wbs
WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

For the Change Makers

Dr. Iman Ahmadi
Assistant Prof. of Marketing

Marketing & Strategy Analytics:

Exploratory Data Analysis II

Visualizing Numeric Variables: Univariate

Visualizing Data in R: CRM Dataset

- **Load a new dataset on CRM dataset**

("CRM" R code)

```
cust.df <- read.csv("http://goo.gl/PmPkaG")
str(cust.df)
```

```
'data.frame': 1000 obs. of 12 variables:
$ cust.id : int 1 2 3 4 5 6 7 8 9 10 ...
$ age : num 22.9 28 35.9 30.5 38.7 ...
$ credit.score : num 631 749 733 830 734 ...
$ email : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 1 1 ...
$ distance.to.store: num 2.58 48.18 1.29 5.25 25.04 ...
$ online.visits : int 20 121 39 1 35 1 1 48 0 14 ...
$ online.trans : int 3 39 14 0 11 1 1 13 0 6 ...
$ online.spend : num 58.4 756.9 250.3 0 204.7 ...
$ store.trans : int 4 0 0 2 0 0 2 4 0 3 ...
$ store.spend : num 140.3 0 0 95.9 0 ...
$ sat.service : int 3 3 NA 4 1 NA 3 2 4 3 ...
$ sat.selection : int 3 3 NA 2 1 NA 3 3 2 2 ...
```

Chapman, C. and Feit, E.M. (2015) *R for Marketing Research and Analytics*. Springer. Chapter 4.

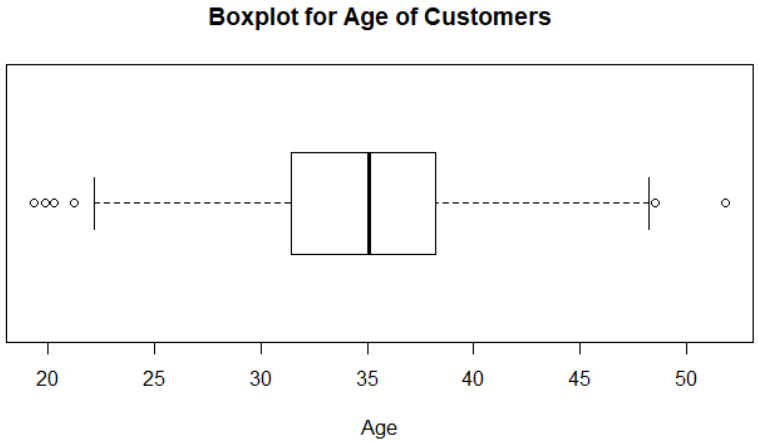
Creating Boxplots in R

- Basic boxplot is good for data exploration:

```
summary(cust.df$age)
```

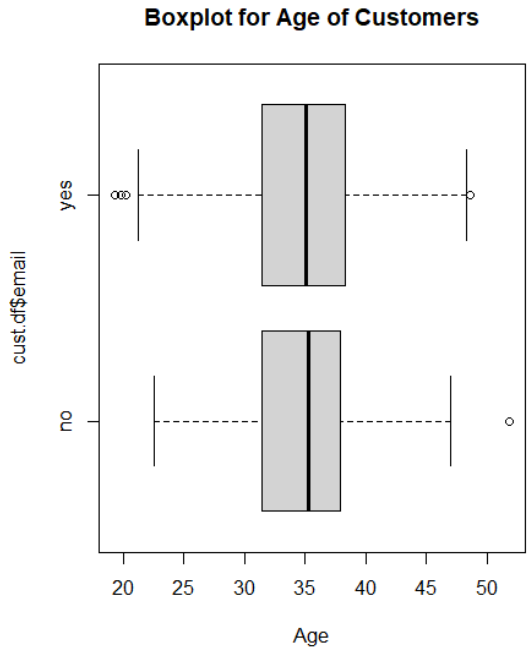
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.34   31.43   35.10   34.92   38.20   51.86
```

```
boxplot(cust.df$age, xlab="Age",
      main="Boxplot for Age of Customers", horizontal=TRUE)
```



Creating Boxplots in R

