Lab 8

Ramin Jabbarialghanab

October 27, 2017

Using your own dataset (which may include more than one table) carry out the following data cleaning steps. Knit together the PDF document and commit both the Lab 8 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

Before you begin: as many of you have large datasets, you're going to want to select only the variables you're interested in utilizing for this project (ideally no more than twenty columns but perhaps much smaller) so you don't have R Studio's memory working on the entire dataset. The example code provided below can be modified to allow you to subset your data to only the variables you wish to use. First, read in your complete dataset and save it as `data`. Then, add the names of the variables you wish to use for your poster project to the `select` function, separated by commas. Run the two lines of code to save this new, smaller version of your data to `data_subset`. Use this smaller dataset to complete the rest of the lab

```
# Read in your data with the appropriate function
library(readxl)
unemployment_men <- read_excel("~/Desktop/Autumn 2017/Statistics 321/unemployment rate/unemployment_men
    skip = 2)

unemployment_women <- read_excel("~/Desktop/Autumn 2017/Statistics 321/unemployment rate/unemployment_w
    skip = 2)

unemployment_total <- read_excel("~/Desktop/Autumn 2017/Statistics 321/unemployment rate/unemployment_t
    skip = 2)
```

1. To get a feel for its structure, look at the class, dimensions, column names, structure, and basic summary statistics of your data.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
class(unemployment_men)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
dim(unemployment_men)
```

```
## [1] 32 10
```

```
names(unemployment_men)
```

```
##  [1] "Province | Year" "2001"            "2006"
##  [4] "2008"            "2009"            "2010"
```

```
##  [7] "2011"            "2012"            "2013"
## [10] "2014"
```

```
str(unemployment_men)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    32 obs. of  10 variables:
##  $ Province | Year: chr  "Total" "East Azarbayejan" "West Azarbayejan" "Ardebil" ...
##  $ 2001           : chr  "13.2" "-" "-" "-" ...
##  $ 2006           : chr  "10" "5.3" "10.8" "11.1" ...
##  $ 2008           : chr  "9.1" "6.1" "10.7" "9.8000000000000007" ...
##  $ 2009           : chr  "10.8" "9.5" "11" "11.3" ...
##  $ 2010           : chr  "11.9" "10.1" "12.4" "12.9" ...
##  $ 2011           : num  10.5 8 13 12.5 10.8 16.3 12.6 10.7 9.6 12.1 ...
##  $ 2012           : num  10.4 12.1 11.4 12.2 11.1 12.3 13.9 10.9 9.7 8.9 ...
##  $ 2013           : num  8.6 7.8 9.1 11.4 9.2 8.2 10.9 8.6 7.6 9 ...
##  $ 2014           : num  8.8 6.8 9.7 10.6 9.9 9.7 8.3 8.3 6.5 13.1 ...
```

```
summary(unemployment_men)
```

```
##  Province | Year        2001              2006
##  Length:32          Length:32          Length:32
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      2008               2009               2010              2011
##  Length:32          Length:32          Length:32          Min.   : 4.500
##  Class :character   Class :character   Class :character   1st Qu.: 8.675
##  Mode  :character   Mode  :character   Mode  :character   Median :10.450
##                                                           Mean   :10.525
##                                                           3rd Qu.:12.525
##                                                           Max.   :17.300
##       2012            2013             2014
##  Min.   : 5.60   Min.   : 4.000   Min.   : 5.600
##  1st Qu.: 8.80   1st Qu.: 7.150   1st Qu.: 8.250
##  Median : 9.85   Median : 8.600   Median : 8.950
##  Mean   :10.31   Mean   : 8.866   Mean   : 9.306
##  3rd Qu.:12.12   3rd Qu.:10.825   3rd Qu.:10.375
##  Max.   :18.10   Max.   :15.700   Max.   :14.300
```

```
glimpse(unemployment_men)
```

```
## Observations: 32
## Variables: 10
## $ `Province | Year` <chr> "Total", "East Azarbayejan", "West Azarbayej...
## $ `2001`            <chr> "13.2", "-", "-", "-", "-", "-", "-", "-", "...
## $ `2006`            <chr> "10", "5.3", "10.8", "11.1", "9.6", "-", "12...
## $ `2008`            <chr> "9.1", "6.1", "10.7", "9.8000000000000007", ...
## $ `2009`            <chr> "10.8", "9.5", "11", "11.3", "9.9", "-", "11...
## $ `2010`            <chr> "11.9", "10.1", "12.4", "12.9", "13.1", "-",...
## $ `2011`            <dbl> 10.5, 8.0, 13.0, 12.5, 10.8, 16.3, 12.6, 10....
## $ `2012`            <dbl> 10.4, 12.1, 11.4, 12.2, 11.1, 12.3, 13.9, 10...
## $ `2013`            <dbl> 8.6, 7.8, 9.1, 11.4, 9.2, 8.2, 10.9, 8.6, 7....
## $ `2014`            <dbl> 8.8, 6.8, 9.7, 10.6, 9.9, 9.7, 8.3, 8.3, 6.5...
```

```
class(unemployment_women)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
dim(unemployment_women)
```

```
## [1] 32 10
```

```
names(unemployment_women)
```

```
##  [1] "Province | Year" "2001"            "2006"
##  [4] "2008"            "2009"            "2010"
##  [7] "2011"            "2012"            "2013"
## [10] "2014"
```

```
str(unemployment_women)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    32 obs. of  10 variables:
##  $ Province | Year: chr  "Total" "East Azarbayejan" "West Azarbayejan" "Ardebil" ...
##  $ 2001           : chr  "19.899999999999999" "-" "-" "-" ...
##  $ 2006           : chr  "16.2" "5.4" "7.3" "11.4" ...
##  $ 2008           : chr  "16.7" "9.1999999999999993" "9.1" "10.4" ...
##  $ 2009           : chr  "16.8" "11.6" "9.6" "13.8" ...
##  $ 2010           : chr  "20.5" "17.2" "12.4" "18.2" ...
##  $ 2011           : num  20.9 12.1 12.9 13.5 24.1 36.4 27.5 12.6 21 20.4 ...
##  $ 2012           : num  19.7 13.5 9.4 17.1 25 30.8 31.5 15.3 20.5 21.3 ...
##  $ 2013           : num  19.8 17 9.9 15.8 18.4 27 26.8 10.9 23.1 22.3 ...
##  $ 2014           : num  19.7 11.4 11.1 14 23.2 24.5 21 12.7 18.1 25.2 ...
```

