Old Mutual Insure

Data Science Assessment

Lundi Mlanduli

Email:lvmlanduli@gmail.com

Cell: 0609149159

Year: 2023

Question 1

The dataset contains 9 198 records. Here is a view of the top 5 results:

	ID	Title	Description	Location	Bedrooms	Latitude	Longitude	ListingDate	Class
0	2	Take a Break; Unplug at the BNB on Milagro Farms	You'll love the Back Door de Milagro, a Bed &a...	Forreston	1	32.233543	-96.862790	October 2016	1
1	3	Your Home Away From Home - Suite B	This private room is located in South Arlingto...	Arlington	1	32.623912	-999999.000000	October 2016	1
2	4	Quiet peaceful location	Quiet neighborhood very peaceful at night with...	Frisco	1	33.179972	-999999.000000	March 2016	1
3	5	Home Sweet Home	My place is close to parks, the city center, a...	Boerne	2	29.779891	-98.694755	October 2016	2
4	7	Hideaway Guest Suite-Travis Heights	Welcome to The Hideaway Guest Suite, a luxury ...	Austin	1	30.244405	-999999.000000	September 2012	2

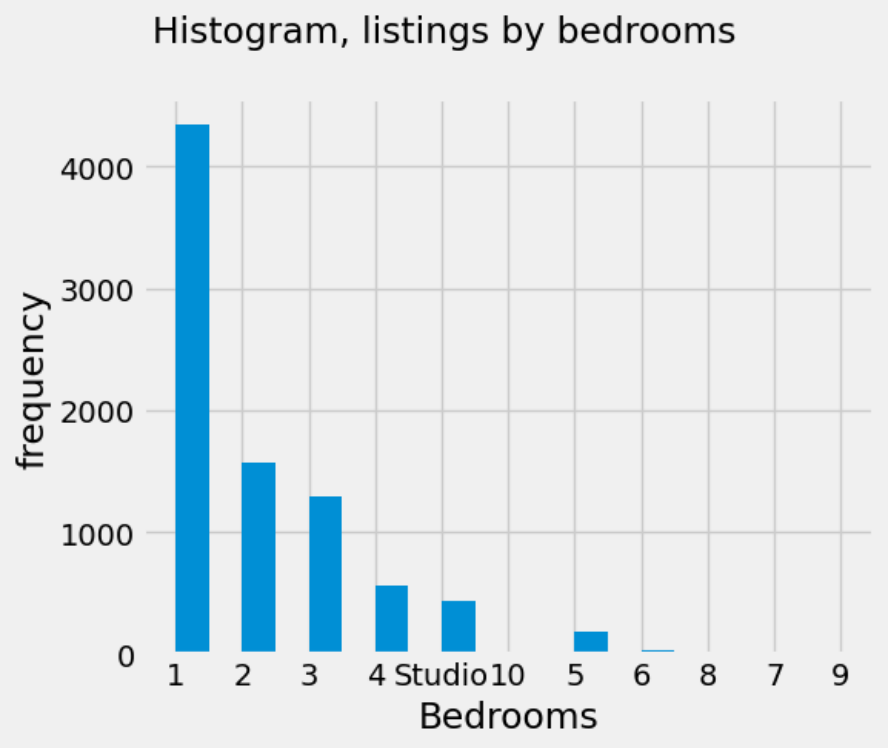
The latitude and longitude columns are numeric; the class column is an integer representing the three levels of classes. The bedrooms column is an object because the some bedrooms are not numbers but 'Studio'.

```
RangeIndex: 9198 entries, 0 to 9197
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   ID           9198 non-null   int64  
 1   Title        9196 non-null   object 
 2   Description  9181 non-null   object 
 3   Location     9198 non-null   object 
 4   Bedrooms     8468 non-null   object 
 5   Latitude     9194 non-null   float64
 6   Longitude    9195 non-null   float64
 7   ListingDate  9198 non-null   object 
 8   Class        9198 non-null   int64  
```

