Exercise2_HaoranXu

Haoran Xu

2025-01-24

1. Why is EDA an important part of a modeling exercise?

EDA allows us get a brief glimpse of the dataset to understand its structure, characteristics, and quality before conducting data analysis. It also helps us to possibly spot patterns, trends, and correlations within and between variables so that we can detect dataset defects in advance and choose the right model.

2. How many aesthetic elements of visualization can you think of? List them and explain how they could be mapped to variables in different scales.

x-axis, y-axis

Quantitative variables: mapped to precise positions Categorical variables: mapped to discrete positions
e.g.,

```
ggplot(aes(x = mode, y = pollutants))
```

color

Quantitative variables: use distinct and discrete colors Categorical variables: use gradients or sequential colors
e.g.,

```
geom_bar(color = "black",
           fill = "white")
```

size

Adjust the sizes of points, lines, or areas.
e.g.,

```
geom_segment(size = 1)
```

shape

Adjust the shapes of points.
e.g.,

```
geom_sf(data = data,
          size = 2,
          shape = 1)
```

line type / thickness

Adjust the shapes or width of lines.

transparency (alpha)

Higher values are more opaque, lower values more transparent.

labels

Texts for axises or legends.
e.g.,

```
labs(y="Mode",
     x = expression("Sidewalk density (km/km"^2*")"),
     # Add a label for the fill
     fill = "Mode")
```

faceting

Create multiple subplots.
e.g.,

```
facet_wrap(~ available.Walk,
           labeller = label_both)
```

fill

Quantitative variables: use distinct and discrete colors Categorical variables: use gradients or sequential colors
e.g.,

```
scale_fill_brewer(palette = "Dark2")
```

3. The data set explored in this chapter was in wide format: each row was a single respondent. However, discrete choice data is often found in long format: each row is a choice situation, and each individual now appears in the table as many times as they faced a choice situation. Would this change in any way how you approach data analysis?

I think this would change someway about how to conduct data analysis. Wide format data have rows corresponding to a single respondent, which is better for individual-level summary statistics. While long format data have rows corresponding to a single choice situation, which is better for discrete choice models.

Long format datasets have advantages in choice-specific analysis and are needed for models like Multinomial Logit Models, Mixed Logit Models, and Nested Logit Models. They are also good at visualizing choice-specific data analysis.

```
library(mlogit)
data("Car")
```

4. How many variables are there in this data set and of which type (i.e., categorical/quantitative)?

```
summary(Car)
```

