The life-changing magic of tidying up

Min Fang

"80% of data analysis is spent on the process of cleaning and preparing the data."

— Conventional wisdom

Tidy data

“

Each variable forms a column.

Each observation forms a row.

Each type of observational unit forms a table.

— *Hadley Wickham*

R packages I use regularly to handle tidy data seamlessly

- tidyverse (https://www.tidyverse.org/)
  - readr
  - tidyr
  - dplyr
  - ggplot2
- stringr

Doing things easily with tidy data

Q1: What is the distribution of aspect labels from the model?

```
In [6]:  1  head(sentiment_eval)
```

review_id	review	bathroom	QA_bathroom	breakfast	QA_breakfast	internet	QA_internet	staff	QA_staff	swimming_pool	QA_swimming_pool
1	Was ment to stay for 2 nights lasted cnel The bath [...]	3	2	NA	NA	NA	NA	NA	NA	NA	NA
2	On arrival walking to reception doors met by cig e [...]	3	2	NA	NA	NA	NA	1	2	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	NA	NA	NA	NA	NA	NA	1	3	3	3
4	Went with open mind after reading recent reviews b [...]	3	1	3	3	NA	NA	3	3	NA	NA
5	Expensive rooms did not match expectations. Breakf [...]	NA	NA	NA	3	NA	NA	NA	NA	NA	NA
6	Overall, great hotel with big rooms and toilets, g [...]	3	2	NA	NA	NA	NA	NA	NA	NA	NA

Q1: What is the distribution of aspect labels from the model?

In [21]:

```
predictions <- sentiment_eval %>% select(-starts_with("QA_"))
head(predictions)
```

review_id	review	bathroom	breakfast	internet	staff	swimming_pool
1	Was ment to stay for 2 nights lasted one! The bath [...]	3	NA	NA	NA	NA
2	On arrival walking to reception doors met by cig e [...]	3	NA	NA	1	NA
3	Had read some reviews before my visit on 17th Oct' [...]	NA	NA	NA	1	3
4	Went with open mind after reading recent reviews b [...]	3	3	NA	3	NA
5	Expensive rooms did not match expectations. Breakf [...]	NA	NA	NA	NA	NA
6	Overall, great hotel with big rooms and toilets, g [...]	3	NA	NA	NA	NA

