

# Socio-Informatics 348

## Data Transformation with the Tidyverse Visualising the Transformations

Dr Lisa Martin

Department of Information Science  
Stellenbosch University

# Useful Visualisations

- Visualisations created by or adapted from Garrick Aden-Buie
- Useful for understanding how some dplyr transformations work

# Mutate

**df** |>

cat1	cat2	x
a	j	1
a	j	2
a	k	3
b	j	4
b	k	5
b	k	6
c	j	7
c	j	8
c	k	9

**mutate(y = x - min(x))**

cat1	cat2	x	y
a	j	1	0
a	j	2	1
a	k	3	2
b	j	4	3
b	k	5	4
b	k	6	5
c	j	7	6
c	j	8	7
c	k	9	8

```
df |>  
  mutate(  
    y = x - min(x)  
  )
```

```
df |>  
  mutate(  
    y = x - min(x),  
    .after = cat2)
```

cat1	cat2	y	x
a	j	0	1
a	j	1	2

# Group and mutate

df			group_by(cat1)  > mutate(y = x - min(x))		
cat1	cat2	x	cat1	cat2	x
a	j	1	a	j	1
a	j	2	a	j	2
a	k	3	a	k	3
b	j	4	b	j	4
b	k	5	b	k	5
b	k	6	b	k	6
c	j	7	c	j	7
c	j	8	c	j	8
c	k	9	c	k	9

cat1	cat2	x	y
a	j	1	0
a	j	2	1
a	k	3	2
b	j	4	0
b	k	5	1
b	k	6	2
c	j	7	0
c	j	8	1
c	k	9	2

```
df |>  
  group_by(cat1) |>  
  mutate(  
    y = x - min(x)  
  )
```

# Group and summarise I

**df** |>

cat1	cat2	x
a	j	1
a	j	2
a	k	3
b	j	4
b	k	5
b	k	6
c	j	7
c	j	8
c	k	9

**group\_by(cat1)** |> **summarize(...)**

cat1	cat2	x
a	j	1
a	j	2
a	k	3
b	j	4
b	k	5
b	k	6
c	j	7
c	j	8
c	k	9

  

cat1	avg	total	n
a	2	6	3

  

cat1	avg	total	n
b	5	15	3

  

cat1	avg	total	n
c	8	24	3

```
df |>
  group_by(cat1) |>
  summarize(
    avg = mean(x),
    total = sum(x),
    n = n()
  )
```

cat1	avg	total	n
a	2	6	3
b	5	15	3
c	8	24	3

# Group and summarise II

**df** |>

cat1	cat2	x
a	j	1
a	j	2
a	k	3
b	j	4
b	k	5
b	k	6
c	j	7
c	j	8
c	k	9

**group\_by(cat2)** |> **summarize(...)**

cat1	cat2	x
a	j	1
a	j	2
b	j	4
c	j	7
c	j	8

  

cat2	avg	total	n
j	4.4	22	5

  

cat1	cat2	x
a	k	3
b	k	5
b	k	6
c	k	9

  

cat2	avg	total	n
k	5.75	23	4

```
df |>
  group_by(cat2) |>
  summarize(
    avg = mean(x),
    total = sum(x),
    n = n()
  )
```

cat2	avg	total	n
j	4.4	22	5
k	5.75	23	4

# Group and summarise III

```
df |> group_by(cat1, cat2) |> summarize(...) |> ungroup()
```

cat1	cat2	x
a	j	1
a	j	2
a	k	3
b	j	4
b	k	5
b	k	6
c	j	7
c	j	8
c	k	9

cat1	cat2	x
a	j	1
a	j	2
a	k	3
b	j	4
b	k	5
b	k	6
c	j	7
c	j	8
c	k	9

cat1	cat2	avg	total	n
a	j	1.5	3	2
a	k	3	3	1
b	j	4	4	1
b	k	5.5	11	2
c	j	7.5	15	2
c	k	9	9	1

```
df |> group_by(cat1, cat2) |> summarize(
  avg = mean(x),
  total = sum(x),
  n = n()
) |> ungroup()
```

cat1	cat2	avg	total	n
a	j	1.5	3	2
a	k	3	3	1
b	j	4	4	1
b	k	5.5	11	2
c	j	7.5	15	2
c	k	9	9	1

# Group and slice - min

df  >			group_by(cat1)  >			slice_min(x, n=1)		
cat1	cat2	x	cat1	cat2	x	cat1	cat2	x
a	j	1	a	j	1	a	j	1
a	j	2	a	j	2	b	j	4
a	k	3	a	k	3	c	j	7
b	j	4	b	j	4			
b	k	5	b	k	5			
b	k	6	b	k	6			
c	j	7						
c	j	8	c	j	7			
c	k	9	c	j	8			
			c	k	9			

```
df |>  
  group_by(cat1) |>  
  slice_min(x, n=1)
```



# Group and slice - max

df  >			group_by(cat1)  >			slice_max(x, n=1)		
cat1	cat2	x	cat1	cat2	x	cat1	cat2	x
a	j	1	a	j	1	a	k	3
a	j	2	a	j	2	b	k	6
a	k	3	a	k	3	c	k	9
b	j	4	b	j	4			
b	k	5	b	k	5			
b	k	6	b	k	6			
c	j	7	c	j	7			
c	j	8	c	j	8			
c	k	9	c	k	9			

```
df |>  
  group_by(cat1) |>  
  slice_max(x, n=1)
```