ON

BROADWAY


my boys
JERSEY BOYS
The story of Frankie Valli & The Four Seasons
HAIR
ROCK OF AGES
A NEW MUSICAL
THE BEGINNING IS NEAR
JUNE 2
TheSims3.com
PATRICIA CORNWELL
273 LAFAYETTE ST.
49 MERCER ST.
ALSO AVAILABLE AT
So Much Happened Before Dorothy Dropped In
A New Musical
WICKED
The Untold Story of the Witches of Oz
Tonight belongs to...
PHANTOM
TEXT "ROCKFAM" to 56643 for FREE STUFF!
TADS Steaks
sbarro
sbarro
sbarro
sbarro
sbarro
sbarro

SHOW-SCORE

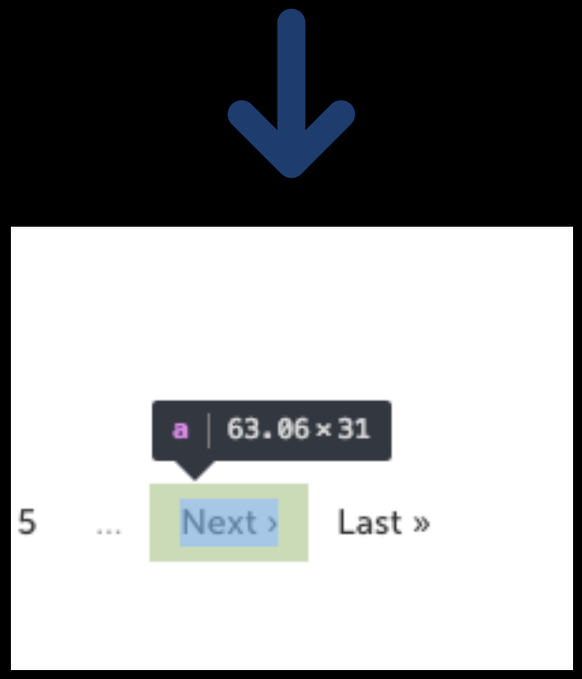
https://www.show-score.com/

- Scraped all reviews for the 29 shows currently on broadway
- Over 90,000 reviews scraped
- Mainly looking for score and descriptions

SCRAPING SHOW-SCORE

 tag containing each review

<a> tag containing href for the next page



urls followed a simple pattern

(?page+1)

BROADWAY.COM

http://www.broadway.com/

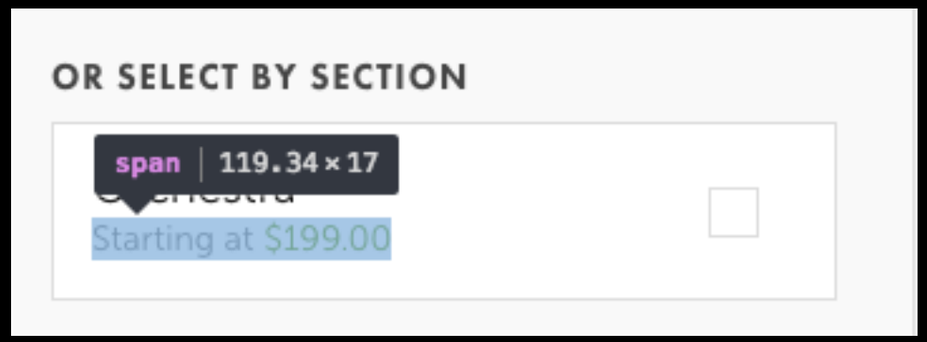
- Looking for price data for the shows I scraped from Show-Score

- Ended up being much harder to scrape than initially anticipated

- Built two spiders to scrape this site:
  - First spider to extract all the links needed
  - Second spider to extract the items from those links

SCRAPING BROADWAY.COM

 tag has the price
I want to extract




```
In [5]: response.xpath('//span')
Out[5]: []

In [6]:
```

???????

the content is being dynamically generated by javascript

SCRAPING BROADWAY.COM

Surprisingly, almost all of the information I want is in this mess of a script

Everything except for the next link for my spider to follow. So I need to find them

```
▼<script type="text/javascript">
            // IE9 fix
            if(!window.console) {
                var console = {}
            }
```

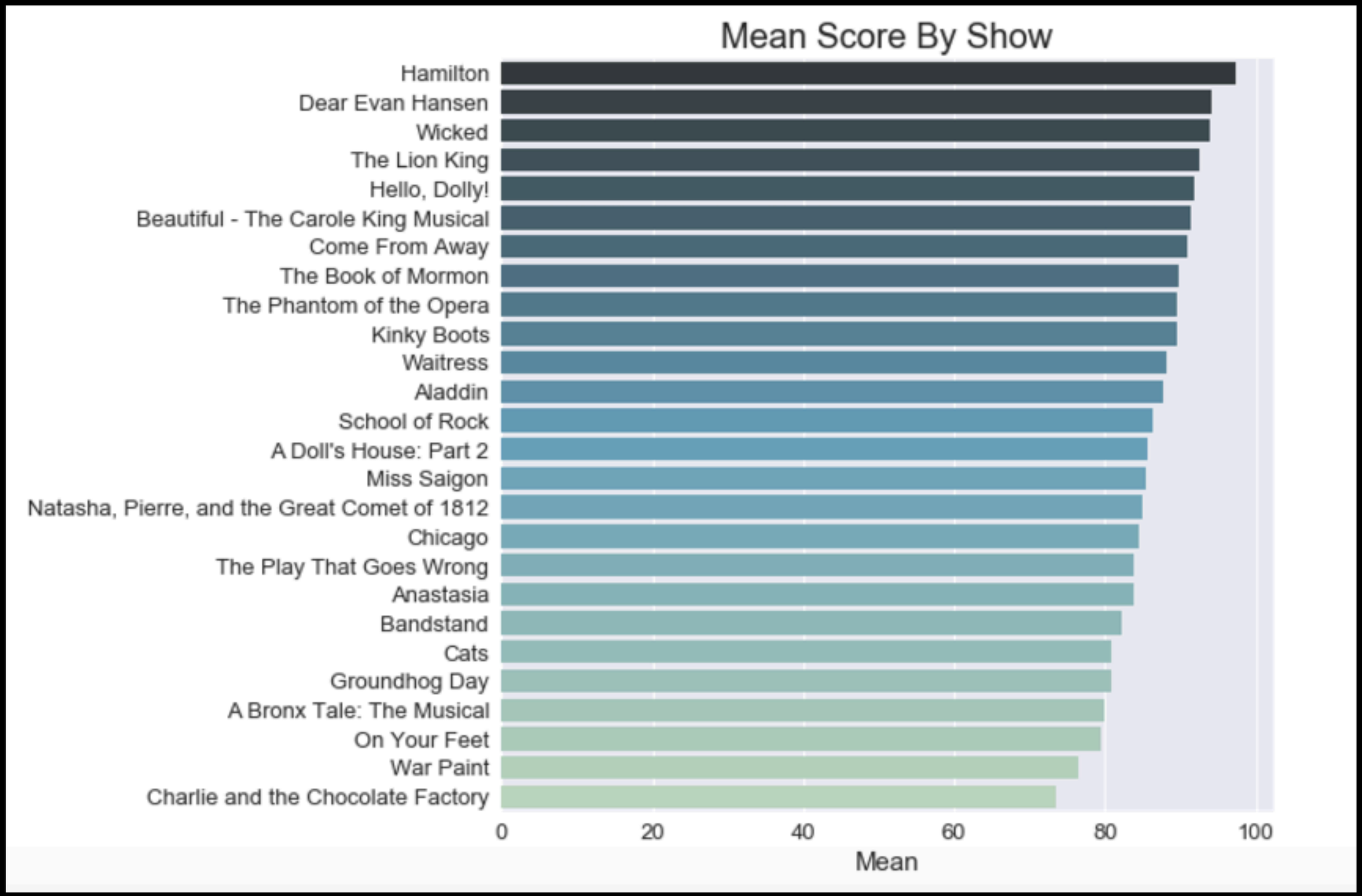

```
            var json = {"status": 0, "data": {"next_performance_time": "2017-09-
02T00:00:00+00:00", "first_date": "2017-08-02T14:00:00-04:00", "performances": [{"date":
"2017-08-02", "times": [{"status": "On Sale", "availability": "Sold Out", "id": 724533,
"time": "2017-08-02T14:00:00-04:00"}, {"status": "On Sale", "availability": "Sold Out",
"id": 724534, "time": "2017-08-02T20:00:00-04:00"}]}, {"date": "2017-08-03", "times":
[{"status": "On Sale", "availability": "Sold Out", "id": 724535, "time": "2017-08-
03T19:00:00-04:00"}]}, {"date": "2017-08-04", "times": [{"status": "On Sale",
"availability": "Sold Out", "id": 724536, "time": "2017-08-04T20:00:00-04:00"}]}, {"date":
"2017-08-05", "times": [{"status": "On Sale", "availability": "Sold Out", "id": 724537,
