数据挖掘大作业一：数据探索性分析与数据预处理

姓名：孙澈

学号：2120171054

1. 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。

2. 数据说明

- 数据集 1: NFL Play-by-Play 2009-2017
- 数据集 2: San Francisco Building Permits

下载数据: 地址

数据集中属性解释：

- 数据集 1: 参考
- 数据集 2: 见下载地址中 `DataDictionaryBuildingPermit.xlsx`

3. 数据分析要求

3.1 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

3.2 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。

4. 提交内容

4.1 分析过程

数据集 1: NFL Play-by-Play 2009-2017

1）数据摘要

NFL Play-by-Play 2009-2017（下简称为 NFL 数据集）本身保存为 csv 格式，是美国橄榄球大联盟 2009-2017 赛季各场比赛的数据统计，根据数据集的属性解释得知该数据集有 102 个属性，包含比赛日期、比赛队伍、得分等信息。根据属性解释文件分析后可以得出，数值属性包括：'TimeUnder'，'TimeSecs'，'PlayTimeDi'，'yrdline100','ydstogo','ydsnet','Yards.Gained','AirYards','YardsAfterCatch','FieldGoal-Distance','Penalty.Yards','PosTeamScore','DefTeamScore','ScoreDiff','AbsScoreDiff'等 42 个属性。标称属性有'GameID','Driver','qtr','down','SideoField','GoalToGO'等 60 个属性。

标称属性的频数：因为标称属性过多，且个别标称属性的类别也过多，在实验中，只统计了类别数小于 1000 的标称属性的频数，存储在\NFL\result\ATTRIBUTES_Nominal 文件夹下，在报告中只列举部分标称属性的统计量。

frequence of RunGap attribute

Type	Description	Count	Percent
	end	31265	35.76%
	guard	27074	30.97%
	tackle	29089	33.27%

frequence of down attribute

Type Description	Count	Percent
1	138878	40.08%
2	104089	30.04%
3	67398	19.45%
4	36169	10.44%

frequence of qtr attribute

Type Description	Count	Percent
1	89176	21.87%
2	112317	27.55%
3	90682	22.24%
4	112641	27.63%
5	2872	0.70%

frequence of PlayType attribute

Type Description	Count	Percent
End of Game	1973	0.48%
Extra Point	10063	2.47%
Field Goal	8928	2.19%
Half End	40	0.01%
Kickoff	23403	5.74%
No Play	21414	5.25%
Pass	159353	39.09%
Punt	22003	5.40%
QB Kneel	3530	0.87%
Quarter End	4914	1.21%
Run	120831	29.64%
Sack	10649	2.61%
Spike	640	0.16%
Timeout	16206	3.98%
Two Minute Warning	3741	0.92%

frequence of PassLength attribute

Type Description	Count	Percent
20	1	0.00%
Deep	32438	19.40%
Short	134729	80.59%

frequence of FieldGoalResult attribute

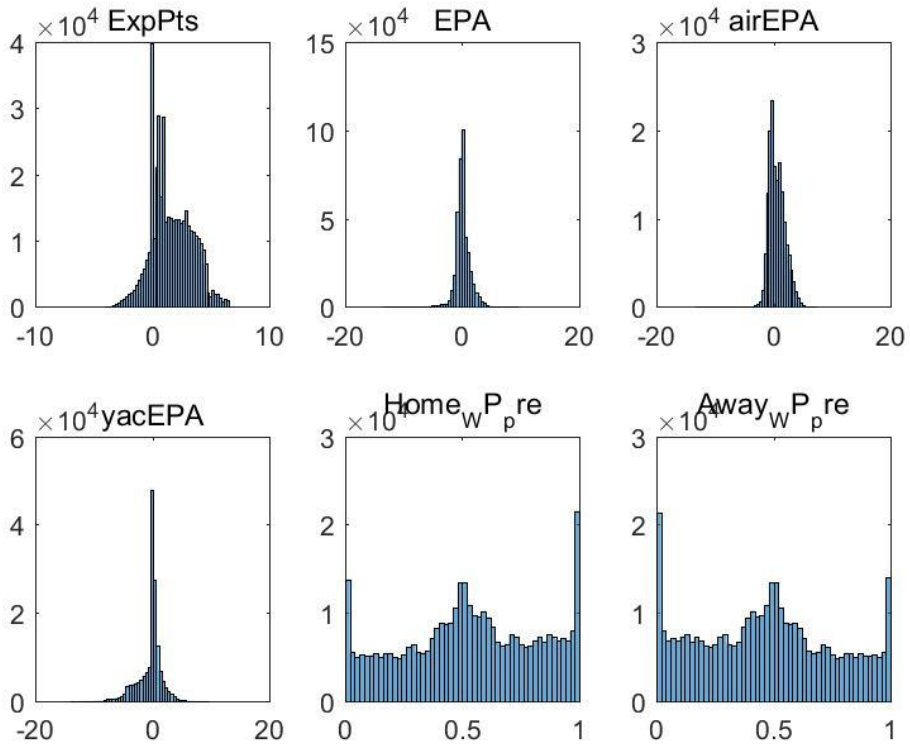
Type Description	Count	Percent
Blocked	205	2.26%
Good	7582	83.70%
No Good	1272	14.04%