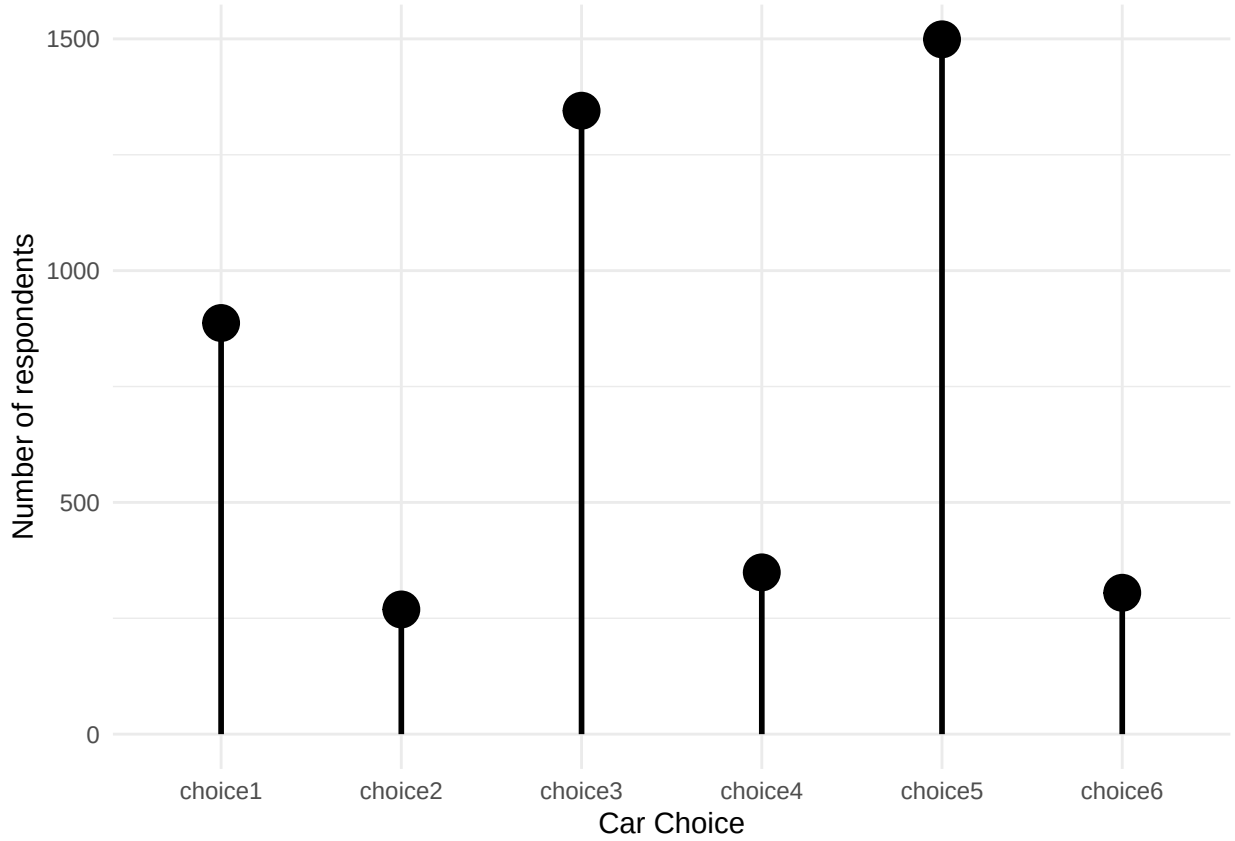
There are 70 variables of different variable types: - Categorical - Nominal (16): choice, college, hsg2, coml5, typez, fuelz, and among them college, hsg2, and coml5 are dummy variables coded in {0, 1}. - Ordinal (6): sizez - Quantitative - Ratio (48): pricez, rangez, accz, speedz, pollutionz, spacez, costz, stationz

5. Choose four relevant categorical variables from this data set and describe them using univariate, bivariate, and multivariate techniques. Discuss your results.

