3.9 Exercises

1. The table below shows the age of each employee at a cardboard box factory.

ID	1	2	3	4	5	6	7	8	9	10
AGE	51	39	34	27	23	43	41	55	24	25

ID	11	12	13	14	15	16	17	18	19	20
AGE	38	17	21	37	35	38	31	24	35	33

Based on this data calculate the following **summary statistics** for the AGE feature:

a. Minimum, maximum and range

b. Mean and median

c. Variance and standard deviation

d. 1^{st} quartile (25^{th} percentile) and 3^{rd} quartile (75^{th} percentile)

e. Inter-quartile range

f. 12^{th} percentile

2. The table below shows the policy type held by customers at a life assurance company.

ID	POLICY
1	Silver
2	Platinum
3	Gold
4	Gold
5	Silver
6	Silver
7	Bronze
8	Silver
9	Platinum
10	Platinum
11	Silver

12	Gold
13	Platinum
14	Silver
15	Platinum
16	Silver
17	Platinum
18	Platinum
19	Gold
20	Silver

a. Based on this data calculate the following **summary statistics** for the POLICY feature:

i. Mode and 2nd mode

ii. Mode % and 2nd mode %

b. Draw a **bar plot** for the POLICY feature.

3. An analytics consultant at an insurance company has built an **ABT** that will be used to train a model to predict the best communications channel to use to contact a potential customer with an offer of a new insurance product.[15] The following table contains an extract from this ABT—the full ABT contains 5,200 instances.

ID	OCC	GENDER	AGE	LOC	MOTOR INS	MOTOR VALUE	HEALTH INS	HEALTH TYPE	HEALTH DEPS ADULTS	HEALTH DEPS KIDS	PREF CHANNEL
1	Student	female	43	urban	yes	42,632	yes	PlanC	1	2	sms
2		female	57	rural	yes	22,096	yes	PlanA	1	2	phone
3	Doctor	male	21	rural	yes	27,221	no				phone
4	Sheriff	female	47	rural	yes	21,460	yes	PlanB	1	3	phone
5	Painter	male	55	rural	yes	13,976	no				phone
⋮					⋮				⋮		
14		male	19	rural	yes	48,66	no				email
15	Manager	male	51	rural	yes	12,759	no				phone
16	Farmer	male	49	rural	no		no				phone
17		female	18	urban	yes	16,399	no				sms
18	Analyst	male	47	rural	yes	14,767	no				email
⋮					⋮				⋮		
2747		female	48	rural	yes	35,974	yes	PlanB	1	2	phone
2748	Editor	male	50	urban	yes	40,087	no				phone
2749		female	64	rural	yes	156,126	yes	PlanC	0	0	phone
2750	Reporter	female	48	urban	yes	27,912	yes	PlanB	1	2	email
⋮					⋮				⋮		
4780	Nurse	male	49	rural	no		yes	PlanB	2	2	email
4781		female	46	rural	yes	18,562	no				phone
4782	Courier	male	63	urban	no		yes	PlanA	2	0	email
4783	Sales	male	21	urban	no		no				sms
4784	Surveyor	female	45	rural	yes	17,840	no				sms
⋮					⋮				⋮		
5199	Clerk	male	48	rural	yes	19,448	yes	PlanB	1	3	email
5200	Cook	47	female	rural	yes	16,393	yes	PlanB	1	2	sms

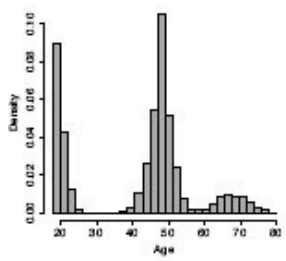
The descriptive features in this dataset are defined as follows:

- AGE: The customer's age
- GENDER: The customer's gender (*male* or *female*)
- LOC: The customer's location (*rural* or *urban*)
- OCC: The customer's occupation
- MOTORINS: Whether the customer holds a motor insurance policy with the company (*yes* or *no*)
- MOTORVALUE: The value of the car on the motor policy
- HEALTHINS: Whether the customer holds a health insurance policy with the company (*yes* or *no*)
- HEALTHTYPE: The type of the health insurance policy (*PlanA*, *PlanB*, or *PlanC*)
- HEALTHDEPSADULTS: How many dependent adults are included on the health insurance policy
- HEALTHDEPSKIDS: How many dependent children are included on the health insurance policy
- PREFCHANNEL: The customer's preferred contact channel (*email*, *phone*, or *sms*)

