stat430_finalproject

STAT 430 Final Project

Course ID : 15

12/9/19

```r
library(plyr)
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::arrange()   masks plyr::arrange()
## x purrr::compact()   masks plyr::compact()
## x dplyr::count()     masks plyr::count()
## x dplyr::failwith()  masks plyr::failwith()
## x dplyr::filter()    masks stats::filter()
## x dplyr::id()        masks plyr::id()
## x dplyr::lag()       masks stats::lag()
## x dplyr::mutate()    masks plyr::mutate()
## x dplyr::rename()    masks plyr::rename()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()
```

```r
library(dplyr)
```

```r
austin = read.csv("~/Desktop/stat430_final/austin_lots.csv")
```

1.1 What are the initial dimensions of the dataset?

```r
dim(austin)
```

```
## [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?

Some of the columns that I think will be the least helpful include created_by, date_creat, modified_b, date_modif. ##### Why do you think these are less helpful? I dont think these are helpful because the information about the employee who created and updated the record does not affect the quality of the site. The employee should not affect it. The date the record was created and updated does not affect the overall quality of the site. The date is just a timestamp and the quality of the site should not be affected by the date the record was created or modified.

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

```
df = subset(austin, select = -c(created_by,date_creat,modified_b,date_modif) )
dim(df)
```

```
## [1] 26284    40
```

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

It helps us remove any unnecessary columns and focus on the data we need for our analysis. Subsetting the data allows us to work with a smaller version of the dataset which makes it easier to perform modeling, regression, and other analysis. Instead of trying to run these functions on a large dataset we can run them on a smaller dataset with variables that matter.

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.

```
names(df)[names(df) == "zoning_o_3"] <- "zoning"
names(df)[names(df) == "zcta5ce10"] <- "zip"
names(df)[names(df) == "Med_HH_Inc"] <- "med_income"
names(df)[names(df) == "Aff_rent_t"] <- "aff_rent"
names(df)[names(df) == "Aff_own_te"] <- "aff_own"

head(df,1)
```

```
##   FID block_id land_base_ land_base1 lot_id objectid City_dist Airpt_dist
## 1   0             1876887     PARCEL     14   356102   3208.66    10007.6
##   district Shape_Area zoning   zip LAND_USE_2 GENERAL_LA EWC_dist NSC_dist
## 1       14   7022.986        78704     143809        100  2395.46  5935.48
##   Mopac_dist X130_dist X35_dist ExTrail_1m PpTrail_1m conf bike_lanes Bus_area
## 1    3625.73   12690.6  1453.61          1         17    2         22        1
##   TotBdgArea Num_Bldgs MaxBdgArea tax_break2 bk_tx_brk    GEOID Housing__
## 1    238.829         3    179.237  9.5203304         0 4.85e+11  Very Low
##   Education Economic__ Comprehens med_income Med_rent Med_home aff_rent aff_own
## 1  Very Low   Moderate   Very Low      50248      940   338200       99      33
##
## 1 Change of use  Interior remodel from convenience store to cafï¿½ retail  Scope of work to include a
```

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

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

I would first test for missing values in each column using the functions above. I would then use the na.omit function to create a new dataset excluding the missing values. I chose this function because based on the test done above there is not that high of a percentage of missing data in the variables. Na.omit is a good enough function for this dataset. If the dataset had more missing values and more complicated missing values I would try to perform different tactics on dealing with missing values.

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

My choice for dealing with missing values may affect my later analysis because the na.omit function does not delete all the missing values. In question 2.1 I checked the sum of missing values which was 173 and based on the number of rows the new dimension should be 26111, but the number of rows is 26115 which shows that it doesn't work for every row.

2.4 Implement your methods for dealing with the missing values.

```r
ds <- na.omit(df)
```

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

```r
dim(ds)
```

```
## [1] 26115    40
```

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.

```r
# count the unique values of a variable or a set of variables
sapply(ds, function(x) length(unique(x)))
```

```
##        FID   block_id land_base_ land_base1     lot_id   objectid  City_dist
##      26114        195      26114         10        422      26114      25300
## Airpt_dist   district Shape_Area     zoning        zip LAND_USE_2 GENERAL_LA
##      23866          2      25954        173         15      12352         17
##   EWC_dist   NSC_dist Mopac_dist  X130_dist   X35_dist ExTrail_1m PpTrail_1m
##      25588      25553      25455      23033      25515         22         47
##       conf bike_lanes   Bus_area TotBdgArea  Num_Bldgs MaxBdgArea tax_break2
##          5         59          2      18203         58      14826        104
##   bk_tx_brk      GEOID  Housing__  Education Economic__ Comprehens med_income
##       6705          1          4          5          5          5         13
##   Med_rent   Med_home   aff_rent    aff_own Descriptio
##         13         13          7         11       1113
```

```r
# land_base1
unique(ds$land_base1)
```

```
##  [1] PARCEL LOT    Lot           TRACT  Parcel lott   PCL    Tract  OTHER
## Levels:   Lot LOT lott OTHER Parcel PARCEL PCL Tract TRACT
```

```r
levels(ds$land_base1)
```

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

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
ds$land_base1 <-revalue(ds$land_base1, c("PARCEL"="parcel", "Parcel"="parcel", "PCL"="parcel"))
ds$land_base1 <-revalue(ds$land_base1, c("Lot"="lot", "LOT"="lot", "lott"="lot"))
ds$land_base1 <-revalue(ds$land_base1, c("Tract"="tract", "TRACT"="tract"))
ds$land_base1 <-revalue(ds$land_base1, c("OTHER"="other"))

levels(ds$land_base1)
```

```
## [1] " "      "lot"    "other"  "parcel" "tract"
```

```r
unique(ds$land_base1)
```

```
## [1] parcel lot           tract  other
## Levels:   lot other parcel tract
```

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.

```r
#levels(mydata$tax_break2)
ds$tax_break2 <-revalue(ds$tax_break2, c("$0.98 "="0.98", "$3.11 "="3.11", "$7.73 "="7.73", "$9.52 "="9
#levels(mydata$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.

```r
options(scipen=999)
unique(ds$GEOID)
```

```
## [1] 485000000000
```

```r
new_data = subset(ds, select = -c(GEOID) )
```

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
which(duplicated(new_data))
```

```
## [1] 378
```

```r
newdata <-new_data[-c(378),]
dim(newdata)
```

```
## [1] 26114    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.

Metadata is needed to keep the records and prevent inconsistencies in data, metadata allows data to be used and valued in the long term.

```r
mydata_final <-subset(newdata, select=-c(LAND_USE_2))
```

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

Data cleaning is needed because it helps imperative the quality of the data we are working on which in turn helps improve the overall productivity of the analysis. We could always add the columns back into the data frame.

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).

Another step that could be implemented is cleaning out data that would be considered outliers and standardizing the data into a singular format.

4.1 Please display the initial variable classes for each column

```r
str(mydata_final)
```

```
## 'data.frame':    26114 obs. of  38 variables:
##  $ FID       : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ block_id  : Factor w/ 195 levels " ","1","10","100",..: 1 1 1 165 1 1 1 1 1 1 ...
##  $ land_base_: int  1876887 1676746 1839096 1909677 1650609 1647428 1880381 1741600 1726221 1659892
##  $ land_base1: Factor w/ 5 levels " ","lot","other",..: 4 2 2 2 4 4 4 4 4 2 ...
##  $ lot_id    : Factor w/ 426 levels " ","0","1","1 1/2",..: 82 110 309 99 1 1 1 1 1 357 ...
##  $ objectid  : int  356102 296037 319082 333367 270888 160741 266031 344624 318570 147975 ...
##  $ City_dist : num  3209 3203 3188 3090 6319 ...
##  $ Airpt_dist: num  10008 12721 12793 12715 4681 ...
##  $ district  : int  14 14 14 14 14 14 14 14 14 14 ...
##  $ Shape_Area: num  7023 7717 18297 5605 63141 ...
##  $ zoning    : Factor w/ 174 levels " ","AV","B-H",..: 1 134 134 121 76 76 80 165 134 165 ...
##  $ zip       : int  78704 78705 78705 78705 78742 78742 78742 78702 78702 78702 ...
##  $ GENERAL_LA: int  100 100 100 100 300 500 900 500 200 400 ...
##  $ EWC_dist  : num  2395 8688 8641 8548 824 ...
```

```
##  $ NSC_dist  : num  5935.5 5656.7 5769.8 5790.1 90.8 ...
##  $ Mopac_dist: num  3626 3321 3188 3161 9799 ...
##  $ X130_dist : num  12691 12050 12163 12142 5659 ...
##  $ X35_dist  : num  1454 311 425 402 5295 ...
##  $ ExTrail_1m: int  1 0 0 0 0 0 0 4 4 2 ...
##  $ PpTrail_1m: int  17 5 6 6 8 9 10 25 24 13 ...
##  $ conf      : int  2 2 2 2 3 0 4 2 1 1 ...
##  $ bike_lanes: int  22 9 18 13 14 0 9 19 18 28 ...
##  $ Bus_area  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ TotBdgArea: num  239 137 295 138 506 ...
##  $ Num_Bldgs : int  3 3 1 1 6 2 0 2 5 1 ...
##  $ MaxBdgArea: num  179 123 295 138 249 ...
##  $ tax_break2: Factor w/ 104 levels "0.98","3.11",..: 99 15 15 15 5 5 5 5 5 5 ...
##  $ bk_tx_brk : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Housing__ : Factor w/ 6 levels ""," ","Low","Moderate",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ Education : Factor w/ 7 levels ""," ","High",..: 7 3 3 3 7 7 7 4 4 4 ...
##  $ Economic__: Factor w/ 7 levels ""," ","High",..: 5 6 6 6 4 4 4 4 4 4 ...
##  $ Comprehens: Factor w/ 7 levels ""," ","High",..: 7 5 5 5 7 7 7 7 7 7 ...
##  $ med_income: int  50248 11917 11917 11917 34076 34076 34076 34734 34734 34734 ...
##  $ Med_rent  : int  940 1088 1088 1088 639 639 639 766 766 766 ...
##  $ Med_home  : int  338200 292500 292500 292500 54400 54400 54400 175400 175400 175400 ...
##  $ aff_rent  : int  99 94 94 94 100 100 100 99 99 99 ...
##  $ aff_own   : int  33 79 79 79 100 100 100 67 67 67 ...
##  $ Descriptio: Factor w/ 1121 levels "","100 amp service on cell tower pole small cell",..: 99 798 79
```

