starwars
Exploratory Analysis

Ben Gierhart, wgierhartt@bellarmine.edu

I. INTRODUCTION

This is a dataset of characters from the Star Wars universe and their characteristics. It is one of the built-in datasets from R. I chose it because I thought it would be a fun way to learn and complete this project.

II. DATA SET DESCRIPTION

This data set contains 87 samples with 14 columns with various data types before cleaning and 68 samples with 14 columns after cleaning. A complete listing is shown in **Table 1**. **Note: Missing Data (%) only applies to data before cleaning. After cleaning, Missing Data (%) is 0% for all variables. For the 'vehicles' and 'starships' variables, the value 'character(0)' is not an empty value. It indicates that that character is not shown to have a vehicle or starship of their own. Similarly, some characters have no hair, so their 'hair_color' is 'none.'**

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
name	chr	0%
height	int	6.90%
mass	num	32.18%
hair_color	chr	5.75%
skin_color	chr	0%
eye_color	chr	0%
birth_year	num	50.57%
sex	chr	4.60%
gender	chr	4.60%
homeworld	chr	11.49%
species	chr	4.60%
films	chr	0%
vehicles	chr	0%
starships	chr	0%

III. Data Set Summary Statistics

The information below explores the numerical and categorical data in the dataset after cleaning.

Table 2: Summary Statistics for starwars

Variable Name	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max*
height	79.0	170.0	183.0	179.1	191.5	264.0
mass	15.00	75.00	79.00	77.37	80.00	159.00
birth_year	8.00	48.00	52.00	53.49	52.00	200.00

The 'name' variable is not present in this table as each value is unique.

Table 3: Proportions for XXX (n=yyy)

Category	*Frequency*	*Proportion (%)*
hair_color	Auburn (1), auburn/grey (1), auburn/white (1), black (11), blond (3), blonde (1), brown (12), brown/grey (1), grey (1), none (33), white (3)	Auburn (1.47%), auburn/grey (1.47%), auburn/white (1.47%), black (16.18%), blond (4.41%), blonde (1.47%), brown (17.65%), brown/grey (1.47%), grey (1.47%), none (48.53%), white (4.41%)

Category	*Frequency*	*Proportion (%)*
skin_color	Blue (2), blue/grey (2), brown (4), brown mottle (1), brown/white (1), dark (4), fair (14), fair/green/yellow (1), green (4), green/grey (1), grey (6), grey/blue (1), grey/green/yellow (1), grey/red (1), light (8), mottled green (1), orange (2), pale (4), red (1), red/blue/white (1), tan (2), unknown (1), white (2), white/blue (1), yellow (2)	Blue (2.94%), blue/grey (2.94%), brown (5.88%), brown mottle (1.47%), brown/white (1.47%), dark (5.88%), fair (20.59%), fair/green/yellow (1.47%), green (5.88%), green/grey (1.47%), grey (8.82%), grey/blue (1.47%), grey/green/yellow (1.47%), grey/red (1.47%), light (11.76%), mottled green (1.47%), orange (2.94%), pale (5.88%), red (1.47%), red/blue/white (1.47%), tan (2.94%), unknown (1.47%), white (2.94%), white/blue (1.47%), yellow (2.94%)
eye_color	Black (8), blue (17), blue/gray (1), brown (16), gold (1), green/yellow (1), hazel (2), orange (7), pink (1), red (2), unknown (2), yellow (10)	Black (11.76%), blue (25%), blue/gray (1.47%), brown (23.53%), gold (1.47%), green/yellow (1.47%), hazel (2.94%), orange (10.29%), pink (1.47%), red (2.94%), unknown (2.94%), yellow (14.71%)
sex	Female (14), male (54)	Female (20.59%), male (79.41%)
gender	Female (14), male (54)	Female (14), male (54)
homeworld	Alderaan (3), Aleen Minor (1), Bespin (1), Cato Neimoidia (1), Cerea (1), Champala (1), Chandrila (1), Concord Dawn (1), Corellia (2), Coruscant (3), Dathomir (1), Dorin (1), Endor (1), Eriadu (1), Geonosis (1), Glee Anselm (1), Haruun Kal (1), Iktotch (1), Iridonia (1), Kalee (1), Kamino (3), Kashyyyk (2), Malastare (1), Mirial (2), Mon Cala (1), Muunilinst (1), Naboo (8), Ojom (1), Quermia (1), Ryloth (2), Serenno (1), Shili (1), Shako (1) Socorro (1), Stewjon (1), Sullust (1), Tatooine (8), Toydaria (1), Trandosha (1), Troiken (1), Tund (1), Utapau (1), Vulpter (1), Zolan (1)	Alderaan (4.41%), Aleen Minor (1.47%), Bespin (1.47%), Cato Neimoidia (1.47%), Cerea (1.47%), Champala (1.47%), Chandrila (1.47%), Concord Dawn (1.47%), Corellia (2.94%), Coruscant (4.41%), Dathomir (1.47%), Dorin (1.47%), Endor (1.47%), Eriadu (1.47%), Geonosis (1.47%), Glee Anselm (1.47%), Haruun Kal (1.47%), Iktotch (1.47%), Iridonia (1.47%), Kalee (1.47%), Kamino (4.41%), Kashyyyk (2.94%), Malastare (1.47%), Mirial (2.94%), Mon Cala (1.47%), Muunilinst (1.47%), Naboo (11.76), Ojom (1.47%), Quermia (1.47%), Ryloth (2.94%), Serenno (1.47%), Shili (1.47%), Shako (1.47%) Socorro (1.47%), Stewjon (1.47%), Sullust (1.47%), Tatooine (11.76), Toydaria (1.47%), Trandosha (1.47%), Troiken (1.47%), Tund (1.47%), Utapau (1.47%), Vulpter (1.47%), Zolan (1.47%)

