所有表格都在 描述统计.xlsx 这一文件中

数据在modified.csv中

直方图、散点图、线箱图在各自文件夹中

1. 用describe简单查看数据情况，由于变量名中部分含有特殊符号，被自动替换，但Variable lable中有标注，后边描述统计中已换回来（即与原变量名保持一致）。共有143个变量，2921条数据。（详见Describe表格）

2. 查看重复值：无

3. 查看缺失值：发现Nation和Age缺失，查看存在缺失的数据所在行，刚好在同一行（1532 Javier Llabrés），自行搜索了该球员信息并修改填充。在填充过程中发现该球员出生年份为“0”，直接按照查到的信息修改为“2002”。

4. 查看字符型变量

（1）Player（球员）：检测到有重复值，涉及到部分球员转会，但这些人在不同的队伍中位置和赛季均不相同，我暂时未对他们做处理，如需要处理可再戳我。

Duplicates in terms of Player

--------------------------------------

Copies | Observations Surplus

---------+----------------------------

1 | 2652 0

2 | 266 133

3 | 3 2

--------------------------------------

这些球员的Rk和名字已存入Transferred player表格中。

（2）Nation（球员国籍）：可以看到我们的数据集中有来自102个国家的球员，其中最多的是来自西班牙、法国、意大利、德国和英格兰的球员。（详见Nation表格）



（3）Pos（球员位置）：发现有一个特殊值“GKMF”，怀疑是脏数据，使用维基百科搜到的内容，将该球员的位置改为“GK”，得到共计10个不同的值。

【DF: Defender（后卫）

MF: Midfielder（中场）

FW: Forward（前锋）

MFFW: Midfielder/Forward（中场/前锋）

FWMF: Forward/Midfielder（前锋/中场）

GK: Goalkeeper（守门员）

DFMF: Defender/Midfielder（后卫/中场）

FWDF: Forward/Defender（前锋/后卫）

MFDF: Midfielder/Defender（中场/后卫）

DFFW: Defender/Forward（后卫/前锋）】

（4）Squad（球队）：我们的数据来自98支不同的球队，平均每支球队为我们提供了约30条数据。其中最多的是Salernitana，提供了42条数据。

（5）Comp（联赛）：我们的数据来自Bundesliga、La Liga、Ligue 1、Premier League、Serie A这五类联赛。其中La Liga联赛生成的数据量最大，有617条。

5. 接下来查看数值型变量中是否存在明显的异常值（已略过），但绘制的直方图和箱线图留下了。

6. 群里举例的三个变量类型不属于一套分类体系，所以我猜所谓变量类型其实是想要Measure（度量标准）：1个Ordinal（定序变量），1个Interval（定距变量），5个Norninal（定类变量），136个Scale/Ratio（定比变量）。

7. 直方图与分布情况

对136个变量进行分布情况观察。由于没有用核密度估计/统计检验/Q-Q图/参数拟合等方法，肉眼直接看直方图且有一些分布图象上可能很接近，我将大致使用以下几个简单类别。（该结果存在个人主观因素导致的误差）。

正态分布（Normal Distribution）：1

偏态分布（Skewed Distribution）：67

双峰分布（Bimodal Distribution）：1

均匀分布（Uniform Distribution）：1

指数分布（Exponential Distribution）：37

幂律分布（Powerlaw Distribution）：6

混合分布（Mixture Distribution）：23