CS544 Module1

Suresh Kalathur

- Course Outline
  - Module1
    - Review basics in statistics
    - R - Data types and structures
  - Module2
    - Probability, Conditional Probability
    - R – Programming constructs
  - Module3
    - Data – Univariate, Bivariate, Multivariate
    - Data Visualization (Base R, Plotly, ggplot2)
  - Module4
    - Distributions – Discrete, Continuous
    - Modeling Random variables
  - Module5
    - Central Limit Theorem, Sampling,
    - RMarkdown and Dashboards
  - Module6
    - Strings, Regular Expressions, Data Wrangling

Grading

- Programming assignments – 20%
  - Six (One for each module)
- Quizzes – 30%
  - First 4 modules only
- Group Project – 20%
  - Ready to start after Module3
- Final Exam – 30%

Lecture 1 - Statistics

- Measures of Central Tendency
  - Mean, Median, Mode
- Measures of Variation
  - Range
  - Variance, Standard deviation
  - Quartiles
  - Inter-quartile range (IQR)

- Percentiles
  - Divide data into 100 equal parts
  - https://dqydj.com/household-income-percentile-calculator/
- Quartiles
  - Divide data into 4 equal parts
    - Q1 – bottom 25% from the top 75%
    - Q2 – bottom 50% from the top 50% (Median)
    - Q3 – bottom 75% from the top 25%

- IQR – Inter Quartile Range
    - Q3 - Q1
- Five Number Summary
    - Min, Q1, Q2, Q3, Max
- Variations in each quarter
    - Q1 - Min, Q2 - Q1, Q3 - Q2, Max - Q3
- Outliers
    - Outside the range
        - (Q1 - 1.5*IQR, Q3 + 1.5*IQR)
        - (Mean - 3*SD, Mean + 3*SD)

- Population versus Sample
- Standardized Variables
  - Mean 0 and Standard Deviation 1
  - *z*-score for variables
    - Negative score – below the mean
      - How many SD below the mean
    - Positive score – above the mean
      - How many SD above the mean
  - Most values in the range -3 to 3
    - Otherwise, outliers

Z-Scores Application

Suppose we are analyzing the dataset with income and age attributes.
Consider a subset of three people, say, A, B, and C, from the dataset.

Person	**Income**	**Age**
A	90000	50
B	80000	40
C	100000	20

For ML applications which involve similarity comparison (clustering, etc.), a common measure used is the distance metric. From geometry, the distance between two points (x1,y1) and (x2,y2) is

$$\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

Now, let us compute the distance metric between any pair of people in the dataset:

Pair	**Distance**
(A,B)	10000
(A,C)	10000.04
(B,C)	20000.01

If you notice, the value of the *age* is not playing any role if the calculation and the *income* with large values is entirely dominating in the calculation. **If the person A is fixed, B and C are considered as equal when compared with A, i.e., (A,B) and (A,C) have the same distance measure.**

Now, let us assume that the *income* attribute for the entire dataset has a mean of 60000 and standard deviation of 23000.

Similarly, let us assume that the *age* attribute for the entire dataset has a mean of 45 and standard deviation of 15.

The z-scores of the *income* and *age* attributes calculated with these means and standard deviations are as follows:

Person	**Income**	**Age**
A	1.3043	0.3333
B	0.8696	-0.3333
C	1.7391	-1.6667

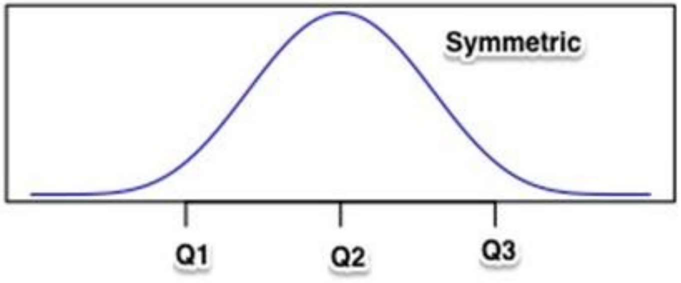
Now, let us re-compute the distance metric between any pair of people in the dataset using the z-scores:

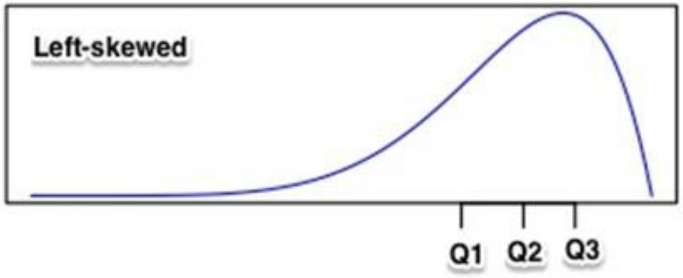
Pair	**Distance**
(A,B)	0.8
(A,C)	2.0
(B,C)	1.6

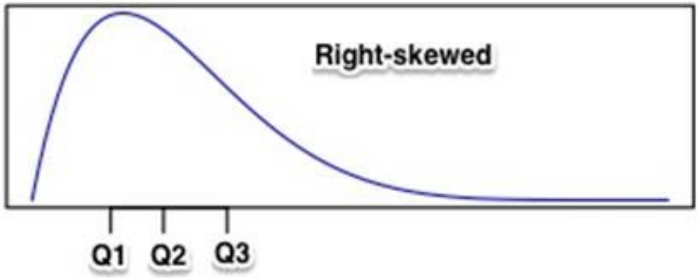
With these calculations**, if the person A is fixed, B and C are much different when compared with A.**

