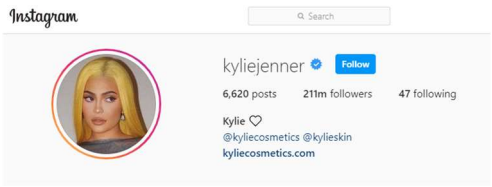
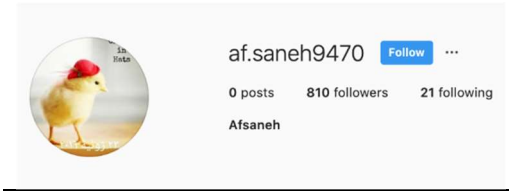
Which one is fake?



Vs.

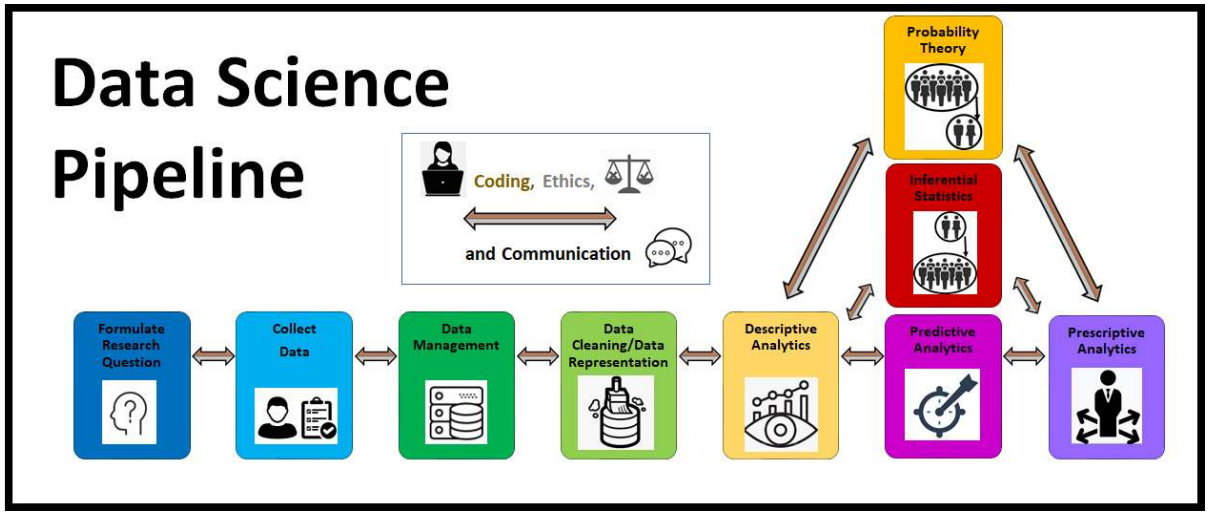


Unit 2: Introduction to the Data Science Pipeline

Case Study:

- *We will explore the relationship between categorial variables in a dataset of fake and real Instagram accounts to demonstrate how we might consider the "full data science pipeline" approach to making informed, business decisions when it comes to identifying fake and real accounts.*

Purpose of this Lecture:



In this lecture we will cover the following topics.

- Case study introduction
- What are two main types of **variables** we could have in a dataset?
- **Types of data science analyses**: question first or dataset first?
- What kind of **research questions can we ask** (and be able to answer) given how this **dataset was collected**?
- Common Python packages, what they are useful for, and how to **import packages**.
- How to read **csvs** into a **Pandas dataframe?**
- Python objects types: **what is a Pandas dataframe?** What are some dataframe functions and attributes that quickly **describe the dataframe**?
- How can we determine what **type of object** an object is in Python?
- What are the two main types of ***basic* descriptive analytics**?
- What are two common types of **summary statistics** we can use to describe **categorical data**?
- What are some common **visualizations** we can use to describe **categorical data**?
- Is there an association between having/not having a profile picture and an Instagram account being fake/not fake IN THIS DATASET?

