Post1-kaiqian-zhu

kaiqianzhu

10/21/2017

Data Frames Real World Application

Motivation

WHAT is a R data frame exactly, **WHY** data frame is so important in data analysis, and **How** do we utilize its special properties? In this post, we will go through the useful features of data frame and use them to analyzes real world data.

Data frame is one of the most important **Data Structure** in R that allow us to effortlessly handle **tabular data sets**. And we all know that tabular data sets like **csv** files are the primary data set we use in data analysis. To make it more easy to understand, data frame is a **container** in which we store the **columns** of our tabular input as **vectors** for specific varables(e.g height, weight, age). And normally, the **row** of the data frame corresponds to the measurements of one of the instance or sample.(Tim's measurements, John's measurements). Now, you can imagine we created a **overviewable, rectangular grid** in R by inputting a tabular data sets.

Here is a R build in data frame that we can visiualize what a data frame looks like:

```
head(mtcars,5) # showing the first 5 rows
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

We see that the column like **hp** is a vector for the **variable: horse power** of a car and the first row **Mazda RX4** is the measurements for a **instance: car called Mazda RX4**.

We now see that data frame is an indispensable 'tool' in R and is the first step for almost any data analysis, since we need a container to apply all kinds of operation(functions) on. Hopefully, we have a comprehensive understandign on data frame now and below we are going to see more clearly on data frames' **Why** and **How** by applying it in real world application.

Application

First, we want to read in the **tabular data sets: Salaries.csv** to store in **Data Frame: salary**. And take a glimpse in that data frame.

```
salary <- read.csv('data/Salaries.csv')
head(glimpse(salary),5)
```

```
## Observations: 148,654
## Variables: 13
## $ Id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
## $ EmployeeName     <fctr> NATHANIEL FORD, GARY JIMENEZ, ALBERT PARDINI...
## $ JobTitle         <fctr> GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORI...
## $ BasePay          <fctr> 167411.18, 155966.02, 212739.13, 77916.0, 13...
## $ OvertimePay      <fctr> 0.0, 245131.88, 106088.18, 56120.71, 9737.0,...
## $ OtherPay         <fctr> 400184.25, 137811.38, 16452.6, 198306.9, 182...
## $ Benefits         <fctr> , , , , , , , , , , , , , , , , , , , , , ...
## $ TotalPay         <dbl> 567595.4, 538909.3, 335279.9, 332343.6, 32637...
## $ TotalPayBenefits <dbl> 567595.4, 538909.3, 335279.9, 332343.6, 32637...
## $ Year             <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 201...
## $ Notes            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ Agency           <fctr> San Francisco, San Francisco, San Francisco,...
## $ Status           <fctr> , , , , , , , , , , , , , , , , , , , , , ...
```

```
##   Id      EmployeeName                                      JobTitle
## 1  1    NATHANIEL FORD GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY
## 2  2      GARY JIMENEZ                CAPTAIN III (POLICE DEPARTMENT)
## 3  3    ALBERT PARDINI                CAPTAIN III (POLICE DEPARTMENT)
## 4  4 CHRISTOPHER CHONG           WIRE ROPE CABLE MAINTENANCE MECHANIC
## 5  5   PATRICK GARDNER   DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)
##     BasePay OvertimePay  OtherPay Benefits TotalPay TotalPayBenefits Year
## 1 167411.18         0.0 400184.25          567595.4         567595.4 2011
## 2 155966.02   245131.88 137811.38          538909.3         538909.3 2011
## 3 212739.13   106088.18   16452.6          335279.9         335279.9 2011
## 4   77916.0    56120.71  198306.9          332343.6         332343.6 2011
## 5  134401.6      9737.0 182234.59          326373.2         326373.2 2011
##   Notes        Agency Status
## 1    NA San Francisco
## 2    NA San Francisco
## 3    NA San Francisco
## 4    NA San Francisco
## 5    NA San Francisco
```

Notice that this is a real world tabular data sets which has 148,654 rows in it, meaning we have measurments of over 140,000 people in SF! Let only focus on the people work as full-time and we see that only 22,334 people remains:

```
salary_ft = filter(salary, Status == "FT")
head(salary_ft,5)
```

```
##       Id         EmployeeName               JobTitle   BasePay OvertimePay
## 1 110533           Amy P Hart      Asst Med Examiner 318835.49    10712.95
## 2 110535       Gregory P Suhr        Chief of Police 307450.04        0.00
## 3 110536 Joanne M Hayes-White Chief, Fire Department 302068.00        0.00
## 4 110537      Ellen G Moffatt      Asst Med Examiner 270222.04     6009.22
## 5 110538        John L Martin            Dept Head V 311298.55        0.00
##   OtherPay Benefits TotalPay TotalPayBenefits Year Notes        Agency
## 1 60563.54 89540.23 390112.0         479652.2 2014    NA San Francisco
## 2 19266.72 91302.46 326716.8         418019.2 2014    NA San Francisco
## 3 24165.44 91201.66 326233.4         417435.1 2014    NA San Francisco
## 4 67956.20 71580.48 344187.5         415767.9 2014    NA San Francisco
## 5     0.00 89772.32 311298.5         401070.9 2014    NA San Francisco
##   Status
## 1     FT
## 2     FT
## 3     FT
## 4     FT
## 5     FT
```

The Second step is to assign them into different group so that we can works with smaller number of rows and get a cleaner data frame. Notice that salary for different jobs is more interesting to examine than salary for different individuals. Thus we group the instances by job:

```
salary_job = salary_ft %>% group_by(JobTitle)
salary_job %>%
    summarise(Frequency = n()) %>%
    arrange(desc(Frequency)) %>%
    head(5)
```

```
## # A tibble: 5 x 2
##           JobTitle Frequency
##              <fctr>     <int>
## 1 Transit Operator      1524
## 2      Firefighter       738
## 3 Police Officer 3       642
## 4        Custodian       565
## 5   Deputy Sheriff       552
```

What we just did is **putting people with the same jobtitle together** in the data frame salary and assign it to a new data frame called **salary_job**. Notice that if we view group of people with the same job as an instance(row), we now only have 916 rows! Then we **sorted** them by the job by its frequency and see that there are most people works as **Transit Operator**, then **Firefighter**. What about how each job get paid(your guess)? Probably Engineer?

```
salary_meanpay = salary_job %>%
    summarise(Mean_Salary = mean(TotalPay)) %>%
    arrange(desc(Mean_Salary))
head(salary_meanpay, 5)
```

```
## # A tibble: 5 x 2
##                      JobTitle Mean_Salary
##                         <fctr>       <dbl>
## 1           Asst Med Examiner    339664.8
## 2             Chief of Police    326716.8
## 3      Chief, Fire Department    326233.4
## 4  Gen Mgr, Public Trnsp Dept    294000.2
## 5 Dep Chf of Dept (Fire Dept)    285575.8
```

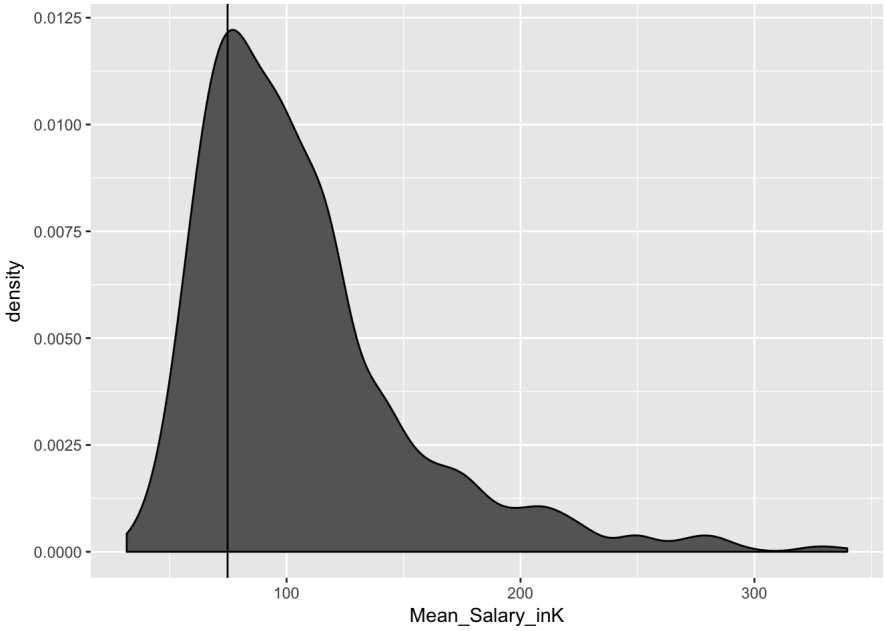
Wrong! It is actually **Asst Med Examiner**. How much does **engineers** earn, you may ask:

```
head(filter(salary_meanpay, grepl("Engineer", JobTitle, fixed = TRUE)),10)
```

```
## # A tibble: 10 x 2
##                         JobTitle Mean_Salary
##                            <fctr>       <dbl>
##  1 Marine Engineer of Fire Boats    207420.1
##  2  Engineer/Architect Principal    177995.9
##  3       Building Plans Engineer    144502.0
##  4         IS Engineer-Principal    143178.3
##  5       Administrative Engineer    140187.7
##  6           Structural Engineer    140042.8
##  7      Fire Protection Engineer    138023.9
##  8                      Engineer    135451.3
##  9            IS Engineer-Senior    132631.5
## 10     Chief Stationary Engineer    119961.1
```

Not bad. Let see the **density** plot of the **mean total pay** of each jobs and get a better visiualization.

```
salary_meanpay = mutate(salary_meanpay, Mean_Salary_inK = salary_meanpay$Mean_Salary / 1000)
salary_meanpay %>%
    ggplot(aes(x = Mean_Salary_inK)) +
    geom_density(fill = "grey40") + geom_vline(xintercept=mean(salary$TotalPay)/1000)
```



Let do something more interesting! How much do we need to earn in order to live **WELL** in San Francisco. First, we need to analyze the cost of living in SF.

Food

Food	[Edit]
Basic lunchtime menu (including a drink) in the business district	$16
Combo meal in fast food restaurant (Big Mac Meal or similar)	$8
500 gr (1 lb.) of boneless chicken breast	$7
1 liter (1 qt.) of whole fat milk	$1.16
12 eggs, large	$4.97
1 kg (2 lb.) of tomatoes	$5.75
500 gr (16 oz.) of local cheese	$11
1 kg (2 lb.) of apples	$4.66
1 kg (2 lb.) of potatoes	$2.97
0.5 l (16 oz) domestic beer in the supermarket	$2.77
1 bottle of red table wine, good quality	$17
2 liters of Coca-Cola	$2.21
Bread for 2 people for 1 day	$2.40

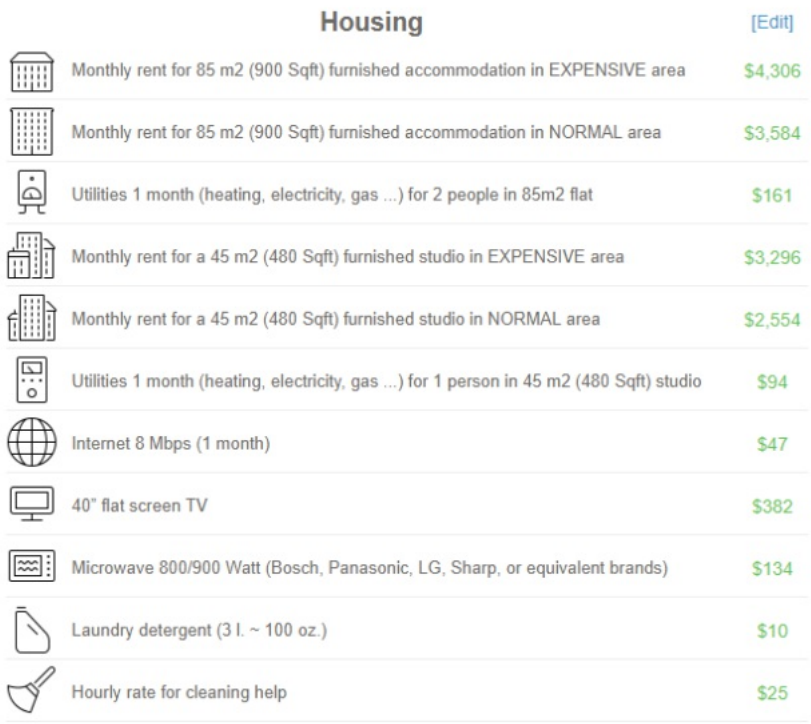
food

How much are we going to spend on food? According to the above data from Expatistan.com(normal meal w/ nothing special):

