Distributions and outliers

EXPLORATORY DATA ANALYSIS IN POWER BI

Jacob H. Marquez
Data Scientist at Microsoft

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?

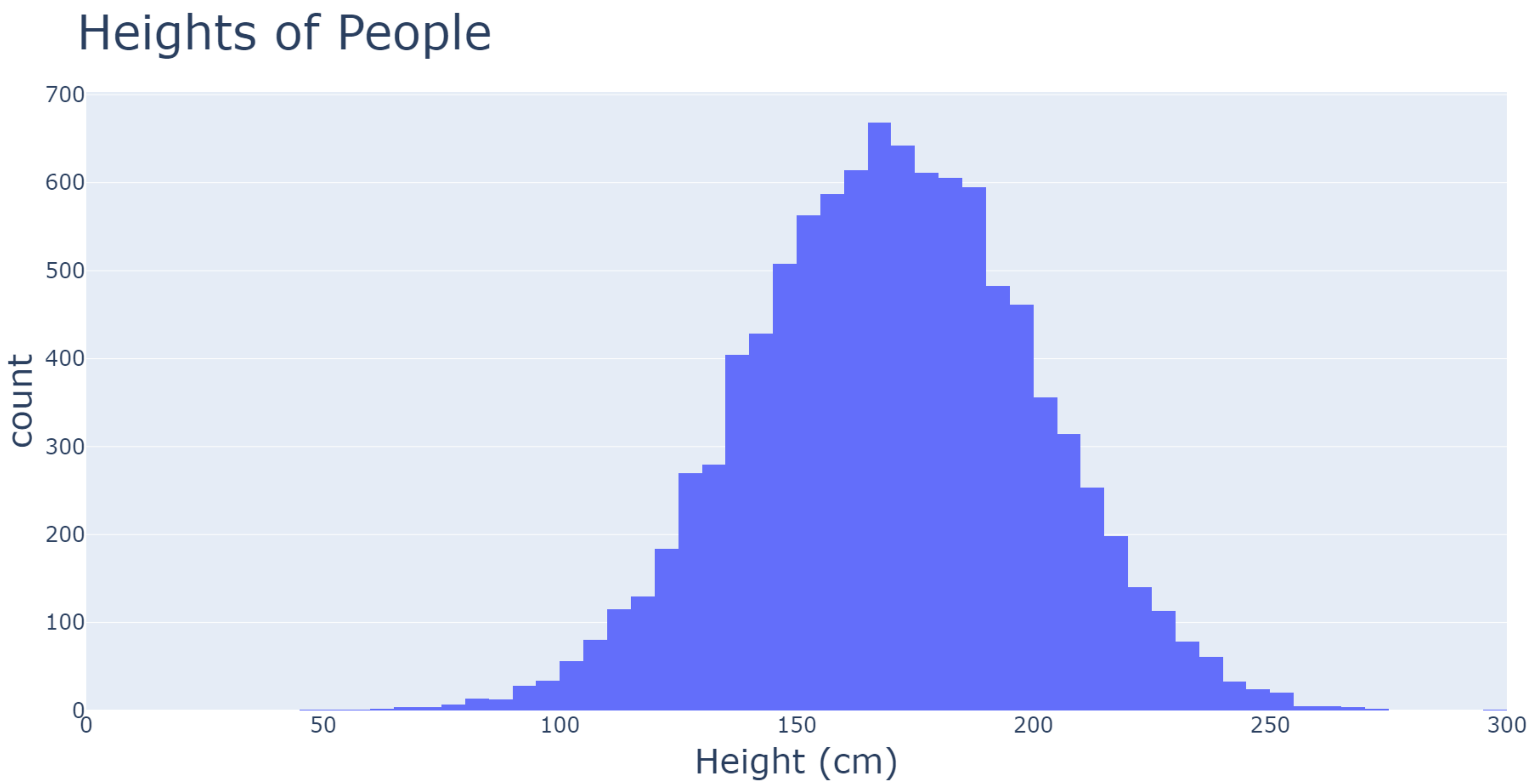
Continuous

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

Categorical

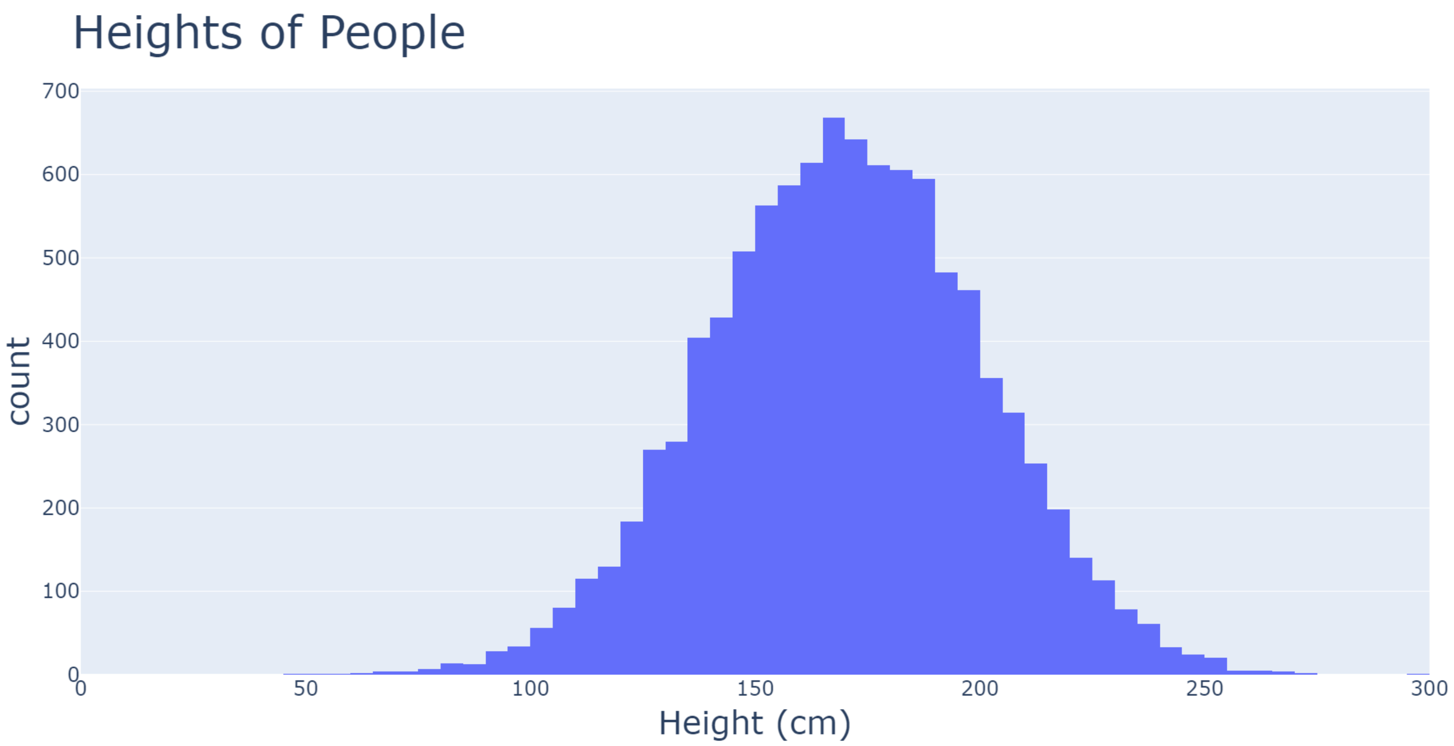
Hair Color	Frequency
Blonde	30
Brown	50
Black	40
Red	20
Grey	20

What are histograms?

