IS457_finalProject_48.R

2019-12-09

```
#
**************************************************************************
*



#
**************************************************************************
*
#install.packages("tidyverse")
#install.packages("tidyverse")
#install.packages("qdap")
library(tidyverse)

## -- Attaching packages ---------------------------------------------------
--------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------
-------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(dplyr)
library(qdap)

## Loading required package: qdapDictionaries

## Loading required package: qdapRegex

##
## Attaching package: 'qdapRegex'

## The following object is masked from 'package:dplyr':
##
##     explain

## The following object is masked from 'package:ggplot2':
##
##     %+%

## Loading required package: qdapTools
```

```
##
## Attaching package: 'qdapTools'

## The following object is masked from 'package:dplyr':
##
##     id

## Loading required package: RColorBrewer

## Registered S3 methods overwritten by 'qdap':
##   method               from
##   t.DocumentTermMatrix tm
##   t.TermDocumentMatrix tm

##
## Attaching package: 'qdap'

## The following object is masked from 'package:forcats':
##
##     %>%

## The following object is masked from 'package:stringr':
##
##     %>%

## The following object is masked from 'package:dplyr':
##
##     %>%

## The following object is masked from 'package:purrr':
##
##     %>%

## The following object is masked from 'package:tidyr':
##
##     %>%

## The following object is masked from 'package:base':
##
##     Filter

library(grid)
library(naniar)

##
## Attaching package: 'naniar'

## The following object is masked from 'package:qdap':
##
##     %>%

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

### Part 1.
##Q1 Import the data using the read.csv() function.
#setwd("D:/UIUC/Fall 2019/Stat 430/final projects")
austin_lots = read.csv("Austin_Lots.csv")

##1.1 What are the initial dimensions of the dataset?
dim(austin_lots)

## [1] 26284    44

##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?

# Answer: created_by,date_creat, modified_b, date_modif
# These four columns seems least helpful because they are just log info for
the database, without containing any information about the location.

##1.3Subset your data by removing the unnecessary columns you identified.
What are the new dataset dimensions?
austin_lots$created_by = NULL
austin_lots$date_creat = NULL
austin_lots$modified_b = NULL
austin_lots$date_modif = NULL
dim(austin_lots)

## [1] 26284    40

##1.4 Why is it useful to subset your data before starting your analysis?
#Subsetting the data allows us just focus the parts of large files which are
of interest for a specific purpose.
#It's especially useful when there are lots of irrelevant data.

##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.
#old names
colnames(austin_lots)

##  [1] "FID"        "block_id"   "land_base_" "land_base1" "lot_id"
##  [6] "objectid"   "City_dist"  "Airpt_dist" "district"   "Shape_Area"
## [11] "zoning_o_3" "zcta5ce10"  "LAND_USE_2" "GENERAL_LA" "EWC_dist"
## [16] "NSC_dist"   "Mopac_dist" "X130_dist"  "X35_dist"   "ExTrail_1m"
## [21] "PpTrail_1m" "conf"       "bike_lanes" "Bus_area"   "TotBdgArea"
```

```
## [26] "Num_Bldgs"  "MaxBdgArea" "tax_break2" "bk_tx_brk"  "GEOID"
## [31] "Housing__"  "Education"  "Economic__" "Comprehens" "Med_HH_Inc"
## [36] "Med_rent"   "Med_home"   "Aff_rent_t" "Aff_own_te" "Descriptio"

#updated names
colnames(austin_lots)[1] = "row_id"
colnames(austin_lots)[3] = "land_base_id"
colnames(austin_lots)[4] = "land_base_type"
colnames(austin_lots)[11] = "zoning_designation"
colnames(austin_lots)[12] = "zipcode"
colnames(austin_lots)[22] = "bike_confLevel"
colnames(austin_lots)
```

```
##  [1] "row_id"             "block_id"           "land_base_id"
##  [4] "land_base_type"     "lot_id"             "objectid"
##  [7] "City_dist"          "Airpt_dist"         "district"
## [10] "Shape_Area"         "zoning_designation" "zipcode"
## [13] "LAND_USE_2"         "GENERAL_LA"         "EWC_dist"
## [16] "NSC_dist"           "Mopac_dist"         "X130_dist"
## [19] "X35_dist"           "ExTrail_1m"         "PpTrail_1m"
## [22] "bike_confLevel"     "bike_lanes"         "Bus_area"
## [25] "TotBdgArea"         "Num_Bldgs"          "MaxBdgArea"
## [28] "tax_break2"         "bk_tx_brk"          "GEOID"
## [31] "Housing__"          "Education"          "Economic__"
## [34] "Comprehens"         "Med_HH_Inc"         "Med_rent"
## [37] "Med_home"           "Aff_rent_t"         "Aff_own_te"
## [40] "Descriptio"
```

```
##Q2 Dealing with missing values.
##2.1 What columns in the dataset contain missing values?

#Amswer:
#Blank is used to indicate missing values in block_id, lot_id,
land_base_type, zoning_designation, Housing__, Education, Economic__,
Comprehens, Descriptio.
#0 is used to indicate missing values in LAND_USE_2 , GENERAL_LA.
colSums(is.na(austin_lots))
```

```
##             row_id           block_id       land_base_id
##                  0                  0                  0
##     land_base_type             lot_id           objectid
##                  0                  0                  0
##          City_dist         Airpt_dist           district
##                  0                  0                  0
##         Shape_Area zoning_designation            zipcode
##                  0                  0                  0
##         LAND_USE_2         GENERAL_LA           EWC_dist
##                  0                  0                  0
##           NSC_dist         Mopac_dist          X130_dist
##                  0                  0                  0
##           X35_dist         ExTrail_1m         PpTrail_1m
```

```
##                   0                   0                   0
##     bike_confLevel         bike_lanes           Bus_area
##                   0                   0                   0
##         TotBdgArea          Num_Bldgs         MaxBdgArea
##                   0                   0                   0
##         tax_break2          bk_tx_brk              GEOID
##                   0                   0                 168
##          Housing__          Education         Economic__
##                   0                   0                   0
##         Comprehens         Med_HH_Inc           Med_rent
##                   0                   1                   1
##           Med_home         Aff_rent_t         Aff_own_te
##                   1                   1                   1
##         Descriptio
##                   0
```

```
#NA is used to indicate missing values in GEOID, and there is one instance
(row_id: 470) of NA in Med_HH_Inc, Med_rent, Med_home, Aff_rent_t,
Aff_own_te.
filter(austin_lots, austin_lots$row_id == 470)
```

```
##   row_id block_id land_base_id land_base_type lot_id objectid City_dist
## 1    470        2      1888291            LOT      4    22124   1659.12
## 2    470        2      1888291            LOT      4    22124   1659.12
##   Airpt_dist district Shape_Area zoning_designation zipcode LAND_USE_2
## 1    10473.6       14   7835.848                 NP   78702      95482
## 2    10473.6       14   7835.848                 NP   78702      95482
##   GENERAL_LA EWC_dist NSC_dist Mopac_dist X130_dist X35_dist ExTrail_1m
## 1        100  6654.54  3874.39    4427.11   10564.3  1180.94          0
## 2        100  6654.54  3874.39    4427.11   10564.3  1180.94          0
##   PpTrail_1m bike_confLevel bike_lanes Bus_area TotBdgArea Num_Bldgs
## 1          8              1         22        1    442.029         2
## 2          8              1         22        1    442.029         2
##   MaxBdgArea tax_break2 bk_tx_brk    GEOID Housing__ Education Economic__
## 1    242.695          0         0 4.85e+11
## 2    242.695          0         0 4.85e+11  Very Low  Moderate   Moderate
##   Comprehens Med_HH_Inc Med_rent Med_home Aff_rent_t Aff_own_te
## 1                   NA       NA       NA         NA         NA
## 2   Very Low      34734      766   175400         99         67
##                         Descriptio
## 1
## 2 tenant finishout to create retail
```

```
##2.2 Briefly describe how you deal will with these missing values and
justify why you chose these methods.
#Answer:
#Let's take a look at different levels of block_id as an example.
levels(austin_lots$block_id)
```

```
##   [1] " "       "1"       "10"      "100"     "101"     "102"     "103"
##   [8] "104"     "105"     "106"     "107"     "108"     "109"     "11"
```

```
##  [15] "110"     "111"     "112"     "113"     "114"     "115"     "116"
##  [22] "117"     "119"     "12"      "120"     "121"     "122"     "123"
##  [29] "124"     "125"     "126"     "128"     "129"     "12I"     "12L"
##  [36] "13"      "130"     "131"     "132"     "134"     "135"     "136"
##  [43] "137"     "138"     "14"      "140"     "141"     "142"     "143"
##  [50] "146"     "147"     "148"     "149"     "15"      "150"     "152"
##  [57] "153"     "154"     "156"     "157"     "158"     "159"     "16"
##  [64] "160"     "161"     "162"     "164"     "165"     "166"     "167"
##  [71] "168"     "17"      "170"     "171"     "172"     "173"     "174"
##  [78] "175"     "176"     "178"     "179"     "18"      "185"     "188"
##  [85] "19"      "2"       "20"      "21"      "22"      "23"      "24"
##  [92] "25"      "26"      "27"      "28"      "29"      "2A"      "3"
##  [99] "30"      "31"      "32"      "33"      "34"      "35"      "36"
## [106] "38"      "39"      "3A"      "4"       "40"      "41"      "42"
## [113] "43"      "43A\"\"" "44"      "44A"     "45"      "46"      "47"
## [120] "48"      "49"      "5"       "50"      "51"      "52"      "53"
## [127] "57"      "58"      "6"       "6-A"     "6.5"     "62"      "63"
## [134] "64"      "69"      "7"       "70"      "72"      "74"      "75"
## [141] "76"      "77"      "78"      "79"      "8"       "80"      "82"
## [148] "83"      "84"      "85"      "86"      "87"      "88"      "89"
## [155] "9"       "91"      "92"      "93"      "94"      "95"      "96"
## [162] "97"      "98"      "99"      "A"       "AB"      "AC"      "b"
## [169] "B"       "B5"      "B6"      "C"       "D"       "E"       "F"
## [176] "G"       "H"       "I"       "J"       "K"       "L"       "M"
## [183] "N"       "O"       "P"       "Q"       "R"       "S"       "T"
## [190] "U"       "V"       "W"       "X"       "Y"       "Z"
```

#Because we do not have more information about specific block_id meaning, I would like to keep the missing values blank, since it makes no sense to replace by mean value or drop the observation (may be useful in the future).
#For the same reason, I would like to keep the cells blank in block_id, lot_id, land_base_type, zoning_designation, and Descriptio.

