UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Preparation of Data

Transformation of Variables: Normal – Type Conversion

Transformation of Variables: Outliers

...

Converting variables

- Please watch the video through this link

https://media.ed.ac.uk/media/Converting+Variables/1_yjgpuhwd/117185481

- Study the following file
  11 - numeric_and_categorical_variable_conversions_reading.ipynb

Activity: Converting variables

Transformation of Variables: Normal – Type Conversion

- This transformation is necessary because models constrain the types of variables that can be used.
- **Standardisation**:
- for each observation $x_i, i = 1 \dots n$, we do

$$z_i = \frac{x_i - \mu}{\sigma}$$

Where μ is the estimated mean and σ is the estimated std.

- The result is dimensionless and can be compared among variables.
- It has a mean of 0 and a standard deviation of 1 while retains the shape of the variable distribution
- In practice, it is used as a pre-processing step for almost all modelling approaches, especially neural networks and principle component analysis.

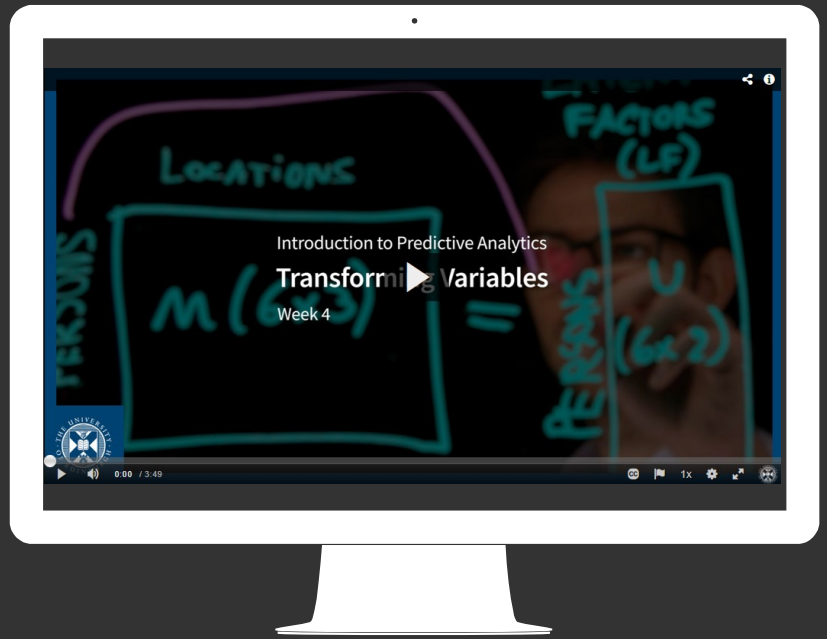
Transformation of Variables: Normal – Type Conversion

- Min-max scaling (feature scaling, or normalisation)
- for each observation $x_i, i = 1 \dots n$, we do

$$x_{i,norm} = \frac{x_i - min(\mathbf{x})}{max(\mathbf{x}) - min(\mathbf{x})}$$

- It rescales the whole variable to values between 0 and 1 while retains the shape of the variable distribution.
- Sensitive to outliers (note that standardization does not have this issue)

Transforming variables



- Please watch the video through this link

https://media.ed.ac.uk/media/Transforming+Variables/1_eyj9w37g/117185481

Quiz

In what case would you convert a numeric variable to a different type?

- A. To convert a regression into a classification.
- B. To convert a regression into a time series analysis.
- C. To convert a classification into a regression.
- D. None of the above

Quiz

When is standardisation preferred over min-max scaling?

- A. When there are a lot of negative values in the dataset.
- B. When there are no negative values in the dataset.
- C. When outliers are present.
- D. When it performs similarly to min-max scaling.

Quiz

Select the best TWO options for completing the following sentence:

“Normalisation is required because…”

- A. Predictive algorithms require particular input ranges.
- B. It can transform continuous variables into a normal distribution.
- C. It can transform categorical variables into continuous ones.
- D. Variables have different dimensions.

