**Team 1: Adil A Kumar, Jake Arendsen, Ivy Zhou, Zihan Wang**

1. **Business problem:**

Rossmann is a German drug store chain that currently operates over 3,790 stores across Europe. As of this Kaggle competition, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied. Therefore, Rossmann wants a centralized model that can be used to project sales for all their stores.

1. **Business Value:**

Obviously, there’s great value to Rossmann in being able to project sales in advance. It can help shape corporate strategy in terms of promotions, marketing, and expansion/closing decisions. There’s truly numerous things Rossmann could do with the information to help drive sales even further.

1. **Use case:**

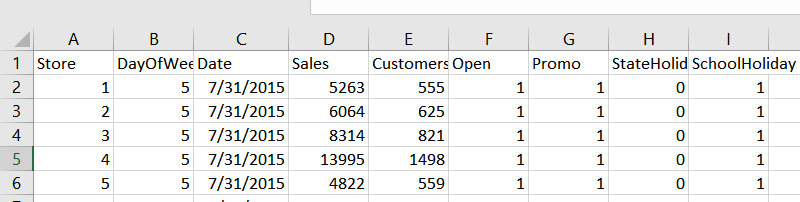
Again, there’s any number of things Rossmann could want to do with the data. Most likely, Rossmann would want to know this in advance to help project cash flows and know which stores might need support and which should achieve just fine without intervention.

1. **Data Science problem:** Predict the daily sales for different Rossman stores given the dependent variables.

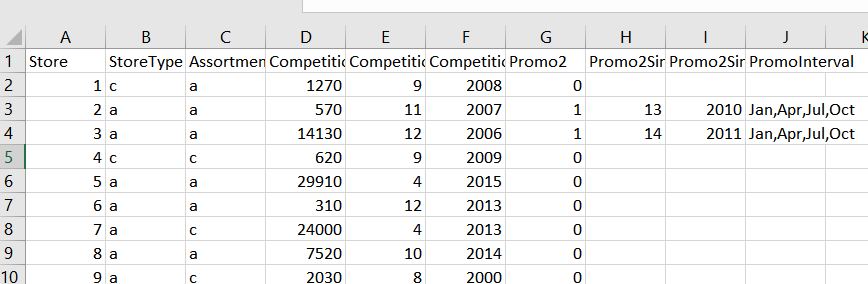
The problem is a regression problem. Thus, it is a supervised learning problem.

1. **Data instance:** Each row of data indicates daily sales by store with no of customers visited. We also have store, promotion, assortment, competition, and promotion attributes by store. Below are snapshots

Training data:



Store attributes:



Data Description:

* Id - an Id that represents a (Store, Date) duple within the test set
* Store - a unique Id for each store
* Sales - the turnover for any given day (this is what you are predicting)
* Customers - the number of customers on a given day
* Open - an indicator for whether the store was open: 0 = closed, 1 = open
* StateHoliday - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
* SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools
* StoreType - differentiates between 4 different store models: a, b, c, d
* Assortment - describes an assortment level: a = basic, b = extra, c = extended
* CompetitionDistance - distance in meters to the nearest competitor store
* CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
* Promo - indicates whether a store is running a promo on that day
* Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
* Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2
* PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

This data description is taking straight from Kaggle. The remaining columns (Day of week, date, customers) are very intuitive, though we will likely need to confirm that day of week is done using the European system (1 for Monday rather than Sunday)

1. **Useful features:** All columns in data look like useful features for modeling. However, we would need to explore the data in further detail and drop features based on importance. We may have to also create some new variables. Additionally, we’ll need to find a way to incorporate store data into the training and test data as we believe some of those variables will be key to helping predict sales. Finally, customers are not included in the training set (as it’s obviously not known). We highly suspect that the number of customers visiting the store I might be one of the most important variables in predicting the sales for each store. We’ll have to either drop it from modeling or try to forecast customers to help forecast sales.
2. **Data Source** - <https://www.kaggle.com/c/rossmann-store-sales/data>

The data is taken from the above Kaggle competition.

1. **Data instance count:** train data has 1017210 rows of daily sales data and store attribute data has 1115 rows
2. **Data Structure** – 1 ID column (in test data only), 17 Independent variables, 1 Target Variable - Sales