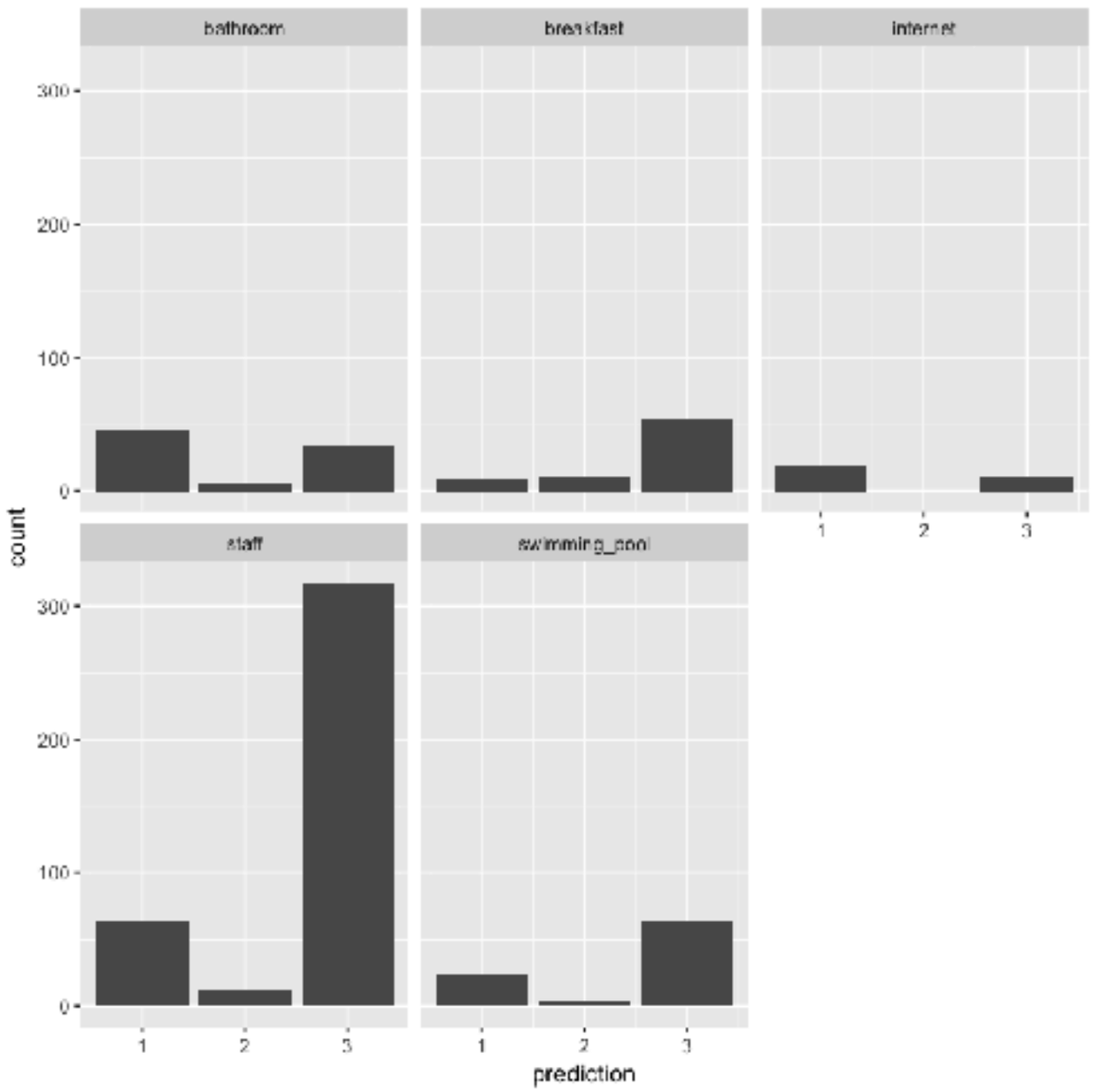
Q1: What is the distribution of aspect labels from the model?

In [25]:
```
sentiment_eval %>% select(-starts_with("QA_")) %>%
gather("aspect", "prediction", 3:7) %>% arrange(review_id)
```

review_id	review	aspect	prediction
1	Was ment to stay for 2 nights lasted one! The bath [...]	bathroom	3
1	Was ment to stay for 2 nights lasted one! The bath [...]	breakfast	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	internet	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	staff	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	swimming_pool	NA
2	On arrival walking to reception doors met by cig e [...]	bathroom	3
2	On arrival walking to reception doors met by cig e [...]	breakfast	NA
2	On arrival walking to reception doors met by cig e [...]	internet	NA
2	On arrival walking to reception doors met by cig e [...]	staff	1
2	On arrival walking to reception doors met by cig e [...]	swimming_pool	NA
3	Had read some reviews before my visit on 17th Oct' [...]	bathroom	NA
3	Had read some reviews before my visit on 17th Oct' [...]	breakfast	NA
3	Had read some reviews before my visit on 17th Oct' [...]	internet	NA
3	Had read some reviews before my visit on 17th Oct' [...]	staff	1
3	Had read some reviews before my visit on 17th Oct' [...]	swimming_pool	3
4	Went with open mind after reading recent reviews b [...]	bathroom	3
4	Went with open mind after reading recent reviews b [...]	breakfast	3
4	Went with open mind after reading recent reviews b [...]	internet	NA
4	Went with open mind after reading recent reviews b [...]	staff	3

Q1: What is the distribution of aspect labels?

```
predictions.tidy <- sentiment_eval %>% select(-starts_with("QA_")) %>%
gather("aspect", "prediction", 3:7) %>% arrange(review_id)
ggplot(predictions.tidy) + facet_wrap(~aspect) +
geom_bar(aes(x=prediction, y=..count..))
```



Q2: What is the number of aspects labels* per review?

