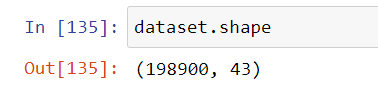
1. **Description of the dataset**

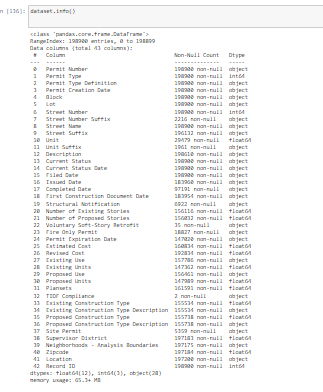
This data set pertains to all types of structural permits from Jan 1, 2013-Feb 25th 2018. Data includes details on application/permit numbers, job addresses, supervisorial districts, and the current status of the applications (Dataset URL: https://www.kaggle.com/code/celesteell/exercise-handling-missing-values/data?select=Building\_Permits.csv).

Before deal with missing values, first we need to identify the dataset.

**1.1 Shape**



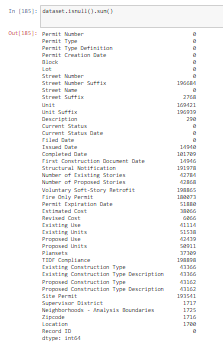
**1.2 Data types**



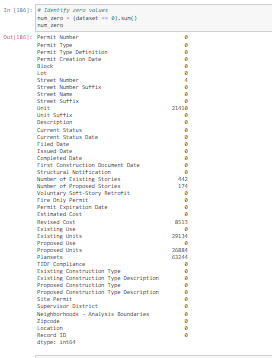
* 1. **Identify data**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Notes** |
| Permit Number | Number assigned while filing |  |
| Permit Type | Type of the permit represented numerically. |  |
| Permit Type Definition | Description of the Permit type, for example  new construction, alterations |  |
| Permit Creation Date | Date on which permit created, later than  or same as filing date |  |
| Block | Related to address | Cannot Predict/replace/Non Zero |
| Lot | Related to address | Cannot Predict/replace/Non Zero |
| Street Number | Related to address | Cannot Predict/replace/Non Zero |
| Street Number Suffix | Related to address | Cannot Predict/replace/Non Zero |
| Street Name | Related to address | Cannot Predict/replace |
| Street Name Suffix | Related to address | Cannot Predict/replace |
| Unit | Unit of a building |  |
| Unit suffix | Suffix if any, for the unit | Cannot Predict/replace/Non Zero |
| Description | Details about purpose of the permit.  Example: reroofing, bathroom renovation | Cannot Predict/replace/Non Zero |
| Current Status | Current status of the permit application. | Cannot Predict/replace/Non Zero |
| Current Status Date | Date at which current status was entered | Cannot Predict/replace/Non Zero |
| Filed Date | Filed date for the permit | Cannot Predict/replace/Non Zero |
| Issued Date | Issued date for the permit | Cannot Predict/replace/Non Zero |
| Completed Date | The date on which project was completed,  applicable if Current Status = “completed” | Cannot Predict/replace/Non Zero |
| First Construction Document Date | Date on which construction was documented | Cannot Predict/replace/Non Zero |
| Structural Notification | Notification to meet some legal need, given or not | Yes or No |
| Number of Existing Stories | Number of existing stories in the building.  Not applicable for certain permit types | Numerical |
| Number of Proposed Stories | Number of proposed stories for the construction/alteration | Numerical |
| Voluntary Soft-Story  Retrofit | Soft story to meet earth quake regulations | Yes or No |
| Fire Only Permit | Fire hazard prevention related permit | Yes or No |
| Permit Expiration Date | Expiration date related to issued permit. | Cannot Predict/replace/Non Zero |
| Estimated Cost | Initial estimation of the cost of the project | Numerical |
| Revised Cost | Revised estimation of the cost of the project | Numerical |
| Existing Use | Existing use of the building | Non Zero |
| Existing Units | Existing number of units | Numerical |
| Proposed Use | Proposed use of the building | Non Zero |
| Proposed Units | Proposed number of units | Numerical |
| Plansets | Plan representation indicating the general design  intent of the foundation.. | Numerical |
| TIDF Compliance | TIDF compliant or not, this is a new legal requirement | Yes or No |
| Existing Construction Type | Construction type, existing as categories  represented numerically | Numerical (Identification Number) |
| Existing Construction Type Description | Description of the above, for example,  wood or other construction types | Non Zero |
| Proposed Construction Type | Construction type, proposed, as categories  represented numerically | Numerical (Identification Number) |
| Proposed Construction Type Description | Description of the above | Non Zero |
| Site Permit | Permit for site | Yes or No |
| Supervisor District | Supervisor District to which the building location belongs to | Cannot Predict/replace/Non Zero |
| Neighborhoods - Analysis Boundaries | Neighborhood to which the building location belongs to | Cannot Predict/replace/Non Zero |
| Zipcode | Zipcode of building address | Cannot Predict/replace/Non Zero |
| Location | Location in latitude, longitude pair. | Cannot Predict/replace/Non Zero |
| Record ID | Some ID, not useful for this |  |

**1.4 Identify Null values**

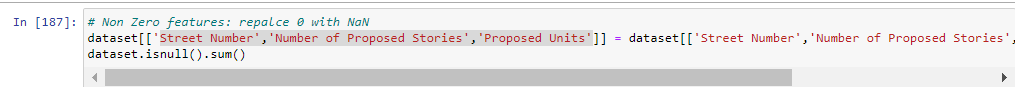


**1.5 Zeros**

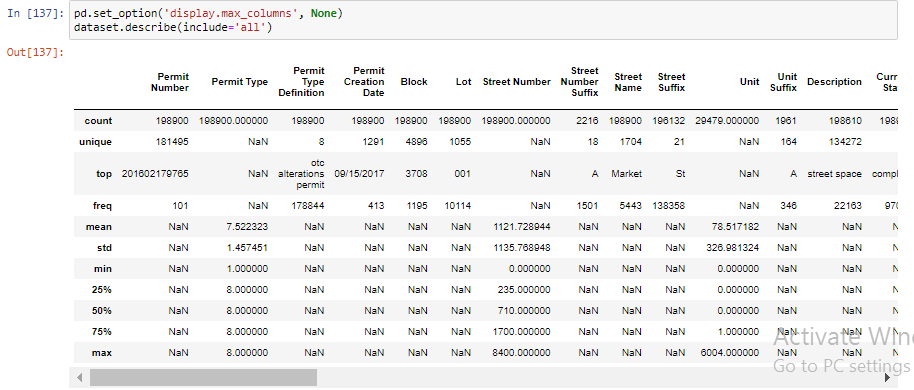


Following none zero futures are replaces with NaN.

* Street Number
* Number of Proposed Stories
* Proposed Units



* 1. **Important figures of the dataset**

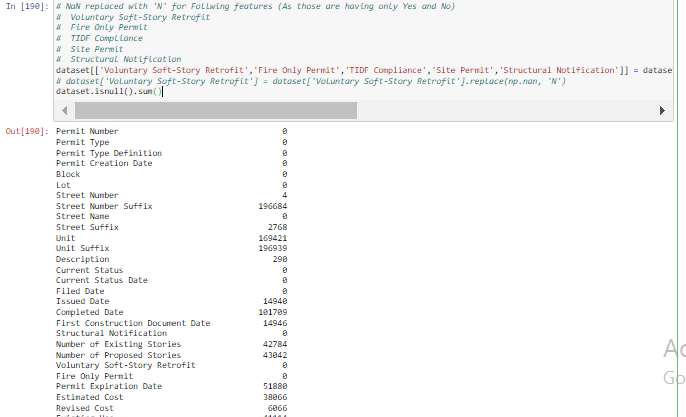


1. **Handling Missing Values**

**2.1 Imputation**

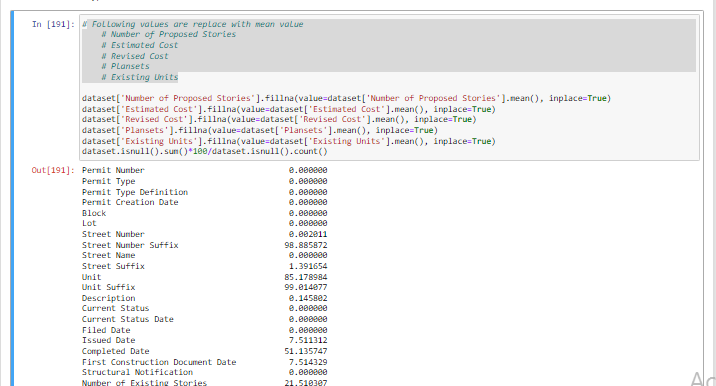
Following columns should have ‘Yes’ or ‘No’ values. Therefore null values are replaced with 'N'.