```
boxplot(cust.df$age ~ cust.df$email, xlab="Age",
      main="Boxplot for Age of Customers", horizontal=TRUE)     #boxplot by email
```



Measuring Spread – Variance

- The normal distribution, which describes many types of real-world data, can be defined with:
  - **mean**: the center of normal distribution
  - **spread**: is measured by a statistic called the **standard deviation**
- In order to calculate the standard deviation, we must first obtain the **variance**

$$Var(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

n: sample size
i: index for observation (i = 1, …., n)
$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$

(the average of the squared differences between each value and the mean value)

Lantz, B. (2015) *Machine Learning with R (Second edition)*. Birmingham: Packt Publishing. Chapter 2.

Measuring Spread –Standard Deviation

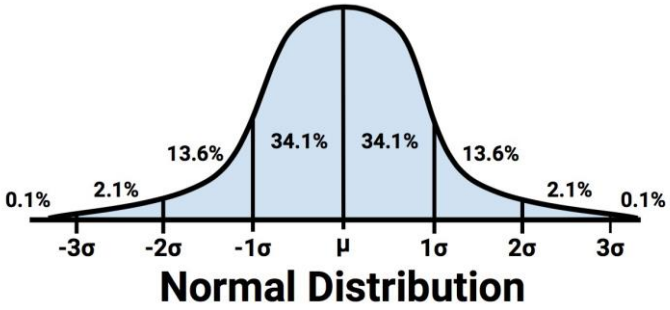
- The standard deviation is the square root of the variance:

$$SD(x) = \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

n: sample size
i: index for observation (i = 1, …., n)
$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$

- σ can be used to quickly estimate how extreme a given value is under the assumption that it came from a normal distribution

- The **68–95–99.7** rule:



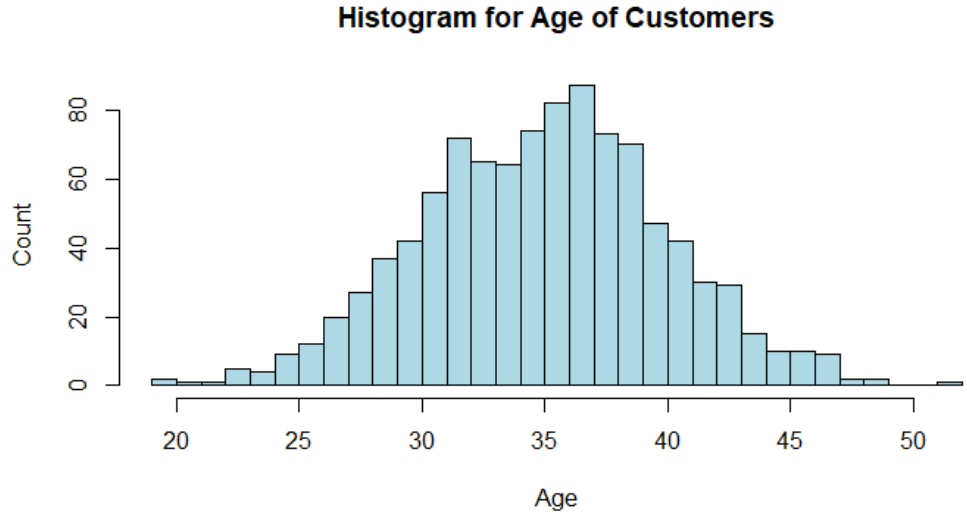
(Note: $\mu = \bar{x}$)

Lantz, B. (2015) *Machine Learning with R (Second edition)*. Birmingham: Packt Publishing. Chapter 2.

Towards Histograms in R (II/III)

- Make it more granular and colorful

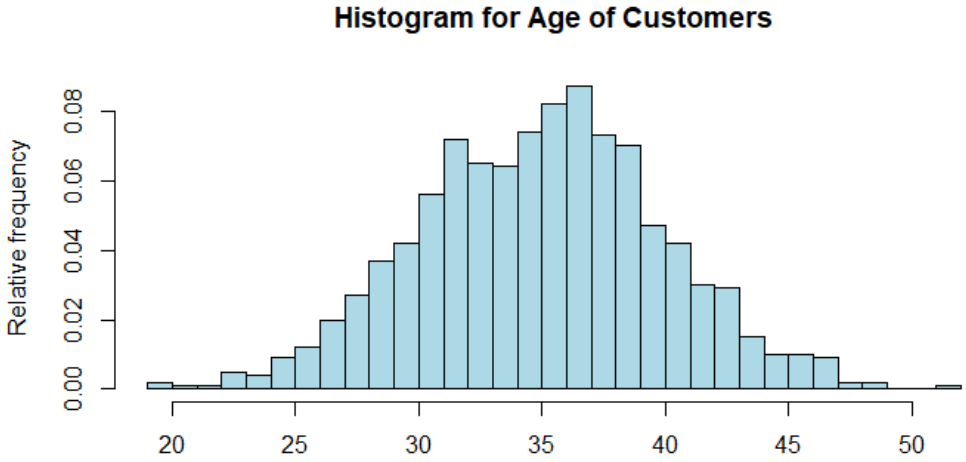
```
hist(cust.df$age,
    main="Histogram for Age of Customers",
    xlab="Age",
    ylab="Count",
    breaks=30,          # more columns
    col="lightblue")    # color the bars
```



Towards Histograms in R (III/III)

- Change counts to proportions (i.e., %)

