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。

2. 数据说明

数据集: NFL Play-by-Play 2009-2017

3. 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

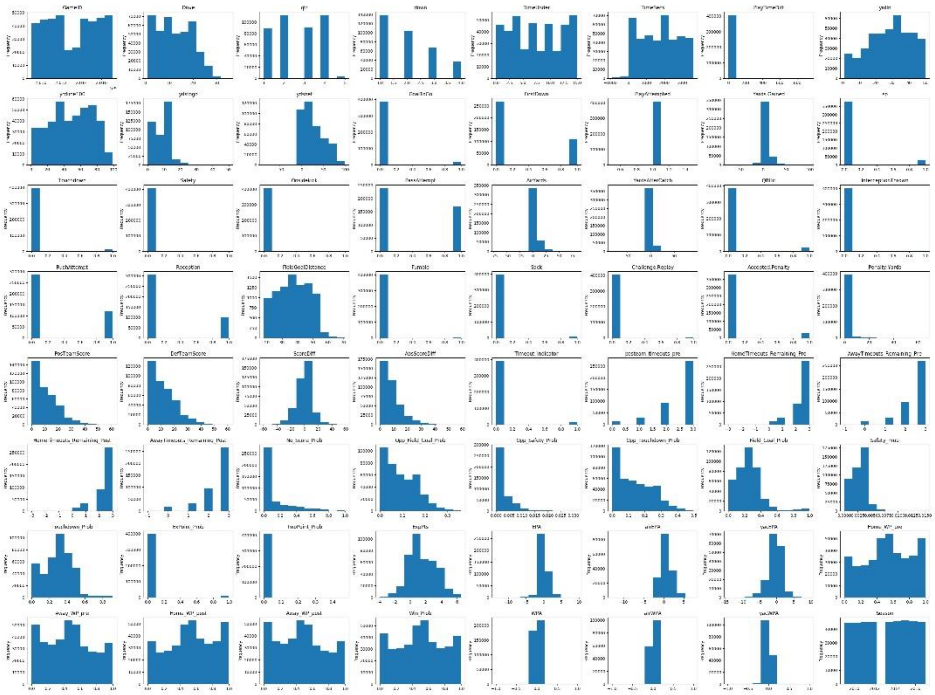
	Season
count	407688.000000
mean	2013.018985
std	2.576962
min	2009.000000
25%	2011.000000
50%	2013.000000
75%	2015.000000
max	2017.000000

[8 rows x 64 columns]

