Star Wars Jedi

A look at lightsabers, species, gender, rank, date of death, master, and apprentice.

```
import numpy as np
import urllib.request
import pandas as pd
import matplotlib.pyplot as plt
import re
```

First, let's grab the .csv file, which I downloaded from https://docs.google.com/spreadsheets/d/1tL__nCzQcQiNWqle-ej-we7krecKtopWDKOG0N7yMrl/edit#gid=177428702

Thank you to CanePlayz, who compiled this data and posted it in a discussion on https://starwars.fandom.com/f/u/27777468

```
jedi_csv=pd.read_csv(r"https://raw.githubusercontent.com/theRealJennie/therealjennie.github.io/main/Star%20Wars/List%20of%20All%20Jedi%20%5BC
```

```
#This chops the extra rows from the dataframe. From this point on, the csv file contained unverified Jedi.
jedi_csv=jedi_csv[0:170]
```

```
#Here is a list of the existing columns in this data set.
jedi_csv.columns
```

```
    Index(['Name', 'Rank', 'Lightsaber', 'Death', 'Species and gender',
           'Leaving/getting banned from the Jedi Order', 'Jedi Master(s)',
           'Jedi Apprentice(s)', 'Wookieepedia article'],
          dtype='object')
```

```
#We're not going to look into whether they left or got banned from the Jedi order. We'll drop that column, as well as the link to the Wookiee
#because we aren't using that information in this analysis.

jedi_csv=jedi_csv.drop('Leaving/getting banned from the Jedi Order',axis=1)
jedi_csv=jedi_csv.drop('Wookieepedia article',axis=1)
```

```
#Look at the column list again to make sure it worked correctly.
jedi_csv.columns
```

```
    Index(['Name', 'Rank', 'Lightsaber', 'Death', 'Species and gender',
           'Jedi Master(s)', 'Jedi Apprentice(s)'],
          dtype='object')
```

```
#Break gender into its own column, separate from species
jedi_csv['Gender']=jedi_csv["Species and gender"].str.rsplit(" ", 1).str[-1]
```

```
    <ipython-input-73-e712a7daf3b9>:2: FutureWarning: In a future version of pandas all arguments of StringMethods.rsplit except for the arg
      jedi_csv['Gender']=jedi_csv["Species and gender"].str.rsplit(" ", 1).str[-1]
```

```
#Make sure all entries are capitalized properly
jedi_csv['Gender']=jedi_csv['Gender'].str.capitalize()
```

```
#Replace NaN in the Gender column with 'Unspecified'
jedi_csv['Gender']=jedi_csv['Gender'].fillna("Unspecified")
```

```
#Let's take a quick look to see how our Gender column is showing up
np.unique(jedi_csv['Gender'])
```

```
    array(['Female', 'Male', 'Species', "Twi'lek", 'Unspecified'],
          dtype=object)
```

```
#We can see we have a couple of entries there that we don't want, and it now seems 'Unknown' is better than 'Unspecified', so let's take care
jedi_csv['Gender']=jedi_csv['Gender'].replace('Unspecified','Unknown')
jedi_csv['Gender']=jedi_csv['Gender'].replace('Species','Unknown')
jedi_csv['Gender']=jedi_csv['Gender'].replace("Twi'lek",'Unknown')
```

```
#Let's take another look at how our Gender column looks to make sure that worked
np.unique(jedi_csv['Gender'])
```

```
array(['Female', 'Male', 'Unknown'], dtype=object)
```

```
#Let's try to get rid of female and male in Species and Gender, since we put those in their own column
jedi_csv['Species and gender']=jedi_csv['Species and gender'].str.replace("female"," ")
jedi_csv['Species and gender']=jedi_csv['Species and gender'].str.replace("Female"," ")
jedi_csv['Species and gender']=jedi_csv['Species and gender'].str.replace("male"," ")
jedi_csv['Species and gender']=jedi_csv['Species and gender'].str.replace("Male"," ")

#While we're at it, let's take care of NaN in this field and blanks and replace them with "Unknown"
jedi_csv['Species and gender']=jedi_csv['Species and gender'].replace(np.nan,"Unknown")
jedi_csv['Species and gender'] = jedi_csv['Species and gender'].replace(r'^\s*$', "Unknown", regex=True)


#And to verify that worked, let's look at the data again
jedi_csv.sample(10)
```

