

M1 Mini Assignment 1

Lars Nielsen

9/9/2019

Contents

M1 Mini-Assignment 1	1
1. Preprocessing	1
2. People	2
3. Trips	4

M1 Mini-Assignment 1

You are given 2 datasets from <https://nomadlist.com/> - A community page for remote workersworldwide.

- The trips data holds ~46k individual trips of travellers on the platform <https://github.com/SDS-AAU/M1-2019/raw/master/data/trips.csv>
- People contains some personal information on 4k travelers <https://github.com/SDS-AAU/M1-2019/raw/master/data/people.csv>
- Finally, you find a countrylist file that holds countrycodes, contrynames and region-associations <https://github.com/SDS-AAU/M1-2019/raw/master/data/countrylist.csv>

Your solution approach is more important than the results obtained! Comment your notebook well, explaining all the steps of your analysis. Small technical explanations can go as comments in the code. Broader explanations should be inserted as markdown cells. Remember that notebooks execute sequentially.

Submission: Wednesday 11.9. 12:00.

Peergrade.io (link + submission details will be sent out on Monday, 9.9)

1. Preprocessing

a. Trips: transform dates into timestamps

(note: in Python, you will have to ‘coerce’ errors for faulty dates)

```
trips[,c("date_end", "date_start") := .(as.Date(date_end, "%Y-%m-%d"), as.Date(date_start, "%Y-%m-%d"))
head(trips$date_end, 5)
```

```
## [1] "2018-06-15" "2018-06-03" "2017-11-05" "2017-08-07" "2017-03-18"
```

b. Calculate trip duration in days

(you can use loops, list comprehensions or map-lambda-functions (python) to create a column that holds the numerical value of the day. You can also use the “datetime” package.)

```
trips[, dur_days := date_end - date_start]
head(trips$dur_days, 5)
```

```
## Time differences in days
## [1] 11  3  4 14 29
```

c. Filter extreme (fake?) observations

for durations as well as dates - start and end (trips that last 234565 days / are in the 17th or 23rd century) The minimum duration of a trip is 1 day! Hint: use percentiles/quantiles to set boundaries for extreme values - between 1 and 97, calculate and store the boundaries before subsetting. Rhint: Use `percent_rank(as.numeric(variable))` to create percentiles

```
trips[, quantile := dplyr::percent_rank(as.numeric(trips$dur_days))]  
trips_s <- trips[quantile >= 0.01 & quantile <= 0.97]  
  
range(na.omit(trips$dur_days))
```

```
## Time differences in days  
## [1] -730484 731122
```

```
range(trips_s$dur_days)
```

```
## Time differences in days  
## [1] 1 208
```

d. Join the countrylist data

to the trips data-frame using the countrycode as a key

```
countrylist[, country_code := alpha_2]  
trips_s$country_code[trips_s$country_code == "UK"] <- "GB"  
trips_s[countrylist, on = "country_code", c("region", "sub_region")] := .(i.region, i.sub_region)  
  
head(trips_s[,.(country_code, dur_days, region, sub_region)])
```

```
##   country_code dur_days  region                sub_region  
## 1:          MX   11 days Americas Latin America and the Caribbean  
## 2:          MX    3 days Americas Latin America and the Caribbean  
## 3:          MX    4 days Americas Latin America and the Caribbean  
## 4:          JO   14 days      Asia                Western Asia  
## 5:          CN   29 days      Asia                Eastern Asia  
## 6:          VN  167 days      Asia                South-eastern Asia
```

2. People

a. How many people have at least a “High School” diploma?

