Rolling Sales Data Analysis - Manhattan

Shrey Patel
Data 607

Contents

- Introduction
  - Data set
  - Task
- Getting data ready for analysis
  - Handling missing data
  - Type conversion
- Exploratory analysis
- Predictive analysis

Introduction

Manhattan data set

- The data set contains rolling sales data for properties that were sold over the 12-month period between Dec 2020 and Nov 2021.
- 21 attributes: borough, neighborhood, building class category, tax class at present, tax class at sale, building class at present, building class at sale, block, lot, easement, address, apartment number, zipode, residential units, commercial units, total units, land square feet, gross square feet, year built, sale price, and sale date.
- 20853 observations

Task

Using exploratory analysis, discovering and understanding a) the sale prices for each tax class b) the number of sales for each tax class, and c) the distribution of sale prices over time for each tax class. Finally, using multinomial logisitc regression, predict the land square feet area for the buildings sold.

Getting data ready for analysis

First Impressions

- Borough has the same value for all observations

- Easement is an empty column

- Almost 3200 buildings were sold at a price $250 or lower
  - 97% of them being sold at a price less than $10

- Around 88 buildings were sold with 0 residential and 0 commercial units.
  - Each of these buildings belong to tax class #4, includes warehouses, factories, offices, etc.
  - About 75% of those observations were either sold for $0, $1, or $2 – NMAR?

- Land and gross square feet information is missing for 93% of the observations
  - Information regarding the number of residential and commercial units is missing for about half of those observations
  - These properties belong tax class #2

- There are observations where gross square feet is less than the land square feet. Most of those observations belong to tax class #4

Getting data ready for analysis

Columns Removed

- Borough, Easement, Address, Apartment number, Gross square feet columns were removed.

Type Conversion

- Land square feet and sale price variables were converted to numeric.
- The rest of the variables like neighborhood, building class category, tax class, block, lot, and so on were converted into factors.
- The sales date column was converted to a date type column

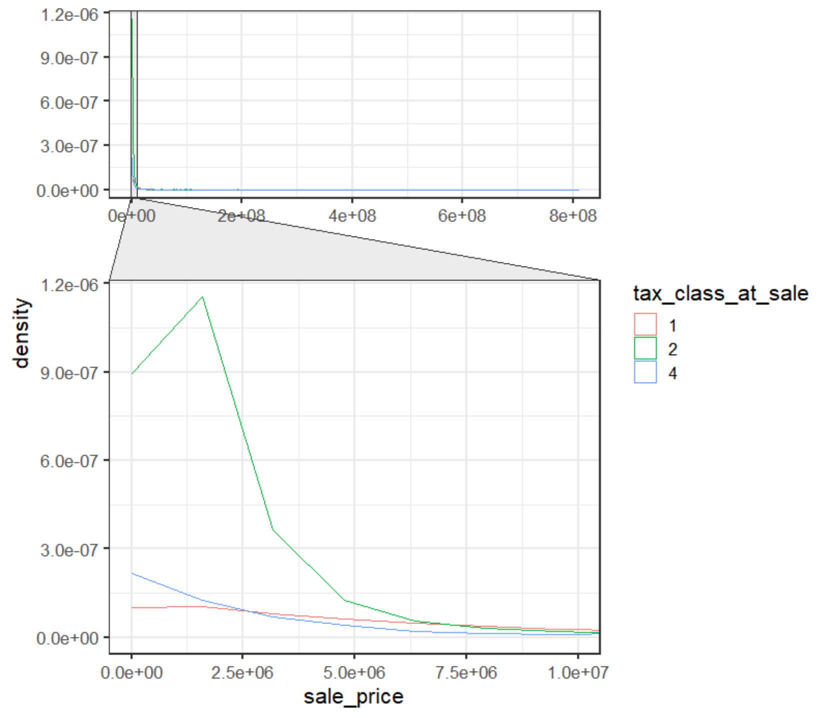
Handling Missing Data

- All observations with sale price $1000 or lower were removed.
- All observations with missing data with respect to land square feet area and year built were removed.
- Duplicate rows were removed, if any.

The resulting data set consists of 947 observations.