```
summary(unemployment_women)
```

```
##  Province | Year        2001              2006
##  Length:32          Length:32          Length:32
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      2008               2009               2010                2011
##  Length:32          Length:32          Length:32          Min.   : 7.60
##  Class :character   Class :character   Class :character   1st Qu.:14.32
##  Mode  :character   Mode  :character   Mode  :character   Median :19.55
##                                                           Mean   :20.77
##                                                           3rd Qu.:24.70
##                                                           Max.   :42.50
##       2012            2013            2014
##  Min.   : 8.60   Min.   : 6.20   Min.   :10.70
##  1st Qu.:16.20   1st Qu.:15.55   1st Qu.:14.00
##  Median :20.10   Median :17.80   Median :19.40
##  Mean   :20.47   Mean   :19.14   Mean   :19.63
##  3rd Qu.:24.38   3rd Qu.:23.30   3rd Qu.:22.60
##  Max.   :44.40   Max.   :37.30   Max.   :35.70
```

```
library(dplyr)
glimpse(unemployment_women)
```

```
## Observations: 32
```

```
## Variables: 10
## $ `Province | Year` <chr> "Total", "East Azarbayejan", "West Azarbayej...
## $ `2001`            <chr> "19.899999999999999", "-", "-", "-", "-", "-...
## $ `2006`            <chr> "16.2", "5.4", "7.3", "11.4", "16", "-", "18...
## $ `2008`            <chr> "16.7", "9.1999999999999993", "9.1", "10.4",...
## $ `2009`            <chr> "16.8", "11.6", "9.6", "13.8", "20.7", "-", ...
## $ `2010`            <chr> "20.5", "17.2", "12.4", "18.2", "25.1", "-",...
## $ `2011`            <dbl> 20.9, 12.1, 12.9, 13.5, 24.1, 36.4, 27.5, 12...
## $ `2012`            <dbl> 19.7, 13.5, 9.4, 17.1, 25.0, 30.8, 31.5, 15....
## $ `2013`            <dbl> 19.8, 17.0, 9.9, 15.8, 18.4, 27.0, 26.8, 10....
## $ `2014`            <dbl> 19.7, 11.4, 11.1, 14.0, 23.2, 24.5, 21.0, 12...
```

```
class(unemployment_total)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
dim(unemployment_total)
```

```
## [1] 32 10
```

```
names(unemployment_total)
```

```
##  [1] "Province | Year" "2001"            "2006"
##  [4] "2008"            "2009"            "2010"
##  [7] "2011"            "2012"            "2013"
## [10] "2014"
```

```
str(unemployment_total)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    32 obs. of  10 variables:
##  $ Province | Year: chr  "Total" "East Azarbayejan" "West Azarbayejan" "Ardebil" ...
##  $ 2001           : chr  "14.2" "6.7" "10.6" "11.6" ...
##  $ 2006           : chr  "11.3" "5.3" "10" "11.1" ...
##  $ 2008           : chr  "10.4" "6.8" "10.3" "9.9" ...
##  $ 2009           : chr  "11.9" "10" "10.7" "12" ...
##  $ 2010           : chr  "13.5" "11.7" "12.4" "14.2" ...
##  $ 2011           : num  12.3 8.8 13 12.7 13.2 19.3 15.7 11 11.3 13.3 ...
##  $ 2012           : num  12.1 12.4 11 13.3 13.7 14.9 17.2 11.6 11.6 10.8 ...
##  $ 2013           : num  10.4 9.6 9.3 12.3 10.9 10.7 13.8 9 9.9 10.7 ...
##  $ 2014           : num  10.6 7.8 9.9 11.3 12.4 11.7 11.1 9 8.3 15 ...
```

```
summary(unemployment_total)
```

```
##  Province | Year        2001              2006
##  Length:32          Length:32          Length:32
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      2008              2009              2010              2011
##  Length:32          Length:32          Length:32          Min.   : 6.00
##  Class :character   Class :character   Class :character   1st Qu.:10.12
##  Mode  :character   Mode  :character   Mode  :character   Median :12.10
##                                                           Mean   :12.22
##                                                           3rd Qu.:13.47
##                                                           Max.   :19.30
##       2012            2013            2014
```

```
##  Min.   : 6.30   Min.   : 5.800   Min.   : 6.90
##  1st Qu.:10.28   1st Qu.: 7.975   1st Qu.: 9.15
##  Median :11.55   Median :10.350   Median :11.00
##  Mean   :11.98   Mean   :10.516   Mean   :10.94
##  3rd Qu.:13.40   3rd Qu.:12.575   3rd Qu.:12.40
##  Max.   :20.00   Max.   :17.100   Max.   :15.70
```

```
library(dplyr)
glimpse(unemployment_total)
```

```
## Observations: 32
## Variables: 10
## $ `Province | Year` <chr> "Total", "East Azarbayejan", "West Azarbayej...
## $ `2001`            <chr> "14.2", "6.7", "10.6", "11.6", "13.1", "-", ...
## $ `2006`            <chr> "11.3", "5.3", "10", "11.1", "11", "-", "13....
## $ `2008`            <chr> "10.4", "6.8", "10.3", "9.9", "9.4", "-", "1...
## $ `2009`            <chr> "11.9", "10", "10.7", "12", "12", "-", "12.6...
## $ `2010`            <chr> "13.5", "11.7", "12.4", "14.2", "15.3", "-",...
## $ `2011`            <dbl> 12.3, 8.8, 13.0, 12.7, 13.2, 19.3, 15.7, 11....
## $ `2012`            <dbl> 12.1, 12.4, 11.0, 13.3, 13.7, 14.9, 17.2, 11...
## $ `2013`            <dbl> 10.4, 9.6, 9.3, 12.3, 10.9, 10.7, 13.8, 9.0,...
## $ `2014`            <dbl> 10.6, 7.8, 9.9, 11.3, 12.4, 11.7, 11.1, 9.0,...
```

2. Preview the first and last 15 rows of your data. Is you dataset tidy? If not, what principles of tidy data does it seem to be violating?

