# **Title: Rossman Store Sales Data**

|  |  |  |
| --- | --- | --- |
| First name | Last Name | IIT Email |
| Vishnu | Pajjuri | [vpajjuri1@hawk.iit.edu](mailto:vpajjuri1@hawk.iit.edu) |
| Gauri | Khatri | [gkhatri@hawk.iit.edu](mailto:gkhatri@hawk.iit.edu) |
| Aditya | Dubey | [adubey6@hawk.iit.edu](mailto:adubey6@hawk.iit.edu) |
| Amrutha | Gowda | alakshmanegowda@hawk.iit.edu |

**Type of Your Projects (select one or more):** Multiple linear regression, Hypothesis and ANOVA

1. **Introduction**

Our Project is based on Rossmann Store Sales data which describes the various features related to Store, Sales, Customers, StoreType, Open, StateHoliday, SchoolHoliday, Assortment, CompetitionDistance, Promo.Store Sales are influenced by many factors. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity.

1. **Data Sets**

The dataset hosted by Rossmann company.

This data has been collected from kaggle.com(<https://www.kaggle.com/c/rossmann-store-sales/data>) and is available in csv format (7 MB). There are 10,17,209 instances(records) and 15 attributes(columns) in the dataset.

Attributes in this dataset are as follows:

* 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

1. **Research Problems**

Based on the dataset, the interested research problems are:

1. How does sales are affected based on -School Holiday, Promo, Promo Interval, state holidays.

2. Analyse sales in each store based on product sold.

3. Analyse affect of promotion on sales.

1. **Potential Solutions**

1. We will use Multiple linear regression to predict the sales on given day based on the above mentioned factors.

2. ANOVA is performed to compare average sales of different store types .

1. We are performing hypothesis testing to decide if stores participating in promotion has high number of sales or not?
2. **Evaluations**

We will take following steps in evaluating our model:

* Since our database contains 1.0 million records, we will use Hold-out evaluation for our model. We will split the data into 75-25 ratio i.e. training set (75%) and test set (25%) and will start building the model using training data.
* We will build different models and evaluate them using coefficient of determination on training dataset to check the accuracy of model. We will then apply the prediction methods on test dataset and evaluate the results by comparing all the models and choose the best out of them.

1. **Expected Outcomes**

* Predicting number of sales in advance on given day can provide enough time to store managers to focus on customers and their teams.
* Store participating in promo has high number of sales or not? We will get this thing to be predicted by hypothesis testing.
* Predicting number of sales in advance on given day can help in better management of staff schedules.
* Predicting number of sales for given day can help in increasing efficiency of employees.