Web Scraping for (I/O) Psychologists

Ben Listyg
USF I/O Psychology Brown Bag

Fall 2017

Overview

What

Why

When

Where

Who

How

Web scraping is…

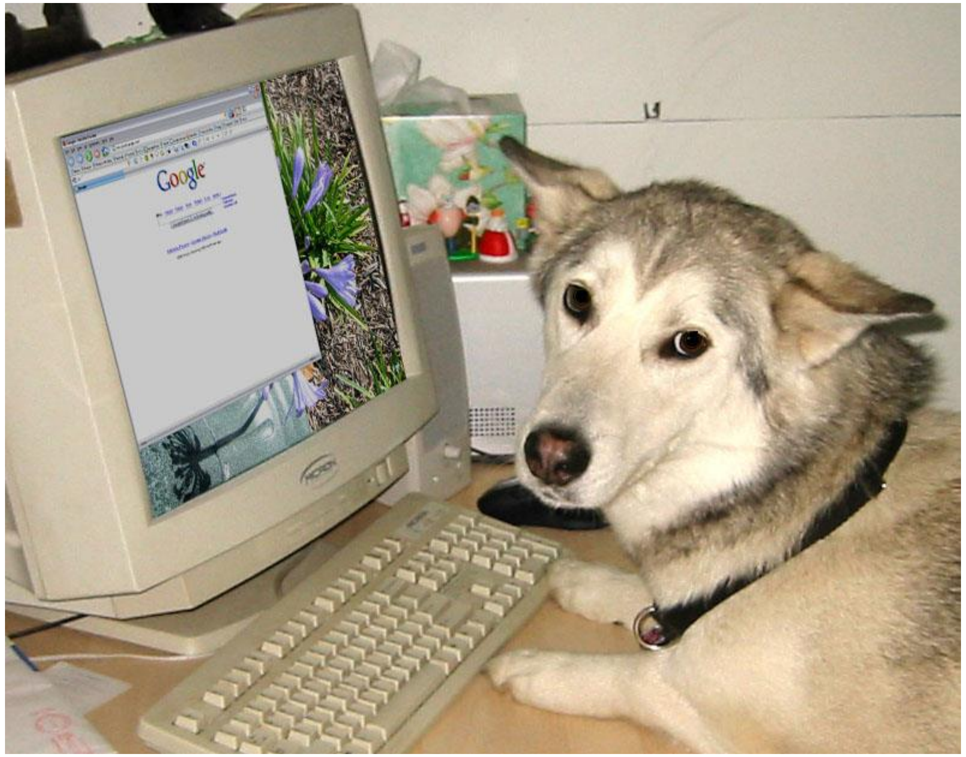
a way to collect data directly from the internet

“Data harvesting”

“Data scraping”

“Extraction”

“Pulling”



Why web scrap?

The internet generates massive amounts of data

Human behavior is frequently being recorded on the internet

- Insights that supplement existing research methods

Salganik (2016)

Why web scrap?

The internet generates massive amounts of data

Human behavior is frequently being recorded on the internet

- Insights that supplement existing research methods

Reproducibility

Who (should/could/does) web scrap?

Should?

Research question that can benefit from pre-existing data on the internet

There is available data on the internet that can address your question

Have other tasks undergrad RA’s could be doing…

Who (should/could/does) web scrap?

Could?

- Team performance
    - Basketball Reference
    - NBA Website
- Occupation / Job / Career Mobility
    - LinkedIn
- Information Sharing
    - Twitter
- Teacher Performance
    - RateMyProfessor

Who (should/could/does) web scrap?

Does?

Gentry, W. A., Hoffman, B. J., & Lyons, B. D. (2017). Box Scores and Bottom Lines: Sports Data Can Inform Research and Practice in Organizations.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological methods*, *21*(4), 475.

Wu, A.H. 2017. Gender stereotyping in academia: Evidence from economics job market rumors forum. Unpublished manuscript.

Boehmer, D. M., & Wood, W. C. (2017). Student vs. faculty perspectives on quality instruction: Gender bias, “hotness”, and “easiness” in evaluating teaching. *Journal of Education for Business*, *92*(4), 173-178.

Klug M, Bagrow JP. 2016 Understanding the group dynamics and success of teams. R. Soc. open sci.3: 160007. http://dx.doi.org/10.1098/rsos.160007

Where

The “Web”

