## LTAT.02.002: Introduction to Data Science

# Bike sales in Europe

*A machine learning approach*

Elias Elias & Georg-kennet Gutman

# Business understanding

**Background:**

With the increase in the gas prices, people are preferably aiming to use alternative transportation methods for their daily translocations. Although a considerable number of people are interested in electric cars, countless people fancy having a bike to use for their short-distance or healthy trips.

The team members are willing to exploit the chance to achieve a great outcome and healthy status by meeting the expectations and demands of the customers' bikes preferences while enhancing the motivation to sell bikes (as a result of profitability). However, speaking of preferences, demand on bikes varies greatly across regions, sexes, age-groups, and even months. We see that if we don't have the appropriate and ideal bike available at the time of demand, this will lead to decrease in the likeliness of customers referring to the company and consequently a loss in the revenue (and vice versa!).

Provision of the favored bike depends on measurements and considerations that must be taken for accurate goals realization. This is where the drive is built – the need to serve the customer and not to lose profitability due to bike model unavailability. This approach depends on the time and efficiency of the analysis of the sales made considering the aforementioned variables.

**Business goals:**

The aim of this project is to implement a machine learning strategy that predicts the revenues (highest in focus) generated by the different variables. We are interested in building the big, complete picture of what contributes to a higher revenue generation, ultimately improving the decision making of when, where, and how much of the bikes models are best needed to be sold. This is because we want the business to continuously and increasingly grow through, and achieve, profitable sales.

**Business success criteria:**

1) Successful prediction of the preferred bike models.
2) Revenue increase realization
    a) basically due to having a clear idea of what needs to be sold.

**Inventory of resources:**

1) dataset:
    a) The dataset includes all the required features to be used for the analysis and prediction processes.

    b) We have detailed records of all the bike models and their quantities purchased by each age-group and gender and the dates purchased.

2) Python programming language will be used due to its free access and with all the important libraries required for analysis and prediction models.
3) The project is carried by two members.

**Requirements, assumptions, and constraints:**

The data has no restriction on the usage. License requirements stated to be unknown, and the publisher of the project (Sadiq Shaq) doesn't seem to treat the project as a confidential one. It is clearly stated the "*This Notebook has been released under the Apache 2.0 open source license*". Consequently, there are no assumptions or further requirements from the contributors. We believe that it is because this is not a competition, and will merely be citing the owner's name.

**Risk and contingencies:**

The only issue is the insufficient time availability for the complete focus on the project. This is due to the obligatory and inevitable balance we have to maintain between this project and the interference of other courses' obligations (like preparation for exams and other projects). However, we aim to allocate time at the beginning of the day to the other obligations, and have the rest of the day uninterrupted for focusing on this project.

**Terminology:**

So far, we only thought of the following:

1) **MSE**: Mean Standard Error
2) **MAE**: Mean Absolute Error
3) **Sub_Category (*dataset column*)**: The type of the item
4) **Product (*dataset column)***: The model of the item with its attributes (color, height).
5) **RMSLE**: Root Mean Square Log Error
6) **R2**: R-squared (a prediction evaluation method)

**Costs and benefits:**

We believe that the only resources we will be using are our own laptops. No other resources will be used, no outer materials will be utilized for the study. Hence, we don't see there is a way of performing cost-benefit analysis. The only benefit is our grade and the complacency we achieve with our skills enhanced by the end of this project. Nevertheless, speaking of the contributions we make with this project, a company will benefit a lot with the data-driven decision making.

**Data-mining goals:**

Data mining and analytics will be applied to study the relationship between the features of the dataset and how they relate to, and help us predict the revenue. We will be using Pandas, Numpy, and Matplotlib for data analysis and visualization. For the machine learning phase, we aim to train data using Ridge and Random Forest models, followed by an evaluation method.

**Data-mining success criteria:**

For the accuracy of the models and predictions, we want to evaluate the models by the Mean Square Error (so far). This also has us mentioning that we may go into Root Mean Square Error, and possibly MAE or RMSLE for perfect evaluation. We believe these three will result in an accurate modeling evaluation and predictive improvement compared to the MSE or R2.

# Data understanding

## Gathering data:

In our project we will look into bike sales in Europe, North-America and Australia. Our dataset comes from the Kaggle platform. The dataset comes from a Kaggle data scientist Sadiq Shah. The data comes as a 15.24 MB ".csv" (comma separated value) file and it is compatible with both Office Excel and Jupyter Notebook, which are the tools that will be used in the current project. Due to the goals of the project, the previously mentioned dataset is the only necessary data source that will be used in the project.

## Describing data:

The bike sales dataset consists of 113036 rows and 18 columns of which 9 are Integers, 7 Strings, 1 DateTime and 1 Other. Even though the dataset is named "Bike Sales in Europe" it also includes sales of different biking accessories and clothing, which will be excluded. The dataset columns consist of all the necessary features to accomplish the goals of this project - date of the sale, customer age and gender, location of the transaction, bought products and information about the unit sold. The dataset does not have any missing or "weird" values. On the other hand the dataset has 87054 rows of information that we are not interested in, for example “Accessories” from the “Product_Category” column. That is because our goal is to study only bike sales.

## Exploring data:

The features in the dataset are quite straight-forward and easy to understand. That makes the exploration of the dataset quite simple. The data has been collected from 2011 to 2016. Most of the bike sales were done in 2014 and 2016 by mostly Adults (35-64) age group. In the 6 years (2011-2016) it is possible to see a decline in sales during winter months at the end of the year. At first, the data exploration step did not bring out any data quality issues.

## Verifying data quality:

For our intended use the data did not seem to have major quality issues. It is questionable whether the data is from real-life or not, because the metadata does not include any license or general info whatsoever. On the other hand, for this project the source of the data is not so important. The dataset itself has the relevancy for the project goals. Also, the data is complete and does not include any missing values or data records. Some minor issues could be that the data is not up-to-date and it does not have the information about every European country. Because of that it is impossible to generalize the results to the whole of Europe. In general the quality of the dataset is sufficient for the project and can be used to proceed to next steps.

## Task 4. Planning your project

Our primary goals are to 1) analyze the dataset to determine preferred bike models per each age group and gender, 2) study the contributions of the models to the revenue and use the data to predict the revenue and 3) use data visualization to understand the changes in bike popularity in the society.

Task 1: Clean the dataset from anomalies and remove “accessories” and other product categories to focus only on preferred bike models category.
Task 2: Visualization of the dataset: plotting of features grouped by gender, revenue, countries, dates and products. For plotting we are going to use Pandas and Matplotlib.
Task 3: Finding correlations between features.
Task 4: Performing feature engineering with the aim of creating relevant features and transforming features for modelling.

Task 5: Split the dataset into two datasets: training (90%) and testing (10%).

Task 6: Performing the training using the Ridge and Random Forest models. Then, we perform evaluation methods (MSE/MAE, RMSLE) to compare the efficiency of the models to be used for later prediction/analysis tasks.

Task 7: Analyze the trained model for its prediction efficiency and accuracy.

Some tasks may be repeated for accuracy (like task 5 and 6 for example)

Georg performs tasks 1-3 and Elias performs tasks 4-6. Work hours are estimated around 30 hours for both students.