# Be a Hawk not a Turkey

How a Bird's Eye View of your Data Can Streamline Data Analysis

Nicholas Tierney

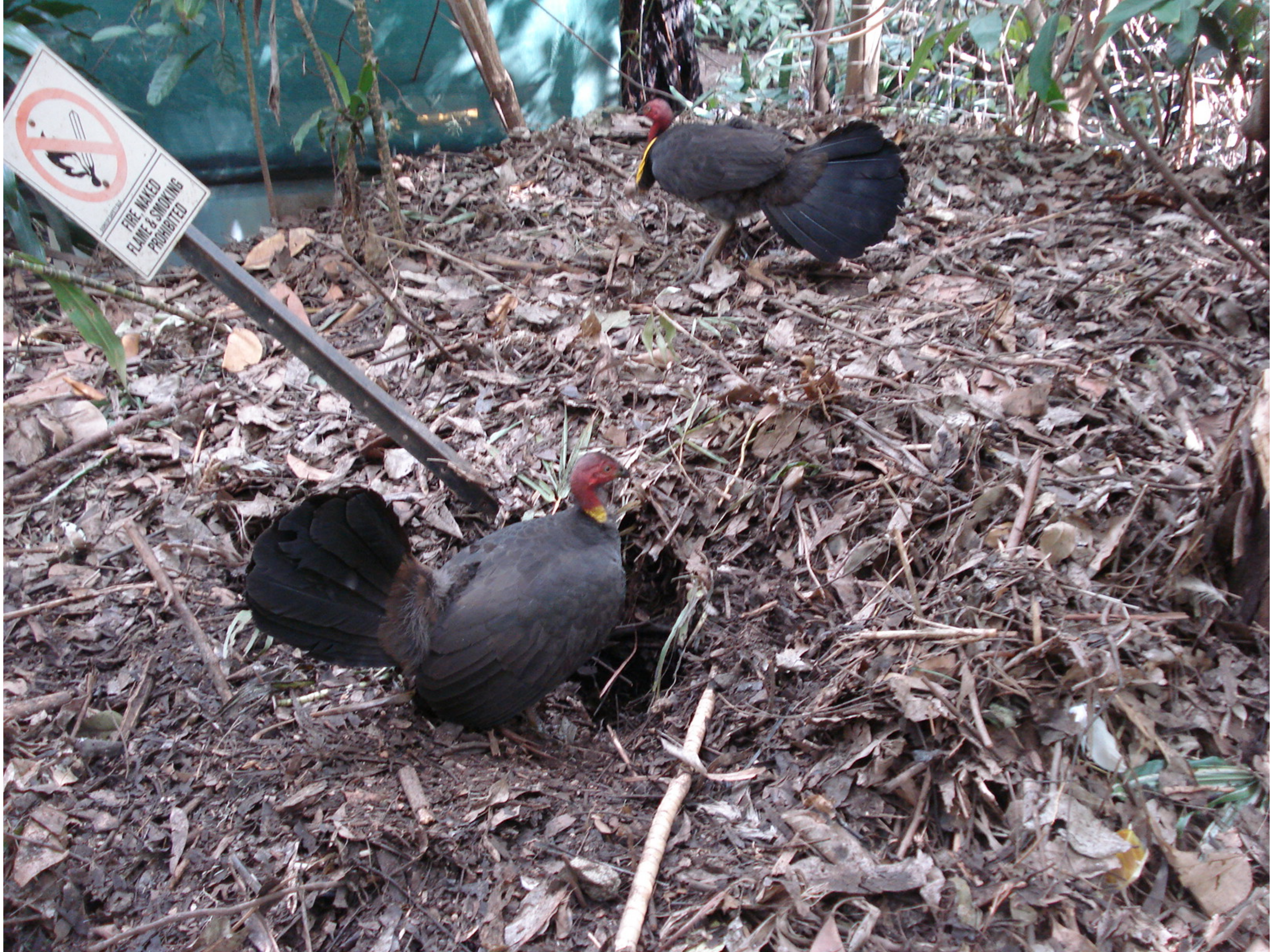
PhD Candidate QUT

WOMBAT, Melbourne Zoo

19/02/2016

# The Project

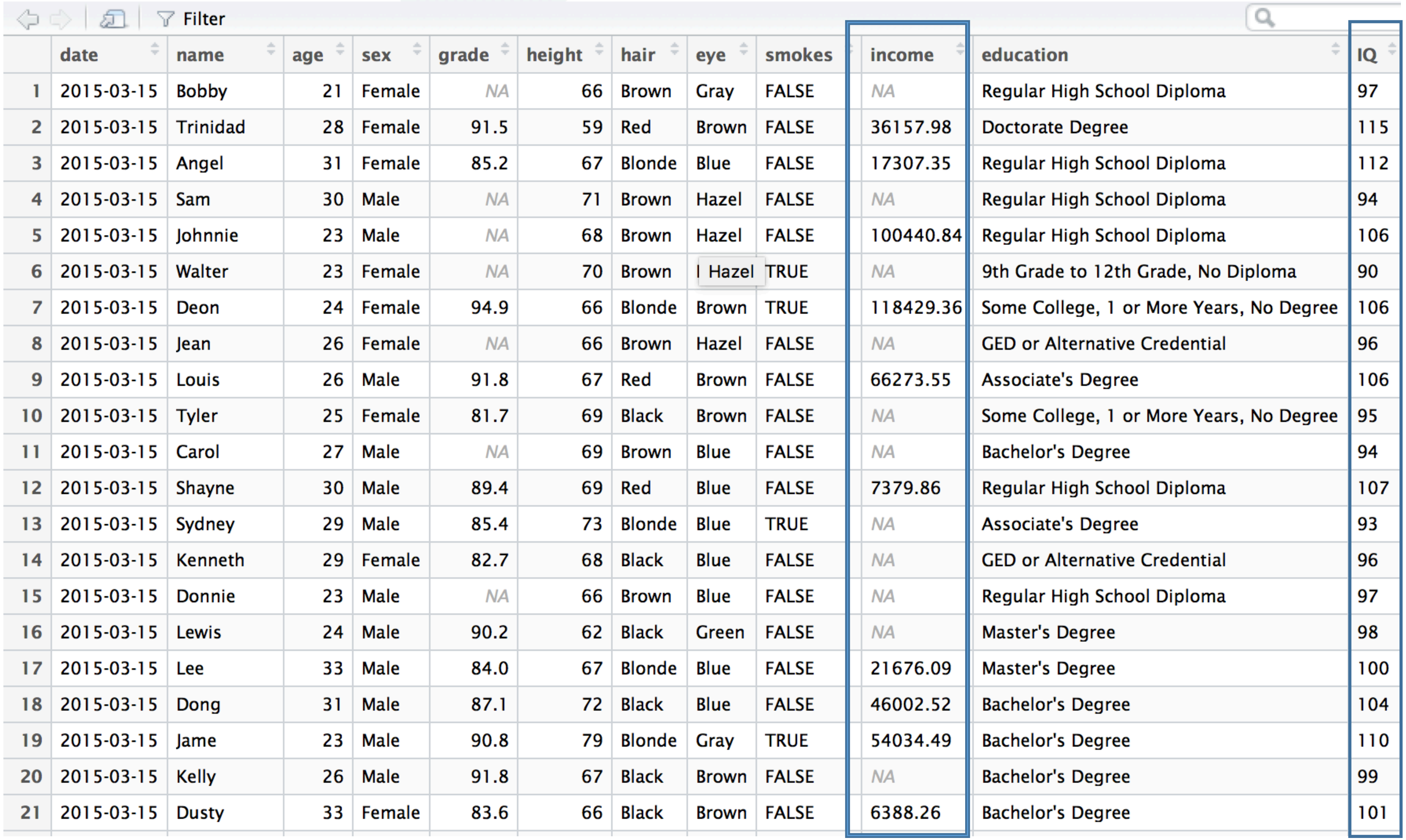
HUNTER Industrial Medicine


FIRE, NAKED FLAME & SMOKING PROHIBITED

# “Can you have a look at the data?”

What does that mean?

