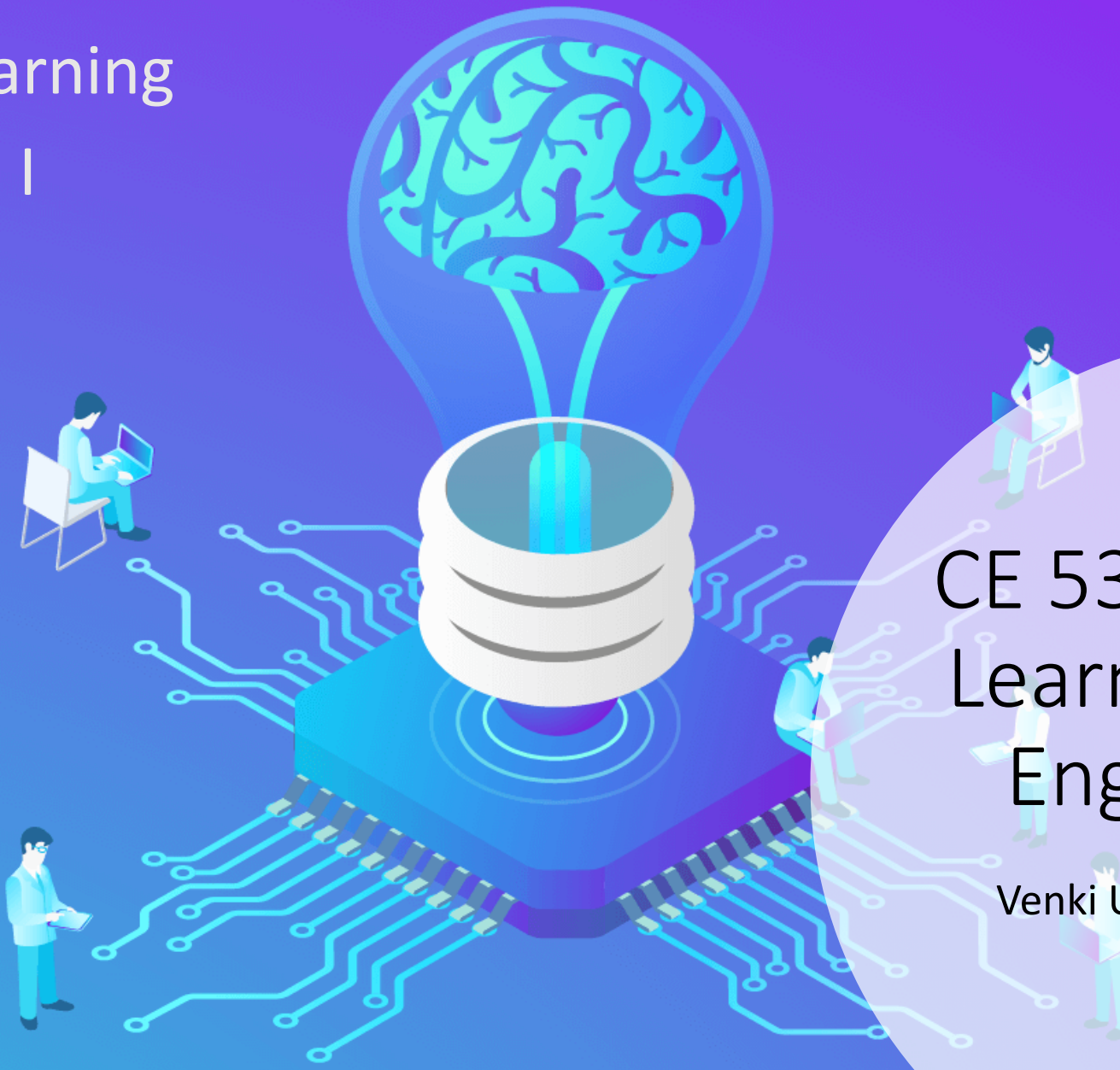
