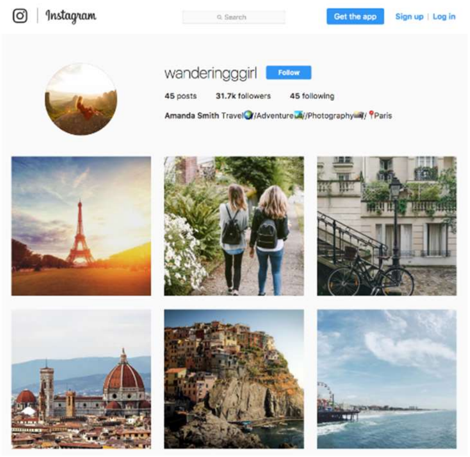
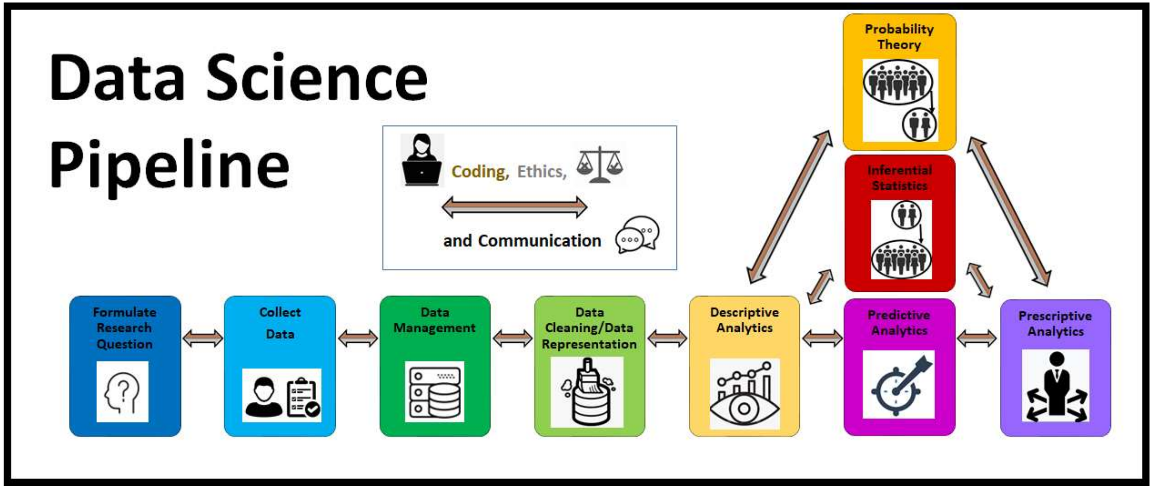
Unit 4: Descriptive Analytics for Numerical Variables



***Case Study*: Fake and Real Instagram Accounts**

- *We will use descriptive analytics to compare the distributions of the number of accounts followed by fake and real Instagram accounts.*
- *Is there an association between the number of posts made by fake and real Instagram accounts in this dataset?*

Purpose of this Lecture:



Data Science Pipeline

Coding, Ethics, and Communication

Formulate Research Question

Collect Data

Data Management

Data Cleaning/Data Representation

Descriptive Analytics

Probability Theory

Inferential Statistics

Predictive Analytics

Prescriptive Analytics

In this lecture we will cover the following topics.

Case Study: Is there an association between number of follows and fake and real Instagram accounts *in the dataset*?

1. Missing Data Checking: How do we check for missing data observations *that have not yet been "detected" and "coded" as NaN yet?*
2. Visualizations for a Single Numerical Variable: What are three plots that we can use to visualize the distribution of a single numerical variable?
    1. Histograms
    2. Boxplots
    3. Violin plots
3. Histograms: Types of histograms
4. Using Frequency Histograms: How to estimate the proportion of observations that are within a given range?
5. Describing a Single Numerical Variable Distribution: What are four things we should be ready to describe about the distribution of a numerical variable?
    1. Shape
        i. Modality

ii. Skew

2. Measure of Center (Summary Statistics)
   i. Mean
   ii. Median
   iii. When to use mean vs. median?
3. Measure of Spread (Summary Statistics)
   i. Standard deviation
   ii. IQR
   iii. Range
   iv. When to use standard deviation vs. IQR vs. range?
4. Any outliers?

