While waiting for [*Star Wars: The Force Awakens*](https://en.wikipedia.org/wiki/Star_Wars:_The_Force_Awakens) to come out, the team at [FiveThirtyEight](http://fivethirtyeight.com/) became interested in answering some questions about *Star Wars* fans. In particular, they wondered: **does the rest of America realize that “The Empire Strikes Back” is clearly the best of the bunch?**

The team needed to collect data addressing this question. To do this, they surveyed *Star Wars* fans using the online tool SurveyMonkey. They received 835 total responses, which you download from [their GitHub repository](https://github.com/fivethirtyeight/data/tree/master/star-wars-survey).

For this project, you'll be cleaning and exploring the data set in Jupyter notebook. To see a sample notebook containing all of the answers, visit [the project's GitHub repository](https://github.com/dataquestio/solutions/blob/master/Mission201Solution.ipynb).

The following code will read the data into a pandas dataframe:



import pandas as pd

star\_wars = pd.read\_csv("star\_wars.csv", encoding="ISO-8859-1")

We need to specify an encoding because the data set has some characters that aren't in Python's default utf-8 encoding. You can read more about character encodings [on developer Joel Spolsky's blog](http://www.joelonsoftware.com/articles/Unicode.html).

The data has several columns, including:

* RespondentID - An anonymized ID for the respondent (person taking the survey)
* Gender - The respondent's gender
* Age - The respondent's age
* Household Income - The respondent's income
* Education - The respondent's education level
* Location (Census Region) - The respondent's location
* Have you seen any of the 6 films in the Star Wars franchise? - Has a Yes or No response
* Do you consider yourself to be a fan of the Star Wars film franchise? - Has a Yes or Noresponse

There are several other columns containing answers to questions about the *Star Wars* movies. For some questions, the respondent had to check one or more boxes. This type of data is difficult to represent in columnar format. As a result, this data set needs a lot of cleaning.

First, you'll need to remove the invalid rows. For example, RespondentID is supposed to be a unique ID for each respondent, but it's blank in some rows. You'll need to remove any rows with an invalid RespondentID.

Instructions

* Read the data set into a dataframe.
* Explore the data by entering star\_wars.head(10). Look for any strange values in the columns and rows.
* Review the column names with star\_wars.columns.
* Remove any rows where RespondentID is NaN. You can use the [pandas.notnull()](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.notnull.html" \t "_blank) function to accomplish this. Only select rows where the RespondentID column is not null.
* When you're finished, star\_wars should only consist of rows where RespondentID is not NaN.