```
breadfast = 8
meal = 16
food_cost_annual = (breadfast + meal*2)*30*12
sprintf("People in SF normally spend $%s on food annually", food_cost_annual)
```

```
## [1] "People in SF normally spend $14400 on food annually"
```

Housing



Housing	[Edit]
Monthly rent for 85 m2 (900 Sqft) furnished accommodation in EXPENSIVE area	$4,306
Monthly rent for 85 m2 (900 Sqft) furnished accommodation in NORMAL area	$3,584
Utilities 1 month (heating, electricity, gas ...) for 2 people in 85m2 flat	$161
Monthly rent for a 45 m2 (480 Sqft) furnished studio in EXPENSIVE area	$3,296
Monthly rent for a 45 m2 (480 Sqft) furnished studio in NORMAL area	$2,554
Utilities 1 month (heating, electricity, gas ...) for 1 person in 45 m2 (480 Sqft) studio	$94
Internet 8 Mbps (1 month)	$47
40" flat screen TV	$382
Microwave 800/900 Watt (Bosch, Panasonic, LG, Sharp, or equivalent brands)	$134
Laundry detergent (3 l. ~ 100 oz.)	$10
Hourly rate for cleaning help	$25

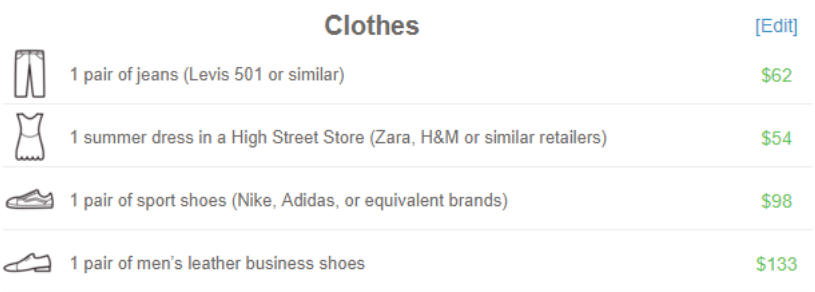
housing

Housing is a huge part in cost of living. We spend most of the salary on that! Let say we stay in a small studio in **NORMAL** area.

```
utilities = 94
rent = 2554
internet = 47

housing_cost_annual = (utilities + rent + internet)*12
sprintf("People in SF normally spend $%s on housing annually", housing_cost_annual)
```

```
## [1] "People in SF normally spend $32340 on housing annually"
```

clothes



Clothes	[Edit]
1 pair of jeans (Levis 501 or similar)	$62
1 summer dress in a High Street Store (Zara, H&M or similar retailers)	$54
1 pair of sport shoes (Nike, Adidas, or equivalent brands)	$98
1 pair of men's leather business shoes	$133

housing

Let say on average we buy ourselves a new outfit every 3 monthes. Just basic clothes, not Gucci, Prada:

```
jean = 62
top = 50
shoes = 98

clothes_cost_annual = (jean + top + shoes)*4
sprintf("People in SF normally spend $%s on clothes annually", clothes_cost_annual)
```

```
## [1] "People in SF normally spend $840 on clothes annually"
```

Transportation

Transportation [Edit]

Volkswagen Golf 1.4 TSI 150 CV (or equivalent), with no extras, new	$22,734
1 liter (1/4 gallon) of gas	$0.80
Monthly ticket public transport	$79
Taxi trip on a business day, basic tariff, 8 km. (5 miles)	$19

transportation

Let say our generous parent grant us a car for transportation:

