KOÇ ÜNİVERSİTESİ

INTRODUCTION TO DATASCIENCE USING PYTHON
ENGR350

PREVIOUS PROJECTS

Project : The main aim of the project is to provide a classification system for identification of humpback whales using machine learning.
Data source : Regarding test and training datasets, they have already been published on Kaggle by the competition host “HappyWhale” [3]. Since the dataset is not sufficient to build and train a CNN, applying several data augmentation techniques such as image rotation, gray scale noise adding etc. is planning to be carried out to expand the dataset size.

Project : Pesticides Effect on Honeybee Colonies and Honey Production in USA between 1998-2015
In this project, the change in yearly number of honeybee colonies and their honey production due to the amount of pesticides used was investigated. So the data frame used should have the data of number of honeybee colonies, their total honey production and the amount of pesticides used from a selected range of years.
Data source: Kaggle

Project : Student instructor relations and student satisfaction reasons. Our analytic approach is mostly descriptive and diagnostic where we are trying to comprehend why students answered the survey questions the way they did and how it can be interpreted.
Data source: The data set we used was found from UC Irvine Machine Learning Repository. It involved 28 survey questions that is responded by 5820 Gazi University students.

Project : Our project aims to check whether there is causality between Ramadan month and traffic accidents. If there is we will have a solid argument for authorities about this issue.
Data source: There are 2 different data sources. First, from trafik.gov.tr, one can access yearly reports regarding traffic accidents until 1978. There reports are in pdf format, and their structure change every year. In order to test our hypothesis, we need high frequency data. The other data source is the traffic data at Emniyet Genel Müdürlüðü. We have made a formal application to retrieve the data as a researcher from Koç University

Project : After Microsoft acquiring the leading collaborative version control system Github started lurking on the internet, the developer community quickly started to look for alternatives. As future developers, we thought it might be interesting to see what the trends in the developer community were currently like. We thought it would be a reliable way to get and compare data from Github and a competitor, Bitbucket, in order to get a more clear understanding of the current state of the developer community.
Data source : Most of our data comes from Github and Bitbucket, obtained via their APIs. The

Project : The aim of the project is solving a Traveling Salesman Problem (TSP) using optimization packages in Python. TSP is the problem of determining the shortest possible route that visits each city and returns to the origin city with using a list of cities and the distances between each pair of cities. With our program, we aim to implement the techniques of data importation and transformation that we learned in class to a well-known optimization problem and after solving the problem, we aim to visualize the solution using the data visualization packages in Python.

INTRODUCTION TO DATASCIENCE USING PYTHON

ENGR350

Project : this project will provide consumers and producers with nutrition levels of food products and offer healthier alternatives for products which are considered unhealthy to humans and environment. Orders, products and their nutrition values of an online grocery store “Instacart” will be used to provide customers with data that would help them find healthier alternatives and influence their consuming behaviours.

Data source : Data to be used in this experiment are taken from “Instacart Market Basket Analysis” competition from Kaggle which contains a list of product names/IDs, names/numbers of aisles of the Instacart market, food categories and customer orders/IDs. This data will be used alongside the food product names and nutrition values data taken from the API (application programming interface) of United States Department of Agriculture (USDA) Branded Food Products Database. Since the data is obtained from two different sources, they will need additional adjustments / modifications; unnecessary data will be eliminated and food product names from Instacart data will be matched with corresponding nutrition values inside USDA database which will form the data of this project alongside orders of customers. Instacart data input is in csv format and USDA input will be downloaded using an API. Data will be read into Python using the pandas library. 49688 distinct products and 3.4 million orders will be stored inside pandas data frames.