Univariate analysis

```
Car %>%
  group_by(choice) %>%
  summarize(n = n()) %>%
  ggplot(aes(x = choice,
             xend = choice,
             y = n,
             yend = 0)) +
  # Add geometric featues of type point
  geom_point(color = "black",
             fill = "white",
             size = 6) +
  # Add geometric features of type segment (line segments) --> width of lines
  geom_segment(size = 1) +
  # Label the axes
  labs(x = "Car Choice",
       y = "Number of respondents") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



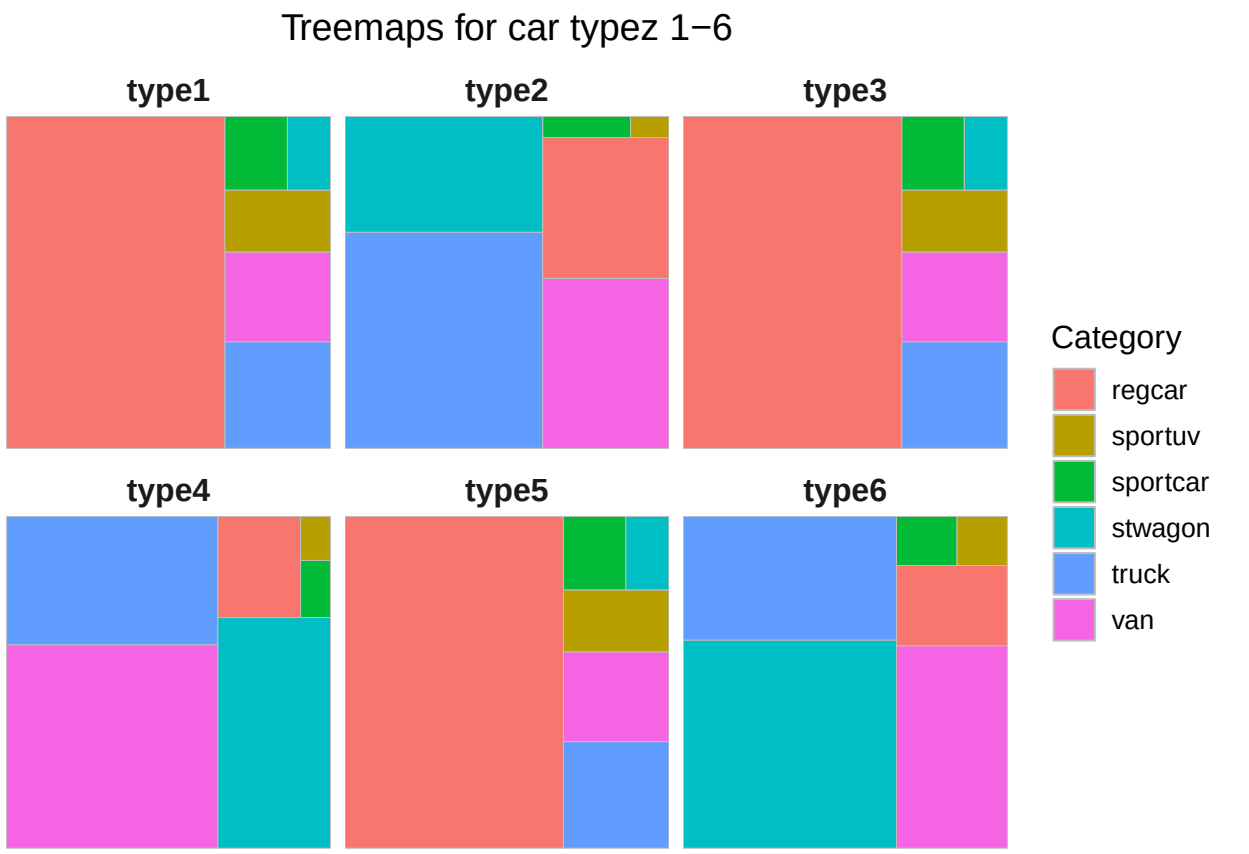
Regarding choices, most of people made choice5, choice3, choice1, while for the other three choices there are less quantities.

```
library(treemapify)
library(tidyr)

# Combine type1 to type6 into a long format
Car_long <- Car %>%
  pivot_longer(cols = starts_with("type"),
               names_to = "type",
               values_to = "category")

