NASA Astronauts Dataset Analysis

Frantz Alexander

November 27, 2022

Project Introduction

The dataset used was obtained from the NASA official website and published as the Astronaut April 2013 fact book.

The goal of the project was to acquire key insights of astonauts in the human space flight program at NASA from 1959 to 2013.

Libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_style('ticks')

%matplotlib inline
```

Data Preparation

Create Data Wrangling Function

In [2]:
```
def wrangle(data_set):
    """
        This function is meant to import and wrangle the dataset.
    """
    df = pd.read_csv(data_set)

    return df
```

Import CSV File

In [3]:
```
df = wrangle("astronauts.csv")
```

Exploring the Characteristics of the Dataset.

How Many Rows and Columns are in the Dataset?

In [4]:
```
rows = df.shape[0]
columns = df.shape[1]

print("This dataset contains: {} rows and {} columns".format(rows,columns))
```

```
This dataset contains: 357 rows and 19 columns
```

What are the Column Names?

In [5]:
```
for col in df.columns:
    print(col)
```

```
Name
Year
Group
Status
Birth Date
Birth Place
Gender
Alma Mater
Undergraduate Major
Graduate Major
Military Rank
Military Branch
Space Flights
Space Flight (hr)
Space Walks
Space Walks (hr)
Missions
Death Date
Death Mission
```

What are the Column Data Types?

In [6]:
```
df.dtypes
```

Out[6]:
```
Name                    object
Year                   float64
Group                  float64
Status                  object
Birth Date              object
Birth Place             object
Gender                  object
Alma Mater              object
Undergraduate Major     object
Graduate Major          object
Military Rank           object
Military Branch         object
Space Flights            int64
Space Flight (hr)        int64
Space Walks              int64
Space Walks (hr)       float64
Missions                object
Death Date              object
Death Mission           object
dtype: object
```

What are the Characteristics of each Column Feature?

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 357 entries, 0 to 356
Data columns (total 19 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Name                 357 non-null    object 
 1   Year                 330 non-null    float64
 2   Group                330 non-null    float64
 3   Status               357 non-null    object 
 4   Birth Date           357 non-null    object 
 5   Birth Place          357 non-null    object 
 6   Gender               357 non-null    object 
 7   Alma Mater           356 non-null    object 
 8   Undergraduate Major  335 non-null    object 
 9   Graduate Major       298 non-null    object 
 10  Military Rank        207 non-null    object 
 11  Military Branch      211 non-null    object 
 12  Space Flights        357 non-null    int64  
 13  Space Flight (hr)    357 non-null    int64  
 14  Space Walks          357 non-null    int64  
 15  Space Walks (hr)     357 non-null    float64
 16  Missions             334 non-null    object 
 17  Death Date           52 non-null     object 
 18  Death Mission        16 non-null     object 
dtypes: float64(3), int64(3), object(13)
memory usage: 53.1+ KB
```

Displaying the Number of Unique Values in each Column

In [8]:
```
df.nunique()
```

Out[8]:
```
Name                   357
Year                    20
Group                   20
Status                   4
Birth Date             348
Birth Place            272
Gender                   2
Alma Mater             280
Undergraduate Major     83
Graduate Major         143
Military Rank           12
Military Branch         14
Space Flights            8
Space Flight (hr)      270
Space Walks             11
Space Walks (hr)        52
Missions               305
Death Date              38
Death Mission            3
dtype: int64
```

Displaying the Number of Null Values in each Column

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
Name                     0
Year                    27
Group                   27
Status                   0
Birth Date               0
Birth Place              0
Gender                   0
Alma Mater               1
Undergraduate Major     22
Graduate Major          59
Military Rank          150
Military Branch        146
Space Flights            0
Space Flight (hr)        0
Space Walks              0
Space Walks (hr)         0
Missions                23
Death Date             305
Death Mission          341
dtype: int64
```

Display of the Percentage of Missing Values in each Column

In [10]:
```
for col in df.columns:
    percentage_missing = np.mean(df[col].isnull())
    print("{} - {}%".format(col, round(percentage_missing*100)))
```

