Coasting on Couches

0. Loading Data

0.1 Raw

```
sin_listing <- read.csv("./data/sin_listings.csv")
```

0.2 Data Wrangling

This is a function that wrangles AirBnb data into an analysable chunk. Because we will be doing the same for multiple cities, we will do a function out of this. The function is based on top of code shared in the lecture for Module 2.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(readr)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(knitr)

wrangle_airbnb_dataset <- function (raw_listing_full)
{
  listing.raw <- raw_listing_full  %>%
                 select(price,number_of_reviews,beds,bathrooms,accommodates,reviews_per_month, property_
                 rename(Reviews = number_of_reviews) %>%
                 rename(Beds = beds) %>%
                 rename(Baths = bathrooms) %>%
                 rename(Capacity = accommodates) %>%
                 rename(Monthly_Reviews = reviews_per_month) %>%
                 rename(Property_Type = property_type) %>%
                 rename(Room_Type = room_type) %>%
                 rename(Price = price) %>%
```

```
                rename(Rating = review_scores_rating) %>%
                rename(Neighbourhood = neighbourhood_cleansed) %>%
                rename(Region = neighbourhood_group_cleansed)


  listing.raw <-  listing.raw %>%
                mutate(Price = str_replace(Price, "[$]", "")) %>%
                mutate(Price = str_replace(Price, "[,]", "")) %>%
                mutate(Price = as.numeric(Price)) %>%
                mutate(Room_Type = factor(Room_Type, levels = c("Shared room", "Private room", "Entire h
                mutate(Capacity_Sqr = Capacity * Capacity) %>%
                mutate(Beds_Sqr = Beds * Beds) %>%
                mutate(Baths_Sqr = Baths * Baths) %>%
                mutate(ln_Price = log(1+Price)) %>%
                mutate(ln_Beds = log(1+Beds)) %>%
                mutate(ln_Baths = log(1+Baths)) %>%
                mutate(ln_Capacity = log(1+Capacity)) %>%
                mutate(ln_Rating = log(1+Rating)) %>%
                mutate(Shared_ind = ifelse(Room_Type == "Shared room",1,0)) %>%
                mutate(House_ind = ifelse(Room_Type == "Entire home/apt",1,0)) %>%
                mutate(Private_ind = ifelse(Room_Type == "Private room",1,0)) %>%
                mutate(Capacity_x_Shared_ind = Shared_ind * Capacity) %>%
                mutate(H_Cap = House_ind * Capacity) %>%
                mutate(P_Cap = Private_ind * Capacity) %>%
                mutate(ln_Capacity_x_Shared_ind = Shared_ind * ln_Capacity) %>%
                mutate(ln_Capacity_x_House_ind = House_ind * ln_Capacity) %>%
                mutate(ln_Capacity_x_Private_ind = Private_ind * ln_Capacity)

