Introduction – Motivation

When a person is researching a property, one of the biggest challenges for users is to efficiently gather information. Making the best decisions and buying/renting at the right price is about research and knowing your market inside and out [1]. The current practice for searching information about a property is manual and time inefficient. According to the National Association of Realtors (NAR) , "The share of home buyers who used the internet to search for a home increased to an all-time high of 97%" [2]. Additionally, the NAR states, "Buyers typically searched for eight weeks and looked at a median of 9 homes and viewed 5 of these homes only online". If we treat homebuyers as a proxy for property researchers, this research indicates that most property research is done online and for about two months. Two months does not seem satisfactory for homebuyers, as many still report that they have difficulties finding the right one [3]. The housing market is currently on an upward trend, with September 2021 experiencing an 8.6% increase over the last year [4]. Additionally, houses are on the market for a shorter period, down from 54 days in 2020 to 43 days in 2021. The change in these metrics indicate that property researchers need to not only reduce the time they spend searching, but they must also explore more options within that shortened period.

Problem Definition

The main reason most innovations fail is that most current market research methods focus on consumer needs but ignore consumers' personalities and the different ways they make purchase decisions [5]. DataRealtor focuses on providing the user with a multi-dimensional perspective on a property's geographic location. Modern real estate research tools offer information about a given property but oftentimes do not provide any information on the kind of environment the user might be living in. Zillow, one of the main real estate websites, uses their users by the demand they place on the market to increase the value of all property in an area and limit the availability of affordable options [6]. This tool will provide a comprehensive view of the property's environment, as we are gathering data describing local markets, schools, transportation, and insurance, from a variety of sources that would take the user a much longer time to gather by hand. We are also providing a review scorer that will allow the user to check the trustworthiness of an apartment review.

There are some important risks with our application that must be considered. One risk is that the user will need to be engaged and critically think about the information presented to them. Data integrity cannot be guaranteed to the user so they will need to pay attention to the quality of the information presented to them. The major payoff to our application is the amount of user questions the tool can answer and the time efficiency improvement spent on user research. The main source of our costs was the time and computational cost.

Proposed Method

We expect a wide variety of user classes to find this application useful because of its ability to deep dive into multiple dimensions of information. Our features are:

- Review Scorer – Helps apartment renters determine which reviews are trustworthy by allowing the user to input a sequence of review text and return a trustworthiness score.
- Local Economy Snapshot – Helps homebuyers understand the long-term income and demographic data trends for long term analysis and regional comparison. The National Association of Realtors (NAR) estimates that one job is generated for every two home sales [7] and demographics are often overlooked but significantly affects how real estate is priced and what types of properties are in demand [8].

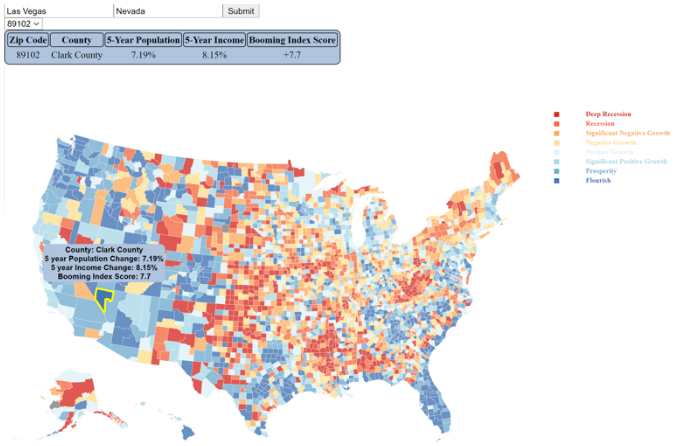
- Car Insurance Local Landscape - Helps homebuyers understand the insurance implications of living at that property. Homeowners pay more premium in expensive zip codes since insurance companies take location into account for setting up the insurance rate [9]. Also, interestingly, according to Commercial Real Estate Sales (CRES), failing to disclose the possibility of an accident and impact of the busy road to the property buyer can potentially lead to legal consequences [10].
- Local K-12 Education Snapshot – Catered to property researchers with children who are interested in having their children go to a good school. Even in a down market, an excellent school can be the rising tide that lifts all nearby home prices [11].
- Hourly Uber Accessibility Snapshot – Catered to property researchers who plan to use Uber in lieu of self or public transportation. This feature could also serve to inform drivers which hours of the day have the least traffic. The main goal of an urban transport system is to provide accessibility for the inhabitants on an urban region and an integrated network that provides seamless access between all points is ideal [12].