#For LAND_USE_2 , GENERAL_LA, I would like to keep the values 0 since there's no expalination for code 0 in applendix.
#For Housing__, Education, Economic__, and Comprehens, I would like to replace the blank values by "Moderate", because it is "Moderate" makes seem when we do not have the information.
#We will drop the colume GEOID in step 3.4, so we don't care about the NA values in GEOID.
#I would like to drop the instance (row_id: 470) of NA values in Med_HH_Inc, Med_rent, Med_home, Aff_rent_t, Aff_own_te with 0.

##2.3 Describe how your choice of method to deal with missing values may affect your later analysis.
#Answer: Depend on how we make use of the values later, keeping the value NA may require us to omit some entries.
#For the values that we replaced by "Moderate", we are pushing more lavues toward the average, this may lead our result less extreme than it may actaully be.

```r
#We will deal with the value row_id = 470 in step 3.5 because it's a
duplicate.

##2.4 Implement your methods for dealing with the missing values.
#Answer:
austin_lots$LAND_USE_2[austin_lots$LAND_USE_2 %in% c("0")] = NA
austin_lots$GENERAL_LA[austin_lots$GENERAL_LA %in% c("0")] = NA
austin_lots$Housing__[austin_lots$Housing__ %in% c("", " ")] = "Moderate"
austin_lots$Education[austin_lots$Education %in% c("", " ")] = "Moderate"
austin_lots$Economic__[austin_lots$Economic__ %in% c("", " ")] = "Moderate"
austin_lots$Comprehens[austin_lots$Comprehens %in% c("", " ")] = "Moderate"


##2.5 After dealing with missing values, once again show the new dimensions
of the dataset.
#Answer:
dim(austin_lots)

## [1] 26284    40

##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.
#Answer:
#"land_base1" was renamed as "land_base_type".
land_levels = levels(austin_lots$land_base_type)
length(land_levels)

## [1] 10

#10 unique values exist.
out = sapply(land_levels, function(x)
length(which(austin_lots$land_base_type==x)))
out

##           Lot    LOT   lott  OTHER Parcel PARCEL    PCL  Tract  TRACT
##    121   3433  20442      1      2     56   2166      1      4     58

# Lot, LOT, lott, should be treat as the same thing; Parcel, PARCEL, PCL
shoulde be treated as the same thing; Tract, TRACT shoule be treated as the
same thing.

#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?)
#Answer:
austin_lots = within(austin_lots, land_base_type[land_base_type %in%
c("Lot","LOT","lott")] <- "LOT")
austin_lots = within(austin_lots, land_base_type[land_base_type %in%
```

```
c("Parcel","PCL", "PARCEL")] <- "PARCEL")
austin_lots = within(austin_lots, land_base_type[land_base_type %in%
c("TRACT","Tract")] <- "TRACT")
out = sapply(land_levels, function(x)
length(which(austin_lots$land_base_type==x)))
out
```

```
##           Lot    LOT   lott  OTHER Parcel PARCEL    PCL  Tract  TRACT
##    121      0  23876      0      2      0   2223      0      0     62
```

```
#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.
#Answer:
austin_lots$tax_break2 = as.numeric(gsub("\\$", "", austin_lots$tax_break2))

#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.
#Answer:
#4.85e+11 will be like 485000000000.
n_distinct(austin_lots$GEOID)
```

```
## [1] 2
```

```
#There is only 2 distinct number (NA and 4.85e+11) in this column. Thus, it
is not giving us enough information to identify each of them, becausue they
are suppose to be unique.
austin_lots$GEOID = NULL

#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.
#Answer:
#
austin_lots$row_id[duplicated(austin_lots$row_id)]
```

```
## [1] 376 470
```

```
filter(austin_lots, austin_lots$row_id == 376)
```

```
##   row_id block_id land_base_id land_base_type lot_id objectid City_dist
## 1    376        2      1970747            LOT     29   353636   2194.53
## 2    376        2      1970747            LOT     29   353636   2194.53
##   Airpt_dist district Shape_Area zoning_designation zipcode LAND_USE_2
## 1    9078.47       14   5770.635                NP   78702      98593
## 2    9078.47       14   5770.635                NP   78702      98593
```

```
##   GENERAL_LA EWC_dist NSC_dist Mopac_dist X130_dist X35_dist ExTrail_1m
## 1        100  5255.58  2534.36    5644.46   9460.05  2252.81          0
## 2        100  5255.58  2534.36    5644.46   9460.05  2252.81          0
##   PpTrail_1m bike_confLevel bike_lanes Bus_area TotBdgArea Num_Bldgs
## 1         12              2          7        1    194.749         3
## 2         12              2          7        1    194.749         3
##   MaxBdgArea tax_break2 bk_tx_brk Housing__ Education Economic__
## 1    176.742    3.94564 0.0923577  Very Low  Very Low   Moderate
## 2    176.742    3.94564 0.0923577  Very Low  Very Low   Moderate
##   Comprehens Med_HH_Inc Med_rent Med_home Aff_rent_t Aff_own_te
## 1   Very Low      34734      766   175400         99         67
## 2   Very Low      34734      766   175400         99         67
##                                       Descriptio
## 1 Install irrigation system around the whole property
## 2 Install irrigation system around the whole property

#row_id = 376 is having fully duplicated rows
filter(austin_lots, austin_lots$row_id == 470)

##   row_id block_id land_base_id land_base_type lot_id objectid City_dist
## 1    470        2      1888291            LOT      4    22124   1659.12
## 2    470        2      1888291            LOT      4    22124   1659.12
##   Airpt_dist district Shape_Area zoning_designation zipcode LAND_USE_2
## 1    10473.6       14   7835.848                NP   78702      95482
## 2    10473.6       14   7835.848                NP   78702      95482
##   GENERAL_LA EWC_dist NSC_dist Mopac_dist X130_dist X35_dist ExTrail_1m
## 1        100  6654.54  3874.39    4427.11   10564.3  1180.94          0
## 2        100  6654.54  3874.39    4427.11   10564.3  1180.94          0
##   PpTrail_1m bike_confLevel bike_lanes Bus_area TotBdgArea Num_Bldgs
## 1          8              1         22        1    442.029         2
## 2          8              1         22        1    442.029         2
##   MaxBdgArea tax_break2 bk_tx_brk Housing__ Education Economic__
## 1    242.695          0         0  Moderate  Moderate   Moderate
## 2    242.695          0         0  Very Low  Moderate   Moderate
##   Comprehens Med_HH_Inc Med_rent Med_home Aff_rent_t Aff_own_te
## 1   Moderate         NA       NA       NA         NA         NA
## 2   Very Low      34734      766   175400         99         67
##                          Descriptio
## 1
## 2 tenant finishout to create retail

#row_id = 470 is having partially duplicated rows, drop the first instance
since there is more info in the second
austin_lots = austin_lots[-c(377,473),]
#the dropped instances were at row 377, 473
dim(austin_lots)

## [1] 26282    39

#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.
#Answer:
austin_lots$LAND_USE_2 = NULL
#Metadata is the data about data -- metadata describes data containing
specific information like type, length, textual description and other
characteristics.
#Thus, it's very important to help understand the data.

#3.7 Describe why these cleaning steps are necessary. What would happen if
you needed to use these columns in later analyses?
#Answer:
#Data cleansing is important because it improves your data quality and
overall productivity. Removing duplicates and standeardizing the values allow
us get better results in later steps.
#If I needed to use deleted columns in later analyses, I'll read the csv file
again and add back the columns needed.



#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).
#Answer:
#Bus_area should be converted from integer to factor, because we only allow 0
or 1.
#Aff_rent_t and Aff_own_te may be converted to numeric, because they mean to
be percentage and then it would be more accurate.
#Descriptio should be convert from factor to characters, because they are too
many different Descriptions.



#Q4 Transform columns into proper formats.
#4.1 Please display the initial variable classes for each column.
#Answer:
sapply(austin_lots,class)
```

```
##             row_id           block_id       land_base_id
##          "integer"           "factor"          "integer"
##     land_base_type             lot_id           objectid
##           "factor"           "factor"          "integer"
##          City_dist         Airpt_dist           district
##          "numeric"          "numeric"          "integer"
##         Shape_Area zoning_designation            zipcode
##          "numeric"           "factor"          "integer"
##         GENERAL_LA           EWC_dist           NSC_dist
##          "integer"          "numeric"          "numeric"
##         Mopac_dist          X130_dist           X35_dist
##          "numeric"          "numeric"          "numeric"
##         ExTrail_1m         PpTrail_1m     bike_confLevel
##          "integer"          "integer"          "integer"
##         bike_lanes           Bus_area         TotBdgArea
```

```
##          "integer"          "integer"          "numeric"
##          Num_Bldgs         MaxBdgArea         tax_break2
##          "integer"          "numeric"          "numeric"
##          bk_tx_brk          Housing__          Education
##          "numeric"           "factor"           "factor"
##         Economic__         Comprehens         Med_HH_Inc
##           "factor"           "factor"          "integer"
##           Med_rent           Med_home         Aff_rent_t
##          "integer"          "integer"          "integer"
##         Aff_own_te         Descriptio
##          "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.
#Answer:
#Descriptio shoulde be changed from factor to character, because there are
too many instances of Descriptio and it should not be a few choices as in
factor.
#Bus_area should be converted from integer to factor, because we only allow 0
or 1.
```

```
#4.3 Change the variable class(es) to one that is more fitting. Then display
the new class(es) for those columns.
#Answer:
austin_lots$Descriptio = as.character(austin_lots$Descriptio)