# “Looking” at the data

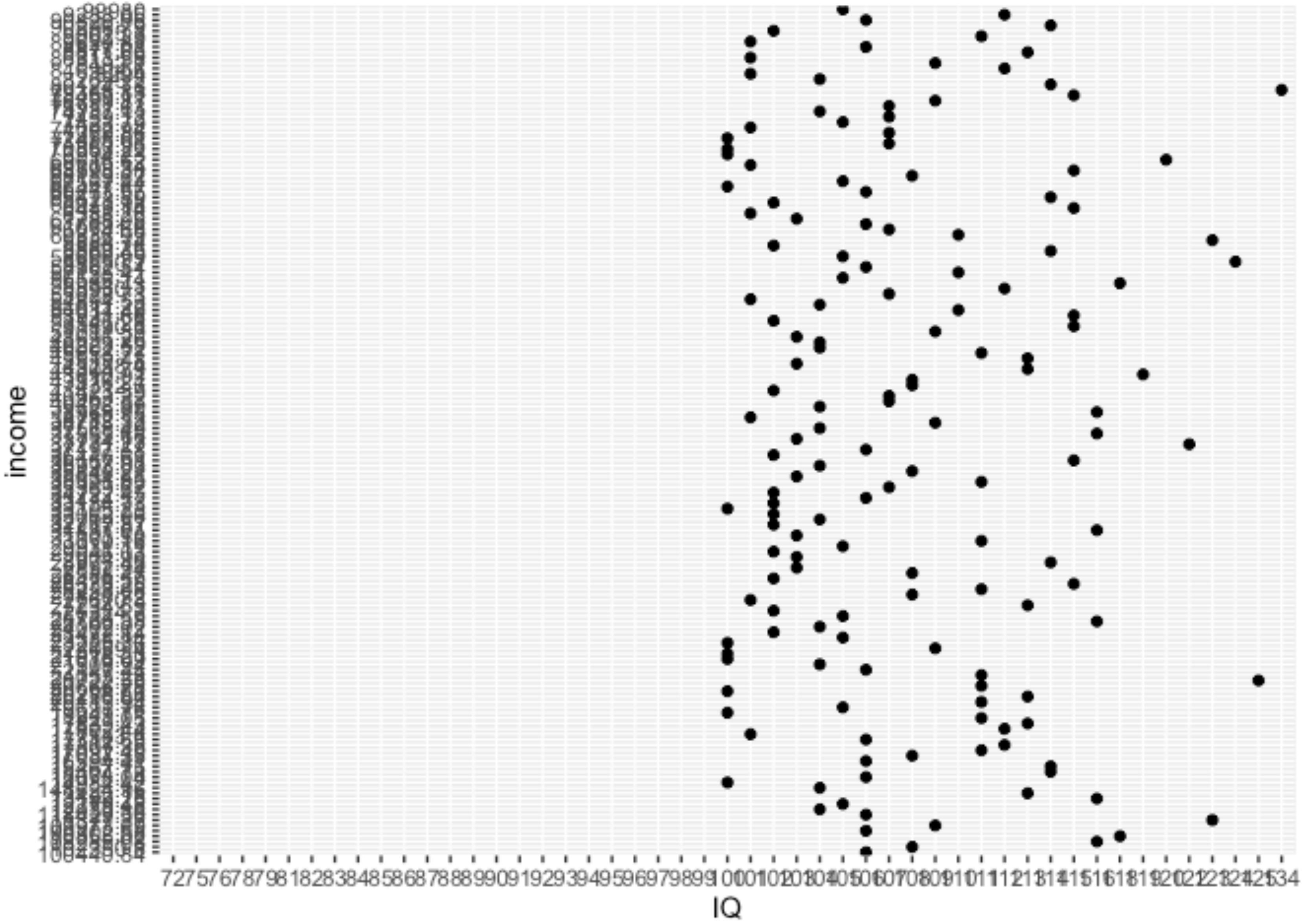


| | date | name | age | sex | grade | height | hair | eye | smokes | income | education | IQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2015-03-15 | Bobby | 21 | Female | NA | 66 | Brown | Gray | FALSE | NA | Regular High School Diploma | 97 |
| 2 | 2015-03-15 | Trinidad | 28 | Female | 91.5 | 59 | Red | Brown | FALSE | 36157.98 | Doctorate Degree | 115 |
| 3 | 2015-03-15 | Angel | 31 | Female | 85.2 | 67 | Blonde | Blue | FALSE | 17307.35 | Regular High School Diploma | 112 |
| 4 | 2015-03-15 | Sam | 30 | Male | NA | 71 | Brown | Hazel | FALSE | NA | Regular High School Diploma | 94 |
| 5 | 2015-03-15 | Johnnie | 23 | Male | NA | 68 | Brown | Hazel | FALSE | 100440.84 | Regular High School Diploma | 106 |
| 6 | 2015-03-15 | Walter | 23 | Female | NA | 70 | Brown | Hazel | TRUE | NA | 9th Grade to 12th Grade, No Diploma | 90 |
| 7 | 2015-03-15 | Deon | 24 | Female | 94.9 | 66 | Blonde | Brown | TRUE | 118429.36 | Some College, 1 or More Years, No Degree | 106 |
| 8 | 2015-03-15 | Jean | 26 | Female | NA | 66 | Brown | Hazel | FALSE | NA | GED or Alternative Credential | 96 |
| 9 | 2015-03-15 | Louis | 26 | Male | 91.8 | 67 | Red | Brown | FALSE | 66273.55 | Associate's Degree | 106 |
| 10 | 2015-03-15 | Tyler | 25 | Female | 81.7 | 69 | Black | Brown | FALSE | NA | Some College, 1 or More Years, No Degree | 95 |
| 11 | 2015-03-15 | Carol | 27 | Male | NA | 69 | Brown | Blue | FALSE | NA | Bachelor's Degree | 94 |
| 12 | 2015-03-15 | Shayne | 30 | Male | 89.4 | 69 | Red | Blue | FALSE | 7379.86 | Regular High School Diploma | 107 |
| 13 | 2015-03-15 | Sydney | 29 | Male | 85.4 | 73 | Blonde | Blue | TRUE | NA | Associate's Degree | 93 |
| 14 | 2015-03-15 | Kenneth | 29 | Female | 82.7 | 68 | Black | Blue | FALSE | NA | GED or Alternative Credential | 96 |
| 15 | 2015-03-15 | Donnie | 23 | Male | NA | 66 | Brown | Blue | FALSE | NA | Regular High School Diploma | 97 |
