# Data Mining

Lecture-2

Data

Dr. Salem Othman

Summer 2023

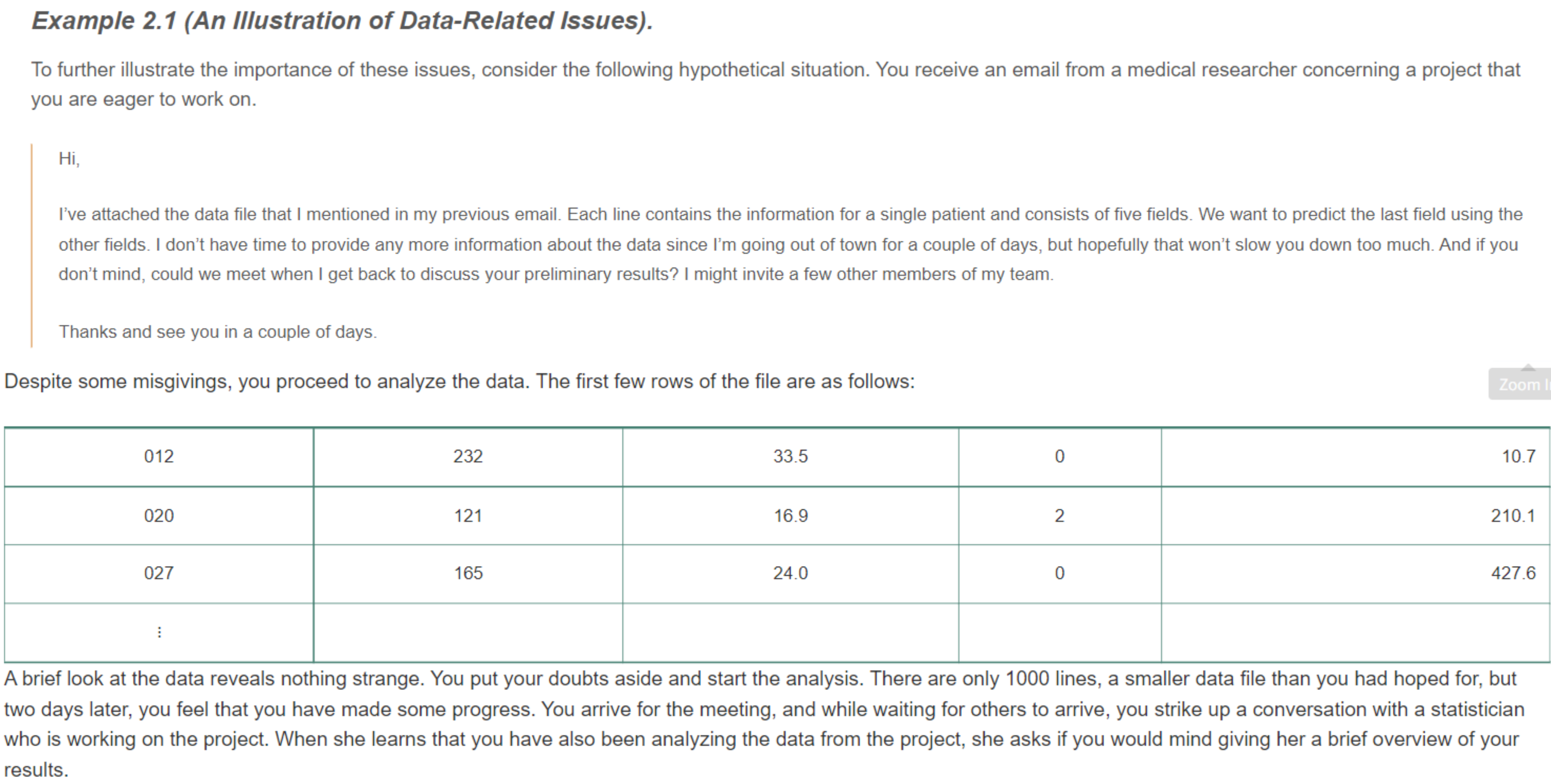
# Outline

- Attributes and Objects

- Types of Data

- Data Quality

# An Illustration of Data-Related Issues



***Example 2.1 (An Illustration of Data-Related Issues).***

To further illustrate the importance of these issues, consider the following hypothetical situation. You receive an email from a medical researcher concerning a project that you are eager to work on.

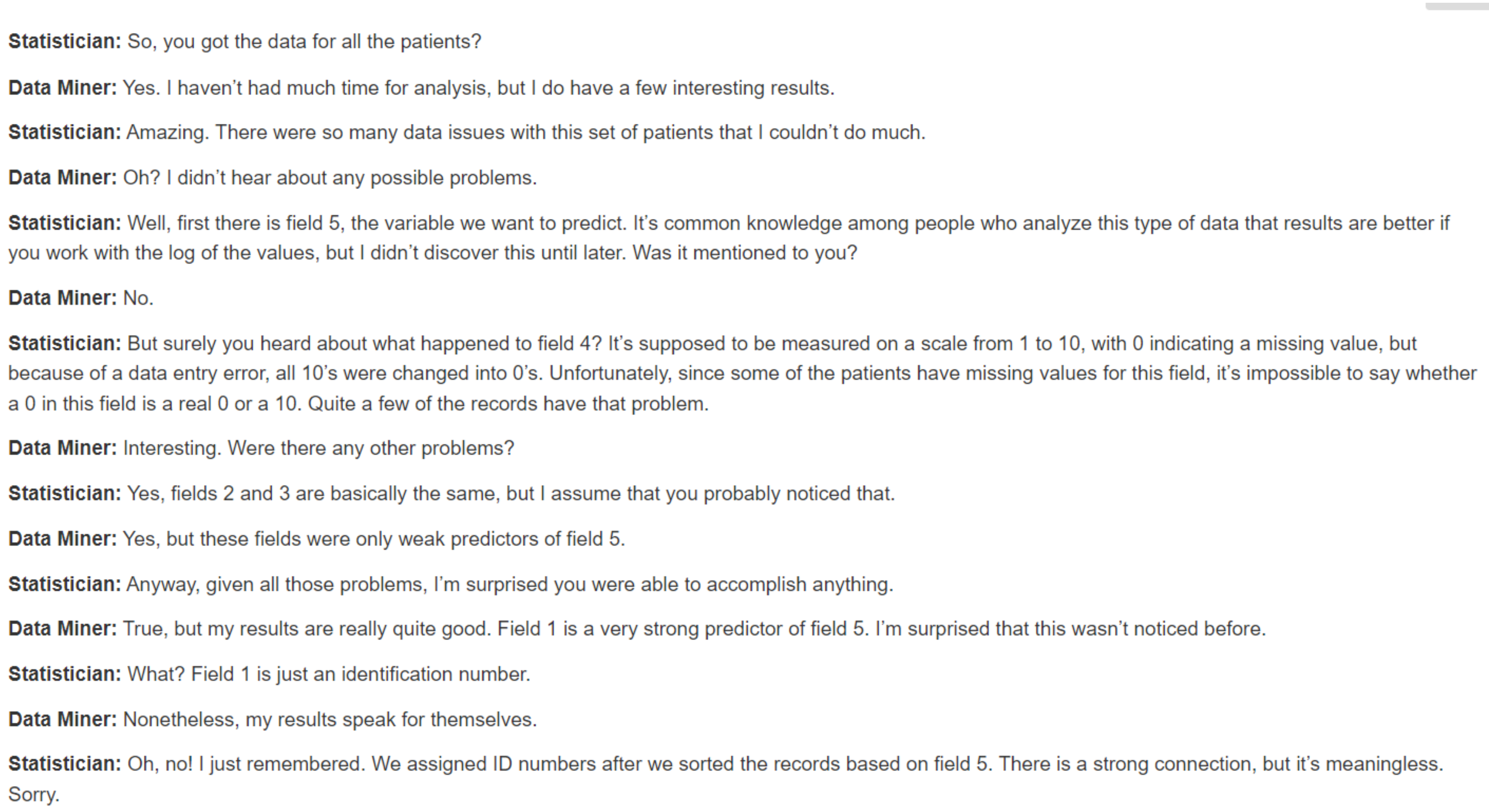
> Hi,
>
> I've attached the data file that I mentioned in my previous email. Each line contains the information for a single patient and consists of five fields. We want to predict the last field using the other fields. I don't have time to provide any more information about the data since I'm going out of town for a couple of days, but hopefully that won't slow you down too much. And if you don't mind, could we meet when I get back to discuss your preliminary results? I might invite a few other members of my team.
>
> Thanks and see you in a couple of days.

Despite some misgivings, you proceed to analyze the data. The first few rows of the file are as follows:

| | | | | |
|---|---|---|---|---|
| 012 | 232 | 33.5 | 0 | 10.7 |
| 020 | 121 | 16.9 | 2 | 210.1 |
| 027 | 165 | 24.0 | 0 | 427.6 |
| ⋮ | | | | |

A brief look at the data reveals nothing strange. You put your doubts aside and start the analysis. There are only 1000 lines, a smaller data file than you had hoped for, but two days later, you feel that you have made some progress. You arrive for the meeting, and while waiting for others to arrive, you strike up a conversation with a statistician who is working on the project. When she learns that you have also been analyzing the data from the project, she asks if you would mind giving her a brief overview of your results.

# An Illustration of Data-Related Issues



**Statistician:** So, you got the data for all the patients?

**Data Miner:** Yes. I haven't had much time for analysis, but I do have a few interesting results.

**Statistician:** Amazing. There were so many data issues with this set of patients that I couldn't do much.

**Data Miner:** Oh? I didn't hear about any possible problems.

**Statistician:** Well, first there is field 5, the variable we want to predict. It's common knowledge among people who analyze this type of data that results are better if you work with the log of the values, but I didn't discover this until later. Was it mentioned to you?

**Data Miner:** No.

**Statistician:** But surely you heard about what happened to field 4? It's supposed to be measured on a scale from 1 to 10, with 0 indicating a missing value, but because of a data entry error, all 10's were changed into 0's. Unfortunately, since some of the patients have missing values for this field, it's impossible to say whether a 0 in this field is a real 0 or a 10. Quite a few of the records have that problem.

