**РЕЗУЛЬТАТЫ ПРЕДВАРИТЕЛЬНОГО АНАЛИЗА ДАННЫХ**

Данные представляют собой таблицу из 69 объектов, которые описываются 99 переменными.

**Структура переменных**

1. Информационные (адрес, долгота…)
2. Внутренние условия функционирования ТТ (площадь, число касс…)
3. Внешние условия функционирования (человеко-поток, средний доход…)
4. Результаты функционирования (прибыль, средний чек, рейтинг гугл…)
5. Гео-характеристики локации (Парковка до 50м…)

**Типы переменных**

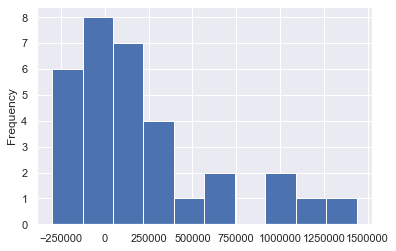
1. Количественные (требуют заполнения пропусков, проверки распределений)
2. Качественные (требуют кодирования)
3. Булевы (возможно, нужно строить искусственные переменные)
4. Даты (требуют преобразования, возможно, нужно строить искусственные переменные)

**Эндогенные переменные**

В этом качестве могут выступать переменные «Результаты функционирования» (Прибуток підрозділу, EBITDA, Успех…)

Проблема 1 – Прибуток підрозділу, EBITDA заданы только для 32 из 69 объектов

Проблема 2 – Диапазон изменения «Прибуток підрозділу» очень широк, распределение негладкое, ассиметричное

