作业一：数据探索性分析与数据预处理

曹倩雯
3120170497

一. 问题描述
本次作业中，将对 2 个数据集进行探索性分析与预处理。

二. 数据说明
数据集 1: NFL Play-by-Play 2009-2017
数据集 2: San Francisco Building Permits

三. 数据分析要求
1． 数据可视化和摘要
数据摘要
对标称属性，给出每个可能取值的频数，
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。
数据的可视化
针对数值属性，
绘制直方图，用 qq 图检验其分布是否为正态分布。
绘制盒图，对离群值进行识别

2． 数据缺失的处理
观察数据集中缺失数据，分析其缺失的原因。
分别使用下列四种策略对缺失值进行处理:
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
处理后，可视化地对比新旧数据集。

四．实验环境及语言
语言及环境依赖
语言： python2
依赖的包：xlrd, pylab, matplotlib, scipy, numpy
xlrd: 数据摘要处理时用到
pylab, matplotlib, scipy, numpy：数据可视化时用于生成图

五．具体问题

1. 数据摘要

1.1 题目要求：

对标称属性，给出每个可能取值的频数，

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

1.2 问题分析：

标称属性：标称型目标变量的结果只在有限目标集中取值，如真与假(标称型目标变量主要用于分类)

数值属性：数值型目标变量则可以从无限的数值集合中取值，如 0.100，42.001 等 (数值型目标变量主要用于回归分析)

因此，对于数据集一，标称型属性有：sp, qtr, down, time, ydstogo, posteam, desc 等;

