

In the following parts, the author of this document is referred to as “the analyst”. For readability, instructions are printed in black, whereas the analyst’s responses are in blue.

Part 1: Data Pre-processing

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.

```
finalProjData = read.csv("D:/UIUC MISIM Studies/2019 Fall Courses/IS 457 Intro Data  
Sci/Assignments/Final/Austin_Lots.csv")  
  
summary(finalProjData)  
dim(finalProjData)
```

The initial dimensions of the datasets from several Austin city departments are obtained using the dim() command, and are returned as 26284 rows and 44 columns: meaning there are a total of 26284 data with a maximum of 44 possible attributes for each.

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?

Intuitively speaking, data that do not demonstrate or are unrelated to a site’s relevance as tech company headquarters are to be discarded first. These are identified by the student as:

(1) created_by: ID of employee who created initial record

- Whoever created the record should not have any impact on a site’s feasibility as headquarters. The quality of the data should matter more than whoever created them. This to prevent favoritism between the analyst and other staff members from diluting the objectiveness of the report.

(2) date_creat: Date record was initially created in Austin public database

- Without any knowledge of when this Austin public database was created, any site that existed before its existence would likely have identical date of creation, which would not be helpful in determining the age of any parcel or location. On a side note, a site’s being old or new does not necessarily impact its feasibility, and as such date created is determined as safe to delete.

(3) modified_b: ID of employee who last updated record

- While “date modified” would certainly be important, the person behind the modification should not matter as much. Suffice to say, if there were data which tell us the credibility/reliability of each staff, this column would be somewhat relevant (data modified by colleagues with good records are likely to be

trustworthy whereas those by people with poor records should be inspected cautiously~), but this dataset has no such thing. Therefore, the analyst believes it is safe to rule out this column with a reason similar to that of “created_by”.

(4) GEOID: US Census block group ID

- Supposedly this is a column that should be qualitative rather than quantitative, as “group ID” indicate a group of unique identifiers for data. However, for reasons unknown to the analyst, all the data in this column look identical and disturbingly numerical, which contradicts the way this column should be utilized. The analyst suspects a data corruption and therefore removes this column.

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

```
projDataSubsetted = subset(finalProjData,  
                           select = -c(created_by, date_creat, modified_b, GEOID))  
dim(projDataSubsetted)
```

The new dimensions are (26,284, 40) which means there are a total of 26284 data with a maximum of 40 possible attributes for each.

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

Aside from getting rid of irrelevant variables, subsetting data also creates a new dataset separate from the original which serves as an extra layer of safety. If, in the process of analysis, the dataset was to be wrongly manipulated and corrupted, the original dataset remains untouched and therefore allows the analyst to restore what was lost and remain on track.

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.

The column names are either left as is, renamed based on the reasoning detailed in the next paragraph, or already deleted in the previous process of subsetting.

The renaming decided by the analyst is done in such a way that:

- (1) The new names describe what all the columns do to the best extent possible within a string, without getting unnecessarily lengthy. When applicable, units of measurement are also indicated in the names (for example, “m” means “meters”, “m2” means “meters squared”, etc.).
- (2) Abbreviations which should indicate different things, do.
- (3) The new names are uniform in terms of capitalization. Only decapitalized characters with underscores between words expect abbreviations for landmarks (EWC for East-

- West Connector Highway, for instance). Underscores which do not separate words in sensible, grammatical ways are also deemed distracting and thus removed.
- (4) The 4 deleted columns are left with “DELETED” in cells that denote their new variable names.

For instance, “dis” should denote “distance” and should no longer be confused with “dist” which denotes “district”; the underscore at the end of “land_base_” accomplishes nothing syntactically, and as such the column would have been renamed as “land_base”, were it not renamed as something more indicative;

```
library(plyr)
projDataRenamed <- rename(projDataSubsetted, c(
  "block_id"="id_block",
  "land_base_"="id_austin_DB",
  "land_base1" = "land_type",
  "lot_id" = "id_lot",
  "date_modif" = "date_modified",
  "objectid" = "id_sec",
  "City_dist" = "dis_m_city_center",
  "Airpt_dist" = "dis_m_intl_airpt", |
  "district" = "dist_number",
  "Shape_Area" = "area_parcel_m2",
  "zoning_o_3" = "code_zoning",
  "zcta5ce10" = "code_zip",
  "LAND_USE_2" = "code_land_use_spec",
  "GENERAL_LA" = "code_land_use_inv",
  "EWC_dist" = "dis_m_EWC",
  "NSC_dist" = "dis_m_NSC",
  "Mopac_dist" = "dis_m_Mopac",
  "X130_dist" = "dis_m_high130",
  "X35_dist" = "dis_m_inter35",
  "ExTrail_1m" = "num_1mi_ex_trail",
  "PpTrail_1m" = "num_1mi_pp_trail",
  "conf" = "lvl_bike_comf",
  "bike_lanes" = "num_1mi_bike_ln",
  "Bus_area" = "bool_bus_sys",
  "TotBdgArea" = "area_total_bldgs",
  "Num_Bldgs" = "num_bldgs_on_parcel",
  "MaxBdgArea" = "area_largest_bldg",
  "tax_break2" = "pct_tx_brk_dist",
  "bk_tx_brk" = "pct_tx_brk_block",
  "Housing_" = "idx_housing_opp",
  "Education" = "idx_ed_opp",
  "Economic_" = "idx_econ_opp",
  "Comprehens" = "idx_comp_opp",
  "Med_HH_Inc" = "med_HHI_perzip",
  "Med_rent" = "med_rent_perzip",
  "Med_home" = "med_home_price_perzip",
  "Aff_rent_t" = "pct_aff_units_perzip",
  "Aff_own_te" = "pct_aff_homes_perzip",
  "Descriptio" = "descr_const_nearby"
))
```

Variable name (Original)	Description	Variable name (New)
FID	Unique row ID	FID
block_id	Identifies Austin city block if applicable	id_block
created_by	ID of employee who created initial record	DELETED (irrelevant)
date_creat	Date record was initially created in Austin public database	DELETED (irrelevant)
land_base_	Austin parcel record database ID	id_austin_DB
land_base1	Land unit type designation	land_type
lot_id	Austin lot ID	id_lot

IS 457 FA19
FINAL REPORT
CLASS ID: 104

modified_b	ID of employee who last updated record	DELETED (irrelevant)
date_modif	Date record was last modified in Austin public database	date_modified
objectid	Austin secondary ID	id_sec
City_dist	Distance in meters from parcel to the Austin City Center	dis_m_city_center
Airpt_dist	Distance in meters from parcel to the Austin International Airport	dis_m_intl_airpt
district	Texas service district number	dist_number
Shape_Area	Area of parcel in meters squared	area_parcel_m2
zoning_o_3	Austin zoning designation. See Appendix for code description	code_zoning
zcta5ce10	Zip code	code_zip
LAND_USE_2	Specific land use designation	code_land_use_spec
GENERAL_LA	General land use designation. See Appendix for code description	code_land_use_inv
EWC_dist	Distance in meters from parcel to East-West Connector Highway	dis_m_EWC
NSC_dist	Distance in meters from parcel to North-South Connector Highway	dis_m_NSC
Mopac_dist	Distance in meters from parcel to Mopac Freeway	dis_m_Mopac
X130_dist	Distance in meters from parcel to Highway 130	dis_m_high130
X35_dist	Distance in meters from parcel to Interstate 35	dis_m_inter35
ExTrail_1m	Number of existing urban trails within 1 mile of parcel	num_1mi_ex_trail
PpTrail_1m	Number of proposed urban trails within 1 mile of parcel	num_1mi_pp_trail

IS 457 FA19
FINAL REPORT
CLASS ID: 104

conf	Average bike lane comfort level (0 is most comfortable, 4 is least)	lvl_bike_comf
bike_lanes	Number of bike lanes within 1 mile of parcel	num_1mi_bike_ln
Bus_area	1 if parcel is in Austin Bus system service area, otherwise 0	bool_bus_sys
TotBdgArea	Total area of buildings on parcel	area_total_bldgs
Num_Bldgs	Number of buildings on parcel	num_bldgs_on_parcel
MaxBdgArea	Area of largest building on parcel	area_largest_bldg
tax_break2	District wide construction perk – percentage of parcel purchase cost waived (9.5 = 9.5%)	pct_tx_brk_dist
bk_tx_brk	Block wide construction perk – percentage of construction fees waived in the first 5 years (0.026 = 2.6%)	pct_tx_brk_block
GEOID	US Census block group ID	DELETED (data corruption)
Housing__	Housing opportunity index: a higher value indicates an individual in this block group has more opportunity to find affordable housing	idx_housing_opp
Education	Education opportunity index: a higher value indicates an individual in this block group has more opportunity to obtain a high level of education	idx_ed_opp
Economic__	Economic opportunity index: a higher value indicates an individual in this block group has more opportunity to achieve economic stability	idx_econ_opp
Comprehens	Comprehensive opportunity index: a higher value	idx_comp_opp

IS 457 FA19
FINAL REPORT
CLASS ID: 104

	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_HHI_perzip
Med_rent	Median rent per zip code	med_rent_perzip
Med_home	Median home price per zip code	med_home_price_perzip
Aff_rent_t	Percentage of rental units per zip code that are affordable for an average worker in tech to rent	pct_aff_units_perzip
Aff_own_te	Percentage of homes per zip code that are affordable for an average worker in tech to rent	pct_aff_homes_perzip
Descriptio	Description of 2019 construction near the parcel	descr_const_nearby

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.

Using an “sapply” function on the renamed data frame, any columns with missing values, empty string, 0, string-fied missing values or “-” are listed out.

```
##{r}
# Q2:

# 2.1
#columnsWithNA <- colnames(projDataRenamed)[colSums(is.na(projDataRenamed)) > 0]
#print(columnsWithNA)

allMissingColumns <- sapply(projDataRenamed, function(x) any(is.na(x) | x == "" | x == 0 | x ==
"N/A" | x == "NA" | x == "-"))
allcolumnNamesWhichMiss <- names(allMissingColumns[allMissingColumns>0])
print("columns with any values missing or recognized as NA, empty string, numerical 0,
string-fied NA, stringfied NA, or '-'")
print(allcolumnNamesWhichMiss)
...
```

```
[1] "columns with any values missing or recognized as NA, empty string, numerical 0, string-fied
NA, stringfied NA, or '-'"
[1] "FID" "id_lot" "date_modified"
[4] "dis_m_city_center" "code_land_use_spec" "code_land_use_inv"
[7] "dis_m_EWC" "num_lmi_ex_trail" "num_lmi_pp_trail"
[10] "lvl_bike_comf" "num_lmi_bike_ln" "bool_bus_sys"
[13] "area_total_bldgs" "num_bldgs_on_parcel" "area_largest_bldg"
[16] "pct_tx_brk_dist" "pct_tx_brk_block" "idx_housing_opp"
[19] "idx_ed_opp" "idx_econ_opp" "idx_comp_opp"
[22] "med_HHI_perzip" "med_rent_perzip" "med_home_price_perzip"
[25] "pct_aff_units_perzip" "pct_aff_homes_perzip" "descr_const_nearby"
```

The search terms are then made more lenient, particularly without “0” and empty strings:

```
# 2.1
#columnsWithNA <- colnames(projDataRenamed)[colSums(is.na(projDataRenamed)) > 0]
#print(columnsWithNA)

allMissingColumns <- sapply(projDataRenamed, function(x) any(is.na(x) | x == "-" | x == "N/A" | x
== "NA" | x == "" | x == "0"))
allcolumnNamesWhichMiss <- names(allMissingColumns[allMissingColumns>0])
print("columns with any values containing NA, empty string, string-fied NA, stringfied NA, or
hyphen")
print(allcolumnNamesWhichMiss)
...
```

```
[1] "columns with any values containing NA, empty string, string-fied NA, stringfied NA, or
hyphen"
[1] "id_block" "land_type" "id_lot"
[4] "date_modified" "code_zoning" "idx_housing_opp"
[7] "idx_ed_opp" "idx_econ_opp" "idx_comp_opp"
[10] "med_HHI_perzip" "med_rent_perzip" "med_home_price_perzip"
[13] "pct_aff_units_perzip" "pct_aff_homes_perzip" "descr_const_nearby"
```

The resulting 15 columns are therefore recognized as missing values or containing empty strings.

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.

The columns are each listed in plain English as follow:

First off, the 5 numerical columns which would be put into quantitative analysis very soon, having missing values could impede the analyst's abilities to conduct meaningful research. At the same time, they are too important to be deleted entirely, so deletion of rows that contain such values is more feasible.

(1) `med_HHI_perzip`: Median household income per zip code

(2) `med_rent_perzip`: Median rent per zip code

(3) `med_home_price_perzip`: Median home price per zip code

Households having median income as missing value is unlikely. Even households entirely made up by students should at least have 0 income documented. Similar reasoning can be applied to assert that median rent and prices cannot be empty. As such, parcels with unknown values in this part should be discarded.

(4) `pct_aff_units_perzip`: Percentage of rental units per zip code that are affordable for an average worker in tech to rent

(5) `pct_aff_homes_perzip`: Percentage of homes per zip code that are affordable for an average worker in tech to rent

For percentages of rental units and homes affordable for an average tech worker, either containing missing values would be problematic, as such parcels would not possess enough information for discerning its feasibility in the latter section.

Moving on to qualitative columns, we have the following:

(6) `id_block`: Identifies Austin city block if applicable

(7) `land_type`: Land unit type designation

(8) `id_lot`: Austin lot ID

(9) `date_modified`: Date record was last modified in Austin public database

(10) `descr_const_nearby`: Description of 2019 construction near the parcel

The instruction does note “if applicable” so missing values in block IDs are more than fine; Date last modified do not impact analysis as much so empty values and strings would be fine. Lastly, description of 2019 construction nearby is purely optional. There could be nothing happening near the parcel that is worth documenting. As such, these 4 columns should be fine as they are if any contains missing value or empty strings, but for data cleanliness it would be best to fill empty spaces with “Not Applicable” labels. However, missing Lot IDs is far less excusable ---- if it was, the instruction would have explicitly stated so as it did for Block IDs. These are possibly of equal importance as land unit type records, which contain information that complement Austin zoning designations so empty. Any parcel missing such is likely not legitimate.

(11) `code_zoning`: Austin zoning designation. See Appendix for code description

(12) `idx_housing_opp`: Housing opportunity index

(13) `idx_ed_opp`: Education opportunity index

(14) `idx_econ_opp`: Economic opportunity index

(15) `idx_comp_opp`: Comprehensive opportunity index

The appendix which explains zoning designation is very important to determine the use of land. Any parcel with missing values or empty strings in this part is at least dangerous and risky.

All the four opportunity indexes are crucial indicators of individuals in a parcel. Any land without any of them will be impossible to evaluate for the client. Therefore, any row missing values in any of the 5 will be subject to deletion.

In summary, the analyst will implement:

Step 1. Delete all rows with missing values or empty strings in any of the 7 qualitative columns:

`id_lot` , `code_zoning`, `idx_housing_opp`, `idx_ed_opp`, `idx_econ_opp`, `idx_comp_opp`, `land_type`

Step 2. Further delete all rows missing values in any of the 5 quantitative columns:

`med_HHI_perzip`, `pct_aff_units_perzip`, `pct_aff_homes_perzip`, `med_rent_perzip`,
`med_home_price_perzip`

Step 3. Replace empty strings in the following columns with “Not Applicable” labels:

`id_block`, `date_modified`, `descr_const_nearby`

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

Ultimately, the choice of method(s) will ensure that only data that are complete and relevant will stay, and that the analysis will be clean and free from corruption (as best as manageable).

This relevance is defined by possessing all the opportunity indexes, land use designation, and quantitative indicators of land affordability. All the aforementioned are necessary for assessing any parcel, and missing values in these cannot be redeemed.

