Lecture 02:
Summarizing and Graphing Data

PKN STAN: Class 5-37 & 5-38
Lecturer: Erika Siregar, SST, M.S.

Review

1. What is population?
2. What is sample?
3. What is sampling?
4. What is data?
5. 2 types of data
6. 4 Levels of data measurement

Today’s Agenda

- Frequency Table
- Histogram
- Scatter Plot

Summarizing & Graphing: Get The Sense

SalesKey	DateKey	channelKey	StoreKey	ProductKey	PromotionKey	CurrencyKey	UnitCost	UnitPrice	SalesQuantity	ReturnQuantity	ReturnAmount	DiscountQuantity	DiscountAmount	TotalCost	SalesAmount	ETLLoadID	LoadDate	UpdateDate
1	00:00.0	1	209	956	10	1	91.05	198	8	0	0	1	33.6	728.4	1544.4	1	00:00.0	00:00.0
2	00:00.0	4	308	766	2	1	10.15	19.9	4	0	0	1	0.995	40.6	78.605	1	00:00.0	00:00.0
3	00:00.0	1	156	1175	11	1	203.03	410	9	0	0	3	61.5	1881.27	3628.5	1	00:00.0	00:00.0
4	00:00.0	2	306	1429	10	1	132.9	289	8	0	0	1	57.8	1063.2	2254.2	1	00:00.0	00:00.0
5	00:00.0	2	306	1133	10	1	144.52	436.2	24	0	0	3	261.72	3468.48	10207.08	1	00:00.0	00:00.0
6	00:00.0	3	200	2365	3	1	183.94	399.99	36	0	0	10	399.99	6621.84	13999.65	1	00:00.0	00:00.0
7	00:00.0	4	310	1016	5	1	68.06	148	6	0	0	2	44.4	408.36	843.6	1	00:00.0	00:00.0
8	00:00.0	2	307	138	15	1	229.93	499.99	9	0	0	1	99.998	2069.37	4399.912	1	00:00.0	00:00.0
9	00:00.0	2	199	1731	12	1	33.32	72.45	24	0	0	5	36.225	799.68	1702.575	1	00:00.0	00:00.0
10	00:00.0	4	310	497	24	1	50.47	99	18	0	0	4	79.2	908.46	1702.8	1	00:00.0	00:00.0
11	00:00.0	2	199	1825	2	1	16.31	32	4	0	0	0	0	65.24	128	1	00:00.0	00:00.0
12	00:00.0	1	119	543	1	1	116.75	229	10	0	0	0	0	1167.5	2290	1	00:00.0	00:00.0
13	00:00.0	1	171	739	3	1	78.19	236	12	0	0	0	0	938.28	2832	1	00:00.0	00:00.0
14	00:00.0	1	16	1269	13	1	25.47	49.96	13	0	0	1	9.992	331.11	639.488	1	00:00.0	00:00.0
15	00:00.0	2	199	1788	1	1	21.92	43	10	0	0	0	0	219.2	430	1	00:00.0	00:00.0
16	00:00.0	1	183	2082	2	1	71.37	139.99	9	0	0	2	13.998	642.33	1245.911	1	00:00.0	00:00.0
17	00:00.0	1	161	1655	11	1	96.08	289.99	9	0	0	0	0	864.72	2609.91	1	00:00.0	00:00.0
18	00:00.0	2	199	1724	3	1	28.55	56	3	0	0	2	11.2	85.65	156.8	1	00:00.0	00:00.0
19	00:00.0	1	292	519	1	1	205.09	619	12	0	0	0	0	2461.08	7428	1	00:00.0	00:00.0
20	00:00.0	1	263	47	1	1	76.45	149.95	20	0	0	0	0	1529	2999	1	00:00.0	00:00.0
21	00:00.0	1	59	43	13	1	91.95	199.95	26	1	199.95	1	39.99	2298.75	5158.71	1	00:00.0	00:00.0
22	00:00.0	1	155	189	11	1	58.36	126.9	9	0	0	2	12.69	525.24	1129.41	1	00:00.0	00:00.0
23	00:00.0	2	199	703	20	1	69.25	209	9	0	0	0	0	623.25	1881	1	00:00.0	00:00.0
24	00:00.0	2	199	572	11	1	87.37	190	9	1	190	4	38	698.96	1672	1	00:00.0	00:00.0
25	00:00.0	1	108	2351	1	1	183.94	399.99	10	0	0	0	0	1839.4	3999.9	1	00:00.0	00:00.0
26	00:00.0	1	144	2226	4	1	61.17	119.99	6	0	0	4	95.992	367.02	623.948	1	00:00.0	00:00.0
27	00:00.0	1	186	2483	1	1	160.95	350	10	0	0	0	0	1609.5	3500	1	00:00.0	00:00.0
28	00:00.0	1	4	1368	22	1	18.48	40.19	13	0	0	0	0	240.24	522.47	1	00:00.0	00:00.0
29	00:00.0	1	3	1290	22	1	121.45	366.55	6	0	0	0	0	728.7	2199.3	1	00:00.0	00:00.0
30	00:00.0	1	302	363	14	1	321.44	699	13	0	0	1	104.85	4178.72	8982.15	1	00:00.0	00:00.0
31	00:00.0	3	200	1759	11	1	34.75	104.89	9	0	0	4	20.978	312.75	923.032	1	00:00.0	00:00.0
32	00:00.0	1	225	346	10	1	303.05	659	8	0	0	4	527.2	2424.4	4744.8	1	00:00.0	00:00.0
33	00:00.0	1	178	824	20	1	6.07	11.9	18	0	0	3	1.785	109.26	212.415	1	00:00.0	00:00.0
34	00:00.0	1	214	1494	1	1	95.65	208	10	0	0	0	0	956.5	2080	1	00:00.0	00:00.0
35	00:00.0	1	55	1332	1	1	71.37	139.99	10	1	139.99	0	0	642.33	1399.9	1	00:00.0	00:00.0
36	00:00.0	2	199	589	1	1	321.44	699	10	0	0	0	0	3214.4	6990	1	00:00.0	00:00.0
37	00:00.0	1	263	1103	1	1	164.63	358	10	0	0	0	0	1646.3	3580	1	00:00.0	00:00.0
38	00:00.0	4	308	2060	12	1	48.43	94.99	12	0	0	2	18.998	581.16	1120.882	1	00:00.0	00:00.0
39	00:00.0	1	196	1051	21	1	155.43	338	12	0	0	3	101.4	1865.16	3954.6	1	00:00.0	00:00.0
40	00:00.0	1	83	1946	13	1	412.42	899	13	0	0	4	719.2	5374.46	10967.8	1	00:00.0	00:00.0

