COURSE 21-22

Data Analytics with R MIBA, 2022

Challenge 1: General overview and Data Treament

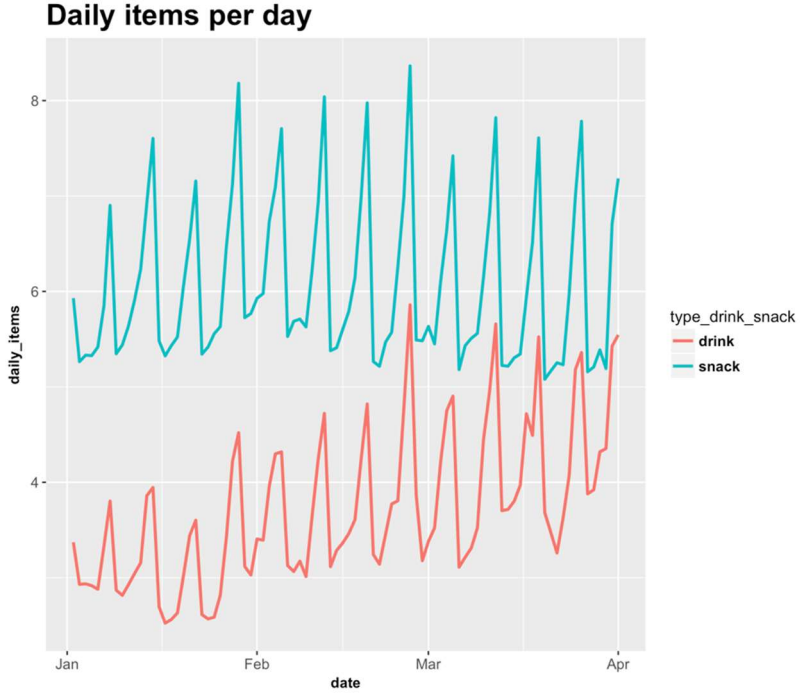
Gonzalo de la Torre
Rubén Coca

1. **[2 points]** General overview of the data:
   Machines:
   a. How many machines are there?
   b. What percentage of them are small?
      Hint: `dt %>% group_by(column_name) %>% n()` *will tell you how many rows in data table dt fulfill each column_name value*
   c. How do they distribute in terms of location type i.e. transport, petrol station?

   Products
   d. How many products are there? Which category has the highest number of products?
   e. Which category has the highest and lowest average price? And within snacks or drinks?

   Transactional data
   f. Restricting the transactional data to March 2017, what's the average daily items among small and big machines?. Why do you think there is such a difference? Give at least 2 possible reasons.
      Note: To calculate daily sales consider only the "active days" of a machine to exclude machine failures. For that, divide the number of items sold by a machine by the total number of "distinct" days.
      Hint: Check function `month()` *to do the date restriction.*

2. **[2 points]** Consider the following plot of the number of items sold per machine and day



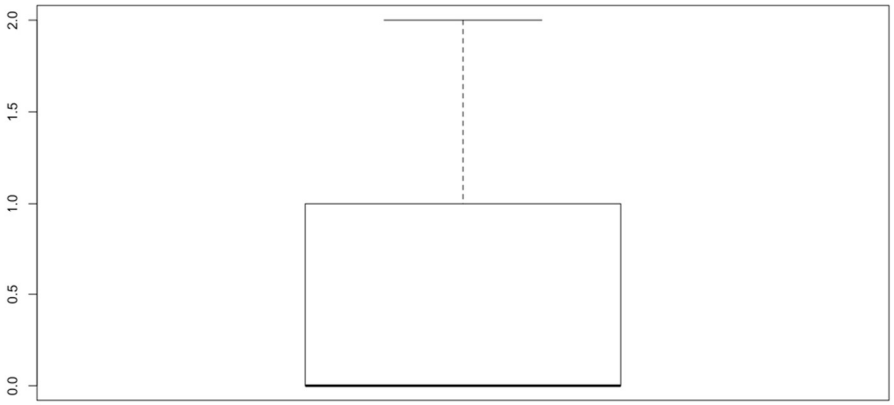
Let us focus on the different patterns present in the data:
   a. Is there a general trend in the number of snacks and drinks as the months progress from January to April? Is it the same for snacks and drinks? Why do you think that might be so?
   b. Is there shorter time period trend as well? Is it the same for snacks and drinks? What do you might be the cause?

3. **[2 points]** Given the following distribution of average income:

```
> summary(d$income_average)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  25540   48860   54100  162900   60960 9777000      97
```

a) Are there outliers? How would you treat them? Provide code with your answer
b) Can you give three possibilities on how to treat the NA cases? Which option would you choose and why? Provide code with your answer
Hint: Take a look at the relation between having NA in the income average and the average daily items.

4. **[2 points]** According to the following boxplot, what is the median number of hotels in the machine area?



5. **[2 points]** In this exercise we will build a location score that tells us what's the average daily items per machine depending on the location it is placed. This model could be used to a) Decide in which locations to place new machines b) Construct a benchmark for each machine: how much should it sell according to its location? This can be used to detect problems in machines (i.e. illumination, bad placement within a station etc.)

   For that, you will build a linear model to predict machine daily items using the following features:

   1. Machine size (big or small)
   2. Income of the area
   3. Number of routes in the area
   4. Number of hotels with 4 and 5 stars in the area
   5. 0-1 Indicator of whether the machine has train_AvgDailyPassengers informed meaning it is a petrol station or other type of location
      Hint: You can generate an indicator variable of this sort using the syntax
      `dt <- dt %>% mutate(new_column_name = ifelse(is.na(column_name),1,0))`
   6. Number of other Vendex machines in the area

   Answer the following questions:

   a. Do all variables show statistical significance? Which ones doesn't? How do you know?
      Hint: Recall that to check the parameters of a glm model called "model1" you need to run `summary(model1)`

   b. Build another linear model but this time instead of using the variables "total_number_of_routes_600 use the log of that variable in base 10 calling it "log_transport". Does this new variable show statistical significance?

Train the model constructed in the previous question (b) removing the variables that do NOT show statistical significance calling it final_model and with that model answer the following questions

c. How many daily items less do small machines sell all other factors remaining equal?
d. What's effect on machine sales does having other nearby machines all other factors remaining equal?
e. Ranking all machines according to the final_model, what are the real daily sales of the top20% machines with respect to your model prediction? And the real daily sales of the bottom20% machines according to your model? What's the top20%/bottom20% ratio?
f. Given the following 2 locations for a big machine:
    i. Supermarket entrance, 2 nearby hotels of 4 stars, 20 transport routes, no nearby machines
    ii. Transport station, no nearby hotels of 4 or 5 stars, 10 transport routes nearby, 3 nearby Vendex machines

   Which location would you choose and why?