What is exploratory data analysis?

EXPLORATORY DATA ANALYSIS IN POWER BI

Jacob H. Marquez
Data Scientist at Microsoft

datacamp

What is exploratory data analysis?

> "An approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods."

[1] https://en.wikipedia.org/wiki/Exploratory_data_analysis

Six steps to EDA

1. Understanding the data structure
2. Identifying missing data
3. Describing the data with descriptive statistics & distributions
4. Identifying outliers
5. Examining and quantifying relationships between variables
6. Forming hypothesis

Six steps to EDA

1. **Understanding the data structure**
2. **Identifying missing data**
3. **Describing the data with descriptive statistics & distributions**
4. Identifying outliers
5. Examining and quantifying relationships between variables
6. Forming hypothesis

1. Understanding the data structure

Continuous

Numerical variables often able to take an infinite set of values

- Number of stars in space
- Click-through rates
- Distance between two cities

Categorical

Non-numerical variables, usually text, with two or more groups

- House types
- Country
- Company

2. Identifying missing data

Missing at random

CITY	Rainfall (inches)			
SEATTLE	2.03	1.13	0.52	4.59
	4.67		2.09	4.53
	0.42	2.60	1.90	
NYC	1.35	3.40	3.75	1.75
		3.93	0.07	3.14
	3.96	3.95		3.60
PARIS	4.72		2.27	2.68
	2.33	2.07	1.06	1.38
		4.29	4.29	1.47

Missing not at random

CITY	Rainfall (inches)			
SEATTLE	4.67	1.75	2.09	4.53
	0.42	2.60	1.90	3.14
	1.35	3.40	3.75	1.75
NYC	2.68	3.93	0.07	3.14
	3.96	3.95	0.52	3.60
	4.72	4.72	2.27	2.68
PARIS	2.33	2.07	1.06	1.38
	2.07	4.29	4.29	1.47

2. Addressing missing data

CITY	Rainfall (inches)			
SEATTLE	4.67	1.75	2.09	4.53
	0.42	2.60	1.90	3.14
	1.35	3.40	3.75	1.75
NYC	2.68	3.93	0.07	3.14
	3.96	3.95	0.52	3.60
	4.72	4.72	2.27	2.68
PARIS	2.33	2.07	1.06	1.38
	2.07	4.29	4.29	1.47

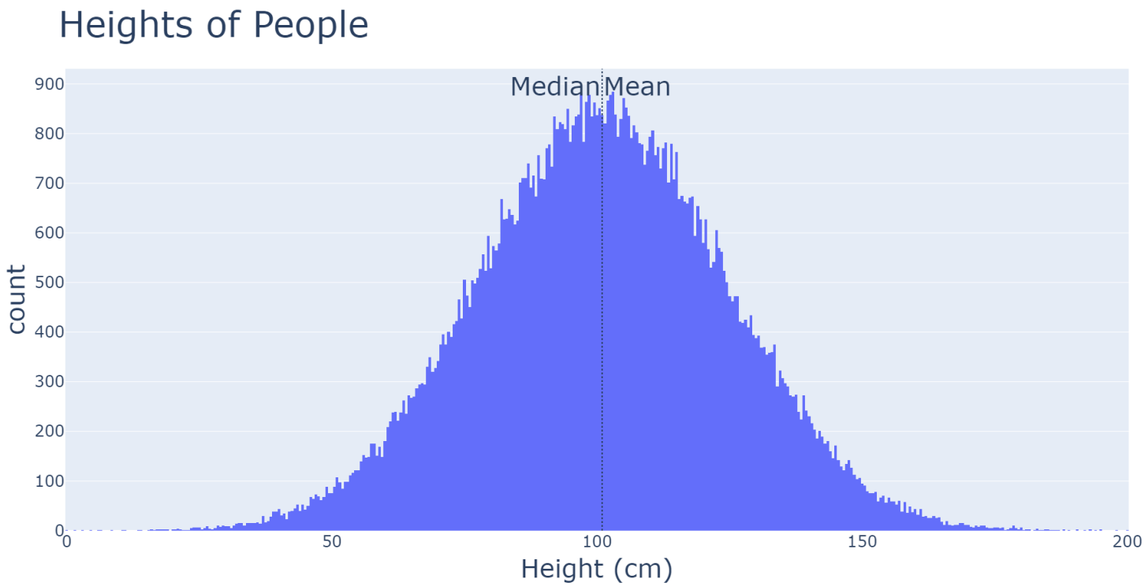
CITY	Rainfall (inches)			
SEATTLE	4.67	1.75	2.09	4.53
	0.42	2.60	1.90	3.14
	1.35	3.40	3.75	1.75
NYC	2.68	3.93	0.07	3.14
	3.96	3.95	0.52	3.60
	4.72	4.72	2.27	2.68
PARIS	2.33	2.07	1.06	1.38
	2.07	4.29	4.29	1.47

CITY	Rainfall (inches)			
	2.54	2.54	2.54	2.54
SEATTLE	4.67	1.75	2.09	4.53
	0.42	2.60	1.90	3.14
	1.35	3.40	3.75	1.75
NYC	2.68	3.93	0.07	3.14
	3.96	3.95	0.52	3.60
	4.72	4.72	2.27	2.68
PARIS	2.33	2.07	1.06	1.38
	2.07	4.29	4.29	1.47