数值型属性有：Drive, yrdln, AirYards, YardsAfterCatch 等

1.3 结果与分析

见 git

对于数据集一

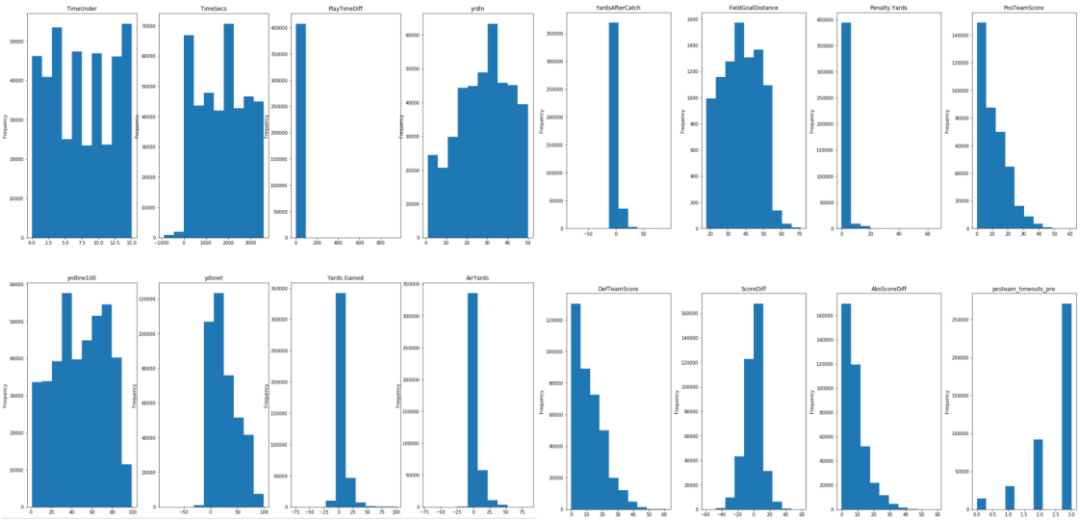
2. 数据可视化

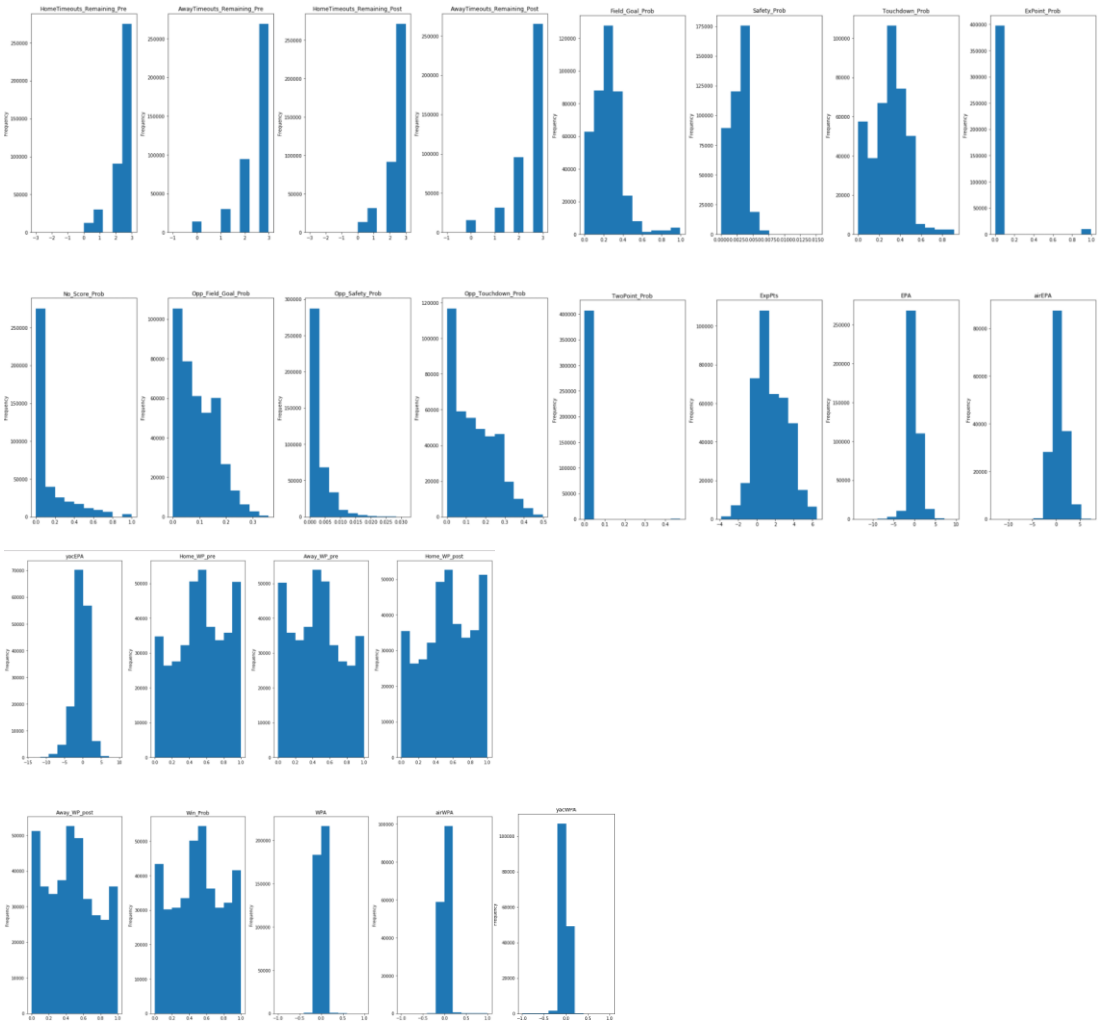
2.1 题目要求：

针对数值属性，绘制直方图，用 qq 图检验其分布是否为正态分布。绘制盒图，对离群值进行识别

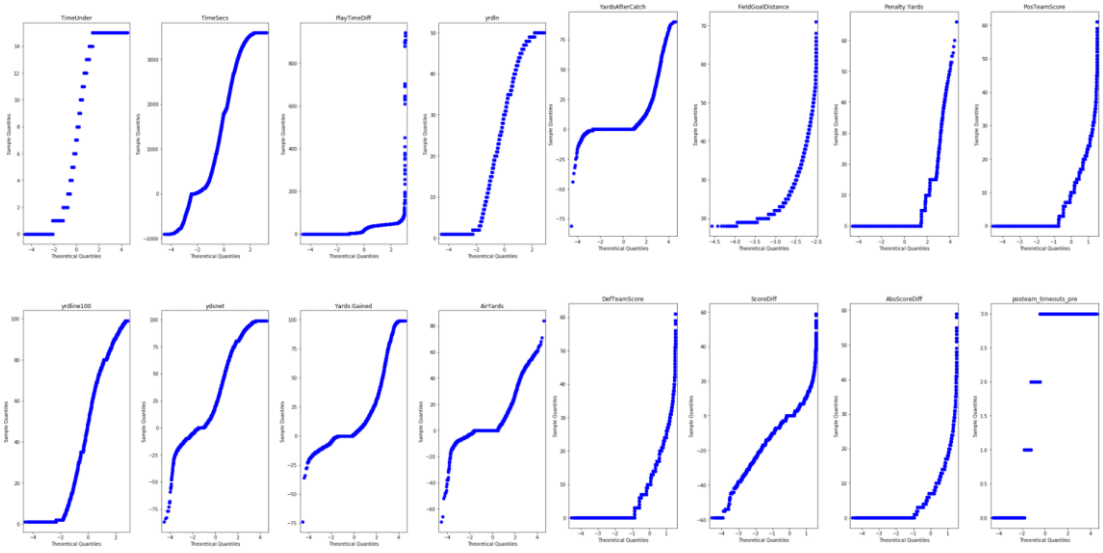
2.2 结果与分析

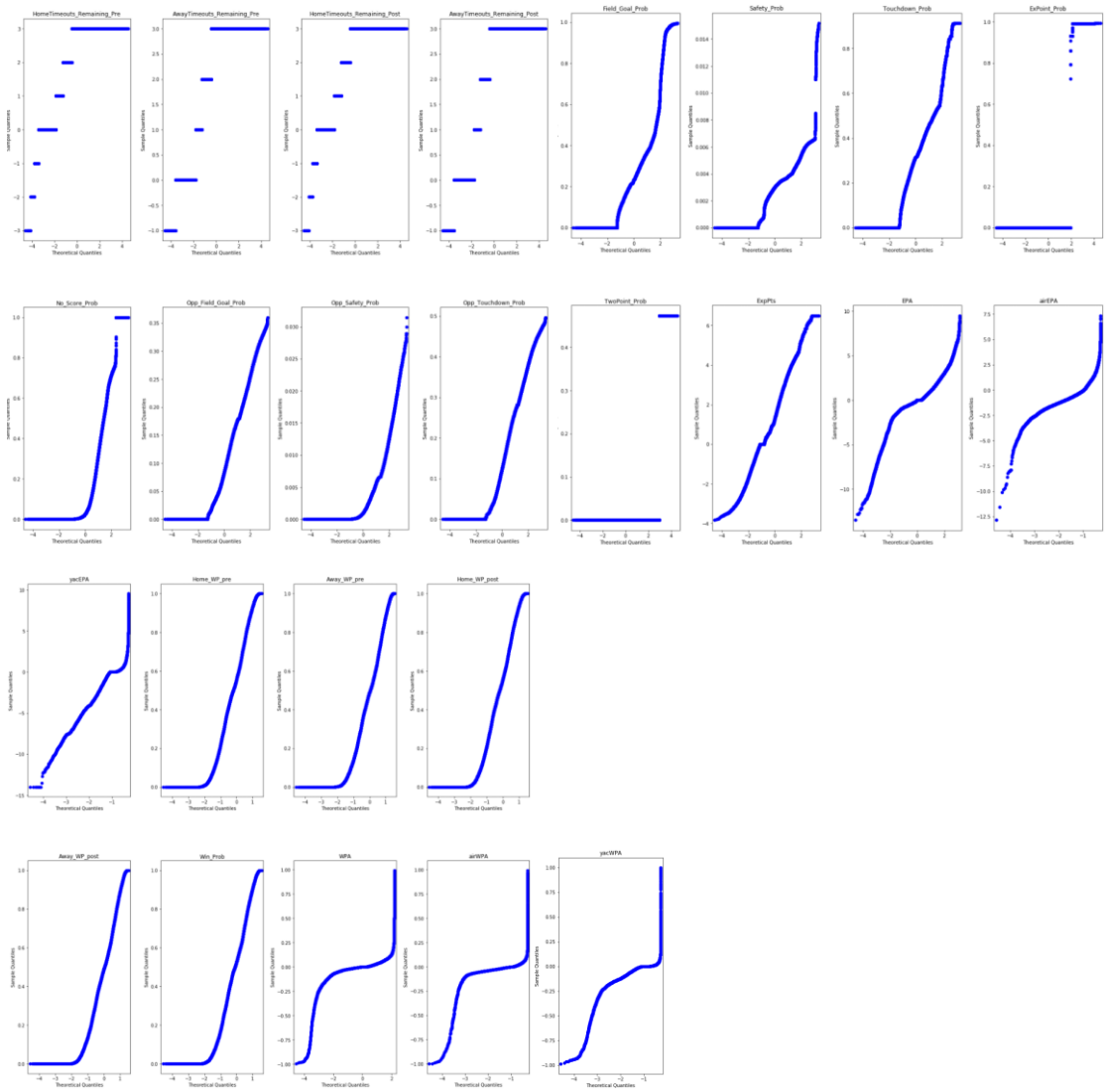
2.3.1 直方图：



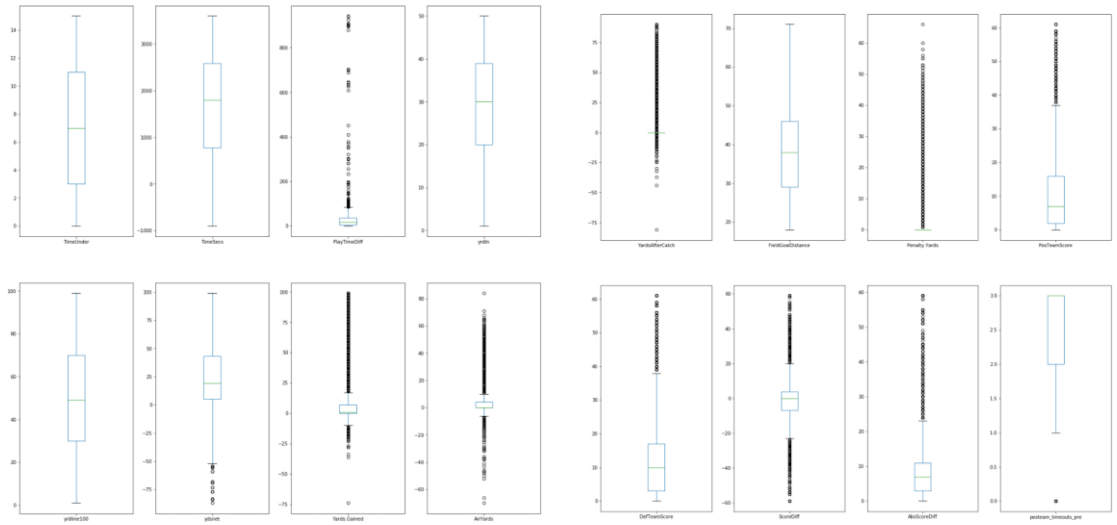

HomeTimeouts_Remaining_Pre
AwayTimeouts_Remaining_Pre
HomeTimeouts_Remaining_Post
AwayTimeouts_Remaining_Post
Field_Goal_Prob
Safety_Prob
Touchdown_Prob
ExPoint_Prob
No_Score_Prob
Opp_Field_Goal_Prob
Opp_Safety_Prob
Opp_Touchdown_Prob
TwoPoint_Prob
ExpPts
EPA
airEPA
yacEPA
Home_WP_pre
Away_WP_pre
Home_WP_post
Away_WP_post
Win_Prob
WPA
airWPA
yacwpa

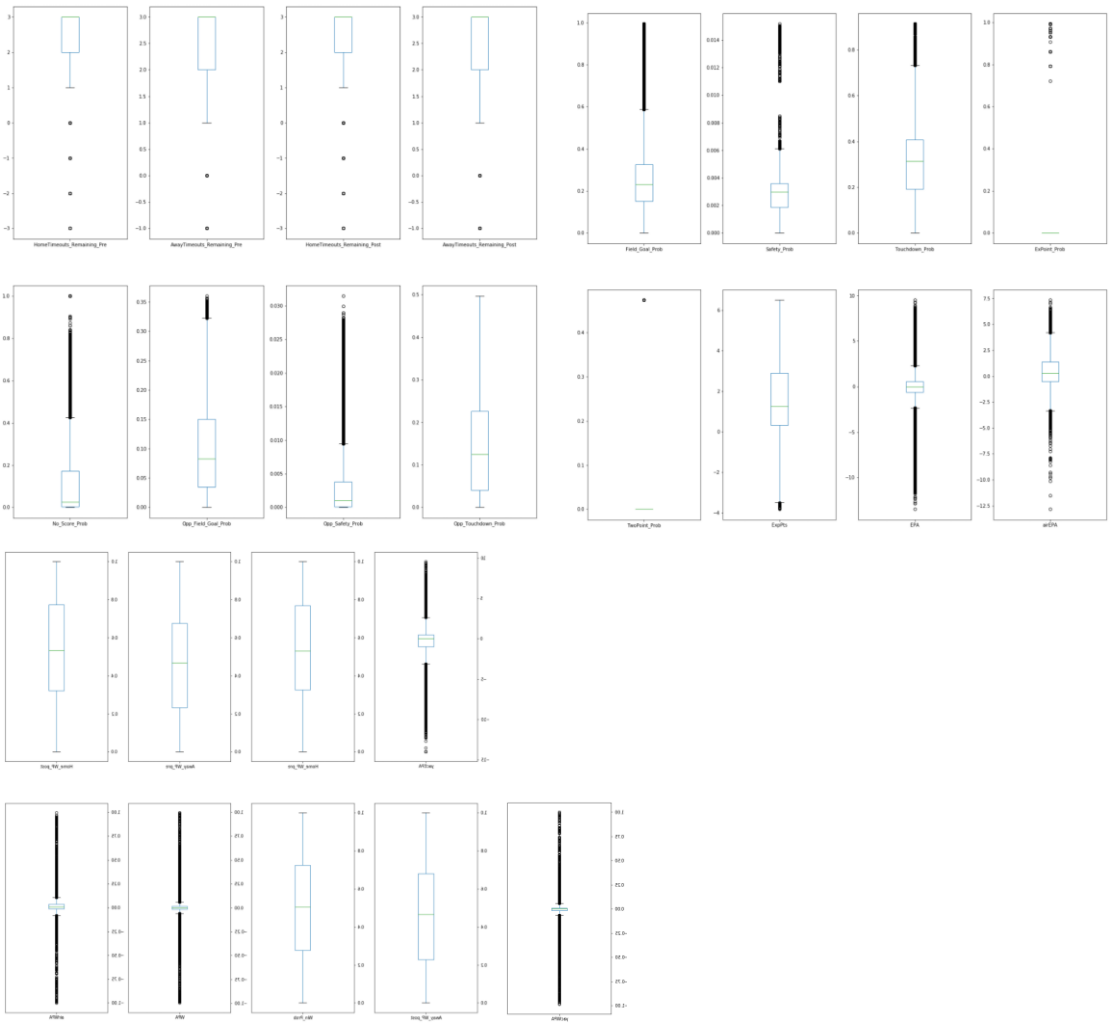
2.3.2 qq 图


TimeUnder
TimeSecs
PlayTimeDiff
ydln