- Hypertext Markup Language (HTML)
- APIs

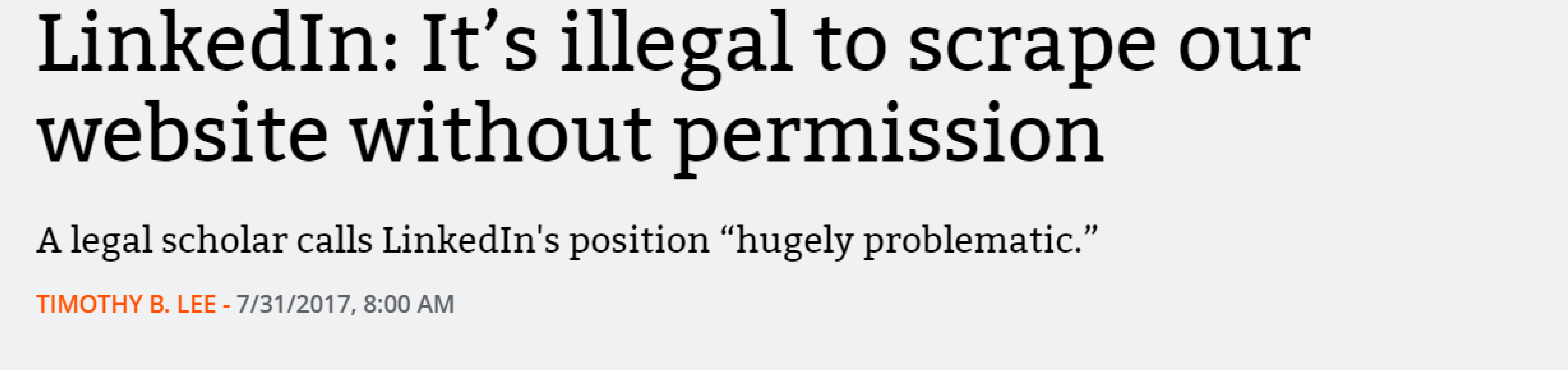
Web scraping is **not**…

a method for analyzing data

a way of generating new data

Potential issues

Legality



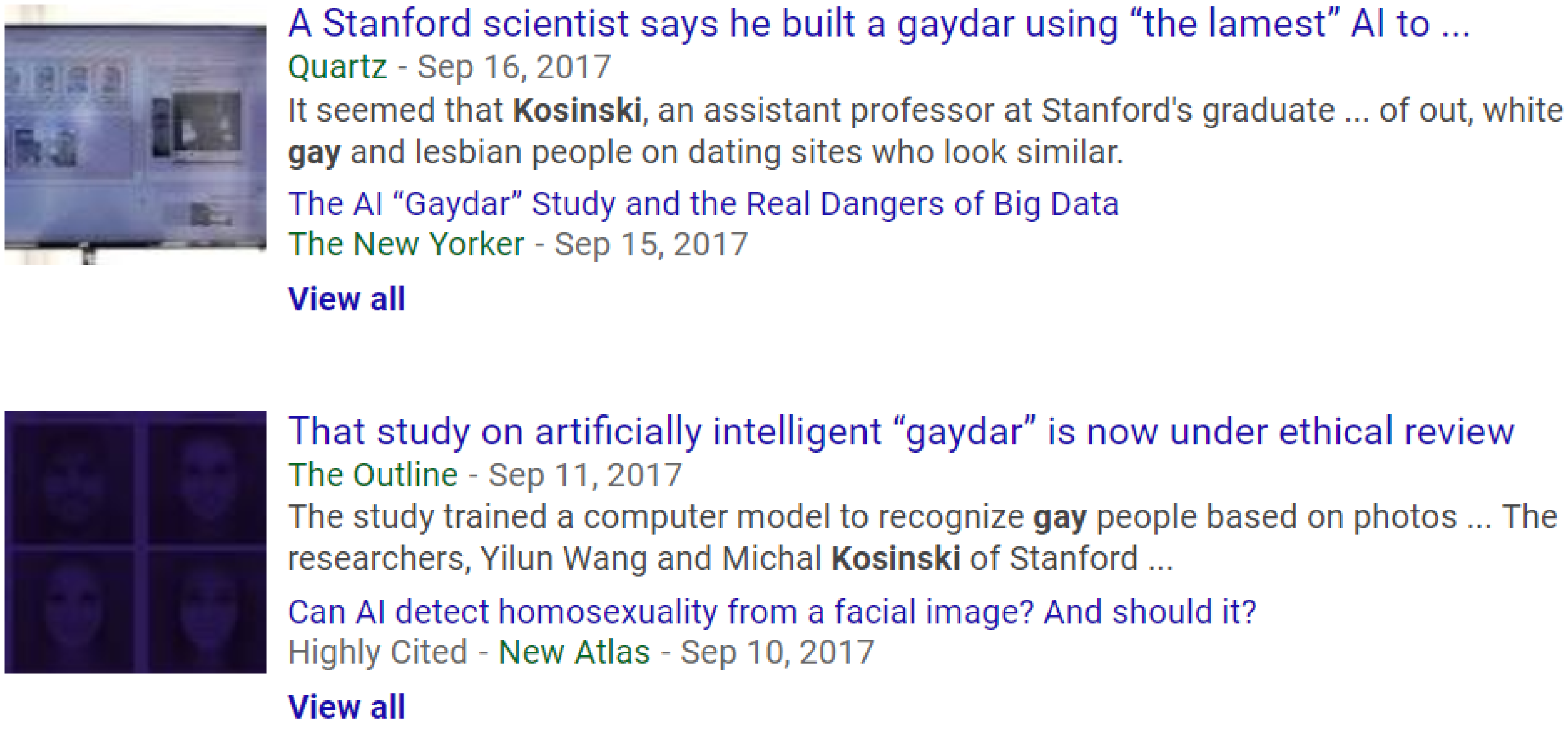
LinkedIn: It’s illegal to scrape our website without permission

A legal scholar calls LinkedIn's position “hugely problematic.”