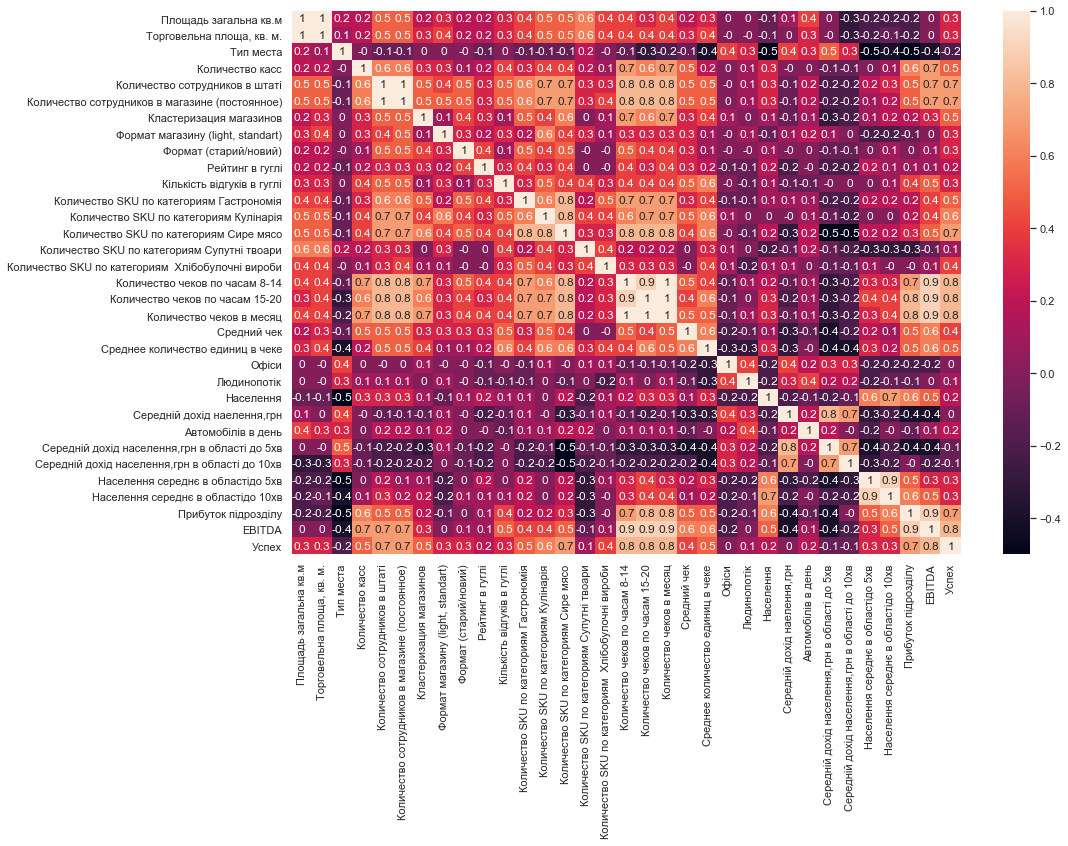


**Экзогенные переменные**

В этом качестве могут выступать переменные «Внутренние условия функционирования ТТ», «Внешние условия функционирования», «Гео-характеристики локации»

**Задачи:**

1. Есть слабо-вариативные переменные, нужно чистить / создавать искусственные
2. Есть пропущенные значения, нужно чистить / заполнять
3. Нужно тестировать данные на наличие ошибок, выбросов
4. Есть коррелированные переменные, нужно чистить



**Моделирование**

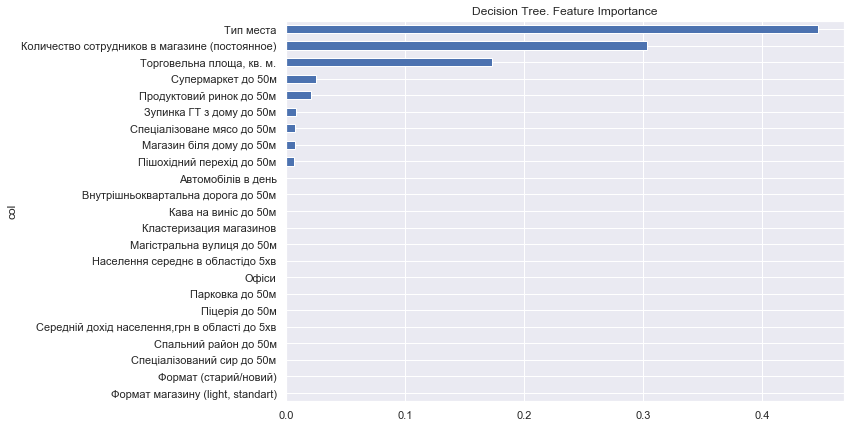
**Регрессия**

Проблема 1 – число переменных существенно превышает число наблюдений, нужно чистить список, формировать искусственные. В любом случае, стоить сложные нелинейные модели нецелесообразно

Проблема 2 – Количество некоррелированных переменных, описывающих внутренние и внешние условия функционирования – порядка 15, число вариативных гео-характеристик локаций – порядка 25 \* 2. Однако гео-характеристики локаций – булевы переменные, они не способны существенно повысить качество описания объясняемой переменной.

Тест 1 – При построении базовой модели множественной регрессии значимых переменных не выявлено, оценка качества на тестовой выборке – низкое.

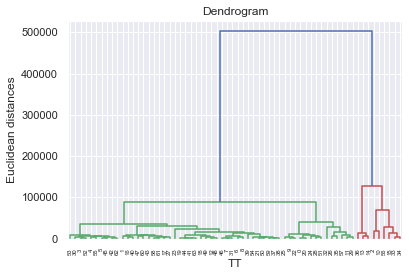
Тест 2 – При построении дерева решений, оценка качества на тестовой выборке – низкое, наиболее значимые переменные:



**ВЫВОД 1** – учитывая соотношения числа объектов и их характеристик, количественных и булевых переменных, диапазон и характер распределения объясняемой переменной, построение качественных моделей прогнозирования представляется маловероятным

**Кластеризация**

Тест 3 – можно выделить как минимум 2 достаточно однородные группы объектов



В данных можно выделить более однородные структуры, но они представлены слишком маленьким числом объектов, чтоб эти результаты можно было использовать для моделирования внутри кластеров

**Классификация**

Тест 4 – Базовый классификатор демонстрирует достаточно высокое качество как на обучающей, так и на тестовой выборке

0.9607843137254902, 0.9230769230769231

**ВЫВОД 2** – предлагаем строить классификатор – успех/провал ТТ