# MATH1401

Fall 2021

# Lecture 8

Visualization – Part II

# Class Checklist

- **HW - 2 Due Date** : Tuesday: 9/14 – 9 PM
  - Graded Questions:1.1-1.3, 2.1-2.5, 3.1-3.5, 4.1-4.2, 6.1-6.7
- **Lab 3 – Due Date** : Friday 9/17 – 9 PM
  - Graded Questions : 1.1,1.1.1, 3.1-3.2,4.1-4.3, all questions from section 2

- **Quiz 6** – Tuesday: 9/14 – Covers Chapter 7

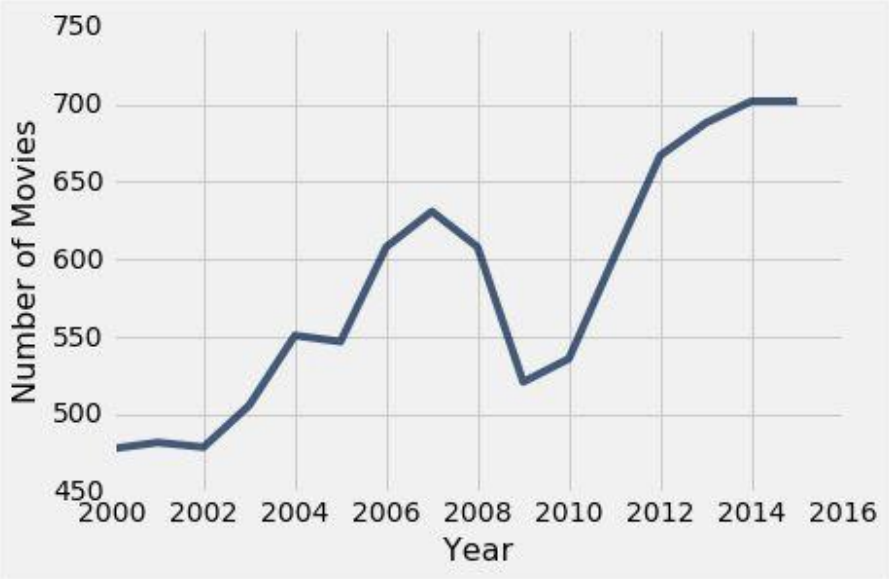
# Review

# Lecture 7 - Review

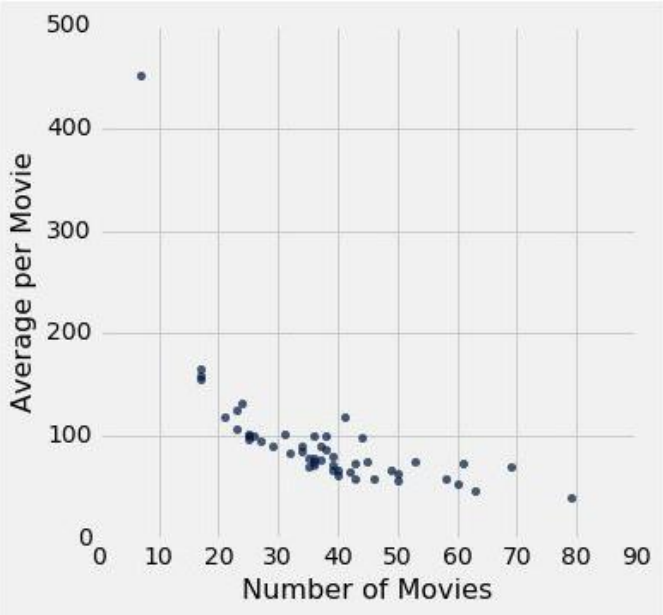
- Identify Numerical Data
- Identify Categorical Data
- Plot numerical data with Line plot, Scatter Plot
- Plot categorical data with bar graphs

