DQR of NY Property

# File description

## file name is NY\_property\_data.csv

## source is internet

ny = fread('NY\_property\_data.csv', header = T, sep = ',')

##   
Read 23.8% of 1048575 rows  
Read 34.3% of 1048575 rows  
Read 45.8% of 1048575 rows  
Read 60.1% of 1048575 rows  
Read 78.2% of 1048575 rows  
Read 1048575 rows and 30 (of 30) columns from 0.145 GB file in 00:00:10

attach(ny)

## recoreds and fields

str(ny)

## Classes 'data.table' and 'data.frame': 1048575 obs. of 30 variables:  
## $ RECORD : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ BBLE : chr "3046020035" "5046820019" "3074790028" "4027980132" ...  
## $ BLOCK : int 4602 4682 7479 2798 695 3181 5186 8202 5257 7078 ...  
## $ LOT : int 35 19 28 132 27 7 1001 64 8 50 ...  
## $ EASEMENT: chr "" "" "" "" ...  
## $ OWNER : chr "DESMOND CAMPBELL" "CINISOMO MARIO" "GANGICHIODO DONALD" "DCAS" ...  
## $ BLDGCL : chr "B1" "A5" "V0" "V0" ...  
## $ TAXCLASS: chr "1" "1" "1B" "1B" ...  
## $ LTFRONT : int 18 25 16 21 0 20 0 24 40 24 ...  
## $ LTDEPTH : int 100 100 19 75 0 100 0 100 96 100 ...  
## $ STORIES : num 2 3 NA NA NA 2 6 2 2 2 ...  
## $ FULLVAL :Class 'integer64' num [1:1048575] 2.01e-318 2.05e-318 6.32e-319 5.56e-319 0.00 ...  
## $ AVLAND :Class 'integer64' num [1:1048575] 6.10e-320 6.57e-320 4.00e-322 9.58e-321 0.00 ...  
## $ AVTOT :Class 'integer64' num [1:1048575] 9.65e-320 1.05e-319 4.00e-322 9.58e-321 0.00 ...  
## $ EXLAND :Class 'integer64' num [1:1048575] 8e-321 8e-321 0e+00 0e+00 0e+00 ...  
## $ EXTOT :Class 'integer64' num [1:1048575] 8e-321 8e-321 0e+00 0e+00 0e+00 ...  
## $ EXCD1 : int 1017 1017 NA NA NA NA NA NA NA 1017 ...  
## $ STADDR : chr "140 EAST 49 STREET" "537 AMHERST AVENUE" "COYLE STREET" "MAZEAU STREET" ...  
## $ ZIP : int 11203 10306 NA NA NA 11375 11355 11236 11358 11223 ...  
## $ EXMPTCL : chr "X7" "" "" "" ...  
## $ BLDFRONT: int 18 14 0 0 0 20 0 20 21 18 ...  
## $ BLDDEPTH: int 36 51 0 0 0 37 0 44 49 65 ...  
## $ AVLAND2 :Class 'integer64' num [1:1048575] 0 0 0 0 0 ...  
## $ AVTOT2 :Class 'integer64' num [1:1048575] 0 0 0 0 0 ...  
## $ EXLAND2 :Class 'integer64' num [1:1048575] 0 0 0 0 0 0 0 0 0 0 ...  
## $ EXTOT2 :Class 'integer64' num [1:1048575] 0 0 0 0 0 0 0 0 0 0 ...  
## $ EXCD2 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ PERIOD : chr "FINAL" "FINAL" "FINAL" "FINAL" ...  
## $ YEAR : chr "2010/11" "2010/11" "2010/11" "2010/11" ...  
## $ VALTYPE : chr "AC-TR" "AC-TR" "AC-TR" "AC-TR" ...  
## - attr(\*, ".internal.selfref")=<externalptr>

## Time frame

unique(ny$YEAR)

## [1] "2010/11"

## column names

names(ny)

## [1] "RECORD" "BBLE" "BLOCK" "LOT" "EASEMENT" "OWNER"   
## [7] "BLDGCL" "TAXCLASS" "LTFRONT" "LTDEPTH" "STORIES" "FULLVAL"   
## [13] "AVLAND" "AVTOT" "EXLAND" "EXTOT" "EXCD1" "STADDR"   
## [19] "ZIP" "EXMPTCL" "BLDFRONT" "BLDDEPTH" "AVLAND2" "AVTOT2"   
## [25] "EXLAND2" "EXTOT2" "EXCD2" "PERIOD" "YEAR" "VALTYPE"

In total, there are 30 variables. There are 17 categoriacal variables and 13 numeric variables.

