Analysis Of StarWars Characters Dataset Using R Programming: Part1 - Exploratory Data Analysis

Siddharth Prajapati

2023-06-10

Setting up my environment

```
#Load the 'tidyverse' package & 'star wars' data set

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data("starwars")
```

A Quick note on Conflicts:

Conflicts aren't necessarily a bad thing! Because R is an open source language and anyone can create a package, it's common for different packages to use the same name for similar functions. In our conflicts we see that the *filter()* function from the *dplyr* package masks the *filter()* function from the *stats* package. We know this because the package name comes before the double colon and the function name comes after, like this:*package::function()*

Understanding the Dataset:

Here's where I like to get a handle on what I'm working with. I'll use various functions to make sure my data imported correctly, and start to get an understanding of the data structure and data types

```
#To check how many rows and columns available in the data frame
dim(starwars)
```

```
## [1] 87 14
```

```
#To get the glimps about the data frame like variable names & their datatypes
#Also look at the first few rows of each variable.
glimpse(starwars)
```

```
## Rows: 87
## Columns: 14
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Or~
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 2~
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.~
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", N~
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "~
## $ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue",~
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, ~
## $ sex        <chr> "male", "none", "none", "male", "female", "male", "female",~
## $ gender     <chr> "masculine", "masculine", "masculine", "masculine", "femini~
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "T~
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Huma~
## $ films      <list> <"The Empire Strikes Back", "Revenge of the Sith", "Return~
## $ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imp~
## $ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1",~
```

Quick note on the 'List' data type:

Lists are the R objects which contain elements of different types like - numbers, strings, vectors etc. In our dataset, there are three variables which have a data type 'list'. The *films* variable contains the list of star war movies in which the character was appeared, similarly the *vehicles* variable tells us that, what all vehicles was used by the character.

```
#To see the first few rows of data set in  nicely-formatted table.
head(starwars)
```

```
## # A tibble: 6 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue             19   male  mascu~
## 2 C-3PO        167    75 <NA>       gold       yellow          112   none  mascu~
## 3 R2-D2         96    32 <NA>       white, bl~ red              33   none  mascu~
## 4 Darth Va~    202   136 none       white      yellow           41.9 male  mascu~
## 5 Leia Org~    150    49 brown      light      brown            19   fema~ femin~
## 6 Owen Lars    178   120 brown, gr~ light      blue             52   male  mascu~
## # i 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
#To view the complete data set
view(starwars)
```

	name	height	mass	hair_color	skin_color	eye_color	birth_year	sex	gender	homeworld	species	films	vehicles	starships
1	Luke Skywalker	172	77.0	blond	fair	blue	19.0	male	masculine	Tatooine	Human	c("The Empire Strikes Back", "Revenge of the Sith" [...]	c("Snowspeeder", "Imperial Speeder Bike")	c("X-wing", "
2	C-3PO	167	75.0	NA	gold	yellow	112.0	none	masculine	Tatooine	Droid	c("The Empire Strikes Back", "Attack of the Clones [...]	character(0)	character(0)
3	R2-D2	96	32.0	NA	white, blue	red	33.0	none	masculine	Naboo	Droid	c("The Empire Strikes Back", "Attack of the Clones [...]	character(0)	character(0)
4	Darth Vader	202	136.0	none	white	yellow	41.9	male	masculine	Tatooine	Human	c("The Empire Strikes Back", "Revenge of the Sith" [...]	character(0)	TIE Advance
5	Leia Organa	150	49.0	brown	light	brown	19.0	female	feminine	Alderaan	Human	c("The Empire Strikes Back", "Revenge of the Sith" [...]	Imperial Speeder Bike	character(0)
6	Owen Lars	178	120.0	brown, grey	light	blue	52.0	male	masculine	Tatooine	Human	c("Attack of the Clones", "Revenge of the Sith", " [...]	character(0)	character(0)
7	Beru Whitesun lars	165	75.0	brown	light	blue	47.0	female	feminine	Tatooine	Human	c("Attack of the Clones", "Revenge of the Sith", " [...]	character(0)	character(0)
8	R5-D4	97	32.0	NA	white, red	red	NA	none	masculine	Tatooine	Droid	A New Hope	character(0)	character(0)
9	Biggs Darklighter	183	84.0	black	light	brown	24.0	male	masculine	Tatooine	Human	A New Hope	character(0)	X-wing
10	Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	57.0	male	masculine	Stewjon	Human	c("The Empire Strikes Back", "Attack of the Clones [...]	Tribubble bongo	c("Jedi starfi
11	Anakin Skywalker	188	84.0	blond	fair	blue	41.9	male	masculine	Tatooine	Human	c("Attack of the Clones", "The Phantom Menace", "R [...]	c("Zephyr-G swoop bike", "XJ-6 airspeeder")	c("Trade Fed
12	Wilhuff Tarkin	180	NA	auburn, grey	fair	blue	64.0	male	masculine	Eriadu	Human	c("Revenge of the Sith", "A New Hope")	character(0)	character(0)
13	Chewbacca	228	112.0	brown	unknown	blue	200.0	male	masculine	Kashyyyk	Wookiee	c("The Empire Strikes Back", "Revenge of the Sith" [...]	AT-ST	c("Millenniu
14	Han Solo	180	80.0	brown	fair	brown	29.0	male	masculine	Corellia	Human	c("The Empire Strikes Back", "Return of the Jedi", [...]	character(0)	c("Millenniu
15	Greedo	173	74.0	NA	green	black	44.0	male	masculine	Rodia	Rodian	A New Hope	character(0)	character(0)

