

Lecture 3

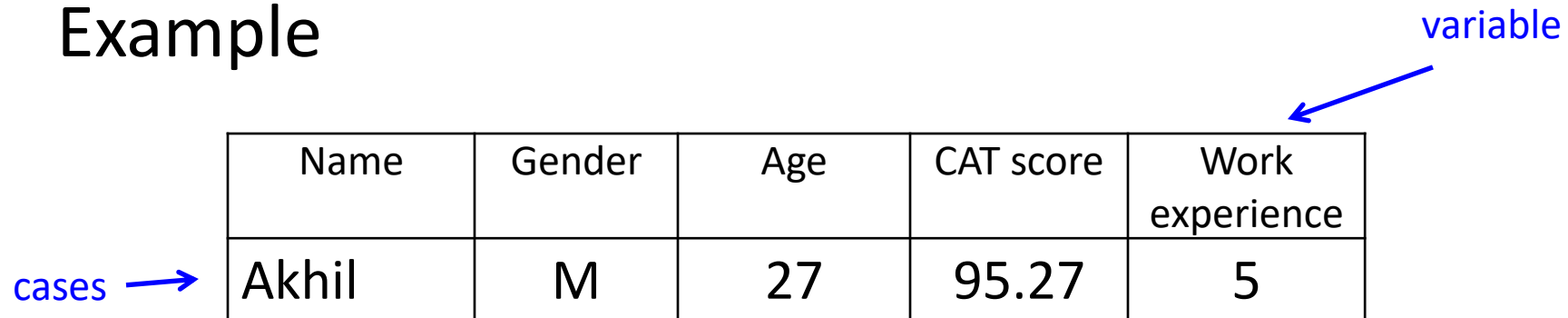
Discussion

Data

Cases and variables

In a data table, rows are called cases or observations while columns are variables.

Example



Name	Gender	Age	CAT score	Work experience
Akhil	M	27	95.27	5

Exercise

For the following variables, give a name and indicate the type of variable (categorical, ordinal, numerical)

1. Car owned by ten friends
2. Income of 20 employees
3. Size of clothes as S, M, L, XL
4. Number of students absent for class
5. Education of people as High School, Graduate, PG, PhD

No.	Description	Variable Name	Variable type
1	Car owned by ten friends	Model (or Brand)	categorical
2	Income of 20 employees	Salary	numerical
3	Size of clothes as S, M, L, XL	Cloth size	ordinal
4	Number of students absent for class	Absentees	ratio
5	Education of people	Education level	Categorical

We can say that Salary of x is Rs1000 more than y or can also say that salary of X is 30% greater than y so it can be Ratio and Interval both thus we categorised it as a Numerical type of data

True or false

1. Pin codes are examples of numerical data F - Categorical
2. Cases represent columns in a data table F - Rows
3. Frequency of time series is the **time spacing** between data T
4. **Likert scale** represents numerical data F - Ordinal data
5. **Aggregation of data** adds more cases F - Reduces

Time series data

- Data measured over time. Say students in an MBA class, year 2020 we had 20 students, Year 2021 we had 40 students, Year 2022 we had 70 students in the class.
- Example (Stock prices over a period, Sales of a shop over period, Price of petrol over period)

Likert scale - A Likert scale is a close-ended, forced-choice scale used in a questionnaire that provides a series of answers that go from one extreme to another. For example, a scale might have five choices that start at one end with "strongly agree" and end at the other with "strongly disagree,"

Scale of Like and Dislike
It represents Ordinal data and not numerical data

Aggregation of data is adding more data but this does not increase the cases but decrease them.
It makes the data more precise and accurate with addition of data thus we results in the reduction of cases

Cross sectional
data - Looking
at data at a
certain point in
time



Cross sectional or time series?

1. Company has data on number of employees who are in PF scheme and the amount in PF CS
2. 1000 people are asked if India would win the cricket world cup CS
3. The number of people who shopped for more than Rs 5000 on five days of the week TS
4. 100 customers of a hotel give feedback. 60 ticked excellent, 30 ticked average while 10 said poor CS
5. Number of sedans and small cars parked in front of a supermarket on 7 days of a week TS

Lecture 4

Describing categorical data

Number of votes polled when asked “Who will score most runs?”

(Imaginary data)

Name	Votes polled	Fraction	Percentage
Chris Gayle	45276	0.097732	9.77
Ajinkya Rahane	39825	0.085966	8.60
Virat Kohli	32419	0.069979	7.00
A B de villiers	29666	0.064037	6.40
Suresh Raina	48977	0.105721	10.57
Brendon McCullum	41678	0.089966	9.00
Shikar Dhawan	26423	0.057036	5.70
Rohit Sharma	30912	0.066726	6.67
J P Duminy	19627	0.042367	4.24
Glenn Maxwell	27555	0.05948	5.95
Ben Stokes	28432	0.061373	6.14
Ambati Rayudu	17666	0.038134	3.81
Angelo Mathews	15487	0.03343	3.34
Gautam Gambhir	22723	0.04905	4.91
Manish Pandey	14900	0.032163	3.22
Virender Sehwag	21700	0.046841	4.68

Total = **463266**

-Fractions in column C will add to "1"
-Percentages in column D will add up to 100

