- Now that we know more about the complexity of our data

knowledge about the application,

such as how the data was collected,

the user population,

the intended users of the application etc.

are important.

This domain knowledge

is essential to making informed decisions

on how to handle incomplete

or incorrect data.

You also need to be careful

about the changes you make

to avoid coming to incorrect conclusions

and be sure to keep records

of the changes you make.

The second part of preparing data

is to manipulate the clean data

into a format needed for analysis.

This step is known by many names,

data manipulation,

data pre-processing,

data wrangling

and probably my favorite, data munging.

Some operations for this data munging,

wrangling, pre-processing

includes scaling, transformation,

feature selection, dimensionality reduction

and data manipulation.

Let's look at these in further detail.

Scaling involves changing range of values

to be between a specified range

such as from zero to one.

This is done to avoid having certain features

with large values from dominating the results.

For example, in analyzing data with height and weight

the magnitude of the weight values

is much greater than the magnitude

of the height values.

So, scaling all values

to be between zero and one

will equalize contributions

from both height and weight features.

Various transformations can be performed on the data

to reduce noise and variability.

One such transformation is called aggregation.

Aggregate data generally results

in data with less variability

which may help with the analysis in the long term.

For example,

daily sales figures may have many spurious changes.

Aggregating values to weekly or monthly sales figures

will result in smoother data.

Other filtering techniques

can also be used to remove variability in the data.

Of course, this comes at the cost

of less detailed data,

so these factors must be weighed

for the specific application.

Feature selection can involve removing redundant

or irrelevant features,

combining features

and creating new features.

During the exploring data step,

you may have discovered that two features

after exploring with it,

we'll talk about pre-processing data

or transforming it to make it ready for analysis.

After this video,

you will be able to identify some problems

with real-world data,

and describe what is needed to transform raw data

to data that can be used for analysis.

The raw data that you get directly

from your sources

are never in the format that you need

to perform analysis on.

There are two main goals in the data pre-processing step.

The first is to clean the data

to address data quality issues.

The second is sort of transform the data

to make it suitable for analysis.

A very important part of data preparation

is to address quality issues in your data.

Real-world data is messy.

There are many examples of quality issues

with data from real applications,

including inconsistent late data,

like a customer with two different addresses

recorded at two different sales locations

but these recordings don't agree

or missing customer age

in demographic studies

or an invalid step code,

for example, a six-digit zip code,

and outliers like a sensor failure

that cause values to be much higher

or lower than expected for a period of time.

Since we get the data downstream,

we usually have little control over

how the data is collected.

Preventing data quality problems

as the data is being collected

is often not an option.

So, we have the data that we get

and we have to address quality issues

by detecting and correcting them.

Here are some approaches we can take

to address these data quality issues.

We can remove the data records with missing values,

we can merge duplicate records.

This would record a way to determine

how to resolve conflicting values.

Perhaps it makes sense to retain the nearer value

whenever there's a conflict.

For invalid values,

a best estimate for a reasonable value

can be used as a replacement.

For example,

a missing age value for an employee

can be filled in

based on a reasonable estimate

on the employee's length of employment.

Outliers can also be removed

if they are not important to the task.

In order to address

all these data quality issues effectively,

are very correlated.

In that case,

one of these features can be removed

without negatively affecting the analysis results.

For example, the purchase price of a product

and the amount of sales tax paid

are likely to be very correlated.

Eliminating the sales tax amount then will be beneficial.

Removing redundant or irrelevant features

will make the subsequent analysis simpler.

In other cases,

you may want to combine features or create new ones.

For example, adding applicants' education level

as a feature to a loan approval application

would make sense.

There are also algorithms to automatically determine

the most relevant features

based on various mathematical properties.

Dimensionality reduction

is useful when the dataset

has a large number of dimensions.

It involves finding a smaller subset of dimensions

that captures most of the variation in the data.

This reduces the dimensions of the data

while eliminating irrelevant features

and makes analysis simpler.

A technique commonly used for dimensionality reduction

is called principal component analysis.

Raw data often has to be manipulated

to be in the correct format for the analysis.

For example,

from samples recording daily changes in stock prices,

we may want to capture the price changes

for a particular market segment,

for example, real estate or healthcare.

This would require determining

which stocks belong to which market segment,

grouping them together,

and perhaps computing the mean, range and standard deviation

for each group.

In summary,

data preparation is a very important part

of the data science process.

In fact,

this is where you will spend most of your time

on any data science effort.

It can be a tedious process

but it is a crucial step.

Always remember,

when it comes to data processing,

garbage in is garbage out.

If you do not spend the time and effort

to create good data for the analysis,

you will not get good results

no matter how sophisticated

your data analysis techniques are.