We used an integration of Python Flask and D3 to make our web application. Each feature has its own dataset. Each feature's dataset was cleaned and transformed in its own separate file and power the web application separately. As a proof of concept, we limited the cities tab to select cities in the US:

- Atlanta, Georgia
- Los Angeles, California
- Chicago, Illinois
- Seattle, Washington

In order to create comprehensive view, we combined information from different features into a combined page where applicable. When the user searches by major city, they can see an aggregated view of car insurance, school and Uber information. When the user searches by zip code, they can see an aggregated view of economic and school information. Our searching by zip code tab is more comprehensive and thus, the user can search across the US and is not limited to the four core cities. Due to the nature of our reviews feature, we created a separate tab just for it.

For our economy feature, demographic and income data was gathered at the county level from 2015 to 2019 from the United States Census Bureau where each year's data was downloaded as a CSV file. We extracted population and income from 2015 and 2019 datasets, compared them in a 5-year window and formed a trending data frame to show the yearly differences of those indicators. The trending data frame was then used to shape an interactive UI that the users can select and view. The web application has a Choropleth map (at the United States County level) to show the demographic and income change trends and a "Booming Index Score" which is calculated based on the trends. The booming index score is calculated as weighted average of 5-year population change and 5-year income change. The selected County is then highlighted, and a list of details is displayed. When a user searches for a particular area by Zip code, the web application will automatically highlight the county which the area belongs to and show the score.



Las Vegas | Nevada | Submit

89102

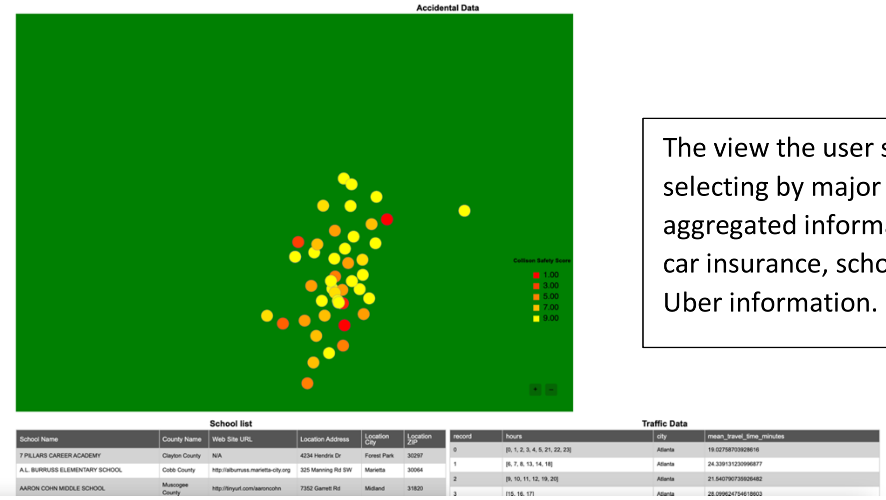
Zip Code	County	5-Year Population	5-Year Income	Booming Index Score
89102	Clark County	7.19%	8.15%	+7.7

County: Clark County
5 year Population Change: 7.19%
5 year Income Change: 8.15%
Booming Index Score: 7.7

School information data includes the school rating, user review and other information. The data source was the integrated Greatschools.org API with scrapping from nces.ed.gov. The data was transformed into Json to use the D3 interactive UI. The information is displayed in a table format with the following columns: school name, county name, web site URL, location address, location city, location zip. Thus, the user can easily look to see which schools are available in a given location, what other nearby school options are and the school website should they decide they want further information.