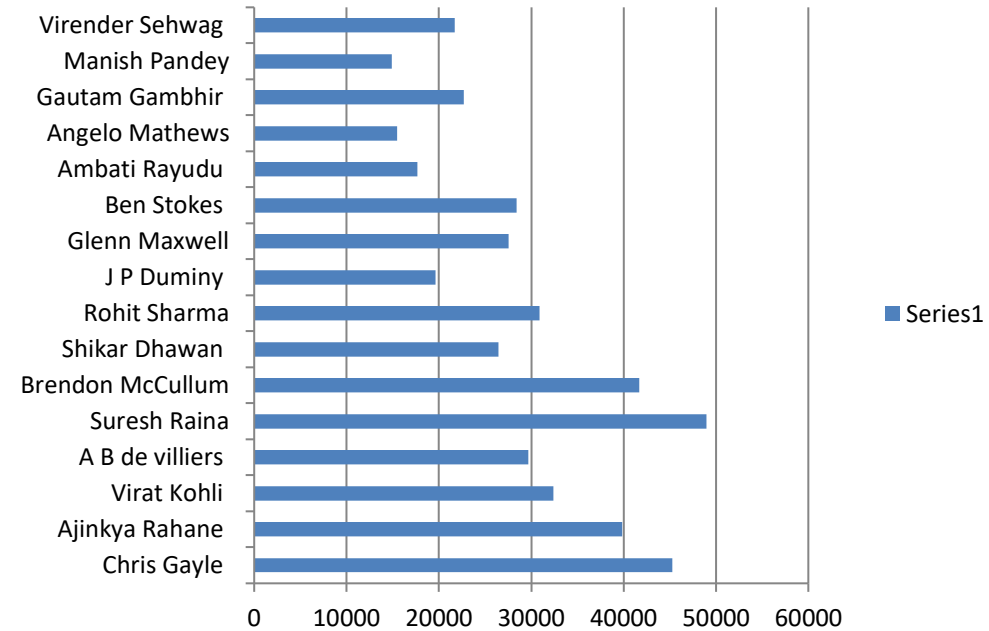
We can present all the data in the Frequency distribution table
But, the table will get large as the data increases thus it will become difficult to analyse the data

Frequency table – represents the distribution of a categorical variable as a table

Can become hard to compare as the table gets large

Same data in Bar chart

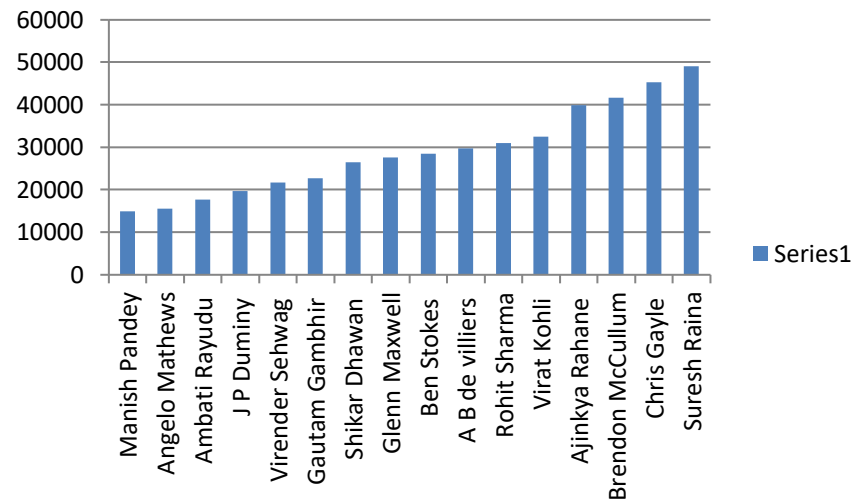
Horizontal
barchart



Advantage - Easier to
present the data

Disadvantage - It is
difficult to have the exact
data in this kind of
representation.

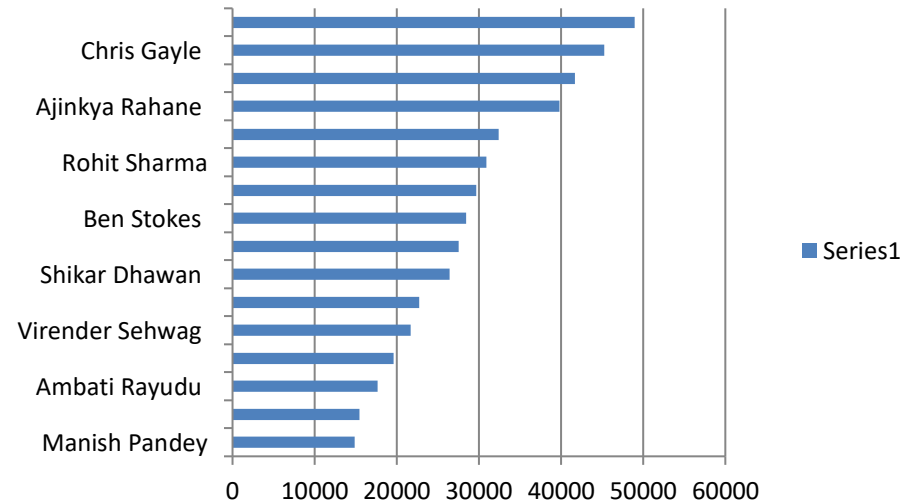
Vertical bar cahrt



Pareto charts

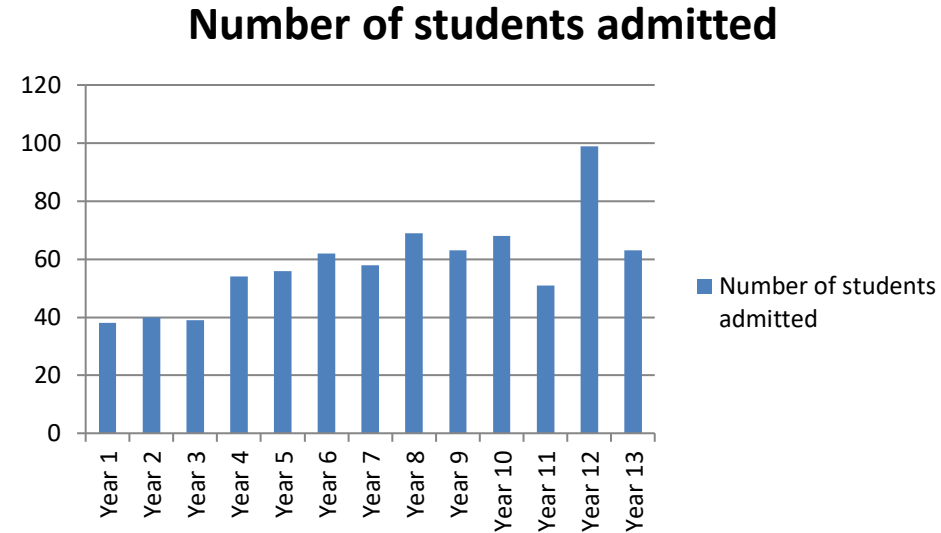
- Bar charts where the class with highest frequency is showed at the top (or on the right) and subsequently the classes with next highest frquencies are shown.
- In this way the chart visually depicts which situations are more significant.

