数据挖掘大作业一：数据探索性分析与数据预处理

姓名:康杨

学号:2120171024

1. 问题描述

对数据集进行探索性分析与预处理。

1. 数据说明

数据集: NFL Play-by-Play 2009-2017

1. 数据分析

3.1数据可视化和摘要

3.1.1数据摘要

* 标称属性,给出其每个可能取值的频数

由于数据量较大,数据结果保存在/NFL/result\_NFL\_nominal.txt中

* 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。

GameID Drive qtr down \

count 4.076880e+05 407688.000000 407688.000000 346534.000000

mean 2.013158e+09 12.316158 2.577412 2.002476

std 2.572839e+06 7.149527 1.129750 1.006353

min 2.009091e+09 1.000000 1.000000 1.000000

25% 2.011101e+09 6.000000 2.000000 1.000000

50% 2.013111e+09 12.000000 3.000000 2.000000

75% 2.015121e+09 18.000000 4.000000 3.000000

max 2.017123e+09 35.000000 5.000000 4.000000

TimeUnder TimeSecs PlayTimeDiff yrdln \

count 407688.000000 407464.000000 407244.000000 406848.000000

mean 7.374200 1695.268944 20.576762 28.488327

std 4.642388 1062.801012 17.969326 12.946471

min 0.000000 -900.000000 0.000000 1.000000

25% 3.000000 778.000000 5.000000 20.000000

50% 7.000000 1800.000000 17.000000 30.000000

75% 11.000000 2585.000000 37.000000 39.000000

max 15.000000 3600.000000 943.000000 50.000000

yrdline100 ydstogo ... yacEPA \

count 406848.000000 407688.000000 ... 159190.000000

mean 48.644081 7.309403 ... -0.386086

std 25.070416 4.869987 ... 1.972715

min 1.000000 0.000000 ... -14.000000

25% 30.000000 3.000000 ... -0.961115

50% 49.000000 9.000000 ... 0.000000

75% 70.000000 10.000000 ... 0.485508

max 99.000000 50.000000 ... 9.559834

Home\_WP\_pre Away\_WP\_pre Home\_WP\_post Away\_WP\_post \

count 382734.000000 382734.000000 381101.000000 381101.000000

mean 0.534488 0.465965 0.534791 0.465613

std 0.285574 0.285629 0.287818 0.287867

min 0.000000 0.000000 0.000000 0.000000

25% 0.325123 0.231411 0.321701 0.227694

50% 0.531274 0.469052 0.533609 0.466670

75% 0.769232 0.675530 0.772882 0.678833

max 1.000000 1.000000 1.000000 1.000000

Win\_Prob WPA airWPA yacWPA \

count 382679.000000 402147.000000 159187.000000 158926.000000

mean 0.501320 0.002099 0.015135 -0.010480

std 0.287445 0.045363 0.056490 0.068139

min 0.000000 -0.997214 -0.999881 -0.986673

25% 0.276472 -0.014728 -0.011518 -0.018683

50% 0.504470 0.000000 0.003441 0.000000

75% 0.725477 0.014684 0.035792 0.011431

max 1.000000 0.994848 0.994848 1.000000

Season

count 407688.000000

mean 2013.018985

std 2.576962

min 2009.000000

25% 2011.000000

50% 2013.000000

75% 2015.000000

max 2017.000000

[8 rows x 64 columns]