```r
lapply(mydata_final, class)
```

```
## $FID
## [1] "integer"
##
## $block_id
## [1] "factor"
##
## $land_base_
## [1] "integer"
##
## $land_base1
## [1] "factor"
##
## $lot_id
## [1] "factor"
##
## $objectid
## [1] "integer"
##
## $City_dist
## [1] "numeric"
##
## $Airpt_dist
## [1] "numeric"
##
## $district
## [1] "integer"
##
```

```
## $Shape_Area
## [1] "numeric"
##
## $zoning
## [1] "factor"
##
## $zip
## [1] "integer"
##
## $GENERAL_LA
## [1] "integer"
##
## $EWC_dist
## [1] "numeric"
##
## $NSC_dist
## [1] "numeric"
##
## $Mopac_dist
## [1] "numeric"
##
## $X130_dist
## [1] "numeric"
##
## $X35_dist
## [1] "numeric"
##
## $ExTrail_1m
## [1] "integer"
##
## $PpTrail_1m
## [1] "integer"
##
## $conf
## [1] "integer"
##
## $bike_lanes
## [1] "integer"
##
## $Bus_area
## [1] "integer"
##
## $TotBdgArea
## [1] "numeric"
##
## $Num_Bldgs
## [1] "integer"
##
## $MaxBdgArea
## [1] "numeric"
##
## $tax_break2
## [1] "factor"
##
```

```
## $bk_tx_brk
## [1] "numeric"
##
## $Housing__
## [1] "factor"
##
## $Education
## [1] "factor"
##
## $Economic__
## [1] "factor"
##
## $Comprehens
## [1] "factor"
##
## $med_income
## [1] "integer"
##
## $Med_rent
## [1] "integer"
##
## $Med_home
## [1] "integer"
##
## $aff_rent
## [1] "integer"
##
## $aff_own
## [1] "integer"
##
## $Descriptio
## [1] "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

I think tax_break2 should be a numeric class but it is a factor that does not make sense since tax_break2 is a percentage of parcel purchase cost waived.

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

```r
levels(mydata_final$tax_break2)
```

```
##   [1] "0.98"      "3.11"      "7.73"      "9.52"      "0"         "0.0213235"
##   [7] "0.122498"  "0.24291"   "0.255072"  "0.256544"  "0.364723"  "0.632385"
##  [13] "0.67714"   "0.785669"  "0.983745"  "1.14823"   "1.2197"    "1.2659"
##  [19] "1.50366"   "1.51293"   "1.64689"   "1.7621"    "2.00828"   "2.03444"
##  [25] "2.0594499" "2.20244"   "2.2168901" "2.27542"   "2.2965901" "2.3626299"
##  [31] "2.4546101" "2.4780099" "2.49894"   "2.5534999" "2.56007"   "2.72223"
##  [37] "2.7440901" "2.75211"   "2.8912899" "2.91031"   "3.0353301" "3.11113"
##  [43] "3.2132199" "3.3429899" "3.5039101" "3.58902"   "3.6100199" "3.8810999"
##  [49] "3.9456401" "4.0345502" "4.04175"   "4.0861602" "4.1902199" "4.3583498"
```

```
##  [55] "4.4771199" "4.75669"   "4.8830099" "4.8898802" "5.0180602" "5.0945802"
##  [61] "5.1057501" "5.2039299" "5.2713199" "5.3537402" "5.4383998" "5.45436"
##  [67] "5.7270298" "6.0208702" "6.08325"   "6.1143799" "6.1303601" "6.1350899"
##  [73] "6.1810398" "6.18437"   "6.37288"   "6.8348198" "6.9305401" "7.1226602"
##  [79] "7.1391201" "7.3952098" "7.4348001" "7.7320199" "7.7928801" "7.8884702"
##  [85] "8.1156397" "8.2697401" "8.27598"   "8.5271997" "8.5521202" "8.5772896"
##  [91] "8.5944996" "8.6483297" "8.88099"   "9.1087303" "9.1929197" "9.2534199"
##  [97] "9.4080095" "9.5174198" "9.5203304" "9.6627502" "9.6882296" "9.7939796"
## [103] "9.95261"   "9.9884796"
```

```
# I think this variable should be numeric and not a factor.
mydata_final$tax_break2 =as.numeric(mydata_final$tax_break2)
```

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?

Sometimes data could be imported in a way that is not useful for further analysis. Sometimes numeric values could be imported as a factor. This step is useful because it puts the data into a useful class. Without this step, we would not know what to do with variables that were in the wrong class.

Part 2: Data Exploration

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(mydata_final)
```

```
##       FID           block_id       land_base_         land_base1
##  Min.   :    0          :12777   Min.   :  1635655         :  120
##  1st Qu.: 6528   A       : 1295   1st Qu.:  1712075   lot    :23766
##  Median :13056   1       : 1076   Median :  1788348   other :    2
##  Mean   :13063   3       :  934   Mean   : 28715871   parcel: 2170
##  3rd Qu.:19585   2       :  912   3rd Qu.:  1863451   tract :   56
##  Max.   :26277   B       :  895   Max.   :400842667
##                  (Other): 8225
##     lot_id          objectid        City_dist       Airpt_dist
##         : 4428   Min.   :     3   Min.   :    0   Min.   :   31.63
##  1      : 1831   1st Qu.: 93173   1st Qu.: 1793   1st Qu.: 7409.22
##  2      : 1644   Median :186048   Median : 2661   Median : 9800.97
##  3      : 1425   Mean   :186440   Mean   : 3350   Mean   : 9013.64
##  4      : 1319   3rd Qu.:279242   3rd Qu.: 4062   3rd Qu.:11236.60
##  5      : 1232   Max.   :375410   Max.   :13573   Max.   :13745.40
##  (Other):14235
##     district       Shape_Area           zoning           zip
##  Min.   :14.00   Min.   :      19   NP       :15105   Min.   :78617
##  1st Qu.:14.00   1st Qu.:    5705            : 6015   1st Qu.:78702
##  Median :14.00   Median :    6927   UNO      :  587   Median :78704
##  Mean   :15.19   Mean   :   31494   TOD      :  563   Mean   :78712
##  3rd Qu.:14.00   3rd Qu.:    9256   AV       :  469   3rd Qu.:78722
##  Max.   :21.00   Max.   :27533199   SF-4A-NP:  279   Max.   :78746
##                                     (Other) : 3096
```

```
##    GENERAL_LA       EWC_dist       NSC_dist          Mopac_dist      
##   Min.   :  0.0   Min.   :   0   Min.   :   6.676   Min.   :   11.15  
##   1st Qu.:100.0   1st Qu.:2587   1st Qu.:2440.747   1st Qu.: 3178.74  
##   Median :100.0   Median :4666   Median :4138.185   Median : 4506.77  
##   Mean   :303.7   Mean   :4357   Mean   :3967.698   Mean   : 5308.79  
##   3rd Qu.:400.0   3rd Qu.:6020   3rd Qu.:5417.370   3rd Qu.: 6581.71  
##   Max.   :940.0   Max.   :8726   Max.   :7786.030   Max.   :16494.80  
##                                                                       
##    X130_dist          X35_dist          ExTrail_1m       PpTrail_1m   
##   Min.   :   54.03   Min.   :   17.92   Min.   : 0.000   Min.   : 0.0  
##   1st Qu.: 8899.83   1st Qu.:  787.93   1st Qu.: 0.000   1st Qu.: 7.0  
##   Median :10737.90   Median : 1687.34   Median : 1.000   Median :14.0  
##   Mean   :10220.35   Mean   : 2247.10   Mean   : 2.957   Mean   :15.3  
##   3rd Qu.:12268.23   3rd Qu.: 2789.59   3rd Qu.: 4.000   3rd Qu.:24.0  
##   Max.   :14789.30   Max.   :11544.00   Max.   :21.000   Max.   :47.0  
##                                                                        
##       conf        bike_lanes       Bus_area        TotBdgArea     
##   Min.   :0.00   Min.   : 0.00   Min.   :0.0000   Min.   :    0.0  
##   1st Qu.:1.00   1st Qu.:11.00   1st Qu.:1.0000   1st Qu.:  109.6  
##   Median :2.00   Median :15.00   Median :1.0000   Median :  219.0  
##   Mean   :1.68   Mean   :15.08   Mean   :0.9784   Mean   :  840.2  
##   3rd Qu.:2.00   3rd Qu.:19.00   3rd Qu.:1.0000   3rd Qu.:  402.2  
##   Max.   :4.00   Max.   :64.00   Max.   :1.0000   Max.   :64263.8  
##                                                                   
##    Num_Bldgs         MaxBdgArea         tax_break2       bk_tx_brk       
##   Min.   :  0.000   Min.   :    0.00   Min.   :  1.00   Min.   :0.000000  
##   1st Qu.:  1.000   1st Qu.:   93.39   1st Qu.:  5.00   1st Qu.:0.000000  
##   Median :  1.000   Median :  163.65   Median : 29.00   Median :0.000000  
##   Mean   :  1.753   Mean   :  710.43   Mean   : 38.05   Mean   :0.012954  
##   3rd Qu.:  2.000   3rd Qu.:  269.19   3rd Qu.: 67.00   3rd Qu.:0.003374  
##   Max.   :114.000   Max.   :47366.60   Max.   :104.00   Max.   :0.099999  
##                                                                          
##      Housing__         Education         Economic__       Comprehens   
##            :    0            :    0            :   0            :    0  
##            :    0            :    0            :   0            :    0  
##   Low      : 5478   High     : 3114   High     :5554   High     : 4162  
##   Moderate :  698   Low      : 5705   Low      :3817   Low      : 5744  
##   Very High:    2   Moderate : 3724   Moderate :5681   Moderate : 2966  
##   Very Low :19936   Very High: 1979   Very High:9594   Very High: 1168  
##                     Very Low :11592   Very Low :1468   Very Low :12074  
##    med_income        Med_rent         Med_home          aff_rent     
##   Min.   :     0   Min.   :   0.0   Min.   :     0   Min.   :  0.00  
##   1st Qu.: 34734   1st Qu.: 766.0   1st Qu.:120200   1st Qu.: 97.00  
##   Median : 34734   Median : 835.0   Median :175400   Median : 99.00  
##   Mean   : 41285   Mean   : 925.8   Mean   :229198   Mean   : 95.58  
##   3rd Qu.: 50248   3rd Qu.: 946.0   3rd Qu.:338200   3rd Qu.: 99.00  
##   Max.   :125327   Max.   :1590.0   Max.   :621900   Max.   :100.00  
##                                                                      
##     aff_own   
##   Min.   :  0  
##   1st Qu.: 33  
##   Median : 67  
##   Mean   : 59  
##   3rd Qu.: 79  
```