Horizontal bar chart



Year	Number of students admitted
Year 1	38
Year 2	40
Year 3	39
Year 4	54
Year 5	56
Year 6	62
Year 7	58
Year 8	69
Year 9	63
Year 10	68
Year 11	51
Year 12	99
Year 13	63

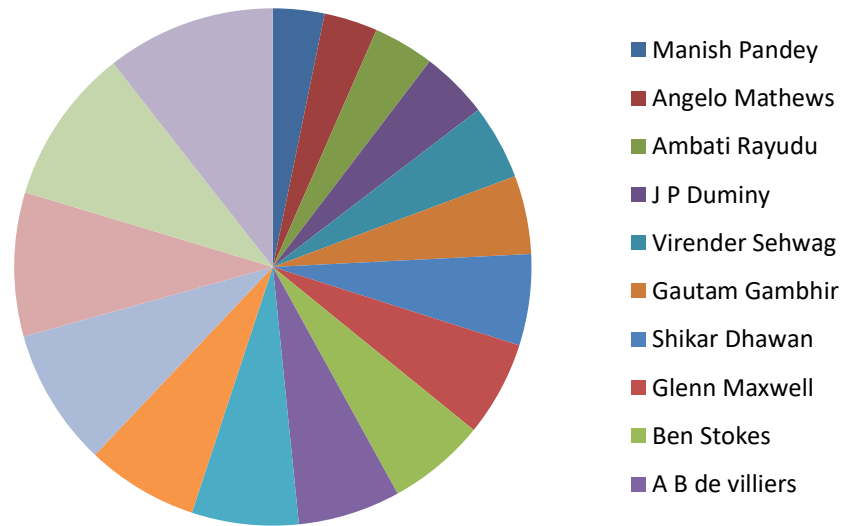
Example of a Vertical bar chart



Is this a frequency table?

Bar charts used to show frequency of categorical variable

Pie Chart



Shows distribution of a categorical variable as wedges of a circle

Useful for proportions particularly when some are close to $\frac{1}{2}$, $\frac{1}{4}$ etc

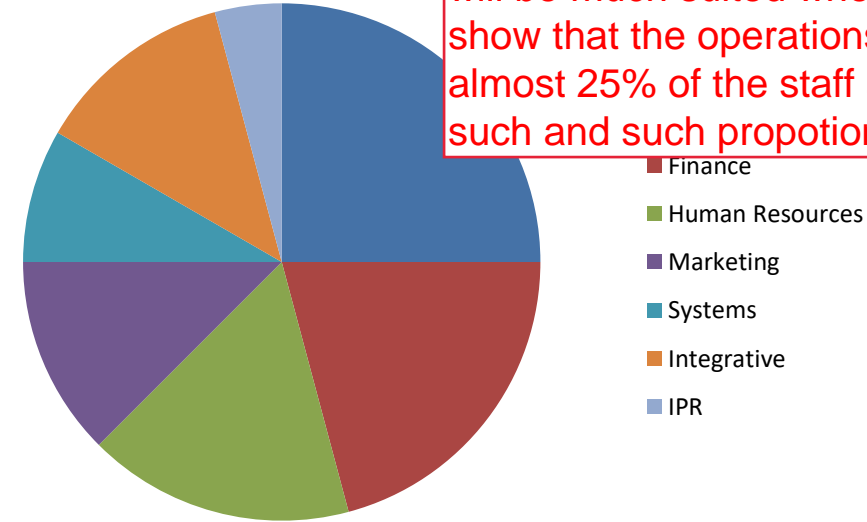
It is easier to
compare bars
(in a bar chart)
than areas (in
a pie chart)

Easier to compare bars than areas

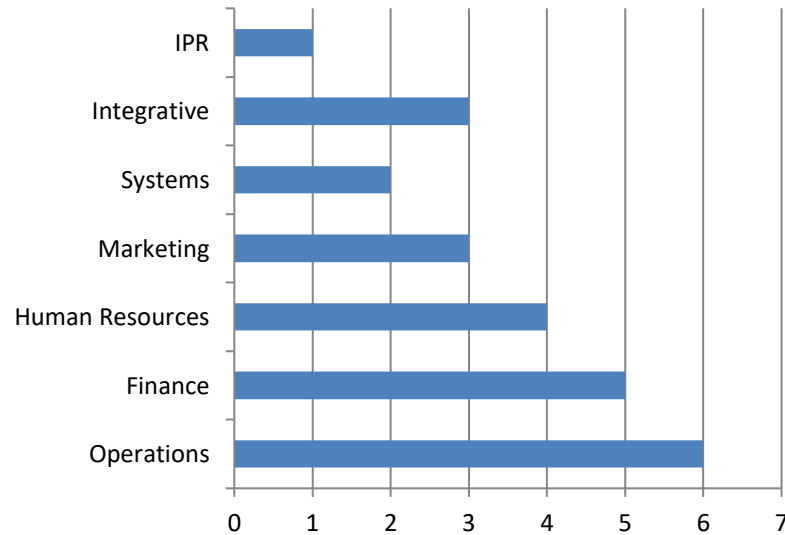
Frequency
distribution
table

Number of faculty (area wise)

Operations	6
Finance	5
Human Resources	4
Marketing	3
Systems	2
Integrative	3
IPR	1



But Pie charts as can be seen that they represent only proportions and not the exact value of data. So thus Pie chart in here is not suitable. It will be much suited where we need to show that the operations department holds almost 25% of the staff and the rest hold such and such proportions.



Which is better?

Bar or Pie? **Bar chart**

Bar chart illustrates the entire data in the exact manner in this case

It is always easier to see from the bar chart that which class is bigger by looking at the bars rather than from the areas in the pie chart

Exercise

(Interpret a bar chart and a pie chart?)

Ask 20 students their mother tongue. Interpret a bar chart and a pie chart?

Bar chart - unless we generalise that these many percentage of students use this and this mother tongue

The pay package given to 50 MBA students are available. Interpret a bar chart and a pie chart?

Bar chart if we are talking only about the exact packages.
but we generalise and consider a range in which different packages fall then

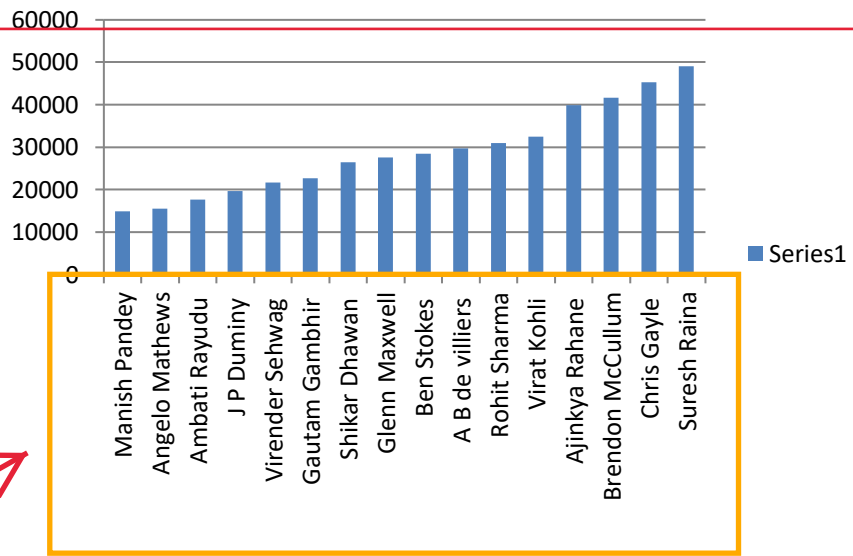
The colour of the shirt worn by 50 students is available.

The specializations taken by 40 second year MBA students

The number of students who start their own companies in the last 10 years

Only bar chart,
as it is a time
series data so
we need both
the X and Y
Axis here

We have used same colour for th bars in the graph only because we have just 1 variable that is name of the players.
But if we have more than 1 variable then we would show the bars in 2 different colours



The area principle

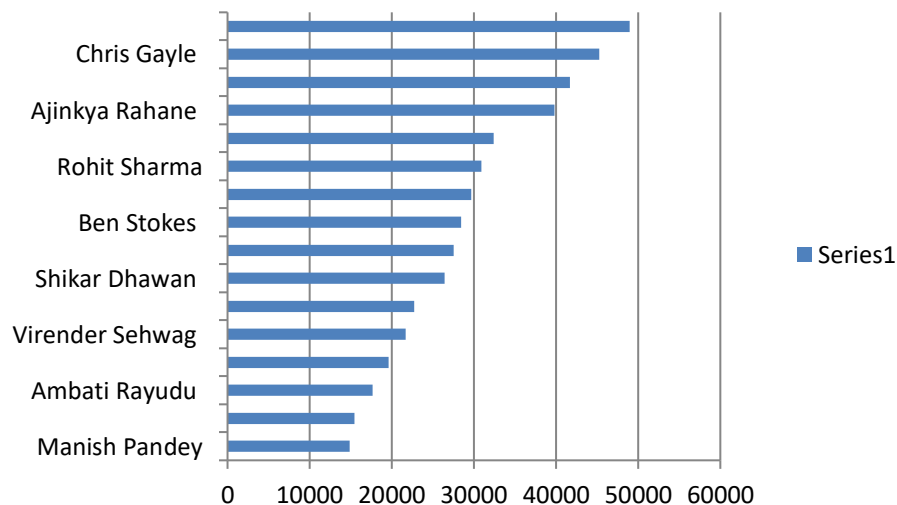
- Primarily means that the width of the bar is consistent throughout the bar chart even though the length varies. It was more relevant when the graphs were drawn with hand. But now with computer it is already been taken care of.
- It also means that the spacing b/w the bars is consistent so that it is pleasing to the eye.

Horizontal V/S Vertical bar chart :-

-This is a disadvantage of the vertical bar chart as it is diff to read the names here compared to the horizontal bar chart.

- Data is much easily interpreted in the horizontal bar chart compared to the vertical bar cahrt

