DATA 245: Machine Learning

Homework -1

Submitted By: Poojan Gagrani

SJSU ID: 016795285

Q1. 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?

Ans: The prediction model subject is 'Potential Drilling Sites' and ABT is the "Likelihood of Exploration Success" for each potential drilling site. In other words, the model will aim to predict the probability or likelihood that drilling at a particular site will lead to finding viable oil wells or exploration success.

Base Table: Potential Drilling Sites

Attribute	Description
Site_ID	Unique identifier for each potential site
Location	Geographic coordinates or address
Geological_Features	Characteristics of rock and soil samples
Seismic_Data	Measurements from seismic instruments
Gravitational_Data	Measurements from gravitational instruments
Exploration_Cost	Cost of exploratory drilling at the site
Other_Relevant_Features	Additional site-specific data
Historical_Success	Binary indicator (Yes or No) for past success

As the target variable in the predictive model, the "Historical_Success" attribute in the base table serves as the label. Models will rely on this attribute, along with geological features, seismic information, and gravitational data, to predict future drilling success or failure at new potential sites.

As a binary classification problem, each potential site can be assigned a probability of success, and a threshold can be set to classify sites as either likely to succeed or not. An oil exploration company will benefit from this prediction since it allows them to prioritize and allocate resources to the most promising drilling sites, enabling them to reduce costs and improve efficiency.

b) Describe the domain concepts for this ABT?

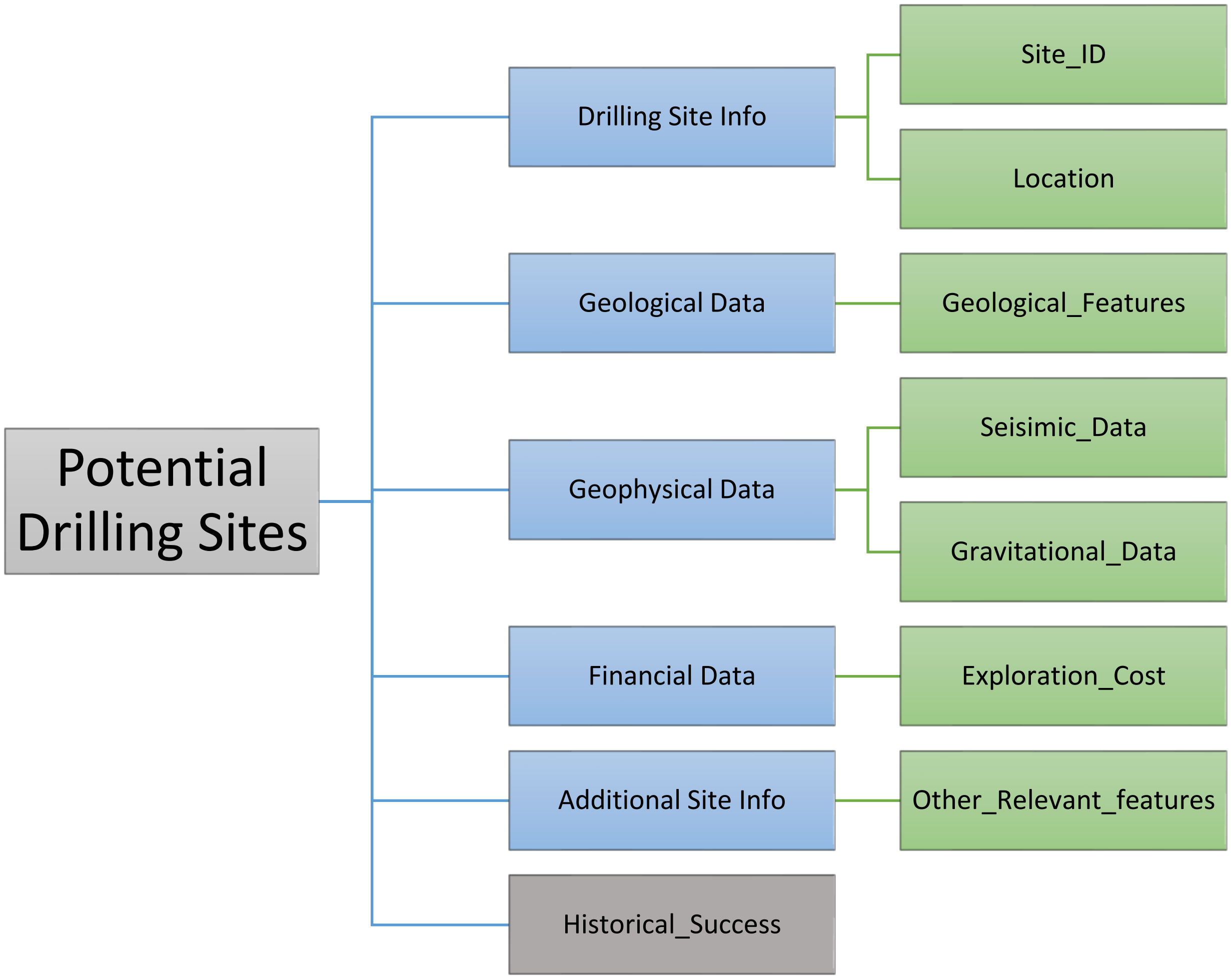
Ans: In oil exploration, the "Potential Drilling Sites" base table includes a set of key terms and concepts related to identifying, exploring, and selecting drilling sites. Data in the table can be understood and manipulated using these concepts. Following are the domain concepts:

- **Site_ID:** A unique identifier for each potential drilling site, allowing for easy reference and tracking of site-specific data.
- **Location:** The geographic coordinates or address of the drilling site, providing information about the site's physical location on the Earth's surface.
- **Geological_Features:** Characteristics of rock and soil samples obtained from the potential site, which may include data on rock composition, porosity, permeability, and other geological properties.
- **Seismic_Data:** Measurements obtained from seismic instruments, which help in understanding subsurface structures, seismic waves, and potential oil reservoirs.
- **Gravitational_Data:** Measurements collected from gravitational instruments, providing information about subsurface density variations and geological anomalies.
- **Exploration_Cost:** The cost associated with conducting exploratory drilling at the site, which includes expenses related to equipment, labor, permits, and other exploration activities.
- **Other_Relevant_Features:** Additional site-specific data that may be relevant to the drilling site selection process. This can include environmental factors, regulatory constraints, proximity to infrastructure, or any other information deemed important.
- **Historical_Success:** A binary indicator (Yes or No) that represents whether previous exploratory drilling at the site was successful (Yes) in finding viable oil wells or not (No).

These domain ideas work together to create the framework for comprehending and interpreting data pertaining to probable drilling sites. Making educated choices about where to allocate resources and carry out exploratory drilling in the oil exploration process depends on the knowledge included in these concepts.

c) Draw a domain concept diagram for the ABT?

Ans: The domain concept diagram for the ‘Potential Drilling Sites’ can be represented as following:



d) Are there likely to be any legal issues associated with the domain concepts you have included?

Ans: There is likelihood of legal and regulatory issues associated with the domain concepts included in the ‘Potential Drilling Sites’, as the exploration and extraction of oil and natural resources are subject to various laws and regulations in most countries. Some of the potential issues associated could be as follows:

1. **Environmental Regulations:** These regulations often require companies to conduct environmental impact assessments, obtain permits, and adhere to specific environmental protection measures.

2. **Property Rights and Land Ownership:** Land ownership and property rights can pose a legal issue. Hence, companies must ensure to have the appropriate rights and permissions to access and drill on private or public lands.

3. **Data Privacy and Security:** Handling of sensitive geological, geophysical, and location data must adhere to data privacy laws, especially if the data contains personally identifiable information or confidential business information.

4. **Financial Reporting and Taxation:** Accurate financial reporting and adherence to taxation laws are essential. Exploration costs, revenue, and taxes associated with drilling sites may have legal implications.