```
##  Max.   :100
##
##                                                          Descriptio
##  new solar installation for new residence                    :  630
##  Adding equipment to existing wireless telecommunication tower:  301
##  total demo of small medical office and gymnasium            :  296
##  interior remodel to existing AISD school                    :  290
##  total demo of church                                        :  285
##  Installation of new 200A service  for Athletic Field Lighting:  279
##  (Other)                                                     :24033
```

```r
colnames(mydata_final) <- tolower(colnames(mydata_final)) #decided to make them lower case for easier a
typeof(mydata_final)
```

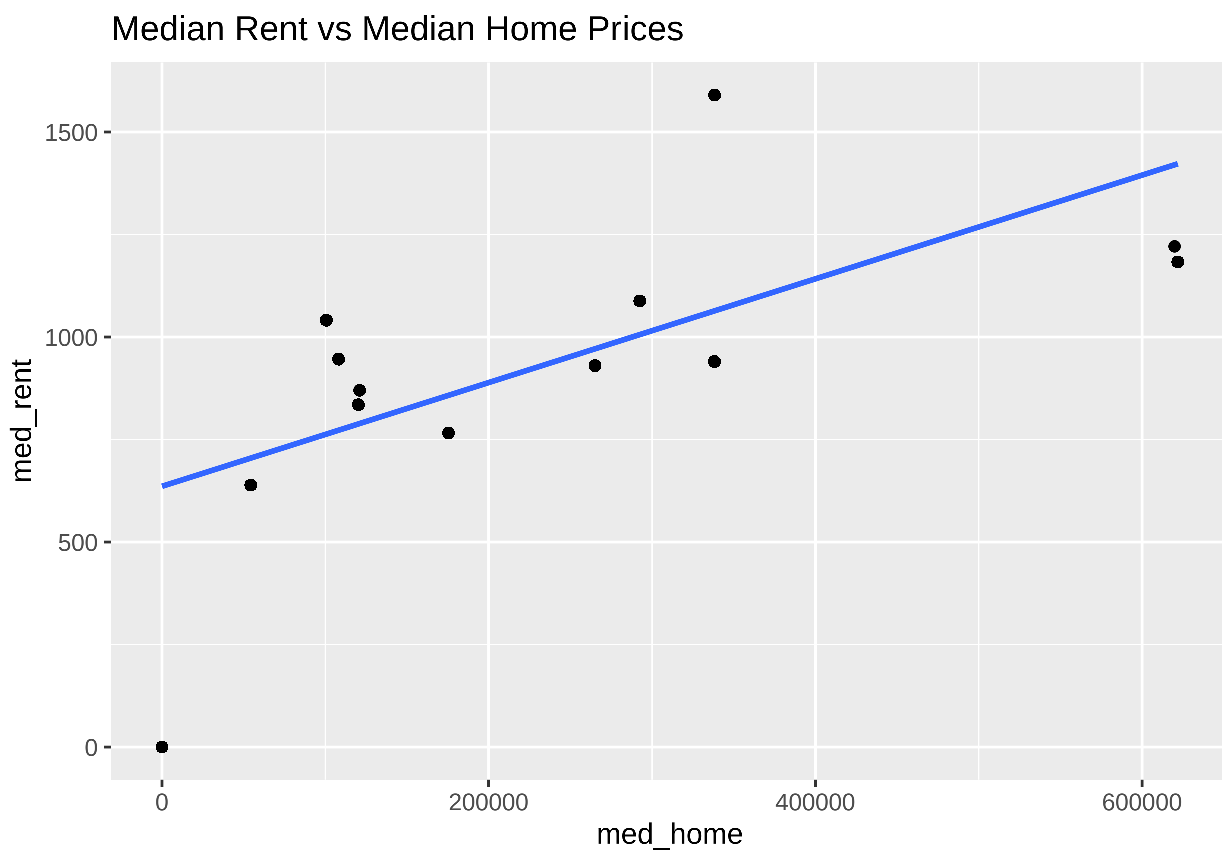
```
## [1] "list"
```

5.2 Describe anything you find that is unexpected or interesting.

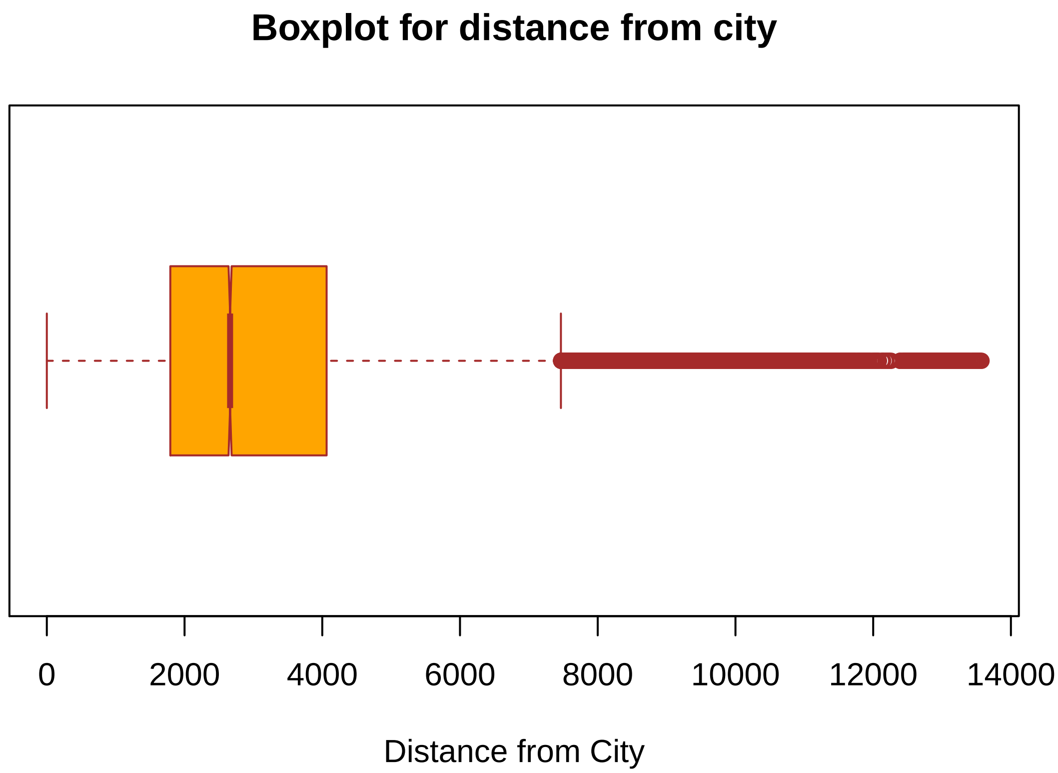
Some interesting points found in the descriptive statistics show that land_base_ has a large range of data which may mean that it has many outliers in the dataset. The max value is very large compared to the mean. City_dist also seems to have values that are far away from the mean with the max value being 13,573 and the mean being 3350. The airpt_dist variables also seem to contain many outliers. Mopac_dist and x35_dist seem to have a large range of values as well.

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

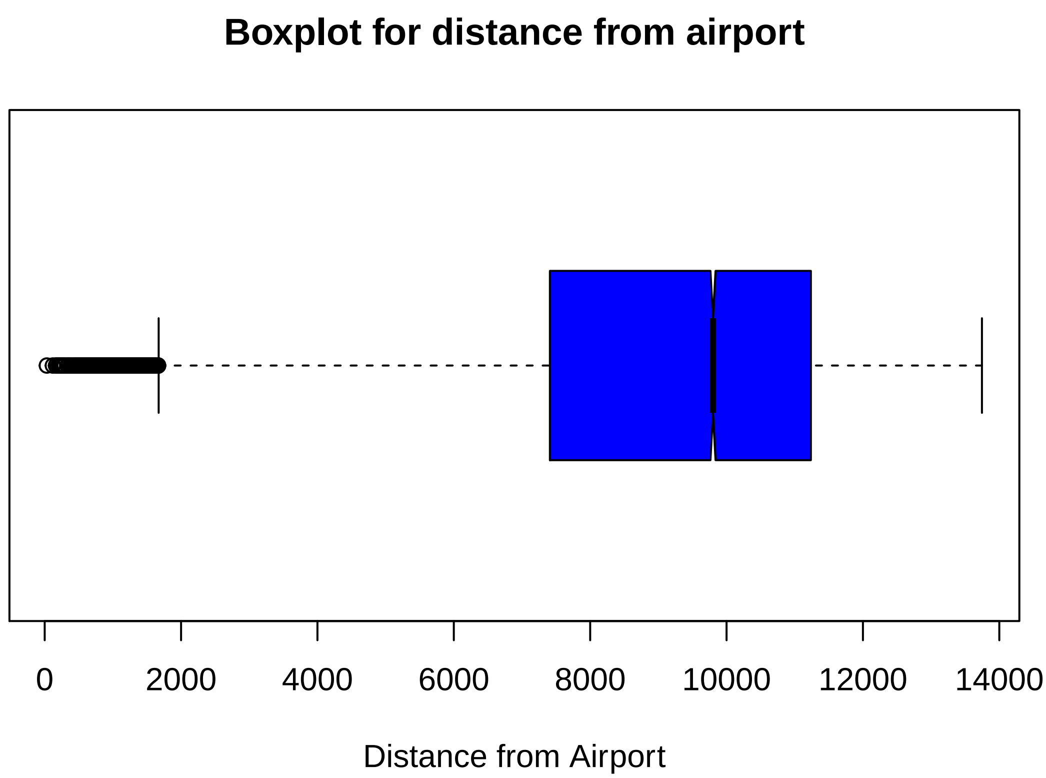
```r
library(ggplot2)
g <- ggplot(mydata_final, aes(med_home, med_rent))
# Scatterplot
g + geom_point() +
  geom_smooth(method="lm", se=F) +
  labs(title="Median Rent vs Median Home Prices",
       y="med_rent",
       x="med_home" )
```



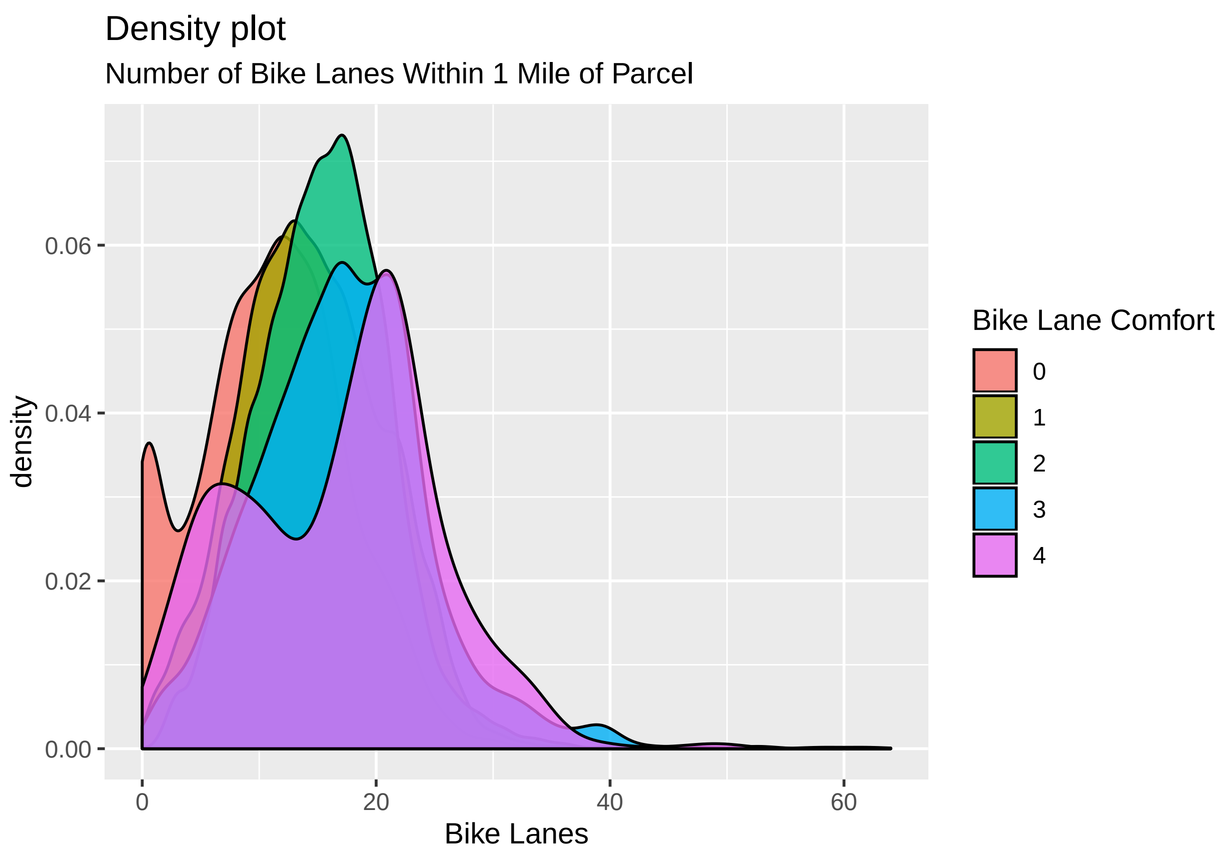