Hint: For this calculation remove missing value-rows or fill with “False”

```
people_s <- people[education_raw != ""]  
table(people_s$education_raw)
```

```
##  
##                                Bachelor's Degree  
##                                197  
##    Bachelor's Degree, Master's Degree  
##                                9
```

```
##                High School
##                58
##            High School, Bachelor's Degree
##                43
## High School, Bachelor's Degree, Master's Degree
##                29
##                Master's Degree
##                115
```

```
people_s[,.N]
```

```
## [1] 451
```

```
people[, high_school := education_raw %like% "High|Master|Bachelor"]
```

b. How many people working with “Software Dev” have a “Master’s Degree”?

```
people_s[work_raw %like% "Software Dev" & education_raw %like% "Master", .N]
```

```
## [1] 57
```

c. Who is the person ...

with a Master’s Degree that has the highest number of followers?[Explore who this person is. :-)]

```
c <- people_s[education_raw %like% "Master"][order(-followers)]
```

```
head(c)
```

```
##      V1      username followers following
## 1: 2043      @levelsio      2182       353
## 2: 1298          @aaz        259        46
## 3: 1949      @neosilky       102         6
## 4:  609      @zackllnyoung     60        92
## 5: 3036      @html5cat        32        18
## 6: 1829 @siddharthkshetrapal    29        24
```

```
##
## 1:                Software Dev, Startup Founder, Creative
## 2:                Web Dev, Software Dev, Startup Founder, Blogging, SaaS
## 3:                Web Dev, Software Dev
## 4:                Startup Founder, Blogging, Coach, Marketing
## 5: Software Dev, Mobile Dev, Startup Founder, Product Manager, Education, Community
## 6:                Web Dev, Software Dev, UI/UX Design, Creative
```

```
##      education_raw
## 1: High School, Bachelor's Degree, Master's Degree
## 2: High School, Bachelor's Degree, Master's Degree
## 3: High School, Bachelor's Degree, Master's Degree
## 4: High School, Bachelor's Degree, Master's Degree
## 5:                Master's Degree
## 6:                Master's Degree
```

```
people_s[username == "@levelsio"]
```

```
##      V1 username followers following
## 1: 2043 @levelsio      2182       353
```

```
##                                work_raw
## 1: Software Dev, Startup Founder, Creative
##                                education_raw
## 1: High School, Bachelor's Degree, Master's Degree
```

3. Trips

a. Which country received the highest number of trips?

```
a <- trips_s[, .N, by = country][order(-N)]
head(a)
```

```
##          country      N
## 1: United States 6963
## 2:      Thailand 3278
## 3: United Kingdom 2053
## 4:          Spain 1875
## 5:          Germany 1814
## 6:          France 1390
```

b. Which country received the highest number of trips in 2017?

Use the start of trips as a timereference. (python: use datetimeindex created in 1 as a selector)?Rhint: Use functions from lubridate package to extract year.

```
b <- trips_s[year(date_start) == 2017, .N, by = country][order(-N)]
head(b)
```

```
##          country      N
## 1: United States 1823
## 2:      Thailand  894
## 3: United Kingdom  612
## 4:          Spain  598
## 5:          Germany 456
## 6:          France 391
```