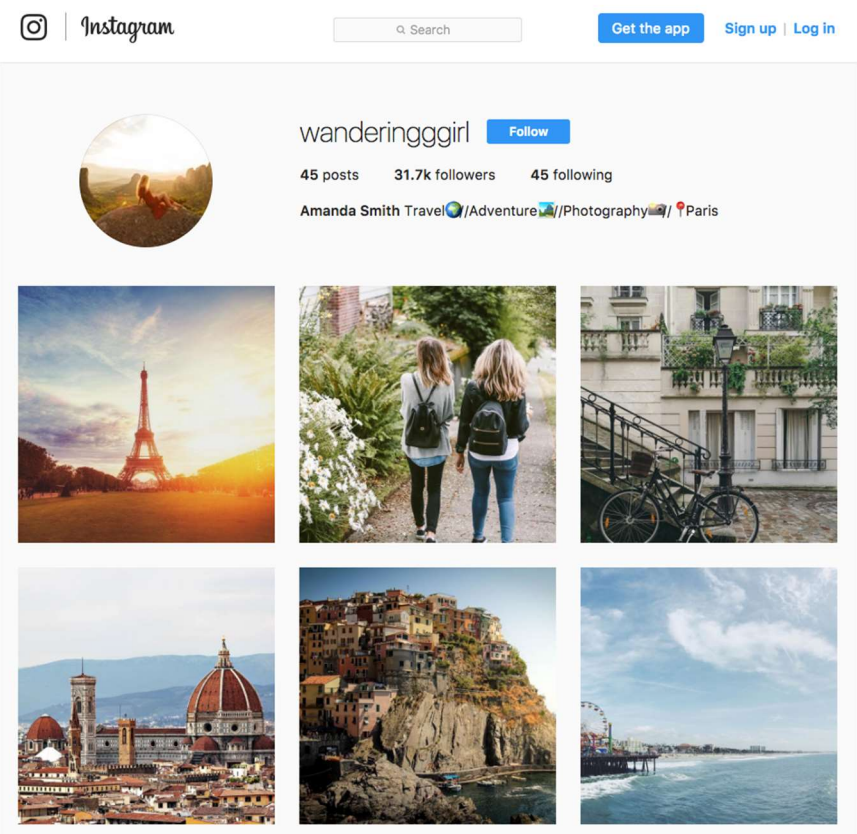
- What are some things that we would want to consider when asking the following question: do we have enough evidence to suggest that having/not having a profile picture and an Instagram account being fake/not fake *OUT OF ALL INSTAGRAM ACCOUNTS?*
- What are some thing we would want to consider, when **building a classifier** to predict whether an Instagram is fake or real?
- How could we use this **classifier to make good business decisions?**

Additional resources:

- Fake vs. real Instagram accounts:
    - https://neoreach.com/how-to-spot-fake-instagram-accounts/
    - https://sproutsocial.com/insights/fake-influencers/

Case Study Introduction

Do you think this is a real or fake Instagram account?



Dataset

Let's assume that this dataset of fake and real Instagram accounts was **randomly sampled** from the **population** of ALL INSTAGRAM ACCOUNTS.

https://www.kaggle.com/free4ever1/instagram-fake-spammer-genuine-accounts

	has_a_profile_pic	number_of_words_in_name	num_characters_in_bio	number_of_posts	number_of_followers	number_of_follows	account_type
0	yes	1	30	35	488	604	real
1	yes	5	64	3	35	6	real
2	yes	2	82	319	328	668	real
3	yes	1	143	273	14890	7369	real
4	yes	1	76	6	225	356	real
...	...	...	...	...	...	...	...
115	yes	1	0	13	114	811	fake
116	yes	1	0	4	150	164	fake
117	yes	2	0	3	833	3572	fake
118	no	1	0	1	219	1695	fake
119	yes	1	0	3	39	68	fake

Research goal:

Data Representation: What are two main types of variables we could have in a dataset?

