Exploratory Data Analysis in Power BI

Power BI

Table of Content

1. Initial Exploratory Data Analysis in Power BI

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   Categorical

Numerical variables often able to take an infinite set of values

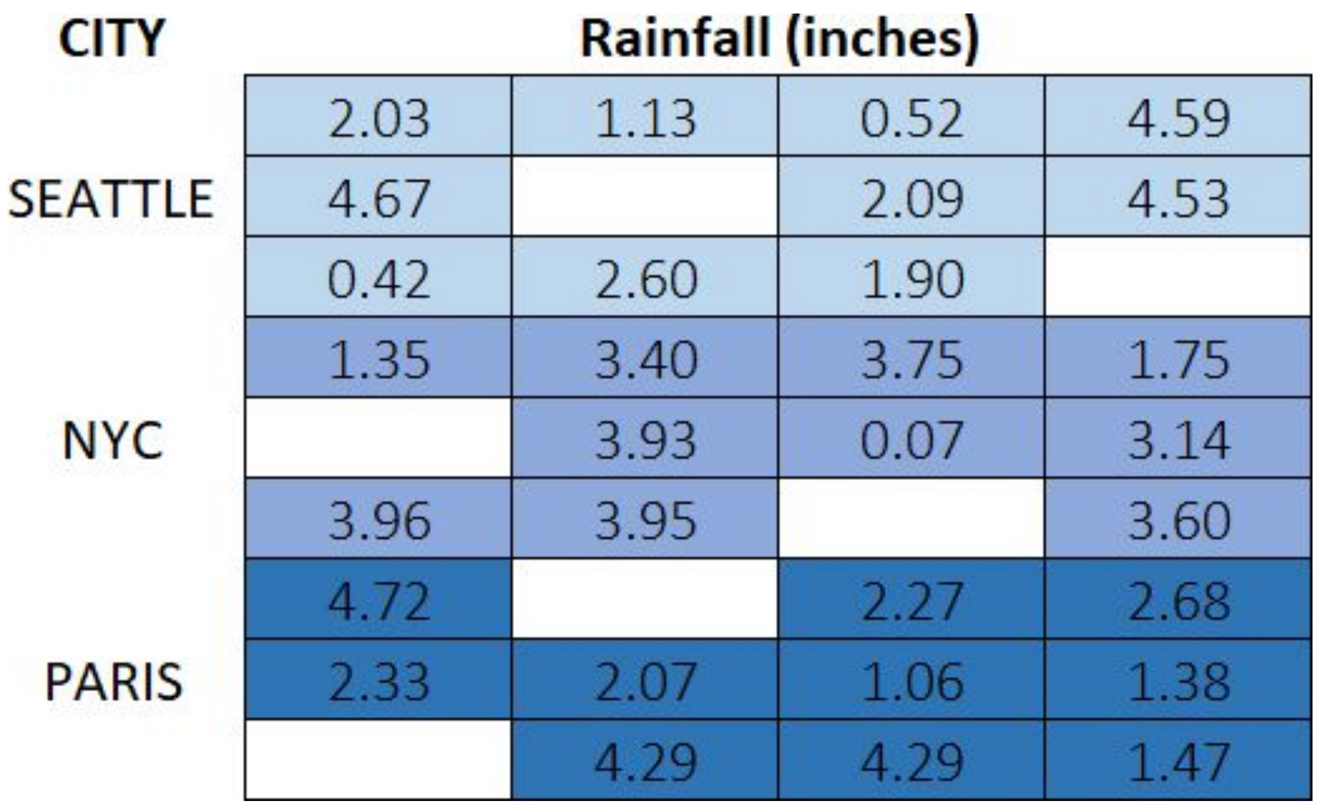
- Number of stars in space
- Click-through rates Distance
- between two cities

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

- House types
- Country
- Company

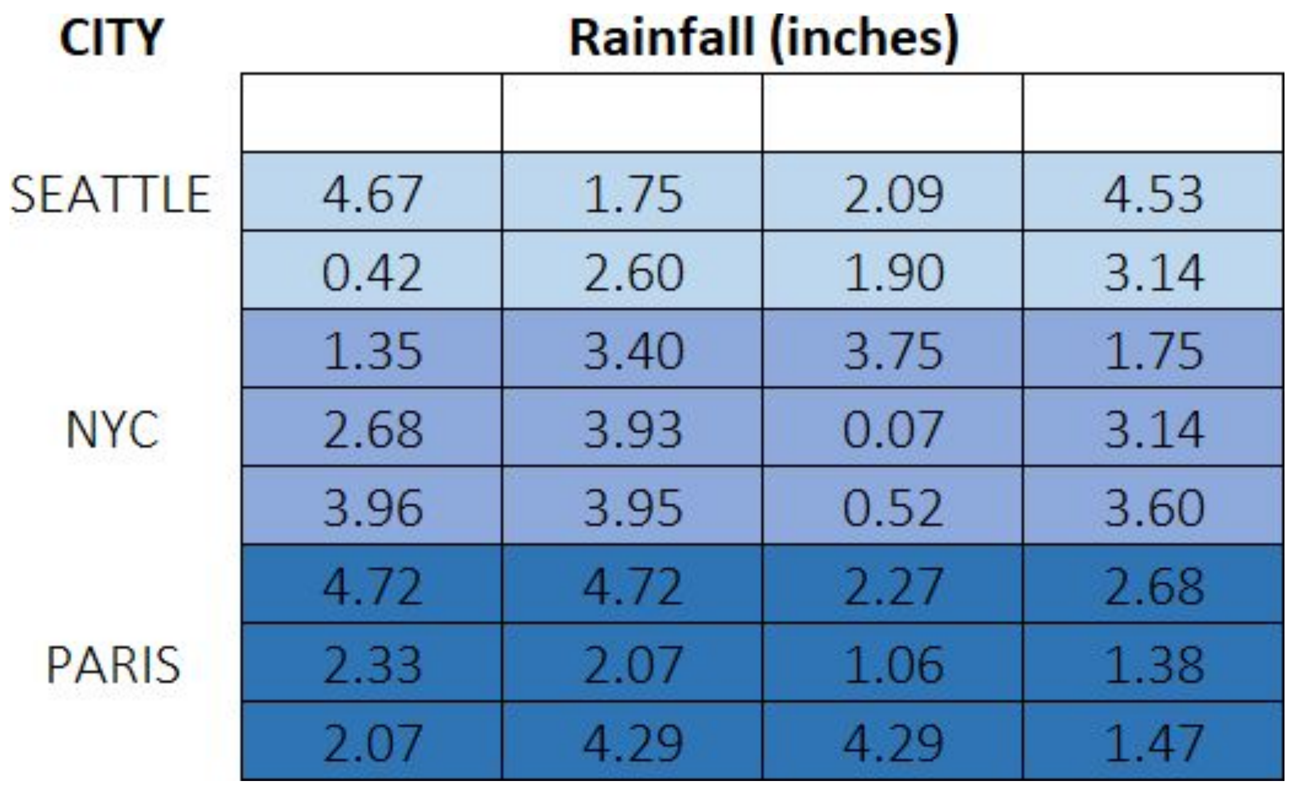
2. Identifying missing data

Missing at random



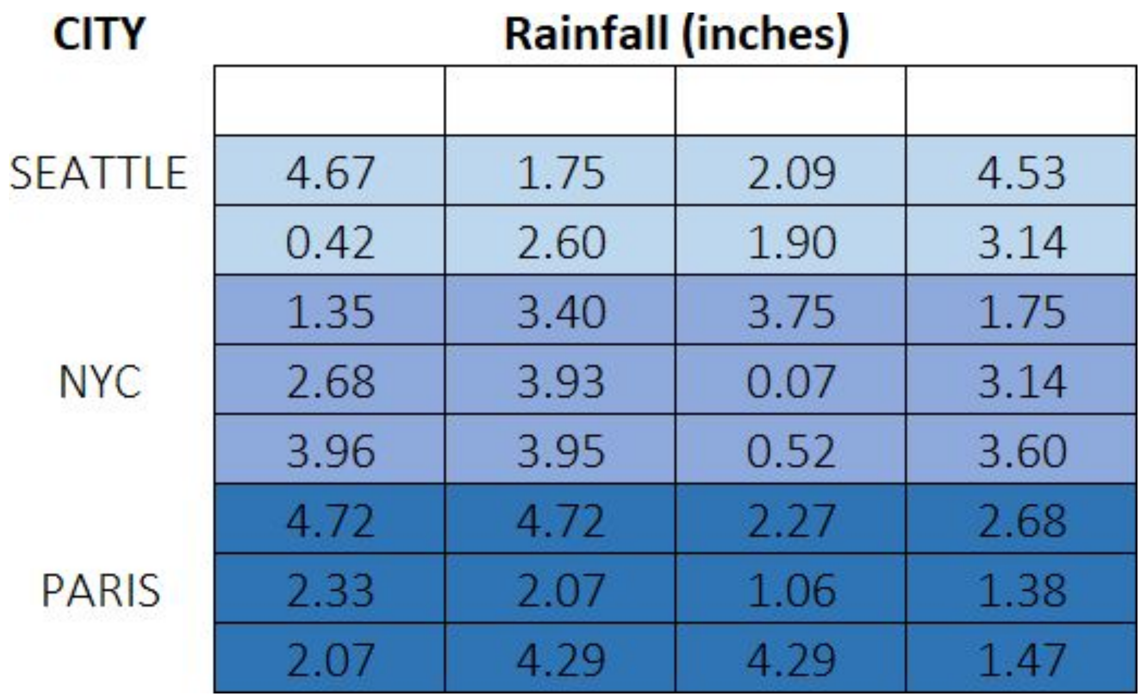
CITY	Rainfall (inches)			
SEATTLE	2.03	1.13	0.52	4.59
	4.67		2.09	4.53
	0.42	2.60	1.90	
NYC	1.35	3.40	3.75	1.75
		3.93	0.07	3.14
	3.96	3.95		3.60
PARIS	4.72		2.27	2.68
	2.33	2.07	1.06	1.38
		4.29	4.29	1.47

Missing not at random

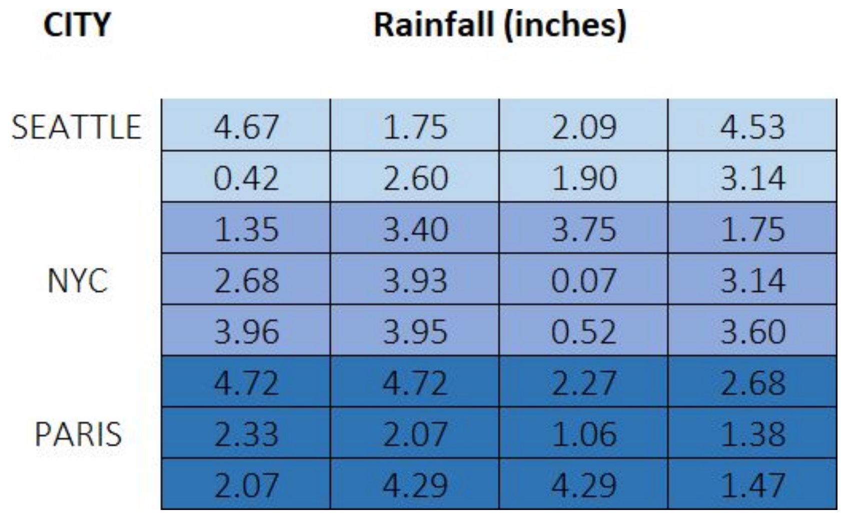


CITY	Rainfall (inches)			
SEATTLE	4.67	1.75	2.09	4.53
	0.42	2.60	1.90	3.14
NYC	1.35	3.40	3.75	1.75
	2.68	3.93	0.07	3.14
	3.96	3.95	0.52	3.60
PARIS	4.72	4.72	2.27	2.68
	2.33	2.07	1.06	1.38
	2.07	4.29	4.29	1.47

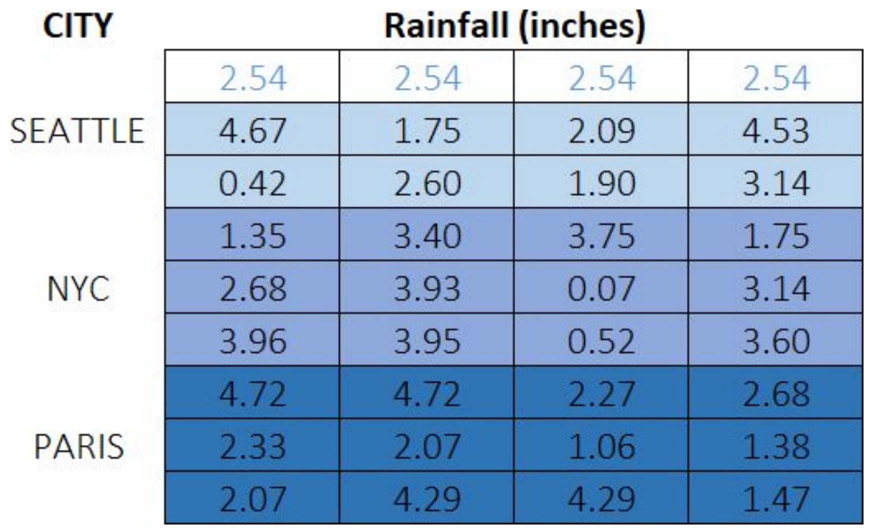
2. Addressing missing data



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

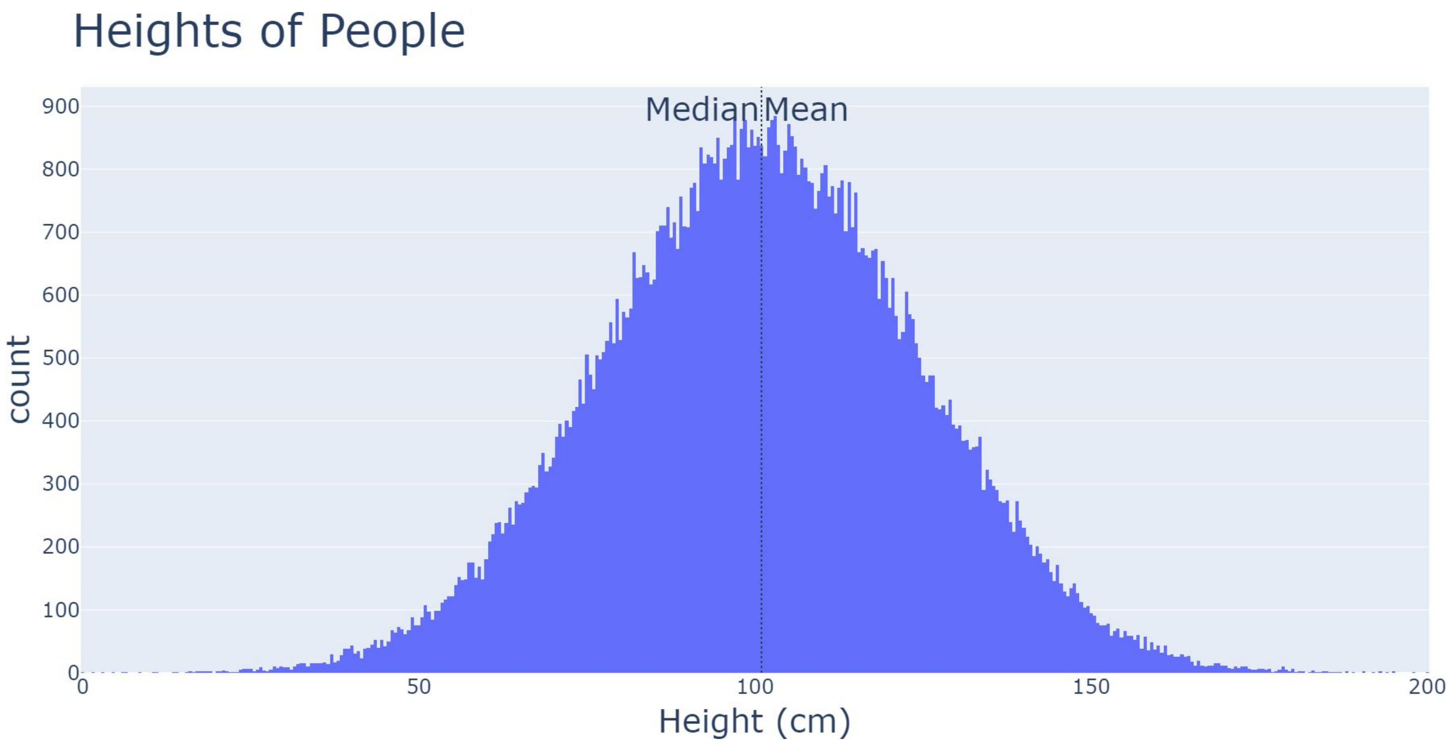


CITY	Rainfall (inches)			
	2.54	2.54	2.54	2.54
SEATTLE	4.67	1.75	2.09	4.53
	0.42	2.60	1.90	3.14
NYC	1.35	3.40	3.75	1.75
	2.68	3.93	0.07	3.14
	3.96	3.95	0.52	3.60
PARIS	4.72	4.72	2.27	2.68
	2.33	2.07	1.06	1.38
	2.07	4.29	4.29	1.47