	Name	Rank	Lightsaber	Death	Species and gender	Jedi Master(s)	Jedi Apprentice(s)	Gender
157	Veleckra	Jedi Master	NaN	NaN	Unknown	NaN	NaN	Unknown
135	Sora Bulq	NaN	Blue	NaN	Weequay	NaN	NaN	Male
51	Huulik	Jedi Knight	Blue	19 BBY, killed by clone troopers during Order ...	Rodian	NaN	NaN	Male
18	Byph	Jedi youngling	Blue	NaN	Ithorian	NaN	NaN	Male
164	Yoda	Jedi Grand Master	Green	4 ABY on Dagobah	Yoda's species	NaN	NaN	Male

```
#And let's rename that column, since it now only contains species
jedi_csv.rename(columns={'Species and gender':'Species'}, inplace=True)


#Now let's look at that Lightsaber column and see if we can clean that up.
#Let's replace those Nan values with Unknown
jedi_csv['Lightsaber']=jedi_csv['Lightsaber'].replace(np.nan,"Unknown")


#Let's see what unique values we have in there now
np.unique(jedi_csv['Lightsaber'])
```

```
array(['2 (unkn. colours), later 2 red', '2 green -> 2 blue -> 2 white',
       'Black (Darksaber)', 'Blue', 'Blue and green (double)',
       'Blue and green (hybrid)', 'Blue, later green', 'Blue, later red',
       'Blue, later yellow', 'Green', 'Green, later 2 red', 'Later red',
       'Later red (IQ)', 'Lightsaber sniper rifle', 'Purple', 'Unknown',
       'Yellow (double), later red (IQ)'], dtype=object)
```

```
#We have a lot to fix up there, so let's get started

#Let's create a field to specify what type of saber they have.
#We'll go with Single Blade, Double Blade, Two Single, Sniper Rifle, Hybrid Single Blade, or a mix of these separated by comma for our values

#First, let's add the new column
jedi_csv['Lightsaber Type']="Unknown" #Unknown is our default value

#Now let's take care of each of these entries
jedi_csv.loc[jedi_csv['Lightsaber']=="2 (unkn. colours), later 2 red",'Lightsaber Type']="Two Single Blades"
jedi_csv.loc[jedi_csv['Lightsaber']=='2 green -> 2 blue -> 2 white','Lightsaber Type']="Two Single Blades"
jedi_csv.loc[jedi_csv['Lightsaber']=='Black (Darksaber)','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue and green (double)','Lightsaber Type']="Dual Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue and green (hybrid)','Lightsaber Type']="Hybrid Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue, later green','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue, later red','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue, later yellow','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Green','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Green, later 2 red','Lightsaber Type']="Single Blade, Two Single Blades"
jedi_csv.loc[jedi_csv['Lightsaber']=='Later red','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Later red (IQ)','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Lightsaber sniper rifle','Lightsaber Type']="Sniper Rifle"
jedi_csv.loc[jedi_csv['Lightsaber']=='Purple','Lightsaber Type']="Single Blade"
jedi_csv.loc[jedi_csv['Lightsaber']=='Yellow (double), later red (IQ)','Lightsaber Type']="Dual Blade"
```

```
np.unique(jedi_csv['Lightsaber Type'])
```

```
array(['Dual Blade', 'Hybrid Single Blade', 'Single Blade',
       'Single Blade, Two Single Blades', 'Sniper Rifle',
       'Two Single Blades', 'Unknown'], dtype=object)
```

```
#Now let's see what we can do with those colors. Maybe a new field, 'Lightsaber Color'"List of All Jedi [Canon] - List of Jedi.csv"
jedi_csv['Lightsaber Color']="Unknown" #Unknown is our default value

#Our values are going to be Blue, Red, Green, White, Yellow, Purple, Dark Saber, and Multiple

jedi_csv.loc[jedi_csv['Lightsaber']=='2 green -> 2 blue -> 2 white','Lightsaber Color']="Multiple"
jedi_csv.loc[jedi_csv['Lightsaber']=='Black (Darksaber)','Lightsaber Color']="Dark Saber"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue','Lightsaber Color']="Blue"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue and green (double)','Lightsaber Color']="Multiple"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue and green (hybrid)','Lightsaber Color']="Multiple"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue, later green','Lightsaber Color']="Multiple"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue, later red','Lightsaber Color']="Multiple"
jedi_csv.loc[jedi_csv['Lightsaber']=='Blue, later yellow','Lightsaber Color']="Multiple"
jedi_csv.loc[jedi_csv['Lightsaber']=='Green','Lightsaber Color']="Green"
jedi_csv.loc[jedi_csv['Lightsaber']=='Green, later 2 red','Lightsaber Color']="Multiple"
jedi_csv.loc[jedi_csv['Lightsaber']=='Later red','Lightsaber Color']="Red"
jedi_csv.loc[jedi_csv['Lightsaber']=='Later red (IQ)','Lightsaber Color']="Red"
jedi_csv.loc[jedi_csv['Lightsaber']=='Purple','Lightsaber Color']="Purple"
jedi_csv.loc[jedi_csv['Lightsaber']=='Yellow (double), later red (IQ)','Lightsaber Color']="Multiple"
```

```
np.unique(jedi_csv['Lightsaber Color'])
```

```
array(['Blue', 'Dark Saber', 'Green', 'Multiple', 'Purple', 'Red',
       'Unknown'], dtype=object)
```

```
#So now let's take a look at that Death column.
#We should split that into Date of Death and Cause of Death

jedi_csv['Date of Death']="Unknown" #Unknown is our default value
jedi_csv['Cause of Death']="Unknown" #Unknown is our default value

#Let's get that Cause of death
jedi_csv['Cause of Death']=jedi_csv["Death"].str.rsplit(",", 1).str[-1]
jedi_csv['Date of Death']=jedi_csv["Death"].str.rsplit(",", 1).str[0]

#Replace NaN with unknown
jedi_csv['Cause of Death']=jedi_csv['Cause of Death'].replace(np.nan,"Unknown")
jedi_csv['Date of Death']=jedi_csv['Date of Death'].replace(np.nan,"Unknown")

#And let's capitalize those fields to normalize everything
jedi_csv['Cause of Death']=jedi_csv['Cause of Death'].str.strip()
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.strip()

jedi_csv['Cause of Death']=jedi_csv['Cause of Death'].str.capitalize()
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.capitalize()
```

```
<ipython-input-88-46aa6bbf14cc>:8: FutureWarning: In a future version of pandas all arguments of StringMethods.rsplit except for the arg
  jedi_csv['Cause of Death']=jedi_csv["Death"].str.rsplit(",", 1).str[-1]
<ipython-input-88-46aa6bbf14cc>:9: FutureWarning: In a future version of pandas all arguments of StringMethods.rsplit except for the arg
  jedi_csv['Date of Death']=jedi_csv["Death"].str.rsplit(",", 1).str[0]
```

```
#Let's see how we stand now
jedi_csv.sample(20)
```

	Name	Rank	Lightsaber	Death	Species	Jedi Master(s)	Jedi Apprentice(s)	Gender	Lightsaber Type
164	Yoda	Jedi Grand Master (HCM)	Green	4 ABY on Dagobah	Yoda's species	NaN	NaN	Male	Single Blade
16	Bolla Ropal	Jedi Master	Unknown	21 BBY, killed by Cad Bane in the Devaron system	Rodian	NaN	NaN	Male	Unknown
160	Wom-Nii Gnaden	Jedi Master	Unknown	NaN	Unknown	NaN	NaN	Male	Unknown
151	Trilla Suduri	Jedi Padawan	Later red	14 BBY, killed by Darth Vader on Nur	Human	NaN	NaN	Female	Single Blade
102	Oslord	Jedi Master	Unknown	NaN	Unknown	NaN	NaN	Male	Unknown
35	Elio	Jedi Master	Unknown	NaN	Unknown	NaN	NaN	Unknown	Unknown
85	Melik Galerha	Jedi Knight	Unknown	NaN	Human	NaN	NaN	Male	Unknown
163	Yeeda	NaN	Unknown	NaN	Unknown	NaN	NaN	Unknown	Unknown
135	Sora Bulq	NaN	Blue	NaN	Weequay	NaN	NaN	Male	Single Blade
93	Niobaya	Jedi Master	Unknown	NaN	Unknown	NaN	NaN	Unknown	Unknown
117	Radaki	NaN	Later red	NaN	Unknown	NaN	NaN	Male	Single Blade
88	Nahdar Vebb	Jedi Knight	Blue	22 BBY, killed by Grievous on	Mon Calamari	NaN	NaN	Male	Single Blade

```
#Let's take a closer look at that Date of Death column
np.unique(jedi_csv['Date of Death'])
```

```
    array(['0 bby', '14 bby', '18 bby', '19 bby', '20 bby', '21 bby',
           '22 bby', '3 bby', '32 bby', '32 bby - 22 bby', '34 aby', '35 aby',
           '35 aby, killed by darth sidious on exegol', '4 aby',
           '4 aby on dagobah', '4 bby', 'Around 30 aby', 'Around 32 bby',
           'Around 48 bby', 'Between 18 bby and 14 bby',
           'Between 19 bby and 18 bby', 'Between 22 bby and 20 bby',
           'By 9 bby', 'Prior to 22 bby',
           'Prior to the establishment of the galactic republic',
           'Sometime prior to 0 bby', 'Sometime prior to 19 bby',
           'Sometime prior to 22 bby', 'Unknown',
           'Within a decade of the corsair wars'], dtype=object)
```

```
#There's a lot of data we need to break apart there if we're going to be able to graph it and make any sense of it, so let's get started.

#First, let's replace all the "between x bby and y bby" with x-y bby
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace(" bby and ","-")

#Get rid of "Between"
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace("Between ","")

#There are still some " bby - " entries to take care of
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace(" bby - ","-")

#Let's get rid of "Around","Prior to", "By", and "Sometime"
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace("Around","") #Don't need "around" because we are close enough for our uses by
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace("Sometime","") #Sometime is just a waste of space for us here
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace("Prior to","Before") #Before covers this more succinctly
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace("prior to","Before") #Before covers this more succinctly
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace("By","") #See Around above
```

```
#Now let's clean up that reference to the Corsair wars, 'Before the establishment of the galactic republic', and " on dagobah"
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace('Within a decade of the corsair wars',"By 19 BBY") #Corsair Wars were prior t
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace('Before the establishment of the galactic republic',"By 25,000 BBY") #Per Sta
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace(" on dagobah","")
```

```
#Let's see where we're at now
np.unique(jedi_csv['Date of Death'])
```

```
     array([' 30 aby', ' 32 bby', ' 48 bby', ' 9 bby', ' Before 0 bby',
            ' Before 19 bby', ' Before 22 bby', '0 bby', '14 bby', '18 bby',
            '18-14 bby', '19 bby', '19-18 bby', '20 bby', '21 bby', '22 bby',
            '22-20 bby', '3 bby', '32 bby', '32-22 bby', '34 aby', '35 aby',
            '35 aby, killed by darth sidious on exegol', '4 aby', '4 bby',
            'Before 22 bby', 'By 19 BBY', 'By 25,000 BBY', 'Unknown'],
           dtype=object)
```

```
#You can see we still have a litte work to do. We have lower case bby and aby, spaces to start some values, and that pesky one that talks abo

#Let's upper case all those bby and aby first
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace("bby","BBY")
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace("aby","ABY")

#Now let's strip away that extra space
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.strip()
```

```
#But we still have the killed by darth sidious entry that can be problematic. Before we just stip that part off, let's check to make sure tha
#Cause of Death column properly
jedi_csv[jedi_csv['Date of Death']=='35 ABY, killed by darth sidious on exegol']
```

	Name	Rank	Lightsaber	Death	Species	Jedi Master(s)	Jedi Apprentice(s)	Gender	Lightsaber Type	Lightsa Co
				35 ABY,						

```
#We can see that the cause of death didn't get that, so let's set that real quick. It's only one row, so we'll dot it the quick way.
jedi_csv['Cause of Death']=jedi_csv['Cause of Death'].str.replace("Later resuscitated","Killed by Darth Sidious on Exegol. Later resuscitated
```

```
jedi_csv[jedi_csv['Date of Death']=='35 ABY, killed by darth sidious on exegol']
```

	Name	Rank	Lightsaber	Death	Species	Jedi Master(s)	Jedi Apprentice(s)	Gender	Lightsaber Type	Lightsa Co
				35 ABY,						

```
#And now we can clean up the Date of Death field with sidious in it
jedi_csv['Date of Death']=jedi_csv['Date of Death'].str.replace('35 ABY, killed by darth sidious on exegol',"35 BBY")
```

```
#Let's see where we're at with that Date of Death now
np.unique(jedi_csv['Date of Death'])
```

```
     array(['0 BBY', '14 BBY', '18 BBY', '18-14 BBY', '19 BBY', '19-18 BBY',
            '20 BBY', '21 BBY', '22 BBY', '22-20 BBY', '3 BBY', '30 ABY',
            '32 BBY', '32-22 BBY', '34 ABY', '35 ABY', '35 BBY', '4 ABY',
            '4 BBY', '48 BBY', '9 BBY', 'Before 0 BBY', 'Before 19 BBY',
            'Before 22 BBY', 'By 19 BBY', 'By 25,000 BBY', 'Unknown'],
           dtype=object)
```

```
#That looks pretty clean, so let's now take a look at the Cause of Death field
np.unique(jedi_csv['Cause of Death'])
```

```
     array(['21 bby', '32 bby - 22 bby', '4 aby on dagobah', 'By 9 bby',
            'Defeated by kanan jarrus aboard the sovereign',
            'Died on ahch-to after projecting himself to crait',
            'Died on ajan kloss after reaching out to ben solo on kef bir',
            "Died on pam'ba",
            'Killed by Darth Sidious on Exegol. Later resuscitated.',
            'Killed by ahsoka tano on raada',
            'Killed by an anooba on lola sayu',
            'Killed by anakin skywalker aboard the invisible hand',
            'Killed by asajj ventress',
            'Killed by cad bane in the devaron system',
            'Killed by cal kestis and merrin on dathomir',
            'Killed by clone troopers during order 66 in the bracca system',
            'Killed by clone troopers during order 66 on coruscant',
            'Killed by clone troopers during order 66 on felucia',
            'Killed by clone troopers during order 66 on his way to rodia',
            'Killed by clone troopers during order 66 on kaller',
            'Killed by clone troopers during order 66 on mygeeto',
            'Killed by clone troopers during order 66 on saleucami',
            'Killed by clone troopers during order 66 on zeffo',
            'Killed by clone troopers during order 66 over cato neimodia',
            'Killed by darth maul', 'Killed by darth maul on naboo',
            'Killed by darth maul on the moon of drazkel',
            'Killed by darth sidious and anakin skywalker',
            'Killed by darth sidious on coruscant', 'Killed by darth vader',
            'Killed by darth vader on coruscant',
            'Killed by darth vader on mon cala',
            'Killed by darth vader on nur',
            'Killed by darth vader on the death star',
            "Killed by darth vader on the river moon of al'doleem",
            'Killed by dogma on umbara', 'Killed by dooku in vizsla keep 09',
            'Killed by dooku on christophisis', 'Killed by dooku on sullust',
            'Killed by grievous', 'Killed by grievous on coruscant',
            'Killed by grievous on vassek 3',
            'Killed by jango fett on geonosis',
            'Killed by kanan jarrus on malachor', 'Killed by maul on malachor',
            'Killed by purge troopers on mon cala',
            'Killed by quinlas vos aboard the vigilance',
            'Killed by savage opress on devaron',
            'Killed by savage opress on florrum',
            'Killed by separatist battle droids',
            'Killed by separatist battle droids on geonosis',
            'Killed by separatist battle droids on mimban',
            'Killed by separatist battle droids on ryloth',
            'Killed by tup aboard the ringo vinda space station',
            'Killed by weequay raiders on rattatak',
            "Killed during barris offee's bombing of the jedi temple hangar",
            'Killed on the spire',
            'Killed when his shuttle got shot over the oba diah moon',
            'Killed when the crew of the freighter advent mutinied',
            'Prior to the establishment of the galactic republic',
            'Purged by other jedi',
            'Sacrificed his life for luke skywalker on the death star ii',
            'Sacrificed his life for rey on exegol',
            'Sacrificed his life for the ghost crew on lothal',
            'Sometime prior to 0 bby', 'Unknown', 'Utapau'], dtype=object)
```

```
#We're more interested in way they died, not where, so let's get rid of all of those "on ..."
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.rsplit(" on", 1).str[0]

#We also have a lot of "in the ... system", so let's get rid of that
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.rsplit(" in the", 1).str[0]

#We also have a lot of "over ... system", so let's get rid of that
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.rsplit(" over ", 1).str[0]

#Let's start to clean up and normalize some of those names
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('darth',"Darth")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('sidious',"Sidious")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('maul',"Maul")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('skywalker',"Skywalker")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('anakin',"Anakin")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('luke',"Luke")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('vader',"Vader")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('grievous',"Grievous")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('dooku',"Dooku")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('cad bane',"Cad Bane")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('dogma',"Dogma")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('ahsoka tano',"Ahsoka Tano")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('anooba',"Anooba")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('kanan jarrus',"Kanan Jarrus")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('jango fett',"Jango Fett")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('vizsla keep',"Vizsla Keep")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('cal kestis',"Cal Kestis")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('merrin',"Merrin")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('the sovereign',"the Sovereign")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('asajj ventress',"Asajj Ventress")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('barris offee',"Barriss Offee")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('advent ',"Advent ")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace(' rey'," Rey")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('the invisible hand',"the Invisible Hand")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('quinlan vos',"Quinlan Vos")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('quinlas vos',"Quinlan Vos")
```

```
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('quinlas vos',"Quinlan Vos")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('sidious',"Sidious")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('jedi',"Jedi")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('the vigilance',"the Vigilance")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('savage opress',"Savage Opress")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('ringo vinda',"Ringa Vinda")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('weequay',"Weequay")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('the ghost',"the Ghost")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('order 66',"Order 66")
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace(' tup'," Tup")
```

```
<ipython-input-100-95dfbcff54e1>:2: FutureWarning: In a future version of pandas all arguments of StringMethods.rsplit except for the ar
  jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.rsplit(" on", 1).str[0]
<ipython-input-100-95dfbcff54e1>:5: FutureWarning: In a future version of pandas all arguments of StringMethods.rsplit except for the ar
  jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.rsplit(" in the", 1).str[0]
<ipython-input-100-95dfbcff54e1>:8: FutureWarning: In a future version of pandas all arguments of StringMethods.rsplit except for the ar
  jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.rsplit(" over ", 1).str[0]
```

```
np.unique(jedi_csv['Cause of Death'])
```

```
array(['21 bby', '32 bby - 22 bby', '4 aby', 'By 9 bby',
       'Defeated by Kanan Jarrus aboard the Sovereign', 'Died',
       'Died on ajan kloss after reaching out to ben solo', 'Killed',
       'Killed by Ahsoka Tano',
       'Killed by Anakin Skywalker aboard the Invisible Hand',
       'Killed by Asajj Ventress', 'Killed by Cad Bane',
       'Killed by Cal Kestis and Merrin', 'Killed by Darth Maul',
       'Killed by Darth Sidious',
       'Killed by Darth Sidious and Anakin Skywalker',
       'Killed by Darth Vader', 'Killed by Dogma', 'Killed by Dooku',
       'Killed by Dooku in Vizsla Keep 09', 'Killed by Grievous',
       'Killed by Jango Fett', 'Killed by Kanan Jarrus', 'Killed by Maul',
       'Killed by Quinlan Vos aboard the Vigilance',
       'Killed by Savage Opress',
       'Killed by Tup aboard the Ringa Vinda space station',
       'Killed by Weequay raiders', 'Killed by an Anooba',
       'Killed by clone troopers during Order 66',
       'Killed by purge troopers', 'Killed by separatist battle droids',
       "Killed during Barriss Offee's bombing of the Jedi temple hangar",
       'Killed when his shuttle got shot',
       'Killed when the crew of the freighter Advent mutinied',
       'Prior to the establishment of the galactic republic',
       'Purged by other Jedi', 'Sacrificed his life for Luke Skywalker',
       'Sacrificed his life for Rey',
       'Sacrificed his life for the Ghost crew',
       'Sometime prior to 0 bby', 'Unknown', 'Utapau'], dtype=object)
```

```
#That looks somewhat better, so now let's look at the rows of some of those and make sure some of that data, like years, appeared in the righ
print(jedi_csv[jedi_csv['Cause of Death'].isin( ['21 bby','32 bby - 22 bby','4 aby', 'By 9 bby','Sometime prior to 0 bby'])])
```

```
             Name                     Rank Lightsaber  \
90       Nes Ukul             Jedi Padawan    Unknown   
101  Ord Enisence              Jedi Knight    Unknown   
143   Tera Sinube              Jedi Master       Blue   
161   Yarael Poof        Jedi Master (HCM)       Blue   
164          Yoda  Jedi Grand Master (HCM)      Green   

                       Death           Species Jedi Master(s)  \
90   Sometime prior to 0 BBY           Unknown            NaN   
101                   21 BBY       Skrilling              NaN   
143                 By 9 BBY          Cosian              NaN   
161          32 BBY - 22 BBY        Quermian              NaN   
164         4 ABY on Dagobah  Yoda's species              NaN   

    Jedi Apprentice(s) Gender Lightsaber Type Lightsaber Color Date of Death  \
90                 NaN   Male         Unknown          Unknown  Before 0 BBY   
101                NaN   Male         Unknown          Unknown        21 BBY   
143                NaN   Male    Single Blade             Blue         9 BBY   
161                NaN   Male    Single Blade             Blue     32-22 BBY   
164                NaN   Male    Single Blade            Green         4 ABY   

              Cause of Death  
90   Sometime prior to 0 bby  
101                   21 bby  
143                 By 9 bby  
161          32 bby - 22 bby  
164                    4 aby  
```

```
#We can just set all of those to Unknown, since the values showed up correctly in the Date of Death field
#jedi_csv[jedi_csv['Cause of Death'].isin( ['21 bby','32 bby - 22 bby','4 aby', 'By 9 bby','Sometime prior to 0 bby'])]
```

```
jedi_csv['Cause of Death']=np.where(jedi_csv['Cause of Death'].isin(['21 bby','32 bby - 22 bby','4 aby', 'By 9 bby','Sometime prior to 0 bby'
```

```
#Let's see how we stand now
np.unique(jedi_csv['Cause of Death'])
```

```
    array(['Defeated by Kanan Jarrus aboard the Sovereign', 'Died',
           'Died on ajan kloss after reaching out to ben solo', 'Killed',
           'Killed by Ahsoka Tano',
           'Killed by Anakin Skywalker aboard the Invisible Hand',
           'Killed by Asajj Ventress', 'Killed by Cad Bane',
           'Killed by Cal Kestis and Merrin', 'Killed by Darth Maul',
           'Killed by Darth Sidious',
           'Killed by Darth Sidious and Anakin Skywalker',
           'Killed by Darth Vader', 'Killed by Dogma', 'Killed by Dooku',
           'Killed by Dooku in Vizsla Keep 09', 'Killed by Grievous',
           'Killed by Jango Fett', 'Killed by Kanan Jarrus', 'Killed by Maul',
           'Killed by Quinlan Vos aboard the Vigilance',
           'Killed by Savage Opress',
           'Killed by Tup aboard the Ringa Vinda space station',
           'Killed by Weequay raiders', 'Killed by an Anooba',
           'Killed by clone troopers during Order 66',
           'Killed by purge troopers', 'Killed by separatist battle droids',
           "Killed during Barriss Offee's bombing of the Jedi temple hangar",
           'Killed when his shuttle got shot',
           'Killed when the crew of the freighter Advent mutinied',
           'Prior to the establishment of the galactic republic',
           'Purged by other Jedi', 'Sacrificed his life for Luke Skywalker',
           'Sacrificed his life for Rey',
           'Sacrificed his life for the Ghost crew', 'Unknown', 'Utapau'],
          dtype=object)
```

```
#Let's take a look at the remaining problematic entries

#Prior to the establishment of the galactic republic
jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('Prior to the establishment of the galactic republic',"Unknown")

jedi_csv['Cause of Death']=jedi_csv["Cause of Death"].str.replace('Utupau',"Unknown")
```

```
#Let's see how we stand now
np.unique(jedi_csv['Cause of Death'])
```

```
    array(['Defeated by Kanan Jarrus aboard the Sovereign', 'Died',
           'Died on ajan kloss after reaching out to ben solo', 'Killed',
           'Killed by Ahsoka Tano',
           'Killed by Anakin Skywalker aboard the Invisible Hand',
           'Killed by Asajj Ventress', 'Killed by Cad Bane',
           'Killed by Cal Kestis and Merrin', 'Killed by Darth Maul',
           'Killed by Darth Sidious',
           'Killed by Darth Sidious and Anakin Skywalker',
           'Killed by Darth Vader', 'Killed by Dogma', 'Killed by Dooku',
           'Killed by Dooku in Vizsla Keep 09', 'Killed by Grievous',
           'Killed by Jango Fett', 'Killed by Kanan Jarrus', 'Killed by Maul',
           'Killed by Quinlan Vos aboard the Vigilance',
           'Killed by Savage Opress',
           'Killed by Tup aboard the Ringa Vinda space station',
           'Killed by Weequay raiders', 'Killed by an Anooba',
           'Killed by clone troopers during Order 66',
           'Killed by purge troopers', 'Killed by separatist battle droids',
           "Killed during Barriss Offee's bombing of the Jedi temple hangar",
           'Killed when his shuttle got shot',
           'Killed when the crew of the freighter Advent mutinied',
           'Purged by other Jedi', 'Sacrificed his life for Luke Skywalker',
           'Sacrificed his life for Rey',
           'Sacrificed his life for the Ghost crew', 'Unknown', 'Utapau'],
          dtype=object)
```

```
#That looks pretty good, so let's see what else we need to do to our data
jedi_csv.sample(20)
```

