NFL Play by Play 数据分析与预处理

唐正 2120171060

1 数据可视化和摘要

1.1 数据摘要

1.1.1 标称数据

每个可能取值的频数（由于内容过多，选取四个显示部分）：

标称属性 <RushAttempt> 频数统计

value	count
0	286857
1	120831

标称属性 <RunLocation> 频数统计

value	count
-NaN-	288178
right	43532
left	43157
middle	32821

标称属性 <RunGap> 频数统计

value	count
-NaN-	320260
end	31265
tackle	29089
guard	27074

标称属性 <Receiver> 频数统计

value	count
-NaN-	246127
J.Jones	1725
C.Johnson	1517
L.Fitzgerald	1368
D.Thomas	1293
B.Marshall	1267
A.Brown	1237
S.Smith	1213
A.Green	1023

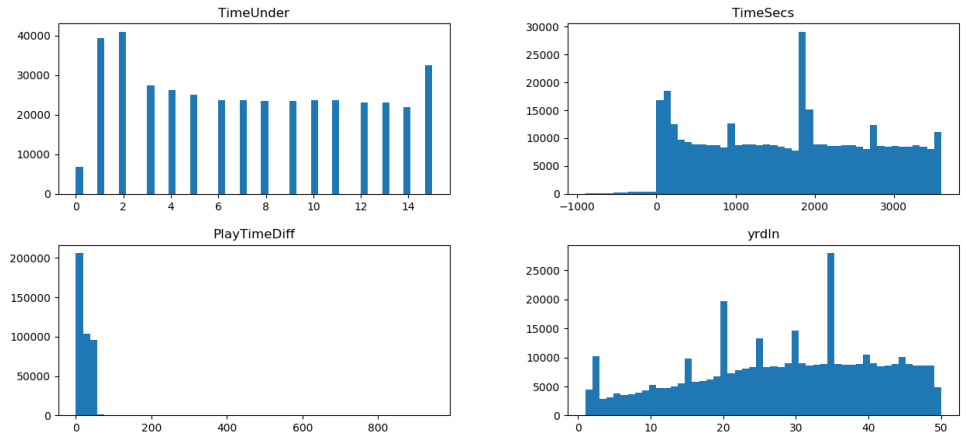
1.1.2 数值属性

数值属性的最大、最小、均值、中位数、四分位数及缺失值的个数：

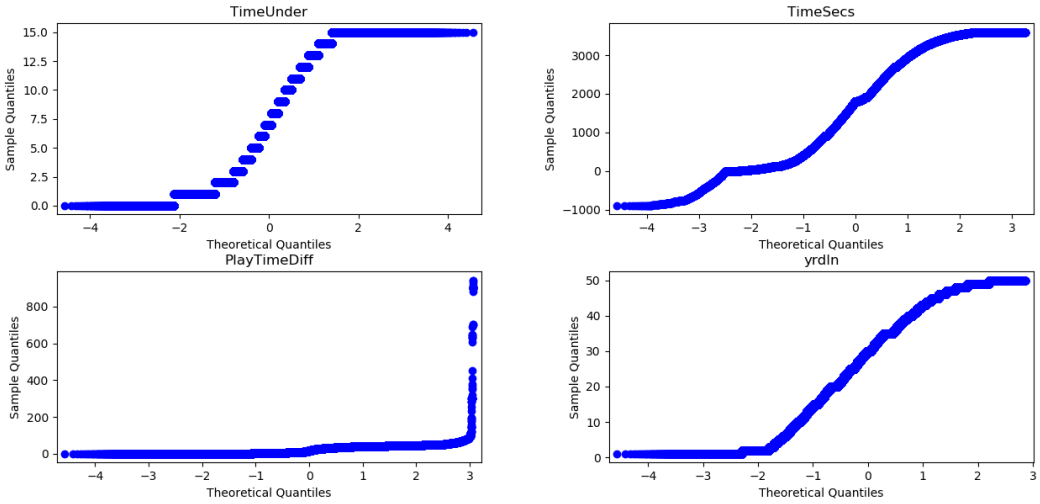
	max	min	mean	50%	25%	75%	NaN
TimeUnder	15.000000	0.000000	7.374200	7.000000	3.000000	11.000000	0
TimeSecs	3600.000000	-900.000000	1695.268944	1800.000000	778.000000	2585.000000	224
PlayTimeDiff	943.000000	0.000000	20.576762	17.000000	5.000000	37.000000	444
yrdln	50.000000	1.000000	28.488327	30.000000	20.000000	39.000000	840
yrdline100	99.000000	1.000000	48.644081	49.000000	30.000000	70.000000	840
ydsnet	99.000000	-87.000000	25.945517	19.000000	5.000000	43.000000	0
Yards.Gained	99.000000	-74.000000	4.994221	1.000000	0.000000	7.000000	0
AirYards	84.000000	-70.000000	3.264006	0.000000	0.000000	4.000000	0
YardsAfterCatch	90.000000	-81.000000	1.252598	0.000000	0.000000	0.000000	0
FieldGoalDistance	71.000000	18.000000	37.465132	38.000000	29.000000	46.000000	398740
Penalty.Yards	66.000000	0.000000	0.613673	0.000000	0.000000	0.000000	0
PosTeamScore	61.000000	0.000000	10.201424	7.000000	2.000000	16.000000	26904
DefTeamScore	61.000000	0.000000	11.414484	10.000000	3.000000	17.000000	26904
ScoreDiff	59.000000	-59.000000	-1.186590	0.000000	-7.000000	4.000000	24988