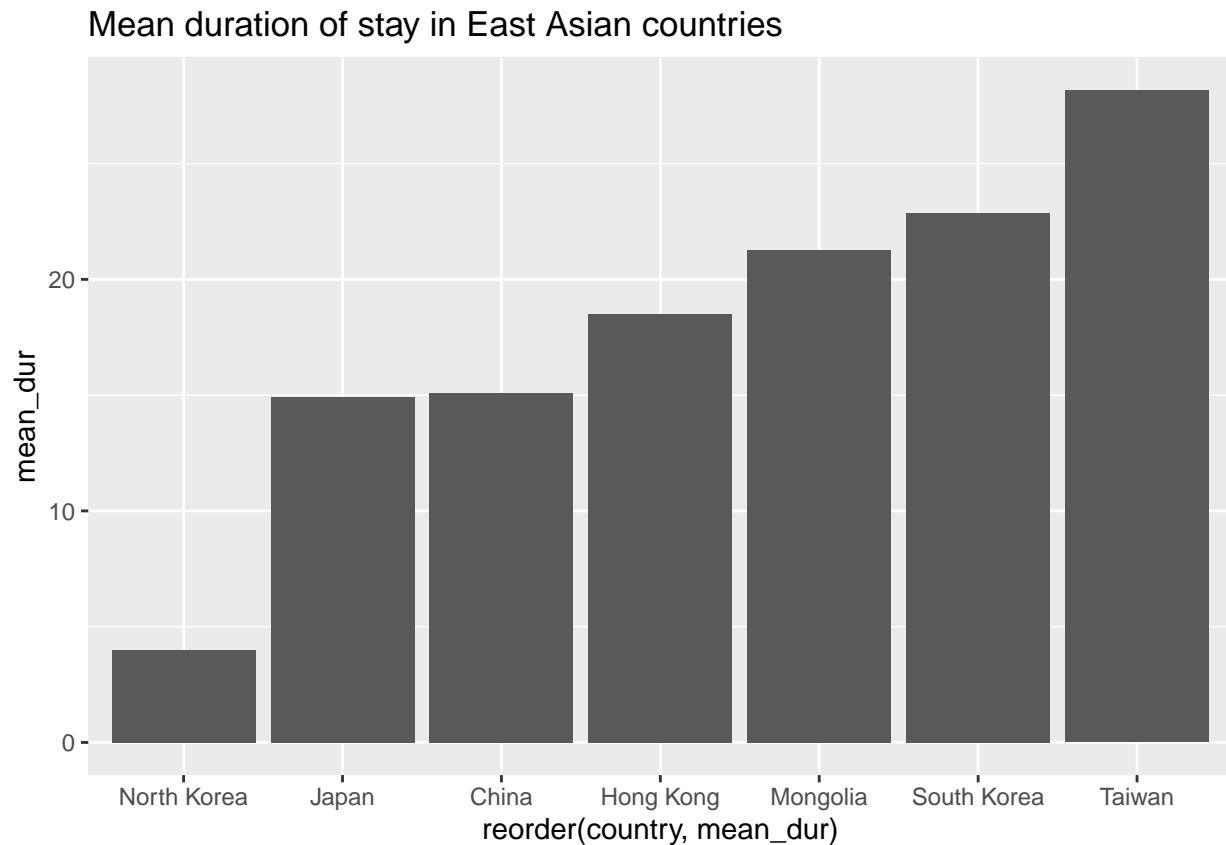
c. Which is the country in 'Eastern Asia' ...

where travellers spent on average least time whengoing there? Provide a visualization.

```
c <- trips_s[sub_region == 'Eastern Asia', .(mean_dur = mean(dur_days), total = .N), by = country][order(mean_dur)]
c
```

```
##          country    mean_dur total
## 1: North Korea  4.00000 days      9
## 2:      Japan 14.90731 days    971
## 3:      China 15.07692 days   1066
## 4:  Hong Kong 18.50000 days      2
## 5:  Mongolia 21.27778 days     18
## 6: South Korea 22.85465 days    344
## 7:      Taiwan 28.15599 days    359
```

```
ggplot(c, aes(reorder(country, mean_dur), mean_dur)) +
  geom_col() +
  scale_y_continuous() +
  labs(title = "Mean duration of stay in East Asian countries")
```



d. Do nomads that ...

indicate working in “Software Dev” tend to have shorter or longer trips on average?

```
trips_s[people, on = "username", c("work_raw") := .(i.work_raw)]
trips_s[, dev := work_raw %like% "Software Dev"]
trips_s[, (mean_dur = mean(dur_days)), by = dev]
```

```
##      dev      mean_dur
## 1: TRUE 16.88626 days
## 2: FALSE 18.36551 days
```

