IS457 Final Project - Fall 2019

Due in class December 9

Notice: From Nov 18 to Nov 22, and Dec 2 to Dec 6, there will be five office hours for each TA, we will use an online sign-up sheet (we will reset for the form for the week after break). https://docs.google.com/spreadsheets/d/1uDCFJLqeGaj28f2WTpYPiPFFvL84t9NNigth5_932ho/edit?usp=sharing . Everyone can sign up for 10 minute meetings (maximum of 3) each week. If you no show for a meeting, we'll cancel your subsequent meetings. Do not make any change for the sheet – recall Google maintains the history of changes.

GlobalTechSync, a highly regarded fortune 500 tech company, is looking for a site to build their new corporate headquarters. You have been asked by the city of **Austin, TX**, to help select potential sites to include in their bid that they submit to the GlobalTechSync location scouts.

You will apply the data science skills you have learned in this course to help select the best locations and will follow a Lifecycle of Data Science as you explore the options.

As always in this class, the work you prepare and turn in must be your own work. **This is not a group project**. All code used in your final project must be written by you and all analysis must also be done by you.

Please read all the instructions before you begin work on the project.

We will use a dataset of parcels (areas of land) in Austin, TX to conduct this analysis called Austin_Lots.csv (make sure you download it from our course page).

Here are some details and documentation about the dataset you will be using:

Variable name	**Description**
FID	Unique row ID
block_id	Identifies Austin city block if applicable
created_by	ID of employee who created initial record
date_creat	Date record was initially created in Austin public database
land_base_	Austin parcel record database ID
land_base1	Land unit type designation
lot_id	Austin lot ID
modified_b	ID of employee who last updated record
date_modif	Date record was last modified in Austin public database
objectid	Austin secondary ID
City_dist	Distance in meters from parcel to the Austin City Center
Airpt_dist	Distance in meters from parcel to the Austin International Airport
district	Texas service district number
Shape_Area	Area of parcel in meters squared
zoning_o_3	Austin zoning designation. See Appendix for code description
zcta5ce10	Zip code

LAND_USE_2	Specific land use designation
GENERAL_LA	General land use designation. See Appendix for code description
EWC_dist	Distance in meters from parcel to East-West Connector Highway
NSC_dist	Distance in meters from parcel to North-South Connector Highway
Mopac_dist	Distance in meters from parcel to Mopac Freeway
130_dist	Distance in meters from parcel to Highway 130
35_dist	Distance in meters from parcel to Interstate 35
ExTrail_1m	Number of existing urban trails within 1 mile of parcel
PpTrail_1m	Number of proposed urban trails within 1 mile of parcel
conf	Average bike lane comfort level (0 is most comfortable, 4 is least)
bike_lanes	Number of bike lanes within 1 mile of parcel
Bus_area	1 if parcel is in Austin Bus system service area, otherwise 0
TotBdgArea	Total area of buildings on parcel
Num_Bldgs	Number of buildings on parcel
MaxBdgArea	Area of largest building on parcel
tax_break2	District wide construction perk – percentage of parcel purchase cost waived (9.5 = 9.5%)
bk_tx_brk	Block wide construction perk – percentage of construction fees waived in the first 5 years (0.026 = 2.6%)
GEOID	US Census block group ID
Housing__	Housing opportunity index: a higher value indicates an individual in this block group has more opportunity to find affordable housing
Education	Education opportunity index: a higher value indicates an individual in this block group has more opportunity to obtain a high level of education
Economic__	Economic opportunity index: a higher value indicates an individual in this block group has more opportunity to achieve economic stability
Comprehens	Comprehensive opportunity index: a higher value indicates an individual in this block group has more opportunity overall in regard to housing, education, and the economy
Med_HH_Inc	Median household income per zip code
Med_rent	Median rent per zip code
Med_home	Median home price per zip code
Aff_rent_t	Percentage of rental units per zip code that are affordable for an average worker in tech to rent
Aff_own_te	Percentage of homes per zip code that are affordable for an average worker in tech to rent
Descriptio	Description of 2019 construction near the parcel

What you need to submit: Note that you might need to try many R commands (especially in your exploratory analysis), but only need to report meaningful steps that you used.

