Project talks about the Super Bowl Halftime Show

Insights after finishing project, we can answer these questions

1. What are the most extreme game outcomes?
2. How does the game affect television viewership?
3. How have viewership, TV ratings, and ad cost evolved over time?
4. Who are the most prolific musicians in terms of halftime show performances?

In [1]:

```
# Load pandas library
import pandas as pd

# Load the CSV data
df1 = pd.read_csv('super_bowls.txt') # super bowls data
df2 = pd.read_csv('tv.txt') # tv data
df3 = pd.read_csv('halftime_musicians.txt') # musicians data
```

In [2]:

```
display(df1.head())
display(df2.head())
display(df3.head())
```

	date	super_bowl	venue	city	state	attendance	team_winner	winning_pts	qb_winner_1	qb_winner_2	coach_winner	tea
0	2018-02-04	52	U.S. Bank Stadium	Minneapolis	Minnesota	67612	Philadelphia Eagles	41	Nick Foles	NaN	Doug Pederson	
1	2017-02-05	51	NRG Stadium	Houston	Texas	70807	New England Patriots	34	Tom Brady	NaN	Bill Belichick	
2	2016-02-07	50	Levi's Stadium	Santa Clara	California	71088	Denver Broncos	24	Peyton Manning	NaN	Gary Kubiak	
3	2015-02-01	49	University of Phoenix Stadium	Glendale	Arizona	70288	New England Patriots	28	Tom Brady	NaN	Bill Belichick	S
4	2014-02-02	48	MetLife Stadium	East Rutherford	New Jersey	82529	Seattle Seahawks	43	Russell Wilson	NaN	Pete Carroll	

	super_bowl	network	avg_us_viewers	total_us_viewers	rating_household	share_household	rating_18_49	share_18_49	ad_cost
0	52	NBC	103390000	NaN	43.1	68	33.4	78.0	5000000
1	51	Fox	111319000	172000000.0	45.3	73	37.1	79.0	5000000
2	50	CBS	111864000	167000000.0	46.6	72	37.7	79.0	5000000
3	49	NBC	114442000	168000000.0	47.5	71	39.1	79.0	4500000
4	48	Fox	112191000	167000000.0	46.7	69	39.3	77.0	4000000

	super_bowl	musician	num_songs
0	52	Justin Timberlake	11.0
1	52	University of Minnesota Marching Band	1.0
2	51	Lady Gaga	7.0
3	50	Coldplay	6.0
4	50	Beyoncé	3.0

In [3]:

```
# Summary of the TV data
df2.info()

print('\n')

# Summary of the halftime musician data to inspect
df3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53 entries, 0 to 52
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   super_bowl        53 non-null     int64  
 1   network           53 non-null     object 
 2   avg_us_viewers    53 non-null     int64  
 3   total_us_viewers  15 non-null     float64
 4   rating_household  53 non-null     float64
 5   share_household   53 non-null     int64  
 6   rating_18_49      15 non-null     float64
 7   share_18_49       6 non-null      float64
 8   ad_cost           53 non-null     int64  
dtypes: float64(4), int64(4), object(1)
memory usage: 3.9+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 134 entries, 0 to 133
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   super_bowl  134 non-null    int64  
 1   musician    134 non-null    object 
 2   num_songs   88 non-null     float64
dtypes: float64(1), int64(1), object(1)
memory usage: 3.3+ KB
```

From TV data, the following columns have missing values and a lot of them:

- *total_us_viewers:* (amount of U.S. viewers who watched at least some part of the TV program)
- *rating_18_49:* (average % of U.S. adults 18-49 who watch entire TV program)
- *share_18_49:* (average % of U.S. adults 18-49 who watch entire TV program with TV in use)

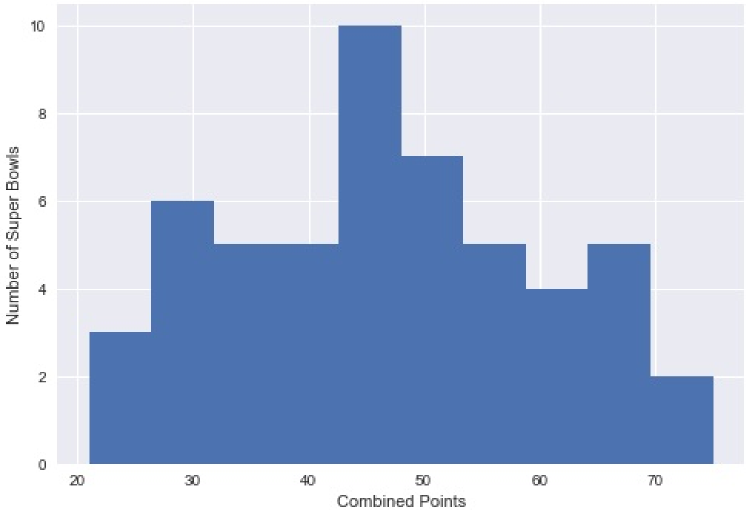
Visualize data to look at combine at point with histogram and check the highest and lowest.