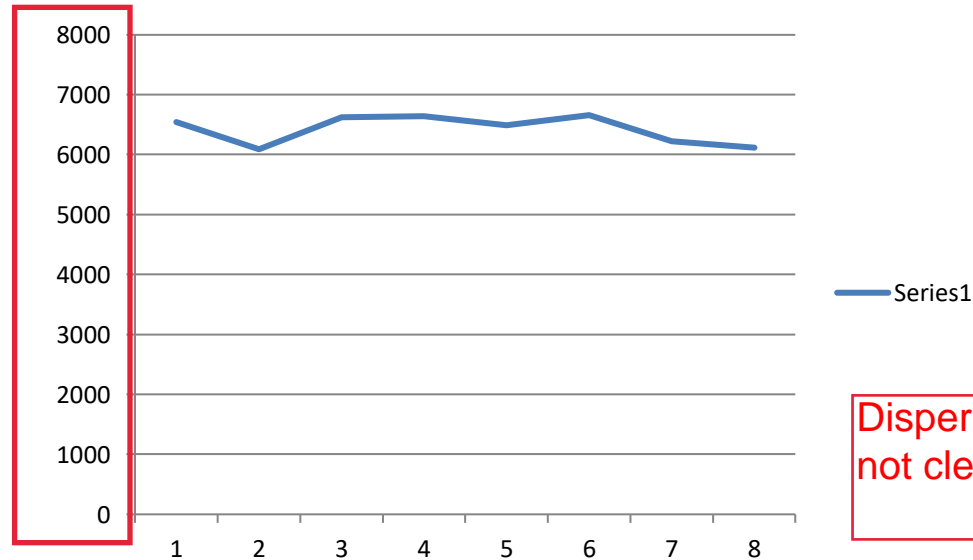
Area occupied by a part of the graph should correspond to the amount of data it represents



6543
6087
6619
6643
6489
6653
6222
6114

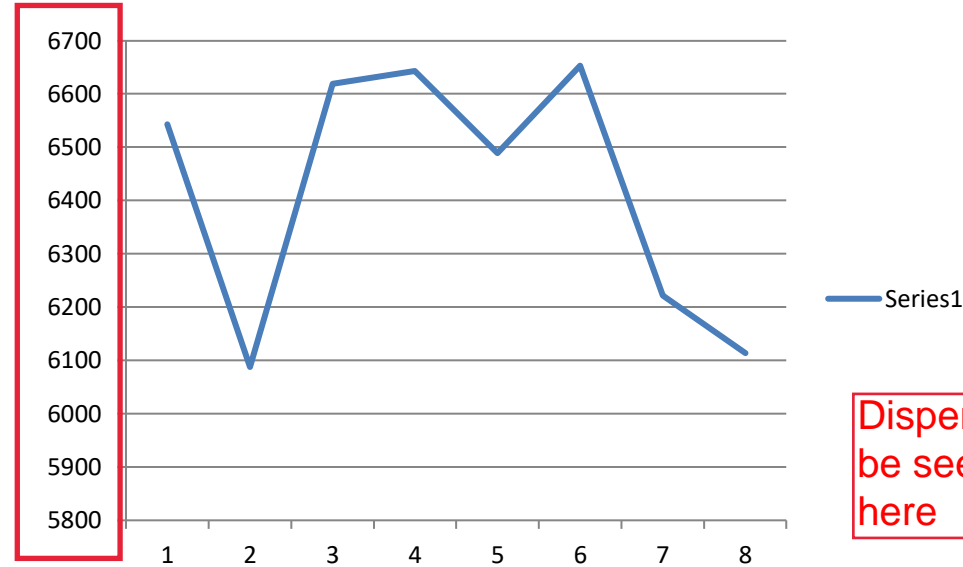
Demand data

Graph with a narrower Y axis, thus we see an almost constant graph



Dispersion is not clear

Graph with a wider Y axis, thus we see such spread



Dispersion can be see clearly here

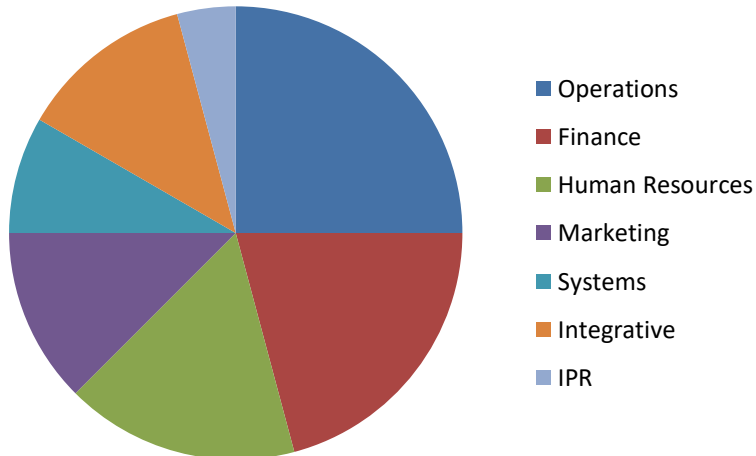
Both Graphs made on the same data of Deamnd as given in table

Types of Pie charts

MBA faculty

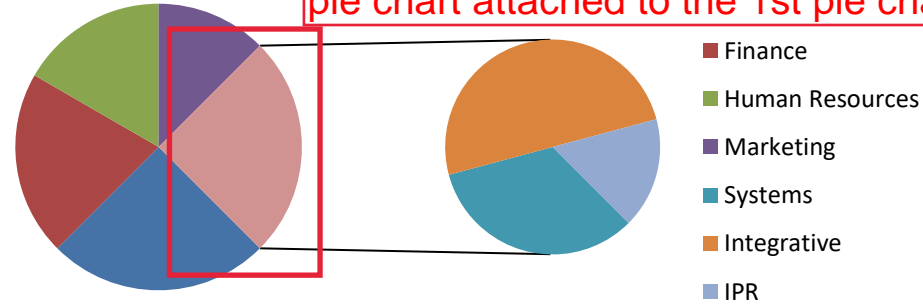
Operations	6
Finance	5
Human Resources	4
Marketing	3
Systems	2
Integrative	3
IPR	1

Original Pie chart

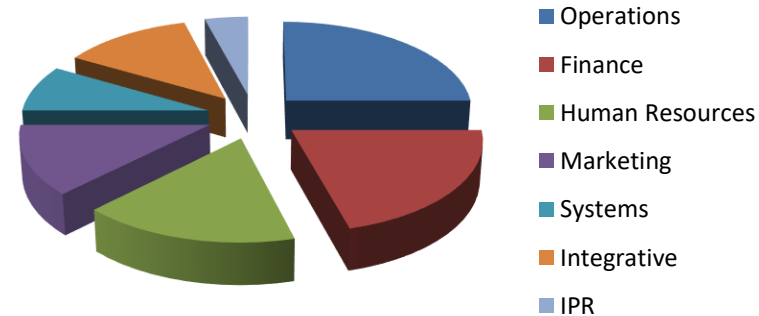


A cleaner looking pie chart.

