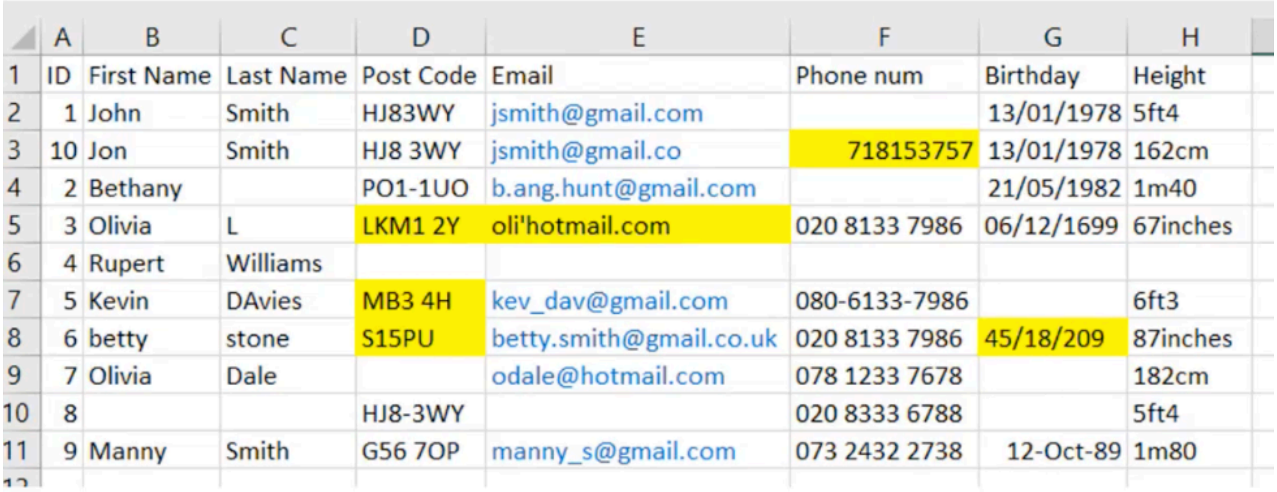
W2 - Data Quality

Data Quality - Degree of excellence exhibited by data in relation to the portrayal of the actual scenario.

- **Validity**
- **Accuracy**
- **Completeness**
- **Consistency**
- **Uniformity**
- **Redundancy**

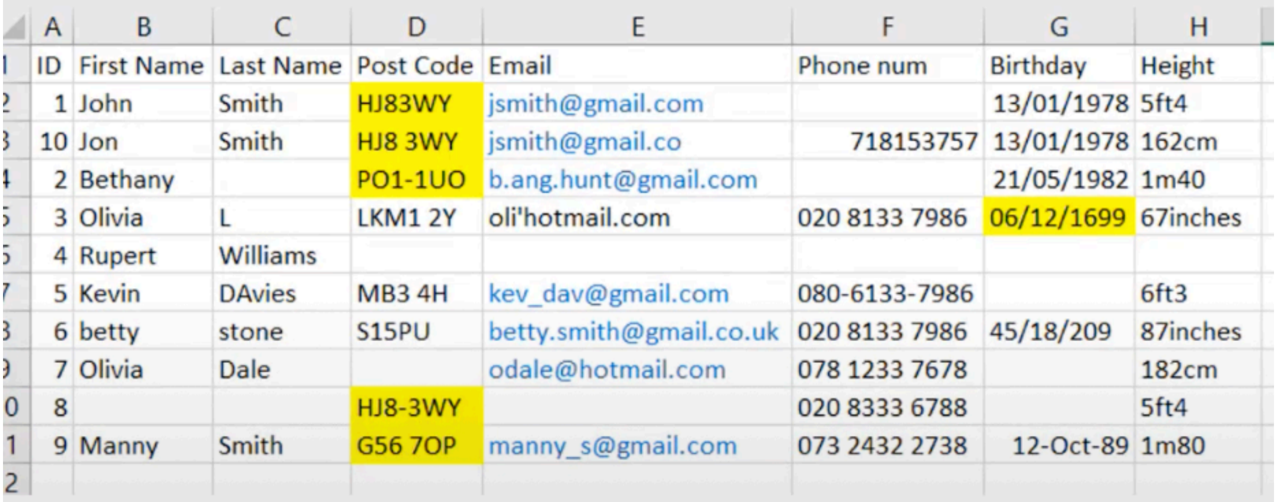
Validity - the degree to which data comply with the defined rules or constraints. **How valid is it?**