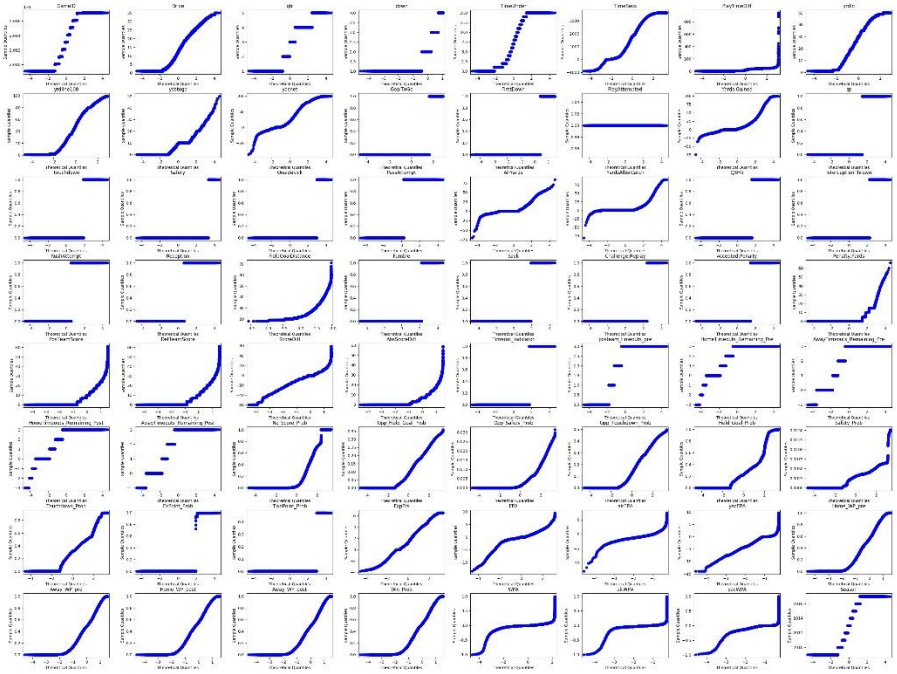
3.1.2 数据可视化

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:

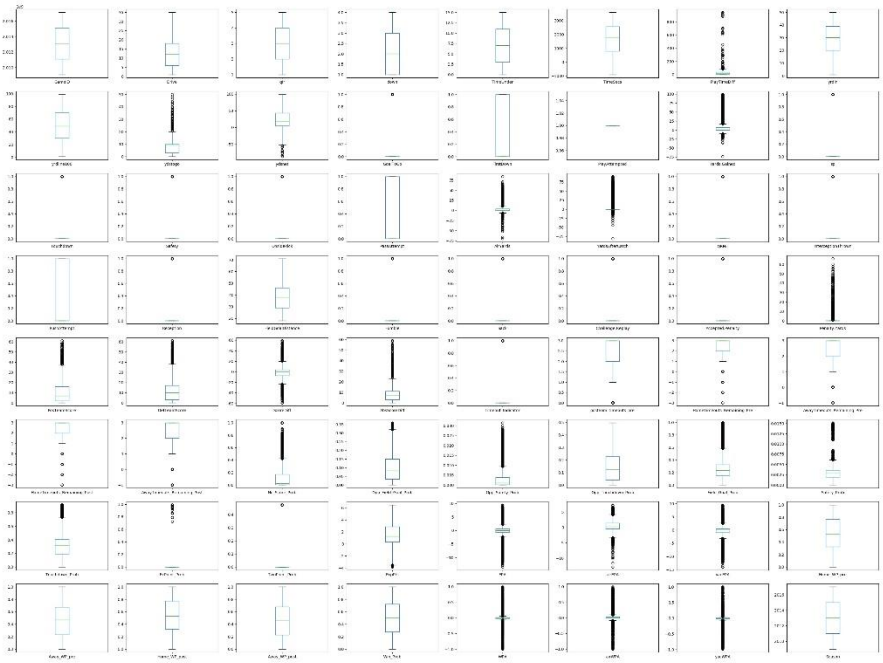


