vidyākośā
La Place For Your Transform

# Data Visualization

***Dr. Arulalan Rajan,***

*Founder & Director, vidyākośā, Bangalore.*

*Faculty, Proficience Programme, CCE, IISc, Bangalore.*

*Adjunct Professor, Dept of AIML, RVCE, Bangalore.*

January 1, 2024

# Topics to be covered[1]

- Exploring the data distribution
  - Percentiles and boxplots
  - Frequency tables and histograms
  - Density Plots and Estimates
- Exploring Binary and Categorical Data
  - Mode
  - Expected Value
  - Probability
  - Correlation
  - Scatterplots
- Exploring two or more variables
  - Hexagonal Binning and Contours
  - Two categorical variables
  - Categorical and Numeric Data
  - Visualizing multiple variables

[1]The Instructor acknowledges authors of various articles available on the web and other resources from which some of the materials presented here are taken

# Exploring Data Distribution

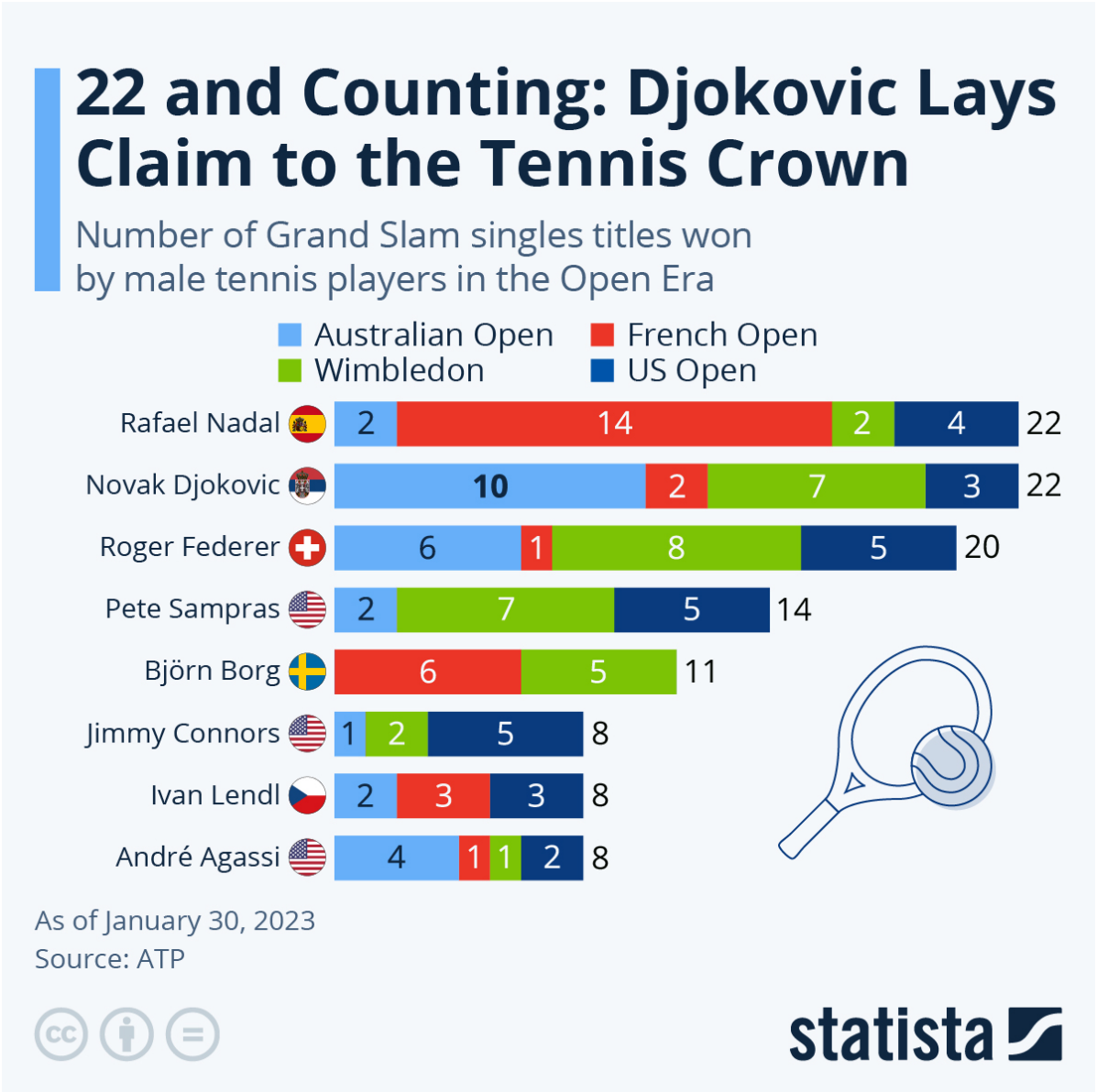
## Here is some interesting Statistics - Tennis this time, folks!!!!

Table: Grand Slam Titles

| Rafa | Djoko | Fedex | Pete | Borg | Connors | Lendl | Agassi |
|---|---|---|---|---|---|---|---|
| 22 | 22 | 20 | 14 | 11 | 8 | 8 | 8 |

# Grand Slam Court Surface

# Finer details of Grand Slam Titles



# Data Visualization

- Distribution of data
- Better presentation of data
- Inclusion of finer details
- Drawing conclusions and inferences!
- Analyze
  - Identify Patterns, Trends, etc.,
  - Formulate/Test Hypothesis
- Provides evidence and support!

# Frequency Table

- Tally of the count of numeric data values falling into a set of intervals/bins
- Divides up the variable range into equally spaced segments or individual data set
- Tells us how many values fall within each segment!

# Frequency Table - Examples!

| Letter | Percentage |
|---|---|
| a | 8.2 |
| b | 1.5 |
| c | 2.8 |
| d | 4.3 |
| e | 12.7 |
| f | 2.2 |
| g | 2.0 |
| h | 6.1 |
| i | 7.0 |
| j | 0.2 |
| k | 0.8 |
| l | 4.0 |
| m | 2.4 |

| Letter | Percentage |
|---|---|
| n | 6.7 |
| o | 7.5 |
| p | 1.9 |
| q | 0.1 |
| r | 6.0 |
| s | 6.3 |
| t | 9.1 |
| u | 2.8 |
| v | 1.0 |
| w | 2.4 |
| x | 0.2 |
| y | 2.0 |
| z | 0.1 |

Based on passages taken from newspapers and novels, and total sample of 100,362 alphabetic characters. *Beker and Piper, 1982*

# Frequency Table - Examples!

| Grades | A+ | A | B+ | B | C | D | F | Drop-outs | Total |
|---|---|---|---|---|---|---|---|---|---|
| 10 Points | 85+ | 75-84 | 65-74 | 55-64 | 45-54 | 35-44 | < 35 | | |
| Total | 3 | 3 | 5 | 5 | 2 | - | - | 5 | 23 |

