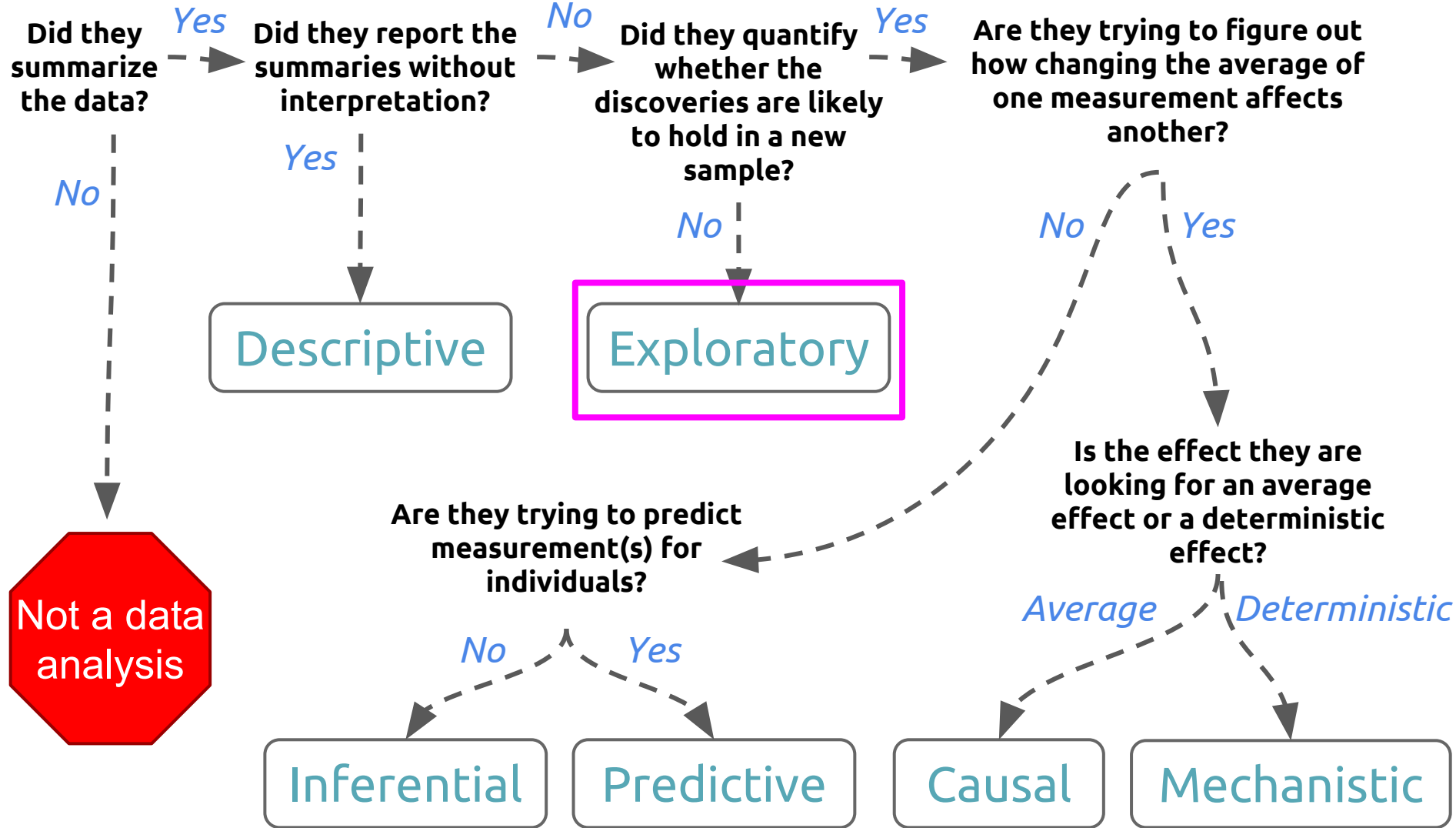


Exploratory Analysis



Data Analysis



Exploratory analysis helps us ...

- understand data properties such as nonlinear relationships, the existence of missing values, the existence of outliers, etc.
- find patterns in data such as associations, group differences, confounders, etc.
- suggest modeling strategies such as linear vs. nonlinear models, transformation
- debug" analyses
- communicate results



The general principles of exploratory analysis:

- Look for missing values
- Look for outlier values
- Use plots to explore relationships
- Use tables to explore relationships
- If necessary, transform variables



12
★ Featured Dataset

120 years of Olympic history: athletes and results

basic bio data on athletes and medal results from Athens 1896 to Rio 2016

Randi H Griffin • last updated a month ago

Data Overview Kernels Discussion Activity

Download (5 MB)

New Kernel

Data

API

?

