Project #8
Adversarial Robustness in Machine Learning Models

Data Understanding Report

Members:
Sai Coumar - 95%
Patrick Florendo - 5%
Nischay Uppal - 0%
Rishabh Pandey - 0%
Supriya Dixit - 0%

1. COLLECT INITIAL DATA

Acquire the data (or access to the data) listed in the project resources. This initial collection includes data loading, if necessary for data understanding. For example, if you intend to use a specific tool for data understanding, it is logical to load your data into this tool.
Describe all the various data used for the project, and include any selection requirements for more detailed data. The data collection report should also define whether some attributes are relatively more important than others.
Remember that any assessment of data quality should be made not just of the individual data sources but also of any data that results from merging data sources. Because of inconsistencies between the sources, merged data may present problems that do not exist in the individual data sources.

Data requirements planning

- Plan which information is needed (e.g., only for given attributes, or specific additional information)
- Check if all the information needed (to solve the data mining goals) is actually available

Attributes in the dataset represent image pixels. All pixels are needed to recreate the images and use them as training input. One attribute in each dataset represents the labeled value of the image. This attribute is necessary for supervised learning.

Our datasets - MNIST, CIFAR10, and SVHN - are all cleaned and include all necessary attributes. No missing data exists in our dataset.

Selection criteria

- Specify selection criteria (e.g., Which attributes are necessary for the specified data mining goals? Which attributes have been identified as being irrelevant? How many attributes can we handle with the chosen techniques?)
- Select tables/files of interest
- Select data within a table/file
- Think about how long a history one should use (e.g., even if 18 months of data are available, only 12 months may be needed for the exercise)

All attributes are needed and all files are needed, as they are image data

Note: SVHN is split into two subsets: One which is cropped around numbers in a .mat table and one which uses the original dataset of raw images. Using both may be disadvantageous/redundant and one might be preferable to use based on testing in the modeling phase. Labels could not be found in the raw image dataset so the cropped dataset seems to be preferable.

History does not affect our data quality so we should use all available data.

Be aware that data collected from different sources may give rise to quality problems when merged (e.g., address files merged with a customer database may show inconsistencies of format, invalidity of data, etc.).

Insertion of data

- If the data contain free text entries, do we need to encode them for modeling or do we want to group specific entries?
- How can missing attributes be acquired?
- How can we best extract the data?

The MNIST and CIFAR10 datasets come as one-hot encoding vectors which can be used in modeling. The SVHN dataset came as image data, but while creating summary statistics functions to convert image data to one-hot encoding vectors and vice versa were made. We now have standardized data structures of a one-hot encoding vector and a label associated with it.

There is no missing data to acquire.

Pandas/numpy can be used to extract the MNIST data. Pickle and helper functions are used to extract the CIFAR10 data into Pandas/numpy. Scipy.io can be used to load the SVHN data into Pandas/numpy.

Remember that some knowledge about the data may be available from non-electronic sources (e.g., from people, printed text, etc.).
Remember that it may be necessary to preprocess the data (time-series data, weighted averages, etc.).

2. DESCRIBE DATA

Examine the "gross" properties of the acquired data and report on the results.
Describe the data that has been acquired, including the format of the data, the quantity of the data (e.g., the number of records and fields within each table), the identities of the fields, and any other surface features that have been discovered.

Volumetric analysis of data

- Identify data [Optional] and method of capture
- Access data sources
- Use statistical analyses if appropriate
- Report tables and their relations
- Check data volume [Optional], number of multiples, complexity
- Note if the data contain free text entries

We have 3 dataframes (post-processing to standardize SVHN into one-hot vector format like MNIST and CIFAR10

1. MNIST
    - The MNIST dataframe has one row for each image. There are 784 columns for each pixel in the flattened 28x28 image and one column for the label
2. CIFAR10
    - The CIFAR10 dataframe has one row for each image. There are 3072 columns for each pixel in the flattened 32x32 image and one column for the label
3. SVHN
    - The SVHNdataframe has one row for each image. There are 3072 columns for each pixel in the flattened 32x32 image and one column for the label

Each of the dataframes has 10 labels that the images are classified into. For MNIST and SVHN these are numbers, but for CIFAR10 they are index values that are mapped to a label dictionary. Classes for CIFAR10 include automobiles, horses, and cats, for example.