2.4 Implement your methods for dealing with the missing values.

```
##{r}
# 2.4

unwantedValues = c("-", "N/A", "NA", "", " ")
pj <- projDataRenamed


projDataCleansed1 <- pj[!(
  # Delete rows with unwanted qualitative values
  is.na(pj$id_lot) | pj$id_lot %in% unwantedValues |
  is.na(pj$code_zoning) | pj$code_zoning %in% unwantedValues |
  is.na(pj$idx_housing_opp) | pj$idx_housing_opp %in% unwantedValues |
  is.na(pj$idx_ed_opp) | pj$idx_ed_opp %in% unwantedValues |
  is.na(pj$idx_econ_opp) | pj$idx_econ_opp %in% unwantedValues |
  is.na(pj$idx_comp_opp) | pj$idx_comp_opp %in% unwantedValues |
  is.na(pj$land_type) | pj$land_type %in% unwantedValues |
  # Delete rows with unwanted quantitative values
  is.na(pj$med_HHI_perzip) | pj$med_HHI_perzip %in% unwantedValues |
  is.na(pj$pct_aff_units_perzip) | pj$pct_aff_units_perzip %in% unwantedValues |
  is.na(pj$pct_aff_homes_perzip) | pj$pct_aff_homes_perzip %in% unwantedValues |
  is.na(pj$med_rent_perzip) | pj$med_rent_perzip %in% unwantedValues |
  is.na(pj$med_home_price_perzip) | pj$med_home_price_perzip %in% unwantedValues), ]

#assign "Not Applicable" labels to missing Block IDs
projDataCleansed1$id_block[projDataCleansed1$id_block %in% unwantedValues] <- NA
projDataCleansed1$date_modified[projDataCleansed1$date_modified %in% unwantedValues] <- NA
projDataCleansed1$land_type[projDataCleansed1$land_type %in% unwantedValues] <- NA
projDataCleansed1$descr_const_nearby[projDataCleansed1$descr_const_nearby %in% unwantedValues] <- NA

dim(projDataCleansed1)
View(projDataCleansed1)
head(projDataCleansed1)
```

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

```
dim(projDataCleansed1)
View(projDataCleansed1)
head(projDataCleansed1)
```



[1] 16491 40

```
dim(projDataCleansed1)
View(projDataCleansed1)
head(projDataCleansed1)
```

	FID	id_block	id_austin_DB	land_type	id_lot	date_modified	id_sec
	<int>	<chr>	<int>	<chr>	<chr>	<chr>	<int>
2	1	NA	1676746	LOT	18	6/3/2008 0:00	296037
3	2	NA	1839096	LOT	6	NA	319082
4	3	A	1909677	LOT	15B-1A	NA	333367
10	9	NA	1659892	Lot	8	12/12/2007 0:00	147975
11	10	1	1820935	Lot	15	12/5/2008 0:00	216517
12	11	2	1691045	Lot	16	11/13/2008 0:00	259788

6 rows | 1-8 of 40 columns

The new dimensions are 16491 records with 40 columns each. In addition, a “head()” to inspect the first 6 data.

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.

land_base1 has been renamed as land_type, which has its unique value set and occurrences of each displayed using the command below:

```
##{r}
#Q3:
# 3.1
table(projDataCleansed1$land_type)
##

```

	Lot	LOT	lott	OTHER	Parcel	PARCEL	PCL	Tract	TRACT	
	0	1383	15037	1	0	0	47	0	0	23

```
##{r}
```

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?)

```
##{r}
# 3.2
class(projDataCleansed1$land_type)
levels(projDataCleansed1$land_type)
##

```

```
[1] "factor"
[1] " " "Lot" "LOT" "lott" "OTHER" "Parcel" "PARCEL" "PCL" "Tract" "TRACT"
```

With the command above, it was deduced that the column is a “factor” which means it takes predefined, finite categorical data. There was only one “lott” entry in the data which is probably a typo, the same can be said for the “Lot”, “Parcel” and “Tract”, which are merely different capitalization of “LOT”, “PARCEL” and “TRACT”.

The following command is therefore used to standardize the values while verifying:

```
projDataCleansed1$land_type[projDataCleansed1$land_type == "Lot" ] <- "LOT"
projDataCleansed1$land_type[projDataCleansed1$land_type == "lott" ] <- "LOT"
projDataCleansed1$land_type[projDataCleansed1$land_type == "Parcel" ] <- "PARCEL"
projDataCleansed1$land_type[projDataCleansed1$land_type == "Lot" ] <- "LOT"
projDataCleansed1$land_type[projDataCleansed1$land_type == "Tract" ] <- "TRACT"
table(projDataCleansed1$land_type)
```

	Lot	LOT	lott	OTHER	Parcel	PARCEL	PCL	Tract	TRACT	
	0	0	16421	0	0	0	47	0	0	23

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.

This column has been renamed as pct_tx_brk_dist to demonstrate that it denotes tax break perks based on districts. The column's class and current instances which contain dollar signs are obtained through the following commands:

```
## {r}
# 3.3

class(projDataCleansed1$pct_tx_brk_dist)
table(projDataCleansed1$pct_tx_brk_dist)|

##
```

```
[1] "factor"

$0.98      $3.11      $7.73      $9.52      0 0.0213235 0.122498
3          1          0          1          6307      42         4
0.24291    0.255072   0.256544   0.364723   0.632385   0.67714    0.785669
411        4          45         0          2          8         1
0.983745   1.14823    1.2197     1.2659     1.50366    1.51293    1.64689
213        59        14         106        0          440        113
1.7621     2.00828    2.03444    2.0594499  2.20244    2.2168901  2.27542
0          134       36         29         0          206        194
```

For the four cases where various values containing, the following command is used to replace them with actual numbers of the same values. In addition, the class of this column is changed to be "numerical":

```
# 3.3 cont

pct_tx_brk_dist <- sapply(projDataCleansed1, is.factor)
projDataCleansed1[pct_tx_brk_dist] <- lapply(projDataCleansed1[pct_tx_brk_dist],
function(x)
      as.factor(gsub("\\$", "", x)))

projDataCleansed2 <- transform(projDataCleansed1, pct_tx_brk_dist =
as.numeric(paste(pct_tx_brk_dist)))

table(projDataCleansed2$pct_tx_brk_dist)

##
```

```
[1] "numeric"
[1] "numeric"

0 0.0213235 0.122498 0.24291 0.255072 0.256544 0.632385 0.67714
6307 42 4 411 4 45 2 8
0.785669 0.98 0.983745 1.14823 1.2197 1.2659 1.51293 1.64689
1 3 213 59 14 106 440 113
2.00828 2.03444 2.0594499 2.2168901 2.27542 2.2965901 2.3626299 2.4780099
134 36 29 206 194 2 83 238
2.49894 2.5534999 2.56007 2.7440901 2.75211 2.8912899 2.91031 3.0353301
506 16 34 267 376 57 294 9
3.11 3.11113 3.2132199 3.3429899 3.5039101 3.58902 3.6100199 3.8810999
3 303 34 338 34 105 286 3
```

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.

Because GEOID was among the first column to have been deleted by the analyst, the original data is required. They are inspected as follow:

```
## {r}
# 3.4
table(finalProjData$GEOID)
#>
#> 4.85e+11
#> 26116
```

This means that there were 26,116 data recorded as 4.85e+11 in scientific notation. As integer, they would appear as 485,000,000,000 ---- that is only 1 unique value. This is a bad thing because it goes against the purpose of a qualitative column like GEOID. The column is supposed to be contain identifiers, not numeric values to be calculated. As such, the column is rendered useless and is thus already deleted prior to this sub-question.

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.

```
## {r}
# 3.5
projDataCleansed2$FID[duplicated(projDataCleansed2$FID)]
projDataCleansed2$id_austin_DB[duplicated(projDataCleansed2$id_austin_DB)]
#>
#> [1] 376
#> [1] 1970747
```

Using the commands here, it is verified that the row going by the FID 376 and ID 1970747 in the Austin Database is duplicated:

A subsequent command is then used to obtain the duplicates. However, the queried results show that values in the two rows are entirely identical. There is no difference in deleting either:

```
projDataCleansed2[projDataCleansed2$FID == 376,]
```

FID	id_block	id_austin_DB	land_type	id_lot	date_modified	id_sec	dis_m_city_center	dis_m_intl_airprt	dist_number	area_parcel_m2
377	376	2	1970747	LOT	29	NA	2194.53	9078.47	14	5770.635
378	376	2	1970747	LOT	29	NA	2194.53	9078.47	14	5770.635

2 rows | 1-8 of 40 columns

code_zoning	code_zip	code_land_use_spec	code_land_use_inv	dis_m_FWC	dis_m_NSC	dis_m_Mopac	dis_m_high130	dis_m_inter35
NP	78702	98593	100	5255.58	2534.36	5644.46	9460.05	2252.81
NP	78702	98593	100	5255.58	2534.36	5644.46	9460.05	2252.81

2 rows | 17-21 of 40 columns

num_1mi_ex_trail	num_1mi_pp_trail	lvl_bike_comf	num_1mi_bike_ln	bool_bus_sys	area_total_bldgs	num_bldgs_on_parcel
0	12	2	7	1	194.749	3
0	12	2	7	1	194.749	3

2 rows | 26-28 of 40 columns

area_largest_bldg	pct_tx_brk_dist	pct_tx_brk_block	idx_housing_opp	idx_ed_opp	idx_econ_opp	idx_comp_opp	med_HHI_perzip
176.742	3.94564	0.0923577	Very Low	Very Low	Moderate	Very Low	34734
176.742	3.94564	0.0923577	Very Low	Very Low	Moderate	Very Low	34734

2 rows | 29-32 of 40 columns

2 rows | 33-36 of 40 columns

IS 457 FA19
FINAL REPORT
CLASS ID: 104

med_rent_perzip <int>	med_home_price_perzip <int>	pct_aff_units_perzip <int>	pct_aff_homes_perzip <int>
766	175400	99	67
766	175400	99	67

Therefore, the analyst defaults to using the method generally agreed to be a safe method of removing duplicates: creating a new data frame using the following commands, source:

<https://stackoverflow.com/questions/13967063/remove-duplicated-rows>

```
projDataCleansed3 <- projDataCleansed2[!duplicated(projDataCleansed2), ]
dim(projDataCleansed2)
dim(projDataCleansed3)
```

data.frame
2 x 40

R Console

```
[1] 16491 40
[1] 16490 40
```

As shown here, the dimension after the calibration is exactly 1 row less than the previous: 16490 rows and 40 columns.

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.

This column has been renamed as code_land_use_spec prior to this section. The command to drop it is done through subsets, with commands to confirm that the new dimension is now down by 1 column, with 39 columns remaining:

```
# 3.6
projDataCleansed4 <- subset(projDataCleansed3, select = -c(code_land_use_spec))
dim(projDataCleansed4)
head(projDataCleansed4)
```

R Console

data.frame
6 x 39

```
[1] 16490 39
```

	FID <int>	id_block <fctr>	id_austin_DB <int>	land_type <fctr>	id_lot <fctr>	date_modified <fctr>	id_sec <int>	dis_m_city_center <dbl>
2	1	NA	1676746	LOT	18	6/3/2008 0:00	296037	3203.180
3	2	NA	1839096	LOT	6	NA	319082	3187.940
4	3	A	1909677	LOT	15B-1A	NA	333367	3089.870
10	9	NA	1659892	LOT	8	12/12/2007 0:00	147975	209.848
11	10	1	1820935	LOT	15	12/5/2008 0:00	216517	147.358
12	11	2	1691045	LOT	16	11/13/2008 0:00	259788	159.042

6 rows | 1-9 of 39 columns

The metadata are long-term records so ensuring their consistency is vital for keeping any data science activity reproducible. Keeping consistent entries also enables the dataset to be used in

other projects, or when the same project needs to be replicated for demonstration or other purposes.

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

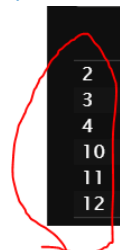
Any data out of sync would impede later analysis and jeopardize the entire task. In the case of duplicated rows, depending on the number of duplicates fundamental statistics like mode can be distorted. Even 1 duplicate can distort the means of all numerical columns. A sufficiently great number of duplicates can even mess with the entire distribution, making skewed ones deceptively normal and vice versa.

In the case of missing values, they can contribute to trends being wrongly interpreted, depending on how the the functions designed to analyze them handles null values and/or empty strings.

In summary, without data cleaning, analyses will produce distorted results.

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).

In the process of data cleaning, the analyst has already deleted quite a few rows and columns. As a result, the indexes R uses to track the remaining 16,490 data are no longer consecutive (shown by the screenshot below).



	FID <int>	id_block <fctr>	id_austin_DB <int>	land_type <fctr>	id_lot <fctr>	date_modified <fctr>	id_sec <int>	dis_m_city_center <dbl>
2	1	NA	1676746	LOT	18	6/3/2008 0:00	296037	3203.180
3	2	NA	1839096	LOT	6	NA	319082	3187.940
4	3	A	1909677	LOT	15B-1A	NA	333367	3089.870
10	9	NA	1659892	LOT	8	12/12/2007 0:00	147975	209.848
11	10	1	1820935	LOT	15	12/5/2008 0:00	216517	147.358
12	11	2	1691045	LOT	16	11/13/2008 0:00	259788	159.042

If, in any later section, this interferes with the quality of any analysis or prevents any command from working, a method to refresh the order and make it consecutive once more.

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.

This command displays the initial variable classes:

```
# 4.1
sapply(projDataCleansed4, class)
...
```

FID	id_block	id_austin_DB	land_type
"integer"	"factor"	"integer"	"factor"
id_lot	date_modified	id_sec	dis_m_city_center
"factor"	"factor"	"integer"	"numeric"
dis_m_intl_airpt	dist_number	area_parcel_m2	code_zoning
"numeric"	"integer"	"numeric"	"factor"
code_zip	code_land_use_inv	dis_m_EWC	dis_m_NSC
"integer"	"integer"	"numeric"	"numeric"
dis_m_Mopac	dis_m_high130	dis_m_inter35	num_1mi_ex_trail
"numeric"	"numeric"	"numeric"	"integer"
num_1mi_pp_trail	lvl_bike_comf	num_1mi_bike_ln	bool_bus_sys
"integer"	"integer"	"integer"	"integer"
area_total_bldgs	num_bldgs_on_parcel	area_largest_bldg	pct_tx_brk_dist
"numeric"	"integer"	"numeric"	"numeric"
pct_tx_brk_block	idx_housing_opp	idx_ed_opp	idx_econ_opp
"numeric"	"factor"	"factor"	"factor"
idx_comp_opp	med_HHI_perzip	med_rent_perzip	med_home_price_perzip
"factor"	"integer"	"integer"	"integer"
pct_aff_units_perzip	pct_aff_homes_perzip	descr_const_nearby	
"integer"	"integer"	"factor"	

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.

The two columns that stand out as unusual for the analyst are:

(1) **pct_aff_units_perzip**: Percentage of rental units per zip code that are affordable for an average worker in tech to rent

(2) **pct_aff_homes_perzip**: Percentage of homes per zip code that are affordable for an average worker in tech to rent

These two have the class of "integer" which prevents decimals from being assigned to them. Considering how the other two columns of similar nature are numerical (pct_tx_brk_dist and pct_tx_brk_block; both represent tax breaks) and that these two columns are not obligated to be added up as exactly 100% (for example, if an analysis is about election, the votes in percentages must be aggregated as such), limiting them to be integers lowers the precision these variables can be. In the real world, it is certainly possible that the ratio of affordable houses to people have decimals (1 to 3 is easily 33.33%, for instance).

Therefore, the two columns should be made numerical.

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

The aforementioned columns which record affordability are changed with these commands:

```
##{r}
# 4.3

projDataCleansed5 <- projDataCleansed4
pctTobeConverted <- c("pct_aff_units_perzip", "pct_aff_homes_perzip")
projDataCleansed5[pctTobeConverted] <- lapply(projDataCleansed5[pctTobeConverted], as.numeric)
head(projDataCleansed5)
sapply(projDataCleansed5, class)
##
```

num_1mi_pp_trail	lvl_bike_comf	num_1mi_bike_ln	bool_bus_sys
"integer"	"integer"	"integer"	"integer"
area_total_bldgs	num_bldgs_on_parcel	area_largest_bldg	pct_tx_brk_dist
"numeric"	"integer"	"numeric"	"numeric"
pct_tx_brk_block	idx_housing_opp	idx_ed_opp	idx_econ_opp
"numeric"	"factor"	"factor"	"factor"
idx_comp_opp	med_HHI_perzip	med_rent_perzip	med_home_price_perzip
"factor"	"integer"	"integer"	"integer"
pct_aff_units_perzip	pct_aff_homes_perzip	descr_const_nearby	
"numeric"	"numeric"	"factor"	

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?

(1) `read.table()`

<https://www.rdocumentation.org/packages/utils/versions/3.6.1/topics/read.table>

Supposedly, `read.table()` achieves the same thing as `read.csv()` but it cannot read tab-delimited files. In the file for this project, information is separated as one record per line. Therefore, this alternative is not useful here.

(2) `scan()`

<https://www.rdocumentation.org/packages/base/versions/3.6.1/topics/scan>

The function `scan()` can be handy when data vectors are small. However, the data read in would be numeric by default unless customized by the user. The file in question is one that possesses a mix of numeric and string-fied values in rather large quantity. Much specification would have been required, making this method more difficult to use.

Once a analysis starts, it will be costly to return and correct the way data are imported. It is for this reason that the analyst believes it important to reflect on alternatives after data cleaning and before committing to deeper analysis because some of them could be beneficial be it it terms of performance or otherwise.

IS 457 FA19
FINAL REPORT
CLASS ID: 104

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.