数值属性的最大值、最小值、均值、中位数、四分位数、及缺失值的个数(同样由于数据过多，这里只展示部分数值数据，其他的存储在\NFL\data\result\ATTRIBUTES_Nominal\路径下)：

attribute	Maximium	Minimium	Average	Median	Quartile	Missing data
statistics of TimeUnder attribute						
TimeUnder:	15.00000	0.00000	7.37420	7.00000	3.00000,11.00000	0.00000
statistics of TimeSecs attribute						
TimeSecs:	3600.00000	-900.00000	1695.26894	1800.00000	778.00000,2585.00000	224.00000
statistics of PlayTimeDiff attribute						
PlayTimeDiff:	943.00000	0.00000	20.57676	17.00000	5.00000,37.00000	444.00000
statistics of yrdln attribute						
yrdln:	50.00000	1.00000	28.48833	30.00000	20.00000,39.00000	840.00000
statistics of yrdline100 attribute						
yrdline100:	99.00000	1.00000	48.64408	49.00000	30.00000,70.00000	840.00000
statistics of ydstogo attribute						
ydstogo:	50.00000	0.00000	7.30940	9.00000	3.00000,10.00000	0.00000
statistics of ydsnet attribute						
ydsnet:	99.00000	-87.00000	25.94552	19.00000	5.00000,43.00000	0.00000
statistics of Yards Gained attribute						