	Name	Rank	Lightsaber	Death	Species	Jedi Master(s)	Jedi Apprentice(s)	Gender	Lightsaber Type
110	Pong Krell	Jedi Master	Blue and green (double)	20 BBY, killed by Dogma on Umbara	Besalisk	NaN	NaN	Male	Dual Blade
164	Yoda	Jedi Grand Master (HCM)	Green	4 ABY on Dagobah	Yoda's species	NaN	NaN	Male	Single Blade
41	Fifth Brother	NaN	Later red (IQ)	3 BBY, killed by Maul on Malachor	Humanoid	NaN	NaN	Male	Single Blade
103	Ovana	Jedi youngling	Unknown	NaN	Unknown	NaN	NaN	Unknown	Unknown
157	Veleckra	Jedi Master	Unknown	NaN	Unknown	NaN	NaN	Unknown	Unknown
32	Eeth Koth	Jedi Master (HCM)	Green	Between 18 BBY and 14 BBY, killed by Darth Vader	Zabrak (Iridonian)	NaN	NaN	Male	Single Blade
33	Eight Brother	NaN	Later red (IQ)	3 BBY, killed by Kanan Jarrus on Malachor	Terrelian Jango Jumper	NaN	NaN	Male	Single Blade
160	Wom-Nii Gnaden	Jedi Master	Unknown	NaN	Unknown	NaN	NaN	Male	Unknown
23	Chon Actrion	Jedi Master	Unknown	NaN	Unknown	NaN	NaN	Unknown	Unknown
62	Kanan	Jedi	Blue	0 BBY, sacrificed his life for	Human	NaN	NaN	Male	Single

```
#I'm going to take care of a couple of things here.

#first, "Eight Brother" should be "Eighth Brother"
jedi_csv['Name']=jedi_csv['Name'].str.replace("Eight Brother","Eighth Brother")

#Then let's replace NaN with "Sith" if they are a known sith, such as Eighth Brother or Fifth Brother
jedi_csv['Rank']=np.where(jedi_csv['Name'].str.contains('Brother'),"Sith",jedi_csv['Rank'])
jedi_csv['Rank']=np.where(jedi_csv['Name'].str.contains('Sister'),"Sith",jedi_csv['Rank'])

#Jedi Master(s) NaN need fixed
jedi_csv['Jedi Master(s)']=jedi_csv['Jedi Master(s)'].replace(np.nan,"Unknown")

#Jedi Apprentice(s) NaN need fixed
jedi_csv['Jedi Apprentice(s)']=jedi_csv['Jedi Apprentice(s)'].replace(np.nan,"Unknown")


#We know Padawans and younglings don't have apprentices, so let's take care of that
jedi_csv['Jedi Apprentice(s)']=np.where(jedi_csv['Rank'].str.contains('Padawan'),"None",jedi_csv['Jedi Apprentice(s)'])
jedi_csv['Jedi Apprentice(s)']=np.where(jedi_csv['Rank'].str.contains('youngling'),"None",jedi_csv['Jedi Apprentice(s)'])


#And I can see that the rank column contains NaN values, so let's take care of that
jedi_csv['Rank']=jedi_csv['Rank'].replace(np.nan,"Unknown")


#Let's see where that rank column stands
jedi_csv['Rank'].unique()
```

```
    array(['Jedi Master', 'Jedi Master (HCM)', 'Jedi Padawan', 'Jedi Knight',
           'Jedi Knight (HCM)', 'Jedi youngling', 'Unknown', 'Sith',
           'Founder', '\n', 'Jedi doctor', 'Jedi Temple Guard',
           'Jedi Grand Master (HCM)'], dtype=object)
```

```
#I can see we need to get rid of a "/n", need to upper case 'doctor' and 'youngling', and I'm not sure what(HCM) is, but let's get rid of tha
jedi_csv['Rank']=jedi_csv['Rank'].str.replace("\n","Unknown")
```

```
jedi_csv['Rank']=jedi_csv['Rank'].str.replace("doctor","Doctor")
jedi_csv['Rank']=jedi_csv['Rank'].str.replace("youngling","Youngling")
jedi_csv['Rank']=jedi_csv['Rank'].str.replace("(HCM)","", regex=False) #Have to set regex to false or it detects that as regular expression b
```

```
#Let's see where that rank column stands now
jedi_csv['Rank'].unique()
```

```
array(['Jedi Master', 'Jedi Master ', 'Jedi Padawan', 'Jedi Knight',
       'Jedi Knight ', 'Jedi Youngling', 'Unknown', 'Sith', 'Founder',
       'Jedi Doctor', 'Jedi Temple Guard', 'Jedi Grand Master '],
      dtype=object)
```

```
#So now let's take another look at our current data
jedi_csv.sample(25)
```

	Name	Rank	Lightsaber	Death	Species	Jedi Master(s)	Jedi Apprentice(s)	Gender	Lightsaber Type
89	Naq Med	Jedi Padawan	Unknown	Around 30 ABY, died on	Human	Unknown	None	Male	Unknown

```
#The only column with any NaN values remainins is the Death column, which I don't intend to use but am leaving because I may.
#Because of this, I'm getting rid of the NaN values in those fields

jedi_csv['Death']=jedi_csv['Death'].replace(np.nan,"Unknown")
```

