Tidy Tuesday

Week 33

AUTHOR
Cristian T

PUBLISHED
August 20, 2024

```
# Load the tidytuesday package
suppressMessages(library(tidytuesdayR))
suppressMessages(library(skimr))
suppressMessages(library(tidyverse))
suppressMessages(library(dplyr))


# Load the current week's dataset
tuesdata <- tidytuesdayR::tt_load('2024-08-13')
```

```
Downloading file 1 of 1: `worlds_fairs.csv`
```

```
worlds_fairs <- tuesdata$worlds_fairs

# Explore the structure of the dataset
str(worlds_fairs)
```

```
spc_tbl_ [70 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ start_month        : num [1:70] 4 5 5 4 5 5 5 10 4 5 ...
 $ start_year         : num [1:70] 1851 1855 1862 1867 1873 ...
 $ end_month          : num [1:70] 10 11 11 11 10 11 11 4 12 10 ...
 $ end_year           : num [1:70] 1851 1855 1862 1867 1873 ...
 $ name_of_exposition : chr [1:70] "The Great Exhibition" "Exposition Universelle / Paris
International" "International Exhibition" "Exposition Universelle / Paris International"
...
 $ country            : chr [1:70] "United Kingdom" "France" "United Kingdom" "France"
...
 $ city               : chr [1:70] "London" "Paris" "London" "Paris" ...
 $ category           : chr [1:70] "World Expo" "World Expo" "World Expo" "World Expo"
...
 $ theme              : chr [1:70] "Industry of all Nations" "Agriculture, Industry and
Arts" "Industry and Arts" "Agriculture, Industry and Arts" ...
 $ notables           : chr [1:70] "The Crystal Palace" "Palais d'Industrie, Bordeaux
Wine classification" NA "Champ de Mars" ...
 $ visitors           : num [1:70] 6 5 6 15 7.25 10 16 1.3 2.3 32 ...
 $ cost               : num [1:70] 165 2 2 45 95 8 11 16 17 3 ...
 $ area               : num [1:70] 10 15 15 69 233 115 75 25 47 96 ...
 $ attending_countries: num [1:70] 25 25 39 42 35 35 36 33 30 35 ...
 - attr(*, "spec")=
  .. cols(
  ..   start_month = col_double(),
  ..   start_year = col_double(),
  ..   end_month = col_double(),
  ..   end_year = col_double(),
  ..   name_of_exposition = col_character(),
  ..   country = col_character(),
  ..   city = col_character(),
  ..   category = col_character(),
  ..   theme = col_character(),
  ..   notables = col_character(),
  ..   visitors = col_double(),
  ..   cost = col_double(),
  ..   area = col_double(),
  ..   attending_countries = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
skim(worlds_fairs)
```

Name	worlds_fairs
Number of rows	70
Number of columns	14
_______________________	
Column type frequency:	
character	6
numeric	8
________________________	
Group variables	None

Data summary

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
name_of_exposition	0	1.00	5	101	0	67	0
country	0	1.00	5	26	0	24	0
city	0	1.00	4	14	0	51	0
category	0	1.00	10	16	0	2	0
theme	0	1.00	5	77	0	68	0
notables	11	0.84	3	83	0	58	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	his
start_month	0	1.00	5.41	1.88	2.00	4.00	5.0	6.00	12.00	
start_year	0	1.00	1947.19	44.26	1851.00	1911.50	1953.5	1983.50	2021.00	
end_month	0	1.00	9.16	2.38	1.00	9.00	10.0	11.00	12.00	
end_year	0	1.00	1947.29	44.26	1851.00	1911.50	1953.5	1983.50	2022.00	
visitors	14	0.80	16.64	16.59	0.80	5.64	10.0	21.63	73.08	
cost	34	0.51	559.94	998.43	2.00	26.75	125.5	448.25	4200.00	
area	5	0.93	105.82	116.99	0.08	30.00	70.0	120.00	523.00	
attending_countries	5	0.93	49.22	45.04	8.00	23.00	33.0	55.00	192.00	

```
# View the first few rows of the dataset
head(worlds_fairs)
```

start_month	start_year	end_month	end_year
<dbl>	<dbl>	<dbl>	<dbl>
4	1851	10	1851
5	1855	11	1855
5	1862	11	1862
4	1867	11	1867
5	1873	10	1873
5	1876	11	1876

6 rows | 1-4 of 14 columns

```
#tidytuesdayR::use_tidytemplate()
```

Missing Values

Notables: 11 missing values (about 15.7% of the rows). Visitors: 14 missing values (about 20% of the rows). Cost: 34 missing values (about 48.6% of the rows). Area: 5 missing values (about 7.1% of the rows). Attending Countries: 5 missing values (about 7.1% of the rows). Most of the cost fetched from https://jdpecon.com/expo/expolist.html