Quiz (optional)

Select the best TWO options for completing the following sentence:

“When converting the dependent categorical value of a dataset,...”

- A. We can obtain regression
- B. We should use dummies
- C. We should use label encoding
- D. We can do time series analysis

Preparation of Data

Transformation of Variables: Normal - Type Conversion

Transformation of Variables: Outliers

...

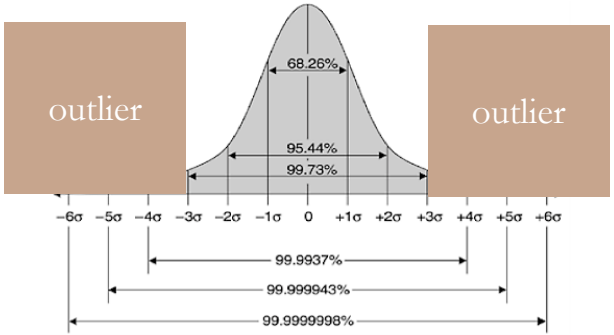
Dealing with outliers in a dataset

- An outliers is an observations that differs significantly from other observations.
  - They typically lies outside the boundaries of majority of observations
  - They infer that being generated by a different underlying system.
  - They are sometimes, but not always, made up of extreme values.
- Example 1: if a customer typically buys 1 ticket for a movie, or 20 when doing group tours, then buying 10 tickets is the outlier, as 1 and 20 are the most common values.
- Example 2: if a person wants to submit a school loan application for 1 billion pounds, we might get suspicious about whether this value reflects reality.

Dealing with outliers – how to detect 1

- Extreme values are typically found by making use of probability distributions.
- Sufficiently large dataset: we can use the standard normal distribution (after we normalised our data) and see how far away our observation is from the mean.
  - 3-sigma rule (Rule of thumb) : | observation - μ | > 3σ are considered outliers.
- Small dataset:

we can use the student t-distribution to find outliers instead.



Dealing with outliers – how to detect 2

- Multivariate outliers: use a multivariate Gaussian distribution
- The parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$ multivariate Gaussian model can be estimated on the dataset given n p-dimensional observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$ where $x_i \in R^p$, parameter fitting:

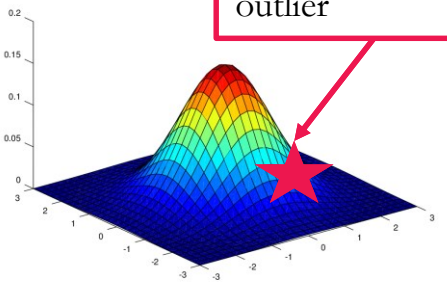
$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \,, \Sigma = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

- Then calculate the likelihood of a certain observation $\boldsymbol{x}'$ given the above Gaussian distribution,

$$P(\boldsymbol{x}'; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\,(\boldsymbol{x}' - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}' - \boldsymbol{\mu})\right)$$

Flag an outlier if $P(\boldsymbol{x}'; \boldsymbol{\mu}, \Sigma) < \epsilon$.

The probability of having an obs here is very low ➔ outlier

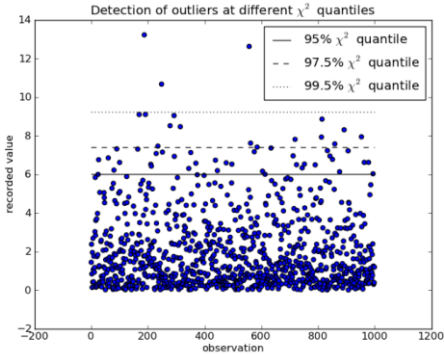


Dealing with outliers – how to detect 3

- Multivariate outliers: **The Mahalanobis distance**
- We can calculate the Mahalanobis distance of obs $\boldsymbol{x}'$ to the mean (vector) $\boldsymbol{\mu}$

