Data Analysis Y3 -Week 4 Lab- Data Preprocessing

- In this lab you will be applying what we covered in our Tuesday's class in pre-processing data including: Handling missing value.
- We'll be using a dataset for bulding permits in the US. You can downlaod the datset from the Moodel page.

1. Import pandas and numpy, and read in the Buliding permits csv file and set it to a DataFrame called "permits".

```
import pandas as pd
import numpy as np
```

```
# Read the CSV file into a DataFrame
permits = pd.read_csv("/content/Building_Permits (1).csv")
```

```
<ipython-input-60-89fe5d8251a7>:2: DtypeWarning: Columns (22,32) have mixed types. Specify dtype option on import or set low_memory
  permits = pd.read_csv("/content/Building_Permits (1).csv")
```

2. Check the head of the first 10 rows in the DataFrame

```
permits.head(10)
```

	Permit Number	Permit Type	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Number Suffix	Street Name	Street Suffix	...	Existing Construction Type	Existing Construction Type Description	Proposed Construction Type
0	2.02E+11	4	sign - erect	05/06/2015	326	23	140	NaN	Ellis	St	...	3.0	constr type 3	NaN
1	2.02E+11	4	sign - erect	04/19/2016	306	7	440	NaN	Geary	St	...	3.0	constr type 3	NaN
2	2.02E+11	3	additions alterations or repairs	05/27/2016	595	203	1647	NaN	Pacific	Av	...	1.0	constr type 1	1.0
3	2.02E+11	8	otc alterations permit	11/07/2016	156	11	1230	NaN	Pacific	Av	...	5.0	wood frame (5)	5.0
4	2.02E+11	6	demolitions	11/28/2016	342	1	950	NaN	Market	St	...	3.0	constr type 3	NaN
5	2.02E+11	8	otc alterations permit	06/14/2017	4105	9	800	NaN	Indiana	St	...	1.0	constr type 1	1.0
6	2.02E+11	8	otc alterations permit	06/30/2017	1739	20	1291	NaN	11th	Av	...	5.0	wood frame (5)	5.0
7	M803667	8	otc alterations permit	06/30/2017	4789	14	1465	NaN	Revere	Av	...	NaN	NaN	NaN
8	M804227	8	otc alterations permit	07/05/2017	1212	54	2094	NaN	Fell	St	...	NaN	NaN	NaN
9	M804767	8	otc alterations permit	07/06/2017	1259	16	89	NaN	Alpine	Tr	...	NaN	NaN	NaN

10 rows × 43 columns

3. Change the Datatype of the following columns into appropriate datatype:

1. Permit type
2. Permit Creation Date
3. Issued Date

```
permits['Permit Type'] = permits['Permit Type'].astype('int')  # Change to integer datatype
permits
# permit.info()
```

	Permit Number	Permit Type	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Number Suffix	Street Name	Street Suffix	...	Existing Construction Type	Existing Construction Type Description	Con
0	2.02E+11	4	sign - erect	2015-05-06	326	23	140	NaN	Ellis	St	...	3.0	constr type 3	
1	2.02E+11	4	sign - erect	2016-04-19	306	7	440	NaN	Geary	St	...	3.0	constr type 3	
2	2.02E+11	3	additions alterations or repairs	2016-05-27	595	203	1647	NaN	Pacific	Av	...	1.0	constr type 1	
3	2.02E+11	8	otc alterations permit	2016-11-07	156	11	1230	NaN	Pacific	Av	...	5.0	wood frame (5)	
4	2.02E+11	6	demolitions	2016-11-28	342	1	950	NaN	Market	St	...	3.0	constr type 3	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
168317	M862628	8	otc alterations permit	2017-12-05	113	017A	1228	NaN	Montgomery	St	...	NaN	NaN	
168318	2.02E+11	8	otc alterations permit	2017-12-05	271	14	580	NaN	Bush	St	...	5.0	wood frame (5)	
168319	M863507	8	otc alterations permit	2017-12-06	4318	19	1568	NaN	Indiana	St	...	NaN	NaN	
168320	M863747	8	otc alterations permit	2017-12-06	298	29	795	NaN	Sutter	St	...	NaN	NaN	
168321	M864287	8	otc alterations permit	2017-12-07	160	6	838	NaN	Pacific	Av	...	NaN	NaN	

