

ON

BROADWAY



SHOW-SCORE

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

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

SCRAPING SHOW-SCORE

<div> tag containing
each review



The screenshot shows a single movie review card. At the top, it says "Sort by: Default". Below that is a header with the class "div#user_review_246333.member-tile.show-review-tile.user_review" and dimensions "445 x 306". The review itself has a green circular icon with the number "95". To the right of the score is the text "Ambitious, Masterful, Great writing, Entertaining, Great staging". Below the score is the name "Megan M 3" followed by three small icons. A blue button labeled "+ Get Alerts" is below the name. The date "June 12th, 2017" is shown in light gray. The review text reads "See it if you are alive! Don't see it if you don't like things simply because most other people do." Below the review are two buttons: "Helpful?" with a count of "3 votes" and social sharing icons for Facebook, Email, and Twitter. At the bottom is a "Report Abuse" link.

<a> tag containing
href for the next page



The screenshot shows a navigation bar with the number "5" and an ellipsis "...". Below the ellipsis are two buttons: "Next >" and "Last »". A tooltip over the "Next >" button shows the URL "a | 63.06 x 31".

urls followed a
simple pattern
(?page+1)



```
disabled">...</li>
▼<li class="next">
  <a href="/broadway-shows/hamilton?page=2">
    Next ></a> == $0
  </li>
▶<li class="last">...</li>
```

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



OR SELECT BY SECTION

span | 119.34 × 17

Starting at \$199.00

```
[In [5]: response.xpath('//span')  
Out[5]: []  
  
In [6]:
```

???????

```
SectionListItem__body__1qTyz">  
► <span>...</span> == $0  
  </dd>  
  </dl>  
  <div class="BoxListItems__right-aligned-item__rhkyK"></div>  
  </button>  
</div>  
► <div class="row">...</div>
```

