LLAMA 3.1

VISUAL ASPECTS	PRC	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colum
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	0	0	1	1	0	1	1	1	0	1	1	1	1	1	0	1	0	1	0	1	0	1	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0
3. Are the axis labels clear and readable?	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	1	0	1	1	1	0	1	0	1	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	0	1	1	1	0	0	1	1	0	0	1	1	1	0	0	0	0	0	1	1	0	0	1
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	1	0	1	1	1	0	1	0	0	0	0	1	1	1	1	0	1	1	1	1	0	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	0	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
8. Is there a correct data-ink ratio to aid readability?	0	0	1	1	1	1	1	0	0	1	0	1	0	0	1	1	1	0	1	1	1	1	1	1	0
TOTAL	5	6	4	7	7	8	3	3	5	5	0	4	5	5	7	6	1	5	5	6	5	7	3	5	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	1	1	1	1	1	1	1	0
3. Does the code correctly reference the column names in the dataset?	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	1	0	1	0	1	0	1	1	0	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	0	0	1	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	1	0	1	1	0	0
7. Does the code avoid redundant or unnecessary computations?	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TOTAL	7	5	5	7	4	5	5	7	7	6	4	5	6	6	4	7	4	6	6	7	5	7	7	5	3

VISUAL ASPECTS	PRC	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colur	Colum	Colum
DATA ASPECTS																											
1. Is the selected graph aligned with the nature of the prompt and used data?	1	0	1	1	1	1	0	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	0		
2. Does the code select the appropriate columns relevant to the user prompt?	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
3. Does the filtering process correctly reflect the user's intent?	0	1	1	1	1	0	0	0	1	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0		
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0		
5. Does the generated graph represent the correct subset of data requested?	1	0	0	1	0	1	0	1	1	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	1		
6. Does the code correctly handle missing, null or unexpected values in the selected column(s)?	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0		
7. Does the selected data cover all different aspects of the user prompt?	1	1	1	1	0	0	0	1	1	1	0	1	1	1	0	1	1	0	1	1	0	1	0	0	0		
TOTAL	5	4	5	5	3	3	2	4	5	4	1	3	4	5	3	5	2	2	5	4	2	3	2	2	2		

	SUM	AVG			EASY	MED	HARD
Visual	119	4.76		**VIZ**	5.285	5.090	3.714
Code	140	5.6		**DATA**	6.571	5.636	4.571
Data	85	3.4		**CODE**	4.285	3.454	2.428
				AVG	5.380	4.727	3.571

SCALED OUT OF 10

	AVG
Llama	
Visual	5.95
Code	8
Data	4.857

Llama	EASY	MED	HARD
VIZ	6.607	6.363	4.642
DATA	9.387	8.051	6.530
CODE	6.122	4.935	3.469
AVG	7.372	6.450	4.880

GEMMA 2

VISUAL ASPECTS	PRC	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	0	1	0	1	0	0	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	0	1	1	1	0	1	1	0	1
3. Are the axis labels clear and readable?	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
5. Are accurate legends included if necessary?	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	1	0	1	1	0	0	1	0	0	0	1	1	1	0	1	0	1	1	1	0	1	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	1	1	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
8. Is there a correct data-ink ratio to aid readability?	1	1	1	1	0	1	1	0	1	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1
TOTAL	6	6	6	6	7	5	2	3	5	7	0	5	6	5	6	5	1	5	5	6	2	6	3	1	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	0	1	1	0	0
3. Does the code correctly reference the column names in the dataset?	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	1	1	0	1	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	0	1	1	0	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	0	1	1	1	0	0	1	1	1	0	1	0	1	1	1	0	1	1	1	0	1	1	1	0
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
TOTAL	7	5	7	6	6	5	5	7	7	7	3	6	5	6	7	7	4	6	6	7	3	7	7	5	3
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	1	0	0	1	1	1	0	1	0	1	1	1	0	1	1	1	0	1	1	0	0
2. Does the code select the appropriate columns relevant to the user prompt?	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0	1	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5. Does the generated graph represent the correct subset of data requested?	1	0	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0
6. Does the code correctly handle missing or null values in the selected column(s)?	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
7. Does the selected data cover all different aspects of the user prompt?	1	0	0	1	1	1	0	1	1	1	0	0	0	1	1	1	0	0	1	1	0	1	0	0	0
TOTAL	5	3	4	5	5	3	2	4	6	4	1	3	2	5	6	4	1	2	5	4	1	4	2	3	2

