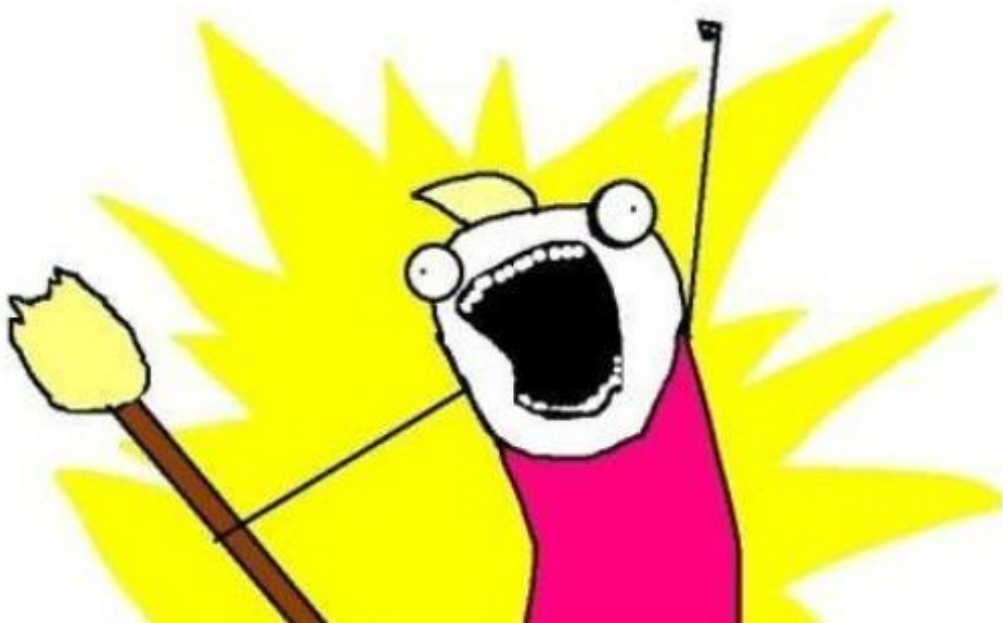


# Introduction to Exploratory Data Analysis

## What is Exploratory Data Analysis?

Exploratory Data Analysis or EDA is understanding the data sets by summarizing their main characteristics. This involves discovering patterns, spotting anomalies, testing hypotheses, and checking assumptions. This is achieved with the help of summary statistics and graphical representations. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. It often takes much time to explore the data. However, through the process of EDA, we can ask to define the problem statement or definition on our data set which is very important.

**EXPLORE ALL THE DATA!**



## How to perform Exploratory Data Analysis?

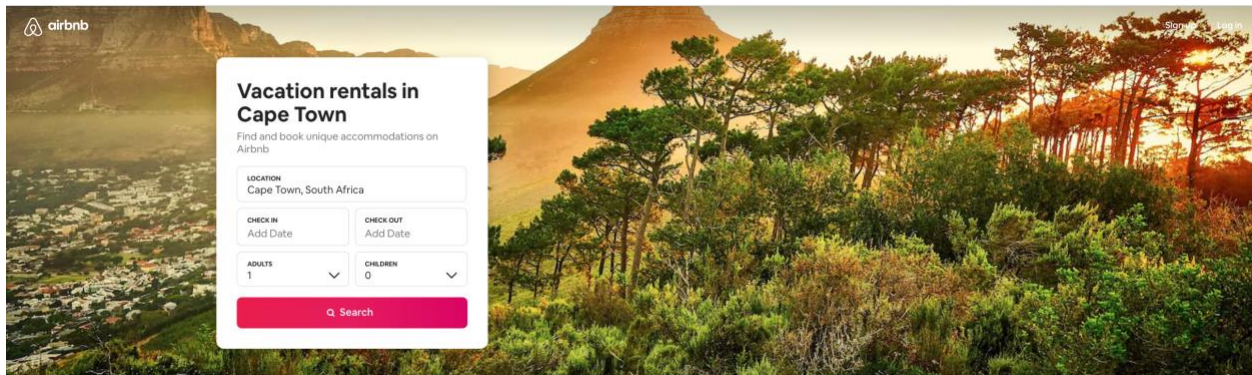
This is one a question that everyone is keen on knowing the answer to. Unfortunately, it is not that straightforward. The answer depends on the data set that you are working on. There is no one method or

common methods in order to perform EDA. However, there are a few set steps that this generally includes:

1. Loading and inspecting your data
2. Cleaning the data which includes:
  - 2.1. dropping data points and columns we don't need.
  - 2.2. checking data types and fixing if needed
  - 2.3. removing duplicates
  - 2.4. dealing with missing values
  - 2.5. looking for outliers and deciding how to deal with these
  - 2.6. reformatting columns if needed
3. Some visual exploration to look at relationships between variables or interesting insights that jump out
4. How can you add, change or remove features to get more out of your data? (aka Feature Engineering)

## AirBnB Cape Town Demo Data


For the purposes of demonstration, we are going to be working with some of AirBnB Demo data. There will be two parts that we will be going through – Data Cleaning and EDA. The code for both parts has been written and well documented on Jupyter Notebooks, which is shared at the top of the article. The rest of the article below will provide some explanation behind the choices I made during the data cleaning and EDA stages.



The image shows the Airbnb search interface for Cape Town. The background is a scenic view of a mountain and trees. The search bar is white with a pink 'Search' button. The text 'Vacation rentals in Cape Town' is displayed, along with the location 'Cape Town, South Africa'. The search criteria are set to 'CHECK IN Add Date', 'CHECK OUT Add Date', 'ADULTS 1', and 'CHILDREN 0'.


### Top-rated vacation rentals in Cape Town

Guests agree: these stays are highly rated for location, cleanliness, and more.




**SUPERHOST** Entire loft - 3 guests - 1 bed - 1.5 baths  
**Modern, Chic Penthouse with Mountain, City & Sea Views...**  
Sit back in a private plunge pool and enjoy an uninterrupted vista of Table Mountain, the city skyline, and the ocean beyond. The views are just as good from...

WHAT GUESTS ARE SAYING:



**SUPERHOST** Entire apartment - 2 guests - 1 bed - 1 bath  
**Unwind in a Bright, Airy Space with Rustic Accents**  
**QUARANTINE-FRIENDLY** - cosy at this quiet and secure retreat in the centre of Cape Town. Open the sliding doors to breathe in the ocean air after preparing a meal...

WHAT GUESTS ARE SAYING:



