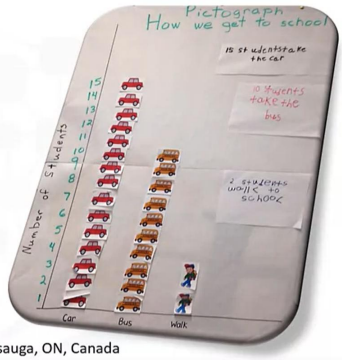
# Week 2 - Data Visualization

## 1. Visualization Fundamentals

Visualization Fundamentals




Kindergarten students at
St. Luke Catholic School, Mississauga, ON, Canada

### Variable type Matters

Chart types should be based on the type of variables being depicted

Examples:

- Categorical variables:
  - Counts, e.g., How many instructors are male and how many are female
- Continuous variables:
  - Scatter plots
  - Statistical properties (averages)

Let us begin with the fundamentals of visualization. What you see on the right side is a picture graph or a bar chart prepared by students in kindergarten class, four-year-olds at the St. Luke Catholic school in Mississauga. This bar chart presents the frequency for transportation modes the kids have used to come to school. Some have been dropped to school by car, others by bus, and two students walked to school. In the brave, big world of big data, we can see that children are being trained at the earliest possible age with data and data science.

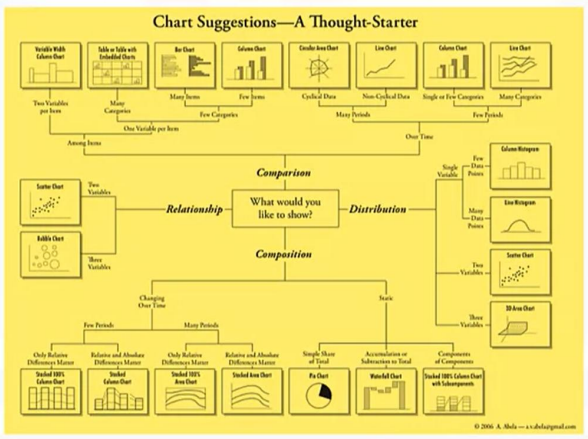
The most important thing to realize is that the type of visualization you will use depends upon the type of variables you're trying to analyze. For instance, if you're working with categorical variables such as gender, you have to rely on a certain type of charting tools than if you were to be working with continuous variables such as age and income. In one case, you may be using bar charts, in another case, you will be required to use scatter plots.

### The Extreme Presentation Method

Dr. Andrew Abela:

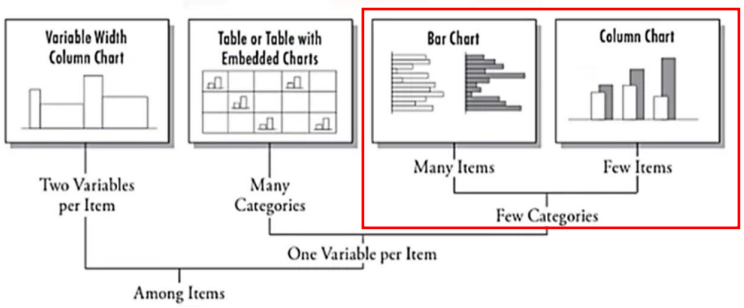
- The Extreme Presentation method is a step-by-step approach for designing presentations of complex or controversial information in ways that drive people to action.
- https://extremepresentation.com/

### Chart suggestions

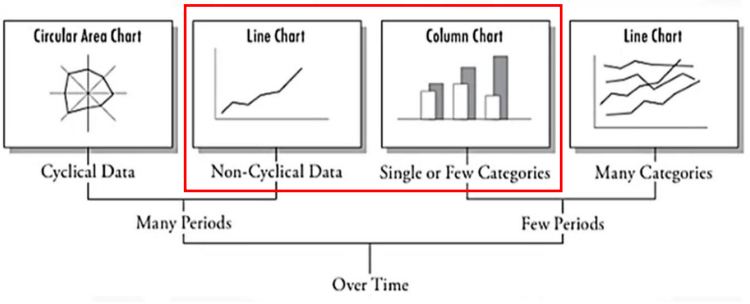



I would like to draw your attention to the Extreme Presentation Method developed by Dr. Andrew Abela. Essentially, it lays out the possible ways of depicting data based on what kind of variables you have at your disposal. This visualization, this graphic is developed by Dr. Abela that shows that if you are interested in comparing variables or demonstrating their distribution or composition or the relationship between two variables or more, you have to rely on specific type of graph.

### Comparison – among items




### Comparison – over time

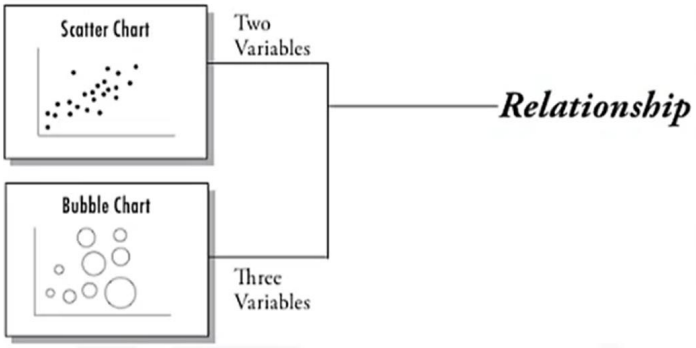



If you're comparing items with few categories, you can use **bar charts** or **column charts**.

