Data Visualization

Individual Assignment Solution

Name: Barun Kumar

**Table of Contents**

[1 The Assignment 3](#_Toc24734811)

[2 The Solution 4](#_Toc24734812)

[2.1 Loading & Understanding the Data 4](#_Toc24734813)

[2.2 Data Exploration 4](#_Toc24734814)

[2.3 Feature Engineering 11](#_Toc24734815)

[2.4 Putting it all together 13](#_Toc24734816)

# The Assignment

In this individual assignment, we are required to work on the dataset provided “Boston Condo Info” in excel format:



The dataset contains following variables:

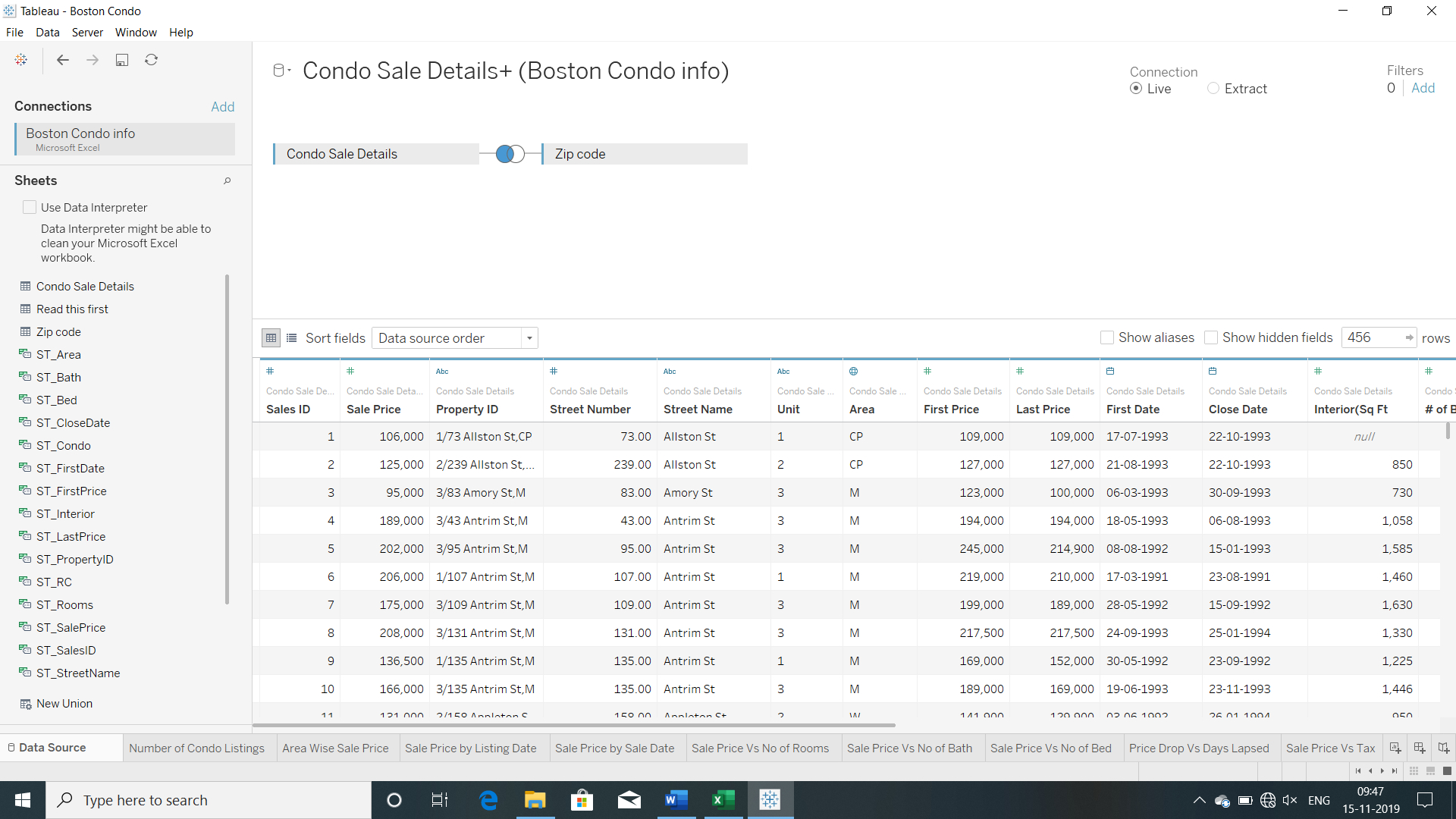
|  |  |
| --- | --- |
| Name | Description |
| Sales\_ID | S No of transaction |
| Sale Price | Price the house was sold for (USD) |
| Property\_ID | Address |
| Street Number | Street number |
| Street Name | Street name |
| Unit | The unit number of the house (Taken from the address) |
| Area | Area / Locality |
| First Price | First price put up by the seller (USD) |
| Last Price | Last price the buyer had decided to hold on to (USD) |
| First Date | Date at which the condo first came for sale |
| Close Date | Date at which the condo was sold |
| Interior(Sq Ft | Sq Feet of condo (Total) |
| # of Bed | No of bed rooms |
| # of Bath | No of bathrooms |
| # of Rooms | Total number of rooms |
| Condo Fee | The fee associated for maintaince, public common property etc |
| Tax | Property tax (USD) |
| RC | This is a boolean value. RC -0 allows home owners to stictly use the condo for residential purposes only, RC - 1 allows home owners to use the property aas residential as well as Small commerical establishments (Like having a boutique, florist, boulangere/Bakery in their home) |