6. Coding: Create a function in Python
7. Coding: Using a function that we have created in Python.
8. Dataframe Manipulation: Subsetting a dataframe with indices that have names (not numbers).
9. Boxplots:
   1. How to make one by hand?
   2. How to interpret?
10. Violin Plots:
    1. How to interpret?
    2. Violin plot vs. boxplot vs. histogram
11. Categorical and Numerical Variable: How to visualize the relationship between a numerical and categorical variable.
12. Putting it all together: Is there an association between number of follows and fake and real Instagram accounts *in the dataset*?

Additional resources:

- Chapter 4 in J. VanderPlas (2016) *Python Data Science Handbook*
  - https://jakevdp.github.io/PythonDataScienceHandbook/04.05-histograms-and-binnings.html
  - https://jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html

1. Missing Data Checking: How do we check for missing data observations *that have not been "detected" or "coded" as NaN yet?*

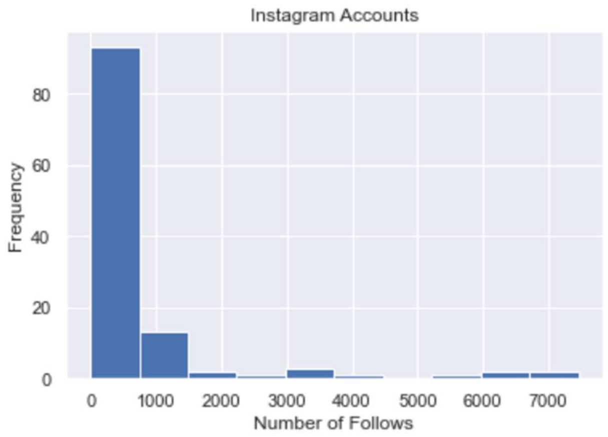
****Go to Unit 4, section 1 Jupyter notebook for explanation and code.****

2. Visualizations for a Single Numerical Variable: What are three plots that we can use to visualize the distribution of a single numerical variable?