Clean the data

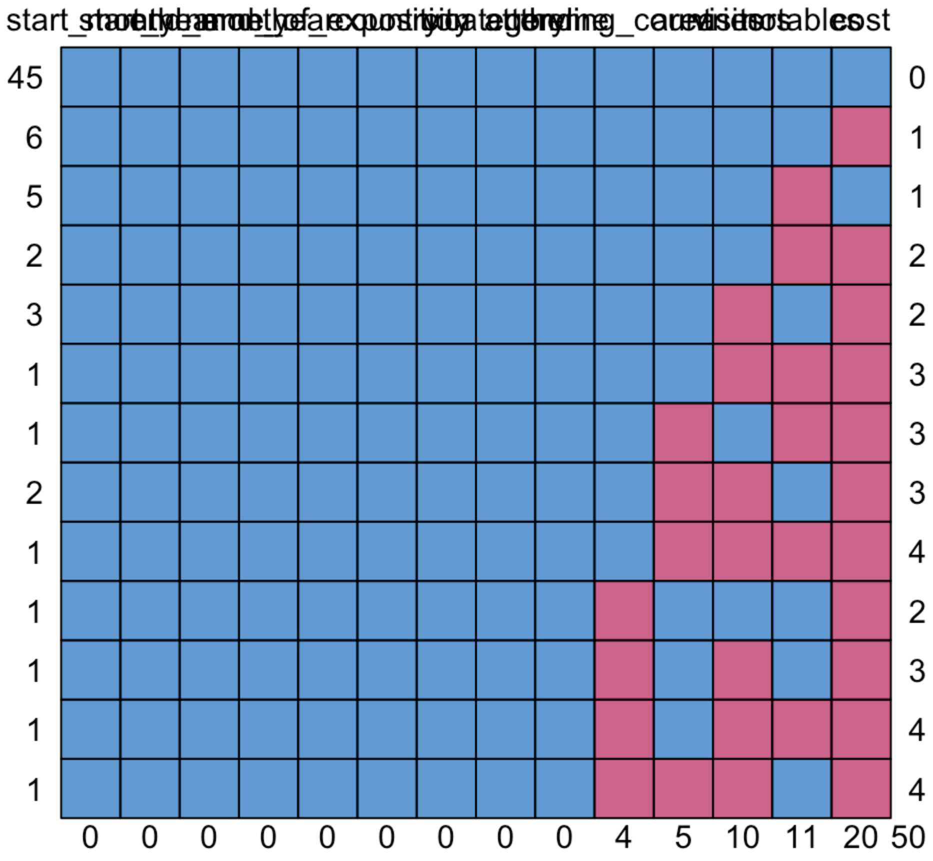
```
#selected_data <- worlds_fairs %>% dplyr::select("country", "start_year", "city",
#selected_data

# add missing values
# https://jdpecon.com/expo/expolist.html
worlds_fairs[70, "cost"] <- 14600 # Dubai 2022
worlds_fairs[69, "cost"] <- 3000 # Kazakhstan 2017
```

```
worlds_fairs[68, "cost"] <- 1640 # Italy 2015
worlds_fairs[65, "cost"] <- 1070 # Spain 2008
worlds_fairs[64, "cost"] <- 3130 # Japan 2005
worlds_fairs[62, "cost"] <- 2952 # Portugal 1998 | Estimated $1753 to $2398 milli
worlds_fairs[61, "cost"] <- 2800 # South Korea 1993
worlds_fairs[60, "cost"] <- 568.75 # Italy 1992 $487.5 to $650 million according
worlds_fairs[59, "cost"] <- 2332 # Spain 1992
worlds_fairs[54, "cost"] <- 763.7 # Japan 1985
worlds_fairs[50, "cost"] <- 304 # Japan 1975
worlds_fairs[12, "cost"] <- 1.14 # Belgium 1897


# Corrections
# https://jdpecon.com/expo/expolist.html
worlds_fairs[1, "cost"] <- 1.67 # London 1851
worlds_fairs[2, "cost"] <- 5.7 # France 1855
worlds_fairs[3, "cost"] <- 2.3 # London 1862
worlds_fairs[4, "cost"] <- 4.5 # Paris 1867
worlds_fairs[5, "cost"] <- 12 # Vienna 1873
worlds_fairs[6, "cost"] <- 9 # Philadelphia 1876
# worlds_fairs[7, "cost"] <- 11 # France 1878
worlds_fairs[8, "cost"] <- 1.56 # Melbourne 1880
worlds_fairs[9, "cost"] <- 1.7 # Spain 1881
worlds_fairs[10, "cost"] <- 8.01 # France 1889 41.5 million French francs would b
worlds_fairs[11, "cost"] <- 27.3 # Chicago 1893
worlds_fairs[12, "cost"] <- 1.14 # Belgium  1897
worlds_fairs[13, "cost"] <- 18.75 # Paris 1900
worlds_fairs[14, "cost"] <- 26.5 # St Louis 1904
worlds_fairs[15, "cost"] <- 2.89 # Belgium 1905
worlds_fairs[16, "cost"] <- 2.6 # Milan 1906
worlds_fairs[17, "cost"] <- 3.55 # Belgium  1910
# worlds_fairs[18, "cost"] <- NA # Italy    1911
worlds_fairs[19, "cost"] <- 3.3 # Belgium   1913
worlds_fairs[20, "cost"] <- 24.7 # San Francisco    1915
# worlds_fairs[21, "cost"] <- 25 # Spain    1929
worlds_fairs[22, "cost"] <- 31 # Chicago 1933
worlds_fairs[23, "cost"] <- 22.4 # Belgium 1935
# worlds_fairs[24, "cost"] <- NA # Sweden   1936
worlds_fairs[25, "cost"] <- 57.25 # Paris 1937
# worlds_fairs[26, "visitors"] <- NA # Finland  1938
worlds_fairs[27, "cost"] <- 160 # New York City 1939
worlds_fairs[27, "attending_countries"] <- 52 # New York City 1939

# https://en.worldfairs.info/expodonnees.php?expo_id=32
worlds_fairs[28, "area"] <- 80 # Belgium    1939
worlds_fairs[28, "visitors"] <- 1.7 # Belgium   1939
worlds_fairs[28, "cost"] <- 58.51 # Belgium 1939 | 150 million Belgian francs wou

# worlds_fairs[29, "cost"] <- NA # France   1947
# worlds_fairs[30, "cost"] <- NA # Sweden   1949
# worlds_fairs[31, "cost"] <- NA # France   1949
# https://www.bie-paris.org/site/en/latest/blog/entry/cultivating-tourism-in-hait
worlds_fairs[32, "cost"] <- 4 # Expo Port-au-Prince 1949
worlds_fairs[32, "visitors"] <- 0.07 # Expo Port-au-Prince 1949
# worlds_fairs[33, "cost"] <- NA # France   1951
# worlds_fairs[34, "cost"] <- NA # Italy    1953
worlds_fairs[35, "visitors"] <- 0.6 # Expo 1953 Jerusalem
# worlds_fairs[36, "cost"] <- NA # Italy    1954
worlds_fairs[37, "visitors"] <- 0.12 # Italy    1955
# worlds_fairs[38, "cost"] <- NA # Sweden   1955
# worlds_fairs[39, "cost"] <- NA # Israel   1956
# worlds_fairs[40, "cost"] <- NA # West Germany 1957
worlds_fairs[41, "cost"] <- 43.4 # Belgium  1958
# worlds_fairs[42, "cost"] <- NA # Italy    1961
worlds_fairs[43, "cost"] <- 22.8 # Seattle 1962
# worlds_fairs[44, "cost"] <- NA # West Germany 1965
worlds_fairs[45, "cost"] <- 384.7 # 1967 Canada 415.920 CAD million would be appr
# worlds_fairs[46, "cost"] <- 156.00 # San Antonio 1968
worlds_fairs[47, "cost"] <- 247.54 # Japan  1970
# worlds_fairs[48, "cost"] <- NA # Hungary  1971
worlds_fairs[49, "cost"] <- 78.40 # Spokane 1974
worlds_fairs[50, "cost"] <- 155.2 # Japan   1975
# worlds_fairs[51, "cost"] <- NA # Bulgaria 1981
worlds_fairs[52, "cost"] <- 111 # Knoxville 1982
worlds_fairs[53, "cost"] <- 442.5 # New Orleans 1984
worlds_fairs[54, "cost"] <- 763.7 # Japan   1985
# worlds_fairs[55, "cost"] <- NA # Bulgaria 1985
worlds_fairs[56, "cost"] <- 609.52 # Canada 1986
worlds_fairs[57, "cost"] <- 642.2 # Australia   1988
# worlds_fairs[58, "cost"] <- NA # Bulgaria 1991
# worlds_fairs[59, "cost"] <- 2332 # Spain 1992
# worlds_fairs[60, "cost"] <- 568.75 # Italy 1992
# worlds_fairs[61, "cost"] <- 2800 # South Korea 1993
# worlds_fairs[62, "cost"] <- 2952 # Portugal 1998
worlds_fairs[63, "cost"] <- 2238 # Germany  2000
# worlds_fairs[64, "cost"] <- 3130 # Japan 2005
# worlds_fairs[65, "cost"] <- 1070 # Spain 2008
worlds_fairs[66, "cost"] <- 8800  # China   2010

#https://www.exhibitionworld.co.uk/crossing-oceans-expo-2012-yeosu-korea
worlds_fairs[67, "cost"] <- 2000 # South Korea 2012

worlds_fairs[68, "cost"] <- 1640 # Italy 2015
worlds_fairs[69, "cost"] <- 3000 # Kazakhstan 2017
worlds_fairs[70, "cost"] <- 14600 # Dubai 2022



# we still have 21 null values for cost
# visitors  10
# area  4 and attending_countries   5
```