```
hist(cust.df$age,
    main="Histogram for Age of Customers",
    xlab="Age",
    ylab="Relative frequency",
    breaks=30,
    col="lightblue",
    freq=FALSE)        # freq=FALSE for density
```

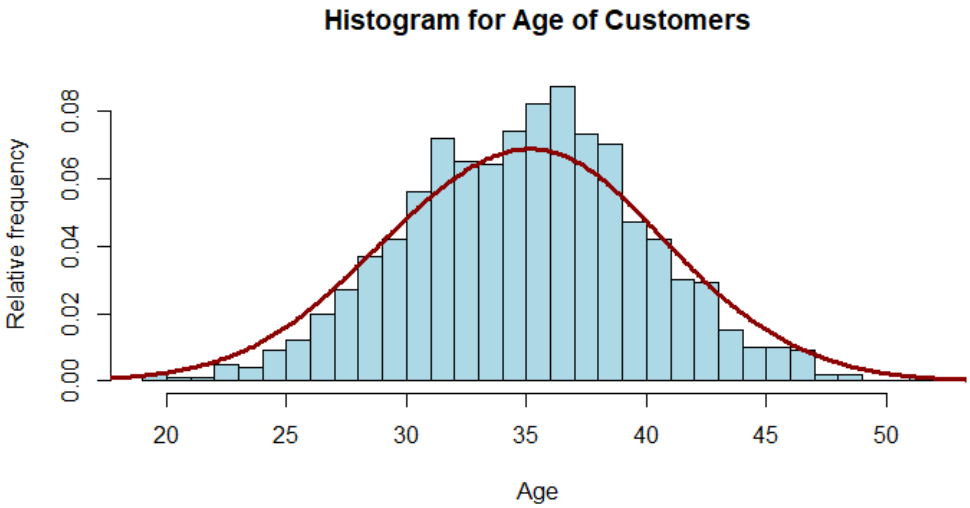


Towards Histograms in R (IV/IV)

- Add density curve

```
lines(density(store.df$age, bw=3),  # bw = smoothing
    type="l", col="darkred", lwd=3)       # lwd = line width
```

("CRM" R code)



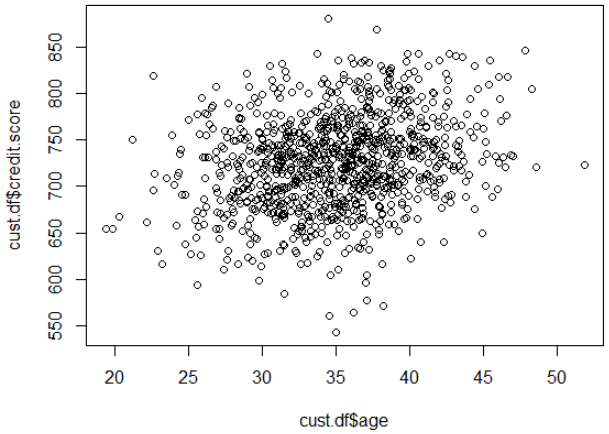
- A **density curve** is a graph that shows probability. The area under the density curve is equal to 100 percent of all probabilities.

Visualizing Numeric Variables: Bivariate

Scatter Plots (I/II)

- **Scatterplot** is a diagram that visualizes a bivariate relationship
- Patterns in the placement of observations (usually by dots) reveal the **underlying associations** between the two variables
- How does age relate to credit score?

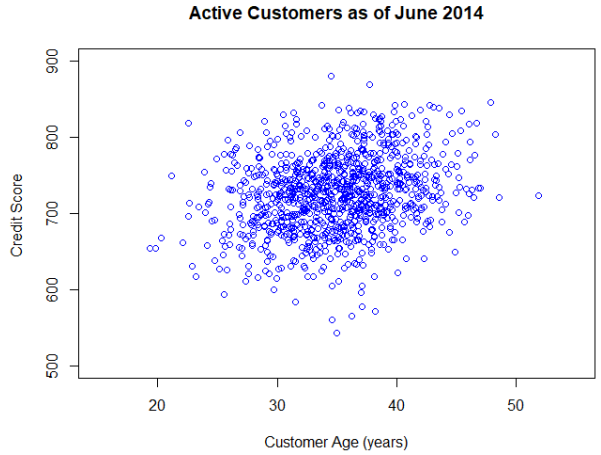
```
plot(x=cust.df$age, y=cust.df$credit.score)
```



Scatter Plots (II/II)

- Add color, labels, and adjust the axis limits:

```
plot(cust.df$age, cust.df$credit.score,
    col="blue",
    xlim=c(15, 55), ylim=c(500, 900),
    main="Active Customers as of June 2014",
    xlab="Customer Age (years)", ylab="Credit Score")
```



Linear Associations: Correlation Analysis (III/IV)

$-1 \leq r \leq 1$, **where:**

- positive values denote positive linear correlation;
- negative values denote negative linear correlation;
- value of 0 denotes no linear correlation;
- the closer the value is to 1 or –1, the stronger the linear correlation

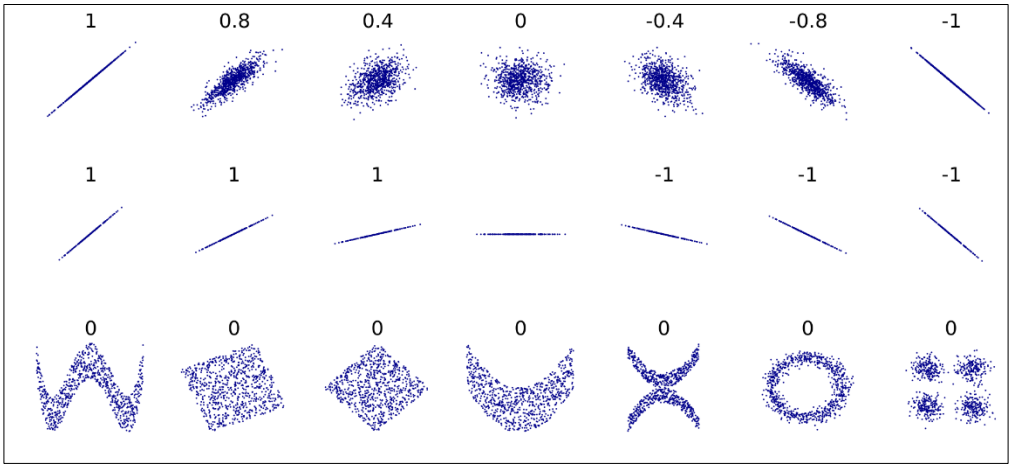
Rule of Thumb for Interpreting the Size of a Correlation Coefficient

Size of Correlation (r)	**Interpretation**
0.90 to 1.00 (–0.90 to –1.00)	Very high positive (negative) correlation
0.70 to 0.90 (–0.70 to –0.90)	High positive (negative) correlation
0.50 to 0.70 (–0.50 to –0.70)	Moderate positive (negative) correlation
0.30 to 0.50 (–0.30 to –0.50)	Low positive (negative) correlation
0.00 to 0.30 (0.00 to –0.30)	Little (if any) correlation

Linear Associations: Correlation Analysis (IV/IV)

$-1 \leq r \leq 1$, where:

- positive values denote positive linear correlation;
- negative values denote negative linear correlation;
- value of 0 denotes no linear correlation;
- the closer the value is to 1 or –1, the stronger the linear correlation



Examples of correlation coefficient

Source of figure: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Correlation in R

- A basic inferential test of Pearson's r can be done with cor.test():