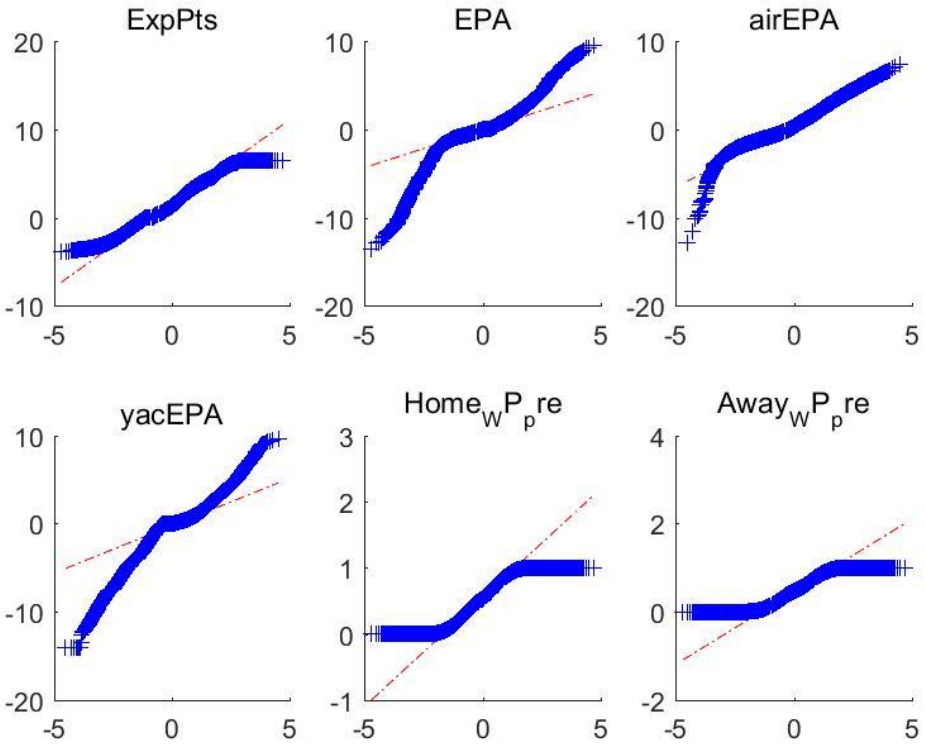
2）数据可视化

这里，我们随机选取 ExpPts、EPA、airEPA、yacEPA、HomewPpre 和 AwaywPpre 六个数值属性进行分析。因此，在数值属性本身存在的缺失情况下的直方图（存储在\NFL\result\figure）：

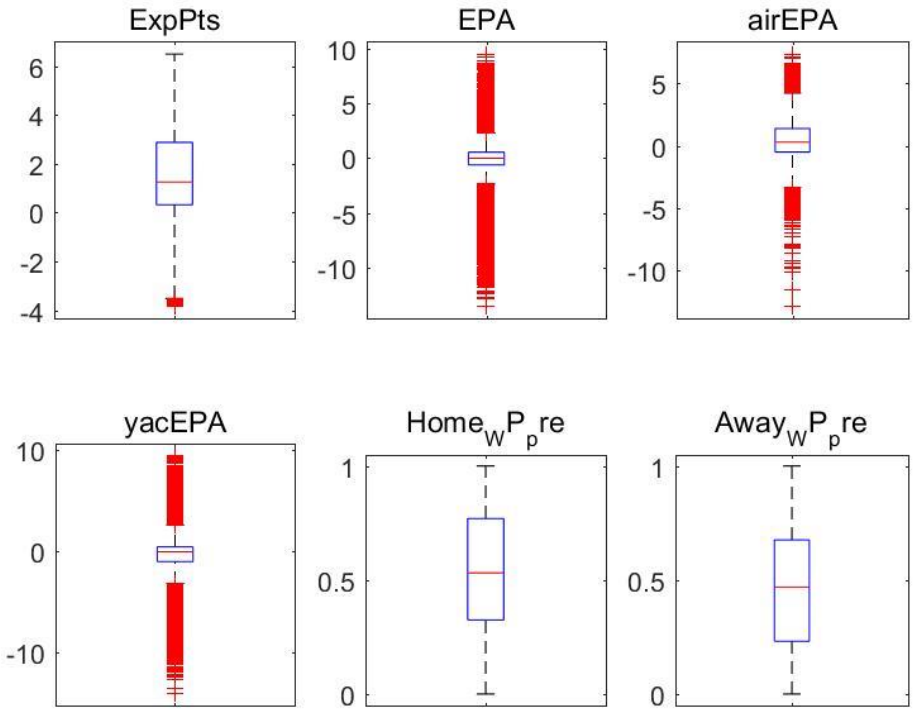


从上图可以看出 ExpPts、EPA、airEPA 和 yacEPA 均接近正态分布，并且均值大概都在 0 附近，HomewPpre 和 AwaywPpre 的分布类似，类似均匀分布，但是在最大值和最小值附近分布较为密集。 从 QQ 图也可以判定其和正态分布的相似度。