```
Name - 0%
Year - 8%
Group - 8%
Status - 0%
Birth Date - 0%
Birth Place - 0%
Gender - 0%
Alma Mater - 0%
Undergraduate Major - 6%
Graduate Major - 17%
Military Rank - 42%
Military Branch - 41%
Space Flights - 0%
Space Flight (hr) - 0%
Space Walks - 0%
Space Walks (hr) - 0%
Missions - 6%
Death Date - 85%
Death Mission - 96%
```

Display of the First 5 Rows of each Column

In [11]:
```
df.head()
```

Out[11]:

	Name	Year	Group	Status	Birth Date	Birth Place	Gender	Alma Mater	Undergraduate Major	
0	Joseph M. Acaba	2004.0	19.0	Active	5/17/1967	Inglewood, CA	Male	University of California-Santa Barbara; Univer...	Geology	
1	Loren W. Acton	NaN	NaN	Retired	3/7/1936	Lewiston, MT	Male	Montana State University; University of Colorado	Engineering Physics	So
2	James C. Adamson	1984.0	10.0	Retired	3/3/1946	Warsaw, NY	Male	US Military Academy; Princeton University	Engineering	Er
3	Thomas D. Akers	1987.0	12.0	Retired	5/20/1951	St. Louis, MO	Male	University of Missouri-Rolla	Applied Mathematics	Ma
4	Buzz Aldrin	1963.0	3.0	Retired	1/20/1930	Montclair, NJ	Male	US Military Academy; MIT	Mechanical Engineering	As

The Descriptive Statstics of the Numerical Column Features

In [12]:
```
df.describe()
```

Out[12]:

	Year	Group	Space Flights	Space Flight (hr)	Space Walks	Space Walks (hr)
count	330.000000	330.000000	357.000000	357.000000	357.000000	357.000000
mean	1985.106061	11.409091	2.364146	1249.266106	1.246499	7.707283
std	13.216147	5.149962	1.428700	1896.759857	2.056989	13.367973
min	1959.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1978.000000	8.000000	1.000000	289.000000	0.000000	0.000000
50%	1987.000000	12.000000	2.000000	590.000000	0.000000	0.000000
75%	1996.000000	16.000000	3.000000	1045.000000	2.000000	12.000000
max	2009.000000	20.000000	7.000000	12818.000000	10.000000	67.000000

Mission Status

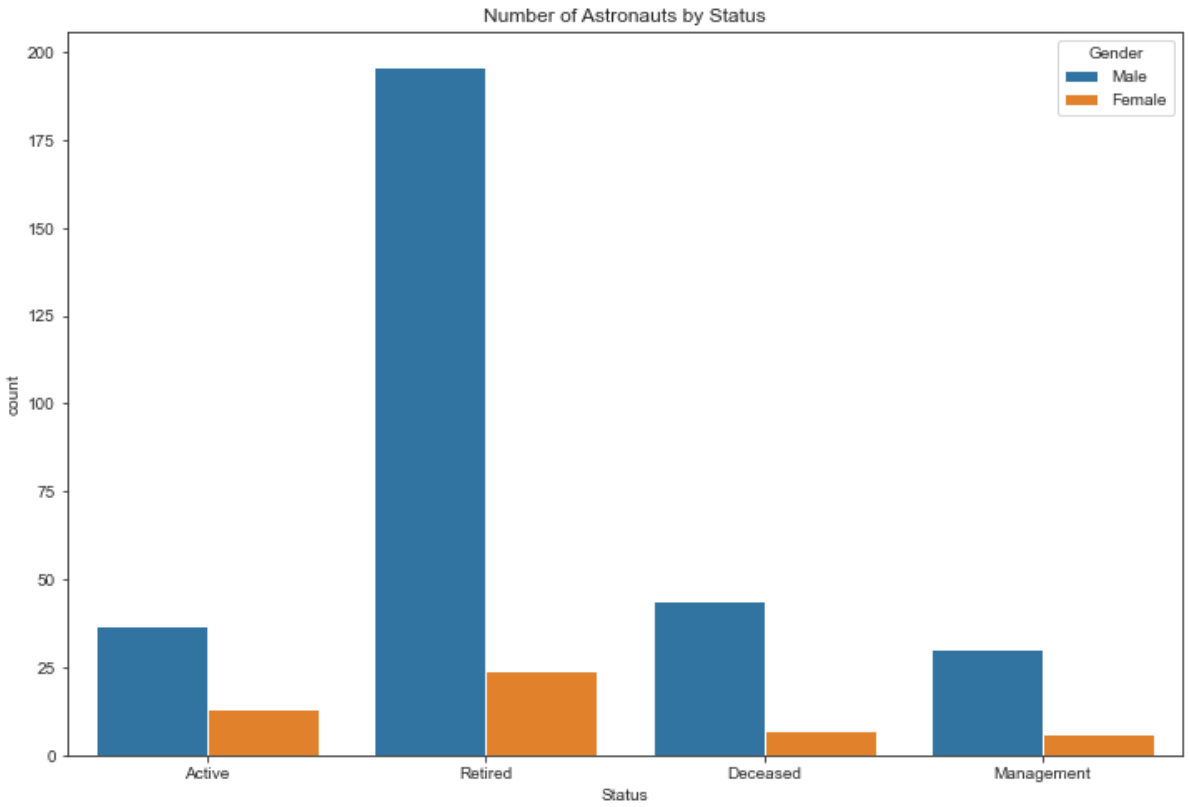
What is the Status of the Astronauts?

In [13]:
```
df["Status"].value_counts()
```

Out[13]:
```
Retired       220
Deceased       51
Active         50
Management     36
Name: Status, dtype: int64
```

Visualizing the Number of Astronauts by Status

In [14]:
```
plt.figure(figsize = (12,8))
sns.countplot(
    x = "Status",
    data = df,
    hue = "Gender"
)
plt.title("Number of Astronauts by Status");
```



Display of Astronaut Status Based on Gender

In [15]:
```
df.groupby("Gender")["Status"].value_counts()
```

Out[15]:
```
Gender  Status
Female  Retired        24
        Active         13
        Deceased        7
        Management      6
Male    Retired       196
        Deceased       44
        Active         37
        Management     30
Name: Status, dtype: int64
```

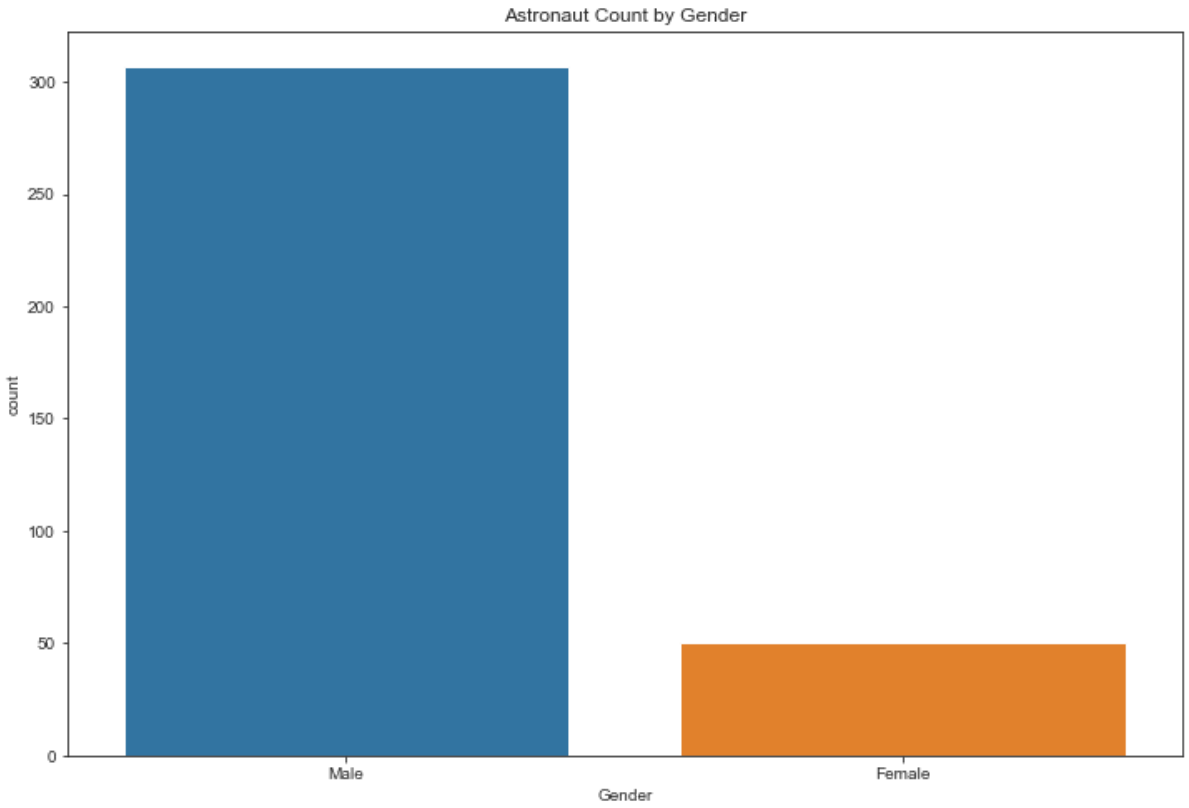
How Many Astronauts were of each Gender?

In [16]:
```
df["Gender"].value_counts()
```

Out[16]:
```
Male      307
Female     50
Name: Gender, dtype: int64
```

Visualizing the Count of Astronauts by Gender

In [17]:
```
plt.figure(figsize = (12,8))
sns.countplot(
    x = "Gender",
    data = df
)
plt.title("Astronaut Count by Gender");
```



Subset for Active Astronauts

In [18]:
```
active_mask = df["Status"] == "Active"
active_astronauts = df[active_mask]
```

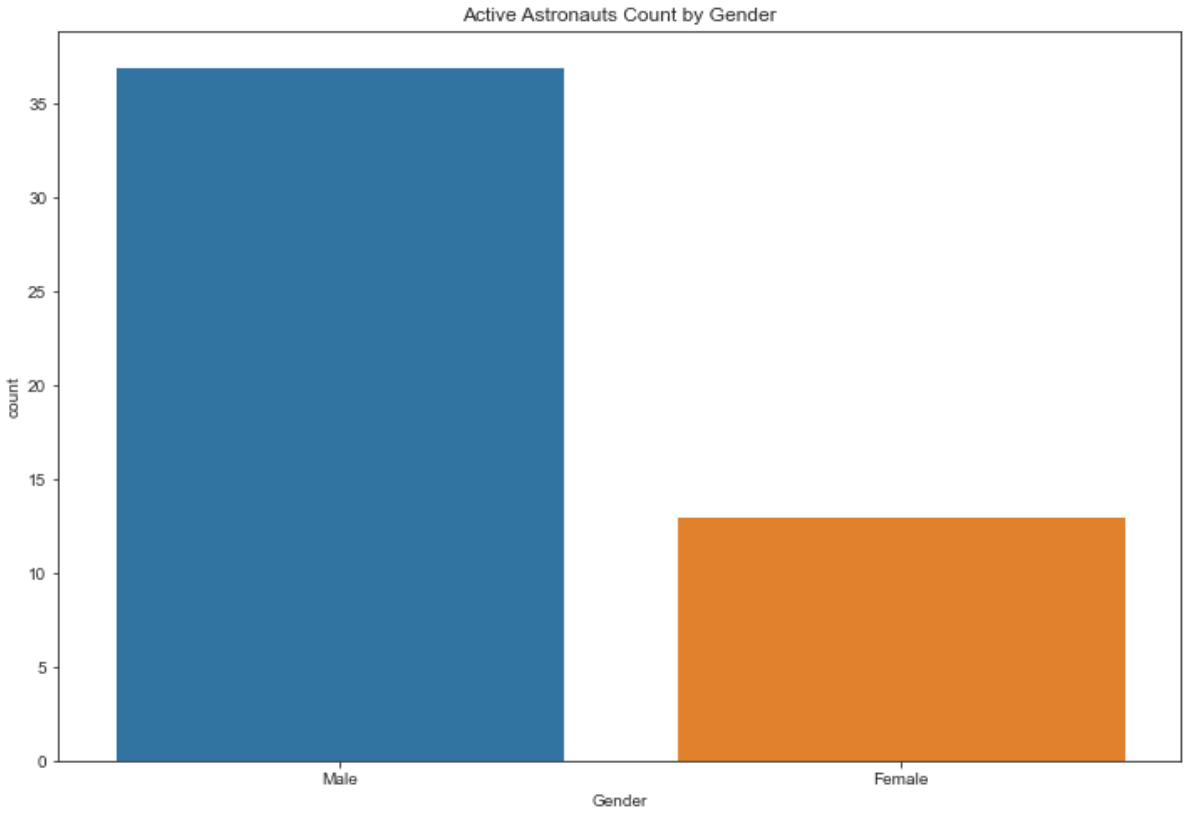
How Many Active Astronauts are of each Gender?

In [19]:
```
active_astronauts["Gender"].value_counts()
```

Out[19]:
```
Male      37
Female    13
Name: Gender, dtype: int64
```

Visualizing the Active Astronauts Count by Gender

In [20]:
```
plt.figure(figsize = (12,8))
sns.countplot(
    x = "Gender",
    data = active_astronauts
)
plt.title("Active Astronauts Count by Gender");
```



Mission Accidents

Create Mission Accidents Subset

In [21]:
```
accidents_mask = ~df["Death Mission"].isnull()
mission_accidents = df[accidents_mask]
```

Which Missions had Accidents?

In [22]:
```
for mission in mission_accidents["Death Mission"].unique():
        print(mission)
```

```
STS-107 (Columbia)
Apollo 1
STS 51-L (Challenger)
```

How Many Men and Women Passed Away from an Accident?

In [23]:
```
male, female = mission_accidents["Gender"].value_counts()

print("There were {} men and {} women that passed away from \
an accident in space flight.".format(male, female))
```

```
There were 12 men and 4 women that passed away from an accident in space flight.
```

How Many People Passed Away in a Space Flight Accident?

In [24]:
```
people_accidents = len(mission_accidents)

print("There were {} astronauts who passed away in space \
flight accidents throughout NASA's history.".format(people_accidents))
```