TIMOTHY B. LEE - 7/31/2017, 8:00 AM

Potential issues

Ethics



A Stanford scientist says he built a gaydar using “the lamest” AI to ...

Quartz - Sep 16, 2017

It seemed that **Kosinski**, an assistant professor at Stanford's graduate ... of out, white **gay** and lesbian people on dating sites who look similar.

The AI “Gaydar” Study and the Real Dangers of Big Data

The New Yorker - Sep 15, 2017

View all

That study on artificially intelligent “gaydar” is now under ethical review

The Outline - Sep 11, 2017

The study trained a computer model to recognize **gay** people based on photos ... The researchers, Yilun Wang and Michal **Kosinski** of Stanford ...

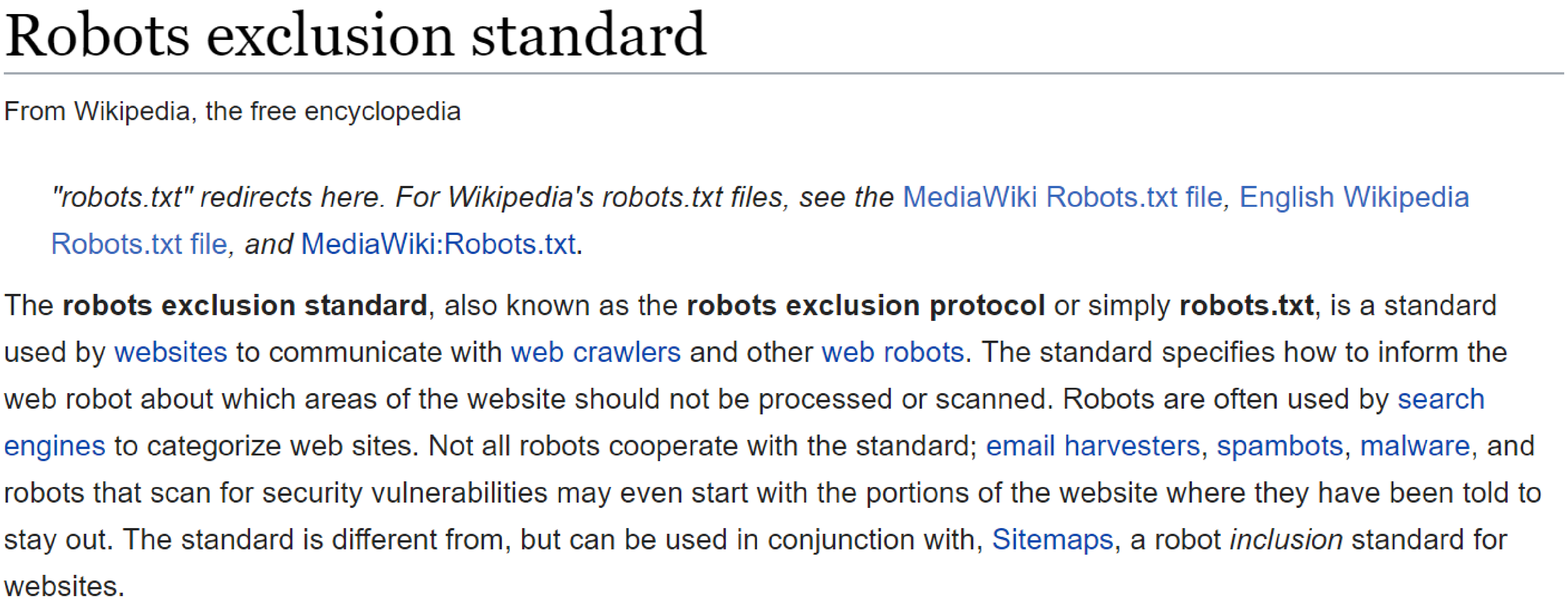
Can AI detect homosexuality from a facial image? And should it?

Highly Cited - New Atlas - Sep 10, 2017

View all

Potential issues

Technical blocks



Robots exclusion standard

From Wikipedia, the free encyclopedia

> *"robots.txt" redirects here. For Wikipedia's robots.txt files, see the* MediaWiki Robots.txt file*,* English Wikipedia Robots.txt file*, and* MediaWiki:Robots.txt.

The **robots exclusion standard**, also known as the **robots exclusion protocol** or simply **robots.txt**, is a standard used by websites to communicate with web crawlers and other web robots. The standard specifies how to inform the web robot about which areas of the website should not be processed or scanned. Robots are often used by search engines to categorize web sites. Not all robots cooperate with the standard; email harvesters, spambots, malware, and robots that scan for security vulnerabilities may even start with the portions of the website where they have been told to stay out. The standard is different from, but can be used in conjunction with, Sitemaps, a robot *inclusion* standard for websites.

How?