austin_lots$Bus_area = factor(austin_lots$Bus_area)
class(austin_lots$Descriptio)
```

```
## [1] "character"
```

```
class(austin_lots$Bus_area)
```

```
## [1] "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?
#Answer:
#R oftern have errors importing data when there are values with blank spaces,
commas, so each word will be interpreted as a separate variable, resulting in
errors that are related to the number of elements per line in your data set
#R also have difficulty to differente betweeen intergers and numeric values,
R doesn't know if we really want to put those values as interger without
decimals or not.
#Checking the data type/class is important because this allows us to operate
on the values easily in the future, and it puts some limits on the infomation
we can put in the column, and thus help us avoid inputting wrong data when
updating the table.
```

```r
##Part 2: Data Exploration
#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.
#Answer:
head(austin_lots, n=10)
```

```
##    row_id block_id land_base_id land_base_type lot_id objectid City_dist
## 1       0               1876887         PARCEL     14   356102  3208.660
## 2       1               1676746            LOT     18   296037  3203.180
## 3       2               1839096            LOT      6   319082  3187.940
## 4       3        A      1909677            LOT 15B-1A   333367  3089.870
## 5       4               1650609         PARCEL          270888  6319.060
## 6       5               1647428         PARCEL          160741  6583.260
## 7       6               1880381         PARCEL          266031  6736.650
## 8       7               1741600         PARCEL          344624  1005.030
## 9       8               1726221         PARCEL          318570  1126.080
## 10      9               1659892            LOT      8   147975   209.848
##    Airpt_dist district  Shape_Area zoning_designation zipcode GENERAL_LA
## 1    10007.60       14    7022.986                      78704        100
## 2    12720.80       14    7716.545                 NP   78705        100
## 3    12793.40       14   18296.797                 NP   78705        100
## 4    12714.60       14    5604.736         MF-6-CO-NP   78705        100
## 5     4680.63       14   63141.045               I-RR   78742        300
## 6     4655.52       14  469659.576               I-RR   78742        500
## 7     3852.66       14 2625995.232                 LI   78742        900
## 8     9451.75       14   94264.494                TOD   78702        500
## 9     9345.15       14   37557.439                 NP   78702        200
## 10   10330.80       14    3006.199                TOD   78702        400
##     EWC_dist NSC_dist Mopac_dist X130_dist X35_dist ExTrail_1m PpTrail_1m
## 1  2395.4600 5935.480    3625.73  12690.60 1453.610          1         17
## 2  8688.3398 5656.730    3321.37  12049.60  311.009          0          5
## 3  8640.8799 5769.830    3187.75  12163.20  424.619          0          6
## 4  8548.2598 5790.150    3160.69  12141.90  402.334          0          6
## 5   823.6870   90.777    9798.61   5658.94 5294.910          0          8
## 6   905.1980  470.644   10089.60   5059.46 5644.550          0          9
## 7    73.6846  238.516   10202.20   4911.38 5643.280          0         10
## 8  5069.7900 3496.590    4552.48  10534.60 1027.800          4         25
## 9  4997.7998 3402.990    4656.52  10442.20 1118.190          4         24
## 10 5398.2900 4469.820    3667.17  11503.90  155.830          2         13
##    bike_confLevel bike_lanes Bus_area TotBdgArea Num_Bldgs MaxBdgArea
## 1               2         22        1   238.8290         3   179.2370
## 2               2          9        1   136.6550         3   122.6870
## 3               2         18        1   295.2630         1   295.2630
## 4               2         13        1   137.7780         1   137.7780
## 5               3         14        1   505.5580         6   248.5640
## 6               0          0        1    29.4237         2    17.6246
## 7               4          9        1     0.0000         0     0.0000
## 8               2         19        1  7220.5298         2  5038.9800
```

```
## 9               1         18        1  1393.8101         5   364.6880
## 10              1         28        1  3630.6001         1  3630.6001
##    tax_break2 bk_tx_brk Housing__ Education Economic__ Comprehens
## 1    9.520330         0  Very Low  Very Low   Moderate   Very Low
## 2    0.983745         0  Very Low      High  Very High   Moderate
## 3    0.983745         0  Very Low      High  Very High   Moderate
## 4    0.983745         0  Very Low      High  Very High   Moderate
## 5    0.000000         0  Very Low  Very Low        Low   Very Low
## 6    0.000000         0  Very Low  Very Low        Low   Very Low
## 7    0.000000         0  Very Low  Very Low        Low   Very Low
## 8    0.000000         0  Very Low       Low        Low   Very Low
## 9    0.000000         0  Very Low       Low        Low   Very Low
## 10   0.000000         0  Very Low       Low        Low   Very Low
##    Med_HH_Inc Med_rent Med_home Aff_rent_t Aff_own_te
## 1       50248      940   338200         99         33
## 2       11917     1088   292500         94         79
## 3       11917     1088   292500         94         79
## 4       11917     1088   292500         94         79
## 5       34076      639    54400        100        100
## 6       34076      639    54400        100        100
## 7       34076      639    54400        100        100
## 8       34734      766   175400         99         67
## 9       34734      766   175400         99         67
## 10      34734      766   175400         99         67
##
Descriptio
## 1  Change of use  Interior remodel from convenience store to
cafÃÂ¯ÃÂ¿ÃÂ½ retail  Scope of work to include a 155sf 1 story addition
## 2                                                                Remodel
to add 2 exterior doors to existing religious assembly
## 3                                                                Remodel
to add 2 exterior doors to existing religious assembly
## 4                                                                Remodel
to add 2 exterior doors to existing religious assembly
## 5                                                          AddingRemoving
Equipment tofrom Existing TowerEquipment Configuration
## 6                                                          AddingRemoving
Equipment tofrom Existing TowerEquipment Configuration
## 7                                                          AddingRemoving
Equipment tofrom Existing TowerEquipment Configuration
## 8
FINISHOUT TO CREATE PERSONAL SERVIES
## 9
FINISHOUT TO CREATE PERSONAL SERVIES
## 10
Tenant finish out to create retail

summary(austin_lots)
```

```
##      row_id         block_id      land_base_id       land_base_type
##  Min.   :    0          :12911   Min.   :  1635655   LOT    :23874
##  1st Qu.: 6570   A      : 1310   1st Qu.:  1712106   PARCEL : 2223
##  Median :13140   1      : 1076   Median :  1788364          :  121
##  Mean   :13140   3      :  934   Mean   : 28756094   TRACT  :   62
##  3rd Qu.:19711   2      :  912   3rd Qu.:  1863476   OTHER  :    2
##  Max.   :26281   B      :  899   Max.   :400842667   Lot    :    0
##                  (Other): 8240                       (Other):    0
##      lot_id         objectid        City_dist       Airpt_dist
##         : 4497   Min.   :     3   Min.   :    0   Min.   :   31.63
##  1      : 1851   1st Qu.: 93157   1st Qu.: 1796   1st Qu.: 7399.48
##  2      : 1655   Median :186121   Median : 2663   Median : 9800.56
##  3      : 1430   Mean   :186481   Mean   : 3352   Mean   : 9015.16
##  4      : 1323   3rd Qu.:279304   3rd Qu.: 4057   3rd Qu.:11242.33
##  5      : 1236   Max.   :375410   Max.   :13573   Max.   :13745.40
##  (Other):14290
##     district      Shape_Area        zoning_designation    zipcode
##  Min.   :14.0   Min.   :      19   NP      :15189     Min.   :78617
##  1st Qu.:14.0   1st Qu.:    5713           : 6047     1st Qu.:78702
##  Median :14.0   Median :    6940   UNO     :  587     Median :78704
##  Mean   :15.2   Mean   :   34659   TOD     :  564     Mean   :78712
##  3rd Qu.:14.0   3rd Qu.:    9336   AV      :  472     3rd Qu.:78722
##  Max.   :21.0   Max.   :27533199   SF-4A-NP:  282     Max.   :78746
##                                    (Other) : 3141
##    GENERAL_LA       EWC_dist       NSC_dist          Mopac_dist
##  Min.   :100.0   Min.   :   0   Min.   :   6.676   Min.   :   11.15
##  1st Qu.:100.0   1st Qu.:2582   1st Qu.:2441.665   1st Qu.: 3167.51
##  Median :100.0   Median :4662   Median :4139.355   Median : 4506.77
##  Mean   :305.3   Mean   :4356   Mean   :3969.737   Mean   : 5306.69
##  3rd Qu.:500.0   3rd Qu.:6021   3rd Qu.:5428.107   3rd Qu.: 6584.86
##  Max.   :940.0   Max.   :8726   Max.   :7786.030   Max.   :16494.80
##  NA's   :34
##    X130_dist           X35_dist          ExTrail_1m       PpTrail_1m
##  Min.   :   54.03   Min.   :   17.92   Min.   : 0.000   Min.   : 0.0
##  1st Qu.: 8903.22   1st Qu.:  788.13   1st Qu.: 0.000   1st Qu.: 7.0
##  Median :10737.75   Median : 1684.43   Median : 1.000   Median :14.0
##  Mean   :10221.83   Mean   : 2245.82   Mean   : 2.963   Mean   :15.3
##  3rd Qu.:12276.30   3rd Qu.: 2785.11   3rd Qu.: 4.000   3rd Qu.:24.0
##  Max.   :14789.30   Max.   :11544.00   Max.   :21.000   Max.   :47.0
##
##  bike_confLevel    bike_lanes     Bus_area    TotBdgArea
##  Min.   :0.000   Min.   :  0.00   0:  563   Min.   :    0.0
##  1st Qu.:1.000   1st Qu.: 11.00   1:25719   1st Qu.:  109.1
##  Median :2.000   Median : 15.00             Median :  219.2
##  Mean   :1.684   Mean   : 15.09             Mean   :  850.9
##  3rd Qu.:2.000   3rd Qu.: 19.00             3rd Qu.:  404.2
##  Max.   :4.000   Max.   :144.00             Max.   :66073.6
##
##    Num_Bldgs        MaxBdgArea          tax_break2      bk_tx_brk
##  Min.   :  0.00   Min.   :    0.00   Min.   :0.000   Min.   :0.000000
```

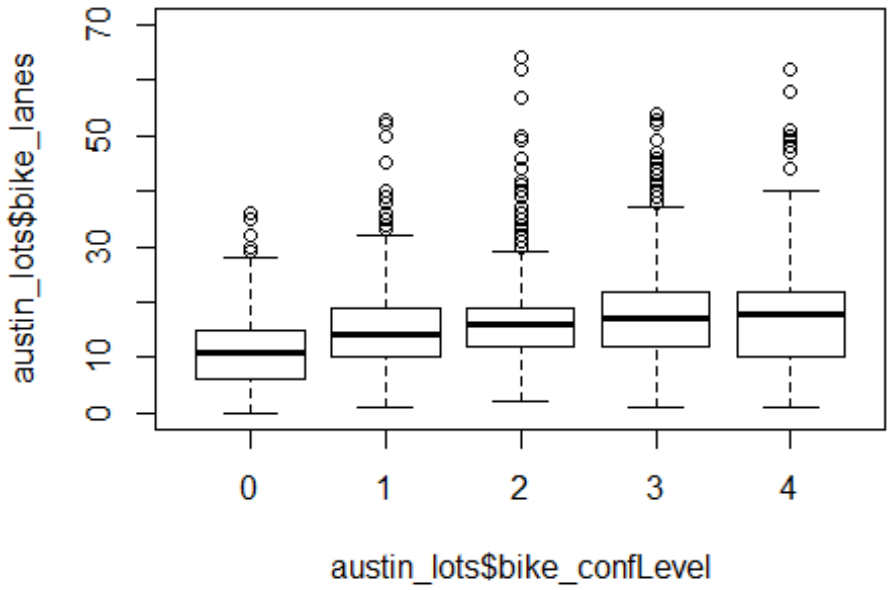
```
##  1st Qu.:  1.00   1st Qu.:   93.09   1st Qu.:0.000   1st Qu.:0.000000
##  Median :  1.00   Median :  163.73   Median :2.297   Median :0.000000
##  Mean   :  1.78   Mean   :  711.33   Mean   :3.189   Mean   :0.012906
##  3rd Qu.:  2.00   3rd Qu.:  270.70   3rd Qu.:5.727   3rd Qu.:0.003026
##  Max.   :137.00   Max.   :47366.60   Max.   :9.988   Max.   :0.099999
##
##      Housing__          Education         Economic__        Comprehens
##           :    0            :    0            :   0            :    0
##           :    0            :    0            :   0            :    0
##  Low      : 5478   High     : 3114   High     :5554   High     : 4162
##  Moderate :  867   Low      : 5705   Low      :3817   Low      : 5744
##  Very High:    2   Moderate : 3892   Moderate :5849   Moderate : 3135
##  Very Low :19935   Very High: 1979   Very High:9594   Very High: 1168
##                    Very Low :11592   Very Low :1468   Very Low :12073
##    Med_HH_Inc        Med_rent       Med_home        Aff_rent_t
##  Min.   :     0   Min.   :   0   Min.   :     0   Min.   :  0.00
##  1st Qu.: 34076   1st Qu.: 766   1st Qu.:120200   1st Qu.: 97.00
##  Median : 34734   Median : 835   Median :175400   Median : 99.00
##  Mean   : 41272   Mean   : 926   Mean   :229400   Mean   : 95.56
##  3rd Qu.: 50248   3rd Qu.: 946   3rd Qu.:338200   3rd Qu.: 99.00
##  Max.   :125327   Max.   :1590   Max.   :621900   Max.   :100.00
##  NA's   :1        NA's   :1      NA's   :1        NA's   :1
##    Aff_own_te      Descriptio
##  Min.   :  0.00   Length:26282
##  1st Qu.: 33.00   Class :character
##  Median : 67.00   Mode  :character
##  Mean   : 59.03
##  3rd Qu.: 79.00
##  Max.   :100.00
##  NA's   :1
```

```
#5.2 Describe anything you find that is unexpected or interesting.
#Answer:
levels(austin_lots$Education)
```

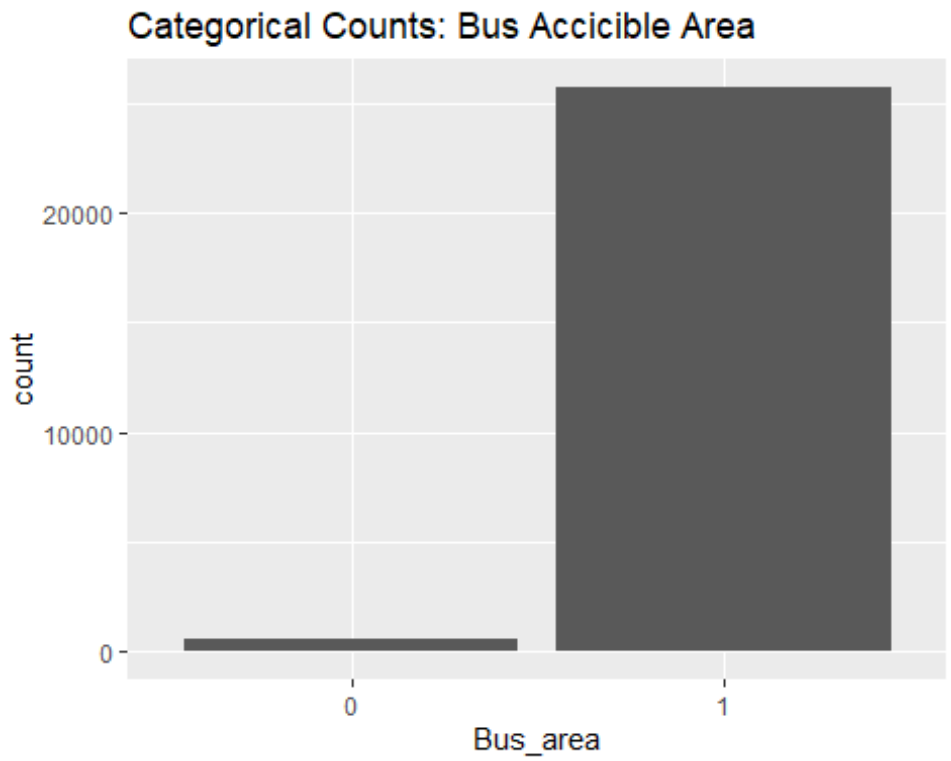
```
## [1] ""          " "         "High"      "Low"       "Moderate"  "Very
High"
## [7] "Very Low"
```

```
#For "Housing__", "Education", "Economic__", "Comprehens", there are two
types of blank cells: "" and " ". We shuold clean the data.
austin_lots$Housing__[austin_lots$Housing__ %in% c("", " ")] = NA
austin_lots$Education[austin_lots$Education %in% c("", " ")] = NA
austin_lots$Economic__[austin_lots$Economic__ %in% c("", " ")] = NA
austin_lots$Comprehens[austin_lots$Comprehens %in% c("", " ")] = NA


#Median of bk_tx_brk is 0.000000, meaning that more than half of the
bk_tx_brk values are 0!
#1st Qu. of Aff_rent_t is 97, which means the large majority of worker in
```

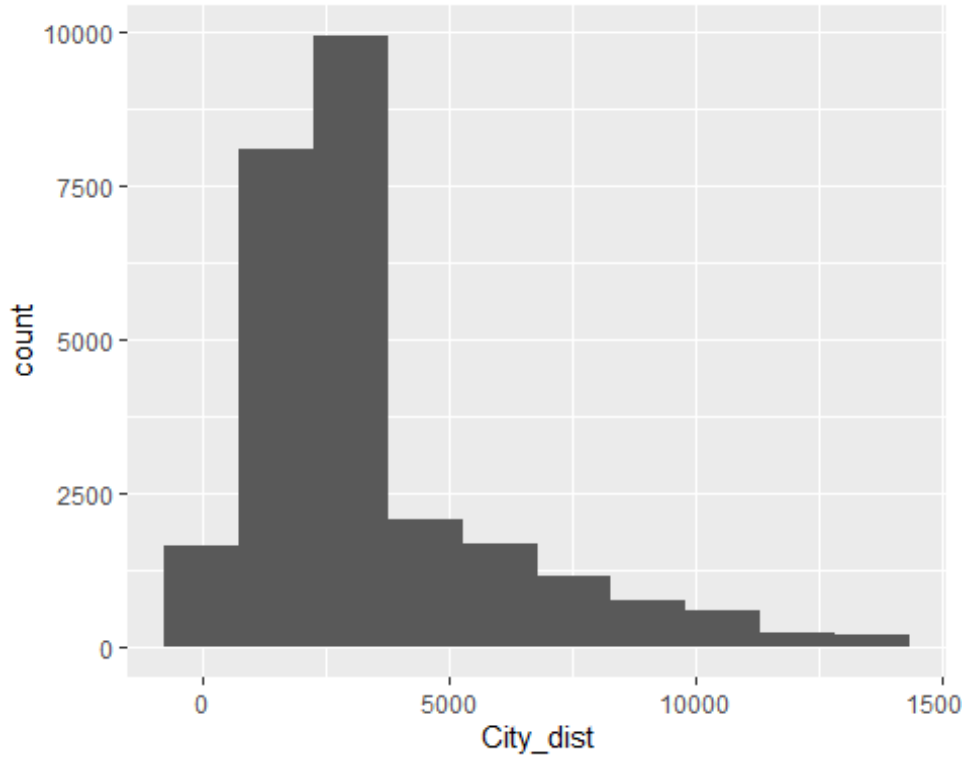
```
tech can afford to rent!

#Q6 Visualize the data.
#6.1 Think about the types of variables in the Austin dataset. Then choose
appropriate graphs to display distributions and trends for multiple
variables.
#Answer: some examples of the variables
boxplot(austin_lots$bike_lanes ~ austin_lots$bike_confLevel, ylim=c(0,70))
```