# Create treemaps with facets for type1 to type6
Car_long %>%
  group_by(type, category) %>%
  summarize(n = n(), .groups = "drop") %>%
  ggplot(aes(fill = category, area = n)) +
  geom_treemap() +
  facet_wrap(~ type, scales = "free") +
  labs(title = "Treemaps for car typez 1-6",
       fill = "Category") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 12, face = "bold"),
    plot.title = element_text(size = 14, hjust = 0.5),
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10)
```

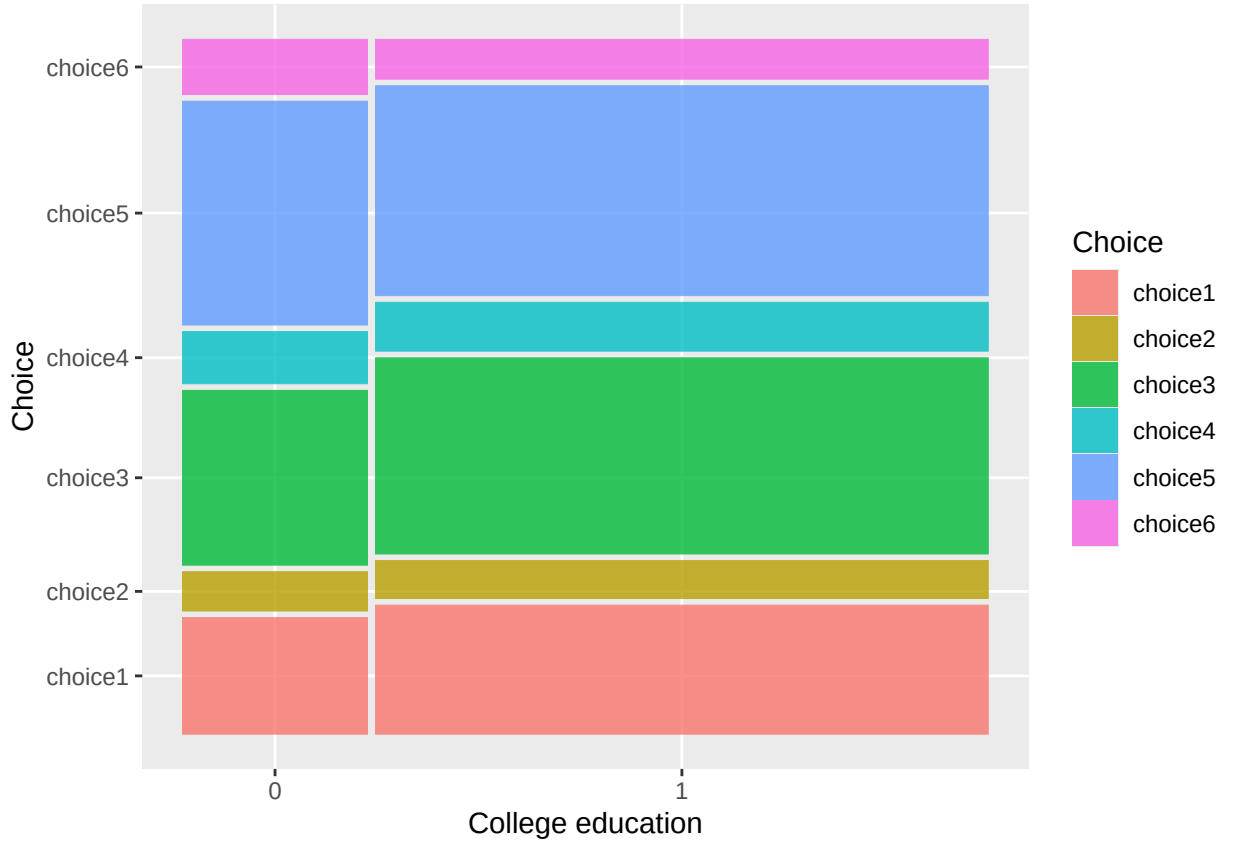
```
)
```



The six propositions of car body types, along with the summary results of these variables, show that the type1, type3, and type5 are exactly the same, in which regular car account for most choices, followed by truck, van, sport utility vehicle, sportcar, and station wagon. While for type2, type4, and type6, truck, van, station wagon account respectively for the most types chosen.

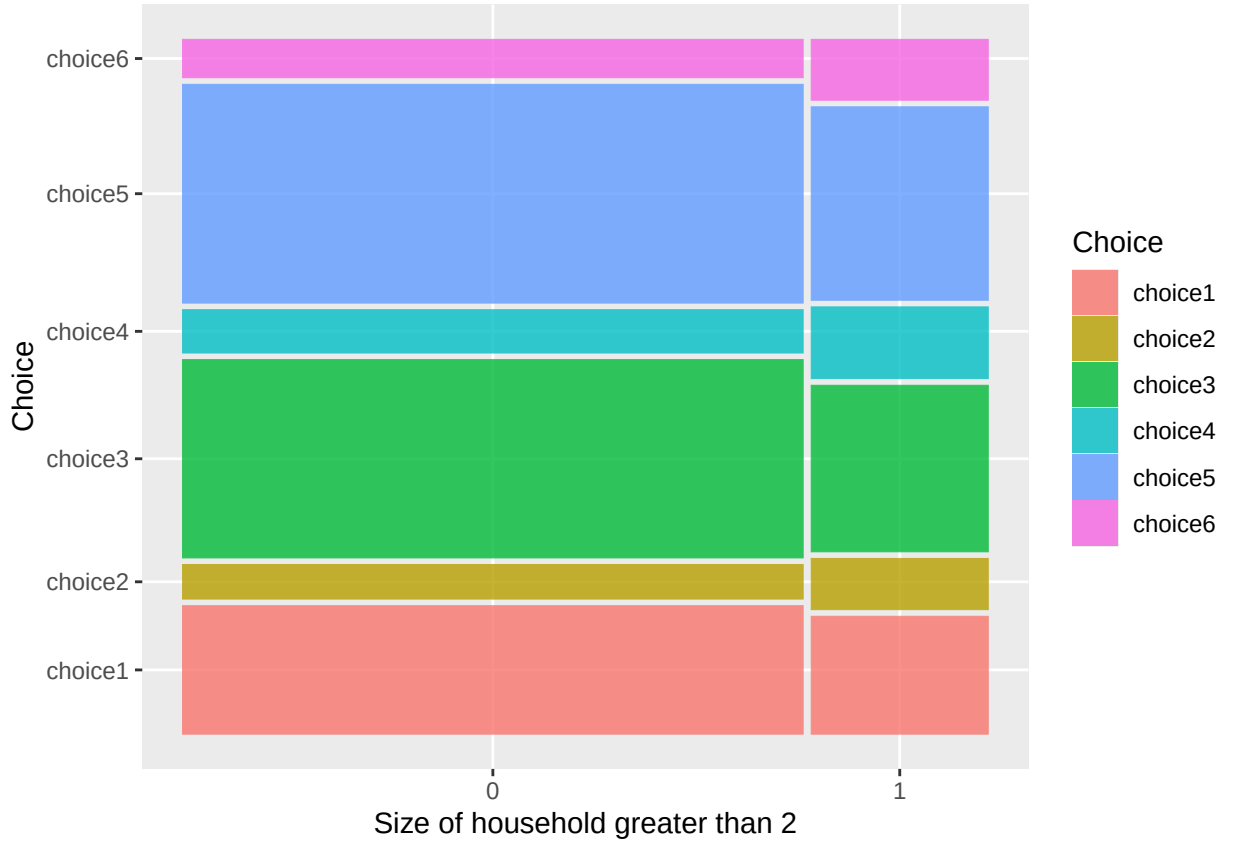
Bivariate analysis