Huang, J. L., & Pearce, M. (2013). The other side of the coin: Vocational interests, interest differentiation and annual income at the occupation level of analysis. *Journal of Vocational Behavior*, *83*(3), 315-326.

github.com/blistyg/webscrapbb

RIASEC and Income

Interest	**Correlation**
R	−0.35
I	0.62
A	0.26
S	0.17
E	0.14
C	−0.10

Tutorial

Install and load “rvest” package

```
Install.packages(“rvest”)
library(rvest)
riasec.scrape = function(x,y) {read_html(x) %>%
                html_nodes(y) %>%
                html_text()}
```

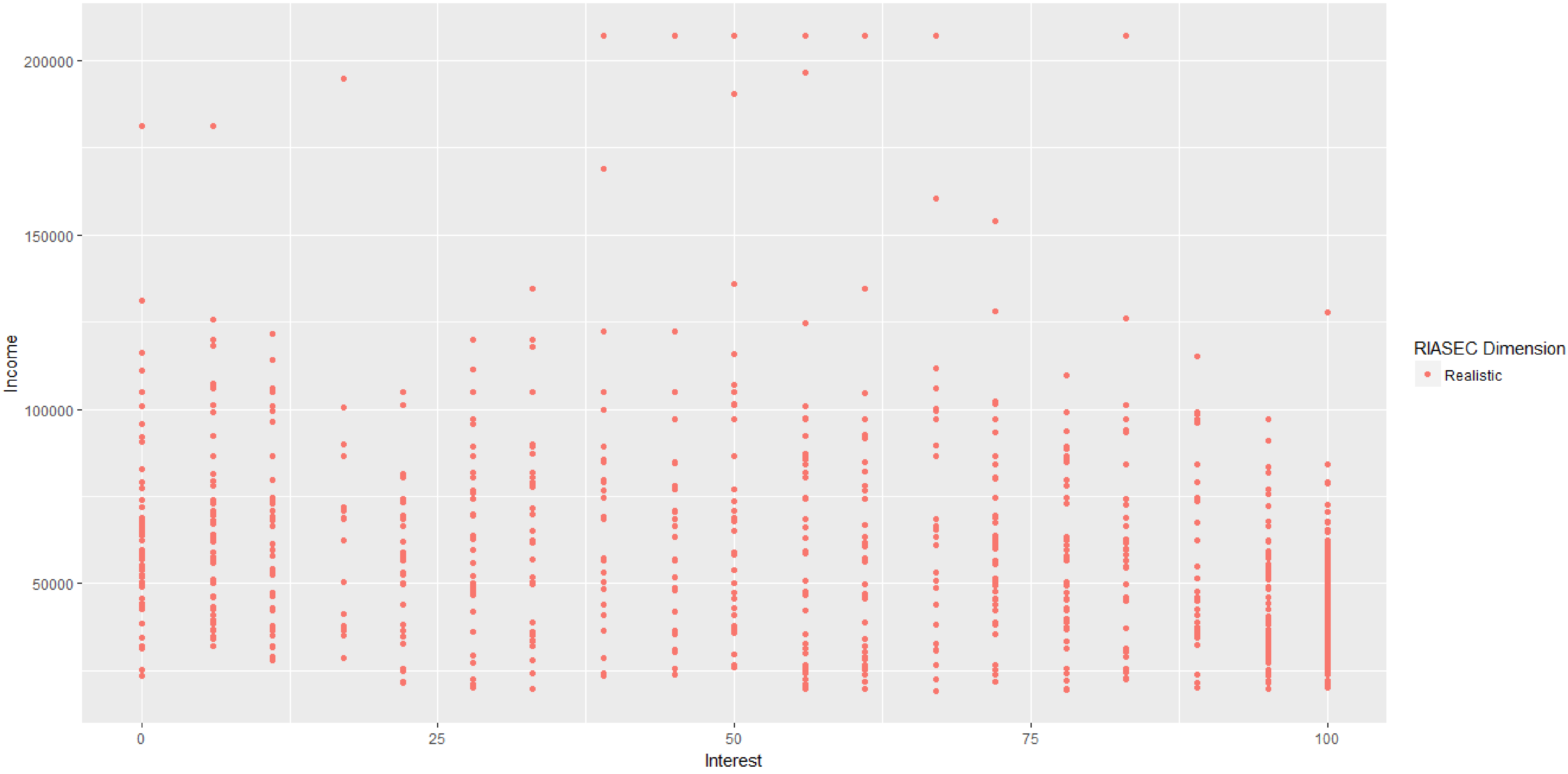
```
cbind(
riasec.scrape(x = "https://www.onetonline.org/link/details/11-9013.01", y = "#wrapper_Interests .report2a b"),
riasec.scrape(x = "https://www.onetonline.org/link/details/11-9013.01", y = "#wrapper_Interests .moreinfo b")
)
```