  return(listing.raw)
}

sin_listing.clean <- wrangle_airbnb_dataset(sin_listing)
head(sin_listing.clean)
```

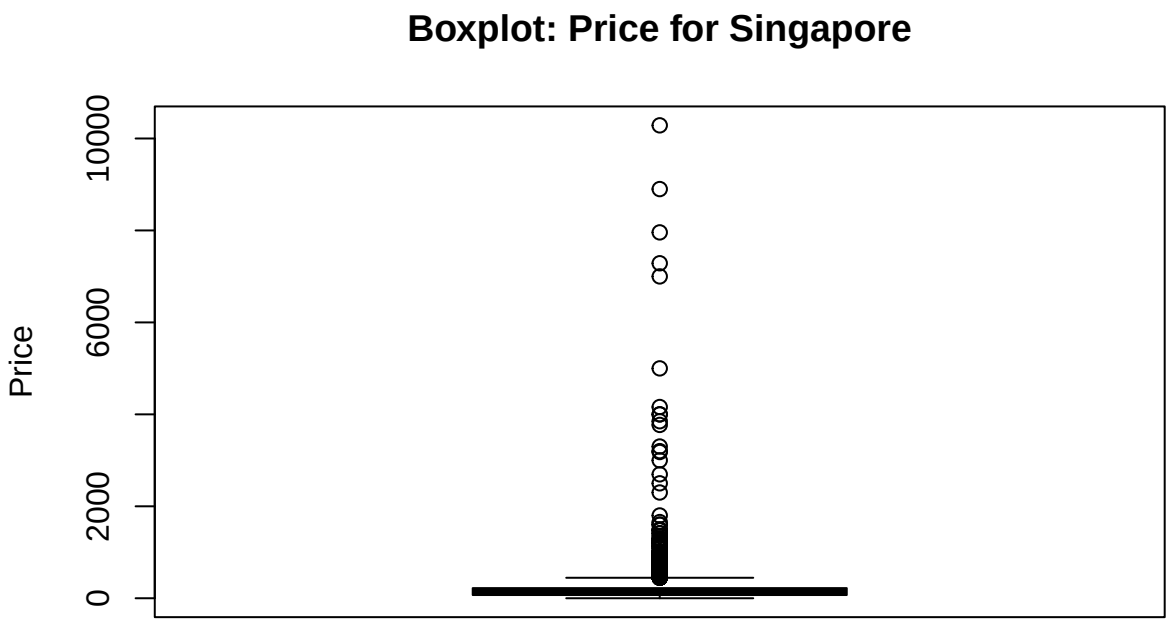
```
##   Price Reviews Beds Baths Capacity Monthly_Reviews
## 1    80      18    1    NA        2            0.19
## 2   179      20    3    NA        6            0.16
## 3    82      24    1    NA        3            0.19
## 4    82      47    2    NA        3            0.36
## 5    52      20    1    NA        1            0.19
## 6    40      13    1    NA        1            0.11
##                       Property_Type    Room_Type Rating Neighbourhood
## 1      Private room in rental unit Private room   4.56   Bukit Timah
## 2            Private room in villa Private room   4.44      Tampines
## 3 Private room in residential home Private room   4.16      Tampines
## 4 Private room in residential home Private room   4.41      Tampines
## 5      Private room in rental unit Private room   4.39   Bukit Merah
## 6      Private room in rental unit Private room   4.55   Bukit Merah
##           Region latitude longitude Capacity_Sqr Beds_Sqr Baths_Sqr ln_Price
## 1 Central Region  1.33432  103.7852           4        1        NA 4.394449
## 2    East Region  1.34537  103.9589          36        9        NA 5.192957
## 3    East Region  1.34754  103.9596           9        1        NA 4.418841
## 4    East Region  1.34531  103.9610           9        4        NA 4.418841
## 5 Central Region  1.29015  103.8081           1        1        NA 3.970292
```

