LocsOfSpatialData_Analysis

```
library(RColorBrewer)
```

```
## Warning: package 'RColorBrewer' was built under R version 4.3.3
```

```
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
```

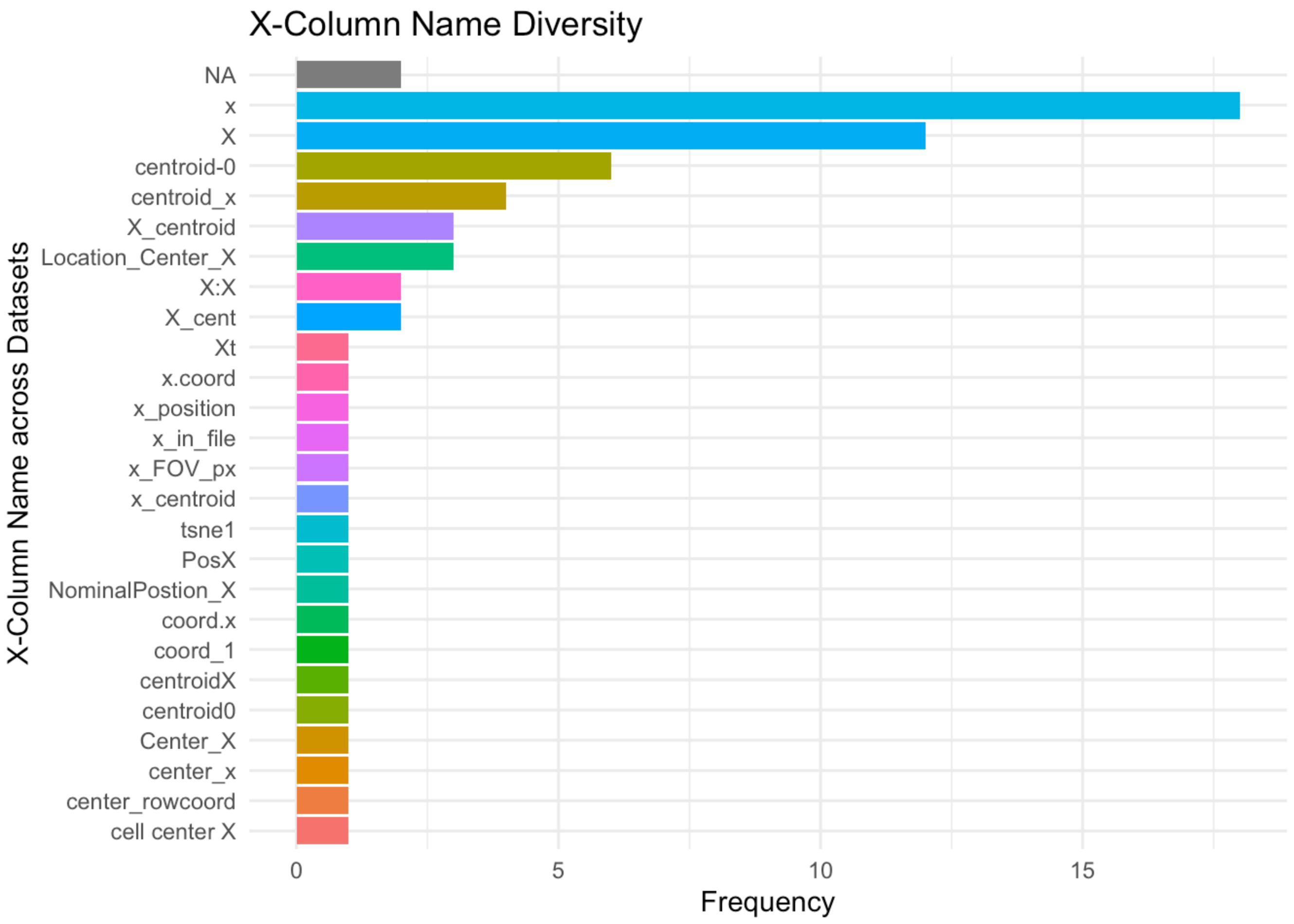
```
library(wordcloud)
library(dplyr)
library(ggplot2)
library(tidytext)
library(treemap)
library(paletteer)
library(stringr)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

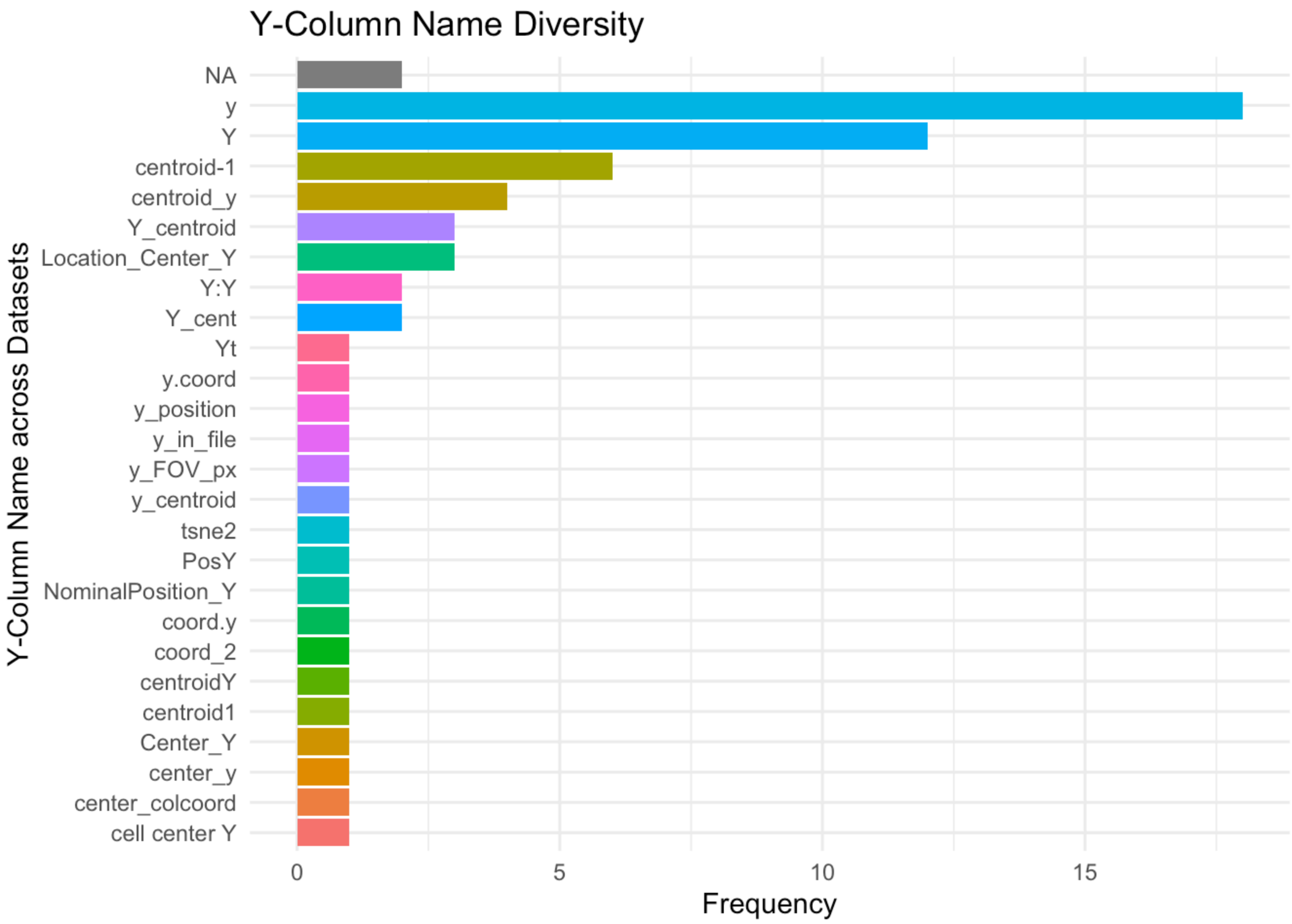
```
data = read.csv("/Users/riahcul/Downloads/020425_locsofspatialdata.csv")
data[data == ""] <- NA
```

```
data <- data |>
  mutate(year = str_extract(folder_name, "\\d{4}"))

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, year) |>
  summarize(Frequency = n(), .groups = "drop") |>
  arrange(desc(Frequency))