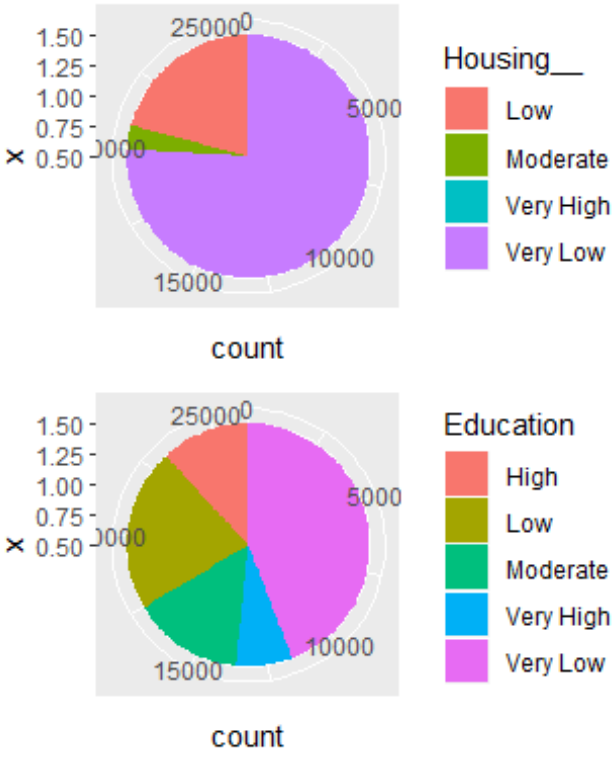

```
ggplot(data = austin_lots, aes(x = Bus_area)) + geom_histogram(stat="count")
+ ggtitle("Categorical Counts: Bus Accicible Area")

## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
ggplot(data = austin_lots, aes(x = City_dist) ) + geom_histogram(bins =10)
```

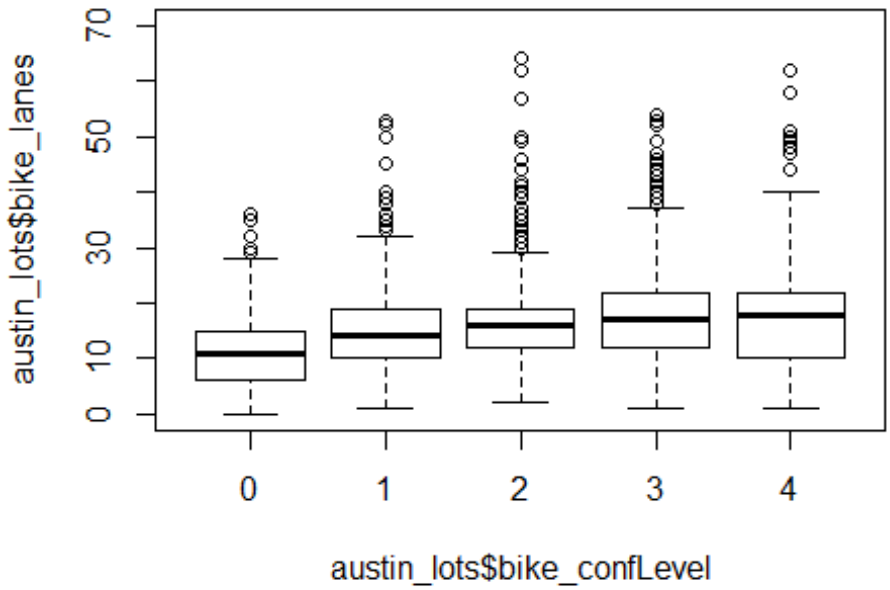


```r
pie1 = ggplot(austin_lots, aes(x=1, fill=Housing__))+
  geom_bar(width = 1)+
  coord_polar("y")
pie2 = ggplot(austin_lots, aes(x=1, fill=Education))+
  geom_bar(width = 1)+
  coord_polar("y")
grid.newpage()
pushViewport(viewport(layout = grid.layout(2,1)))
vplayout <- function(x,y){
  viewport(layout.pos.row = x, layout.pos.col = y)
}
print(pie1, vp = vplayout(1,1))
print(pie2, vp = vplayout(2,1))
```

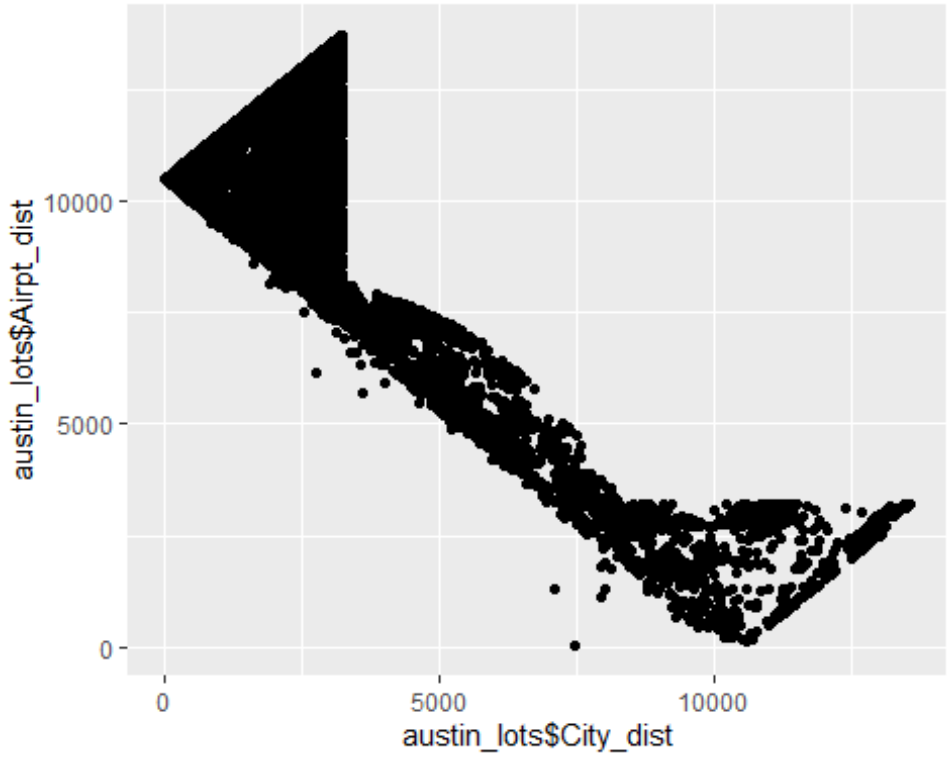



```r
#6.2 Compare different graph types to see which ones best convey trends, outliers, and patterns in the data.
#Answer:
#For compositional static data, pie chart are easy to see the percentage of different types.
#For distributional data, histogram are useful to see frequency, and scatter plots are useful to see outliers, and patterns in the data.
#For comparison data, scatter plots and line Charts are useful to see trend or see relationship between data.


#6.3 Describe what you find from the graphs.
#Answer:
```

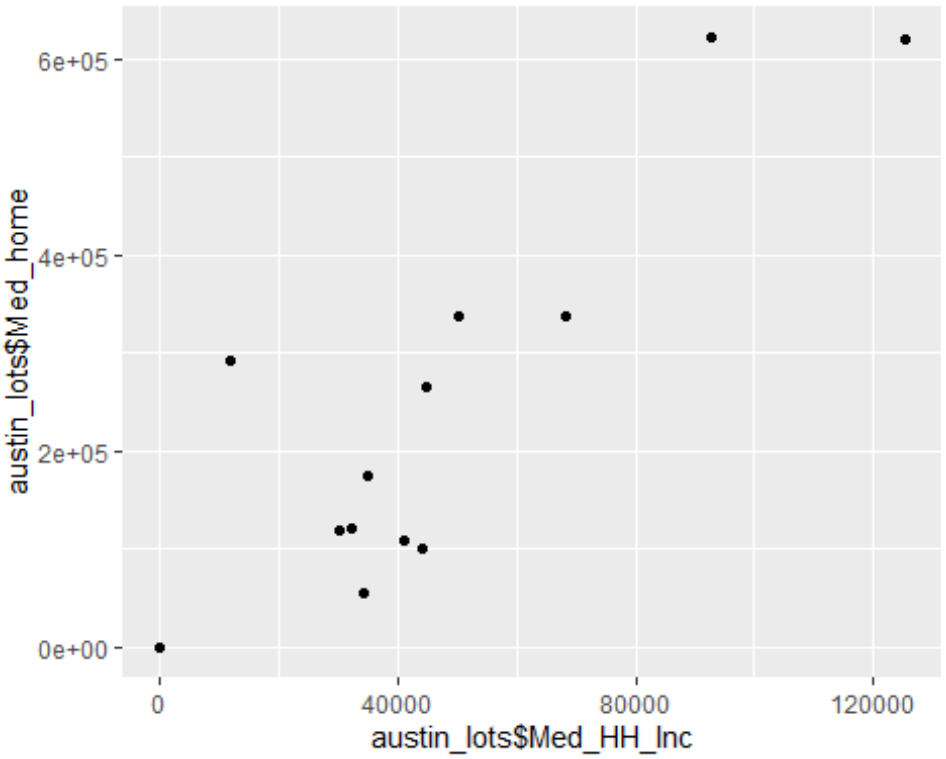
```r
#Q7 Now look at the relationships among several variables.
#7.1 For example, look at the original "conf" and "bike_lanes" columns. They are both indicators of ease of bicycle transportation, but each column conveys different information. What different information and what similar information can you get from these variables? How are the two variables related? Explain what you find.

boxplot(austin_lots$bike_lanes ~ austin_lots$bike_confLevel, ylim=c(0,70))
```



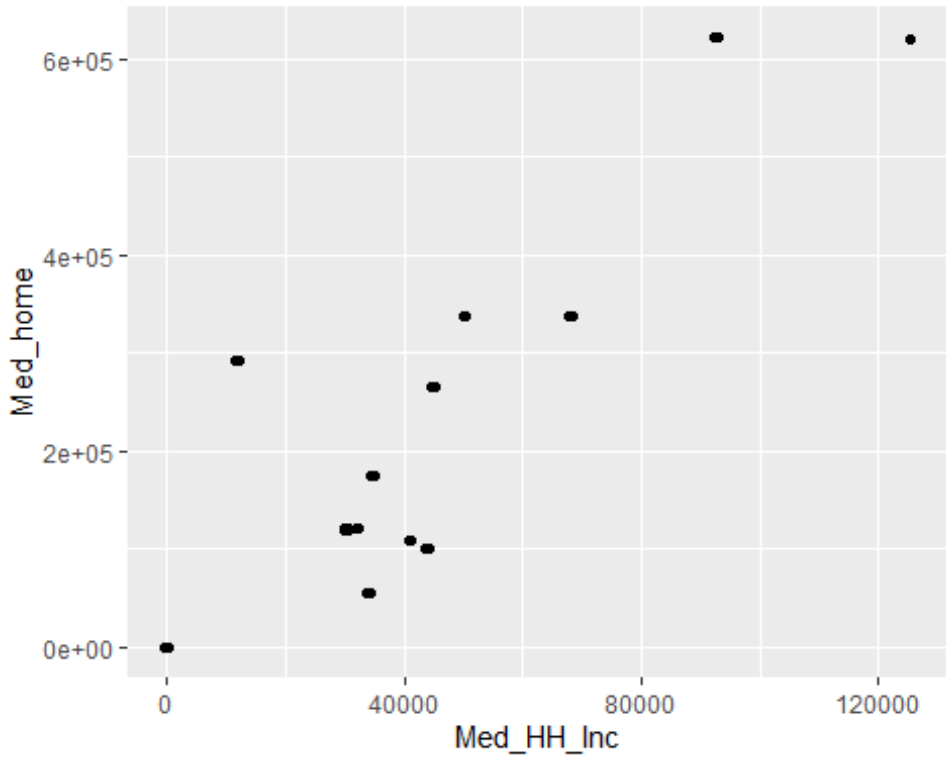

```r
#Answer: conf level is likely to be more subjective by people who did the survey, while number of lanes is objective by the fact.
#They are both telling the information for bikers, and both of them mean a more comfortable zone with higher value.
#From the boxplot, we can see the more lanes there are, the higher conft level in general, although it was not significant increase.

#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).
#Answer:
qplot(austin_lots$City_dist, austin_lots$Airpt_dist)
```




```
cor.test(austin_lots$City_dist, austin_lots$Airpt_dist)

##
##  Pearson's product-moment correlation
##
## data:  austin_lots$City_dist and austin_lots$Airpt_dist
## t = -235.77, df = 26280, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8278565 -0.8200940
## sample estimates:
##        cor
## -0.8240139

#We can see there is a pattern bewtween austin_lots$City_dist,
austin_lots$Airpt_dist, and they are strongly negatively related.

qplot(austin_lots$Med_HH_Inc, austin_lots$Med_home)

## Warning: Removed 1 rows containing missing values (geom_point).
```




```
p <- ggplot(austin_lots, aes(Med_HH_Inc, Med_home))
p + geom_jitter()

## Warning: Removed 1 rows containing missing values (geom_point).
```




```
cor.test(austin_lots$Med_HH_Inc,austin_lots$Med_home)

##
##  Pearson's product-moment correlation
##
## data:  austin_lots$Med_HH_Inc and austin_lots$Med_home
## t = 200.35, df = 26279, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7725715 0.7821392
## sample estimates:
##       cor
## 0.7774003

#WE can tell that austin_lots$Med_HH_Inc,austin_lots$Med_home are positively
correlated.

##Q8 Find areas that could be attractive to future employees.
#8.1Convert the letters in the "Descriptio" column to lower case. Why is this
helpful? Do you lose information by doing this?
#Answer:
austin_lots$Descriptio = lapply(austin_lots$Descriptio, tolower)
#it helps us to avoid count one unique word as two in different cases. We
lose very few infomation, except some word when different meaning when it's
capitilized.
#For example, We maycount some "A"s with special as "a"s, which is a stop
word.
```

```r
#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.
#Answer:
austin_lots$Descriptio = as.character(austin_lots$Descriptio)
wordlist = unique(unlist(strsplit(austin_lots$Descriptio, "\\s+")))
stopwords = 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", "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")
wordlist = wordlist[!wordlist %in% stopwords]
head(wordlist, n=50)
```

```
##  [1] "change"                "use"
##  [3] "interior"              "remodel"
##  [5] "convenience"           "store"
##  [7] "cafÃ£Â¯Ã¢Â¿Ã¢Â½"              "retail"
##  [9] "scope"                 "work"
## [11] "include"               "155sf"
## [13] "1"                     "story"
## [15] "addition"              "add"
## [17] "2"                     "exterior"
## [19] "doors"                 "existing"
## [21] "religious"             "assembly"
## [23] "addingremoving"        "equipment"
## [25] "tofrom"                "towerequipment"
## [27] "configuration"         "finishout"
## [29] "create"                "personal"
## [31] "servies"               "tenant"
## [33] "finish"                "out"
## [35] "new"                   "leasing"
## [37] "office"                "accessory"
## [39] "garage"                "residenceofficehistoric"
## [41] "structure"             "museum"
## [43] "chage"                 "laundmat"
## [45] "ligour"                "sale"
## [47] "residential"           "admin"
## [49] "1801"                  "e"
```

```r
#8.3 Preform a similar function to 8.2 but this time finding unique words and
their frequency. What are the 10 most frequent non stop words, i.e. which are
```

```
frequent words that give you meaningful information about the type of
construction occurring? How can these help you finding a good site for
GlobalTechSync?
#Answer:
wordlist = strsplit(as.character(austin_lots$Descriptio), " ")
word_freq = as.data.frame(table(tolower(unlist(wordlist))))
ord_freq = word_freq[-(word_freq$Var1 == ''),]

stop_word_position = c()
for(i in 1:length(stopwords)){
  stop_word_position = c(stop_word_position,which(word_freq$Var1 ==
stopwords[i]))
}
freq_without_sw = word_freq[-stop_word_position,]
arrange(freq_without_sw, desc(Freq))[1:10,]

##        Var1 Freq
## 1  existing 8912
## 2       new 7968
## 3  interior 4638
## 4   remodel 4345
## 5        lf 3919
## 6    office 2928
## 7  sidewalk 2879
## 8   replace 2505
## 9   service 2459
## 10  install 2415

#new, remodel, interior, sidewalk seems to be used since

#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?
#Answer:
arrange(freq_without_sw, desc(Freq))[1:30,]

##           Var1 Freq
## 1      existing 8912
## 2           new 7968
## 3      interior 4638
## 4       remodel 4345
## 5            lf 3919
## 6        office 2928
## 7      sidewalk 2879
## 8       replace 2505
## 9       service 2459
## 10      install 2415
## 11         city 2357
```

```
## 12         meet 2331
## 13        shall 2310
## 14    construct 2155
## 15    standards 2110
## 16 construction 2103
## 17         demo 1857
## 18     building 1725
## 19       change 1653
## 20   commercial 1534
## 21 installation 1526
## 22           cg 1466
## 23         site 1439
## 24            2 1403
## 25          per 1401
## 26         work 1370
## 27          ada 1301
## 28          out 1252
## 29   electrical 1219
## 30    equipment 1213
```

```
#"city"2357 ,"sidewalk"2879 may be useful to determine the place of the
parcel, meaning somewhere close to city center with a walking distance.
#Low frequency words may help determine what worker do not like about, thus
may also be useful.

#8.5 What additional word processing steps or stop words do you think would
be useful for further text analysis of this variable? You don't have to
implement these ideas.
#Answer:
# numbers (such as "2" 1403 times) doesn't contain too much info.
# Becasue it is about chosing new place for the office,  "new" 7968,"remodel"
4345, "replace"2505,"change"1653, are likely meaningless for deciding a
place.

###Part 3: Site Selection
##Q9 Filter out unsuitable parcels.
#9.1 Remove any parcels that are not in the metro bus service area.
austin_lots = austin_lots[austin_lots$Bus_area == 1,]
dim(austin_lots)
```

```
## [1] 25719    38
```

```
#9.2 Remove any parcels that have an area under 300 square meters.
austin_lots = austin_lots[austin_lots$Shape_Area >= 300,]
dim(austin_lots)
```

```
## [1] 25579    38
```

```
#9.3 Remove any parcels with a residential zoning area (use the zoning_o_3
column and the residential general zoning category).
`%notin%` <- Negate(`%in%`)
```

```r
austin_lots = austin_lots[austin_lots$zoning_designation %notin% 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"),]

#9.4 What are your new dataset dimensions after removing these rows?
dim(austin_lots)
```

```
## [1] 25530    38
```

```r
##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.
#Answer:
# Enrichment: for GENERAL_LA = 640 and 740
## undeveloped: GENERAL_LA = 900
austin_lots = austin_lots[(austin_lots$GENERAL_LA %in% c(640,740,900)),]

# Tax breaks or discounts:  higher tax_break2 or higher bk_tx_brk
austin_lots = austin_lots[(austin_lots$tax_break2 > 6.0 |
austin_lots$bk_tx_brk > 0.07),]

# Easy acces to interstate or highway: set 900 meter-distance as easy access
austin_lots = austin_lots[(austin_lots$EWC_dist<900 |
austin_lots$NSC_dist<900 | austin_lots$X130_dist<900 |
austin_lots$X35_dist<900),]

# Easy access by bike or on foot: If there exsists urban trail or bike lane
of a total of 15
austin_lots = austin_lots[(austin_lots$ExTrail_1m + austin_lots$bike_lanes) >
15,]

# Education: above moderate education opportunities
austin_lots = austin_lots[(austin_lots$Education ==
"Moderate"|austin_lots$Education == "High" | austin_lots$Education == "Very
High"),]

# Own houses: Aff_own_te > 80%
austin_lots = austin_lots[austin_lots$Aff_own_te > 80,]

#10.2 Using your ranking system, determine the top 10 best parcels to submit
to GlobalTechSync and record the parcel FIDs below.
dim(austin_lots)
```