Uber data is gathered at the hourly level from Uber Movement, which Uber initiated to provide data and tools for cities to understand and address urban transportation challenges [13] more deeply. We used one year's worth of data ranging from quarter 2 of 2019 to quarter 1 of 2020 since the data currently available to the public was only updated until quarter 1 of 2020. Each quarter's worth of data was downloaded via csv. The dataset contains for each source and district the mean and standard deviation of the travel time. This feature is displayed as a table with the average wait time for each cluster of hours, allowing the user visibility on best times to travel.

Car Insurance data consists of accident data collected from February 2016 to December 2020 from 49 US states and has 47 attributes. The Zip code information was downloaded from Simple Maps and cleaned to produce the following columns: Country, zipcode, City, StateName, StateCode, Municipality, Latitude and Longitude. US accidents information was downloaded from Kaggle. As per the application requirement, we cleaned the dataset to 11 required attributes for the four cities listed above. To calculate an accurate weighted score, it is critical that all zip codes are listed for a city with or without accident data. "The more car accidents there are in ZIP code, the higher insurance rate will be. This is because the risk of getting into an accident and filing a claim is increased [14]". To address the above need, we used the file from Simple Maps which contains all the zip codes and cities in the United States. We used python SQLite to load the accident file and zip code file. The weighted score was calculated by dividing the zip code accident count over the total accident count in a given city. The visualization of the weighted score was done in the D3 interactive UI. The user selects the city and then sees an aggregate weighted score of 1-10 calculated from the accident dataset for that zip code. The less the accident count, the higher the weighted score. The color legends for this Map ranges from Red (lowest score) to yellow (highest score). This will help user to estimate insurance rate difference compared to nearby zip codes.



School list

School Name	County Name	Web Site URL	Location Address	Location City	Location ZIP
7 PILLARS CAREER ACADEMY	Clayton County	N/A	4234 Hendrix Dr	Forest Park	30297
A.L. BURRUSS ELEMENTARY SCHOOL	Cobb County	http://alburruss.marietta-city.org	325 Manning Rd SW	Marietta	30064
AARON COHN MIDDLE SCHOOL	Muscogee County	http://tinyurl.com/aaroncohn	7352 Garrett Rd	Midland	31820

Traffic Data

record	hours	city	mean_travel_time_minutes
0	[0, 1, 2, 3, 4, 5, 21, 22, 23]	Atlanta	19.02758703928616
1	[6, 7, 8, 13, 14, 18]	Atlanta	24.339131230996877
2	[9, 10, 11, 12, 19, 20]	Atlanta	21.540790735926482
3	[15, 16, 17]	Atlanta	28.099624754618603

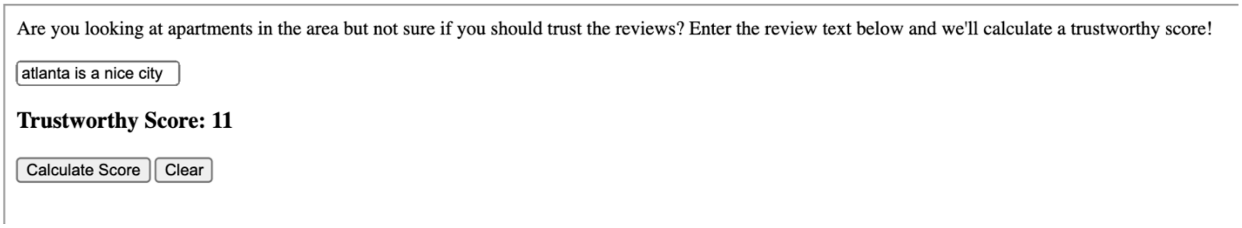
The view the user sees when selecting by major city, aggregated information of car insurance, school and Uber information.

