

Data Visualization Showcase

Sierra Stryker

Introduction

This file represents data management and visualization skills. I will demonstrate a small selection of data cleaning and then showcase numerous types of graphs and explain their usage for understanding the data.

This file will use the Discrimination in Developmental Disability Support dataset (`DDS.discr`) from the OpenIntro repository. This is a publicly available repository of both real and fabricated datasets primarily for teaching advanced statistics.

The `DDS.discr` dataset contains 1000 California consumers who received financial support related to a developmental disability. Each recipient has data for their age, gender, ethnicity, and the support amount spent by the state of California annually (expenditures).

Libraries and Dataset

I will use 3 libraries in this R file to give myself access to more functions: - tidyverse will be used for general data management - ggplot2 will be used for generating graphs - scales is an addendum to ggplot2 that will be used for additional graphing capabilities

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.3      v readr      2.1.4
v forcats    1.0.0      v stringr    1.5.0
v ggplot2    3.4.3      v tibble     3.2.1
v lubridate  1.9.2      v tidyr      1.3.0
v purrr      1.0.2
```

```
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
library(ggplot2)
library(scales)
```

Attaching package: 'scales'

The following object is masked from 'package:purrr':

discard

The following object is masked from 'package:readr':

col_factor

```
dds <- read.csv("https://www.openintro.org/data/csv/dds.discr.csv")
```

Data Management: Demographics

In this section, categories for variables will be combined and changed in order to make the data easier to work with and manipulate. I have made notes about where and why data was changed and the appropriateness of these data management strategies.

Additionally, Ethnicity as a key variable in this dataset will be visualized and better understood.

Visualize Ethnicity

To better understand the demographics of the funding recipients, a table and a donut graph were generated to visualize percentages.

```
# Table of Ethnicity Percentages
donutgraph <- dds %>%
  count(ethnicity) %>%
  mutate(percent = n/sum(n)*100)
donutgraph
```

	ethnicity	n	percent
1	American Indian	4	0.4
2	Asian	129	12.9
3	Black	59	5.9
4	Hispanic	376	37.6
5	Multi Race	26	2.6
6	Native Hawaiian	3	0.3
7	Other	2	0.2
8	White not Hispanic	401	40.1

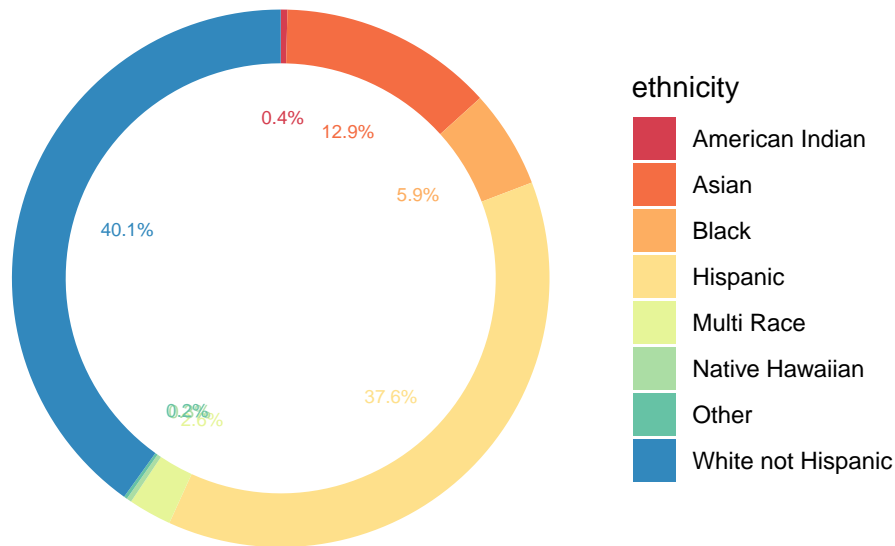
```
# Donut Graph of Ethnicity Percentages
```

