Anna Harvey
DSC 530
08/08/20
Final project

Statistical Question:
"Do physical variables of planets affect the attributes of the dominant species in the Star Wars universe?"

Dataset source:
https://www.kaggle.com/jsphyg/star-wars
Originally sourced from Wookiepedia. www.starwars.fandom.com

Datasets used (included in GitHub repo):
species.csv
planets.csv

The results of my EDA of these datasets do not show enough evidence for any absolute correlations between planetary variables (diameter, orbital period, rotation period, surface water, and climate) and species attributes (average height, classification). This is most likely due to the incomplete nature of the data. Information regarding technical specifications within the Star Wars universe varies in quality. Although the Star Wars fandom is one of the most thorough when it comes to documenting details about the fictional universe, they can only work with data that has been given through official sources. That being the case, the methods of this analysis would most likely work better with a more extensive dataset of planetary and species information.

Including the terrain variables in the analysis could have possibly helped find a more robust conclusion. It also would be beneficial to have additional variables for the species to compare. These could include a more detailed division of classification information, average weight, number of limbs, and primary diet. The dataset really suffered from having planets with 0 km diameter and 0% surface water (for those planets that were not actually 0% surface water planets). I assumed that surface water percentage would affect the classification of species present on the planet. That does not seem to be the case with this data. This could suggest that the dataset needs to be more robust. It could also indicate that having mammals present on planets with a high percentage of surface water is just as likely as amphibians. This makes sense considering the nature of the earth's oceans and the biodiversity present there.

Overall, the dataset was a good practice set to work with. The benefit of exercising new knowledge on a dataset of fictional information is that you do not have to worry about any negative effects arising from the analysis. The curators of the Star Wars universe put an immense amount of effort into making the fictional construct as logical as possible within the realms of science fantasy. However, it is also likely that even with extensive data and details available about planets and species, that no natural correlations may emerge. Since writers and creators have a certain amount of freedom in the creation of new worlds and species, it is likely that getting to the detail level of planet diameter affecting species height would be the least of anyone's concerns.