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

tissue_freqs <- data |>
  group_by(tissue, year) |>
  summarize(Frequency = n(), .groups = "drop") |>
  arrange(desc(Frequency))
```

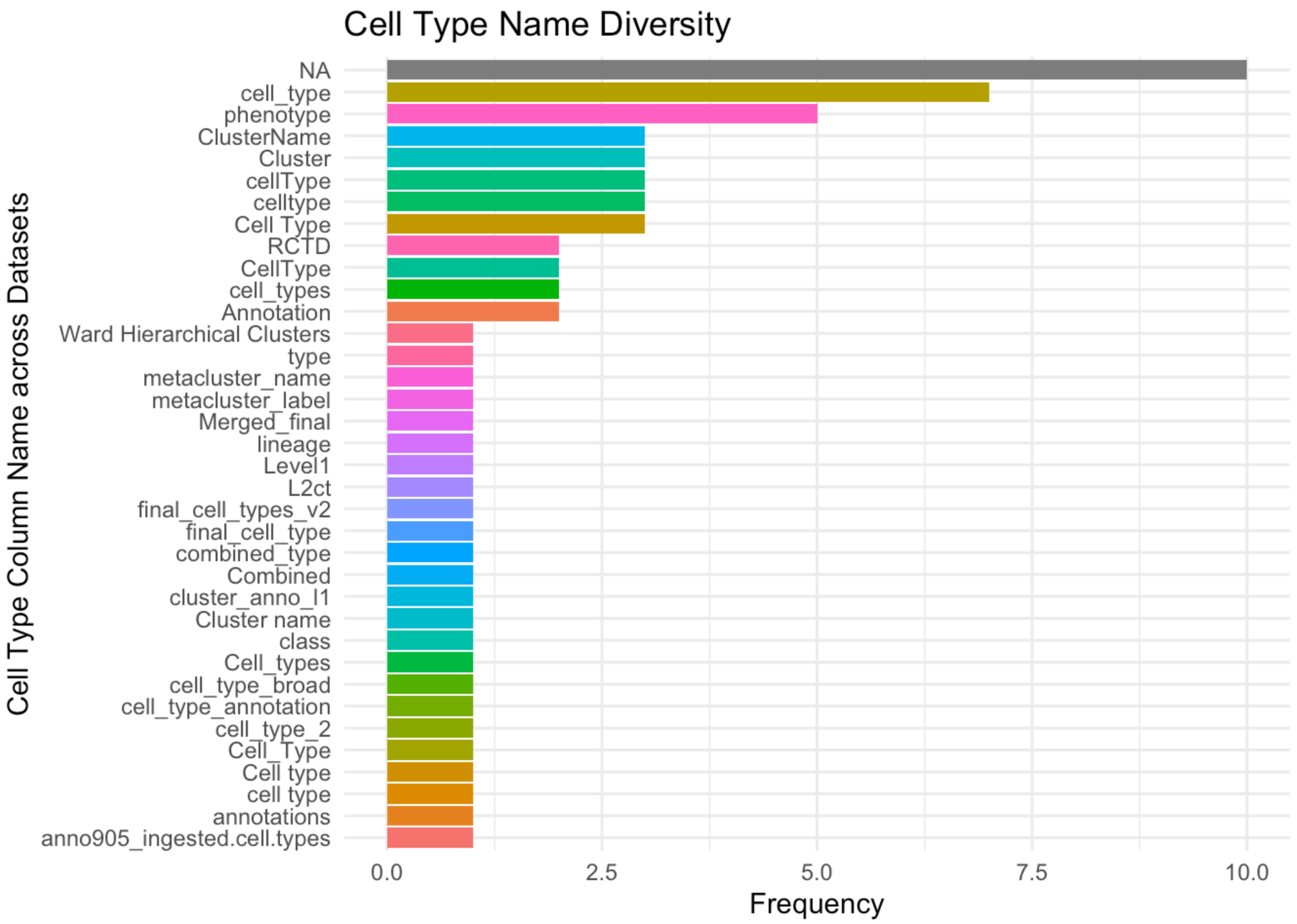
```
x_freqs |>
  ggplot(aes(x = reorder(x_col_name, Frequency), y=Frequency, fill = x_col_name)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "X-Column Name Diversity",
    x = "X-Column Name across Datasets",
  ) +
  coord_flip()
```



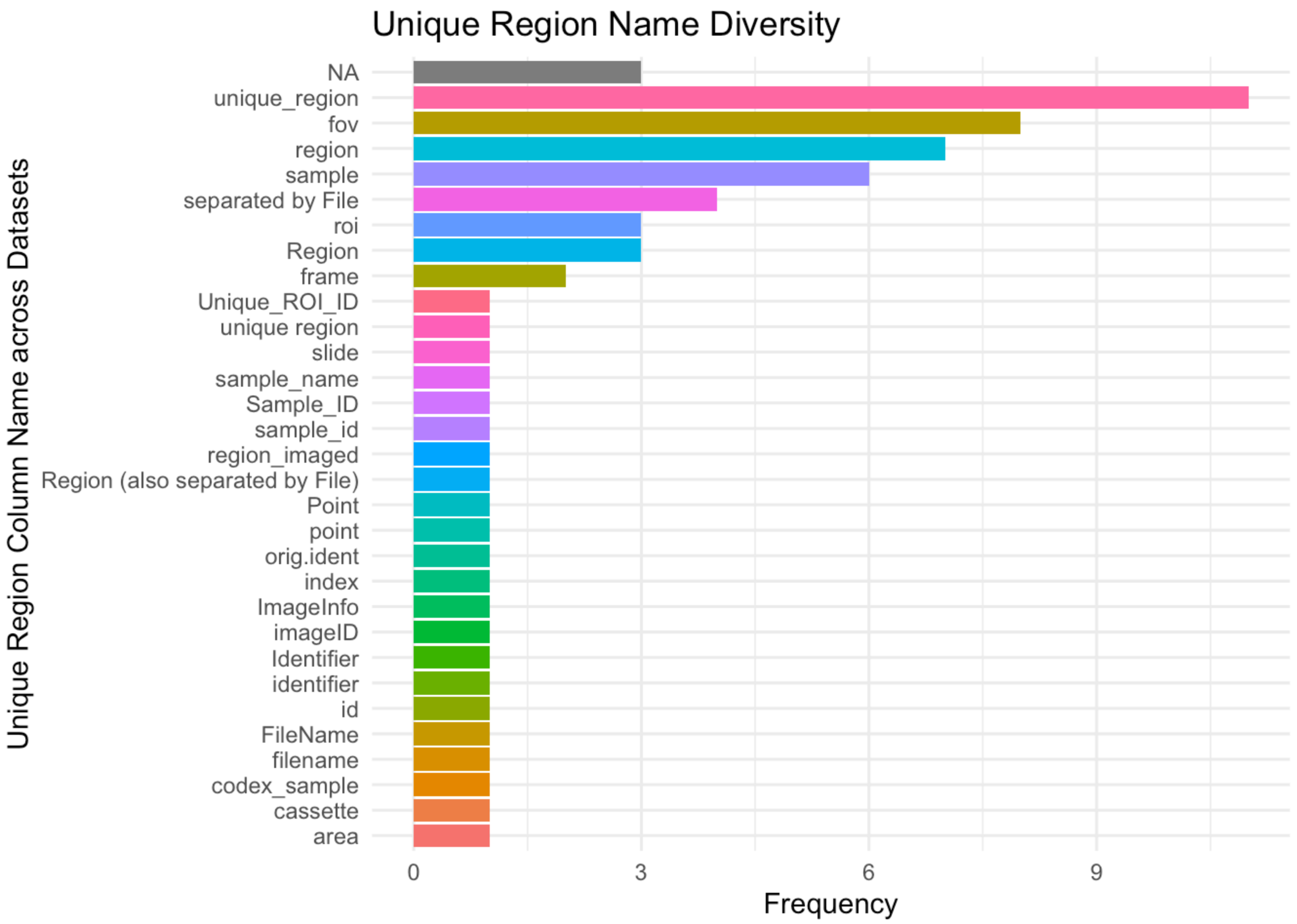
```
y_freqs |>
  ggplot(aes(x = reorder(y_col_name, Frequency), y=Frequency, fill = y_col_name)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Y-Column Name Diversity",
    x = "Y-Column Name across Datasets",
  ) +
  coord_flip()
```



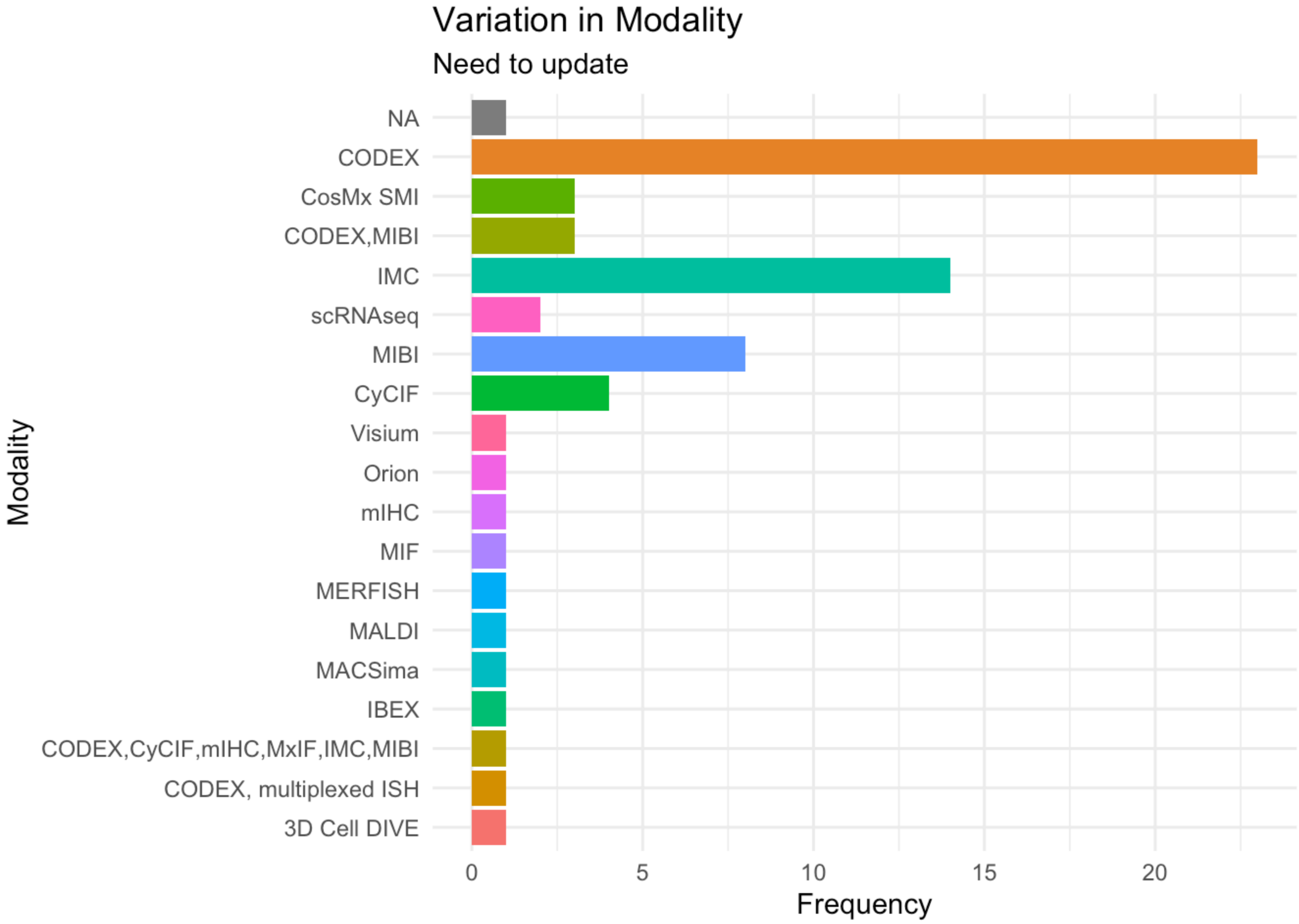
```
celltype_freqs |>
  ggplot(aes(x = reorder(celltype_col_name, Frequency), y=Frequency, fill = celltype_col
_name)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Cell Type Name Diversity",
    x = "Cell Type Column Name across Datasets",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



