Fraud Detection Report

-The New York City Property Dataset

DSO562-Fraud Analytics

Team 5:

Rui Xin Wu
Zhuolin Ouyang
Xiaoxuan Feng
Feifei Shao
Zihao Geng
Zimeng Cao
Zijian Wang

Date: 02/22/2018

Table of Contents

Part I. Executive Summary

This report examines the New York City Property data to detect the abnormality and potential fraud events using an unsupervised machine learning method. Our team used Python and R as major tools, and Principle Component Analysis (PCA), Heuristic Algorithm, and Autoencoder as featured analysis methods.

The steps we took were:

1. Data cleaning and variable construction
2. Dimensionality reduction using PCA
3. Fraud score building using Heuristic Algorithm and Autoencoder
4. Algorithm combining to improve accuracy

The original dataset contains more than one million properties' data across the city of New York, with information about the property features and locations.

Using the PCA, we reduced dimensionality by retaining only PC1-PC7 as the principal components for further analysis. With the heuristic fraud algorithm, we found the score range of the properties to be right skewed with 99.37% of all records between 0 and 1, and the high of 1921.29. For autoencoder, we calculated the fraud scored based on mean squared error between the input and the reconstructed output. The graph is also right-skewed, with 99.25% of all records between 0 and 0.001, and a high of 1.927. Finally, we combined the results from both the Heuristic Algorithm and Autoencoder to get the combined ranking.

By comparing the complete dataset with the top 0.1% records from the combined ranking, we found that potentially fraudulent properties are more concentrated in Manhattan and Staten Island, in buildings of class Z, Q or D, and on Tax Class 4, which are non-residence buildings. By checking the top 10 records with highest probability of fraud, we concluded that these properties are concentrated in Queens with significantly large size and either extremely higher or lower stories than average.

Part II. Description of Data

Dataset Name: Property Valuation and Assessment Data
File Name: NY property 1 million.xlsx
Data Sources: NYC Open Data-Department of Finance (DOF)
Dataset Overview:

- Category: Housing & Development
- Dimensions: 1,048,575 records and 30 variables
  - Categorical Variables: 13
  - Numeric Variables: 14
  - Text Variables: 2
  - Date Variable: 1
- Variables Names: RECORD, BBLE, BLOCK, LOT, EASEMENT, OWNER, BLDGCL, TAXCLASS, LTFRONT, LTDEPTH, STORIES, FULLVAL, AVLAND, AVTOT, EXLAND, EXTOT, EXCD1, STADDR, ZIP, EXMPTCL, BLDFRONT, BLDDEPTH, AVLAND2, EXLAND2, EXLAND2, EXTOT2, EXCD2, PERIOD, YEAR, VALTYPE

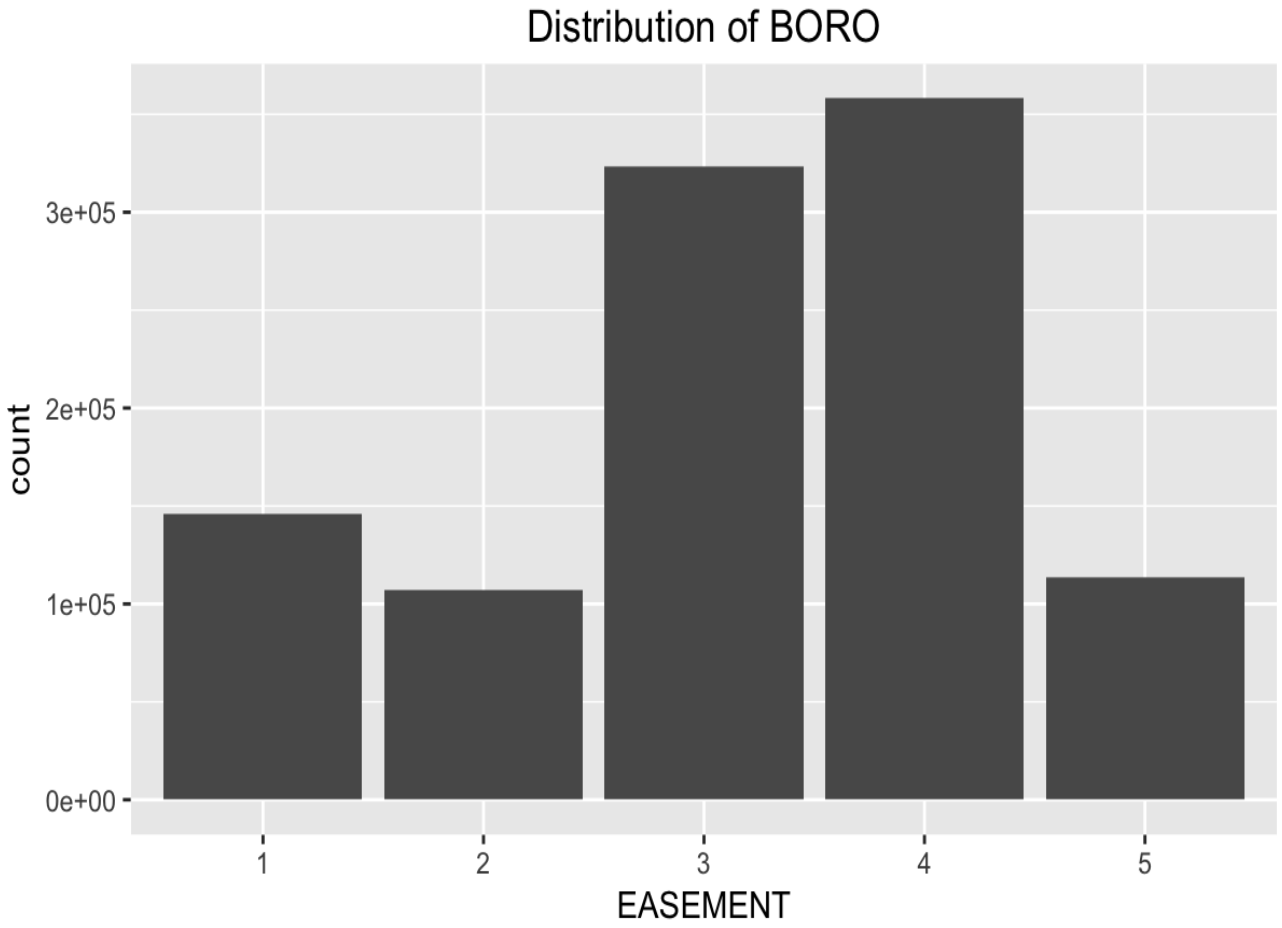
Most Important Variables in Consideration:

Name	Type	Description
RECORD	categorical	Unique Identifier of each record
BLOCK	categorical	Block Number Index
LOT	categorical	Unique number of the property within BORO/BLOCK
BLDGCL	categorical	Property Building class
TAXCLASS	categorical	Tax class
LTFRONT	numerical	Lot Frontage in feet
LTDEPTH	numerical	Lot Depth in feet
STORIES	numerical	Number of stories for the building
FULLVAL	numerical	Total market value of the property
AVLAND	numerical	Total Land Area
AVTOT	numerical	Assessed Value of the property
ZIP	categorical	Postal zip code of the property
BLDFRONT	numerical	Frontage in feet
BLDDEPTH	numerical	Depth in feet

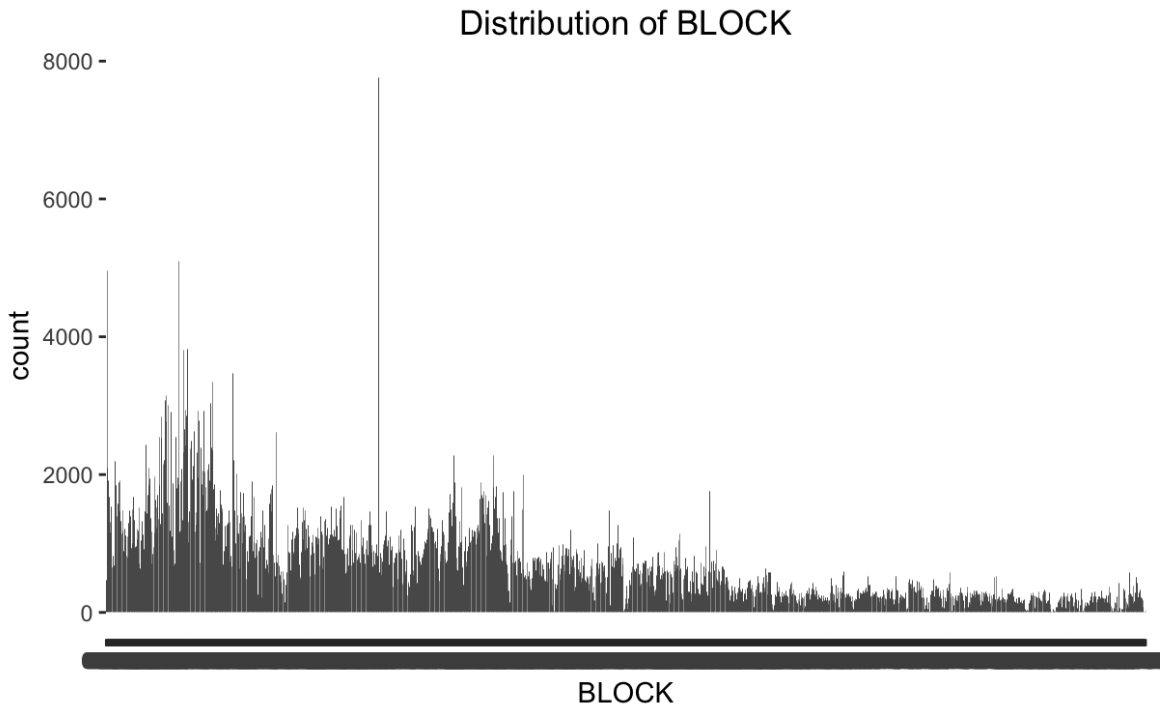
Distributions of each important variables: (the full Data Quality Report will be provided in the appendix)

Name	Description
RECORD	A unique key of record, from 1 to 1,048,575

Name	Description
BBLE	A file key to uniquely identify each record. Concatenation of BBLE_BORO, BBLE_BLOCK, BBLE_LOT, and BBLE_EASEMENT. (length: 11 alphanumeric)
BBLE_BORO	1 = MANHATTAN 2 = BRONX 3 = BROOKL YN 4 = QUEENS 5 = STATEN ISLAND

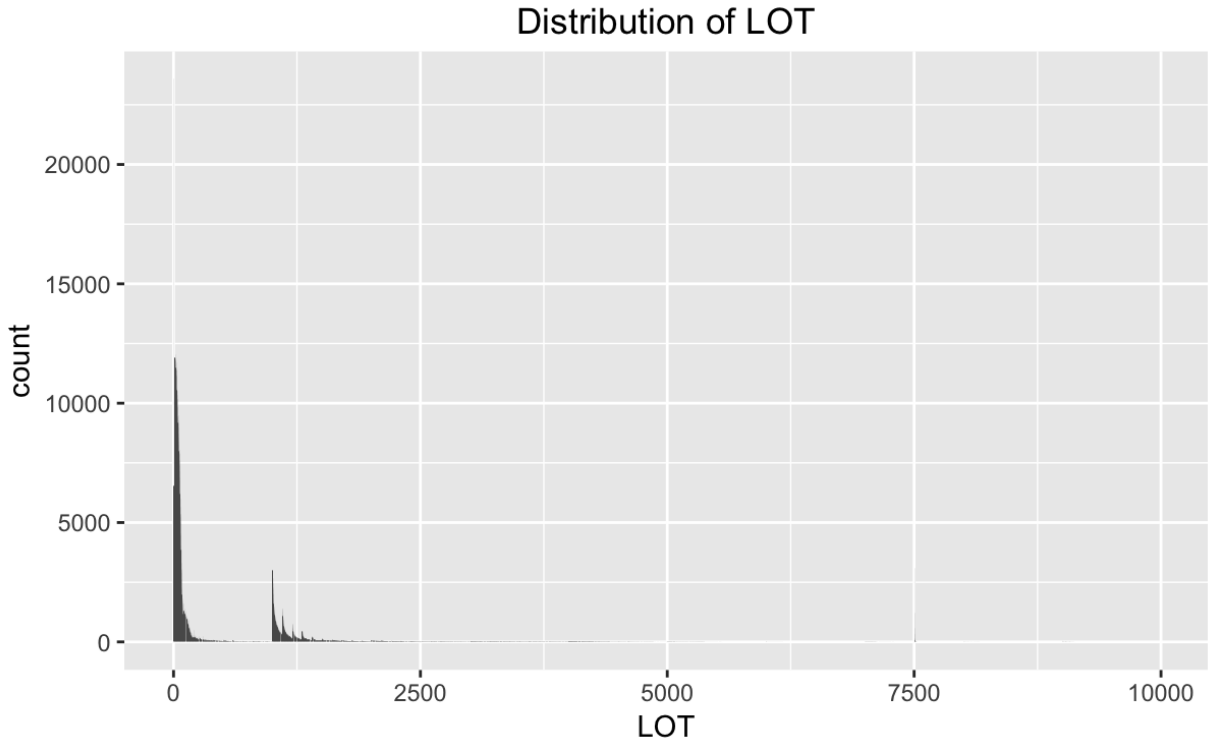


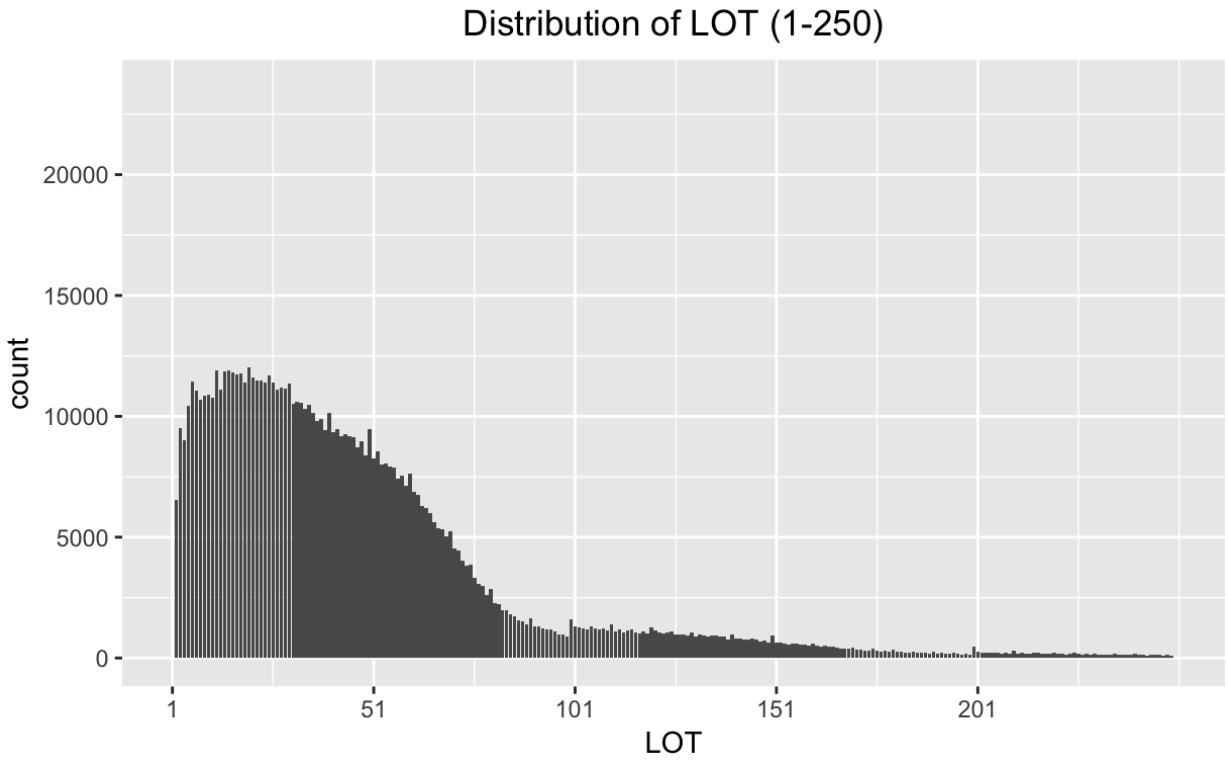
Name	Description
BLOCK	Valid block ranges by BORO. (length: 5 numeric)



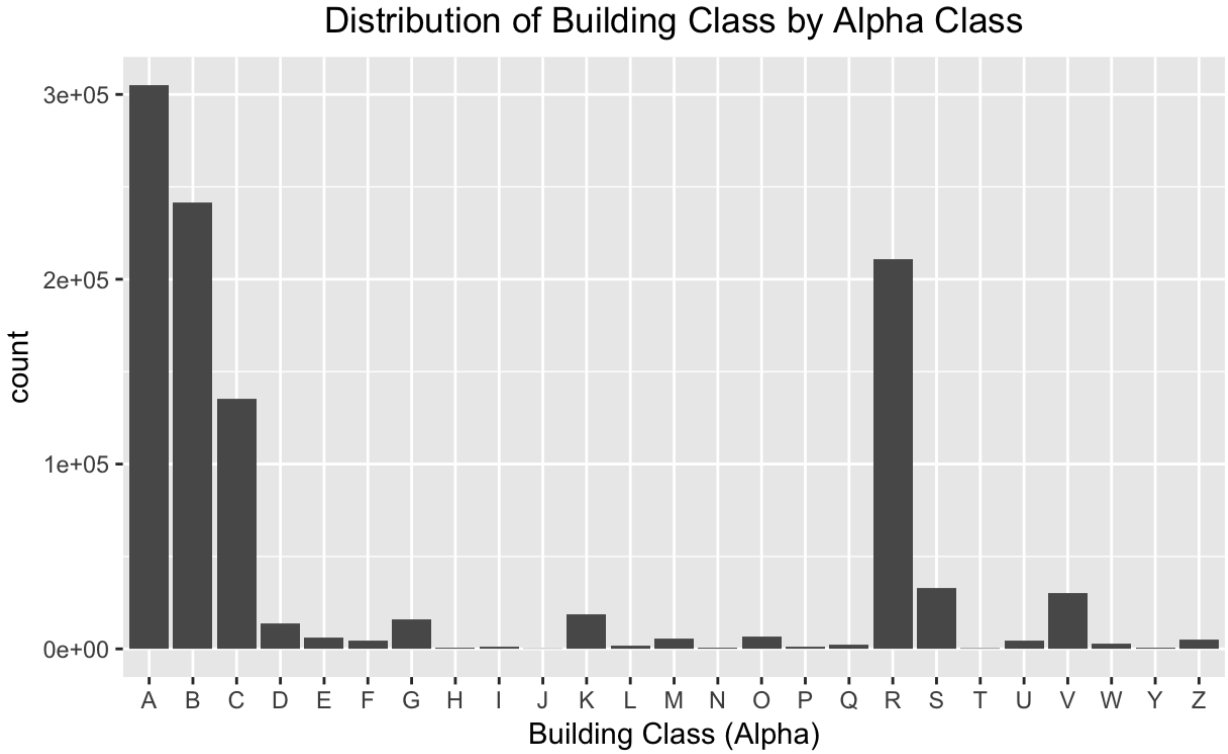
Name	Description
LOT	Unique # Within BORO/BLOCK. (length: 4 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*

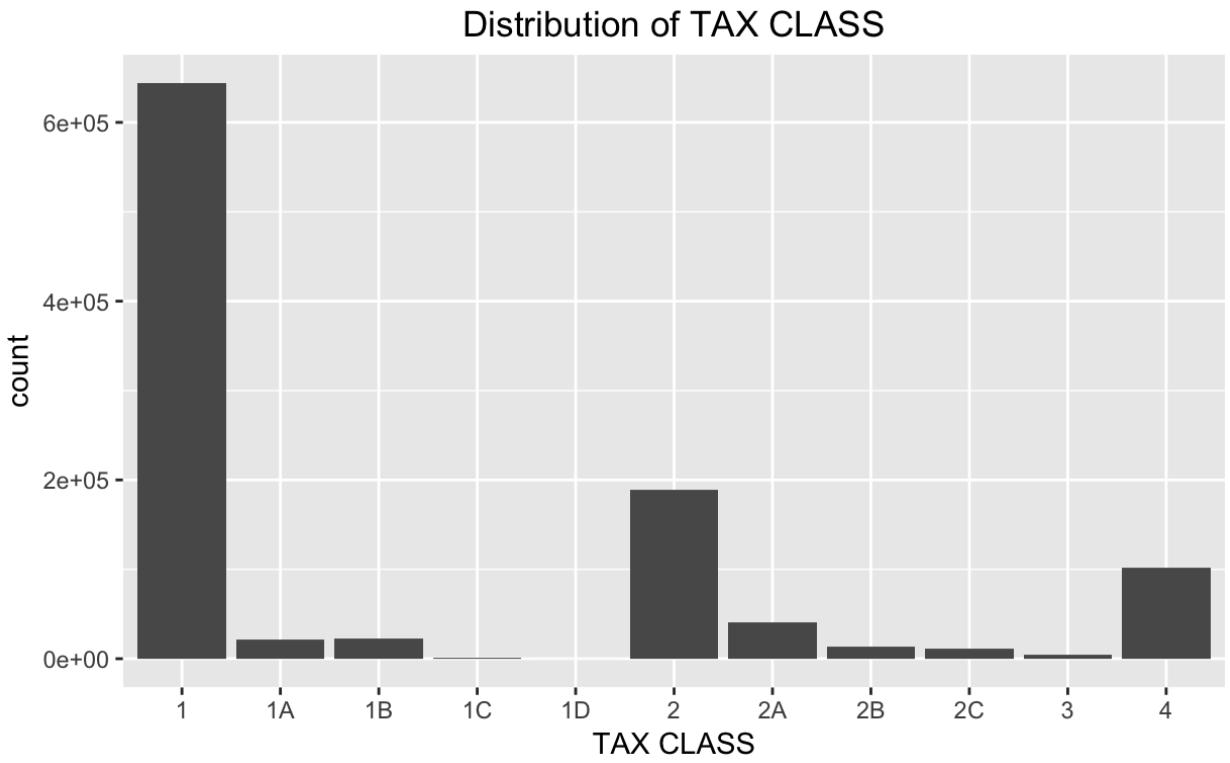




Name	Description
BLDGCL	Position 1 = Alphas & Position 2 = Numeric (length: 2 Character)

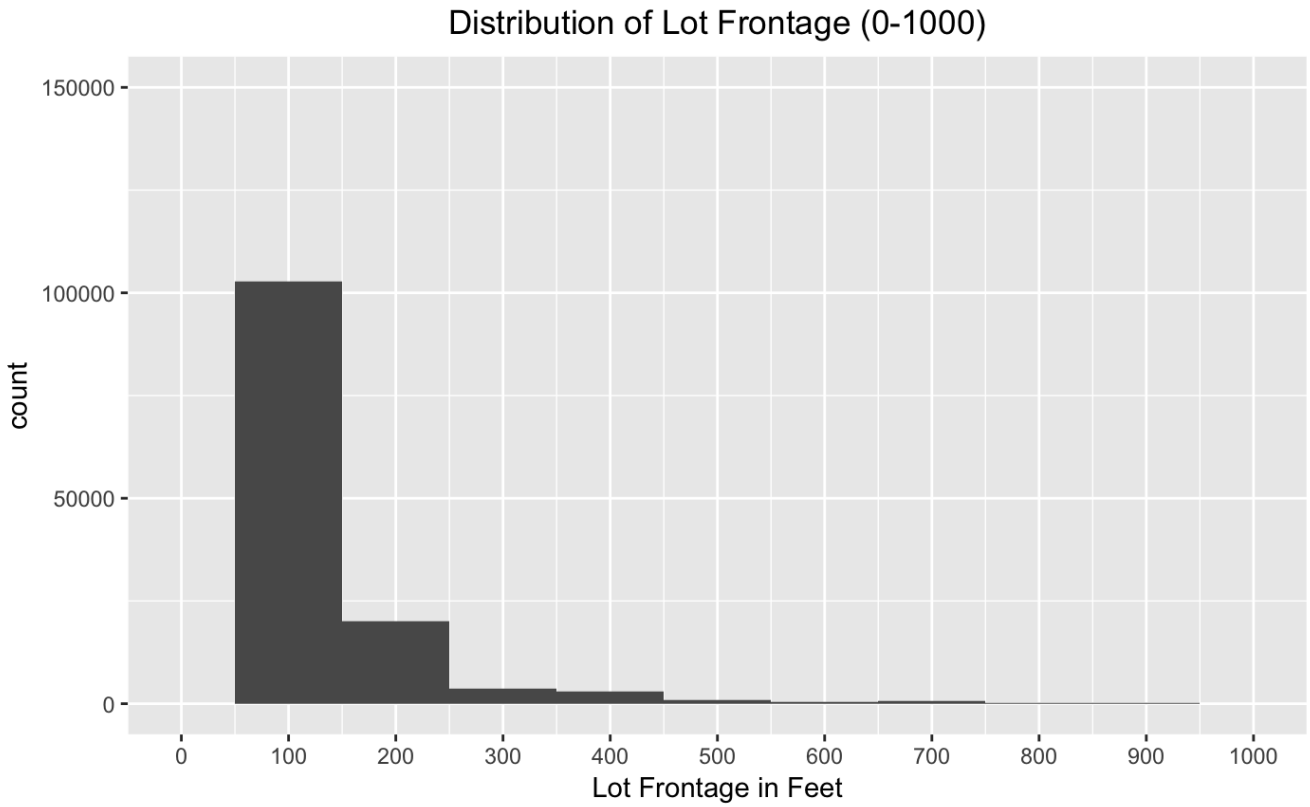


Name	Description
TAXCLASS	Current Property Tax Class Code (NYS Classification) 1 = 1-3-unit residences 1A = 1-3 story condominiums originally a condo 1B = residential vacant land 1C = unit condominiums originally tax class 1 1D = select bungalow colonies 2 = apartments 2A = apartments with 4-6 units 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

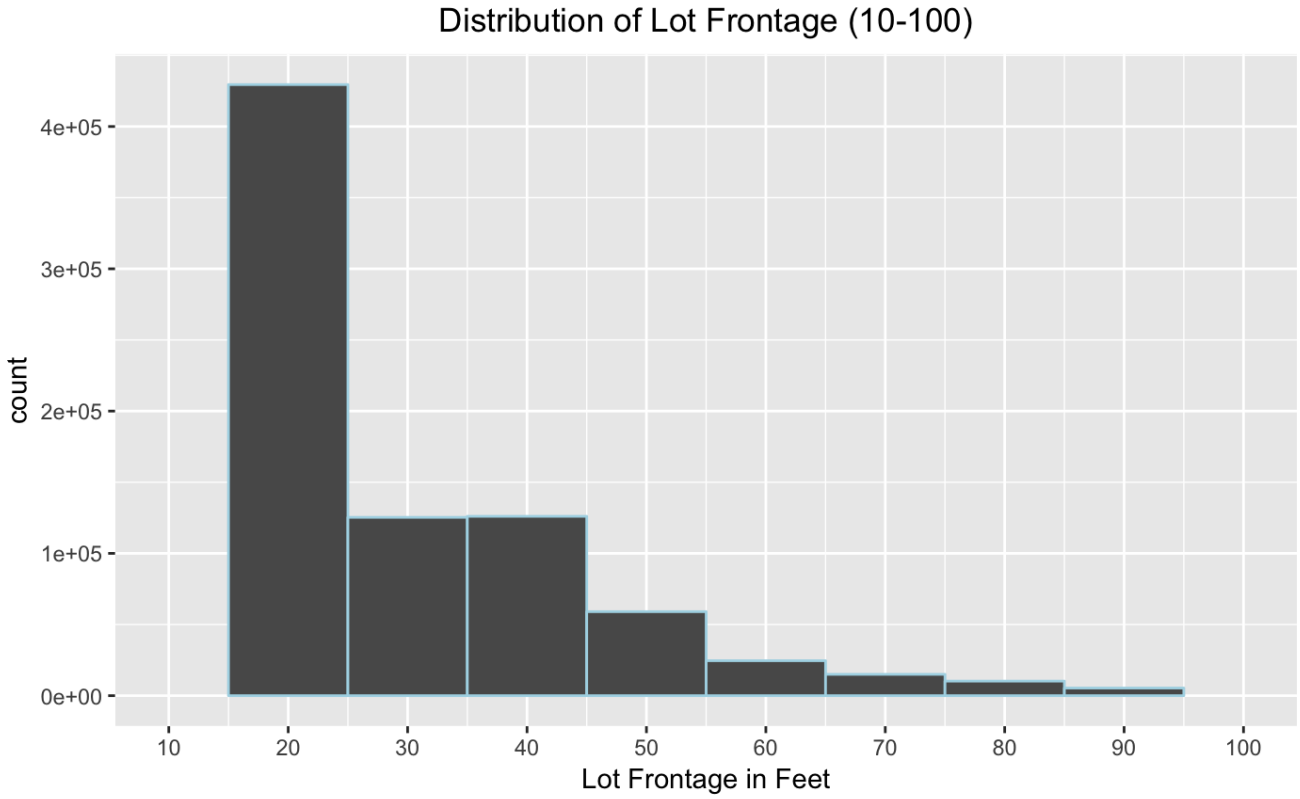


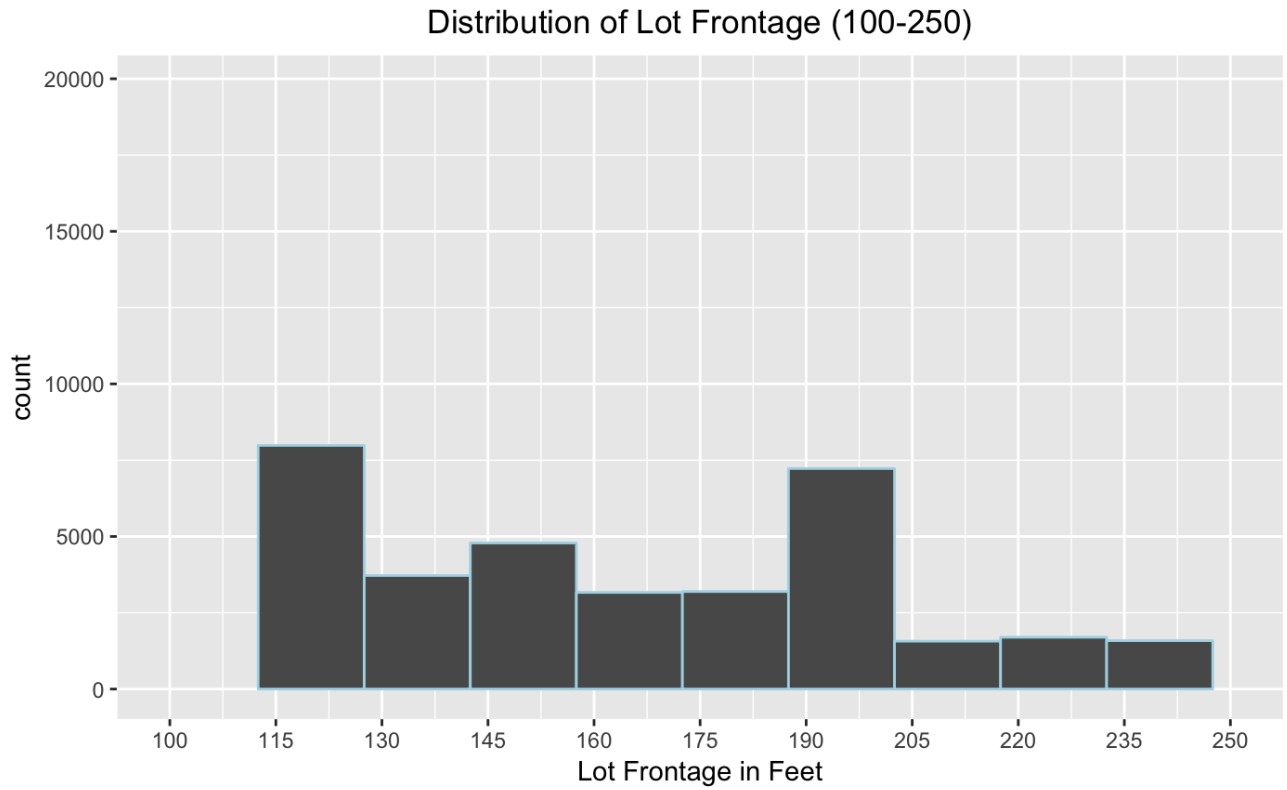
Name	Description
LTFRONT	Lot frontage in feet (length 7 numeric)

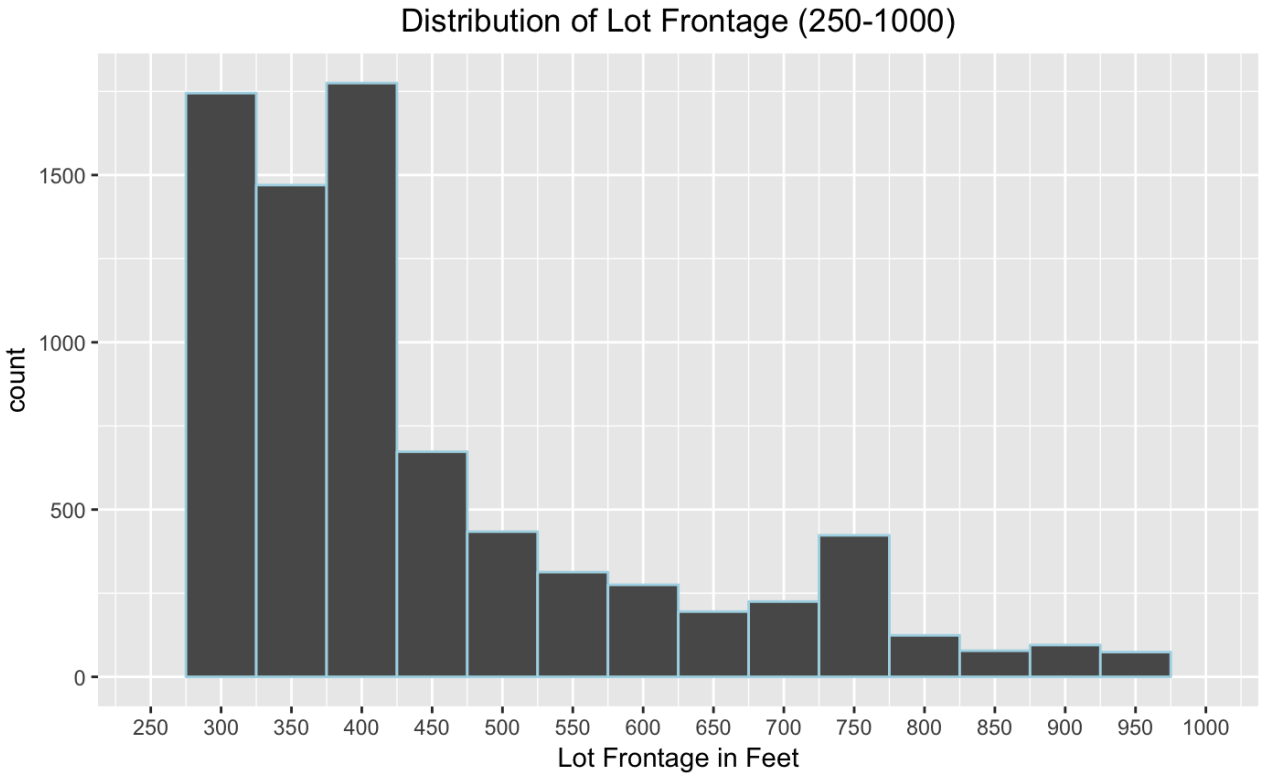
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*



Zoom in the distribution as shown figures on the next page:

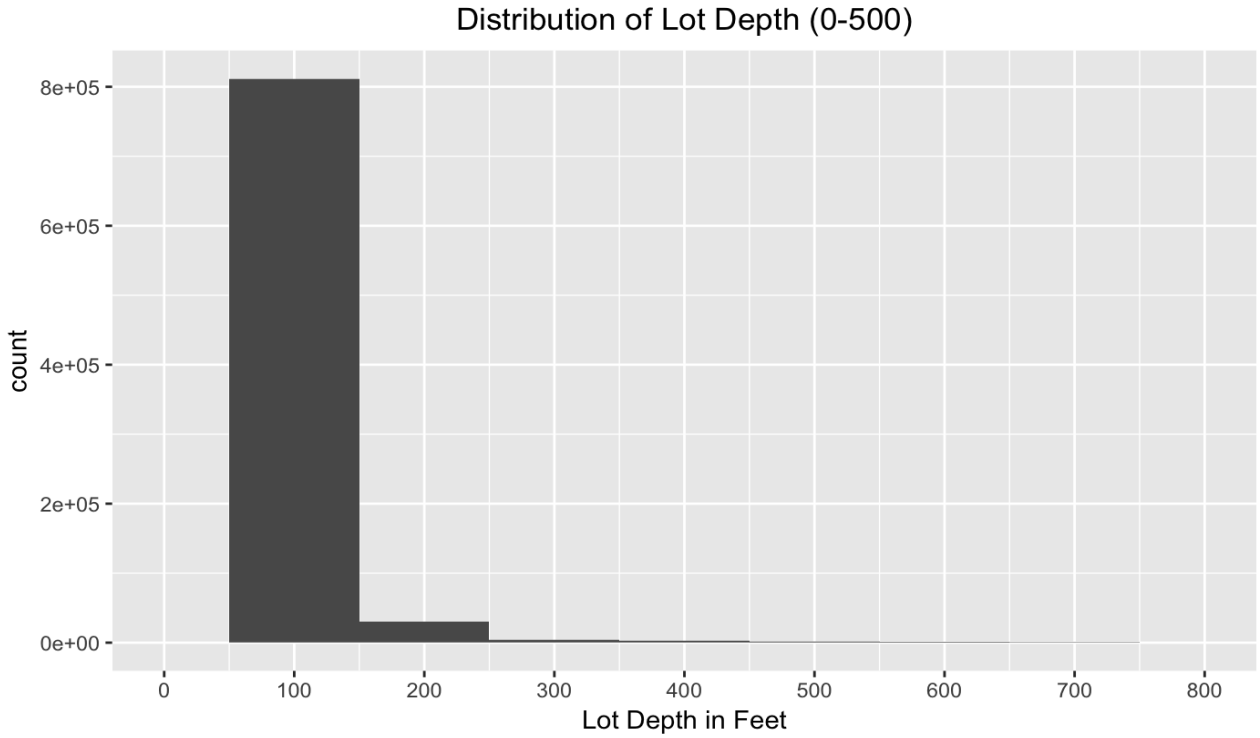



Distribution of Lot Frontage (100-250)
count
20000
15000
10000
5000
0
100
115
130
145
160
175
190
205
220
235
250
Lot Frontage in Feet

