Recommendation of OFF-CAMPUS housing for international students based on preferences

BY GROUP 4 – BADREESH, ISHAN, SAMAR, YASH L

TOPIC AND QUESTIONS

The topic we chose to go with is "***Recommendation of OFF-CAMPUS housing for international students based on preferences***". The reason for the same is a lot of students, more specifically international students, often have trouble looking for relevant housing. This would be a useful tool not just for international students but also for others as it will analyze where students tend to stay and the location where to stay; brokers don't share the data regarding safety, as they are only interested in the broking fees.

The business problem is challenges faced by international students with respect to off-campus housing based on preferences. Therefore, the unique value proposition for this task on recommendation is based on individual level of analysis based on preference.

Recommendation preferences will be based on price, location, and size of unit.

Recommendation for new students with respect to off-campus housing locations, the data wrangling will mainly comprise Northeastern students' off-campus housing location and situation in Boston. It analyses the spending cost with respect to rent, household spends and the kind of majority demographics located in each neighborhood

Why is it important

This is important to analyze the best location for students to stay and the recommendations for the same to help them with the kind of location individuals would like to stay in.

Who will be interested

Students on a budget as well as those who would like to reside in a neighborhood following on their liking and preferences. For eg: People from a specific community would like to stay in an area where the maximum number of people from the community reside.

Who is the analysis ultimately for

The analysis is mainly for students, to help with recommendation on the neighborhood based on liking and preference. For instance, students would like to stay in a safe neighborhood where the crime rate is low to none and has restaurants and grocery stores nearby.

DATA COLLECTION AND SOURCES

The datasets that were utilized were of 2 different types to help answer the business problem, this included the primary dataset which was survey from students currently staying and secondary dataset which was a combination of 3 public dataset and 1 web scraped dataset from Zillow. The amalgamation of these dataset assisted with creation and inference on recommendation for the best spots for OFF-CAMPUS housing for international students.

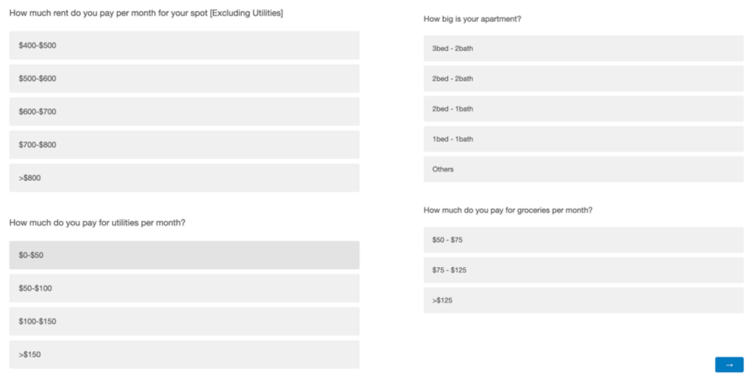
Data Sources include the following

PRIMARY DATASET (SURVEY – QUALTRICS)

