

1. Transaction data can also be represented as a transaction-item matrix or document-term matrix. They go by the name Sparse Data Matrix. A sparse data matrix only takes into account non-zero values and contains data of the same type.

Advantage: Since there are more zeros in a sparse data matrix than non-zero values, we may take advantage of this and focus more on preserving space for non-zero values. For instance, the zeros cancel out a lot of data when a mathematical operation like multiplication is involved among matrices, making the compilation process simpler.

Disadvantage: The biggest drawback is that it has a substantial amount of zeroes that might not be helpful and provides very little information that is valuable (non-zero values). Take a sparse matrix, for instance, whose dimensions are equal to the total number of words in a novel. In this instance, only one position in the matrix will be 1 for a single word, with the remaining locations being 0. This demonstrates how, when looking at this concept as a whole, space and time have a very evident disadvantage.

Transaction-item matrix is an example of a data set that has asymmetric discrete features.

Discrete: Because the elements of transaction-item matrices do not fall in a fixed order and there are clear spaces between the values, they have discrete characteristics. Element spacing will vary between transactions because each transaction has a unique set of objects. For instance, different people bring varying numbers of stuff.

Asymmetric: Transaction-item matrices are asymmetric because the transposed version of the matrix is different from the original matrix. In this scenario, transposing the matrix will change both the overall matrix and the item list for each transaction. Consider a matrix where the rows represent transactions and the columns represent item lists. Here, if we transpose the matrix, the row 1 that originally represented transaction 1 (for all items) is transformed to the number of item A in all transactions, changing the matrix's original meaning. As a result, transaction-item matrices are asymmetric.

2. **(a) Brightness as measured by a light meter**

Continuous - Because it varies over time and can take on various values at various points in time

Quantitative and ratio: Because the data measured is a numeric value and temperature can have zero point.

- (b) Brightness as measured by people's judgments.**

Discrete: Because these values are finite and measurable

Qualitative and Ordinal: Because it can be measured with a name or symbol and variables can be in a specific order

- (c) Number of customers in a grocery store.**

Discrete: Because we can count the number of customers

Quantitative and Ratio: Because the lower number of customers is 0.

- (d) Letter grades (A, B, C, D and F).**

Discrete: Because we have a finite number of letter grades

Qualitative and Ordinal: because the grades are ranked according to the order mentioned above

(e) Distance from the Monroe County Courthouse.

Continuous: Because it can have different values with respect to a different point of reference.

Quantitative, interval/ratio: Interval because the distance between two points can be measured and the variables are constant. Ratio because of the existence of zero value when the distance is measured with respect to itself.

3. Noise is defined as inaccuracies in attribute values and incorrectly tagged examples. On the other hand, outliers include both errors and population-wide variations in values.

Is noise ever interesting or desirable? Outliers?

=> Because noise changes the attributes, it is undesirable. The noise in attributes renders them meaningless. As a result, it offers no insights into the data.

On the other hand, outliers are desirable since they assist in identifying variations in attributes that may be significant data elements that aid in problem-solving.

In conclusion, noise is indeed not interesting or desirable, whereas outliers are.

Can noise objects be outliers?

=> Yes. Outliers are just fluctuations in attribute values, and noise in attribute value distributions is unpredicted, unpredictable, and meaningless. So sometimes the noise in the values of the attributes can get identified as outliers.

Are noise objects always outliers?

=> No. because the noise in a value might appear to be true value with little or no fluctuation from genuine values. Since outliers differ from the true values of the attributes, this situation cannot be classified as an outlier.

Are outliers always noise objects?

=> No, because outliers may represent significant data characteristics that aid in the resolution of many data mining-related issues. So, not all outliers are noise objects.

Can noise make a typical value into an unusual one, or vice versa?

=> Yes, sometimes a true variation in attribute values can appear as noise because noise is unpredictable and erratic. Similar to how noise can resemble an expected value.

