Exploratory Data Analysis

A first look at the data

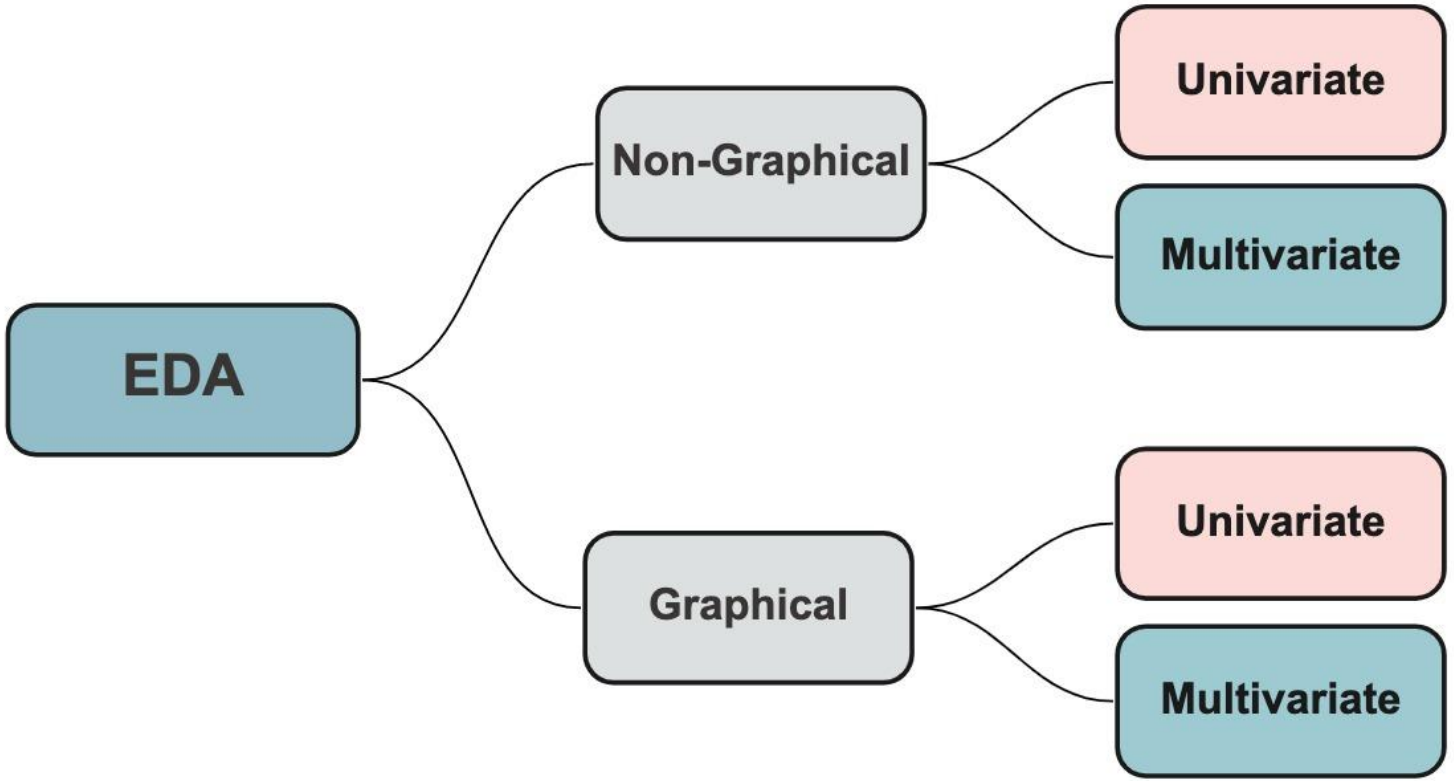
Kaustubh Kulkarni

Definition of EDA

- Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data.
- EDA is used by data scientists to analyze and investigate data sets and summarize their main characteristics including outliers.
- EDA provides a better understanding of data set variables and the relationships between them.
- It helps determine how best to manipulate data sources to get the answers you need.
- Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

Types of EDA

- Exploratory data analysis is generally cross-classified in two ways.
- First, each method is either **non-graphical** or **graphical**.
- Second, each method is either **univariate** or **multivariate** (usually just bivariate).