QQ 图：可以看出前四个个属性更符合正态分布。



盒图可以看出这些数值属性的分布情况，可以看出 ExpPts、HomewPpre 和 AwaywPp 的离散值比较小。

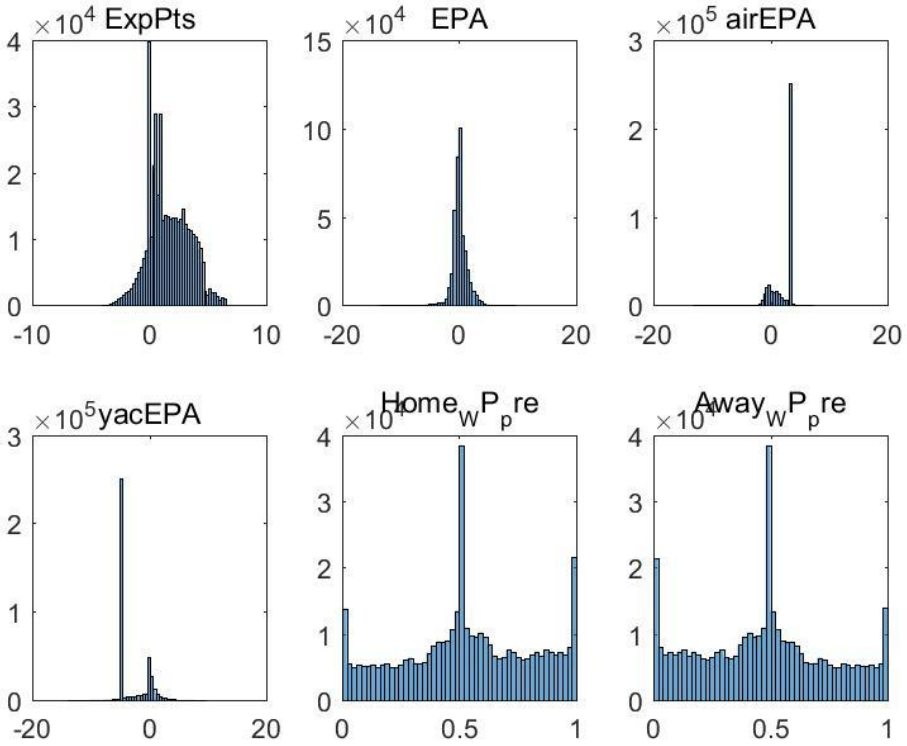


：

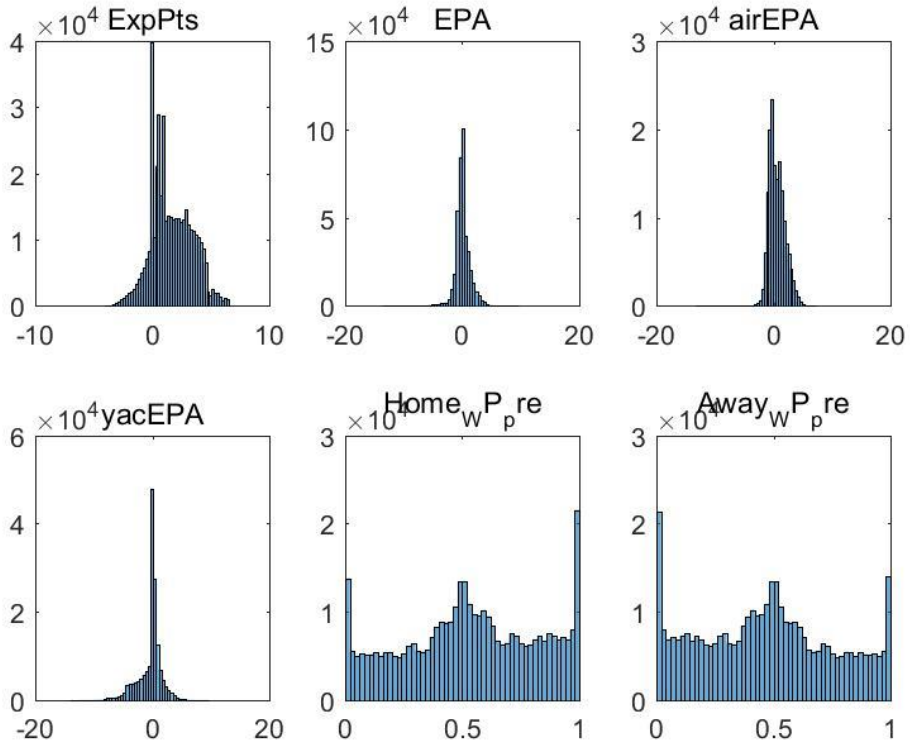
3）数据集预处理

1. 剔除缺省值的操作已经在之前做好， 在此不再赘述。

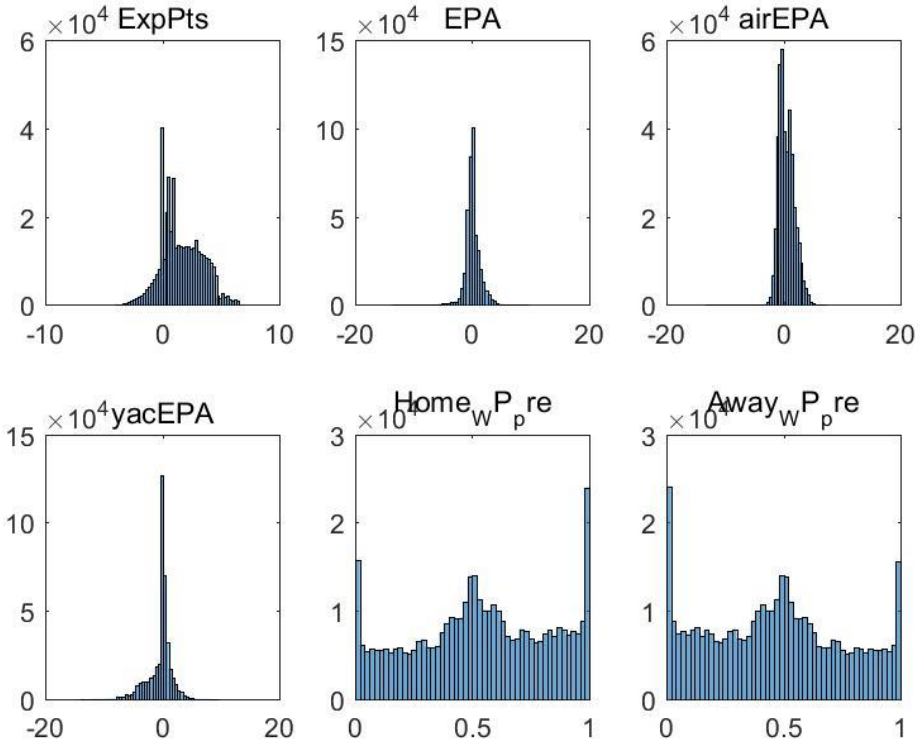
2. 用最高频率值来填补缺失值： 在此属性中的缺失值用此属性中所计算出的最高频率值填补。原始直方图以图如下，可以看出，填补后的几个属性在某个区域的分布会激增，这是因为缺失值较多的原因，用众数填补显然会导致这种情况。



3. 通过属性的相关关系来填补缺失值：计算两个属性的相关性，相关性越大越可以根据另一个属性推断缺失属性的值。 通过另一属性的回归分析，计算当前的缺失值。直方图如下：



4. 通过数据对象之间的相似性来填补缺失值： 计算两个样本的相似程度， 越相似证明越可以使用该样本推断当前含缺失。直方图如下：



由以上四幅图比较可知，按相似性填补的结果和处理前更相似（前四个属性更符合正态分布），相似性填补尤其是第三种填补方法，符合原来的分布情况，从原始的直方图也可以看出来按用最高频率值来填补缺失值的结果和处理前差别最大。

4.2 分析程序

程序由三部分组成，NFL\main.m、NFL\mypreprocessing.m、NFL\myvisualization.m。下面分别简单介绍它们的功能（具体见代码文件）。

1. main.m
   （1） 读入数据并保存成字符串元胞格式。
   （2） 标记数值属性和标称属性。
   （3） 处理标称数据（数据处理如格式变换、去除缺失值、统计标称属性有哪些取值 unique_current_attribute，存在 NFL\result\ATTRIBUTES_Nominal 文件夹下、输出统计量）。标称属性对应的标签在\NFL\data\ Nominal_Label\
   （4） 处理数值属性（数据处理如格式变换、去除缺失值、输出统计量），存储在\NFL\resultATTRBUTES_Number
   （5） myvisualization(total_attribute_number,attribute);可视化原始数据。
   （6） 填补缺失值并存储及可视化，输入 1：最大频率填补（存储在 NFL\result\building_permits_filled_by_maximium.txt），2：属性的相关关系来填补（存储在 NFL\result\building_permits_filled_by_attribute.txt），3：数据对象之间的相似性来填补（存储在 NFL\result\building_permits_filled_by_similarity.txt），其他：退出。

2．myvisualization.m 绘制 6 个属性的直方图、QQ 图、盒图,存储在 NFL\result\figure

3. mypreprocessing ,PREPROCESSING 对数据进行预处理，支持三种方式（method 取值为 1~3）。三种方式分别为：最高频值替代，属性相关关系，对象相似性。将缺失部分剔除是缺省方式，不在预处理函数中。处理后的数据库存储在 NFL\result\的三个 txt 文件中。