- As can be seen the 3 different variables are clubbed together in the 1st pie and then show separately in 2nd pie chart attached to the 1st pie chart



3-D Pie cchart

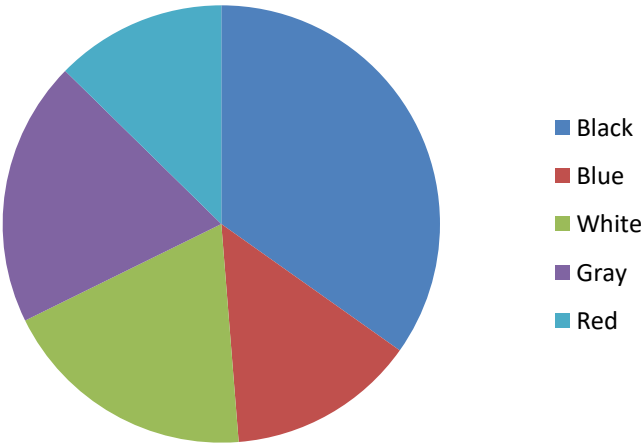
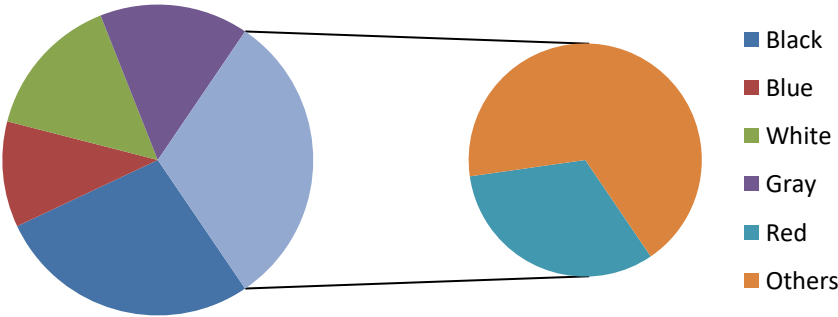
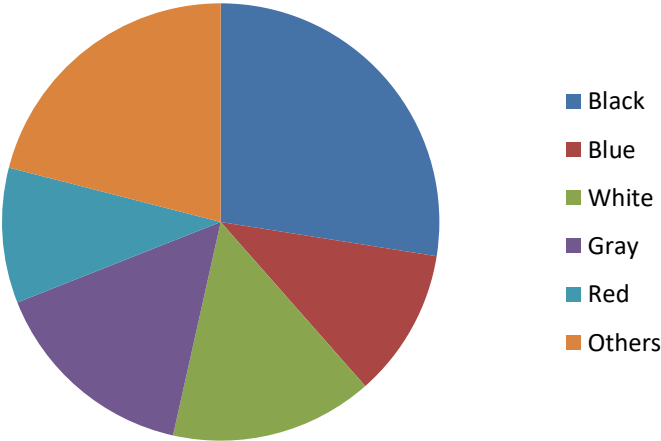


Same as the Old pie but in 3D. But this is not much used as the clarity is even less than the original pie chart.

- In old pie chart we can clearly see that operations (Blue) is 1/4 of the entire pie but we cannot see the same in the 3D Pie chart

Colour of cars parked

Black	55
Blue	22
White	30
Gray	31
Red	20
Others	42



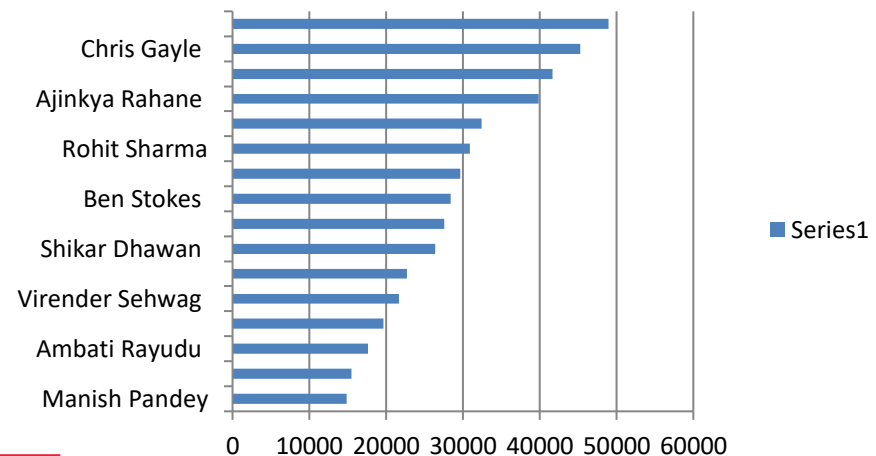
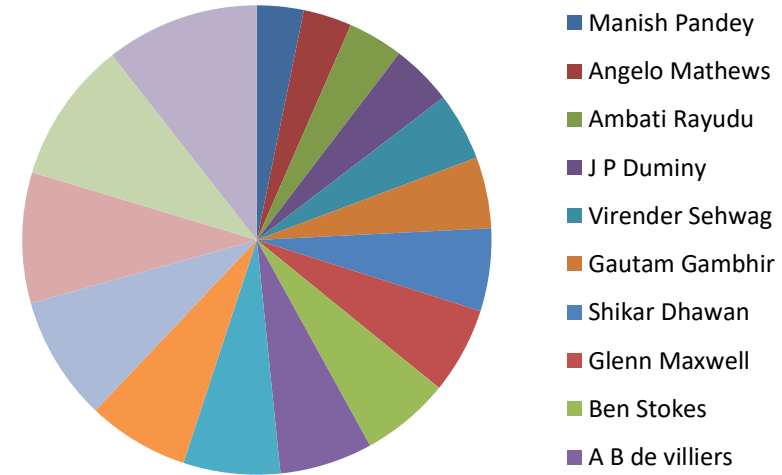
Lecture 5

Summary of categorical data

Summaries of categorical data

MODE

- Most common category
 - Category with the highest frequency
 - Longest bar in a bar chart
 - Largest slice of area in a pie chart
 - First category shown in a Pareto chart
-
- Sometimes bi modal or multi modal



Which player represents the mode?