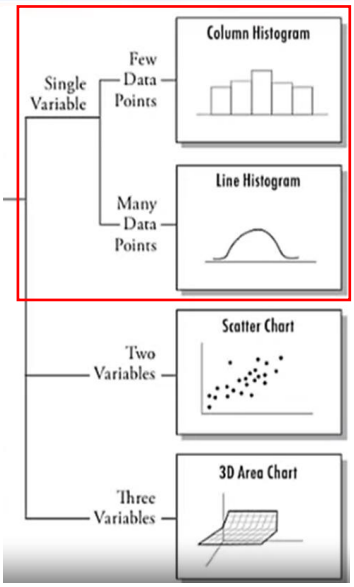
If you are comparing behaviors over time and if you have the time periods running into several months, you may have to use a **line chart**. If the time periods are not that many, then you can use **columns** and other approaches.

**总结:**

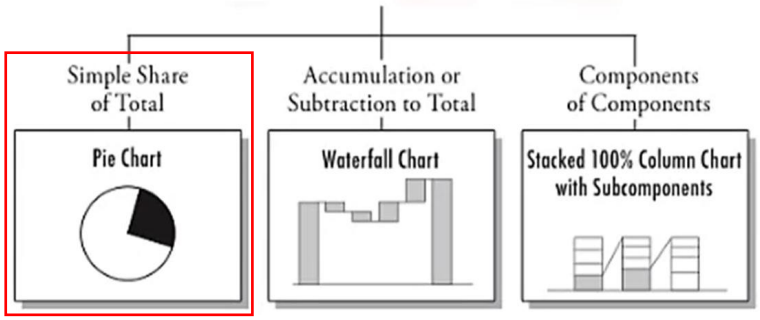
## Relationship



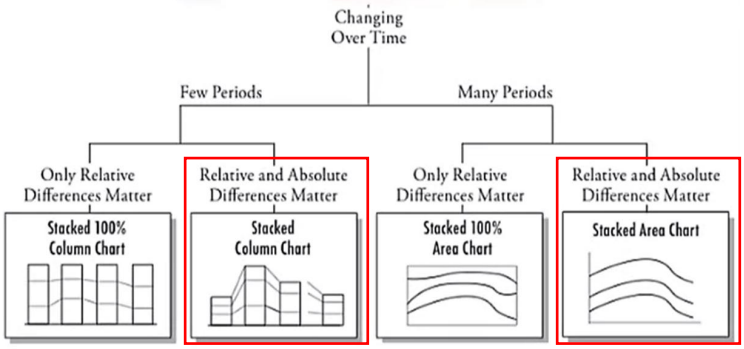

## Distribution




## Composition - Static




## Composition – Changing over time




If you're trying to depict relationships between two continuous variables, your choice is a **scatterplot**. The **bubble chart** will depict two variables on x and y axis and the third variable will be depicted by the size of the circle. So you can essentially have three variables.

If you're interested in depicting the distribution of the dataset, you can use the **histogram**, which could be a bar chart type of histogram or a line histogram. The distribution could also be shown through scatter plots.

If you would like to show the composition, then if it's a static data, you can use **pie charts**.
If you're showing the composition that changes over time for few periods, you can use **stacked columns**. If you have several periods, you can use the **stacked area charts**.

For hands-on in Python, we will be using the Seaborn library and the Matplotlib library to create visualizations in the labs. We will learn how to use different functions within the library to create different kinds of charts.

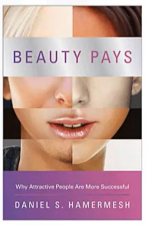
**总结:**

## 2. Statistics by Groups



### Statistics by groups

- We are often interested in how a statistic differs between groups in our data set
- For instance, do teaching evaluations and beauty scores differ by:
  - Gender
  - Tenure
  - English proficiency
  - Visible minority status
- Data set:
  - University of Texas
  - Survey data from 463 courses
  - Teaching evaluations of instructors




We have learned to compute averages and standard deviations, but now we will use the same information of same knowledge to make comparisons between groups.
So we will use the same dataset that we have used so far. That is the teaching evaluation data from University of Texas, comprising 463 courses.

