Images haven't loaded yet. Please exit printing, wait for images to load, and try to print again.

Home Remodeling Analysis Turned Excel Data Handling in Python

Why cleaning data is the most important step

Manjula Mishra

Dec 14, 2018 · 7 min read

Original Project Mission: Find interesting insights to see where the remodeling market is headed

Project Mission (Twist): How to handle well manicured excel data in Python (spoiler: neat is a deceptive word)

Timeline: One week (I tell you if you're new to DS like me, one week is **not** enough)

Project Findings for the Original Goal :

- The numbers look stable with some reshuffle in priorities
- Average costs are considerably higher than national average in/around Palo Alto, CA
- States where the biggest Full Services remodeling companies are located

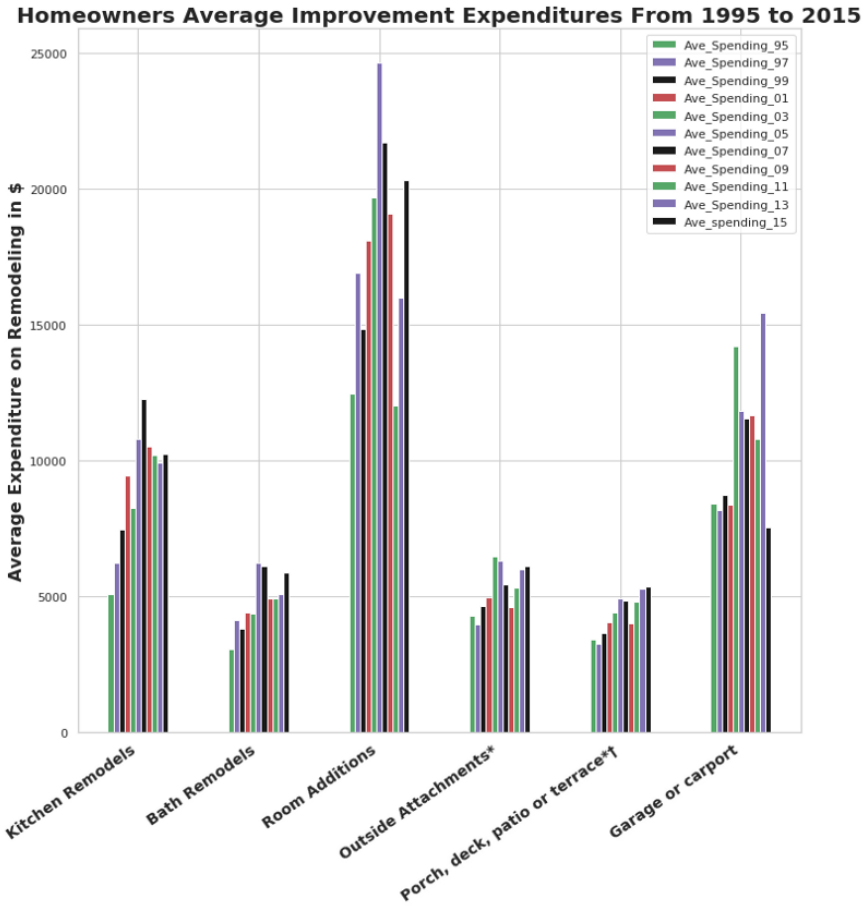
Finding on Data Handling: There is a solution in Pandas (Python), life is not hopeless yet!

Data: Exploring publicly available data from different sources—Joint Center for Housing Studies at Harvard, Remodeling magazine, and obtaining average project costs for local area in and around Palo Alto, CA from Hinkle Construction Inc's web page.

Method: Python—data wrangling, basic statistics and visualization using mainly Pandas, Matplotlib, Seaborn.

The sequence for this post will be to first talk about the results we got from the home remodeling data. Later, we will go through the data cleaning struggles as well.