```
gas = 53
insurance = 55
car_wash = 8

transportation_cost_annual = (gas + insurance + car_wash)*12
sprintf("People in SF normally spend $%s on transportation annually", transportation_cost_annual)
```

```
## [1] "People in SF normally spend $1392 on transportation annually"
```

Personal Care

Personal Care [Edit]

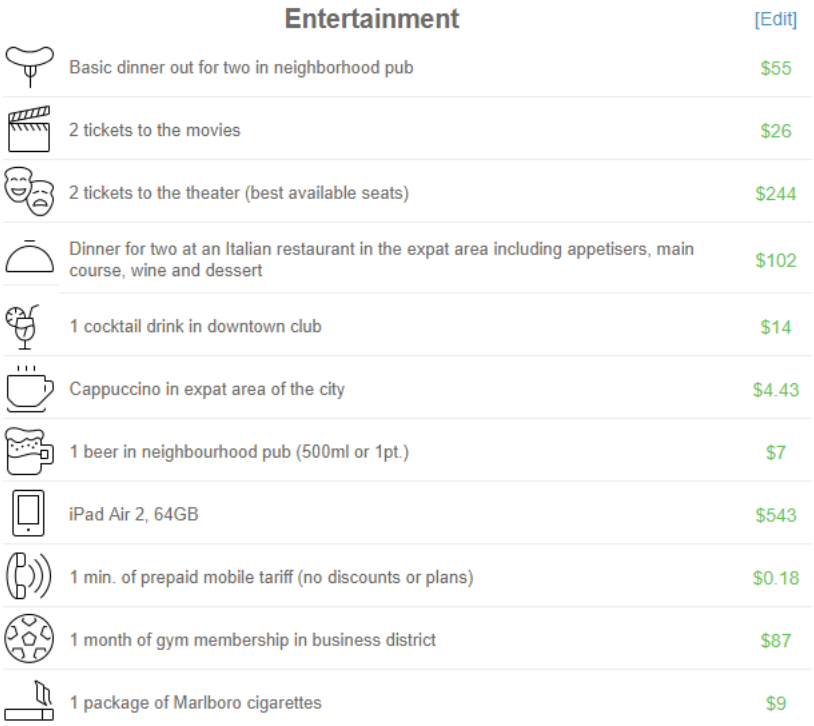
Cold medicine for 6 days (Tylenol, Frenadol, Coldrex, or equivalent brands)	$7
1 box of antibiotics (12 doses)	$17
Short visit to private Doctor (15 minutes)	$137
1 box of 32 tampons (Tampax, OB, ...)	$7
Deodorant, roll-on (50ml ~ 1.5 oz.)	$3.84
Hair shampoo 2-in-1 (400 ml ~ 12 oz.)	$7
4 rolls of toilet paper	$3.26
Tube of toothpaste	$2.19
Standard men's haircut in expat area of the city	$44

Personal Care

```
hair = 44
cosmetic = 30
necessities = 20


pc_cost_annual = (necessities + cosmetic + hair)*12
sprintf("People in SF normally spend $%s on Personal Care annually", pc_cost_annual)
```

```
## [1] "People in SF normally spend $1128 on Personal Care annually"
```

Entertainment



Entertainment	[Edit]
Basic dinner out for two in neighborhood pub	$55
2 tickets to the movies	$26
2 tickets to the theater (best available seats)	$244
Dinner for two at an Italian restaurant in the expat area including appetisers, main course, wine and dessert	$102
1 cocktail drink in downtown club	$14
Cappuccino in expat area of the city	$4.43
1 beer in neighbourhood pub (500ml or 1pt.)	$7
iPad Air 2, 64GB	$543
1 min. of prepaid mobile tariff (no discounts or plans)	$0.18
1 month of gym membership in business district	$87
1 package of Marlboro cigarettes	$9

Entertainment

```
moive = 26
gym = 87
phone_bill = 50
one_special_dinner = 102


entertainment_cost_annual = (moive + gym + phone_bill+one_special_dinner)*12
sprintf("People in SF normally spend $%s on Entertainment annually", entertainment_cost_annual)
```