Reviews data was webscraped from Apartments.com using BeautifulSoup as the main tool for extraction. On a high-level, the process uses city and state to gather URLs to properties. Once a list of

URLs is obtained, the URLs are entered individually into another module that extracts all the reviews for a specific property, then fed into a regex function to extract all relevant text and eliminate trailing and leading whitespaces. Information such as address are also stored. To avoid being blocked or banned, the module will only call the apartments.com API once every 2 minutes, with random variance added.

This process yielded around 60k reviews across 915 cities. We hand-labeled 4k reviews as fake or real using techniques covered in [15], such as looking for vague language and comparing reviews across websites. The reasoning behind hand-labeled reviews is that a quantified approach towards labeling reviews as fake or real would involve developing a trustworthiness framework, which is out of scope for a project with this short of a timeline. Once labeled and preprocessed to remove undesirable contents, such as leading/trailing whitespaces and non-ascii characters, we randomly selected 1500 fake reviews and 1500 real reviews, to allow for a balanced dataset. The dataset is then split, 70% for training, 20 % for validation, and 10% for testing.

For the modeling of review scores, we utilized the ASGD Weight-Dropped Long-Short-Term-Memory (AWD LSTM) [16] model from the Fastai package, applying tokenization, stopword removal, and character masking prior to any fitting. AWD LSTM is a recurrent neural network based-architecture that applies valuable generalization techniques such as DropConnect and Non-monotonically Triggered Asynchronous Stochastic Gradient Descent (NT-ASGD) optimization, making it an extremely powerful building block for transfer learning. We apply transfer learning by first training the model's layers individually, for one epoch, using a learning rate optimizer to guide us towards the best learning rate to use for each layer's training. The goal for this model is to learn on the review text corpus and be able to predict the ending of a sentence when given a starting block of text. This model is then saved and used as an encoder for a text classification model. To train the classifier, we first load in the fine-tuned model, then train the head of the model for one epoch, freezing the weights in the remaining layers, then one layer at a time, we unfreeze the remaining layers and train them using discriminative learning rates. To evaluate the performance, we test the model against a test set consisting of 600 review texts. Once the model is built, we save it as a pickle file and create a function that takes a string input and returns the model's probability for class 0, which we establish as the 'fake' class. The output is then multiplied by 100 and rounded.



Are you looking at apartments in the area but not sure if you should trust the reviews? Enter the review text below and we'll calculate a trustworthy score!