4. Summary:

Here, the data frame used is of TikTok popular songs. There are a total of 223 entries in it. Each song and its attributes are listed in rows. The data frame contains 18 properties. Track name, artist name, artist pop, etc. are a few of them. Nine of the attributes have float64 datatypes, six have int64 datatypes, and three have object datatypes. The object datatype is for attributes that contain strings for example track_name, artist_name, and album in our data frame. The overall memory usage of data is 31.5 KB. To understand more about each attribute distribution, they are plotted using bar plots, histograms, box plots, and bivariate plots. These give insight into the range/distribution of attribute values, type of value, finding outliers in them, etc. Loudness, track pop, duration ms, and other attributes with outliers were discovered using box plots.

Missing Data: There are no missing values in this data frame. In Jupyter Notebook, I have used 2 approaches to calculate missing values.

Approach 1: Determine how many rows there are in the data frame. After that, delete the rows that have missing values and count the number of rows once more. The number of rows with missing values is determined by the difference between these two values.

Approach 2: Calculate the number of attributes whose value is null and take a sum of it.

Correlation: Yes, there is a correlation between attributes. The value of correlation ranges from -1 to 1. If the correlation is negative, then the two attributes are inversely proportional. If the correlation is positive, then they are proportional. If the correlation is near zero, they are not related.

For example,

Inverse proportional(Pearson):

1. Speechless and track_pop: -0.203603

2. Acousticness and energy: -0.408780

Proportional(Pearson):

1. Valence and energy: 0.312315

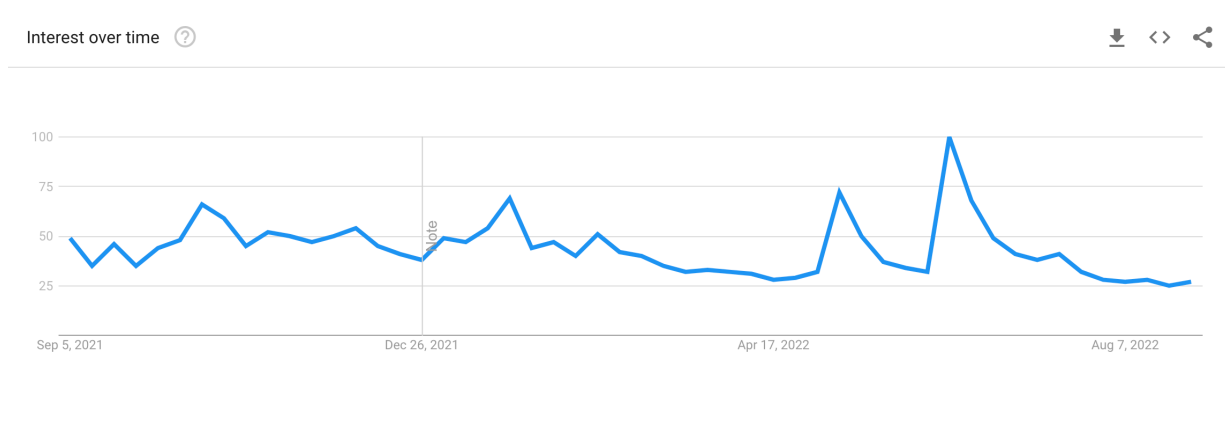
2. Loudness and energy: 0.681296

5. Options Selected: Bitcoin interest in the United States in the past 12 months:

Highlight 1: Interest Over Time

An illustration of interest over time in Google Trends is a line graph. The Y-axis depicts time, and the X-axis reflects the interest in relation to the chosen place and time that has the greatest number of search interests. Here, we can observe that the period between June 12 and June 18, 2022, in the United States, has the most search interest. On the other hand, from August 28 to September 3, 2022, there is only a 25 percent search interest. It is clear from the graph that interest in bitcoin searches is consistently higher than 25.

In my opinion, we can use the platform to gather real-time data because Google Trends' data is updated hourly. We can convert the line graph into a data frame for data mining tasks.