```
## 6 Central Region  1.28836  103.8114            1        1        NA 3.713572
##     ln_Beds ln_Baths ln_Capacity ln_Rating Shared_ind House_ind Private_ind
## 1 0.6931472       NA   1.0986123  1.715598          0         0           1
## 2 1.3862944       NA   1.9459101  1.693779          0         0           1
## 3 0.6931472       NA   1.3862944  1.640937          0         0           1
## 4 1.0986123       NA   1.3862944  1.688249          0         0           1
## 5 0.6931472       NA   0.6931472  1.684545          0         0           1
## 6 0.6931472       NA   0.6931472  1.713798          0         0           1
##   Capacity_x_Shared_ind H_Cap P_Cap ln_Capacity_x_Shared_ind
## 1                     0     0     2                        0
## 2                     0     0     6                        0
## 3                     0     0     3                        0
## 4                     0     0     3                        0
## 5                     0     0     1                        0
## 6                     0     0     1                        0
##   ln_Capacity_x_House_ind ln_Capacity_x_Private_ind
## 1                       0                1.0986123
## 2                       0                1.9459101
## 3                       0                1.3862944
## 4                       0                1.3862944
## 5                       0                0.6931472
## 6                       0                0.6931472
```

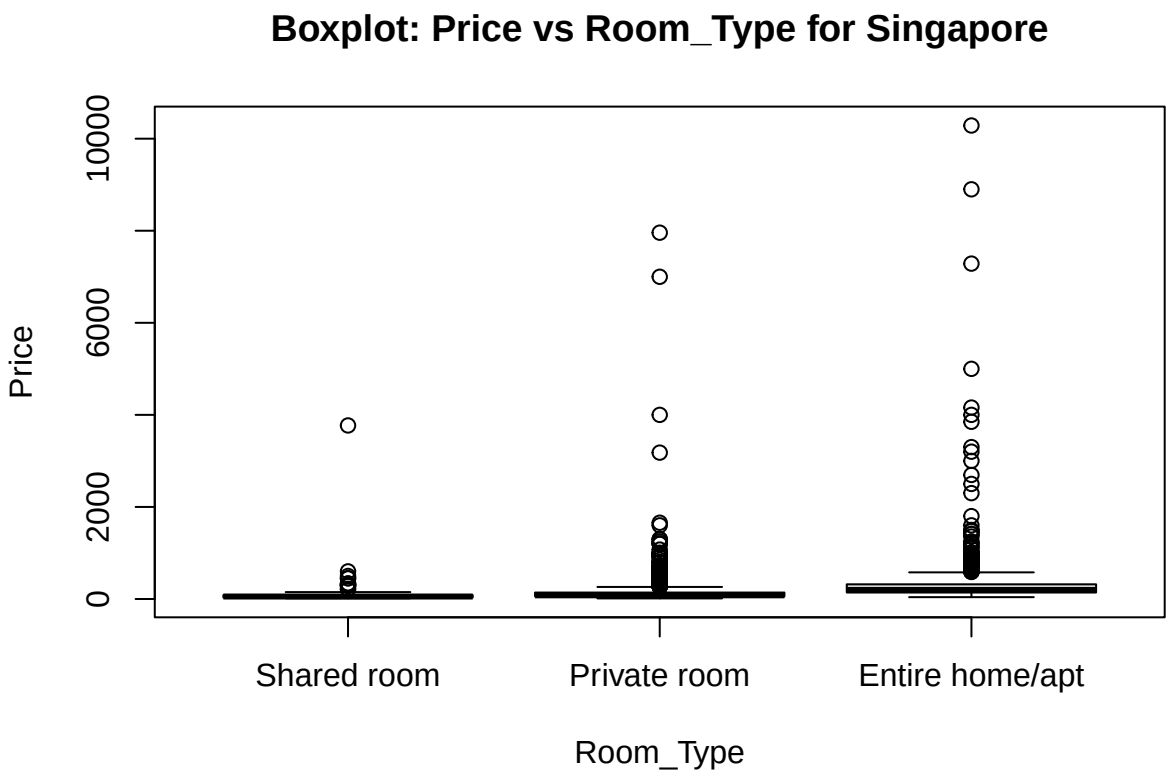
0.3 Outlier Detection

We will now check out outliers in our data for various parameters.

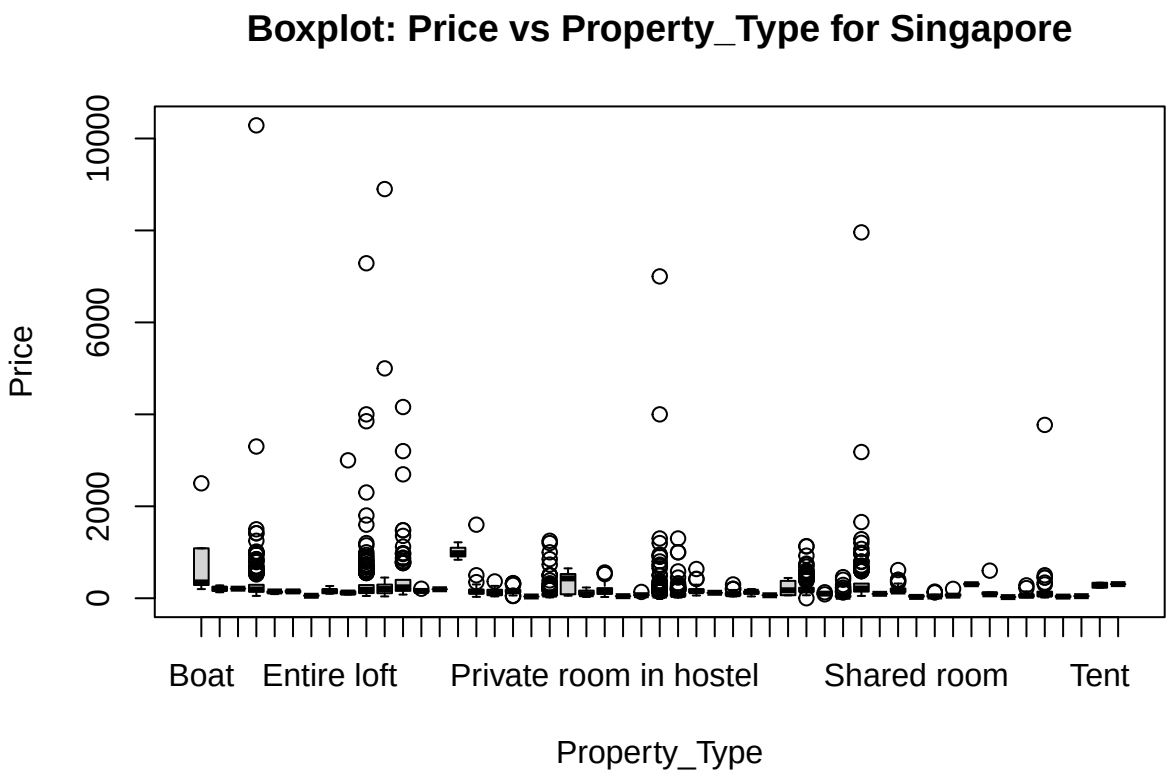
```
generate_price_boxplot <- function (listing.clean, city, comparison_col = "")
{
  # png(file = "./graphs/boxplot.png")
  if (comparison_col == "")
  {
    boxplot(listing.clean$Price, data = listing.clean, ylab="Price", main=paste("Boxplot: Price for", ci
  }
  else
    boxplot(listing.clean$Price ~ listing.clean[[comparison_col]], data = listing.clean, ylab="Price", m
  # dev.off()
}