```
## [1] "People in SF normally spend $3180 on Entertainment annually"
```

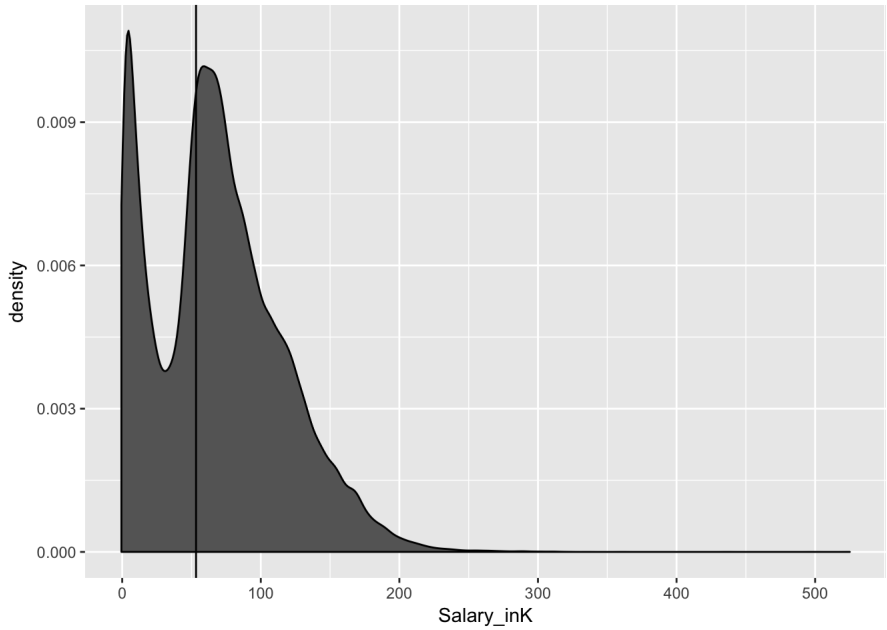
Above is my estimation of a normal person's **annual cost of living** in SF based on real data. Do you have a guess on how much?

```
cost_of_living_annual = food_cost_annual + housing_cost_annual + pc_cost_annual + transportation_cost_annual + ent
ertainment_cost_annual + clothes_cost_annual

sprintf("The cost of living annually in SF is $%s", cost_of_living_annual)
```

```
## [1] "The cost of living annually in SF is $53280"
```

That is expensive!! But how much do we earn? of course, after tax!

```
salary_taxed = salary %>% mutate(TotalPayTaxed = TotalPay*0.925) %>% select(JobTitle, TotalPayTaxed)


salary_taxed = mutate(salary_taxed, Salary_inK = salary_taxed$TotalPayTaxed / 1000)
salary_taxed %>%
    ggplot(aes(x = Salary_inK)) +
    geom_density(fill = "grey40") + geom_vline(xintercept=cost_of_living_annual/1000)
```




```
prop = nrow(filter(salary_taxed, TotalPayTaxed>=cost_of_living_annual)) / nrow(salary_taxed)
prop
```

```
## [1] 0.630242
```

We see that only 63% of the samples can afford the **basic** cost of living!!

Can you earn more than $53280(after tax)?

Conclusions

From above real world application, we see that **data frame** provide us a platform to explore the pattern in the real world tabular data sets. With the help from library like ggplot2, dplyr, we are able to exploit the power of data frame. Filtering, ordering, grouping, ploting and so much more can be done in just few lines of code! Notice that we just worked with a real data sets, which is huge! However, data frame can handle it easily and generate result instantly. Without data frame, we have no way to put together thousands of instances(rows) each with hundreds of variable(cols) and perform operation on them jointly. I hope you learn something interesting.

Take-home message
* Data Frame is not only useful but necessary in analyzing real world data sets.
* It is expensive to live in San Francisco.
* Find a job with salary more than **$53280**(taxed) in order to **survive**!

References

15-easy-solutions-data-frame-problems-r

r-introduction-data-frame

sf-salary

data-frame-slide

cost-of-living-expatistan

cost-of-living-sf

tips-for-data-frame