数据探索性分析与数据预处理
数据集一：NFL Play-by-Play 2009-2017
1. 数据摘要
i. 标称属性
以“FirstDown”属性为例，列举出了所有可能的取值，以及对应的频数：

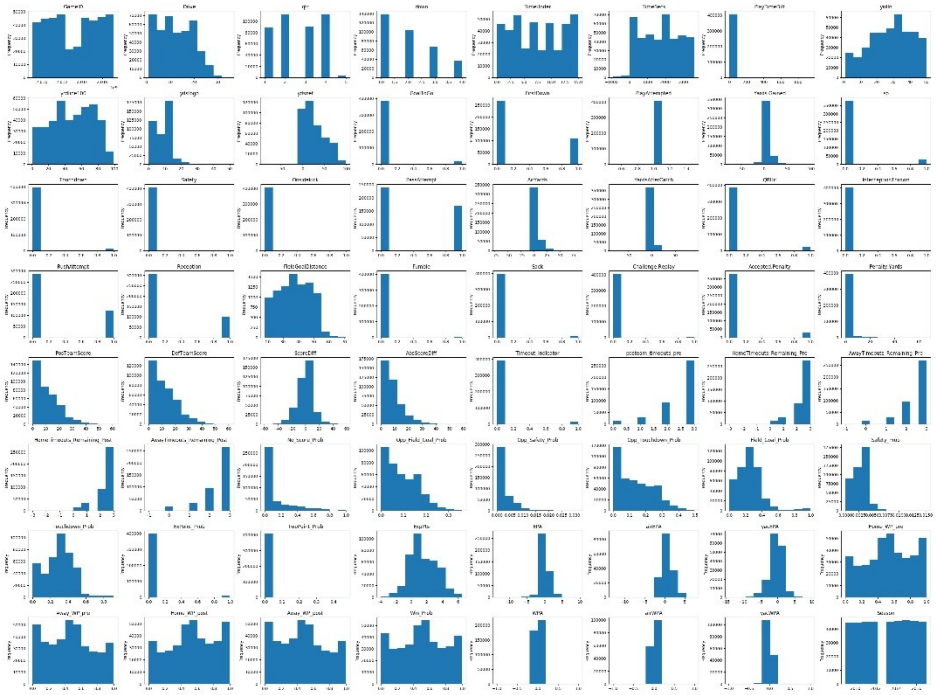
FirstDown	
0:	268810
1:	110067
NA:	28811

由于数据量较大,数据结果保存在/NFL Play by Plays/result_NFL_nominal.txt 中

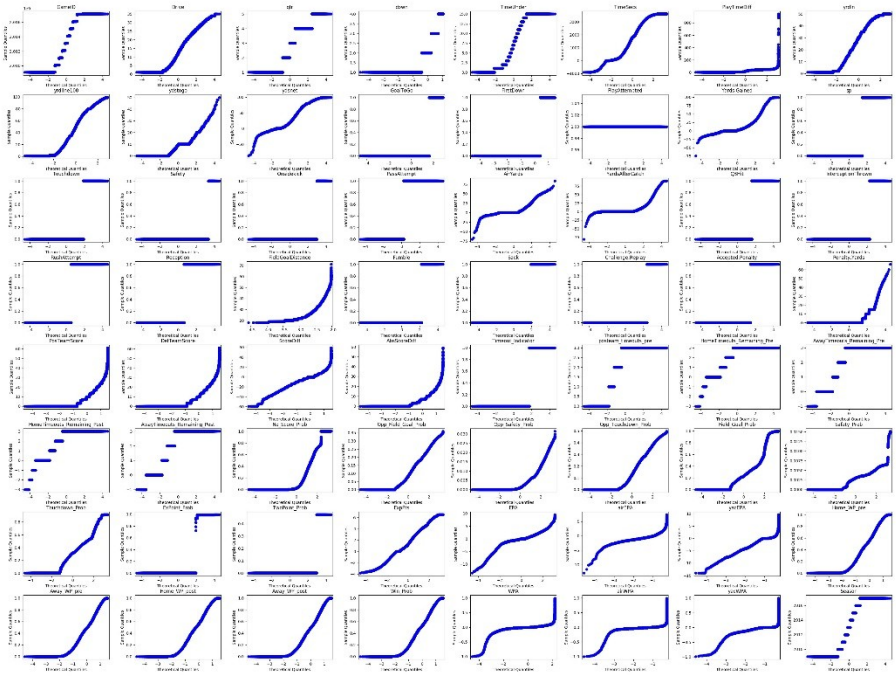
ii. 数值属性
以“Home_WP_pre”属性为例，分别给出了非空值数据的个数（count），平均值（mean），方差（std），最小值（min），四分位数（min，25%，50%，75%，max）以及最大值（max）。