Shape of Data

- Distribution of the data
  - Symmetric
    - Mean and median are the same
    - 32, 41, 50, 52, 56, 60, 64, 68, 70, 79, 88
    - Mean: 60, Median: 60
  - Left-skewed (negatively skewed)
    - An easy quiz/exam
    - Mean is less than the median
    - 12, 15, 80, 81, 84, 85, 86, 87, 88, 91, 94
    - Mean: 73, Median: 85
  - Right-skewed (positively skewed)
    - A hard quiz/exam
    - Mean is greater than the median
    - 41, 45, 48, 50, 51, 54, 57, 60, 94, 96, 97
    - Mean: 63, Median: 54





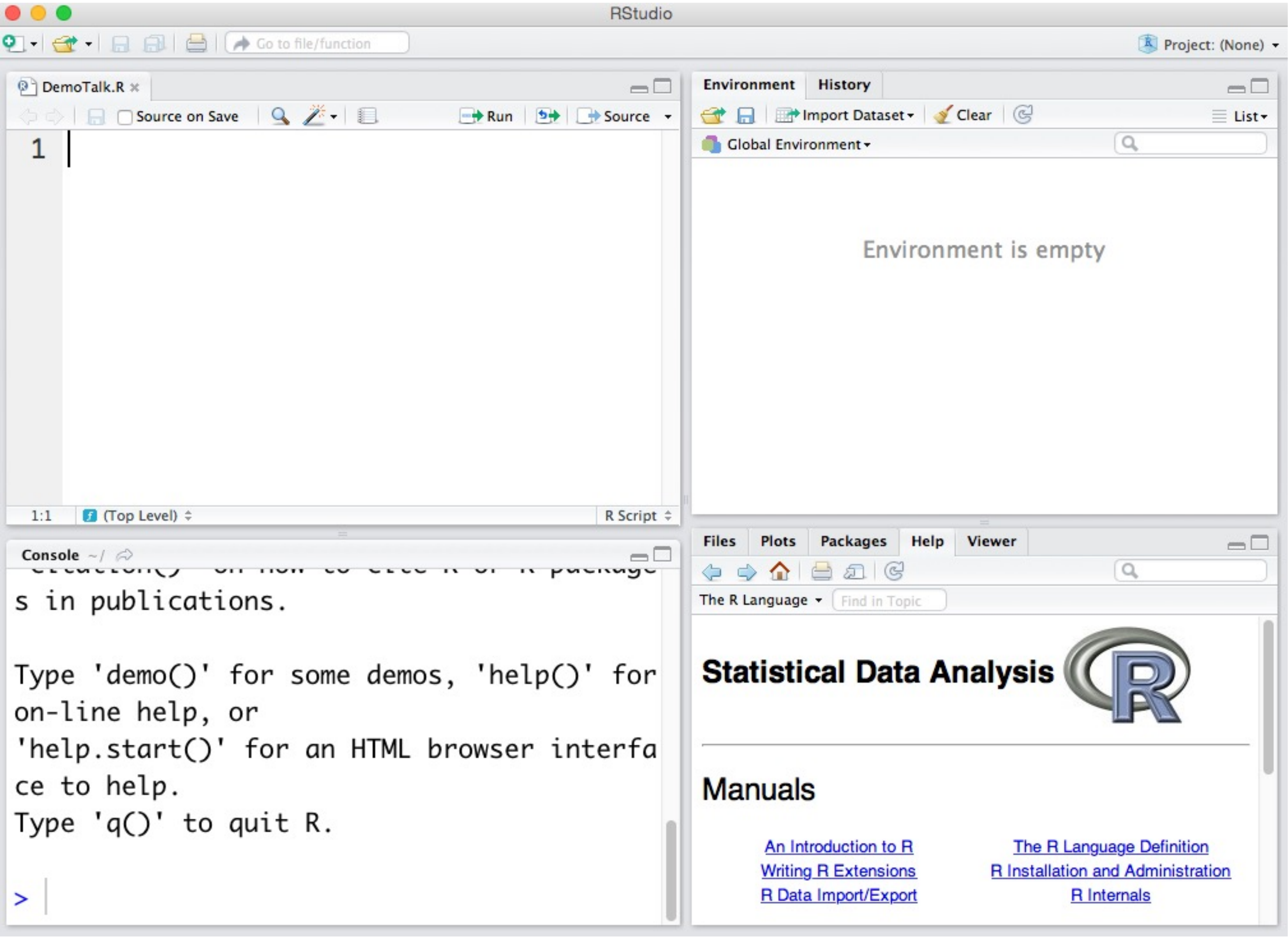


R

- A language and environment for statistical computing and graphics
- GNU General Public License
- Initially written by Robert Gentleman and Ross Ihaka (University of Auckland)
- http://www.r-project.org
- Base version of R
- Rstudio

R

- Statistical techniques
  - Linear and nonlinear modeling
  - Classical statistical tests
  - Time-series analysis
  - Classification
  - Clustering, …
- Graphical techniques



Environment is empty

```
citation()' on how to cite R or R package
s in publications.

Type 'demo()' for some demos, 'help()' for
on-line help, or
'help.start()' for an HTML browser interfa
ce to help.
Type 'q()' to quit R.

>
```

Statistical Data Analysis

Manuals

An Introduction to R
Writing R Extensions
R Data Import/Export

The R Language Definition
R Installation and Administration
R Internals

Data in R

- *Data types frequently used in R*
  - numeric
  - integer
  - logical
  - character
  - complex

...Data in R

- *Data structures*
  - ***vector*** – a collection of values of the same type
  - *factor* – a collection of values from a fixed set of possible values
  - ***matrix*** – a two-dimensional collection of values of the same type
  - *list* – a collection of any of the data structures
  - ***data frame*** – a collection of vectors all of the same length