```r
boxplot(mydata_final$city_dist,
main = "Boxplot for distance from city",
xlab = "Distance from City",
ylab = " ",
col = "orange",
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```



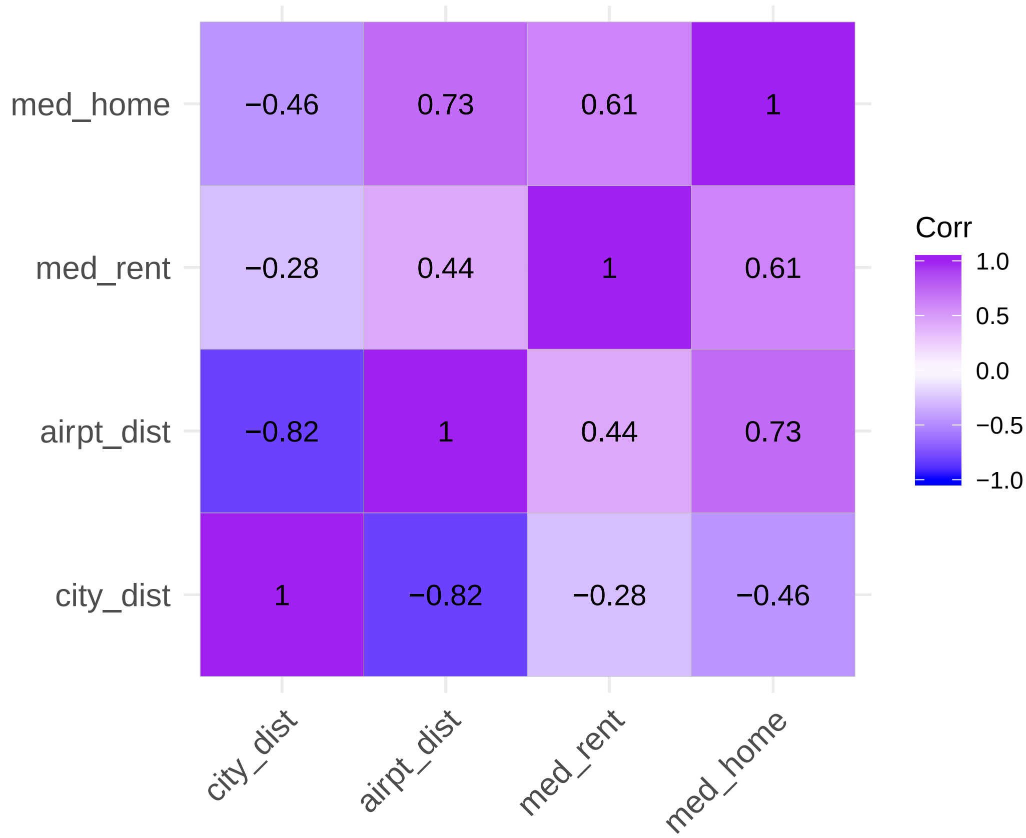

```r
boxplot(mydata_final$airpt_dist,
main = "Boxplot for distance from airport",
xlab = "Distance from Airport",
ylab = " ",
col = "blue",
border = "black",
horizontal = TRUE,
notch = TRUE
)
```




```
g <- ggplot(mydata_final, aes(bike_lanes))
g + geom_density(aes(fill=factor(conf)), alpha=0.8) +
    labs(title="Density plot",
         subtitle="Number of Bike Lanes Within 1 Mile of Parcel",
         x="Bike Lanes",
         fill="Bike Lane Comfort")
```




```
library(ggcorrplot)
corr_dist_price = cor(select(mydata_final, city_dist,airpt_dist, med_rent, med_home))
ggcorrplot(corr_dist_price,lab=TRUE, colors = c("blue", "white", "purple"))
```



6.2 Compare different graph types to see which ones best convey trends, outliers, and patterns in the data.
I believe the best trends where represented in the scatterplot, box plots, and correlation matrix. ##### 6.3 Describe what you find from the graphs. The first graph showed the linear relationship between the median rent and median home prices in the area. I wanted to see if these values had a linear relationship. Then I made two boxplots for the distance from the city and the distance from the airport. The boxplot for distance from the city showed that the data had a median in 2000 which many outliers. The boxplot for the distance from the airport showed that the median was around 10,000 with a couple of outliers as well. The correlation matrix was very interesting I wanted to see if the distance from the airport and the distance from the city were correlated with the median home price and median rent. I found that distance from the city has a slight negative correlation with the median rent and a negative correlation with median home prices. However, the distance from the airport seems to have a positive correlation with median rent and a strong positive correlation with the median home price.

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.

```
summary(mydata_final$conf)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.00    1.00    2.00    1.68    2.00    4.00
```

```r
#head(mydata_final$bike_lanes)
cor(mydata_final$bike_lanes,mydata_final$conf)
```

```
## [1] 0.212934
```

Both the conf variable and the bike lane variables tell us how easy it is to travel by bike. The conf lane is a factor that indicates the average bike lane comfort level(0 is the most comfortable while 4 is the least comfortable). The bike_lanes variable is the number of bike lanes within a 1 mile of the parcel. To see the relationship between the number of bike lanes and comfort level, I found the correlation between the two variables. It looks like the two variables have a slight positive correlation of .2129. ##### 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).

Are the number of trails(ExTrail_1m) near the parcel and the number of bike lanes related in some way?

```r
cor(mydata_final$bike_lanes,mydata_final$extrail_1m)
```

```
## [1] 0.1127457
```

Does distance from the highway and the number of bike lanes have a potential relationship? I will create a correlation matrix that shows the relationship between highway distance and num ber of bike lanes.

```r
corr_dist_price = cor(select(mydata_final, ewc_dist,nsc_dist, mopac_dist, x130_dist, x35_dist,bike_lane
ggcorrplot(corr_dist_price,lab=TRUE, colors = c("pink", "white", "purple"))
```

This correlation matrix shows that bike lanes have a positive relationship with distance from East-West

Connector highway, North-South Connector Highway and highway 130 but have a negative relationship between Mopac freeway and Interstate 35. None of the correlations are very strong.

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

Converting all the letters in the column to the lower case helps us perform further analysis of the words in the descriptio column. Working with just lower case letters makes text mining and information retrieval easier. We might lose significance if capital letters were being used to make a point or emphasize something. However, I do not believe much is lost by converting all the letters to lower case.