species	Aleena (1), Besalisk (1), Cerean (1), Chagrian (1), Clawdite (1), Dug (1), Ewok (1), Geonosian (1), Gungan (3), Human (29), Iktotchi (1), Kaleesh (1), Kaminoan (2), Kel Dor (1), Mirialan (2), Mon Calamari (1), Muun (1), Nautolan (1), Neimodian (1), Pau'an (1), Quermian (1), Skakoan (1), Sullustan (1), Tholothian (1), Togruta (1) Toong (1), Toydarian (1), Twi'lek (2), Vulptereen (1), Wookie (2), Xexto (1), Zabrak (2)	Aleena (1.47%), Besalisk (1.47%), Cerean (1.47%), Chagrian (1.47%), Clawdite (1.47%), Dug (1.47%), Ewok (1.47%), Geonosian (1.47%), Gungan (4.41%), Human (42.65), Iktotchi (1.47%), Kaleesh (1.47%), Kaminoan (2.94%), Kel Dor (1.47%), Mirialan (2.94%), Mon Calamari (1.47%), Muun (1.47%), Nautolan (1.47%), Neimodian (1.47%), Pau'an (1.47%), Quermian (1.47%), Skakoan (1.47%), Sullustan (1.47%), Tholothian (1.47%), Togruta (1.47%) Toong (1.47%), Toydarian (1.47%), Twi'lek (2.94%), Vulptereen (1.47%), Wookie (2.94%), Xexto (1.47%), Zabrak (2.94%)

films	A New Hope (12), Attack of the Clones (33), Return of the Jedi (15), Revenge of the Sith (29), The Empire Strikes Back (12), The Force Awakens (5), The Phantom Menace (29)	A New Hope (8.89%), Attack of the Clones (24.44%), Return of the Jedi (11.11%), Revenge of the Sith (21.48%), The Empire Strikes Back (8.89%), The Force Awakens (3.70%), The Phantom Menace (21.48%)

vehicles	AT-ST (1), Flitknot speeder (1), Imperial Speeder Bike (2), Koro-2 Exodrive airspeeder (1), Sith speeder (1), Snowspeeder (2), Tribubble bongo (1), Tsmeu-6 personal wheel bike (1), XJ-6 airspeeder (1), Zephyr-G swoop bike (1)	AT-ST (8.33%), Flitknot speeder (8.33%), Imperial Speeder Bike (16.67%), Koro-2 Exodrive airspeeder (8.33%), Sith speeder (8.33%), Snowspeeder (16.67%), Tribubble bongo (8.33%), Tsmeu-6 personal wheel bike (8.33%), XJ-6 airspeeder (8.33%), Zephyr-G swoop bike (8.33%)

starships	Belbullab-22 starfighter (2), H-type Nubian yacht (1), Imperial shuttle (3), Jedi Interceptor (2), Jedi starfighter (2), Millennium Falcon (4), Naboo fighter (2), Naboo Royal Starship (1), Naboo star skiff (1), Scimitar (1), Slave 1 (1), TIE Advanced x1 (1), Trade Federation cruiser (2), X-wing (3)	Belbullab-22 starfighter (7.41%), H-type Nubian yacht (3.70%), Imperial shuttle (11.11%), Jedi Interceptor (7.41%), Jedi starfighter (7.41%), Millennium Falcon (7.41%), Naboo fighter (7.41%), Naboo Royal Starship (3.70%), Naboo star skiff (3.70%), Scimitar (3.70%), Slave 1 (3.70%), TIE Advanced x1 (3.70%), Trade Federation cruiser (7.41%), X-wing (11.11%)