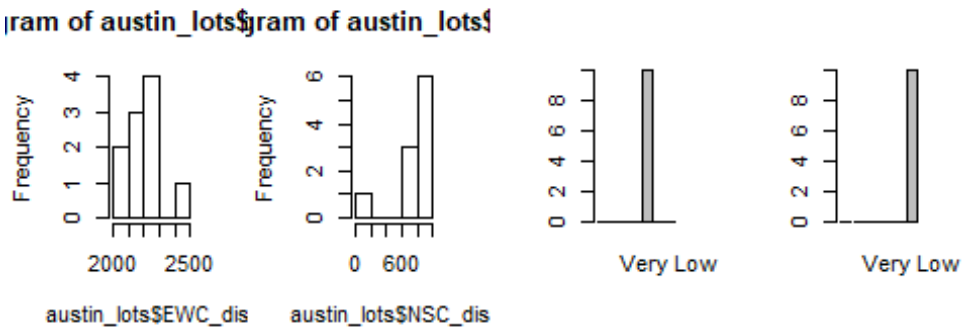
```
## [1] 10 38
```

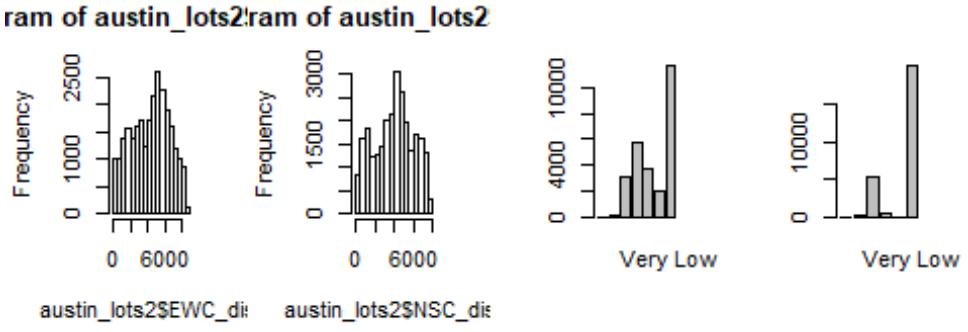
```r
# WE finally recommand parcels with row_id 2015, 7414, 7638, 10020, 10024,
10026,  10037,  15059,  15062,  20264
austin_lots
```

```
##       row_id block_id land_base_id land_base_type lot_id objectid
## 2018    2015               1759970         PARCEL          211231
## 7417    7414               1777537         PARCEL          240305
## 7641    7638        A    400544600            LOT      3   173156
## 10023  10020        A      1647184            LOT      1    19554
## 10027  10024        A      1654412            LOT      2   334595
## 10029  10026               1673646         PARCEL          219137
## 10040  10037               1740197         PARCEL           38694
## 15062  15059               1673648         PARCEL           69414
## 15065  15062               1684513         PARCEL          230997
## 20267  20264        2      1669430            LOT     10   161653
##       City_dist Airpt_dist district Shape_Area zoning_designation zipcode
## 2018    4803.93    5875.32       21   5076.074                 NP   78741
## 7417    4652.82    6024.08       21   2115.422                 NP   78741
## 7641    4701.03    6308.92       21   9239.703                 NP   78741
## 10023   4631.93    6031.13       21   8739.779                 NP   78741
## 10027   4640.00    6022.89       21   3518.152                 NP   78741
## 10029   4657.83    6003.19       21   6040.352                 NP   78741
## 10040   4665.28    5987.28       21   5423.412                 NP   78741
## 15062   4662.05    5995.71       21   4818.170                 NP   78741
## 15065   4807.42    5866.24       21   5964.633                 NP   78741
## 20267   4596.80    6084.54       21   7344.389                 NP   78741
##       GENERAL_LA EWC_dist NSC_dist Mopac_dist X130_dist X35_dist
## 2018         900  2057.81  753.707    8256.01   7201.35  3799.37
## 7417         900  2206.59  803.008    8106.19   7339.81  3668.40
## 7641         900  2443.98  148.038    8238.82   7013.29  4070.26
## 10023        900  2214.21  817.067    8084.29   7366.75  3638.46
## 10027        900  2206.18  849.145    8086.95   7377.75  3634.08
## 10029        900  2186.00  820.268    8107.68   7342.25  3658.29
## 10040        900  2170.52  847.715    8110.03   7352.50  3649.42
## 15062        900  2178.73  834.858    8109.00   7347.64  3654.23
## 15065        900  2048.65  766.489    8257.46   7205.71  3795.23
## 20267        900  2267.15  779.128    8058.84   7375.48  3635.24
##       ExTrail_1m PpTrail_1m bike_confLevel bike_lanes Bus_area TotBdgArea
## 2018           4         17              1         18        1      0.000
## 7417           4         19              1         13        1      0.000
## 7641           4         16              3         13        1      0.000
## 10023          4         19              1         14        1      0.000
## 10027          4         19              1         13        1      0.000
## 10029          4         19              1         14        1    285.629
## 10040          4         19              1         12        1    177.236
## 15062          4         19              1         13        1    154.360
## 15065          4         17              1         16        1      0.000
## 20267          4         19              1         13        1      0.000
##       Num_Bldgs MaxBdgArea tax_break2 bk_tx_brk Housing__ Education
## 2018          0      0.000    8.52720 0.0000000  Very Low  Moderate
## 7417          0      0.000    8.52720 0.0000000  Very Low  Moderate
## 7641          0      0.000    3.58902 0.0773219  Very Low  Moderate
## 10023         0      0.000    8.52720 0.0150272  Very Low  Moderate
## 10027         0      0.000    8.52720 0.0271360  Very Low  Moderate
```

```
## 10029         2    143.260    8.52720 0.0000000  Very Low  Moderate
## 10040         3    155.645    8.52720 0.0000000  Very Low  Moderate
## 15062         2    142.370    8.52720 0.0000000  Very Low  Moderate
## 15065         0      0.000    8.52720 0.0000000  Very Low  Moderate
## 20267         0      0.000    8.52720 0.0247921  Very Low  Moderate
##       Economic__ Comprehens Med_HH_Inc Med_rent Med_home Aff_rent_t
## 2018    Moderate        Low      30183      835   120200        100
## 7417    Moderate        Low      30183      835   120200        100
## 7641    Moderate        Low      30183      835   120200        100
## 10023   Moderate        Low      30183      835   120200        100
## 10027   Moderate        Low      30183      835   120200        100
## 10029   Moderate        Low      30183      835   120200        100
## 10040   Moderate        Low      30183      835   120200        100
## 15062   Moderate        Low      30183      835   120200        100
## 15065   Moderate        Low      30183      835   120200        100
## 20267   Moderate        Low      30183      835   120200        100
##       Aff_own_te                                    Descriptio
## 2018          93                          total demo of church
## 7417          93                          total demo of church
## 7641          93 electric service to montopolis traffic signal
## 10023         93                          total demo of church
## 10027         93                          total demo of church
## 10029         93                          total demo of church
## 10040         93                          total demo of church
## 15062         93                          total demo of church
## 15065         93                          total demo of church
## 20267         93                          total demo of church

##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?
#Answer:
#It was not too hard to select 10 parcels, because of the giving preferences.
I typicallytoo many parcels and then I put more strict limit on the
preference to reduce it to 10.

#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?
#Answer:
#I decided the cut offs by looking at the existing instances and decide based
on my own undertstand to the preferences.
#I'm not alway happy with the options but I changed the cut off multiple
times until I'm satisfied with the result.

#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?
#Answer:
#No, for example, all the result instances only have austin_lots$Education ==
"Moderate" which is not as good as "high".
```