					EASY	MED	HARD					SCALED OUT OF 10												
				VIZ	5.1428	4.7272	3.2857				Gemm	AVG		Gemm	EASY	MED	HARD							
	SUM	AVG		DATA	6.5714	5.8181	4.8571				Visual	5.55		VIZ	6.4285	5.9090	4.1071							
Visual	111	4.44		CODE	4.2857	3.4545	2.5714				Code	8.2285		DATA	9.3877	8.3116	6.9387							
Code	144	5.76		AVG	5.3333	4.6666	3.5714				Data	4.9142		CODE	6.1224	4.9350	3.6734							
Data	86	3.44												AVG	7.3129	6.3852	4.9064							

QWEN 2.5

VISUAL ASPECTS	PRC	Colur	Colun	Colun	Colun	Colur	Colun	Colun	Colur	Colur	Colun	Colun	Colur	Colur	Colun	Colun	Colur	Colun	Colun	Colur	Colur	Colur	Colun	Colur	Colur
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	0	1	0	0	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	1	1	0	1	1	0	0
3. Are the axis labels clear and readable?	0	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	0	1	0	0	0	0	1	1	0	0	1	1	0	0	0	1	1	0	0	1	0	0	1
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	1	0	0	1
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	1	1	1	1	1	1	1	0	0	0	1	1	1	0	1	0	1	1	1	0	0	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0
8. Is there a correct data-ink ratio to aid readability?	0	1	1	1	1	1	1	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	0	0
TOTAL	5	8	7	6	6	6	7	4	5	5	0	4	6	6	0	6	1	8	7	6	2	7	3	0	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	1	1	1	0	1	1	0	0
3. Does the code correctly reference the column names in the dataset?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	1	1	0	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1	0	1	1	1	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	0	1	1	1	1	0	0	1	0	1	1	1	0	1	1	1	0	1	1	1	0
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
TOTAL	7	7	7	6	6	7	7	7	7	6	4	6	6	6	5	7	4	7	7	7	3	7	7	5	3
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	1	0	1	1	1	0	1	1	0	0
2. Does the code select the appropriate columns relevant to the user prompt?	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	0	1	1	1	1	0	1	0	1	0	1	1	1	1	0	1	0	1	1	1	0	0	0	0	0
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0
5. Does the generated graph represent the correct subset of data requested?	1	1	1	0	0	1	1	1	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1
6. Does the code correctly handle missing or null values in the selected column(s)?	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0
7. Does the selected data cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0	1	1	1	0	1	0	1	0
TOTAL	5	6	6	5	4	4	6	4	5	4	1	3	5	5	3	4	1	7	5	5	1	3	2	4	2

					EASY	MED	HARD
	SUM	AVG		VIZ	5.5714	4.7272	3.7142
Visual	117	4.68		DATA	6.7142	6.1818	5.1428
Code	151	6.04		CODE	4.1428	4.0909	3.7142
Data	100	4		AVG	5.4761	5	4.1904

SCALED OUT OF 10

Qwen	AVG		Qwen	EASY	MED	HARD
Visual	5.85		VIZ	8.3928	7.7272	6.4285
Code	8.6285		DATA	5.9183	5.8441	5.3061
Data	5.7142		CODE	7.8231	7.1428	5.9863
			AVG	7.3781	6.9047	5.9070

