

Data Handling and Business Intelligence

Assessment two

STUDENT: MADALINA LARISA VASILE, 2022389

Lecture : Anupam Mazumdar

Module : Data Handling And Business Intelligence

PART 1

✚ Using the "Excel" program :

The basic requirement of this assessment, "part one", is to calculate the crime rate per million population. The calculation must be made as follows: dividing the number of crimes in relation to the total population.

We must use one of the functions: IF, VLOOKUP, PIVOT TABLES or any other function we consider useful. If the formula is not clear to us, we can always find information directly using excel.

E.g :

- IF has the following requirements (checking if a certain condition is met): "`= IF (logical_test, value_if_true, value_if_false)`"
- VLOOKUP has the following requirements (looking for a certain value): "`= VLOOKUP (lookup_value, table_array, col_index_num, [range_lookup])`"

The first step is to download the "crime_rate" data set required in this assessment.

The second step is about what I want to know/found with excel functions. For example: I want to know the relationship between "x" and "y". The basic notion being the crimes, you have to choose the relationship between "CrimeRate" and anything else ("Education", "Man" [maybe even "Gender" but it is not a column named in the table, it can be discovered by a simple subtraction, between the total number of people minus the total number of men], "StateSize" [in a particular city], etc.).

Selected functions are: "VLOOKUP" & "PIVOT TABLE". The calculation of these functions can be done anywhere (there is empty space to initiate a future function) near the end of the initial table.

What will I write down in the empty box? Always put "=" before any function to get a result :

"= VLOOKUP", then (I can type by hand what I need or even choose by clicking with the mouse what I need to finish the function) I choose any number I want from the "CrimeRate" column (because that's where I look for the information), then (to find out exactly where that number of crimes chosen is) click on the whole area of the present table, then you have to choose the exact column where I want to find a result / where to analyze the crimes (put the column number, for example "StateSize" is the number "9" which I have to write down if I want to find results based on it), then I add "FALSE" (to which there is no explanation related to searches or related to the table, simply this is how the function should be written like we to have a valid result in the end).

ScreenShot :

1. In the image below, the "lookup value" was chosen in the "CrimeRate" column: 67.6

Excel Crime_rate_data 1 - Saved

Search (Alt + Q)

File Home Insert Draw Page Layout Formulas Data Review View Automate Help Editing

PivotTable Table Forms Pictures Shapes Recommended Charts Link New Comment

P7 $\times$ $\checkmark$ f_x =VLOOKUP(A7

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	CrimeRate	Youth	Region	Education	Expenditu	LabourFor	Males	MoreMale	StateSize	YouthUn	MatureUn	HighYouth	Wage	BelowWage					
2	45.5	135	0	12.4	69	540	965	0	6	80	22	1	564	139					
3	52.3	140	0	10.9	55	535	1045	1	6	135	40	1	453	200					
4	56.6	157	1	11.2	47	512	962	0	22	97	34	0	288	276					
5	60.3	139	1	11.9	46	480	968	0	19	135	53	0	457	249					
6	64.2	126	0	12.2	106	599	989	0	40	78	25	1	593	171					
7	67.6	128	0	13.5	67	624	972	0	28	77	25	1	507	206					
8	70.5	139	0	14.1	63	641	984	0	14	70	21	1	486	196					
9	73.2	143	0	12.9	66	537	977	0	10	114	35	1	487	166					
10	75	141	0	12.9	56	523	968	0	4	107	37	0	489	170					
11	78.1	133	0	11.4	51	599	1024	1	7	99	27	1	425	225					
12	79.8	142	1	12.9	45	533	969	0	18	94	33	0	318	250					
13	82.3	123	0	12.5	97	526	948	0	113	124	50	0	572	158					
14	83.1	135	0	13.6	62	595	986	0	22	77	27	0	529	190					
15	84.9	121	0	13.2	118	547	964	0	25	84	29	0	689	126					
16	85.6	166	1	11.4	58	521	973	0	46	72	26	0	396	237					
17	88	140	0	12.9	71	632	1029	1	7	100	24	1	526	174					
18	92.3	126	0	12.7	74	602	984	0	34	102	33	1	557	195					
19	94.3	130	0	13.3	128	536	934	0	51	78	34	0	627	135					
20	95.3	125	0	12	90	586	964	0	97	105	43	0	617	163					
21	96.8	151	1	10	58	510	950	0	33	108	41	0	394	261					
22	97.4	152	1	10.8	57	530	986	0	30	92	43	0	405	264					
23	98.7	162	1	12.1	75	522	996	0	40	73	27	0	496	224					
24	99.9	149	1	10.7	61	515	953	0	36	86	35	0	395	251					
25	103	177	1	11	58	638	974	0	24	76	28	0	382	254					
26	104.3	134	0	12.5	75	595	972	0	47	83	31	0	580	172					
27	105.9	130	0	13.4	90	623	1049	1	3	113	40	0	588	160					
28	106.6	157	1	11.1	65	553	955	0	39	81	28	0	421	239					
29	107.2	148	0	13.7	72	601	998	0	9	84	20	1	590	144					
30	108.3	126	0	13.8	97	542	990	0	18	102	35	0	589	166					
31	109.4	135	1	11.4	123	537	978	0	31	89	34	0	631	165					