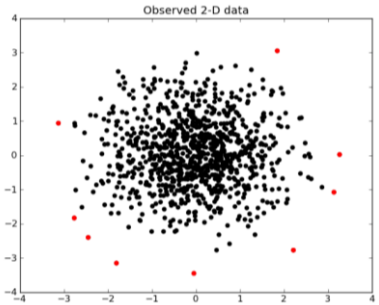
$$d_M(\boldsymbol{x}', \boldsymbol{\mu}) = \sqrt{(\boldsymbol{x}' - \boldsymbol{\mu})^T \cdot \Sigma^{-1} \cdot (\boldsymbol{x}' - \boldsymbol{\mu})}$$

d_M follows a χ^2-distribution. An outlier is flagged if $d_M(\boldsymbol{x}', \boldsymbol{\mu}) > \chi^2_{0.95}(p)$

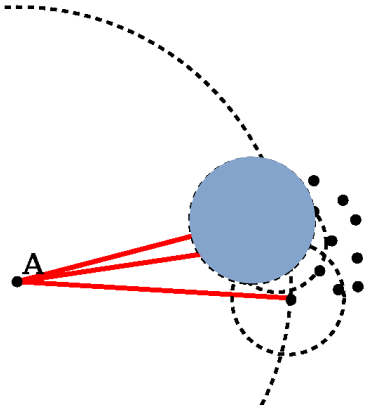


- The Mahalanobis distance is unitless and scale-invariant it considers correlation.
- In case the uncorrelation across dimensions (Σ is diagonal), the Mahalanobis coincides with standardized Euclidean distance, since

$$d(\boldsymbol{x}', \boldsymbol{\mu}) = \sqrt{\sum_{i=1}^{p} \frac{(x_i' - \mu_i)^2}{s_i^2}}$$, where s_i^2 is the sample std of X_i.



Dealing with outliers – how to detect 4.1



The Local Outlier Factor (LOF) is a density-based method that relies on nearest neighbors search. LOF has an already existing function in scikit-learn: `neighbors.LocalOutlierFactor`

Suppose $N_k(\mathrm{A})$ is the set of neighbors of point A, k is the number of points in this set, and $d(\mathrm{A},\mathrm{B})$ is the distance between points A and B.

- The reachability distance of an object A from B is defined as

$$\text{reachability}-\text{distance}_K(A,B) = \max\{d(A,B), \text{k}-\text{distance}(B)\}$$

Where $\text{k}-\text{distance}(B)$ be the distance of the object B to its k-th nearest neighbor.

- The estimated **local reachability density** of an object A is defined as:

$$LRD_A = \frac{1}{\frac{1}{k}\sum_{B\in N_k(\mathrm{A})} \text{reachability}-\text{distance}_K(A,B)}$$

Dealing with outliers – how to detect 4.2

Then the local outlier factor score is calculated as

$$LOF_A = \frac{\frac{1}{k}\sum_{B \in N_k(\mathrm{A})} LRD_B}{LRD_A}$$

which is the average local reachability density of the neighbors divided by the object's own local reachability density.

- $LOF_A < 1$ means Higher density than neighbors ➔ Inlier;
- $LOF_A \sim 1$ means Similar density as neighbors;
- $LOF_A > 1$ means Lower density than neighbors ➔ Outlier

The original paper is given by http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.446.9951&rep=rep1&type=pdf

Dealing with outliers

- Note that typically, outliers are firstly removed before applying any normalization, as they tend to drastically influence those operations.
- Although the above outlier detection methods, you still want to be cautious about the identification – maybe they are informative or belong to an important group with particular characteristics.

- Better to double-check whether the identified outliers are indeed unwanted.