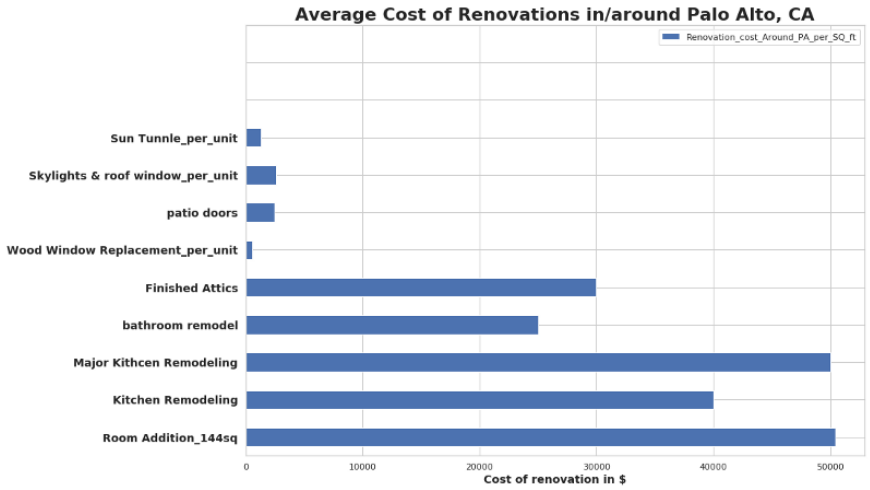
The graphs below show national average spending on different categories of home improvements. The data shows that 'Room Additions' is the most popular category. In last 15 years, the spending in this category peaked in 2005 at ~$25k in average expenditure. The lowest point was in 2011 with spending going down to below $12k. The market seems to be recovering since then. From 1995–2005, homeowners spending on room additions increased by more than 62% in 20 years. Another relatively bigger change is in kitchen remodeling where the expenditure doubled in from 1995 to 2015 and peaking in 2007.



Source: Joint Center for Housing Studies at Harvard

When I got the above results, I was curious to know the average homeowner expenditures in my local area where I live. Given the time constrain, I couldn't find the data directly comparable to the graph above but the average of the those projects in our area.

The bar graph below depicts the average costs of different types of home improvements in and around Palo Alto, California. The costs are significantly higher than the national averages. Adding a new room, 144 square feet in size, costs more than $50k. Major kitchen remodeling is also equally expensive at $50k.



Source: Hinkle Construction Inc' website

For the first two plots, please check the attached python file, if interested.

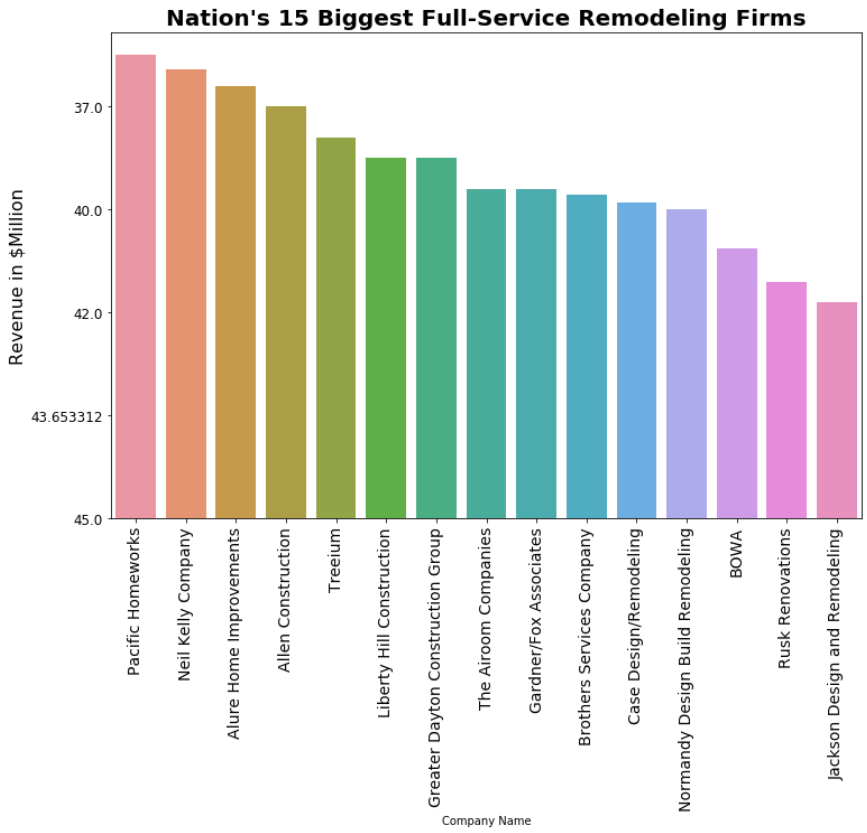
```
In [214]: #lets import pandas and other libraries
          import pandas as pd
          %matplotlib inline
          import matplotlib.pyplot as plt
          import numpy as np
          import seaborn as sns

In [215]: #create a dataframe of average cost of
           rows to Median_sale_price_Palo_Alto.cs
          #http://hinkle-construction.com/project
          df = pd.DataFrame(
                  {"Renovation_categories_Palo_Al
          4sq', 'Kitchen Remodeling', 'Major Kith

          , 'Finished Attics', 'Wood Window Repla
          o doors',
```

