DataSience_FinalProject

Jan Steinwender & Paul Pavlis

2020-10-28

Clean the environment variables (*Temporary*)

Remove this whole header before handing the project in. This is just so that working with the document is easier

```
rm(list = ls())
```

Load needed libraries

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

Data wrangling

Data import from csv file

```
restaurant_data = read.csv(file = "restaurants_data.csv")
restaurant_data = restaurant_data %>% na_if("") # Set empty data entries to NA
```

Clean the data

Rename and remove columns

```
restaurant_data = as_tibble(restaurant_data)

restaurant_data = restaurant_data %>%
  rename("Cuisine_Style" = Cuisine.Style) %>%
  rename("Price_Range" = Price.Range) %>%
  rename("Review_Number" = Number.of.Reviews)

restaurant_data = restaurant_data %>%
  mutate(X = NULL, URL_TA = NULL, ID_TA = NULL) # Remove useless lines
```

Correct the column types

```
restaurant_data = restaurant_data %>% mutate(City = as_factor(City)) # Change from character to factor

restaurant_data = restaurant_data %>% mutate(Price_Range = as_factor(Price_Range)) # Change from charac
restaurant_data = restaurant_data %>% mutate(Price_Range = fct_recode(
  Price_Range,
  "low" = "$",
  "medium" = "$$ - $$$",
  "high" = "$$$$"
)) # Rename the levels
```

Check duplicated and NULL values

```
duplicated(restaurant_data) %>% sum()
```

```
## [1] 286
```

```
restaurant_data = restaurant_data %>% distinct() # Remove duplicates


is.na(restaurant_data) %>% sum()
```

```
## [1] 123992
```

```
sapply(restaurant_data, function(x) sum(is.na(x)))
```

```
##          Name          City Cuisine_Style       Ranking        Rating
##             0             0         31223          9373          9351
##   Price_Range Review_Number       Reviews
##         47643         17065          9337
```

Summary

```
restaurant_data
```

```
## # A tibble: 125,241 x 8
##    Name   City  Cuisine_Style  Ranking Rating Price_Range Review_Number Reviews
##    <chr>  <fct> <chr>            <dbl>  <dbl> <fct>               <dbl> <chr>
##  1 Marti~ Amst~ ['French', 'D~       1    5   medium                136 [['Just~
##  2 De Si~ Amst~ ['Dutch', 'Eu~       2    4.5 high                  812 [['Grea~
##  3 La Ri~ Amst~ ['Mediterrane~       3    4.5 high                  567 [['Sati~
##  4 Vinke~ Amst~ ['French', 'E~       4    5   high                  564 [['True~
##  5 Libri~ Amst~ ['Dutch', 'Eu~       5    4.5 high                  316 [['Best~
##  6 Ciel ~ Amst~ ['Contemporar~       6    4.5 high                  745 [['A tr~
##  7 Zaza's Amst~ ['French', 'I~       7    4.5 medium               1455 [['40th~
##  8 Blue ~ Amst~ ['Asian', 'In~       8    4.5 high                  675 [['Grea~
##  9 Teppa~ Amst~ ['Japanese', ~       9    4.5 high                  923 [['Grea~
## 10 Rob W~ Amst~ ['Dutch', 'Se~      10    4.5 low                   450 [['Exce~
## # ... with 125,231 more rows
```

```
str(restaurant_data)
```

```
## tibble [125,241 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Name         : chr [1:125241] "Martine of Martine's Table" "De Silveren Spiegel" "La Rive" "Vinkel
##  $ City         : Factor w/ 31 levels "Amsterdam","Athens",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Cuisine_Style: chr [1:125241] "['French', 'Dutch', 'European']" "['Dutch', 'European', 'Vegetarian
##  $ Ranking      : num [1:125241] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Rating       : num [1:125241] 5 4.5 4.5 5 4.5 4.5 4.5 4.5 4.5 4.5 ...
##  $ Price_Range  : Factor w/ 3 levels "medium","high",..: 1 2 2 2 2 2 1 2 2 3 ...
##  $ Review_Number: num [1:125241] 136 812 567 564 316 ...
##  $ Reviews      : chr [1:125241] "[['Just like home', 'A Warm Welcome to Wintry Amsterdam'], ['01/03/
```

```
summary(restaurant_data)
```

