***Graphical user interface

Description automatically generated with medium confidence***

***DIPLOMA IN DATA ANALYTICS CO-OP***

**Data Handling and Decision Making (10012022-DIP-321-DHDM-G1)**

**Final Group Project**

**Project Name: Manhattan Housing Market**

|  |  |  |
| --- | --- | --- |
| Student Name | ID | Email |
| Daniel de Jesús Martínez de León | O1023964 | dan3200dm@gmail.com |
| Cesar Romeo Hernandez Luis | O1017528 | cesar.rh.luis@gmail.com |
| Eduardo Blanco Fernandez | O1026654 | edublanco2011@hotmail.com |
| Francisco Ignacio Muñoz Carmona | O1023168 | franciscomc68@gmail.com |
| Jehison Ernesto Castaneda Ibanez | O1030306 | jehison8012@gmail.com |
| María Alejandra Bedoya Mazo | O1025305 | alejandrabedoya00@gmail.com |

# Executive Summary

In the last years the housing market on the Manhattan Island has been really volatile and hard to predict due to different factors, from the COVID-19 pandemic to the variation of the interest rate from the federal reserve, among other factor, this housing market is constantly breaking records of sale volumes. This situation has created complexity and new opportunities to the real state agents of New York that operate on the Manhattan Island.

The objective of this research is to analyse, give a set of proper insights and create a predictive model of the housing market in the Manhattan area using the R language, this allowed us to clean, manipulate and summarize a big data set with detailed information of the market. Using R also allow us to maintain a clean and optimised data base trough time.

We propose to create the base of a predictive model that will improve over time which the real state agents of New York could use to predict the prices of the housing market of Manhattan. This model can become an important tool to find real state opportunities and to give a starting point for the agents to start their negotiations and speed up the research to attend their clients.

It is important for real state agents to understand better the market they are working on and to have the statistics and a predictive model that will improve over time, this gives them an important advantage on this competitive market. With the tools that we propose the real state agents can add a new dimension to their working flow with data analytics and data science.

**Contents**

[Executive Summary 6](#_Toc101211228)

[1. Introduction 9](#_Toc101211229)

[2. Case description and business objectives 10](#_Toc101211230)

[3. Assumptions 12](#_Toc101211231)

[4. Data Analysis Process 12](#_Toc101211232)

[5. Importing Data 14](#_Toc101211233)

[6. Data Sanity Checks 15](#_Toc101211234)

[6.1. Variable Analysis 15](#_Toc101211235)

[6.2. Treating Duplicate Entries 16](#_Toc101211236)

[6.3. Treating Missing Values 17](#_Toc101211237)

[7. Data Summarization, Visualization, Statistical Analysis, and Calculations 17](#_Toc101211238)

[8. Predictive Analysis and Visualization 21](#_Toc101211239)

[9. Summary 23](#_Toc101211240)

[10. Appendix 25](#_Toc101211241)

[10.1. R code 25](#_Toc101211242)

[10.2. Personal reflection (Alejandra) 29](#_Toc101211243)

[10.3. Personal Reflection (Daniel) 30](#_Toc101211244)

[10.4. Personal Reflection (Francisco) 30](#_Toc101211245)

[10.5. Personal Reflection (Eduardo) 31](#_Toc101211246)

[10.6. Personal Reflection (Cesar) 31](#_Toc101211247)

[10.7. 31](#_Toc101211248)

[Figure 1: Data analysis process map 14](#_Toc101211221)

[Figure 2: Overall sales in 2020 18](#_Toc101211222)

[Figure 3: Overall sales by building categories 19](#_Toc101211223)

[Figure 4: Buildings older than 20 years old 20](#_Toc101211224)

[Figure 5: Correlation between tax class and price 20](#_Toc101211225)

[Figure 6: Trained predictive model (manhattansup) 21](#_Toc101211226)

[Figure 7: Predicted prices 22](#_Toc101211227)

# Introduction

The objective of this research is to analyse and give a set of proper insights of the housing market in the Manhattan Island on New York City of the state New York on the United States of America. The data set which was used for the purpose of this research included information about the neighborhoods, tax classes, building classes, residential units, commercial units, land, gross land, year of building, date of sell, and price of sell on USD.

