ps1_answers

September 11, 2023

1 BA 222 PS 1

Ethan Chang

---

1. Describe the data. What does each observation correspond to? How many variables are there? How many observations are there? What different types of variables do we observe? You do not need to be exhaustive, but write a few complete sentences.

Each observation corresponds to a different property sale in NYC, and each property is described by 21 variables. There are a total of 84548 observations/properties. There are a variety of different types of variables, including numeric and categorical. There also also dates and addresses.

---

2. What is the mean of the Year Built column? Look at the Year Built column. What are the minimum and maximum values? Do these seem reasonable? Sort your data by year built, and delete all 0 values. What is the mean of the remaining data? (Bonus: As an advanced move, there are many ways of looking at the mean of nonzero values without sorting or modifying the data.)

The mean of the Year Built column is about 1789, while the minimum and maximum values are 0 and 2017, respectively. The minimum value is not reasonable, but the maximum value is. After finding the average without taking the 0's into account, the mean becomes about 1950, which is more reasonable.

---

3. Using the COUNTIF function, count how many observations in the Zip Code column are from zip code "10011". Look at a map of New York City on google maps. Where is zip code 10011?

There are 1048 observations from zip code 10011. Zip code 10011 is in/near Manhattan.

---

4. What is the sum of the recorded sale prices? Format this value in a cell with commas and no decimal places. Why are some sales prices 0? Read the data glossary to find out.

The sum of Sale Prices is $\$89,335,360,909$. Some of the sales prices are 0 because "there was a transfer of ownership without a cash consideration."

---

5. Consider the total units column. What are the median number of units in a sale? What is the standard deviation of the number of units? What do you think this tells us about the distribution of the number of units? Write a sentence telling us about the distribution.

The median and standard deviation of the number of units in a sale are 1 and roughly 19, respectively. The standard deviation is large, which tells us that the distribution of the number of units is very spread out. The median is 1, which tells us that 50% sales have 1 unit or less, but there are sales with a large number of units that skew the data to the right and increase the standard deviation.

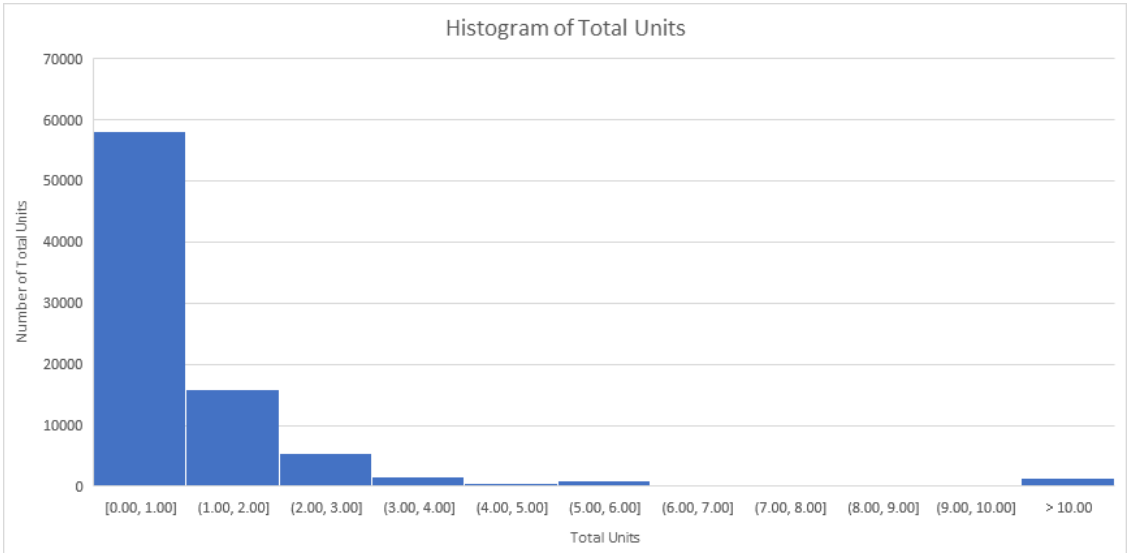
---

6. Now, plot a histogram of the number of the Total Units column. Give the histogram useful labels (X axis title, Y axis title, graph title)

The histogram has an overflow bin as there were extreme outliers in the dataset which made the histogram difficult to read.

[1]:
```
from IPython.display import Image

Image(filename='histogram.png')
```

[1]:


---

7. How many commercial units were sold? How many total units were sold? What fraction of the total units sold were commercial units?

There are 16365 commercial units sold, and 190164 total units sold. The fraction of total units sold that were commercial units is about 0.086 or 8.6%.

Excel Stuff

[2]:
```
Image(filename='excel_stuff.png')
```

[2]:

	A	B
1	Num variables	21
2	Num observations	84548
3		
4	Year Built Mean	1789.322976
5	Year Built Max	2017
6	Year Built Min	0
7	Non-zero Year Built M	1950.084805
8		
9	Zip Code 10011	1048
10		
11	Sum of Sale Prices	89,335,360,909
12		
13	Median Total Units	1
14	Total Units STDev	18.97258443
15		
16	Commercial Units	16365
17	Total Units	190164
18	Commercial/Total	0.086057298