Rossmann Lab – Part I

**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

**Identify the drivers to the sales and impact of promotions**

1. Put on your detective hat and ask lots of questions
   1. What’s the format of this dataset?

csv file

* 1. How many stores / years / store formats / assortment types are in the data?

3345 unique combos of store / year/store type/assortment

[1115 = Stores, 3 = Year, 4 = Store Types, 3 = Assortments]

* 1. Missing values and outliers?

No missing values

Some outliers

* 1. Other idiosyncrasies, e.g. when store is closed, any sales and customers?

No idiosyncrasies as described above

* 1. What are the most interesting variables towards the goal of “identifying the drivers of sales”

Store, Day of Week, Year, Month, Promo, Holiday, Competition Distance

* 1. Any major pieces of information missing?

economic class of the location of its stores

* 1. Is the amount of data sufficient or is it too large? Does it require sampling? Which sampling method is best?

I think it'd be best to sample by year and distance from competition

see if any trends are occurring within the store sales

data is not too large

* 1. Any artificial bias that may exist in the data? E.g. only promo sales are included, data are truncated, certain months are missing?

only contains jan 2013 - july 2015 (not 3 full years)

Competition Distance may be truncated to not include decimal values

non-promo sales are included

data is not missing

**Notes on how to explore univariate data**

1. What kinds of data / variables are there?
   1. Continuous vs. discrete variables

Discrete: Store,day of week, year, month, date, sales, customers, open, promo, holiday, competition info

Continuous: Competition Distance...?

* 1. Coverage of years / regions / .../ any other attributes (store type, e.g.)

Month, Year, Day of Week, Holidays, Store Type, Competition Date Info, Promo

* 1. Time series / cross-sectional data?

Month, Year, Day of Week,

1. Examine the ‘cleanness’ of the data
   1. Missing values <- None
   2. Outliers <- Yes

Exist in Sales and Competition Distance

* 1. Logic integrity <- Yes
     1. E.g. when store is closed, are there any positive sales or customers
  2. Data consistency / common sense

<- collected when data was collected does not seem to be consistently everyday or at the end of each month...

1. Discover the high-level patterns in the data
2. Plot or chart the variable distribution
   1. Histogram
   2. Box-plot
   3. Frequency distribution

**Look at the variables of interest individually**

1. What does the sales’ distribution like?

Right Skew, Unimodal  ~ Normal between 0 and 20000

1. How about customer counts’ distribution?

Right skew, unimodal ~ Normal between 0 and 2000

1. What does the competition distance distribution like?

Right skew, unimodal ~ Normal between 0 and 20000

1. How many different holidays are there?

\*\*\* 3 different types of state holidays, 1 type of school holiday

1. How many store types and assortment types are included? What is their frequency distribution?

store types = 4

assortment types = 3

Part 2: Plotting and descriptive statistics

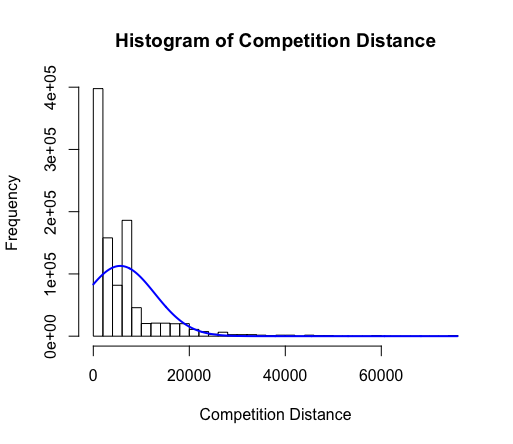
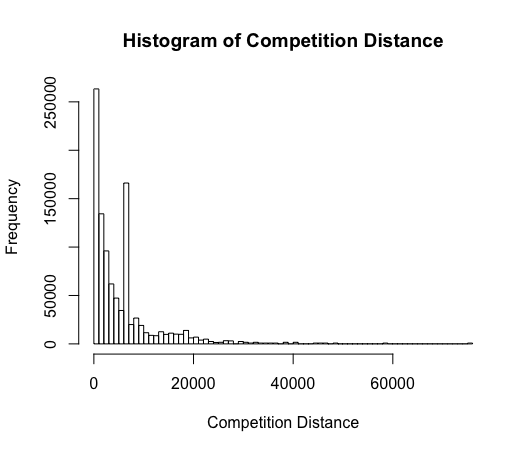
1. Perform descriptive statistics analysis on competitive distance (“CompetitionDistance”). Answer the question:
   1. Based on the values of the descriptive statistics, e.g. mean, mode, median, range, etc, what kind of a shape does this distribution may have?

