

# ETC5521: Exploratory Data Analysis

**Working with a single variable, making transformations, detecting outliers, using robust statistics**

Lecturer: *Emi Tanaka*

Department of Econometrics and Business Statistics

✉ ETC5521.Clayton-x@monash.edu

Week 4 - Session 2

# Categorical variables

This lecture is based on Chapter 4 of  
Unwin (2015) Graphical Data Analysis with R

# **There are two types of categorical variables**

**Nominal** where there is no intrinsic ordering to the categories  
E.g. blue, grey, black, white.

**Ordinal** where there is a clear order to the categories.  
E.g. Strongly disagree, disagree, neutral, agree, strongly agree.

# Categorical variables in R

- In R, categorical variables may be encoded as **factors**.

```
data <- c(2, 2, 1, 1, 3, 3, 3, 1)
factor(data)

## [1] 2 2 1 1 3 3 3 1
## Levels: 1 2 3
```

- You can easily change the labels of the variables:

```
factor(data, labels = c("I", "II", "III"))

## [1] II  II  I   I   III III III I
## Levels: I II III
```

- Order of the factors are determined by the input:

```
# numerical input are ordered in increasing order
factor(c(1, 3, 10))

## [1] 1 3 10
## Levels: 1 3 10
```

```
# character input are ordered alphabetically
factor(c("1", "3", "10"))

## [1] 1 3 10
## Levels: 1 10 3
```

```
# you can specify order of levels explicitly
factor(c("1", "3", "10"), levels = c("1", "3", "10"))

## [1] 1 3 10
## Levels: 1 3 10
```

# Numerical factors in R

```
x <- factor(c(10, 20, 30, 10, 20))
mean(x)

## Warning in mean.default(x): argument is not numeric or logical: returning NA

## [1] NA
```

