Assignment 1-Aparna Bharathi Suresh

Question 1- Chapter 2, exercise 7

Select one of the predictive analytics models that you proposed in your answer

to the previous question about the oil exploration company for exploration of the

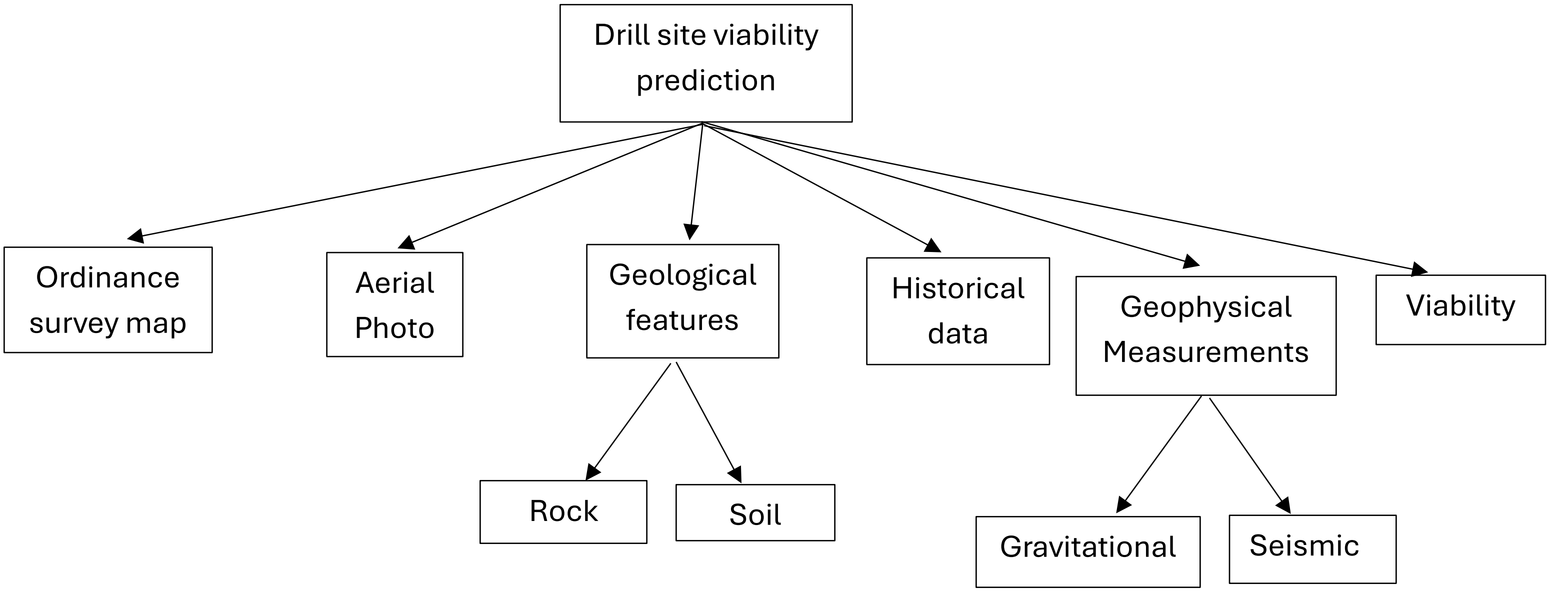
design of its **analytics base table**.

a. What is the prediction subject for the model that will be trained using this ABT?
b. Describe the domain concepts for this ABT.
c. Draw a domain concept diagram for the ABT.
d. Are there likely to be any legal issues associated with the domain concepts you have included?

Answer 1:

a. The prediction subject is the potential drilling site for viable oil wells.
b. Domain concepts for this ABT are:
    - **Ordinance survey maps**: They provide information about the features such as roads, paths, buildings, hills, water bodies, and land boundaries.
    - **Aerial photographs**: Aerial photographs are images of the Earth's surface taken from above the ground, typically from an aircraft, drone, or satellite.
    - **Geological features**: Characteristics of rock and soil samples taken from potential sites.
    - **Historical Success Rate:** Historical data on whether nearby sites yielded viable wells.
    - **Geophysical Measurements:** Gravity and Seismic Measurements from special instruments.
    - **Viability:** Target feature.

