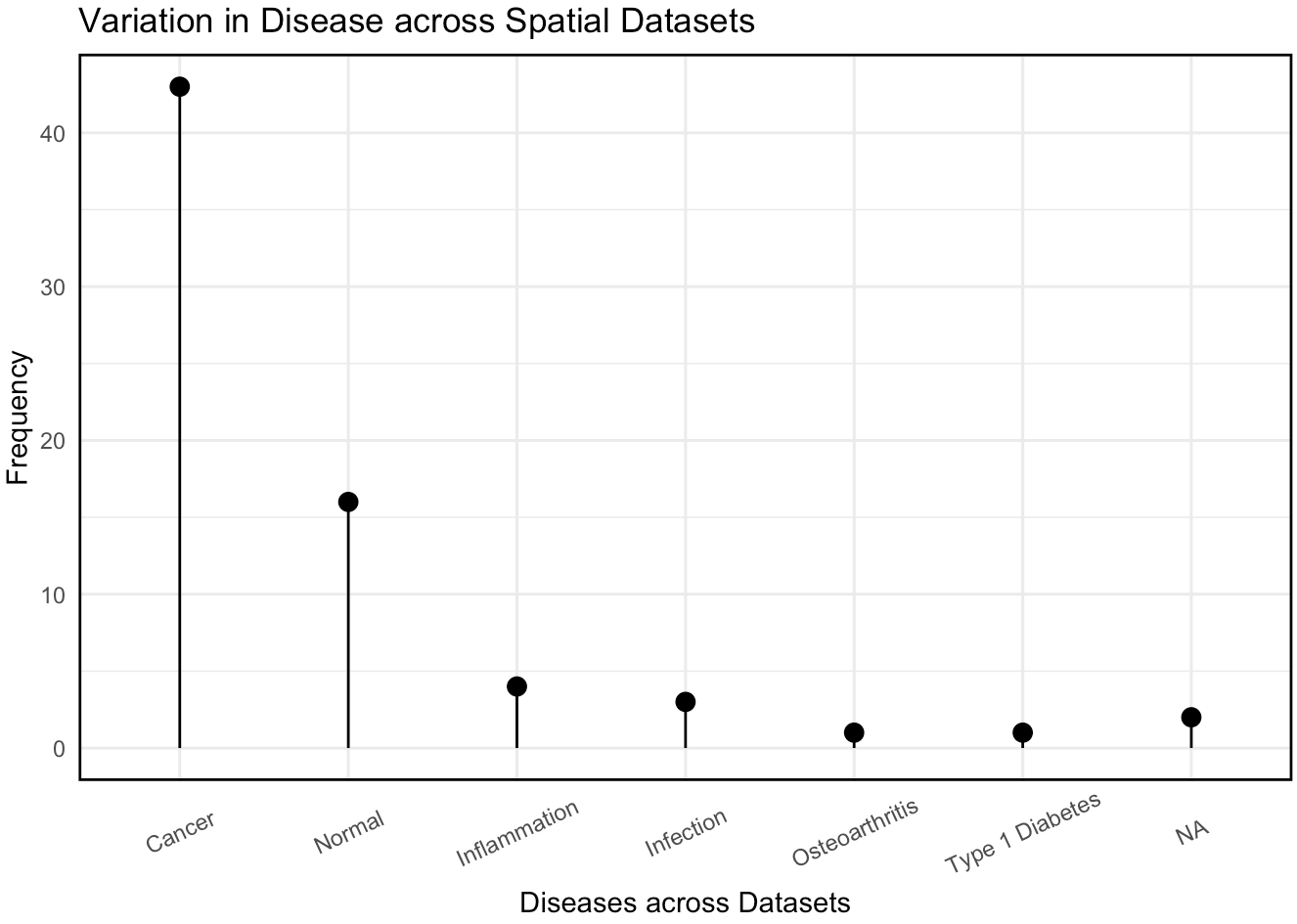



```
tissue_freqs |>
  ggplot(aes(x = reorder(tissue, Frequency), y=Frequency, fill = tissue)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Variation in Disease",
    x = "Diseases across Datasets",
  ) +
  coord_flip()
```




```
data_long <- disease_freqs |>
  mutate(disease_split = strsplit(disease, ",")) |>
  unnest(disease_split)

disease_counts <- data_long |>
  group_by(disease_split) |>
  summarize(Frequency = sum(Frequency)) |>
  arrange(desc(Frequency))

disease_counts |>
  ggplot(aes(x = reorder(disease_split, -Frequency), y=Frequency)) +