YardsAfterCatch
FieldGoalDistance
Penalty Yards
PosTeamScore
ydsnet
Yards.Gained
AirYards
DefTeamScore
ScoreDiff
AbsScoreDiff
posteam_timeouts_pre
Sample Quantiles
Theoretical Quantiles


HomeTimeouts_Remaining_Pre
AwayTimeouts_Remaining_Pre
HomeTimeouts_Remaining_Post
AwayTimeouts_Remaining_Post
Field_Goal_Prob
Safety_Prob
Touchdown_Prob
ExPoint_Prob
No_Score_Prob
Opp_Field_Goal_Prob
Opp_Safety_Prob
Opp_Touchdown_Prob
TwoPoint_Prob
ExpPts
EPA
airEPA
yacEPA
Home_WP_pre
Away_WP_pre
Home_WP_post
Away_WP_post
Win_Prob
WPA
airWPA
yacWPA

2.3.3 盒图


TimeUnder
TimeSecs
PlayTimeDiff
ydsnet
YardsAfterCatch
FieldGoalDistance
Penalty Yards
PosTeamScore
ydsnet100
ydsnet
Yards.Gained
AirYards
DefTeamScore
ScoreDiff
AbsScoreDiff
posteam_timeouts_pre



3. 数据缺失的处理

3.1 题目要求：

观察数据集中缺失数据，分析其缺失的原因。分别使用下列四种策略对缺失值进行处理:

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

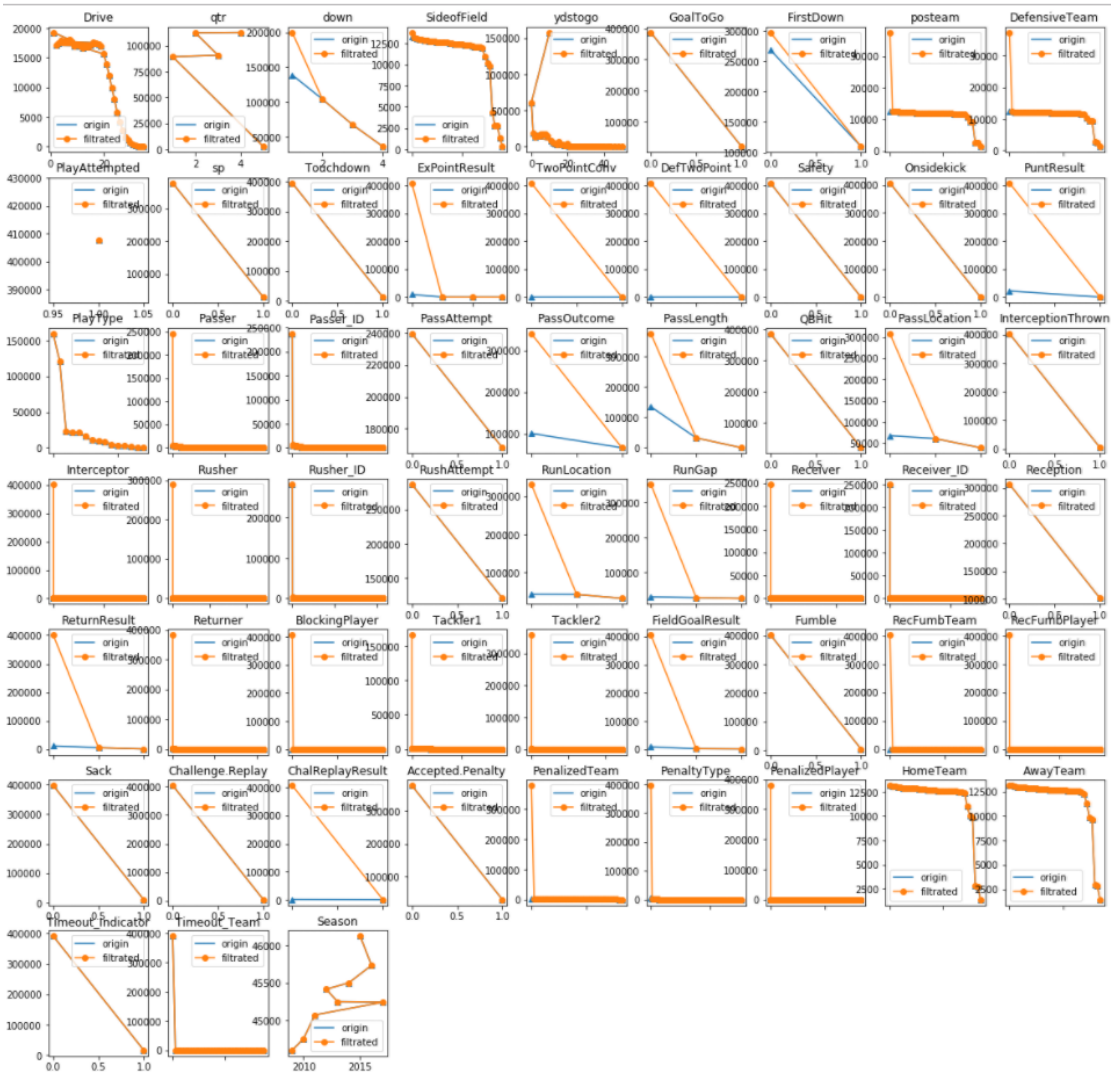
3.2 问题分析：

需要收件将缺失数据剔除，然后依次使用三种策略进行填补

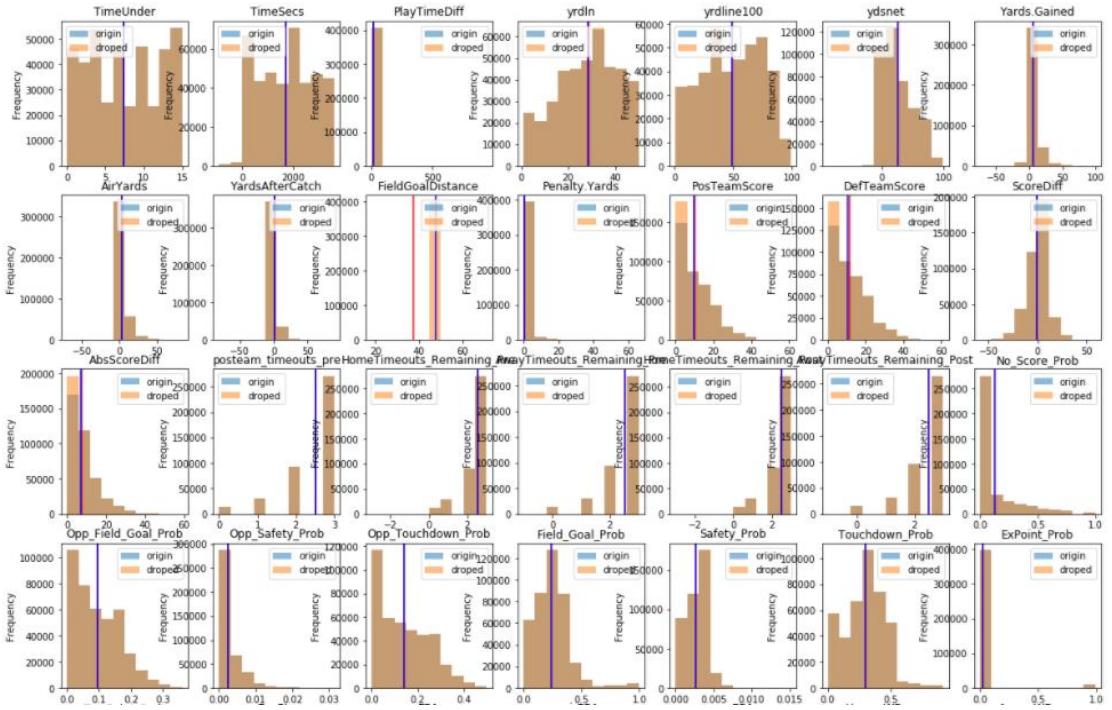
3.3 结果与分析

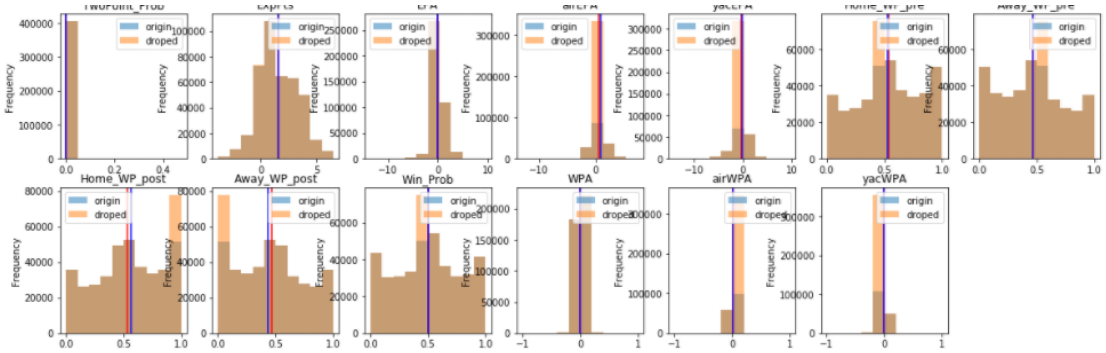