```
head(unemployment_men, n = 15)
```

```
## # A tibble: 15 x 10
##           `Province | Year` `2001`             `2006`             `2008`
##                        <chr>  <chr>               <chr>              <chr>
##  1                   Total   13.2                  10                9.1
##  2         East Azarbayejan      -                 5.3                6.1
##  3         West Azarbayejan      -                10.8               10.7
##  4                  Ardebil      -                11.1 9.8000000000000007
##  5                  Esfahan      -                 9.6                  8
##  6                   Alborz      -                   -                  -
##  7                     Ilam      -                  12               11.3
##  8                  Bushehr      - 9.8000000000000007                 10
##  9                   Tehran      -                10.9 9.3000000000000007
## 10 Chaharmahal & Bakhtiyari      -                11.7               11.3
## 11           South Khorasan      -                 9.1                7.7
## 12        Khorasan-e-Razavi      -                 7.7                7.9
## 13           North Khorasan      -                 5.2                6.1
## 14                Khuzestan      -                11.1 10.199999999999999
## 15                   Zanjan      -                10.6                8.5
## # ... with 6 more variables: `2009` <chr>, `2010` <chr>, `2011` <dbl>,
## #   `2012` <dbl>, `2013` <dbl>, `2014` <dbl>
```

```
tail(unemployment_men, n = 15)
```

```
## # A tibble: 15 x 10
##           `Province | Year` `2001`             `2006`             `2008`
##                        <chr>  <chr>               <chr>              <chr>
##  1                    Fars      -                12.3                 10
##  2                  Qazvin      - 9.8000000000000007 8.1999999999999993
##  3                     Qom      -                10.8                8.4
```

```
##  4               Kordestan      -               10.8               12.5
##  5                  Kerman      -                 10                7.5
##  6              Kermanshah      -               15.2               11.1
##  7 Kohgiluyeh & Boyerahmad      -               13.4               10.8
##  8                Golestan      -                7.9                6.4
##  9                   Gilan      -                9.1               10.6
## 10                Lorestan      -               14.6               13.6
## 11              Mazandaran      -                5.7                5.5
## 12                 Markazi      -               11.1               10.7
## 13              Hormozgan      -                7.6                  8
## 14                 Hamedan      -               13.4               13.5
## 15                    Yazd      -                5.9                5.9
## # ... with 6 more variables: `2009` <chr>, `2010` <chr>, `2011` <dbl>,
## #   `2012` <dbl>, `2013` <dbl>, `2014` <dbl>
```

3. Create a histogram for at least two variables you plan to focus on for your study. Describe what these plots show you about these variables.

```
unemployment_total
```

```
## # A tibble: 32 x 10
##           `Province | Year`             `2001` `2006` `2008` `2009` `2010`
##                       <chr>              <chr>  <chr>  <chr>  <chr>  <chr>
##  1                  Total               14.2   11.3   10.4   11.9   13.5
##  2         East Azarbayejan              6.7    5.3    6.8     10   11.7
##  3         West Azarbayejan             10.6     10   10.3   10.7   12.4
##  4                 Ardebil              11.6   11.1    9.9     12   14.2
##  5                 Esfahan              13.1     11    9.4     12   15.3
##  6                  Alborz                 -      -      -      -      -
##  7                    Ilam 17.100000000000001   13.6   14.6   12.6   15.8
##  8                 Bushehr                12   10.5   10.7   11.7   13.3
##  9                  Tehran              12.2     13     11   11.9   14.2
## 10 Chaharmahal & Bakhtiyari              12.4   12.5   14.1      7   13.6
## # ... with 22 more rows, and 4 more variables: `2011` <dbl>, `2012` <dbl>,
## #   `2013` <dbl>, `2014` <dbl>
```

```
as.numeric(unemployment_total$`2006`)
```

```
## Warning: NAs introduced by coercion
```

```
##  [1] 11.3  5.3 10.0 11.1 11.0   NA 13.6 10.5 13.0 12.5 11.1  8.6  7.0 12.9
## [15] 11.7 10.6 10.9 13.7 10.3 11.1 10.7 13.4 16.6 15.6  9.0 11.4 16.2  8.0
## [29] 12.5  7.7 13.5  7.4
```

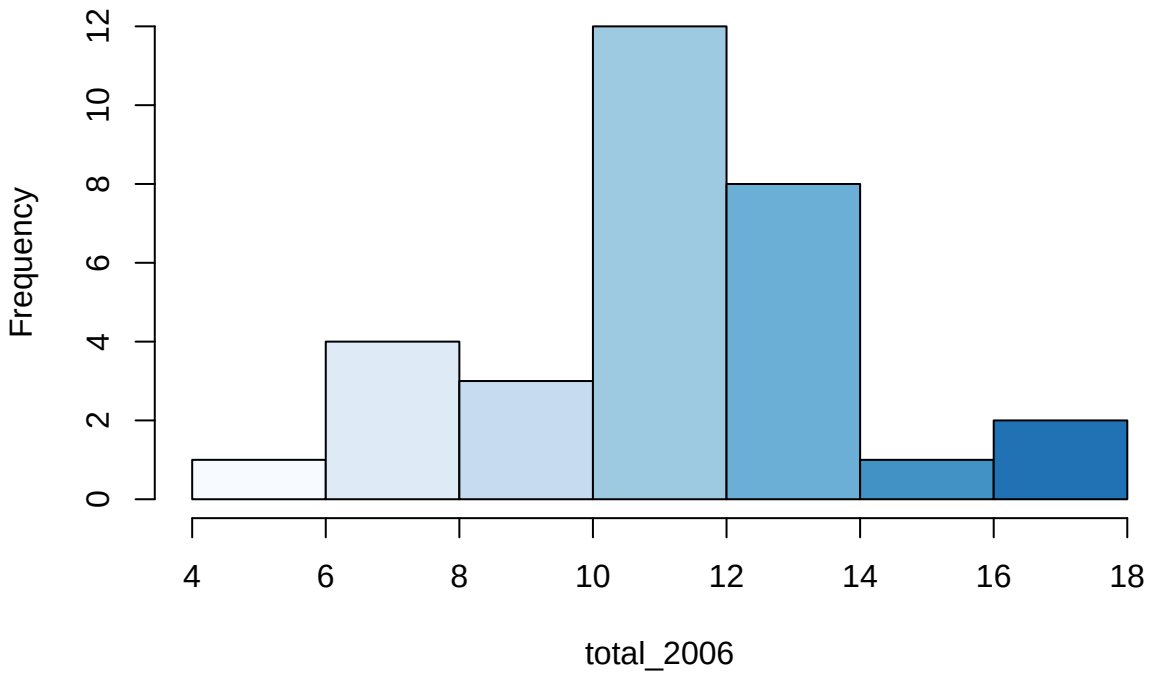
```
total_2006 <- as.numeric(unemployment_total$`2006`)
```

