ExploratoryDataAnalysis

Rajesh Jagannath

August 3, 2016

1. Introduction

Datasets provided by Cincinnati Area Geographic Information System (CAGIS) are being used in the this project. A Street car is being introduced in the City of Cincinnati. Its economic benefit is being analyzed and forecast in this project. A 1000 ft buffer zone has been established around the street-car route. It is further divided into CORE, CENTER and EDGE buffer zones. A subset of featureshave been seleced from the original .csv file. In the following sections, an exploratory analysis of the parcels in the buffer zone has been performed the market Land Value and a distribution of observations based on Existing Land Use Code has been performed.

2. Pre-Procesing

This program reads in the CENTER, EDGE and CORE csv files

```
library(dplyr)
library(RSQLite)
library(tidyr)
library(ggplot2)
library(readr)
library(stringr)
library(scales)
library(mixtools)
library(readxl)
library(ggmap)
```

3. Data Loading

```
# read in the csv
stcar_center <- read_csv("./streetcarbuffer_parcels/parcel_csv_050616/StreetCarParcels_CENTER.csv")
stcar_core   <- read_csv("./streetcarbuffer_parcels/parcel_csv_050616/StreetCarParcels_CORE.csv")
stcar_edge   <- read_csv("./streetcarbuffer_parcels/parcel_csv_050616/StreetCarParcels_EDGE.csv")
```

4. Prepare data

Subset some interesting features - PARCELID, Existing Land Use Code, Mkt value of the land/ improvements, Sale Amt., Area and Acres

```
selected_var_v <- c("PARCELID", "EXLUCODE", "MKTLND", "MKTIMP", "MKT_TOTAL_","ADDRNO","ADDRST", "ADDRSF


parcelid_center<- stcar_center[selected_var_v]
parcelid_edge  <- stcar_core[selected_var_v]
parcelid_core  <- stcar_edge[selected_var_v]
```

```
# Existing land use code  description
exlu_desc <- c("?- Unknown", "C - Commercial", "ED - Educational ", "HI - heavy Industrial", "IN - Insti
```

5a. Exploratory Analysis on Dataset - CORE

```
str(parcelid_core)
```

```
## Classes 'tbl_df' and 'data.frame':   1713 obs. of  8 variables:
##  $ PARCELID  : chr  "007400010001" "007400010002" "007400010003" "007400010004" ...
##  $ EXLUCODE  : chr  "?" "VA" "VA" "VA" ...
##  $ MKTLND    : num  96970 108230 150140 437770 0 ...
##  $ MKTIMP    : num  1.9e+08 0.0 0.0 0.0 0.0 ...
##  $ MKT_TOTAL_: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ADDRNO    : int  405 415 1031 405 NA 1006 NA NA 1000 1008 ...
##  $ ADDRST    : chr  "READING" "READING" "SPRING" "READING" ...
##  $ ADDRSF    : chr  "RD" "RD" "ST" "RD" ...
```