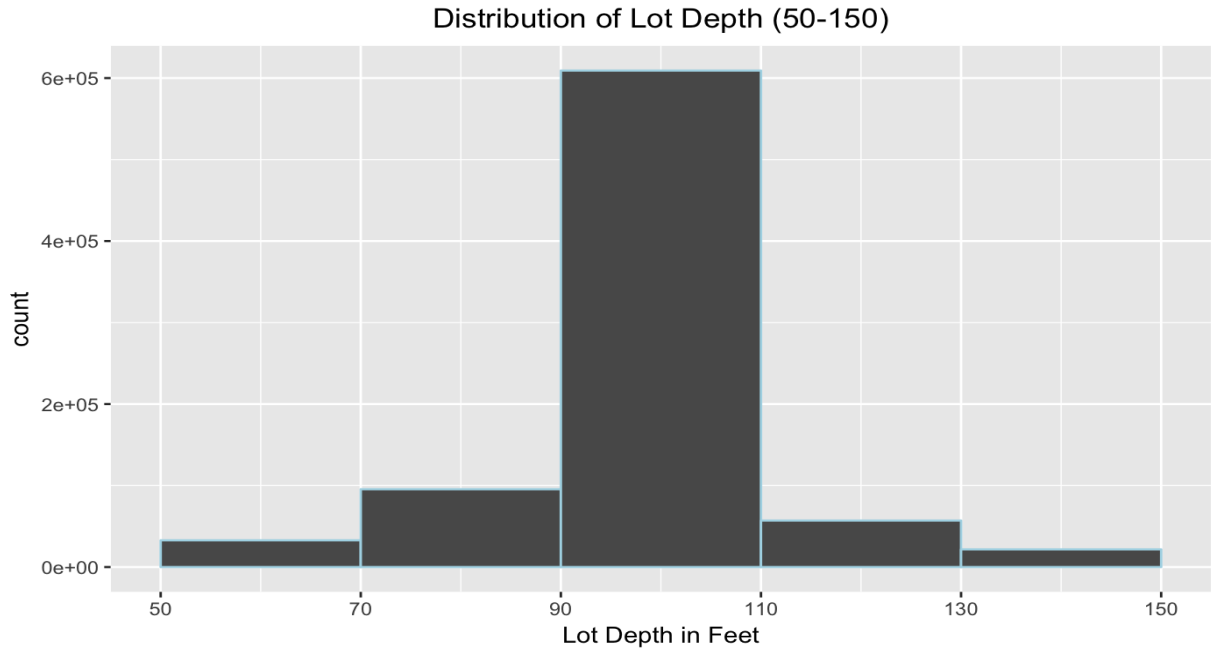

Distribution of Lot Frontage (250-1000)
count
1500
1000
500
0
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000
Lot Frontage in Feet

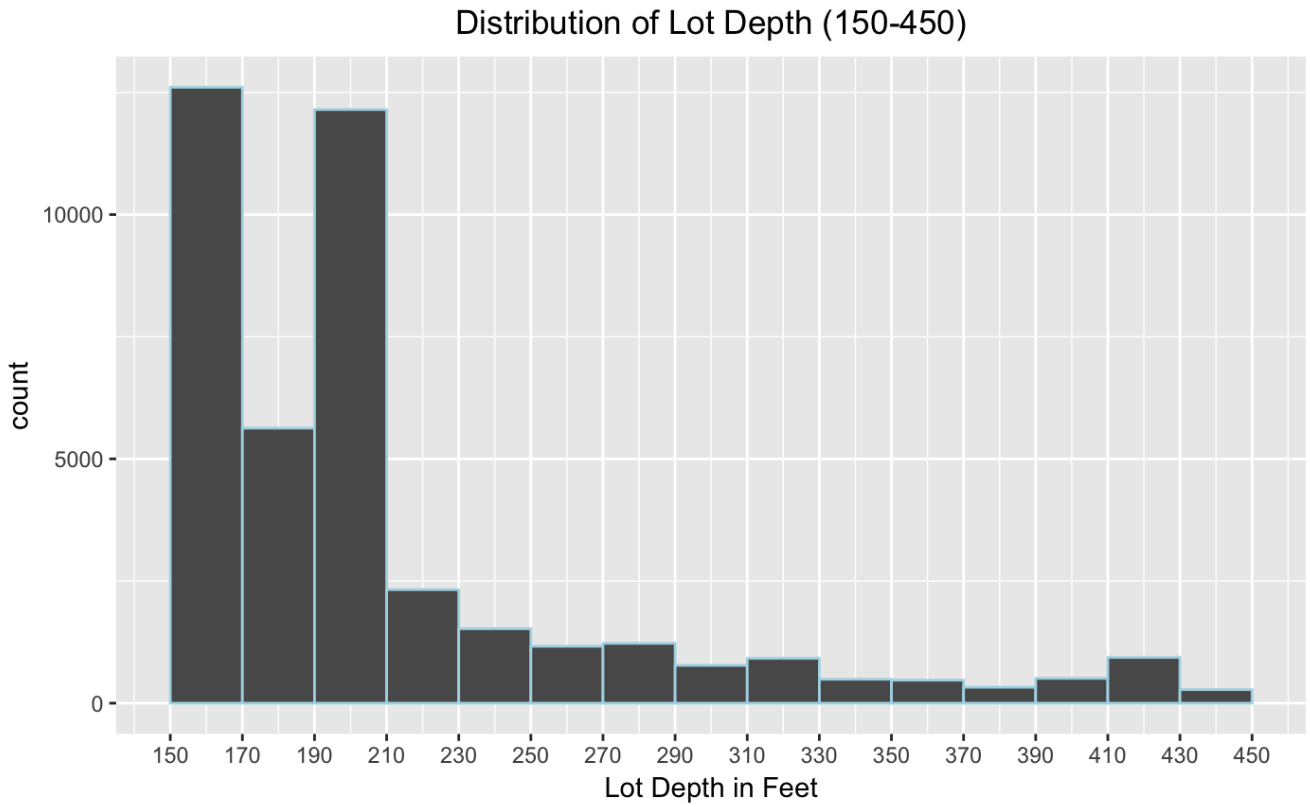
Name	Description
LTDEPTH	Lot depth in feet (length 7 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-interval is selected to specify partial distribution.*



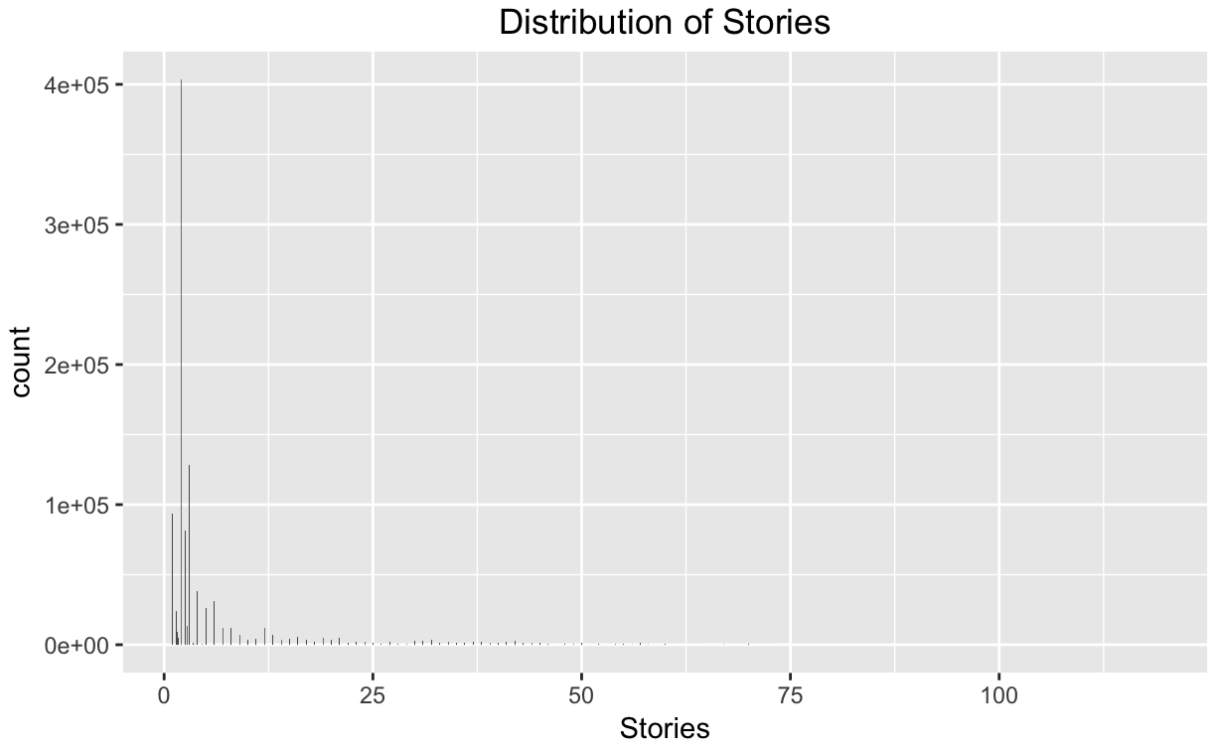
Zoom in the distribution as shown figures below:



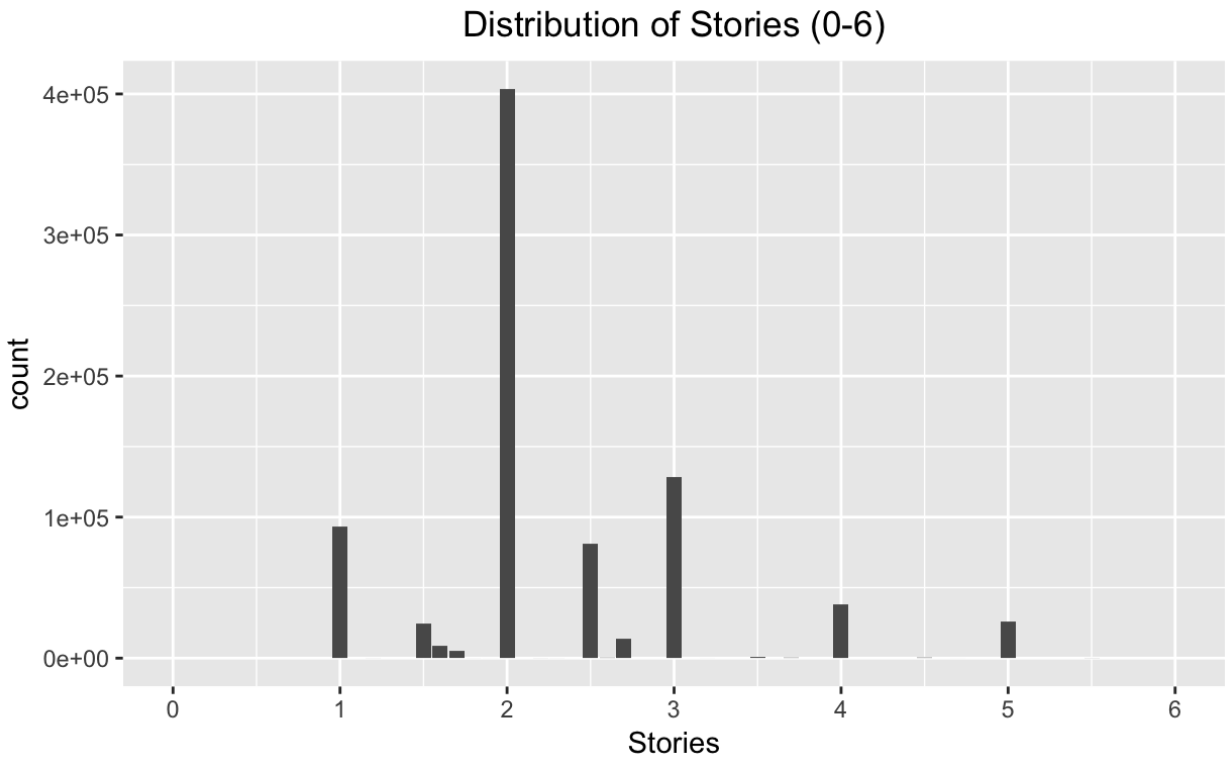


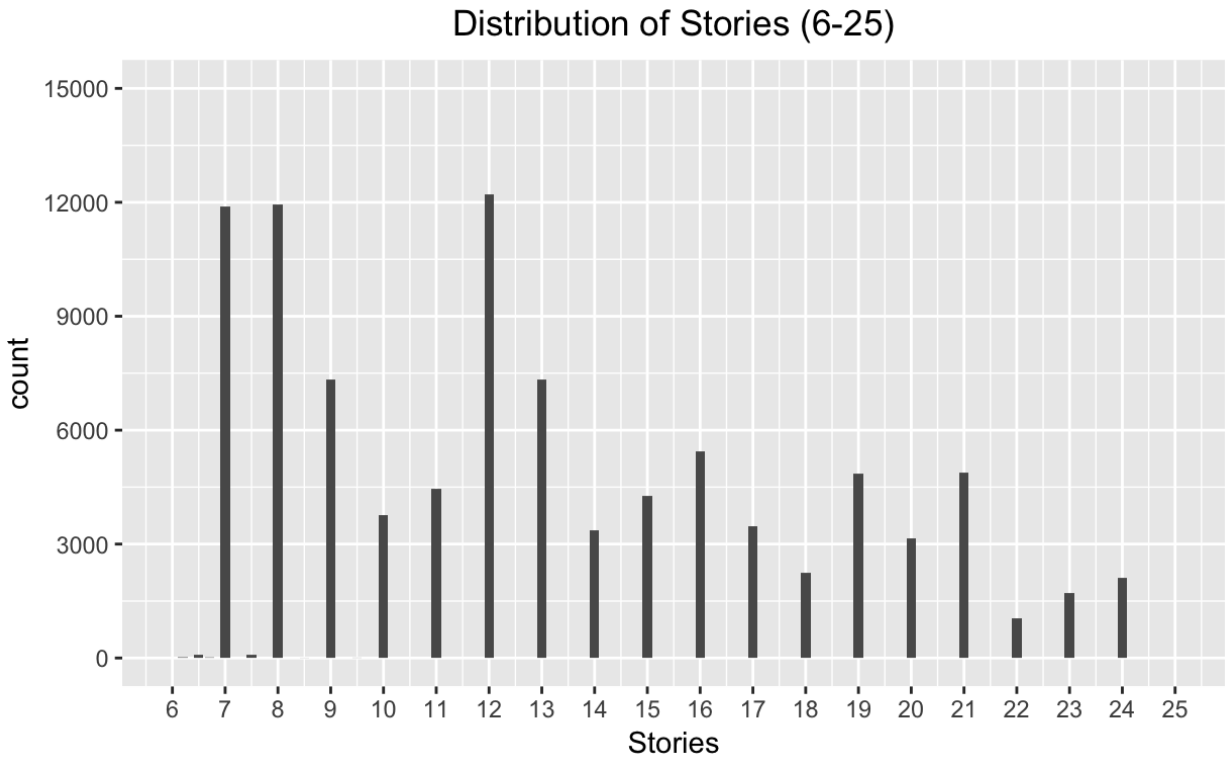
Name	Description
STORIES	The number of stories for the building (# of Floors). (length 5 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*



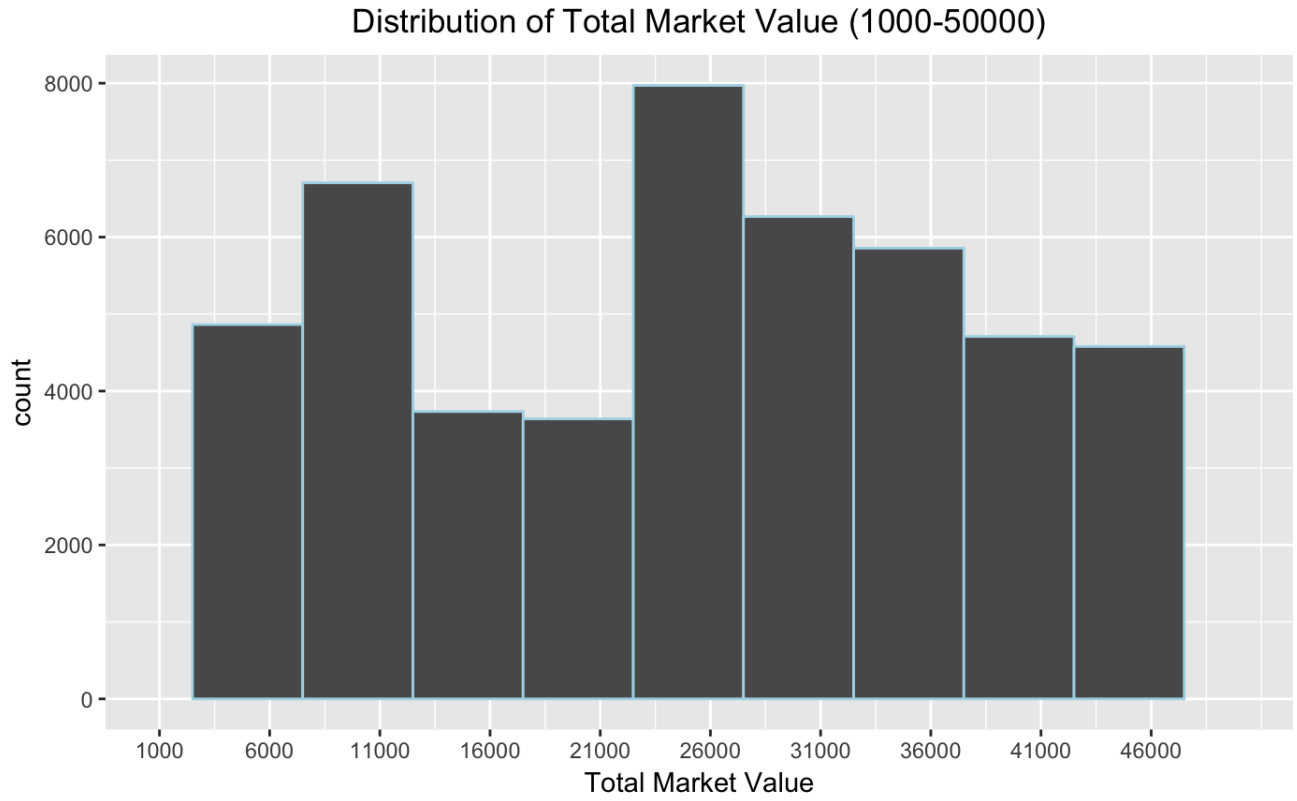
Zoom in the distribution as shown figures below:

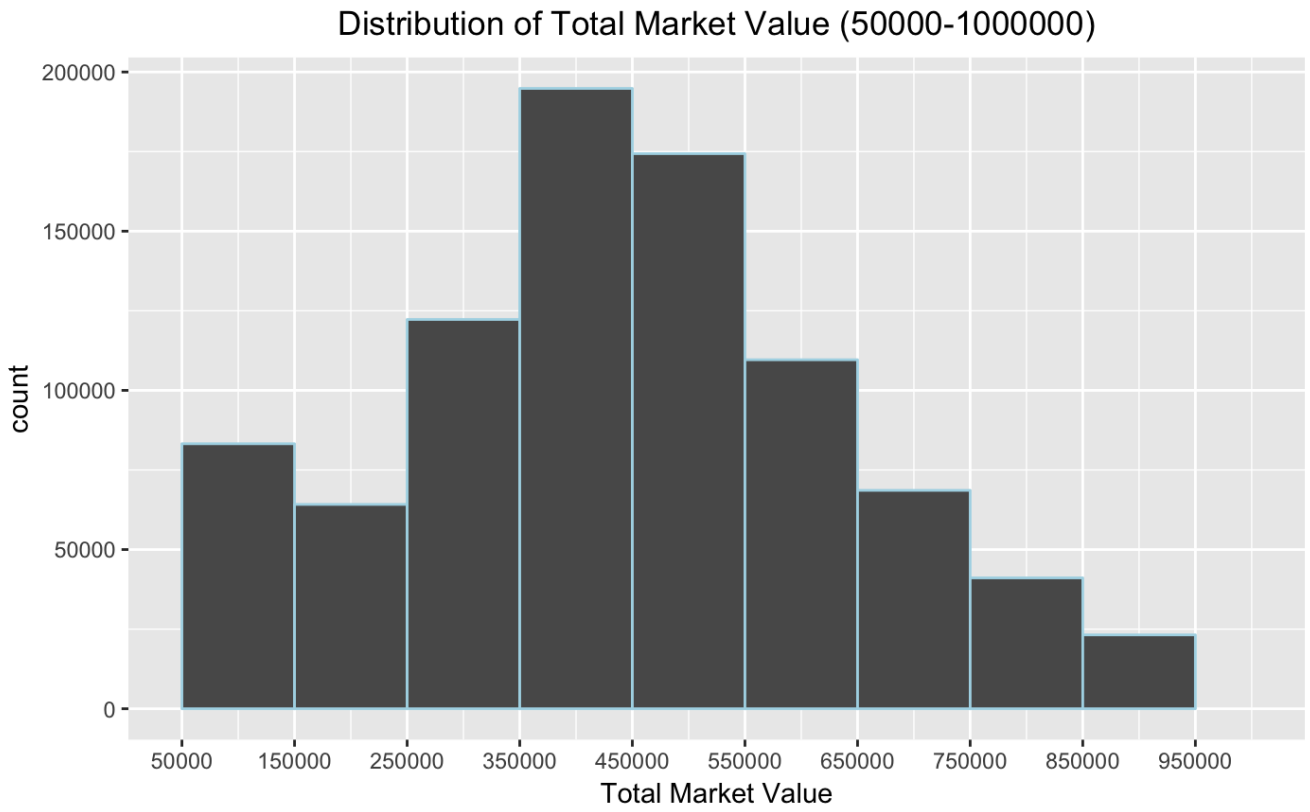


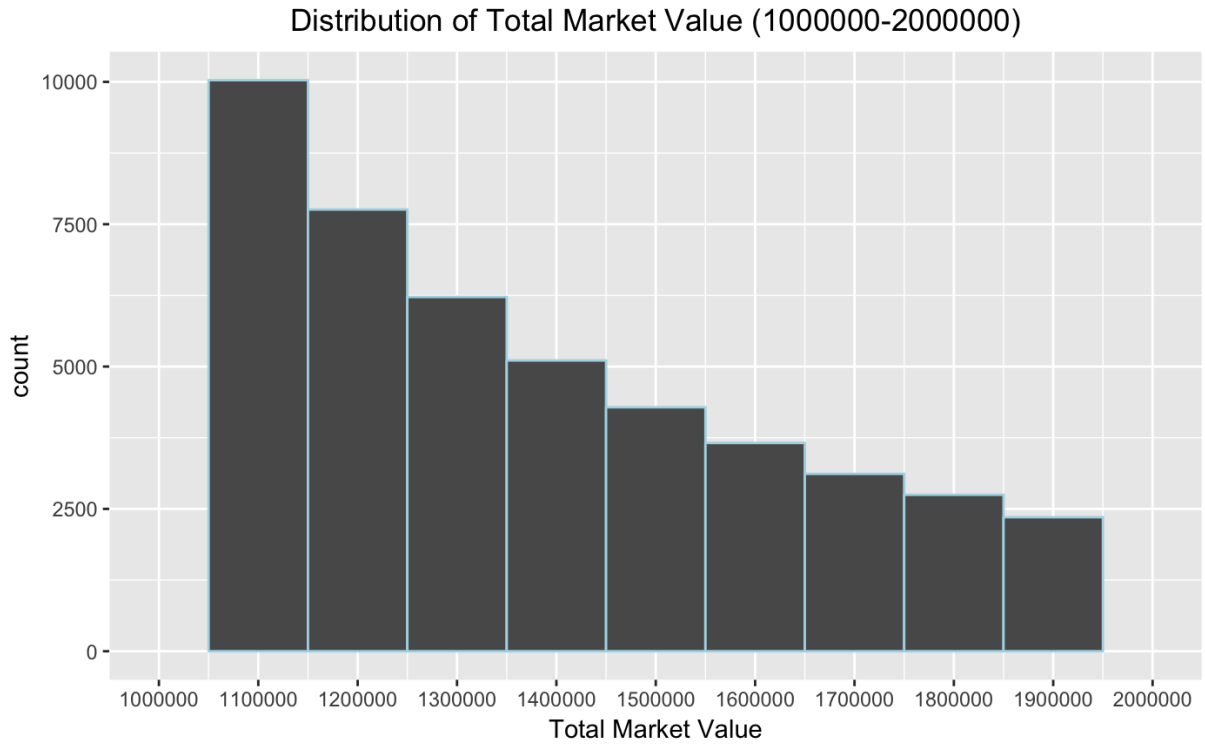


Name	Description
FULLVAL	Total market value (length 11 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*

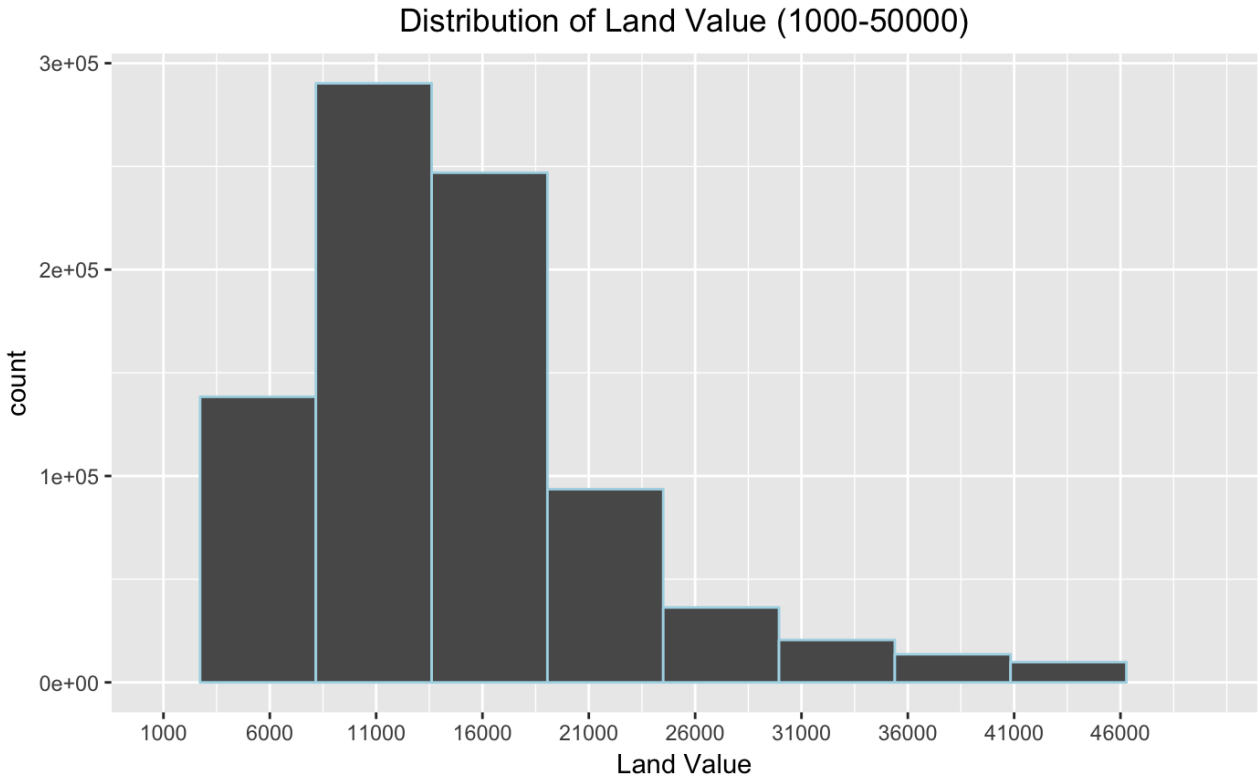


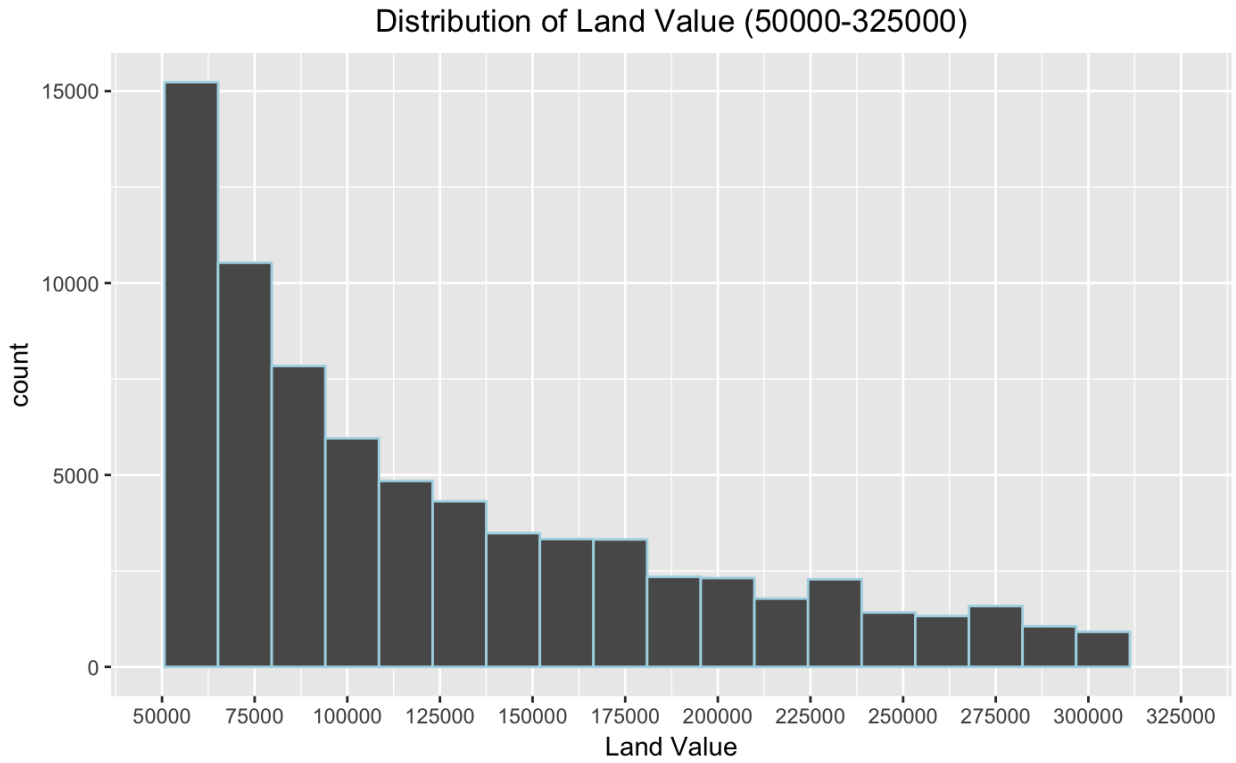




Name	Description
AVLAND	Market value of the land (length 11 numeric)

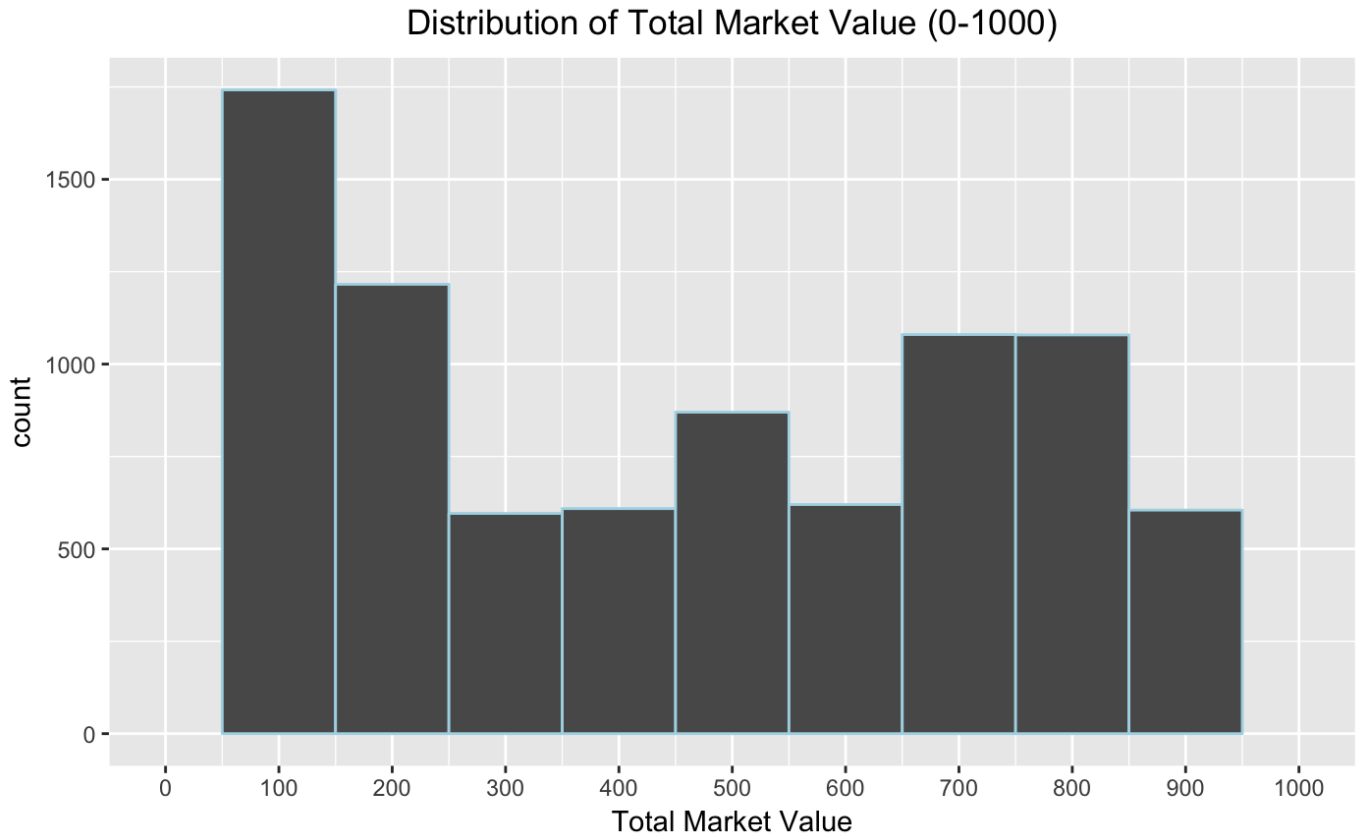
Note: This variable exhibits severe right-skew distribution. Sub-interval is selected to specify partial distribution.*

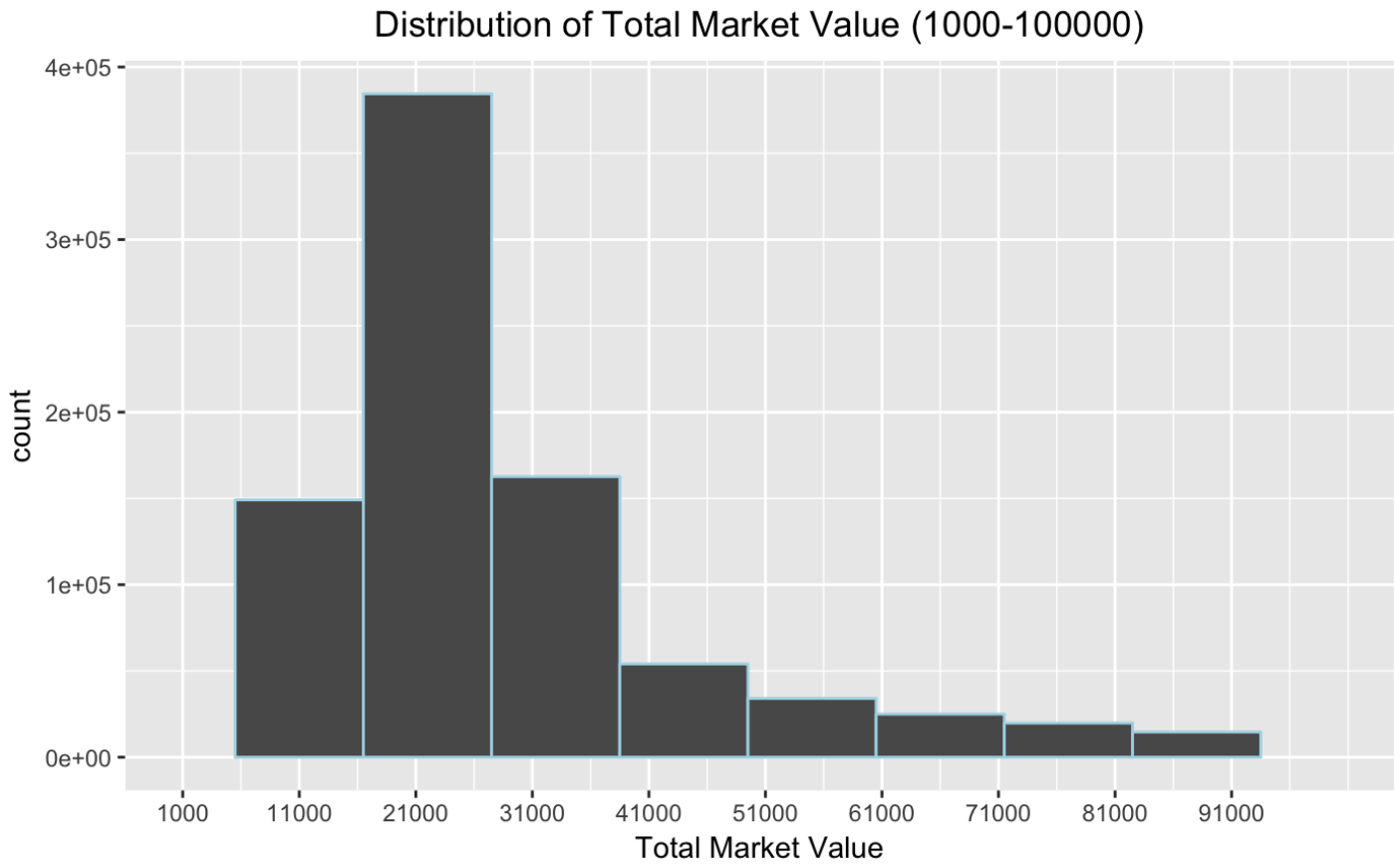


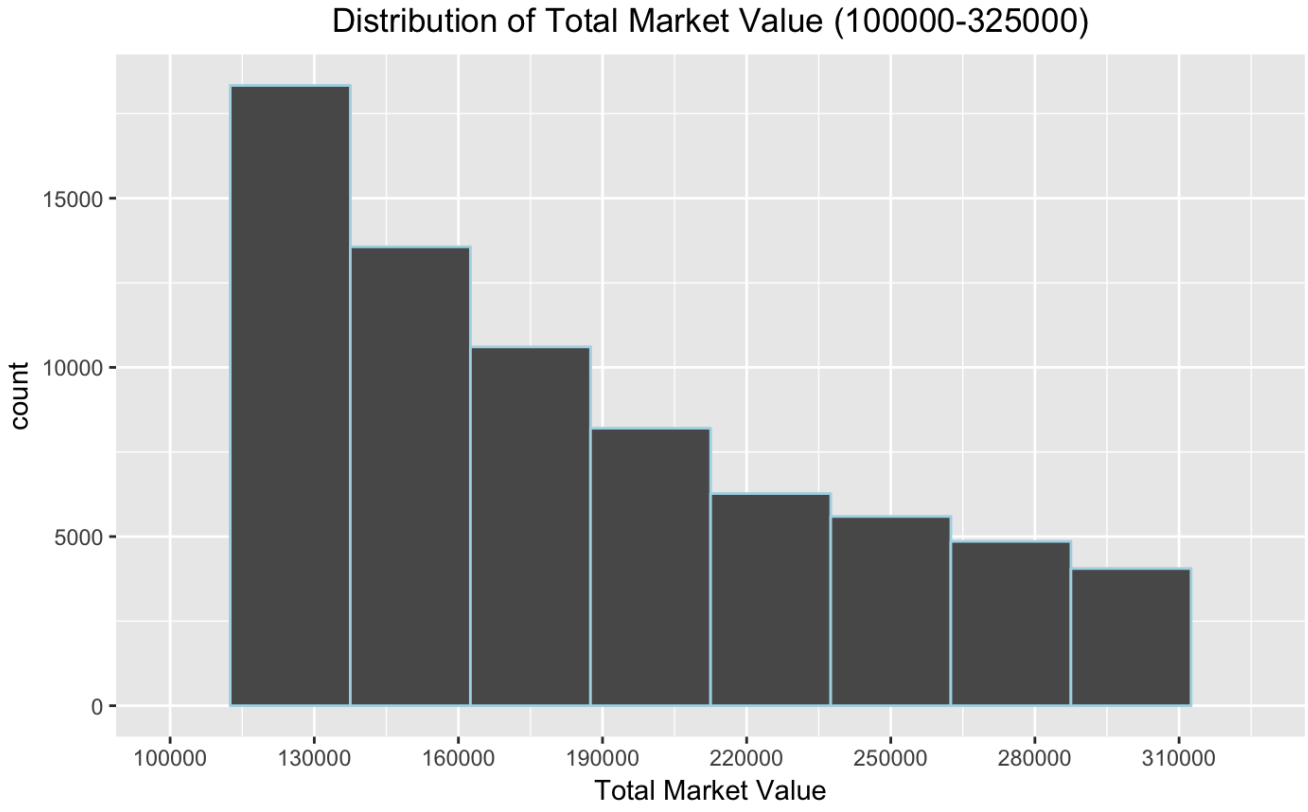


Name	Description
AVTOT	Current year's total market value (length 11 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*