```
## Warning: NAs introduced by coercion
```

```
as.numeric(2006)
```

```
## [1] 2006
```

```
hist(total_2006, col = blues9)
```




```
unemployment_women
```

```
## # A tibble: 32 x 10
##           `Province | Year`             `2001`             `2006`
##                       <chr>              <chr>              <chr>
##  1                   Total 19.899999999999999               16.2
##  2         East Azarbayejan                  -                5.4
##  3         West Azarbayejan                  -                7.3
##  4                 Ardebil                   -               11.4
##  5                 Esfahan                   -                 16
##  6                  Alborz                   -                  -
##  7                    Ilam                   -               18.8
##  8                 Bushehr                   -               14.5
##  9                  Tehran                   -                 24
## 10 Chaharmahal & Bakhtiyari                  - 16.399999999999999
## # ... with 22 more rows, and 7 more variables: `2008` <chr>, `2009` <chr>,
## #   `2010` <chr>, `2011` <dbl>, `2012` <dbl>, `2013` <dbl>, `2014` <dbl>
```

```
as.numeric(unemployment_women$`2006`)
```

```
## Warning: NAs introduced by coercion
```

```
##  [1] 16.2  5.4  7.3 11.4 16.0   NA 18.8 14.5 24.0 16.4 14.9 12.2 13.8 23.1
## [15] 14.7 20.6  8.0 20.6 12.5 13.6 10.2 24.3 22.5 24.6 12.1 17.9 23.6 17.2
## [29] 19.8  8.4 13.9 12.7
```

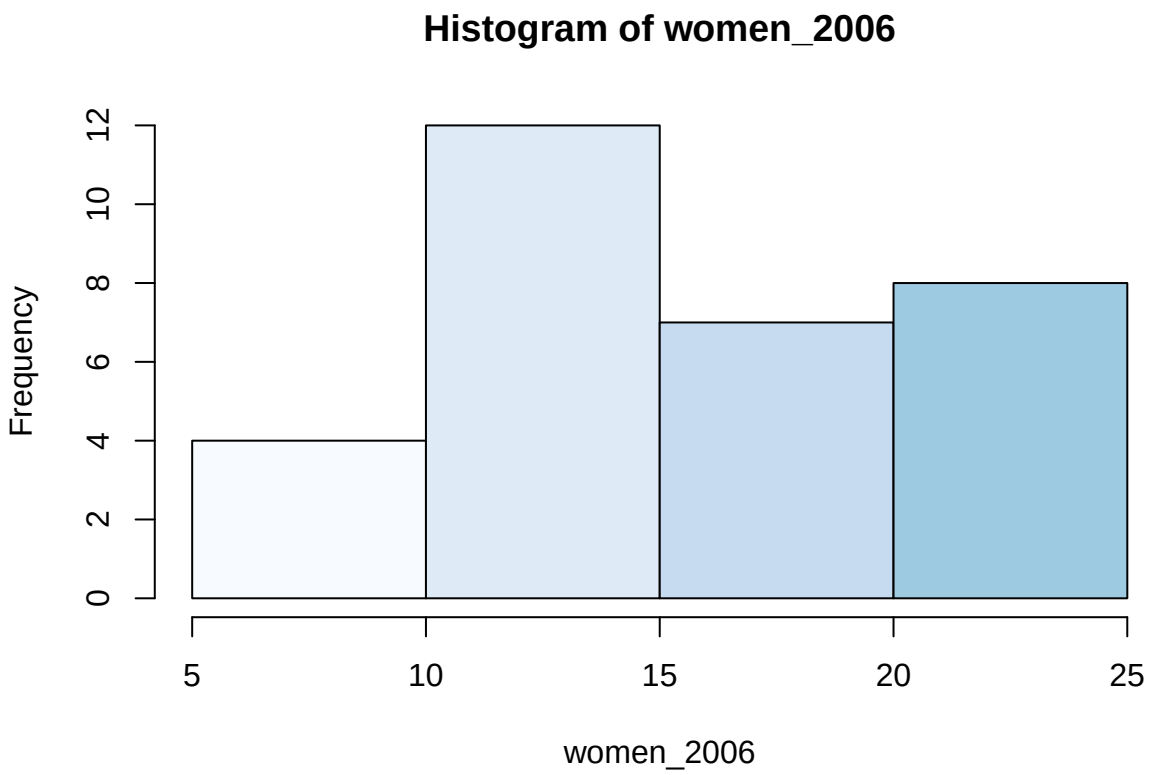
```
women_2006 <- as.numeric(unemployment_women$`2006`)
```

