6.1 Visualizing data with one feature

Learning goals

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Interpret plots of a single categorical feature.
- Interpret plots of a single numerical feature.
- Construct bar charts, histograms, density plots, and box plots using `seaborn`.

Visualizing a categorical feature

The simplest case for visualization is counting items in different groups. A ***categorical feature*** divides the dataset into different groups or categories. A ***bar chart*** has the groups on one axis and then rectangles with heights that represent the number of individuals in that group.

This section uses the country dataset taken from the 2017 Gapminder data, which contains 151 countries and 7 features. The features are given in the table below.

Table 6.1.1: Features of the country dataset.

Feature	Type	Description
Years	Numerical	Years of schooling completed
Emissions	Numerical	CO2 emissions per person
Fertility	Numerical	Births per woman
Internet	Numerical	Percent of population with internet access
Continent	Categorical	Continent where the country is located

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Internet access	Categorical	Low, Moderate, High, Very high
Emissions range	Categorical	Low, Moderate, High, Very high

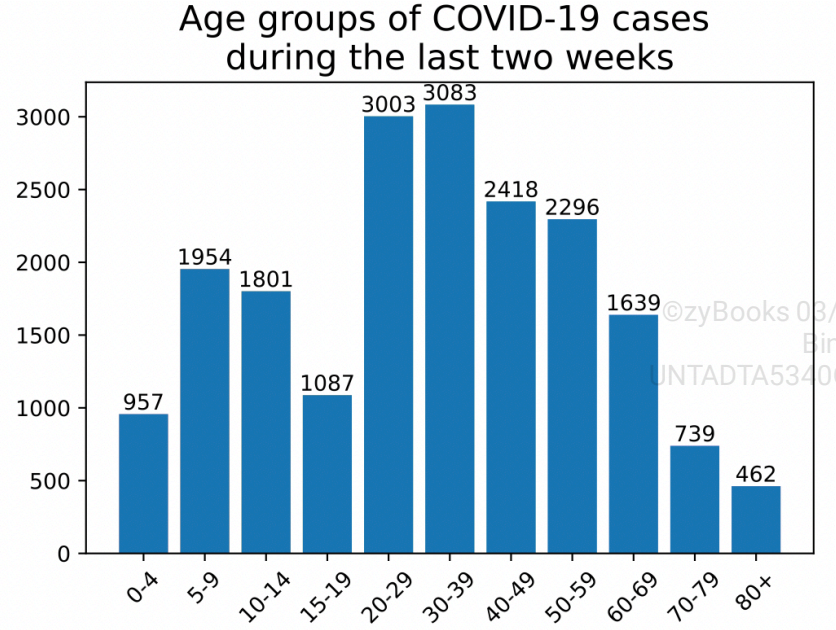
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing a categorical feature from the country dataset.

Use the drop-down menu to select a categorical feature.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.1.1: Interpreting bar charts.



The data used to create the bar chart above was obtained from the Massachusetts Department of Public Health's COVID-19 Dashboard on November 17, 2021. [1] The height of the bars corresponds to the number of positive cases of COVID-19 in the two weeks prior in the age range on the horizontal axis.

1) How many people between the ages of 50 and 59 were diagnosed with COVID-19 in the past two weeks?

- 2,296 cases
- 2,418 cases
- 2,300 cases

2) How many people under the age of 20 tested positive for COVID-19 during the past two weeks?

- 1,087
- 1,954
- 5,799

3) How many people between the ages of 20 and 39 tested positive for COVID-19 during the past two weeks?

- 3,083
- 6,086
- 8,504

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

4) What are the issues with this plot?

- Changing the number of years in each category leads to perceiving fewer cases in children than in adults.
- No labels on the axes
- No labels on the axes and changing the number of years included in each bar.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing proportions of categorical features

A ***relative frequency bar chart*** is a bar chart, but the height of each bar corresponds to the proportion of the dataset in each group. Relative frequency bar charts are useful when the proportion is more important than the count. Ex: Comparing between groups of different sizes.

To find the relative frequency, each count is divided by the total count. Since the total count does not change, dividing by the total count produces a scaling effect on the height of each bar. Thus, both frequency and relative frequency bar charts have the same shape.

Visualizing the relative frequency of a categorical feature from the country dataset.

Use the drop-down menu to select a categorical feature.

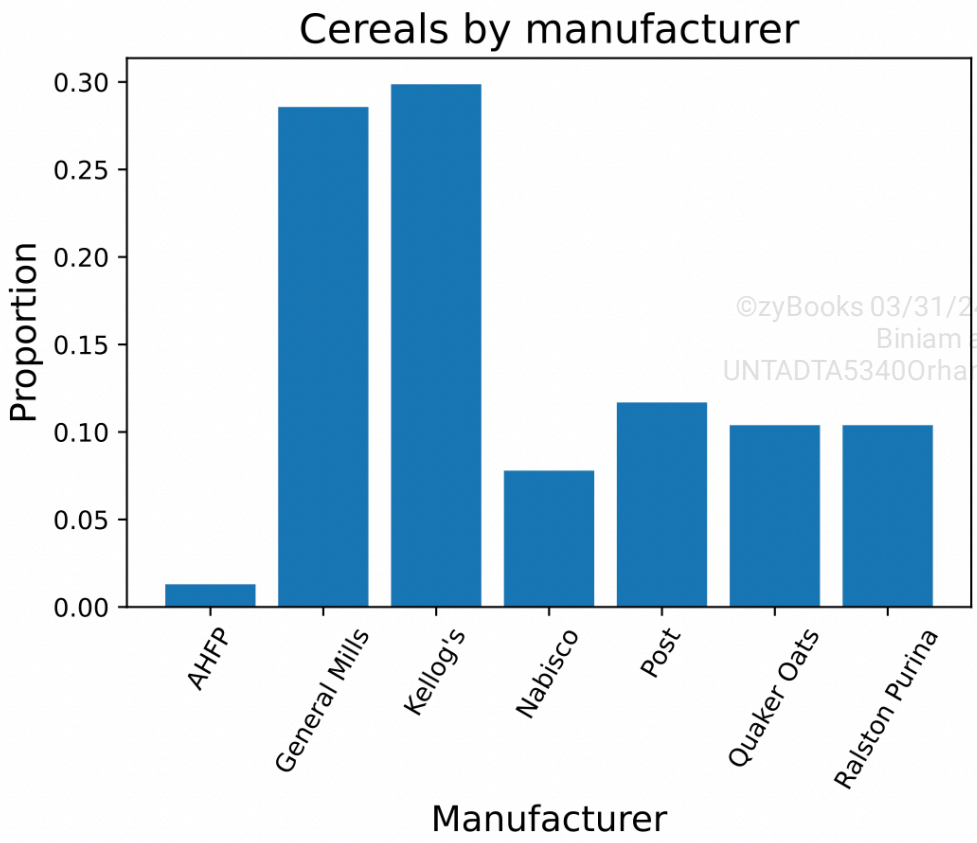
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.1.2: Relative frequency bar charts.

Information about 77 varieties of cereal on the shelves at a grocery store were collected. [2]
The distribution of manufacturers is plotted below.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

1) Estimate the proportion of cereals that were made by Post.

- 0.104
- 0.117
- 11.7

2) How many of the 77 cereals were manufactured by Post?

- 9
- 11
- 12

3) Which manufacturer has the second most cereals at this grocery store?

- General Mills
- Kellogg's
- Nabisco

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing a numerical feature

A ***numerical feature*** contains numbers that vary over many values. When visualizing a numerical feature, this variation is the aspect that must be communicated. Several approaches to visualizing a numerical feature are listed below.

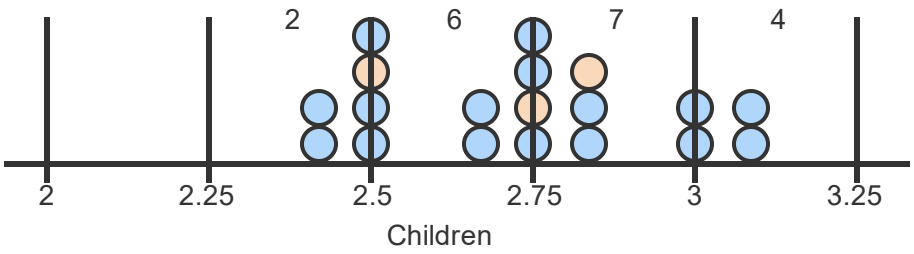
Table 6.1.2: Numerical feature visualizations.

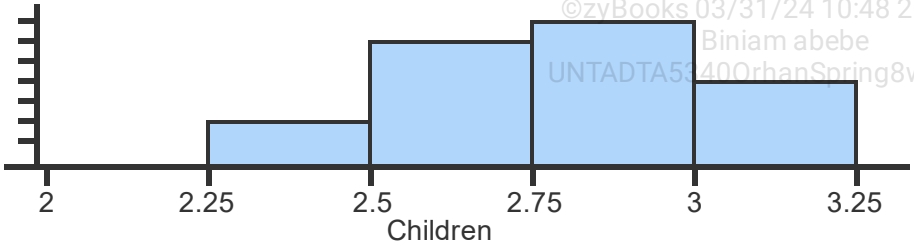
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Type	Description
Histogram	A bar chart that is created by dividing the numerical feature into small regions, or bins, and then counting the number of values in each region.
Density plot	A plot that approximates the density function of the distribution for the feature. Density plots can be thought of as a smoothed histogram. Usually this approximation is done by centering a small normal distribution over each data point and summing all these normal distributions to form the final density curve.
Box plot	A visual representation of the five-number summary; minimum, first quartile, median, third quartile, and maximum of a feature.

PARTICIPATION ACTIVITY 6.1.3: Making histograms and box plots.

Children
2.5
2.67
2.42
2.5
2.5
2.42
2.08
...
2.75
2.08
2.83
2.67





©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Step 1: A table of fertility rates (children per woman) is shown. Step 2: Points fly from the table to the corresponding value on a number line. If points have the same value they stack. Step 3: Dividers are added at quarter foot intervals, and the number of points in each interval between dividers is displayed. Step 4: A bar chart with no space between bars is shown on a duplicate number line. The height of each bar is the number of points in each interval. Step 5: Red slashes appear on the fifth, tenth and 15th points from the left. These occur at 5.5, 5.75, and 5.833. Step 6: A third number line appears the vertical lines fly from the red slashes to the location on the number line and form a box with a line in the middle. Horizontal lines extend from the box down to the minimum fertility rate and up to the maximum fertility rate.

Animation captions:

1. The average number of children born to each woman is shown for a sample of countries.
2. A dot plot is created by representing each country's average as a dot. Dots for countries with the same average are stacked.
3. To make a histogram, averages are placed into bins and counted. In this case, values on the boundary are rounded up.
4. Each bin is transformed into a bar that measures the number of countries in that bin.
5. The three quartiles, which cut the data into quarters, are needed to create a box plot.
6. The first quartile is the left edge of the box, the third quartile is the right edge. The median is the line in the middle. Whiskers extend to the minumum and maximum.

Visualizing a numerical feature from the country dataset.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.1.4: Interpreting numerical feature visualizations.

Use the above tool to answer the following questions.

1) What are the minimum and maximum years of schooling completed for countries in Europe?

- minimum = 9, maximum = 14
- minimum = 0.345, maximum = 0.085
- minimum = 8, maximum = 13

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) What is the median percentage of the population with internet access in the Americas?

- 53 percent
- 63 percent
- 74 percent

3) How many outliers does the data on CO2 emissions per person in Asia have?

- 0
- 3
- 4

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Single feature plots in Python with seaborn

Seaborn is a Python package for visualization that is built on top of and simplifies the syntax of matplotlib. Each type of plot is called with a named function. The dataframe and features to plot are passed as parameters. In the table below, `df` represents the dataframe.

Table 6.1.3: Seaborn single feature plots.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Command	Description
`sns.histplot(df, x='Feature')`	Creates a histogram of the named numerical feature from the dataframe.
`sns.kdeplot(df, x='Feature')`	Creates a density plot of the named numerical feature from the dataframe.
`sns.countplot(df, x='Feature')`	Creates a bar chart of the named categorical feature from the dataframe.
`sns.boxplot(df, x='Feature')`	Creates a box plot of the named numerical feature from the dataframe.
`sns.violinplot(df, x='Feature')`	Creates a violin plot of the named numerical feature from the dataframe.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing categorical features in Python.

Full screen

The code below uses the country dataset taken from the 2017 Gapminder data.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code by changing `x` to `y` to create a horizontal bar chart.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing numerical features in Python.

Full screen

The code below uses the country dataset taken from the 2017 Gapminder data.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify each plot to visualize `Internet` or `Emissions`.

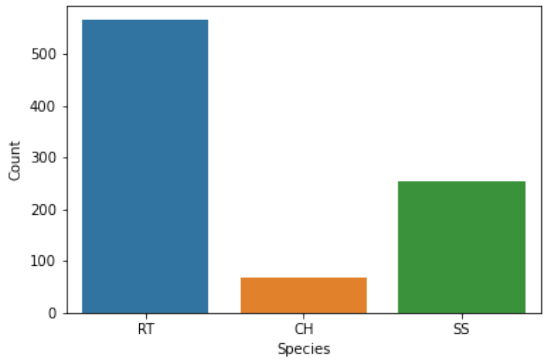
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.1.5: Code for single feature plots in seaborn.

Select the code that generates each plot from a dataset named `hawks` that contains body measurements including age, sex, wing, weight, bill length (culmen) and talon length (hallux) for a sample of hawks observed near Iowa City, Iowa.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

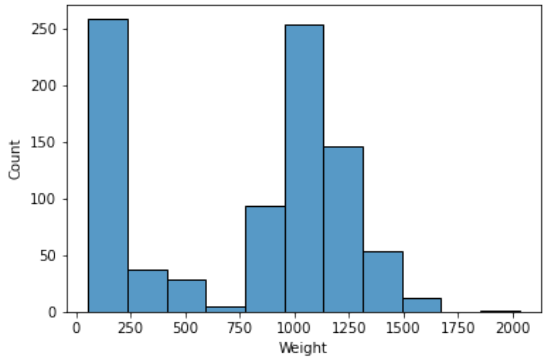
1)



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- `sns.countplot(data=hawks, x='Species')`
- `sns.kdeplot(data=hawks, x='Species')`
- `sns.violinplot(data=hawks, x='Species')`

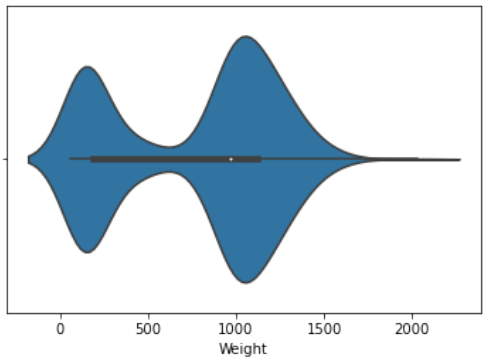
2)



- `sns.countplot(data=hawks, x='Weight')`
- `sns.histplot(data=hawks, x='Weight')`
- `sns.kdeplot(data=hawks, x='Weight')`
- `sns.violinplot(data=hawks, x='Weight')`

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

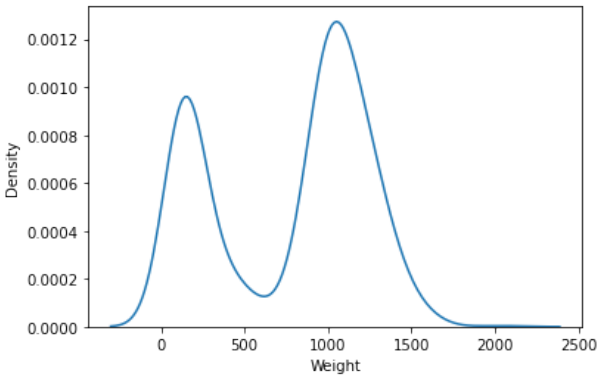
3)



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- `sns.countplot(data=hawks, x='Weight')`
- `sns.histplot(data=hawks, x='Weight')`
- `sns.kdeplot(data=hawks, x='Weight')`
- `sns.violinplot(data=hawks, x='Weight')`

4)



- `sns.countplot(data=hawks, x='Weight')`
- `sns.histplot(data=hawks, x='Weight')`
- `sns.kdeplot(data=hawks, x='Weight')`
- `sns.violinplot(data=hawks, x='Weight')`

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

CHALLENGE ACTIVITY 6.1.1: Visualizing data with a single feature.

537150.4174434.qx3zqy7

(*1) "Covid-19 Response Reporting." Massachusetts Department of Public Health. Accessed November 17, 2021. https://www.mass.gov/info-details/covid-19-response-reporting#covid-19-interactive-data-dashboard-.

(*2) "1993 Statistical Graphics Expo README." Data Expo 1993-Graphics Exposition. Joint Statistical Computing and Statistical Graphics Section, 1993. https://community.amstat.org/jointscsg-section/dataexpo/dataexpo1993 .

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

6.2 Visualizing data with multiple features

Learning goals

- Interpret plots containing two numerical features.
- Interpret plots containing one numerical feature and one categorical feature.
- Interpret plots containing two categorical features.
- Identify appropriate options for plots with more than two features.
- Construct multiple feature plots using `seaborn`.