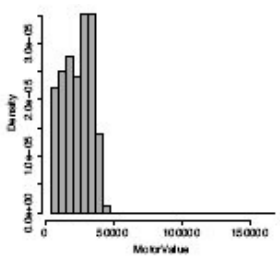
The consultant generated the following **data quality report** from the ABT (visualizations of binary features have been omitted for space saving).

Feature	Count	% Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	Max.	Std. Dev.
AGE	5,200	0	51	18	22	41.59	47	50	80	15.66
MOTORVALUE	5,200	17.25	3,934	4,352	15,089.5	23,479	24,853	32,078	166,993	11,121
HEALTHDEPSADULTS	5,200	39.25	4	0	0	0.84	1	1	2	0.65
HEALTHDEPSKIDS	5,200	39.25	5	0	0	1.77	2	3	3	1.11

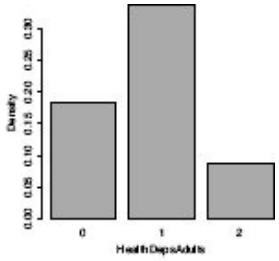
Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
GENDER	5,200	0	2	female	2,626	50.5	male	2,574	49.5
LOC	5,200	0	2	urban	2,948	56.69	rural	2,252	43.30
OCC	5,200	37.71	1,828	Nurse	11	0.34	Sales	9	0.28
MOTORINS	5,200	0	2	yes	4,303	82.75	no	897	17.25
HEALTHINS	5,200	0	2	yes	3,159	60.75	no	2,041	39.25
HEALTHTYPE	5,200	39.25	4	PlanB	1,596	50.52	PlanA	796	25.20
PREFCHANNEL	5,200	0	3	email	2,296	44.15	phone	1,975	37.98



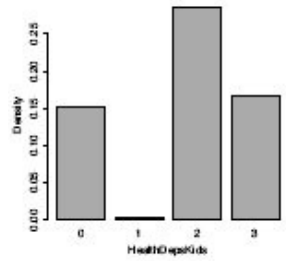
AGE



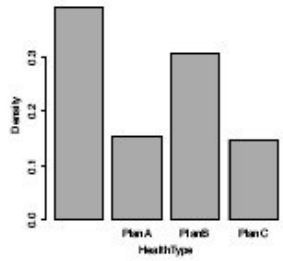
MOTORVALUE



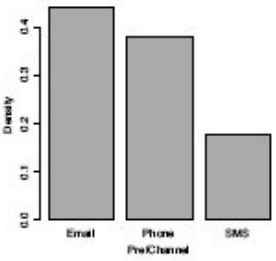
HEALTHDEPSADULTS



HEALTHDEPSKIDS



HEALTHTYPE



PREFCHANNEL

Discuss this data quality report in terms of the following:

a. Missing values

b. Irregular cardinality

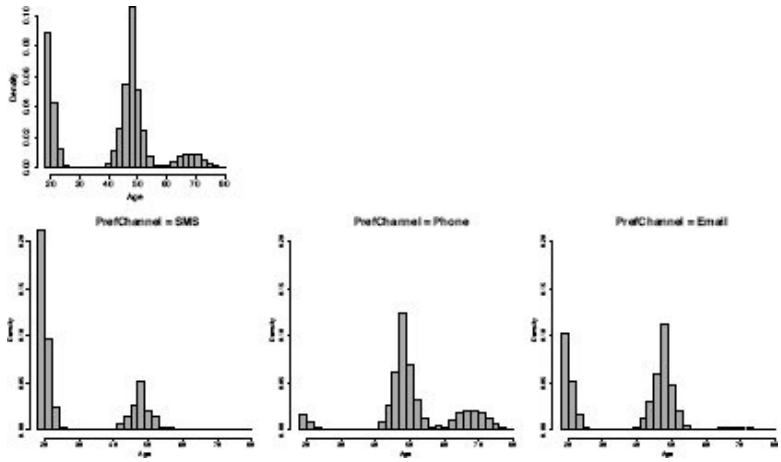
c. Outliers

d. Feature distributions

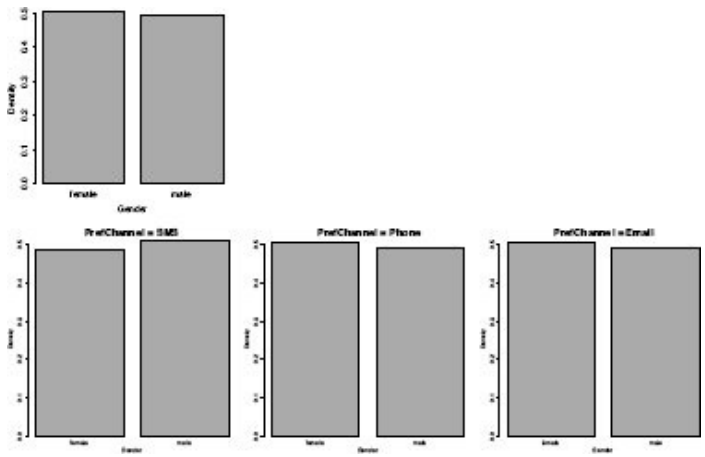
4. The following **data visualizations** are based on the channel prediction dataset given in Question 3. Each visualization illustrates the relationship between a descriptive feature and the target feature, PREFCHANNEL. Each visualization is composed of four plots: one plot of the distribution of the descriptive feature values in

the entire dataset, and three plots illustrating the distribution of the descriptive feature values for each level of the target. Discuss the strength of the relationships shown in each visualizations.

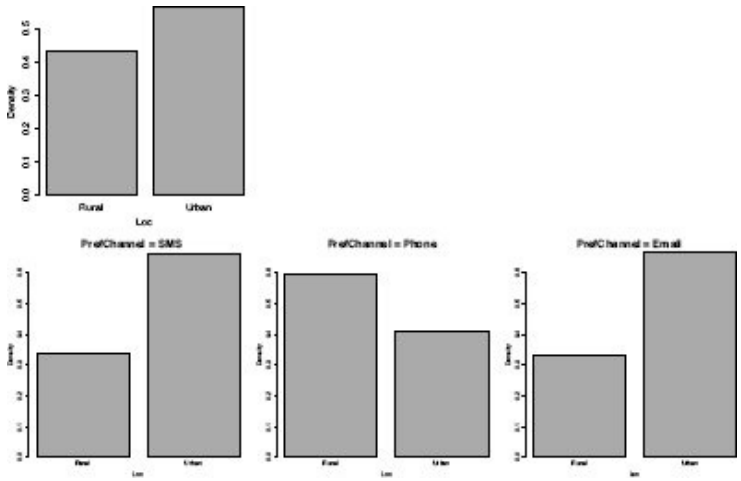
a. The visualization below illustrates the relationship between the continuous feature AGE and the target feature, PREFCHANNEL.



b. The visualization below illustrates the relationship between the categorical feature GENDER and the target feature PREFCHANNEL.



c. The visualization below illustrates the relationship between the categorical feature LOC and the target feature, PREFCHANNEL.



5. The table below shows the scores achieved by a group of students on an exam.

ID	1	2	3	4	5	6	7	8	9	10
SCORE	42	47	59	27	84	49	72	43	73	59

ID	11	12	13	14	15	16	17	18	19	20
SCORE	58	82	50	79	89	75	70	59	67	35

Using this data, perform the following tasks on the SCORE feature:

a. A **range normalization** that generates data in the range (0, 1)

b. A **range normalization** that generates data in the range (−1, 1)

c. A **standardization** of the data

6. The following table shows the IQs for a group of people who applied to take part in a television general knowledge quiz.