Figure: Grades in a course on Statistics

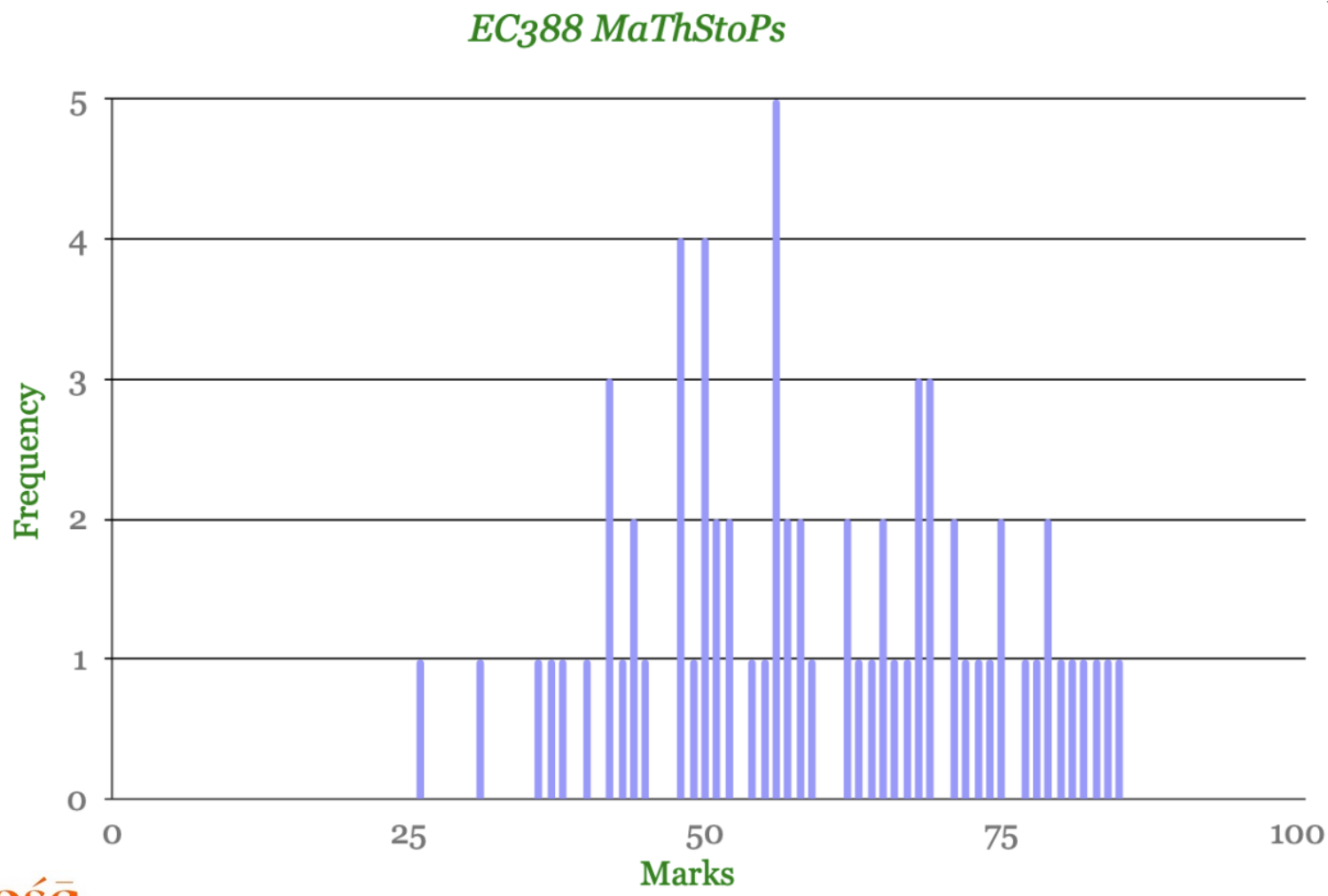
- Observe the bin size
- Bin size - too large - Consequence?
- Bin size - too small -Consequence?

# Types of Frequency Distribution

- Grouped Frequency Distribution - Divide the observations between different intervals (*class intervals* and count the frequencies for each class interval - Large Data set
- Ungrouped Frequency Distribution - List all distinct observations and count individually - Small Data set
- Relative Frequency Distribution - Displays proportion of observations in different intervals; Relative Frequency - *You know this already!!!*
- Cumulative Frequency Distribution - $\leq$ (sum all the frequencies before the current interval or $\geq$ (sum all the frequencies after the current interval)

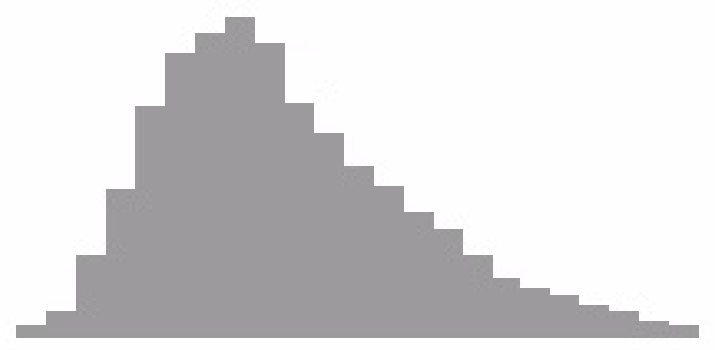
# Histogram

- Visualize a frequency table
- Bins on the $x$-axis and data count on the $y$-axis



# Histogram

Normal distribution



Right-skewed distribution

Bimodal (double-peaked) distribution

Plateau distribution

Skewness and Kurtosis - Discovered through visual displays of data

*Courtesy: https://asq.org/quality-resources/histogram*

# Boxplots

- Percentiles measure spread of data
- Quartiles and Deciles
- Boxplots - Based on percentiles
- Distribution of data based on a five number summary - Minimum score, First Quartile, Median, Third Quartile and Maximum score
- Can also tell about outliers and their values

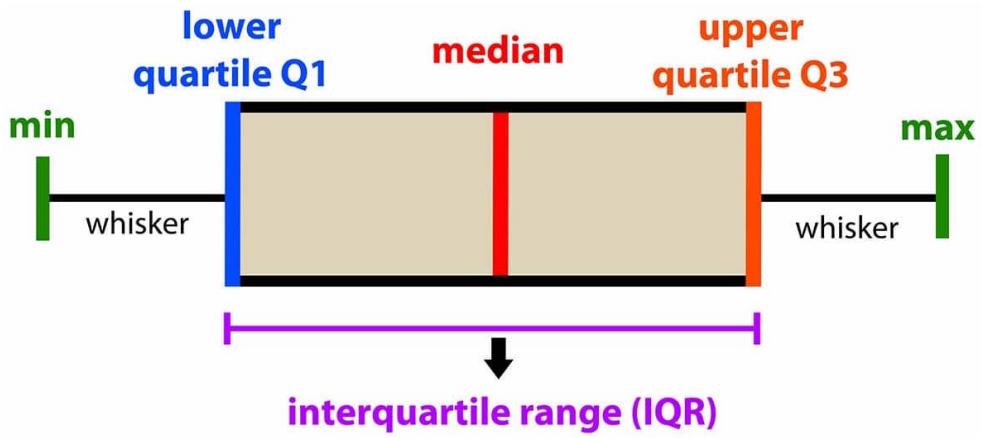



*Image Courtesy:* https://www.simplypsychology.org/wp-content/uploads/box-whisker-plot.jpg