Figure 1: snapshot of the sample dataset

a note on the names():

I have a really hard time remembering what the names of my variables are, and because R is case-sensitive, how the names are formatted. We could fix this by converting all of our variable names to the same case, but for now just know that if you ever need a refresher on the names of the variables in your dataset (and how they're formatted!) you can run names(), like this:

```
names(starwars)
```

```
##  [1] "name"       "height"     "mass"       "hair_color" "skin_color"
##  [6] "eye_color"  "birth_year" "sex"        "gender"     "homeworld"
## [11] "species"    "films"      "vehicles"   "starships"
```

```r
#To know the unique values from particular variable
unique(starwars$hair_color)
```

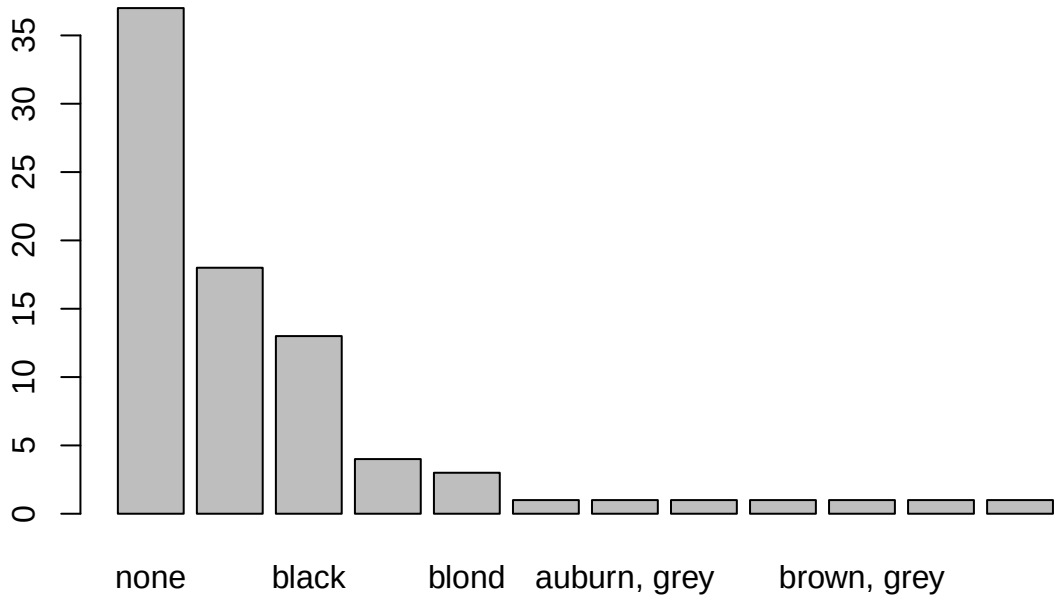
```
##  [1] "blond"         NA              "none"          "brown"
##  [5] "brown, grey"   "black"         "auburn, white" "auburn, grey"
##  [9] "white"         "grey"          "auburn"        "blonde"
## [13] "unknown"
```

```r
#To know how many observations are there against the unique values in the variable(categorical)
#also sort it in descending order
#and view in the neat table format like Pivot Table from Excel

View(sort(table(starwars$hair_color), decreasing = TRUE))
```