# Plotting Two Numerical Variables

Line graph: `plot`




Scatter plot : `scatter`

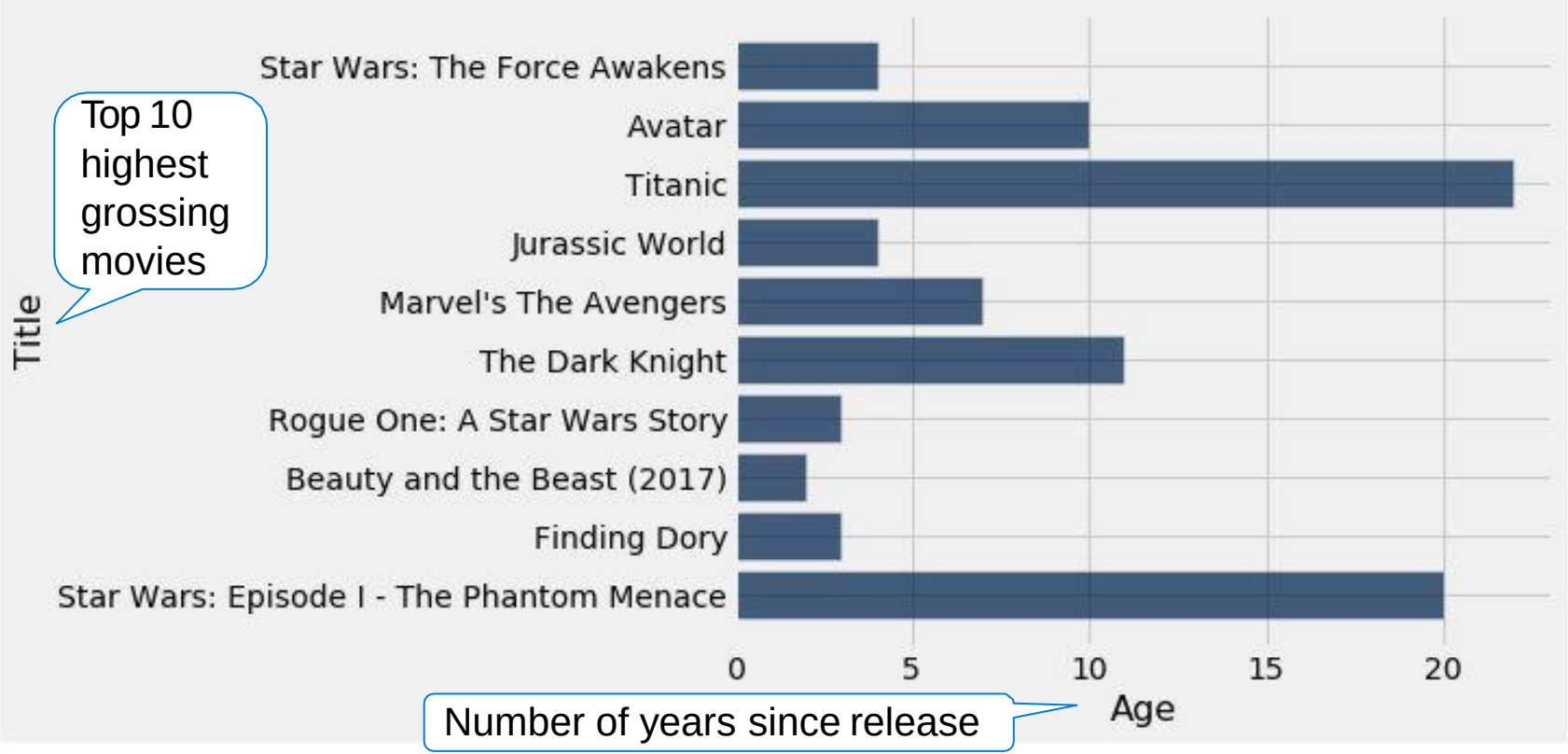


# When to use a line vs scatter plot?

- Use line plots for sequential data: if...
  - ...your x-axis has an order
  - ...sequential differences in y values are meaningful
  - Usually: x-axis is **time** or **distance**

- Use scatter plots for non-sequential data
  - When you’re looking for associations

# How Do You Generate This Chart?



# Distributions!

# Lecture 8 - Overview

- **Distributions** – Numerical Vs. Categorical
- **Bar Graph**
  - display categorical distributions
  - lengths of the bars are the value for each category
- **Histograms**
  - display numerical distributions
  - heights of bars measure densities

# Lecture 8 – Programming Checklist

- **Table.group("Column")** – Count Frequency of each individual returns result as a table
- **Table.barh('categorical','numerical') –** Create a bar graph with 'categorical' on x-axis and numerical on y-axis
- **Table.barh('numerical',bins,unit) –** Create a histogram 'numerical1' on x-axis
- **Table.bin('numerical',bins) –** Count frequency of each item in each bin

# Terminology

- **Individuals**: those whose features are recorded
- **Variable**:  an attribute
- A variable has different **values**
- Values can be **numerical** or **categorical**, and of many sub-types within these
- Each **individual has exactly one value** of the variable
- **Distribution**: For each different value of the variable, the frequency of individuals that have that value

(Demo)

# Distributions of Categorical Variables

# Visualization

- Bar charts are commonly used to visualize categorical distributions
- One axis is categorical, one numerical

(Demo)