To address the missing values, we will use an imputation method to help fill in the remaining missing values using a more sophisticated technique than the median. Using the mice package (https://amices.org/mice/), we will consider the relationships between variables to fill in the missing data.

Instead of using KNN, we decided to try PMM (Predictive Mean Matching). This matches each missing value with the nearest observed values and imputes from those values. Effective for maintaining the distribution of the data. PMM does multiple imputations by default when you set m > 1, creating multiple versions of the dataset with imputed values, reflecting the uncertainty of the missing data.

```
suppressMessages(library(mice))

#  Show the missing data pattern
md.pattern(worlds_fairs) # Blue is observed, red is missing
```



```
   start_month start_year end_month end_year name_of_exposition country city
45           1          1         1        1                  1       1    1
6            1          1         1        1                  1       1    1
5            1          1         1        1                  1       1    1
2            1          1         1        1                  1       1    1
3            1          1         1        1                  1       1    1
```

```
1            1          1         1        1                  1       1    1
1            1          1         1        1                  1       1    1
2            1          1         1        1                  1       1    1
1            1          1         1        1                  1       1    1
1            1          1         1        1                  1       1    1
1            1          1         1        1                  1       1    1
1            1          1         1        1                  1       1    1
1            1          1         1        1                  1       1    1
             0          0         0        0                  0       0    0
   category theme attending_countries area visitors notables cost
45        1     1                   1    1        1        1    1  0
6         1     1                   1    1        1        1    0  1
5         1     1                   1    1        1        0    1  1
2         1     1                   1    1        1        0    0  2
3         1     1                   1    1        0        1    0  2
1         1     1                   1    1        0        0    0  3
1         1     1                   1    0        1        0    0  3
2         1     1                   1    0        0        1    0  3
1         1     1                   1    0        0        0    0  4
1         1     1                   0    1        1        1    0  2
1         1     1                   0    1        0        1    0  3
1         1     1                   0    1        0        0    0  4
1         1     1                   0    0        0        1    0  4
          0     0                   4    5       10       11   20 50
```

```r
imputed_data <- mice(worlds_fairs, method = 'pmm', m = 5, maxit = 50)
```

```
 iter imp variable
  1   1  visitors  cost  area  attending_countries
  1   2  visitors  cost  area  attending_countries
  1   3  visitors  cost  area  attending_countries
  1   4  visitors  cost  area  attending_countries
  1   5  visitors  cost  area  attending_countries
  2   1  visitors  cost  area  attending_countries
  2   2  visitors  cost  area  attending_countries
  2   3  visitors  cost  area  attending_countries
  2   4  visitors  cost  area  attending_countries
  2   5  visitors  cost  area  attending_countries
  3   1  visitors  cost  area  attending_countries
  3   2  visitors  cost  area  attending_countries
  3   3  visitors  cost  area  attending_countries
  3   4  visitors  cost  area  attending_countries
  3   5  visitors  cost  area  attending_countries
  4   1  visitors  cost  area  attending_countries
  4   2  visitors  cost  area  attending_countries
  4   3  visitors  cost  area  attending_countries
  4   4  visitors  cost  area  attending_countries
  4   5  visitors  cost  area  attending_countries
  5   1  visitors  cost  area  attending_countries
```

```
5   2  visitors  cost  area  attending_countries
5   3  visitors  cost  area  attending_countries
5   4  visitors  cost  area  attending_countries
5   5  visitors  cost  area  attending_countries
6   1  visitors  cost  area  attending_countries
6   2  visitors  cost  area  attending_countries
6   3  visitors  cost  area  attending_countries
6   4  visitors  cost  area  attending_countries
6   5  visitors  cost  area  attending_countries
7   1  visitors  cost  area  attending_countries
7   2  visitors  cost  area  attending_countries
7   3  visitors  cost  area  attending_countries
7   4  visitors  cost  area  attending_countries
7   5  visitors  cost  area  attending_countries
8   1  visitors  cost  area  attending_countries
8   2  visitors  cost  area  attending_countries
8   3  visitors  cost  area  attending_countries
8   4  visitors  cost  area  attending_countries
8   5  visitors  cost  area  attending_countries
9   1  visitors  cost  area  attending_countries
9   2  visitors  cost  area  attending_countries
9   3  visitors  cost  area  attending_countries
9   4  visitors  cost  area  attending_countries
9   5  visitors  cost  area  attending_countries
10   1  visitors  cost  area  attending_countries
10   2  visitors  cost  area  attending_countries
10   3  visitors  cost  area  attending_countries
10   4  visitors  cost  area  attending_countries
10   5  visitors  cost  area  attending_countries
11   1  visitors  cost  area  attending_countries
11   2  visitors  cost  area  attending_countries
11   3  visitors  cost  area  attending_countries
11   4  visitors  cost  area  attending_countries
11   5  visitors  cost  area  attending_countries
12   1  visitors  cost  area  attending_countries
12   2  visitors  cost  area  attending_countries
12   3  visitors  cost  area  attending_countries
12   4  visitors  cost  area  attending_countries
12   5  visitors  cost  area  attending_countries
13   1  visitors  cost  area  attending_countries
13   2  visitors  cost  area  attending_countries
13   3  visitors  cost  area  attending_countries
13   4  visitors  cost  area  attending_countries
13   5  visitors  cost  area  attending_countries
14   1  visitors  cost  area  attending_countries
14   2  visitors  cost  area  attending_countries
14   3  visitors  cost  area  attending_countries
14   4  visitors  cost  area  attending_countries
14   5  visitors  cost  area  attending_countries
15   1  visitors  cost  area  attending_countries
15   2  visitors  cost  area  attending_countries
```

```
15   3  visitors  cost  area  attending_countries
15   4  visitors  cost  area  attending_countries
15   5  visitors  cost  area  attending_countries
16   1  visitors  cost  area  attending_countries
16   2  visitors  cost  area  attending_countries
16   3  visitors  cost  area  attending_countries
16   4  visitors  cost  area  attending_countries
16   5  visitors  cost  area  attending_countries
17   1  visitors  cost  area  attending_countries
17   2  visitors  cost  area  attending_countries
17   3  visitors  cost  area  attending_countries
17   4  visitors  cost  area  attending_countries
17   5  visitors  cost  area  attending_countries
18   1  visitors  cost  area  attending_countries
18   2  visitors  cost  area  attending_countries
18   3  visitors  cost  area  attending_countries
18   4  visitors  cost  area  attending_countries
18   5  visitors  cost  area  attending_countries
19   1  visitors  cost  area  attending_countries
19   2  visitors  cost  area  attending_countries
19   3  visitors  cost  area  attending_countries
19   4  visitors  cost  area  attending_countries