  geom_point(size=3) +
  geom_segment(aes(x= disease_split, xend=disease_split, y=0, yend = Frequency)) +
  labs(
    title = "Variation in Disease across Spatial Datasets",
    x = "Diseases across Datasets",
  ) +  theme_minimal() +
  theme(panel.border = element_rect(color = "black", fill = NA, size = 1), axis.text.x=
element_text(angle=25, vjust=.6))
```

```
## Warning: The `size` argument of `element_rect()` is deprecated as of ggplot2 3.4.0.
## ℹ Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



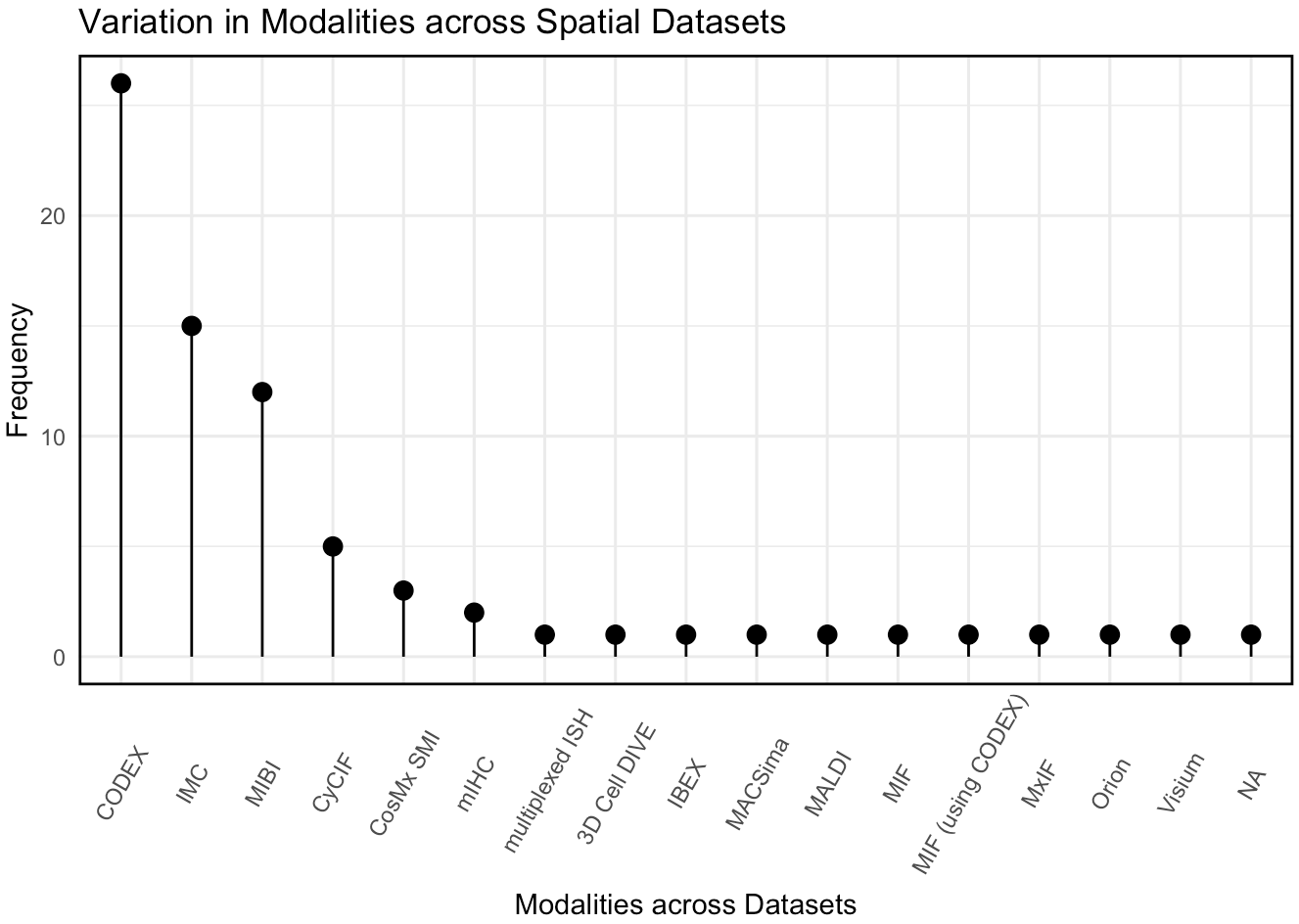

