Frequency Distributions: Takeaways

Syntax

- Generating a frequency distribution table for a column (col):

```
df %>%
  group_by(col) %>%
  summarize(Freq = n())
```

- Sorting the values of frequency distribution table in ascending order (default):

```
df %>%
  group_by(col) %>%
  summarize(Freq = n())
```

- Sorting in ascending order (being explicit):

```
df %>%
  group_by(col) %>%
  summarize(Freq = n()) %>%
  arrange(col)
```

- Sorting in descending order:

```
df %>%
  group_by(col) %>%
  summarize(Freq = n()) %>%
  arrange(desc(col))
```

- Finding proportions and percentages in a frequency distribution table:

```
df %>%
  group_by(col) %>%
  summarize(Freq = n()) %>%
  mutate(Prop = Freq / nrow(df)) %>%
  mutate(Percentage = Freq / nrow(df) * 100) %>%
  arrange(desc(Freq))
```

- Finding the percentile rank of a value (score) in a column:

```
mean(df$col <= value) * 100
```

- Finding percentiles – only the quartiles:

```
quantile(df$col)
```

- Finding any percentile we designate:

```
quantile(df$col,
            probs = c(0, 0.1, 0.25, 0.33, 0.5, 0.66, 0.75, 0.9, 1))
```

- Generate percentiles for each value in a dataframe:

```
df %>%
  mutate(cume_dist_col = cume_dist(col))
```

- Generating a grouped frequency table:

```
df <- df %>%
  mutate(categories = cut(col, breaks = 5, dig.lab = 4))
df %>%
  group_by(categories) %>%
  summarize(Freq = n())
```

- Generating a grouped frequency table with custom class intervals:

```
df <- df %>%
  mutate(categories =
           cut(col,
               breaks = c(0, 150, 300, 450,
               600, 750, 900, 1050),
               dig.lab = 4))
```

- Showing frequency and percentage

```
df %>%
  group_by(categories) %>%
  summarize(Freq = n()) %>%
  mutate(Percentage = Freq / nrow(df) * 100)
```

Concepts

- A table that shows the frequency for each unique value in a distribution is called a **frequency distribution table**.
- The frequencies can be expressed as:
    - Absolute counts (**absolute frequencies**).
    - Proportions or percentages (**relative frequencies**).
- **Quantiles** provide us with the value of a random variable for a specified probability.
- The three percentiles that divide the distribution in *four* equal parts are also known as **quartiles**. The lower quartile is the value of the quantile at probability 0.25.
- The percentage of values that are equal or less than a value is called the **percentile rank** of . For instance, if the percentile rank of a value of 32 is 57%, 57% of the values are equal to or less than 32.
- If a value has a percentile rank of , we say that is the **percentile**. For instance, if 32 has a percentile rank of 57%, we say that 32 is the 57th percentile.
- Frequency distribution tables can be grouped in **class intervals** to form **grouped frequency distribution tables**. As a rule of thumb, 10 is a good number of class intervals to choose because it offers a good balance between information and comprehensibility.

Resources

- An intuitive introduction to frequency distribution tables.
- An intuitive introduction to grouped frequency distribution tables.
- The Wikipedia entry on frequency distributions.