Crime_data Column Label

Calculation Mode: Automatic Workbook Statistics

2. In the image below, the whole area of the table containing the information was chosen (ie without the header of the table in which the information is named): from A2 to N48 which is notable "A2: N48"

Excel Crime_rate_data 1 - Saved Search (Alt + Q)

File Home Insert Draw Page Layout Formulas Data Review View Automate Help Editing

PivotTable Table Forms Pictures Shapes Recommended Charts Link New Comment

P7 fx =VLOOKUP(A7,A2:N48)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	CrimeRateYouth	Region	Education	Expenditu	LabourFor	Males	MoreMale	StateSize	YouthUn	MatureUn	HighYouth	Wage	BelowWage						
2	45.5	135	0	12.4	69	540	965	0	6	80	22	1	564	139					
3	52.3	140	0	10.9	55	535	1045	1	6	135	40	1	453	200					
4	56.6	157	1	11.2	47	512	962	0	22	97	34	0	288	276					
5	60.3	139	1	11.9	46	480	968	0	19	135	53	0	457	249					
6	64.2	126	0	12.2	106	599	989	0	40	78	25	1	593	171					
7	67.6	128	0	13.5	67	624	972	0	28	77	25	1	507	206					
8	70.5	130	0	14.1	63	641	984	0	14	70	21	1	486	196					
9	73.2	143	0	12.9	66	537	977	0	10	114	35	1	487	166					
10	75	141	0	12.9	56	523	968	0	4	107	37	0	489	170					
11	78.1	133	0	11.4	51	599	1024	1	7	99	27	1	425	225					
12	79.8	142	1	12.9	45	533	969	0	18	94	33	0	318	250					
13	82.3	123	0	12.5	97	526	948	0	113	124	50	0	572	158					
14	83.1	135	0	13.6	62	595	986	0	22	77	27	0	529	190					
15	84.9	121	0	13.2	118	547	964	0	25	84	29	0	689	126					
16	85.6	166	1	11.4	58	521	973	0	46	72	26	0	396	237					
17	88	140	0	12.9	71	632	1029	1	7	100	24	1	526	174					
18	92.3	126	0	12.7	74	602	984	0	34	102	33	1	557	195					
19	94.3	130	0	13.3	128	536	934	0	51	78	34	0	627	135					
20	95.3	125	0	12	90	586	964	0	97	105	43	0	617	163					
21	96.8	151	1	10	58	510	950	0	33	108	41	0	394	261					
22	97.4	152	1	10.8	57	530	986	0	30	92	43	0	405	264					
23	98.7	162	1	12.1	75	522	996	0	40	73	27	0	496	224					
24	99.9	149	1	10.7	61	515	953	0	36	86	35	0	395	251					
25	103	177	1	11	58	638	974	0	24	76	28	0	382	254					
26	104.3	134	0	12.5	75	595	972	0	47	83	31	0	580	172					
27	105.9	130	0	13.4	90	623	1049	1	3	113	40	0	588	160					
28	106.6	157	1	11.1	65	553	955	0	39	81	28	0	421	239					
29	107.2	148	0	13.7	72	601	998	0	9	84	20	1	590	144					
30	108.3	126	0	13.8	97	542	990	0	18	102	35	0	589	166					
31	109.4	135	1	11.4	123	537	978	0	31	89	34	0	631	165					

