IS607 Assignment 7

Derek G. Nokes

Saturday, March 21, 2015

Contents

```
## Warning: package 'tidyr' was built under R version 3.1.3
```

```
## Warning: package 'dplyr' was built under R version 3.1.3
```

1. Write down 3 questions that you might want to answer based on this data.

1. What proportion of the total number of people surveyed across both cities were in each city?
2. Was the proportion of 'no' responses higher in the '16-24' or '+25' age group?
3. What was the proportion of responses in Glasgow by age group and response type?

2. Create an R data frame with 2 observations to store this data in its current "messy" state. Use whatever method you want to re-create and/or load the data.

We create a data frame with the data in its messy state.

```
# create the messy data set
city<-c('Edinburgh','Edinburgh','Glasgow','Glasgow')
age<-c('16-24','+25','16-24','+25')
yes<-c(80100,143000,99400,150400)
no<-c(143000,214800,43000,207000)
messy_t<-data.frame(city=city,age=age,yes=yes,no=no)
# create the transposed messy table
knitr::kable(messy_t, caption = 'Transposed Messy Data set')
```

city	age	yes	no
Edinburgh	16-24	80100	143000
Edinburgh	+25	143000	214800
Glasgow	16-24	99400	43000
Glasgow	+25	150400	207000

Table 1: Transposed Messy Data set

Notice that the data is transposed, but otherwise the same as the original table

3. Use the functionality in the tidyr package to convert the data frame to be "tidy data."

In a tidy dataset, each variable forms a column, each observation forms a row, and each type of observational unit forms a table.

In our messy dataset above, our data is tabular (probably designed for presentation). Column headers are values, not variable names (i.e., the cities - Edinburgh and Glasgow - where the survey takes place are headers rather than values of the variable 'city'). Multiple variables are stored in one column (i.e., the age and frequency of a response). Variables are also stored in both rows and columns.

We tidy the messy dataset as follows:

```r
# tidy the data set by reshaping
tidy<-messy_t %>% gather(response,frequency,yes:no)
# create the tidy table
knitr::kable(tidy, caption = 'Tidy Data set')
```

city	age	response	frequency
Edinburgh	16-24	yes	80100
Edinburgh	+25	yes	143000
Glasgow	16-24	yes	99400
Glasgow	+25	yes	150400
Edinburgh	16-24	no	143000
Edinburgh	+25	no	214800
Glasgow	16-24	no	43000
Glasgow	+25	no	207000

Table 2: Tidy Data set

Each row represents an observation, the response of the survey in one city, age group, and response type. Each column is a variable (i.e., city, age group, response type, and frequency of response).

4. Use the functionality in the dplyr package to answer the questions that you asked in step 1.

1. What proportion of the total number of people surveyed across both cities were in each city?

```r
# compute the total number of responses by city and response type
responses_by_city<-tidy %>%  group_by(city) %>%
  summarise(sum(frequency, na.rm = TRUE))
# add column names
colnames(responses_by_city)<-c('city','number')
# add the proportion
responses_by_city<-mutate(responses_by_city,proportion=number/sum(number))
# display the table
knitr::kable(responses_by_city, caption = 'Responses By City')
```

city	number	proportion
Edinburgh	580900	0.537522
Glasgow	499800	0.462478

Table 3: Responses By City

2. Was the proportion of 'no' responses higher in the '16-24' or '+25' age group?

```r
#
total_responses<-tidy %>%  group_by(age,response) %>%
  summarise(sum(frequency, na.rm = TRUE))
# name the columns
colnames(total_responses)<-c('age','response','frequency')
# add the proportions
total_proportion<-mutate(total_responses,proportion=frequency/sum(frequency))
# display the table
knitr::kable(total_proportion, caption = 'Responses By Age And Type')
```

age	response	frequency	proportion
+25	yes	293400	0.4102349
+25	no	421800	0.5897651
16-24	yes	179500	0.4911081
16-24	no	186000	0.5088919

Table 4: Responses By Age And Type

We can see that 58.9765101 % of '+25' responded 'no', compared to 50.8891929 % of '16-24'.

3. What was the proportion of responses in Glasgow by age group and response type?

```r
# compute the proportion of responses by type and age group as a function the
# total Glasgow responses
response_proportion_glasgow<-select(mutate(filter(tidy,city == 'Glasgow'),
                                    proportion=frequency/sum(frequency)),age,
                                    response,frequency,proportion)
# display table
knitr::kable(response_proportion_glasgow, caption = 'Glasgow Responses By Age And Type')
```

age	response	frequency	proportion
16-24	yes	99400	0.1988796
+25	yes	150400	0.3009204
16-24	no	43000	0.0860344
+25	no	207000	0.4141657

Table 5: Glasgow Responses By Age And Type

5. Having gone through the process, would you ask different questions and/or change the way that you structured your data frame?

I would not ask different questions and/or change the way that I structured the data frame.