Summarizing is meant to get the general idea of the dataset

General Idea = Important Characteristics of Data

1. **Center**
   a. Where the middle of the data set is located
   b. Secara general, datanya gimana sih? Mean, median, modus
2. **Variation**
   a. Amount that the data values vary.
   b. Datanya mirip2 atau significantly variatif?
3. **Distribution**
   a. The **shape of the spread** of data over the range of values (such as bell-shaped, uniform, or skewed).
   b. Kalau digambarkan dalam bentuk histogram, bentuknya bagaimana?
4. **Outliers**
   Values that lie very far away from the vast majority of other sample values. → ada pencilan?
5. **Time**
   Changing characteristics of the data over time. → **trend**?

1. Frequency table

2. Graph/Visual

   Histogram
   Scatter plot
   Bar graph
   Pie Chart
   etc

Frequency Table

Why frequency table?

1. Fast summary
2. Easier to create a graph

Method of Travelling	Number of children
Walking	8
Car	9
Bus	4
Cycle	5
Train	1
Taxi	3

Single value

Table 2-2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

interval

Terminology in Frequency Table

Table 2-2

IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Lower class limit

Upper class limit

width

Class Boundaries

the numbers used to separate classes, but without the gaps created by class limits

Table 2-2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Class Boundaries

49.5
69.5
89.5
109.5
129.5

Class Midpoints

the values in the middle of the classes and can be found by adding the lower class limit to the upper class limit and dividing the sum by two

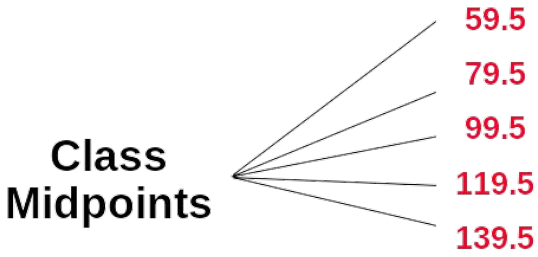
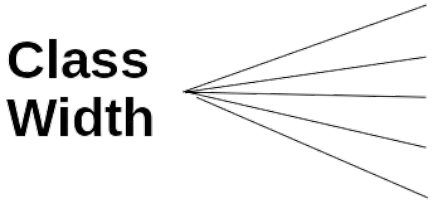


Table 2-2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Class Width

the difference between two consecutive lower class limits



20
20
20
20
20

Table 2-2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

How to Construct a Frequency Table

1. **Arrange/sort the data**
2. **Decide how many classes we want to build**
   a. **Recommended number of classes: between 5 and 20.**
   b. **Or use the Sturges' rule: k = 1 + 3.322 log n. → n = jumlah data**
3. **Calculate the class width (round up).**

$$\text{class width} \approx \frac{\text{(maximum value) – (minimum value)}}{\text{number of classes}}$$

4. **Starting point: Choose the minimum data value or a convenient value below it as the first lower class limit.**
5. **Using the first lower class limit and class width, proceed to list the other lower class limits.**
6. **List the lower class limits in a vertical column and proceed to enter the upper class limits.**
7. **Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to get the frequency.**

Table 2-2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Let’s Get into practice

Table 2-1 Full IQ Scores of Low Lead Group and High Lead Group															
Low Lead Level (Group 1)															
70	85	86	76	84	96	94	56	115	97	77	128	99	80	118	86
141	88	96	96	107	86	80	107	101	91	125	96	99	99	115	106
105	96	50	99	85	88	120	93	87	98	78	100	105	87	94	89
80	111	104	85	94	75	73	76	107	88	89	96	72	97	76	107
104	85	76	95	86	89	76	96	101	108	102	77	74	92		
High Lead Level (Group 3)															
82	93	85	75	85	80	101	89	80	94	88	104	88	88	83	104
96	76	80	79	75											

- Number of classes = 5
- Find the min value = 50
- Find the max value = 141

Let’s Get into practice

Table 2-1 Full IQ Scores of Low Lead Group and High Lead Group

Low Lead Level (Group 1)															
70	85	86	76	84	96	94	56	115	97	77	128	99	80	118	86
141	88	96	96	107	86	80	107	101	91	125	96	99	99	115	106
105	96	50	99	85	88	120	93	87	98	78	100	105	87	94	89
80	111	104	85	94	75	73	76	107	88	89	96	72	97	76	107
104	85	76	95	86	89	76	96	101	108	102	77	74	92		
High Lead Level (Group 3)															
82	93	85	75	85	80	101	89	80	94	88	104	88	88	83	104
96	76	80	79	75											

- Number of classes = 5
- Find the min value = 50
- Find the max value = 141

$$\text{class width} \approx \frac{\text{(maximum value) – (minimum value)}}{\text{number of classes}}$$

Class width = (141 - 50) / 5 = 18,..
⇒ roundup = **19**

Build the Frequency Table

Table 2-1 Full IQ Scores of Low Lead Group and High Lead Group

Low Lead Level (Group 1)															
70	85	86	76	84	96	94	56	115	97	77	128	99	80	118	86
141	88	96	96	107	86	80	107	101	91	125	96	99	99	115	106
105	96	50	99	85	88	120	93	87	98	78	100	105	87	94	89
80	111	104	85	94	75	73	76	107	88	89	96	72	97	76	107
104	85	76	95	86	89	76	96	101	108	102	77	74	92		
High Lead Level (Group 3)															
82	93	85	75	85	80	101	89	80	94	88	104	88	88	83	104
96	76	80	79	75											

IQ Score	Freq	Tally
50 - 69	2	II
70 - 89	33	IIIII IIIII II
90 - 109	35	IIIII IIIII IIII
110 - 129	7	IIIII
130 - 149	1	I

Things to Calculate from Frequency Table

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}} \times 100\%$$

Table 2-2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Total Frequency = 78

Table 2-4 Relative Frequency Distribution of IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2.6%
70–89	42.3%
90–109	44.9%
110–129	9.0%
130–149	1.3%

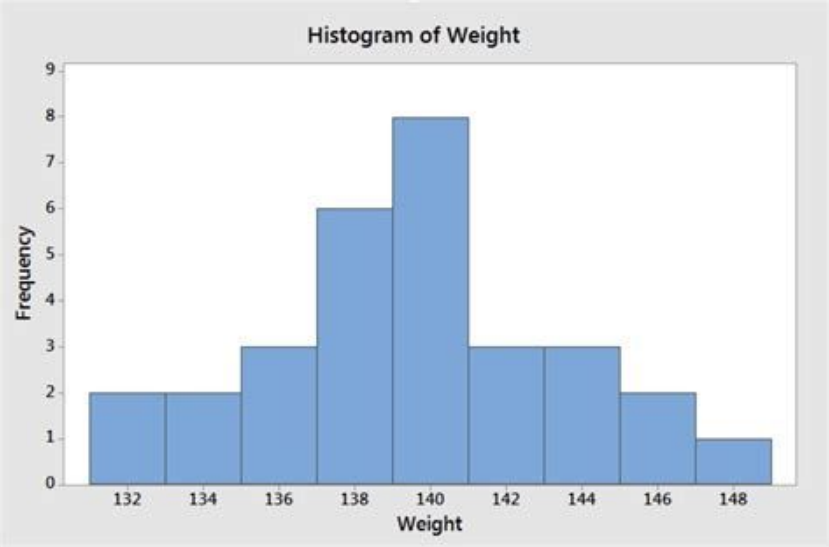
* $2/78 \times 100 = 2.6\%$

Complete Frequency Table

No	IQ Score	Frequency	Cumulative Frequency	Relative Frequency
1.	50 - 69	2	2	2.6%
2.	70 - 89	33	35	42.3%
3.	90 - 109	35	70	44.9%
4.	110 - 129	7	77	9.0%
5.	130 - 149	1	78	1.3%