```
print(disease_counts)
```

```
## # A tibble: 7 × 2
##   disease_split   Frequency
##   <chr>               <int>
## 1 Cancer                 43
## 2 Normal                 16
## 3 Inflammation            4
## 4 Infection               3
## 5 <NA>                    2
## 6 Osteoarthritis          1
## 7 Type 1 Diabetes         1
```

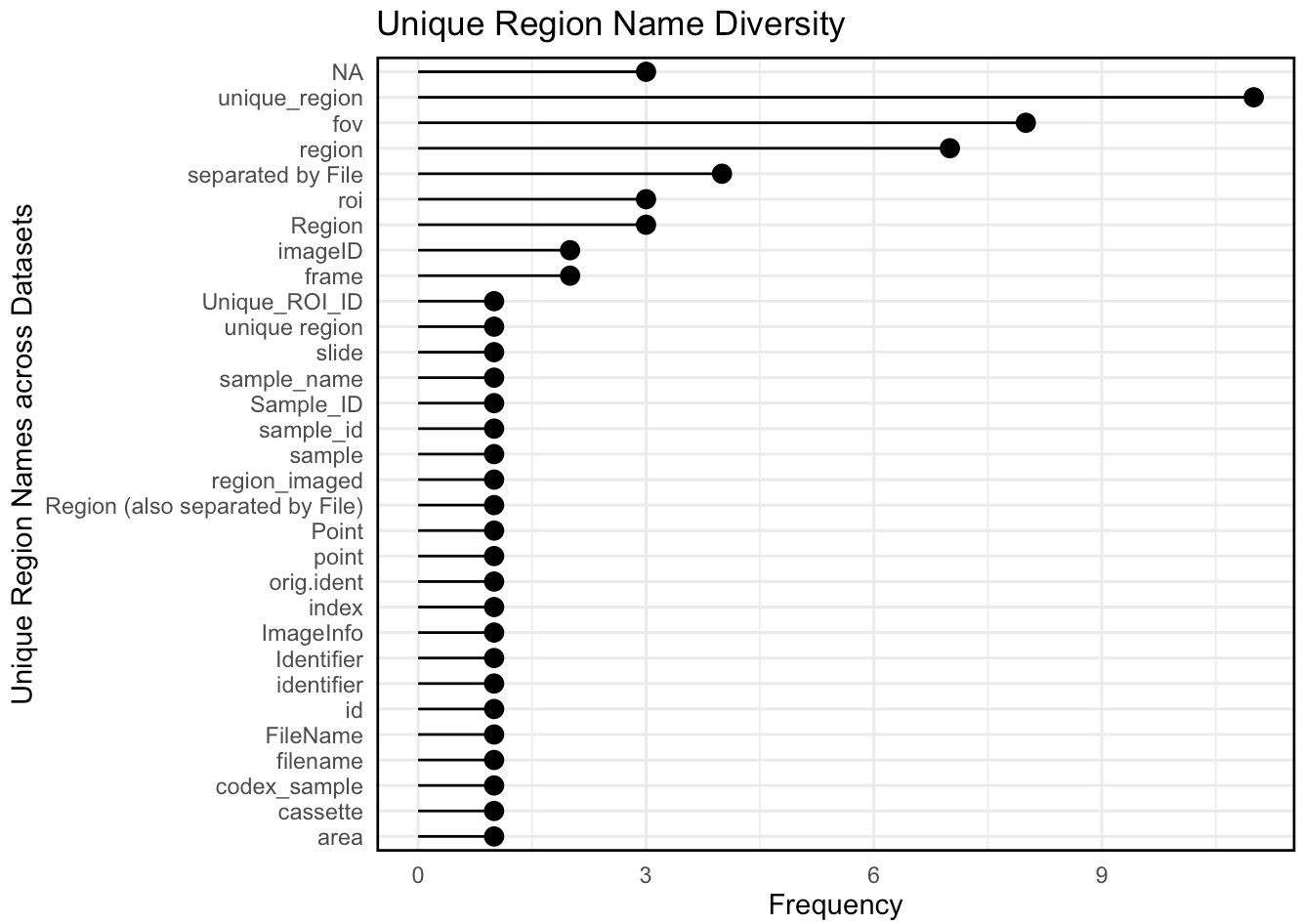
```
moddata_long <- modality_freqs |>
  mutate(modality_split = strsplit(modality, ",")) |>
  unnest(modality_split)

modality_counts <- moddata_long |>
  group_by(modality_split) |>
  summarize(Frequency = sum(Frequency)) |>
  arrange(desc(Frequency))

