# Achievement 6 Data Source

## Option 1: NYC OpenData: NYC Parks Event Listing database

Agency: Department of Parks and Recreation

Link: <https://data.cityofnewyork.us/City-Government/NYC-Parks-Events-Listing-Event-Listing/fudw-fgrp/about_data>

Pros:

* Updated October 2023, features data back to 2013.
* Features data dictionary and seems well documented.
* 74.5K rows far exceeds requirement of 1.5K.
  + Kind of a con too, would need to limit the db somehow. Thinking about focusing on pre-during and post COVID, so like 2019-2023 or so.
* Spatial component baked in; contains Borough for categories, lat/longs, addresses, and specific parks for visualizations.
* Time component also available in start time and end time. Could derive event lengths for insights.
* 3+ categorical variables: cost\_free, must\_see, borough, accessible, event category.
* 3+ continuous variables: start\_time, end\_time, cost\_description.
  + this one is a little tighter/ could be better imo
  + Could derive some continuous variables maybe? Event length is the only one I can think of.
* Interesting and doesn’t appear to be played out in other portfolios – would be unique.

Cons:

* Major con off the top is that there doesn’t appear to be any attendance figures.
* Sparse on continuous variables.
  + Could make it difficult to make regression model

## Option 2: AirBnB in Austin TX or Barcelona, ESP

Dataset from Inside Airbnb

Links: <http://insideairbnb.com/get-the-data/>

<http://insideairbnb.com/austin/>

<http://insideairbnb.com/barcelona/>

Pros:

* Updated December 2023. Very up to date
* Interesting to me from Real Estate and travel perspective
* Barcelona data a bit more robust (features licensing statistics, more private rooms, more longer term rentals), but Austin market more familiar to me (I lived there).
  + Barca 18K listings, Austin 15k listings
* Geographic component is obvious.
* Time component: quarterly data is readily available, would need to request older than that.
* Categorical variables: Neighborhoods, private room vs entire home, license vs unlicensed
* Continuous variables: price, occupancy(?), income, review score

Cons:

* Again, short on continuous variables. Could make options for modeling more difficult.

## Option 3: Medical Insurance datasets

Dataset from Kaggle

Link: <https://www.kaggle.com/datasets/harishkumardatalab/medical-insurance-price-prediction>

Link: <https://www.kaggle.com/datasets/thedevastator/insurance-claim-analysis-demographic-and-health>

Link: <https://www.kaggle.com/datasets/mirichoi0218/insurance>

Seen a bunch of versions of this kind of study, linking health factors and insurance payouts, but this one is relatively recent and seems to be well documented. From an R textbook

Pros:

* Applicable for getting a job in healthcare industry, which I have family connections in.
* Regions check off the geographic component.
  + Not as robustly as the aforementioned options though.
* Plenty of continuous variables:
  + BMI, blood pressure, Charges
* Categories: Children, Gender, Regions, age (range), Smokers, etc.

Cons:

* The major con with this one is that its lacking a “time” component. Yes, there’s age but that’s not really the same thing.
* A bit of a textbook example. Wouldn’t necessarily distinguish me from the pack.