⚠ as.numeric function returns the internal integer values of the factor

```
mean(as.numeric(x))

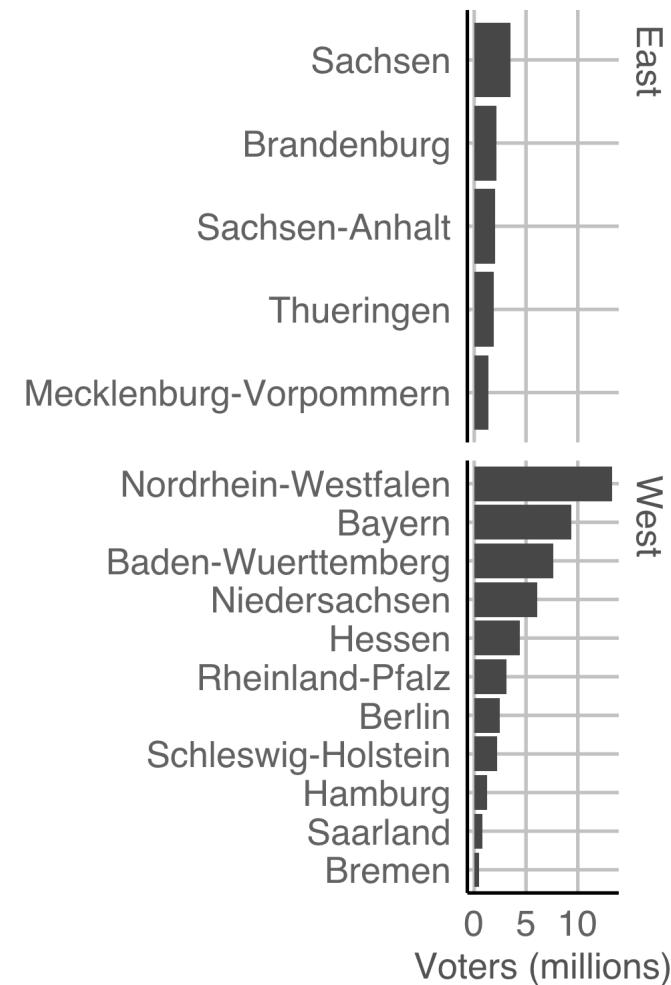
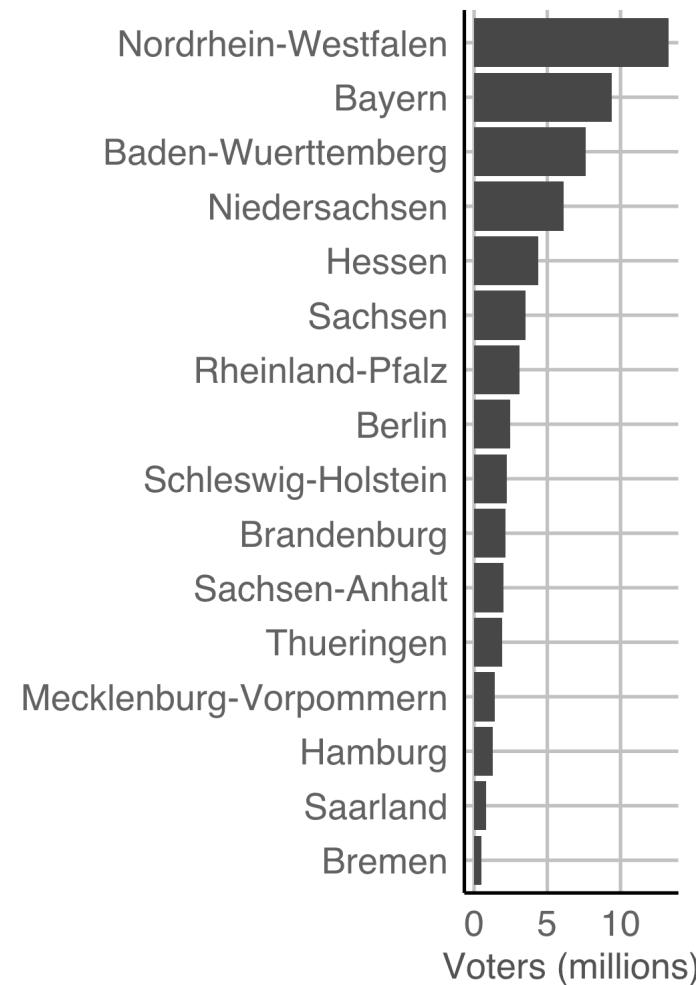
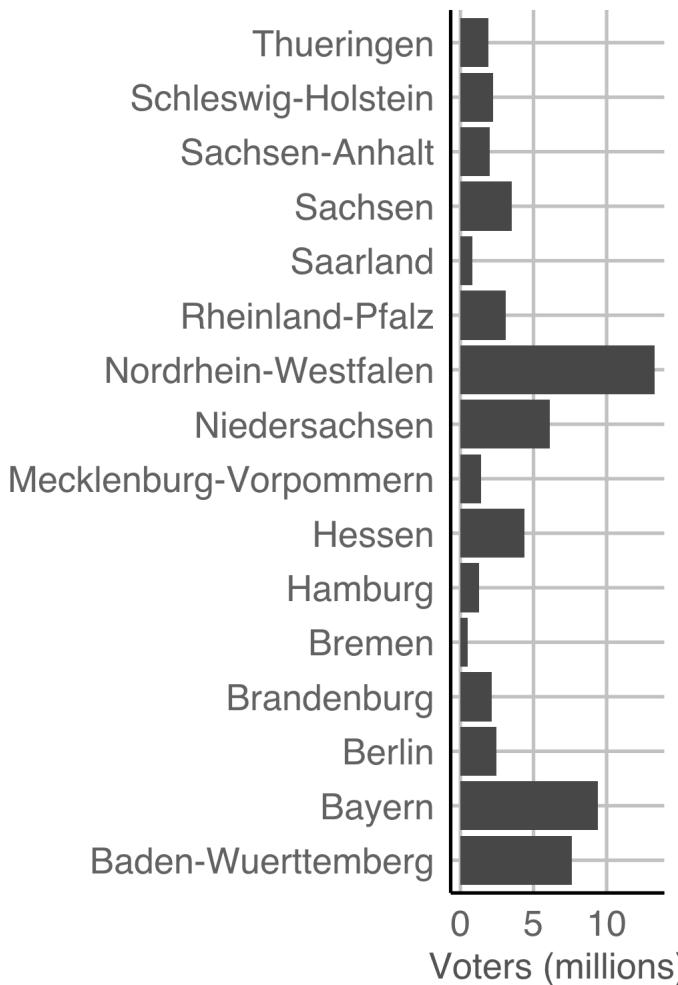
## [1] 1.8
```