Total number of parcels and the mean Market value of the Land in CORE

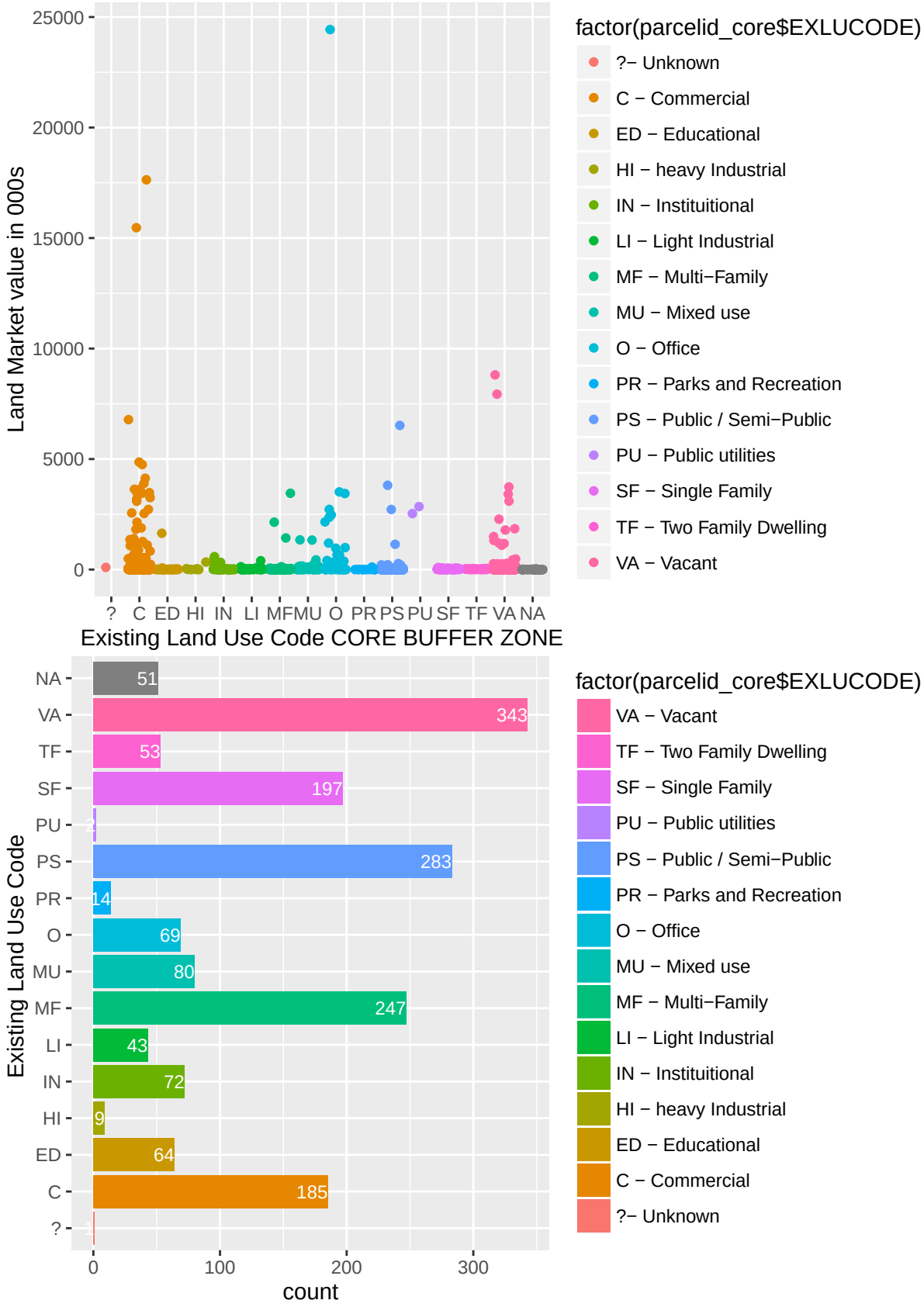
```
## [1] "Total number of parcels in CORE is          1713"
```

```
## [1] "The mean Market value of the land in CORE   155471.593695271"
```

```
## [1] "The count of parcels with 0 MKT value is     427"
```

Plot- CORE

Most High Value properties are in the Core buffer area are C- Commercial, VA - , PS - and O- Office We need to remove or impute the lone data with LandUse classified as "?" One office building has Mkt Land value of 25Million


Land Market value in 000s
25000
20000
15000
10000
5000
0
? C ED HI IN LI MF MU O PR PS PU SF TF VA NA
Existing Land Use Code CORE BUFFER ZONE
factor(parcelid_core$EXLUCODE)
?− Unknown
C − Commercial
ED − Educational
HI − heavy Industrial
IN − Instituitional
LI − Light Industrial
MF − Multi−Family
MU − Mixed use
O − Office
PR − Parks and Recreation
PS − Public / Semi−Public
PU − Public utilities
SF − Single Family
TF − Two Family Dwelling
VA − Vacant
Existing Land Use Code
NA 51
VA 343
TF 53
SF 197
PU 2
PS 283
PR 14
O 69
MU 80
MF 247
LI 43
IN 72
HI 9
ED 64
C 185
? 1
0
100
200
300
count
factor(parcelid_core$EXLUCODE)
VA − Vacant
TF − Two Family Dwelling
SF − Single Family
PU − Public utilities
PS − Public / Semi−Public
PR − Parks and Recreation
O − Office
MU − Mixed use
MF − Multi−Family
LI − Light Industrial
IN − Instituitional
HI − heavy Industrial
ED − Educational
C − Commercial
?− Unknown