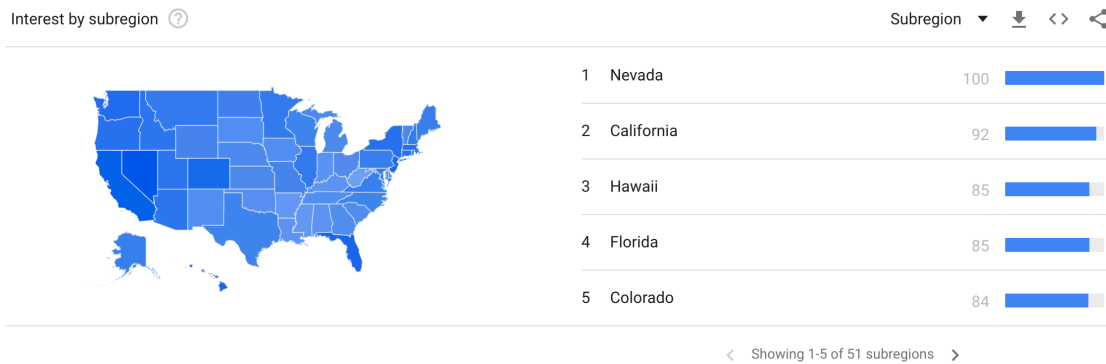


Interest Over Time

Highlight 2: Interest by subregion

Here, search interest on the map of the United States is depicted using a gradient plot. The level of blue color intensity corresponds to the level of search interest in that area, and vice versa. We can observe from the map that the majority of the US's western areas have darker blue hues, while the majority of its eastern regions have lighter hues. Nevada has the biggest search interest in this case, followed by California, Hawaii, etc. West Virginia, which has a search interest of 37, has the lowest search interest out of the 51 regions.

So, for finding the search interest from the map, we can extract the intensity of the blue hue and determine the intensity of the search in that area



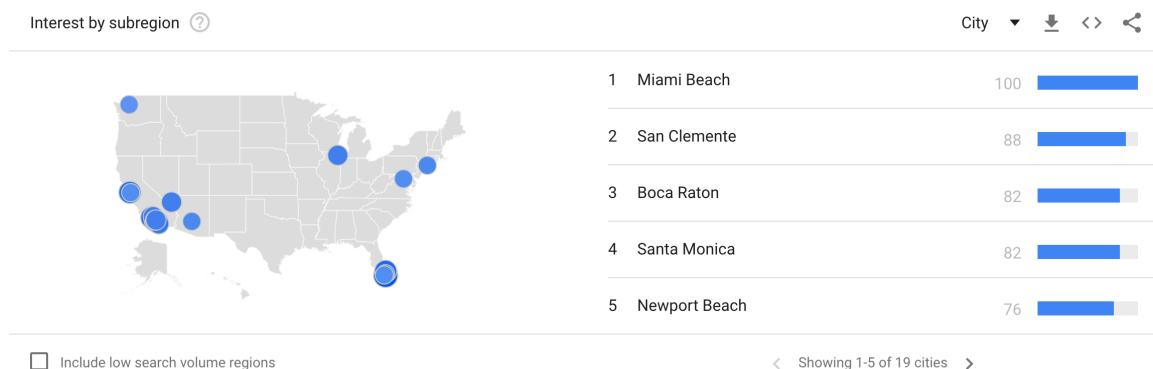
Interest by subregion

Highlight 3: Interest by subregion: City

Here, a scatter plot is used to represent the cities on the United States map. To fully comprehend the map, it is important to keep two things in mind: 1. Size of the circle 2. The intensity of the blue color.

The circle gets bigger and the blue hue gets more intense as the search interest rises. The intensity of the blue hue and the size of the circle both diminish when the search interest is low.

Miami Beach is the city in this case with the most search interest, followed by San Clemente, Thousand Oaks, etc. Fishers, with a search intensity of 43, has the lowest search interest among the top 18 cities.