In [4]:

```
# Import matplotlib and set plotting style
from matplotlib import pyplot as plt
%matplotlib inline
plt.style.use('seaborn')

# Plot a histogram of combined points
df1['combined_pts'].hist()
plt.xlabel('Combined Points')
plt.ylabel('Number of Super Bowls')
plt.show()


# Display the Super Bowls with the highest and lowest combined scores
display(df1[df1['combined_pts'] > 70])
display(df1[df1['combined_pts'] < 25])
```



	date	super_bowl	venue	city	state	attendance	team_winner	winning_pts	qb_winner_1	qb_winner_2	coach_winner	tea
0	2018-02-04	52	U.S. Bank Stadium	Minneapolis	Minnesota	67612	Philadelphia Eagles	41	Nick Foles	NaN	Doug Pederson	
23	1995-01-29	29	Joe Robbie Stadium	Miami Gardens	Florida	74107	San Francisco 49ers	49	Steve Young	NaN	George Seifert	Sa C

	date	super_bowl	venue	city	state	attendance	team_winner	winning_pts	qb_winner_1	qb_winner_2	coach_winner	team_l
43	1975-01-12	9	Tulane Stadium	New Orleans	Louisiana	80997	Pittsburgh Steelers	16	Terry Bradshaw	NaN	Chuck Noll	Minne Vi
45	1973-01-14	7	Memorial Coliseum	Los Angeles	California	90182	Miami Dolphins	14	Bob Griese	NaN	Don Shula	Washir Red
49	1969-01-12	3	Orange Bowl	Miami	Florida	75389	New York Jets	16	Joe Namath	NaN	Weeb Ewbank	Balti

The histogram shows that most scores are in range between 40 and 55

The highest combined points 74 (in 2018) and 75 (1995)

The lowest combined points is 21 (in1973)

And the I will check difference points

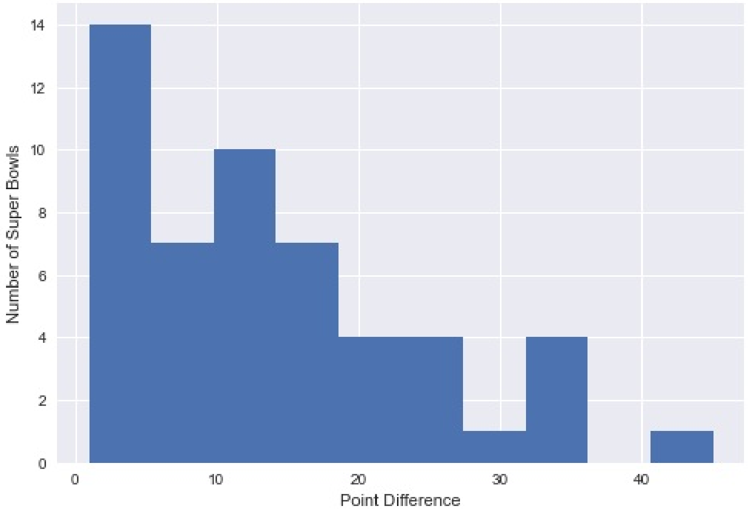
In [5]:

```
# Plot a histogram of point differences
plt.hist(df1.difference_pts)
plt.xlabel('Point Difference')
plt.ylabel('Number of Super Bowls')

# Display the closest game(s) and biggest blowouts
display(df1[df1.difference_pts == 1])
display(df1[df1.difference_pts >=35])
```

	date	super_bowl	venue	city	state	attendance	team_winner	winning_pts	qb_winner_1	qb_winner_2	coach_winner	team_loser
27	1991-01-27	25	Tampa Stadium	Tampa	Florida	73813	New York Giants	20	Jeff Hostetler	NaN	Bill Parcells	Buffalo Bills

	date	super_bowl	venue	city	state	attendance	team_winner	winning_pts	qb_winner_1	qb_winner_2	coach_winner	te
4	2014-02-02	48	MetLife Stadium	East Rutherford	New Jersey	82529	Seattle Seahawks	43	Russell Wilson	NaN	Pete Carroll	
25	1993-01-31	27	Rose Bowl	Pasadena	California	98374	Dallas Cowboys	52	Troy Aikman	NaN	Jimmy Johnson	Bu
28	1990-01-28	24	Louisiana Superdome	New Orleans	Louisiana	72919	San Francisco 49ers	55	Joe Montana	NaN	George Seifert	
32	1986-01-26	20	Louisiana Superdome	New Orleans	Louisiana	73818	Chicago Bears	46	Jim McMahon	NaN	Mike Ditka	



The most difference points is between 0 and 13. There is only a game which have the 45 (!) difference points where Hall of Famer Joe Montana's led the San Francisco 49ers to victory in 1990, one year before the closest game ever.

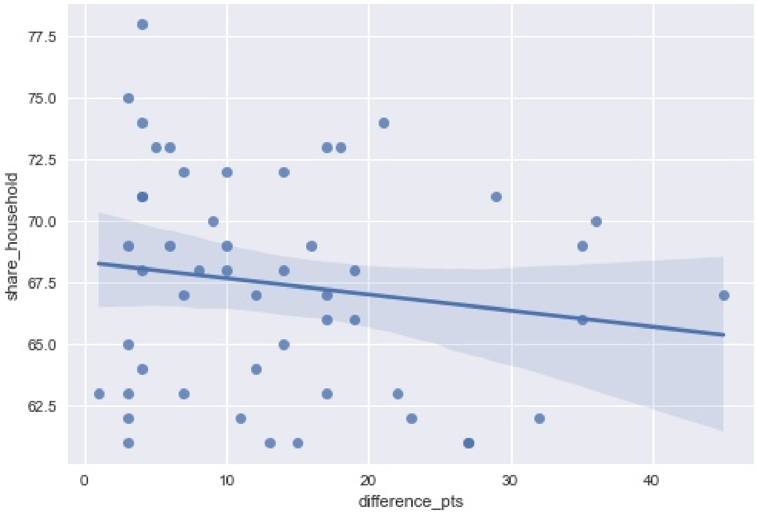
Now I want to check if games which have huge difference points can decrease viewers.

```
In [6]: # Join game and TV data, filtering out SB I because it was split over two networks
games_tv = pd.merge(df2[df2['super_bowl'] > 1], df1, on='super_bowl')

# Import seaborn
import seaborn as sns

# Create a scatter plot with a linear regression model fit
sns.regplot(x='difference_pts', y='share_household', data=games_tv)
```

```
Out[6]: <AxesSubplot:xlabel='difference_pts', ylabel='share_household'>
```



Overall, the graph shows that there was a downward trend when this match was blowout.

The relationship of Superbowl with average Number of US Viewers, household rating, Ad Cost

```
In [7]: # Create a figure with 2x1 subplot and activate the top subplot
```

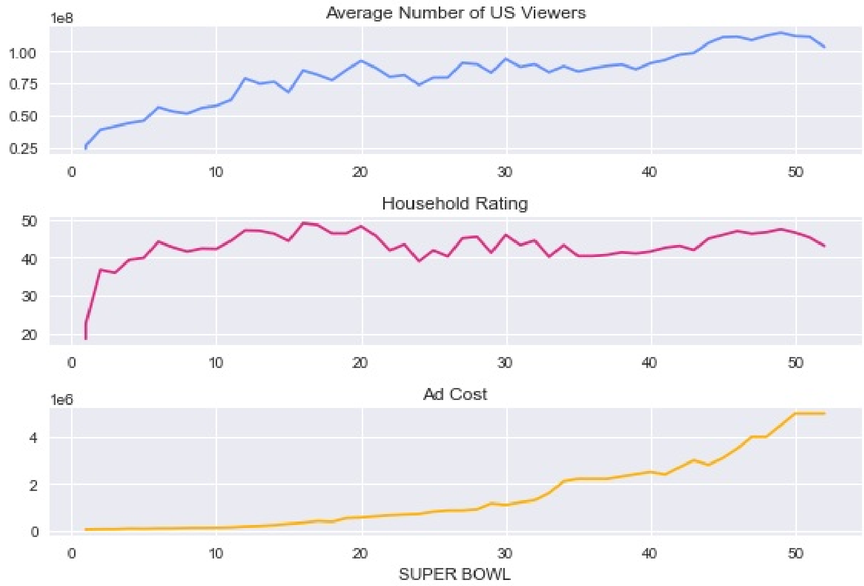
In [7]:

```
# Create a figure with 3x1 subplot and activate the top subplot
plt.subplot(3, 1, 1)
plt.plot(df2.super_bowl, df2.avg_us_viewers, color='#648FFF')
plt.title('Average Number of US Viewers')

# Activate the middle subplot
plt.subplot(3, 1, 2)
plt.plot(df2.super_bowl, df2.rating_household, color='#DC267F')
plt.title('Household Rating')

# Activate the bottom subplot
plt.subplot(3, 1, 3)
plt.plot(df2.super_bowl, df2.ad_cost, color='#FFB000')
plt.title('Ad Cost')
plt.xlabel('SUPER BOWL')

# Improve the spacing between subplots
plt.tight_layout()
```