3. Describing the data

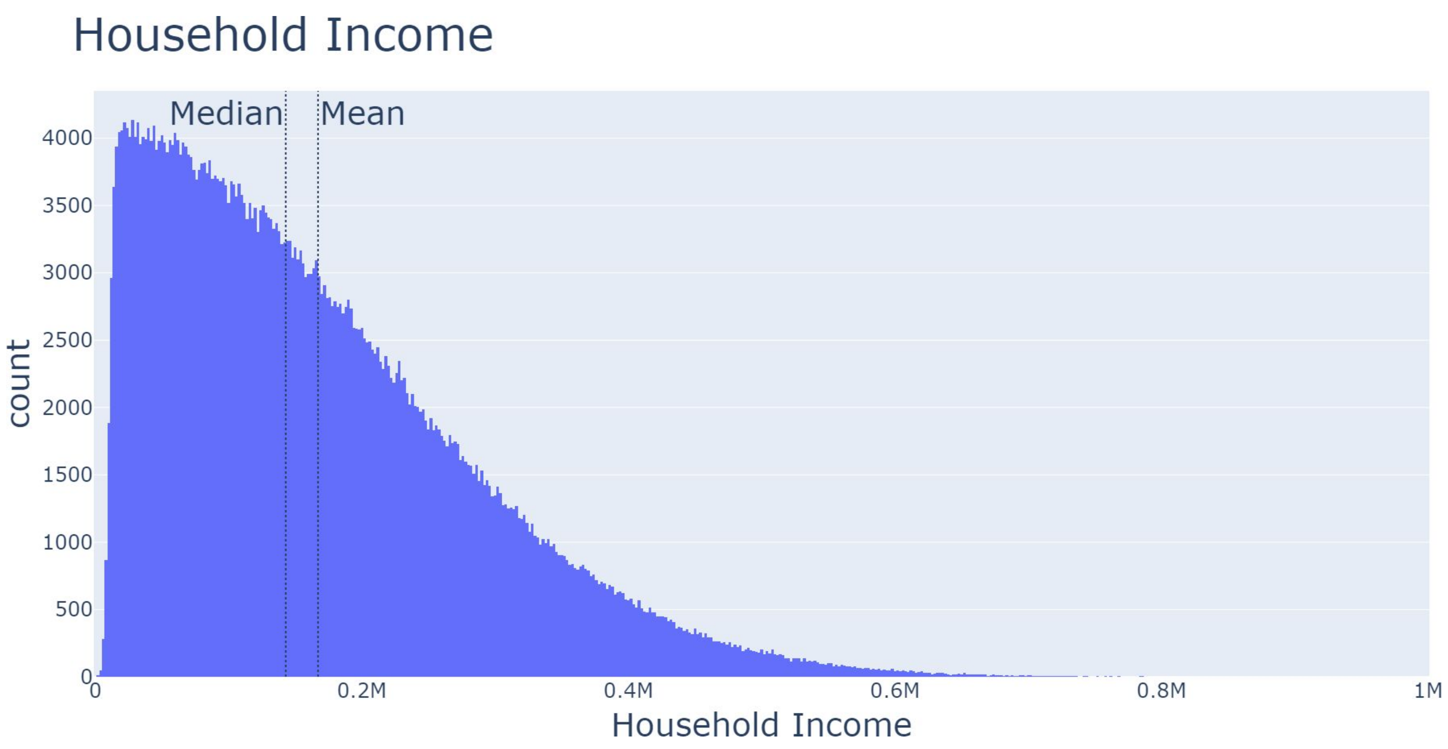
- Minimum
- Maximum
- Mean: sum of all values divided by the number of observations
- Median: the value in the center of a range of values
- Standard Deviation: average amount of difference from the mean of a variable observed across all data points

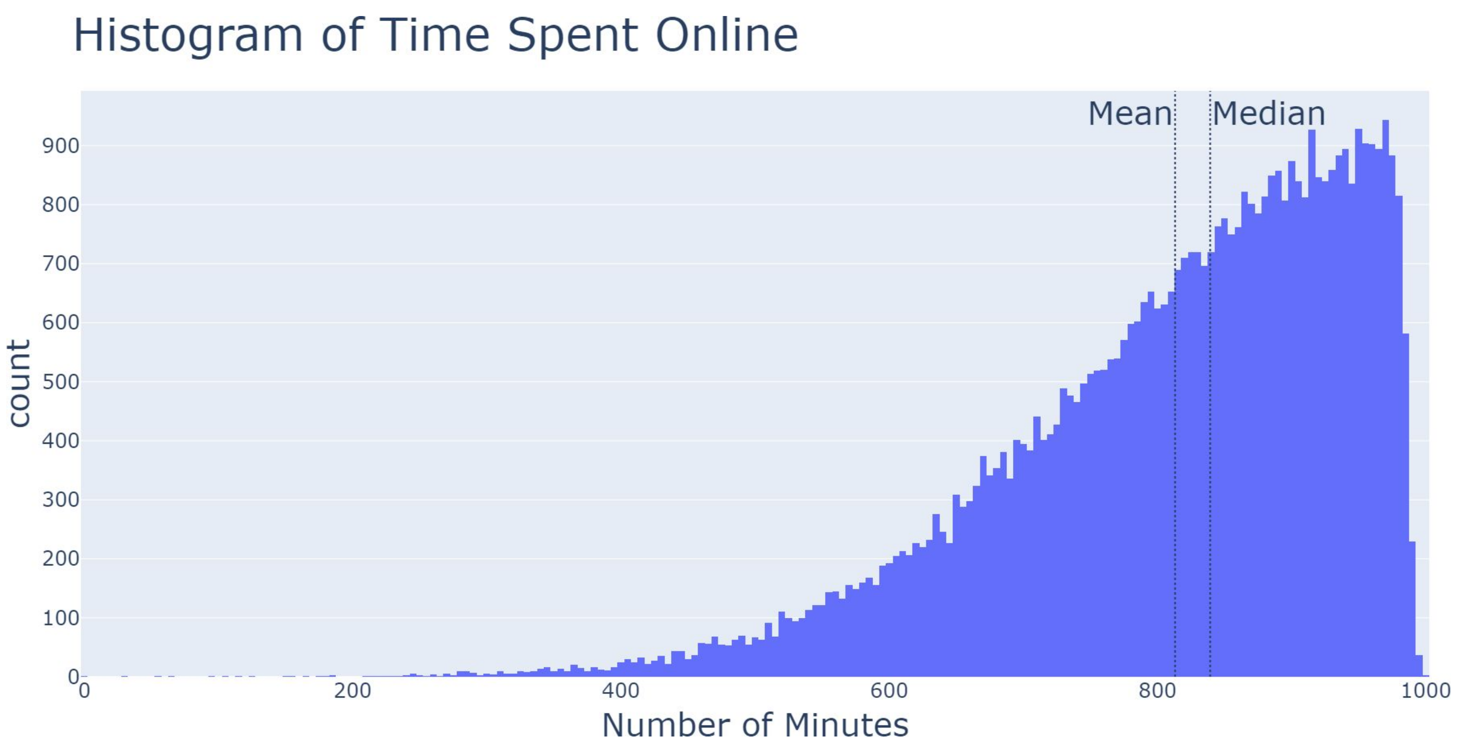
3. Describe the data with distributions.



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

3. Describing the data with distributions

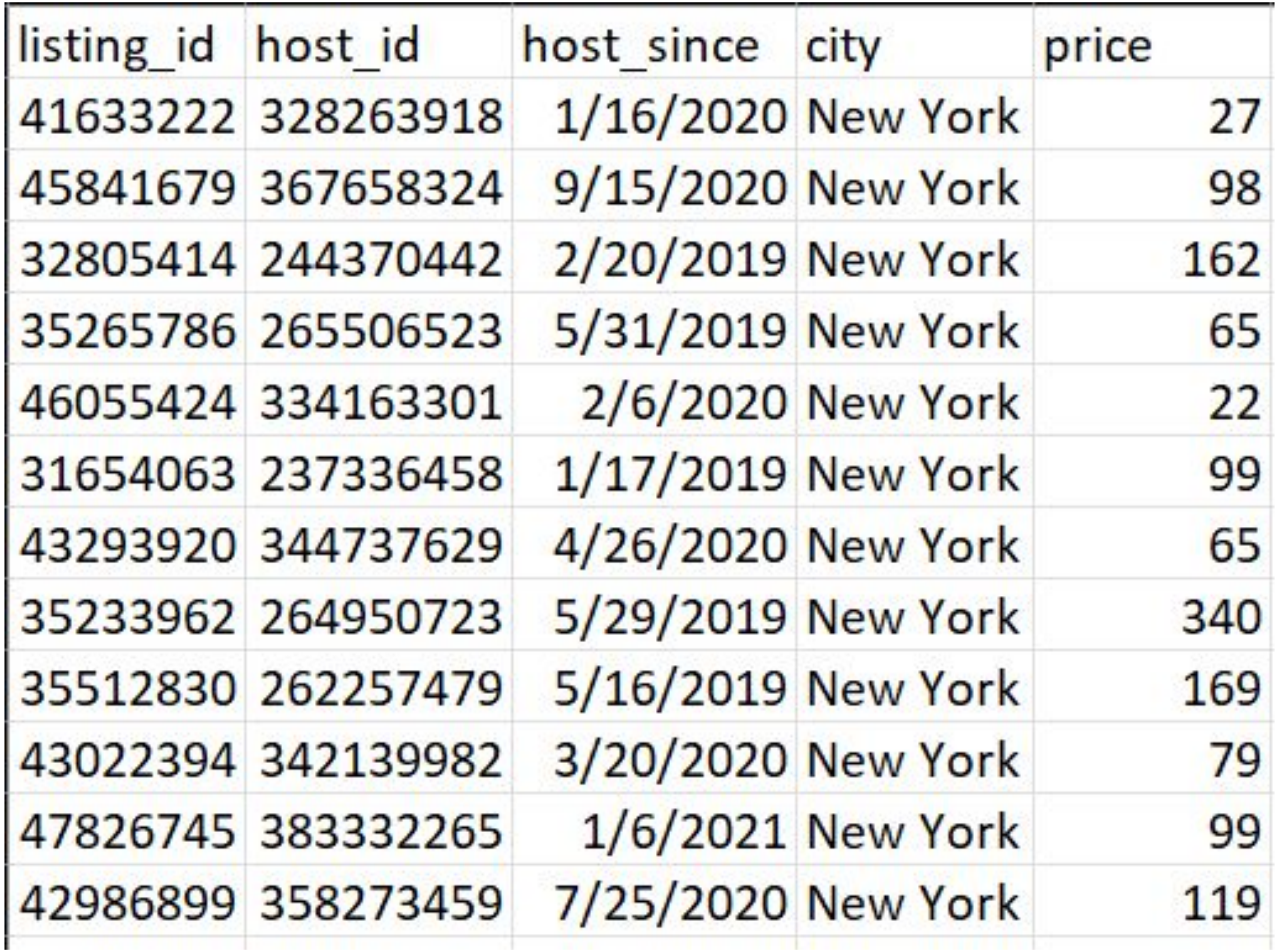




- Median < Mean
- "Right-skewed": the tail is to the right

- 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

Demo

2. Distributions and outliers

What are distributions?

Definition: *set of all possible values of the variable and the associated frequencies.*

What are distributions?

Continuous

Age	Frequency
18	7
19	11
20	13
21	19
22	12

What are distributions?

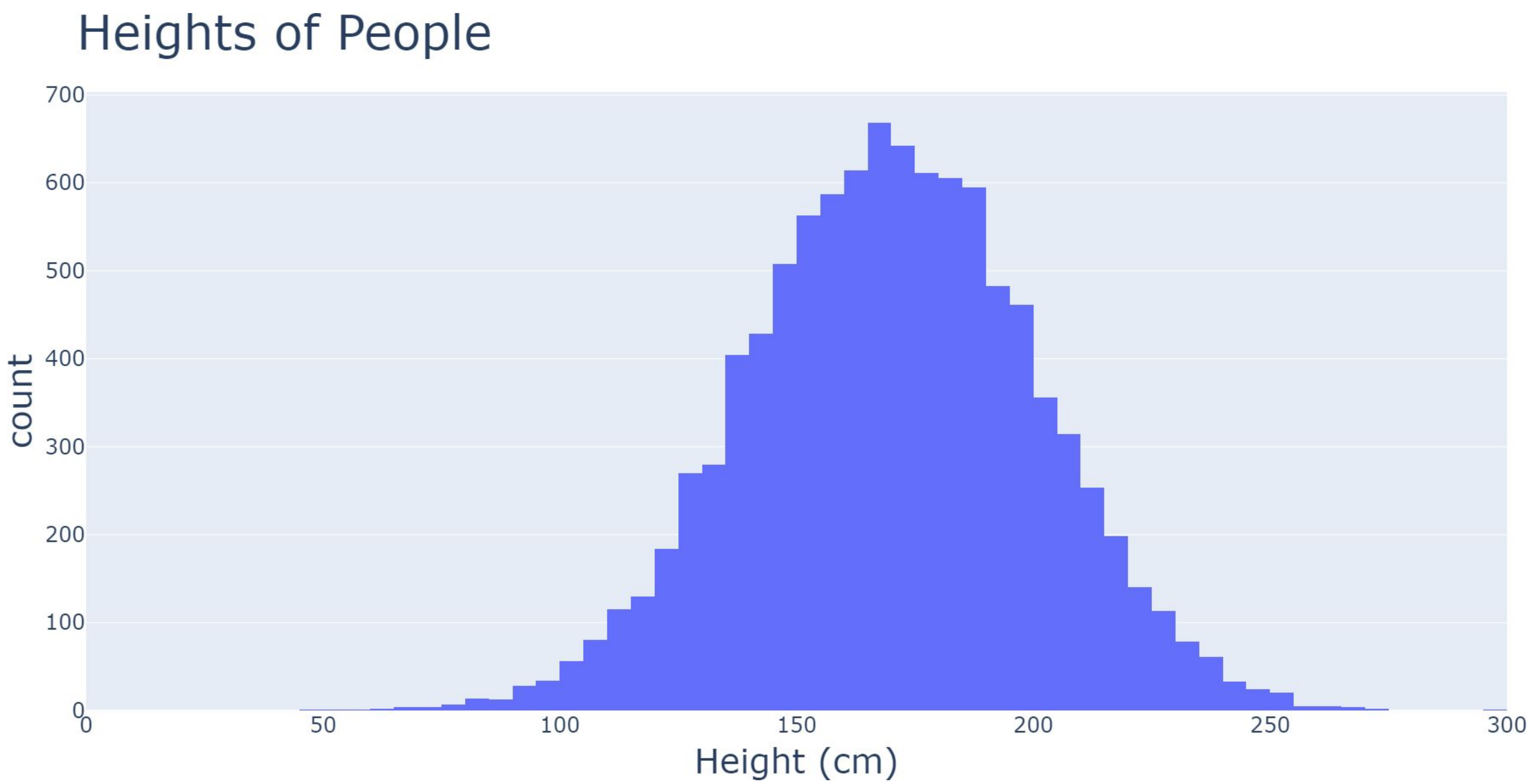
Continuous

Age	Frequency
18	7
19	11
20	13
21	19
22	12

Categorical

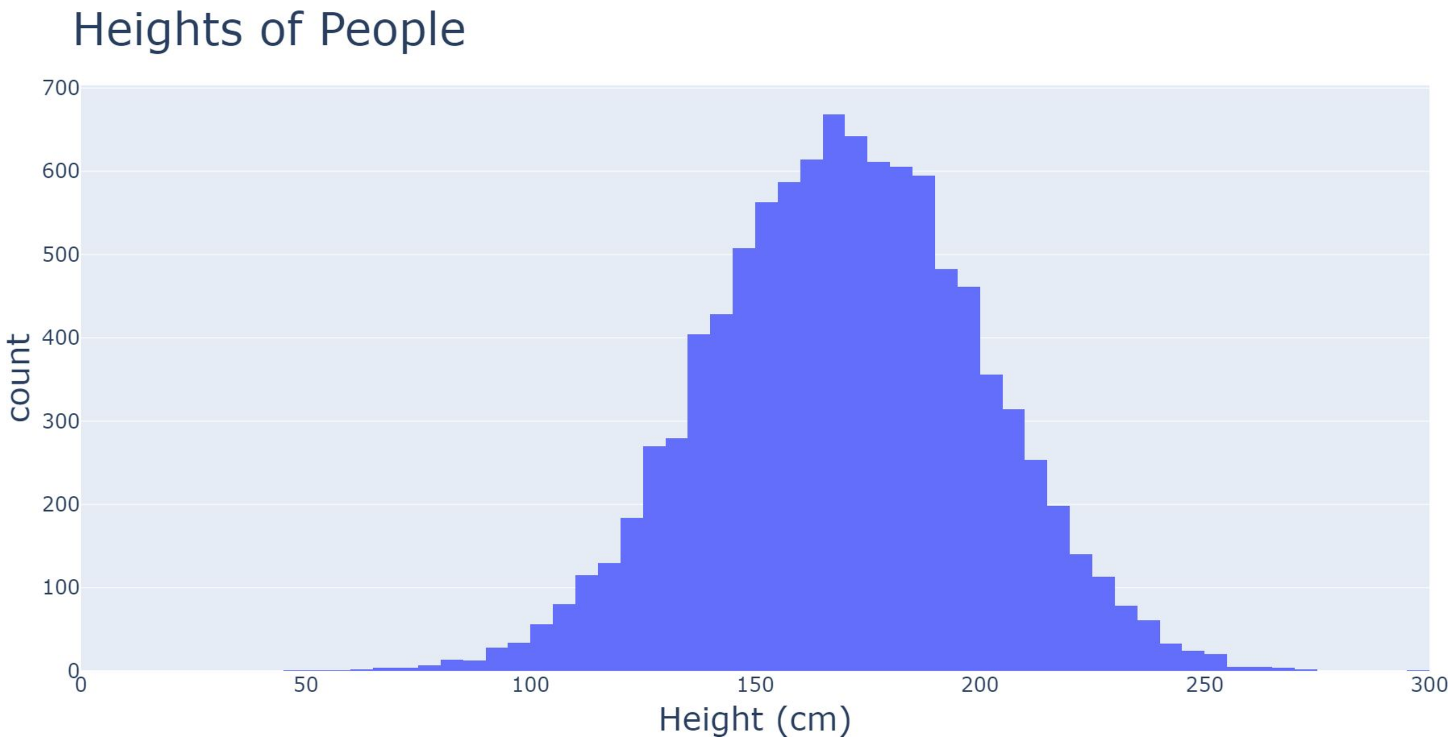
Hair Color	Frequency
Blonde	30
Brown	50
Black	40
Red	20
Grey	20

What are histograms?