# Fields Description

## RECORD and BBLE

They are catogarical variables. RECORD and BBLE are 100% unique.

unq\_record = data.frame( unique(RECORD) )  
nrow(unq\_record) / nrow(ny)

## [1] 1

unq\_BBLE = data.frame( unique(BBLE) )  
nrow(unq\_BBLE) / nrow(ny)

## [1] 1

## BLOCK

BLOCK is Categorical variable. Length 5 numeric.MANHATTAN 1 TO 2255,BRONX 2260 TO 5958, BROOKLYN 1 TO 8955, QUEENS 1 TO 16350, STATEN ISLAND 1 TO 8050.

the percent of unique BLOCK is:

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

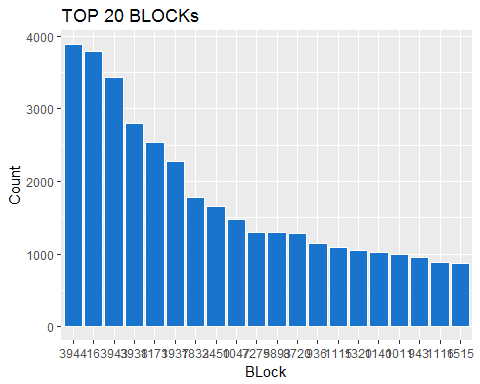
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

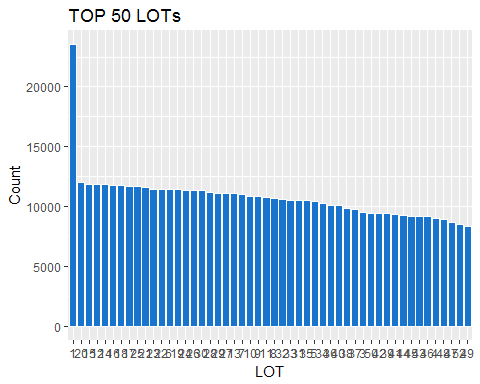
nrow(data.frame(unique(BLOCK)))/ nrow(ny)

## [1] 0.01330282

ny %>% group\_by(BLOCK) %>%   
 summarize(cnt = n()) %>%   
 arrange(desc(cnt) ) %>%   
 slice(1:20) %>%  
 ggplot(aes( x = reorder(as.factor(BLOCK),-cnt), y = cnt) )+  
 geom\_bar(stat = "identity", color = "white" , fill = "dodgerblue3")+  
 xlab("BLock")+  
 ylab("Count")+  
 ggtitle("TOP 20 BLOCKs")

 ## LOT LOT is a categorical varible. To uniquely identify each record of the AV Master Key consists of BBLE, BORO, BLOCK, LOT & EASEMENT CODE. So LOT here is a categorical variable.

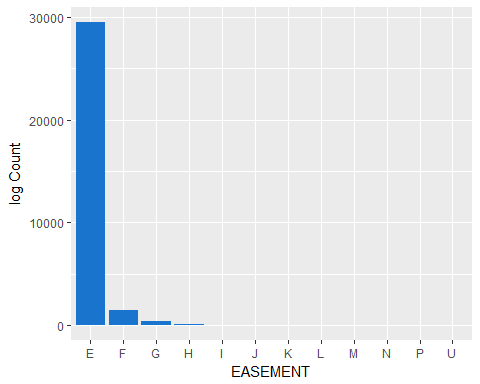
## [1] 0.006071096

 ## EASEMENT EASEMENT is a categorical varible that is used to describe easement. Excepting blank, "E" is most common value, which indicates the portion of the lot that has a Land Easement.

table(EASEMENT)

## EASEMENT  
## E F G H I J K L   
## 1044532 3603 265 95 30 14 7 4 3   
## M N P U   
## 2 17 2 1

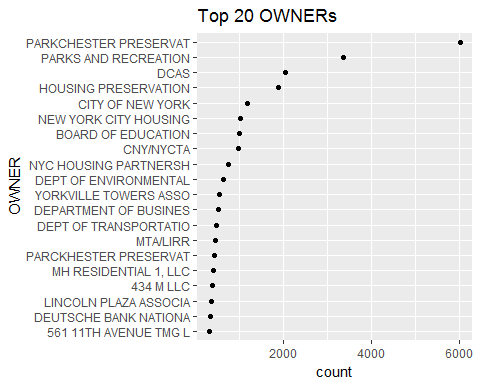
ny[ny$EASEMENT != ""]%>%  
 group\_by(EASEMENT) %>%  
 mutate(cnt1 = n()) %>%  
 ggplot( aes( x = EASEMENT, y = log( cnt1)) )+  
 geom\_bar(stat = "identity",fill = "dodgerblue3")+  
 xlab("EASEMENT")+  
 ylab("log Count")



## OWNER

Categorical variable.OWNER PARCKHESTER PRESERVAT is wrongly spelled which should be PARKCHESTER PRESERVAT.