19   5  visitors  cost  area  attending_countries
20   1  visitors  cost  area  attending_countries
20   2  visitors  cost  area  attending_countries
20   3  visitors  cost  area  attending_countries
20   4  visitors  cost  area  attending_countries
20   5  visitors  cost  area  attending_countries
21   1  visitors  cost  area  attending_countries
21   2  visitors  cost  area  attending_countries
21   3  visitors  cost  area  attending_countries
21   4  visitors  cost  area  attending_countries
21   5  visitors  cost  area  attending_countries
22   1  visitors  cost  area  attending_countries
22   2  visitors  cost  area  attending_countries
22   3  visitors  cost  area  attending_countries
22   4  visitors  cost  area  attending_countries
22   5  visitors  cost  area  attending_countries
23   1  visitors  cost  area  attending_countries
23   2  visitors  cost  area  attending_countries
23   3  visitors  cost  area  attending_countries
23   4  visitors  cost  area  attending_countries
23   5  visitors  cost  area  attending_countries
24   1  visitors  cost  area  attending_countries
24   2  visitors  cost  area  attending_countries
24   3  visitors  cost  area  attending_countries
24   4  visitors  cost  area  attending_countries
24   5  visitors  cost  area  attending_countries
25   1  visitors  cost  area  attending_countries
25   2  visitors  cost  area  attending_countries
25   3  visitors  cost  area  attending_countries
```

```
25   4  visitors  cost  area  attending_countries
25   5  visitors  cost  area  attending_countries
26   1  visitors  cost  area  attending_countries
26   2  visitors  cost  area  attending_countries
26   3  visitors  cost  area  attending_countries
26   4  visitors  cost  area  attending_countries
26   5  visitors  cost  area  attending_countries
27   1  visitors  cost  area  attending_countries
27   2  visitors  cost  area  attending_countries
27   3  visitors  cost  area  attending_countries
27   4  visitors  cost  area  attending_countries
27   5  visitors  cost  area  attending_countries
28   1  visitors  cost  area  attending_countries
28   2  visitors  cost  area  attending_countries
28   3  visitors  cost  area  attending_countries
28   4  visitors  cost  area  attending_countries
28   5  visitors  cost  area  attending_countries
29   1  visitors  cost  area  attending_countries
29   2  visitors  cost  area  attending_countries
29   3  visitors  cost  area  attending_countries
29   4  visitors  cost  area  attending_countries
29   5  visitors  cost  area  attending_countries
30   1  visitors  cost  area  attending_countries
30   2  visitors  cost  area  attending_countries
30   3  visitors  cost  area  attending_countries
30   4  visitors  cost  area  attending_countries
30   5  visitors  cost  area  attending_countries
31   1  visitors  cost  area  attending_countries
31   2  visitors  cost  area  attending_countries
31   3  visitors  cost  area  attending_countries
31   4  visitors  cost  area  attending_countries
31   5  visitors  cost  area  attending_countries
32   1  visitors  cost  area  attending_countries
32   2  visitors  cost  area  attending_countries
32   3  visitors  cost  area  attending_countries
32   4  visitors  cost  area  attending_countries
32   5  visitors  cost  area  attending_countries
33   1  visitors  cost  area  attending_countries
33   2  visitors  cost  area  attending_countries
33   3  visitors  cost  area  attending_countries
33   4  visitors  cost  area  attending_countries
33   5  visitors  cost  area  attending_countries
34   1  visitors  cost  area  attending_countries
34   2  visitors  cost  area  attending_countries
34   3  visitors  cost  area  attending_countries
34   4  visitors  cost  area  attending_countries
34   5  visitors  cost  area  attending_countries
35   1  visitors  cost  area  attending_countries
35   2  visitors  cost  area  attending_countries
35   3  visitors  cost  area  attending_countries
35   4  visitors  cost  area  attending_countries
```

```
35   5  visitors  cost  area  attending_countries
36   1  visitors  cost  area  attending_countries
36   2  visitors  cost  area  attending_countries
36   3  visitors  cost  area  attending_countries
36   4  visitors  cost  area  attending_countries
36   5  visitors  cost  area  attending_countries
37   1  visitors  cost  area  attending_countries
37   2  visitors  cost  area  attending_countries
37   3  visitors  cost  area  attending_countries
37   4  visitors  cost  area  attending_countries
37   5  visitors  cost  area  attending_countries
38   1  visitors  cost  area  attending_countries
38   2  visitors  cost  area  attending_countries
38   3  visitors  cost  area  attending_countries
38   4  visitors  cost  area  attending_countries
38   5  visitors  cost  area  attending_countries
39   1  visitors  cost  area  attending_countries
39   2  visitors  cost  area  attending_countries
39   3  visitors  cost  area  attending_countries
39   4  visitors  cost  area  attending_countries
39   5  visitors  cost  area  attending_countries
40   1  visitors  cost  area  attending_countries
40   2  visitors  cost  area  attending_countries
40   3  visitors  cost  area  attending_countries
40   4  visitors  cost  area  attending_countries
40   5  visitors  cost  area  attending_countries
41   1  visitors  cost  area  attending_countries
41   2  visitors  cost  area  attending_countries
41   3  visitors  cost  area  attending_countries
41   4  visitors  cost  area  attending_countries
41   5  visitors  cost  area  attending_countries
42   1  visitors  cost  area  attending_countries
42   2  visitors  cost  area  attending_countries
42   3  visitors  cost  area  attending_countries
42   4  visitors  cost  area  attending_countries
42   5  visitors  cost  area  attending_countries
43   1  visitors  cost  area  attending_countries
43   2  visitors  cost  area  attending_countries
43   3  visitors  cost  area  attending_countries
43   4  visitors  cost  area  attending_countries
43   5  visitors  cost  area  attending_countries
44   1  visitors  cost  area  attending_countries
44   2  visitors  cost  area  attending_countries
44   3  visitors  cost  area  attending_countries
44   4  visitors  cost  area  attending_countries
44   5  visitors  cost  area  attending_countries
45   1  visitors  cost  area  attending_countries
45   2  visitors  cost  area  attending_countries
45   3  visitors  cost  area  attending_countries
45   4  visitors  cost  area  attending_countries
45   5  visitors  cost  area  attending_countries
```

```
 46   1  visitors  cost  area  attending_countries
 46   2  visitors  cost  area  attending_countries
 46   3  visitors  cost  area  attending_countries
 46   4  visitors  cost  area  attending_countries
 46   5  visitors  cost  area  attending_countries
 47   1  visitors  cost  area  attending_countries
 47   2  visitors  cost  area  attending_countries
 47   3  visitors  cost  area  attending_countries
 47   4  visitors  cost  area  attending_countries
 47   5  visitors  cost  area  attending_countries
 48   1  visitors  cost  area  attending_countries
 48   2  visitors  cost  area  attending_countries
 48   3  visitors  cost  area  attending_countries
 48   4  visitors  cost  area  attending_countries
 48   5  visitors  cost  area  attending_countries
 49   1  visitors  cost  area  attending_countries
 49   2  visitors  cost  area  attending_countries
 49   3  visitors  cost  area  attending_countries
 49   4  visitors  cost  area  attending_countries
 49   5  visitors  cost  area  attending_countries
 50   1  visitors  cost  area  attending_countries
 50   2  visitors  cost  area  attending_countries
 50   3  visitors  cost  area  attending_countries
 50   4  visitors  cost  area  attending_countries
 50   5  visitors  cost  area  attending_countries
```

