作业1 —**Consumer & Visitor Insights For Neighborhoods**

通过观察数据集，发现该数据集标称属性包括：

visitor\_home\_cbgs

visitor\_work\_cbgs,

related\_same\_day\_brand,

related\_same\_month\_brand,

top\_brands,

popularity\_by\_hour,

popularity\_by\_day

数值属性,包括：

census\_block\_group

date\_range\_start

date\_range\_end

raw\_visit\_count

raw\_visitor\_count

distance\_from\_home

1 以‘visitor\_home\_cbgs’为例，计算标称属性的频数如下（只展示部分结果）：

{} 28412

{"220710133021":95} 4

{"060750601001":57} 4

{"220870302082":59} 4

{"420210111001":50} 3

{"720915702012":51} 3

{"191114907001":59} 3

{"060552017002":58} 3

…

2 计算全部数值属性的最大最小值等，如：

census\_block\_group 7.803099e+11

date\_range\_start 1.538352e+09

date\_range\_end 1.541030e+09

raw\_visit\_count 7.179900e+06

raw\_visitor\_count 6.113949e+06

distance\_from\_home 6.297845e+06

dtype: float64

census\_block\_group 1.001020e+10

date\_range\_start 1.538352e+09

date\_range\_end 1.541030e+09

raw\_visit\_count 6.000000e+01

raw\_visitor\_count 5.000000e+01

distance\_from\_home 7.060000e+02

dtype: float64

census\_block\_group 2.870864e+11

date\_range\_start 1.538352e+09

date\_range\_end 1.541030e+09

raw\_visit\_count 4.793066e+04

raw\_visitor\_count 1.182032e+04

distance\_from\_home 3.511280e+04

dtype: float64

census\_block\_group 2.901900e+11

date\_range\_start 1.538352e+09

date\_range\_end 1.541030e+09

raw\_visit\_count 3.064000e+04

raw\_visitor\_count 6.541000e+03

distance\_from\_home 1.461400e+04

dtype: float64

census\_block\_group 1.312101e+11

date\_range\_start 1.538352e+09

date\_range\_end 1.541030e+09

raw\_visit\_count 1.704200e+04

raw\_visitor\_count 3.430000e+03

distance\_from\_home 8.584000e+03

Name: 0.25, dtype: float64

census\_block\_group 1

date\_range\_start 0

date\_range\_end 0

raw\_visit\_count 106

raw\_visitor\_count 106

distance\_from\_home 217

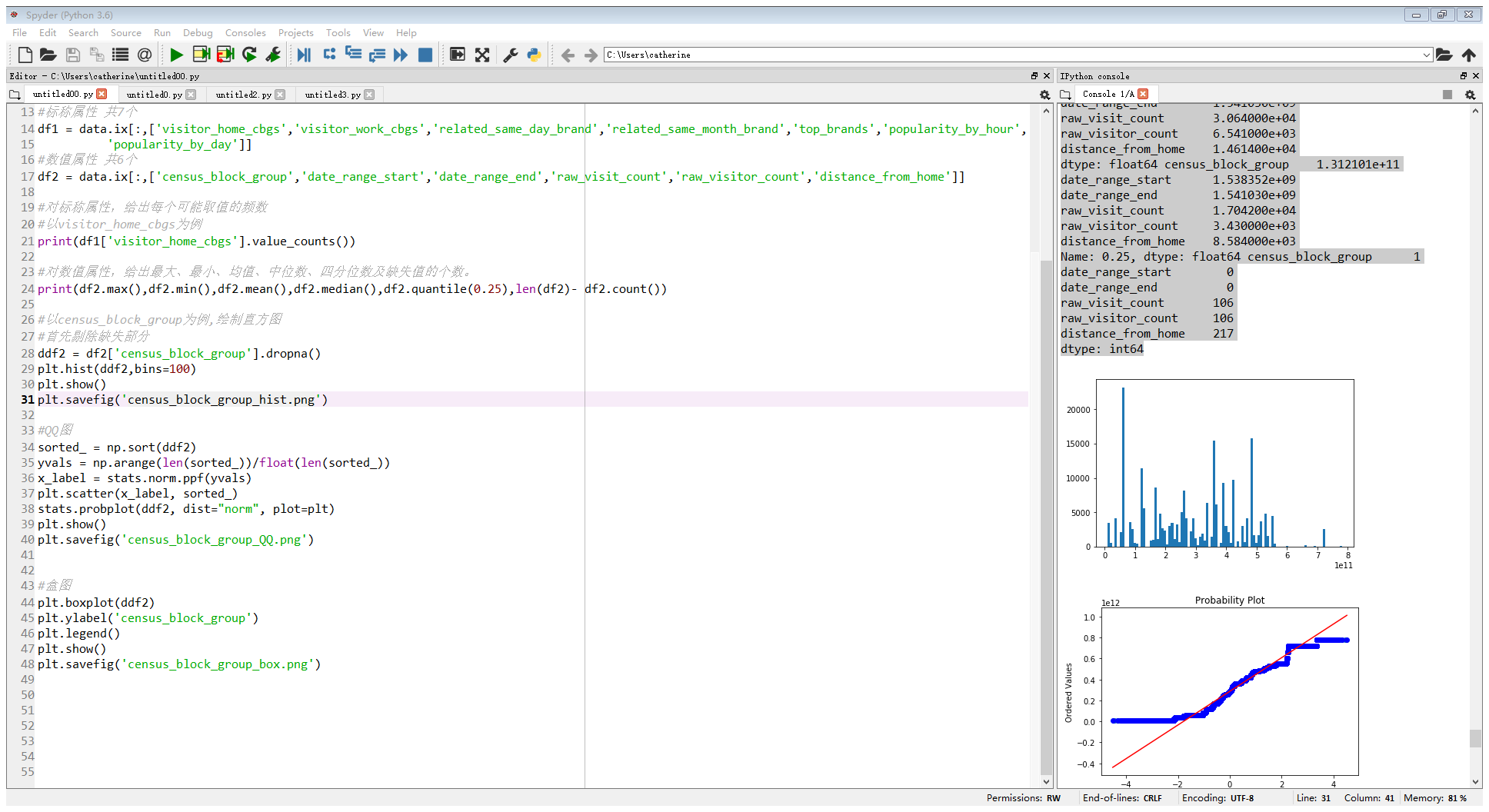
dtype: int64

3 直方图+qq图+盒图

以‘census\_block\_group’为例

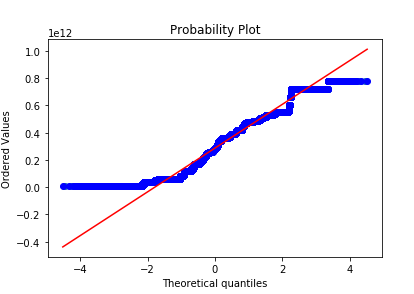
3.1 剔除缺失数据

**直方图（以census\_block\_group\_hist\_1.png存储）：**

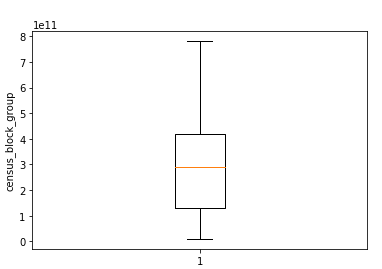


**QQ图（'census\_block\_group\_QQ\_1.png'存储）：**

并不是所有点基本上在一条直线上，所以这两个分布不是同一分布，也即不服从正太分布。

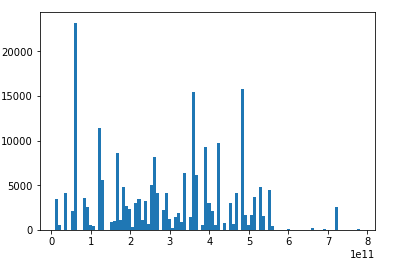


**盒图（'census\_block\_group\_box\_1.png'存储）：**

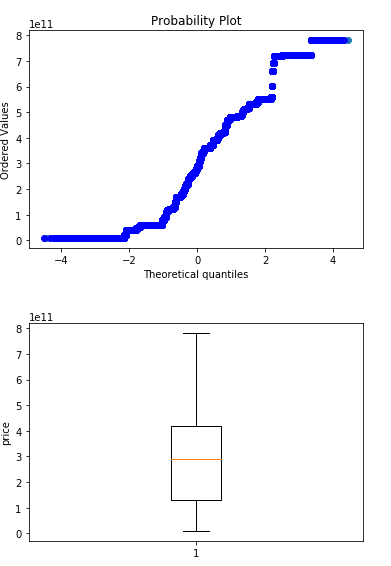


3.2 以最高频率值来填补缺失值

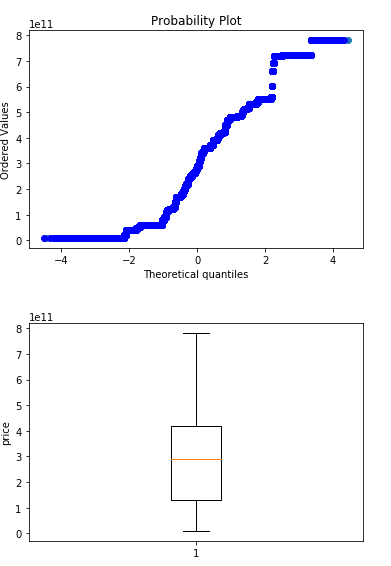
直方图：



QQ图：

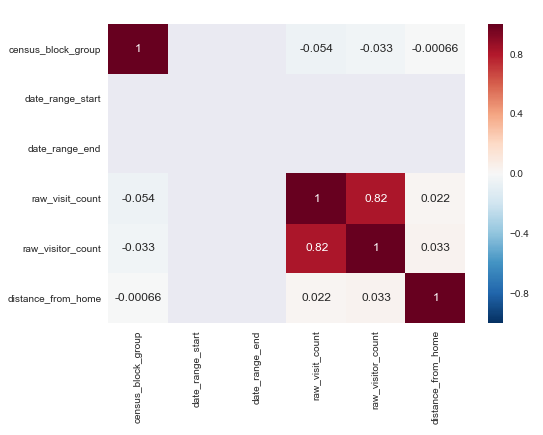


盒图：

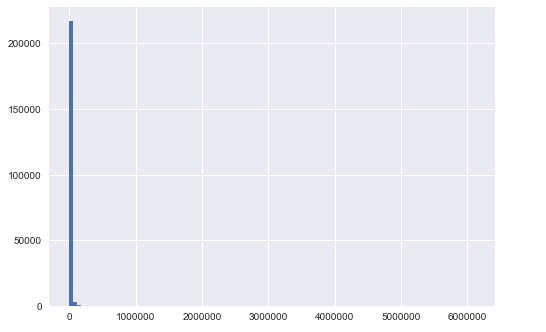


3.3 以属性间的相关关系来填补缺失值（根据热图，知census和raw\_visit\_count还算相关，但使用随机森林方法预测 数据集1中**census\_ block\_ group**属性值，需要将其转换为数值形式，而其本身当作数值来看超出内存，故使用date\_range\_start预测raw\_visitor\_count，效果可能不是很理想）

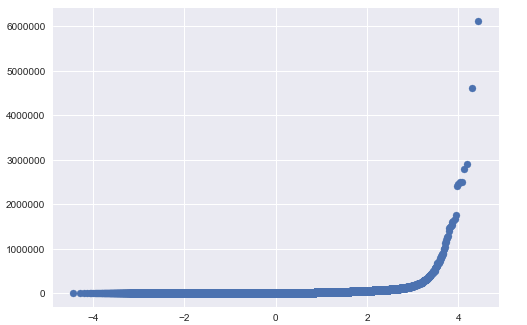
**（热图：）**



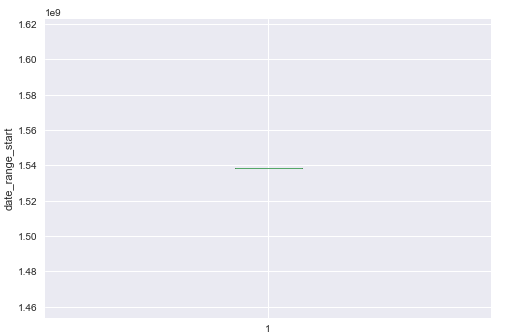
直方图：



QQ图：



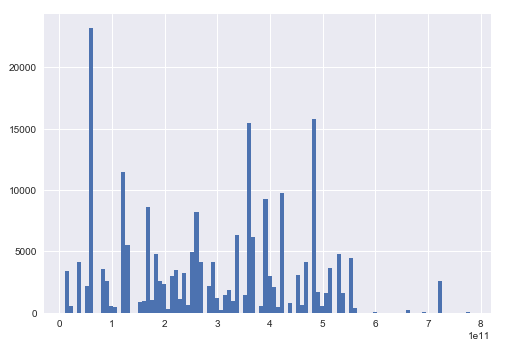
盒图：



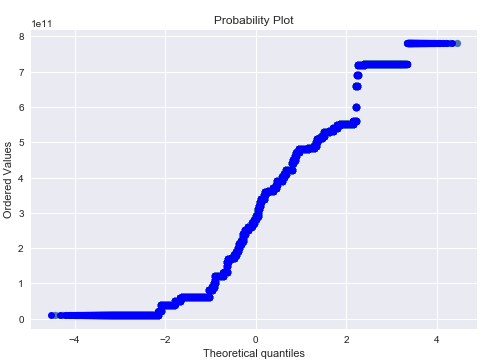
3.4 通过数据对象之间的相似性来填补缺失值

同3.2，但使用mean值

直方图：



QQ图：



盒图：