| 16 | 2015-03-15 | Lewis | 24 | Male | 90.2 | 62 | Black | Green | FALSE | NA | Master's Degree | 98 |
| 17 | 2015-03-15 | Lee | 33 | Male | 84.0 | 67 | Blonde | Blue | FALSE | 21676.09 | Master's Degree | 100 |
| 18 | 2015-03-15 | Dong | 31 | Male | 87.1 | 72 | Black | Blue | FALSE | 46002.52 | Bachelor's Degree | 104 |
| 19 | 2015-03-15 | Jame | 23 | Male | 90.8 | 79 | Blonde | Gray | TRUE | 54034.49 | Bachelor's Degree | 110 |
| 20 | 2015-03-15 | Kelly | 26 | Male | 91.8 | 67 | Black | Brown | FALSE | NA | Bachelor's Degree | 99 |
| 21 | 2015-03-15 | Dusty | 33 | Female | 83.6 | 66 | Black | Brown | FALSE | 6388.26 | Bachelor's Degree | 101 |

# “...Looking?” at the data?

```
ggplot(data = data,
       aes(x = IQ,
           y = income)) +
       geom_point()
```

# “…Looking?” at the data?



# So…

What if the data is all weird, and stuff?

# Real data is generally real messy

Dates are not dates

Gender is not Categorical

Rows are supposed to be columns

Missing data

# Data Cleaning...janitorial work...munging...

Data Wrangling

- dplyr
- plyr
- data.table

Testing Data

- assertr
- testdat

# Data inspection: `dplyr::glimpse(dat)`

```
Observations:  300
Variables:  15
$ date       (date) 2015-03-15, 2015-03-...
$ name       (chr) "Bobby", "Trinidad", ...
$ age        (int) 21, 28, 31, 30, 23, 2...
$ sex        (fctr) Female, Female, Fema...
$ grade      (int) NA, 4, 3, NA, NA, NA,...
$ height     (dbl) 66, 59, 67, 71, 68, 7...
$ hair       (fctr) Brown, Red, Blonde, ...
$ eye        (fctr) Gray, Brown, Blue, H...
$ smokes     (lgl) FALSE, FALSE, FALSE, ...
$ income     (chr) NA, "36157.98", "17307.35”
$ education  (fctr) Regular High School ...
$ IQ         (fctr) 97, 115, 112, 94, 106...
$ employment (int) NA, 1, 4, NA, 1, NA, ...
$ race       (fctr) Hispanic, Black, Bla...
$ religion   (fctr) Muslim, Christian, N...
```