Download All

✕

Data Sources

athlete_events.csv 271k x 15

noc_regions.csv 230 x 3

About this file

Edit

Help us describe this file

Columns

Edit

ID
A Name
A Sex
Age
A Height
A Weight
A Team
A NOC
A Games


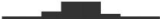



<https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>

```
library(readr)
# Loading data as a data frame
df <- read_csv("athlete_events.csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_integer(),
##   Name = col_character(),
##   Sex = col_character(),
##   Age = col_integer(),
##   Height = col_integer(),
##   Weight = col_double(),
##   Team = col_character(),
##   NOC = col_character(),
##   Games = col_character(),
##   Year = col_integer(),
##   Season = col_character(),
##   City = col_character(),
##   Sport = col_character(),
##   Event = col_character(),
##   Medal = col_character()
## )
```



```
library(skimr)
# summary of the dataset
skim(df)
```

```
## Skim summary statistics
##  n obs: 271116
##  n variables: 15
##
## — Variable type:character —————
##  variable missing complete      n min max empty n_unique
##    City         0   271116 271116   4  22    0         42
##    Event         0   271116 271116  15  85    0        765
##    Games         0   271116 271116  11  11    0         51
##    Medal 231333   39783 271116   4   6    0          3
##    Name          0   271116 271116   2 108    0       134731
##    NOC           0   271116 271116   3   3    0         230
##    Season        0   271116 271116   6   6    0          2
##    Sex           0   271116 271116   1   1    0          2
##    Sport         0   271116 271116   4  25    0         66
##    Team          0   271116 271116   2  47    0        1184
##
## — Variable type:integer —————
##  variable missing complete      n   mean    sd   p0   p25   p50   p75
##    Age      9474   261642 271116   25.56   6.39  10    21    24    28
##    Height   60171  210945 271116  175.34  10.52 127   168   175   183
##    ID        0    271116 271116 68248.95 39022.29   1 34643 68205 1e+05
##    Year      0    271116 271116 1978.38   29.88 1896  1960  1988  2002
##
##  p100    hist
##    97    
##    226    
##    135571 
##    2016    
##
## — Variable type:numeric —————
##  variable missing complete      n mean    sd p0 p25 p50 p75 p100    hist
##    Weight   62875   208241 271116 70.7 14.35 25  60  70  79  214 
```



```
library(skimr)
# summary of the dataset
skim(df)
```

```
## Skim summary statistics
```

```
## n obs: 271116
```

```
## n variables: 15
```

```
##
```

```
## - Variable type:character -
```

##	variable	missing	complete	n	min	max	empty	n_unique
##	City	0	271116	271116	4	22	0	42
##	Event	0	271116	271116	15	85	0	765
##	Games	0	271116	271116	11	11	0	51
##	Medal	231333	39783	271116	4	6	0	3
##	Name	0	271116	271116	2	108	0	134731
##	NOC	0	271116	271116	3	3	0	230
##	Season	0	271116	271116	6	6	0	2
##	Sex	0	271116	271116	1	1	0	2
##	Sport	0	271116	271116	4	25	0	66
##	Team	0	271116	271116	2	47	0	1184

```
##
```

```
## - Variable type:integer -
```

##	variable	missing	complete	n	mean	sd	p0	p25	p50	p75
##	Age	9474	261642	271116	25.56	6.39	10	21	24	28
##	Height	60171	210945	271116	175.34	10.52	127	168	175	183
##	ID	0	271116	271116	68248.95	39022.29	1	34643	68205	1e+05
##	Year	0	271116	271116	1978.38	29.88	1896	1960	1988	2002

```
## p100 hist
```

```
## 97 
```

```
## 226 
```

```
## 135571 
```

```
## 2016 
```

```
##
```

```
## - Variable type:numeric -
```