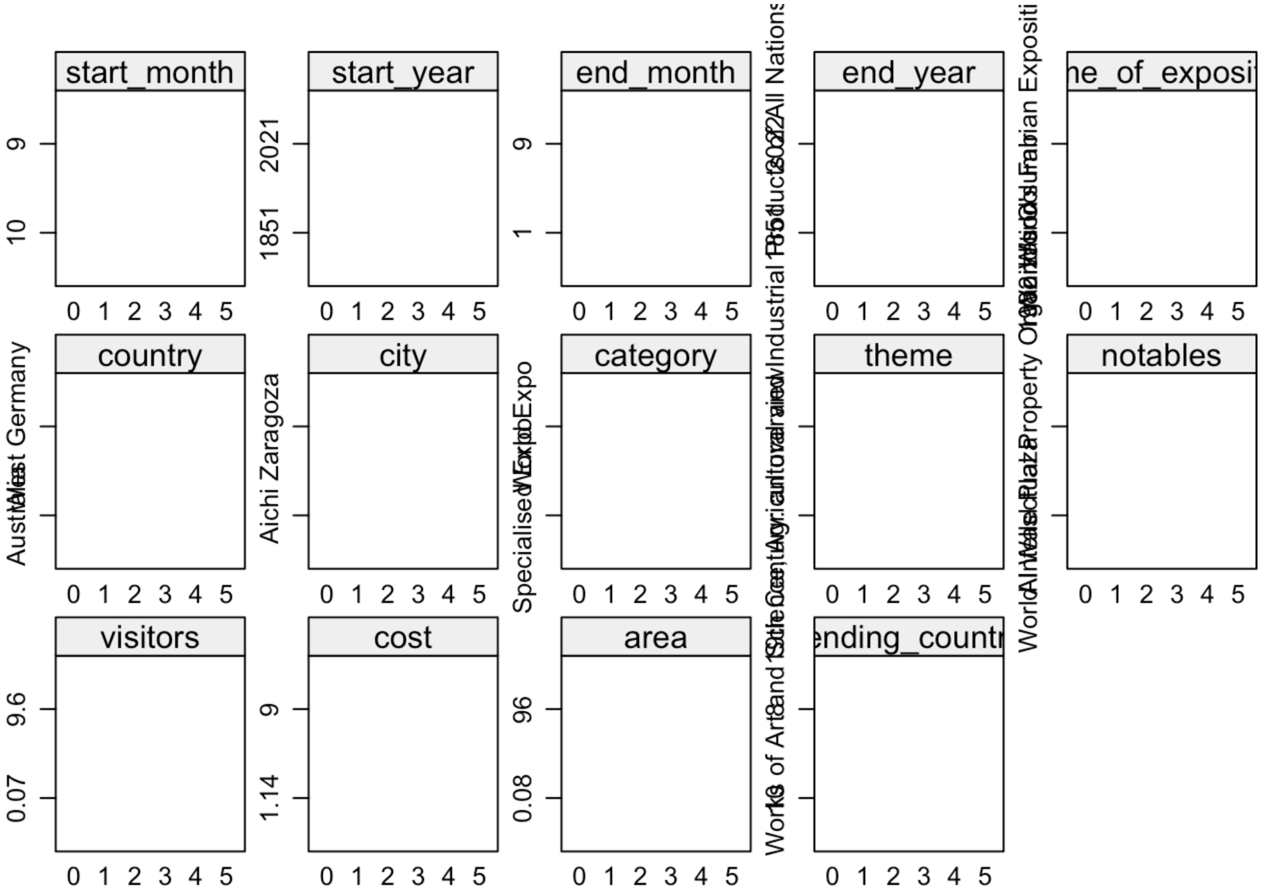
```r
#imputed_data <- mice(worlds_fairs, method = 'norm', m = 5, maxit = 50) # Gives r
#imputed_data <- mice(worlds_fairs, method = 'norm.predict', m = 5, maxit = 50) #
#imputed_data <- mice(worlds_fairs, method = 'mean', m = 5, maxit = 50) # Gives r
#imputed_data <- mice(worlds_fairs, method = 'pmm', m = 5, maxit = 50, seed = 184
# we are using the Predictive Mean Matching method and we generated 5 different v

# Inspect the imputed data.
#summary(imputed_data)

# Matrix of the imputed values for the cost variable
imputed_data$imp$cost
```