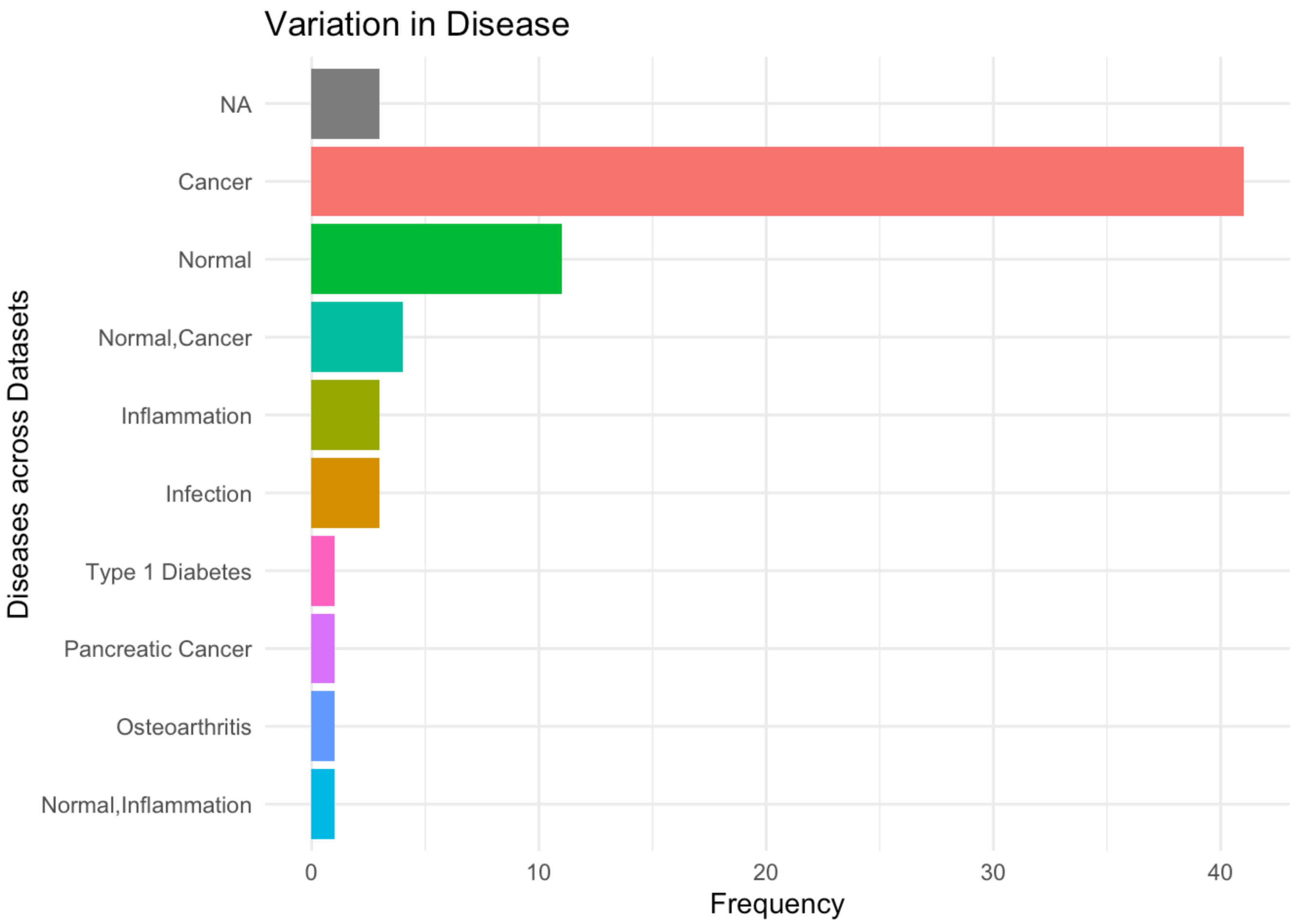
```
uniqueregion_freqs |>
  ggplot(aes(x = reorder(uniqueregion_col_name, Frequency), y=Frequency, fill = uniquere
gion_col_name)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Unique Region Name Diversity",
    x = "Unique Region Column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



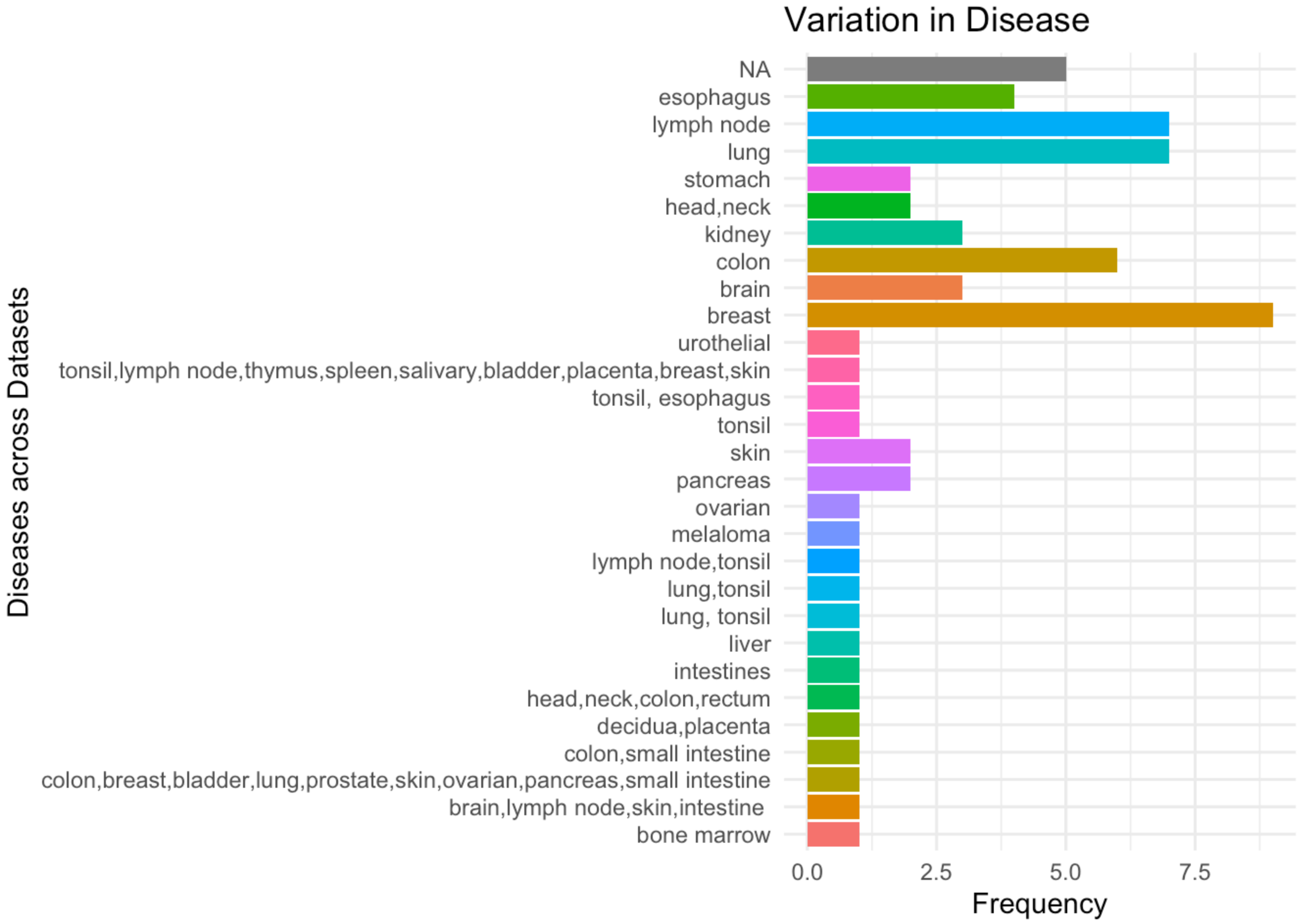
```
modality_freqs |>
  ggplot(aes(x = reorder(modality, Frequency), y=Frequency, fill = modality)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Variation in Modality",
    x = "Modality",
    subtitle = "Need to update"
  ) +
  coord_flip()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



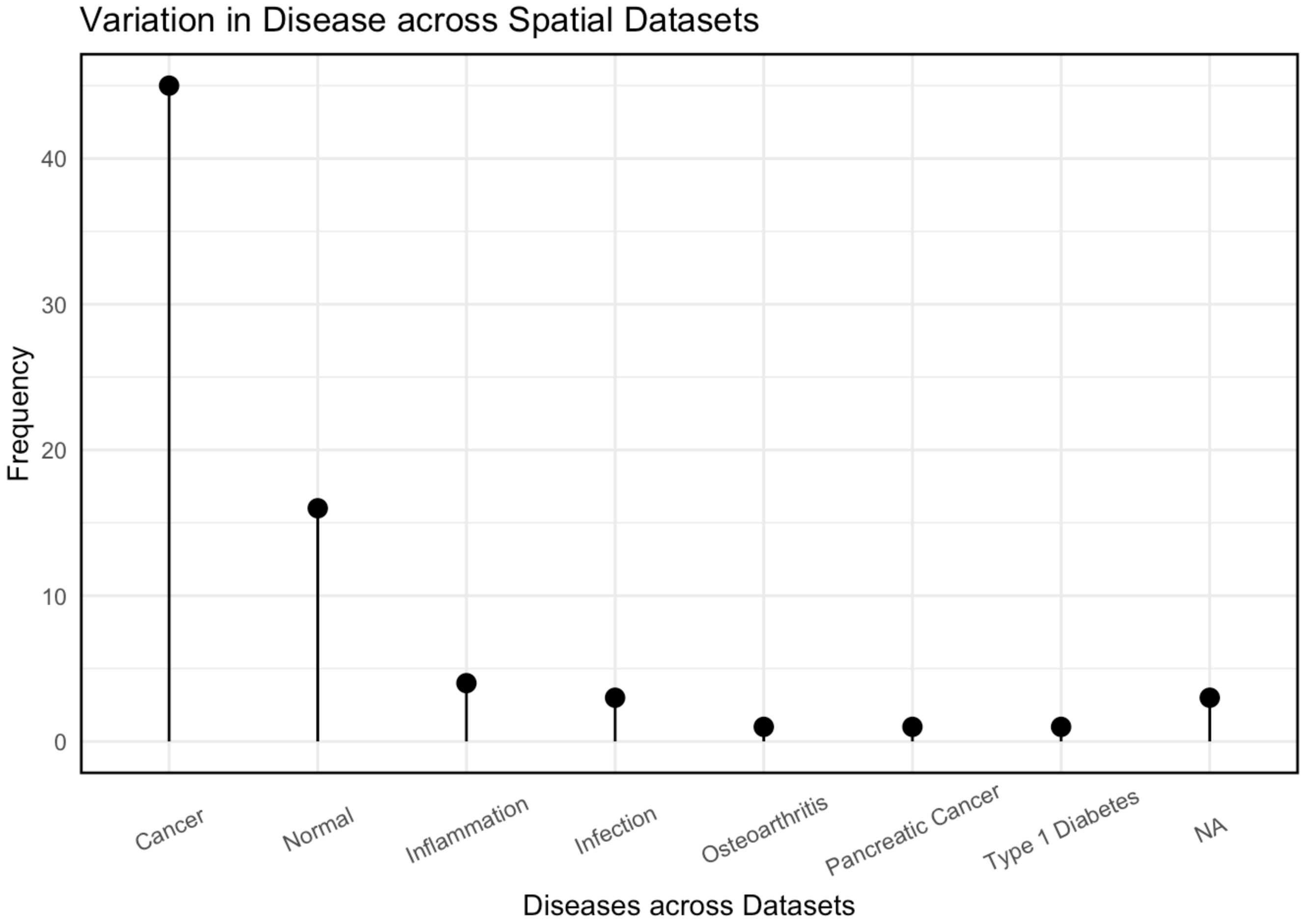
```
disease_freqs |>
  ggplot(aes(x = reorder(disease, Frequency), y=Frequency, fill = disease)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title = "Variation in Disease",
    x = "Diseases across Datasets",
  ) +
  coord_flip()
```



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



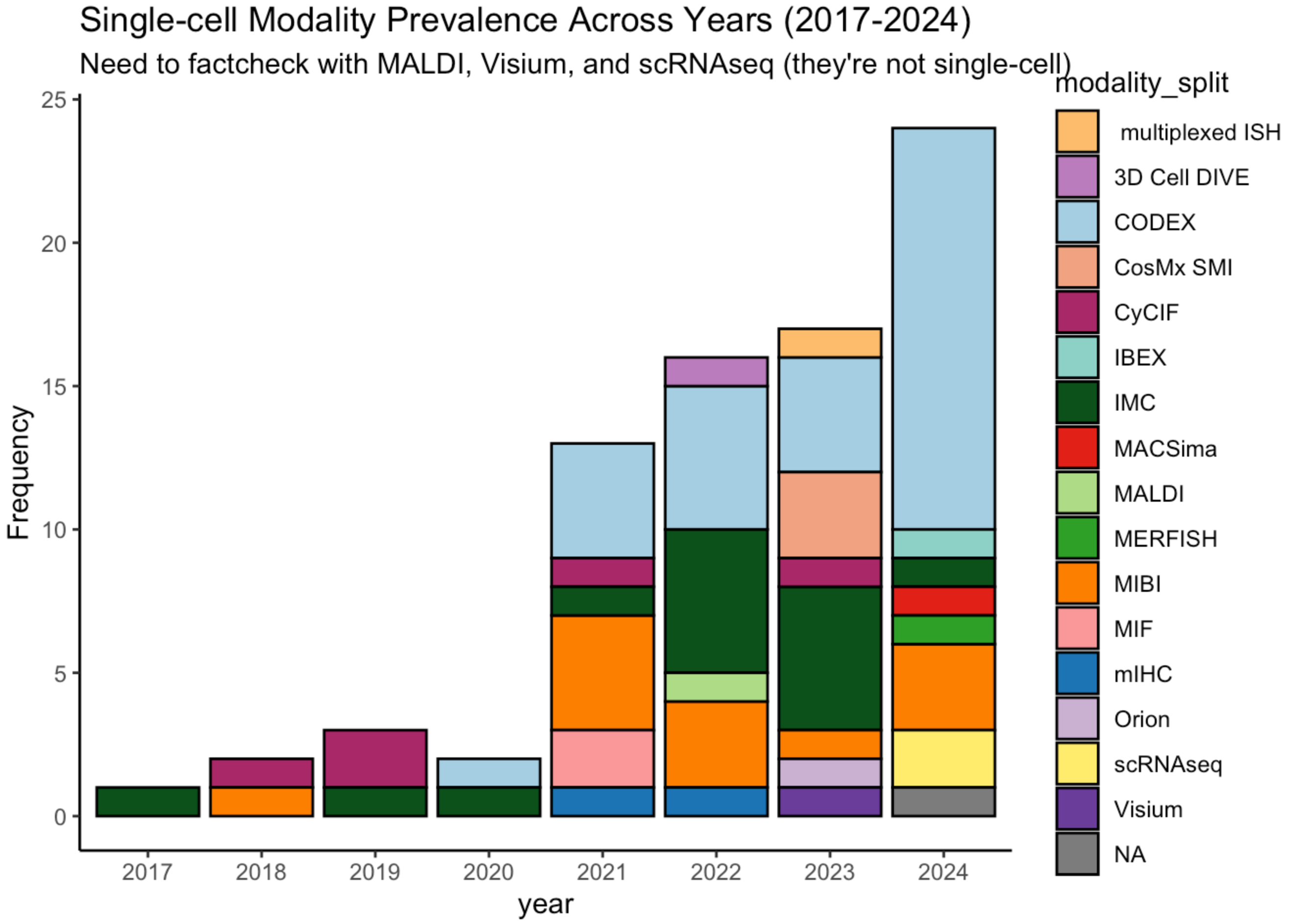
```
data_long <- disease_freqs |>
  mutate(disease_split = strsplit(disease, ",")) |>
  unnest(disease_split)

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, linewidth = 1),
        axis.text.x= element_text(angle=25, vjust=.6))
```



