

# Báo cáo bài tập 4

1612174 - Phùng Tiến Hào - [tienhaophung@gmail.com](mailto:tienhaophung@gmail.com)

07/04/2019

# Contents

<b>1</b>	<b>Mô tả tổng quan dataset (nguồn gốc, mục đích nghiên cứu, tổng thể nghiên cứu, cách thu thập dữ liệu, cỡ mẫu, số lượng biến).</b>	<b>1</b>
<b>2</b>	<b>Chọn ra 5 biến quan tâm, trong đó có ít nhất 2 biến định tính và 2 biến định lượng. Mô tả sơ lược ý nghĩa 5 biến này và nêu lí do chọn.</b>	<b>1</b>
<b>3</b>	<b>Phân tích thăm dò riêng từng biến đã chọn:</b>	<b>2</b>
<b>4</b>	<b>Chọn ra 2 biến định tính (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.</b>	<b>9</b>
<b>5</b>	<b>Chọn ra 1 biến định tính, 1 biến định lượng (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.</b>	<b>10</b>
<b>6</b>	<b>Chọn ra 2 biến định lượng (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.</b>	<b>11</b>
<b>7</b>	<b>(Cộng điểm) Phân tích thăm dò quan hệ giữa nhiều hơn 2 biến (từ 5 biến quan tâm).</b>	<b>13</b>
<b>8</b>	<b>Tham khảo</b>	<b>15</b>

# **1 Mô tả tổng quan dataset (nguồn gốc, mục đích nghiên cứu, tổng thể nghiên cứu, cách thu thập dữ liệu, cỡ mẫu, số lượng biến).**

- Nguồn gốc: Gelman and Hill phân tích dữ liệu sử dụng hồi quy và mô hình đa cấp /phân cấp của đại học Cambridge tại New York năm 2007
- Mục đích nghiên cứu: Khảo sát và phân tích các đánh giá của các học sinh về bạn khác giới của mình trong cuộc gặp gỡ 4 phút của trường Columbia về việc tham dự sự kiện "SpeedDating".
- Tổng thể nghiên cứu: Những người tham dự sự kiện "SpeedDating" là học sinh của trường Columbia được chọn lọc bởi các trợ lý nghiên cứu.
- Cách thu thập dữ liệu:
  - Lấy dữ liệu ngày đầu tiên của cuộc hội giữa người tham dự và bạn tình của họ
  - Các cuộc gặp gỡ được chọn ngẫu nhiên và thời lượng 4 phút
  - Sau đó, người tham dự đánh giá các thuộc tính trên thang điểm 1-10.
- Cỡ mẫu: 276 quan sát
- Số lượng biến: 22 biến

# **2 Chọn ra 5 biến quan tâm, trong đó có ít nhất 2 biến định tính và 2 biến định lượng. Mô tả sơ lược ý nghĩa 5 biến này và nêu lí do chọn.**

## **a) Các biến định tính:**

- DecisionMale (Yes/No): Quyết định nam có muốn 1 ngày hẹn nào khác không?
  - RaceF (Asian, Black,...): chủng tộc của bạn nữ
- Lý do: vì cái kết quả quan trọng nhất là bạn nam tham dự có muốn tiến đến cuộc hẹn hò thật sự vào một ngày khác không. Thường con trai họ sẽ thích những người cùng chủng tộc với họ.

## **b) Các biến định lượng:**

- AttractiveM (num): Nam đánh giá về sức quyến rũ của bạn nữ.
  - LikeM (num): Mức độ thích của người nam đối với nữ.
  - SincereM (num): Nam đánh giá về độ chân thành của nữ.
- Lý do: Đây các yếu tố mang tính cảm tính để quyết định người nam có ấn tượng ban đầu tốt đối với người phụ nữ và ảnh hưởng đến DecisionMale.

### 3 Phân tích thăm dò riêng từng biến đã chọn:

1 2

a) Biến định tính:

- DecisionMale:

```
1 tab1 = table(DecisionMale) # Count so luong nam yes va no
2 # Them total
3 addmargins(tab1)
4 > DecisionMale
5 No Yes Sum
6 130 146 276
7
8 prop.table(tab1) # Proportions
9 > DecisionMale
10 No Yes
11 0.4710145 0.5289855
12
13 barplot(tab1) # Ve barchart
14
```

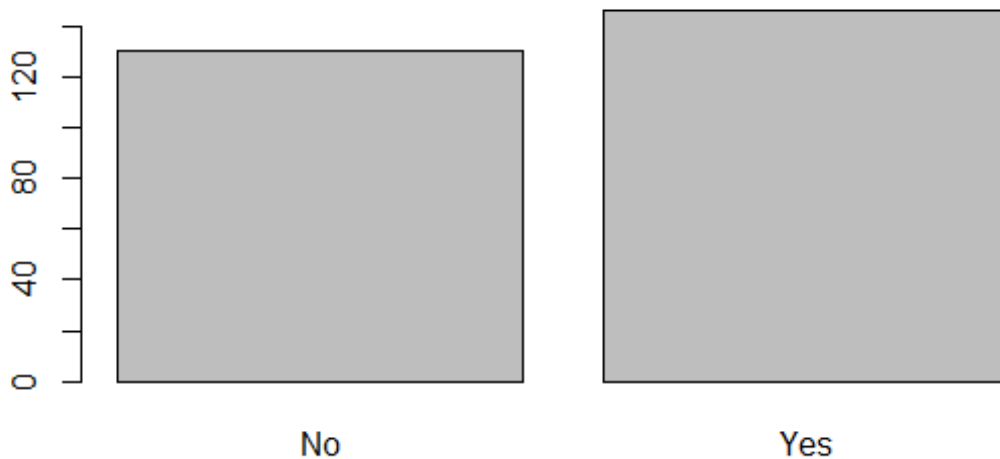


Figure 1: DecisionMale barchar

NX: Tỷ lệ nam đồng ý nhiều hơn là không đồng ý. ( $0.529 - 0.471 = 0.058$ )

