Part A

Dataset description 1

Name: Marit van den Helder
Student number: 13970097

Dataset: https://doi.org/10.24432/C58D0H

Records: 666
Variables: 11
Dataset organization: Structured

Description:
The dataset contains characteristics of candidates that were admitted to medical colleges of Assam. Its aim was to find correlations between the characteristics and the performance on the exam. Note that these are characteristics of people who were admitted, not of people who failed the entrance exam.

The variables in this dataset are:
- Performance (discrete, ordinal)
- Gender (binary)
- Coaching (discrete, categorical)
- Time (discrete, interval)
- Class X education (discrete, categorical)
- Class XII education (discrete, categorical)
- Medium (discrete, categorical)
- Class X percentage (discrete, interval)
- Class XII percentage (discrete, interval)
- Father occupation (discrete, categorical)
- Mother occupation (discrete, categorical)

Table 1. Descriptive statistics for each variable

Variable	Value	Frequency	Proportion	Percentage	Mode	Median
Performance	Excellent	101	101/666	15.16%	Good	Good
	Very good	198	198/666	29.73%		
	Good	210	210/666	31.53%		
	Average	157	157/666	23.57%		
Gender	Male	355	355/666	53.03%	Male	n/a
	Female	311	311/666	46.70%		
Coaching	None	150	150/666	22.52%	With Assam	n/a
	With Assam	449	449/666	67.42%		
	Outside	67	67/666	10.06		

	Assam					
Class X percentage	Excellent	511	511/666	76.73%	Excellent	Excellent
	Very good	101	101/666	15.17%		
	Good	41	41/666	6.16%		
	Average	13	13/666	1.95%		
Father occupation	Doctor	55	55/666	8.26%	Other	n/a
	School teacher	109	109/666	16.37%		
	Business	103	103/666	15.47%		
	College teacher	27	27/666	4.05%		
	Other	277	277/666	41.59%		
	Bank official	23	23/666	3.45%		
	Engineer	45	45/666	6.76%		
	Cultivator	27	27/666	4.05%		

Research question:
Is the occupation of the father correlated with the performance in the entrance exam of Assam medical school?

Dataset description 2

Name: Marit van den Helder
Student number: 13970097

Dataset: Coffee quality database from the Coffee Quality Institute (CQI) (https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi?select=merged_data_cleaned.csv)

Records: 1336
Variables: 44
Dataset organization: Structured

Description:
The dataset contains flavour descriptions, quality measures, processing elements and farm metadata (such as country of origin) of different kinds of coffee in 2018. The reviews are based on reviews from specialized reviewers for two kinds of coffee: arabica and robusta.

The variables in this dataset are:

- Species (discrete, categorical)
- Owner (discrete, categorical)
- Country of origin (discrete, categorical)
- Farm name (discrete, categorical)
- Lot number (discrete, categorical)
- Mill (discrete, categorical)
- ICO number (discrete, categorical)
- Company (discrete, categorical)
- Altitude (continuous, ratio)
- Region (discrete, categorical)
- Producer (discrete, categorical)
- Number of bags (discrete, ratio)
- Bag weight (continuous, ratio)
- In country partner (discrete, categorical)
- Harvest year (discrete, ordinal)
- Grading date (discrete, categorical)
- Owner one (discrete, categorical)
- Variety (discrete, categorical)
- Processing method (discrete, categorical)
- Aroma (discrete, categorical)
- Flavor (discrete, interval)
- Aftertaste (discrete, interval)
- Acidity (discrete, interval)
- Body (discrete, interval)
- Balance (discrete, interval)
- Uniformity (discrete, interval)
- Clean cup (discrete, interval)
- Sweetness (discrete, interval)
- Cup per points (discrete, interval)
- Total cup points (discrete, interval)

- Moisture (discrete, interval)
- Category one defects (discrete, ratio)
- Quakers (discrete, interval)
- Color (discrete, categorical)
- Category two defects (discrete, ratio)
- Expiration (discrete, categorical)
- Certification body (discrete, categorical)
- Certification address (discrete, categorical)
- Certification contact (discrete, categorical)
- Unit of measurement (discrete, categorical)
- Altitude low meters (continuous, ratio)
- Altitude high meters (continuous, ratio)
- Altitude mean meters (continuous, ratio)