Q2. 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	11	12	13	14	15	16	17	18	19	20
SCORE	42	47	59	27	84	49	72	43	73	59	58	82	50	79	89	75	70	59	67	35

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

Ans: The formula to perform range normalization in the range (0,1) is:

Normalized Score = (X - X_min) / (X_max - X_min)

Where,

X = Original Score

X_min = Minimum Score

X_max = Maximum Score

The range normalization for each score in the range of (0,1) are:

ID	Score	Normalized Score
1	42	0.242
2	47	0.323
3	59	0.516
4	27	0.0
5	84	0.919
6	49	0.355

7	72	0.726
8	43	0.258
9	73	0.742
10	59	0.516
11	58	0.5
12	82	0.887
13	50	0.371
14	79	0.839
15	89	1.0
16	75	0.774
17	70	0.694
18	59	0.516
19	67	0.645
20	35	0.129

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

Ans: The formula to perform range normalization in the range (-1,1) is:

Normalized Score = 2 * (X - X_min) / (X_max - X_min) - 1

Where,

X = Original Score

X_min = Minimum Score

X_max = Maximum Score

The range normalization for each score in the range of (-1,1) are:

ID	Score	Normalized Score
1	42	-0.516
2	47	-0.354
3	59	0.032
4	27	-1.0
5	84	0.838
6	49	-0.29
7	72	0.451
8	43	-0.483
9	73	0.483
10	59	0.032
11	58	0.0
12	82	0.774
13	50	-0.258
14	79	0.6774

15	89	1.0
16	75	0.548
17	70	0.387
18	59	0.032
19	67	0.290
20	35	-0.741

c) A standardization of the data?

Ans: The formula to perform standardization or z-score normalization is:

Standardization (z-score) = (X -µ) / σ

Where,

X = Original Score

µ = Mean of the score

σ = Standard deviation of the scores

The standardization for each score is:

ID	**Score**	**Standardized Score**
1	42	-1.098
2	47	-0.809
3	59	-0.113
4	27	-1.968
5	84	1.336
6	49	-0.693
7	72	0.641
8	43	-1.04
9	73	0.698
10	59	-0.113
11	58	-0.171
12	82	1.22
13	50	-0.635
14	79	1.046
15	89	1.626
16	75	0.814
17	70	0.525
18	59	-0.113
19	67	0.351
20	35	-1.504

Q3. 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

11	12	13	14	15	16	17	18	19	20
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?

Ans: Equal width binning can be defined as a preprocessing technique in which primary goal is to divide a range of continuous data values into intervals or bins of equal width.

It can be calculated as following,

Width = range/b

Where,

Range = max value – min value

b = number of bins selected

Hence, equal-width binning of IQ can be defined as following:

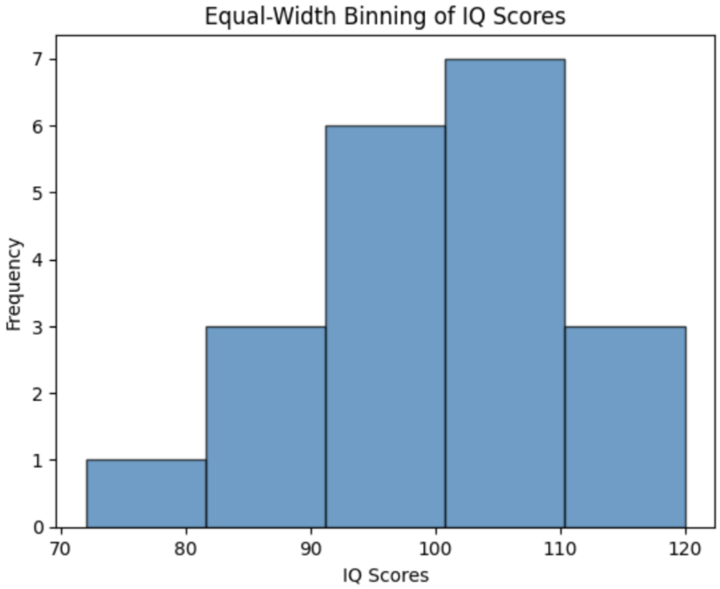
Width = 9.6

Range = 120 – 72

b = 5

Bin No.	Range
1	72-81.6
2	81.6-91.2
3	91.2-100.8
4	100.8-110.4
5	110.4-120

It can be visually represented as,



b) An equal-frequency binning using 5 bins?

Ans: Equal frequency can be defined as a preprocessing technique in which the primary goal is to divide a dataset into bins or intervals in such a way that each bin contains approximately the same number of data points.