- RaceF:

<sup>1</sup>Nên dùng attach(SpeedDating) để khỏi phải gõ \$ trước tên biến mỗi khi truy cập

<sup>2</sup>Nếu attach rồi thì phải detach(SpeedDating) khi đã dùng xong

```

1  tab2 = table(RaceF) #Count so luong nu cho tung chung toc
2  # Them total
3  addmargins(tab2)
4  > RaceF
5      Asian  Black Caucasian  Latino  Other  Sum
6  4    70    15    148    23    16    276
7
8  prop.table(tab2) # Proportions
9  > RaceF
10     Asian  Black Caucasian  Latino  Other
11 0.01449275 0.25362319 0.05434783 0.53623188 0.08333333 0.05797101
12
13 barplot(tab2)
14

```

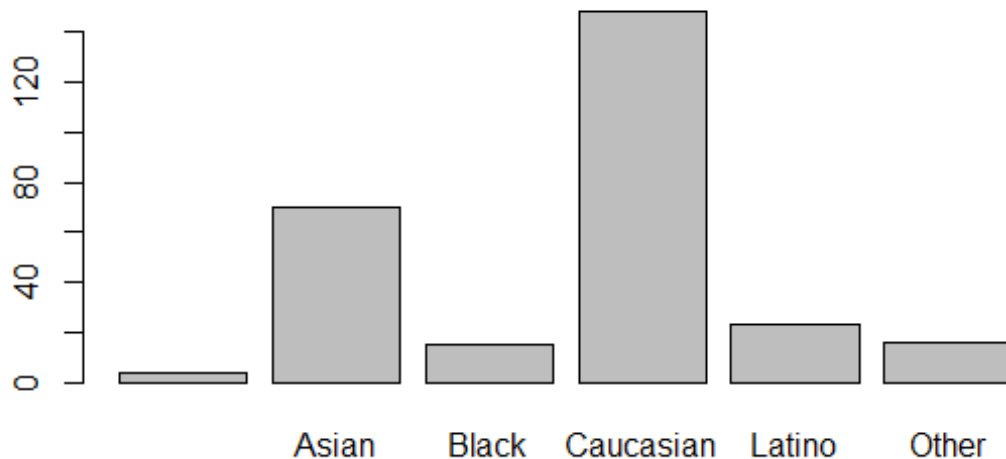


Figure 2: RaceF barchar

NX chung: Tỷ lệ nữ trắng nhiều nhất (0.536), tiếp đến là nữ châu Á (0.254).

b) Biến định lượng:

Mặc dù trung bình (mean) thường được sử dụng để mô tả trọng tâm của phân phối nhưng nó lại rất nhạy cảm với ngoại lệ (outlier). Do đó, trung vị (median) được dùng thay thế vì ít chịu ảnh hưởng của outlier.

Thêm vào đó, interquartile range (IQR):  $IQR = Q_3 - Q_1$ . Đây là đại lượng dùng để đo độ lan ra của dữ liệu (measure of spread) và ít bị ảnh hưởng bởi nhiễu.

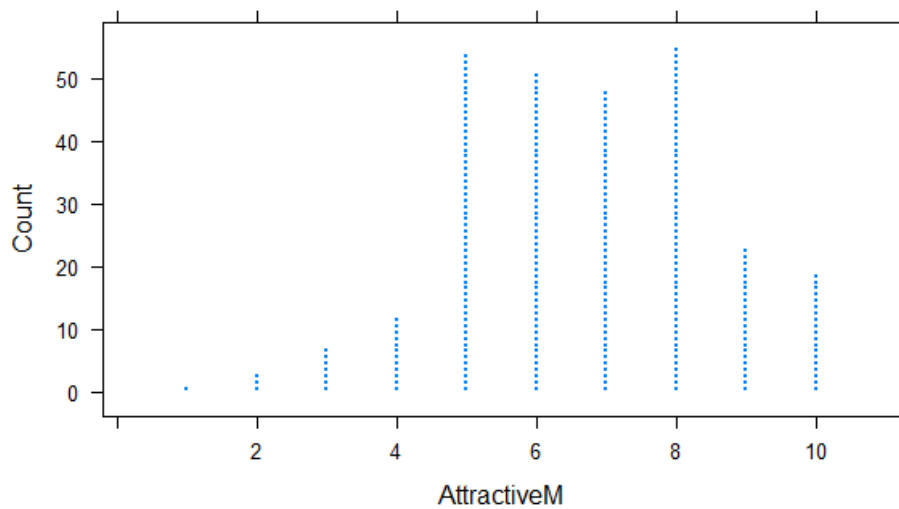
Thông thường, người ta sử dụng median kèm với IQR như là 1 measure of spread khi mà dữ liệu có nhiễu nhiều.