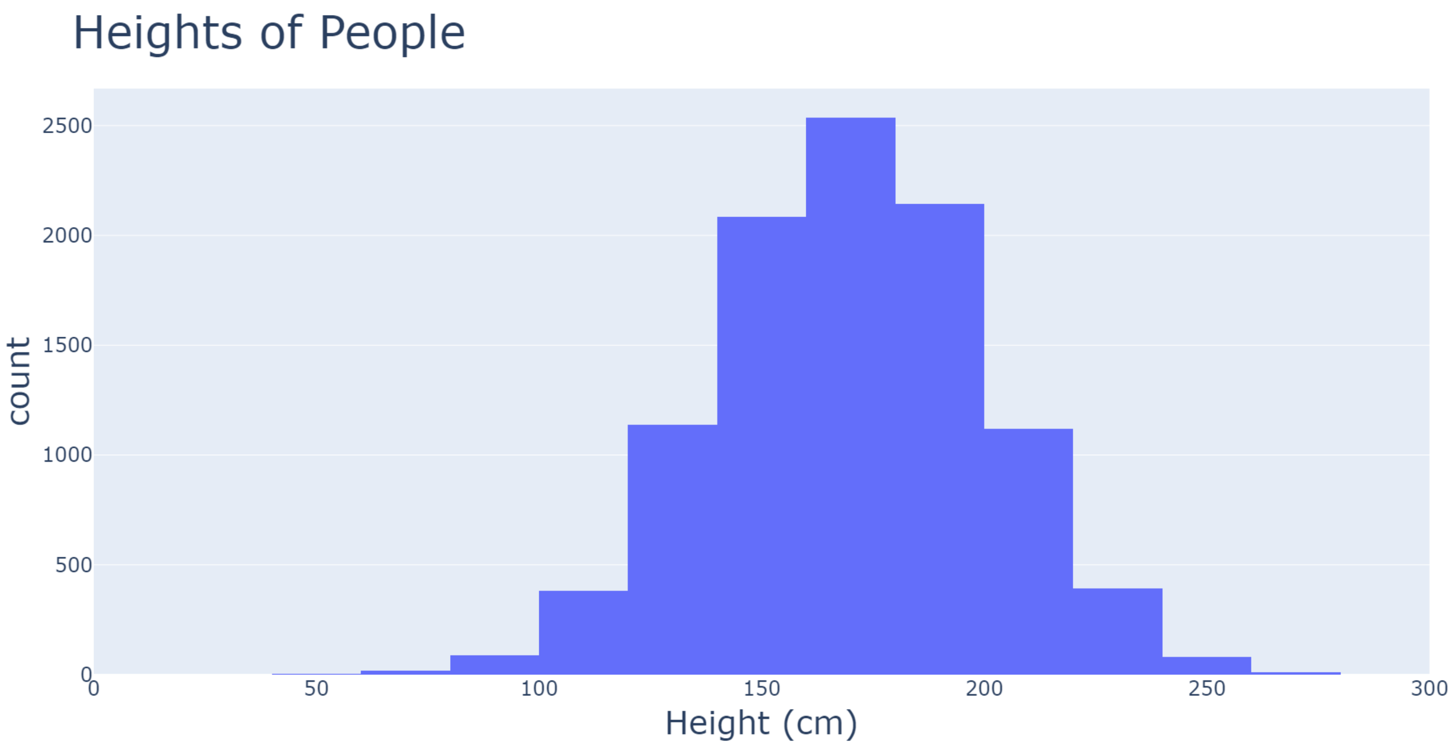


What are histogram? - bins

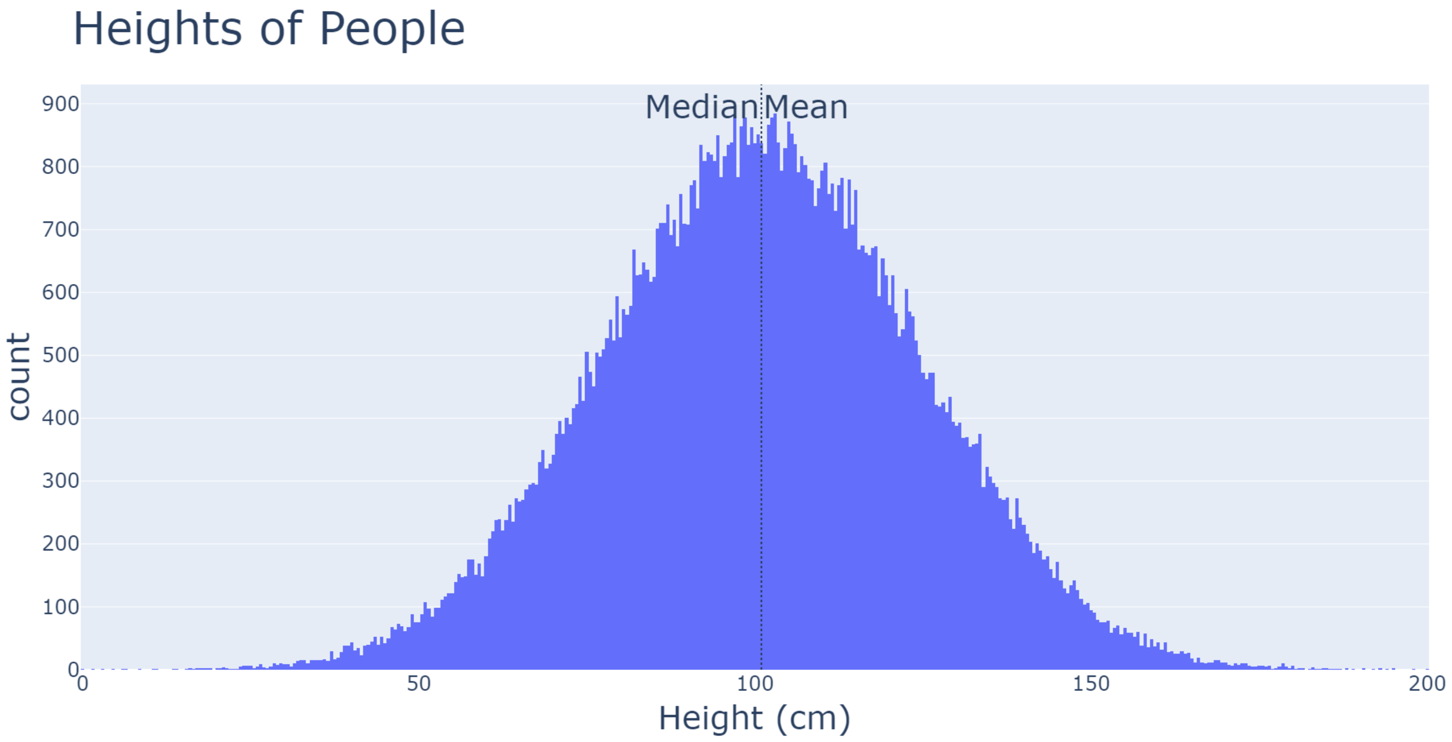
Histogram with 100 bins

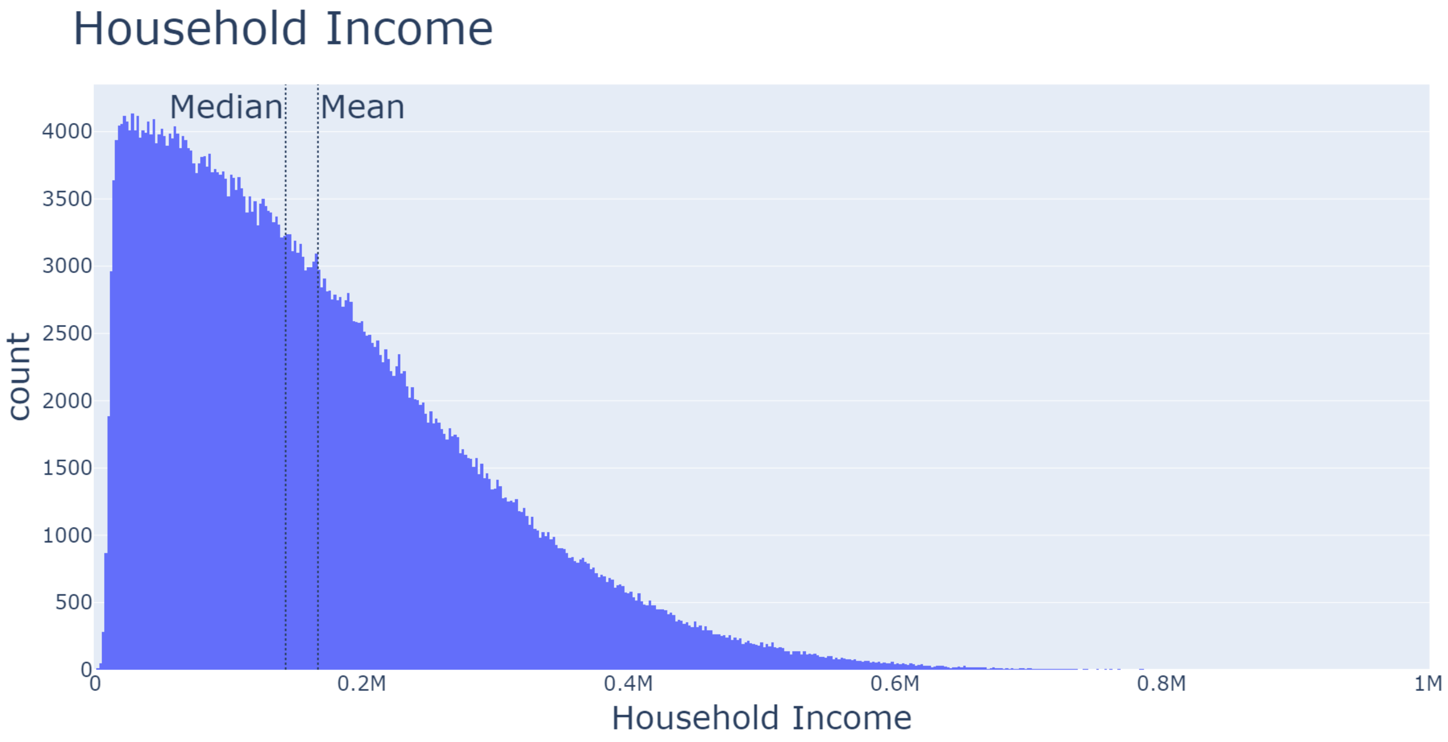


Histogram with 20 bins



Reading histograms - centrality and skewness



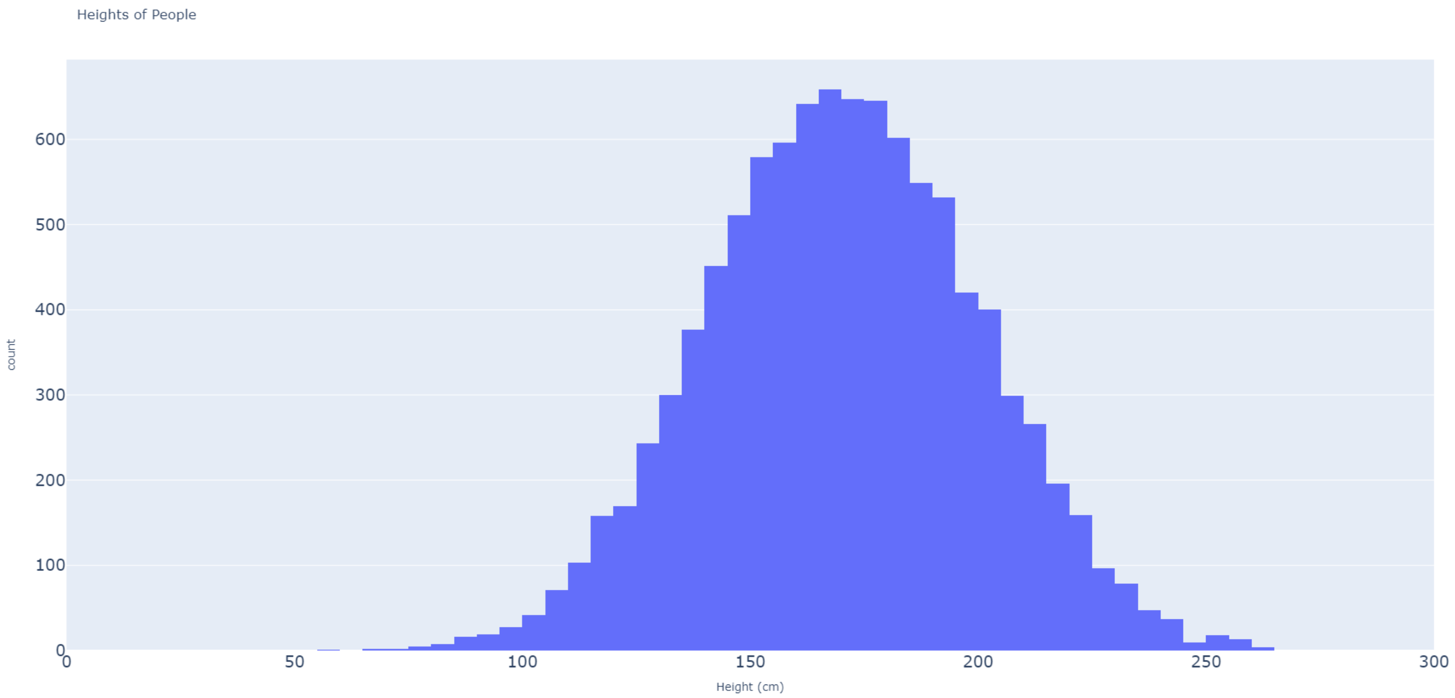


Normal distribution

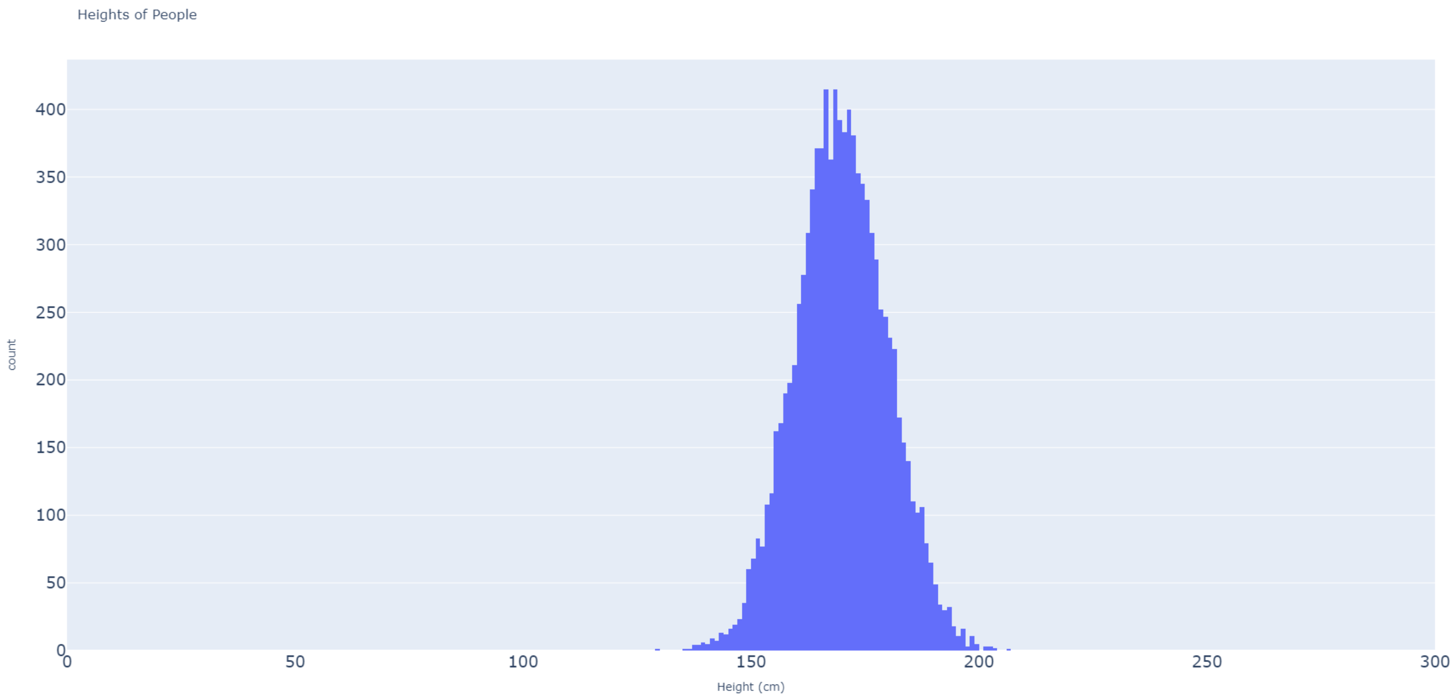
Right-skewed distribution

Reading histograms - spread