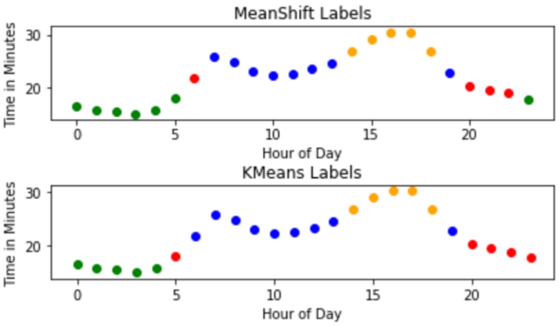
atlanta is a nice city

Trustworthy Score: 11

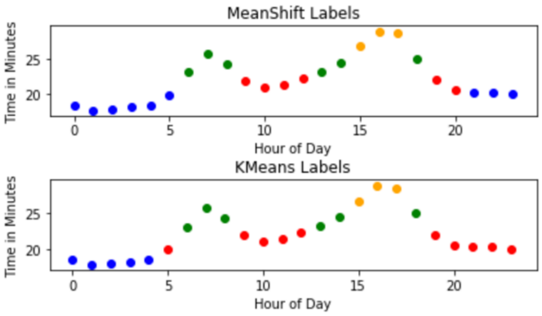
Experiments/Evaluation

We ran an experiment on the clustering algorithm for the Uber data. For each hour of each city, the Uber data listed the average travel times for each hour over a period of one year. We took the average of the average travel times aggregated by hour and our goal was to cluster the hours with similar travel times so that the end user could see best travel times to call an Uber and whether or not that hour of the day was similar to another. We tested to see which of the KMeans or MeanShift algorithms was better. The test was run across all cities. We compared the process and output of both algorithms. The MeanShift algorithm is a centroid based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region. These candidates are then filtered in a post-processing stage to eliminate near-duplicates to form the final set of centroids. Unlike the KMeans, MeanShift automatically sets the amount of clusters. The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified [17]. The number of clusters was determined via the elbow method for this experiment. As we can see below, the output of both clustering models proved to be very similar for all 4 cities. As a result, we decided to use the output of the MeanShift model as it would allow us to skip the elbow method step of specifying the number of clusters.

