D2: Wrangling

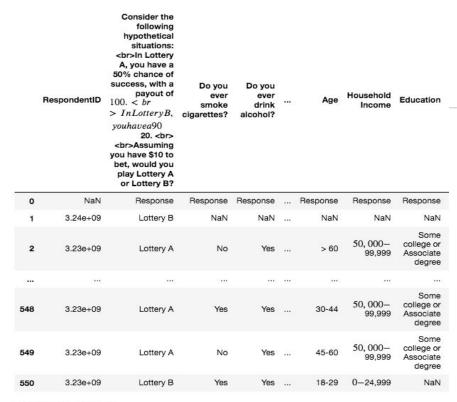


Simple import functions

Import numpy as np Import pandas as pd



survey = pd.read_csv('LINK')



551 rows x 15 columns

Quite simple functions that you can search up on documentation!



Part 2: iloc

 Delete the first row because it is not useful for the data.

Use the function .iloc and use for row 1

survey = survey.iloc[1:]
survey.head()

Res	spondentID	success, with a payout of 100. < br > InLotteryB, youhavea90 20. cbr>Assuming you have \$10 to bet, would you play Lottery A	Do you ever smoke cigarettes?	Do you ever drink alcohol?	
		or Lottery B?			
0	NaN	Response	Response	Response	

Lottery B

NaN

NaN

3.24e+09

Household Income	Age	 ever drink alcohol?	ever smoke cigarettes?	payout of 100. < br > InLotteryB, youhavea90 20. cbr>Assuming you have \$10 to bet, would you play Lottery A or Lottery B?	espondentID	R
NaN	NaN	 NaN	NaN	Lottery B	3.24e+09	1
50,000- 99,999	> 60	 Yes	No	Lottery A	3.23e+09	2



Part 2: List

 Print a list of all the column names in this DataFrame.

list(survey)

```
[ 'RespondentID',
 'Consider the following hypothetical situat
ions: <br > In Lottery A, you have a 50% chanc
e of success, with a payout of $100. <br>In
Lottery B, you have a 90% chance of success,
with a payout of $20. <br>Assuming you h
ave $10 to bet, would you play Lottery A or
Lottery B?',
 'Do you ever smoke cigarettes?',
 'Do you ever drink alcohol?',
 'Do you ever gamble?',
 'Have you ever been skydiving?',
 'Do you ever drive above the speed limit?',
 'Have you ever cheated on your significant
other?',
 'Do you eat steak?',
 'How do you like your steak prepared?',
 'Gender',
 'Age',
 'Household Income',
 'Education'.
 'Location (Census Region)']
```



Part 2: Specifying Dataset

- Who cheats more on their significant other males or females?
- Are cigarette smokers less likely to skydive?
- Do people in New England gamble more than other parts of the country?



Part 2: Drop (.iloc)

 Drop the first two columns from the dataset. This should still be assigned to the variable survey

We choose all rows and drop first 2 columns using slicing

survey = survey.iloc[:,2:]



Part 2: Rename

Assign new column names

```
survey.columns = ['smoking', 'alcohol', 'gambling',
  'skydiving', 'speeding', 'cheated',
  'steak', 'steak_preference', 'gender',
  'age', 'income', 'education', 'region']
```

	smoking	alcohol	gambling	skydiving		age	incom
1	NaN	NaN	NaN	NaN		NaN	Nal
2	No	Yes	No	No		> 60	50,000- 99,99
3	No	Yes	Yes	No	***	> 60	\$150,000
4	Yes	Yes	Yes	No		> 60	50, 000- 99,99
5	No	Yes	No	No		> 60	50, 000- 99,99

5 rows x 13 columns



null_rows = survey.isnull().any(axis=1).sum()

If we print null_rows, we get 217 null values



Part 3: dropna

survey = survey.dropna(how='all')
survey.head()

smoking	alcohol	gambling	skydiving		age	incom
No	Yes	No	No	***	> 60	50, 000- 99,99
No	Yes	Yes	No		> 60	\$150,000-
Yes	Yes	Yes	No		> 60	50, 000- 99,999
No	Yes	No	No		> 60	50, 000- 99,99!
No	No	No	No		18- 29	0-24,99
	No No Yes No	No Yes No Yes Yes Yes Yes No Yes	No Yes No No Yes Yes Yes Yes Yes No Yes No	No Yes No No No Yes Yes No Yes Yes No No Yes No No	No Yes No No No Yes Yes No Yes Yes No No Yes No No	No Yes No No \$0 No Yes Yes No \$0 Yes Yes Yes No \$0 No Yes No No \$0

5 rows x 13 columns

Finished! Work Time and Specific Questions