# Displaying a Categorical Distribution

- The distribution of a variable (a column, e.g. Studios) describes the frequencies of its different values
- The `group` method counts the number of rows for each value in the column (e.g. the number of top movies released by each studio)
- Bar charts can display the distribution of a categorical variable (e.g. studios):
  - One bar for each category
  - Length of bar is the count of individuals in that category
  - You can choose the order of the bars

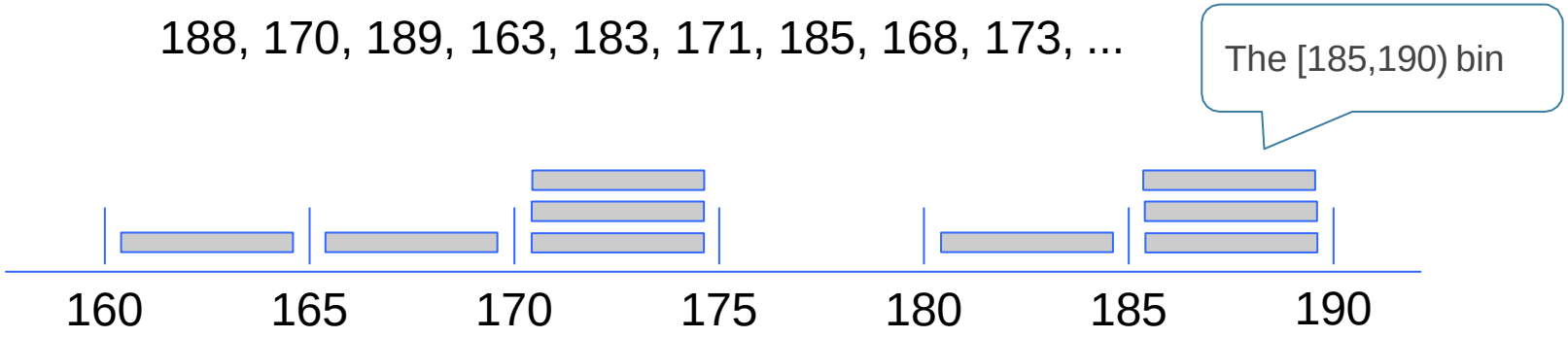
# Distributions of Numerical Variables

# Binning Numerical Values

Binning is counting the number of numerical values that lie within ranges, called bins.

- Bins are defined by their lower bounds (inclusive)
- The upper bound is the lower bound of the next bin

188, 170, 189, 163, 183, 171, 185, 168, 173, ...



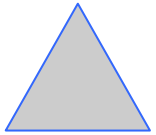

(Demo)

# Area Principle

# Area Principle

**Areas** should be proportional to the values they represent.

For example

- If you represent 20% of a population by 
- Then 40% can be represented by: 
- But not by: 

# Drawing Histograms

# Histogram

- Chart that displays the distribution of a numerical variable
- Uses bins; there is one bar corresponding to each bin
- Uses the area principle:
  - The ***area*** of each bar is the percent of individuals in the corresponding bin

(Demo)

# Density

# Histogram Axes

- By default, `hist` uses a scale (`normed=True`) that ensures the area of the chart sums to 100%
- The **area** of each bar is a percentage of the whole
- The horizontal axis is a number line (e.g., years), and the bins sizes don't have to be equal to each other
- The vertical axis is a rate (e.g., percent per year)

(Demo)

# How to Calculate Height

The [40, 65) bin contains 51 out of 200 movies

- “52 out of 200” is 25.5%
- The bin is 65 - 40 = 25 years wide

$$\text{Height of bar} = \frac{25.5 \text{ percent}}{25 \text{ years}}$$

$$= 1.02 \text{ percent per year}$$

# Height Measures Density

$$\text{Height} = \frac{\text{\% in bin}}{\text{width of bin}}$$

- The height measures the percent of data in the bin ***relative to the amount of space in the bin*.**
- Height measures crowdedness, or **density**.
- Units: percent per unit on the horizontal axis

# Area Measures Percent

**Area of bar = % in bin = Height x width of bin**

- “How many individuals in the bin?” Use area.
- “How crowded is the bin?” Use height.

# Bar Chart or Histogram?

To display a distribution:

## Bar Chart

- Distribution of categorical variable
- Bars have arbitrary (but equal) widths and spacings
- **height (or length)** and **area** of bars proportional to the percent of individuals

## Histogram

- Distribution of numerical variable
- Horizontal axis is numerical: to scale, no gaps, bins can be unequal
- **Area** of bars proportional to the percent of individuals; **height** measures density