Summary of the current dataset should provide a clue which statistics may be of interest:

FID	id_block	id_austin_DB	land_type	id_lot
Min. : 1	A : 1000	Min. : 1635655	LOT : 16420	1 : 1386
1st Qu.: 6533	1 : 796	1st Qu.: 1715402	PARCEL : 47	2 : 1249
Median : 13226	2 : 668	Median : 1792824	TRACT : 23	3 : 1059
Mean : 13237	B : 651	Mean : 30803123		4 : 971
3rd Qu.: 19951	3 : 631	3rd Qu.: 1871003		5 : 895
Max. : 26277	(Other) : 5670	Max. : 400842667		6 : 862
	NA's : 7074			(Other) : 10068
date_modified	id_sec	dis_m_city_center	dis_m_intl_airpt	
12/12/2007 0:00 : 563	Min. : 3	Min. : 0	Min. : 1178	
3/28/2012 0:00 : 346	1st Qu.: 93111	1st Qu.: 1773	1st Qu.: 7736	
9/19/2012 0:00 : 299	Median : 186071	Median : 2581	Median : 9364	
4/2/2009 0:00 : 241	Mean : 186316	Mean : 3152	Mean : 8818	
4/1/2009 0:00 : 185	3rd Qu.: 278509	3rd Qu.: 3401	3rd Qu.: 10600	
(Other) : 4480	Max. : 375405	Max. : 13573	Max. : 13745	
NA's : 10376				
dist_number	area_parcel_m2	code_zoning	code_zip	code_land_use_inv
Min. : 14.00	Min. : 104	NP : 12710	Min. : 78617	Min. : 0.0
1st Qu.: 14.00	1st Qu.: 5710	UNO : 524	1st Qu.: 78702	1st Qu.: 100.0
Median : 14.00	Median : 6842	TOD : 495	Median : 78704	Median : 100.0
Mean : 15.26	Mean : 18278	SF-4A-NP : 276	Mean : 78714	Mean : 277.4
3rd Qu.: 14.00	3rd Qu.: 8792	ERC : 179	3rd Qu.: 78741	3rd Qu.: 400.0
Max. : 21.00	Max. : 5789225	NCCD-NP : 162	Max. : 78744	Max. : 940.0
		(Other) : 2144		
dis_m_EWC	dis_m_NSC	dis_m_Mopac	dis_m_high130	dis_m_inter35
Min. : 0	Min. : 6.676	Min. : 587.6	Min. : 544.6	Min. : 21.81
1st Qu.: 2484	1st Qu.: 2431.612	1st Qu.: 3954.3	1st Qu.: 9073.1	1st Qu.: 675.53
Median : 4738	Median : 3804.105	Median : 4884.8	Median : 10393.0	Median : 1348.34
Mean : 4457	Mean : 3588.572	Mean : 5544.6	Mean : 9965.9	Mean : 1920.94
3rd Qu.: 6349	3rd Qu.: 4689.675	3rd Qu.: 6387.3	3rd Qu.: 11398.0	3rd Qu.: 2508.41
Max. : 8726	Max. : 7576.530	Max. : 16296.5	Max. : 14457.2	Max. : 10607.40
num_lmi_ex_trail	num_lmi_pp_trail	lvl_.bike_comf	num_lmi_bike_ln	bool_bus_sys
Min. : 0.000	Min. : 1.00	Min. : 0.000	Min. : 0.00	Min. : 0.0000
1st Qu.: 0.000	1st Qu.: 7.00	1st Qu.: 1.000	1st Qu.: 11.00	1st Qu.: 1.0000
Median : 0.000	Median : 13.00	Median : 1.000	Median : 15.00	Median : 1.0000
Mean : 1.643	Mean : 14.21	Mean : 1.593	Mean : 15.32	Mean : 0.9992
3rd Qu.: 4.000	3rd Qu.: 20.00	3rd Qu.: 2.000	3rd Qu.: 20.00	3rd Qu.: 1.0000
Max. : 20.000	Max. : 45.00	Max. : 4.000	Max. : 52.00	Max. : 1.0000
area_total_bldgs	num_bldgs_on_parcel	area_largest_bldg	pct_tx_brk_dist	pct_tx_brk_block
Min. : 0.0	Min. : 0.000	Min. : 0.0	Min. : 0.000	Min. : 0.00000
1st Qu.: 123.1	1st Qu.: 1.000	1st Qu.: 102.6	1st Qu.: 0.000	1st Qu.: 0.00000
Median : 213.9	Median : 2.000	Median : 161.6	Median : 2.275	Median : 0.00000
Mean : 558.7	Mean : 1.768	Mean : 441.1	Mean : 3.054	Mean : 0.01447
3rd Qu.: 358.0	3rd Qu.: 2.000	3rd Qu.: 242.8	3rd Qu.: 5.438	3rd Qu.: 0.01384
Max. : 42389.8	Max. : 92.000	Max. : 33872.0	Max. : 9.953	Max. : 0.10000

IS 457 FA19
FINAL REPORT
CLASS ID: 104

idx_housing_opp	idx_ed_opp	idx_econ_opp	idx_comp_opp	med_HHI_perzip
Low : 4118	High : 937	High :3033	High :1422	Min. : 0
Moderate: 326	Low :3829	Low :2786	Low :2899	1st Qu.:30183
Very Low:12046	Moderate :2285	Moderate :4356	Moderate :2640	Median :34734
	Very High:1090	Very High:5318	Very High: 879	Mean :36533
	Very Low :8349	Very Low : 997	Very Low :8650	3rd Qu.:41056
				Max. :92606

med_rent_perzip	med_home_price_perzip	pct_aff_units_perzip	pct_aff_homes_perzip
Min. : 0.0	Min. : 0	Min. : 0.00	Min. : 0.00
1st Qu.: 766.0	1st Qu.:120200	1st Qu.: 99.00	1st Qu.: 67.00
Median : 835.0	Median :175400	Median : 99.00	Median : 67.00
Mean : 878.7	Mean :197114	Mean : 98.35	Mean : 67.35
3rd Qu.: 940.0	3rd Qu.:265100	3rd Qu.:100.00	3rd Qu.: 93.00
Max. :1590.0	Max. :621900	Max. :100.00	Max. :100.00

```
descr_const_nearby
new solar installation for new residence
: 623
Adding equipment to existing wireless telecommunication tower
: 256
Install subpanel and outlet for food truck run existing service from meter to food truckREFER 969712J
: 230
interior renovations through out school campus bringing stage access ramp ADTTAS compliant ADTTAS COMPLICATION TO RESTROOM
LOCKER ROOM BASKETS: 222
Interior Remodel to Existing Restrooms to ADA Replace Park Lighting
: 221
Replace Tub to Existing Multi Family UNITS 101103104
: 180
(Other)
:14758
```

5.2 Describe anything you find that is unexpected or interesting.

- (1) Despite being entirely qualitative, there are many identical entries in the `descr_const_nearby` (Descriptio in the original data) column. It can be inferred that many parcels are near the same construction sites or places where the same events occur.
- (2) The three (median) columns for household income, home price and rent all have means higher than medians, indicating that they are skewed to the right, and there are outliers of exceptionally high values that drag the mean upwards. The same can be said for prices of immovable assets. The distribution of wealth in the city of Austin is, similar to most metropolises, in such a way that the minority rich hold more than the majority poor --- though not unexpected, this is worth pointing out.
- (3) Of the four opportunity indexes, the Economic one stands out as an anomaly. It has “Very High” as mode whereas the other three have “Very Low” as modes; this does not quite add up as the Comprehensive one is supposedly contributed to by the others.
- (4) There are several identical records for Austin Lot IDs and land types. They have modes which can analyzed if needs be. Suffice to say, the appendix will have to be referenced to deduce the actual meanings are.

(5) The percentages of affordable rental units and homes per zip imply that rental units is more feasible for an average tech worker than purchasing homes. The median and mean for the former are 99% and 98.35%, whereas the ones for the latter are 67% and 67.35%: almost every average tech worker can afford rents, but only about 2 out of 3 can afford homes.

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.

Picking two sets of variables, the analyst utilizes different graphs to demonstrate them in manners most suitable for each:

(1) Distance(s) to landmarks (Histogram and Boxplot) (numerical, quantitative):

Since all parcels should have distances to all landmarks in the real world, we can use the following code to map the distribution of the 7 columns recording them:

dis_m_city_center: Distance in meters from parcel to the Austin City Center

dis_m_EWC: Distance in meters from parcel to East-West Connector Highway

dis_m_high130: Distance in meters from parcel to Highway 130

dis_m_inter35: Distance in meters from parcel to Interstate 35

dis_m_intl_airpt: Distance in meters from parcel to the Austin City Center

dis_m_NSC: Distance in meters from parcel to North-South Connector Highway

dis_m_Mopac: Distance in meters from parcel to Mopac Freeway

```
# Q6

# 6.1
library(plotrix)
require(plotrix)
distanceToPlaces <- list(
  projDataCleansed5$dis_m_city_center,
  projDataCleansed5$dis_m_EWC,
  projDataCleansed5$dis_m_high130,
  projDataCleansed5$dis_m_inter35,
  projDataCleansed5$dis_m_intl_airpt,
  projDataCleansed5$dis_m_NSC,
  projDataCleansed5$dis_m_Mopac)

colorsForMultiHist7 = c(
  "red",
  "dark orange",
  "gold",
  "green",
  "light blue",
  "dark blue",
  "purple")

namesForThese7 = c(
  "City Center",
  "East-West Connector Highway",
  "Highway 130",
  "Interstate 35",
  "International Airport",
  "North-South Connector Highway",
  "Mopac Freeway")

par(mar=c(5.1, 4.1, 4.1, 8.1), xpd=TRUE)

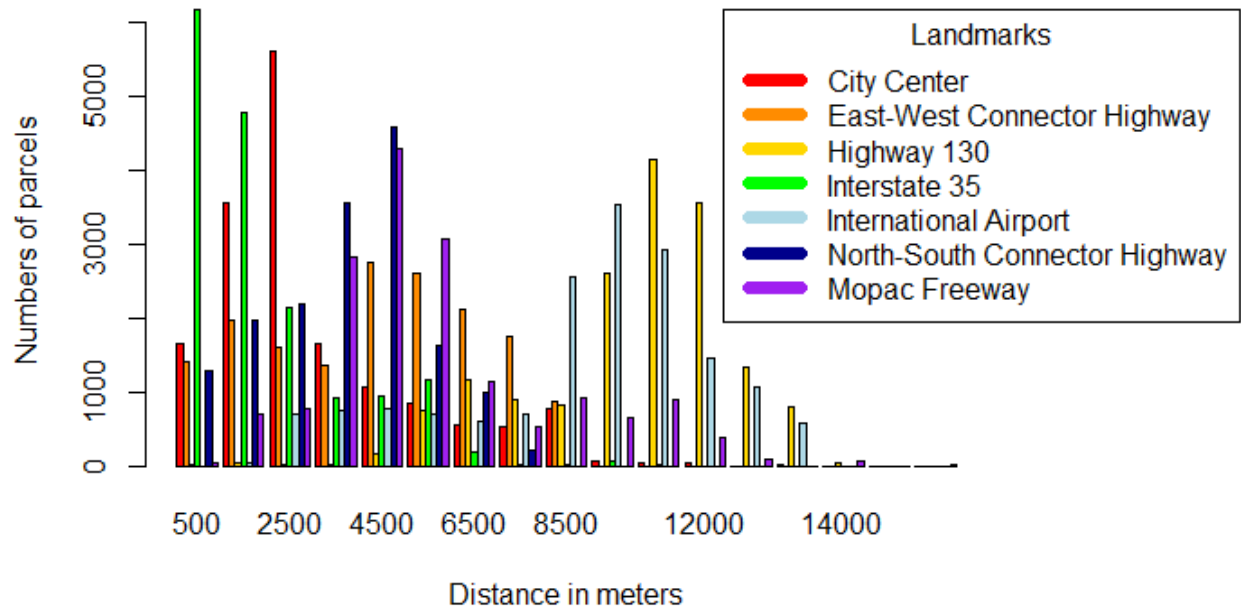
multihist(distanceToPlaces, col= colorsForMultiHist7,
  main = "Distances to Landmarks in histograms",
  xlab = "Distance in meters",
  ylab = "Numbers of parcels",
  color.legend = colorsForMultiHist7)
legend("topright", namesForThese7,
  col = colorsForMultiHist7, lwd = 7,
  title = "Landmarks",
  inset=c(-0.3,0))

par(mar=c(5.1, 4.1, 4.1, 13.1), xpd=TRUE)

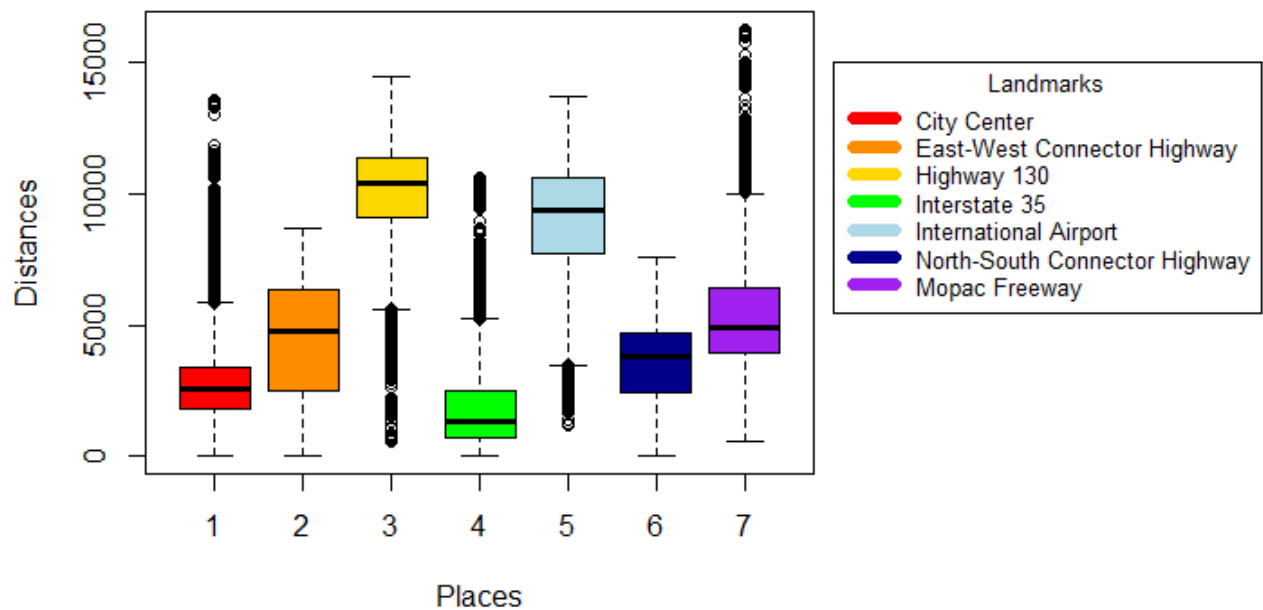
boxplot(distanceToPlaces,
  main="Distance to Landmarks in boxplots",
  xlab = "Places", ylab = "Distances",
  col=colorsForMultiHist7,
  border="black"
)

legend(8, 15000,
  namesForThese7,
  title = "Landmarks",
  col = colorsForMultiHist7,
  cex = 0.8,
  lwd = 7, lty = 1)
```

Distances to Landmarks in histograms



Distance to Landmarks in boxplots



(2) Opportunity indexes (Pie Charts) (categorical, qualitative):

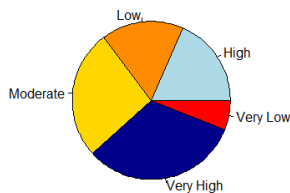
The 4 opportunity indexes which denote individuals' self-perceived odds of different qualities of life can be fitted into pie charts:

```
library(MASS)

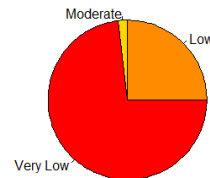
redIsBadBlueIsGood <- c("Light Blue", "Dark Orange", "Gold", "Dark Blue", "Red")
redIsBadBlueIsGood2 <- c("Dark Orange", "Gold", "Red", "Light Blue", "Dark Blue")

pie(table(projDataCleansed5$idx_econ_opp), main = "Pie Chart for Opportunity Indexes: Economic",
    col = redIsBadBlueIsGood)
pie(table(projDataCleansed5$idx_housing_opp), main = "Pie Chart for Opportunity Indexes: Housing",
    col = redIsBadBlueIsGood2)
pie(table(projDataCleansed5$idx_ed_opp), main = "Pie Chart for Opportunity Indexes: Education",
    col = redIsBadBlueIsGood)
pie(table(projDataCleansed5$idx_comp_opp), main = "Pie Chart for Opportunity Indexes: Comprhensive",
    col = redIsBadBlueIsGood)
```

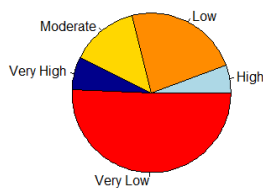
Pie Chart for Opportunity Indexes: Economic



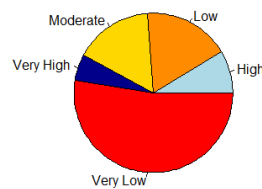
Pie Chart for Opportunity Indexes: Housing



Pie Chart for Opportunity Indexes: Education



Pie Chart for Opportunity Indexes: Comprhensive



Because opportunity indexes are recorded in seemingly ordinal but non-numerical manners, pie charts allow them to be inspected quite intuitively.