```
## Creating Labels & Formatting
```

```
donutgraph$ymax <- cumsum(donutgraph$percent)
donutgraph$ymin <- c(0, head(donutgraph$ymax, n=-1))
donutgraph$labelPosition <- (donutgraph$ymax + donutgraph$ymin) / 2
donutgraph$label <- paste0(donutgraph$percent, "%")
```

```
## Generating Graph
```

```
ggplot(data=donutgraph, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=ethnicity)) +
  geom_rect() +
  geom_text(x=2, aes(y=labelPosition, label=label, color=ethnicity), size=2.5) +
  scale_fill_brewer(palette="Spectral") +
  scale_color_brewer(palette="Spectral") +
  coord_polar(theta="y") +
  xlim(c(-1, 4)) +
  theme_void()
```



There were multiple ethnicity groups that had fewer than 30 total recipients. When these categories are collapsed into an overarching category of “Other”, their combined value was less than 5% of the sample.

For actual data analyses, these groups would ideally remain separate. However, purely for the purposes of this demonstration, the understandability of graphs is improved by combining the groups.

```
# Combine American Indian, Multi Race, Native Hawaiian, and Other into one category; also
dds <- dds %>%
  mutate(ethnicity_simple =
    case_when(
      ethnicity == "American Indian" ~ "Other",
      ethnicity == "Multi Race" ~ "Other",
      ethnicity == "Native Hawaiian" ~ "Other",
      ethnicity == "Other" ~ "Other",
      ethnicity == "White not Hispanic" ~ "White",
      TRUE ~ as.character(ethnicity)
    )
  )
```

Using the simplified groups, the data is more readable and the percentages are easier to understand.

```
# Table of Simplified Ethnicity Percentages
donutgraph_simple <- dds %>%
```

```

count(ethnicity_simple) %>%
mutate(percent = n/sum(n)*100)
donutgraph_simple

```

	ethnicity_simple	n	percent
1	Asian	129	12.9
2	Black	59	5.9
3	Hispanic	376	37.6
4	Other	35	3.5
5	White	401	40.1

```

# Donut Graph of Simplified Ethnicity Percentages

```

```

## Creating Labels & Formatting

```

```

donutgraph_simple$ymax <- cumsum(donutgraph_simple$percent)
donutgraph_simple$ymin <- c(0, head(donutgraph_simple$ymax, n=-1))
donutgraph_simple$labelPosition <- (donutgraph_simple$ymax + donutgraph_simple$ymin) / 2
donutgraph_simple$label <- paste0(donutgraph_simple$ethnicity, "\n", donutgraph_simple$per

```