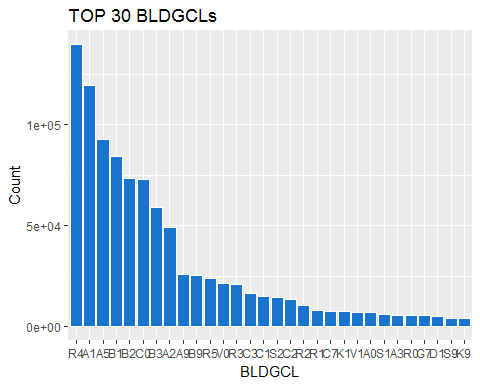
ny %>%   
 filter(OWNER!="") %>%  
 group\_by(OWNER) %>%   
 summarize(cnt = n()) %>%   
 arrange(desc(cnt) ) %>%   
 slice(1:20) %>%  
 ggplot(aes( x = cnt, y = reorder(as.factor(OWNER),cnt)) )+  
 geom\_point()+  
 xlab("count")+  
 ylab("OWNER")+  
 ggtitle("Top 20 OWNERs")



## BLDGCL

Building Class. Position 1 = ALPHA & Position 2 = NUMERIC. There is a direct correlation between the Building Class and the Tax Class.

ny %>%   
 select (BLDGCL) %>%  
 group\_by(BLDGCL) %>%  
 mutate(cnt = n()) %>%  
 arrange(-cnt) %>%  
 distinct() %>%  
 arrange(desc(cnt)) %>%  
 slice(1:30) %>%  
   
 ggplot(aes(x = reorder(BLDGCL,-cnt) , y = cnt ) ) +  
 geom\_bar(stat = "identity", fill = "dodgerblue3" , colour = "white")+  
 xlab("BLDGCL")+  
 ylab("Count")+  
 ggtitle("TOP 30 BLDGCLs")



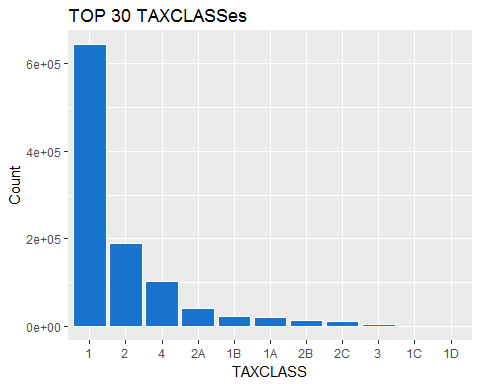
## TAXCLASS

Categorical variable. Building Class. Position 1 = ALPHA & Position 2 = NUMERIC. There is a direct correlation between the Building Class and the Tax Class.1 = 1-3 UNIT RESIDENCES, 1A = 1-3 STORY CONDOMINIUMS ORIGINALLY A CONDO, 1B = RESIDENTIAL VACANT LAND. 1C = 1-3 UNIT CONDOMINUMS ORIGINALLY TAX CLASS 1. 1D = SELECT BUNGALOW COLONIES. 2 = APARTMENTS, 2A = APARTMENTS WITH 4-6 UNITS, 2B = APARTMENTS WITH 7-10 UNITS, 2C = COOPS/CONDOS WITH 2-10 UNITS.  
3 = UTILITIES (EXCEPT CEILING RR), 4A = UTILITIES - CEILING RAILROADS, 4 = ALL OTHERS.

table(TAXCLASS)

## TAXCLASS  
## 1 1A 1B 1C 1D 2 2A 2B 2C 3   
## 643774 20899 22193 946 29 188592 40558 13962 10795 4546   
## 4   
## 102281

ny %>%   
 select (TAXCLASS) %>%  
 group\_by(TAXCLASS) %>%  
 mutate(cnt = n()) %>%  
 arrange(-cnt) %>%  
 distinct() %>%  
 arrange(desc(cnt)) %>%  
   
 ggplot(aes(x = reorder(TAXCLASS,-cnt) , y = cnt ) ) +  
 geom\_bar(stat = "identity", fill = "dodgerblue3" , colour = "white")+  
 xlab("TAXCLASS")+  
 ylab("Count")+  
 ggtitle("TOP 30 TAXCLASSes")



## LTFRONT

numeric variable.

# statistical description:  
summary(LTFRONT)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 19.00 25.00 36.17 40.00 9999.00

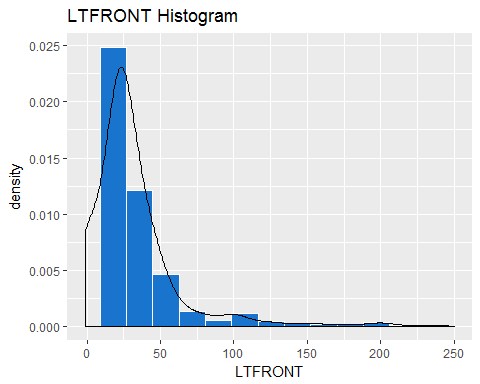
stdev = sd(LTFRONT) ;stdev

## [1] 73.73356

#variance = var(LTFRONT) ;variance  
mode0 = names(sort(desc(table(LTFRONT))))[1]; mode0

## [1] "0"

ggplot(ny, aes(x = LTFRONT , y = ..density.. ) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 15)+  
 geom\_density(adjust = 10)+  
 xlim(-1,250) +  
 xlab("LTFRONT")+  
 ggtitle("LTFRONT Histogram")