```
## Warning: NAs introduced by coercion
```

```
as.numeric(2006)
```

```
## [1] 2006
```

```r
hist(women_2006, col = blues9)
```




```r
unemployment_men
```

```
## # A tibble: 32 x 10
##           `Province | Year` `2001`             `2006`             `2008`
##                       <chr>  <chr>               <chr>              <chr>
##  1                    Total   13.2                  10                9.1
##  2         East Azarbayejan      -                 5.3                6.1
##  3         West Azarbayejan      -                10.8               10.7
##  4                  Ardebil      -                11.1 9.8000000000000007
##  5                  Esfahan      -                 9.6                  8
##  6                   Alborz      -                   -                  -
##  7                     Ilam      -                  12               11.3
##  8                  Bushehr      - 9.8000000000000007                 10
##  9                   Tehran      -                10.9 9.3000000000000007
## 10 Chaharmahal & Bakhtiyari      -                11.7               11.3
## # ... with 22 more rows, and 6 more variables: `2009` <chr>, `2010` <chr>,
## #   `2011` <dbl>, `2012` <dbl>, `2013` <dbl>, `2014` <dbl>
```

```r
as.numeric(unemployment_men$`2006`)
```

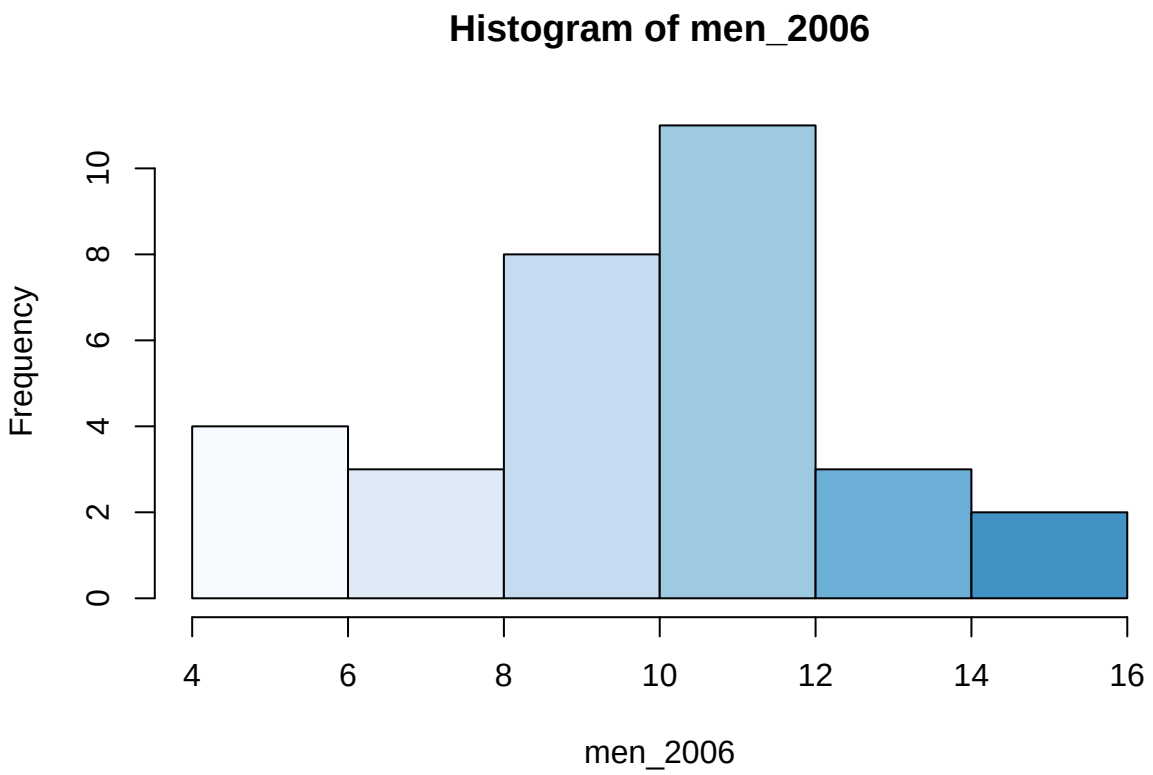
```
## Warning: NAs introduced by coercion
```

```
##  [1] 10.0  5.3 10.8 11.1  9.6   NA 12.0  9.8 10.9 11.7  9.1  7.7  5.2 11.1
## [15] 10.6  9.0 12.0 12.3  9.8 10.8 10.8 10.0 15.2 13.4  7.9  9.1 14.6  5.7
## [29] 11.1  7.6 13.4  5.9
```

```r
men_2006 <- as.numeric(unemployment_men$`2006`)
```

```
## Warning: NAs introduced by coercion
```

```
hist(men_2006, col = blues9)
```



4. Create at least one bivariate plot showing the relationship between two variables of interest. What does/do the(se) plot(s) tell you about the association between these two variables?

That shows the corrosponding value for each province in different years. Since for comparing the variables like men and women I need to merge different rows of the three datasets, I could not compare the genders to each other (we will learn th mergin in the next classes). I did create plots seprately.

```
as.numeric(unemployment_men$`2006`)
```

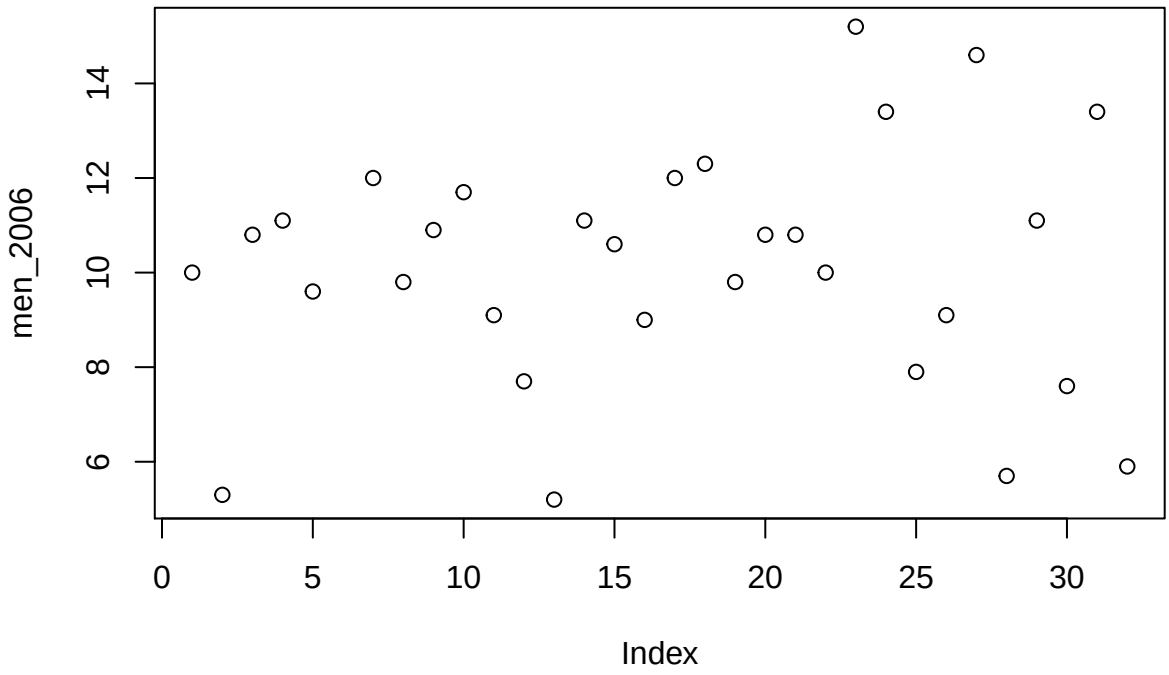
```
## Warning: NAs introduced by coercion
```