```
Car %>%
  ggplot() +
  geom_mosaic(aes(x = product(choice,
                              college),
                  fill = choice)) +
  # Add labels
  labs(x = "College education",
       y = "Choice",
       fill = "Choice")
```



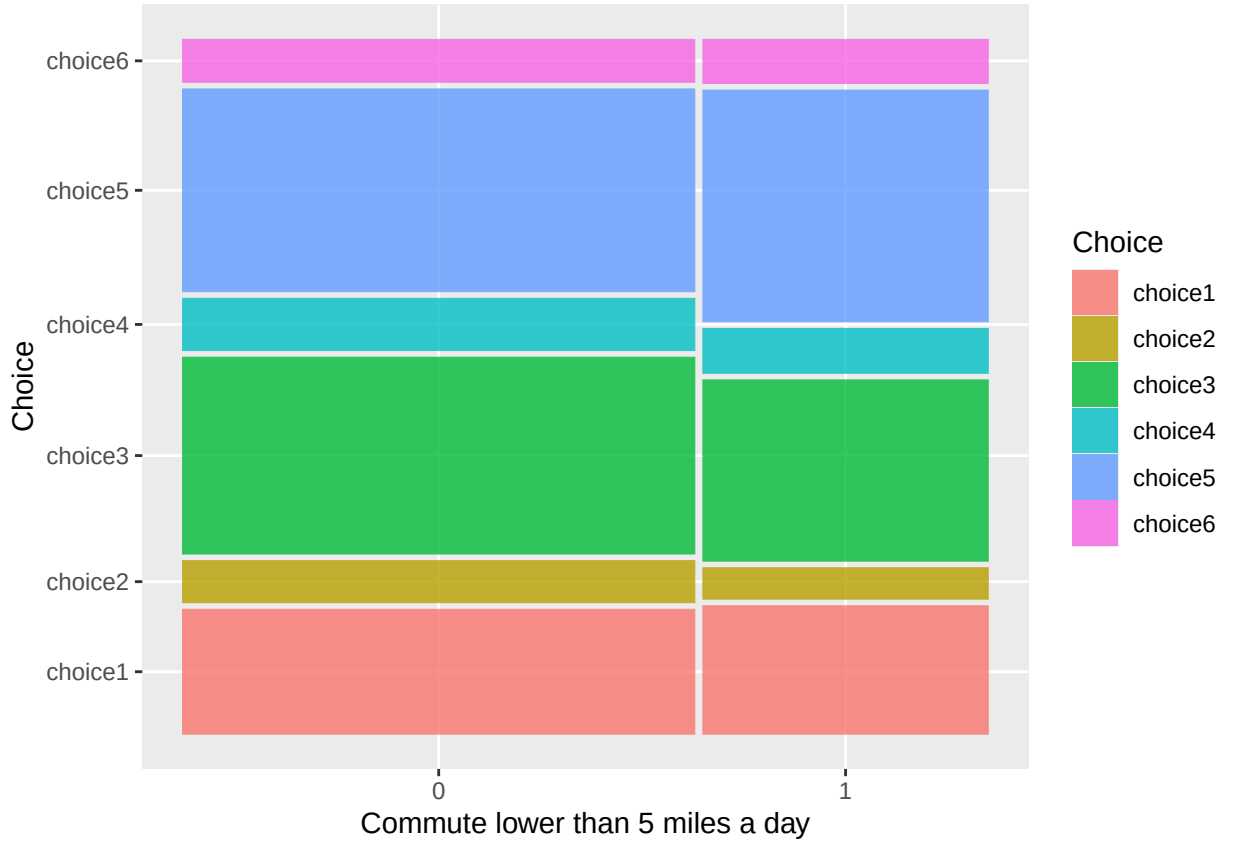
The bivariate plot of size of college education and choice shows that most people have college education. For for those who have it, people tend to choose choice1 and choice3 more often than people who do not.