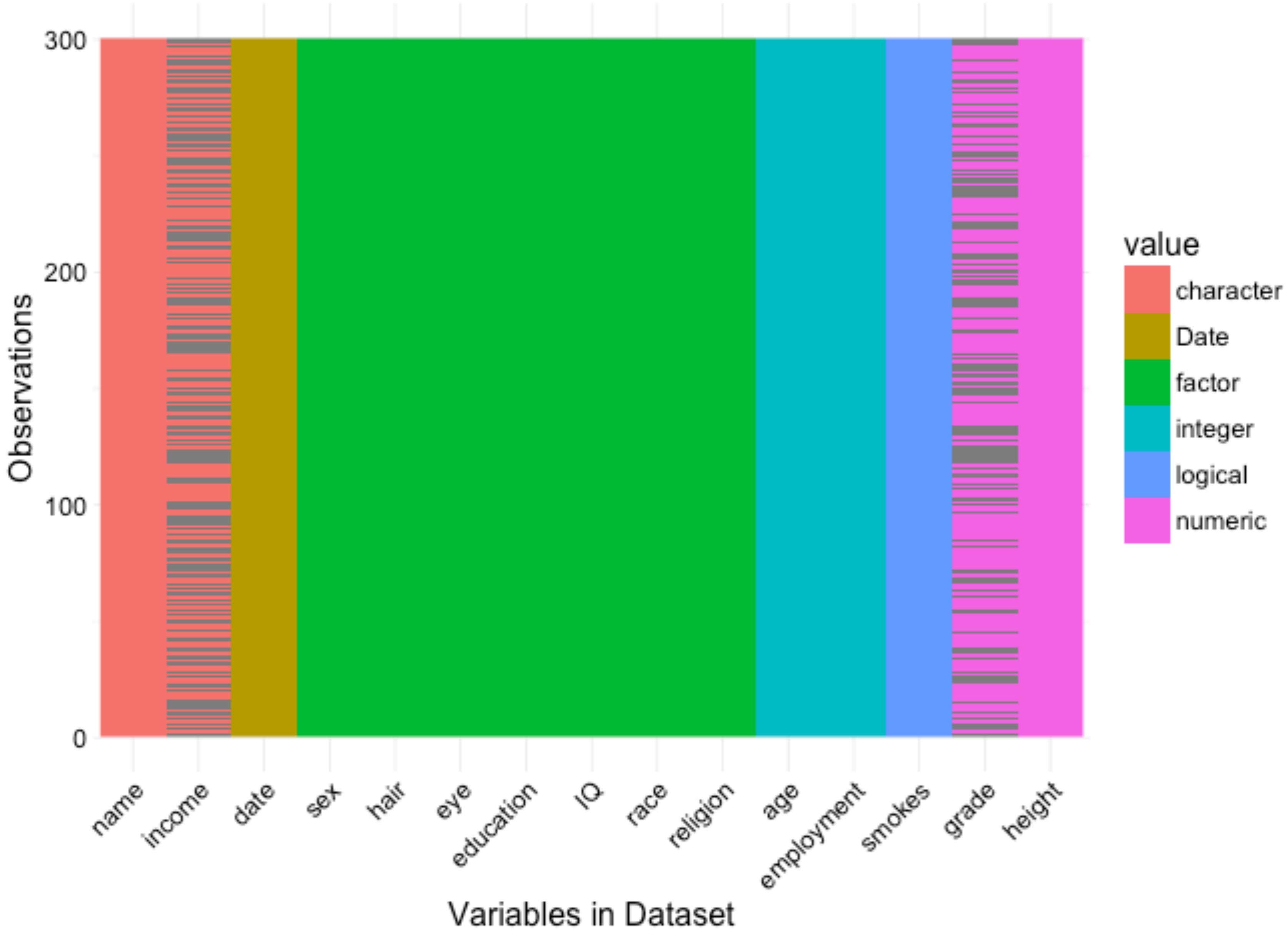
# Pre-exploratory Visualisations?

Visualisation methods for Checking Data?