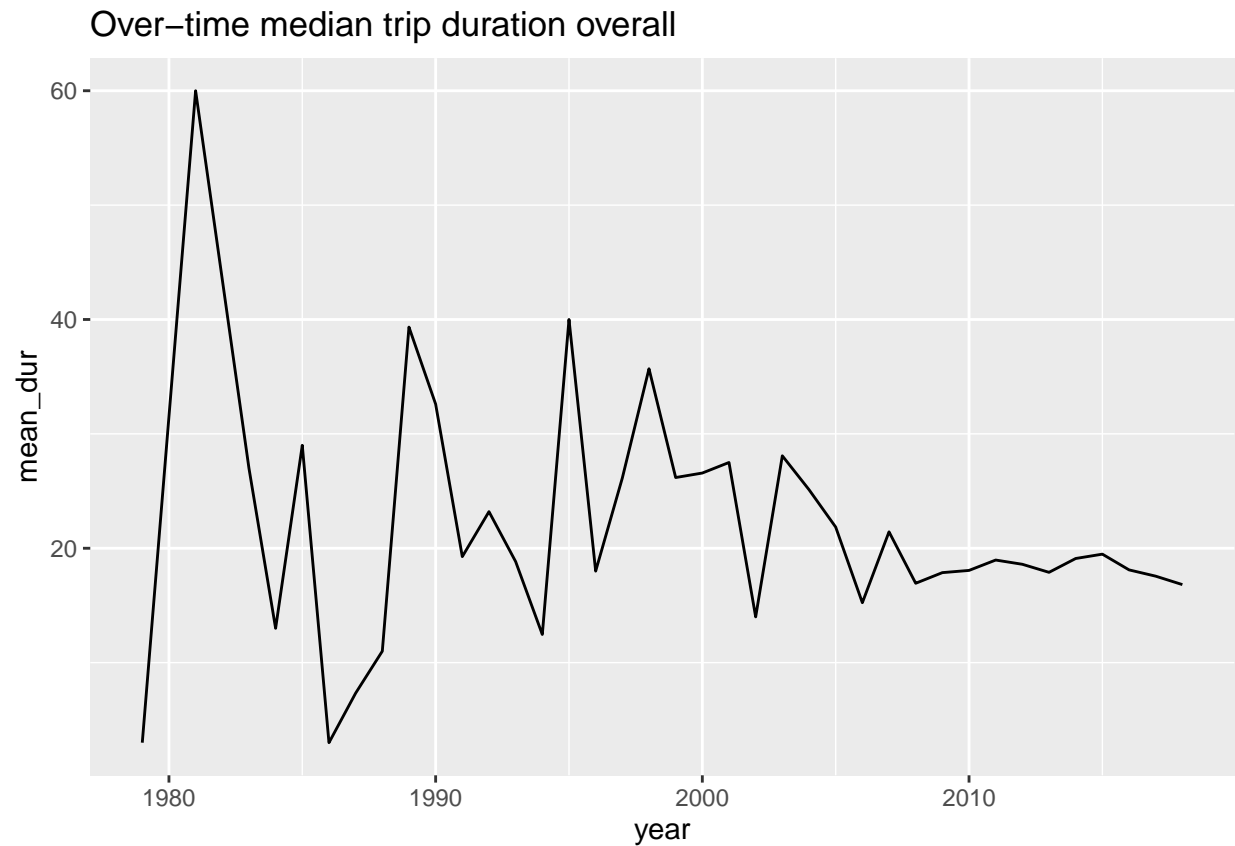
e. Visualize over-time median trip duration

overall (bonus: and split by world-region). You will get a weird looking plot :-)

```
e1 <- trips_s[, (mean_dur = mean(dur_days)), by = year(date_start)][year > 1970 & year < 2019]
```