```

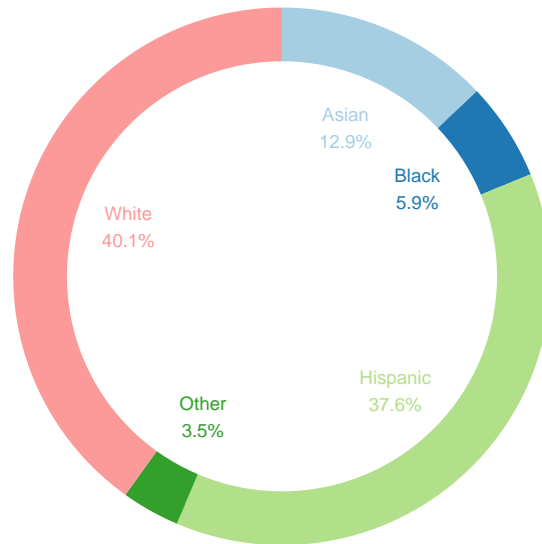
## Generating Graph

```

```

ggplot(data=donutgraph_simple, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=ethnicity_si
geom_rect() +
geom_text(x=2, aes(y=labelPosition, label=label, color=ethnicity_simple), size=2.5) +
scale_fill_brewer(palette="Paired") +
scale_color_brewer(palette="Paired") +
coord_polar(theta="y") +
xlim(c(-1, 4)) +
theme_void() +
theme(legend.position = "none")

```



Another common data transformation is to generate two overarching categories for ethnicity: “White” and “Non-White”. This is known as dichotomizing a variable and can be very useful in certain types of data analysis.

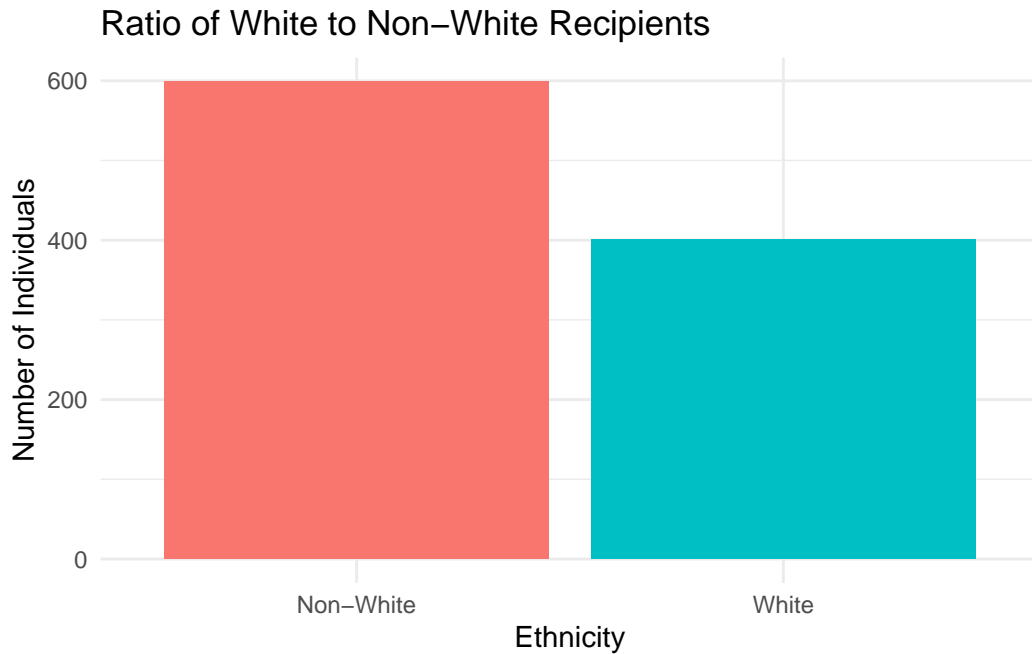
```
# Dichotomize Ethnicity
dds <- dds %>%
  mutate(ethnicity_dichotomous =
    case_when(
      ethnicity != "White not Hispanic" ~ "Non-White",
      ethnicity == "White not Hispanic" ~ "White"
    )
  )
```

A table and a bar graph were generated to represent the percentages and number of recipients for dichotomized ethnicity.

```
# Table of Dichotomized Ethnicity Percentages
dds %>%
  count(ethnicity_dichotomous) %>%
  mutate(percent = n/sum(n)*100)
```

	ethnicity_dichotomous	n	percent
1	Non-White	599	59.9
2	White	401	40.1