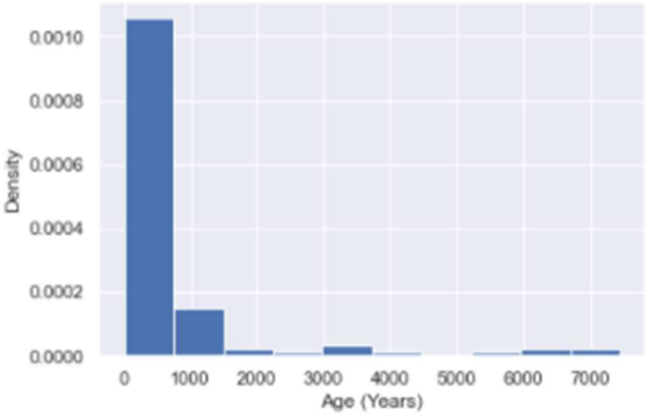
3. Histograms: Types of Histograms

****See Unit 4, section 3 Jupyter notebook for code****

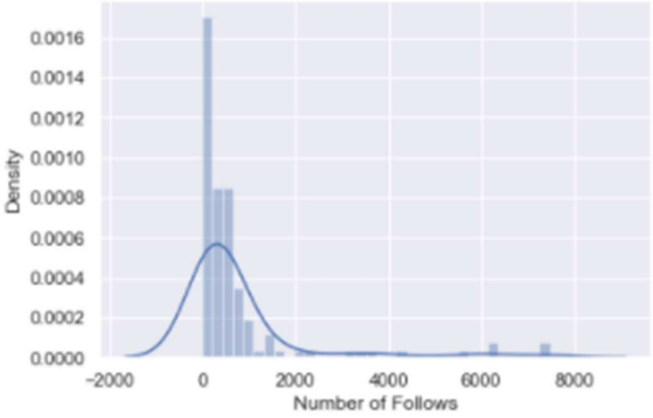
3.1 Frequency Histogram



3.2 Density Histogram

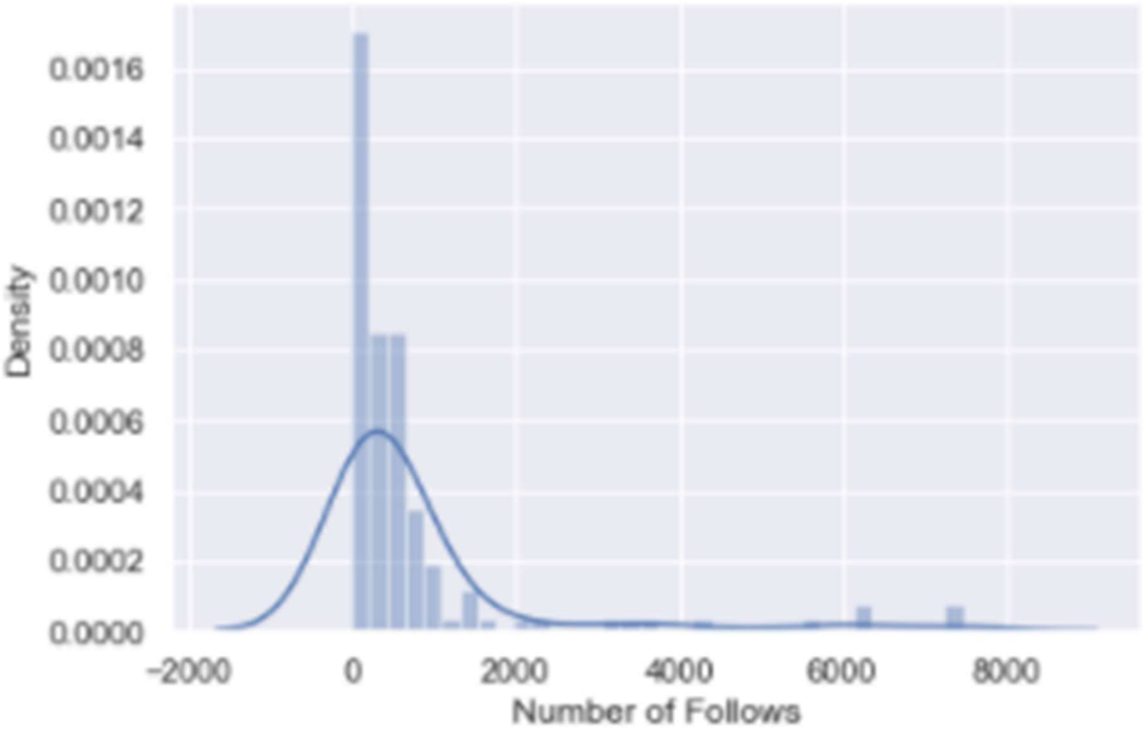


3.3 Density Histogram Fitted with a Density Curve



4. Using Frequency Histograms: How to estimate the proportion of observations that are within a given range?

Ex: Estimate the proportion of accounts in the dataset that follow between 0 to 400 accounts.



Way 1:

1. Find the boxes that are approximately within this range.
2. Add up the areas of these boxes.

Way 2:

1. Find the area underneath of the density curve in this range.

5. Describing a Single Numerical Variable Distribution: What are four things we should always be ready to describe about the distribution of a numerical variable?

