Cleaning Data: Python Data Playbook

UNDERSTANDING YOUR DATA

Chris Achard

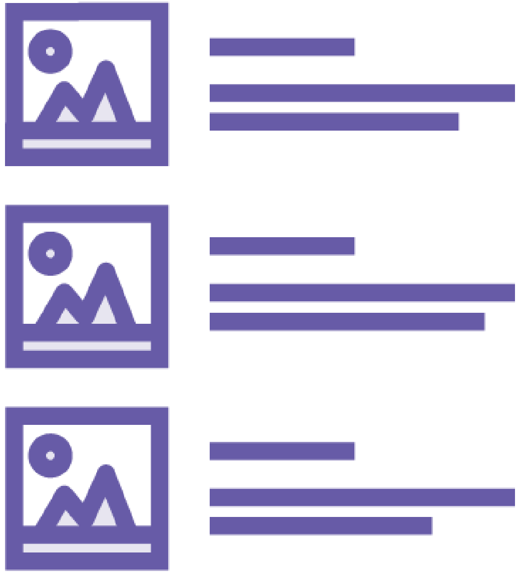
@nanohopdev www.nanohop.com

Overview

Understanding Your Data

- Understand and convert types
- Aggregate data
- Normalize data
- Transform data
- Filter data

Tate Gallery Artwork Dataset



https://www.tate.org.uk/about-us/digital/collection-data

> 69,000 pieces of artwork

Sample data set of 10 pieces

Demo

Understand the types of data we have

View it in aggregate, and in summary

Transform and filter

Know what we have, and how to limit and change it

Viewing and Converting Types

Aggregating Data

Normalizing Data

Transforming Data

Filtering Data

Review

```
df.dtypes
```

◄ View the data types of columns

```
df.year.astype(float)
```

◄ Convert column to new type

```
df.year = df.year.astype(float)
```

◄ Assign converted type back to dataframe

```
pd.to_numeric(df.height, errors=“coerce”)
```

◄ Coerce errors

```
df.year.min()
```

- Call functions on series

```
df.agg([‘min’, ‘max’])
```

- Use .agg to call multiple functions on the dataframe

```
(c – c.mean()) / c.std()
```

◀ Standardize around 0

```
(c – c.min()) / (c.max() – c.min())
```

◀ Normalize between 0 and 1

```
df[‘normalized_column’] = new_values
```

◀ Assign normalized values back to dataframe as a new column

```
df.height.transform(lambda x: x / 10)
```

◄ Transform a single column

```
df.groupby(’artist’).transform(‘nunique’)
```

◄ View data summary by group

```
df.groupby(‘artist’)[‘height’]
      .transform(‘mean’)
```

◄ Transform a single column grouped by another column

`df.filter(items=[‘id’, ‘artist’])`	◀ View only certain columns
`df.filter(regex=“(?i)year”)`	◀ View columns that match a regex
`df.filter(axis=0, like=’100’, case=False)`	◀ Switch the axis to filter rows