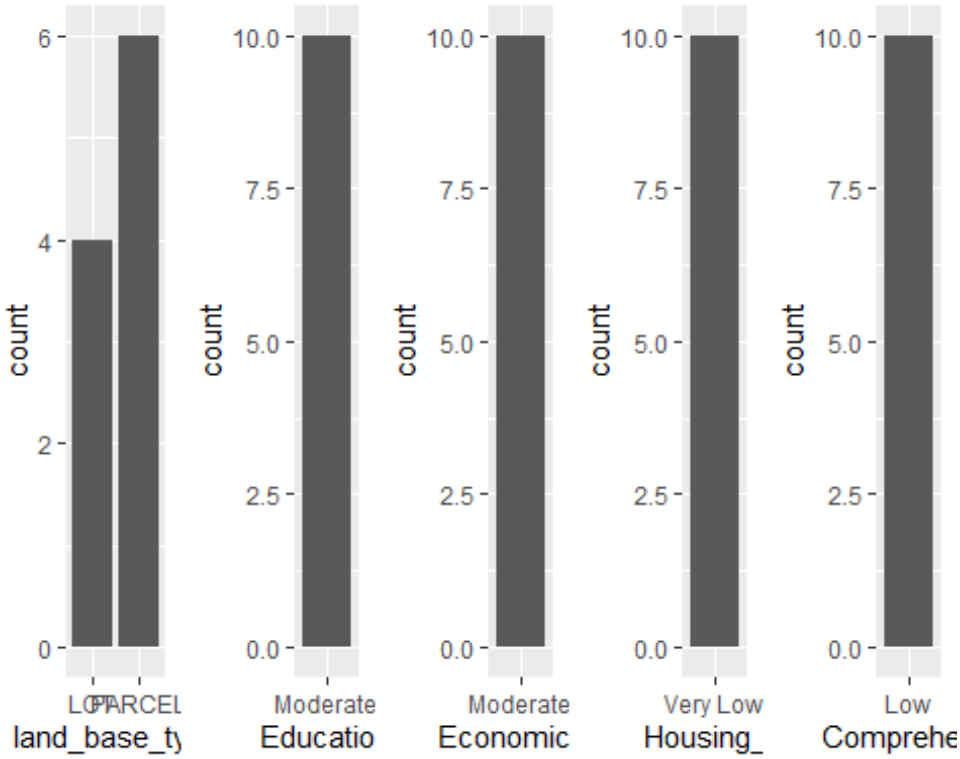
```
#I believe it's the trade offs that we have to decide on which preference is
of higher priority.
#I would like to know the percentage of busy traffic time in those areas,
because I believe it's very important to worker who need to drive to the
company.

###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.
#Answer:
austin_lots2 = read.csv("Austin_Lots.csv")
par(mfrow = c(2,4))
hist(austin_lots$EWC_dist)
hist(austin_lots$NSC_dist)
plot(austin_lots$Education)
plot(austin_lots$Housing__)
hist(austin_lots2$EWC_dist)
hist(austin_lots2$NSC_dist)
plot(austin_lots2$Education)
plot(austin_lots2$Housing__)
```






```
#12.2 Create a chart showing qualitative variables for each of the 10 final
parcels.
#Answer:
grid.arrange(ggplot(austin_lots)+geom_bar(aes(x=land_base_type)),
ggplot(austin_lots)+geom_bar(aes(x=Education)),
```

```
            ggplot(austin_lots)+geom_bar(aes(x=Economic__)),
ggplot(austin_lots)+geom_bar(aes(x=Housing__)),
            ggplot(austin_lots)+geom_bar(aes(x=Comprehens)),
            nrow = 1)
```



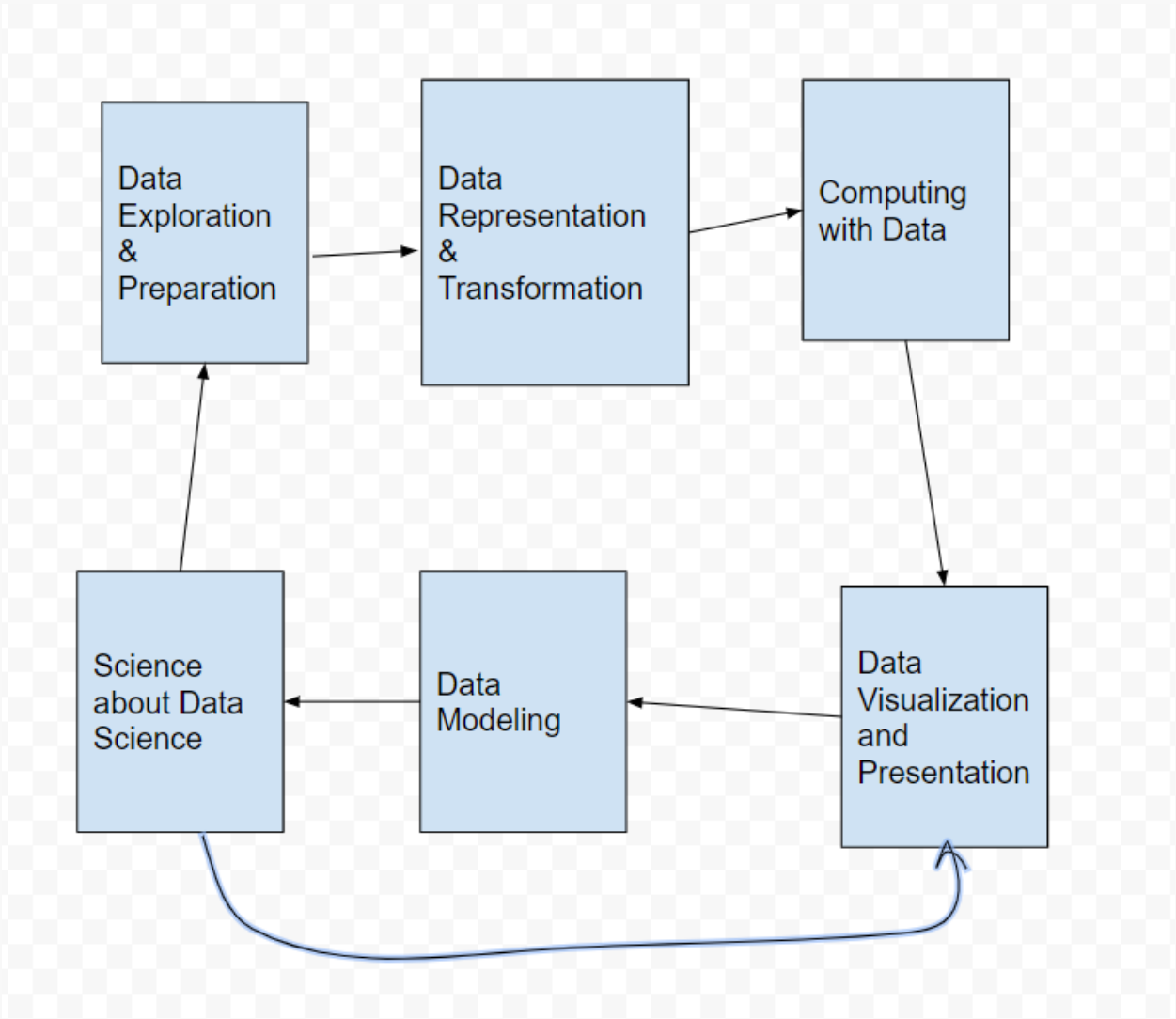

```
#Education, Economic__,Housing__,Comprehens are the same for all selected
parcels.


#12.3 For each of the 10 final parcels list their strengths and weaknesses.
If the parcels end up very similar to each other, propose a system to further
rank each parcel and back up your decision.
#Answer:
#AS shown above, qualitative variables for each of the 10 final parcels are
the same; thus, we will develop a system to further rank each parcel.
#From Question 8.4, we can assume emploees love work at a place near the city
center. Thus, we use City_dist the the metrics to rank the 10 parcels.
#The closer the distance to city center, the better.
rank(austin_lots$City_dist)
```

```
##  [1]  9  4  8  2  3  5  7  6 10  1
```

```
#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).
#Answer:
```

#I believe the busy traffic time, number of restaurants, number of gas stations shoulde be included in order tohelp make the decision.
#Because those factors are unlikely to be similar or tie with each other, and thus they are useful to different the parcels.

###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.
#Answer:



#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.
#Answer:
#1. Data Exploration and Preparation.

```
# This is where we explore and understand the data. We were giving the data
and the metadata.
#And it's our task to understand the purpose of the project(why), where the
data was from, what the data is about, and for whom. We did this by reading
the giving
#background info before we started, and we understand some basic info about
the data set in Q1.
#
#2. Data Representation and Transformation
# This is where we take care of data formats and the quality about our data.
We deal with missing values in Q2 and we further cleaned the data in Q3, Q8,
taking care about the formats.
# Also, in Q4 (change variable data class) and Q5 (eliminate blanks to NA for
pie charts), I further tranformed some data after further understanding about
the data.

#3. Computing with Data
#Data were processed in R here. In Q5, we calculates descriptive and
distributional statistics, try to understnad data set
#comprehensively about different colomns, recarding the properties of the
parcel.

#4. Data Visualization and Presentation
#Almost every project need data visualization, bacause it is the most effiect
way for ourselves to understand the data through graphs, and we can also
demostrate our research process better thru data visualization.
# We did this in Q6 Q7, where we compate many graph bewtween different
vaiables. In Q12, we demostrated our conclution thru visulizations.
#Also, a good Data Visualization should be repeatable.

#5. Data Modeling
# We did a traditinal way of modeling based on existing data, instead of the
predictive way (AI, machine learning). In Q8 andin Q9, we did
# the analysis and modeled based on parcels' property and employee
preferences, and we try to select the best site for theGlobalTechSync.

#6. Science about Data Science
#This is how we data scientists are doing data science as science. In Q11 and
Q12, we commented on the process and then made recommadation based on all
previous analysis.



#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?
#Answer:
# I believe generating graphs and show changes in data step by step in the
report will definitely help explain the workflow.
# One of advantages is that graphs are esay to understand by anyone even with
```

```
limited data science background.
# One of disadvantages will be the graphs generated will likely be biased to
the result, because it's created for the result.
#Using repeated envirnment such as Wholetale may also help the GlobalTechSync
location scouts check my methods.


#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?
#Answer:
#I will talk with the ones who created the dataset in order to undertstand
the limits and meaning of the coloumns. Also, I will try to avoid importing
the NA values into the dataset.
#I will suggues keep the documentation and metadata more detailed, and keep
the naming of the columns accurate and meaningful.

#Bonus: https://dashboard.wholetale.org/run/5dee8d1e7bf5ca3bf54ab06c
```