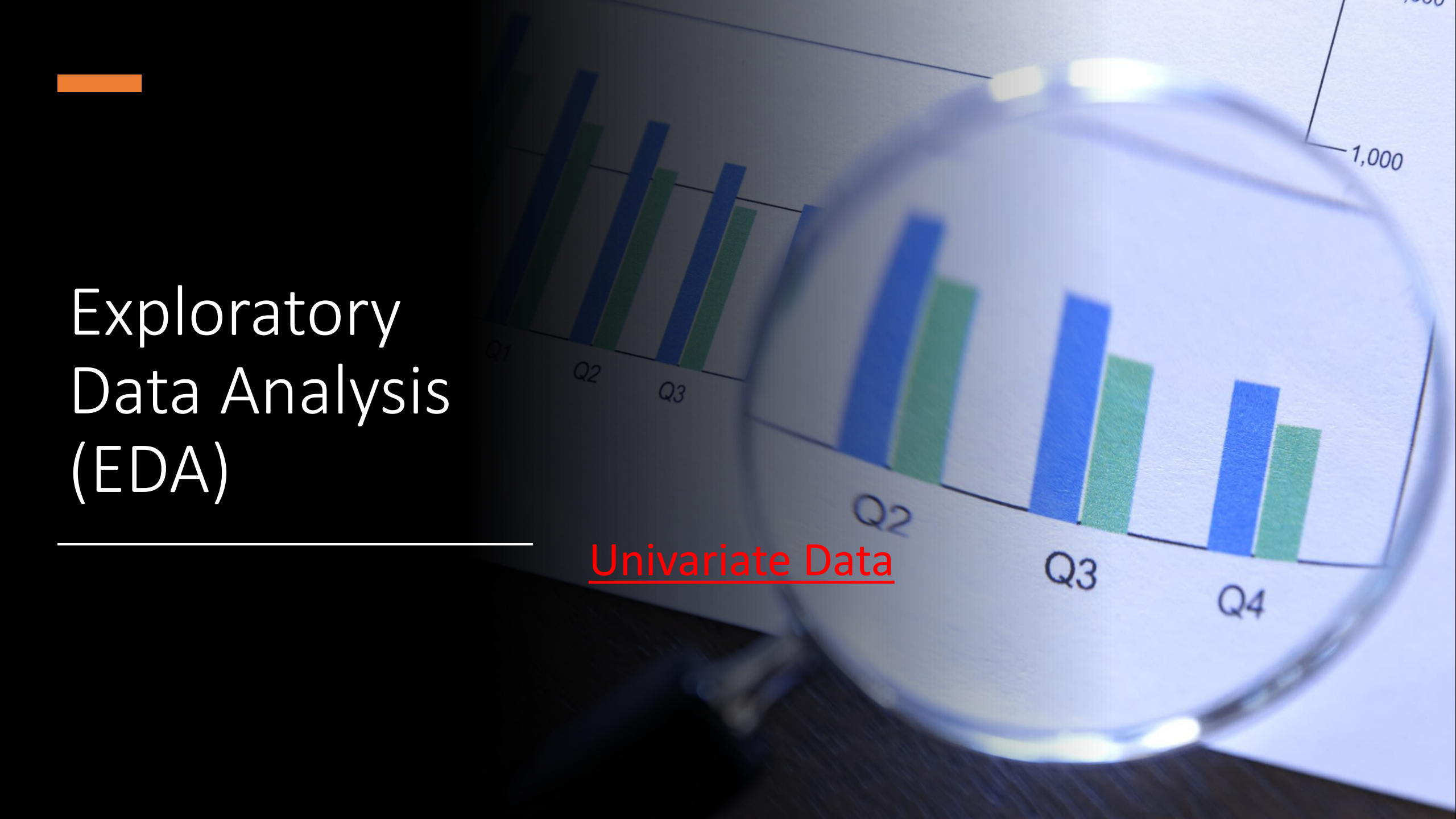
Machine Learning Basics - I