Table 4: Correlation Table/Tables

	height	mass	birth_year
height	1.0000000	0.6222929	0.2661701
mass	0.6222929	1.0000000	0.2308662
birth_year	0.2661701	0.2308662	1.0000000

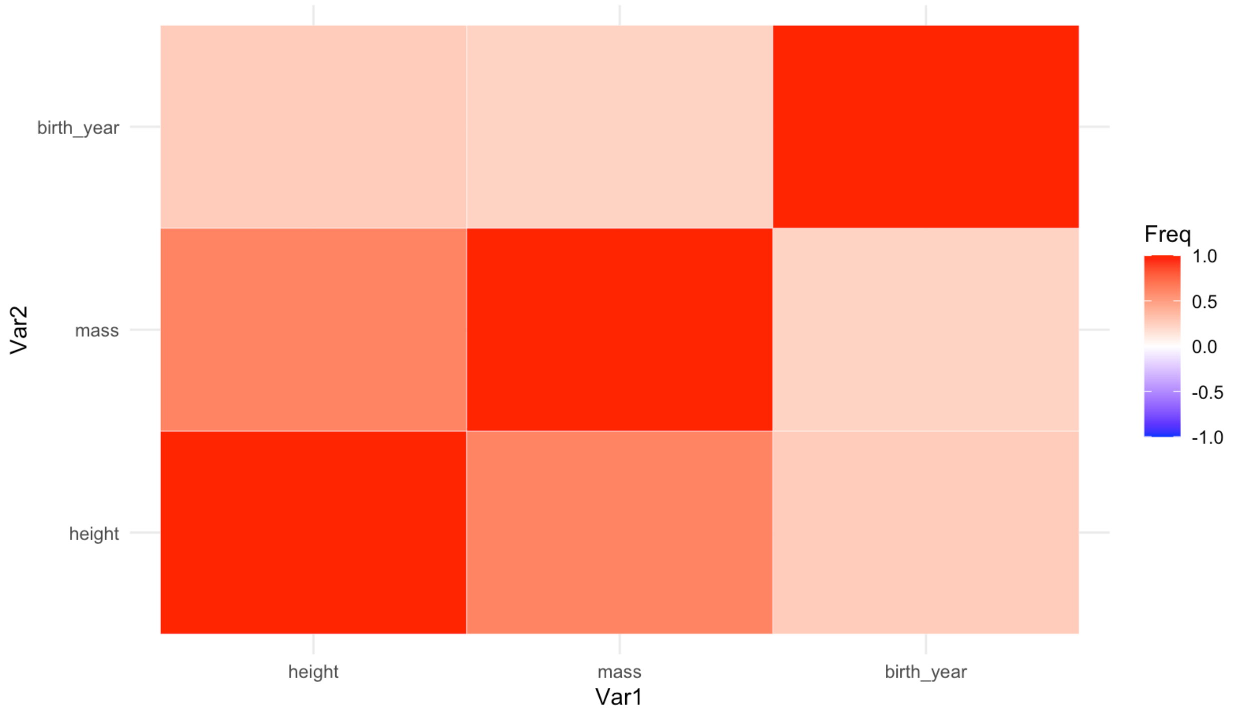

Figure 1: Heat Map of height, mass, and birth_year

From the table and heat map, we can see that, as expected, there is a strong positive correlation between height and mass.

IV. DATA SET GRAPHICAL EXPLORATION

This section will illustrate some of the distributions of and relationships between traits of Star Wars characters.

A. Distributions

Below are histograms to measure distribution of height and mass of Star Wars characters. Both histograms show that the distribution is approximately normal; however, many of the datapoints lie at the mean/median.