In [25]:
```
1 sentiment_eval %>% select(-starts_with("QA_")) %>%
2 gather("aspect", "prediction", 3:7) %>% arrange(review_id)
```

review_id	review	aspect	prediction
1	Was ment to stay for 2 nights lasted one! The bath [...]	bathroom	3
1	Was ment to stay for 2 nights lasted one! The bath [...]	breakfast	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	internet	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	staff	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	swimming_pool	NA
2	On arrival walking to reception doors met by cig e [...]	bathroom	3
2	On arrival walking to reception doors met by cig e [...]	breakfast	NA
2	On arrival walking to reception doors met by cig e [...]	internet	NA
2	On arrival walking to reception doors met by cig e [...]	staff	1
2	On arrival walking to reception doors met by cig e [...]	swimming_pool	NA
3	Had read some reviews before my visit on 17th Oct' [...]	bathroom	NA
3	Had read some reviews before my visit on 17th Oct' [...]	breakfast	NA
3	Had read some reviews before my visit on 17th Oct' [...]	internet	NA
3	Had read some reviews before my visit on 17th Oct' [...]	staff	1
3	Had read some reviews before my visit on 17th Oct' [...]	swimming_pool	3
4	Went with open mind after reading recent reviews b [...]	bathroom	3
4	Went with open mind after reading recent reviews b [...]	breakfast	3
4	Went with open mind after reading recent reviews b [...]	internet	NA
4	Went with open mind after reading recent reviews b [...]	staff	3

* or any other interesting aggregate function

Q2: What is the number of aspects labels* per review?

```
predictions.tidy %>% group_by(review) %>%
summarise(
    num_predictions = sum(!is.na(prediction))
)
```

review	num_predictions
1st my wife asked for help with lugagges and told [...]	2
A very good hotel with a wonderful pool,, excellen [...]	2
After our problem with check-in our stay was quite [...]	1
After three weeks in Vietnam, spending a few days [...]	2
Air conditioners in the rooms are very loud other [...]	1
All fine, its a huge hotel really nice location, v [...]	1
All I can say is "WOW"....the staff is amazing!! [...]	1
All of the staff at this hotel were amazingly frie [...]	1
All staff friendly & helpful. Room was comfortable [...]	1
Almost everything about this place is 5 star. Our [...]	1
Amazing grounds. Amenities and staff were amazing. [...]	1
An error was made on the 2 rooms booked in but the [...]	0
An old property, but the pool is adequate, the bed [...]	2
Animal friendly, staff was friendly and helpful. W [...]	1

Q2: What is the number of aspects labels* per review?

```r
predictions.tidy %>% group_by(review) %>%
summarise(
    num_predictions = sum(!is.na(prediction)),
    max_prediction = max(prediction, na.rm = T),
    min_prediction = min(prediction, na.rm = T),
    avg_prediction = mean(prediction, na.rm = T),
    max_min_diff = max_prediction - min_prediction,
    sd_prediction = sd(prediction, na.rm = T)
)
```

review	num_predictions	max_prediction	min_prediction	avg_prediction	max_min_diff	sd_prediction
1st my wife asked for help with lugagges and told [...]	2	3	3	3.0	0	0.0000000
A very good hotel with a wonderful pool,, excellen [...]	2	3	3	3.0	0	0.0000000
After our problem with check-in our stay was quite [...]	1	3	3	3.0	0	NA
After three weeks in Vietnam, spending a few days [...]	2	3	2	2.5	1	0.7071068
Air conditioners in the rooms are very loud other [...]	1	2	2	2.0	0	NA
All fine, its a huge hotel really nice location, v [...]	1	3	3	3.0	0	NA
All I can say is "WOW"....the staff is amazing!! [...]	1	3	3	3.0	0	NA
All of the staff at this hotel were amazingly frie [...]	1	3	3	3.0	0	NA
All staff friendly & helpful. Room was comfortable [...]	1	3	3	3.0	0	NA
Almost everything about this place is 5 star. Our [...]	1	1	1	1.0	0	NA
Amazing grounds. Amenities and staff were amazing. [...]	1	3	3	3.0	0	NA
An error was made on the 2 rooms booked in but the [...]	0	-Inf	Inf	NaN	-Inf	NA
An old property, but the pool is adequate, the bed [...]	2	3	3	3.0	0	0.0000000