88 Infil'a Master Green Vader on Unknown Unknown Unknown Unknown Blade

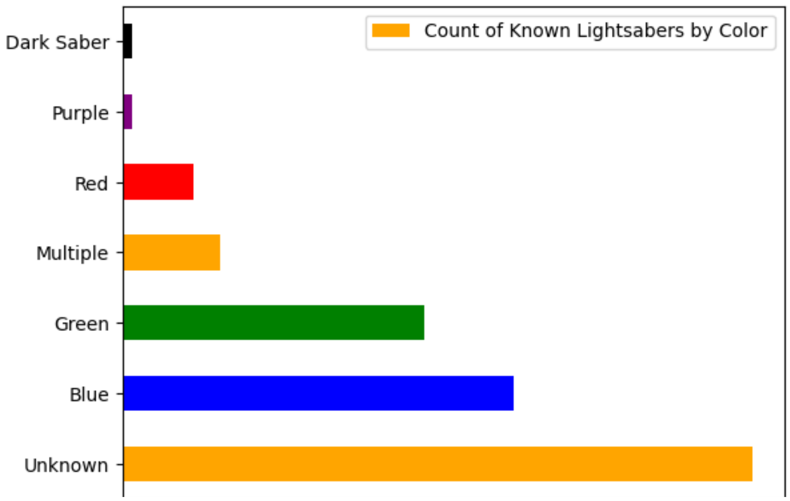
```
#So now let's look and see if we see any other issues
#jedi_csv.sample(25)
print(jedi_csv)
```

```
            Name           Rank                   Lightsaber  \
0    Aayla Secura     Jedi Master                         Blue
1      Adi Gallia    Jedi Master                         Blue
2      Agen Kolar    Jedi Master        Blue and green (hybrid)
3     Ahsoka Tano    Jedi Padawan  2 green -> 2 blue -> 2 white
4     Akar-Deshu      Jedi Knight                         Blue
..           ...            ...                          ...
165  Yula Braylon    Jedi Master                      Unknown
166  Zang Arraira    Jedi Padawan                      Unknown
167          Zatt   Jedi Youngling                        Green
168  Zett Jukassa    Jedi Padawan                         Blue
169   Zharva Kall         Unknown                      Unknown

                                                 Death            Species  \
0    19 BBY, killed by clone troopers during Order ...            Twi'lek
1           20 BBY, killed by Savage Opress on Florrum         Tholothian
2         19 BBY, killed by Darth Sidious on Coruscant  Zabrak (Iridonian)
3                                              Unknown            Togruta
4    19 BBY, killed by Quinlas Vos aboard the Vigil...             Mahran
..                                                 ...                ...
165                                            Unknown            Unknown
166                                            Unknown            Unknown
167                                            Unknown           Nautolan
168  19 BBY, killed by clone troopers during Order ...              Human
169                                            Unknown            Unknown

       Jedi Master(s) Jedi Apprentice(s)   Gender      Lightsaber Type  \
0         Quinlan Vos            Unknown   Female         Single Blade
1             Unknown            Unknown   Female         Single Blade
2             Unknown         Tan Yuster     Male  Hybrid Single Blade
3    Anakin Skywalker               None   Female    Two Single Blades
4             Unknown            Unknown     Male         Single Blade
..                ...                ...      ...                  ...
165           Unknown            Unknown   Female              Unknown
166           Unknown               None   Female              Unknown
167           Unknown               None     Male         Single Blade
168           Unknown               None     Male         Single Blade
169           Unknown               None  Unknown              Unknown

    Lightsaber Color Date of Death                            Cause of Death
0               Blue        19 BBY    Killed by clone troopers during Order 66
1               Blue        20 BBY                   Killed by Savage Opress
2           Multiple        19 BBY                   Killed by Darth Sidious
3           Multiple       Unknown                                   Unknown
4               Blue        19 BBY  Killed by Quinlan Vos aboard the Vigilance
..               ...           ...                                       ...
165          Unknown       Unknown                                   Unknown
166          Unknown       Unknown                                   Unknown
167            Green       Unknown                                   Unknown
168             Blue        19 BBY    Killed by clone troopers during Order 66
169          Unknown       Unknown                                   Unknown

[170 rows x 12 columns]
```

on

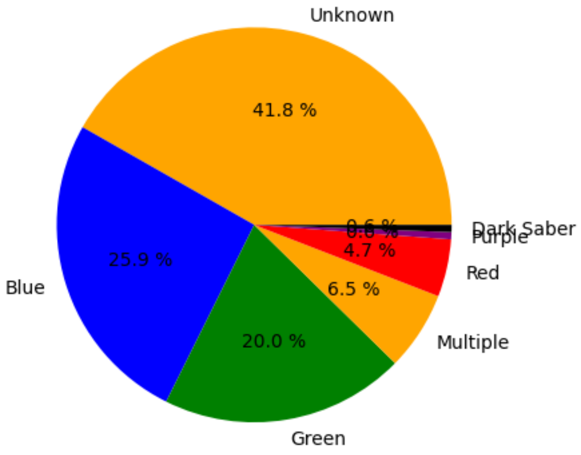
```
#That looks good, so now we can start craeting some visuals to show the results of all this hard work.

#First, let's look at a chart of lightsaber colors
df_sabers=pd.DataFrame(jedi_csv['Lightsaber Color'], columns=['Lightsaber Color'])

df_sabers['Lightsaber Color'].value_counts()[:20].plot(kind='barh', color=['orange','blue','green', 'orange','red','purple','black'], xticks=
plt.legend(("Count of Known Lightsabers by Color",))
```

```
<matplotlib.legend.Legend at 0x7f10d4aa3820>
```



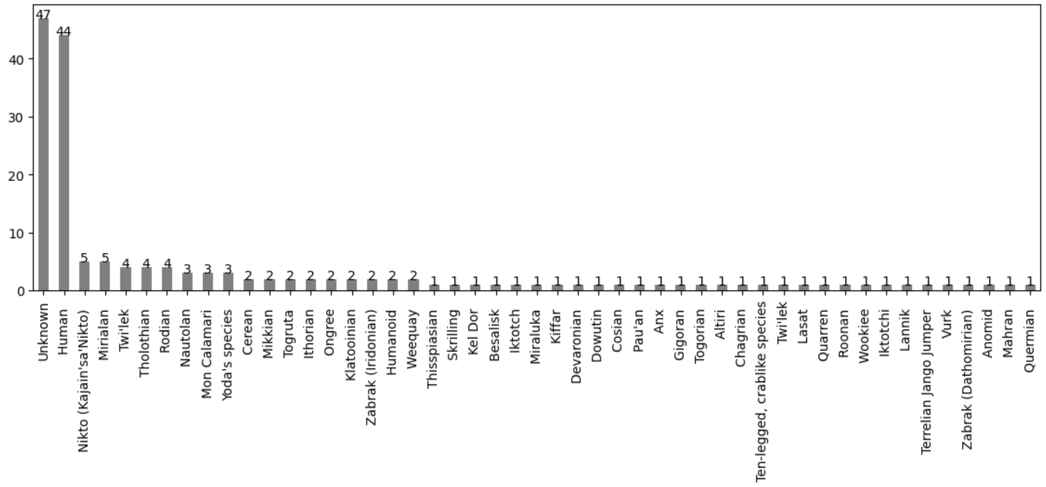
```
#And we can see what percentage of the total each color makes up using a pie chart.
s=jedi_csv['Lightsaber Color'].squeeze()
s.value_counts(normalize=True).plot.pie(autopct='%.1f %%', ylabel='', legend=False, colors=['orange','blue','green','orange','red','purple','
```

```
<Axes: >
```