print(modality_counts)
```

```
## # A tibble: 17 × 2
##    modality_split      Frequency
##    <chr>                   <int>
##  1 "CODEX"                    26
##  2 "IMC"                      15
##  3 "MIBI"                     12
##  4 "CyCIF"                     5
##  5 "CosMx SMI"                 3
##  6 "mIHC"                      2
##  7 " multiplexed ISH"          1
##  8 "3D Cell DIVE"              1
##  9 "IBEX"                      1
## 10 "MACSima"                   1
## 11 "MALDI"                     1
## 12 "MIF"                       1
## 13 "MIF (using CODEX)"         1
## 14 "MxIF"                      1
## 15 "Orion"                     1
## 16 "Visium"                    1
## 17  <NA>                       1
```

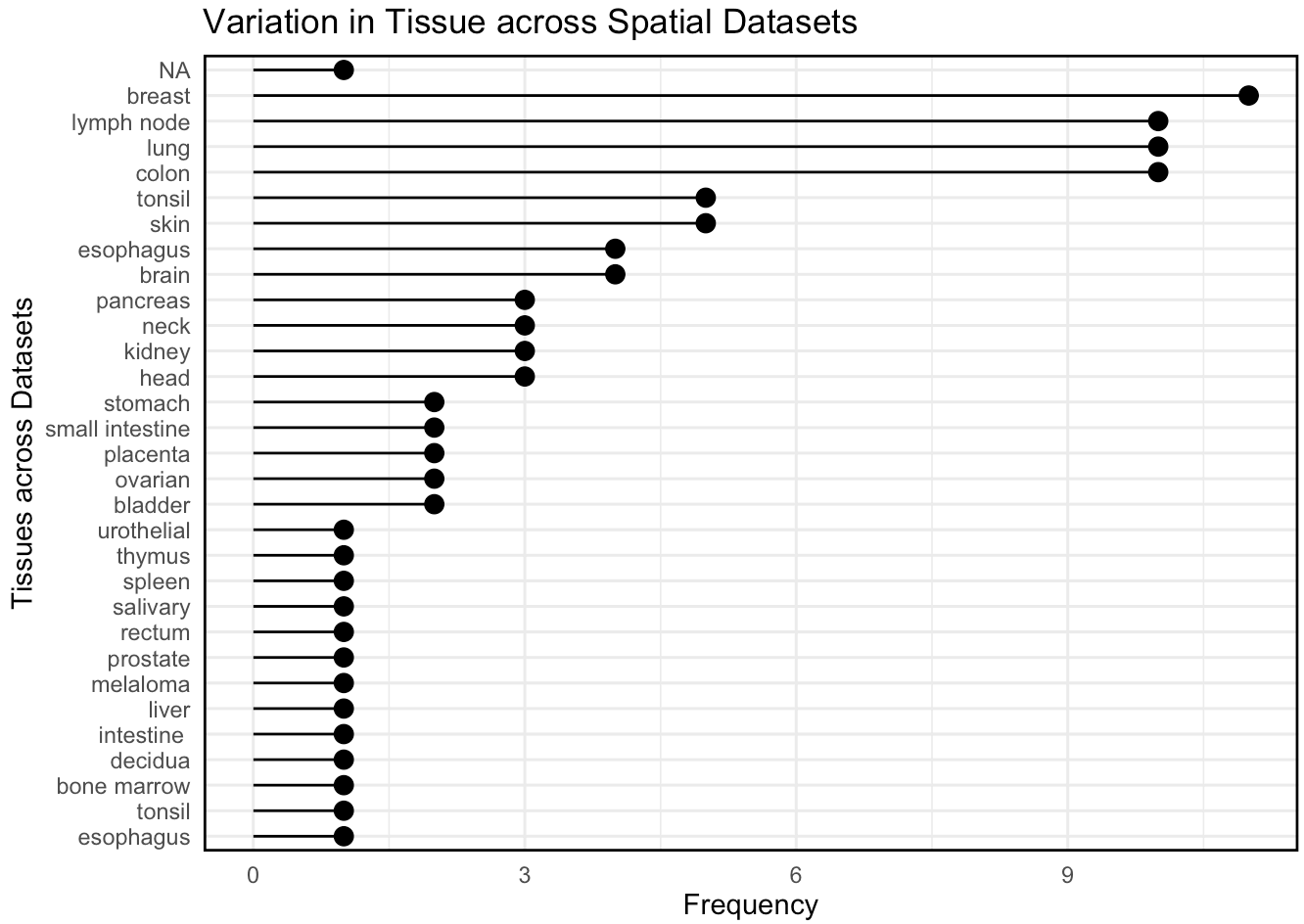
```
modality_counts |>
  ggplot(aes(x = reorder(modality_split, -Frequency), y=Frequency)) +
  geom_point(size=3) +
  geom_segment(aes(x= modality_split, xend=modality_split, y=0, yend = Frequency)) +
  labs(
    title = "Variation in Modalities across Spatial Datasets",
    x = "Modalities across Datasets",
  ) +  theme_minimal() +
  theme(panel.border = element_rect(color = "black", fill = NA, size = 1), axis.text.x=
element_text(angle=60, vjust=.55))
```