CE 5331 Machine Learning for Civil Engineers

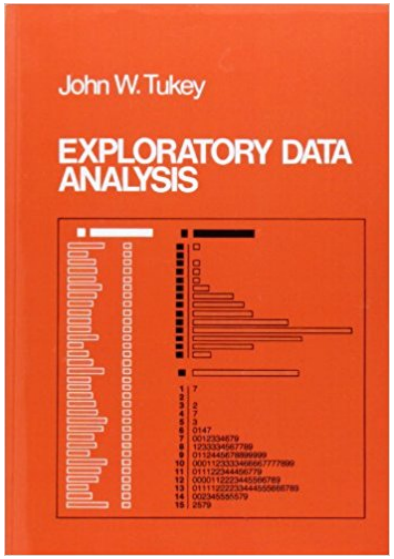
Venki Uddameri, Ph.D. , P.E.

Exploratory Data Analysis (EDA)



Univariate Data

Exploratory Data Analysis (EDA)



- Before performing any advanced calculations, it is important to understand the data
- EDA was advocated by statistician John Tukey (Tukey, 1977)
- EDA seeks to provide initial insights into the data
- EDA uses summary measures and visualization to understand data
  - There are no rigorous analyses carried out
- The objective of EDA is to look for patterns and unique features in the dataset
  - Try to see things that we cannot think about
- How to do EDA is subjective (there are no set rules)
- There are no fixed set of tools or procedures for EDA
  - Knowing some common data summary and visualization tools come in handy

EDA – Summary Measures (Parametric)

- Measures of location
  - Mean or average
  - A measure of central tendency
    - It need not be at the center of the data
- Measures of spread or variability
  - Variance / standard deviation
    - Spread around the mean
    - Measure of dispersion or spread
- Measure of asymmetry
  - Skewness
    - How symmetric is the data distribution

$$\bar{x} = \sum_{i=1}^{N}$$

Mean

$$s^2 = \frac{1}{N}\sum_{n=1}^{N}(x_i - \bar{x})^2$$

Variance

$$s = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_i - \bar{x})^2}$$

Standard Deviation

Coefficient of Variation

$$v = \frac{s}{\bar{x}}$$

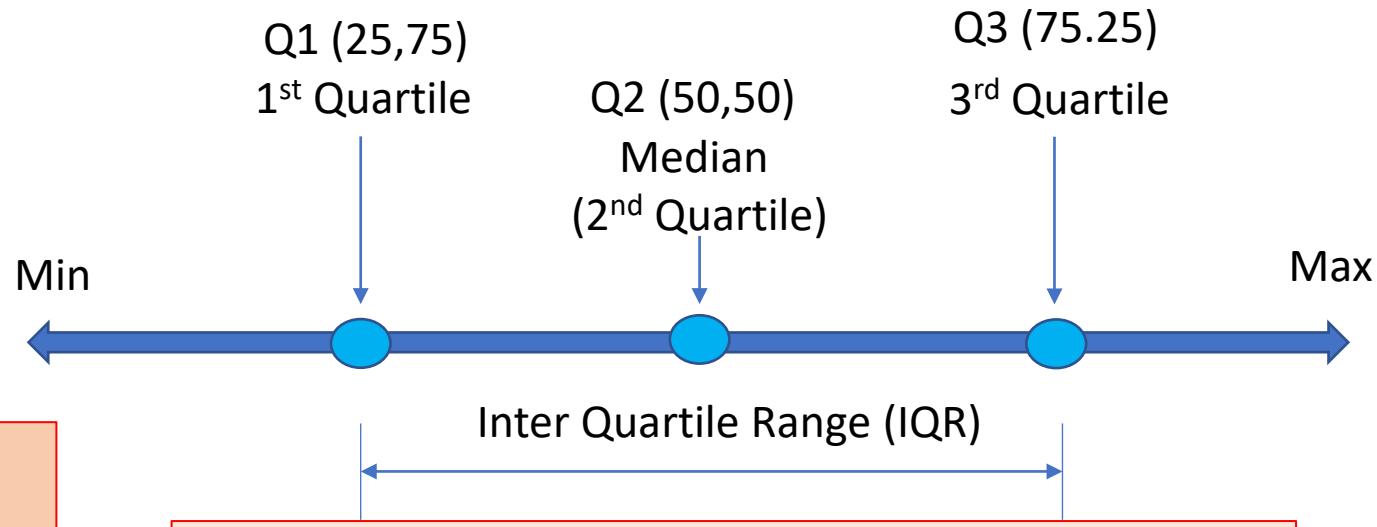
$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{s^3}$$