"time": "2017-08-05T14:00:00-04:00"}, {"status": "On Sale", "availability": "Sold Out",
"id": 724538, "time": "2017-08-05T20:00:00-04:00"}]}, {"date": "2017-08-06", "times":
[{"status": "On Sale", "availability": "Sold Out", "id": 724539, "time": "2017-08-
06T15:00:00-04:00"}]}, {"date": "2017-08-08", "times": [{"status": "On Sale",
"availability": "Sold Out", "id": 724540, "time": "2017-08-08T19:00:00-04:00"}]}, {"date":
"2017-08-09", "times": [{"status": "On Sale", "availability": "Sold Out", "id": 724541,
"time": "2017-08-09T14:00:00-04:00"}, {"status": "On Sale", "availability": "Sold Out",
"id": 724542, "time": "2017-08-09T20:00:00-04:00"}]}, {"date": "2017-08-10", "times":
[{"status": "On Sale", "availability": "Sold Out", "id": 724543, "time": "2017-08-
10T19:00:00-04:00"}]}, {"date": "2017-08-11", "times": [{"status": "On Sale",
"availability": "Sold Out", "id": 724544, "time": "2017-08-11T20:00:00-04:00"}]}, {"date":
"2017-08-12", "times": [{"status": "On Sale", "availability": "Sold Out", "id": 724545,
"time": "2017-08-12T14:00:00-04:00"}, {"status": "On Sale", "availability": "Sold Out",
"id": 724546, "time": "2017-08-12T20:00:00-04:00"}]}, {"date": "2017-08-13", "times":
```


I used ids like these to generate all 4,390 links I needed

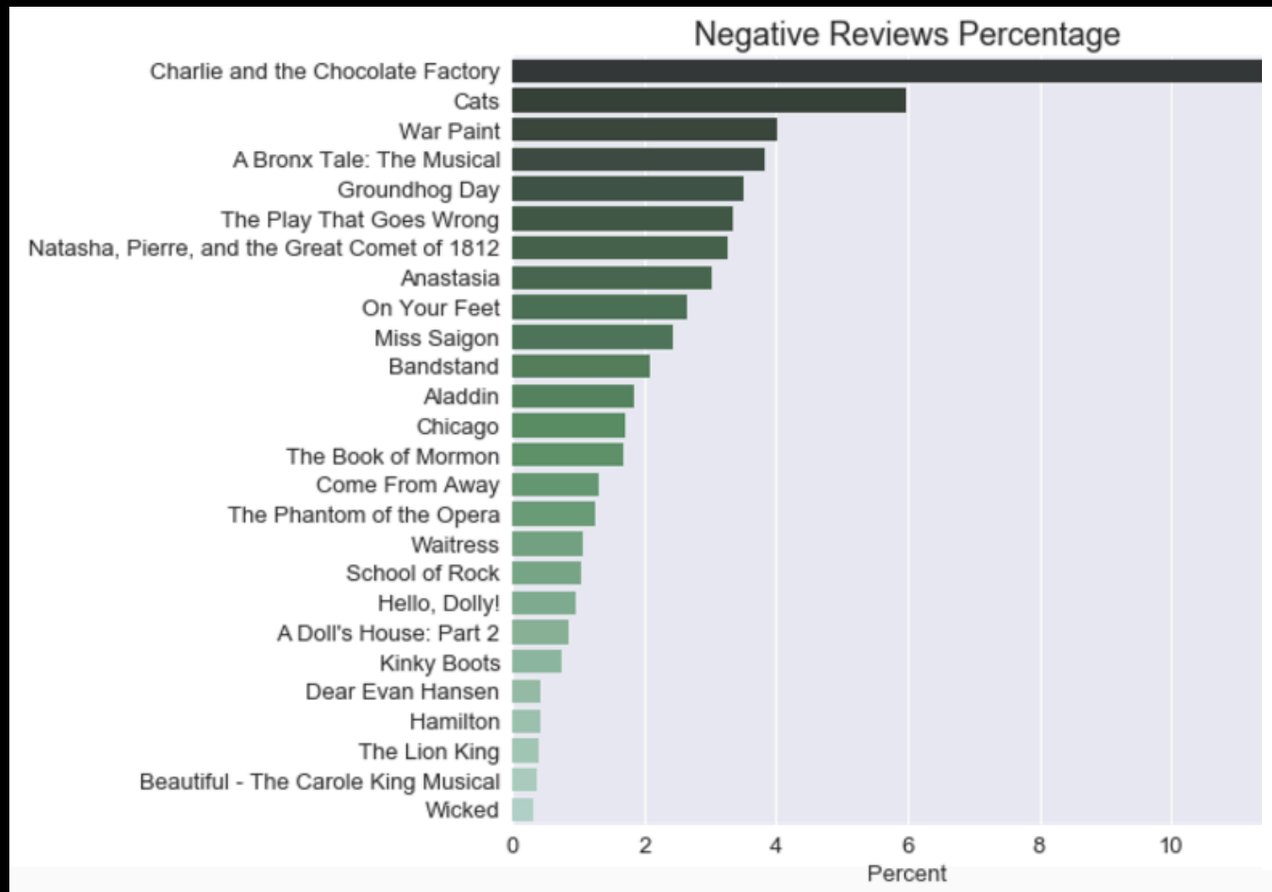
WHAT TO LOOK FOR?

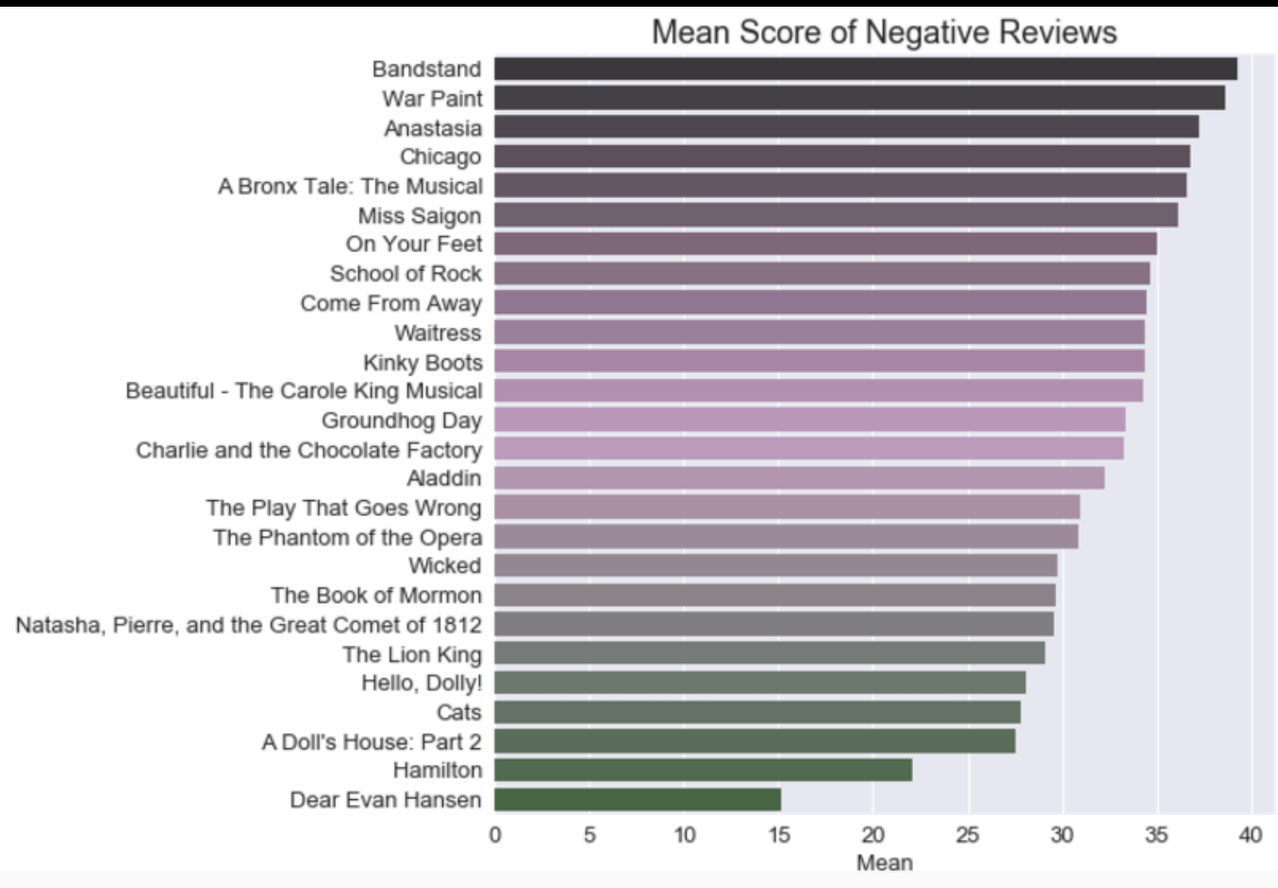
- Explore the scores- more than just the mean
- What word(s) are most used to describe the top shows?
- Price data- any trends/factors explaining the huge difference between the top shows and the rest?
- Price vs Score?

MEAN SCORE


Mean Score By Show
Hamilton
Dear Evan Hansen
Wicked
The Lion King
Hello, Dolly!
Beautiful - The Carole King Musical
Come From Away
The Book of Mormon
The Phantom of the Opera
Kinky Boots
Waitress
Aladdin
School of Rock
A Doll's House: Part 2
Miss Saigon
Natasha, Pierre, and the Great Comet of 1812
Chicago
The Play That Goes Wrong
Anastasia
Bandstand
Cats
Groundhog Day
A Bronx Tale: The Musical
On Your Feet
War Paint
Charlie and the Chocolate Factory
0
20
40
60
80
100
Mean

NEGATIVE REVIEWS





Positive: 70-100

Mixed: 50-69

Negative: 0-49

Lower-score shows are not as heavily affected by negative reviews as I had initially anticipated

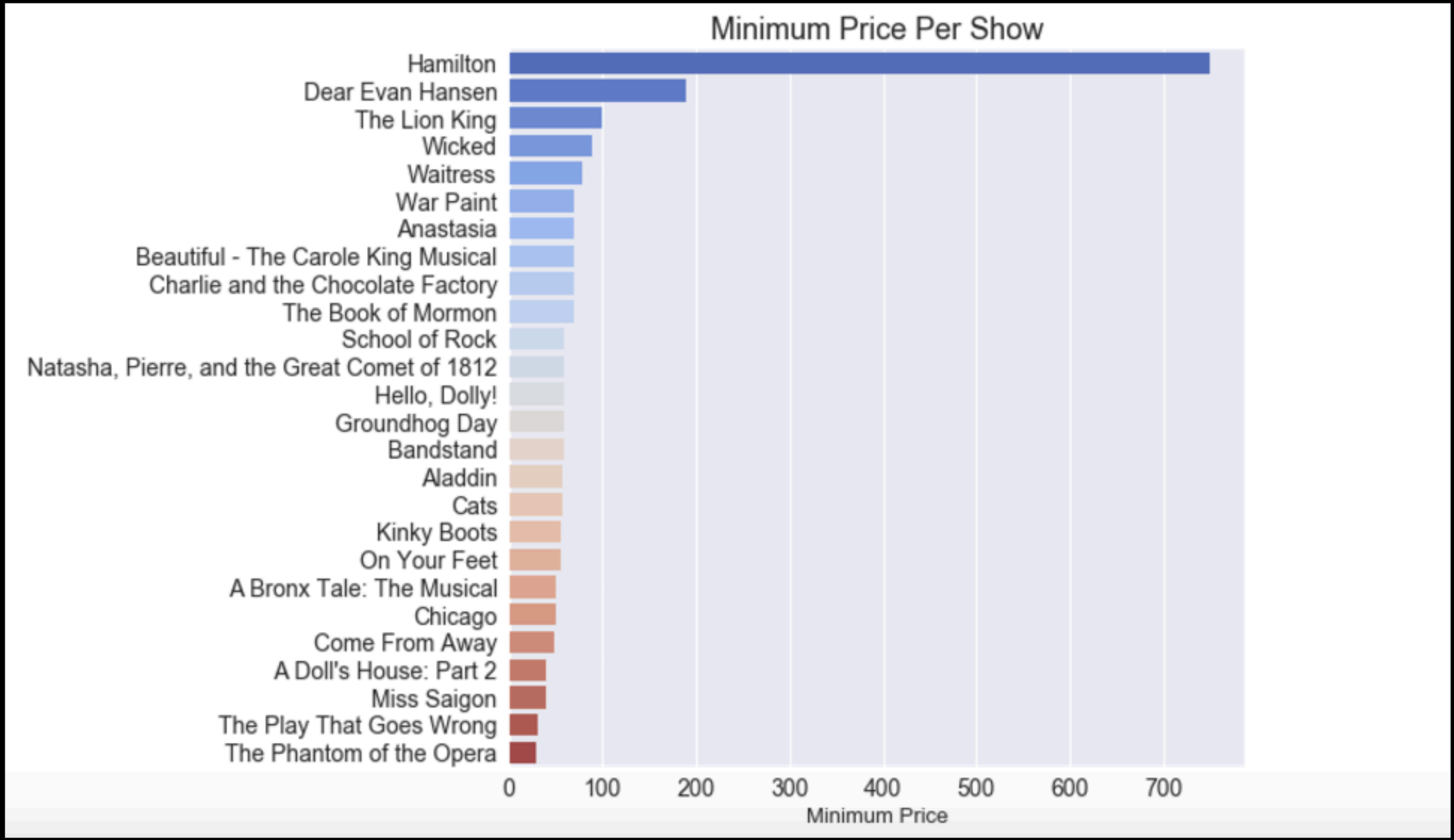
Hamilton

Dear Evan Hansen

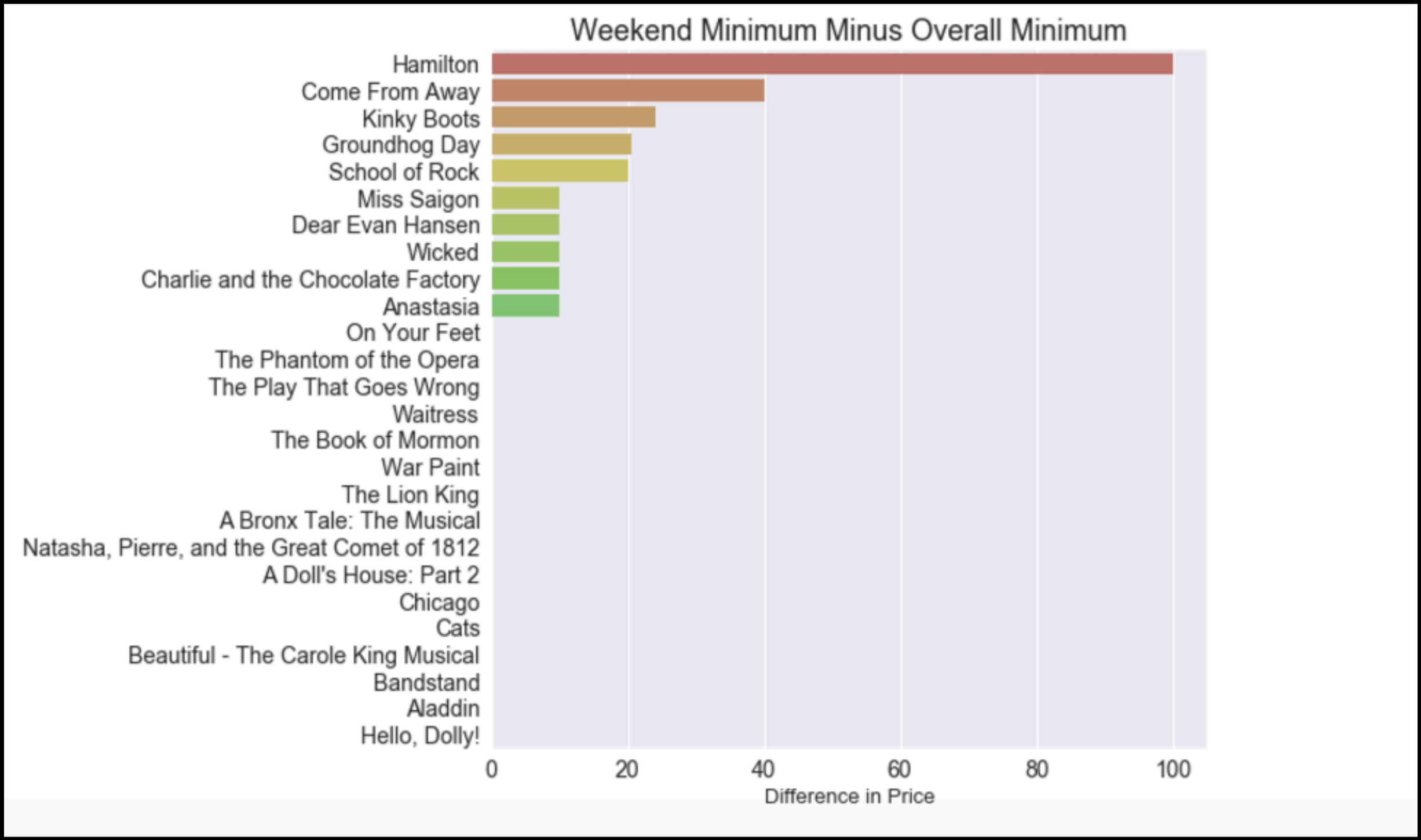
