

L1 Introducⁿ to Datasets

- we instruct the computer, to achieve a certain goal.
- seq. of procedures
- the problems basically have a pattern, they can be used with procedures interchangeably.
- Dataset → scores, shopping bills, words from a paragraph.

L2 Concept of Variables, Iterators &

Filtering

- keep track of something \Rightarrow helps to remember.
e.g. count = $\& \times 2 \dots \rightarrow$ it keeps on changing
 \hookrightarrow variable (iterators)
- mostly, in variables we are interested in its value at the end of the flow
Avg. Maths marks =
$$\frac{\text{sum of maths marks for } 30 \text{ student}}{30}$$

- iteration, variables, take the values of variables and do some operations on it.
- Filtering
 - ↳ let's say from the word dataset I want to find the count of words that are only verbs. We can keep count of filtered words while doing the iteration. Thus, saving the time. \Rightarrow Filtered Iteration.
Eg. getting the count of verbs & punctuations while doing 1 iteration.

L: Tutorial

- Eg. of shopping bill
Count the no. of bills which have
 - (a) no food item purchase
 - (b) only food item purchase.Process \rightarrow we do the iteration over the entire shopping bills dataset. We initialize 2 variables (allFood and nonFood) with 0. We check & update these 2 variables accordingly.

- calculate the average word len. of text data.

count of cards seen =

letter count (given on card) =

once the iteration is over, we will divide letter count total by total count of words seen to get the average word length.

L1.3 Iteration using combination of Filtering conditions

- find all the girls (gender) from chennai (city)
 1. first find all the girls, keep them separately. In second iteration, we find the specific city.
 2. or filter once on chennai, then, find the girls.
 3. do double conditions (F & Chennai) in one iteration itself.
both should be true to satisfy the condition
and operator.

- find the no. of boys & girls who are born in the first half of the year (before 30th June)
 1. keep track of gender and at the same time, check the DOB w.r.t. 30th June.
 - ↳ exactly half
- as a teacher, I want to find whether girls or boys, who are doing better in Maths \Rightarrow find Math's avg for girls & boys.
We have 4 variables \rightarrow girls, boys count, girls, boys maths marks.
 - ↳ find the average \Rightarrow do it in 1 scan.
 - ↳ algorithm \Rightarrow to validate the hypothesis

L: Tutorial

Ex. Find the no. of adjectives which have letter count ≥ 5 .
 \Rightarrow go through the words dataset, find all the adjectives and check whether its word letter count is ≥ 5 or not. Based on it, update the

variable accordingly.

Eg. Calculate the amount spent by a particular person in each of the stores.

⇒ go through shopping bills dataset.

find the particular person along with that check the store name!

Then, here you will see there are 3 unique stores. Update the variables accordingly.

Eg. Compare the size of pronoun and and prepositions.

⇒ We now have to compare, so first better to have a avg. lengths of pronouns and prepositions. Find that using 4 variables. Then compare the avg. lengths.

L1.4 Intro. to Flowcharts

- we should formalize these procedures what we have done so far.

