DataDive Week 3: Data Preparation

SheffDataSoc

Data Science Society

Datasoc Promo!

Datasoc is hosting a **career panel** session **Next Wednesday** (4/3) at **3pm** (**Diamond LT9**)! Come along to have the chance to hear from speakers from industry, and get great insights into data science work!

Check out Datasoc on instagram (**@sheffdatasoc**) for updates and information on our events!

Tell your friends and classmates about the Data Dive! It's a fun way to work on a project, get experience for your CV, and win actual prizes!

Data Dive Information

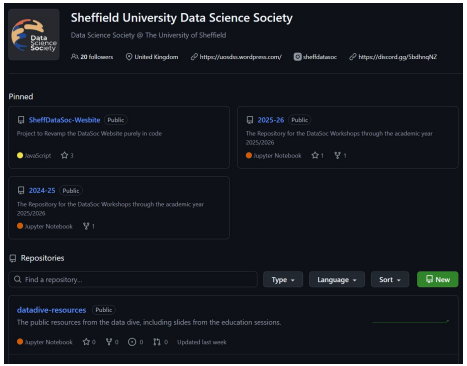
Project Lead Sign-Up information will come at the end of this session, so stick around if this interests you!

It might be a good idea to start thinking about what sort of topic you may want to tackle in the Data Dive, or who you may want to form a team with.

You can access the Data Dive resources at this link:

https://github.com/sheffdatasoc/datadive-resources

It may help to star the repository for future access! ☆



Session Outline

- Why Do We Need To Prepare Our Data?
- How To Spot Issues With The Data
- Dealing With Missing Data: Different Approaches
- Data Cleaning: Numerical Data
- Data Cleaning: Textual Data
- Project Lead Information!

Access the resources for today's session:

https://github.com/sheffdatasoc/datadive-resources

Why should we prepare our data?

Datasets often contain missing data, inaccurate data, or data that is not fit for analysis or machine learning. This may be an issue, even if the source of the data is trustworthy.

“**Garbage in, garbage out**” - using inaccurate data will lead to inaccurate results!

Machine learning also generally requires data to be numeric, if the dataset contains other forms of data it will have to be changed.

Messy Data - Example

color	director_name	duration	gross	movie_title	anguage	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall StreetÂ	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3Â	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful EightÂ	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	MargaretÂ	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of SmaugÂ	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of JusticeÂ	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected JourneyÂ	English	USA	180000000	2012	7.9
Color	Edward Hall	180		RestlessÂ	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud AtlasA	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon TattooÂ	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of GodÂ	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Battle of the Five ArmiesÂ	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6

source:
https://medium.com/well-red/cleaning-a-messy-dataset-using-python-7d7ab0bf199b

Messy Data - Example

What issues can you spot with this data?