### Summary statistics for groups

Case Summaries

| | gender | eval | | | beauty | | | age | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | mean | std | var | mean | std | var | mean | std | var |
| 0 | female | 3.901026 | 0.538803 | 0.290308 | 0.116109 | 0.81781 | 0.668813 | 45.092308 | 8.532031 | 72.795559 |
| 1 | male | 4.069030 | 0.556652 | 0.309861 | -0.084482 | 0.75713 | 0.573246 | 50.746269 | 9.993396 | 99.867964 |

| | tenure | eval | | | beauty | | | age | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | mean | std | var | mean | std | var | mean | std | var |
| 0 | no | 4.133333 | 0.556747 | 0.309967 | 0.028359 | 0.876656 | 0.768525 | 50.186275 | 6.946372 | 48.252087 |
| 1 | yes | 3.960111 | 0.549104 | 0.301516 | -0.008013 | 0.763074 | 0.582282 | 47.850416 | 10.420056 | 108.577562 |

### Let's think for a second

- Average teaching evaluation is an attribute of the course
  - There are 463 courses
- Average age or beauty is an attribute of the instructor
  - There are 94 instructors in the data set who taught 463 courses
- Eliminate 'duplicates' to avoid repeated measure for the age and beauty variable
- Identify duplicates

```
1 no_duplicates_ratings_df = ratings_df.drop_duplicates(subset =['prof'])
```

We are looking at the teaching evaluation, beauty, and age. We're comparing the averages for these three variables for female instructor variable. So those who are females, their average teaching evaluation was 3.9 compared to those of men, 4.06.
Here we're looking at the average teaching evaluation for tenured professors, 3.96 verses untenured, 4.13. The average age of untenured professors was 50.2 years, and that for tenured professors, 47.85 years.

One thing that is very important in statistical analysis is to think about the question and to think about the population or sample that you are working with. We are computing averages across 463 courses. We find the average age or beauty, but these are the attributes of instructors. We know from our data to their 94 instructors who have collectively taught 463 courses, and we know that there are duplicates, that is the same instructor who has taught multiple courses. So when I compute the average age using 463 courses, it's not necessarily the average age of the instructors because it could be true that older aged individuals may have taught more courses than younger individuals, resulting in an higher average. That is not necessarily the average age of the instructors. So to avoid this problem, we have to subset the data so that we remove the duplicates and have only one observation per individual instructor in the dataset. Instead of 463 observations, you should have just 94 observations.

### Age and Beauty corrected

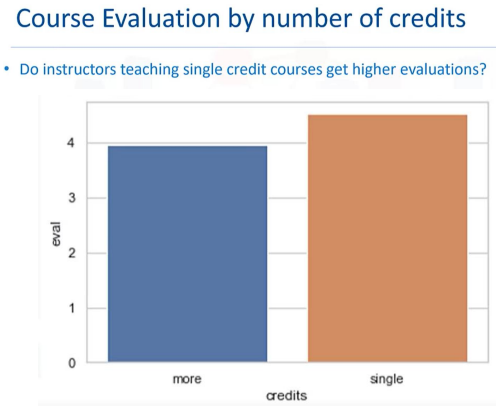
With 94 observations

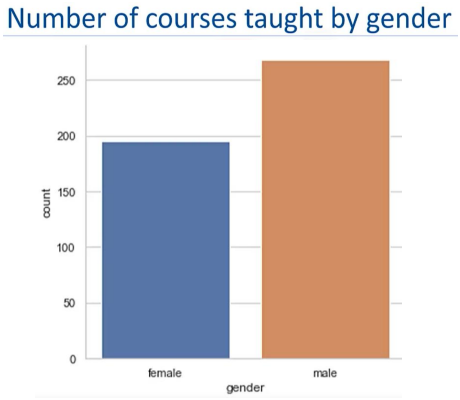
| | gender | beauty | | | age | | |
|---|---|---|---|---|---|---|---|
| | | mean | std | var | mean | std | var |
| 0 | female | 0.252303 | 0.843667 | 0.711774 | 44.950000 | 8.935524 | 79.843590 |
| 1 | male | -0.033098 | 0.801559 | 0.642497 | 49.481481 | 10.813585 | 116.933613 |

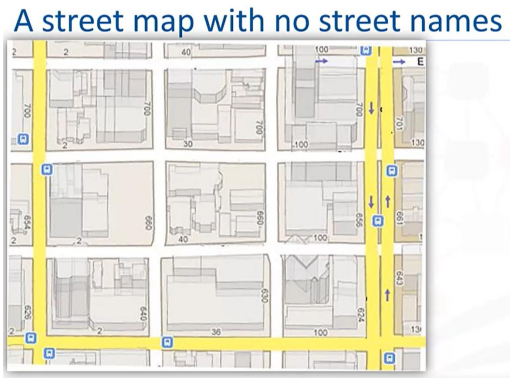
| | gender | eval | | | beauty | | | age | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | mean | std | var | mean | std | var | mean | std | var |
| 0 | female | 3.901026 | 0.538803 | 0.290308 | 0.116109 | 0.81781 | 0.668813 | 45.092308 | 8.532031 | 72.795559 |
| 1 | male | 4.069030 | 0.556652 | 0.309861 | -0.084482 | 0.75713 | 0.573246 | 50.746269 | 9.993396 | 99.867964 |

Now let's look at the comparison. When we use 94 observations where no instructor is repeated in the dataset, the average age or average beauty score is 0.25. When we look at the 463 courses, the average value is 0.11. Let's compare the age. The average age using 94 observations for males is 49.4 and for females is 44.9. You see here that as for age, we don't see much difference whether we use 463 observations or 94. But we certainly see much difference in the beauty scores if you were to use the wrong dataset. That is the dataset where individuals are repeated multiple times.