ID	1	2	3	4	5	6	7	8	9	10
IQ	92	107	83	101	107	92	99	119	93	106

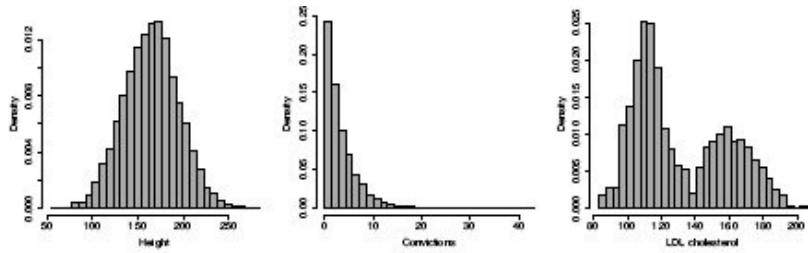
ID	11	12	13	14	15	16	17	18	19	20
IQ	105	88	106	90	97	118	120	72	100	104

Using this dataset, generate the following **binned** versions of the IQ feature:

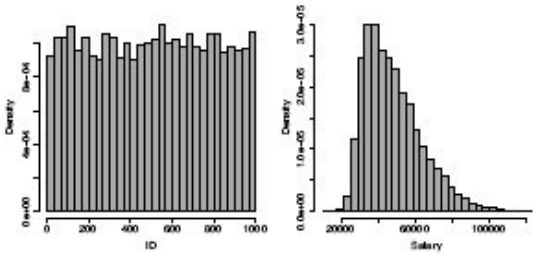
a. An **equal-width binning** using 5 bins.

b. An **equal-frequency binning** using 5 bins

✻ 7. Comment on the **distributions** of the features shown in each of the following histograms.



(a) (b) (c)



(d) (e)

a. The height of employees in a truck driving company.

b. The number of prior criminal convictions held by people given prison sentences in a city district over the course of a full year.

c. The LDL cholesterol values for a large group of patients, including smokers and non-smokers.

d. The employee ID numbers of the academic staff at a university.

e. The salaries of motor insurance policy holders.

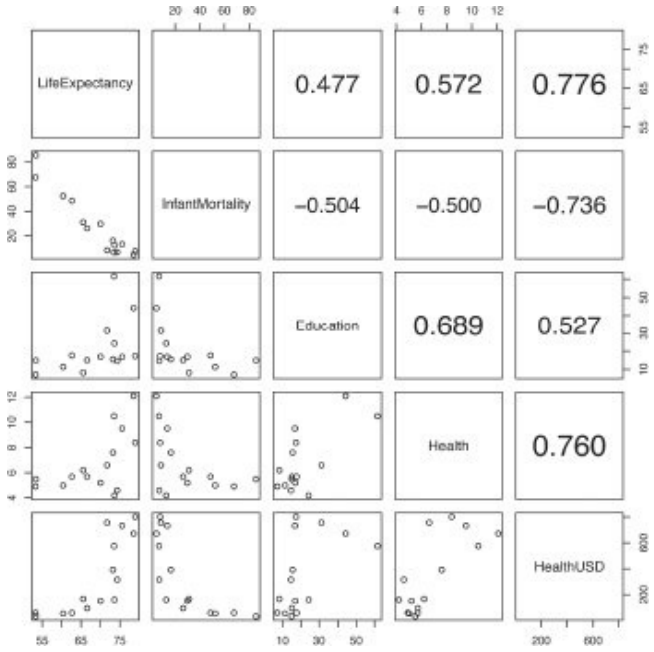
✻ 8. The table below shows socio-economic data for a selection of countries for the year 2009,[16] using the following features:

- COUNTRY: The name of the country
- LIFEEXPECTANCY: The average life expectancy (in years)
- INFANTMORTALITY: The infant mortality rate (per 1,000 live births)
- EDUCATION: Spending per primary student as a percentage of GDP
- HEALTH: Health spending as a percentage of GDP
- HEALTHUSD: Health spending per person converted into US dollars

