106 Data Science Final Report

105356002 MIS 蔡佑晟

GITHUB & SHINYAPPS.IO

- GitHub: https://github.com/sxrogerex/DataScienceFinal

- SHINYAPPS.IO: http://sxrogerex.shinyapps.io/final

INPUT

- Data Source: 279 online questionnaire

- Input Format: csv

- Preprocessing:
  1. Transform Mandarin into English
  2. Transform ordinal variables into numbers, ex. 非常滿意 > 5
  3. Use mean to replace range of some ordinal variables,
     ex. $0 - $300 > $150

INPUT - DATA

Variable	Values
Sex	Male: 1, Female: 0
Income	[$0 - $5,000], [$5000 - $10,000], [$10,000 - $15,000] [$15,000 - $20,000], [$20,000 up]
Purchase Frequency	[1 week], [1 month], [3 months], [6 months], [6 months up]
Price of Short Sleeves	[$0 - $300], [$300 - $500], [$500 - $1000], [$1000 - $1500], [$1500 - $2000], [$2000 up]
Price of Long-Sleeved Shirt	
Price of Shorts	
Price of Trousers	
Price of Hoodies	
Price of Coats	

INPUT – RAW DATA

	A	B	C	D	E	F	G	H
1	時間戳記	請問您是否購買過「平價	請問您平常購買「平價服	請問您平均購買「平價服	請問對於下列「平價服飾」的品項，您願意支付	請問對於下列「平價服飾」	請問對於下列「平價服飾」	請問對於下列「平價
2	2015/12/16 上午 2:51:46	是，我購買過平價服飾。	實體服飾專賣店	一季(3個月)購買一次	300元-500元	501元-700元	501元-700元	701元-1000元
3	2015/12/16 上午 3:43:24	是，我購買過平價服飾。	實體服飾專賣店	每半年購買一次	300元-500元	300元-500元	300元-500元	501元-700元
4	2015/12/16 上午 10:41:10	是，我購買過平價服飾。	實體服飾專賣店	一個月購買一次	300元-500元	300元-500元	501元-700元	1001元-1500元
5	2015/12/16 上午 11:52:26	是，我購買過平價服飾。	實體服飾專賣店	我沒有固定頻率/忘記了	300元-500元	501元-700元	300元-500元	501元-700元
6	2015/12/16 下午 10:19:56	是，我購買過平價服飾。	網路購物平台	一個月購買一次	300元以下	300元-500元	300元以下	300元-500元
7	2015/12/16 下午 10:20:43	是，我購買過平價服飾。	百貨公司, 實體服飾專賣店	一個月購買一次	501元-700元	501元-700元	501元-700元	1001元-1500元
8	2015/12/16 下午 10:23:53	是，我購買過平價服飾。	百貨公司, 實體服飾專賣店	超過半年購買一次	300元以下	300元-500元	300元-500元	300元以下
9	2015/12/16 下午 10:24:11	是，我購買過平價服飾。	實體服飾專賣店, 網路購物	一季(3個月)購買一次	300元-500元	300元-500元	300元-500元	701元-1000元
10	2015/12/16 下午 10:28:01	是，我購買過平價服飾。	實體服飾專賣店	一個月購買一次	300元以下	300元以下	300元以下	300元以下
11	2015/12/16 下午 10:30:28	是，我購買過平價服飾。	實體服飾專賣店, 網路購物	每半年購買一次	300元-500元	300元-500元	300元-500元	701元-1000元
12	2015/12/16 下午 10:30:36	是，我購買過平價服飾。	實體服飾專賣店	我沒有固定頻率/忘記了	300元-500元	501元-700元	501元-700元	701元-1000元
13	2015/12/16 下午 10:31:45	是，我購買過平價服飾。	網路購物平台	一季(3個月)購買一次	300元以下	300元以下	300元-500元	501元-700元
14	2015/12/16 下午 10:36:03	是，我購買過平價服飾。	百貨公司	一季(3個月)購買一次	300元-500元	501元-700元	501元-700元	701元-1000元
15	2015/12/16 下午 10:36:58	是，我購買過平價服飾。	百貨公司	一個月購買一次	300元以下	300元-500元	300元-500元	701元-1000元
16	2015/12/16 下午 10:38:37	是，我購買過平價服飾。	百貨公司, 實體服飾專賣店	我沒有固定頻率/忘記了	300元-500元	501元-700元	300元-500元	501元-700元
17	2015/12/16 下午 10:39:42	是，我購買過平價服飾。	實體服飾專賣店, 網路購物	我沒有固定頻率/忘記了	300元-500元	300元-500元	300元-500元	501元-700元
18	2015/12/16 下午 10:41:27	是，我購買過平價服飾。	百貨公司	每半年購買一次	300元以下	300元以下	300元-500元	501元-700元
19	2015/12/16 下午 10:44:31	是，我購買過平價服飾。	實體服飾專賣店, 網路購物	一周購買一次	300元以下	300元-500元	300元以下	300元-500元
20	2015/12/16 下午 10:45:27	是，我購買過平價服飾。	百貨公司, 實體服飾專賣店	我沒有固定頻率/忘記了	300元-500元	501元-700元	501元-700元	1001元-1500元
21	2015/12/16 下午 10:48:26	是，我購買過平價服飾。	百貨公司, 實體服飾專賣店	一季(3個月)購買一次	300元以下	300元-500元	300元-500元	501元-700元