	IMBD title ID	Original titlÊ	Release year	Genrë¨	Duration	Country	Content Rating	Director	Unnamed: 8	Income	Votes	Score
12	tt0060196	Il buono, il brutto, il cattivo	23rd December of 1966	Western	161	Italy	Approved	Sergio Leone	NaN	$ 25252481	672.499	8.8
39	tt0110357	The Lion King	1994-11-25	Animation, Adventure, Drama	88	USA	G	Roger Allers, Rob Minkoff	NaN	$ 968511805	917.248	8.4
31	tt0172495	Gladiator	2000-05-19	Action, Adventure, Drama	155	USA	R	Ridley Scott	NaN	$ 465361176	1.308.193	8.5
13	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
62	tt0047396	Rear Window	1955-04-14	Mystery, Thriller	112	USA	NaN	Alfred Hitchcock	NaN	$ 37032034	432.390	8.1
80	tt0180093	Requiem for a Dream	2000-12-15	Drama	102	USA	R	Darren Aronofsky	NaN	$ 7390108	748.291	7.8
100	tt0045152	Singin' in the Rain	1953-02-05	Comedy, Musical, Romance	103	USA	NaN	Stanley Donen	NaN	$ 1864182	213.152	7.4
28	tt6751668	Gisaengchung	2019-11-07	Comedy, Drama, Thriller	132	South Korea	NaN	Bong Joon Ho	NaN	$ 257604912	470.931	8.6
17	tt0099685	Goodfellas	1990-09-20	Biography, Crime, Drama	146	USA	R	Martin Scorsese	NaN	$ 46879633	991.505	8.7
37	tt0103064	Terminator 2: Judgment Day	1991-12-19	Action, Sci-Fi	137	USA	R	James Cameron	NaN	$ 520884847	974.970	8.4
73	tt0112573	Braveheart	1995-12-01	Biography, Drama, History	178	USA	R	Mel Gibson	NaN	$ 213216216	941.683	7.9
75	tt0105236	Reservoir Dogs	1992-10-09	Crime, Drama, Thriller	99	USA	R	Quentin Tarantino	NaN	$ 2889963	896.551	7.9
42	tt0253474	The Pianist	2002-10-25	Biography, Drama, Music	150	UK	R	Roman Polanski	NaN	$ 120072577	707.942	8.4
41	tt1675434	Intouchables	2012-02-24	Biography, Comedy, Drama	112	France	NaN	Olivier Nakache, Ã□ric Toledano	NaN	$ 426588510	736.691	8.4
94	tt2106476	Jagten	2012-11-22	Drama	115	Denmark	R	Thomas Vinterberg	NaN	$ 15843274	269.616	7.5

source: https://www.kaggle.com/datasets/davidfuenteherraiz/messy-imdb-dataset/data

Activity - Identifying Data Issues

Use the dataset and notebook on GitHub, and import and explore the data to see if you can spot any patterns or issues!

(**Note**: this is a modified version of the AirBnB dataset from last week)

```
1 import pandas as pd
```

```
1 df = pd.read_csv("[redacted].csv")
2 df.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	lat
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.6
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.7
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.8
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.6

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48896 entries, 0 to 48895
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Unnamed: 0                      48896 non-null  int64
 1   id                              48896 non-null  int64
 2   name                            48880 non-null  object
 3   host_id                         48896 non-null  int64
 4   host_name                       48875 non-null  object
 5   neighbourhood_group             48896 non-null  object
 6   neighbourhood                   48896 non-null  object
 7   latitude                        48896 non-null  float64
 8   longitude                       48896 non-null  float64
 9   room_type                       48588 non-null  object
 10  price                           48896 non-null  int64
 11  minimum_nights                  48896 non-null  int64
 12  number_of_reviews               48896 non-null  int64
 13  last_review                     38843 non-null  object
 14  reviews_per_month               38843 non-null  float64
 15  calculated_host_listings_count  48896 non-null  int64
 16  availability_365                48522 non-null  float64
dtypes: float64(4), int64(7), object(6)
memory usage: 6.3+ MB
```

Missing Data

This is a common and crucial issue to tackle with any datasets being used.

Analysis and visualisations will be misguided if there is data missing, and machine learning algorithms will not be able to work.

It is also important to consider *why* the data is missing. It might be:

- Missing Completely At Random (MCAR), there is no pattern to the data being missing
- Missing At Random (MAR), the data is missing in relation to other features
- Missing Not At Random (MNAR), the data is missing in relation to the missing feature itself

Missing Data - Implicit vs Explicit

It is worth noting that some missing data might not initially be obvious.

Using isna().sum() will identify NaN/None missing values...

	x	y
0	1.0	NaN
1	2.0	2.0
2	NaN	NaN
3	4.0	4.0

```
temp.isna().sum()

x    1
y    2
dtype: int64
```

	x	y
0	1	NaN
1	2	2.0
2	,	3.0
3	4	4.0

	0
x	0
y	1

```
temp.dtypes

0
x    object
y    float64
```

...however other data may be 'missing' by being incorrect, replaced with 0s, etc. A way of checking this could be by examining feature datatypes with .dtypes

Missing Data - Different Approaches

There are multiple ways to deal with missing data, with pros and cons:

- Remove the row/datapoint (if there are few rows to remove, and the missingness is random)
- Remove the feature (if it has a significant number of missing values)
- Fill with a constant value
- Imputation with mean/median/mode
- Imputation with an algorithm (such as KNN)

Activity: On the worksheet, see if you can find any missing values in the dataset. How would you go about fixing this?

