Data Management and Analysis Workshop: Part 3

Data visualization

We will walk through some basic data cleaning, analysis, and visualization tasks using per pupil expenditures as a running example. The repository contains all the data and code we need, so you can just follow along! This notebook introduces data visualization.

Set up

```
# Install required packages.
req <- c("tidyverse")
new <- req[!(req %in% installed.packages()[, "Package"])]
if (length(new)) install.packages(new)

# Load libraries.
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(forcats)
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```r
# Identify inputs and outputs.
PWD <- getwd()
DTA <- file.path(PWD, "out", "data.rda")
OUT <- file.path(PWD, "out")
```

```r
# Load analysis file.
# This loads the dataset we created into the same namespace.
load(DTA)
head(dta)
```

```
   leaid             name  stname   tot       exp      ppe pct_asian pct_black
1 100005 Albertville City Alabama  5842  59207000 10134.71 0.4621705  4.193769
2 100006  Marshall County Alabama  5758  68866000 11960.06 0.5036471  1.128864
3 100007      Hoover City Alabama 13640 192421000 14107.11 7.1554252 23.453079
4 100008     Madison City Alabama 11804 184180000 15603.19 9.1070823 19.290071
5 100011       Leeds City Alabama  2097  24080000 11483.07 0.7629948 24.034335
6 100012        Boaz City Alabama  2431  28483000 11716.58 0.6581654  2.591526
   pct_hisp pct_white pct_other maj_group
1 52.601849  39.86648  2.875727      hisp
2 26.432789  70.47586  1.458840     white
3  8.541056  54.95601  5.894428     white
4  7.285666  58.42087  5.896306     white
5 15.069146  58.17835  1.955174     white
6 37.885644  55.49157  3.373097     white
```

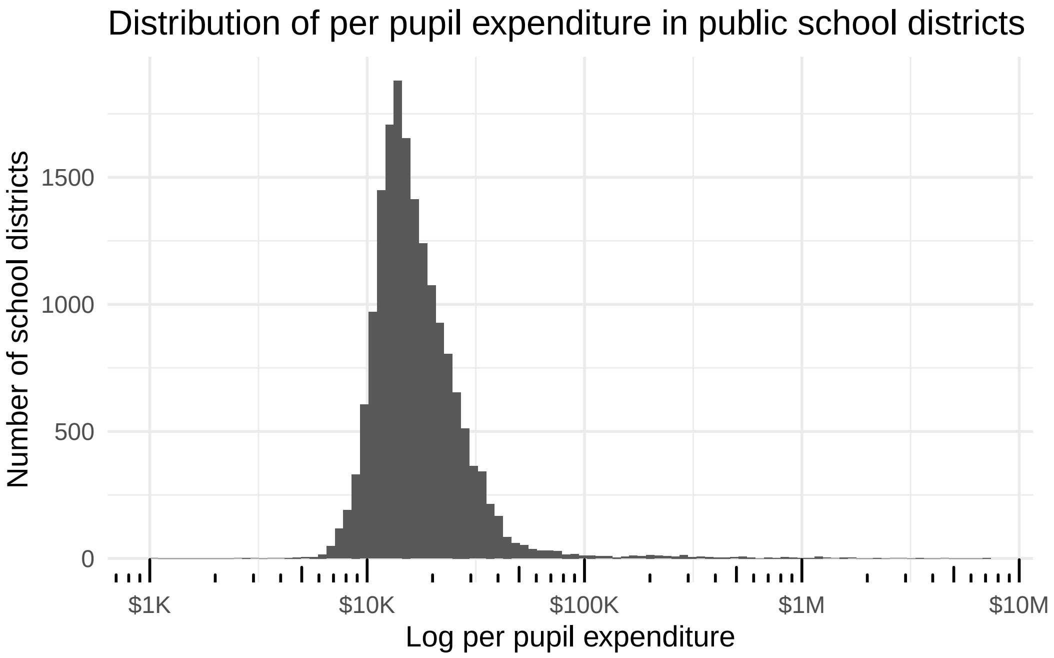
Plot distribution of per pupil expenditure

```r
# Select breaks for x-axis.
x_breaks <- c(1e3, 1e4, 1e5, 1e6, 1e7)
x_labels <- c("$1K", "$10K", "$100K", "$1M", "$10M")
```

```r
# Draw the figure.
ggplot(dta, aes(x = ppe)) +
  geom_histogram(bins = 100) +
  scale_x_log10(breaks = x_breaks, labels = x_labels) +
  annotation_logticks(side = "b") +
  theme_minimal() +
  labs(
    title = "Distribution of per pupil expenditure in public school districts",
    x = "Log per pupil expenditure",
    y = "Number of school districts"
  )
```

```
Warning in scale_x_log10(breaks = x_breaks, labels = x_labels): log-10
transformation introduced infinite values.
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_bin()`).
```



```r
# Save the figure.
ggsave(file.path(OUT, "ppe-hist.png"))
```

```
Saving 5.5 x 3.5 in image
```

```
Warning in scale_x_log10(breaks = x_breaks, labels = x_labels): log-10 transformation introdu
Removed 5 rows containing non-finite outside the scale range (`stat_bin()`).
```