```
There were 16 astronauts who passed away in space flight accidents throughout NASA'
s history.
```

There were 3 missions that had accidents which resulted in the deaths of 16 astronauts:

- STS-107 (Columbia)
- Apollo 1
- STS 51-L (Challenger)

Who were the Astronauts who Passed Away in an Accident?

In [25]:
```
print("These brave men and women who have passed away\
have worked to make our world a better place:")

for name in mission_accidents["Name"]:
    print(name)
```

```
These brave men and women who have passed awayhave worked to make our world a bette
r place:
Michael P. Anderson
David M. Brown
Roger B. Chaffee
Kalpana Chawla
Laurel B. Clark
Virgil I. Grissom
Rick D. Husband
Gregory B. Jarvis
S. Christa McAuliffe
William C. McCool
Ronald E. McNair
Ellison S. Onizuka
Judith A. Resnik
Francis R. Scobee
Michael J. Smith
Edward H. White II
```

Mission Accidents by Gender

In [26]:
```
pd.crosstab(
    index = mission_accidents["Gender"],
    columns = mission_accidents["Death Mission"]
)
```

Out[26]:

Death Mission	Apollo 1	STS 51-L (Challenger)	STS-107 (Columbia)
Gender			
Female	0	2	2
Male	3	5	4

Military Ranks Value Counts

In [27]:
```
mission_accidents["Military Rank"].value_counts()
```

Out[27]:
```
Lieutenant Colonel      4
Captain                 3
Lieutenant Commander    1
Colonel                 1
Commander               1
Major                   1
Name: Military Rank, dtype: int64
```

Mission Accidents by Military Branch

In [28]:
```
pd.crosstab(
    index = mission_accidents["Military Branch"],
    columns = mission_accidents["Death Mission"]
)
```

Out[28]:

Death Mission	Apollo 1	STS 51-L (Challenger)	STS-107 (Columbia)
Military Branch			
US Air Force	2	1	2
US Air Force (Retired)	0	1	0
US Navy	1	1	3

Military Rank

The List of Military Ranks

In [29]:
```
for rank in df["Military Rank"].unique():
    print(rank)
```

```
nan
Colonel
Lieutenant Colonel
Captain
Major General
Commander
Lieutenant Commander
Brigadier General
Major
Lieutenant General
Chief Warrant Officer
Rear Admiral
Vice Admiral
```

The List of Military Ranks for Active Astronauts

In [30]:
```
for rank in active_astronauts["Military Rank"].unique():
    print(rank)
```

```
nan
Commander
Colonel
Captain
```

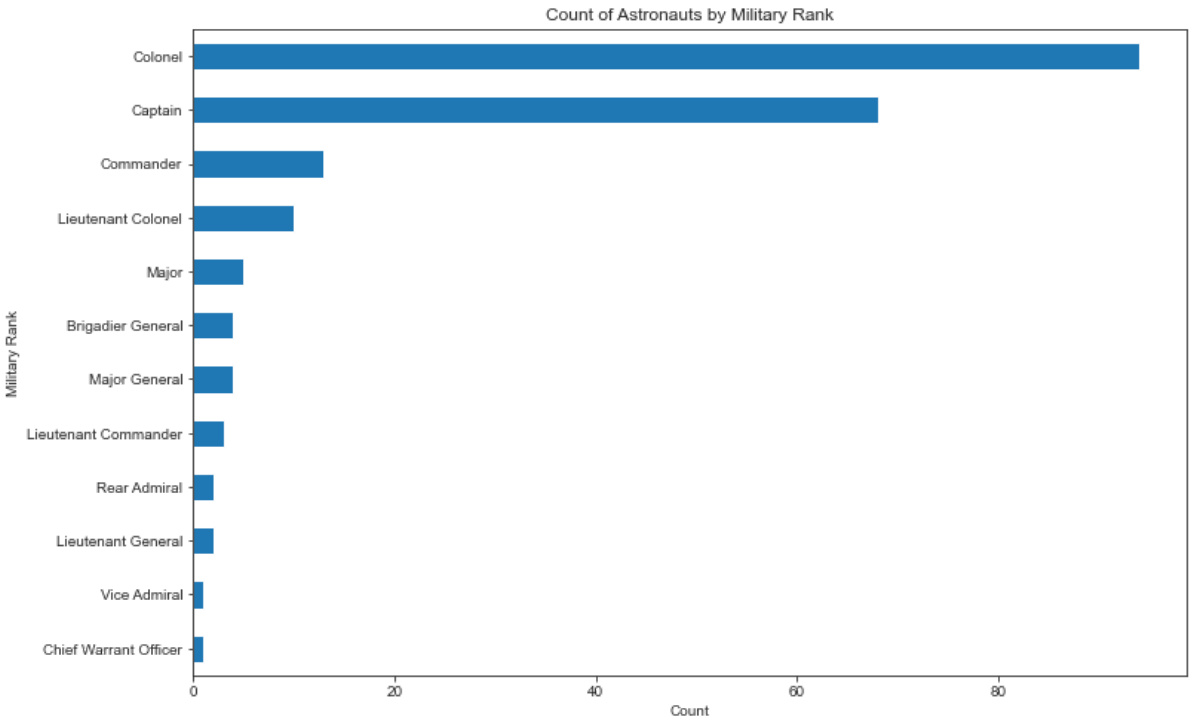
What are the Most Common Military Rank Among Astronauts?

In [31]:
```
df["Military Rank"].value_counts()
```

Out[31]:
```
Colonel                 94
Captain                 68
Commander               13
Lieutenant Colonel      10
Major                    5
Major General            4
Brigadier General        4
Lieutenant Commander     3
Lieutenant General       2
Rear Admiral             2
Chief Warrant Officer    1
Vice Admiral             1
Name: Military Rank, dtype: int64
```

Visualizing the count of Astronauts by Military Rank

In [32]:
```
plt.figure(figsize = (12,8))
(df["Military Rank"]
     .value_counts()
     .sort_values()
     .plot(kind = "barh")
)
plt.xlabel("Count")
plt.ylabel("Military Rank")
plt.title("Count of Astronauts by Military Rank");
```



Who was the Highest Ranking Military Service Member who Participated in Space Flight?

In [33]:
```
vice_admiral = df[df["Military Rank"] == "Vice Admiral"]

for name in vice_admiral["Name"]:
    print("Vice Admiral {}, was the highest ranking Military\
    service member to participate in space flight.".format(name))
```

```
Vice Admiral Richard H. Truly, was the highest ranking Military    service member t
o participate in space flight.
```

What are the Most Commmon Military Ranks Among Active Astronauts?

In [34]:
```
active_astronauts["Military Rank"].value_counts()
```

Out[34]:
```
Colonel      17
Captain       5
Commander     4
Name: Military Rank, dtype: int64
```

Astronaut Military Rank Breakdown by Gender

In [35]:
```
pd.crosstab(
    index = df["Military Rank"],
    columns = df["Gender"]
)
```

Out[35]:

Gender	Female	Male
Military Rank		
Brigadier General	0	4
Captain	6	62
Chief Warrant Officer	0	1
Colonel	5	89
Commander	1	12
Lieutenant Colonel	0	10
Lieutenant Commander	0	3
Lieutenant General	1	1
Major	0	5
Major General	0	4
Rear Admiral	0	2
Vice Admiral	0	1

Breakdown of Active Astronauts Military Ranks by Gender

In [36]:
```
pd.crosstab(
    index = active_astronauts["Military Rank"],
    columns = active_astronauts["Gender"]
)
```

Out[36]:

Gender	Female	Male
Military Rank		
Captain	1	4
Colonel	1	16
Commander	0	4

How Many Astronauts did not Serve in the Military?

In [37]:
```
df["Military Rank"].isnull().sum()
```

Out[37]:
```
150
```

Military Branch

Breakdown of Military Branch by Gender?

In [38]:
```
pd.crosstab(
    index = df["Military Branch"],
    columns = df["Gender"]
)
```

Out[38]:

Gender	Female	Male
Military Branch		
US Air Force	2	19
US Air Force (Retired)	3	58
US Air Force Reserves	0	2
US Air Force Reserves (Retired)	0	3
US Army	0	4
US Army (Retired)	1	12
US Coast Guard (Retired)	0	2
US Marine Corps	0	3
US Marine Corps (Retired)	0	17
US Marine Corps Reserves	0	2
US Naval Reserves	1	1
US Naval Reserves (Retired)	0	1
US Navy	3	18
US Navy (Retired)	3	56

Space Flight statistics for each Military Branch by Gender

In [39]:
```
pd.crosstab(
    index = df["Military Branch"],
    columns = df["Gender"],
    values = df["Space Flights"],
    aggfunc = ["mean", "std", "max"]
)
```

Out[39]:

	mean		std		max	
Gender	Female	Male	Female	Male	Female	Male
Military Branch						
US Air Force	2.500000	1.631579	3.535534	1.011628	5.0	4.0
US Air Force (Retired)	3.333333	2.775862	0.577350	1.351323	4.0	7.0
US Air Force Reserves	NaN	1.500000	NaN	0.707107	NaN	2.0
US Air Force Reserves (Retired)	NaN	2.000000	NaN	1.000000	NaN	3.0
US Army	NaN	2.000000	NaN	1.414214	NaN	3.0
US Army (Retired)	4.000000	2.416667	NaN	1.378954	4.0	5.0
US Coast Guard (Retired)	NaN	2.500000	NaN	0.707107	NaN	3.0
US Marine Corps	NaN	1.000000	NaN	1.000000	NaN	2.0
US Marine Corps (Retired)	NaN	2.588235	NaN	1.175735	NaN	4.0
US Marine Corps Reserves	NaN	2.500000	NaN	2.121320	NaN	4.0
US Naval Reserves	2.000000	5.000000	NaN	NaN	2.0	5.0
US Naval Reserves (Retired)	NaN	3.000000	NaN	NaN	NaN	3.0
US Navy	1.666667	1.277778	0.577350	0.894792	2.0	3.0
US Navy (Retired)	2.333333	2.875000	1.527525	1.322016	4.0	6.0

What Ranks did Astronauts Achieve within each Military Branch?

In [40]:
```
pd.crosstab(
    index = df["Military Branch"],
    columns = df["Military Rank"]
)
```

Out[40]:

Military Rank	Brigadier General	Captain	Chief Warrant Officer	Colonel	Commander	Lieutenant Colonel	Lieutenant Commander	Lieutenant General
Military Branch								
US Air Force	0	2	0	11	0	5	0	1
US Air Force (Retired)	3	0	0	53	0	1	0	1
US Air Force Reserves	0	0	0	0	0	0	0	0
US Air Force Reserves (Retired)	0	0	0	1	0	0	0	0
US Army	0	0	0	3	0	1	0	0
US Army (Retired)	1	1	1	9	0	1	0	0
US Coast Guard (Retired)	0	1	0	0	1	0	0	0
US Marine Corps	0	0	0	2	0	0	0	0
US Marine Corps (Retired)	0	0	0	14	0	2	0	0
US Marine Corps Reserves	0	0	0	1	0	0	0	0
US Naval Reserves	0	1	0	0	0	0	0	0
US Naval Reserves (Retired)	0	1	0	0	0	0	0	0
US Navy	0	12	0	0	7	0	2	0
US Navy (Retired)	0	50	0	0	5	0	1	0

Education

Education Subset Function

In [41]:
```
def education_wrangle(dataframe):

    # Create copy of dataframe
    edu_df = dataframe.copy()

    # Add "Undergraduate Alma Mater" Column
    edu_df["Undergraduate Alma Mater"] = (
        edu_df["Alma Mater"].str.split(";", expand = True)[0]
    )

    # Add "Graduate Alma Mater" Column
    edu_df["Graduate Alma Mater"] = (
        edu_df["Alma Mater"].str.split(";", expand = True)[1]
    )

    # Add "Post-Graduate Alma Mater" Column
    edu_df["Post-Graduate Alma Mater"] = (
        edu_df["Alma Mater"].str.split(";", expand = True)[2]
    )

    # Drop old "ALma Mater" Column
    edu_df.drop(columns = ["Alma Mater"], inplace = True)

    return edu_df
```

In [42]:
```
education_df = education_wrangle(df)
```

List of Columns in the Education Dataframe

In [43]:
```
for col in education_df.columns:
    print(col)
```

```
Name
Year
Group
Status
Birth Date
Birth Place
Gender
Undergraduate Major
Graduate Major
Military Rank
Military Branch
Space Flights
Space Flight (hr)
Space Walks
Space Walks (hr)
Missions
Death Date
Death Mission
Undergraduate Alma Mater
Graduate Alma Mater
Post-Graduate Alma Mater
```

How Many Astronauts did not have a Graduate Degree?

In [44]:
```
graduate_degree = education_df["Graduate Major"].isnull().sum()

print("There were {} astronauts that did \
not have a graduate degree.".format(graduate_degree))
```

```
There were 59 astronauts that did not have a graduate degree.
```

What Percentage of Astronauts did not have a Graduate Degree?

In [45]:
```
graduate_percentage = (
    np.mean(education_df["Graduate Major"].isnull())
)

print("{}% of Astronauts did not\
have a Graduate Degree.".format(round(graduate_percentage*100)))
```

```
17% of Astronauts did nothave a Graduate Degree.
```

How Many Different Types of Majors did Astonauts have?

In [46]:
```
num_majors = education_df["Undergraduate Major"].nunique()

print("Astronauts had {} unique undergraduate majors.".format(num_majors))
```

```
Astronauts had 83 unique undergraduate majors.
```

Top 10 Undergraduate Majors for Astronauts

In [47]:
```
education_df["Undergraduate Major"].value_counts().head(10)
```

Out[47]:
```
Physics                      35
Aerospace Engineering        33
Mechanical Engineering       30
Aeronautical Engineering     28
Electrical Engineering       23
Engineering Science          13
Engineering                  12
Mathematics                  11
Chemistry                    10
Chemical Engineering          9
Name: Undergraduate Major, dtype: int64
```

Top 10 Universities for Undergraduate Studies

In [48]:
```
education_df["Undergraduate Alma Mater"].value_counts().head(10)
```

Out[48]:
```
US Naval Academy                    52
US Air Force Academy                38
US Military Academy                 18
Purdue University                   15
MIT                                 12
University of Colorado               8
Stanford University                  7
University of Texas                  6
University of Illinois               5
University of California-Berkeley    5
Name: Undergraduate Alma Mater, dtype: int64
```

Top 10 Universities for Graduate Studies.

In [49]:
```
education_df["Graduate Alma Mater"].value_counts().head(10)
```

Out[49]:
```
US Naval Postgraduate School          29
MIT                                   22
Stanford University                   14
Georgia Institute of Technology       10
California Institute of Technology     8
Purdue University                      8
University of Southern California      7
University of Tennessee                6
University of Colorado                 6
George Washington University           5
Name: Graduate Alma Mater, dtype: int64
```

Top 10 Universities for Post-Graduate Studies

In [50]:
```
education_df["Post-Graduate Alma Mater"].value_counts().head(10)
```

Out[50]:
```
University of Houston-Clear Lake        5
University of Florida                   3
University of Texas                     2
US Naval War College                    2
Rice University                         2
Harvard University                      2
University of California-Los Angeles    2
University of Washington                2
University of Houston                   2
US Naval Postgraduate School            2
Name: Post-Graduate Alma Mater, dtype: int64
```

Space Flight

Subset for Space Flight Hours

In [51]:
```
space_flight_mask = df["Space Flight (hr)"] > 0
space_flights = df[space_flight_mask]
```

What was the Greatest Amount of Time in Space?

In [52]:
```
space_flight_max = space_flights["Space Flight (hr)"].max()

print(
    "The highest amount of time in space \
was {} hours.".format(space_flight_max)
)
```

```
The highest amount of time in space was 12818 hours.
```

Who was the Astronaut that had Spent the Most Amount of Time in Space?

In [53]:
```
for name in space_flights[space_flights["Space Flight (hr)"] == 12818]["Name"]:
    print(name + ", is the astronaut that spent the most time in space flight.")
```

```
Jeffrey N. Williams, is the astronaut that spent the most time in space flight.
```

Which Active Astronaut had the Most Space Flights?

In [54]:
```
active_space_flights = (
    space_flights[space_flights["Status"] == "Active"]
)
```

In [55]:
```
active_space_flights["Space Flights"].max()
```

Out[55]:
```
6
```

In [56]:
```
most_space_flights = (
    active_space_flights[active_space_flights["Space Flights"] == 6]
)

for name in most_space_flights["Name"]:
    print(name +", is the astronaut that performed the most space flights.")
```

```
C. Michael Foale, is the astronaut that performed the most space flights.
```

Which Astronaut has Performed the Most Space Flights?

In [57]:
```
df["Space Flights"].max()
```

Out[57]:
```
7
```

In [58]:
```
most_flights_mask = df["Space Flights"] == 7

print("There were {} astronauts that participated \
in the most space flights:".format(len(df[most_flights_mask])))

for name in df[most_flights_mask]["Name"]:
    print(name)
```

```
There were 2 astronauts that participated in the most space flights:
Franklin R. Chang-Diaz
Jerry L. Ross
```

Which Female Astronauts had the Most Space Flights?

In [59]:
```
df[df["Gender"] == "Female"]["Space Flights"].max()
```

Out[59]:
```
5
```

In [60]:
```
female_flights = df[
    (df["Gender"] ==  "Female") & (df["Space Flights"] == 5)
]

print("These are the Female astronauts \
who participated in the most space flights:")

for name in female_flights["Name"]:
    print(name)
```

```
These are the Female astronauts who participated in the most space flights:
Bonnie J. Dunbar
Susan J. Helms
Marsha S. Ivins
Tamara E. Jernigan
Shannon W. Lucid
Janice E. Voss
```

Which Active Female Astronauts had the Most Space Flights?

In [61]:
```
(
    active_space_flights
    [active_space_flights["Gender"] == "Female"]
    ["Space Flights"]
    .max()
)
```

Out[61]:
```
3
```

In [62]:
```
most_active_female_space_flights = active_space_flights[
    (active_space_flights["Gender"] ==  "Female")
    & (active_space_flights["Space Flights"] ==  3)
]

print("This is a listing of the current active \
female astronauts with the most space flights:")

for name in most_active_female_space_flights["Name"]:
    print(name)
```

```
This is a listing of the current active female astronauts with the most space fligh
ts:
Catherine G. Coleman
Peggy A. Whitson
Stephanie D. Wilson
```

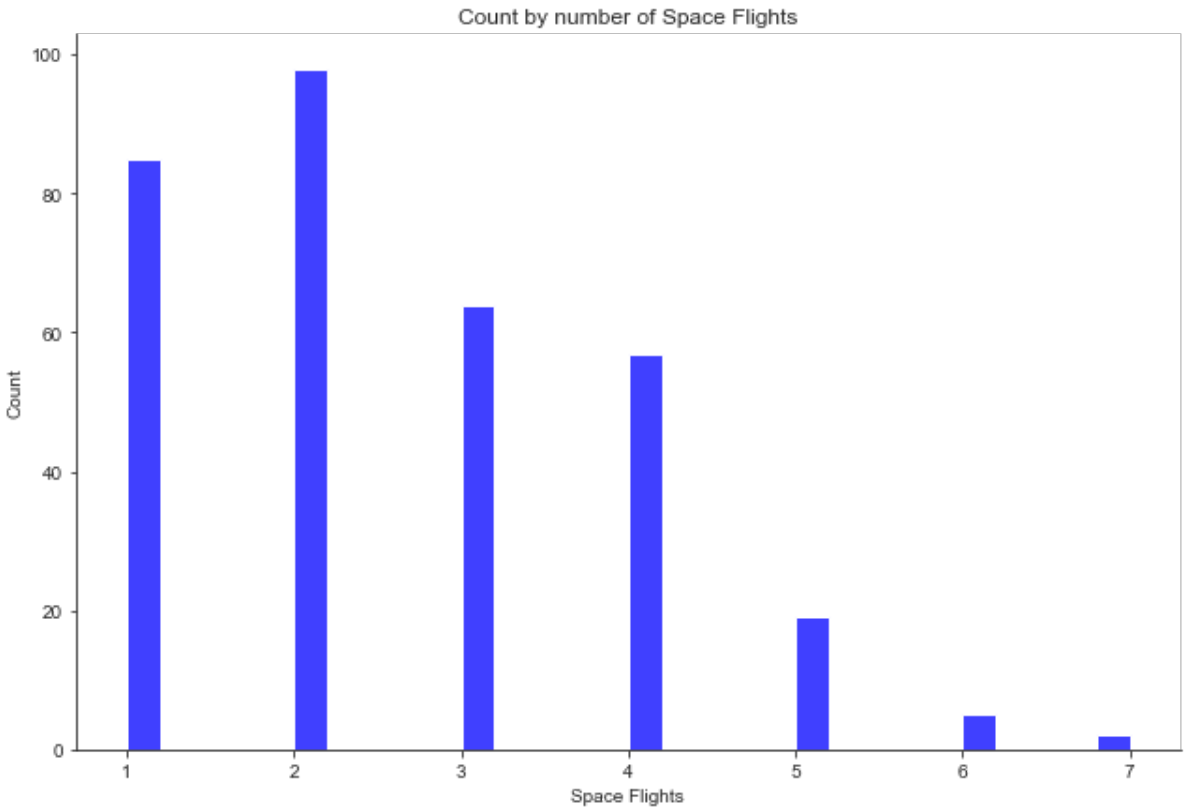
What is the Count of Space Flights Performed?

In [63]:
```
space_flights["Space Flights"].value_counts(ascending = False)
```

Out[63]:
```
2    98
1    85
3    64
4    57
5    19
6     5
7     2
Name: Space Flights, dtype: int64
```

Visualizing the Count of Space flights

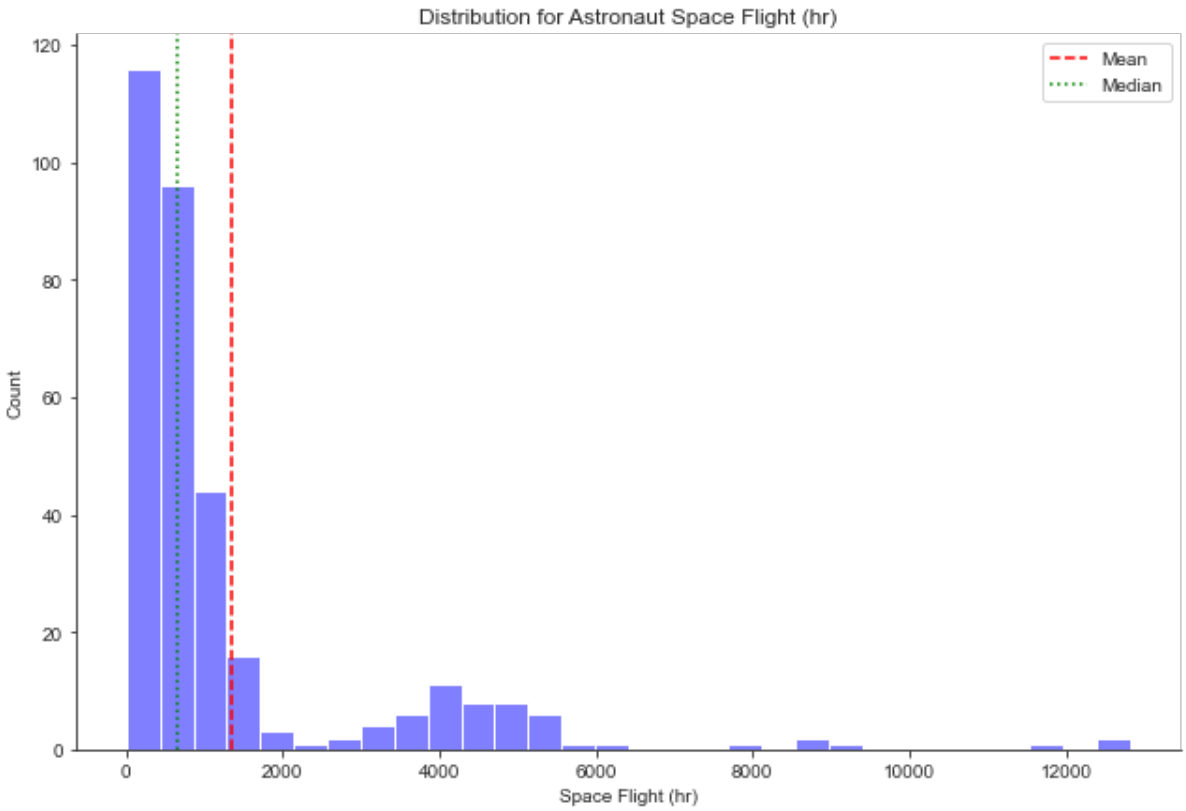
In [64]:
```
sns.displot(
    space_flights["Space Flights"],
    bins = 30,
    color = "blue",
    height = 6,
    aspect = 1.5
)

plt.title("Count by number of Space Flights");
```



Visualizing the Distribution of Astronaut Space Flight Hours

In [65]:
```
space_flight_hours = space_flights["Space Flight (hr)"]
```

In [66]:
```
sns.displot(
    space_flight_hours,
    bins = 30,
    color = "blue",
    height = 6,
    aspect = 1.5,
    alpha = 0.5
)
plt.axvline(
    np.mean(space_flight_hours),
    ls = "--",
    label = "Mean",
    color = "red"
)
plt.axvline(
    np.median(space_flight_hours),
    ls = ":",
    label = "Median",
    color = "green"
)
plt.legend()
plt.title("Distribution for Astronaut Space Flight (hr)");
```



Breakdown of Space Flights by Military Rank

In [67]:
```
pd.crosstab(
    index = space_flights["Military Rank"],
    columns = space_flights["Space Flights"]
)
```

Out[67]:

Military Rank \ Space Flights	1	2	3	4	5	6	7
Brigadier General	1	2	1	0	0	0	0
Captain	13	17	15	15	3	2	0
Chief Warrant Officer	1	0	0	0	0	0	0
Colonel	14	32	19	20	4	1	1
Commander	7	3	1	0	0	0	0
Lieutenant Colonel	2	5	3	0	0	0	0
Lieutenant Commander	0	0	1	0	0	0	0
Lieutenant General	0	0	0	1	1	0	0
Major	2	1	0	0	0	0	0
Major General	2	1	0	1	0	0	0
Rear Admiral	0	1	1	0	0	0	0
Vice Admiral	0	1	0	0	0	0	0

Number of Space Flights by Gender

In [68]:
```
pd.crosstab(
    index = space_flights["Gender"],
    columns = space_flights["Space Flights"]
)
```

Out[68]:

Space Flights	1	2	3	4	5	6	7
Gender							
Female	11	12	10	6	6	0	0
Male	74	86	54	51	13	5	2

Pivot Table for Space Flights based on Gender

In [69]:
```
pd.pivot_table(
    data = space_flights,
    index = "Gender",
    values = ["Space Flights", "Space Flight (hr)"],
    aggfunc = ["mean","median", "std"]
)
```

Out[69]:

	mean		median		std	
	Space Flight (hr)	Space Flights	Space Flight (hr)	Space Flights	Space Flight (hr)	Space Flights
Gender						
Female	1752.555556	2.644444	890	2	2305.561296	1.351019
Male	1288.150877	2.529825	614	2	1869.817804	1.325530

Space Walks

Subset for Astronauts that have Taken Space Walks

In [70]:
```
space_walks_mask = (
    df["Space Walks (hr)"] > 0) & (df["Space Flights"] > 0
)
space_walks = df[space_walks_mask]
```

What was the Highest Number of Space Walks Performed?

In [71]:
```
space_walks["Space Walks"].max()
```

Out[71]:
```
10
```

Which Astronaut Performed the Most Space Walks?

In [72]:
```
most_space_walks = space_walks[space_walks["Space Walks"] == 10]

for name in most_space_walks["Name"]:
    print(name + ", performed the greatest number of \
    space walks at {}.".format(space_walks["Space Walks"].max()))
```

