Data wrangling

Working with multiple data frames

Prof. Dr. Jan Kirenz

Inputs

professions dates works

```
dates
```

```
## # A tibble: 8 x 3
##   name                birth_year death_year
##   <chr>                    <dbl>      <dbl>
## 1 Janaki Ammal              1897       1984
## 2 Chien-Shiung Wu           1912       1997
## 3 Katherine Johnson         1918       2020
## 4 Rosalind Franklin         1920       1958
## 5 Vera Rubin                1928       2016
## 6 Gladys West               1930         NA
## 7 Flossie Wong-Staal        1947         NA
## 8 Jennifer Doudna           1964         NA
```

Inputs

professions dates works

```
works
```

```
## # A tibble: 9 x 2
##   name              known_for                                             
##   <chr>             <chr>                                                 
## 1 Ada Lovelace      first computer algorithm                              
## 2 Marie Curie       theory of radioactivity,  discovery of element…
## 3 Janaki Ammal      hybrid species, biodiversity protection               
## 4 Chien-Shiung Wu   confim and refine theory of radioactive beta d…
## 5 Katherine John… calculations of orbital mechanics critical to …
## 6 Vera Rubin        existence of dark matter                              
## 7 Gladys West       mathematical modeling of the shape of the Eart…
## 8 Flossie Wong-S… first scientist to clone HIV and create a map …
## 9 Jennifer Doudna one of the primary developers of CRISPR, a gro…
```

Student records

In class

Survey missing

Dropped

```
enrolment %>%
  anti_join(survey, by = "id")
```

```
## # A tibble: 1 x 2
##      id name       
##   <dbl> <chr>      
## 1     1 Dave Friday
```

Student records

In class

Survey missing

Dropped

```r
survey %>%
  anti_join(enrolment, by = "id")
```

```
## # A tibble: 2 x 3
##      id name  username
##   <dbl> <chr> <chr>
## 1     4 Peter peter_bakes
## 2     5 Mark  thebakingbuddha
```

Grocery sales

Total revenue

Revenue per customer

```
purchases %>%
  left_join(prices)
```

```
## # A tibble: 5 x 3
##   customer_id item         price
##         <dbl> <chr>        <dbl>
## 1           1 bread         1   
## 2           1 milk          0.8 
## 3           1 banana        0.15
## 4           2 milk          0.8 
## 5           2 toilet paper  3   
```

```
purchases %>%
  left_join(prices) %>%
  summarise(total_revenue = sum(price))
```

```
## # A tibble: 1 x 1
##   total_revenue
##           <dbl>
## 1          5.75
```

Grocery sales

Total revenue

Revenue per customer

```
purchases %>%
  left_join(prices)
```

```
## # A tibble: 5 x 3
##   customer_id item          price
##         <dbl> <chr>         <dbl>
## 1           1 bread          1   
## 2           1 milk           0.8 
## 3           1 banana         0.15
## 4           2 milk           0.8 
## 5           2 toilet paper   3   
```

```
purchases %>%
  left_join(prices) %>%
  group_by(customer_id) %>%
  summarise(total_revenue = sum(price))
```

```
## # A tibble: 2 x 2
##   customer_id total_revenue
##         <dbl>         <dbl>
## 1           1          1.95
## 2           2          3.8 
```