# Boxplots

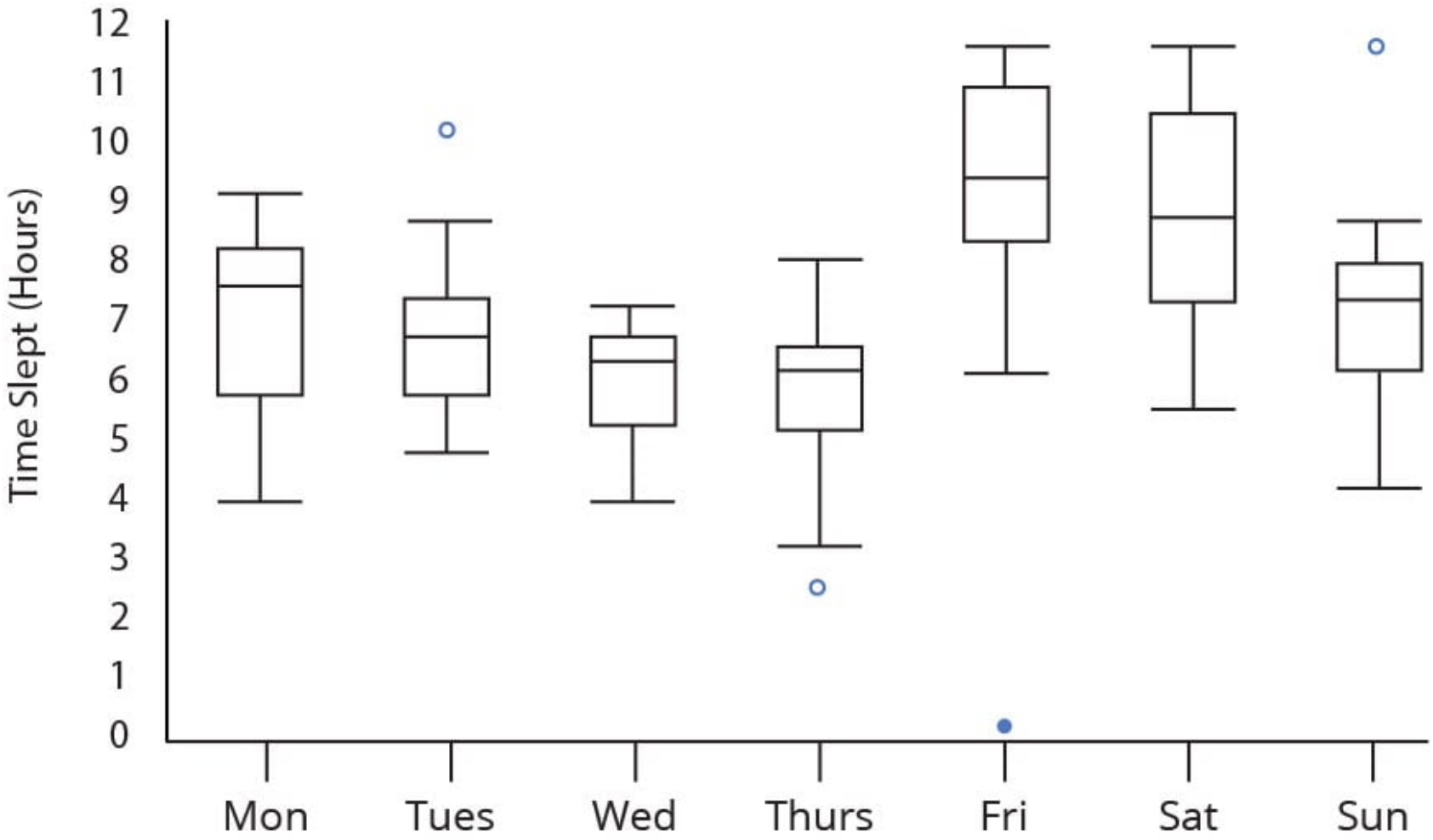



*Image Source:* https://www.simplypsychology.org/boxplots.html

# Boxplots

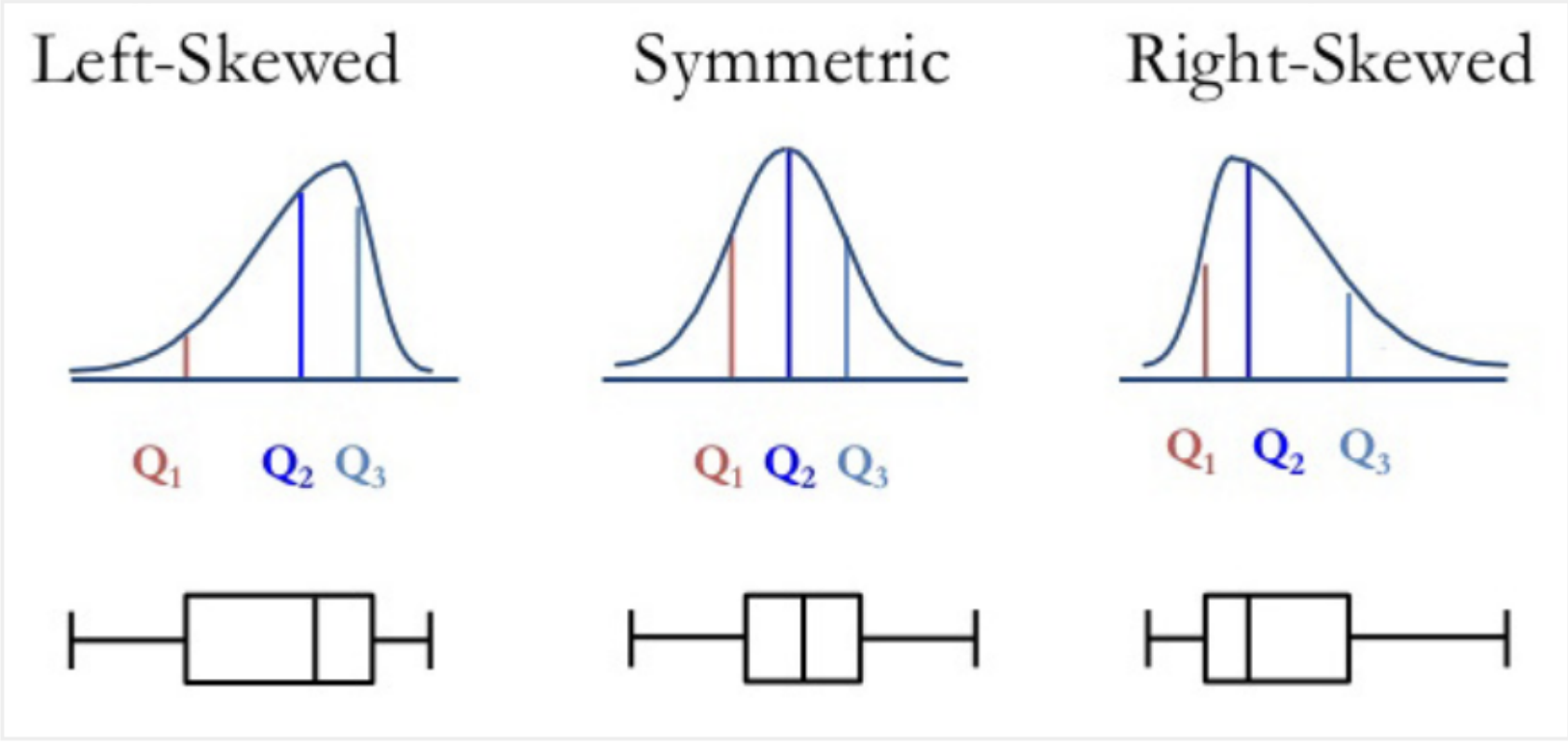



*Image Source:* https://www.simplypsychology.org/boxplots.html

# Density plots and Estimates

- Distribution of data values as a continuous line
- Kind of smoothened histogram
- Recall: The total area under the density curve is 1
- $y$- axis of histogram - Frequency; $y$- axis of density plot - Proportion!
- Area under the density curve between any two values $x_1$ and $x_2$ - Proportion of the distribution lying between the two points!