```
Michael E. Lopez-Alegria, performed the greatest number of     space walks at 10.
```

What was the most amount of time Spent During Space Walks?

In [73]:
```
space_walks["Space Walks (hr)"].max()
```

Out[73]:
```
67.0
```

Which Astronaut Spent the Most Time Performing Space Walks

In [74]:
```
most_time_space_walks = space_walks[space_walks["Space Walks (hr)"] == 67]

for name in most_time_space_walks["Name"]:
    print(name + ", was the astronaut that spent \
the most time during space walks.")
```

```
Michael E. Lopez-Alegria, was the astronaut that spent the most time during space w
alks.
```

What is the Average Amount of Time Spent Performing Space Walks?

In [75]:
```
round(space_walks["Space Walks (hr)"].mean(),2)
```

Out[75]:
```
20.38
```

Pivot Table for Space Walks and Space Walk (hr) based on Gender

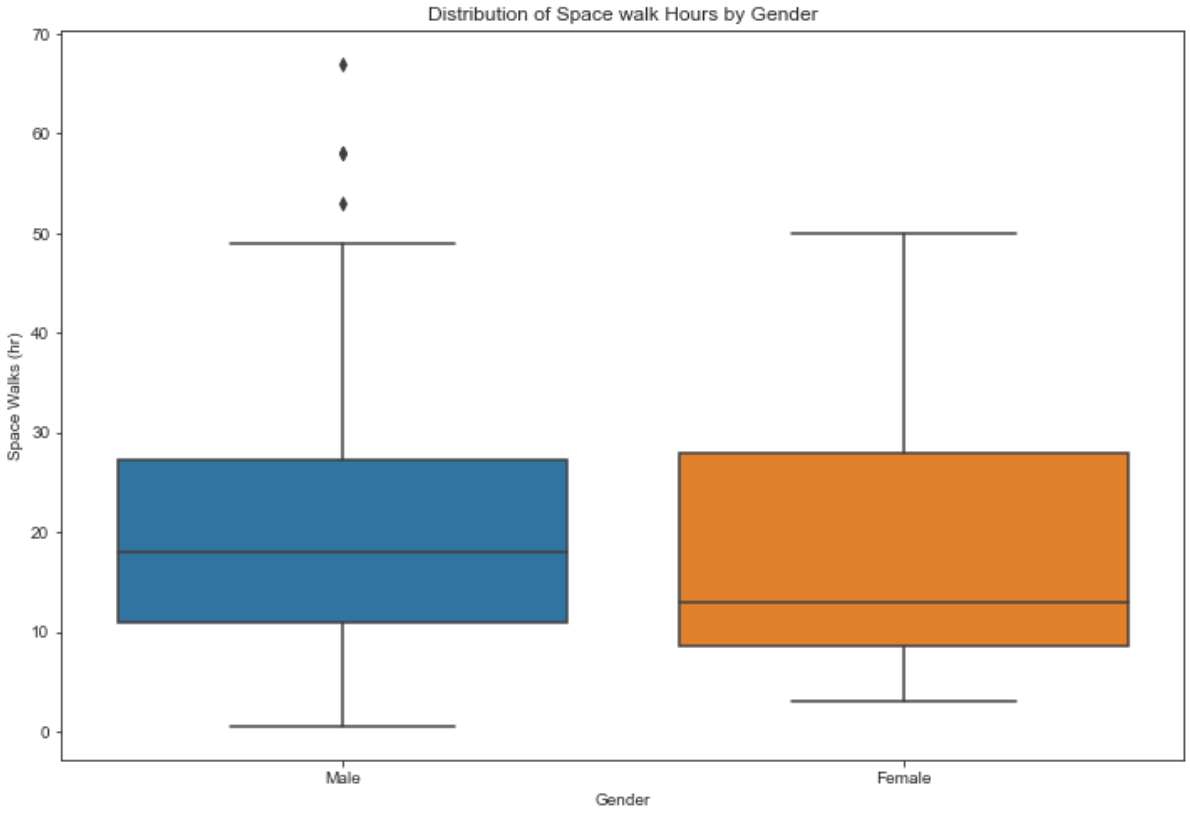
In [76]:
```
pd.pivot_table(
    data = space_walks,
    index = "Gender",
    values = ["Space Walks", "Space Walks (hr)"],
    aggfunc = ["mean","median", "std"]
)
```

Out[76]:

	mean		median		std	
	Space Walks	Space Walks (hr)	Space Walks	Space Walks (hr)	Space Walks	Space Walks (hr)
Gender						
Female	2.727273	20.181818	2	13.0	2.240130	16.289986
Male	3.346774	20.399194	3	18.0	2.095463	14.564404

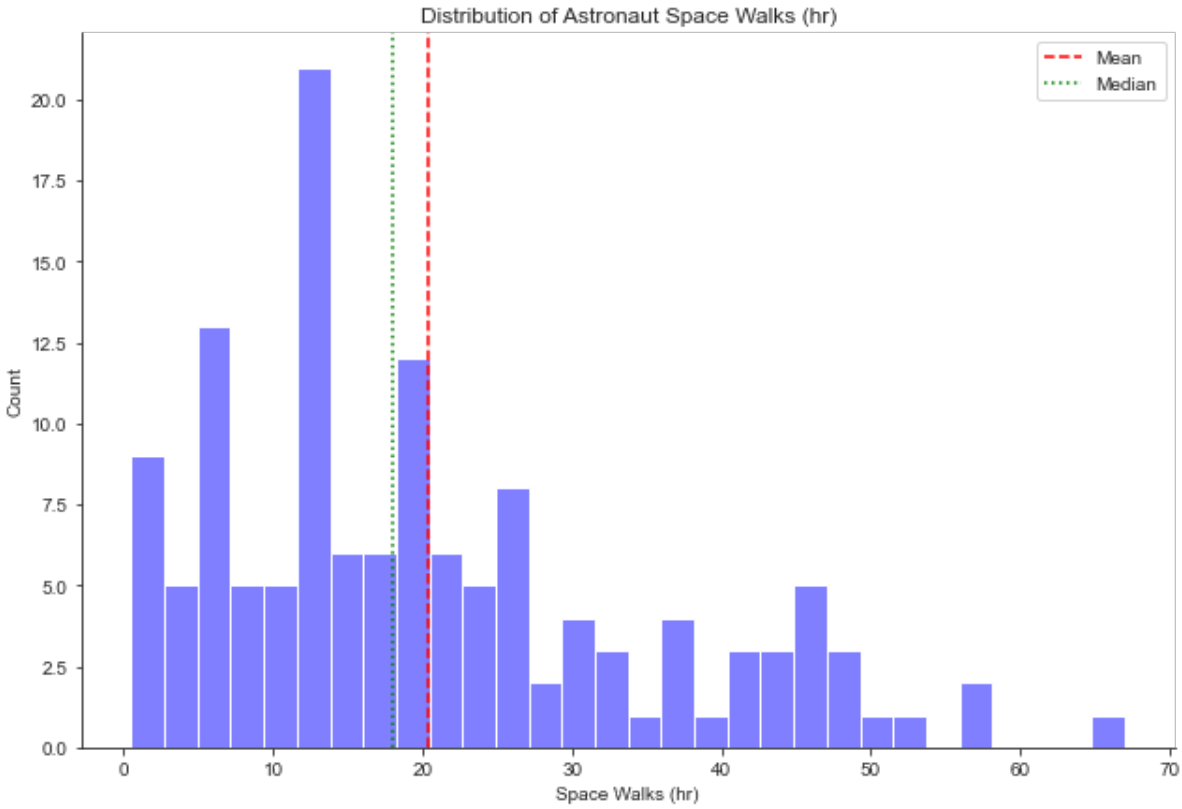
Boxplot based on Gender to Examine the Distribution and Identify Outliers

In [77]:
```
plt.figure(
    figsize = (12,8)
)
sns.boxplot(
    x = "Gender",
    y = "Space Walks (hr)",
    data = space_walks
)
plt.title("Distribution of Space walk Hours by Gender");
```



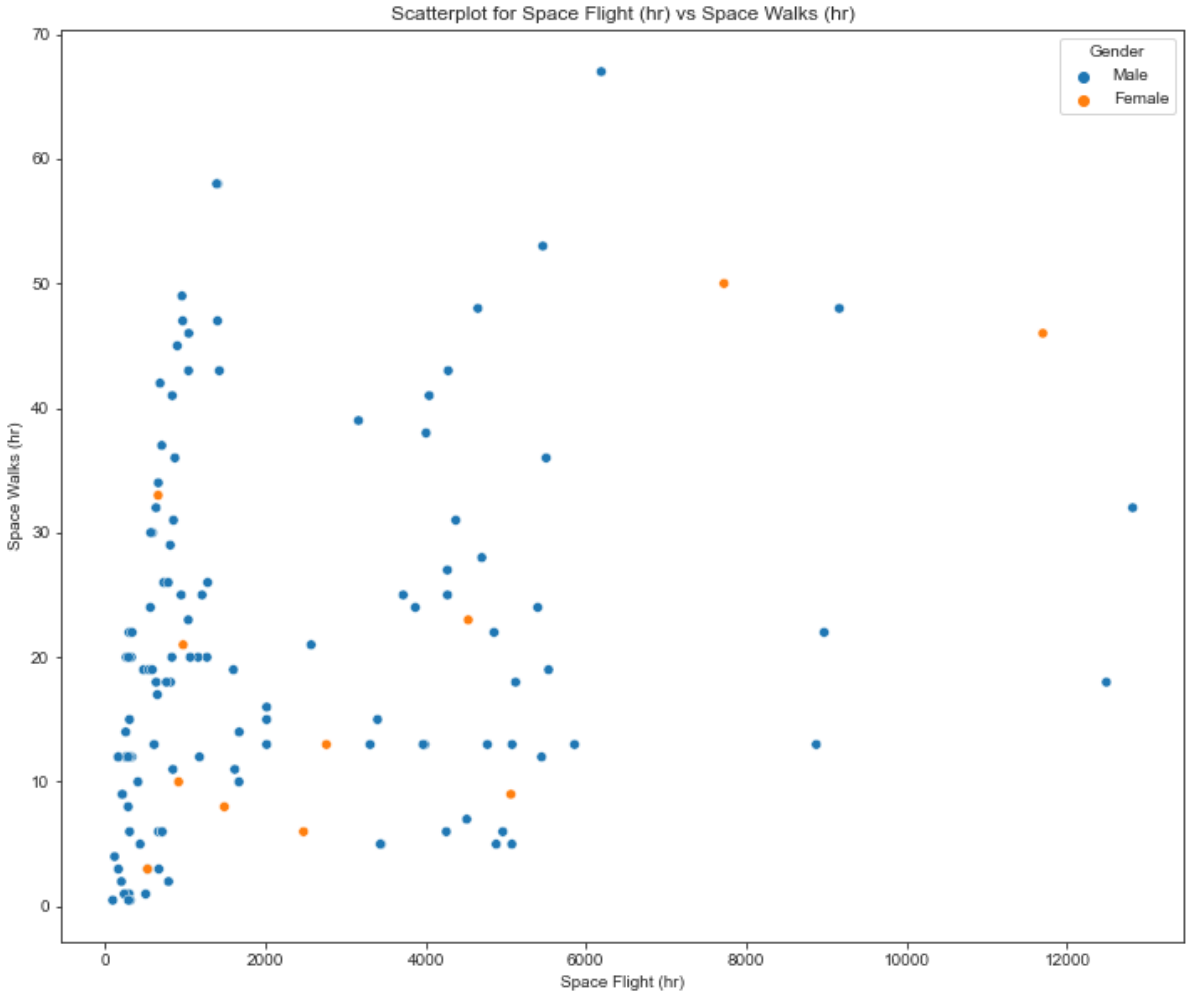
Histogram used to Examine the Distribution of Space Walks (hr)

In [78]:
```
sns.displot(
    space_walks["Space Walks (hr)"],
    bins = 30,
    color = "blue",
    height = 6,
    aspect = 1.5,
    alpha = 0.5
)
plt.axvline(
    np.mean(space_walks["Space Walks (hr)"]),
    ls = "--",
    label = "Mean",
    color = "red"
)
plt.axvline(
    np.median(space_walks["Space Walks (hr)"]),
    ls = ":",
    label = "Median",
    color = "green"
)
plt.legend()
plt.title("Distribution of Astronaut Space Walks (hr)");
```



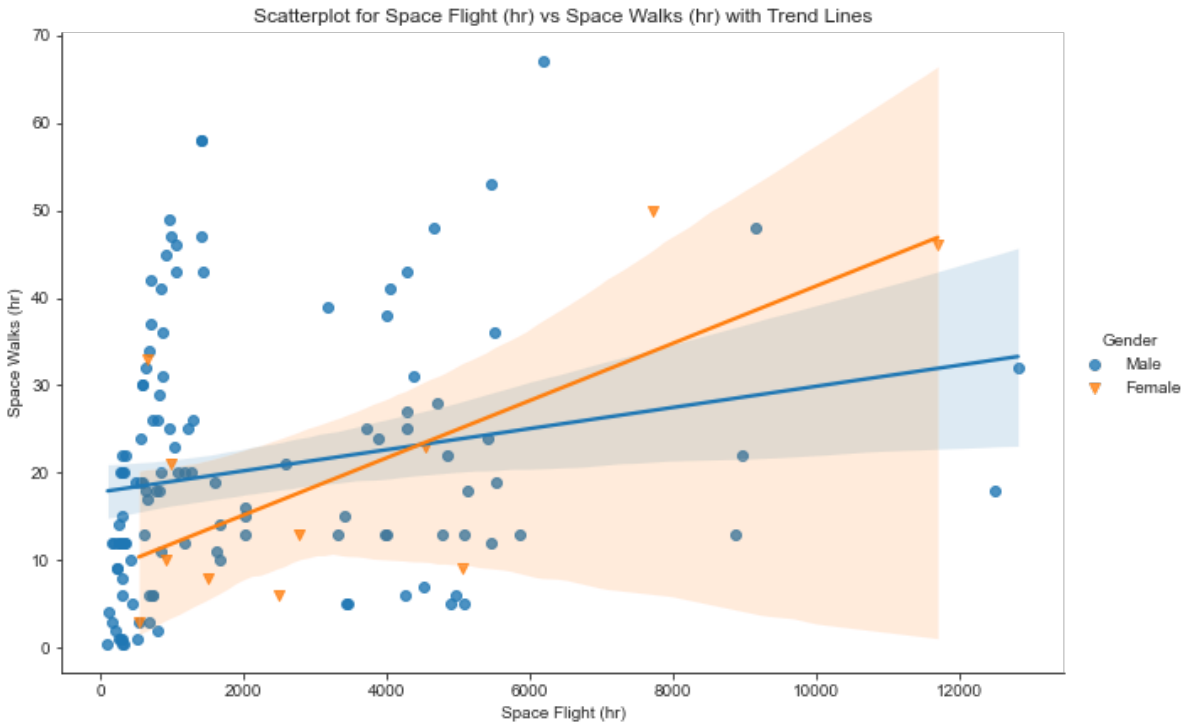
Visualizing the Scatterplot of the Relationship between Space Flight (hr) and Space Walks (hr)

In [79]:
```
plt.figure(
    figsize = (12,10)
)