Missing Data - Additional tips and tricks

Avoid deleting too much data (Potentially ruin the data population)

Avoid deleting variables of high importance

Try to limit imputation to avoid reliance on weak assumptions

If a column has significant amounts of missing data, consider removing it entirely!

At the end of the day context, domain, data pattern and methods used are the most important factors to consider as well when it comes to deletion and imputation

Data Cleaning - Numeric Data

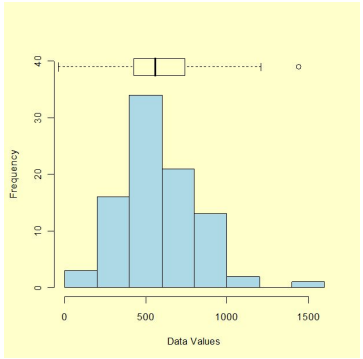
Another data issue to be considered is outliers.

These can normally be spotted by plotting the feature in a histogram, boxplot or scatter plot.

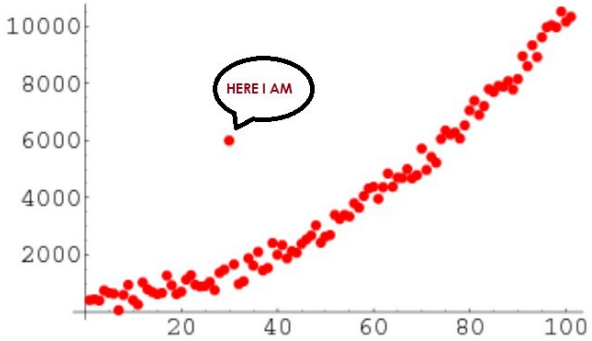
Consider where the outlier may have come from, it may be a data reporting mistake, or a genuine extreme value.

Outliers can also be handled using a handful of approaches:

- Remove the row/datapoint
- Clip values using:
  - **IQR Method** (eg 1.5 IQRs below Q1 and above Q3)
  - **Z-score Method** (3 standard deviations below/above mean)



source: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm



source: https://medium.com/@aa.adwan72/understanding-outliers-strategies-for-addressing-unusual-data-points-c66072199ffa

Data Cleaning - Numeric Data

For preparation for machine learning models, it is often useful (or even necessary) to scale numerical features. This can improve algorithmic efficiency, and ensure features with massively different scales can be equally important.

The typical methods of scaling are:

- **Normalisation** / Min-Max Scaling (making all values fall within the range [0, 1]
- **Standardisation** / Z-score Scaling (transforming the data to fit the standard normal distribution, with a mean of 0 and standard deviation of 1)

When scaling, consider the distribution of the feature, and whether there are outliers. This may inform the technique you use!

$$x^{(i)}_{std} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

$$x^{(i)}_{norm} = \frac{x^{(i)} - \mathbf{x}_{min}}{\mathbf{x}_{max} - \mathbf{x}_{min}}$$

standardization

min-max scaling ("normalization")

	input	standardized	normalized
0	0	-1.46385	0.0
1	1	-0.87831	0.2
2	2	-0.29277	0.4
3	3	0.29277	0.6
4	4	0.87831	0.8
5	5	1.46385	1.0

source: https://levelup.gitconnected.com/mastering-data-normalization-and-standardization-a-practical-guide-c6a707a042a2

Data Cleaning - Textual Data

Textual data features could include things like names, addresses, regions (as seen in the AirBnB dataset).

This data is often messy, as it may be the result of user input and poor integrity checks.