the content is being dynamically
generated by javascript



<script type="text/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": [{"status": "On Sale", "availability": "Sold Out", "id": 724547, "time": "2017-08-13T14:00:00-04:00"}]}]
```

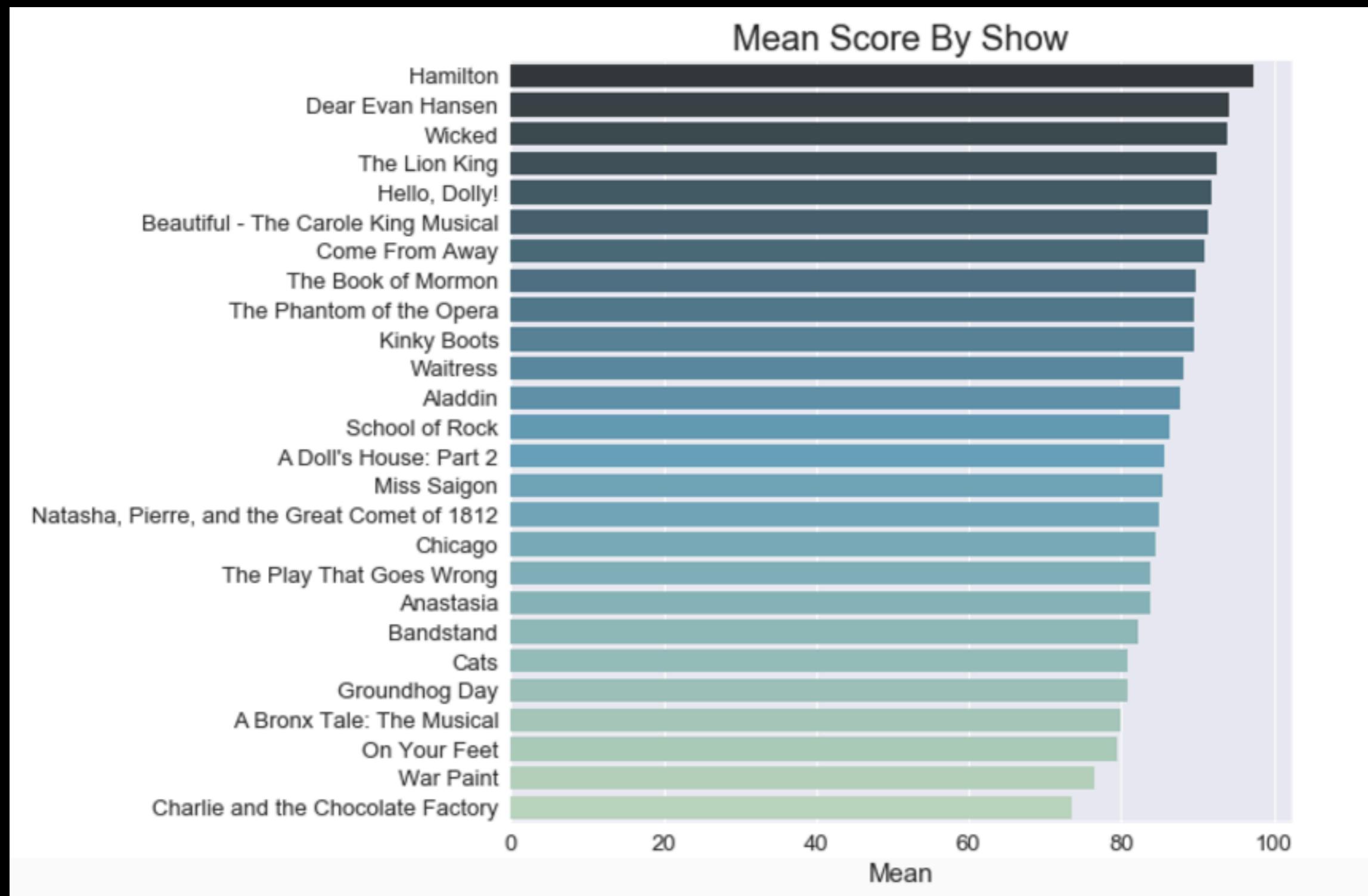
I used ids like these to generate all 4,390 links I needed



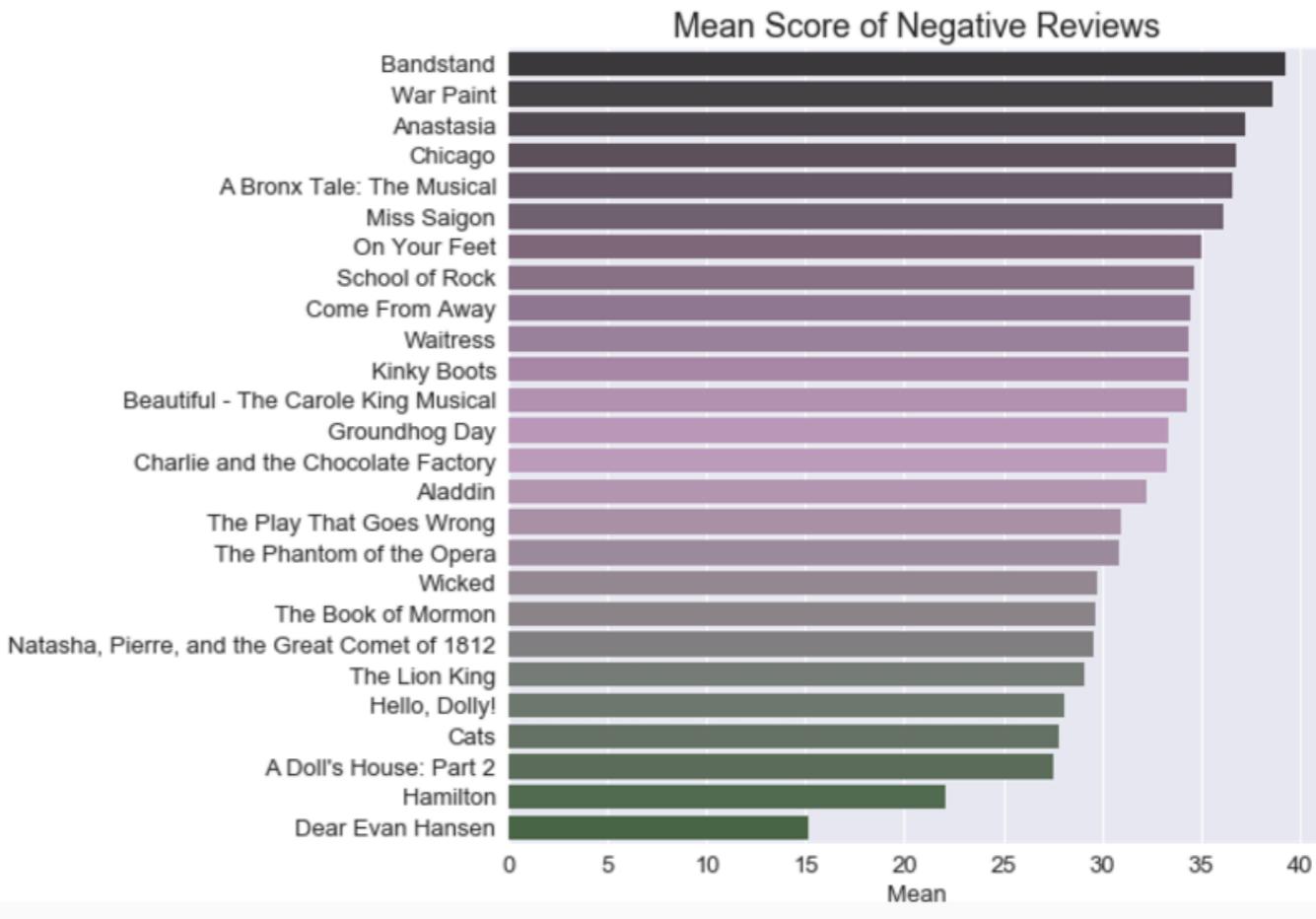
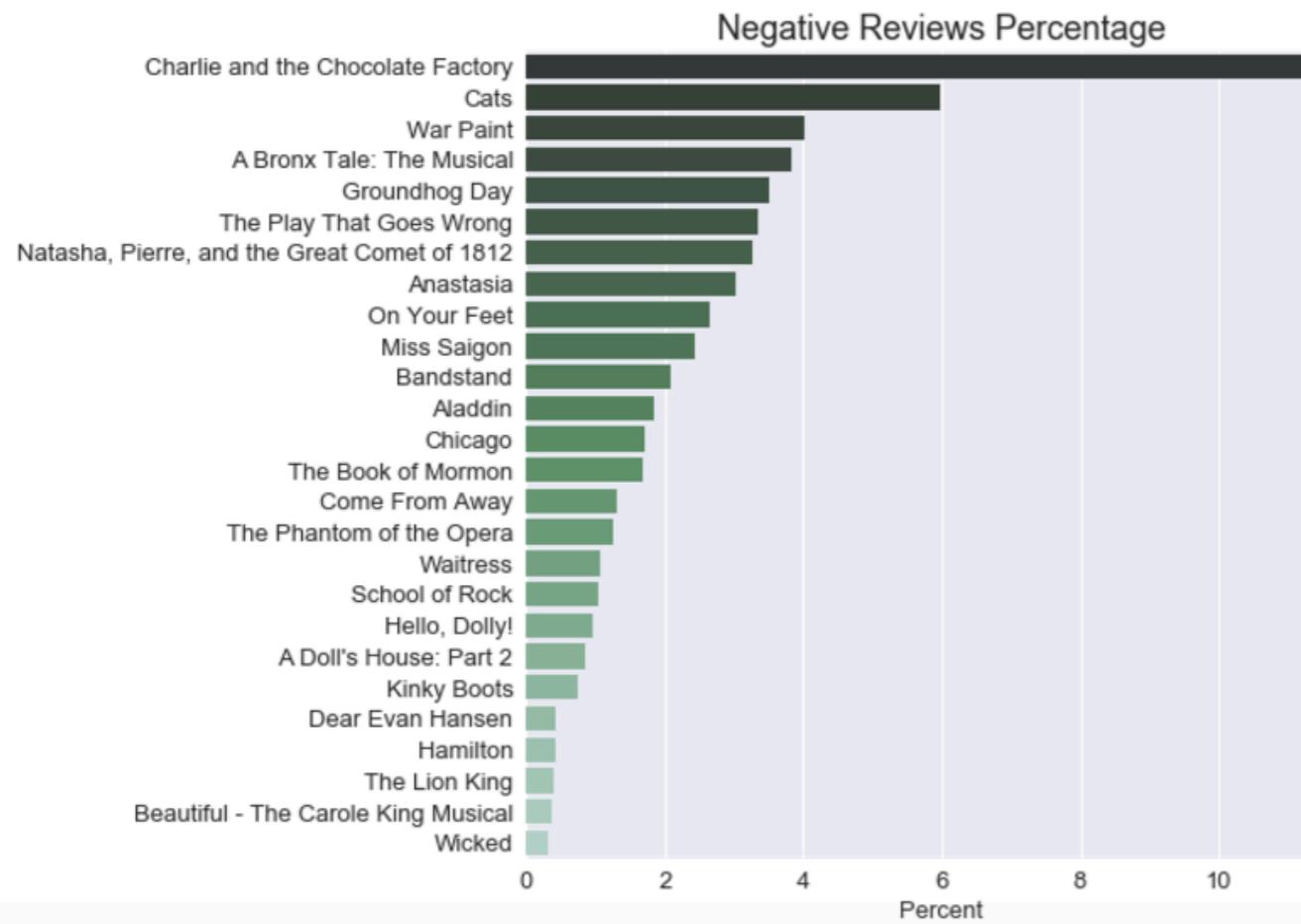
WHAT TO LOOK FOR?

- 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



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

Riveting
Exquisite
Enchanting Intelligent
Brilliant Clever Dizzying
Must see Edgy Ambitious
Absorbing
Original Delightful