Crime_data Column Label +

Calculation Mode: Automatic Workbook Statistics

3. In the image below, the relationship between "CrimeRate" and "StateSize" was chosen, which was found to be the -9th column in the table, being notable with: 9

Excel Crime_rate_data 1 - Saved

Search (Alt + Q)

File Home Insert Draw Page Layout Formulas Data Review View Automate Help Editing

PivotTable Table Forms Pictures Shapes Recommended Charts Link New Comment

P7 fx =VLOOKUP(A7,A2:N48,9

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	CrimeRate	Youth	Region	Education	Expenditu	LabourFor	Males	MoreMale	StateSize	YouthUn	MatureUn	HighYouth	Wage	BelowWage					
2	45.5	135	0	12.4	69	540	965	0	6	80	22	1	564	139					
3	52.3	140	0	10.9	55	535	1045	1	6	135	40	1	453	200					
4	56.6	157	1	11.2	47	512	962	0	22	97	34	0	288	276					
5	60.3	139	1	11.9	46	480	968	0	19	135	53	0	457	249					
6	64.2	126	0	12.2	106	599	989	0	40	78	25	1	593	171					
7	67.6	128	0	13.5	67	624	972	0	28	77	25	1	507	206					
8	70.5	130	0	14.1	63	641	984	0	14	70	21	1	486	196					
9	73.2	143	0	12.9	66	537	977	0	10	114	35	1	487	166					
10	75	141	0	12.9	56	523	968	0	4	107	37	0	489	170					
11	78.1	133	0	11.4	51	599	1024	1	7	99	27	1	425	225					
12	79.8	142	1	12.9	45	533	969	0	18	94	33	0	318	250					
13	82.3	123	0	12.5	97	526	948	0	113	124	50	0	572	158					
14	83.1	135	0	13.6	62	595	986	0	22	77	27	0	529	190					
15	84.9	121	0	13.2	118	547	964	0	25	84	29	0	689	126					
16	85.6	166	1	11.4	58	521	973	0	46	72	26	0	396	237					
17	88	140	0	12.9	71	632	1029	1	7	100	24	1	526	174					
18	92.3	126	0	12.7	74	602	984	0	34	102	33	1	557	195					
19	94.3	130	0	13.3	128	536	934	0	51	78	34	0	627	135					
20	95.3	125	0	12	90	586	964	0	97	105	43	0	617	163					
21	96.8	151	1	10	58	510	950	0	33	108	41	0	394	261					
22	97.4	152	1	10.8	57	530	986	0	30	92	43	0	405	264					
23	98.7	162	1	12.1	75	522	996	0	40	73	27	0	496	224					
24	99.9	149	1	10.7	61	515	953	0	36	86	35	0	395	251					
25	103	177	1	11	58	638	974	0	24	76	28	0	382	254					
26	104.3	134	0	12.5	75	595	972	0	47	83	31	0	580	172					
27	105.9	130	0	13.4	90	623	1049	1	3	113	40	0	588	160					
28	106.6	157	1	11.1	65	553	955	0	39	81	28	0	421	239					
29	107.2	148	0	13.7	72	601	998	0	9	84	20	1	590	144					
30	108.3	126	0	13.8	97	542	990	0	18	102	35	0	589	166					
31	109.4	135	1	11.4	123	537	978	0	31	89	34	0	631	165					

Crime_data Column Label +

Calculation Mode: Automatic Workbook Statistics

4. In the image below it is the final result that is correct (result that you can also search for manually not necessarily by using the function): 28

Excel Crime_rate_data 1 - Saved

Search (Alt + Q)

FileHomeInsertDrawPage LayoutFormulasDataReviewViewAutomateHelpEditing

PivotTableTableFormsPicturesShapesRecommended ChartsLinkNew Comment

P7fx=VLOOKUP(A7,A2:N48,9,FALSE)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	CrimeRate	Youth	Region	Education	Expenditu	LabourFor	Males	MoreMale	StateSize	YouthUne	MatureUn	HighYouth	Wage	BelowWage					
2	45.5	135	0	12.4	69	540	965	0	6	80	22	1	564	139					
3	52.3	140	0	10.9	55	535	1045	1	6	135	40	1	453	200					
4	56.6	157	1	11.2	47	512	962	0	22	97	34	0	288	276					
5	60.3	139	1	11.9	46	480	968	0	19	135	53	0	457	249					
6	64.2	126	0	12.2	106	599	989	0	40	78	25	1	593	171					
7	67.6	128	0	13.5	67	624	972	0	28	77	25	1	507	206		28			
8	70.5	130	0	14.1	63	641	984	0	14	70	21	1	486	196					
9	73.2	143	0	12.9	66	537	977	0	10	114	35	1	487	166					
10	75	141	0	12.9	56	523	968	0	4	107	37	0	489	170					
11	78.1	133	0	11.4	51	599	1024	1	7	99	27	1	425	225					
12	79.8	142	1	12.9	45	533	969	0	18	94	33	0	318	250					
13	82.3	123	0	12.5	97	526	948	0	113	124	50	0	572	158					
14	83.1	135	0	13.6	62	595	986	0	22	77	27	0	529	190					
15	84.9	121	0	13.2	118	547	964	0	25	84	29	0	689	126					
16	85.6	166	1	11.4	58	521	973	0	46	72	26	0	396	237					
17	88	140	0	12.9	71	632	1029	1	7	100	24	1	526	174					
18	92.3	126	0	12.7	74	602	984	0	34	102	33	1	557	195					
19	94.3	130	0	13.3	128	536	934	0	51	78	34	0	627	135					
20	95.3	125	0	12	90	586	964	0	97	105	43	0	617	163					
21	96.8	151	1	10	58	510	950	0	33	108	41	0	394	261					
22	97.4	152	1	10.8	57	530	986	0	30	92	43	0	405	264					
23	98.7	162	1	12.1	75	522	996	0	40	73	27	0	496	224					
24	99.9	149	1	10.7	61	515	953	0	36	86	35	0	395	251					
25	103	177	1	11	58	638	974	0	24	76	28	0	382	254					
26	104.3	134	0	12.5	75	595	972	0	47	83	31	0	580	172					
27	105.9	130	0	13.4	90	623	1049	1	3	113	40	0	588	160					
28	106.6	157	1	11.1	65	553	955	0	39	81	28	0	421	239					
29	107.2	148	0	13.7	72	601	998	0	9	84	20	1	590	144					
30	108.3	126	0	13.8	97	542	990	0	18	102	35	0	589	166					
31	109.4	135	1	11.4	123	537	978	0	31	89	34	0	631	165					

Crime_dataColumn Label+

Calculation Mode: AutomaticWorkbook Statistics

- In "A" column is "CrimeRate". Using a sum function, we found out the sum of all the crimes : "=SUM(A2:A48)"

A51 $\times$ $\checkmark$ f_x =SUM(A2:A48)

	A	B
47	157.7	136
48	161.8	131
49		
50		
51	4832	Sum of crime
52	0.04832	Sum of crime divide by population
53	4.832	Rate crime
54		

- In "A52" is the total of the number of crimes (A51) divide to the total population of cities (100000)