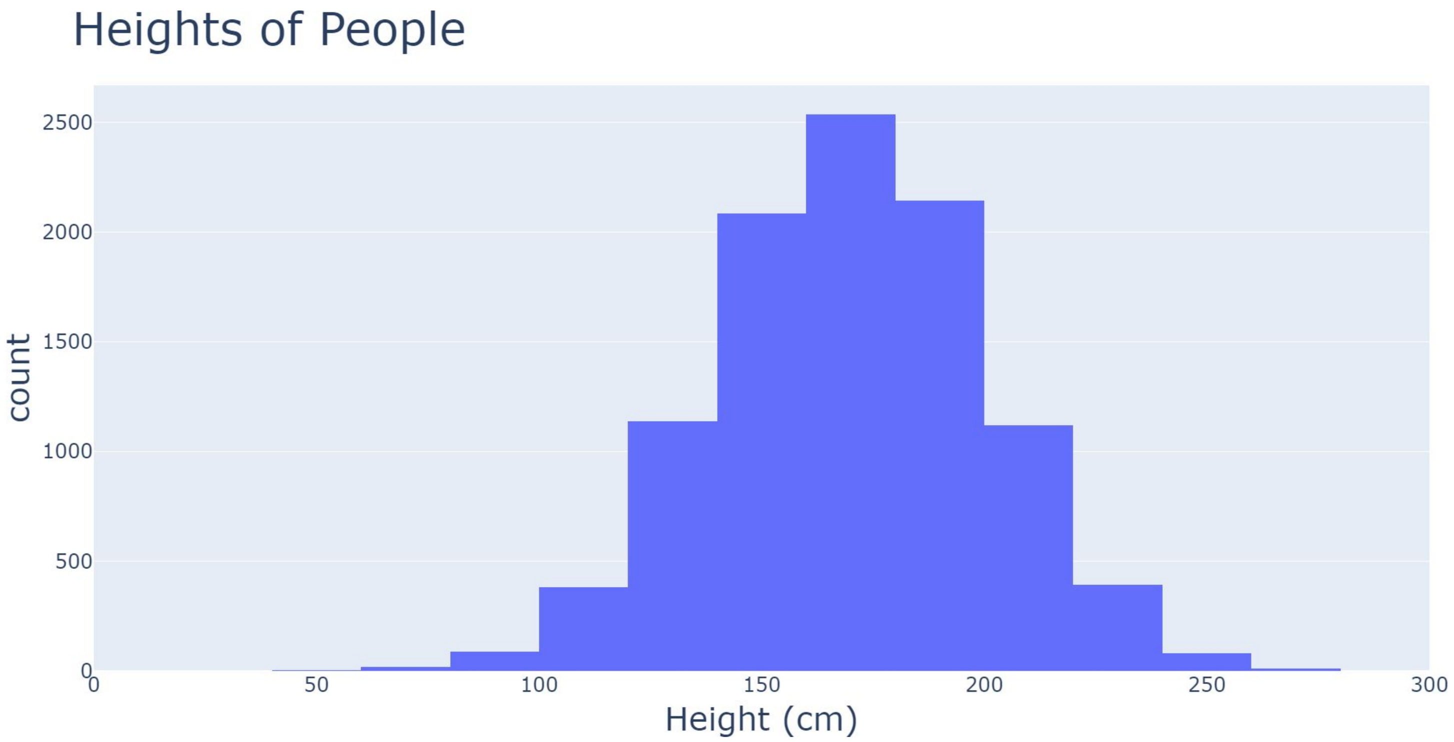


What are histogram? - bins

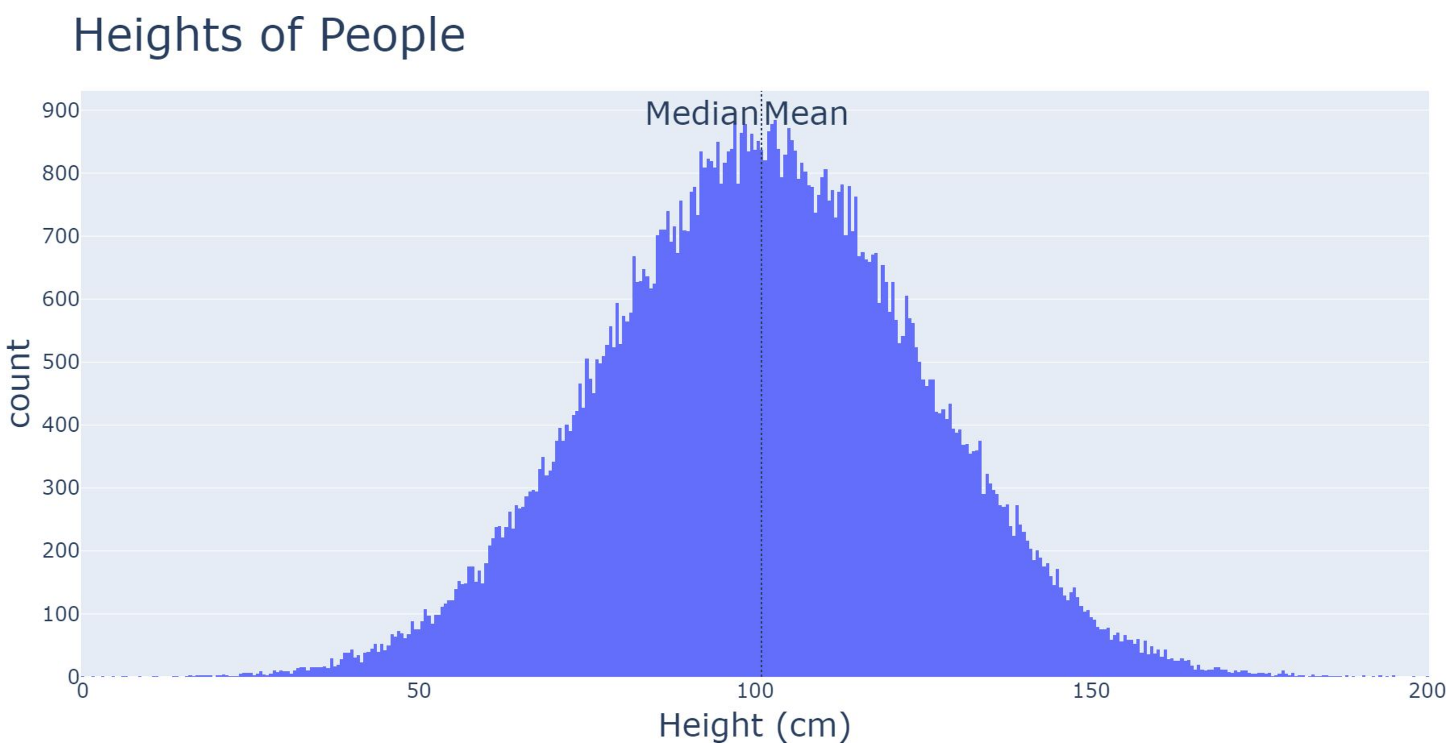
Histogram with 100 bins

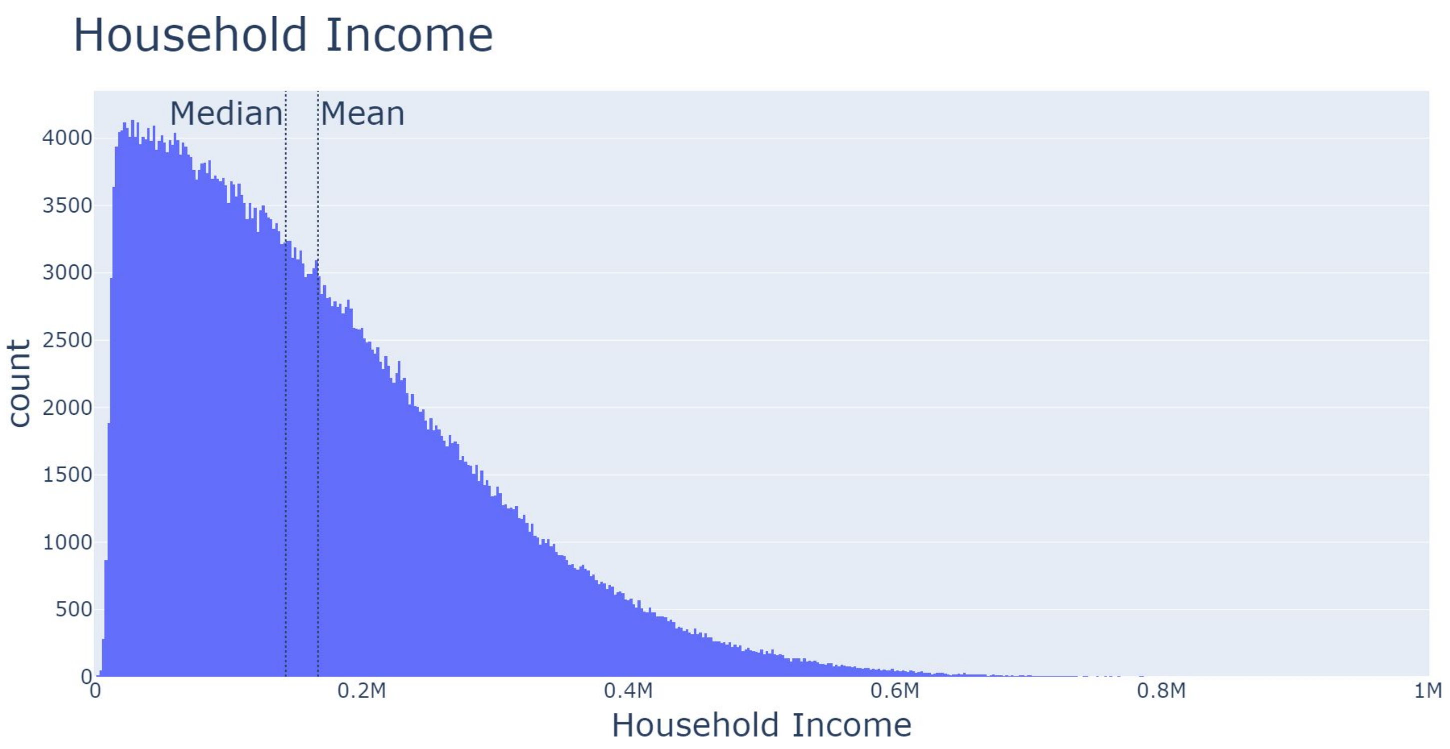


Histogram with 20 bins



Reading histograms - centrality and skewness



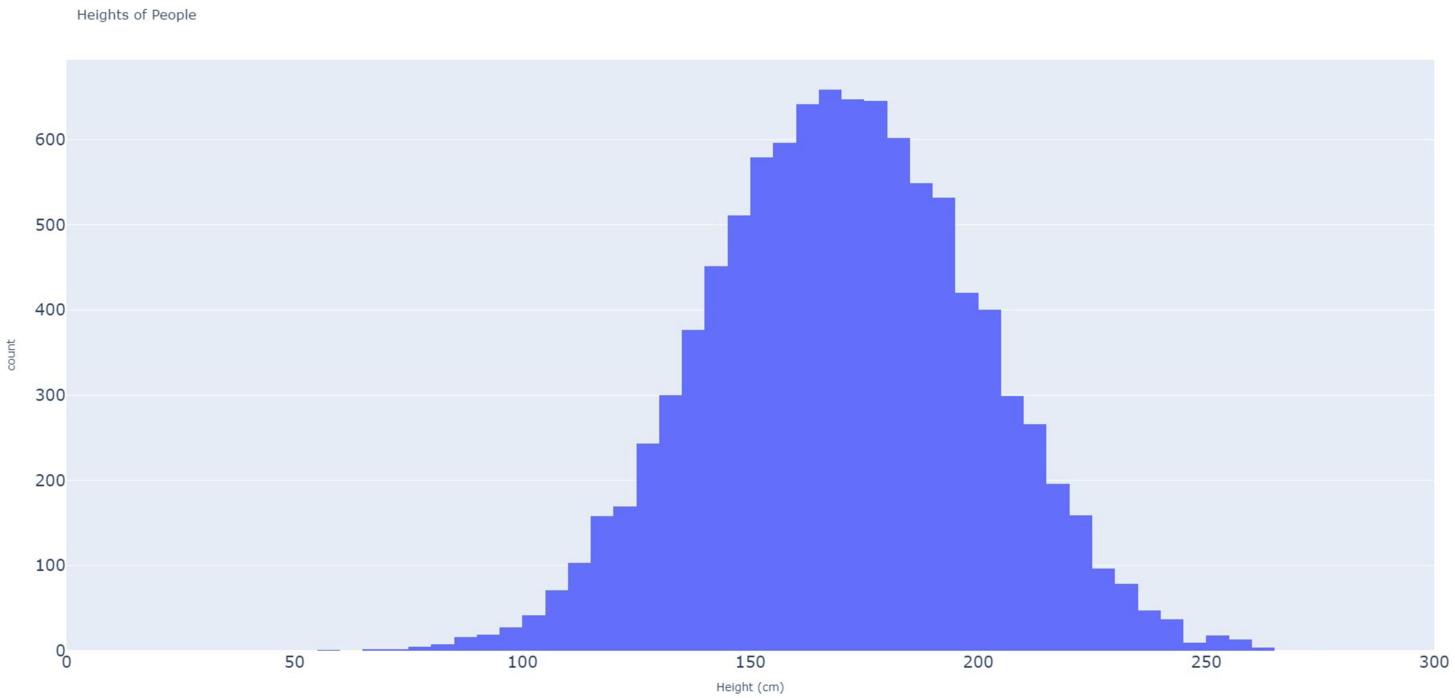


Normal distribution

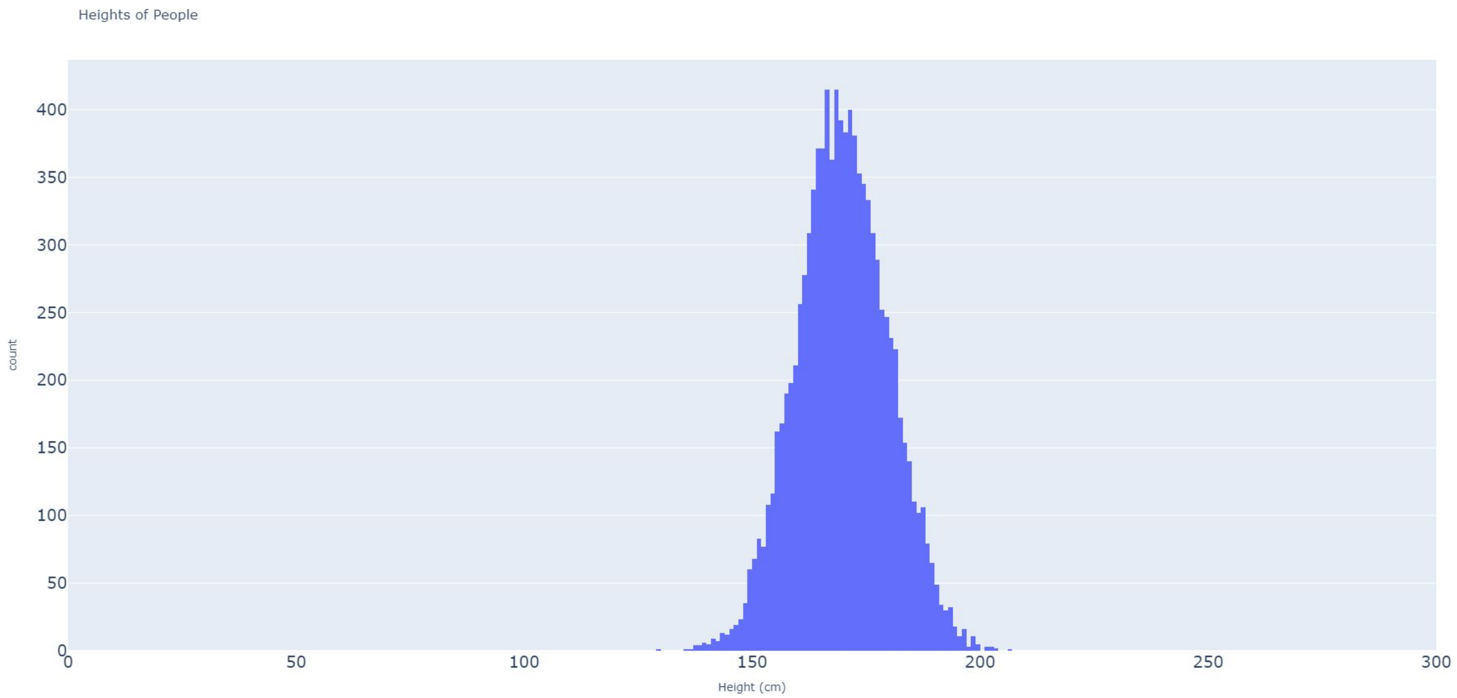
Right-skewed distribution

Reading histograms - spread

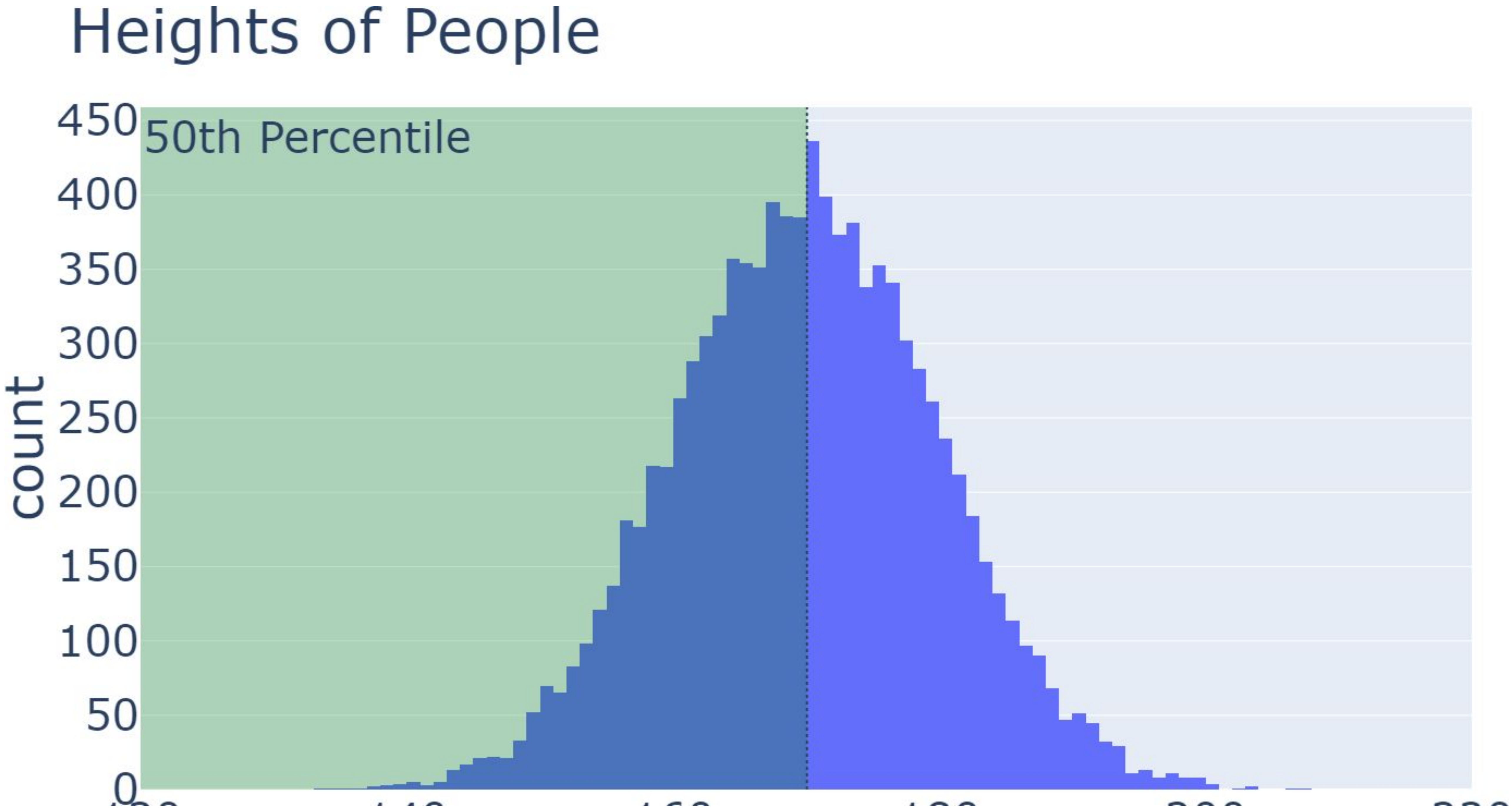
Larger standard deviation



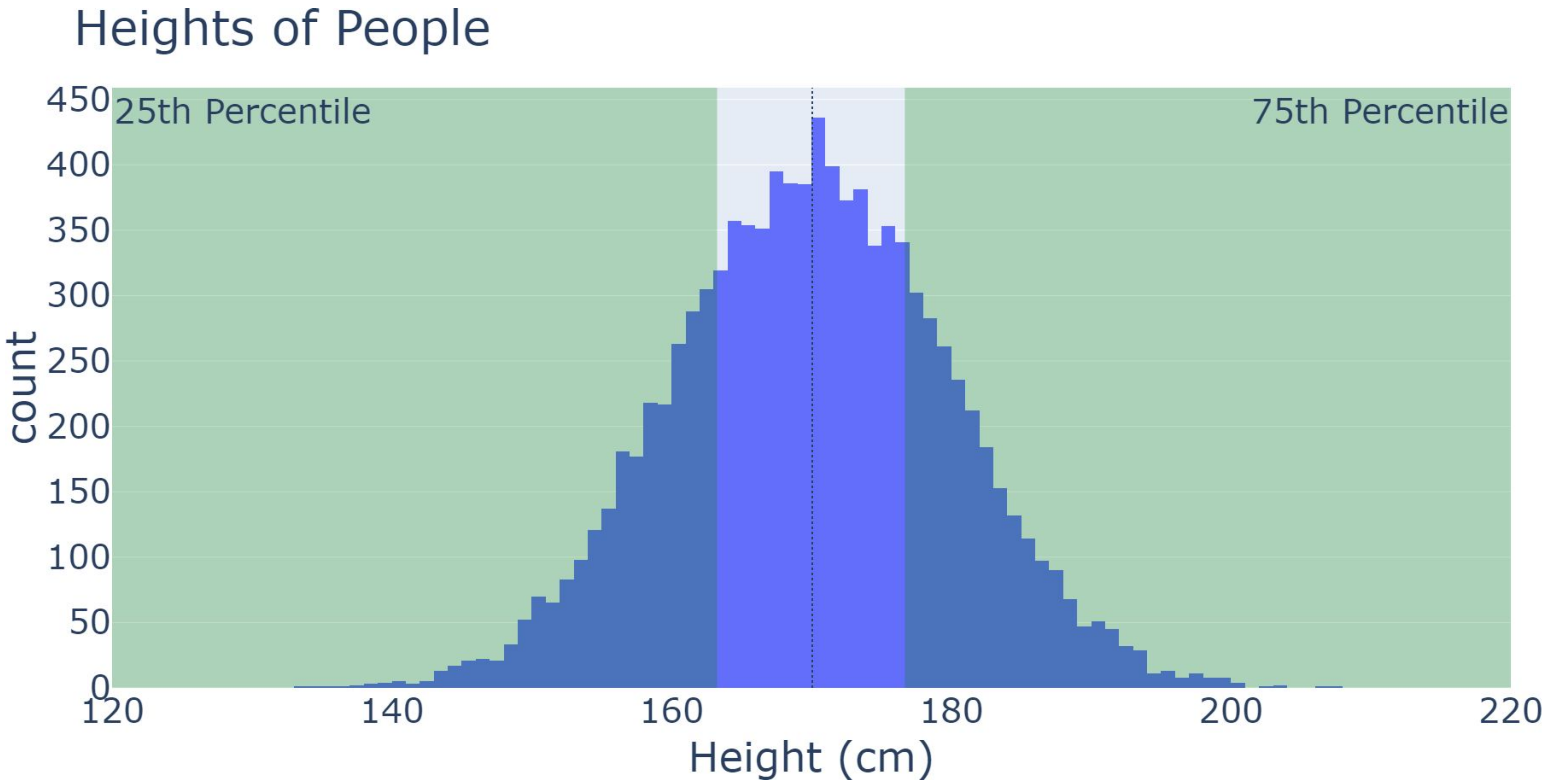
Smaller standard deviation



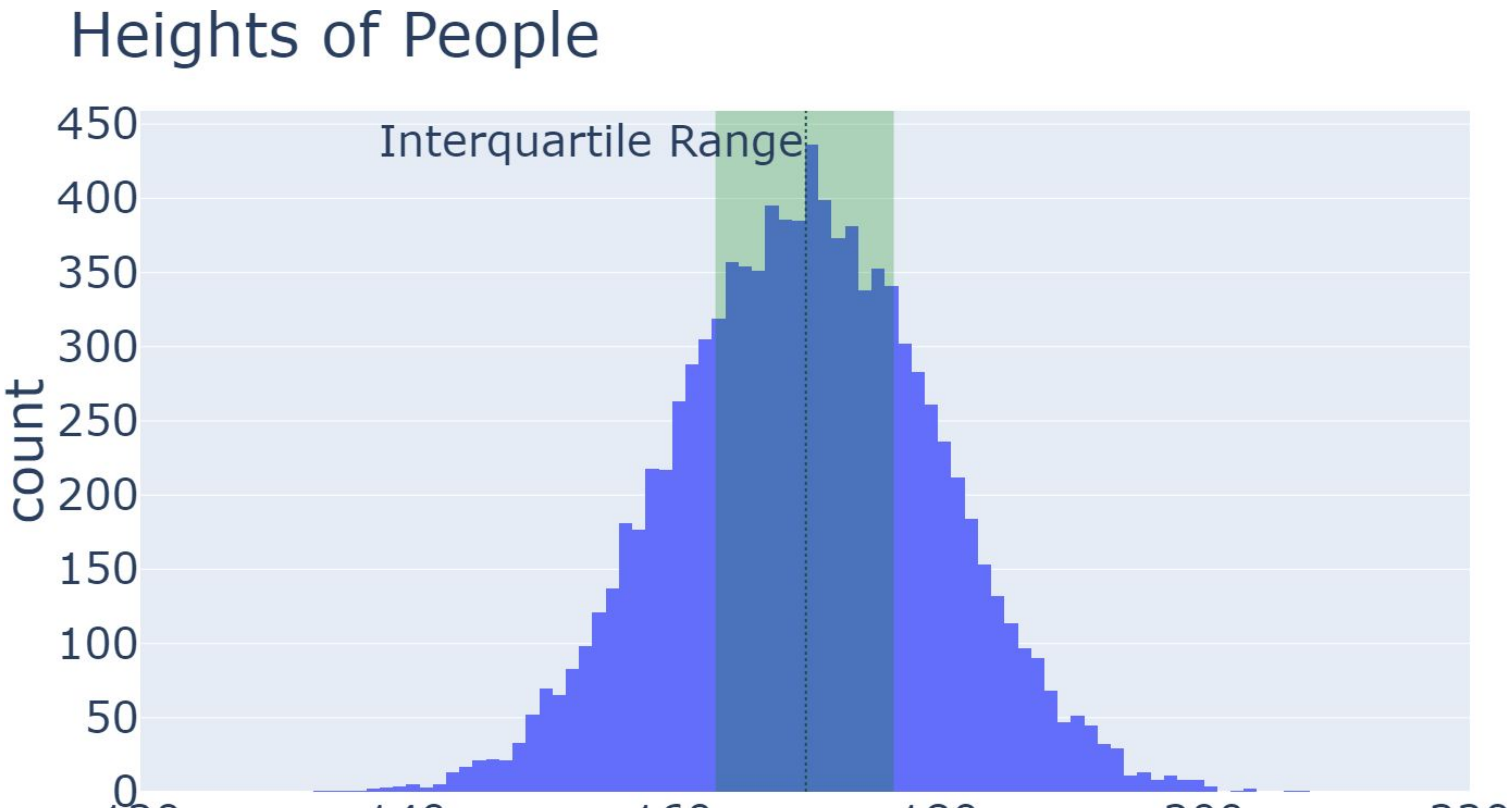
Reading histograms - percentiles