It can be done as following,

First, we have to sort the values in ascending order and can be calculated as following:

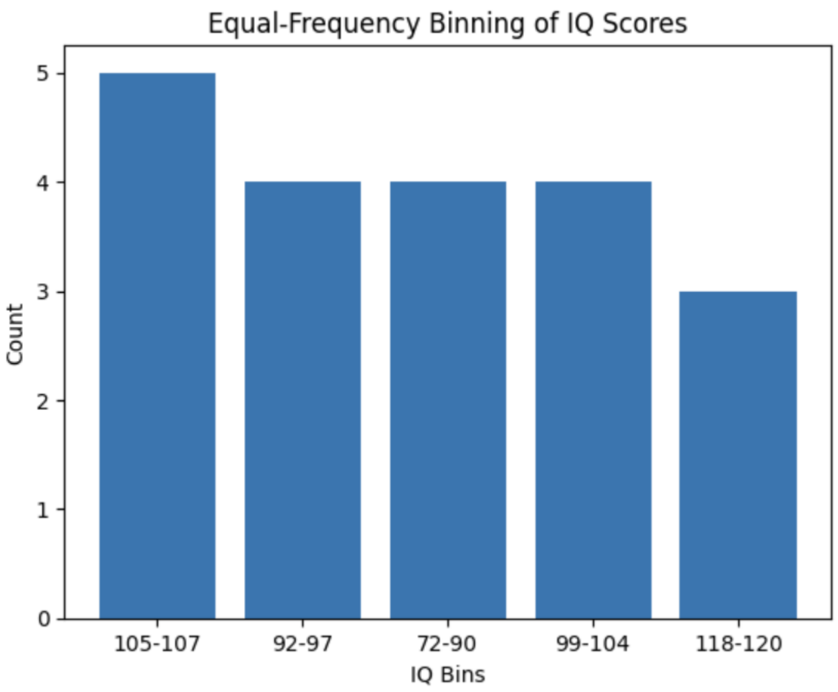
Width = Total Values/number of bins

Hence, equal frequency binning of IQ can be defined as following:

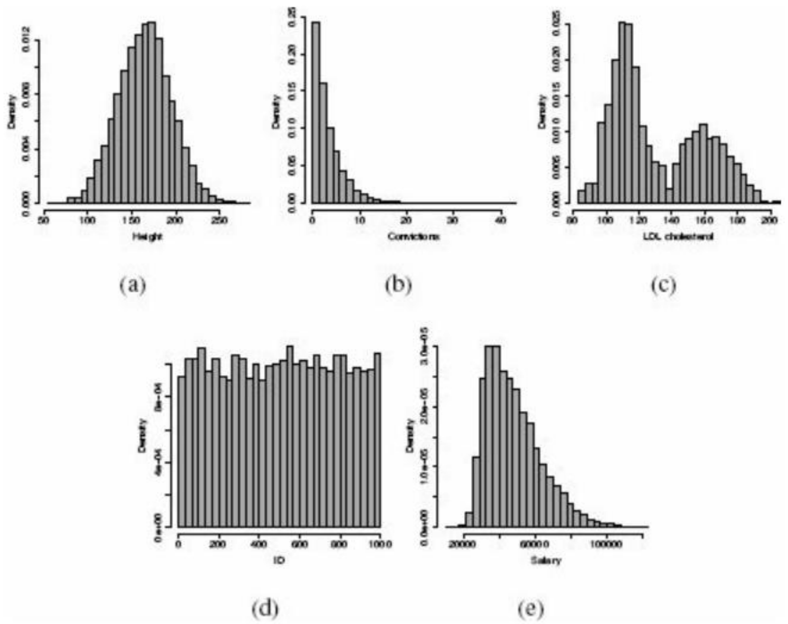
Sorted IQ = 72, 83, 88, 90, 92, 92, 93, 97, 99, 100, 101, 104, 105, 106, 106, 107, 107, 118, 119, 120

Bin No.	**Values**
1	72, 83, 88, 90
2	92, 92, 93, 97
3	99, 100, 101, 104
4	105, 106, 106, 107
5	107, 118, 119, 120

It can be visually represented as,



Q4. Comment on the distributions of the features shown in each of the following histograms?



a) The height of employees in a truck driving company?

Ans: The height of the employees demonstrates following qualities:

1. It’s a normally distributed histogram.
2. It’s a bell-shaped curve with maximum central tendency at 150-175 and is further symmetrically distributed.