qq 图如下所示:



由各个属性的 qq 图可以看出,属性 ExpPts 和 EPA 满足正态分布

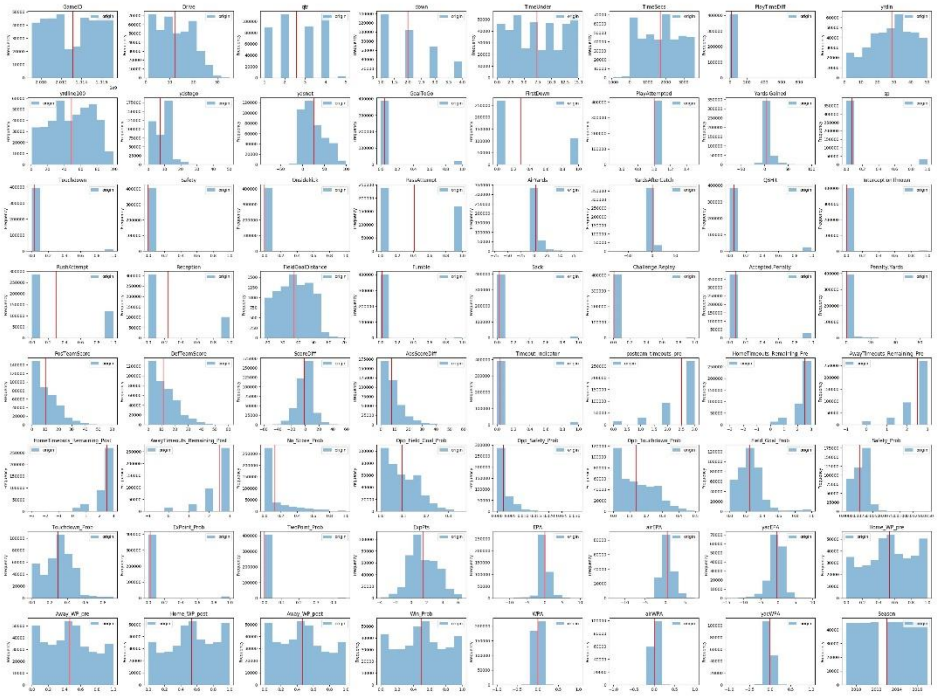
- 绘制盒图，对离群值进行识别
  盒图如下所示:



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

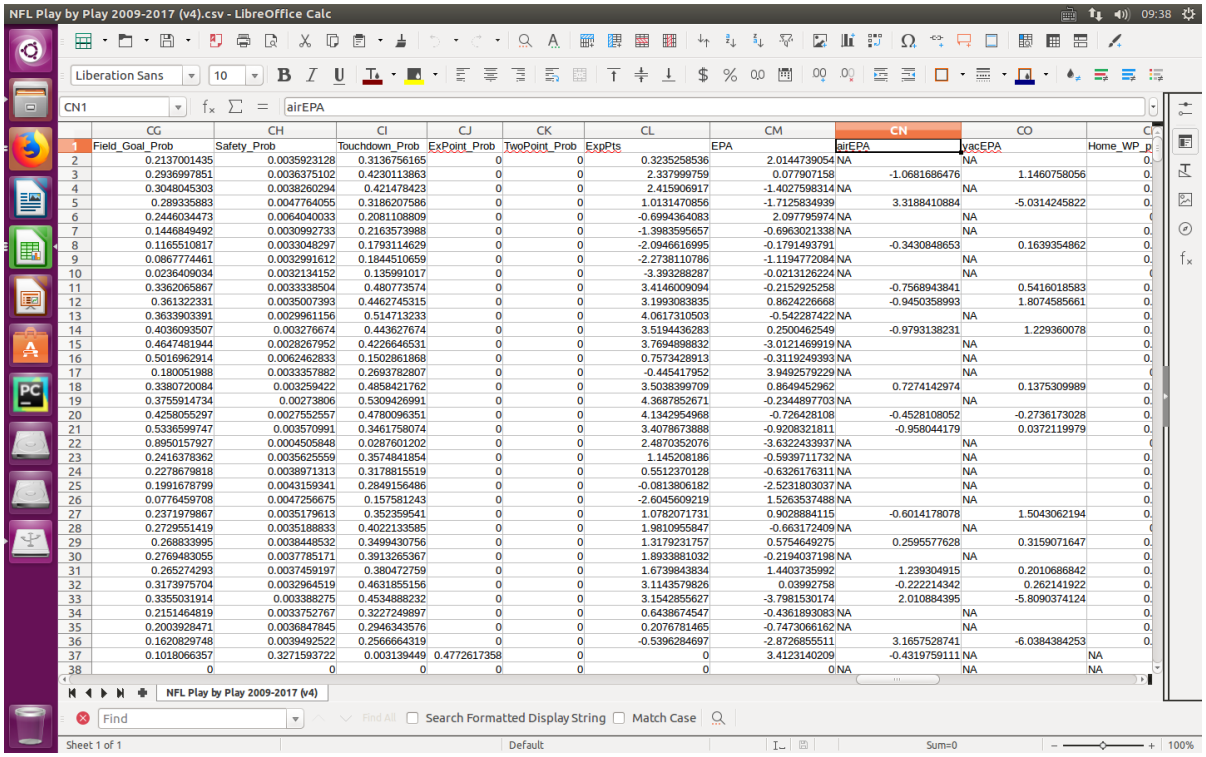
3.2 数据缺失处理

- 将缺失部分剔除



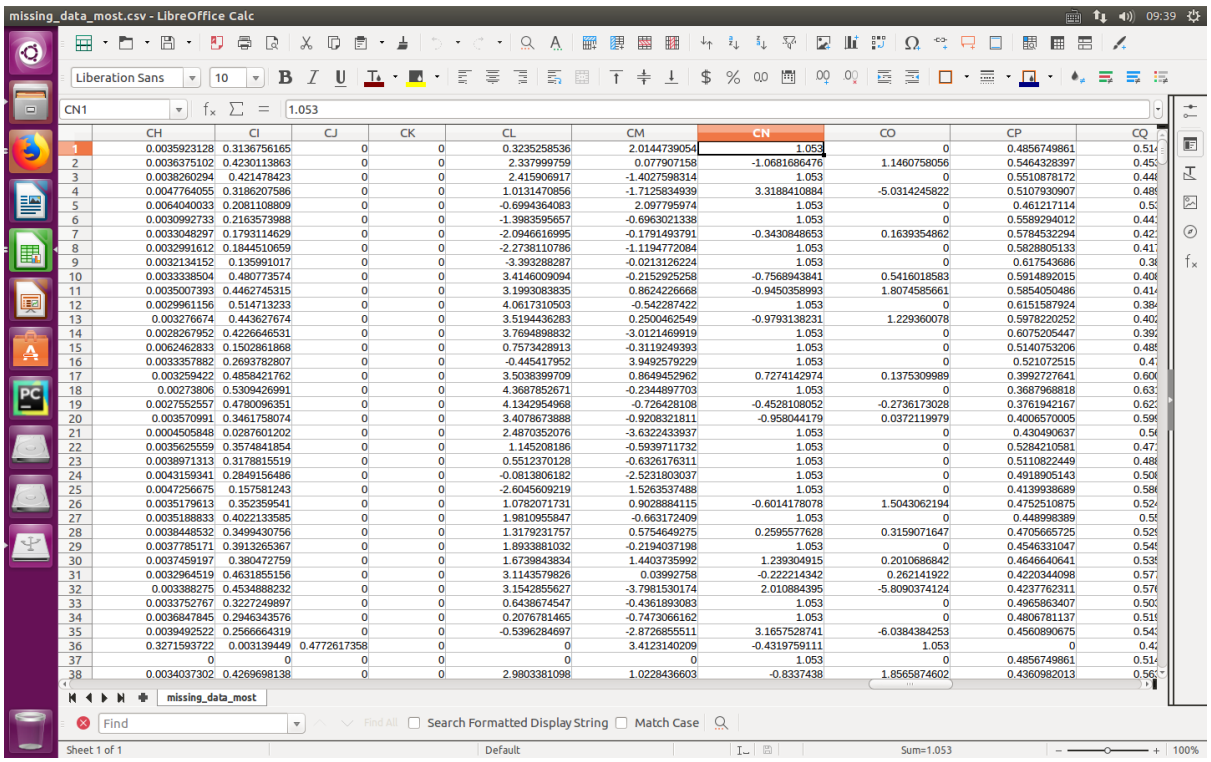
- 用最高频率值来填补缺失值

填补前,csv 数据文件部分情况如下所示:

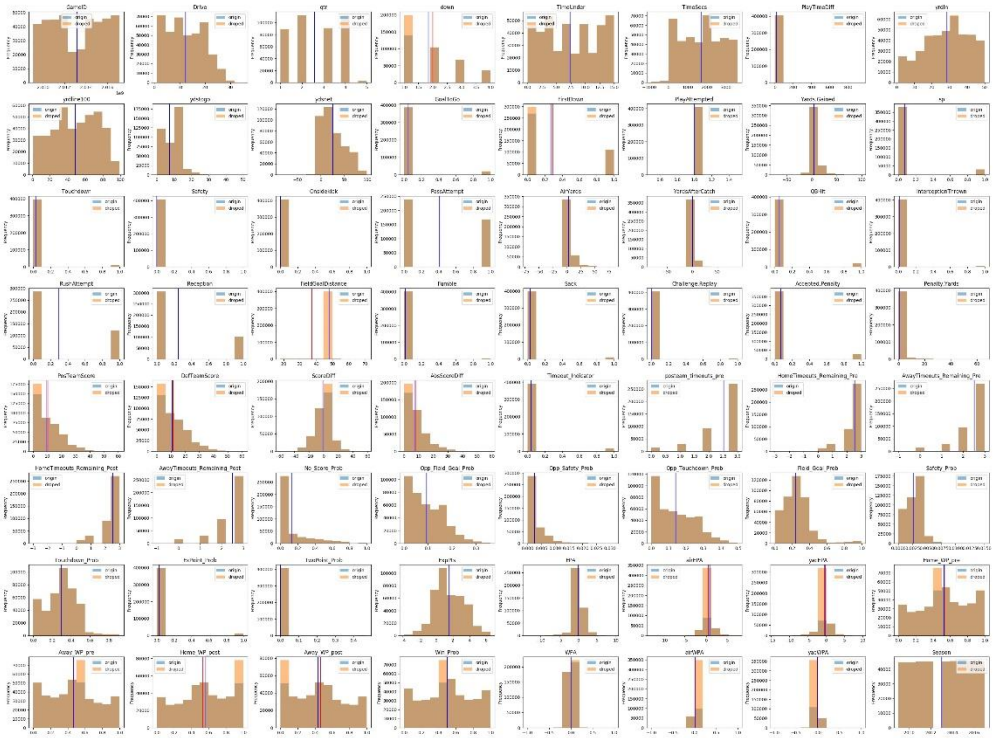


	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP
1	Field_Goal_Prob	Safety_Prob	Touchdown_Prob	ExPoint_Prob	TwoPoint_Prob	ExpPts	EPA	airEPA	yacEPA	Home_WP_p
2	0.2137001435	0.0035923128	0.3136756165	0	0	0.3235258536	2.0144739054	NA	NA	0.
3	0.2936997851	0.0036375102	0.4230113863	0	0	2.337999759	0.077907158	-1.0681686476	1.1460758056	0.
4	0.3048045303	0.0038260294	0.421478423	0	0	2.415906917	-1.4027598314	NA	NA	0.
5	0.289335883	0.0047764055	0.3186207586	0	0	1.0131470856	-1.7125834939	3.3188410884	-5.0314245822	0.
6	0.2446034473	0.0064040033	0.2081108809	0	0	-0.6994364083	2.097795974	NA	NA	0
7	0.1446849492	0.0030992733	0.2163573988	0	0	-1.3983595657	-0.6963021338	NA	NA	0.
8	0.1165510817	0.0033048297	0.1793114629	0	0	-2.0946616995	-0.1791493791	-0.3430848653	0.1639354862	0.
9	0.0867774461	0.0032991612	0.1844510659	0	0	-2.2738110786	-1.1194772084	NA	NA	0.
10	0.0236409034	0.0032134152	0.135991017	0	0	-3.393288287	-0.0213126224	NA	NA	0
11	0.3362065867	0.0033338504	0.480773574	0	0	3.4146009094	-0.2152925258	-0.7568943841	0.5416018583	0.
12	0.361322331	0.0035007393	0.4462745315	0	0	3.1993083835	0.8624226668	-0.9450358993	1.8074585661	0.
13	0.3633903391	0.0029961156	0.514713233	0	0	4.0617310503	-0.542287422	NA	NA	0.
14	0.4036093507	0.003276674	0.443627674	0	0	3.5194436283	0.2500462549	-0.9793138231	1.229360078	0.
15	0.4647481944	0.0028267952	0.4226646531	0	0	3.7694898832	-3.0121469919	NA	NA	0.
16	0.5016962914	0.0062462833	0.1502861868	0	0	0.7573428913	-0.3119249393	NA	NA	0.
17	0.180051988	0.0033357882	0.2693782807	0	0	-0.445417952	3.9492579229	NA	NA	0
18	0.3380720084	0.003259422	0.4858421762	0	0	3.5038399709	0.8649452962	0.7274142974	0.1375309989	0.
