1)

Trans 1	{Milk, Bread, Jam}
Trans 2	{Milk, Cookie}.
⋮	⋮

Transaction data

	Milk	Bread	Jam	Cooki
Trans1	1	1	1	0
Trans2	1	0	0	1
⋮				

Transaction matrix.

Advantage: Since string format is converted into numerical format, calculations become easier & also various mathematical formulas or operations can be applied on it.

Disadvantage: The matrix becomes sparse many times. i.e there are much more 0's than 1's.

eg: a grocery store ~~with~~ which sells 200 different products & a customer buys only 2 from 200, i.e 198 col vals will be 0 or useless.

Asymetric features ~~means where the~~ data points ~~with~~ one value ~~are regarded more~~ impor

In Asymetric features, there is not equal importance given to different values, like in this example, if the attribute has value 1, it is useful & when the attribute for an object is 0, it ~~can not be used~~ or is meaningless.

Discrete because it can take only specific values & not from a range of values. Also, in this case it would be Binary as well

Thats why this is an example of discrete asymmetric features or attributes

2) (a) Brightness by light meter. eg: { 0, 1.1, 100, ... }

- continous, because the value can be anything between a range. eg: 30.1, 300.226, etc (real values)

- ratio, because it has a meaningful zero point/value. & there is no hierarchy & two readings can have same value. also ratio = a x old(value) is true here.

(b) Brightness by people's judgement. eg: { very low, high, low }

- discrete, bec there are only selective values. here different people can have different names for the same brightness like one can say low & other "some other word" but they mean the same thing

- Ordinal, bec two readings can be same & also there is a certain order.

(c) No of customers in a grocery store. eg: { 0, 1, 2, ... }.

- discrete, bec no of customers can be only integers & not 1.5 or 2.4 or must be countable

- Ratio, bec no of customers can not be negative. hence theres a true zero point. & saying that Today there are twice as customers than yesterday makes sense hence follows. new value = x x old value, ratio operations

(d) Grades: A, B, C, D, E, F.

these have order and also are categorical.
Hence these are discrete & ordinal.

(e) distance ~~between~~ from Monroe County Courthouse.
Distance will be continuous because
it can be any number between 0 to x miles.
depending how accurately it has been measured.
Also, 0 is meaningful while calculating/measuring distance
hence this quantity would be Ratio., ratio operations are allowed.

3) - Noise is a datapoint with incorrect values or random values, due to human error while recording or due to other errors of the sensor/machienary which is used to record data.

= Outlier is a point which deviats significantly from the other data points. This again can be created from Equipment error or Data entry error.

- Noise can be interesting or desirable depending on the problem statement. eg: finding or detecting or analyzing noise in data can help determine errors in data recording OR noise in data which is used to train a model. can result in lower accuracy because the model keeps on learning from the incorrect (noise) data points. Also, if the data instances are huge with very less noise then the accuracy wouldn't get affected or reduced much. Conversely when training a model with considerable instances of incorrect readings than actual correct (useful) data will degrade the performance on real-world data

= Outliers can be useful or not again depending on the case. Considering an classification problem with 1000 instances for class A & 10 instances for A' class. It may seem that the later 10 points are outliers, but they are valuable in this case. Also in some cases outliers can change/shift the means, hence they must be eleminated.

- Yes, noise objects can be outliers, eg: true value is. 6' 12" for height but incorrectly recorded as 61' 2" by human error.

- Noise objects are not always outliers because the error is not predictable it can be either too much, too low or unnoticable. eg: min & max weight is 50 to 120 kgs for age 22 to 25, then error could be either of value ± 10 or ± 50, true value assumption = 60, error -10, reading = 50 (does not seem outlier). true value = 100, error +40, reading 140 (seems like an outlier). Again here, we don't know for sure if an outlier is an outlier because it could be the actual value recorded that is a rare case

- Outliers can be. true but rare data points but noise is something that holds incorrect values, hence outliers are not always noise.

- Yes noise can make an actual value into unusual one or (outlier) as the error changes, the value can be either ~~an~~ seems like normal or like unusual. Also, an unusual value recorded with ~~an~~ errors can make it seem like normal.