Is it bimodal? No, bec we do not find 2 bars of the same lenth

MEDIAN of an ordinal variable

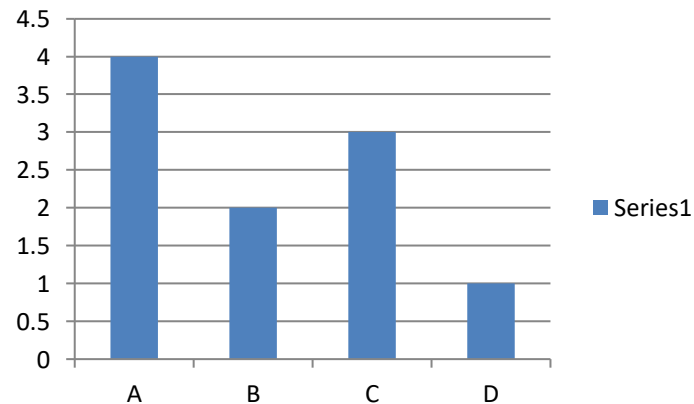
But there is an order in the ordinal data types, and median depends on the order

- Data has to be put in order
- Category of the middle observation when ordered
- Odd number Vs even number

In the Ordinal data we can calculate both Mode and median but when the data is Categorical/ Nominal then we can only calculate the Mode.

Grades of 10 students in Operations Management
A, C, B, A, A, C, B, D, A, C

A	4
B	2
C	3
D	1



Mode is A

Mode is the longest bar

Median is B

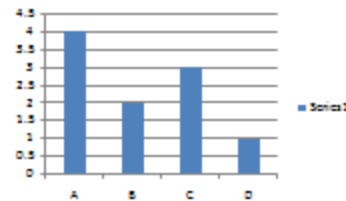
10 observations. -> so the even median = $10 + 1/2 = 11/2 = 5.5$ = 5th and 6th term. Thus we see that both 5th and 6th term fall under grade B so the median is B

MEDIAN of an ordinal variable

- Data has to be put in order
- Category of the middle observation when ordered
- Odd number Vs even number

Grades of 10 students in Operations Management
A, C, B, A, A, C, B, D, A, C

A	4
B	2
C	3
D	1



Mode is A
Median is B

Relate mode and median to the following?

Ask 20 students their mother tongue. Categorical data, thus only Mode

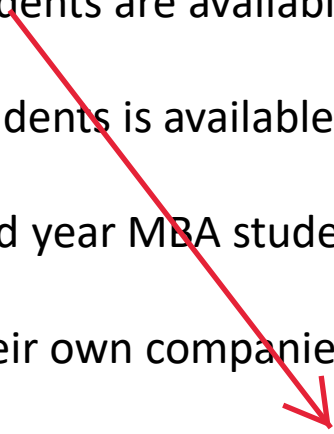
The pay package given to 50 MBA students are available.

The colour of the shirt worn by 50 students is available. Categorical data, thus only Mode

The specializations taken by 40 second year MBA students Categorical data, thus only Mode

The number of students who start their own companies in the last 10 years

Categorical data, thus only Mode



if we can bring in a range of salaries and arrange the range in an order then we can also find the median.
- Also the range with highest frequency will be the mode

Discussion

Describing categorical data

Bar chart or Pie chart?

- Proportion of men and women students in a class PC

- Number of different types of defects in manufacturing BC

- Number of visits in a website on 5 days in a week BC - Time series data

- Number of journal publications of faculty of a department BC

- .

- Fours and sixes hit by a batsman out of his total career score PC

- Number of customers rating a hotel service as VG, G and poor PC

These can also take a bar chart/ Pie chart depending on how you present the data. Wether you generalise it and show propotions out of it then use a Pie chart otherwise use a bar chart

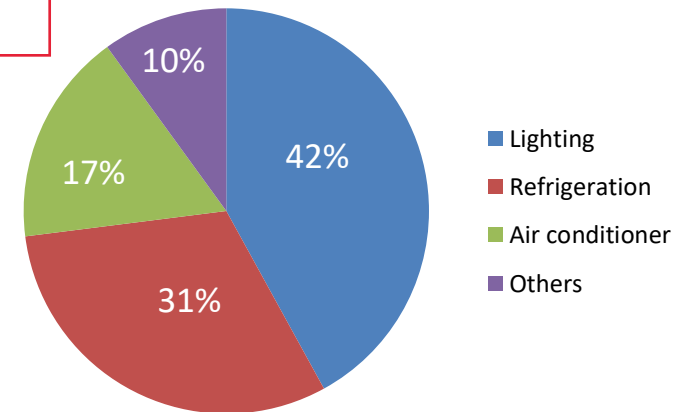
True or false

- Charts are better than tables to summarize categorical data
- The frequency is the money value of the observations in a group
- We use bar charts to show proportions and a pie to show the actual numbers for an categorical variable
- It is important to write the variables in an order while making bar chart for ordinal variable
- Share of purchases for saree, dress material and jeans in a ladies showroom