Q3: How much do the two opinions agree per aspect?

```
In [6]:  1  head(sentiment_eval)
```

review_id	review	bathroom	QA_bathroom	breakfast	QA_breakfast	internet	QA_internet	staff	QA_staff	swimming_pool	QA_swimming_pool
1	Was ment to stay for 2 nights lasted one! The bath [...]	3	2	NA	NA	NA	NA	NA	NA	NA	NA
2	On arrival walking to reception doors met by cig e [...]	3	2	NA	NA	NA	NA	1	2	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	NA	NA	NA	NA	NA	NA	1	3	3	3
4	Went with open mind after reading recent reviews b [...]	3	1	3	3	NA	NA	3	3	NA	NA
5	Expensive rooms did not match expectations. Breakf [...]	NA	NA	NA	3	NA	NA	NA	NA	NA	NA
6	Overall, great hotel with big rooms and toilets, g [...]	3	2	NA	NA	NA	NA	NA	NA	NA	NA

Q3: How much do the two opinions agree per aspect?

review_id	review	aspect	prediction	qa
1	Was ment to stay for 2 nights lasted one! The bath [...]	bathroom	3	2
1	Was ment to stay for 2 nights lasted one! The bath [...]	breakfast	NA	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	internet	NA	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	staff	NA	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	swimming_pool	NA	NA
2	On arrival walking to reception doors met by cig e [...]	bathroom	3	2
2	On arrival walking to reception doors met by cig e [...]	breakfast	NA	NA
2	On arrival walking to reception doors met by cig e [...]	internet	NA	NA
2	On arrival walking to reception doors met by cig e [...]	staff	1	2
2	On arrival walking to reception doors met by cig e [...]	swimming_pool	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	bathroom	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	breakfast	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	internet	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	staff	1	3
3	Had read some reviews before my visit on 17th Oct' [...]	swimming_pool	3	3

Goal dataframe

Q3: How much do the two opinions agree per aspect?

review_id	review	aspect	prediction
1	Was ment to stay for 2 nights lasted one! The bath [...]	bathroom	3
1	Was ment to stay for 2 nights lasted one! The bath [...]	breakfast	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	internet	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	staff	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	swimming_pool	NA
2	On arrival walking to reception doors met by cig e [...]	bathroom	3

```
qa.tidy <- sentiment_eval %>% select(review_id, review, starts_with("QA_")) %>%
gather("aspect", "qa", 3:7) %>% arrange(review_id)
head(qa.tidy)
```

review_id	review	aspect	qa
1	Was ment to stay for 2 nights lasted one! The bath [...]	QA_bathroom	2
1	Was ment to stay for 2 nights lasted one! The bath [...]	QA_breakfast	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	QA_internet	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	QA_staff	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	QA_swimming_pool	NA
2	On arrival walking to reception doors met by cig e [...]	QA_bathroom	2

Q3: How much do the two opinions agree per aspect?

review_id	review	aspect	prediction
1	Was ment to stay for 2 nights lasted one! The bath [...]	bathroom	3
1	Was ment to stay for 2 nights lasted one! The bath [...]	breakfast	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	internet	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	staff	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	swimming_pool	NA
2	On arrival walking to reception doors met by cig e [...]	bathroom	3

```
qa.tidy <- qa.tidy %>% mutate(aspect = str_replace(aspect, "QA_", ""))
head(qa.tidy)
```

review_id	review	aspect	qa
1	Was ment to stay for 2 nights lasted one! The bath [...]	bathroom	2
1	Was ment to stay for 2 nights lasted one! The bath [...]	breakfast	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	internet	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	staff	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	swimming_pool	NA
2	On arrival walking to reception doors met by cig e [...]	bathroom	2