**Data Miner:** Interesting. Were there any other problems?

**Statistician:** Yes, fields 2 and 3 are basically the same, but I assume that you probably noticed that.

**Data Miner:** Yes, but these fields were only weak predictors of field 5.

**Statistician:** Anyway, given all those problems, I'm surprised you were able to accomplish anything.

**Data Miner:** True, but my results are really quite good. Field 1 is a very strong predictor of field 5. I'm surprised that this wasn't noticed before.

**Statistician:** What? Field 1 is just an identification number.

**Data Miner:** Nonetheless, my results speak for themselves.

**Statistician:** Oh, no! I just remembered. We assigned ID numbers after we sorted the records based on field 5. There is a strong connection, but it's meaningless. Sorry.

# What is Data?

- Collection of ***data objects*** and their ***attributes***

- An ***attribute*** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature

- A collection of attributes describe an ***object***
  - Object is also known as record, point, case, sample, entity, or instance

**Attributes**

**Objects**

| *Tid* | **Refund** | **Marital Status** | **Taxable Income** | **Cheat** |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 95K | **Yes** |
| 6 | No | Married | 60K | **No** |
| 7 | Yes | Divorced | 220K | **No** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 75K | **No** |
| 10 | No | Single | 90K | **Yes** |

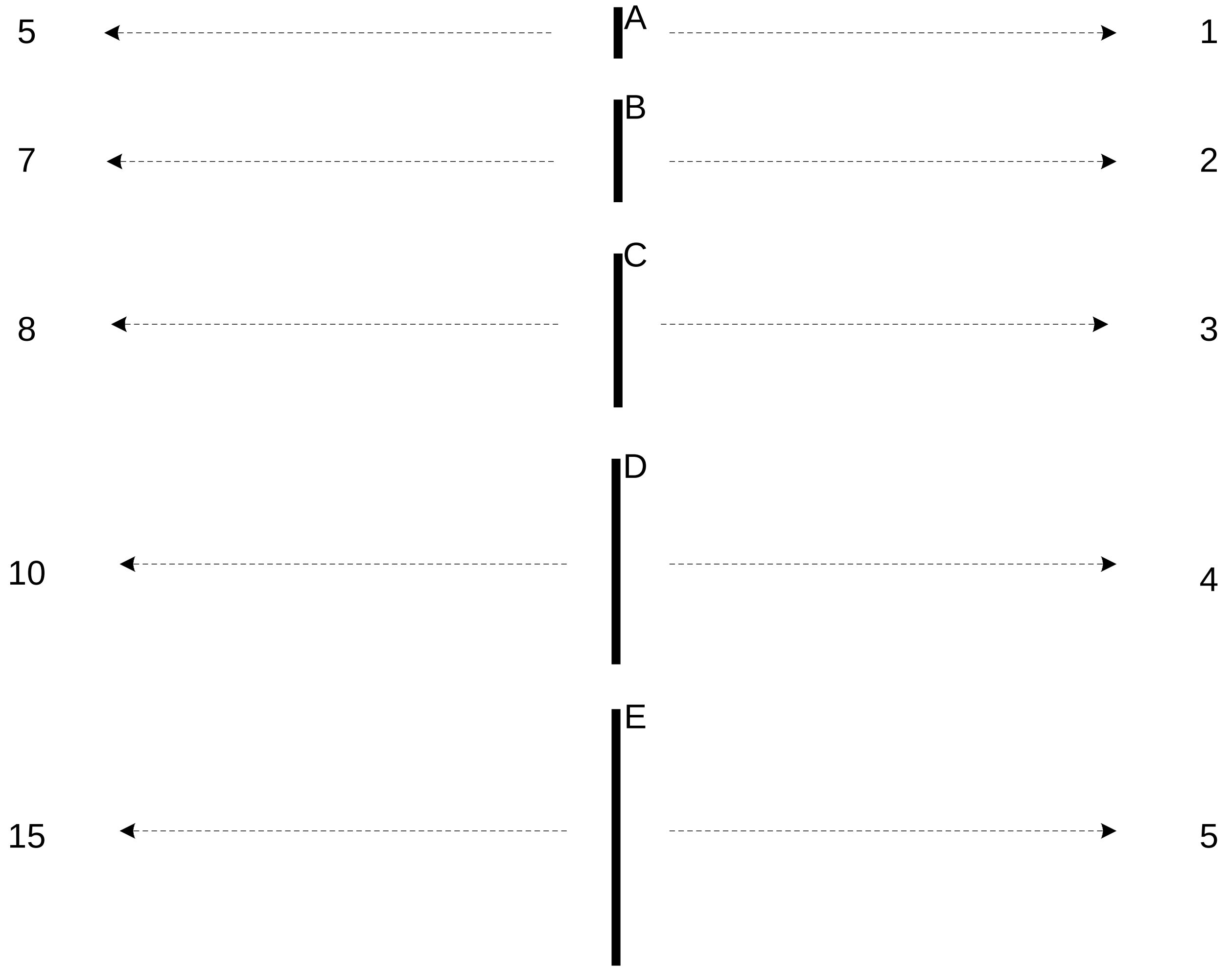
# Attribute Values

- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
  - But properties of attribute can be different than the properties of the values used to represent the attribute

# Measurement of Length

- The way you measure an attribute may not match the attributes properties.



**This scale preserves only the ordering property of length.**

**This scale preserves the ordering and additvity properties of length.**

# Types of Attributes

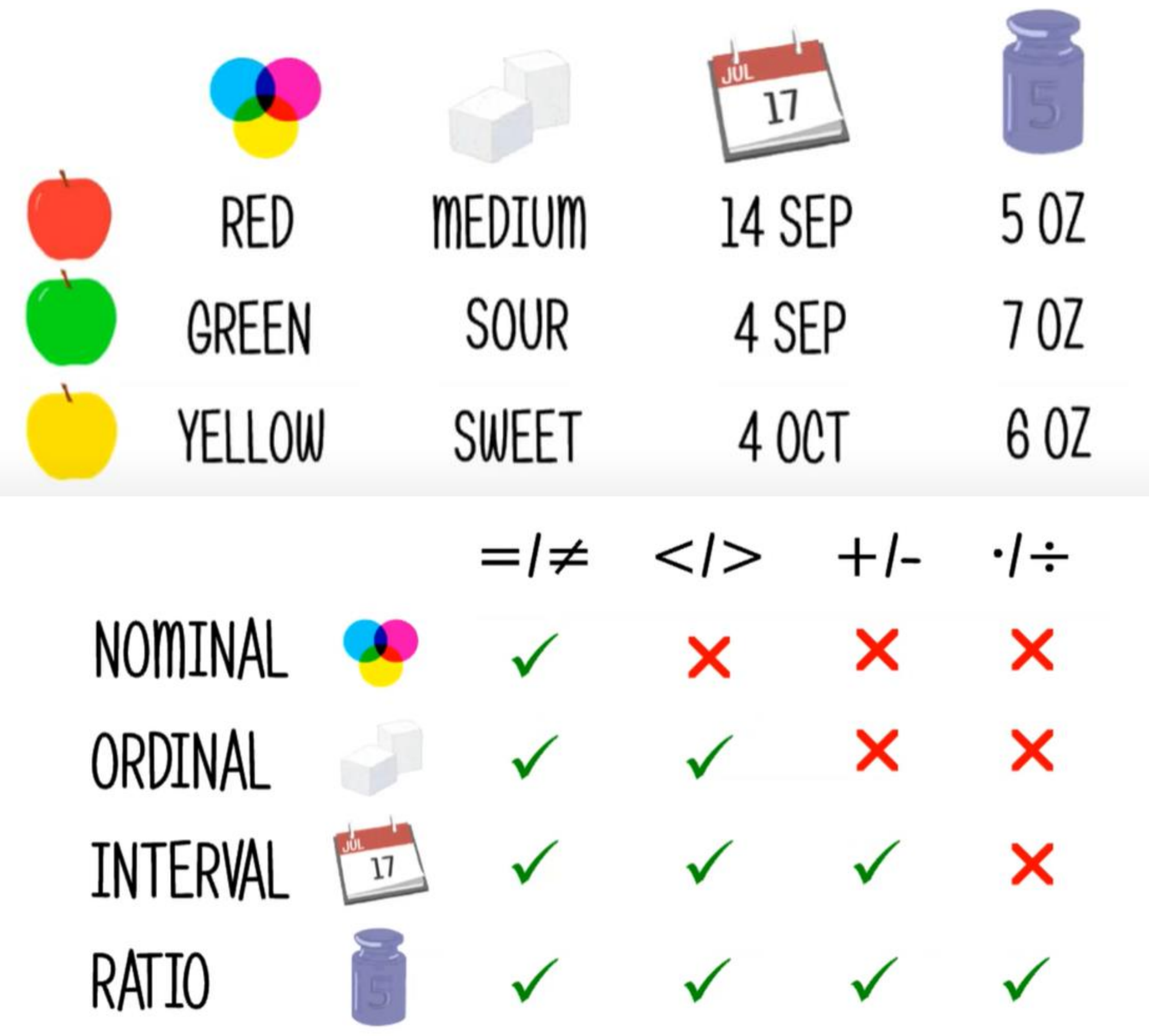
- There are different types of attributes
  - Nominal (numbers have no real meaning other than differentiating between objects)
    - Examples: ID numbers, eye color, zip codes
  - Ordinal (numbers have meaningful order)
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval (Numbers have order but there are also equal intervals between adjacent categories)
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio (Differences are meaningful (like interval) plus ratios are meaningful and there is a true zero point)
    - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

# Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful : + -
  - Ratios are meaningful * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & meaningful differences
  - Ratio attribute: all 4 properties/operations