# Histogram and Density Plots

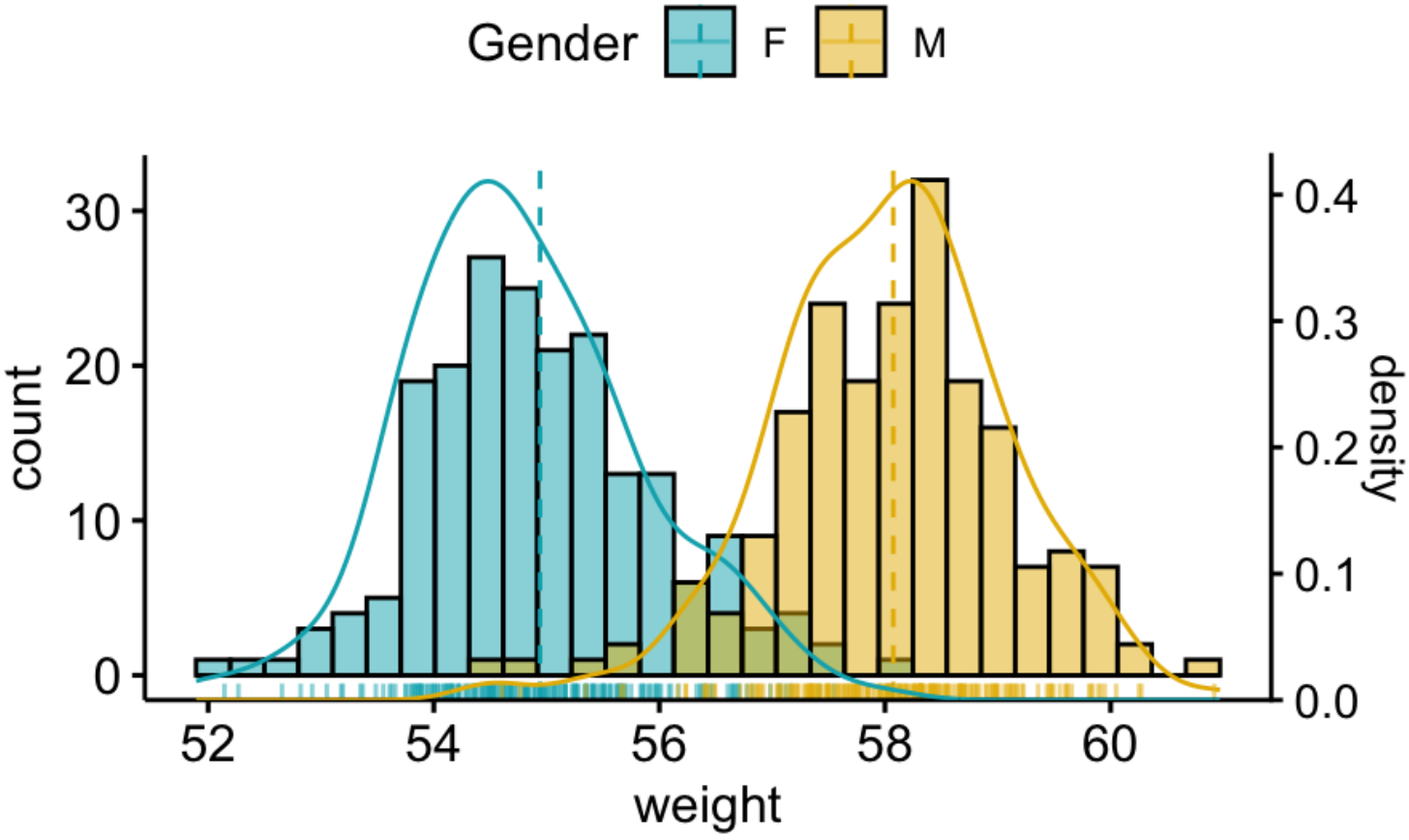



Figure: Density of weights across 400 samples, 200 F + 200 M

# Exploring Binary and Categorical Data

- Takes on fixed and limited number of possible values
- Example: Musical Instruments, Blood Group type, Trains type etc.,
- Logical order and categorical order are not the same
- However, sorting uses logical order
- Allows assignment of categories, but cannot let you order the categories

# Terms for Exploring Categorical Data

## Mode

Most commonly occurring category or value in a dataset

## Expected Value

Associating the categories with numerical values, $E[X]$ gives the average based on the probability of occurrence of the category

$$E[X] = \sum_{x=1}^{n} p(x)x$$

where $x$ is the category and $p(x)$ is the probability of occurrence of the category

# Terms for Exploring Categorical Data - Bar Chart

Frequency or Proportion for each category plotted as bars

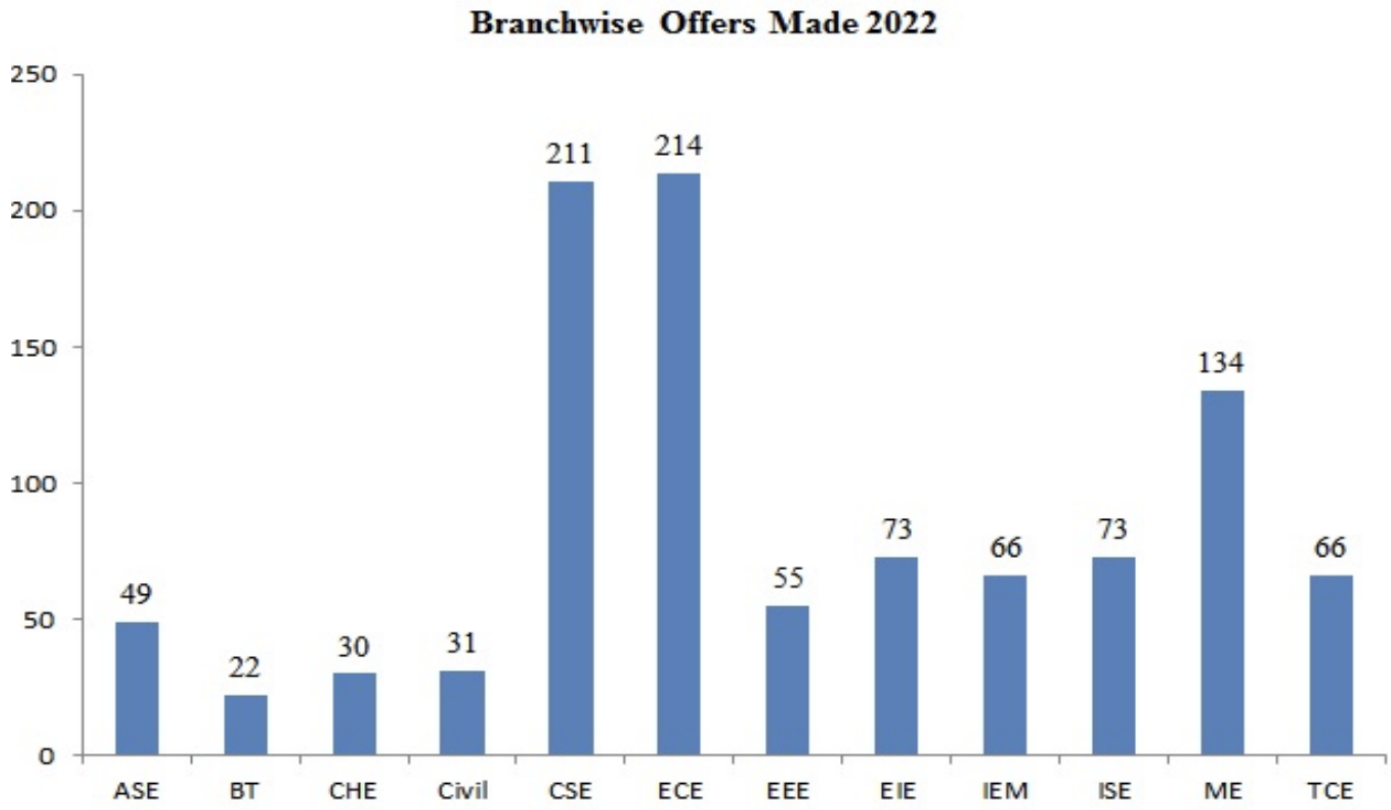



Figure: RVCE Placement Bar Chart[2]

[2]Courtesy: RVCE Placement Webpage

# Terms for Exploring Categorical Data - Pie Chart

Frequency or Proportion for each category plotted as wedges in a pie!