```
# Bar Graph of Dichotomized Ethnicity Percentages
ggplot(data=dds, aes(x=ethnicity_dichotomous, fill=ethnicity_dichotomous)) +
  geom_bar() +
  labs(title="Ratio of White to Non-White Recipients", x = "Ethnicity", y="Number of Individuals") +
  theme_minimal() +
  theme(legend.position = "none")
```



Understanding the Data

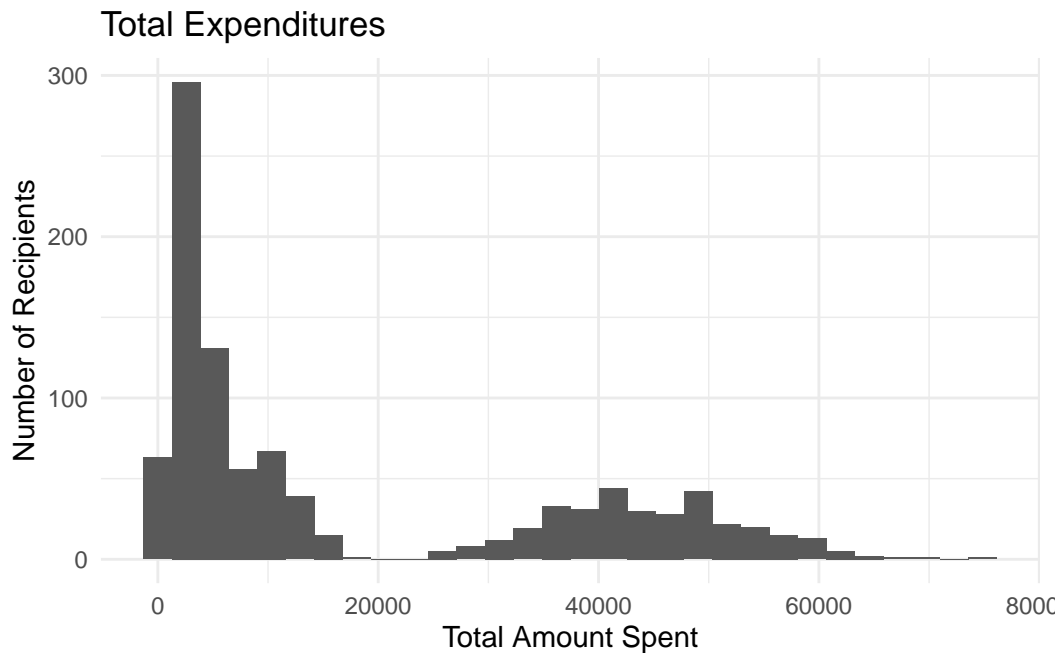
Expenditures

To begin to understand the award amounts, this first graph is a histogram of the overall expenditures allotted to applicants. This presents information about the most common reward amounts for recipients and the range of values.

Most applicants received between \$0-10,000 in rewards.

Another less common range for applicants was an award amount between \$30,000-60,000.

```
ggplot(data=dds, aes(x=expenditures)) +
  geom_histogram(bins=30) +
  labs(title="Total Expenditures", x="Total Amount Spent", y="Number of Recipients") +
  theme_minimal()
```



Now that there is a baseline understanding of the general range of common expenditures, the next step is to investigate demographic variables that may affect expenditure amounts.

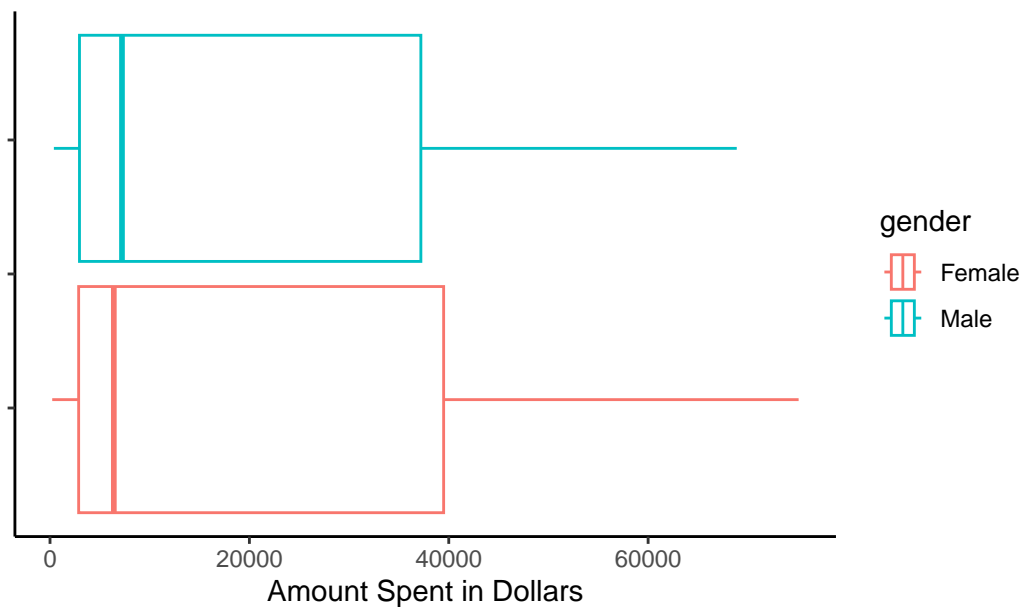
Effects of Gender

To better understand the role of gender on award amounts, gender and expenditure will be visualized in a box and whisker plot.

Box and whisker plots represent information about the minimum, maximum, median, and first and third quartiles for numeric data.