Skewness

Parameteric summary measures make direct use of the data

EDA – Nonparametric Summary Measures

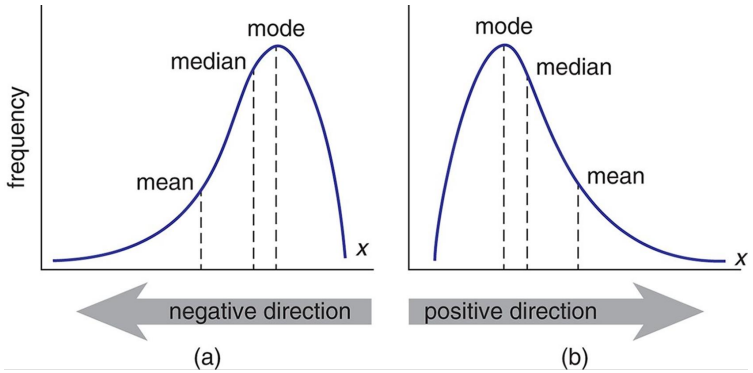
- Central tendency or location
  - Median
- Variability
  - Inter-quartile range
    - Q3 – Q1
  - Range (Max - Min)



Quartiles divide the range of the data into 4 groups.
There are three quartiles Q1, Q2, Q3
25% of the data are ≤ Q1;
50% of the data are ≤ Q2 and 75% of the data are ≤ Q3

Percentiles divide the data into 100 groups
25th percentile is Q1

Mean and Median values can be used to establish skewness in the data



Nonparametric Measures are based on Ranks of the actual data but their values are actual numbers (not ranks)

Calculation of summary measures

- R software provides several functions to calculate the summary statistics
- Basic steps:
  - Read the data from a csv file
  - Call the necessary functions
  - Evaluate the results.

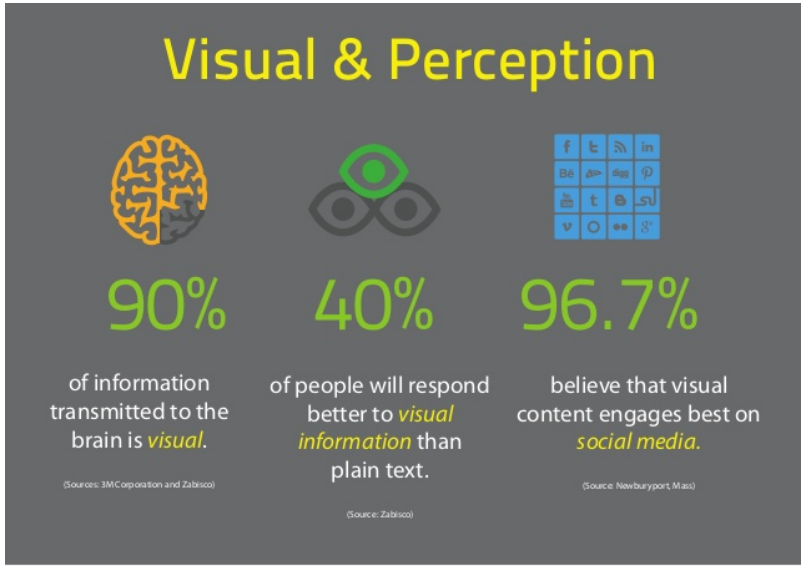
```
# Compute nonparametric summary
shear.med <- median(shear)
shear.iqr <- IQR(shear)
shear.Q1 <- quantile(shear,0.25) # 25th percentile is 1st quartile
shear.Q2 <- quantile(shear,0.5)  # 50th percentile is 2nd quartile
shear.Q3 <- quantile(shear,0.75) # 75th percentile is 3rd quartile

# Write the output to the console
shear.npar <- c(shear.med,shear.iqr,shear.Q1,shear.Q2, shear.Q3)
names(shear.npar) <- c('Median','IQR','Q1','Q2','Q3')
shear.npar

# R can also be used compute mean and 5 point summary in one command
shear.sum <- summary(shear) # Provides a 5 point summary
shear.sum # will write this to the console
```