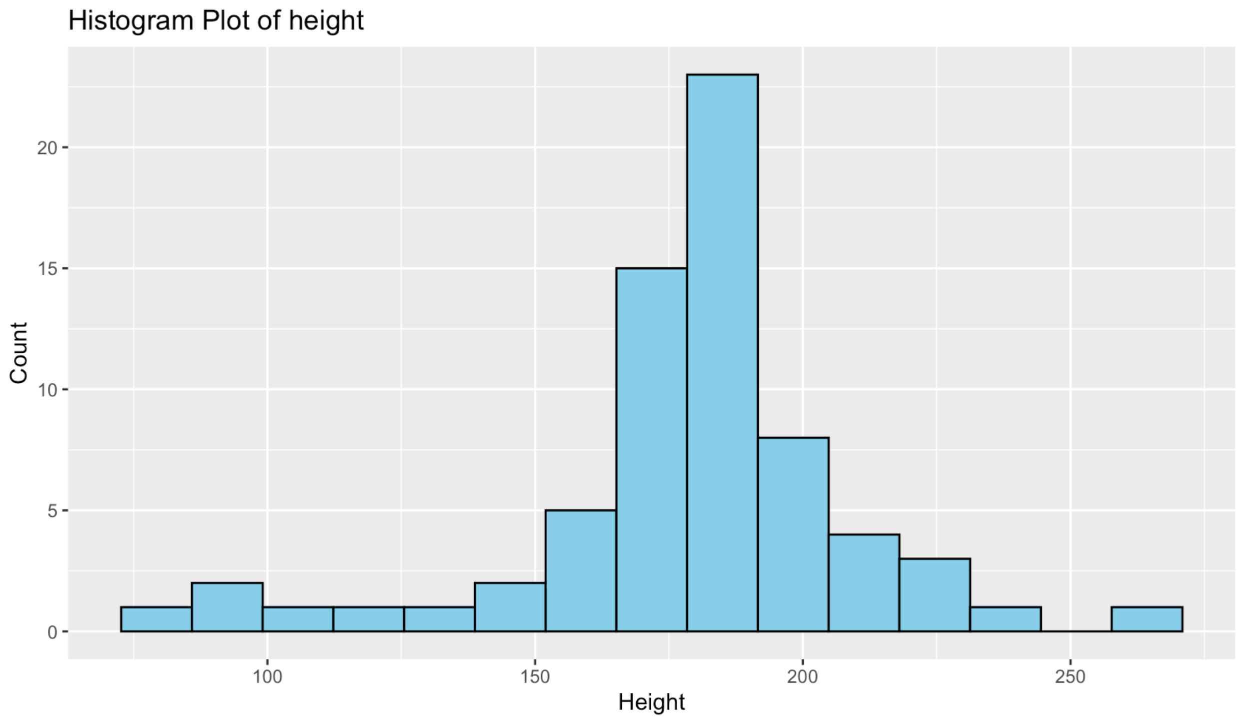

Figure 2: Histogram Plot of height

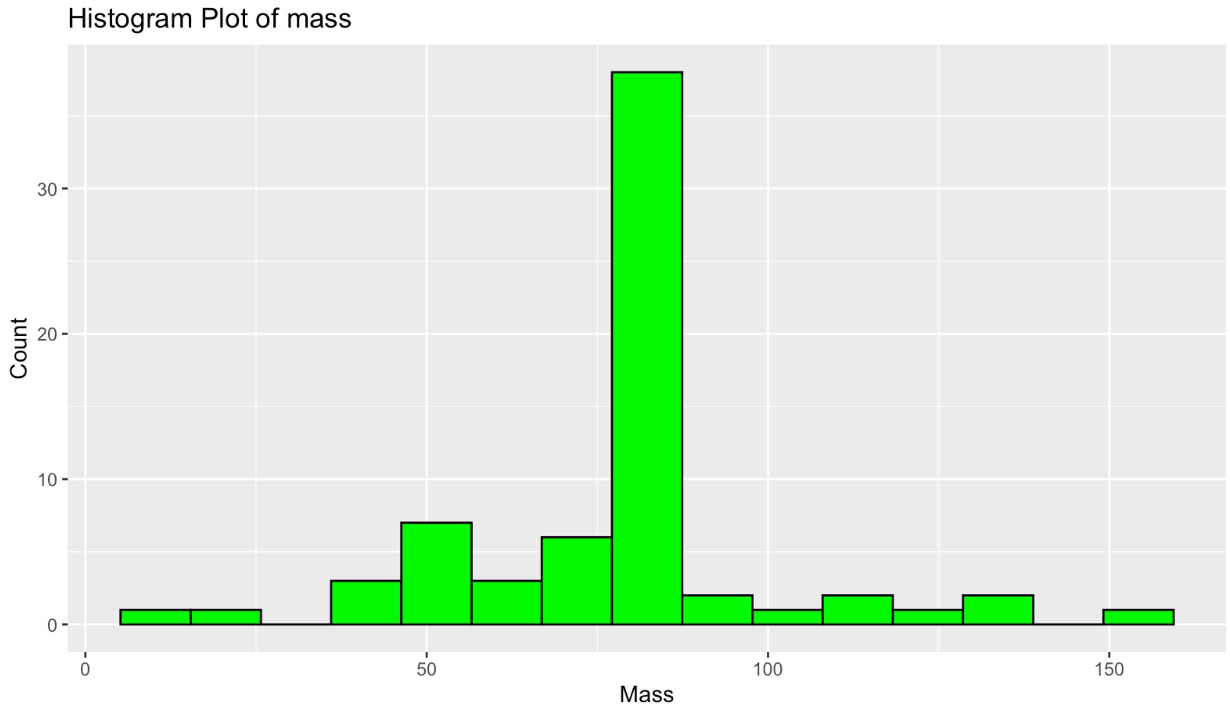

Figure 3: Histogram Plot of mass

B. Scatter Plot

Below is a scatter plot of height vs. mass of Star Wars characters. Many characters are clustered in the means of height and mass, which makes sense given the distribution shown earlier. There is