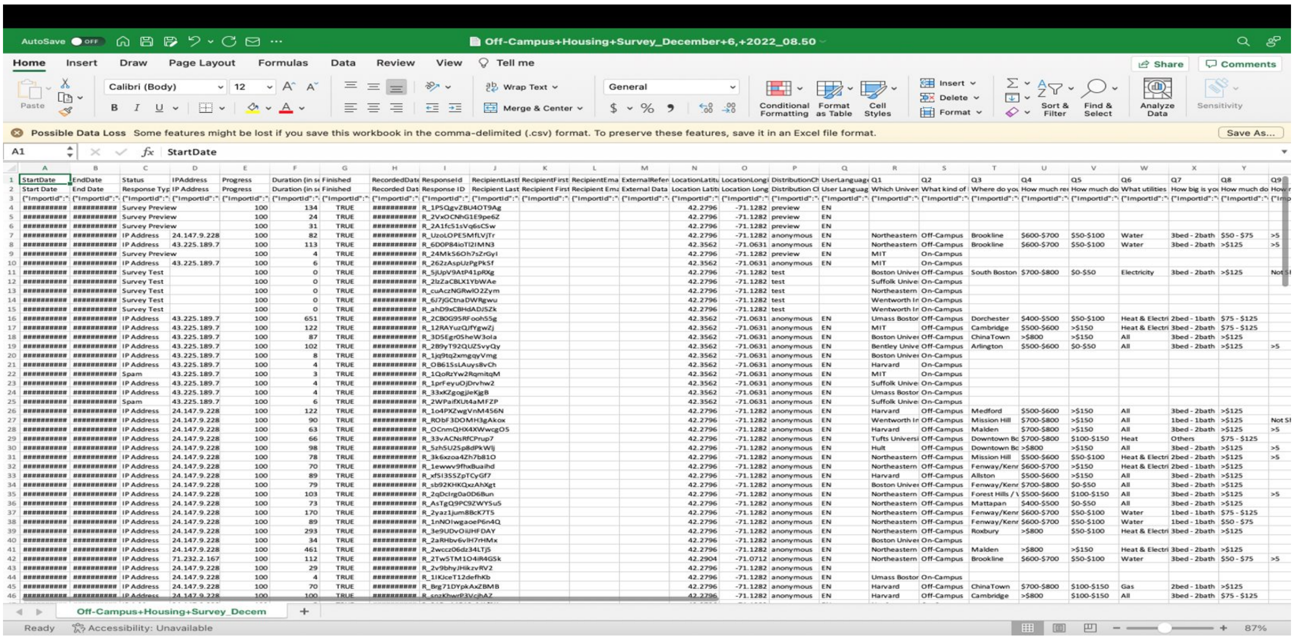
URL:https://qfreeaccountssjc1.az1.qualtrics.com/jfe/form/SV_exPoRUCmgFKab2u

Questions asked in the survey:

- Which University you go to?
- What kind of housing do you live in?
- Where do you stay? (Eg: Brookline, Jamaica Plain, Roxbury, etc.)
- How much rent do you pay per month for your spot [Excluding Utilities]
- How much do you pay for utilities per month?
- What utilities do you have included in your rent?
- How big is your apartment? (3b2b/2b2/2b1b)
- How much do you pay for groceries per month?
- How many people do you share your apt with?
- Do you have a private room?
- What kind of house are you staying in? (Condo/Townhouse/apartment)
- What is your ethnicity? - Selected Choice
- What is your ethnicity? - Others – Text
- How do you self identify?
- What mode of transport do you travel from
- How far do you stay from university?
- Is your university's shuttle service (Eg: RedEye for Northeastern) accessible to you?
- On a scale of 1-5 , How important is it to have restaurants/bars around you?
- Name your favorite restaurant where you stay?
- How many grocery stores are near you?



SURVEY RAW DATA EXPORTED TO EXCEL (162 rows, 37 columns)



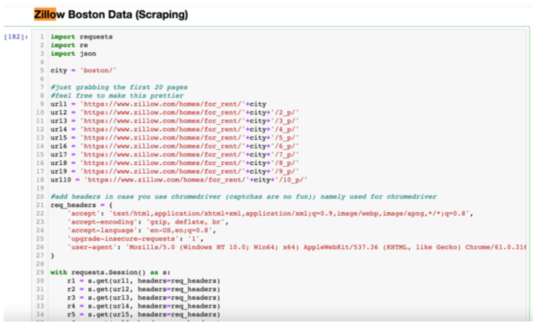
SECONDARY DATASETS (ANALYZE BOSTON – CRIME, HOSPITAL, RESTAURANT)

URL:

1. Crime Dataset - 58429 rows and 17 columns
   a. https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
2. Hospital dataset - 3030 rows and 14 columns
   a. https://data.boston.gov/dataset/hospital-locations
3. Restaurant dataset - 26 rows 7 columns
   a. https://data.boston.gov/dataset/active-food-establishment-licenses

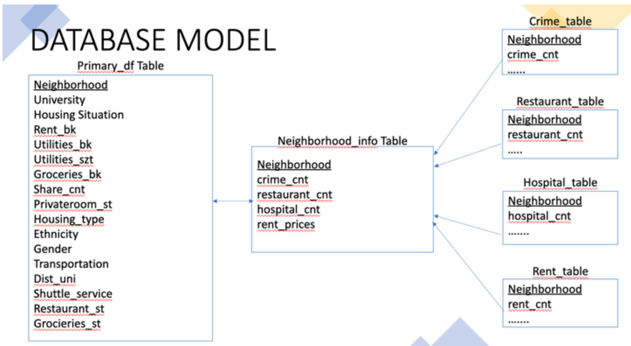
WEB-SCRAPING (ZILLOW's Housing Dataset - 400 rows and 47 columns)