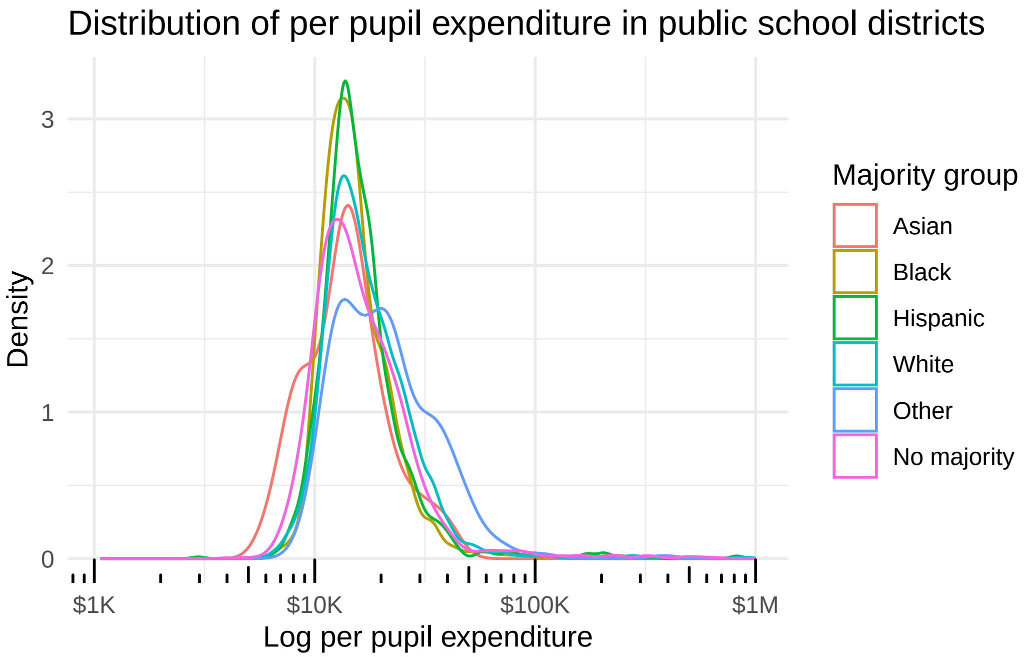
```r
# Map new to old race category labels.
race_recode <- c(
  "Asian" = "asian",
  "Black" = "black",
  "Hispanic" = "hisp",
  "White" = "white",
  "Other" = "other",
  "No majority" = "none"
)

# Relabel and reorder race categories.
dta <- mutate(dta,
  maj_group2 = fct_recode(as.factor(maj_group), !!!race_recode),
  maj_group2 = fct_relevel(maj_group2, names(race_recode))
)
```

```r
# Draw the figure.
ggplot(dta, aes(x = ppe, color = maj_group2)) +
  geom_density() +
  scale_x_log10(breaks = x_breaks, labels = x_labels, limits = c(NA, 1e6)) +
  annotation_logticks(side = "b") +
  theme_minimal() +
  labs(
    title = "Distribution of per pupil expenditure in public school districts",
    x = "Log per pupil expenditure",
    y = "Density",
    color = "Majority group"
  )
```

```
Warning in scale_x_log10(breaks = x_breaks, labels = x_labels, limits = c(NA, :
log-10 transformation introduced infinite values.
```

```
Warning: Removed 35 rows containing non-finite outside the scale range
(`stat_density()`).
```



```
# Save the figure.
ggsave(file.path(OUT, "ppe-dens.png"))
```

```
Saving 5.5 x 3.5 in image
```

```
Warning in scale_x_log10(breaks = x_breaks, labels = x_labels, limits = c(NA, : log-10 trans
Removed 35 rows containing non-finite outside the scale range
(`stat_density()`).
```

Plot membership by revenue

```
# Select breaks for PPE.
quantile(dta$ppe, c(0, 0.05, 0.25, 0.5, 0.75, 0.95))
```

```
       0%         5%        25%        50%        75%        95% 
    0.000  9510.663 12498.941 15509.335 21105.393 36613.729 
```

```
breaks <- c(0, seq(10000, 20000, 2000), 30000, 40000)
ppe_breaks <- seq(from = 0, to = 40000, by = 10000)
ppe_labels <- c("$0", "$10K", "$20K", "$30K", "$40K")
```

```r
# Select breaks for x-axis.
x_breaks <- c(1, 1e1, 1e2, 1e3, 1e4, 1e5)
x_labels <- c("0", "10", "100", "1K", "10K", "100K")
```