	Home_WP_pre
count	382734.000000
mean	0.534488
std	0.285574
min	0.000000
25%	0.325123
50%	0.531274
75%	0.769232
max	1.000000

数据结果保存在/NFL Play by Plays/result_NFL_numerical.txt 中
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：

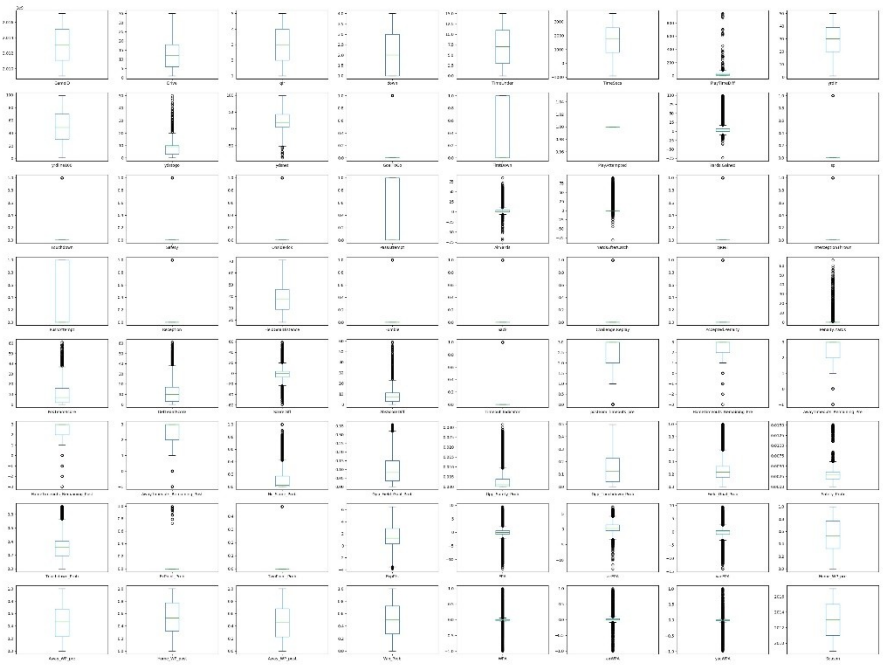


qq 图如下所示：



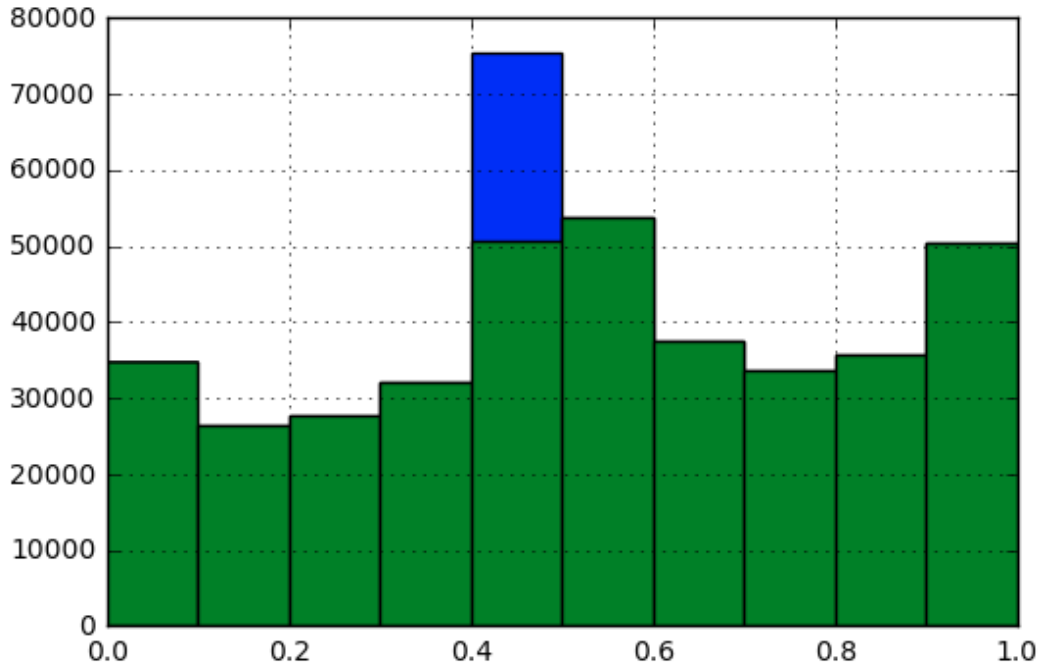
由各个属性的 qq 图可以看出，属性 ExpPts 和 EPA 满足正态分布

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

2. 数据缺失
i. 数据缺失原因
观察数据集中缺失的数据，原因主要是：
ii. 处理缺失数据
剔除缺失部分（绿色） vs 用最高频率值来填补缺失值（蓝色）,下面都以属性“TimeUnder”为例



对于数值属性，可以通过计算协方差矩阵，来判断数据之间的相似度，利用属性的相关关系来填补缺失值。下图截取部分协方差矩阵值，观察可以发现，”Drive“属性和”qtr“属性相关系数为 0.91，二者之间的正相关性很高，因此当其中一个数据缺失时，可以使用另一个数据值进行填充。同理，”TimeSecs“属性和”Drive“、”qtr“属性之间的负相关性很高，它们之间也可以相互填补缺失值。

	GameID	Drive	qtr	down	TimeUnder	TimeSecs
GameID	1.000000	-0.016707	0.000594	-0.003281	-0.007028	-0.002367
Drive	-0.016707	1.000000	0.917050	-0.006638	-0.249329	-0.942744