generate_price_boxplot(sin_listing.clean, "Singapore") #, sin_listing.clean$)
```



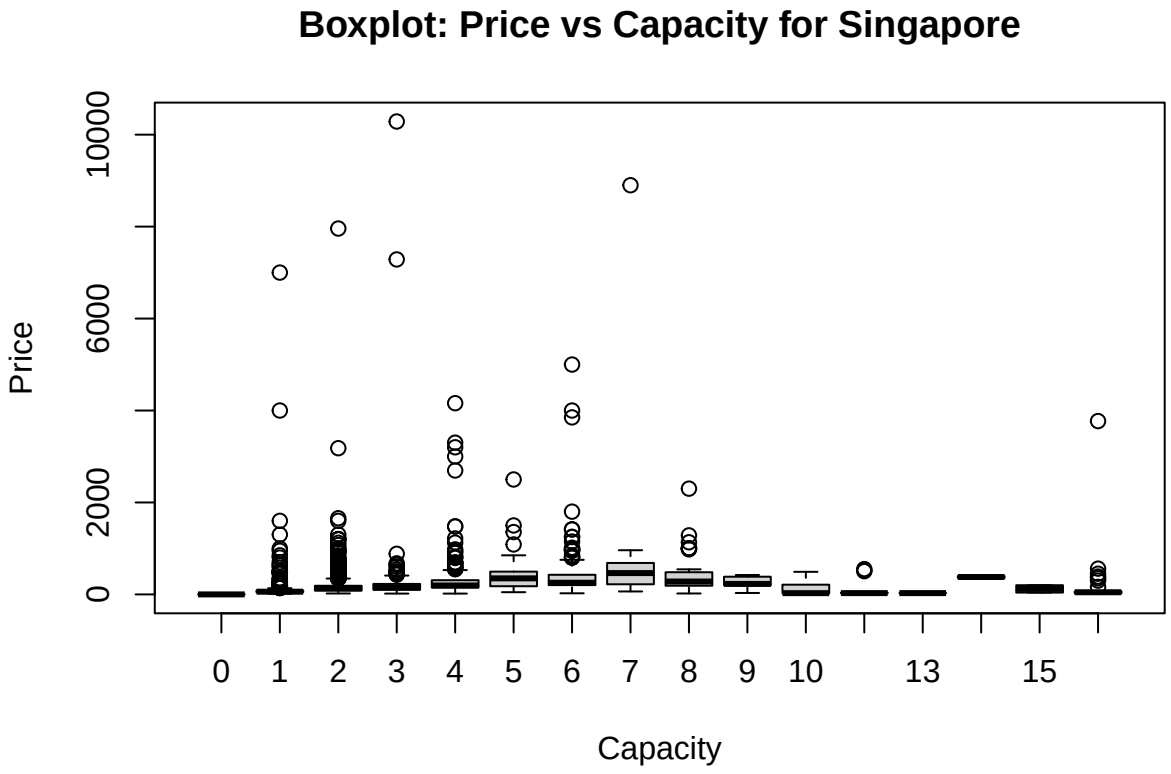