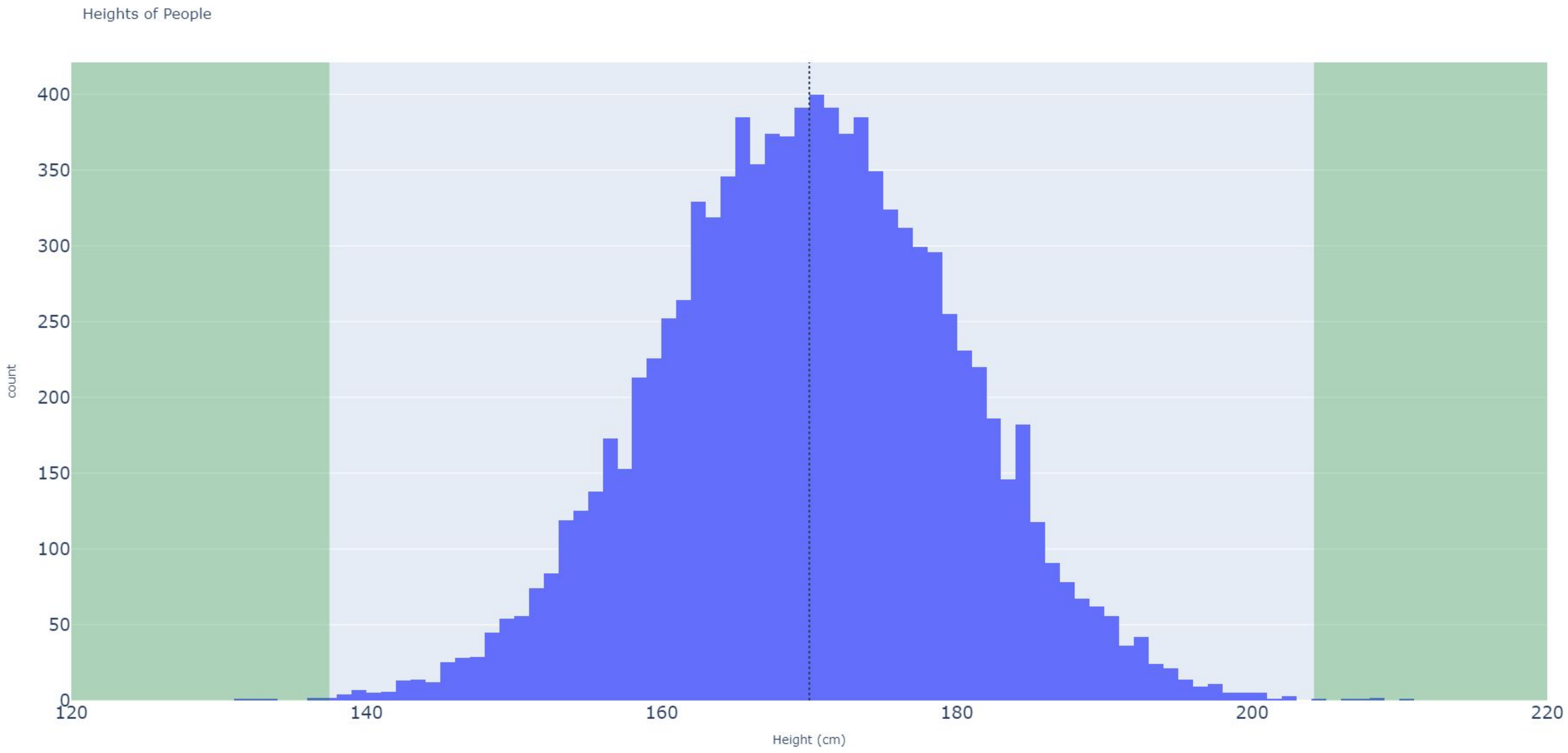
Reading histograms - 25th & 75th percentiles



Reading histograms - interquartile range



What is an outlier?



Finding outliers

Using standard deviation

$lower = -3 * SD$

$upper = 3 * SD$

Outlier when

$value < lower$ OR $upper < value$

Interquartile Range (IQR)

$lower = 25percentile - (1.5 * IQR)$ $upper = 75percentile + (1.5 * IQR)$

Outlier when

$value < lower$ OR $upper < value$

Addressing outliers

1. Remove observations
2. Imputation

Winsorizing

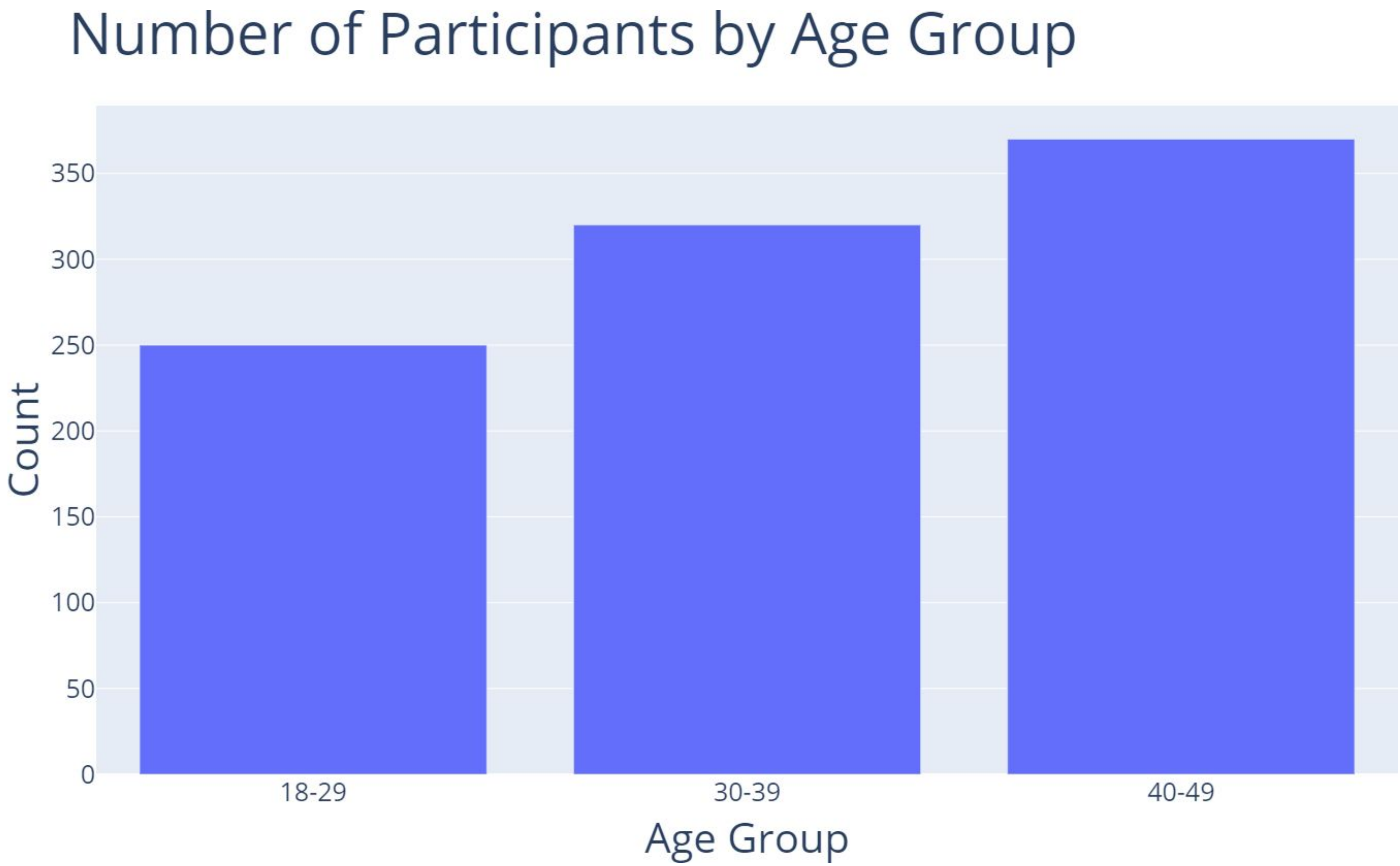
IF *value* < *5th percentile* **THEN** *value* = *5th percentile*