**总结:**

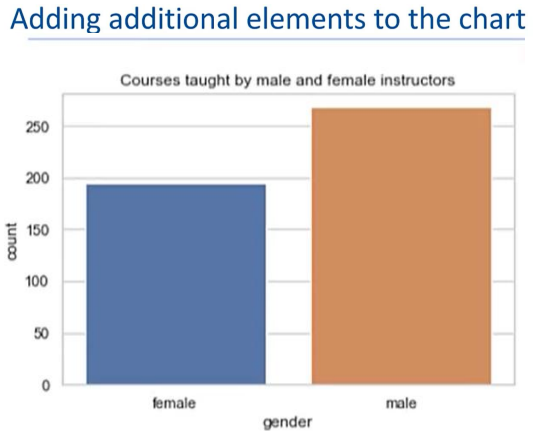


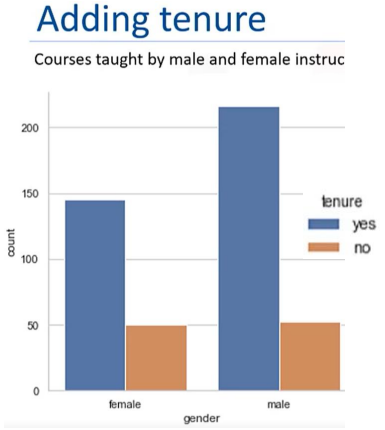



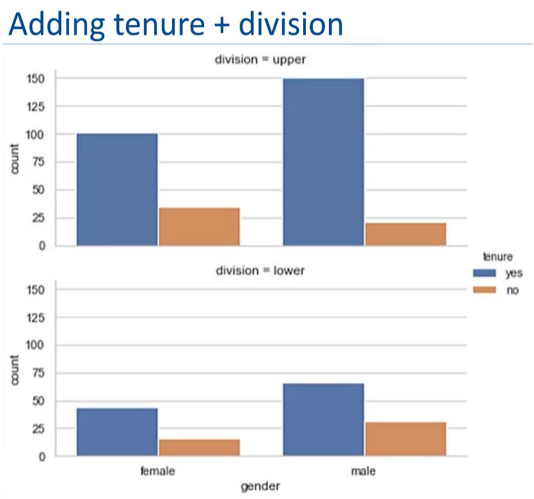




Data visualization is a critical piece of modern-day statistical analysis. Their staples are helpful, so you don't have to eyeball the output to figure out what the trends are. The visual displays are much easier to understand.
We will use the same datasets of teaching evaluations and ask this question, do instructors teaching single credit courses get higher evaluations? We see that, yes, they do. By Mean evaluation, when plotted as the chart, you see that instructors who teach single credit courses have a slightly higher average teaching evaluation.

Let us start by determining how many courses were taught by male instructors and how many by female instructors. For this, we can use a bar chart. Notice that the information is complete from a statistical point of view in that we know how many courses were taught by males versus females. But we do not have some critical information from this chart as it relates to communication.
Therefore, we can say this chart serves as **statistical purpose**, but it doesn't serve a **communication purpose**.
Let me illustrate this with an example. Here you are looking at a street map. You can see the streets and the buildings and the highways, but you don't see the street names. Without street names, it is hard to determine where you are and in which direction you should be heading. Even though it is according to scale, it may be accurate in its depiction of the streets in the neighborhood, but it's still lacks the ability to communicate information to you. To add communication value to this map, you can simply add the street names.




```
ax = sns.countplot(x='gender', data=ratings_df)
ax.set_title("Courses taught by male and female instructors")
```

```
ax = sns.countplot(x='gender', data=ratings_df, hue = 'tenure')
ax.set_title("Courses taught by male and female instructors")
```

```
ax = sns.catplot(x='gender', data=ratings_df, kind ='count', hue = 'tenure', row = 'division')
```

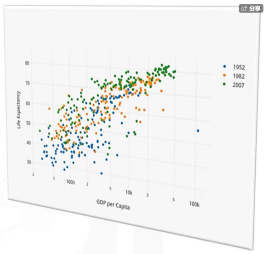
Let us apply the same philosophy to our graphic. But once we add information about this infographic, for example, adding just a title makes this chart more informative.

- To do this in Python, we'll use the ***countplot*** function in the seaborne library and set the **title label**. This helps your graph to be more informative.
- We can also add more dimensions to the data. In addition to the gender of the instructors, we could add the tenure status of the instructors as well to the graphic. To do that in Python, you add **hue argument** to the countplot.
- We can add another dimension to the data, regenerating the same graphic with the same information. That is, the number of courses taught by gender and tenure. Then adding the dimension of courses being upper-division and lower division, and presenting them in two rows or columns. To do this in Python, we can specify the **row argument** using the ***countplot*** function.

