Lecture 2

Pros of csv

1. Dataset size – csv
2. Human readable
3. Human writable
4. Readable by any programming language

Cons

1. Non-encrypted
2. Jtype or audio file – if there is a csv file with unclosed column containing comma

Code

Read from working directory or url

While reading the dataset it looks at the subset of data and makes a guess about the datatype.

.data – it stores a data dictionary

**Describing Data**

1. Observe characteristics of the sample and extrapolate for the population – statistical inference
2. One observation should be one entire row.

Discrete vs contionous

1. Continuous- even though there is a range of height values but even a decimal value makes a lot of difference hence it is continuous

Classification based on classes

1. Nominal - Car model cannot be ordered.
2. Ordinal variable - Categorical variable which can be ordered
3. Interval variable – temperature varies

**Data Dictonary**

NA vs null.

Factor variable – R stores nominal/ordinal as factor

True is 1 and False is 0, sum will add 1 and 0.

Data issues

1. Median and mean are significantly different
2. Bimodality exists

Measure of shape

1. Skewness

Note: if value of alpha is increased then the confidence interval will becom wider

Confidence interval

1. Alpha
2. Two sided or one sided
3. 0.05 or 0.05/2 = find the prob on chart, 0.025(alpha/2) and .0975(1-alpha/2)

Zvalue computed.

Find – 0.975

url+’/robotics.txt

1. User agent – what machine specs one is using, so that machine knows what to send.
2. Depending whether we have wget as our user agent.