1. Shape
    - i. Modality
    - ii. Skew

2. Measure of Center (Summary Statistics)
    - **i.** Mean ****See Unit 4, section 5 Jupyter notebook for code****
    - i. Median ****See Unit 4, section 5 Jupyter notebook for code****
    - *ii. When to use mean vs. median?*

3. Measure of Spread (Summary Statistics)
    - iii. Standard deviation ****See Unit 4, section 5 Jupyter notebook for code****
    - iv. IQR ****See Unit 4, section 5 Jupyter notebook for code****
    - v. Range ****See Unit 4, section 5 Jupyter notebook for code****
    - *vi. When to use standard deviation vs. IQR vs. range?*

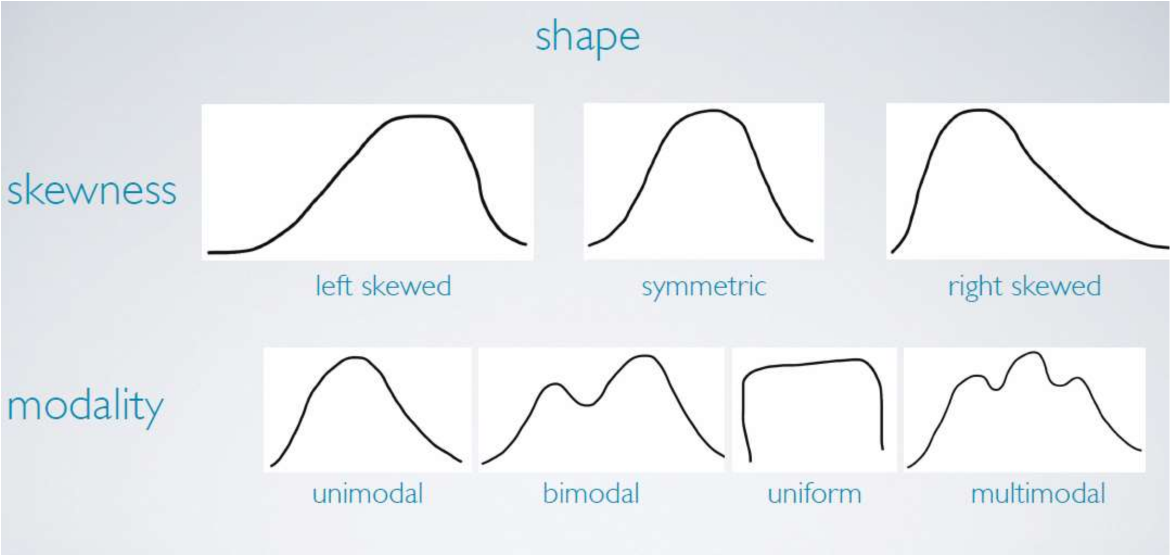
4. Any outliers?

5.1. Shape of a Distribution

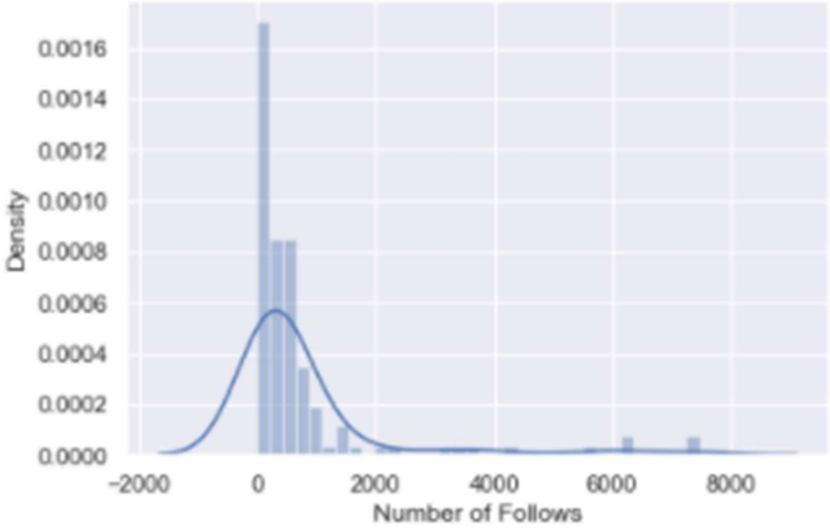
Shape of the distribution

a) Modality: unimodal, bimodal, multimodal, or uniform

b) Skewness: symmetric, left-skewed, or right-skewed



Ex: Describe the shape of the number of follows distribution.



