Pre-processing task

Student number: 8520000

Introduction

In this report I will be looking at sales data from a large drug story chain. I will begin by looking at the data provided and do some initial pre processing. I will then conduct some exploratory analysis of the data. Based on the results, I will conduct further pre-processing. Finally I will fit a very basic model and interpret the results.

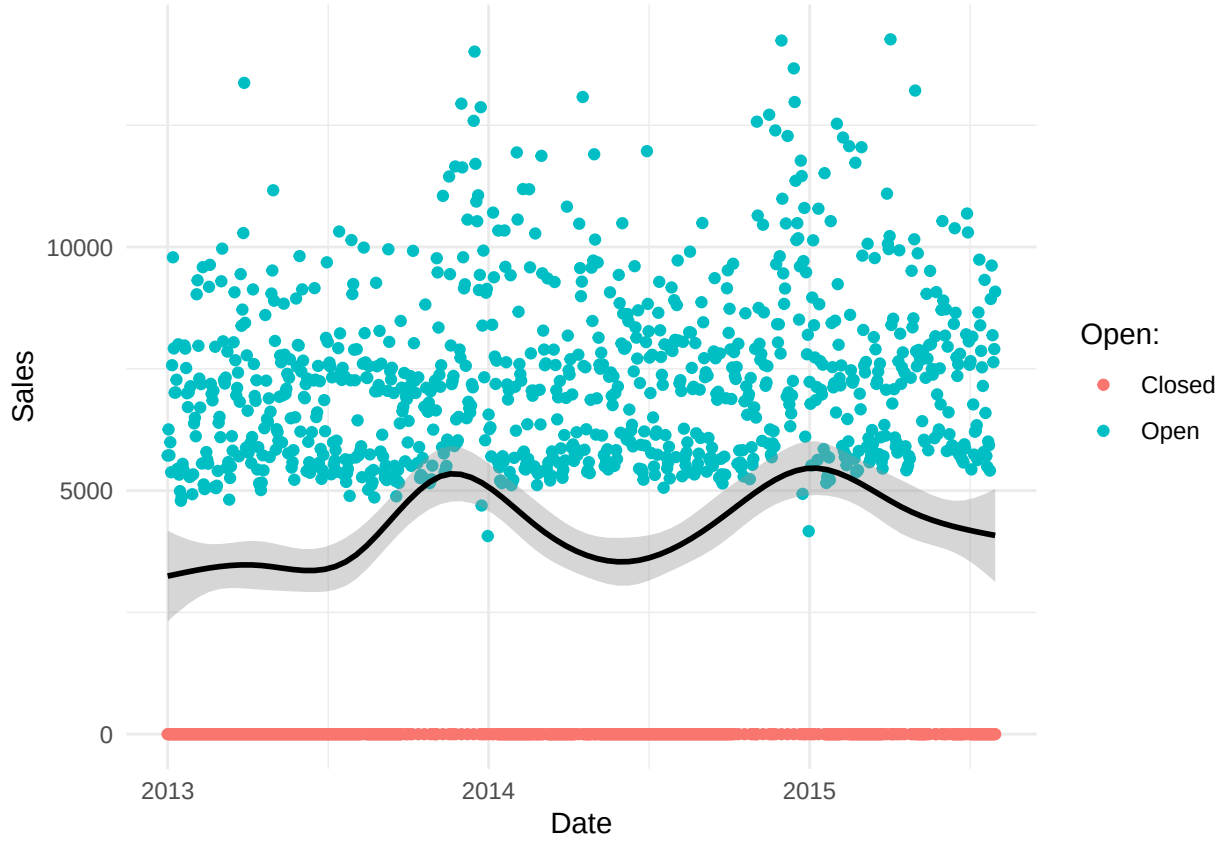
Pre-processing

We are provided with three data sets, one with information on various stores, and a training and testing data set with information on sales, customers. The first preprocessing step is to join these data sets, on the store variable.

Exploratory analysis

First we can visually inspect how sales vary over time.

To do so the mean amount of daily sales was plotted over time. To visually aid interpretation, a generalised additive model was fit to the sales data and it prediction line was plotted in black. Aditionally, we can look at whether stores being open or not has an effect by colouring the points in our graph according to whether the stores were open or not.