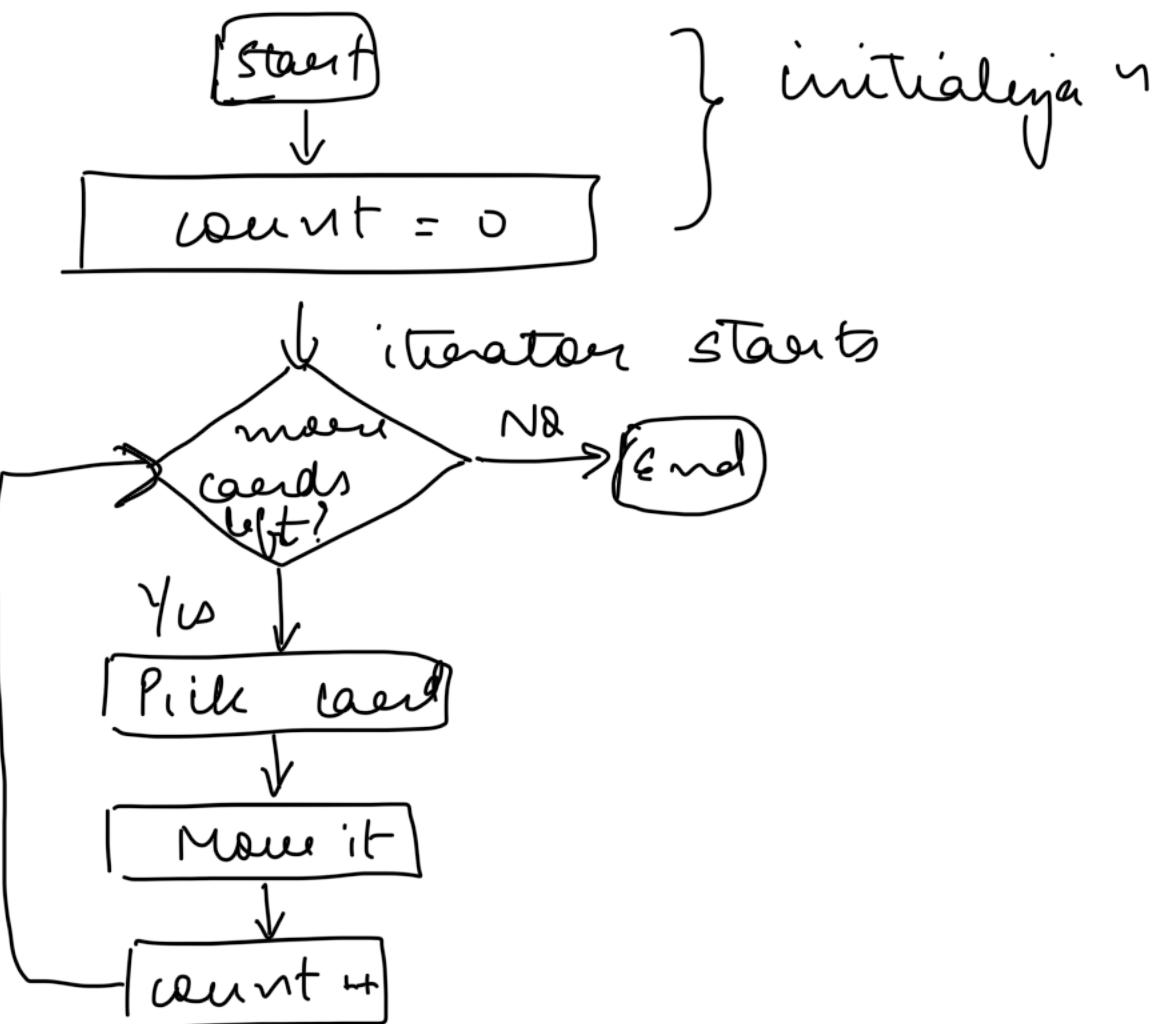


⇒ process / activity

→ → flowline / arrow \Rightarrow order of execution

 ⇒ decisions ⇒ determine the program path
 ⇒ Terminal ⇒ indicates the start or end of prog.