## LTDEPTH

numeric variable

summary(LTDEPTH)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 80.00 100.00 88.28 100.00 9999.00

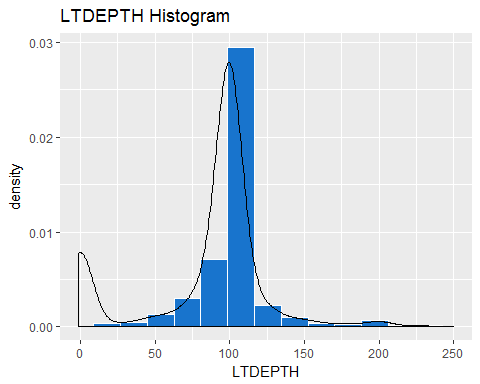
stdev = sd(LTDEPTH) ;stdev

## [1] 75.47885

#variance = var(LTDEPTH) ;variance  
mode0 = names(sort(desc(table(LTDEPTH))))[1]; mode0

## [1] "100"

ggplot(ny, aes(x = LTDEPTH, y = ..density.. ) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 15)+  
 geom\_density(adjust = 10) +  
 xlim(-1,250) +  
 xlab("LTDEPTH")+  
 ggtitle("LTDEPTH Histogram")



## STORIES

numeric variable

summary(STORIES)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.00 2.00 2.00 5.06 3.00 119.00 52142

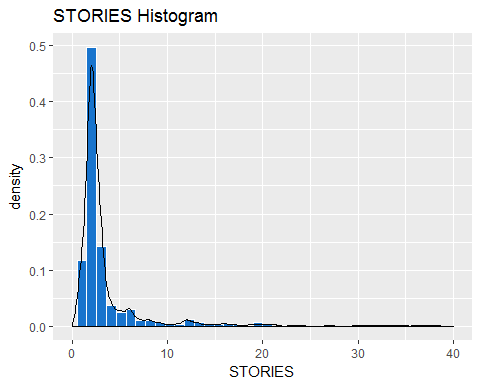
stdev = sd(STORIES, na.rm = TRUE) ;stdev

## [1] 8.431372

#variance = var(STORIES, na.rm = TRUE) ;variance  
mode0 = names(sort(desc(table(STORIES))))[1]; mode0

## [1] "2"

ggplot(ny, aes(x = STORIES, y = ..density..) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 40 )+  
 geom\_density(adjust = 10) +  
 xlim(0,40) +  
 xlab("STORIES")+  
 ggtitle("STORIES Histogram")



## FULLVAL

numeric variable

summary(FULLVAL)

## integer64  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 303000 446000 880487 619000 6150000000

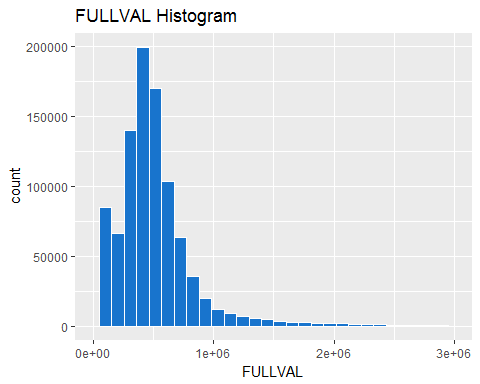
stdev = sd(FULLVAL, na.rm = TRUE) ;stdev

## [1] 11702927

#variance = var(FULLVAL/100000, na.rm = TRUE) ;variance  
mode0 = names(sort(desc(table(FULLVAL))))[1]; mode0

## [1] "0"

ggplot(ny, aes(x = FULLVAL) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2 ) +  
 xlim(0,3000000) +  
 xlab("FULLVAL")+  
 ggtitle("FULLVAL Histogram")



## AVLAND

numeric variable

summary(AVLAND)

## integer64  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 9160 13646 85995 19706 2668500000

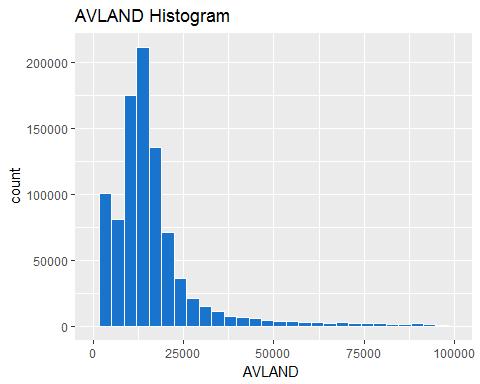
stdev = sd(AVLAND, na.rm = TRUE) ;stdev

## [1] 4100755

#variance = var(AVLAND/10000, na.rm = TRUE) ;variance  
mode0 = names(sort(desc(table(AVLAND))))[1]; mode0

## [1] "0"

ggplot(ny, aes(x = AVLAND) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2 ) +  
 xlim(0,10\*10000) +  
 xlab("AVLAND ")+  
 ggtitle("AVLAND Histogram")