- Data type constraints
- Range constraints
- Mandatory constraints
- Unique constraints



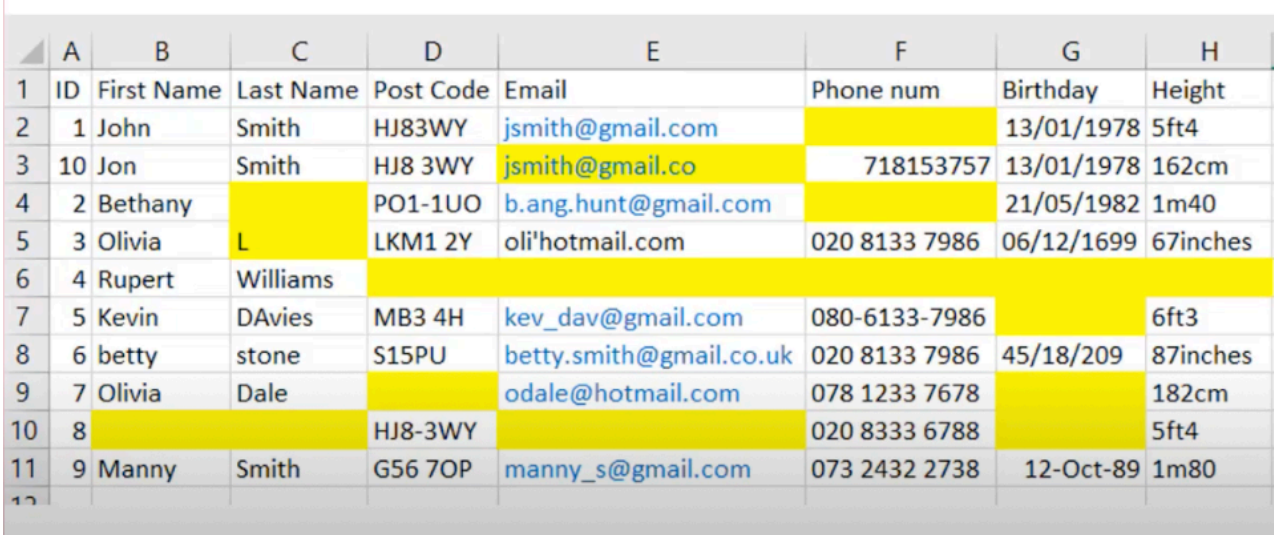
	A	B	C	D	E	F	G	H
1	ID	First Name	Last Name	Post Code	Email	Phone num	Birthday	Height
2	1	John	Smith	HJ83WY	jsmith@gmail.com		13/01/1978	5ft4
3	10	Jon	Smith	HJ8 3WY	jsmith@gmail.co	718153757	13/01/1978	162cm
4	2	Bethany		PO1-1UO	b.ang.hunt@gmail.com		21/05/1982	1m40
5	3	Olivia	L	LKM1 2Y	oli'hotmail.com	020 8133 7986	06/12/1699	67inches
6	4	Rupert	Williams					
7	5	Kevin	DAvies	MB3 4H	kev_dav@gmail.com	080-6133-7986		6ft3
8	6	betty	stone	S15PU	betty.smith@gmail.co.uk	020 8133 7986	45/18/209	87inches
9	7	Olivia	Dale		odale@hotmail.com	078 1233 7678		182cm
10	8			HJ8-3WY		020 8333 6788		5ft4
11	9	Manny	Smith	G56 7OP	manny_s@gmail.com	073 2432 2738	12-Oct-89	1m80
12								