```
Car %>%
  ggplot() +
  geom_mosaic(aes(x = product(choice,
                              hsg2),
                  fill = choice)) +
  labs(x = "Size of household greater than 2",
       y = "Choice",
       fill = "Choice")
```



The bivariate plot of size of household greater than 2 and choice shows that most people do not have a household greater than 2. For households greater than 2, people tend to choose choice2, choice4 and choice 6 more often than people who do not.

```
Car %>%
  ggplot() +
  geom_mosaic(aes(x = product(choice,
                              coml5),
                  fill = choice)) +
  labs(x = "Commute lower than 5 miles a day",
       y = "Choice",
       fill = "Choice")
```



The bivariate plot of people commuting lower than 5 miles a day and choice shows that most people commute lower than 5 miles a day. For for those who commute higher than 5 miles a day, people tend to choose choice2 and choice3 more often than people who commute less distance.

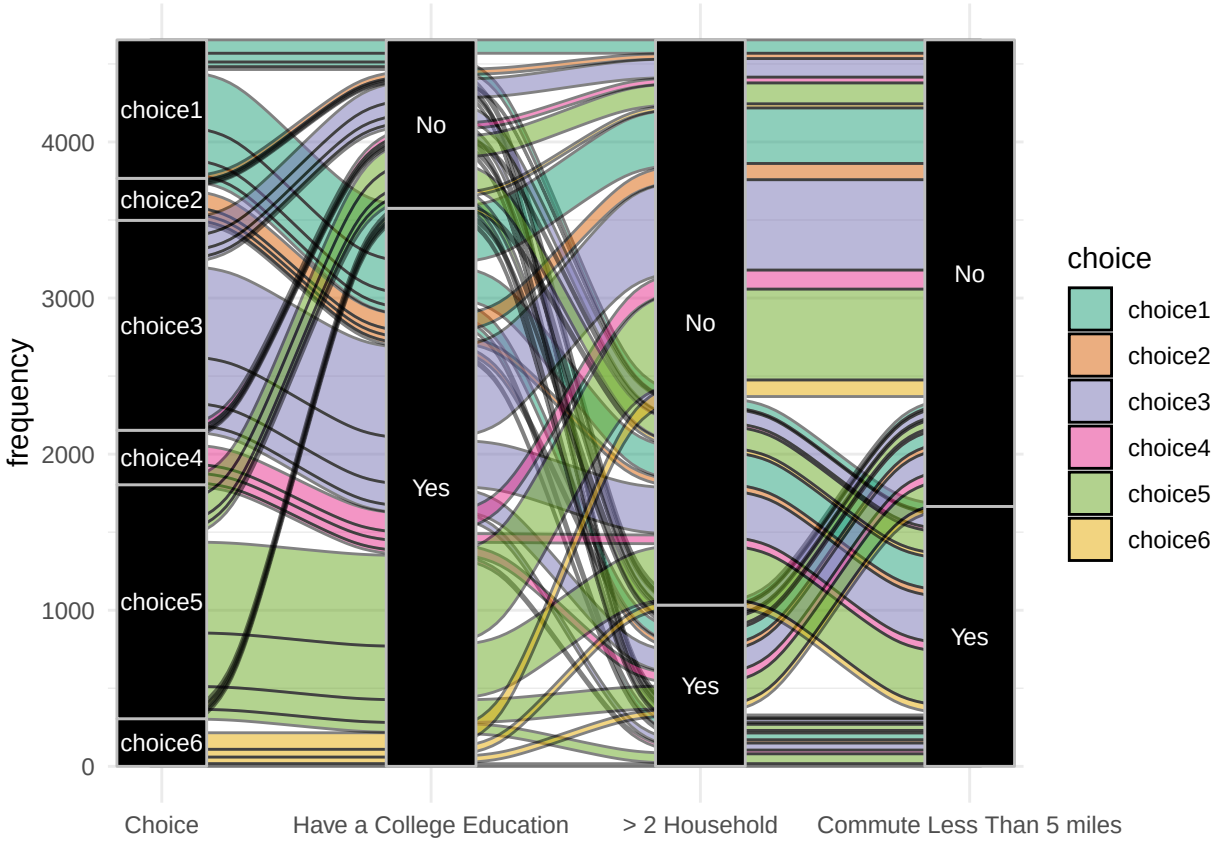
Multi-variate analysis

```
# data preparation
car_alluvia <- Car %>%
  mutate(collegeEducation = case_when(college == "1" ~ "Yes",
                                      TRUE ~ "No"),
         householdGreaterThan2 = case_when(hsg2 == "1" ~ "Yes",
                                      TRUE ~ "No"),
         commuteLowerThan5 = case_when(coml5 == "1" ~ "Yes",
                                      TRUE ~ "No")
         ) %>%
  select(choice, collegeEducation, householdGreaterThan2, commuteLowerThan5) %>%
  group_by(choice, collegeEducation, householdGreaterThan2, commuteLowerThan5) %>%
  summarize(frequency = n(),
            .groups = "drop")
```

```
car_alluvia %>%
  ggplot(aes(y = frequency,
             axis1 = choice,
             axis2 = collegeEducation,
             axis3 = householdGreaterThan2,
             axis4 = commuteLowerThan5)) +
```

```
  geom_alluvium(aes(fill = choice),
                width = 1/12,
                color = "black") +
  geom_stratum(width = 1/3,
               fill = "black",
               color = "grey") +
  geom_text(stat = "stratum",
            aes(label = after_stat(stratum)),
            color = "white",
            size = 3) +
  scale_x_discrete(limits = c("Choice",
                              "Have a College Education",
                              "> 2 Household",
                              "Commute Less Than 5 miles"),
                   expand = c(.05, .05)) +
  scale_fill_brewer(type = "qual",
                    palette = "Dark2") +
  theme_minimal()
```

```
## Warning in to_lodes_form(data = data, axes = axis_ind, discern =
## params$discern): Some strata appear at multiple axes.
## Warning in to_lodes_form(data = data, axes = axis_ind, discern =
## params$discern): Some strata appear at multiple axes.
## Warning in to_lodes_form(data = data, axes = axis_ind, discern =
## params$discern): Some strata appear at multiple axes.
```



According to the multi-variate alluvial graph drawn above, a majority of people who chose choice3 and choice5 have a college education, does not have a household more than 2 people, and commute less than 5 miles a day.

6. What ideas about individuals' choices regarding car do you develop from this EDA exercise?

People who chose choice1 or choice3 (regular car) have a higher tendency of having a college education and live in a household of no more than 2 people. While the variable of commuting less than 5 miles does not have a significant influence over people's choice over cars.