A52		$\times$	$\checkmark$	f_x	$= (A51/100000)$
	A			B	
46	134.3				124
47	157.7				136
48	161.8				131
49					
50					
51	4832			Sum of crime	
52	0.04832			Sum of crime divide by population	
53	4.832			Rate crime	
54					

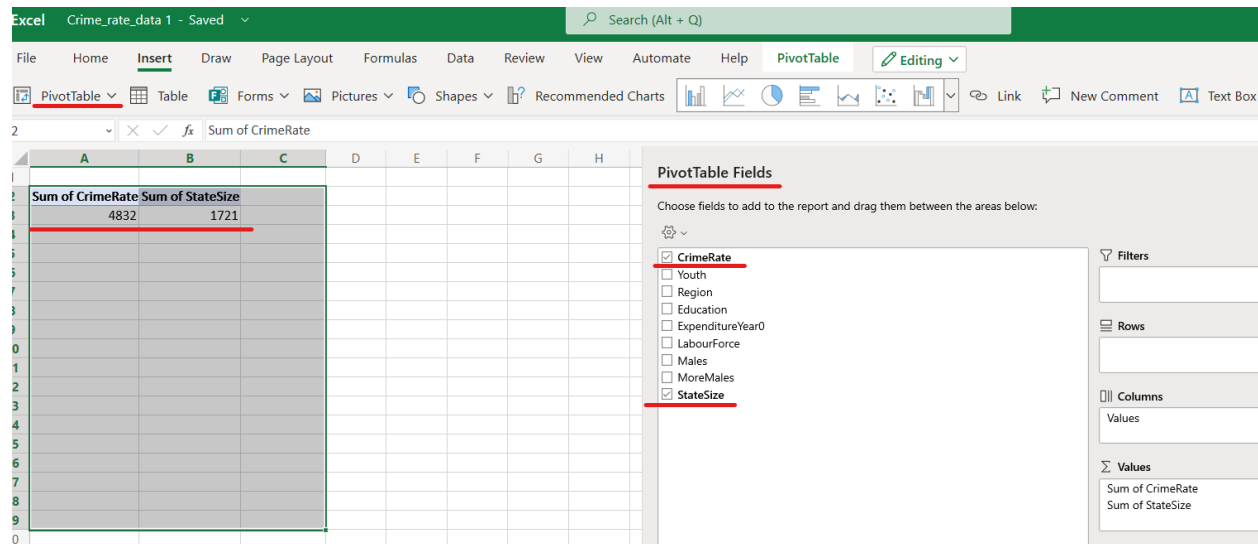
- In "A53" is the sum of crime divide by population, then multiply the result by 100.000 (for using a tiny little decimal)

A53		$\times$	$\checkmark$	f_x	$= (A52*"100.000")$
	A			B	
46	134.3				124
47	157.7				136
48	161.8				131
49					
50					
51	4832			Sum of crime	
52	0.04832			Sum of crime divide by population	
53	4.832			Rate crime	
54					

For "Pivot Table" (creation of a table showing the total number of crimes / incidents):

- Click on "insert", then click on "pivot table", then click on "from table / range", then select the entire table.
- On the right side you will receive notions about "fields" (which are at our choice depending on what we are looking for), and the answer comes in the form of the total amount.

ScreenShot :



The total amount coincides with the one above, where "VLOOKUP" was used. In both cases an extra calculation is required, but from the initial table "crime_rate" data set both "VLOOKUP" & "PIVOT TABLE" gave the same results for "sum of crime rate".

PART 2

Using the "SPSS" program :

In the request we need the data of 200 students, these can be found in the "ice_cream" data set. In our table there are some notions necessary to remember, but at the same time necessary to delete from our calculation area.