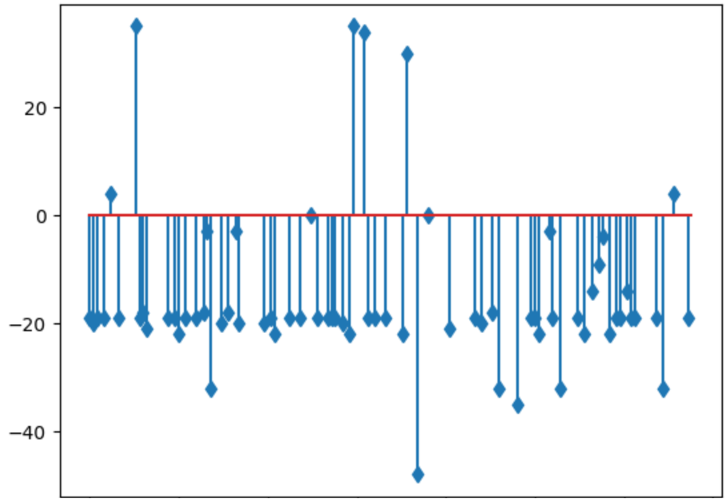
```
# Let's take a look at what the breakdown by gender of the Jedi are
jedi_csv['Gender'].value_counts().plot(kind='bar', color=['blue','red','brown'], yticks=jedi_csv['Gender'].value_counts())
```

```
#We can see that there are over twice as many male Jedi as female, based on those whose gender is known.
#This is not necessarily the case, because all of the Unknown may be female. If that were the case then
#there would be closer to a 4:3 ratio. I'd still like to see more balance there, personally.

#Let's take a look at species
plt.figure(figsize=(14,4))
jedi_csv['Species'].value_counts().plot(kind='bar', color='gray')
y=jedi_csv['Species'].value_counts()
x=jedi_csv['Species'].unique()

#plt.bar(x,y)
plt.xticks(rotation=90)
#plt.bar(range(len(y)), sorted(y))
for i in range(len(x)):
    plt.text(i,y[i], y[i],ha='center')
```



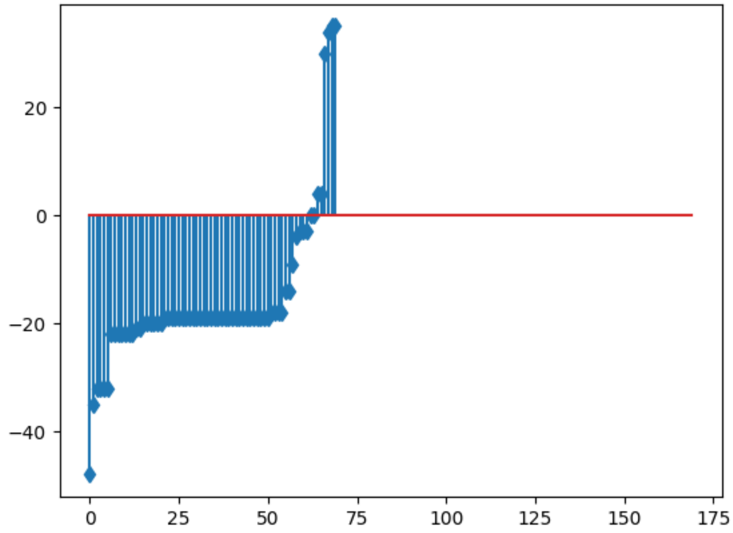
```
#We can see that we don't know the species of most of the Jedi, but of those we do know, most are human.
#This could mean the unknown continue at the same ratio across the species, but they may all be Qermian or Zabrak,
#So we can't draw concrete conclusions. We can, however, take this as an inference that humans are more likely to become Jedi.

#Let's take a look at the date of death

#To properly plot this we'll need it in a numerical format, so we'll go about doing that with BBY as a negative and ABY as a positive
#plt.stem(jedi_csv['Date of Death'].value_counts())

jedi_csv['Date of Death Numerical']=None

import re

for index,row in jedi_csv.iterrows():
    if row['Date of Death'][0].isdigit():
        row['Date of Death Numerical']=re.findall(r'\d+',row['Date of Death'])
        row['Date of Death Numerical']=int(row['Date of Death Numerical'][0])
        if "BBY" in row['Date of Death']:
            row['Date of Death Numerical']=row['Date of Death Numerical']-(2*row['Date of Death Numerical'])

#SO now we should have numerical years they died. Let's try to plot that.

plt.stem(jedi_csv['Date of Death Numerical'], markerfmt='d')
#That looks ok, but we really need to spruce this up a little. Get rid of the stems? Maybe group like numbers, and add BBY and ABY labels to
```

```
<StemContainer object of 3 artists>
```



