

Data_exploration

Group_4

2024-03-23

Data exploration

This is an R Markdown document where the exploration of the raw data can be found.

Summary statistics

We start with some summary statistics based on the raw data to see for example which variables are characters and to see some basic statistical insights, such as means and medians.

```
summary(all_data)
```

```
##      host_id      host_url      host_name      host_since
## Min.      :    275  Length:184734  Length:184734  Min.      :2008-04-17
## 1st Qu.: 28033691  Class :character  Class :character  1st Qu.:2015-03-01
## Median :105263166  Mode  :character  Mode  :character  Median :2016-12-02
## Mean      :176862777
## 3rd Qu.:310121433
## Max.      :552019444
##                                     NA's      :17
## host_location      host_about      host_response_time host_response_rate
## Length:184734      Length:184734      Length:184734      Length:184734
## Class :character    Class :character    Class :character    Class :character
## Mode  :character    Mode  :character    Mode  :character    Mode  :character
##
##
##
## host_acceptance_rate host_is_superhost host_thumbnail_url host_picture_url
## Length:184734      Mode :logical      Length:184734      Length:184734
## Class :character    FALSE:137282      Class :character    Class :character
## Mode  :character    TRUE :47241      Mode  :character    Mode  :character
##                                     NA's :211
##
##
##
## host_neighbourhood host_listings_count host_total_listings_count
## Length:184734      Min.      : 1.00      Min.      : 1.00
## Class :character    1st Qu.: 1.00      1st Qu.: 1.00
## Mode  :character    Median : 2.00      Median : 2.00
##                                     Mean      : 36.21      Mean      : 58.29
```

```
##          3rd Qu.: 7.00      3rd Qu.: 8.00
##          Max.    :3981.00    Max.    :9722.00
##          NA's    :17        NA's    :17
## host_verifications host_has_profile_pic host_identity_verified
## Length:184734      Mode :logical      Mode :logical
## Class :character   FALSE:5540         FALSE:16347
## Mode  :character   TRUE :179177        TRUE :168370
##                   NA's :17          NA's :17
##
##
##
## review_scores_rating review_scores_accuracy review_scores_cleanliness
## Min.    :0.00        Min.    :0.00        Min.    :0.00
## 1st Qu.:4.67        1st Qu.:4.73        1st Qu.:4.60
## Median :4.86        Median :4.89        Median :4.86
## Mean   :4.74        Mean   :4.78        Mean   :4.71
## 3rd Qu.:5.00        3rd Qu.:5.00        3rd Qu.:5.00
## Max.    :5.00        Max.    :5.00        Max.    :5.00
## NA's    :38943      NA's    :38981      NA's    :38973
## review_scores_checkin review_scores_communication review_scores_location
## Min.    :0.00        Min.    :0.00        Min.    :0.00
## 1st Qu.:4.80        1st Qu.:4.83        1st Qu.:4.68
## Median :4.94        Median :4.96        Median :4.88
## Mean   :4.83        Mean   :4.84        Mean   :4.77
## 3rd Qu.:5.00        3rd Qu.:5.00        3rd Qu.:5.00
## Max.    :5.00        Max.    :5.00        Max.    :5.00
## NA's    :39000      NA's    :38980      NA's    :38998
## review_scores_value reviews_per_month Region_Dataset Country_Dataset
## Min.    :0.00        Min.    : 0.01      Length:184734    Length:184734
## 1st Qu.:4.53        1st Qu.: 0.20      Class :character  Class :character
## Median :4.75        Median : 0.54      Mode  :character  Mode  :character
## Mean   :4.65        Mean   : 1.01
## 3rd Qu.:4.92        3rd Qu.: 1.28
## Max.    :5.00        Max.    :61.92
## NA's    :39003      NA's    :39096
```

Observations per region

Several datasets have been merged to create the final France and Greece dataset. Therefore, we look at how many observations each country and each region has in the dataset.

```
# Observations per country
country_counts <- table(all_data$Country_Dataset)
print(country_counts)
```

```
##
## France Greece
## 109363 75371
```

```
# Observations per region
region_counts <- table(all_data$Region_Dataset)
print(region_counts)
```

```
##
##      Athens      Bordeaux      Crete      Lyon      Paris      Pays Basque
##      13182      11854      25416      10331      74329      12849
## South Aegean Thessaloniki
##      32698      4075
```

