AP CSP

Unit 2 – Review

Name: Key

1. Note: A helpful resource to read through as you study is the Data Visualization 101 sheet (see my weebly). You don't need to memorize the whole thing, but it is helpful to recognize, in general, what graphs are more helpful for what types of data. Below is some space to write notes about some graphs we have seen in class: * look at design best practices *

- Bar Chart:
  - good for change over time, to show different categories, or to compare parts of a whole
  - there are vertical, horizontal, and stacked
- Pie Chart:
  - best used for making part-to-whole comparisons with discrete or continuous data
  - better w/ small data sets (no more than 5 categories)

* make sure it adds up to 100%.

- Line Chart:
  - best used for time-series relationships with continuous data
  - don't plot more than 4 lines (gets messy)
- Scatter Plot:
  - show relationships between items based on two sets of variables
  - best used to show correlation in a large amount of data

2. What is the digital divide?

- Access to the Internet differs by income, race, education, age, disability, and geography
- as a result, some groups are over- or under-represented when looking at activity online

3. What is the purpose of a README file?

- a plain-text document that gives some background information about the dataset, how it was collected, and what the column headings mean

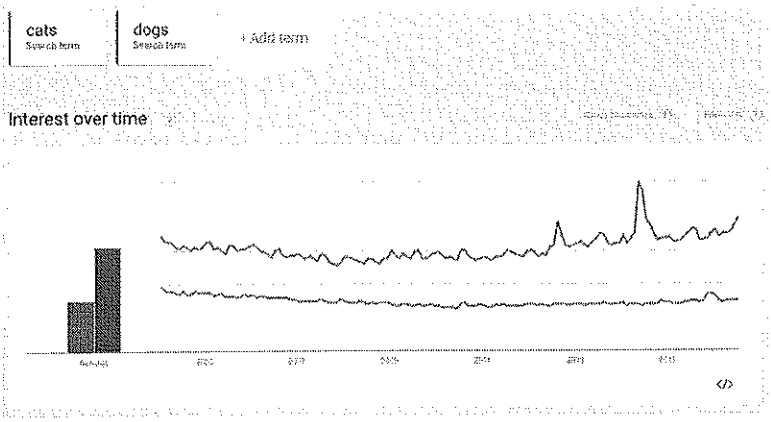
4. Why do we need to "clean" data?

- raw data is often not ready for analysis
- examples: "7" vs. "7 and a half," incorrect inputs, different units of measure

5. What is the purpose of a pivot table?

- pivot table = summary table
- they provide a way to visualize data and see things you might otherwise not see and allow you to manipulate and create new data

6. Below is an image from Google Trends that plots Cats and Dogs. Choose the most accurate description of what this data is actually showing based on what you know about how Google Trends works.



a) People like dogs more than cats

(b) People search for "dogs" more frequently than "cats"

c) There was a sharp increase in the dog population sometime between 2014 and 2015

d) The popularity of dogs as pets is slightly increasing over time, while the popularity of cats is relatively flat

7. Based on the same graph above (from question 1), give a plausible explanation or hypothesis for the spike in dog searches that occurred between 2014 and 2015 that would lead to further investigation or research. Give your explanation and what you would want to investigate next. * many possible answers *

example: maybe a funny dog video went viral

further investigation: use google trends to narrow the explanation down.

8. Which of the following is the most accurate description of what is known as the "digital divide". The digital divide is about how...

a) ... people's access to computing and digital technology increases over time through a process of dividing and growing quickly – it is often likened to the biological processes of cell growth.
(b) ... people's access to computing and the Internet differs based on socioeconomic or geographic characteristics
c) ... people's access to computing technology is affected by the fact that newer devices that use new protocols makes it more difficult for them to communicate with older devices and technology
d) ... the amount of data on the Internet is growing so fast that the amount of computing power and time we have to process it is lagging behind.

9. Which of the following statements are true about pivot tables? Select two answers.

- Pivot tables are used to quickly remove errors and inconsistencies from a dataset
- (o) Pivot tables are used to quickly perform aggregate computations and groupings on a set of raw data
- Pivot tables are used because they automatically detect and highlight potential trends or patterns in the underlying raw data
- (o) Pivot tables are used to generate a summarized view of a large dataset which is helpful for gaining insight.

10. Which of the following is the most accurate statement about cleaning and filtering data?

a) Using computing tools to filter and clean raw data makes it impossible to analyze or draw accurate conclusions
b) Filtering and cleaning data is a fully automated process that should not require human input or intervention
c) Filtering and cleaning data is a human process that does not require the use of computers
d) Filtering and cleaning data is necessary to ensure that data is in a form that is better for computers to process

11. In order to analyze data with a computer, we need to clean the data first. Based on your experience in lesson 13, would you say that data analysis is a perfectly objective process? Why or why not?

Key ideas:

- data cleaning usually requires a human to make decisions about the data
- there often will not be one "right" way to clean the data and different people will do it differently
- any categorizing in particular is quite subjective
- just because it is subjective doesn't make it 'bad' necessarily

12. Consider the following statement from the CS Principles Course Framework (the list of things this course is supposed to cover):

7.4.1C The Global distribution of computing resources raises issues of equity, access, and power.

Briefly describe one of these issues that you learned about in the lesson → on the digital divide (lesson 9) and how it affects your life or the lives of people you know. Keep your response to about 100 words (about 3-5 sentences).

→ many answers here

13. Describe two properties that make a great visualization of data and two properties that make a poor visualization.

great

- easy to compare
- one color to represent each category
- labels (an appropriate amount)

poor

- don't add "chart junk" (unnecessary illustrations, designs)
- don't use more than 6 colors