Q3: How much do the two opinions agree per aspect?

```
evaluation <- inner_join(predictions.tidy, qa.tidy, by=c('review_id', 'review', 'aspect'))
evaluation
```

review_id	review	aspect	prediction	qa
1	Was ment to stay for 2 nights lasted one! The bath [...]	bathroom	3	2
1	Was ment to stay for 2 nights lasted one! The bath [...]	breakfast	NA	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	internet	NA	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	staff	NA	NA
1	Was ment to stay for 2 nights lasted one! The bath [...]	swimming_pool	NA	NA
2	On arrival walking to reception doors met by cig e [...]	bathroom	3	2
2	On arrival walking to reception doors met by cig e [...]	breakfast	NA	NA
2	On arrival walking to reception doors met by cig e [...]	internet	NA	NA
2	On arrival walking to reception doors met by cig e [...]	staff	1	2
2	On arrival walking to reception doors met by cig e [...]	swimming_pool	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	bathroom	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	breakfast	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	internet	NA	NA
3	Had read some reviews before my visit on 17th Oct' [...]	staff	1	3
3	Had read some reviews before my visit on 17th Oct' [...]	swimming_pool	3	3

Q3: How much do the two opinions agree per aspect?

```r
evaluation %>% mutate(agree = prediction == qa %>%
group_by(aspect, agree) %>% summarise(
    count = n()
)
```

aspect	agree	count
bathroom	FALSE	20
bathroom	TRUE	15
bathroom	NA	535
breakfast	FALSE	6
breakfast	TRUE	37
breakfast	NA	527
internet	FALSE	4
internet	TRUE	11
internet	NA	555
staff	FALSE	36
staff	TRUE	172
staff	NA	362
swimming_pool	FALSE	9
swimming_pool	TRUE	41
swimming_pool	NA	520

```r
evaluation %>% filter(!is.na(prediction) | !is.na(qa)) %>%
mutate(agree = if_else(
        is.na(prediction == qa), F, prediction == qa
    )
) %>%
group_by(aspect, agree) %>% summarise(
    count = n()
)
```

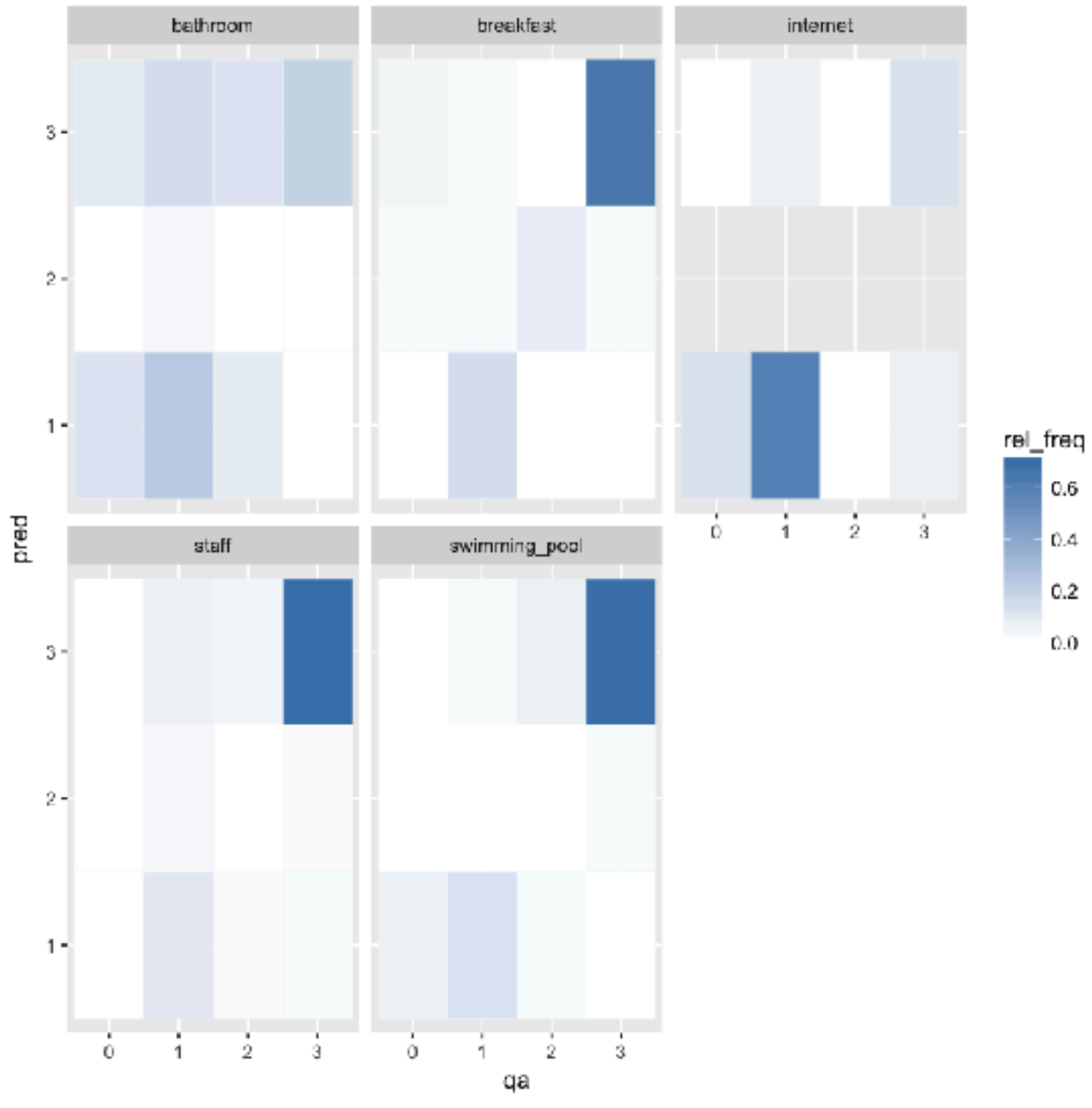
aspect	agree	count
bathroom	FALSE	90
bathroom	TRUE	15
breakfast	FALSE	49
breakfast	TRUE	37
internet	FALSE	27
internet	TRUE	11
staff	FALSE	249
staff	TRUE	172
swimming_pool	FALSE	65
swimming_pool	TRUE	41