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

As demonstrated above, histograms are effective in demonstrating patterns in numerical distributions, while boxplots are useful finding outliers and other fundamental statistics including max values, min values, medians, 1st quantiles, 3rd quantiles. On the other hand, when

a vector of data is categorical where said statistics are not relevant, pie charts make the distributions easy to understand by clearly demonstrating the ratio of each category, as observed in the ones showing opportunity indexes.

In addition, though not yet plotted in this section, scatterplots and regression lines are useful in expressing correlation between two or multiple columns. These will be explored further in the section regarding relationships (Q7).

6.3 Describe what you find from the graphs.

(1) It appears in the histogram that most parcels are rather close to Interstate 35 and far away from Highway 130: the distribution of these two columns are in quite opposite trends with some outliers. The former is right skewed while the latter is left skewed.

The accompanying boxplot shows that distances between these 16,490 parcels to Interstate 35 indeed have lower mean and quantiles (Q1, Q3) than distances to other landmarks, and its outliers are on the upper end of the distribution; this is the opposite for distances to Highway 130. In short, this boxplot complements the information which can be discerned from the histogram.

(2) The max values, min values, medians, 1st quantiles, 3rd quantiles are quite similar for distances to East-West Connector Highway and North-South Connector Highway. In addition, these two columns do not have outliers. It can be inferred that the majority of parcels are (close to) equally distanced from both connectors.

(3) The pie charts show that, except for economic stability, most individuals on these parcels have very low opportunities to achieve any of the following: affordable housing, high level of education, and overall wellness expressed by these indexes. This once again adds to the general impression that, while people in big cities have higher and more stable income than residents elsewhere, most of them do not have satisfactory lives.

Q7 Now look at the relationships among several variables.

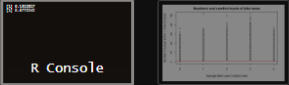
7.1 For example, look at the original “conf” and “bike_lanes” columns. They are both indicators of ease of bicycle transportation, but each column conveys different information. What different information and what similar information can you get from these variables? How are the two variables related?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

To start off, the correlation and covariance of these two columns have similar takeaways. The range for correlation coefficients is between -1 and +1, while that for covariance is between $-\infty$ and $+\infty$. Knowing this, the following results can be interpreted: there may be a weak positive correlation between the number of bike lanes within 1 mile and the average comfort level of bike lanes. The degree of this weakness is further illustrated by the scatterplot:

```
# 7.1
cor(projDataCleansed5$lv1_bike_comf, projDataCleansed5$num_1mi_bike_ln)
cov(projDataCleansed5$lv1_bike_comf, projDataCleansed5$num_1mi_bike_ln)

plot(projDataCleansed5$lv1_bike_comf, projDataCleansed5$num_1mi_bike_ln,
     main = "Numbers and comfort levels of bike lanes",
     xlab = "Average Bike Lane Comfort Level",
     ylab = "Number of Lanes within 1 mile of parcel")

abline(lm(projDataCleansed5$lv1_bike_comf ~ projDataCleansed5$num_1mi_bike_ln), col = "Red")
...

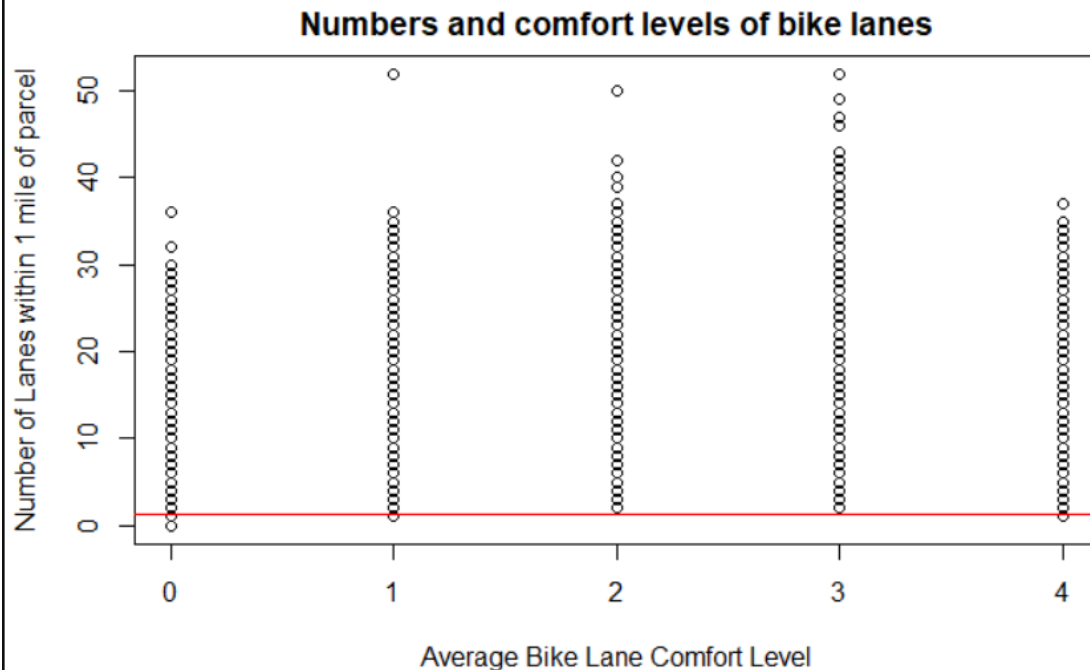

R Console
[1] 0.1512817
[1] 0.8772242
```

```
Call:
lm(formula = lv1_bike_comf ~ num_1mi_bike_ln, data = projDataCleansed5)

Residuals:
    Min       1Q   Median       3Q      Max
-2.0541 -0.6304 -0.3406  0.4588  2.7263

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.251375   0.018783   66.62  <2e-16 ***
num_1mi_bike_ln 0.022297   0.001135   19.65  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9138 on 16488 degrees of freedom
Multiple R-squared:  0.02289,    Adjusted R-squared:  0.02283
F-statistic: 386.2 on 1 and 16488 DF,  p-value: < 2.2e-16
```

As observed, though there appears to be some more points denoting high numbers of bike lanes at the comfort level of 3, there does not for level 4; the red regression line is also almost flat, indicating that this relation is weak, but existent.

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).

(1) Numbers of existing urban trails and median rent per zip (Scatterplot):

```
cor(projDataCleansed5$num_1mi_ex_trail, projDataCleansed5$med_rent_perzip)
reg1 <- lm(num_1mi_ex_trail ~ med_rent_perzip, data = projDataCleansed5)
coeff = coefficients(reg1)
summary(reg1)

plot(projDataCleansed5$num_1mi_ex_trail, projDataCleansed5$med_rent_perzip,
     main = "Numbers of Urban Trails and Median Rent Per Zip",
     xlab = "Numbers of Urban Trails",
     ylab = "Median Rent Per Zip")
abline(reg1, col="blue")
```

```
[1] 0.3340881
```

```
Call:
```

```
lm(formula = num_1mi_ex_trail ~ med_rent_perzip, data = projDataCleansed5)
```

```
Residuals:
```

```
    Min       1Q   Median       3Q      Max  
-4.952 -1.617 -1.119   2.071  16.941
```

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -2.4443596   0.0913955  -26.75  <2e-16 ***  
med_rent_perzip  0.0046520   0.0001022   45.51  <2e-16 ***
```

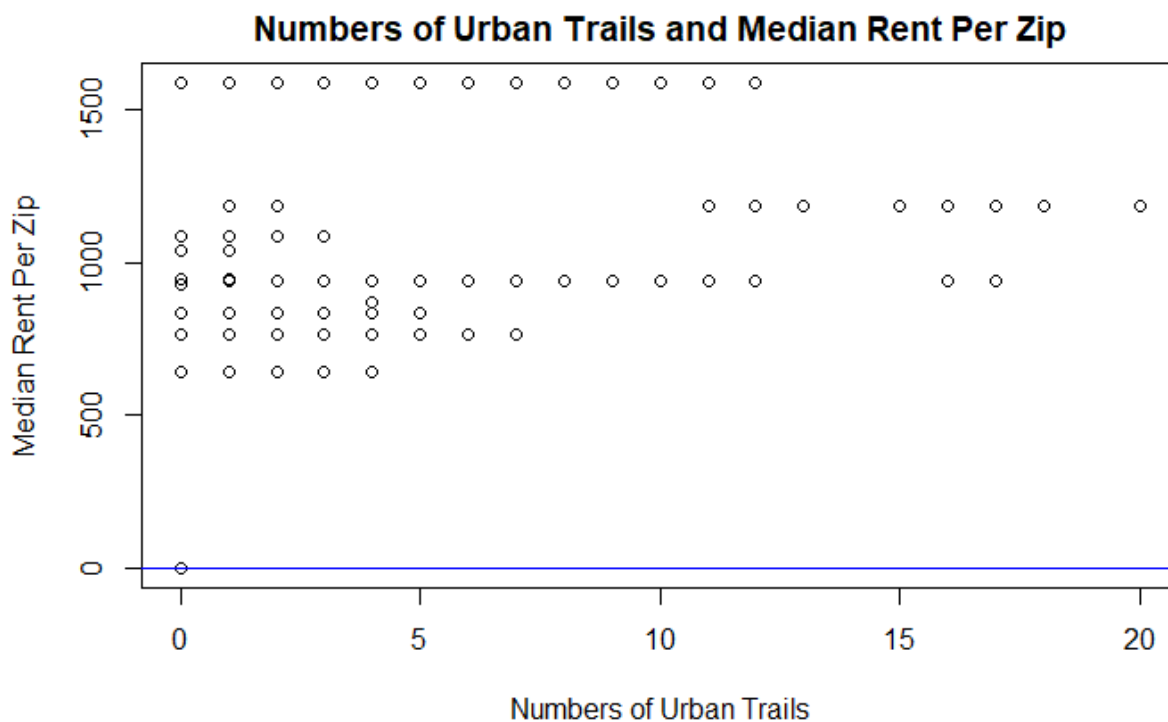
```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.178 on 16488 degrees of freedom
```

```
Multiple R-squared:  0.1116,    Adjusted R-squared:  0.1116
```

```
F-statistic: 2072 on 1 and 16488 DF,  p-value: < 2.2e-16
```



There is a weak positive relationship between number of existing urban trails and median rent per zip. That is, the more trails exist nearby, the higher the median rent is for a parcel.

(2) Distances to Highway 130 and Interstate 35(Scatterplot):

```
cor(projDataCleansed5$dis_m_high130, projDataCleansed5$dis_m_inter35)
reg2 <- lm(dis_m_high130 ~ dis_m_inter35, data = projDataCleansed5)
coeff2 = coefficients(reg2)
summary(reg2)

plot(projDataCleansed5$dis_m_high130, projDataCleansed5$dis_m_inter35,
     main = "Distances to Highway 130 and Interstate 35",
     xlab = "Distance to Highway 130",
     ylab = "Distance to Interstate 35")
abline(reg2, col = "magenta")
```

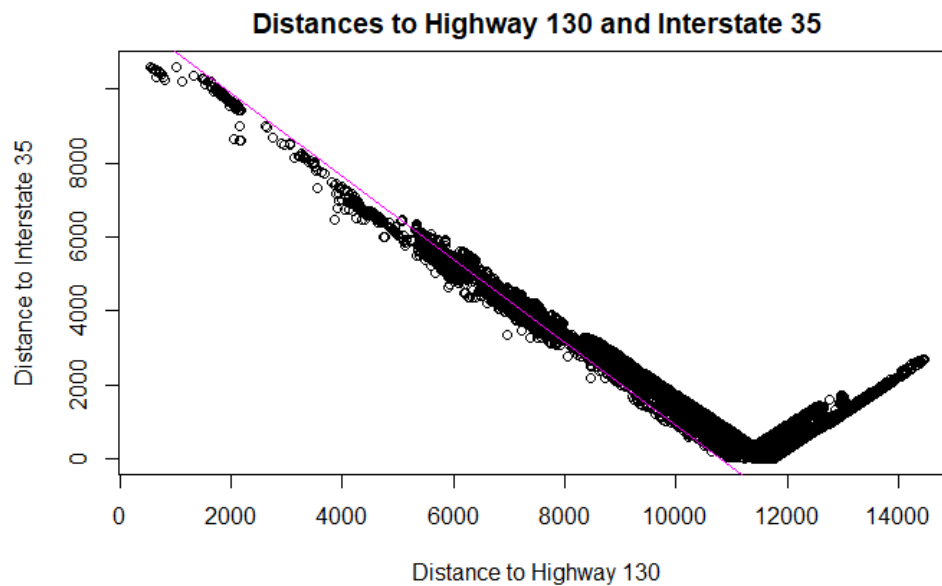
```
[1] -0.8975906

Call:
lm(formula = dis_m_high130 ~ dis_m_inter35, data = projDataCleansed5)

Residuals:
    Min       1Q   Median       3Q      Max
-1369.7  -433.8  -269.8   -99.2   5374.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.212e+04  1.115e+01  1087.5  <2e-16 ***
dis_m_inter35 -1.123e+00  4.293e-03  -261.5  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 962.9 on 16488 degrees of freedom
Multiple R-squared:  0.8057,    Adjusted R-squared:  0.8057
F-statistic: 6.836e+04 on 1 and 16488 DF,  p-value: < 2.2e-16
```



The is a very strong negative relationship between distances to Interstate 35 and Highway 130; that is, the closer to either a parcel in Austin is, the more far away it is from another.

(3) Median household income per zip and comprehensive opportunity indexes

```
cor(projDataCleansed5$dis_m_NSC, projDataCleansed5$med_home_price_perzip)
reg3 <- lm(dis_m_NSC ~ med_home_price_perzip, data = projDataCleansed5)
coeff3 = coefficients(reg3)
summary(reg3)
plot(projDataCleansed5$dis_m_NSC, projDataCleansed5$med_home_price_perzip,
      main = "Distances to North-South Connector Highway and Median Home Price per ZIP",
      xlab = "Distance to North-South Connector Highway",
      ylab = "Median Home Price per ZIP")
abline(reg3, col = "green")
```

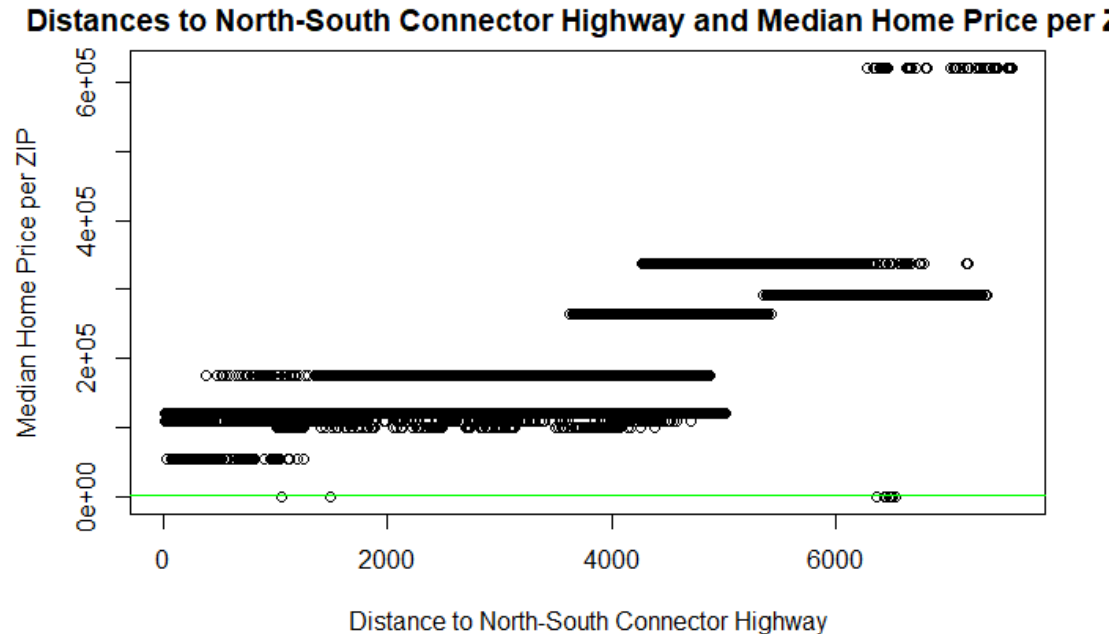
```
[1] 0.7538556

Call:
lm(formula = dis_m_NSC ~ med_home_price_perzip, data = projDataCleansed5)

Residuals:
    Min       1Q   Median       3Q      Max
-3807.0  -861.6   -36.9    807.8   5965.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.694e+02  2.215e+01   25.7  <2e-16 ***
med_home_price_perzip  1.532e-02  1.040e-04  147.3  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1080 on 16488 degrees of freedom
Multiple R-squared:  0.5683,    Adjusted R-squared:  0.5683
F-statistic: 2.171e+04 on 1 and 16488 DF,  p-value: < 2.2e-16
```



There is a strong positive relationship between the distance to North-South Connector Highway and median Home price per zip; that is, the more far away from the NS Connector Highway a parcel, the more expensive homes on that parcel generally are.

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?

