THINKFUL

2021

# Data Science Fundamentals: The Pandas Library

Information and practical exercises to add to your current toolkit or take the first step in launching a new career.

# Welcome to Thinkful!

**We teach tech skills that lead to fulfilling, high-paying careers.**

Our students learn **in-demand** industry tools through **100% online programs** as they work toward a **job-ready portfolio** with the help of an **expert mentor**.

## Let's get started.

# Workshop Rundown

**We're going to talk about:**

- ❏ What is Data Science?
- ❏ Why pandas?
- ❏ What is a pandas DataFrame?
- ❏ Pandas data manipulation
- ❏ Further resources

# What is Data Science?

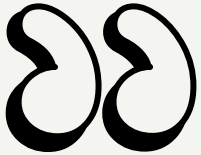
**Data Science combines several disciplines from Math, Statistics, Computer Science, and Communications to solve problems and make predictions.**

Some specific steps in that process:

- ❏ Data Wrangling
- ❏ Analytics
- ❏ Predictions

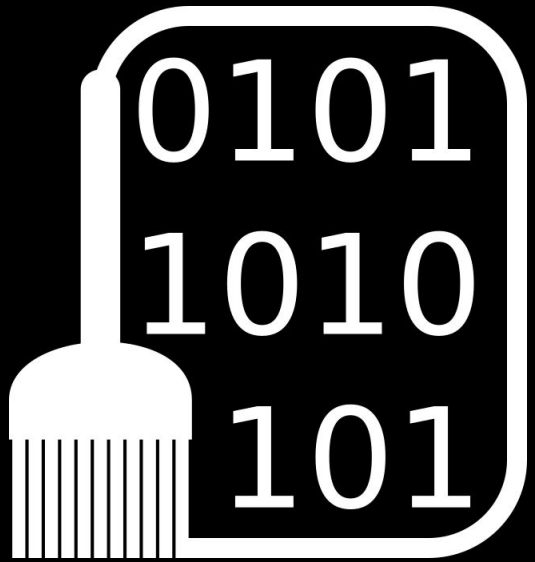
# The Less Sexy Side of Data Science

You've got to know where the value is with your data, and to trust it before you put it in your model. The best tech out there for cleaning data is to hammer through it yourself...



**Scott Nicholson**
Chief Data Scientist
Accretive Health

# Data Cleaning Tasks



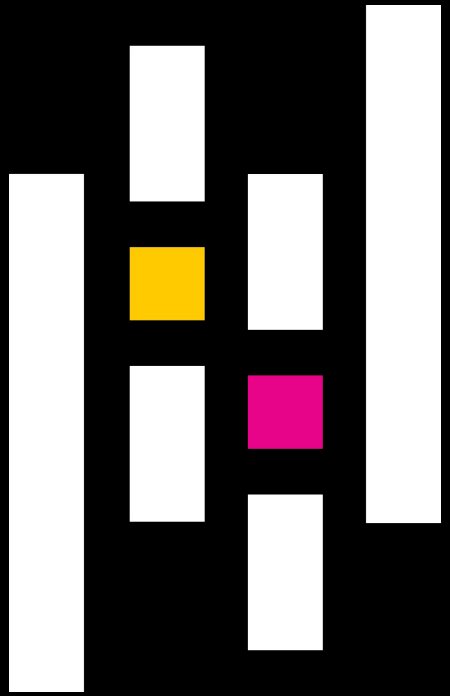

## What do we need to clean?

- ❑ Fill in missing values
- ❑ Identify outliers and reduce noise
- ❑ Correct inconsistent data
- ❑ Resolve redundancies

# Pandas

**Pandas is a library for data analysis**

- ❑ Data cleaning
- ❑ Data manipulation
- ❑ Data analysis
- ❑ Works well with large data sets



# Pandas DataFrame

rows characters

| | first_name | last_name | houses_owned | castles_owned | prefered_temp |
|---|---|---|---|---|---|
| 0 | Robert | Baratheon | 68 | 9 | 72 |
| 1 | Tywin | Lannister | 43 | 5 | 68 |
| 2 | Eddard | Stark | 23 | 3 | 35 |

columns

index

# Starter Code

**Starter Code**
bit.ly/colab_pandas

**We'll be using a Google-hosted Python notebook called Colaboratory**

- ❑ Click **File**
- ❑ Select **Save a Copy in Drive**
- ❑ This is your personal version of the notebook – let's get started!

# Configure Our Environment

## Importing Libraries

```
# Data analysis packages
import numpy as np
import pandas as pd

# Default settings for pandas
pd.set_option('mode.chained_assignment', None)
pd.set_option('display.float_format', '{:,.2f}', None)
```

## Importing Data

```
# Import data from github
data = pd.read_csv('url_to_csv_file')
```

# Clean The Data

## Rename Columns

```
# Rename columns to include units
data.rename(columns = {'duration' :'duration_mins',
                       'budget':'budget_usd',
                       'gross':'gross_usd'},
            inplace = True)
```

## Removing Duplicates

```
# Checking to see what movies are duplicates.
# Sorting by movie title to see duplicates
data[
      data.duplicated(
            subset=['movie_title','title_year'],
            keep=Bool)
      ].sort_values('movie_title').head()
```

# Clean The Data (Continued)

## Missing Values

```
# Show how many values are missing from each column
data.isna().sum()
```

```
# Drop the aspect ratio column (axis = 1)
data.drop('aspect_ratio', axis = 1, inplace = True)
```

```
# Drop the row (axis = 0) for the remaining missing value
data.dropna(subset = 'budget_usd', axis = 0, inplace = True)
```