Try It with R

Histogram

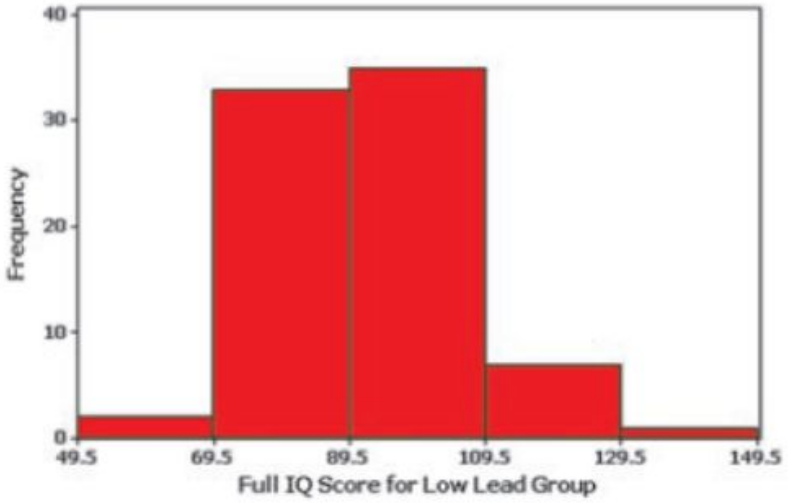


a visual tool used for analyzing the **shape of the distribution** of the data.

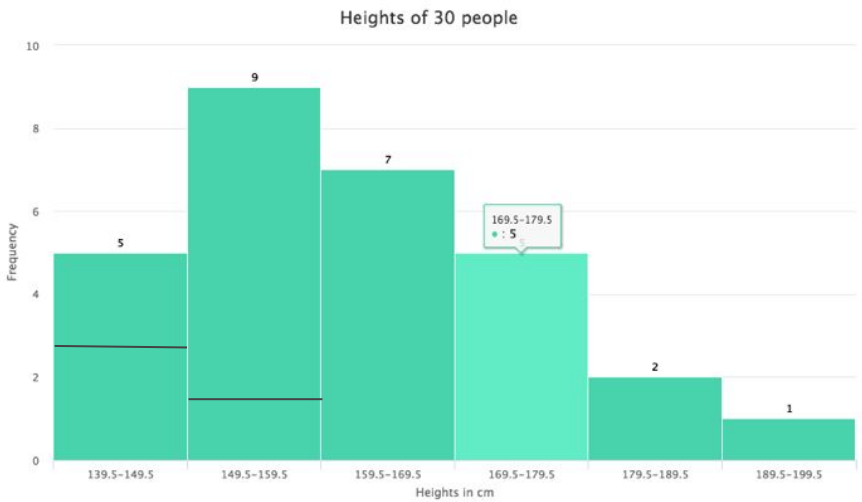
Histogram is a graphic version of frequency table

Table 2-2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1



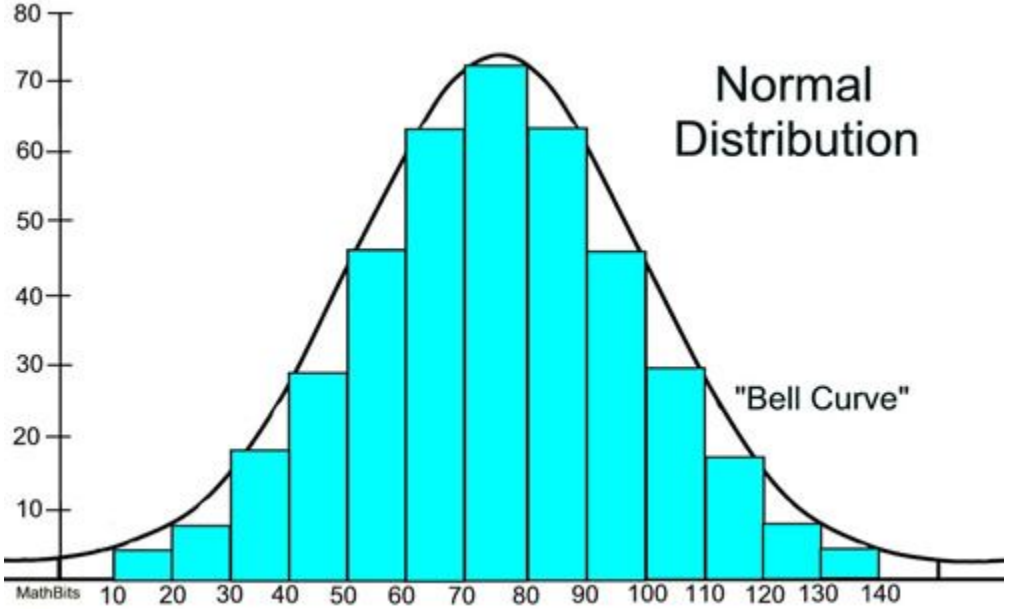
Histogram Characteristics



- consisting of bars of equal width
- No **gaps** between the bar
- Horizontal scale = the classes of quantitative data values
  - Classes could be both single values or intervals.
- Horizontal label could be one of the following:
  - Class boundaries
  - Class midpoints
  - Lower class limits (introduces a small error)
- Vertical scale = frequency