Definitions

_____ **variable**: can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with

_____ **variable**: values represent categories, where each category is called a _____. It does *not* make sense to add, subtract, or take average with this variable.

Conceptual Quiz

Which of the following is NOT a numerical variable?

a.) Temperature

b.) How many hours a week you spend on Instagram

c.) Numerically coded political data where: 0 = Independent, 1 = Republican, 2 = Democrat.

Additional resources:

- Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php
  - Section 1.2

Research Questions: Types of data science analyses: question first or dataset first?

- **Question first:**

- **Dataset first:**

Research Questions/Data Collection: What kinds of research questions can we ask (and be able to answer) given how a dataset was collected?

Circle all of the following research questions that you think we can feasibly ask and answer with *this particular Instagram dataset.*

a.) Is there an *association* between having/not having a profile picture and being/not being a fake account ***in this dataset?***

b.) Do fake accounts **make it more likely** for the account to not have a profile picture ***in this dataset?***

c.) Is there an *association* between having/not having a profile picture and being/not being a fake account ***in the population of all Instagram accounts***

d.) Do fake accounts **make it more likely** for the account to not have a profile picture ***in the population of all Instagram accounts?***

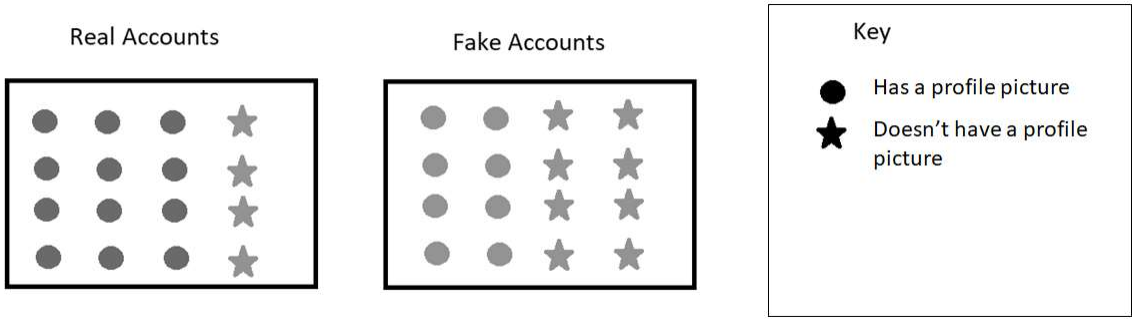
We can use a **sample** of data to **make an inference** about some aspect of a **population** of data when the sample is _____ of the population.

A sample can be _____ of a population when:

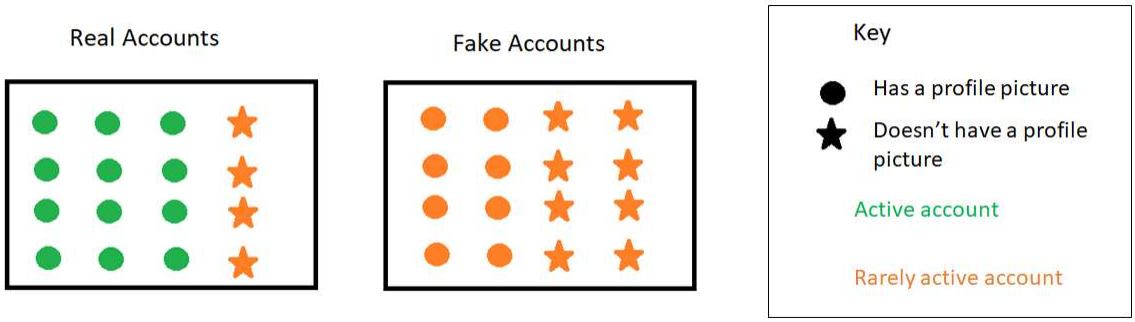
- the sample is _____ collected from the population,
- there is no _____ when the data is collected.

Example: If the person that collected this dataset randomly sampled Instagram accounts that follow Kylie Jenner, would this be a representative sample of all Instagram users (ie. our population of interest)?

Example: If we saw this result in our data, does this show that creating a fake account **causes/makes it less likely for** the creator to not have a profile picture?

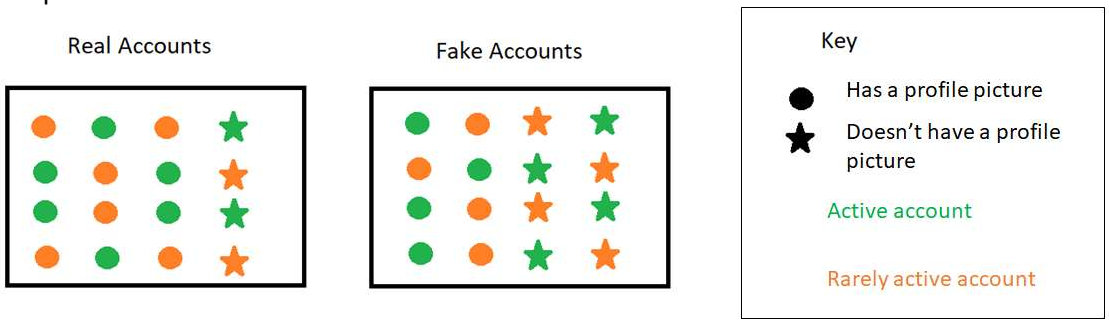


Example: What if Instagram provided us with back-end information about these accounts we discovered this relationship (shown below)? Would we feel more or less confident about saying that creating a fake account **causes/*makes* it less likely** for the creator to not have a profile picture?



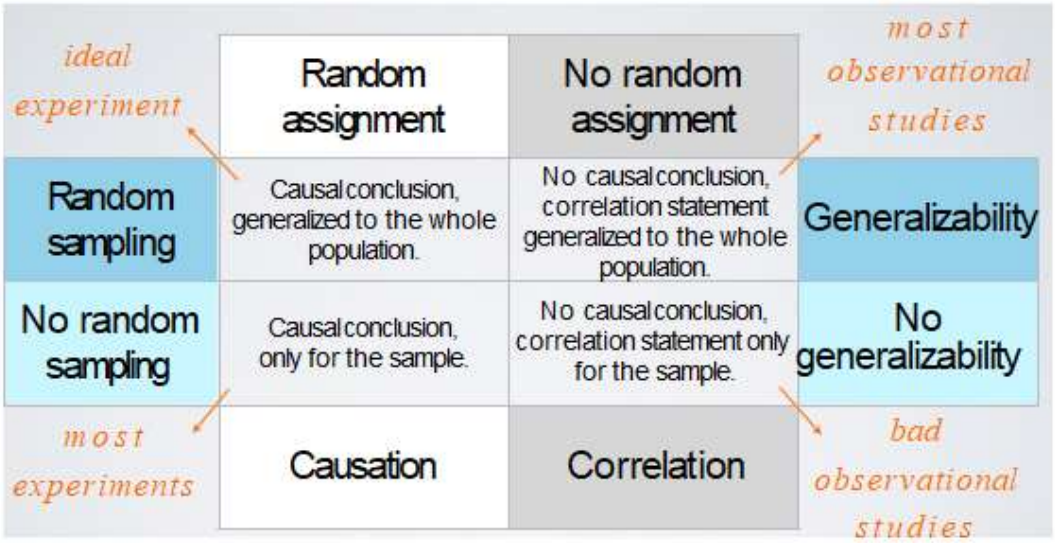
_____ is a **confounding variable** in this analysis because it is an extraneous variable that could affect both the explanatory and the response variable and that make it seem like there's a relationship between them.

Example:
Let's say we randomly selected 100 Instagram users. Then we **randomly assign** 50 of these users to provide us with information about their real accounts, and then told the other 50 to create fake accounts. We then told them to use the account (fake/real) the same way that they would normally. Suppose we got the following results below. Would we feel more or less confident about saying that creating a fake account **causes/*makes* it less likely** for the creator to not have a profile picture?



We can use a **sample** of data to say that one variable in the data **causes** some effect in another variable in the data (ie. we can make a _____), when the data is collected using **random assignment.** This involves randomly assigning observations from the already collected sample (this sample can be a.) random or b.) not random) to each of the levels of the _____ (ie. the variable that you "suspect" causes the other variable (ie. the _____). Using random assignment helps to "wash away" the effects that any _____ might have on the response variable.

Four types of analyses that involve at least two variables.



<u>Additional resources:</u>

- Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php
    - Section 1.3-1.4

Coding: Common Python packages, what are they useful for, and how do you import them?

Python comes with a basic **standard library** which contains a variety of basic functions and attributes. To do anything interesting with Python you will most likely need to download some **packages** which basically provides a collection of modules or _____.

Most common packages:

- **pandas** used for _____
- **matplotlib.pyplot** used for _____
- **seaborn** used for_____
- **numpy** used for _____.

How to import packages

One way: This imports _____ from the pandas package.

```
#Importing all functions from the pandas package
import pandas as pd

#Using any function you want from the package
df = pd.read_csv('fake_insta.csv')
```

One way: This imports _____ from the pandas package.

```
#Importing just the read_csv function from pandas package
from pandas import read_csv

#Using the read_csv function from pandas
df = read_csv('fake_insta.csv')
```

Data Management: How to read csvs into a Python dataframe?

```
1.  df = pd.read_csv('fake_insta.csv')
```

	has_a_profile_pic	number_of_words_in_name	num_characters_in_bio	number_of_posts	number_of_followers	number_of_follows	account_type
0	yes	1	30	35	488	604	real
1	yes	5	64	3	35	6	real
2	yes	2	82	319	328	668	real
3	yes	1	143	273	14890	7369	real
4	yes	1	76	6	225	356	real
...	...	...	...	...	...	...	...
115	yes	1	0	13	114	811	fake
116	yes	1	0	4	150	164	fake
117	yes	2	0	3	833	3572	fake
118	no	1	0	1	219	1695	fake
119	yes	1	0	3	39	68	fake

Data Representation: What is a pandas dataframe? What are some dataframe functions and attributes that quickly describe the dataframe?

Dataframe a 2-dimensional labeled data structure with columns of potentially different **types.**

I. Display the **whole dataframe** (usually abbreviated in the middle if it's large).

```
1.  #Display the dataframe object that we called df
2.  df
```

	has_a_profile_pic	number_of_words_in_name	num_characters_in_bio	number_of_posts	number_of_followers	number_of_follows	account_type
0	yes	1	30	35	488	604	real
1	yes	5	64	3	35	6	real
2	yes	2	82	319	328	668	real
3	yes	1	143	273	14890	7369	real
4	yes	1	76	6	225	356	real
...	...	...	...	...	...	...	...
115	yes	1	0	13	114	811	fake
116	yes	1	0	4	150	164	fake
117	yes	2	0	3	833	3572	fake
118	no	1	0	1	219	1695	fake
119	yes	1	0	3	39	68	fake

II. General structure of the dataframe

III. Visualize just the **first few rows** of the dataframe.

```
1. #Display the first 5 rows (5=default)
2. df.head()
```

	has_a_profile_pic	number_of_words_in_name	num_characters_in_bio	number_of_posts	number_of_followers	number_of_follows	account_type
0	yes	1	30	35	488	604	real
1	yes	5	64	3	35	6	real
2	yes	2	82	319	328	668	real
3	yes	1	143	273	14890	7369	real
4	yes	1	76	6	225	356	real

```
1. #Display the first 12 rows
2. df.head(12)
```

	has_a_profile_pic	number_of_words_in_name	num_characters_in_bio	number_of_posts	number_of_followers	number_of_follows	account_type
0	yes	1	30	35	488	604	real
1	yes	5	64	3	35	6	real
2	yes	2	82	319	328	668	real
3	yes	1	143	273	14890	7369	real
4	yes	1	76	6	225	356	real
5	yes	1	0	6	362	424	real
6	yes	1	132	9	213	254	real
7	yes	2	0	19	552	521	real
8	yes	2	96	17	122	143	real
9	yes	1	78	9	834	358	real
10	yes	1	0	53	229	492	real
11	yes	1	78	97	1913	436	real