Question 1

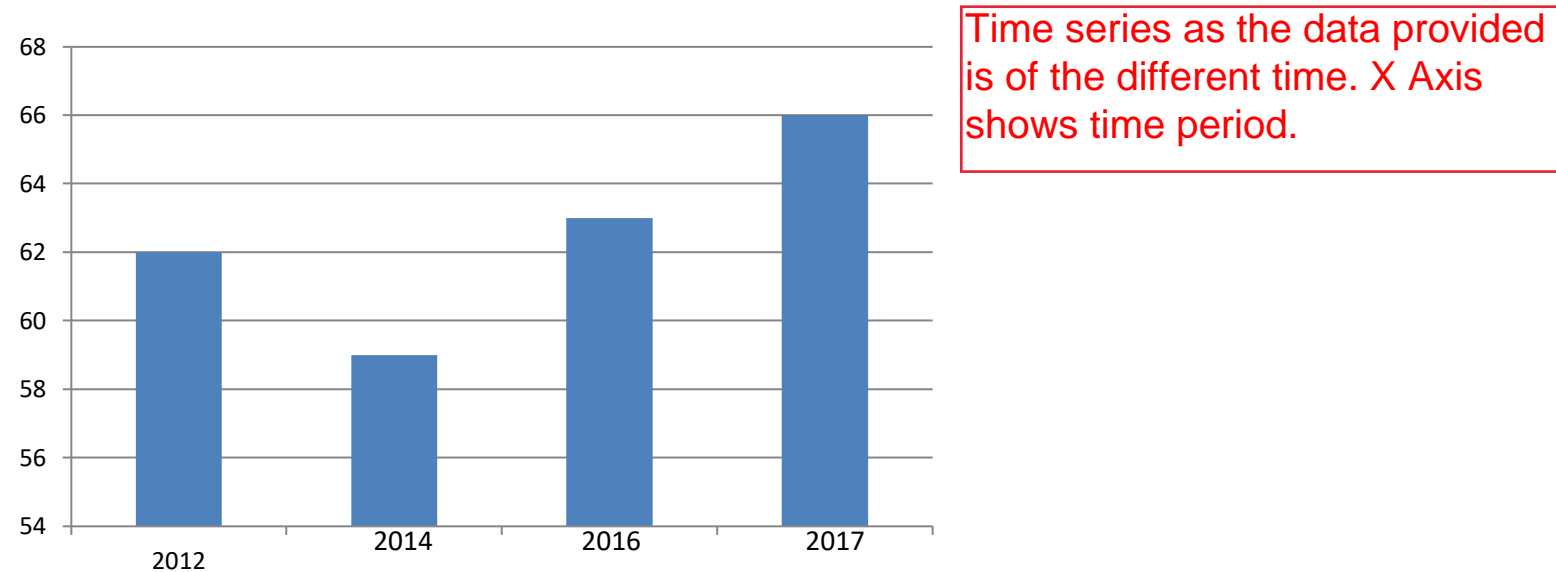
There is a move to replace incandescent bulbs with energy efficient bulbs. The following chart shows average energy consumption of 100 households. Would using energy efficient bulbs reduce energy consumption?

Yes, as we can see 42% is towards lighting thus the change in bulbs will account for a reduction in the energy consumption as lighting occupies significantly large are of the energy consumption.



Question 2

Explain the following figure that shows the number of publications of professors in a college. Is it a bar chart or time series?



Question 3

A categorical variable has only two values – Male and Female. Would you represent this with a bar chart or a pie chart or a frequency table?

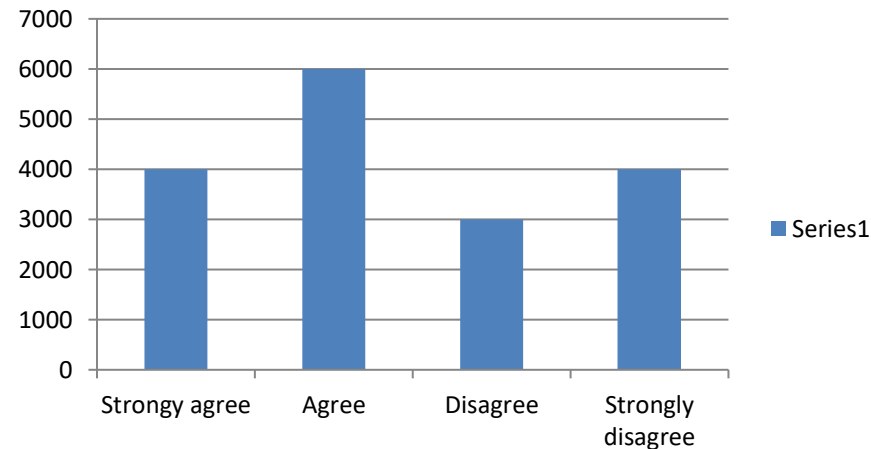
Answer: Frequency table

- We can use a bar chart to show the exact numbers.
- we can also use the pie chart if the data is given in proportions

Question 4

The following number of responses were received for a sensitive question

Strongly agree = 4000, Agree = 6000, Disagree = 3000, Strongly disagree = 4000. Is this ordinal or categorical. Would you use a pie chart to represent this data



Ordinal data

Pie chart is used to represent categorical data. So the Bar chart is fine.

Question 5

The sale of beverages in a shop in a week is given below

No.	Brand	Company	Sale
1	Mirinda	Pepsi	350
2	Maaza	Coke	600
3	Slice	Pepsi	200
4	Frooti	Parle	500
5	Fizz	Parle	250
6	Tropicana	Pepsi	300
7	Tang	Cadbury	180

Figures are imaginary

Share is indicative of using a bar chart

Maybe no, as the shop could also sell other brands

1. Does the table have a row of every case of soft drinks sold?
2. Prepare a chart that represents share of each brand? PC, by calculating the proportions
3. Prepare a chart to represent share of each company? How can you use the previous chart? PC, by calculating the proportions. We can Aggregate the values from the previous chart and then can use that
4. Prepare a chart presenting the amount of each brand sold?

BC, as we will represent the actual sale figures