# Data Preprocessing (UDatE)

## Assessed Report Task Description

Eghbal Rahimikia
Alliance Manchester Business School
The University of Manchester
Email: eghbal.rahimikia@manchester.ac.uk

# Introduction

## Data pre-processing for a sales forecasting problem

Real-world business problem of forecasting sales is one of the most difficult challenges faced by retailers worldwide

- Numerous factors (e.g. promotions, competition, holidays, seasonality, locality) affect sales

Your analyse historical sales data collected from a large drug store chain in Europe--ROSSMANN

- Gain understanding / insight about some of the ways in which data can be fully prepared to optimise its analytical value

# Description of the business context

**Overall business objective is to predict 6 weeks of daily sales for 1,115 stores located across Germany, by building a sale forecasting model**

Challenges and tasks:

- Sales is affected by various factors
- Major data preparation tasks
  - Data integration
  - Visualisation
  - Cleaning and transformation
  - Missing values
  - …

# Available datasets: stores.csv

| Column | Description |
|---|---|
| Store | the anonymised store number |
| StoreType | 4 different store models: a, b, c, d |
| Assortment | an assortment level: a = basic, b = extra, c = extended |
| CompetitionDistance | distance in meters to the nearest competitor store |
| CompetitionOpenSinceMonth | the approximate month of the time when the nearest competitor was opened |
| CompetitionOpenSinceYear | the approximate year of the time when the nearest competitor was opened |
| Promo2 | a continuing and consecutive promotion, e.g., a coupon based mailing campaign, for some stores: 0 = store is not participating, 1 = store is participating |
| Promo2SinceWeek | the calendar week when the store started participating in Promo2 |
| Promo2SinceYear | the year when the store started participating in Promo2 |
| PromoInterval | the consecutive intervals in which Promo2 is re-started, naming the months the promotion is started anew. e.g., "Feb,May,Aug,Nov" means each round of the coupon based mailing campaign starts in February, May, August, November of any given year for that store, as the coupons, mostly for a discount on certain products are usually valid for three months, and a new round of mail needs to be sent to customers just before those coupons have expired |

# Available datasets: train.csv, test.csv

train.csv: Historical sales data from 01/01/2013 to 31/07/2015

| Column | Description |
|---|---|
| Store | the anonymised store number |
| DayOfWeek | the day of the week: 1 = Monday, 2 = Tuesday, … |
| Date | the given date |
| Sales | the turnover on a given day |
| Customers | the number of customers on a given day |
| Open | an indicator for whether the store was open on that day: 0 = closed, 1 = open |
| Promo | indicates whether a store is running a store-specific promo on that day |
| 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 |

test.csv: Identical to train.csv, except that Sales and Customers are unknown for the period of 01/08/2015 to 17/09/2015.

# General requirements

- Collaborate with your group members to understand the business problem and lay out the data pre-processing plan
- Please email you slides and a ten-minute video on your group's analysis plan to eghbal.rahimikia@manchester.ac.uk by 3:00pm, 12th December 2023
- Attend a 15-minute slot during class time on 13th December to receive feedback on your analysis plan.

- After that you have to work individually on a report of 1500 words (weight: 55%)
- Deadline for individual report submission: 9th February 2024

# General requirements

Your work should cover (but not be limited to) the following:

- Review the available data and describe it in terms of its variables, quality, and relevance to the sales forecasting
- Link data sets together as appropriate,
- Pre-process the data as appropriate for further analytics
  - Encoding categorical data, creating new variables, dealing with MVs, …
- Identify the key factors affecting sales
- Build a forecasting model (e.g. regression model, neural networks) using the variables you identified.

# Indicative breakdown of marks

| Assessed report | % |
|---|---|
| Introduction | 15 |
| Methodology (major data pre-processing tasks) | 35 |
| Results (description, discussion, analysis, etc.) | 25 |
| Conclusion, implications and recommendation | 15 |
| Layout and presentation | 10 |