- AttractiveM:

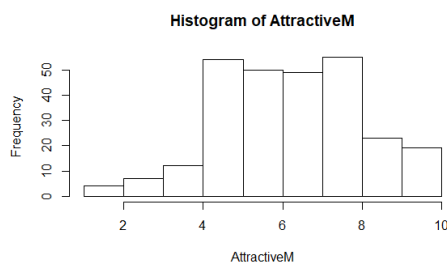
```

1      # AttractiveM
2      five_num = summary(AttractiveM) # 5-number summary
3      >
4      Min. 1st Qu. Median Mean 3rd Qu. Max. NA
5      1.000 5.000 7.000 6.687 8.000 10.000 3
6
7      # Range:
8      range = five_num[6] - five_num[1]; range
9      > 9
10
11     # Interquartile range
12     IQR = five_num[5] - five_num[2]; IQR
13     > 3
14
15     # Detect outlier: smaller than Q1 - 1.5(IQR) or greater than Q3 + 1.5(IQR)
16     (t1 <- five_num[2] - 1.5*IQR); (t2 <- five_num[5] + 1.5*IQR)
17     > 0.5
18     12.5
19
20     # Dotplot de dem so luong cho tung diem tuong ung
21     dotPlot(~AttractiveM, width = 1, cex = 0.35)
22     # Ve histogram
23     hist(AttractiveM)
24     # Ve phan bo cua du lieu
25     densityplot(AttractiveM)
26
27

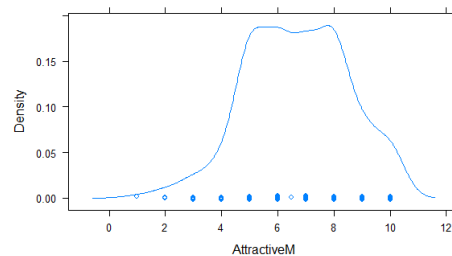
```



(a) Dotplot of AttractiveM



(b) Histogram of AttractiveM



(c) Density of AttractiveM

- Đồ thị có dạng ”xấp xỉ” bell-shape, gần đối xứng hai bên.
- 25% Nam đánh giá sức quyến rũ của bạn nữ ít nhất 1 - 5 điểm
- 50% Nam đánh giá sức quyến rũ của bạn nữ trong khoảng 5 - 8 điểm
- 25% Nam đánh giá sức quyến rũ của bạn nữ nhiều nhất từ 8 - 10 điểm
- Kết luận:
  - \* 50% Nam đánh giá sức quyến rũ của bạn nữ ít hơn 7 điểm.
  - \* 50% Nam đánh giá sức quyến rũ của bạn nữ trên 7 điểm.
- Ta thấy  $range = Max - Min = 9$ , phân bố từ 1 - 10.
- Ta tính được interquartile range (IQR):  $IQR = Q_3 - Q_1 = 3$ .
- Dựa vào IQR, ta sẽ phát hiện được phạm vi phân bố của outliers: Nhỏ hơn  $Q_1 - 1.5 * IQR = 0.5$  hay lớn hơn  $Q_3 + 1.5 * IQR = 12.5$ . Ta thấy rằng không có outlier nào lớn hơn 12.5 và cũng không có outlier nào bé hơn 0.5 trong dữ liệu. Điều này cho thấy dữ liệu đã được thu thập rất tốt.

• LikeM:

```

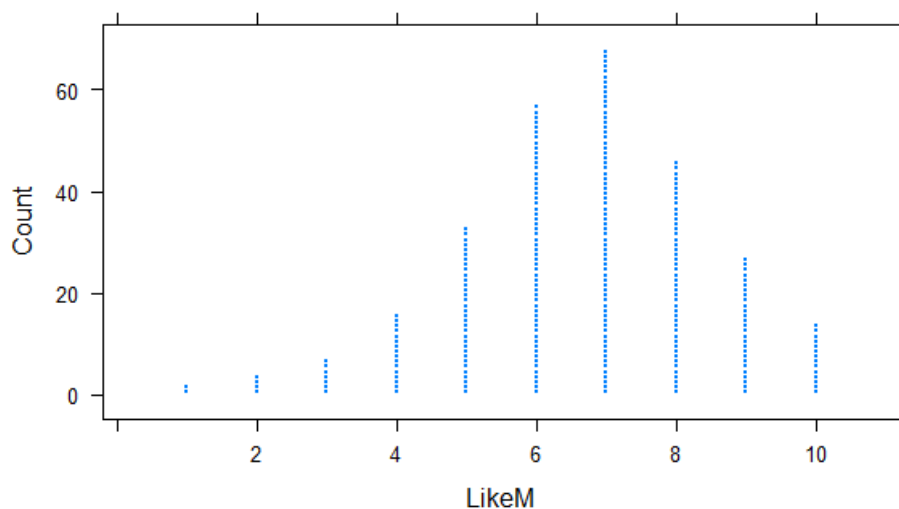
1 # LikeM
2 five_num = summary(LikeM) # 5-number summary
3 >
4 Min. 1st Qu. Median Mean 3rd Qu. Max. NA
5 1.000 6.000 7.000 6.682 8.000 10.000 2
6
7 # Range:

```

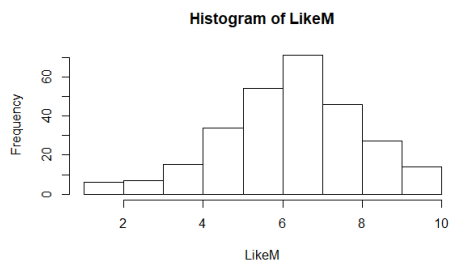
```

8  range = five_num[6] - five_num[1]; range
9  > 9
10
11  # Interquartile range
12  IQR = five_num[5] - five_num[2]; IQR
13  > 2
14
15  # Detect outlier: smaller than  $Q1 - 1.5(IQR)$  or greater than  $Q3 + 1.5(IQR)$ 
16  (t1 <- five_num[2] - 1.5*IQR); (t2 <- five_num[5] + 1.5*IQR)
17  > 3
18  11
19
20  # Tim so luong cac doi tuong outlier
21  # TH: < t1
22  count(subset(SpeedDating, LikeM < t1))
23  > 6
24
25  # Dotplot de dem so luong cho tung diem tuong ung
26  dotPlot(~LikeM, width = 1, cex = 0.35)
27  # Ve histogram
28  hist(LikeM)
29  # Ve phan bo cua du lieu
30  densityplot(LikeM)
31

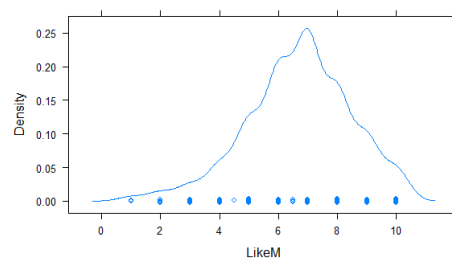
```



(a) Dotplot of LikeM



(b) Histogram of LikeM



(c) Density of LikeM



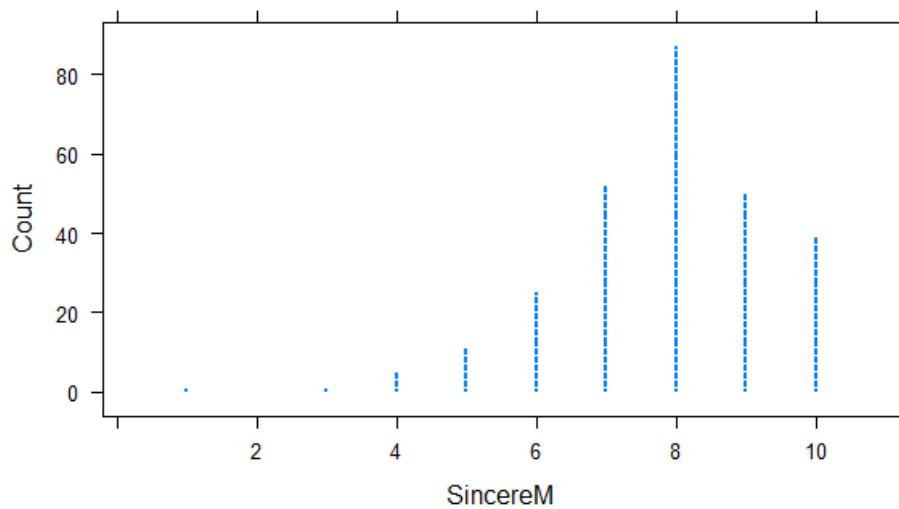
- Đồ thị có dạng bell-shape, gần đối xứng 2 bên
- 25% mức độ thích thú của nam ít nhất 1 - 6 điểm
- 50% mức độ thích thú của nam trong khoảng 6 - 8 điểm
- 25% mức độ thích thú của nam nhiều nhất từ 8 - 10 điểm
- Kết luận:
  - \* 50% mức độ thích thú của nam ít hơn 7 điểm.
  - \* 50% mức độ thích thú của nam trên 7 điểm.
- Ta thấy  $range = Max - Min = 9$ , phân bố từ 1 - 10.
- Ta tính được interquartile range (IQR):  $IQR = Q_3 - Q_1 = 2$ .
- Dựa vào IQR, ta sẽ phát hiện được phạm vi phân bố của outliers: Nhỏ hơn  $Q_1 - 1.5 * IQR = 3$  hay lớn hơn  $Q_3 + 1.5 * IQR = 11$ . Ta thấy rằng có 6 outliers dưới 3 và không có outlier nào lớn hơn 11 trong dữ liệu.