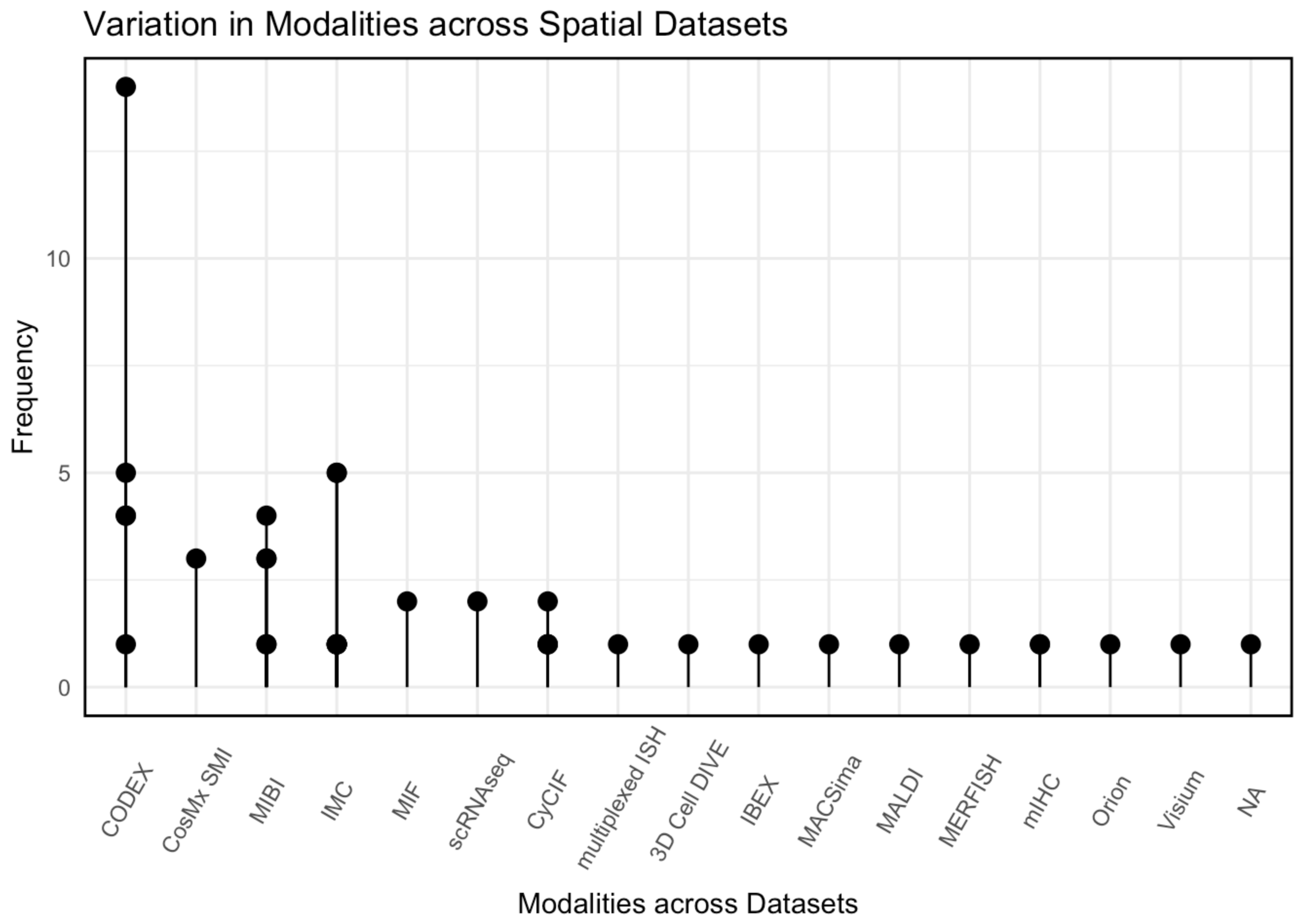
```
col_vector <- c("#FDBF6F", "#BC80BD", "#A6CEE3", "#F4A582", "#ab296a", "#8DD3C7", "#0c541f", "#E31A1C", "#B2DF8A", "#33A02C", "#FF7F00", "#FB9A99", "#1F78B4", "#CAB2D6", "#FFED6F", "#6A3D9A")

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

modality_counts <- moddata_long |>
  group_by(modality_split, year) |>
  summarize(Frequency = sum(Frequency), .groups = "drop") |>
  arrange(desc(Frequency))

print(modality_counts)
```

```
## # A tibble: 35 × 3
##    modality_split year  Frequency
##    <chr>          <chr>     <int>
##  1 CODEX          2024         14
##  2 CODEX          2022          5
##  3 IMC            2022          5
##  4 IMC            2023          5
##  5 CODEX          2021          4
##  6 CODEX          2023          4
##  7 MIBI           2021          4
##  8 CosMx SMI      2023          3
##  9 MIBI           2022          3
## 10 MIBI           2024          3
## # ℹ 25 more rows
```

```
#my_palette <- paletteer_dynamic("cartography::green.pal", 17)
#my_palette <- colorRampPalette(c("darkorchid4", "darkslateblue", "deeppink3", "deepskyblue2", "blue2"))(17)

modality_counts |>
  ggplot(
  aes(x = year, y = Frequency, fill = modality_split)) +
  geom_bar(stat = "identity", color = "black") +
  scale_fill_manual(values = col_vector) +
theme_classic() +
  labs(
    title = "Single-cell Modality Prevalence Across Years (2017-2024)",
    subtitle = "Need to factcheck with MALDI, Visium, and scRNAseq (they're not single-cell)"
  )
```

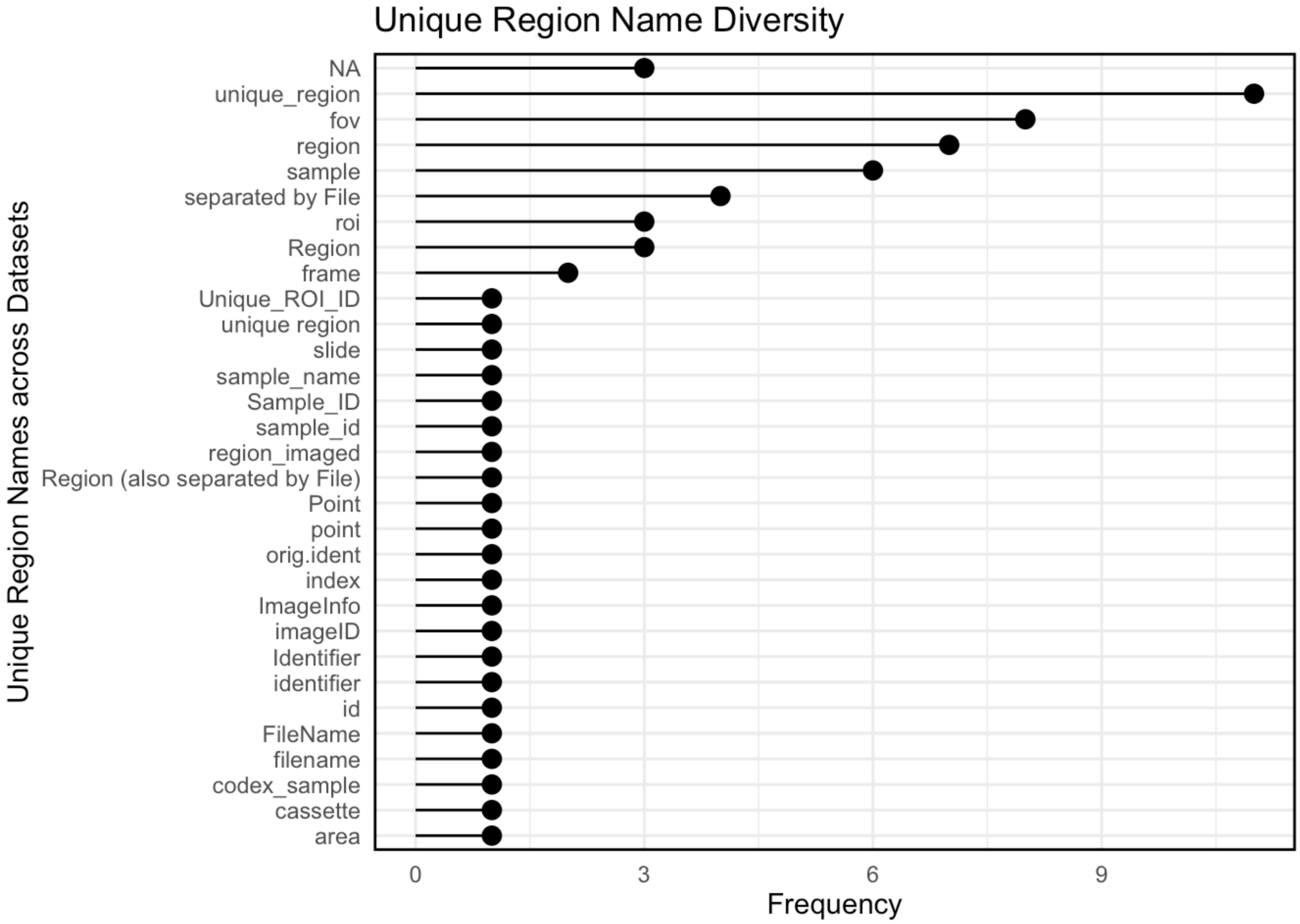


```
modality_counts |>
  ggplot(aes(x = reorder(modality_split, -Frequency), y=Frequency)) +
  geom_point(size=3) +
  geom_segment(aes(x= reorder(modality_split, -Frequency), xend=modality_split, y=0, yend = Frequency)) +
  labs(
    title = "Variation in Modalities across Spatial Datasets",
    x =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

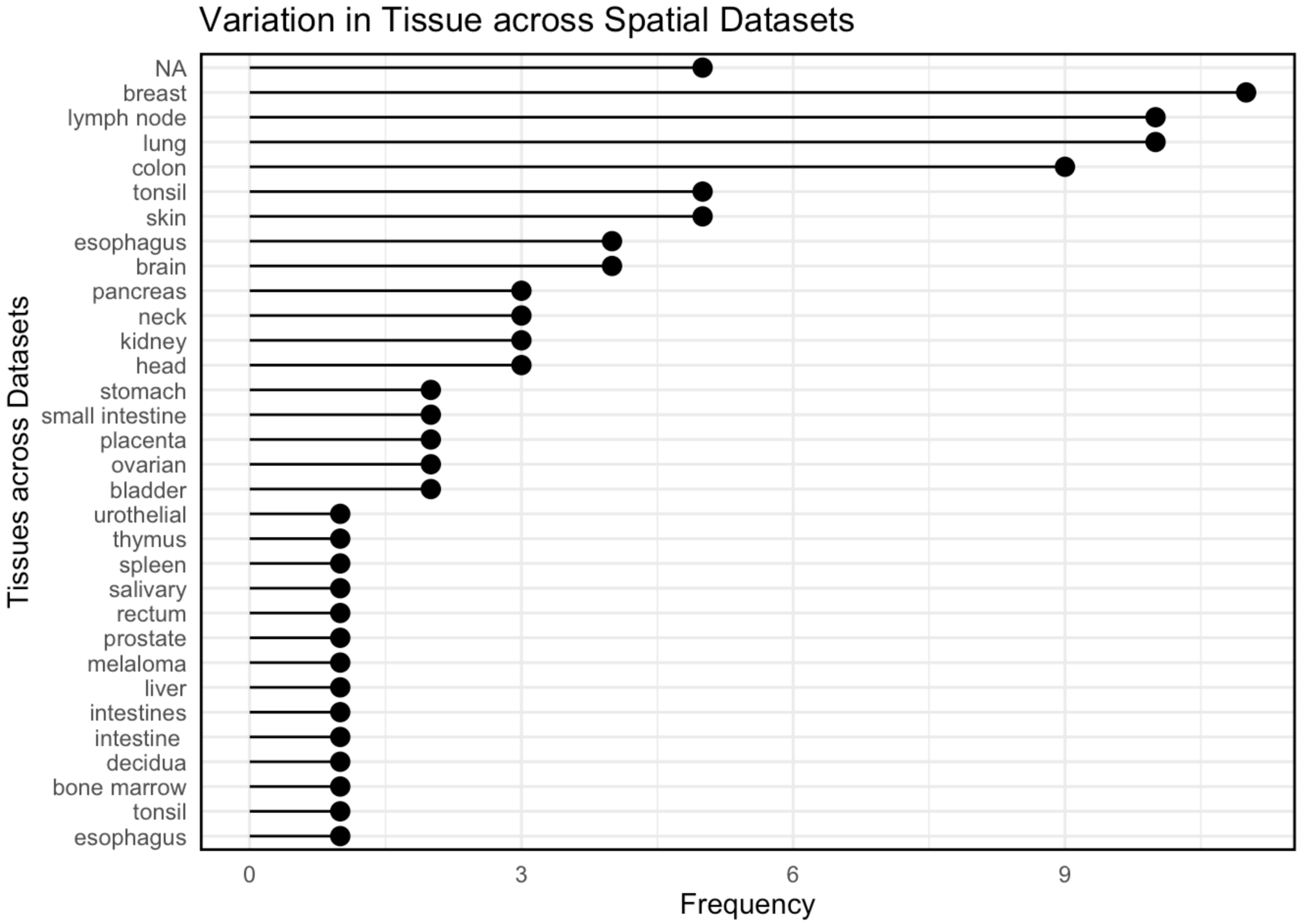