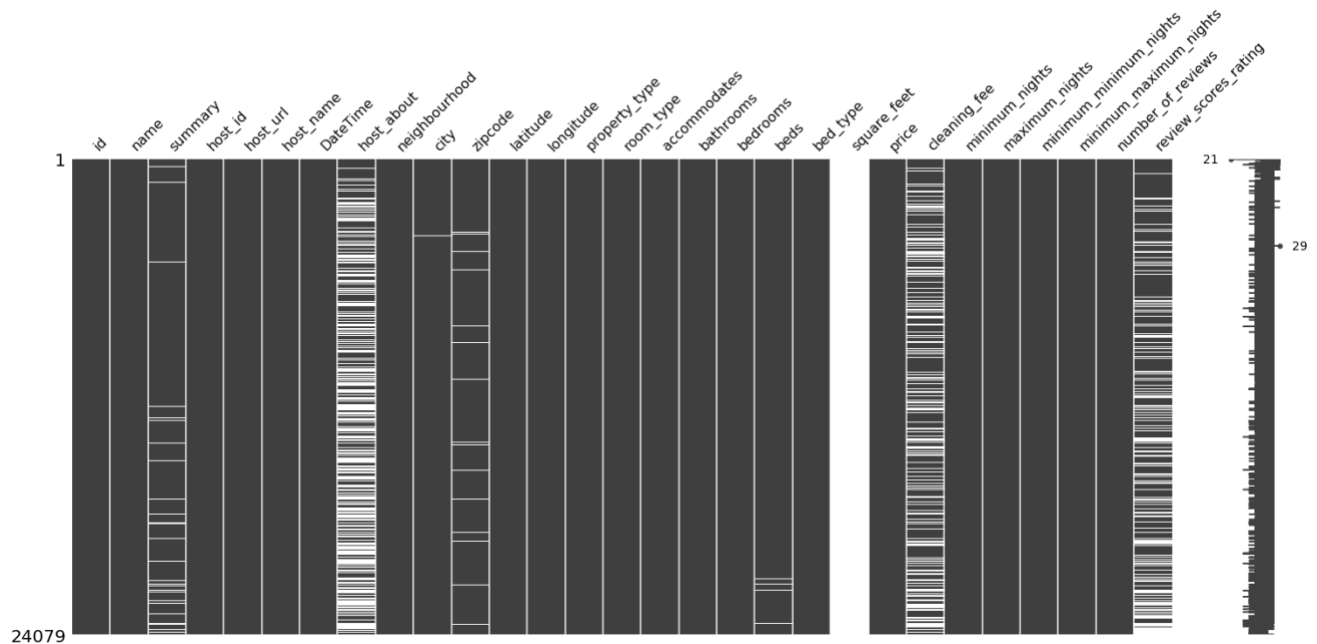
**SUPERHOST** Entire apartment - 2 guests - 1 bed - Half-bath  
**Primaview, Camps Bay, Cape Town**  
Primaview is situated in beautiful Camps Bay, Cape Town. Offering comfortable accommodation, alongside an inviting pool and surrounded by panoramic views of...

WHAT GUESTS ARE SAYING:

## Data Cleaning

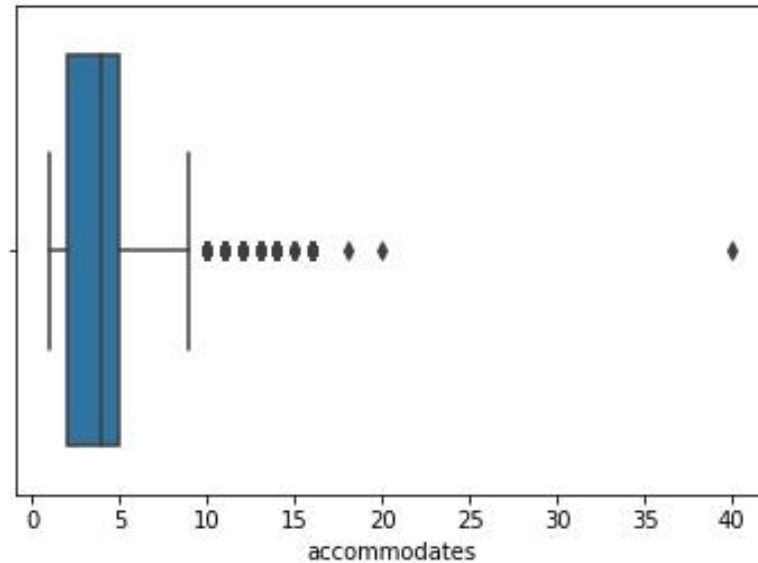
The data cleaning process can start off in many ways. This can involve simply previewing a part of the dataset to get an idea of what the data looks like, deleting highly correlated variables, checking number of unique values for the each variable in the dataset, dropping columns we don't think are important to the analysis, removing duplicated data, changing data types appropriately, dealing with missing values and outliers and much more.

Some plots can come in handy when making data cleaning related decisions such as the one below.



For example, the plot above provides an overview of the presence of data in the dataset. The white gaps represent the missing data in the columns.

Another example is by using a boxplot to find outliers. Below, I plot a boxplot for the variable 'accommodates' and find an outlier immediately.



A good tip is that, if the dataset is big, you can divide up the dataset into categorical and numerical data and work from there. When assessing categorical variables, it is good to get an idea of the cardinality:

- High cardinality = variables with few repeated values (ie all different)
- Low cardinality = many repeated values (ie almost all one type)

This will help make decisions regarding groupings or modifying data in predictive variables.

