Introduction to Big Data

- Developed by Dr. Keungoui KIM
- https://awekim.github.io/portfolio/

Lecture 6. Sampling

```
import pandas as pd
housing = pd.read_csv('/content/drive/MyDrive/[Lecture]/IntBigData/BigData_Python/06_Sampling/housing.csv
housing.head()
```

```
housing.sample(n=10000).shape
```

```
housing.sample(frac=0.3).shape
```

```
housing.sample(n=2, axis=1).head()
```

```
housing.sample(n=2, axis=1).shape
```

```
housing.sample(n=10000).head()
```

```
housing.sample(n=10000).head()
```

```
housing.sample(n=10000, random_state=1).head()
```

```
housing.sample(n=10000, random_state=1).head()
```

Sampling with California Housing Data

1. longitude: A measure of how far west a house is; a higher value is farther west
2. latitude: A measure of how far north a house is; a higher value is farther north
3. housingMedianAge: Median age of a house within a block; a lower number is a newer building
4. totalRooms: Total number of rooms within a block
5. totalBedrooms: Total number of bedrooms within a block
6. population: Total number of people residing within a block
7. households: Total number of households, a group of people residing within a home unit, for a block
8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. medianHouseValue: Median house value for households within a block (measured in US Dollars)
10. oceanProximity: Location of the house w.r.t ocean/sea

```
import pandas as pd
housing = pd.read_csv('/content/drive/MyDrive/[Lecture]/IntBigData/BigData_Python/06_Sampling/housing.cs
housing.head()
```

```
housing.shape
```

```
housing.info()
```

```
housing.describe()
```

```
housing.isnull().any()
```

```
housing['total_bedrooms'].unique()
```

```
housing['total_bedrooms'].isnull()
```

```
housing[housing['total_bedrooms'].isnull()]
```

```
housing['ocean_proximity'].unique()
```

```
housing['ocean_proximity'].value_counts()
```

```
housing['ocean_proximity'].value_counts(normalize=True)
```

```
housing['median_income'].mean()
```

```
housing_op = housing[housing.ocean_proximity != 'ISLAND']
housing_op['ocean_proximity'].value_counts()
```

```
housing_op['ocean_proximity'].value_counts(normalize=True)
```

Random Sampling

```
housing_op = housing[housing.ocean_proximity != 'ISLAND']
housing_op_1000 = housing_op.sample(n=1000, random_state=1)
```

```
housing_op_1000['ocean_proximity'].value_counts()
```

```
housing_op_1000['ocean_proximity'].value_counts(normalize=True)
```

Groupby -> Random Sampling

```
housing_op_gr = housing_op.groupby('ocean_proximity')
housing_op_gr.head(1)
```

```
housing_op_gr_1000 = housing_op_gr.sample(n=1000, random_state=17)
```

```
housing_op_gr_1000.ocean_proximity.value_counts()
```

```
housing_op_gr_1000.ocean_proximity.value_counts(normalize=True)
```

Comparison of Random Sampling and Groupby-Random Sampling

```
housing_op_1000 = housing_op.sample(n=1000, random_state=1)
(
    housing_op_1000.groupby('ocean_proximity')['median_income'].
 mean().reset_index()
)
```

```
housing_op_gr_1000 = housing_op_gr.sample(n=1000, random_state=17)
(
    housing_op_gr_1000.groupby('ocean_proximity')['median_income'].
 mean().reset_index()
)
```

Visualization

```
import seaborn as sns
sns.set(rc={'figure.figsize':(12,9)})

housing_op_1000_sum = (
    housing_op_1000.groupby('ocean_proximity')['median_income'].
 mean().reset_index()
)
sns.barplot(x='ocean_proximity',y='median_income',
            data=housing_op_1000_sum).set_title('Sampling')

housing_op_gr_1000_sum = (
    housing_op_gr_1000.groupby('ocean_proximity')['median_income'].
 mean().reset_index()
)
sns.barplot(x='ocean_proximity',y='median_income',
            data=housing_op_gr_1000_sum).set_title('Groupby-Sampling')

housing_op_1000_sum['type'] = 'Sampling'
housing_op_gr_1000_sum['type'] = 'GroupbySampling'

housing_op_1000_sum_all = (
    housing_op_1000_sum.append(housing_op_gr_1000_sum)
)

sns.barplot(x='ocean_proximity',y='median_income',
            data=housing_op_1000_sum_all,
            hue='type').set_title('ALL')
```