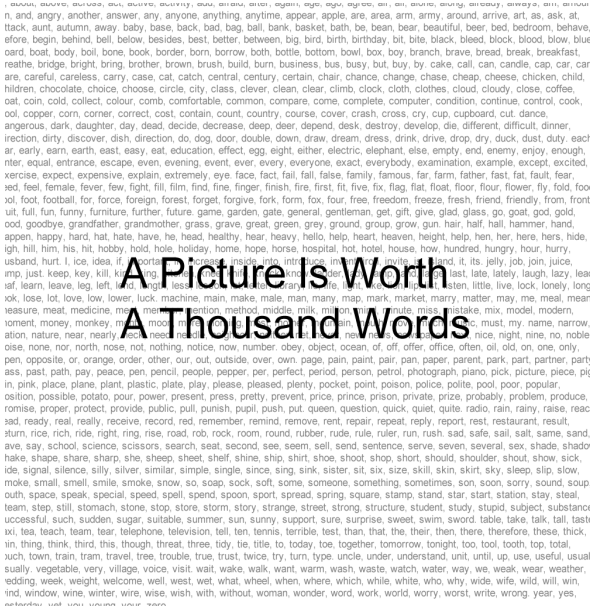
```
# Script to compute summary statistics
# Venki Uddameri, Ph.D., P.E.

# load necessary libraries
library(e1071)

# Set working Directory
setwd('Your Working directory')

# Read the data file
a <- read.csv('Lecture1-sheardata.csv')  # this is a data.frame

# See the first few values in the data frame to make sure
# Everything came in correctly
head(a)

# store the shear data in a separate variable
shear <- a$shear.tsf
head(shear)  # this is a vector

# Compute parametric summary
shear.mean <- mean(shear)
shear.sd <- sd(shear)
shear.cv <- shear.sd/shear.mean
shear.sk <- skewness(shear)  # needs library e1071 loaded

Bind the data and write to the computer console
shear.par <- c(shear.mean,shear.sd,shear.cv,shear.sk)
names(shear.par) <- c('Mean','Std. Dev','Coeff Var','Skewness')
shear.par
```


Data Visualization

Data Visualization

- Data visualization is an important step of EDA
- There are a variety of visualization methods available:
  - Whether the data is univariate (single variable) or multivariate (more than one variable)
  - Whether the data is collected in time
  - Whether the data is circular (seasonal or repetitive)
  - Whether the data is collected in space.





Visualization methods for Univariate Data

- Some basic approaches to visualize univariate data are:
  - Histograms
    - Sensitive to Binning size
  - Cumulative Frequency plot
    - Plots cumulative frequency
  - Box Plot
    - Summarizes the range and quantiles
    - Useful to identify extreme values
      - Extreme values vs. outliers
  - Violin Plot
    - Combines density and box plot into a single graph

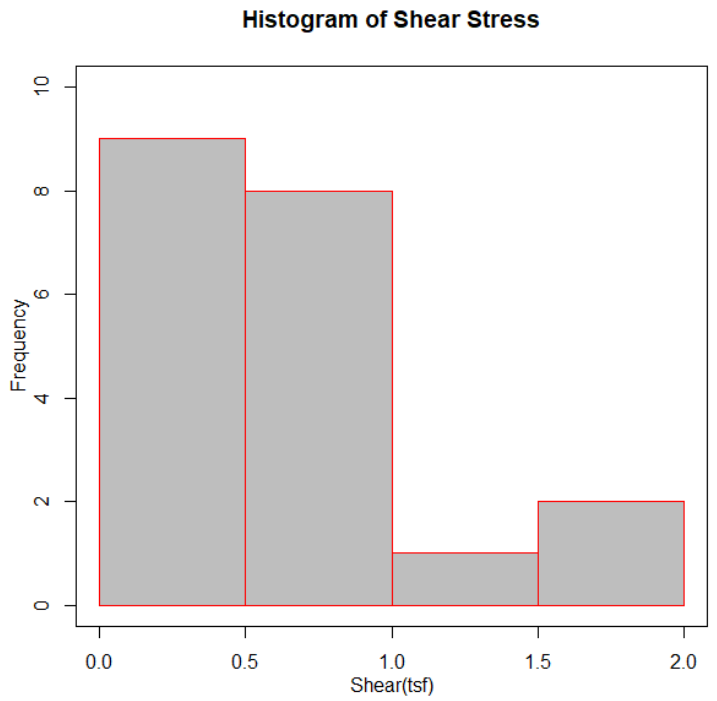
- Empirical Density
  - Smoothed histogram
  - Can be affected by smoothing
  - Alternative to histogram
- Empirical Cumulative Distribution
  - Many plotting position formulas