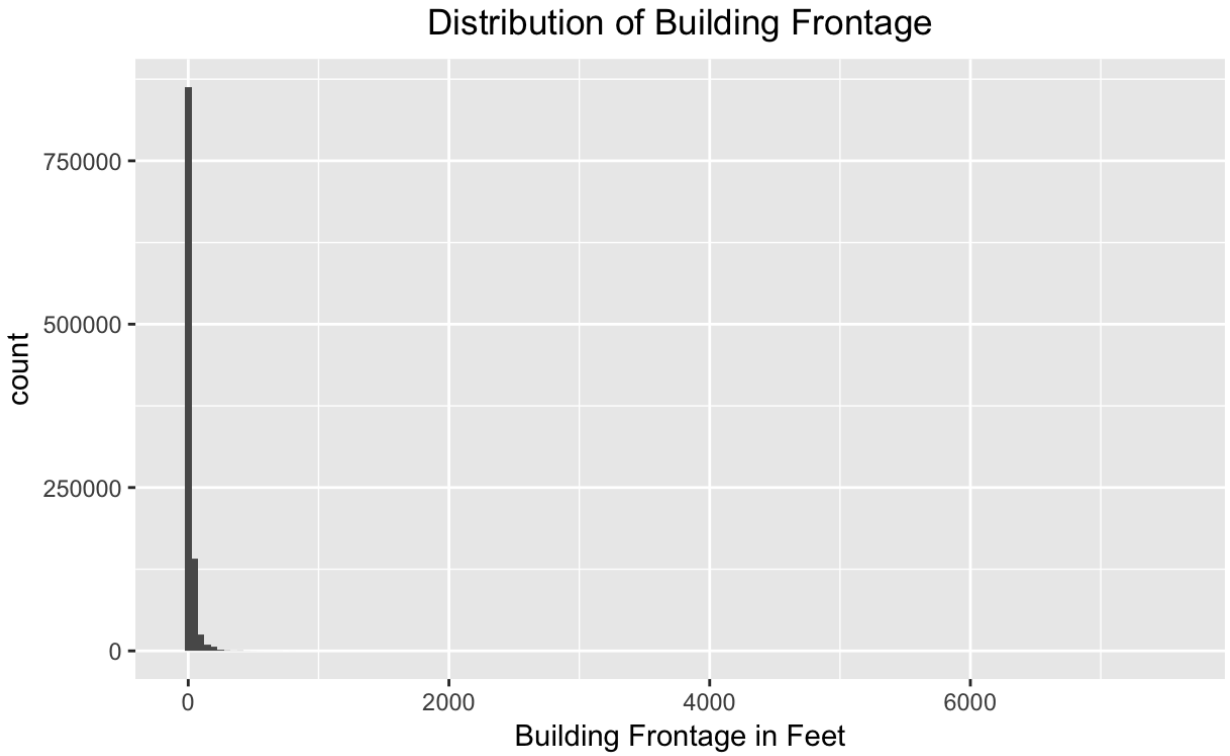


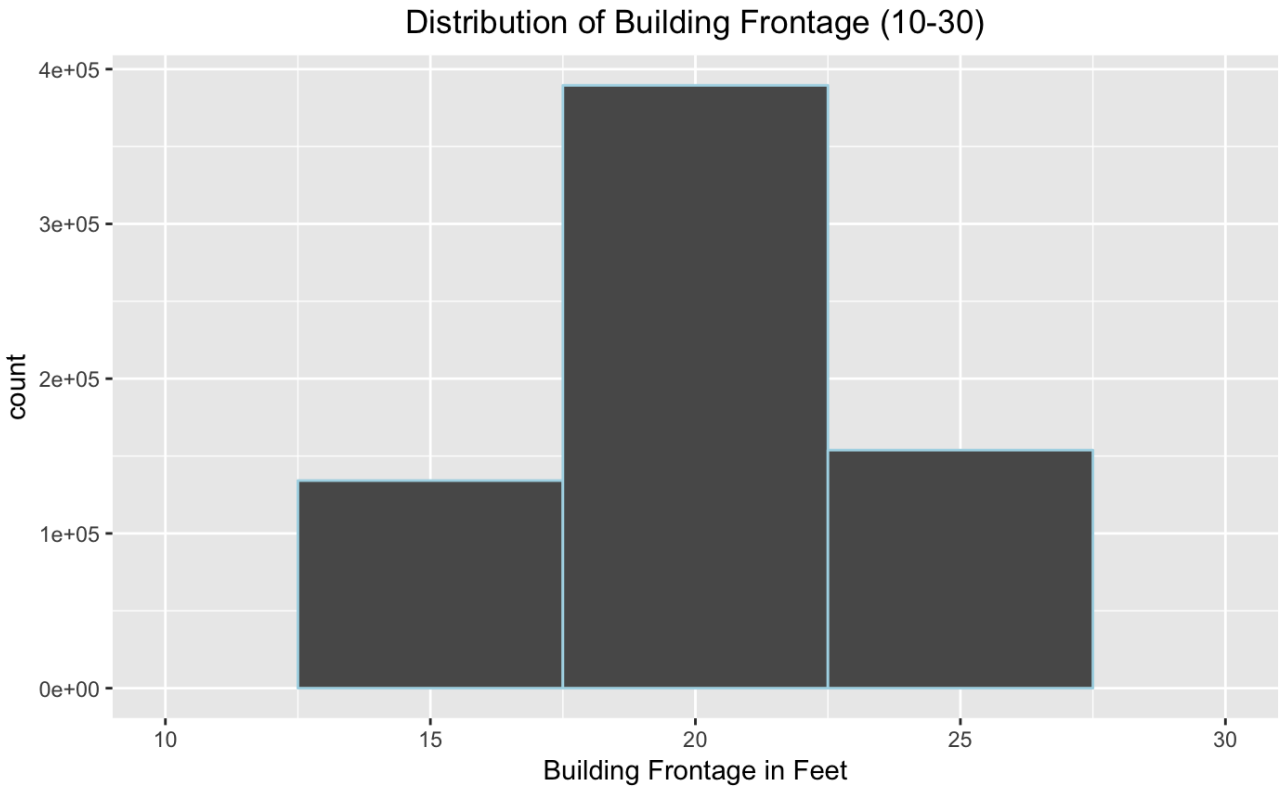

Distribution of Total Market Value (1000-100000)
4e+05
3e+05
2e+05
1e+05
0e+00
count
1000
11000
21000
31000
41000
51000
61000
71000
81000
91000
Total Market Value

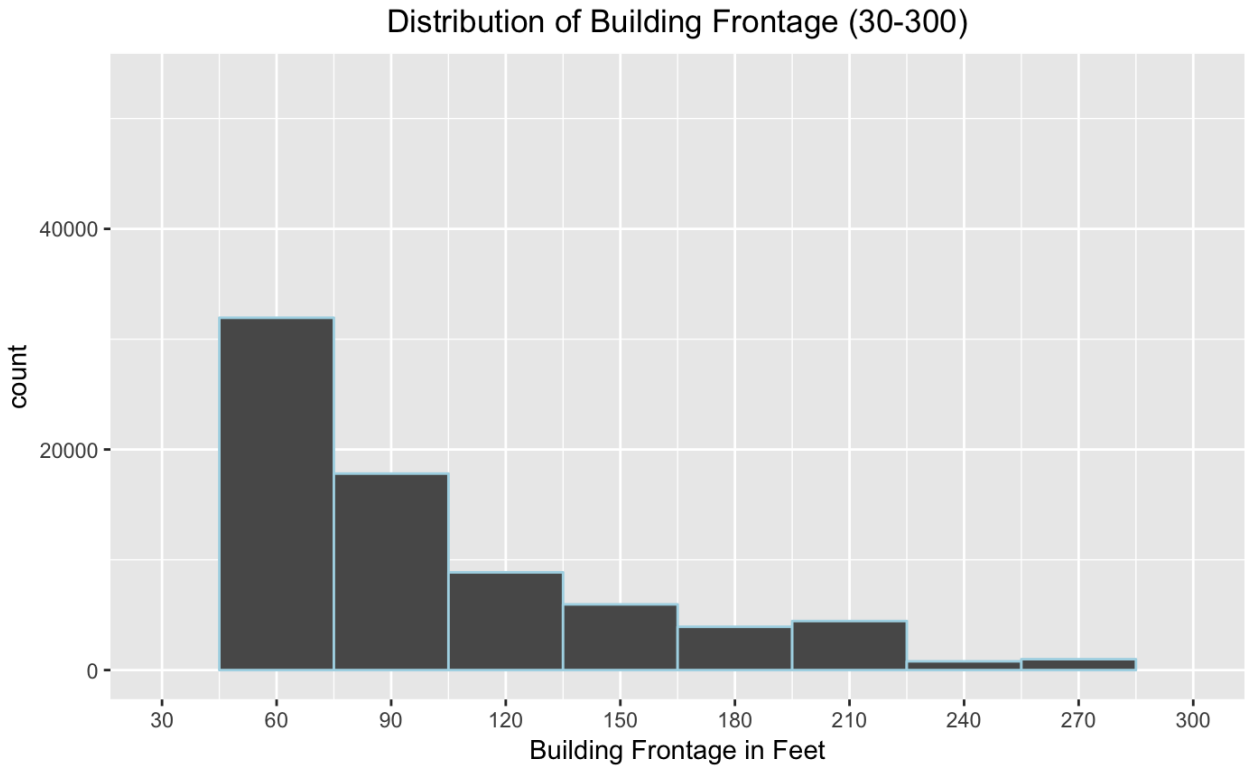

Distribution of Total Market Value (100000-325000)
15000
10000
5000
0
count
100000
130000
160000
190000
220000
250000
280000
310000
Total Market Value

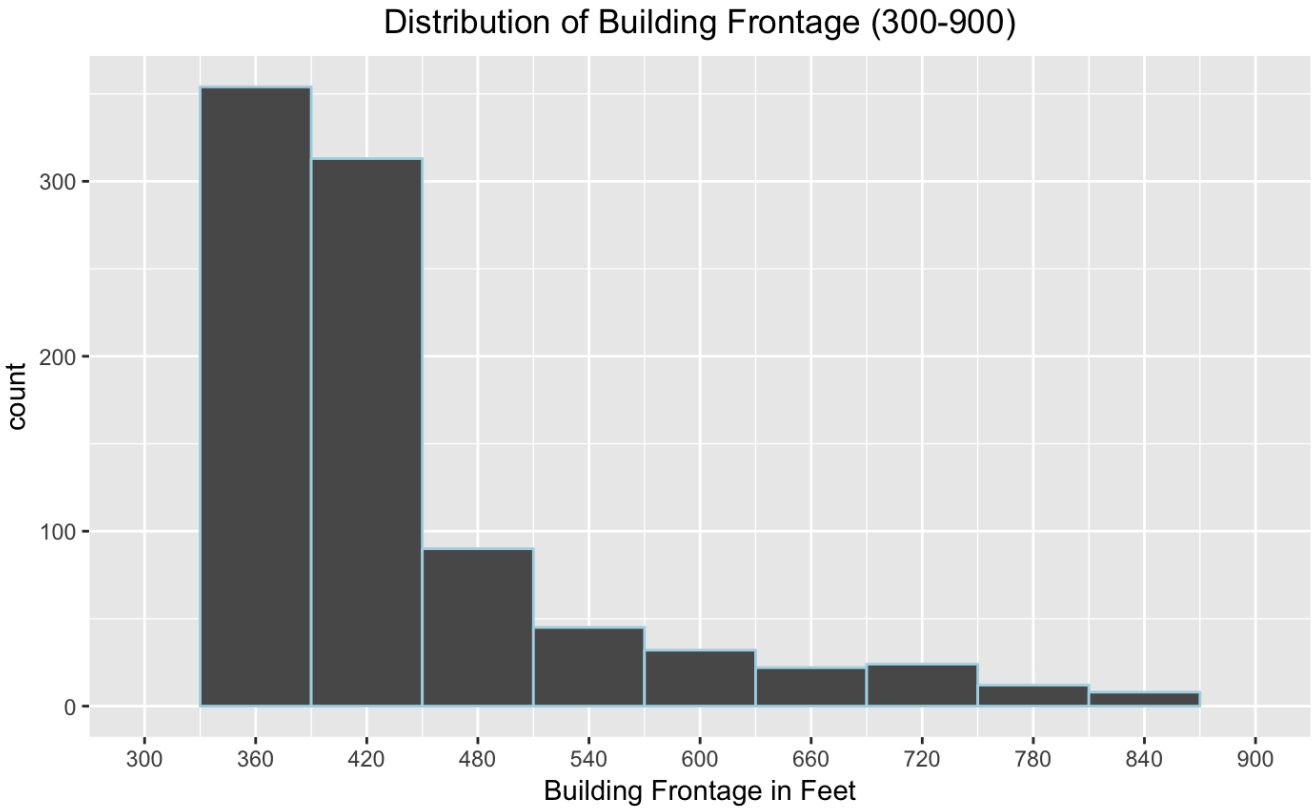
Name	Description
BLDFRONT	Building frontage in feet (length 7 numeric)

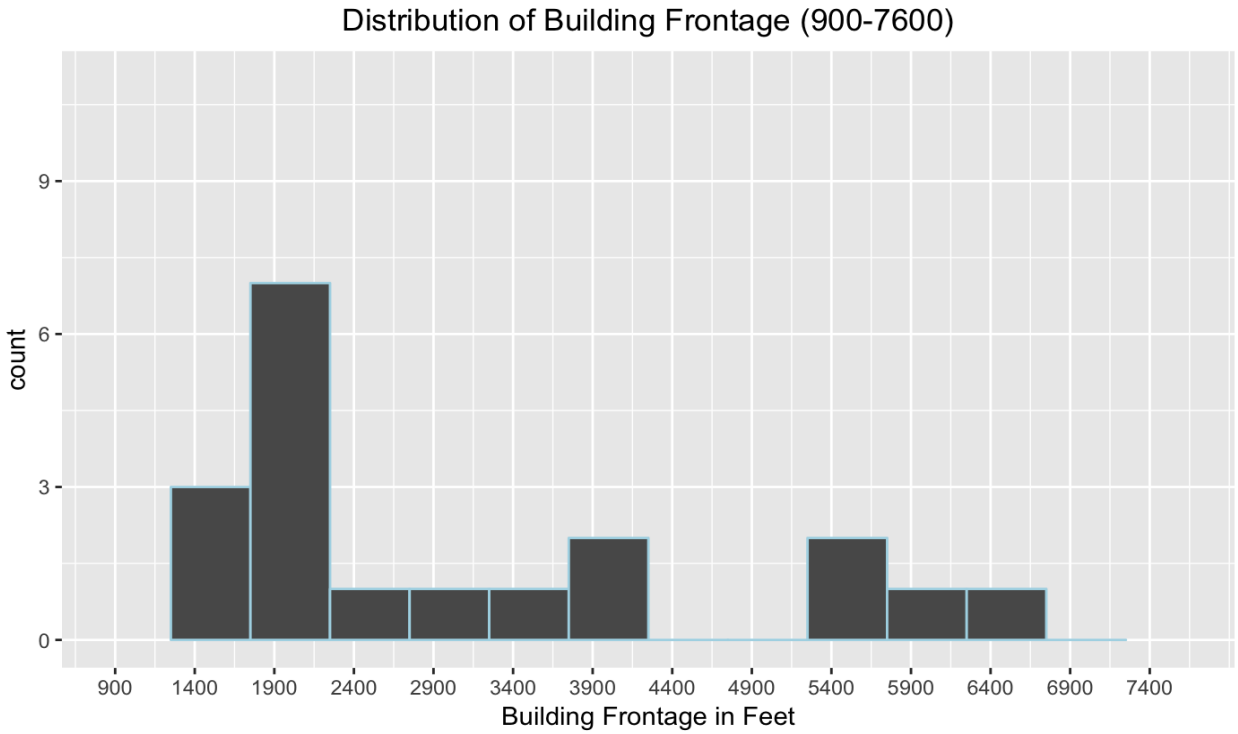
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution*





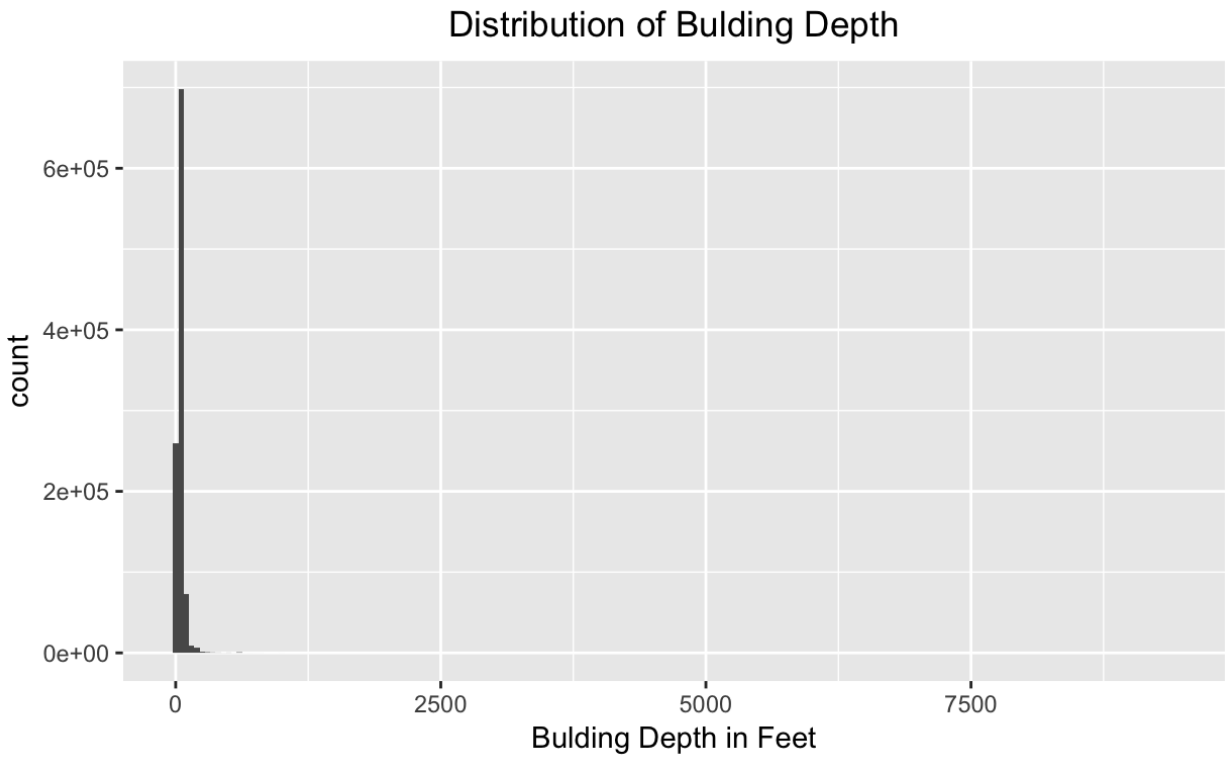

Distribution of Building Frontage (30-300)
count
40000
20000
0
30
60
90
120
150
180
210
240
270
300
Building Frontage in Feet

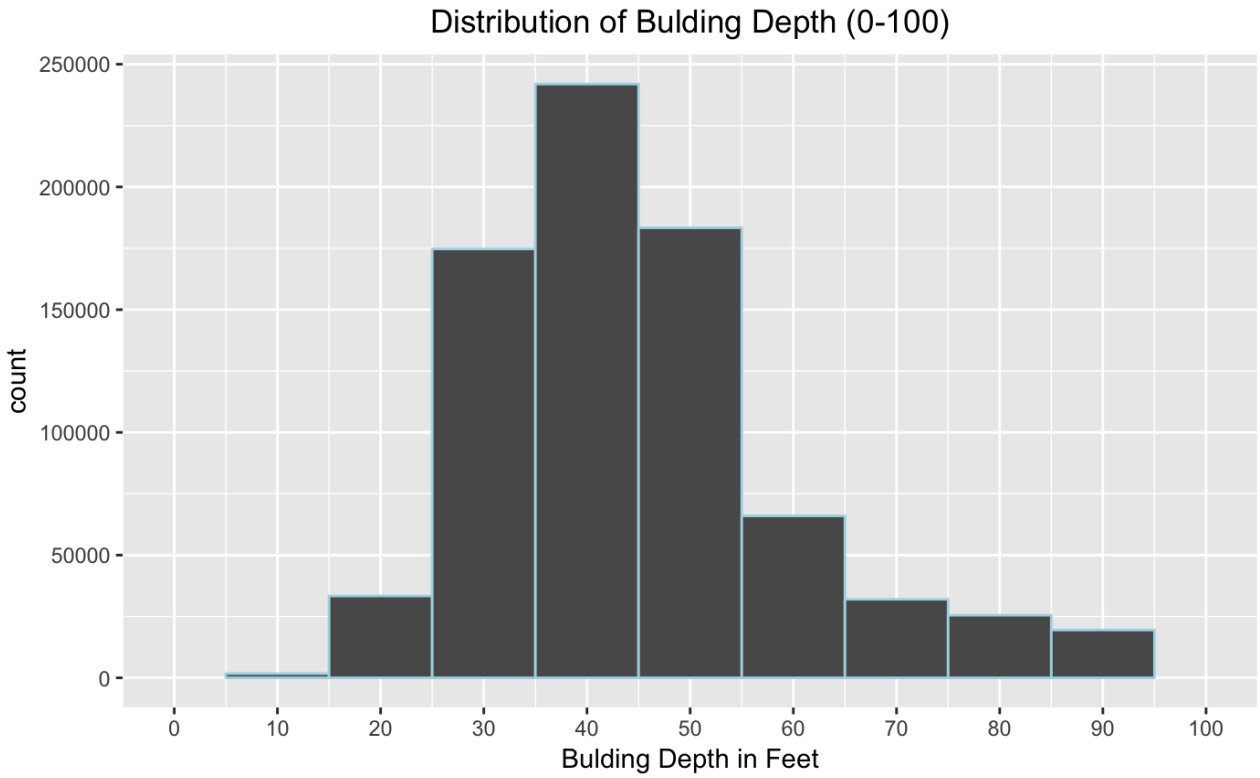

Distribution of Building Frontage (300-900)
count
300
200
100
0
300
360
420
480
540
600
660
720
780
840
900
Building Frontage in Feet

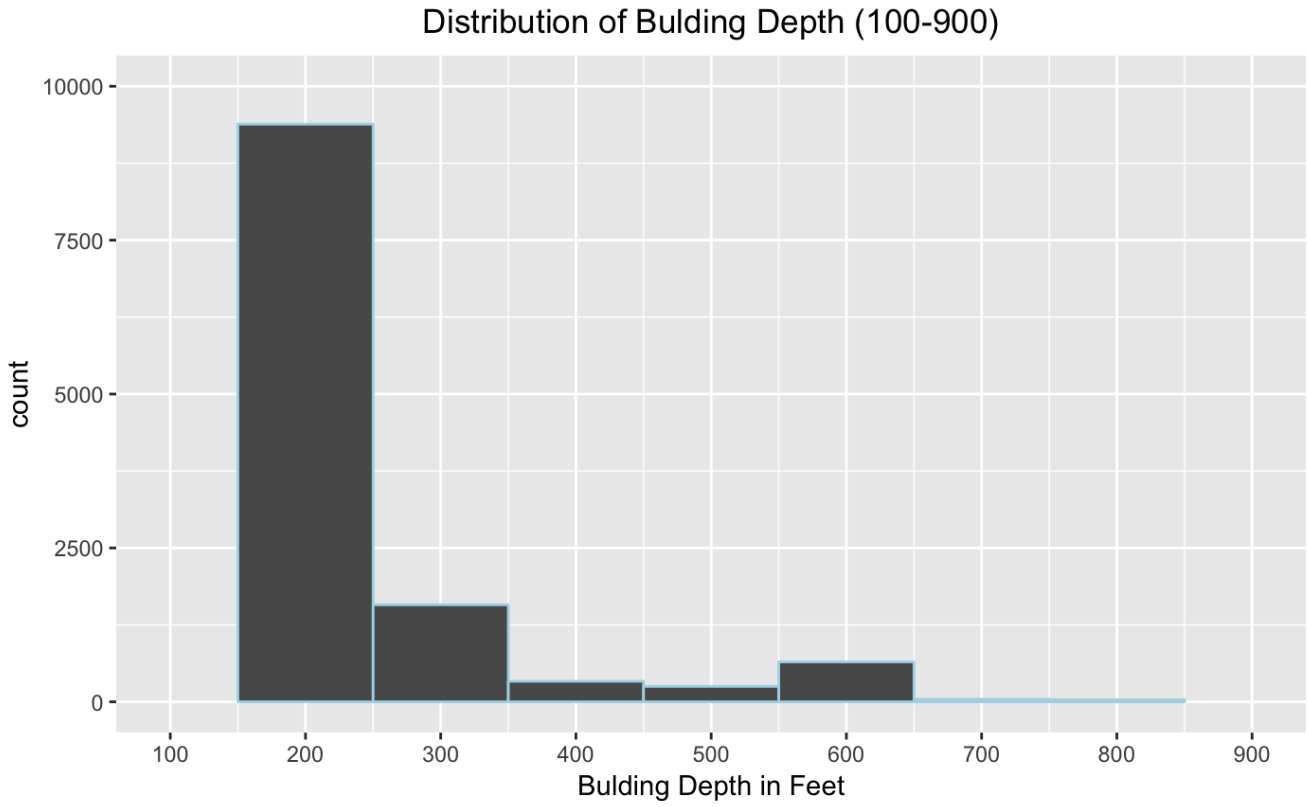


Name	Description
BLDDEPTH	Building Depth in feet (length 7 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*



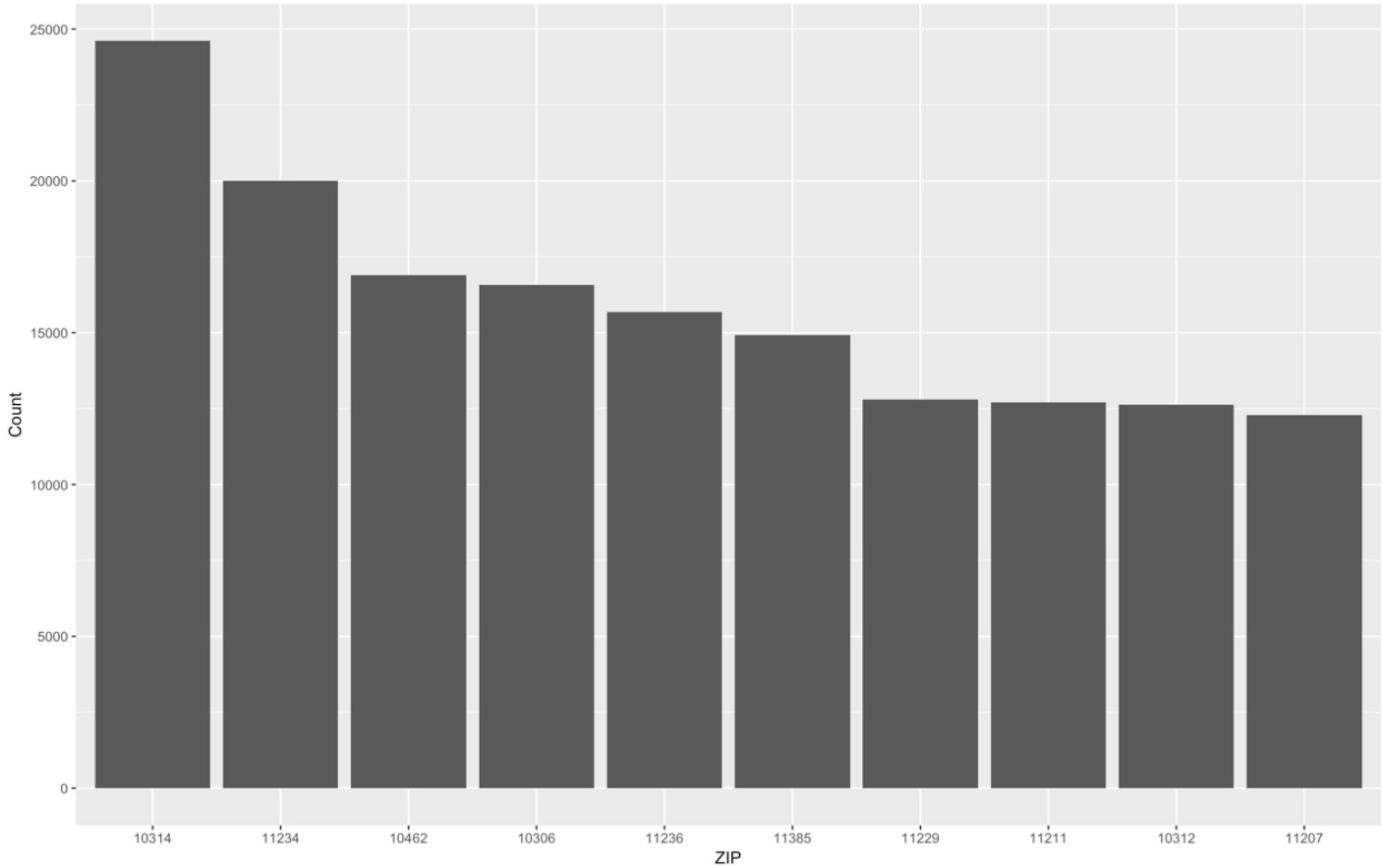

Distribution of Bulding Depth (0-100)
250000
200000
150000
100000
50000
0
count
0
10
20
30
40
50
60
70
80
90
100
Bulding Depth in Feet


Distribution of Bulding Depth (100-900)
10000
7500
5000
2500
0
count
100
200
300
400
500
600
700
800
900
Bulding Depth in Feet

Name	Description
ZIP	Postal ZIP code of the property (length 5 numeric) #NA: 26356

Top 10 ZIP Frequency and Distribution:

Zip	Weight (%)
10314	2.407018
11234	1.956626
10462	1.653755
10306	1.62157
11236	1.533722
11385	1.459668
11229	1.251493
11211	1.243373
10312	1.235939
11207	1.20258



Part III. Data Cleaning

Step 1: We deleted some less important variables from this dataset: STADDR, PERIOD, YEAR and VALTYPE. Although some of them can provide us with some general information about the property, these variables are still not very helpful for our detecting algorithms.

Step 2: We deleted some less populated variables from this dataset: EXCD1, EXMPTCL, AVLAND2, AVTOT2, EXLAND2, EXTOT2, and EXCD2. Since there are too many missing values in these variables, we cannot draw much information from them; then we decide to get rid of them.

Names	Percentage Filled
EXCD 1	59.40%
EXMPTCL	1.40%
AVLAND2	26.80%
AVTOT2	26.80%
EXLAND2	8.30%
EXTOT2	12.40%
EXCD2	8.70%

Step 3: We extracted the first letter of variable BBLE and make a new variable called “BOROUGH” indicating the borough number for each record

Step 4: For variable ZIP, we filled missing values with the corresponding central borough zip code of the record.

Borough Name	Central Zip Code to Fill
MANHATTAN	10000
BRONX	10400
BROOKLYN	11200
QUEENS	11300
STATEN ISLAND	10300

Step 5: For variables: LEFTFRONT, LTDEPTH, STORIES, FULLVAL, AVLAND, AVTOT, BLDFRONT, BLDDEPTH, we filled the missing value or zero values (0 or 0000) with the average values by each TAX CLASS.

TAX CLASS	Avg. of LTFRONT	Avg. of LTDEPTH	Avg. of STORIES	Avg. of FULLVAL	Avg. of AVLAND	Avg. of AVTOT	Avg. of BLDFRONT	Avg. of BLDDEPTH
1	30.45	99.07	2.12	570486.02	14833.97	24897.33	21.20	41.61
2	113.26	132.62	16.10	799812.36	90830.64	359916.87	92.21	108.02
3	137.25	278.55	1.33	111276.50	43368.35	50074.44	19.33	33.75
4	92.84	132.66	5.47	3254843.23	606593.74	1508998.42	64.42	87.93
1A	54.24	93.17	1.67	337564.68	2244.77	14421.87	25.34	45.91
1B	51.23	108.46	4.00	548322.84	14746.50	14749.66	39.50	82.67
1C	27.25	96.78	3.05	761535.95	8225.66	28984.47	23.37	60.09
1D	338.28	497.83	1.07	22336137.93	709303.79	1166866.31	19.59	36.86
2A	25.90	97.26	2.84	864085.44	31364.01	79617.81	22.98	62.44
2B	27.25	95.75	4.00	1253078.34	54443.67	178586.62	26.04	66.59
2C	29.99	95.74	4.75	772879.87	25639.07	117044.36	27.89	68.66

Part IV. Variables Construction:

To begin with, we selected 3 property monetary value variables and calculated 3 property size variables as the bellows:

3 property monetary value variables:

- FULLVAL: full value of building
- AVLAND: assessed value of land
- AVTOT: assessed value of property

3 property size variables:

- LOTAREA = LOTFRONT * LOTDEPTH: measurement of lot area
- BLDAREA = BLDFRONT * BLDDEPTH: measurement of building area
- BLDVOL = BLDAREA * STORIES: measurement of building volume

Then, we divided 3 monetary value variables by the 3 property size variables and got 9 numerators as the bellows:

- FULLVAL/LOTAREA
- FULLVAL/BLDAREA
- FULLVAL/BLDVOL
- AVLAND/LOTAREA
- AVLAND/BLDAREA
- AVLAND/BLDVOL
- AVTOT/LOTAREA
- AVTOT/BLDAREA
- AVTOT/BLDVOL

After that, we used the following 8 denominator variables to classify numerators. Denominator 1-4 are original variables from the dataset, and denominator 5-7 are expert variables constructed to further refine the classification.

1. ZIP5: zip code
2. ZIP3: first 3 digits of zip code
3. BOROUGH: borough code
4. TAXCLASS: tax class
5. STORIES, ZIP3: number of stories for the building, first 3 digits of zip code
6. STORIES, TAXCLASS: number of stories for the building, tax class
7. ZIP3, TAXCLASS: first 3 digits of zip code, tax class
8. ALL: original value

After grouping 9 numerators by 8 denominators, we standardized these numerators, that is, calculating mean in every group and further dividing the numerators value by the corresponding mean to get the ratio. For instance, we defined the variable

FULLVAL_LOTAREA_ZIP5 with a certain ZIP5 as: FULLVAL_LOTAREA_ZIP5 = (FULLVAL/LOTAREA) / (mean of FULLVAL/LOTAREA in that group).

In total, we created an 8*9 metric, that is, 72 expert variables.

Catalog of Expert Variables

#	Variable Name	Variable Description
1	r_FV_lotarea5	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of buildings grouped by ZIP CODE
2	r_FV_bldarea5	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of buildings grouped by ZIP CODE
3	r_FV_bldvol5	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of buildings grouped by ZIP CODE
4	r_AL_lotarea5	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of buildings grouped by ZIP CODE
5	r_AL_bldarea5	Ratio of AVLAND /BLDAREA to Average AVLAND/BLDAREA of buildings grouped by ZIP CODE
6	r_AL_bldvol5	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of buildings grouped by ZIP CODE
7	r_AT_lotarea5	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of buildings grouped by ZIP CODE
8	r_AT_bldarea5	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of buildings grouped by ZIP CODE
9	r_AT_bldvol5	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of buildings grouped by ZIP CODE
10	r_FV_lotarea3	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of buildings grouped by first 3 digits of ZIP CODE
11	r_FV_bldarea3	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of buildings grouped by first 3 digits of ZIP CODE
12	r_FV_bldvol3	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of buildings grouped by first 3 digits of ZIP CODE
13	r_AL_lotarea3	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of buildings grouped by first 3 digits of ZIP CODE
14	r_AL_bldarea3	Ratio of AVLAND /BLDAREA to Average AVLAND/BLDAREA of buildings grouped by first 3 digits of ZIP CODE
15	r_AL_bldvol3	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of buildings grouped by first 3 digits of ZIP CODE
16	r_AT_lotarea3	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of buildings grouped by first 3 digits of ZIP CODE
17	r_AT_bldarea3	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of buildings grouped by first 3 digits of ZIP CODE
18	r_AT_bldvol3	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of buildings grouped by first 3 digits of ZIP CODE
19	r_FV_lotareaT	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of buildings grouped by TAX CLASS
20	r_FV_bldareaT	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of buildings grouped by TAX CLASS

21	r_FV_bldvolT	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of buildings grouped by TAX CLASS
22	r_AL_lotareaT	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of buildings grouped by TAX CLASS
23	r_AL_bldareaT	Ratio of AVLAND /BLDAREA to Average AVLAND/BLDAREA of buildings grouped by TAX CLASS
24	r_AL_bldvolT	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of buildings grouped by TAX CLASS
25	r_AT_lotareaT	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of buildings grouped by TAX CLASS
26	r_AT_bldareaT	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of buildings grouped by TAX CLASS
27	r_AT_bldvolT	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of buildings grouped by TAX CLASS
28	r_FV_lotareaB	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of buildings grouped by BOROUGH CODE
29	r_FV_bldareaB	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of buildings grouped by BOROUGH CODE
30	r_FV_bldvolB	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of buildings grouped by BOROUGH CODE
31	r_AL_lotareaB	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of buildings grouped by BOROUGH CODE
32	r_AL_bldareaB	Ratio of AVLAND /BLDAREA to Average AVLAND/BLDAREA of buildings grouped by BOROUGH CODE
33	r_AL_bldvolB	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of buildings grouped by BOROUGH CODE
34	r_AT_lotareaB	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of buildings grouped by BOROUGH CODE
35	r_AT_bldareaB	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of buildings grouped by BOROUGH CODE
36	r_AT_bldvolB	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of buildings grouped by BOROUGH CODE
37	r_FV_lotareaS3	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of buildings grouped by STORIES and first 3 digits of ZIP CODE
38	r_FV_bldareaS3	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of buildings grouped by STORIES and first 3 digits of ZIP CODE
39	r_FV_bldvolS3	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of buildings grouped by STORIES and first 3 digits of ZIP CODE
40	r_AL_lotareaS3	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of buildings grouped by STORIES and first 3 digits of ZIP CODE
41	r_AL_bldareaS3	Ratio of AVLAND /BLDAREA to Average AVLAND/BLDAREA of buildings grouped by STORIES and first 3 digits of ZIP CODE

42	r_AL_bldvolS3	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of buildings grouped by STORIES and first 3 digits of ZIP CODE
43	r_AT_lotareaS3	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of buildings grouped by STORIES and first 3 digits of ZIP CODE
44	r_AT_bldareaS3	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of buildings grouped by STORIES and first 3 digits of ZIP CODE
45	r_AT_bldvolS3	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of buildings grouped by STORIES and first 3 digits of ZIP CODE
46	r_FV_lotareaST	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of buildings grouped by STORIES and TAX CLASS
47	r_FV_bldareaST	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of buildings grouped by STORIES and TAX CLASS
48	r_FV_bldvolST	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of buildings grouped by STORIES and TAX CLASS
49	r_AL_lotareaST	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of buildings grouped by STORIES and TAX CLASS
50	r_AL_bldareaST	Ratio of AVLAND /BLDAREA to Average AVLAND/BLDAREA of buildings grouped by STORIES and TAX CLASS
51	r_AL_bldvolST	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of buildings grouped by STORIES and TAX CLASS
52	r_AT_lotareaST	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of buildings grouped by STORIES and TAX CLASS
53	r_AT_bldareaST	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of buildings grouped by STORIES and TAX CLASS
54	r_AT_bldvolST	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of buildings grouped by STORIES and TAX CLASS
55	r_FV_lotarea3T	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS
56	r_FV_bldarea3T	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS
57	r_FV_bldvol3T	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS
58	r_AL_lotarea3T	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS
59	r_AL_bldarea3T	Ratio of AVLAND /BLDAREA to Average AVLAND/BLDAREA of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS
60	r_AL_bldvol3T	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS

61	r_AT_lotarea3T	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS
62	r_AT_bldarea3T	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS
63	r_AT_bldvol3T	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of buildings grouped by first 3 digits of ZIP CODE and TAX CLASS
64	r_FV_lotareaAL	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of buildings
65	r_FV_bldareaAL	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of buildings
66	r_FV_bldvolAL	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of buildings
67	r_AL_lotareaAL	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of buildings
68	r_AL_bldareaAL	Ratio of AVLAND /BLDAREA to Average AVLAND/BLDAREA of buildings
69	r_AL_bldvolAL	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of buildings
70	r_AT_lotareaAL	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of buildings
71	r_AT_bldareaAL	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of buildings
72	r_AT_bldvolAL	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of buildings

Part V. Principal Component Analysis (PCA)

Introduction of PCA

Principal component analysis (PCA) is an unsupervised dimension-reduction technique used to transform the high-dimensional dataset into a lower-dimensional dataset, which still retains most of the information (or variation) in the original dataset but is significantly reduced in dimensional complexity. Mathematically, PCA is performed via linear algebra functions called Eigen-decomposition or Svd-decomposition, which calculates the principal components and order them by the decreasing order of their eigenvalues. It can be intuitively understood as a projection of the original dataset into a new coordinate system so that the highest variance in the dataset can be best explained by the first several coordinates (or the first several principal components).