5.2. Measures of Center of a Distribution

Measures of Center: an estimate of a typical observation in the data

i. **Mean:** average ****See Unit 4, section 5 Jupyter notebook for code****

- **Best to use when**:

i. **Median: **See Unit 4, section 5 Jupyter notebook for code****

- **observation that is higher than 50% of the data and lower than 50% of the data**
- **Best to use when**:

Ex: Sorted Number of Hours Spent Studying:
[1, 1, 1, 2, 4, 5, 5, 6, 7]

Odd # of observations?

- **Median** = middle number

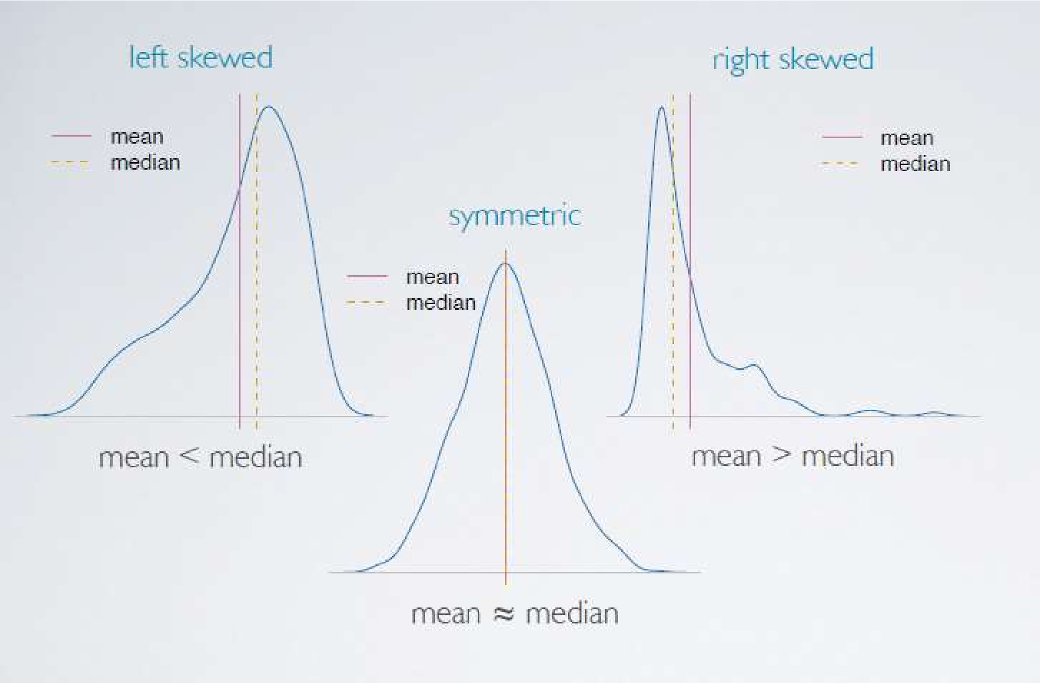
Ex: Sorted Number of Hours Spent Studying:
[1, 1, 1, 2, 3, 5, 5, 6, 7, 7]

Avg(3,5) = 4

Odd # of observations?

- **Median** = average of middle two numbers

ii. **How to Mean and Median Interact:**



5.3. Measures of Spread of a Distribution

Spread: a measure of variability of the data ****See Unit 4, section 5 Jupyter notebook for code****

a) **Standard Deviation (of a Population)** = $\sqrt{\frac{(x_1-mean)^2+(x_2-mean)^2+\cdots(x_n-mean)^2}{n}}$