```
# Box and Whisker Plot of Expenditures by Gender
ggplot(data=dds, aes(x=expenditures, color=gender)) +
  geom_boxplot() +
  labs(title="Expenditures by Gender", x="Amount Spent in Dollars") +
  theme_classic() +
  theme(axis.text.y=element_blank())
```


Expenditures by Gender



Key Takeaways for Gender:

From this box and whisker plot, it can be seen that female recipients had a slightly higher maximum award value, but a slightly lower median value.

This indicates that female and male award amounts were fairly equal, but the average female recipient was awarded less in expenditures than the average male recipient. However, there is also some evidence that the highest rewards tended to go to women compared to men.

Effects of Ethnicity

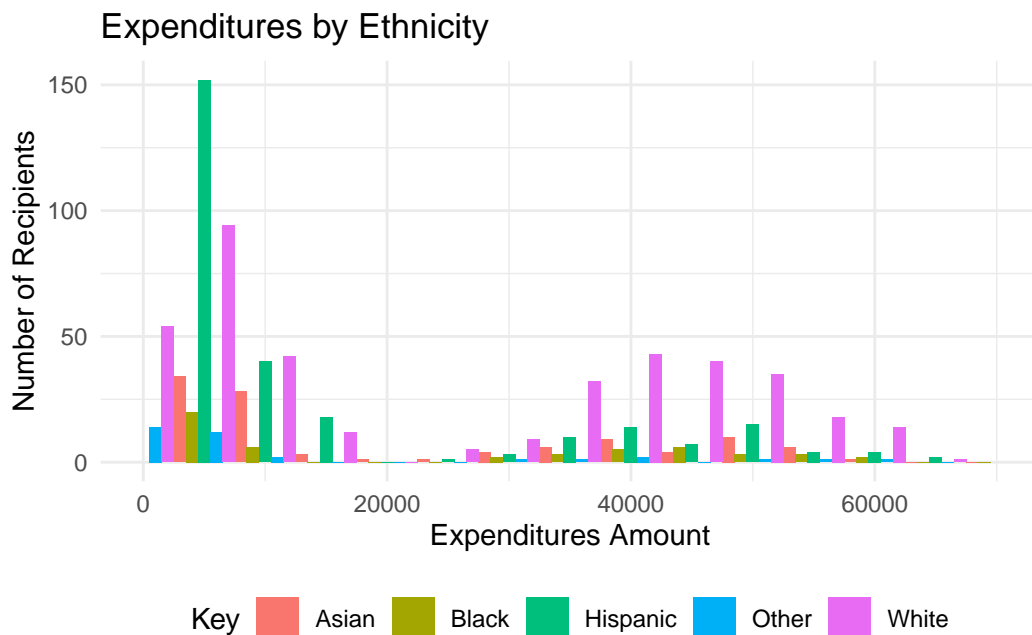
This dataset highlights ethnicity-based discrimination in state-awarded developmental disability support. Therefore, understanding the differences between ethnic groups is imperative in visualizing this dataset.

Simplified Ethnicity

A histogram will be used to examine the effects of ethnicity on expenditures. This is ideal for showing which ethnic groups were more likely to be awarded certain amounts.

To improve readability, data were slightly condensed (using bins) and a maximum amount was set at \$70,000 because greater values were rare. Furthermore, this graph used the simplified ethnicity category created earlier.