We need to produce an interactive story about the given dataset and publish it to tableau public.

# The Solution

## Loading & Understanding the Data

The dataset contains information about condo sales made in Boston, US from 1988 to 1994. We begin by importing the dataset in Tableau and checked for correctness of data types. The excel sheet contains two sheets apart from the data description sheet, “Condo Sale Details” and “Zip Code”. Sheet “Zip Code” is basically a look-up table providing Zip Codes for the Area listed in sheet “Condo Sale Details”. We joined both sheets before importing in the Tableau.



We confirmed and corrected data types for all variables. Now, we begin with the exploration of each variable one-by-one.

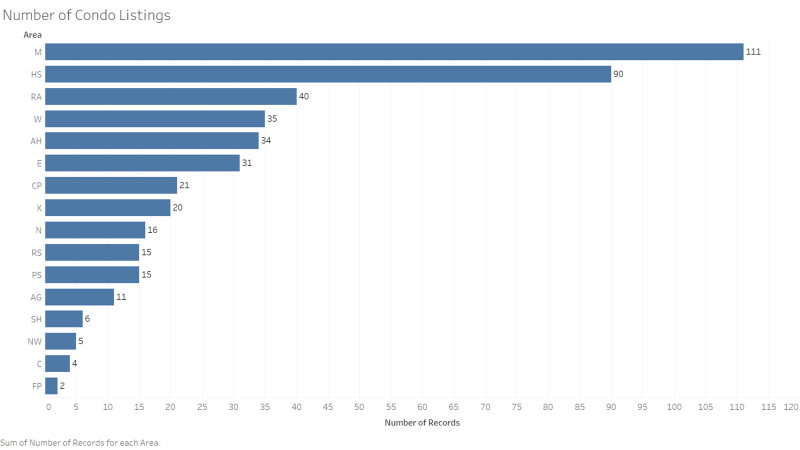
## Data Exploration

We can see that there are **10 dimensions** and **13 measures** in the dataset.

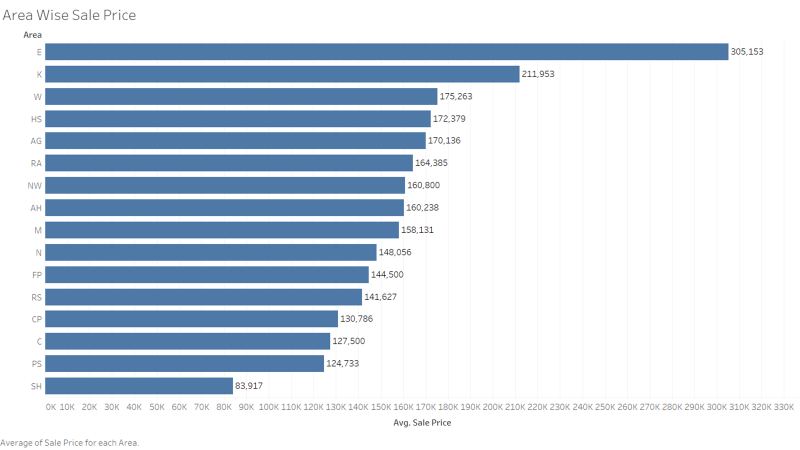
“Sale Price” is the measure of our interest. We will visualize each variable with respect to Sale Price and try to build a story out of it. Let’s begin!