19	0.3755914734	0.00273806	0.5309426991	0	0	4.3687852671	-0.2344897703	NA	NA	0.
20	0.4258055297	0.0027552557	0.4780096351	0	0	4.1342954968	-0.726428108	-0.4528108052	-0.2736173028	0.
21	0.5336599747	0.003570991	0.3461758074	0	0	3.4078673888	-0.9208321811	-0.958044179	0.0372119979	0.
22	0.8950157927	0.0004505848	0.0287601202	0	0	2.4870352076	-3.6322433937	NA	NA	0
23	0.2416378362	0.0035625559	0.3574841854	0	0	1.145208186	-0.5939711732	NA	NA	0.
24	0.2278679818	0.0038971313	0.3178815519	0	0	0.5512370128	-0.6326176311	NA	NA	0.
25	0.1991678799	0.0043159341	0.2849156486	0	0	-0.0813806182	-2.5231803037	NA	NA	0.
26	0.0776459708	0.0047256675	0.157581243	0	0	-2.6045609219	1.5263537488	NA	NA	0.
27	0.2371979867	0.0035179613	0.352359541	0	0	1.0782071731	0.9028884115	-0.6014178078	1.5043062194	0.
28	0.2729551419	0.0035188833	0.4022133585	0	0	1.9810955847	-0.663172409	NA	NA	0
29	0.268833995	0.0038448532	0.3499430756	0	0	1.3179231757	0.5754649275	0.2595577628	0.3159071647	0.
30	0.2769483055	0.0037785171	0.3913265367	0	0	1.8933881032	-0.2194037198	NA	NA	0.
31	0.265274293	0.0037459197	0.380472759	0	0	1.6739843834	1.4403735992	1.239304915	0.2010686842	0.
32	0.3173975704	0.0032964519	0.4631855156	0	0	3.1143579826	0.03992758	-0.222214342	0.262141922	0.
33	0.3355031914	0.003388275	0.4534888232	0	0	3.1542855627	-3.7981530174	2.010884395	-5.8090374124	0.
34	0.2151464819	0.0033752767	0.3227249897	0	0	0.6438674547	-0.4361893083	NA	NA	0.
35	0.2003928471	0.0036847845	0.2946343576	0	0	0.2076781465	-0.7473066162	NA	NA	0.
36	0.1620829748	0.0039492522	0.2566664319	0	0	-0.5396284697	-2.8726855511	3.1657528741	-6.0384384253	0.
37	0.1018066357	0.3271593722	0.009139449	0.4772617358	0	0	3.4123140209	-0.4319759111	NA	NA
38	0	0	0	0	0	0	0	NA	NA	NA