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

Ans: The number of prior criminal convictions held by people given prison demonstrates following qualities:

1. It’s an exponentially distributed histogram.
2. For exponentially distributed histograms the likelihood of low values occurring is very high, but it declines rapidly for higher values, hence likelihood of outliers is high.
3. Hence, we can say that prior criminal convictions held by people are high at the beginning but later they tend to diminish.

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

Ans: The LDL cholesterol values for large group of patients demonstrates following qualities:

1. It’s a multimodal distributed histogram.
2. We can observe two groups of patients clearly smokers and non-smokers, it’s quite significant that smokers have peak LDL cholesterol values of about 0.025 whereas non-smokers tend to have lower LDL cholesterol values of about 0.015

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

Ans: The employee ID of university staff demonstrates following qualities:

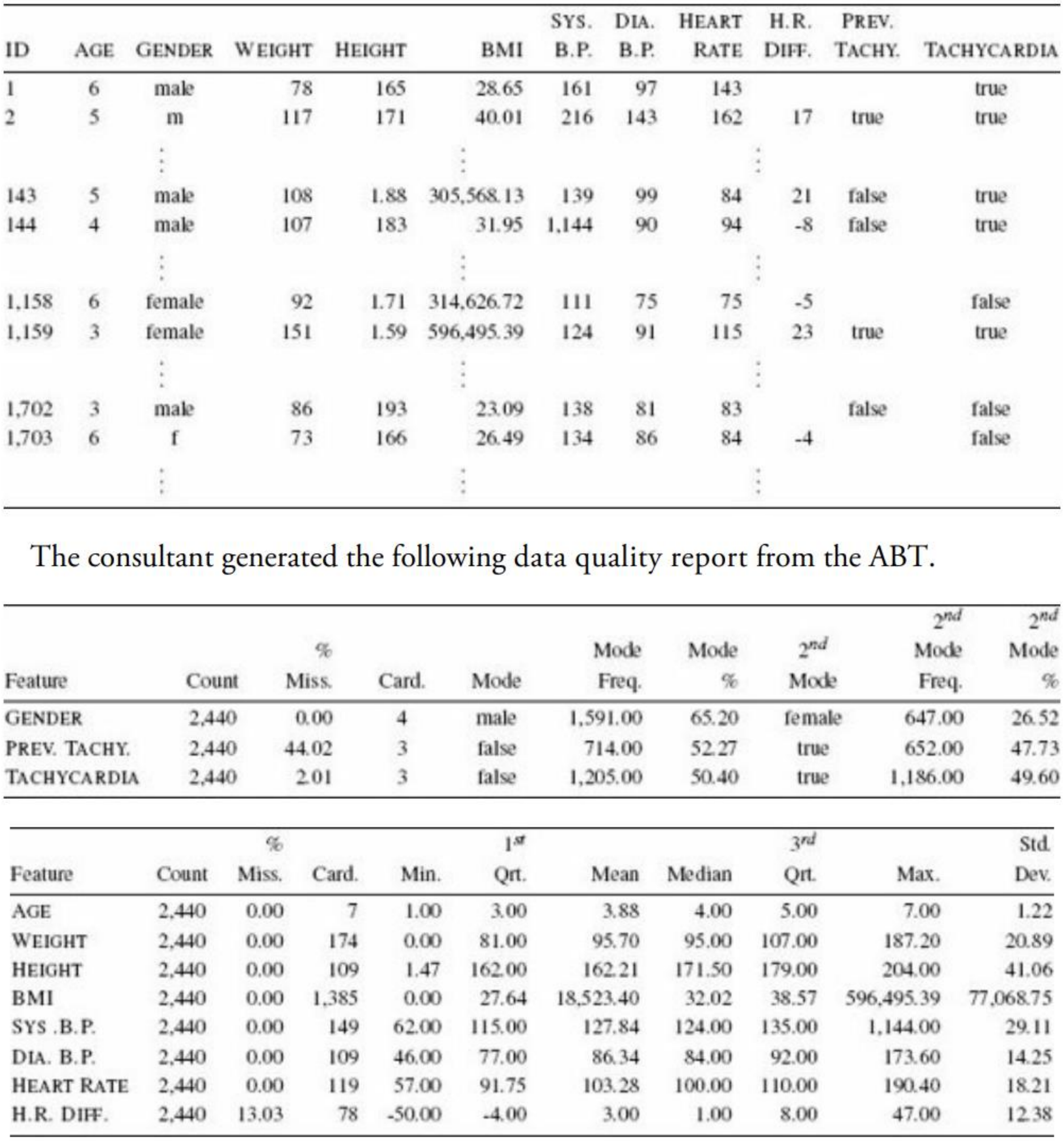
1. It’s a uniformly distributed histogram.
2. We can see that employee ID are uniformly distributed in all range values.

e) The salaries of motor insurance policy holders?