INPUT – AFTER PROCESSING

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	physical_st	department	online_stor	frequency_	short_sleev	long_sleev	shorts	trousers	hoodies	coat	satisfaction	satisfaction	satisfaction	satisfaction	satisfaction	satisfaction	satisfaction	satisfaction
2	1	0	0	12	400	600	600	850	600	850	4	3	4	4	4	3	4	4
3	1	0	0	24	400	400	400	600	850	1250	4	4	4	4	4	3	4	4
4	1	0	0	4	400	400	600	1250	850	1750	4	4	4	4	5	5	5	4
5	1	0	0	0	400	600	400	600	850	850	2	3	3	3	3	3	3	3
6	0	0	1	4	150	400	150	400	600	600	4	4	3	3	4	3	4	4
7	1	1	0	4	600	600	600	1250	850	1750	3	4	3	4	3	2	4	4
8	1	1	0	30	150	400	400	150	600	850	3	4	3	4	3	2	3	4
9	1	0	1	12	400	400	400	850	600	1250	4	4	4	3	4	3	4	4
10	1	0	0	4	150	150	150	150	400	600	4	4	4	4	3	3	4	3
11	1	0	1	24	400	400	400	850	850	850	3	3	3	3	3	3	3	3
12	1	0	0	0	400	600	600	850	600	850	3	3	4	3	2	2	2	2
13	0	0	1	12	150	150	400	600	400	850	3	4	4	4	4	3	4	3
14	0	1	0	12	400	600	600	850	600	850	4	5	1	3	3	2	4	4
15	0	1	0	4	150	400	400	850	400	850	4	4	3	4	4	2	3	4
16	1	1	0	0	400	600	400	600	600	850	4	4	3	3	3	3	3	4
17	1	0	1	0	400	400	400	600	600	850	4	4	3	3	3	3	3	3
18	0	1	0	24	150	150	400	600	600	850	4	3	4	3	4	3	4	2
19	1	0	1	1	150	400	150	400	600	850	4	5	5	4	2	2	3	4
20	1	1	0	0	400	600	600	1250	850	1250	3	4	4	4	4	3	4	3
21	1	1	0	12	150	400	400	600	600	1750	3	4	4	4	5	3	4	4

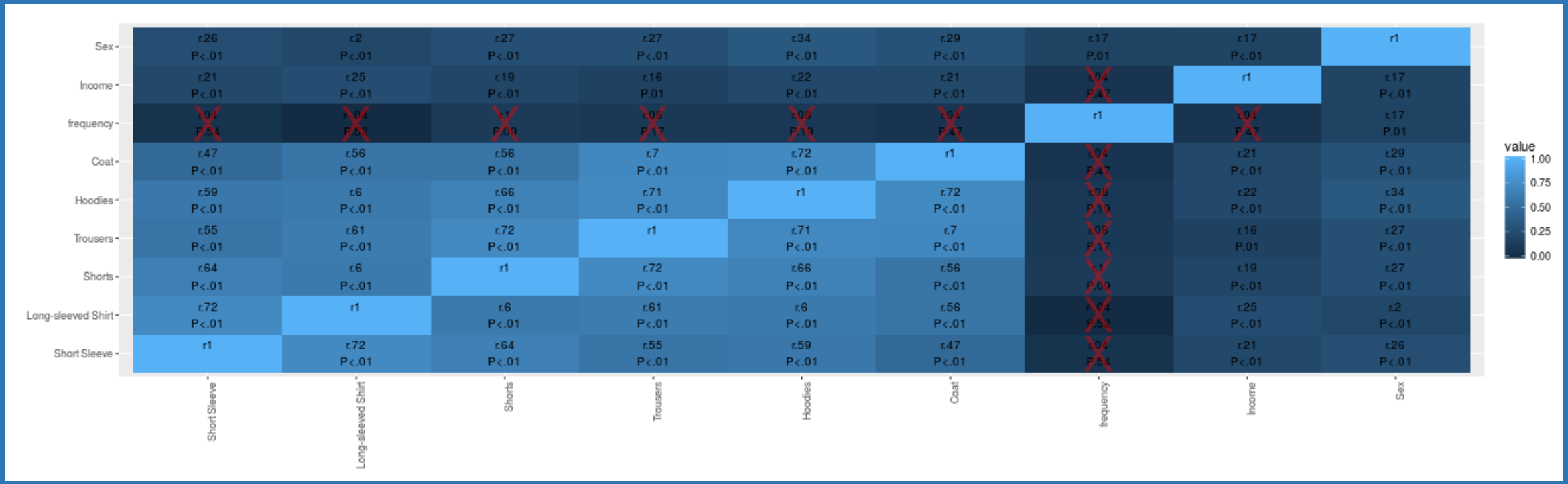
MODELING

- Spearman rank correlation
  -null model: all variables have no relation between each other