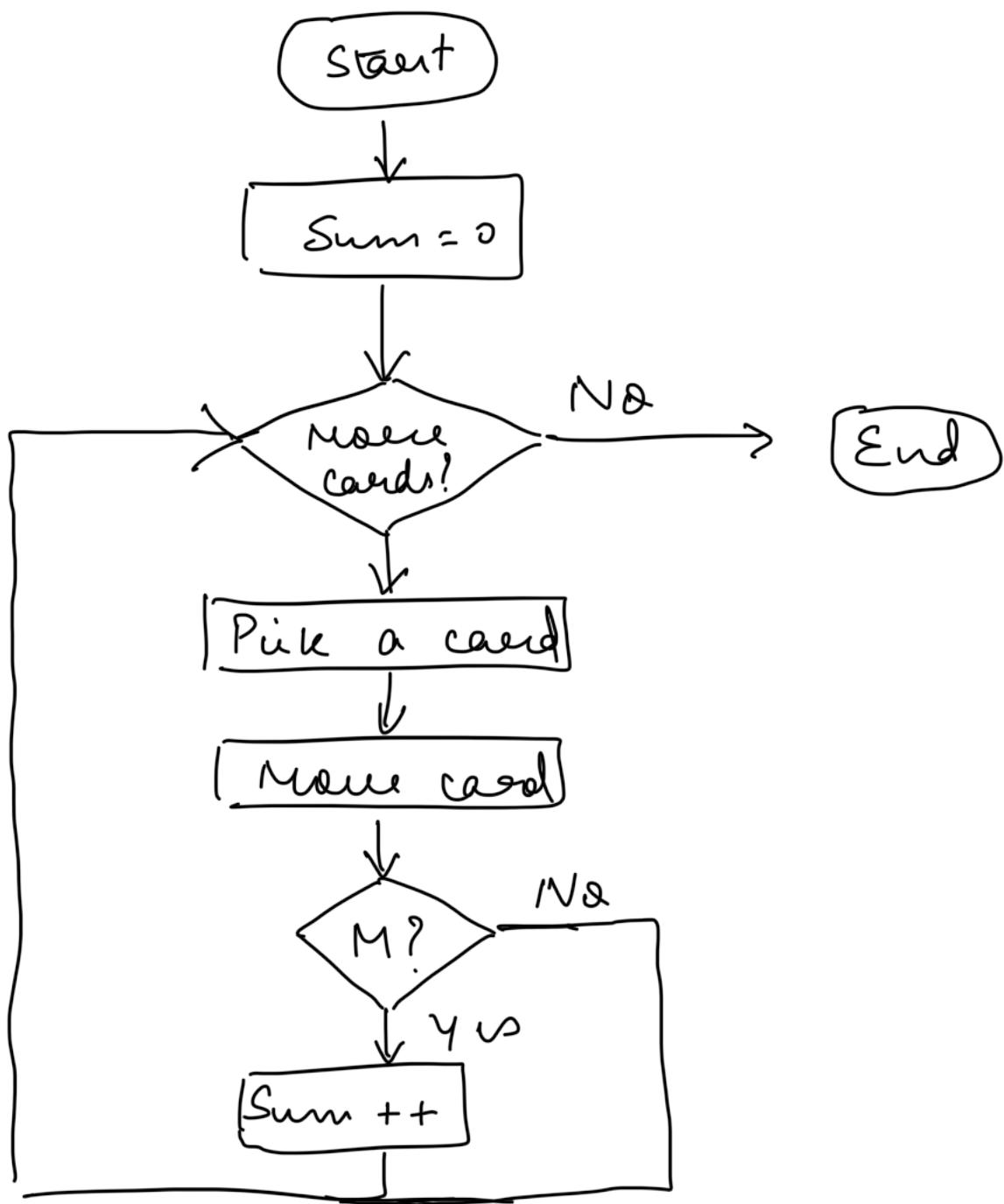
Eg. to count cards



Eg. to calculate the sum of maths marks using flow chart.

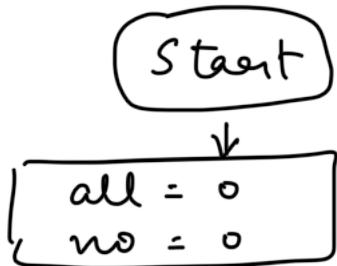
from these 2 flowcharts we can create a generalized flowchart for iteration.