IF *95th percentile* > *value* **THEN** *value* = *95th percentile*

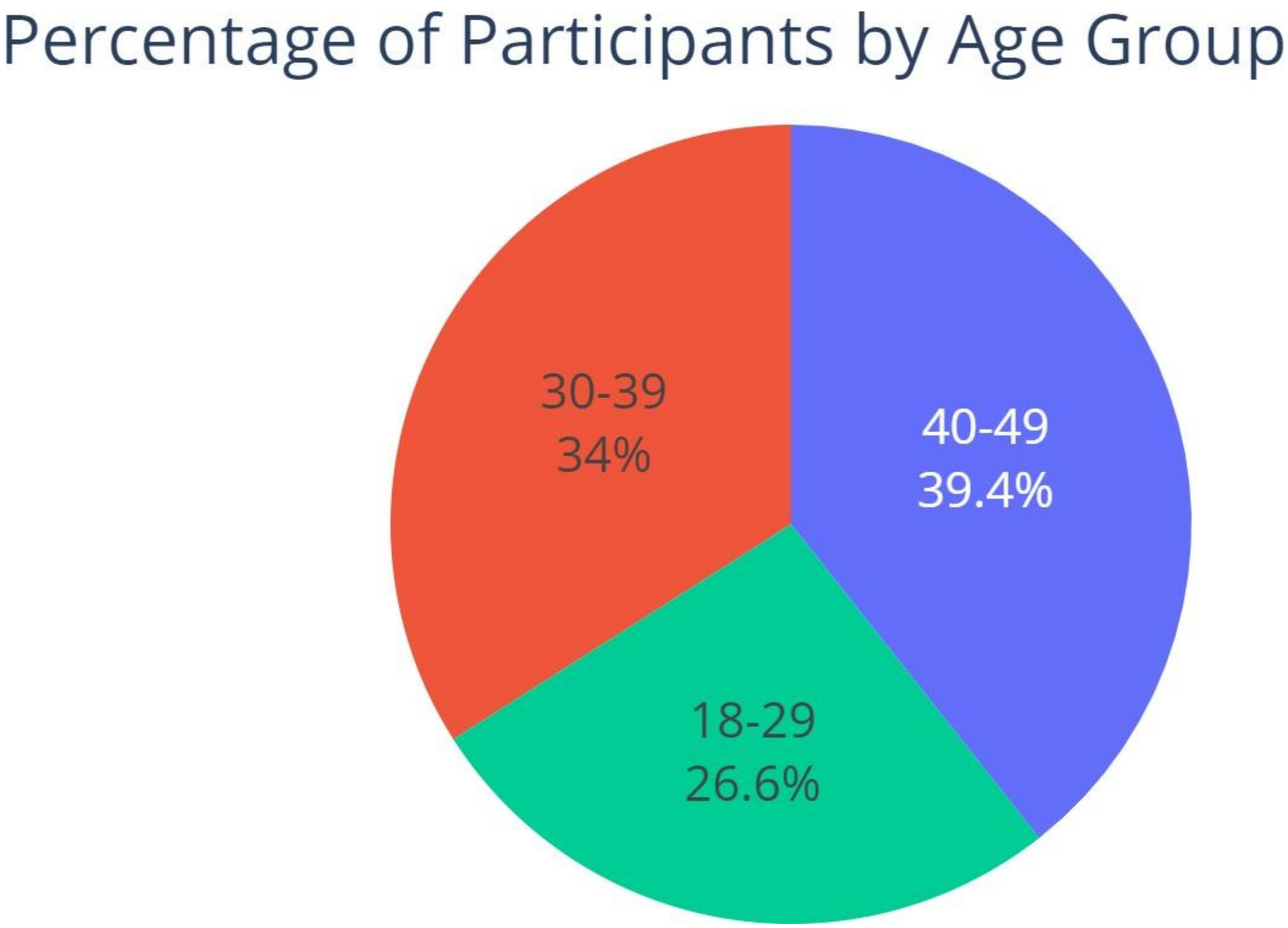
Demo

3. EDA with categorical variables

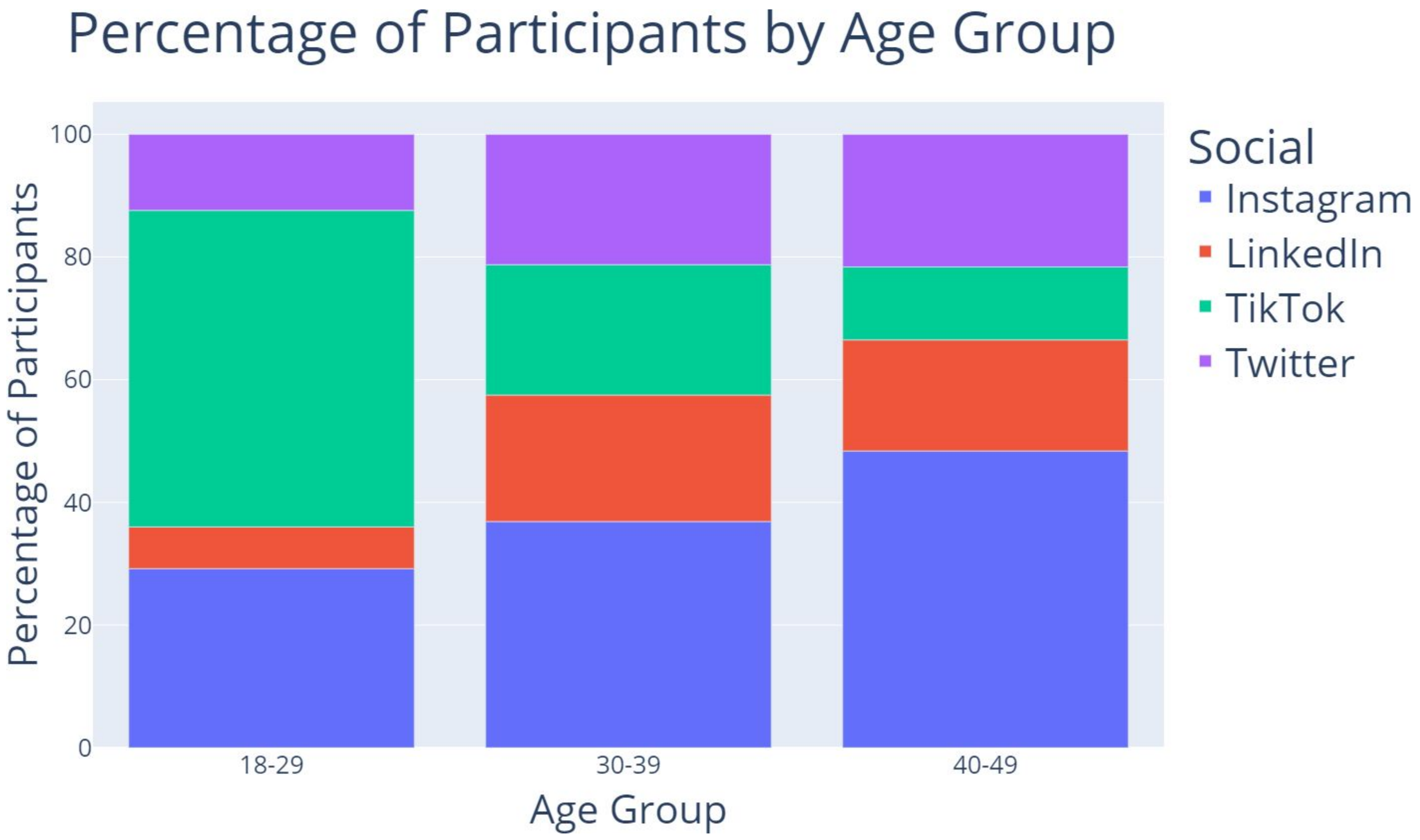
Categorical variables and frequency



Categorical variables and percentages



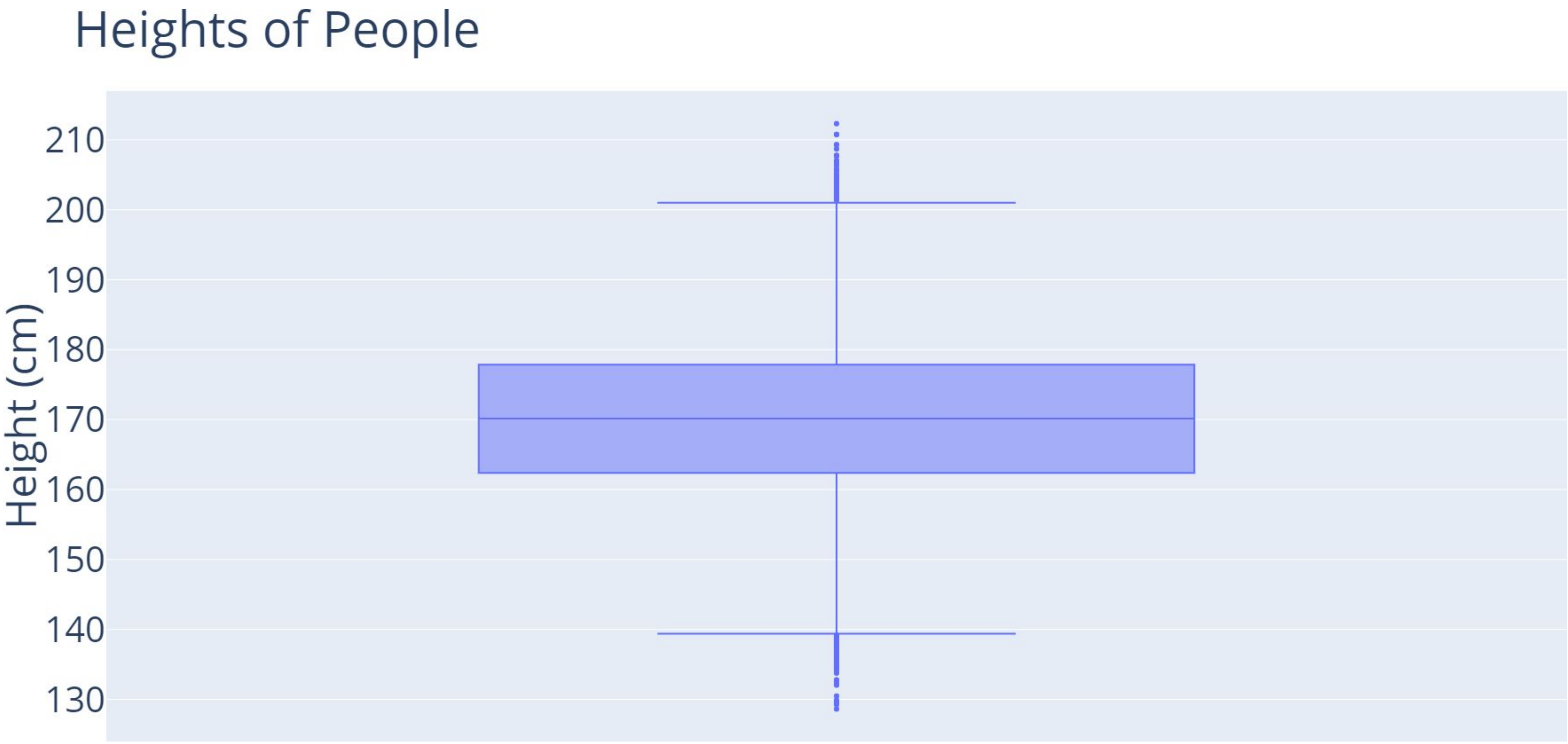
Proportions across multiple categorical variables



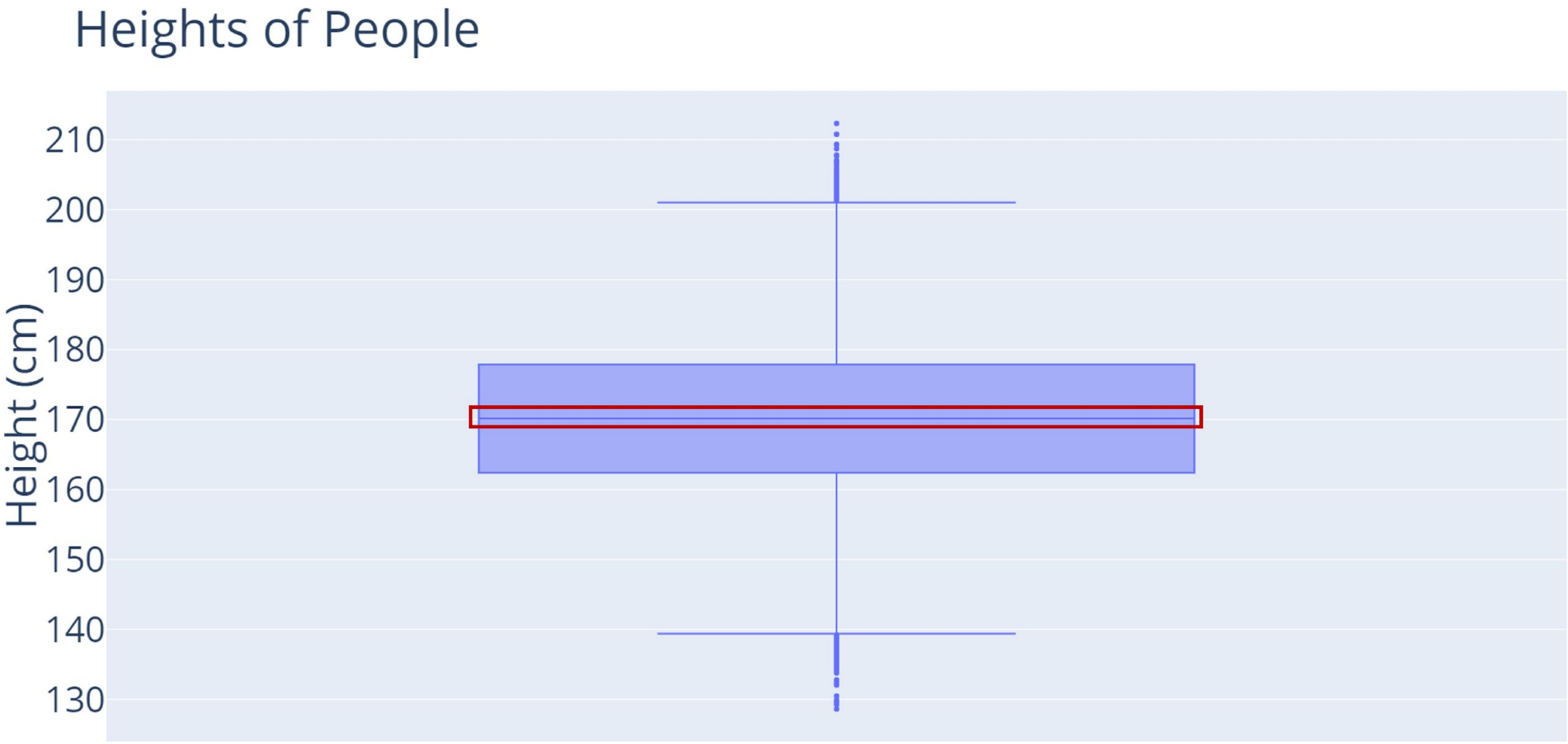
Categorical variables with descriptive statistics

Age Group	Median Hours per Day on Social Media
18-29	6
30-39	3
40-49	3

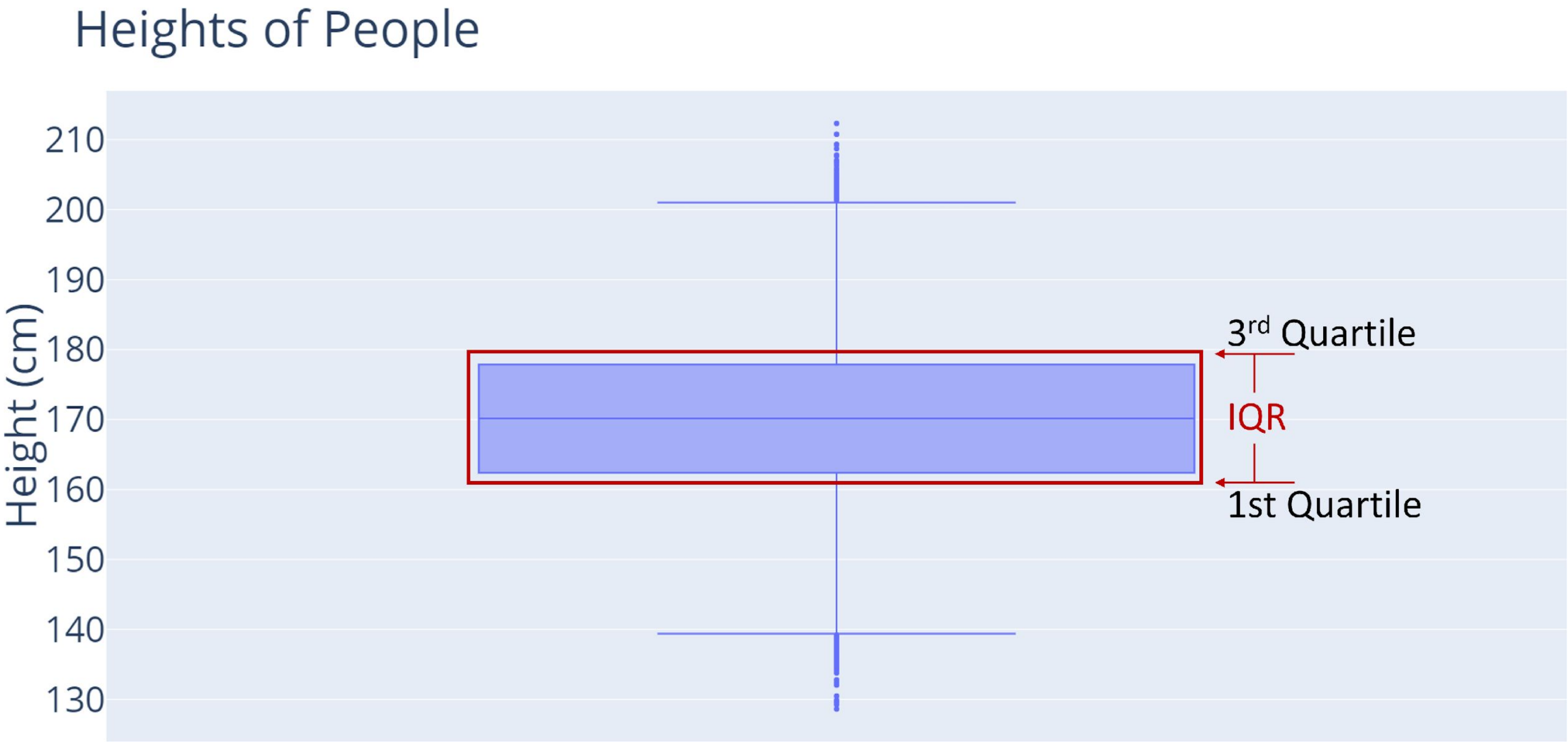
What are boxplots?