# Properties of Attribute Values Example



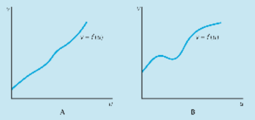
| | | | |
|---|---|---|---|
| RED | MEDIUM | 14 SEP | 5 OZ |
| GREEN | SOUR | 4 SEP | 7 OZ |
| YELLOW | SWEET | 4 OCT | 6 OZ |

| | =/≠ | </> | +/- | ·/÷ |
|---|---|---|---|---|
| NOMINAL | ✓ | ✗ | ✗ | ✗ |
| ORDINAL | ✓ | ✓ | ✗ | ✗ |
| INTERVAL | ✓ | ✓ | ✓ | ✗ |
| RATIO | ✓ | ✓ | ✓ | ✓ |

|  | **Attribute Type** | **Description** | **Examples** | **Operations** |
|---|---|---|---|---|
| Categorical<br>Qualitative | Nominal | Nominal attribute values only distinguish. (=, ≠) | zip codes, employee ID numbers, eye color, sex: {*male, female*} | mode, entropy, contingency correlation, $\chi$2 test |
|  | Ordinal | Ordinal attribute values also order objects.<br>(<, >) | hardness of minerals, {*good, better, best*}, grades, street numbers | median, percentiles, rank correlation, run tests, sign tests |
| Numeric<br>Quantitative | Interval | For interval attributes, differences between values are meaningful. (+, - ) | calendar dates, temperature in Celsius or Fahrenheit | mean, standard deviation, Pearson's correlation, *t* and *F* tests |
|  | Ratio | For ratio variables, both differences and ratios are meaningful. (*, /) | temperature in Kelvin, monetary quantities, counts, age, mass, length, current | geometric mean, harmonic mean, percent variation |

**This categorization of attributes is due to S. S. Stevens**

| | **Attribute Type** | **Transformation** | **Comments** |
|---|---|---|---|
| Categorical<br>Qualitative | Nominal | Any permutation of values | If all employee ID numbers were reassigned, would it make any difference? |
| | Ordinal | An order preserving change of values, i.e.,<br>*new_value = f(old_value)*<br>where *f* is a monotonic function | An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}. |
| Numeric<br>Quantitative | Interval | *new_value = a * old_value + b*<br>where a and b are constants | Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree). |
| | Ratio | *new_value = a * old_value* | Length can be measured in meters or feet. |



A positive monotonic transformation. Panel A illustrates a monotonic function one that is always increasing. Panel B illustrates a function that is not monotonic, since it sometimes increases and sometimes decreases.

**This categorization of attributes is due to S. S. Stevens**

# Difference Between Ratio and Interval

- Imagine you have a bag of candies. In your bag, you have 10 candies and your friend has 5 candies. You have twice as many candies as your friend, right? That's because when it comes to candies, having zero candies means you really have none. This is like the Kelvin temperature scale, where zero really means zero.
- Now, let's imagine you and your friend are playing a game where you start with 10 points, and your friend starts with 5 points. In this game, having 0 points doesn't mean you're out of the game, it's just the starting point. If your friend gets 5 more points and goes up to 10 points, we can't say your friend has doubled their points, because the game didn't start at zero points. This is like temperature in degrees Celsius or Fahrenheit, where zero doesn't really mean there is no temperature.
- So, in the game, your points and your friend's points are like an interval scale, where you can't really say that one amount is 'double' the other. But in the candies example, your candies and your friend's candies are like a ratio scale, where you can say that one amount is 'double' the other. This is why knowing the difference between ratio and interval is important, it helps you make sense of things in a better way.

# Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10 ° is twice that of 5° on
  - the Celsius scale?
  - the Fahrenheit scale?
  - the Kelvin scale?

- Consider measuring the height above average
  - If Bill’s height is three inches above average and Bob’s height is six inches above average, then would we say that Bob is twice as tall as Bill?
  - Is this situation analogous to that of temperature?

# Discrete and Continuous Attributes

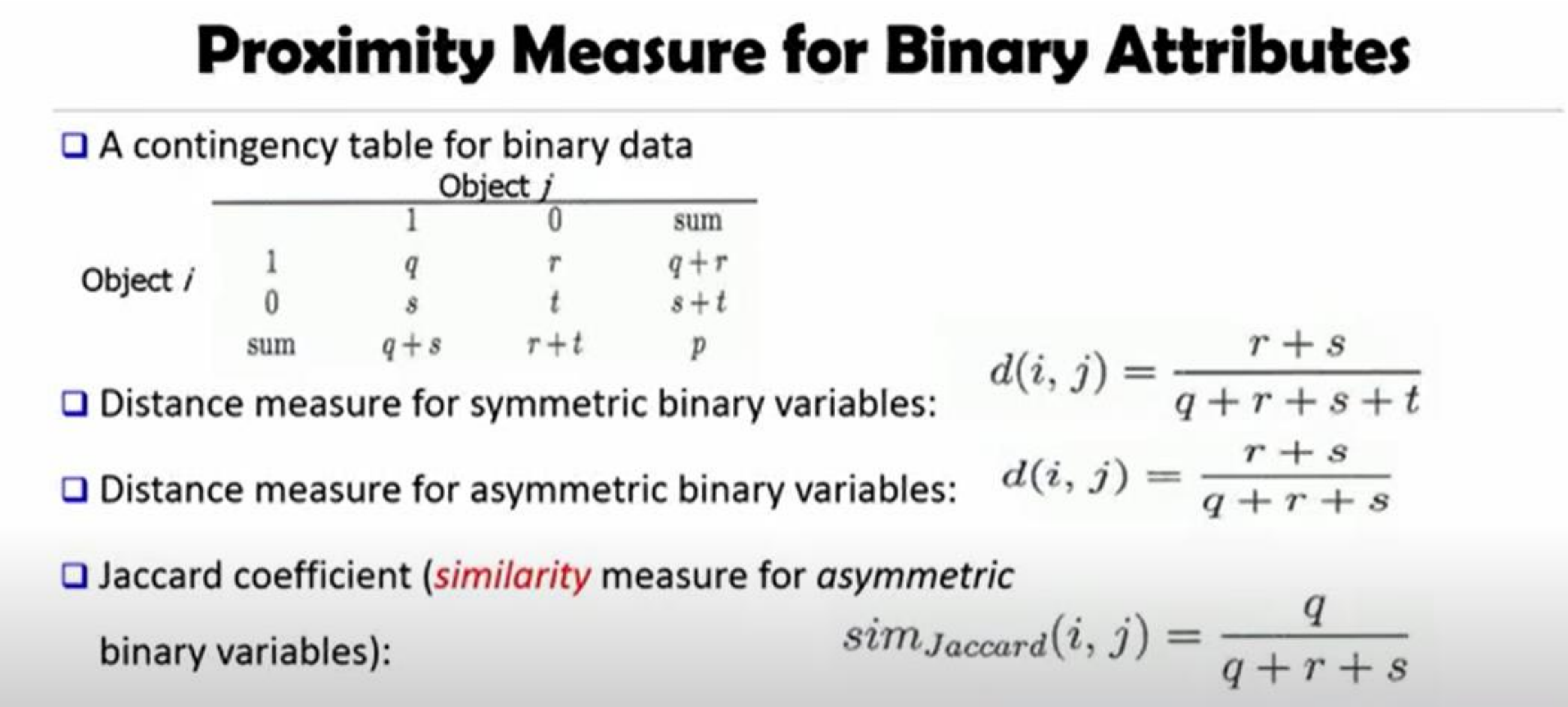
- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

# Asymmetric Attributes

- Asymmetric is the attribute which the two states are not equally important, for example, the medical test (positive or negative), here, the positive result is more significant than the negative one.
- Only presence (a non-zero attribute value) is regarded as important (only non-zero values are important)
  - Words present in documents
  - Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

  *"I see our purchases are very similar since we didn't buy most of the same things."*

# Asymmetric Attributes



## Proximity Measure for Binary Attributes

- A contingency table for binary data

|  |  | Object *j* |  |  |
|---|---|---|---|---|
|  |  | 1 | 0 | sum |
| Object *i* | 1 | $q$ | $r$ | $q+r$ |
|  | 0 | $s$ | $t$ | $s+t$ |
|  | sum | $q+s$ | $r+t$ | $p$ |

- Distance measure for symmetric binary variables:

$$d(i, j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables:

$$d(i, j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$$

# Critiques of the attribute categorization

- Incomplete
  - Asymmetric binary
  - Cyclical
  - Multivariate
  - Partially ordered
  - Partial membership
  - Relationships between the data

- Real data is approximate and noisy
  - This can complicate recognition of the proper attribute type
  - Treating one attribute type as another may be approximately correct

# Key Messages for Attribute Types

- The types of operations you choose should be “meaningful” for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are only four (among many possible) properties of data
  - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not present
  - Analysis may depend on these other properties of the data
    - Many statistical analyses depend only on the distribution
  - In the end, what is meaningful can be specific to domain