- Study the following file
  12 - normalising_variables_and_detecting_outliers_reading.ipynb

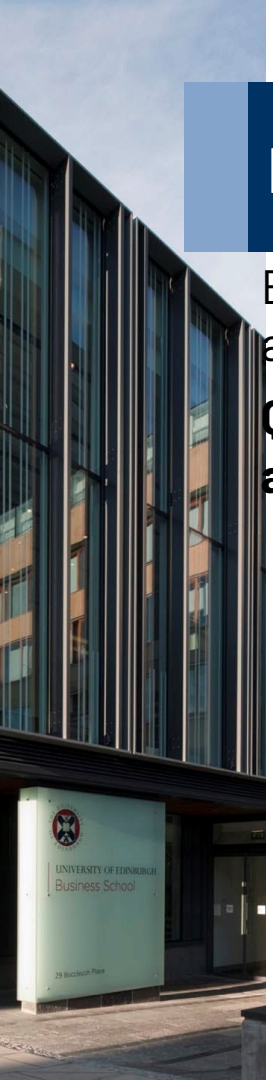
Activity: Normalising variables and detecting outliers

Homework

Based on '13 - Assessment_can_you_spot_the_fraudsters.ipynb' and dataset 'credit.csv', answer the following question:

Q1. Which of the two variables did you remove from the dataset?

CC, NO, CITY, PHONE

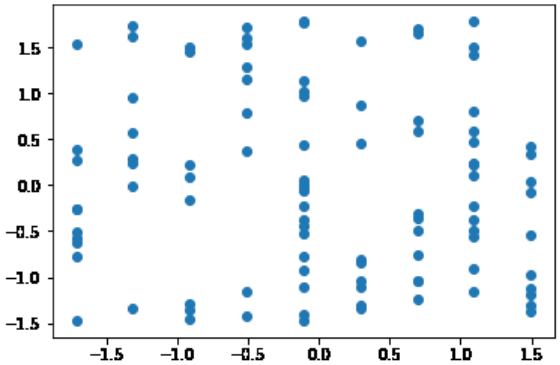


Homework

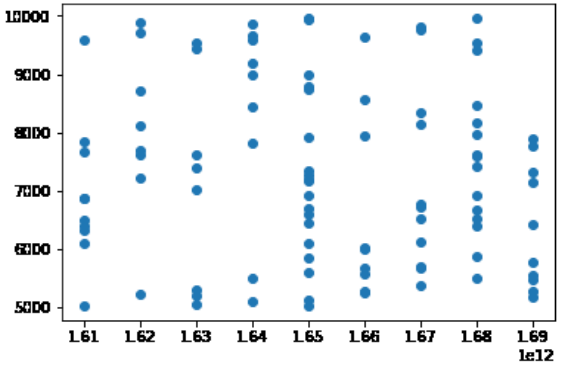
Based on '13 - Assessment_can_you_spot_the_fraudsters.ipynb' and dataset 'credit.csv', answer the following question:

Q2. Which graph contains the standardised representation of both No and Money?

Graph A



Graph B

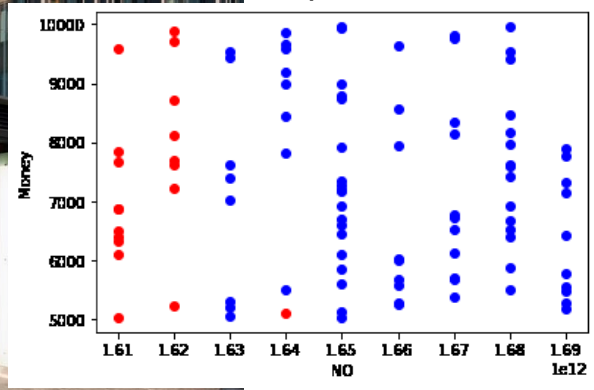


Homework

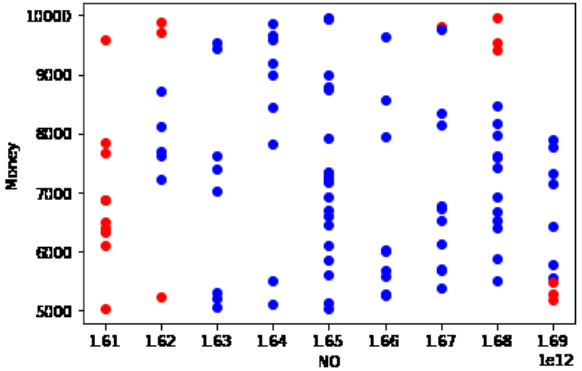
Based on '13 - Assessment_can_you_spot_the_fraudsters.ipynb' and dataset 'credit.csv', answer the following questions:

Q3. Which graph represents the result of LOF with 2 neighbours, 20 neighbours, and the elliptic envelope procedure, all at a contamination rate of 20%? The outliers are indicated in red.