With this data set we were able to provide the following information:

* More common building class for residential units.
* Categorize the number of buildings according to its price.
* Number of buildings that were build before the year of 2002.
* Neighborhoods with the more residential buildings on it.
* Relationship between the tax class and the price.
* Months that are more common to sell buildings.

In addition to the previous summary of the housing market we try to create a multiple linear regression model that will help us to predict an estimate of the housing prices in the Manhattan area that could be used to create a tool that will help the real state agents to estimate the initial price for buildings as part their negotiations.

For the data handling process we imported the data set from a excel worksheet to a data frame in R, after that to avoid redundancy, we deleted the columns that were not needed for this research and renamed the variables so they could be easily manipulated. Then we assigned the right class to each variable which resulted in four categorical variables, five numerical variables, one character string and one date variable.

For missing values and observations that could introduce a bias to the analysis we created four key arguments that will help us get rid of them, removing buildings that were presold but were not build yet, removing building classes that included hospitals, asylums, warehouses, factories, hotels, studios, offices, convents, parks, schools, among others. We also only considered observations that had at least one residential unit and that had gross land to ensure they were proper housing buildings.

Finally, to test the predictive multiple linear regression model we divided the data set into two different ones to exclude the outsiders that were clearly input errors and would have introduced bias to the model. We later used these outliers to test the predictive model and fill this missing values.

# Case description and business objectives

For this case study, this has for objective the analysis of different data corresponding of The Department of Finance’s Rolling Sales files which lists properties that sold in the last twelve-month period in New York City for tax class 1, 2, and 4. These files include:

* the neighborhood.
* building type.
* square footage.
* other data.

It is important to also note the tax class for each of these properties, in which Property in NYC is divided into 4 classes:

* Class 1: Most residential property of up to three units (family homes and small stores or offices with one or two apartments attached), and most condominiums that are not more than three stories.
* Class 2: All other property that is not in Class 1 and is primarily residential (rentals, cooperatives, and condominiums). Class 2 includes:
  + Sub-Class 2A (4 - 6-unit rental building).
  + Sub-Class 2B (7 - 10-unit rental building).
  + Sub-Class 2C (2 - 10-unit cooperative or condominium); and
  + Class 2: 11 units or more).
* Class 3: Most utility property.
* Class 4: All commercial and industrial properties, such as office, retail, factory buildings and all other properties not included in tax classes 1, 2 or 3.

With this information clear, the analysis will be focused on the different variables that revolve around the properties that are being sold in the city of New York. This will be directed mainly towards individual people or real state organizations that focus on finding properties for residential purposes, so the findings that this case study may get will help to determine which factors and variables will affect the prices of this residential properties for future sales.

This new real state organization has projected an uprise in the amount of people looking for residential properties in different areas of New York City, so one of the main objectives for the organization is to have a better view of the current state of real state across the neighbourhoods of New York City, studying the class of tax, class of building, dates of sales and many other variables to better determine the prices of this properties and give fair deals to their costumers. That is why this organization is looking for this data analysis to set the foundations for a better understanding of all these variables that may affect prices of properties, to even get to the point of predicting said prices. So, the main business objectives are:

* Predict the price for different properties.
* Analyze how many properties are over 20 years old.
* See how much tax classes affect the price.
* Which is the most common building level.
* Relation of the area with the number of apartments.
* Which months has the most sales.
* Categorize the most common apartments, the normal ones, and the most expensive ones.

This way, the case study will address these objectives and analyze them through data and different approaches towards said dataset, applying different models, calculations, filtering, and cleaning to get the most accurate amount of data that will get the organizations the optimal results to give efficient responses to their customers needs.

# Assumptions

For this analysis, it was necessary to assume:

1) Inference had to be made on how the size of the properties influenced the price of the property and the interrelation between the cost of each space with respect to the type of use that each property had, we were able to visualize that each property was categorized depending on the use that was given to the property (family, commercial, industrial use) so we made the assumption that this classification had a direct impact on the sale price of the property.

2) The relationship between the availability of apartments by area and how this was related to the sale price had to be assumed, in detail the number of limited properties available (supply) in each area was directly related to the sale price of the property and it did not depend specifically on the area in which it was located the property.

3) We assumed that there were months in which the sale of properties was more frequent, we inferred this part because in a quick visualization of our database that there were periods during the year in which there could be an increase in the acquisition of estate.