- **What it represents:**

- **Best to use when:**

b) **IQR = Q3 – Q1**

- **What it represents:**

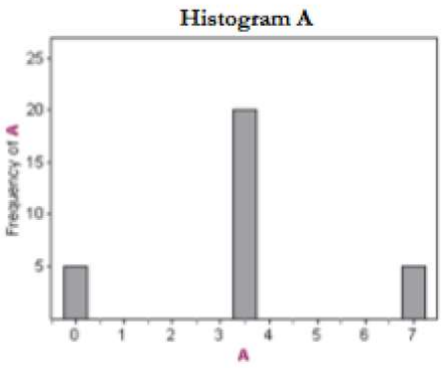
- **Best to use when:**

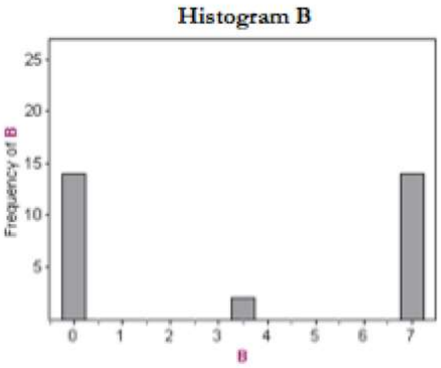
c) **Range = max- min**

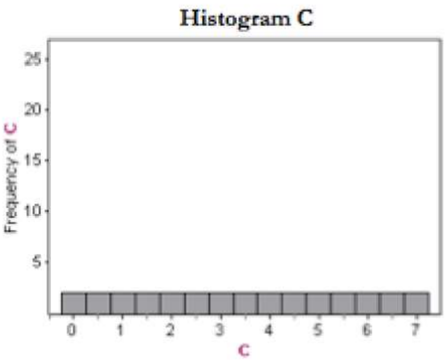
- **What it represents:**

- **Best to use when:**

Ex: Which of the following three datasets (each plotted in a histogram below) has the highest standard deviation? Which has the lowest standard deviation?







5.4. Any outliers?

How to **classify an outlier** from a single numerical variable. ****See Unit 4, section 5 Jupyter notebook for code****

A point is an outlier if either:

- it is $> Q3 + 1.5(IQR)$ or
- it is $< Q1 - 1.5(IQR)$

6. Coding: Create a function in Python

****See Unit 4, section 6 Jupyter notebook for code****

7. Coding: Using a function that we have created in Python.

****See Unit 4, section 7 Jupyter notebook for code****

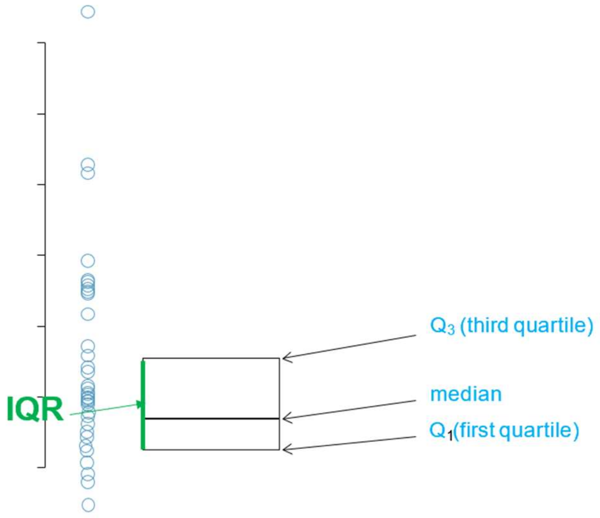
8. Dateframe Manipulation: Subsetting a Dataframe with Indices that have names (not numbers).

