

Stratified Sampling in Pandas

- Use `groupby` on the label column to create sub-frames for each label and then use the `sample()` function.
- Passing an integer gives an exact sample (e.g. `sample(5)` gives 5 rows).
- Passing `frac=0.1` gives a percentage (i.e. 10%)
- Remember to set `random_state` for reproducible results

```
df = pd.read_csv("path/to/data.csv")
```

```
df.groupby('Category', group_keys=False).apply(lambda x: x.sample(frac=0.1, random_state=42))
```

Revision #2

Created 10 November 2022 11:21:06 by James

Updated 11 February 2024 16:27:03 by James