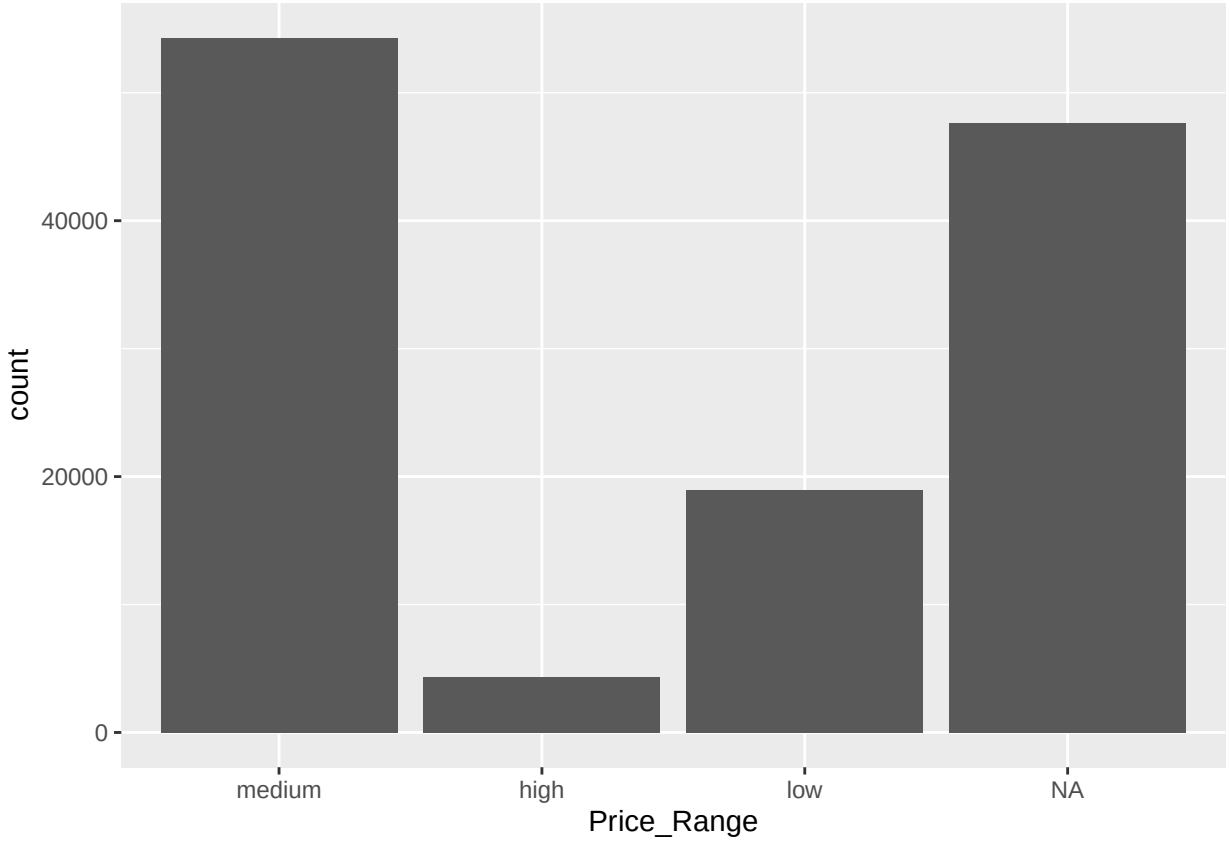
```
##      Name                 City       Cuisine_Style         Ranking
##  Length:125241      London   :18113   Length:125241      Min.   :    1
##  Class :character   Paris    :14867   Class :character   1st Qu.:  965
##  Mode  :character   Madrid   : 9524   Mode  :character   Median : 2256
##                     Barcelona: 8390                      Mean   : 3658
##                     Berlin   : 7073                      3rd Qu.: 5237
##                     Milan    : 6681                      Max.   :16444
##                     (Other)  :60593                      NA's   :9373
##      Rating       Price_Range    Review_Number       Reviews
##  Min.   :-1.000   medium:54303   Min.   :    2.0   Length:125241
##  1st Qu.: 3.500   high  : 4306   1st Qu.:    9.0   Class :character
##  Median : 4.000   low   :18989   Median :   32.0   Mode  :character
##  Mean   : 3.987   NA's  :47643   Mean   :  125.2
##  3rd Qu.: 4.500                  3rd Qu.:  114.0
##  Max.   : 5.000                  Max.   :16478.0
##  NA's   :9351                    NA's   :17065
```

Visualisation

...