MNIST has 60,000 training records and 10,000 testing records
CIFAR10 has 50,000 training records and 10,000 testing records
SVHN has 73,257 training records and 25,032 records. SVHN has 531,131 "extra" records

All data is in the form of integers. Labels range from 0-9 for MNIST and CIFAR10, and 1-10 for SVHN. Pixel values range from 0-255. MNIST uses grayscale images while CIFAR and SVHN use RGB channels.

Attribute types and values

- Check accessibility and availability of attributes
- Check attribute types (numeric, symbolic, taxonomy, etc.)
- Check attribute value ranges
- Analyze attribute correlations
- Understand the meaning of each attribute and attribute value in business terms
- [Optional] For each attribute, compute basic statistics (e.g., compute distribution, average, max, min, standard deviation, variance, mode, skewness, etc.)
- [Optional] Analyze basic statistics and relate the results to their meaning in business terms
- Decide if the attribute is relevant for the specific data mining goal
- [Optional] Determine if the attribute meaning is used consistently
- Interview domain experts to obtain their opinion of attribute relevance
- Decide if it is necessary to balance the data (based on the modeling techniques to be used)

All data is in the form of integers. Labels range from 0-9 for MNIST and CIFAR10, and 1-10 for SVHN. Pixel values range from 0-255. MNIST uses grayscale images while CIFAR and SVHN use RGB channels.

Correlations are not easily analyzed with image data. In truth, all pixels matter in relation to each other as the relationships between them are what define the underlying distribution that is learned by a CNN.

Data transformations may be done for image processing purposes. Gaussian Blurring, Histogram Equalization, Noise Reduction, and resizing may all prove beneficial to making strong classifiers.

Keys

- Analyze key relationships
- Check amount of overlaps of key attribute values across tables

Table format is consistent with all 3 tables (post-processing). Therefore we can likely use the same architecture and plug in different datasets interchangeably to create multiple classifiers. Pixel attributes do not have direct correlations because we are working with image data.

Review assumptions/goals

- Update list of assumptions, if necessary

None.

3. EXPLORE DATA

This task tackles the data mining questions that can be addressed using querying, visualization, and reporting techniques. These analyses may directly address the data mining goals. However, they may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed before further analysis can occur.
Describe the results of this task, including first findings or initial hypotheses and their impact on the remainder of the project. The report may also include graphs and plots that indicate data characteristics or point to interesting data subsets worthy of further examination.

Data exploration

- Analyze properties of interesting attributes in detail (e.g., basic statistics, interesting sub populations)
- [Optional] Identify characteristics of sub-populations

To get an idea of the “average” image within a class we compute the mean image vector by grouping the data by label and then computing the mean one-hot encoding vector column-wise. We can then convert that to a human-interpretable image to get an idea of what this looks like.

For MNIST this worked very well due to the lack of noise and structured framing.