still a slight positive linear relationship present, indicating that, generally, taller characters have more mass.

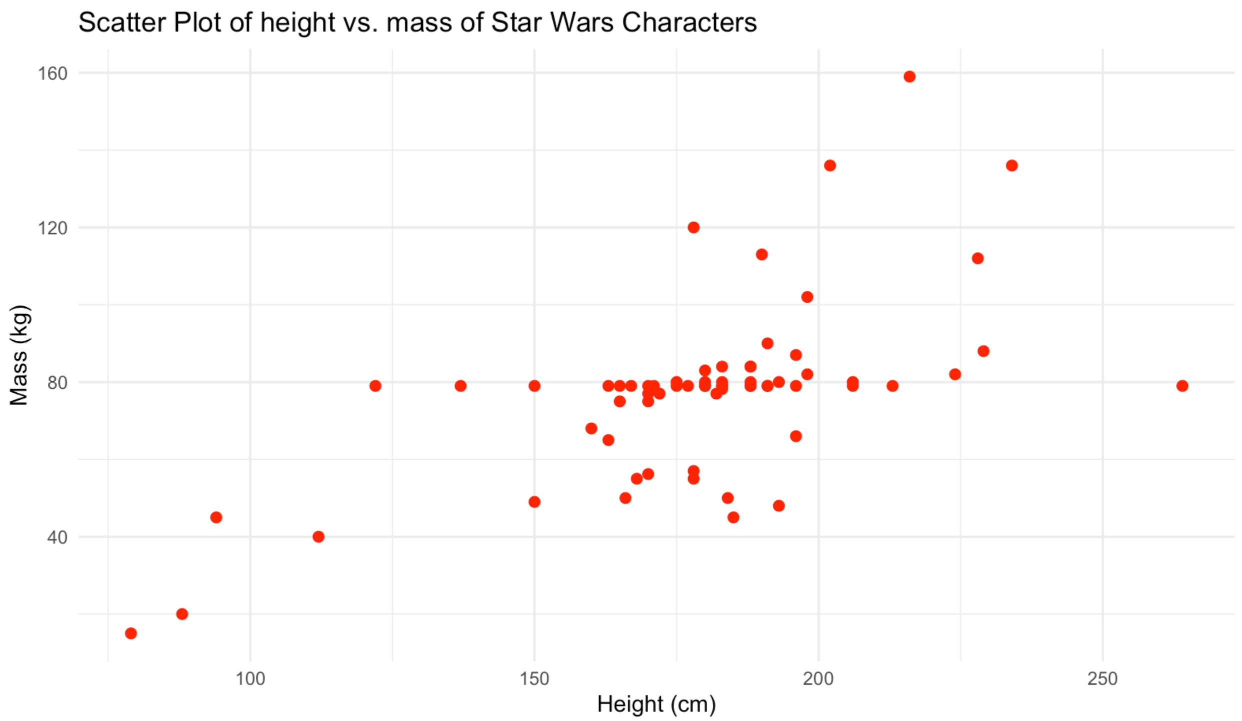

Figure 4: Scatter Plot of height vs. mass

C. Bar Charts of Categorical Variables

We can see from this bar chart that there are nearly four times as many male characters than female characters in the dataset.