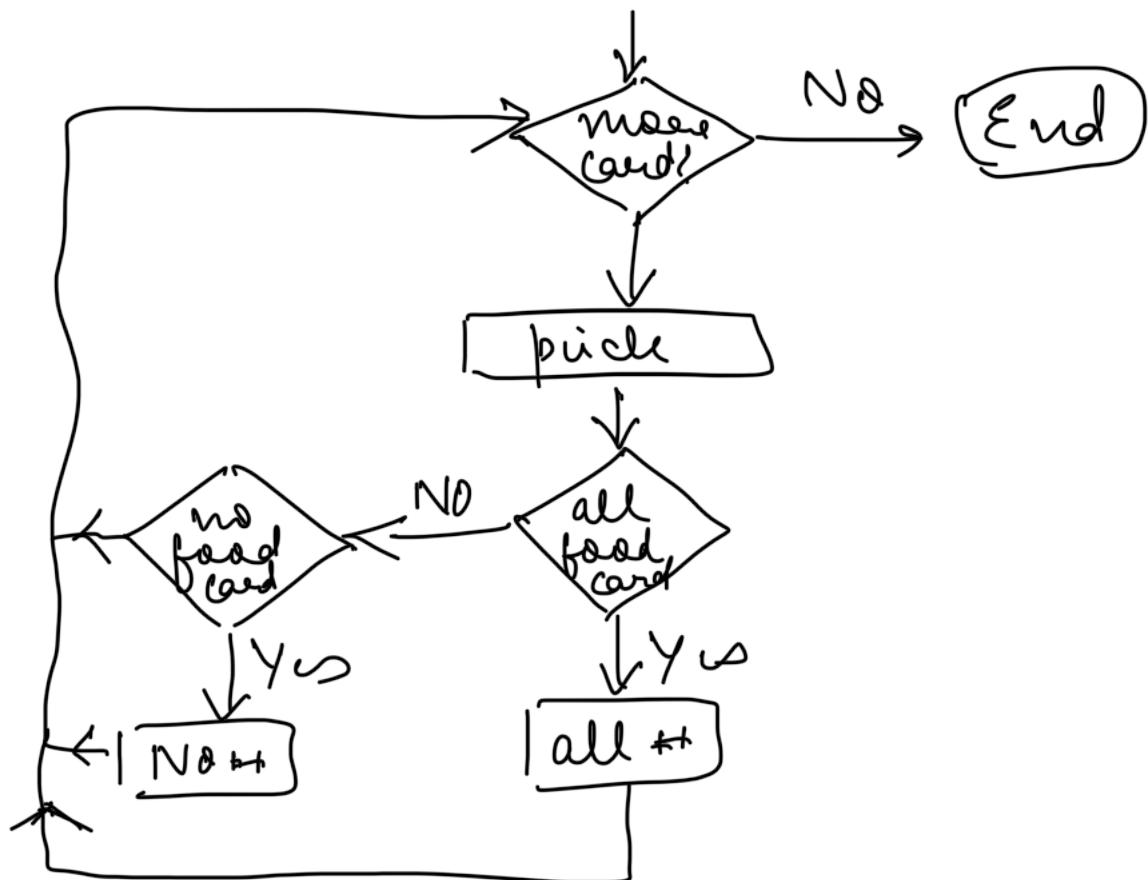
L5 Flowchart for Sum with Filtering



L : Tutorial

- convert allfood and no food purchase into a flowchart .





L6 Sanity of Data

- every field will have a certain datatype only.
- even if a no. \Rightarrow there should be range specified.
(legal values)
- typing mistakes are common.
- overridden problem by just copy-pasting wrong value at a wrong field.
- maybe a swapped entry, otherwise data is correct.
- a constraint can be applied.

Eg. Total = sum total of all the subj. marks

Eg. dates = {1.3.}, {1.30}, {1.29/28}

- name credentials (first last or last first)
- in shopping data -
 1. in a particular column, same datatype should be used.
 2. dep. on context, you have to check certain items can be brought as a whole.
 3. there can't be a quantity with 2 decimals used.
 4. maybe sometime internal inconsistency, total could be wrong.
 5. items and category should be relevant to each other.
- in words data
 1. check for spellings
 2. per word, it should match the given word length.
 3. every words belongs to a larger class like verb, pronoun, etc. By using a grammar checker / dictionary, we need to verify whether or not they have been defined properly.

L7 Intro to Datatypes

- in during the sanity of data we saw

about constraint satisfaction like a particular column should have a particular datatype, then each of them have a range and other constraints as well.

↳ concept of datatype

it tells the intention of that particular var./column, also with that it specifies operations and range.

- Boolean $\rightarrow T \text{ or } F$. Op["] $\rightarrow \text{AND, OR}$
- Integer $\rightarrow -\infty \text{ to } +\infty$ Op["] $\rightarrow +, -, \times, \%, <, >, =$
- Character \rightarrow alpha, digits, spl. char.

L8 Subtypes of Basic Datatypes

- Integer
 - 1. Seq. No \rightarrow a range ($0 \leq i \leq \text{max}$)
 \rightarrow no negative values.
 \rightarrow no integer op["] make sense.
 - 2. Marks \rightarrow a range ($0 \leq i \leq 100$)
 \rightarrow Op["] $\rightarrow +, -, >, <, =$
- Character
 - 1. Gender $\rightarrow M \text{ or } F$
 - String \rightarrow sequence of charac. (no

restrictions on length, can be
alpha numeric)

Opⁿ → in, =,

1. Names → string with no spl. chars.
also, \no digits
2. Category → take some specific values
and not others.

L9 Transformation of Sub - Datatypes

- Integer

1. dates → can't allow strings, the
maths operations won't work.

{1, ..., 365} corresponding
uniquely to each day.

Opⁿ → Print, >, <, =

2. Marks → $0 < i < 100$

Opⁿ → +, -, >, <, =

3. Fractional Marks → float (nos.
with decimal places) \therefore , new
constraint for float has to be
introduced.

4. Amount → $0 \rightarrow \infty$ (+ float)

Opⁿ → +, -, >, <, =

5. Quantity \rightarrow int + float

$Op^n \rightarrow +, -, >, <, =$

L10 Intro. to Complex Datatypes

- basic data types \rightarrow int, bool, charac.
- record (\rightarrow struct, tuple)
 - ↳ data type with multiple fields \rightarrow each of them have a name & value.
- marks = {
 - sig : signo
 - name : name
 - gender : M or F
 - marks : $0 < i < 100$
 - total : $0 < i < 100$}
- if I get marks cards, they can be added as a list (seq. of data elements (eg. a seq. of records)).
- but in shopping bill, items can have diff. values. so, a list is good. create a item tuple \rightarrow collect them into itemlist. Use that in the shopping bill dataset.