For this final project we were asked to work in a real-life project (available in public sources or not) and make use of the concepts learned and used during the semester. While it is a great way to deep on this knowledge it also involves some research and learning how to deal with an specific problem from the industry. In your case, predict 6 weeks of daily sales for 1,115 drug stores located across Germany.

# Project Understanding

The challenge proposed by Rossmann Stores seems to be well understood by the team. A good explanation of the company's needs is given and also an explanation of the expected outcome.

Do you have any additional knowledge about the business after doing some research? E.g. insights on how to deal with this specific problem according to success-cases from academic papers.

At the same time, the *head()* output for the train dataset doesn't add much useful information to the problem description; we think it's more adequate for the data understanding section.

The link to the dataset/competition on Kaggle isn't working. By removing the final "s" in "stores" should work: <https://www.kaggle.com/c/rossmann-store-sales>

# Data Understanding

In respect to the document organization, the Exploratory Data Analysis seems to be pretty late within the document. We think it would suit better immediately after the Problem Description section. Also the PCA analysis should be included as part of it.

On a first glance of the section, it looks like there're to much images, some of them oversized, and a poor explanation of them.

Additionally, a description/explanation of each feature is missing. What data do we have available?

Great analysis of the sales by store type and day of week. However, an explicit say of the day that you're considering as day 1. For example, if Monday is treated as the first day of the week you can reference the international standard ISO 8601. The same for promo and day of week, and at the same time we noticed that your promo variable is Boolean while the legend at the side of the plot shows a continuous Promo variable. By setting this variable as factor and assigning it to the *group* parameter in ggplot it can be solved.

The last graph isn't explained, and actually, does it add important information or it's superfluous? The customers per day distributions seems to be very skewed according to its density plot, are you going to mention something about this?

Due that the PCA analysis isn't too meaningful for your model, we suggest not to spend that much space in the report for that. We have limited space so it's important to keep it for the most meaningful information. You mentioned "we will explore more feature engineering as the analysis continues" during the PCA analysis, be careful, PCA is not feature engineering.

You didn't mention this, but do you have sufficient data for modeling or would you have to come with new features by doing feature engineering?

Any variable isn't relevant to solve the problem? We suggest to create a basic linear regression model to analyze the P values and evaluate the importance of each attribute for the model.

# Data Preparation

What do you mean by "the dataset was very clean"? There were not missing values, NAs, infinite values, etc.? There were not string variables? Outliers? It should reflect your findings from the EDA and use appropriate technical language that we get during the lecture.

Are you planning to do some feature engineering or transform the data to get the most out of it for modeling?

# Modeling

When you select lambda for your LASSO model, the support graphs are using log(lambda), but it's never mention/explained explicitly. We think it should be mentioned to make the report clearer.

How many folds are you using for cross-validation for LASSO? It isn't stated in the report either.

The explanation of the random forest model could be improved to reflect the concepts given in class. We didn't understand why you decided to use non-technical vocabulary in this case for that ("a group of 'weak' models combine to form a more powerful model").

The report states that changing the seed for random numbers generation is part of the tuning of a model. It isn't. Setting a seed for random numbers generation is just a way to make sure that you will be able to get the same results every time you execute your program, even when make use of random numbers.

You didn't mention how did you get the parameters for the Random Forest model. Did you do hyper-parameter tuning?

# Evaluation

In the Kaggle competition it says that submissions are evaluated using the Root Mean Square Percentage Error (RMSPE) but in your report it looks like you're using Root Mean Square Error (RMSE). We suggest to use the same metric that Kaggle uses in order to fit the models for the best performance.

Instead of including a screenshot of your Kaggle submission we think that a comparative table with the tuning and performance summary of the different models built (like in Homework #5) would improve the quality of the report.

# Overall/Summary

On the whole, your project could be benefited from:

* Reorganization of the document (Project understanding, Data understanding, Data preparation, Modeling and Evaluation);
* Deepen on Exploratory Data Analysis;
* Decreasing the size of images and plots, while adding more explanation;
* Be careful with the redaction, first-person singular shouldn't be used since you're a team; and
* The Executive Summary should be written as the final part of the report, summarizing all the most important part of the report in a way that reflects it content as a whole and be useful for making decisions.