168322 rows × 43 columns

```
permits['Permit Creation Date'] = pd.to_datetime(permits['Permit Creation Date'])  # Change to datetime

print(permits)
```

```
     168318        Bush           St  ...                      5.0
     168319     Indiana           St  ...                      NaN
     168320      Sutter           St  ...                      NaN
     168321     Pacific           Av  ...                      NaN
```

```
168317                                  NaN      NaN
168318                                  NaN      NaN
168319                                  NaN      NaN
168320                                  NaN      NaN
168321                                  NaN      NaN

                                       Location     Record ID
0       (37.785719256680785, -122.40852313194863)  1.380000e+12
1        (37.78733980600732, -122.41063199757738)  1.420000e+12
2         (37.7946573324287, -122.42232562979227)  1.420000e+12
3        (37.79595867909168, -122.41557405519474)  1.440000e+12
4        (37.78315261897309, -122.40950883997789)  1.450000e+11
...                                           ...           ...
168317                                       NaN  1.490000e+12
168318                                       NaN  1.490000e+12
168319                                       NaN  1.490000e+12
168320                                       NaN  1.490000e+12
168321                                       NaN  1.490000e+12

[168322 rows x 43 columns]
```

```
permits['Issued Date'] = pd.to_datetime(permits['Issued Date'])  # Change to datetim
```

```
permits
```

	Permit Number	Permit Type	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Number Suffix	Street Name	Street Suffix	...	Existing Construction Type	Existing Construction Type Description	Con
0	2.02E+11	4	sign - erect	2015-05-06	326	23	140	NaN	Ellis	St	...	3.0	constr type 3	
1	2.02E+11	4	sign - erect	2016-04-19	306	7	440	NaN	Geary	St	...	3.0	constr type 3	
2	2.02E+11	3	additions alterations or repairs	2016-05-27	595	203	1647	NaN	Pacific	Av	...	1.0	constr type 1	
3	2.02E+11	8	otc alterations permit	2016-11-07	156	11	1230	NaN	Pacific	Av	...	5.0	wood frame (5)	
4	2.02E+11	6	demolitions	2016-11-28	342	1	950	NaN	Market	St	...	3.0	constr type 3	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
168317	M862628	8	otc alterations permit	2017-12-05	113	017A	1228	NaN	Montgomery	St	...	NaN	NaN	
168318	2.02E+11	8	otc alterations permit	2017-12-05	271	14	580	NaN	Bush	St	...	5.0	wood frame (5)	
168319	M863507	8	otc alterations permit	2017-12-06	4318	19	1568	NaN	Indiana	St	...	NaN	NaN	
168320	M863747	8	otc alterations permit	2017-12-06	298	29	795	NaN	Sutter	St	...	NaN	NaN	
168321	M864287	8	otc alterations permit	2017-12-07	160	6	838	NaN	Pacific	Av	...	NaN	NaN	