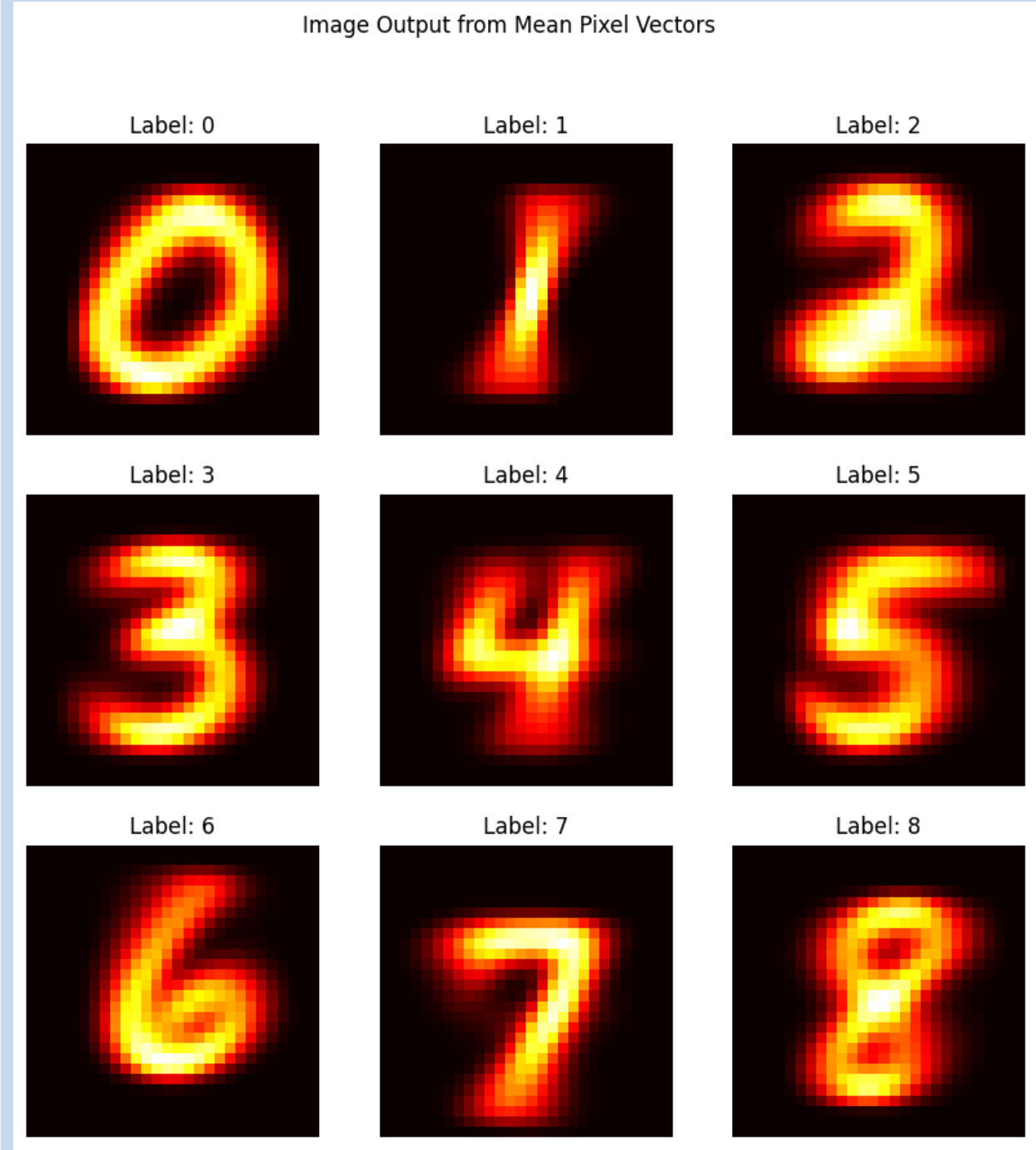
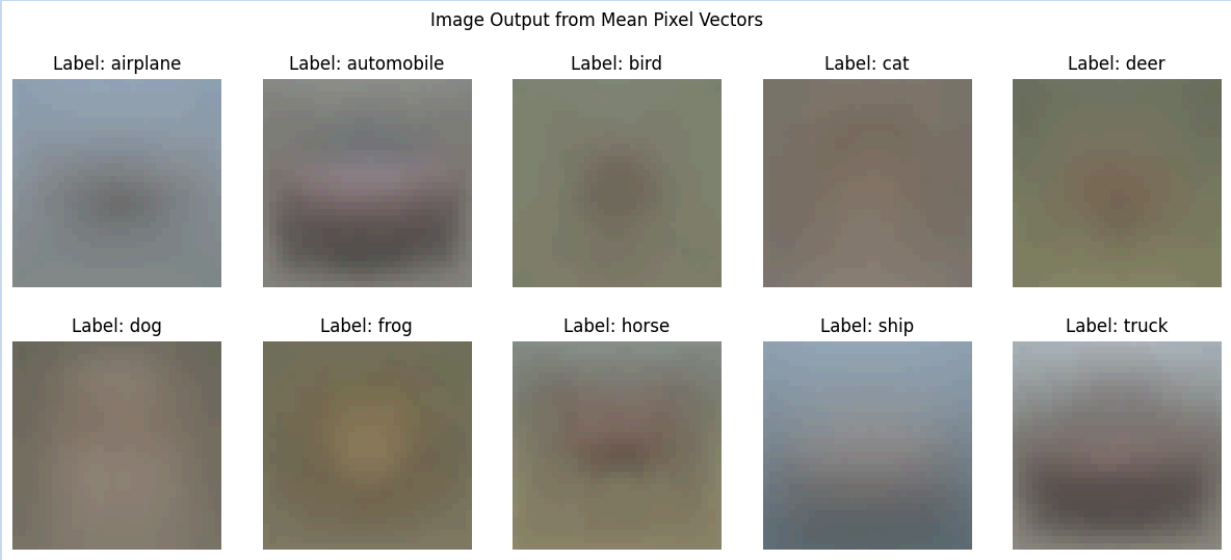