AbsScoreDiff	59.000000	0.000000	7.783541	7.000000	3.000000	11.000000	26904
posteam_timeouts_pre	3.000000	0.000000	2.521239	3.000000	2.000000	3.000000	0
HomeTimeouts_Remaining_Pre	3.000000	-3.000000	2.540479	3.000000	2.000000	3.000000	0
AwayTimeouts_Remaining_Pre	3.000000	-1.000000	2.517222	3.000000	2.000000	3.000000	0
HomeTimeouts_Remaining_Post	3.000000	-3.000000	2.520118	3.000000	2.000000	3.000000	0
AwayTimeouts_Remaining_Post	3.000000	-1.000000	2.496367	3.000000	2.000000	3.000000	0
No_Score_Prob	1.000000	0.000000	0.127816	0.024771	0.002791	0.172509	176
Opp_Field_Goal_Prob	0.360177	0.000000	0.094614	0.083088	0.034599	0.149943	176
Opp_Safety_Prob	0.031461	0.000000	0.002495	0.000988	0.000104	0.003845	176
Opp_Touchdown_Prob	0.496874	0.000000	0.139973	0.124032	0.039834	0.226408	176
Field_Goal_Prob	0.994605	0.000000	0.243906	0.231311	0.152443	0.326130	176
Safety_Prob	0.015177	0.000000	0.002634	0.002990	0.001883	0.003582	176
Touchdown_Prob	0.912963	0.000000	0.295940	0.313676	0.191206	0.407684	176
ExPoint_Prob	0.993128	0.000000	0.024072	0.000000	0.000000	0.000000	0
TwoPoint_Prob	0.473500	0.000000	0.000703	0.000000	0.000000	0.000000	0
ExpPts	6.500900	-3.836488	1.565415	1.257967	0.323526	2.882048	176
EPA	9.508015	-13.494136	0.019116	0.000000	-0.599034	0.563047	369
airEPA	7.346969	-12.849594	0.524818	0.295977	-0.502895	1.386486	248394
yacEPA	9.559834	-14.000000	-0.386086	0.000000	-0.961115	0.485508	248498
Home_WP_pre	1.000000	0.000000	0.534488	0.531274	0.325123	0.769232	24954
Away_WP_pre	1.000000	0.000000	0.465965	0.469052	0.231411	0.675530	24954
Home_WP_post	1.000000	0.000000	0.534791	0.533609	0.321701	0.772882	26587
Away_WP_post	1.000000	0.000000	0.465613	0.466670	0.227694	0.678833	26587