Funny Masterful
Great acting Epic
Thought-provoking Entertaining
Entertaining Intense

Dear Evan Hansen

Disappointing Relevant
Delightful Intelligent
Overrated Thought-provoking
Great acting
Absorbing
Funny Clever Refreshing
Ambitious Edgy Epic
Exquisite Intense Profound
Enchanting Must see
Masterful

Wicked

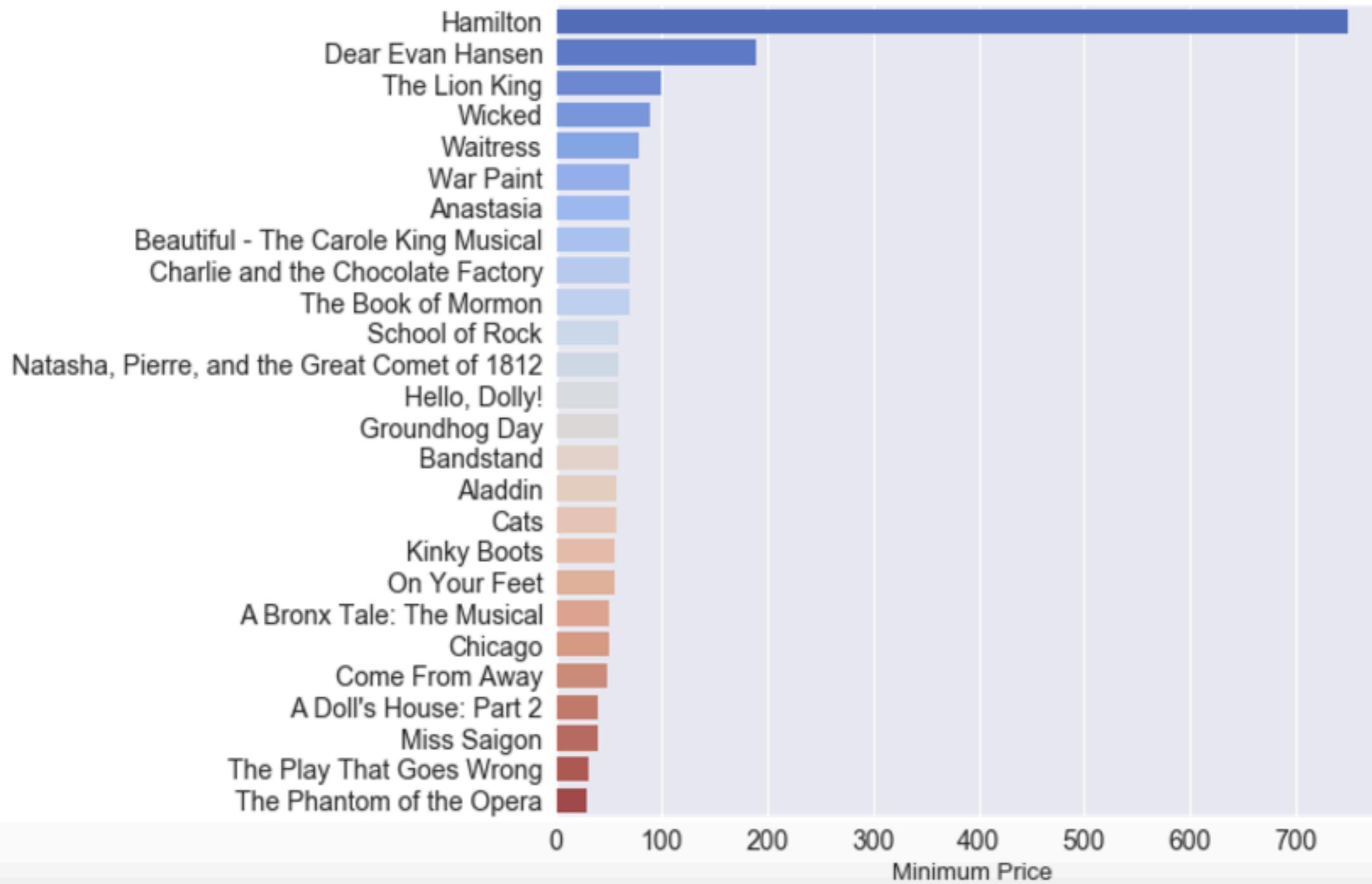
Disappointing Great acting
Thought-provoking Intelligent
Epic Intelligent
Overrated Banal Must see
Masterful
Delightful
Funny
Absorbing
Clever Ambitious
Dizzying Edgy Entertaining
Enchanting Original Exquisite

The Lion King

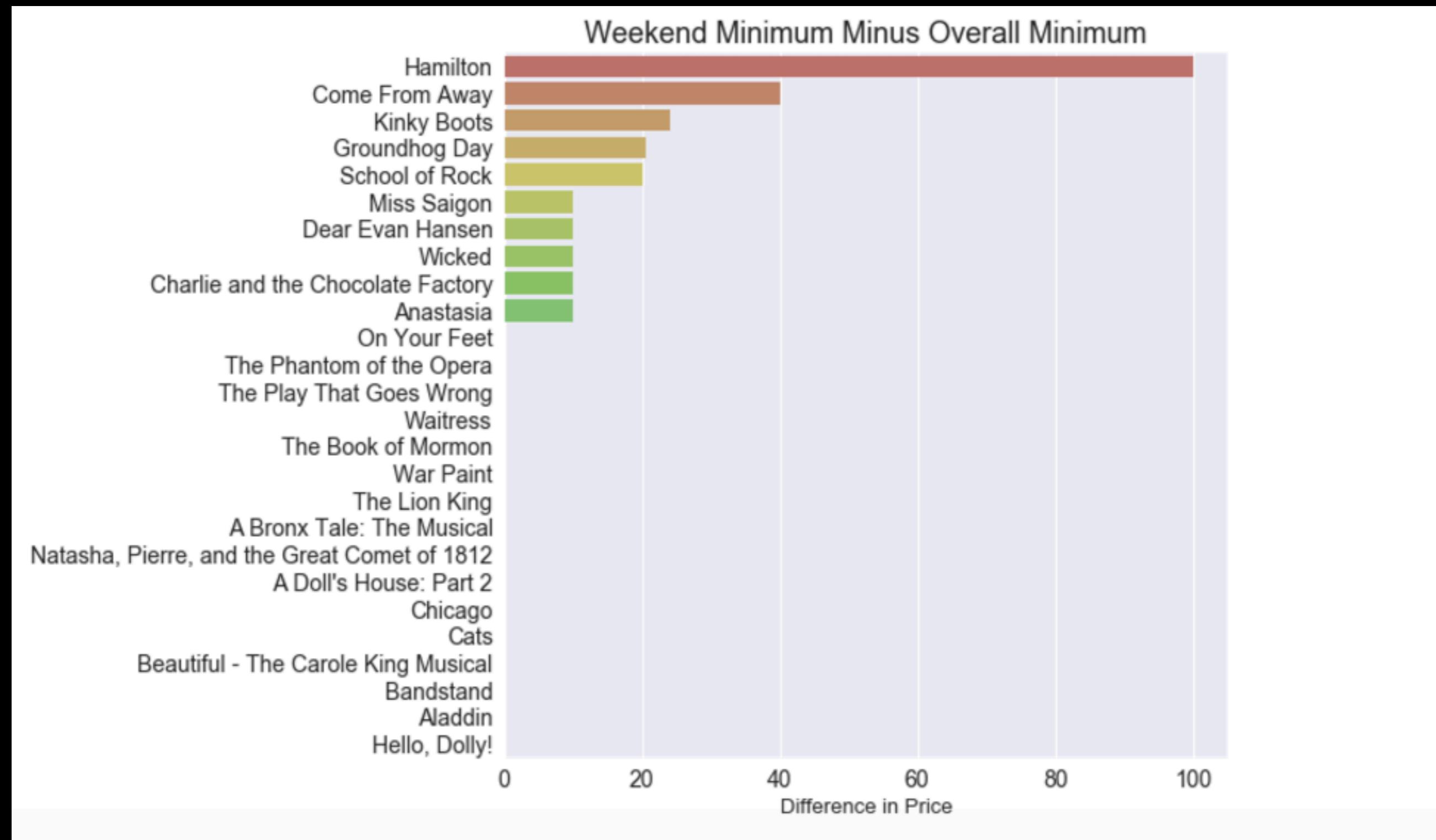
Intense Disappointing
Banal Overrated
Masterful