Histogram

- Plots frequency of occurrence of a given value or bin
- If the variable is categorical it can be used directly
- Small number of discrete values can be plotted directly
- A large range of integer and continuous values need to be binned
- The choice of the binning can affect the shape of the graph
  - Will not affect the data per se but can hinder interpretation

Histogram tells us where the data are clustered
Can help identify gaps in the data

Bins need not be of equal width but commonly assumed to be so
Some trial-and-error is necessary to make a good histogram



Common formulas for estimating Number of Bins (Round up to the next integer)

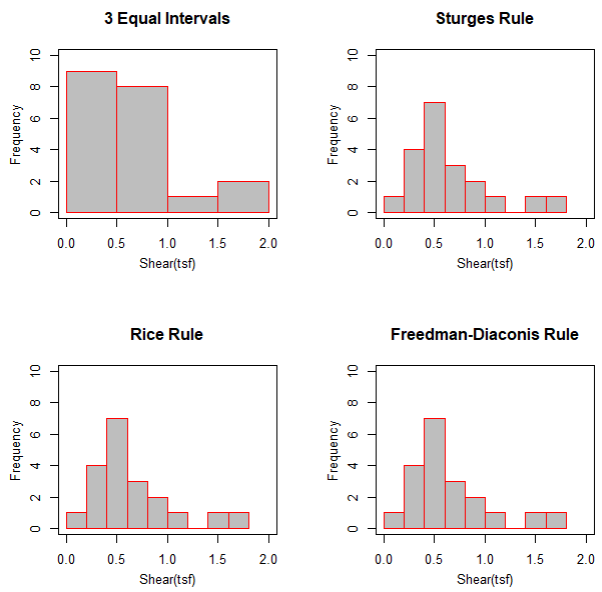
Sturges' formula: $k = log_2[n] + 1$

Rice Rule: $k = 2\sqrt[3]{n}$

Freedman-Diaconis: $k = 2\frac{IQR(x)}{\sqrt[3]{n}}$

Histogram

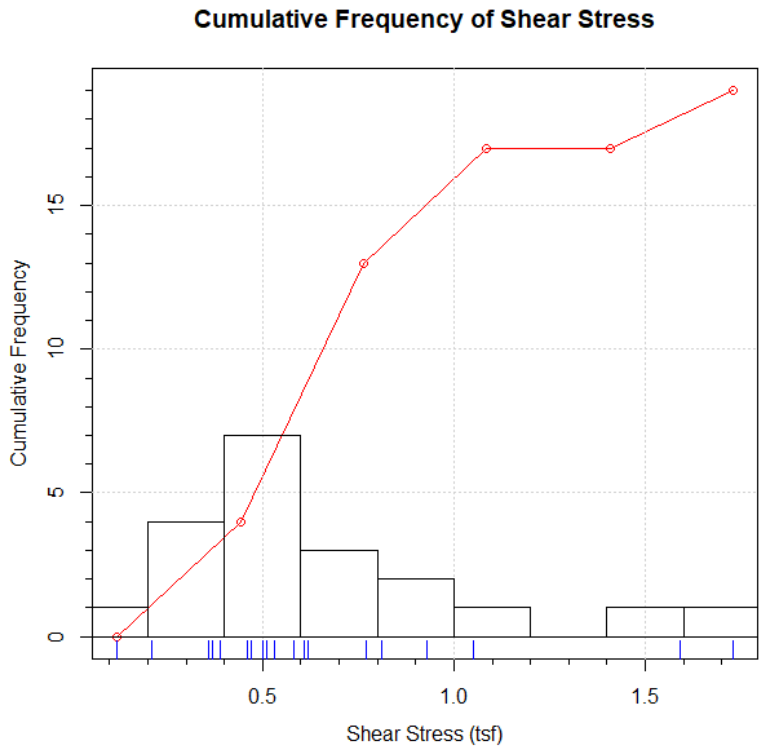
- Create histograms of shear stress data using
  - 3 equal breaks
  - Sturges rule
  - Rice rule
  - Freedman-Diaconis rule
- Plot them on a 2 x 2 graph




```
# Script for univariate plots
# Venki Uddameri, Ph.D., P.E.

# Set working Directory
setwd('Your working directory')

# Read the data file
a <- read.csv('Lecture1-sheardata.csv')  # this is a data.frame
shear <- a$shear.tsf

# Plot various histograms
# Set up a 2 x 2 canvas
par(mgp=c(2.25,1,0))
par(mfrow= c(2,2))

# Histogran with equal intervals
hist(shear,breaks=3,ylim=c(0,10),xlim=c(0,2),xlab='Shear(tsf)',col='grey',border='red',lty=1,main="")
box()
title('3 Equal Intervals')

# Histogran with sturges rule intervals
hist(shear,breaks='sturges',ylim=c(0,10),xlim=c(0,2),xlab='Shear(tsf)',col='grey',border='red',lty=1,main=
"")
box()
title('Sturges Rule')

# Histogran with Rice rule
k <- ceiling(2*(length(shear))^(1/3))
hist(shear,breaks=k,ylim=c(0,10),xlim=c(0,2),xlab='Shear(tsf)',col='grey',border='red',lty=1,main="")
box()
title('Rice Rule')

# Histogram Freedman-Diaconis
k <- nclass.FD(shear)
hist(shear,breaks=k,ylim=c(0,10),xlim=c(0,2),xlab='Shear(tsf)',col='grey',border='red',lty=1,main="")
box()
title('Freedman-Diaconis Rule')
```

