Data Component Analysis Update

1. *What data do you need to make your decision?*

   Data required for our Real Estate DSS can broadly be defined in two categories – Housing and Environment data. Housing data are data related to specific features of an individual real estate listing. This includes elements such as square footage, housing type (apartment, house, condo, etc.), number of bedrooms, and price.

   Environmental data will cover a much larger scope. Under this group, we will look at schools, restaurants, nightlife, community centers, and public transportation. For restaurants and bars, we will report the count of matching items within a certain radius and return the aggregated ratings for these places. Public transportation, community centers, and schools will be evaluated on distance to the nearest matching location.

   Finally there will also be user preference data. This will come in the form of a pre-set profile that a user will select that best describes themselves as a potential home buyer. Each of these profiles will be assigned a profile with evaluation criteria set to levels that would correspond to typical interests of that group. For example, the retiree profile would weight more heavily on community centers and restaurants against other considerations such as schools and bars.

2. *Where will you get the data?*

   For this project we will mainly be relying on APIs to collect data from various web sources. Housing data will be collected from the Realtor.com API, which contains information on housing units currently listed for sale, their pricing, and various features about the house that may serve as criteria for prospective buyers.

   Google APIs will provide data on restaurants, bars, and bus stations. Since our focus for the prototype is on Arlington County, Virginia, we will also utilize publicly available information provided by the county – which offers a listing of locations of schools, community centers, and metro stops in the county.

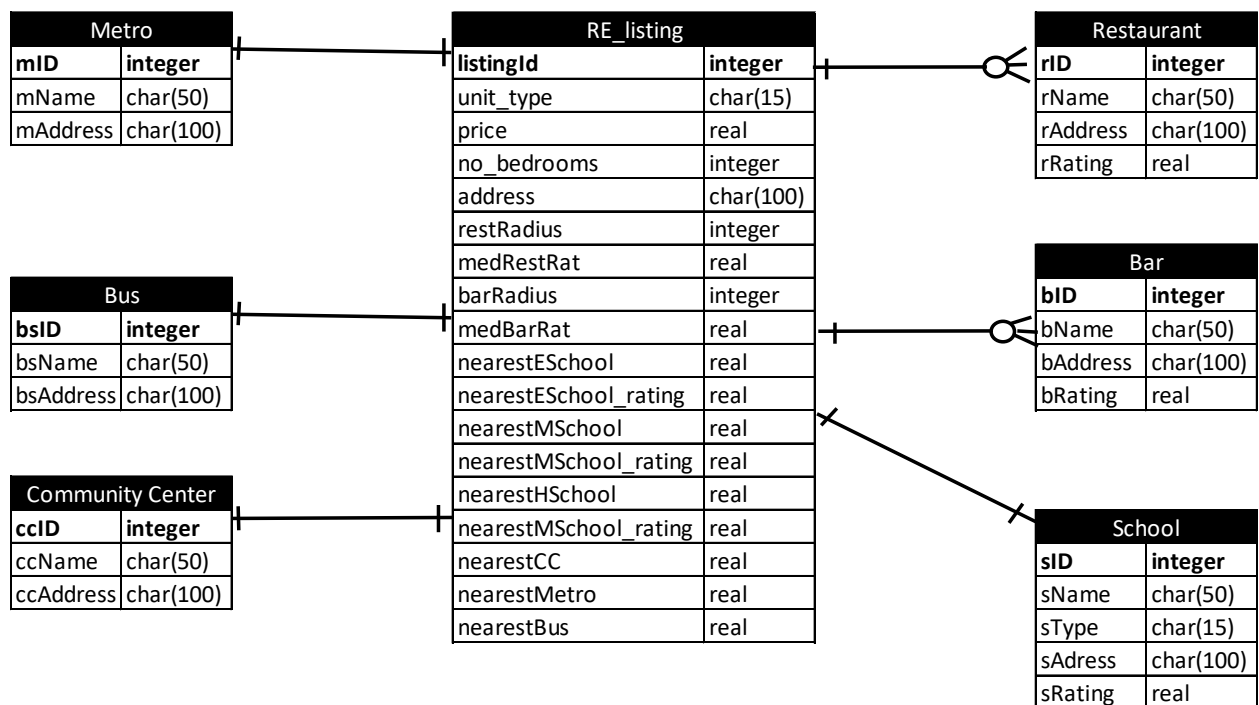3. *How does the data meet the characteristics required for your decision?*

   The Housing data will provide the information specifically relating to housing units. Homebuyers may be interested in housing of certain sizes, or within a price range, the housing data will allow the system to return listings that meet these minimum criteria.