Try Histogram with R

Histogram of Normal Distribution



Characteristics:

1. Bell-Shaped distribution
2. The frequencies increase to a maximum, and then decrease.
3. symmetry, with the left half of the graph **roughly** a mirror image of the right half.

Scatter Plot

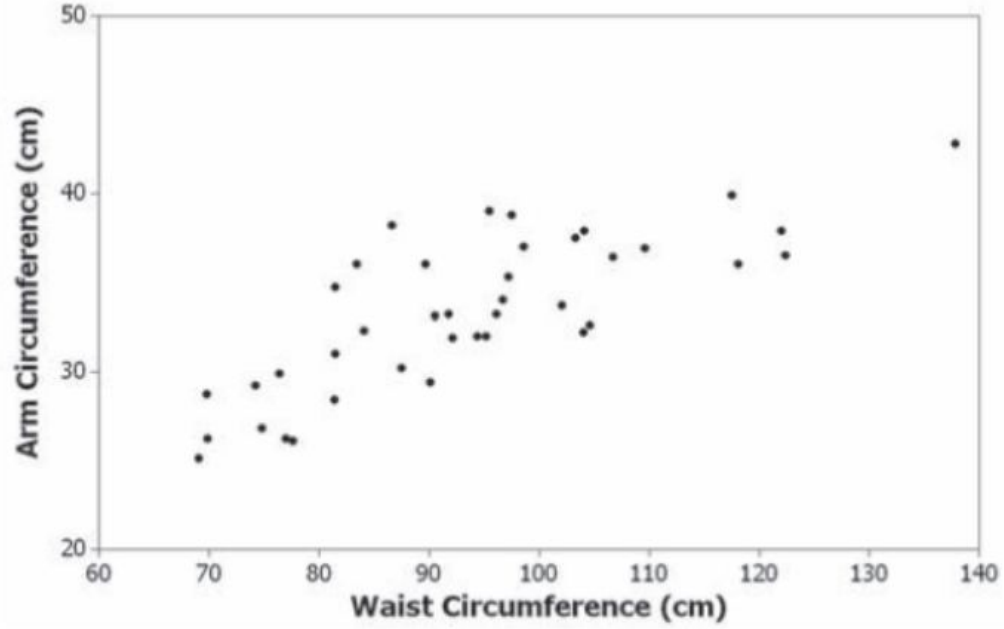


Figure 2-6 Waist Circumference and Arm Circumference in Males

Time Series Graph

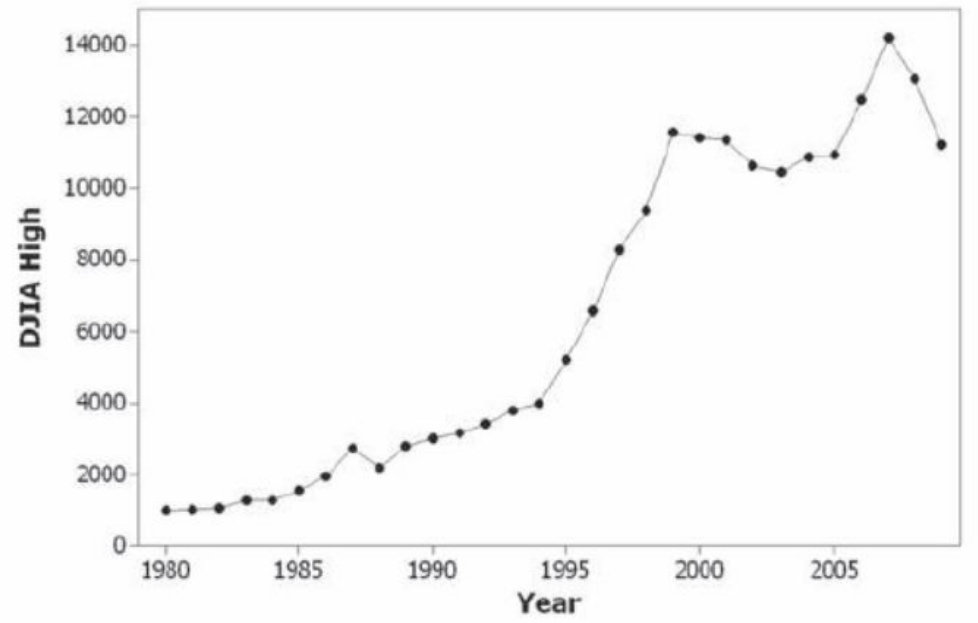


Figure 2-9 Dow Jones Industrial Average

Dot Plot

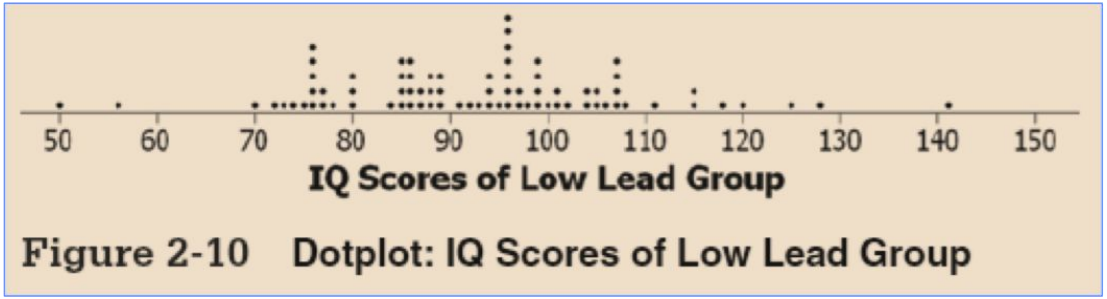