• SincereM:

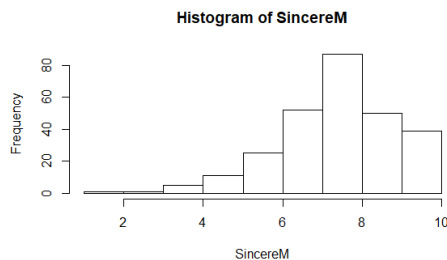
```

1  # SincereM
2  summary(SincereM) # 5-number summary
3  >
4  Min. 1st Qu. Median Mean 3rd Qu. Max. NA
5  1.000 7.000 8.000 7.856 9.000 10.000 5
6
7  # Range:
8  range = five_num[6] - five_num[1]; range
9  > 9
10
11 # Interquartile range
12 IQR = five_num[5] - five_num[2]; IQR
13 > 2
14
15 # Detect outlier: smaller than  $Q_1 - 1.5(IQR)$  or greater than  $Q_3 + 1.5(IQR)$ 
16 (t1 <- five_num[2] - 1.5*IQR); (t2 <- five_num[5] + 1.5*IQR)
17 > 4
18 12
19
20 # Tìm số lượng các đối tượng outlier
21 # TH: < t1
22 count(subset(SpeedDating, LikeM < t1))
23 > 13
24
25 # Dotplot để đếm số lượng cho từng điểm tương ứng
26 dotPlot(~SincereM, width = 1, cex = 0.35)
27 # Vẽ histogram
28 hist(SincereM)
29 # Vẽ phân bố của dữ liệu
30 densityplot(SincereM)
31

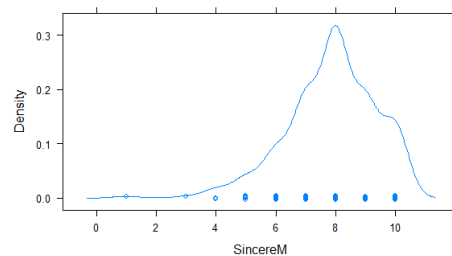
```



(a) Dotplot of SincereM



(b) Histogram of SincereM



(c) Density of SincereM

- Đồ thị hơi nghiêng về bên trái do đó  $mean < median$  ( $7.856 < 8$ )
- 25% nam đánh giá độ chân thành của nữ ít nhất 1 - 7 điểm
- 50% nam đánh giá độ chân thành của nữ trong khoảng 7 - 9 điểm
- 25% nam đánh giá độ chân thành của nữ nhiều nhất từ 9 - 10 điểm
- Kết luận:
  - \* 50% nam đánh giá độ chân thành của nữ ít hơn 8 điểm.
  - \* 50% nam đánh giá độ chân thành của nữ trên 8 điểm.
- Ta thấy  $range = Max - Min = 9$ , phân bố từ 1 - 10.
- Ta tính được interquartile range (IQR):  $IQR = Q_3 - Q_1 = 2$ .
- Dựa vào IQR, ta sẽ phát hiện được phạm vi phân bố của outliers: Nhỏ hơn  $Q_1 - 1.5 * IQR = 4$  hay lớn hơn  $Q_3 + 1.5 * IQR = 12$ . Ta thấy rằng có 13 outliers dưới 4 (khá nhiều) và không có outlier nào lớn hơn 12 trong dữ liệu.

NX cảm tính: Độ chân thành và sức quyến rũ của bạn nữ dễ ảnh hưởng đến DecisionM và LikeM của nam.

## 4 Chọn ra 2 biến định tính (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.

Chọn 2 biến định tính: DecisionMale (Yes/No), RaceF (Asian, Black, Caucasian, Latino, or Other)

```
1 # 2 biến định tính
2 tab1 = table(D DecisionMale, RaceF)
3 # Thêm margin
4 addmargins(tab1)
5
6 >
7      RaceF
8 DecisionMale Asian Black Caucasian Latino Other Sum
9   No      2    32     7     72     10   130
10  Yes      2    38     8     76     16   146
11  Sum      4    70    15    148     23   276
12
13 # 2-way table
14 # Tỷ lệ nam (yes/no) điều kiện chủng tộc nữ (Asian, Black, ...)
15 prop.table(tab1, margin = 1)
16
17 >
18      RaceF
19 DecisionMale Asian Black Caucasian Latino Other
20   No 0.01538462 0.24615385 0.05384615 0.55384615 0.05384615 0.07692308
21   Yes 0.01369863 0.26027397 0.05479452 0.52054795 0.10958904 0.04109589
22
23 # Segmented barchart
24 barplot(tab1, legend = TRUE)
25
```

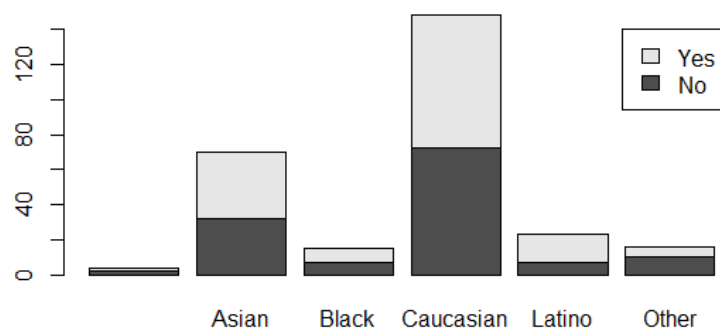


Figure 6: Segmented barchart of 2 categorical variables

- Ta thấy rằng tỷ lệ phản hồi (no/yes) của nam đối với chủng tộc nữ da trắng (Caucasian) là cao nhất (0.553, 0.52). Tiếp đến là nữ châu Á (Asian) (0.246, 0.26).

- Tỷ lệ phản hồi "yes" và "no" đối với nữ da trắng:

$$p_{yes} \approx 0.52$$

$$p_{no} \approx 0.553$$

- Tỷ lệ khác biệt (difference proportion) giữa tỷ lệ phản hồi "yes" và phản hồi "no" đối với nữ da trắng:

$$p_{yes} - p_{no} = 0.52 - 0.553 = -0.033$$

→ Tỷ lệ người nam phản hồi 'no' cao hơn phản hồi 'yes' đối với nữ da trắng 0.033.