```
# 8.1  
projDataCleansed5$descr_const_nearby <- tolower(projDataCleansed5$descr_const_nearby)  
View(projDataCleansed5)
```

descr_const_nearby
remodel to add 2 exterior doors to existing religious assem...
remodel to add 2 exterior doors to existing religious assem...
remodel to add 2 exterior doors to existing religious assem...
tenant finish out to create retail
tenant finish out to create retail
tenant finish out to create retail
tenant finish out to create retail
new leasing office an accessory to garage
new leasing office an accessory to garage

Given that this is a purely descriptive column, converting all characters to lowercase should eliminate errors that could arise from failing to discern “i” and “l” ---- the capitalization of the former could be easily confused with the latter. Considering how all words are passed over, there should be no information loss in this process.

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, 7 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

The stop words given above are stripped from the column using the command below. After stripping is complete, the first 10 values of the remaining values are shown:

```
# 8.2
library(tm)
stopwordsForNow = c("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", "7 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")

textMined1 <- removeWords(projDataCleansed5$descr_const_nearby, stopwordsForNow)
TMResult1 <- sort(table(textMined1), decreasing = TRUE)[1:10]
TMResult1
```

```
textMined1

new solar installation new residence
623
existing wireless telecommunication tower adding equipment
256
service meter food truckrefer 969712j install subpanel outlet food truck run existing
230
interior renovations through out school campus bringing stage access ramp adttas compliant adttas
complication restroom locker room baskets
222
restrooms ada replace park lighting interior remodel existing
221
existing multi family units 101103104 replace tub
180
200a service athletic field lighting installation new
170
city standards crtask order 24bon03500 construct 1500 lf new sidewalk 1500 lf cg shall meet
169
reroof existing commercial building nonstructural
167
residential admin office interior remodel change use
166
```

The most frequent value is “new solar installation new residence” which has 623 entries, followed by “adding equipment existing wireless telecommunication tower” which has 256. The 10th most frequent is “change use residential admin office interior remodel” which has 166 occurrences.

8.3 Perform 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?

The following commands strips the stop words from the description column as previous but this time displays the top 10 non stop words in the remaining list.

Specifically, because the method used produces empty spaces as the most frequent occurrences, a vector position of 2 to 11 was specified instead of 1 to 10 to obtain the relevant result (empty spaces cannot really be considered word, after all)

```
## {r}
# 8.3
library(tm)
textMined2 <- removeWords(projDataCleansed5$descr_const_nearby, stopwordsForNow)
textMined2 <- strsplit(textMined2, " ", fixed = T)
words2 <- unlist(textMined2)
TMResult2 <- sort(table(words2), decreasing = TRUE)[2:11]
TMResult2
```

words2	new	lf interior	remodel sidewalk	office	meet	city	shall
existing 5711	5605	2975	2812	2590	2272	1925	1837
						1830	1828

As shown here, the most frequent non stop word is “existing” which has 5711 entries, and the 9th most frequent word is “city” with 1830 entries, etc.

The several frequent words that could give useful info are “new”, “remodel”, “office” and “city”. These could indicate new constructions that involve subjects of interest to an owner of tech office; it could be a newer site which often means sturdier infrastructure, which means less long term maintenance expense for whoever owns the unit. Suffice to say, more information is required in all cases to discern the actual meaning of such words.

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?

(1) List from 8.2:

To start off, “existing wireless telecommunication tower” which has 256 entries (2nd highest in the list) could entice tech worker as it indicates good and functional wireless signal; “food truck run existing service” which has 230 could also be attractive because tech workers usually require supper that are fast and easily accessible; on the contrary, the influence of words such

as “school campus” (222 entries) and “multi family units” (180 entries) will heavily depend on the composition of a corporation. The general stereotype is that tech workers do not have families or children so school and multi-family units are likely to be less relevant to them, though it may be quite the opposite for those in managerial positions.

(2) List from 8.3:

The word “sidewalk” could be of interest to any tech worker who are commuters as the presence of sidewalks often dictates where s/he can or cannot go through. Though it is of lower frequency, “city” in this list could mean proximity to any place in the city, which means convenience. Though such may indicate higher prices for property owners or managers, a tech worker would appreciate such qualities of life.

In sum, words at various frequency which indicate fast access to specific necessities of live could become attractions for tech workers. This is true even for those words that are low frequency words. After all, the dataset was collected without such interests (or any interests) so the frequency of words only represents the events happening, not how intriguing such words are to individual workers.

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.

An additional list of words which would be collected from surveys conducted to ask workers what resources they like should be created. Steps like the ones above should then be used to find the most frequent occurrences of words specified in the list instead of filtering them out.

Once the most frequent occurrences have been found, the rows that contain such words should be listed out and marked as sites of interest.

The reason why this additional step would help is because selecting the most frequent non stop words only produce data that represent construction events, not how much such events are of interests to workers or GlobalTechSync managers (who will decide which site to situate the company). Considering how

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.

The `Bus_area` column has been renamed as `bool_bus_sys`. As requested, the commands below filters the 14 parcels which do not have bus service:

```
{r}
# 9.1

projDataFiltered1 <- projDataCleansed5[which(projDataCleansed5$bool_bus_sys != 0),]
dim(projDataCleansed5)
dim(projDataFiltered1)
head(projDataFiltered1)
```

num_1mi_pp_trail <int>	lvl_bike_comf <int>	num_1mi_bike_ln <int>	bool_bus_sys <int>	area_total_bldgs <dbl>
5	2	9	1	136.655
6	2	18	1	295.263
6	2	13	1	137.778
13	1	28	1	3630.600
15	4	32	1	353.614
12	3	23	1	0.000

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

The column in question is `Shape_Area` has been renamed as `area_parcel_m2`. The following commands filters the parcels as instructed. Only 1 parcel was removed:

```
##{r}
# 9.2
projDataFiltered2 <- projDataFiltered1[which(projDataFiltered1$area_parcel_m2 >= 300),]
dim(projDataFiltered1)
dim(projDataFiltered2)
head(projDataFiltered2)
```

dis_m_intl_airpt	dist_number	area_parcel_m2	code_zoning	code_zip	code_land_use_inv
12720.8	14	7716.545	NP	78705	100
12793.4	14	18296.797	NP	78705	100
12714.6	14	5604.736	MF-6-CO-NP	78705	100
10330.8	14	3006.199	TOD	78702	400
10443.8	14	5563.109	TOD	78702	300
10368.2	14	6307.889	TOD	78702	300

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

The column `zoning_o_3` has been renamed as `code_zoning`. Any parcel with codes included in the table is removed with the subsequent commands:

```
##{r}
# 9.3
residentialLands = c("LA", "RR", "SF-1", "SF-2", "SF-3", "SF-4A", "SF-4B", "SF-5", "SF-6", "MF-1", "MF-2", "MF-3", "MF-4", "MF-5", "MF-6", "MH")
projDataFiltered3 = subset(projDataFiltered2, !(code_zoning %in% residentialLands))
dim(projDataFiltered2)
dim(projDataFiltered3)
head(projDataFiltered3)
```

```
[1] 16475 39
[1] 16445 39
```

dis_m_intl_airpt	dist_number	area_parcel_m2	code_zoning	code_zip	code_land_use_inv
12720.8	14	7716.545	NP	78705	100
12793.4	14	18296.797	NP	78705	100
12714.6	14	5604.736	MF-6-CO-NP	78705	100
10330.8	14	3006.199	TOD	78702	400
10443.8	14	5563.109	TOD	78702	300
10368.2	14	6307.889	TOD	78702	300

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

As shown by the end of 9.3, the new dataset dimensions are (16445, 39). That is, there are 16445 parcels which are not within residentials, have areas greater than 300 square meters, and also bus services, each with up to 39 attributes being recorded.

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.

A custom R function is created to represent each of the following preferences as equally as manageable within the language's known capability:

1. An undeveloped site is preferred.

A parcel gains 1.0 score if its land use inventory code is "900", which means "Undeveloped"s
900

Undeveloped - Parcels without structures that have the potential for development

2. Ease of access to a major interstate or highway is preferred.

A parcel gains 0.5 score if the summation of its distances to the following 5 places is less than the mean summation of distances to those: East-West Connector Highway, North-South Connector Highway, Mopac Freeway, Highway 130 and Interstate 35; it gains another 0.5 score if its distances is less than the 25th quantile (Q1) of the mean summation.

3. Easy access to the site by bike or foot is preferred.

A parcel gains 0.5 score if its number of bike lanes within 1 mile is equal or greater than the median number of bike lanes; it gains another 0.5 score if its number is greater than the 75th quantile (Q3) of overall numbers.

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.

A parcel gains 0.5 score if its area occupied by buildings is less than the mean area occupied by buildings, as such indicates that it has more open and empty spaces; it gains another 0.5 if its area occupied is less than the 25th quantile (Q1).

5. Higher tax breaks or discounts at both the district and block levels is preferred.

A parcel gains 1.0 score if **both** its tax break percentages at district and block levels are greater than the mean percentages.

6. High education opportunity in the area and strong nearby university systems are preferred.

A parcel gains 0.5 score if index of education opportunity is "High" or "Very High". It also gains 0.5 if its land use inventory code is "640", which means "Educational".

640

Educational - Day care, primary and secondary education, colleges, universities, business trade schools

7. Ability for tech workers to own their own houses is preferred.

A parcel gains 1.0 score if its index of housing opportunity is “Very High”, and 0.2 less for every level below (“High” = 0.8, “Moderate” = 0.6, etc.).

8. Fast reliable internet needs to be easily accessible at the site.

A parcel gains 1.0 if its description of construction nearby mentions anything similar to “existing wireless telecommunication”.

9. Nearby active construction of office type structures is preferred.

A parcel gains 1.0 if its description of construction nearby mentions anything including “office”.

With the preferences listed above, the analyst ensures that each is represented as equally as possible while preferences that could be satisfied by multiple conditions have their scores divided so that the difficulty of satisfying that preference is kept on a similar level as those with only singular requirements.

The system will become more comprehensive with the implementation in the following subquestion.

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

With each of the preferences represented, the implementation of R code is as follow:

```
library(data.table)
rankParcelsWithPref <- function(x, y){
  #Function: Give scores to each parcel in the dataset according to how much they match
  #preference specified by the client.
  # Input 1: x; a dataframe conforming to the standards of Austin parcel record database
  # Input 2: y; an integer denoting how many records should be shown after ranking
  # Output: A list of parcels which is a subset of the input dataframe

  # 1. An undeveloped site is preferred
  x$rank_pref1 <- x$code_land_use_inv == 900

  # 2. Ease of access to a major interstate or highway is preferred
  x$rank_pref2A <- x$dis_m_EWC + x$dis_m_NSC + x$dis_m_Mopac + x$dis_m_high130 +
  x$dis_m_inter35 < mean(x$dis_m_EWC + x$dis_m_NSC + x$dis_m_Mopac + x$dis_m_high130 +
  x$dis_m_inter35)

  x$rank_pref2B <- x$dis_m_EWC + x$dis_m_NSC + x$dis_m_Mopac + x$dis_m_high130 +
  x$dis_m_inter35 < quantile(x$dis_m_EWC + x$dis_m_NSC + x$dis_m_Mopac + x$dis_m_high130 +
  x$dis_m_inter35, 0.25)

  # 3. Easy access to the site by bike or foot is preferred
  # Improvement required: No existing variable can represent convenience on foot
  x$rank_pref3A <- x$num_1mi_bike_ln >= median(x$num_1mi_bike_ln)
  x$rank_pref3B <- x$num_1mi_bike_ln >= quantile(x$num_1mi_bike_ln, 0.75)

  # 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
  # Improvement required: Only open spaces are represented thus far. No known variable can
  # convey green spaces and areas for enrichment.
  x$rank_pref4A <- x$area_total_bldgs < mean(x$area_total_bldgs)
  x$rank_pref4B <- x$area_total_bldgs < quantile(x$area_total_bldgs, 0.25)
```

IS 457 FA19
FINAL REPORT
CLASS ID: 104

```
# 5. Higher tax breaks or discounts at both the district and block levels is preferred
x$rank_pref5 <- x$pct_tx_brk_block > mean(x$pct_tx_brk_block) & x$pct_tx_brk_dist >
mean(x$pct_tx_brk_dist)

# 6. High education opportunity in the area and strong nearby university systems are
preferred
x$rank_pref6A <- x$idx_ed_opp == "High" | x$idx_ed_opp == "Very High"
x$rank_pref6B <- x$code_land_use_inv == 640

# 7. Ability for tech workers to own their own houses is preferred
# Improvement required: Variables only convey housing opportunities
x$rank_pref7A <- x$idx_housing_opp == "Very High"
x$rank_pref7B <- x$idx_housing_opp == "High"
x$rank_pref7C <- x$idx_housing_opp == "Moderate"
x$rank_pref7D <- x$idx_housing_opp == "Low"
x$rank_pref7E <- x$idx_housing_opp == "Very Low"

# 8. Fast reliable internet needs to be easily accessible at the site
# Needs improvement. Description of nearby construction should not be the only way to
represent easily accessible internet. Would not make sense.
x$rank_pref8 <- x$descr_const_nearby %like% "existing wireless telecommunication"
# 9. Nearby active construction of office type structures is preferred
x$rank_pref9 <- x$descr_const_nearby %like% "office"

#Summation of the scores
x$rank_score <- as.integer(as.logical(x$rank_pref1)) +
  as.integer(as.logical(x$rank_pref2A)) * 0.5 +
  as.integer(as.logical(x$rank_pref2B)) * 0.5 +
  as.integer(as.logical(x$rank_pref3A)) * 0.5 +
  as.integer(as.logical(x$rank_pref3B)) * 0.5 +
  as.integer(as.logical(x$rank_pref4A)) * 0.5 +
  as.integer(as.logical(x$rank_pref4B)) * 0.5 +
  as.integer(as.logical(x$rank_pref5)) +
  ((as.integer(as.logical(x$rank_pref6A)) +
    as.integer(as.logical(x$rank_pref6B))) / 2) +
  as.integer(as.logical(x$rank_pref7A)) * 1 +
  as.integer(as.logical(x$rank_pref7B)) * 0.8 +
  as.integer(as.logical(x$rank_pref7C)) * 0.6 +
  as.integer(as.logical(x$rank_pref7D)) * 0.4 +
  as.integer(as.logical(x$rank_pref7E)) * 0.2 +
  as.integer(as.logical(x$rank_pref8)) +
  as.integer(as.logical(x$rank_pref9))

rankedParcels = head(x[order(x$rank_score, decreasing=TRUE), ], y)
return(rankedParcels)
}
```

With the function in place, it will be called with the parameters put in as follow:

	FID <int>	id_block <fctr>	id_austin_DB <int>	land_type <fctr>	id_lot <fctr>	date_modified <fctr>	id_sec <int>
6011	6008	A	1935294	LOT	2	6/26/2015 0:00	67497
6008	6005	A	1969749	LOT	1	6/26/2015 0:00	66305
15824	15821	E	1912663	LOT	TRACT B	NA	65467
21726	21723	A	400457988	LOT	5	12/9/2013 0:00	208167
21742	21739	A	1670885	LOT	1	9/19/2012 0:00	202909
3555	3552	NA	1754711	PARCEL	3	6/26/2015 0:00	348990
3563	3560	NA	1637168	LOT	1	NA	166646
5430	5427	B	1782273	LOT	1	9/19/2012 0:00	66289
8046	8043	B	1707826	LOT	2	9/19/2012 0:00	61001
11370	11367	NA	1782948	LOT	1-A	12/9/2013 0:00	362439

1-10 of 10 rows | 1-8 of 57 columns

The FID of the top 10 parcels with the highest ranking scores according to the system specifications above is highlighted by the light green rectangle.

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?

It was indeed hard to formulate a function that can represent all the preferences given by the clients, especially when not every preference can be quantified by existing columns. Specifically, easy access by bike or on foot in a real-world sense could not just be defined by the number of bike lanes within 1 mile, but it was the closet variable that could come close to describing it. Ability for tech workers to own house was another difficulty one as the definition for said ability was abstract. Using percentage of affordable homes and housing opportunity index was more an improvisation.

If the dataset were to be updated with newer columns, these should be marked as most needing attention and improvement.

In terms of result, the analyst could discern the top 5 parcels (FID 6008, 6005, 15821, 21723, and 21739) with confidence because their scores of 6.4 and 5.9 were distinctively high. On the contrary, starting from the 6th parcel (FID 3159) there had been too many parcels to choose from, as the number of parcels with scores of 5.4 goes all the way up to 16th.

rank_pref7C <lg>	rank_pref7D <lg>	rank_pref7E <lg>	rank_pref8 <lg>	rank_pref9 <lg>	rank_score <dbl>
FALSE	TRUE	FALSE	TRUE	FALSE	6.4
FALSE	TRUE	FALSE	TRUE	FALSE	5.9
FALSE	TRUE	FALSE	FALSE	TRUE	5.9
FALSE	TRUE	FALSE	TRUE	FALSE	5.9
FALSE	TRUE	FALSE	TRUE	FALSE	5.9
FALSE	TRUE	FALSE	TRUE	FALSE	5.4
FALSE	TRUE	FALSE	TRUE	FALSE	5.4
FALSE	TRUE	FALSE	FALSE	TRUE	5.4
FALSE	TRUE	FALSE	FALSE	FALSE	5.4
FALSE	TRUE	FALSE	TRUE	FALSE	5.4

1-10 of 20 rows | 53-58 of 57 columns

Previous 2 Next

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?

For continuous numerical fields, it was an intuition to give the score of 1.0 (which indicates that a preference is satisfied in whole) to whichever parcel has a preferable value than half the remaining data, hence the comparison with means. Worth mentioning is the comparison with median instead of mean in the terms of number of bike lanes ---- it occurred that such numbers can only be integers, so it would be more sensible to compare an individual with median.

Though much investment of time and effort was involved, the analyst is not happy with the available options, as several variables were not the best representations of “satisfying” the preferences to which they are related. Preference 8 stated “fast reliable internet” which could not be discerned by the simple inclusion of keywords recording constructions nearby alone, but a variable had to be put in place to represent the preference, nonetheless.

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?

Parcel 6008 is leading with a score of 6.4, which means that it satisfies all requirements and relatively the most preferences. Then the 4 parcels (FID 6005, 15821, 21723, and 21739) with score 5.9. While none of them is perfect in the sense that any could satisfy all preferences (which would result in a score of 9.0), Parcel 6008 is the “current best” option that the analyst can recommend according to the ranking system.

It is worth noting that the variables/columns used to represent the latter preferences are not optimal. Ability for workers to own houses should be better represented by more than just housing indexes. Similarly, fast internet and newly furnished offices should be better conveyed than the current workaround which is simply the inclusion of such words in constructions nearby. Access to site on foot is also non-existent in the current dataset.

Overall, if new columns regarding the lacking information above could be added, much better decisions can be made. Parcels after the top 5 can then be further differentiated.

As most preferences ended up being measured using composite columns which are derived from existing columns, evaluation of each will also be recommended before the implementation can be regarded as legitimate.

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.

A total of 5 numeric variables are compared using scatterplots and boxplots, resulting in a total of 10 graphs as follow.

Because the creation of graphs involves similar codes, only the ones used to the first two graphs are shown here. The others are identical in structure and can be further inspected in the R file, also part of the submitted documents.

```
# Enumeration of the Top 10 points
# Tier 1 includes Parcel 6008 alone which has the highest score of 6.4
# Highlighted as Green Diamond
dotsToHighlightFIDTier1 = c(6008)
# Tier 2 includes 4 parcels which have the second highest score of 5.9
# Highlighted as Blue Triangles
dotsToHighlightFIDTier2 = c(6005, 15821, 21723, 21739)
# Tier 3 includes 5 parcels which have the third highest score of 5.4
# Highlighted as Red Dots
dotsToHighlightFIDTier3 = c(3552, 3560, 5427, 8043, 11367)

# All top 10 parcels are referred to as a separate subset in subsequent boxplots
boxplotSubTop10Parcels = c(dotsToHighlightFIDTier1, dotsToHighlightFIDTier2,
dotsToHighlightFIDTier3)
projDataFiltered3$isTop10 <- ifelse(projDataFiltered3$FID %in% boxplotSubTop10Parcels,
"Top 10 Parcels", "Other Parcels")

# Graph 1: Total area of buildings on parcel
plot(projDataFiltered3$FID, projDataFiltered3$area_total_bldgs, col = "grey", cex = 0.5,
      main = "Total area of buildings on parcel",
      xlab = "FID", ylab = "Area occupied by buildings in square meters",
      ylim = c(0, 9999))
points(6008, projDataFiltered3$area_total_bldgs[projDataFiltered3$FID == 6008], col =
"green", pch = 18, cex = 3)

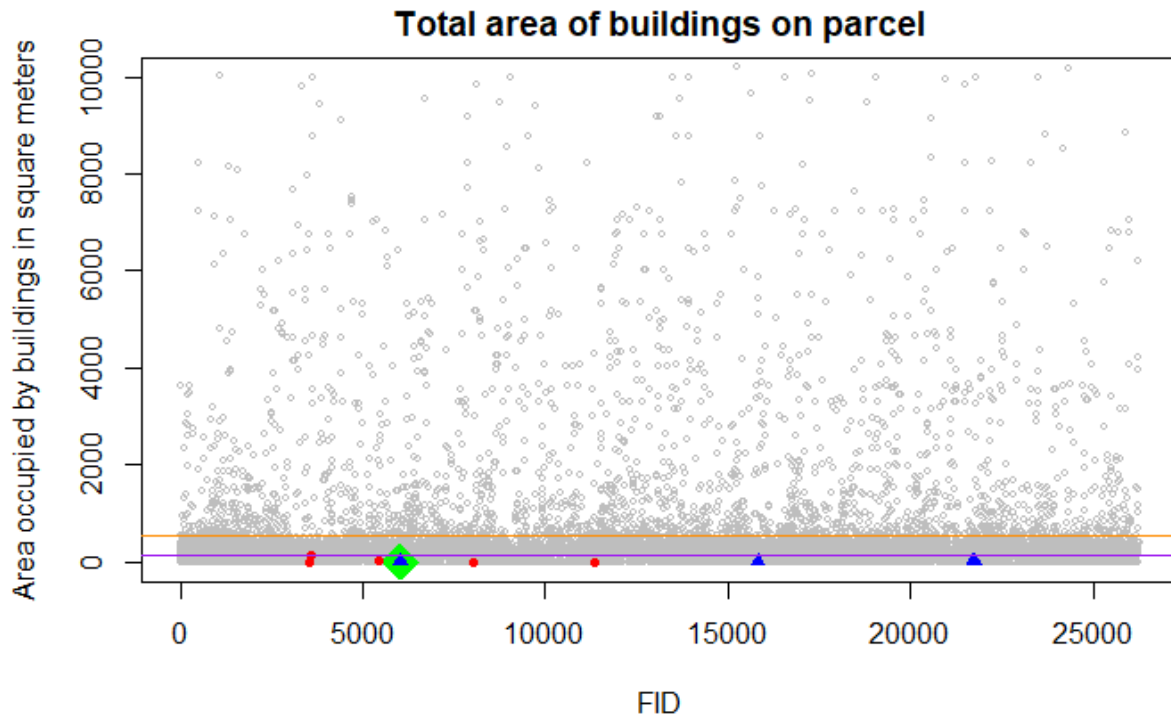
points(dotsToHighlightFIDTier2, projDataFiltered3$area_total_bldgs[projDataFiltered3$FID
%in% dotsToHighlightFIDTier2], col = "blue", pch = 17, cex = 1)

points(dotsToHighlightFIDTier3, projDataFiltered3$area_total_bldgs[projDataFiltered3$FID
%in% dotsToHighlightFIDTier3], col = "red", pch = 16, cex = 0.75)
# Mean denoted by the line in dark orange
abline(h = mean(projDataFiltered3$area_total_bldgs), col = "dark orange")
# Q1 denoted by the line in purple
abline(h = quantile(projDataFiltered3$area_total_bldgs, 0.25), col = "purple")