	Var1	Freq
1	none	37
2	brown	18
3	black	13
4	white	4
5	blond	3
6	auburn	1
7	auburn, grey	1
8	auburn, white	1
9	blonde	1
10	brown, grey	1
11	grey	1
12	unknown	1

```
#To get a frequency bar plot for the the particular variable(categorical)
barplot(sort(table(starwars$hair_color), decreasing = TRUE))
```



```
#Using pipes to get the same result

starwars %>%
  select(hair_color) %>%
  count(hair_color) %>%
  arrange(desc(n)) %>%
  view()
```

	hair_color	n
1	none	37
2	brown	18
3	black	13
4	*NA*	5
5	white	4
6	blond	3
7	auburn	1
8	auburn, grey	1
9	auburn, white	1
10	blonde	1
11	brown, grey	1
12	grey	1
13	unknown	1

Figure 2: Frequency of variable values using 'Pipe'

The NAs!:

NA stands for "Not Available", meaning data that is missing.If we don't handle our NA values we're going to be in for a bad time

```
#To know the number of missing values in the variable
starwars %>%
  select(hair_color) %>%
  is.na() %>%
  sum()
```

```
## [1] 5
```

```
#To know the number of missing values from all the variables
sapply(starwars,function(x) sum(is.na(x)))
```

```
##       name     height       mass hair_color skin_color  eye_color birth_year
##          0          6         28          5          0          0         44
##        sex     gender  homeworld    species      films   vehicles  starships
##          4          4         10          4          0          0          0
```

```
#exclude NA values while calculations
#na.rm is like asking the question, "Should we remove NAs from our code?"

starwars %>%
  summarise(avg_height = mean(height,na.rm = TRUE))
```

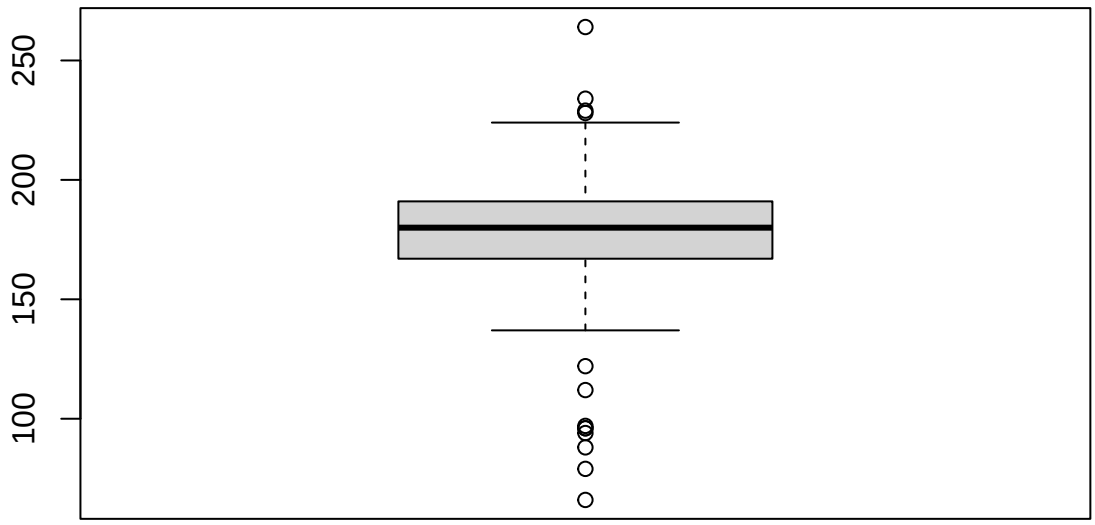
```
## # A tibble: 1 x 1
##   avg_height
##        <dbl>
## 1       174.
```

Working with the Numeric Variables

```
#To Know the Minimum, Maximum, Mean etc. values
#For Numeric type variable
summary(starwars %>%
          select(height))
```

```
##      height
##  Min.   : 66.0
##  1st Qu.:167.0
##  Median :180.0
##  Mean   :174.4
##  3rd Qu.:191.0
##  Max.   :264.0
##  NA's   :6
```

```
#To get the box plot
boxplot(starwars %>%
           select(height))
```



Skills Practice/Learn From This Analysis:

1. How to load the packages and data sets in R.
2. Now I know that 'conflicts' error is not an bad thing.
3. How to get a sense of the data using functions like 'dim', 'glimps','head'.
4. Understand the 'list' data types & it is different from the other data types.
5. How to use the 'names' function to know what the names of variables available in the data set.
6. How to use 'unique' function to know the unique values available in the categorical variable.
7. How to use the pipes operator along with the functions like 'select','count','arrange' to drill down the data set.
8. How to get the number of missing values 'NA' available in the variables from the data set.
9. How to work with the 'Numeric' data type variable to know their 'Min', 'Max', 'mean' values.
10. How to plot the box plot & understand how to read it.