168322 rows × 43 columns

```
permits.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 168322 entries, 0 to 168321
Data columns (total 43 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   Permit Number                         168322 non-null  object
 1   Permit Type                           168322 non-null  int64
 2   Permit Type Definition                168322 non-null  object
 3   Permit Creation Date                  168322 non-null  datetime64[ns]
 4   Block                                 168322 non-null  object
 5   Lot                                   168322 non-null  object
 6   Street Number                         168322 non-null  int64
 7   Street Number Suffix                  1849 non-null    object
 8   Street Name                           168322 non-null  object
 9   Street Suffix                         165996 non-null  object
 10  Unit                                  24833 non-null   float64
```

```
 11  Unit Suffix                           1687 non-null    object
 12  Description                           168088 non-null  object
 13  Current Status                        168322 non-null  object
 14  Current Status Date                   168322 non-null  object
 15  Filed Date                            168322 non-null  object
 16  Issued Date                           155332 non-null  datetime64[ns]
 17  Completed Date                        81890 non-null   object
 18  First Construction Document Date      155339 non-null  object
 19  Structural Notification               5915 non-null    object
 20  Number of Existing Stories            132426 non-null  float64
 21  Number of Proposed Stories            132257 non-null  float64
 22  Voluntary Soft-Story Retrofit         25 non-null      object
 23  Fire Only Permit                      15720 non-null   object
 24  Permit Expiration Date                124400 non-null  object
 25  Estimated Cost                        136429 non-null  float64
 26  Revised Cost                          163115 non-null  float64
 27  Existing Use                          133815 non-null  object
 28  Existing Units                        125139 non-null  float64
 29  Proposed Use                          132637 non-null  object
 30  Proposed Units                        125602 non-null  float64
 31  Plansets                              137096 non-null  float64
 32  TIDF Compliance                       1 non-null       object
 33  Existing Construction Type            131963 non-null  float64
 34  Existing Construction Type Description  131963 non-null  object
 35  Proposed Construction Type            132011 non-null  float64
 36  Proposed Construction Type Description  132011 non-null  object
 37  Site Permit                           4363 non-null    object
 38  Supervisor District                   166878 non-null  float64
 39  Neighborhoods - Analysis Boundaries   166872 non-null  object
 40  Zipcode                               166879 non-null  float64
 41  Location                              166892 non-null  object
 42  Record ID                             168322 non-null  float64
dtypes: datetime64[ns](2), float64(13), int64(2), object(26)
memory usage: 55.2+ MB
```

4. How many total missing values in each colum

```
missing_values = permits.isnull().sum()
print(missing_values)
```

```
Permit Number                             0
Permit Type                               0
Permit Type Definition                    0
Permit Creation Date                      0
Block                                     0
Lot                                       0
Street Number                             0
Street Number Suffix                 166473
Street Name                               0
Street Suffix                          2326
Unit                                 143489
Unit Suffix                          166635
Description                             234
Current Status                            0
Current Status Date                       0
Filed Date                                0
Issued Date                           12990
Completed Date                        86432
First Construction Document Date      12983
Structural Notification              162407
Number of Existing Stories            35896
Number of Proposed Stories            36065
Voluntary Soft-Story Retrofit        168297
Fire Only Permit                     152602
Permit Expiration Date                43922
Estimated Cost                        31893
Revised Cost                           5207
Existing Use                          34507
Existing Units                        43183
Proposed Use                          35685
Proposed Units                        42720
Plansets                              31226
TIDF Compliance                      168321
Existing Construction Type            36359
Existing Construction Type Description    36359
Proposed Construction Type            36311
Proposed Construction Type Description    36311
Site Permit                          163959
Supervisor District                    1444
Neighborhoods - Analysis Boundaries    1450
Zipcode                                1443
Location                               1430
Record ID                                 0
dtype: int64
```

```
missing_values = permits.isnull().sum()

missing_columns = missing_columns[missing_columns > 0]

print(missing_values.round(1))
```

```
Permit Number                                 0
Permit Type                                   0
Permit Type Definition                        0
Permit Creation Date                          0
Block                                         0
Lot                                           0
Street Number                                 0
Street Number Suffix                     166473
Street Name                                   0
Street Suffix                              2326
Unit                                     143489
Unit Suffix                              166635
Description                                 234
Current Status                                0
Current Status Date                           0
Filed Date                                    0
Issued Date                               12990
Completed Date                            86432
First Construction Document Date          12983
Structural Notification                  162407
Number of Existing Stories                35896
Number of Proposed Stories                36065
Voluntary Soft-Story Retrofit            168297
Fire Only Permit                         152602
Permit Expiration Date                    43922
Estimated Cost                            31893
Revised Cost                               5207
Existing Use                              34507
Existing Units                            43183
Proposed Use                              35685
Proposed Units                            42720
Plansets                                  31226
TIDF Compliance                          168321
Existing Construction Type                36359
Existing Construction Type Description    36359
Proposed Construction Type                36311
Proposed Construction Type Description    36311
Site Permit                              163959
Supervisor District                        1444
Neighborhoods - Analysis Boundaries        1450
Zipcode                                    1443
Location                                   1430
Record ID                                     0
dtype: int64
```

```
permits.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 168322 entries, 0 to 168321
Data columns (total 43 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   Permit Number                         168322 non-null  object
 1   Permit Type                           168322 non-null  int64
 2   Permit Type Definition                168322 non-null  object
 3   Permit Creation Date                  168322 non-null  datetime64[ns]
 4   Block                                 168322 non-null  object
 5   Lot                                   168322 non-null  object
 6   Street Number                         168322 non-null  int64
 7   Street Number Suffix                  1849 non-null    object
 8   Street Name                           168322 non-null  object
 9   Street Suffix                         165996 non-null  object
 10  Unit                                  24833 non-null   float64
 11  Unit Suffix                           1687 non-null    object
 12  Description                           168088 non-null  object
 13  Current Status                        168322 non-null  object
 14  Current Status Date                   168322 non-null  object
 15  Filed Date                            168322 non-null  object
 16  Issued Date                           155332 non-null  datetime64[ns]
 17  Completed Date                        81890 non-null   object
 18  First Construction Document Date      155339 non-null  object
 19  Structural Notification               5915 non-null    object
 20  Number of Existing Stories            132426 non-null  float64
 21  Number of Proposed Stories            132257 non-null  float64
 22  Voluntary Soft-Story Retrofit         25 non-null      object
 23  Fire Only Permit                      15720 non-null   object
 24  Permit Expiration Date                124400 non-null  object
 25  Estimated Cost                        136429 non-null  float64
 26  Revised Cost                          163115 non-null  float64
 27  Existing Use                          133815 non-null  object
```

```
 28  Existing Units                          125139 non-null  float64
 29  Proposed Use                            132637 non-null  object
 30  Proposed Units                          125602 non-null  float64
 31  Plansets                                137096 non-null  float64
 32  TIDF Compliance                         1 non-null       object
 33  Existing Construction Type              131963 non-null  float64
 34  Existing Construction Type Description  131963 non-null  object
 35  Proposed Construction Type              132011 non-null  float64
 36  Proposed Construction Type Description  132011 non-null  object
 37  Site Permit                             4363 non-null    object
 38  Supervisor District                     166878 non-null  float64
 39  Neighborhoods - Analysis Boundaries     166872 non-null  object
 40  Zipcode                                 166879 non-null  float64
 41  Location                                166892 non-null  object
 42  Record ID                               168322 non-null  float64
dtypes: datetime64[ns](2), float64(13), int64(2), object(26)
memory usage: 55.2+ MB
```

5. Print out the percentage of the missing values in the dataset

```
percentage_missing = (missing_values / len(permits)) * 100
percentage_missing.round(1)
```

```
Permit Number                             0.0
Permit Type                               0.0
Permit Type Definition                    0.0
Permit Creation Date                      0.0
Block                                     0.0
Lot                                       0.0
Street Number                             0.0
Street Number Suffix                     98.9
Street Name                               0.0
Street Suffix                             1.4
Unit                                     85.2
Unit Suffix                              99.0
Description                               0.1
Current Status                            0.0
Current Status Date                       0.0
Filed Date                                0.0
Issued Date                               7.7
Completed Date                           51.3
First Construction Document Date          7.7
Structural Notification                  96.5
Number of Existing Stories               21.3
Number of Proposed Stories               21.4
Voluntary Soft-Story Retrofit           100.0
Fire Only Permit                         90.7
Permit Expiration Date                   26.1
Estimated Cost                           18.9
Revised Cost                              3.1
Existing Use                             20.5
Existing Units                           25.7
Proposed Use                             21.2
Proposed Units                           25.4
Plansets                                 18.6
TIDF Compliance                         100.0
Existing Construction Type               21.6
Existing Construction Type Description   21.6
Proposed Construction Type               21.6
Proposed Construction Type Description   21.6
Site Permit                              97.4
Supervisor District                       0.9
Neighborhoods - Analysis Boundaries       0.9
Zipcode                                   0.9
Location                                  0.8
Record ID                                 0.0
dtype: float64
```

```
# total_percentage_missing = permits.isna().sum().sum() / ()
# total_percentage_missing

total_percentage_missing = (permits.isna().sum().sum() / permits.size) * 100
total_percentage_missing
```

```
26.23099469096192
```

```
percentage_missing = (missing_values / len(permits)) * 100

missing_columns = percentage_missing[percentage_missing > 0]
```

```
print(missing_columns.round(1))
```

```
Street Number Suffix                      98.9
Street Suffix                              1.4
Unit                                      85.2
Unit Suffix                               99.0
Description                                0.1
Issued Date                                7.7
Completed Date                            51.3
First Construction Document Date           7.7
Structural Notification                   96.5
Number of Existing Stories                21.3
Number of Proposed Stories                21.4
Voluntary Soft-Story Retrofit            100.0
Fire Only Permit                          90.7
Permit Expiration Date                    26.1
Estimated Cost                            18.9
Revised Cost                               3.1
Existing Use                              20.5
Existing Units                            25.7
Proposed Use                              21.2
Proposed Units                            25.4
Plansets                                  18.6
TIDF Compliance                          100.0
Existing Construction Type                21.6
Existing Construction Type Description    21.6
Proposed Construction Type                21.6
Proposed Construction Type Description    21.6
Site Permit                               97.4
Supervisor District                        0.9
Neighborhoods - Analysis Boundaries        0.9
Zipcode                                    0.9
Location                                   0.8
dtype: float64
```