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.

```
library(tm)
```

```
## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
description <- VectorSource(mydata_final$descriptio)
descrip_corpus <- VCorpus(description)
descrip_corpus
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 26114
```

```
all_stop_words <- c("a", "about", "across", "after","all",  "almost",   "also", "am",   "among","an","a
```

```
descrip_corpus <- tm_map(descrip_corpus, content_transformer(tolower))
descrip_corpus <- tm_map(descrip_corpus, removeWords, all_stop_words)
descrip_corpus <- tm_map(descrip_corpus, removePunctuation)
descrip_corpus <- tm_map(descrip_corpus, PlainTextDocument)
descrip_corpus <- tm_map(descrip_corpus, removeNumbers)




library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(SnowballC)

wordcloud(descrip_corpus, max.words = 10, random.order = FALSE)
```

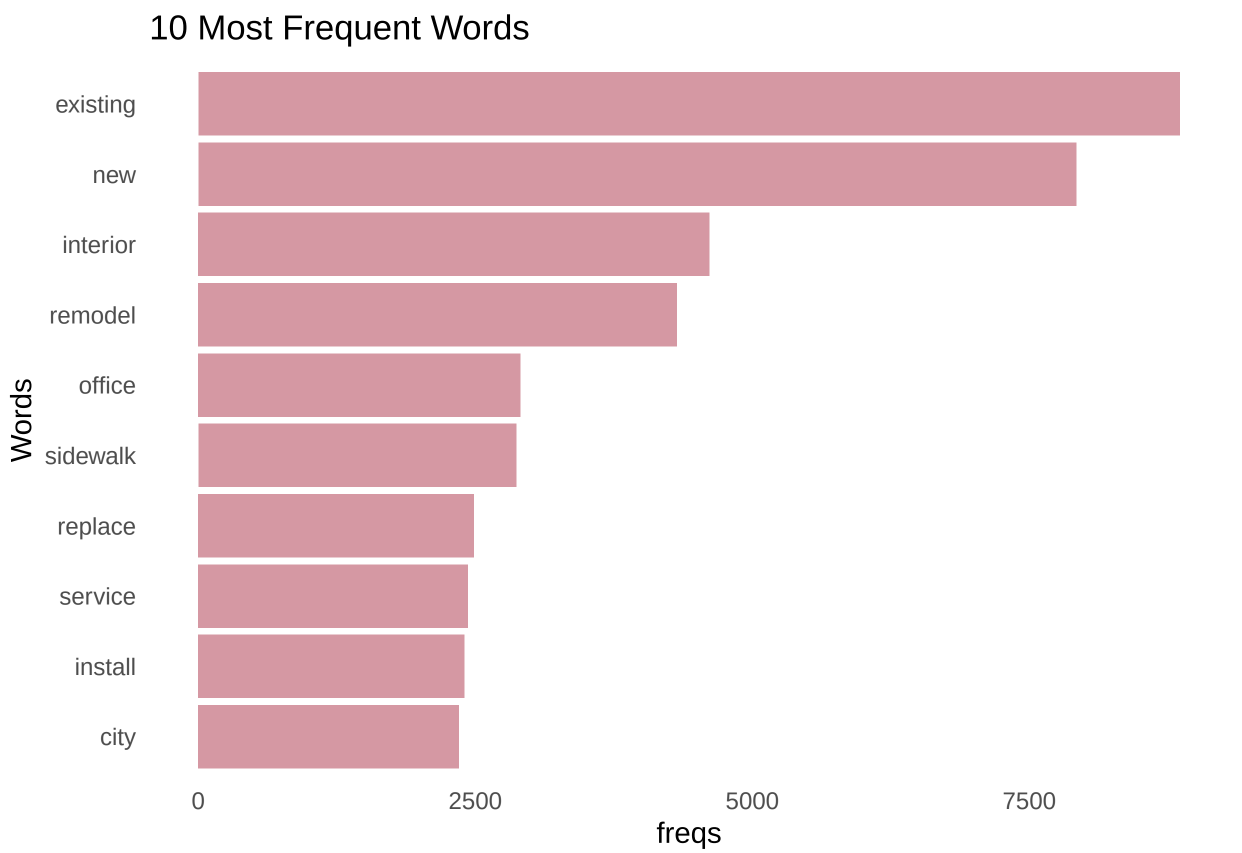
8.3 Preform a similar function to 8.2 but this time finding unique words and their frequency. What are the 10 most frequent non stop words, i.e. which are frequent words that give you meaningful information about the type of construction occurring? How can these help you finding a good site for GlobalTechSync?

```
tdm<-TermDocumentMatrix(descrip_corpus, control=list(weighting=weightTf))
tdm.desc <- as.matrix(tdm)
sfq <- data.frame(words=names(sort(rowSums(tdm.desc),decreasing = TRUE)), freqs=sort(rowSums(tdm.desc),
head(sfq,10)
```

```
##        words freqs
## 1   existing  8858
## 2        new  7924
## 3   interior  4613
## 4    remodel  4319
## 5     office  2907
## 6   sidewalk  2870
## 7    replace  2487
## 8    service  2435
## 9    install  2402
## 10      city  2352
```

```
ggplot(sfq[1:10,], mapping = aes(x = reorder(words, freqs), y = freqs)) +
  geom_bar(stat= "identity", fill="#d598a3") +
```

```
  coord_flip() +
  scale_colour_hue() +
  labs(x= "Words", title = "10 Most Frequent Words") +
  theme(panel.background = element_blank(), axis.ticks.x = element_blank(),axis.ticks.y = element_blank
```



It shows that the most frequent word is existing followed by new, interior, remodel, office, sidewalk, replace, service, install and city. I think this shows that maybe eisting construction is important to consider when choosing a good site.

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?

I believe words like city, remodel, interior and new would be the most useful to determine places to attract tech workers. I believe that for the working environment, tech workers tend to look for location, activities around, the interior design and how new/clean the building would be. I think the word interior would be the best high-frequency word to determine places to attract tech workers. The interior of the building is very important since that is the place most workers would spend their time. a low-frequency word that tech employees probably do not care about is install/replace. These are construction terms that are not necessarily appealing to tech employees,

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.

I think possibly stemming the words so there are not repetitive types of words like install and installing. I think this process would be useful so we could eliminate a lot of repetitive words from the corpus. This would make the analysis easier.

Part 3: Site Selection

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.

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

```
mydata_final<-mydata_final[!(mydata_final$bus_area=="0"),]
```

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

```
mydata_final<-mydata_final[!(mydata_final$shape_area < 300.00),]
```

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

```
mydata_final<-mydata_final[!(mydata_final$zoning == "LA" & mydata_final$zoning == "RR" & mydata_final$zo
```

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

```
dim(mydata_final)
```

```
## [1] 25411    38
```

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.

```
#1. An  undeveloped site    is  preferred
mydata_final$rank1 <-with(mydata_final, ifelse(mydata_final$general_la == 900, 1, 0))

#2. Ease    of access   to  a   major   interstate  or  highway is  preferred.
mean_ewc_dist = mean(mydata_final$ewc_dist)
mydata_final$rank20 <-with(mydata_final, ifelse(mydata_final$ewc_dist < mean_ewc_dist, .2, 0))
mean_nsc_dist = mean(mydata_final$nsc_dist)
mydata_final$rank21 <-with(mydata_final, ifelse(mydata_final$nsc_dist < mean_nsc_dist, .2, 0))
mean_mopac_dist = mean(mydata_final$mopac_dist)
mydata_final$rank22 <-with(mydata_final, ifelse(mydata_final$mopac_dist < mean_mopac_dist, .2, 0))
mean_130_dist = mean(mydata_final$x130_dist)
mydata_final$rank23 <-with(mydata_final, ifelse(mydata_final$x130_dist < mean_130_dist, .2, 0))
```

```
mean_35_dist = mean(mydata_final$x35_dist)
mydata_final$rank24 <-with(mydata_final, ifelse(mydata_final$x35_dist < mean_35_dist, .2, 0))

#3. Easy    access  to  the site    by  bike    or  foot    is  preferred
mean_bike = mean(mydata_final$bike_lanes)
mydata_final$rank30 <-with(mydata_final, ifelse(mydata_final$bike_lanes > mean_bike, .5, 0))
mean_trail = mean(mydata_final$extrail_1m)
mydata_final$rank31 <-with(mydata_final, ifelse(mydata_final$extrail_1m > mean_trail, .5, 0))

#4. Close   access  to  green   spaces  and areas   that    offer   opportunities   for employee    enr
mean_city = mean(mydata_final$city_dist)
mydata_final$rank4 <-with(mydata_final, ifelse(mydata_final$city_dist < mean_city, 1, 0))

#5. Higher  tax breaks or   discounts at    both    the district    and block   levels  is  preferred.
mean_tax = mean(mydata_final$tax_break2)
mydata_final$rank5 <-with(mydata_final, ifelse(mydata_final$tax_break2 > mean_tax, 1, 0))

#6. High    education   opportunity in  the area and    strong  nearby  university  systems are preferr
mydata_final = transform(mydata_final, education = factor(education,
       levels = c("Very Low", "Low", "Moderate", "High", "Very High"),
       labels = c(1, 2, 3, 4, 5)))

mydata_final$education =as.numeric(mydata_final$education)

mydata_final$rank6 <-with(mydata_final, ifelse(mydata_final$education > 3, 1, 0))

#7. Ability for tech    workers to  own their   own houses  is preferred.
mean_aff = mean(mydata_final$aff_own)
mydata_final$rank7 <-with(mydata_final, ifelse(mydata_final$aff_own > mean_aff, 1, 0))

#8. Fast    reliable    internet    needs   to  be  easily  accessible  at  the site.
mydata_final = transform(mydata_final, comprehens = factor(comprehens,
       levels = c("Very Low", "Low", "Moderate", "High", "Very High"),
       labels = c(1, 2, 3, 4, 5)))

mydata_final$comprehens =as.numeric(mydata_final$comprehens)
mydata_final$rank8 <-with(mydata_final, ifelse(mydata_final$comprehens > 3, 1, 0))

#9. Nearby  active  construction    of  office  type    structures  is  preferred.
descriptio_existing = str_detect(mydata_final$descriptio, "existing")
mydata_final$rank9 <-with(mydata_final, ifelse(descriptio_existing == TRUE, 1, 0))
```

```
mydata_final$rank10 <- mydata_final$rank1 + mydata_final$rank20 + mydata_final$rank21 +mydata_final$ranl
```

I created a column for rankings for each preference.