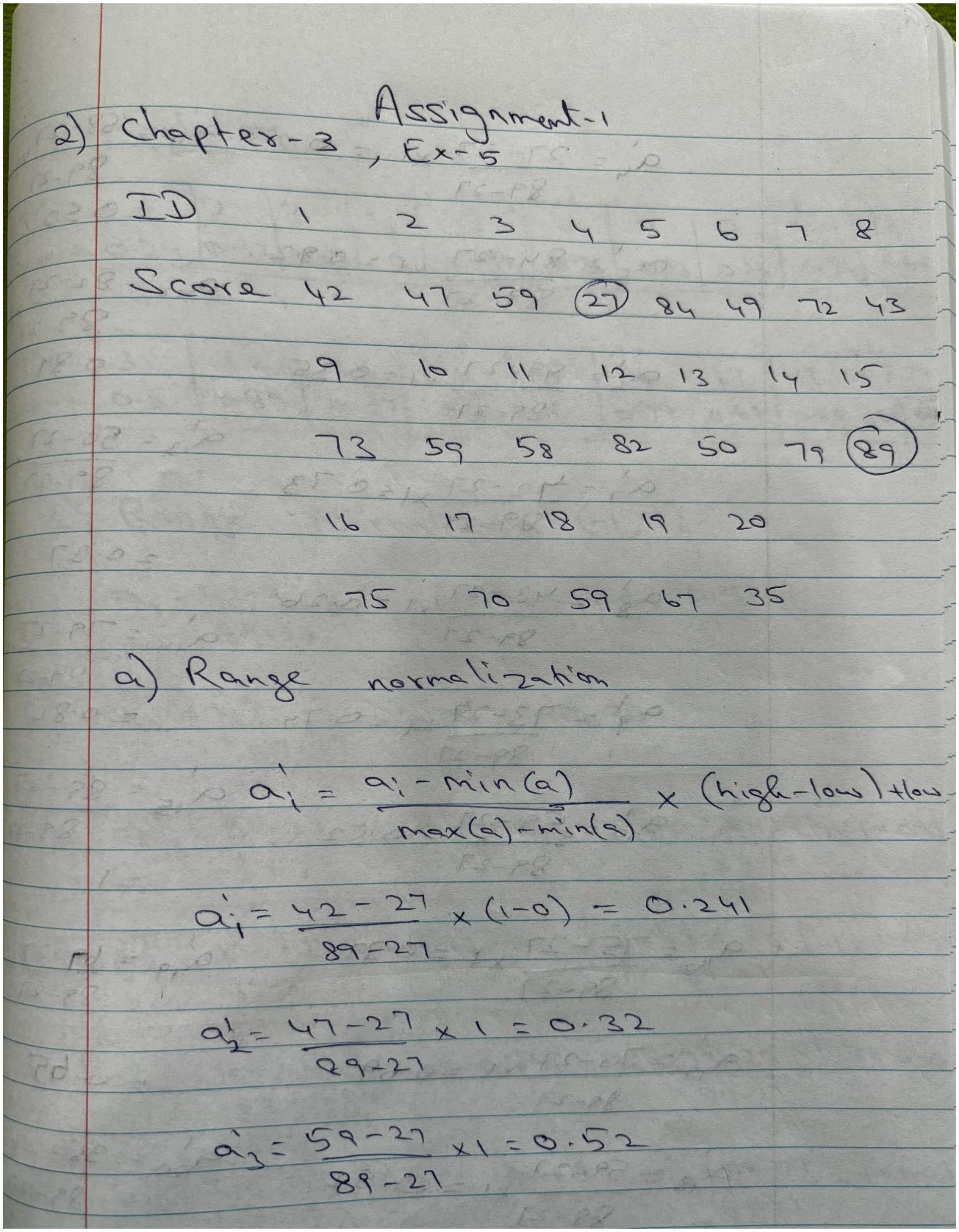
c. Domain concept diagram for the ABT



d. Legal issues associated:

- Depending on the location of the sites, there may be legal constraints regarding the collection and use of geological and seismic data from privately owned land.
- Drilling operations can have environmental consequences. So, the model may need to consider factors related to environmental impact assessments and regulations.