Accuracy - the degree to which data reflect the true value or a standard



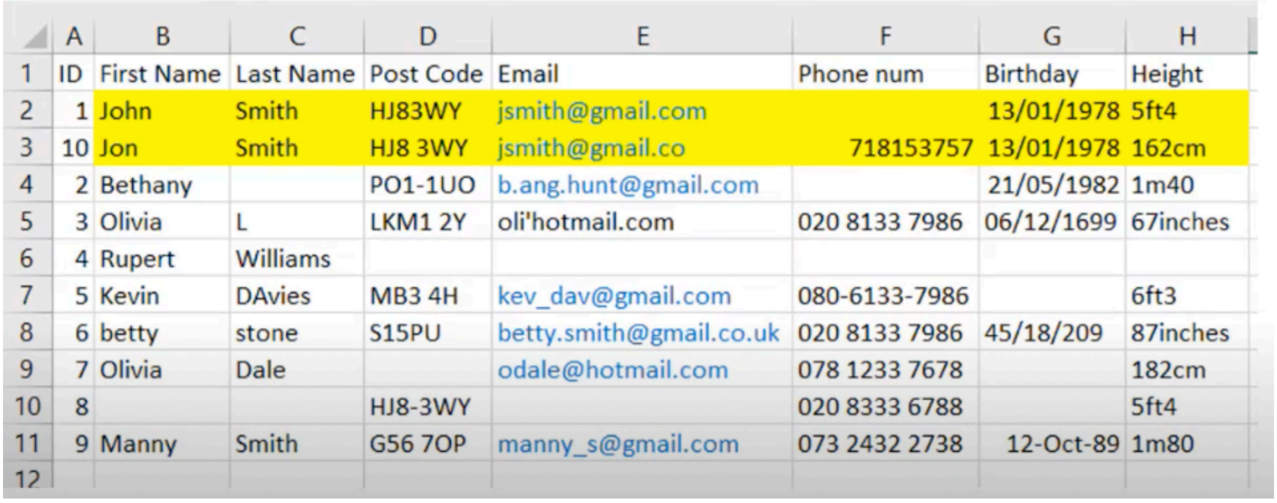
	A	B	C	D	E	F	G	H
l	ID	First Name	Last Name	Post Code	Email	Phone num	Birthday	Height
2	1	John	Smith	HJ83WY	jsmith@gmail.com		13/01/1978	5ft4
3	10	Jon	Smith	HJ8 3WY	jsmith@gmail.co	718153757	13/01/1978	162cm
4	2	Bethany		PO1-1UO	b.ang.hunt@gmail.com		21/05/1982	1m40
5	3	Olivia	L	LKM1 2Y	oli'hotmail.com	020 8133 7986	06/12/1699	67inches
5	4	Rupert	Williams					
7	5	Kevin	DAvies	MB3 4H	kev_dav@gmail.com	080-6133-7986		6ft3
3	6	betty	stone	S15PU	betty.smith@gmail.co.uk	020 8133 7986	45/18/209	87inches
9	7	Olivia	Dale		odale@hotmail.com	078 1233 7678		182cm
0	8			HJ8-3WY		020 8333 6788		5ft4
1	9	Manny	Smith	G56 7OP	manny_s@gmail.com	073 2432 2738	12-Oct-89	1m80
2								

Completeness - the degree to which all required data points are known. Missing values, truncated information



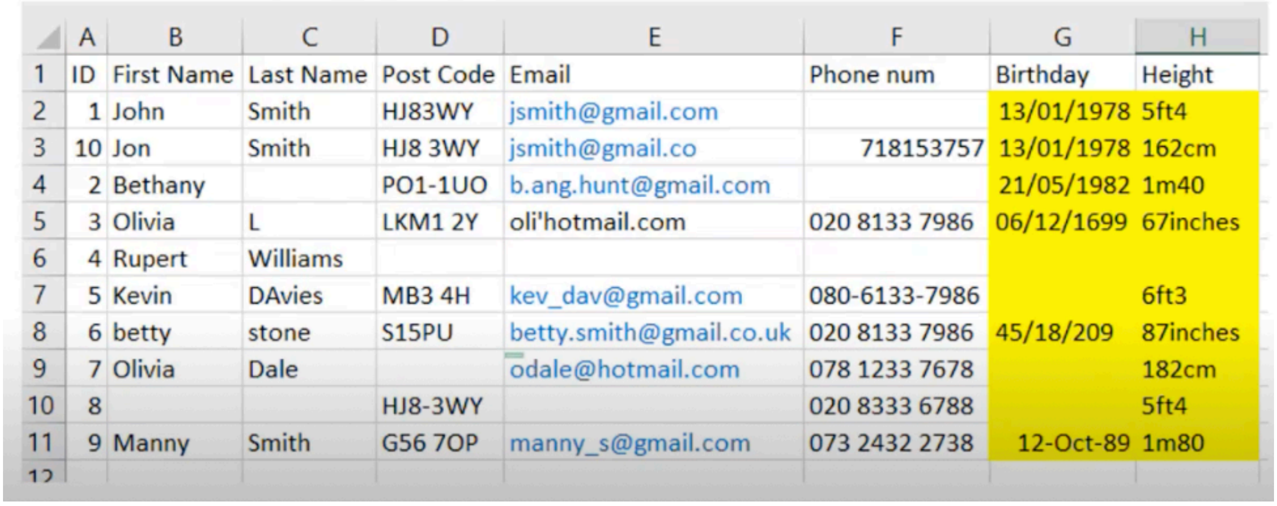
	A	B	C	D	E	F	G	H
1	ID	First Name	Last Name	Post Code	Email	Phone num	Birthday	Height
2	1	John	Smith	HJ83WY	jsmith@gmail.com		13/01/1978	5ft4
3	10	Jon	Smith	HJ8 3WY	jsmith@gmail.co	718153757	13/01/1978	162cm
4	2	Bethany		PO1-1UO	b.ang.hunt@gmail.com		21/05/1982	1m40
5	3	Olivia	L	LKM1 2Y	oli'hotmail.com	020 8133 7986	06/12/1699	67inches
6	4	Rupert	Williams					
7	5	Kevin	DAvies	MB3 4H	kev_dav@gmail.com	080-6133-7986		6ft3
8	6	betty	stone	S15PU	betty.smith@gmail.co.uk	020 8133 7986	45/18/209	87inches
9	7	Olivia	Dale		odale@hotmail.com	078 1233 7678		182cm
10	8			HJ8-3WY		020 8333 6788		5ft4
11	9	Manny	Smith	G56 7OP	manny_s@gmail.com	073 2432 2738	12-Oct-89	1m80