Visualizing two numerical features

Visualizing a single feature uses one axis to display the feature value and another axis to display the value's frequency. But visualizing more than one feature eliminates the axis to show frequency. Visualizing more than one feature should highlight the relationship between those features to be used in a model or when communicating to others. With two numerical features, the standard plot is a scatter plot. A ***scatter plot*** displays an instance in a dataset as a point in a two-dimensional plane. The point's coordinates are the two features' values.

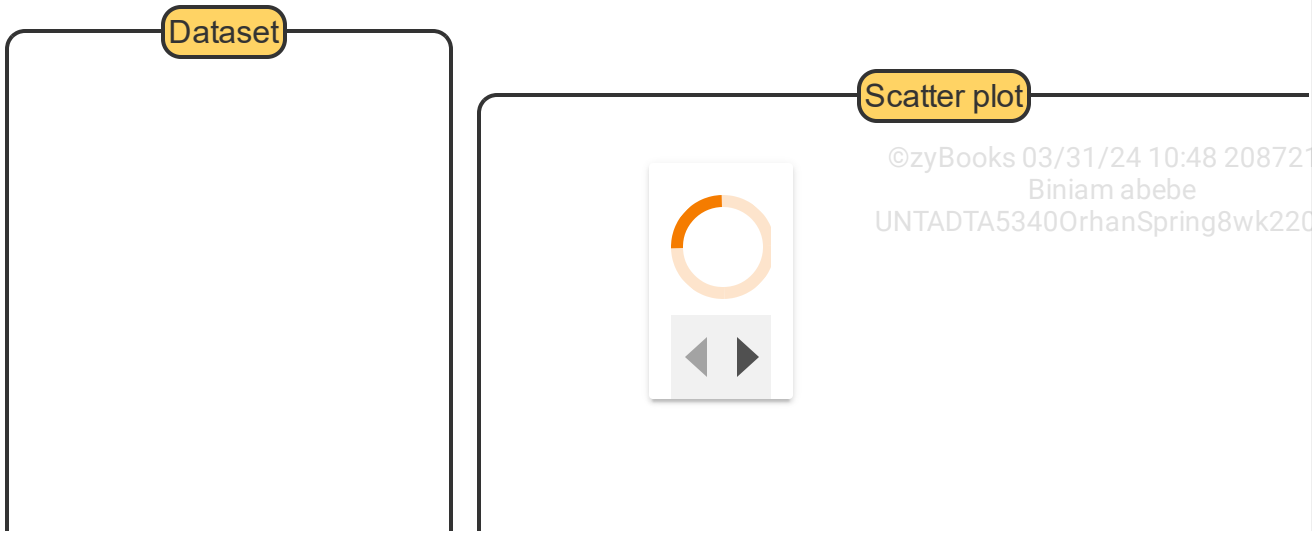
This section uses the tips dataset, which is a simulated dataset that has 244 rows and 7 columns. The features are given in the table below.

Table 6.2.1: Features of the tips dataset.

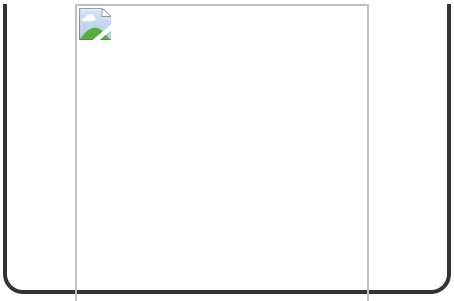
Feature	Type	Description
Total bill	Numerical	The total bill for the party in dollars
Tip	Numerical	The tip paid by the party in dollars
Sex	Categorical	The sex of the bill payer as identified by the waiter
Smoker	Categorical	Whether the party contained smokers
Day	Categorical	The day of the week
Time	Categorical	Whether the meal was lunch or dinner
Party size	Categorical	The size of the party

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.2.1: Making a scatter plot.



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

Step 1: A table showing the total bill and tips appears with values. Step 2: Axes with total bill on the horizontal axis and tips on the vertical axis appear. Tips go from around 1 to 10 and total bills go from around 1 to 50. Points appear at each pair of values and fly to the corresponding point (Total bill, tips) on the axes. Step 3: More points appear and the trend is that tips increase as the total bill increases.

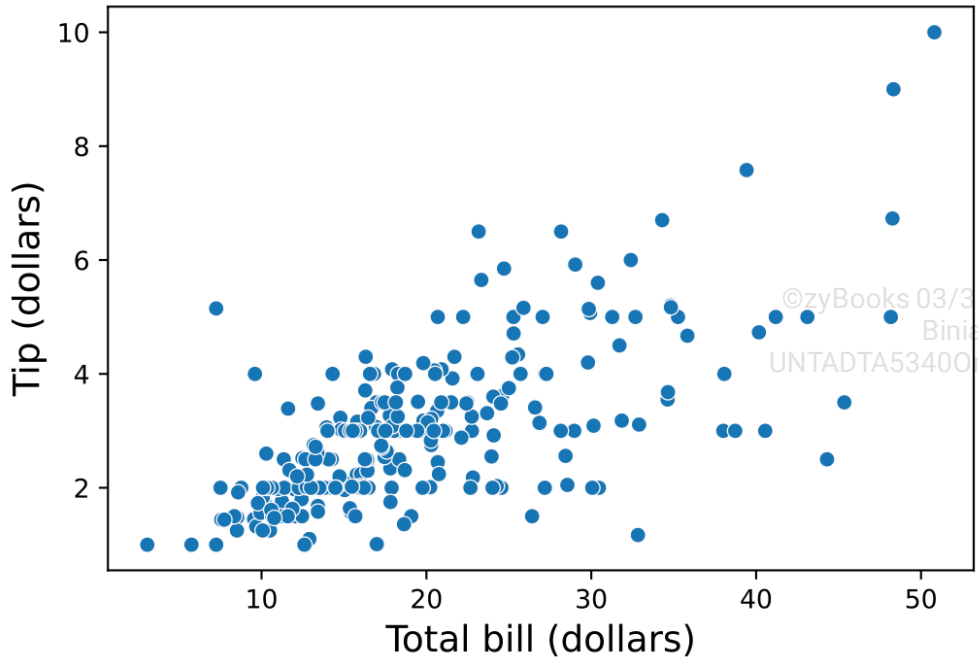
Animation captions:

1. Data on 244 diners at a restaurant were collected. The dataset has two features: total bill and tip.
2. Each diner is represented in the scatter plot as a point with total bill as the horizontal coordinate and tip as the vertical coordinate.
3. The scatter plot shows that as the total bill increases, so does the tip. In many places, tips are customarily given as a percentage of the total bill, so the trend makes sense.

PARTICIPATION ACTIVITY 6.2.2: Interpreting scatter plots.

One restaurant waiter recorded information about each tip received over a period of a few months. The waiter collected the bill and tip in dollars, sex of the bill payer, whether there were smokers in the party, day of the week, time of day, and size of the party.[1] The plot below shows the relationship between the total bill and the tip the waiter received.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

1) What is the largest tip's value?

- $10
- $50.81
- 71%

2) Which is the largest tip percentage \(\left (\dfrac{\text{tip}}{\text{bill}} \times 100\right) \)?

- 14.4%
- 19.7%
- 71%

3) What approximate values of total bill and tip occur most often?

- Total bill ≈ $12.50, tip ≈ $2
- Total bill ≈ $18, tip ≈ $3
- Total bill ≈ $22, tip ≈ $3.50

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing two categorical features

A plot of a single categorical feature uses one axis to separate categories and one axis to display relative frequency. A plot of two categorical features gives more information by grouping one categorical feature according to a second categorical feature. Ex: Grouping male and female customers according to the day the customer visited.

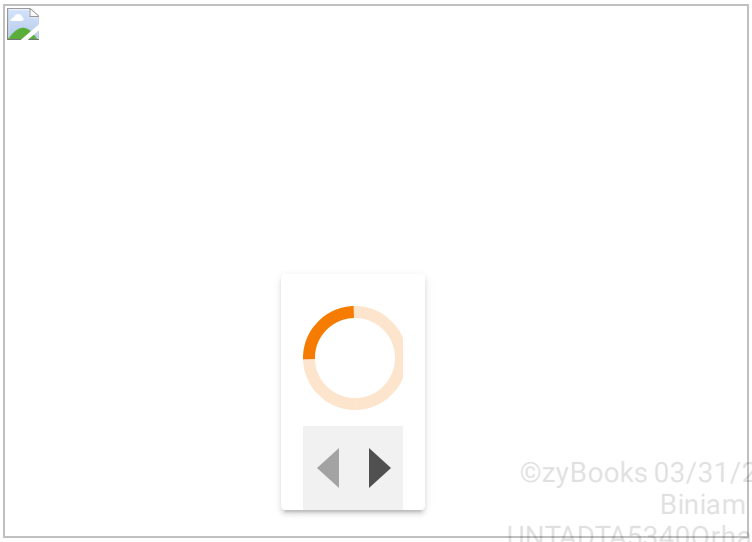
Table 6.2.2: Visualizations for two categorical features.

Name of plot	Description
Stacked bar chart	A bar chart is created with one categorical feature. The second feature divides, with color or shading, each bar. The height of each piece is based on the number of instances with values matching the bar and shading.
Grouped bar chart	A bar is created for each possible pair of values. The height of the bar is based on the number of instances with that pair of values. Bars are grouped by one feature and colored by the other feature.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.2.3: Grouped bar charts.

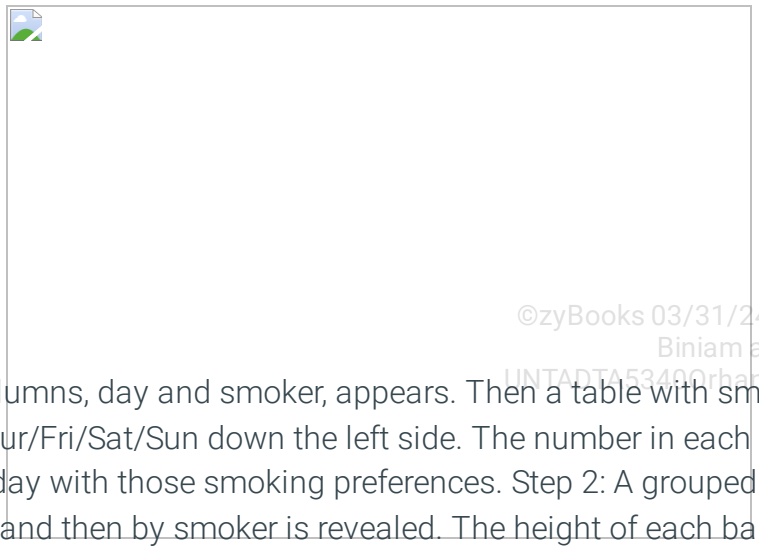
Cross tabulation



Did the restaur
the most smok
on Thursday, F
or Sunday?

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

The restaurant
most smokers
Saturday.



Animation content:

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Step 1: A table with two columns, day and smoker, appears. Then a table with smoker: yes/no across the top and day: Thur/Fri/Sat/Sun down the left side. The number in each cell is the number of parties on that day with those smoking preferences. Step 2: A grouped bar chart that separates first by day and then by smoker is revealed. The height of each bar corresponds to the value in the table in step 2. Step 3: The question "Did the restaurant have the most smokers on Thursday, Friday, or Sunday?" appears. Step 4: A stacked bar chart separated by day and colored by smoker appears. The total height of each bar is the total for each day. The pieces of the bars have heights corresponding to the value in the table in step 2.

Animation captions:

1. The cross tabulation relates the categorical features day and smoker. Bar charts visualize the relationship between categorical features.
2. A stacked bar chart highlights the total for a category.
3. Segments within a stacked bar chart can be difficult to compare, because the groups often do not start at the same position.
4. A grouped bar chart is helpful for comparisons within each day. A grouped bar chart easily shows that more smokers visited on Saturday.

Visualizing two categorical features from the tips dataset.

Use the drop-down menus to select bar chart type, categorical feature, and grouping feature.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.2.4: Interpreting bar charts.

Use the above tool to answer the following questions.

1) Which time had the most smokers?

- Lunch
- Dinner

2) Which day had the most visitors belonging to a party of 2?

- Friday
- Saturday
- Sunday

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

3) Which day was the restaurant serving lunch?

- Friday
- Saturday
- Sunday

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing categorical and numerical features

The commonly used plots for a categorical feature and a numerical feature borrow heavily from scatter plots and single feature plots.

Table 6.2.3: Categorical and numerical feature visualizations.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Name of plot	Description
Strip plot	A strip plot shows a scatter plot with points jittered off of the lines for the categorical feature. Jittering adjusts the horizontal or vertical position of points by a small amount so the points do not plot directly on top of each other.
Swarm plot	A swarm plot is a scatter plot with points jittered off the lines for the categorical feature so the points do not overlap. A swarm plot is useful for small datasets, but with an increasing number of points, the plots get too wide.
Density plot	In a density plot, the numerical feature's distribution is plotted for each of the categorical feature's categories. All the density plots are plotted on the same axes with color used to differentiate categories. A density plot is useful for large datasets but does not display differences in the number of instances in each category.
Violin plot	In a violin plot, the density plot for the numerical feature is plotted for each of the categorical feature's categories. Each density plot is mirrored and plotted offset from the others. A violin plot is useful for large datasets but does not display differences in the number of instances in each category.
Box plots	Box plots show a single box plot of the numerical feature for each value of the categorical feature. Box plots are useful for condensing large datasets and looking for outliers.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing a numerical and a categorical feature from the tips dataset.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Use the drop-down menus to select plot type, numerical feature, and categorical feature.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.2.5: Interpreting categorical and numerical visualizations.

Use the above tool to answer the following questions.

1) Which day had the largest total bill?

- Friday
- Saturday
- Sunday

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) The median tip for smokers is larger than the median tip for non-smokers.

- True
- False

3) What pattern is apparent in the swarm plot but not the other plots?

- Friday has the fewest data points.
- The distribution of tips is positively skewed for every day.
- Tips tend to fall at whole dollar amounts.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing more than two features

With more than two features, the screen cannot display enough directions for the information. For numerical features, color, size, transparency, and animation can be used. For categorical features, color, shading, size, transparency, shape, and faceting can be used. ***Faceting*** is the practice of displaying multiple plots side by side in an array where one feature changes from plot to plot.

Table 6.2.4: Attributes for displaying more features.

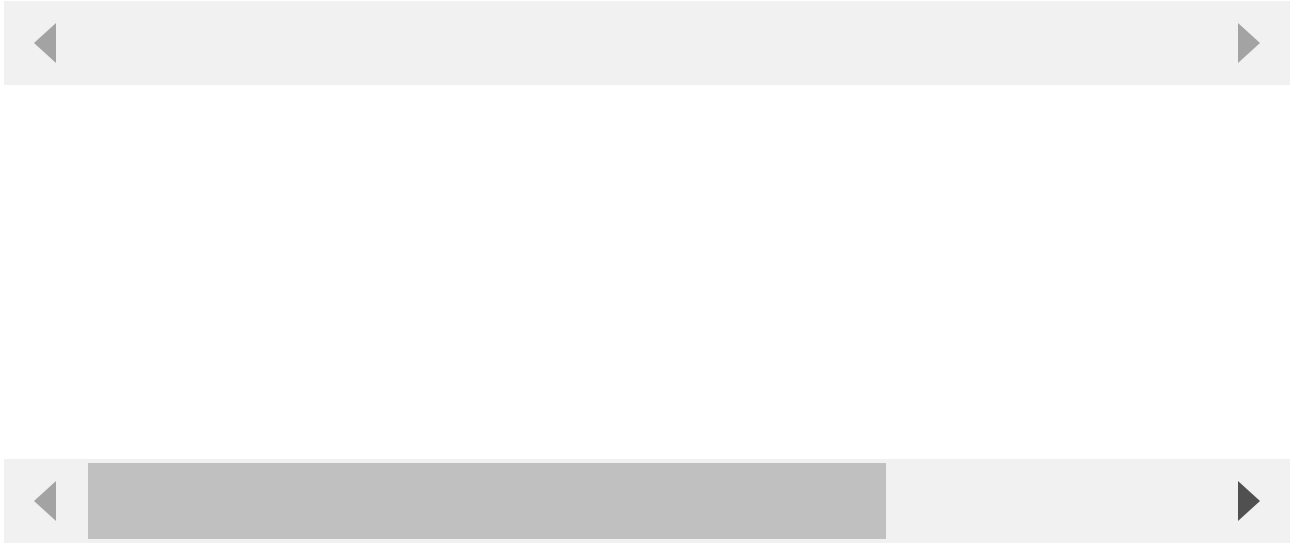
Attribute	Use
Color	Change the color of a point based on the value of categorical or numerical features.
Faceting	Display multiple plots side by side in an array where a categorical feature changes from plot to plot.
Shape	Change the shape of a point based on the value of a categorical feature.
Size	Change the size of a point based on the value of a categorical or numerical feature.
Transparency	Change the transparency of a point based on the value of a categorical or numerical feature.
Shading	Change the shading of a point based on the value of a categorical feature. Ex: Hatching, dotted, dashed.
Animation	Plot the data for each value of a numerical feature. Plots are often animated to display changes over time.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing a numerical feature and two categorical features from the tips dataset.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Use the drop-down menus to select plot type, numerical feature, and two categorical features.



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing two numerical and two categorical features from the tips dataset.

Use the drop-down menus to select features for the point colors and styles.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.2.6: Visualizing multiple features.

1) Which day had the largest median tip among diners belonging to a party size of 3?

- Thursday
- Saturday
- Sunday

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) Which plot shows the relationship between the total bill and tip, highlighting daily variations?

- A scatter plot with the total bill as the x-coordinate, tip as the y-coordinate, faceted by day
- Violin plots of tip for each day, faceted by total bill
- A swarm plot with total bill as the x-coordinate, colored by tip, and faceted by day

3) Which plot shows the relationship between the smoking and nonsmoking parties for each day and meal time?

- A scatter plot of day as the x-coordinate and meal as the y-coordinate, colored by whether the party smokes or not
- A stacked density plot of meal with the color-based day of the week and faceted by smoker or not
- A stacked bar chart with the color based on smoking/nonsmoking with meal on the horizontal axis and faceted by day of the week

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Multiple feature plots in Python using seaborn

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Seaborn is a Python package for visualization that is built on top of and simplifies the syntax of matplotlib. Each type of plot is called with a named function. The dataframe and features to plot are passed as parameters. In the tables below, `df` represents the dataframe. All of the single feature plots can be adjusted to accommodate a second feature using another function attribute. Ex: `y, hue, style, size`. Further information about attributes for each function is found in the seaborn references.

Table 6.2.5: Two feature plots in seaborn.

Command
`sns.scatterplot(df, x='Horizontal feature', y='Vertical feature`
`sns.swarmplot(df, x='Numerical feature', y='Categorical feature`
`sns.stripplot(df, x='Numerical feature', y='Categorical feature`

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing two numerical features in Python.

Full screen

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to produce a scatter plot of `Total bill` and `Party size`.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing two categorical features in Python.

Full screen

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to place `Time` on the x axis and set `Day` as the hue feature.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing numerical and categorical features in Python.

Full screen

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to use `time` as the categorical feature and `tip` as the numerical feature.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Visualizing multiple features in Python.

Full screen

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to use `Time` as the categorical feature.
- `seaborn` has several color palettes built in. Try some in the code below. Ex: "pastel", "dark", "viridis", "magma", "icefire", "Spectral".

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

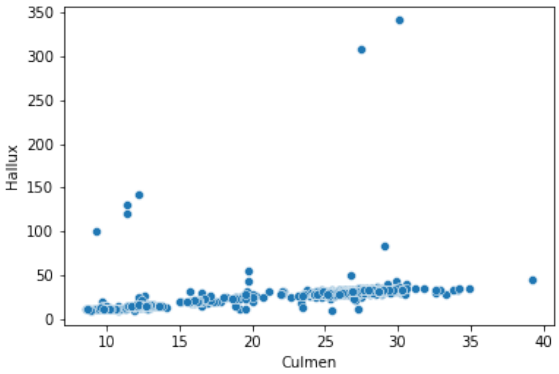
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.2.7: Code for multivariable plots in seaborn.

Select the code that generates each plot from a dataset named `hawks` that contains body measurements including age, sex, wing, weight, bill length (culmen) and talon length (hallux) for a sample of hawks observed near Iowa City, Iowa.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

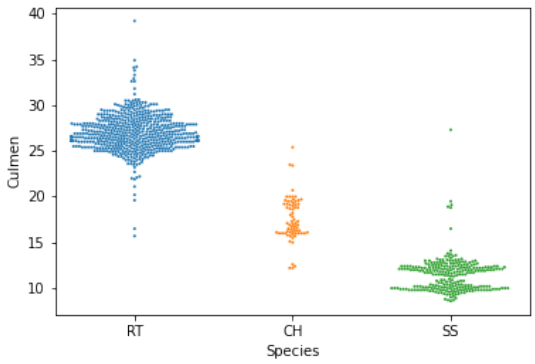
1)



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- `sns.scatterplot(data=hawks, x='Culmen', y='Hallux')`
- `sns.scatterplot(data=hawks, x='Hallux', y='Culmen')`
- `sns.stripplot(data=hawks, x='Culmen', y='Hallux')`

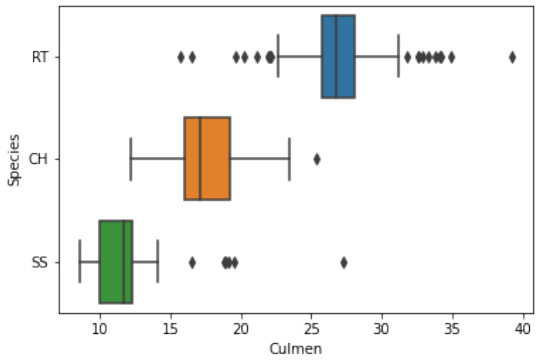
2)



- `sns.stripplot(data=hawks, x='Species', y='Culmen')`
- `sns.violinplot(data=hawks, x='Species', y='Culmen')`
- `sns.swarmplot(data=hawks, x='Species', y='Culmen')`

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

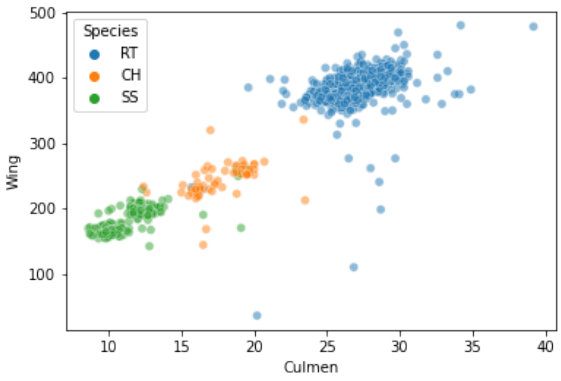
3)



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- `sns.boxplot(data=hawks, y='Species', x='Culmen')`
- `sns.violinplot(data=hawks, y='Species', x='Culmen')`
- `sns.boxplot(data=hawks, x='Species', y='Culmen')`

4)



- `sns.scatterplot(data=hawks, y='Culmen', x='Wing', hue='Species', alpha=0.5)`
- `sns.scatterplot(data=hawks, x='Culmen', y='Wing', hue='Species', alpha=0.5)`
- `sns.scatterplot(data=hawks, x='Culmen', y='Wing')`

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

CHALLENGE ACTIVITY 6.2.1: Visualizing data with multiple features.

537150.4174434.qx3zqy7

(*1) Bryant, P. G. and Smith, M Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing, 1995.

6.3 Best practices for visualizing data

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Learning goals

- Select and interpret an appropriate plot for given feature(s).
- Identify plots with misleading scales and explain how to improve plot scales.
- Compare plot colors and explain how to choose a suitable color palette.
- Explain why pie charts should be avoided.

Choosing base visualizations

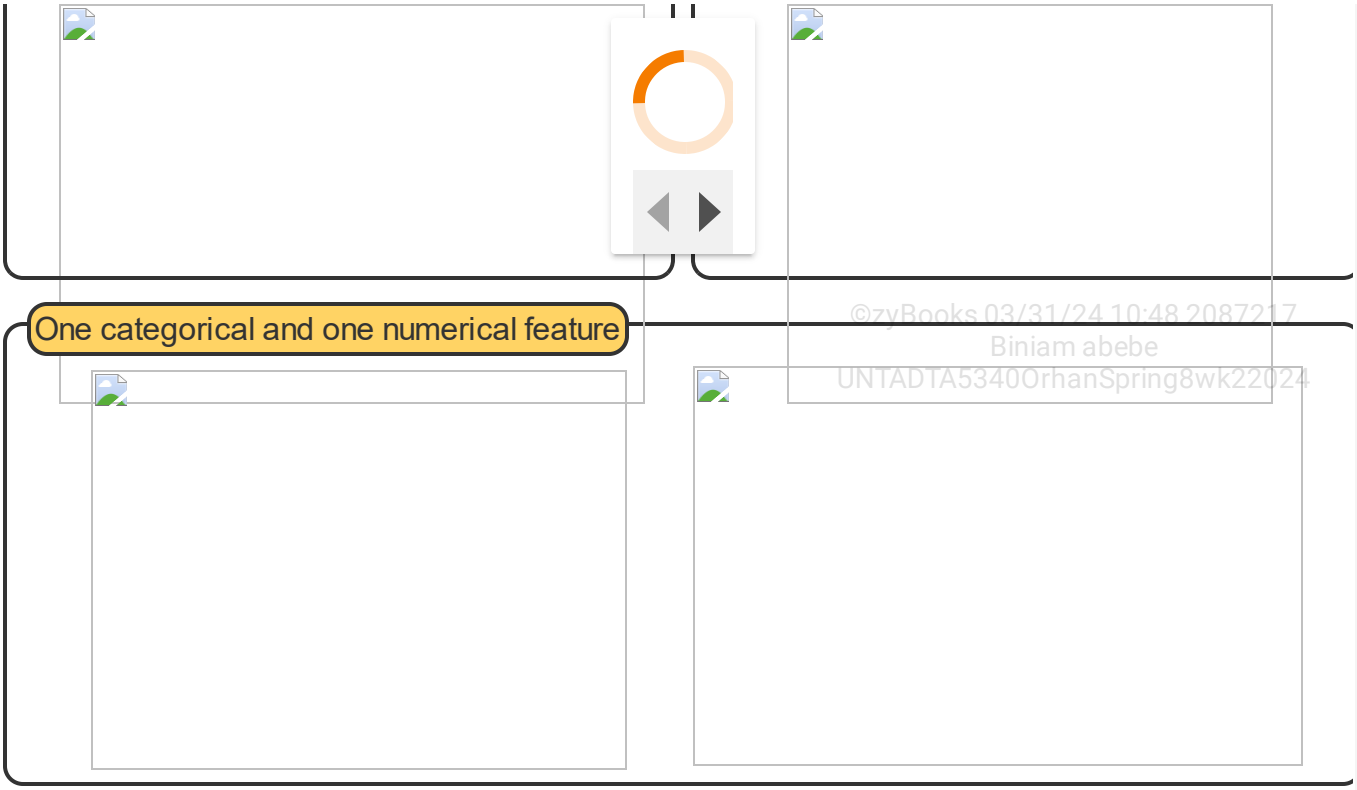
Given a set of features to visualize, the features' levels of measurement determine most choices for which plot to use. But, a data scientist should consider the audience and the story the plots tell. Simpler plots may be better than a complex visualizations that tell the same story but must be explained to the viewer. The norms of visualization choices vary based on audience.

PARTICIPATION ACTIVITY 6.3.1: Guidelines for selecting plots.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Two categorical features

Two numerical features



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

The plots described in the captions appear.

Animation captions:

1. If both features are categorical, a bar chart is appropriate.
2. If both features are numerical, a scatter plot is appropriate.
3. If one feature is categorical and the other feature is numerical, different plots are chosen depending on what effect is studied.

PARTICIPATION ACTIVITY 6.3.2: Determining the type of plot based on feature type.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Match the description of the features to the best plot listed.

If unable to drag and drop, refresh the page.

Violin plot **Scatter plot** **Stacked bar chart**

	Total bill at a restaurant vs. the party's size
	Amount of gas used to heat a house vs. the day of the week
	A student's major vs. the grade the student received in a particular course

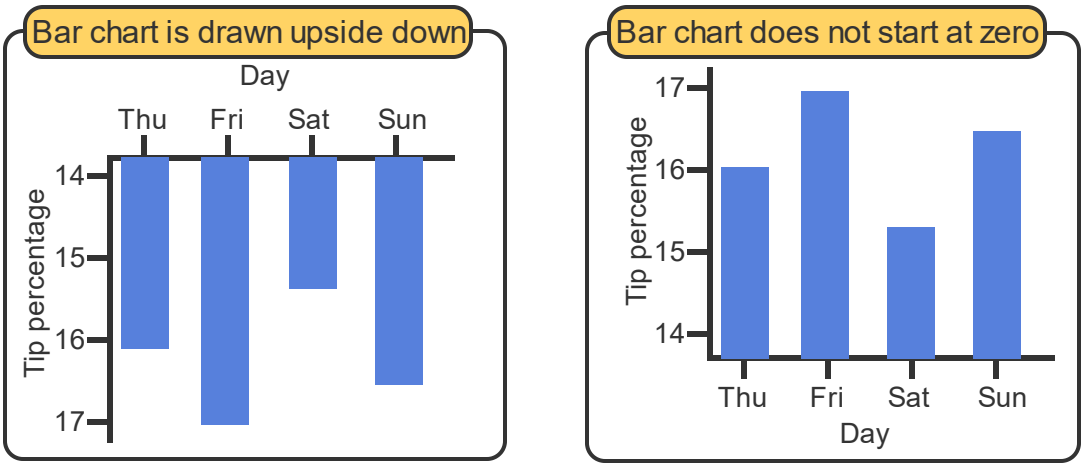
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

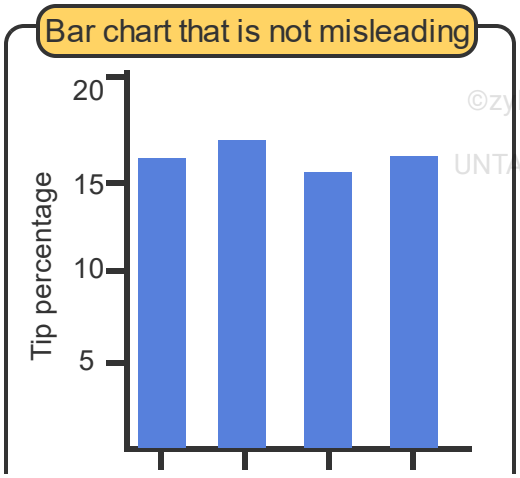
Reset

Words of warning: Scales

Scales should be clear to the viewer and not altered to accentuate differences in a misleading manner. Manipulation of scales is most common when using bar charts. Bar charts should include the value 0 for the bars' length. Scales should also follow the conventions that up and to the right are the positive directions on the axes.

PARTICIPATION ACTIVITY 6.3.3: Bad bar charts.





©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Thu Fri Sat Sun
Day

Animation content:

Three bar charts are shown of day vs. tip percentage. The issues with plots are described in the captions.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

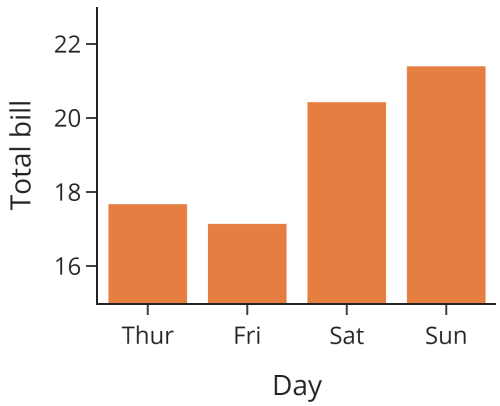
Animation captions:

1. In this bar chart, the vertical scale is upside down making Saturday look like the best day. Flipping the scale fixes that issue.
2. In this bar chart, the vertical scale does not start at zero, making Friday look twice as good as Saturday.
3. A bar chart should be upright and not use misleading scales.

PARTICIPATION ACTIVITY 6.3.4: Identifying issues with scales.

The following plots display the same data about the mean total bill in dollars for each day. For each of the following plots, identify the type of error that was addressed earliest in the animation above.

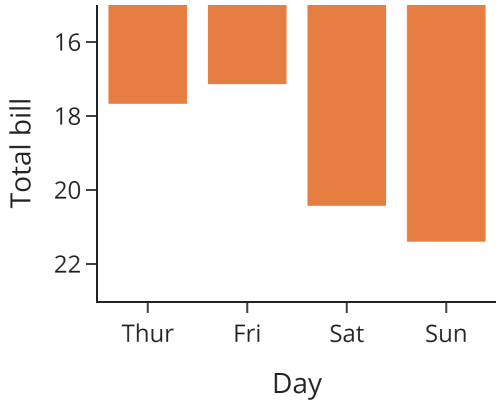
1)



- The scale does not include 0 and should be lengthened.
- The vertical scale is upside down and should be reversed.
- The horizontal labels are not in the appropriate order.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

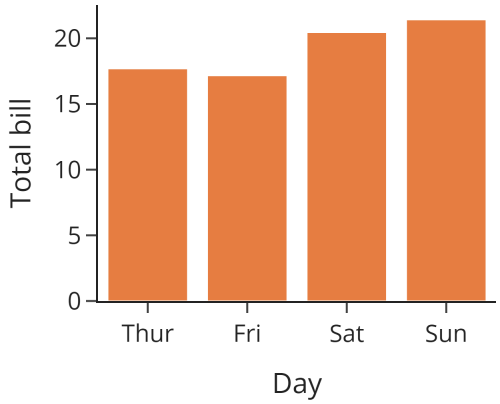
2)



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- The scale does not include 0 and should be lengthened.
- The vertical scale is upside down and should be reversed.

3)



- The horizontal labels are not in the appropriate order.
- The vertical scale is upside down and should be reversed.
- No error

Words of warning: Color

Color-vision deficiency (CVD) occurs in 8% of males, which reduces the person's ability to distinguish between red and green. Given the commonality of this deficiency, visualizations should be designed to accommodate for CVD by avoiding color that incorporates variations of red and green as the only distinguishing characteristic.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Contrast is the difference between the brightness of two colors. The highest contrast exists between black and white. Higher contrast makes differences easier to perceive and will enhance users' ability to distinguish between the two colors under suboptimal conditions. Ex: the screen is illuminated by bright sunshine, the screen is dim, the user has a visual impairment.

Gapminder plot using different color recommendations.

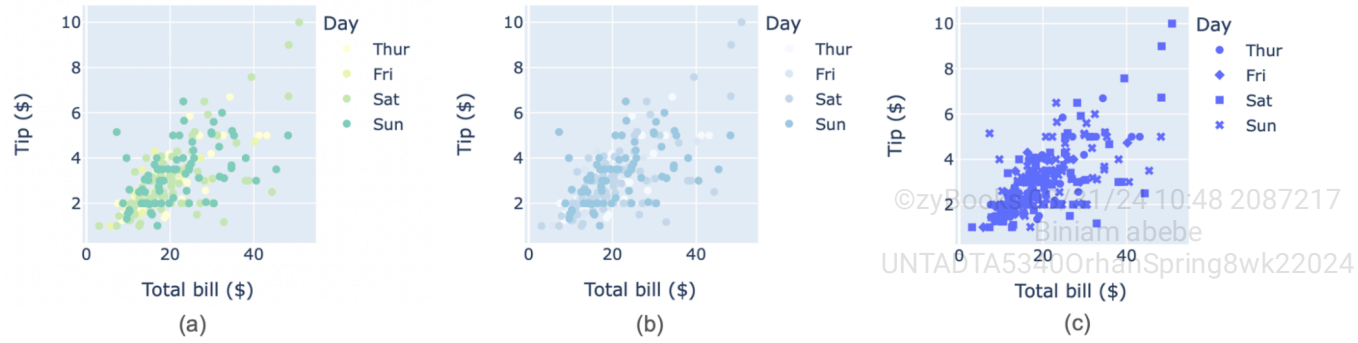
The data for the following plots comes from the free Gapminder dataset, which contains data about economic and health metrics for most countries in the world from 1952 to 2007, under CC-BY license. Each plot contains recommendations of when to use that color palette.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.3.5: Picking visual designs.

Match each plot with the listed design recommendation that was used.



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

If unable to drag and drop, refresh the page.

Differences in contrast **Safe color palette** **Shapes to separate groups**

	Plot (a)
	Plot (b)
	Plot (c)

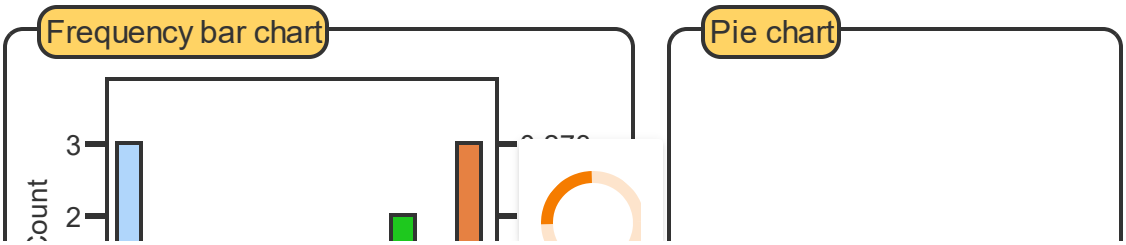
Reset

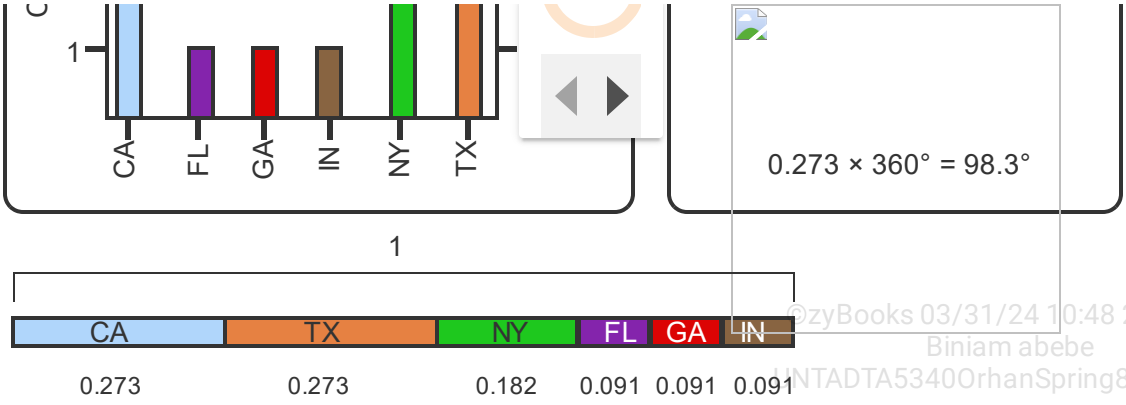
Words of warning: Pie charts

A ***pie chart*** is a circle made of wedges, one wedge for each group. The size of each wedge illustrates the proportion of the dataset in each group. Pie charts are aesthetically pleasing but do a terrible job of displaying information in an easy-to-understand way.

PARTICIPATION ACTIVITY 6.3.6: Making pie charts.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024





©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

On the left, a bar chart is shown with six states and the following counts: CA (3), FL (1), GA (1), IN (1), NY (2), and TX (3). On the right, the same data is represented as a pie chart.

Animation captions:

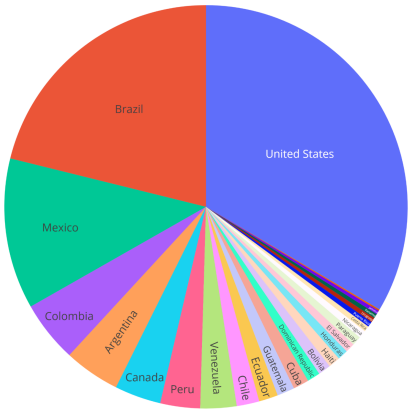
1. A bar chart displays the counts of each category. Scaling by dividing by the total leads to a proportional bar chart.
2. If the bars are laid out end to end, the length of all proportions together is 1.
3. Wrapping that line onto a circle arrives at a pie chart. Each wedge has an angle measure of a category's proportion times 360°, so the angle for California should be 98.3°.

PARTICIPATION ACTIVITY 6.3.7: Interpreting proportion charts.

The data for this question comes from the free Gapminder dataset for the year 2007 under CC-BY license.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

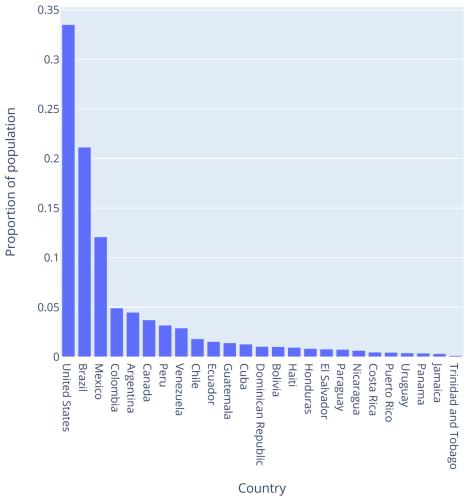
1) Estimate the proportion of people in the Americas who live in Mexico.



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- 10.53%
- 12.09%
- 15.27%

2) Estimate the proportion of people in the Americas who live in Argentina.



- 4.48%
- 5.34%
- 12.35%

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

3) Why is the bar chart better than the pie chart?

- The scale on the vertical axis makes estimating values and noticing differences easier.
- The labels on the horizontal axis are easier to read.
- Both of the above.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

CHALLENGE ACTIVITY 6.3.1: Best practices for visualizing data.

537150.4174434.qx3zqy7

6.4 Tools for visualizing data

Learning goals

- Compare tools for data visualization.
- Explain when each data visualization tool is preferred.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Spreadsheets

Many roles in an organization use data visualization. Ex: Managers, analysts, engineers, data scientists, researchers, and customers all create and consume visualizations while carrying out their roles. Different technologies that have different interfaces and capabilities support users in these different roles.

A ***spreadsheet*** is an application that displays data in a grid and allows calculations and edits within that grid. Ex: Microsoft Excel, Google Sheets, LibreOffice Calc. Spreadsheet applications are in widespread use and effective for quick visualizations, but may be difficult to automate and reproduce. Spreadsheets occasionally have data errors caused by changes from multiple users or automated conversion of values to dates.

Video 6.4.1: Visualizing data with spreadsheets.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

The video shows opening and plotting data energy use and production data for a house collected by the homeowner. This data will be used for each example in the section. For spreadsheets, the steps for plotting are:

1. Open the data using a menu.
2. Highlight the data to be plotted. Here, the features are gas usage and temperature.
3. Insert the desired chart using a menu.

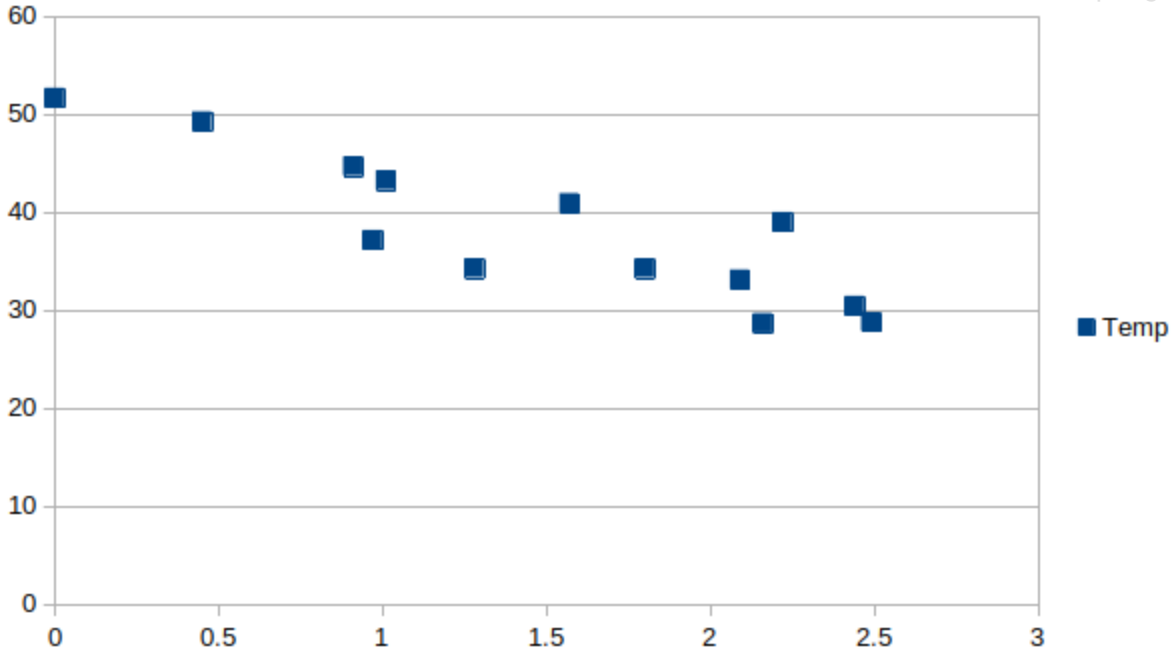
Spreadsheet

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.4.1: Spreadsheet visualization.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

The above video's resulting plot is shown below.



1) What is the most important thing missing from this default plot?

- Labels on the axes
- Coloring the points by the electricity produced by the solar panels
- A trendline

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) What does the plot indicate about the relationship between temperature and gas usage?

- Gas usage is about 51 therms at 0 degrees and decreases as temperature rises.
- Gas usage for heating starts when temperatures are below 50 degrees and increases as the temperature decreases.
- As gas usage for heating increases, the temperature outside decreases.

3) What should be changed about the plot to communicate the effect of outside temperature on gas usage?

- A trendline should be added.
- Other energy sources for heating should be added to the plot.
- The axes should be reversed.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Business intelligence applications

Business intelligence applications (BIAs) are designed to create reports and visualizations from various data sources using a visual interface. Ex: PowerBI, Tableau, Looker. BIAs provide data-wrangling pipeline capabilities that are reproducible. BIAs also provide many standard visualizations. By combining pipelines and visualizations, BIAs are effective for dashboarding. But, BIAs are often not the primary tools used by a data scientist or data engineer.

Video 6.4.2: Visualizing data with PowerBI.

The video below shows opening and plotting data using the same energy data as above. For PowerBI, the steps taken are:

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

1. Open the dataset that is already uploaded.
2. Select the type of plot that is desired.
3. Insert the features in the appropriate features. Here, the features are temperature, natural gas usage, and solar panel electricity production.
4. Turn off the summation of each feature.

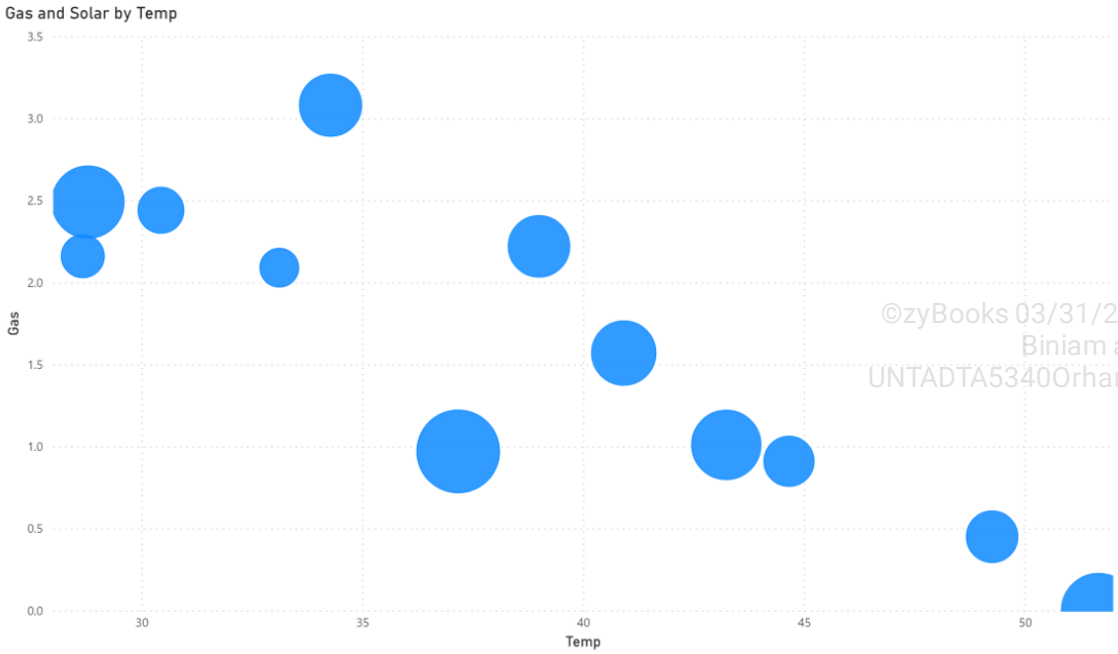
Power BI

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.4.2: Interpreting the plot from PowerBI.

Below is the resulting plot from PowerBI. The size of the points indicates the amount of electricity produced by the solar panels on that day.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

1) Based on this visualization, what impact does electricity production from solar have on gas usage?

- No clear impact
- More production from solar leads to more gas used
- More production from solar leads to less gas used

2) Assuming the data visualized is part of a much larger dataset, which dashboard visualization would help a homeowner understand the impact of recent actions on gas usage?

- Display the scatter plot of gas usage vs. temperature for all the days in the dataset.
- Show the deviation from the linear trend for each day from the most recent week.
- Display the average amount of gas used daily over every day in the dataset.

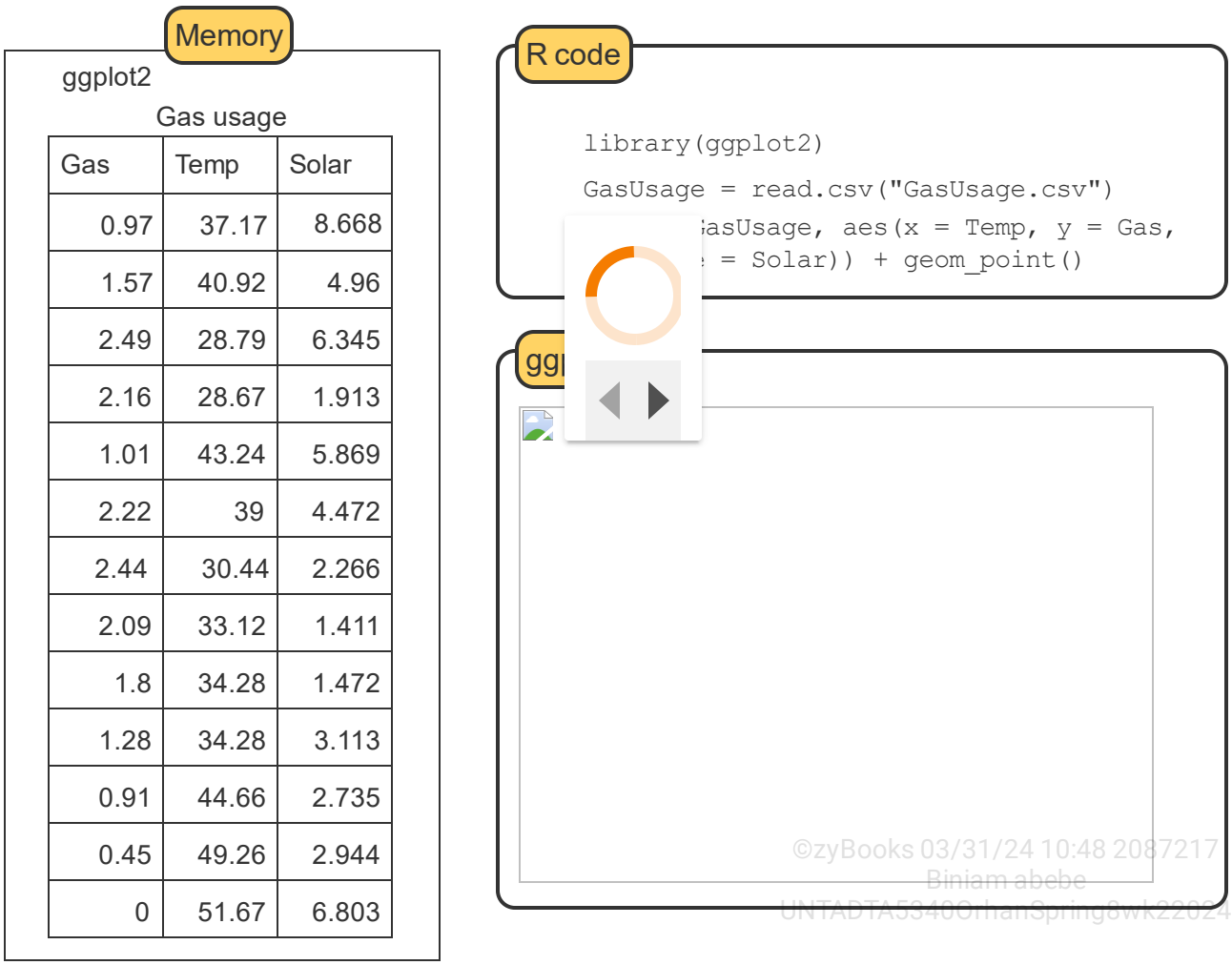
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

R: ggplot2

A ***grammar of graphics*** is a tool that allows the user to define visualizations in a structured fashion. ***ggplot2*** extends the functionality of R to use a grammar of graphics to create visualizations. ggplot2 produces clean, effective visualizations and eases switching from one visualization to another. Many extensions for ggplot2 exist to incorporate other visualizations or types of data. Ex: animations, graphs and networks, ridgeline plots, and interactivity.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.4.3: Visualizing data with ggplot2 in R.



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

Step 1: The code 'library(ggplot2)' appears. The word 'ggplot2' appears in a box labeled memory.
Step 2: The code 'GasUsage = read.csv("GasUsage.csv")' appears. A table of data named 'Gas

usage' with columns Gas, Temp, and Solar appears in the memory box. Step 3: The code 'ggplot(GasUsage, aes(x = Temp, y = Gas, size = Solar)) + geom_point()' appears, then a scatter plot with Temp on the horizontal axis, Gas on the vertical axis, and Solar related to the size of the points appears.

Animation captions:

1. The user loads the package needed for the analysis, ggplot2.
2. The user then loads the data using the read.csv() function.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.4.4: ggplot2 visualization.

1) Which aesthetic highlights the relationship between solar production and gas usage?

- `aes(x=Temp, y=Gas, size=Solar)`
- `aes(x=Solar, y=Gas)`
- `aes(x=Gas, y=Solar)`

2) Which code creates a histogram of the solar production?

- `ggplot(GasUsage, aes(x=Solar)) + geom_point()`
- `ggplot(GasUsage, aes(x=Gas)) + geom_histogram()`
- `ggplot(GasUsage, aes(x=Solar)) + geom_histogram()`

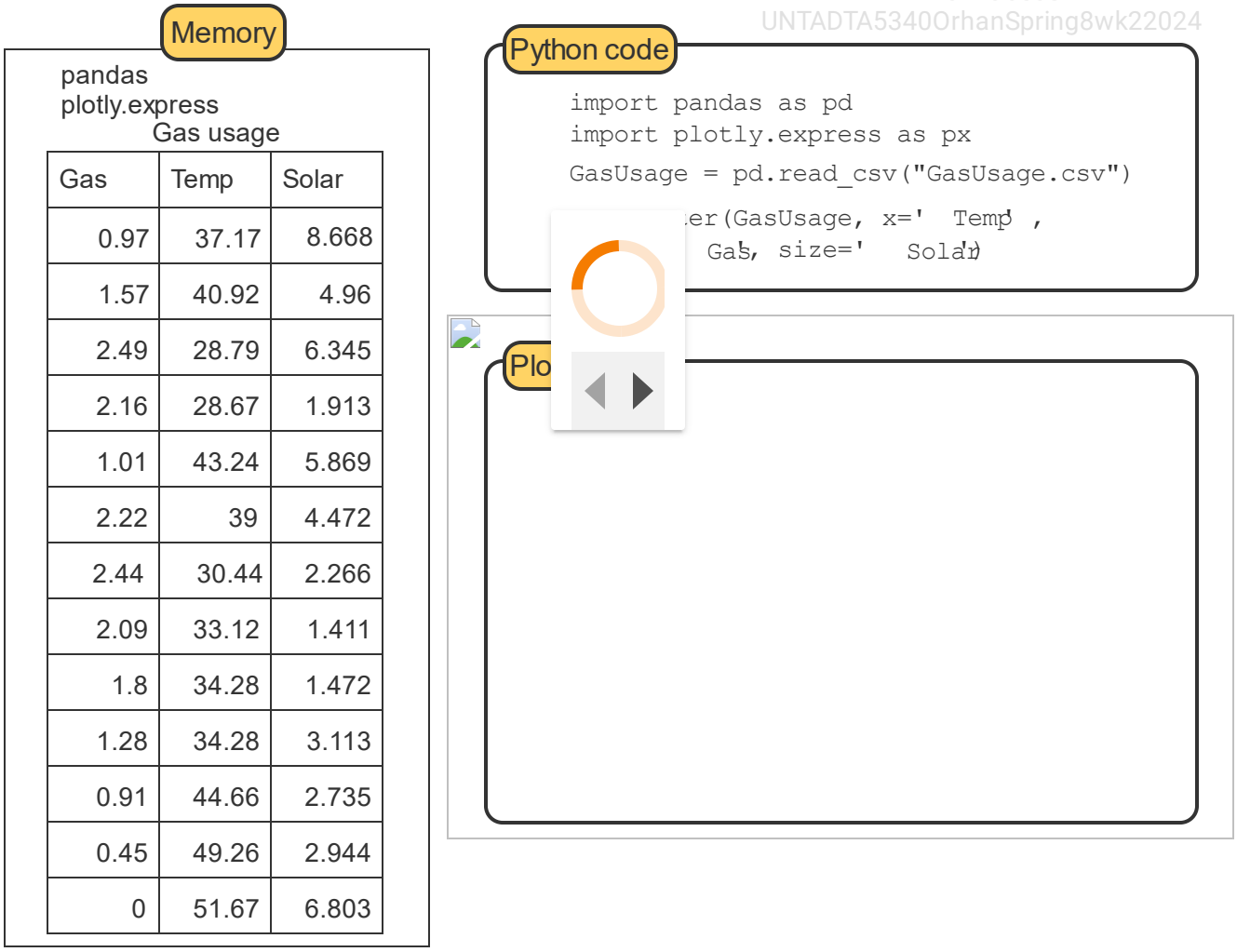
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

R and Python: Plotly

Plotly is a company that provides data visualization services and open source data visualization libraries for Julia, Python and R. Plotly creates visualizations using a javascript library, `plotly.js`, which allows for interaction. This interactivity allows the user to identify interesting data points quicker than `ggplot2` or `matplotlib`.

PARTICIPATION ACTIVITY 6.4.5: Visualizing data with Plotly in Python.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



Gas usage

Gas	Temp	Solar
0.97	37.17	8.668
1.57	40.92	4.96
2.49	28.79	6.345
2.16	28.67	1.913
1.01	43.24	5.869
2.22	39	4.472
2.44	30.44	2.266
2.09	33.12	1.411
1.8	34.28	1.472
1.28	34.28	3.113
0.91	44.66	2.735
0.45	49.26	2.944
0	51.67	6.803

Animation content:

Step 1: The code 'import pandas as pd
import plotly.express as px' appears. The words 'pandas' and 'plotly.express' appear in a box labeled memory. Step 2: The code 'GasUsage = pd.read_csv("GasUsage.csv")' appears. A table of data named 'Gas usage' with columns Gas, Temp, and Solar appears in the memory box. Step 3: The code 'px.scatter(data = GasUsage, x = 'Temp', y = 'Gas', size = 'Solar')' appears, then a scatter plot with Temp on the horizontal axis, Gas on the vertical axis, and Solar related to the size of the points appears.

Animation captions:

1. The user loads the packages needed for the analysis.

PARTICIPATION ACTIVITY 6.4.6: Plotly visualizations.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

1) Which code creates a three-dimensional plot of the same data as above?

- `px.scatter_3d(GasUsage, x='Temp', y='Solar', z='Gas')`
- `px.scatter(GasUsage, x='Temp', y='Solar', z='Gas')`
- `px.scatter_3d(GasUsage, x='Temp', y='Gas', size='Solar')`

2) Which code creates a histogram of temperature?

- `px.histogram(GasUsage, x='Gas')`
- `px.scatter(GasUsage, x='Temp')`
- `px.histogram(GasUsage, x='Temp')`

CHALLENGE ACTIVITY 6.4.1: Tools for visualizing data.

537150.4174434.qx3zqy7

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

6.5 Performing exploratory data analysis

Learning goals

- List the four steps of exploratory data analysis.
- Identify the size of a dataset and classify features as categorical or numerical.
- Explain the relationship between features.
- Categorize a feature's distribution as skewed or symmetric.
- Identify types of missing values.
- Use `pandas` and `seaborn` for exploratory data analysis.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Exploratory data analysis

Exploratory data analysis or ***EDA*** is the process of investigating a dataset to understand what is in the dataset.

Process 6.5.1: Exploratory data analysis.

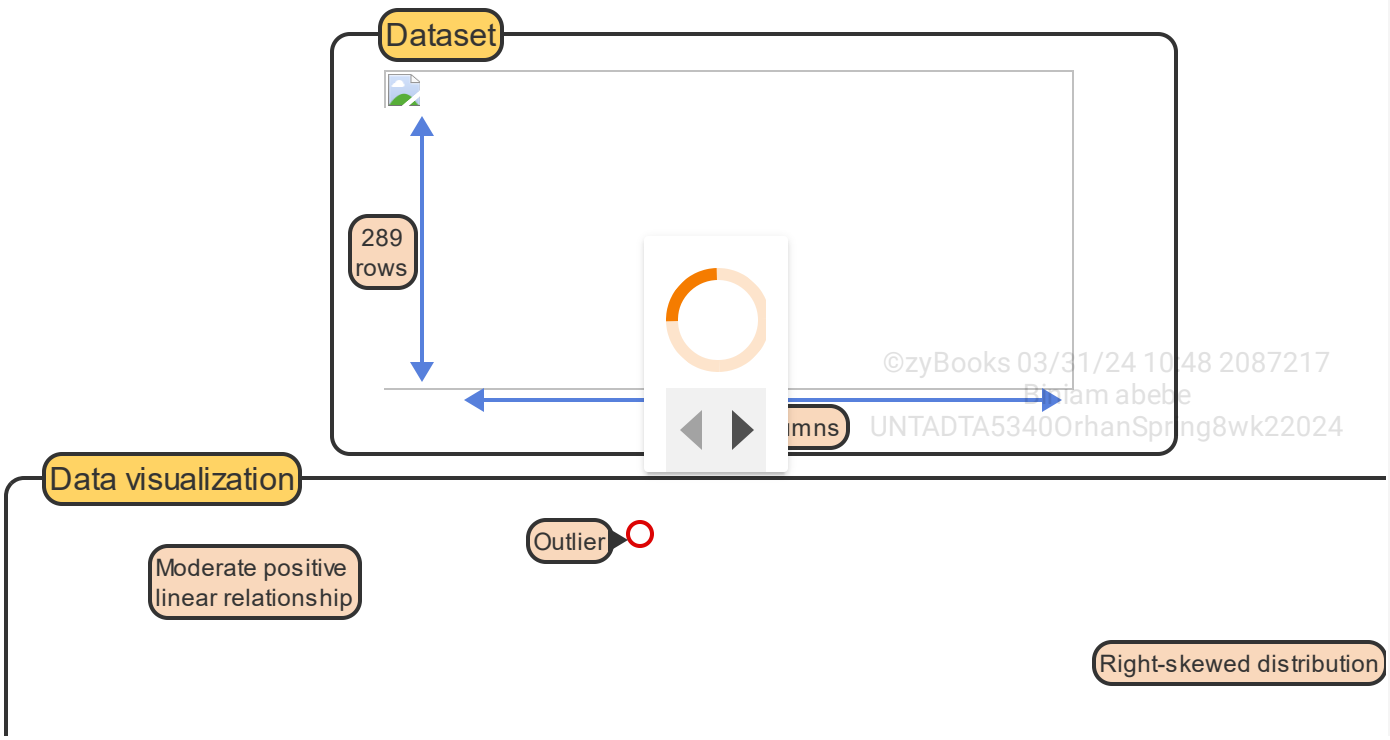
These four steps should be followed in conducting exploratory data analysis:

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

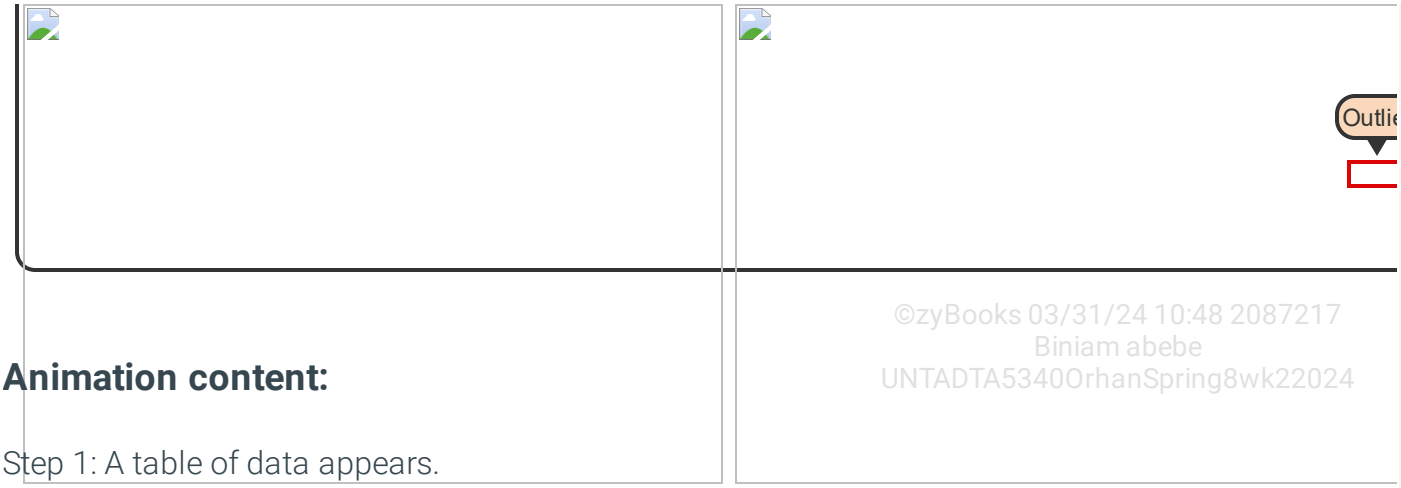
Step	Description
Step 1: Understand the data	Find the size of the dataset (number of rows and columns), identify and categorize the features (categorical, numerical).
Step 2: Identify relationships between features	Find the direction (positive, negative) and strength (strong, moderate, weak) of correlation between the features.
Step 3: Describe the shape of data	Determine the shape of the distribution (symmetric, skewed).
Step 4: Detect outliers and missing data	Find values that are much higher or lower than the rest of the data or values that strongly affect the shape of the data.

©zyBooks 03/31/24 10:48 2087217
Bisiam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.5.1: Simulated data on 289 individuals.



©zyBooks 03/31/24 10:48 2087217
Bisiam abebe
UNTADTA5340OrhanSpring8wk22024



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

Step 1: A table of data appears.
Step 2: Dimensions are added to the table.
Step 3: A scatter plot of Weight vs Height with the best fit line appears.
Step 4: A histogram of income appears.
Step 5: An outlier is highlighted on both the scatter plot and the histogram.

Animation captions:

1. A simulated dataset on 289 individuals has five features: height (in), weight (lbs), income (dollars), handedness (left or right), and heart rate (bpm).
2. (Step 1) The data contains 289 rows and five columns. Handedness is categorical, and heart rate, height, weight, and income are numerical.
3. (Step 2) A positive relationship exists between height and weight. Points fall near a linear trend, but with some scatter, so the relationship is moderate.
4. (Step 3) The distribution of the income feature is skewed to the right, because the tail of the distribution is on the right.
5. (Step 4) The individual with a height of 73 inches and weight of 193 pounds is an outlier. On the histogram, an outlier has an income much larger than the others.

PARTICIPATION ACTIVITY 6.5.2: Exploratory data analysis steps.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

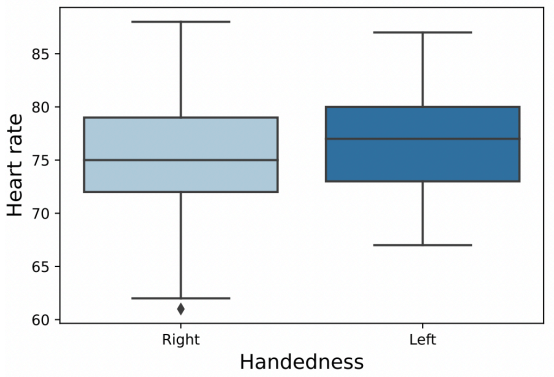
1) The table below is most useful in which step of exploratory data analysis?

	Height (inches)	Weight (pounds)	Income (dollars)	Heart rate (beats per minute)
Min	63.00	93.00	616.78	61
Q1	66.00	131.00	26853.18	72
Median	67.00	151.00	65276.41	75
Mean	67.07	150.76	179509.20	75.30
Q3	68.00	170.00	189069.80	79

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Understanding the data
- Identifying relationships between features
- Detecting outliers and missing data

2) The box plots below are most useful in which step of exploratory data analysis?



- Only detecting outliers
- Both identifying relationships between features and detecting outliers
- Describing the shape of data

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

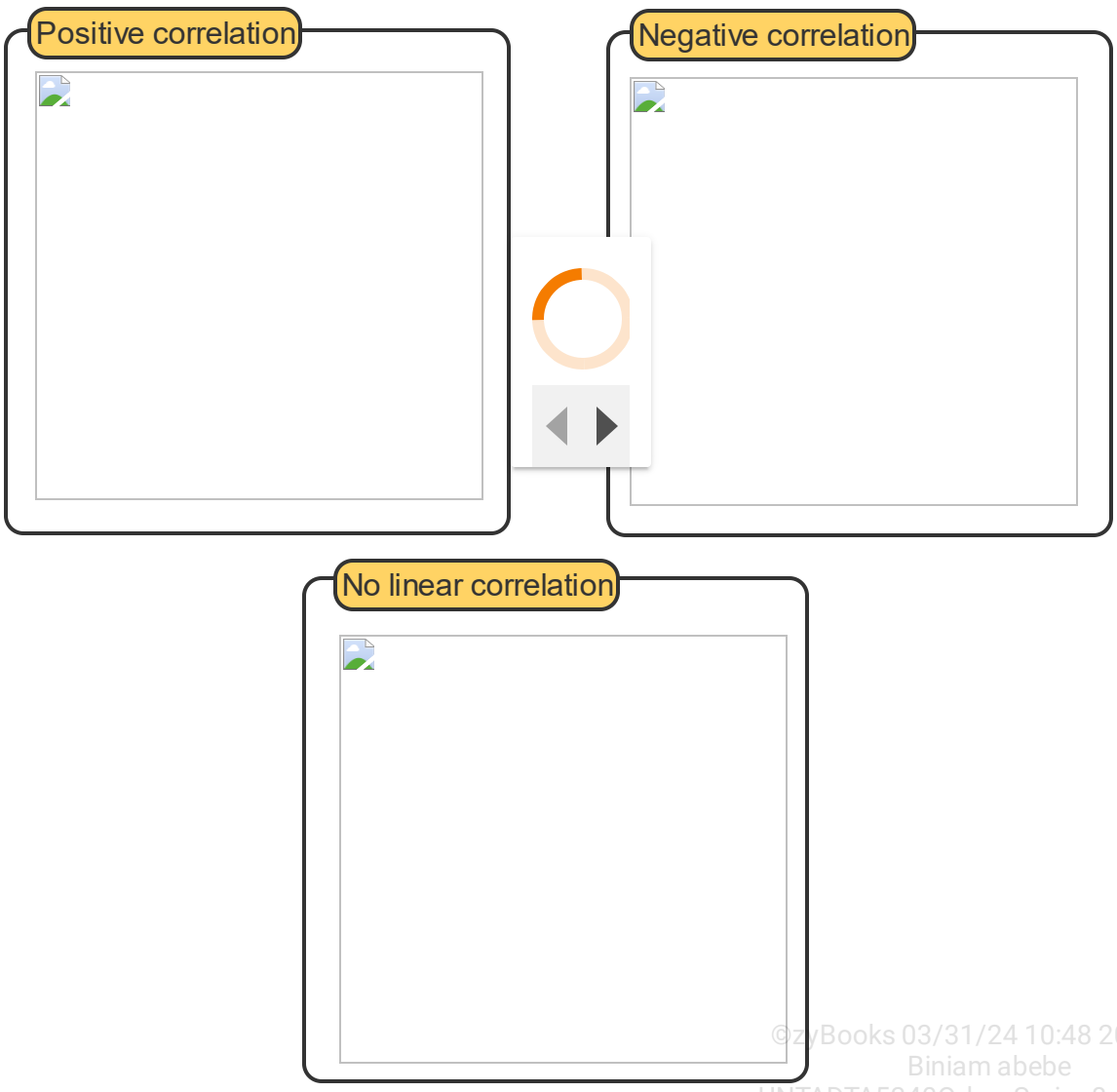
Identifying relationships between features

A data scientist must be aware of how features change with respect to each other. Linear relationships are the simplest type of relationship. Features can also have more complex relationships, especially in the case of categorical features. The most common way to explore these

relationships is with scatter plots. Multiple features being linearly related to each other allow for some models to reduce the amount of data needed to make predictions. ***Correlation*** describes the strength and direction of a linear relationship between numerical features.

PARTICIPATION ACTIVITY 6.5.3: Identifying relationships between features.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

Step 1: A scatter plot with a positive correlation appears. Points tend to follow an upward trend.
Step 2: A scatter plot with a negative correlation appears. Points tend to follow a downward trend.
Step 3: A scatter plot with no pattern. No trend exists in the points.

Animation captions:

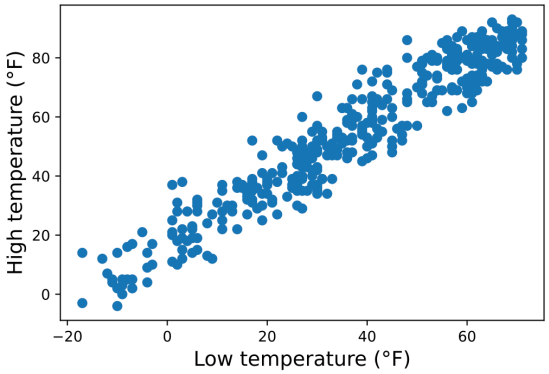
1. Positively correlated features increase as the other feature increases.
2. Negatively correlated features decrease as the other feature increases.
3. Not all features display linear correlation.

PARTICIPATION ACTIVITY 6.5.4: Energy consumption in a home.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

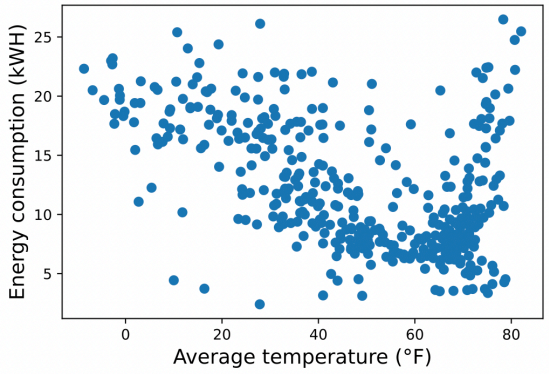
One of the ways to save on electric and gas bills is to identify patterns in energy consumption and temperature in the home.

1) Which description of the relationship plotted below fits best?



- Positively correlated
- Negatively correlated
- No relationship

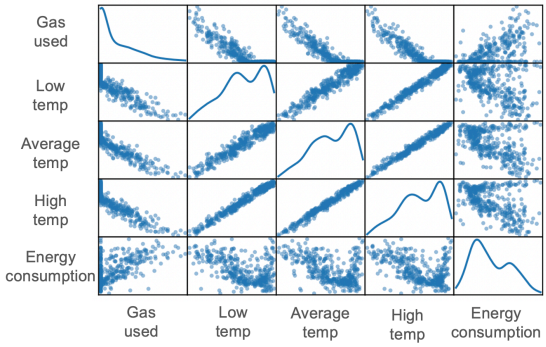
2) Which description of the relationship plotted below fits best?



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Negative and linear
- No relationship
- Nonlinear

3) Using the scatter plot matrix, which features have a negative correlation?



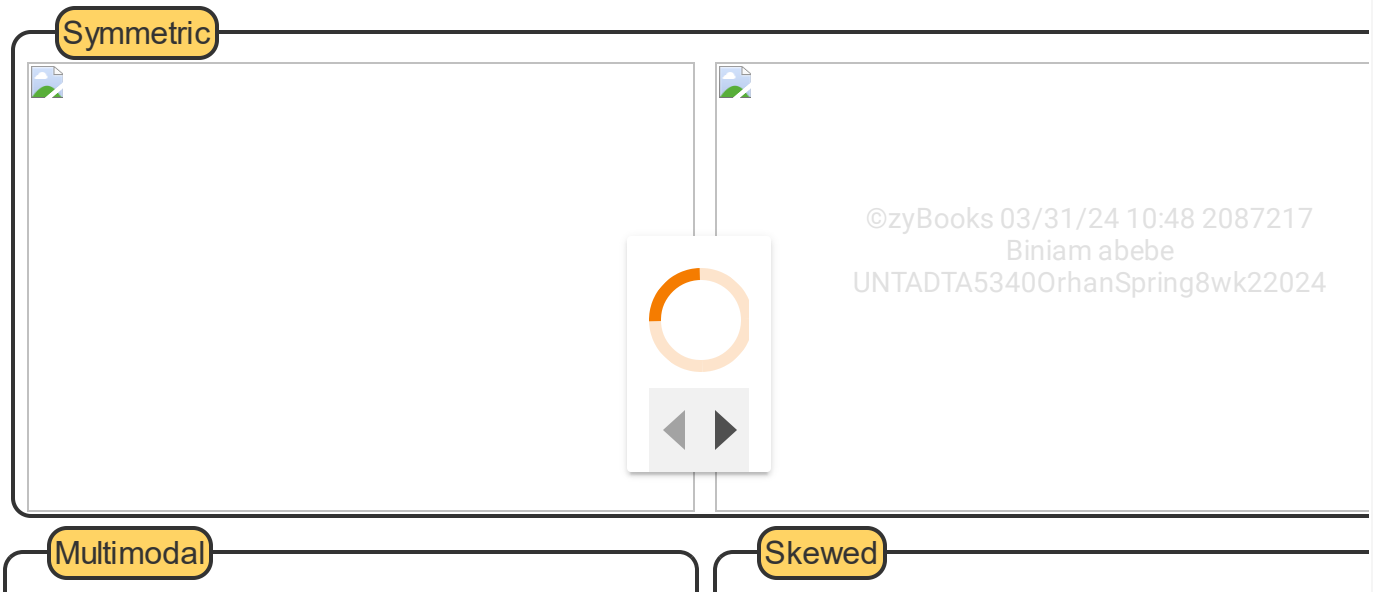
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- ○ High temperature and gas used
- ○ High temperature and low temperature
- ○ Energy consumption and gas used

Describing the shape of data

Many models assume that each feature has a symmetric and unimodal distribution. Deviation from this assumption can affect the model's validity. Determining the distribution's shape can confirm whether certain assumptions are satisfied.

PARTICIPATION ACTIVITY 6.5.5: Describing the shape of data.



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

Four histograms are shown of height, heart rate, weight, and income.

Height: Values range from 63 to 73, with a symmetric distribution. A single peak exists for height between 66 and 70.

Heart rate: Values range from 60 to 90, with a symmetric distribution. A single peak for heart rate exists between 70 and 80.

Weight: Values range from 100 to 200. Two peaks exist: one between 120 and 140, and one between 160 and 180.

Income: Values range from 0 to 5, in hundreds of thousands. Nearly all values are between 0 and 1. A few values are greater than 5.

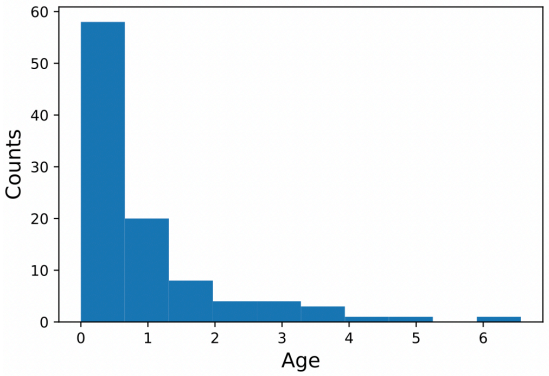
Animation captions:

1. The height and heart rate feature' distributions are symmetric. Symmetric distributions always look the same on the left and right sides.
2. The weight feature's distribution is multimodal. A multimodal distribution has multiple modes or peaks.
3. The income feature's distribution is right-skewed. A skewed distribution has one tail longer

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.5.6: Describing the shape of data.

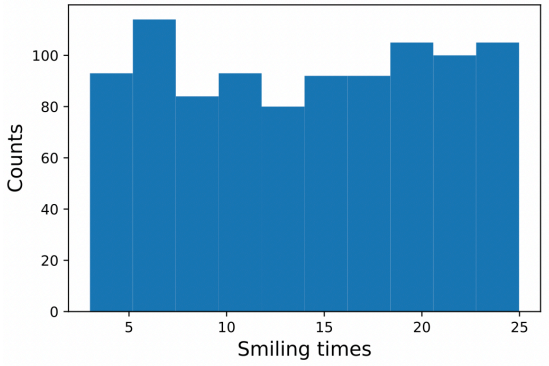
1) The figure below gives the frequency distribution of age for 100 children. What is the best description of the distribution's shape?



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Symmetric
- Right-skewed
- Left-skewed

2) The figure below gives the frequency distribution of simulated data for 958 smiling times of a baby in seconds. What is the best description for the distribution's shape?



- Symmetric and uniform
- Symmetric and multimodal
- Skewed

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Missing data

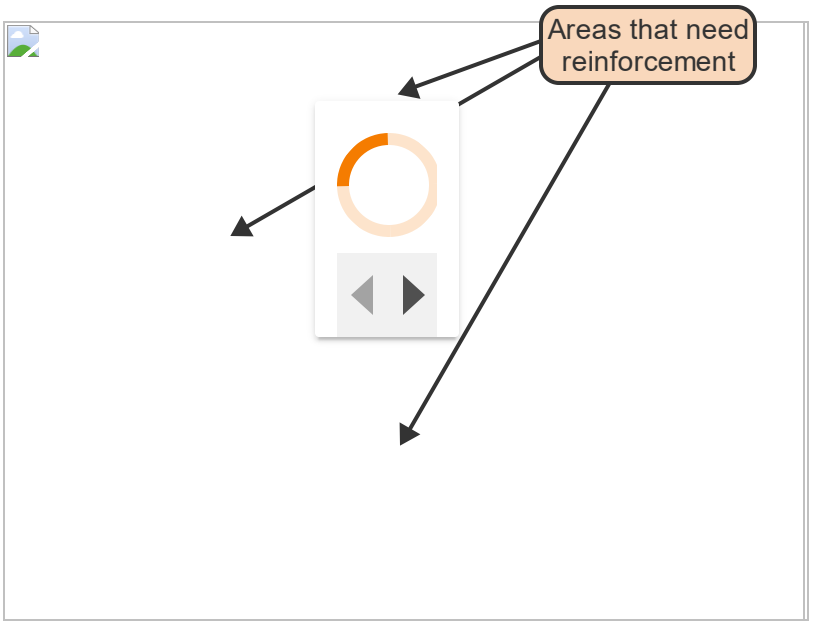
Sometimes, datasets may have missing values, which can come from data entry issues, measurement issues, or processing errors. Missing values can be classified as:

- Values that are ***missing completely at random*** or ***MCAR*** have the same probability of being missing for all cases.

- Values that are ***missing at random*** or ***MAR*** have the same probability of being missing for specific observable cases.
- Values that are ***missing not at random*** or ***MNAR*** have different probabilities of being missing due to unknown reasons.

PARTICIPATION ACTIVITY 6.5.7: Survivorship bias.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



Animation content:

Step 1. Outline of an airplane. Step 2. Red dots are added to the airplane outline.

Animation captions:

1. During WWII, US aircraft were lost due to enemy fire and military leaders considered adding armor to the aircraft.
2. The locations of bullet holes in aircraft that returned were recorded. Military leaders initially thought these areas needed armor.
3. Abraham Wald from the Statistical Research Group correctly reasoned that the aircraft that didn't return were hit in areas that didn't contain many bullet holes. Planes with bullet holes in critical areas were missing not at random.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Image source: Survivorship bias by SlvrKy is licensed under CC BY-SA 4.0, via Wikimedia Commons

PARTICIPATION ACTIVITY 6.5.8: Identify the type of missing data.

If unable to drag and drop, refresh the page.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Missing at random **Missing not at random** **Missing completely at random**

	An accident in processing blood samples at the lab leads to missing data for those patients.
	A subject missed an appointment due to illness for a study that includes information about the subject's general health.
	Customer satisfaction survey responses mostly fall in extremely positive or extremely negative categories.

Reset

Exploratory data analysis in Python

The `pandas` package has three functions in a dataframe to aid with initial exploration of the features in a dataframe and the numerical summaries of the features.

Table 6.5.1: Dataset summary functions.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Function	Behavior
`df.shape`	`.shape` returns the dataframe's dimensions and displays as (number of instances, number of features). `df.shape` is useful when code needs one of these dimensions.

`df.info()`	`.info()` displays the name, number of non-null values, and type of each feature in the dataframe.
`df.describe(include = "all")`	`.describe()` displays summary statistics (count, mean, standard deviation, min/max, and quartiles) for each numerical feature. Including `include = "all"` displays the count, number of categories, and mode's name and frequency for categorical features.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

The `pandas` package also contains several methods for visualizing all the features in a dataframe at once. In the tables below, `df` represents the dataframe.

Table 6.5.2: Many relationship visualization in pandas.

Function	Behavior
`df.hist()`	`df.hist()` plots a histogram for every column in the dataframe.
`df.boxplot()`	`df.boxplot()` plots a box plot for every column in the dataframe.
`pd.plotting.scatter_matrix(df)`	`pd.plotting.scatter_matrix(df)` plots every pair of numerical features as an individual scatter plot. For more control, seaborn provides the function `sns.pairplot(df)`.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Data exploration in Python.

Full screen

The data set below contains data about the energy usage and demands at one of the author's houses.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Change the features in the commented scatter plot code to investigate the relationship between two features.
- Change the feature in the commented histogram code to investigate the relationship between two features.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.5.9: Exploratory data analysis with Python.

1) Which `pandas` method should be used to explore the shape of numerical features in a dataframe?

- `df.boxplot()`
- `df.describe()`
- `df.hist()`

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) Which `pandas` method should be used to detect unusual values of numerical features in a dataframe?

- `df.boxplot()`
- `df.describe()`
- `df.hist()`

3) Which `pandas` method can be used to detect missing values in a dataframe?

- `df.boxplot()`
- `df.info()`
- `df.shape()`

CHALLENGE ACTIVITY 6.5.1: Performing exploratory data analysis.

537150.4174434.qx3zqy7

6.6 Detecting outliers

Learning goals

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Define outliers, high leverage points, and influential points.
- Use Tukey's fences and z-scores to identify outliers.

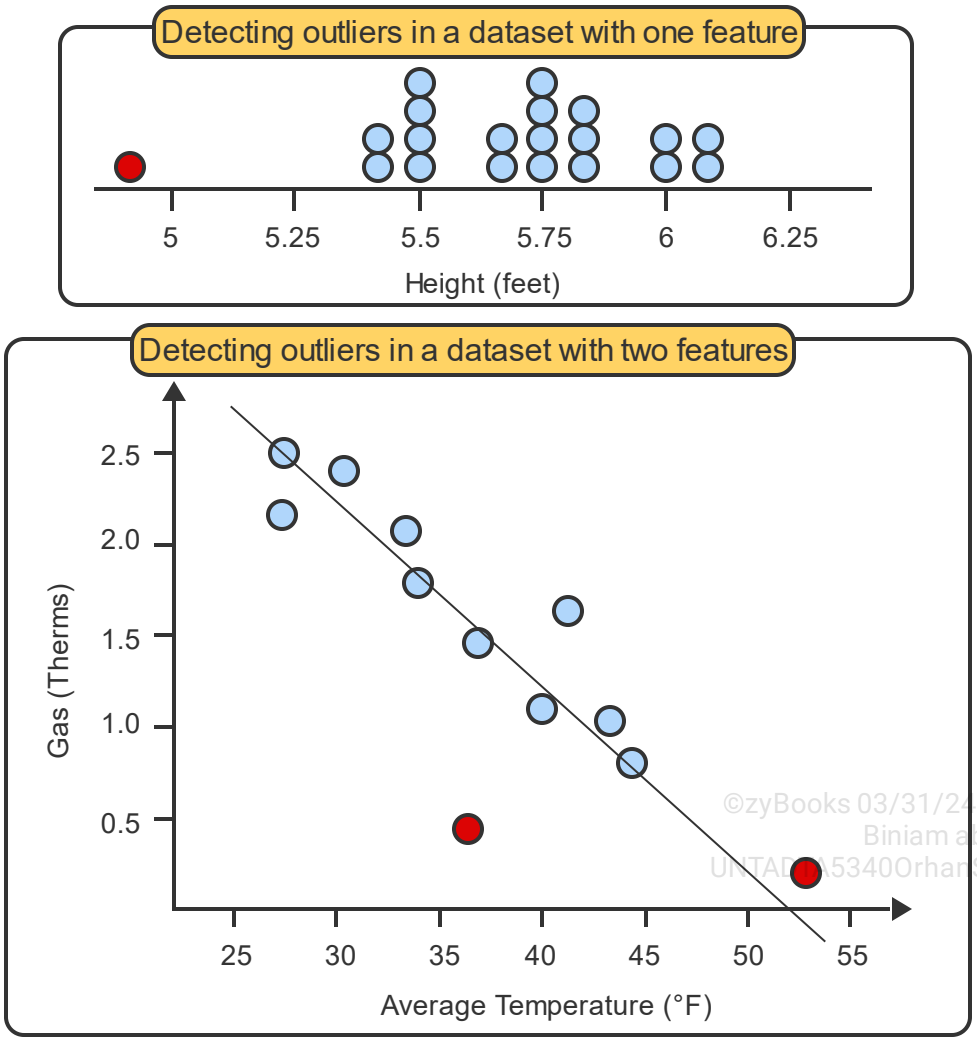
- Explain possible causes of outliers

Overview of outliers

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

An ***outlier*** is an instance that is separated from the rest of the dataset. This separation can be in a single dimension or a combination of dimensions. Outliers have the potential to cause issues with models that use an average or similar function. Hence, identifying outliers is important for understanding their possible effects on resulting models.

PARTICIPATION ACTIVITY 6.6.1: Detecting outliers.



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

Step 1:A dot plot of heights appears with a red dot far to the left. Step 2: A scatter plot with the line of best fit. Step 3 and 4: Two points far away from the line are highlighted in red.

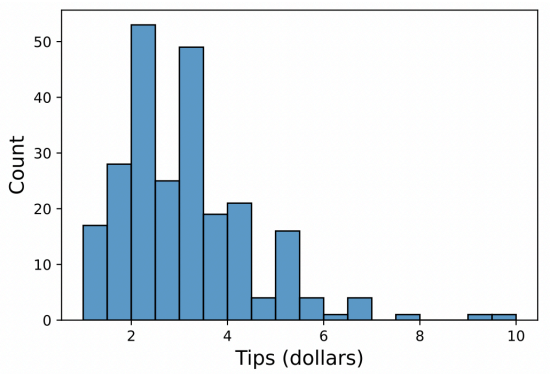
Animation captions:

1. A 4'11" student joins a basketball team and is much shorter than the other students on the team. Thus, the height of 4'11" is an outlier in the dataset.
2. A house has a linear relationship between the temperature and the amount of natural gas used to heat the house.
3. But, two outliers are far from the line. One outlier has low gas use, compared to other instances with similar temperatures.
4. One outlier has both low gas use and high average temperature compared to the other instances.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.6.2: Detecting outliers using graphs.

1)

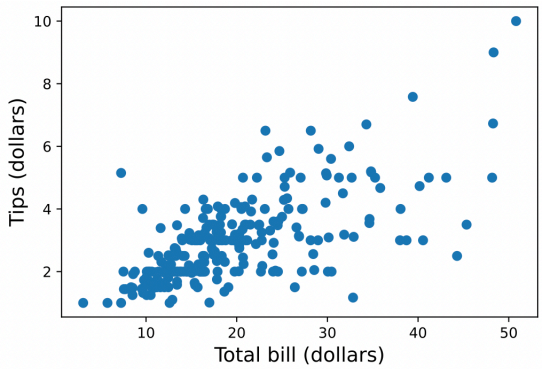


Where are the outliers on this histogram of tips received by a waiter?

- $6.50
- $10
- 53

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2)



Is the point at (7.25, 5.15) an outlier? Why or why not?

- Yes, because this point falls outside the data's general trend.
- No, because this point is not too far from the rest of the data.
- No, because the point at ($50.81, $10) is clearly the outlier for this dataset.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

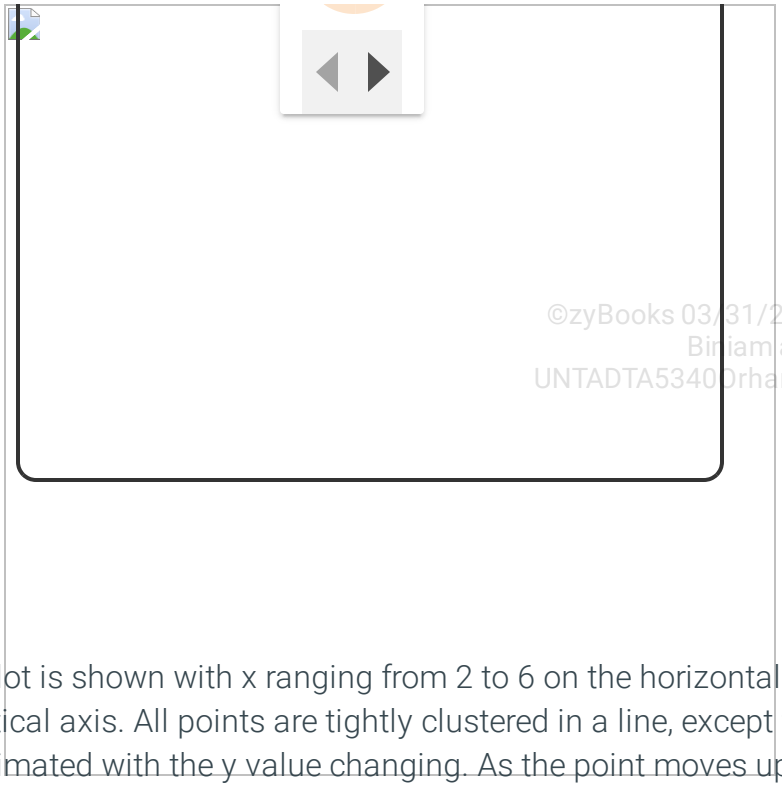
Impact of outliers on models

Outliers can have an oversized impact on statistics and models. Ex: The mean and standard deviation are subject to being affected by outliers. A linear model is subject to change by the values of an outlier. The ***leverage*** of an instance describes that instance's ability to single-handedly change the parameters of a model. A ***high leverage point*** is an instance with an input value that is far out in at least one feature's distribution. A high leverage point could change a model a great deal based on the instance's output value An ***influential point*** is a high leverage point whose presence changes a model a lot. But, not all high-leverage points are influential.

PARTICIPATION ACTIVITY 6.6.3: Leverage points and linear models.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Effect of a high leverage point on a linear model



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation content:

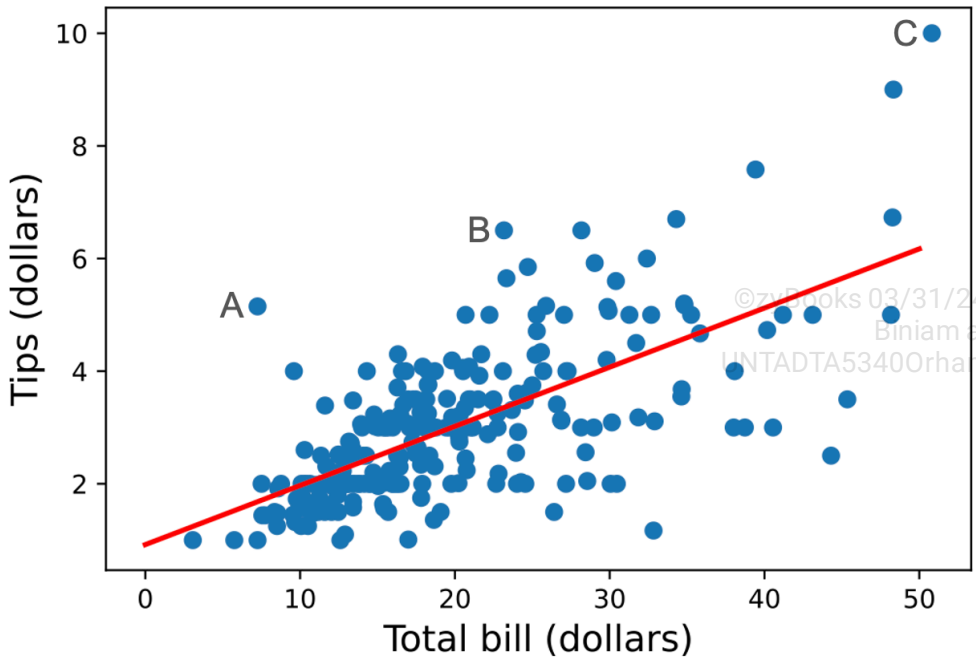
Static image: A scatter plot is shown with x ranging from 2 to 6 on the horizontal axis and y ranging from 15 to 25 on the vertical axis. All points are tightly clustered in a line, except for one. A point with a high x value is animated with the y value changing. As the point moves up and down, the slope of the line changes.

Animation captions:

1. A linear model matches this dataset well.
2. If a point with a large amount of leverage is added, the point can have a strong influence on the model.
3. This effect can be seen as the point moves vertically.

PARTICIPATION ACTIVITY 6.6.4: Leverage and influence.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

1) Order the labeled points from highest leverage to lowest leverage.

- A, B, C
- C, A, B
- B, A, C

2) Is C influential? Why?

- Yes, because C is an outlier in both features and is far from the best-fit line.
- Yes, because C is an outlier.
- No, because C is close to the best-fit line.

Parametric detection of outliers

Two easy to implement methods of detecting outliers are Tukey's Fences and z-scores. Both methods depend on the fact that outliers will lie on the far edge of the feature's distribution. Like with missing data, outliers should be investigated to identify the possible causes of the outlier and the implications the outliers have on the analysis.

Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Table 6.6.1: Outlier detection.

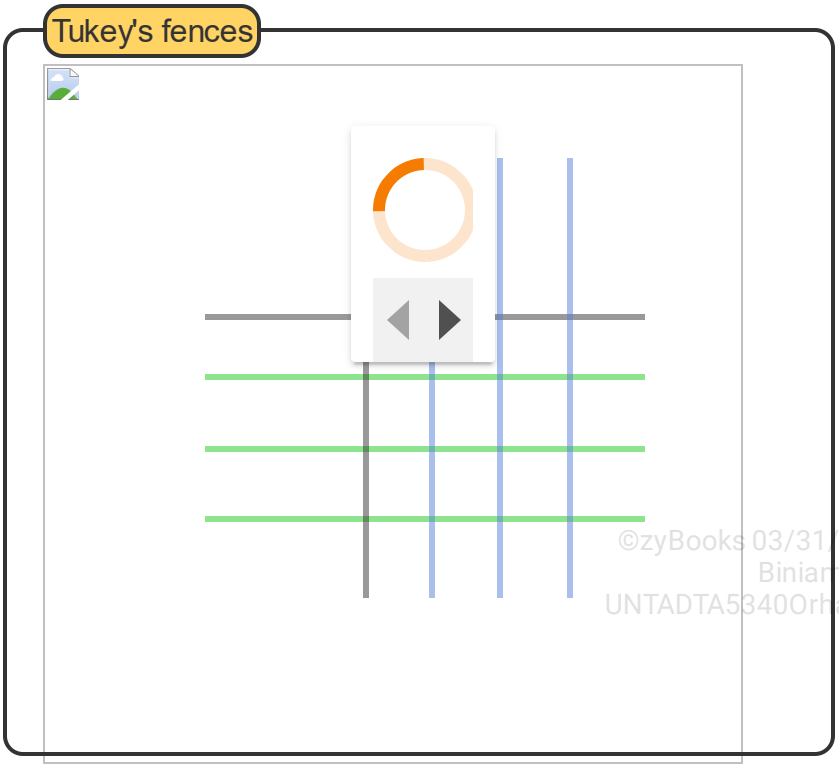
Method	Description

Tukey's Fences	Often used to determine outliers in box plots. 1. Calculate the interquartile range, \(\text{IQR} = Q_3-Q_1 \) for a feature. 2. Classify all points that fall 1.5IQR above \(Q_3 \) or 1.5IQR below \(Q_1\) as outliers.
z-scores	1. Calculate the z-score \(z= \dfrac{\text{value}-\text{mean}}{\text{standard deviation}} \) for each value. 2. Classify all points that have a z-score of \(|z| >3\) as outliers.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.6.5: Detecting outliers using Tukey's fences.

Tukey's fences



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Step 1: A scatter plot of points. One point is well below the rest and one point is well to the right of the rest. A point at (-1.5, 2) is relatively alone in the upper left corner of the plot.
Step 2: Boxplots appear above and to the right of the scatter plot. The point below and the point to the right both appear as outliers.
Step 3: Horizontal lines representing the mean of y and 1,2, and 3 standard deviations below the mean appear. The point below is below the bottom line. Similar vertical lines for x appear, and the point to the right is beyond the 3 standard deviations line.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation captions:

1. A scatter plot is useful in detecting outliers in a dataset with two features.
2. Using Tukey's fences, one point for each feature is identified as an outlier.
3. Plotting the mean and 1, 2 and 3 standard deviations away from the mean for each feature identifies the same two points as outliers.

PARTICIPATION ACTIVITY 6.6.6: Parametric outlier detection.

1) A feature has a mean of 6 and a standard deviation of 3. What is the z-score for the instance with the value 10.5? Should this value be classified as an outlier?

   - 1.25, No
   - 1.5, No
   - 4.5, Yes

2) The first quartile of a feature is 3, and the third quartile is 7. Would the instance with the value 14 be considered an outlier? Why?

   - Yes, because \(14 > 11 \).
   - Yes, because \(14 > 13 \).
   - No, because \(14 < 15 \).

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

3) Is (-1.5, 2) an outlier in the dataset shown in the animation above? Why?

- Yes, because (-1.5, 2) is well-separated from the rest of the data.
- No, because (-1.5, 2) is well within the usual ranges for both x and y.
- No, because other values are close enough.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Dealing with outliers

Identifying the cause of an outlier helps an analyst know how to deal with the outlier during data wrangling. Care should always be taken when dealing with outliers, because some outliers occur naturally and should not automatically be excluded. Ex: Energy production from solar panels in the Northern Hemisphere is much higher in June than in December. Investigating outliers further is a good practice. If no explanation for the occurrence of an outlier can be found, the outlier should probably not be removed.

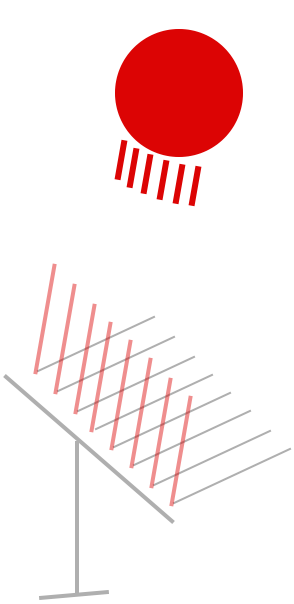
Table 6.6.2: Dealing with outliers.

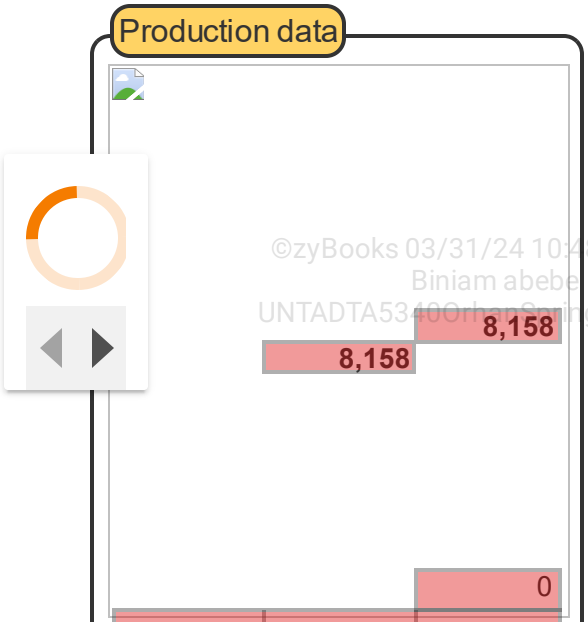
©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Cause of outlier	Description of cause	Method of cleaning
Data entry error	Human error entering a value into the dataset	If the intended value can be inferred, correct the value.
Measurement error	Error in an instrument recording the value	If the error is consistently biased from the actual value and can be identified, remove the bias.
Data processing error	Accident in data manipulation	If feasible, redo the manipulation to remove the error.
Sampling error	Error in combining data from multiple populations or groups	Explore dataset for outliers that may be from different populations or groups.
Natural outlier	Not an error. This value, though extreme, belongs in the data.	Models should be checked for stability of results with and without the outlier.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.6.7: Causes of outliers in solar data.





©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

6/30/2021	12.100	0

Animation content:

Step 1: A sketch of a solar panel and the sun appears.
Step 2: A table of Data and Electricity produced appears. Dates range from November 21 to December 2. 8,158 is highlighted.
©zyBooks 03/31/24 10:48 2087217
Bikram Abebe
UNTADTA5340OrhanSpring8wk22024
Step 3: A new column of data, Electricity produced yesterday appears. All values have moved up a row. 8,158 is still highlighted. 0 is at the bottom of the column and is highlighted in red.
Step 4: The sun moves higher in the sketch. A new row is added with date June 30. This entire row is highlighted in red.

Animation captions:

1. A homeowner in the United States is recording the amount of electricity their solar panels produce.
2. Each day the homeowner records the production for that day. But one day, a data entry error is made when a comma is used instead of a decimal point.
3. Later, a new variable containing the production from the day before is created. However, a data processing error is made, because the first day has no data from the day before.
4. A sampling error is made when production from a day in June is added, when days are longer and energy production is higher.

PARTICIPATION ACTIVITY 6.6.8: Types of outliers.

Match the type of cause to the description.

If unable to drag and drop, refresh the page.

Natural outlier **Data entry error** **Data processing error**

	An outlier is generated when the analyst does a transformation that divides by a small number.
	A survey respondent enters a percentage instead of a decimal.
	One student scores well above the rest on a test.

©zyBooks 03/31/24 10:48 2087217
Bikram abebe
UNTADTA5340OrhanSpring8wk22024

Reset

CHALLENGE ACTIVITY 6.6.1: Detecting outliers.

537150.4174434.qx3zqy7

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

6.7 Case study: Palmer penguins

Learning goals

- Classify features as categorical or numerical.
- Examine relationships between features.
- Examine distributions of features.
- Identify outliers or missing values.
- Complete the exploratory data analysis process using `pandas` and `seaborn`.

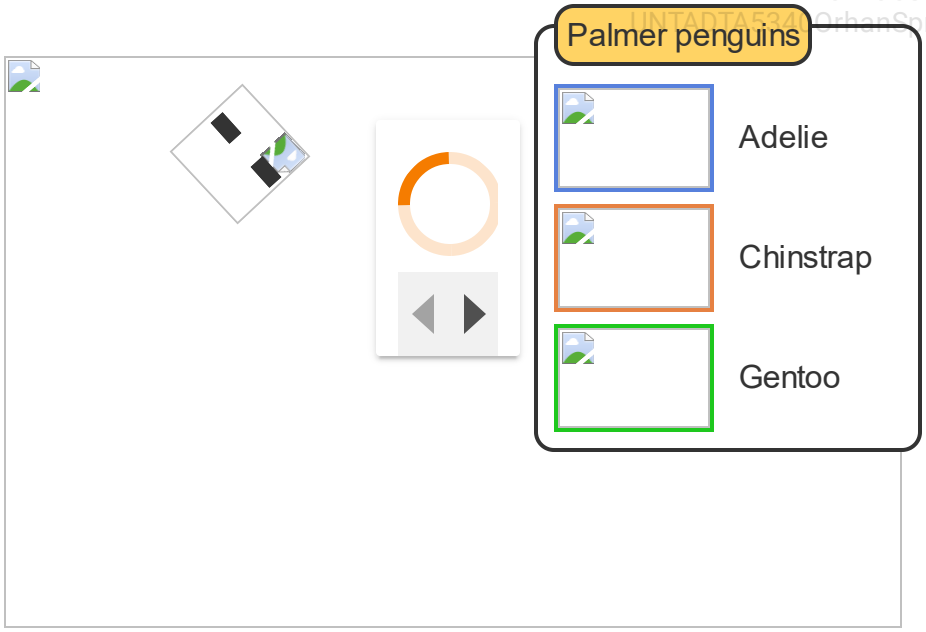
Exploratory data analysis with the Palmer penguins

The Palmer penguins dataset describes body measurements from a sample of 344 penguins living near the Palmer Archipelago in Antarctica. Three species of penguins are included in the dataset: Adelie, Chinstrap, and Gentoo. The Palmer penguins are classified as species of least concern, meaning that these penguin species have large, established populations that are not at risk of extinction. But, changes in global climate and sea levels may threaten penguin populations in Antarctica in the future. Studying penguin populations today gives ecologists a reference point to monitor changes and possible threats.

Exploratory data analysis is an important part of the data science lifecycle. For the Palmer penguins data, exploratory data analysis can highlight relationships between penguin features, such as flipper length and body mass, or identify unique characteristics of each species.

PARTICIPATION ACTIVITY 6.7.1: Studying penguins in the Antarctic.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



Animation content:

Static image: Map of Antarctica with location of Palmer Station highlighted. Images of each penguin species appear.
Step 1: The Palmer penguins dataset was collected by researchers studying penguin populations at Palmer Station in Antarctica.
Step 2: Adelie penguins have a characteristic white ring around their eyes, and are the most widespread penguin species on the Palmer Archipelago.
Step 3: Chinstrap penguins are known for the distinctive narrow band under their heads, resembling a strap around a chin..
Step 4: Gentoo penguins have a wide white stripe across the head.
Step 5: Understanding body characteristics of each penguin species allows for improved population monitoring using satellite images.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Animation captions:

1. The Palmer penguins dataset was collected by researchers studying penguin populations at the Palmer Station in Antarctica.

2. Adelie penguins have a characteristic white ring around their eyes and are the most widespread penguin species on the Palmer Archipelago.
3. Chinstrap penguins are known for the distinctive narrow band under their heads, resembling a strap around a chin.
4. Gentoo penguins have a wide white stripe across the head.
5. Understanding body characteristics of each penguin species allows for improved population monitoring using technology like satellite images.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Data source: Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 0.5281/zenodo.3960218.

Image sources: Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081. https://doi.org/10.1371/journal.pone.0090081 and Google, © 2022.

PARTICIPATION ACTIVITY 6.7.2: Exploratory data analysis.

Researchers plan to use exploratory data analysis on the Palmer penguins dataset. Choose the task or type of plot that matches each step in the exploratory data analysis process.

1) Step 1: Understand the data.

- ○ Identify correlations between features.
- ○ Identify features as categorical or numerical.
- ○ Identify unusual values of a feature.

2) Step 2: Identify relationships between features.

- ○ Bar chart
- ○ Box plot
- ○ Scatter plot

3) Step 3: Describe the shape of data.

- ○ Histogram
- ○ Pie chart
- ○ Stacked bar chart

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

4) Step 4: Detect outliers and missing values.

- ○ Box plot
- ○ Density plot
- ○ Histogram

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Step 1: Understand the data

The first step of exploratory data analysis is to explore the number of instances, number of features, and type of each feature. A ***data dictionary*** contains feature names and descriptions for all features in a dataset. Data dictionaries are often presented as tables or bulleted lists and are important resources for understanding a dataset. The Palmer penguins dataset contains data from 344 penguins using the features listed below.

Table 6.7.1: Data dictionary for the Palmer penguins dataset.

Name	Description
`species`	Penguin species (Adelie, Chinstrap, or Gentoo)
`island`	Island where the penguin was sampled (Biscoe, Dream, or Torgersen)
`bill_length_mm`	Penguin's bill length, in millimeters (mm)
`bill_depth_mm`	Penguin's bill depth, in millimeters (mm)
`flipper_length_mm`	Penguin's flipper length, in millimeters (mm)
`body_mass_g`	Penguin's body mass, in grams (g)
`sex`	Penguin's sex (male or female)

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Understanding the penguins dataset with Python. Full screen

The Python code below loads and summarizes the penguins dataset.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.7.3: Understanding the penguins dataset.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

1) How many instances are in the penguins dataset?

- 3
- 7
- 344

2) How many features are in the penguins dataset?

- 2
- 7
- 344

3) Which list contains the categorical features in the dataset?

- `species, island`
- `island, body_mass_g, sex`
- `species, island, sex`

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Step 2: Identify relationships between features

Once the dataset's size and features are understood, the relationships between features should be explored. Data visualizations like scatter plots give a visual representation of how pairs of features change together. Scatter plots and other visualizations may suggest possible models for making predictions or describing relationships. Ex: If two features have a strong linear relationship, then a model based on a straight line might be a good fit.

Identifying relationships in the penguins dataset with Python.

Full screen

The Python code below creates data visualizations for features in the penguins dataset.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to explore the relationship between `island` and `species`.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.7.4: Identifying relationships in the penguins dataset.

1) Which island has only Adelie penguins?

- Biscoe
- Dream
- Torgersen

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) Which is the only island to have Gentoo penguins?

- Biscoe
- Dream
- Torgersen

3) As bill length tends to increase, body mass tends to _____.

- decrease
- increase
- stay the same

4) The correlation between bill length and body mass _____ on species.

- depends
- does not depend

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Step 3: Describe the shape of data

The shape of a feature's distribution is another important consideration when exploring a dataset. Some models in data science make assumptions about the distribution of a feature. Ex: A common assumption is that features in a model are approximately normally distributed, or have a symmetric, bell-shaped distribution. Most models can handle some departures from symmetry, but features with extreme skewness are more difficult to model.

Describing distributions in the penguins dataset with Python.

Full screen

The Python code below creates data visualizations for features in the penguins dataset.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.7.5: Describing the shape of features in the penguins dataset.

1) Which species tends to have the smallest bill depths?

- Adelie
- Chinstrap
- Gentoo

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) How should the distribution of bill depth by species be described?

- Some species have a symmetric distribution, but not all.
- All species have a skewed right distribution.
- All species have a symmetric distribution.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

3) Which feature has the most overlap between Chinstrap and Gentoo penguins?

- Bill depth
- Bill length
- Body mass
- Flipper length

Step 4: Detect outliers and missing data

Outliers and missing values can both impact a model. If only a few instances have missing data, those instances may be set aside without having a negative impact on a model. But, if many instances have missing data, techniques like imputation may be needed.

Outliers often have a disproportionate impact on a model: one instance with an extreme value can shift a model away from the rest of the dataset. Outliers may be identified using Tukey's rule with box plots, or by transforming features to z-scores. Once identified, outliers should not be discarded from a dataset. Instead, data scientists often fit models with and without outliers to determine whether the outliers are having a large impact or not.

Identifying missing values in the penguins dataset with Python. Full screen

The Python code below searches features in the penguins dataset for missing values.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to identify penguins with missing values for sex.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.7.6: Missing values in the penguins dataset.

1) How many penguins are missing information about bill length?

- 0 penguins
- 2 penguins
- 342 penguins

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) What type of missing data are the numerical measurements for these two penguins?

- Missing completely at random
- Missing at random
- Missing not at random

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

3) What type of missing data are the missing values for sex?

- Missing completely at random
- Missing at random
- Missing not at random

Detecting outliers in the penguins dataset with Python.

Full screen

The Python code below creates data visualizations and calculates z-scores for features in the penguins dataset.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 6.7.7: Detecting outliers in the penguins dataset.

1) Which penguin would be considered an outlier based on bill length?

- The Chinstrap penguin with a 58 mm long bill
- The Adelie penguin with a bill that is 21.5 mm deep
- The Gentoo penguin with a 59.6 mm long bill

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

2) Which Gentoo penguin would be considered an outlier based on bill length and bill depth?

- A
- B
- C

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

6.8 LAB: Visualizing mpg data using matplotlib

The dataset `mpg` contains information on miles per gallon (mpg) and engine size for cars sold from 1970 through 1982. The dataset has the features `mpg`, `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `model_year`, `origin`, and `name`.

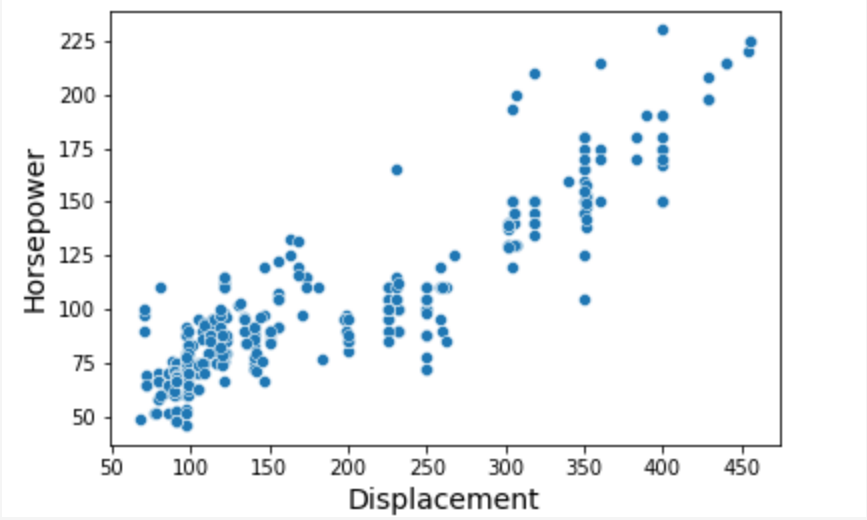
- Load the dataset `mpg`
- Create a new dataframe using the columns `weight` and `mpg`
- Use `matplotlib` to make a scatter plot of `weight` vs `mpg` labelling the x-axis `Weight` and the y-axis `MPG`

If `displacement` and `horsepower` were used instead of `weight` and `mpg`, the output would be:

```
     displacement  horsepower
0           307.0       130.0
1           350.0       165.0
2           318.0       150.0
3           304.0       150.0
4           302.0       140.0
..            ...         ...
393         140.0        86.0
394          97.0        52.0
395         135.0        84.0
396         120.0        79.0
397         119.0        82.0

[398 rows x 2 columns]
```

©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024



©zyBooks 03/31/24 10:48 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

537150.4174434.qx3zqy7

LAB ACTIVITY | 6.8.1: LAB: Visualizing mpg data using matplotlib | 0 / 1

main.py

```
1 Loading latest submission...
```

Develop mode | **Submit mode**

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

If your code requires input values, provide them here.