Cumulative Frequency Plots

- Plots cumulative frequency plot for the shear-stress data
  - Add a histogram and a data rug to the plot

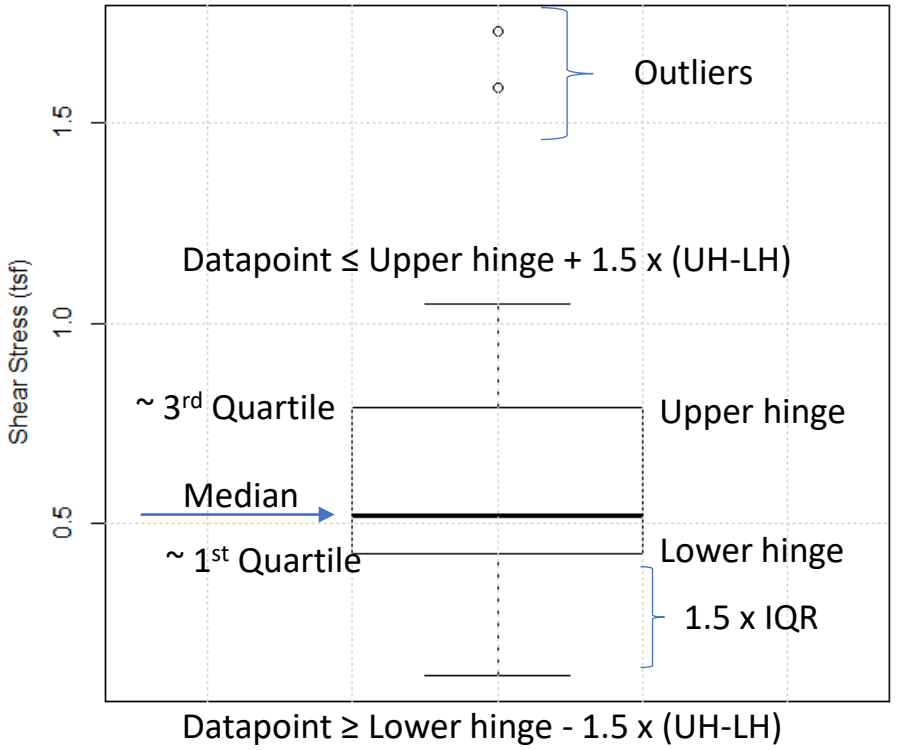



```
# Script for univariate plots
# Venki Uddameri, Ph.D., P.E.

# Load library
library(Hmisc)  # Required for rug plots

# Set working Directory
setwd('Your Working Directory')

# Read the data file
a <- read.csv('Lecture1-sheardata.csv')  # this is a data.frame
shear <- a$shear.tsf

#Step 1: Calculate the break values (6 breaks)
breaks <- seq(min(shear),max(shear),length.out=6)
# Step 2: Arrange the data in bins
shear.bin <- cut(shear,breaks=breaks)
# Step 3: Compute Frequencies
shear.freq <- table(shear.bin)

# Step 4: Calculate Cumulative Frequency
shear.cum <- c(0,cumsum(shear.freq))

# Step 5: Plot cumulative frequency
par(mgp=c(2.5,1,0))
plot(breaks,shear.cum,col='red',xlab='Shear Stress (tsf)',
ylab='Cumulative Frequency') # plot the cumulative frequency
curve
lines(breaks,shear.cum,col='red') # join the points with a line
minor.tick(nx=5,ny=5)
grid()
title('Cumulative Frequency of Shear Stress')

# Step 6: Add a Histogram if necessary
hist(shear,breaks=6,add=T)
# Step 7: Add a data rug
rug(shear,col='blue')
```