1.2 数据可视化

1.2.1 直方图

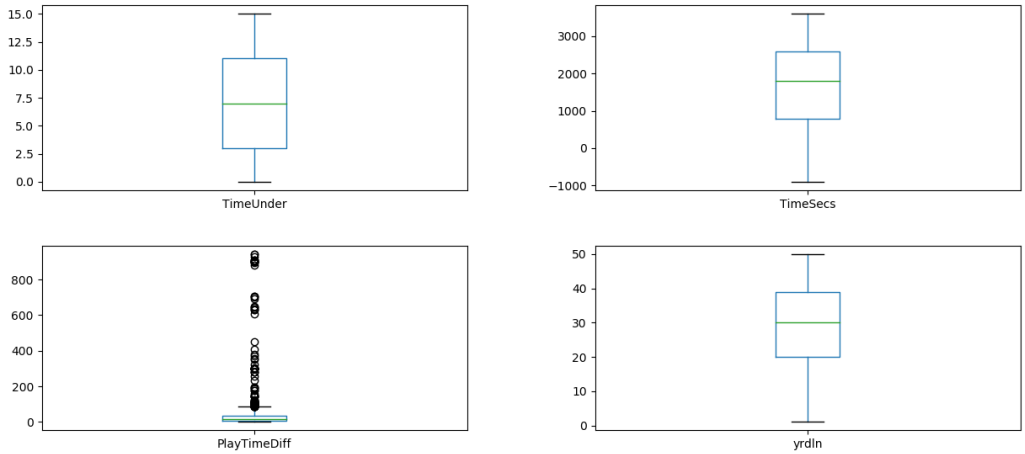


1.2.2 qq 图



根据 qq 图可知图像 1、2 和 4 是近似直线的，其对应属性（TimeUnder、TimeSecs、yrdln）为正态分布态。

1.2.3 盒图



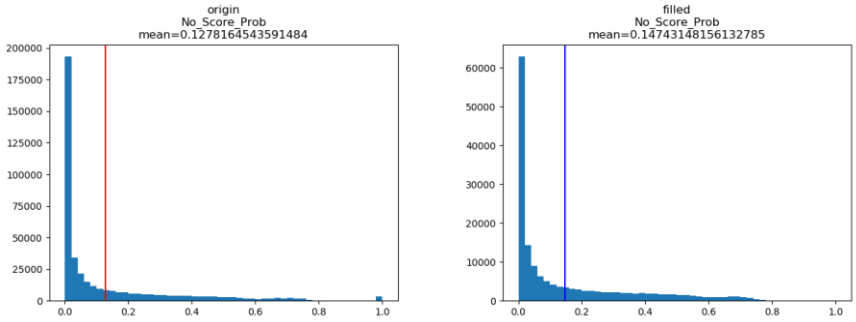
2 数据缺失处理

2.1 将缺失部分剔除

根据分析，可填充的数值属性字段有：No_Score_Prob, Opp_Field_Goal_Prob, Opp_Safety_Prob, Opp_Touchdown_Prob,

Field_Goal_Prob, 'Safety_Prob', Touchdown_Prob, ExpPts, EPA, airEPA, yacEPA, Home_WP_pre, Away_WP_pre, Home_WP_post, Away_WP_post, Win_Prob, WPA, airWPA, yacWPA。对缺失部分进行剔除。

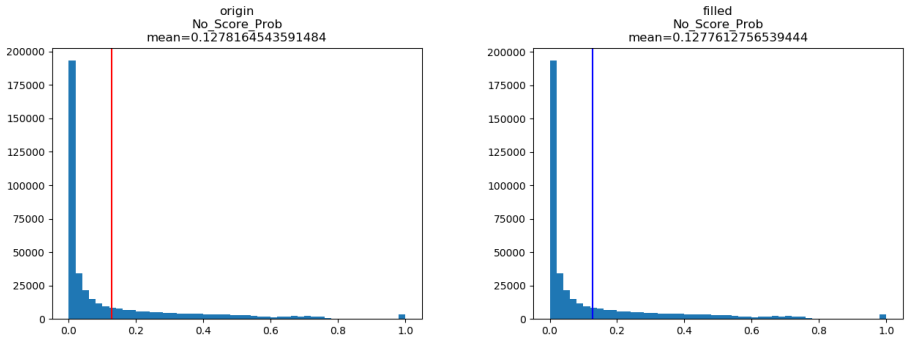
通过直方图比较新旧数据集的数值属性：



在直方图中，左边的红色垂线表示旧数据集的均值，右边的蓝色垂线表示剔除有缺失的数据得到的新数据集的均值。

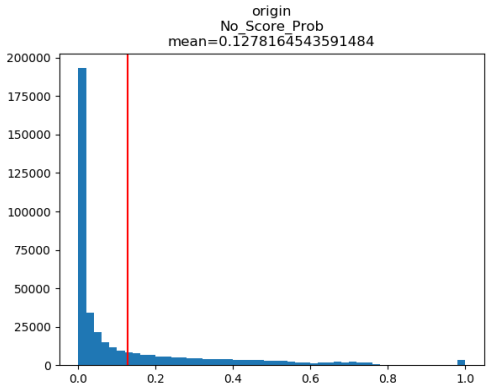
2.2 用最高频率值来填补缺失值

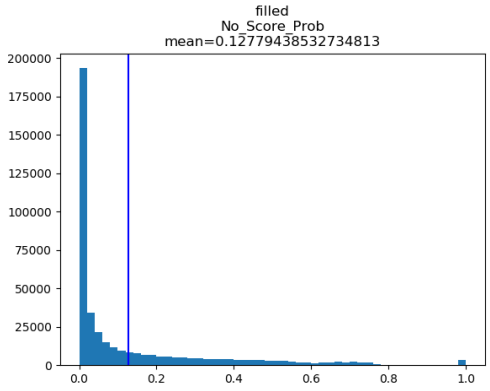
找到每个属性中出现次数最多的值，用这个值填充这个属性中所有的缺失值。在直方图中，左边的红色垂线表示旧数据集的均值，右边的蓝色垂线表示剔除有缺失的数据得到的新数据集的均值。



2.3 通过属性的相关关系来填补缺失值

对于每个数值属性进行插值计算，利用得到的插值填充缺失值。在直方图中，左边的红色垂线表示旧数据集的均值，右边的蓝色垂线表示剔除有缺失的数据得到的新数据集的均值。





2.4 通过数据对象之间的相似性来填补缺失值

无