5) (I) The score that is allotted by Google Trends (Interest over Time) is directly proportional to the searches related to bitcoin & inversely proportional to the total number of searches. in a region. Hence this does not certainly or directly depend on the population of region

(II) When changing the time frame of plot, the value for every time/day changes because it is scaled based on the maximum no of searches for BTC. Hence we can not directly compare scores from plots associated within different time frames, since there will be different highest no of searches or the value which will be considered as 100. like, popularity of BTC in State Indiana on 1 Dec 2018 will be different for 2 different time frames. viewed eg: Year ~~200~~ 2017 to 2019 and for 2018 to 2020. Bec highest value is projected as 100 & all other values are calculated using this 100 as reference.

(III) Also, this helps us analyze ~~how~~ the interest of state as well as cities in a more general way or based on proportion or distribution. I guess, multiplying the population ~~out~~ factor with the interest score can give us the total no of queries with bitcoin keyword for that sub-region

(IV) The score here states the percent of queries related to bitcoin from the total no of searches. This implies the ~~popularity of bitcoin~~ will reduce either if other topics are ~~more searched than it or~~.

(V) We will have data, but while visualizing we must take care to visualize it accurately & as fully as possible. Also how the data is visualized or what transformations are performed must be clearly stated else this may lead to incorrect inference. ~~of~~ from the analysis.

6) Notion of fairness: proportional group representation in all clusters.

I think, this means that theres some kind of similarity in the proportions of one or ~~more~~ multiple classes in every single cluster. eg: assume two classes & no of clusters formed be 2 then cluster A will have 90% of class A instances & 10% of B & cluster B will have 90% of class B & 10% of A instances or something similar to this!

Problem is to reduce under representation or over representation, which results in different proportions of ~~each class~~ same class in different clusters. under representation is the very small proportion of instances that should not be there in a cluster. Solution is to use group fair clustering which will impose demographic representation or ~~pop~~ proportionality within classes in all clusters. which will cause clusters to deform & spread out. on to the other solution, where instead of clusters, fairness is ensured for labels which helps reducing the deformation of clusters.

In the example of hiring candidates, classes: hire, short-list, scrutinize further, reject. here the author states that demographic parity is much more important here, ~~which means~~ in labels rather than clusters. i.e it means that consider only. hire & reject class, if company hires an unworthy employee, it will cost them company time & resources to train that employee, so its company's loss. If the company does not hire / reject an worthy employee again its their loss. So, to reduce the loss- more % worthy candidates must be in cluster with label Hired & ~~more~~ % very low % unworthy. Similarly for the ~~see~~ Rejected case this improves accuracy. by reducing the loss I guess.

Fairness here also means probability of sucess for each class protected or non-protected must be same. i.e no bias in entire population

To ensure fairness, two settings are used:
LCAL: center labels are pre-assigned
LCUL: labels are to be choosed.
~~Also, here the author says the~~ Since the centers are fixed for each ~~problem~~ cluster, it can be considered as a routing problem where each point is directed to the center of cluster. By this cost can be calculated which is ~~the goa~~ one of the goal of the paper (to reduce cost)

Using LCAL, the authors are able to assign classes in much less time than traditional approaches (polynomial) & for binary classification with a linear time complexity which is even better. Hence, fairness can be guranteed in less time cost & this algo can also be used for large datasets.

The cost is measured as POF which is calcutated from the cost of fair approach & unfair approach where lower cost is better.

Datasets like Credit Card, Adult is used to calculate POF for ~~Labeled~~ Fair clustering approach, fairness (labeled) approach. where The Fair Labeled Clustering has the lowest cost.

Also, the cost or time, to compute classes for larger datasets like Census ~~is not much compared to~~ still scalability ~~remains an issue here~~ The time taken by the Fair Labeled Clustering is relatively very less than Fair Custering.

I think, talking from the scalability point of view the research is pretty amazing. I believe optimal solutions & algorithms ~~are~~ will be for more important now because the data collection & generation speeds have increased and so are the model complexities which can feed large data & provide better results