Experiment No.1

Title: Understanding of the Data

Name: Chandana Ramesh Galgali

Batch: B-1 **Roll No.:16010422234**

Experiment No.:1

Aim: Understanding of the Data

Resources needed: Any RDBMS, EXCEL, Data storage tool

Theory:

In order to make data ready for data mining process, data exploration is essential step to develop a high-level understanding of the data. Data exploration includes in detail analysis of attributes and their data values and visualization. It aimed at identifying possible relationship between two or more variables/objects.
Broadly classifying, there are two types of attributes, numeric and categorical.

Categorical Attribute:

In categorical, each value represents some kind of category, code, or state. Categorical variables are either nominal or ordinal, depending on the extent of information the numerical coding provides.
The values of a nominal attribute are symbols or names of things. Nominal means "relating to names."
E.g. hair color and occupation are two attributes describing person objects.
Possible values for hair color are black, brown, blond, red, auburn, gray, and white. For occupation, possible values are teacher, dentist, programmer, farmer etc.

An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known. For example, grade attribute with values A+, A,A-, B, C; Student_progress attribute with values Good, average , poor. The central tendency of an ordinal attribute can be represented by its mode and its median (the middle value in an ordered sequence), but the mean cannot be defined.
Nominal, binary, and ordinal attributes are qualitative. That is, they describe a feature of an object without giving an actual size or quantity. The values of such qualitative attributes are typically words representing categories.

Numeric Attributes:

A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.

Interval-Scaled Attributes:
Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values.
For example, temperature, humidity attributes

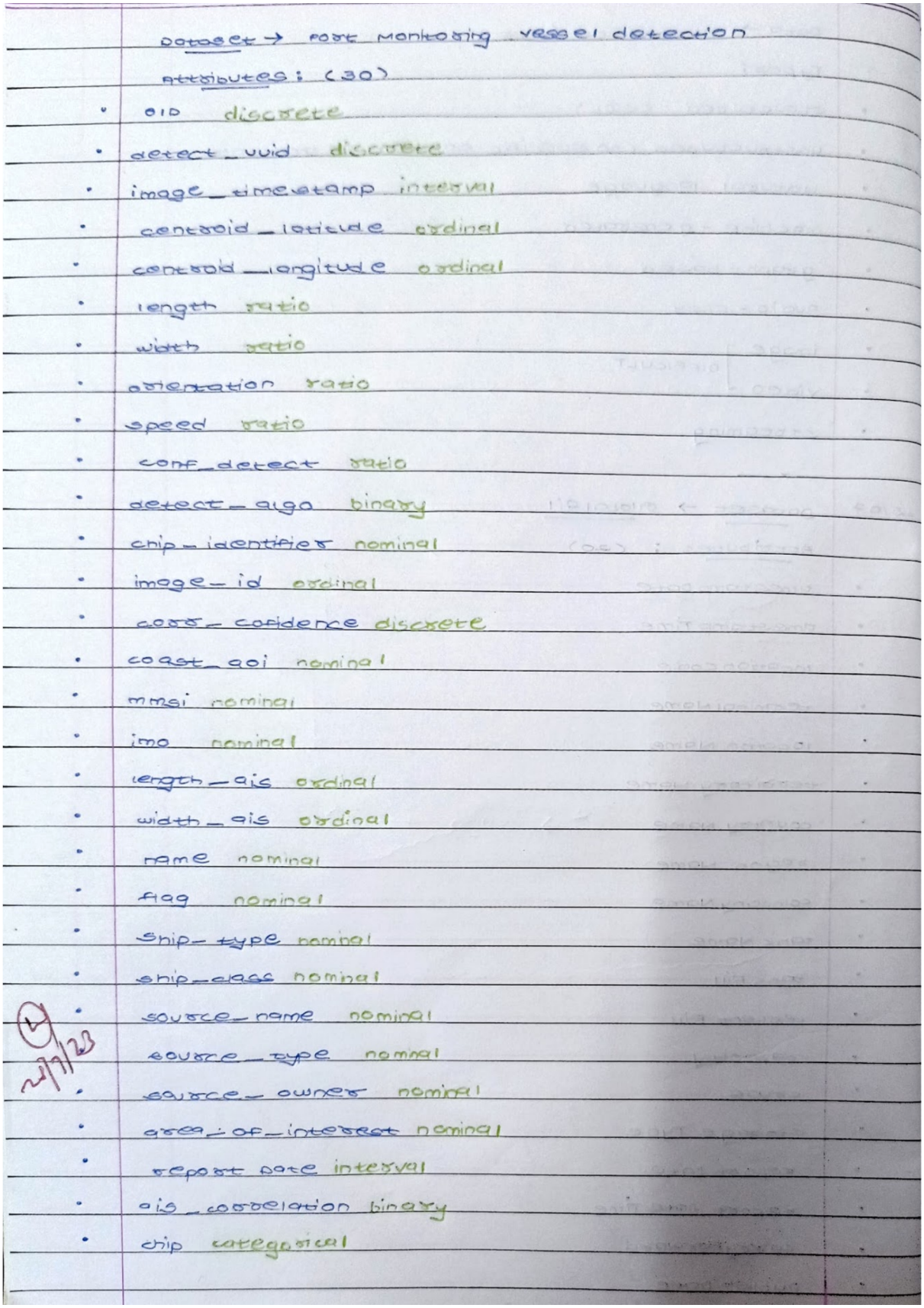
Ratio-Scaled Attributes:
A ratio-scaled attribute is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value. In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.
For example, *years of experience*