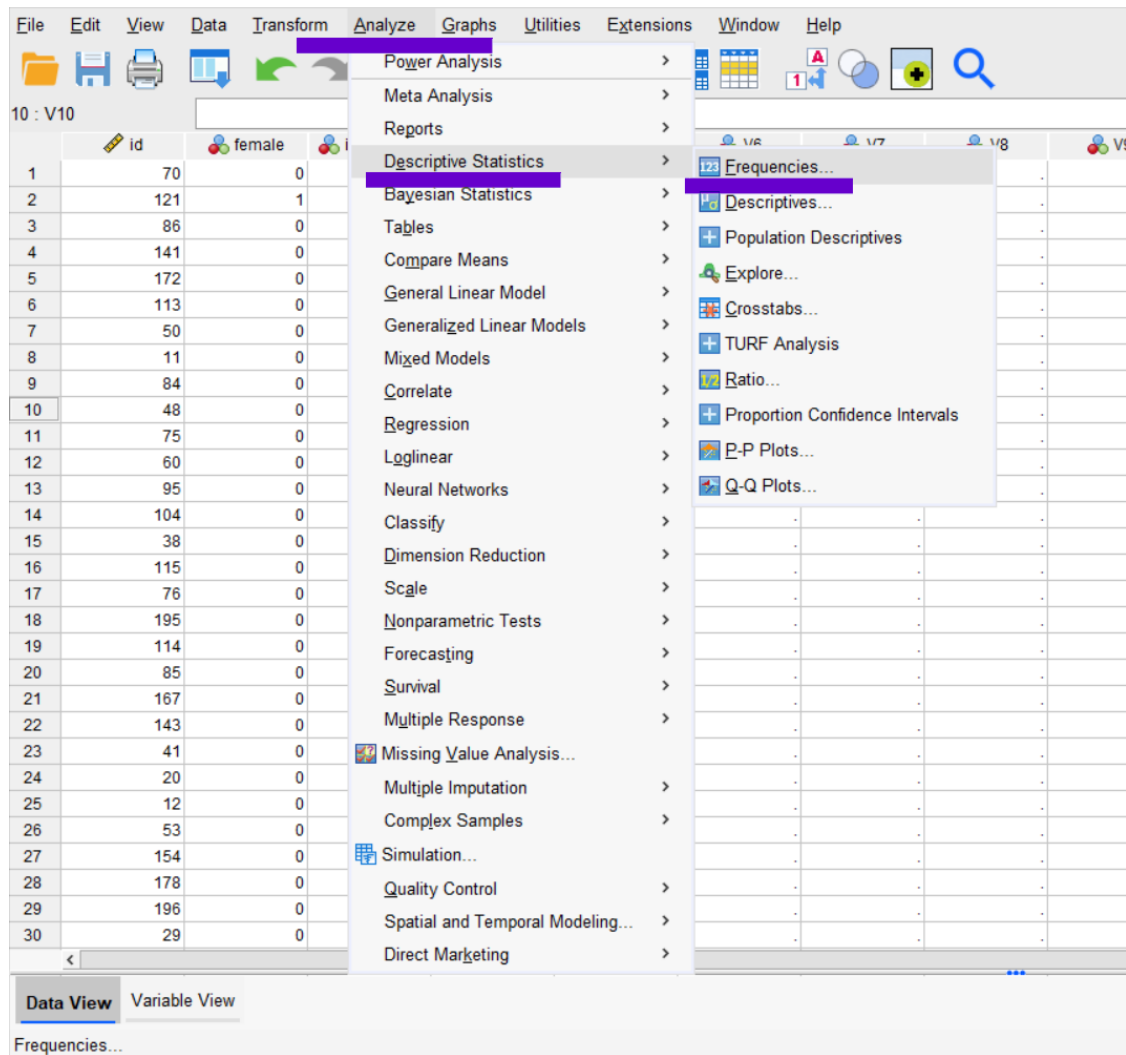
.	.
Male	0
Female	1
.	.
Favorite Flavor (1: Vanilla, 2: Chocolate, 3: Strawberry)	.
.	.
.	.