**Variable “Area”**

This variable contains information about areas in Boston where the condo belongs to. Let’s see the number of condo listings in each of these areas.



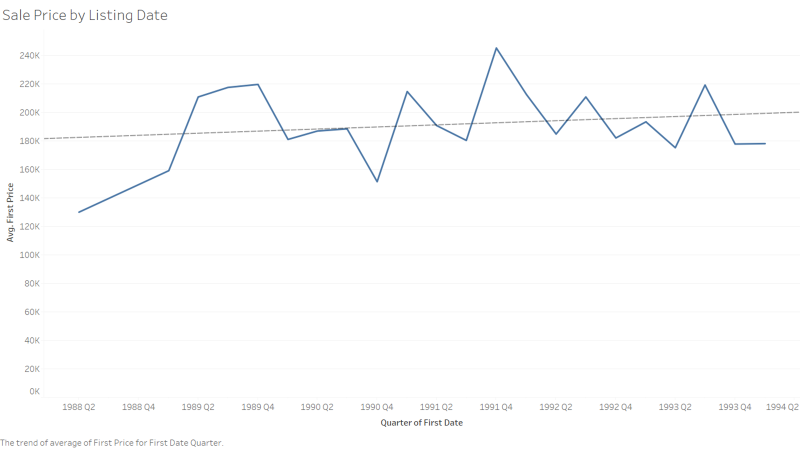
We can see that maximum number of condos were sold in area M, followed by area HS. Now let’s see the average sale price in all these areas.



Area E has the maximum average sale price followed by area K. Area M, which had maximum number of condo sales, is at 9th position out of total 16 areas in Boston.

**Variable “First Date”**

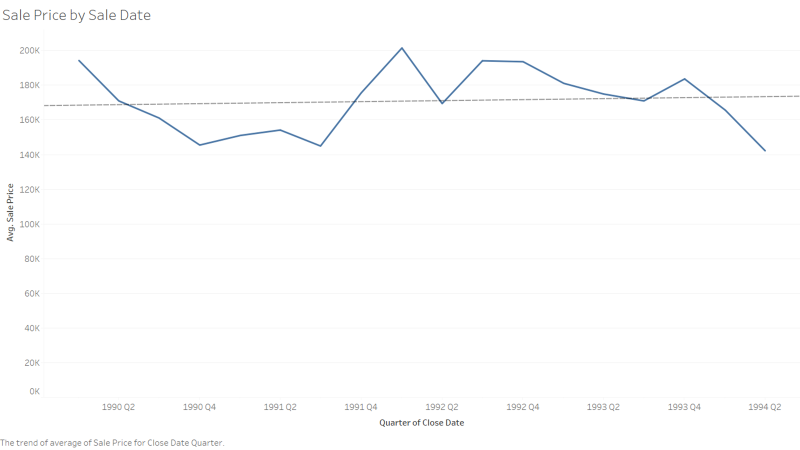
This variable contains date on which the condo first came for sale. The seller lists its condo on a certain initial price. Let’s visualize First Date w.r.t First Price (First price put up by the seller).



There is a almost no trend in average first price put-up by sellers over the quarters.

**Variable “Close Date”**

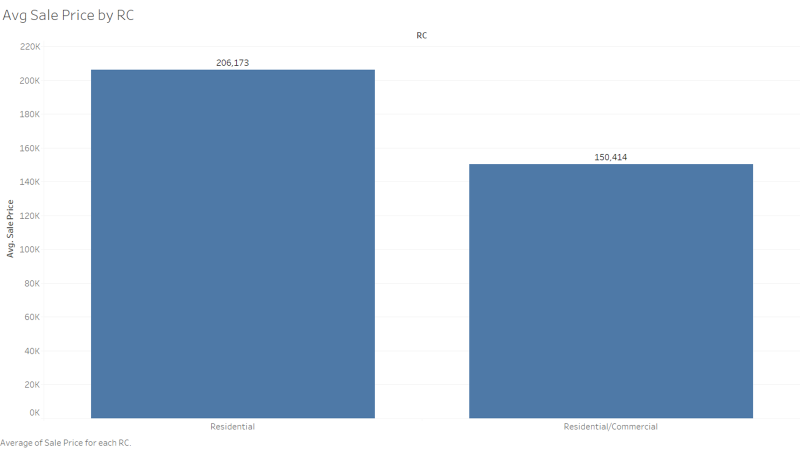
This variable contains date on which the condo was finally sold. Let’s visualize this w.r.t Sale Price.