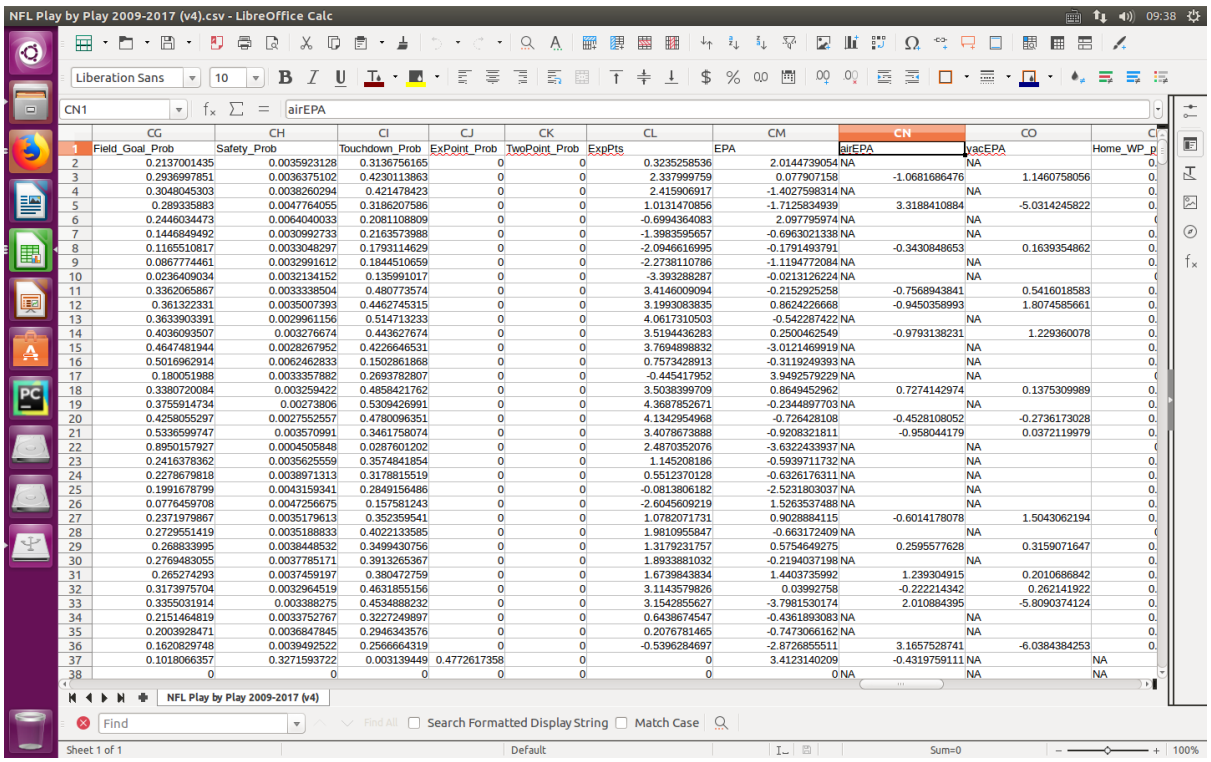
填补后,csv 数据文件部分情况如下所示:



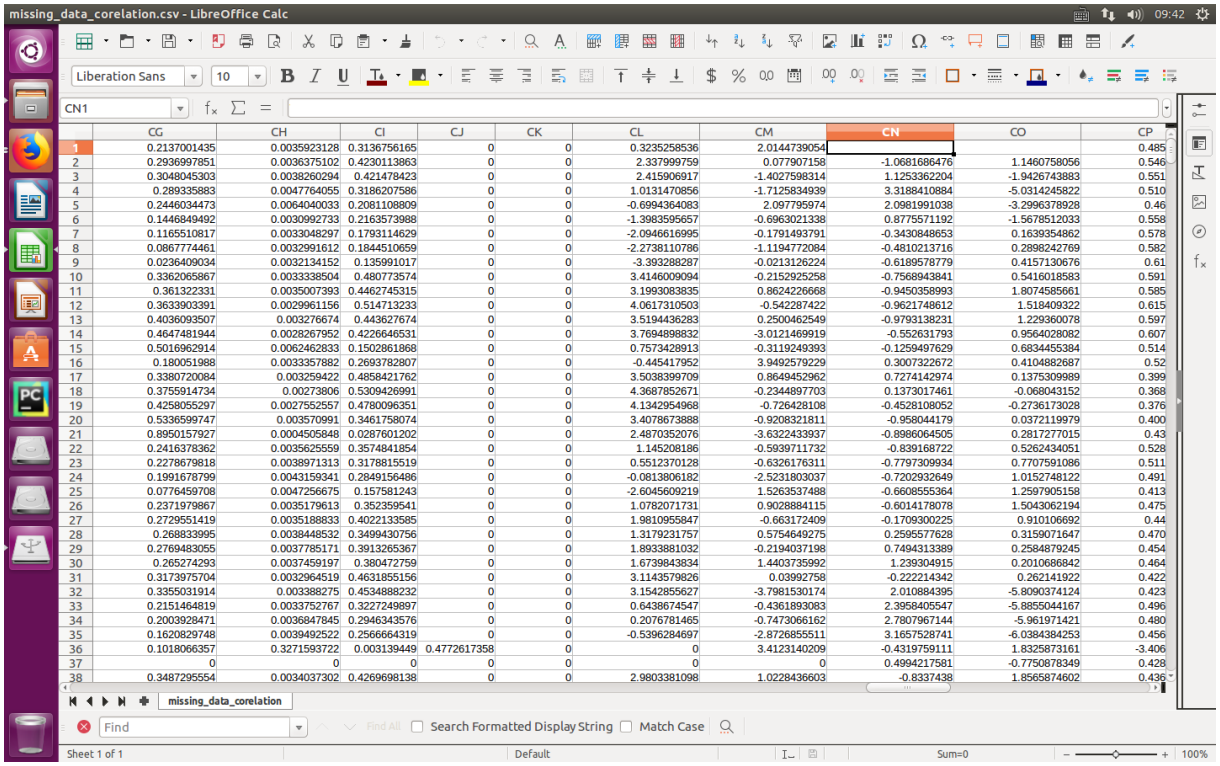
可视化对比填补前后数据,直方图如下所示:



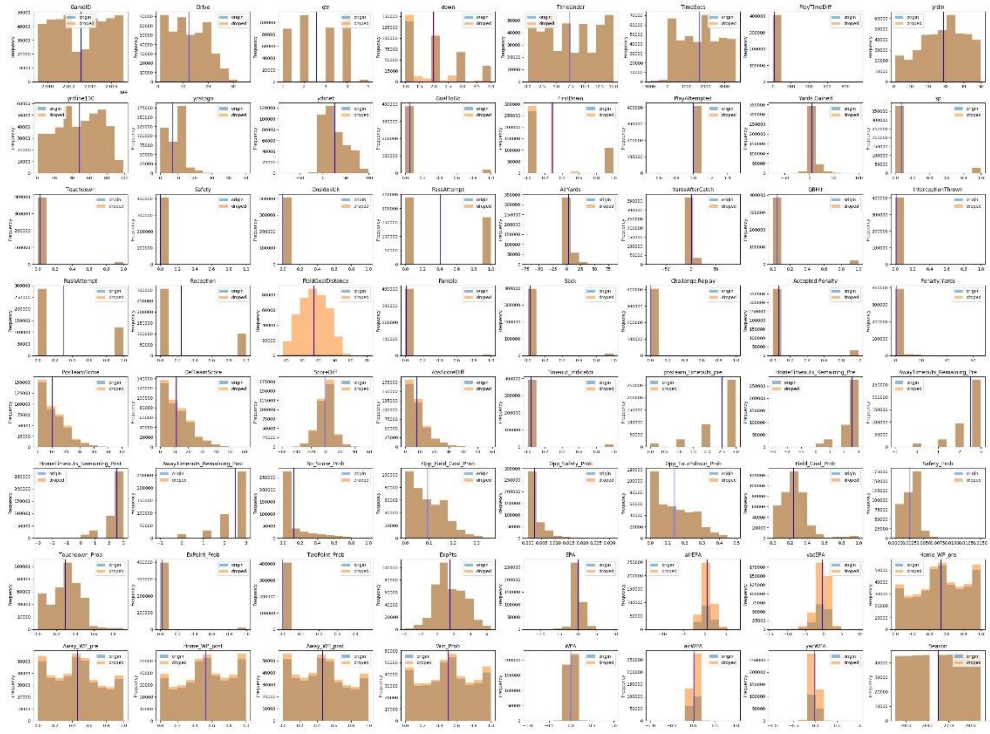
- 通过属性的相关关系来填补缺失值
  填补前,csv 数据文件部分情况如下所示:



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