CLAUDE 3.5

VISUAL ASPECTS	PRC	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
3. Are the axis labels clear and readable?	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	1	0	1	0	0	0	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0	1	1	0	1	1	0	0	0
8. Is there a correct data-ink ratio to aid readability?	0	1	0	1	1	1	1	0	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1	0	1
TOTAL	5	6	5	6	7	6	7	3	5	8	6	6	6	5	6	6	5	7	7	6	7	8	3	0	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
3. Does the code correctly reference the column names in the dataset?	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	0	0	0	1	0	1	1	1	1	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	0	1	1	1	0
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
TOTAL	7	5	5	6	7	6	7	7	7	6	5	5	5	3	6	5	3	7	7	7	6	7	7	5	3
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
2. Does the code select the appropriate columns relevant to the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	0	0	1	1	0	1	0	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0
5. Does the generated graph represent the correct subset of data requested?	1	1	0	1	1	1	1	1	1	0	1	1	1	0	0	1	0	1	0	0	1	1	0	0	0
6. Does the code correctly handle missing or null values in the selected column(s)?	0	0	1	0	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1	1	1	0	1	0
7. Does the selected data cover all different aspects of the user prompt?	1	0	0	1	1	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	0	1	0
TOTAL	5	3	4	5	5	5	6	4	5	5	6	6	5	3	5	5	3	7	5	5	5	7	2	4	2

					EASY	MED	HARD																		
	SUM	**AVG**		**VIZ**	5.8571	5.1818	5.7142																		
Visual	138	5.52		**DATA**	6.2857	5.5454	5.5714																		
Code	144	5.76		**CODE**	5.2857	4.0909	5																		
Data	117	4.68		**AVG**	5.8095	4.9393	5.4285																		

SCALED OUT OF 10																									
Claude	**AVG**		Claude	**EASY**	**MED**	**HARD**																			
Visual	6.9		**VIZ**	7.3214	6.4772	7.1428																			
Code	8.2285		**DATA**	8.9795	7.9220	7.9591																			
Data	6.6857		**CODE**	7.5510	5.8441	7.1428																			
			AVG	7.9506	6.7478	7.4149																			

DEEPSEEK-CHAT

VISUAL ASPECTS	PRO	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
3. Are the axis labels clear and readable?	0	0	0	1	1	1	1	0	1	1	0	0	0	0	0	1	0	1	1	1	1	1	0	0	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	1	0	1	0	0	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	0	0
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values'	1	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	0	1	1	0	1	1	0	0	0
8. Is there a correct data-ink ratio to aid readability?	0	1	0	1	1	1	1	0	1	1	1	0	0	0	0	1	0	1	1	1	1	1	1	0	1
	5	6	5	6	8	6	7	3	7	8	6	4	6	5	6	6	5	8	7	6	7	8	3	0	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	1	0	1	1	1	1	1	1	0	0
3. Does the code correctly reference the column names in the dataset?	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)'	1	0	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	0
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
	7	5	5	6	7	6	7	7	7	7	5	6	7	3	6	7	3	7	7	7	6	7	7	5	3
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
2. Does the code select the appropriate columns relev to the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	0	0	1	1	1	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	0	0	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	1	0	1	0
5. Does the generated graph represent the correct subset of data requested?	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	0	0	1	1	0	0	0
6. Does the code correctly handle missing or null value in the selected column(s)?	0	0	1	0	1	0	1	0	0	1	1	0	0	1	1	1	1	1	0	1	1	1	0	1	0
7. Does the selected data cover all different aspects of the user prompt?	1	0	0	1	1	1	1	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	0	1	0
	5	3	4	5	7	5	6	4	5	6	6	3	6	3	5	5	3	7	5	5	5	7	2	4	2

					EASY	MED	HARD					SCALED OUT OF 10													
	SUM	AVG		VIZ	5.5714	5.4545	5.8571				Deeps	AVG		Deeps	EASY	MED	HARD								
Visual	140	5.6		DATA	6.7142	5.8181	5.5714		Does pretty w		Visual	7		VIZ	6.9642	6.8181	7.3214								
Code	150	6		CODE	4.8571	4.4545	5				Code	8.5714		DATA	9.5918	8.3116	7.9591								
Data	118	4.72		AVG	5.7142	5.2424	5.4761				Data	6.7428		CODE	6.9387	6.3636	7.1428								
														AVG	7.8316	7.1645	7.4744								