```
##  [1] 10.0  5.3 10.8 11.1  9.6   NA 12.0  9.8 10.9 11.7  9.1  7.7  5.2 11.1
## [15] 10.6  9.0 12.0 12.3  9.8 10.8 10.8 10.0 15.2 13.4  7.9  9.1 14.6  5.7
## [29] 11.1  7.6 13.4  5.9
```

```
men_2006 <- as.numeric(unemployment_men$`2006`)
```

```
## Warning: NAs introduced by coercion
```

```
plot(men_2006)
```




```
as.numeric(unemployment_women$`2006`)
```

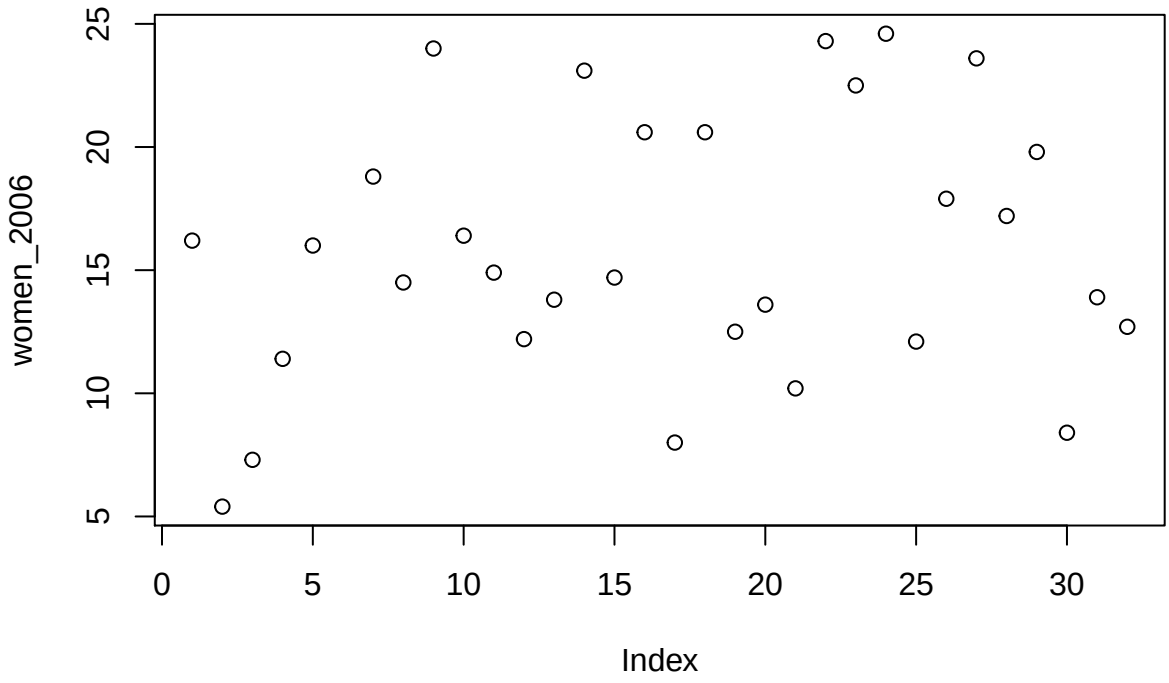
```
## Warning: NAs introduced by coercion
```

```
##  [1] 16.2  5.4  7.3 11.4 16.0   NA 18.8 14.5 24.0 16.4 14.9 12.2 13.8 23.1
## [15] 14.7 20.6  8.0 20.6 12.5 13.6 10.2 24.3 22.5 24.6 12.1 17.9 23.6 17.2
## [29] 19.8  8.4 13.9 12.7
```