1. An undeveloped site is preferred: –I assigned 1 if the general_la value was equal to 900 and a 0 if it was not.
2. Ease of access to a major interstate or highway is preferred.: –I calculate the mean distance for each highway and gave it a value of .2 if the distance was lower than the mean and 0 if it was higher. .2 was used as a rank since there were five major highways under the same preference.
3. Easy access to the site by bike or foot is preferred. : –For this value, I calculated the mean of bike_lanes and existing trails and if the value was greater than the mean it was ranked with .5 if it was less than

the mean it was given a 0. .5 was used since there were two variables under the same preference.

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.: –I used city distance for this ranking if the distance to the city center was lower than the mean the rank was 1 if it was higher it was given a 0.
5. Higher tax breaks or discounts at both the district and block levels is preferred. : –If the tax breaks were higher than the mean I assigned the rank value of 1 but if it was lower than the mean it got a value of 0.
6. High education opportunity in the area and strong nearby university systems are preferred.: –First, I had to convert the factors to a numeric value. Very low =1, low=2,moderate=3,high=4, very high=5. I then said that if the value was above a 3 it would be ranked with a 1 if it was below a 3 it would get a 0.
7. Ability for tech workers to own their own houses is preferred. : –If the average affordable own percentage was higher than the mean it was assigned a 1 if it was lower it was assigned a 0.
8. Fast reliable internet needs to be easily accessible at the site.: –For this, I used the comprehens variable. The comprehensive opportunity index shows the overall opportunity. First, I had to convert the factors to a numeric value. Very low =1, low=2,moderate=3,high=4, very high=5. I then said that if the value was above a 3 it would be ranked with a 1 if it was below a 3 it would get a 0.
9. Nearby active construction of office type structures is preferred.: –For this, I looked at the descriptio column and looked for word "existing". If the description contained the word "existing" I assigned it a 1 if it did not I assigned it a 0.

I then summed these ranks and sorted them from highest to lowest taking the top 10 ranks to determine the best parcels.

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

```r
df <-mydata_final[order(-mydata_final$rank10),]
head(df,10)
```

```
##         fid block_id land_base_ land_base1 lot_id objectid city_dist airpt_dist
## 17068 17065        A    1942209        lot      1   183140   2659.88    13105.6
## 19146 19143             2001479        lot      3   331123   3128.74    13596.5
## 11683 11680             1680487        lot     18   236196   2748.70    12793.1
## 365     364        A    1925931        lot      8   127621   2648.86    13089.2
## 880     877             1793499        lot      6   152253   3084.54    13544.2
## 883     880             2001480        lot      2   187038   3103.32    13573.8
## 929     926             1706234        lot     23    58221   2341.16    12826.2
## 930     927             1758586        lot     16   100932   2324.10    12819.6
## 1018   1015             1826319        lot     11    92108   2805.18    12896.6
## 1279   1276             1824933        lot     15   193963   2549.34    13002.5
##       district shape_area zoning   zip general_la ewc_dist nsc_dist mopac_dist
## 17068       14   2599.879    UNO 78705        900  7373.47  6737.65    1651.38
## 19146       14  15453.596    UNO 78705        900  7616.57  7226.08    1226.80
## 11683       14   7604.475     NP 78705        900  8022.51  6188.41    2593.72
## 365         14   2595.443    UNO 78705        900  7380.96  6716.15    1677.42
## 880         14   9073.244    UNO 78705        200  7613.56  7165.62    1290.54
## 883         14  15201.121    UNO 78705        200  7589.39  7208.77    1231.23
## 929         14   8428.477    UNO 78705        800  7031.60  6531.55    1718.86
## 930         14   8437.459    UNO 78705        800  6978.07  6543.09    1690.80
## 1018        14   7081.877 SF-3-H 78705        100  8026.43  6303.85    2476.79
## 1279        14   8739.496    UNO 78705        800  7269.64  6650.79    1670.09
```

```
##       x130_dist x35_dist extrail_1m pptrail_1m conf bike_lanes bus_area
## 17068   13397.4 1624.310          0          8    2         24        1
## 19146   13847.9 2086.840          0          9    3         21        1
## 11683   12599.8  856.421          0          5    2          9        1
## 365     13372.5 1600.070          0          8    2         24        1
## 880     13783.4 2022.810          0          9    3         18        1
## 883     13837.3 2074.600          0          9    3         20        1
## 929     13274.4 1472.270          0          6    2         20        1
## 930     13300.4 1496.050          0          7    2         19        1
## 1018    12722.4  979.013          0          7    1         18        1
## 1279    13333.0 1553.500          0          7    3         26        1
##       totbdgarea num_bldgs maxbdgarea tax_break2 bk_tx_brk housing__ education
## 17068      0.000         0      0.000         92 0.0214310  Very Low         5
## 19146   1238.980         1   1238.980         92 0.0000000  Very Low         5
## 11683    251.909         2    170.393         70 0.0000000  Very Low         5
## 365      486.368         1    486.368         92 0.0331512  Very Low         5
## 880     3189.300         3   2390.580         92 0.0000000  Very Low         5
## 883     1238.980         1   1238.980         92 0.0000000  Very Low         5
## 929     2368.310         1   2368.310         42 0.0000000  Very Low         5
## 930     1170.920         3    932.802         42 0.0000000  Very Low         5
## 1018     378.077         2    190.502         70 0.0000000  Very Low         5
## 1279    4544.410         1   4544.410         42 0.0000000  Very Low         5
##       economic__ comprehens med_income med_rent med_home aff_rent aff_own
## 17068        Low          5      11917     1088   292500       94      79
## 19146        Low          5      11917     1088   292500       94      79
## 11683  Very High          5      11917     1088   292500       94      79
## 365          Low          5      11917     1088   292500       94      79
## 880          Low          5      11917     1088   292500       94      79
## 883          Low          5      11917     1088   292500       94      79
## 929         High          5      11917     1088   292500       94      79
## 930         High          5      11917     1088   292500       94      79
## 1018   Very High          5      11917     1088   292500       94      79
## 1279         Low          5      11917     1088   292500       94      79
##
## 17068
## 19146
## 11683
## 365   TCP AMENDED 82219 izz CPOATD Muniz will repair approx 200 LF of sidewalk on South side of 25th
## 880
## 883
## 929
## 930
## 1018
## 1279
##       rank1 rank20 rank21 rank22 rank23 rank24 rank30 rank31 rank4 rank5 rank6
## 17068     1      0      0    0.2      0    0.2    0.5      0     1     1     1
## 19146     1      0      0    0.2      0    0.2    0.5      0     1     1     1
## 11683     1      0      0    0.2      0    0.2    0.0      0     1     1     1
## 365       1      0      0    0.2      0    0.2    0.5      0     1     1     1
## 880       0      0      0    0.2      0    0.2    0.5      0     1     1     1
## 883       0      0      0    0.2      0    0.2    0.5      0     1     1     1
## 929       0      0      0    0.2      0    0.2    0.5      0     1     1     1
## 930       0      0      0    0.2      0    0.2    0.5      0     1     1     1
## 1018      0      0      0    0.2      0    0.2    0.5      0     1     1     1
```

```
## 1279      0      0      0    0.2      0    0.2    0.5      0      1      1      1
##       rank7 rank8 rank9 rank10
## 17068     1     1     1    7.9
## 19146     1     1     1    7.9
## 11683     1     1     1    7.4
## 365       1     1     0    6.9
## 880       1     1     1    6.9
## 883       1     1     1    6.9
## 929       1     1     1    6.9
## 930       1     1     1    6.9
## 1018      1     1     1    6.9
## 1279      1     1     1    6.9
```

1. 17065
2. 19143
3. 11680
4. 364
5. 877
6. 880
7. 926
8. 927
9. 1015
10. 1276

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?

I think it was fairly easy to select the best parcels. It was easier was I was able to quantify the preferences. After adding up it ranks it was just a matter of choosing the top 10 ranks.

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?

I decided on the values to use as cut-offs by using the mean. I figured the ranking system would be fair if I used the mean for all numerical data as a cutoff. I think I was a little unhappy with the available options because using the mean as a cutoff isn't necessarily always fair in a ranking system. It would be better if there were more constraints for these variables.

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?

I do not think I can find a parcel that perfectly satisfies all the requirements and preferences. I think that to satisfy all the preferences we would need to find a better way to quantify and rank more categorical preferences. I believe that preference 9. which was if there was active construction should have more constraints to determine if the parcel was ranked fairly for that constraint. That constraint was ranked with slight bias since I had to determine what word/words would determine if there were active construction.