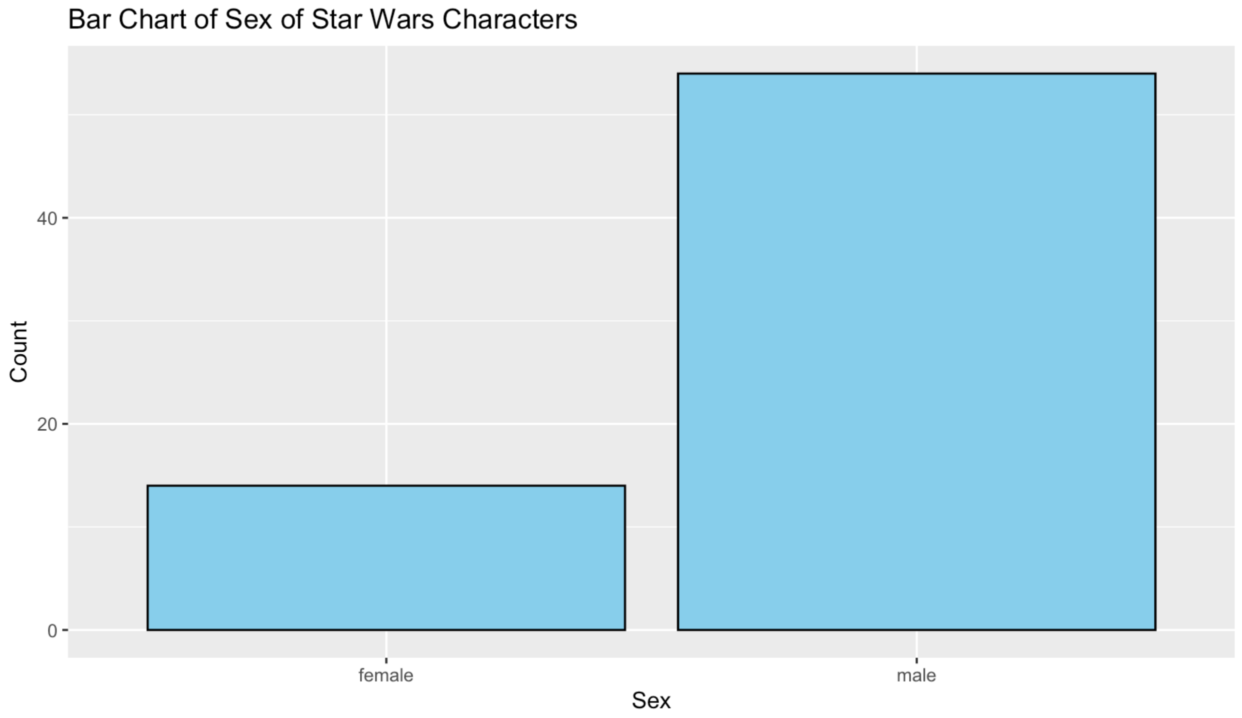

Figure 5: Bar Chart of sex of Star Wars characters

This bar chart shows the number of characters representative of different species in the dataset. By far the most represented species. In this franchise, many characters are tokens of their race.