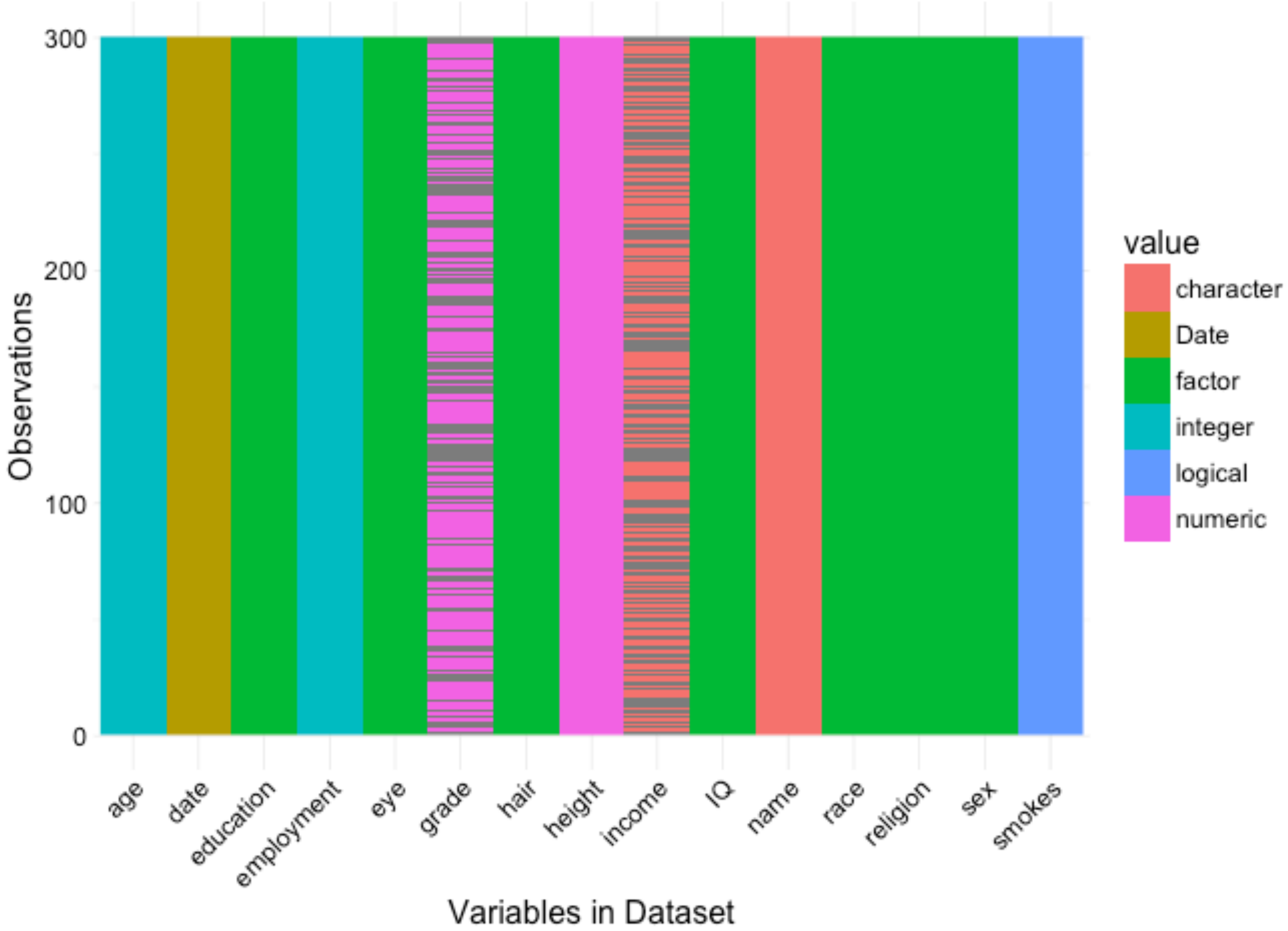
# visdat

Visualise whole data frames at once

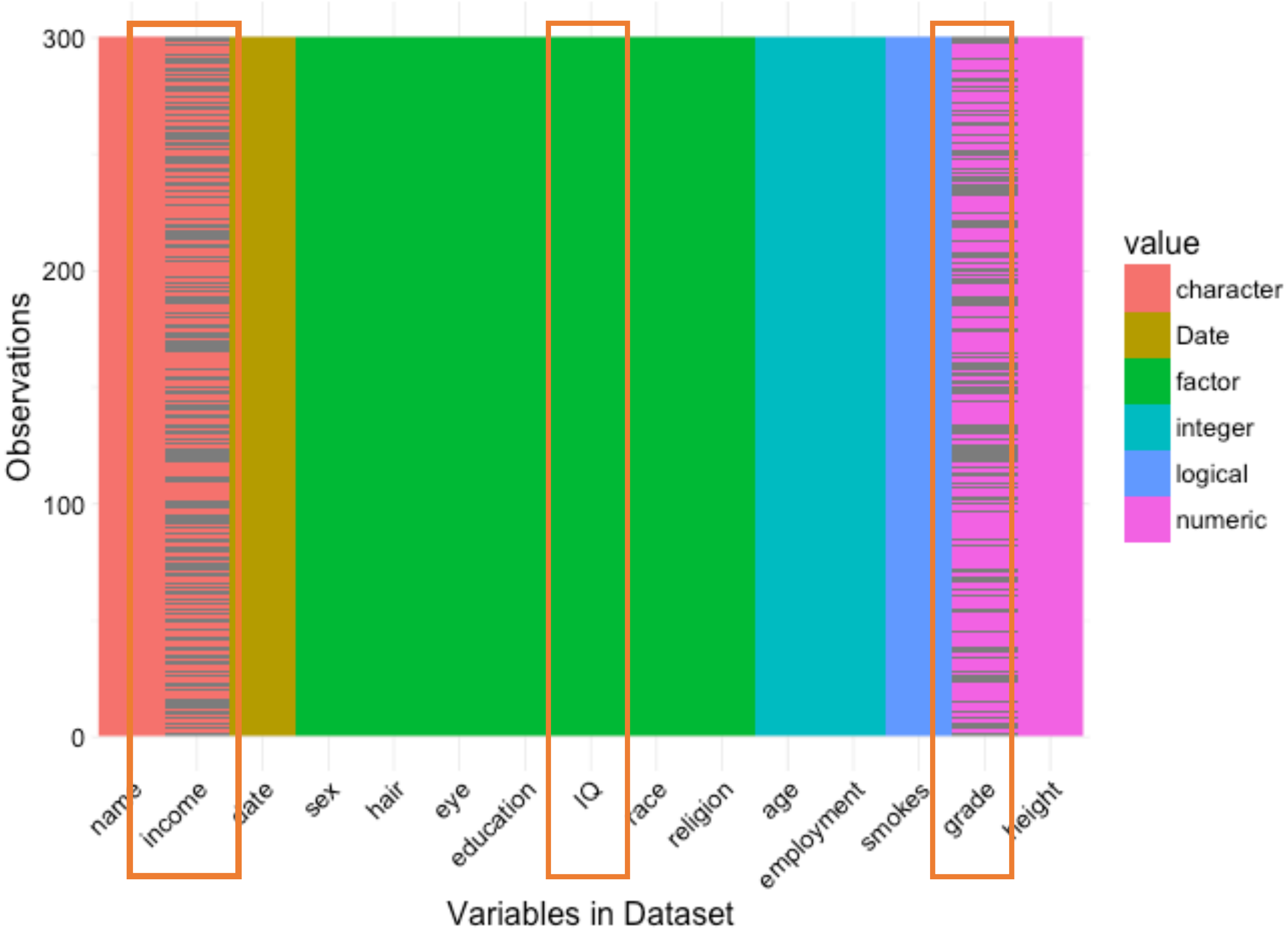
# vis_dat(data)


300
200
100
0
Observations
name
income
date
sex
hair
eye
education
IQ
race
religion
age
employment
smokes
grade
height
Variables in Dataset
value
character
Date
factor
integer
logical
numeric