Observations for hosts

As this research is about the impact of hosts on guest reviews, we will provide the observation count for some of the host variables.

```
# Observations for presence of profile pictures
profilepic_counts <- table(all_data$host_has_profile_pic)
print(profilepic_counts)
```

```
##
## FALSE TRUE
## 5540 179177
```

```
# Observations for presence of verified identity
identity_verified_counts <- table(all_data$host_identity_verified)
print(identity_verified_counts)
```

```
##
## FALSE TRUE
## 16347 168370
```

```
# Observations for host response time
responsetime_counts <- table(all_data$host_response_time)
print(responsetime_counts)
```

```
##
## a few days or more      N/A      within a day within a few hours
##      4037      68377      14055      17659
##      within an hour
##      80589
```

```
# Observations for presence of superhost status
superhost_counts <- table(all_data$host_is_superhost)
print(superhost_counts)
```

```
##
## FALSE TRUE
## 137282 47241
```

Review scores per region

We will provide the average review rating per country and region

```
# Average review score per country
review_score_country <- all_data %>%
  group_by(Country_Dataset) %>%
  summarise(mean_review_score = mean(review_scores_rating, na.rm = TRUE))

review_score_country
```

```
## # A tibble: 2 x 2
##   Country_Dataset mean_review_score
##   <chr>           <dbl>
## 1 France         4.71
## 2 Greece         4.79
```

```
# Average review score per region
review_score_region <- all_data %>%
  group_by(Country_Dataset, Region_Dataset) %>%
  summarise(mean_review_score = mean(review_scores_rating, na.rm = TRUE))
```

'summarise()' has grouped output by 'Country_Dataset'. You can override using the '.groups' argument.

```
review_score_region
```

```
## # A tibble: 8 x 3
## # Groups:   Country_Dataset [2]
##   Country_Dataset Region_Dataset mean_review_score
##   <chr>           <chr>           <dbl>
## 1 France         Bordeaux         4.75
## 2 France         Lyon             4.72
## 3 France         Paris            4.69
## 4 France         Pays Basque      4.75
## 5 Greece         Athens           4.74
## 6 Greece         Crete            4.82
## 7 Greece         South Aegean     4.81
## 8 Greece         Thessaloniki     4.71
```

Host variables + review scores

We will explore some key statistics related to some host variables and their review scores.

```
# Average review scores for the presence of a profile pic
review_score_profilepic <- all_data %>%
  group_by(host_has_profile_pic) %>%
  summarise(mean_review_score = mean(review_scores_rating, na.rm = TRUE))

review_score_profilepic
```

```
## # A tibble: 3 x 2
##   host_has_profile_pic mean_review_score
##   <lgl>                 <dbl>
```

```
## 1 FALSE 4.75
## 2 TRUE 4.74
## 3 NA 4.72
```

```
# Average review scores for the presence of a verified identity
review_score_verification <- all_data %>%
  group_by(host_identity_verified) %>%
  summarise(mean_review_score = mean(review_scores_rating, na.rm = TRUE))

review_score_verification
```

```
## # A tibble: 3 x 2
##   host_identity_verified mean_review_score
##   <lgl>                  <dbl>
## 1 FALSE 4.66
## 2 TRUE 4.75
## 3 NA 4.72
```

```
# Average review scores for the presence of a superhost status
review_score_superhost <- all_data %>%
  group_by(host_is_superhost) %>%
  summarise(mean_review_score = mean(review_scores_rating, na.rm = TRUE))

review_score_superhost
```