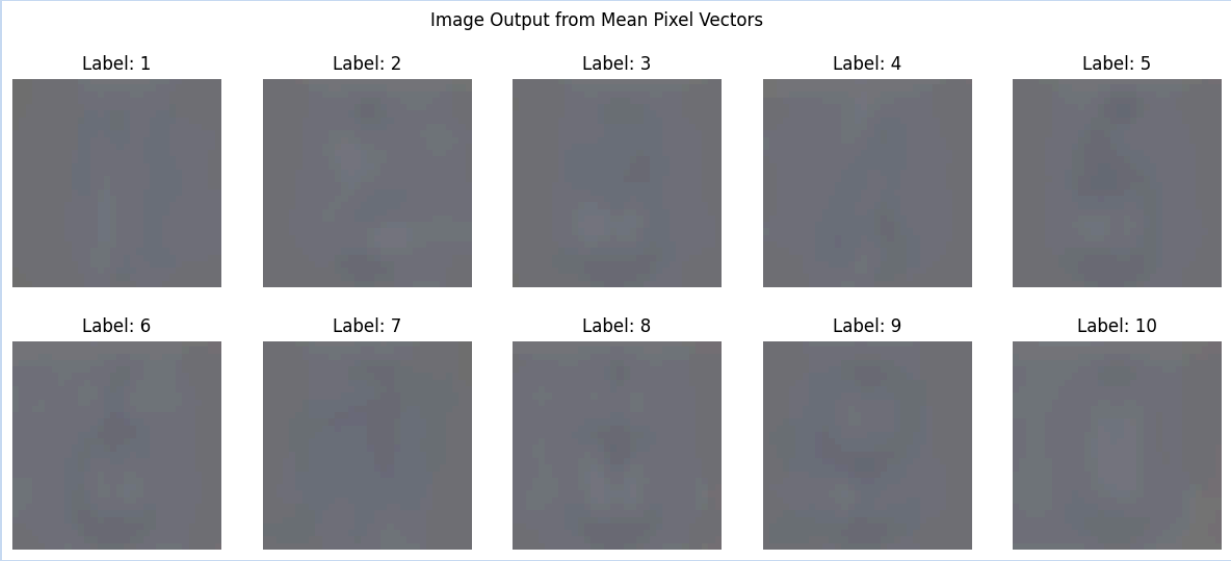


