5 Main Steps of Data Analysis

1. Define the questions that you want to answer through the analysis

- clearly define the problem
- make hypothesis or research questions
- identify what types of data you will need and where it'll come from

2. Collect the Data

- primary sources: internal data that the company already has (eg: CRM software, email marketing tools etc.)
- secondary sources: public/external data (eg: government, Google Trends, WHO etc.)

3. Clean the Data

- remove duplicates
- handle anomalies and missing data

4. Analyze the Data

- regression analysis
- cluster analysis
- time-series analysis

5. Interpret and share the results

- visualization

Terms

- ETL = Extract, Transform, Load
- Dax = Data analyst expressions
- Copilot = Feature in Excel, like ChatGPT powered by Microsoft

Notes

1. Concatenate values from 2 string columns:

=A2&B2 if A2='x' and B2='y', then this will produce "xy"

=A2&"-"&B2 adds a hyphen (-) between the two cell values

- the quotation marks are necessary to put text characters between two values

2. Range = collection of cells

Functions

1. Logical

a. IF(condition, returned_value_if_true, returned_value_if_false)

b. AND(condition1, condition2)

c. OR(condition1, condition2)

d. IFS(condition1, value1_if_cond1_is_true, condition2, value2_if_cond2_is_true)

2. Math

a. COUNT(B:B) - counts all non-blank NUMERICAL cells of the given range (column B)

b. COUNTA(B:B) - counts all non-blank cells of the given range

c. COUNTIF(B:B, $A2) - counts how many values of B is equal to the value of A2

d. COUNTIFS(range1, condition1, range2, condition2)

- counts how many cells of range1 match condition1 & how many cells of range 2 match condition2

e. SUM(B:B)

f. SUMIF(range_to_assess, condition_to_assess, range_to_add)

g. SUMIFS(range_to_add, range1_to_assess, condition1_to_assess, range2_to_assess, condition2_to_assess)
h. AVERAGE(B:B)
i. AVERAGEIF
j. AVERAGEIFS
k. MIN(B:B)
l. MAX(B:B)

3. Statistical

Median

It is the number in the middle of the sorted set (smallest to largest)

Median is better than average for statistics because average is affected by the outliers in the dataset

Standard Deviation (σ)

= how much the data point deviates from the standard (mean)

= how spread out the data is

If SD is low, that means the data is closely clustered around the mean/average

If SD is high, that means the data is dispersed over a wider range of values

SD is used to understand if a specific data point is standard and expected, or unusual and unexpected

A data point's distance from the mean can be measured by the number of standard deviations (i.e. if it is above or below the mean)

1σ = 68% = 68% of data fall within 1 SD of the mean

2σ = 95%

3σ = 99.7%

SD is used when the distribution of data is approximately normal (like a bell curve)

SD calculation:

Population	Sample
$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{n}}$	$S = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$
μ - Population Average **xi - Individual Population Value** **n - Total Number of Population**	**x - Sample Average** **x_i - Individual Population Value** **n - Total Number of Sample**

Quartile

25% = 25th percentile = 1st Quartile = first 25% of the sorted data or the histogram?

50% = 50th percentile = 2nd Quartile = Median

75% = 75th percentile = 3rd Quartile

100% = 100th percentile = 4th Quartile

Interquartile Range = Q3-Q1

a. MEDIAN(B:B)
b. STDEV.P(B:B) - standard deviation of the population
c. STDEV.S(B:B) - standard deviation of a sample
d. QUARTILE.INC(range, quartile_value)
    i. eg: QUARTILE.INC(A:A, 2) = 2nd quartile (median)
    ii. Inclusive (percentile range of 0 to 1 inclusive)
    iii. Considers the median

iv. Used often for odd-numbered sample size

e. QUARTILE.EXC(range, quartile_value)

i. Exclusive (percentile range of 0 to 1 exclusive)

ii. Doesn’t consider the median

iii. Used for even-numbered sample size

iv.

f. RANK(value_to_rank, entire_range, 0)

i. 0 = descending order, 1 = ascending order, this is optional argument

g. MODE(A:A)

i. The most frequently occurring value in the given range

4. Array

5. Lookup

6. Text

7. Date & Time

8. V