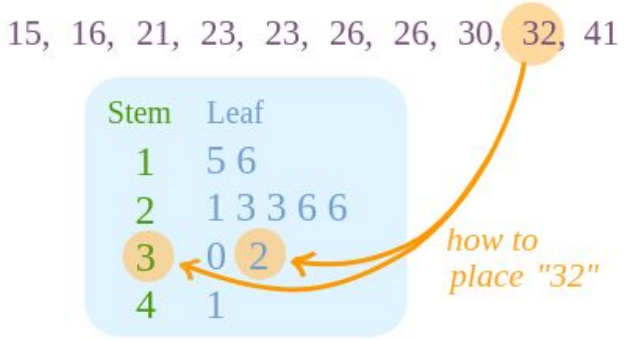
Figure 2-10 Dotplot: IQ Scores of Low Lead Group

each data value is plotted as a point (or dot) along a scale of values. Dots representing equal values are stacked.

Stemplot

Each values are separated into 2 parts:

1. The stem → leftmost digits → arrange vertically
2. The leaf → rightmost digits → arrange horizontally



5	06 ← Lowest IQ scores are 50 and 56.
6	
7	0234566666778
8	00045555666677888999
9	1234445666666677899999
10	01124455677778
11	1558 ← These IQ scores are 111, 115, 115, and 118.
12	058
13	
14	1