	1 <dbl>	2 <dbl>	3 <dbl>	4 <dbl>	5 <dbl>
18	442.50	1.14	3.55	160.00	22.80
24	1.70	111.00	3.55	3.55	78.40
26	1.14	442.50	24.70	22.40	1.70
29	3.55	57.25	24.70	2.30	4.50
30	609.52	4.00	12.00	43.40	568.75
31	12.00	3130.00	25.00	12.00	3130.00
33	160.00	5.70	8.01	160.00	5.70
34	58.51	25.00	27.30	642.20	2.89

	1	2	3	4	5
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
35	58.51	1070.00	763.70	8.01	27.30
36	22.40	609.52	8.01	57.25	156.00

1-10 of 20 rows Previous **1** 2 Next

```r
# Visual summary of the imputation process
stripplot(imputed_data)
```



```r
# helps you understand how the imputed values compare to the observed data and wh

# Extract the Dataset
worlds_fairs_imputed <- complete(imputed_data,3)

# Compare Before and After
summary(worlds_fairs$cost)
```

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
   1.140     8.258    50.325   986.281   634.030 14600.000        20 
```

```r
summary(worlds_fairs_imputed$cost)
```

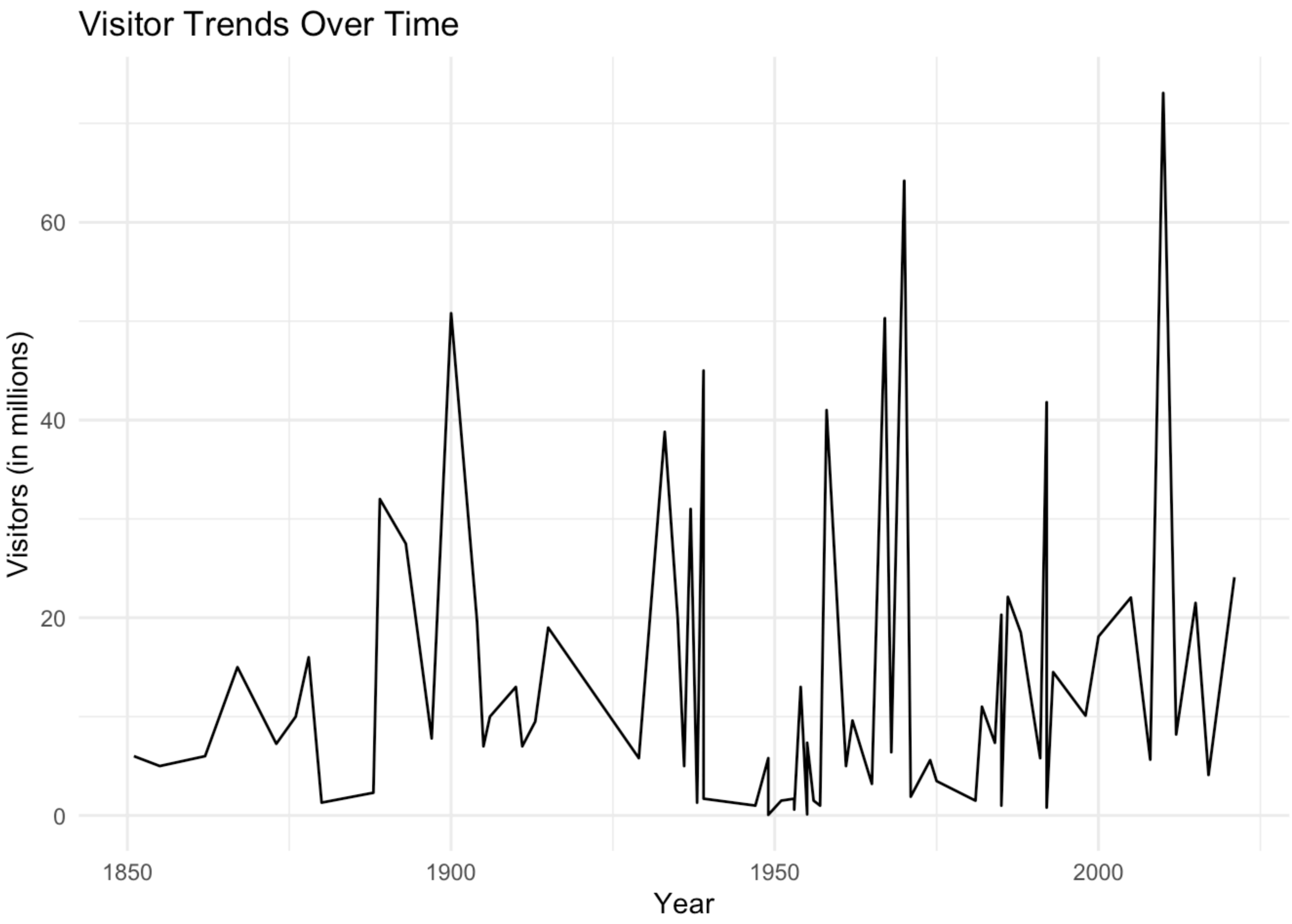
```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
   1.14     8.01    25.00   758.60   537.19 14600.00 
```

```r
#skim(worlds_fairs)
#skim(worlds_fairs_imputed)

