Missing Data: I

What does missing data look like?

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35.0	NaN	33.6	0.627	50	1
1	1	85	66	29.0	NaN	26.6	0.351	31	0
2	8	183	64	NaN	NaN	23.3	0.672	32	1
3	1	89	66	23.0	94.0	28.1	0.167	21	0
4	0	137	40	35.0	168.0	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48.0	180.0	32.9	0.171	63	0
764	2	122	70	27.0	NaN	36.8	0.340	27	0
765	5	121	72	23.0	112.0	26.2	0.245	30	0
766	1	126	60	NaN	NaN	30.1	0.349	47	1
767	1	93	70	31.0	NaN	30.4	0.315	23	0

Data based on https://raw.githubusercontent.com/susanli2016/Machine-Learning-with-Python/master/diabetes.csv

What does missing data look like?

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	<NA>	33.6	0.627	50	1
1	1	85	66	29	<NA>	26.6	0.351	31	0
2	8	183	64	<NA>	<NA>	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	<NA>	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	<NA>	<NA>	30.1	0.349	47	1
767	1	93	70	31	<NA>	30.4	0.315	23	0

What does missing data look like?

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	--	33.6	0.627	50	1
1	1	85	66	29	<NA>	26.6	0.351	31	0
2	8	183	64	--	NaN	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	<NA>	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	N/A	<NA>	30.1	0.349	47	1
767	1	93	70	31	<NA>	30.4	0.315	23	0

Unify missing data markers

```
data = pd.read_csv("data.csv",
                   na_values = ['--', 'N/A', 'na'])
```

Count missing data

```
38] data.isnull().sum()

    Pregnancies                 0
    Glucose                     5
    BloodPressure              35
    SkinThickness             227
    Insulin                     0
    BMI                        11
    DiabetesPedigreeFunction    0
    Age                         0
    Outcome                     0
    dtype: int64
```

- `isnull` returns True when something does not have a value.
- `notnull` returns True when something does have a value.
- Can use these methods to select rows or to count as shown here.

What does missing data look like?

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns