We’ve long known there is a difference in what people pay for the same experience, but outside of going to the event and asking people, there hasn’t been a way to quantify that. And what is the role of the $5 billion secondary market (such as StubHub) on ticket prices?

As a consumer, what’s the best strategy: buy early? Buy only from the venue? Or does it vary by genre, venue, location, artist, etc.?

**Compiling the data:**

So far I have created three scripts:

* The scraper, which has (to-date) pulled in about 300,000 files, each of which represents anywhere from one to 40 tickets transactions for a particular artist, venue and date. As it turns out, a \*lot\* of these are transactions for parking passes, and not useful for this analysis. Separately, I am scraping an ‘event’ file into SQL. That holds one row = one record for:
* Event ID, artist, event\_date, venue, city, state
* A script that converts the thousands of .js files into a .csv These are known as the ‘ticket files’ which have the basic information:
* Date scraped, days till event, seat location, current price, list price (more on that below), transaction ID, event ID, etc. It also contains some minimal seller information
* A script that joins a .csv version of the event file with the ticket .csv so artist, venue name, etc are combined in memory. I’m also creating a count of the ticket transactions to simplify whether this is the first case of the ticket for sale, the second, etc.

**Analysis – methodology**

Data exploration – descriptive statistics

Total transactions: 1,693

Total tickets available 7,004

Tickets with complete concert cycles (at least 2 weeks of

sales prior to concert: 0

Total events (artist/date/venue) active: 85

Percent large sellers (vs small, though some 'small'

sellers have dozens of ticket): 42

Problem: I can’t verify that ‘list\_price’ is actually the face value. I’ve been going through the developer documentation and it’s unclear: both fields are referenced but I can’t find ‘face value’ as a field I can actually scrape.

Proposed solution: I plan to buy and sell a few tickets and trace the transaction. If ‘list\_price’ is actually the price asked by the seller \*plus\* the fees, the variance will be predictable and I have to find a new source for the face value. This would require choosing a subset of events to track.

**Analysis - creating a model**

Create a code/class for:

* Urban vs suburban vs rural venues. This may include a flag for outdoor venues and/or semi-permanent acts such as Las Vegas shows (‘Legends of Rock!’). *Cluster analysis?*
* Genre of music. (country, pop, rock, alt; perhaps groups that have toured 30+ years-ie geezer rock, etc. Should also have a flag for a ‘hot’ ticket, such as Taylor Swift. I think both would require research unless there’s a way to do it computationally.
* Type of seller. The data includes minimal information such as ‘large seller’ and something called ‘www.’ I can track number of transactions to expand that category or, perhaps use cluster analysis based on #tickets for sale/price of ticket.
* Others, TBD

I will run a linear regression on the variance (continuous) against base variables and separately, the recoded variables (above: venue type, genre, etc.) If there is little explanation of the variance (low R-squared), I will need to create a supervised class for each coded variable, as well as ‘time until event’? Or does it makes sense to create a class of high-variance/not high variance and look for explanations within each class?

**Other questions (to do list):**

Create an actual time series analysis

I’m not sure the best way to measure price variance that occurs \*between\* the start and end point of the transaction. I know there are ways to do this, but I need 1) office hours and 2) more data of transactions that last longer than three weeks. Perhaps more importantly: are there situations where mid-transaction variance \*matters\* in the final sales price?