Data types

1) Provide a URL to the dataset. I downloaded my dataset from http://data.insideairbnb.com/united-states/ca/san-francisco/2017-10-02/data/listings.csv.gz (http://data.insideairbnb.com/united-states/ca/san-francisco/2017-10-02/data/listings.csv.gz)

2) Explain why you chose this dataset.

I have used this dataset before for a project i was interested in.

3) What are the entities in this dataset? How many are there? The entities are specific Airbnb listings in San Francisco. There are 8706 listings.

4) How many attributes are there in this dataset? There are 95 attributes.

5) What is the datatype of each attribute (categorical -ordered or unordered-, numeric -discrete or continuous-, datetime, geolocation, other)? Write a short sentence stating how you determined the type of each attribute. Do this for at least 5 attributes, if your dataset contains more than 10 attributes, choose 10 of them to describe.

Num	Name	Type	Description
1	'ID'	Categorical	Identifier of each listing which is a finite set.
2	'Neighborhood'	Categorical Unordered	Finite set of number of neighborhoods
3	'Latitude'	Numerical Continuous	Infinite precision can be applied to yield an infinite set.
4	'Longitude'	Numerical Continuous	Infinite precision can be applied to yield an infinte set.
5	'bedrooms'	Numerical Discrete	Finite set of number of beds.
6	'Maximum nights'	Numerical Discrete	Finite set of nights listing allows.

Num	Name	Type	Description
7	'Host listings count'	Numerical Discrete	Hosts can only list a finite number of listings.
8	'Accommodates'	Numerical Discrete	listings can only accomoodate a finite number of people
9	'Bed Type'	Categorical Unordered	Finite set of bed types
10	'Number of Reviews'	Numerical Discrete	Finite set of reviews.

6) Write R code that loads the dataset using function `read_csv` *. Were you able to load the data successfully? If no, why not?* Yes but ther were some parsing errors.

```
library(tidyverse)
# loading code goes here
mydata <- read_csv("listings.csv")
```

```
## Warning: 2 parsing failures.
## row # A tibble: 2 x 5 col     row col     expected               actual    file
expected   <int> <chr>   <chr>                  <chr>     <chr>          actual 1  19
19 zipcode no trailing characters -1273     'listings.csv' file 2  2285 zipcode no tr
ailing characters "\n94158" 'listings.csv'
```

```
mydata
```

```
## # A tibble: 8,706 x 95
##         id listing_url scrape_id last_scraped name  summary    space
##      <int> <chr>           <dbl> <date>       <chr> <chr>      <chr>
##  1  1.10e7 https://ww~   2.02e13 2017-04-02   Simp~ I signed ~ This is cl~
##  2  8.05e6 https://ww~   2.02e13 2017-04-02   Spac~ Enjoy the~ "We love o~
##  3  1.40e7 https://ww~   2.02e13 2017-04-02   Priv~ Awesome l~ <NA>
##  4  1.61e7 https://ww~   2.02e13 2017-04-02   Spac~ Quiet 1 b~ <NA>
##  5  9.08e6 https://ww~   2.02e13 2017-04-02   Edwa~ Bedroom w~ <NA>
##  6  9.36e5 https://ww~   2.02e13 2017-04-02   Sunn~ Come stay~ Located ju~
##  7  7.05e6 https://ww~   2.02e13 2017-04-02   Bout~ Boutique ~ <NA>
##  8  8.00e6 https://ww~   2.02e13 2017-04-02   Pret~ Quiet one~ <NA>
##  9  1.28e7 https://ww~   2.02e13 2017-04-02   Spac~ Built in ~ <NA>
## 10  1.04e7 https://ww~   2.02e13 2017-04-02   Brig~ Located i~ Located in~
## # ... with 8,696 more rows, and 88 more variables: description <chr>,
## #   experiences_offered <chr>, neighborhood_overview <chr>, notes <chr>,
## #   transit <chr>, access <chr>, interaction <chr>, house_rules <chr>,
## #   thumbnail_url <chr>, medium_url <chr>, picture_url <chr>,
## #   xl_picture_url <chr>, host_id <int>, host_url <chr>, host_name <chr>,
## #   host_since <date>, host_location <chr>, host_about <chr>,
## #   host_response_time <chr>, host_response_rate <chr>,
## #   host_acceptance_rate <chr>, host_is_superhost <chr>,
## #   host_thumbnail_url <chr>, host_picture_url <chr>,
## #   host_neighbourhood <chr>, host_listings_count <int>,
## #   host_total_listings_count <int>, host_verifications <chr>,
## #   host_has_profile_pic <chr>, host_identity_verified <chr>,
## #   street <chr>, neighbourhood <chr>, neighbourhood_cleansed <chr>,
## #   neighbourhood_group_cleansed <chr>, city <chr>, state <chr>,
## #   zipcode <int>, market <chr>, smart_location <chr>, country_code <chr>,
## #   country <chr>, latitude <dbl>, longitude <dbl>,
## #   is_location_exact <chr>, property_type <chr>, room_type <chr>,
## #   accommodates <int>, bathrooms <dbl>, bedrooms <int>, beds <int>,
## #   bed_type <chr>, amenities <chr>, square_feet <int>, price <chr>,
## #   weekly_price <chr>, monthly_price <chr>, security_deposit <chr>,
## #   cleaning_fee <chr>, guests_included <int>, extra_people <chr>,
## #   minimum_nights <int>, maximum_nights <int>, calendar_updated <chr>,
## #   has_availability <chr>, availability_30 <int>, availability_60 <int>,
## #   availability_90 <int>, availability_365 <int>,
## #   calendar_last_scraped <date>, number_of_reviews <int>,
## #   first_review <date>, last_review <date>, review_scores_rating <int>,
## #   review_scores_accuracy <int>, review_scores_cleanliness <int>,
## #   review_scores_checkin <int>, review_scores_communication <int>,
## #   review_scores_location <int>, review_scores_value <int>,
## #   requires_license <chr>, license <chr>, jurisdiction_names <chr>,
## #   instant_bookable <chr>, cancellation_policy <chr>,
## #   require_guest_profile_picture <chr>,
## #   require_guest_phone_verification <chr>,
## #   calculated_host_listings_count <int>, reviews_per_month <dbl>
```

Wrangling

1)I am trying to find a listing on Airbnb that can accommodate at least a family of 4 (if it accommodates more, great I can bring more people but 4 is the minimum) that also has really good ratings. Price is not a concern here, so I left it off but ideally someone else would want to compare prices to get the best deal.

```
# pipeline goes here
data <- mydata %>%
select(id,review_scores_rating, accommodates) %>%
filter(review_scores_rating > 90, accommodates > 4) %>%
arrange(desc(review_scores_rating))
```

Plotting

1. I wanted to see if a review score rating was based on how many people the listing accommodated and from my results it does not appear so for review scores above 90 and accommodations above 4. On the X-axis I have the different accommodations of listings and I have the dependent variable on the Y-Axis (review score)

```
# plot goes here
data %>%
ggplot(mapping=aes(x = accommodates, y = review_scores_rating)) + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```