# Data Wrangling Report for Building Data

A Part of the Capstone Project Submissions

## Introduction

A building permit is an official approval document issued by a governmental agency that allows you or your contractor to proceed with a construction or remodeling project on one's property. For more details click here. Each city or county has its own office related to buildings, that can do multiple functions like issuing permits, inspecting buildings to enforce safety measures, modifying rules to accommodate needs of the growing population etc. For the city of San Francisco, permit issuing is taken care by Permit Services wing of Department of Building Inspection (henceforth called DBI). The delays in permit issuance pose serious problems to construction industries and later on real estate agencies. Read this. Trulia study and Vancouver city article.

### This Document

This document particularly describes the data wrangling steps that I undertook to clean the San Francisco building permit data set. It explains what are tricky parts of data wrangling phase in general. What kind of cleaning steps were performed on this particular set, how the missing values or the outliers handled. The code for data wrangling is here

## Data retrieval

Data used to get the results explained in next sections is available in San Francisco city open data portal. It is updated every Saturday.

Step by step process to download:

- Go to the link: SF data portal.
- Click on Filter and "Add a Filter Condition". A drop down menu appears.
- Select, "Filed Date" and "is after".
- Enter date as 12/31/2012, because I wanted to do analysis of last 4-5 years. I think most recent data is important in matters such as this, the city council policies could

change, there might be new rules, new employers to expedite process etc. Old data may not be too useful in modeling.

CSV format is chosen because it is less than 100MB size and easy to load into notebook. There are other methods like downloading json, or using socrata. This is found to be more reliable and less dependent on any extra libraries.

**Date of download for this analysis:** The file as of Feb 25, 2018 (Sunday) has been downloaded and kept locally for easy access. Size is about 75MB. The results of this analysis can be reproduced only if one more filter is used in the second step above, to select "Filed Date" "is before" and put Feb 26th, 2018.

**List of attributes in the data:**

```
RangeIndex: 198900 entries, 0 to 198899
Data columns (total 43 columns):
Permit Number                                198900 non-null object
Permit Type                                  198900 non-null int64
Permit Type Definition                       198900 non-null object
Permit Creation Date                         198900 non-null object
Block                                        198900 non-null object
Lot                                          198900 non-null object
Street Number                                198900 non-null int64
Street Number Suffix                         2216 non-null object
Street Name                                  198900 non-null object
Street Suffix                                196132 non-null object
Unit                                         29479 non-null float64
Unit Suffix                                  1961 non-null object
Description                                  198610 non-null object
Current Status                               198900 non-null object
Current Status Date                          198900 non-null object
Filed Date                                   198900 non-null object
Issued Date                                  183960 non-null object
Completed Date                               97191 non-null object
First Construction Document Date             183954 non-null object
Structural Notification                      6922 non-null object
Number of Existing Stories                   156116 non-null float64
Number of Proposed Stories                   156032 non-null float64
Voluntary Soft-Story Retrofit                35 non-null object
Fire Only Permit                             18827 non-null object
Permit Expiration Date                       147020 non-null object
Estimated Cost                               160834 non-null float64
Revised Cost                                 192834 non-null float64
Existing Use                                 157786 non-null object
Existing Units                               147362 non-null float64
Proposed Use                                 156461 non-null object
Proposed Units                               147989 non-null float64
Plansets                                     161591 non-null float64
TIDF Compliance                              2 non-null object
Existing Construction Type                   155534 non-null float64
Existing Construction Type Description       155534 non-null object
Proposed Construction Type                   155738 non-null float64
Proposed Construction Type Description       155738 non-null object
Site Permit                                  5359 non-null object
Supervisor District                          197183 non-null float64
Neighborhoods - Analysis Boundaries          197175 non-null object
Zipcode                                      197184 non-null float64
Location                                     197200 non-null object
Record ID                                    198900 non-null int64
dtypes: float64(12), int64(3), object(28)
memory usage: 65.3+ MB
```

## Tricky Part!

The Tricky part of data wrangling,

1. Knowing what is present in the 'null' cells, is it NaN or simply ' ' or whitespace(s)
2. In the non-null cells, if the all values are meaningful.
3. Recognizing that even if a column has all non-null and meaningful values, the future updates to the column may have problems. Hence need to expect it and handle it
4. See the data and think if certain value make sense for the business and decide to drop those which are not relevant
5. Data Imputation: Filling null cells with non-trivial substitutes-values other than mean, mode or median of that column
6. Ensuring not to introduce bias while imputation or dropping.