For Scraping Zillow Data for Boston Housing Data, requests package was used to get all the scapred data row and columns. The code was used to get 10 pages of Boston Rent Data. It resulted in 400 rows and the rent price data accordingly for each apartment. Below is a snippet for scraping. More details can be found in the Ipynb file.



- Raw Data Columns - 'zpid', 'id', 'lotId', 'imgSrc', 'hasImage', 'statusType', 'statusText', 'detailUrl', 'latLong', 'units', 'variableData', 'badgeInfo', 'buildingName', 'isBuilding', 'address', 'addressStreet', 'addressState', 'addressCity', 'addressZipcode', 'providerListingId', 'canSaveBuilding', 'has3DModel', 'isFeaturedListing', 'isSaved', 'list', 'relaxed', 'area', 'availabilityDate', 'baths', 'beds', 'countryCurrency', 'hasAdditionalAttributions', 'hasVideo', 'isHomeRec', 'isUndisclosedAddress', 'isUserClaimingOwner', 'isUserConfirmedClaim', 'isZillowOwned', 'pgapt', 'price', 'sgapt', 'shouldShowZestimateAsPrice', 'unformattedPrice', 'zestimate', 'streetViewMetadataURL', 'streetViewURL', 'best_deal'

DATA PROFILING AND INFORMATION QUALITY



- Most of the data here is of excellent quality but a few columns, especially in the Boston building database.
- Missing data will be cross-referred to a similar data set in hand if the missing values are less. If more, we will have to wrangle the data or search for a different dataset from where we can scoop the data out.
- Missing values can be 0's and Nulls can be dealt with by using python.
  As for credibility, most of the data is credible as it comes from the public database. But the survey will definitely be of less credibility as survey data can always be fake or just input by the person doing the survey for the fun of it.
- Data Profiling ssisted with creation of DATABASE MODEL to help in connecting the dots and doing some validation checks

METHODS AND TOOLS USED FOR DATA WRANGLING PROCESS

Python tools include

- Pandas for Data Wrangling and Analysis of Crime, Hospitals.
- Requests for web scraping rent data for Boston through Zillow Website (https://www.zillow.com/homes/for_rent/boston).
- Geopy for getting the Latitude, Longitude of places mentioned in the crime locations, hospital locations, restaurant locations.
- Plotly-Dash for plotting bar graphs of survey data.
- Streamlit for creating the Webapp and Heroku to host the website.
- RapidFuzz for similar string matching of Restaurants based on User Input.

- Folium for building Maps and populating locations on the Webapp

Excel for manipulating and formatting the data.

Tableau for visualizing the results. Tableau was chosen by us as the last stop for data, from where we could derive inferences as Tableau offers a very user-friendly interface and is super-fast and powerful. We majorly deployed multiple basic visualizations in a dashboard and on clubbing them together in a view, we can get a very clear picture of what/where we want to be.

All our visualizations were made interactive and as filters, so that when clicked upon a certain attribute, that attribute serves as a filter to every other visualization on the dashboard and spits out results based on it. Looking at all of these together, we can derive inferences about a lot of things in the area chosen.

The problems that we came across were that of changing data types due to Tableau's in-built data processing tool called "Tableau Prep". This tool as an example, changed Latitude to a Numeric data type than a Geographical data type. A few fields were set as dimensions whereas they should've been Measures. Which merging multiple data sources, the Primary Keys were identical but not same. Challenges Python include in Latitude and Longitude there were many places where we can convert it to suitable format before mapping it on folium. So various operations like adding – to variables to get the correct information.

DATA WRANGLING PROCESS

Our Extract Transform Load (ETL) process comprised of scouting out our target dataset and based on it, creating a survey which could complement the primary dataset.

This comprised our Extraction process.

All these data sets were then cleaned and transformed into a usable format, and it was made sure that "Neighborhood" would be our Unit of Analysis and our Foreign key which shall link all data sets together based on a one-to-many cardinality.