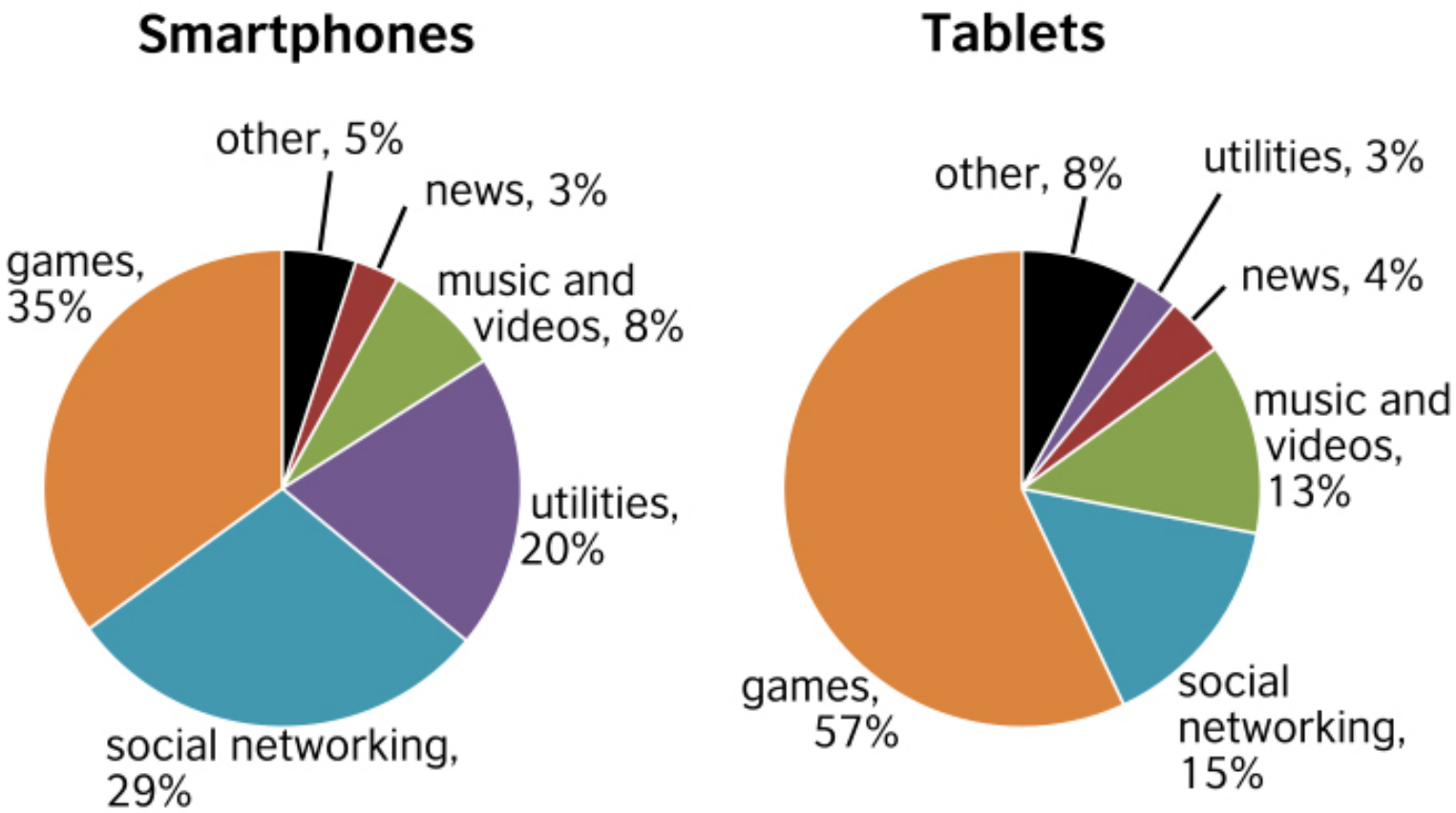



Figure: Time Spent on Smartphones and Tablets[3]

[3] Courtesy: https://learnenglish.britishcouncil.org/skills/writing/b2-writing/comparing-two-charts

# Terms for Exploring Categorical Data - Bar Chart

Frequency or Proportion for each category plotted as bars

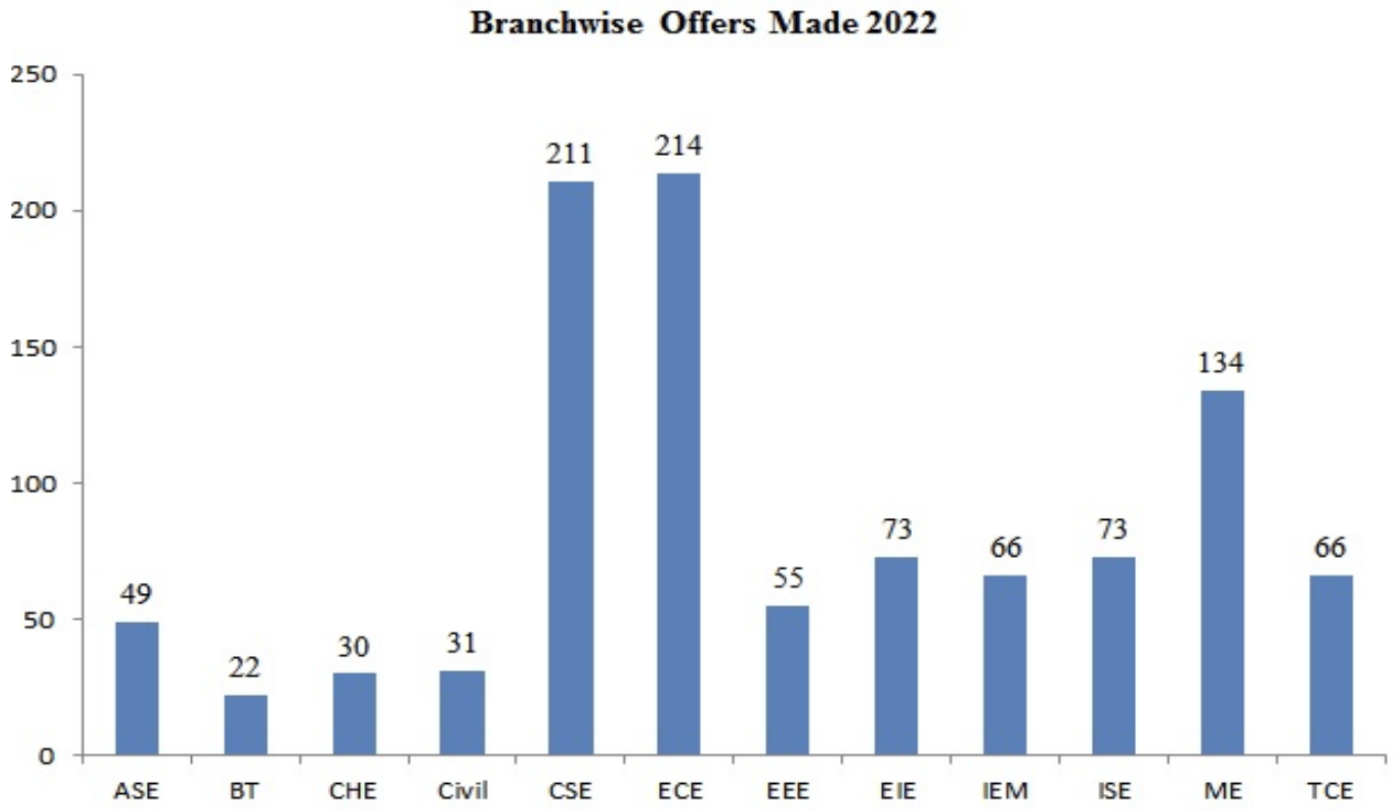



Figure: RVCE Placement Bar Chart[4]

[4]Courtesy: RVCE Placement Webpage

# Terms for Exploring Categorical Data - Probability

Already done in class!!! So u guys escape!!!!