Table 2. Descriptive statistics for each variable

Variable	**Value**	**Frequency**	**Proportion**	**Percentage**	**Mode**	**Median**	**Mean**
Species	Arabica	1308	1308/1336	97.90%	Arabica	n/a	n/a
	Robusta	28	28/1336	2.10%			
Country of origin	Ethiopia	44	44/1336	3.29%	Mexico	n/a	n/a
	Guatemala	181	181/1336	13.55%			
	Brazil	131	131/1336	9.81%			
	Peru	10	10/1336	0.75%			
	United States	10	10/1336	0.75%			
	United States (Hawaii)	73	73/1336	5.46%			
	Indonesia	20	20/1336	1.50%			
	China	16	16/1336	1.20%			
	Costa Rica	51	51/1336	3.82%			
	Mexico	236	236/1336	17.66%			
	Uganda	36	36/1336	2.69%			
	Honduras	53	53/1336	3.97%			
	Taiwan	73	73/1336	5.46%			
	Nicaragua	26	26/1336	1.95%			
	Tanzania	40	40/1336	2.99%			

	Kenya	25	25/1336	1.87%			
	Thailand	32	32/1336	2.40%			
	Colombia	183	183/1336	13.70%			
	Panama	4	4/1336	0.30%			
	Papua New Guinea	1	1/1336	0.07%			
	El Salvador	21	21/1336	1.57%			
	Japan	1	1/1336	0.07%			
	Ecuador	3	3/1336	0.22%			
	United States (Puerto Rico)	4	4/1336	0.30%			
	Haiti	6	6/1336	0.45%			
	Burundi	2	2/1336	0.15%			
	Vietnam	8	8/1336	0.60%			
	Phillipines	5	5/1336	0.37%			
	Rwanda	1	1/1336	0.07%			
	Malawi	11	11/1336	0.82%			
	Laos	3	3/1336	0.22%			
	Zambia	1	1/1336	0.07%			
	Myanmar	8	8/1336	0.60%			
	Mauritius	1	1/1336	0.07%			
	Côte D'Ivore	1	1/1336	0.07%			
	India	1	1/1336	0.07%			
	None	1	1/1336	0.07%			
Color	Green	868	868/1336	64.97%	Green	n/a	n/a
	Bluish-green	114	114/1336	8.53%			

	Blue-green	84	84/1336	6.29%			
	None	270	270/1336	20.21%			
Processing method	Washed / wet	813	813/1336	60.85	Washed / wet	n/a	n/a
	Natural / dry	357	357/1336	19.24			
	Pulped natural / honey	14	14/1336	1.05			
	Semi - washed / semi - pulped	56	56/1336	4.19			
	Other	26	26/1336	1.95			
	None	170	170/1336	12.72			
Aroma	8.75	1	1/1336	0.07%	7.67	7.58	7.57
	8.67	2	2/1336	0.15%			
	8.58	1	1/1336	0.07%	**Q1**	**Q3**	**IQR**
	8.5	3	3/1336	0.22%	7.42	7.75	0.33
	8.42	9	9/1336	0.67%			
	8.33	7	7/1336	0.52%	**Standard Deviation**		
	8.25	9	9/1336	0.67%	0.378		
	8.17	20	20/1336	1.50%			
	8.08	20	20/1336	1.50%			
	8.0	18	18/1336	3.59%			
	7.92	59	59/1336	4.41%			
	7.83	103	103/1336	7.71%			
	7.81	2	2/1336	0.15%			
	7.75	125	125/1336	9.36%			
	7.67	179	179/1336	13.40%			
	7.58	152	152/1336	11.38%			

	7.5	164	164/1336	12.28%	
	7.42	121	121/1336	9.06%	
	7.33	98	98/1336	7.34%	
	7.25	77	77/1336	6.76%	
	7.17	45	45/1336	3.37%	
	7.08	28	28/1336	2.10%	
	7.0	23	23/1336	1.72%	
	6.92	14	14/1336	1.05%	
	6.83	9	9/1336	0.67%	
	6.75	7	7/1336	0.52%	
	6.67	3	3/1336	0.22%	
	6.5	2	2/1336	0.15%	
	6.42	1	1/1336	0.07%	
	6.33	1	1/1336	0.07%	
	6.17	1	1/1336	0.07%	
	5.08	1	1/1336	0.07%	
	0.0	1	1/1336	0.07%	

Research question:
Is there a correlation between the rating of the aroma and the way coffee beans are processed?

Part B

Name: Marit van den Helder
Student number: 13970097

Dataset: Jobs (https://databank.worldbank.org/source/jobs#)

Description search process:
I started looking on google for global data on wealth, since that can have a correlation on the general health of the population in a country. I found a website with a lot of different datasets, customisable to the subjects you want data in. I chose a variety of wealth indicators, to hopefully find a correlation.

Chosen option: Option 1 (comparison based on country)

Table 3. Correlation overview

Analysis description:
Unfortunately I was unable to complete the Pearson correlation tests. I had trouble with the data types. I was able to aggregate the two databases using the pandas library. Specifically I used the join command.

Correlation reflection:
Not applicable.