Ans: The salaries of motor insurance policy holders demonstrate following qualities:

1. It’s a Unimodal right skewed histogram.
2. It’s evident that a largen number of holders have a salary between 30K to 50K and only a few holders have salary greater than 75k.

Q5. The following table contains an extract from this ABT—the full ABT contains 2,440 instances.

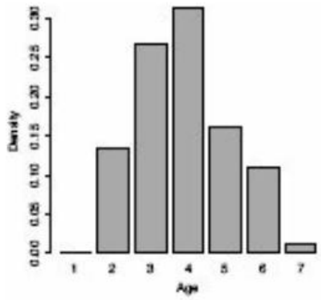


ID	AGE	GENDER	WEIGHT	HEIGHT	BMI	SYS. B.P.	DIA. B.P.	HEART RATE	H.R. DIFF.	PREV. TACHY.	TACHYCARDIA
1	6	male	78	165	28.65	161	97	143			true
2	5	m	117	171	40.01	216	143	162	17	true	true
		⋮			⋮			⋮			
143	5	male	108	1.88	305,568.13	139	99	84	21	false	true
144	4	male	107	183	31.95	1,144	90	94	-8	false	true
		⋮			⋮			⋮			
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
		⋮			⋮			⋮			

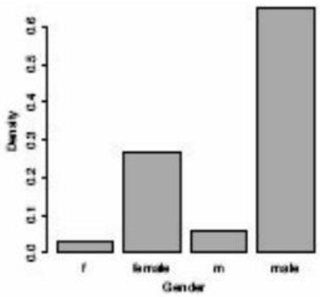
The consultant generated the following data quality report from the ABT.

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

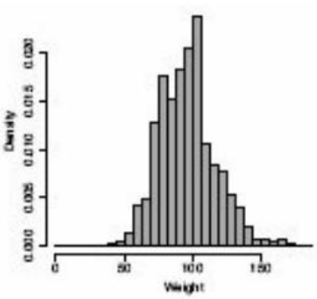
Feature	Count	% Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	Max.	Std. Dev.
AGE	2,440	0.00	7	1.00	3.00	3.88	4.00	5.00	7.00	1.22
WEIGHT	2,440	0.00	174	0.00	81.00	95.70	95.00	107.00	187.20	20.89
HEIGHT	2,440	0.00	109	1.47	162.00	162.21	171.50	179.00	204.00	41.06
BMI	2,440	0.00	1,385	0.00	27.64	18,523.40	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