- Tỷ lệ nam phản hồi "yes" đối với người da trắng cao hơn với người châu Á:  $0.553 - 0.246 = 0.307$
- Từ bảng 2-way table, ta thấy rằng có 4 người nữ không có chủng tộc: 2 nhận phản hồi "yes" và 2 nhận phản hồi no "no". Đây là nhóm nhận được phản hồi ít nhất, cũng dễ hiểu vì trong nhóm này chỉ có duy nhất 4 người.

NX: Người da trắng (Caucasian) và người da màu (Asian) nhận được sự phản hồi cao hơn các tộc còn lại.

## 5 Chọn ra 1 biến định tính, 1 biến định lượng (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.

Chọn 1 biến định tính và 1 biến định lượng: DecisionMale (yes/no), AttractiveM (1-10)

```

1 # 1 quantitative and 1 categorical variables
2 # statistics for the quantitative variable within each category
3 by(AttractiveM, DecisionMale, mean, na.rm=TRUE)
4 >
5 DecisionMale: No
6 [1] 5.641732
7 -----
8 DecisionMale: Yes
9 [1] 7.59589
10
11 # Tính favorite statistics
12 favstats(~AttractiveM | DecisionMale)
13 >
14 DecisionMale min Q1 median Q3 max mean sd n missing
15 1 No 1 5 5 6 10 5.641732 1.694877 127 3
16 2 Yes 5 7 8 8 10 7.595890 1.357375 146 0
17
18 # side-by-side boxplots
19 boxplot(AttractiveM ~ DecisionMale, xlab = 'DecisionMale', ylab = 'AttractiveM')
20

```

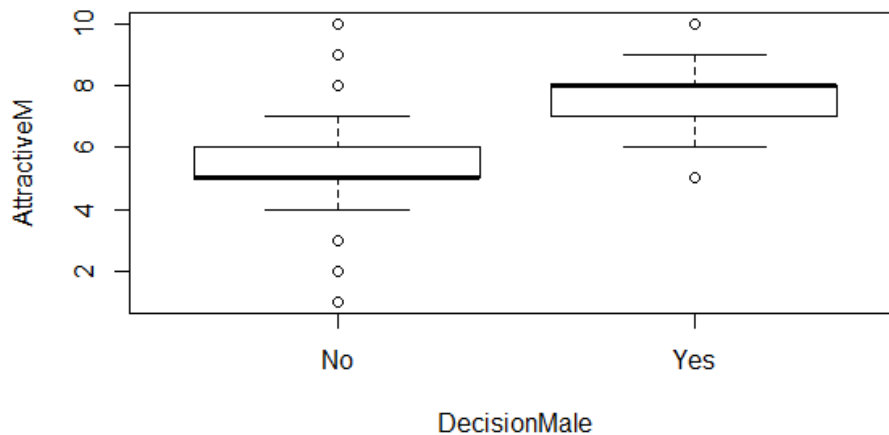


Figure 7: Side-by-side boxplots

- Ở trường hợp phản hồi "no", Ta thấy median trùng với 1st quartile bằng 5.
- Ở trường hợp phản hồi "yes", ta thấy median trùng với 3rd quartile bằng 8.
- Do  $se_{no} = 1.695 > se_{yes} = 1.357$  nên các outlier của phản hồi "no" xuất hiện nhiều hơn của phản hồi "yes". Điều này làm tần suất của dữ liệu trải dài đều từ 1 - 10.
- Ta thấy rằng có TH outlier bên phản hồi "no" có attractiveM = 10. Điều này khá bất thường. Còn bên phản hồi "yes" thì các outlier không đáng kể lắm, phân bố cũng gần IQR của dữ liệu.
- Thêm vào đó, ta thấy rằng mean của DecisionMale: No (5.641) < mean của DecisionMale: Yes (7.59). Do đó ta có thể nói mức điểm quyến rũ của phản hồi "yes" cao hơn của phản hồi "no".
- Đối với phản hồi "no", ta thấy rằng median < mean do có nhiều outlier trải dài từ 1-10. Ngược lại, phản hồi "yes" có median > mean do ít outlier hơn và phân bố của outlier cũng không quá xa IQR của dữ liệu.
- Interquartile range của phản hồi "yes" và "no" bằng nhau ( $IQR = 1$ ).
- Chúng ta thấy có 1 liên kết khi AttractiveM càng cao thì khả năng phản hồi yes cũng cao tuy rằng liên kết này không quá mạnh.

