Introduction to Big Data

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

Lecture 8. Descriptive Statistics

Descriptive Analysis

```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
```

Mean & median comparison

```
sampling1 = np.random.normal(100,20,10000)
sampling1
```

```
len(sampling1)
```

```
sns.histplot(sampling1)
```

```
### mean
# sampling1.mean()
np.mean(sampling1)
```

```
### median
np.median(sampling1)
```

```
# Add outlier
sampling1 = np.append(sampling1, [30000])
sampling1
```

```
np.mean(sampling1)
```

```
np.median(sampling1)
```

```
sns.histplot(sampling1)
```

```
sns.distplot(sampling1)
```

Descriptive analysis with tips

```
tips = sns.load_dataset('tips')
```

```
tips.head()
```

```
# sns.histplot(tips['total_bill'])
sns.histplot(x='total_bill',data=tips)
```

```
tips['total_bill'].mode()
```

```
tips['total_bill'].mode().iloc[0]
```

```
tips.query('total_bill==13.42')
```

```
tips_mean=tips['total_bill'].mean()
tips_mean
```

```
tips_median=tips['total_bill'].median()
tips_median
```

```
tips_mode=tips['total_bill'].mode().values[0]
tips_mode
```

```
sns.histplot(tips['total_bill'])
plt.axvline(tips_mean, color='r', linestyle='--', label="Mean")
plt.axvline(tips_median, color='g', linestyle='-', label="Median")
plt.axvline(tips_mode, color='y', linestyle='-', label="Mode")
```

Descriptive analysis with dataframe

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

```
grade_df = pd.DataFrame({'student':['Kim','Lee','Park','Choi','Jang','Wang','Ha
                         'midterm':[88,93,67,75,52,78,99,34,74,83],
                         'final':[92,94,56,78,34,25,90,np.nan,63,85]})
grade_df
```

```
grade_df['final'].min()
```

```
grade_df.final.min()
```

```
grade_df['midterm'].max()
```

```
grade_df.sum()
```

```
# grade_df.sum(axis='columns')
grade_df.sum(axis=1)
```

```
# grade_df.mean(axis='rows')
grade_df.sum()
```

```
grade_df.mean()
```

```
# grade_df.mean(axis='rows',skipna=False)
grade_df.mean(skipna=False)
```

```
# grade_df.mean(axis='columns')
grade_df.mean(axis=1)
```

```
# grade_df.mean(axis='columns', skipna=False)
grade_df.mean(axis=1, skipna=False)
```

```
# grade_df.median(axis='rows')
grade_df.median()
```

```
grade_df.sort_values(by=['final'],
                     ascending=False)
```

```
grade_df.sort_values(by=['final'],
                     ascending=False)
```

```
grade_df.std()
```

```
grade_df.var()
```

```
grade_df.describe()
```

```
sns.distplot(grade_df['midterm'])
```

```
sns.distplot(grade_df['final'])
```

```
midterm_mean=grade_df['midterm'].mean()
```

```
midterm_median=grade_df['midterm'].median()
midterm_mode=grade_df['midterm'].mode()
```

```
sns.histplot(grade_df['midterm'])
plt.axvline(midterm_mean, color='r', linestyle='--', label="Mean")
plt.axvline(midterm_median, color='g', linestyle='-', label="Median")
#plt.axvline(midterm_mode, color='y', linestyle='-', label="Mode")
```

```
final_mean=grade_df['final'].mean()
final_median=grade_df['final'].median()
final_mode=grade_df['final'].mode()
```

```
sns.histplot(grade_df['final'])
plt.axvline(final_mean, color='r', linestyle='--', label="Mean")
plt.axvline(final_median, color='g', linestyle='-', label="Median")
#plt.axvline(midterm_mode, color='y', linestyle='-', label="Mode")
```