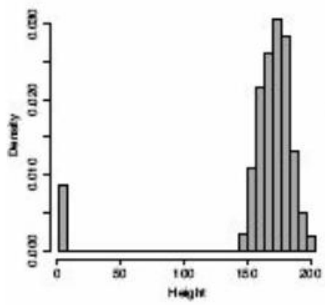
AGE



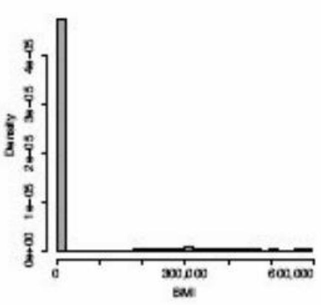
GENDER



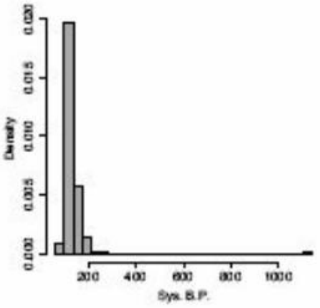
WEIGHT



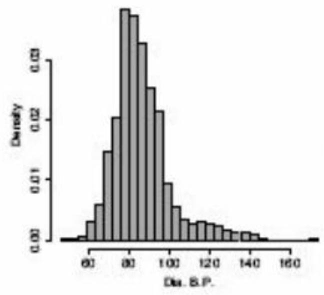
HEIGHT



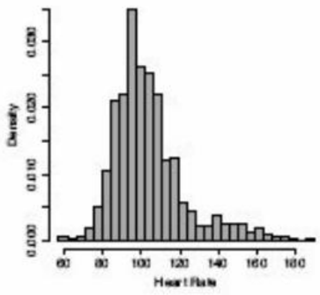
BMI



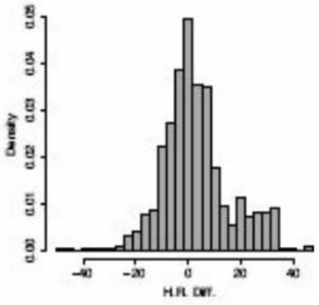
SYS. B.P.



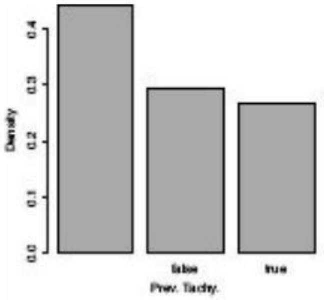
DIA. B.P.



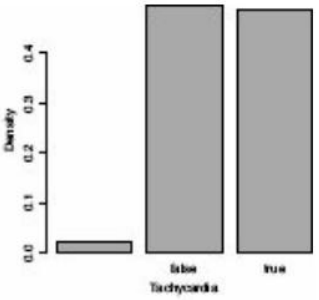
HEART RATE



H.R DIFF.



PREV. TACHY.



TACHYCARDIA

Discuss this data quality report in terms of the following?

a) **Missing values?**

Ans: The following conclusions on missing values can be made based on the report generated:

1. There's 44.02% of missing data of patients who previously suffered tachycardia.
2. Approximately 2.01% of missing data of patients who suffered tachycardia.
3. Also, there's almost 13.03% of missing data of H.R diff i.e., people whose heart rate was different in the current visit and their last visit to the clinic.

b) **Irregular cardinality?**

Ans: The following conclusions on irregular cardinality can be made based on the report generated:

1. The gender column has cardinality of 4 meaning along with male and female they have sometimes used m or f to represent gender instead, which is unnecessary and can be standardized.
2. The prev. tachy field is a Boolean field but shows the cardinality of 3 meaning it has some null or incorrect values, Boolean values must only contain 2 values and must be standardized.
3. Similarly, Tachycardia field is a Boolean field but shows the cardinality of 3 meaning it has some null or incorrect values, Boolean values must only contain 2 values and must be standardized.

c) **Outliers?**

Ans: The following conclusions on outliers can be made based on the report generated:

1. The minimum age is 1.00, which seems unusually low. It might be an entry error or a specific case.
2. The minimum weight is 0.00, which is impossible for an adult human. This could be a data entry error or missing data.
3. The minimum BMI is 0.00, which is also unrealistic. BMI should typically not be below 18.5, as it indicates underweight.
4. The maximum value of 596,495.39 is extremely high and likely an error.
5. The maximum value of 77.068.75 is also very high and likely an error.
6. The negative minimum value of -50.00 might indicate an issue or error in data collection. Heart rate difference is typically a positive value, representing changes between resting and active states.

d) **Feature distributions?**

Ans: The following conclusions on feature distributions can be made based on the report generated:

1. The distribution of ages appears to be slightly positively skewed. This suggests that most individuals in your dataset are in the higher age range, with relatively few younger individuals.
1. The distribution of weights appears to have a relatively small standard deviation. This suggests that the weights in your dataset are relatively consistent, with most individuals having similar weights.
2. The distribution of height suggests that there may be some variation in height within the dataset, but it's not extremely skewed in either direction.
3. The distribution of BMI values has a wide range. This indicates a considerable variation in BMI values within your dataset, with some individuals having very high or very low BMI values.
4. The distribution suggests that systolic blood pressure values are somewhat normally distributed, with most individuals falling around the mean.
5. Similarly, the distribution appears somewhat normal, indicating that diastolic blood pressure values are clustered around the mean.
6. The distribution of heart rate differences appears to have a negative minimum value of -50.00. A negative heart rate difference is unusual and might indicate measurement errors or inconsistencies in data collection.