```
generate_price_boxplot(sin_listing.clean, "Singapore", "Room_Type") #, sin_listing.clean$)
```



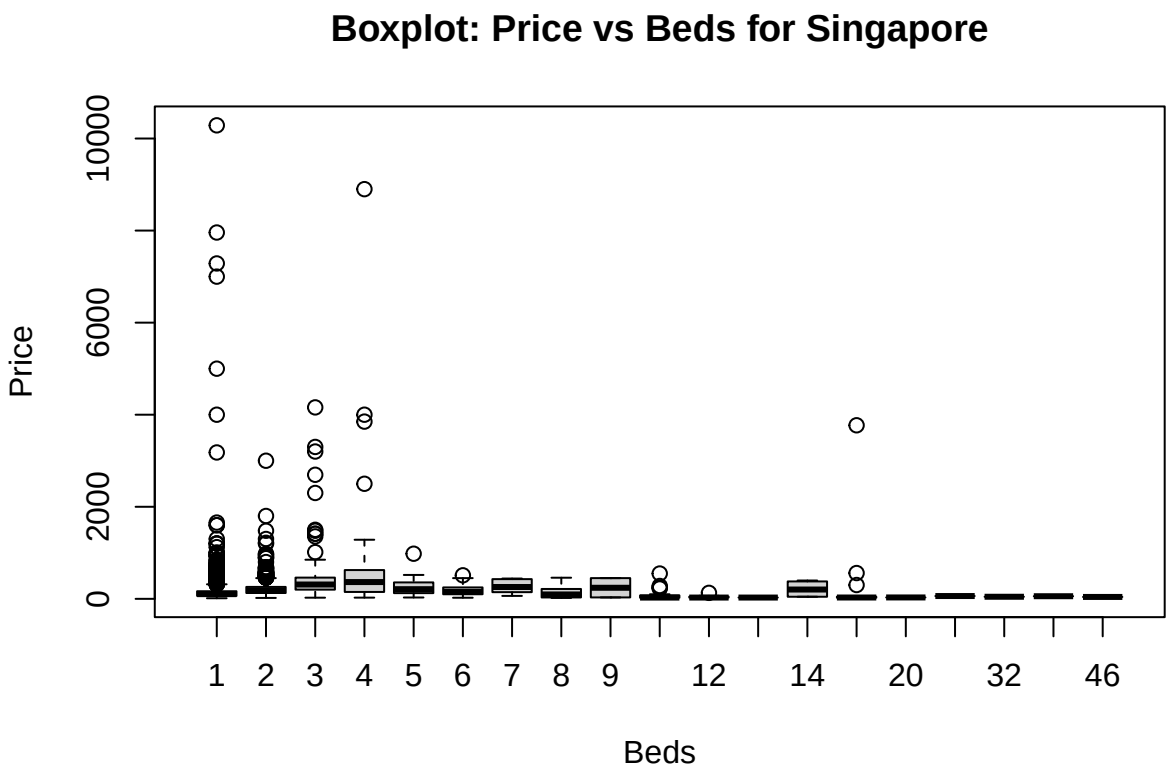