# Important Characteristics of Data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges

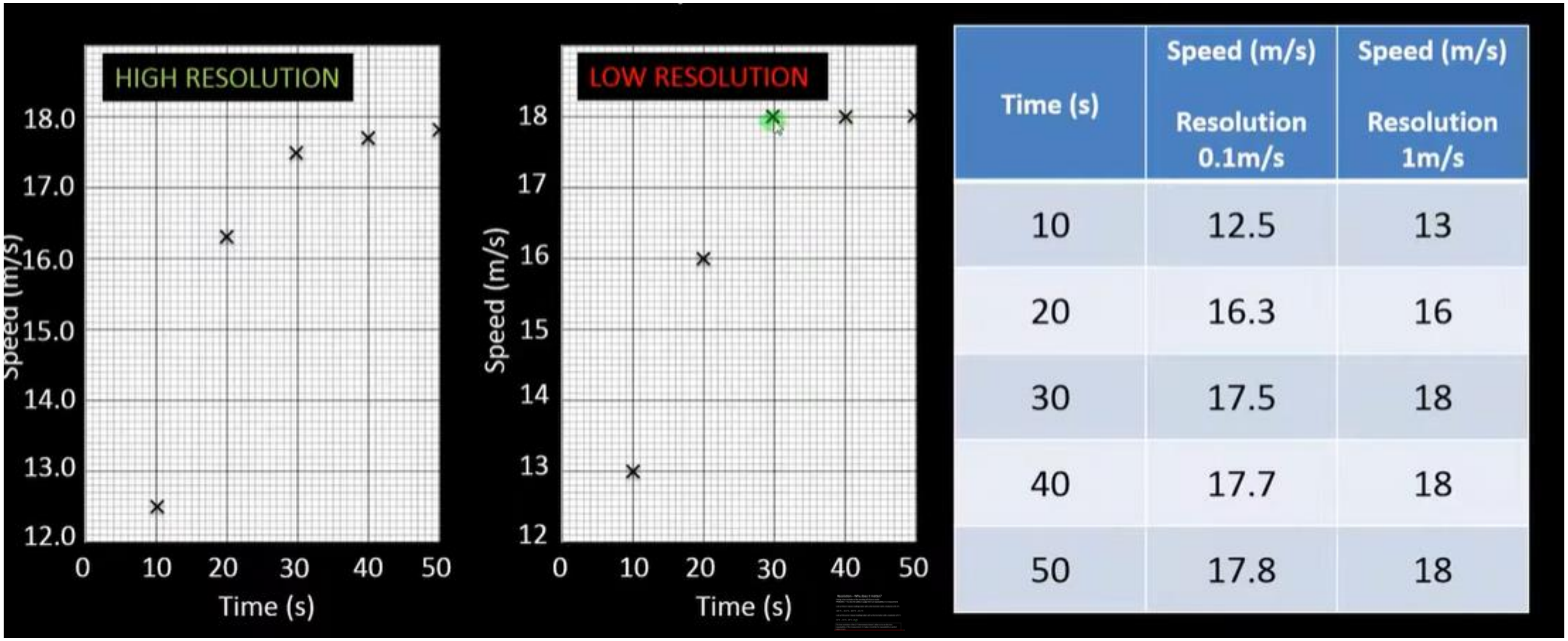
- Sparsity
  - Only presence counts, most attributes of an object has a value of zero, 1% of the values are non-zero, those values need to be stored and manipulated.

- Resolution
  - Patterns depend on the scale, if the resolution is too fine, a pattern may not be visible.

- Size
  - Type of analysis may depend on size of data

# Resolution

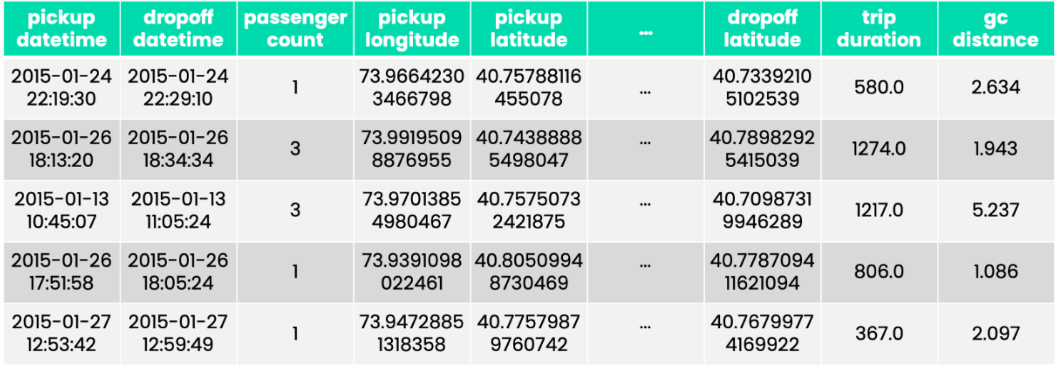
- Having a low resolution is like rounding off all your result.
- You could lose the ability to see small differences in a trend. (problem 1)



| Time (s) | Speed (m/s)<br>Resolution 0.1m/s | Speed (m/s)<br>Resolution 1m/s |
|---|---|---|
| 10 | 12.5 | 13 |
| 20 | 16.3 | 16 |
| 30 | 17.5 | 18 |
| 40 | 17.7 | 18 |
| 50 | 17.8 | 18 |

# Types of data sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data: it is any type of data that directly or indirectly references a specific geographical area or location.
  - Temporal Data: it is simply data that represents a state in time.
  - Sequential Data: A common example of this is a Timeseries such as a stock price or a sensor data where each point represents an observation at a certain point in time. There are other examples of sequential data like sequences, gene sequences, and weather data.
  - Genetic Sequence Data



| pickup datetime | dropoff datetime | passenger count | pickup longitude | pickup latitude | ... | dropoff latitude | trip duration | gc distance |
|---|---|---|---|---|---|---|---|---|
| 2015-01-24 22:19:30 | 2015-01-24 22:29:10 | 1 | 73.96642303466798 | 40.75788116455078 | ... | 40.73392105102539 | 580.0 | 2.634 |
| 2015-01-26 18:13:20 | 2015-01-26 18:34:34 | 3 | 73.99195098876955 | 40.74388885498047 | ... | 40.78982925415039 | 1274.0 | 1.943 |
| 2015-01-13 10:45:07 | 2015-01-13 11:05:24 | 3 | 73.97013854980467 | 40.75750732421875 | ... | 40.70987319946289 | 1217.0 | 5.237 |
| 2015-01-26 17:51:58 | 2015-01-26 18:05:24 | 1 | 73.9391098022461 | 40.80509948730469 | ... | 40.778709411621094 | 806.0 | 1.086 |
| 2015-01-27 12:53:42 | 2015-01-27 12:59:49 | 1 | 73.94728851318358 | 40.77579879760742 | ... | 40.76799774169922 | 367.0 | 2.097 |

# Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 95K | **Yes** |
| 6 | No | Married | 60K | **No** |
| 7 | Yes | Divorced | 220K | **No** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 75K | **No** |
| 10 | No | Single | 90K | **Yes** |

# Document Data

- Each document becomes a ‘term’ vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document.

| | team | coach | play | ball | score | game | win | lost | timeout | season |
|---|---|---|---|---|---|---|---|---|---|---|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

# Transaction Data

- A special type of data, where
  - Each transaction involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
  - Can represent transaction data as record data

| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

# Data Matrix

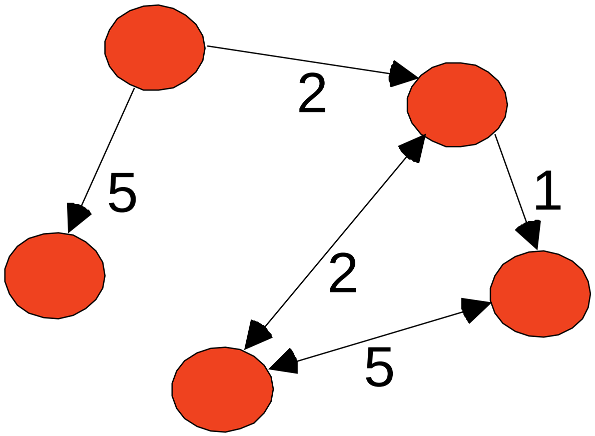
- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

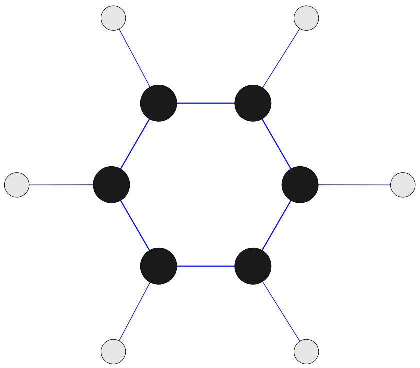
- Such a data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute

| **Projection of x Load** | **Projection of y load** | **Distance** | **Load** | **Thickness** |
|---|---|---|---|---|
| 10.23 | 5.27 | 15.22 | 2.7 | 1.2 |
| 12.65 | 6.25 | 16.22 | 2.2 | 1.1 |

