LINE 239 >

- Do we have excessive number of categories? Do we want to combine some?

```{r}

# Pie chart: Category Distribution

category\_distribution <- card\_fraud %>%

count(category, sort = TRUE) %>%

mutate(perc = n / sum(n))

# Extract top 10 categories and combine remaining categories into "Other" category

top\_categories <- head(category\_distribution, 10)

other\_category <- data\_frame(category = "Other", n = sum(category\_distribution[-(1:10), "n"]), perc = sum(category\_distribution[-(1:10), "perc"]))

combined\_categories <- bind\_rows(top\_categories, other\_category)

# Pie chart with top 10 categories and "Other" category

ggplot(combined\_categories, aes(x = "", y = perc, fill = category)) +

geom\_bar(width = 1, stat = "identity") +

coord\_polar("y", start = 0) +

labs(title = "Category Distribution",

fill = "Category",

x = NULL, y = NULL) +

theme\_void() +

theme(legend.position = "right")

# Pie chart: Job Distribution of Card Fraud Holders

job\_distribution <- card\_fraud %>%

count(job, sort = TRUE) %>%

mutate(perc = n / sum(n))

# Extract top 20 jobs and combine remaining jobs into "Other" category

top\_jobs <- head(job\_distribution, 20)

other\_jobs <- data\_frame(job = "Other", n = sum(job\_distribution[-(1:20), "n"]), perc = sum(job\_distribution[-(1:20), "perc"]))

combined\_jobs <- bind\_rows(top\_jobs, other\_jobs)

# Pie chart with top 20 jobs and "Other" category

ggplot(combined\_jobs, aes(x = "", y = perc, fill = job)) +

geom\_bar(width = 1, stat = "identity") +

coord\_polar("y", start = 0) +

labs(title = "Job Distribution",

fill = "Job",

x = NULL, y = NULL) +

theme\_void() +

theme(legend.position = "right")

```

# Group1: Yes, both category and job have too many categories, espeially job variable. It does not make sense to group Others for jobs because the largest group is very small so there is no good threshold. we just drop this var. However, we can group category var.

```{r}

card\_fraud %>%

count(category, sort=TRUE)%>%

mutate(perc = n/sum(n))

card\_fraud %>%

count(job, sort=TRUE) %>%

mutate(perc = n/sum(n))

```

```{r}

#Visualization: Category Distribution

card\_fraud %>%

count(category, sort = TRUE) %>%

mutate(perc = n / sum(n)) %>%

ggplot(aes(x = reorder(category, perc), y = perc, fill = category)) +

geom\_bar(stat = "identity") +

labs(title = "Category Distribution of Card Fraud", x = "Category", y = "Percentage") +

theme\_bw() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

```