

R によるデータ分析入門：練習問題・回答例 第2章

1. 東京城南地区および川崎市の賃貸物件データ (rent-jonan-kawasaki.csv) を用いて以下の表を作成したい.

- (1) 鉄道路線別オートロック付物件の比率

```
> prop.table(table(dataf$auto_lock, dataf$line), margin=2)
```

	JR	keikyu	tokyu
NO	0.6292683	0.6666667	0.6474576
YES	0.3707317	0.3333333	0.3525424

JR 沿線でオートロック付き物件の比率がやや高い

- (2) 鉄道沿線別の平均賃貸料, 平均築年数, 平均占有面積

```
> dataf %>% group_by(line) %>% summarize(mean(rent_total), mean(floor), mean(age))
```

```
# A tibble: 3 × 4
```

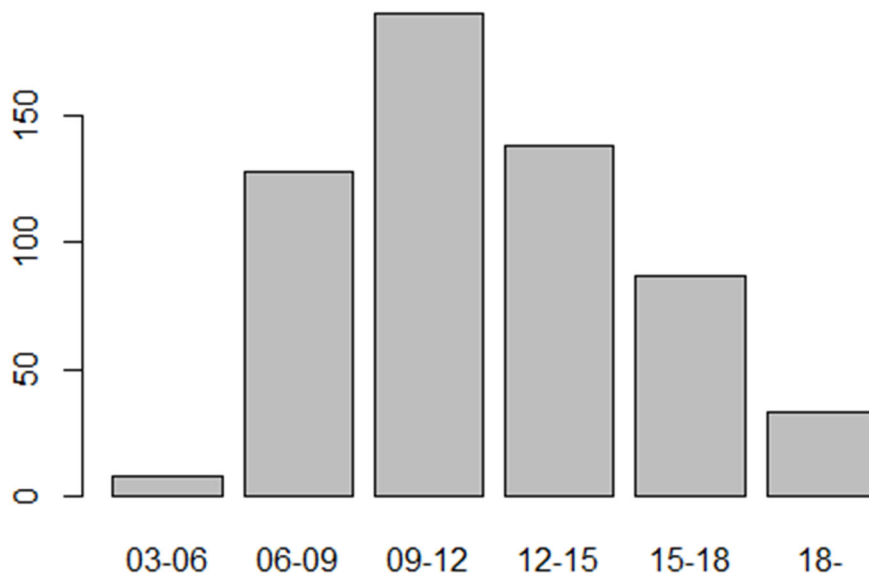
line	`mean(rent_total)`	`mean(floor)`	`mean(age)`
<chr>	<dbl>	<dbl>	<dbl>
1 JR	11.6	35.4	12.9
2 keikyu	10.7	31.3	11.6
3 tokyu	12.1	29.2	14.0

家賃は東急沿線で高い、専有面積は JR 沿線が広い、築年数は京急沿線が低い (新しい)

- (3) 賃貸料を 3 万円刻みの度数分布表を作成し, また鉄道沿線別にヒストグラムを作成せよ.

```
> table(dataf$r_category)
```

03-06	06-09	09-12	12-15	15-18	18-
8	128	190	138	87	33



(4) 東急沿線物件に限定し、賃貸料, 1 平方メートル当たり賃貸料と駅からの時間距離 (徒歩分数とバス所要時間の合計), ターミナルからの所要時間の相関係数行列 なお, 賃貸料は, 賃貸料 (rent) と管理費 (service) の合計として再定義して計算すること.

```
> dataf %>% dplyr::filter(line=="tokyu") %>% dplyr::select(rent_total, rent_floor, dist, terminal) %>%
+   cor(., use="pairwise.complete.obs")
```

	rent_total	rent_floor	dist	terminal
rent_total	1.00000000	0.1387697	-0.0281793	-0.01940692
rent_floor	0.13876975	1.0000000	-0.2541803	-0.12762813
dist	-0.02817930	-0.2541803	1.0000000	0.10841366
terminal	-0.01940692	-0.1276281	0.1084137	1.0000000