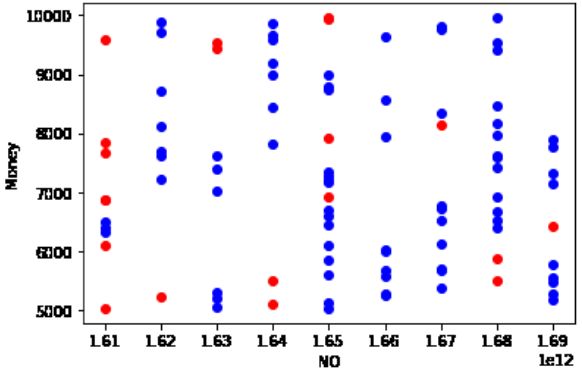
Graph A



Graph B



Graph C



Share your experiences on discussion board

- Have you taken any other approaches?
- Did you use a simpler or more convenient code?
- Were there any particular parts you struggled with?

Quiz

A company wants to perform the analysis of its yearly sales data to predict which customers will leave the company.

What approach should the company use?

- A. Time series analysis
- B. Classification
- C. Regression
- D. All of them can be used

Quiz

Which of the following techniques is/are normalisation used for?

- A. Regression
- B. Classification
- C. Pre-processing
- D. All of them above

Quiz

Which of the following parts of the data mining process can clustering be used for?

- A. Data selection
- B. Data pre-processing
- C. Data transformation
- D. All of them above

Quiz

When can predictive modelling achieve the best results?

- A. When the data is normalised.
- B. When the model is performing very well.
- C. When the mean and median of all variables are close together.
- D. When the mean and median of all variables are close together.
- E. None of them above

Data Quality

Data Quality

- Recall our Edinburgh Castle example.
  - If, the visitor count is available monthly for the last year, and bi-annually from the years before. This means that we have a different granularity, and we cannot apply the analysis directly.
  - Or consider the scenario in which you obtain the data on people visiting Edinburgh, which is incompatible with our own data because of how variables are recorded (for example,“F” and “M” instead of “male”/ “female”).

- **Causes of data quality problems**: human error, imprecise measurement, improper conversion of data sources due to encoding, language differences, translation or lacking technical infrastructure.

Data Quality: Edinburgh Castle

Obtained from Edinburgh Castle

	Name	Sales	Gender
1	PARKER Susan	2	Female
2	MCCAIDE Hamish	9999	Male
3	PARESH Ranjan	5	Male
4	CANTERBURY Pablo	-1	Male

Obtained from Edinburgh visitor centre

			Longitude	Latitude	Sex
1	Susan	Parker	98°57'08.7"N	3°11'35.7"W	F
2	Hamish	McCaid	55°56'54.9"N	3°11'59.6"W	F
3	Maria	Benevolente	55°56'28.1"N	3°11'30.6"W	F
4	Georgina	Albatross			F
5	Hamish	McCaide	55°56'54.9"N	3°11'59.6"W	M

Try to identify and discuss what data quality problems are present and most importantly how you would solve them. Next, consider how you would combine both tables to make a prediction on what type of customer is more likely to visit Edinburgh castle.