COUNTRY	LIFE EXPECTANCY	INFANT MORTALITY	EDUCATION	HEALTH	HEALTH USD
Argentina	75.592	13.500	16.841	9.525	734.093
Cameroon	53.288	67.700	7.137	4.915	60.412
Chile	78.936	7.800	17.356	8.400	801.915
Colombia	73.213	16.500	15.589	7.600	391.859
Cuba	78.552	4.800	44.173	12.100	672.204
Ghana	60.375	52.500	11.365	5.000	54.471
Guyana	65.560	31.200	8.220	6.200	166.718
Latvia	71.736	8.500	31.364	6.600	756.401
Malaysia	74.306	7.100	14.621	4.600	316.478
Mali	53.358	85.500	14.979	5.500	33.089
Mongolia	66.564	26.400	15.121	5.700	96.537
Morocco	70.012	29.900	16.930	5.200	151.513
Senegal	62.653	48.700	17.703	5.700	59.658
Serbia	73.532	6.900	61.638	10.500	576.494
Thailand	73.627	12.700	24.351	4.200	160.136

a. Calculate the **correlation** between the LIFEEXPECTANCY and INFANT-MORTALITY features.

b. The image below shows a **scatter plot matrix** of the continuous features from this dataset (the correlation between LIFEEXPECTANCY and INFANTMORTALITY has been omitted). Discuss the relationships between the features in the dataset that this scatter plot highlights.



✻ 9. Tachycardia is a condition that causes the heart to beat faster than normal at rest. The occurrence of tachycardia can have serious implications including increased risk of stroke or sudden cardiac arrest. An analytics consultant has been hired by a major hospital to build a predictive model that predicts the likelihood that a patient at a heart disease clinic will suffer from tachycardia in the month following a visit to the clinic. The hospital will use this model to make predictions for each patient when they visit the clinic and offer increased monitoring for those deemed to be at risk. The analytics consultant has generated an ABT to be used to train this model.[17] The descriptive features in this dataset are defined as follows:

- AGE: The patient's age
- GENDER: The patient's gender (*male* or *female*)
- WEIGHT: The patient's weight
- HEIGHT: The patient's height
- BMI: The patient's body mass index (BMI) which is calculated as $\frac{weight}{height^2}$ where weight is measured in kilograms and height in meters.
- SYS. B.P.: The patient's systolic blood pressure
- DIA. B.P.: The patient's diastolic blood pressure
- HEART RATE: The patient's heart rate
- H.R. DIFF.: The difference between the patient's heart rate at this visit and at their last visit to the clinic
- PREV. TACHY.: Has the patient suffered from tachycardia before?
- TACHYCARDIA: Is the patient at high risk of suffering from tachycardia in the next month?