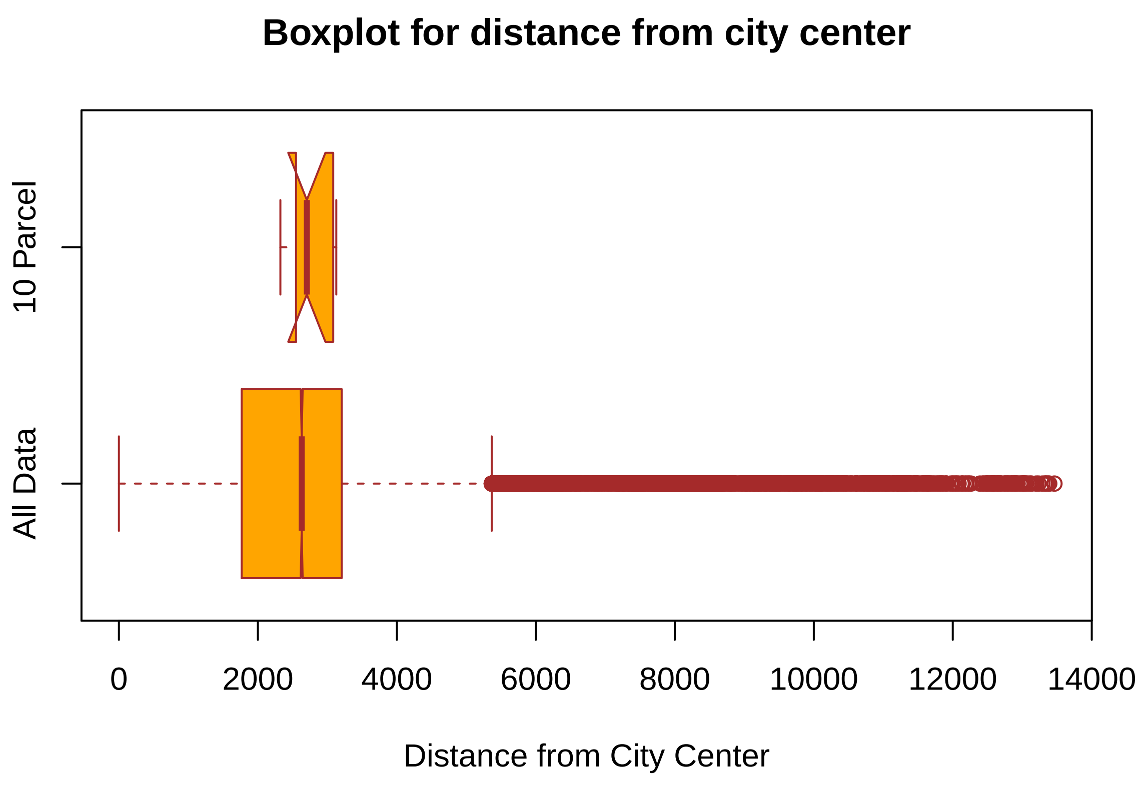
12.1 Display graphs highlighting where your 10 final parcels are compared to the rest of the dataset for at least 3 numeric variables.

```
my_10_parcels = subset(mydata_final, mydata_final$fid == 17065 | mydata_final$fid == 19143  | mydata_fi

parcels_city = mean(my_10_parcels$city_dist)
parcels_extrail = mean(my_10_parcels$extrail_1m)
parcels_bike = mean(my_10_parcels$bike_lanes)
```

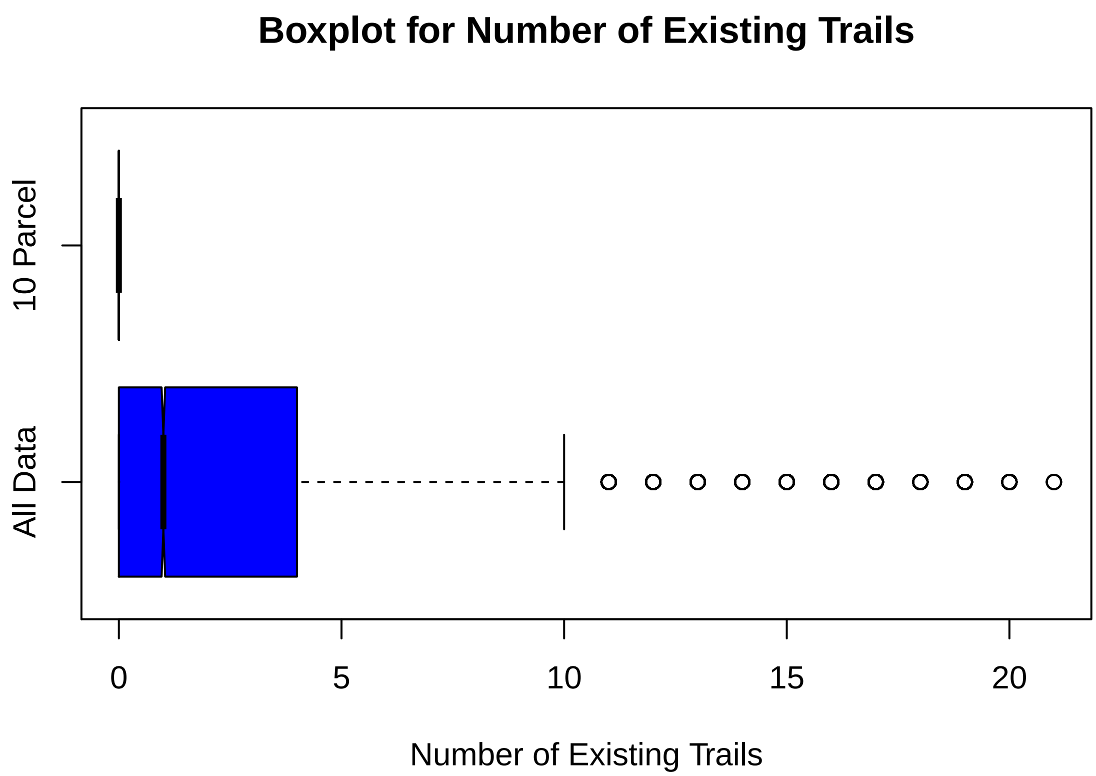
```
boxplot(mydata_final$city_dist,my_10_parcels$city_dist, names=c("All Data","10 Parcel"),
main = "Boxplot for distance from city center",
xlab = "Distance from City Center",
ylab = " ",
col = "orange",
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```

```
## Warning in bxp(list(stats = structure(c(0, 1766.5700075, 2630.23999,
## 3206.075073, : some notches went outside hinges ('box'): maybe set notch=FALSE
```



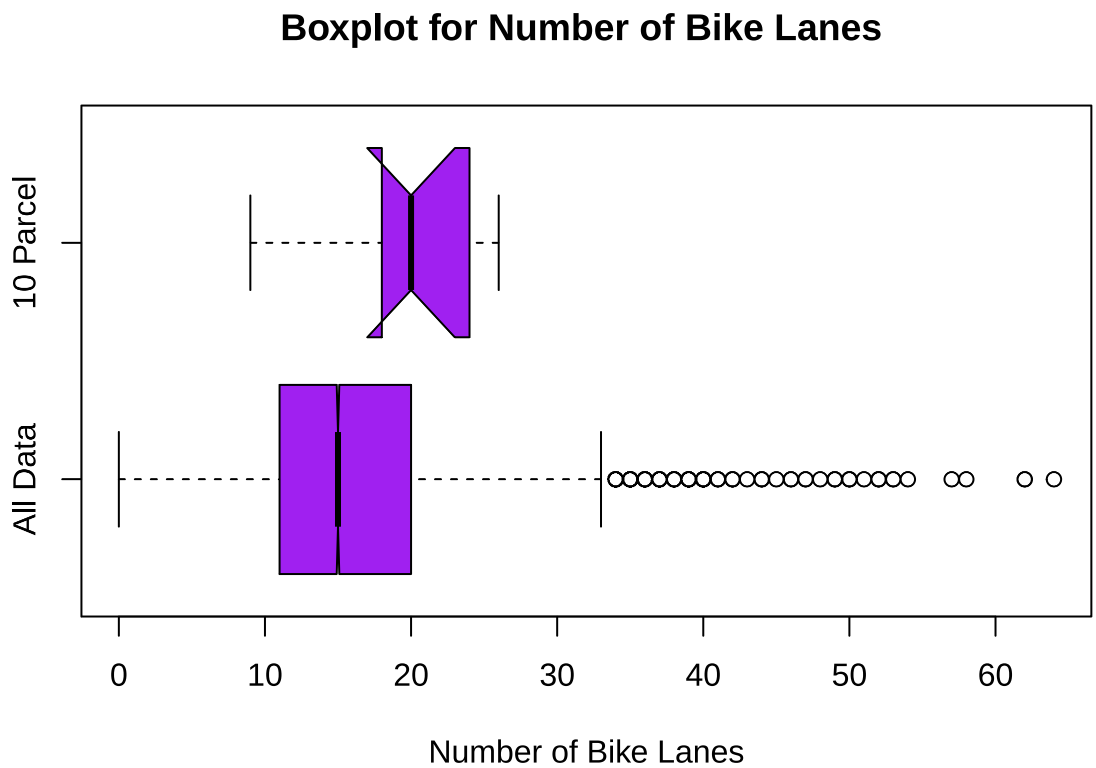

```
boxplot(mydata_final$extrail_1m,my_10_parcels$extrail_1m, names=c("All Data","10 Parcel"),
main = "Boxplot for Number of Existing Trails",
xlab = "Number of Existing Trails",
ylab = " ",
col = "blue",
```

```
border = "black",
horizontal = TRUE,
notch = TRUE
)
```



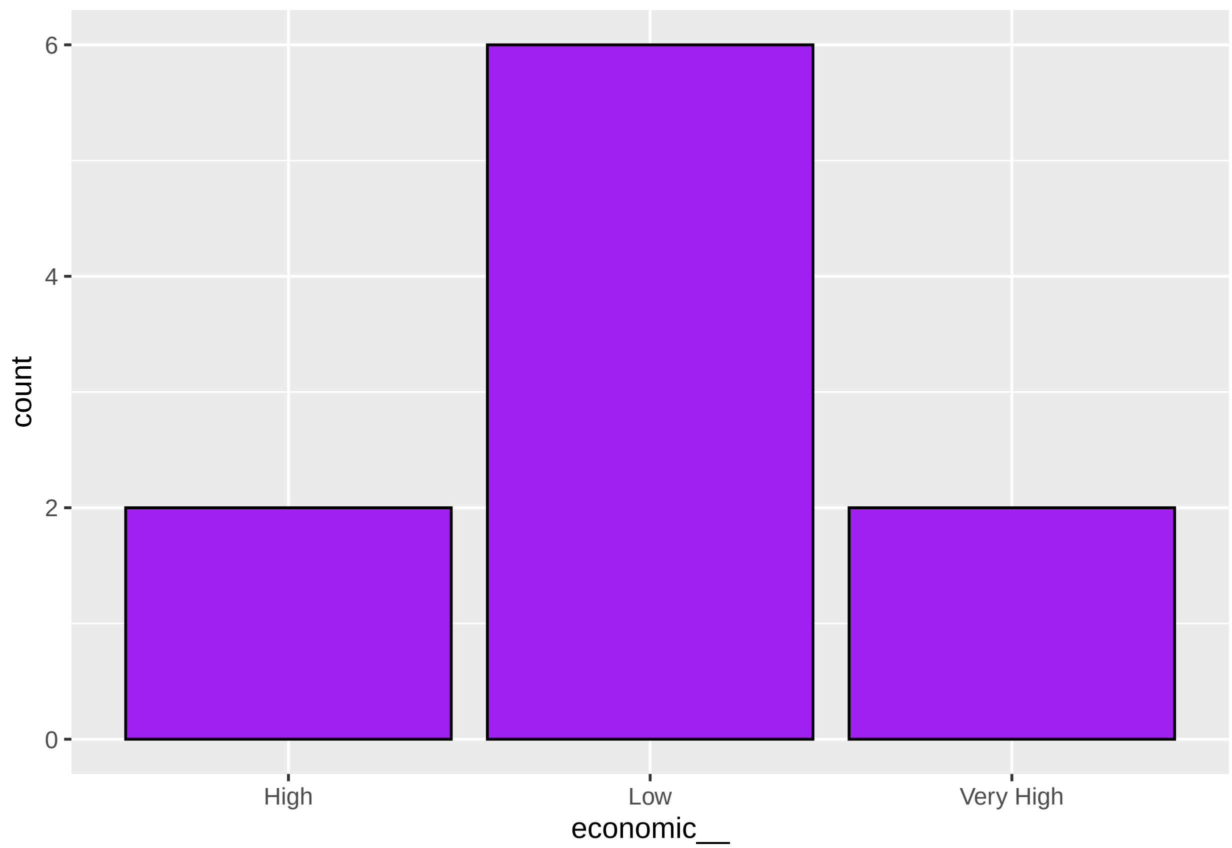