We see that average sale price remains almost same over the quarters.

**Variable “RC”**

This is a Boolean variable. RC -0 allows home owners to strictly use the condo for residential purposes only, RC - 1 allows home owners to use the property as residential, as well as small commercial establishments (Like having a boutique, florist, boulangerie/Bakery in their home). Let’s visualize the average sale price for these two types of condos.



Clearly, average sale price for residential condos are high compared to the other one.

**Variable “Street Name”**

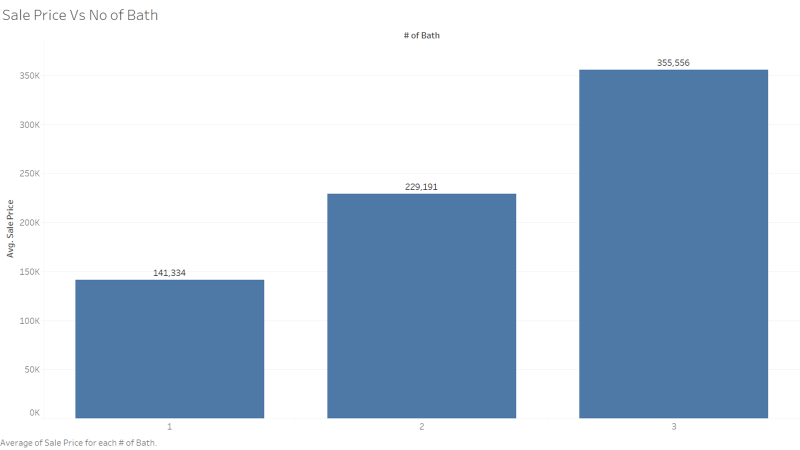
This variable contains name of the street where the condo is located. Let’s visualize the street wise average sale price.



It is clear from above word cloud that **Cambridge Pky** is the costliest street followed by **Dana St** and **Lancaster Ter**.

**Variable “# of Bath”**

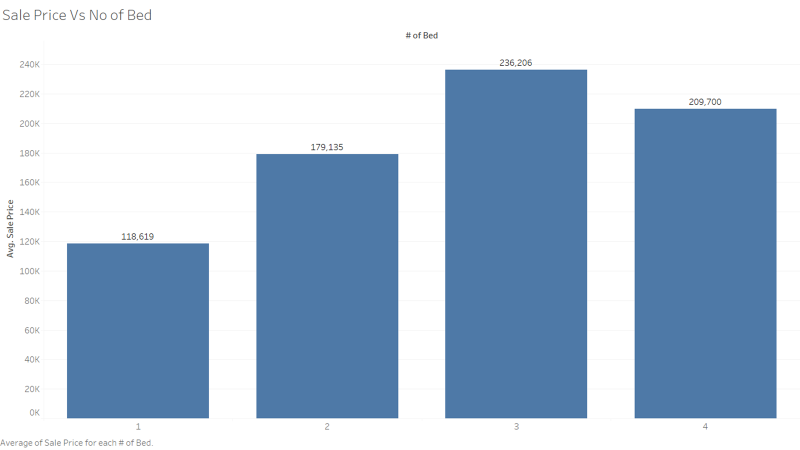
This variable contains information about number of bathrooms a condo has. Though this variable is numeric we convert this to dimension since statistical measures doesn’t make any sense here. Let’s see how average sale price depends on this variable.



We see that the average sale price increases as the number of bathrooms increases. Maximum number of bathrooms a condo has is 3.

**Variable “# of Bed”**

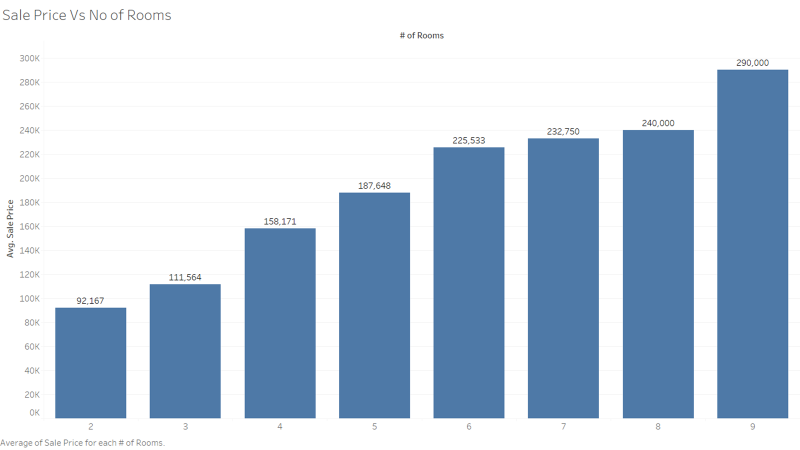
This variable contains information about number of bedrooms a condo has. We convert this into a dimension as well, going by our previous logic. Let’s see how average sale price depends on this variable.



