Sampling

While dimension elimination is reducing the number of attributes, sampling is working on the number of records.

Setup

```
import numpy as np
import pandas as pd
```

```
df = pd.read_csv('/content/sample_data/california_housing_train.csv')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17000 entries, 0 to 16999
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           17000 non-null  float64
 1   latitude            17000 non-null  float64
 2   housing_median_age  17000 non-null  float64
 3   total_rooms         17000 non-null  float64
 4   total_bedrooms      17000 non-null  float64
 5   population          17000 non-null  float64
 6   households          17000 non-null  float64
 7   median_income       17000 non-null  float64
 8   median_house_value  17000 non-null  float64
dtypes: float64(9)
memory usage: 1.2 MB
```

Sampling by numbers

```
df.sample(n=5)
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	po
15268	-122.28	37.90	49.0	3191.0	516.0	
762	-117.06	32.77	18.0	2269.0	682.0	
11262	-121.10	35.60	20.0	3389.0	704.0	
4084	-117.98	34.06	33.0	1353.0	228.0	
7767	-118.38	33.80	36.0	4421.0	702.0	

Sampling by percentage

```
df.sample(frac=0.001)
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	po
352	-116.90	32.79	21.0	3770.0	491.0	
2408	-117.56	33.88	40.0	1196.0	294.0	
10461	-120.37	40.17	21.0	789.0	141.0	
16562	-122.69	39.02	27.0	2199.0	527.0	
922	-117.08	32.62	28.0	2468.0	506.0	
10882	-120.84	38.77	11.0	1013.0	188.0	
12997	-121.85	37.28	17.0	4208.0	954.0	
9818	-119.69	36.41	38.0	1016.0	202.0	
1432	-117.19	34.27	16.0	7961.0	1147.0	
13195	-121.90	38.00	14.0	2677.0	368.0	
13253	-121.91	37.33	52.0	2212.0	563.0	
13422	-121.94	37.26	43.0	2104.0	388.0	
5897	-118.20	33.81	45.0	944.0	178.0	
8433	-118.48	35.14	4.0	8417.0	1657.0	
1769	-117.25	32.74	40.0	2186.0	549.0	
16978	-124.18	40.79	39.0	1836.0	352.0	
2349	-117.49	33.99	21.0	2050.0	392.0	

```
df.loc[:10].sample(frac=0.9)
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	popul
10	-114.60	33.62	16.0	3741.0	801.0	
1	-114.47	34.40	19.0	7650.0	1901.0	
4	-114.57	33.57	20.0	1454.0	326.0	
7	-114.59	34.83	41.0	812.0	168.0	
8	-114.59	33.61	34.0	4789.0	1175.0	
2	-114.56	33.69	17.0	720.0	174.0	
5	-114.58	33.63	29.0	1387.0	236.0	
0	-114.31	34.19	15.0	5612.0	1283.0	
3	-114.57	33.64	14.0	1501.0	337.0	
9	-114.60	34.83	46.0	1497.0	309.0	

Sampling with replacement

```
df.loc[:10].sample(15,replace=True)
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	popul
5	-114.58	33.63	29.0	1387.0	236.0	
4	-114.57	33.57	20.0	1454.0	326.0	
1	-114.47	34.40	19.0	7650.0	1901.0	
4	-114.57	33.57	20.0	1454.0	326.0	
3	-114.57	33.64	14.0	1501.0	337.0	
10	-114.60	33.62	16.0	3741.0	801.0	
9	-114.60	34.83	46.0	1497.0	309.0	
1	-114.47	34.40	19.0	7650.0	1901.0	
4	-114.57	33.57	20.0	1454.0	326.0	
4	-114.57	33.57	20.0	1454.0	326.0	
1	-114.47	34.40	19.0	7650.0	1901.0	
1	-114.47	34.40	19.0	7650.0	1901.0	
6	-114.58	33.61	25.0	2907.0	680.0	
3	-114.57	33.64	14.0	1501.0	337.0	
1	-114.47	34.40	19.0	7650.0	1901.0	

Colab paid products - Cancel contracts here