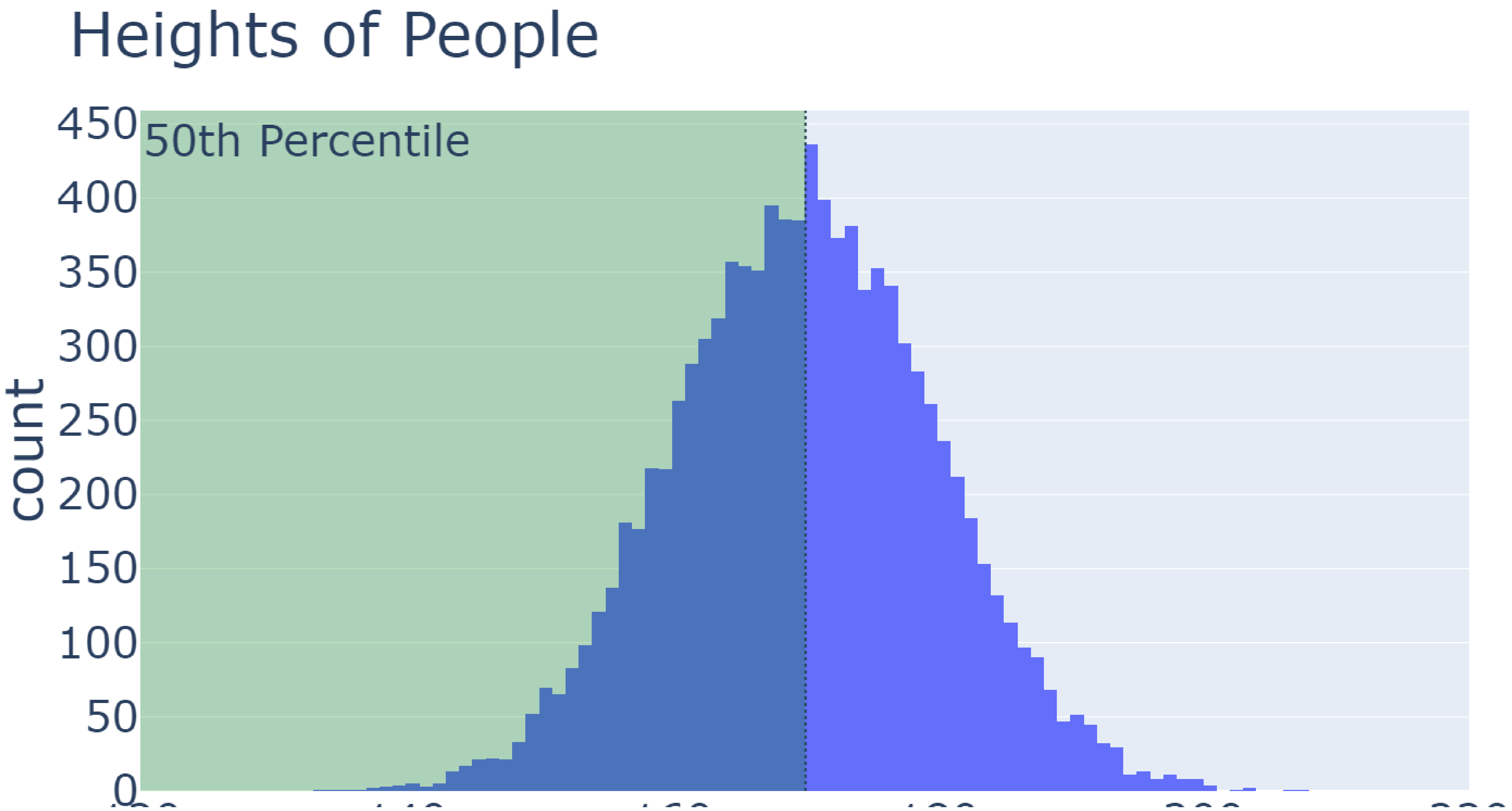
Larger standard deviation



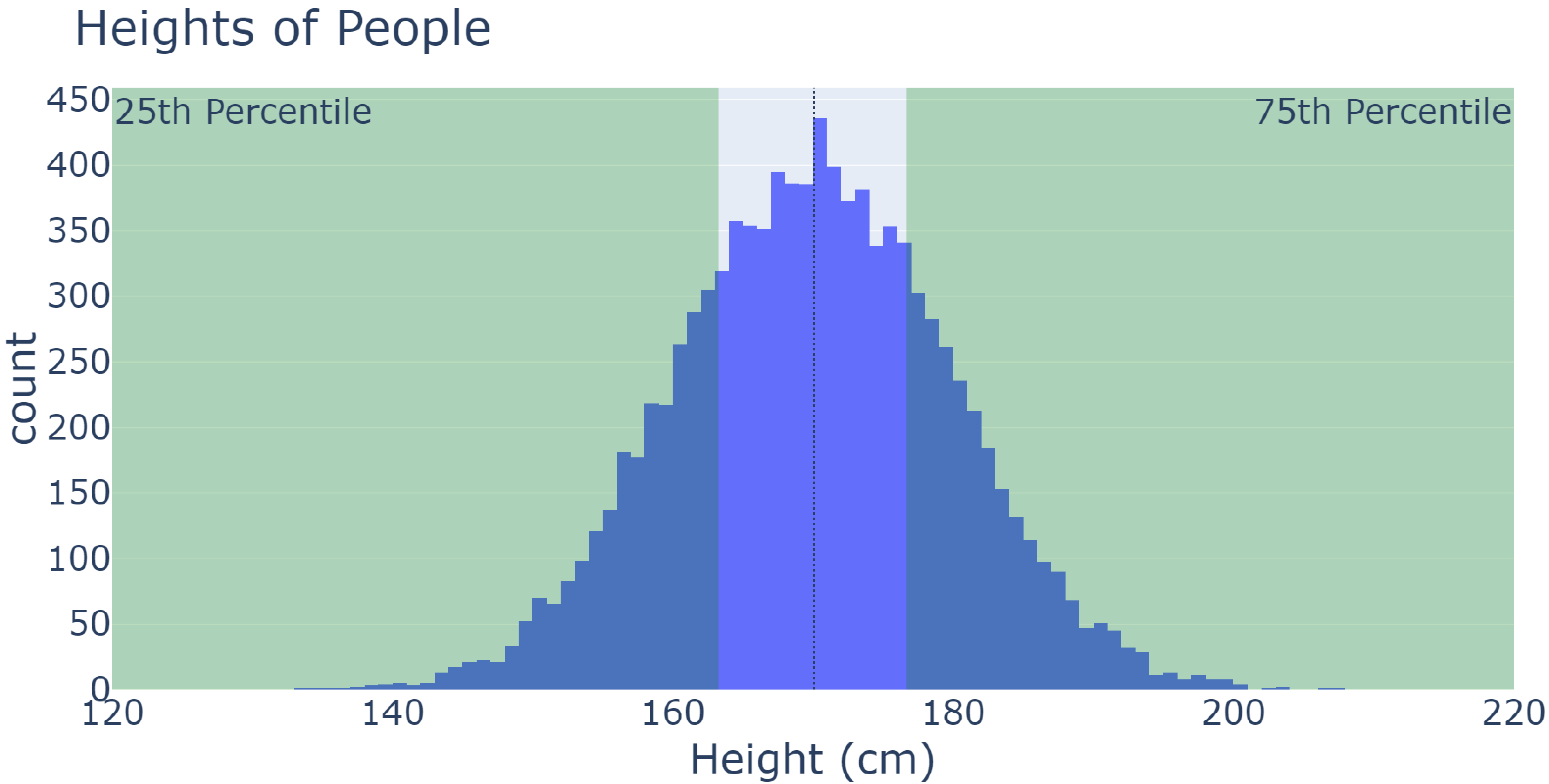
Smaller standard deviation



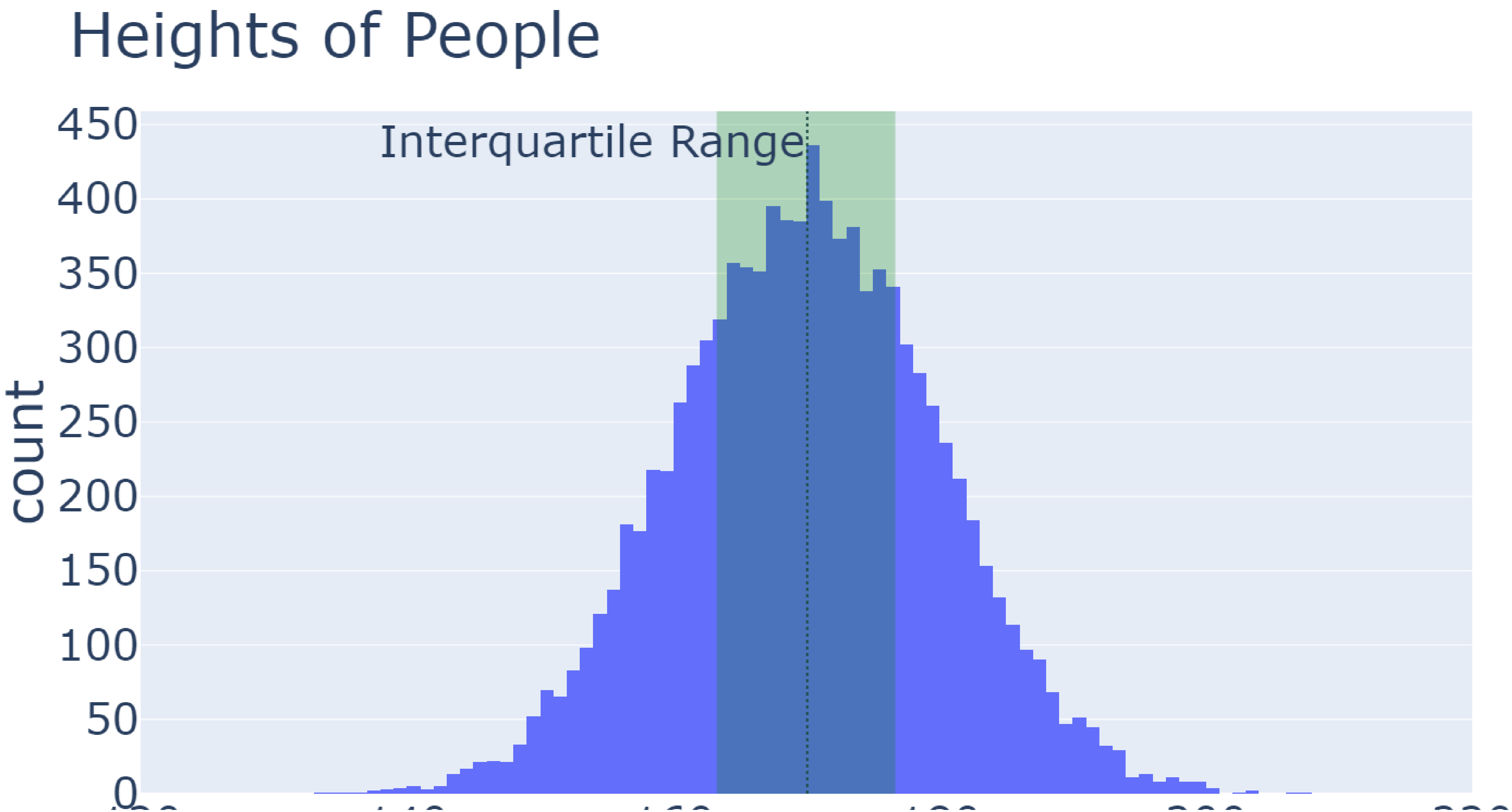
Reading histograms - percentiles