Bar Graph

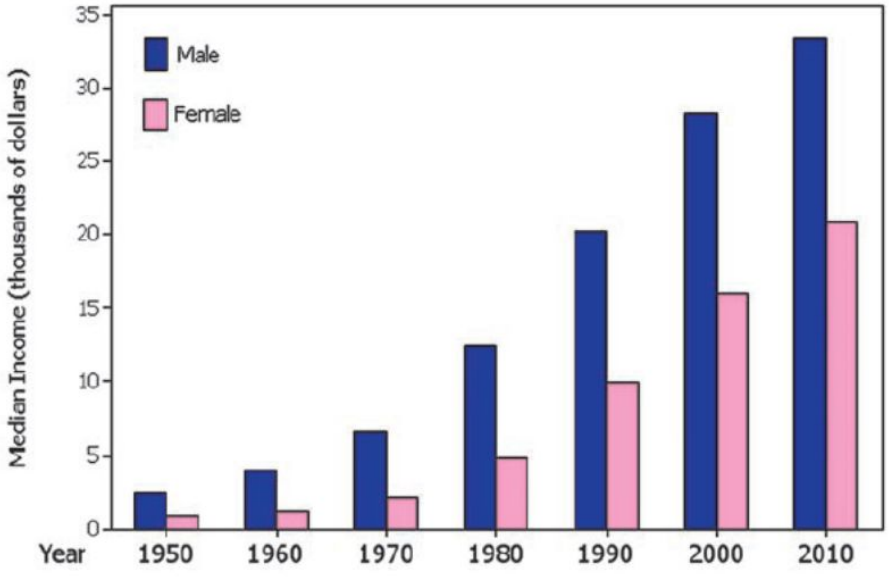
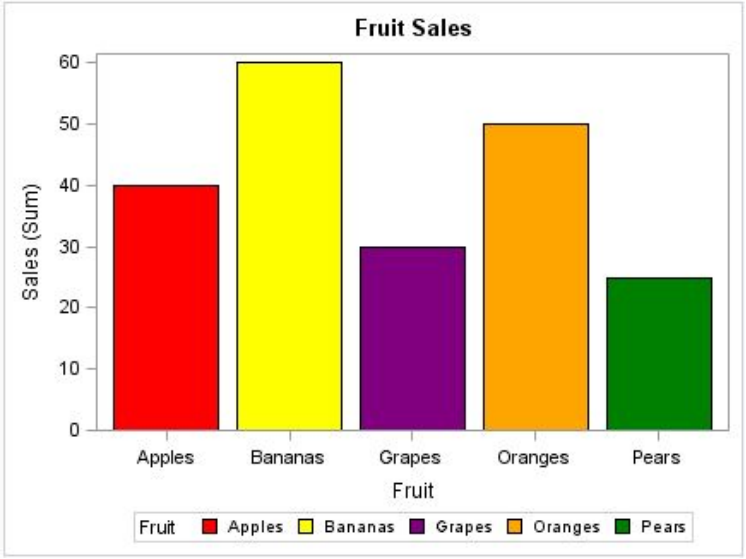
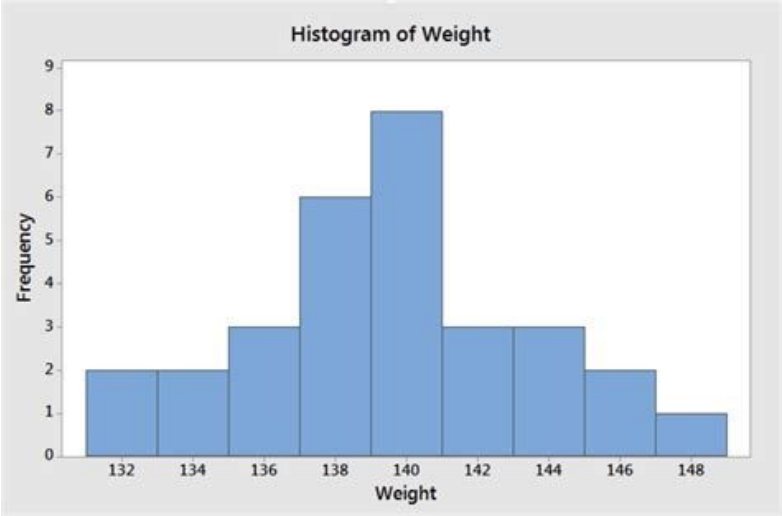


Figure 2-11 Multiple Bar Graph: Median Income by Gender

- Horizontal scale = **categories** of qualitative data.
- Vertical scale = **frequencies** or **Values**

Histogram vs bar chart

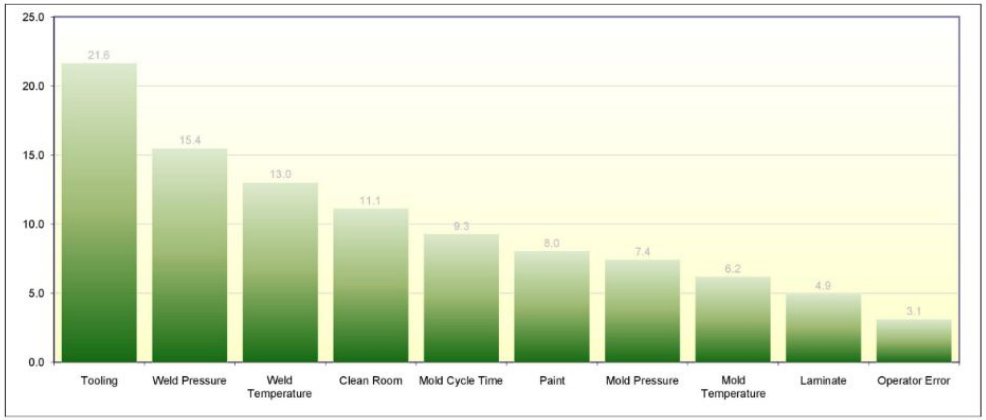




What's the key difference between the 2 graphs above?

1. Histogram, x = quantitative
2. Histogram, x = interval
3. Bar graph, x = category
4. Histogram, y = pasti frequency, bar graph = frequency or values