EDA
Non-Graphical
Univariate
Multivariate
Graphical
Univariate
Multivariate

Types of EDA (continued)

- Beyond the four categories created by the above cross-classification, each of the categories of EDA have further divisions based on
  - role (**outcome** or **explanatory**) of the variable(s)
  - type (**categorical** or **quantitative**) of the variable(s) being examined.

A quick review of types of variables

Types of variables

1. Categorical variables

A categorical variable (also called qualitative variable) refers to a characteristic that can't be quantifiable. Categorical variables can be either nominal or ordinal.

1a. Nominal variables

A nominal variable is one that describes a name, label or category without natural order, eg, gender

1b. Ordinal variables

An ordinal variable is a variable whose values are defined by an order relation between the different categories, eg, customer ratings

Types of variables (continued)

2. Numeric variables

A numeric variable (also called quantitative variable) is a quantifiable characteristic whose values are numbers (except numbers which are codes standing up for categories).

2a. Interval

A variable measured on an interval scale gives information about more or betterness as ordinal scales do, but interval variables have an equal distance between each value.

2b. Ratio

Something measured on a ratio scale has the same properties that an interval scale has except, with a ratio scaling, there is an absolute zero point.

Examples of interval data

- Examples of **interval level data** include temperature measured in degrees Celsius or Fahrenheit and year.
  - There is the exact same difference between 100 degrees and 90 as there is between 42 and 32.
- Other examples:
  - Time of day in a 12-hour clock.
  - Temperature in degrees Fahrenheit or Celsius (not Kelvin).
  - IQ test.
  - SAT and GRE scores.
  - Age.
  - Income range.
  - Year.
  - Voltage.

Example (continued)

- For instance, zero degrees Fahrenheit and zero degrees Celsius are different temperatures, and neither indicates the absence of temperature.
- Zero meters and zero feet mean exactly the same thing, however.
- An implication of this difference is that a quantity of 20 measured at the ratio scale is twice the value of 10, a relation that does not hold true for quantities measured at the interval level (20 degrees is not twice as warm as 10 degrees).
- In contrast, a ratio scale, like the Kelvin scale, has a meaningful zero point. On the Kelvin scale, 20 Kelvin is twice as warm as 10 Kelvin because 0 Kelvin represents absolute zero, the complete absence of thermal energy.

Types of variables (continued)

3. Explanatory Variable

Also known as the *independent* or *predictor variable*, it explains variations in the response variable; in an experimental study, it is manipulated by the researcher

4. Response Variable

Also known as the *dependent* or *outcome variable,* its value is predicted or its variation is explained by the explanatory variable; in an experimental study, this is the outcome that is measured following manipulation of the explanatory variable

Example

Outcome Variable / Response Variable: Blood Pressure

In a study examining the effects of a new drug on blood pressure, the outcome variable or response variable could be the participants' blood pressure readings. This is the variable of interest that researchers want to observe.

Explanatory Variable: Drug Administration (Placebo vs. New Drug)

Participants are either given a placebo or the new drug. Researchers manipulate this variable to observe its impact on the participants' blood pressure, the outcome variable.

Univariate non-graphical EDA : Categorical Data

- The characteristics of interest for a categorical variable are simply the range of values and the frequency (or relative frequency) of occurrence for each value.
- Therefore the only useful univariate non-graphical techniques for categorical variables is some form of **tabulation** of the frequencies, usually along with calculation of the fraction (or percent) of data that falls in each category.

Example

Statistic/College	H&SS	MCS	SCS	other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