Complete data set

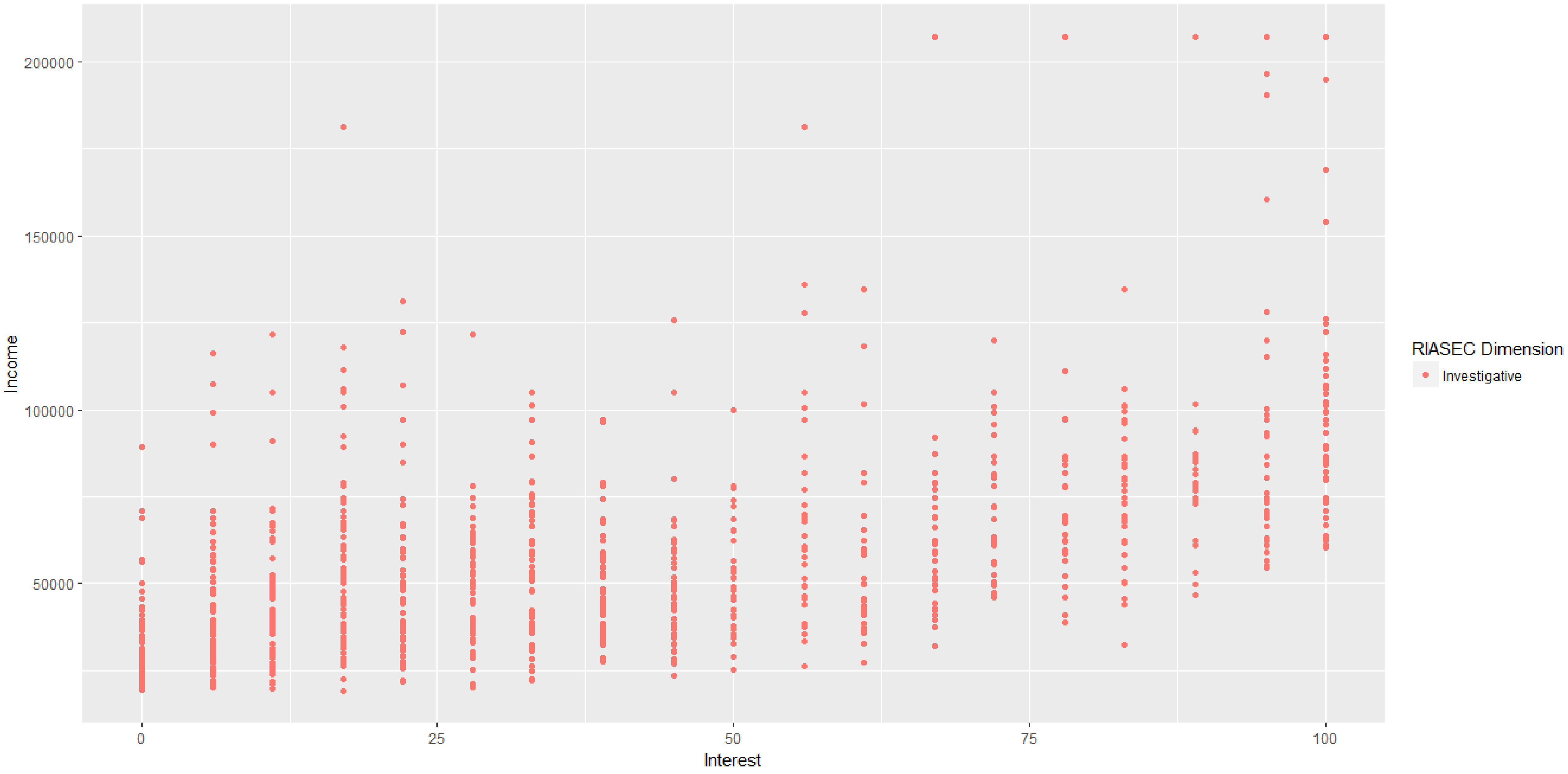
- “final.data.csv”