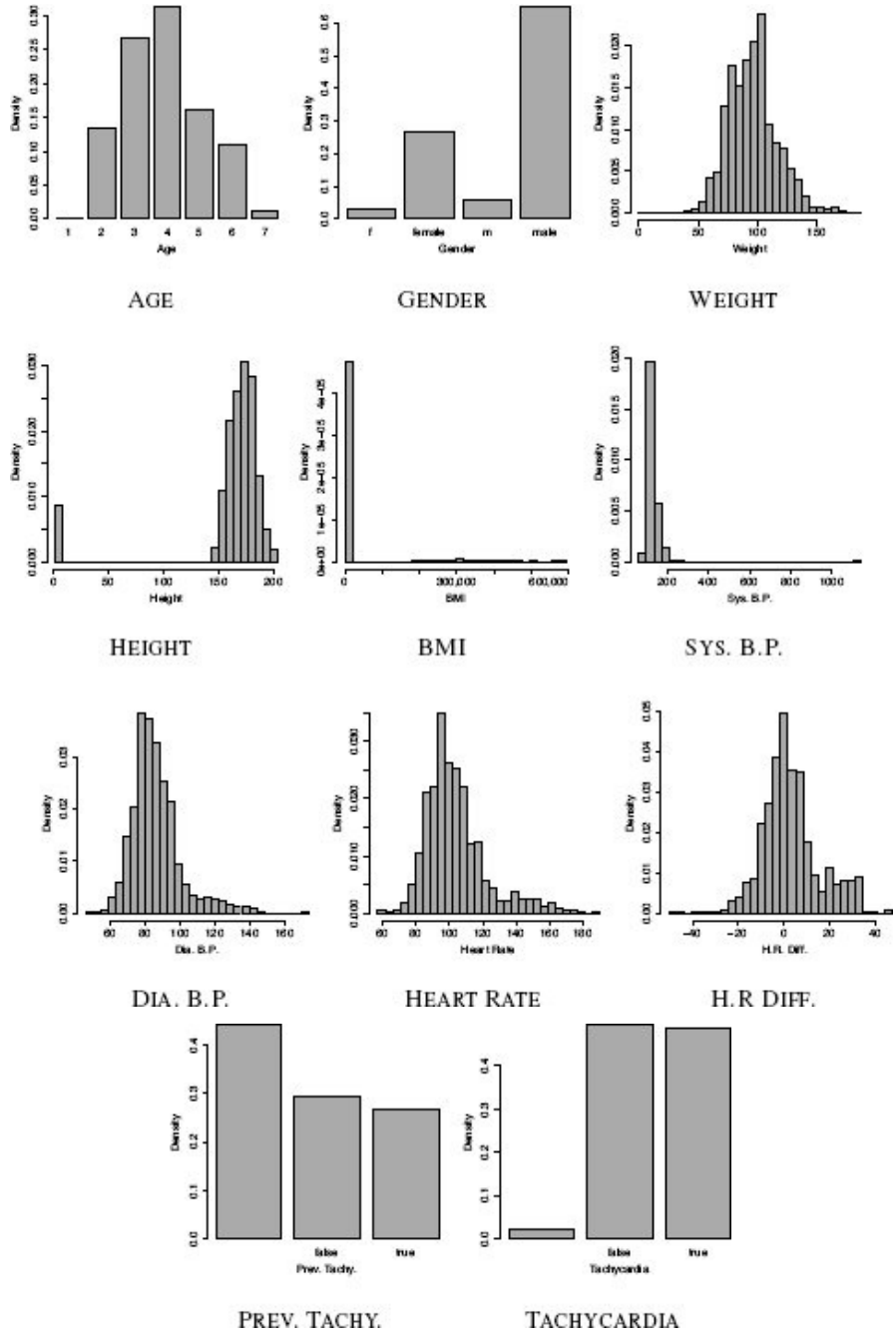
The following table contains an extract from this ABT—the full ABT contains 2,440 instances.

ID	AGE	GENDER	WEIGHT	HEIGHT	BMI	SYS. B.P.	DIA. B.P.	HEART RATE	H.R. DIFF.	PREV. TACHY.	TACHYCARDIA
1	6	male	78	165	28.65	161	97	143			true
2	5	m	117	171	40.01	216	143	162	17	true	true
		⋮			⋮				⋮		
143	5	male	108	1.88	305,568.13	139	99	84	21	false	true
144	4	male	107	183	31.95	1,144	90	94	-8	false	true
		⋮			⋮				⋮		
1,158	6	female	92	1.71	314,626.72	111	75	75	-5		false
1,159	3	female	151	1.59	596,495.39	124	91	115	23	true	true
		⋮			⋮				⋮		
1,702	3	male	86	193	23.09	138	81	83		false	false
1,703	6	f	73	166	26.49	134	86	84	-4		false
		⋮			⋮				⋮		

The consultant generated the following **data quality report** from the ABT.

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
GENDER	2,440	0.00	4	male	1,591.00	65.20	female	647.00	26.52
PREV. TACHY.	2,440	44.02	3	false	714.00	52.27	true	652.00	47.73
TACHYCARDIA	2,440	2.01	3	false	1,205.00	50.40	true	1,186.00	49.60

Feature	Count	% Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	Max.	Std. Dev.
AGE	2,440	0.00	7	1.00	3.00	3.88	4.00	5.00	7.00	1.22
WEIGHT	2,440	0.00	174	0.00	81.00	95.70	95.00	107.00	187.20	20.89
HEIGHT	2,440	0.00	109	1.47	162.00	162.21	171.50	179.00	204.00	41.06
BMI	2,440	0.00	1,385	0.00	27.64	18,523.40	32.02	38.57	596,495.39	77,068.75
SYS .B.P.	2,440	0.00	149	62.00	115.00	127.84	124.00	135.00	1,144.00	29.11
DIA. B.P.	2,440	0.00	109	46.00	77.00	86.34	84.00	92.00	173.60	14.25
HEART RATE	2,440	0.00	119	57.00	91.75	103.28	100.00	110.00	190.40	18.21
H.R. DIFF.	2,440	13.03	78	-50.00	-4.00	3.00	1.00	8.00	47.00	12.38



