Supera - Data analysis challenge

Using the following instructions and dataset, proceed as far as you can within 60 minutes. **Do not collaborate directly with anyone else during this challenge**. Don't worry if you can't complete the entire challenge**.** If you have any questions during the challenge, please contact us by email.

Part 1: Data setup

You have a dataset called "Bechdel.xlsx" in Excel format. Import it into Python or R and prepare the dataset for analysis.

Learn about what the Bechdel test is
https://en.wikipedia.org/wiki/Bechdel_test

See the data dictionary for details on the variable names.

Part 2: Data analysis

i) Summarise the data using any descriptive statistics, inferential statistics, or visualizations that you deem appropriate. Address the following questions as well:

– What percentage of movies passed the test, based on the binary definition?
– How does the percentage of movies passing change over time?

ii) Propose a model that will predict whether a movie passes or fails (the binary variable) based on whatever other variables in the data are available that you think are useful. There is a dataset called test.csv which contains a new set of movies that has every variable except the binary pass/fail variable. Using your predictive model, classify each of the new movies as pass or fail. (1 = pass, 0 = fail). Simply save your predictions in a csv file with the variable 'mdb' to indicate the movie and your classification of 0 or 1.

Part 3: Version control

Put your code into a new public repo on github or a similar service. Provide the link.