```
uniqueregion_freqs |>
  ggplot(aes(x = reorder(uniqueregion_col_name, Frequency), y=Frequency)) +
  geom_point(size=3) +
  geom_segment(aes(x= uniqueregion_col_name, xend=uniqueregion_col_name, y=0, yend = Fre
quency)) +
  labs(
    title = "Unique Region Name Diversity",
    x = "Unique Region Names 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()
```

```
## 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.
```



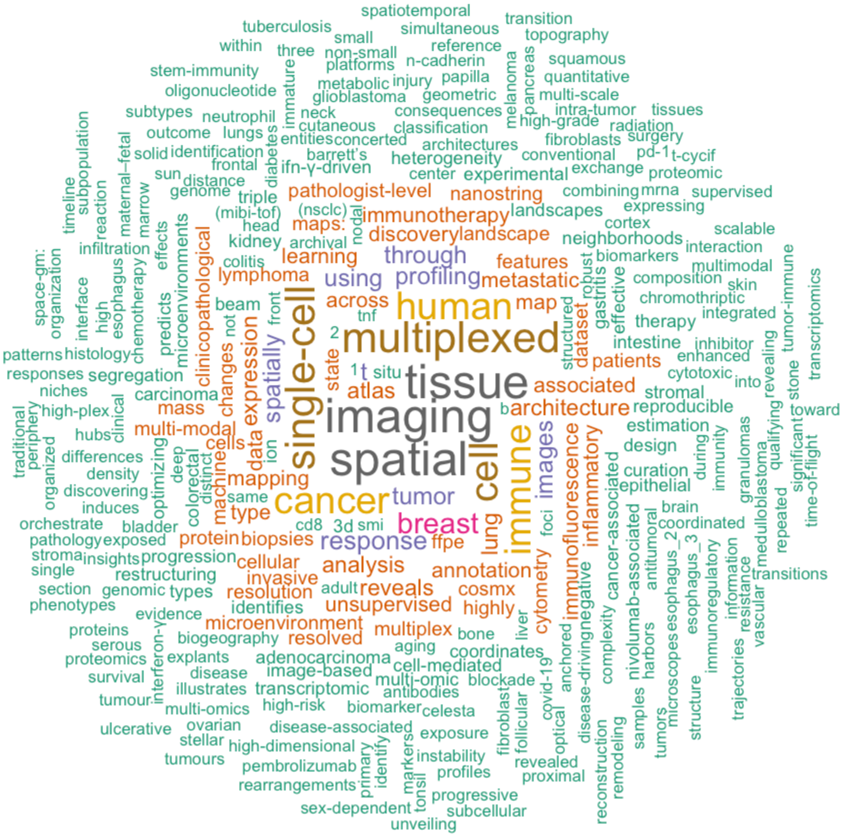
```
tissuedata_long <- tissue_freqs |>
  mutate(tissue_split = strsplit(tissue, ",")) |>
  unnest(tissue_split)

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()
```


Variation in Tissue across Spatial Datasets
Tissues across Datasets
NA
breast
lymph node
lung
colon
tonsil
skin
esophagus
brain
pancreas
neck
kidney
head
stomach
small intestine
placenta
ovarian
bladder
urothelial
thymus
spleen
salivary
rectum
prostate
melaloma
liver
intestines
intestine
decidua
bone marrow
tonsil
esophagus
0
3
6
9
Frequency

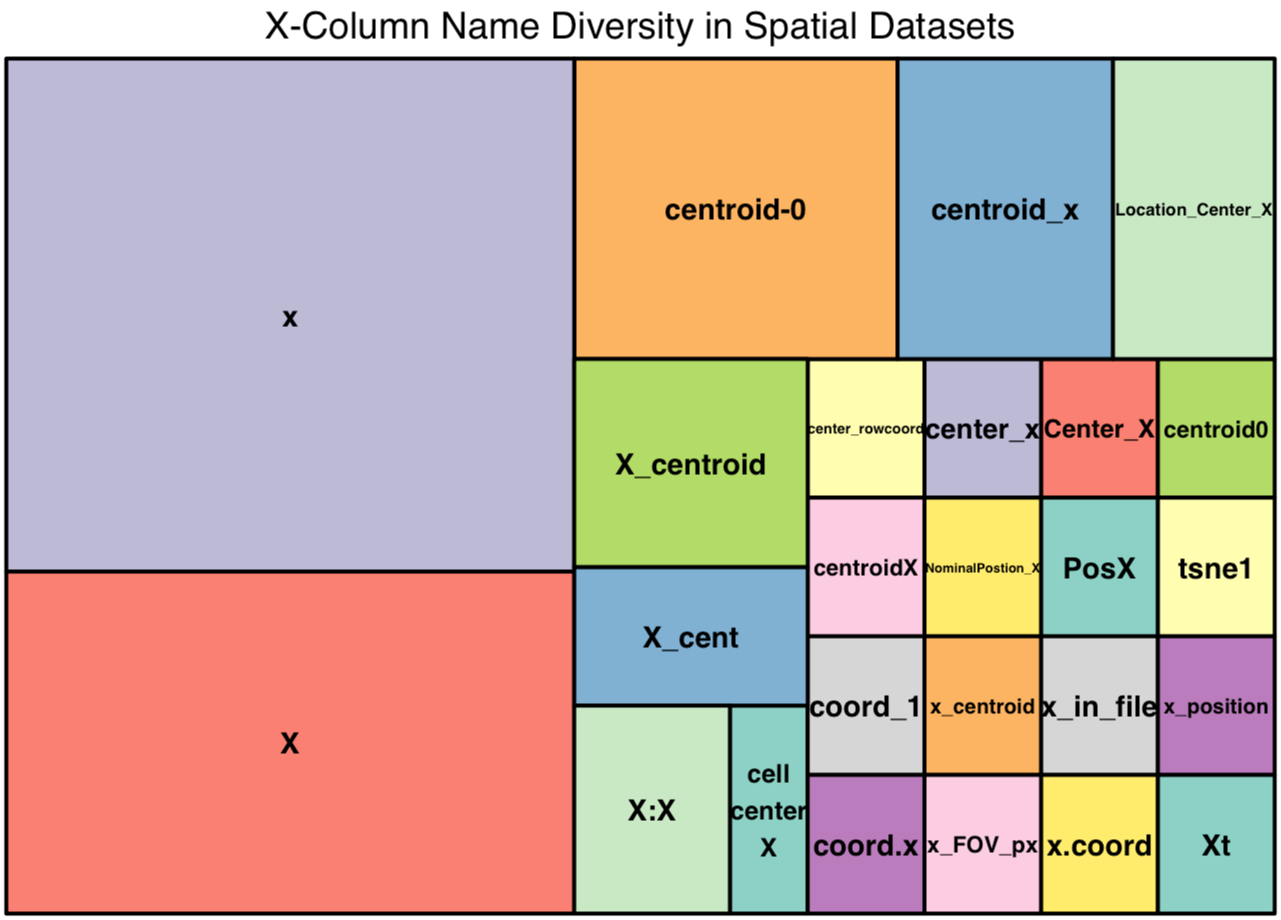
```
title_freqs <- data |>
  group_by(title, year, modality) |>
  summarize(Frequency = n(), .groups = "drop") |>
  arrange(desc(Frequency))