IV. Print the number of **rows** and **columns** in a dataframe.

```
1. df.shape
```

```
(120, 7)
```

V. Print all of the **variables** (**columns** in the dataframe).

```
1. print(df.columns.values)
```

```
['has_a_profile_pic' 'number_of_words_in_name' 'num_characters_in_bio'
 'number_of_posts' 'number_of_followers' 'number_of_follows'
 'account_type']
```

VI. Print the **index of all the rows** in the dataframe.

```
1. df.index
```

```
RangeIndex(start=0, stop=120, step=1)
```

Coding: How can we determine what type of object an object is in Python?

Almost everything in Pyhon is an object. We just created an **object** that we called df**.** We claimed that this **type of object** was a **pandas dataframe object,** but let's double check using the **type()** function.

```
1.  type(df)
```

```
pandas.core.frame.DataFrame
```

Different types of objects in Python will have different types of **attributes** and **functions** that correspond to it. It's important to know what type of object you're dealing with.

Descriptive Analytics: What are the two main types of "basic" descriptive analytics?

- **Summary statistics:** a _____ used to summarize a set of observations.

- **Visualizations:** a _____ used to summarize a set of observations.

Descriptive Analytics: What are two common types of summary statistics we can use to describe a single categorical variable?

- **Counts** of each _____ of a single categorical variable.

```
df['has_a_profile_pic'].value_counts()
```

```
yes    91
no     29
Name: has_a_profile_pic, dtype: int64
```

- **Proportions** of each _____ of a single categorical variable.

```
df['has_a_profile_pic'].value_counts(normalize=True)
```

```
yes    0.758333
no     0.241667
Name: has_a_profile_pic, dtype: float64
```

Descriptive Analytics: What is a common visualization we can use to describe a single categorical variable?

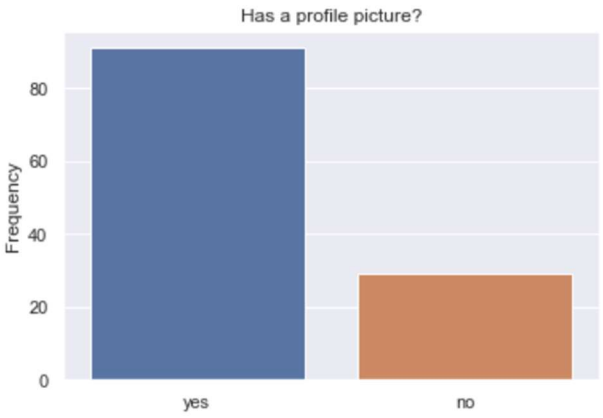
Barplot with frequency (counts)

```
#First we create a pandas series object called 'has_pic_counts'
has_pic_counts = df['has_a_profile_pic'].value_counts()
display(counts.shape, counts)
```

```
(2,)

yes    91
no     29
Name: has_a_profile_pic, dtype: int64
```

```
sns.barplot(x=has_pic_counts.index, y=has_pic_counts)
plt.title('Has a profile picture?')
plt.ylabel('Frequency')
plt.show()
```

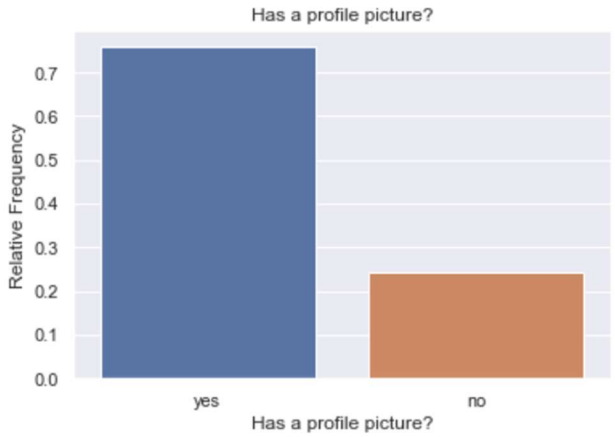


Barplot with relative frequency (percentages)

```
1. #We create a new pandas series object called 'has_pic_counts_perc'
2. has_pic_counts_perc = df['has_a_profile_pic'].value_counts(normalize=True)
3. display(has_pic_counts_perc.shape, has_pic_counts_perc)
```

```
yes    0.758333
no     0.241667
Name: has_a_profile_pic, dtype: float64
```

```
1. sns.barplot(x=has_pic_counts_perc.index, y=has_pic_counts_perc)
2. plt.title('Has a profile picture?')
3. plt.xlabel('Has a profile picture?')
4. plt.ylabel('Relative Frequency')
5. plt.show()
```



Descriptive Analytics: What are two common types of summary statistics we can use to describe at least two categorical variables?

I. **Counts** of each _____ of levels for two (or more categorical variables).

```
1. pd.crosstab(df['account_type'], df['has_a_profile_pic'])
```

has_a_profile_pic	no	yes
account_type		
fake	29	31
real	0	60

II. **Percentages** of observations that are of each level type of variable 2 for each level of variable 1

```
1. pd.crosstab(df['account_type'], df['has_a_profile_pic'], normalize='index')
```

has_a_profile_pic	no	yes
account_type		
fake	0.483333	0.516667
real	0.000000	1.000000

```
1. pd.crosstab(df['has_a_profile_pic'], df['account_type'], normalize='index')
```

account_type	fake	real
has_a_profile_pic		
no	1.000000	0.000000
yes	0.340659	0.659341

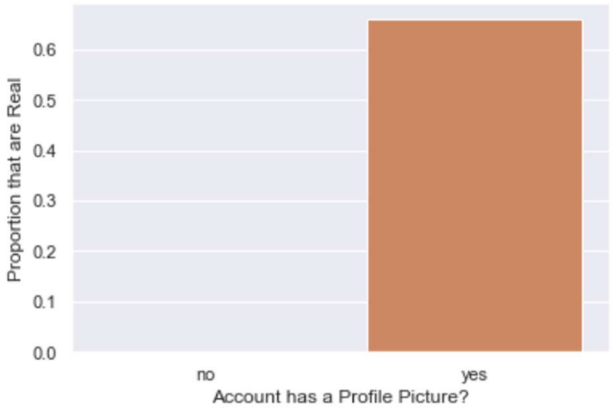
Descriptive Analytics: What are two common visualizations we can use to summarize at least two categorical variables?

I. **Single barplot:** Visualize each level of first variable on the x-axis, visualize the proportion that are of just one type of level of the second variable on the y-axis.

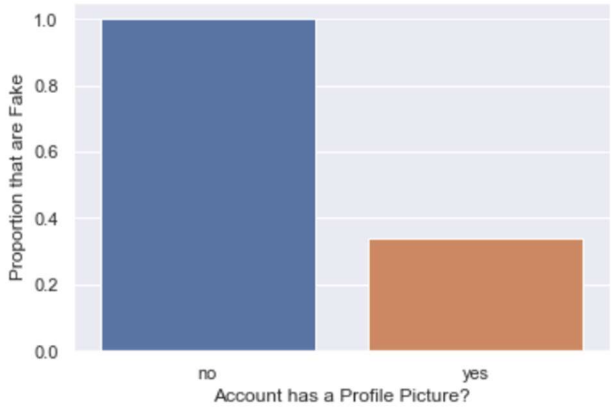
```
1. temp = pd.crosstab(df['has_a_profile_pic'], df['account_type'], normalize='index')
2. temp
```

account_type	fake	real
has_a_profile_pic		
no	1.000000	0.000000
yes	0.340659	0.659341

```
# graph the normalized crosstabs
sns.barplot(x=temp.index, y="real", data=temp)
plt.ylabel("Proportion that are Real")
plt.xlabel('Account has a Profile Picture?')
plt.show()
```




```
sns.barplot(x=temp.index, y="fake", data=temp)
plt.ylabel("Proportion that are Fake")
plt.xlabel('Account has a Profile Picture?')
plt.show()
```

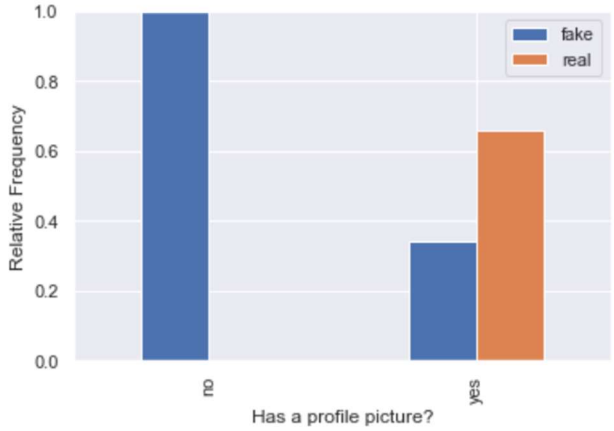


II. **Side-by-side barplots:** Visualize each level of first variable on the x-axis, visualize the proportion of each level of the second variable on the y-axis.

```
temp = pd.crosstab(df['has_a_profile_pic'], df['account_type'], normalize='index')
temp
```

account_type	fake	real
has_a_profile_pic		
no	1.000000	0.000000
yes	0.340659	0.659341

```
temp.plot.bar()
plt.legend(loc='upper right')
plt.xlabel('Has a profile picture?')
plt.ylabel('Relative Frequency')
plt.ylim([0,1])
plt.show()
```




```
temp2 = pd.crosstab(df['account_type'], df['has_a_profile_pic'], normalize='index')
temp
```

account_type	fake	real
has_a_profile_pic		
no	1.000000	0.000000
yes	0.340659	0.659341

```
temp2.plot.bar()
plt.legend(loc="upper center")
plt.xlabel('Account type')
plt.ylabel('Relative Frequency')
plt.show()
```

DESCRIPTIVE ANALYTICS/RESEARCH QUESTION: Is there an association between the variable of having/not having a profile picture and the variable of fake/real Instagram accounts IN THIS DATASET?

Probability:

What is the probability that a randomly selected account with a profile picture (in this dataset) is fake?

What is the probability that a randomly selected account without a profile picture (in this dataset) is fake?

INFERENCE: Do we have enough evidence to suggest that there an association between the variable of having/not having a profile picture and the variable of fake/real Instagram accounts IN THE POPULATION OF ALL INSTAGRAM ACCOUNTS?

What are some things we would want to consider when answering this question?

PREDICTIVE ANALYTICS: Building a basic classifier

has_a_profile_pic	no	yes
account_type		
fake	0.483333	0.516667
real	0.000000	1.000000

has_a_profile_pic	no	yes
account_type		
fake	29	31
real	0	60

Example: Using this analysis, suppose we built a basic fake/real Instagram classifier using the following rule/model.

Classifier Model:

- If the Instagram account doesn't have a profile picture, label it as fake.
- If the Instagram account does have a profile picture, label it as real.

How well would this classifier do at classifying the real accounts in this dataset? Try to numerically quantify this performance.

How well would this classifier do at classifying the fake accounts in this dataset? Try to numerically quantify this performance.

How well would this classifier do at classifying all the accounts in this dataset? Try to numerically quantify this performance.

PRESCRIPTIVE ANALYTICS: Using your analysis to make the "best" decision.

Which of the following two "clients" would you be the best to try to "sell" this classifier model to?

- **Client 1: This client would like to identify as many fake accounts as possible, but suffers a high financial penalty for any real accounts that are labeled "fake."**

- **Client 2: This client would like select a group of Instagram accounts that they can feel very confident are real accounts.**