   The Environmental data will provide potential buyers a better idea of neighborhood area in which they are buying into. Whereas Housing data can only give potential buyers a

view of the individual unit they, the environmental data will give purchasers a broader view of the area they are moving into. This will correspond to various preferences for different user groups to purchase a house in an ideal area for their current circumstances.

4. *What type of database will you use? Draw Entity Relationship*

We will plan to use a relational database for this exercise. Housing data will represent a single entity, where each row is a single listing, with various features such as bedroom count, size, and type as features. There will be several other entities to represent the different facets of environmental features. The most important data feature to these entities will be their location, as it will be used to calculate the distance to individual listings.

Metro

mID	integer
mName	char(50)
mAddress	char(100)

Bus

bsID	integer
bsName	char(50)
bsAddress	char(100)

Community Center

ccID	integer
ccName	char(50)
ccAddress	char(100)

RE_listing

listingId	integer
unit_type	char(15)
price	real
no_bedrooms	integer
address	char(100)
restRadius	integer
medRestRat	real
barRadius	integer
medBarRat	real
nearestESchool	real
nearestESchool_rating	real
nearestMSchool	real
nearestMSchool_rating	real
nearestHSchool	real
nearestMSchool_rating	real
nearestCC	real
nearestMetro	real
nearestBus	real

Restaurant

rID	integer
rName	char(50)
rAddress	char(100)
rRating	real

Bar

bID	integer
bName	char(50)
bAddress	char(100)
bRating	real

School

sID	integer
sName	char(50)
sType	char(15)
sAdress	char(100)
sRating	real

5. *How will the data in your system be kept current?*

API calls can be refreshed to pull in new listings and new environment features. Based on the current housing market in Arlington, these refreshes would have to occur frequently as housing options are not listed for long.

6. *Will you have data available for your prototype, or will you need to use constructed data?*

For nearly all of our data, we will have real data from the sources described in the above paragraphs. For the School Ratings feature, we will use constructed data. We have opted

to use constructed ratings for this because we do not currently have the experience on the team to know how to objectively rate schools. Further calculating such a measure would be an entire project to itself using features like faculty to student ratio, test scores, graduation rates, etc. In order to keep our project scope manageable, we plan to use a constructed 1 through 5 ratings for each school.

Follow up from Update #1 Feedback

Here we address comments provided on our first update. In response to the agency issue where real estate agents may have different criteria and weights than the actual homebuyer, we plan to market this tool directly to the homebuyers themselves. We assume that real estate agents will already have a good understanding of the area in which they are working, with a base knowledge of how close a listing is to each of our features. Our system does not seem like it would provide much value to those real estate agents, therefore we are planning to design this for a novice user who may be unfamiliar with the Arlington area and all it has to offer.

In response to a "community-wide direct benefit"; this may better be described as a soft benefit where we assume our tool will match people to living areas where they will be more likely to take advantages of local offerings (e.g. the foodie living near an area with many restaurants will likely eat out more, benefiting the restaurant industry, than if that foodie had ended up in an area that is far from restaurants).

In response to the broad range of interests; we are moving towards a model where systems users will select among a set of pre-defined user profiles, which will include groups such as "Young Family", "Retirees", "Power Couple", etc. Each of these profiles will be assigned different weightings for the evaluation considerations. We believe this may ease the user experience, removing the need for them to have to decide how much they value having restaurants nearby than schools – which may be too much of a thought intensive process to frustrate users.