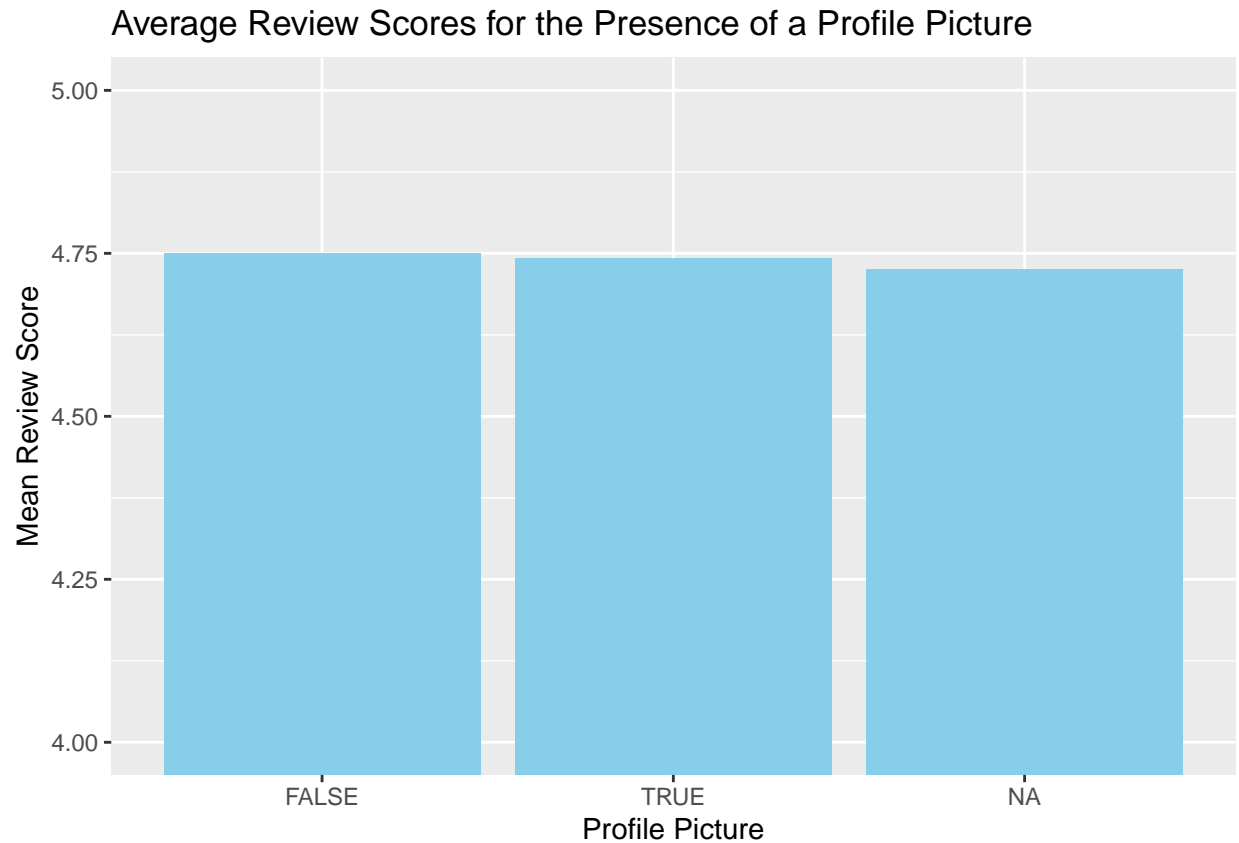
```
## # A tibble: 3 x 2
##   host_is_superhost mean_review_score
##   <lgl>                  <dbl>
## 1 FALSE 4.69
## 2 TRUE 4.87
## 3 NA 4.75
```

Visualizations for the host variables + review scores

To create a visual oversight of the host variables and their respective review scores, the following visualization have been created:

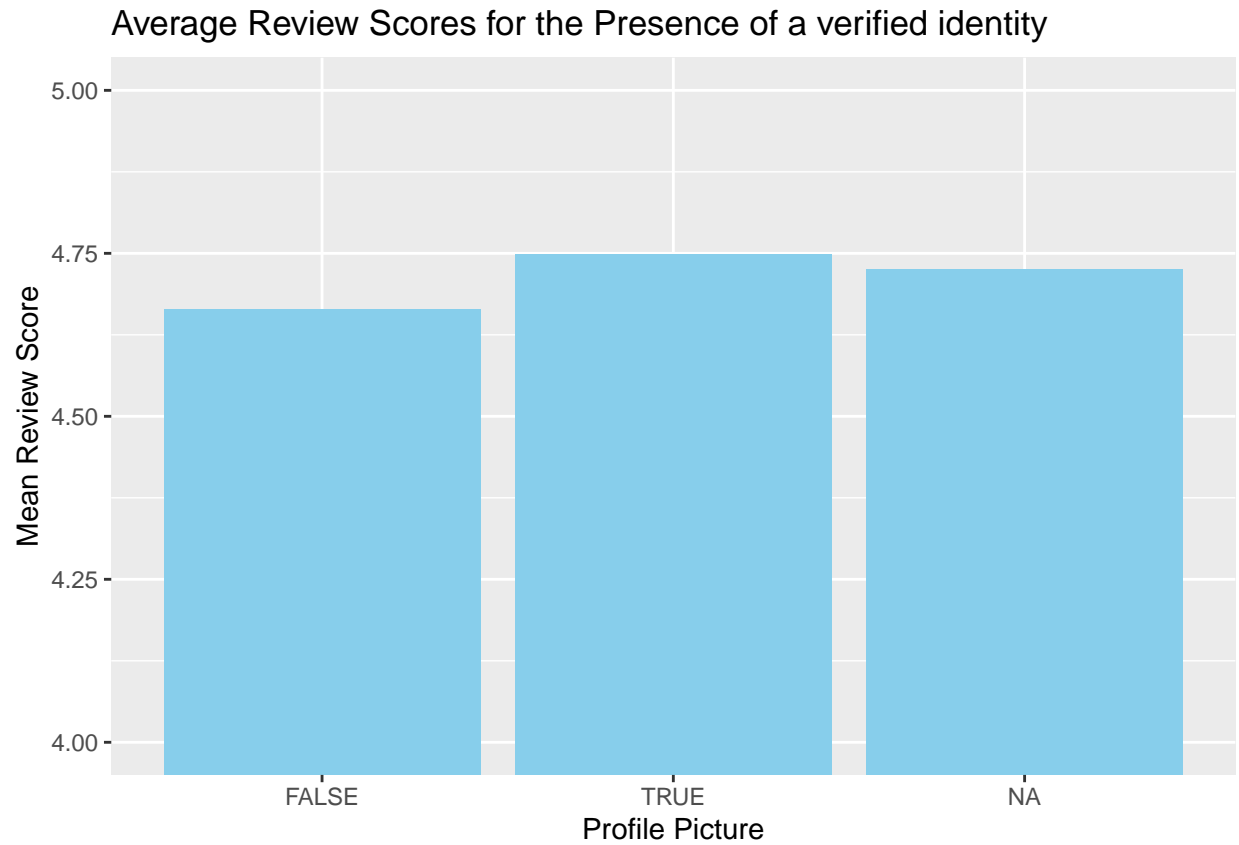
```
# Average review scores for the presence of a profile pic
bar_plot_profilepic <- ggplot(review_score_profilepic, aes(x = host_has_profile_pic, y = mean_review_score)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_cartesian(ylim = c(4, 5)) +
  labs(x = "Profile Picture", y = "Mean Review Score") +
  ggtitle("Average Review Scores for the Presence of a Profile Picture")

bar_plot_profilepic
```



```
# Average review scores for the presence of a verified identity
bar_plot_identity <- ggplot(review_score_verification, aes(x = host_identity_verified, y = mean_review_score)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_cartesian(ylim = c(4, 5)) +
  labs(x = "Profile Picture", y = "Mean Review Score") +
  ggtitle("Average Review Scores for the Presence of a verified identity")

bar_plot_identity
```



```
# Average review scores for the presence of a verified identity
bar_plot_superhost <- ggplot(review_score_superhost, aes(x = host_is_superhost, y = mean_review_score))
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_cartesian(ylim = c(4, 5)) +
  labs(x = "Profile Picture", y = "Mean Review Score") +
  ggtitle("Average Review Scores for the Presence of a superhost status")

bar_plot_superhost
```