## EXTOT

numeric variable

summary(EXTOT)

## integer64  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 1620 92543 2090 4668308947

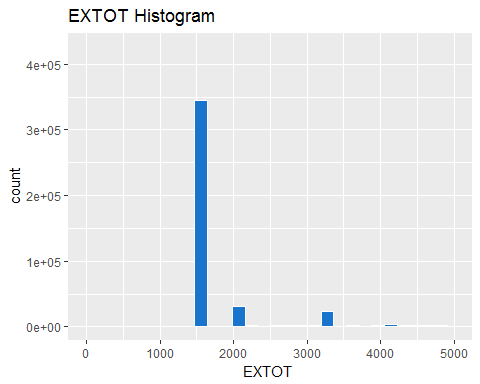
stdev = sd(EXTOT, na.rm = TRUE) ;stdev

## [1] 6578281

#variance = var(EXTOT/1000, na.rm = TRUE) ;variance  
mode0 = names(sort(desc(table(EXTOT))))[1]; mode0

## [1] "0"

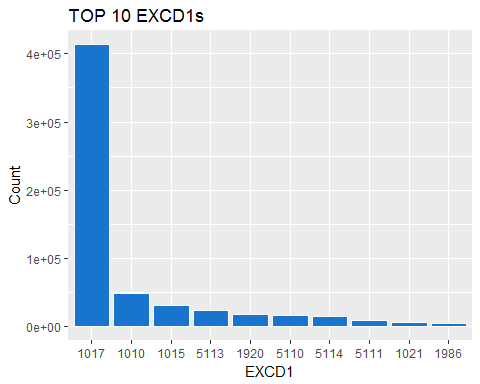
ggplot(ny, aes(x = EXTOT) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2 ) +  
 xlim(0,5000) +  
 xlab("EXTOT ")+  
 ggtitle("EXTOT Histogram")

 ## EXCD1 Categorical variable

attach(ny)  
levels(as.factor(EXCD1))

## [1] "1010" "1011" "1015" "1016" "1017" "1019" "1021" "1022" "1023" "1101"  
## [11] "1102" "1200" "1301" "1401" "1402" "1403" "1404" "1501" "1504" "1505"  
## [21] "1511" "1521" "1522" "1523" "1561" "1562" "1571" "1572" "1601" "1602"  
## [31] "1603" "1604" "1605" "1606" "1620" "1630" "1640" "1650" "1660" "1700"  
## [41] "1840" "1841" "1850" "1870" "1871" "1872" "1880" "1881" "1882" "1891"  
## [51] "1920" "1925" "1950" "1951" "1961" "1963" "1985" "1986" "1990" "1992"  
## [61] "2100" "2120" "2131" "2132" "2133" "2134" "2151" "2152" "2171" "2172"  
## [71] "2191" "2198" "2201" "2202" "2220" "2231" "2232" "2233" "2251" "2252"  
## [81] "2261" "2262" "2280" "2310" "2350" "2400" "2500" "3360" "3380" "3390"  
## [91] "3400" "3410" "3500" "4500" "4520" "4530" "4540" "4550" "4600" "4650"  
## [101] "5101" "5102" "5103" "5104" "5105" "5106" "5107" "5108" "5109" "5110"  
## [111] "5111" "5112" "5113" "5114" "5116" "5118" "5129" "5130" "6120" "6200"  
## [121] "6320" "6400" "6600" "6800" "7120" "7150" "7160" "7165" "7170"

ny[!is.na(EXCD1)] %>%   
 select (EXCD1) %>%  
 group\_by(EXCD1) %>%  
 mutate(cnt = n()) %>%  
 arrange(-cnt) %>%  
 distinct() %>%  
 arrange(desc(cnt)) %>%  
 slice(1:10) %>%  
 ggplot(aes(x = reorder(EXCD1,-cnt) , y = cnt ) ) +  
 geom\_bar(stat = "identity", fill = "dodgerblue3" , colour = "white")+  
 xlab("EXCD1")+  
 ylab("Count")+  
 ggtitle("TOP 10 EXCD1s")



## STADDR

Categorical variable

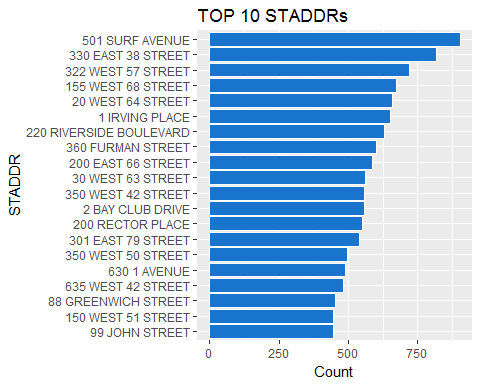
head(STADDR)

## [1] "140 EAST 49 STREET" "537 AMHERST AVENUE" "COYLE STREET"   
## [4] "MAZEAU STREET" "WEST 23 STREET" "90-07 68 AVENUE"