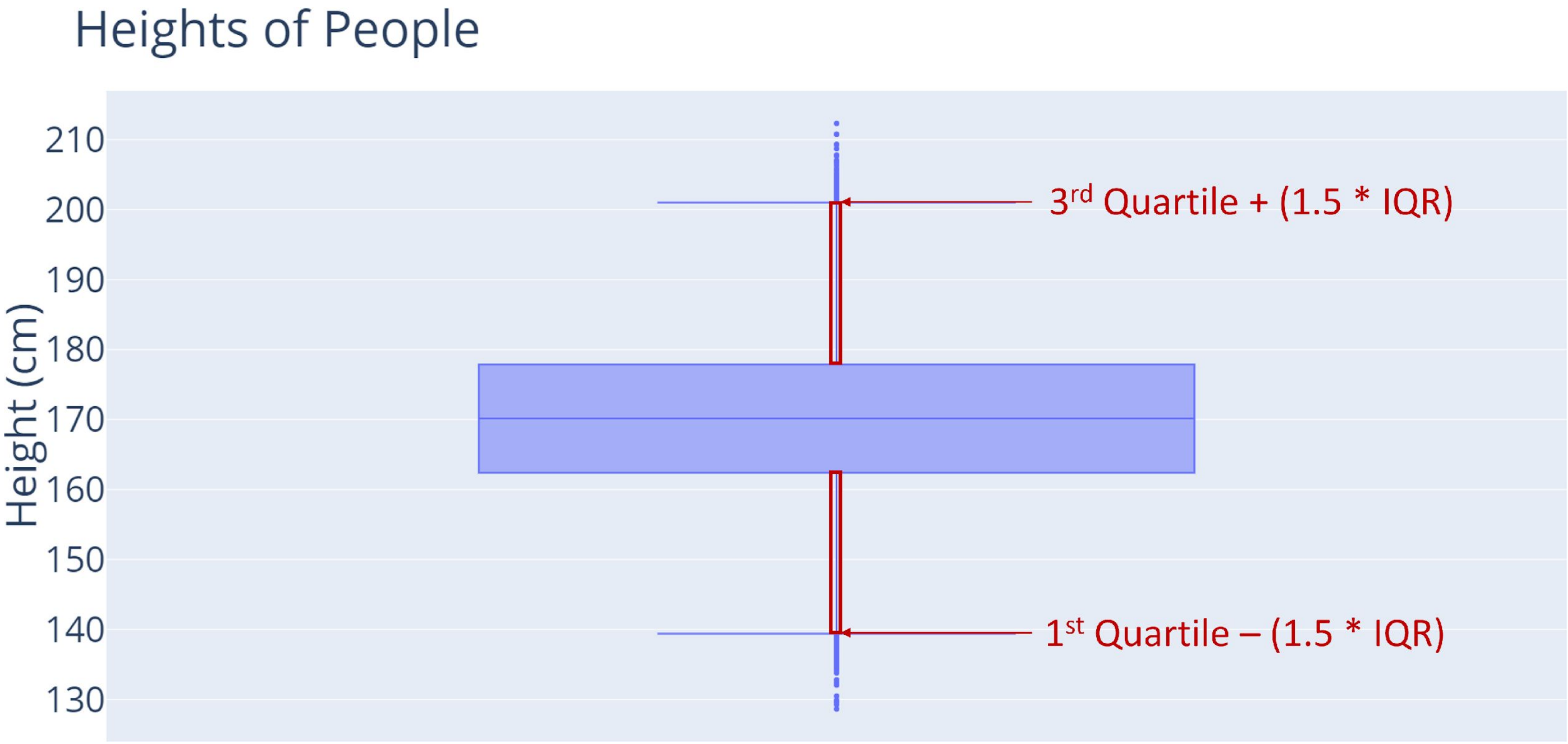
What are boxplots?



What are boxplots?



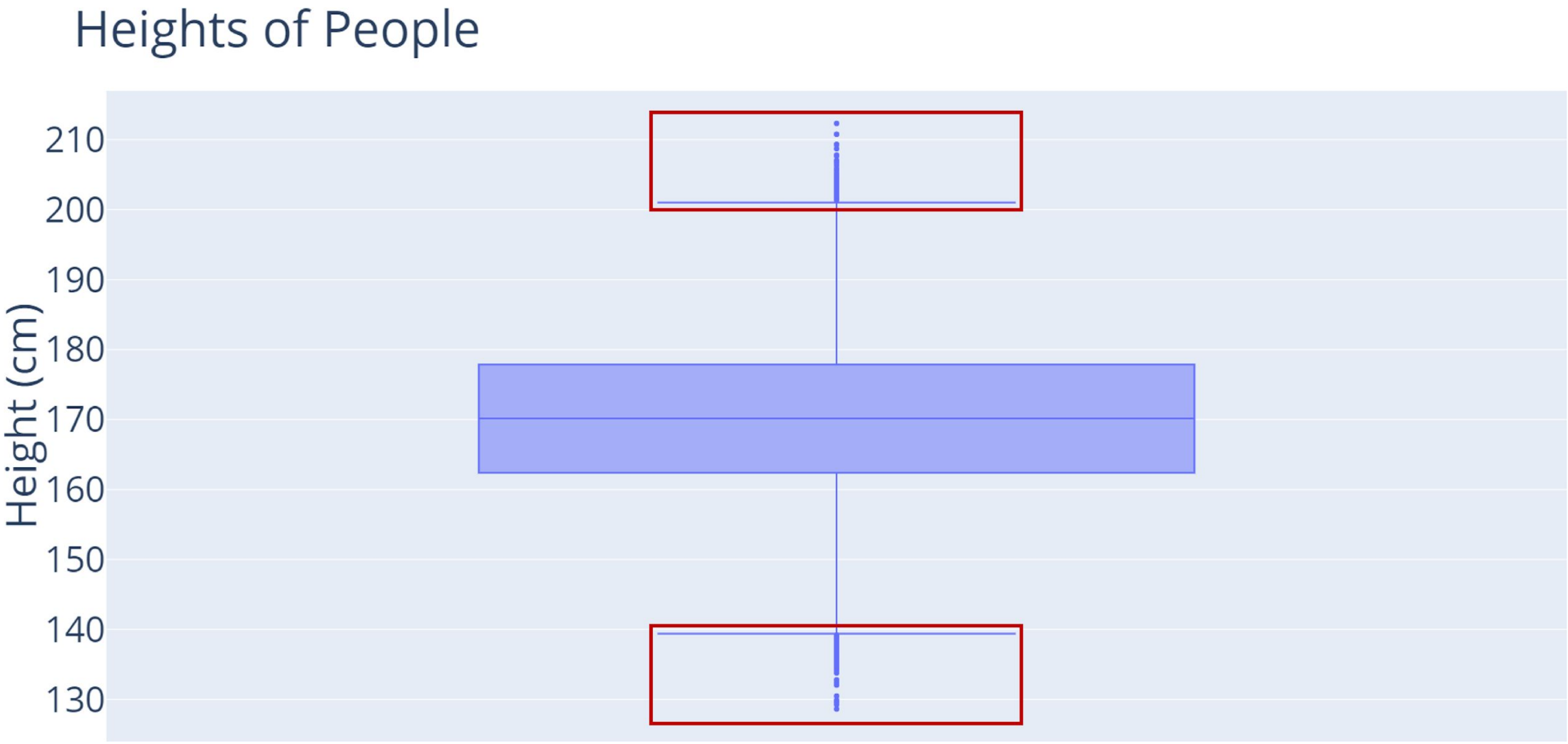
What are boxplots?



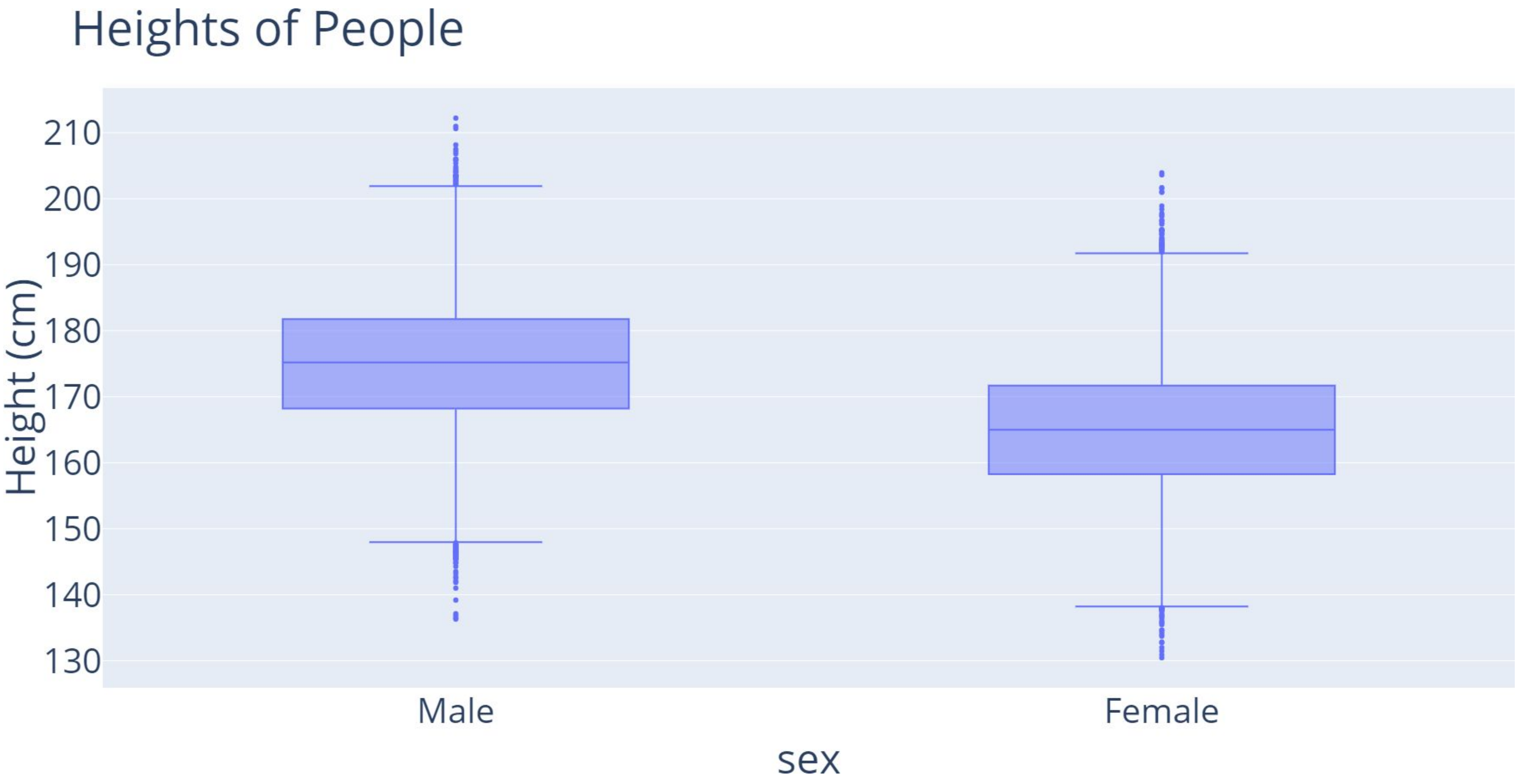
3rd Quartile + (1.5 * IQR)

1st Quartile – (1.5 * IQR)

What are boxplots?



Comparing distributions with categorical variables



Creating new variables

Data mutation: creating new variables to refine an analysis or visualization

Creating new variables

Data mutation: creating new variables to refine an analysis or visualization

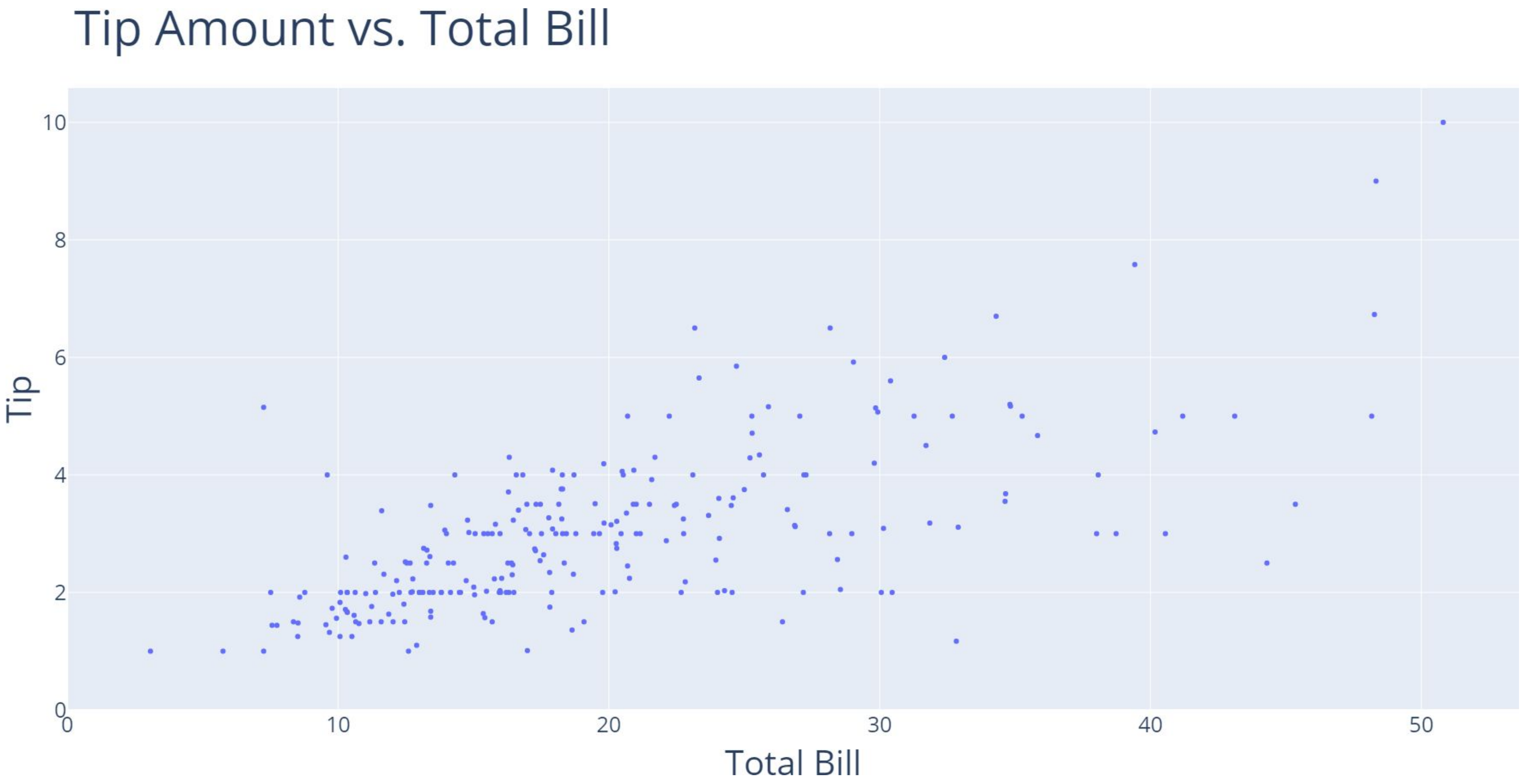
Age	Age Group
18	Teen
19	Teen
20	Early Adult
21	Early Adult
30	Adult
31	Adult
40	Middle Age
41	Middle Age

Course Title	Course Type
Introduction to Power BI	Power BI
Unsupervised Learning in R	R
DAX in Power BI	Power BI
Introduction to Python	Python