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: 315 × 2
##    title_split Frequency
##    <chr>           <int>
##  1 imaging            21
##  2 spatial            21
##  3 tissue             21
##  4 multiplexed        18
##  5 single-cell        17
##  6 cell               16
##  7 cancer             15
##  8 human              15
##  9 immune             14
## 10 breast              9
## # ℹ 305 more rows
```

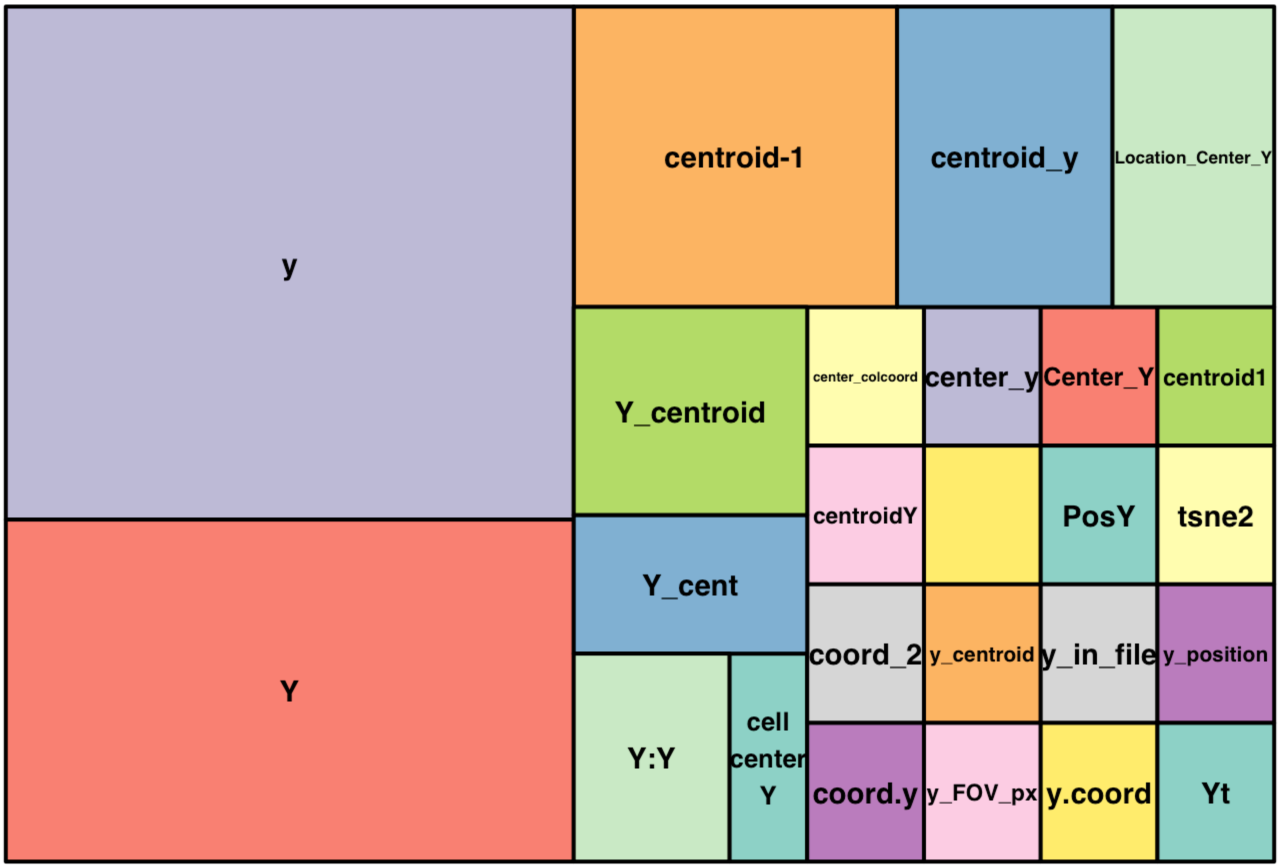
```
#why is quality so bad???
wordcloud(words = title_clean_counts$title_split, freq = title_clean_counts$Frequency,
          min.freq = 1, scale = c(1.5, 0.475), random.order = FALSE, rot.per = .25, colo
rs=brewer.pal(8,"Dark2"))
```



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

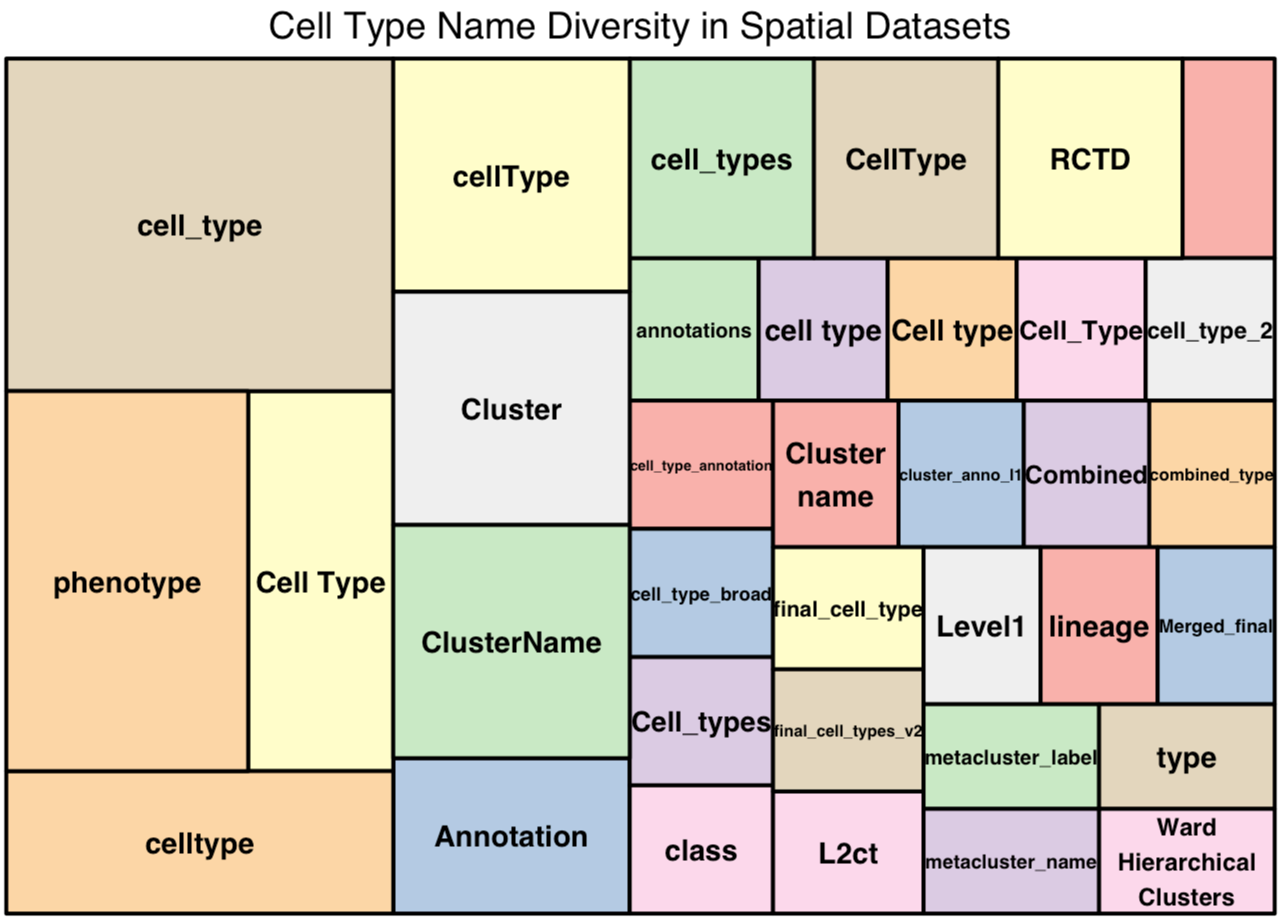


```
treemap(y_freqs,
        index = c("y_col_name"),
        vSize = "Frequency",
        vColor = "Frequency",
        title = "Y-Column Name Diversity in Spatial Datasets",
        palette = "Set3",
        draw = TRUE)
```

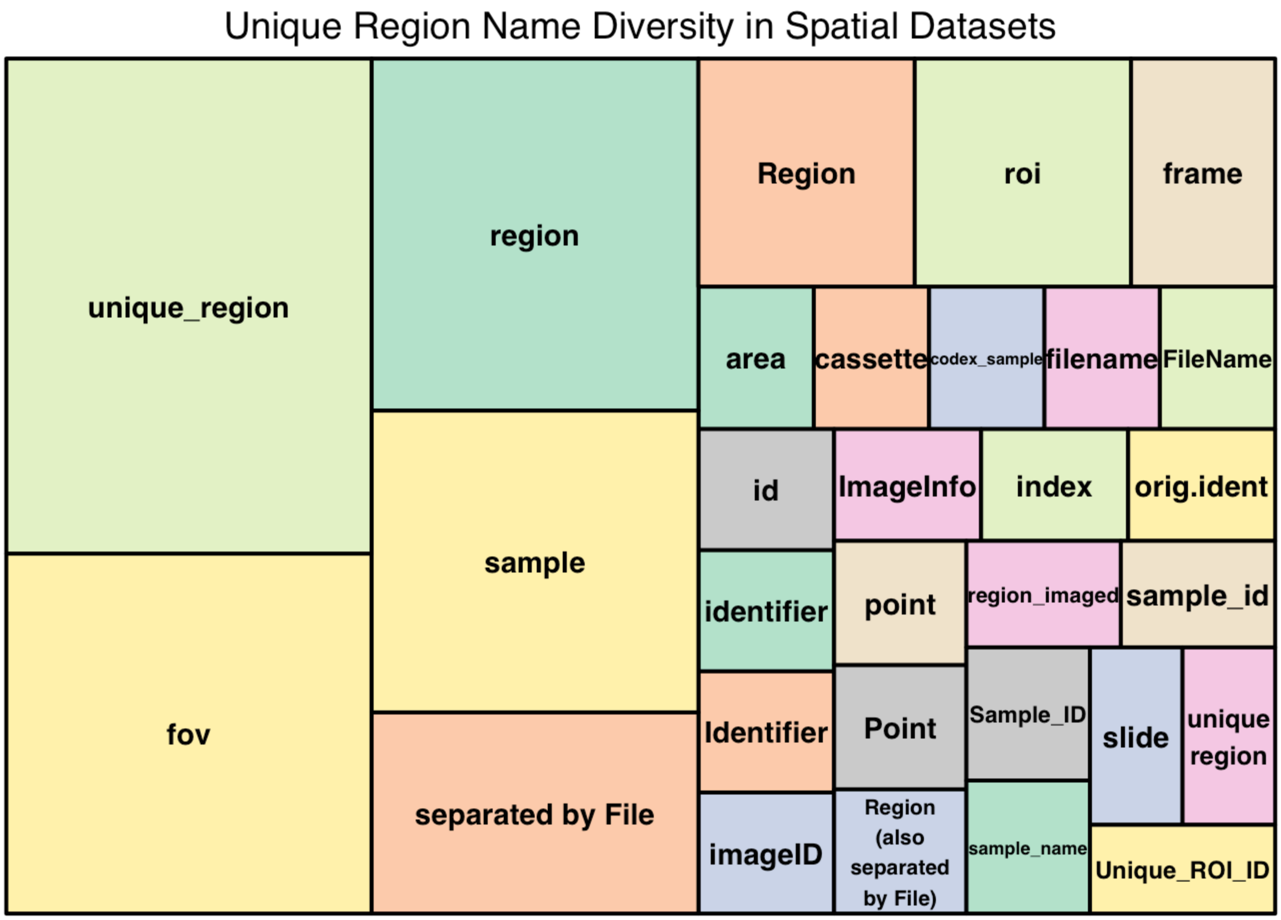


Y-Column Name Diversity in Spatial Datasets

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



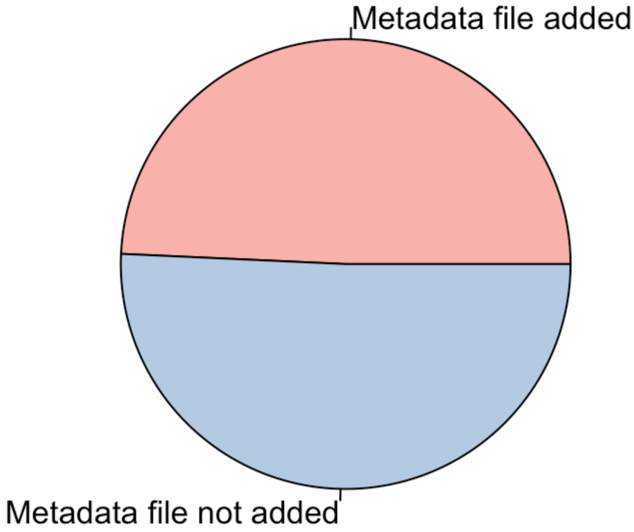
```
treemap(uniqueregion_freqs,
        index = c("uniqueregion_col_name"),
        vSize = "Frequency",
        vColor = "Frequency",
        title = "Unique Region Name Diversity in Spatial Datasets",
        palette = "Pastel2",
        draw = TRUE)
```



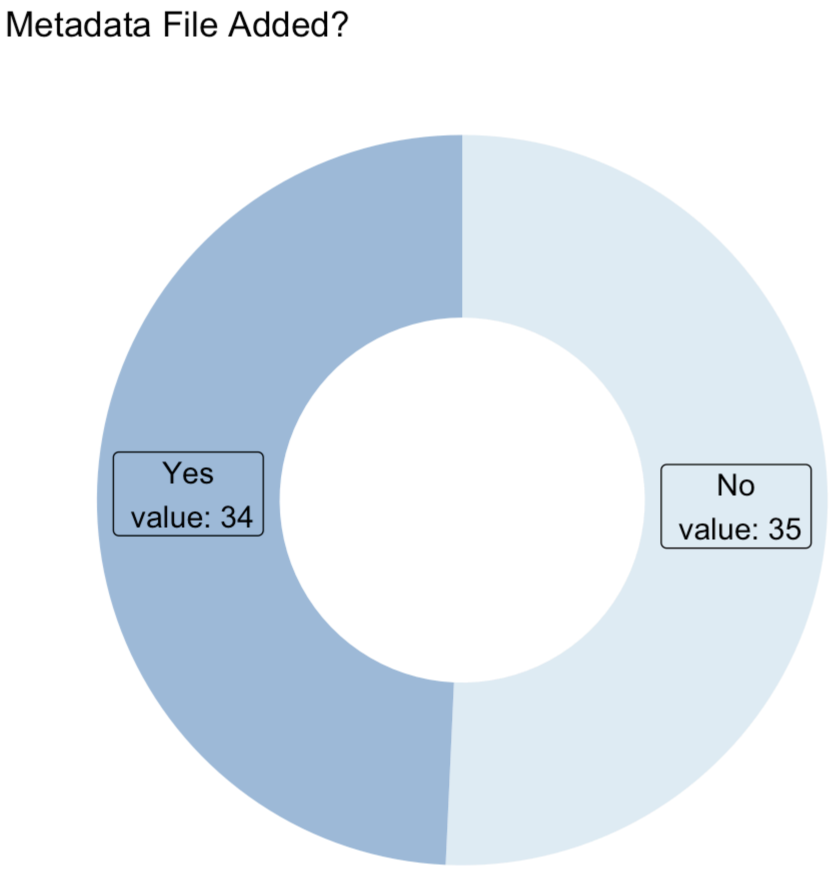
```
metaadded_freqs <- data |>
  group_by(metadata_file_added) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

metaadded_freqs <- metaadded_freqs |> mutate(
  metadata_file_added = if_else(is.na(metadata_file_added), "No", metadata_file_added)
  )

print(metaadded_freqs)
```

```
## # A tibble: 2 × 2
##   metadata_file_added Frequency
##   <chr>                   <int>
## 1 No                         35
## 2 Yes                        34
```

```
mypalette <- brewer.pal(3, "Pastel1")

Prop <- c(34, 35)
pie(Prop, labels = c("Metadata file added", "Metadata file not added"), col=mypalette)
```



```
metaadded_freqs$fraction = metaadded_freqs$Frequency / sum(metaadded_freqs$Frequency)
metaadded_freqs$ymax = cumsum(metaadded_freqs$fraction)
metaadded_freqs$ymin = c(0, head(metaadded_freqs$ymax, n=-1))
metaadded_freqs$labelPosition <- (metaadded_freqs$ymax + metaadded_freqs$ymin) / 2
metaadded_freqs$label <- paste0(metaadded_freqs$metadata_file_added, "\n value: ", metaa
dded_freqs$Frequency)

metaadded_freqs |>
  ggplot(
    aes(ymax = ymax, ymin = ymin, xmax = 4, xmin = 3, fill = metadata_file_added)
  ) +
  geom_rect() +
  geom_label(x=3.5, aes(y=labelPosition, label =label), size=4) +
  scale_fill_brewer(palette = "BuPu") +
 # scale_fill_paletteer_d("lisa::JohnSingerSargent_2") +
  coord_polar(theta = "y") +
  xlim(c(2,4)) +
  theme_void() +
  theme(legend.position = "none") +
  labs(
    title = "Metadata File Added?"
  )
```



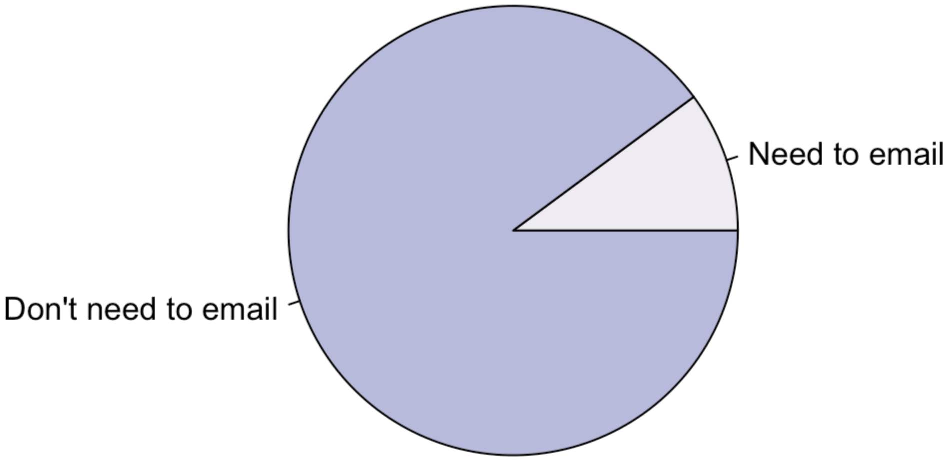
```
email_freqs <- data |>
  group_by(need_to_email) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

email_freqs <- email_freqs |>
  mutate(
    need_to_email = if_else(is.na(need_to_email), "No", need_to_email)
  )

print(email_freqs)
```

```
## # A tibble: 3 × 2
##   need_to_email Frequency
##   <chr>             <int>
## 1 No                   61
## 2 Yes                   7
## 3 No                    1
```

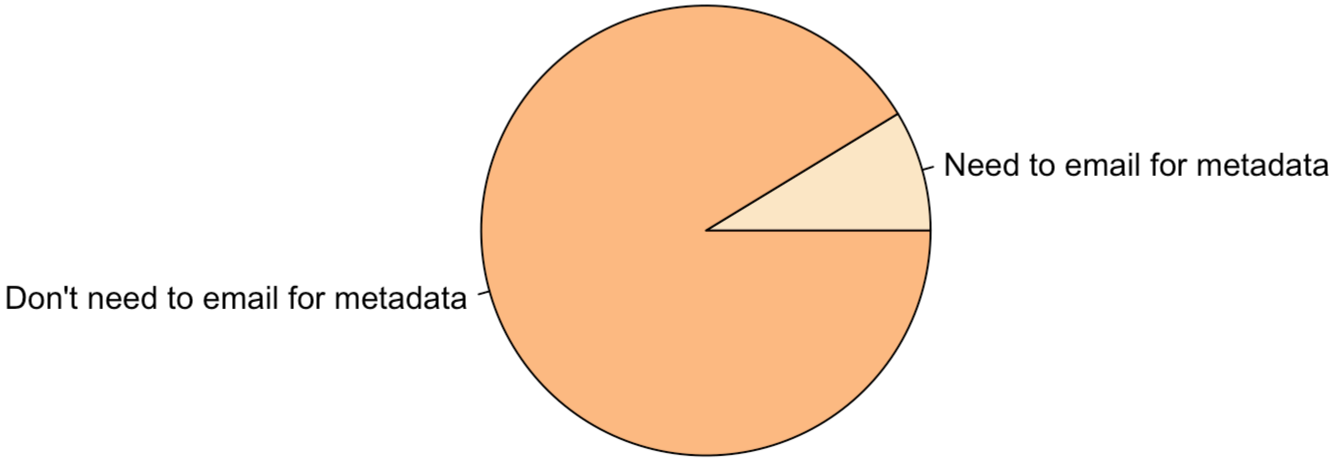
```
email_md_freqs <- data |>
  group_by(need_to_email_md) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

email_md_freqs <- email_md_freqs |>
  mutate(
    need_to_email_md = if_else(is.na(need_to_email_md), "No", need_to_email_md)
  )

print(email_md_freqs)
```

```
## # A tibble: 2 × 2
##   need_to_email_md Frequency
##   <chr>                <int>
## 1 No                      63
## 2 Yes                      6
```

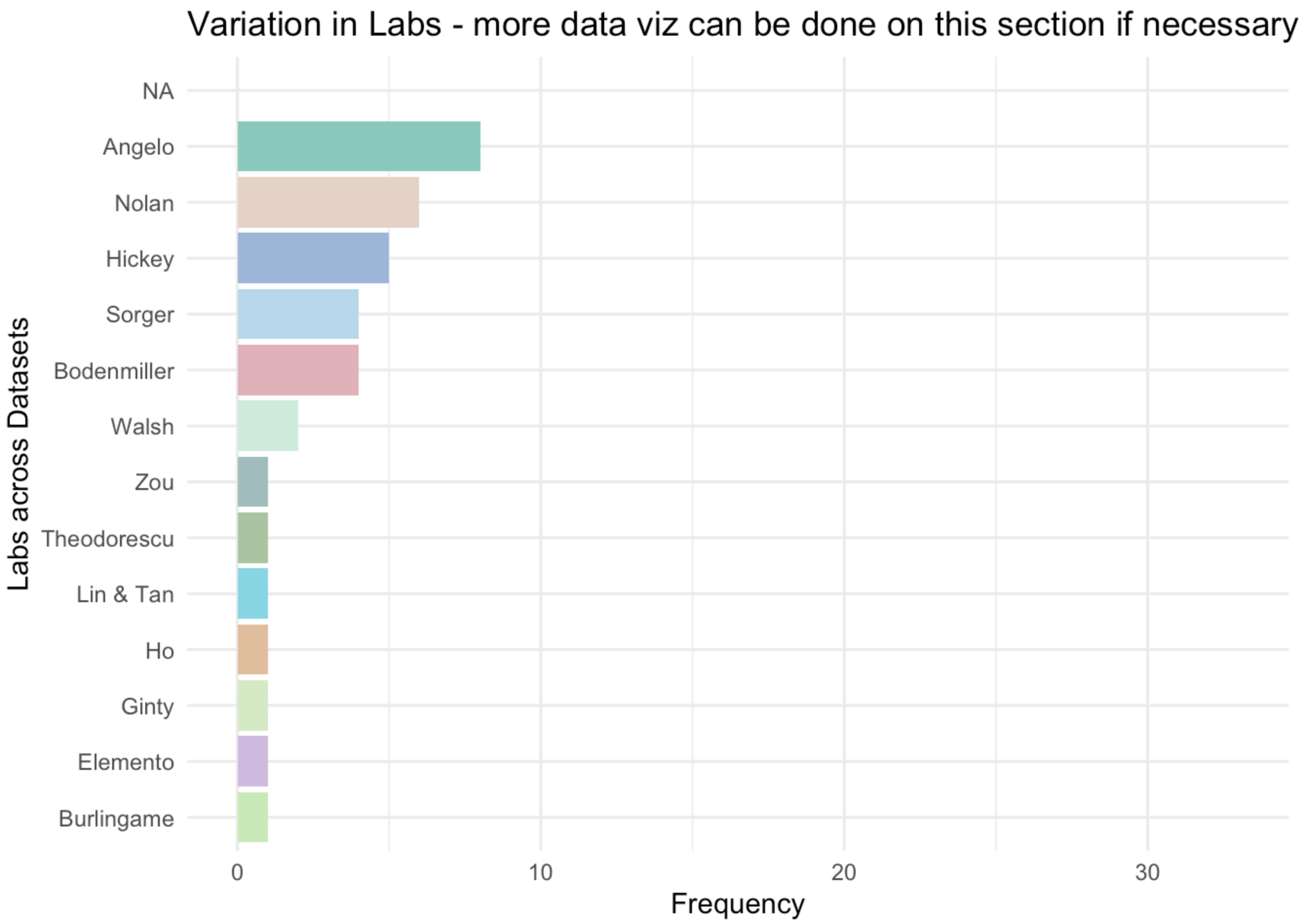
```
mypalette1 <- brewer.pal(3, "Purples")
mypalette2 <- brewer.pal(3, "OrRd")

needtoemail <- c(7, 62)
pie(needtoemail, labels = c("Need to email", "Don't need to email"), col=mypalette1)
```



```
needtoemailmd <- c(6, 63)
pie(needtoemailmd, labels = c("Need to email for metadata", "Don't need to email for met
adata"), col=mypalette2)
```



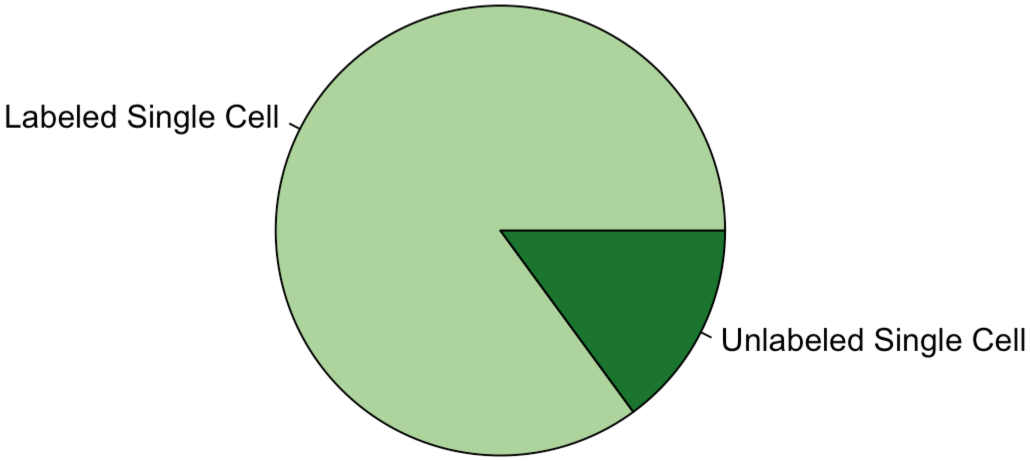
```
lab_freqs <- data |>
  group_by(lab) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

lab_freqs |>
  ggplot(aes(x = reorder(lab, Frequency), y=Frequency, fill = lab)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_paletteer_d("cartography::pastel.pal", dynamic = TRUE) +
  labs(
    title = "Variation in Labs - more data viz can be done on this section if necessar
y",
    x = "Labs across Datasets",
  ) +
  coord_flip()
```



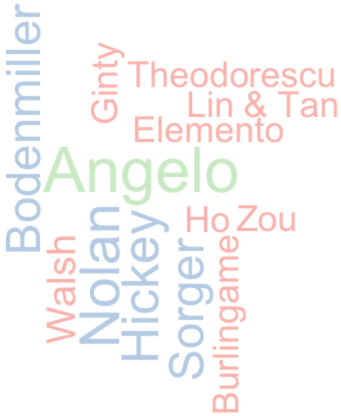
```
priority_freqs <- data |>
  group_by(priority) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

print(priority_freqs)
```

```
## # A tibble: 2 × 2
##   priority            Frequency
##   <chr>                   <int>
## 1 labeled_singlecell         59
## 2 unlabeled_singlecell       10
```

```
mypalette <- paletteer_dynamic("cartography::green.pal", 2)
priority_pie <- c(57, 10)
pie(priority_pie, labels = c("Labeled Single Cell", "Unlabeled Single Cell"), col=mypale
tte)
```



```
wordcloud(words = lab_freqs$lab, freq = lab_freqs$Frequency,
          min.freq = 1, scale = c(5, 1), random.order = FALSE, rot.per = .25, colors=brewer.pal(9,"Pastel1"))
```

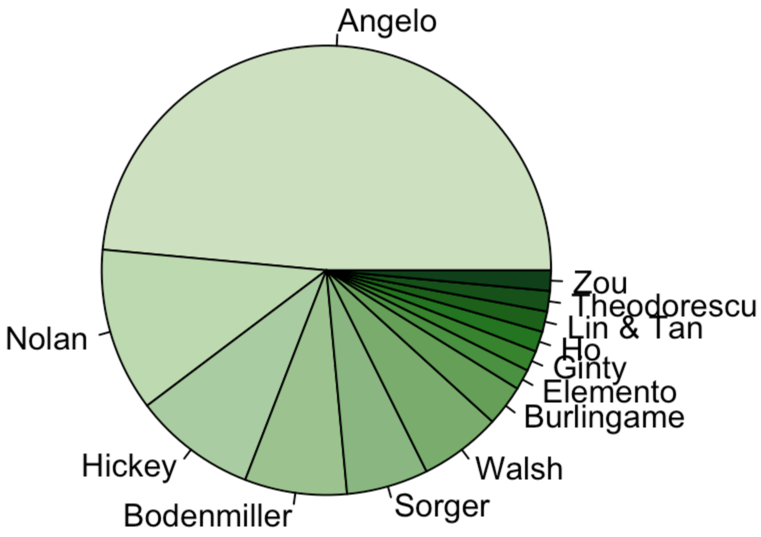


```
#
# data <- data |>
#   mutate(tissue_new = strsplit(tissue, ",")) |>
#   unnest(tissue_new)
#
# data <- data |>
#   mutate(modality_new = strsplit(modality, ",")) |>
#   unnest(modality_new)
#
# data <- data |>
#   mutate(disease_new = strsplit(disease, ",")) |>
#   unnest(disease_new)
```

```
print(lab_freqs)
```

```
## # A tibble: 14 × 2
##    lab         Frequency
##    <chr>           <int>
##  1 <NA>               33
##  2 Angelo              8
##  3 Nolan               6
##  4 Hickey              5
##  5 Bodenmiller         4
##  6 Sorger              4
##  7 Walsh               2
##  8 Burlingame          1
##  9 Elemento            1
## 10 Ginty               1
## 11 Ho                  1
## 12 Lin & Tan           1
## 13 Theodorescu         1
## 14 Zou                 1
```

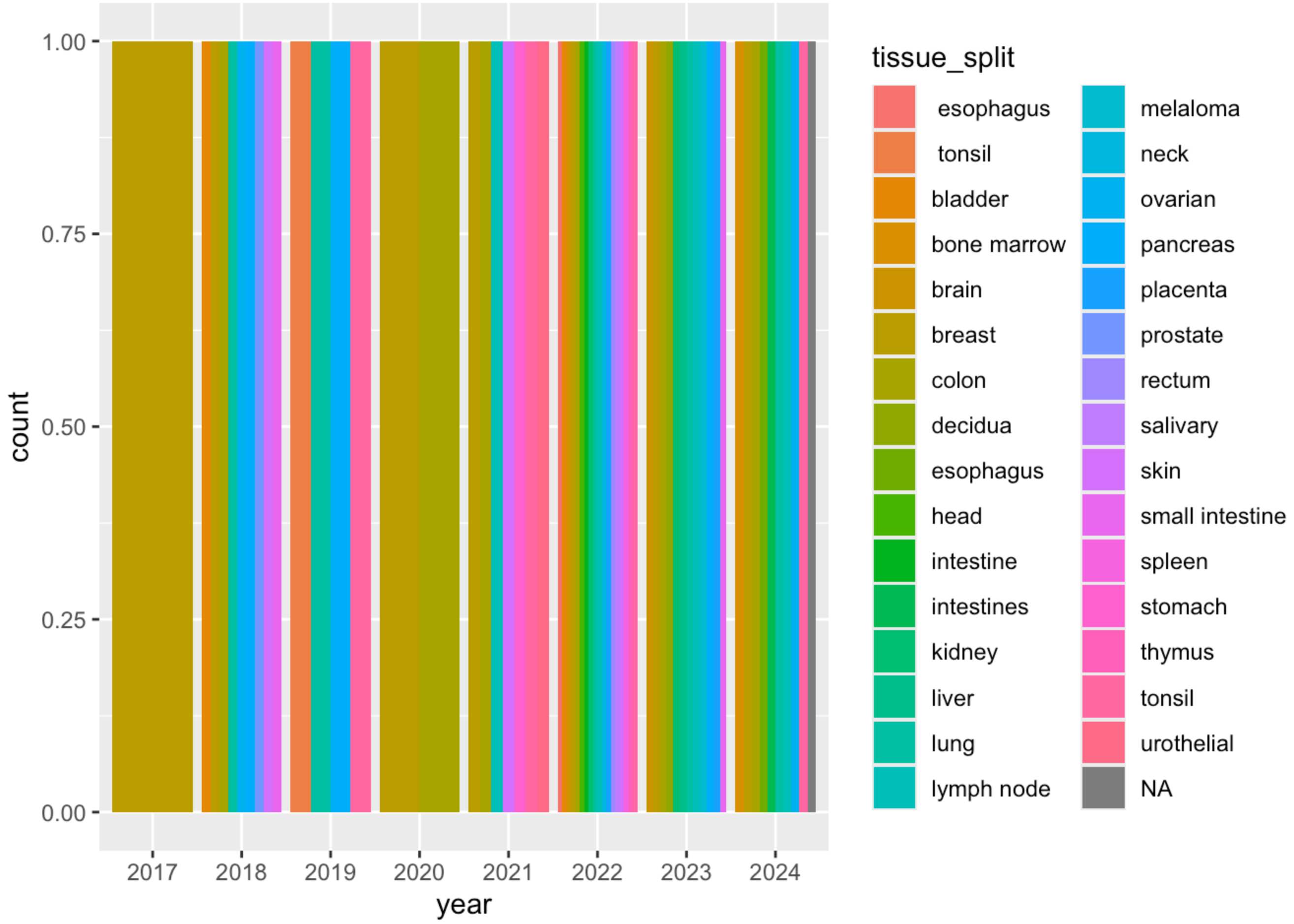
```
palette_lab <- paletteer_dynamic("cartography::green.pal", 13)

labpie <- c(33, 8, 6, 5, 4, 4, 2, 1, 1, 1, 1, 1, 1)
pie(labpie, labels = c("Angelo", "Nolan", "Hickey", "Bodenmiller", "Sorger", "Walsh", "B
urlingame", "Elemento", "Ginty", "Ho", "Lin & Tan", "Theodorescu", "Zou"), col=palette_l
ab)
```



```
tissue_year <- tissuedata_long |>
  group_by(tissue_split, year) |>
  summarize(Frequency = sum(Frequency)) |>
  arrange(desc(Frequency))
```

```
## `summarise()` has grouped output by 'tissue_split'. You can override using the
## `.groups` argument.
```

```
# tissue_freqs |>
#   ggplot(
#   aes(x = year, y = Frequency, fill = modality_split)) +
#   geom_bar(stat = "identity", color = "black") +
#   scale_fill_manual(values = col_vector) +
# theme_classic() +
#
tissue_year |>
  ggplot(
    aes(x = year, fill = tissue_split)) +
    geom_bar(position = "dodge")
```


count
year
tissue_split
esophagus
tonsil
bladder
bone marrow
brain
breast
colon
decidua
esophagus
head
intestine
intestines
kidney
liver
lung
lymph node
melaloma
neck
ovarian
pancreas
placenta
prostate
rectum
salivary
skin
small intestine
spleen
stomach
thymus
tonsil
urothelial
NA