Procedure / Approach /Algorithm / Activity Diagram:

1. Download the large dataset for the purpose of exploration and ensure that dataset has variety of attributes; number of attributes must be at least 25.
2. Identify the category of each attribute from the dataset which you have created.
3. Identify the attributes which can provide any kind of useful information either collectively or as an individual. Also, discuss the about the information provided by the attribute and how it will be computed?

	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	U	X	Y	Z	AA	AB	AC
1	OID_	detect_uu	image_tin	centroid_	centroid_	length	width	orientatio	conf_dete	detect_al	chip_iden	image_id	corr_confi	coast_aoi	mmsi	flag	source_na	source_ty	source_ov	area_of_ir	report_da	ais_corr
2	0	a26211aa-	19:32.0	35.20868	129.5547	14.9992	14.7442	36.29461	0.573245	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
3	1	cb1b9c17-	19:32.0	35.29412	129.5691	66.4337	8.69964	8.648654	0.680874	CFAR	{"dataset"	02210e31-	8	Ulsan	6.36E+08	LR	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
4	2	6720aabf-	19:32.0	35.1622	129.5135	15.5012	5.21978	8.648654	0.505599	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
5	3	89e120da-	19:32.0	35.48483	129.4399	5.21978	4.42891	98.64865	0.313818	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
6	4	29044658-	19:32.0	35.43646	129.4302	171.759	80.9192	137.442	1	CFAR	{"dataset"	02210e31-	82	Ulsan	4.77E+08	HK	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
7	5	7c10dafb-	19:32.0	35.48138	129.4385	55.6052	12.7416	147.1743	0.498037	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
8	6	513691f1-	19:32.0	35.39159	129.4216	94.5807	15.7909	146.7261	0.624826	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
9	7	fe208a41-	19:32.0	35.46775	129.4358	50.9325	19.1392	8.648654	0.866154	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
10	8	a837701c-	19:32.0	35.40315	129.4238	126.514	16.3243	9.606396	0.770171	CFAR	{"dataset"	02210e31-	78	Ulsan	4.4E+08	KR	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
11	9	ae839aca-	19:32.0	35.6288	129.4658	8.85782	5.21978	8.648654	0.401447	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
12	10	39e7bad1-	19:32.0	35.44948	129.431	132.477	32.8091	152.8858	0.989712	CFAR	{"dataset"	02210e31-	93	Ulsan	5.64E+08	SG	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
13	11	424ef500-	19:32.0	35.42909	129.4259	166.868	49.5184	155.9134	1	CFAR	{"dataset"	02210e31-	82	Ulsan	3.71E+08	PA	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
14	12	0dacdbcd-	19:32.0	35.52422	129.444	28.1107	12.4107	105.2486	0.669726	CFAR	{"dataset"	02210e31-	38	Ulsan	2.16E+08	MT	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
15	13	698217ba-	19:32.0	35.15993	129.5129	6.95971	6.64337	98.64865	0.406516	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
16	14	97c8a005-	19:32.0	35.38388	129.4179	48.718	20.8791	8.648654	0.771753	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