6. How many missing values in the columns "Street Number Suffix" and "Zipcode" ?

```
missing_street_suffix = permits['Street Number Suffix'].isnull().sum()

print("Resuslt:",missing_street_suffix)
```

```
Resuslt: 166473
```

```
missing_zipcode = permits['Zipcode'].isnull().sum()

print("Result is: ", missing_zipcode)
```

```
Result is:  1443
```

7. Create new dataFrame named “pemits_dropped” that has all the column with empty values.

```
permits_dropped = permits.dropna(axis=1)

permits_dropped.head()
```

	Permit Number	Permit Type	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Name	Current Status	Cu S
0	2.02E+11	4	sign - erect	2015-05-06	326	23	140	Ellis	expired	12/21
1	2.02E+11	4	sign - erect	2016-04-19	306	7	440	Geary	issued	08/03
2	2.02E+11	3	additions alterations or repairs	2016-05-27	595	203	1647	Pacific	withdrawn	09/26
3	2.02E+11	8	otc alterations	2016-11-	156	11	1230	Pacific	complete	07/24

8. Calculate all number of the dropped columns.

```
dropped_columns_count = len(permits.columns) - len(permits_dropped.columns)

print("================================================")
```

```
print("Number of dropped columns:", dropped_columns_count)
print("===================================================")
```

```
==================================================
Number of dropped columns: 31
==================================================
```

9. Replacing all missing value in Street Number Suffix with “0”

```
permits['Street Number Suffix'].fillna("0")

permits
```

	Permit Number	Permit Type	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Number Suffix	Stre Na
0	2.02E+11	4	sign - erect	2015-05-06	326	23	140	NaN	E
1	2.02E+11	4	sign - erect	2016-04-19	306	7	440	NaN	Ge
2	2.02E+11	3	additions alterations or repairs	2016-05-27	595	203	1647	NaN	Pac
3	2.02E+11	8	otc alterations permit	2016-11-07	156	11	1230	NaN	Pac
4	2.02E+11	6	demolitions	2016-11-28	342	1	950	NaN	Mar
...	...	...	...	...	...	...	...	...	
168317	M862628	8	otc alterations permit	2017-12-05	113	017A	1228	NaN	Montgom
168318	2.02E+11	8	otc alterations permit	2017-12-05	271	14	580	NaN	Bu
168319	M863507	8	otc alterations permit	2017-12-06	4318	19	1568	NaN	India
168320	M863747	8	otc alterations permit	2017-12-06	298	29	795	NaN	Sut
168321	M864287	8	otc alterations permit	2017-12-07	160	6	838	NaN	Pac

168322 rows × 43 columns

10. what is the existing use of the units with the following permits No: M805907 and M839987

```
existing = permits.loc[permits['Permit Number'] == "M805907", 'Existing Use']

print("Result is:", existing)
```

```
Result is: 11    NaN
Name: Existing Use, dtype: object
```

```
existing_use = permits.loc[permits['Permit Number'] == "M839987", 'Existing Use']

print("Result is: ", existing_use)
```

```
Result is:  326    NaN
Name: Existing Use, dtype: object
```