Univariate non-graphical EDA : Quantitative Data

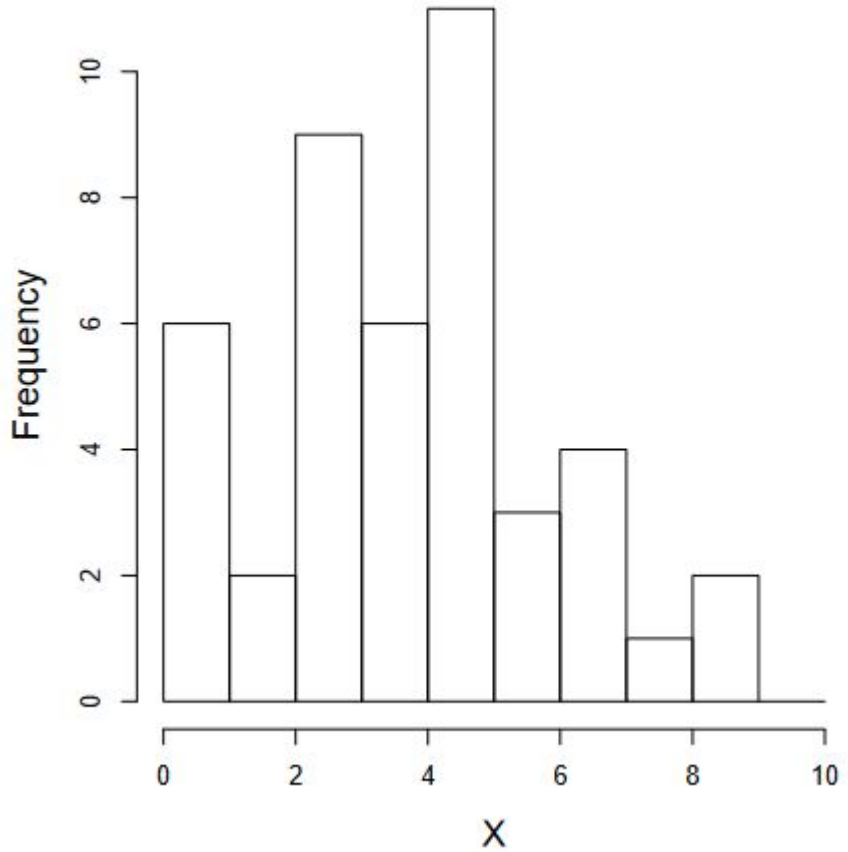
- Measures of central tendency
- Measures of spread

(covered in lecture on module 1.3 descriptive statistics)