3. Describing the data

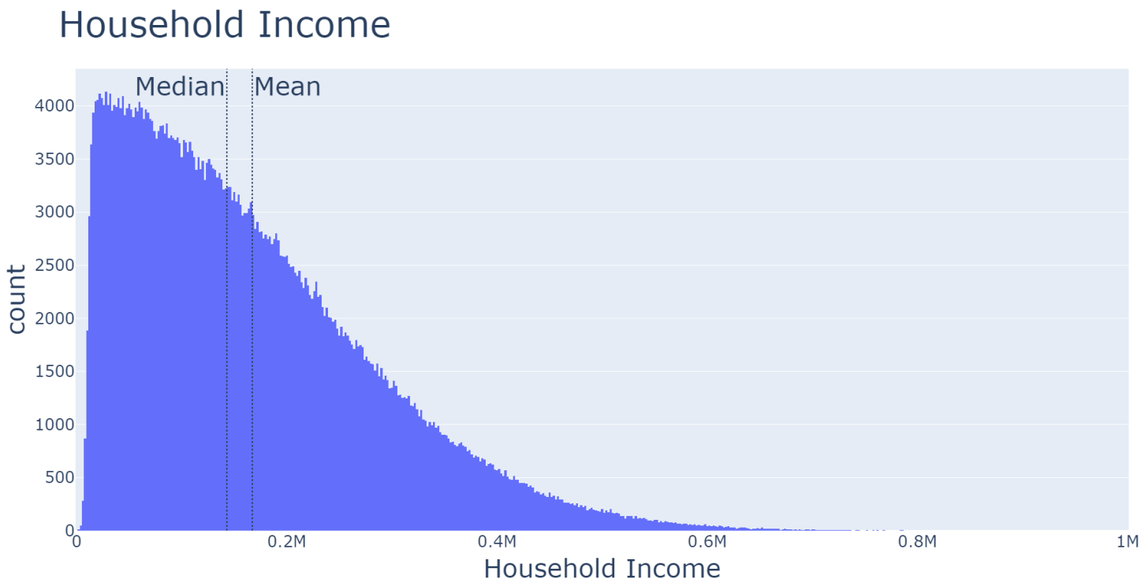
- Minimum
- Maximum
- Mean: sum of all values divided by the number of observations
- Median: the value in the center of a range of values
- Standard Deviation: average amount of difference from the mean of a variable observed across all data points

3. Describe the data with distributions.

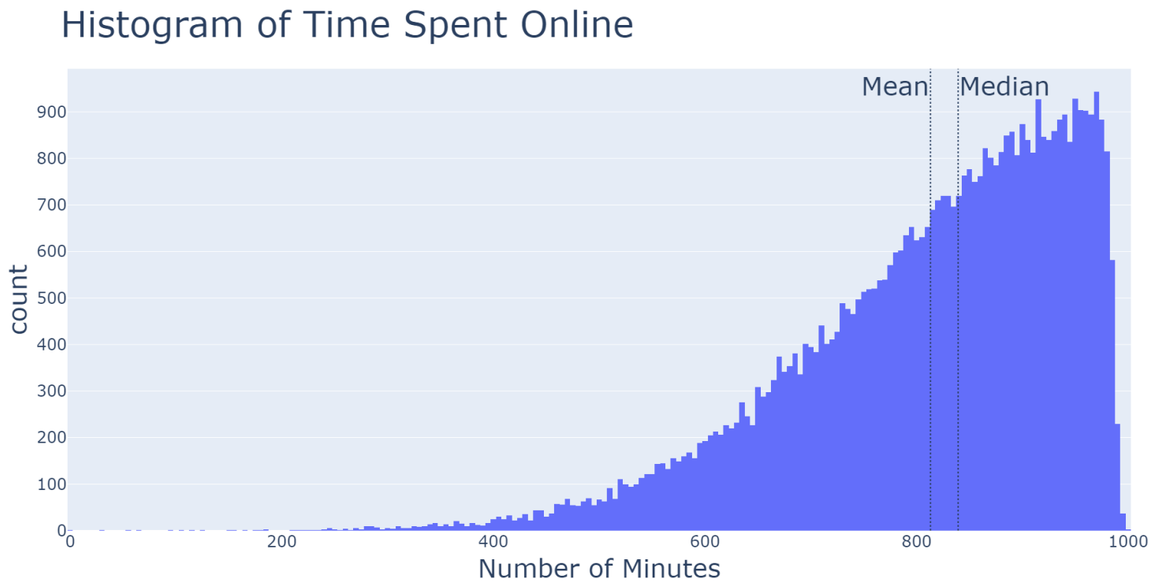


- Median and the mean are the same value
- A symmetrical curve

3. Describing the data with distributions



- Median < Mean
- "Right-skewed": the tail is to the right



- Median > Mean
- "Left-skewed": the tail is to the left

The dataset: AirBnB listings

listing_id	host_id	host_since	city	price
41633222	328263918	1/16/2020	New York	27
45841679	367658324	9/15/2020	New York	98
32805414	244370442	2/20/2019	New York	162
35265786	265506523	5/31/2019	New York	65
46055424	334163301	2/6/2020	New York	22
31654063	237336458	1/17/2019	New York	99
43293920	344737629	4/26/2020	New York	65
35233962	264950723	5/29/2019	New York	340
35512830	262257479	5/16/2019	New York	169
43022394	342139982	3/20/2020	New York	79
47826745	383332265	1/6/2021	New York	99
42986899	358273459	7/25/2020	New York	119

Let's practice!

EXPLORATORY DATA ANALYSIS IN POWER BI

Initial EDA of AirBnB listings

EXPLORATORY DATA ANALYSIS IN POWER BI

Jacob H. Marquez
Data Scientist at Microsoft

datacamp

Let's practice!

EXPLORATORY DATA ANALYSIS IN POWER BI