Project (Python): Analyzing TV Data

by Jenn Lai

Introduction

Whether or not you like football, the Super Bowl is a spectacle. There's a little something for everyone at your Super Bowl party. Drama in the form of blowouts, comebacks, and controversy for the sports fan. There are the ridiculously expensive ads, some hilarious, others gut-wrenching, thought-provoking, and weird. The half-time shows with the biggest musicians in the world, sometimes riding giant mechanical tigers or leaping from the roof of the stadium. It's a show, baby. And in this notebook, we're going to find out how some of the elements of this show interact with each other. After exploring and cleaning our data a little, we're going to answer questions like:

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

1. TV, halftime shows, and the Big Game

Import pandas then load the data

```
import pandas as pd

super_bowls = pd.read_csv('datasets/super_bowls.csv')
tv = pd.read_csv('datasets/tv.csv')
halftime_musicians = pd.read_csv('datasets/halftime_musicians.csv')

display(super_bowls.head())
display(tv.head())
display(halftime_musicians.head())
```

Result :

super_bowls

	date	super_bowl	venue	city	state	attendance	team_winner
0	2018-02-04	52	U.S. Bank Stadium	Minneapolis	Minnesota	67612	Philadelphia Eagle
1	2017-02-05	51	NRG Stadium	Houston	Texas	70807	New England Patrio
2	2016-02-07	50	Levi's Stadium	Santa Clara	California	71088	Denver Broncos
3	2015-02-01	49	University of Phoenix Stadium	Glendale	Arizona	70288	New England Patrio
4	2014-02-02	48	MetLife Stadium	East Rutherford	New Jersey	82529	Seattle Seahawks

tv

	super_bowl	network	avg_us_viewers	total_us_viewers	rating_household	share_household	rating_18_49	share_18_49
0	52	NBC	103390000	null	43.1	68	33.4	7
1	51	Fox	111319000	172000000	45.3	73	37.1	7
2	50	CBS	111864000	167000000	46.6	72	37.7	7
3	49	NBC	114442000	168000000	47.5	71	39.1	7
4	48	Fox	112191000	167000000	46.7	69	39.3	7

halftime_musicians

	super_bowl	musician	num_songs
0	52	Justin Timberlake	11
1	52	University of Minnesota Marching Band	1
2	51	Lady Gaga	7
3	50	Coldplay	6
4	50	Beyoncé	3

2. Taking note of dataset issues

Display and inspect the summaries of the TV and halftime musician DataFrames for issues.

```
1  tv.info()
2
3  print()
4
5  halftime_musicians.info()
```

Result :

```
<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
```

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

total_us_viewers, rating_18_49, share_18_49

3. Combined points distribution

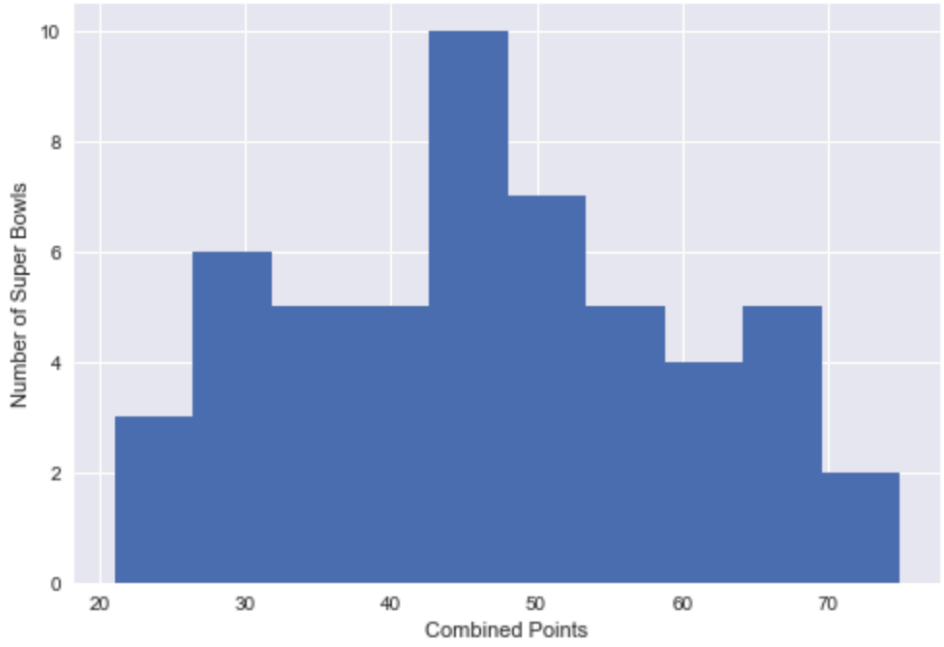
Plot a histogram of combined points then display the rows with the most extreme combined point outcomes.

```
from matplotlib import pyplot as plt
plt.style.use('seaborn')

plt.hist(super_bowls.combined_pts)
plt.xlabel('Combined Points')
plt.ylabel('Number of Super Bowls')
plt.show()

display(super_bowls[super_bowls['combined_pts'] > 70])
display(super_bowls[super_bowls['combined_pts'] < 25])
```

Result :



	date	super_bowl	venue	city	state	attendance	team_winner	winning_pts	qb_winner_1	qb_winner_2	coach_winner	team_loser	losing_pts	qb_loser_1	qb_loser_2	coach_loser	combined_pts	difference_pts
0	2018-02-04	52	U.S. Bank Stadium	Minneapolis	Minnesota	67612	Philadelphia Eagles	41	Nick Foles	null	Doug Pederson	New England Patriots	33	Tom Brady	null	Bill Belichick	74	8
23	1995-01-29	29	Joe Robbie Stadium	Miami Gardens	Florida	74107	San Francisco 49ers	49	Steve Young	null	George Seifert	San Diego Chargers	26	Stan Humphreys	null	Bobby Ross	75	23

	date	super_bowl	venue	city	state	attendance	team_winner	winning_pts	qb_winner_1	qb_winner_2	coach_winner	team_loser	losing_pts	qb_loser_1	qb_loser_2	coach_loser	combined_pts	difference_pts
43	1975-01-12	9	Tulane Stadium	New Orleans	Louisiana	80997	Pittsburgh Steelers	16	Terry Bradshaw	null	Chuck Noll	Minnesota Vikings	6	Fran Tarkenton	null	Bud Grant	22	10
45	1973-01-14	7	Memorial Coliseum	Los Angeles	California	90182	Miami Dolphins	14	Bob Griese	null	Don Shula	Washington Redskins	7	Bill Kilmer	null	George Allen	21	7
49	1969-01-12	3	Orange Bowl	Miami	Florida	75389	New York Jets	16	Joe Namath	null	Weeb Ewbank	Baltimore Colts	7	Earl Morrall	Johnny Unitas	Don Shula	23	9

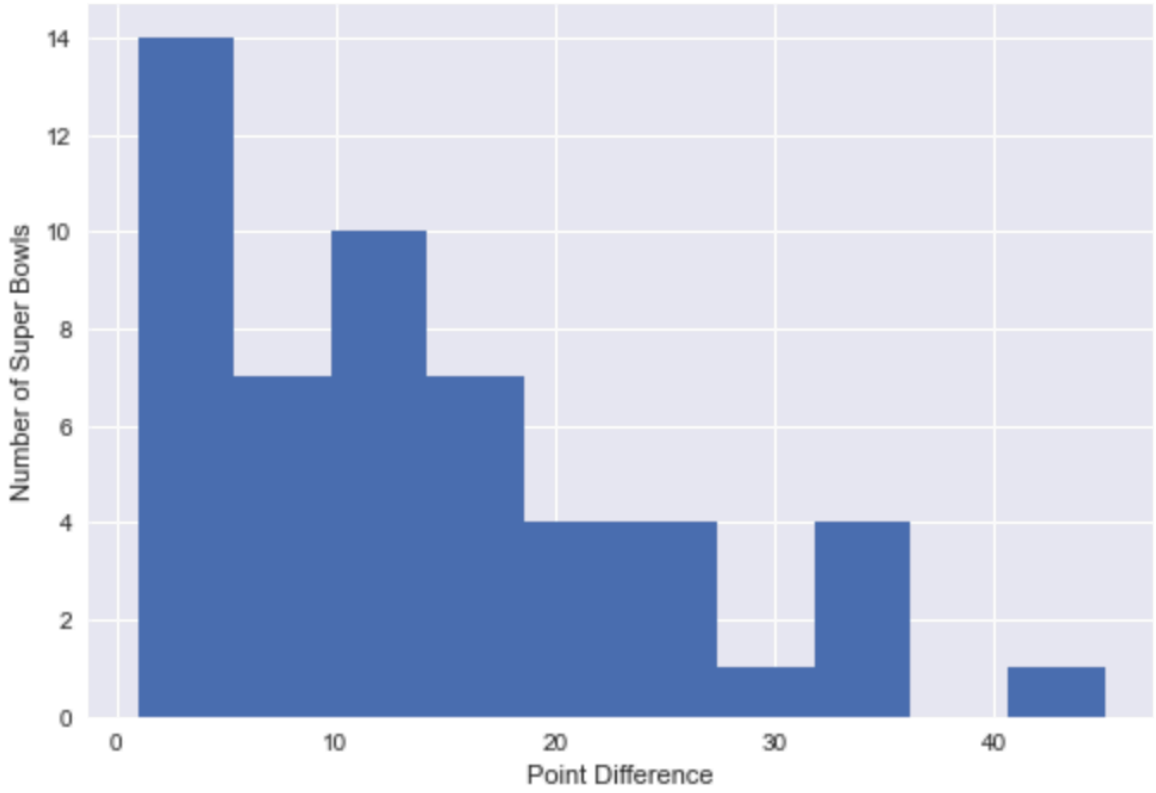
*Most combined scores are around 40-50 points, with the extremes being roughly equal distance away in opposite directions. With the highest combined scores at 74 and 75 and the lowest at 21.

4. Point difference distribution

Modify and display the histogram of point differences, then display the rows with the most extreme point difference outcomes.

```
plt.hist(super_bowls.difference_pts)
plt.xlabel('Point Difference')
plt.ylabel('Number of Super Bowls')
plt.show()
```

Result :



* The vast majority of Super Bowls are close games

5. Do blowouts translate to lost viewers?

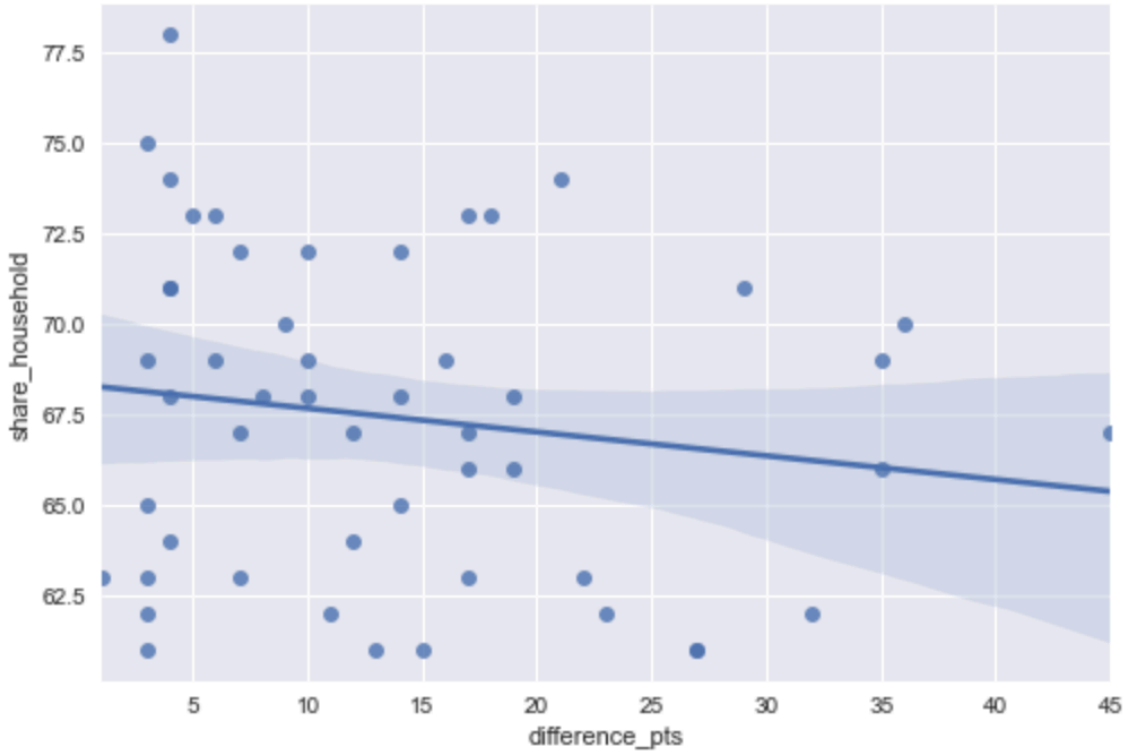
Import seaborn and plot household share vs. point difference.

```
games_tv = pd.merge(tv[tv['super_bowl'] > 1], super_bowls, on='super_bowl')

import seaborn as sns

sns.regplot(x='difference_pts', y='share_household', data=games_tv)
```

Result :



*The downward sloping regression line and the 95% confidence interval for that regression suggest that bailing on the game if it is a blowout is common.

6. Viewership and the ad industry over time

Create three line plots using the tv DataFrame to compare viewers, rating, and ad cost.

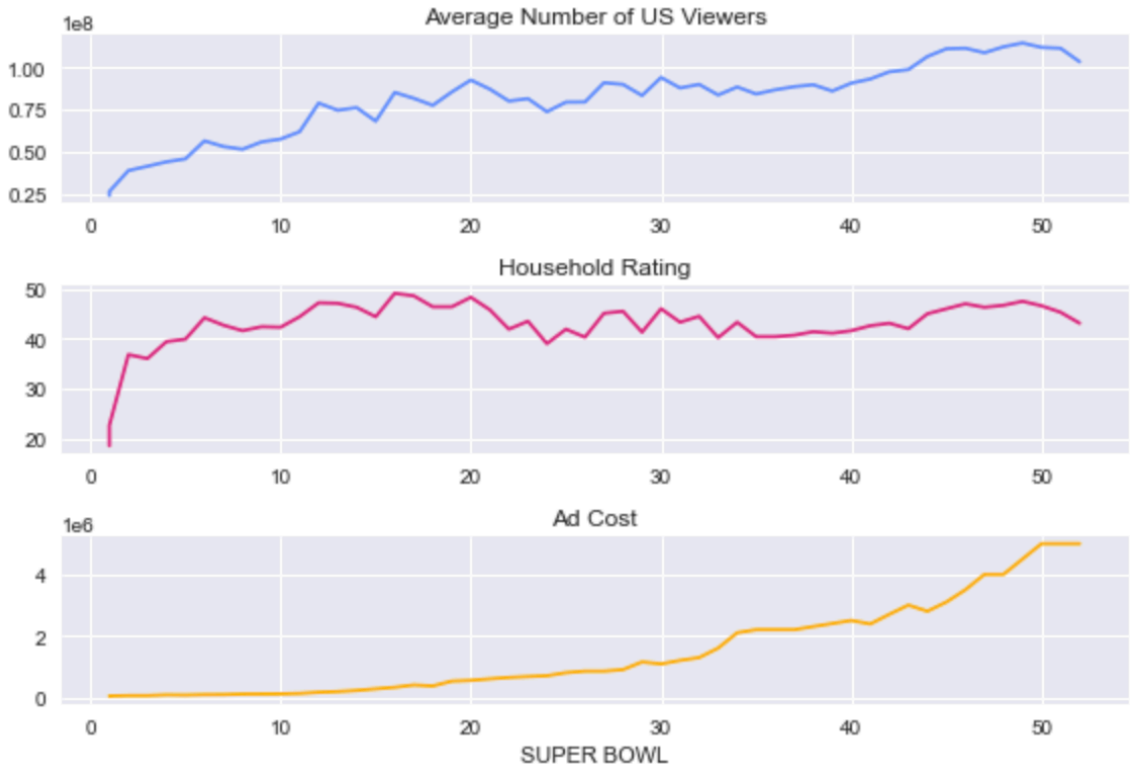
```
plt.subplot(3, 1, 1)
plt.plot(tv.super_bowl, tv.avg_us_viewers, color='#648FFF')
plt.title('Average Number of US Viewers')

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

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

plt.tight_layout()
```

Result :



* We can see viewers increased before ad costs did.
*Maybe halftime shows weren't that good in the earlier years?

7. Halftime shows weren't always this great

Filter and display the musicians for halftime shows up to and including Super Bowl 27.

```
halftime_musicians[halftime_musicians.super_bowl <= 27]
```

Result :

Collapse — Rows per page 10 — 54 rows

	super_bowl	musician	num_songs
80	27	Michael Jackson	5
81	26	Gloria Estefan	2
82	26	University of Minnesota Marching Band	null
83	25	New Kids on the Block	2
84	24	Pete Fountain	1
85	24	Doug Kershaw	1
86	24	Irma Thomas	1
87	24	Pride of Nicholls Marching Band	null
88	24	The Human Jukebox	null
89	24	Pride of Acadiana	null

Page 1 of 6 — Rows per page 10 — 54 rows

*It turns out Michael Jackson's Super Bowl XXVII performance, one of the most watched events in American TV history, was when the NFL realized the value of Super Bowl airtime and decided they needed to sign big name acts from then on out.

8. Who has the most halftime show appearances?

Select and display the musicians with more than one halftime show appearance.

```
halftime_appearances = halftime_musicians.groupby('musician').count()['super_bowl'].reset_index()
halftime_appearances = halftime_appearances.sort_values('super_bowl', ascending=False)
halftime_appearances[halftime_appearances['super_bowl'] > 1]
```

Result :

Collapse — Rows per page 10 — 14 rows

	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

Page 1 of 2 — Rows per page 10 — 14 rows

9. Who performed the most songs in a halftime show?

Modify the histogram of number of songs performed for non-band musicians.

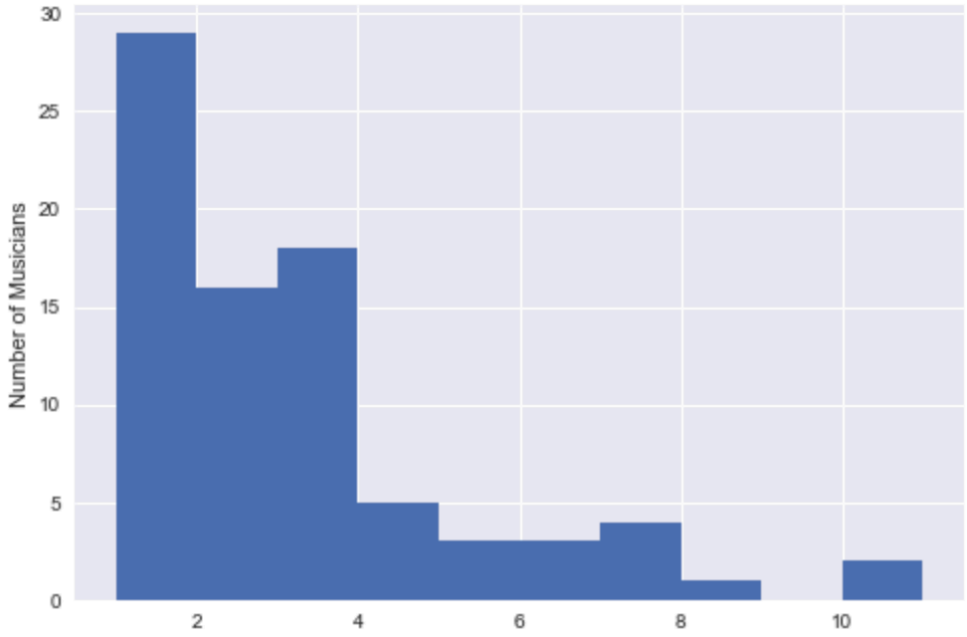
```
no_bands = halftime_musicians[~halftime_musicians.musician.str.contains('Marching')]
no_bands = no_bands[~no_bands.musician.str.contains('Spirit')]

most_songs = int(max(no_bands['num_songs'].values))
plt.hist(no_bands.num_songs.dropna())
plt.ylabel('Number of Songs Per Halftime Show Performance')
plt.ylabel('Number of Musicians')
plt.show()

no_bands = no_bands.sort_values('num_songs', ascending=False)

display(no_bands.head(15))
```

Result :



Collapse — Rows per page 10 — 15 rows

	super_bowl	musician	num_songs
0	52	Justin Timberlake	11
70	30	Diana Ross	10
10	49	Katy Perry	8
2	51	Lady Gaga	7
90	23	Elvis Presto	7
33	41	Prince	7
16	47	Beyoncé	7
14	48	Bruno Mars	6
3	50	Coldplay	6
25	45	The Black Eyed Peas	6

Page 1 of 2 — Rows per page 10 — 15 rows

*So most non-band musicians do 1-3 songs per halftime show, yet Justin Timberlake performed 11 songs in 2018.