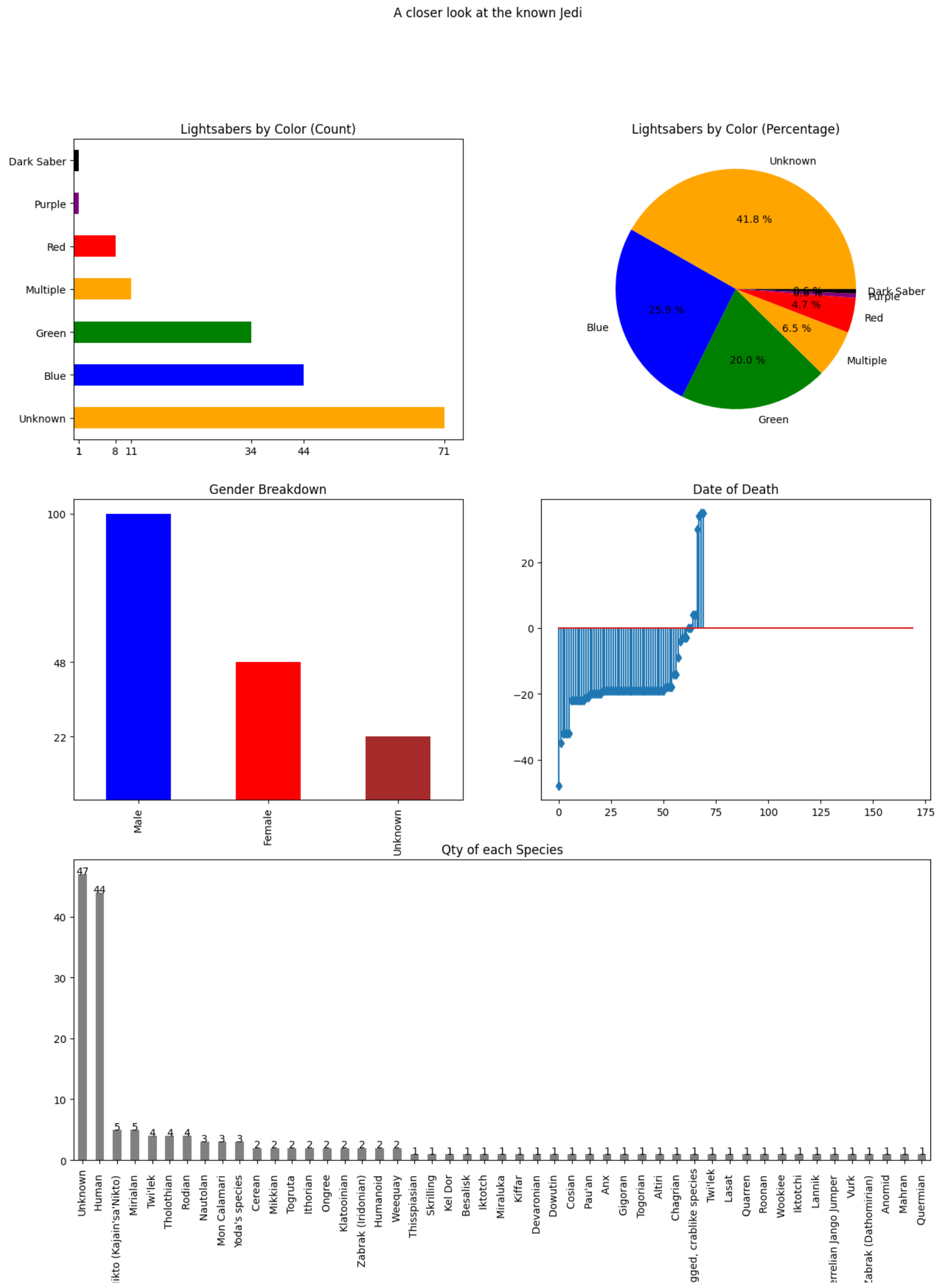
```
jedi_csv['Date of Death Numerical']=None

import re

for index,row in jedi_csv.iterrows():
    if row['Date of Death'][0].isdigit():
        row['Date of Death Numerical']=re.findall(r'\d+',row['Date of Death'])
        row['Date of Death Numerical']=int(row['Date of Death Numerical'][0])
        if "BBY" in row['Date of Death']:
            row['Date of Death Numerical']=row['Date of Death Numerical']-(2*row['Date of Death Numerical'])

#SO now we should have numerical years they died. Let's try to plot that.

plt.stem(jedi_csv['Date of Death Numerical'].sort_values(axis=0), markerfmt='d')
```

```
<StemContainer object of 3 artists>
```



```
#We can see that the great majority of Jedi died about 19 BBY, which is buring Order 66

#Now let's put this all together into one chart
figs=plt.figure(figsize=(15,18))
figs.suptitle("A closer look at the known Jedi")


gs=figs.add_gridspec(3,2)

ax1=figs.add_subplot(gs[0,0])
ax1.title.set_text('Lightsabers by Color (Count)')
df_sabers['Lightsaber Color'].value_counts()[:20].plot(kind='barh', color=['orange','blue','green', 'orange','red','purple','black'], xticks=
#plt.legend(("Count of Known Lightsabers by Color",))

ax2=figs.add_subplot(gs[0,1])
ax2.title.set_text('Lightsabers by Color (Percentage)')
```

```
s=jedi_csv['Lightsaber Color'].squeeze()
s.value_counts(normalize=True).plot.pie(autopct='%.1f %%', ylabel='', legend=False, colors=['orange','blue','green','orange','red','purple','

ax3=figs.add_subplot(gs[1,0])
ax3.title.set_text('Gender Breakdown')
jedi_csv['Gender'].value_counts().plot(kind='bar', color=['blue','red','brown'], yticks=jedi_csv['Gender'].value_counts())

ax4=figs.add_subplot(gs[1,1])
ax4.title.set_text('Date of Death')

jedi_csv['Date of Death Numerical']=None

import re

for index,row in jedi_csv.iterrows():
    if row['Date of Death'][0].isdigit():
        row['Date of Death Numerical']=re.findall(r'\d+',row['Date of Death'])
        row['Date of Death Numerical']=int(row['Date of Death Numerical'][0])
        if "BBY" in row['Date of Death']:
            row['Date of Death Numerical']=row['Date of Death Numerical']-(2*row['Date of Death Numerical'])

#SO now we should have numerical years they died. Let's try to plot that.

plt.stem(jedi_csv['Date of Death Numerical'].sort_values(), markerfmt='d')
#That looks ok, but we really need to spruce this up a little. Get rid of the stems? Maybe group like numbers, and add BBY and ABY labels to

ax5=figs.add_subplot(gs[2,:])
ax5.title.set_text('Qty of each Species')

jedi_csv['Species'].value_counts().plot(kind='bar', color='gray')
y=jedi_csv['Species'].value_counts()
x=jedi_csv['Species'].unique()


for i in range(len(x)):
    plt.text(i,y[i], y[i],ha='center')

plt.show()
```



A closer look at the known Jedi

Lightsabers by Color (Count)

Lightsabers by Color (Percentage)

Gender Breakdown

Date of Death

Qty of each Species

Observations

Lightsaber Color

While there are a lot of Jedi whose lightsaber color is unknown, of those we do know blue is the most common, with green slightly behind it in popularity. Of the known saber colors, of the color isn't blue or green then there is a high chance the Jedi has had multiple colors of saber.

Jedi Gender

We can also see that there are about twice as many male Jedi as female, at least of those whose gender we know. While the unknown number may continue at a similar rate, we don't know they will. This suggests more Jedi are men than are women overall, but it is only a suggestion and not proof. We don't know the total number of Jedi, which makes this hard to predict. Maybe we know of more male than female because these particular male Jedi happened to be involved in something well known, or maybe those who created the stories of the Jedi were biased toward males. The data is suggestive, not conclusive.

Date of Death

A majority of the known Jedi die around 19 BBY (Before the battle of Yavin). For those who know Star Wars history, that is when Order 66 was implemented. During Order 66 the Jedi were declared enemies of The Republic by Chancellor Palpatine and he ordered the clones to kill the Jedi. Additionally, many Jedi and their Padowans died at the hands of Darth Vader at this time. That explains this clumping of data. The data also shows that a small number of Jedi survived to live many years beyond Order 66.

Species

Of all the known Jedi, an overwhelming majority of them are human, according to this data. While the total number of Twi'lek, Zabrak, Togruta, and others may add up to more than the total number of humans, the numbers are so lopsided as to show a definite trend toward human Jedi. This is a limited data set, and it may not indicate the number overall, but it is suggestive.

Additional Thoughts

This was an incomplete view of the Jedi, based on a small sampling of data. While the point of this was to explore a little about the Jedi, I didn't expect any solid conclusions. And I didn't find any.

There is more that can be done with this data, including showing the master-apprentice relationships, but I'm happy with what I've found here.

✓ 3s completed at 1:30 PM ● ✕