Cmd 1

databricks Academy

Cmd 2

# Descriptive Statistics

**Objective**: *Use descriptive statistics to gather information about a data set.*

In this lab, you will complete a series of exercises to calculate various summary statistics of the `dsfda.ht_daily_metrics` data set.

Cmd 3

```
%run "../../Includes/Classroom-Setup"
```

Cmd 4

## Part 1: Measures of Central Tendency

Cmd 5

### Exercise 1: Mean

Use the `AVG` SQL function to calculate the average resting heartrate, as an integer, for each lifestyle.

**Hint:** If needed, check out the Spark SQL documentation on `avg`.

Cmd 6

```
%sql
-- ANSWER

SELECT CAST(AVG(resting_heartrate) AS int) as avg_resting_heartrate, lifestyle
FROM dsfda.ht_daily_metrics
GROUP BY lifestyle
```

Cmd 7

### Exercise 2: Median

Use the `percentile` SQL function to calculate the median resting heartrate, as an integer, for the "Athlete" lifestyle.

Note that this code block takes longer to run; this is because calculating the median requires shuffling across worker nodes.

**Hint:** If needed, check out the Spark SQL documentation on `percentile`.

Cmd 8

```
%sql
-- ANSWER

SELECT CAST(percentile(resting_heartrate, 0.5) AS int) AS median_resting_heartrate
FROM dsfda.ht_daily_metrics
WHERE lifestyle = 'Athlete'
```

Cmd 9

### Exercise 3: Mode

Calculate the mode, as an integer, of the resting heartrate for each lifestyle.

**Hint:** Cast the heartrate to the nearest integer first. You can use the Spark SQL documentation on `cast` for reference.

**Hint:** Try to find the `mode` Spark SQL function in the documentation on your own.

Cmd 10

```
%sql
-- ANSWER

SELECT COUNT(*) AS count, CAST(resting_heartrate AS int) AS resting_heartrate
FROM dsfda.ht_daily_metrics
GROUP BY resting_heartrate
ORDER BY count DESC
```

Cmd 11

## Part 2: Measures of Dispersion

Cmd 12

### Exercise 4: Standard Deviation

Calculate the standard deviation of the resting heartrate for each lifestyle.

**Hint:** Cast the heartrate to the nearest integer first.

**Hint:** Use the Spark SQL documentation to determine which function computes the standard deviation.

Cmd 13

```sql
%sql
-- ANSWER

SELECT CAST(STD(resting_heartrate) AS int) as std_resting_heartrate, lifestyle
FROM dsfda.ht_daily_metrics
GROUP BY lifestyle
```

Cmd 14

### Exercise 5: Interquartile Range

Calculate the interquartile range of resting heartrate, as an integer, for each lifestyle.

**Hint:** Recall that the interquartile range is the difference between the 75th percentile and the 25th percentile.

**Hint:** Refer to previous exercises in this lab to determine which Spark SQL function can be used to compute the value at a given percentile.

Cmd 15

```sql
%sql
-- ANSWER

SELECT CAST(percentile(resting_heartrate, 0.75) - percentile(resting_heartrate, 0.25) AS int) AS iqr_resting_heartrate
FROM dsfda.ht_daily_metrics
GROUP BY lifestyle
```

Cmd 16

Congratulations! You've completed the Descriptive Statistics Lab.

If you have any trouble with any of this course's labs, be sure to check out the solutions to the labs in the Solutions folder.

Cmd 17



Shift+Enter to run