Question 2- Chapter 3, exercise 5



Assignment-1

2) Chapter-3, Ex-5

ID	1	2	3	4	5	6	7	8
Score	42	47	59	27	84	49	72	43
	9	10	11	12	13	14	15	
	73	59	58	82	50	79	89	
	16	17	18	19	20			
	75	70	59	67	35			

a) Range normalization

$$a_i' = \frac{a_i - \min(a)}{\max(a) - \min(a)} \times (\text{high} - \text{low}) + \text{low}$$

$$a_1' = \frac{42 - 27}{89 - 27} \times (1-0) = 0.241$$

$$a_2' = \frac{47 - 27}{89 - 27} \times 1 = 0.32$$

$$a_3' = \frac{59 - 27}{89 - 27} \times 1 = 0.52$$

$$a'_4 = \frac{27-27}{89-27} \times 1 = 0$$

$$a'_5 = \frac{84-27}{89-27} \times 1 = 0.92$$

$$a'_6 = \frac{49-27}{89-27} \times 1 = 0.35$$

$$a'_7 = \frac{72-27}{89-27} \times 1 = 0.73$$

$$a'_8 = \frac{43-27}{89-27} \times 1 = 0.26$$

$$a'_9 = \frac{73-27}{89-27} \times 1 = 0.74$$

$$a'_{10} = \frac{59-27}{89-27} = 0.52$$

$$a'_{11} = \frac{58-27}{89-27} \times 1$$

$$= 0.50$$

$$a'_{12} = \frac{82-27}{89-27} \times 1$$

$$= 0.89$$

$$a'_{13} = \frac{50-27}{89-27} \times 1$$

$$= 0.37$$

$$a'_{14} = \frac{79-27}{89-27} \times 1$$

$$= 0.84$$

$$a'_{15} = \frac{89-27}{89-27} \times 1$$

$$= 1$$

$$a'_{16} = \frac{75-27}{89-27} \times 1 = 0.77$$

$$a'_{17} = \frac{70-27}{89-27} \times 1 = 0.69$$

$$a'_{18} = \frac{59-27}{89-27} \times 1 = 0.52$$

$$a'_{19} = \frac{69-27}{89-27} \times 1$$

$$= 0.65$$

$$a'_{20} = \frac{35-27}{89-27} \times 1$$

$$= 0.13$$

Solution:

ID	1	2	3	4	5	6	7	8	9	10
Score	0.24	0.32	0.52	0	0.92	0.35	0.73	0.26	0.74	0.52

ID	11	12	13	14	15	16	17	18	19	20
Score	0.5	0.89	0.37	0.84	1	0.77	0.69	0.52	0.65	0.13

b) Range Normalization (-1, 1)

$$a'_1 = \frac{42-27}{89-27} \times (1-(-1))+(-1) = -0.5161$$

$$a'_2 = \frac{47-27}{89-27} \times 2 - 1 = -0.35$$

$$a'_3 = \frac{59-27}{89-27} \times 2 - 1 = 0.03 \qquad a'_7 = \frac{72-27}{89-27} \times 2 - 1 = 0.45$$

$$a'_4 = \frac{27-27}{89-2} \times 2 - 1 = -1 \qquad a'_8 = \frac{43-27}{89-27} \times 2 - 1 = -0.48$$

$$a'_5 = \frac{84-27}{89-27} \times 2 - 1 = 0.84 \qquad a'_9 = \frac{73-27}{89-27} \times 2 - 1 = 0.48$$

$$a'_6 = \frac{49-27}{89-27} \times 2 - 1 = -0.29 \qquad a'_{10} = \frac{59-27}{89-27} \times 2 - 1 = 0.03$$

$a'_{11} = \frac{58-27}{89-27} \times 2 - 1 = 0$

$a'_{16} = \frac{75-27}{89-27} \times 2 - 1 = 0.55$

$a'_{12} = \frac{82-27}{89-27} \times 2 - 1 = 0.7$

$a'_{17} = \frac{70-27}{89-27} \times 2 - 1 = 0.39$

$a'_{13} = \frac{50-27}{89-27} \times 2 - 1 = -0.26$

$a'_{18} = \frac{59-27}{89-27} \times 2 - 1 = 0.03$

$a'_{14} = \frac{79-27}{89-27} \times 2 - 1 = 0.68$

$a'_{19} = \frac{67-27}{89-27} \times 2 - 1 = 0.29$

$a'_{15} = \frac{89-27}{89-27} \times 2 - 1 = 1$

$a'_{20} = \frac{35-27}{89-27} \times 2 - 1 = -0.74$

Solution:

ID	1	2	3	4	5	6	7	8	9	10
Score	-0.52	-0.35	0.03	-1	0.84	-0.29	0.45	-0.48	0.48	0.03

ID	11	12	13	14	15	16	17	18	19	20
Score	0	0.77	-0.26	0.68	1	0.55	0.39	0.03	0.29	-0.74

c) Standardization

$$a'_i = \frac{a_i - \bar{a}}{sd(a)}$$

$\bar{a} = \Sigma \text{scores} / N = 60.95$

$$Sd(a) = \sqrt{\frac{\Sigma(a_i - \bar{a})^2}{N-1}} = 17.2519$$

$$= \sqrt{\frac{5654.95}{19}} = \sqrt{297.628} = 17.2519$$

$$a_1 = \frac{42 - 60.95}{17.2519} = -1.0984$$

$$a_2 = \frac{47 - 60.95}{17.2519} = -0.81$$

$$a_3 = \frac{59 - 60.95}{17.2519} = -0.11$$

$$a_4 = \frac{27 - 60.95}{17.2519} = -1.97$$

$$a_5 = \frac{84 - 60.95}{17.2519} = 1.34$$

$$a_6 = \frac{49 - 60.95}{17.2519} = -0.69$$

$$a_7 = \frac{72 - 60.95}{17.2519} = 0.64$$

$$a_8 = \frac{43 - 60.95}{17.2519} = -1.04$$

$$a_9 = \frac{73 - 60.95}{17.2519} = 0.70$$

$$a_{10} = \frac{59 - 60.95}{17.2519} = -0.11$$

$$a_{11} = \frac{58 - 60.95}{17.2519} = -0.17$$

$a_{12} = \frac{82-60.95}{17.2519} = 1.22$

$a_{13} = \frac{50-60.95}{17.2519} = -0.63$

$a_{14} = \frac{79-60.95}{17.2519} = 1.05$

$a_{15} = \frac{89-60.95}{17.2519} = 1.63$

$a_{16} = \frac{75-60.95}{17.2519} = 0.81$

$a_{17} = \frac{70-60.95}{17.2519} = 0.52$

$a_{18} = \frac{59-60.95}{17.2519} = -0.11$

$a_{19} = \frac{67-60.95}{17.2519} = 0.35$

$a_{20} = \frac{35-60.95}{17.2519} = -1.50$

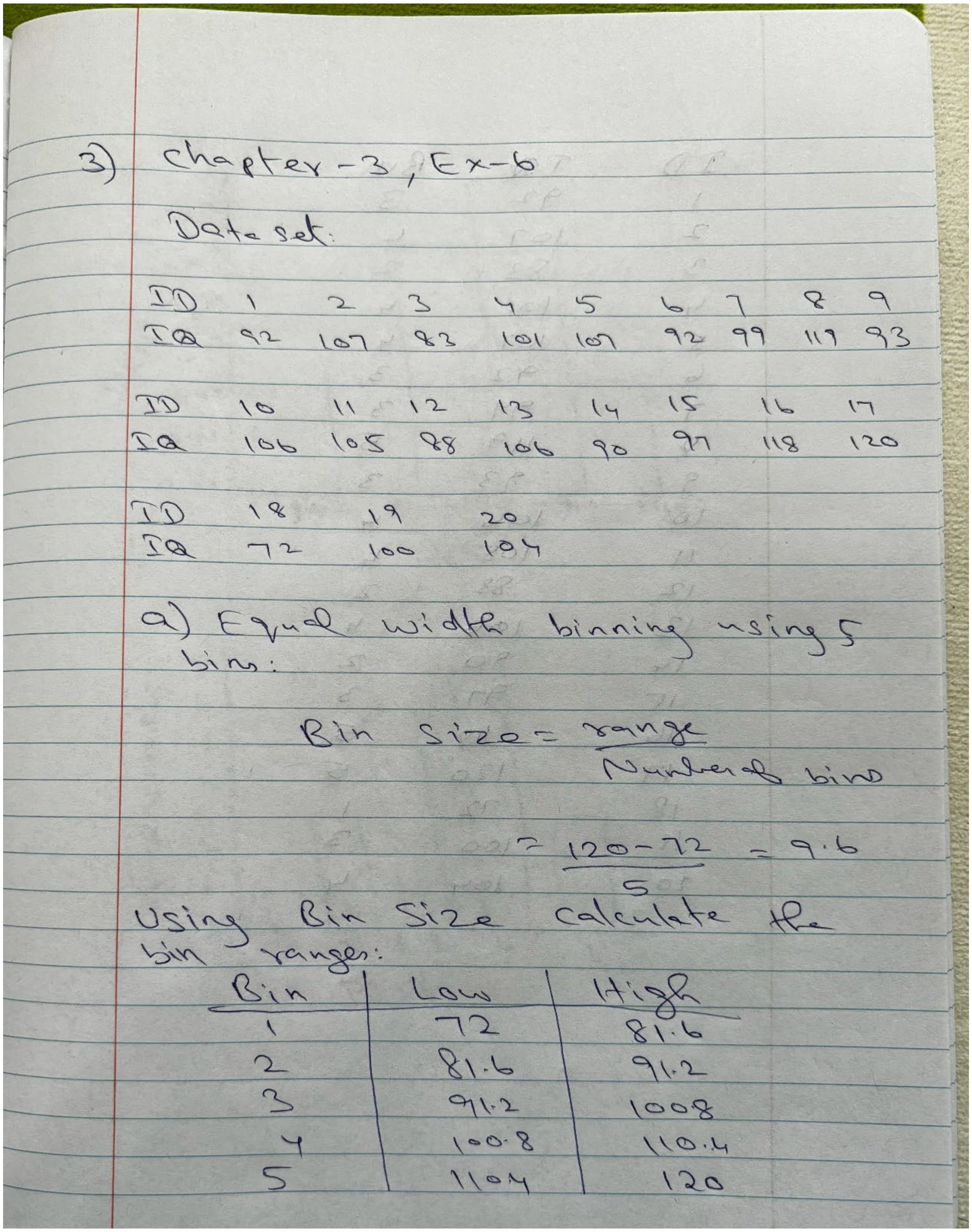
Solution:

ID	1	2	3	4	5	6	7	8	9	10
Score	-1.10	-0.81	-0.11	-1.97	1.34	-0.69	0.64	-1.04	0.70	-011

ID	11	12	13	14	15	16	17	18	19
Score	-0.17	1.22	-0.63	1.05	1.63	0.81	0.52	-0.11	0.35

20
-1.50

Question 3- Chapter 3, exercise 6



3) chapter-3, Ex-6

Data set:

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

ID	10	11	12	13	14	15	16	17
IQ	106	105	88	106	90	97	118	120

ID	18	19	20
IQ	72	100	104

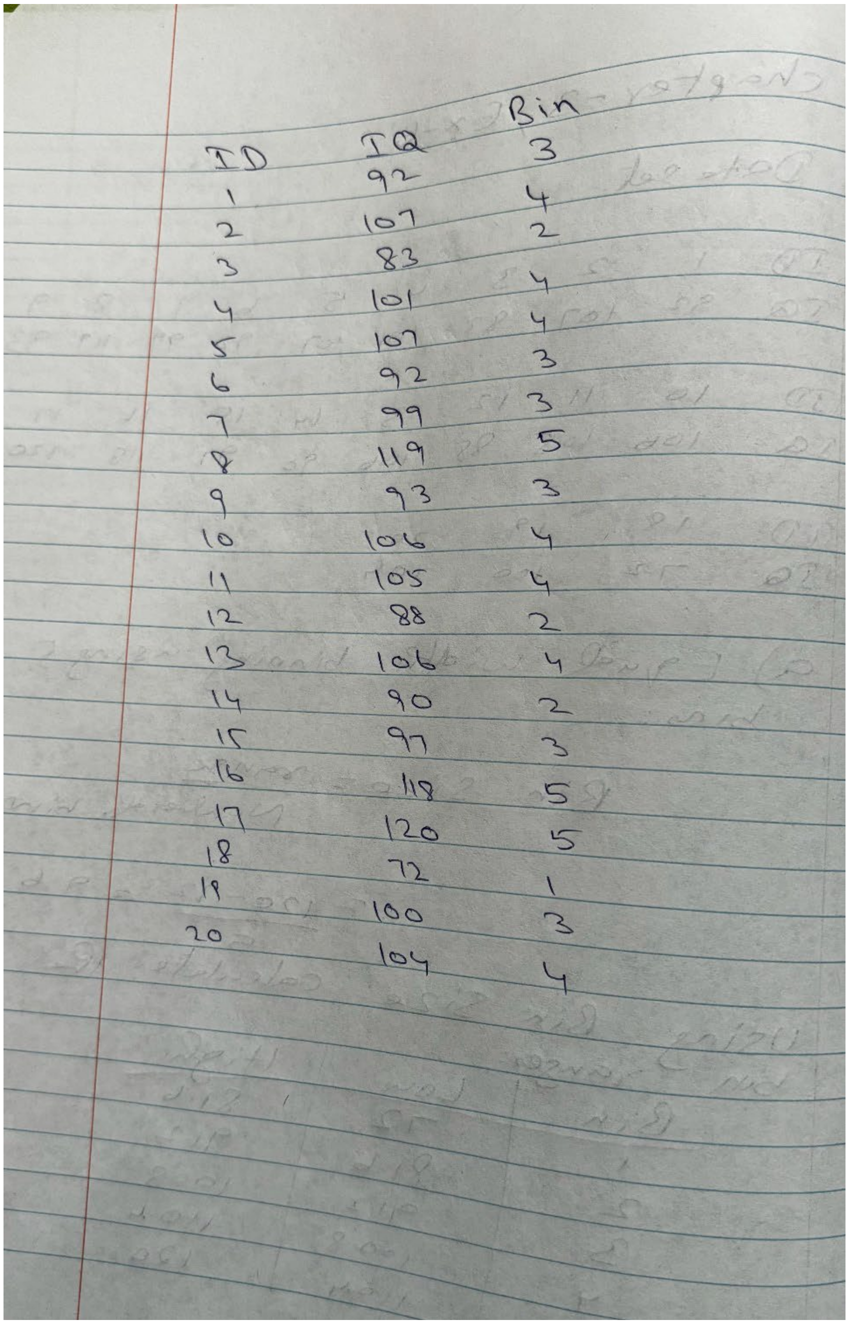
a) Equal width binning using 5 bins:

$$\text{Bin Size} = \frac{\text{range}}{\text{Number of bins}}$$

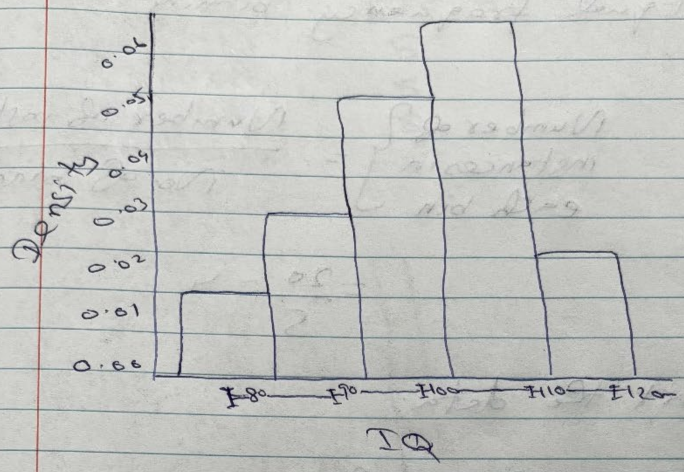
$$= \frac{120-72}{5} = 9.6$$

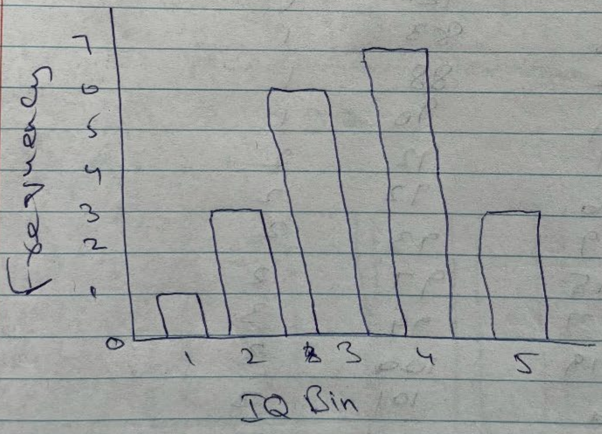
Using Bin Size calculate the bin ranges:

Bin	Low	High
1	72	81.6
2	81.6	91.2
3	91.2	1008
4	100.8	110.4
5	1104	120



ID	IQ	Bin
1	92	3
2	107	4
3	83	2
4	101	4
5	107	4
6	92	3
7	99	3
8	119	5
9	93	3
10	106	4
11	105	4
12	88	2
13	106	4
14	90	2
15	97	3
16	118	5
17	120	5
18	72	1
19	100	3
20	104	4


Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
80
90
100
110
120
IQ


Frequency
7
6
5
4
3
2
1
0
1
2
3
4
5
IQ Bin

b) Equal frequency binning

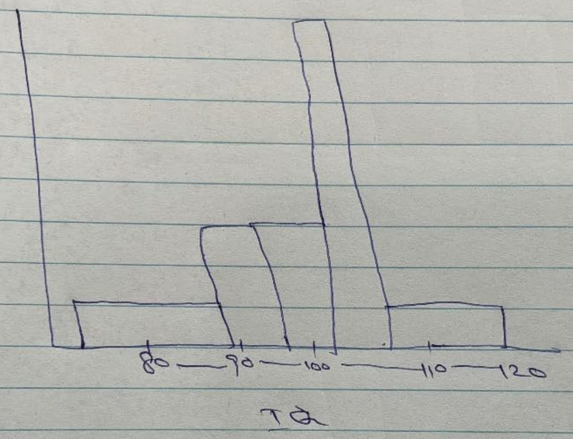
$$\left.\begin{array}{l}\text{Number of}\\ \text{instances in}\\ \text{each bin}\end{array}\right\} = \frac{\text{Number of instances}}{\text{No. of bins}}$$

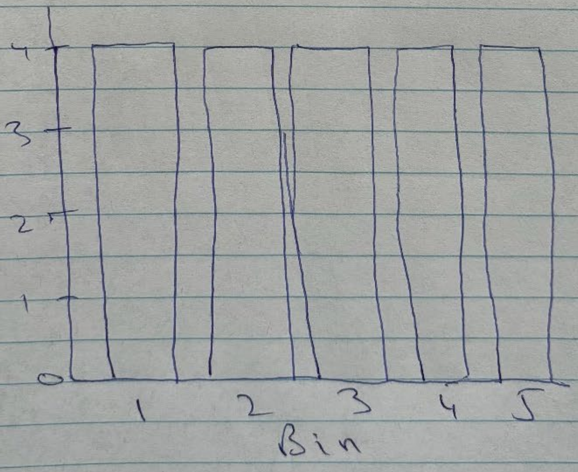
$$= \frac{20}{5} = 4$$

Sort the data:

ID	IQ	BIN
18	72	1
3	83	1
12	88	1
14	90	1
1	92	2
6	92	2
9	93	2
15	97	2
7	99	3
19	100	3
4	101	3
20	104	3
11	105	4
10	106	4
13	106	4
2	107	4

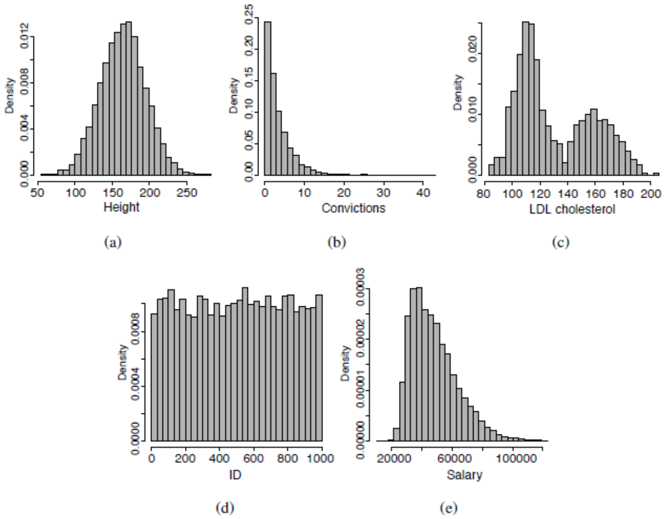
ID	IQ	Bin
5	107	5
16	118	5
8	119	5
17	120	5





Question 4- Chapter 3, exercise 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.

Answer 4:

a. The histogram of height of employees appears to be a bell curve in the given diagram, so it follows normal distribution with most of the heights around the mean value, the mean value would be approximately 175 from the diagram. Only a few heights will deviate from the mean value as it follows a normal distribution.
b. The histogram of prior criminal convictions appears to be rightly skewed in the given diagram, most of the people have very few convictions and a few people

have large number of convictions. There may be outliers too, having a very large number.

c. The histogram of LDL cholesterol is bimodal in the given diagram, indicating that there are two groups' Smokers and Nonsmokers. The central tendency of one group is around 110 and the central tendency of another is 160. Smoking is known to lower HDL cholesterol, which is considered "good" cholesterol. So, the average LDL cholesterol of smoking group is 110 and the average LDL cholesterol of nonsmoking group is 160. According to the given diagram, the number of smokers is greater than the number of nonsmokers in the company.

d. The histogram of employee ID number appears to have uniform distribution according to the given diagram. Each ID occurs with the same frequency across the dataset.

e. The histogram of Salary appears to be rightly skewed in the given diagram, with most policy holders earning lower salaries having a central tendency around 35000, but a few policy holders having higher salaries.