Q3: How much do the two opinions agree per aspect?

```r
evaluation %>% mutate(agree = case_when(
    is.na(prediction) & is.na(qa) ~ T,
    is.na(prediction == qa) ~ F,
    !is.na(prediction == qa) ~ prediction == qa
)) %>%
group_by(aspect, agree) %>% summarise(
    count = n()
) %>% mutate(rel_freq = count/sum(count))
```

aspect	agree	count	rel_freq
bathroom	FALSE	90	0.15789474
bathroom	TRUE	480	0.84210526
breakfast	FALSE	49	0.08596491
breakfast	TRUE	521	0.91403509
internet	FALSE	27	0.04736842
internet	TRUE	543	0.95263158
staff	FALSE	249	0.43684211
staff	TRUE	321	0.56315789
swimming_pool	FALSE	65	0.11403509
swimming_pool	TRUE	505	0.88596491

Q3: How much do the two opinions agree per aspect?



Q3: How much do the two opinions agree per aspect?

```
crosstab <- evaluation %>% group_by(aspect) %>% do(
    data_frame(crosstab=list(table(qa=.$qa, pred=.$prediction)))
)
crosstab
```

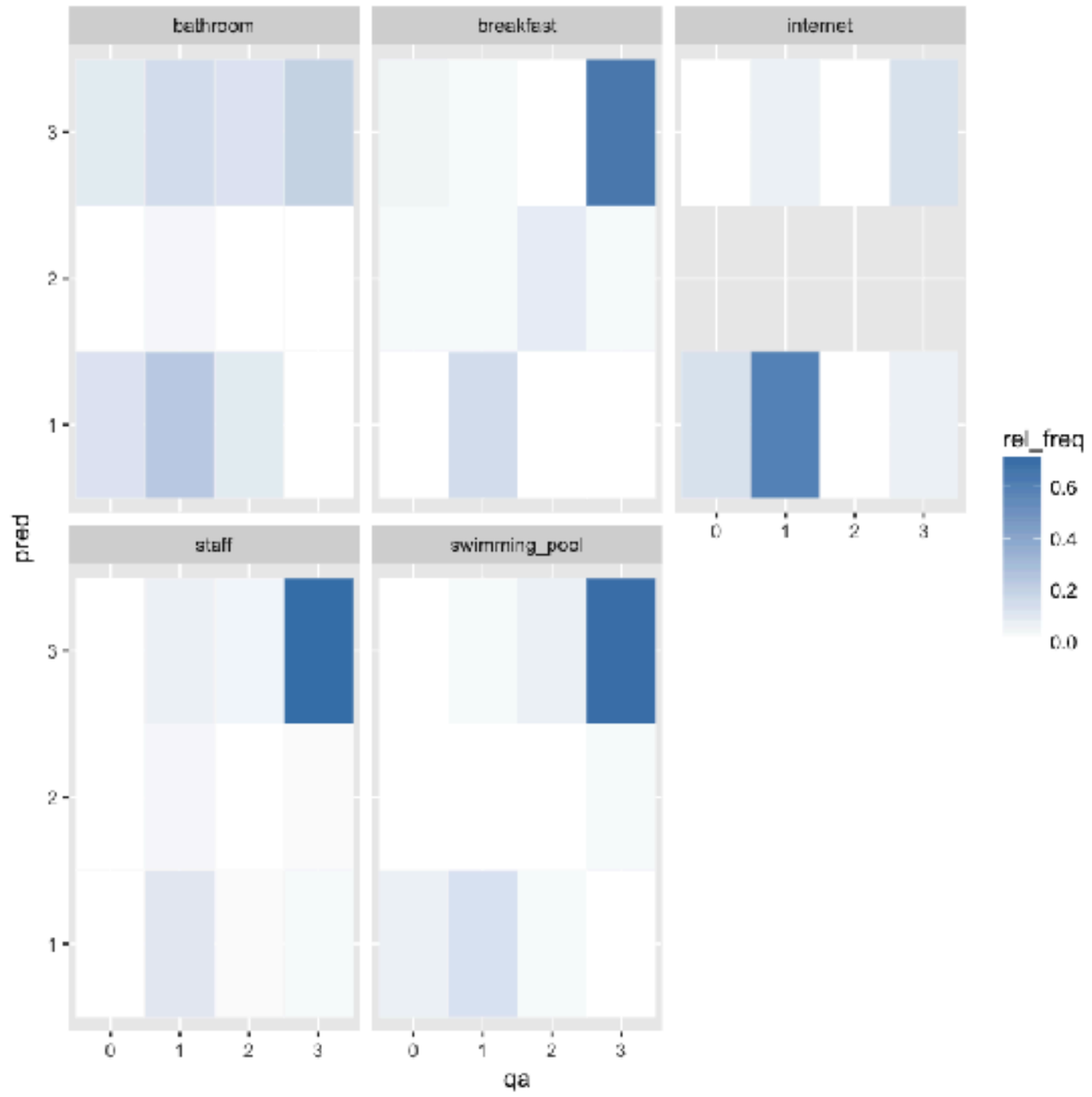
aspect	crosstab
bathroom	4, 8, 3, 0, 0, 1, 0, 0, 3, 5, 4, 7
breakfast	0, 6, 0, 0, 1, 1, 3, 1, 2, 1, 0, 28
internet	2, 9, 0, 1, 0, 1, 0, 2
staff	0, 21, 3, 5, 0, 6, 0, 3, 0, 11, 8, 151
swimming_pool	3, 6, 1, 0, 0, 0, 0, 1, 0, 1, 3, 35

```
crosstab.tidy <- crosstab %>% group_by(aspect) %>%
do(as.data.frame(.$crosstab)) %>%
group_by(aspect) %>% mutate(rel_freq = Freq/sum(Freq))
crosstab.tidy
```

aspect	qa	pred	Freq	rel_freq
bathroom	0	1	4	0.11428571
bathroom	1	1	8	0.22857143
bathroom	2	1	3	0.08571429
bathroom	3	1	0	0.00000000
bathroom	0	2	0	0.00000000
bathroom	1	2	1	0.02857143
bathroom	2	2	0	0.00000000
bathroom	3	2	0	0.00000000
bathroom	0	3	3	0.08571429
bathroom	1	3	5	0.14285714
bathroom	2	3	4	0.11428571
bathroom	3	3	7	0.20000000
breakfast	0	1	0	0.00000000