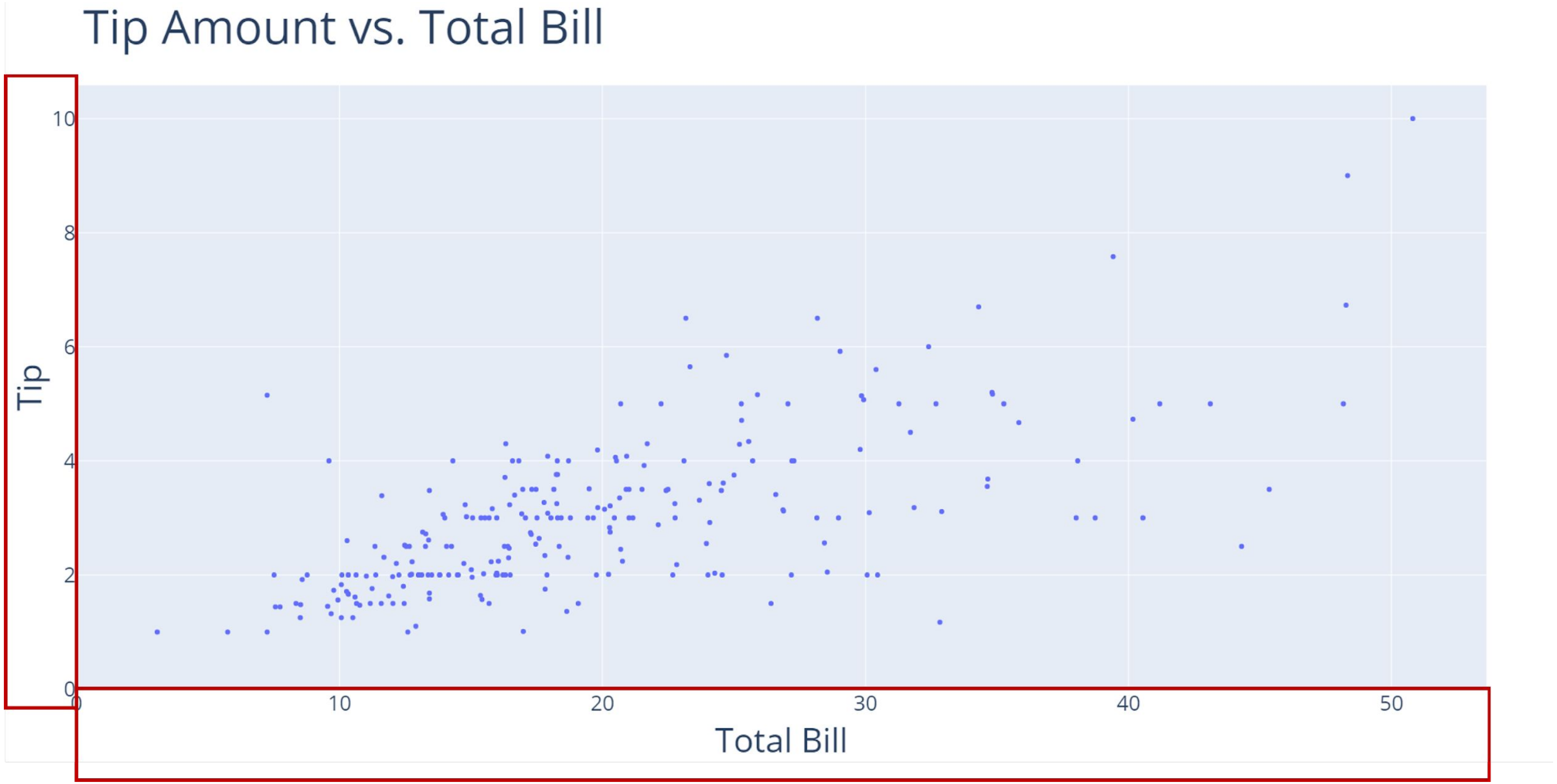
Demo

4. Relationships between continuous variables

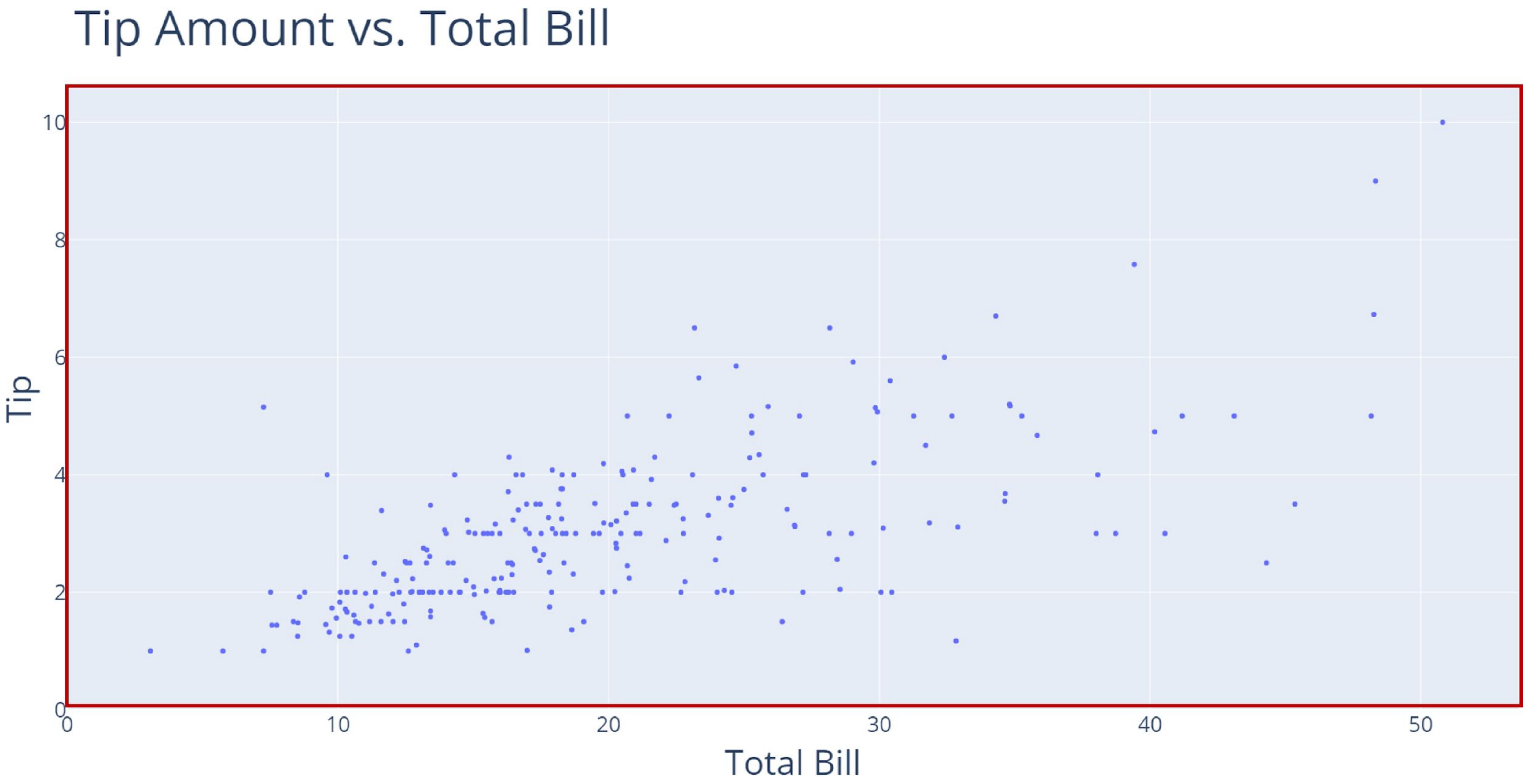
What are scater plots?



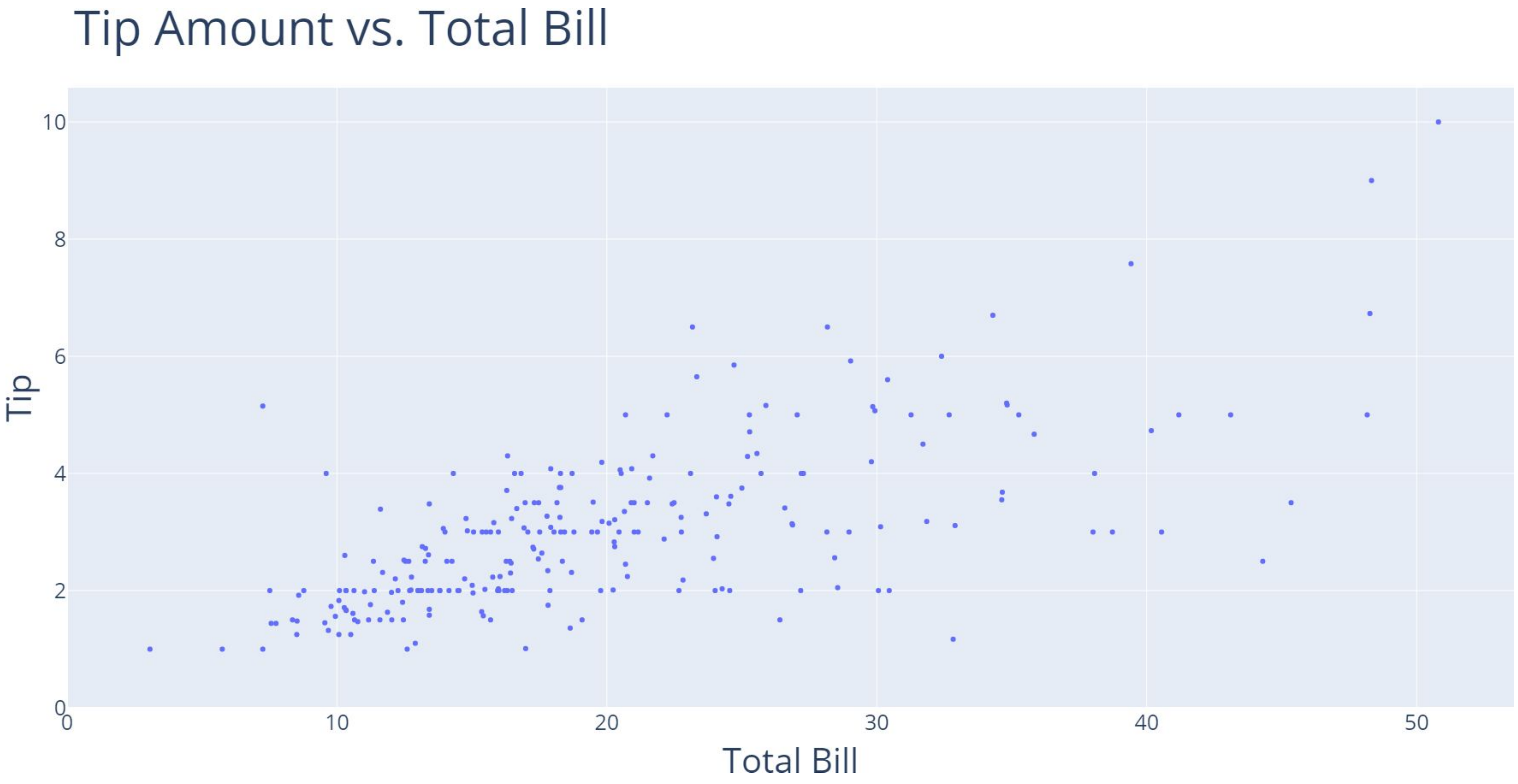
What are scater plots?



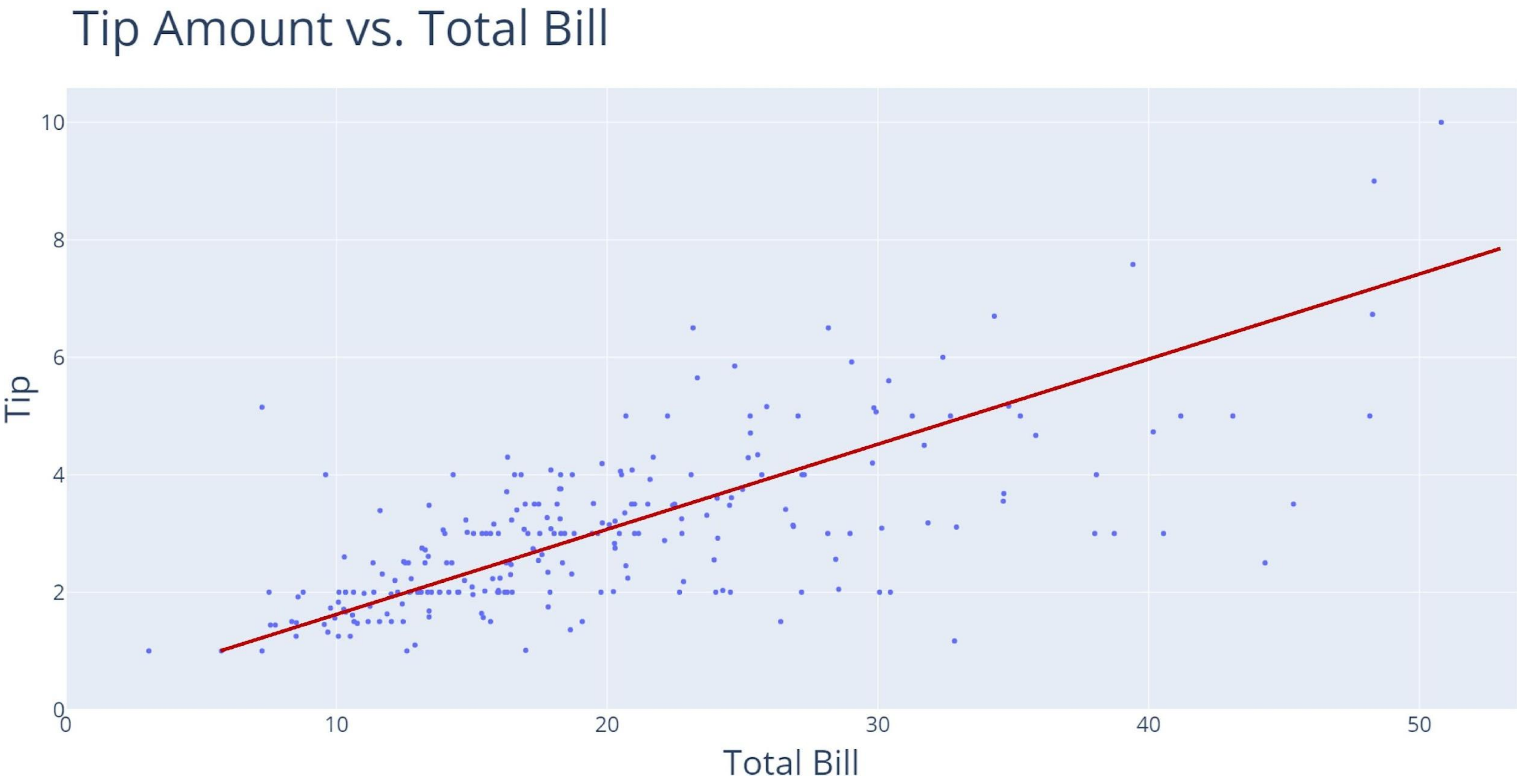
What are scater plots?