- Linear regression + k-fold cross validation
  -null model: y = a + 0 * x1 + 0 * x2 + e
  -5-fold cross validation
  -use 2 variables (income, sex) to predict the price customer willing to pay for different kinds of clothes

SPEARMAN RANK CORRELATION

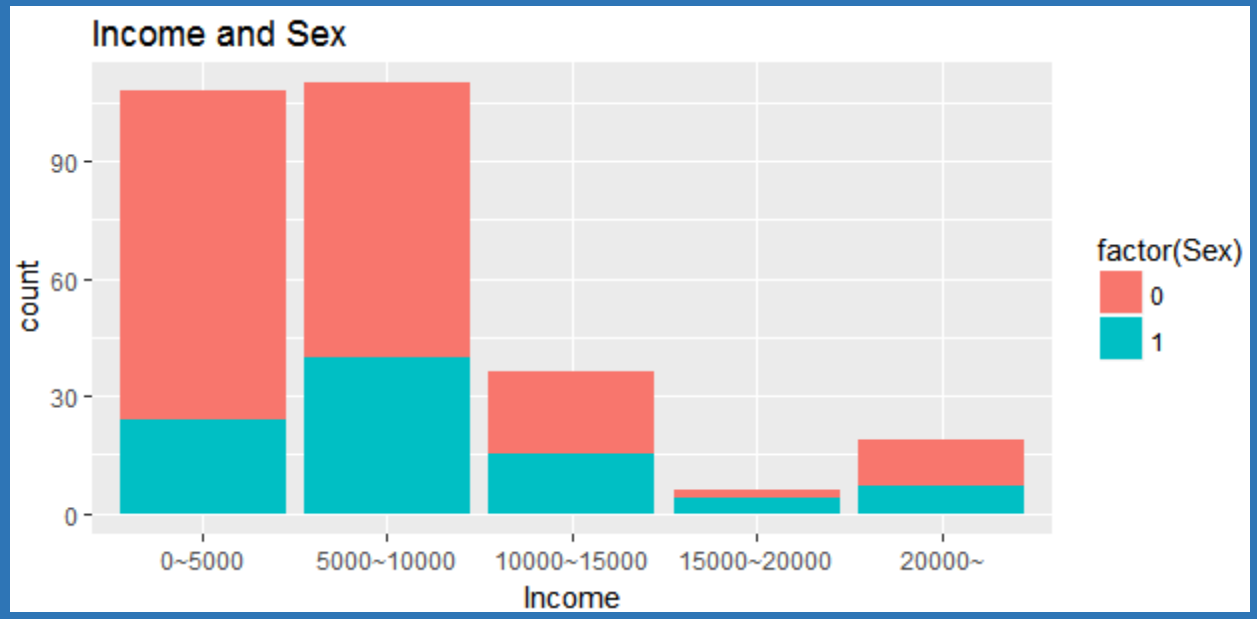
Except of purchase frequency, the correlation between all variables are statistically significant.



LINEAR REGRESSION

- Focus on 3 variables: Income, Sex, and Price (Which are correlated)

Income, Sex of sample data



The price willing to pay

	Average ($)
Short Sleeve	332
Long-sleeved Shirt	496
Shorts	464
Trousers	729
Hoodies	697
Coat	1041

FORMULAS

1. Price of Short Sleeves = a + b1 * Sex + b2 * Income + e
2. Price of Long-sleeved shirts = a + b1 * Sex + b2 * Income + e
3. Price of Shorts = a + b1 * Sex + b2 * Income + e
4. Price of Trousers = a + b1 * Sex + b2 * Income + e
5. Price of Hoodies = a + b1 * Sex + b2 * Income + e
6. Price of Coats = a + b1 * Sex + b2 * Income + e

ADJUSTED R SQUARE

- Low: 0.01 – 0.09
- Medium: 0.09 – 0.25
- ~~High: 0.25 up~~

	1	2	3	4	5
Short Sleeve	0.12	0.1	0.07	0.09	0.1
Long-sleeved Shirt	0.06	0.08	0.06	0.05	0.05
Shorts	0.08	0.08	0.08	0.08	0.07
Trousers	0.04	0.04	0.05	0.08	0.04
Hoodies	0.13	0.1	0.12	0.2	0.14
Coat	0.12	0.07	0.08	0.11	0.11

OUTPUT

- Performance:
  -Spearman rank correlation: Modestly correlated
  -Linear regression: R-Square not high, bad predict power

- Most challenging part:
  1. Hard to design a good questionnaire
  2. Preprocessing data takes lots of time
  3. Result may be bad because of many reasons, ex. Data quality, sample size, modelling…