**总结:**

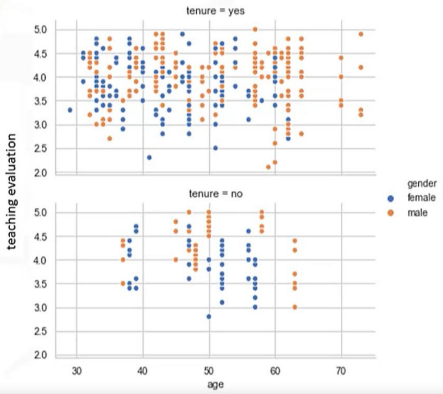
## Scattered plots

Continuous variables



## Does age affect teaching evaluations?

- Continuous Variables
  - Age
  - Evaluation Scores
- Categorical Variables
  - Gender
  - Tenure



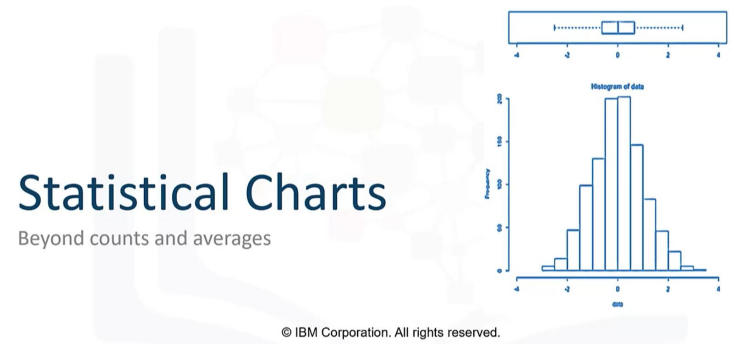

```
g = sns.FacetGrid(ratings_df, row="tenure",  hue="gender")
g = (g.map(plt.scatter, "age", "eval")
     .add_legend())
```

Now let's look at the situation where our primary variables of interest are continuous variables. We would like to explore the relationship between the two while adding further categorical variables as an additional dimension.

Using the teaching evaluation data we ask this question, does age effect teaching evaluations? We then add two additional dimensions, which are gender and tenure. So our dataset consists of age and teaching evaluation, which are the two primary variables of interest and are continuous. Then we add two other dimensions, i.e., gender and tenure. These are categorical variables. Age is on the X axis and the teaching evaluation scores on the Y axis. The orange colored circles represent males and the blue colored circles represent females. The top panel is for tenured professors and the bottom panel is for the untenured instructors.
To do this in Python, we use the FacetGrid option, which works for multiplot gridding and allows tweaking the plot. You create the row and hue for the categorical variables, in our case, tenure and gender. Then we use the map to apply a plotting function to each subset of the data.

**总结:**

# 3. Statistical Charts



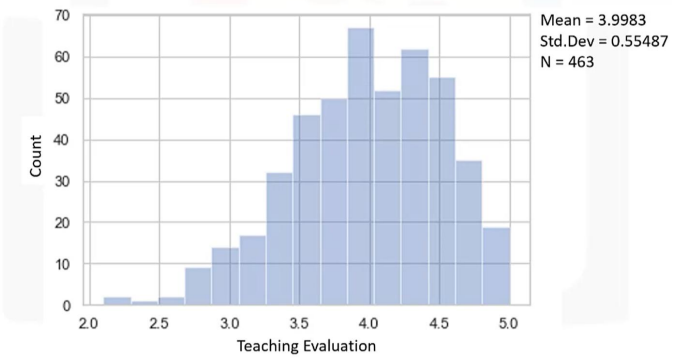

## Better than the average

Statistical parameters illustrated in charts:

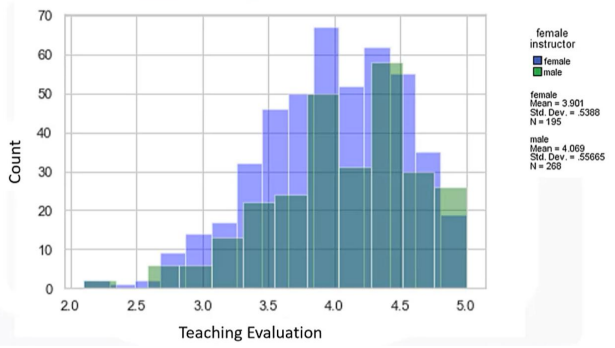
- Distribution and variance: Histogram
- Box plot: Displaying mean, median quartile and outliers

So far we have displayed data as averages and counts. Now let's look at some other statistical parameters that we will illustrate as graphics. We have not yet shown anything about variance and I think the first thing that one should look into beyond averages, is to look at variance or how the data are distributed.
A good way of looking at the distribution of data, especially if it were to be a continuous variable, is to look at histograms. If you are interested in displaying something more than the average, maybe the median and the quartiles, then perhaps boxplots should be our choice.

## Simple Histogram – teaching evaluation score



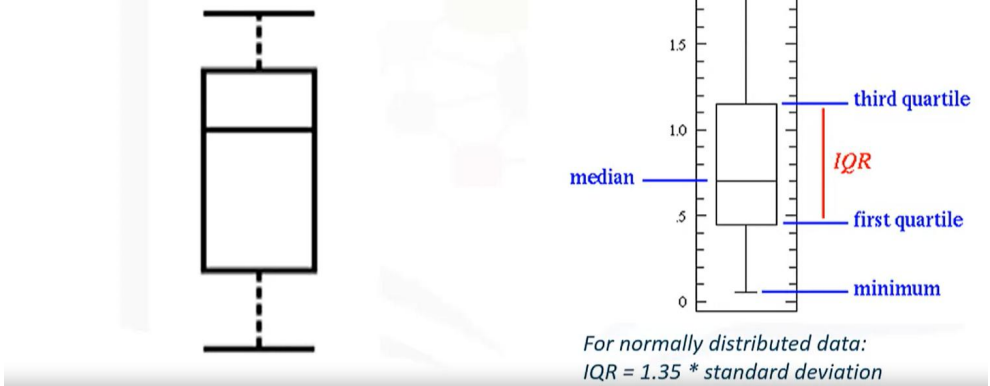

## Evaluation by gender-Histogram




Using the teaching evaluation data, we have plotted a histogram of teaching evaluation scores. You could see that the mean score is around four, but then you could see very low teaching evaluation scores, not many frequently, but most frequently is the around the average and then you'd see that some have lower teaching evaluation scores and some have fairly high teaching evaluation scores. The histogram approximates the normal distribution curve. Essentially you have 3.99 to four as the mean, the standard deviation of 0.55, looking at 463 records. This gives you a good idea of how your data are distributed.

You can in fact plot multiple histograms such that you can see the difference between the subgroups. Here you have the histograms overlaid for males and females. These frequent lower teaching evaluations for females is likely to influence the average teaching evaluation score for females versus the males.

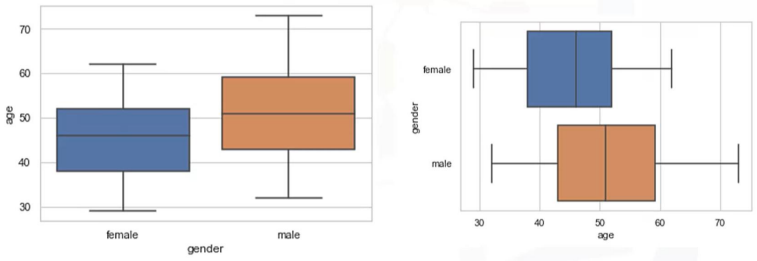
## Box plots




A box plot essentially looks like this. The thick line in the box represents the median. The top part of the box is the third quartile. The bottom part of the box as the first quartile. The line at the bottom is the minimum value, and the line at the top is the maximum value. The range between the first quartile and the third quartile is called the **inter-quartile range**.

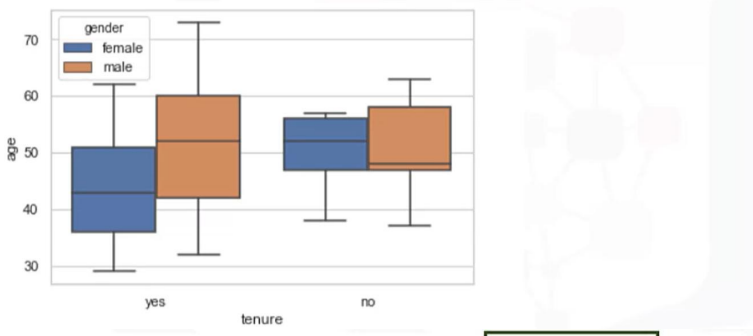
**总结:**

## Box plots – age of the instructor by gender




```
ax = sns.boxplot(x="gender", y="age", data=ratings_df)
```

## Comparing age along tenure and gender

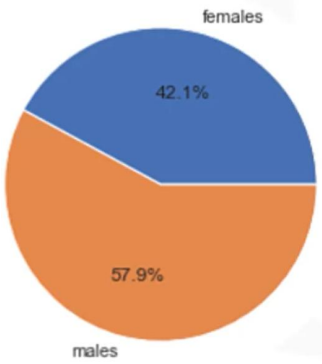



```
ax = sns.boxplot(x="tenure", y="age", hue="gender",
                 data=ratings_df)
```

In this graphic, we have created the box plots for the age variable. We can see that the median age of males is higher than the median age of females. Also, the maximum age of the males is higher than the maximum age of females. To do this in Python, we use the boxplot function in the Seaborne library. We will put the gender on the Y axis and the age of the instructor on the X axis. You can play around with the X and Y axis.
If you want a horizontal style boxplot for readability, I like to use vertical boxplots.

We can also add another dimension. Here we will add tenure. So those who are tenured are plotted on the right, and those who were not tenured are plotted on the left. The blue color represents the female instructors, and the orange color represents the male instructors. We can see the differences between male and female. Instructors, male tenured instructors are older than male untenured instructors, whereas female tenured instructors are younger than female untenured instructors. To do this in Python, add the hue argument at the box plot function.

## Pie in the sky



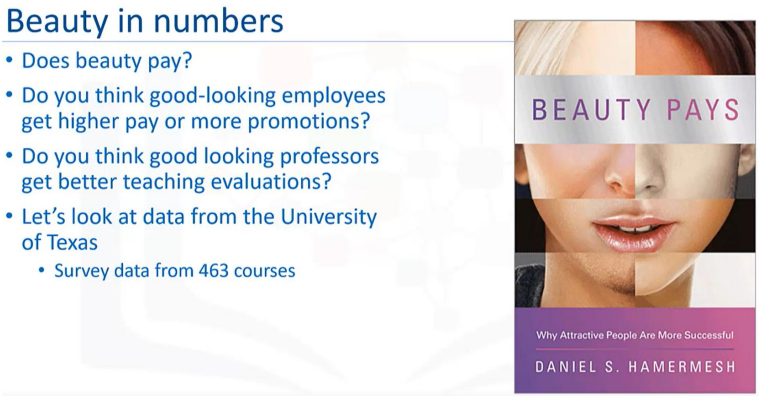

```
import matplotlib.pyplot as plt
```

```
labels = ['females', 'males']
sizes = [ratings_df['gender'].value_counts()[1],
         ratings_df['gender'].value_counts()[0]
        ]
fig1, ax1 = plt.subplots()
ax1.pie(sizes, labels=labels, autopct='%1.1f%%')
plt.show()
```

A pie chart is another way of looking at your data. You can see here in this graphic that the number of courses taught by male instructors is larger than the number of courses taught by female instructors. To do this in Python, we will use the matplotlib library. First we specify the labels, get the number of courses taught both by male and females, and assign it to a size is variable. Create a subplot, insert the sizes, labels and percentage to one decimal place in the pie function, and print out the pie chart with the show function.

**总结:**

# 4. Introducing the teacher's rating data




In this video I will introduce the Teachings Ratings Data. We have been working with teaching ratings data from University of Texas. And the underlying question is, if students teaching evaluations are influenced by the looks of individual instructors. Or you can ask if they're teaching evaluations differ by gender, or if good looking instructors get higher teaching evaluations.

I obtained this data from Professor Daniel Hamermesh, who has written a paper about how beauty may impact in an instructors teaching evaluation. In fact, he has written an amazing book called Beauty Pays, in which he answers these questions. Such as do you think good looking employees get higher pay or faster promotions? Do you think good looking instructors get higher teaching evaluations and the data comes from University of Texas? It's a survey. And data obtained from 463 courses.

## Meta data: The data about data

- Details are available in Chapter 4 of *Getting Started with Data Science*
- Variable type:
  - Real number/Integer
  - Continuous/ categorical variables
  - Dichotomous variable

| Variable | Description |
|---|---|
| beauty | Rating of the instructor's physical appearance by a panel of six students, averaged across the six panelists, transformed to have a mean of zero. |
| eval | Course overall teaching evaluation score, on a scale of 1 (very unsatisfactory) to 5 (excellent). |
| minority | Factor variable. Does the instructor belong to a minority (non-Caucasian)? |
| age | The professor's age. |
| gender | Factor indicating instructor's gender (male/female). |
| native | Factor variable. Is the instructor a native English speaker? |
| tenure | Factor variable. Is the instructor on tenure track? |
| credits | Factor variable. Is the course a single-credit elective (for example, yoga, aerobics, dance)? |
| division | Factor variable. Is the course an upper or lower division course? (Lower division courses are mainly large freshman and sophomore courses.) |
| students | Number of students who participated in the evaluation. |
| allstudents | Number of students enrolled in the course. |
| prof | Factor variable indicating instructor identifier. |

So the data is first referenced in the book, Getting Started with Data Science, making sense of data with Analytics in Chapter 4. They're variables that essentially define the attributes of instructors and characteristics of the courses. Some variables are continuous, others that dichotomous or categorical variables.
So the primary two variables of our interests are beauty score, which is basically the physical appearance of an instructor which was ranked by a panel of six students. And I think that they have normalized the beauty score such that it had the mean of zero and variance of one or a standard deviation of one the same as **Z transformation**.
The dependent variable, the variable of interest is evaluation, which is basically the teaching evaluation ranging between the scale of 1 to 5, one being very unsatisfactory and the student found the course to be excellent, then it's 5.

And then there are other dichotomous or binary or categorical variables such as minority. If the instructor was non Caucasian, age is a continuous variable as the professors age or instructors age, gender being male or female. Native stands for native English speaker. If the instructor with a native speaker of English language10, otherwise. If the professor was tenured, 10 otherwise.

**总结:**

# Descriptive Statistics

| | age | beauty | eval | students | allstudents |
|---|---|---|---|---|---|
| count | 463.000000 | 4.630000e+02 | 463.000000 | 463.000000 | 463.000000 |
| mean | 48.365011 | 6.271140e-08 | 3.998272 | 36.624190 | 55.177106 |
| std | 9.802742 | 7.886477e-01 | 0.554866 | 45.018481 | 75.072800 |
| min | 29.000000 | -1.450494e+00 | 2.100000 | 5.000000 | 8.000000 |
| 25% | 42.000000 | -6.562689e-01 | 3.600000 | 15.000000 | 19.000000 |
| 50% | 48.000000 | -6.801430e-02 | 4.000000 | 23.000000 | 29.000000 |
| 75% | 57.000000 | 5.456024e-01 | 4.400000 | 40.000000 | 60.000000 |
| max | 73.000000 | 1.970023e+00 | 5.000000 | 380.000000 | 581.000000 |

native

| | | count |
|---|---|---|
| 0 | no | 28 |
| 1 | yes | 435 |

minority

| | | count |
|---|---|---|
| 0 | no | 399 |
| 1 | yes | 64 |

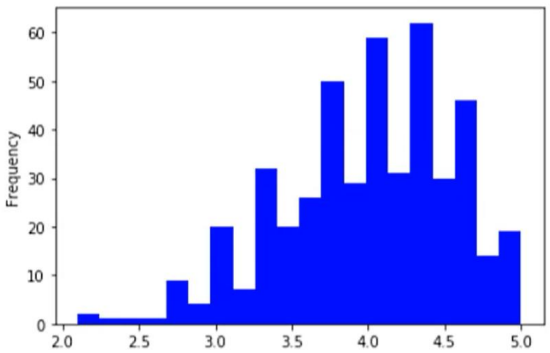
gender

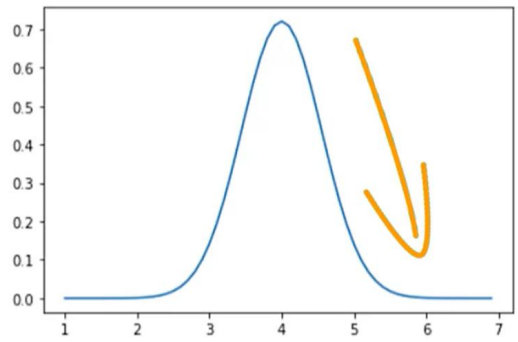
| | | count |
|---|---|---|
| 0 | female | 195 |
| 1 | male | 268 |

tenure

| | | count |
|---|---|---|
| 0 | no | 102 |
| 1 | yes | 361 |

I have produced some descriptors statistics for your reference. For instance:
For the continuous variables such as age and beauty and teaching evaluation and the number of students who are enrolled in the course on the number of students who performed the teaching evaluations. I produce the descriptive statistics such as the minimum, the maximum mean, and standard deviation.
For categorical variables such as gender, female yes or no, visible minority yes or no person being a tenured professor or otherwise. I produced the frequency distributions and percentage of individuals falling in one category or otherwise.
Notice that the teaching evaluation score is 3.99, with the standard deviation of 0.55.




Let's see if I were to produce a histogram of this variable teaching evaluation, how will it look like with raw data? And if I were to use the normal distribution and feed the two parameters, that is the mean and the standard deviation, how will the same distribution look like using a normal distribution? I am presenting here the distribution of the **raw data** on left side and the presentation of the same data with the same parameters of the mean and standard deviation using normal distribution.
You could see that the data not exactly following a Bell curve. This is the raw data and it seldom does. But then the theoretical distribution looks like this. Essentially the same data set with a mean of 3.998 in standard deviation is presented here and then a normal distribution drawn from these two characters appear step, so it's much smoother. Theoretical distributions are much smoother than the raw data.

**总结:**

1. What's the best way to display median and outliers?

- ○ A time series plot
- ○ A bubble chart
- ○ A scatter plot
- ◉ A box plot

✓ **正确**

Correct! Boxplots are a way of displaying the distribution of data based on a five number summary ("minimum", first quartile, median, third quartile, and "maximum"). It also displays the outliers of the dataset

2. What is a suitable way to display the average basketball scores between two teams?

- ◉ A bar chart
- ○ A pie chart
- ○ A histogram
- ○ A scatter plot

✓ **正确**

Correct! A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent

**总结:**

2. What is a suitable way to display relationship between two continuous variables?

- ○ A bar chart
- ◉ A scatter plot
- ○ A histogram
- ○ A pie chart

✓ **正确**

Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another.

3. When the sum of two or more categories equals 100, what chart type is ideally suited for displaying data?

- ○ A line chart
- ○ A box plot
- ○ A histogram
- ◉ A pie chart

✓ **正确**

Correct! A Pie Chart is a chart that uses pie slices to show relative sizes of data.

5. When multiple observations are reported for each respondent in the data set, to compute statistics for variables about the respondents, one must:

- ○ Ignore the presence of duplicates and compute statistics as usual
- ○ Weight data by duplicates
- ◉ Remove duplicates before running analysis
- ○ None of the above

✓ **正确**

Correct!

**总结:**