This comprised our Transformation Process.

Lastly, we used StreamLit and Tableau to derive inferences from the Transformed data, by injecting it into the latter and filtering based on our needs and constraints.

STEPS FOR DATA WRANGLING PROCESS (McGregor, 2021).

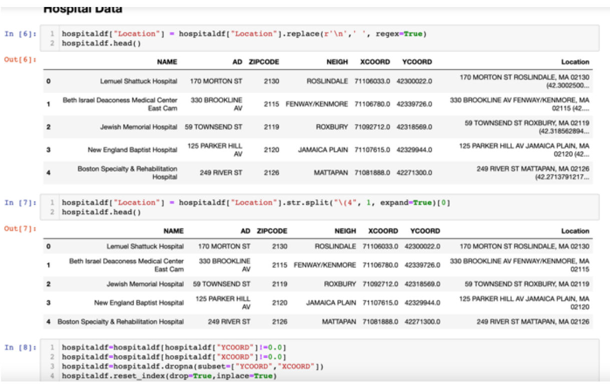
- DATA COLLECTION
- DATA PREPROCESSING
  - DATA FORMATTING
    - Unit level of analysis – Individual level

- Structured Data – Primary dataset
- Unstructured Data – Secondary Dataset
- Normalized Irregular Names - Validation check
- Corrected incorrect entries
- Check for cardinalities
- Mitigated data discrepancies
- Renaming of columns
- Normalized names to match a common value (Eg: “data” and “Data” are the same, so we normalized both as ”Data” - Validation check

- DATA PROFILING
  - Understanding the data through EXCEL and Python files loaded
  - Assisted with creation of DATABASE MODEL
- DATA CLEANING
  - Removing Null and Missing Values from both Primary and Secondary Data set.

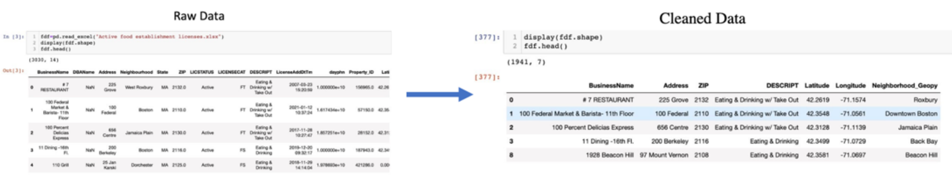
    Normalizing and standardizing data into a single acceptable format.

    Using feature engineering deriving a few values



    Converting parameters from one data type to another for better tangibility and usage.

    Data scraping to generate more accurate results.
  - Reverse encoding using Geopy to Neighborhood
  - PRIMARY DATA
    - Usage of alias in TABLEAU for purpose of dashboarding visualization
  - SECONDARY DATA
    - ANALYZE BOSTON (PUBLIC DATA)
      - Columns cleaned from RESTAURANT DATA: 'BusinessName','Address', 'ZIP','DESCRIPT','Latitude', 'Longitude', 'Neighborhood_Geopy'



- Columns cleaned from HOSPITAL DATA : 'NAME', 'XCOORD', 'YCOORD', 'ZIPCODE', 'Location', 'Neighborhood_Geopy'

- Columns cleaned from CRIME DATA: 'INCIDENT_NUMBER', 'OFFENSE_DESCRIPTION', 'SHOOTING', 'OCCURRED_ON_DATE', 'MONTH', 'DAY_OF_WEEK', 'HOUR', 'STREET', 'Lat', 'Long', 'Neighborhood_Geopy'

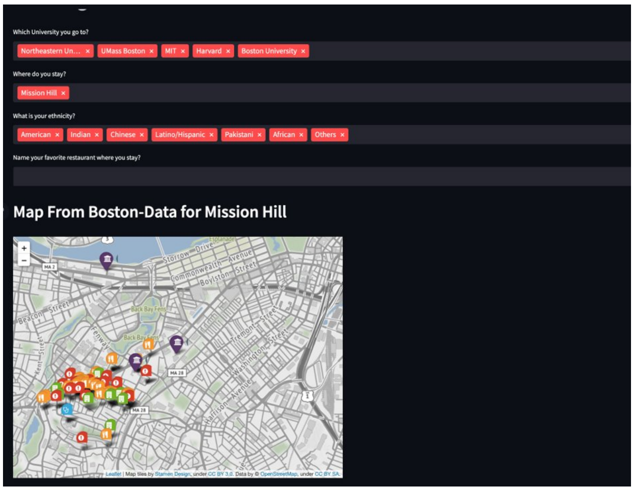
- Columns cleaned from HOUSING DATA

	id	address	beds	baths	area	price
273	318226130	1650 Commonwealth Ave UNIT 606, Brighton, MA 0...	3.0	2.0	1440	$5,500/mo
332	59166982	201 Newbury St APT 508, Boston, MA 02116	2.0	2.0	1126	$6,000/mo
207	59177974	3 bed, 2.0 bath, 1100 sqft, $5,400, 11 Bay Sta...	3.0	2.0	1100	$5,400/mo
335	59104307	28 Sumner St, Dorchester, MA 02125	7.0	2.0	2509	$7,675/mo
212	59177983	23 Bay State Rd APT 6, Boston, MA 02215	2.0	1.0	813	$3,200/mo
211	59167365	25 Bay State Rd, Boston, MA 02215	2.0	1.0	1000	$3,200/mo
288	59094576	39 Snow St, Brighton, MA 02135	4.0	2.5	1100	$4,500/mo
319	59178532	501 Beacon St APT 9, Boston, MA 02215	1.0	1.0	575	$2,875/mo
263	59096315	(undisclosed Address), Brighton, MA 02135	5.0	4.0	2800	$7,200/mo
226	59167359	17 Bay State Rd, Boston, MA 02215	4.0	2.0	999	$6,200/mo
247	2104772572	263 Shawmut Ave #4, Boston, MA 02118	1.0	1.0	535	$2,800/mo
277	333606922	36 Montvale St #1, Boston, MA 02131	3.0	2.0	1541	$3,200/mo
320	59130788	64 Weld Hill St #2, Jamaica Plain, MA 02130	2.0	1.0	1029	$3,200/mo
316	59095012	370 Chestnut Hill Ave, Brighton, MA 02135	1.0	1.0	567	$2,650/mo
168	2096507205	346 Hyde Park Ave #1, Roslindale, MA 02131	3.0	1.0	1200	$2,850/mo
355	59094977	374 Chestnut Hill Ave APT 34, Brighton, MA 02135	1.0	1.0	567	$2,650/mo
256	81853945	1810 Dorchester Ave APT 8, Dorchester, MA 02124	1.0	1.0	450	$1,900/mo
128	59089414	Studio, 1.0 bath, $1,500, 1673 Commonwealth Av...	0.0	1.0	400	$1,550/mo
218	2060511654	7 Barrows St #5C, Allston, MA 02134	0.0	1.0	None	$1,800/mo
185	2118413674	1455 Commonwealth Ave APT 202, Boston, MA 02135	1.0	1.0	450	$1,825/mo

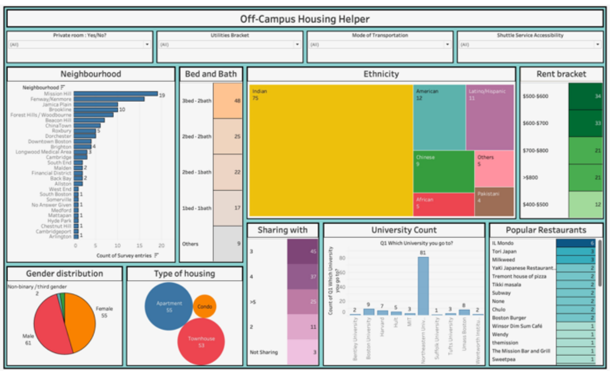
- DATA ANALYSIS
  - We have performed a comprehensive analysis of every attribute that has been included in our survey, along with our other secondary datasets.
  - Tableau has been deployed to perform individual level analysis of parameters to be finally clubbed together to form a bigger picture which shall give us the inference that is desired.
  - Filtration is something that our analysis heavily relies upon, setting conditional filters for each and every attribute is imperative to gettign the desired result.

- We leverage data analysis from parameters like; Rent Bracket, Ethnicity, Utilities Included etc to enable the user to have a highly granular and customized output, as well as crime data

- VISUALIZATION FROM STREAMLIT
  - Website: https://housing-streamlit-heroku.herokuapp.com/
  - Crime, Hospital, Restaurant Visualization based on user input for neighborhood
  - Zillow Price Heatmap around the locality
  - Crime Shooting Heatmap around the locality
  - Count of Crimes, Hospitals, Restaurants



- VISUALIZATION FROM TABLEAU
  - Demographic visualization based on survey input data.
  - Ethnicity concentration based on survey data
  - Conditional filtering based on visualization and chosen elements
  - Based on survey, recommendation of most popular restaurants
  - All-comprehensive dashboard built upon filtering and selection.

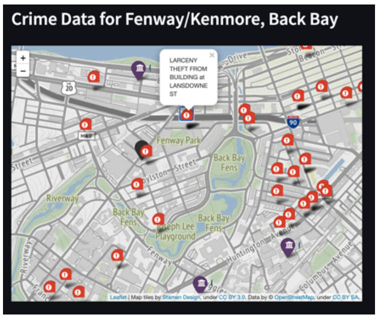


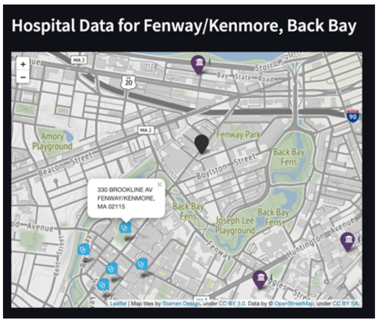
ANALYSIS AND RESULTS

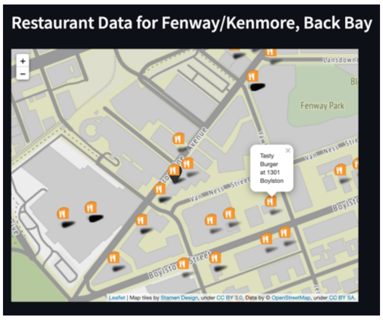
The analysis through Stream LIT is explained through an example/filter selection for the following

Suppose I have a restaurant in mind to reside near Saloniki Restaurant in Fenway/Kenmore. So, I along with it select the neighborhoods to stay as Fenway/Kenmore. In the above website we can select multiple neighborhoods.

- Crime, Hospital, Restaurant Visualization based on user input for neighborhood

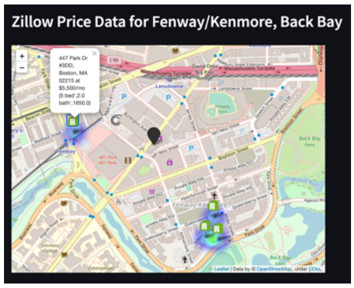






From the above maps, we can see the closest crime, hospital, and restaurant data. The tooltip lets us get more information on the location.

- Zillow Price Heatmap around the locality



The plot at the right allows you to look at the least-priced properties around the area. If the area is greener then the prices are lower compared to the other properties. Vice versa if the area is red then the price will be higher compared to the other properties. So here we strike a deal in Zillow for 5 beds and 2 baths (1650 sqft) at 447 Park Drive.

- Crime Shooting Heatmap around the locality

The visualization at the right shows the crime scenes with shooting occurring in mind. It shows that Fenway/Kenmore and Back Bay is quite safe compared to Roxbury, Dorchester and Mattapan where a lot of shootings occur.

- Count of Crimes, Hospitals, Restaurants in Boston Neighborhood

From the data table at right, we see the count level statistics for the neighborhoods.