Summarizing Variability – Box Plot

- Box Plot summarizes the main characteristics of the data
- Provides a way to identify extreme values (outliers) in the data

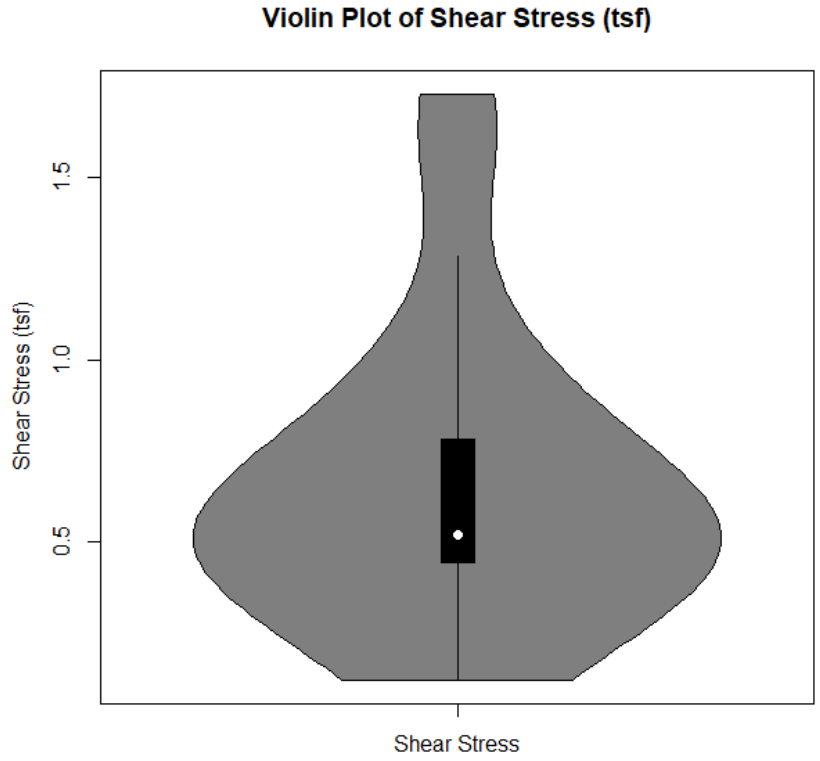
```
# Script for Box plot
# Venki Uddameri, Ph.D., P.E.

# Set working Directory
setwd('Your Working Directory')

# Read the data file
a <- read.csv('Lecture1-sheardata.csv')  # this is a data.frame
shear <- a$shear.tsf

# Plot the box plot
par(mgp=c(2.5,1,0))
boxplot(shear,ylab='Shear Stress (tsf)')

# Boxplot with no outliers plotted
boxplot(shear,ylab='Shear Stress (tsf)',outline=F)
```



Violin Plot

- Violin Plot contains both the density of data and boxplot summaries
  - Densities are plotted on either side of the boxplot for aesthetics
- Can see median in relation to the mode
- Use ‘vioplot’ package in R to plot it

```
# Script for Violin plot
# Venki Uddameri, Ph.D., P.E.

# Load library
library(vioplot)

# Set working Directory
setwd('Your Working Directory')

# Read the data file
a <- read.csv('Lecture1-sheardata.csv')  # this is a data.frame
shear <- a$shear.tsf
vioplot(shear,names='Shear Stress',ylab='Shear Stress (tsf)')
grid()
title('Violin Plot of Shear Stress (tsf)')
```



Key concepts

- First steps towards Understanding Data
  - Data collection
  - Data visualization
- Exploratory data analysis (EDA)
  - Summary statistics (parametric and nonparametric)
  - Histogram and Cumulative Histogram
  - Box plots
  - Violin Plots
- How to use R for exploratory data analysis