##	variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
##	Weight	62875	208241	271116	70.7	14.35	25	60	70	79	214	


```
library(skimr)
# summary of the dataset
skim(df)
```

```
## Skim summary statistics
##  n obs: 271116
##  n variables: 15
##
## — Variable type:character —————
##  variable missing complete      n min max empty n_unique
##    City         0   271116 271116    4  22     0        42
##    Event         0   271116 271116   15  85     0       765
##    Games         0   271116 271116   11  11     0        51
##    Medal 231333  39783 271116    4   6     0         3
##    Name          0   271116 271116    2 108     0     134731
##    NOC           0   271116 271116    3   3     0        230
##    Season        0   271116 271116    6   6     0         2
##    Sex           0   271116 271116    1   1     0         2
##    Sport         0   271116 271116    4  25     0        66
##    Team          0   271116 271116    2  47     0     1184
##
## — Variable type:integer —————
##  variable missing complete      n   mean    sd  p0  p25  p50  p75
##    Age  9474  261642 271116   25.56   6.39  10   21   24   28
##    Height 60171 210945 271116  175.34  10.52 127  168  175  183
##    ID      0   271116 271116 68248.95 39022.29  1 34643 68205 1e+05
##    Year    0   271116 271116  1978.38   29.88 1896  1960  1988  2002
##
##  p100  hist
##    97  █
##    226 █
##  135571 █
##    2016 █
##
## — Variable type:numeric —————
##  variable missing complete      n mean    sd p0 p25 p50 p75 p100  hist
##    Weight  62875  208241 271116  70.7 14.35 25  60  70  79  214  █
```

Missing values

```
library(skimr)
# summary of the dataset
skim(df)
```

```
## Skim summary statistics
##  n obs: 271116
##  n variables: 15
##
## — Variable type:character —
##  variable missing complete      n min max empty n_unique
##    City         0   271116 271116    4  22     0         42
##    Event         0   271116 271116   15  85     0        765
##    Games         0   271116 271116   11  11     0         51
##    Medal 231333   39783 271116    4   6     0          3
##    Name          0   271116 271116    2 108     0       134731
##    NOC           0   271116 271116    3   3     0         230
##    Season        0   271116 271116    6   6     0          2
##    Sex           0   271116 271116    1   1     0          2
##    Sport         0   271116 271116    4  25     0         66
##    Team          0   271116 271116    2  47     0        1184
##
## — Variable type:integer —
##  variable missing complete      n   mean    sd  p0  p25  p50  p75
##    Age      9474  261642 271116   25.56   6.39  10   21   24   28
##    Height   60171 210945 271116  175.34  10.52 127  168  175  183
##    ID        0   271116 271116 68248.95 39022.29  1 34643 68205 1e+05
##    Year      0   271116 271116  1978.38   29.88 1896  1960  1988  2002
##    p100      97  135571 271116  2016.00  29.88 1896  1960  1988  2002
##    97         97  135571 271116  2016.00  29.88 1896  1960  1988  2002
##    226        97  135571 271116  2016.00  29.88 1896  1960  1988  2002
##    135571     97  135571 271116  2016.00  29.88 1896  1960  1988  2002
##    2016       97  135571 271116  2016.00  29.88 1896  1960  1988  2002
##
## — Variable type:numeric —
##  variable missing complete      n mean    sd p0 p25 p50 p75 p100  hist
##    Weight   62875  208241 271116  70.7 14.35 25  60  70  79  214  [hist]
```

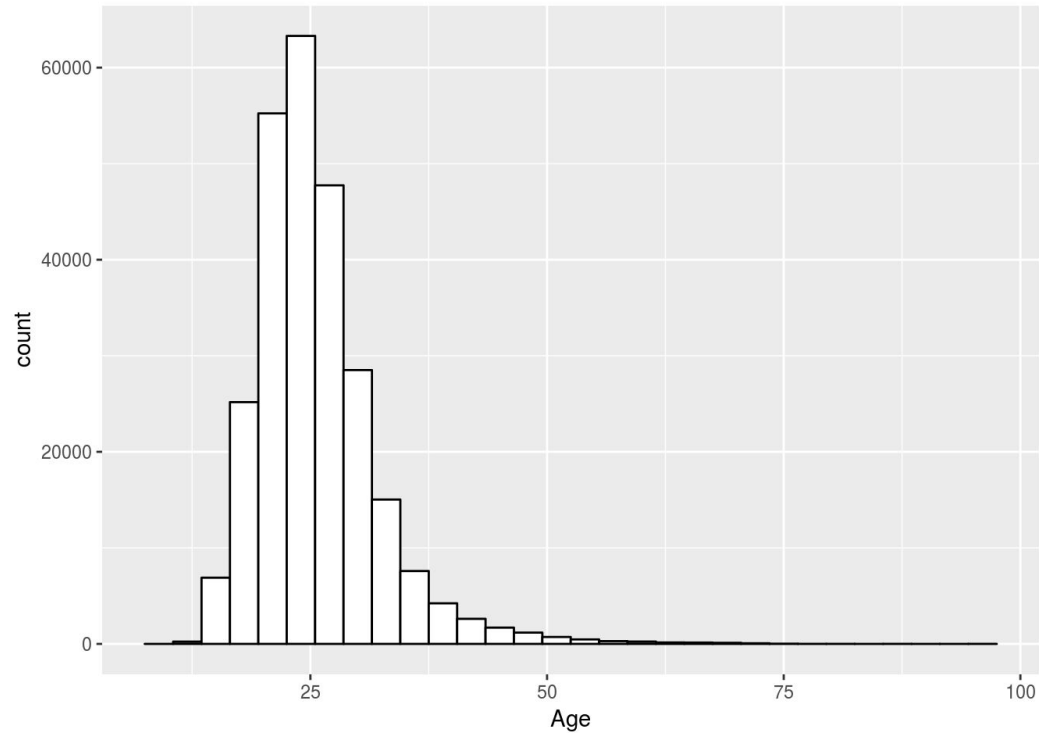
Outlier



```
library(ggplot2)
ggplot(df, aes(x=Age)) +
  geom_histogram(color="black", fill="white")
```