imputed_fairs <- worlds_fairs_imputed %>% dplyr::select("start_year","country", "
imputed_fairs
```

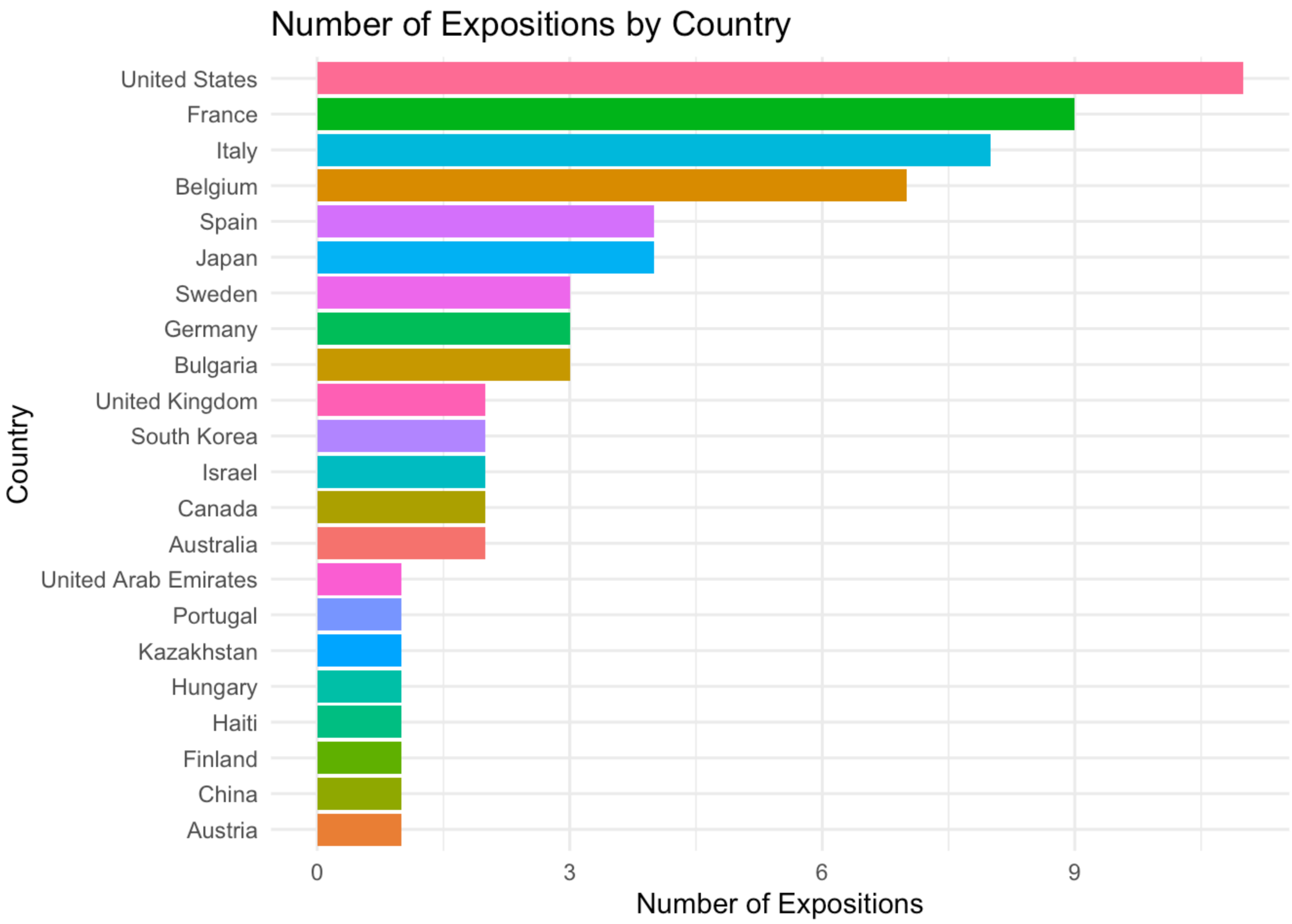
start_year	country	city	cost
<dbl>	<chr>	<chr>	<dbl>
1851	United Kingdom	London	1.67
1855	France	Paris	5.70
1862	United Kingdom	London	2.30
1867	France	Paris	4.50
1873	Austria-Hungary	Vienna	12.00
1876	United States	Philadelphia	9.00
1878	France	Paris	11.00
1880	Colony of Victoria	Melbourne	1.56
1888	Spain	Barcelona	1.70
1889	France	Paris	8.01

1-10 of 70 rows Previous **1** 2 3 4 5 6 7 Next

```r
library(ggplot2)

ggplot(worlds_fairs_imputed, aes(x = start_year, y = visitors)) +
  geom_line() +
  labs(title = "Visitor Trends Over Time",
       x = "Year",
       y = "Visitors (in millions)") +
  theme_minimal()
```



```r
# Rename counties to
worlds_fairs_imputed <- worlds_fairs_imputed %>%
  mutate(country = recode(country, "Colony of Victoria" = "Australia", "West Germ


ggplot(worlds_fairs_imputed %>%
         count(country) %>%
         arrange(n),
       aes(x = reorder(country, n), y = n, fill = country)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Number of Expositions by Country",
       x = "Country",
       y = "Number of Expositions") +
  theme_minimal() +
  theme(legend.position = "none")
```

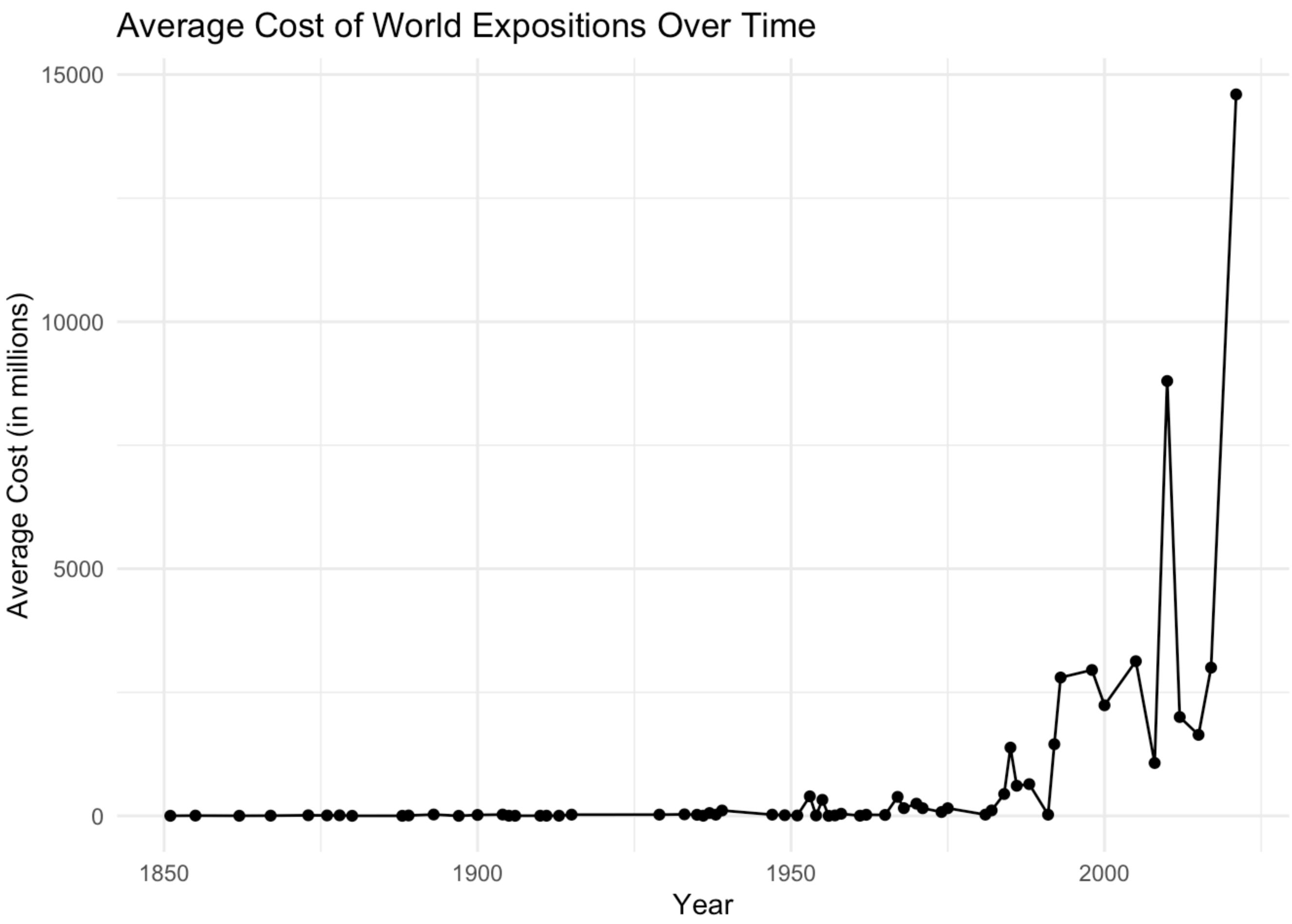