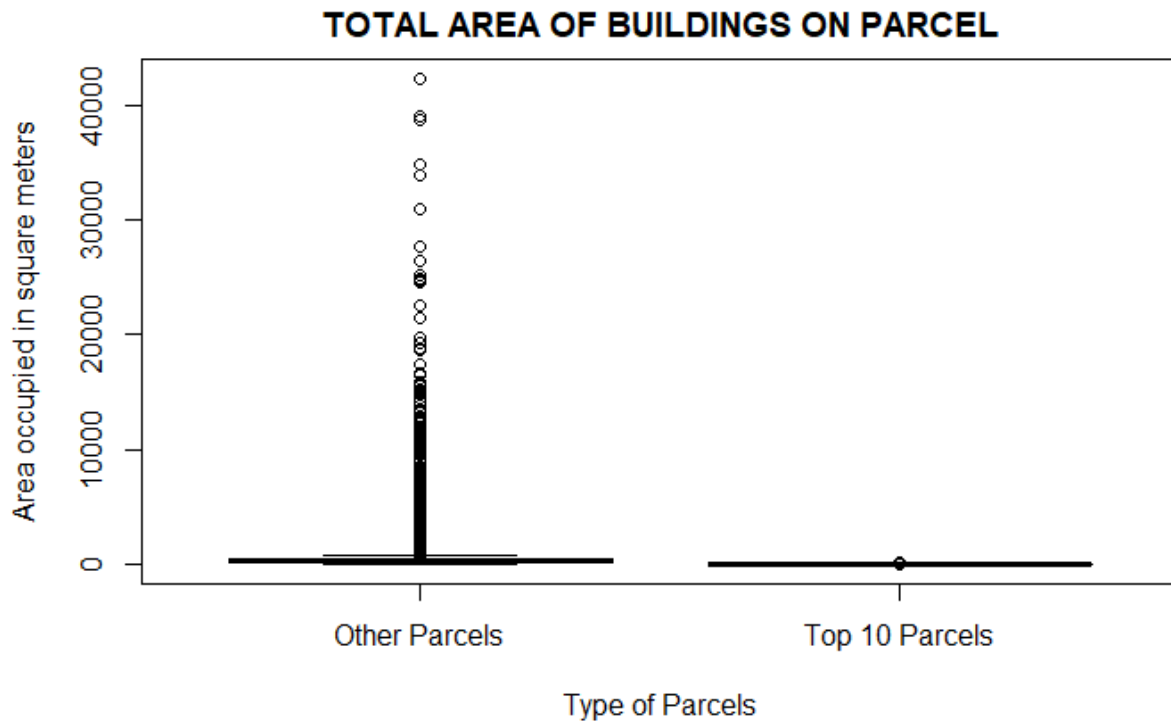
boxplot(area_total_bldgs~isTop10,
      data = projDataFiltered3,
      col = c("skyblue", "green"),
      main = toupper("Total area of buildings on parcel"),
      xlab = "Type of Parcels",
      ylab = "Area occupied in square meters")
```

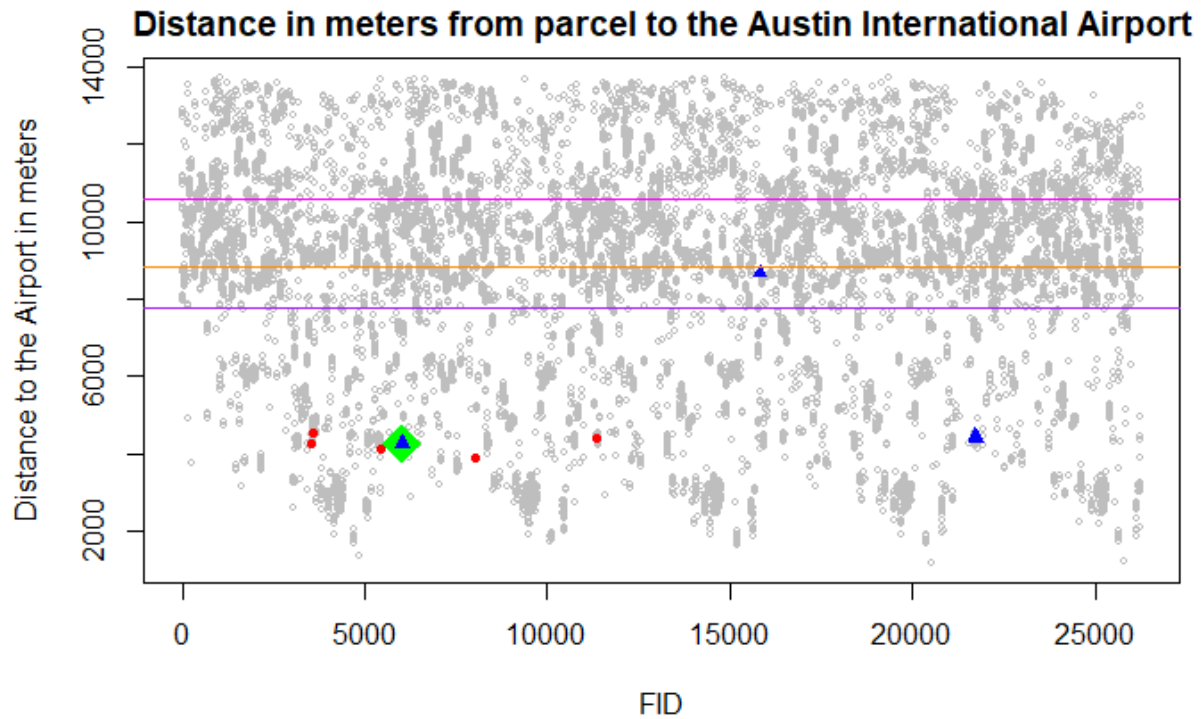
In the following graphs that are scatterplots:

- The Great Green Diamond represents Parcel 6008, the single best parcel.
- Blue Triangles represent Parcels 6005, 15821, 21723, 21739, the 2nd best parcels.
- Red Dots represent Parcels 3552, 3560, 5427, 8043, 11367, the 3rd best parcels.
- Mean/Median is represented by orange lines, Q1 by purple, and Q3 by magenta.

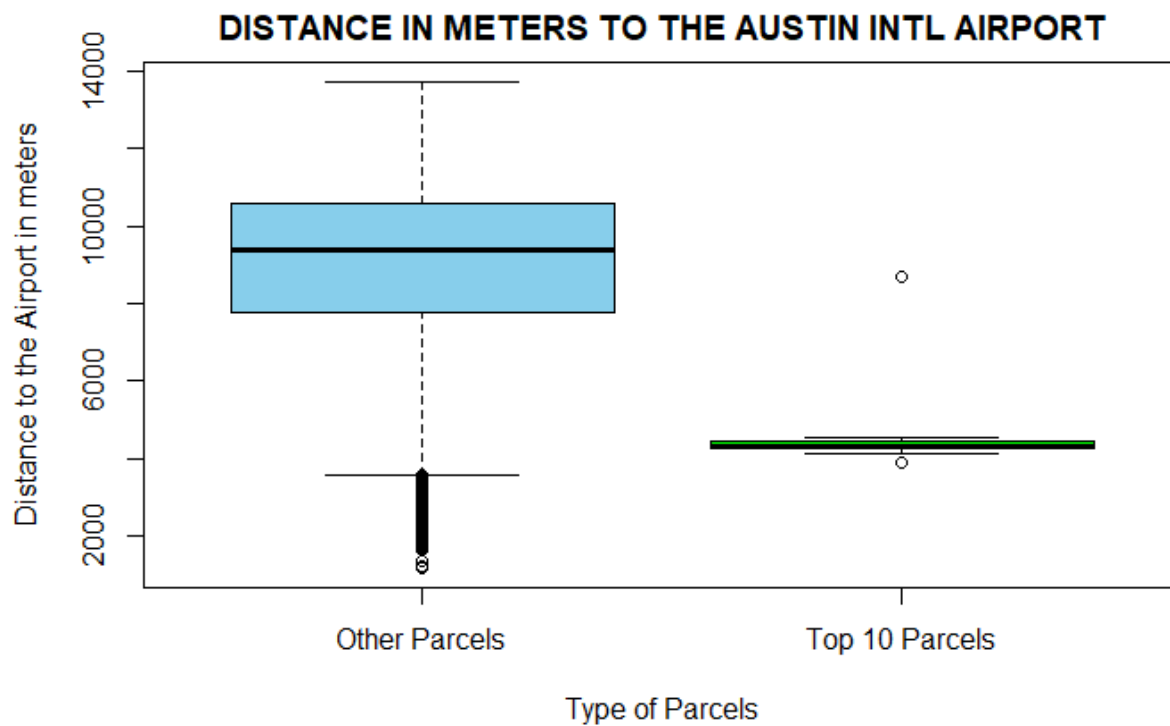


The Top 10 have less total area of buildings on their parcels, even less than the Q1.



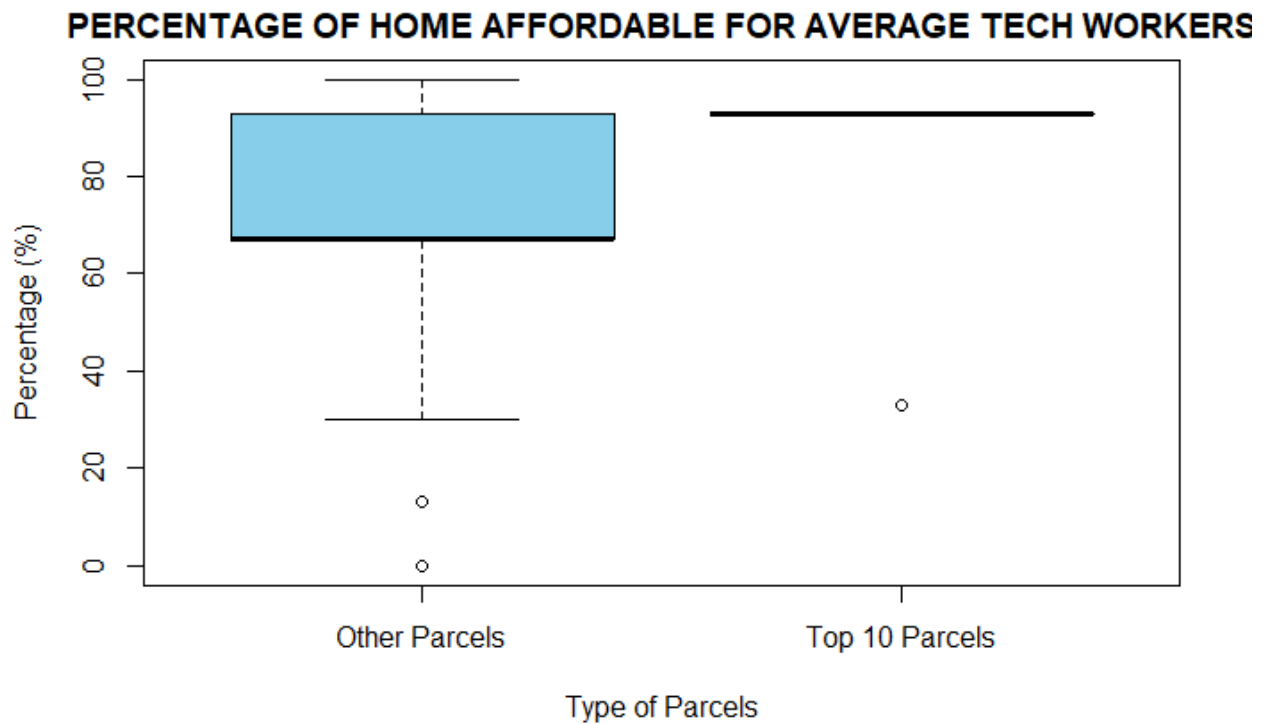


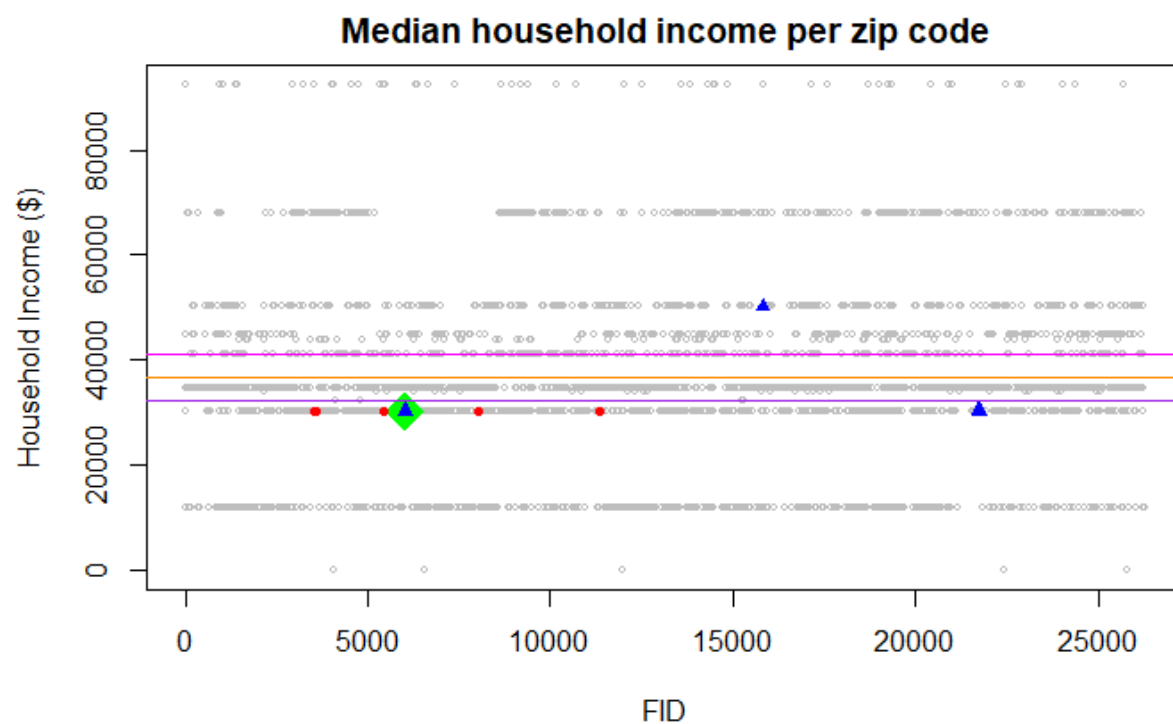
Except Parcel 15821, the top 10 are much closer to the International Airport (closer than Q1).



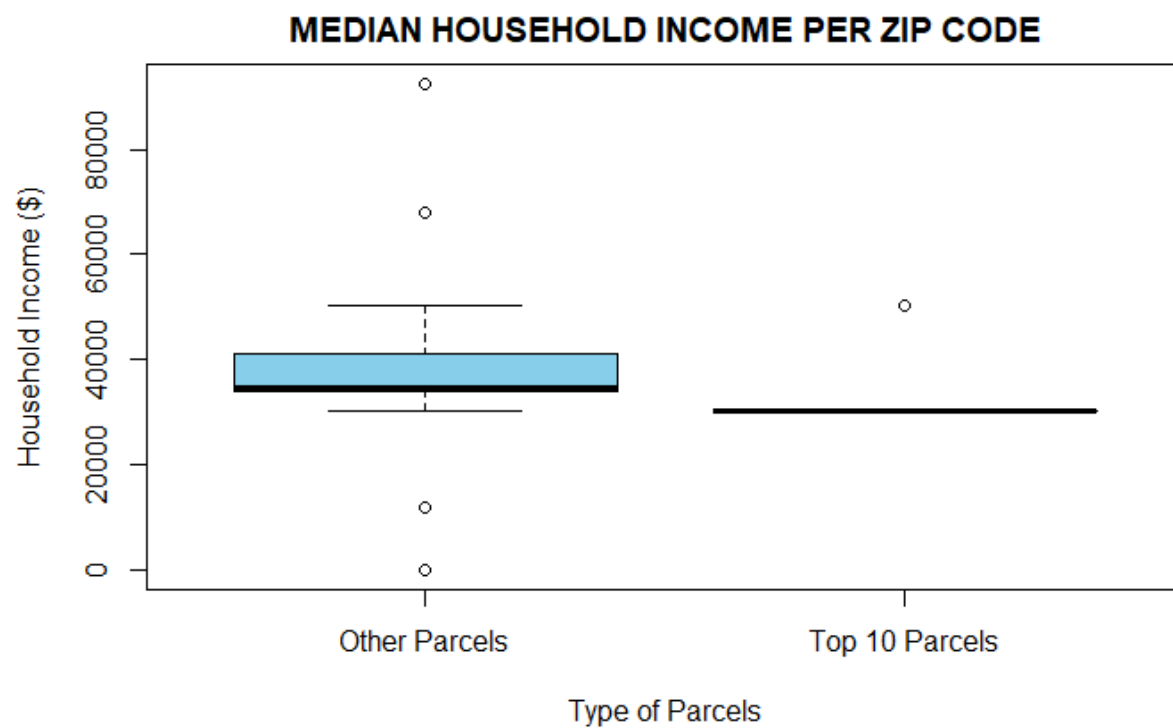


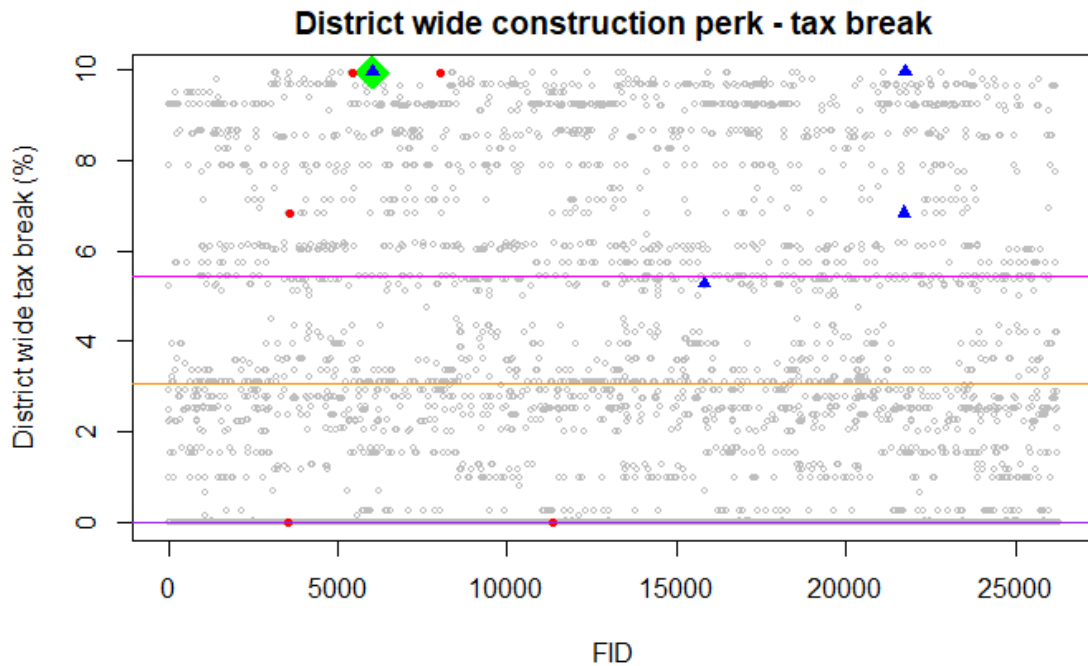
Except Parcel 15821, the top 10 have more affordable homes (about the percentage of Q3).



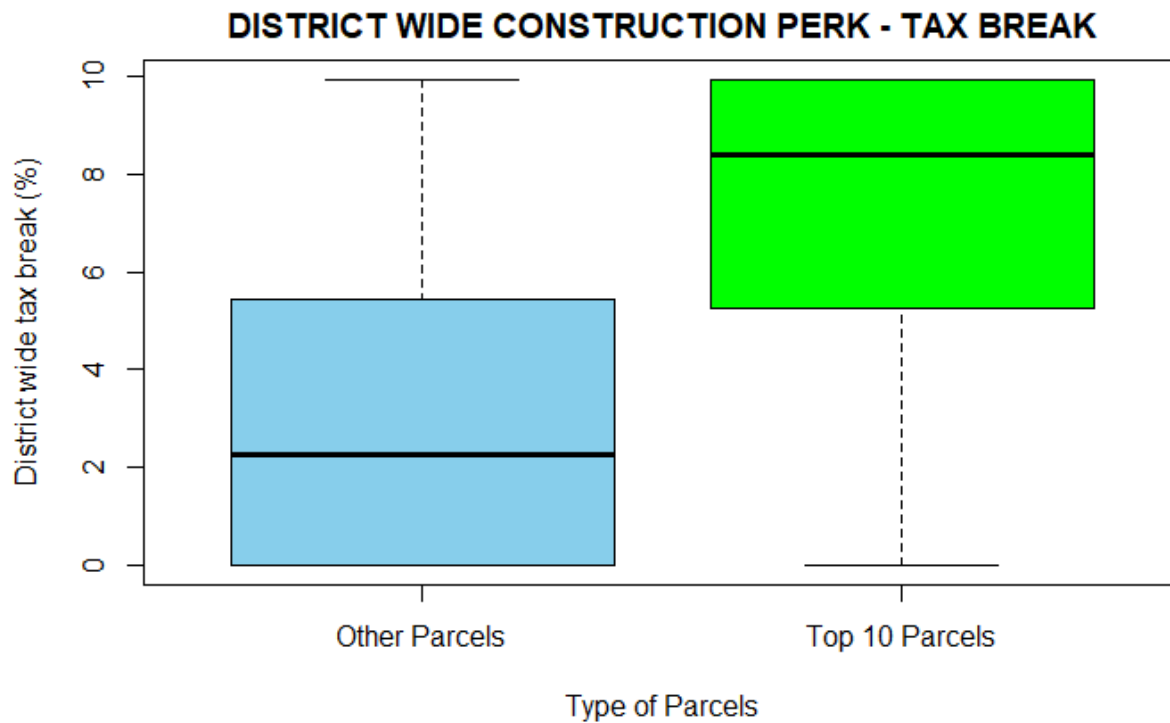


Except Parcel 15821, the top 10 have similar median household income to others.





The Top 10 mostly have way better tax break perks than others, above the entire distribution's Q3, though Parcel 3552 and 11370 have 0% which is Q1.



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

Within the limit of just one chart, all qualitative variables will have to fit in a purely text-based chart. In addition, bar plots have been included to support these statistics where applicable.

```
## [R]
# 12.2

projDataTop10 <- subset(projDataFiltered3, isTop10 == "Top 10 Parcels")
#summary(projDataTop10)