The output of the PCA in our analysis is a matrix composed of principal components (PCs), which corresponds to a linear combination of the original variables, on its columns, and the 72 expert variables on its rows. After examining the scree plot of decaying variance on PCs, we only retained PC1 to PC7, which can already explain over 90% of the variation of the dataset.

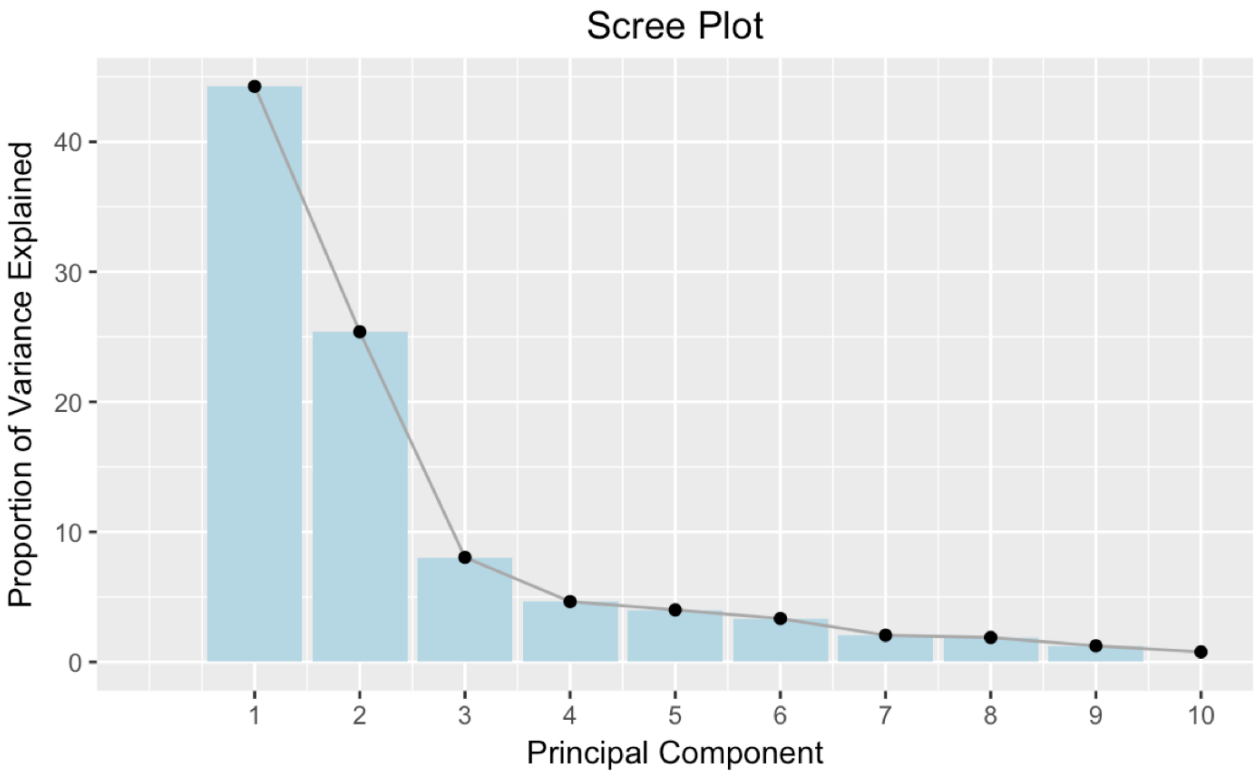
Functions of PCA

PCA plays an indispensable role in our fraud detection analysis. Its specific functionalities are listed as follows:

1. Dimensionality and Complexity Reduction
   In the variable creation step, in order to fully explore each variable and the interaction effect behind them, we created 72 expert variables, which inevitably overlap or correlate with each other regarding their measurements and properties. PCA is an extremely effective tool to summarize these variables in an efficient and representative way so that the dimensionality and correlation in the variables are significantly reduced. Additionally, the simplified results obtained from the PCA can be directly used in our following steps of fraud algorithms, and significantly reduced the complexity and computation time.
2. Originality Retention
   Besides allowing us to capture the variables that best explain the variation in the dataset by providing a set of principal components, PCA also retains the original expert variables to minimize the loss of information when processing the data.

Analysis and Visualizations

In this project, we used the prcomp() function in “stats” package in R to perform the PCA. We inputted the 72 expert variables we created into the function and specified “center” and “scale” to be “TRUE” in the function to normalize all variables. The PCA result can be visualized as the following scree plot.



The scree plot clearly shows how well each principal component explains the variation in the data. We can tell from the graph that PC1 can explain nearly 45% of the variance, while the succeeding explanation wellness drops sharply from PC2, which can explain approximately 25% of the variance. Through calculations, it is found that at from PC1 to PC7, the cumulative percentage of variance explained is 91.75%. Therefore, we set the cut-off point at PC7 and use these 7 principal components to perform the Heuristic algorithm and the Autoencoder and calculate the final fraud score.

Part VI. Fraud Algorithms

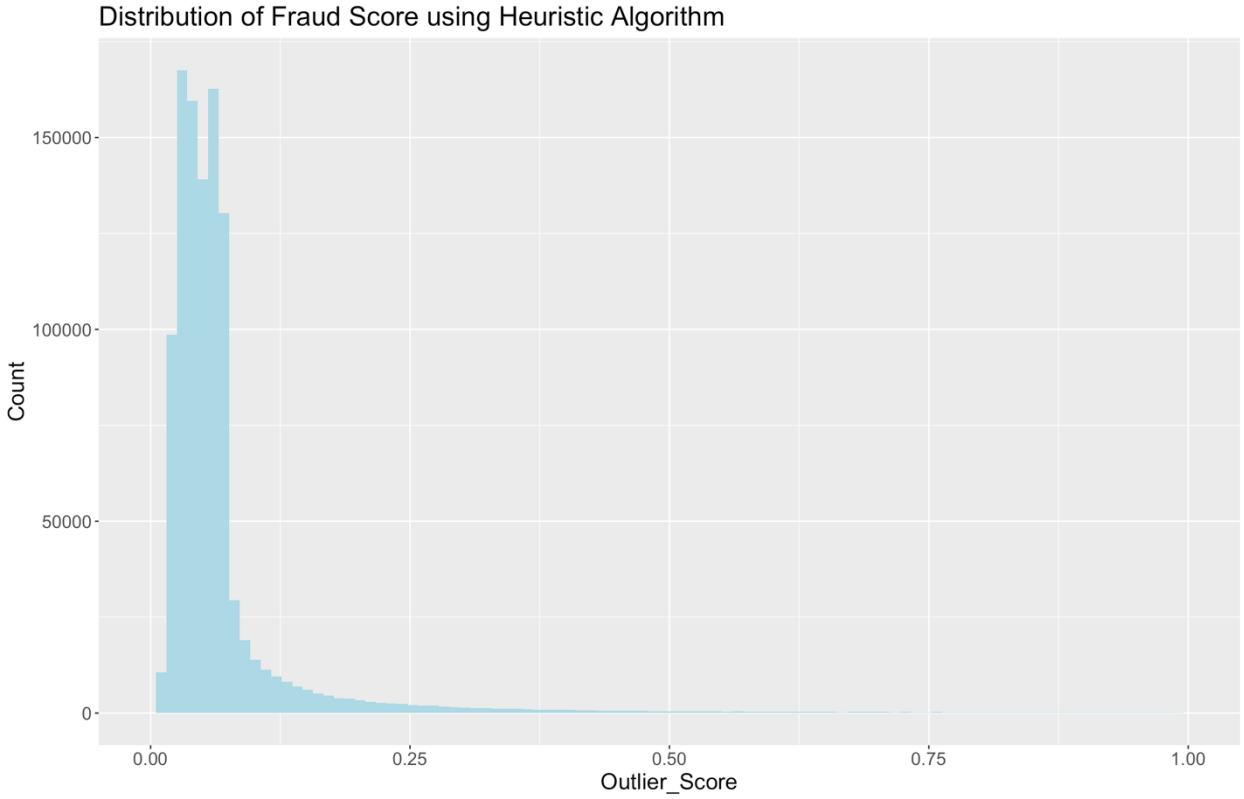
To start building fraud algorithms, we first used the 7 principal components we got from our Principal Component Analysis to reduce the dimensions of the original dataset by multiplying the PCA matrix and the original data matrix. Then, we z-scaled that matrix to obtain the final data dataset for further calculation of the fraud scores.

Heuristic Algorithm

Since we have performed z-scaling after PCA, it saves us the effort to calculate Mahalanobis Distance (the same as first performing PCA then z-scaling). Our heuristic algorithm calculates the fraud score using outlier detection via z-scores. It takes the following form:

$$S_H = \sqrt[2]{\sum_{i=1}^{k} {z_i}^2}$$

The graph below shows the distribution of this fraud score. It is right-skewed with a very long tail.



The score ranges from 0.0069 to 1921.29, and most of the values (99.37% of all records) are distributed between 0 and 1.

Autoencoder

An autoencoder is a neural network that is trained to reconstruct its own input. For example, we input some value (say, (4.5,1.7,-2.3)) and penalize it if it returns anything but (4.5,1.7,-2.3). Once the network can reliably reconstruct its input, the hidden layer must contain enough information to represent the output. If the hidden layer is smaller than the input and output layers, what it represents is the same information in a lower dimensionality.

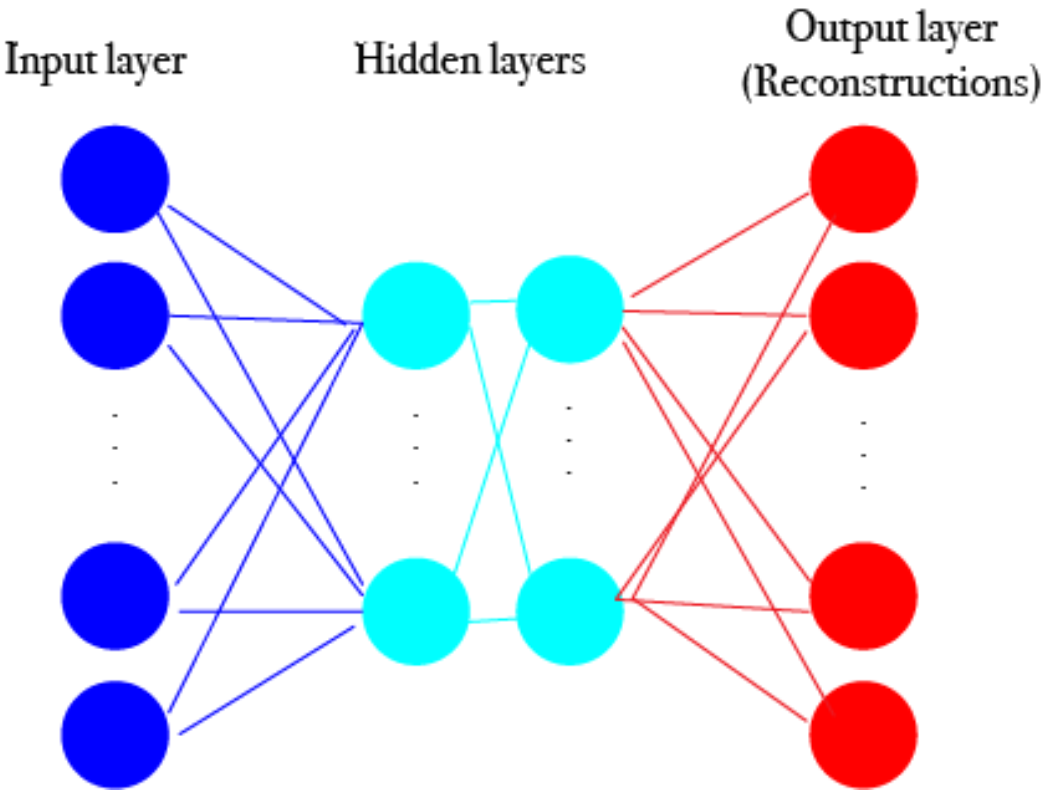


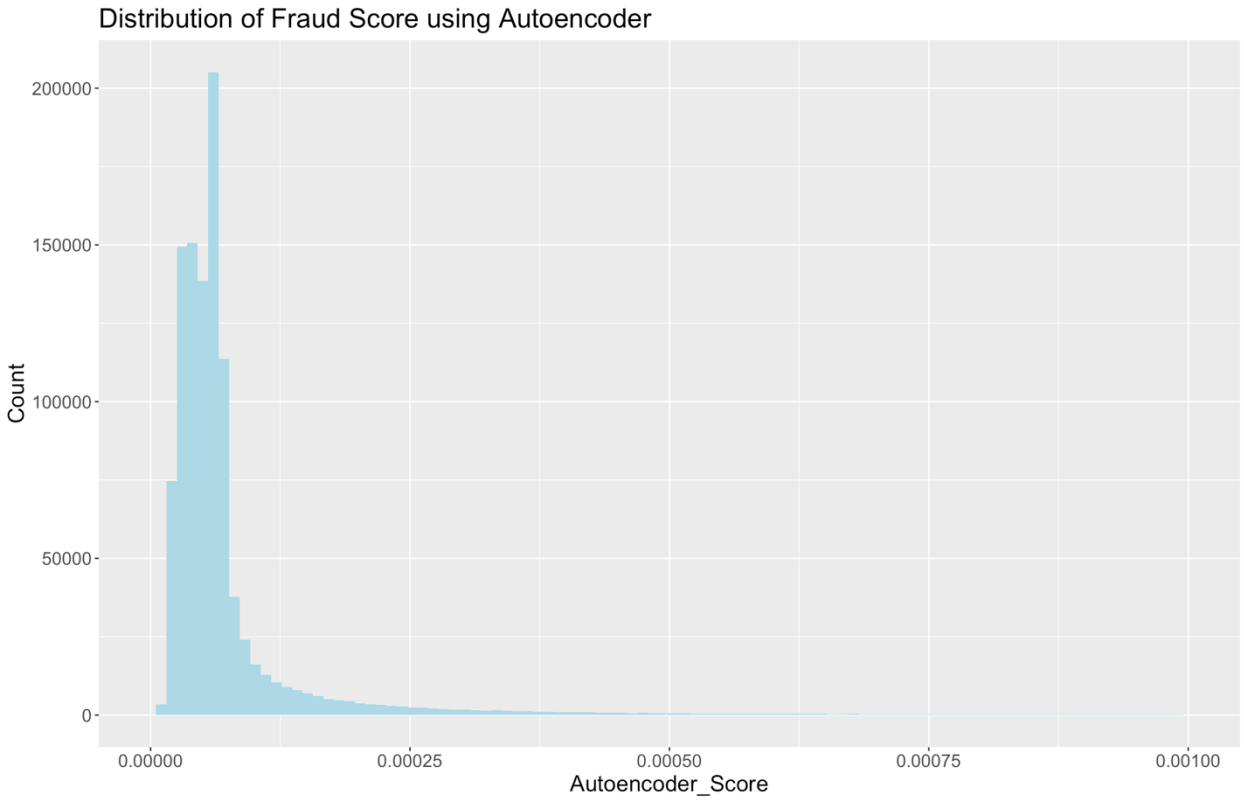
Figure. Illustration of Autoencoder

In fraud study, we can use autoencoder to find anomaly records. Based on the autoencoder model that was trained before, the input data will be reconstructed, and for each instance, the mean squared error (MSE) between the actual value and reconstruction value is calculated. We can use MSE to detect fraudulent behaviors since the MSE is going to be higher for fraudulent records than for regular ones.

We trained our final dataset with an R package called “h2o”. We used the default “h2o” autoencoder function which has 2 hidden layers, and each hidden layer contains 50 nodes. After training the model, we called ‘h20.anomaly’ function to calculate the squared errors for each record. In the end, we computed the MSE for each record as the fraud score.

$$S_A = \sqrt[2]{\sum_{i=1}^{k} (z_i - z_i')^2}$$

The graph below shows the distribution of this fraud score. It is right-skewed with a very long tail.



The score ranges from 0.0000068 to 1.9274842, and most of the values (99.25% of all records) are distributed between 0 and 0.001.

Combined Score

For each of the two scores:

- We sorted the records by the score in descending order and binned them into 1000 bins with equal bin size.
- For records in the highest bin, we replaced the scores by the number 1000. For the next bin, we replaced the scores by the number 999 and so on.

After we obtained the new, scaled scores for each of the two scores, we calculated the combined score using the weighted average.

$$S_{combined} = 0.5 * S'_H + 0.5 * S'_A$$

The percentage of records with overlapping scaled scores is 2.4%, which means that around 25,000 records have the same scaled score.

Part VII. Results

Based on the combined ranking, we selected the top 0.1% records to capture the general characteristics. More specifically, we selected 1035 records with a combined ranking of 1000. Below is the comparison between selected records with a high score and all the records for important numerical variables. To make the comparison more consistent, we use the clean dataset for comparison.

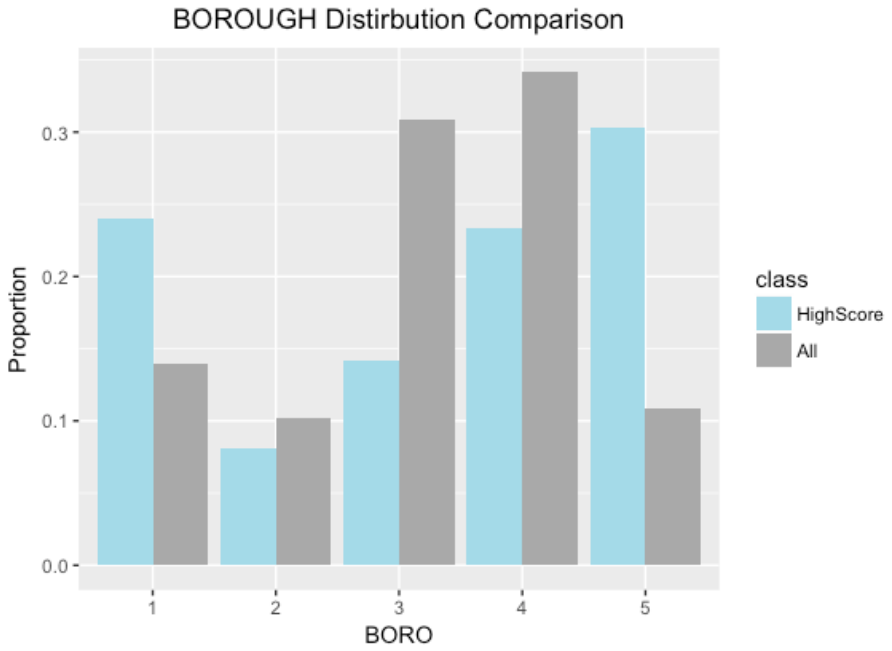
	Complete Data			**Top 0.1% Records**		
Variable	mean	median	SD	mean	median	SD
STORIES	5.0664	2	8.24905	8.995695	5.474805	13.07284
LTFRONT	52.59101	30	75.77305	386.6807	100	831.67
LTDEPTH	109.0974	100	65.72018	360.8573	121	682.4034
BLDFRONT	38.85578	22	40.0639	58.60882	64.41877	63.16458
BLDDEPTH	60.4261	48	42.68128	70.37368	87.92779	54.92998
LOTAREA	8075.638	3000	154736.7	483764.3	12500	2227150
BLDAREA	3451.8	1078	99653.3	6131.358	5664.2	15289.38
BLDVOLUME	44161.09	2256	2316931	109001.2	31010.39	341139.7
FULLVAL	889777	451000	11703708	78080729	4910000	321454324
AVTOT	234748.4	25740	6951304	38206707	2127060	203928970
AVLAND	87386.54	13857	4100781	20356954	1260000	125984771

Insights for numerical values:

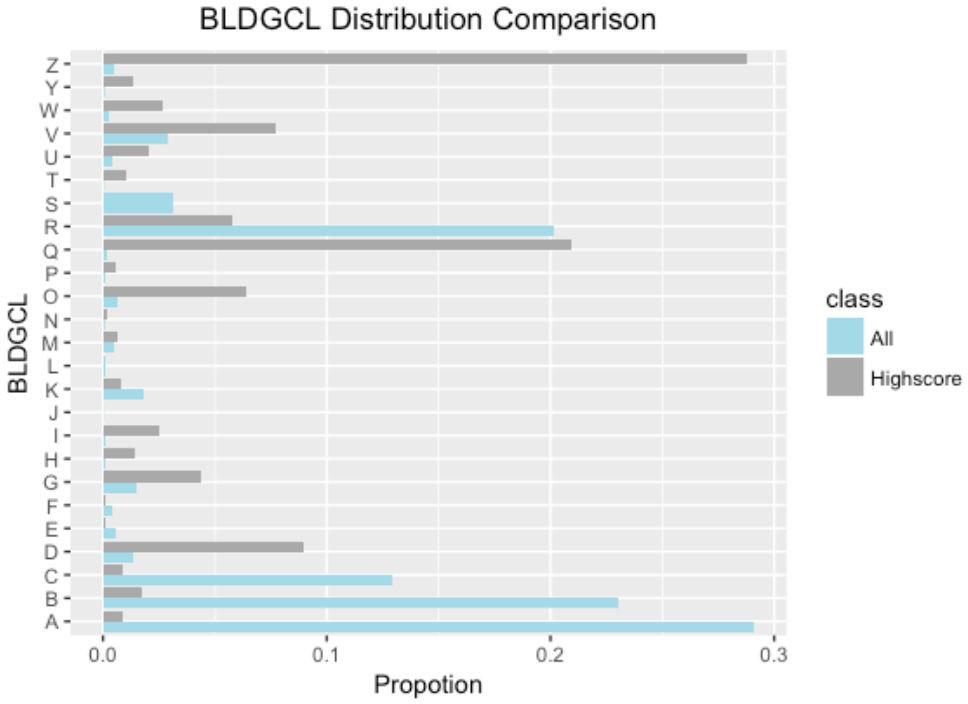
1) Compared with the complete dataset, records with higher fraud score have significantly larger STORIES, LTFRONT, BLDFRONT, BLDDEPTH, LOTAREA, BLDAREA, BLDVOLUME. According to the average and median value of BLDAREA and BLDVOLUME, potential fraud properties are around twice the sizes of the complete data. This indicates that bigger buildings have higher propensity to fraud.
2) The potential fraud properties have significant higher values including FULLVAL, AVTOT and AVLAND. According to the average value and median value of FULLVAL, AVTOT and AVLAND, the potential fraud properties have more than 100 times the value of complete data.

Besides numerical variables, we also compare the distribution of top 0.1% properties and the whole dataset with respect to categorical variables. The comparisons and insights are as below.

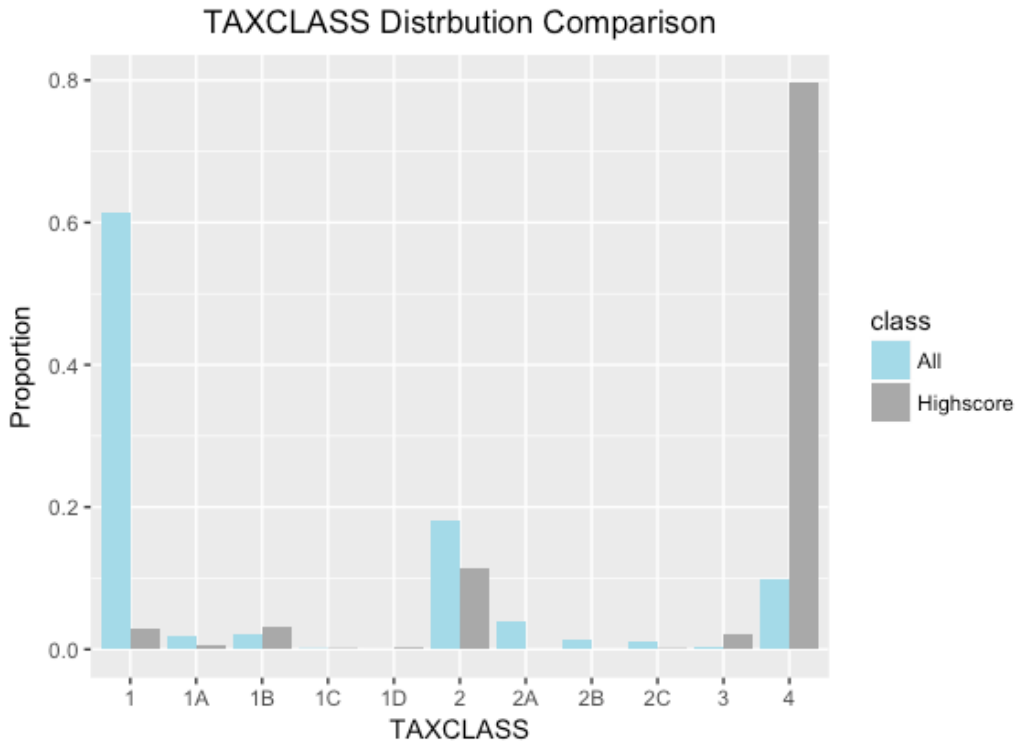
1) The potential fraud properties are more concentrated in BOROUGH 1 and BOROUGH 5, which are Manhattan and Staten Island.



2) The potential fraud properties are mostly in Building Class Z, Q and D. Compared with the whole dataset, the potential fraud properties are more concentrated on Building Class Z, D and O, and less concentrated on Building Class A, B, and C.



3) According to the distribution of tax class, the potential fraud properties are more concentrated on Tax Class 4, and less concentrate on Tax Class 1. Tax Class 1 mainly includes 1-3 unit residences, which are small buildings and not the main types of potential fraud properties. Tax Class 4 includes utilities-ceiling railroads and all other building types. This indicates that the potential fraud properties are usually less common building types.



We further select the top 10 properties with respect to the score computed by the heuristic algorithm. The detailed information about the 10 properties are as below.

Borough, Borough Name, Building Class, Tax Class, LTFRONT, LOTAREA for top 10 records

RECORD	BOROUGH	BORO_NAME	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	LOTAREA
5393	4	QUEENS	D9	2	157	95	14915
977471	4	QUEENS	O3	4	298	402	119796
78804	3	BROOKLYN	V9	4	117	108	12636
1046264	4	QUEENS	Q9	4	6	1	6
787892	4	QUEENS	O3	4	139	342	47538
24586	4	QUEENS	H9	4	94	165	15510
376243	4	QUEENS	T1	4	4910	132.6583	651352.253
224352	5	STATEN ISLAND	D3	2	136	132	17952

970081	1	MANHATTAN	Q1	4	4000	150	600000
894415	4	QUEENS	Q1	4	610	534	325740

Stories, FULLVAL, AVALAND, BLDFRONT, BLDDEPTH, BLDAREA, BLDVOLUME for top 10 records:

RECORD	STORIES	FULLVAL	AVLAND	BLDFRONT	BLDDEPTH	BLDAREA	BLDVOLUME
5393	1	2930000	1318500	1318500	1	1318500	1318500
977471	20	3443400	1549530	1549530	1	1549530	30990600
78804	5.474805	4326303700	1946836665	1946836665	64.41877	1.25413E+11	6.86611E+11
1046264	1	3254843	606593.7	1508998	64.41877	97207795.09	97207795.09
787892	20	2151600	968220	968220	1	968220	19364400
24586	10	3712000	252000	1670400	1	1670400	16704000
376243	3	374019883	1792808947	4668308947	64.41877	3.00727E+11	9.0218E+11
224352	8	1040000	236250	468000	1	468000	3744000
970081	1	70214000	31455000	31596300	8	252770400	252770400
894415	3	242000000	103500000	108900000	20	2178000000	6534000000

Comparing the summary statistics of the entire dataset and the top 0.1% records with the top 10 records, we can draw the following insights:

1) Though the top 0.1% potential fraud properties are more concentrated on Manhattan and Staten Island, 7 of the top 10 potential fraud properties are in Queens. This indicates that the number of fraud properties in Queens might not be significantly large, but there're some properties with a high probability of fraud in Queens.
2) Among the top 10 potential fraud properties, 8 of them are in tax class 4, which is consistent with the result shown in top 0.1% potential fraud properties.
3) The top 10 potential fraud properties have large variance in STORIES. Record 5393, 1046264 and 970081 only have 1 story, while record 977471 and 787892 have 20 stories, and record 24586 have 10 stories. Most of the top 10 fraud properties have either extremely higher or extremely lower stories compared with the average value, indicating that an unusual number of stories might be relevant to fraud.
4) The building area and building volume of the top 10 potential fraud properties are all far beyond the average of the top 0.1% potential fraud properties. It makes sense that such large buildings have higher values, but the value might also be fraudulent due to the lack of comparison with similar properties.

Appendix: Data Quality Report

1. Summary Statistics for Numeric Variables

	Min	1stIQR	Median	Mean	3rdIQR	Max	# NA	Mode	SD	% populated	Unique #
LTFRONT	0	19	25	36.17	40	9999	0	0	73.73	100%	1277
LTDEPTH	0	80	100	88.28	100	9999	0	100	75.48	100%	1336
STORIES	1	2	2	5.06	3	119	52142	2	8.43	95.03%	111
FULLVAL	4	3.107e+05	4.5e+05	8.913e+05	6.23e+05	6.15e+09	12762	502000	11774390	98.78%	108276
AVLAND	1	9.425e+03	1.375e+04	8.706e+04	1.983e+04	2.668e+09	12764	45000	4125933	98.78%	70528
AVTOT	1	1.866e+04	2.556e+04	2.336e+05	4.688e+04	4.668e+09	12762	16588	6993850	98.78%	112293
EXLAND	0	0	1.620e+03	3.681e+04	1.620e+03	2.668e+09	0	0	4.024e+06	100%	33186
EXTOT	0	0	1.620e+03	9.254e+04	2.090e+03	4.668e+09	0	0	6.578e+06	100%	63805
EXCD1	1010	1017	1017	1604	1017	7170	425933	1017	1.388e+03	59.38%	129
BLDFRONT	0	15	20	23.02	24	7575	0	0	35.79	100%	610
BLDDEPTH	0	26	39	40.07	51	9393	0	0	43.04	100%	620
AVLAND2	3	5.705e+03	2.006e+04	2.464e+05	6.234e+04	2.371e+09	767609	2408	6.199e+06	26.80%	58169
AVTOT2	3	3.401e+04	8.001e+04	7.161e+05	2.408e+05	4.501e+09	767603	750	1.169e+07	26.80%	110890
EXLAND2	1	2.090e+03	3.053e+03	3.518e+05	3.142e+04	2.371e+09	961900	2090	1.085e+07	8.27%	21996
EXTOT2	7	2.889e+03	3.712e+04	6.581e+05	1.066e+05	4.501e+09	918642	2090	1.613e+07	12.40%	48106
EXCD2	1011	1017	1017	1372	1017	7160	957634	1017	1.105e+03	8.67%	60

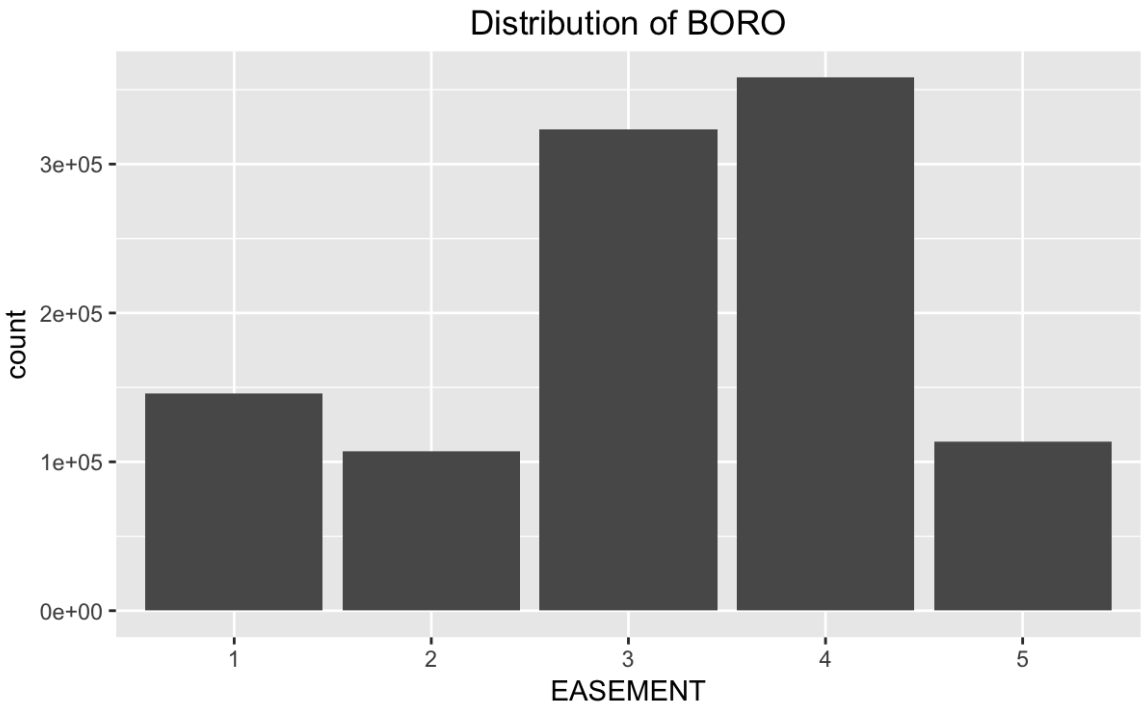
Note: Value 0 for some variables, including FULLVAL, AVLAND, and AVTOT, indicates missing value and thus excluded when calculating statistics.*

- Findings:
  1) Outliers exist in all variables, some maxima are too large, and some minima are meaningless.
  2) In addition to LTDEPTH, other variables have mean larger than median, indicating a right-skew distribution.
  3) Variables labeled with "2", including AVLAND2, AVTOT2, EXLAND2, EXTOT2, and EXCD2, have a significantly larger percent of missing values.

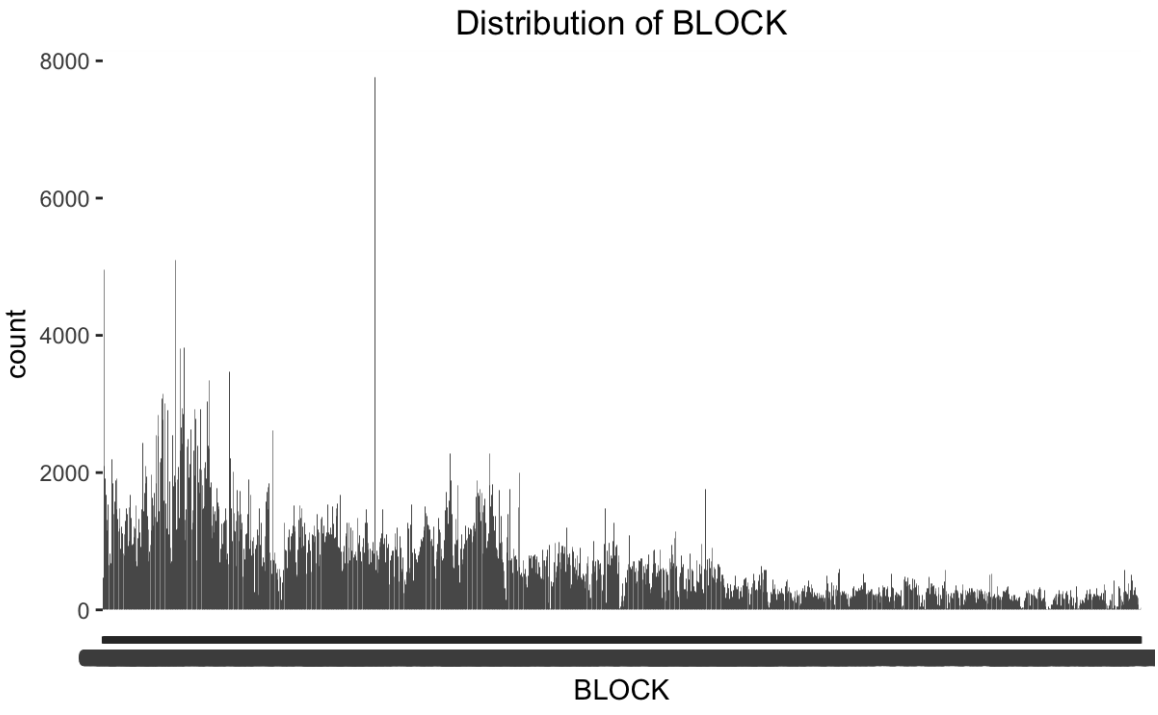
2. Detailed Information for Each Field

Variable 1	Description
RECORD	A unique key of record, from 1 to 1,048,575

Variable 2	Description
BBLE	A file key to uniquely identify each record. Concatenation of BBLE_BORO, BBLE_BLOCK, BBLE_LOT, and BBLE_EASEMENT. (length: 11 alphanumeric)
BBLE_BORO	1 = MANHATTAN 2 = BRONX 3 = BROOKL YN 4 = QUEENS 5 = STATEN ISLAND

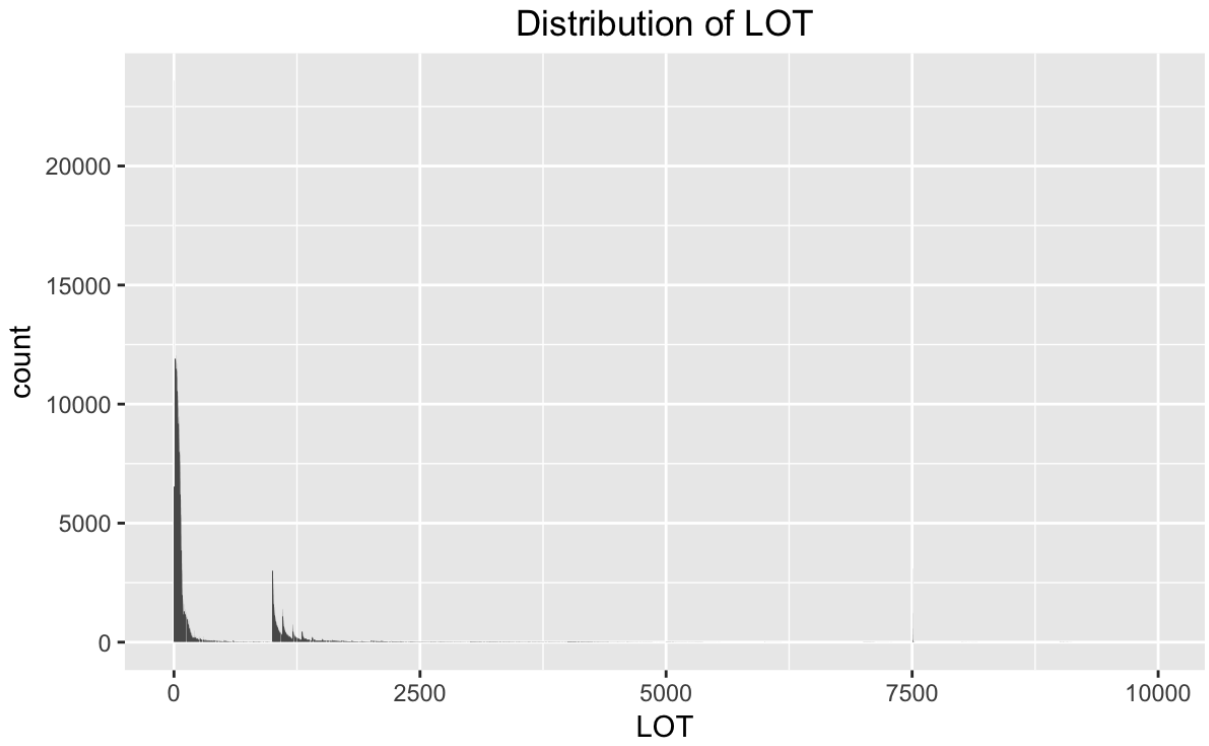


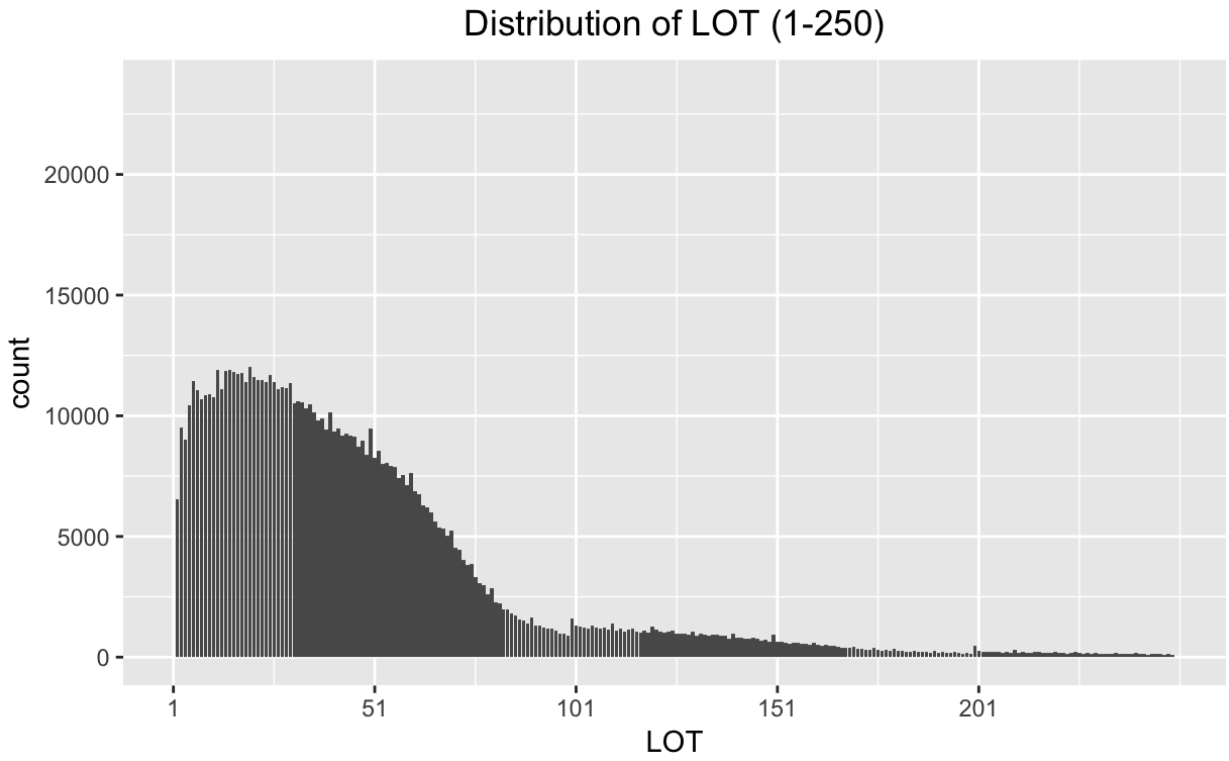
Variable 3	Description
BLOCK	Valid block ranges by BORO. (length: 5 numeric)



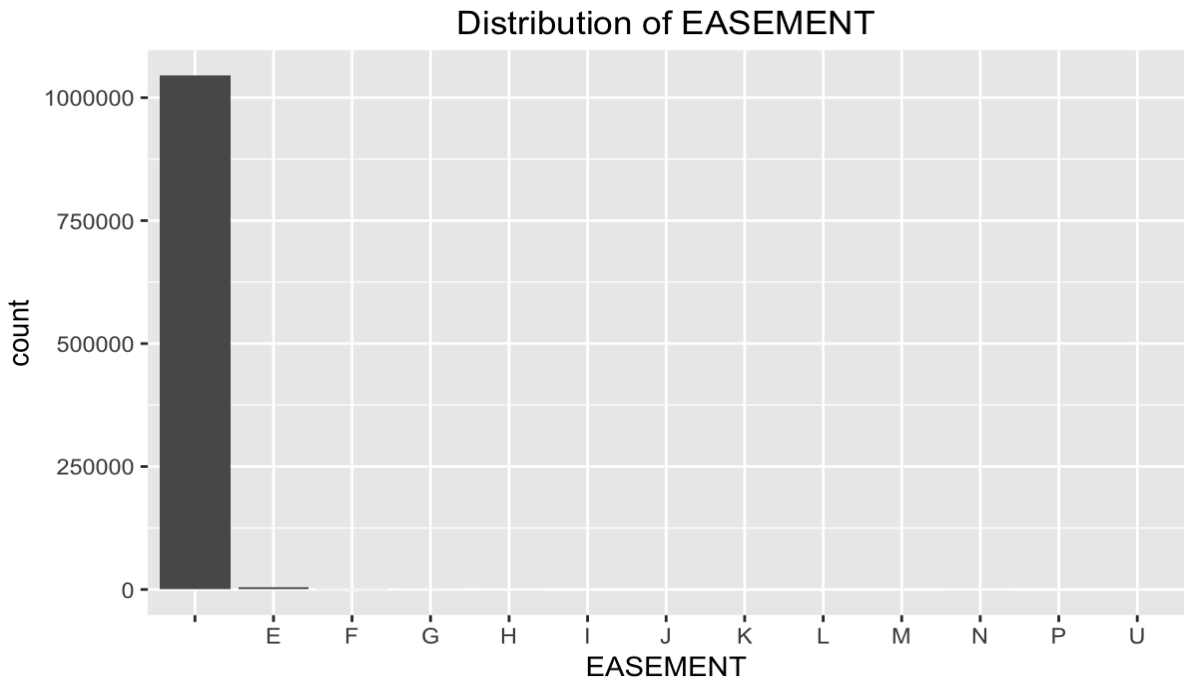
Variable 4	Description
LOT	Unique # Within BORO/BLOCK. (length: 4 numeric)

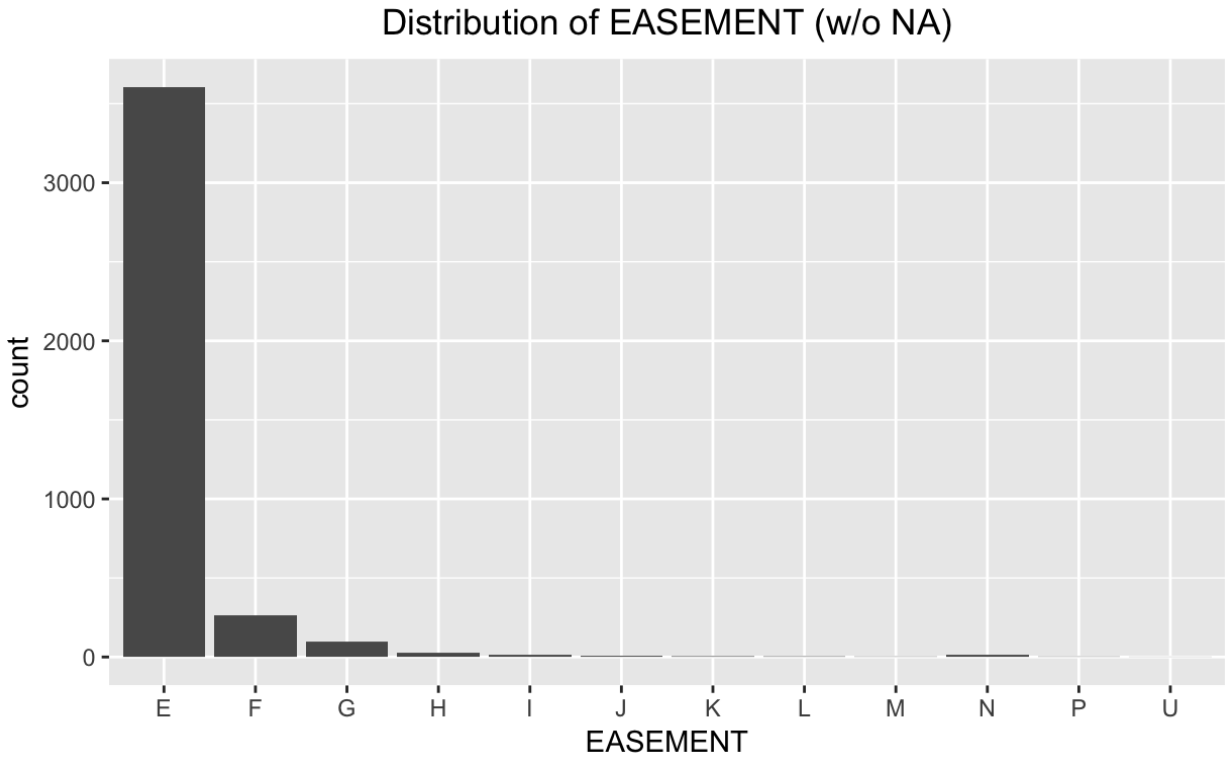
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*





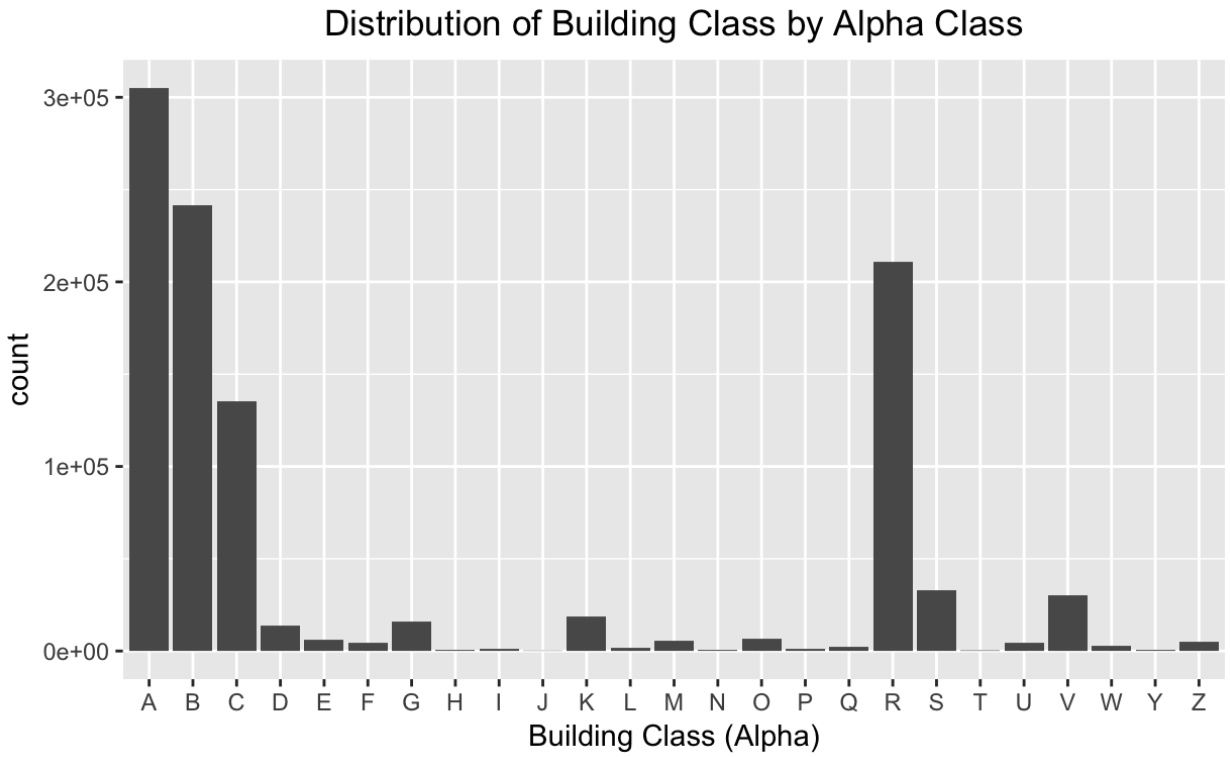
Variable 5	Description
EASEMENT	SPACE: No Easement. 'A' Indicates the portion of the Lot that has an Air Easement 'B' Indicates Non-Air Rights. 'E' Indicates the portion of the lot that has a Land Easement 'F' THRU 'M' Are duplicates of 'E'. 'N' Indicates Non-Transit Easement 'P' Indicates Piers.



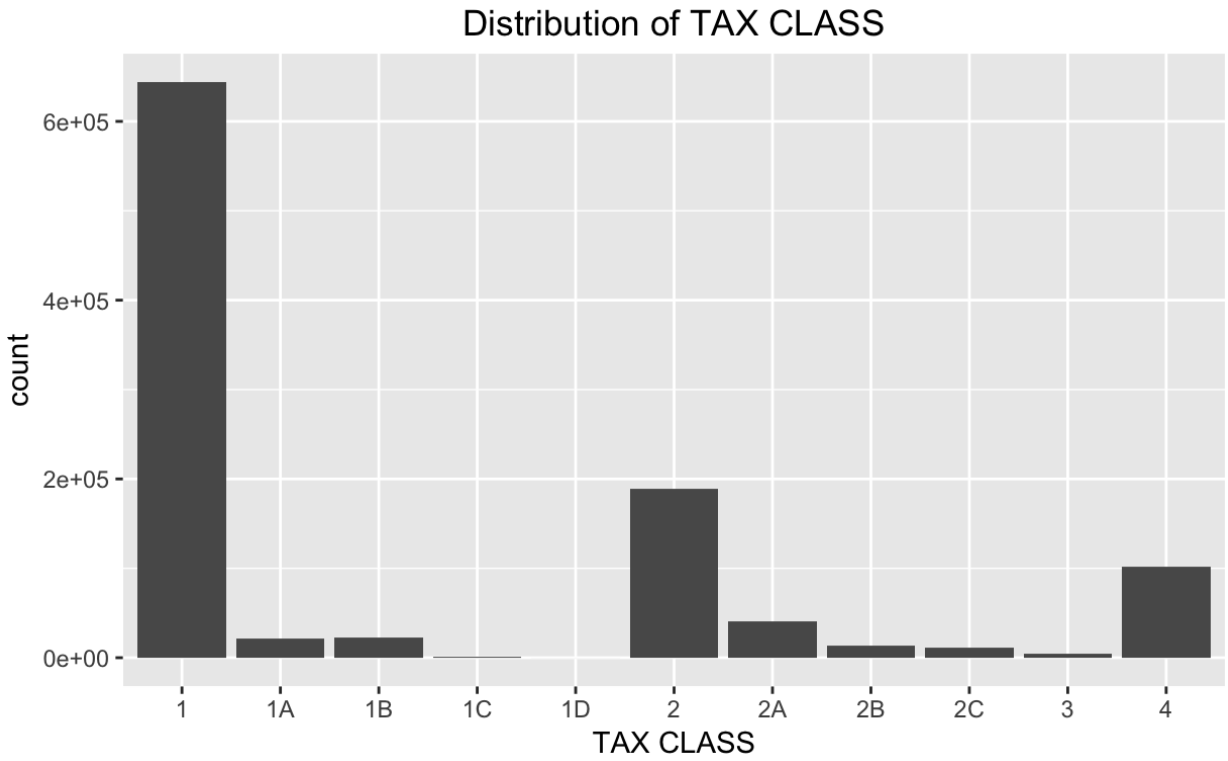


Variable 6	Description
OWNER	The owner's name; # NA: 31,081

Variable 7	Description
BLDGCL	Position 1 = Alphas & Position 2 = Numeric (length: 2 Character)

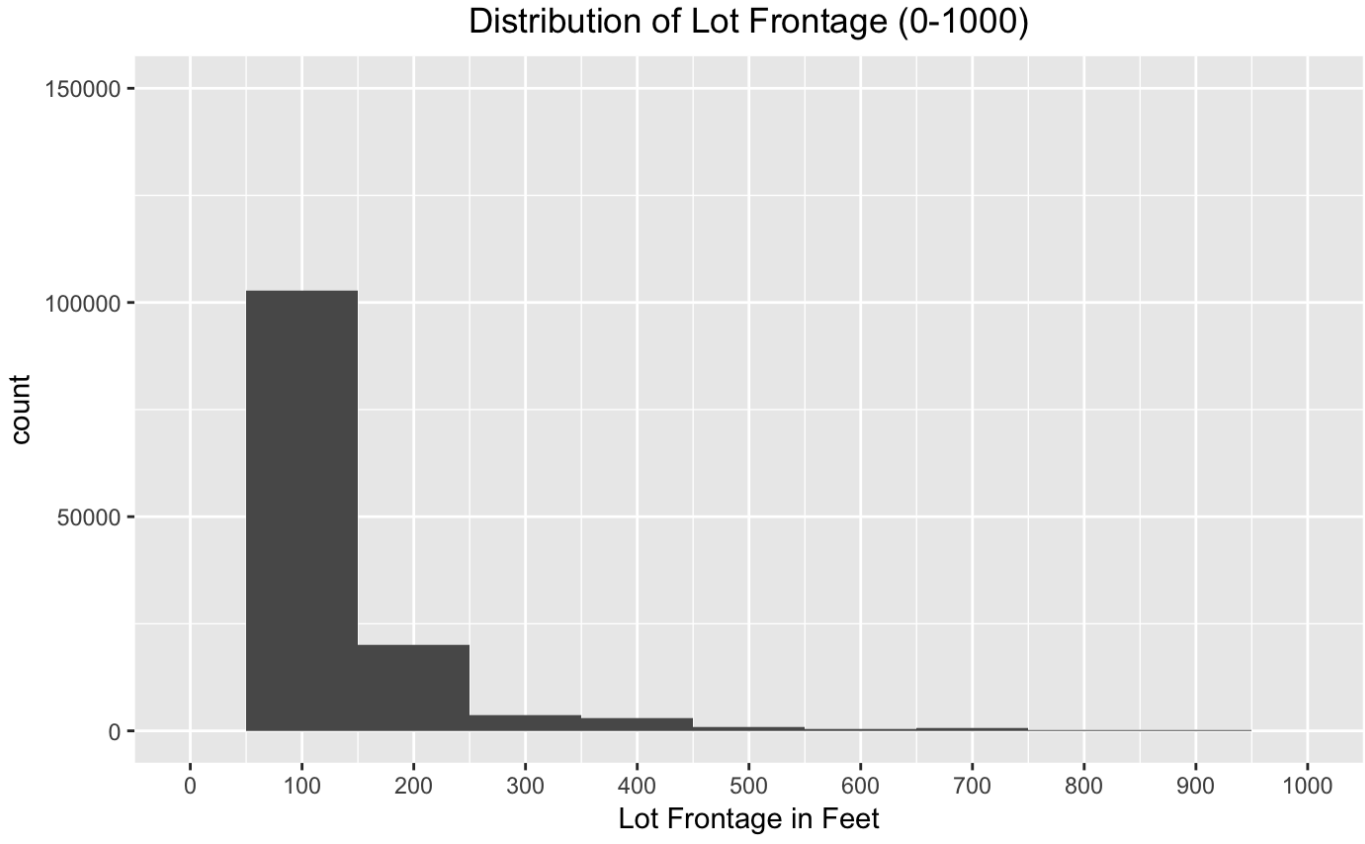


Variable 8	Description
TAXCLASS	Current Property Tax Class Code (NYS Classification) 1 = 1-3 unit residences 1A = 1-3 story condominiums originally a condo 1B = residential vacant land 1C = unit condominiums originally tax class 1 1D = select bungalow colonies 2 = apartments 2A = apartments with 4-6 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

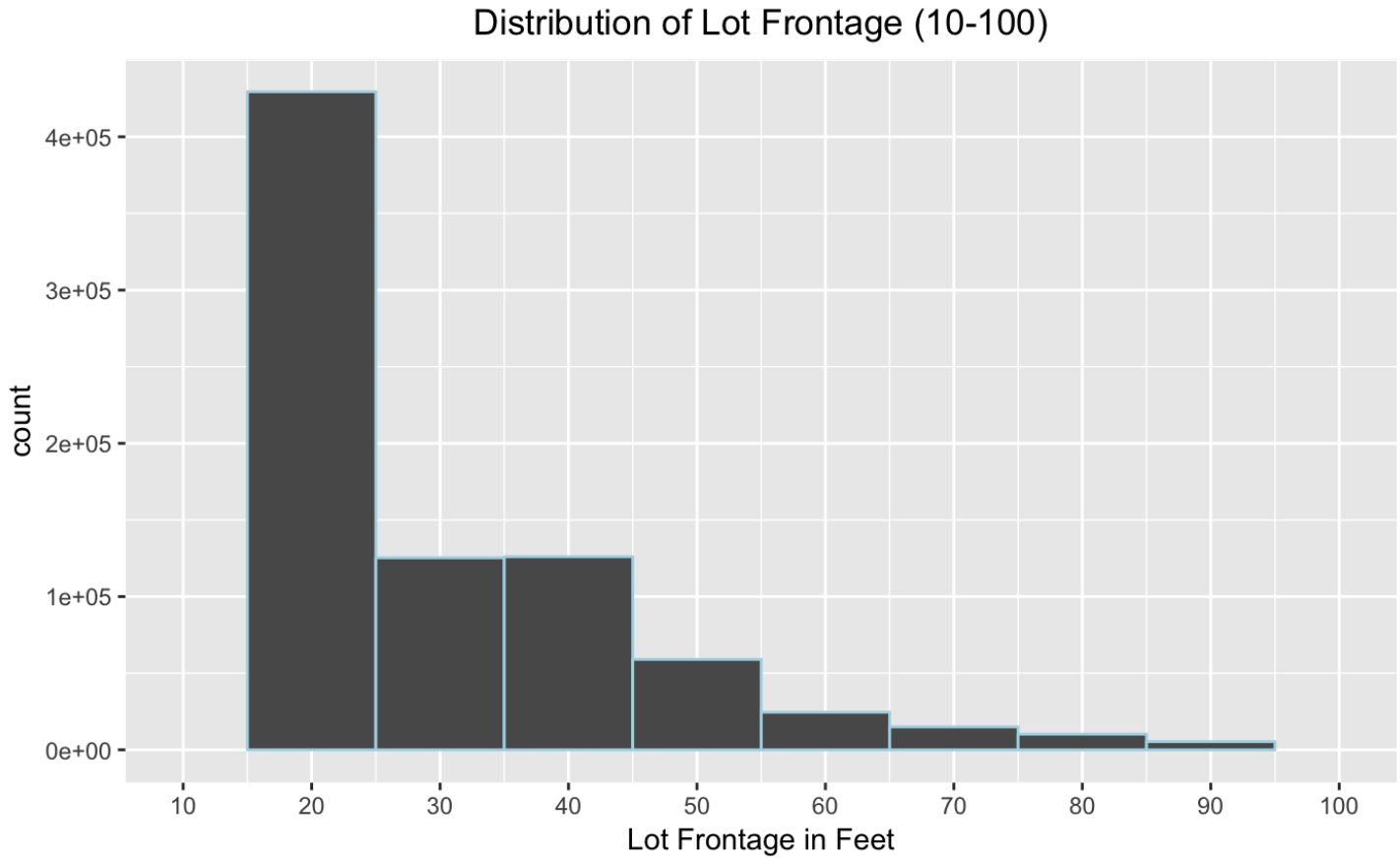


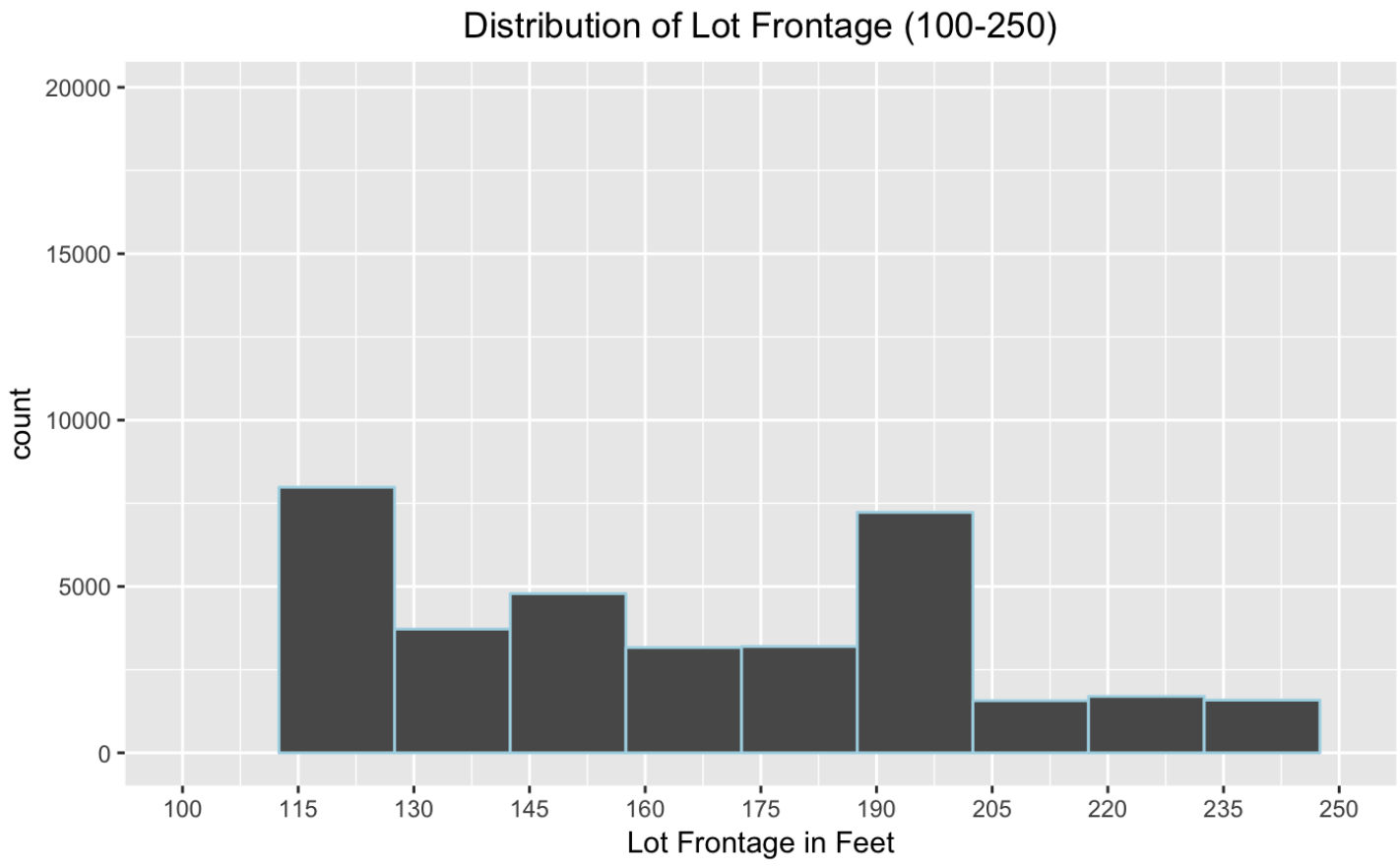
Variable 9	Description
LTFRONT	Lot frontage in feet (length 7 numeric)

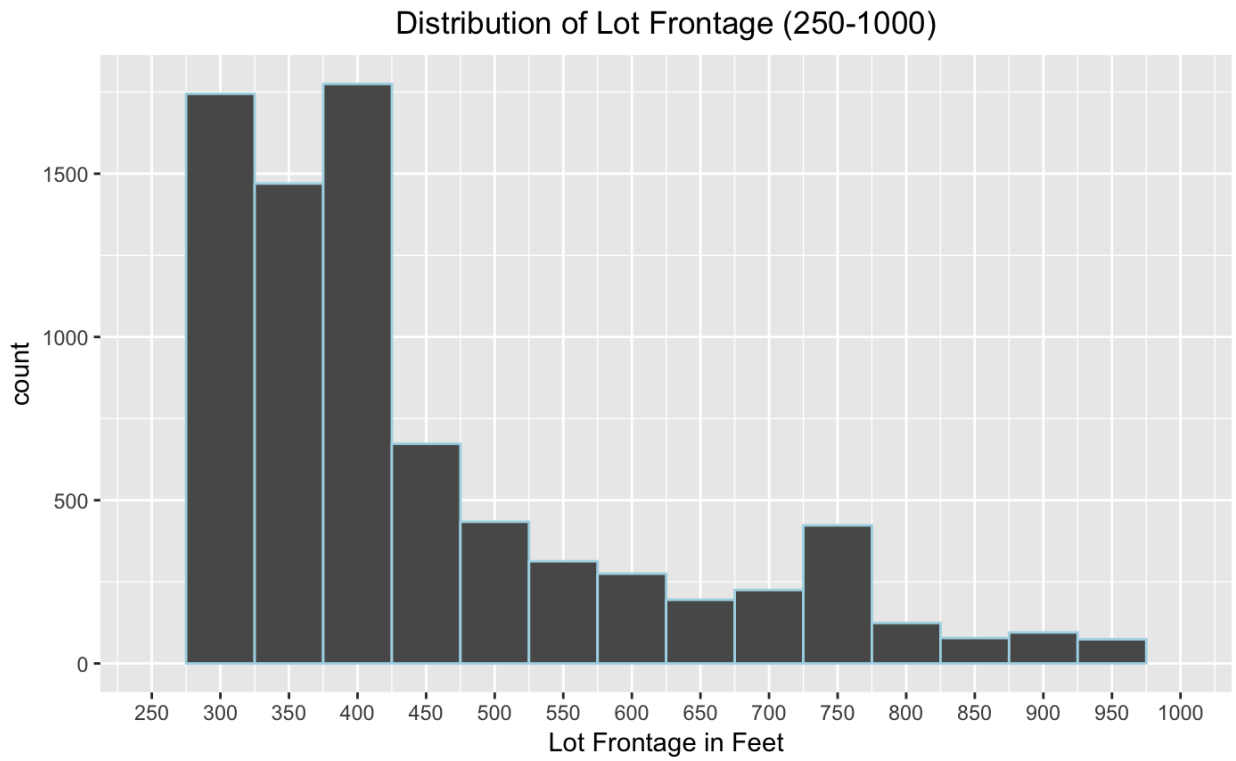
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*



Zoom in the distribution as shown figures on the next page：

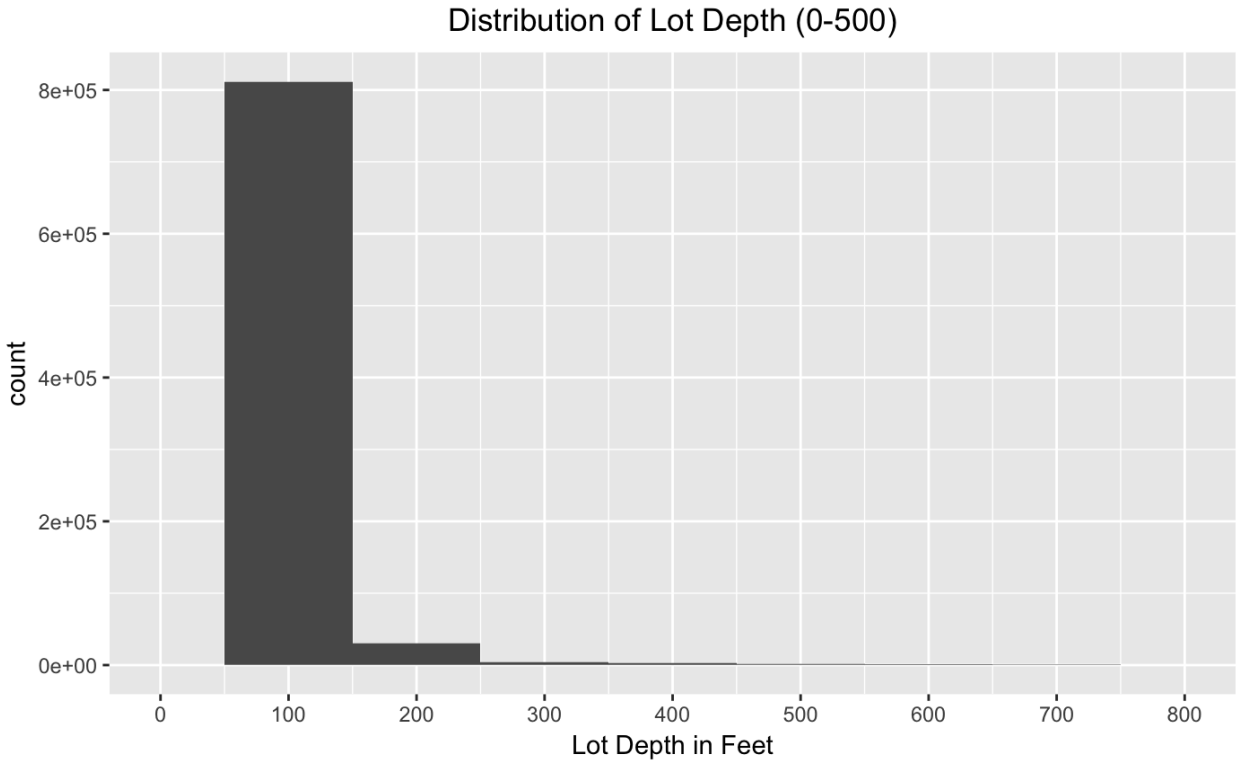




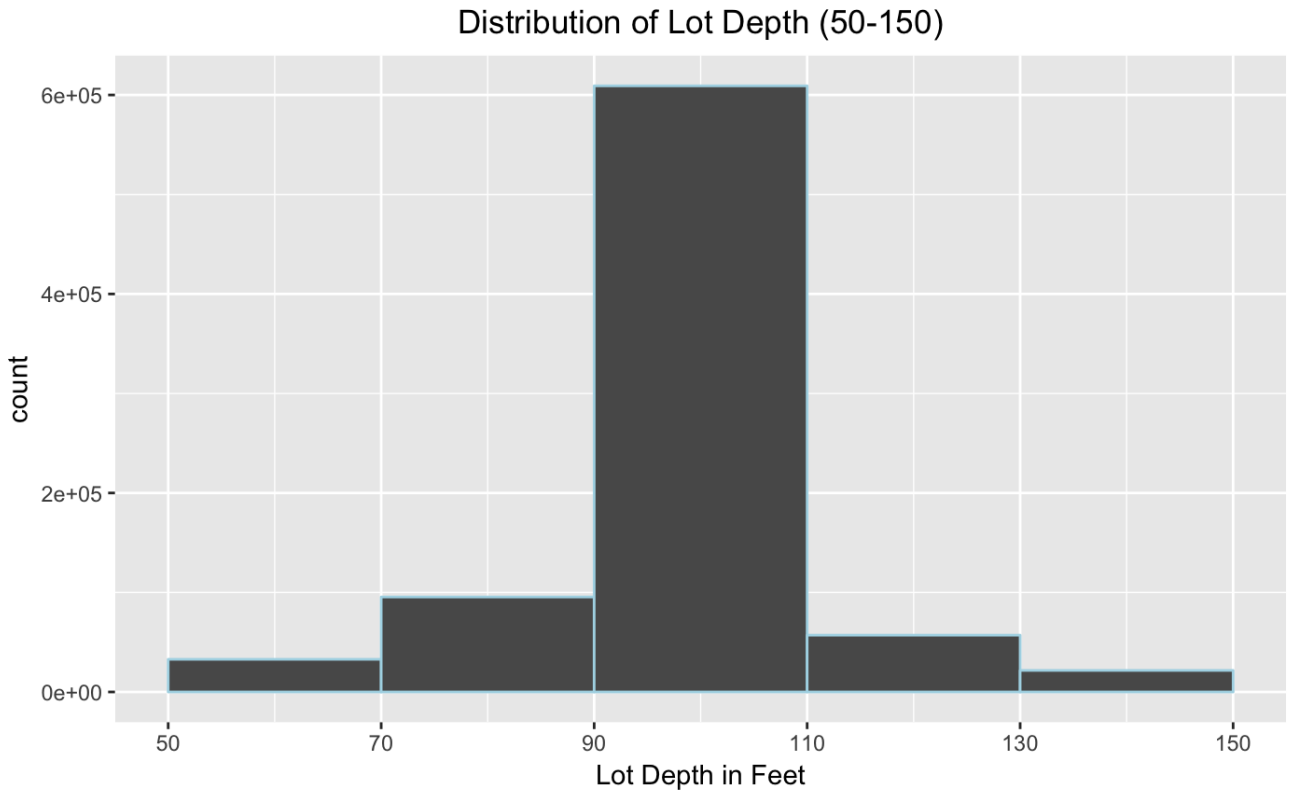


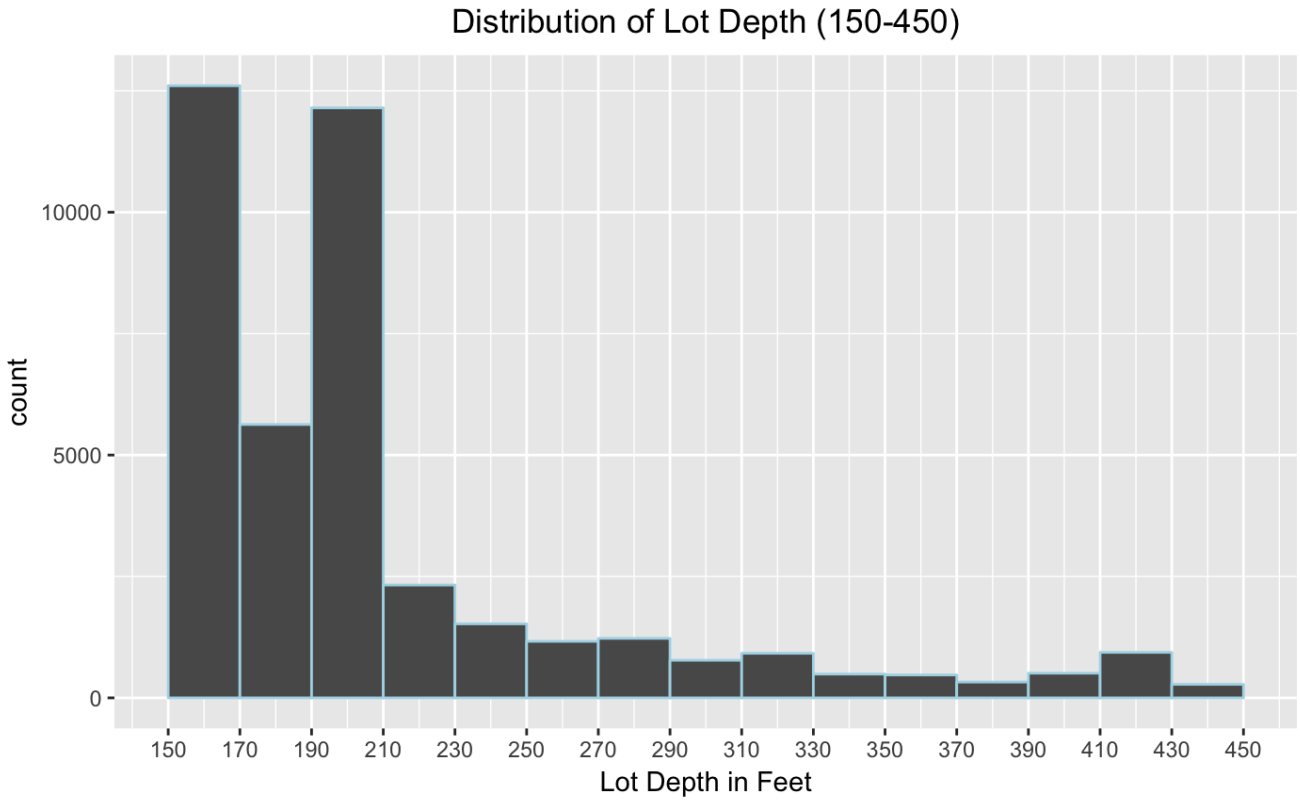
Variable 10	Description
LTDEPTH	Lot depth in feet (length 7 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-interval is selected to specify partial distribution.*



