Week 5: Pandas!

DSUA111: Data Science for Everyone, NYU, Fall 2020

TA Jeff, `jpj251@nyu.edu`

- This slideshow: https://jjacobs.me/dsua111-sections/week-05 (https://jjacobs.me/dsua111-sections/week-05)
- All materials: https://github.com/jpowerj/dsua111-sections (https://github.com/jpowerj/dsua111-sections)

Outline

1. HW1 Review
2. Pandas!

HW1 Review

The 4 most-missed questions (by far):

- q5b (~40% wrong)
- q7a (~35% wrong)
- q7c (~30% wrong)
- q7b (~25% wrong)

(Q5B) What is it about selecting on the dependent variable here that limits our ability to draw inferences about the causal relationship between the independent variables and the outcome?

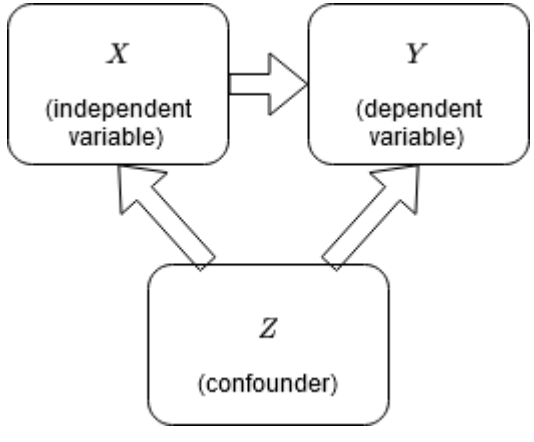
A. There may other countries where people are happy, but which do not have easy access to the outdoors
B. Happiness may be difficult to measure for some countries
---> C. We do not observe whether the factors common to the top performers might also be present in non-top performers
D. We must never draw causal conclusions from observational data

A and B: Unrelated to selecting on the independent variable (i.e., they are potential issues *whether or not the researcher selects on the independent variable)

D is just false. We should be *more careful* when drawing causal conclusions from observational data (relative to experimental data), but it's not this all-or-nothing. Sometimes experimental data is flawed to the point that we shouldn't draw causal conclusions from it, and sometimes observational data is good enough that we *should* draw causal conclusions after a careful analysis. (see, e.g., https://www.journals.uchicago.edu/doi/abs/10.1086/700936?mobileUi=0& (https://www.journals.uchicago.edu/doi/abs/10.1086/700936?mobileUi=0&))

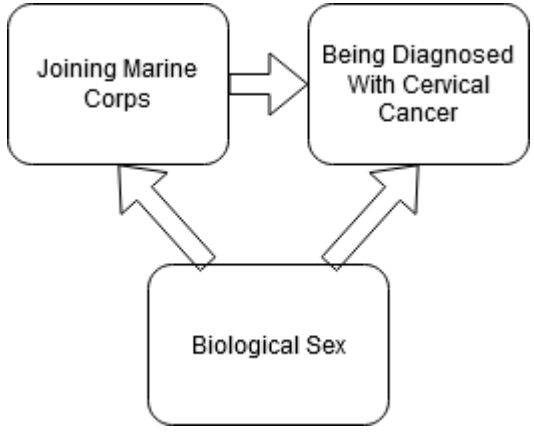
q7: Confounders

For all of these, PLEASE remember: Z is a confounder if (and only if) it has a causal impact on **BOTH** the **independent AND** the **dependent** variable.



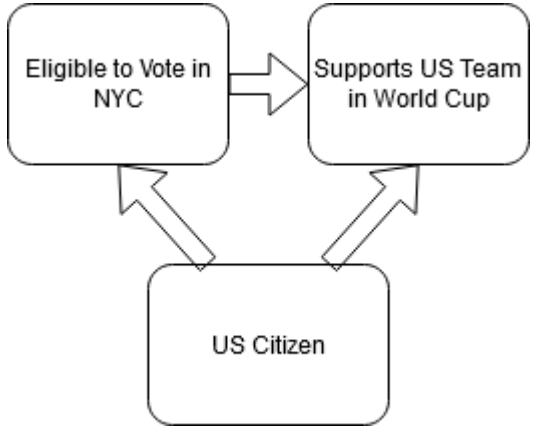
(Q7A) On average, people who join the Marine Corps are less likely than those that don't to be diagnosed with cervical cancer. Before we recommend joining up as a preventive treatment, what's a likely confounder?

A. Military health care plan
B. Exposure to radiation in civilian life
---> C. Biological sex
D. Recruit training at Parris Island



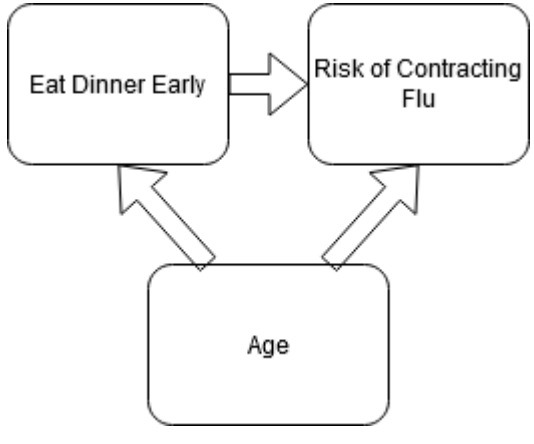
(Q7B) On average, people who can vote in New York City are more likely to support the US team in the World Cup than people who can't. Before we recommend sending out voting materials to improve support for the US soccer team, what is a likely confounder?