****See Unit 4, section 8 Jupyter notebook for code****

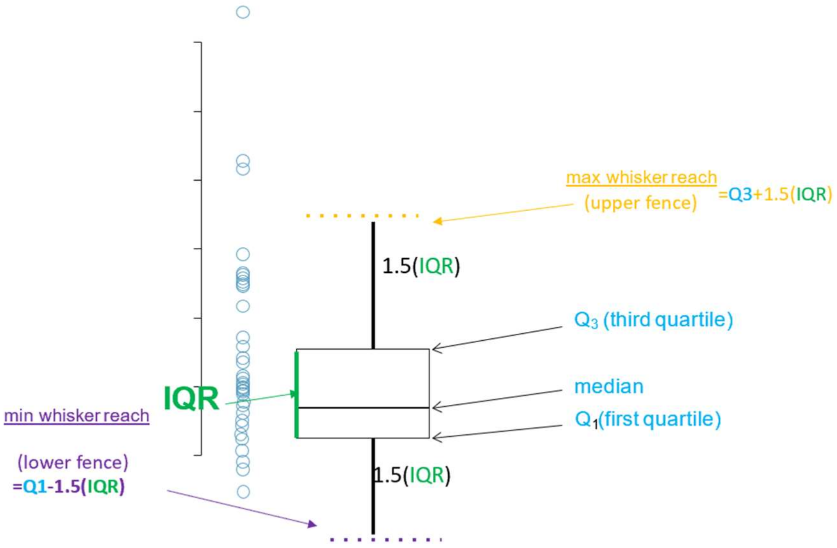
9. Boxplots: How to create one "by hand"? How to interpret?

****See Unit 4, section 9 Jupyter notebook for code****

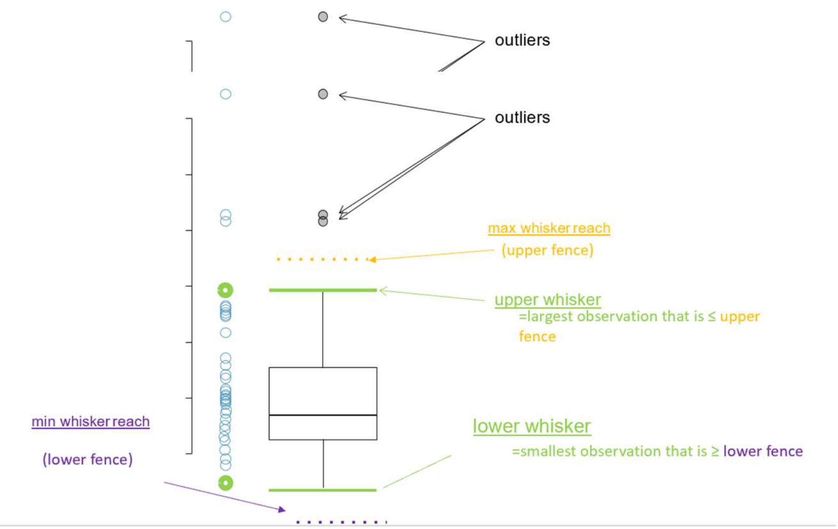
1. Sort the data
2. Calculate the median, Q1, Q3, and IQR.
3. **Draw a thick line at the median.**
4. **Draw two** other lines at the **Q1** and **Q3** value.
5. Make a box connecting these three lines.



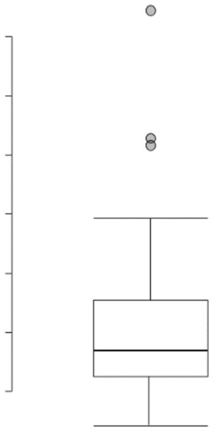
6. Draw a temporary line at **Q3+1.5(IQR)** (called the **max whisker reach)**
7. Draw a temporary line at **Q1-1.5(IQR)** (called the **min whisker reach)**



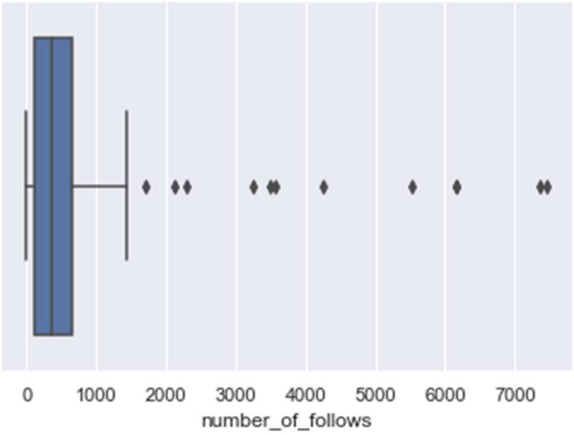
8. Draw dots for any points above the max whisker reach. These are **outliers.**
9. Draw dots for any points below the min whisker reach. These are **outliers.**
10. Draw a solid line at largest point that is less than or equal to the max whisker reach. This is the **upper whisker.**
11. Draw a solid line at smallest point that is greater than or equal to the min whisker reach. This is the **lower whisker.**