Common issues that need to be check for and fixed are:

- **Case** (ensuring all values are upper/lower/title case)
- **Punctuation** & **Whitespace** (avoiding unnecessary punctuation and leading/trailing spaces)
- **Spelling**

	A	B	C
0	john	masters	27
1	bODAY	graduate	23
2	minA	graduate	21
3	Peter	Masters	23
4	nicky	Graduate	24

source: https://www.geeksforgeeks.org/python/capitalize-first-letter-of-a-column-in-pandas-dataframe/

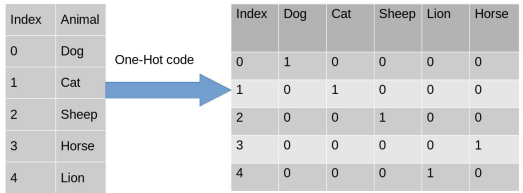
If you want to perform things like sentiment analysis, you will have to perform further steps such as tokenisation (breaking text into words), lemmatisation (converting words to their base meaning), etc...

Data Cleaning - Categorical Data

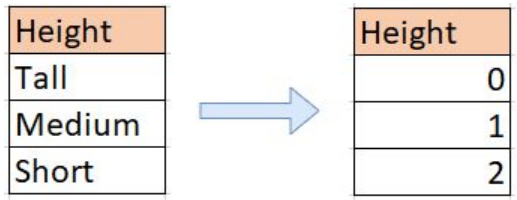
Machine learning models cannot interpret text! In order to use this data, it must be encoded.

When dealing with categorical data, this is generally done in two ways:

- **One-Hot Encoding**: binary columns are created for each category (this is best for *nominal* data, where there is no clear order of categories)
- **Label Encoding**: category values are assigned integer values (this is best for *ordinal* data, where the categories have a clear order)



source: https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

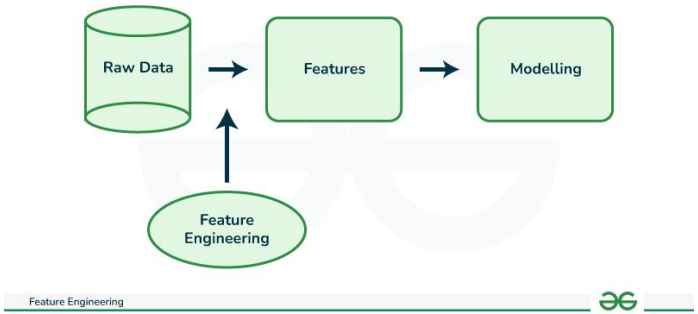


source: https://ai-ml-analytics.com/categorical-encoding-label-encoding/

Data Cleaning - Feature Engineering

What is Feature Engineering

- Using variables you have to create a new feature
- Can be in the form of aggregation/derivation
- Context and Correlation are important in this step!



source:
https://www.geeksforgeeks.org/machine-learning/what-is-feature-engineering/

Activity - Clean And Prepare Your Data!

Using the notebook, have a go at cleaning and preparing the dataset using some of the techniques provided.

If you are stuck, it may be useful to look at pandas and scikit-learn documentation, or check out Datasoc's guides!

https://www.sheffdatasoc.org/guides

If you haven't accessed them yet, you can find the resources for the session here:

https://github.com/sheffdatasoc/datadive-resources

Project Lead Sign-Up!

Are you interested in becoming a **Project Lead**?

This will involve being the first point of contact for a team, advising them and working through any issues they are encountering with their project. It will be your responsibility to check that the team is staying on track, and you will have access to their project's GitHub repository and their work to see how they are getting on.

This is an additional role, you can be in a team *and* be PL for other team(s).

Sign up below! (closes in 1 week)

https://tinyurl.com/datadiveprojectlead

What comes next?

Thank you for coming! Make sure to spread the word and get your friends involved in the Data Dive!

Join us at **3pm** next week for the career panel!

Sign-up below if you are considering becoming a project lead:

https://tinyurl.com/datadiveprojectlead

Week	Session	Where/When
4	Creating ML Models	**Wednesday 3pm (for Career Panel) Diamond LT9**
5	Evaluating & Interpreting ML Results	**Wednesday, 4-6pm Diamond LT9**