Image Output from Mean Pixel Vectors
Label: 0
Label: 1
Label: 2
Label: 3
Label: 4
Label: 5
Label: 6
Label: 7
Label: 8

For CIFAR10 this did not work as well. You can vaguely make out blurry images of objects that look like the class label — horse for example is almost perceivable — but it doesn't really mean much to a human. However, these images do give a good idea of the average background for each class.



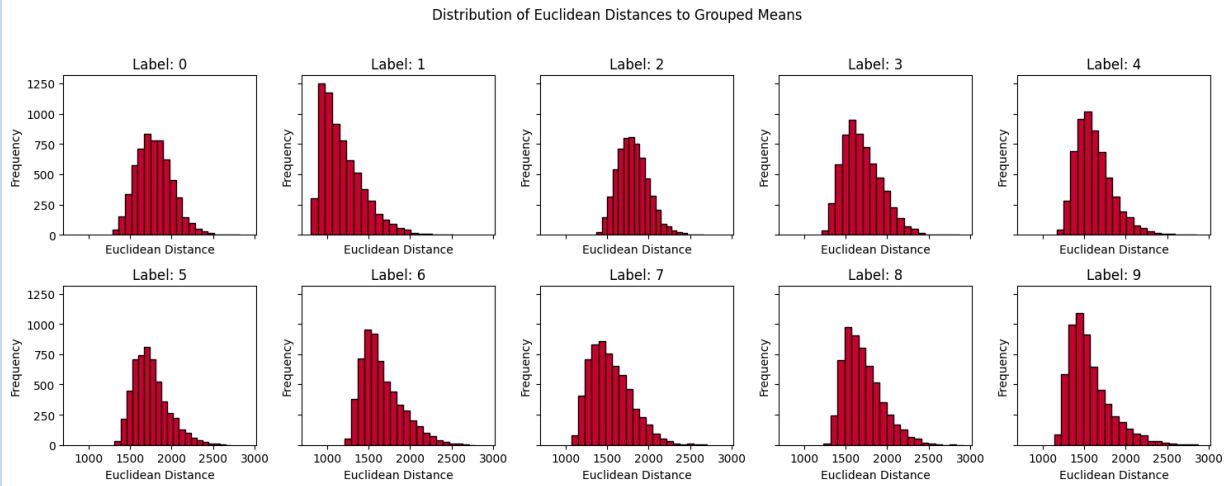
SVHN mean vectors were very "Grey" because of the lighting of the images and pixel intensity range being very small.



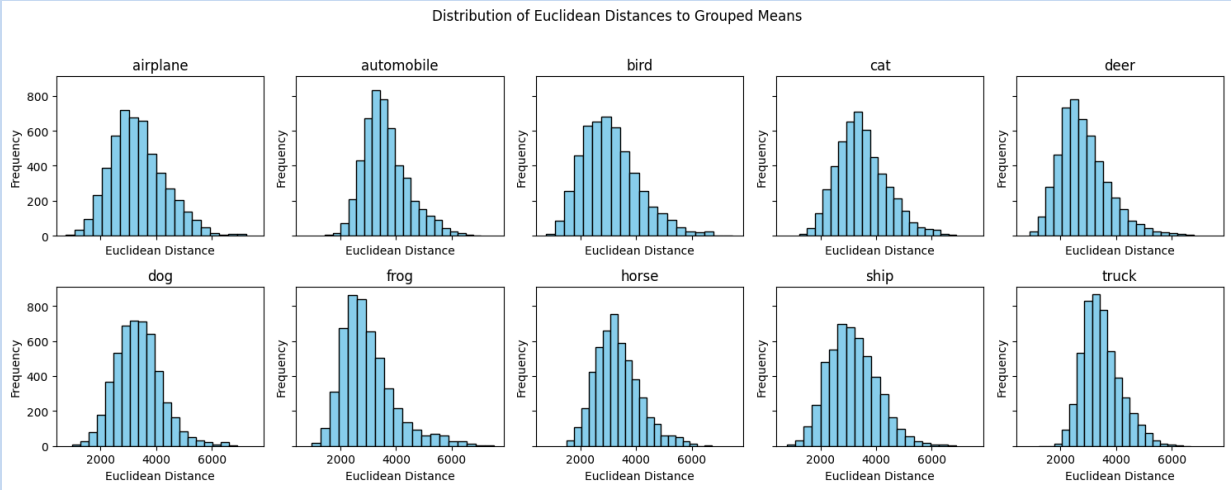
You can almost make out vague shapes. To make this more perceivable to humans, we can grayscale and binarize the image. The output is quite noisy but you can make out shapes that look similar to the class

label. 2, 4, and 5 are particularly well-shaped.