```
generate_price_boxplot(sin_listing.clean, "Singapore", "Property_Type") #, sin_listing.clean$)
```



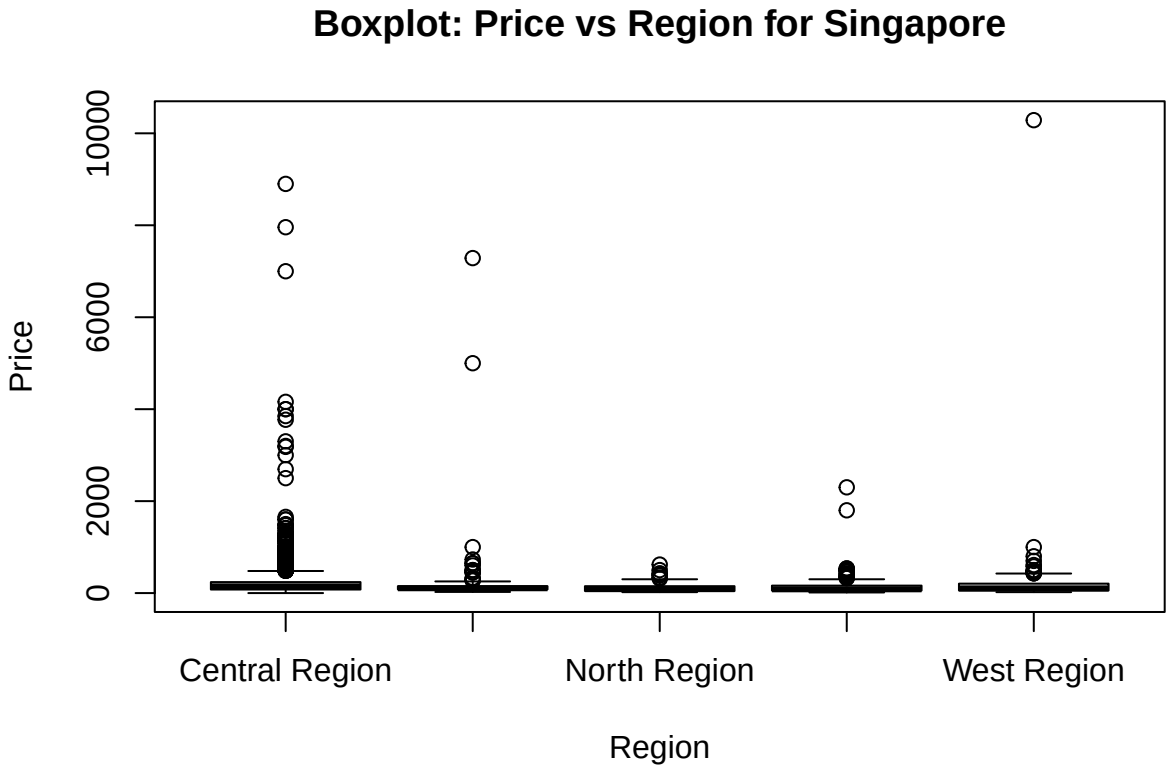

```
generate_price_boxplot(sin_listing.clean, "Singapore", "Capacity") #, sin_listing.clean$)
```




```
generate_price_boxplot(sin_listing.clean, "Singapore", "Beds") #, sin_listing.clean$)
```




```
generate_price_boxplot(sin_listing.clean, "Singapore", "Region") #, sin_listing.clean$)
```



That was a visual analysis of outliers. Clearly, there are a few offerings that are very highly priced. Let's look at them in a bit more depth, especially those units with rental prices above $5000.

```
filter(sin_listing.clean, Price > 5000)
```

```
##   Price Reviews Beds Baths Capacity Monthly_Reviews               Property_Type
## 1  7000       5    1    NA        1            0.09 Private room in rental unit
## 2  7286       5    1    NA        3            0.08          Entire rental unit
## 3  8900       0    4    NA        7              NA     Entire residential home
## 4 10286       5    1    NA        3            0.10  Entire condominium (condo)
## 5  7958       0    1    NA        2              NA               Room in hotel
##         Room_Type Rating    Neighbourhood         Region latitude longitude
## 1    Private room   4.67          Outram Central Region  1.28348  103.8414
## 2 Entire home/apt   4.50       Pasir Ris    East Region  1.37620  103.9617
## 3 Entire home/apt     NA Southern Islands Central Region  1.25312  103.8237
## 4 Entire home/apt   5.00            Tuas    West Region  1.31947  103.6483
## 5    Private room     NA          Orchard Central Region  1.30448  103.8269
##   Capacity_Sqr Beds_Sqr Baths_Sqr ln_Price   ln_Beds ln_Baths ln_Capacity
## 1            1        1        NA 8.853808 0.6931472       NA   0.6931472
## 2            9        1        NA 8.893847 0.6931472       NA   1.3862944
## 3           49       16        NA 9.093919 1.6094379       NA   2.0794415
## 4            9        1        NA 9.238636 0.6931472       NA   1.3862944
## 5            4        1        NA 8.982059 0.6931472       NA   1.0986123
##   ln_Rating Shared_ind House_ind Private_ind Capacity_x_Shared_ind H_Cap P_Cap
## 1  1.735189          0         0           1                     0     0     1
## 2  1.704748          0         1           0                     0     3     0
## 3        NA          0         1           0                     0     7     0
## 4  1.791759          0         1           0                     0     3     0
## 5        NA          0         0           1                     0     0     2
##   ln_Capacity_x_Shared_ind ln_Capacity_x_House_ind ln_Capacity_x_Private_ind
## 1                        0               0.000000                 0.6931472
## 2                        0               1.386294                 0.0000000
## 3                        0               2.079442                 0.0000000
## 4                        0               1.386294                 0.0000000
## 5                        0               0.000000                 1.0986123
```

The listing for Southern Islands was curious enough for us to want to look it up on a map.

The given coordinates point to a location in Universal Studios Singapore, specifically the Jurassic Park Rapids Adventure ride. Now the price to USS is indeed expensive, and the Jurassic Park Rapids Adventure ride is a family favourite, but whether it is worth spending $8,900 to spend a night there may indeed be debatable.

Realistically however, we suspect this is a listing from somewhere further south at Sentosa Cove, an exclusive community designed for high net-worth individuals.

Which is a great lesson for analysing place data from Singapore. We are a very compact country, and a few decimal points' worth of difference in lat-long coordinates can indeed be the difference between a much loved amusement part ride or a high-end dwelling.

We definitely need to filter by price. Let's eliminate the listings with price > 1000, as we did in the lecture.