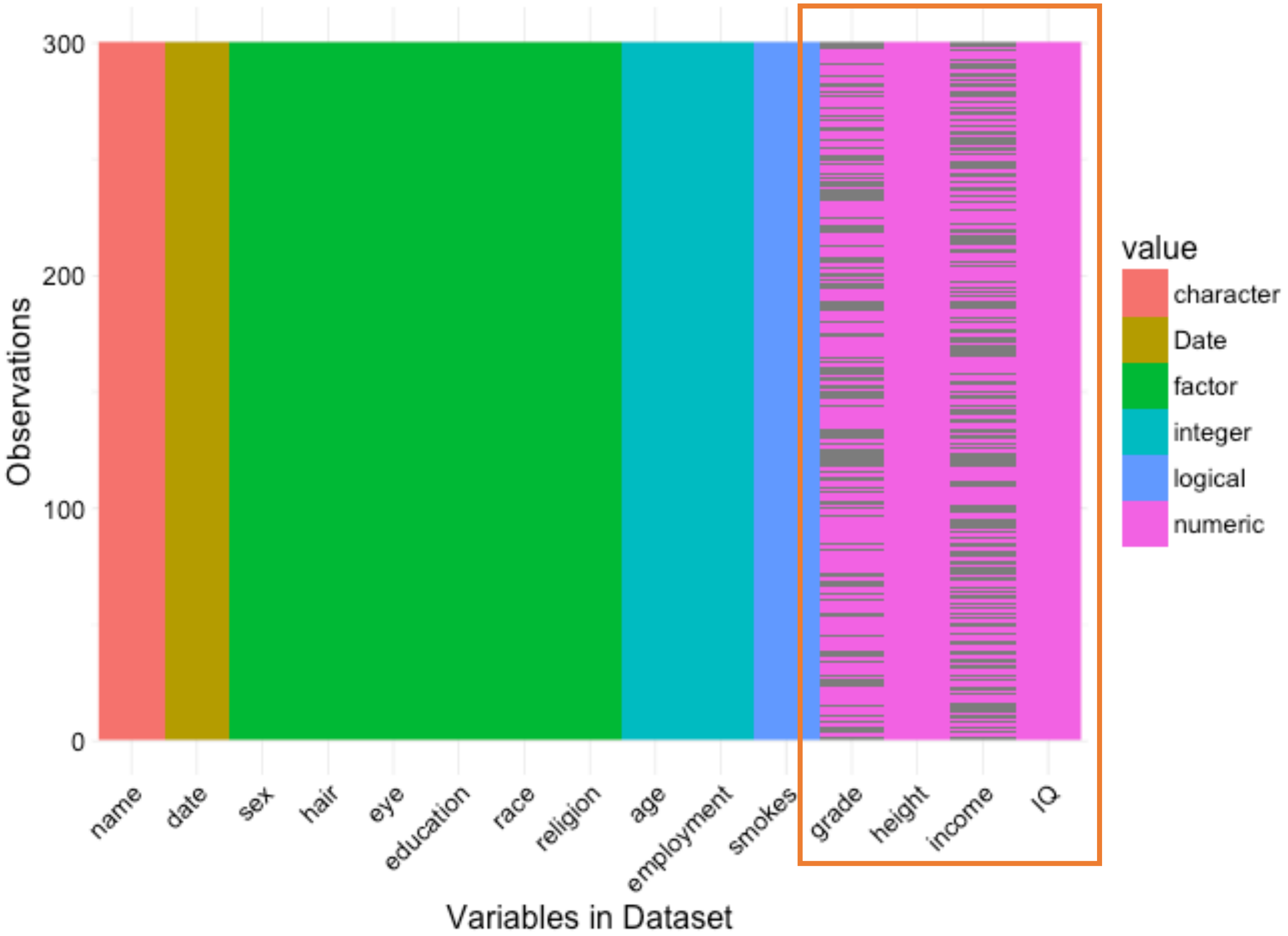
# vis_dat(data, sort_type = F)



