数据集San Francisco Building Permits

**一、数据可视化和摘要**

1. 数据摘要：

• 对标称属性，给出每个可能取值的频数，

|  |  |  |
| --- | --- | --- |
| **Permit Type**  1 349  2 950  3 14663  4 2892  5 91  6 600  7 511  8 178844  **Existing Construction Type**  1.0 28072  2.0 4068  3.0 9663  4.0 381  5.0 113350 | **Street Suffix**  Al 83  Av 43219  Bl 3555  Cr 97  Ct 667  Dr 3267  Hl 1  Hy 240  Ln 354  No 2  Pk 128  Pl 538  Pz 210  Rd 389  Rw 5  So 2  St 138358  Sw 2  Tr 1466  Wk 9  Wy 3540 | **Proposed Construction Type**  1.0 27841  2.0 3778  3.0 9360  4.0 377  5.0 114382  **Proposed Construction Type Description**  constr type 1 27841  constr type 2 3778  constr type 3 9360  constr type 4 377  wood frame (5) 114382  **Existing Construction Type Description**  constr type 1 28072  constr type 2 4068  constr type 3 9663  constr type 4 381  wood frame (5) 113350 |

分析代码：

g1=df.groupby(‘Revised Cost’ )

g1.size()

• 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of Existing Stories | Number of Proposed Stories | Estimated Cost | Revised Cost |
| mean | 5.705773 | 5.745043 | 1.689554e+05 | 1.328562e+05 |
| Min | 0 | 0 | 1.000000e+00 | 0.000000e+00 |
| 25% | 2 | 2 | 3.300000e+03 | 1.000000e+00 |
| 50% | 3 | 3 | 1.100000e+04 | 7.000000e+03 |
| 75% | 4 | 4 | 3.500000e+04 | 2.870750e+04 |
| max | 78 | 78 | 5.379586e+08 | 7.805000e+08 |
| 中位数 | 3 | 3 | 11000 | 7000 |
| 缺失数 | 42784 | 42868 | 38066 | 6066 |

分析代码：

#平均数、最大、最小、四分位：

df.describe()

#缺失数：

len(df['Number of Existing Stories'])-df['Number of Existing Stories'].count()

#或,显示所有列的缺失数

total =df.isnull().sum().sort\_values(ascending=False)

print(total)

#中位数：

df['Number of Existing Stories'].median()

#或

c1=df[np.isnan(df['Number of Existing Stories'])==False]

s1=c1['Number of Existing Stories']

s2= s1.sort\_values().reset\_index(drop=True)

s2[len(s2) //2]

1. 数据的可视化

安装anaconda包

conda update conda #更新Anaconda包 （必须以管理员身份进入 命令提示符）

conda install numpy #安装过程慢

conda install matplotlib

conda list

针对数值属性

（1）绘制直方图，用qq图检验其分布是否为正态分布

#直方图

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

df=pd.read\_csv('e:\dmjob\job1\Building\_Permits.csv')

c1=df[np.isnan(df['Number of Existing Stories'])==False]

fig = plt.figure()

ax = fig.add\_subplot(111)

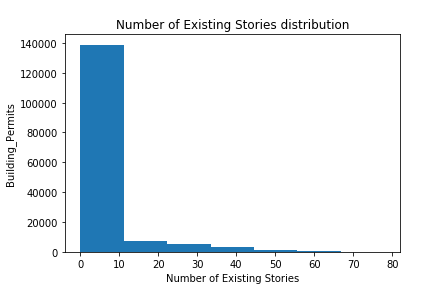
ax.hist(c1['Number of Existing Stories'], bins=7)

plt.title('Number of Existing Stories distribution')

plt.xlabel('Number of Existing Stories')

plt.ylabel('Building\_Permits')

plt.show()



（2）绘制盒图，对离群值进行识别

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

df=pd.read\_csv('e:\dmjob\job1\Building\_Permits.csv')

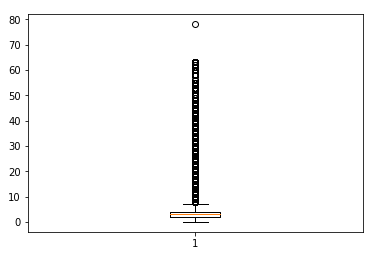
c1=df[np.isnan(df['Number of Existing Stories'])==False]

fig = plt.figure()

ax = fig.add\_subplot(111)

ax.boxplot(c1[' Number of Existing Stories '])

plt.show()



二、数据缺失的处理

• 将缺失部分剔除

• 用最高频率值来填补缺失值

• 通过属性的相关关系来填补缺失值

• 通过数据对象之间的相似性来填补缺失值

#统计缺失数据

total =df.isnull().sum().sort\_values(ascending=False)

print(total)

|  |  |  |
| --- | --- | --- |
| 列名 | 缺失数 | 处理办法 |
| TIDF Compliance | 198898 | 删除列 |
| Voluntary Soft-Story Retrofit | 198865 | 删除列 |
| Unit Suffix | 196939 | 删除列 |
| Street Number Suffix | 196684 | 删除列 |
| Site Permit | 193541 | 删除列 |
| Structural Notification | 191978 | 删除列 |
| Fire Only Permit | 180073 | 删除列 |
| Unit | 169421 | 删除列 |
| Completed Date | 101709 | 删除列 |
| Permit Expiration Date | 51880 | 删除列 |
| Existing Units | 51538 | -1 |
| Proposed Units | 50911 | -1 |
| Existing Construction Type | 43366 | 用最高频率值来填补缺失值 |
| Existing Construction Type Description | 43366 |  |
| Proposed Construction Type | 43162 |  |
| Proposed Construction Type Description | 43162 |  |
| Number of Proposed Stories | 42868 | -1 |
| Number of Existing Stories | 42784 | -1 |
| Proposed Use | 42439 |  |
| Existing Use | 41114 |  |
| Estimated Cost | 38066 |  |
| Plansets | 37309 |  |
| First Construction Document Date | 14946 |  |
| Issued Date | 14940 |  |
| Revised Cost | 6066 | 均值 |
| Street Suffix | 2768 |  |
| Neighborhoods - Analysis Boundaries | 1725 |  |
| Supervisor District | 1717 |  |
| Zipcode | 1716 |  |
| Location | 1700 |  |
| Description | 290 | 忽略 |

#去掉一列

df = df.drop(['TIDF Compliance'], axis = 1)

#填-1

df['Existing Units'] = df['Existing Units'].fillna(-1)

#使用平均数填补

df[' Revised Cost'] = df[' Revised Cost'].fillna(df[' Revised Cost'].mean())

#使用出现次数最多的值填补

df['Existing Construction Type'] = df['Existing Construction Type'].fillna('5.0')

#返回已经去掉重复行的数据集

df.drop\_duplicates()

#去掉这个特征为空的行

df\_new = df.drop(df[df['Description'].isnull()].index)