Lecture 6
Distances between Observations

Dennis Sun
Stanford University
DATASCI / STATS 112

January 23, 2023

1 Selecting Rows

2 Distances between Observations

3 Reminders

```
df_titanic = pd.read_csv(
    "http://dlsun.github.io/stats112/data/titanic.csv",
    index_col="name")
df_titanic
```

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29.0000	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
Allison, Miss. Helen Loraine	1	0	female	2.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
Allison, Mr. Hudson Joshua Creighton	1	0	male	30.0000	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	1	0	female	25.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Zabour, Miss. Hileni	3	0	female	14.5000	1	0	2665	14.4542	NaN	C	NaN	328.0	NaN
Zabour, Miss. Thamine	3	0	female	NaN	1	0	2665	14.4542	NaN	C	NaN	NaN	NaN
Zakarian, Mr. Mapriededer	3	0	male	26.5000	0	0	2656	7.2250	NaN	C	NaN	304.0	NaN
Zakarian, Mr. Ortin	3	0	male	27.0000	0	0	2670	7.2250	NaN	C	NaN	NaN	NaN

We've seen how to select *columns* from a **`DataFrame`**:

```
df_titanic["fare"]
```

But how do we select *rows*?

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29.0000	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
Allison, Miss. Helen Loraine	1	0	female	2.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
Allison, Mr. Hudson Joshua Creighton	1	0	male	30.0000	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	1	0	female	25.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Zabour, Miss. Hileni	3	0	female	14.5000	1	0	2665	14.4542	NaN	C	NaN	328.0	NaN
Zabour, Miss. Thamine	3	0	female	NaN	1	0	2665	14.4542	NaN	C	NaN	NaN	NaN
Zakarian, Mr. Mapriededer	3	0	male	26.5000	0	0	2656	7.2250	NaN	C	NaN	304.0	NaN
Zakarian, Mr. Ortin	3	0	male	27.0000	0	0	2670	7.2250	NaN	C	NaN	NaN	NaN

We can refer to rows by their *name* using `.loc`:

```
df_titanic.loc["Allison, Miss. Helen Loraine"]
```

or by their *position* using `.iloc`:

```
df_titanic.iloc[2]
```

Selecting One Row

```
df_titanic.loc["Allison, Miss. Helen Loraine"]
```

```
pclass                                      1
survived                                    0
sex                                    female
age                                       2.0
sibsp                                       1
parch                                       2
ticket                                 113781
fare                                   151.55
cabin                                  C22 C26
embarked                                    S
boat                                      NaN
body                                      NaN
home.dest     Montreal, PQ / Chesterville, ON
Name: Allison, Miss. Helen Loraine, dtype: object
```

Selecting Multiple Rows

```
df_titanic.loc[["Allison, Master. Hudson Trevor",
                "Allison, Miss. Helen Loraine",
                "Allison, Mr. Hudson Joshua Creighton"]]
```

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.55	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
Allison, Miss. Helen Loraine	1	0	female	2.0000	1	2	113781	151.55	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
Allison, Mr. Hudson Joshua Creighton	1	0	male	30.0000	1	2	113781	151.55	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON

We can achieve the same result using slice notation.

```
df_titanic.loc["Allison, Master. Hudson Trevor":
               "Allison, Mr. Hudson Joshua Creighton"]
```

Selecting Rows and Columns

Here is the general syntax for selecting both rows and columns:

```
df_titanic.loc[ROWS, COLS]
```

Example:

```
df_titanic.loc["Allison, Master. Hudson Trevor":
               "Allison, Mr. Hudson Joshua Creighton",
               ["age", "fare"]]
```

name	age	fare
Allison, Master. Hudson Trevor	0.9167	151.55
Allison, Miss. Helen Loraine	2.0000	151.55
Allison, Mr. Hudson Joshua Creighton	30.0000	151.55

1 Selecting Rows

2 Distances between Observations

3 Reminders

Distances between Observations

The remainder of this lesson is in this Colab:

1 Selecting Rows

2 Distances between Observations

3 Reminders

Reminders

- Assignment 2 due Friday. Upload to Gradescope before midnight.
- Exam 1 is in class next Monday. Make sure you look at the info and practice on the course website.