Week 3 Tutoring

Visualization, Functions, and Grouping

Histograms vs. Bar Charts

- Uses numerical variables
  - Heights of dogs
  - Weights of dogs
  - Time dog takes to run 100 meters

- Uses categorical variables
  - Breeds of dogs
  - Colors of dogs
  - Moods of dogs

More histogram review

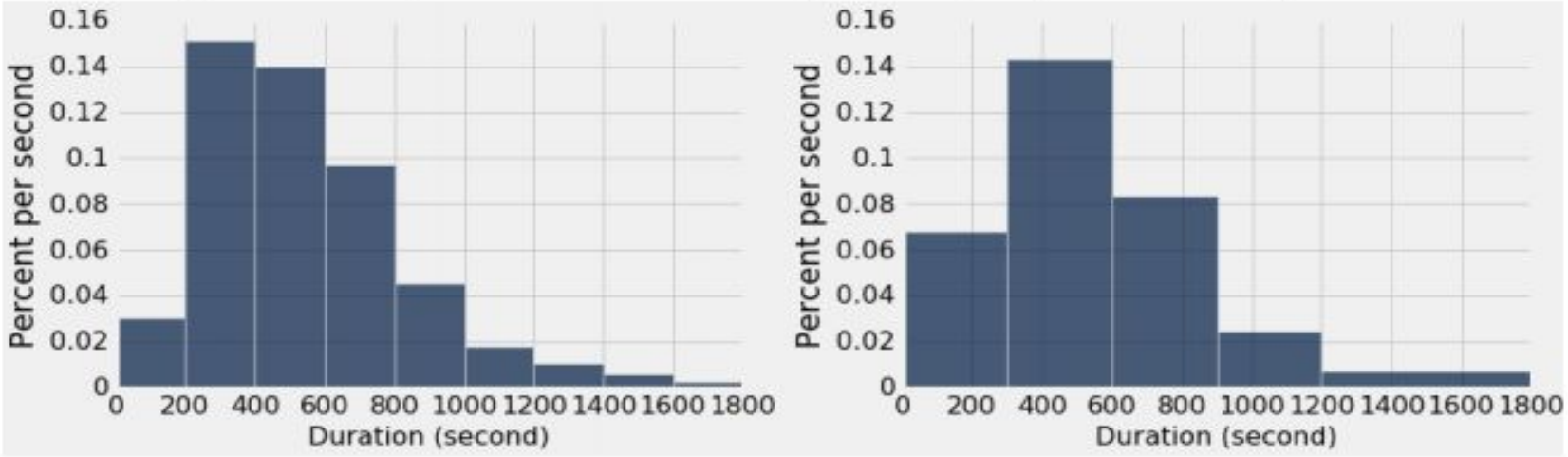
- X-axis:
  - Bins are groupings of the numerical variable into continuous intervals
    - E.g. ages could be split into three bins: 0-45, 45-70, 70-120
  - Note that bin widths may not be equal
- Y-axis:
  - P**roportion of items relative to width of bin (density)**
  - From textbook: **area of bar = height of bar x width of bar**
  - So:
    - Height of bar = area of bar / width of bar

Histogram Challenge Question

- From Professor DeNero's Spring 2016 midterm

3. (15 points) Distributions

The two histograms of bike trip durations below were both generated by `trip.hist(...)` using different bins.



(a) (8 pt) Write the proportion of trips that fall into each range of durations below. *Show your work.* If it is not possible to tell from the histograms, instead write **Not enough information**.

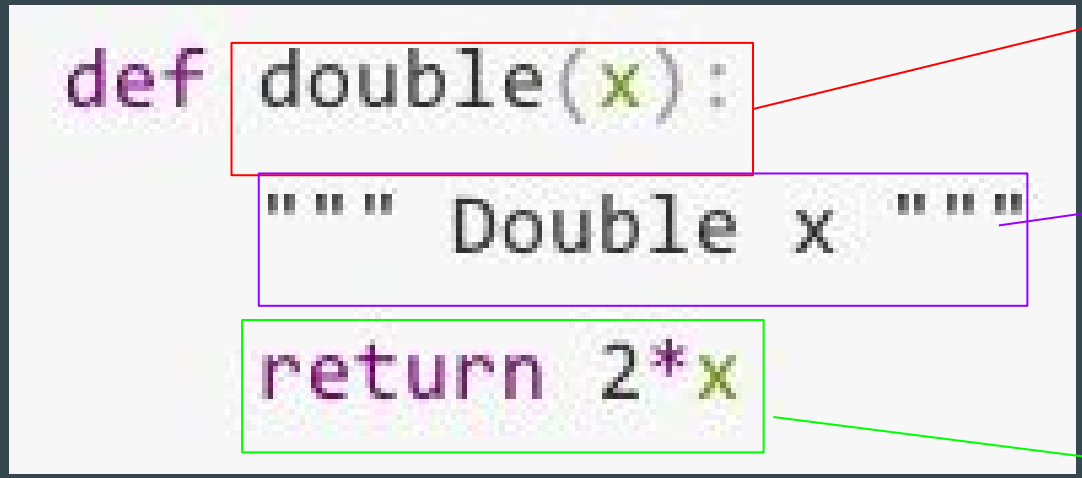
- Between 200 (inclusive) and 400 (exclusive) seconds

- Between 300 (inclusive) and 900 (exclusive) seconds

- Between 400 (inclusive) and 900 (exclusive) seconds

Functions

- What does each component of the function definition do?



How to use: Apply

- Calls a function on each element of a column, forming a new array of return values.
- Syntax:
  - <table name>.apply(<function>, ‘<column name>’)
    - Note that the name of the column goes inside quotes

How to use: Group with one argument

- Counts the number of rows for each category in a column
- Syntax:
  - <table name>.group(‘<column name>’)
  - Note that the name of the column goes inside quotes

How to use: Group with two arguments

- Aggregates values by category with whatever function is given
- Syntax:
  - <table name>.group('<column name>', <function>)

How to use: Pivot

- Makes a new table that groups together rows that share a combination of values
- Syntax:
  - <table name>.pivot('<column1>', <column2>)
    - Column 1: these values will form the new columns
    - Column 2: these values will form the new rows

Extra Pivot Arguments

- Two extra optional arguments
- Syntax:
  - <table name>.pivot('<column1>', <column2>, <values>, <collect>)
    - Values: these values will replace the counts
    - Collect: a function that will be used to aggregate the values