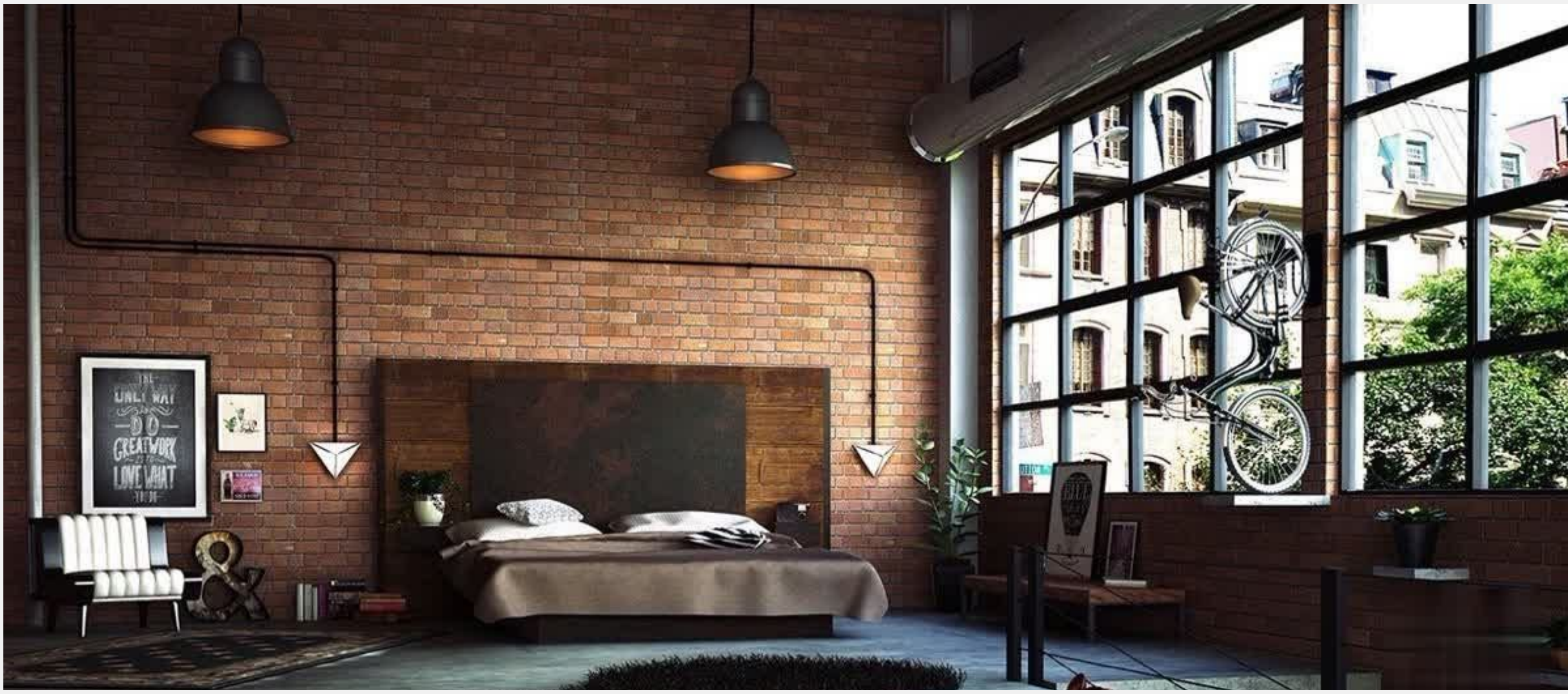
THE
ONLY WAY
DO
GREATWORK
LOVE WHAT
YOU DO

MILESTONE 1 : PROJECT PROPOSAL

1. LIST OF GROUP MEMBERS

NAME	EMAIL	GITHUB
Ramanpreet Bhatia	rpbhatia@seas.upenn.edu	rpbhatia
Jasmine Jian	yuejian@seas.upenn.edu	JasmineYJ
Cayde Roothoff	croot22@seas.upenn.edu	croot22
John Hentrich	hentrich@seas.upenn.edu	johnhentrich