5b.Exploratory Analysis on Dataset - CENTER

Total number of parcels and the mean Market value of the Land in CENTER

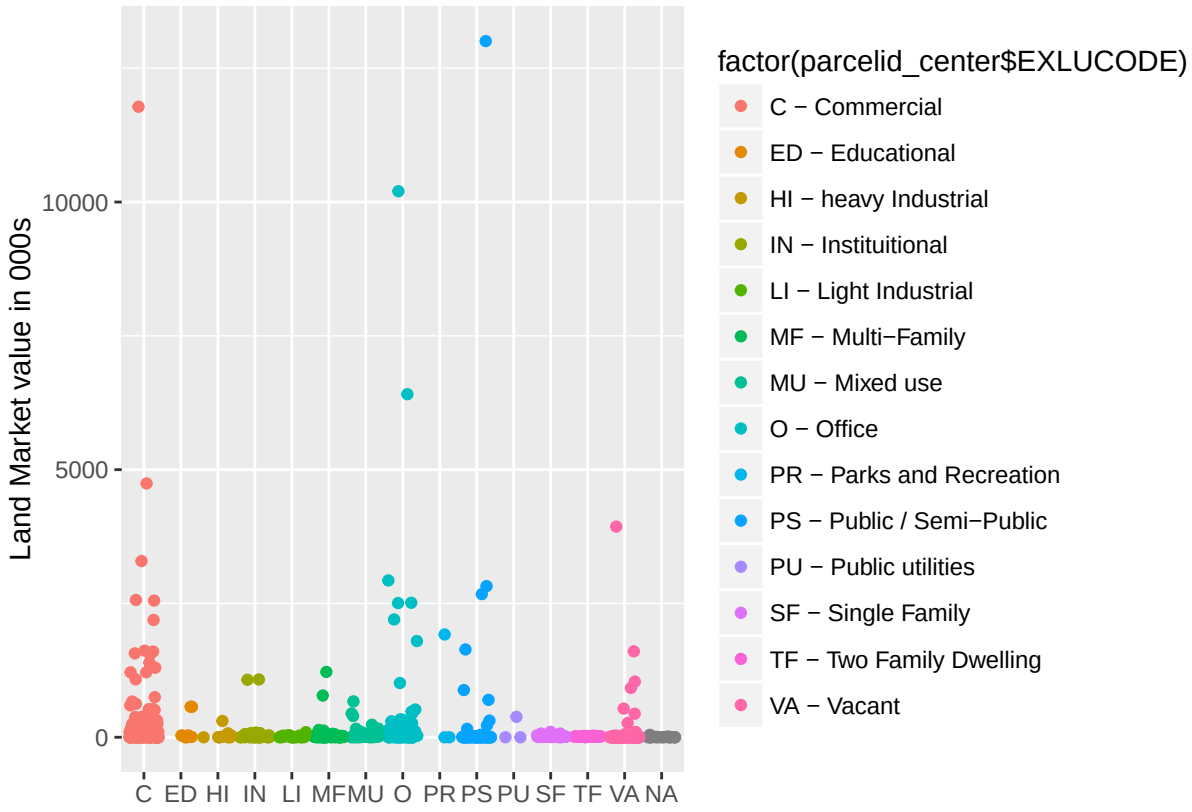
```
## Classes 'tbl_df' and 'data.frame':   946 obs. of  8 variables:
##  $ PARCELID  : chr  "007500010007" "007500010005" "007500010008" "007500010009" ...
##  $ EXLUCODE  : chr  "C" "0" "C" "C" ...
##  $ MKTLND    : num  0 28180 24250 25350 23880 ...
##  $ MKTIMP    : num  0 90810 1570 1960 1570 ...
##  $ MKT_TOTAL_: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ADDRNO    : chr  NA "1208" "312" "314" ...
##  $ ADDRST    : chr  NA "SYCAMORE" "TWELFTH" "TWELFTH" ...
##  $ ADDRSF    : chr  NA "ST" "ST" "ST" ...
```

```
## [1] "Total number of parcels in CENTER is                  946"
```