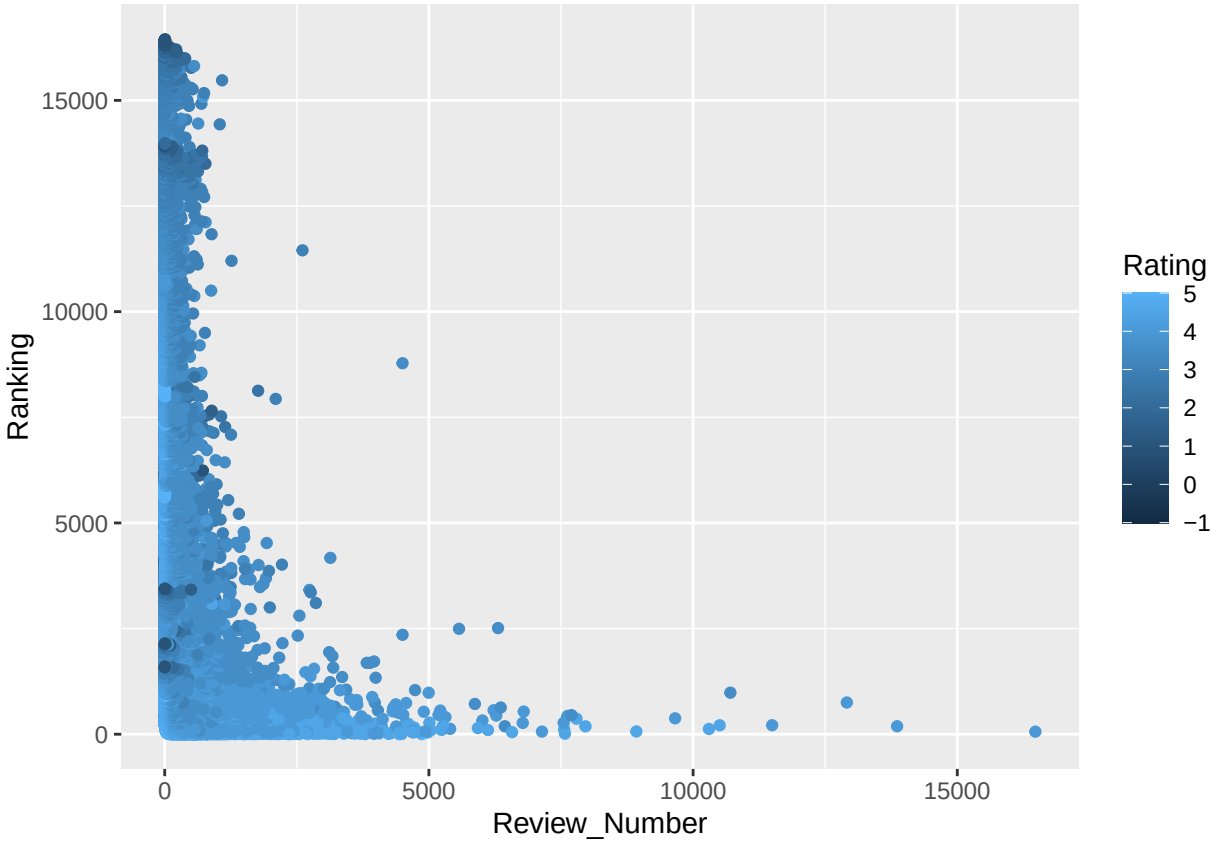
Beispiele zum testen ob es eh funktioniert:

```
ggplot(restaurant_data, aes(x = Price_Range)) + geom_bar()
```




```
ggplot(restaurant_data, aes(x = Review_Number, y = Ranking)) + geom_point(aes(col = Rating))
```

```
## Warning: Removed 17093 rows containing missing values (geom_point).
```


15000
10000
5000
0
Ranking
0
5000
10000
15000
Review_Number
Rating
5
4
3
2
1
0
−1