```
uniqueregion_freqs |>
  ggplot(aes(x = reorder(uniqueregion_col_name, Frequency), y=Frequency)) +
  geom_point(size=3) +
  geom_segment(aes(x= uniqueregion_col_name, xend=uniqueregion_col_name, y=0, yend = Fre
quency)) +
  labs(
    title = "Unique Region Name Diversity",
    x = "Unique Region Names across Datasets",
  ) +  theme_minimal() +
  theme(panel.border = element_rect(color = "black", fill = NA, size = 1), axis.text.x=
element_text(angle=0, vjust=.6)) +
  coord_flip()
```




```
tissuedata_long <- tissue_freqs |>
  mutate(tissue_split = strsplit(tissue, ",")) |>
  unnest(tissue_split)

tissue_counts <- tissuedata_long |>
  group_by(tissue_split) |>
  summarize(Frequency = sum(Frequency)) |>
  arrange(desc(Frequency))

tissue_counts |>
  ggplot(aes(x = reorder(tissue_split, Frequency), y=Frequency)) +
  geom_point(size=3) +
  geom_segment(aes(x= tissue_split, xend=tissue_split, y=0, yend = Frequency)) +
  labs(
    title = "Variation in Tissue across Spatial Datasets",
    x = "Tissues across Datasets",
  ) +  theme_minimal() +
  theme(panel.border = element_rect(color = "black", fill = NA, size = 1), axis.text.x=
element_text(angle=0, vjust=.6)) +
  coord_flip()
```



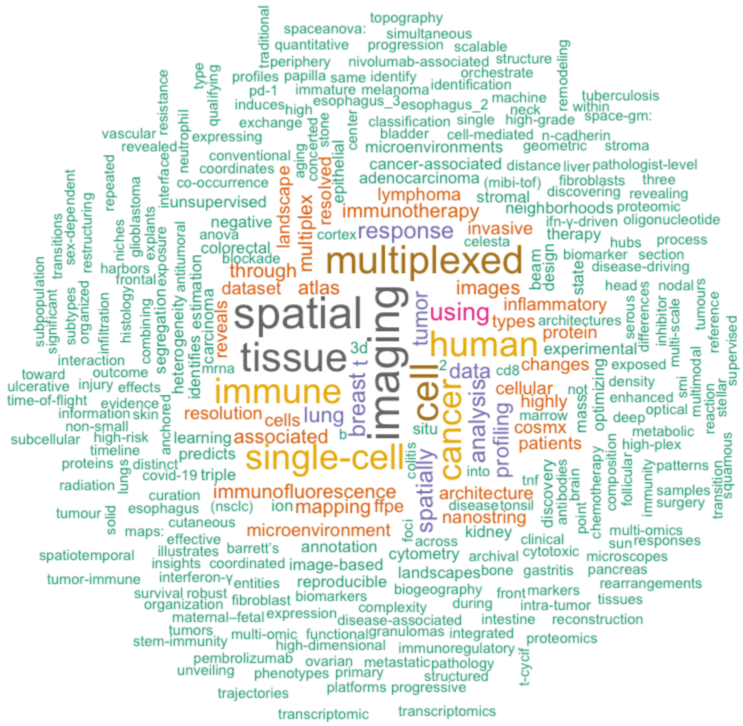

```
print(tissue_counts)
```

```
## # A tibble: 31 × 2
##    tissue_split Frequency
##    <chr>            <int>
##  1 breast              11
##  2 colon               10
##  3 lung                10
##  4 lymph node          10
##  5 skin                 5
##  6 tonsil               5
##  7 brain                4
##  8 esophagus            4
##  9 head                 3
## 10 kidney               3
## # ℹ 21 more rows
```

```
library(wordcloud)

title_freqs <- data |>
  group_by(title) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

title_freqs <- title_freqs |>
  rename(titles = title)

titledata_long <- title_freqs |>
  mutate(title_split = strsplit(titles, " ")) |>
  unnest(title_split)

titledata_long <- titledata_long |>
  mutate(title_split = gsub(",", "", title_split)) |>
  mutate(title_split = tolower(title_split))

title_counts <- titledata_long |>
  group_by(title_split) |>
  summarize(Frequency = sum(Frequency)) |>
  arrange(desc(Frequency))

extra_words <- c("of", "and", "in", "the", "by", "to", "with", "a", "for", "from", "an",
"at", "the", "is", "but")
title_clean_counts <- title_counts |>
  filter(!title_split %in% extra_words)