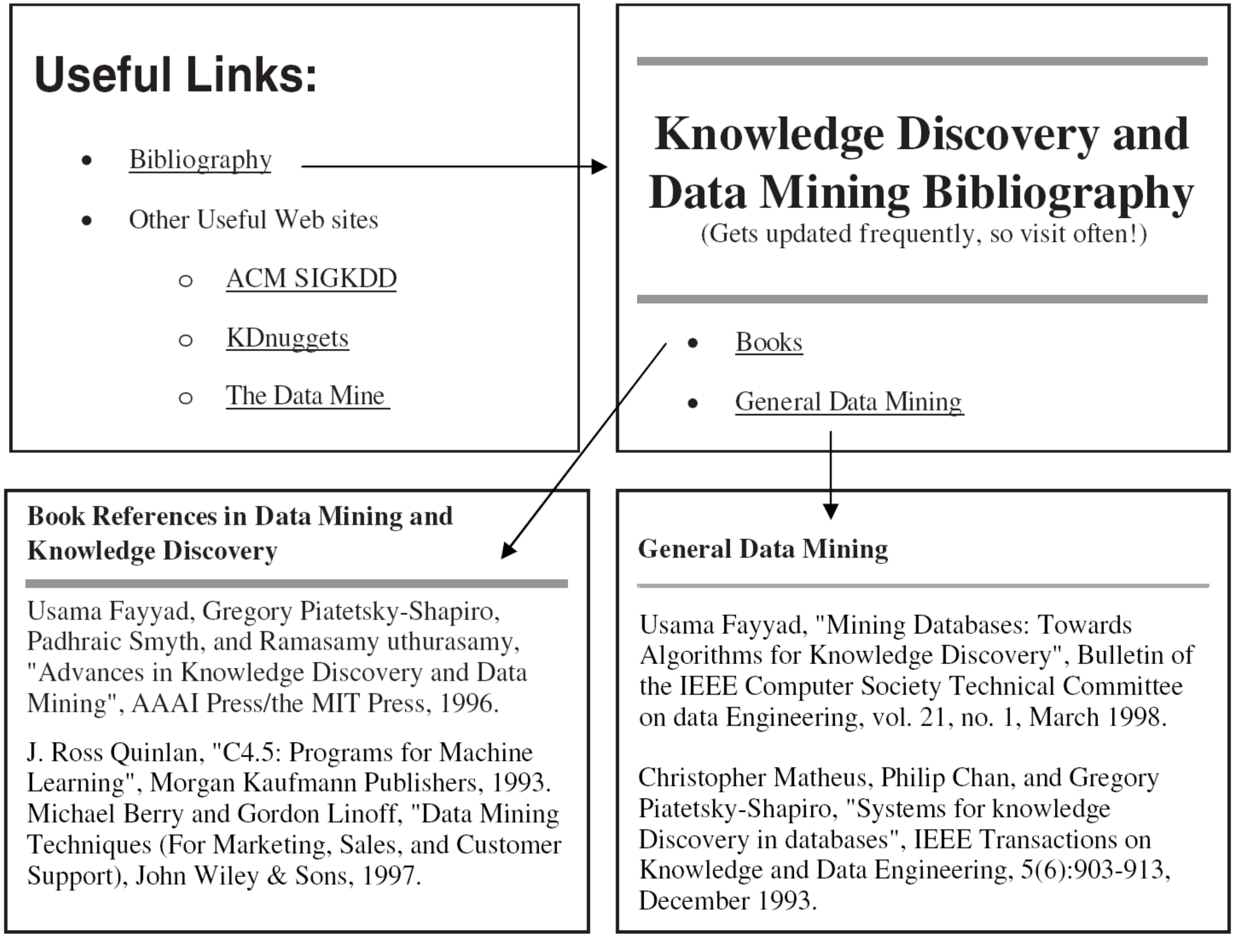
# Graph Data

- Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6



**Useful Links:**

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

**Knowledge Discovery and Data Mining Bibliography**

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

**Book References in Data Mining and Knowledge Discovery**

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

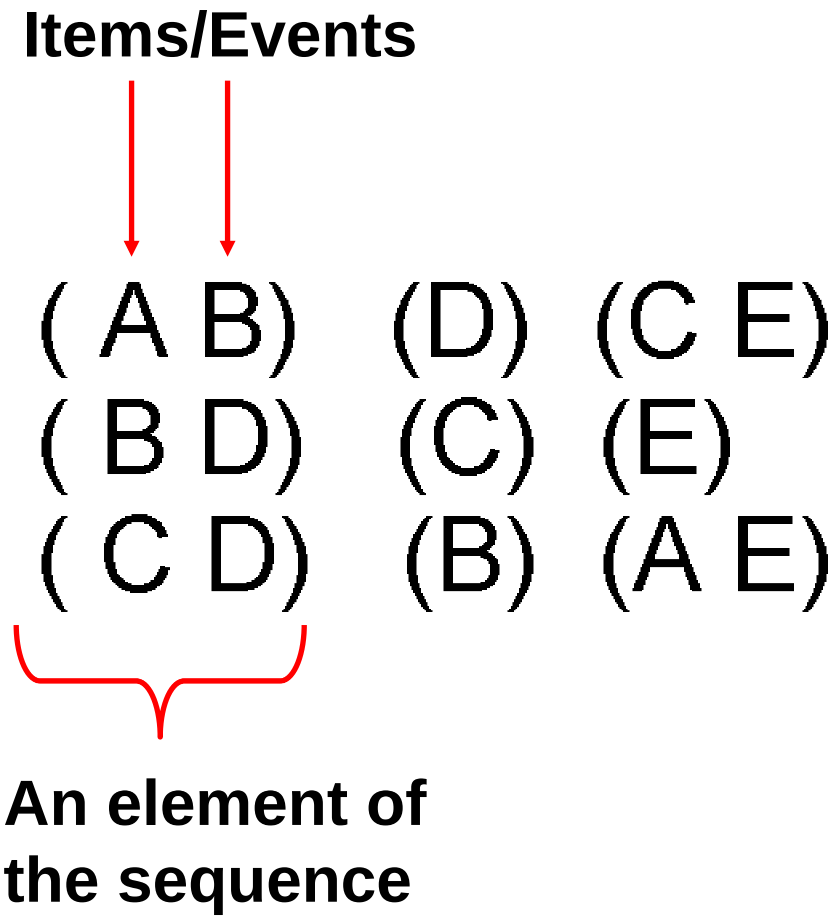
**General Data Mining**

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

# Ordered Data

- Sequences of transactions



Items/Events

( A B) (D) (C E)
( B D) (C) (E)
( C D) (B) (A E)

An element of the sequence

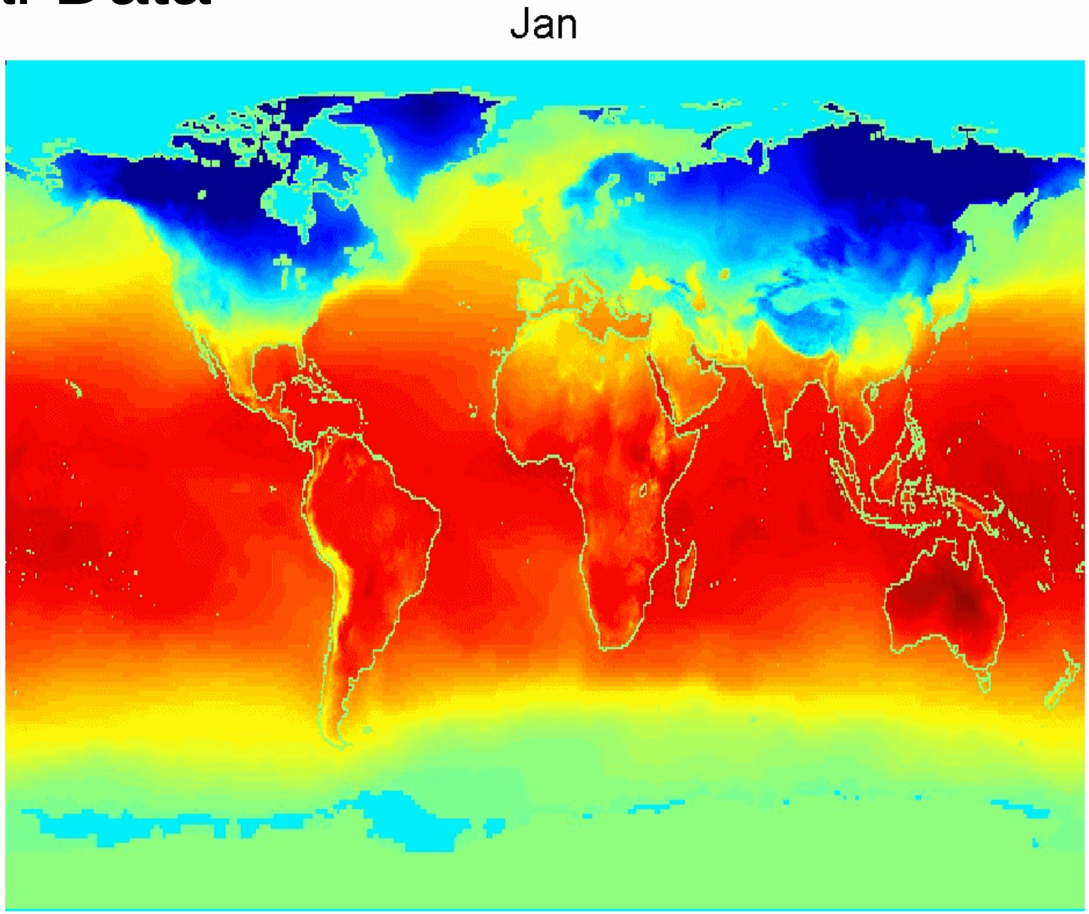
# Ordered Data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

# Ordered Data

- Spatio-Temporal Data



**Average Monthly Temperature of land and ocean**

# Data Quality

- Poor data quality negatively affects many data processing efforts

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

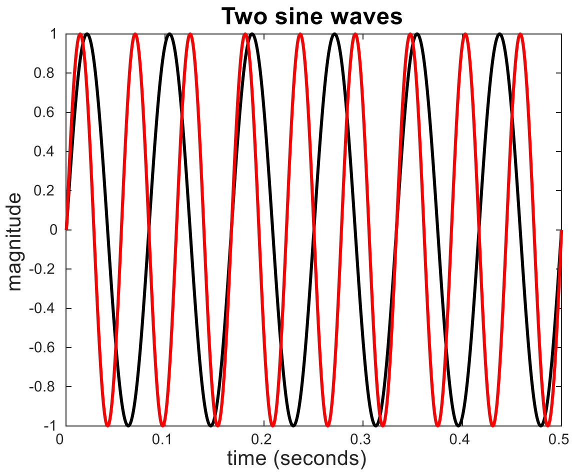
# Data Quality …

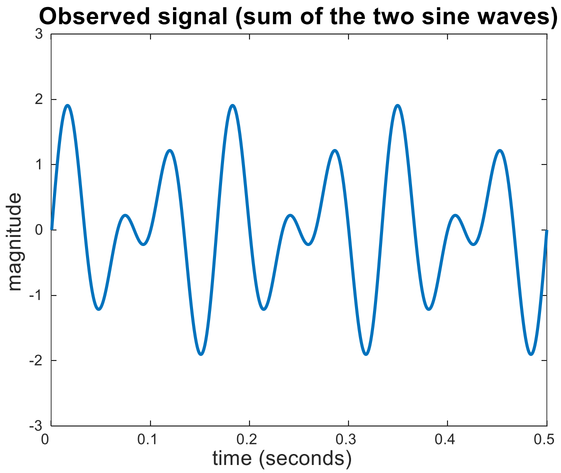
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

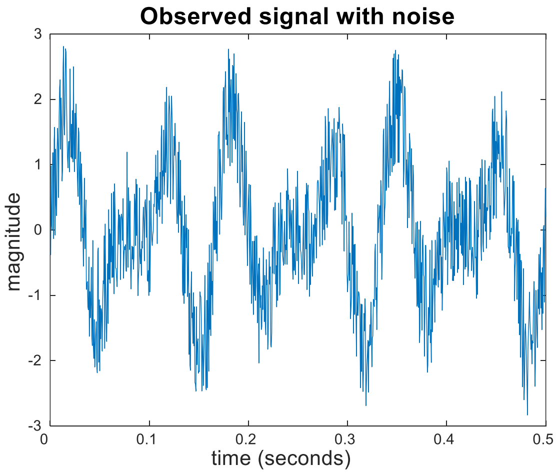
- Examples of data quality problems:
  - Noise and outliers
  - Wrong data
  - Fake data
  - Missing values
  - Duplicate data

# Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person’s voice when talking on a poor phone and “snow” on television screen
  - The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise
    - The magnitude and shape of the original signal is distorted