Pareto Chart



A descending-order bar graph

Pie Chart

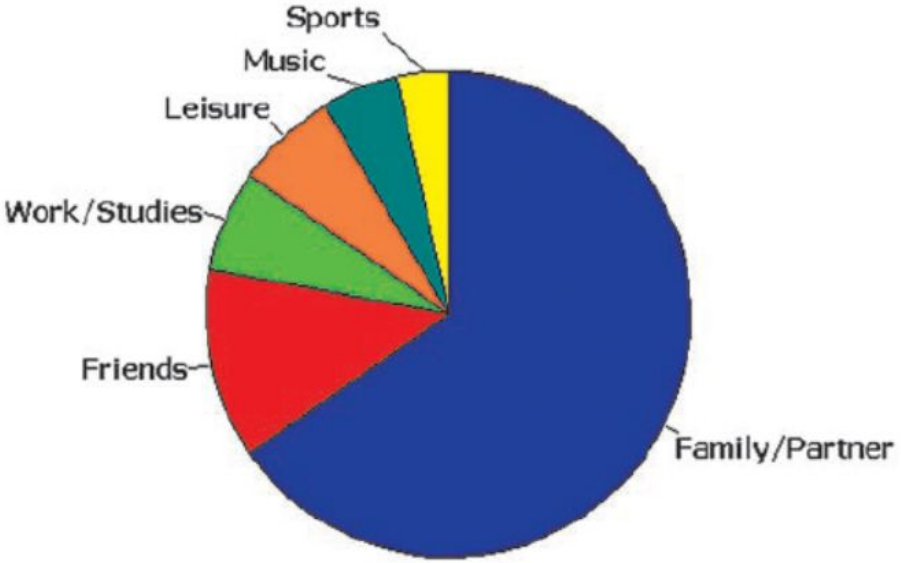
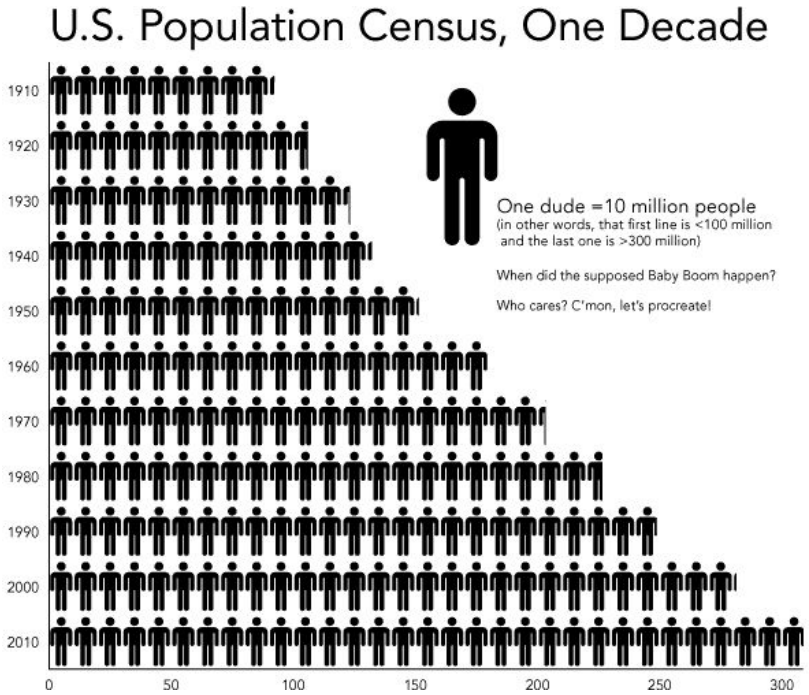


Figure 2-13 Pie Chart: What Contributes Most to Happiness?

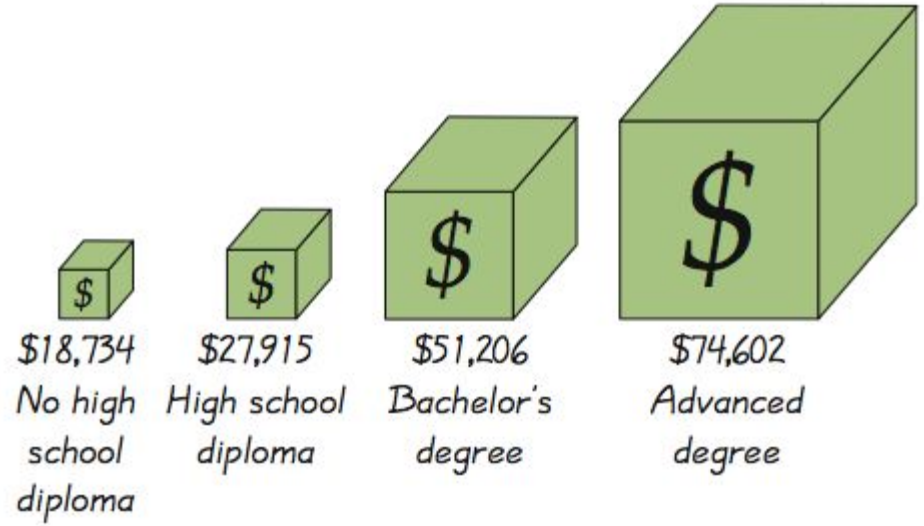
Each slice represents percentage or share

Pictograph



Representing values through objects

The Risk of Using Pictograph



1. Misleading
2. Not easy to gain insight quickly.

Misleading. Depicts one-dimensional data with three-dimensional boxes. Last box is 64 times as large as first box, but income is only 4 times as large.