We have to analyze 3 different objectives to find out the number:

- students who like the vanilla flavor of ice cream

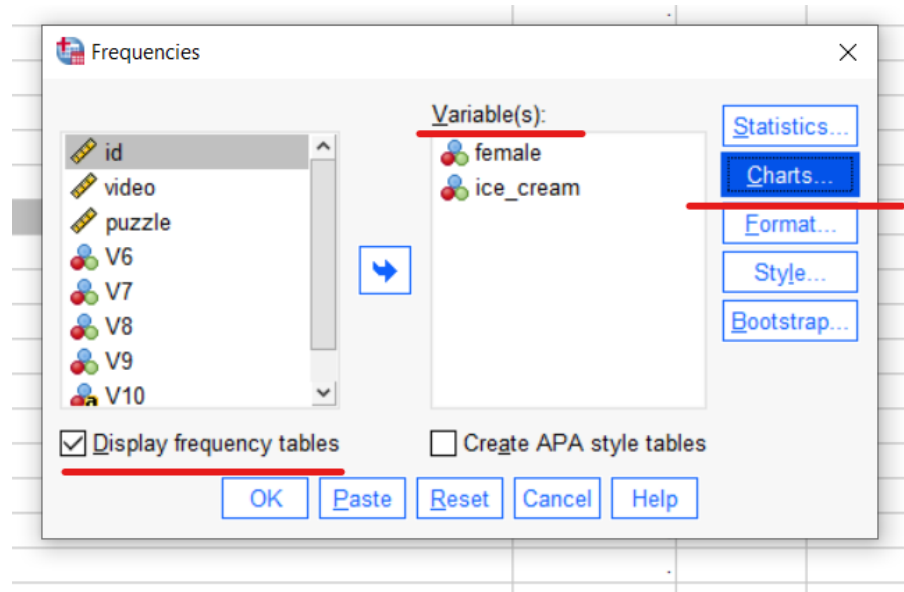
- b. students who are men and those who are women
- c. the average of those who prefer the chocolate / strawberry flavor

We do this data analysis by (see also screenshot / image below): we click on "analyze", then we click on "descriptive statistics", then we click on "frequencies" where we will be asked the data we want to analyze them.



The necessary notions are in the tables called "female" (because we need to find how many students are men and how many are women, and in this column is all the information) and "ice_cream" (because we want to know the flavor preferences ice cream for all genders).

These two must be moved to the right column "variable (s)" after which we check that there is no automatic check in any to "statistic (s)" (we check that it is not because we do not need "statistic"), our choice must be called "charts" (below), after we choose the desired type of "chart". Before confirming, we must also make sure that the "display frequency tables" check mark is present.



After all this, a new page will open, where our previous choice is automatically presented. From there we take the necessary information to meet the requirements.

We have a table that answers "the total number of students" (200) and "the total number of students who answered what flavor of ice cream they like" (200).

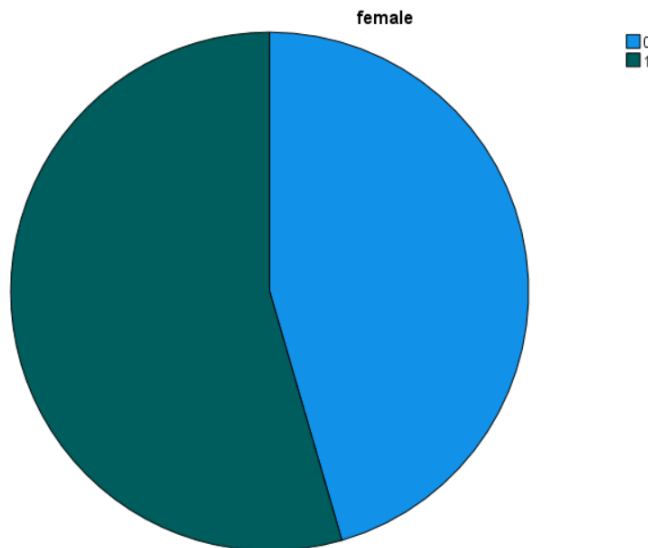
Frequencies			
Statistics			
		female	ice_cream
N	Valid	200	200
	Missing	0	0

Another table shows "genders". Recalling from the above (previous) that "0" are "men" and "1" are "women". Men = 91, Women = 109.