Question 5- Chapter 3, exercise 9

Discuss this data quality report in terms of the following:

a. Missing values

b. Irregular cardinality

c. Outliers

d. Feature distributions

Answer 5:

a. Missing values:

According to the data quality report from the ABT mentioned in the question there are 3 features that have missing data:

- PREV. TACHY - 44.02% of missing data - For PREV. TACHY, the missing percentage is too high so to handle the missing data we must remove the feature instead of doing imputations, as imputations will make lots of changes to the data.
- TACHYCARDIA – 2.01% of missing data - Since TACHYCARDIA is the target feature, we can't do imputation, so we must remove the rows that have missing values for TACHYCARDIA.

- H.R. DIFF. - 13.03% of missing data - To handle the missing data in H.R. DIFF., we can do mean or median imputation as the missing data is not very high.

b. Cardinality:

- Most of the numeric features have cardinalities less than 2440, because most of the features have defined ranges, for example diastolic blood pressure values should fall between 40 and 100, systolic blood pressure values should fall between 90 to 200, weight should fall between 50 and 200, height falls between 150 and 200, heart rate falls between 60 and 180.
- The cardinality for AGE is 7, it is very low for a numerical feature like age which suggests that AGE in this data set is not the actual values, age is categorized according to this data set.
- Gender has a cardinality of 4, from the bar plot we see that there are 4 values of gender Male, Female, M and F.
- The cardinality of PREV. TACHY and TACHYCARDIA are 3, which looks correct, as they have True, False and Null values.

c. Outliers:

- From the visualization diagrams and the table, we can see that there are 3 features with outliers that are HEIGHT, BMI, and SYS. B.P.
- The minimum value of height mentioned in the table is 1.47 which is an outlier.
- The values of BMI for many rows are incredibly high, the maximum value of BMI in the given data set is 596,495.39 which is clearly an outlier. The mean value is also very high 18,523.40. The visualization also shows that BMI has lots of outliers. The normal range of BMI is somewhere between 15 to 40.
- The max value of SYS. B.P is 1,144 which is an outlier. The normal range of SYS.B. P is 90 to 200.

d. Feature distributions:

- AGE:
  The histogram for age shows a distribution that is not symmetric. There appear to be peaks for specific age ranges. It appears to be lightly skewed with most of the data concentrated around certain age groups.
- GENDER:
  The distribution for gender is highly skewed. There are far more males than females.
- WEIGHT:
  The histogram for weight shows almost a normal distribution, which indicates that most individuals have a weight close to the mean, with fewer individuals who have either very low or very high weight. Has outliers.

- HEIGHT:
  The histogram for height is relatively normal but slightly skewed, most individuals fall within a specific height range, with fewer outliers who are much taller or shorter. Has outliers.
- BMI:
  The Body Mass Index (BMI) distribution shows a peak around a specific range. And there is a tail to the right, indicating that there are fewer individuals with much higher BMI values.
- SYSTOLIC BLOOD PRESSURE (SYS. B.P.):
  The systolic blood pressure values appear to be skewed, with most people having blood pressure around a particular range. There are fewer individuals with very high or very low blood pressure.
- DIASTOLIC BLOOD PRESSURE (DIA. B.P.):
  The distribution of diastolic blood pressure is also approximately normal, but slightly skewed. There is a central peak where most of the values lie, with fewer data points on either side (high and low blood pressure).
- HEART RATE:
  Heart rate data shows a roughly normal distribution, with most individuals' heart rates clustering around a central value. Fewer individuals have extreme heart rates.
- HEART RATE DIFFERENCE:
  The histogram for the heart rate difference shows a skewed distribution, with most values concentrated around lower values and a tail extending towards higher values. This distribution suggests that a few individuals experience a larger heart rate difference, but for most, the difference is smaller.
- PREV. TACHY:
  The distribution is almost uniform if the null values are removed.
- TACHYCARDIA:
  The distribution is almost uniform if the null values are removed, both values have almost the same density.