4) The value of the land has experienced, on average, an increase since the exit from convertibility, if the time elapsed has determined the appreciation of the cost of the property over time, regardless of the year of construction of the property.

5) The demand for real estate is concentrated mainly in the South and Center areas of the city and it is there where the highest prices are registered, contrary to assumption number two, we wanted to explore the possibility that not only the number of available properties determined the rise or fall in real estate prices, but rather the reputation of the areas themselves determined the price regardless of whether the offer was wide or scarce.

# Data Analysis Process

As we know, the process of Data Analysis involves the collection, transformation, cleaning, and modeling of data to discover useful information of interest to a company or organization. All the data obtained are transformed into conclusions and used for decision making.

The data analysis process is based on several steps and phases. Findings from later phases may require reworking an earlier phase, implying a cyclical rather than a linear process. Most importantly, the success of data analysis processes depends on the repeatability and automation of each of these steps.

The phases are:

* Data Requirements Specification
* Data Collection
* Data Processing
* Data Cleaning
* Data Analysis
* Communication

In this case we did not have to develop the first two stages (Data Requirements Specification & Data Collection) because we have a data set already built and ready to process. However, for clarity, Data Requirements Specification determines what data and under what characteristics will be collected. This involves a bit of investigative work, such as talking to stakeholders, finding out who is responsible for the data and getting access to the data.

Data collection is the process of gathering information on specific variables that have been identified as data needs. It is important to ensure that the data collected is accurate, truthful and meets the requirements set forth in the requirements analysis. This is one of the most important steps since poor data collection could lead the study to biased or erroneous conclusions.

Data Processing was the first stage of the Data Analysis Process since we had an excel file previously collected and published by a company or organization. This file was converted to csv to facilitate and speed up its reading with R. An important fact is that there were fields that included commas in their content, so it was necessary to generate the csv with, as separator.

Graphical user interface, text, application, chat or text message

Description automatically generated

Figure 1: Data analysis process map

Data cleaning was performed by removing unnecessary columns, filtering, and organizing the remaining rows and then performing the analysis, which will be detailed later in this paper. The communication of our analysis is detailed in this paper where each of the conclusions obtained are explained.

# Importing Data

First, it was necessary to look for the correct data source off which it was possible to obtain the data referring to sales of properties for this project. At first there were some difficulties of finding the correct data fil since we had trouble finding suitable data sources, encountering data that had basically no flaws in its structure (no duplicates or NA values) or data that was inaccurate, but finally we found an interesting topic in the New York City Department of Finance’s website with data related to real state in Manhattan New York, which had not only reliable and accurate information, but it also had a structure suitable to applying different types of data handling process.

With this Excel sheet at hand, at first there was some simple eyesight analysis regarding the information that was contained in this file and there was also a discussion amongst the group members to determine exactly which were the useful fields of data and how we would import this dataset in R. But before importing the file to R, thanks to the first analysis, the first 4 rows that contained a brief and irrelevant description of the document were eliminated manually in Excel to make it easier to read for R, and then, the Excel sheet was saved as a csv file.

With the file converted, it was just a matter of computing the right input on R using the read.csv() function, writing down the correct directory and telling the function to read the csv file considering the separators as semicolon, saving the information in a data frame.

And so, after this process was done, we had a proper data source with a suitable structure to start working on it and give it the correct data treatment and processing to prepare it for calculations.

# Data Sanity Checks

## Variable Analysis

There are multiple objectives that we are trying to analyze with this dataset, the main one being the price of any residential property excluding commercial ones. The dataset that we chose to work on was one with very complete information about residences in Manhattan, but it was extremely dirty and with lots of variables we wouldn’t use, therefore I will explain each variable and why we use it or not on the final dataset.