# NA and blank values is   
nrow(ny[is.na(STADDR) | STADDR ==""] )/nrow(ny)

## [1] 0.0006113058

ny[!is.na(STADDR) & STADDR !=""] %>%   
 select (STADDR) %>%  
 group\_by(STADDR) %>%  
 mutate(cnt = n()) %>%  
 arrange(-cnt) %>%  
 distinct() %>%  
 arrange(desc(cnt)) %>%  
 slice(1:20) %>%  
 ggplot(aes(x = reorder(STADDR,cnt) , y = cnt ) ) +  
 geom\_bar(stat = "identity", fill = "dodgerblue3" , colour = "white")+  
 coord\_flip() +  
 xlab("STADDR")+  
 ylab("Count")+  
 ggtitle("TOP 10 STADDRs")



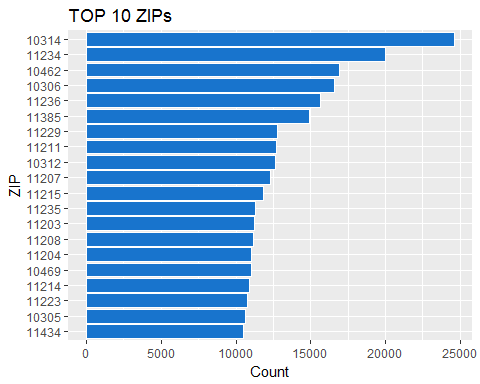
## ZIP

Categorical variable

# NA and blank values is   
nrow(ny[is.na(ZIP) | ZIP ==""] )/nrow(ny)

## [1] 0.02513506

ny[!is.na(ZIP) & STADDR !=""] %>%   
 select (ZIP) %>%  
 group\_by(ZIP) %>%  
 mutate(cnt = n()) %>%  
 arrange(-cnt) %>%  
 distinct() %>%  
 arrange(desc(cnt)) %>%  
 slice(1:20) %>%  
 ggplot(aes(x = reorder(ZIP,cnt) , y = cnt ) ) +  
 geom\_bar(stat = "identity", fill = "dodgerblue3" , colour = "white")+  
 coord\_flip() +  
 xlab("ZIP")+  
 ylab("Count")+  
 ggtitle("TOP 10 ZIPs")



## EXMPTCL

Categorical variable

levels(as.factor(EXMPTCL))

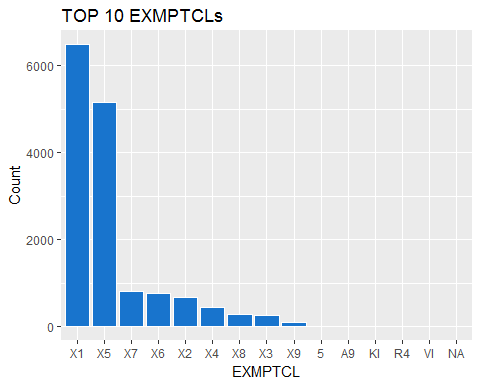
## [1] "" "5" "A9" "KI" "R4" "VI" "X1" "X2" "X3" "X4" "X5" "X6" "X7" "X8"  
## [15] "X9"

# NA and blank values is   
nrow(ny[is.na(EXMPTCL) | EXMPTCL ==""] )/nrow(ny)

## [1] 0.9857025

ny[!is.na(EXMPTCL) & EXMPTCL !=""] %>%   
 select (EXMPTCL) %>%  
 group\_by(EXMPTCL) %>%  
 mutate(cnt = n()) %>%  
 arrange(-cnt) %>%  
 distinct() %>%  
 arrange(desc(cnt)) %>%  
 slice(1:20) %>%  
 ggplot(aes(x = reorder(EXMPTCL,-cnt) , y = cnt ) ) +  
 geom\_bar(stat = "identity", fill = "dodgerblue3" , colour = "white")+  
 #coord\_flip() +  
 xlab("EXMPTCL")+  
 ylab("Count")+  
 ggtitle("TOP 10 EXMPTCLs")

## Warning: Removed 6 rows containing missing values (position\_stack).



## BLDFRONT

# statistical description:  
summary(BLDFRONT)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 15.00 20.00 23.02 24.00 7575.00

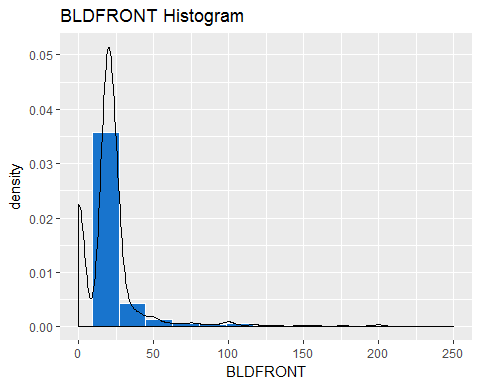
stdev = sd(BLDFRONT) ;stdev

## [1] 35.78847

#variance = var(BLDFRONT) ;variance  
mode0 = names(sort(desc(table(BLDFRONT))))[1]; mode0

## [1] "0"

ggplot(ny, aes(x = BLDFRONT ,y = ..density.. ) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 15)+  
 geom\_density(adjust = 10) +  
 xlim(0,250) +  
 xlab("BLDFRONT")+  
 ggtitle("BLDFRONT Histogram")



## BLDDEPTH

numeric variable.