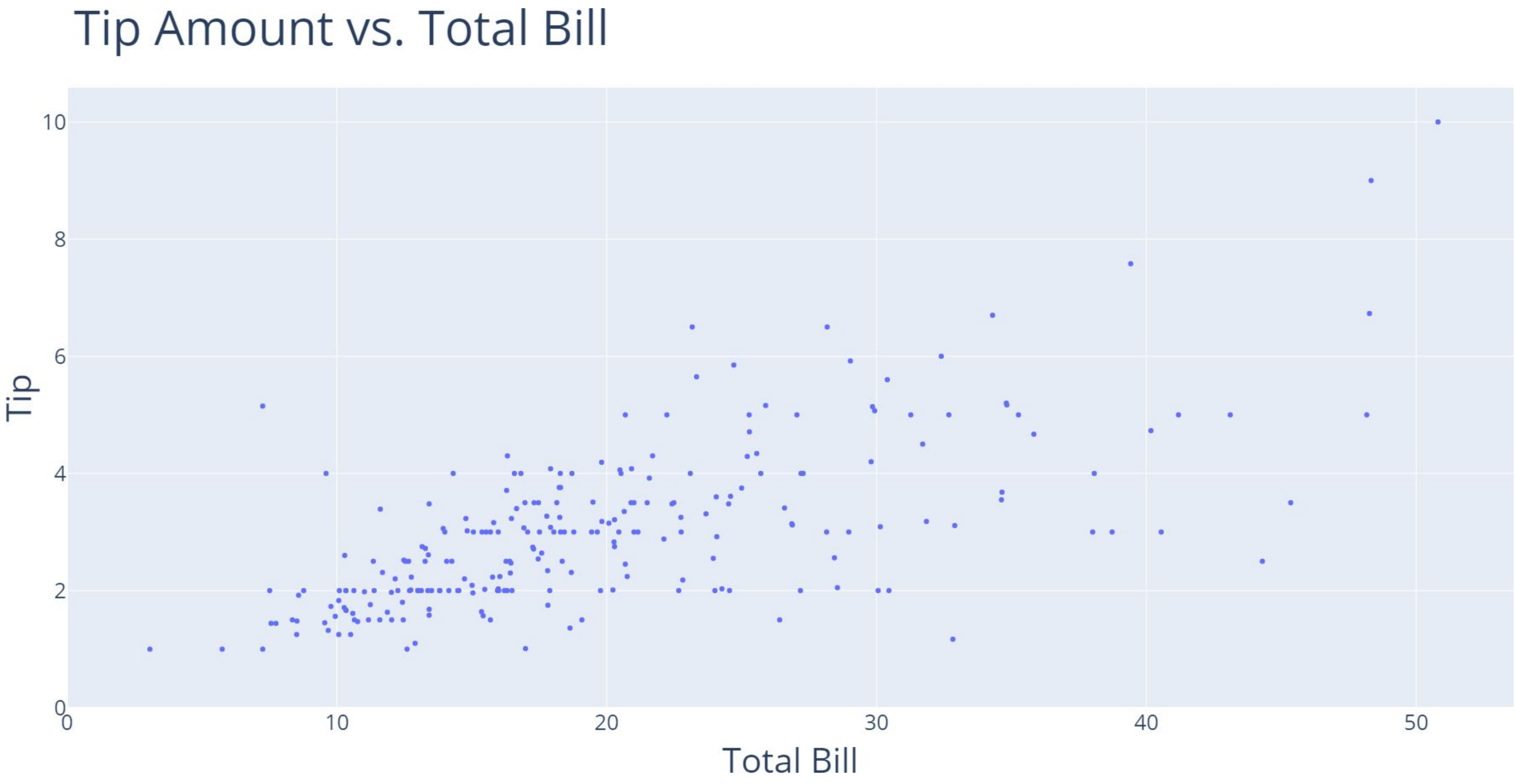
Interpreting a scater plot



Interpreting a scater plot

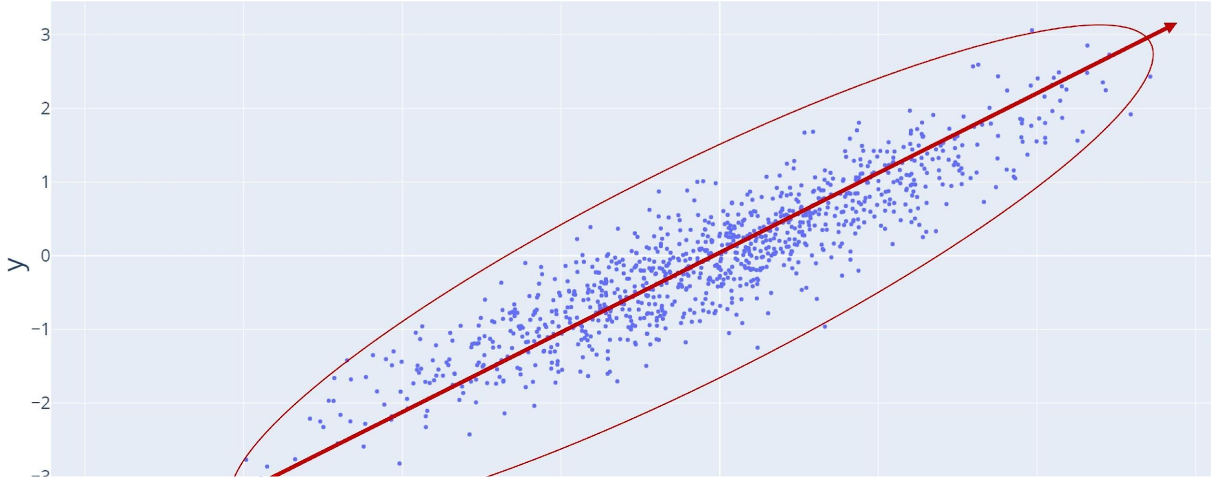


Interpreting a scater plot

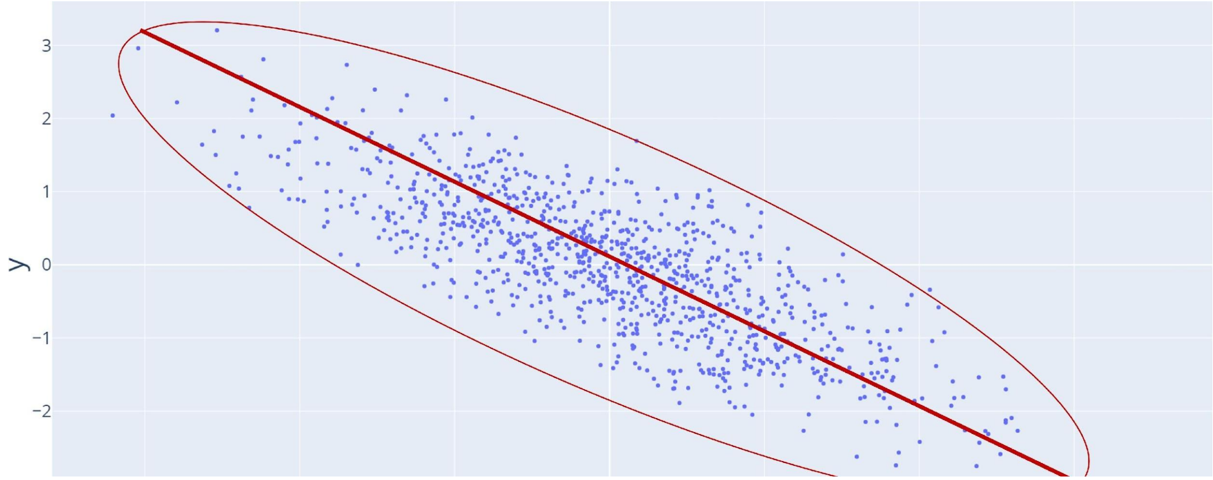


Interpreting a scater plot

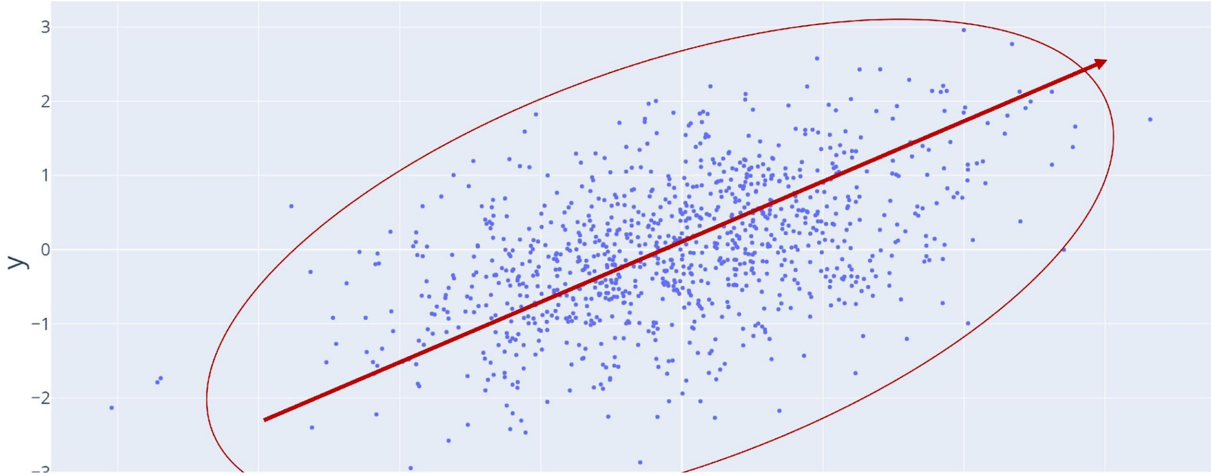
Strong-positive



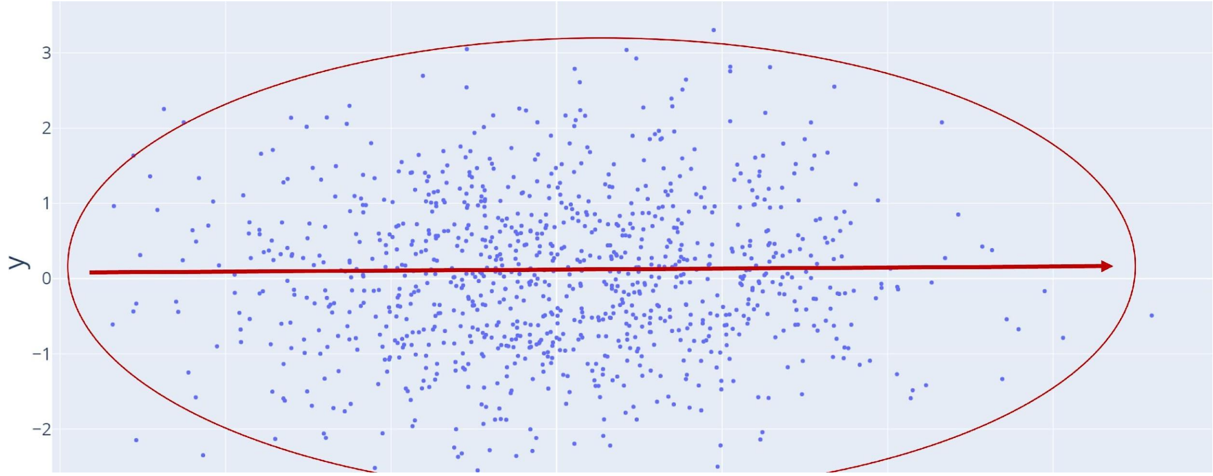
Strong-negative



Weak-positive



No relationship



Correlation coefficient

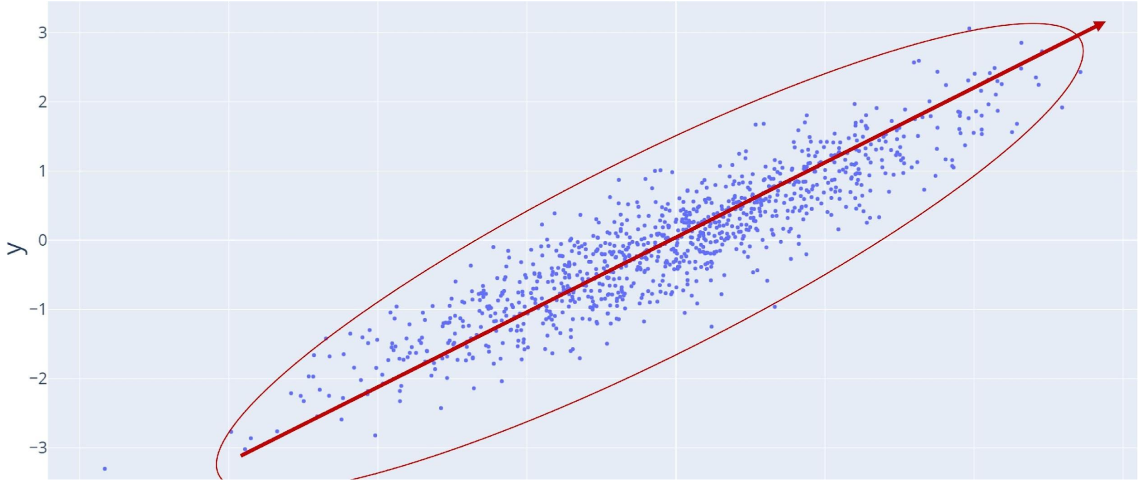
- Used to quantify the relationship
- Represented by the letter, *r*

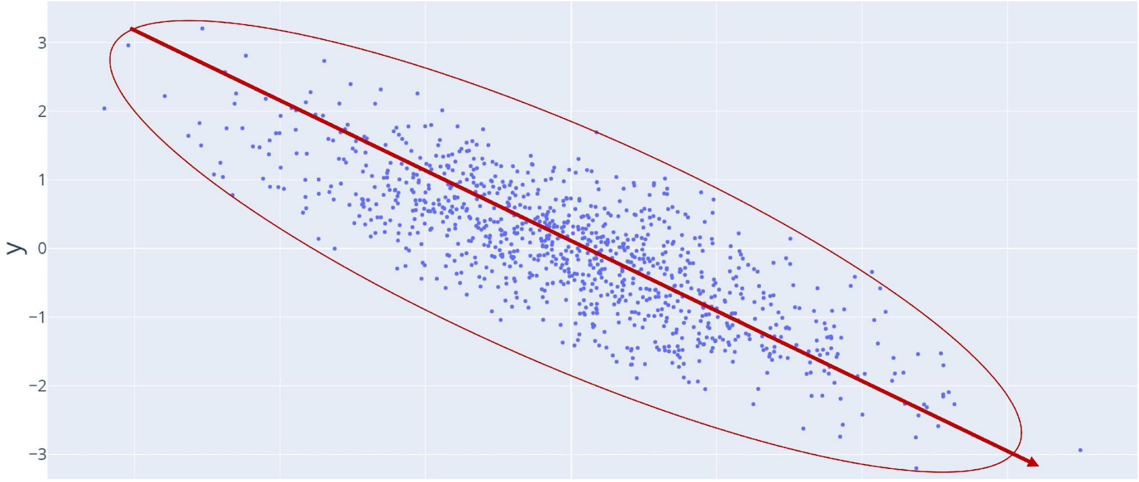
r =	**Relationship description**
-1	Strong-negative
0	No relationship
1	Strong-positive

Calculating the correlation coefficient is beyond the scope of this course

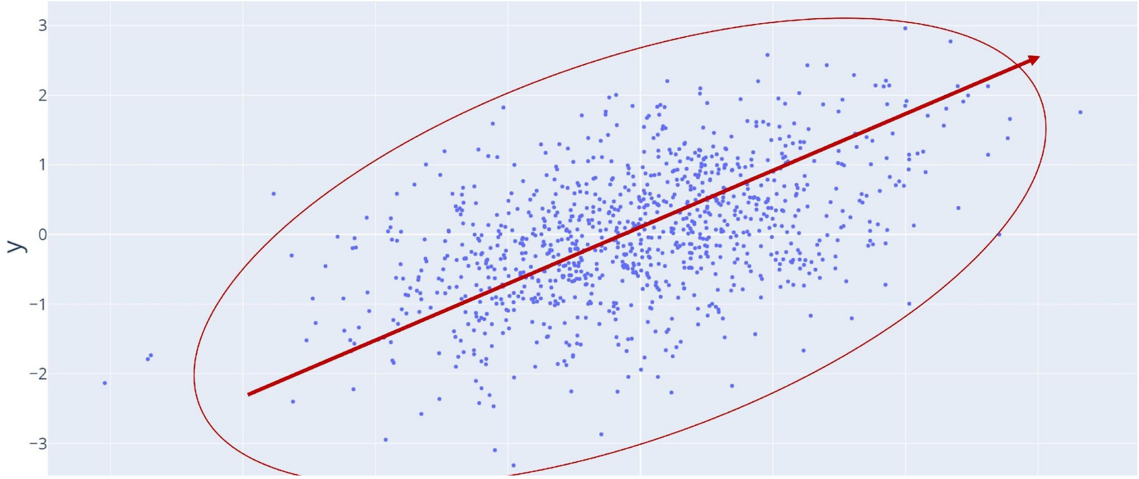
Correlation coefficient and scater plots

Strong-positive **r=0.9**Strong-negative **r=-0.9**

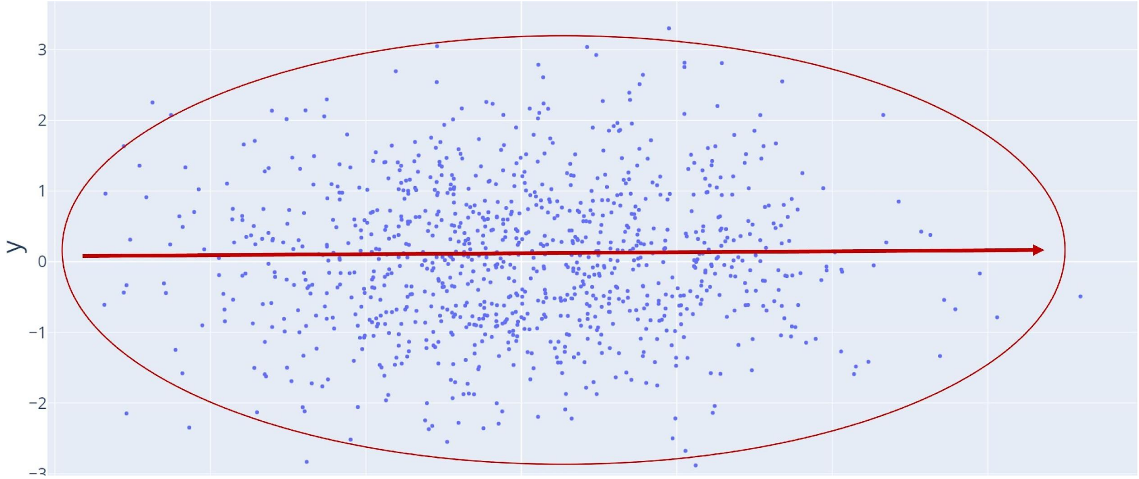




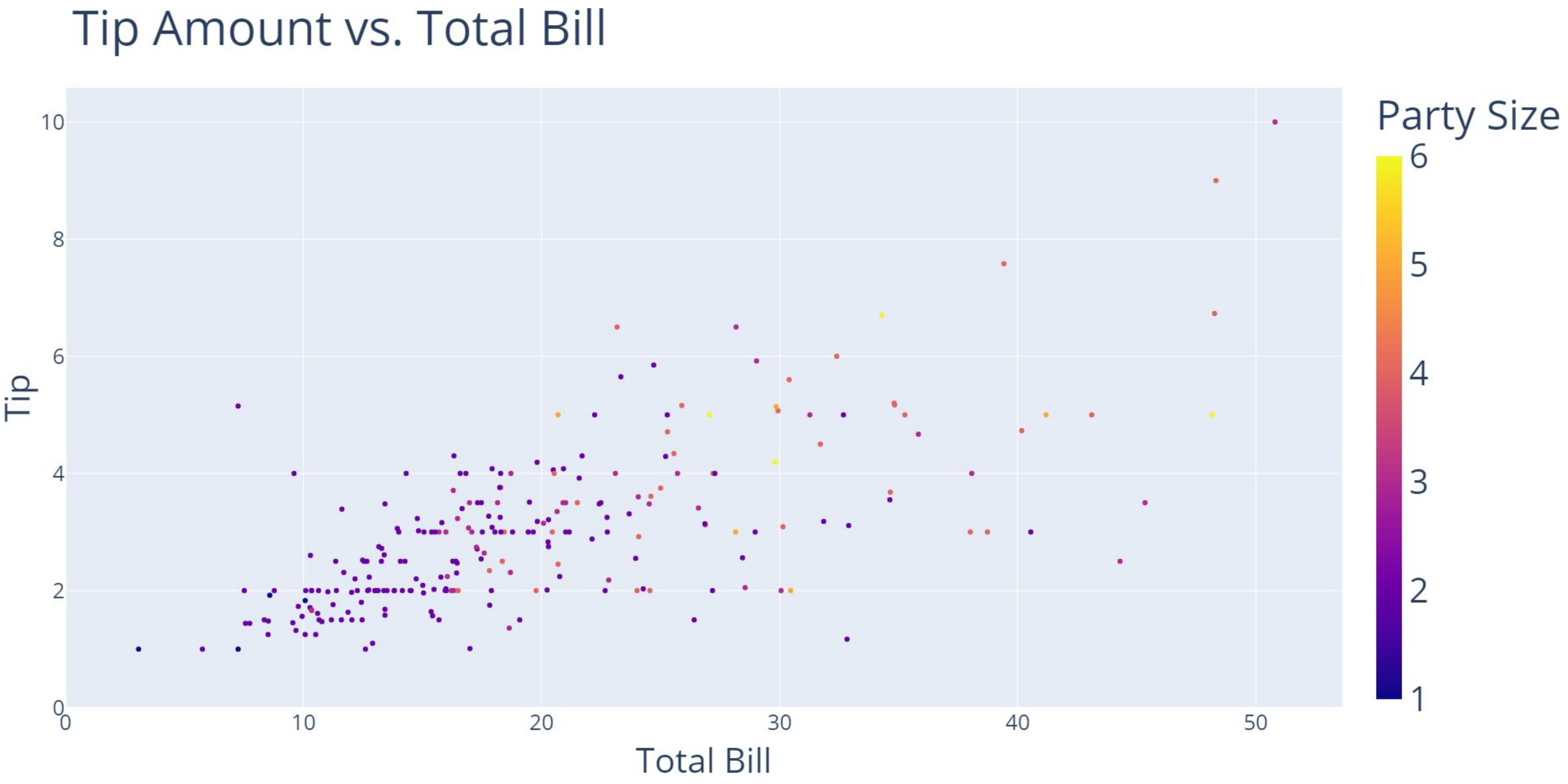
Weak-positive **r=0.35**



No relationship **r=0.0**



Adding context to a scater plot



Demo

Congratulations!

Your first steps with EDA

- Identifying and imputation of missing data
- Address outliers
- EDA with categorical variables
- EDA with continuous variables

- Histograms
- Box plots
- Scatter plots