qtr	0.000594	0.917050	1.000000	0.009883	-0.032128	-0.964949
down	-0.003281	-0.006638	0.009883	1.000000	-0.021469	-0.015410
TimeUnder	-0.007028	-0.249329	-0.032128	-0.021469	1.000000	0.292694
TimeSecs	-0.002367	-0.942744	-0.964949	-0.015410	0.292694	1.000000

数据集二：San Francisco Building Permits

1. 数据摘要

i. 标称属性

以“Permit Type Definition”属性为例，列举出了所有可能的取值，以及对应的频数：

```
{'otc alterations permit': 178844,
 'new construction wood frame': 950,
 'sign - erect': 2892,
 'additions alterations or repairs': 14663,
 'grade or quarry or fill or excavate': 91,
 'demolitions': 600,
 'new construction': 349,
 'wall or painted sign': 511}
```

ii. 数值属性

以“Existing Construction Type”属性为例，分别给出了非空值数据的个数（count），平均值（mean），方差（std），最小值（min），四分位数（min，25%，50%，75%，max）以及最大值（max）。

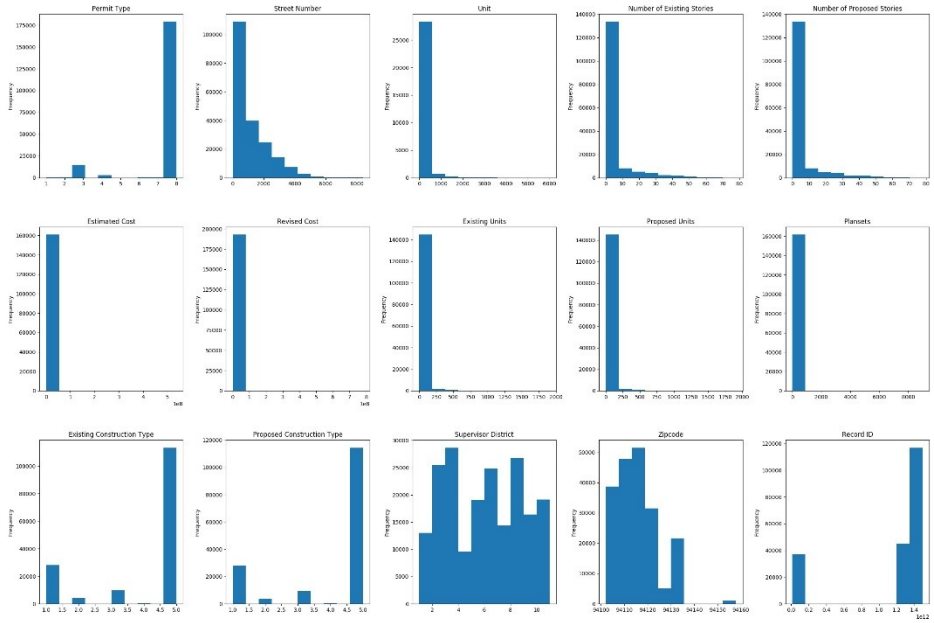
	Existing Construction Type
count	155534.000000
mean	4.072878
std	1.585756
min	1.000000
25%	3.000000
50%	5.000000
75%	5.000000
max	5.000000