print(title_clean_counts)
```

```
## # A tibble: 306 × 2
##    title_split Frequency
##    <chr>           <int>
##  1 imaging            18
##  2 spatial            18
##  3 tissue             16
##  4 cell               15
##  5 multiplexed        15
##  6 human              13
##  7 immune             13
##  8 single-cell        13
##  9 cancer             12
## 10 using               7
## # ℹ 296 more rows
```

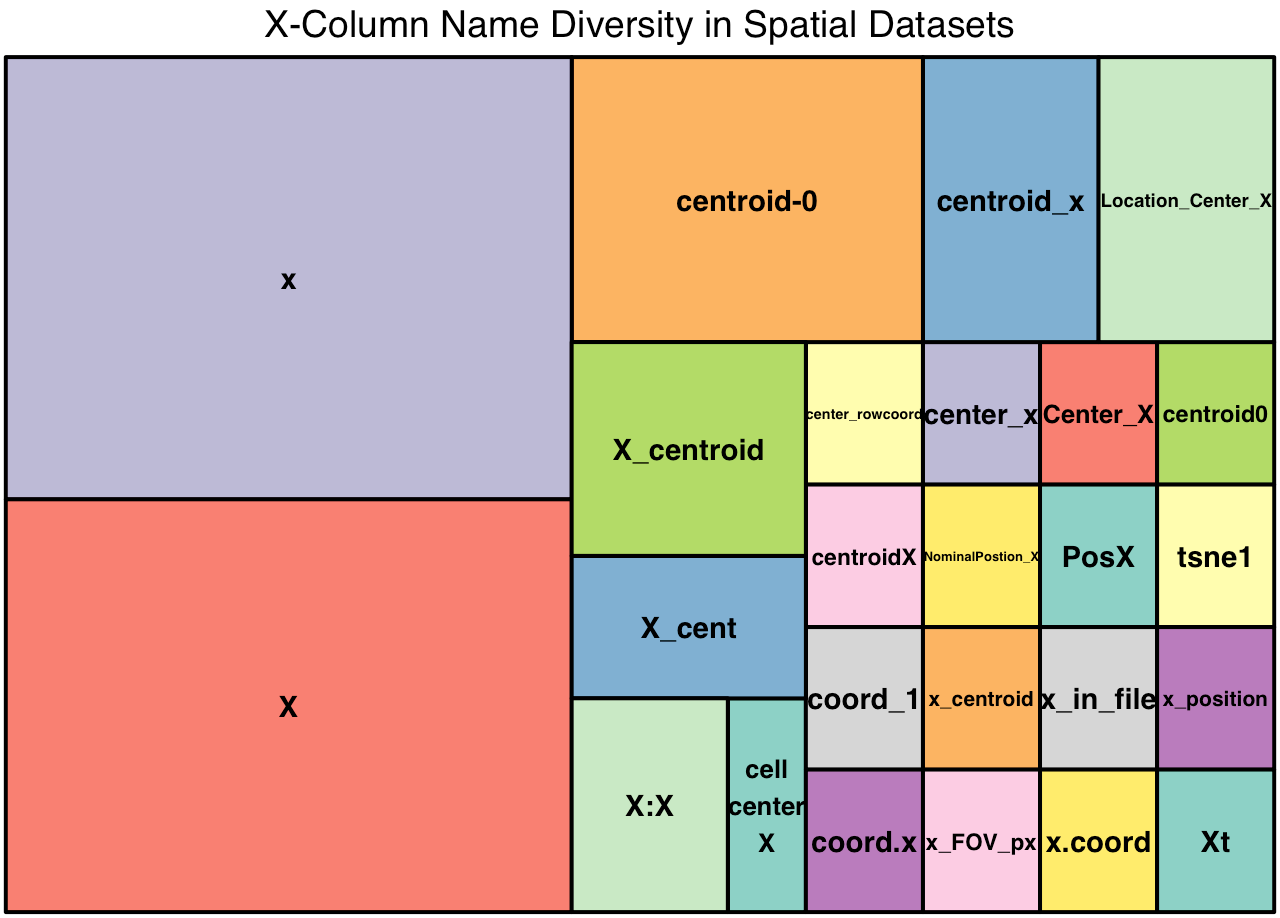
```
wordcloud(words = title_clean_counts$title_split, freq = title_clean_counts$Frequency, m
in.freq = 1, scale = c(1.4, 0.4),
          random.order = FALSE, rot.per = .25, colors=brewer.pal(8, "Dark2"))
```