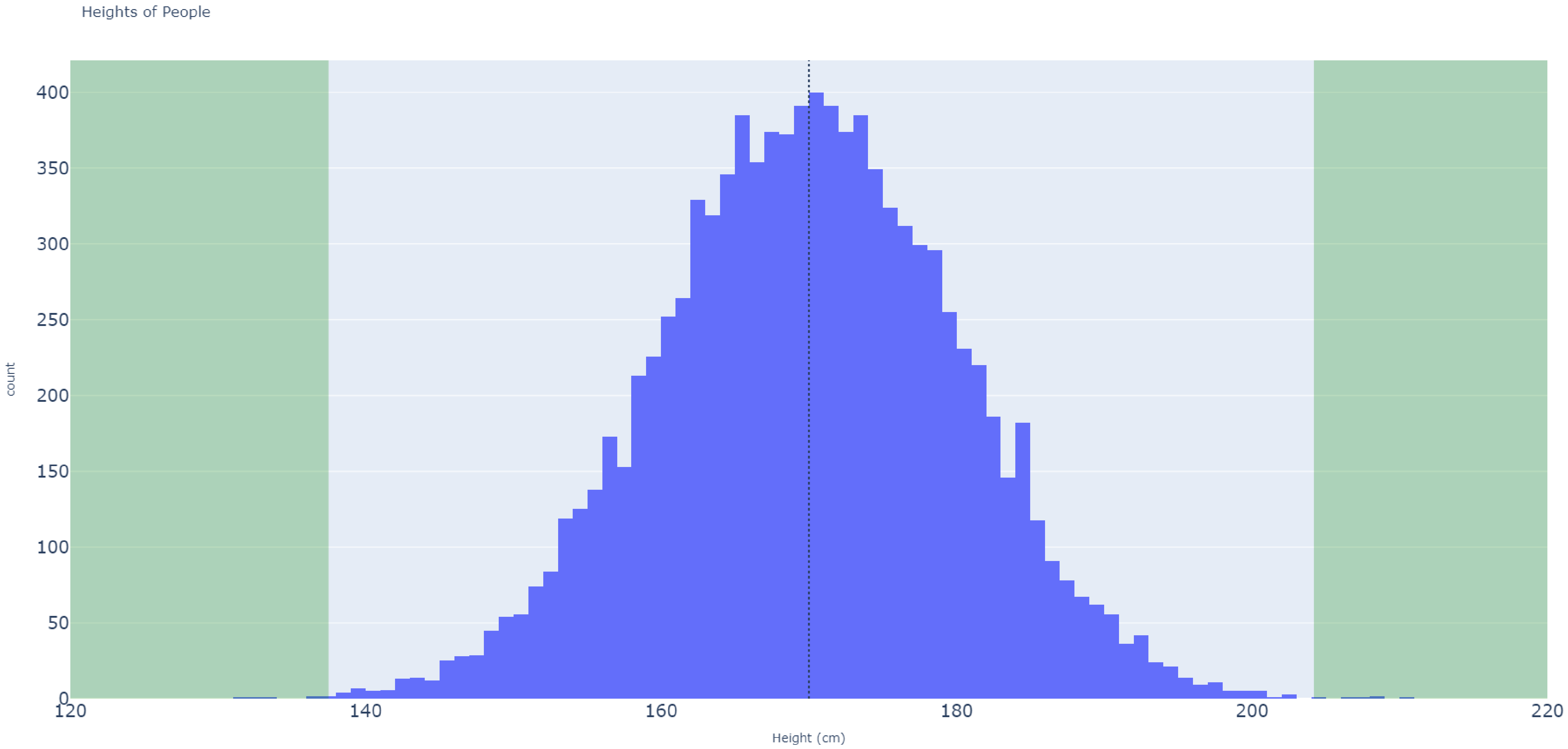
Reading histograms - 25th & 75th percentiles



Reading histograms - interquartile range



What is an outlier?



Finding outliers

Using standard deviation

$$lower = -3 * SD$$

$$upper = 3 * SD$$

Outlier when

$value < lower$ OR $value > upper$

Interquartile Range (IQR)

$$lower = 25percentile - (1.5 * IQR)$$

$$upper = 75percentile + (1.5 * IQR)$$

Outlier when

$value < lower$ OR $value > upper$

Addressing outliers

1. Remove observations
2. Imputation

Winsorizing

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

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

Let's practice!

EXPLORATORY DATA ANALYSIS IN POWER BI

Histograms and outliers in AirBnB listings

EXPLORATORY DATA ANALYSIS IN POWER BI

Jacob H. Marquez
Data Scientist at Microsoft

Let's practice!

EXPLORATORY DATA ANALYSIS IN POWER BI