You probably want to use:

mean(as.numeric(levels(x)[x]))	mean(as.numeric(as.character(x)))	•
## [1] 18	## [1] 18	•

# Revisiting case study ① German Bundestag Election 2009

data R



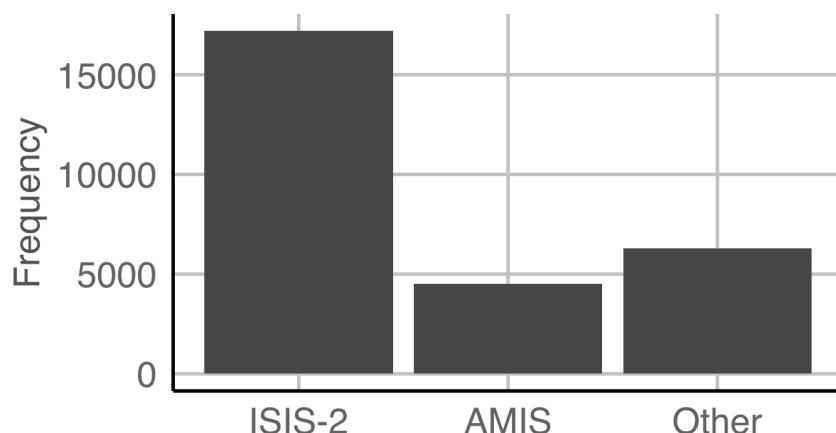
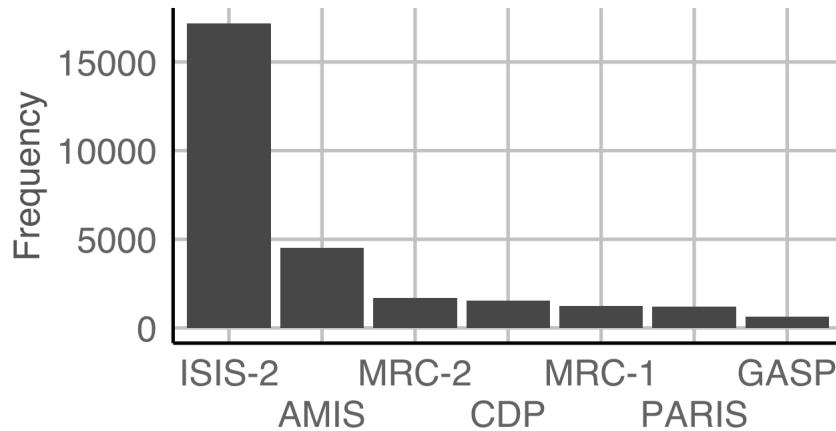
# Order nominal variables meaningfully

**</> Coding tip:** use below functions to easily change the order of factor levels

```
stats::reorder(factor, value, mean)  
forcats::fct_reorder(factor, value, median)  
forcats::fct_reorder2(factor, value1, value2, func)
```

# Case study 8 Aspirin use after heart attack

data R



- Meta-analysis is a statistical analysis that combines the results of multiple scientific studies.
- This data studies the use of aspirin for death prevention after myocardial infarction, or in plain terms, a heart attack.
- The ISIS-2 study has more patients than all other studies combined.
- You could consider lumping the categories with low frequencies together.