```
women_2006 <- as.numeric(unemployment_women$`2006`)
```

```
## Warning: NAs introduced by coercion
```

```
plot(women_2006)
```



```
as.numeric(unemployment_total$`2006`)
```

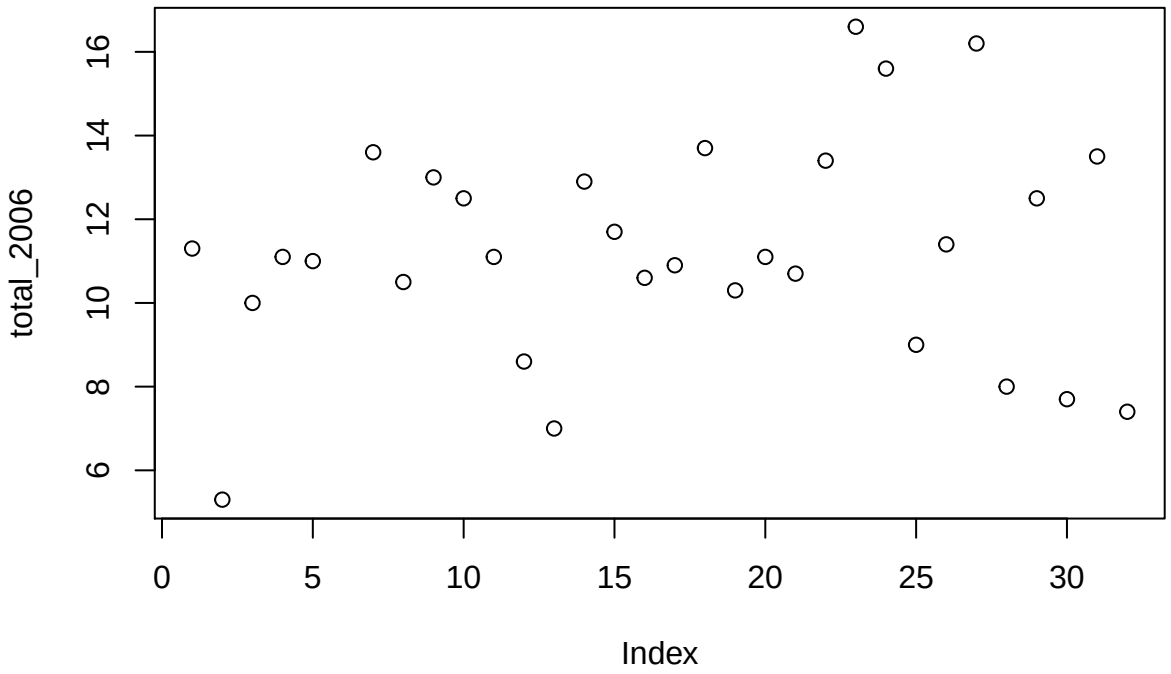
```
## Warning: NAs introduced by coercion
```

```
##  [1] 11.3  5.3 10.0 11.1 11.0   NA 13.6 10.5 13.0 12.5 11.1  8.6  7.0 12.9
## [15] 11.7 10.6 10.9 13.7 10.3 11.1 10.7 13.4 16.6 15.6  9.0 11.4 16.2  8.0
## [29] 12.5  7.7 13.5  7.4
```

```
total_2006 <- as.numeric(unemployment_total$`2006`)
```

```
## Warning: NAs introduced by coercion
```

```
plot(total_2006)
```



5. Load the `tidyr` package. Do all of your columns correspond to variables? Do any columns represent multiple variables? If your answer is yes to either question, carry out the appropriate `tidyr` function (`gather()` or `spread()` respectively) to tidy your data.

```
library(tidyr)
gather(data = unemployment_total,key = "year", value = "value", ...= -`Province | Year`)
```

```
## # A tibble: 288 x 3
##           `Province | Year`  year              value
##                       <chr> <chr>             <chr>
##  1                    Total  2001              14.2
##  2         East Azarbayejan  2001               6.7
##  3         West Azarbayejan  2001              10.6
##  4                  Ardebil  2001              11.6
##  5                  Esfahan  2001              13.1
##  6                   Alborz  2001                 -
##  7                     Ilam  2001 17.100000000000001
##  8                  Bushehr  2001                12
##  9                   Tehran  2001              12.2
## 10 Chaharmahal & Bakhtiyari  2001              12.4
## # ... with 278 more rows
```

```
library(tidyr)
gather(data = unemployment_men,key = "year", value = "value", ...= -`Province | Year`)
```

```
## # A tibble: 288 x 3
##           `Province | Year`  year value
##                       <chr> <chr> <chr>
##  1                    Total  2001  13.2
##  2         East Azarbayejan  2001     -
##  3         West Azarbayejan  2001     -
##  4                  Ardebil  2001     -
##  5                  Esfahan  2001     -
##  6                   Alborz  2001     -
##  7                     Ilam  2001     -
##  8                  Bushehr  2001     -
##  9                   Tehran  2001     -
## 10 Chaharmahal & Bakhtiyari  2001     -
## # ... with 278 more rows
```

```
library(tidyr)
gather(data = unemployment_women,key = "year", value = "value", ...= -`Province | Year`)
```

```
## # A tibble: 288 x 3
##           `Province | Year`  year              value
##                       <chr> <chr>              <chr>
##  1                    Total  2001 19.899999999999999
##  2         East Azarbayejan  2001                   -
##  3         West Azarbayejan  2001                   -
##  4                  Ardebil  2001                   -
##  5                  Esfahan  2001                   -
##  6                   Alborz  2001                   -
##  7                     Ilam  2001                   -
##  8                  Bushehr  2001                   -
##  9                   Tehran  2001                   -
## 10 Chaharmahal & Bakhtiyari  2001                   -
## # ... with 278 more rows
```

6. Do any columns need to be separated into two or more? Do any columns need to be combined into one? If so, carry out the appropriate the appropriate `tidyr` function (`separate()` or `unite()` respectively) to tidy your data.

I think I do not have any column to be seperated or united.

At this stage each row in your data should represent one observation, each column should be a variable, and each table should be observational unit.

7. What is the class of each of the variables in your analysis? Are these classes appropriate for the type of measurement they purport to capture? Explain your reasoning.

The variable of provinces' class is character, and the class of year is numeric.

```
class(2013)
```

```
## [1] "numeric"
```

```
class("women")
```

```
## [1] "character"
```

```
class("men")
```

```
## [1] "character"
```

```
class("total")
```

```
## [1] "character"
```

8. Do any of your variables need to be coerced into a different data type? If so, carry out the appropriate coercion methods below. (This includes transformation of any date objects using the `lubridate` package)

I do not think that coertion is needed for my data.

9. Are there any strings you need to manipulate for your analysis? If so, use the appropriate function from the `stringr` package.

Manipulation is not needed.

10. Do you have any missing values in your dataset? How many and how are they coded? **Be sure to look out for specific codebook values for missing values (i.e. `-1` for `NA`) as well as empty strings or other software-specific values for `NA`.** Don't worry about removing NAs yet - we'll tackle this question later once discern whether they're random or systematically distributed.

Yes, there are missing values. There are coded with "-". The summary function does show the number of the missing values.

```
summary(unemployment_total)
```

```
##  Province | Year        2001              2006
##  Length:32          Length:32          Length:32
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      2008               2009               2010               2011
##  Length:32          Length:32          Length:32          Min.   : 6.00
##  Class :character   Class :character   Class :character   1st Qu.:10.12
##  Mode  :character   Mode  :character   Mode  :character   Median :12.10
##                                                           Mean   :12.22
##                                                           3rd Qu.:13.47
##                                                           Max.   :19.30
##       2012            2013             2014
##  Min.   : 6.30   Min.   : 5.800   Min.   : 6.90
##  1st Qu.:10.28   1st Qu.: 7.975   1st Qu.: 9.15
##  Median :11.55   Median :10.350   Median :11.00
##  Mean   :11.98   Mean   :10.516   Mean   :10.94
##  3rd Qu.:13.40   3rd Qu.:12.575   3rd Qu.:12.40
##  Max.   :20.00   Max.   :17.100   Max.   :15.70
```

```
summary(unemployment_women)
```

```
##  Province | Year        2001              2006
##  Length:32          Length:32          Length:32
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##      2008               2009               2010               2011
##  Length:32          Length:32          Length:32          Min.   : 7.60
##  Class :character   Class :character   Class :character   1st Qu.:14.32
##  Mode  :character   Mode  :character   Mode  :character   Median :19.55
##                                                           Mean   :20.77
##                                                           3rd Qu.:24.70
##                                                           Max.   :42.50
##       2012            2013            2014
##  Min.   : 8.60   Min.   : 6.20   Min.   :10.70
##  1st Qu.:16.20   1st Qu.:15.55   1st Qu.:14.00
##  Median :20.10   Median :17.80   Median :19.40
##  Mean   :20.47   Mean   :19.14   Mean   :19.63
##  3rd Qu.:24.38   3rd Qu.:23.30   3rd Qu.:22.60
##  Max.   :44.40   Max.   :37.30   Max.   :35.70
```

```
summary(unemployment_men)
```

```
##  Province | Year        2001               2006
##  Length:32          Length:32          Length:32
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      2008               2009               2010               2011
##  Length:32          Length:32          Length:32          Min.   : 4.500
##  Class :character   Class :character   Class :character   1st Qu.: 8.675
##  Mode  :character   Mode  :character   Mode  :character   Median :10.450
##                                                           Mean   :10.525
##                                                           3rd Qu.:12.525
##                                                           Max.   :17.300
##       2012            2013            2014
##  Min.   : 5.60   Min.   : 4.000   Min.   : 5.600
##  1st Qu.: 8.80   1st Qu.: 7.150   1st Qu.: 8.250
##  Median : 9.85   Median : 8.600   Median : 8.950
##  Mean   :10.31   Mean   : 8.866   Mean   : 9.306
##  3rd Qu.:12.12   3rd Qu.:10.825   3rd Qu.:10.375
##  Max.   :18.10   Max.   :15.700   Max.   :14.300
```

11. Are there any special values in your dataset? If so, what are they and how do you think they got there? *The presence of special values is less likely if you haven't performed any data manipulation yet so you should remember to return to this step each time you carry out a mathematical transformation of any values in your dataset.*

There are not special values.

12. Create a boxplot of your data (you can create an individual boxplot for each variable if there are too many variables in your dataset to meaningfully visualize them all in one plot). Are there any outliers? If so, what are they and to which variable do they correspond? Do any of these outliers seem like obvious errors? If so, why?

The is no outlier. The funtion of summary also shows that I do not have any outlier in the data.

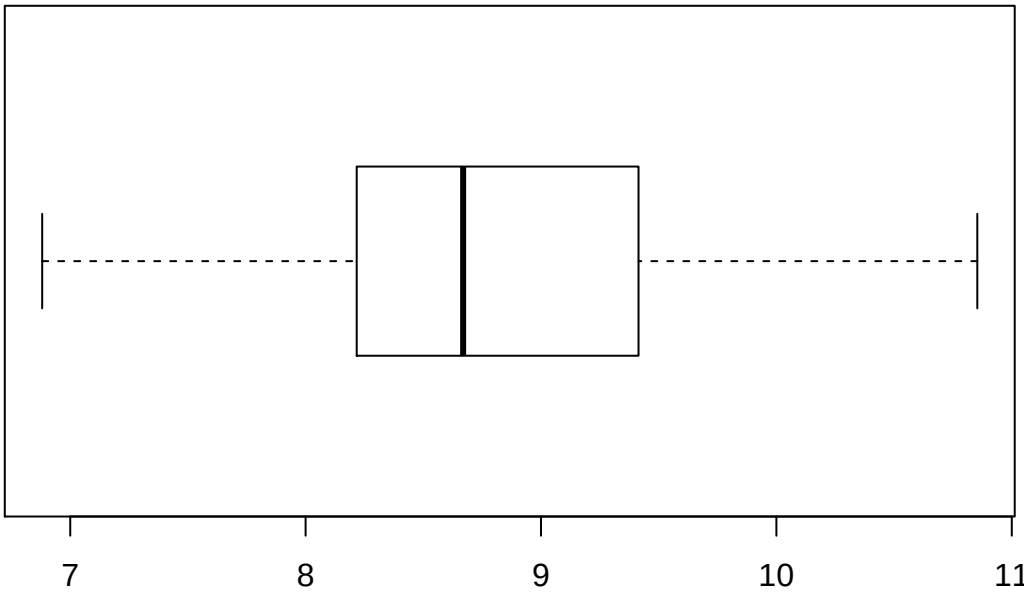
```
c(rnorm(11.7, mean = 8.866), 7.150, 8.600, 10.825)
```

```
##  [1]  9.626955  8.983870  9.362880 10.021301  9.583887 10.161364  7.268016
```

```
##  [8]  9.762096  9.249006  9.612078 10.010029  7.150000  8.600000 10.825000
```

```
men_2013 <- c(rnorm(11.7, mean = 8.866), 7.150, 8.600, 10.825)

boxplot(men_2013, horizontal = TRUE)
```



13. For any outliers and/or obvious errors, what do you think is the best way to handle them (i.e. remove them entirely, run analyses including and excluding them and compare the results, manually change them to an appropriate measure of center, or something else?).

Here the is not any outlier. So, the solution is not needed.