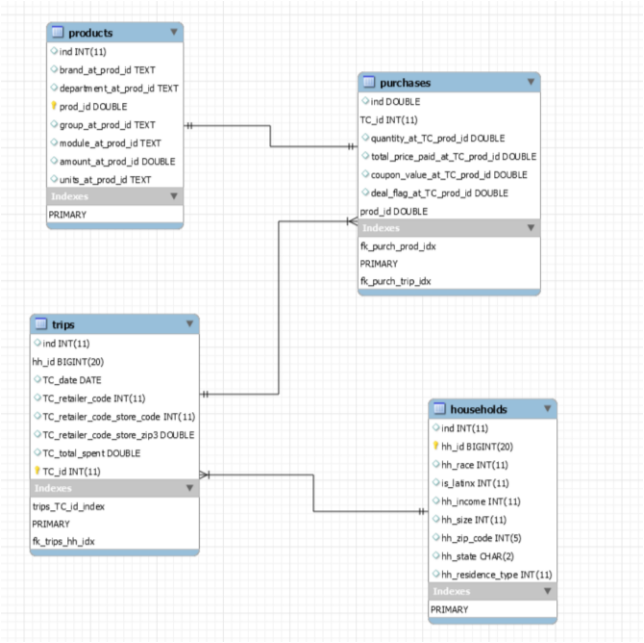
超市购物行为分析项目

此项目为大数据 I（研究生阶段课程）的期末项目。
超市购物行为分析基于现实生活中的一个零售数据集。该数据集包含一年中的 37916 个不同家庭的 7395060 次购物行程中的 39296405 条购物记录。

我们为四个独立的表设计了一个 MySQL 数据库的框架：
1.**产品**（产品的详细信息表，例如品牌、数量、单位等）(3844900 行）
2.**购买**（每条购买记录的详细信息表，例如产品 ID、每件商品的金额、优惠券价值等）（39296405 行）
3.**行程**（购物行程的详细信息表，例如哪个家庭、何时、去哪家零售商等）（7395060 行）
4.**家庭**（家庭的信息表，例如种族、收入水平、住所等）（37916 行）

然后我们将数据集导入上述设计好的数据库中。下面展示的是我们获得的 ER 图：



基于此数据库，我们通过 sql 查询操作以提取这些约 38,000 户家庭显示的购物模式。这些汇总结果为我们提供了有关购买行为的有意义的商业见解。我们了解到，客户的购物习惯或行为背后总有一个故事或特殊的原因。
有趣的问题主要包括（我们针对这些特定的问题写 sql code）：
1. 在家庭月度水平上，有多少个家庭在三个月内至少没有购物一次？为什么？那合理吗？
2. 有多少家庭将大部分商店（支出超过 80%）集中在某个零售商上；在两个特定的零售商之间？
3. 哪个忠诚度最高的零售商？（百分比和总数）
4. 这些忠实的客户住在哪里？（每个州分别有多少个）
5. 两次连续购物旅行之间的平均天数。
6. 每月购物旅行的次数是否与购买的平均物品数量相关？
7. 每件商品的平均支付价格是否与购买的商品数量相关？
8. 哪些产品类别包含更多超市自有品牌的产品？
9. 超市自有品牌的产品的支出份额在几个月内是否保持不变？
10. 比较高收入，中收入和低收入家庭组在每月支出中超市自有品牌的产品消费份额的占比。

示例 sql 查询代码如下：
##购物旅行次数和每月平均购买的物品数量

```
DROP TABLE IF EXISTS tb1;
CREATE TABLE tb1
   SELECT *, SUBSTR(T1.TC_date,3,5) as month
   FROM (SELECT hh_id,
                              TC_id as TC_id_1,
                              TC_date
               FROM trips) AS T1
   LEFT JOIN (SELECT TC_id as TC_id_2,
                                      SUM(quantity_at_1
                        FROM purchases
       GROUP BY TC_id) AS T2
   ON T1.TC_id_1 = T2.TC_id_2;

DROP TABLE IF EXISTS tb2;
CREATE TEMPORARY TABLE tb2
   SELECT hh_id,
                month,
                COUNT(TC_id_1) as num_trips
   FROM tb1
   GROUP BY hh_id, month;

DROP TABLE IF EXISTS tb3;
CREATE TEMPORARY TABLE tb3
   SELECT month,
                round(AVG(num_trips),2) as avg_trips
   FROM tb2
   GROUP BY month
   ORDER BY month;

DROP TABLE IF EXISTS tb4;
CREATE TEMPORARY TABLE tb4
   SELECT month,
          round(AVG(num_item),2) as avg_items
   FROM tb1
   GROUP BY month
   ORDER BY month;

DROP TABLE IF EXISTS tb5;
CREATE TEMPORARY TABLE tb5
   SELECT T1.month, avg_trips, avg_items
   FROM tb3 as T1
      INNER JOIN tb4 as T2
      ON T1.month = T2.month;

SELECT * FROM tb5;
```

除此之外，我们还使用 Python 对一些有趣的结论进行可视化的展示。我们使用的软件包包括：matplotlib，seaborn。

您可以点击在随附的 PDF，查看上述问题的答案以及其他的有趣发现。

由于该项目的数据集是保密的（属于布兰迪斯大学），因此数据集不能共享。