Here are some possible entities (groupings) that can be carried out with the dataset:

- Most common zipcode is 8001 and ward is 115.
- City was almost all Cape Town, so not very informative for differentiation (ie low cardinality).
- Property\_type - whole houses and apartments are the most common type.
- Room\_type - a lot of entire apartments and shared rooms.
- Bed type - predominantly real beds. Not much value in this variable.
- Accommodates - good range of sizes of properties.

## Comparisons and Aggregations

Now that the data is clean and more understandable, we arrive at the final stage of the EDA journey - “Comparisons and Aggregations” stage. This will look at the relationship between variables in the data. The goal of this stage is to get a better understanding of how our data interacts with each other.

Here are some things that can be done in this stage.

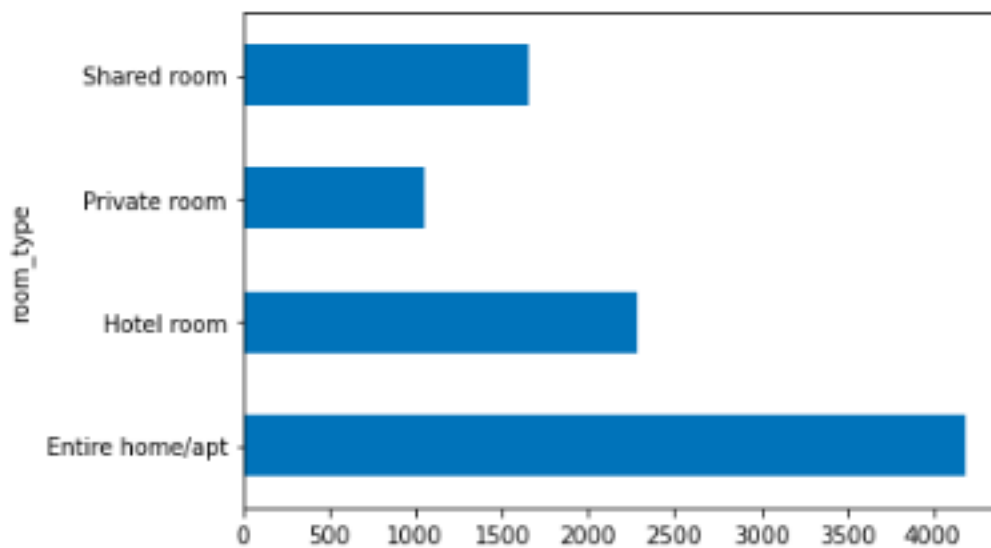
### 1. Multiple Group Counts

This involves making a pivot count with different variables in the dataset. The example below shows what a multiple group counts table looks like for the variables: “accommodates” and “room\_type”.