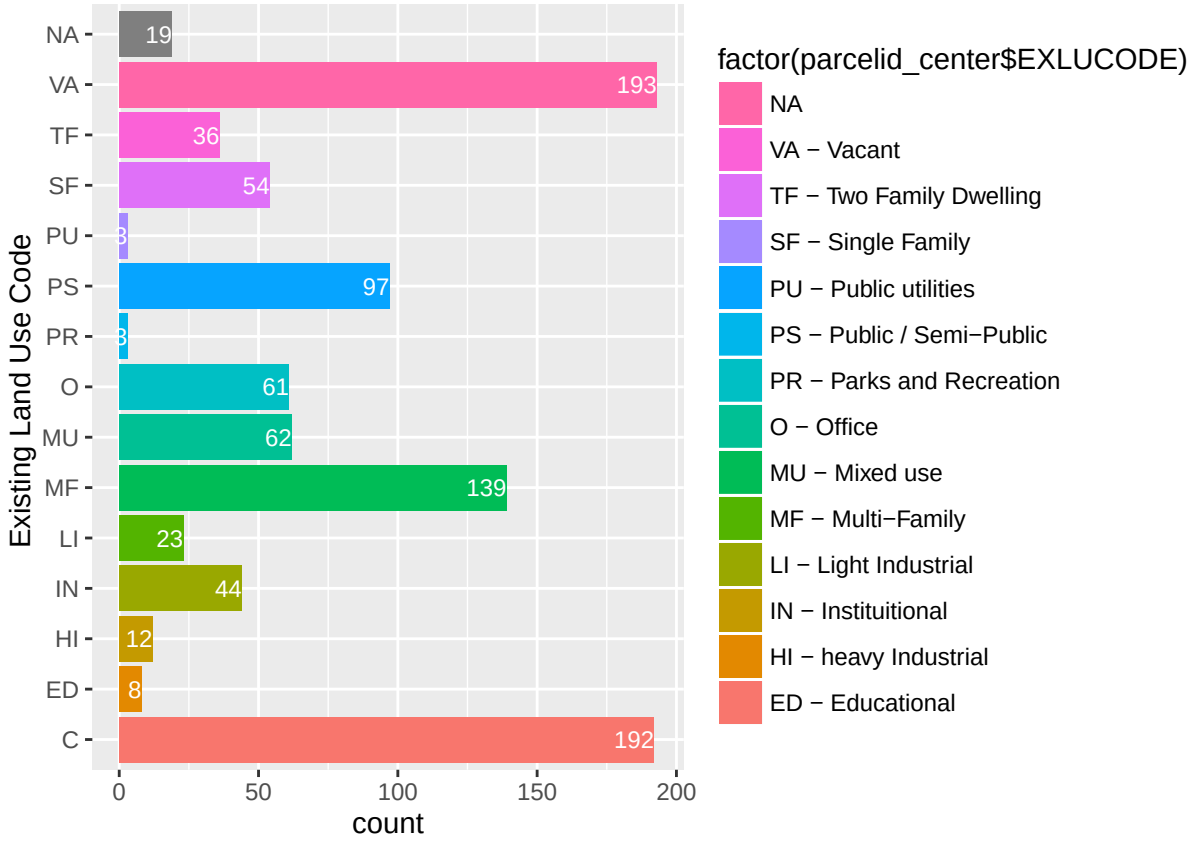
```
## [1] "The mean Market value of the land in ENTER            146591.532769556"
```

```
## [1] "The count of parcels in CENTER  with 0 MKT value is   175"
```