* Neighborhood – As the name says it shows what neighborhood the property is, as a string categorical value so we treated as factor.
* Building class category – It shows what type of building is, for example, if it is a condo, an apartment with elevators, no elevators, a duplex, etc. As a factor.
* Tax class- The tax type class that affects the property, as a num categorical value, must be a factor.
* Block- The block where the property belongs, as a num value.
* LOT- The Lot.
* EASEMENT.
* Building Class at present – The type of building class that is registered as in the city, as a string but it is also classified already in subgroups, so we treated as factor.
* Address – The address of the place, as a string.
* Apartment Number- The number of the apartment, we decided to remove this variable because it didn’t apport anything to the analysis.
* ZIP code- The zip code, as a numeric categorical value. Should be a factor
* Residential units – It indicates the number of residential units on this address, we decided to remove this variable because we are only working on residential places so every place would be a residential place, it was redundant.
* Commercial units- Although we delete this column, it was helpful to let us know which places were not residential places.
* Total Units – It shows when there where more than one unit in one address, which for us wasn’t helpful because we weren’t working with that kind of places.
* Land Square Feet – It shows the total square feet of the place, is a numeric variable.
* Gross Square Feet - It shows the square feet of the construction in that place, is a numeric variable.
* Year Build- the year of construction, it helps us know how old the place was.
* Tax class at the time of sale – tax at the time of sale, is a factor.
* Building Class at Time of Sale - we decided to remove this class because we already had a building class.
* Sale Price – This is the target; this is what we want to find out at the end, and we use this variable to train the model.
* Sale Date – The date of sale, must be a Date variable.

## Treating Duplicate Entries

The dataset that we used was full of missing values and duplicates, spotting the duplicated values wasn’t that simple because the complete row didn’t repeat itself, but instead, we realized that duplicated values, in this case, were all the apartments that had the same characteristics on the same building, for example maybe all the apartment on the second floor of a building are exactly the same so instead of keeping all that information we could stay with just one of those apartments.

After cleaning the dataset of the columns, we didn't care about empty values, we decided what characteristics were what made a place a duplicate place, once we got that information, we proceeded

## Treating Missing Values

In this case, all the missing data were caused by missingness when they were creating the dataset. The CSV had a lot of information, but it was also recorded in a bad way, mixing some data in the same columns, and leaving a lot of empty values. After making a weekly analysis of which data is the one that needs cleaning we decided to combine the address with the number, to remove all rows that had missing data on the characteristics that were important to us, and in this case we couldn’t fill it up with data created by us, like the mean of all the column, because this was independent data in most cases or data that wasn’t affected by the other rows.

# Data Summarization, Visualization, Statistical Analysis, and Calculations

To show and analyze sales by month, it was necessary to analyze the data available. The first analysis was that we only had data corresponding to 2020, so it was not necessary to consider the year. Knowing this, a copy of the original dataset was made to avoid any type of alteration or damage to the data and a new column was added corresponding only to the month of the date of sale, this operation was performed with the format() function since the input data were in date format.

After having the data classified by month, the aggregate() function was used to count the number of sales records corresponding to each month of 2020. With the plot() function, a graph was generated and it was concluded that the months with the most sales were June and December. However, we had a column corresponding to the type of building that had data such as A1, B3, C6… (Property Shark, 2022). We wanted to analyze the sales behavior for each category, so we generated them by eliminating the number and obtained as a result only the letters of the 6 categories in function. A graph was made with the behavior of each category in question, and we realized that there is a significant difference in sales according to the class of building.

****

Figure 2: Overall sales in 2020

In Figure 2 we have a graph representing the overall sales carried out in 2020. You can clearly see how June and December have peaks in sales, so you could conclude that these are the appropriate months to offer a property. However, we must go further and analyze each of the building categories for which we have data in Figure 3.

When we look in detail, we realize that sales behave differently depending on the construction category, so it is of utmost importance to analyze the property to be sold in order to make decisions about the offer or other aspects.

****

Figure 3: Overall sales by building categories

To know which are the properties that were of interest for our analysis, it was necessary to be able to filter the database that we managed, so the first part was through the “filter” function, it was decided that only the buildings would remain in the database. that have been built before 2002 since these would be the ones that would have an age greater than 20 years of construction, with this result a new data frame was generated, this new data frame is used to perform a count, through the "count" function ” for each year to know how many buildings were built respectively each year, and then use the “barplot” function to view it in a bar graph, this being a visual aid for future analysis.