```
# Histogram of Expenditures by Ethnicity
ggplot(data=dds, aes(x=expenditures, fill=ethnicity_simple)) +
  geom_histogram(bins = 15, position = "dodge") +
  labs(title="Expenditures by Ethnicity", x="Expenditures Amount", y="Number of Recipients") +
  xlim(0, 70000) +
  theme_minimal() +
  theme(legend.position="bottom")
```



Key Takeaways for Simplified Ethnicity:

Across all award amounts, White recipients were the largest group.

All ethnicities were most likely to receive amounts under \$20,000.

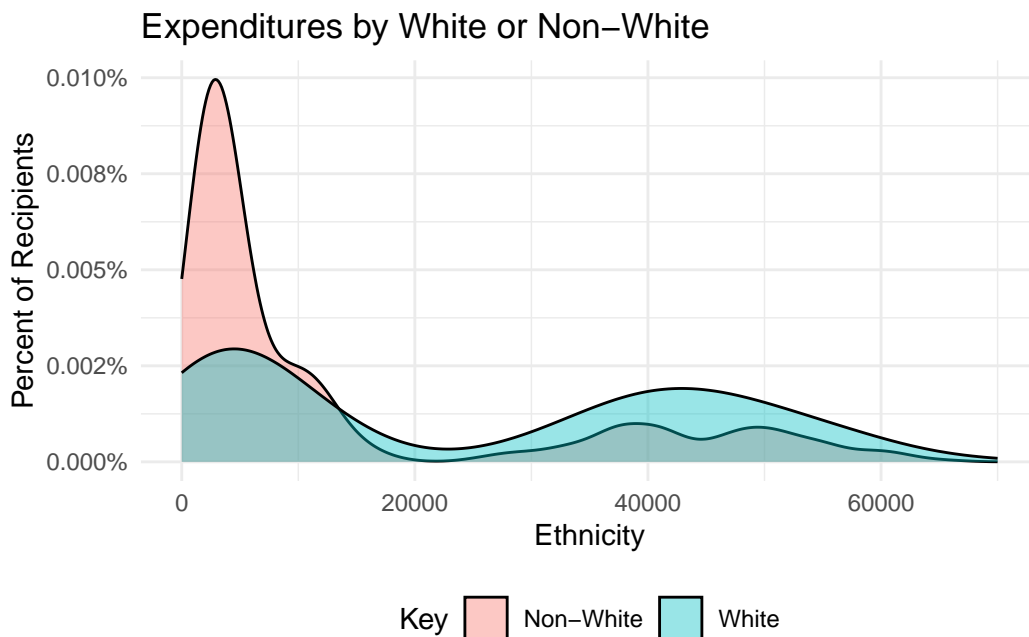
However, there were many more White recipients awarded over \$30,000 compared to all other ethnicities. To determine if this is due to the number of White recipients or a true statistical difference, an ANOVA should be performed.

Dichotomized Ethnicity

As noted earlier, there were more Non-White recipients than White recipients. Using a density graph, the difference in expenditures will be assessed for White compared to Non-White recipients.

A density graph is essentially a smoothed-out representation of the information from a histogram. Transparency can be utilized with density graphs to make visual comparisons between groups.

```
# Density Graph of Expenditures by Dichotomous Ethnicity
p1 <- ggplot(data=dds, aes(x=expenditures, fill=ethnicity_dichotomous)) +
  geom_density(alpha=0.4) +
  labs(title="Expenditures by White or Non-White", x="Ethnicity", y="Percent of Recipients") +
  xlim(0, 70000) +
  theme_minimal() +
  theme(legend.position="bottom")
p1 + scale_y_continuous(labels = percent_format(accuracy = .001))
```



Key Takeaways for Dichotomized Ethnicity:

This density graph clearly displays that Non-White recipients were much more likely to receive amounts under \$10,000 compared to White recipients.

Furthermore, White recipients were more likely to receive amounts above \$20,000, even though White recipients were outnumbered by Non-White recipients.

Without performing any statistical analyses (i.e., a t-test), this graph can give audiences a strong understanding of the discrepancy between expenditure amounts for White compared to Non-White recipients.