Funny
Enchanting
Original Clever Dizzying
Beautiful
Absorbing
Delightful
Ambitious Exquisite
Edgy Epic Great acting
Entertaining Must see

MINIMUM PRICE

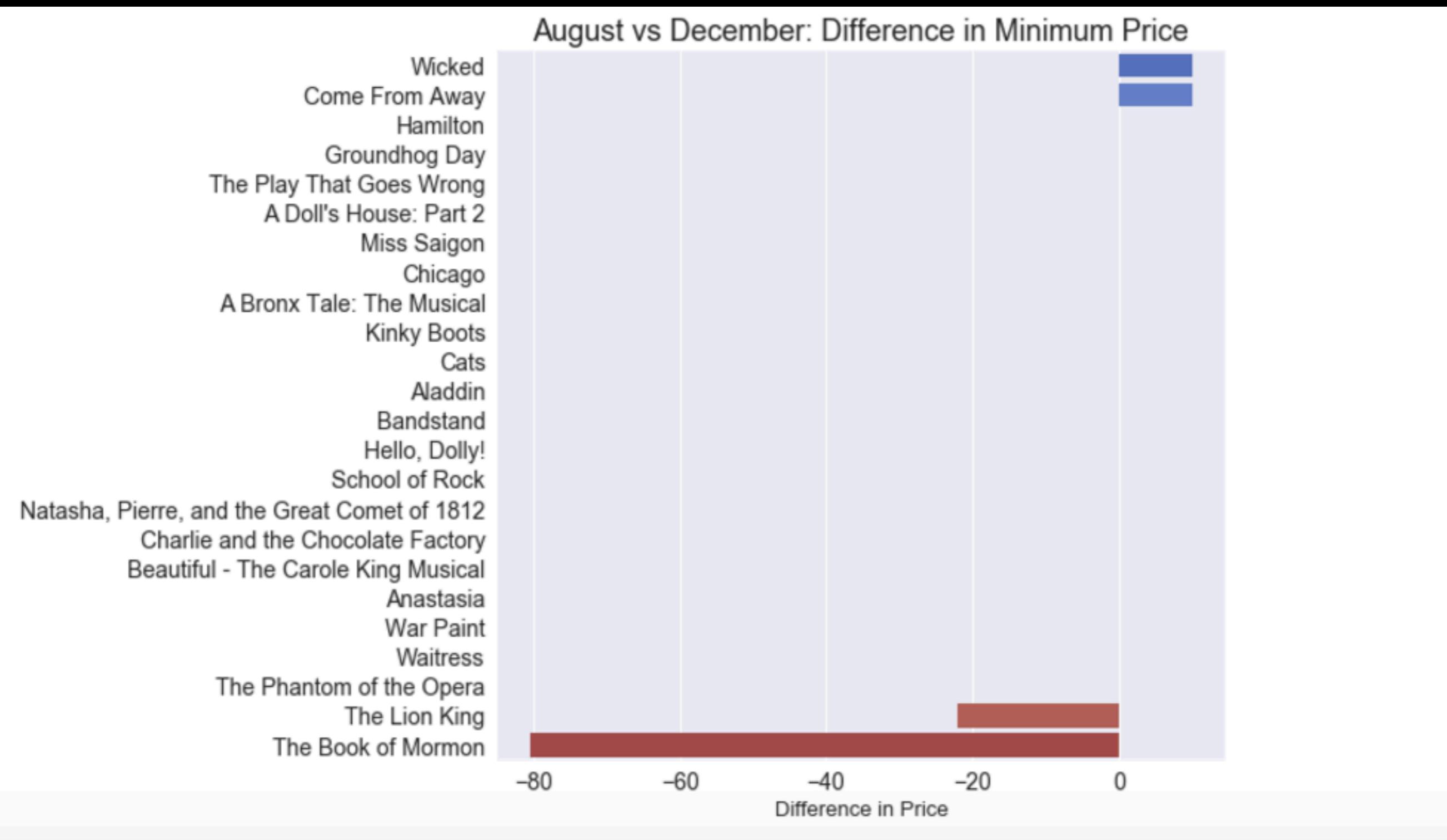
Minimum Price Per Show



WEEKEND MINUS MIN



AUGUST VS DECEMBER



PRICE VS SCORE



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