A. Interest in soccer
B. Living in NYC
C. Age
---> D. US citizenship status



(Q7C) On average, people who eat dinner earlier in the day tend to be more at risk from flu. Before we suggest waiting until 11pm to ward off sickness, what is a likely confounder?

A. Income
B. Region of country
---> C. Age
D. Degree of danger in occupation



Pandas

Importing a Dataset

In [1]:
```
gift_df = pd.read_csv("ForeignGifts_Universities.csv")
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-1-42a862247680> in <module>
----> 1 gift_df = pd.read_csv("ForeignGifts_Universities.csv")

NameError: name 'pd' is not defined
```

What happened?

In [2]:
```
import pandas as pd
```

In [3]:
```
gift_df = pd.read_csv("ForeignGifts_Universities.csv")
```

Inspecting a Dataset

In [4]:
```
gift_df.head()
```

Out[4]:

	ID	OPEID	Institution Name	City	State	Foreign Gift Received Date	Foreign Gift Amount	Gift Type	Country of Giftor	Giftor Name
0	1	107000	Thunderbird School of Global Management	Glendale	AZ	12/12/2012	$1,220,000.00	Contract	MEXICO	Instituto Technologico y de Estudio
1	2	107000	Thunderbird School of Global Management	Glendale	AZ	11/26/2012	$395,790.00	Contract	CHINA	Intel Products (Chengdu) LTD
2	3	107000	Thunderbird School of Global Management	Glendale	AZ	11/19/2012	$2,769,651.00	Contract	SAUDI ARABIA	Saudi Basic Industries Corporation
3	4	107000	Thunderbird School of Global Management	Glendale	AZ	7/31/2012	$364,128.00	Contract	KUWAIT	Kuwait National Petroleum CO
4	5	108100	Arizona State University	Tempe	AZ	12/29/2017	$180,000.00	Contract	THE NETHERLANDS	Airbus Group SE

In [5]:
```
gift_df['Institution Name'].value_counts()
```

Out[5]:
```
University of California, Los Angeles                          3206
California Institute of Technology                             3127
Johns Hopkins University                                       1157
Columbia University in the City of New York                     791
Ohio State University (The)                                     681
                                                               ...
University of North Texas Health Science Center at Fort Worth     1
Michigan Technological University                                 1
Villanova University                                              1
Colorado School of Mines                                          1
Hult International Business School                                1
Name: Institution Name, Length: 150, dtype: int64
```

In [6]:
```
nyu_df = gift_df[gift_df['Institution Name'] == "New York University"].copy()
nyu_df.head()
```

Out[6]:

	ID	OPEID	Institution Name	City	State	Foreign Gift Received Date	Foreign Gift Amount	Gift Type	Country of Giftor	Giftor Name
14632	14633	278500	New York University	New York	NY	12/22/2017	$350,000.00	Monetary Gift	BERMUDA	Anonymous
14633	14634	278500	New York University	New York	NY	12/15/2017	$325,000.00	Monetary Gift	HONG KONG	Anonymous
14634	14635	278500	New York University	New York	NY	12/13/2017	$1,000,000.00	Monetary Gift	VIRGIN ISLANDS (BRITISH)	Anonymous
14635	14636	278500	New York University	New York	NY	12/8/2017	$1,000,000.00	Monetary Gift	BRAZIL	Anonymous
14636	14637	278500	New York University	New York	NY	11/15/2017	$537,605.00	Contract	THE NETHERLANDS	Government

Cleaning and Organizing a Dataset

In [7]:
```
nyu_df['Foreign Gift Amount'] = nyu_df['Foreign Gift Amount'].str.replace("$","")
nyu_df['Foreign Gift Amount'] = nyu_df['Foreign Gift Amount'].str.replace(",","")
```

In [8]:
```
nyu_df.head()
```

Out[8]:

	ID	OPEID	Institution Name	City	State	Foreign Gift Received Date	Foreign Gift Amount	Gift Type	Country of Giftor	Giftor Name
14632	14633	278500	New York University	New York	NY	12/22/2017	350000.00	Monetary Gift	BERMUDA	Anonymous
14633	14634	278500	New York University	New York	NY	12/15/2017	325000.00	Monetary Gift	HONG KONG	Anonymous
14634	14635	278500	New York University	New York	NY	12/13/2017	1000000.00	Monetary Gift	VIRGIN ISLANDS (BRITISH)	Anonymous
14635	14636	278500	New York University	New York	NY	12/8/2017	1000000.00	Monetary Gift	BRAZIL	Anonymous
14636	14637	278500	New York University	New York	NY	11/15/2017	537605.00	Contract	THE NETHERLANDS	Government