3.3.1 用最高频率值来填补缺失值

对标称属性使用折线图可视化新旧数据集变化



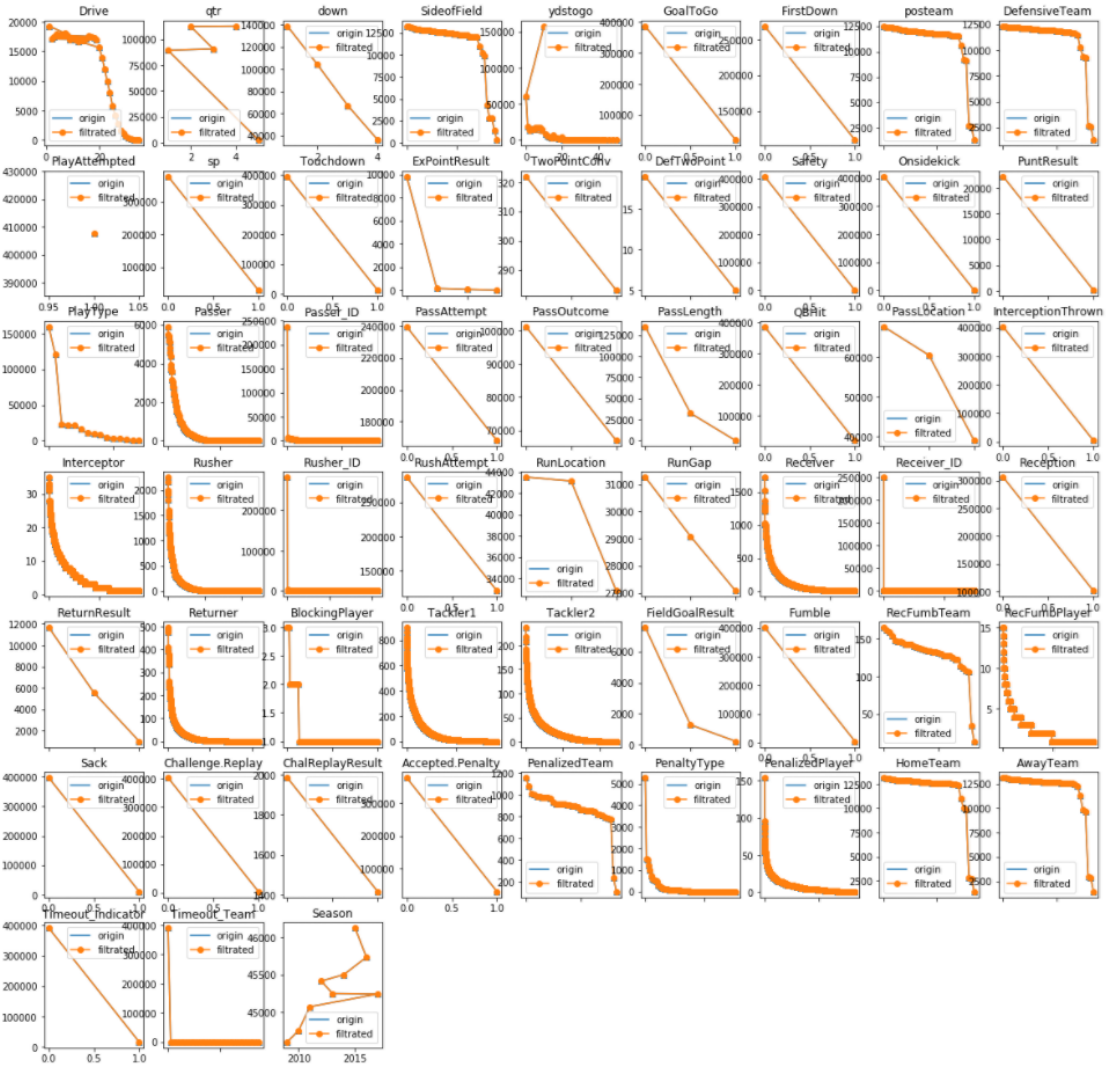
对数值属性使用直方图可视化新旧数据集变化



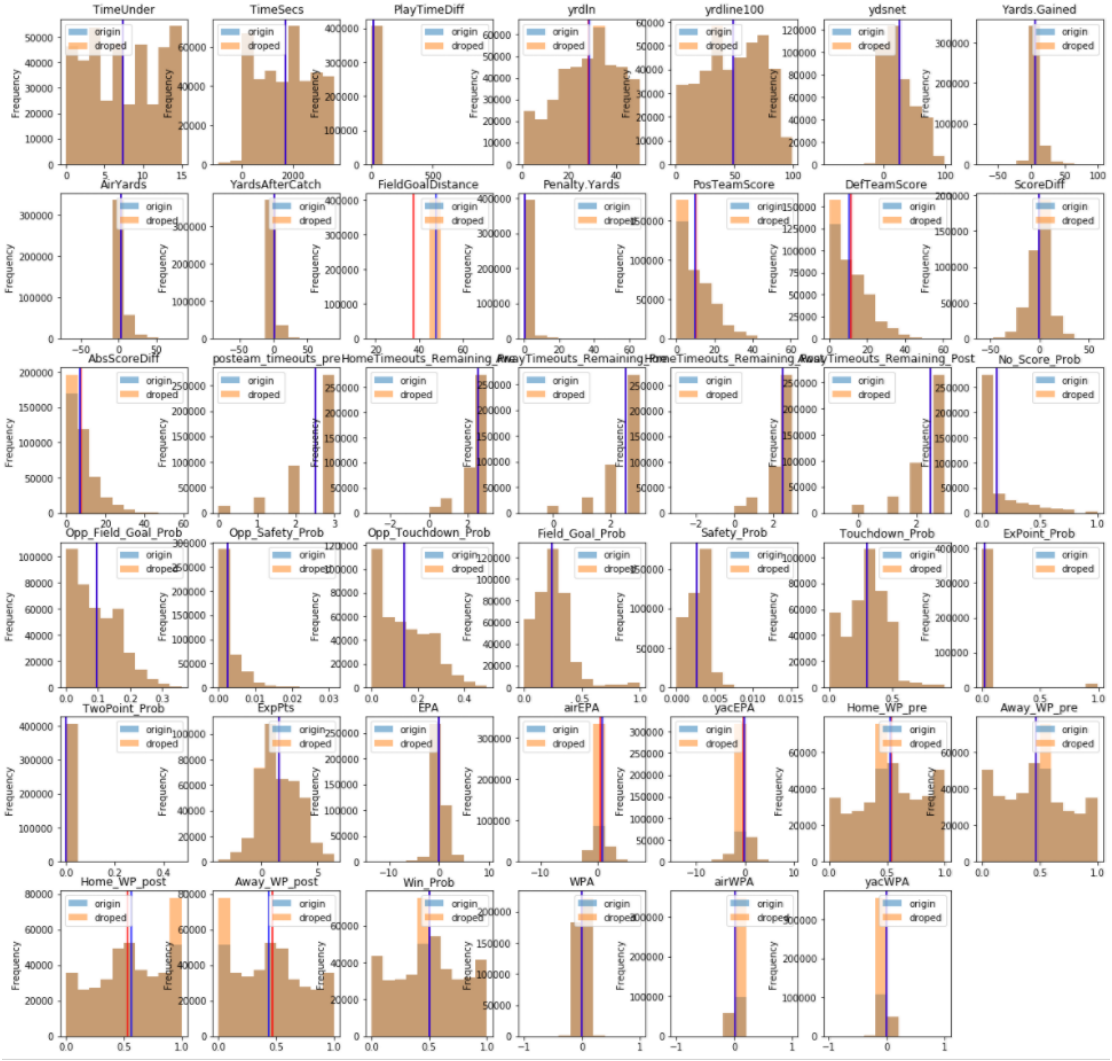


3.3.2 用属性间相关关系来填补缺失值

对标称属性使用折线图可视化新旧数据集变化