```
# la_listing <- la_listing %>%
#               dplyr::filter(Price < 1000 , !is.na(Beds), !is.na(Baths), !is.na(Price), !is.na(Rating)
#                 dplyr::filter(Capacity < 9) %>%
#                 mutate(ln_Reviews = log(1+Reviews)) %>%
#                 mutate(ln_Monthly_Reviews = log(1+Monthly_Reviews))
```

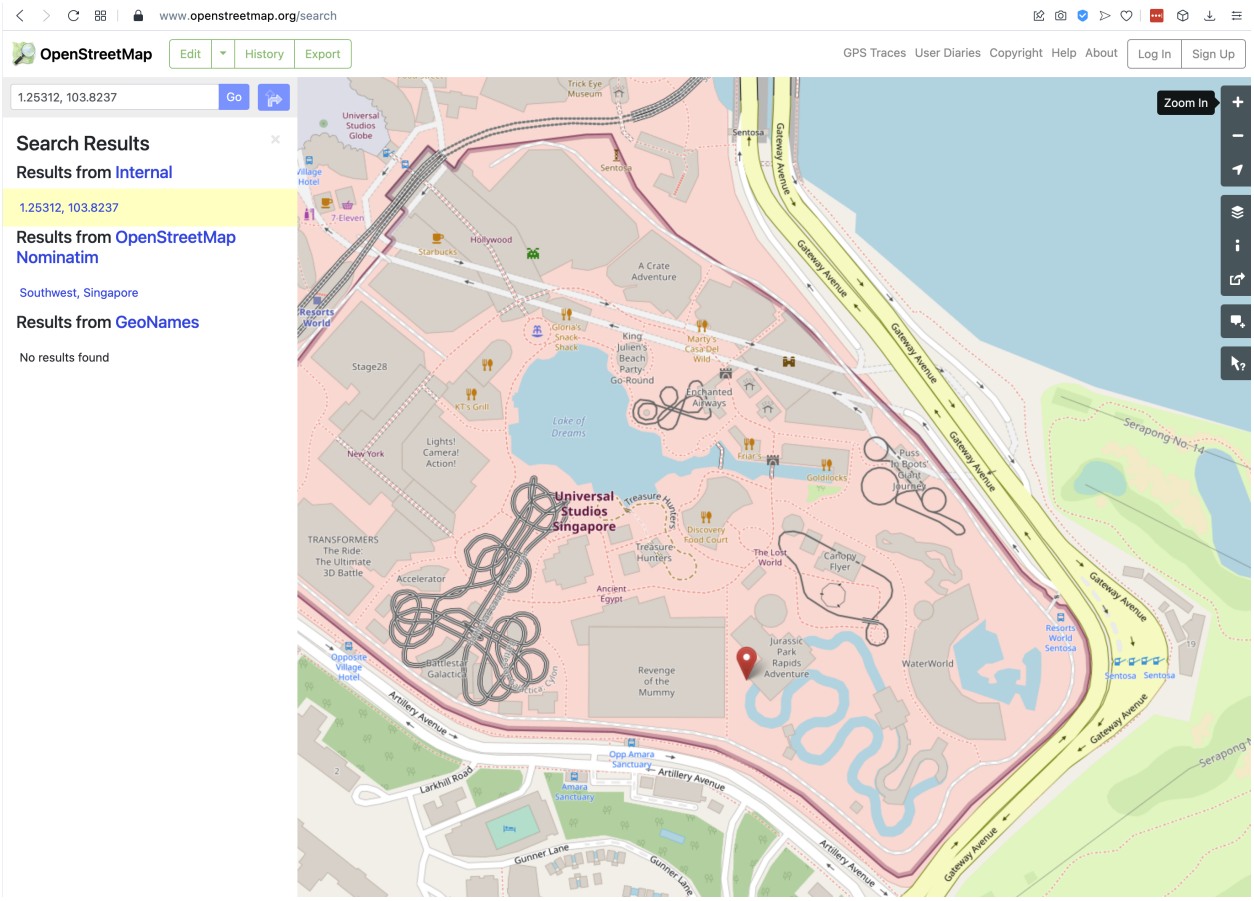


Figure 1: Southern Islands

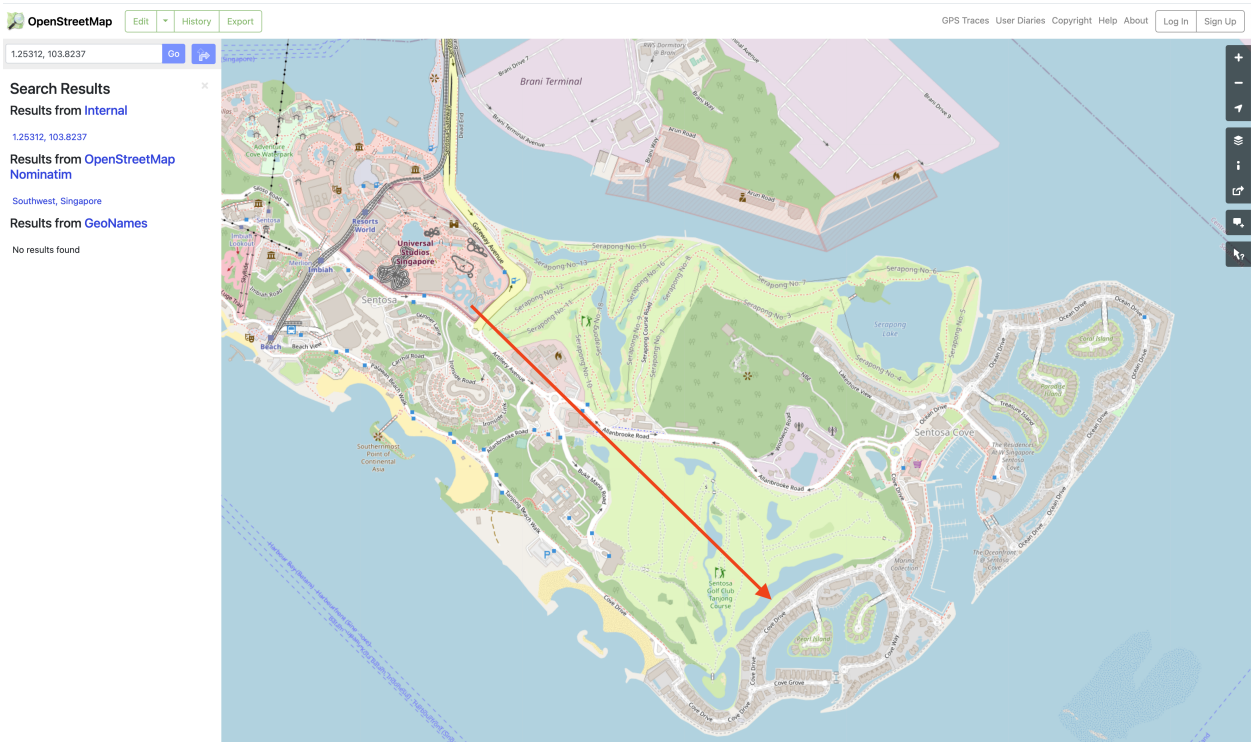


Figure 2: Sentosa Cove