```
cor.test(cust.df$age, cust.df$credit.score)
```

```
        Pearson's product-moment correlation

data: cust.df$age and cust.df$credit.score
t = 8.3138, df = 998, p-value = 4.441e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1955974 0.3115816
sample estimates:
      cor
0.2545045
```

t statistic

degrees of freedom

p value $(= 4.441 \times 10^{-16} = 0.0000000 < 0.05)$

Pearson's correlation

- Age is associated with credit score here, $r = 0.25, p < 0.05$

Correlation Matrix

- Correlation matrix

```
cor(cust.df[, c(2, 3, 5:12)])      # only numeric cols
```

```
                          age credit.score distance.to.store
age               1.000000000  0.254504457        0.00198741
credit.score      0.254504457  1.000000000       -0.02326418
distance.to.store 0.001987410 -0.023264183        1.00000000
online.visits    -0.061381070 -0.010818272       -0.01460036
online.trans     -0.063019935 -0.005018400       -0.01955166
online.spend     -0.060685729 -0.006079881       -0.02040533
store.trans       0.024229708  0.040424158       -0.27673229
store.spend       0.003841953  0.042298123       -0.24149487
sat.service                NA           NA                NA
sat.selection              NA           NA                NA
```

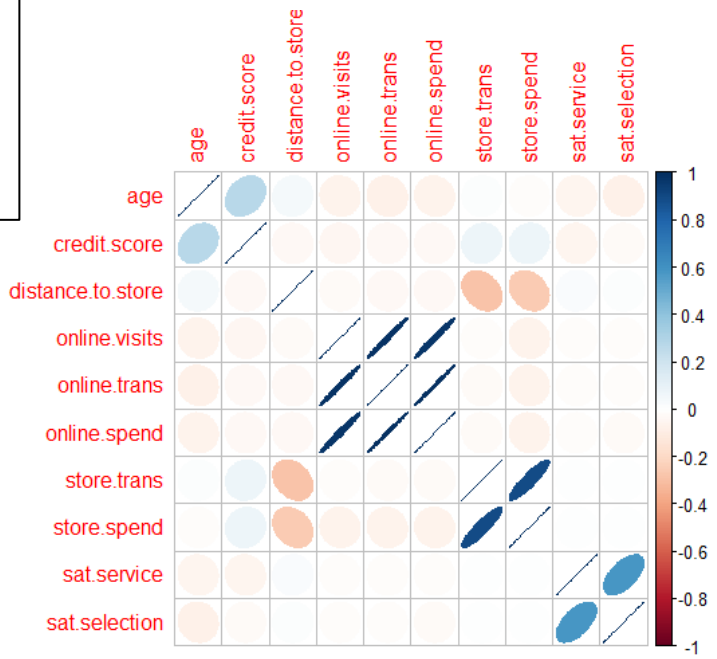
- Note: the full output is not shown here
- You can also get correlations for complete cases only with '**use="complete.obs"**'

```
cor(cust.df[, c(2, 3, 5:12)], use="complete.obs")
```

Visualize Correlation Matrix

- Use the “corrplot” package. See book (and help) for more options.

```
install.packages("corrplot")
library(corrplot)
corrplot(corr=cor(cust.df[ , c(2, 3, 5:12)],
                  use="complete.obs"),
         method ="ellipse")
```



Exercise 3.2 – Salaries for Professors

Access the Salaries data set:

```
install.packages("car")  # search how you can install a library in RStudio
library(car)
data(Salaries)
```

1. Create descriptive statistics for each of the variables. What does the dataset seem to contain?
2. Plot salary vs. years since PhD.
3. What is the correlation for salary vs. years since PhD? … vs. years of service? Are they statistically significant? Explain.
4. Draw a visualization of all bivariate relationships.

Note: make sure that you fully interpret your output and results.

("Salary PhD" R code)

Exercise 3.3 – Salaries for Professors

Access the StoreData data set:

1. What does the dataset seem to contain? What do you notice?
2. Draw a boxplot and histogram for sales of product 1 and 2.
3. Compare the sales of product 1 when it is on promotion and when it is not.
4. What is the correlation for sales of product 1 vs. price of product 1? … vs. sales of product 2? … vs. price of product 2? Are they statistically significant? Explain.
5. Draw a visualization of all bivariate relationships.

Note: make sure that you fully interpret your output and results.

("StoreData_Exercise_3.3_Correlation" R code)

Thank You!

Iman.Ahmadi@wbs.ac.uk
Room No.: 3.207