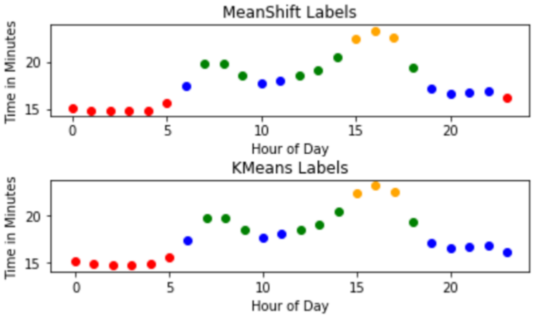
Los Angeles



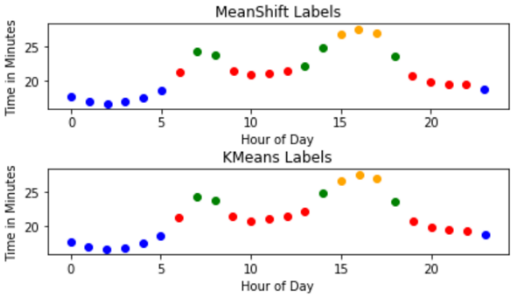
Atlanta



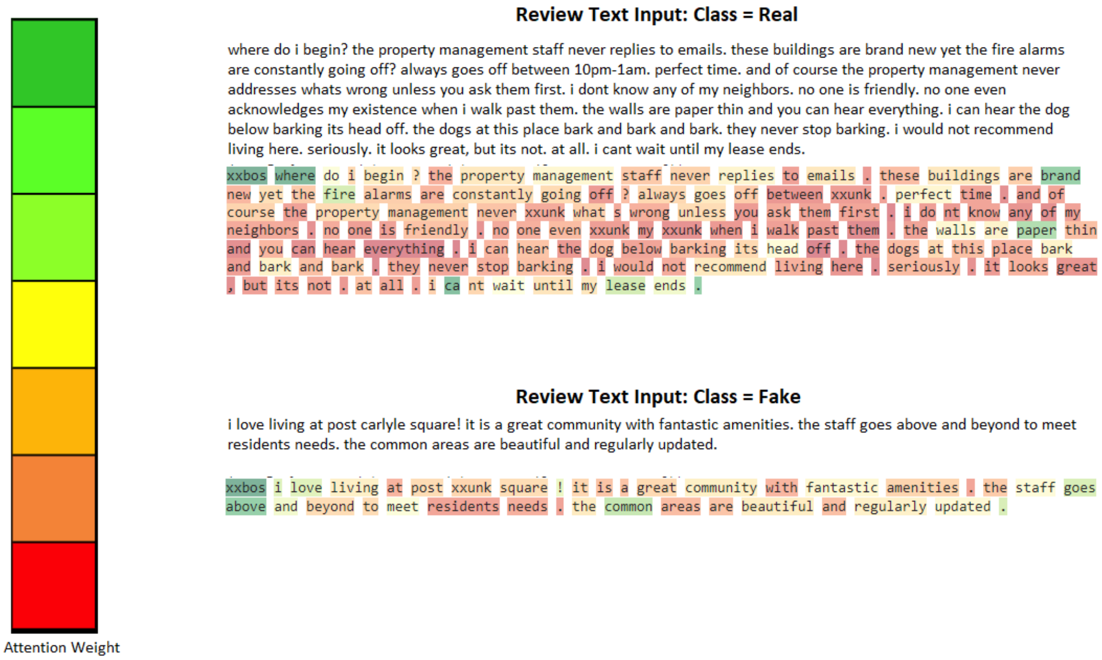
Seattle



New York



The review scorer model was evaluated on 600 labeled review texts, resulting in a 91% accuracy and 85.5% precision. We evaluated the model further by utilizing fastai's intrinsic attention function, which uses sequential Jacobian matrix to analyze the intrinsic attention of the input. This allows us to analyze the impact of individual words in a sentence by calculating gradients for each token with respect to the classification out, then compare the relative magnitude of the gradients by bounding the values 0 – 1. We took 10 out of dataset samples and looked at the scores for the words, one example of which can be found below. Evaluation using intrinsic attention revealed that words that exaggerate the quality of the experience or discuss the property in a vague fashion had the highest relative impact on the classification score. We can speculate that the model learned to classify review text using the same strategies that were used for the labeling process. Another possible interpretation is that the model is learning based on positive or negative sentiment, enough though the labelling process was supposed to be sentiment agnostic.



Review Text Input: Class = Real

where do i begin? the property management staff never replies to emails. these buildings are brand new yet the fire alarms are constantly going off? always goes off between 10pm-1am. perfect time. and of course the property management never addresses whats wrong unless you ask them first. i dont know any of my neighbors. no one is friendly. no one even acknowledges my existence when i walk past them. the walls are paper thin and you can hear everything. i can hear the dog below barking its head off. the dogs at this place bark and bark and bark. they never stop barking. i would not recommend living here. seriously. it looks great, but its not. at all. i cant wait until my lease ends.

```
xxbos where do i begin ? the property management staff never replies to emails . these buildings are brand
new yet the fire alarms are constantly going off ? always goes off between xxunk . perfect time . and of
course the property management never xxunk what s wrong unless you ask them first . i do nt know any of my
neighbors . no one is friendly . no one even xxunk my xxunk when i walk past them . the walls are paper thin
and you can hear everything . i can hear the dog below barking its head off . the dogs at this place bark
and bark and bark . they never stop barking . i would not recommend living here . seriously . it looks great
, but its not . at all . i ca nt wait until my lease ends .
```

Review Text Input: Class = Fake

i love living at post carlyle square! it is a great community with fantastic amenities. the staff goes above and beyond to meet residents needs. the common areas are beautiful and regularly updated.

```
xxbos i love living at post xxunk square ! it is a great community with fantastic amenities . the staff goes
above and beyond to meet residents needs . the common areas are beautiful and regularly updated .
```

We issued a survey that simply asks a rating of the web application from 1-10 and a comment section for feedback. The rating of 1-10 is simply so we can get a rough opinionated estimate of how the first iteration of our prototype fares as a whole to users who don't know the product. Our comment section is used to obtain open ended and hopefully honest critiques of our prototype. We made our questions generic and didn't specify to rank x or comment on x in order to allow the surveyed to base their answers on the unbiased, subconscious KPI they had in their minds. We distributed this survey via Piazza post to classmates and the school slack groups. Our main objective is simply to understand if the application is usable and if any noticeable changes need to be made to the first iteration of the application so we are willing to risk audience biases in our responses. Through the 12 respondents so far on our survey, we learned that Data Realtors is above average (mode rank of 7) and needs to be tuned to provide more user intuition on how to use the website and the information presented. We considered this to be valuable information because developers who work on a product and know it well will not have the same view points as a user who is seeing it for the first time.