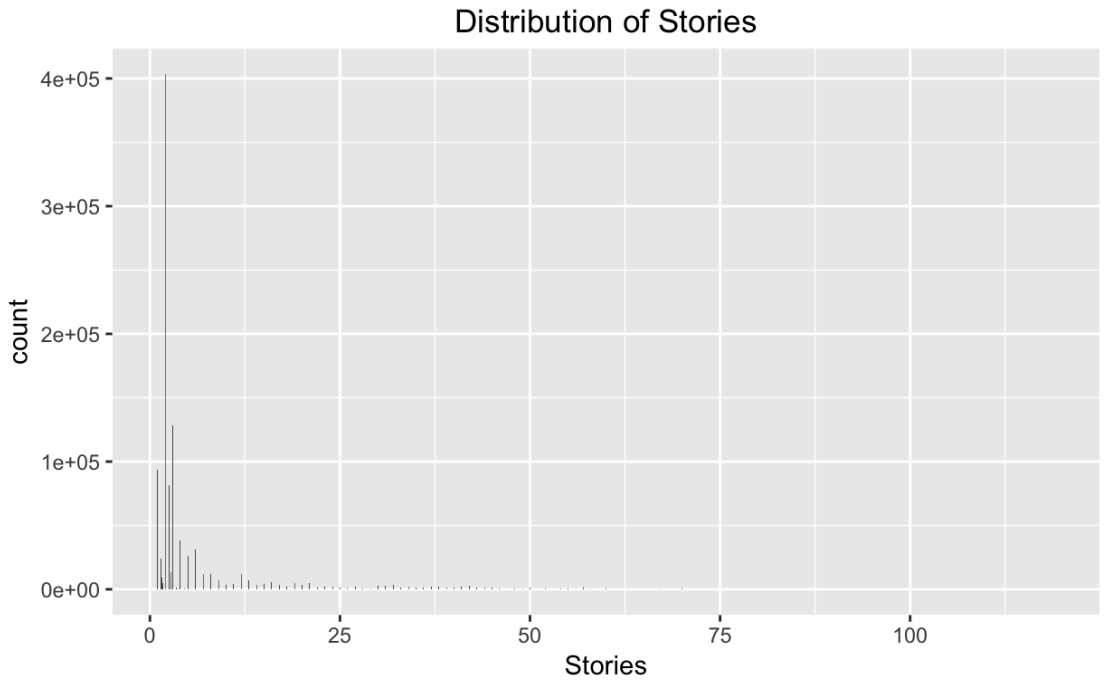
Zoom in the distribution as shown figures below：


Distribution of Lot Depth (50-150)
6e+05
4e+05
2e+05
0e+00
count
50
70
90
110
130
150
Lot Depth in Feet

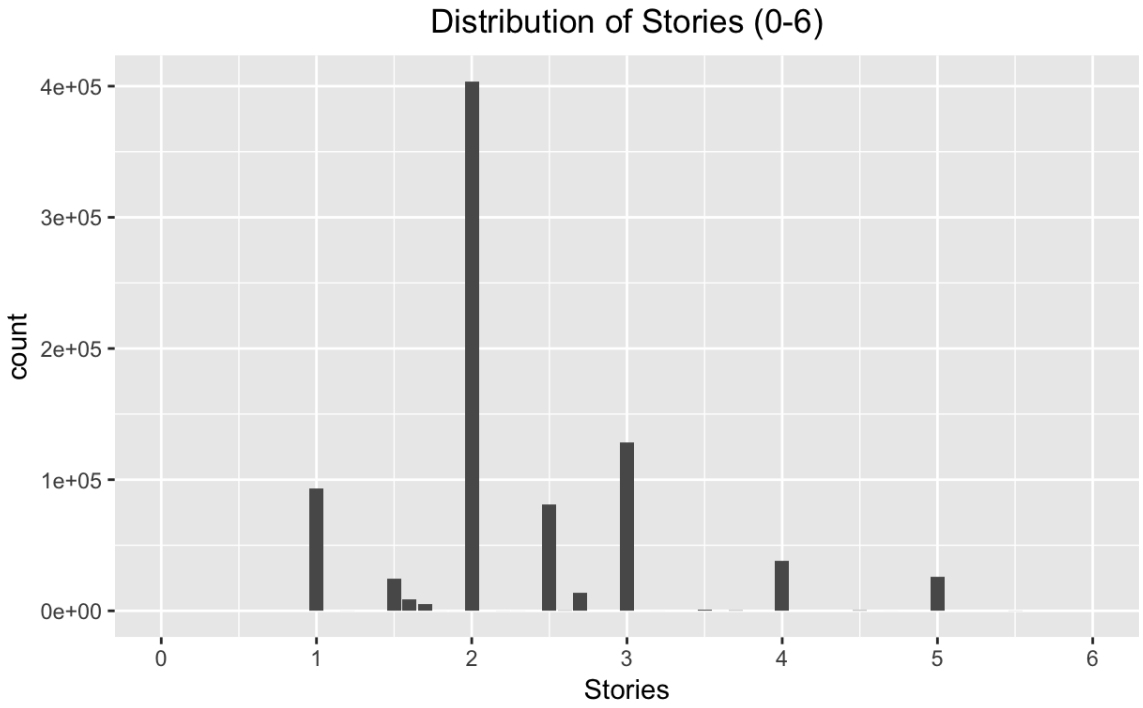

Distribution of Lot Depth (150-450)
10000
5000
0
count
150
170
190
210
230
250
270
290
310
330
350
370
390
410
430
450
Lot Depth in Feet

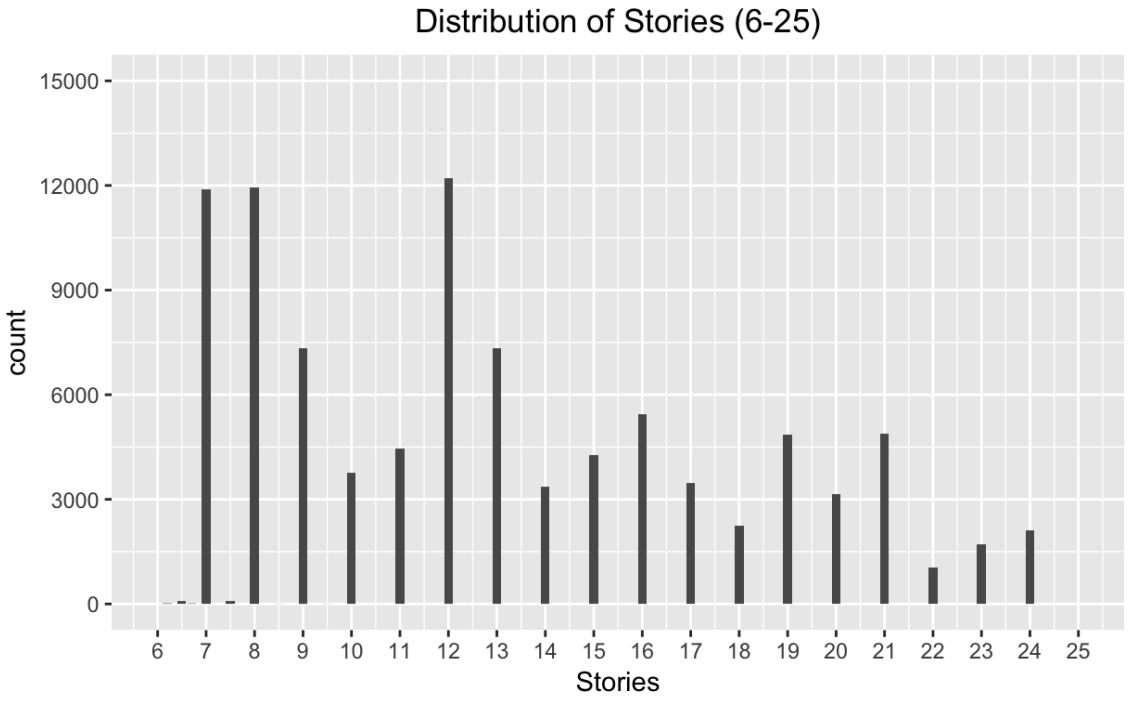
Variable 11	Description
STORIES	The number of stories for the building (# of Floors). (length 5 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*



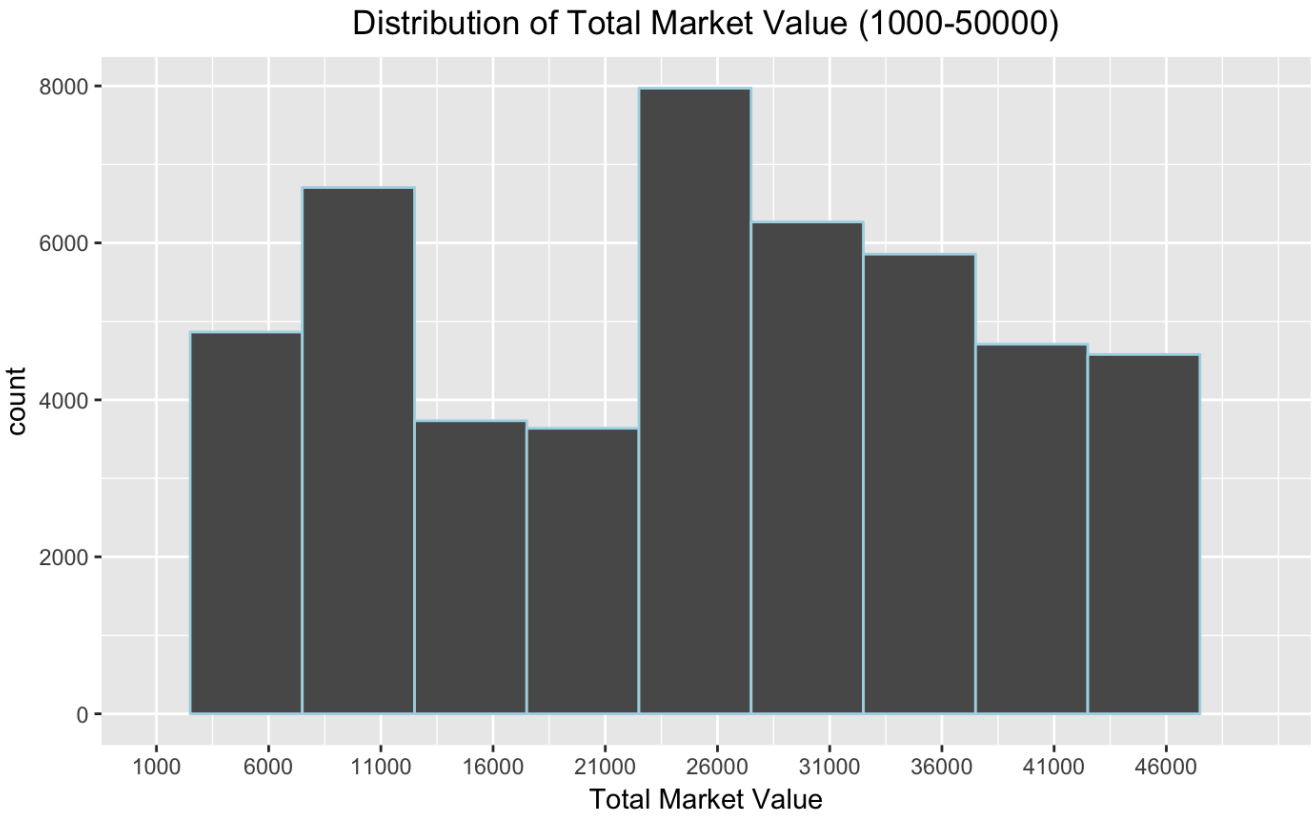
Zoom in the distribution as shown figures below：

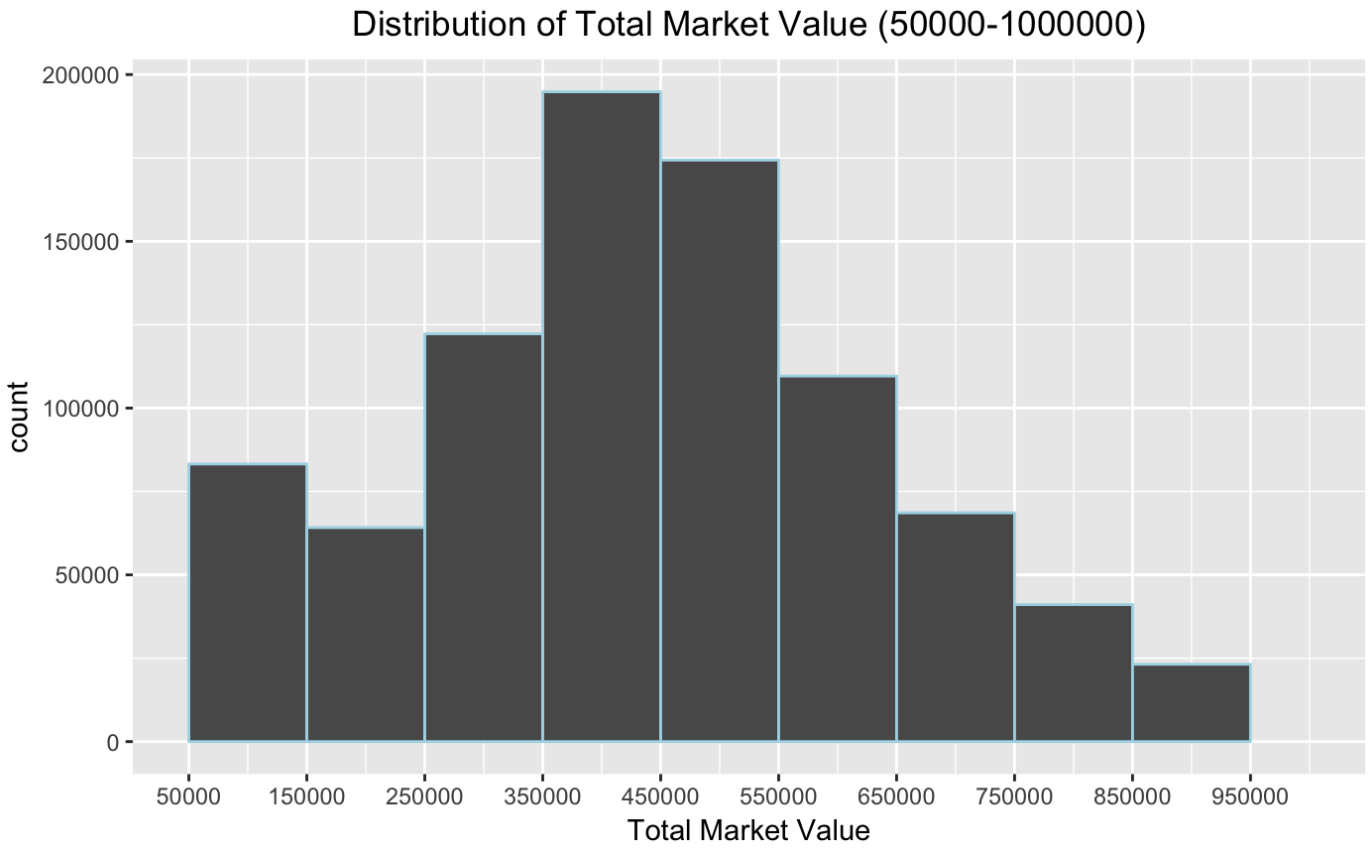


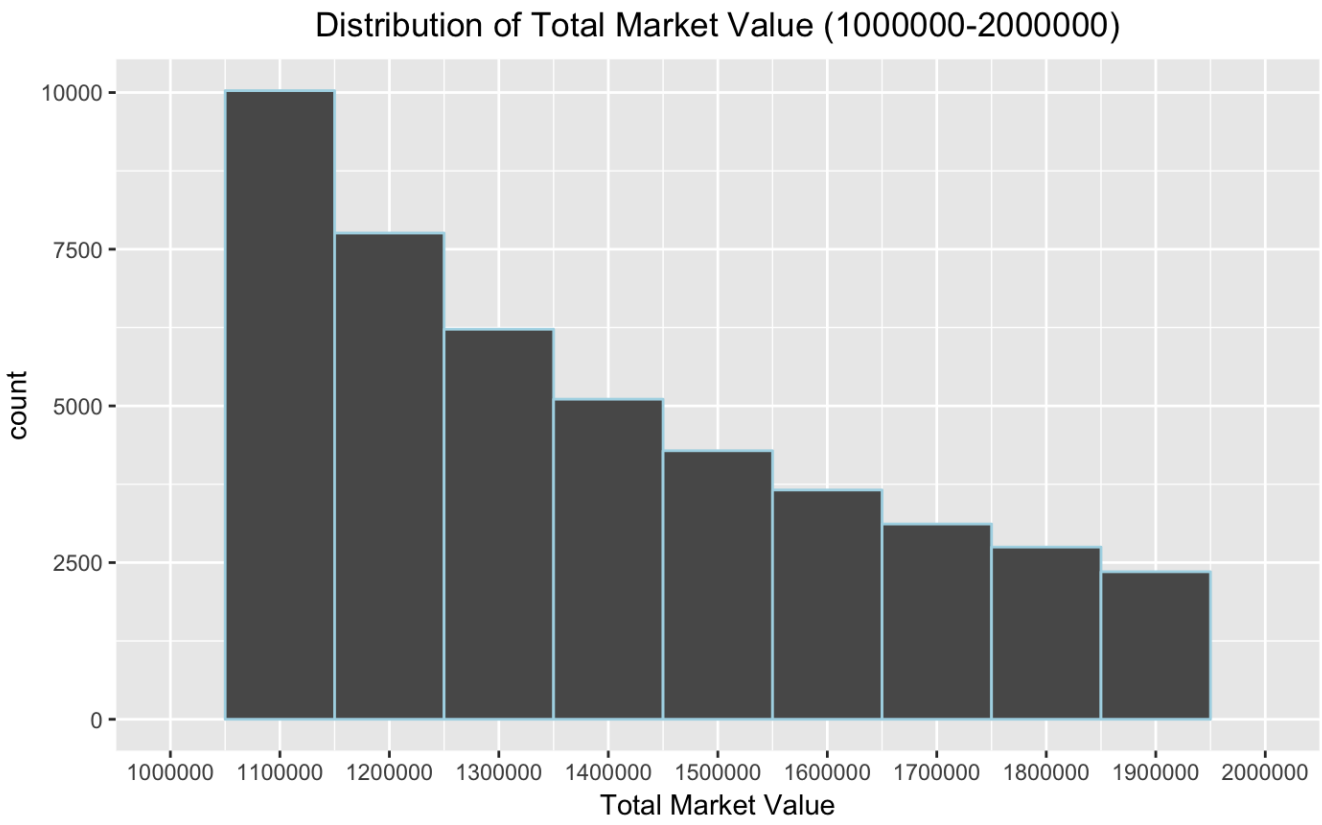


Variable 12	Description
FULLVAL	Total market value (length 11 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*

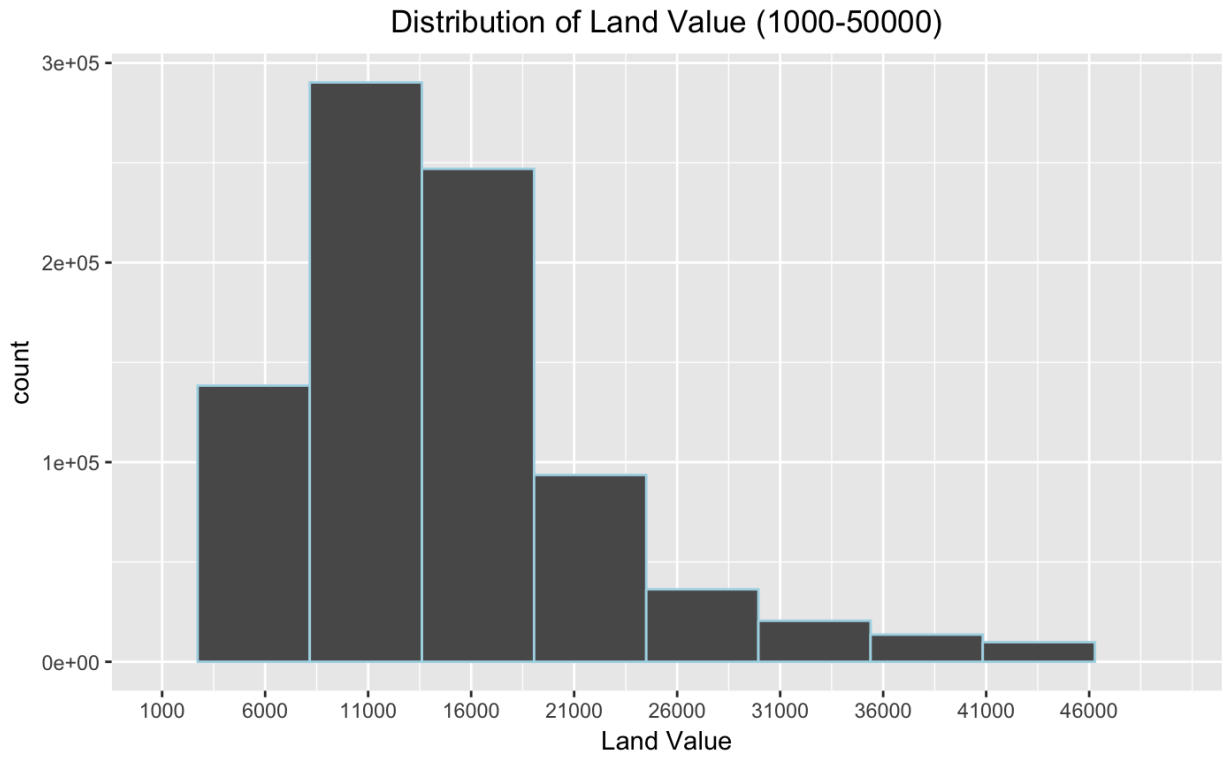


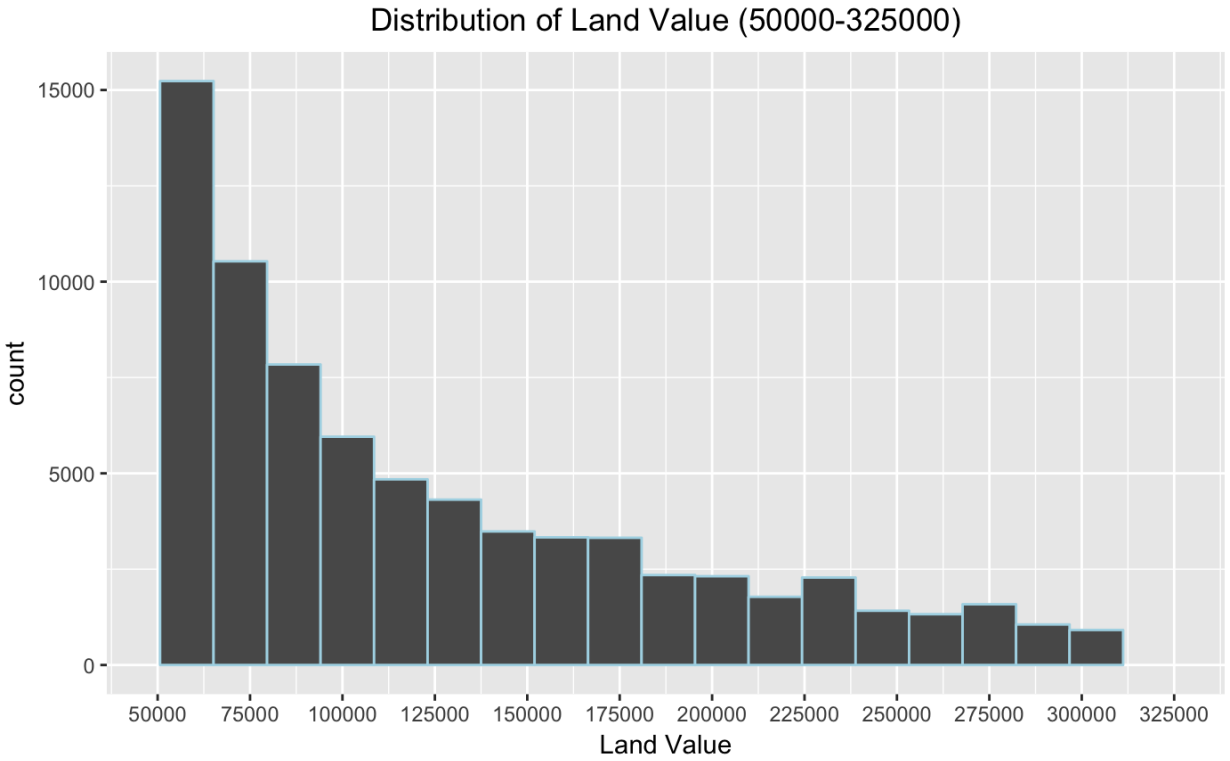

Distribution of Total Market Value (50000-1000000)
count
200000
150000
100000
50000
0
50000
150000
250000
350000
450000
550000
650000
750000
850000
950000
Total Market Value


Distribution of Total Market Value (1000000-2000000)
count
10000
7500
5000
2500
0
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
Total Market Value

Variable 13	Description
AVLAND	Market value of the land (length 11 numeric)

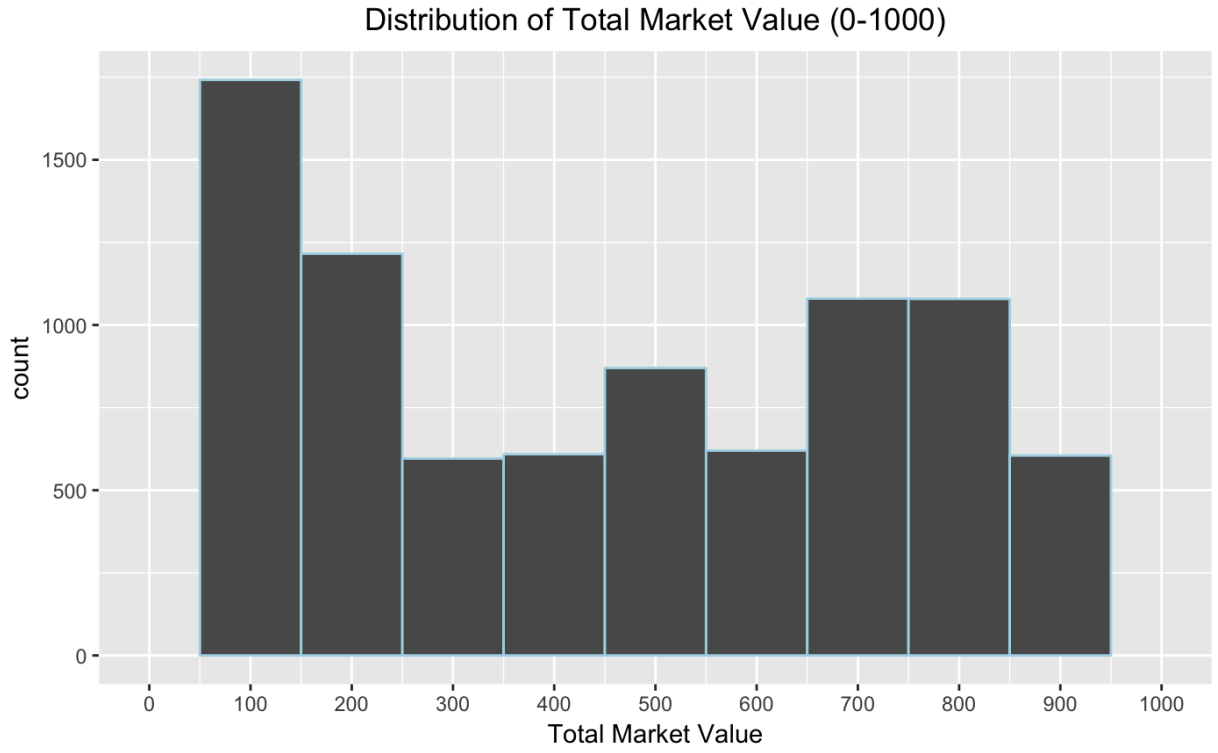
Note: This variable exhibits severe right-skew distribution. Sub-interval is selected to specify partial distribution.*

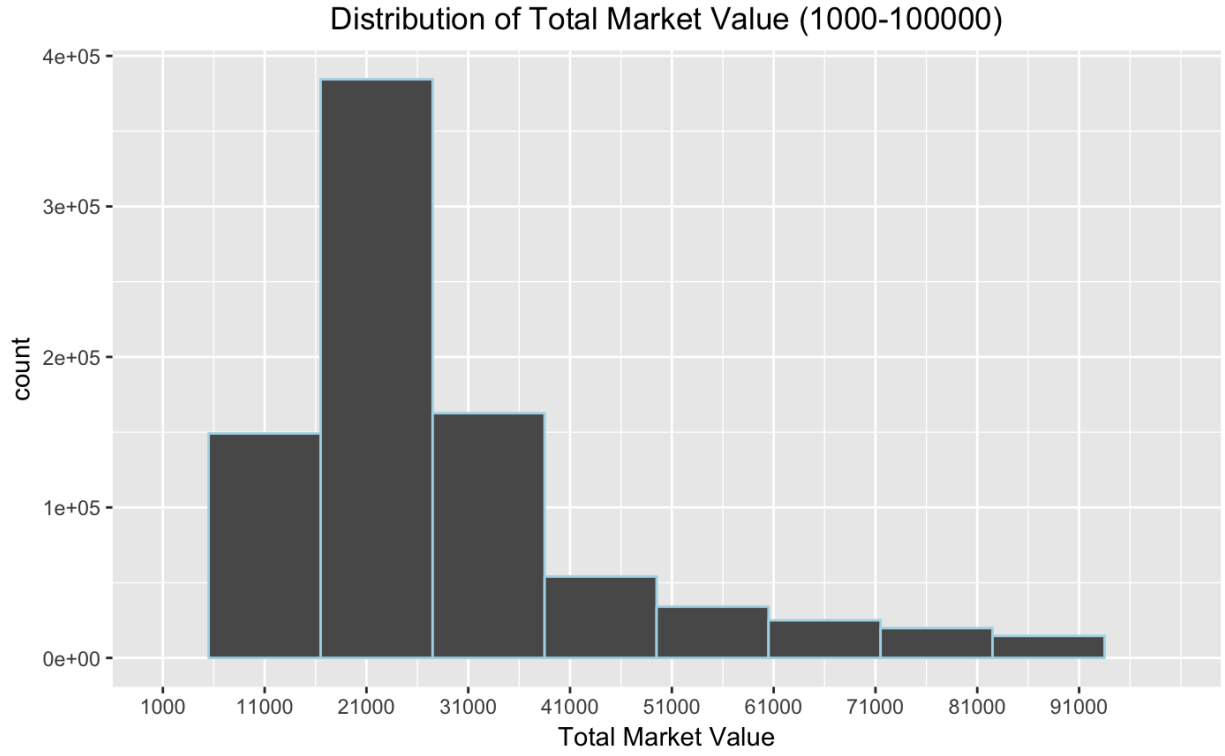


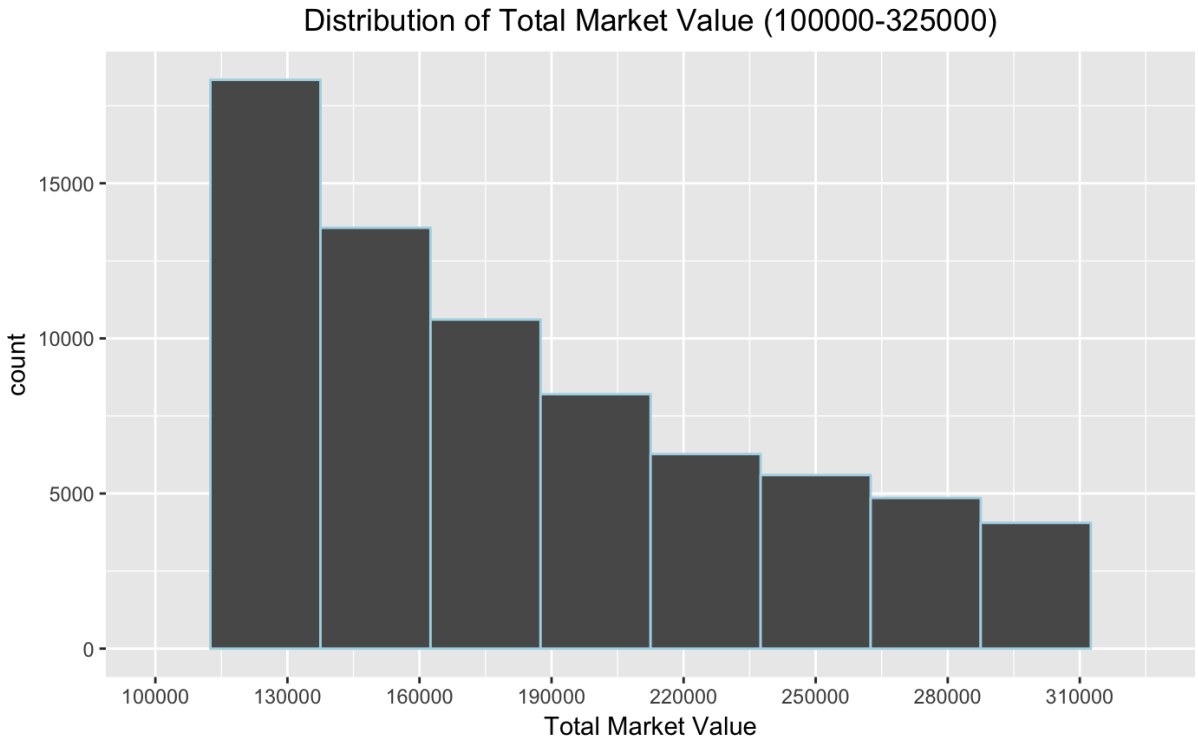


Variable 14	Description
AVTOT	Current year's total market value (length 11 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*

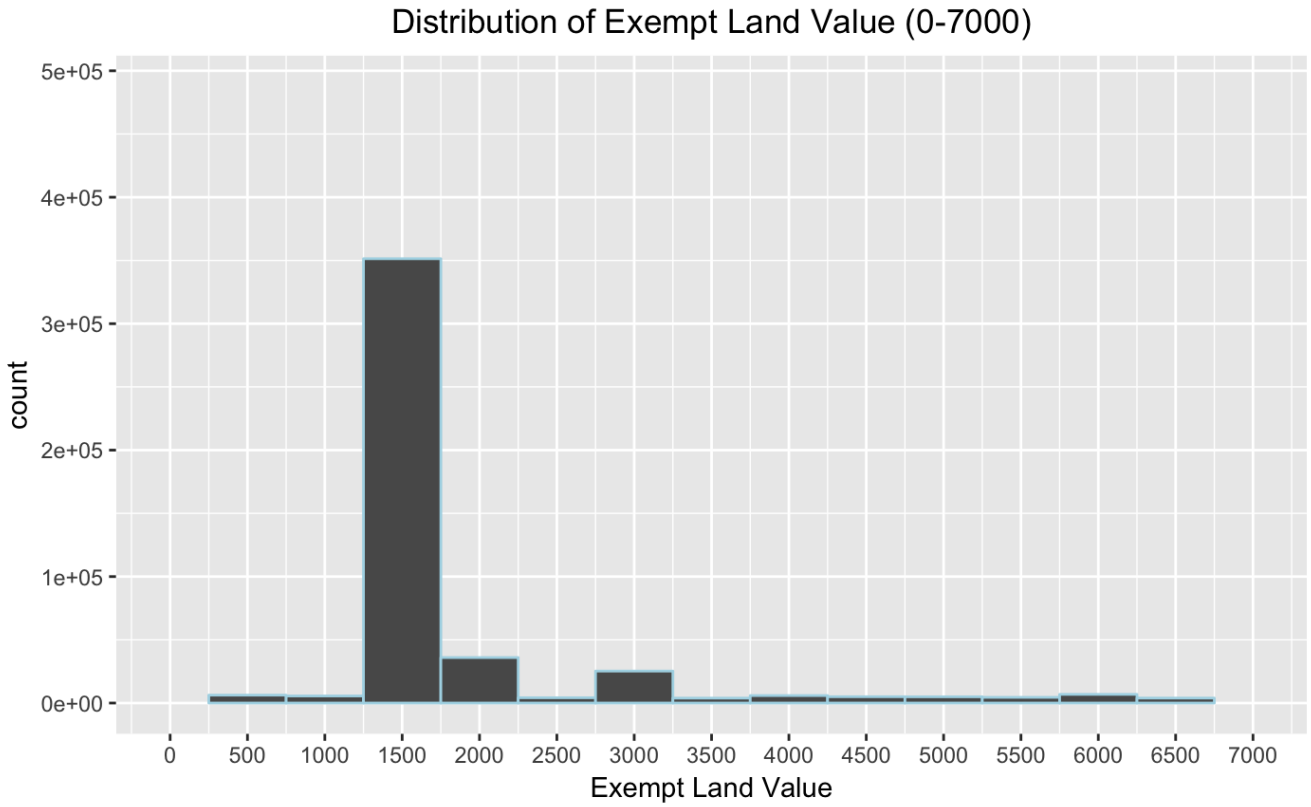


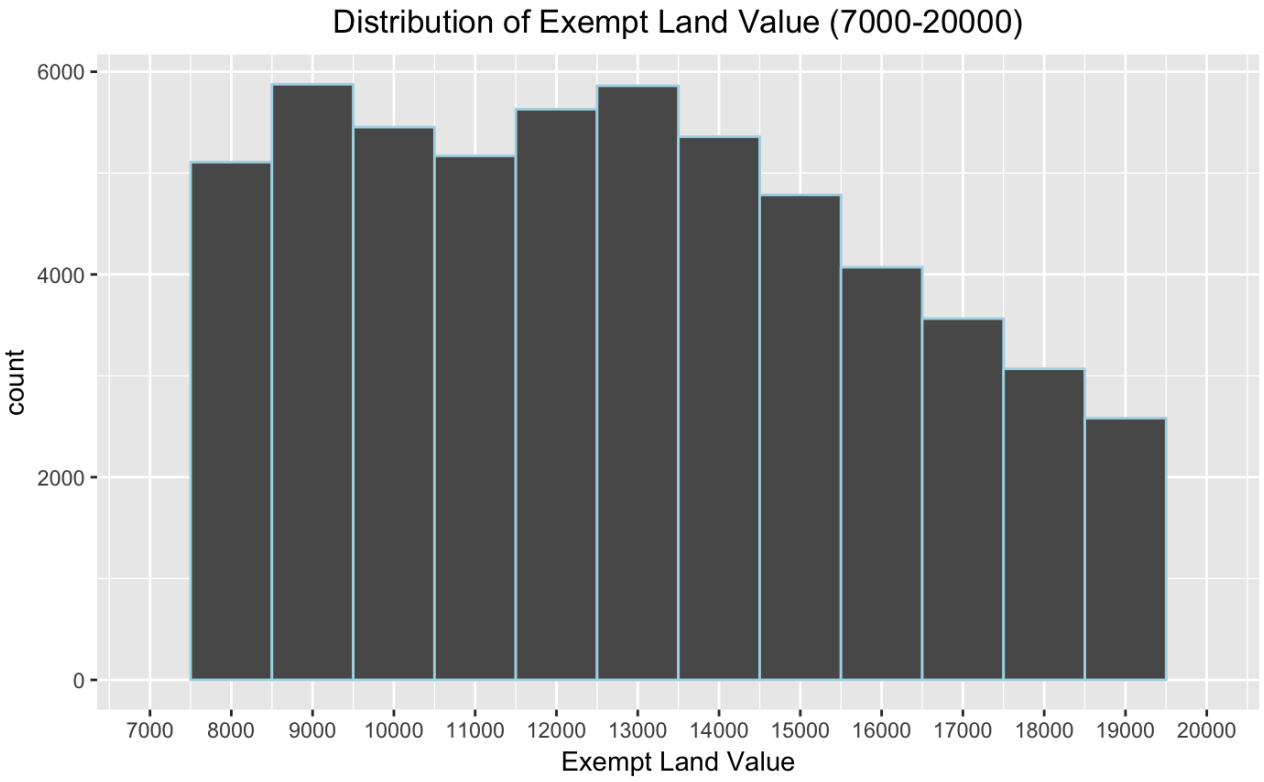




Variable 15	Description
EXLAND	Exempt land value (length 11 numeric)

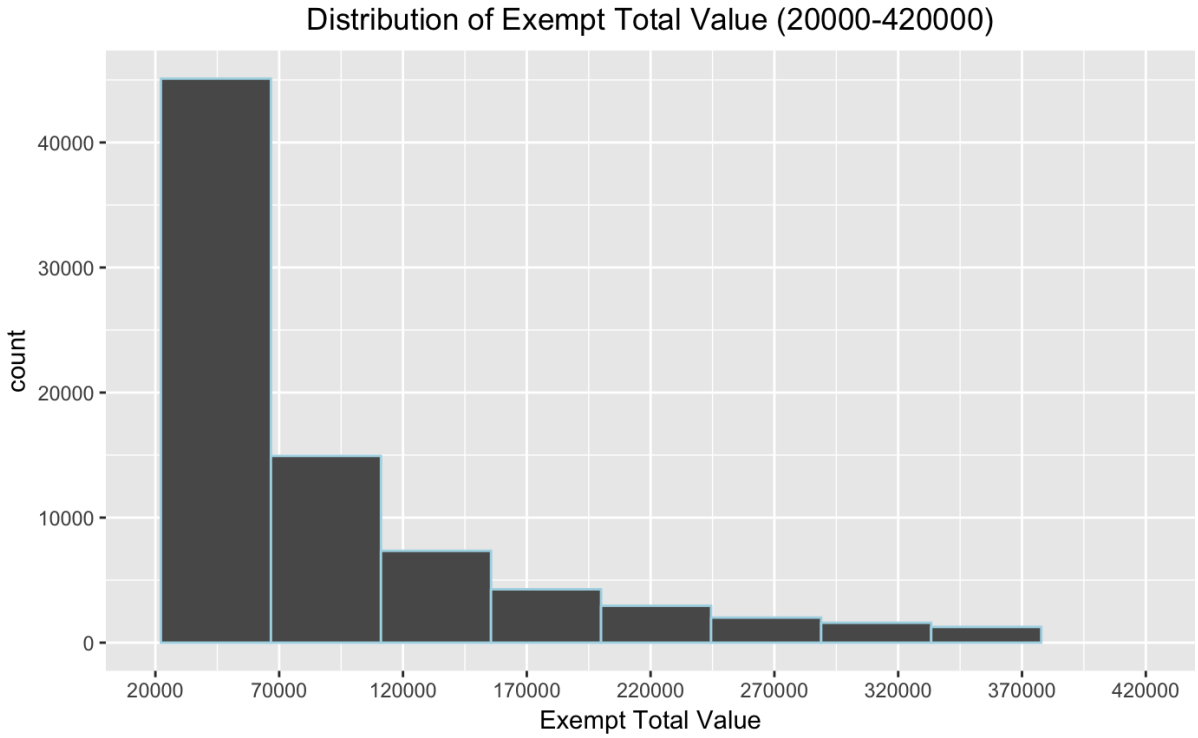
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution*

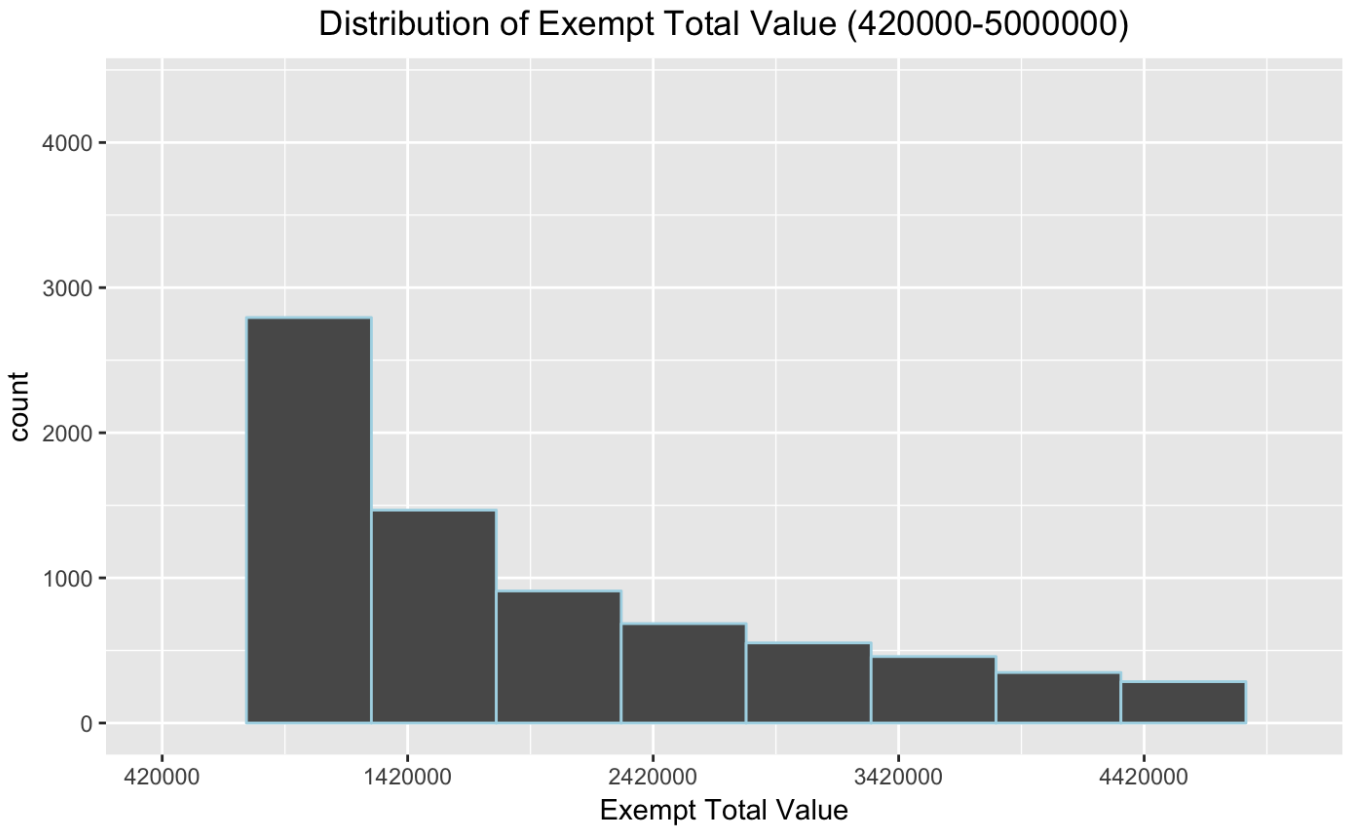




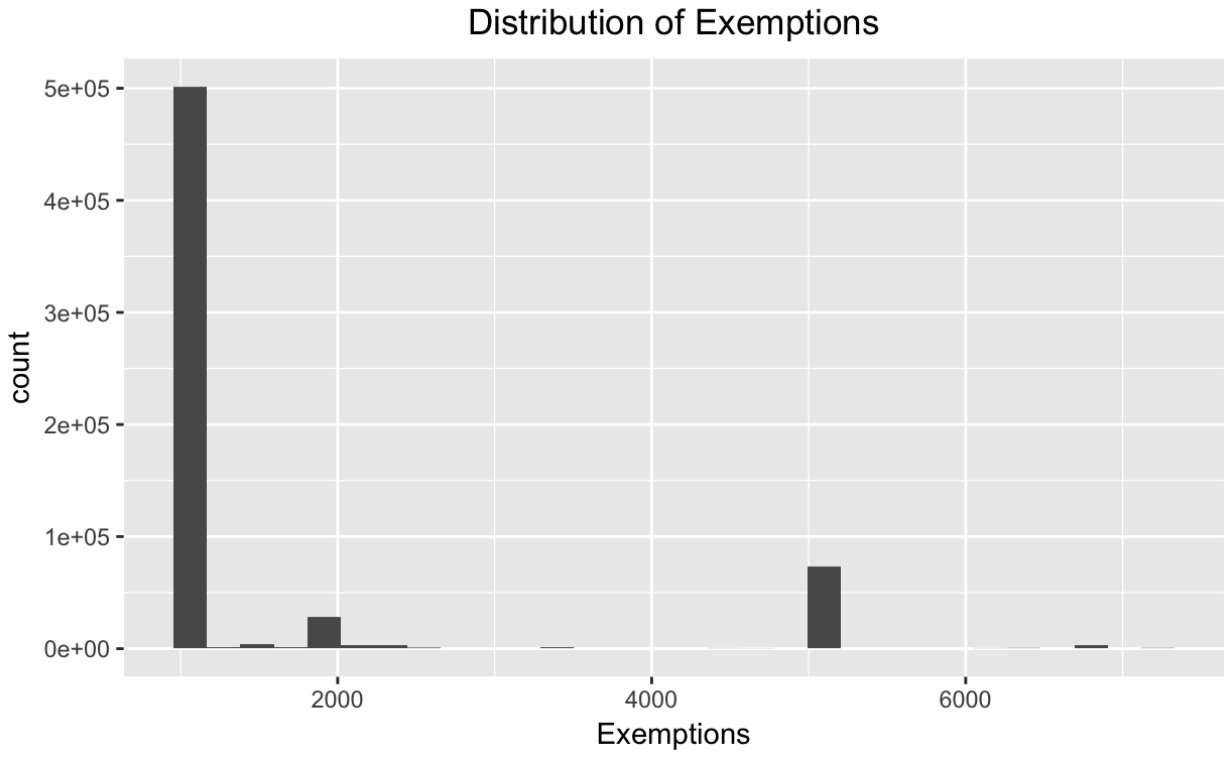
Variable 16	Description
EXTOT	Exempt total value (length 11 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*



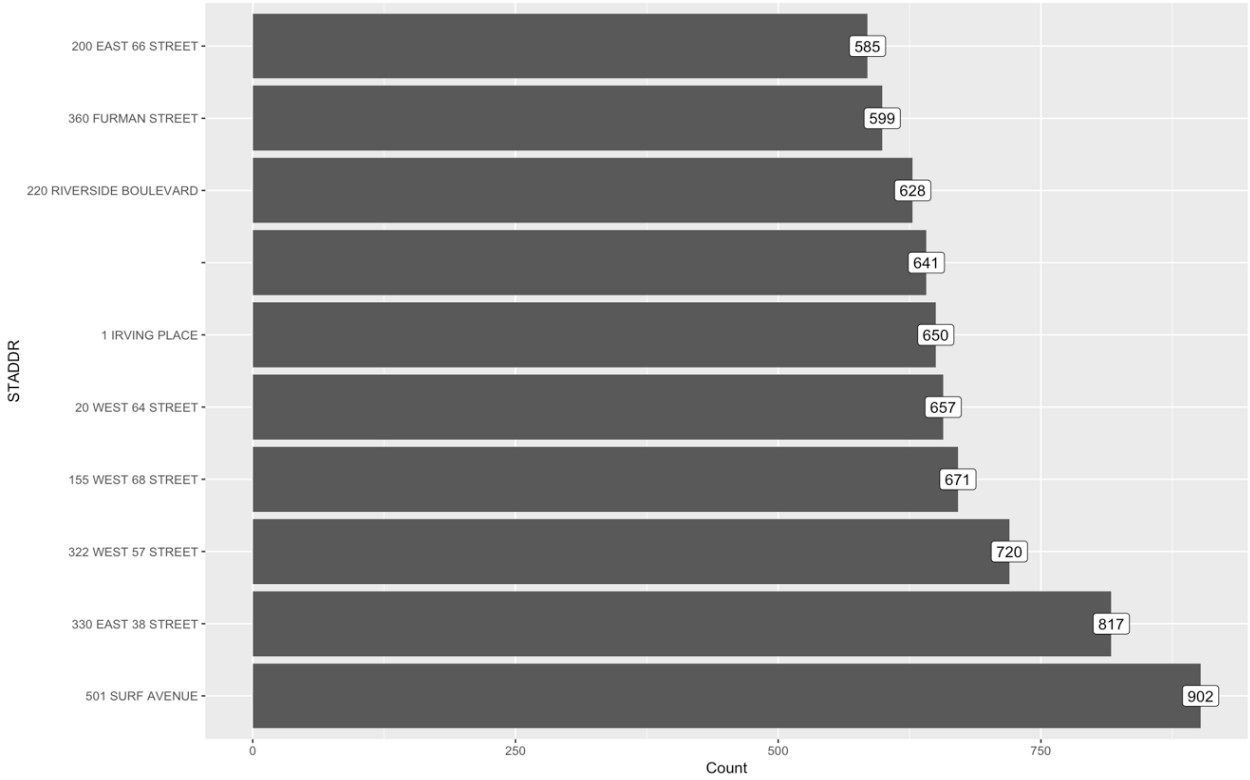


Variable 17	Description
EXCD1	Number of exemptions on the property



Variable 18	Description
STADDR	Street name of the property (length 20 Character) #NA: 641

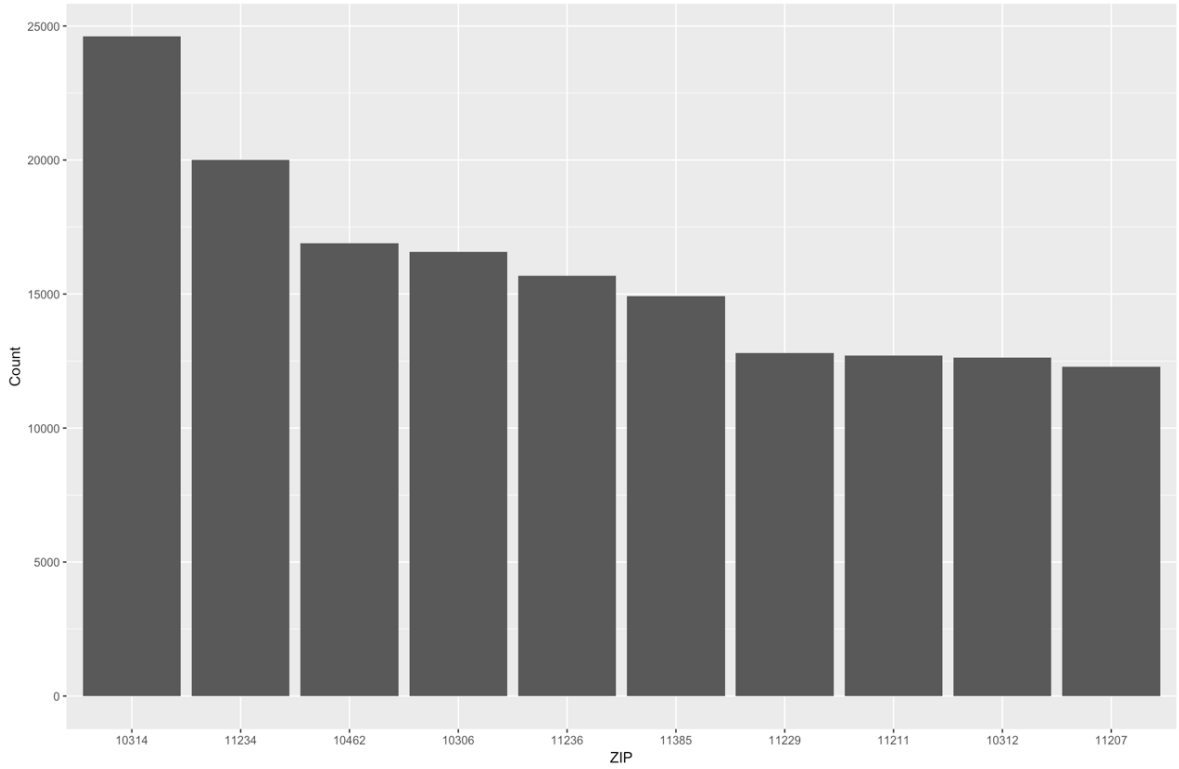
Top 10 Address Chart:



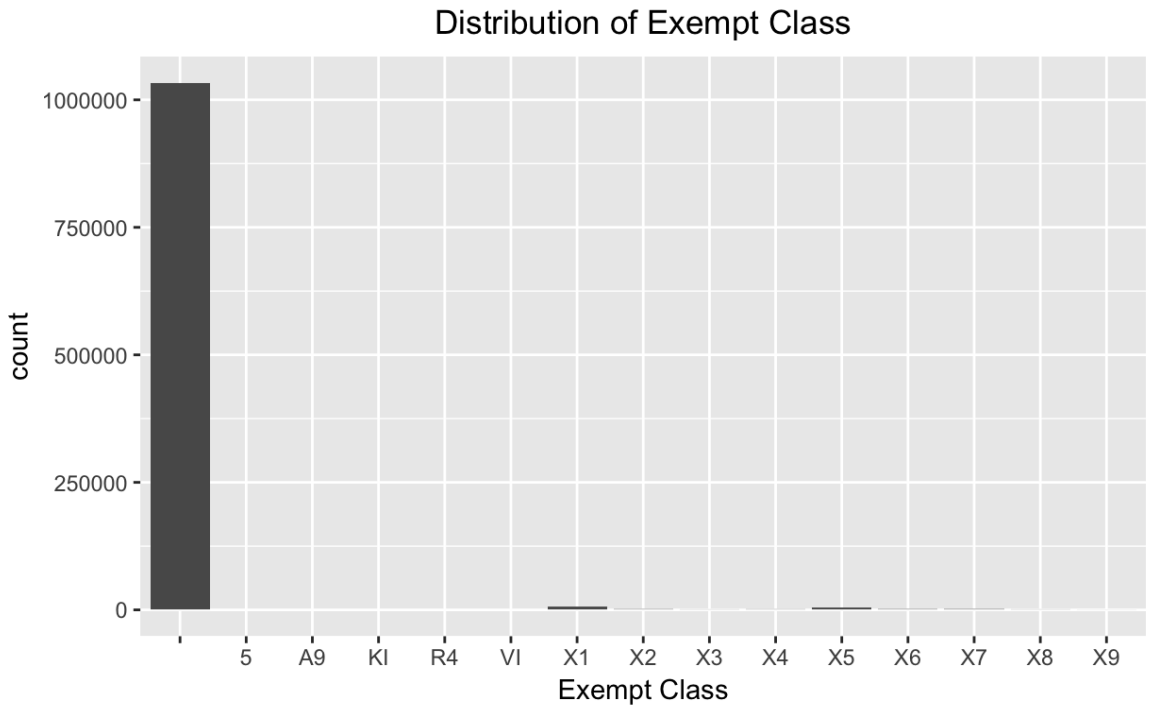
Variable 19	Description
ZIP	Postal ZIP code of the property (length 5 numeric) #NA: 26356

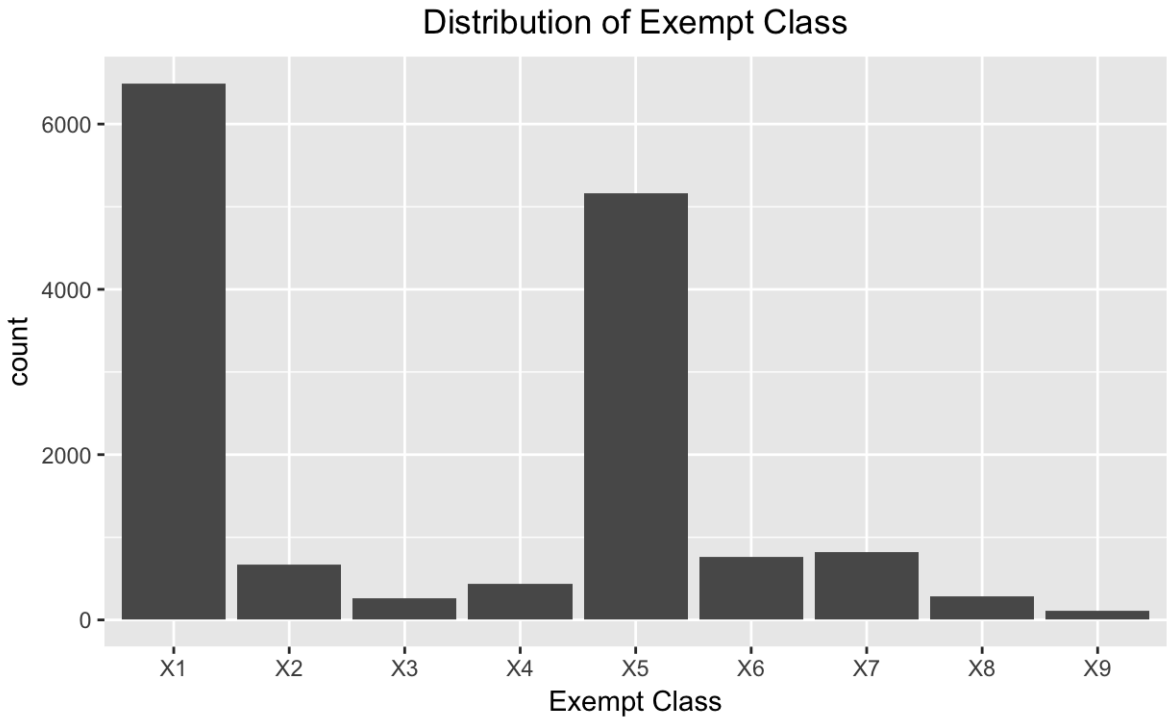
Top 10 ZIP Frequency and Distribution:

Zip	Weight (%)
10314	2.407018
11234	1.956626
10462	1.653755
10306	1.62157
11236	1.533722
11385	1.459668
11229	1.251493
11211	1.243373
10312	1.235939
11207	1.20258



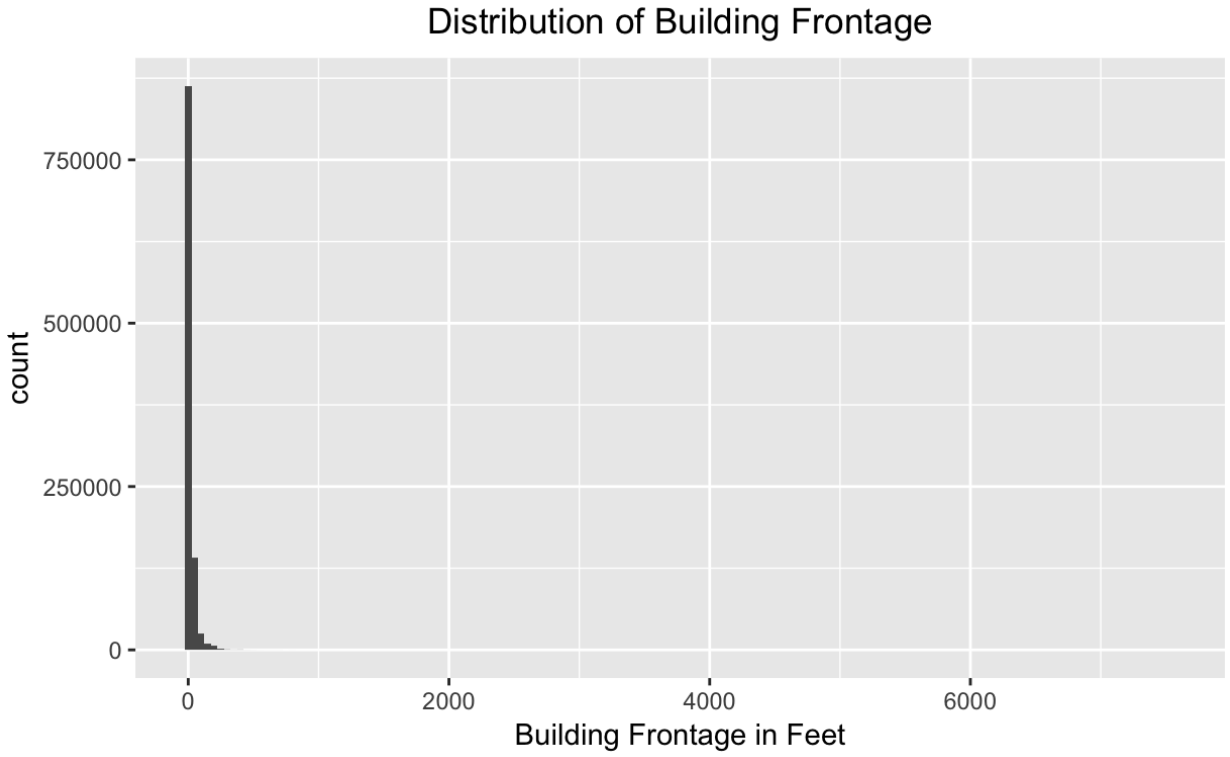
Variable 20	Description
EXEMPTCL	Exempt Class used for fully exempt properties only. (length 2 character) 'X1 - X9'.

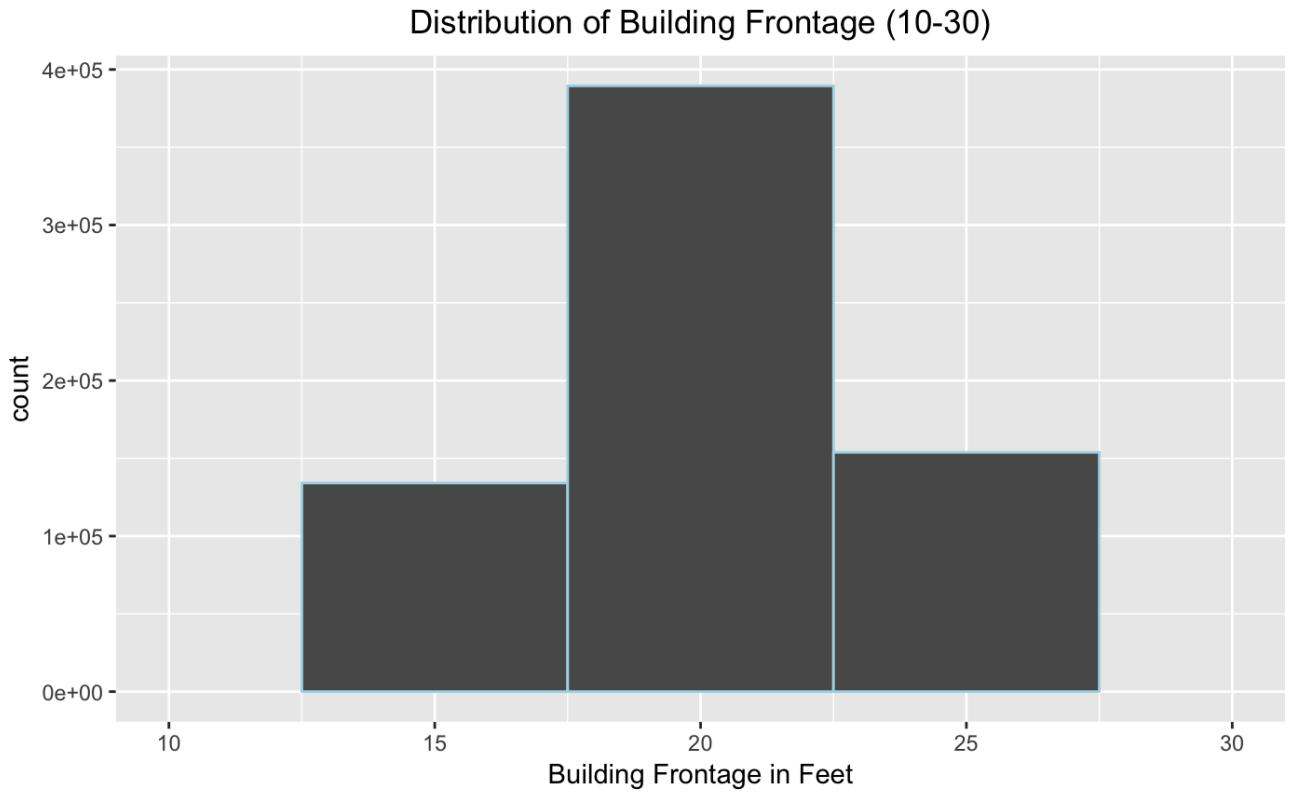


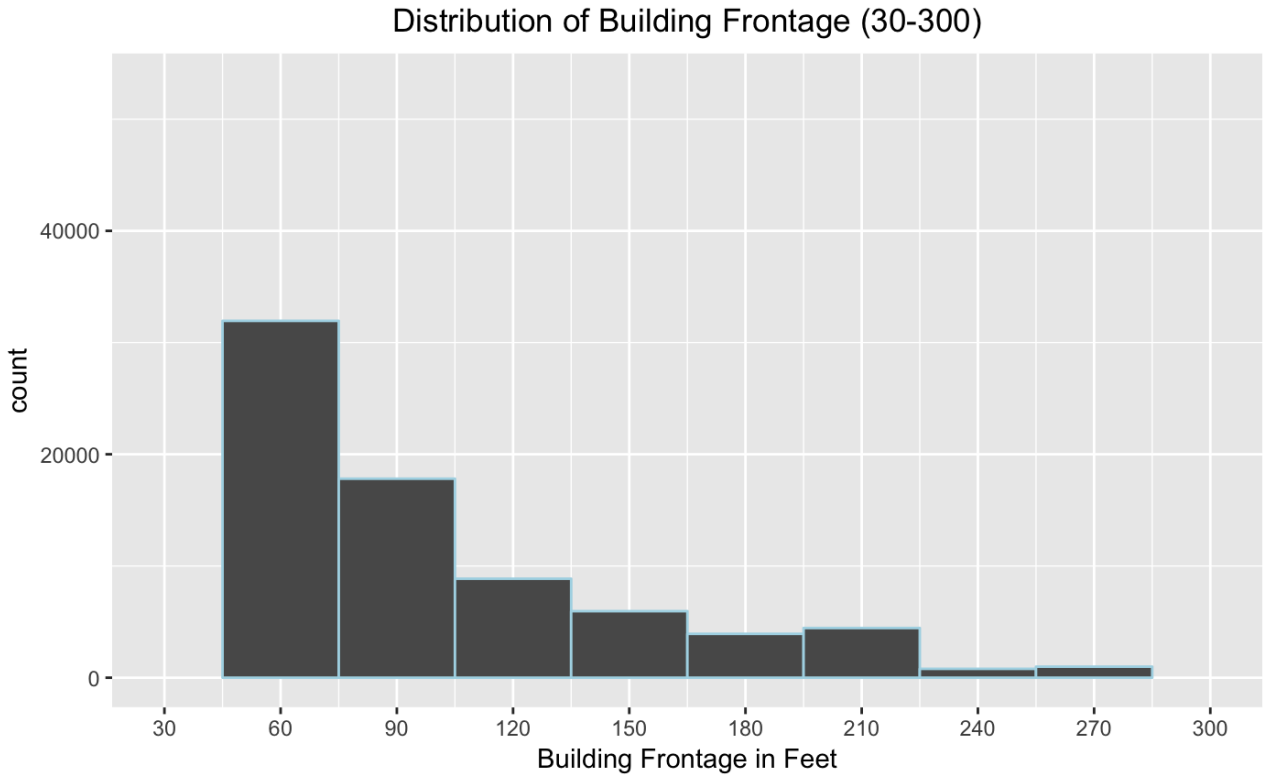


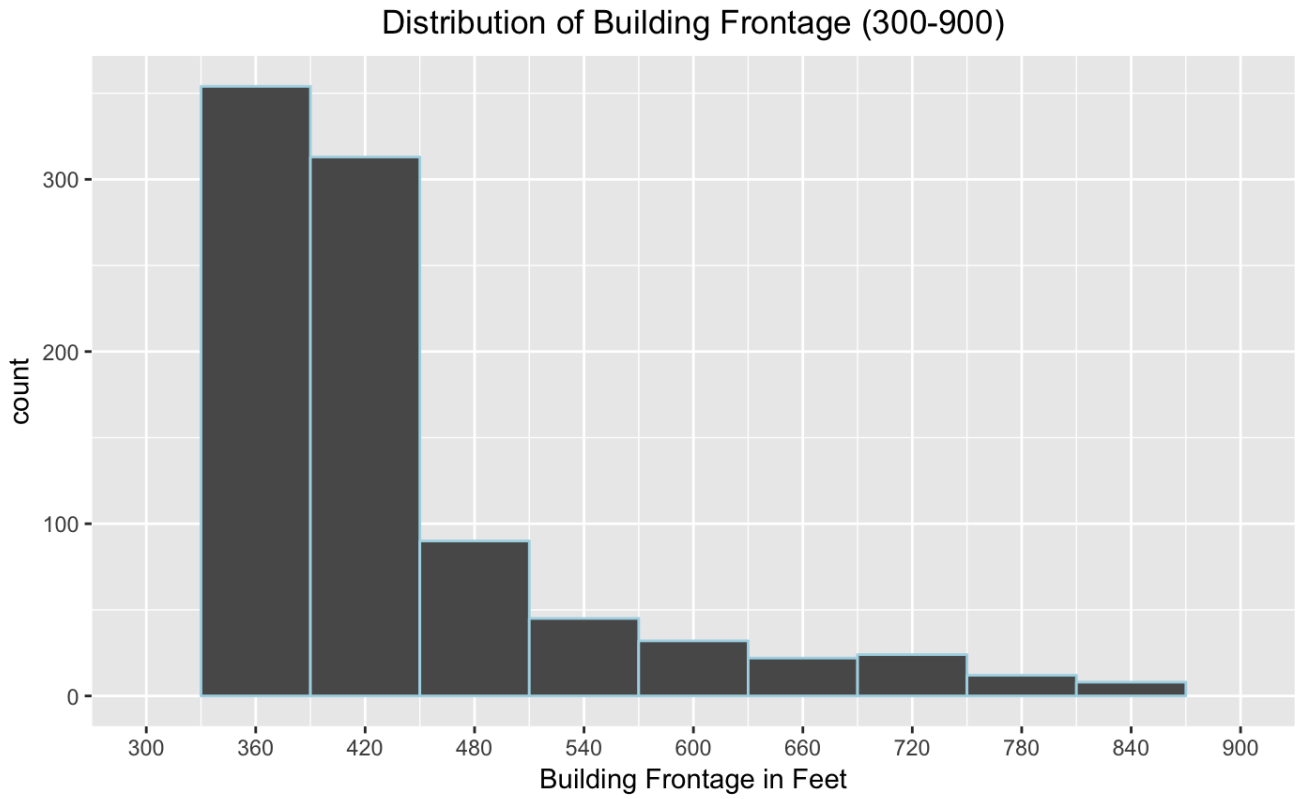
Variable 21	Description
BLDFRONT	Building frontage in feet (length 7 numeric)