## 6 Chọn ra 2 biến định lượng (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.

Chọn 2 biến định lượng: AttractiveM (1-10) và LikeM (1-10)

```

1 # 2 quantitative variables
2 # Summary statistics: correlation, regression line
3 > cor(AttractiveM, LikeM, use = "complete.obs") # avoid missing value NA
4 [1] 0.7240187
5
6 > lm(LikeM~AttractiveM) # Linear regression for 2 variables
7
8 Call:
9 lm(formula = LikeM ~ AttractiveM)
10
11 Coefficients:
12 (Intercept) AttractiveM
13 1.9110      0.7139
14
15 # Graphical display: scatterplot
16 plot(AttractiveM, LikeM, main = "Scatter plot example", pch=19)
17 # Add fit lines
18 abline(lm(LikeM~AttractiveM), col="red") # regression line (y~x)
19
20

```

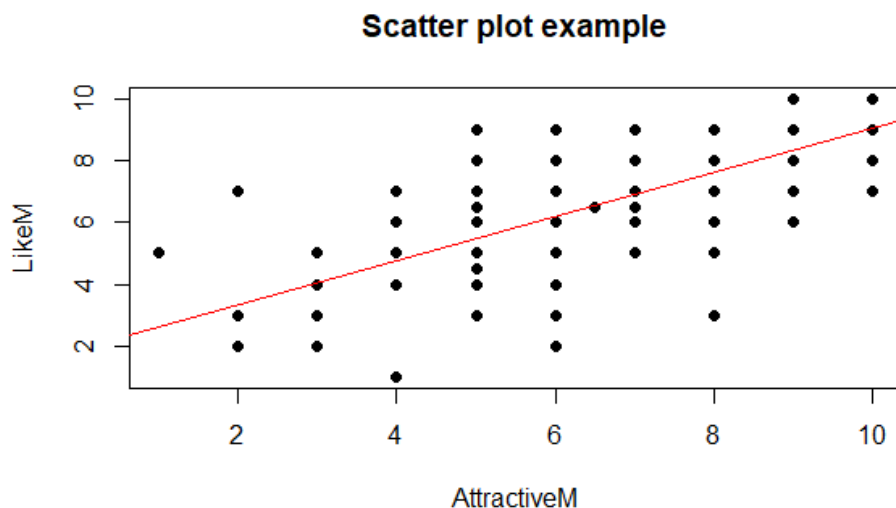


Figure 8: Scatterplot of 2 quantitative variables

- Ta thấy rằng giữa 2 biến định lượng này có liên kết dương khá mạnh (positive association) ( $r = 0.724$ ) do đó nhìn chung khi *AttractiveM* tăng thì *LikeM* cũng tăng.
- Tuy rằng, có liên kết mạnh nhưng điều này không thể khẳng định giữa *LikeM* và *AttractiveM* có quan hệ nhân quả (causation)
- Chúng ta có thể fit 1 đường thẳng (best-fit line)  $y = 1.911 + 0.7139x$  để chia tập dữ liệu thì ta thấy rằng y sẽ tăng 71.39% nếu x tăng lên 1 đơn vị.
- Cụ thể ở đây khi *AttractiveM* tăng 1 đơn vị thì *LikeM* sẽ tăng 71.39%. The intercept 1.911 chỉ rằng  $LikeM = 1.911$  nếu  $AttractiveM = 0$  nhưng hầu như rất hiếm  $AttractiveM = 0$ .

- Ta thấy dữ liệu phân bố không gần best-fit line.
- Việc dùng đường thẳng này để dự đoán cho quần thể (population) ở đây là khả thi nhưng hiệu quả không cao do dữ liệu không phân bố không gần regression line điều này dẫn đến tổng residual trung bình bình phương (square error) trên tập training set sẽ cao.

## 7 (Cộng điểm) Phân tích thăm dò quan hệ giữa nhiều hơn 2 biến (từ 5 biến quan tâm).

Multiple regression:  $\text{LikeM} \sim \text{AttractiveM} + \text{SincereM}$

3

```

1 # Multiple regression
2 > fit <- lm(LikeM~AttractiveM + SincereM)
3 > summary(fit) # show the results
4 Call:
5 lm(formula = LikeM ~ AttractiveM + SincereM)
6
7 Residuals:
8 Min    1Q  Median    3Q   Max
9 -3.9329 -0.5840  0.0905  0.7111  3.3394
10
11 Coefficients:
12 Estimate Std. Error t value Pr(>|t|)
13 (Intercept)  0.06763    0.39650   0.171   0.865
14 AttractiveM  0.62837    0.04165  15.085 < 2e-16 ***
15 SincereM     0.30639    0.05023   6.100 3.7e-09 ***
16 ---
17 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
18
19 Residual standard error: 1.149 on 267 degrees of freedom
20 (6 observations deleted due to missingness)
21 Multiple R-squared:  0.5903, Adjusted R-squared:  0.5872
22 F-statistic: 192.3 on 2 and 267 DF, p-value: < 2.2e-16
23
24 #shows the correlation coefficient of multiple variables
25 #in conjunction with a scatterplot
26 #(including a line of best fit with a confidence interval) and a density plot.
27 ggpairs(SpeedDating,
28 columns = c("AttractiveM", "SincereM", "LikeM"),
29 upper = list(continuous = wrap("cor",
30 size = 10)),
31 lower = list(continuous = "smooth"))
32
33
```

<sup>3</sup>Sử dụng hàm ggpairs của GGally package để hiển thị hệ số tương quan cũng như scatterplot cho từng cặp biến và density plot cho từng biến

