### 1. 问题描述

本次作业中，将对2个数据集进行探索性分析与预处理。

### 2. 数据说明

* 数据集1: NFL Play-by-Play 2009-2017
* 数据集2: San Francisco Building Permits

下载数据: [地址](https://pan.baidu.com/s/1rayD18hDMD8elzcE9GYs1g)

数据集中属性解释：

* 数据集1: [参考](https://github.com/ryurko/nflscrapR-data)
* 数据集2: 见下载地址中DataDictionaryBuildingPermit.xlsx

### 3. 数据分析要求

#### 3.1 数据可视化和摘要

##### 数据摘要

* 对标称属性，给出每个可能取值的频数，
* 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。

##### 数据的可视化

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。
* 绘制盒图，对离群值进行识别

#### 3.2 数据缺失的处理

观察数据集中缺失数据，分析其缺失的原因。

分别使用下列四种策略对缺失值进行处理:

* 将缺失部分剔除
* 用最高频率值来填补缺失值（标称数据）
* 通过属性的相关关系来填补缺失值（C4.5）
* 通过数据对象之间的相似性来填补缺失值（K紧邻）

处理后，可视化地对比新旧数据集。

# 数据挖掘作业一