Two sine waves


Observed signal (sum of the two sine waves)


Observed signal with noise

# Noise: Example

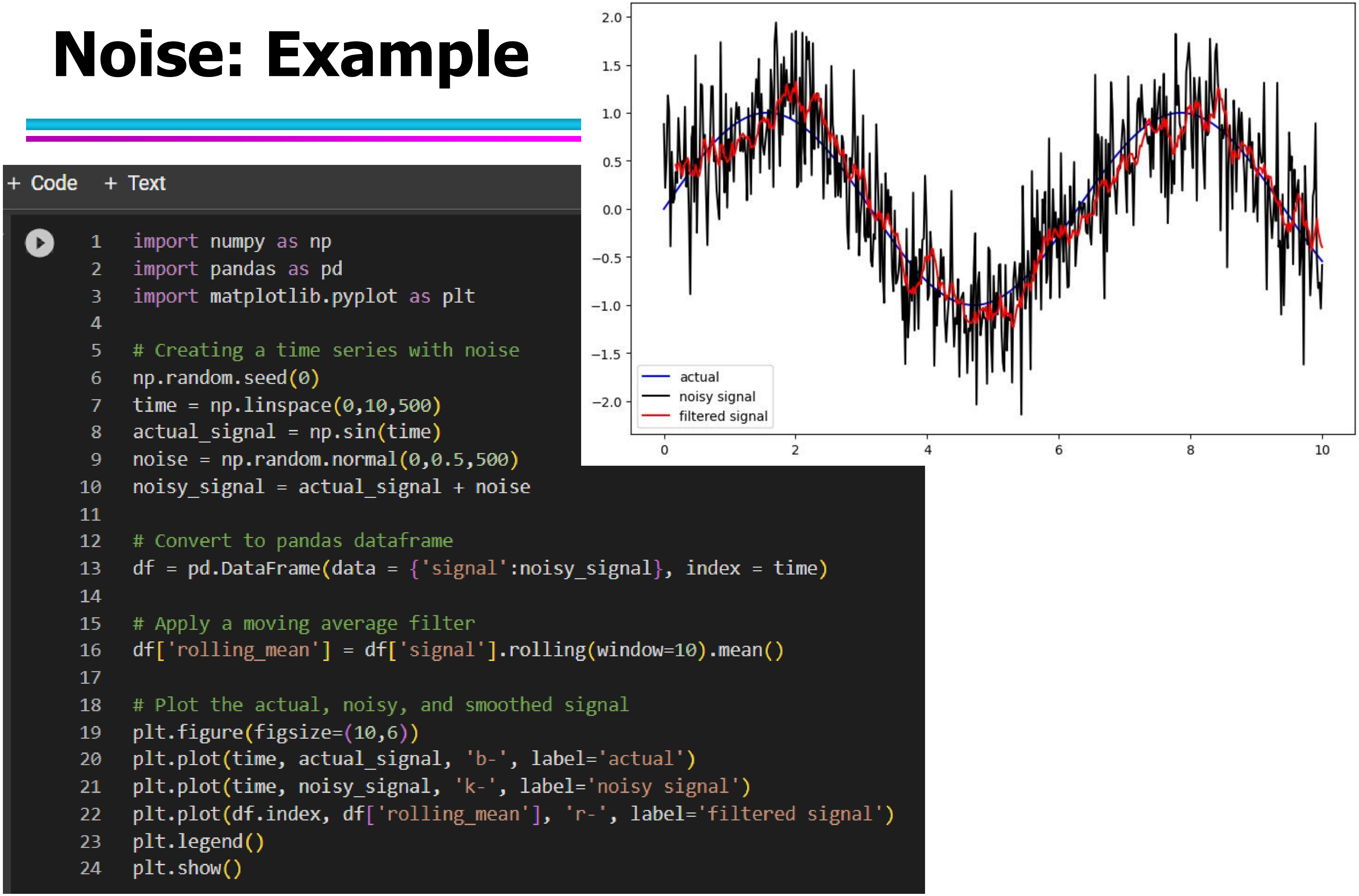
**Question**:

In signal processing or data analysis, we often deal with noisy data. For instance, let's say we have collected sensor data over time, and this data is contaminated (impure) with random noise. How can we reduce or filter out this noise to reveal the true underlying signal?

**Answer**:

We can use a **moving average filter**, a simple and widely used method, to smooth our time series data and reduce noise. Here's a Python example using the pandas library:

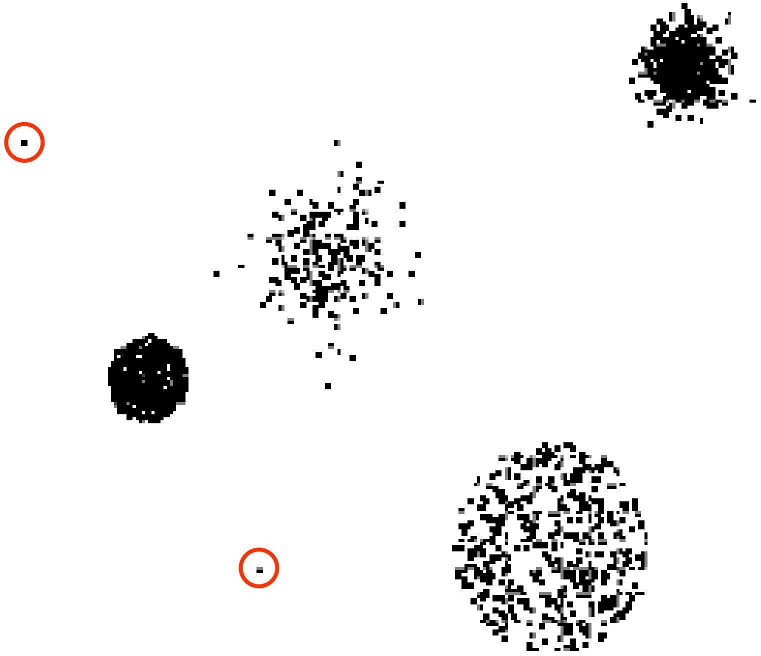
# Noise: Example



```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Creating a time series with noise
np.random.seed(0)
time = np.linspace(0,10,500)
actual_signal = np.sin(time)
noise = np.random.normal(0,0.5,500)
noisy_signal = actual_signal + noise

# Convert to pandas dataframe
df = pd.DataFrame(data = {'signal':noisy_signal}, index = time)

# Apply a moving average filter
df['rolling_mean'] = df['signal'].rolling(window=10).mean()

# Plot the actual, noisy, and smoothed signal
plt.figure(figsize=(10,6))
plt.plot(time, actual_signal, 'b-', label='actual')
plt.plot(time, noisy_signal, 'k-', label='noisy signal')
plt.plot(df.index, df['rolling_mean'], 'r-', label='filtered signal')
plt.legend()
plt.show()
```

# Outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection

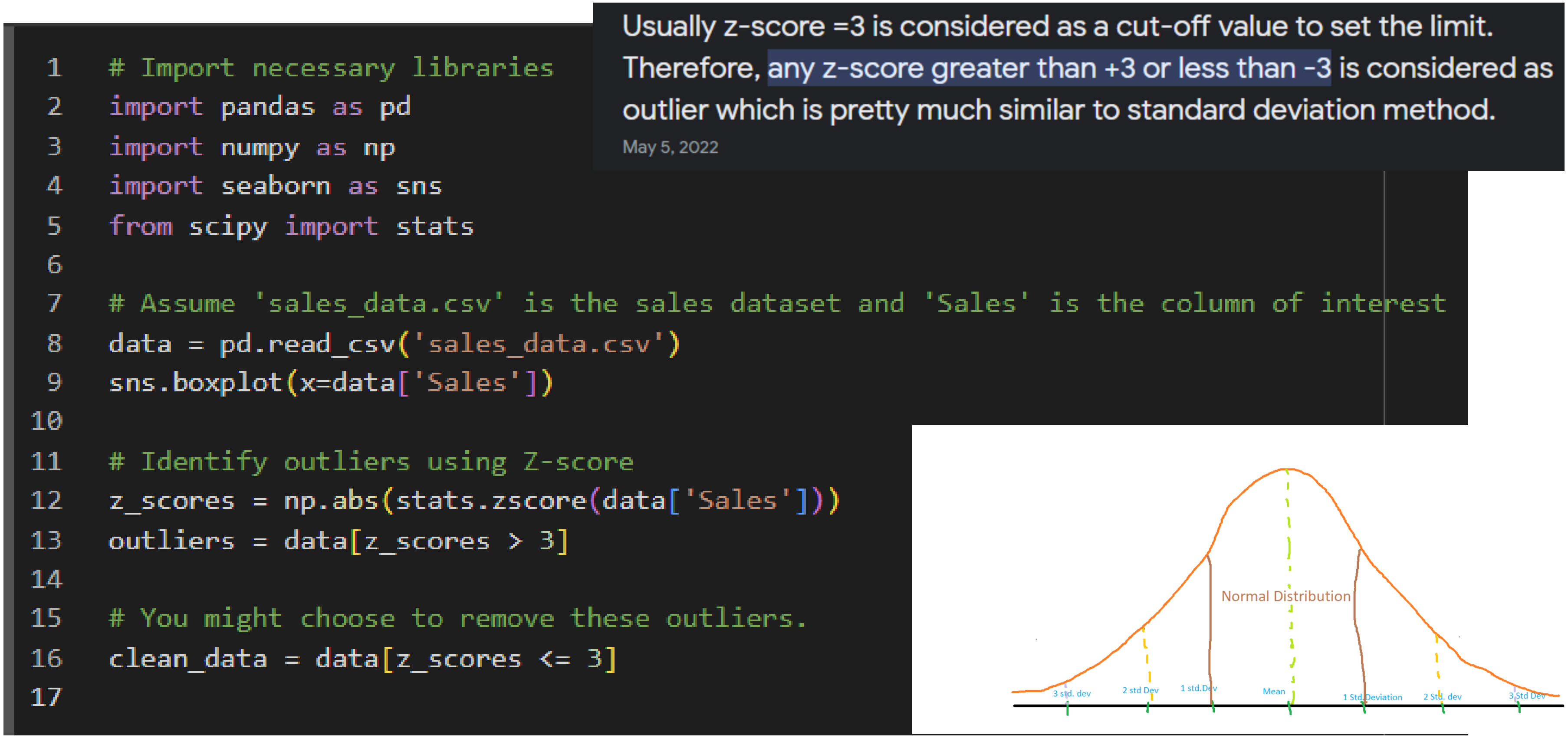


- Causes?