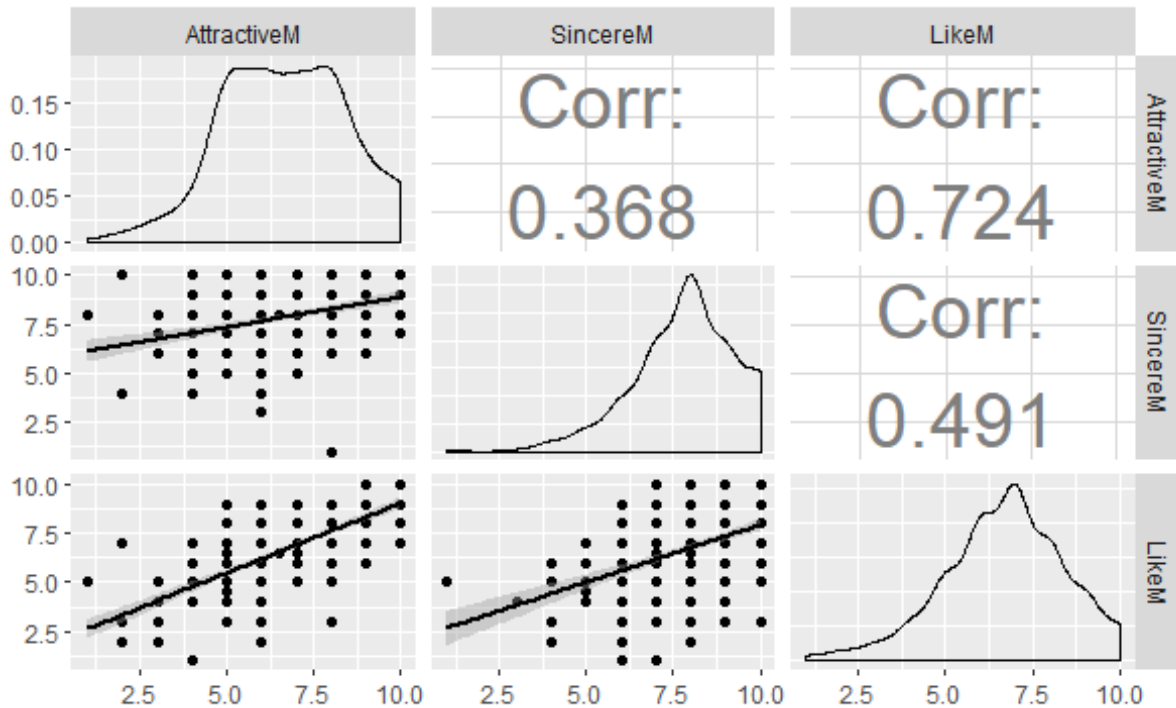


Figure 9: Multivariate plot

- Ta thấy rằng p-value của  $F - statistic < 2.2e - 16$  điều này có nghĩa là ít nhất 1 biến  $x$  thay đổi thì sẽ ảnh hưởng đến output  $y$ .
- R-squared: 59.03% điều này cho thấy mô hình này có square error khá cao do đó khả năng đem đi dự đoán cho quần thể (population) mang lại chính xác không cao.
- Đường thẳng phân tách dữ liệu của ta là:  $y = 0.067 + 0.628x_1 + 0.306x_2$
- Ta thấy cả AttractiveM và LikeM đều ảnh hưởng đáng kể đến output  $y$ . Cụ thể, ta thấy AttractiveM ảnh hưởng nhiều đến LikeM hơn là SincereM ( $0.628 > 0.306$ ).
- Ví dụ: khi AttractiveM tăng 1 thì đầu ra  $y$  sẽ tăng 0.067 lần so với khi SincereM tăng 1 thì  $y$  chỉ tăng 0.306 lần.
- Như đã phân tích ở câu 6 (trang 11), ta thấy rằng giữa AttractiveM và LikeM có liên kết dương mạnh ( $r_{AL} = 0.724$ ). Và giữa SincereM và LikeM cũng có liên kết dương trung tính ( $r_{SL} = 0.491 \approx 0.5$ ). Điều này lại một lần nữa khẳng định sự ảnh hưởng của AttractiveM, SincereM lên LikeM là đáng kể.
- Thêm vào đó, hệ số tương quan của AttractiveM và SincereM  $r_{AS} = 0.368$  không quá mạnh. Điều này cho thấy sự thay đổi của cả 2 không ảnh hưởng đến nhau nhiều.
- Ta thấy rằng các scatter plot giữa (AttractiveM, LikeM), (SincereM, LikeM) và (AttractiveM, SincereM) có phần smooth curve màu xám. Phần xám này khá to ở phần đầu khi số lượng point nhỏ và dần về sau phần xám nhỏ đi và biến mất khi số point nhiều hơn.
- Các smooth curves này giúp chúng ta thấy được số lượng các liên kết không chắc chắn (uncertain association) với regression line. Sự không chắc chắn này tăng khi dữ liệu quan sát nhỏ và giảm khi dữ liệu quan sát lớn.



## 8 Tham khảo

- [1] Lock5withR pdf-file, [Lock5withR](#).
- [2] Quick-R by DataCamp, [Quick-R](#)
- [3] Official blog - R Correlation tutorial, [DataCamp-Blog](#)
- [4] Multiple Linear Regression in R, [STHDA - Articles - Regression Analysis](#)
- [5] Book: Unlocking the power of data, Chapter 2 – Describing Data, [Robin H. Lock](#), [Patti Frazer Lock](#), [Dennis F. Lock](#), [Kari Lock Morgan](#), [Eric F. Lock](#) - Statistics: Unlocking the Power of Data (2012, Wiley)