Consistency - the degree to which data points are consistent across their group. when 2 values in a dataset contradict each other



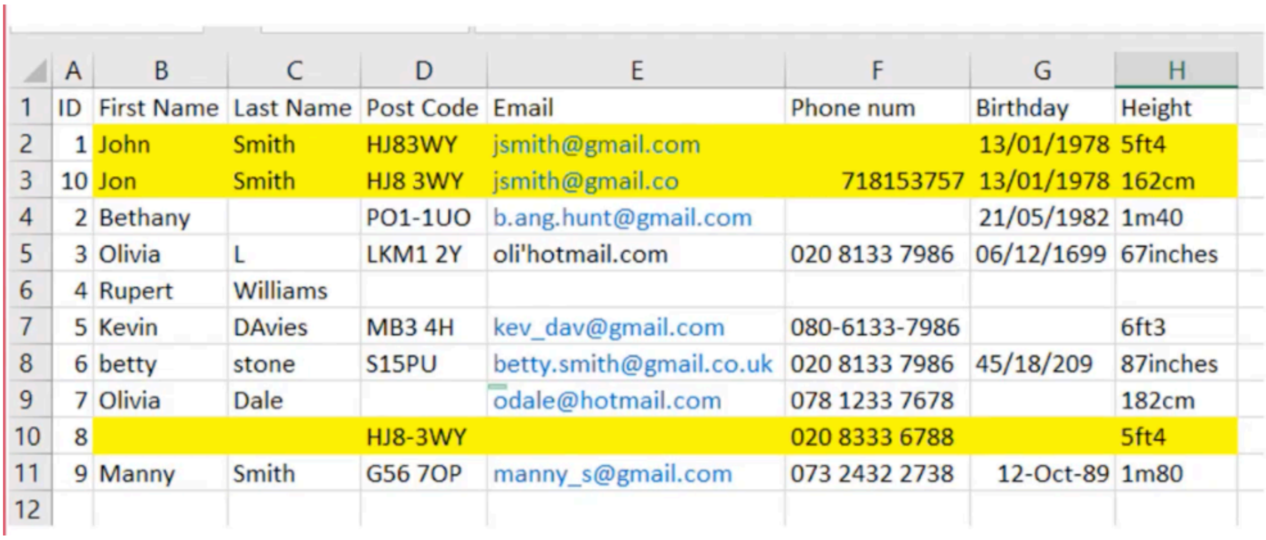
	A	B	C	D	E	F	G	H
1	ID	First Name	Last Name	Post Code	Email	Phone num	Birthday	Height
2	1	John	Smith	HJ83WY	jsmith@gmail.com		13/01/1978	5ft4
3	10	Jon	Smith	HJ8 3WY	jsmith@gmail.co	718153757	13/01/1978	162cm
4	2	Bethany		PO1-1UO	b.ang.hunt@gmail.com		21/05/1982	1m40
5	3	Olivia	L	LKM1 2Y	oli'hotmail.com	020 8133 7986	06/12/1699	67inches
6	4	Rupert	Williams					
7	5	Kevin	DAvies	MB3 4H	kev_dav@gmail.com	080-6133-7986		6ft3
8	6	betty	stone	S15PU	betty.smith@gmail.co.uk	020 8133 7986	45/18/209	87inches
9	7	Olivia	Dale		odale@hotmail.com	078 1233 7678		182cm
10	8			HJ8-3WY		020 8333 6788		5ft4
11	9	Manny	Smith	G56 7OP	manny_s@gmail.com	073 2432 2738	12-Oct-89	1m80
12								

