Rossmann Lab – Part II

**Goals**

* Review the questions in the lab below, and consider the data from the outside, what is important? What can we learn from the data?
* Review the lecture notes for statistical analysis methods for analyzing a dataset
* Perform the exercises below and understand the rationale behind them, ***come up with your own analyses, and document what you did and the results***

**Data description**

Rossmann store data has information about sales, promotions and competitors

Part 0: Rossmann dataset

Open **Rossmann Lab - Data.csv** and inspect the dataset. Make sure you understand the meaning of each column. Please bring any questions to class.

*\* Copy All the files in a folder and set that folder as your working Directory in RStudio.*

Part 1: Explore the data

**See if there are significant difference among the categories of a data variable, e.g.**

1. Do different store types have significantly different average sales, given other factors equal?
2. How about assortment type? Does it affect the sales?
3. Same idea applies to other factors as well, such as month of the sales
4. Are holidays and promos associated?

**Use regression analysis to decide on the predictive factors to sales**

1. From the previous exploratory data analysis, we have been able to identify the potential sales drivers:
   1. Promotions
   2. Store format
   3. Store assortment type (although associated with the store type)
   4. Seasonality (month)
   5. Day of week (week though)
   6. Competition distance (counter intuitive in terms of the direction)
2. Now we want to use regression approach to verify if put together, these factors can jointly explain the variation in store sales

Part 2: Correlation and regression

1. Identify whether store sales may be associated with the competitive distance or not. Perform the flowing
   1. Correlation analysis for store sales ~ competition distance
   2. As you’ve seen that the distance distribution is highly dispersed, i.e. the values are all over the place, you may want to transform the variable first and then perform the analysis
   3. Use Scatter Plot to visualize the relationship between the two, also plot both store sales ~ completion distance and store sales ~ log(competition distance)
   4. Use hypothesis testing to determine the probability that the slopes are zero

**Additional bivariate and multi-variate analysis**

1. Sales time series
   1. Seasonality - plot the sales monthly time series
2. Possible collinearity
   1. Store type ~ assortment type
3. Uncover possible causality:
   1. Sales ~ State promo Flag over time
   2. Sales ~ School promo Flag over time
   3. Promo flag ~ holiday flag
   4. Sales ~ store format
   5. Sales ~ assortment type

Part 3: Multivariate regression and hypothesis testing notes

**Test on the hypothesized comparative relationships among datasets**

1. Compare means (single means, two samples e.g.)
2. Compare variances
3. Test on the goodness of fit

Part 4: Multivariate regression and hypothesis testing analysis

1. Modify the regression models run and see if transforming the variables or adding an additional variable may further improve the model’s fitness.
2. on holidays, can you verify if this is true with this data? Hint:
   1. Filter for one year’s data using one store only to remove the repeated calendar dates, assuming all stores have the same promo schedule
   2. For this type of association analysis between two categorical variables, a Chi-square test is the right tool

Part 5: Identify the drivers to the sales and impact of promotions

1. Use your descriptive statistical analysis, degree of association analysis (covariance, correlation and regression) and comparative analysis (hypothesis testing) to answer the above question in 5 clear slides.