# Question 1 (Case 1 - Outliers as Noise)

*We have a dataset of monthly sales figures for a retail store. The goal is to model and forecast future sales. However, there was a one-time, massive sales event during one particular month that is considered an outlier and might interfere with the analysis. How can we identify and handle this outlier?*

# General Solution for Case 1 - Outliers as Noise



```python
# Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
from scipy import stats

# Assume 'sales_data.csv' is the sales dataset and 'Sales' is the column of interest
data = pd.read_csv('sales_data.csv')
sns.boxplot(x=data['Sales'])

# Identify outliers using Z-score
z_scores = np.abs(stats.zscore(data['Sales']))
outliers = data[z_scores > 3]

# You might choose to remove these outliers.
clean_data = data[z_scores <= 3]

```

Usually z-score =3 is considered as a cut-off value to set the limit. Therefore, any z-score greater than +3 or less than -3 is considered as outlier which is pretty much similar to standard deviation method.

May 5, 2022

https://seaborn.pydata.org/generated/seaborn.boxplot.html

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html

https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/

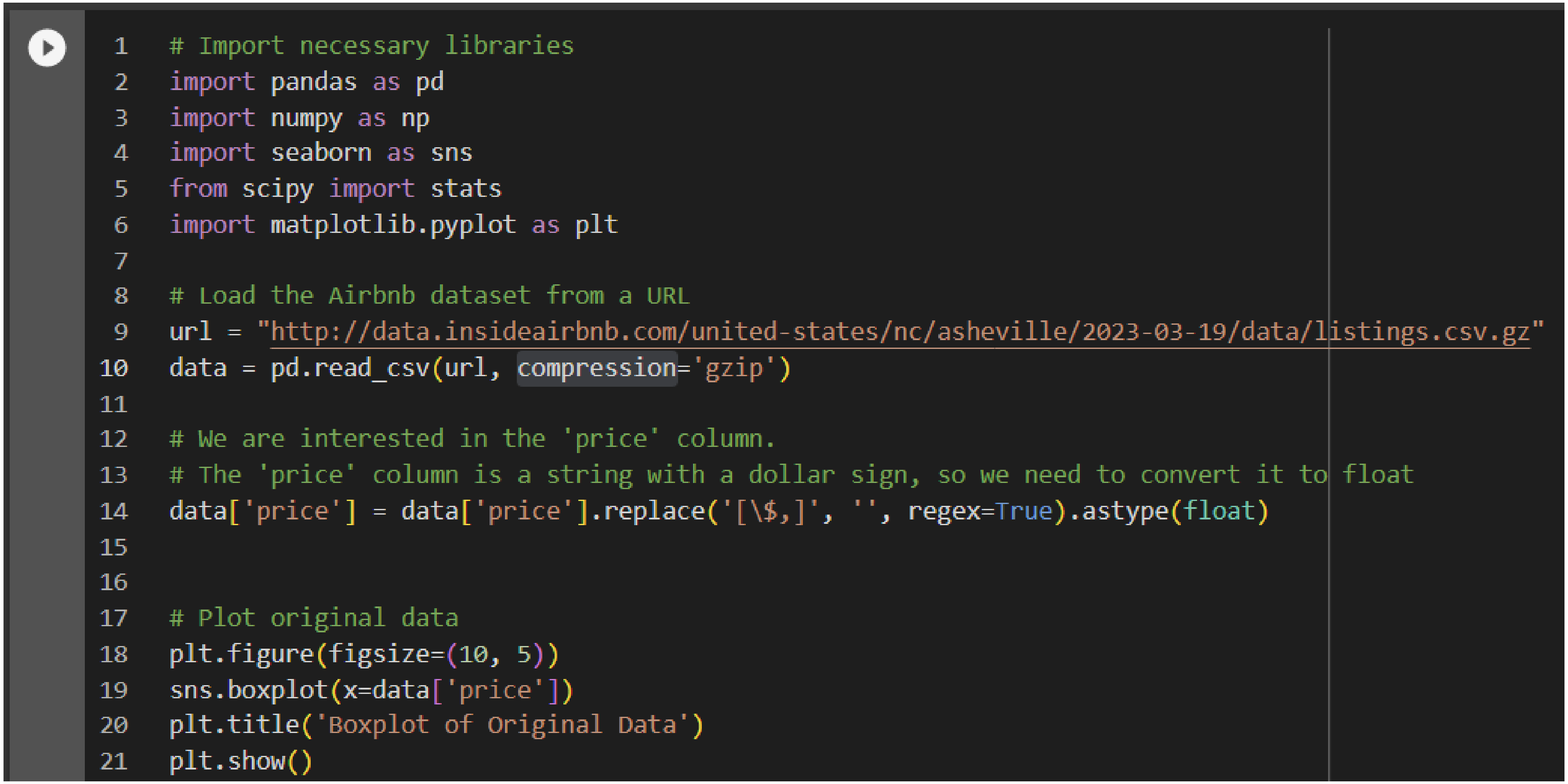
# Question 1 (Case 1 - Outliers as Noise)

Let's consider the Airbnb dataset (publicly available on the internet) containing information about various listings, including the price of the listing per night. However, some listings may have unusually high or low prices that can be considered outliers and might interfere with our analysis. How can we identify and handle these outliers?

http://insideairbnb.com/get-the-data/

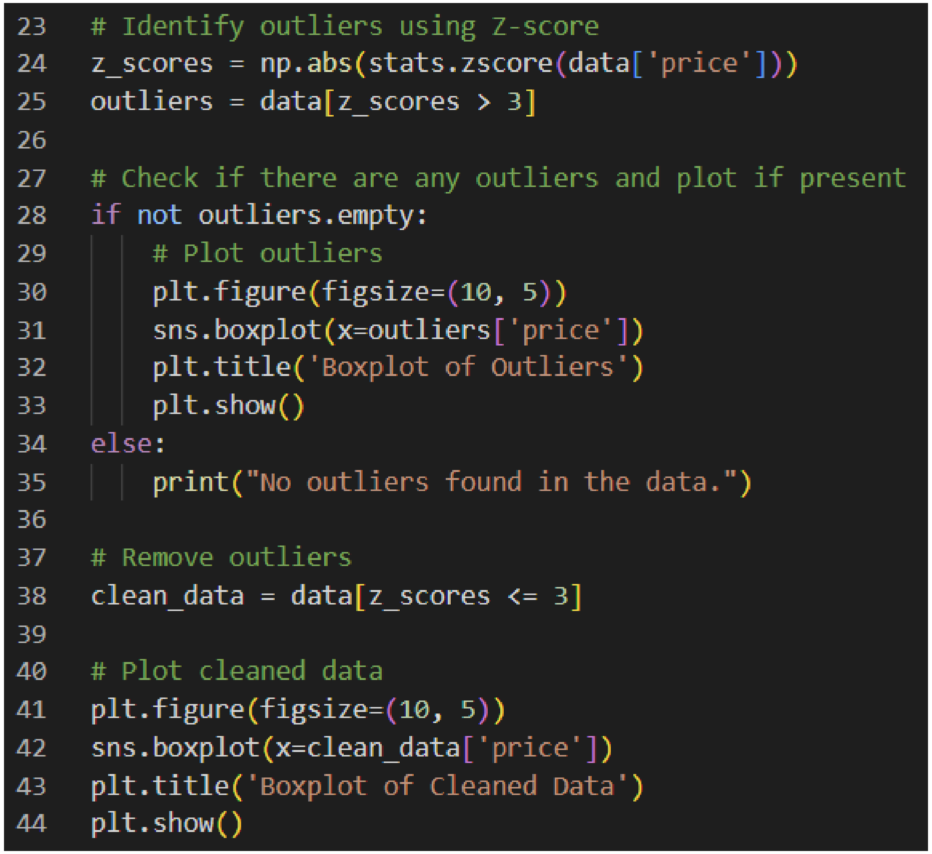
http://data.insideairbnb.com/united-states/nc/asheville/2023-03-19/data/listings.csv.gz

# Question 1: Airbnb

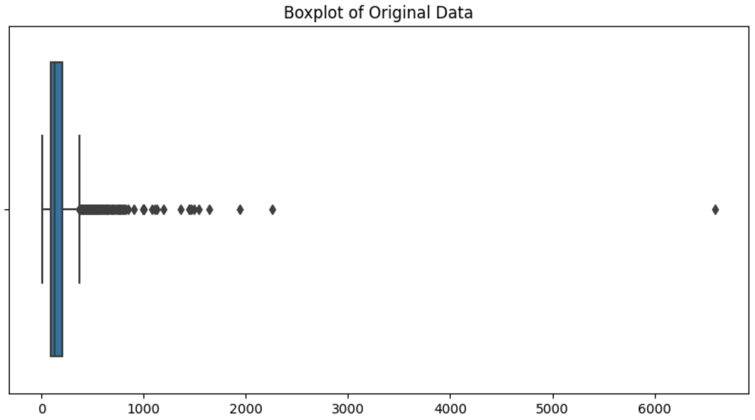


```
# Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
from scipy import stats
import matplotlib.pyplot as plt

# Load the Airbnb dataset from a URL
url = "http://data.insideairbnb.com/united-states/nc/asheville/2023-03-19/data/listings.csv.gz"
data = pd.read_csv(url, compression='gzip')

# We are interested in the 'price' column.
# The 'price' column is a string with a dollar sign, so we need to convert it to float
data['price'] = data['price'].replace('[\$,]', '', regex=True).astype(float)


# Plot original data
plt.figure(figsize=(10, 5))
sns.boxplot(x=data['price'])
plt.title('Boxplot of Original Data')
plt.show()
```