Average sale price increases as the number of bedrooms increases till 3 bedrooms, thereafter the average sale price dips off.

**Variable “# of Rooms”**

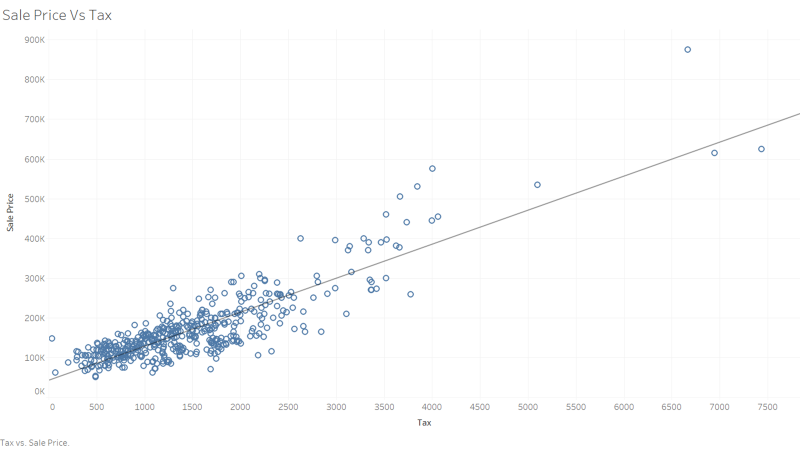
This variable contains information about total number of rooms (including bathrooms and bedrooms) a condo has. This variable is also converted into dimension. Let’s see how average sale price depends on this variable.



We see that the average sale price increases as the total number of rooms increases. It is also worth noting that some of the condos have additional rooms (except bathroom & bedroom), like storage rooms. These kinds of condos might particularly be used for commercial purposes.

**Variable “Tax”**

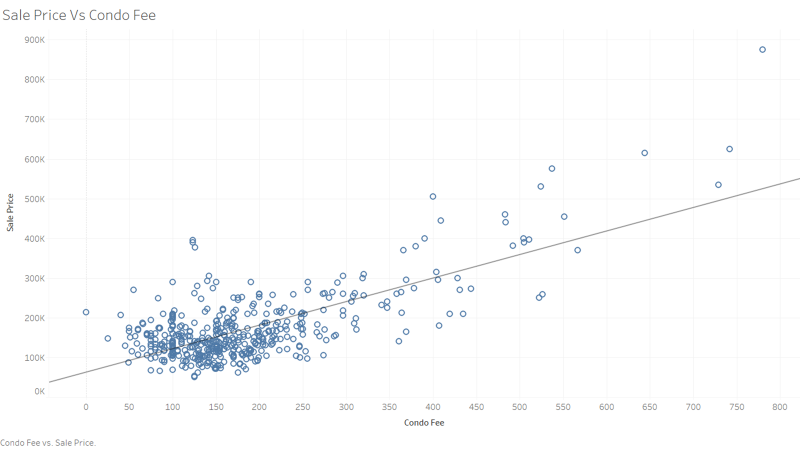
This is the property tax of a condo. Let’s visualize this w.r.t the sale price.



We see that sale price and property tax are highly correlated.