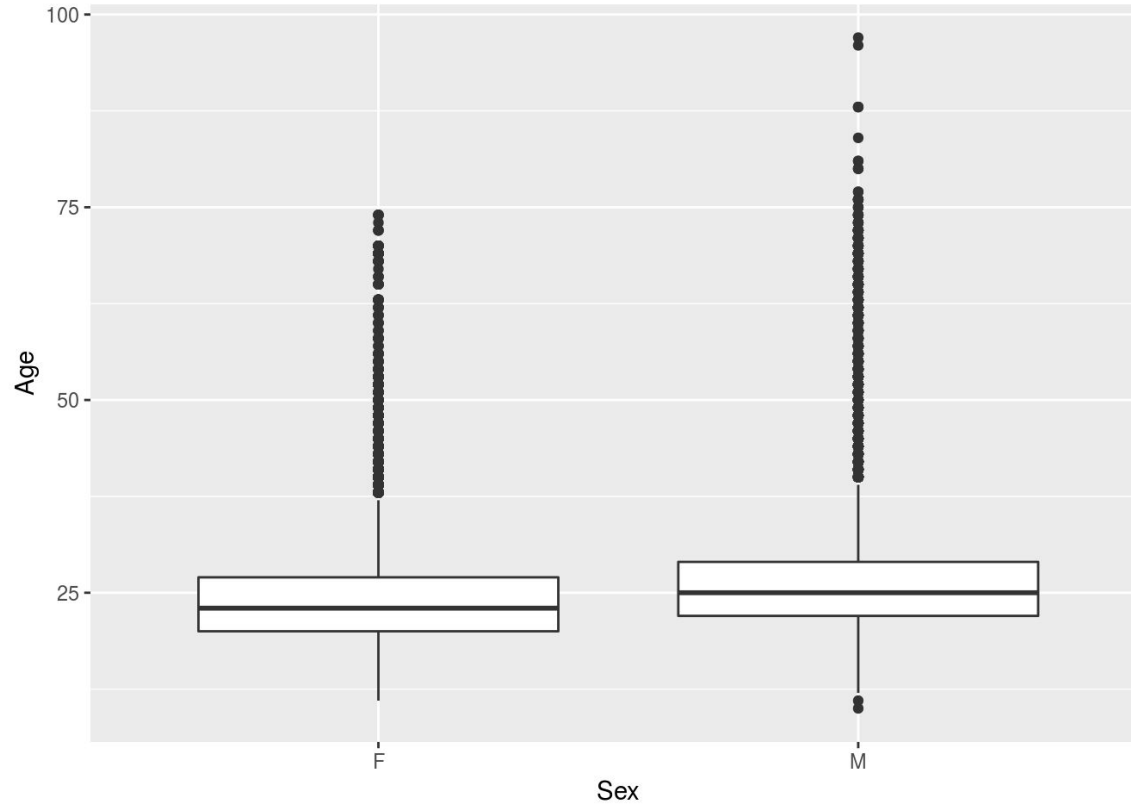
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 9474 rows containing non-finite values (stat_bin).
```



```
ggplot(df, aes(x=Sex, y=Age)) +  
  geom_boxplot()
```

```
## Warning: Removed 9474 rows containing non-finite values (stat_boxplot).
```



