D2: Wrangling

1

Simple import functions

Import numpy as np
Import pandas as pd

Part 1

```
survey = pd.read_csv('LINK')
```

	RespondentID	Consider the following hypothetical situations: In Lottery A, you have a 50% chance of success, with a payout of $100. < br > In Lottery B, you have a 90 20. Assuming you have $10 to bet, would you play Lottery A or Lottery B?	Do you ever smoke cigarettes?	Do you ever drink alcohol?	...	Age	Household Income	Education
0	NaN	Response	Response	Response	...	Response	Response	Response
1	3.24e+09	Lottery B	NaN	NaN	...	NaN	NaN	NaN
2	3.23e+09	Lottery A	No	Yes	...	> 60	50, 000– 99,999	Some college or Associate degree
...	...	...	...	...	...	...	...	...
548	3.23e+09	Lottery A	Yes	Yes	...	30-44	50, 000– 99,999	Some college or Associate degree
549	3.23e+09	Lottery A	No	Yes	...	45-60	50, 000– 99,999	Some college or Associate degree
550	3.23e+09	Lottery B	Yes	Yes	...	18-29	0–24,999	NaN

551 rows × 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()
```

	RespondentID	success, with a payout of 100. < br > InLotteryB, youhavea90 20. Assuming you have $10 to bet, would you play Lottery A or Lottery B?	Do you ever smoke cigarettes?	Do you ever drink alcohol?	...
0	NaN	Response	Response	Response	...
1	3.24e+09	Lottery B	NaN	NaN	...

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

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><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 × 13 columns

Part 3: isnull

```
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
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
6	No	No	No	No	...	18-29	0–24,99

5 rows × 13 columns

Finished! Work Time and Specific Questions