• An R Script (or rmd file) that contains all calculations that would allow us to re-run your analysis and obtain your findings.

• A (human readable) PDF Final Report

Part 1: Data Pre-processing

Q1

Importing the data into R and handling columns.

Import the data using the read.csv() function.

1.1 What are the initial dimensions of the dataset?

The data you are working was combined using datasets from several different Austin city departments and contains some columns that you will not need to make your decision on the best site location.

1.2 Look at the column descriptions above. Which four columns do you think will be the least helpful in selecting an ideal site for the GlobalTechSync headquarters? Why do you think these are less helpful?

1.3 Subset your data by removing the unnecessary columns you identified. What are the new dataset dimensions?

1.4 Why is it useful to subset your data before starting your analysis?

1.5 The current column names can be hard to read and recognize. Rename some of the columns so that the variables are easier to work with. Display your new set of column names.

Q2

Dealing with missing values.

Commonly used methods for dealing with missing values include replacing missing values by the mean/median/mode, keeping NAs, or dropping the observations with NAs, etc.).

2.1 What columns in the dataset contain missing values? What placeholder text is used to indicate that the values are missing (e.g blank, NA, N/A, -, etc.)? List any columns you think appear to have missing values, but actually should not have a value or have a value of 0.

2.2 Briefly describe how you deal will with these missing values **and justify why you chose these methods**. You may decide to use different methods for different data columns. You do not need to use methods beyond those we have discussed in class, however you should be thinking about the data and explain why you chose the steps you did based on observations about the data.

2.3 Describe how your choice of method to deal with missing values may affect your later analysis.

2.4 Implement your methods for dealing with the missing values.

2.5 After dealing with missing values, once again show the new dimensions of the dataset.

Q3

Data cleaning.

3.1 For the column initially called land_base1, how many unique values exist? Display the current value set and how many occurrences there are for each value. Indicate any values you think are errors.

3.2 Please standardize the values for the land_base1 column (so that each value that refers to the same thing has the same format). Then display the current values with how many there are of each. (Hint: what class of variable does R consider this to be?)

3.3 You realize that some of the tax_break2 values contain dollar signs. Find these instances and remove the dollar sign. Do you need to change the variable class? If so, go ahead.

3.4 It's happened again! Someone used Excel to open the files at one point and the values for GEOID (a 12 digit unique block group identifier) have been stored using scientific notation. What does a value in this column look like when you display it as an integer not in scientific notation? How many unique values are in this column? Why is this a bad thing? If you haven't already done so, delete this column.

3.5 Someone from the data department lets you know that there are likely 2 fully or partially duplicated rows in this dataset. Find these two rows and remove the duplicated rows (keep the copy of the duplicated row with the most information). Display the updated data set dimensions.

3.6 It turns out that the specific land use codes (LAND_USE_2) have missing metadata – no one can remember what they actually mean! Delete this column. Explain why metadata is so important.

3.7 Describe why these cleaning steps are necessary. What would happen if you needed to use these columns in later analyses?

3.8 Comment on and explain any other data cleaning or preparation steps you think would be necessary from your inspection of the data (you do not need to carry them out).

Q4

Transform columns into proper formats.

Often when you import data, the variable classes assigned to each column do not match what you would like them to be.

4.1 Please display the initial variable classes for each column.

4.2 Find at least one column where the variable class does not seem to make sense for the type of data. State what that column is and why a different class is more fitting.

4.3 Change the variable class(es) to one that is more fitting. Then display the new class(es) for those columns.

4.4 Give some examples of other ways R could import data as a variable class that is not useful. In general, why is it important to do this after the data cleaning step?