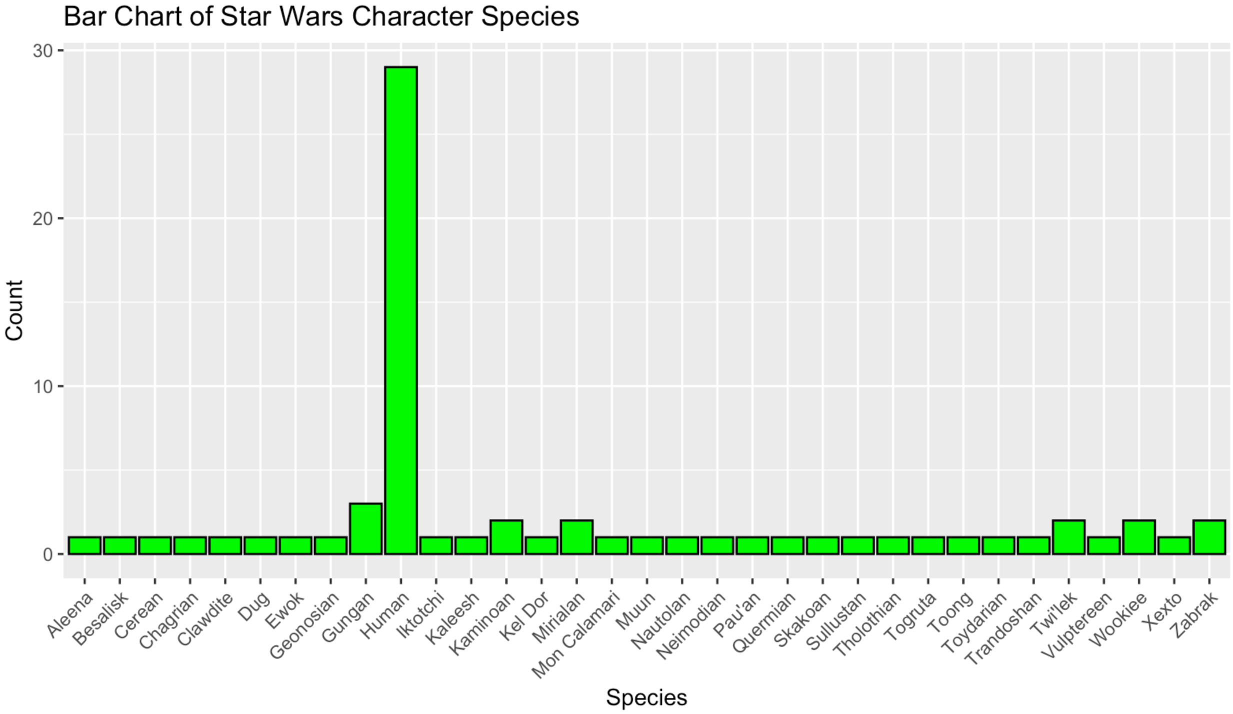

Figure 6: Bar Chart of Star Wars Characters Species

This bar chart shows the number of character appearances for most films in the series. The film with the most character appearances is *Attack of the Clones*. *Revenge of the Sith* and *The Phantom Menace* are closed behind. The prequel trilogy features many more character appearances than the other films in the series. *The Force Awakens*, the only sequel trilogy film represented in this dataset, has by far the least.