```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'
```

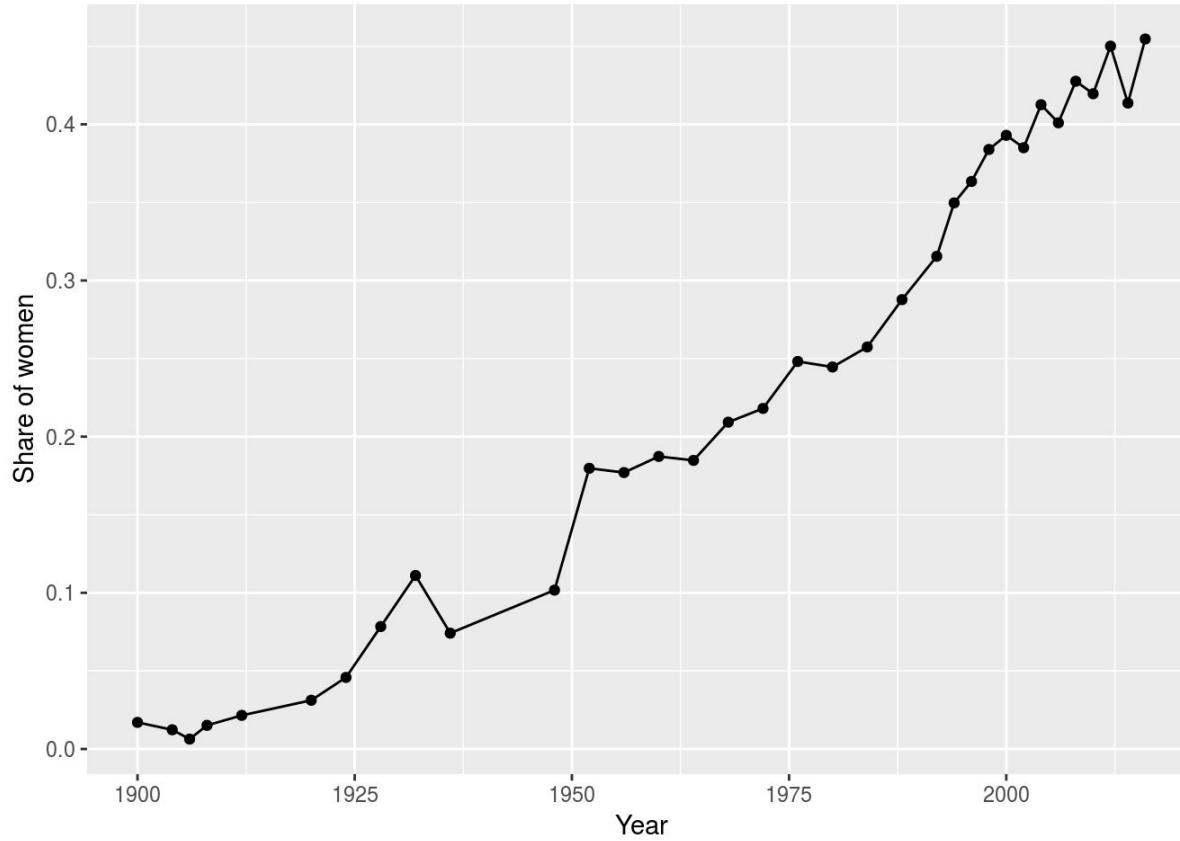
```
## The following objects are masked from 'package:stats':  
##  
##     filter, lag
```

```
## The following objects are masked from 'package:base':  
##  
##     intersect, setdiff, setequal, union
```

```
share <- df %>%  
  group_by(Year, Sex) %>%  
  summarise(n=n()) %>%  
  mutate(freq = n / sum(n)) %>%  
  filter(Sex=="F")
```



```
ggplot(share, aes(x=Year, y=freq)) +  
  geom_line()+  
  geom_point()+  
  ylab("Share of women")
```



The most important plots in exploratory data analysis:

- Scatterplots: `geom_point()`
- Histograms: `geom_histogram()`
- Density plots: `geom_density()`
- Boxplots: `geom_boxplot()`
- Barplots: `geom_bar()`



```
df <- df %>%  
  mutate(has.medal=Medal %in% c("Gold", "Silver", "Bronze"))  
  
table(df$has.medal)
```

```
##  
##  FALSE    TRUE  
## 231333  39783
```




```
ggplot(df, aes(x=has.medal, y=Height)) +  
  geom_boxplot()
```

```
## Warning: Removed 60171 rows containing non-finite values (stat_boxplot).
```