**Variable “Condo Fee”**

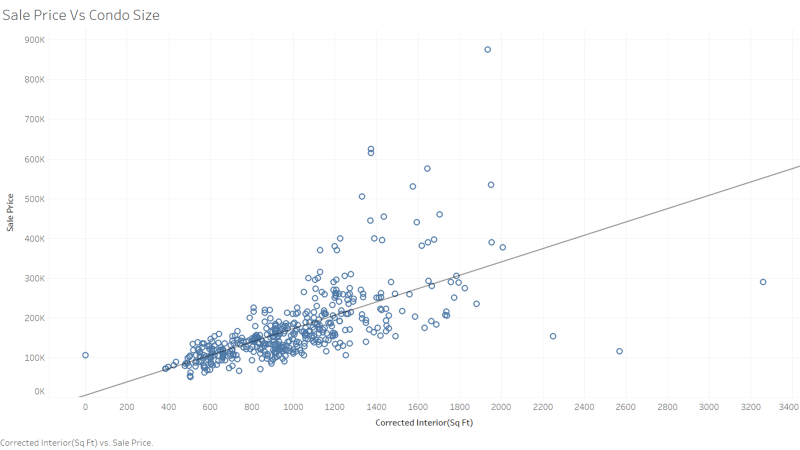
This is the fee associated for maintenance, public common property etc. of the condo. Let’s visualize this w.r.t sale price of the condo.



We can see that sale price and condo fee are highly correlated.

**Variable “Interior (Sq Ft)”**

This variable represents the total square feet area of condo. This variable contains one missing value. We have imputed that missing value with the median square feet area. Let’s visualize this w.r.t sale price.



We see that there is high correlation between sale price and square feet area of the condo.

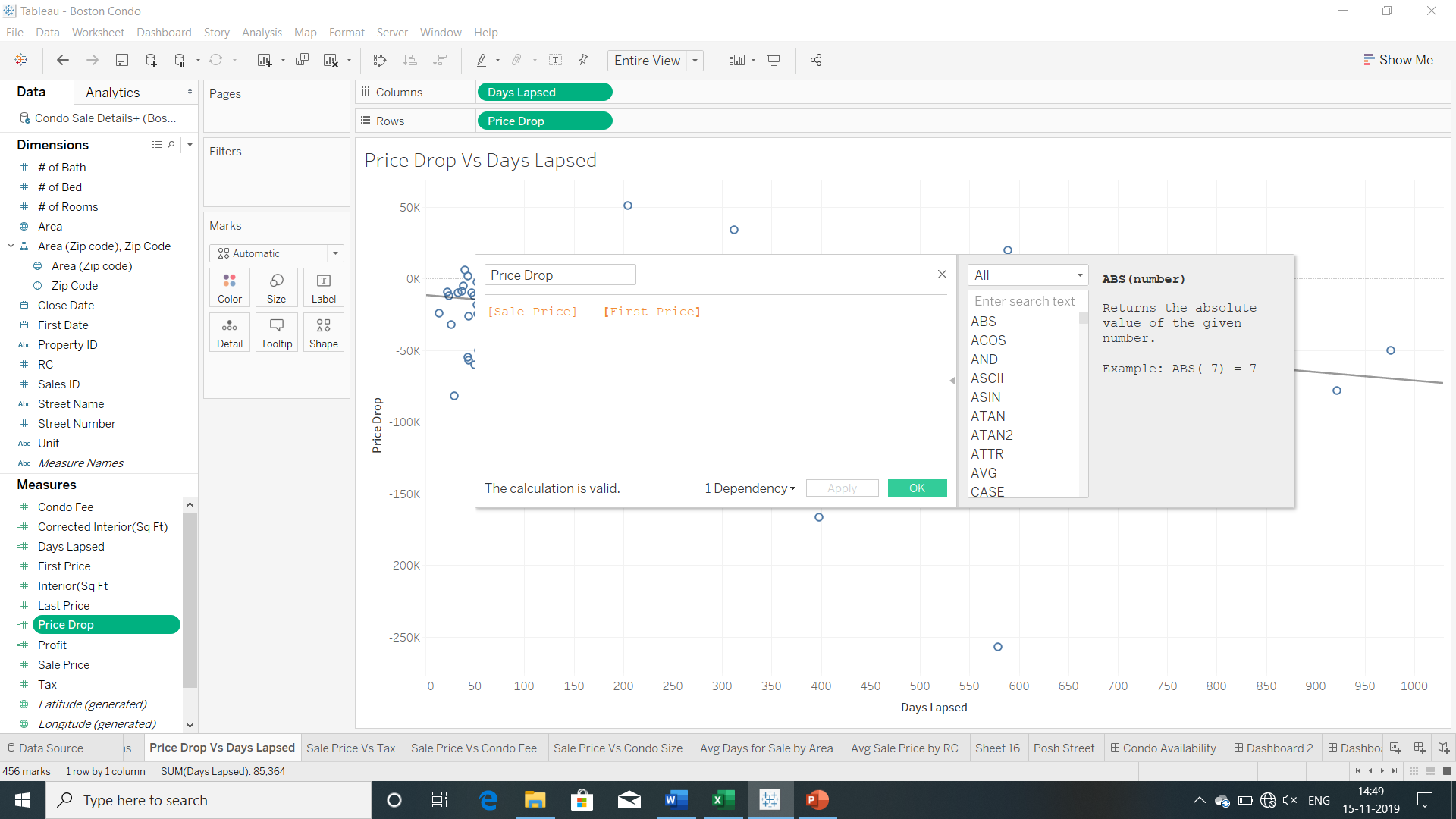
We have covered all our dimensions and measures except the variables “Property ID”, “Sales ID”, “Street Number” and “Unit”. These variables are mostly unique id’s and doesn’t make any sense to analyze. Let’s now try to create some additional features with the given variables.