12. Erase the max whisker reach and min whisker reach lines.



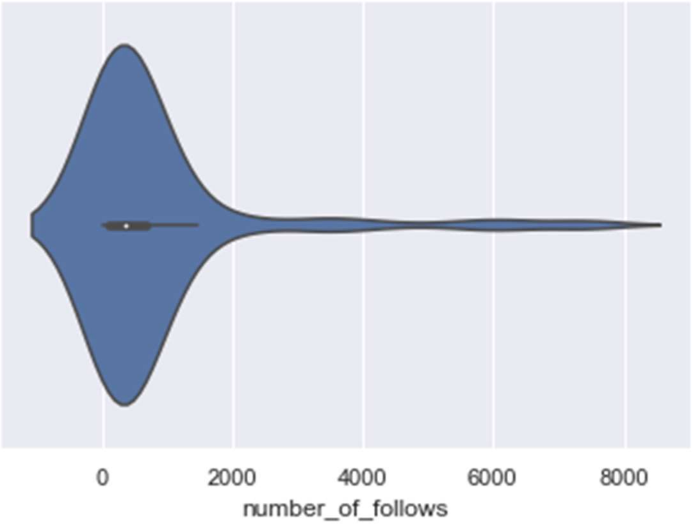
Ex: What is the **skew** of this distribution? Can we determine the **modality** of this distribution by using a boxplot?



10. Violinplots

****See Unit 4, section 10 Jupyter notebook for code****

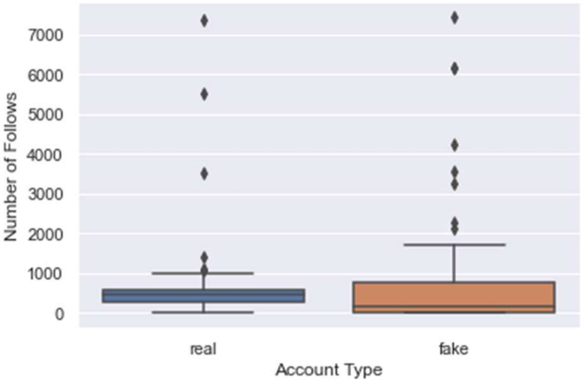
Ex: What is the **skew** and **modality** of this distribution?

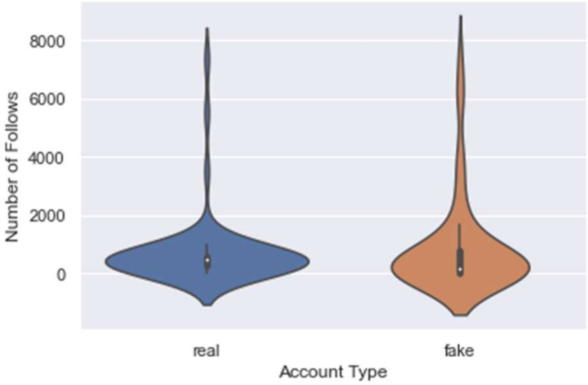


Distinction: What is something that we could identify with a boxplot, but not with a violin plot?

11. Categorical and Numerical Variable: How to visualize the relationship between a numerical and categorical variable.

****See Unit 4, section 11 Jupyter notebook for code****





12. Putting it all together: Is there an association between number of people that fake and real Instagram accounts follow in this dataset?

1. Compare the distribution the numbers of people that real and fake accounts follow in this dataset.

2. Is there a STRONG association between number of people that fake and real Instagram accounts follow in this dataset?