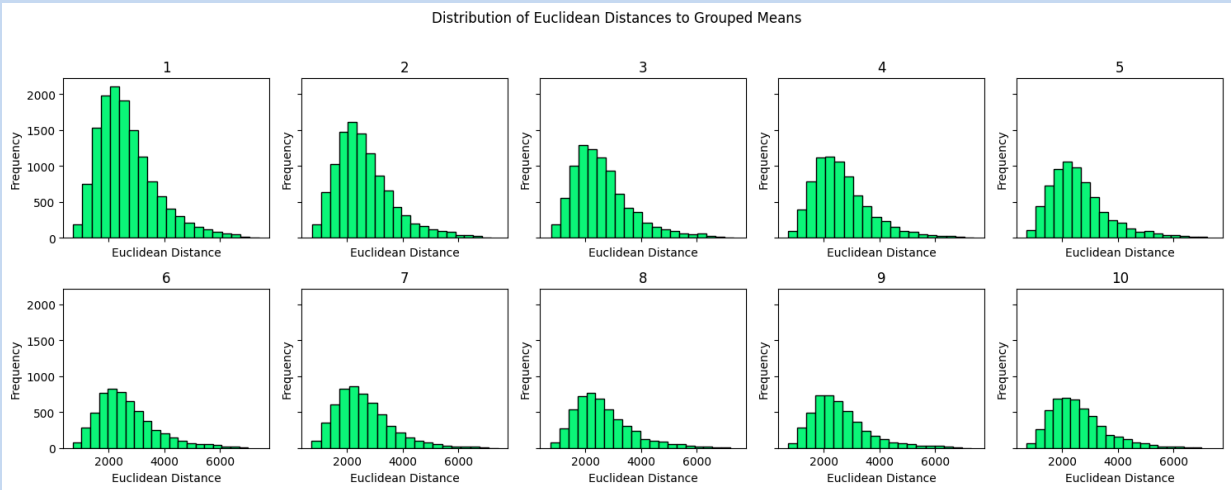
To check variation from records against the mean vector within a class, euclidean distance can be used to measure distance in the higher vector space. These distances can be plotted as histograms.
MNIST:



CIFAR10:



SVHN:



The distribution and variance of these histograms can tell us how much given individual images vary from the mean vector. More detailed summary statistics on these histograms were collected and saved to text files.

Form suppositions for future analysis

- Consider and evaluate information and findings in Section 2: Describe Data •

Form a hypothesis and identify actions

- [Optional] Transform the hypothesis into a data mining goal, if possible
- [Optional] Clarify data mining goals or make them more precise. A "blind" search is not necessarily useless, but a more directed search toward business objectives is preferable.
- Perform basic analysis to verify the hypothesis

We can hypothesize that image data is somehow correlated with class labels. Using Neural Network models we can verify this is true and create classifiers that accurately predict class labels for input images.

4. VERIFY DATA QUALITY

Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct or does it contain errors? If there are errors, how common are they? Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they? List the results of the data quality verification; if there are quality problems, list possible solutions. •

Identify special values and catalog their meaning

There are no special values

Review keys, attributes

- Check coverage (e.g., whether all possible values are represented)
- Check keys
- Verify that the meanings of attributes and contained values fit together
- Identify missing attributes and blank fields
- [Optional] Establish the meaning of missing data
- Check for attributes with different values that have similar meanings (e.g., low fat, diet) • Check spelling and format of values (e.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter)
- Check for deviations, and decide whether a deviation is "noise" or may indicate an interesting phenomenon
- Check for plausibility of values, (e.g., all fields having the same or nearly the same values)

All possible values are covered and represented. Ignoring this curated and historically well-researched dataset that has been cleaned and processed, we can also visually confirm that the images are not corrupted.

Review any attributes that give answers that conflict with common sense (e.g., teenagers with high income levels).
Use visualization plots, histograms, etc. to reveal inconsistencies in the data.
Data quality in flat files

- If data are stored in flat files, check which delimiter is used and whether it is used consistently within all attributes

- If data are stored in flat files, check the number of fields in each record to see if they coincide

In MNIST the delimiter is commas (CSV). CIFAR10 is stored as byte data which cannot be viewed without Python tools. SVHN has both raw images (less usable) and .mat tables which require specialized software or Python tools to view.

Post-loading and processing all datasets are stored in pandas dataframes.

Noise and inconsistencies between sources

- Check consistencies and redundancies between different sources
- Plan for dealing with noise
- Detect the type of noise and which attributes are affected

Of the three sources, the MNIST dataset is the most free of noise. The other datasets represent more realistic and less curated sources of image data but are more noisy. Noisy data in image data are pixels (parts of the image) that contradict the class label. For example, the background in CIFAR10 might present an indicator that the image is a class that it is not.

Remember that it may be necessary to exclude some data since they do not exhibit either positive or negative behavior (e.g., to check on customers' loan behavior, exclude all those who have never borrowed, do not finance a home mortgage, those whose mortgage is nearing maturity, etc.). Review whether assumptions are valid or not, given the current information on data and business knowledge.