I was compelled to find out who were the biggest players in the remodeling industry. 'Remodeling Magazine' conducts intensive research in this field. The Remodeling 550 division of the magazine collects the data and announces the winners in remodeling business based on the previous years revenues. In their own words: "*The 2017 Remodeling 550 comes in four lists, each represented in a tab below and showing who's who among full-service remodelers, replacement contractors, insurance restoration companies, and franchisors. Companies*

are ranked based on their 2016 revenue generated from remodeling; not by the company's gross revenue."

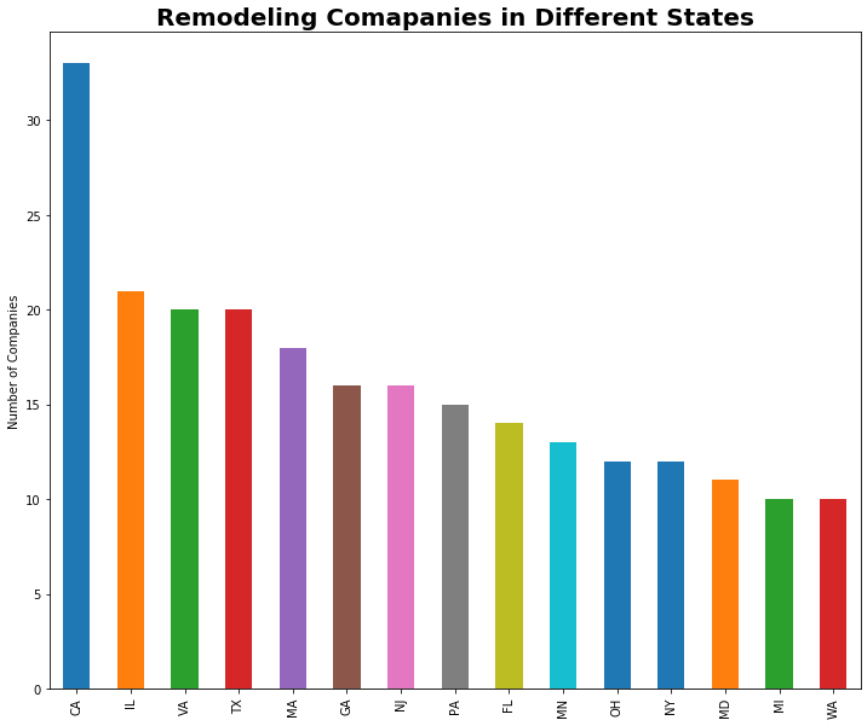
I only looked into 2017's Full service companies. The ranking was based revenues earned in 2016 (the list for 2018 is not out until May, 2019).

The bar chart shows 15 most prominent companies (out of total 340) in remodeling business across the country.