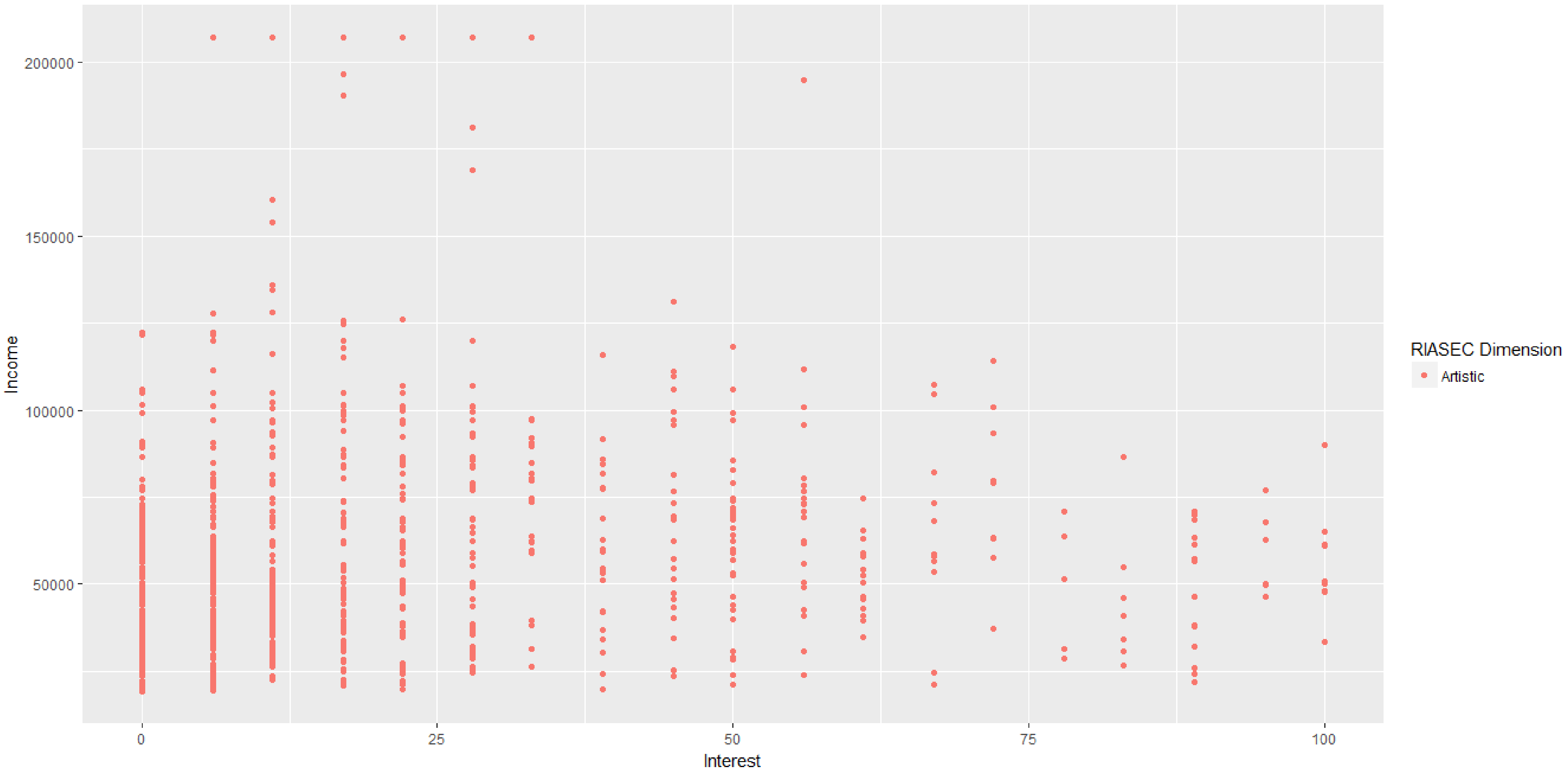
Final code

- “final.code.R”