In general, they all increased over the time. We can observe that the viewers increased before ad cost did.

In [8]:

```
# Display all halftime musicians for Super Bowls up to and including Super Bowl XXVII
df3[df3.super_bowl <= 27]
```

Out[8]:

	super_bowl	musician	num_songs
80	27	Michael Jackson	5.0
81	26	Gloria Estefan	2.0
82	26	University of Minnesota Marching Band	NaN
83	25	New Kids on the Block	2.0
84	24	Pete Fountain	1.0
85	24	Doug Kershaw	1.0
86	24	Irma Thomas	1.0
87	24	Pride of Nicholls Marching Band	NaN
88	24	The Human Jukebox	NaN
89	24	Pride of Acadiana	NaN
90	23	Elvis Presto	7.0
91	22	Chubby Checker	2.0
92	22	San Diego State University Marching Aztecs	NaN
93	22	Spirit of Troy	NaN
94	21	Grambling State University Tiger Marching Band	8.0
95	21	Spirit of Troy	8.0
96	20	Up with People	NaN
97	19	Tops In Blue	NaN
98	18	The University of Florida Fightin' Gator March...	7.0
99	18	The Florida State University Marching Chiefs	7.0
100	17	Los Angeles Unified School District All City H...	NaN

100	17	Los Angeles Unified School District All City H...	NaN
101	16	Up with People	NaN
102	15	The Human Jukebox	NaN
103	15	Helen O'Connell	NaN
104	14	Up with People	NaN
105	14	Grambling State University Tiger Marching Band	NaN
106	13	Ken Hamilton	NaN
107	13	Gramacks	NaN
108	12	Tyler Junior College Apache Band	NaN
109	12	Pete Fountain	NaN
110	12	Al Hirt	NaN
111	11	Los Angeles Unified School District All City H...	NaN
112	10	Up with People	NaN
113	9	Mercer Ellington	NaN
114	9	Grambling State University Tiger Marching Band	NaN
115	8	University of Texas Longhorn Band	NaN
116	8	Judy Mallett	NaN
117	7	University of Michigan Marching Band	NaN
118	7	Woody Herman	NaN
119	7	Andy Williams	NaN
120	6	Ella Fitzgerald	NaN
121	6	Carol Channing	NaN
122	6	Al Hirt	NaN
123	6	United States Air Force Academy Cadet Chorale	NaN
124	5	Southeast Missouri State Marching Band	NaN
125	4	Marguerite Piazza	NaN
126	4	Doc Severinsen	NaN
127	4	Al Hirt	NaN
128	4	The Human Jukebox	NaN
129	3	Florida A&M University Marching 100 Band	NaN
130	2	Grambling State University Tiger Marching Band	NaN
131	1	University of Arizona Symphonic Marching Band	NaN
132	1	Grambling State University Tiger Marching Band	NaN
133	1	Al Hirt	NaN

From 1 to 27, this time we can see that there is a significant increase in number of viewers, and household rating.

Filter the muscian to find the reason for this upward trend.

In [9]:

```
# Count halftime show appearances for each musician and sort them from most to least
halftime_appearances = df3.groupby('musician').count()['super_bowl'].reset_index()
halftime_appearances = halftime_appearances.sort_values('super_bowl', ascending=False)

# Display musicians with more than one halftime show appearance
halftime_appearances[halftime_appearances['super_bowl'] > 1]
```

Out[9]:

	musician	super_bowl
28	Grambling State University Tiger Marching Band	6
104	Up with People	4
1	Al Hirt	4
83	The Human Jukebox	3
76	Spirit of Troy	2
25	Florida A&M University Marching 100 Band	2
26	Gloria Estefan	2
102	University of Minnesota Marching Band	2
10	Bruno Mars	2
64	Pete Fountain	2
5	Beyoncé	2
36	Justin Timberlake	2
57	Nelly	2
44	Los Angeles Unified School District All City H...	2

The world famous Grambling State University Tiger Marching Band takes the crown with six appearances. Beyoncé, Justin Timberlake, Nelly, and Bruno Mars are the only post-Y2K musicians with multiple appearances (two each).

From previous, we saw there are many missing values so now I will solve this problem

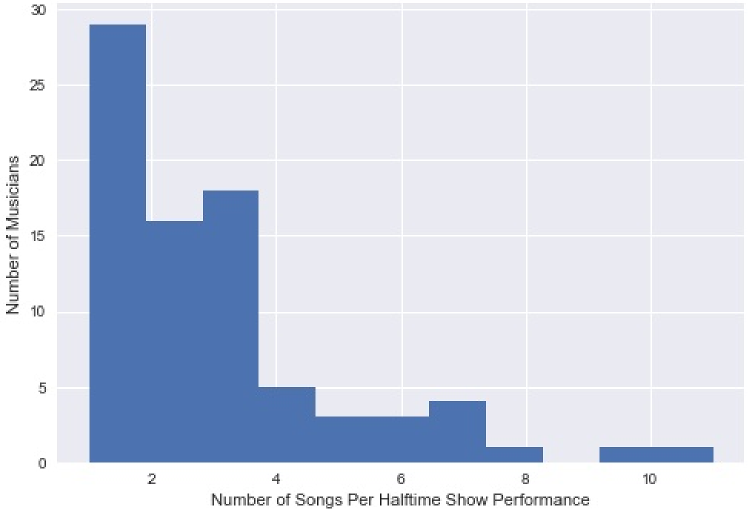
Let's filter out marching bands by filtering out musicians with the word "Marching" in them and the word "Spirit" (a common naming convention for marching bands is "Spirit of [something]"). Then we'll filter for Super Bowls after Super Bowl XX to address the missing data issue, then let's see who has the most number of songs.

In [10]:

```python
# Filter out most marching bands
no_bands = df3[~df3.musician.str.contains('Marching')]
no_bands = no_bands[~no_bands.musician.str.contains('Spirit')]

# Plot a histogram of number of songs per performance
most_songs = int(max(no_bands['num_songs'].values))
plt.hist(no_bands.num_songs.dropna(), bins=most_songs)
plt.xlabel('Number of Songs Per Halftime Show Performance')
plt.ylabel('Number of Musicians')
plt.show()

# Sort the non-band musicians by number of songs per appearance...
no_bands = no_bands.sort_values('num_songs', ascending=False)
# ...and display the top 15
display(no_bands.head(15))
```



	super_bowl	musician	num_songs
0	52	Justin Timberlake	11.0
70	30	Diana Ross	10.0
10	49	Katy Perry	8.0
2	51	Lady Gaga	7.0
90	23	Elvis Presto	7.0
33	41	Prince	7.0
16	47	Beyoncé	7.0
14	48	Bruno Mars	6.0
3	50	Coldplay	6.0
25	45	The Black Eyed Peas	6.0
20	46	Madonna	5.0
30	44	The Who	5.0
80	27	Michael Jackson	5.0
64	32	The Temptations	4.0
36	39	Paul McCartney	4.0

So most non-band musicians do 1-3 songs per halftime show. It's important to note that the duration of the halftime show is fixed (roughly 12 minutes) so songs per performance is more a measure of how many hit songs you have. JT went off in 2018, wow. 11 songs! Diana Ross comes in second with 10 in her medley in 1996.

In this notebook, we loaded, cleaned, then explored Super Bowl game, television, and halftime show data. We visualized the distributions of combined points, point differences, and halftime show performances using histograms. We used line plots to see how ad cost increases lagged behind viewership increases. And we discovered that blowouts do appear to lead to a drop in viewers.

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js