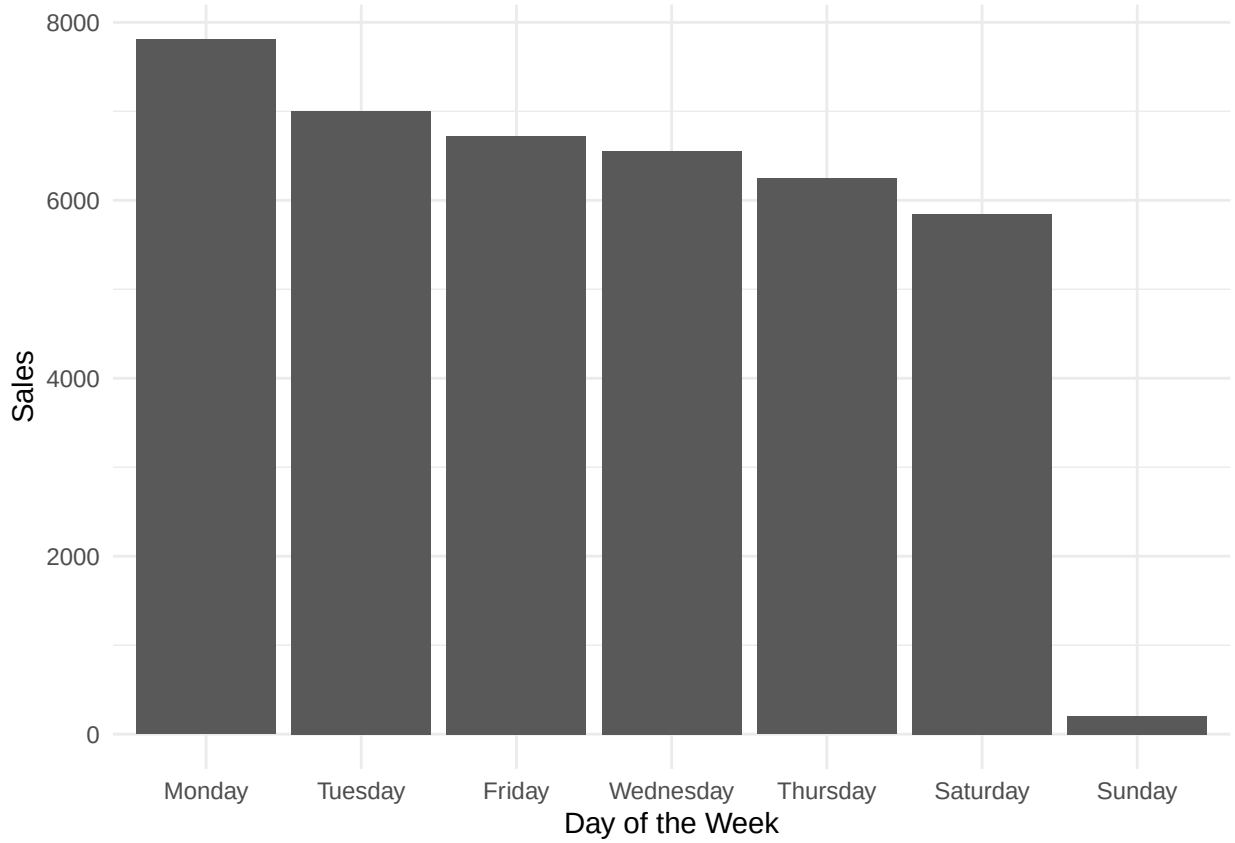


There are a number of insights provided by this graph.

First, looking at the colours, we can see that when stores are closed sales are always equal to zero. Instead of fitting a model to make predictions for sales on days when stores will be closed, we can instead assume that they will always be zero.

Second, looking at the points and the fitted prediction line, we can see that sales seem to be higher winter than in summer, and there appears to be very high sales around the holiday season every year.

We can look in more detail at the variation in sales over time. For example we might also want to find out if certain days of the week tend to have more sales than others. To do so we can create a box plot showing the mean amount of sales for each day of the week.



We can see from Monday to Saturday, daily sales average between approximately 6000 and 8000. However for Sundays the average amount of sales is much lower (≈ 204). We might want to consider fitting a seperate model for Sundays.

Further pre-processing

Conclusion