3.1.2数据可视化

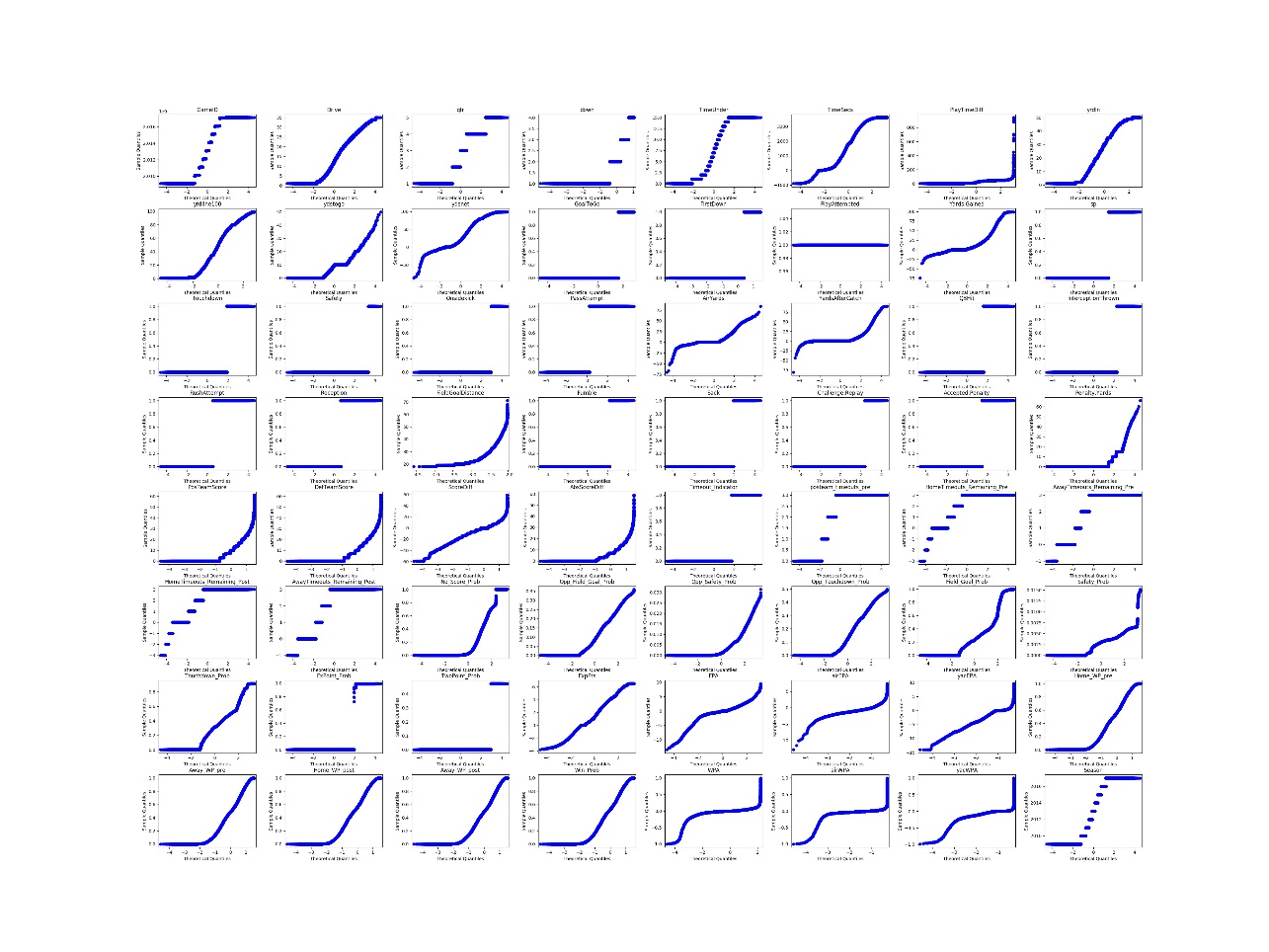
针对数值属性，

* 绘制直方图，用qq图检验其分布是否为正态分布。

直方图如下所示:



qq图如下所示:



由各个属性的qq图可以看出,属性ExpPts和EPA满足正态分布

* 绘制盒图，对离群值进行识别

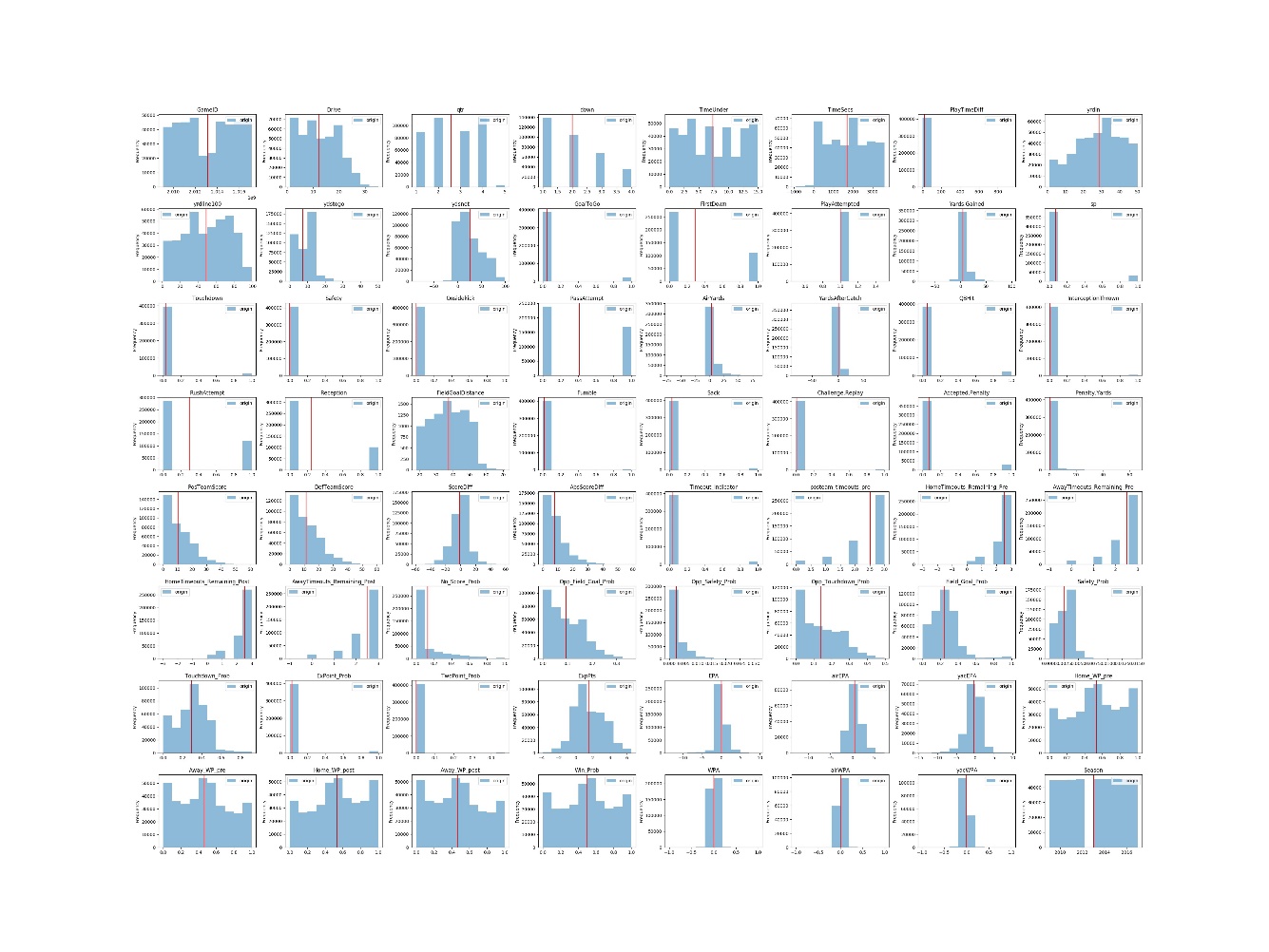
盒图如下所示:



从各个属性的盒图观察可得,属性PlayTimeDiff、ydstogo、ydsnet、GoalToGo、Yards.Gained、sp、Touchdown、Safety、Onsidekick、AirYards、YardsAfterCatch、QBHit、Interception Thrown、Reception、Fumble、Sack、Challenge.Replay、Accepted.Penalty、Penalty.Yards、PosTeamScore、DefTeamScore、ScoreDiff、AbsScoreDiff、Timeout\_Indicator、posteam\_timeouts\_pre、HomeTimeouts\_Remaining\_Pre、AwayTimeouts\_Remaining\_Pre、HomeTimeouts\_Remaining\_Post、AwayTimeouts\_Remaining\_Post、No\_Score\_Prob、Opp\_Field\_Goal\_Prob、Opp\_Safety\_Prob、Field\_Goal\_Prob、Safety\_Prob、Touchdown\_Prob、ExPoint\_Prob、TwoPoint\_Prob、ExpPts、EPA、airEPA、yacEPA、WPA、airWPA、yacWPA存在离群值

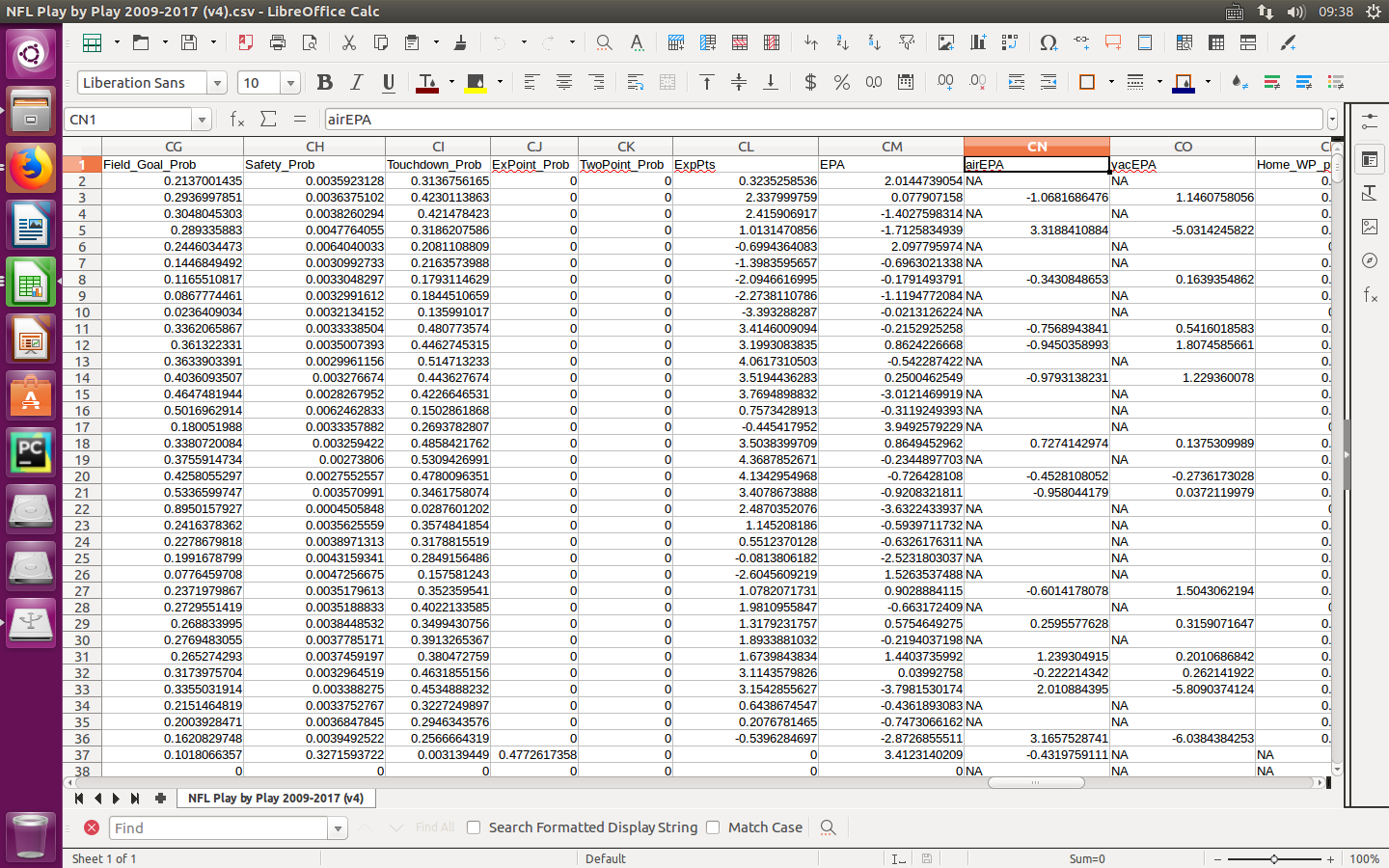
3.2数据缺失处理

* 将缺失部分剔除

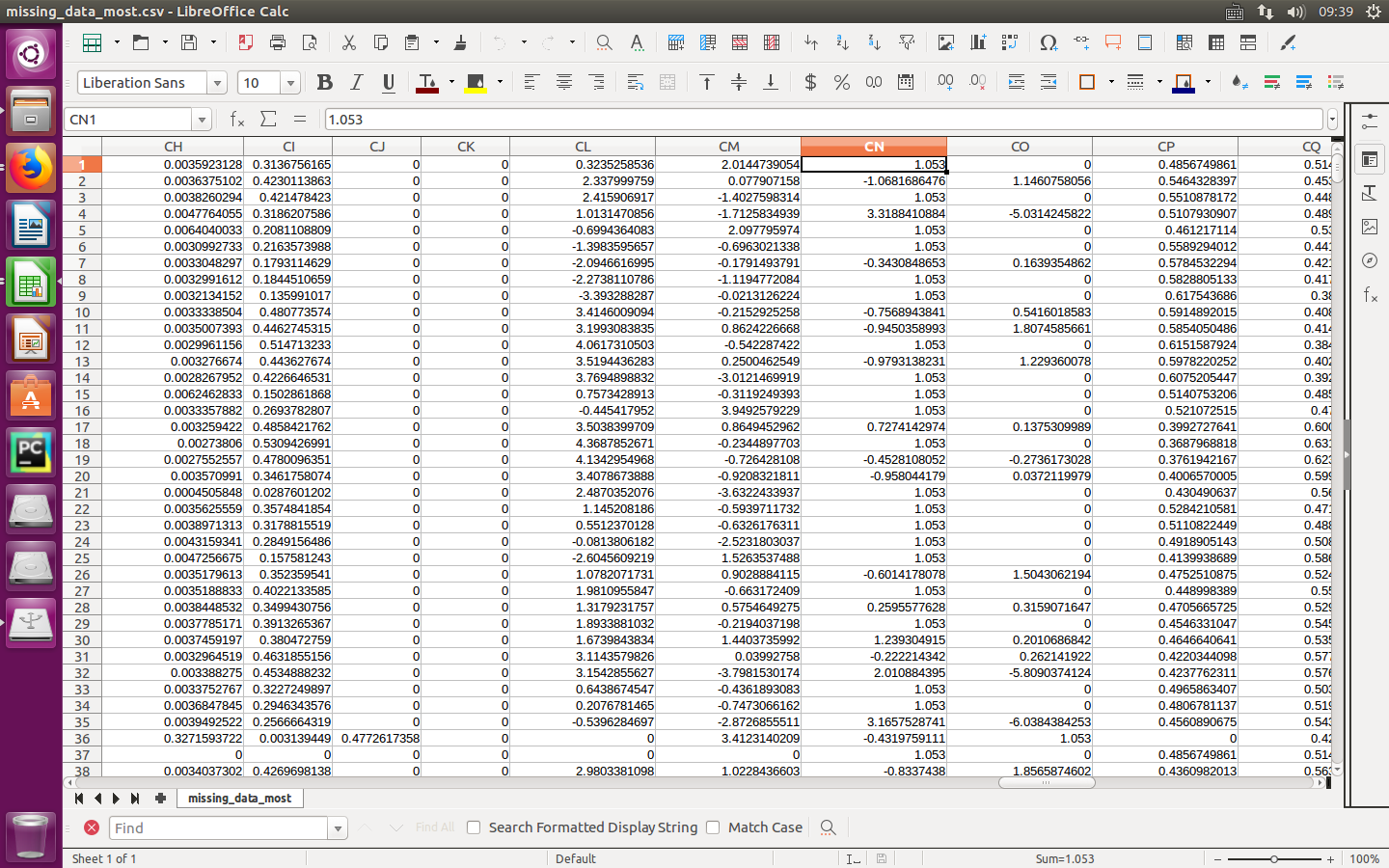


* 用最高频率值来填补缺失值

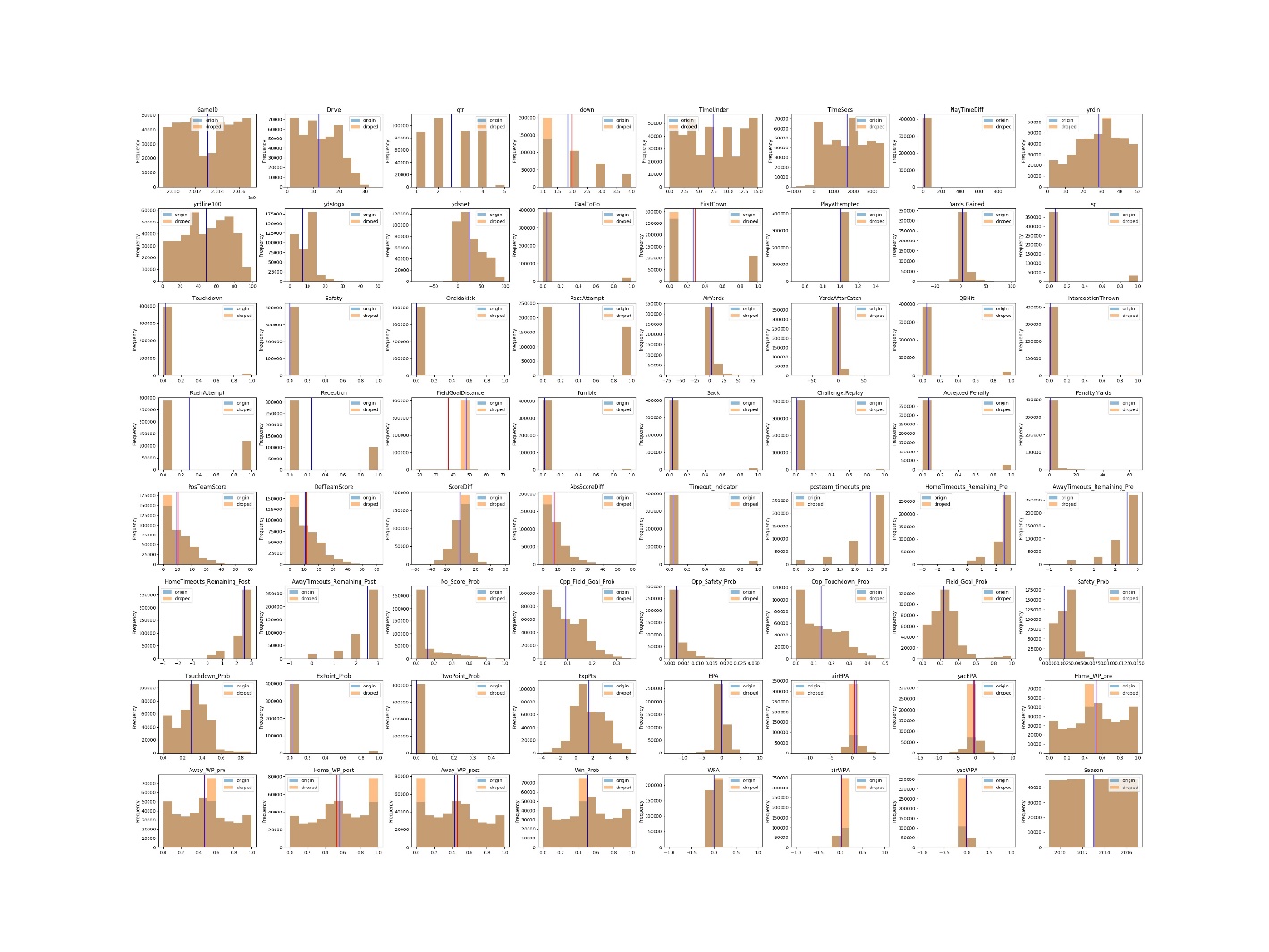
填补前,csv数据文件部分情况如下所示:



填补后,csv数据文件部分情况如下所示:

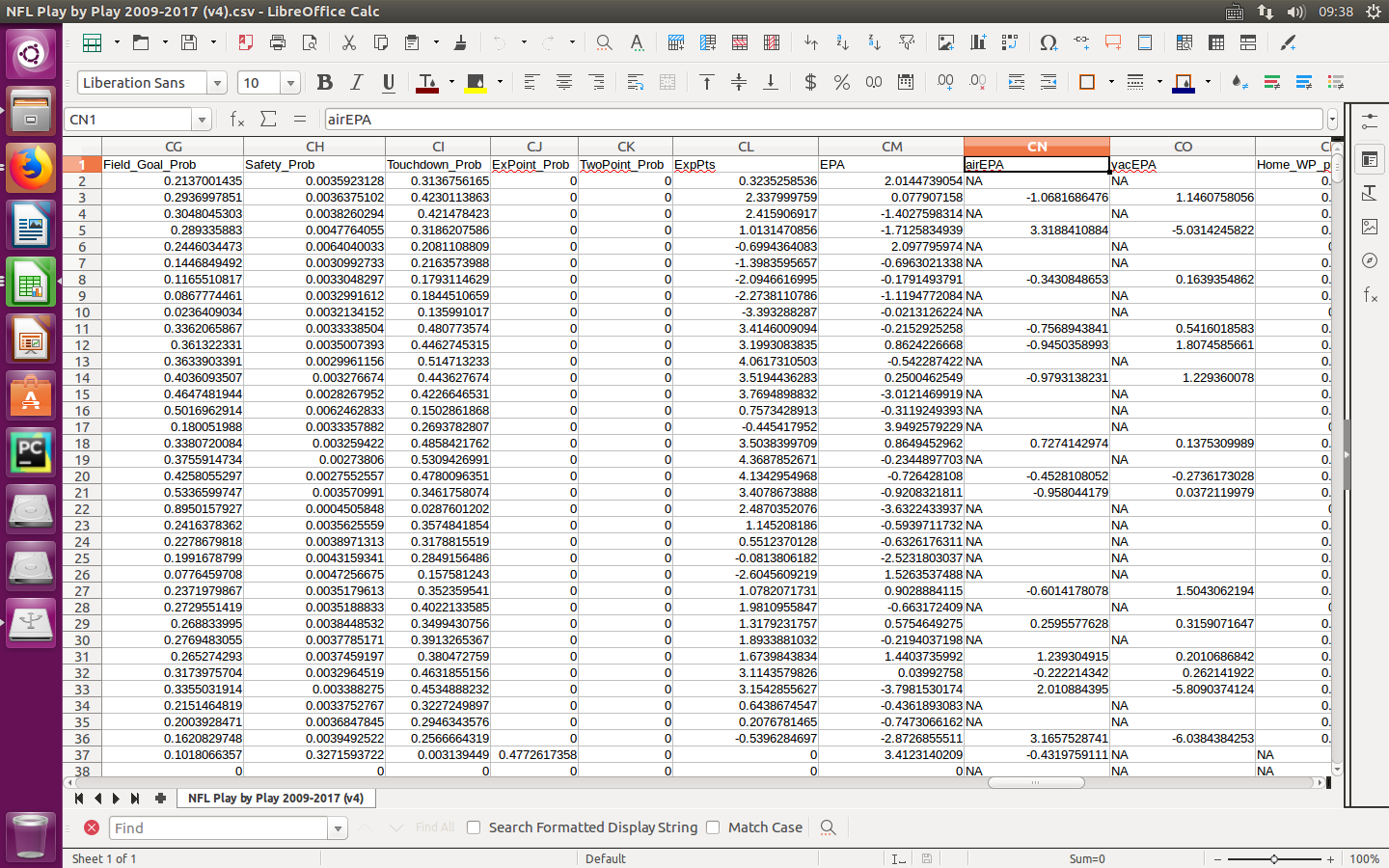


可视化对比填补前后数据,直方图如下所示:

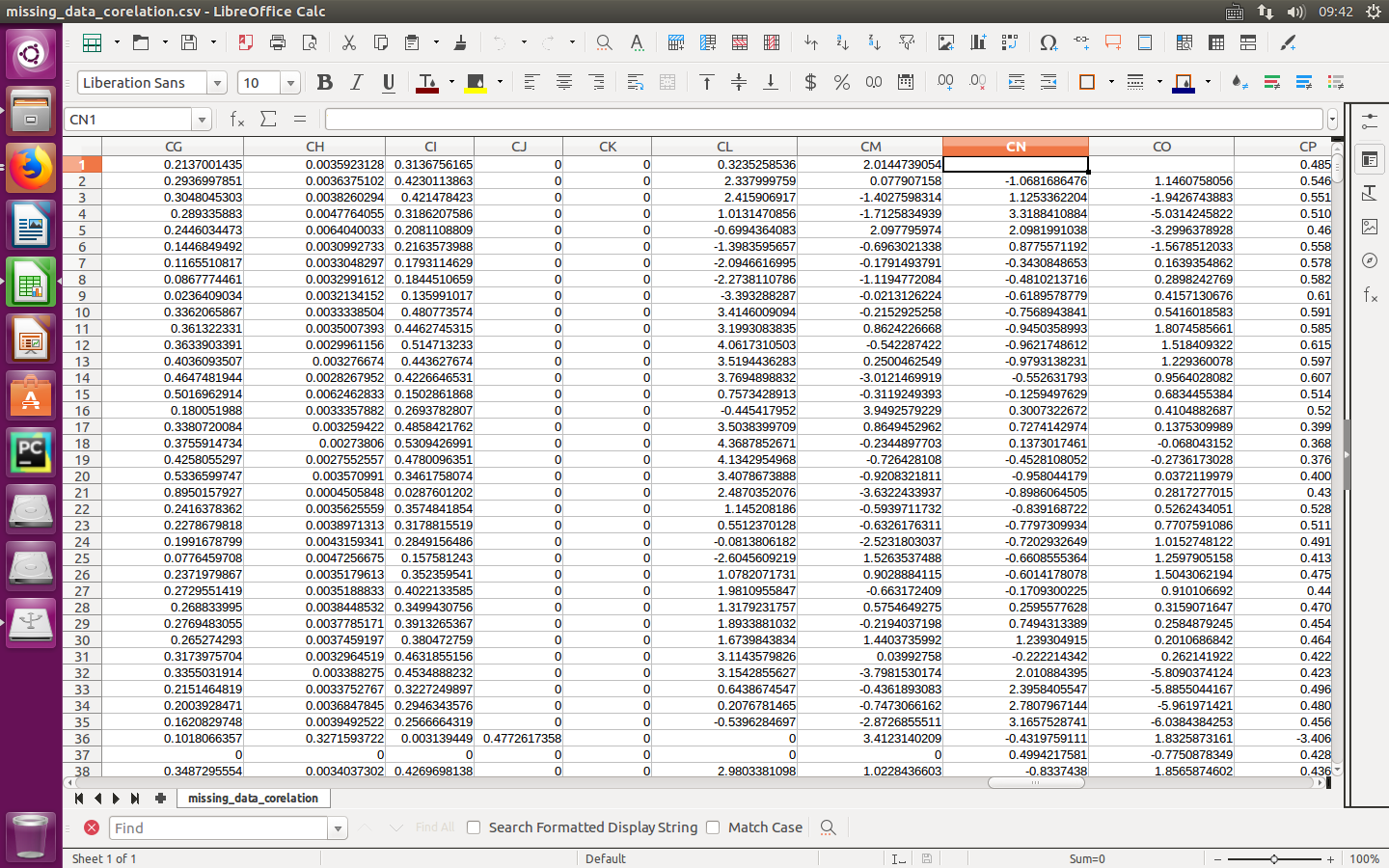


* 通过属性的相关关系来填补缺失值

填补前,csv数据文件部分情况如下所示:



填补后,csv数据文件部分情况如下所示:



可视化对比填补前后数据,直方图如下所示:



* 通过数据对象之间的相似性来填补缺失值