Discuss this data quality report in terms of the following:

a. Missing values

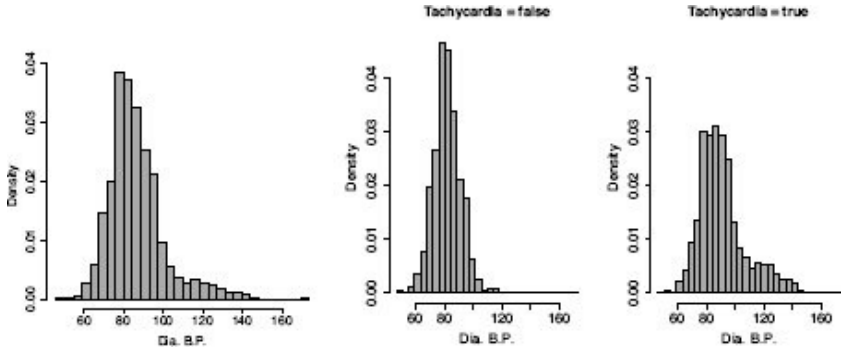
b. Irregular cardinality

c. Outliers

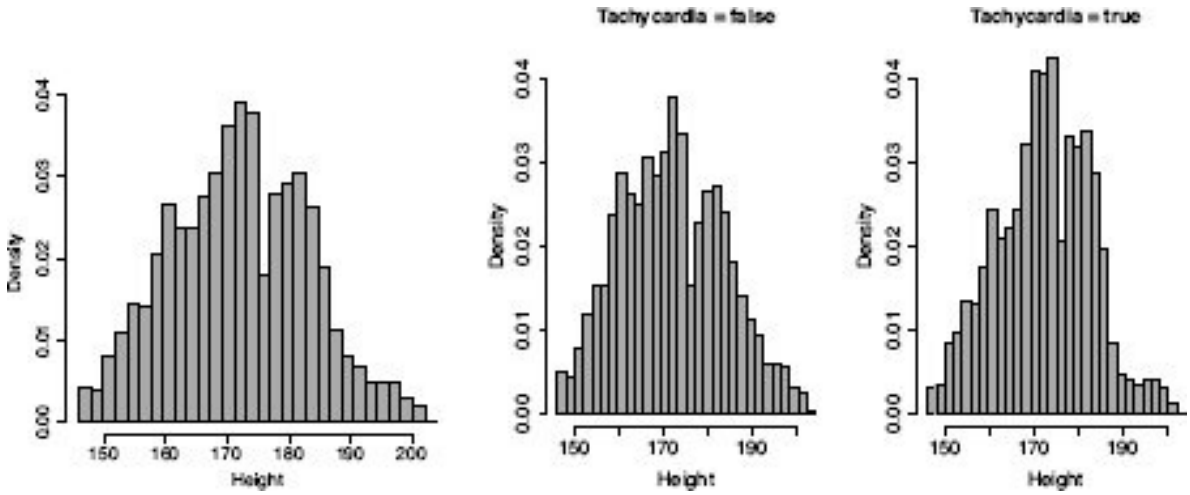
d. Feature distributions

✻ 10. The following data visualizations are based on the tachycardia prediction dataset from Question 9 (after the instances with missing TACHYCARDIA values have been removed and all outliers have been handled). Each visualization illustrates the relationship between a descriptive feature and the target feature, TACHYCARDIA and is composed of three plots: a plot of the distribution of the descriptive feature values in the full dataset, and plots showing the distribution of the descriptive feature values for each level of the target. Discuss the relationships shown in each visualizations.

a. The visualization below illustrates the relationship between the continuous feature DIA. B.P. and the target feature, TACHYCARDIA.



b. The visualization below illustrates the relationship between the continuous HEIGHT feature and the target feature TACHYCARDIA.



c. The visualization below illustrates the relationship between the categorical feature PREV. TACHY. and the target feature, TACHYCARDIA.

Frequency
1000
600
200
0
1
Percent
0.5
0
false
true
Prev. Tachy.
Tachycardia
false
true