17	15	c8fec580-	19:32.0	35.52464	129.4429	4.42891	3.47986	8.648654	0.154966	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
18	16	4f73a3cf-4	19:32.0	35.46793	129.4308	143.111	24.2261	173.0402	0.846774	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
19	17	efd44733-	19:32.0	35.52784	129.4413	88.0362	13.4191	89.60916	0.696392	CFAR	{"dataset"	02210e31-	13	Ulsan	2.16E+08	MT	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
20	18	f0e773c9-9	19:32.0	35.47379	129.4304	142.27	79.2271	3.92067	0.760883	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
21	19	f96b03f1-9	19:32.0	35.46216	129.4277	119.272	44.0264	172.7033	0.880589	CFAR	{"dataset"	02210e31-	89	Ulsan	4.4E+08	KR	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
22	20	eddd6e2b	19:32.0	35.4468	129.4248	19.9301	10.4396	8.648654	0.553998	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
23	21	f4ff053a-f	19:32.0	35.45404	129.425	157.951	31.7525	173.9723	1	CFAR	{"dataset"	02210e31-	97	Ulsan	4.17E+08	TW	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
24	22	cf589dbd-	19:32.0	35.47659	129.4291	8.85782	8.69964	8.648654	0.44145	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
25	23	e25ba399-	19:32.0	35.50714	129.4347	11.0723	6.95971	8.648654	0.508618	CFAR	{"dataset"	02210e31-	97	Ulsan	4.42E+08	KR	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
26	24	ef9db90c-	19:32.0	35.15685	129.5107	68.6481	5.21978	8.648654	0.735815	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
27	25	5f726e2d-	19:32.0	35.45994	129.4256	42.0747	15.6594	8.648654	0.703191	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
28	26	25d30da3-	19:32.0	35.22742	129.382	113	73.49	9.865176	0.980344	CFAR	{"dataset"	02210e31-	87	Ulsan	3.13E+08	BZ	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
29	27	b05d5bf5-	19:32.0	35.41529	129.4167	13.2867	10.4396	8.648654	0.450383	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE
30	28	8f7272a5-3	19:32.0	35.43411	129.4198	106.574	52.6954	128.5381	1	CFAR	{"dataset"	02210e31-	93	Ulsan	4.42E+08	KR	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
31	29	836f9815-(	19:32.0	35.50633	129.4335	6.64337	5.21978	8.648654	0.315381	CFAR	{"dataset"	02210e31-	88	Ulsan	4.4E+08	KR	Spire_AIS	AIS	Spire	Ulsan	3/25/2021	TRUE
32	30	194c8ad6-	19:32.0	35.63007	129.4551	48.2438	22.5362	33.88916	0.691954	CFAR	{"dataset"	02210e31-	2e65-4c14	Ulsan						Ulsan	3/25/2021	FALSE

Results: (Program printout with output / Document printout as per the format)



Page No.. YOUVA
Date:

1. Nominal data } qualitative → categorical data → assigning nos. to qualitative characteristics or groups
2. ordinal data }
3. ratio } quantitative → numerical data → naturally measured as numbers
4. Interval }
5. Binary → whose unit can take only 2 possible states
6. continuous (real) • can be of any value • can change • Almost infinite • broken into fractions, decim
7. Discrete Data • can take on certain values • fixed • only finite values • whole nos. / int

