In this assignment, you are going to explore data from the convenience store retail giant NANSE. NANSE owns over seven hundred small-form convenience stores throughout Canada. This dataset only includes sales from 2015 for food and beverage. It also excludes alcohol and tobacco. Each row represents the weekly average for one store for all of 2015. Thus, each row in the data represents the average weekly results for that one store for the whole year. You are going to use EDA to investigate different features of these data.

**Instructions**

ETL Tasks:

1. Import the `[store\_3HE.csv](https://compass2g.illinois.edu/bbcswebdav/pid-5342188-dt-content-rid-74567746_1/xid-74567746_1)` as dataframe ***df*.**Learn about the data from the [data description file](https://compass2g.illinois.edu/bbcswebdav/pid-5342188-dt-content-rid-74568257_1/xid-74568257_1).
2. (**0.5 points**) Familiarize yourself with the features of the data by using the `str()`, `summary()`, `head()`, `tail()`, `slice\_sample()`, `unique()`, and `n\_distinct()` functions. Give a brief summary of what you learn (just two or three sentences).  
   **Note** - A few of these functions are from the package “dplyr”. Thus, load “dplyr” or the meta-package “tidyverse” into your RStudio notebook. The package “tidyverse” is used in each R notebook we will work with, so, you will see it often.
3. **(0.5 points)** Convert the values in the **`revenue`**column to a numeric data type.
4. **(0.5 points)** Delete rows that contain missing values in any of the columns.

Exploratory Data Analysis Tasks:

1. **(1 point)** We are interested in understanding what factors influence the revenue for an individual store. From dataframe ***df***, create two dataframes ***df\_low* and *df\_high***such that ***df\_low*** contains data on all the stores with ***revenue*** less than the average ***revenue*** for the complete data and ***df\_high***contains data on all the stores with ***revenue*** equal or greater than the average ***revenue*** for the complete data. Calculate the average ***size*** of the stores in the two dataframes (***df\_lows*** and ***df\_high***) and comment on the relationship between ***revenue*** and ***size*** of the stores by discussing the two means (just one or two sentences).
2. **(1.5 points)** Report and comment on the correlation between ***revenue*** and ***size***. First, calculate the correlation between ***revenue***and ***size***in the main dataframe. Next, using the `ggplot()` function, plot the relationship between the ***revenue*** and ***size*** using an appropriate chart. Finally, discuss in one or two sentences.
3. **(1.5 points)**Create boxplots of ***revenue*** for each***region*** and comment on the distribution of ***revenue*** within each ***region*** (in one or two sentences).
4. **(1.5 points)** Calculate the correlation between the ***revenue*** and ***promo\_units*** and comment on whether the value aligns with your expectation. Based on this value of correlation, can you conclude that selling units on promotion helps in increasing revenue? Please explain your reason for the conclusion in a few sentences.
5. **(1.5 points)** Create a bar chart that shows total ***gross\_profit*** for each province. Comment on a few findings from the chart in one or two sentences.
6. **(1.5 points)** Create a correlation matrix for showing correlations among the average store sales for all product categories. Which two product categories sell together the least and which sell together the most?

**Submission\***

1. Answer these question in an RMD file.
2. Keep all code, visuals and comments in this RMD file. Then knit it into an html.
3. Save your submission as netid\_HW3.html. Compress the .html file as a .zip or .rar file for submission.
4. Submit the .zip or .rar file on Compass2g before the deadline.