GEMINI 2.0 FLASH THINKING

VISUAL ASPECTS	PRC	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
3. Are the axis labels clear and readable?	0	0	1	1	1	1	1	1	0	1	1	0	1	0	0	0	0	1	1	1	1	1	0	0	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	1	0	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	0	0	0
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	1	0	1	0	1	1	0	1	0	0	1	1	1	1	1	1	1	0	1	1	0	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	0	1	0	1	1	1	0	0	1	1	0	0	0	1	1	0	0	1	0	1	1	0	0	0
8. Is there a correct data-ink ratio to aid readability?	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1
	5	6	7	6	7	8	7	5	6	7	7	5	6	5	6	6	5	5	6	6	8	7	3	0	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1	0	0
3. Does the code correctly reference the column names in the dataset?	1	0	1	1	1	1	1	0	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1	0	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	0	0	0	1	1	1	0	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	0	1	1	1	1	1	1	0
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1
	7	5	6	6	6	7	7	2	5	7	2	6	3	5	6	5	3	6	7	7	6	7	7	5	3
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
2. Does the code select the appropriate columns relevant to the user prompt?	1	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	0	0	0	1	0	1	0	0	0	1	0	1	0	1	1	1	0	0	1	1	1	1	0	0	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	0	0	0	0	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	1	0
5. Does the generated graph represent the correct subset of data requested?	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0
6. Does the code correctly handle missing or null values in the selected column(s)?	0	0	1	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	1	0	1	0
7. Does the selected data cover all different aspects of the user prompt?	1	0	1	1	0	1	1	0	1	0	0	0	0	1	1	1	0	0	1	1	0	1	0	1	0
	5	3	5	5	3	7	6	3	5	5	2	4	3	4	5	5	3	2	5	5	5	7	2	4	2

	SUM	AVG			EASY	MED	HARD
Visual	141	5.64		**VIZ**	5.7142	5.3636	6
Code	136	5.44		**DATA**	5.7142	5.4545	5.1428
Data	105	4.2		**CODE**	4.8571	4	3.8571
				AVG	5.4285	4.9393	5

SCALED OUT OF 10

			EASY	MED	HARD
Gemin	**AVG**				
Visual	7.05	**VIZ**	7.1428	6.7045	7.5
Code	7.7714	**DATA**	8.1632	7.7922	7.3469
Data	6	**CODE**	6.9387	5.7142	5.5102
		AVG	7.4149	6.7370	6.7857

PROBLEM: HARDCODING

GPT-4O

VISUAL ASPECTS	PRC	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun	Colun
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0	0	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1
3. Are the axis labels clear and readable?	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	0	0	0
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	1	0	1	0	1	1	0	0	1	0	0	1	1	0	1	1	1	0	0	0	0	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0	0	1	0	0	1	0	0	0
8. Is there a correct data-ink ratio to aid readability?	0	1	0	1	1	1	1	1	0	1	1	0	0	1	1	0	1	1	1	1	0	1	1	0	1
	5	6	5	6	7	6	8	5	5	8	5	3	6	5	6	6	5	5	6	4	0	7	3	0	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	0	1	1	1	1	0	1	1	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0
3. Does the code correctly reference the column names in the dataset?	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	0	0	0	1	0	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	1	1	1	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0	1	0	0	1	1	1	1	1	1	0
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	7	5	5	6	6	6	7	5	6	7	5	5	5	5	6	5	3	6	7	5	5	7	7	5	3
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
2. Does the code select the appropriate columns relevant to the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	0	0	1	1	0	1	0	0	0	1	0	1	1	1	1	1	0	0	1	0	0	1	0	0	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0
5. Does the generated graph represent the correct subset of data requested?	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	1	0	0	0	1	0	0	0
6. Does the code correctly handle missing or null values in the selected column(s)?	0	0	1	0	1	0	1	1	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	1	0
7. Does the selected data cover all different aspects of the user prompt?	1	0	0	1	0	1	1	0	1	0	0	0	1	1	0	1	0	0	1	0	1	1	0	1	0
	5	3	4	5	3	5	6	4	5	5	4	3	5	4	5	5	3	4	5	3	4	7	2	4	2

					EASY	MED	HARD					SCALED OUT OF 10				
	SUM	AVG		VIZ	5.4285	5	4.4285				GPT4	AVG		GPT4	EASY	MED	HARD
Visual	124	4.96		DATA	6	5.4545	5.2857				Visual	6.2		VIZ	6.7857	6.25	5.5357
Code	139	5.56		CODE	4.8571	3.8181	4.1428				Code	7.9428		DATA	8.5714	7.7922	7.5510
Data	105	4.2		AVG	5.4285	4.7575	4.6190				Data	6		CODE	6.9387	5.4545	5.9183
														AVG	7.4319	6.4989	6.3350