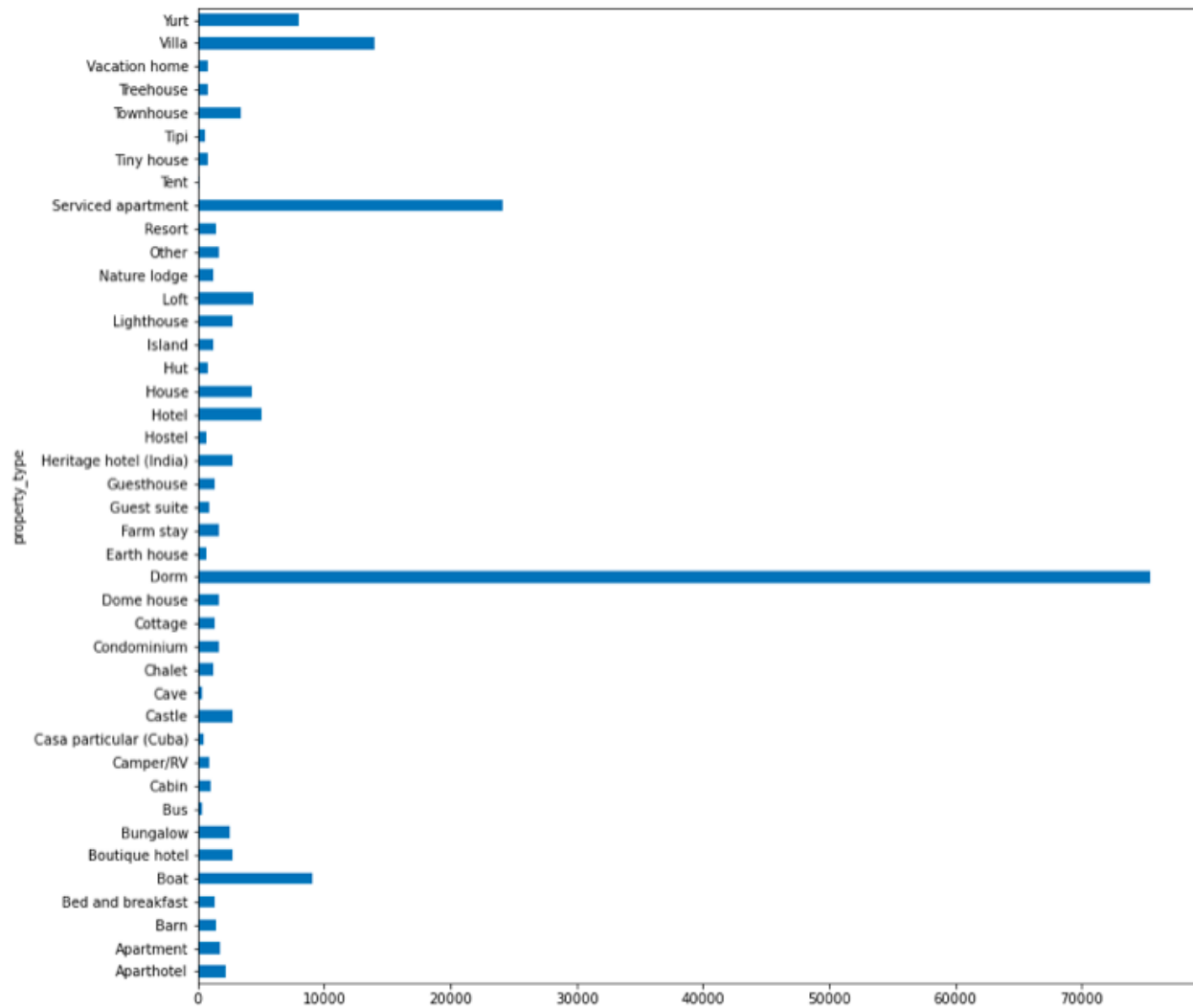
accommodates	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	20	40
room_type																			
Entire home/apt	139.0	5833.0	996.0	5186.0	850.0	2720.0	410.0	1372.0	107.0	499.0	26.0	165.0	17.0	49.0	5.0	49.0	1.0	1.0	NaN
Hotel room	12.0	232.0	36.0	77.0	5.0	31.0	1.0	7.0	NaN	2.0	NaN	1.0	NaN	1.0	NaN	3.0	NaN	NaN	NaN
Private room	712.0	3699.0	234.0	344.0	34.0	71.0	16.0	31.0	11.0	21.0	3.0	16.0	2.0	9.0	3.0	42.0	NaN	NaN	1.0
Shared room	52.0	46.0	5.0	11.0	4.0	5.0	NaN	4.0	NaN	5.0	NaN	4.0	NaN	NaN	NaN	2.0	NaN	NaN	NaN

## 2. Categorical Variable Means

An example of categorical variable means is shown below. Note that this is the average price per room type and property type.



Another very good example is the mean price of properties per property type shown below. You will notice the average price of a dorm room is R70 000 (\$3500 ish!?!). This is a good indication that we missed something there. We might need to look at outliers per room type rather than overall. This goes to show that EDA is not a linear process. Data scientists going back and forth a lot is normal.



There are many more ways to take things to the next level. However, this article on EDA provides a good base to grow from for someone starting out in the field of Data Science.