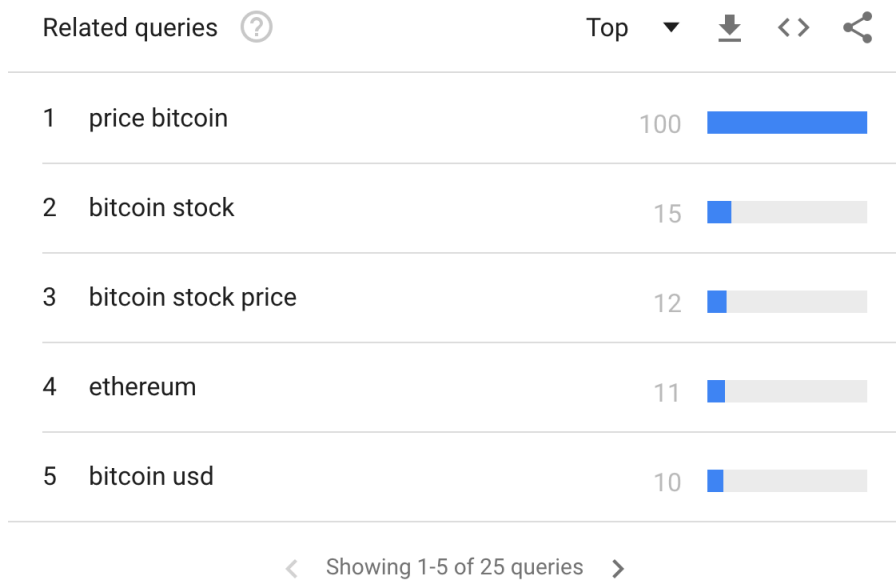
Interest by Subregion: City

Highlight 4: Related Queries: Top

Here, the topmost queries related to bitcoin are ranked with respect to the query that was searched the maximum which has a score of 100, and the other and ranked with respect to it. The queries are just keywords extracted from the complete question. The data is stored in a table format with 2 cols, one being the query and the other being their ranking.

In this case, the most popular query is “bitcoin price” followed by “bitcoin stock”, “bitcoin stock price”, etc. The most unpopular query among the 25 queries listed is “bitcoin value” which is at 2.

This is an amazing platform to extract data to analyze the most prominent queries that the majority of the population has related to bitcoin and find/suggest solutions for the same.



Related queries		Top			
1	price bitcoin	100	<div></div>		
2	bitcoin stock	15	<div></div>		
3	bitcoin stock price	12	<div></div>		
4	ethereum	11	<div></div>		
5	bitcoin usd	10	<div></div>		

< Showing 1-5 of 25 queries >

Related Queries: Top

- The authors of the paper “Fair Labeled Clustering” observe the widespread impact of AI on human beings, and how the concept of fairness becomes more relevant.

They propose a more 'fair' clustering algorithm that introduces fairness at a label level rather than the cluster level. They bring attention to the fact that fair representation at the cluster level could cause deformation of the cluster. There also exists the possibility of under-representation and over-representation in labels.

For the purpose of testing, two datasets are made use of: The Adult dataset and the CreditCard dataset, obtained from the UCI repository.

The Adult dataset consists of 32,561 data points and the 'race' attribute (5 possible values/ 5 colors) is set as the group membership attribute. Only the numeric attributes of the dataset are used as coordinates in space.

The Credit Card dataset consists of 30,000 data points, and the 'marriage' attribute (4 possible values/4 colors) is used to represent group membership. The 'age' and financial

attributes are as spatial coordinates.

The authors implement fairness considerations in two settings: Labeled Clustering with Assigned Labels (LCAL), where label centers are decided by their position in space, and Labeled Clustering with Unassigned Labels (LCUL), where conditionally, the label centers can be selected freely. They propose the calculation of the Price of Fairness (PoF) parameter. PoF is given by the ratio of the fair solution, which is the cost of the fair variant, and the color-blind solution cost, which is the unfair algorithm cost.

In the experiments conducted for the LCAL setting using the Adult dataset, the author's algorithm achieves fairness at a lower PoF when compared to the baseline clustering algorithm. The baseline clustering algorithm has a much larger proportional or color violation.

For the Credit Card dataset, similar behavior is observed, where the author's algorithm has a lower PoF than fair clustering.

The LCUL experiments on both the Adult and Credit Card datasets show the author's algorithm achieving similar performance. The algorithm is compared against two baseline algorithms: Nearest Center with Random Assignment (NCRA), and Fair Clustering (FC). The author's algorithm achieved lower PoF when compared to FC, but in line with NCRA. The author's algorithm however does have a few color violations when compared to FC, but the violation of the number of centers a label receives is less.

The perspective with which the authors approach the problem of fairness and how they could implement clustering is interesting. Their methodology was especially impressive as they measured not only computation time, but also engineered their own PoF metric for the measurement of performance, and took into consideration the scalability of their algorithm and the different parameters and constraints that would be subject to change. This opens up many more interesting approaches to socially responsible and ethical AI.