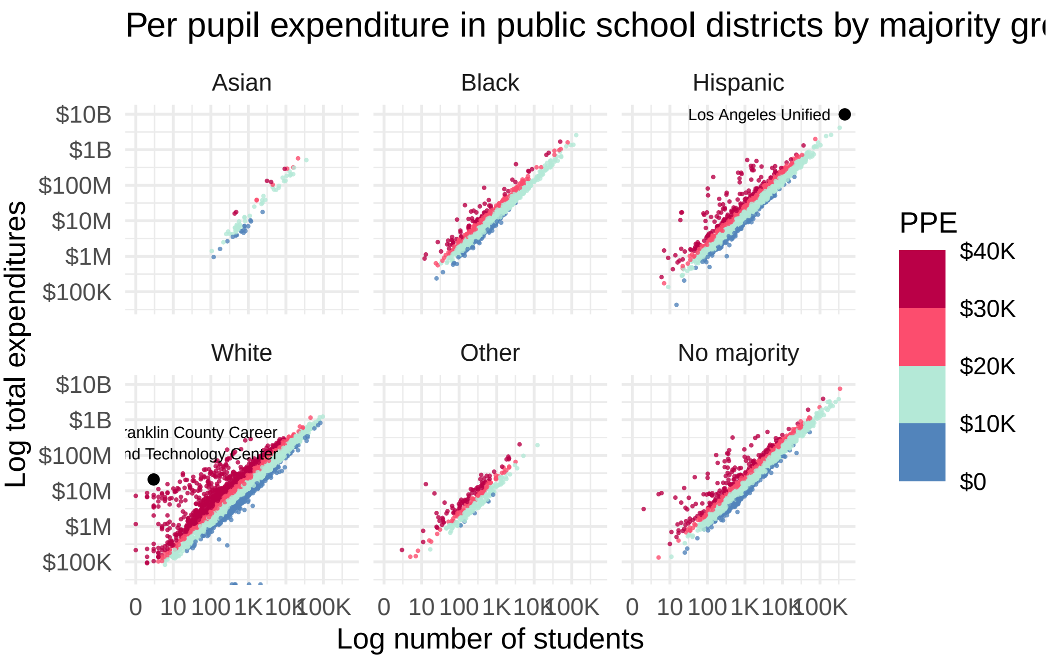
```r
# Select breaks for y-axis.
y_breaks = c(1e5, 1e6, 1e7, 1e8, 1e9, 1e10)
y_labels = c("$100K", "$1M", "$10M", "$100M", "$1B", "$10B")
```

```r
# Set annotations for salient points in the data.
picks <- data.frame(
  leaid = c(4280230, 622710),
  label = c("Franklin County Career\nAnd Technology Center", "Los Angeles Unified"),
  hjust = c(0.25, 1.1),
  vjust = c(-0.6, 0.5)
)

# Merge onto coordinates.
picks <- left_join(picks, dta, by = "leaid", relationship = "one-to-one")
```

```r
# Draw the figure.
ggplot(dta, aes(x = tot, y = exp, color = ppe)) +
  geom_point(size = 0.8, alpha = 0.8, stroke = NA) +
  geom_point(data = picks, color = "black") +
  geom_text(
    data = picks,
    mapping = aes(label = label, hjust = hjust, vjust = vjust),
    size = 2,
    color = "black"
  ) +
  scale_x_log10(breaks = x_breaks, labels = x_labels) +
  scale_y_log10(breaks = y_breaks, labels = y_labels) +
  scale_color_stepsn(
    breaks = ppe_breaks,
    labels = ppe_labels,
    limits = c(min(ppe_breaks), max(ppe_breaks)),
    colors = c("#00429d", "#96ffea", "#ff005e", "#93003a")
  ) +
  facet_wrap(~ maj_group2) +
  theme_minimal() +
  labs(
    title = "Per pupil expenditure in public school districts by majority group",
    x = "Log number of students",
    y = "Log total expenditures",
    color = "PPE"
  )
```

```
Warning in scale_y_log10(breaks = y_breaks, labels = y_labels): log-10
transformation introduced infinite values.
```



```
# Save the figure.
ggsave(file.path(OUT, "pop-exp-ppe.png"))
```

```
Saving 5.5 x 3.5 in image
```

```
Warning in scale_y_log10(breaks = y_breaks, labels = y_labels): log-10
transformation introduced infinite values.
```