## Feature Engineering

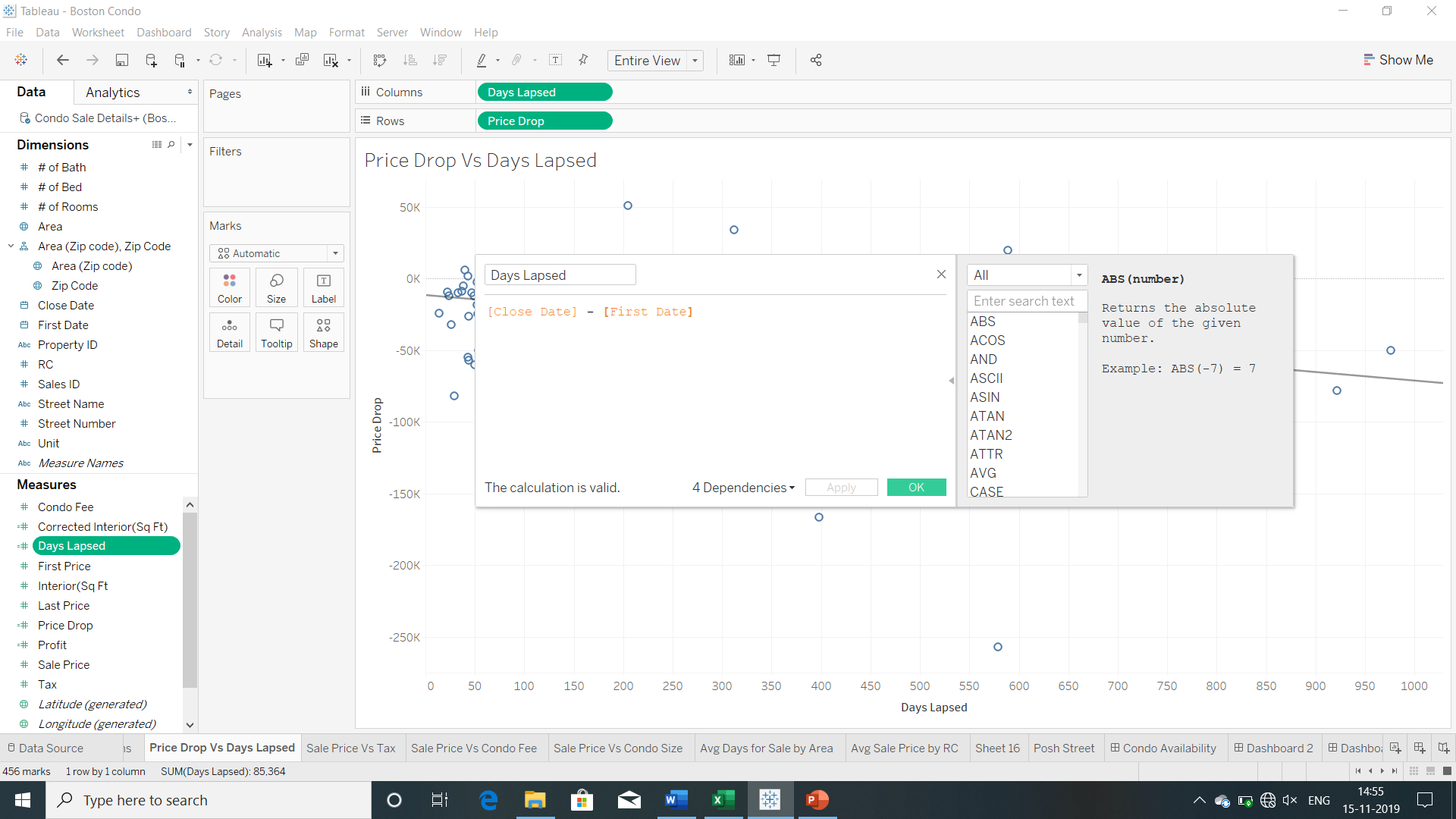
We are given information about First Price and Sale Price of the condo. First Price is the price at which the condo first came for listing while the Sale Price is the final price at which the sale happened. We can create a new feature “Price Drop” based on these two variables.