Frequency Table

		female		
		Frequency	Percent	Cumulative Percent
Valid	0	91	45.5	45.5
	1	109	54.5	100.0
	Total	200	100.0	

Adding a "Pie Chart" that highlights the same thing :



One last important table is related to the flavors of ice cream. We remind you here of the above, "1" is for "vanilla", "2" is for "chocolate", and "3" is for "strawberries".

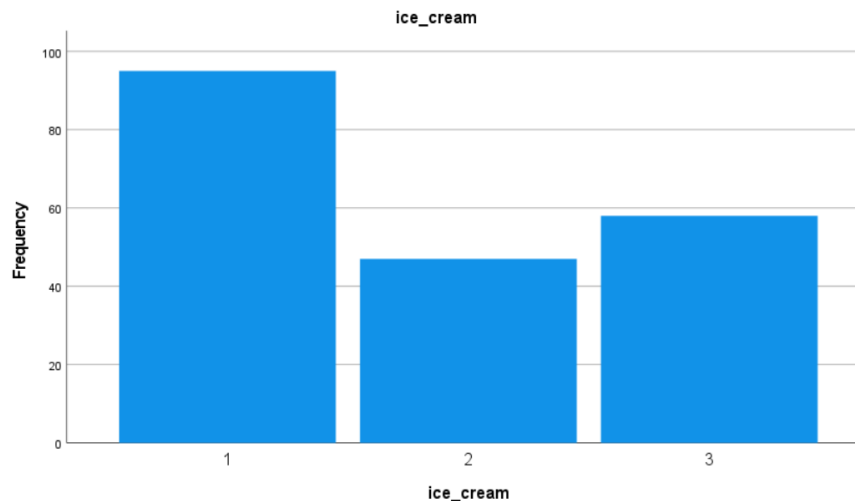
What do we care about in the table below? How many people prefer "vanilla" ice cream (where we find a number of "95" students). Then the second question for us is about an average / a percentage of those who prefer the ice cream flavor with "chocolate" and "strawberries" (we have "chocolate" with "23.5" then we have "strawberries" with "29.0" for which we have to make a sum and from which subtract the maximum percentage from "100" for checking).

$$\text{Answer } 100 - (23.5 + 29.0) = 100 - 52.5 = 47.5$$

Check the table: the percentage for "vanilla" flavor is "47.5" which means that the correct percentage for "chocolate" and "strawberry" flavor is "52.5"

ice_cream				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1	95	47.5	47.5	47.5
2	47	23.5	23.5	71.0
3	58	29.0	29.0	100.0
Total	200	100.0	100.0	

Adding a "Bar Chart" that highlights the same thing :



➤ Advantages and disadvantages of the "Excel" program :

ADVANTAGES	DISADVANTAGES
Data organization	Calculation Errors
Formula and calculation features	Cost of entry
Multiple user access	Lack of security

- Advantages and disadvantages of the "SPSS" program :

ADVANTAGES	DISADVANTAGES
Quick and easy to learn	Very similar to Excel
Can handle large amounts of data	Limited functionality

As a statistical technique for processing any raw data in a structured form, it is good to use "data mining" (more precisely it is a data extraction that involves any processing of text in documents).

When you need to create graphs / tables / visual reports you use "text mining". There are different extraction methods: summarization (keep the main point), categorization (there are documents that need this basis) etc.

How well the methods are shown above. We have a set of data (information entered through text / numbers) that must be extracted and placed according to needs in a simple way to read / explain. The above cases are not necessarily relevant in real life, but they are a good example. That is, ice cream (found and calculated in SPSS) can be replaced with a subject from school, a service provided (which you want to develop), etc. And the case presented by "Excel" is not necessarily relevant, but it may be necessary, like many other similar situations, ie a subject from school, a service provided, etc.