- Nominal data
  - categorical
  - no ranking or natural order / all have same value
  - the most basic level of measurement
  - e.g. gender, ethnicity, eye color, blood type; brand of motor vehicle / television / refrigerator; preferences, favorite meal
- ordinal data
  - categorical
  - ordering or ranking difference between the options
  - e.g. income level, level of agreement, political orientation
- Interval data
  - numerical data
  - naturally quantitative
  - doesn't have a meaningful zero point – zero is arbitrary
  - has an order, spaces betⁿ measurement points are equal
  - e.g. credit scores, GMAT scores, IQ scores, temperature in °F
- Ratio data → most sophisticated level of measurement
  - ordered / ranked
  - zero point reflects an absolute zero
  - numerical distance betⁿ points is consistent & can be measured
  - can meaningfully × & ÷

Questions:

1. Compare Discrete and Continuous Attributes. Give at least 5 examples of each.

Ans: Discrete and continuous attributes are two types of variables used in statistics and data analysis. Here are the definitions and examples of each:

Discrete Attributes:

Gender: This attribute can take on only a limited number of values, such as "male" or "female."

Marital Status: This attribute can have discrete values like "single," "married," or "divorced."

Eye Color: This attribute can have values like "blue," "brown," "green," etc.

Number of Children: This attribute represents the count of children a person has, which can only be a whole number.

Education Level: This attribute can have discrete values like "high school," "bachelor's degree," "master's degree," etc.

Continuous Attributes:

Height: This attribute can take on any value within a range, such as 5.6 feet, 6.2 feet, etc.

Weight: This attribute can have any value within a range, such as 150 lbs, 180 lbs, etc.

Temperature: This attribute can have any value within a range, such as 25.5°C, 30.2°C, etc.

Income: This attribute represents a person's income and can have any value within a range, such as $50,000, $75,000, etc.

Time: This attribute can have any value within a range, such as 10:30 AM, 2:45 PM, etc.

Discrete attributes have a limited number of distinct values, while continuous attributes can take on any value within a range.

2. What are the different jobs associated with Data Analysis? What are the qualities and features needed to do these jobs in an individual?

Ans: There are various job roles associated with data analysis, each with its own specific responsibilities and requirements. Here are some common data analysis job roles along with the qualities and features needed to excel in these positions:

Data Analyst:

Qualities: Strong analytical skills, attention to detail, problem-solving abilities, and proficiency in statistical analysis.

Features: Proficiency in data manipulation and analysis tools (e.g., SQL, Excel, Python, R), ability to interpret and present data insights effectively, and a solid understanding of data visualization techniques.

Business Analyst:

Qualities: Strong business acumen, excellent communication skills, ability to understand and translate business requirements into data-driven solutions.

Features: Proficiency in data analysis tools, knowledge of business intelligence platforms, familiarity with data modeling and database concepts, and the ability to collaborate with stakeholders from different departments.

Data Scientist:

Qualities: Strong mathematical and statistical skills, expertise in machine learning algorithms, programming proficiency, and the ability to think critically and creatively.

Features: Proficiency in programming languages (e.g., Python, R), knowledge of advanced statistical techniques, experience with big data technologies, and the ability to develop predictive models and algorithms.

Data Engineer:

Qualities: Strong programming skills, knowledge of database systems, data architecture expertise, and problem-solving abilities.

Features: Proficiency in programming languages (e.g., Python, SQL), experience with database technologies (e.g., SQL, NoSQL), familiarity with data integration and ETL (Extract, Transform, Load) processes, and the ability to design and optimize data pipelines.
Data Visualization Specialist:
Qualities: Creativity, attention to detail, strong design skills, and the ability to transform complex data into visually appealing and understandable visualizations.
Features: Proficiency in data visualization tools (e.g., Tableau, Power BI), knowledge of design principles and best practices, understanding of user experience (UX) design, and the ability to tell compelling stories through data visualizations.
In addition to these specific qualities and features, individuals in data analysis roles should also possess a curious and inquisitive mindset, a passion for continuous learning, and the ability to work effectively in a team environment. Strong communication and presentation skills are also crucial for effectively conveying insights and recommendations derived from data analysis.

Outcomes:

Understanding of the Data

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)
The experiment provided a solid foundation in working with categorical and numerical attributes. We gained practical skills and knowledge that can be applied in various data analysis tasks, such as exploratory data analysis. Understanding the characteristics and nuances of different attribute types is crucial for extracting meaningful insights and making informed decisions in data science and analytics.

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition