Data Overview

Proprietary Statement

Data from the CourseKata project are not public. These data are proprietary and are only to be used for the purposes of the American Statistical Association's DataFest.

By using this data, you agree to:

1. During participation in ASA's DataFest, store and manage the data securely and privately. This means you may NOT upload the data to websites that may provide access to the data to anyone other than yourselves. (For example, if using Tableux orGithub, make sure privacy settings prevent others from seeing or accessing the data.) If you are not sure, then don't use that service! (Note: Rstudio Cloud does not have access to your data and so you can use it.)
2. Erase all data after your ASA's DataFest participation is complete.
3. Not identify or attempt to identify the information contained in the dataset, nor contact any of the individuals whose information is contained in the dataset.
4. Comply with all applicable U.S. federal and state laws and regulations relating to the maintenance of the dataset, the safeguarding of the confidentiality of the dataset, and the use and disclosure of the dataset.
5. Not publish results of your analysis of the data except that the final products of the competition (video, slide deck, one-page summary) may be displayed on team members' websites and on campus ASA DataFest websites.
6. Not share the data with anyone who is not a participant of ASA DataFest

Finally, please do NOT reveal the source of the data or any features of the data to ANYONE before April 28. This will ensure that all ASA DataFest participants around the globe have the same experience with the data as you. This means to not post anything to social media that might reveal clues, and to make sure that public github repositories are not discoverable (and if they are, that identifying work is removed).

If you are interested in continuing your investigations with the CourseKata data, similar datasets are available by contacting CourseKata and signing a non-disclosure. Please see https://coursekata.org/contact-us

Important Links

Variable definitions for each file.
https://research.coursekata.org/docs/data-download-processing/downloaded-files.html

Preview Book
https://coursekata.org/preview/default/program

Interactive Demo Book
Similar to the preview book, but allows you to interact with it as a student would. This is not necessary, but perhaps might help you understand the experience better. To do this, you'll need to create an account:

1. Sign up with a new account at canvas.instructure.com (Join Code: J88PRD).
2. Select "I am a student"
3. Then go to canvas.instructure.com/enroll/J88PRD to connect to the class.

Challenge

How did you learn Statistics? CourseKata is a platform that creates and publishes a series of e-books for introductory statistics and data science classes that utilize demonstrated learning strategies to help students learn statistics and data science. The developers of CourseKata, Jim Stigler (UCLA) and Ji Son (Cal State Los Angeles) and their team, are cognitive psychologists interested in improving statistics learning by examining students' interactions with online interactive textbooks. Traditionally, much of the research in how students learn is done in a 1-hour lab or through small-scale interviews with students. CourseKata offers the opportunity to peek into the actions, responses, and choices of thousands of students as they are engaged in learning the interrelated concepts and skills of statistics and coding in R over many weeks or months in real classes.

Your challenge is to examine the data and make suggestions to help CourseKata improve the student experience of learning statistics. Your suggestions might be directed to the CourseKata team, and include observations about patterns of engagement, stumbling points, clear successes, missing features, etc. Or they might be directed to students or their classroom instructors and suggest learning strategies that the data signal are underutilized (or are particularly successful).

How to Get Started

First, "flip" through a CourseKata preview book. This preview version of the book does not allow you to interact with it, but shows you the questions asked, the text, and the media.

Note that the titles in the preview page slightly differ from those in your data set. However, each title has a series of letters (for example, ABC or ABCD) and these can be used to match the titles in the data to the titles in the preview.

These books will be a useful reference throughout your weekend, and will help provide context for the data you see.

The students who used CourseKata textbooks accessed the interactive versions of the books through their campus Learning Management System (e.g., Canvas, Blackboard, Moodle, D2L). If you would like to try the interactive version yourself, you can login to this Canvas course that we set up for you (note, even if your campus uses Canvas, this is an unrelated Canvas instance so you'll have to follow the instructions above to access the book. Actually interacting with the textbook may be helpful for you to generate hypotheses about the data. (You should have received instructions for accessing this book before DataFest began.)

Second, read the descriptions below of the four data files and get acquainted with the two codebooks.

Third, we recommend you begin by forming questions that interest you about a *single specific file*. (Maybe do this for each of the four primary files, and then choose the questions your group most likes.) Then analyze that single file to answer your questions, but pull in data from the other files as your curiosity grows. You might abandon some or all of your questions, or change them radically, but starting with a single data file will likely be more productive than trying to find a path of inquiry that unites several files from the start.

Fourth, we STRONGLY advise you to begin with the random sample of data we provided you. This sample was selected by taking a random sample of students and then including all records for those students. These data files are much smaller and so you should be better able to explore and experiment. When you're happy with what you have, you can try it out on the bigger data sets. If those data sets are too large for you to use, (they crash your computer or take too long, etc) then you may use this sample, BUT YOU MUST ACKNOWLEDGE THIS IN YOUR PRESENTATION.

The Data

Where are the data from?

The data were collected in 48 college-level introductory statistics and data science classes taught in 11 different institutions in 2023. Each of these classes used one of the CourseKata textbooks. The data are from a total of 1625 students who allowed for their anonymized data to be used towards improving the textbook. All classes are college classes, although some used the high school version of the text.

Each class used one of three books: Advanced Statistics and Data Science (ABCD) (a college level book), Statistics and Data Science (ABC) (a college level book), and a high-school version of Advanced Statistics and Data Science I (ABC) (a high-school level book). Note that the college and high school textbooks both labeled ABC are identical except that some survey items within the textbook are designed for that population (e.g., high school students are not asked about their major while college students are). We should note that one class used two books!

The labels ABC or ABCD refer to the content included in the books. Section A refers to the first 4 chapters of the book that include an introduction to R and making data visualizations. Section B refers to the next 5 chapters which focus on modeling variation with simple models (goes up through models that include one predictor variable, e.g., regression and grouping models). Section C refers to 3 chapters that cover inferential statistics with a modeling approach (hypothesis testing and confidence intervals with and without simulations). Section D refers to 4 chapters that cover multivariate models (factorial ANOVA, ANCOVA, and multiple regression).

Because the CourseKata textbooks are continuously improved based on student data, they have version numbers (e.g., 5.0, 5.1, 5.2). These version numbers indicate small changes that were made (e.g., new figures, updated code exercises) to improve the student experience. Higher numbers indicate later versions.

book	release
1 College / Advanced Statistics and Data Science (ABCD)	v5.0-exp2
2 College / Advanced Statistics and Data Science (ABCD)	v5.0

3 College / Advanced Statistics and Data Science (ABCD)	v5.1.1
4 College / Statistics and Data Science (ABC)	v5.0-exp1
5 College / Statistics and Data Science (ABC)	v5.0
6 College / Statistics and Data Science (ABC)	v5.1.1
7 College / Statistics and Data Science (ABC)	v5.2
8 High School / Advanced Statistics and Data Science I (ABC)	v5.0

The basic unit of a book is the page, and pages are clustered within chapters (just as in traditional books). Pages and chapters are named and numbered, for example, Chapter 1 has the title "Chapter 1 – Welcome to Statistics: A Modeling Approach", and page 1.04 has the title "1.4 Introduction to R Functions."

The content of a page may contain text, videos, items that require response, items with optional responses. Sometimes, answers are provided in the text that follows a question; other times students are revealed the answer upon answering. Some questions are open-ended and have no correct answer ("What do you feel about.....") and other times they do ("Which of these variables are categorical?")

How were the data generated?

The data given to you are generated by student actions.

Each time a student accesses a page, basic information about that interaction is recorded in the file "page_views.csv", which also contains information to identify the page. (The files are described in more detail below.)

Some pages have items that ask for student input (e.g., multiple choice questions, coding exercises, written response questions). When a student encounters one of these and interacts with it, the interaction is recorded in the "responses.csv" file. This file provides information about the prompt and the student's response and other information as relevent. Additional meta-data about the item is contained in the items.csv file. Note that an "item" may consist of multiple questions.

If the student encounters a video, then information about that encounter is included in the "media_views.csv" file. This file contains data that attempts to measure the student's engagement with the video (how long spent watching, how many times, etc). Note that there aren't many videos in these versions of the textbook.

Finally, at the start of most chapters, students take a Pulse Check survey. This consists of four quick questions that are intended to provide the instructor with a sense of the students' state of mind as they begin the chapter. These responses are provided in "checkpoints.csv". The same file contains summary results of end-of-chapter review questions, which might serve as an assessment of study learning for that chapter.

The Data Files

1. "page_views.csv" (478752 rows X 19 cols)

Each row represents a student's access to a particular page within a chapter. You can learn the date and time a page was accessed (dt_accessed in UTC), some information about the page itself, and some information about the student's level of engagement with the page.

In addition to information about how and when the student traversed the book, you can get information about the content of the chapter and the page. You can learn more about this by looking those pages up in the preview book or the interactive version of the book on Canvas.

Interpreting this file

This file has some particularities and features that you should be aware of when interpreting the data:

review_flag: this variable indicates whether the page includes end-of-chapter review questions. Students might treat these pages differently from others. Also, the questions that appear on these pages provide an assessment that can be used to measure the students' understanding. The percent correct for these review questions is included in the "checkpoints.csv" file.

was_complete: if set to TRUE, this indicates the page was *already* completed when the user accessed it. A "complete" page is one in which all of the items/questions have been answered.

tried_again_clicks and *tried_again_dt*: these help indicate how many times a student reset their answers on a particular page and the date at which the most recent reset occurred. The values for these two variables are reported at the page level and NOT the row level.

> When a user completes all of the items on a page, a "try again" button appears, which, if clicked, resets all of their answers on that page and lets them try again.

> So if, for a particular row, the tried_again_clicks is >0, this indicates that at one point in time (but not necessarily the current time as given by the access date), the student "reset" their answers. The number of "resets" is given by tried_again_clicks, and the time of the *most recent* reset is given in tried_again_dt. Note that a student must answer all questions before they can reset.

Exams: A very few pages have the word "Midterm". However, we do not have the grades for these items, and so we don't recommend you use them as measures of student learning. However, you can view the prompt and the student responses.

2. responses.csv (1585274 X 40)

Each row provides a student's response and supplementary data about one particular question (but see next paragraph for exceptions). Questions are grouped into "items". The item is named within *item_id* and is one of three types as given by *item_type*: learnosity, learnosity-activity, or code. Code questions are uniquely identified by the *item_id* variable, since a code item has only one question. Learnosity and learnosity-activity questions are uniquely identified by the *lrn_question_reference*. Thus, you should see that several different *lrn_question_reference* values are grouped within a single *item_id* value when *item_type* is learnosity or learnosity_activity.

Additional information provided includes the question prompt , the student's response to the prompt (not available for all question types), and, when appropriate, the student's score (and the maximum possible points). More information about the question itself can be found in items.csv.

completes_page: This flag indicates if, by answering the question, the student answered all questions on the page. Thus, this should be FALSE for all but one question on a page and so might provide insight into whether a question was answered out of order.

3. items.csv (1335 X 19)

Each row contains information about a particular question (although it does not provide the prompt). The item to which a question belongs is included. All items/questions are represented. Use this file to go deeper into particular questions that students encounter in the course.

Questions are grouped into items (*item_id*). An item can be one of three *item_type* 's: code, learnosity or learnosity-activity (the distinction between learnosity and learnosity-activity is not important). Code items are a single question and ask for R code as a response. (Responses can be seen in responses.csv.) Learnosity-activities and learnosity items are collections of one or more questions that can be of a variety of lrn_type's:

- association
- choicematrix
- clozeassociation
- formulaV2
- imageclozeassociation
- mcq
- plaintext
- shorttext
- sortlist

Examples of these question types are provided at the end of this document.

The level of detail made available to you in the responses file depends on the *lrn_type*. For example, for multiple choice questions (mcq), you can find the options in the responses file in the columns labeled *lrn_option_0* through *lrn_option_11*, and you can see the chosen option in the results variable.

Assessment Types

In general, assessments, such as the items and questions included in CourseKata, can be used for two purposes. Formative assessments are meant to provide feedback to the student (and instructor), or to serve as a learning aid to help prompt students improve memory and deepen their understanding. Summative assessments are meant to provide a summary of a student's understanding, often for use in assigning a grade. For example, most midterms and final exams that you've taken are summative assessments.

The vast majority of items in CourseKata should be treated as formative assessments. The exceptions are the end-of-chapter Review questions, which can be thought of as summative. The mean number of correct answers for end-of-chapter review questions is provided within the checkpoints file. You might see that some pages have the word "Quiz" or "Exam" or "Midterm" in them. Results from these items and responses to them are not provided to us in this data set.

4. media_views (6149x18)

Each row contains data about a student's interaction with a video.

5. checkpoints (79092 X 12)

The data set is not perfectly "tidy". Each row represents a student's response to either a single pulse item or a summary of end-of-chapter review questions for the relevant chapter. (These summaries are repeated on each relevant row.)

This file contains responses for the "pulse" questions asked at the start of each chapter (EXCEPT for Chapter 1). Each is scored on a scale of 0 (strongly disagree) to 5 (strongly agree). These questions are

1) I am confident about what I learned in the previous chapter. The "expectancy" construct.
2) I think what I have learned in the previous chapter is useful (0 to 5). "General utility value" construct.
3) I think this class is interesting (0 to 5). "Intrinsic Value"
4) I was unable to put in the time needed to do well in the previous chapter (0 to 5) "Course cost" construct.

The constructs refer to psychological frames of mind that students might be in that might explain their engagement in the course.

Pulse questions are not referred to in any of the other data files.

Each chapter also contains end-of-chapter review questions. The percent correct, the total number of questions, the number of correct responses, and the number of attempts at the review are provided on each relevant row.

For example, in Chapter 2 there are 4 pulse questions, and so for each student you should see four rows: one row for each pulse question. On each of these rows, you will see the same summary information for the chapter review.

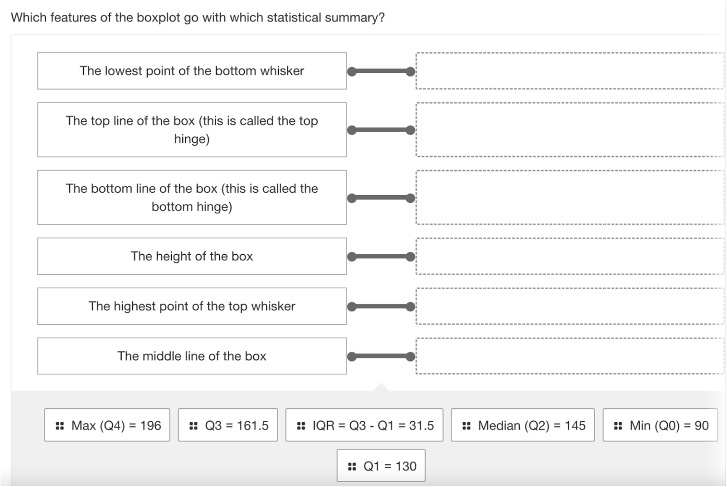
Meta Data

You're provided with two data-dictionaries. One, codebook.csv, is organized by the data file. For each data file, it provides the variables and definitions of the variables. Note that some variables appear in multiple files.

The other, variable_list.csv, is organized by variable name. For each variable, it indicates which data files contain that variable. This is useful for help with merging and comparing data files.

Example `lrn_type`

"association"

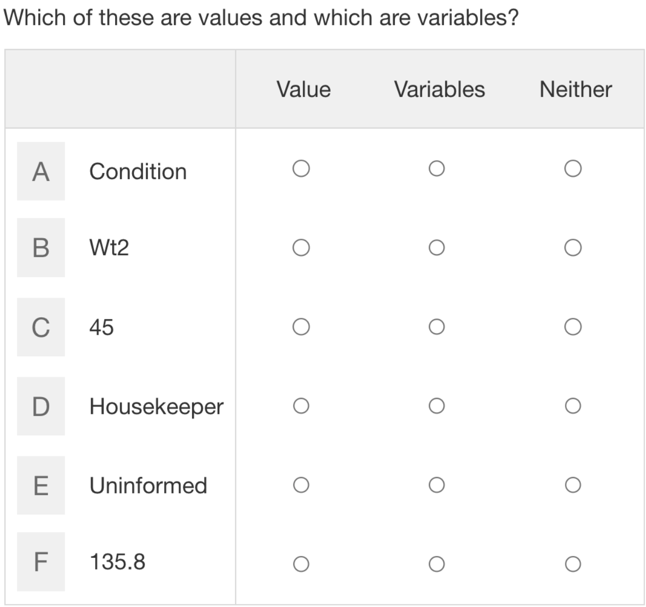


Which features of the boxplot go with which statistical summary?

- The lowest point of the bottom whisker
- The top line of the box (this is called the top hinge)
- The bottom line of the box (this is called the bottom hinge)
- The height of the box
- The highest point of the top whisker
- The middle line of the box

Max (Q4) = 196 | Q3 = 161.5 | IQR = Q3 - Q1 = 31.5 | Median (Q2) = 145 | Min (Q0) = 90 | Q1 = 130

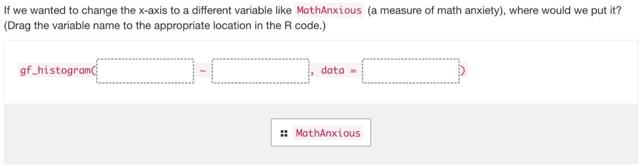
"choicematrix"



Which of these are values and which are variables?

		Value	Variables	Neither
A	Condition	○	○	○
B	Wt2	○	○	○
C	45	○	○	○
D	Housekeeper	○	○	○
E	Uninformed	○	○	○
F	135.8	○	○	○

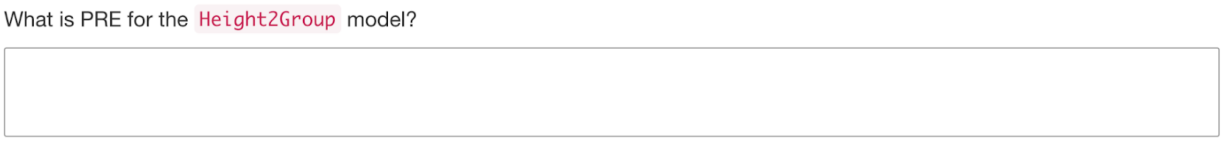
"clozeassociation"



If we wanted to change the x-axis to a different variable like `MathAnxious` (a measure of math anxiety), where would we put it? (Drag the variable name to the appropriate location in the R code.)

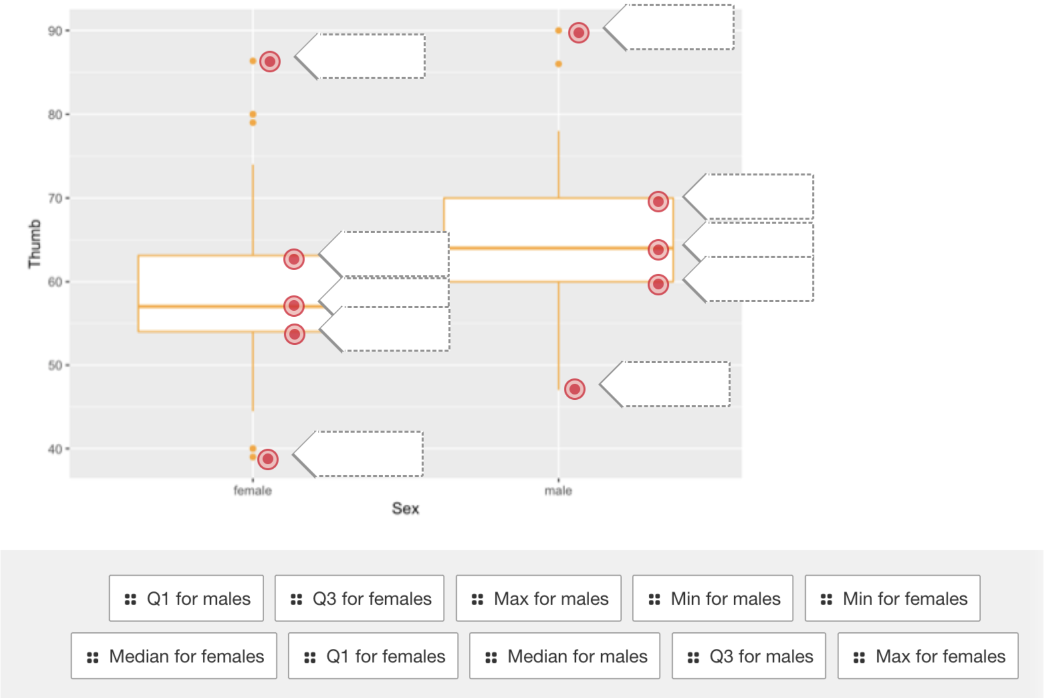
```
gf_histogram(          ~          , data =          )
```

"formulaV2" (a numerical input required)

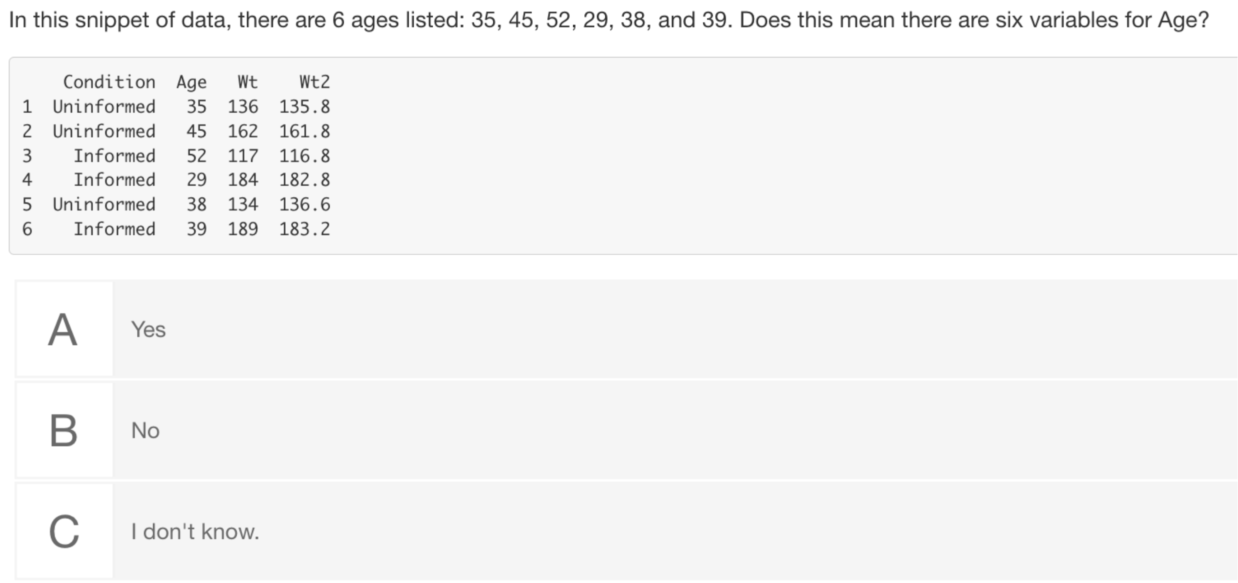


What is PRE for the `Height2Group` model?

"imageclozeassociation"



"mcq"   some mcq items have only 1 correct answer; others are "check all that apply"



In this snippet of data, there are 6 ages listed: 35, 45, 52, 29, 38, and 39. Does this mean there are six variables for Age?

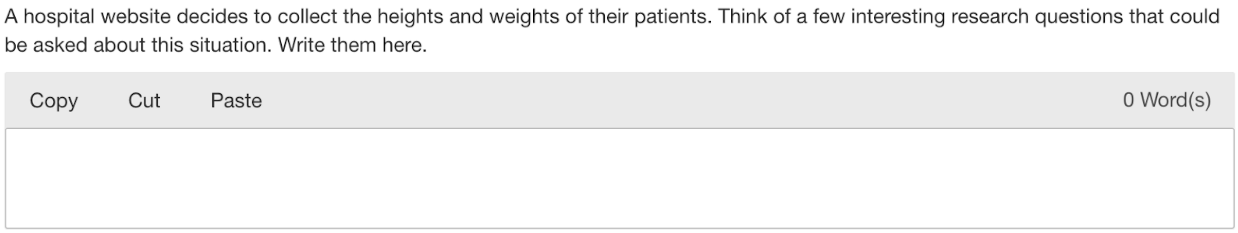
```
   Condition  Age   Wt    Wt2
1 Uninformed   35  136  135.8
2 Uninformed   45  162  161.8
3   Informed   52  117  116.8
4   Informed   29  184  182.8
5 Uninformed   38  134  136.6
6   Informed   39  189  183.2
```

A Yes

B No

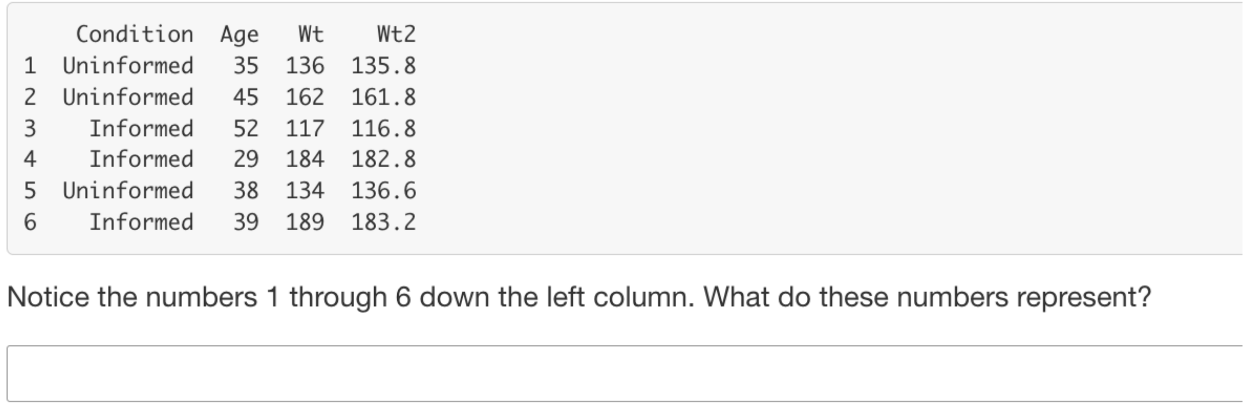
C I don't know.

"plaintext"



A hospital website decides to collect the heights and weights of their patients. Think of a few interesting research questions that could be asked about this situation. Write them here.

"shorttext"



```
   Condition  Age   Wt    Wt2
1 Uninformed   35  136  135.8
2 Uninformed   45  162  161.8
3   Informed   52  117  116.8
4   Informed   29  184  182.8
5 Uninformed   38  134  136.6
6   Informed   39  189  183.2
```

Notice the numbers 1 through 6 down the left column. What do these numbers represent?

"sortlist"

$$\sum_{i=1}^{n} |Y_i - \bar{Y}|$$

The expression above suggests a series of mathematical operations. Place them in their proper order in the target column.

Source		Target	
≡ Take the absolute value	◀ ▶		▲ ▼
≡ For each data point, subtract the mean			
≡ Take the sum across all rows in the data frame			