Conclusions & Discussion

All team members have contributed similar amount of effort. Every member in the group worked on their feature: Xueyuan on Local Economy Snapshot, Richard on Review Scorer, Raghu on Car Insurance Local Landscape, Fira on Hourly Uber Accessibility Snapshot and Dohyung on Local K-12 Education Snapshot. Fira and Richard focused more on the modeling aspects of their features while Xueyuan and Raghu focused more on making their Choropleths for their features more interactive. Dohyung worked extensively on combining all of the features together and publishing the website.

We hope that users are able to walk away with at least one piece of insight that they could not get easily before. Researching property can be difficult. Often times, we see a piece of property with information about it, but we're not sure about what to expect about the outer environment we could potentially living in. We live in a world now where people are more mobile than ever before and moving to a relatively foreign region is becoming more and more popular. Although our application is still a work in progress, we hope users still find satisfaction in the prototype.

References

1. Opie, M. (2015). Find the Right Property, Buy at the Right Price. John Wiley & Sons.
2. Highlights From the Profile of Home Buyers and Sellers. (2021, March 16). Retrieved from https://www.nar.realtor/research-and-statistics/research-reports/highlights-from-the-profile-of-home-buyers-and-sellers
3. “Home Buyers Tapping into Technology to Make Real Estate Search More Personal.” *Bizjournals.com*, https://www.bizjournals.com/philadelphia/news/2019/07/18/home-buyers-tapping-into-technology-to-make-real.html
4. Santarelli, M. (2021, October 14). Housing Market Forecast 2021 & 2022: Crash or Boom Next? Retrieved from https://www.noradarealestate.com/blog/housing-market-predictions/
5. Gurtner, S., Spanjol, J., & Griffin, A. (2018). Leveraging constraints for innovation: New product development essentials from the PDMA. Hoboken, NJ: John Wiley & Sons.
6. Chowdhury, G. G., & Loukissas, Y. (2018). All the Homes: Zillow and the Operational Context of Data. Cham: Springer.
7. Jobs Impact of an Existing Home Purchase. (n.d.). Retrieved from https://www.nar.realtor/jobs-impact-of-an-existing-home-purchase
8. Nguyen, J. (2021, September 13). 4 Key Factors That Drive the Real Estate Market. Retrieved from https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp#demographics
9. Strong, A. (2019, August 14). When Real Estate Agents Sell a Car Crash Prone House. Retrieved from https://www.cresinsurance.com/managing-your-risk-when-selling-a-car-crash-prone-property/
10. Megna, M. (2021, February 12). How your ZIP code affects your car insurance. Retrieved from https://www.carinsurance.com/Articles/zip-code-car-insurance.aspx
11. Why You Need to Research School Districts When Buying a Home. (n.d.). Retrieved from https://www.publicschoolreview.com/blog/why-you-need-to-research-school-districts-when-buying-a-home
12. Carlos, M. A., & Paget-Seekins, L. (2016). Restructuring public transport through bus rapid transit: An international and interdisciplinary perspective. Bristol: Policy Press.
13. Uber movement: Let's find smarter ways forward, together. (n.d.). Retrieved November 4, 2021, from https://movement.uber.com/?lang=en-US.
14. Leslie Kasperowicz Farmers (2021). The Truth About Auto Insurance Rates by ZIP Code: Retrieved from https://www.autoinsurance.org/quoting-auto-insurance-rates-by-zip-code
15. Merity, S., Keskar, N. S., & Socher, R. (2017, August 7). Regularizing and Optimizing LSTM Language Models. Retrieved from https://arxiv.org/pdf/1708.02182v1.pdf.
16. Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016, September 26). Pointer Sentinel Mixture Models. Retrieved from https://arxiv.org/pdf/1609.07843v1.pdf.
17. Sklearn.cluster.AffinityPropagation. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html