Similarly, we have variables First Date and Close Date. First Date is the date at which the condo first came for selling and Close Date is the date on which the final sale was made. These variables can be used to create a new variable “Days Lapsed”, which indicates after how many days a condo gets sold. Let’s create these two variables in Tableau.

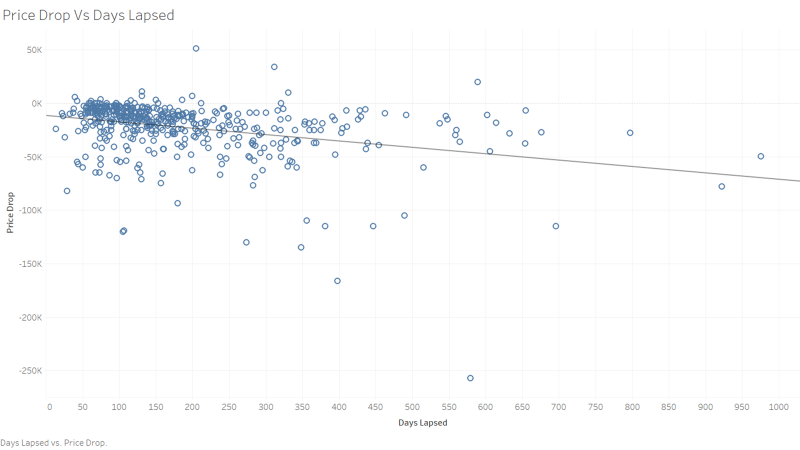
**Creating New Variable “Price Drop”**



**Creating New Variable “Days Lapsed”**

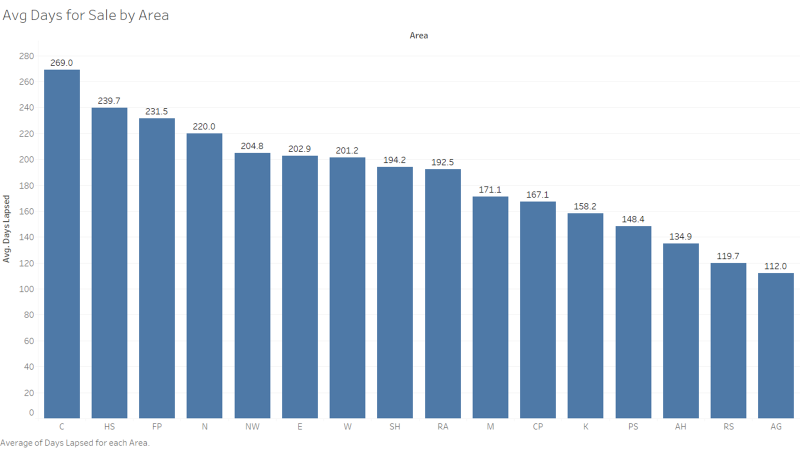


Now, that we have created our variables let’s see how these two are related.



We can see that there is drop in price of the condo as the days lapsed since it was listed for selling.

Let’s also visualize what is the average days taken to sell a condo in a particular area.



We can see that on an average it takes 269 days to sell a condo in area C followed by around 240 days in area HS. It takes minimum 112 days to sell a condo in area AG.

## Putting it all together

Now, that we have done variable exploration and feature engineering let’s try to build a story out of our exploratory analysis. We will try to build our story from a buyer perspective. This would be helpful for a potential buyer in future. We will try to answer following questions through our story:

* Which area is having more availability of condo?
* What is the average condo price in a particular area? Which is the cheapest and costliest area.
* In a particular area, what condo options are available?
* Which factors affects the sale price of the condo?

Following is the link to the tableau story created, which tries to answer the above questions-

<https://public.tableau.com/profile/barun.kumar#!/vizhome/BostonCondoSalePriceAnalysis/Story1?publish=yes>