Boston's Neighborhood Information

	Counts of Crime in Area	percent of crime	Counts of Hospitals in Area	percent of hospitals	Counts of Restaurants in Area	percent of restaurants	Counts of Zillow Properties in Area	percent of zillow properties
Roxbury	9647	17.36%	1	4.17%	115	5.92%	8	5.13%
Dorchester	9451	17.01%	2	8.33%	132	6.8%	13	8.33%
South End	4244	7.64%	2	8.33%	93	4.79%	5	3.21%
Mattapan	4200	7.56%	1	4.17%	33	1.7%	2	1.28%
South Boston	3164	5.69%	0	0.0%	119	6.13%	6	3.85%
Brighton	3105	5.59%	3	12.5%	85	4.38%	25	16.03%
East Boston	2783	5.01%	0	0.0%	105	5.41%	4	2.56%
Hyde Park	2632	4.74%	0	0.0%	29	1.49%	1	0.64%
Jamaica Plain	2164	3.89%	2	8.33%	55	2.83%	1	0.64%
Downtown Boston	1899	3.42%	0	0.0%	225	11.59%	0	0.0%
Back Bay	1752	3.15%	0	0.0%	167	8.6%	3	1.92%
Fenway/Kenmore	1620	2.92%	5	20.83%	187	9.63%	27	17.31%
Roslindale	1605	2.89%	1	4.17%	54	2.78%	5	3.21%
No Details	1484	2.67%	1	4.17%	115	5.92%	12	7.69%
Allston	1339	2.41%	0	0.0%	123	6.34%	31	19.87%
Charlestown	1159	2.09%	0	0.0%	23	1.18%	0	0.0%
Mission Hill	997	1.79%	2	8.33%	32	1.65%	8	5.13%

It is clearly visible from the table that most crimes occur in Roxbury, Dorchester. While from the hospital dataset from Analyze Boston tells us that most of hospitals are located at Fenway/Kenmore which is 5 out of 24 hospitals. Similarly, we can see most restaurants are in Downtown Boston, Back Bay, Fenway/Kenmore.

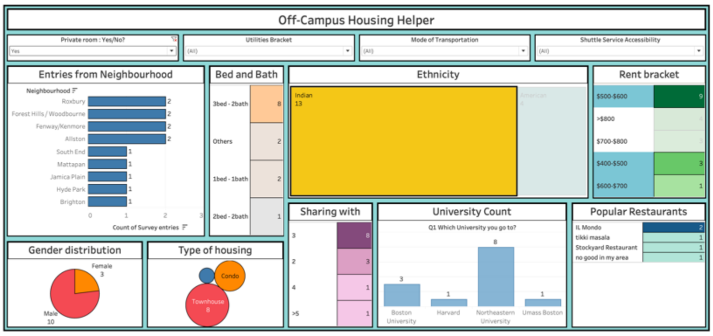
The analysis in TABLEAU recommends neighborhoods to students based on their preferences and likings as presented in the tool.

We analyzed multiple datasets to zero down on the quality and type of a neighborhood and all that it entails, like restaurants, type of buildings, sweeping schedules etc.

This helps the student better understand where they will be going and get a better understanding of the lifestyle they will live/ be exposed to.Our analysis also tells what the average property rate for the area is and whether it is a good buy based on many of the parameters that we have analyzed.

The way we answer out business problem is in a unique way of nudging the user in the direction of their answer by letting them fiddle around with the dashboard and filter according to their needs and preferences. This can be better understood with this example:-

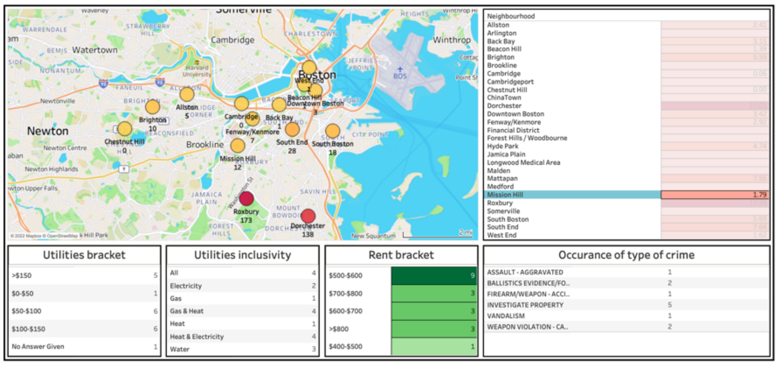
- Ishan wants to live in a private room with Indians and his budget is somewhere around $800 all inclusive. Let's make a personalized dashboard just for him to find out where he could live and what would follow?