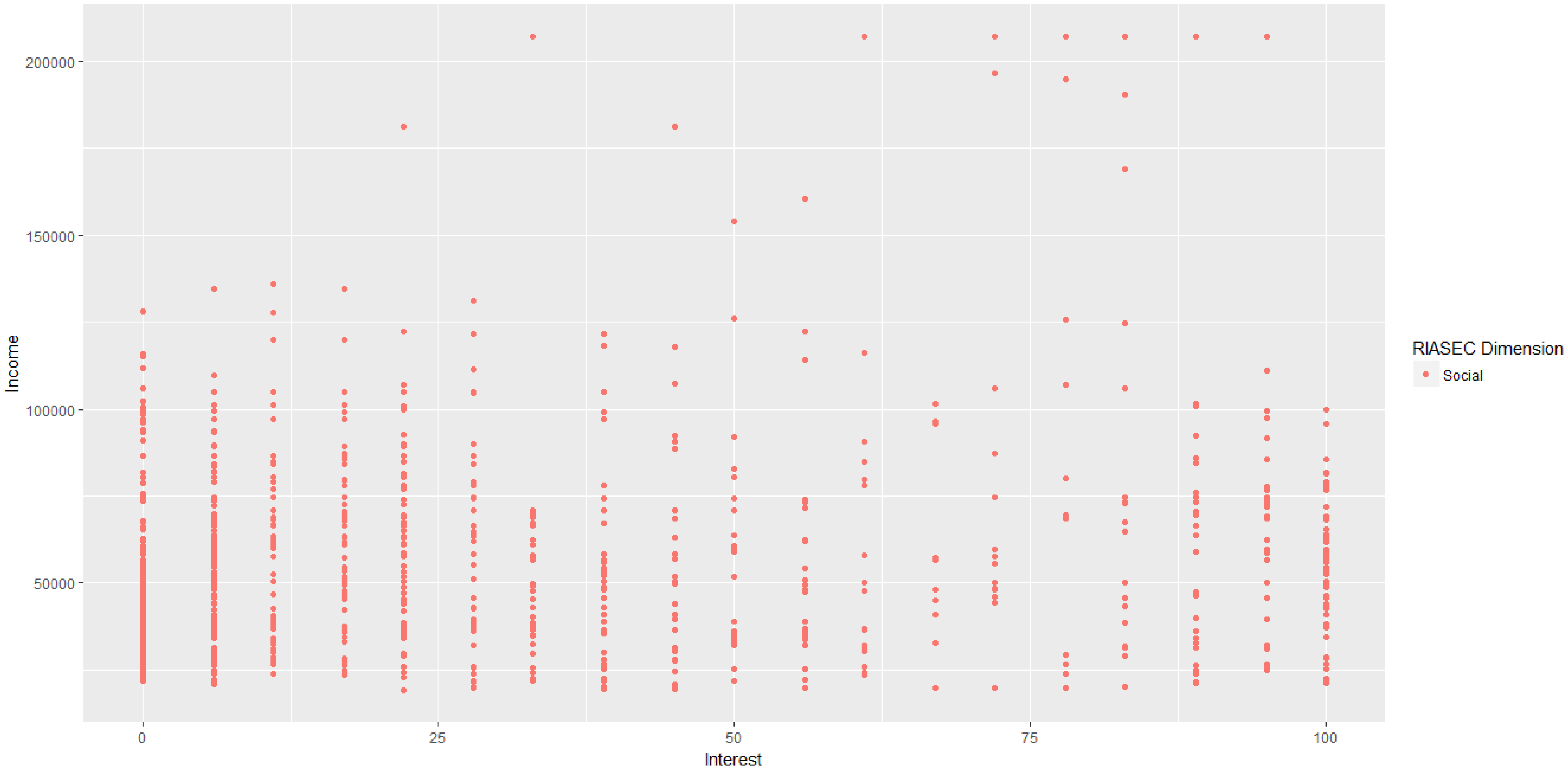


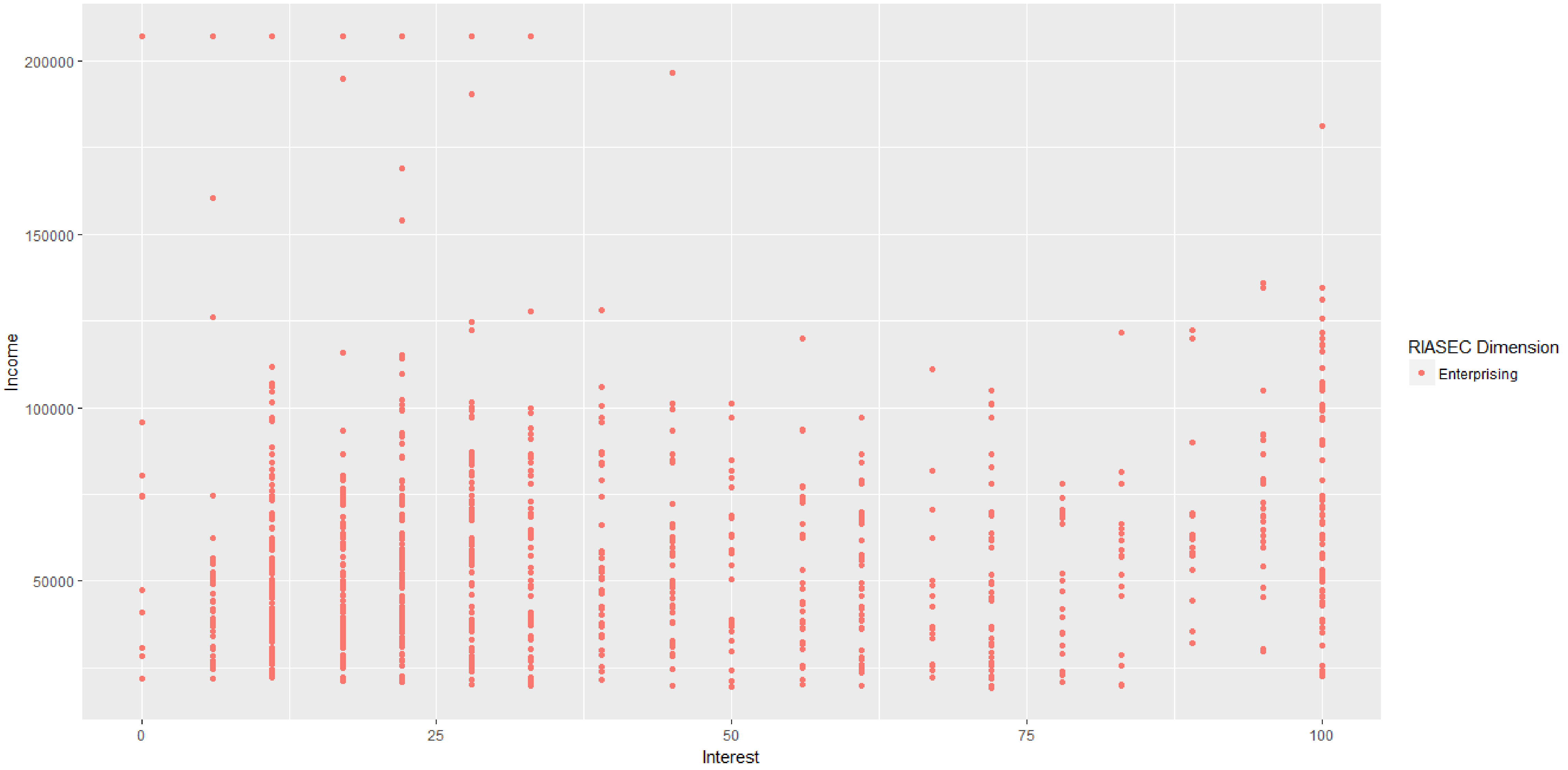
Income
200000
150000
100000
50000
0
25
50
75
100
Interest
RIASEC Dimension
● Realistic