# Exploratory Data Analysis (EDA)

```
# Identify all of the unique countries
data['country'].unique()
```

```
# Get a count for all of the values for each country
data['country'].value_counts()
```

```
# Select just the country names of the three largest
counts.nlargest().index
```

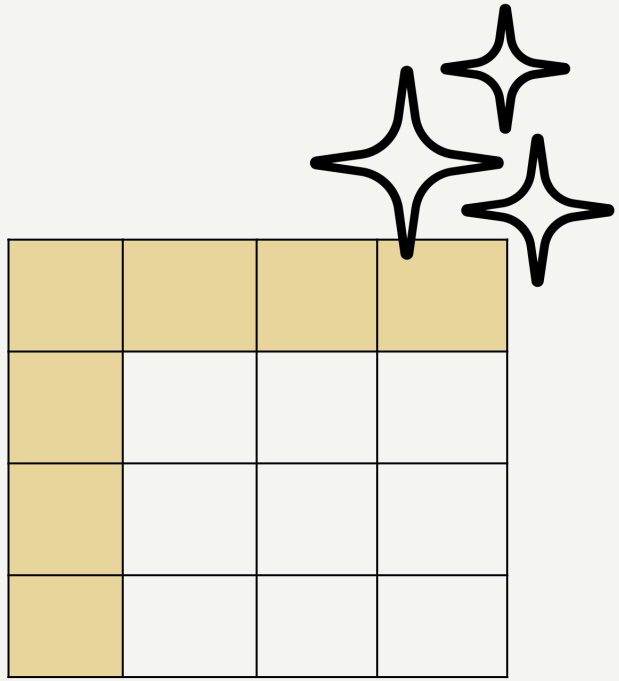
# Feature Engineering

```
# Determine the 'actor_1_name' with the most movies in the set
data['actor_1_name'].value_counts().head
```

```
# Encoding ratings as dummy variables
content_ratings = pd.get_dummies(data['content_rating'])
content_ratings.head(2)
```

```
# Select columns by data type - number
numerical_data = data.select_dtypes(include = 'number')
numerical_data.head()
```

# The Importance Of Clean Data



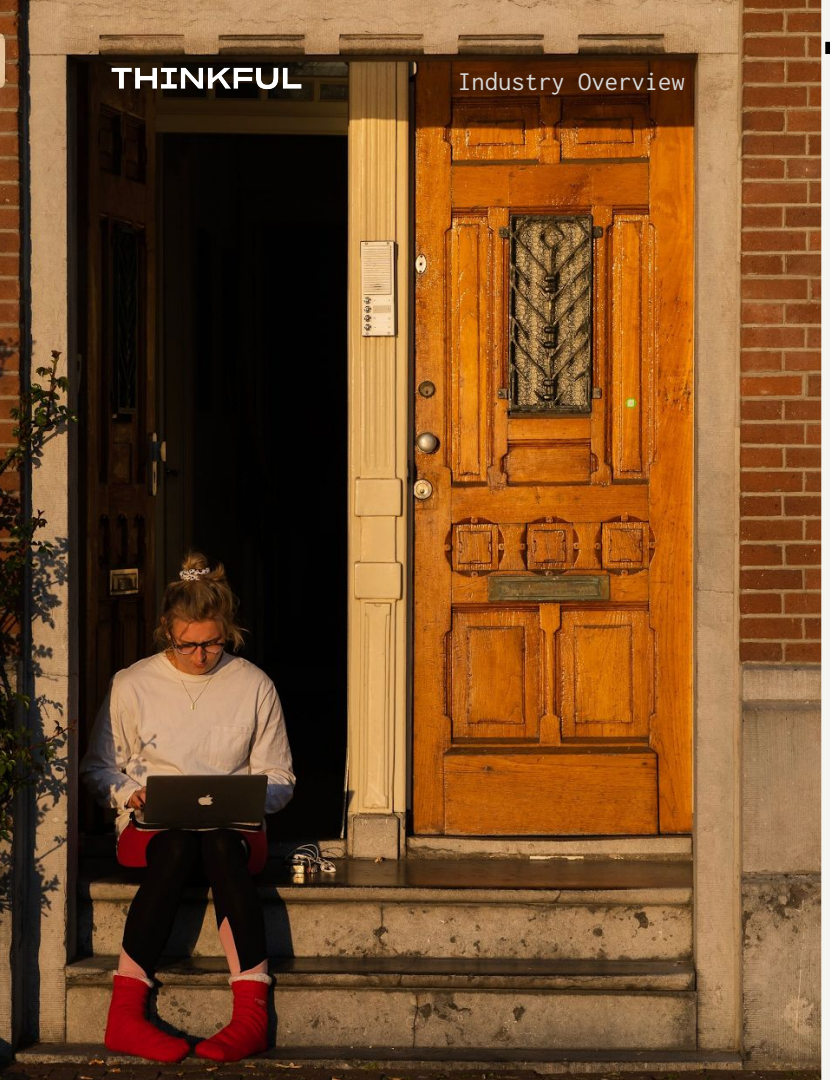
**Random Forest Model**
**$R^2$ score (1.0 = perfect)**

- ❏ Before cleaning > ERROR
- ❏ After cleaning > 0.6118
- ❏ After feature engineering > 0.6332

# Common Questions

## You might also be wondering

- ❏ What are the outcomes of your students for this field?
- ❏ How do I show my work to a potential employer?
- ❏ Is this course entirely online?
- ❏ What should I do from here?






# Take the First Step to A New Career

Anyone who's driven to change their future and achieve a high-earning career is able to enter the world's next workforce. We'll be by your side as you build the skills you need, with personal mentorship and an active, online community of students and educators.

Expand your career opportunities by breaking into tech. Chat with an admissions rep and we'll help you find the perfect fit.

Schedule a Call