Following inferences can be derived from this filtration exercise :-

1. *Roxbury, Forest Hills, Fenway/Kenmore, Allston are a few of the best places* where the majority of the apartment layout is 3 Bed 2 Bath.
2. The rent is selected onto his range, so now we got to see how many people Ishan must share his apartment with, in order to determine room privacy.
   Since *8 entries share their room with 3 people*, it can be safely assumed that *3 beds are shared among 3 people and the private room criteria is **met***.
3. Now *majority of these Indians are from Northeastern*, but a few of them are even from *BU, Harvard and Umass Boston*. This is the extra set of information that our dashboard gives Ishan.
4. Now if Ishan is a foodie, he can also see that *Il Mondo is a restaurant that our target demographic frequents*. So being new to that area, he can go to Il Mondo to have a decent meal with experimenting much.

Now that we have all our data ready, we need to just make sure that the area Ishan lives in is safe or not. And for that we have designed a Crime Monitoring system which tells us just about how safe an area is:-



Over here we can see that mission hill has a crime rate of 1.79% which is 4th last on the list; and that is a good sign.

Now since our data has been double checked, I.e Ishan likes it, and Crime monitor says the choice is safe.

Now Ishan can go ahead and book properties falling in line with these needs.Therefore, the business problem can be solved through the interacting dashboard for the challenges faced by international students with respect to off-campus housing based on preferences.

EXTERNAL MATERIALS

Data Sources

- Boston Bldg Inventory - https://data.boston.gov/dataset/boston-buildings-inventory
- Bldg and property violations - https://data.boston.gov/dataset/building-and-property-violations1
- RentSmart - https://data.boston.gov/dataset/rentsmart
- Hospital Locations - https://data.boston.gov/dataset/hospital-locations
- Crime Incident Reports - https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

Additional Resources

- https://datausa.io/profile/geo/boston-ma

Books

- McGregor, S. (2021). Practical Python Data Wrangling and Data Quality. O'Reilly Media, Inc. ISBN: 9781492091509
- Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., ... & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization, 10(4), 271-288.
- Endel, F., & Piringer, H. (2015). Data Wrangling: Making data useful again. IFAC-PapersOnLine, 48(1), 111-112.

- Koehler, M., Bogatu, A., Civili, C., Konstantinou, N., Abel, E., Fernandes, A. A., ... & Paton, N. W. (2017, December). Data
  context informed data wrangling. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 956-963). IEEE.
- Rattenbury, T., Hellerstein, J. M., Heer, J., Kandel, S., & Carreras, C. (2017). Principles of data wrangling: Practical
  techniques for data preparation. " O'Reilly Media, Inc.".
- Sarkar T., & Roychowdhury S. (2019). Data Wrangling with Python. Packt Publishing. ISBN: 9781789800111
- Furche, T., Gottlob, G., Libkin, L., Orsi, G., & Paton, N. W. (2016, March). Data Wrangling for Big Data: Challenges and Opportunities. In EDBT (Vol. 16, pp. 473-478).

Websites

- https://www.bu.edu/pdpa/files/2019/03/Finding-Housing-in-the-City.pdf
- https://www.google.com/books/edition/International_Encyclopedia_of_Housing_an/1BQLgSXCmmQC?hl=en&gbpv=1&dq=boston+housing+for+international+students&pg=PT6304&printsec=frontcover
- https://www.google.com/books/edition/Babson_College_College_Prowler_Off_the_R/Qn0z20q3PWgC?hl=en&gbpv=1&dq=boston+housing+for+international+students&pg=PA60&printsec=frontcover
- Sciencedirect.com/science/article/abs/pii/S0016718510000850
- https://open.bu.edu/bitstream/handle/2144/25840/Trehan_Surinder_1962_web.pdf;jsessionid=23185CC9D6CE8754BF6620113845C629?sequence=1

CONCLUSION – ADDITIONAL DATA AND ANALYSIS

If we had more time, we would have gathered more data with respect to every neighborhood in Boston through survey collection and some company databases. We could provide a more comprehensive tool for students to get the most accurate property to reside in based on their preference. The Tableau inferences and the Webapp tool can be merged to create a more robust tool that provides a seamless interaction for students recommendation based on the filters provided by them. Additionally, Machine Learning tools could be deployed to give the most desired and preferred property.