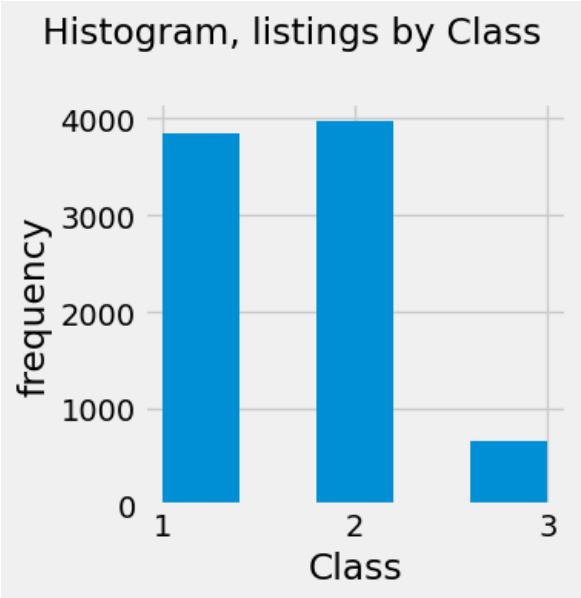
The fields 'Bedrooms', 'Latitude' and 'Longitude' have some missing values. These may be removed as they do not account for a large proportion of the data.

```
ID              0
Title           2
Description    17
Location        0
Bedrooms      730
Latitude        4
Longitude       3
ListingDate     0
Class           0
```

A histogram of the listings by the number of bedrooms shows that 1-bedroom listings are the most common.

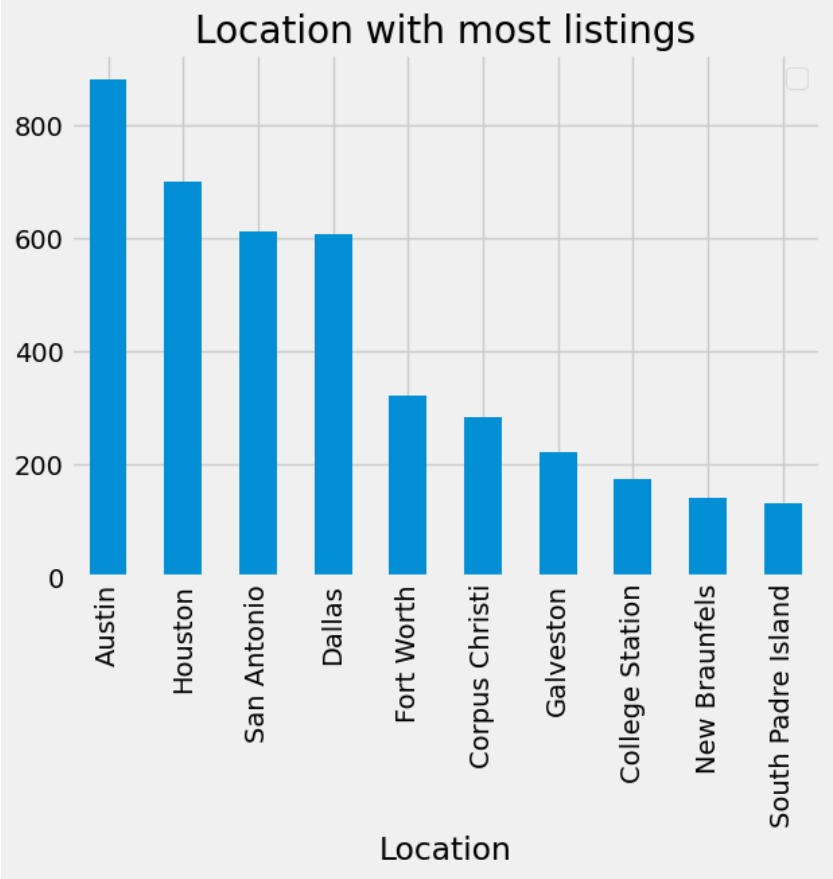


Listings with a class level 2 are most popular in Texas, followed by class 1 and then class 3.

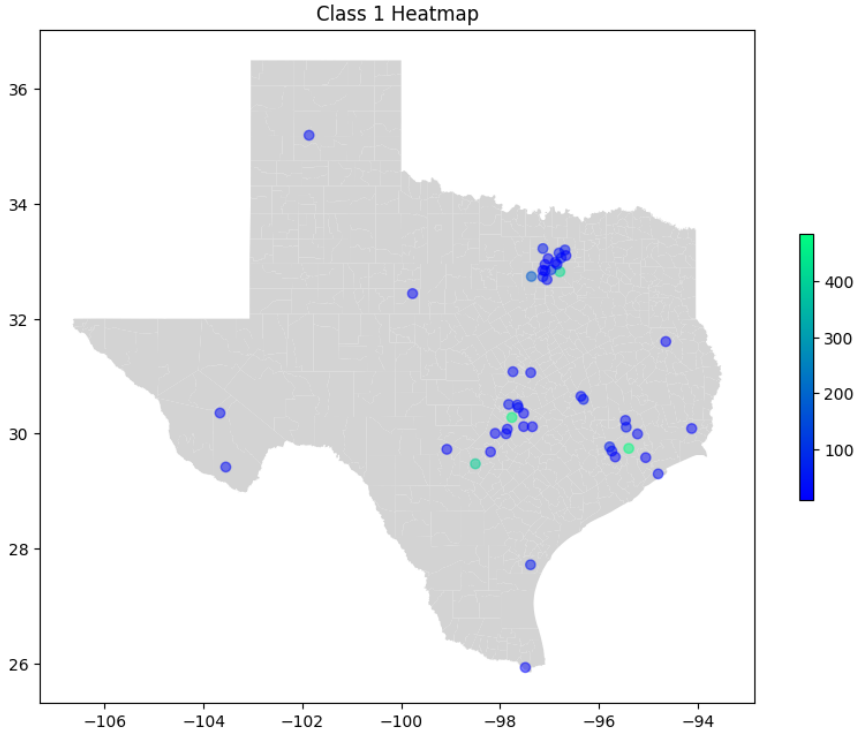


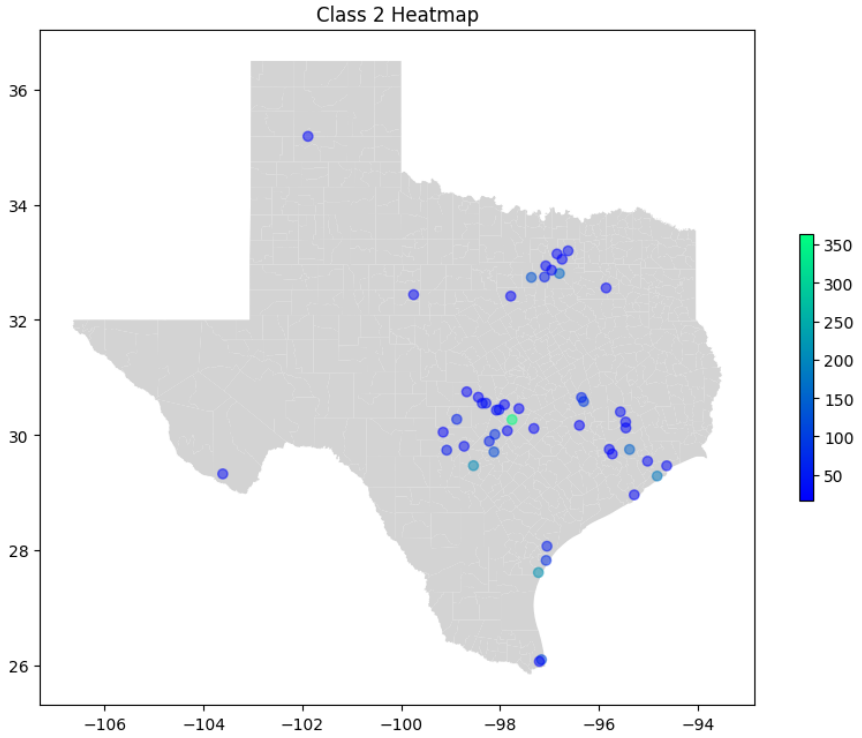
Over the years there has been an increase in the number of listings per year across all classes.

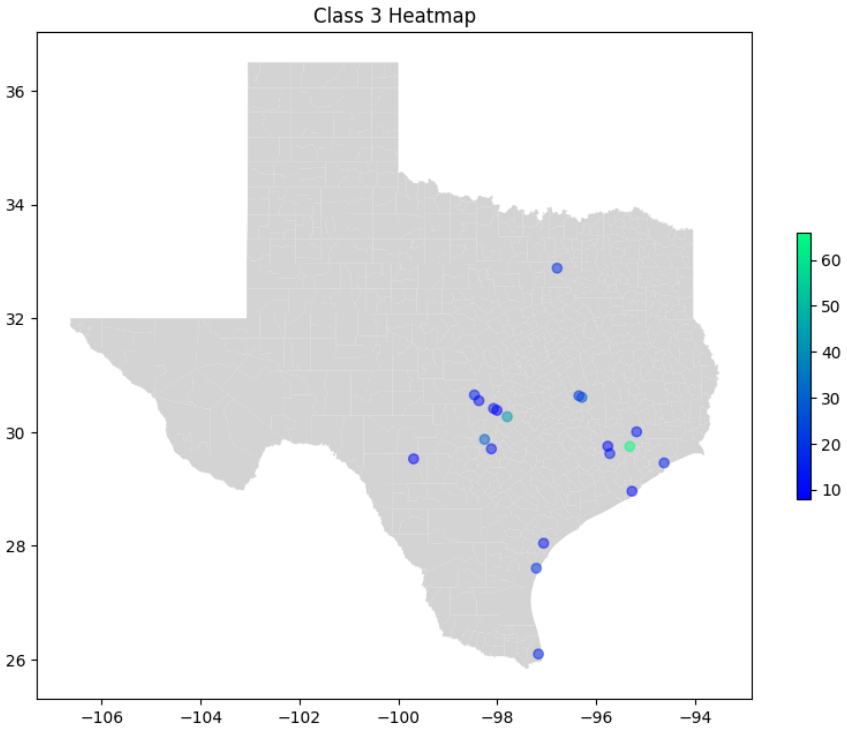
Here is a view of the top 10 cities with the most listings; the city of Austin has the most by far.



The following geoplots illustrate the geographical distributions of the different class listings across the state of Texas.







Question 2

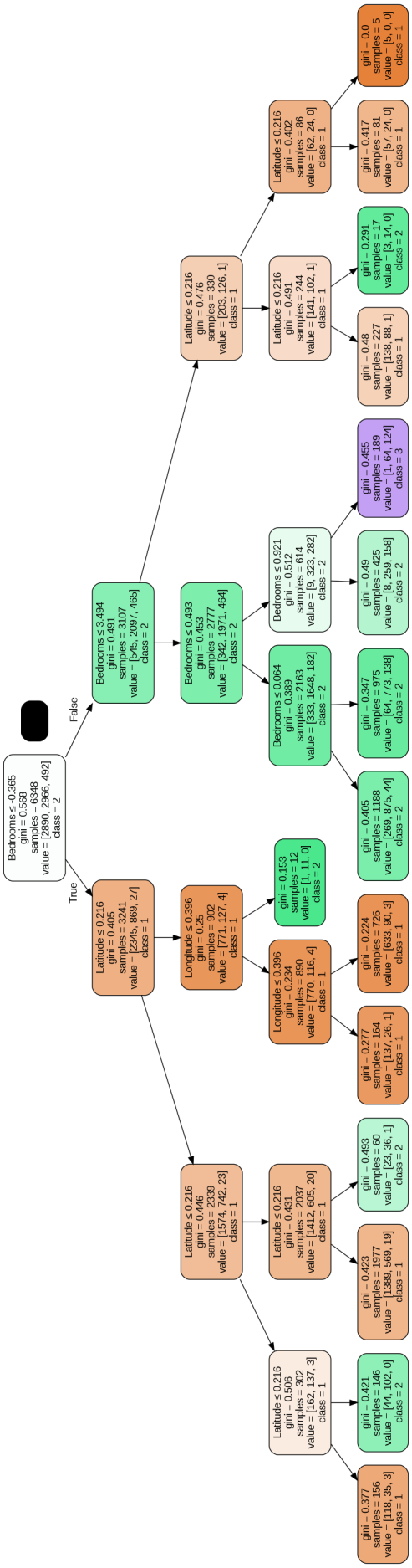
- A Decision Tree Classifier was fitted to the model:

$$\text{Confusion Matrix} = \begin{Bmatrix} 816 & 134 & 0 \\ 311 & 672 & 19 \\ 7 & 101 & 57 \end{Bmatrix}$$

accuracy = 0.7298

recall = 0.7298

precision = 0.732


Bedrooms ≤ 0.365
gini = 0.568
samples = 6348
value = [2890, 2966, 492]
class = 2
True
False
Latitude ≤ 0.216
gini = 0.405
samples = 3241
value = [2345, 869, 27]
class = 1
Bedrooms ≤ 3.494
gini = 0.491
samples = 3107
value = [545, 2097, 465]
class = 2
Latitude ≤ 0.216
gini = 0.446
samples = 2339
value = [1574, 742, 23]
class = 1
Longitude ≤ 0.396
gini = 0.25
samples = 902
value = [771, 127, 4]
class = 1
Bedrooms ≤ 0.493
gini = 0.453
samples = 2777
value = [342, 1971, 464]
class = 2
Latitude ≤ 0.216
gini = 0.476
samples = 330
value = [203, 126, 1]
class = 1
Latitude ≤ 0.216
gini = 0.506
samples = 302
value = [162, 137, 3]
class = 1
Latitude ≤ 0.216
gini = 0.431
samples = 2037
value = [1412, 605, 20]
class = 1
gini = 0.153
samples = 12
value = [1, 11, 0]
class = 2
Longitude ≤ 0.396
gini = 0.234
samples = 890
value = [770, 116, 4]
class = 1
Bedrooms ≤ 0.064
gini = 0.389
samples = 2163
value = [333, 1648, 182]
class = 2
Bedrooms ≤ 0.921
gini = 0.512
samples = 614
value = [9, 323, 282]
class = 2
Latitude ≤ 0.216
gini = 0.491
samples = 244
value = [141, 102, 1]
class = 1
Latitude ≤ 0.216
gini = 0.402
samples = 86
value = [62, 24, 0]
class = 1
gini = 0.377
samples = 156
value = [118, 35, 3]
class = 1
gini = 0.421
samples = 146
value = [44, 102, 0]
class = 2
gini = 0.423
samples = 1977
value = [1389, 569, 19]
class = 1
gini = 0.493
samples = 60
value = [23, 36, 1]
class = 2
gini = 0.277
samples = 164
value = [137, 26, 1]
class = 1
gini = 0.224
samples = 726
value = [633, 90, 3]
class = 1
gini = 0.405
samples = 1188
value = [269, 875, 44]
class = 2
gini = 0.347
samples = 975
value = [64, 773, 138]
class = 2
gini = 0.455
samples = 189
value = [1, 64, 124]
class = 3
gini = 0.49
samples = 425
value = [8, 259, 158]
class = 2
gini = 0.48
samples = 227
value = [138, 88, 1]
class = 1
gini = 0.291
samples = 17
value = [3, 14, 0]
class = 2
gini = 0.417
samples = 81
value = [57, 24, 0]
class = 1
gini = 0.0
samples = 5
value = [5, 0, 0]
class = 1

- An XGBoost Classifier was fitted to the model:

$$\text{Confusion Matrix} = \begin{Bmatrix} 772 & 172 & 6 \\ 189 & 777 & 36 \\ 7 & 83 & 75 \end{Bmatrix}$$

accuracy = 0.7671

recall = 0.7671

precision = 0.7642

- A Random Forest Classifier with max depth = 8, was fitted to the model:

$$\text{Confusion Matrix} = \begin{Bmatrix} 799 & 151 & 0 \\ 270 & 714 & 18 \\ 7 & 115 & 43 \end{Bmatrix}$$

accuracy = 0.735

recall = 0.735

precision = 0.733

- A Support Vector Machine was fitted to the model:

$$\text{Confusion Matrix} = \begin{Bmatrix} 838 & 112 & 0 \\ 356 & 643 & 3 \\ 11 & 138 & 16 \end{Bmatrix}$$

accuracy = 0.7071

recall = 0.7071

precision = 0.7185

The best model based on the highest scores on accuracy, precision and recall is XGBoost.

Accuracy shows how often a classification ML model is correct overall.

Precision shows how often an ML model is correct when predicting the target class.

Recall shows whether an ML model can find all objects of the target class.

Below are the first 5 class predictions followed by the probabilities of each class prediction.

	ID	Title	Description	Location	Bedrooms	Latitude	Longitude	ListingDate	Date	Year	Class_Predictions
0	1	Private Room in Dallas' most central location	This Entire Place, or room depending on what y...	Dallas	1	32.892219	-96.727704	April 2017	2017-04-01	2017	1
1	6	5 BDR LAKE FRONT DELIGHT	This lake front, 5 bdr, 4.5 bathroom home is a...	Richmond	4	29.633127	-95.754159	January 2017	2017-01-01	2017	3
2	18	Updated Sea Horse Inn, 4 bedroom, 3 baths	Sea Horse Inn is perfect for 2 families or 4 c...	Port Aransas	4	27.777523	-97.103504	May 2017	2017-05-01	2017	2
3	37	DKC Farm Bed & Breakfast	Relaxing country getaway equidistant to all ar...	Grand Saline	3	32.628218	-95.679786	March 2017	2017-03-01	2017	2
4	40	26506 Willow	My place is good for couples, business travele...	Katy	4	29.739205	-95.830856	March 2017	2017-03-01	2017	2

	ID	Class_1_Prob	Class_2_Prob	Class_3_Prob
0	1	0.988164	0.011237	0.000599
1	6	0.013723	0.185289	0.800988
2	18	0.003898	0.620197	0.375906
3	37	0.041576	0.720938	0.237486
4	40	0.010313	0.783191	0.206496

Question 3

The proportion of luxury class in College Station is 15.42% whereas the proportion of luxury class across the rest of Texas is 7.59%

	Class	ID
0	1	48
1	2	100
2	3	27

```
[133] print(27/(48+100+27))
      0.15428571428571428
```

	Class	ID
0	1	3792
1	2	3868
2	3	630

```
[134] print(630/(3792+3868+630))
      0.07599517490952955
```

A hypothesis test of the two proportions:

$H_0 : \mu_{Texas} \leq \mu_{CS}$

$H_1 : \mu_{Texas} > \mu_{CS}$

Shows that we cannot reject the null and conclude that College Station has a significantly higher luxury proportion at the 5% significance level.

```
[139] import scipy.stats as stats

[140] ttest,p_value = stats.ttest_ind(0.07599,0.15428)
      print("p value:%.8f" % p_value)
      print("since the hypothesis is one sided >> use p_value/2 >> p_value_one_sided:%.4f" %(p_value/2))
      if p_value/2 <0.05:
          print("Reject null hypothesis")
      else:
          print("Fail to reject null hypothesis")

      p value:nan
      since the hypothesis is one sided >> use p_value/2 >> p_value_one_sided:nan
      Fail to reject null hypothesis
```