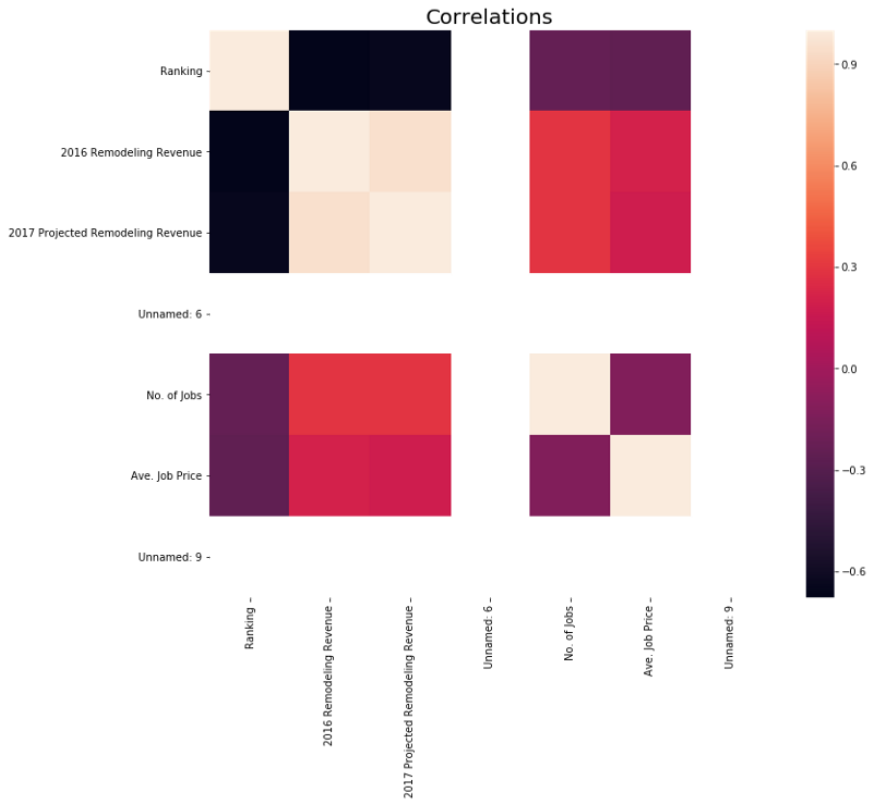
Source: Remodeling magazine

Unsurprisingly, a good number of those companies are based in California.



Source: Remodeling magazine

The last plot is just for fun. I ran a correlation matrix to see how revenues, ranking and other variables are related to each other. Shades show the intensity of the correlation coefficients. For example: ranking ('#') and revenues are highly negatively correlated. Don't get confused here. We all know how ranking works: lower the number, better the ranking (1 is the best). So, higher the revenue, lower the number (say 1) which means higher the ranking.

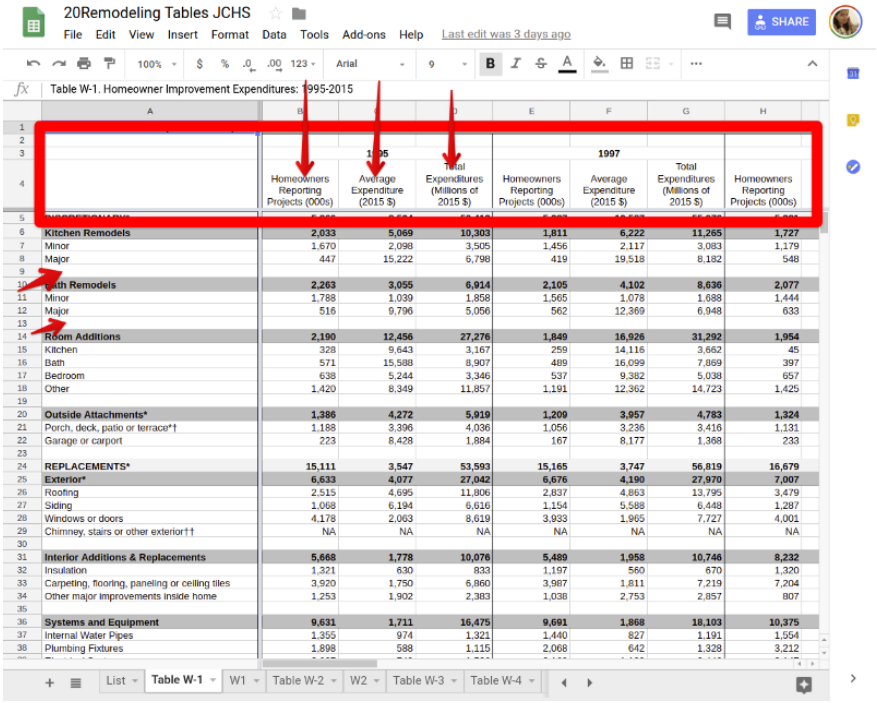


Below is my work in Jupyter Notebook for the above plots.

Now, let's talk about the data wrangling/cleaning struggle part as promised. Some background on the data I collected for this project:

- I downloaded a lot more data then I actually need/could process in time
- I though more data would always mean better results
- Nice, neat looking data in excel would mean less cleaning in Pandas/Python. WRONG! Mistake no. 1.
- My data was spread over 26 Excel sheets in 4 different files(I'm not kidding)

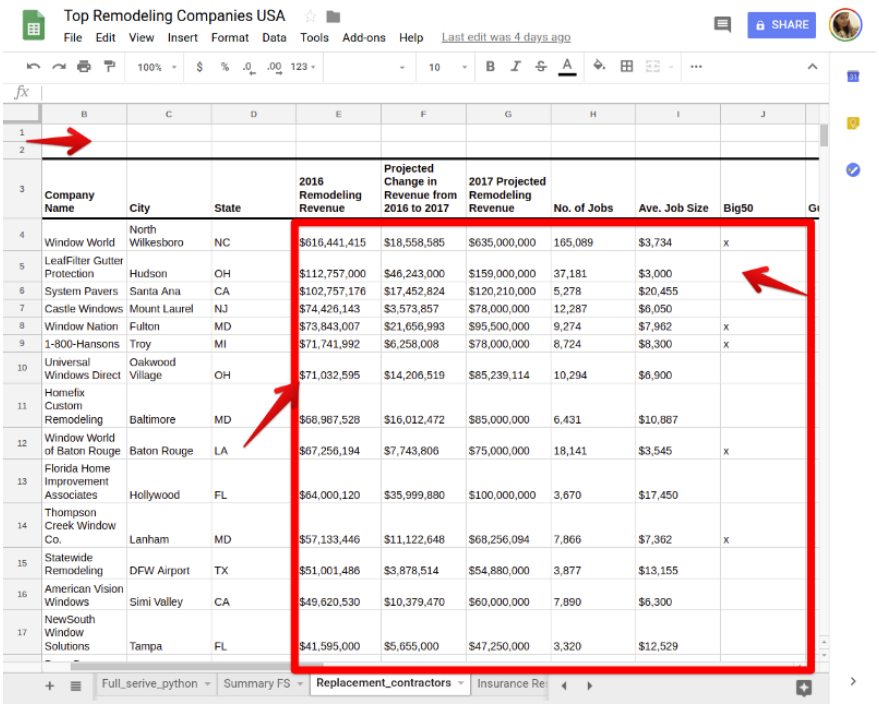
Problems in a neat looking Excel sheet: These are ones I faced, so beware, there might be even more I haven't come across yet.



Impressively neat excel data

- Empty cells: Excel is more flexible when it comes to leaving cell empty but your Pandas read function isn't going to like it. It's laughing at you like "Told ya, Deal with the NaNs now!"
- The header: A clean header in excel doesn't necessarily mean anything in Pandas. You check the actual columns names and don't be surprised that the actual column names were way different than what's being shown in the excel header! Change column names. More the number of columns more the work.

- Another similar problem when you have column names for all columns except for the first one (in excel it will still make sense specially if it's a categorical variable column) but operating on such a column is a problem in Pandas framework.



Company Name	City	State	2016 Remodeling Revenue	Projected Change in Revenue from 2016 to 2017	2017 Projected Remodeling Revenue	No. of Jobs	Ave. Job Size	Big50
Window World	North Wilkesboro	NC	$616,441,415	$18,558,585	$635,000,000	165,089	$3,734	x
LeafFilter Gutter Protection	Hudson	OH	$112,757,000	$46,243,000	$159,000,000	37,181	$3,000	
System Pavers	Santa Ana	CA	$102,757,176	$17,452,824	$120,210,000	5,278	$20,455	
Castle Windows	Mount Laurel	NJ	$74,426,143	$3,573,857	$78,000,000	12,287	$6,050	
Window Nation	Fulton	MD	$73,843,007	$21,656,993	$95,500,000	9,274	$7,962	x
1-800-Hansons	Troy	MI	$71,741,992	$6,258,008	$78,000,000	8,724	$8,300	x
Universal Windows Direct	Oakwood Village	OH	$71,032,595	$14,206,519	$85,239,114	10,294	$6,900	
Homefix Custom Remodeling	Baltimore	MD	$68,987,528	$16,012,472	$85,000,000	6,431	$10,887	
Window World of Baton Rouge	Baton Rouge	LA	$67,256,194	$7,743,806	$75,000,000	18,141	$3,545	x
Florida Home Improvement Associates	Hollywood	FL	$64,000,120	$35,999,880	$100,000,000	3,670	$17,450	
Thompson Creek Window Co.	Lanham	MD	$57,133,446	$11,122,648	$68,256,094	7,866	$7,362	x
Statewide Remodeling	DFW Airport	TX	$51,001,486	$3,878,514	$54,880,000	3,877	$13,155	
American Vision Windows	Simi Valley	CA	$49,620,530	$10,379,470	$60,000,000	7,890	$6,300	
NewSouth Window Solutions	Tampa	FL	$41,595,000	$5,655,000	$47,250,000	3,320	$12,529	