# Terms for Exploring Categorical Data - Correlation

- Statistical relationship between two entities
- Extent to which the variables are linearly related
- Positively Correlated
- Negatively corelated
- Correlation Coefficient : $r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}$
- $-1 \leq r \leq 1$,
- $r = 0$ indicates no correlation
- Non linear relation cannot be captured by $r$

# Terms for Exploring Categorical Data - Scatter Plot

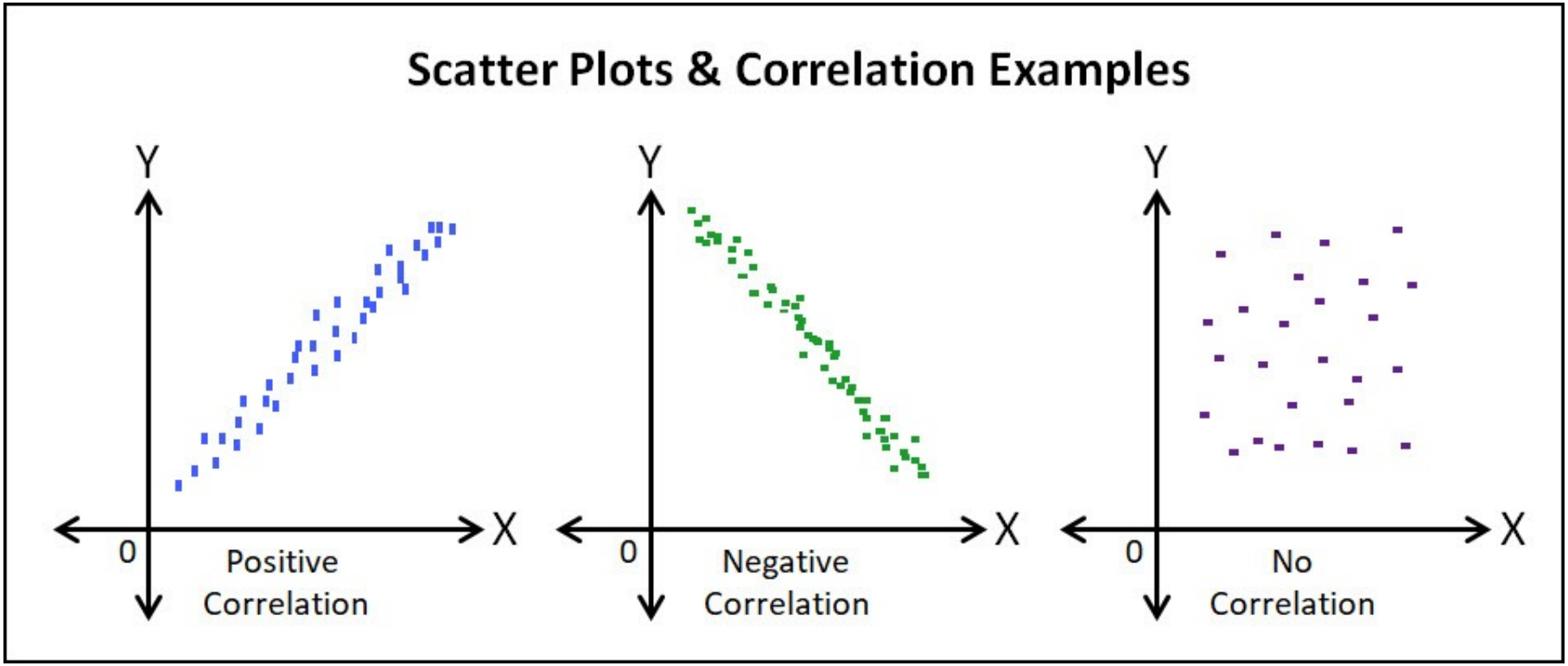
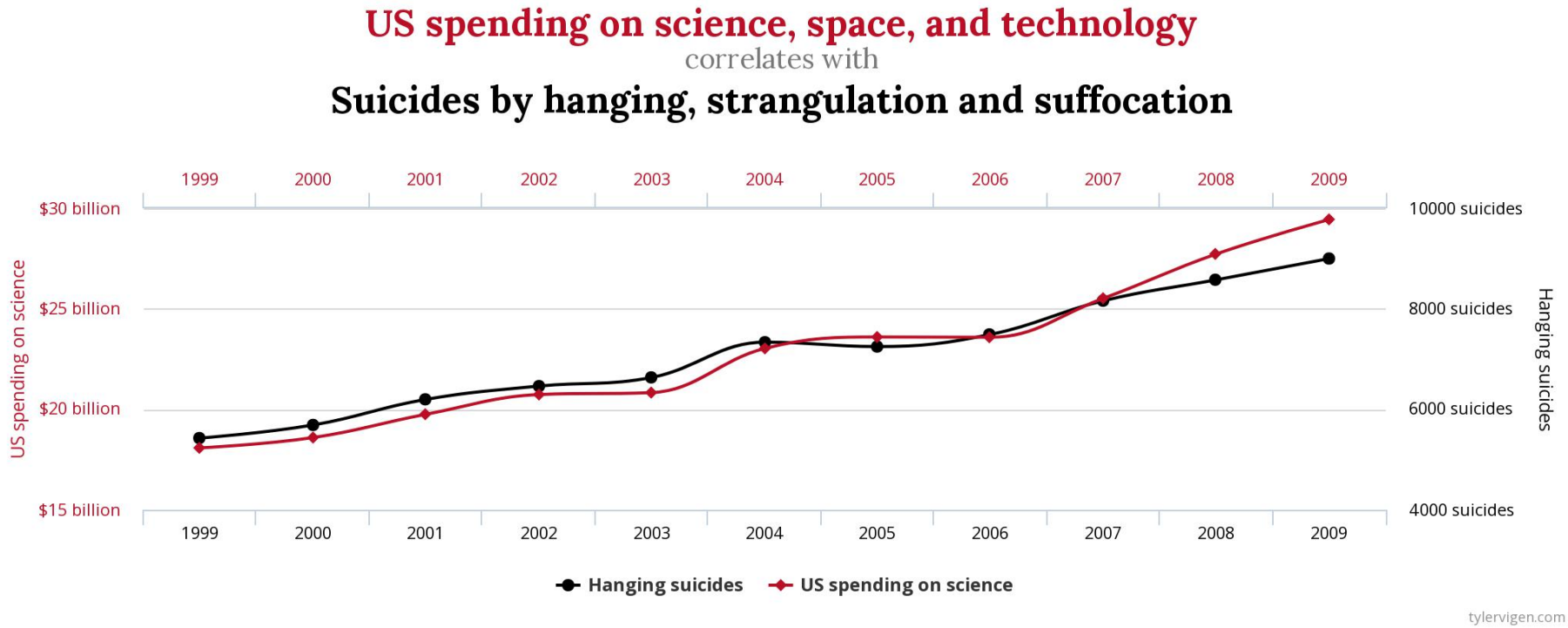
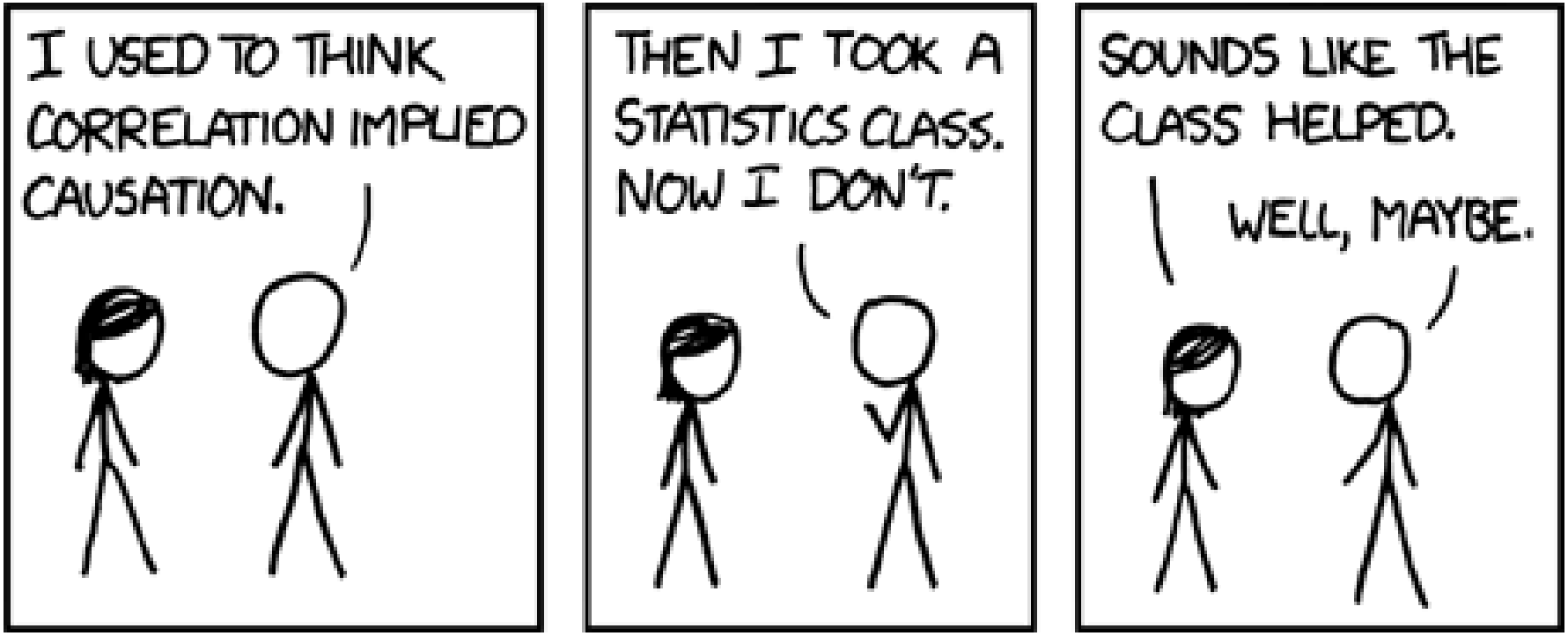