Chart, bar chart, histogram

Description automatically generated

Figure 4: Buildings older than 20 years old

For the analysis regarding the correlation between the tax class and the price of the properties, first it was necessary to create a new data frame with the variables that were going to be analyzed. Then correctly name every column and, since tax class is factorized, convert it to a numeric variable and append this new column to the data frame.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Figure 5: Correlation between tax class and price

Then price was divided by a million to get a number easier to visualize by plotting and finally, use the cor() function to calculate correlation, which gave -0.03 as a result. With this in hand, it was found that the tax class of properties does not necessary affect the price of said properties, due to the low correlation and how this can be seen on a plot which shows that, even though there is a concentration of properties on tax class 2, there is no sign of effect to the prices based on the tax class.

# Predictive Analysis and Visualization

For this analysis we decided to do a multiple linear regression model, to do it we started by creating three new variables, building age where subtracted the building year from 2022, price millions where we divided the price by a million, and gross hectares where we transformed the gross square feet to hectares.

Once we have the new variables, we could see that we have some outliers that were clearly an input error on the price that could induce bias on this model, to see some examples reference the indexes 2, 3, 193, and 2101 from the manhattan2 table. To get rid of the outliers we divided the data set in two different ones, manhattaninf that included the first 32 quantiles, and manhattansup that included the rest of the variables.

Chart, scatter chart

Description automatically generated

Figure 6: Trained predictive model (manhattansup)

Once we had the new data set, we calculated a correlation with the manhattansup data frame where we expected the price to be the dependant variable, the total units, gross hectares, and the building age to be the independent variable. We can see that for each increment in total units had a positive impact by 0.51, the gross hectares had a positive impact of 0.27, and building age had a negative impact of 0.09.

We proceed to create a model that considered the price in millions as the dependant variable and two independent variables, total units and gross hectares, the building age was excluded because the correlation was insignificant.

To visualize the Housing market, we used a 3d scatter plot from the library "scatterplot3d", where the total units are displayed on the x axis, the gross hectares are displayed on the y axis and the price in millions is displayed on the z axis. The limits thar were considered to display the information for both the x axis and z axis is from 0 to 200 and for y axis we considered from 0 to 10.

This model gives us three low p values which means that is a good predictive model that can help us as a starting model, nevertheless the multiple R squared value is 0.28 which means these two variables only explains the 28% of the price. We recommend that to go forward to search for extra variables to improve the accuracy of these model with time.

At the end we decided to use this model to fill the price in millions variable of the data set manhattaninf that had the outliers with the input errors. We now see more reasonable data of these values, before the model we saw a range from 0 to 0.2 million dollars, and after the model we see a range from 5.36 to 119.92 million dollars.

A picture containing chart

Description automatically generated

Figure 7: Predicted prices

# Summary

As it is possible to see, there were multiple processes of data handling that were applied on this dataset and different approaches chosen to tackle the variety of problematics presented in this case study. As for the main findings, these were the following:

* It was possible to predict prices for different properties, going through a range from 5.36 to 119.92 million dollars.
* 1201 buildings are 20 years old or older.
* The tax class does not directly affect the price of the properties.
* June and December being the month with the greatest number of sales.
* The number of sold properties is spread across almost every building category.

As we can see, the main objectives were covered in this analysis, getting valuable and diverse insights into the real state status of New York City, specifically in Manhattan. With this information it is possible for companies to get a better understanding of the properties that are being sold in this area of NYC, which types of buildings and where are these properties being sold the most.

As part of the final conclusions, we can see that the supply of real estate in residential areas is based on the valuations of multiple services, in addition to variables that value the price of the properties. With the use of descriptive and predictive programming, relevant characteristics such as measurement, state of conservation, and the variables involved in defining the interest of the offer presented in the real estate market are specified.

Firstly, because of the analysis, it can be concluded that the realization of this research project has no market, technical, or financial restrictions. This means that it exists, and we were able to verify that there is a real estate demand and a specific market segment for the studied areas of medium-high, and high socioeconomic levels.

As a second conclusion when analyzing and evaluating the real estate offers, it is through this project to be able to generate an idea of ​​the economic feasibility where the variation of the land, the construction site, the quality of the infrastructure, access to immediate services, urbanization, as well as the two most important determining factors in the price, the supply of the area and the valuation of the area.

And finally, the power that R contains in its systems can really help companies, through programming, to get valuable and accurate insights on different real-world topics and case studies that may have a great impact in the operation of these companies, and help them make better decisions.

# Appendix

## R code

manhattan <- read.csv("rollingsales\_manhattan.csv", sep = ";");

#eliminate unnecesary columns

manhattan <- subset(manhattan, select = -c(BOROUGH, BLOCK, BUILDING.CLASS.CATEGORY,

LOT, EASEMENT, APARTMENT.NUMBER,

TAX.CLASS.AT.TIME.OF.SALE,

BUILDING.CLASS.AT.TIME.OF.SALE));

colnames(manhattan);

#rename columns

names(manhattan) <-c("neighborhood", "tax\_class",

"building\_class", "adress",

"zip\_code", "residential\_units",

"commercial\_units", "total\_units",

"land\_sqft", "gross\_sqft", "year\_built",

"price", "sale\_date");

colnames(manhattan);

nrow(manhattan);

#eliminate periods/dots in price, land\_sqft and gross\_sqft

library(stringr);

manhattan$price <- str\_remove\_all(manhattan$price, "\\.")

manhattan$land\_sqft <- str\_remove\_all(manhattan$land\_sqft, "\\.")

manhattan$gross\_sqft <- str\_remove\_all(manhattan$gross\_sqft, "\\.")

# eliminate entries without year\_built

manhattan <- manhattan[!is.na(manhattan$year\_built),];

nrow(manhattan);

# create data frame with building class a:d, s

manhattan1 <-manhattan[grep("A|B|C|D|S", manhattan$building\_class),];

nrow(manhattan1);

#eliminate entries with less than 1 residential unit, land and gross sqft

manhattan1 <- subset(manhattan1,

subset = manhattan1$residential\_units > 0 &

manhattan1$land\_sqft > 0 &

manhattan1$gross\_sqft > 0);

nrow(manhattan1);

object.size(manhattan1);

#assign variable types

manhattan1 <- transform(manhattan1,

neighborhood = as.factor(neighborhood),

tax\_class = as.factor(tax\_class),

building\_class = as.factor(building\_class),

adress = as.character(adress),

zip\_code = as.factor(zip\_code),

residential\_units = as.integer(residential\_units),

commercial\_units = as.integer(commercial\_units),

total\_units = as.integer(total\_units),

land\_sqft = as.numeric(land\_sqft),

gross\_sqft = as.numeric(gross\_sqft),

price = as.numeric(price),

sale\_date = as.Date(sale\_date, "%d-%m-%y"));

object.size(manhattan1);

#eliminate adress duplicates

manhattan2 <- manhattan1[!duplicated(manhattan1$adress),];

nrow(manhattan2);

#quantile to considerer

quantile(manhattan2$price, probs = 0.32);

#data frame inferior and superior based on the quantile

manhattaninf <- manhattan2[manhattan2$price < 221600,];

manhattansup <- manhattan2[manhattan2$price >= 221600,];

#correlation price and tax class

cormanhattan <- data.frame(manhattansup$tax\_class, manhattansup$price)

colnames(cormanhattan) <- c("tax\_class", "price")

cormanhattan <- cbind(cormanhattan, cortax\_class = as.numeric(cormanhattan$tax\_class))

cormanhattan$price <- round(cormanhattan$price/1000000,3)

cormanhattan

cor(cormanhattan$cortax\_class, cormanhattan$price)

plot(cormanhattan$tax\_class, cormanhattan$price,

main = "Correlation Tax Class/Price",

xlab = "Tax Class",

ylab = "Price (in MM$)")

#----PROPERTIES WITH MORE THAN 20 YEARS OF CONSTRUCTION-----

#Install packges and libraries

install.packages("tidyverse")

install.packages("plyr")

library(plyr)

library(tidyverse)

library(datasets)

#Filter the necessary information, and generate a new database

filtro\_year\_built <- manhattan2 %>% filter(year\_built < 2002);

#Make a count by year of construction

Conteo\_year\_built<-count(filtro\_year\_built, "year\_built");

Conteo\_year\_built

#Create a graph with the filtered data.

barplot(table(filtro\_year\_built$year\_built));

# -------------- SALES BY MONTH --------------

manhattan\_sales <- manhattan2

# Obtain only the month of sale date

manhattan\_sales <- transform(manhattan\_sales,

month = as.numeric(format(manhattan2['sale\_date'][,1],'%m')))

# Counting sales per month

sales <- aggregate(manhattan\_sales$month, by = list(manhattan\_sales$month), FUN = NROW)

names(sales) <- c('month', 'value')

sales['month\_name'] <- month.name[sales$month]

summary(sales$value)

par(mfrow = c(1, 1))

plot(sales$month, sales$value, type ="o", main = "Sales per month",

xlab = "", ylab = "Sales", col = 4, lwd=3, pch=19, yaxt = "n",

xaxt = "n")

axis(1, at = sales$month,

labels = sales$month\_name, las = 2, tck = 1, lty = 2, col = "gray")

axis(2, tck = 1, lty = 2, col = "gray")

# Sales by class

# Obtain only the letter of each class

manhattan\_sales['building\_class2'] <- substr(manhattan\_sales[,'building\_class'], 1,1)

# Counting sales per month & class

sales\_by\_bld <- aggregate(manhattan\_sales$month, by = list(manhattan\_sales$month, manhattan\_sales$building\_class2), FUN = NROW)

names(sales\_by\_bld) <- c('month', 'building\_class', 'value')

sales\_by\_bld$building\_class <- factor(sales\_by\_bld$building\_class)

levels(sales\_by\_bld$building\_class)

par(mfrow = c(3, 2))

plot(sales\_by\_bld[sales\_by\_bld['building\_class'] == 'A',][c(1,3)], type = "l", col = 7, main = "Class A", xlab = "", ylab = "")

plot(sales\_by\_bld[sales\_by\_bld['building\_class'] == 'B',][c(1,3)], type = "l", col = 2, main = "Class B", xlab = "", ylab = "")

plot(sales\_by\_bld[sales\_by\_bld['building\_class'] == 'C',][c(1,3)], type = "l", col = 6, main = "Class C", xlab = "", ylab = "")

plot(sales\_by\_bld[sales\_by\_bld['building\_class'] == 'D',][c(1,3)], type = "l", col = 4, main = "Class D", xlab = "", ylab = "")

plot(1:12, replicate(12, 0), col = 0, main = "Class H", xlab = "", ylab = "", ylim = c(0,2), yaxt = "n")

axis(2, at = c(0, 1, 2))

points(sales\_by\_bld[sales\_by\_bld['building\_class'] == 'H',][c(1,3)], col = 1, pch = 19)

plot(sales\_by\_bld[sales\_by\_bld['building\_class'] == 'S',][c(1,3)], type = "l", col = 3, main = "Class S", xlab = "", ylab = "")

#--------------------QTY SOLD BY NEIGHBORHOOD

salesxneigh<-aggregate(manhattansup$isSale, by=list(neigh=manhattansup$neighborhood), FUN=sum)

#-----------DESCENDING ORDER

a<-salesxneigh[rev(order(salesxneigh$x))]

#-----------TOP 10

salesxneigh <-a[1:10,]

#-----------RENAME COLS

names(salesxneigh)<-c("nei","qty")

#-----------PLOT IN BARS

ggplot(data= salesxneigh, aes(x = reorder(nei,-qty), y=qty)) +

geom\_bar(stat="identity", fill="steelblue", alpha = 0.8)+

geom\_text(aes(label= salesxneigh$qty), vjust=-0.3, size=3.5)+

labs(title="Quantity Sold by Neighborhood",

x="Neighborhoods", y = "Quantity Sold")+

theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5, size=6.5))

theme\_minimal()

# -------------- PREDICTING PRICES --------------

#New variables

manhattansup$building\_age <- 2022 - manhattansup$year\_built;

manhattaninf$building\_age <- 2022 - manhattaninf$year\_built;

manhattaninf$price\_millions <- manhattaninf$price / 1000000;

manhattansup$price\_millions <- manhattansup$price / 1000000;

manhattaninf$gross\_hectare <- manhattaninf$gross\_sqft / 107639;

manhattansup$gross\_hectare <- manhattansup$gross\_sqft / 107639;

#Correlation

cor(manhattansup[,c(8, 14:16)]);

#Linear regretion

model <- lm(price\_millions ~ total\_units + gross\_hectare, data = manhattansup);

summary(model);

#Plot

install.packages("scatterplot3d") # Install

library("scatterplot3d") # load

scatterplot3d(manhattansup$total\_units, manhattansup$gross\_hectare, manhattansup$price\_millions,

main = "Housing market", xlab = "Total Units", ylab = "Gross land in hectars", zlab = "Price in millons",

xlim = c(0,200), zlim = c(0,200), color = "blue");

#Predicting price

predicted\_price\_millions <- predict(model, manhattaninf);

#Table with the predicted price

predicted\_price <- cbind(manhattaninf[,c(1,3:5,8,11,16)], predicted\_price\_millions);

predicted\_price;

scatterplot3d(predicted\_price$total\_units, predicted\_price$gross\_hectare, predicted\_price$price\_millions,

main = "Housing market", xlab = "Total Units", ylab = "Gross land in hectars", zlab = "Price in millons",

xlim = c(0,200), zlim = c(0,30), color = "blue");

## Personal Reflection (Alejandra)

During the project we have had to perform multiple analyses with the data worked, although each of them had some complexity and required distinction when dealing with them, the issue that most caught my attention was the cleaning of the data. This is one of the main steps as it opens the way to the rest of the analysis, if we do not have clean, concise, and truthful data we will not be able to reach a good analysis with the same characteristics.

I liked how as a team we analyzed the original dataset and together we defined which data was needed, which to discard, how we could clean the remaining data, among many other things. We were a very committed and willing team for the project, which facilitated communication and task distribution. Sometimes it was difficult to meet because all members had things to do, but this did not prevent the project from running smoothly.

I also found the way we divided the tasks very good, I found it very useful because it allowed us to have our work well organized and the tasks distributed according to the capabilities of each member.

Finally, I would like to thank the team for their patience, commitment and willingness to make this project a success.

## Personal Reflection (Daniel)

Through the course there was a lot of topics that I really enjoy, since learning a new language that is really useful to manipulate large amount of data like R. But if I had to a pick a favorite will be both the linear regression and logistic regression because these predictive models help the business to have healthy proactive approach that helps them to avoid problems.

On this project I was able to practice and improve a lot of the topics that we learned through the course, but two of the things that stuck with me was first to code in a cleaner and more descriptive way, so it was easier to everyone in the group to read. The other thing that I learned about this project is to try to normalize the variables, so it is easier to compare them.

This group has a lot of strengths, but I think our best asset is how we organized on a efficient way all the tasks so we could finish the project at time and with the right output. One area to improve I will say is how we research for information.

## Personal Reflection (Francisco)

The topic that was of most interest for me was the fact that we got to apply all the knowledge that we were given through out the whole course into real data and a possible real-life problem and case study to solve. Applying the different data processing and summarization to the dataset that we chose was something that caught my eye because it gives a better view of how all the things that we were taught can be applied in different ways and approaches.

As for the best thing about executing the project, I would say working as a team is one of the most important things because during the module we mainly did assignments, quizzes and worked by ourselves but in this project we had to know how to tackle the issue as a team and I think that is a key component in the process of learning since, in the real world, working in teams will be a day to day thing.

And about the team we formed, I would say that our best asset was to be organized and the sense of urgency to get through this project in the most efficient way possible, also helping each other in the process and to get the best results.

## Personal Reflection (Eduardo)

The topic that I found most interesting was the part about dealing with duplicates and missing values because that is the most challenging part of the process and in most cases, it is what a lot of jobs will ask you to do. I think that the best thing I learned while making the project was really understanding the problem, and what we needed to do in order to achieve our goals and not just deleting stuff and columns. I think we could improve as a team if we worked in person, but in general, it was a very good experience.

## Personal Reflection (Cesar)

The part that I found most interesting was section L, predictive analysis and visualization since it encompasses most of the development of the project, from the initial approaches through administration and data management and concluding with the generation of a predictive model. thus, having a future vision of the behavior of the market that we investigate in this project.

The best thing I learned through this project is to be able to visualize and unite in a practical way, each of the topics that we studied during this module, as well as to apply the programming knowledge that we acquired in a real case.

As a group we had many strengths, we really knew how to collaborate so that each one could develop the parts of the project in which they felt that their skills were outstanding, personally, I find that the area in which they could improve is related to code generation, specifically in logic. to perform more complex functions.

## Personal Reflection (Jehison)