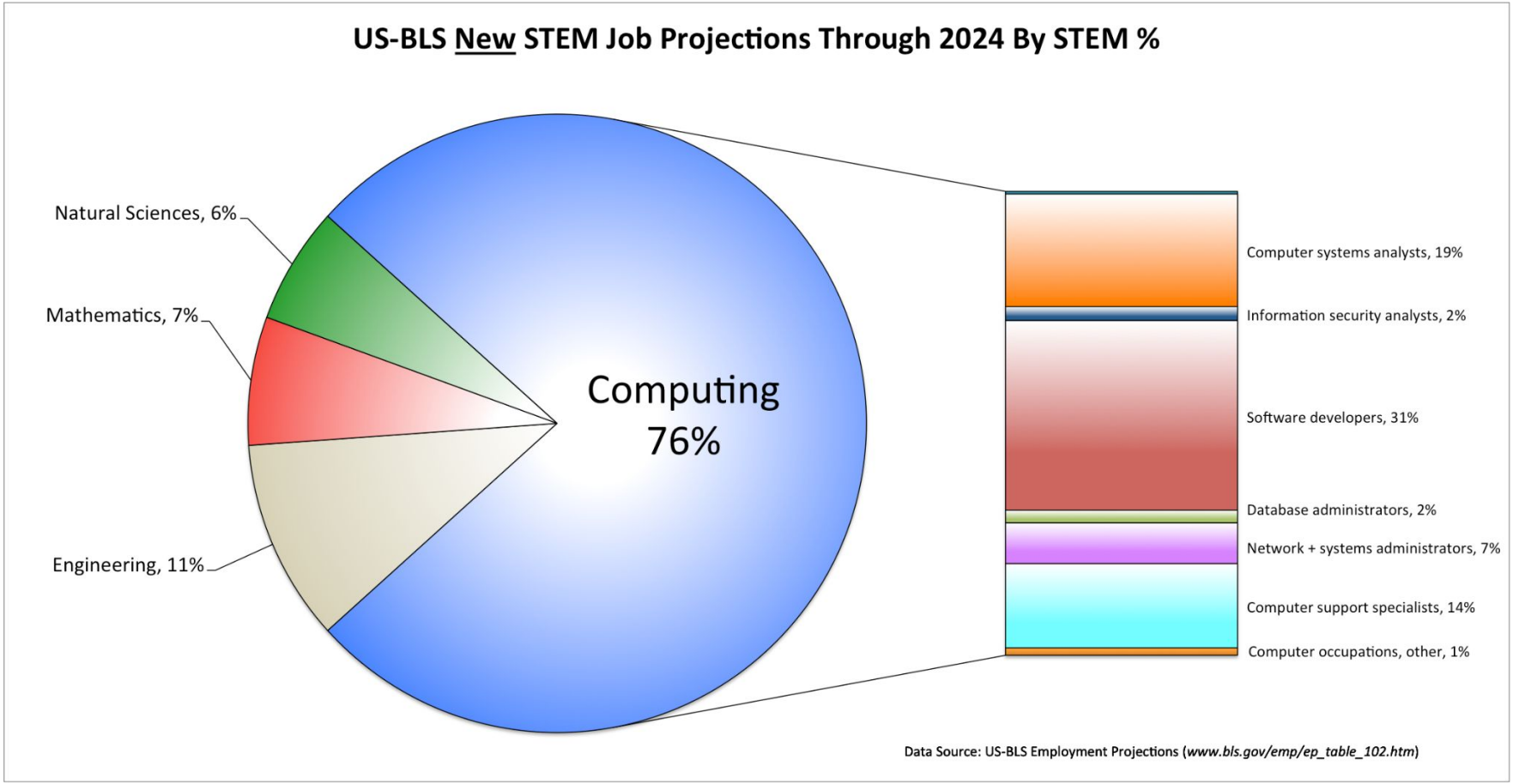
Univariate graphical EDA

- Histogram (Categorical)
- Pie chart (Categorical)
- Box plot (Quantitative)
- Stem and leaf plot (Quantitative)

Histogram



Pie Chart


US-BLS New STEM Job Projections Through 2024 By STEM %
Natural Sciences, 6%
Mathematics, 7%
Engineering, 11%
Computing
76%
Computer systems analysts, 19%
Information security analysts, 2%
Software developers, 31%
Database administrators, 2%
Network + systems administrators, 7%
Computer support specialists, 14%
Computer occupations, other, 1%
Data Source: US-BLS Employment Projections (www.bls.gov/emp/ep_table_102.htm)

Stem and leaf plot

1|000000
2|00
3|000000000
4|000000
5|00000000000
6|000
7|0000
8|0
9|00

Distribution of customer orders processed per day

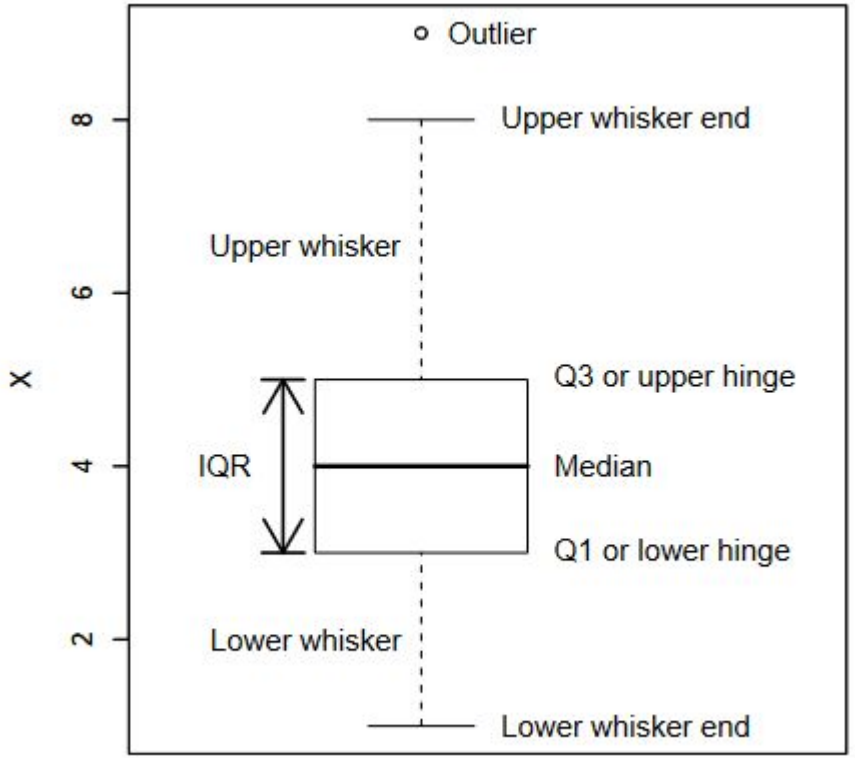
Sample of 80 days during the period
Jan 1, 2020 - Dec 31, 2020