THINK
THINK
THINK

Data Quality: Edinburgh Castle

Obtained from Edinburgh Castle

	Name	Sales	Gender
1	PARKER Susan	2	Female
2	MCCAIDE Hamish	9999	Male
3	PARESH Ranjan	5	Male
4	CANTERBURY Pablo	-1	Male

Obtained from Edinburgh visitor centre

			Longitude	Latitude	Sex
1	Susan	Parker	98°57'08.7"N	3°11'35.7"W	F
2	Hamish	McCaid	55°56'54.9"N	3°11'59.6"W	F
3	Maria	Benevolente	55°56'28.1"N	3°11'30.6"W	F
4	Georgina	Albatross			F
5	Hamish	McCaide	55°56'54.9"N	3°11'59.6"W	M

Data quality problems:

- Incorrect sales number for observations 2 and 4
- No variable names for first name/last name

- Longitude and latitude switched
- Incorrect latitude for first observation (>90)
- Missing data for observation 4
- Probably an incorrect gender for observation 2

Data Quality: Edinburgh Castle

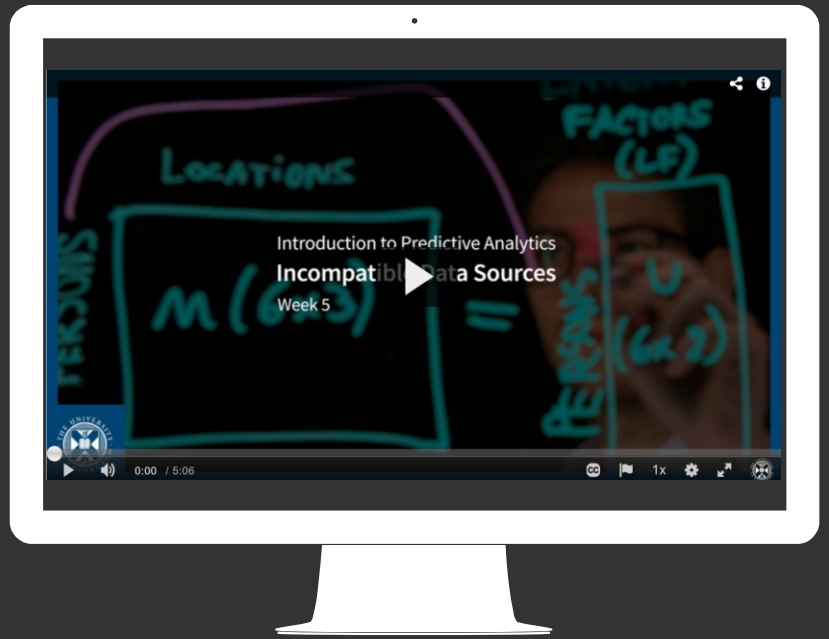
Obtained from Edinburgh Castle			
	Name	Sales	Gender
1	PARKER Susan	2	Female
2	MCCAIDE Hamish	9999	Male
3	PARESH Ranjan	5	Male
4	CANTERBURY Pablo	-1	Male

Obtained from Edinburgh visitor centre					
			Longitude	Latitude	Sex
1	Susan	Parker	98°57'08.7"N	3°11'35.7"W	F
2	Hamish	McCaid	55°56'54.9"N	3°11'59.6"W	F
3	Maria	Benevolente	55°56'28.1"N	3°11'30.6"W	F
4	Georgina	Albatross			F
5	Hamish	McCaide	55°56'54.9"N	3°11'59.6"W	M

Integration problems:

- Names are different: single variable in dataset 1, two in dataset 2, format different as well
- Gender not matching for obs 2
- Gender category not matching
- …

Incompatible Data Sources



Please watch the video through this link

- https://media.ed.ac.uk/media/Incompatible+Data+Sources/1_xs7pwapd/117185481

Missing Values

Missing Values

- Missing values may be caused by lack of entering the values into the system, no particular value to be recorded, impossible to record… e.g.:
- If someone wants to apply for a loan and being asked ‘any previous loans may have had’, the applicant may leave it blank. But we may change the question to ‘whether has had a loan previously or not’, and still get a good predictor.
- Sometimes we are just not allowed to record something because of privacy requirements.
- Some variables such as gender for example, cannot even be used to build predictive models in businesses because of discrimination laws.

Missing Values:How to deal with missing values?