* Voluntary Soft-Story Retrofit
* Fire Only Permit
* TIDF Compliance
* Site Permit
* Structural Notification



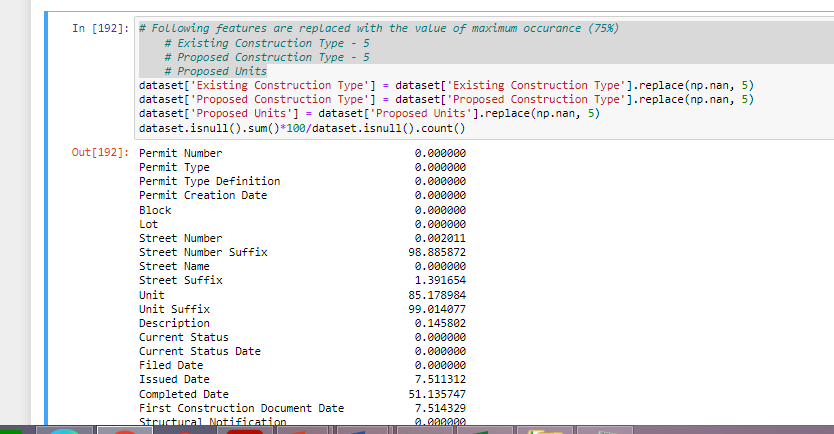
Null values of following columns are replace with mean value (As they have numerical values)

* Number of Proposed Stories
* Estimated Cost
* Revised Cost
* Plansets
* Existing Units



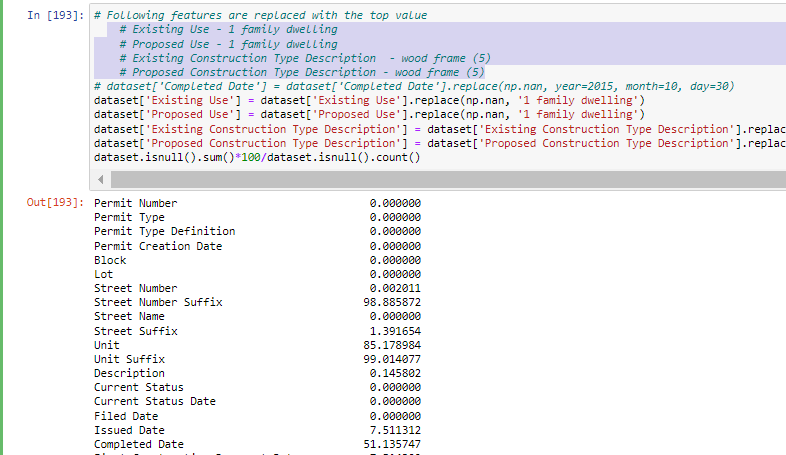
Null values in following features are replaced with the value of maximum occurrence (75%).

* Existing Construction Type
* Proposed Construction Type
* Proposed Units



Null values in following features are replaced with the value of top most occurrence.

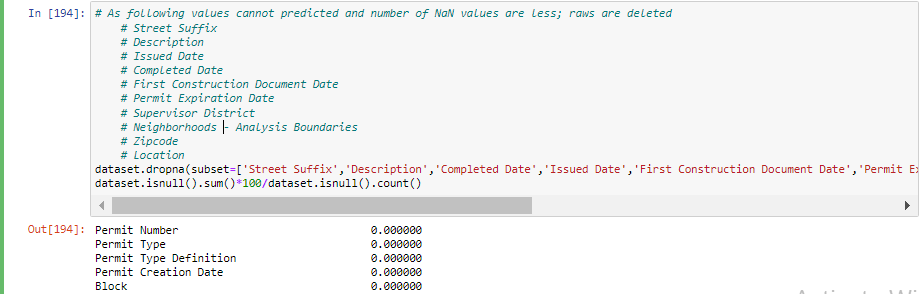
* Existing Use
* Proposed Use
* Existing Construction Type Description
* Proposed Construction Type Description



**2.2 Omission**

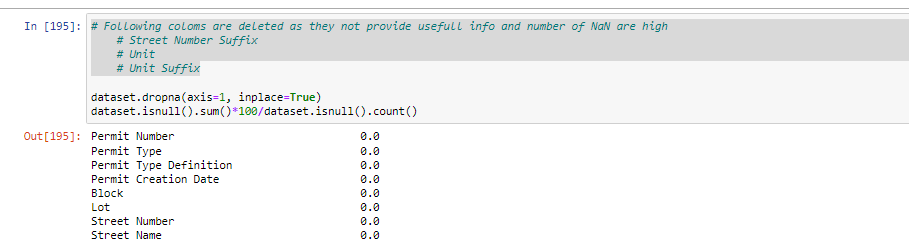
As following values cannot predicted and number of Null values are less (except ‘Completed Date’; Rows are deleted belongs to following columns.

* Street Suffix
* Description
* Issued Date
* Completed Date
* First Construction Document Date
* Permit Expiration Date
* Supervisor District
* Neighborhoods - Analysis Boundaries
* Zipcode
* Location



Following columns are deleted as the number of Null values are high.

* Street Number Suffix
* Unit
* Unit Suffix



Number of rows and columns after deletion as shown below;



1. **Conclusion**

After observing the dataset, high number of missing values were observed in some features. Therefore, those features were deleted. Further, the imputations of data to the Null values were done where possible. However after handling the missing values number of rows are decreased in considerably.