2 数据可视化

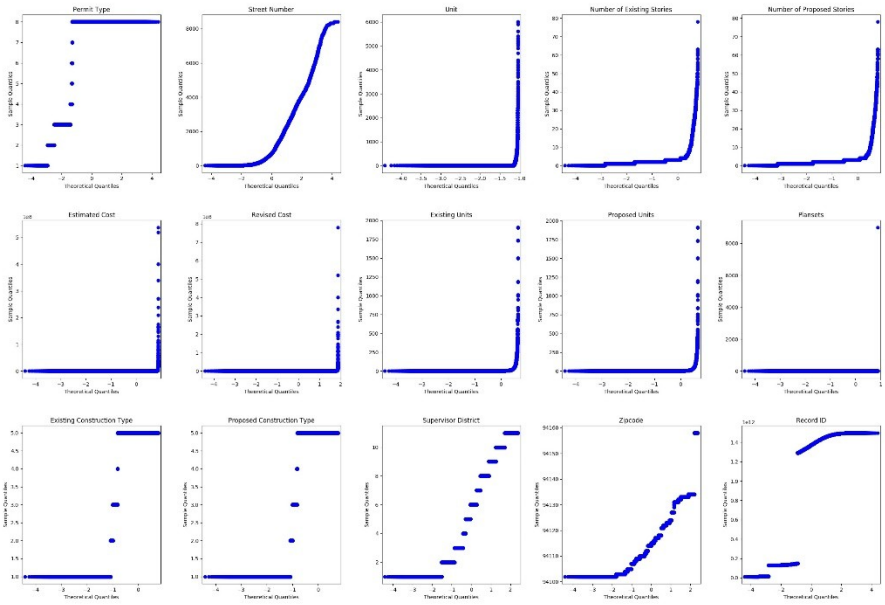
针对数值属性，

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：

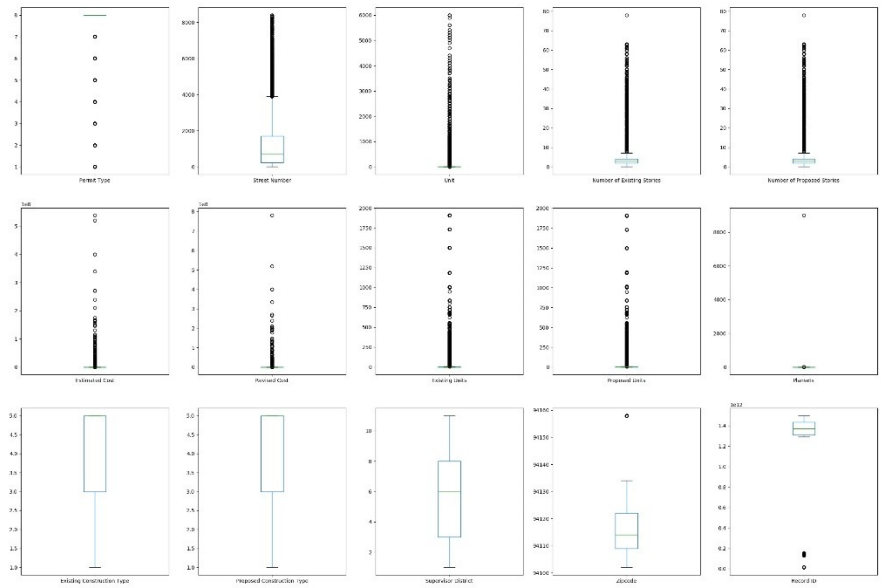


qq 图如下所示：



由各个属性的 qq 图可以看出,无属性满足正态分布

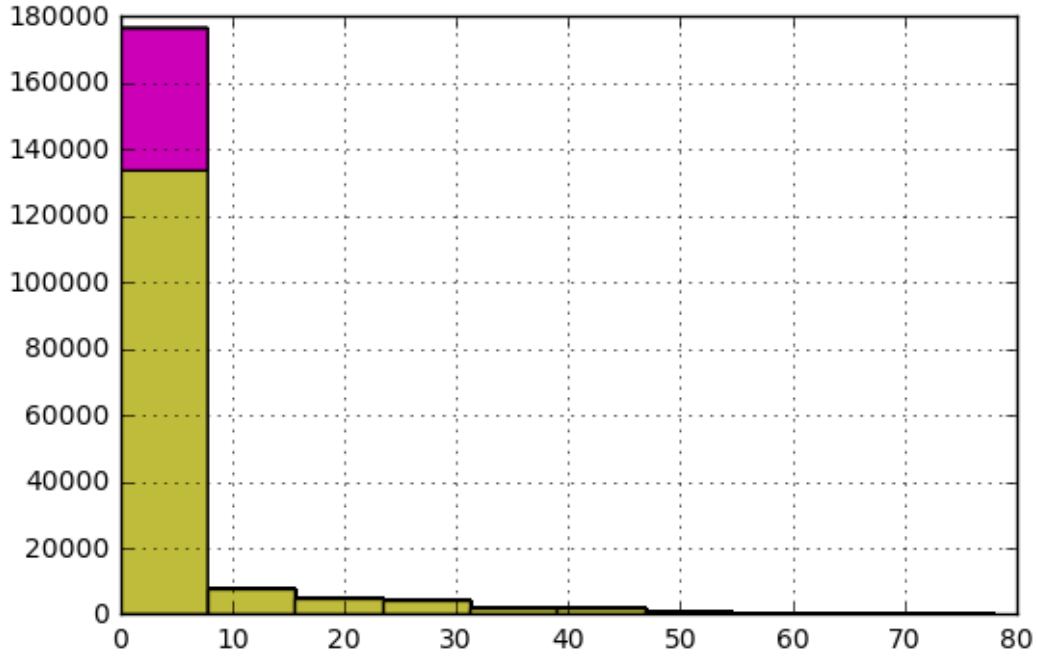
绘制盒图，对离群值进行识别
盒图如下所示：



从各个属性的盒图观察可得,属性 Permit Type、Street Number、Unit、Number of Existing Stories、Number of Proposed Stories、Estimated Cost、Revised Cost、Existing Units、Proposed Units、Plansets、Zipcode、Record ID 存在离群值

3. 数据缺失
i. 数据缺失原因

观察数据集中缺失的数据，原因主要是：
ii. 处理缺失数据
剔除缺失部分（黄色） vs 用最高频率值来填补缺失值（粉色），Number of Existing Stories 为例



对于数值属性，可以通过计算协方差矩阵，来判断数据之间的相似度，利用属性的相关关系来填补缺失值。下图截取部分协方差矩阵值，观察可以发现，”Number of Existing Stories“属性和”Number of Proposed Stories“属性相关系数为 0.99，二者之间的相关性很高，因此当其中一个数据缺失时，可以使用另一个数据值进行填充。同理，”Estimated Cost“属性和”Revised Cost“属性之间的相关系数为 0.97，也可以相互填补缺失值。

	Permit Type	Street Number	Unit	Number of Existing Stories	Number of Proposed Stories	Estimated Cost	Revised Cost
Permit Type	1.000000	-0.002281	0.031978	0.057106	0.055431	-0.120878	-0.120083
Street Number	-0.002281	1.000000	-0.040662	-0.218557	-0.215047	-0.011152	-0.010828
Unit	0.031978	-0.040662	1.000000	0.167038	0.168811	-0.009094	-0.007559
Number of Existing Stories	0.057106	-0.218557	0.167038	1.000000	0.997356	0.030248	0.039181
Number of Proposed Stories	0.055431	-0.215047	0.168811	0.997356	1.000000	0.050336	0.049165
Estimated Cost	-0.120878	-0.011152	-0.009094	0.030248	0.050336	1.000000	0.978798
Revised Cost	-0.120083	-0.010828	-0.007559	0.039181	0.049165	0.978798	1.000000