Figure: Correlation Examples

# Terms for Exploring Categorical Data - Interpretation



# Something Interesting!



# Exploring 2 or more variables - Contingency Tables

- Tabular representation of categorical data
- Tally of counts between two or more categorical variables
- Portraying data that can facilitate calculating probabilities
- Helps in determining conditional probabilities

Table: Gender and Preferred Category of Coffee ... It was beer!!!

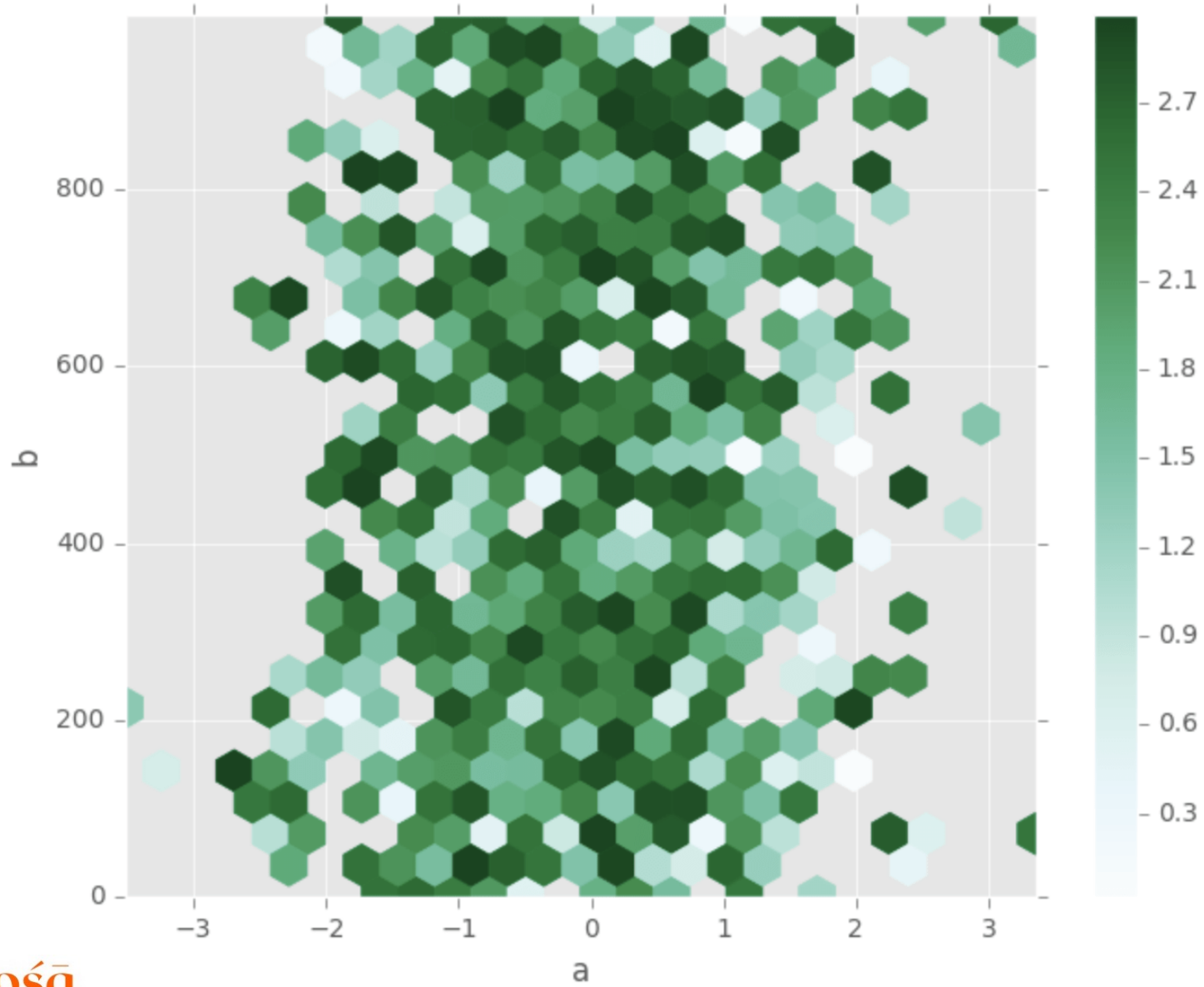
| | Light | Regular | Dark | Total |
|---|---|---|---|---|
| Male | 20 | 40 | 50 | 110 |
| Female | 50 | 20 | 20 | 90 |
| Total: | 70 | 60 | 70 | 200 |

- Used in marketing research etc

# Exploring 2 or more variables - Hexagonal Binning

- Scatter plots - Get dense with very large number of data
- Bivariate histogram for visualising structure in datasets with large $n$
- Plots density rather than points
- $xy$ plane over the set (range($x$), range($y$) is tessellated by grid of hexagons
- Number of points falling in each hexagon are counted and stored in a data structure
- Hexagons with count $> 0$ are plotted using color ramp or varying the radius of the hexagon in proportion to the counts