賃貸料 rent_total, 1 平方メートル当たり賃貸料 rent_floor, 駅からの時間距離 dist, ターミナルからの所要時間 terminal

2. 東京大学社会科学研究所附属社会調査・データアーカイブ研究センターは、東大社研若年者パネル調査を実施しているが、そのデータをもとにした疑似データを公開しています（非制限疑似データ）を使って以下の変数、および表を作成せよ。

(1) 調査票の学歴に関数する質問項目 ZQ23A と ZQ24 を参照して、大卒・大学院卒なら“univ”，それ以外なら“others”の変数，univ を作成せよ。また、支持政党に関する質問項目 ZQ42 を参照しながら自民党支持なら“LDP”，そうでなければ“others”をとる変数 LDP を作成せよ。また、大卒・院卒とそれ以外で、自民党を支持する人の比率を計算せよ。

univ の作成

ZQ23A が最後に通った学校で 4 と 5 が大学・大学院

ZQ24 が卒業か、在学中か中退か示す変数、1 なら卒業。

やや条件式が複雑ですが、case_when を使うと大卒・大学院卒なら“univ”,それ以外なら“others”をとる変数を以下のように作成できます。

```
dataf <- dataf %>%  
  dplyr::mutate(univ = case_when((ZQ23A == 5|ZQ23A ==6) & ZQ24 ==1~"univ",  
                                (ZQ23A != 5&ZQ23A !=6) | ZQ24 !=1~"others"))
```

3 章で紹介する if_else を使うとよりシンプルに書けます

```
dataf <- dataf %>%  
  dplyr::mutate(univ_ifelse = if_else((ZQ23A == 5|ZQ23A ==6) & ZQ24  
==1,"univ","others"))
```

LDP の作成：自民党支持なら“LDP”、それ以外なら“others”をとる変数、

ZQ42 が 1 なら自民支持、ZQ42 > 1 ならそれ以外なので case_when を使うと以下の通り。

```
dataf <- dataf %>%  
  dplyr::mutate(LDP = case_when(ZQ42 == 1~"LDP",  
                                ZQ42>1~"others"))
```

こちらも if_else を使うとよりシンプルに書けます

```
dataf <- dataf %>%  
  dplyr::mutate(LDP_ifelse = if_else(ZQ42 == 1,"LDP","others"))
```

```
> prop.table(table(dataf$univ, dataf$LDP), margin=1)
```

	LDP	others
others	0.1632373	0.8367627
univ	0.2140221	0.7859779

大卒者の自民党支持率は 21.4%、それ以外の人 16.3%で大卒者のほうが自民党支持率が高い

(2) 「普段収入になる仕事している人」(ZQ03==1) かつ「既婚(配偶者)」(ZQ50==2) の「男性」(sex==1) に限定したデータフレームを作成し、大卒・院卒とそれ以外で家事をする人の比率を比べたい。家事については ZQ54A~ZQ54D に注目して「食事の用意」, 「選択」, 「家の掃除」, 「日用品・食用品の買い物」を「毎日する」人とそれ以外の人に分けて、大卒・院卒×家事の実施の有無の比率の表を作成せよ。

```
> # ZQ54A 食事の用意
```

```
> prop.table(table(dataf2$hwork1, dataf2$univ), margin=2)
```

	others	univ
No	0.2558140	0.1884058
Yes	0.7441860	0.8115942

```
> # ZQ54B 洗濯
```

```
> prop.table(table(dataf2$hwork2, dataf2$univ), margin=2)
```

	others	univ
No	0.3837209	0.4927536
Yes	0.6162791	0.5072464

```
> # ZQ54C 家の掃除
```

```
> prop.table(table(dataf2$hwork3, dataf2$univ), margin=2)
```

	others	univ
No	0.5697674	0.7246377
Yes	0.4302326	0.2753623

```
> # ZQ54D 日用品・食料品の買い物
```

```
> prop.table(table(dataf2$hwork4, dataf2$univ), margin=2)
```

	others	univ
--	--------	------

No 0.8372093 0.8115942

Yes 0.1627907 0.1884058

「毎日する人」の比率が大卒者の方が高いのは、食事の用意と日用品・食料品の用意である。
選択と家の掃除を毎日する人の比率は大卒者の方が低い。