projDataTop10[,c("FID", "id_block", "land_type", "id_lot", "code_zoning",
"code_land_use_inv", "bool_bus_sys", "idx_housing_opp", "idx_ed_opp", "idx_econ_opp",
"idx_comp_opp", "descr_const_nearby")]
##
```

	FID <int>	id_block <fctr>	land_type <fctr>	id_lot <fctr>	code_zoning <fctr>	code_land_use_inv <int>
	3555	NA	PARCEL	3	SF-6-CO-NP	900
	3563	NA	LOT	1	ERC	900
	5430	B	LOT	1	ERC	100
	6008	A	LOT	1	NP	900
	6011	A	LOT	2	ERC	900
	8046	B	LOT	2	CS-CO-NP	900
	11370	NA	LOT	1-A	SF-6-CO-NP	900
	15824	E	LOT	TRACT B	NP	900
	21726	A	LOT	5	NP	900
	21742	A	LOT	1	ERC	900

1-10 of 10 rows | 1-7 of 12 columns

	bool_bus_sys <int>	idx_housing_opp <fctr>	idx_ed_opp <fctr>	idx_econ_opp <fctr>	idx_comp_opp <fctr>
	1	Low	Very Low	High	Very Low
	1	Low	Very Low	High	Very Low
	1	Low	Very Low	High	Very Low
	1	Low	Very Low	High	Very Low
	1	Low	Very Low	High	Very Low
	1	Low	Very Low	High	Very Low
	1	Low	Very Low	High	Very Low
	1	Low	Very Low	High	Very Low
	1	Low	Low	Moderate	Low
	1	Low	Very Low	High	Very Low
	1	Low	Very Low	High	Very Low

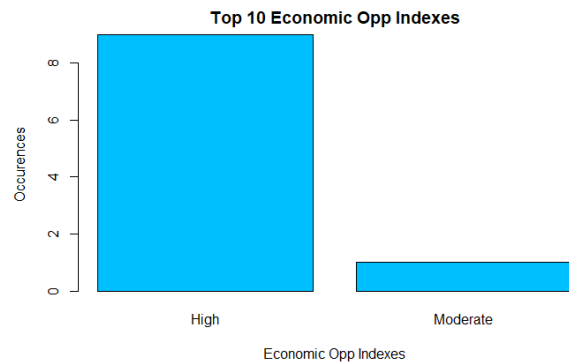
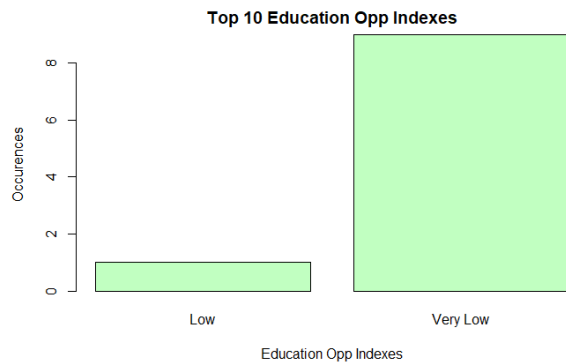
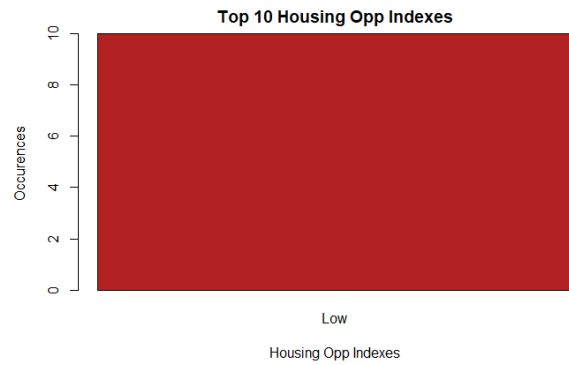
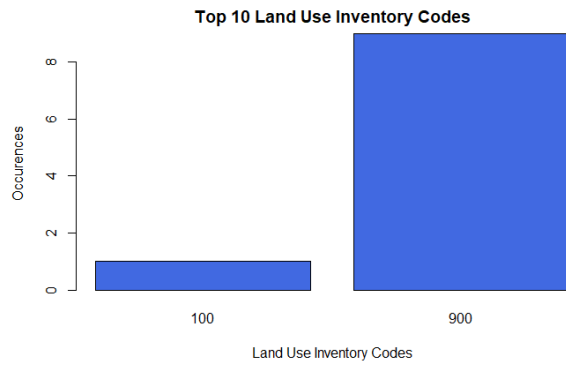
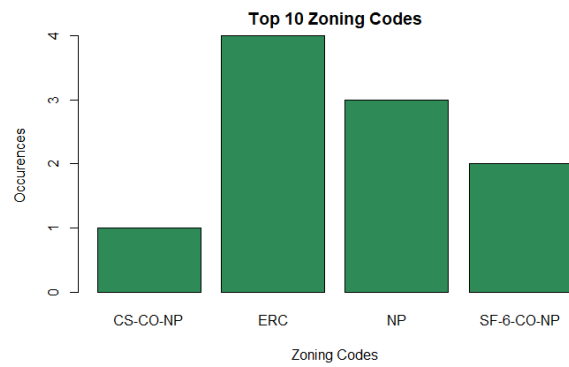
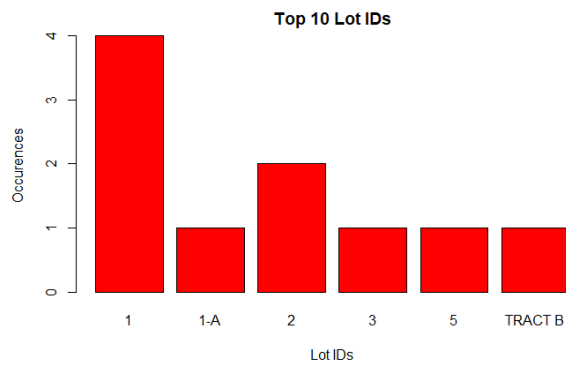
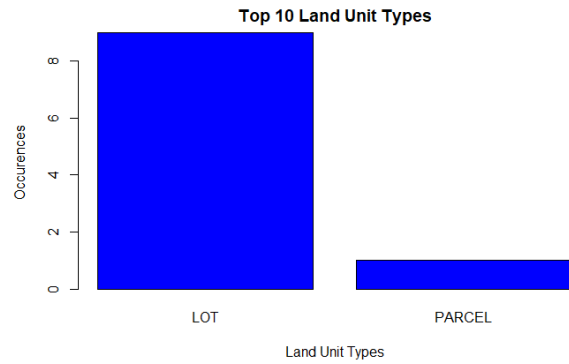
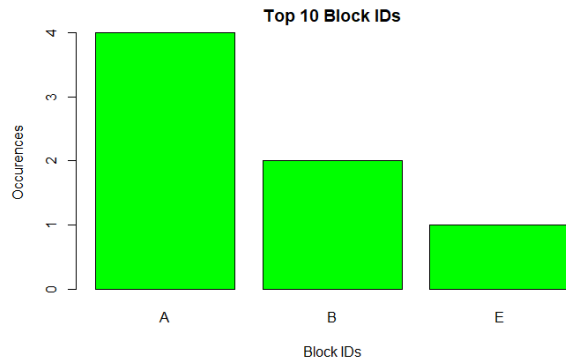
1-10 of 10 rows | 8-12 of 12 columns

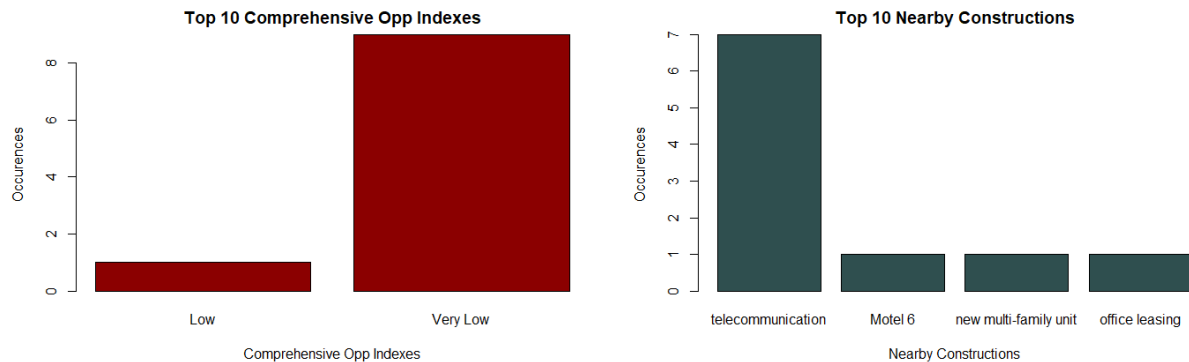
	descr_const_nearby <chr>
	adding equipment to existing wireless telecommunication tower
	adding equipment to existing wireless telecommunication tower
	eplan expedited review new construction of an amenityleasing office for a multifamily complex ph...
	adding equipment to existing wireless telecommunication tower
	adding equipment to existing wireless telecommunication tower
	install new irrigation system at motel 6
	adding equipment to existing wireless telecommunication tower
	expedited review new construction multifamily res dwelling units floors 1 thru 4 work to include 2...
	adding equipment to existing wireless telecommunication tower
	adding equipment to existing wireless telecommunication tower

1-10 of 10 rows | 13-13 of 12 columns

```
barplot(table(droplevels(projDataTop10$id_block)), col = "green",
main = "Top 10 Block IDs", xlab = "Block IDs", ylab = "Occurences")
```

IS 457 FA19
FINAL REPORT
CLASS ID: 104





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.

Parcel 3552:

- Strengths:
 - o Is the sole PARCEL land unit type in the list if the client is interested
 - o Is a townhouse and condominium
 - o Proximity to an existing wireless telecommunication tower
- Weaknesses:
 - o Has no district wide tax break
 - o Ranks low among the top 10 with a score of 5.4, which means it does not fit the client's given preferences very well

Parcel 3660:

- Strengths:
 - o Proximity to an existing wireless telecommunication tower
- Weaknesses:
 - o The context of "ERC" zoning code is unknown
 - o Ranks low among the top 10 with a score of 5.4, which means it does not fit the client's given preferences very well

Parcel 5427:

- Strengths:
 - o Is the only parcel with nearby new constructions of leasing office
- Weaknesses:
 - o Is a Single Family land with "One dwelling in a single building on one lot"
 - o Ranks low among the top 10 with a score of 5.4, which means it does not fit the client's given preferences very well

Parcel 6005:

- Strengths:
 - Proximity to an existing wireless telecommunication tower
- Weaknesses:
 - Is part of a Neighborhood Plan Combining District

Parcel 6008:

- Strengths:
 - Ranks the highest among the top 10 with a score of 6.4, which means it fits the client's specified preferences the most.
 - Proximity to an existing wireless telecommunication tower
- Weaknesses:
 - The context of "ERC" zoning code is unknown

Parcel 8043:

- Strengths:
 - This parcel is the only one close to Motel 6 which indicates its unique entertainment and enrichment values for employees
 - Clearly specified as a Commercial Services district
- Weaknesses:
 - Is part of a Neighborhood Plan Combining District
 - Ranks low among the top 10 with a score of 5.4, which means it does not fit the client's given preferences very well

Parcel 11367:

- Strengths:
 - Is a townhouse and condominium
 - Proximity to an existing wireless telecommunication tower
- Weaknesses:
 - Is part of a Neighborhood Plan Combining District
 - Is part of a Conditional Overlay Combining District
 - Ranks low among the top 10 with a score of 5.4, which means it does not fit the client's given preferences very well

Parcel 15821:

- Strengths:
 - Has the highest Education Opportunity Index among the Top 10
 - Has the highest Comprehensive Opportunity Index among the Top 10
 - Is next to a new construction of new multi-family unit
- Weaknesses:
 - Is part of a Neighborhood Plan Combining District
 - Has the lowest Economic Opportunity Index among the Top 10

Parcel 21723:

- Strengths:
 - Proximity to an existing wireless telecommunication tower
- Weaknesses:
 - Is part of a Neighborhood Plan Combining District

Parcel 21739:

- Strengths:
 - Proximity to an existing wireless telecommunication tower
- Weaknesses:
 - The context of “ERC” zoning code is unknown

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).

(1) As mentioned, 2 of the parcels are a townhouses and condominiums: 3552 and 11367

(2) The value of employee enrichment was one of the specified preferences, but its definition is too vague so there was no way to objectively measure how each parcel satisfies it. In the meantime, 8043’s proximity to Motel 6 makes it stand out as the sole parcel which somewhat fulfills this demand.

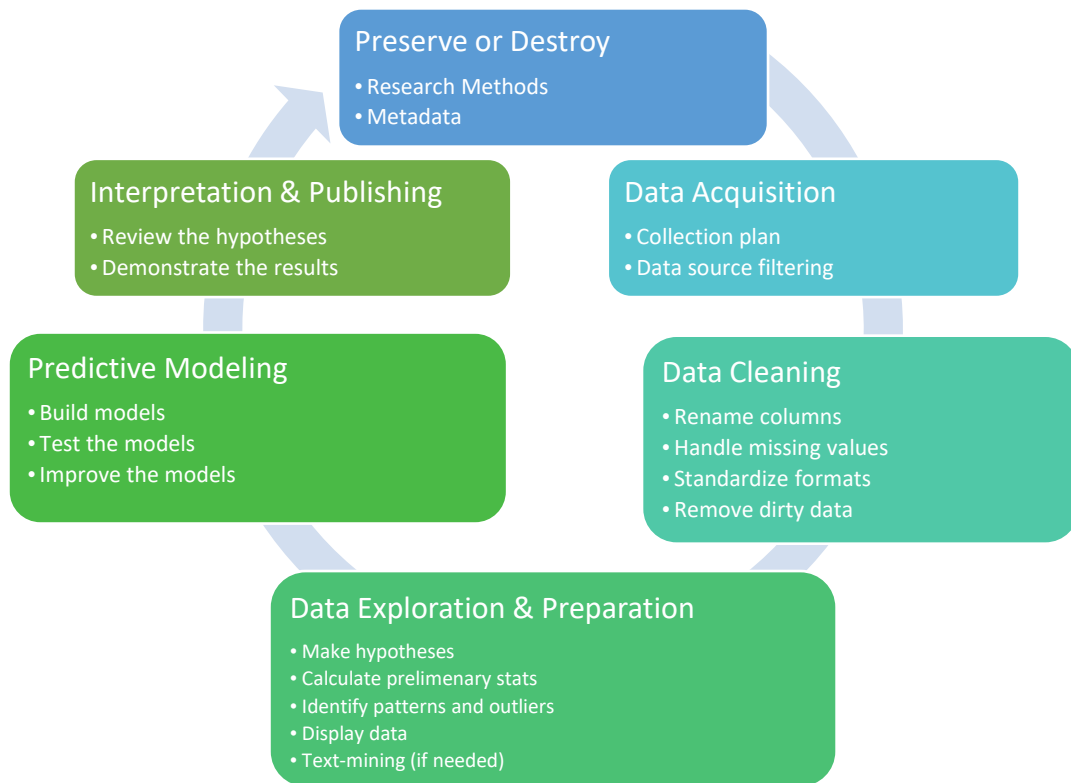
(3) An important factor that could shift the ranks of the Top 10 is the preference of district zoning types. As of now, no such preference has been expressed by the client, save for the one for “undeveloped” land. Each of the 10 has different zoning types, and the system currently has no way to measure their feasibility.

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.

Diagram created using built-in Word SmartArt.



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.

Data Acquisition:

- This step involves gathering data with variables (columns) that the collectors believe to be relevant for the topic under analysis
 - o Deciding which variables to collect, and whether each should be qualitative or quantitative
 - o Collecting from sources that are legitimate and keeping data as objective as manageable
- The data has already been collected for this project, so none of the action on the analyst's side fits in. Nonetheless, it is important to acknowledge its existence

Data Cleaning:

- This step involves preparing the data and making sure it is ready for analysis by getting rid of records out of sync or with wrong formatting
- Part 1: Data Pre-processing (especially Q3: Data Cleaning) fits into this step.
 - o Renaming columns so that they are comprehensible
 - o Dealing with missing values so they do not interfere with later analyses
 - o Altering values that do not conform to their columns formats, such as the dollar sign in tax breaks, and different capitalizations
 - o Removing columns that are no longer meaningful (such as land use codes (LAND_USE_2)) thanks to external developments
 - o Converting column types to handle data properly, such as converting factors to numerical if data are numeric

Data Exploration and Preparation:

- Also called "iterative process", this step is important for formulating hypotheses and can be the most time-consuming:
 - o Testing if there are any relationships
 - o Identifying patterns, if any
- This is almost identical to all steps in Part 2: Data Exploration
 - o Calculating statistics such as mean, median, Q1, Q3, max and min values
 - o Using graphs to display data and finding patterns and outliers
 - o Deciding which graphs to preserve based on how well they convey each of those
 - o Using p-values or other methods for testing relationships
 - o Text-mining which involves:
 - Removing stop words that convey no meaning

- Detecting words with high occurrences and testing whether they could become qualitative advantages

Predictive Modeling:

- This stage could blend with Data Exploration early on because the predictive model is still under construction.
 - Create a model that satisfies the given prompt
 - Adjusts the model based on testing results
- Part 3: Site Selection:
 - Question 9 was arguably where the modeling step blends with the preparation step, as it helped removing unwanted data and improving the quality of the subsequent model
 - The first versions of the model building in Question 10 (not included in the R file) returned many parcels of the same scores and the Top 10 could not be determined.
 - Adjustments were made so that full scores can only be obtained when multiple conditions are satisfied for each preference

Interpretation & Publishing:

- Depending on the scope of the previous steps, this process could vary
- This is in general an interpretation of the data and results conducted throughout the lifecycle
- If the model is ready for inciting decision-making from the clients, it is deployed
- Whether it is deployed or not, continuous upkeep and maintenance will still be performed by the current or future analyst(s) who pass the torches on
- Publishing can take the form of visualizing the results, making the results publicly available, and others
- Part 3: Site Selection:
 - Narrowing down the results to top 10 parcels
 - Determining which parcel best fits the requirements and preferences
 - Reviewing the selection process
- Part 4: Final Report Presentation
 - Using graphs to demonstrate the quantitative advantages of the Top 10
 - Listing the qualitative strengths and weaknesses of the Top 10

Preserve or Destroy:

- Though not a step commonly found in other sources, this was included in the course material. As such, the analyst believes it is necessary to include it
- Measures should be taken to ensure that the current analytic methods are reproducible and openly accessible to those interested in these data
- Part 4: Final Report Presentation
 - o Reviewing the research methods and pointing out parts that need improvement
 - o Documenting whether there are new factors that could impact decision making
 - o Exactly this section of the report which:
 - Reviews the entire process
 - Creates plans for data management
 - Sets up support for future users of this dataset

This subquestion was answered by referencing the course material and an independent source:

http://stodden.net/DataScience/Lectures/9_LifeCycleofDataScience.pdf

<https://community.alteryx.com/t5/Data-Science-Blog/The-Data-Science-Lifecycle/ba-p/408625>

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?

For the analyst, there are several options available:

(1) The crude but simple method: Depositing this document and the R markdown file used to calculate all statistics mentioned into a section of Google Drive and make the public link

*Advantages:

- + Quick deposit and low skill level
- + Original data without any modification from the process of storage

*Disadvantages:

- Those with access will need a locally available RStudio or equivalent tools to reproduce the research process
- No documentation of the workflow by default

(2) The difficult method: Deposit the entire project into wholetale.org

*Advantages:

- + Support for description and illustration of metadata and research methods
- + RStudio environment available on cloud

*Disadvantages:

- Higher skill level to set up
- Building and depositing requires more time

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?

(1) Steps to make things easier:

- A greater distinction between mandatory and optional columns
- Preview the entire workflow before jumping into action
 - o Some amount of time and effort was wasted in deleting columns that turned out to be required in latter sections
- A clear documentation of each column's process
 - o Cases of metadata loss like the LAND_USE_2 should be avoided
 - o Columns that convey overlapping information should either be altered or one of them should be picked at the start

(2) Suggestion to colleague in Seattle:

- Inspect client requirements and preferences before doing data exploration
 - o This can provide a clearer direction of which patterns to pay specific attention to and which hypotheses to make
- Have standard naming conventions for columns before data visualization
 - o Standard naming conventions allows the analyst to intuitively remember and recall columns when creating complex graphs. This eliminates the need to refer back to the definition table and improves efficiency.

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

At the time of this document, the analyst is unable to start or terminate a Whole Tale with RStudio setting still in building phase. This bonus section is therefore passed up.