O1-HIGH

VISUAL ASPECTS	PRO	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
3. Are the axis labels clear and readable?	0	1	1	1	1	1	1	1	1	1	0	1	0	0	0	1	0	0	1	1	1	1	0	0	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	0	0	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	0	0
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	0	0	0	1	0	1	1	0	0	0
8. Is there a correct data-ink ratio to aid readability?	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1
	5	7	7	6	8	7	7	7	7	8	6	6	7	5	6	6	6	6	7	6	8	7	3	0	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	0	0
3. Does the code correctly reference the column names in the dataset?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	0	1	1	0	1	1	1	1	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1
	7	7	7	7	7	7	7	7	7	7	6	6	7	3	6	7	4	7	7	6	6	7	7	5	3
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
2. Does the code select the appropriate columns relevai to the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0	0	1	0	1	1	0	0	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	1	0	1	1	1	1	1	0	1	0	0	1	0	0	0	1	1	0	0	1	1	0	1	0
5. Does the generated graph represent the correct subset of data requested?	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	1	1	1	1	0	0	0
6. Does the code correctly handle missing or null values in the selected column(s)?	0	1	1	1	1	0	1	0	0	1	1	1	0	1	1	1	1	0	1	0	1	1	0	1	0
7. Does the selected data cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	0	1	1	1	0	1	0	1	0
	5	7	6	7	7	6	6	6	5	6	6	4	6	3	5	5	4	5	6	4	5	7	2	4	2

					EASY	MED	HARD																			
	SUM	AVG		VIZ	6.2857	5.8181	6				O1-HI(	AVG		O1-Hig	EASY	MED	HARD									
Visual	150	6		DATA	6.8571	6.0909	6				Visual	7.5		VIZ	7.8571	7.2727	7.5									
Code	157	6.28		CODE	5.7142	4.9090	5				Code	8.9714		DATA	9.7959	8.7012	8.5714									
Data	129	5.16		AVG	6.2857	5.6060	5.6666				Data	7.3714		CODE	8.1632	7.0129	7.1428									
														AVG	8.6054	7.6623	7.7380									

SCALED OUT OF 10

O1-HIGH + ADDITIONAL CONTEXT

VISUAL ASPECTS	PRO	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
3. Are the axis labels clear and readable?	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	0	0	0
4. Are the colors appropriately used to distinguish between different objects?	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	0	0	0
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	0	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	0	0	1	0	0	0
8. Is there a correct data-ink ratio to aid readability?	0	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1
	5	6	5	7	8	8	8	8	7	8	8	6	7	5	6	8	8	8	7	6	4	8	3	0	2
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1
3. Does the code correctly reference the column names in the dataset?	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	0
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
	7	5	5	7	7	7	7	7	7	7	6	6	7	3	6	7	4	7	7	7	6	7	7	6	4
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
2. Does the code select the appropriate columns relevai to the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	1	1	0	1	0	0	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	0	0	1	1	1	1	1	0	1	0	0	1	0	0	0	1	1	0	0	0	1	0	1	0	
5. Does the generated graph represent the correct subset of data requested?	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	1	0	0	1	0	0	0	
6. Does the code correctly handle missing or null values in the selected column(s)?	0	0	1	1	1	0	1	0	0	1	1	1	0	1	1	1	1	0	1	1	0	1	0	1	0	
7. Does the selected data cover all different aspects of the user prompt?	1	0	0	1	1	1	1	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	0	1	0	
	5	3	4	7	7	6	7	6	5	6	6	4	6	3	5	5	4	5	6	5	3	7	2	4	2	

					EASY	MED	HARD					SCALED OUT OF 10														
	SUM	AVG		VIZ	7	6	5.8571					AVG			EASY	MED	HARD									
Visual	156	6.24		DATA	6.8571	6.0909	5.8571					7.8		VIZ	8.75	7.5	7.3214									
Code	156	6.24		CODE	5.7142	4.6363	4.5714					8.9142		DATA	9.7959	8.7012	8.3673									
Data	123	4.92		AVG	6.5238	5.5757	5.4285					7.0285		CODE	8.1632	6.6233	6.5306									
														AVG	8.9030	7.6082	7.4064									