2. WEBSITE IDEA

- Find the preferred zip code to live in and historical listings in that zip code based on housing cost (Real Estate transfers) and factors important to the user, e.g.: educational quality, lifestyle conveniences, and safety.
- Compare historical listings to the FMV for similar properties

3. SELECTED DATA SETS

A. Lifestyle:
   a. School Information:
      i. Description: Data from the School District of Philadelphia that includes metrics on school performance, location, and overall operations.
      ii. Link: link
      iii. Query:
         1. Select and count all schools (High School, K-8) within a given zip code
         2. Calculate average school rating (High School, K-8) within a given zip code
      iv. Size Statistics:
         1. Space: 700 kb
         2. Rows: 322
         3. Attributes: 293
      v. Summary Statistics (for School Ratings Attribute):
         1. Mean: 36.9
         2. Standard Deviation: 19.1
   b. Cuisine based on user interest & Top Review from Yelp
      i. Description: Data from yelp containing the details regarding local business and user reviews, further details regarding attributes can be found here.
      ii. Link: Yelp Dataset : business.json JOIN review.json

- iii. Query:
    1. Ask user to select a type of cuisine
    2. Recommend a restaurant that is opening at the moment, with good average rating and close to the zip code that user's looking at.
- iv. Summary Statistics for Business.json :
    1. Space: 9MB out 153MB (limited to PA)
    2. Rows: 12376
    3. Attributes: 66
- v. Summary Statistics for Review.json :
    1. Space: 6.33GB (will be limited to PA after parsed and joined with Business.json)
    2. Rows: 8021122
    3. Attributes: 9

c. Things to do when & Photo from Yelp

- i. Description: based on the intended schedule, recommend activities available nearby, further details regarding check-in can be found here, reuse dataset from above.
- ii. Link: Yelp Dataset : business.json and check-in.json JOIN ON business_id
- iii. Query:
    1. Ask user to select a range of time
    2. Recommend activities for the given time with good average rating and close to the zip code that user's looking at. Potentially we will attach photos of the business.
- iv. Summary Statistics for CheckIn.json:
    1. Space: 26MB out of 450MB (limited to PA)
    2. Rows: 37127
    3. Attributes: 2

B. Safety:

a. COVID Test Cases:

- i. Description: A list of COVID test counts (positive and negative) per zip code.
- ii. Link: link
- iii. Query:
    1. Count the number of positive test cases per zip code
    2. Calculate the average positive test rate per zip code
- iv. Size Statistics:
    1. Space: 5 kb
    2. Rows: 126
    3. Attributes: 5
- v. Summary Statistics (for Positive Test Cases per Zip Code):
    1. Mean: 2310
    2. Standard Deviation: 1434

C. Real Estate Transfers:
   a. Description: A set of houses sold in 2020 containing the appraised and fair market value.
   b. Queries
      i. Select houses by zip code that match a certain price range
      ii. Find the average of a zip code or multiple zip codes
   c. Housing Costs: link
   d. Rows 140,198
   e. Space 56.2 MB
   f. Attributes 12

4. OPTIONAL FUTURE DATA SETS

A. Lifestyle:
   a. Pre-Ks: Count per zip (JOIN ON zip): link
   b. Parks: Count per zip (JOIN ON zip): link
   c. Historical Landmarks: Count per zip (JOIN ON zip): link
   d. Farmer's Market Locations: Count per zip (JOIN ON zip): link

B. Safety:
   a. Crime Incidents: link
   b. Police Station Locations: link
   c. Fire Department Locations: link
   d. Hospital Locations: link

C. Current Listings
   a. Zillow, CraigsList, Trulia for Philly - link

D. General:
   a. Other Census Data: link