Plot - CENTER



Existing Land Use Code in CENTER BUFFER ZONE



5c. Exploratory Analysis on Dataset - EDGE

```
## Classes 'tbl_df' and 'data.frame':   1418 obs. of  8 variables:
##  $ PARCELID  : chr  "007800020108" "008100040098" "008100040099" "008100040101" ...
##  $ EXLUCODE  : chr  "O" "PS" "LI" "VA" ...
##  $ MKTLND    : num  360650 6000 12080 12040 15030 ...
##  $ MKTIMP    : num  0 0 17090 0 148730 ...
##  $ MKT_TOTAL_: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ADDRNO    : int  250 NA 1415 1409 1405 13 1424 NA 1420 1418 ...
##  $ ADDRST    : chr  "FIFTH" "REPUBLIC" "REPUBLIC" "REPUBLIC" ...
##  $ ADDRSF    : chr  "ST" "ST" "ST" "ST" ...
```

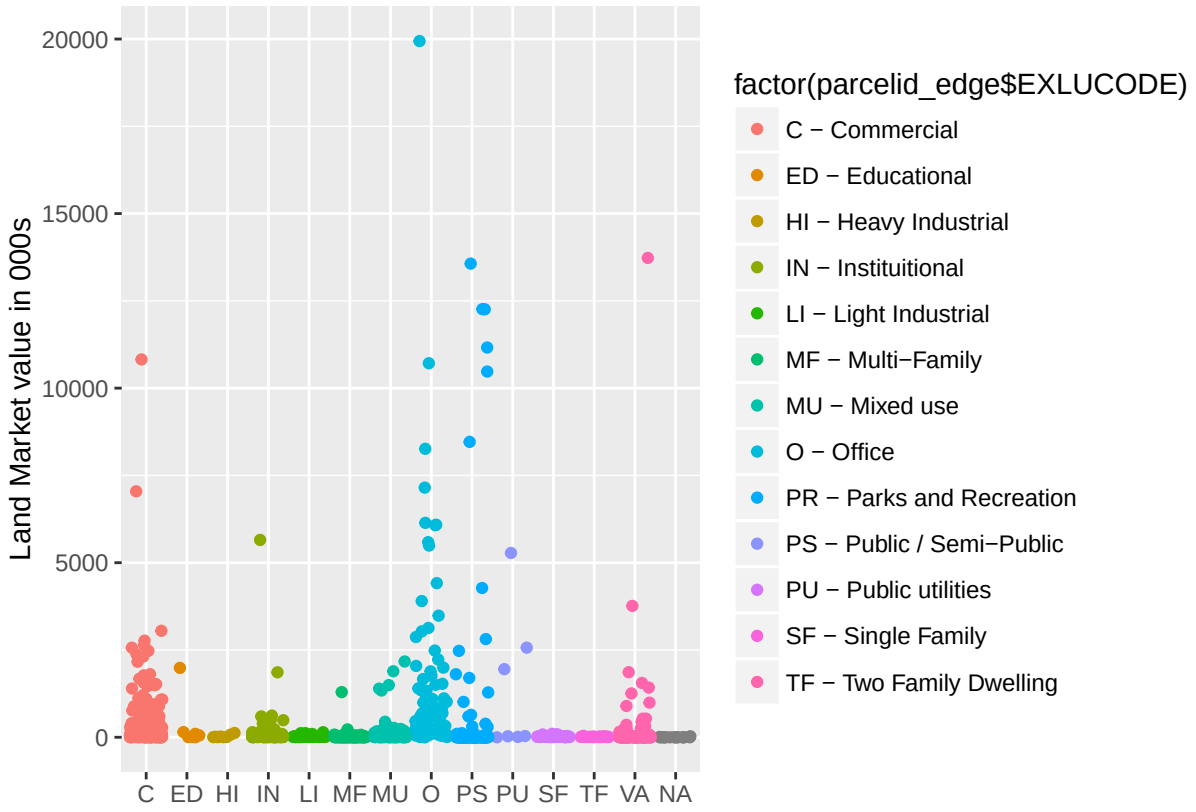
Total records in EDGE

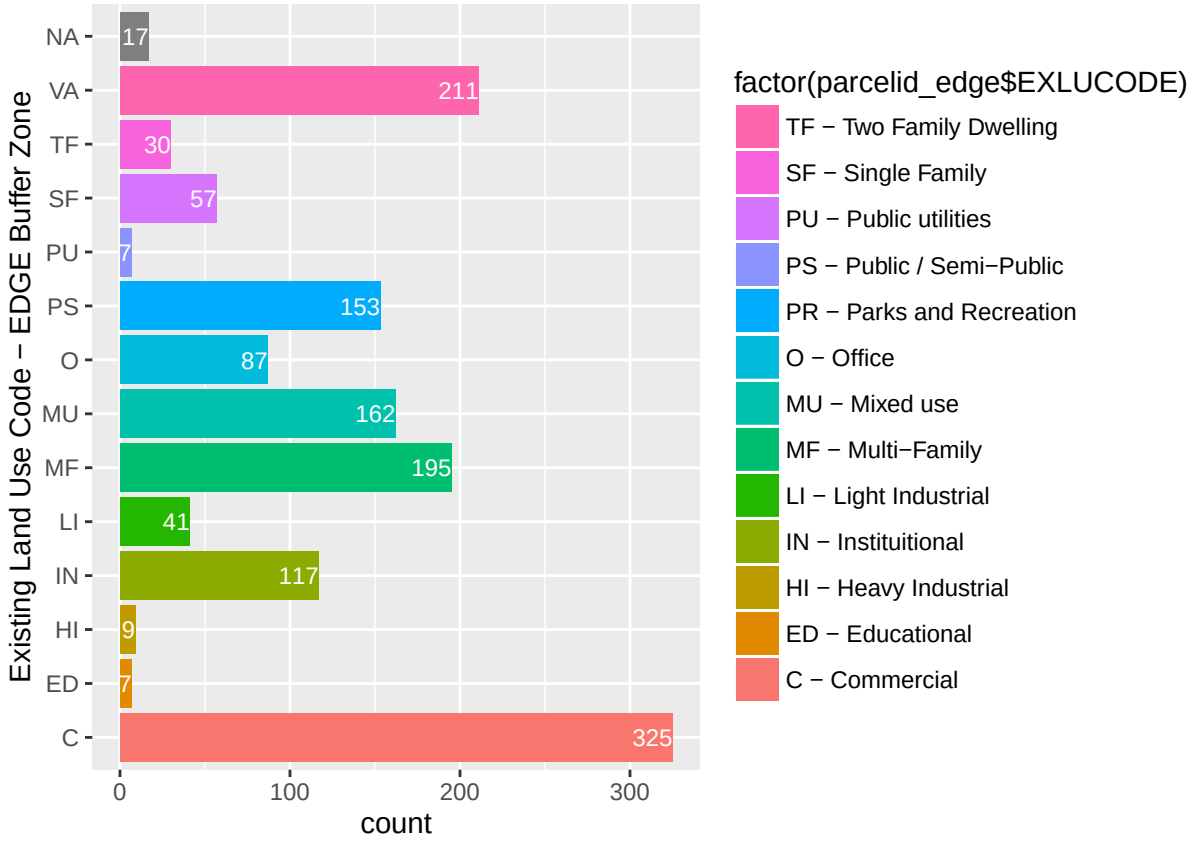
```
## [1] "Total number of parcels in EDGE is            1418"
```

```
## [1] "The mean  Market value of the land in EDGE is 289340.155148096"
```

```
## [1] "The count of parcels with 0 MKTLND value is   69"
```

Plot - EDGE





Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that

generated the plot.

6. Conclusion

- Exploratory analysis of the three buffer zones on the basis of Existing Land Use classification was done
  - Distribution of Parcels vs. Existing Land Use
  - Market Value of the parcels on the basis of Existing Land Use
- The distribution across the Existing Land Use classes is not uniform
- Parcels with High Market Land Value are classified as
  - C - Commercial
  - IN - Instituitional
  - MU - Mixed Use
  - MF - Multi- Family
  - O - Office
  - PS - Public / Semi Public
  - SF - Single Family
  - VA - Vacant
- Parcels with high Market Land values also imply more Property Tax revenue coming into the City of Cincinnati.
- Particular attention should be paid to these Existing Land Use parcels when performing the detailed Analysis/Forecast.