Q3: How much do the two opinions agree per aspect?

```
ggplot(crosstab.tidy, aes(qa, pred))  +
facet_wrap(~aspect) +
geom_tile(aes(fill = rel_freq), colour = "white") +
scale_fill_gradient(low = "white", high = "steelblue")
```



Textual data can be tidy too!


spa
clean
quality
located
staff
positive
service
negative
friendly
amazing
bar
restaurant
lovely
bath
free
sauna
morning
massage
close
people
bit
stayed
city
lot
wonderful
views
front
top
beautiful
hotels
bed
excellent
times
price
enjoyed
time
quiet
4
2
location
swimming
spacious
walking
wifi
dinner
star
nights
booked
return
trip
beach
town
check
stay
guests
buffet
resort
nice
loved
main
beds
kids
feel
food
10
reception
choice
3
minutes
hot
evening
day
visit
facilities
staying
water
floor
fantastic
gym
night
easy
walk
shower
coffee
view
pool
pools
relaxing
5
days
perfect
experience
highly
comfortable
helpful
family
recommend
bathroom
breakfast
restaurants


spa excellent
lovely surroundings
water villas
relations manager
nice spa
amazing helpful
services negative
amazing massage
friendly spa
incredibly accommodating
swedish massage
spa fantastic
relaxing massage
spa perfect
fantastic spa
class spa
relaxing spa
terrific spa
hotel spa
signature massage
spa negative
onsite spa
brilliant spa
spa offers
relaxing getaway
balinese massage
location amazing
comfy clean
free spa
amazing clean
treatments offered
location spa
massive bed
incredible spa
inclusive hotels
enjoyed spa
spa massages
positive pool
complete relaxation
body massage
spa wonderful
superb spa
wellness spa
amazing spa
beautiful spa
spa treatment
spa services
site spa
spa steam
spa treatments
free beach
service minded
coconut trees
lovely spa
fabulous friendly
super spa
body massages
quiet beautiful
hammam treatment
reception helped
enchantment resort
perfect spa
awesome pool
gift shops
wonderful spa
wife loved
day spa
body lotion
spa experience
service lovely
michelin star
pacific hotel
boutique resort
breakfast everyday
warm staff
body scrub
excellent spa
spa lovely
fabulous spa
excellent massage
spa facilities
stone massage
kids enjoyed
spa nice
fabulous massage
wonderful massage
excellent gym
aromatherapy massage
fantastic massage


helpful negative
extremely disappointed
hotel massage
black hair
negative pool
shower screen
negative extra
positive negative
negative spa
hotel advertises
price negative
spa appointment
negative swimming
positive comfortable
normal price
wedding venue
positive location
facilities beautiful
spa prices
ice bucket
positive staff
wash basin
positive bed
minute wait
fire alarm
kerala massage
rock hard
hotel 3
terrible service
staff told
costs 10
night due
bit dirty
mins massage
rated hotel
low quality
positive beds
spa staff
change towels
previous guest
star quality
bit disappointing
superior double
answer questions
bit awkward
2 hotels
held shower
week ago
positive spacious
positive food
negative 1
comfortable location
spacious negative
ladies changing
cold breakfast
comfortable negative
negative sauna
location negative
spa closed

References & Resources

- http://vita.had.co.nz/papers/tidy-data.pdf
- http://tidyr.tidyverse.org/articles/tidy-data.html
- Best cheatsheet ever! https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf
- tidytext book: https://www.tidytextmining.com/
- widyr: https://github.com/dgrtwo/widyr
- dplyr for Python: https://github.com/kieferk/dfply
- Pandas 'equivalents' dplyr: https://pandas.pydata.org/pandas-docs/stable/comparison_with_r.html

trivago

Min Fang
Data Scientist in Hotel Profiling
trivago Amsterdam
Email: min.fang@trivago.com