Part 2: Data Exploration

For each of the questions in this section, make sure you do not only look at a few variables, but explore the data set comprehensively.

Q5

Calculate descriptive and distributional statistics.

5.1 Since it is hard to get a mental picture of large data sets, conduct a preliminary exploration to understand the Austin dataset variables by calculating some descriptive and distributional statistics.

5.2 Describe anything you find that is unexpected or interesting.

Q6

Visualize the data.

To understand large amounts of complex data, it is helpful to use charts, tables, and graphs to visualize the data. Here are general steps you can follow:

6.1 Think about the types of variables in the Austin dataset. Then choose appropriate graphs to display distributions and trends for multiple variables.

6.2 Compare different graph types to see which ones best convey trends, outliers, and patterns in the data.

6.3 Describe what you find from the graphs.

Q7

Now look at the relationships among several variables.

7.1 For example, look at the original "conf" and "bike_lanes" columns. They are both indicators of ease of bicycle transportation, but each column conveys different information. What different information and what similar information can you get from these variables? How are the two variables related? Explain what you find.

7.2 Following this example, analyze at least two other groups of variables where you think there might be a potential relationship (do not pick two variables that are obviously directly related, like total building area and number of buildings).

Q8

Find areas that could be attractive to future employees.

You have access to construction permit description records that are located nearby each parcel which have been entered in the "Descriptio" column. Since multiple parcels could be near the same construction area, there are some duplicates.

8.1 Convert the letters in the "Descriptio" column to lower case. Why is this helpful? Do you lose information by doing this?

8.2 Extract the unique words used in the "Descriptio" column and eliminate the stop words that are in the list below. Displayed the first 10 values of this list.

a, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever,

every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, i, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, is, to, too, was, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

8.3 Preform a similar function to 8.2 but this time finding unique words *and* their frequency. What are the 10 most frequent non stop words, i.e. which are frequent words that give you meaningful information about the type of construction occurring? How can these help you finding a good site for GlobalTechSync?

8.4 Look through both word lists. Which words, at any frequency, do you think will be the most useful to determine places to attract tech workers? Why? Which high frequency words do you think will be the most useful to determine places to attract tech workers? Why? Why might a specific low frequency word be useful?

8.5 What additional word processing steps or stop words do you think would be useful for further text analysis of this variable? You don't have to implement these ideas.

Part 3: Site Selection

GlobalTechSync has several mandatory location requirements and additional things that would be nice but that they do not require.

Mandatory requirements:

1. The site must be in the metro bus service area (in this case the Austin Bus System).
2. The total parcel area must be greater than 300 square meters.
3. The base zoning district must not be residential.

Preferences:

1. An undeveloped site is preferred.
2. Ease of access to a major interstate or highway is preferred.
3. Easy access to the site by bike or foot is preferred.
4. Close access to green spaces and areas that offer opportunities for employee enrichment (such as concerts, public lectures, swimming pools, leisure areas…) is preferred.
5. Higher tax breaks or discounts at both the district and block levels is preferred.

6. High education opportunity in the area and strong nearby university systems are preferred.
7. Ability for tech workers to own their own houses is preferred.
8. Fast reliable internet needs to be easily accessible at the site.
9. Nearby active construction of office type structures is preferred.

Q9

Filter out unsuitable parcels.

9.1 Remove any parcels that are not in the metro bus service area.

9.2 Remove any parcels that have an area under 300 square meters.

9.3 Remove any parcels with a residential zoning area (use the zoning_o_3 column and the residential general zoning category).

9.4 What are your new dataset dimensions after removing these rows?

Q10

Narrow down your options to the 10 best parcels.

10.1 Using the GlobalTechSync preferences, create a ranking system to determine the top 10 parcels. Describe your system and explain how each preference fits in the system relative to the other preferences.