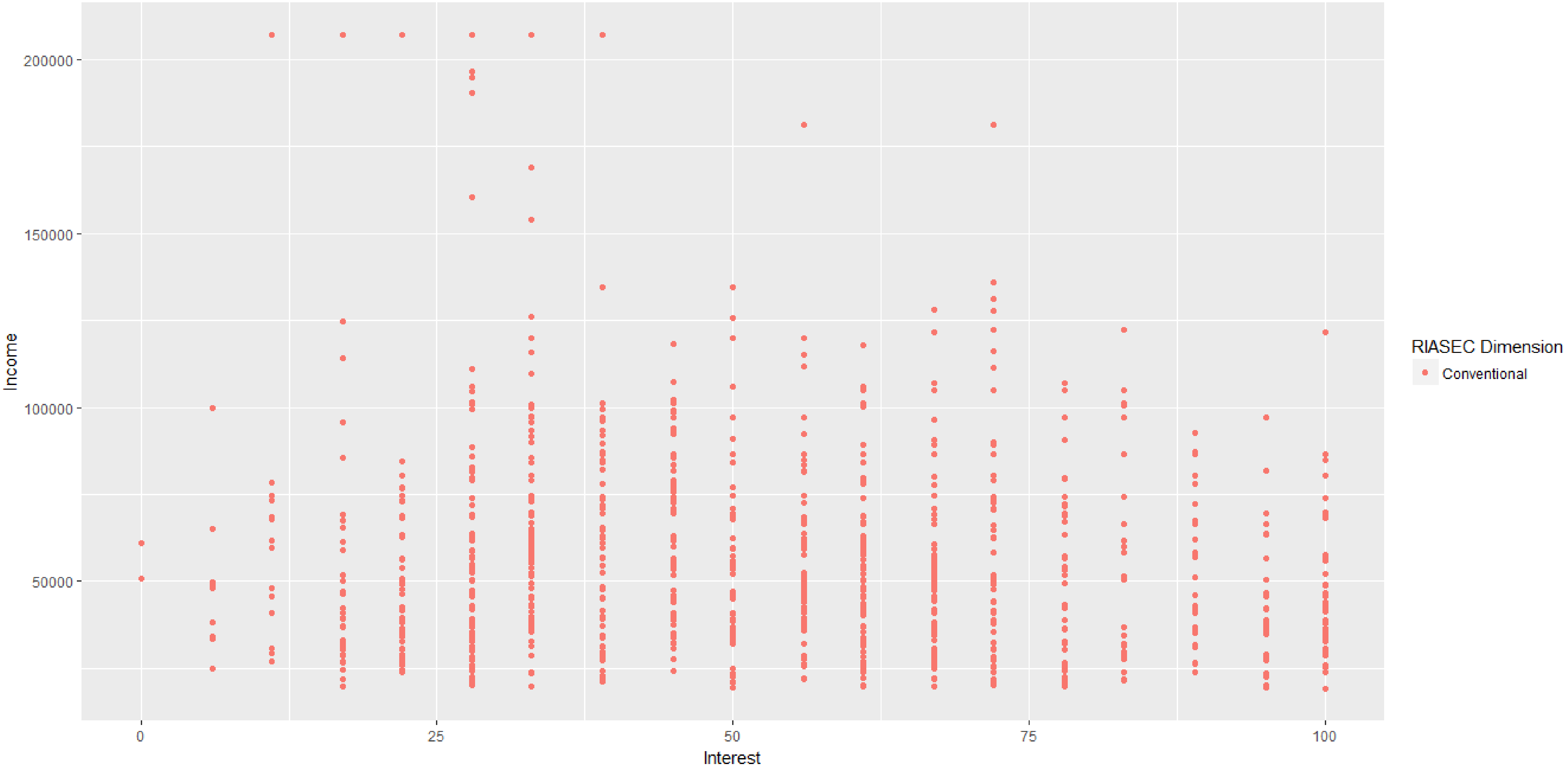

200000
150000
100000
50000
Income
0
25
50
75
100
Interest
RIASEC Dimension
● Investigative



Income
200000
150000
100000
50000
0
25
50
75
100
Interest
RIASEC Dimension
● Artistic


Income
200000
150000
100000
50000
0
25
50
75
100
Interest
RIASEC Dimension
● Social


Income
200000
150000
100000
50000
0
25
50
75
100
Interest
RIASEC Dimension
● Enterprising


Income
200000
150000
100000
50000
0
25
50
75
100
Interest
RIASEC Dimension
● Conventional

Results

Huang and Pearce (2013)

Interest	Correlation
R	–0.35
I	0.62
A	0.26
S	0.17
E	0.14
C	–0.10

USF I/O BB (2017)

Interest	Correlation
R	-0.31
I	0.56
A	0.17
S	0.18
E	0.14
C	-0.09

Questions?

Thank you!