Analytics for Observational Data (IT142IU)

Lab 1-2: Probability distributions

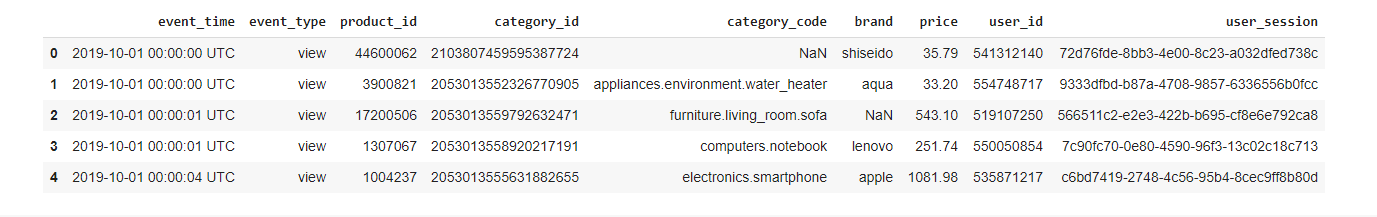
## Objectives

* Apply probability distributions to the provided datasets.
* Apply moment generating functions to find the moments of random variables.
* Dataset sources:
  + <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data>
  + <https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store>
  + <https://archive.ics.uci.edu/ml/datasets/climate+model+simulation+crashes>
* Programming languages: Python/Java

## Exploring the data

* + Data use: <https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store>

Data head:



|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | Ecommerce behavior |
| Identify data objects, attributes, and attribute types. |  |
| Find and choose the data objects changing over time. | - event time  - price |
| Identify and describe the data attributes that are considered as random variables from the chosen data objects.  Note: at least two data attributes chosen. | Variable: Price, Brand |
| Draw boxplots for each numeric attribute, present five-number summaries.  Note: recognize appropriate data areas in the data and draw boxplots | Ln[18] |
| Present the distributions of data regions of random variables using probability functions. |  |
| Find the first and second moments and central moments of the random variables. | Ln[21] |
| Remark random variables if they are useful for modeling or learning (classification or clustering) | Classification:  Price + brand -> product\_id |

## Some references

* Gallery of Distributions: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm>
* NumPy: <https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html>
* Seaborn: <https://seaborn.pydata.org/generated/seaborn.distplot.html>
* SciPy: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.moment.html>