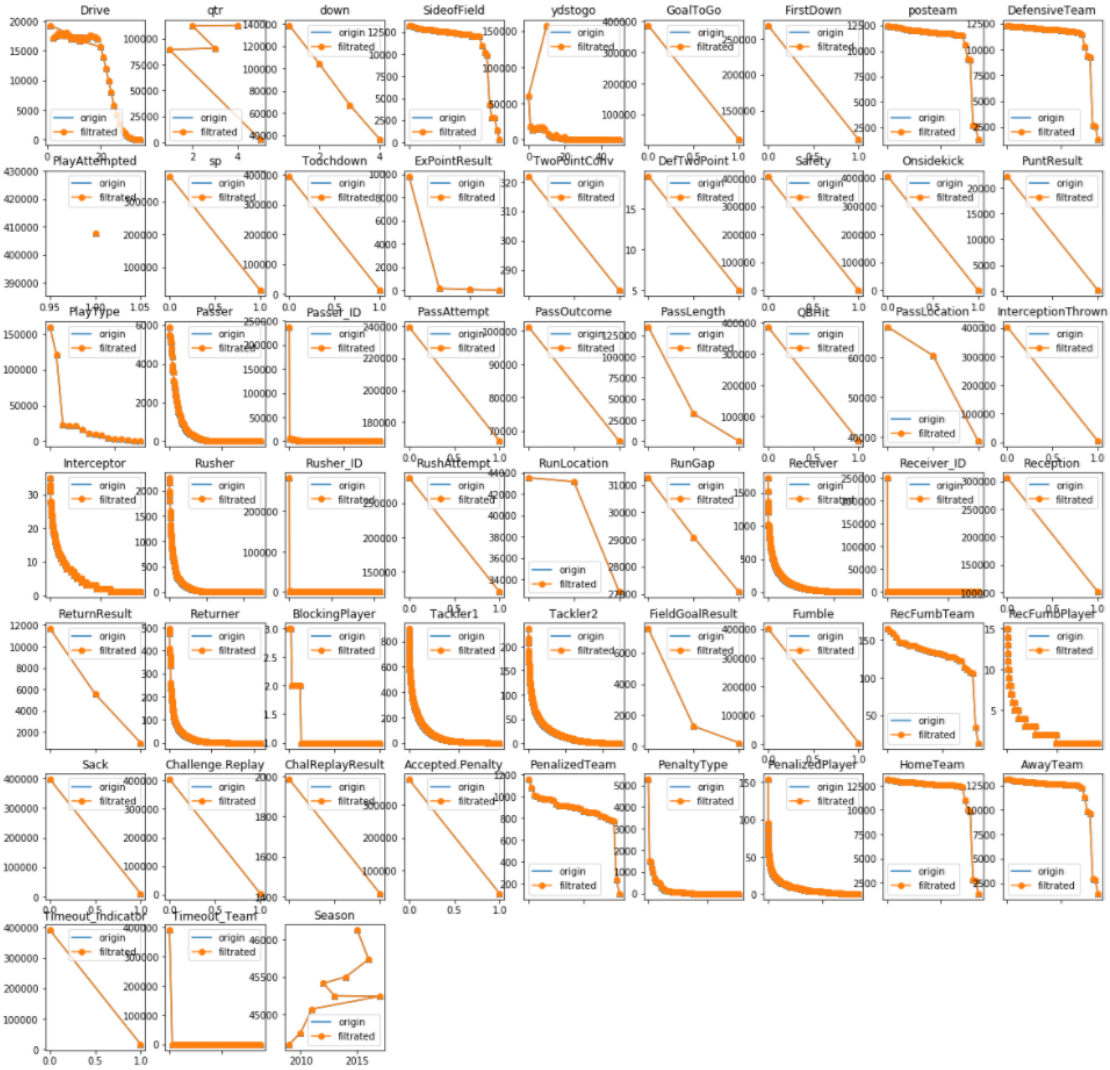


对数值属性使用直方图可视化新旧数据集变化

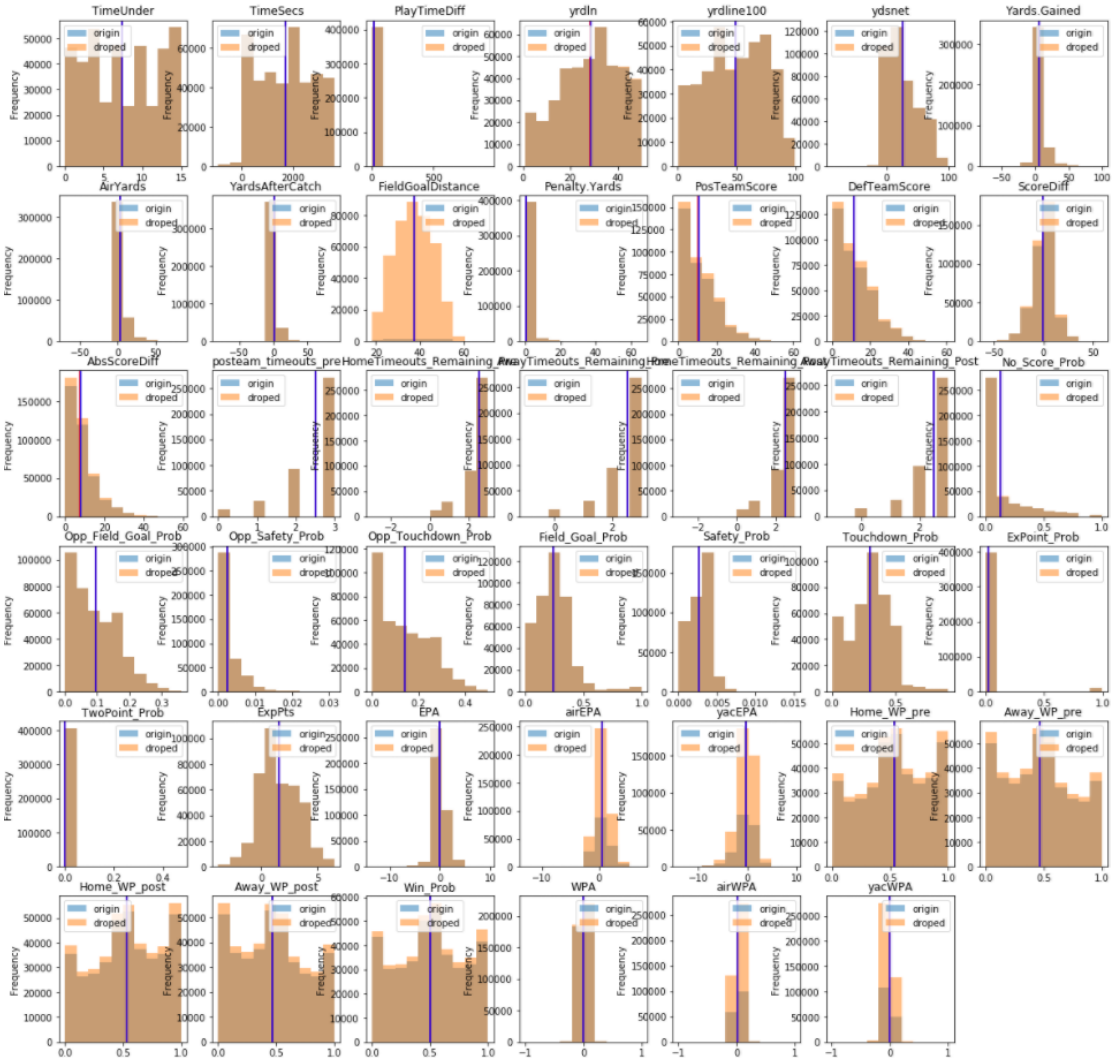


3.3.3 用属性间相关关系来填补缺失值

对标称属性使用折线图可视化新旧数据集变化

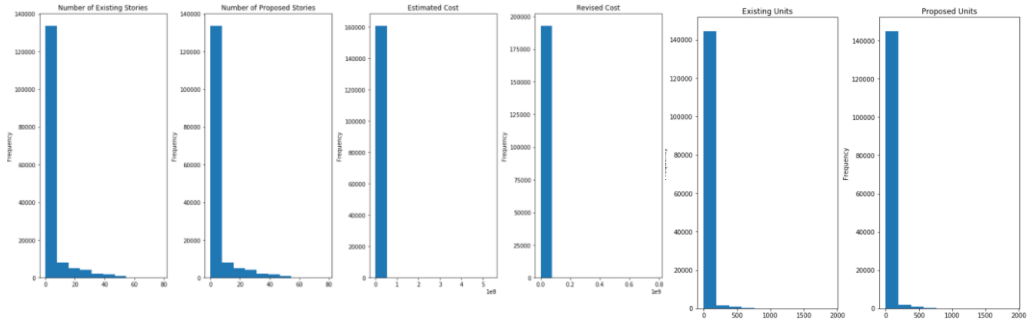


对数值属性使用直方图可视化新旧数据集变化

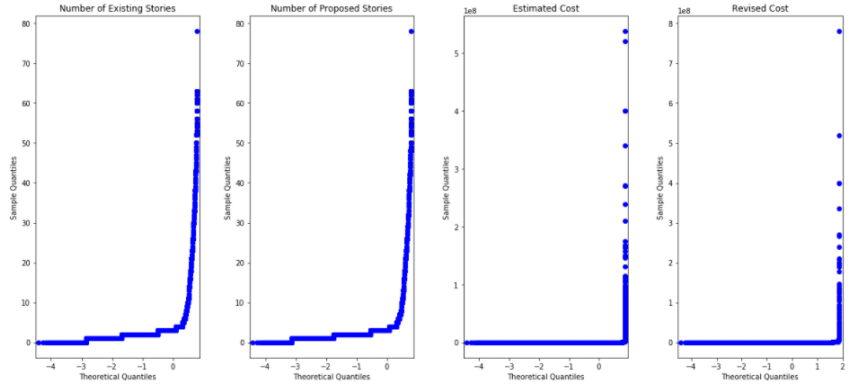


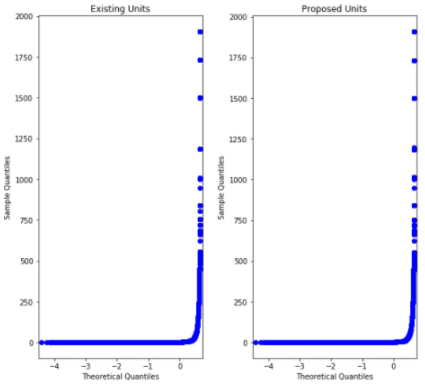
对于数据集二

2.3.1 直方图

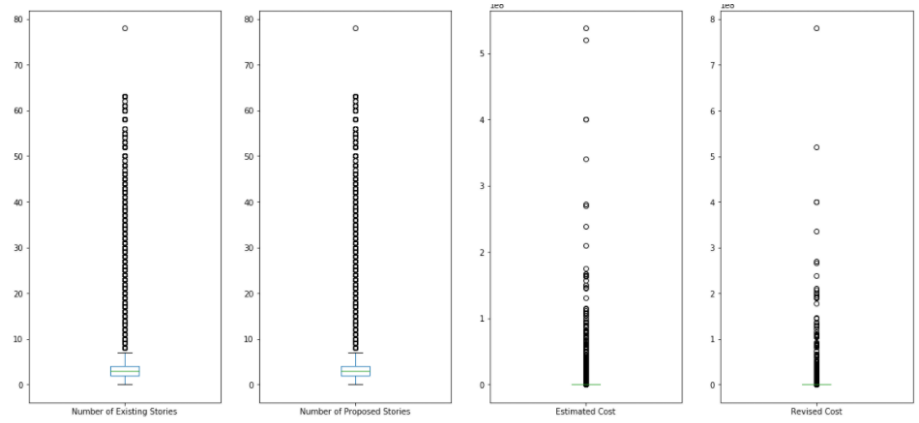


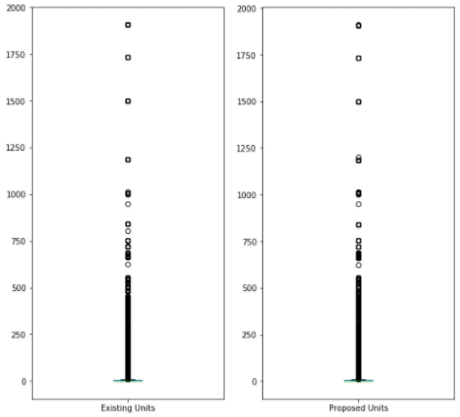
2.3.2 qq 图





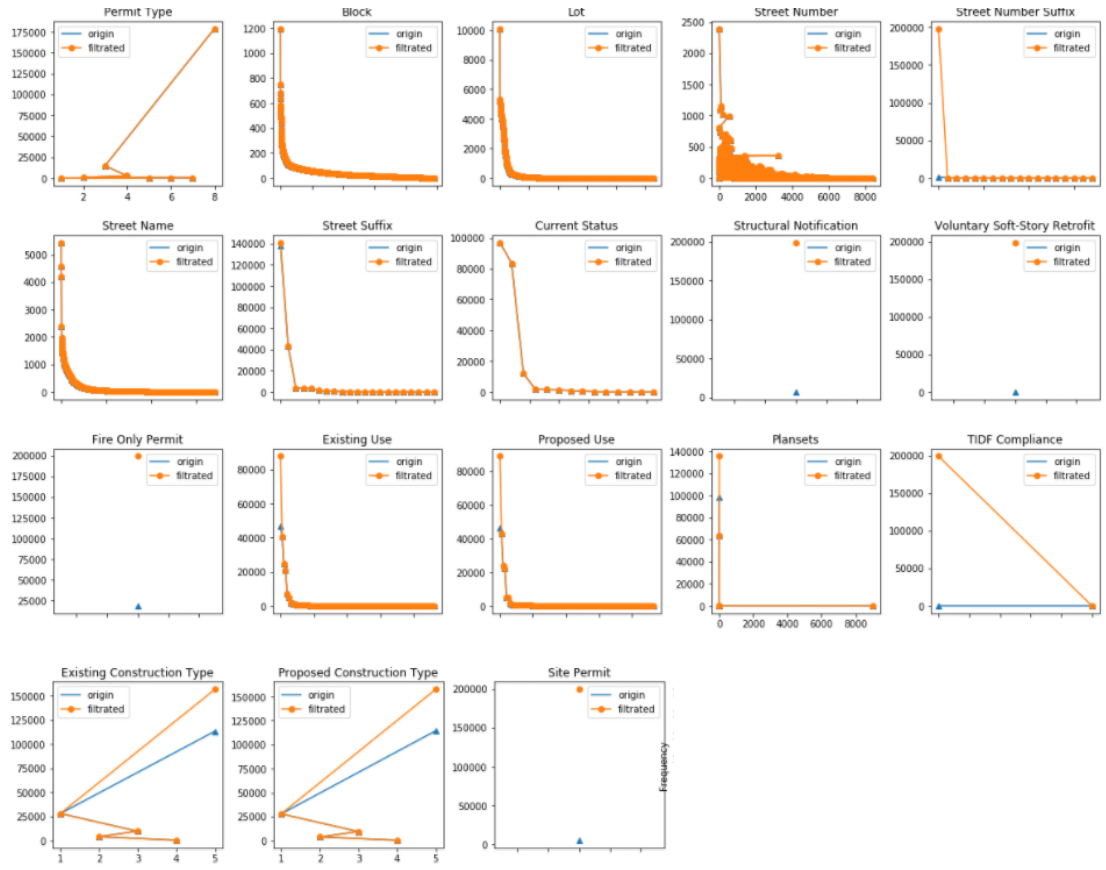