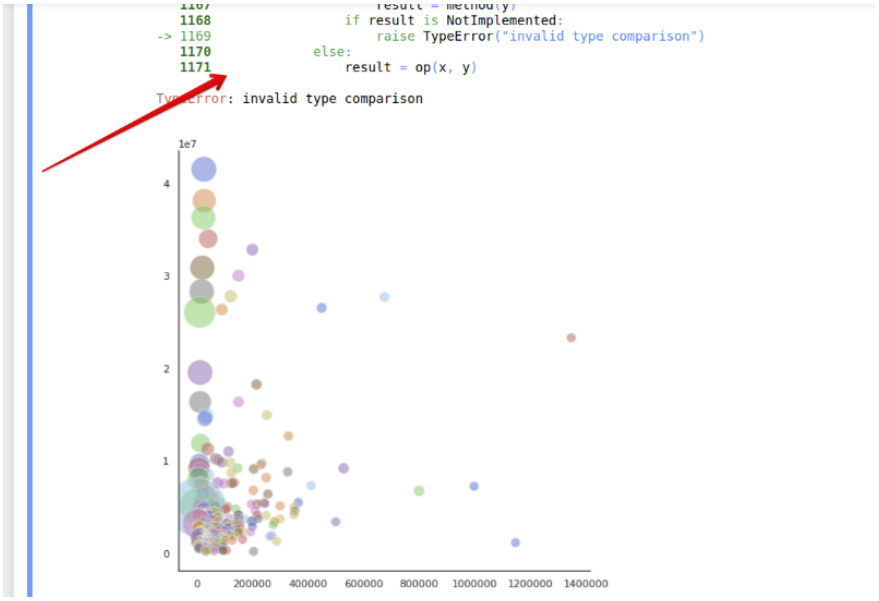
- Ah! those neat looking numeric values! Don't be fooled. You will have deal with Every. Single. Extra. Thing. you have in between the numbers. Be it a dollar ($) sign, comma etc. Pandas reads it as an object and that a problem. You would not be able to achieve any statistical/mathematical goal with that. You have to convert it to int/float to make it usable.

What I Learned from this Project:

- Think about the timeline and don't be extra ambitious. Choose a doable data set. Instead of choosing multiple data sets, choose one and think about how many ways you can manipulate/analyze it.

- Prioritize your work around the project deliverable. When you are stuck in my situation of lots of fragmented data, re-prioritize your objectives and follow inverted pyramid principle: do the work that's most important first.

- You don't always have to convert your excel files into .csv. There is an easier way to import the desired sheet directly. All it needs is a working path name.
- Ask for help. As a beginner in Python, I'm struggling more often than I would like to. But asking your peers/instructors/managers for help might provide you valuable insights and multiple solutions for the single problem.
- When you are stuck, it feels easier to go back and fix the problems in excel than to spend time figuring it out in Python. But, I tell you, it's rewarding to not give and get it done the right way in Pandas.
- You need to know what visualizations are appropriate for your data. There are cool visualization libraries in Python but none of the cool graphs work for your data and that's ok. When see big errors, don't ignore it.



In the graph above, I'm trying to plot all 340 remodeling companies, their revenues, no. of jobs, and average job price. It's invalid comparison type. respect the errors and dig deeper to know what kind of plot is more suitable for your data representation. Simple don't mean bad as fancy doesn't always mean better (ok, may be grapes are sour in this case as I couldn't get a fancies plot, but I mean what I say).

For me, this first project was a big learning curve. But I came out as a stronger, better prepared Data Scientist in training who will never look

at excel data sheets in the same light ever again. :-)