sns.scatterplot(
    data = space_walks,
    x = "Space Flight (hr)",
    y = "Space Walks (hr)",
    hue = "Gender"
)
plt.title("Scatterplot for Space Flight (hr) vs Space Walks (hr)");
```



Visualizing the Relationship of Space Flight (hr) and Space Walks (hr)

In [80]:
```
sns.lmplot(
    x = "Space Flight (hr)",
    y = "Space Walks (hr)",
    data = space_walks,
    hue = "Gender",
    markers = ["o", "v"],
    height = 6,
    aspect = 1.5
)
plt.title("Scatterplot for Space Flight (hr) vs Space Walks (hr) with Trend Lines")
```



What is the Correlation between Space Flight (hr) and Space Walks (hr)?

In [81]:
```
space_walks["Space Flight (hr)"].corr(
    space_walks["Space Walks (hr)"]
)
```

Out[81]:
```
0.2595379056980412
```

Multicollinearity

Checking the Full Dataset Column Correlations for Multicollinearity.

In [82]:
```
df.corr()
```

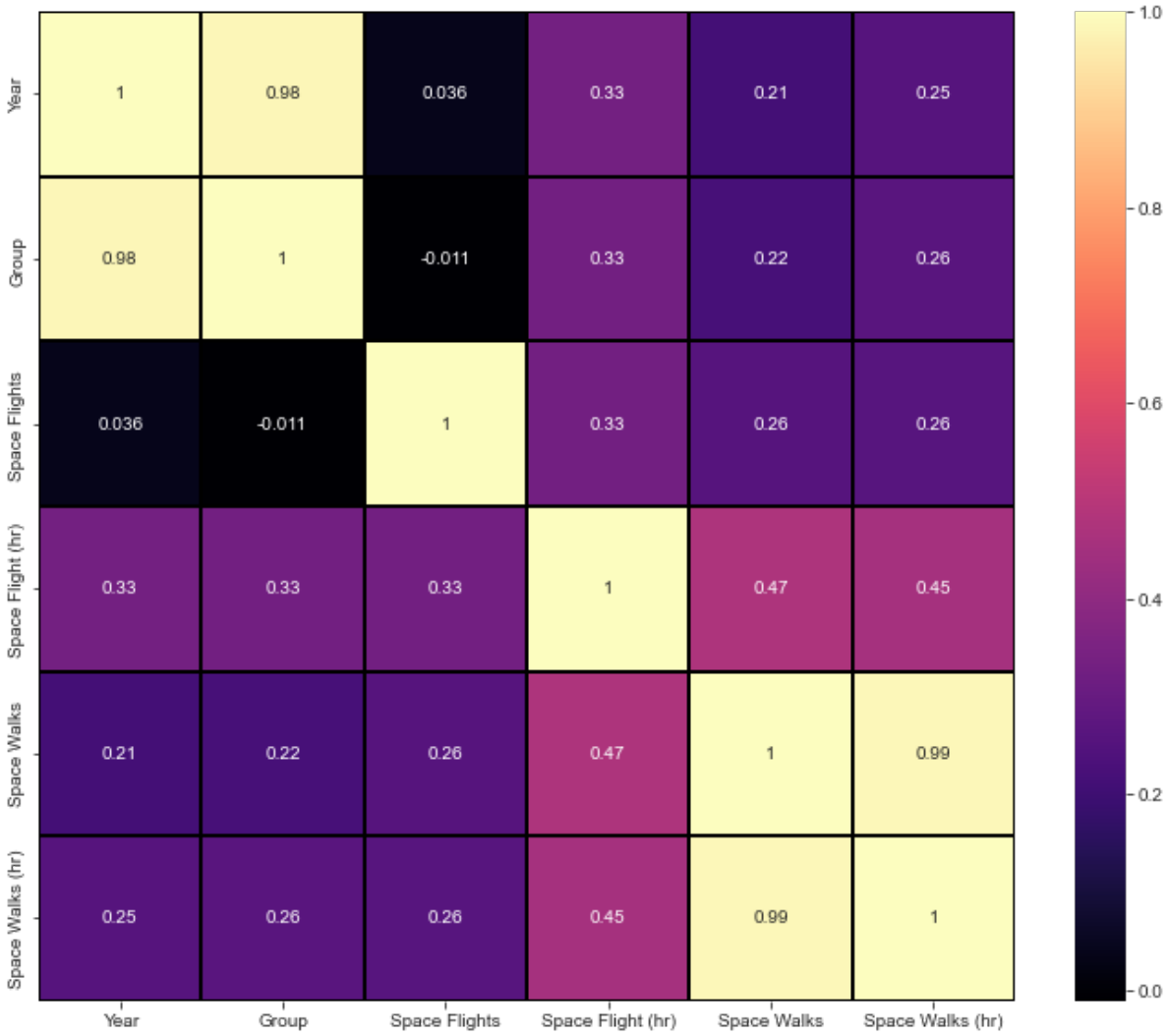
Out[82]:

	Year	Group	Space Flights	Space Flight (hr)	Space Walks	Space Walks (hr)
Year	1.000000	0.980934	0.036420	0.331386	0.210073	0.253502
Group	0.980934	1.000000	-0.011386	0.325683	0.217891	0.261384
Space Flights	0.036420	-0.011386	1.000000	0.325233	0.257073	0.258642
Space Flight (hr)	0.331386	0.325683	0.325233	1.000000	0.472796	0.454408
Space Walks	0.210073	0.217891	0.257073	0.472796	1.000000	0.985755
Space Walks (hr)	0.253502	0.261384	0.258642	0.454408	0.985755	1.000000

Visualizing the Full Dataset Correlation Matrix

In [83]:

```
plt.figure(
    figsize = (12,10)
)
astro_corr = df.corr()
sns.heatmap(
    astro_corr,
    cmap = "magma",
    linecolor = "black",
    linewidths = 2,
    annot = True
);
```



Correlation Matrix for Astronauts that have Performed at Least 1 Space Flight and Space Walk

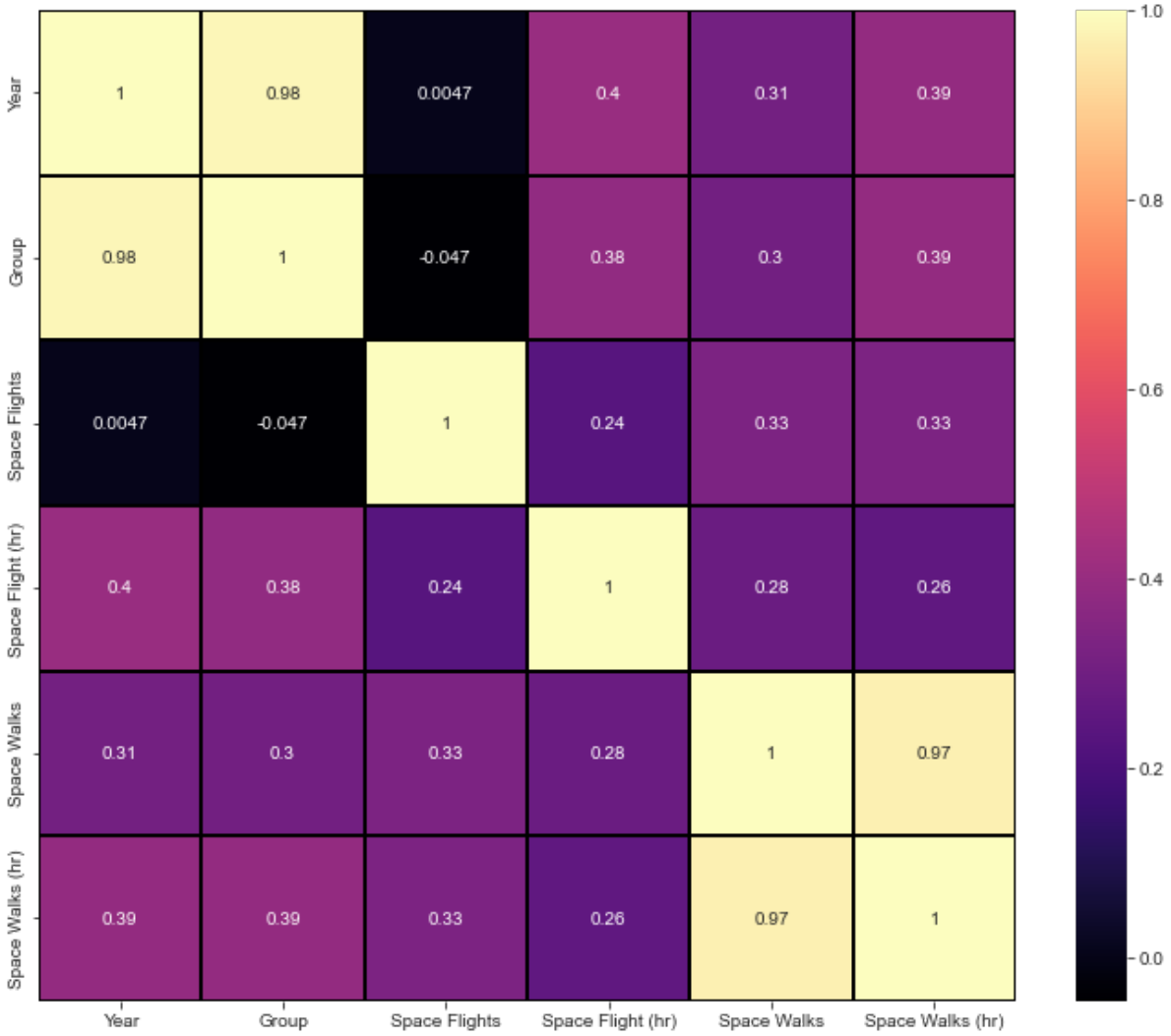
In [84]:
```
space_walks.corr()
```

Out[84]:

	Year	Group	Space Flights	Space Flight (hr)	Space Walks	Space Walks (hr)
Year	1.000000	0.980262	0.004708	0.399003	0.305482	0.390494
Group	0.980262	1.000000	-0.046717	0.384600	0.303513	0.389094
Space Flights	0.004708	-0.046717	1.000000	0.235222	0.333247	0.332081
Space Flight (hr)	0.399003	0.384600	0.235222	1.000000	0.282853	0.259538
Space Walks	0.305482	0.303513	0.333247	0.282853	1.000000	0.970257
Space Walks (hr)	0.390494	0.389094	0.332081	0.259538	0.970257	1.000000

Visualization of the Correlation Matrix for Astronauts that have Performed at Least 1 Space Walk and Space Flight

In [85]:
```
plt.figure(
    figsize = (12,10)
)

astro_corr = space_walks.corr()
sns.heatmap(
    astro_corr,
    cmap = "magma",
    linecolor = "black",
    linewidths = 2,
    annot = True
);
```



Key Insights

The numerical columns are not highly correlated.

- There is evidence for multicollinearity between columns "Group" and "Year", and "Space Walks" and "Space Walk (hr)".

83% of astronauts possessed a graduate degree.

The top 10 undegraduate degrees were all in STEM fields.

Mosts astronauts served in the military with Colonel as the most common rank.

The most common Military Rank for men was Colonel, while for women it was Captain.

The most common number of space flights an astronaut performed was 2.

There were more male astronauts than female astronauts.

- More men served in the military.
- It is a dangerous profession as more men died in accidents.
- Men performed more space walks.

Further Research is required:

Women on average spent more time in space.

- Women generally have less muscle mass and lower bone density than men.
- Women perhaps are more exposed to the detrimental physiological effects of microgravity.

Thank you