Stem	Leaf
2	7 7
3	0 1 2 4 5 6 7 9
4	1 1 2 2 2 5 7 7 8 8 9
5	4 4 6 8 8 8
6	3 4 5 5 5 6 7 7 8 9 9
7	0 0 1 2 3 3 5 5 5 7 7 8 9
8	1 2 4 4 5 5 6 7 7 8 9
9	0 0 0 2 4 5 5 6 8 8
10	2 4 8 9
11	0 4
12	2 5

35	108	109	45	122	36	104
110	98	67	56	65	58	65
37	82	85	47	89	27	70
58	84	73	114	42	71	66
72	75	88	54	90	81	87
102	77	77	67	58	95	34
39	69	98	78	87	85	86
125	49	47	95	27	92	75
31	42	79	94	96	84	
48	73	41	75	63	42	
64	32	90	69	90	48	
54	65	30	41	68	70	

Source: Speedy Dry Cleaners.

Box Plot


Outlier
Upper whisker end
Upper whisker
Q3 or upper hinge
IQR
Median
Q1 or lower hinge
Lower whisker
Lower whisker end
X
8
6
4
2

Multivariate non-graphical EDA : Cross Tabulation

Types of Jobs Across Indian Generations

Among workers who work full time for an employer

2012 first half year	White-collar jobs	Blue-collar jobs (agriculture)	Blue-collar jobs (manu-facturing)	Blue-collar jobs (others)
Gen Y (ages 15 to 30)	22%	44%	8%	26%
Gen X (ages 31 to 47)	19%	46%	7%	28%
Baby Boomers (ages 48 to 66)	16%	57%	5%	22%