- **Missing value imputation: just remove the observation.**
  - Useful when other variables for the same observation are also missing, implying that there was something wrong with the input procedure.
  - Be careful not omit too many observations!
  - Delete the entire observation is called listwise deletion;
    Or we can keep all data that is available but not treating the missing values ➔ not all models allow this; may increase error

- Worthwhile to investigate the cause, or check whether the missing values are correlated with a particular variable
  E.g. the number of previous loans is always missing from people living in a certain area.

Missing Values:How to deal with missing values?

- **Use the central tendency to fill in the gaps:**
  the mean or median of corresponding variable
  - Useful when the predictive model cannot work with missing values, such as Neural Networks.
  - Only apply to numeric variables
  - For categorical values, may create a new category for NA (not available/applicable)
    e.g. assume we have a binary variable with classes 0 and 1, its missing value can be treated as the third class NA

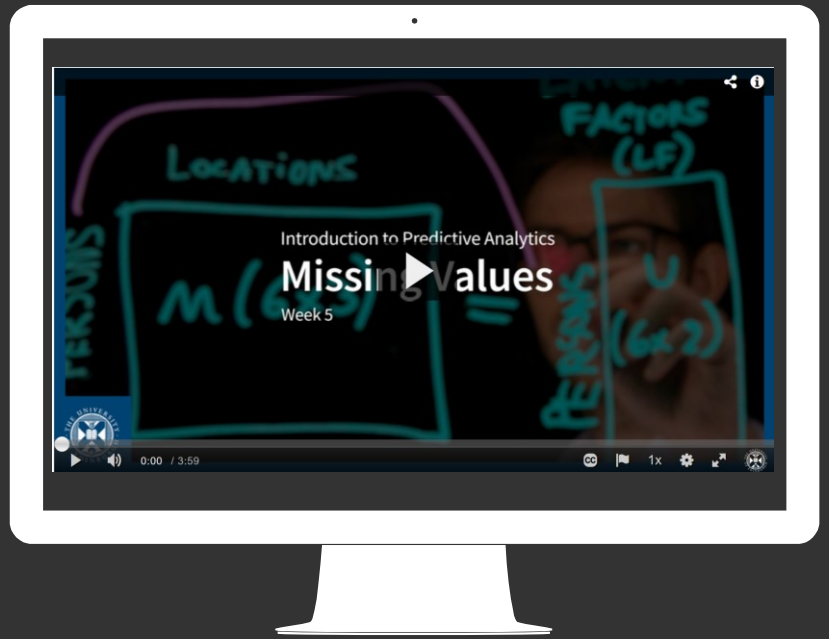
Missing Values:How to deal with missing values?

- **Nearest neighbour correction:**
  find the nearest neighbours and apply central tendency
  - Can be computationally expensive, needs to calculate the neighbours for all observations with the missing values against the whole dataset.
  - The number of neighbours is tricky to set

- **Build a regression (classification) model to predict the missing continuous (categorical) values**

- **If too many values missing, better to omit the whole feature**

Missing Values



Please watch the video through this link

- https://media.ed.ac.uk/media/Missing+Values/1_jar6gk6u/117185481

- Study the following file
  15 - missing_values.ipynb + MV_example.csv

Activity: Missing values and Python

- Study the following file
  16 - merging_datasets_in_python_reading.ipynb+ DM_1.csv + DM_2.csv
- Homework
  17 - can_you_merge_two_datasets_NB_activity5.ipynb + NB_1.csv + NB_2.csv

Activity: Merging datasets in Python

Read the following academic papers and try to identify whether they dealt with any data quality and merging issues:

- Were there any techniques that require missing value correction?
- Were there any techniques that required mixing two (incompatible) datasets?

Van Vlasselaer, Véronique, et al. "Gotcha! Network-based fraud detection for social security fraud." Management Science 63.9 (2016): 3090-3110.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of computational science 2.1 (2011): 1-8.

Yao, Xiao, Jonathan Crook, and Galina Andreeva. "Support vector regression for loss given default modelling." European Journal of Operational Research 240.2 (2015): 528-538.

Reading: Handling real-life data problems (Optional)

UNIVERSITY OF EDINBURGH
Business School