10.2 Using your ranking system, determine the top 10 best parcels to submit to GlobalTechSync and record the parcel FIDs below.

Q11

Comment on the selection process.

11.1 Was it easy or hard select the 10 best parcels? Why? Did you typically have too many parcels to choose from or too few?

11.2 How did you decide which values can be used as cut offs for continuous numerical fields? Are you happy with your available options? Why or why not?

11.3 Can you find a parcel that in your opinion perfectly satisfies all the requirements and preferences? Why or why not? What additional data would you like to have to make this decision?

Part 4: Final Report Presentation

Q12

Present your findings in your report.

12.1 Display graphs highlighting where your 10 final parcels are compared to the rest of the dataset for at least 3 numeric variables.

12.2 Create a chart showing qualitative variables for each of the 10 final parcels.

12.3 For each of the 10 final parcels list their strengths and weaknesses. If the parcels end up very similar to each other, propose a system to further rank each parcel and back up your decision.

12.4 Highlight any other important factors that can help make some of the parcels stand out or help the location scouts make the final decision (you may also mention factors that you do not think are represented in this dataset).

Part 5: Data Science Lifecycle

Q13

13.1 Using your favorite software tool (e.g. Google Draw), create a diagram of the Data Science Lifecycle you used for this project. Make sure that each action you performed to come to your recommendations for GlobalTechSync can be easily assigned to a step of the Lifecycle. Go ahead and make the assignments.

13.2 Clearly explain and describe each step of the Data Science Lifecycle for this project, making sure you indicate how each action you took for the project fits into the lifecycle.

13.3 How do you plan to make your raw data and workflow available to the GlobalTechSync location scouts if they want to check or understand your methods? What are the advantages and disadvantages of the plan you choose?

13.4 What steps can you take to make things easier for yourself for choosing a site in Austin for the next tech headquarters that is looking for a site? What advice can you give your college in Seattle who is undergoing a similar process?

Bonus: Implement your analysis in wholetale.org. Include its URL.

Appendix: Austin Zoning and Land Use Codes

Table 1. Base Zoning Districts

Residential		Commercial	
LA	Lake Austin Residence	NO	Neighborhood Office
RR	Rural Residence	LO	Limited Office
SF-1	Single Family—Large Lot	GO	General Office
SF-2	Single Family—Standard Lot	CR	Commercial Recreation
SF-3	Family Residence	LR	Neighborhood Commercial
SF-4A	Single Family—Small Lot	GR	Community Commercial
SF-4B	Single Family—Condominium	L	Lake Commercial
SF-5	Urban Family Residence	CBD	Central Business District
SF-6	Townhouse & Condominium	DMU	Downtown Mixed Use
MF-1	Multifamily —Limited Density	W/LO	Warehouse/limited Office
MF-2	Multifamily—Low Density	CS	General Commercial Services
MF-3	Multifamily—Medium Density	CS-1	Commercial-Liquor Sales
MF-4	Multifamily—Moderate Density	CH	Commercial Highway Serv
MF-5	Multifamily—High Density	**Special Purpose**	
MF-6	Multifamily—Highest Density	DR	Development Reserve
MH	Mobile Home Residence	AV	Aviation Services
Industrial		AG	Agricultural
IP	Industrial Park	P	Public
LI	Limited Industrial Services	PUD	Planned Unit Development
MI	Major Industry	TN	Traditional Neighborhood
R&D	Research & Development		

Table 2. Some of the Combining Zoning Districts

Code	District name
-CO	Conditional Overlay Combining District
-H	Historic Combining District
-MU	Mixed Use Combining District
-NP	Neighborhood Plan Combining District
-PDA	Planned Development Area
-V	Vertical Mixed Use Building

```
LAND_USE
The land use inventory code.

--------------------------------
100
Single Family -  One dwelling in a single building on one lot.  May be
attached to another unit, as long as there are separate parcels.  Includes
manufactured, non-mobile homes.
113
```