Based on aggregated surveys conducted in Februrary-March and May-June 2012

GALLUP

Two-way contingency tables

```
> HairEyeColor
, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

, , Sex = Female

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

Year	Party	Ideology			
		Liberal	Moderate	Conservative	Total
1980	Democrat	40.6%	43.4%	31.5%	38.6% (548)
	Independent	42.2%	36.0%	35.3%	38.2% (541)
	Republican	17.2%	18.7%	33.2%	23.2% (329)
		(360)	(579)	(479)	n = 1418
				gamma = 0.175**	
2010	Democrat	58.0%	34.3%	16.3%	35.2% (673)
	Independent	35.1%	50.6%	35.3%	41.0% (785)
	Republican	6.9%	15.2%	48.4%	23.8% (456)
		(553)	(724)	(308)	n = 1914
				gamma = 0.569**	

Correlation analysis

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

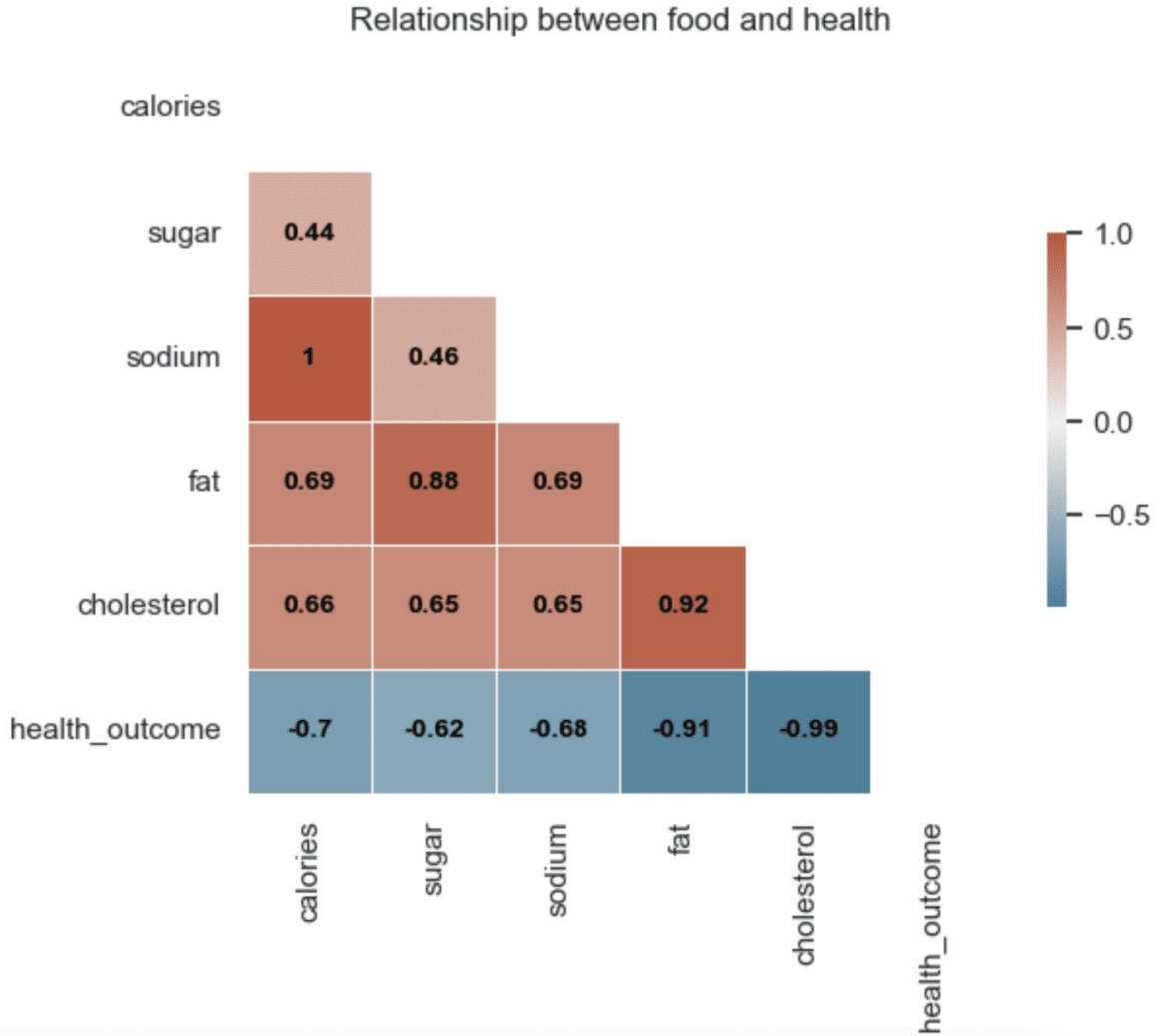
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Heatmap


Relationship between food and health
calories
sugar
sodium
fat
cholesterol
health_outcome
0.44
1
0.46
0.69
0.88
0.69
0.66
0.65
0.65
0.92
-0.7
-0.62
-0.68
-0.91
-0.99
calories
sugar
sodium
fat
cholesterol
health_outcome
1.0
0.5
0.0
−0.5

Case Study (module 2.3 of syllabus)

Company: Online grocery delivery service

Challenge: Increasing customer churn (cancellation of subscriptions)

Data: Transaction history of all customers, including items purchased, frequency of orders, delivery location, demographics, etc.

EDA Process:

Descriptive statistics: Analyzing average order value, purchase frequency, and most popular items per customer segment.

Visualization: Plotting churned vs. non-churned customers by order frequency, purchase amount, and location.

Correlation analysis: Identifying correlations between demographic factors and churn rate.

Insights generated:

EDA revealed that customers who ordered infrequently and spent less were more likely to churn.

Certain product categories like fresh produce and household essentials had higher purchase frequency among non-churned customers.

Customers in suburban areas had a higher churn rate compared to those in urban areas.

Actions taken:

Based on these insights, the company:

- Launched targeted marketing campaigns to re-engage infrequent customers with personalized offers and discounts.
- Introduced subscription plans with guaranteed deliveries of fresh produce and essentials.
- Offered localized promotions and delivery incentives for customers in suburban areas.

Results:

Customer churn rate decreased significantly, leading to increased revenue and customer retention.

You can go through some real world case studies for 2.3 in the separate document uploaded.

Questions?