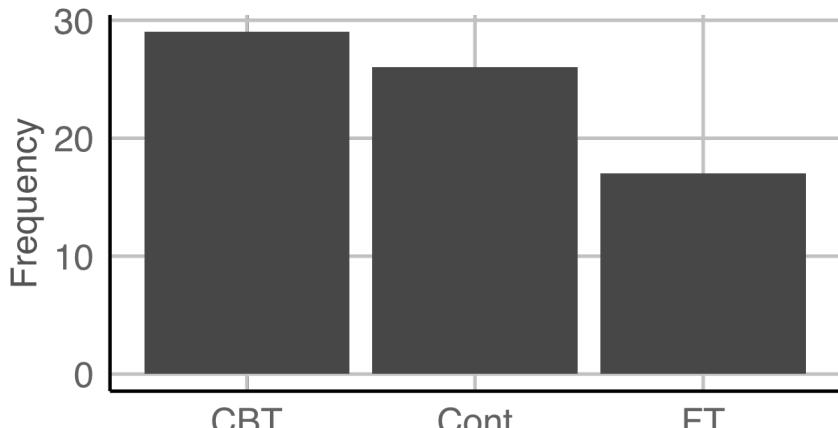
# Consider combining factor levels with low frequencies

🔗 Coding tip: the following family of functions help to easily lump factor levels together:

```
forcats::fct_lump()  
forcats::fct_lump_lowfreq()  
forcats::fct_lump_min()  
forcats::fct_lump_n()  
forcats::fct_lump_prop()  
# if conditioned on another variable  
ifelse(cond, "Other", factor)  
dplyr::case_when(cond1 ~ "level1",  
                 cond2 ~ "level2",  
                 TRUE ~ "Other")
```

# Case study 9 Anorexia

data R

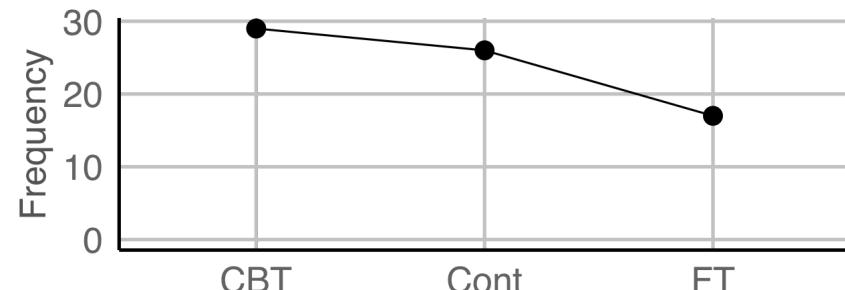


Treatment	Frequency
CBT	29
Cont	26
FT	17

## Table or Plot?

- Table for accuracy, plot for visual communication.

## Why not a point or line?



- This can be appropriate depending on what you want to communicate.
- A barplot occupies more area compared to a point and the area does a better job of communicating size.
- A line is suggestive of a trend.

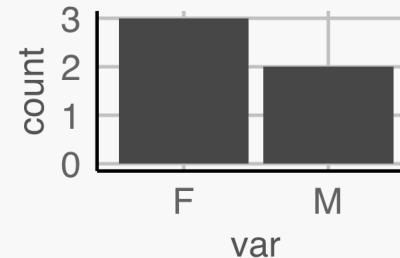
# geom\_bar or geom\_col?

```
df <- data.frame(var = c("F", "F", "M", "M", "F"))
dftab <- as.data.frame(table(df$var))
```

```
df
```

```
ggplot(df, aes(var)) + geom_bar()
```

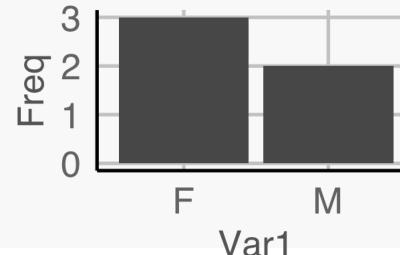
```
##   var
## 1   F
## 2   F
## 3   M
## 4   M
## 5   F
```



```
dftab
```