# statistical description:  
summary(BLDDEPTH)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 26.00 39.00 40.07 51.00 9393.00

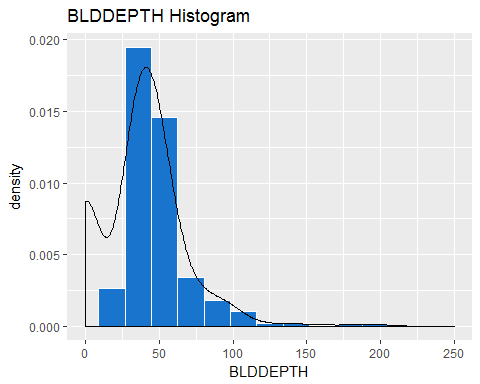
stdev = sd(BLDDEPTH) ;stdev

## [1] 43.0364

#variance = var(BLDDEPTH) ;variance  
mode0 = names(sort(desc(table(BLDDEPTH))))[1]; mode0

## [1] "0"

ggplot(ny, aes(x = BLDDEPTH ,y = ..density.. ) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 15)+  
 geom\_density(adjust = 10) +  
 xlim(0,250) +  
 xlab("BLDDEPTH")+  
 ggtitle("BLDDEPTH Histogram")



## AVLAND2

numeric variable.

# statistical description:  
summary(AVLAND2,na.rm = TRUE)

## integer64  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3 5705 20059 246365 62340 2371005000   
## NA's   
## 767609

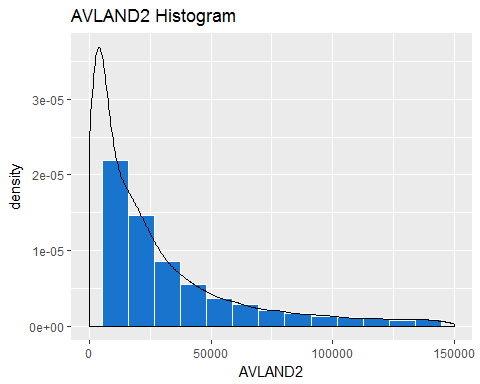
stdev = sd(AVLAND2,na.rm = TRUE) ;stdev

## [1] 6199390

#variance = var(AVLAND2,na.rm = TRUE) ;variance  
mode0 = names(sort(desc(table(AVLAND2))))[1]; mode0

## [1] "2408"

ggplot(ny, aes(x = AVLAND2 ,y = ..density.. ) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 15)+  
 geom\_density(adjust = 2) +  
 xlim(0,150000) +  
 xlab("AVLAND2")+  
 ggtitle("AVLAND2 Histogram")



## AVTOT2

numeric variable.

# statistical description:  
summary(AVTOT2,na.rm = TRUE)

## integer64  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3 34014 80010 716078 240792 4501180002   
## NA's   
## 767603

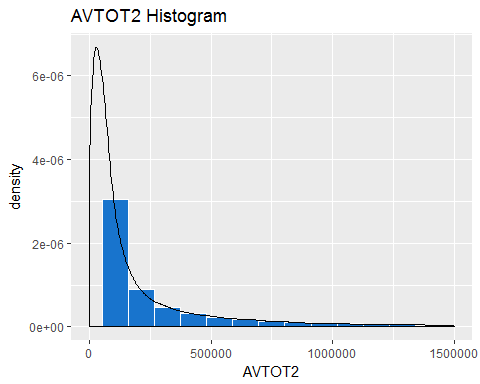
stdev = sd(AVTOT2,na.rm = TRUE) ;stdev

## [1] 11690165

#variance = var(AVTOT2/100000,na.rm = TRUE) ;variance  
mode0 = names(sort(desc(table(AVTOT2))))[1]; mode0

## [1] "750"

ggplot( ny, aes(x = AVTOT2 ,y = ..density.. ) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 15)+  
 geom\_density(adjust = 2) +  
 xlim(0,15\*100000) +  
 xlab("AVTOT2")+  
 ggtitle("AVTOT2 Histogram")



## EXLAND2

numeric variable.