```
boxplot(mydata_final$bike_lanes,my_10_parcels$bike_lanes, names=c("All Data","10 Parcel"),
main = "Boxplot for Number of Bike Lanes ",
xlab = "Number of Bike Lanes",
ylab = " ",
col = "purple",
border = "black",
horizontal = TRUE,
notch = TRUE
)
```

```
## Warning in bxp(list(stats = structure(c(0, 11, 15, 20, 33, 9, 18, 20, 24, : some
## notches went outside hinges ('box'): maybe set notch=FALSE
```

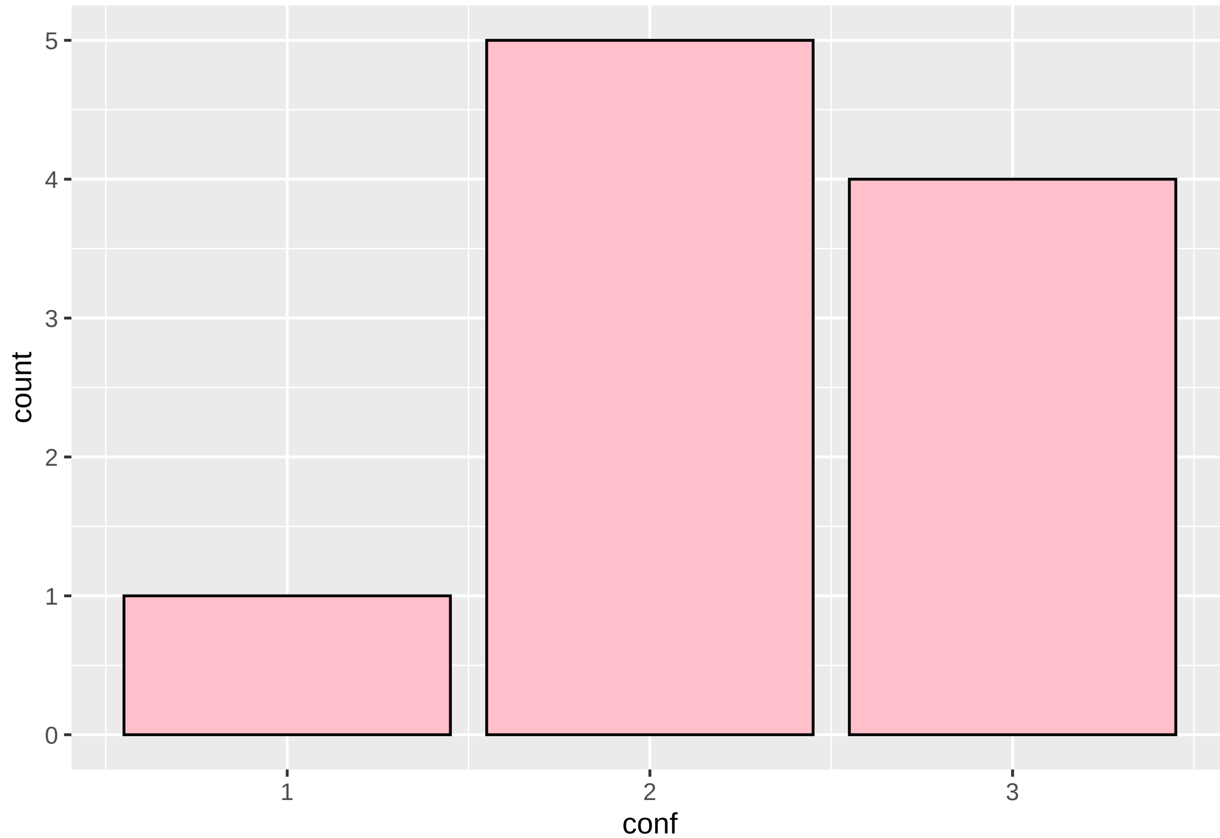


12.2 Create a chart showing qualitative variables for each of the 10 final parcels.

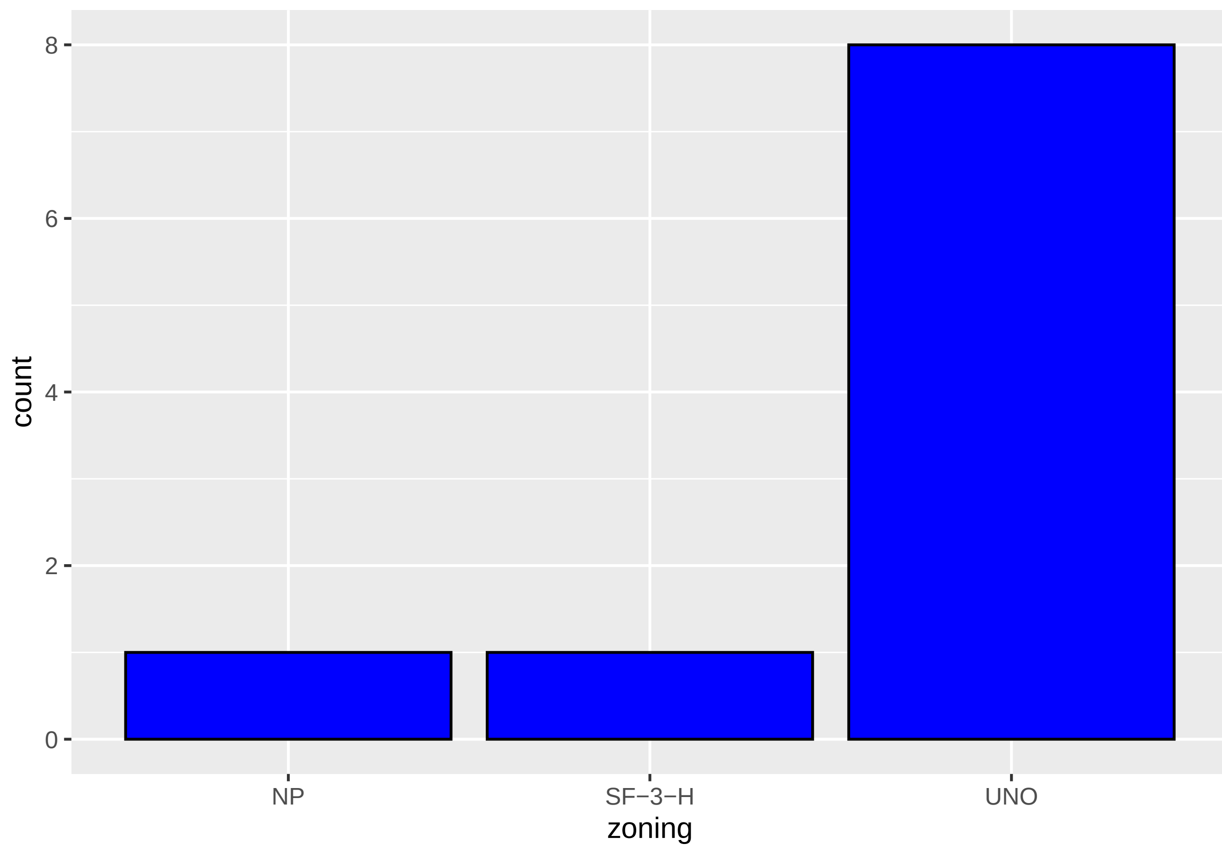
```
ggplot(my_10_parcels) + geom_bar(aes(x = economic__) , color = "black" , fill="purple")
```




```r
ggplot(my_10_parcels) + geom_bar(aes(x = conf) , color = "black" , fill="pink")
```




```r
ggplot(my_10_parcels) + geom_bar(aes(x = zoning) , color = "black" , fill="blue")
```



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.

Overall the 10 parcels seem to have the same average for the distance from the city. The 10 parcels also seem to have a very low number of existing trails which means that the ranking system didn't pick up on that variable, however, the 10 parcels did seem to have a higher average of the number of bike lanes. Overall the 10 parcels seemed to have low economic opportunity, with 7 of them in that category. The majority of the parcels had a bike lane comfort level of 2 or 3 which is not very high so this is something to considering when furthering this ranking system. Most of the parcels fall under the UNO zone which maybe leads us to believe that the majority of the 'good' parcels are in this zone.

I do think the parcels ended up similar to each other based on the ranking system. I think it would be important to include more variables in the ranking system instead of arbitrarily picking which ones fit into the preference. We should include variables like bike lane comfort, into the ranking system. I think it would be good to also lower the ranks based on the importance of preference instead of just assigning 1 and 0.

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).

I think some important factors that help the parcels stand out are variables like economic opportunity and education opportunity. If the parcels have a higher value in these categories then they seem to be good choices. I also think that some variables that are not mentioned in this dataset that should be considered are public school education, like elementary schools. This information would be helpful if the employees had children. They would consider building headquarters near schools.