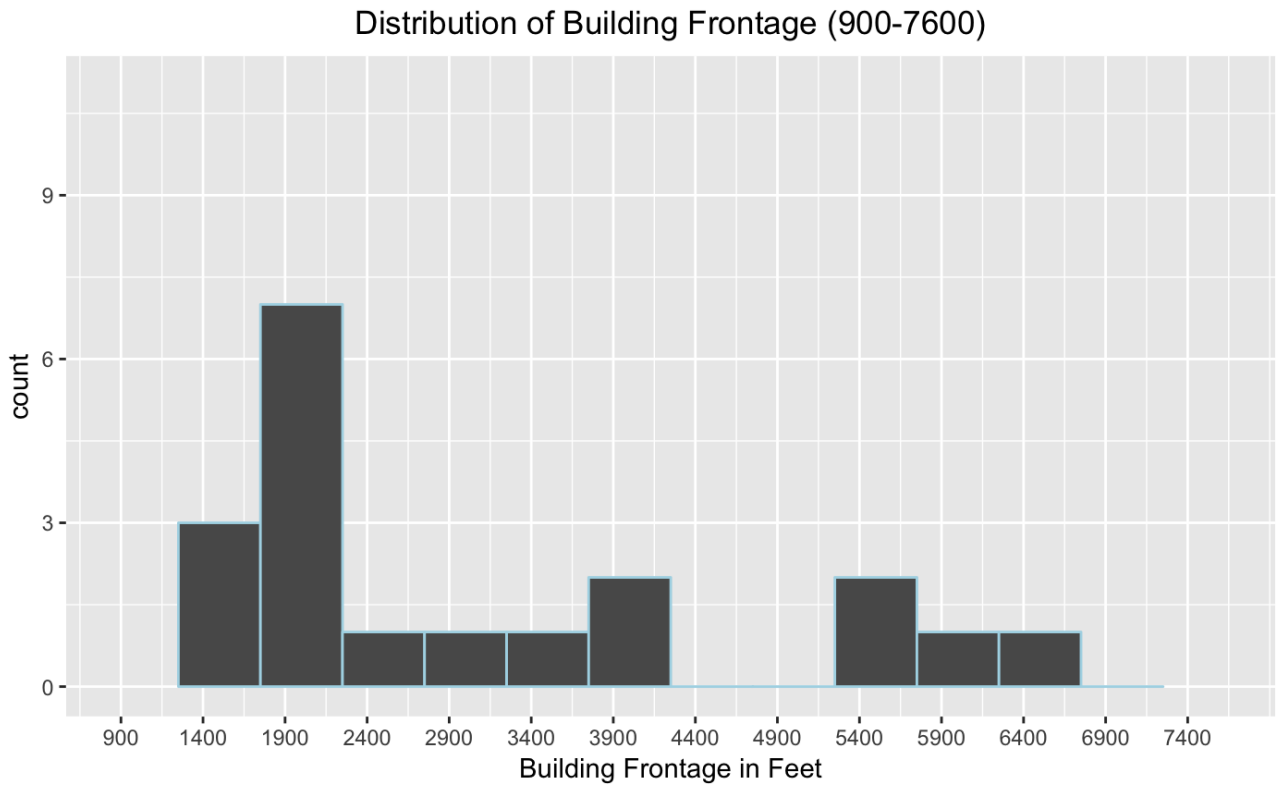
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution*




Distribution of Building Frontage (10-30)
4e+05
3e+05
2e+05
1e+05
0e+00
count
10
15
20
25
30
Building Frontage in Feet

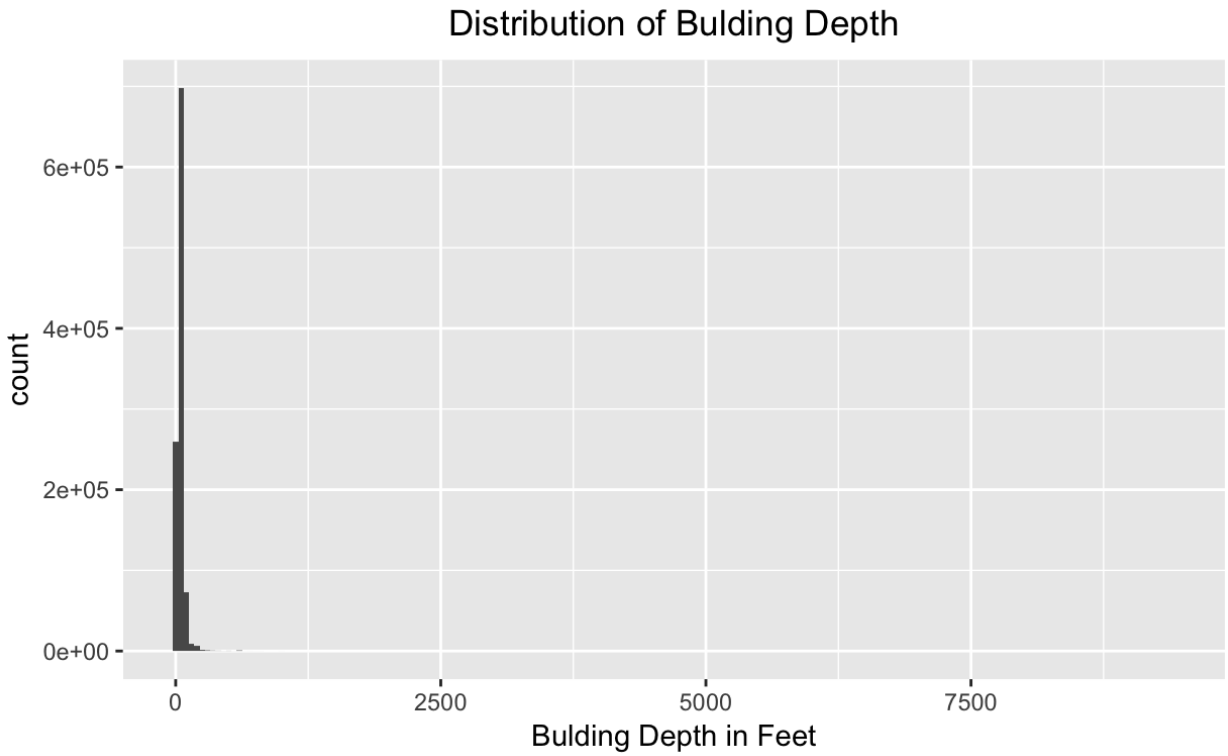

Distribution of Building Frontage (30-300)
40000
20000
0
count
30
60
90
120
150
180
210
240
270
300
Building Frontage in Feet

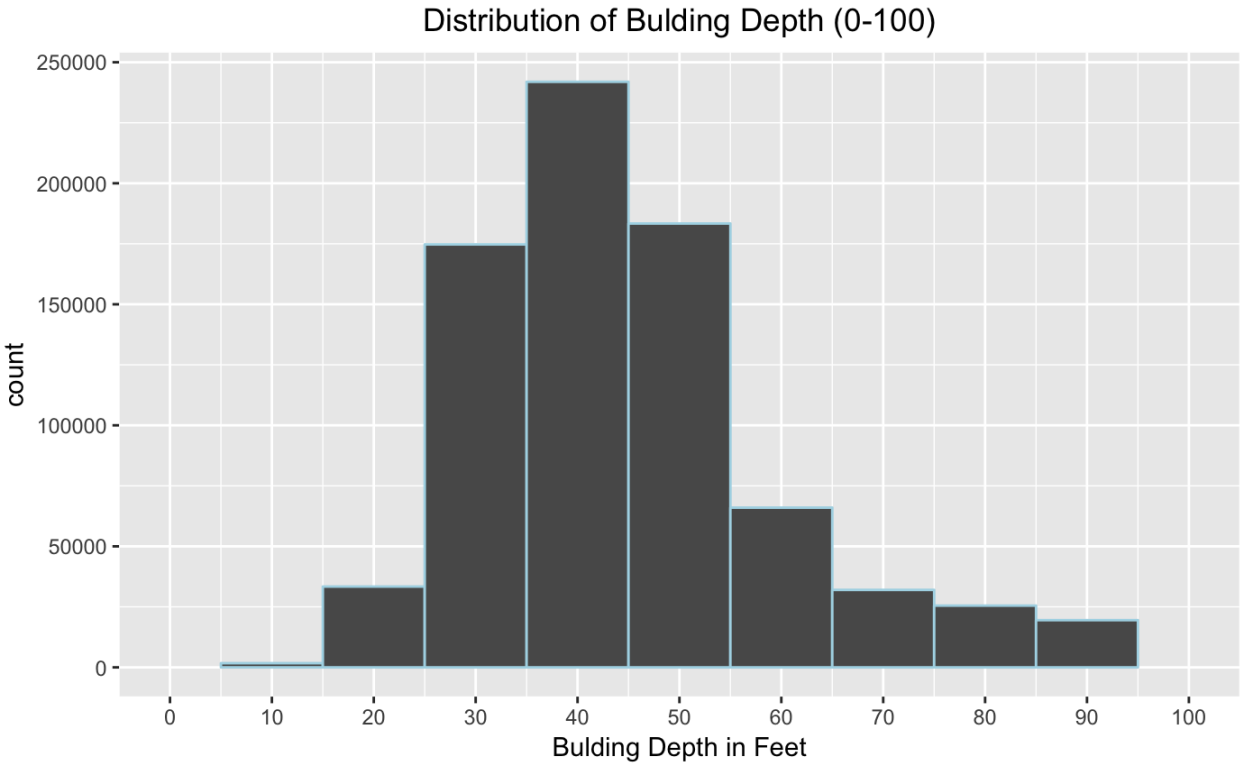


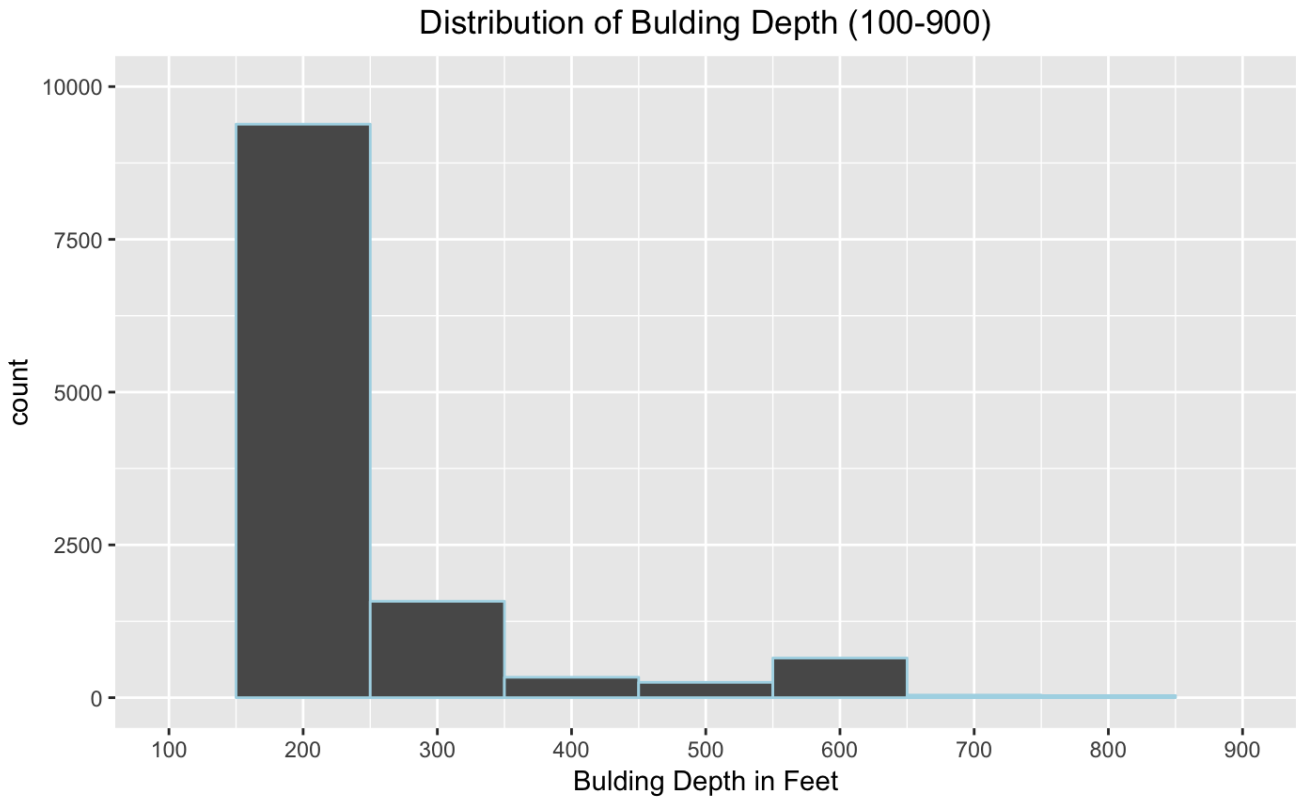


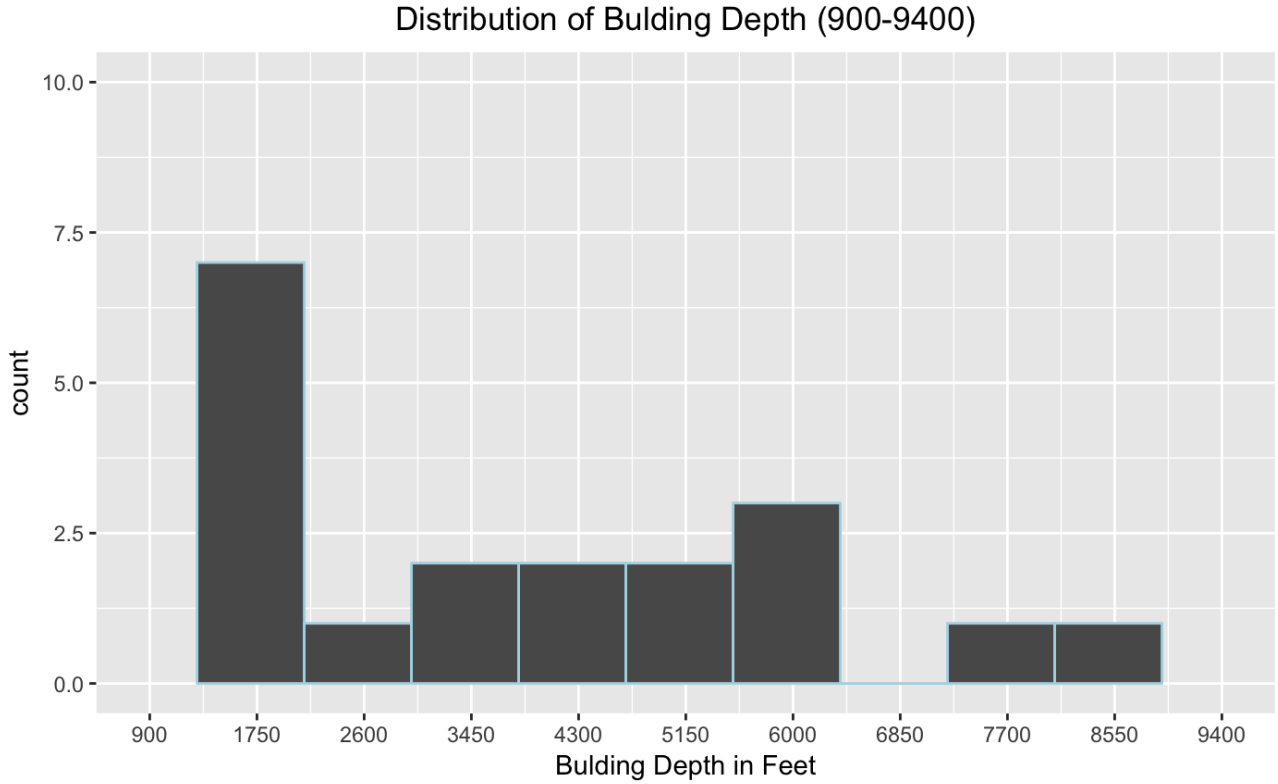
Variable 22	Description
BLDDEPTH	Building Depth in feet (length 7 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*


Distribution of Bulding Depth
count
6e+05
4e+05
2e+05
0e+00
0
2500
5000
7500
Bulding Depth in Feet

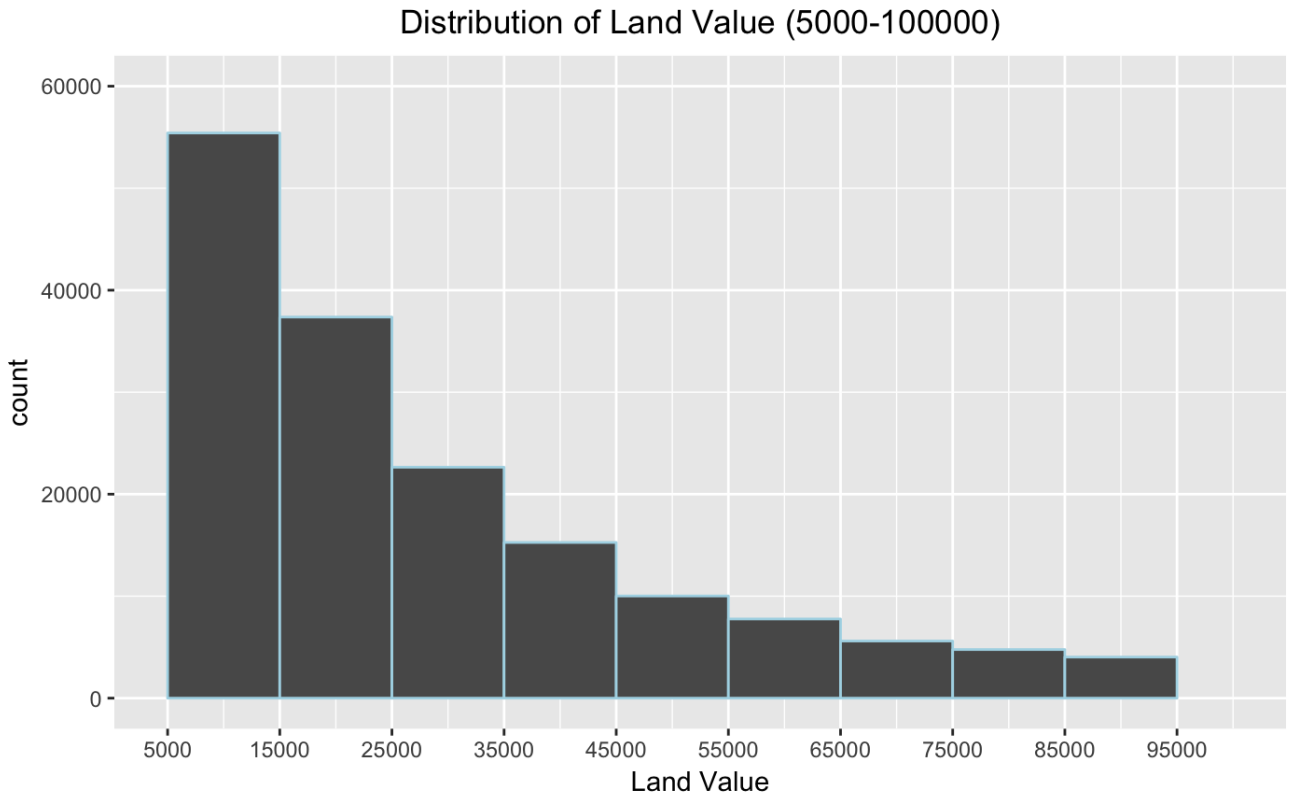

Distribution of Bulding Depth (0-100)
count
250000
200000
150000
100000
50000
0
0
10
20
30
40
50
60
70
80
90
100
Bulding Depth in Feet

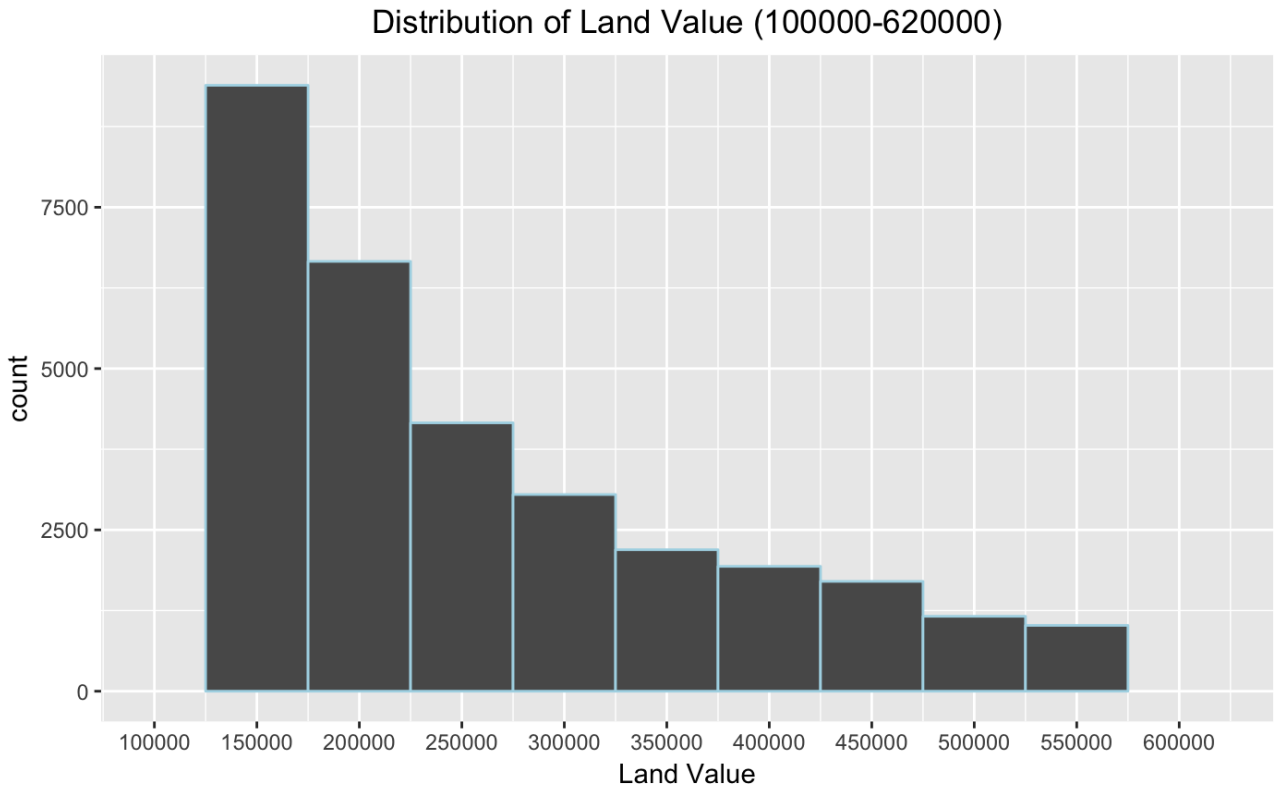




Variable 23	Description
AVLAND2	Second market value of the land (length 11 numeric)

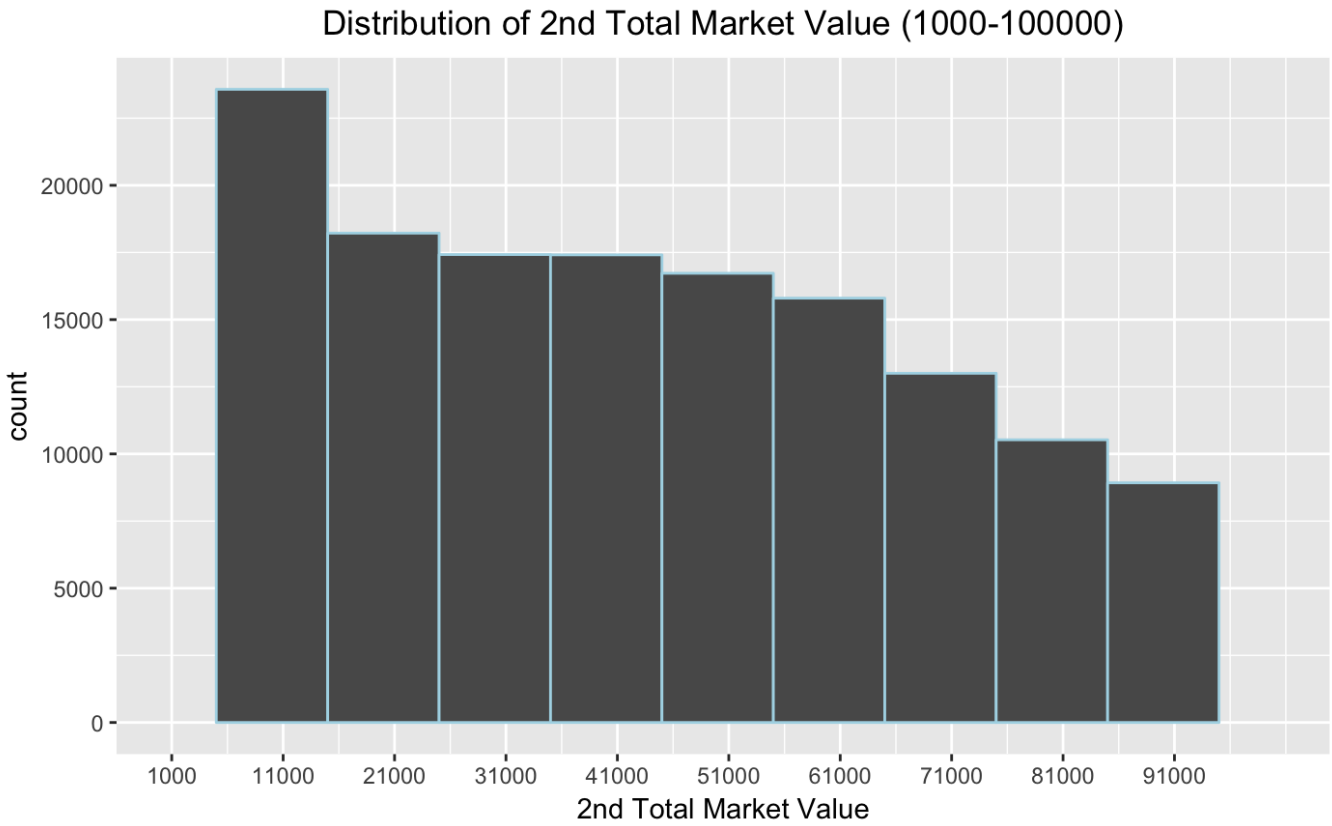
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*

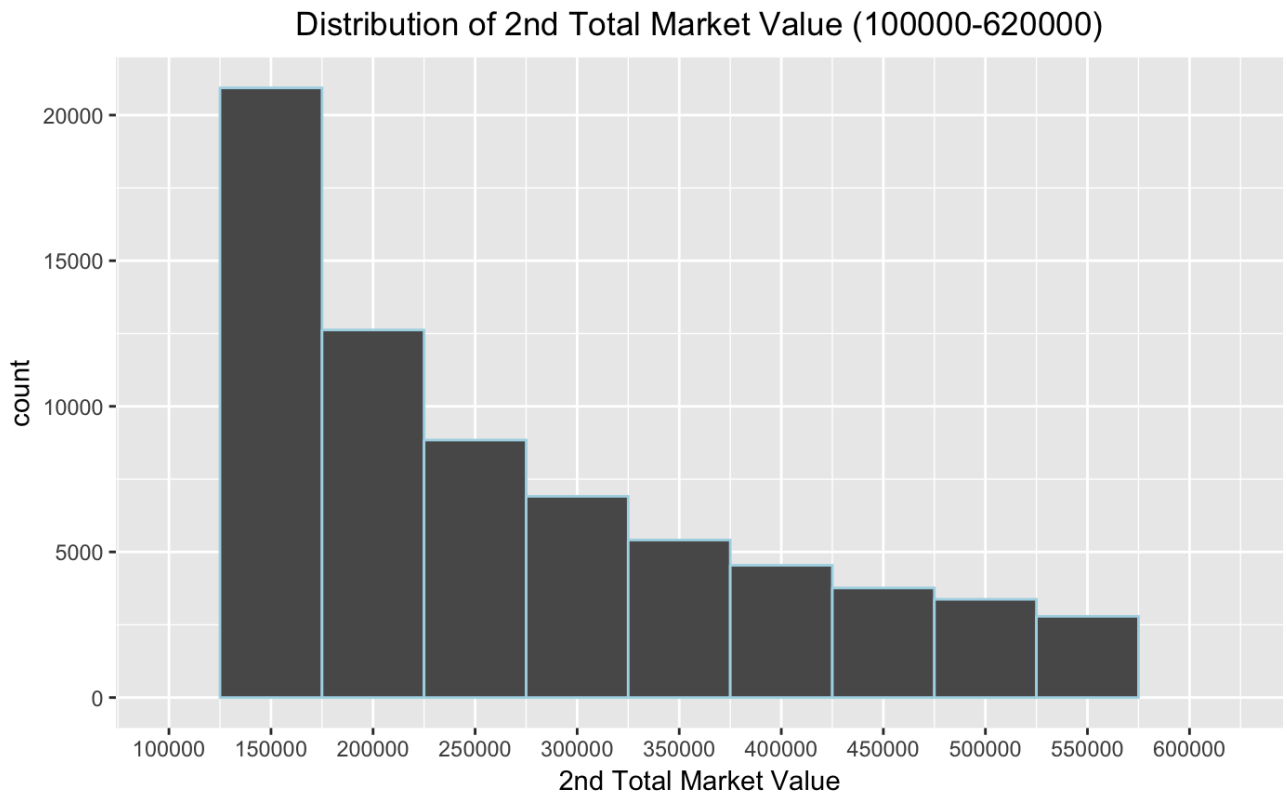




Variable 24	Description
AVTOT2	Second total market value (length 11 numeric)

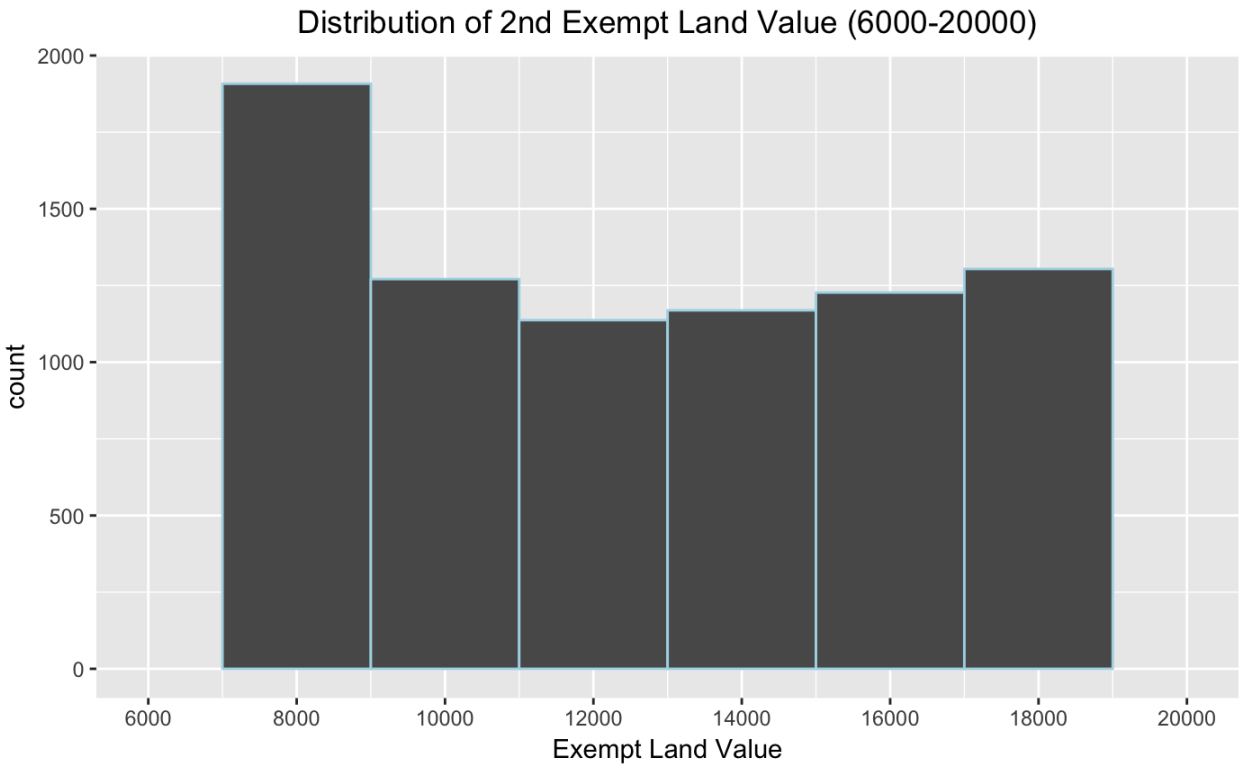
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*





Variable 25	Description
EXLAND2	Second exempt land value (length 11 numeric)

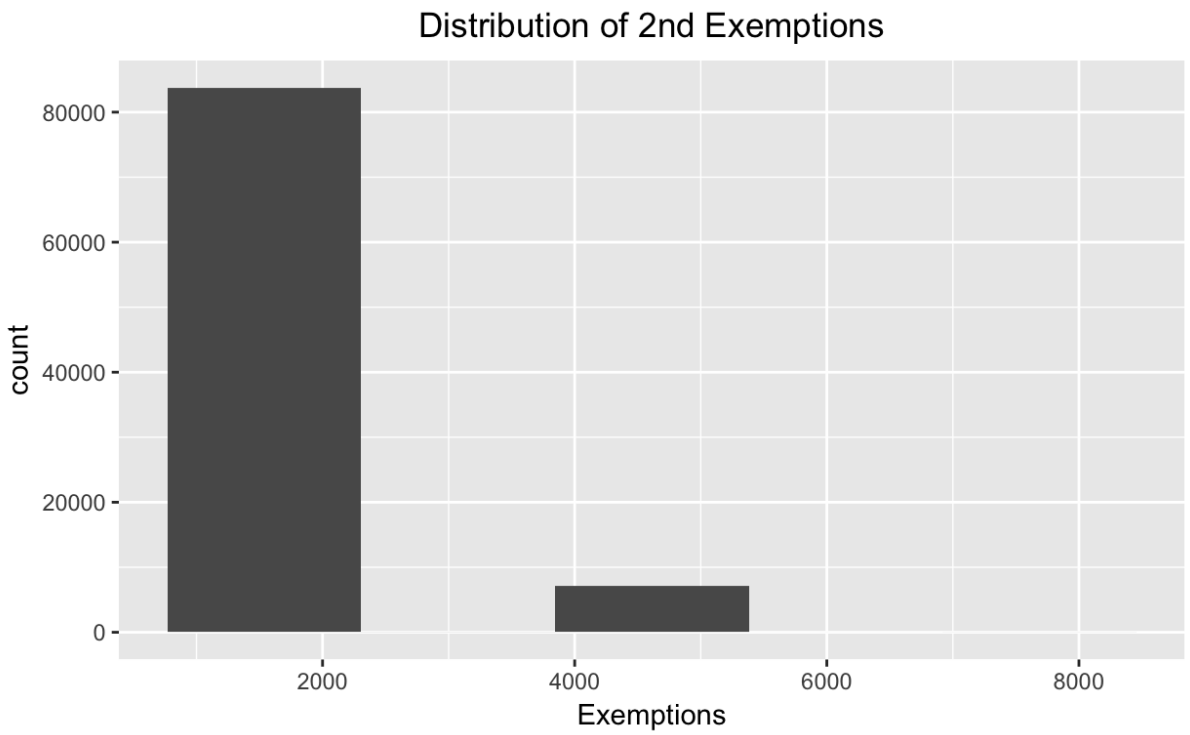
Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*



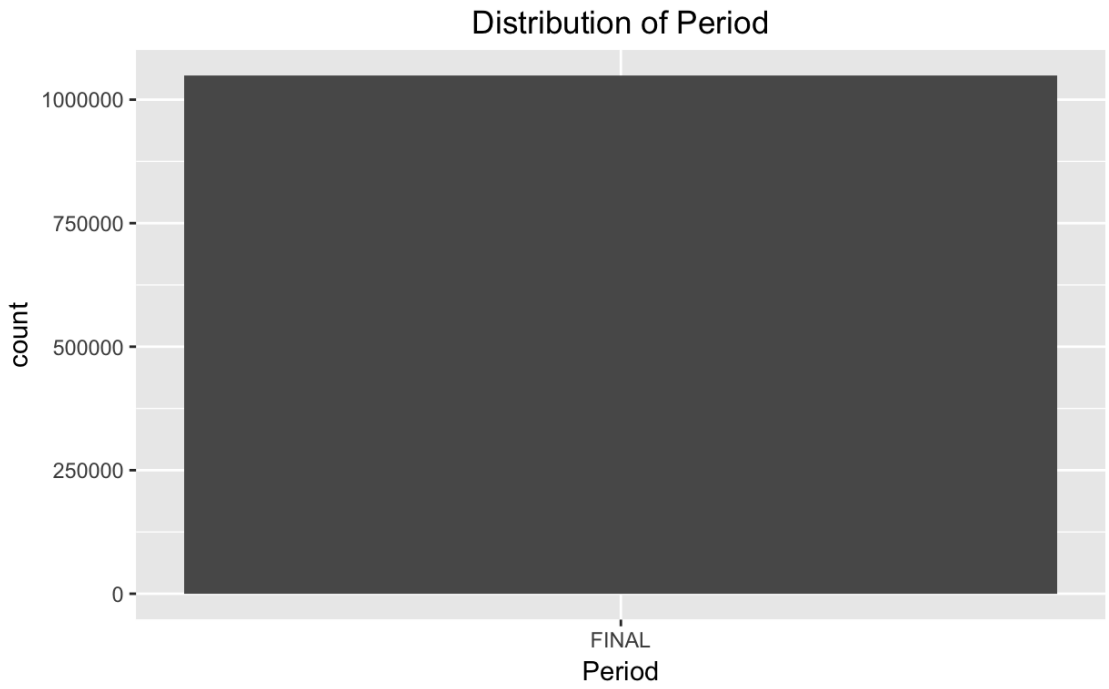
Variable 26	Description
EXTOT2	Second exempt total value (length 11 numeric)

Note: This variable exhibits severe right-skew distribution. Sub-intervals are selected to specify partial distribution.*

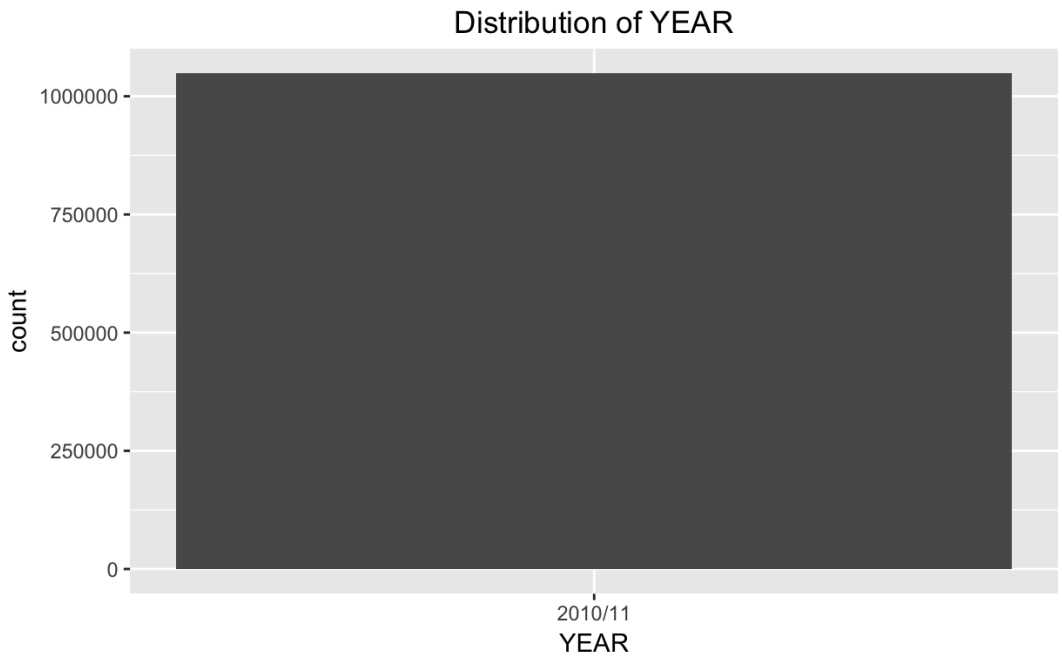
Variable 27	Description
EXCD2	Second number of exemptions on the property



Variable 28	Description
PERIOD	Indicator for the Change Period of the File



Variable 29	Description
YEAR	Year of the file.



Variable 30	Description
VALTYPE	Valid type of the property