# statistical description:  
summary(EXLAND2,na.rm = TRUE)

## integer64  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 2090 3053 351802 31419 2371005000   
## NA's   
## 961900

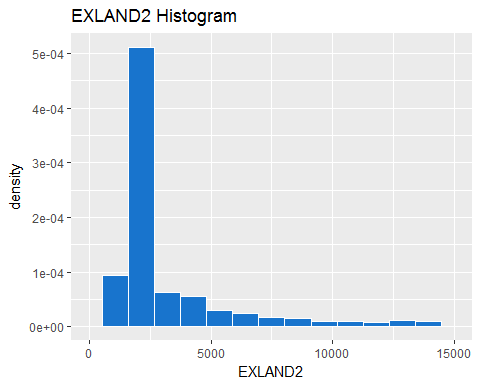
stdev = sd(EXLAND2,na.rm = TRUE) ;stdev

## [1] 10852484

#variance = var(EXLAND2/100000,na.rm = TRUE) ;variance  
mode0 = names(sort(desc(table(EXLAND2))))[1]; mode0

## [1] "2090"

ggplot( ny, aes(x = EXLAND2 ,y = ..density.. ) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 15)+  
 # geom\_density() +  
 xlim(0,15000) +  
 xlab("EXLAND2")+  
 ggtitle("EXLAND2 Histogram")



## EXTOT2

numeric variable.

# statistical description:  
summary(EXTOT2,na.rm = TRUE)

## integer64  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7 2889 37116 658114 106629 4501180002   
## NA's   
## 918642

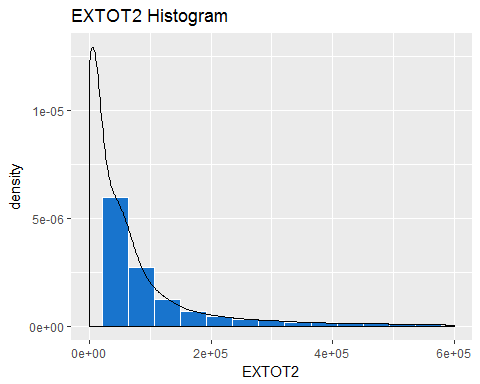
stdev = sd(EXTOT2,na.rm = TRUE) ;stdev

## [1] 16129808

#variance = var(EXTOT2/100000,na.rm = TRUE) ;variance  
mode0 = names(sort(desc(table(EXTOT2))))[1]; mode0

## [1] "2090"

ggplot( ny, aes(x = EXTOT2 ,y = ..density.. ) ) +  
 geom\_histogram(fill = "dodgerblue3" , colour = "white",size = .2, bins = 15)+  
 geom\_density(adjust = 3) +  
 xlim(0,600000) +  
 xlab("EXTOT2")+  
 ggtitle("EXTOT2 Histogram")



## EXCD2

Categorical variables.

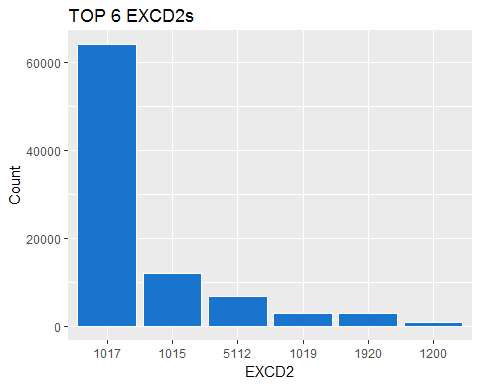
# statistical description:  
levels(as.factor(EXCD2))

## [1] "1011" "1015" "1017" "1019" "1021" "1022" "1023" "1101" "1102" "1200"  
## [11] "1401" "1402" "1404" "1501" "1504" "1505" "1521" "1522" "1523" "1602"  
## [21] "1603" "1604" "1605" "1850" "1891" "1920" "1930" "1985" "1986" "2132"  
## [31] "2134" "2151" "2152" "2201" "2202" "2231" "2232" "2262" "2280" "2310"  
## [41] "2350" "2500" "3390" "3410" "4500" "5101" "5104" "5108" "5109" "5110"  
## [51] "5111" "5112" "5113" "5114" "5116" "5129" "5130" "6320" "6800" "7160"

# NA and blank values is   
nrow(ny[is.na(EXCD2) | EXCD2 ==""] )/nrow(ny)

## [1] 0.9132718

ny[!is.na(EXCD2) & EXCD2 !=""] %>%   
 select (EXCD2) %>%  
 group\_by(EXCD2) %>%  
 mutate(cnt = n()) %>%  
 arrange(-cnt) %>%  
 distinct() %>%  
 arrange(desc(cnt)) %>%  
 slice(1:6) %>%  
 ggplot(aes(x = reorder(EXCD2,-cnt) , y = cnt ) ) +  
 geom\_bar(stat = "identity", fill = "dodgerblue3" , colour = "white")+  
 #coord\_flip() +  
 xlab("EXCD2")+  
 ylab("Count")+  
 ggtitle("TOP 6 EXCD2s")



## PERIOD, YEAR, VALTYPE

These are categorical variables.

# statistical description:  
levels(as.factor(PERIOD))

## [1] "FINAL"

levels(as.factor(YEAR))

## [1] "2010/11"

levels(as.factor(VALTYPE))

## [1] "AC-TR"