2.3.3 盒图



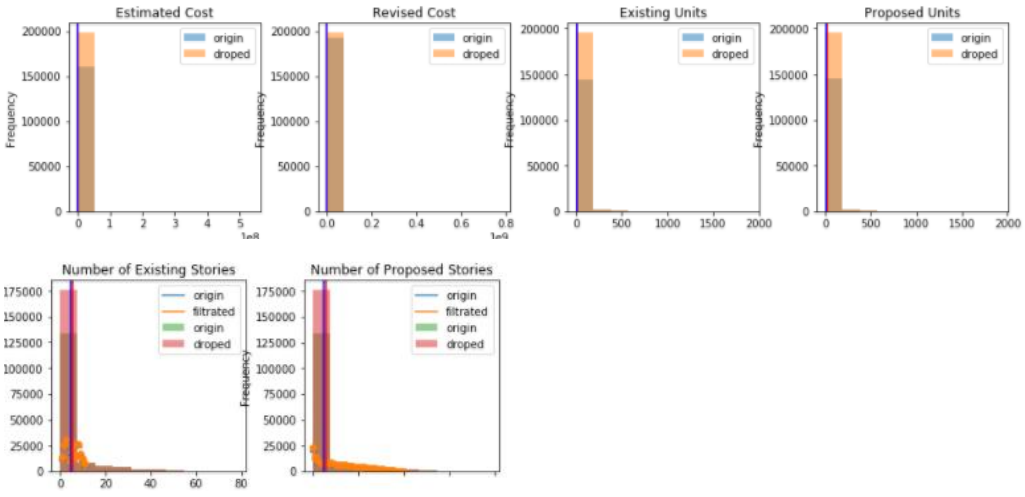


3.1 用最高频率值来填补缺失值

对标称属性使用折线图可视化新旧数据集变化

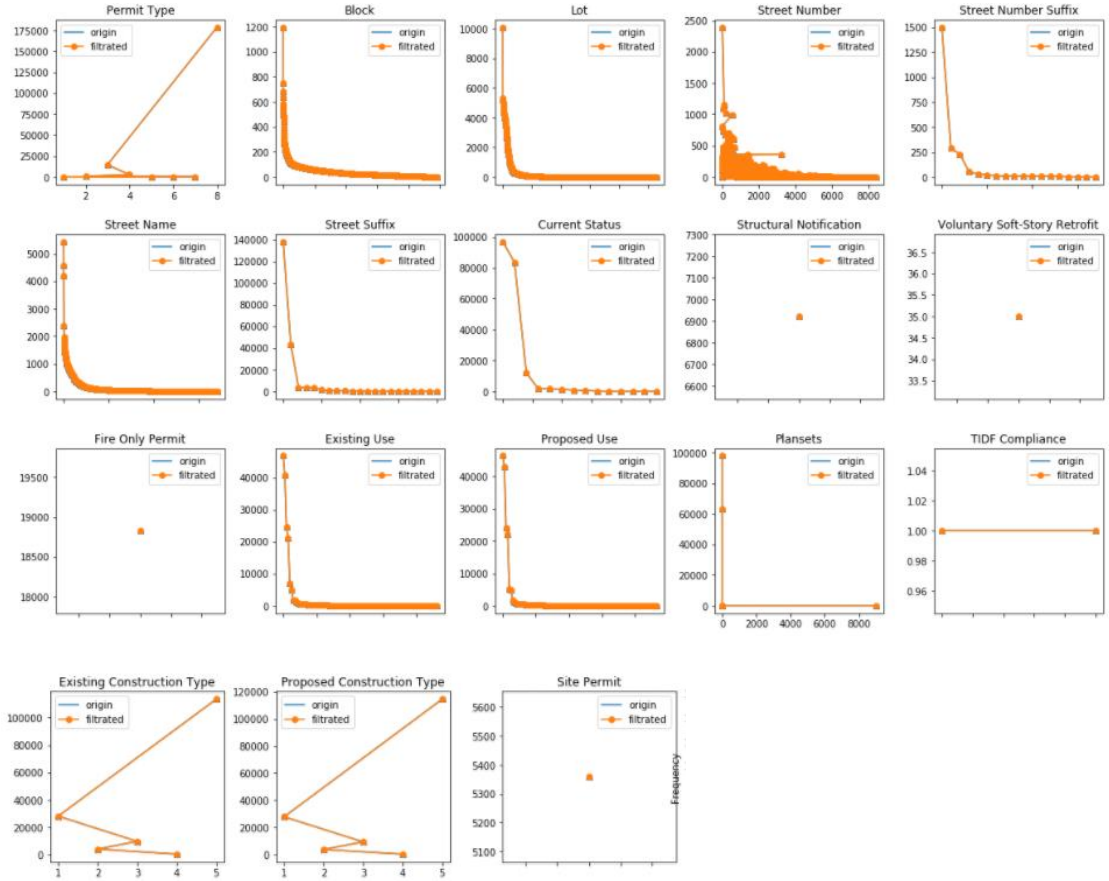


对数值属性使用直方图可视化新旧数据集变化



3.2 用属性间相关关系来填补缺失值

对标称属性使用折线图可视化新旧数据集变化



对数值属性使用直方图可视化新旧数据集变化