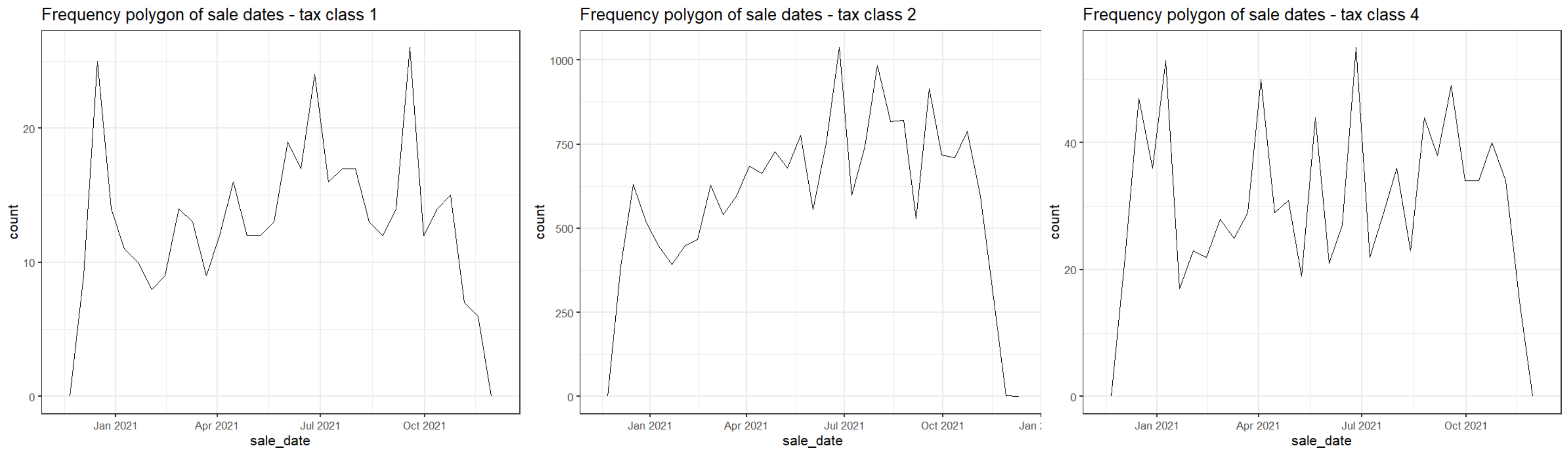
Exploratory Analysis

The distribution of sale prices for each tax class:



Exploratory Analysis

The distribution of sales for each tax class:



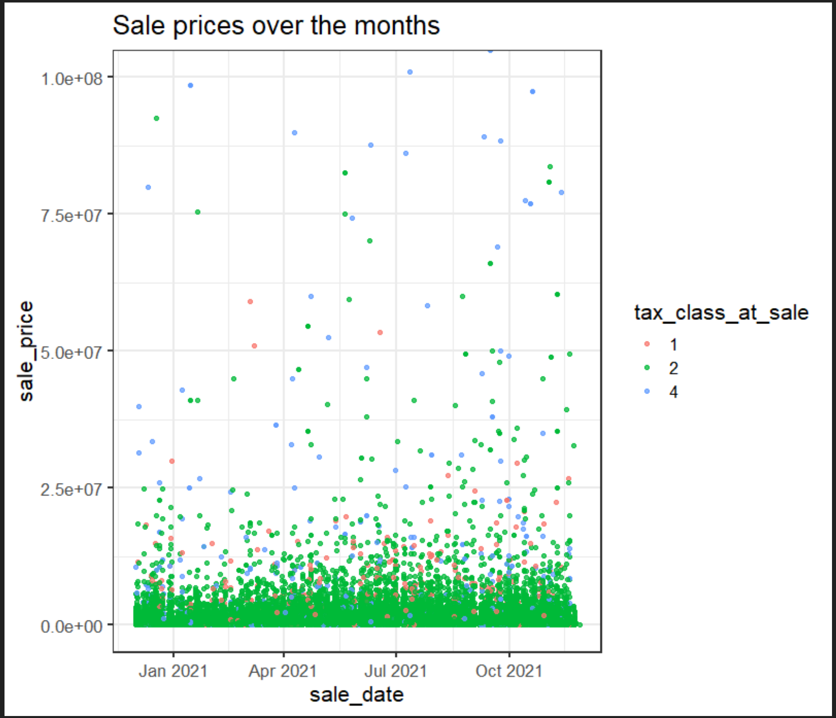
Frequency polygon of sale dates - tax class 1

Frequency polygon of sale dates - tax class 2

Frequency polygon of sale dates - tax class 4

Exploratory Analysis

The distribution of sale prices over time:



Predictive Analysis

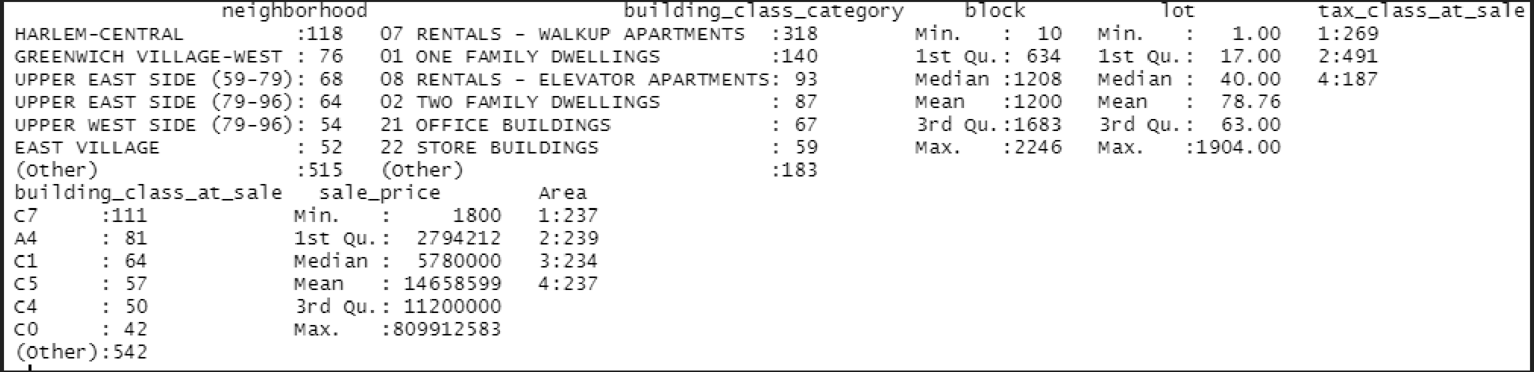
Target Variable: Area

Condition	Value	Number of Observations
Land Square Feet <= 1754	1	237
Land Square Feet <= 2300	2	239
Land Square Feet <= 4112	3	234
Land Square Feet <= 659375	4	237

Features: Neighborhood, Building Class Category, Block, Lot, Tax Class at sale, Building Class at sale, Sale Price

Predictive Analysis

Summary of the data set

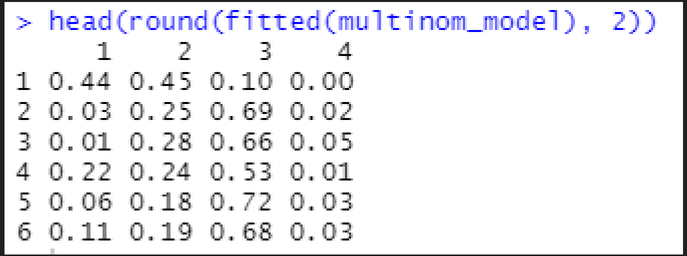


```
                 neighborhood                       building_class_category     block            lot            tax_class_at_sale
 HARLEM-CENTRAL         :118   07 RENTALS - WALKUP APARTMENTS  :318        Min.   :  10   Min.    :   1.00   1:269
 GREENWICH VILLAGE-WEST : 76   01 ONE FAMILY DWELLINGS         :140        1st Qu.: 634   1st Qu.:  17.00   2:491
 UPPER EAST SIDE (59-79): 68   08 RENTALS - ELEVATOR APARTMENTS: 93        Median :1208   Median :  40.00   4:187
 UPPER EAST SIDE (79-96): 64   02 TWO FAMILY DWELLINGS         : 87        Mean   :1200   Mean   :  78.76
 UPPER WEST SIDE (79-96): 54   21 OFFICE BUILDINGS             : 67        3rd Qu.:1683   3rd Qu.:  63.00
 EAST VILLAGE           : 52   22 STORE BUILDINGS              : 59        Max.   :2246   Max.   :1904.00
 (Other)                :515   (Other)                         :183
 building_class_at_sale   sale_price         Area
 C7     :111            Min.   :     1800   1:237
 A4     : 81            1st Qu.:  2794212   2:239
 C1     : 64            Median :  5780000   3:234
 C5     : 57            Mean   : 14658599   4:237
 C4     : 50            3rd Qu.: 11200000
 C0     : 42            Max.   :809912583
 (Other):542
```

Modeling: Multinomial logistic regression model with train/test split: 70/30 - Residual Deviance of 815.033 & AIC of 1457.033

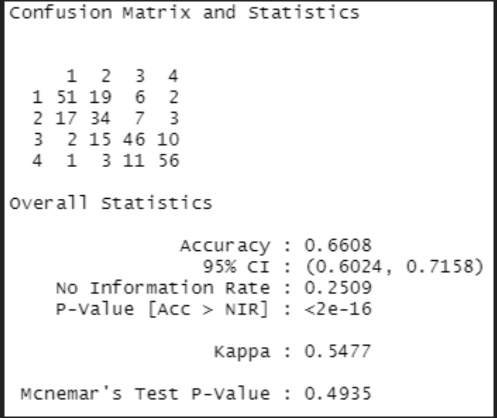
Predictive Analysis

- For each observation, the model chooses class with the highest probability calculated using MLE:



```
> head(round(fitted(multinom_model), 2))
     1    2    3    4
1 0.44 0.45 0.10 0.00
2 0.03 0.25 0.69 0.02
3 0.01 0.28 0.66 0.05
4 0.22 0.24 0.53 0.01
5 0.06 0.18 0.72 0.03
6 0.11 0.19 0.68 0.03
```

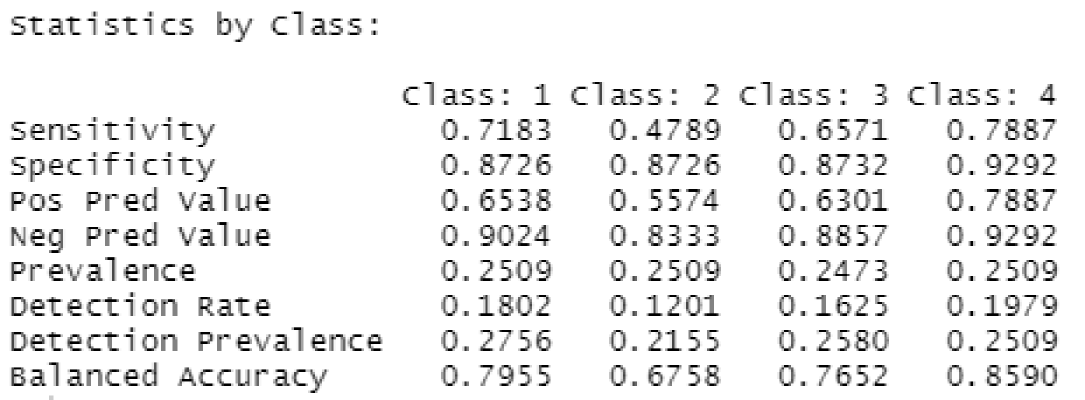
- Confusion Matrix and Statistics



```
Confusion Matrix and Statistics


     1  2  3  4
  1 51 19  6  2
  2 17 34  7  3
  3  2 15 46 10
  4  1  3 11 56

Overall Statistics

               Accuracy : 0.6608
                 95% CI : (0.6024, 0.7158)
    No Information Rate : 0.2509
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5477

 Mcnemar's Test P-Value : 0.4935
```

Predictive Analysis

- Class Statistics:



```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.7183   0.4789   0.6571   0.7887
Specificity            0.8726   0.8726   0.8732   0.9292
Pos Pred Value         0.6538   0.5574   0.6301   0.7887
Neg Pred Value         0.9024   0.8333   0.8857   0.9292
Prevalence             0.2509   0.2509   0.2473   0.2509
Detection Rate         0.1802   0.1201   0.1625   0.1979
Detection Prevalence   0.2756   0.2155   0.2580   0.2509
Balanced Accuracy      0.7955   0.6758   0.7652   0.8590
```

Thank you!