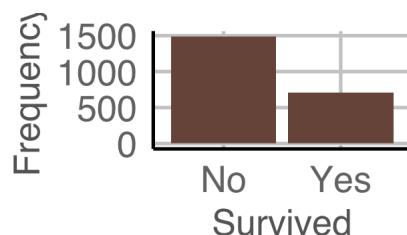
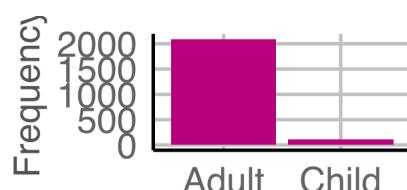
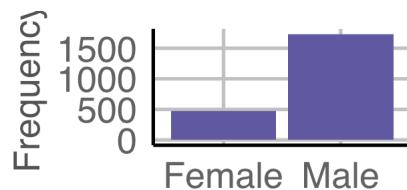
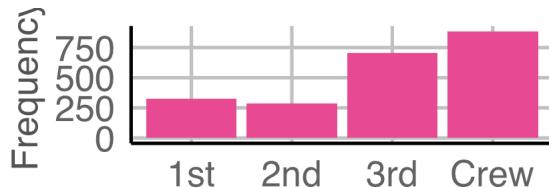
```
ggplot(dftab, aes(Var1, Freq)) + geom_col()
```

```
##   Var1 Freq
## 1   F     3
## 2   M     2
```



# Case study 10 Titanic

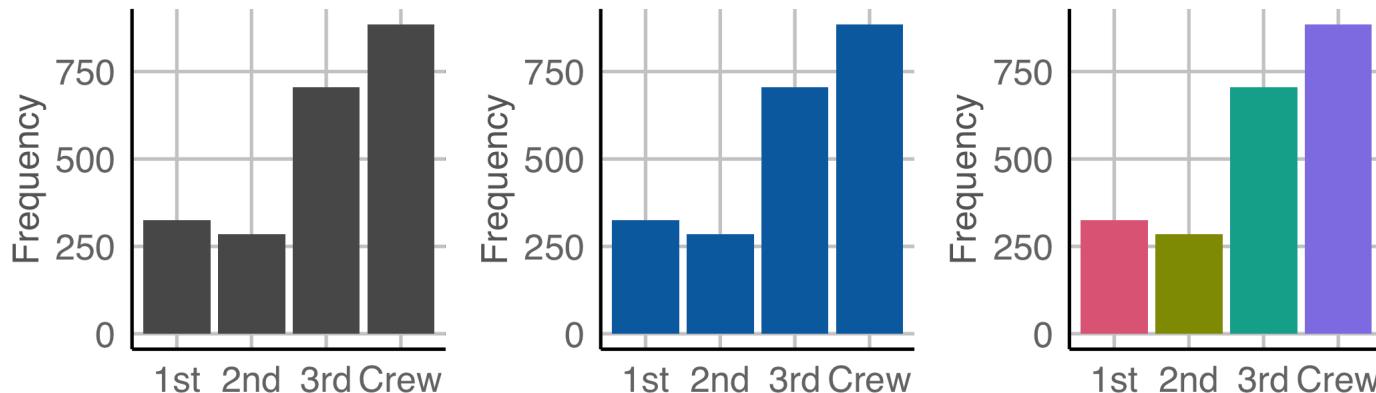
data R



**What does the graphs for each categorical variable tell us?**

- There were more crews than 1st to 3rd class passengers
- There were far more males on ship; possibly because majority of crew members were male. You can further explore this by constructing two-way tables or graphs that consider both variables.
- Most passengers were adults.
- More than two-thirds of passengers died.