In [9]:
```
nyu_df['amount_numeric'] = nyu_df['Foreign Gift Amount'].astype('float')
```

In [10]:
```
nyu_df.head()
```

Out[10]:

	ID	OPEID	Institution Name	City	State	Foreign Gift Received Date	Foreign Gift Amount	Gift Type	Country of Giftor	Giftor Name	amount_numeric
14632	14633	278500	New York University	New York	NY	12/22/2017	350000.00	Monetary Gift	BERMUDA	Anonymous	350000.0
14633	14634	278500	New York University	New York	NY	12/15/2017	325000.00	Monetary Gift	HONG KONG	Anonymous	325000.0
14634	14635	278500	New York University	New York	NY	12/13/2017	1000000.00	Monetary Gift	VIRGIN ISLANDS (BRITISH)	Anonymous	1000000.0
14635	14636	278500	New York University	New York	NY	12/8/2017	1000000.00	Monetary Gift	BRAZIL	Anonymous	1000000.0
14636	14637	278500	New York University	New York	NY	11/15/2017	537605.00	Contract	THE NETHERLANDS	Government	537605.0

In [11]:
```
nyu_df.sort_values(by=['amount_numeric'], ascending=False, inplace=True)
nyu_df.head()
```

Out[11]:

	ID	OPEID	Institution Name	City	State	Foreign Gift Received Date	Foreign Gift Amount	Gift Type	Country of Giftor	Giftor Name	amount_numeric
14778	14779	278500	New York University	New York	NY	5/1/2013	21783526.00	Contract	UNITED ARAB EMIRATES	Omeir Travel	21783526.0
14645	14646	278500	New York University	New York	NY	7/31/2017	13369447.00	Monetary Gift	SWITZERLAND	Anonymous	13369447.0
14720	14721	278500	New York University	New York	NY	12/11/2014	9999973.00	Monetary Gift	UNITED ARAB EMIRATES	Executive Authority of Abu Dhabi	9999973.0
14702	14703	278500	New York University	New York	NY	1/11/2016	9516904.00	Monetary Gift	CHINA	Anonymous	9516904.0
14706	14707	278500	New York University	New York	NY	1/1/2016	8617263.00	Contract	UNITED ARAB EMIRATES	Anonymous	8617263.0

Variable Types

In [12]:
```
nyu_df.dtypes
```

Out[12]:
```
ID                                int64
OPEID                             int64
Institution Name                 object
City                             object
State                            object
Foreign Gift Received Date       object
Foreign Gift Amount              object
Gift Type                        object
Country of Giftor                object
Giftor Name                      object
amount_numeric                  float64
dtype: object
```

Measures of Frequency/Central Tendency

In [13]:
```
nyu_df['amount_numeric'].mean()
```

Out[13]:
```
1041806.82
```

Measures of Dispersion/Position

In [14]:
```
nyu_df['amount_numeric'].std()
```

Out[14]:
```
2014429.729496327
```

Visualization

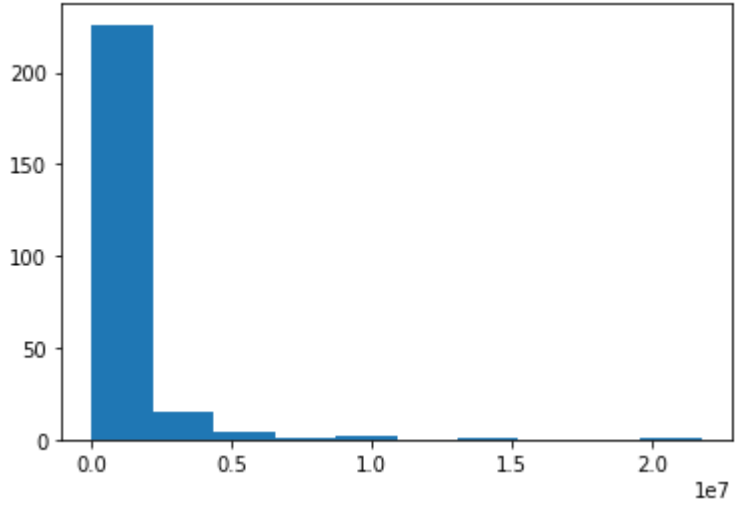
In [15]:
```
plt.hist(nyu_df['amount_numeric'])
plt.show()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-15-9ab9068df8a0> in <module>
----> 1 plt.hist(nyu_df['amount_numeric'])
      2 plt.show()

NameError: name 'plt' is not defined
```

What happened?

In [16]:
```
import matplotlib.pyplot as plt
plt.hist(nyu_df['amount_numeric'])
plt.show()
```



In [17]:
```
plt.boxplot(nyu_df['amount_numeric'], vert=False)
plt.show()
```