String Handling - Part 2

Bryana Benson
Conner Bryan

Data Import and Packages

```
library("readr")
library("stringr")
library("dplyr")
library("knitr")
```

```
url<- 'http://richardtwatson.com/data/chicago.csv'
data<-read_csv(url)
head(data)
```

```
## # A tibble: 6 x 4
##   name            position_title        department     employee_annual_sal~
##   <chr>           <chr>                 <chr>          <chr>
## 1 AARON,  ELVIA J WATER RATE TAKER       WATER MGMNT    $90744.00
## 2 AARON,  JEFFER~ POLICE OFFICER         POLICE         $84450.00
## 3 AARON,  KARINA  POLICE OFFICER         POLICE         $84450.00
## 4 AARON,  KIMBER~ CHIEF CONTRACT EXPED~ GENERAL SERVI~ $89880.00
## 5 ABAD JR,  VICE~ CIVIL ENGINEER IV      WATER MGMNT    $106836.00
## 6 ABARCA,  ANABEL ASST TO THE ALDERMAN   CITY COUNCIL   $70764.00
```

Clean the file

```
typeof(data$name)
```

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

```
typeof(data$position_title)
```

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

```
typeof(data$department)
```

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

```
typeof(data$employee_annual_salary)
```

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

In our dataset, name, position_title, and department should be a character string. employee_annual_salary should be an integar type.

```
#remove $ sign from employee_annual_salary
data$employee_annual_salary <- substring(data$employee_annual_salary,2) #start string from 2nd characte
```

```
#change to numeric data type
data$employee_annual_salary <- as.numeric(data$employee_annual_salary)
str(data$employee_annual_salary)
```

```
##  num [1:32062] 90744 84450 84450 89880 106836 ...
```

```r
head(data)
```

```
## # A tibble: 6 x 4
##   name            position_title        department     employee_annual_sal~
##   <chr>           <chr>                 <chr>                         <dbl>
## 1 AARON,  ELVIA J WATER RATE TAKER       WATER MGMNT                   90744
## 2 AARON,  JEFFER~ POLICE OFFICER         POLICE                        84450
## 3 AARON,  KARINA  POLICE OFFICER         POLICE                        84450
## 4 AARON,  KIMBER~ CHIEF CONTRACT EXPED~ GENERAL SERVI~                 89880
## 5 ABAD JR,  VICE~ CIVIL ENGINEER IV      WATER MGMNT                  106836
## 6 ABARCA,  ANABEL ASST TO THE ALDERMAN   CITY COUNCIL                  70764
```

```r
str(data$department) # character strings repeat
```

```
##  chr [1:32062] "WATER MGMNT" "POLICE" "POLICE" "GENERAL SERVICES" ...
```

```r
# convert to factor
data$department <- as.factor(data$department)
str(data$department) #Now a factor consisting of 35 department levels
```

```
##  Factor w/ 35 levels "ADMIN HEARNG",..: 35 28 28 18 35 10 32 27 10 3 ...
```

```r
str(data$position_title)
```

```
##  chr [1:32062] "WATER RATE TAKER" "POLICE OFFICER" "POLICE OFFICER" ...
```

```r
data$position_title <- as.factor(data$position_title)
str(data$position_title) #Now there are 1093 different position levels
```

```
##  Factor w/ 1093 levels "1ST DEPUTY INSPECTOR GENERAL",..: 1084 761 761 184 227 117 573 1052 952 436
```

Change name, position title, and department to title case

```r
data$name <- str_to_title(data$name)
data$position_title <- str_to_title(data$position_title)
data$department <- str_to_title(data$department)
```

Compute median salary by department

```r
med_dep_salary <- data %>% group_by(data$department) %>% summarize(median(employee_annual_salary))
names(med_dep_salary)[1] <- "Department"
names(med_dep_salary)[2] <- "Median Salary"
med_dep_salary
```

```
## # A tibble: 35 x 2
##    Department        `Median Salary`
##    <chr>                       <dbl>
##  1 Admin Hearng                68028
##  2 Animal Contrl               56928
##  3 Aviation                    72862.
##  4 Board Of Election           48036
##  5 Board Of Ethics             81948
##  6 Budget & Mgmt               89340
##  7 Buildings                   97920
##  8 Business Affairs            75960
##  9 City Clerk                  62004
```

```
## 10 City Council               52950
## # ... with 25 more rows
```

Use kable to format a report, with a comma for thousands and zero decimal places

```
med_dep_salary$`Median Salary`<- round(med_dep_salary$`Median Salary`,0)
med_dep_salary$`Median Salary` <- prettyNum(med_dep_salary$`Median Salary`,big.mark=",",scientific=FALSE
str(med_dep_salary$`Median Salary`)
```

```
##  chr [1:35] "68,028" "56,928" "72,862" "48,036" "81,948" "89,340" ...
```

```
kable(med_dep_salary,col.names = c("Department", "Median Salary"))
```

Department	Median Salary
Admin Hearng	68,028
Animal Contrl	56,928
Aviation	72,862
Board Of Election	48,036
Board Of Ethics	81,948
Budget & Mgmt	89,340
Buildings	97,920
Business Affairs	75,960
City Clerk	62,004
City Council	52,950
Community Development	85,764
Cultural Affairs	84,168
Disabilities	82,044
Doit	96,066
Family & Support	13,520
Finance	68,028
Fire	91,362
General Services	93,600
Health	81,948
Human Relations	89,676
Human Resources	73,170
Inspector Gen	75,036
Ipra	89,880
Law	71,292
License Appl Comm	71,292
Mayor's Office	74,250
Oemc	20,051
Police	87,384
Police Board	79,974
Procurement	75,960
Public Library	57,696
Streets & San	73,840
Transportn	81,536
Treasurer	88,626
Water Mgmnt	82,576