Mobile Homes - One or many dwellings in single buildings, designed to be mobile.
150
Duplexes - Two dwellings within a single building and on one lot, other than a mobile home.
160
Large-lot Single Family - One dwelling in one buidling on a parcel ten acres or greater, usually used as a farm, that can possibly be redeveloped for another use
210
Three/Fourplex - Three or four dwellings within individual buildings on one lot.
220
Apartment/Condo - More than five dwellings within individual buildings on one lot.
230
Group Quarters - One dwelling containing non-family occupants on one lot and providing. Occupants share services.
240
Retirement Housing - One dwelling containing non-family occupants on one lot and providing services to retirees
300
Commercial - Wholesale and retail trade and services. Includes trade of most durable and non-durable goods, building, hardware, garden, general retail mechandise, lumber, grocery, food sales, auto vehicle and gasoline sales, apparel and accessory stores, home furniture and equipment, eating and drinking, commercial art and craft studios, lodging hotels and motels, personal services, mini-warehousing and personal storage, automotive repair, automotive services, entertainment and recreation services, business services, commercial sports recreation and exercise, amusement services.
330
Mixed Use - One building containing both commericial and residential uses
400
Office - Includes accounting, architectural services, design services, engineering, insurance, law offices, organization/association's office, personnel, property management, real estate, secretarial services, telephone answering services, television/film/sound recording studios, travel agency, financial services, banks, savings and loans, credit unions, blood banks, treatment, and guidance centers, doctor, dental, psychological, and other medical offices, electronic, pharmaceutical, chemical, and other research and development services
510
Manufacturing - Basic, light, and custom industry and manufacturing, industrial arts and crafts
520
Warehousing - General, limited and commercial warehousing and distribution, not including mini-warehousing or personal storage
530
Miscellaneous Industrial - Heavy equipment sales and services, including automobiles and recreation vehicles, pool services, cans, paper, plastic, auto and junk recycling facilities, stables, kennels, pet services, and slaughterhouses. Does not include farms and ranches
560
Resource Extraction (Mining) - Quarries and oil and gas drilling facilities
570
Landfills - Processing and storage of garbage and other wastes
610

Semi-institutional Housing - Half-way houses, housing for mentally and psychologically handicapped
620
Hospital -
630
Government Services - Police stations, fire houses, post offices, jails, prisons, military installations
640
Educational - Day care, primary and secondary education, colleges, universities, business trade schools
650
Meeting and Assembly - Club or lodge halls, religious assembly, convention centers
670
Cemetery -
680
Cultural Services - Museums, libraries
710
Parks/Greenbelts - Open spaces set aside for current or future recreation or drainage
720
Golf Courses - Private and public golf courses and driving ranges
730
Camp Grounds - Private and public camp grounds for temporary use. Does not include permanent and semi-permanent RV parks.
740
Common Areas - Areas set aside for common use, typically privately owned areas that serve as drainage but are not registered as such by public agencies
750
Preserves - Open spaces set aside for preservation or protection
810
Railroad Facilities - Railroad stations and right-of-way
820
Transportation Facilities - Bus stations and other transportation facilities not used for aviation, railroads, or marinas
830
Aviation Facilities - Airports and aviation facilities
840
Marinas - Commercial and private marinas
850
Parking - Surface parking for a variety of establishments/parcels, including actual parking garage facilities or pay-for-parking lots. Parcels on separate lots that serve only one establishment are coded with the use of that establishment
860
Streets and Roads - Any road, street, or traffic island in the public right of way
870
Utilities - Electric, water, and wastewater utilities
900
Undeveloped - Parcels without structures that have the potential for development
910
Agricultural - Parcels that are predominantly used for either crops, livestock, animal husbandry, or other farmland but have the potential for development

940
Water - Areas permanently submerged in water
999
Unknown - Parcels where the land use can not be determined from available sources