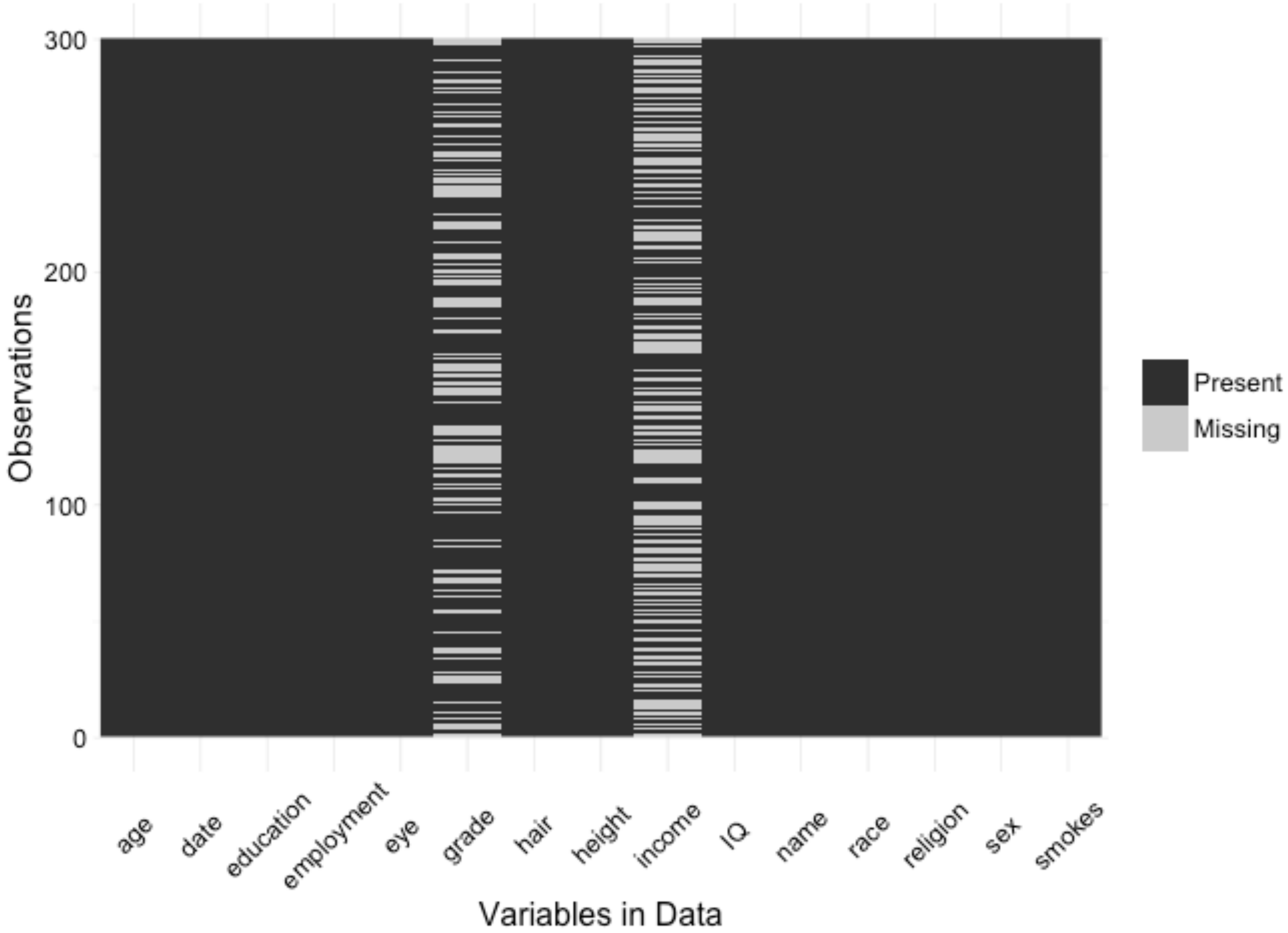
# vis_dat … clean … vis_dat … clean



# vis_dat … clean … vis_dat … clean

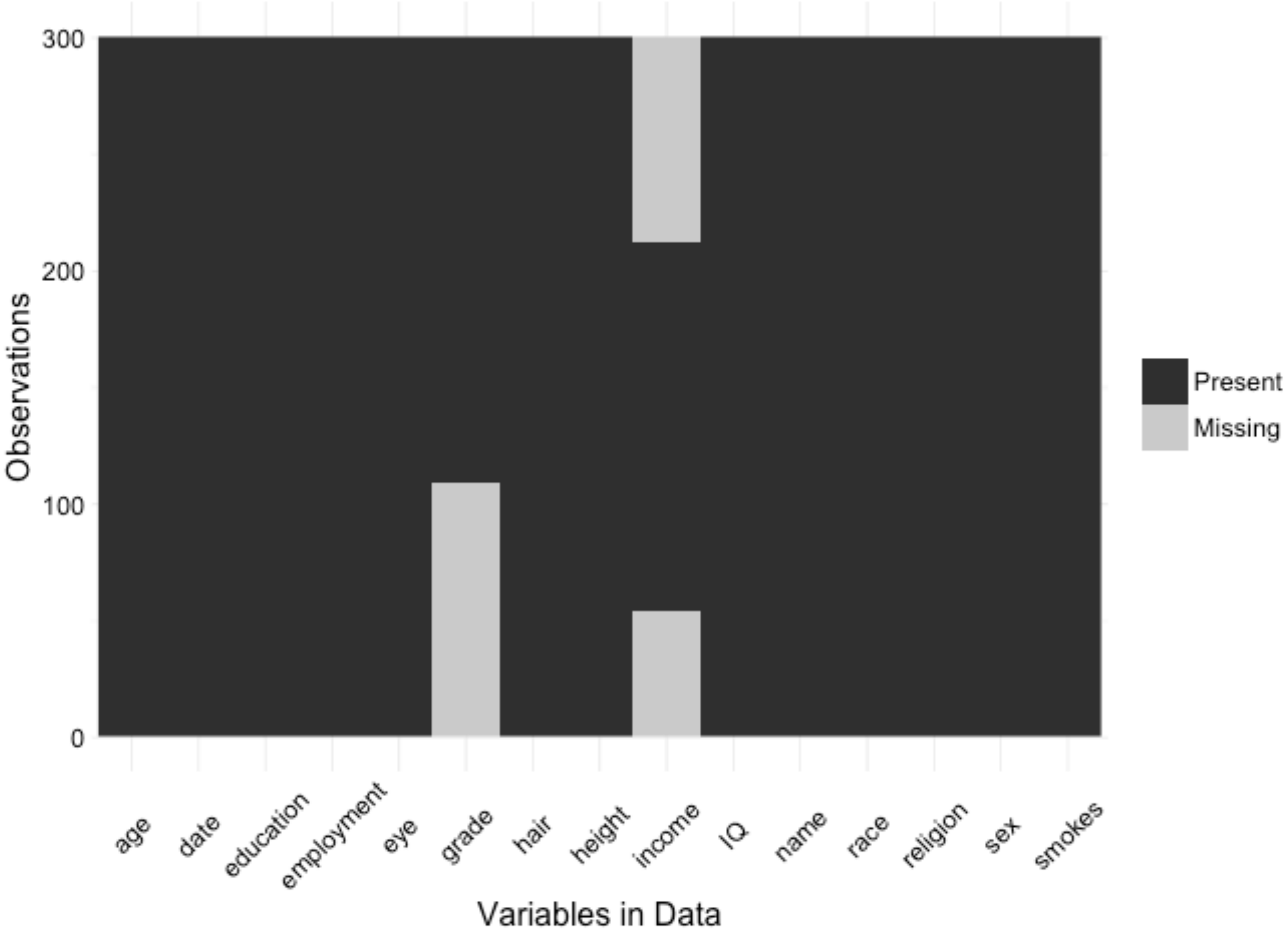


# vis_miss



300
200
100
0
Observations
age
date
education
employment
eye
grade
hair
height
income
IQ
name
race
religion
sex
smokes
Variables in Data
Present
Missing