```r
library(treemap)

head(x_freqs)
```

```
## # A tibble: 6 × 2
##   x_col_name        Frequency
##   <chr>                 <int>
## 1 x                        15
## 2 X                        14
## 3 centroid-0                6
## 4 Location_Center_X         3
## 5 X_centroid                3
## 6 centroid_x                3
```

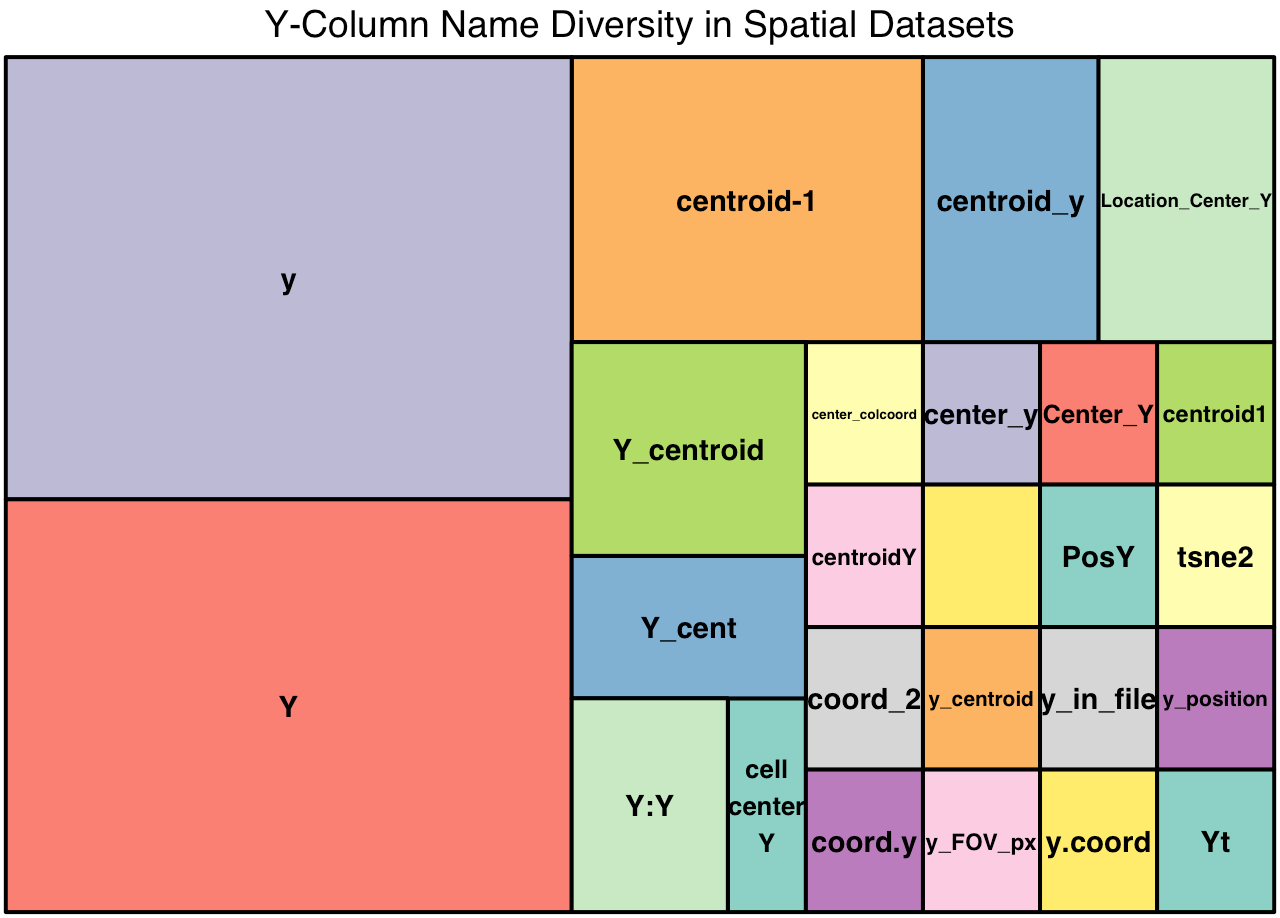
```r
treemap(x_freqs,
        index = c("x_col_name"),
        vSize = "Frequency",
        vColor = "Frequency",
        title = "X-Column Name Diversity in Spatial Datasets",
        palette = "Set3",
        draw = TRUE)
```