Wicked

The Lion King

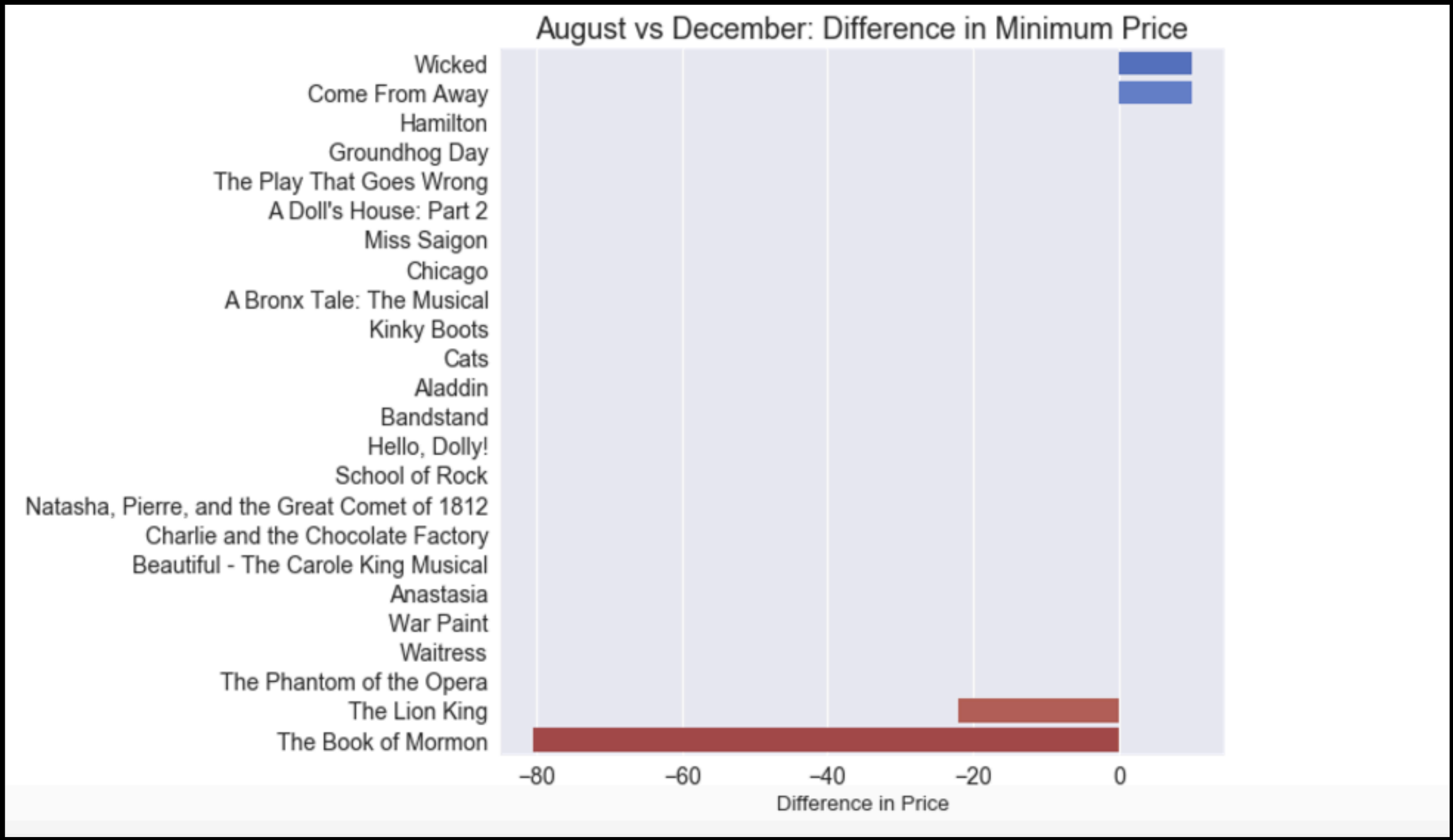
MINIMUM PRICE



WEEKEND MINUS MIN



AUGUST VS DECEMBER



PRICE VS SCORE

- Not what I expecting
- Explanations?
- Price Elasticity of Demand
- Rationing Problem