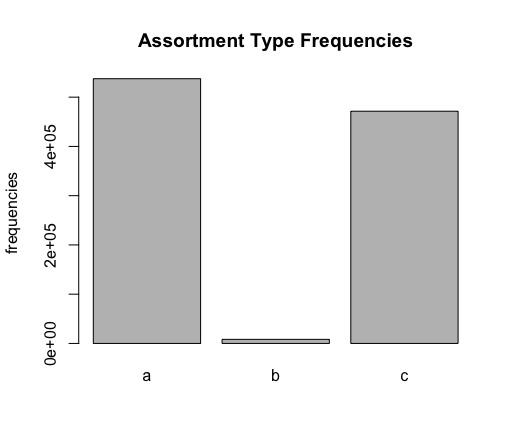
Unimodal, right skew

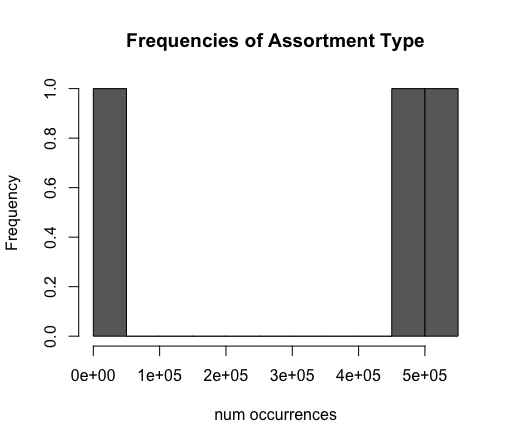
* 1. From the values of Mean, Median and Mode, can you tell whether its distribution is right skewed or left skewed?

Right Skewed

1. Plot the histogram of the variable “CompetitionDistance”
   1. Use different bins and see how that changes the shape of the histogram
   2. Select the best histogram and try plotting a cumulative percentage line over it for the count of stores in each bin

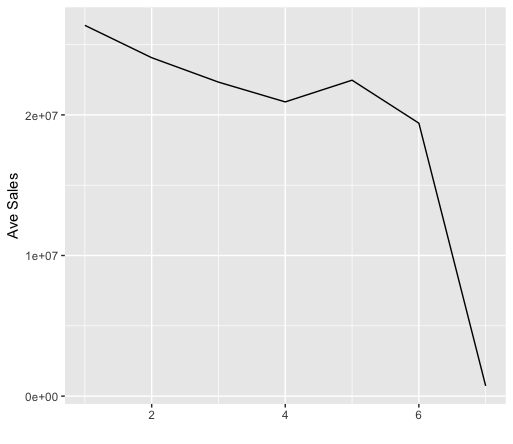


1. For a data field like “AssortmentType”, what are the good ways to represent this variable in order to see what are the most common assortment type, vs. less common etc. Hint:
   1. Table summary
2. asstmtType
3. a b c
4. 537445 8294 471470
   1. Bar chart <-- BEST CHOICE 
   2. Frequency plot (%)



NOTE: assortment type for each store repeatedly occur in the data along time (date)

1. For store sales, we also want to see if day of week may be contributing to the sales variation.
   1. Plot store sales time series to show the average store daily sales for seven days in a week.



* 1. Use data for year 2014 only.
  2. Explain the trends you see in the data

Hint: Randomly select a small number of stores use these stores only in plotting

The Sales decrease as the week continues but on Friday people shop more,

on Sunday sales drop quite low