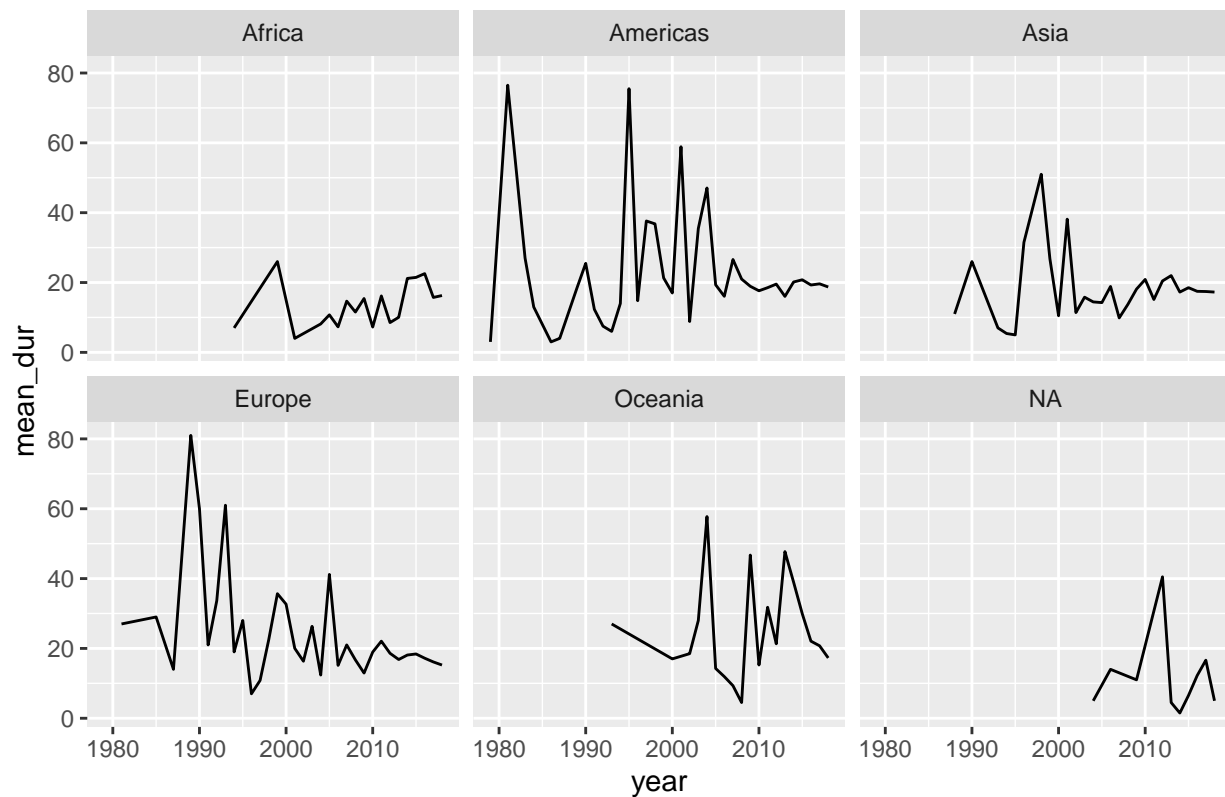
```
ggplot(e1, aes(year, mean_dur)) +
  scale_y_continuous() +
```

```
geom_line() +  
labs(title="Over-time median trip duration overall")
```



```
e2 <- trips_s[, .(mean_dur = mean(dur_days)), by = .(year(date_start), region)][year>1970 & year < 2019]  
  
ggplot(e2, aes(year, mean_dur)) +  
  scale_y_continuous() +  
  geom_line() +  
  facet_wrap(~region) +  
  labs(title="Over-time median trip duration overall - by region")
```

Over-time median trip duration overall – by region



```
table(trips_s$region, useNA="always")
```

```
##
##   Africa Americas      Asia  Europe  Oceania   <NA>
##    1471    11774   12449   16218    1379     44
```

```
vec <- sort(table(trips_s$country_code, is.na(trips_s$region))[,2])
subset(vec, vec > 0)
```

```
## CT OI KS CB VB AA IA
##  1  1  4  5  5 13 15
```