CLAUDE 3.7 + FEEDBACK LOOP

VISUAL ASPECTS	PRO	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Are the axis labels clear and readable?	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4. Are the colors appropriately used to distinguish between different objects?	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	0	1
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
8. Is there a correct data-ink ratio to aid readability?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	7	7	8	7	8	6	8	7	6	8	7	7	7	7	7	6	8	8	6	7	8	7	6	7	8
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the code correctly reference the column names in the dataset?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	5	7	7	7	6	7	7	7	7	7
DATA ASPECTS																									

1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code select the appropriate columns relevant to the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0	1	1	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1
5. Does the generated graph represent the correct subset of data requested?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1
6. Does the code correctly handle missing or null values in the selected column(s)?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
7. Does the selected data cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
	7	7	7	7	7	7	7	7	7	7	7	7	5	7	7	4	6	7	7	5	7	7	6	6	7	

					EASY	MED	HARD					SCALED OUT OF 10															
	SUM	AVG		VIZ	6.7142	7	7.7142					AVG		CLAUI	EASY	MED	HARD									
Visual	178	7.12		DATA	6.7142	6.9090	7					8.9		VIZ	8.3928	8.75	9.6428									
Code	172	6.88		CODE	6.5714	6.3636	7					9.8285		DATA	9.5918	9.8701	10									
Data	165	6.6		AVG	6.6666	6.7575	7.2380					9.4285		CODE	9.3877	9.0909	10									
														AVG	9.1241	9.2370	9.8809									

CLAUDE 3.7 + FEEDBACK LOOP + ADDITIONAL CONTEXT

VISUAL ASPECTS	PRO	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum	Colum
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Is the x-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Is the y-axis correctly assigned and labeled?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Are the axis labels clear and readable?	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
4. Are the colors appropriately used to distinguish between different objects?	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1
5. Are accurate legends included if necessary?	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
6. Is there an appropriate aspect ratio and scaling to avoid misleading information?	1	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1	0
7. Are the marks (data points, bars, lines, etc.) displayed in a way that correctly represents the values?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
8. Is there a correct data-ink ratio to aid readability?	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
	8	8	8	7	6	8	7	7	7	6	8	6	7	7	7	8	8	5	8	8	7	7	8	8	7
CODE ASPECTS																									
1. Does the code correctly import and use the libraries necessary for graph generation?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2. Does the code execute without any syntax errors and successfully generate a graph?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the code correctly reference the column names in the dataset?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4. Does the code correctly filter the dataset based on the user's specified conditions (e.g. location and year)?	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
5. Does the code avoid hardcoding data or column names and instead handle them accordingly?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6. Does the generated code cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
7. Does the code avoid redundant or unnecessary computations?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	7	7	6	7	7	7	7	7	7	7	7	7	7	7	5	7	7	7	7	7	7	7	7	6	6
DATA ASPECTS																									
1. Is the selected graph aligned with the nature of the prompt and used data?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1

2. Does the code select the appropriate columns relevant to the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3. Does the filtering process correctly reflect the user's intent?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
4. Are the correct aggregation operations (e.g. sum, average, median, etc.) applied where necessary?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
5. Does the generated graph represent the correct subset of data requested?	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0
6. Does the code correctly handle missing or null values in the selected column(s)?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1
7. Does the selected data cover all different aspects of the user prompt?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	7	7	6	7	7	7	7	7	7	7	7	7	7	7	7	5	6	6	7	7	7	6	7	7	5	5	

					EASY	MED	HARD
	SUM	AVG		VIZ	7.2857	7.3636	7
Visual	181	7.24		DATA	7	6.9090	6.4285
Code	170	6.8		CODE	6.8571	6.7272	6.1428
Data	165	6.6		AVG	7.0476	7	6.5238

SCALED OUT OF 10

	AVG		CLAUI	EASY	MED	HARD
	9.05		VIZ	9.1071	9.2045	8.75
	9.7142		DATA	10	9.8701	9.1836
	9.4285		CODE	9.7959	9.6103	8.7755
			AVG	9.6343	9.5616	8.9030