Uniformity - the degree to which data follows the same units of measure in all systems. Weight, height, currency.



	A	B	C	D	E	F	G	H
1	ID	First Name	Last Name	Post Code	Email	Phone num	Birthday	Height
2	1	John	Smith	HJ83WY	jsmith@gmail.com		13/01/1978	5ft4
3	10	Jon	Smith	HJ8 3WY	jsmith@gmail.co	718153757	13/01/1978	162cm
4	2	Bethany		PO1-1UO	b.ang.hunt@gmail.com		21/05/1982	1m40
5	3	Olivia	L	LKM1 2Y	oli'hotmail.com	020 8133 7986	06/12/1699	67inches
6	4	Rupert	Williams					
7	5	Kevin	DAvies	MB3 4H	kev_dav@gmail.com	080-6133-7986		6ft3
8	6	betty	stone	S15PU	betty.smith@gmail.co.uk	020 8133 7986	45/18/209	87inches
9	7	Olivia	Dale		odale@hotmail.com	078 1233 7678		182cm
10	8			HJ8-3WY		020 8333 6788		5ft4
11	9	Manny	Smith	G56 7OP	manny_s@gmail.com	073 2432 2738	12-Oct-89	1m80

Redundancy - the information of the same observation is held within a database or storage technology.



	A	B	C	D	E	F	G	H
1	ID	First Name	Last Name	Post Code	Email	Phone num	Birthday	Height
2	1	John	Smith	HJ83WY	jsmith@gmail.com		13/01/1978	5ft4
3	10	Jon	Smith	HJ8 3WY	jsmith@gmail.co	718153757	13/01/1978	162cm
4	2	Bethany		PO1-1UO	b.ang.hunt@gmail.com		21/05/1982	1m40
5	3	Olivia	L	LKM1 2Y	oli'hotmail.com	020 8133 7986	06/12/1699	67inches
6	4	Rupert	Williams					
7	5	Kevin	DAvies	MB3 4H	kev_dav@gmail.com	080-6133-7986		6ft3
8	6	betty	stone	S15PU	betty.smith@gmail.co.uk	020 8133 7986	45/18/209	87inches
9	7	Olivia	Dale		odale@hotmail.com	078 1233 7678		182cm
10	8			HJ8-3WY		020 8333 6788		5ft4
11	9	Manny	Smith	G56 7OP	manny_s@gmail.com	073 2432 2738	12-Oct-89	1m80
12								

Data Cleaning

- **Irrelevant data** - those that are not actually needed, and don't fit under the context of the problem we're solving. Drop unneeded columns/rows
- **Duplicates** - data points that are repeated.
- **Type Conversions** - make sure that numbers are stored as numerical data types. Data should be stored as a date object. Categorical values can be converted into and from numbers. If values can't be converted into the appropriate type, the value is incorrect.
- **Padding** - strings and numbers can be padded with extra characters or digits to ensure they are a certain width.
- **Typos** - strings can be entered in many ways so can have many mistakes. Use a bar plot or histogram to visualize unique values.
- **Standardization** - put each value in the same format, so that it is uniform. Strings/numbers/measurement/currency/date/time
- **Scaling** - scaling data values to a specified range to compare different scores like percentages
- **Normalization/standardization** - type of scaling, rescales data values into a range between 0-1.

Handling Missing Values

- **Drop** - drop rows containing missing values.
- **Impute (Mean/Median/Linear)** - calculate missing values based on other observations. Line of best fit.
- **Hot-Deck** copying values from other similar records. can be applied to numerical/categorical data
- **Flag** - saying that data is missing is informative, missing data can be flagged. numeric data can be filled with a specific number such as 0. Categorical data can be filled with 'Missing'