# vis_miss(cluster = TRUE)



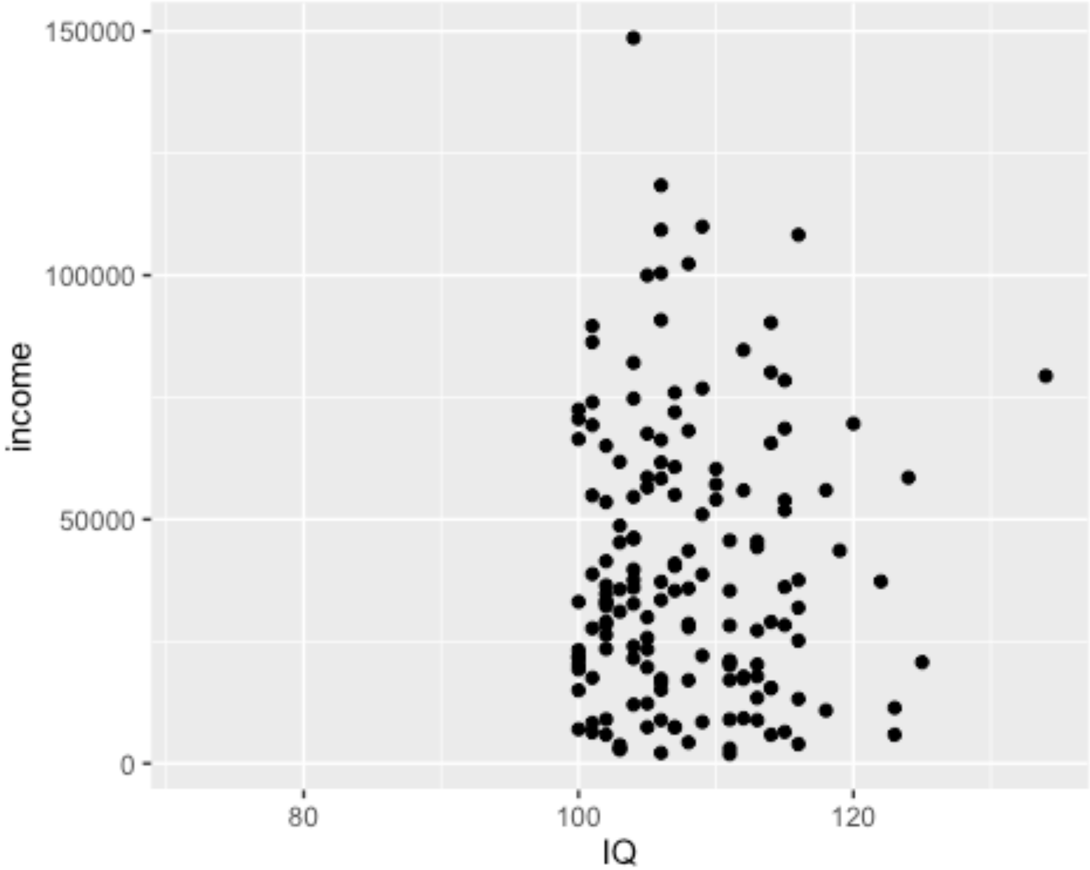
# Slide missing

It’s probably not a big deal

# ggmissing

plotting missing data with ggplot

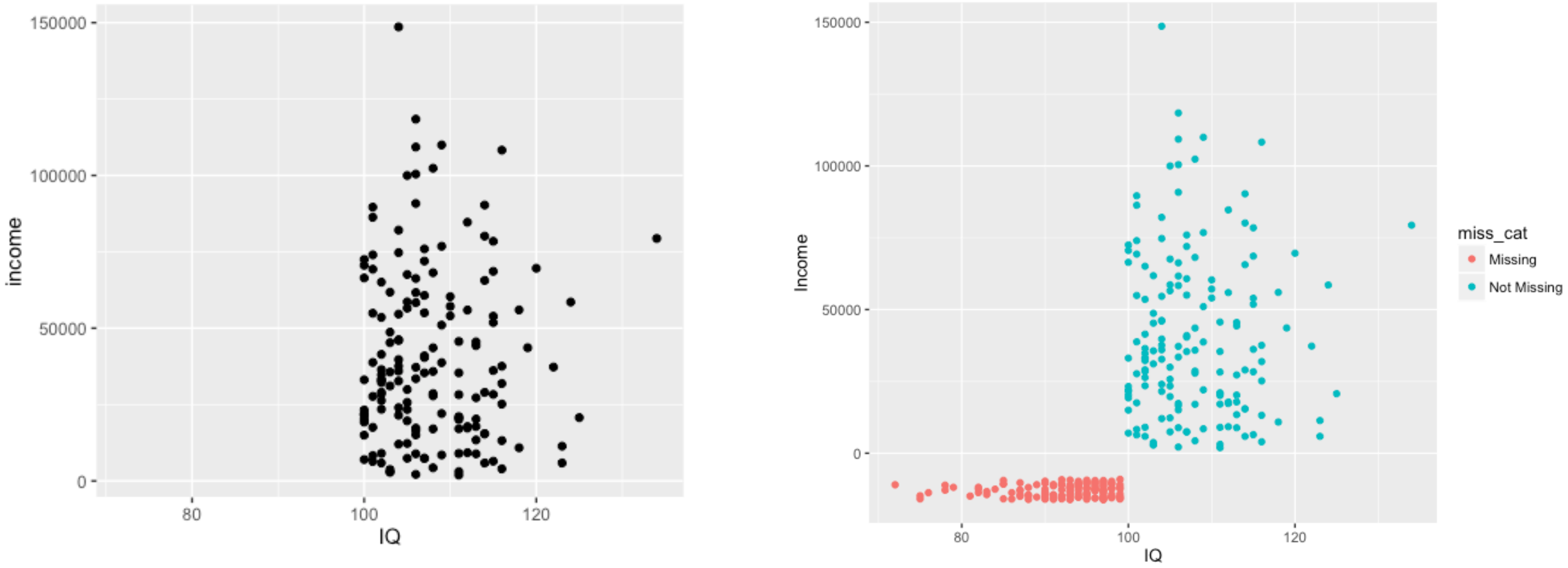
# ggmissing



```
ggplot(data = dat,
       aes(x = IQ ,
           y = income)) +
       geom_point()
```

```
Warning message:
Removed 142 rows
containing missing
values(geom_point).
```

# ggmissing



150000
100000
50000
0
income
80
100
120
IQ
150000
100000
50000
0
Income
80
100
120
IQ
miss_cat
● Missing
● Not Missing

# ggmissing: how to do it

```r
dat %>%
mutate(miss_cat = miss_cat(., "IQ", "income")) %>%
ggplot(data = .,
       aes(x = shadow_shift(IQ),
           y = shadow_shift(income),
           colour = miss_cat)) +
   geom_point()
```

# ggmissing: how we'd like to do it

```
ggplot(data = data,
       aes(x = IQ,
           y = income)) +
   geom_point() +
   geom_missing()
```

```
ggplot(data = data,
       aes(x = IQ,
           y = income)) +
   geom_point(show_missing = T)
```

# Future Work

ggmissing and visdat

# Future Work: visdat

Colour cells intelligently

Guess what kind a variable is

Read in horrible messy data

Include interactivity

Think about ways to sensibly encode summary / value information

Pipe in expectations

# Future Work: ggmissing

Early days yet

Create a philosophy / grammar of missingness

Don’t re-write ggplot

Include rug plot to show missing data

Develop clear/intuitive ways of visualising missing values

# Got an idea or want to help?

Check out our github

`github.com/tierneyn/visdat`

`github.com/tierneyn/ggmissing`

# Thank you

Di Cook

Miles McBain

Jenny Bryan

Kerrie Mengersen

Fiona Harden

Maurice Harden

# Thank you

QUT

HUNTER Industrial Medicine

BRAG

ACEMS

# Questions?

*I caught a glimpse of happiness,*
*And saw it was a bird on a branch,*
*Fixing to take wing*

*- Richard Peck*