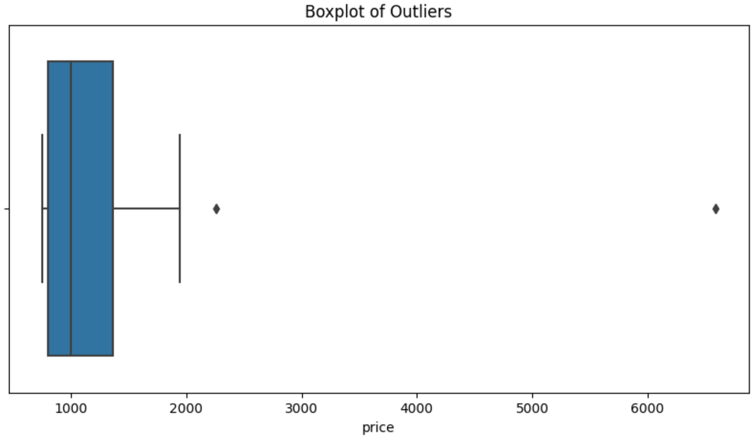
# Question 1: Airbnb



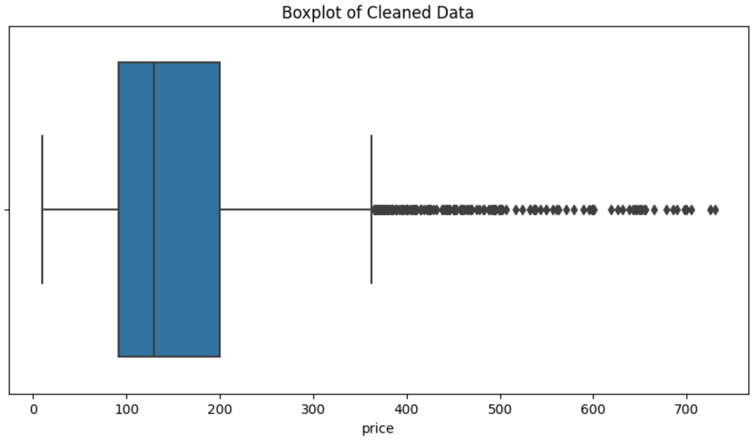
```python
# Identify outliers using Z-score
z_scores = np.abs(stats.zscore(data['price']))
outliers = data[z_scores > 3]

# Check if there are any outliers and plot if present
if not outliers.empty:
    # Plot outliers
    plt.figure(figsize=(10, 5))
    sns.boxplot(x=outliers['price'])
    plt.title('Boxplot of Outliers')
    plt.show()
else:
    print("No outliers found in the data.")

# Remove outliers
clean_data = data[z_scores <= 3]

# Plot cleaned data
plt.figure(figsize=(10, 5))
sns.boxplot(x=clean_data['price'])
plt.title('Boxplot of Cleaned Data')
plt.show()
```



Boxplot of Original Data



Boxplot of Outliers



Boxplot of Cleaned Data

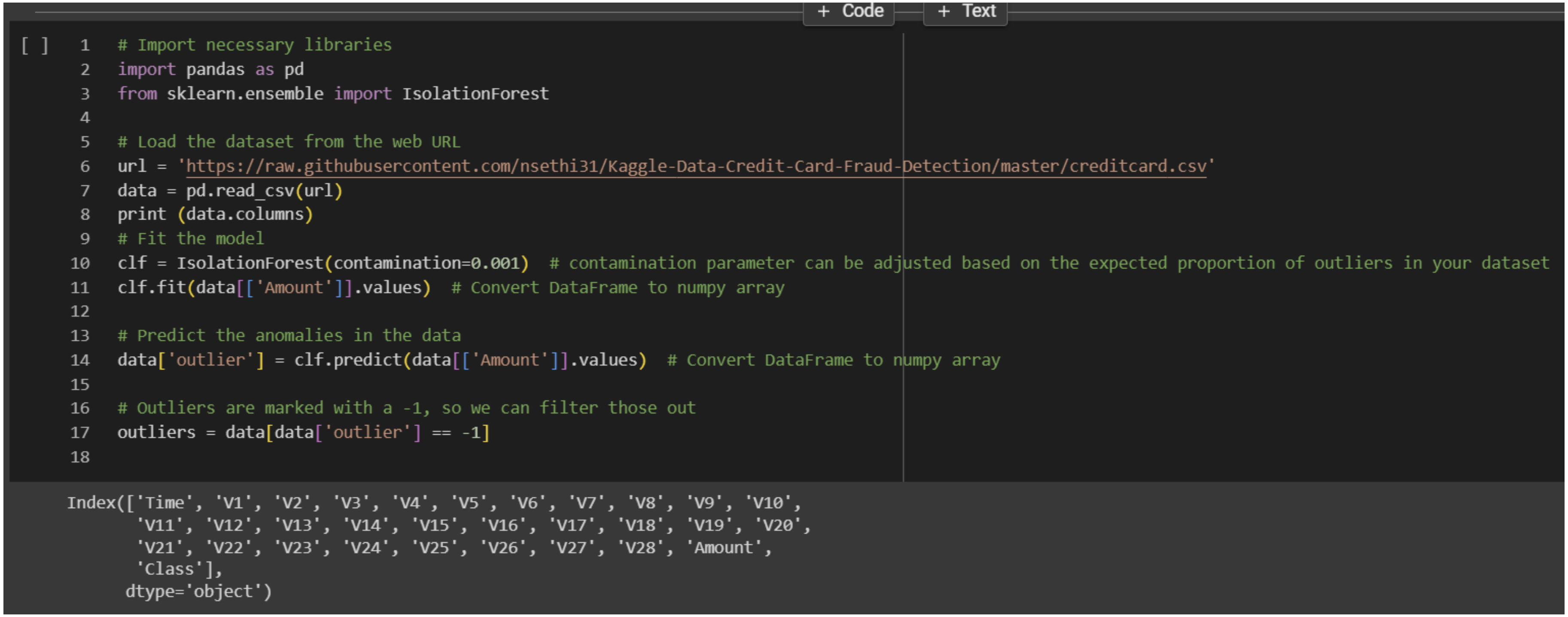
https://www.youtube.com/watch?v=INSIyaZUXIY

# Question 2 (Case 2 - Outliers as Targets)

- We are analyzing credit card transactions to detect fraudulent activities. Outliers in this case could be unusually high transaction amounts that deviate from a user's typical behavior. How can we detect these outliers?
- To solve this problem, we can use the Credit Card Fraud Detection dataset available on Kaggle. This dataset contains transactions made by credit cards, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
- The 'Amount' column in this dataset can be considered for outlier analysis.

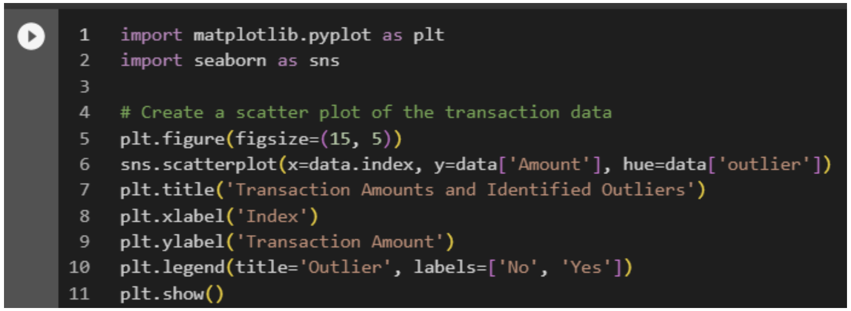
# Question 2: Answer

https://www.youtube.com/watch?v=O9VvmWj-JAk

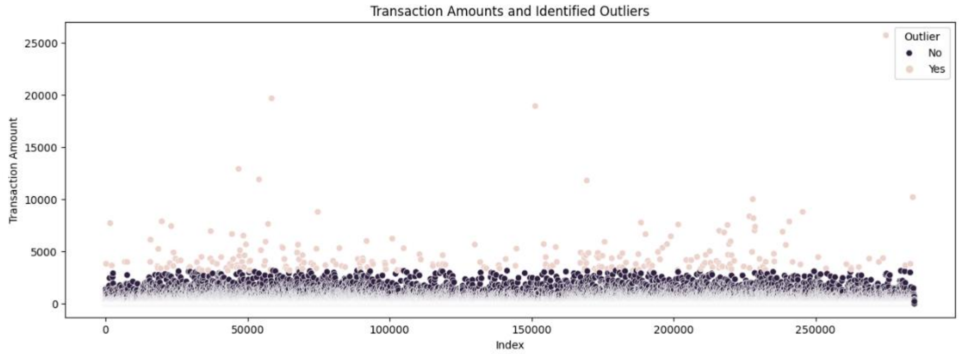


```python
# Import necessary libraries
import pandas as pd
from sklearn.ensemble import IsolationForest

# Load the dataset from the web URL
url = 'https://raw.githubusercontent.com/nsethi31/Kaggle-Data-Credit-Card-Fraud-Detection/master/creditcard.csv'
data = pd.read_csv(url)
print (data.columns)
# Fit the model
clf = IsolationForest(contamination=0.001)  # contamination parameter can be adjusted based on the expected proportion of outliers in your dataset
clf.fit(data[['Amount']].values)  # Convert DataFrame to numpy array

# Predict the anomalies in the data
data['outlier'] = clf.predict(data[['Amount']].values)  # Convert DataFrame to numpy array

# Outliers are marked with a -1, so we can filter those out
outliers = data[data['outlier'] == -1]

```

```
Index(['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10',
       'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20',
       'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Amount',
       'Class'],
      dtype='object')
```



```python
import matplotlib.pyplot as plt
import seaborn as sns

# Create a scatter plot of the transaction data
plt.figure(figsize=(15, 5))
sns.scatterplot(x=data.index, y=data['Amount'], hue=data['outlier'])
plt.title('Transaction Amounts and Identified Outliers')
plt.xlabel('Index')
plt.ylabel('Transaction Amount')
plt.legend(title='Outlier', labels=['No', 'Yes'])
plt.show()
```



# Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate data objects or variables
  - Estimate missing values
    - ◆ Example: time series of temperature
    - ◆ Example: census results
  - Ignore the missing value during analysis

# Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Process of dealing with duplicate data issues

- When should duplicate data not be removed?