By: Timothy Dance

Miami Housing Statics

1) Select a dataset (or datasets) from an online repository or any other relevant source

a) Briefly describe why this dataset interested you

I am someone who really sees the value in owning property. its because of this that I tend to keep a mental note about the housing market in the areas I have lived in or visited. This is also the reason that I find this dataset so interesting since I think that the Miami housing market has a lot of things that can change the value of a property. For instance, the distance to the ocean or how centralized the location of the property plays a big part in price but will these values change with the growing concern of global warming? These are things I would like to look at.

b) Describe the metadata: column names and column data types. Also, if possible, units of measure of the column and possible values that the column can take.

PARCELNO: The property identifier

SALE_PRC: The sales price ($) of the property

LND_SQFOOT: land area (square feet)

TOT_LVG_AREA: floor area (square feet)

SPEC_FEAT_VAL: These are the value of special features swimming pool, gym etc (e.g., swimming pools) ($)

RAIL_DIST: distance to the nearest rail line (an indicator of noise) (feet)

OCEAN_DIST: distance to the ocean (feet)

WATER_DIST: distance to the nearest body of water (feet)

CNTR_DIST: distance to the Miami central business district (feet)

SUBCNTR_DI: distance to the nearest subcenter (feet)

HWY_DIST: distance to the nearest highway (an indicator of noise) (feet)

age: age of the structure

avno60plus: dummy variable for airplane noise exceeding an acceptable level

structure_quality: quality of the structure

month_sold: sale month in 2016 (1 = jan)

LATITUDE

LONGITUDE

2) Formulate at least two Business/ Societal / Personal problems that you want to solve using data.

Examine the relation which variables had on the sales price of a house in Miami.

Look at how the total size of the property changed based on the distance to the ocean.

2a) Success metrics

The information that the answer to these questions will give a business a estimated price for property in Miami. They will be able to do this since, the models that these reports provide will show the market price for a property based on provided factors. This will allow a company to stay competitive in their property prices.

3) Formulate at least three data analytics questions relevant to solve the business/ societal / personal problems defined in 2)

What were the three X factors with the strongest correlation to Y SALE_PRC.

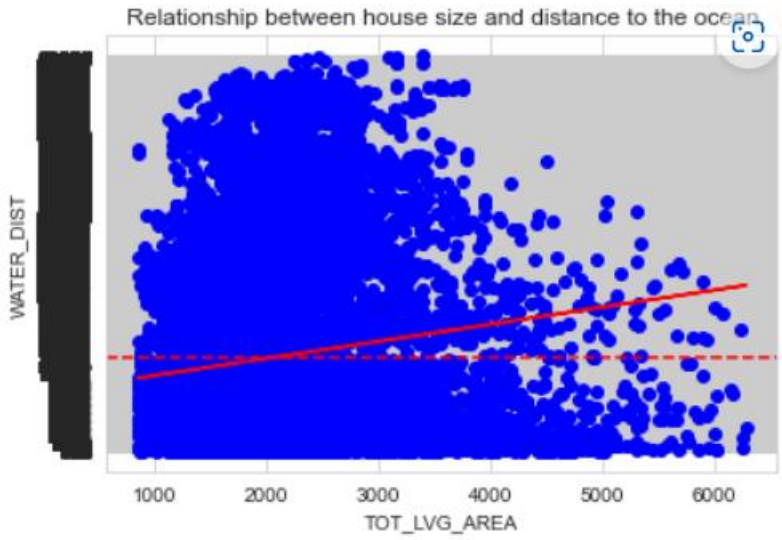
What was the correlation between TOT_LVG_AREA and WATER_DIST

What is the accuracy for predicting if a SALE_PRC of a house would be more than 500000 based on the given factors?

4) Data Analysis step

This can be found in the provided Jupiter notebook. I will be repeating some of the most important aspects of the notebook here but for all the Data Analysis steps, code and notes look in the Jupiter notebook.

To look at the correlation between the two factors TOT_LVG_AREA and WATER_DIST I created a linear regression chart to show the correlation.



This graph shows us that the closer the house is to the water the bigger the property is going to be.

Next, I wanted to look at which of the factors held the most importance for determining price.

	AVG_Importance	STD_Importance
LATITUDE	0.003325	NaN
LONGITUDE	0.003923	NaN
LND_SQFOOT	0.011579	NaN
TOT_LVG_AREA	0.055431	NaN
SPEC_FEAT_VAL	0.005622	NaN
RAIL_DIST	0.000120	NaN
OCEAN_DIST	0.023397	NaN
WATER_DIST	0.002368	NaN
CNTR_DIST	0.010287	NaN
SUBCNTR_DI	0.011148	NaN
HWY_DIST	0.001292	NaN
age	0.000598	NaN
avno60plus	0.000024	NaN
month_sold	-0.000263	NaN
structure_quality	0.008732	NaN

This shows use that the 3 x factors with strongest correlation were

1. TOT_LVG_AREA with a correlation of 0.055431

2. OCEAN_DIST with a correlation of 0.023397

3. LND_SQFOOT with a correlation of 0.011579

This tells us that when it comes to the price of property the floor area of the property is the most important factor for how expensive a property will be.

I then created two prediction models using GaussianNB and logistic regression to see which one gave me a higher accuracy for predicting if the price of a property would be more or less then 500,000$.

GaussianNB classification report

	precision	recall	f1-score	support
0	0.97	0.94	0.95	3400
1	0.76	0.86	0.80	780
accuracy			0.92	4180
macro avg	0.86	0.90	0.88	4180
weighted avg	0.93	0.92	0.92	4180

logistic regression classification report

	precision	recall	f1-score	support
0	0.95	0.98	0.97	3400
1	0.89	0.79	0.84	780
accuracy			0.94	4180
macro avg	0.92	0.89	0.90	4180
weighted avg	0.94	0.94	0.94	4180

After looking at I found that the model using logistic regression gave me a higher accuracy when compared to the GaussianNB model. Using logistic Regression, we got an accuracy of 0.94 with an f1-sore of 0.97 for 0 and 0.84 for 1. This model gave us higher accuracy with a .02 increase from the GaussianNB model it also had higher f1-score for both 0 and 1. Overall this model is the one we will be using to predict our 0 and 1.

5) At least two recommendations based on your analysis

How do the insights you derived help solve the business/societal/individual problem you defined in 2)?

When deciding on what property one should purchase, they should consider the distance that property is to a body of water. The reason for this is due to the fact that a property closer to the ocean is a property that is more likely to retain its value. With that being said one must also understand that if they are looking for a property of a large size it is most likely going to be close to a body of water. If a person is looking for something cheaper and smaller, then they should be looking inland since it is more likely that this property will be cheaper and smaller then those

closer to the bodies of water. Overall, this information can help buyers and sellers get a better understanding of how much they should be expecting to pay for or sell a property based on its distance to the ocean.

If a business or person was to evaluate how much a property is worth the 3 most important factors that should be looked at are the following. The most important factor in determining the value of a property is the floor area of the building of the property. Next to this factor the second most important thing to look at will be the distance from the property to the ocean. The third most important factor is the land area of the property. These are the 3 most important factors for determining the price of a property. When a company or person is going over pricing a property it is important to remember that based on past sales buyers are willing to pay more for properties that are closer to the ocean. that have a lot of land and a large building size .