```r
cost_expositions_df <- worlds_fairs_imputed[, c("cost", "start_year")]

cost_expositions_df
```

cost	start_year
<dbl>	<dbl>
1.67	1851
5.70	1855
2.30	1862
4.50	1867
12.00	1873
9.00	1876
11.00	1878
1.56	1880
1.70	1888
8.01	1889

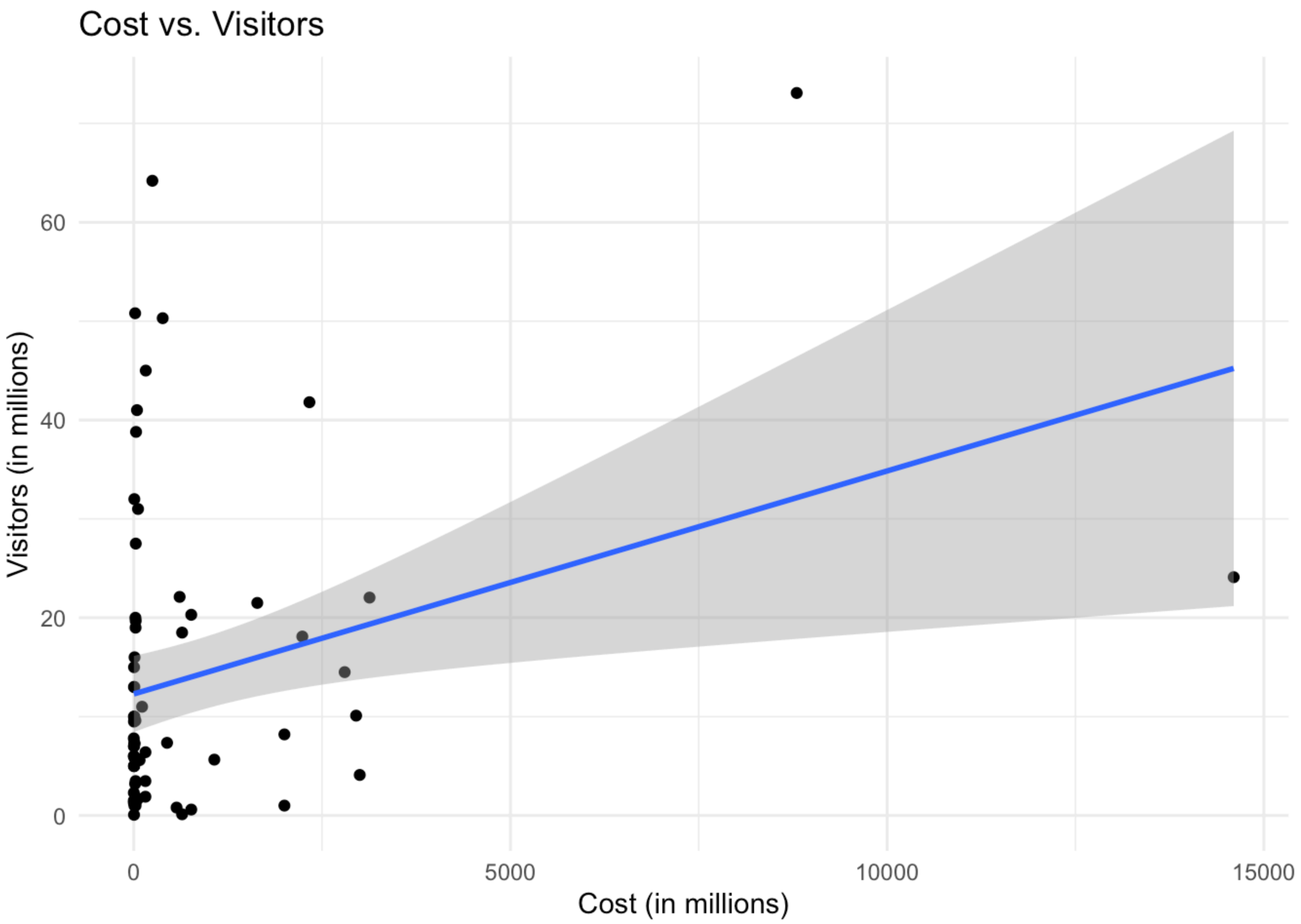
1-10 of 70 rows Previous **1** 2 3 4 5 6 7 Next

```r
worlds_fairs_imputed %>%
  group_by(start_year) %>%
  summarise(avg_cost = mean(cost, na.rm = TRUE)) %>%
  ggplot(aes(x = start_year, y = avg_cost)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Cost of World Expositions Over Time",
       x = "Year",
       y = "Average Cost (in millions)") +
  theme_minimal()
```



```r
#sessioninfo::session_info(include_base = TRUE)
```

```r
ggplot(worlds_fairs_imputed, aes(x = cost, y = visitors)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Cost vs. Visitors",
       x = "Cost (in millions)",
       y = "Visitors (in millions)") +
  theme_minimal()
```



```r
cost_visitors_df <- worlds_fairs_imputed[, c("cost", "visitors")]

cost_visitors_df
```

cost	visitors
<dbl>	<dbl>
1.67	6.000
5.70	5.000
2.30	6.000
4.50	15.000
12.00	7.250
9.00	10.000
11.00	16.000
1.56	1.300
1.70	2.300
8.01	32.000

1-10 of 70 rows Previous **1** 2 3 4 5 6 7 Next