# Hexagonal Binning

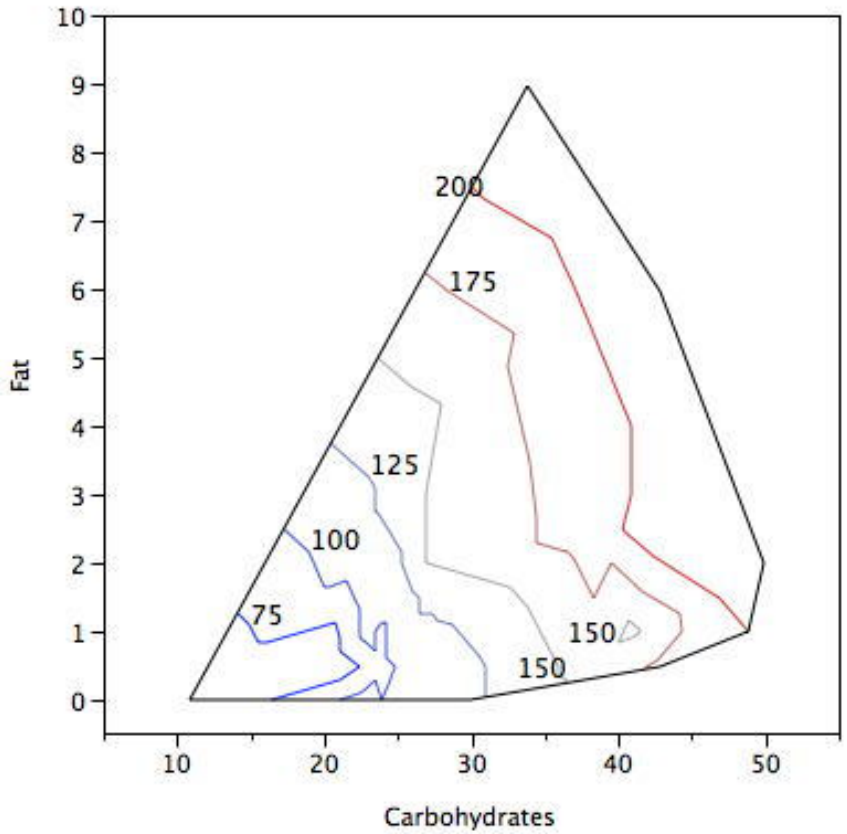


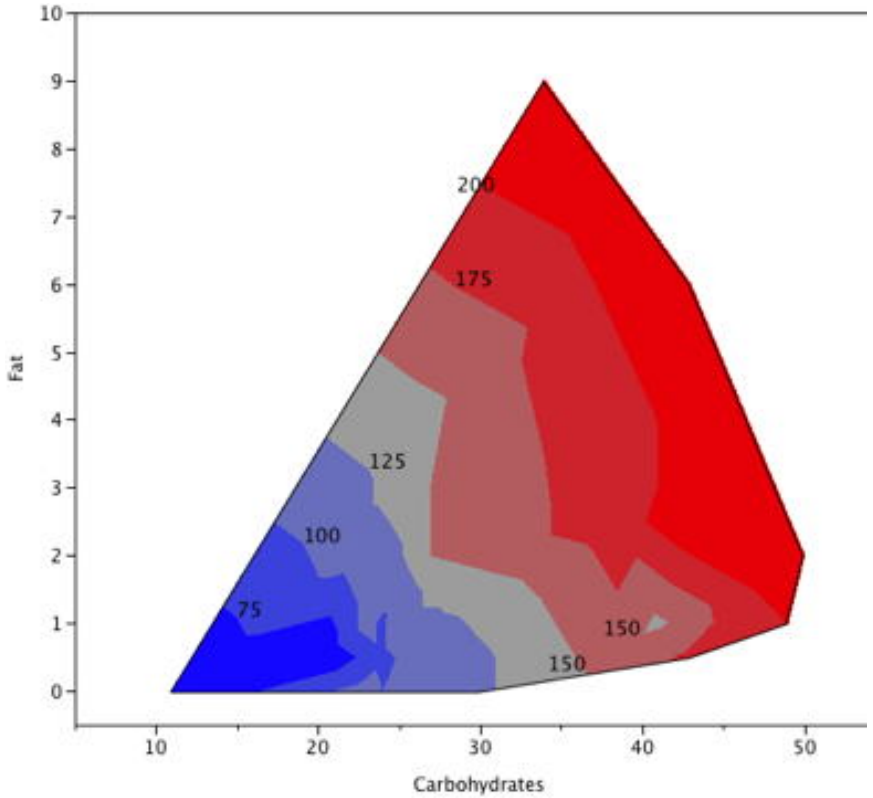
# Contour Plots

- Portray data for 3 variables in 2D
- Each contour line has a constant value on a 3rd variable
- Each band represents a specific density of points, increasing as one nears a peak

# Contour Plot - Example[5]

Calories as a function of fats and carbohydrates



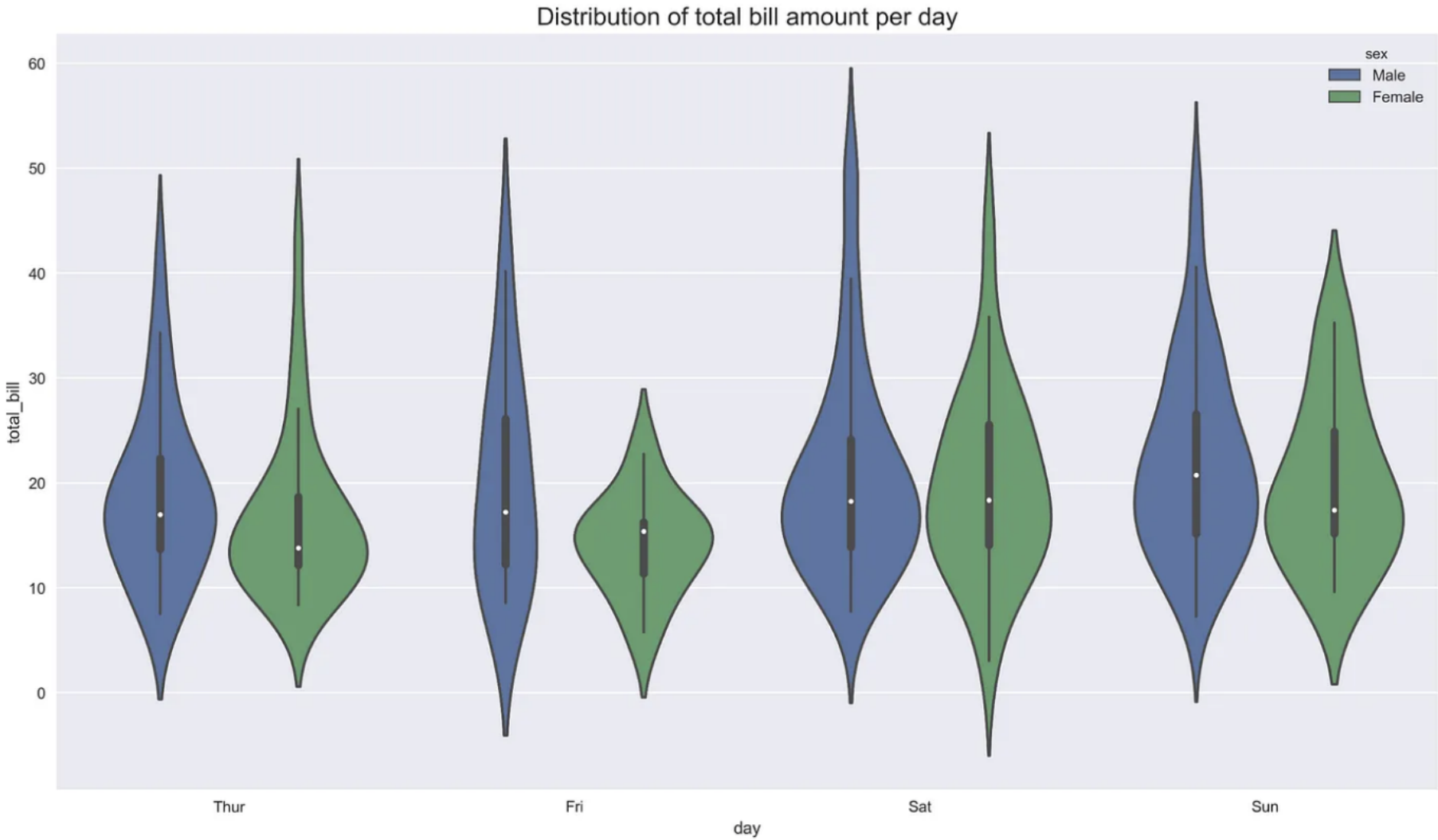




[5]https://onlinestatbook.com/2/advanced_graphs/contour.html

# Categorical and Numerical Data - Violin Plots

- Box plot with additional plot of density estimate with density on the $y$- axis
- Density is mirrored and flipped over and resulting shape is filled in - creating an image similar to a violin
- Shows nuances in the distribution that are not found in boxplot

# Violin Plot



# Violin Plot