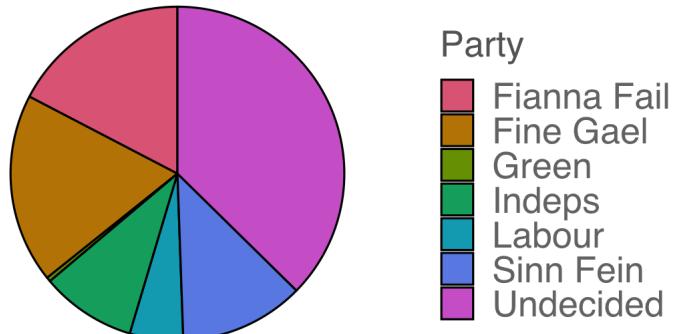
# Coloring bars



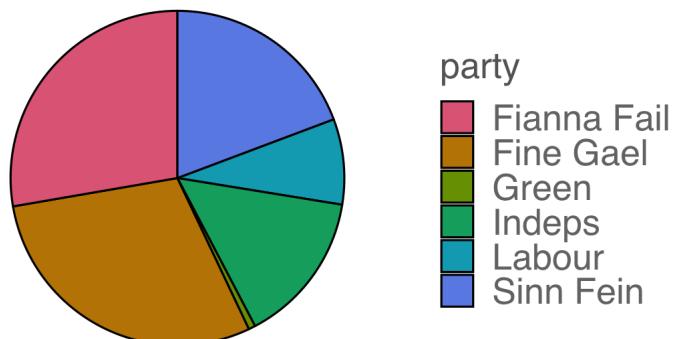
- Colour here doesn't add information as the x-axis already tells us about the categories, but colouring bars can make it more visually appealing.
- If you have too many categories colour won't work well to differentiate the categories.

# Case study 11 Opinion poll in Ireland Aug 2013

 data R

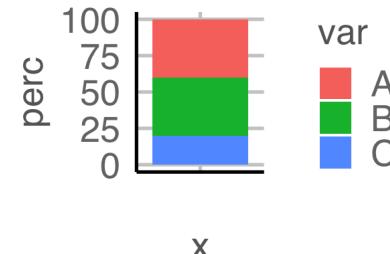


- Pie chart is popular in mainstream media but are not generally recommended as people are generally poor at comparing angles.
- 3D pie charts should definitely be avoided!
- Here you can see that there are many people that are "Undecided" for which political party to support and failing to account for this paints a different picture.

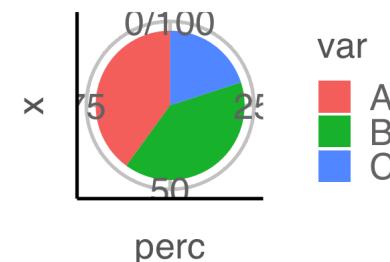


# Piechart is a stacked barplot just with a transformed coordinate system

```
df <- data.frame(var = c("A", "B", "C"), perc = c(40, 40, 20))
g <- ggplot(df, aes("", perc, fill = var)) +
  geom_col()
g
```

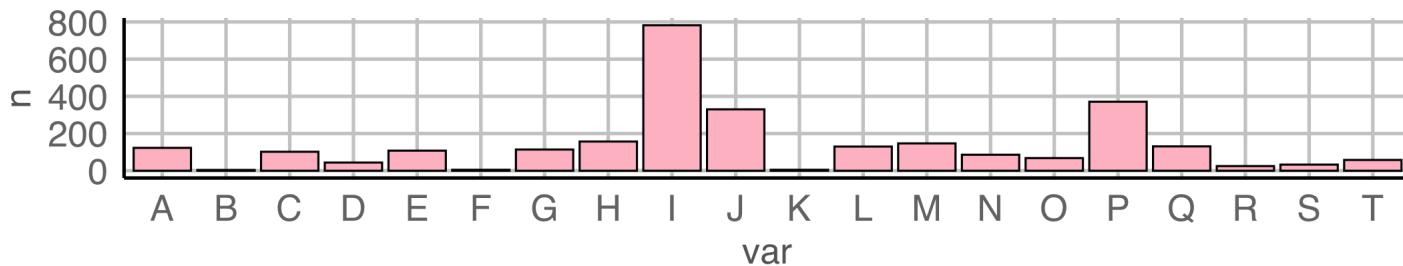


```
g + coord_polar("y")
```



# Roseplot is a barplot just with a transformed coordinate system

```
dummy <- data.frame(var = LETTERS[1:20],  
                      n = round(rexp(20, 1/100)))  
g <- ggplot(dummy, aes(var, n)) + geom_col(fill = "pink", color = "black")  
g
```



```
g + coord_polar("x") + theme_void()
```



# That's it, for this lecture!



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

✉ ETC5521.Clayton-x@monash.edu