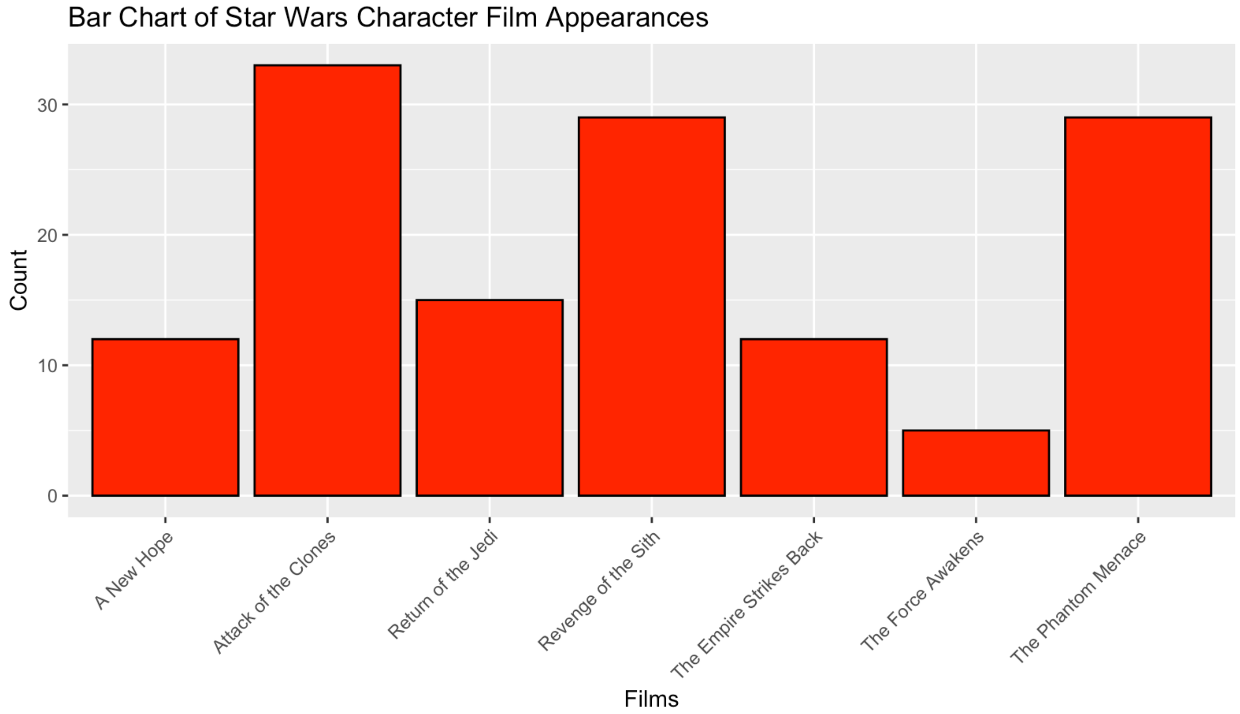

Figure 7: Bar Chart of film appearances of Star Wars characters

D. Other Chart (Box Plot)

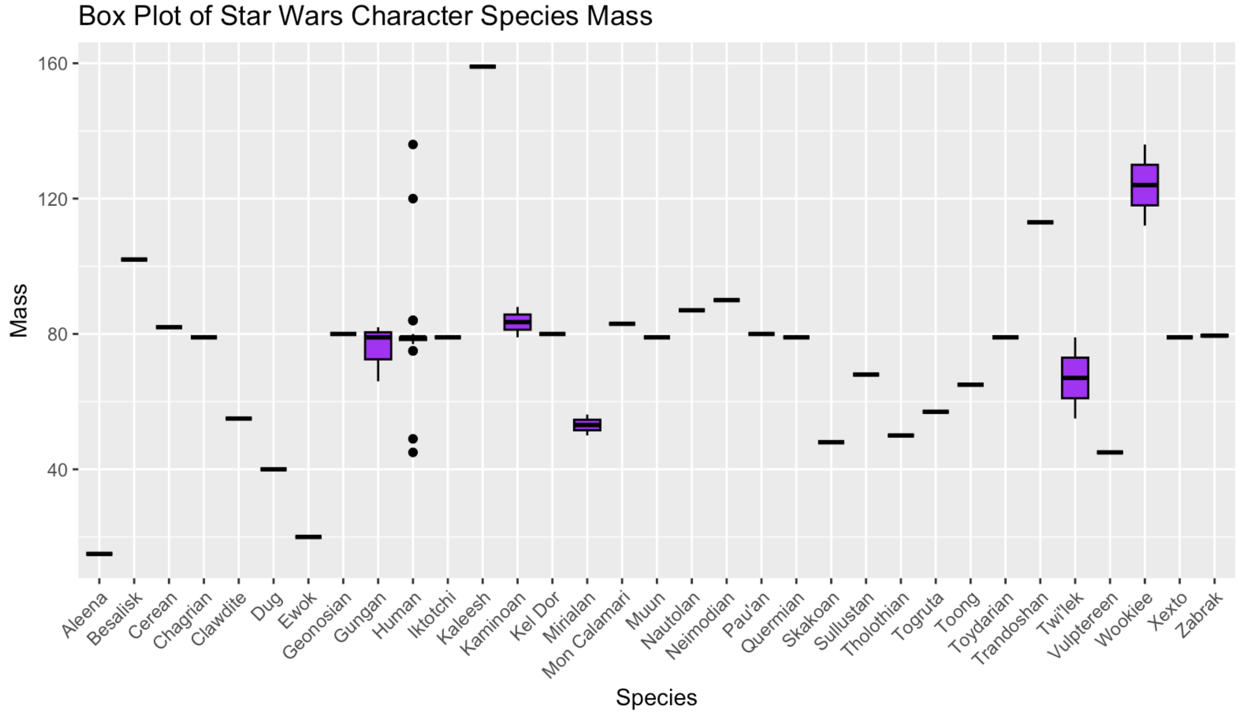

Figure 8: Box Plot of species for mass

This box plot shows the distribution of mass for each species in the dataset. We have the most data on humans, and we can see that most of them are at the mean. That species also has the most outliers. Wookies are the species with the most mass, and they and Twi'leks have means at the center. Gunguns have a mean higher in their distribution, meaning that two of the three gungun characters in the dataset are heavier than the third and influencing the mean.

V. SUMMARY OF FINDINGS

The Star Wars franchise is considered one that is massive and full of imaginative characters. I still believe this is true, especially for its time. The data shows, however, that the majority of characters are human and characters of other species are usually the only ones given significant roles in the narrative. This makes sense as people probably connect more with human characters in a sci-fi story. The majority of the characters are also male. Even when looking at humans, traits like mass and height are homogenous. Popular opinion is that the prequel trilogy is not as good as the original trilogy from an artistic standpoint. However, in terms of imagination and world-building, the prequel films have the highest distribution of different characters.

I think further exploration is warranted with a larger dataset, perhaps including characters from other Star Wars media, to make sure these claims hold true.