## Cleaning Steps Performed:

1. **Identifying the columns to retain:** It is not smart to waste time cleaning up columns that may not turn out useful in modeling. Hence I chose to eliminate some in the first pass, exercising caution.
2. **Identifying the rows to retain**: Current Status had 14 types, of which withdrawn, cancelled and disapproved status and not having issue dates are not relevant for further study. Hence rows corresponding to these records are dropped. This was about 2k in total. After doing this Current Status column is eliminated. Records corresponding to no location also had to be dropped, as it made no sense in the next stages.
3. **Checking permit number,type and file dates:** These three variables are very essential and should be present in all the records.
4. **Converting Invalid weekdays:** DBI is open only from Monday-Friday. I attributed dates corresponding to weekends to typing mistake. Converted them to the nearer valid date. This may not be accurate (as typo may not be from n to n+1 or n-1 in day part alone), however it will at least prevent outliers in the EDA part.
5. **Filling the blanks in variables having only 'Y' as the entries:** In the application forms (both physical or online), normally the applicant is supposed to tick the option if applicable. Otherwise nothing needs to be done. Hence it is understandable that

blanks mean not applicable, a "No". Filled Fire only permit, Site Permit and Structural notification blank entries with “N”

6. **Filling nulls in qualitative variables with > 2 categories, which has numbers as entries:** 5 variables, existing construction type, proposed construction type, existing number of stories, proposed number of stories and plansets have NaN’s. 18-20% of the records are NaN’s in all the 4-5 variables. A few thousands of records have any one to 3 of the 4 variables with NaNs. It is best to leave them as they are, since it is building permit application. NaN corresponds to “not applicable” option for that permit type.
7. **Filling  nulls in qualitative variables with > 2 categories, which has strings as entries:** Replaced existing use, proposed use variables with a more explicit “Unknown” instead of general NaN.
8. **Converting Location (Longitude / Latitude) into numbers:** NaN’s in this variables are substituted with 0’s. They are the same rows for which supervisor district and neighborhoods are unknown. We might end up retaining only only one of these three in the next stages. Longitude and Latitude were read as string types, so converted them to float numpy array of 2 elements.
9. **Interpreting too small values for costs:** Cost of the project is an essential part of the application according to this post. Still, this dataset does not follow that rule. There seem to be invalid entries.  Cost columns are retained even if they are too small, because there are about 30% of them less than 500$, and median cost is found to be 7500$. NaNs are replaced with zeros and retained. Imputation can be done with EDA after understanding the data in detail.
10. **Dealing with rows without issue date:** The rows that do not have valid entries for issue date will yield NaN when we evaluate the wait time. But then they can not be dropped out of the EDA because that would lead to *selection bias* and give the perception of having less wait time in an average sense and would give less value for maximum wait times. We can not assume they were all issued today, because this would still introduce the bias. The only option seems to be leaving it as is and handle it during next stage. Perhaps, some recently filed ones can go into test set to simulate real life scenario.

11. **Computing time taken variable:** The time taken variable was computed by subtracting datetime object column file date from date time object column issue date. This is the variable that we want to analyze, model and predict.
12. **Saving the Clean file:** After all the cleaning steps, I wrote the dataframe to csv file so that the next step can pick it from there if needed.

| Author | Revision History<br>Version / Date | Comments |
|---|---|---|
| Aparna Shastry | 0.1/02.25.2018 | Initial Draft |
| | 1.0/0.3.05.2018 | 1.One more step of "retaining rows" added<br>2.Added data download details<br>3. Screenshot of attributes added. |