**Introduction**

Practice problems or data science projects are one of the best ways to learn data science. You don’t learn data science until you start working on problems yourself.

### The Problem Statement

Understanding the problem statement is the first and foremost step. You can view this in the [competition page](http://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii) but I’ll iterate the same here:

*The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.*

*Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.*

### The Hypotheses

I came up with the following hypothesis while thinking about the problem. These are just my thoughts and you can come-up with many more of these. Since we’re talking about stores and products, lets make different sets for each.

**Store Level Hypotheses:**

1. **City type:** Stores located in urban or Tier 1 cities should have higher sales because of the higher income levels of people there.
2. **Population Density:** Stores located in densely populated areas should have higher sales because of more demand.
3. **Store Capacity:** Stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place
4. **Competitors:** Stores having similar establishments nearby should have less sales because of more competition.
5. **Marketing:** Stores which have a good marketing division should have higher sales as it will be able to attract customers through the right offers and advertising.
6. **Location:** Stores located within popular marketplaces should have higher sales because of better access to customers.
7. **Customer Behavior:** Stores keeping the right set of products to meet the local needs of customers will have higher sales.
8. **Ambiance:** Stores which are well-maintained and managed by polite and humble people are expected to have higher footfall and thus higher sales.

**Product Level Hypotheses:**

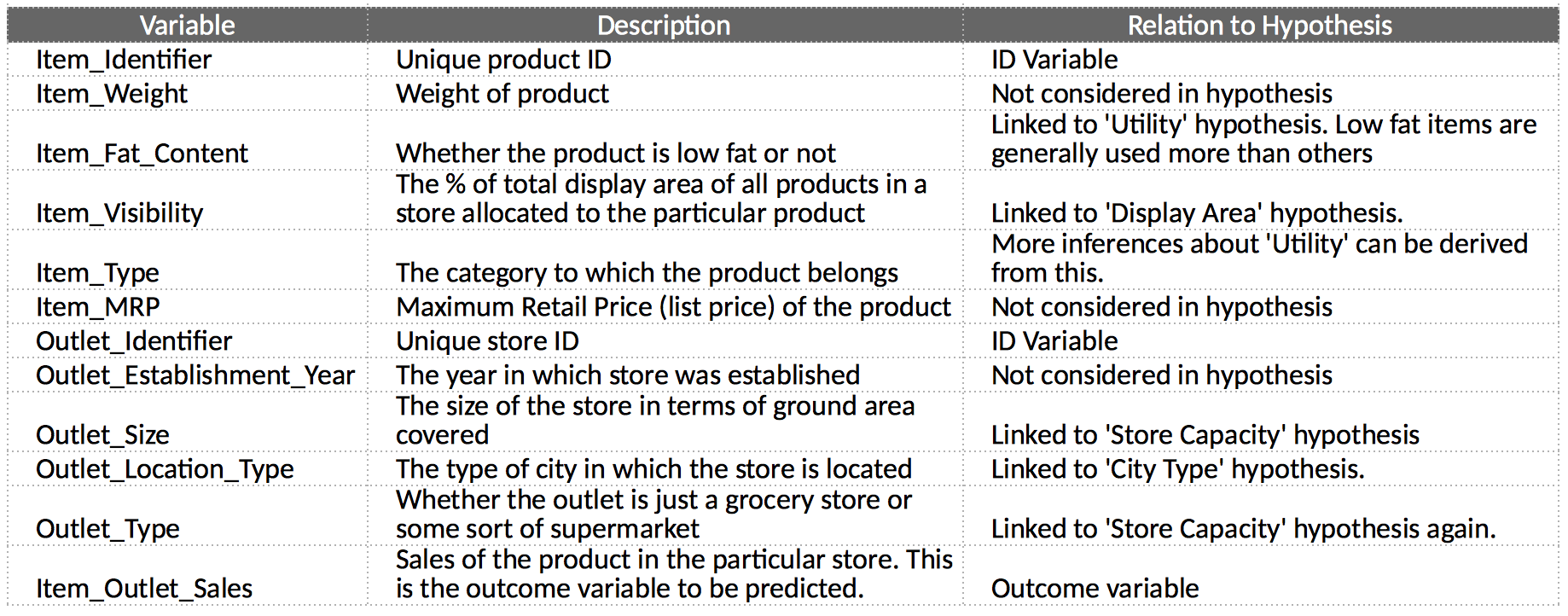
1. **Brand:** Branded products should have higher sales because of higher trust in the customer.
2. **Packaging:** Products with good packaging can attract customers and sell more.
3. **Utility:** Daily use products should have a higher tendency to sell as compared to the specific use products.
4. **Display Area:** Products which are given bigger shelves in the store are likely to catch attention first and sell more.
5. **Visibility in Store:** The location of product in a store will impact sales. Ones which are right at entrance will catch the eye of customer first rather than the ones in back.
6. **Advertising:** Better advertising of products in the store will should higher sales in most cases.
7. **Promotional Offers:** Products accompanied with attractive offers and discounts will sell more.

These are just some basic 15 hypothesis I have made, but you can think further and create some of your own. Remember that the data might not be sufficient to test all of these, but forming these gives us a better understanding of the problem and we can even look for open source information if available.

## 2. Data Exploration

We’ll be performing some basic data exploration here and come up with some inferences about the data. We’ll try to figure out some irregularities and address them in the next section. If you are new to this domain, please refer our [Data Exploration Guide](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/).

The first step is to look at the data and try to identify the information which we hypothesized vs the available data. A comparison between the data dictionary on the competition page and out hypotheses is shown below:

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/02/0.-data-dictionary-1.png)