```
head(y_freqs)
```

```
## # A tibble: 6 × 2
##   y_col_name        Frequency
##   <chr>                 <int>
## 1 y                        15
## 2 Y                        14
## 3 centroid-1                6
## 4 Location_Center_Y         3
## 5 Y_centroid                3
## 6 centroid_y                3
```

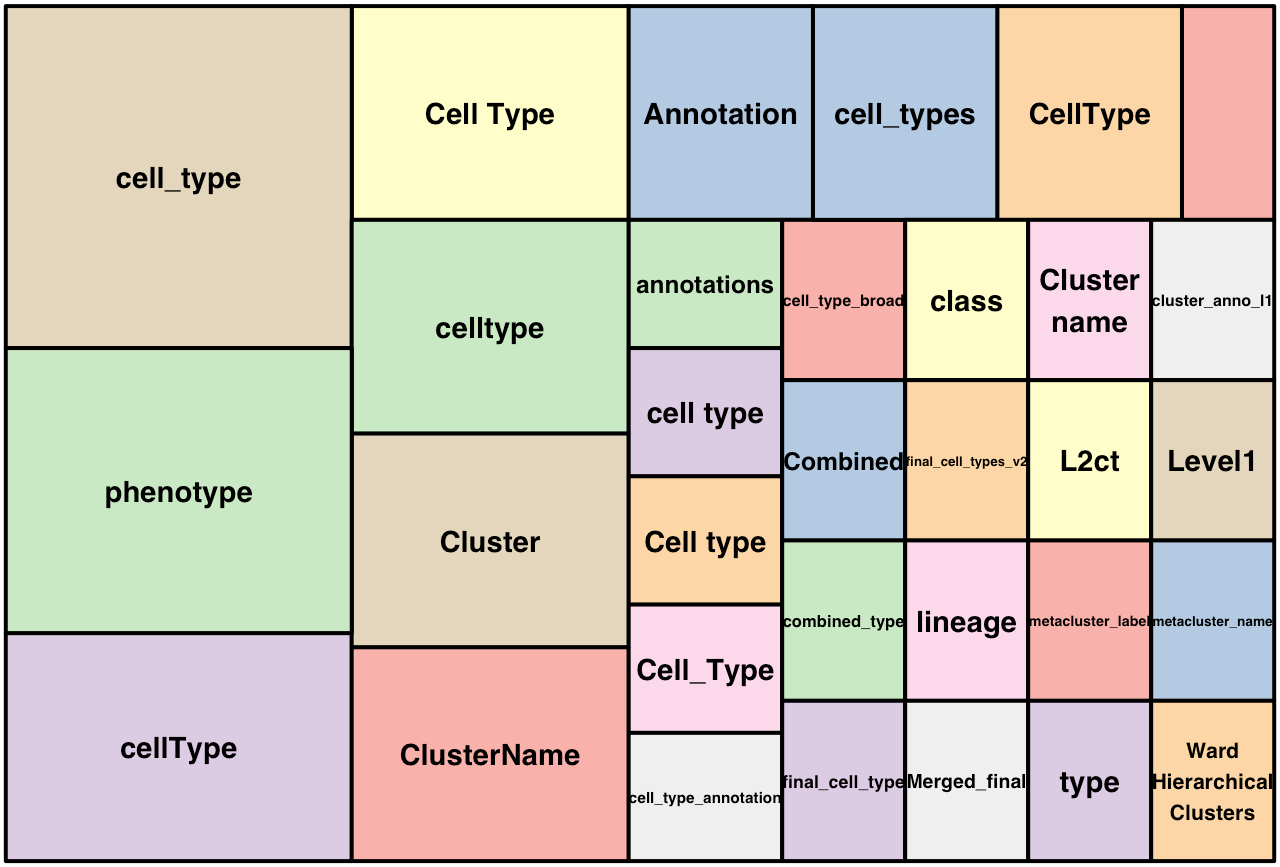
```
treemap(y_freqs,
        index = c("y_col_name"),
        vSize = "Frequency",
        vColor = "Frequency",
        title = "Y-Column Name Diversity in Spatial Datasets",
        palette = "Set3",
        draw = TRUE)
```




```
head(celltype_freqs)
```

```
## # A tibble: 6 × 2
##   celltype_col_name Frequency
##   <chr>                 <int>
## 1 <NA>                     10
## 2 cell_type                 6
## 3 phenotype                 5
## 4 cellType                  4
## 5 Cell Type                 3
## 6 Cluster                   3
```

```
treemap(celltype_freqs,
        index = c("celltype_col_name"),
        vSize = "Frequency",
        vColor = "Frequency",
        title = "Cell Type Name Diversity in Spatial Datasets",
        palette = "Pastel1",
        draw = TRUE)
```




```
head(uniqueregion_freqs)
```

```
## # A tibble: 6 × 2
##   uniqueregion_col_name Frequency
##   <chr>                     <int>
## 1 unique_region                11
## 2 fov                           8
## 3 region                        7
## 4 separated by File             4
## 5 Region                        3
## 6 roi                           3
```

```
treemap(uniqueregion_freqs,
        index = c("uniqueregion_col_name"),
        vSize = "Frequency",
        vColor = "Frequency",
        title = "Unique Region Name Diversity in Spatial Datasets",
        palette = "Pastel2",
        draw = TRUE)
```

Unique Region Name Diversity in Spatial Datasets