# EDA - Car sale Advertisement Dataset

Exploratory data analysis using Python, NumPy, Pandas, Seaborn

## Task

* Data Loading and Description
* Understanding the Dataset
* Preprocessing
* Univariate analysis, Bivariate analysis and multi - variate analysis.
* Data Visualization

## Data Loading and Description

* The dataset consists information collected from car sale advertisements for study/practice purpose where most of their used cars.
* The dataset comprises of 9576 observations of 10 columns.

**Some Background Information**

* This data was collected from private car sale advertisements in Ukraine and provided by INSAID team to perform Exploratory Data Analysis.
* This dataset has real raw data which has all inconvenient moments (as NA’s for example).
* This dataset contains data for more than 9.5K cars sale in Ukraine. Most of them are used cars so it opens the possibility to analyze features related to car operation.

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| car | Manufacturer brand |
| price | Seller’s price in advertisement (in USD) |
| body | Car body type |
| mileage | as mentioned in advertisement (‘000 Km) |
| engV | rounded engine volume (‘000 cubic cm) |
| engType | type of fuel (“Other” in this case should be treated as NA) |
| registration | whether car registered in Ukraine or not |
| year | year of production |
| model | specific model name |
| drive | drive type |

**Dataset link:**

<https://raw.githubusercontent.com/insaid2018/Term-1/master/Data/Projects/car_sales.csv>

1. EDA in mobile price range prediction

Task:

The objective is to do a complete analysis EDA. Provide valuable insights using python

About Dataset:

Price range : This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

The file contains 21 fields that impact the price range prediction and they are as follows.

## **Data fields**

* **id**- Unique ID for each phone
* **battery\_power**- Total energy a battery can store in one time measured in mAh
* **bluetooth**- Has bluetooth or not
* **clock\_speed**- speed at which microprocessor executes instructions
* **dual\_sim**- Has dual sim support or not
* **fc**- Front Camera mega pixels
* **four\_g**- Has 4G or not
* **int\_memory**- Internal Memory in Gigabytes
* **m\_dep**- Mobile Depth in cm
* **mobile\_wt**- Weight of mobile phone
* **n\_cores**- Number of cores of processor
* **pc**- Primary Camera mega pixels
* **px\_height**- Pixel Resolution Height
* **px\_width**- Pixel Resolution Width
* **ram**- Random Access Memory in Megabytes
* **sc\_h**- Screen Height of mobile in cm
* **sc\_w**- Screen Width of mobile in cm
* **talk\_time**- longest time that a single battery charge will last when you are
* **three\_g**- Has 3G or not
* **touch\_screen**- Has touch screen or not
* **wifi**- Has wifi or not
* **price\_range**- \*\*(Target)\*\* A number denoting the price range of the phone [0-4]

Dataset:

<https://www.kaggle.com/c/mobile-price-range-prediction-is2020/data?select=train_data.csv>

# 3. House-Sales-in-King-County-USA

Analyse and predict housing prices using attributes or features such as square footage, number of bedrooms, number of floors and so on.

# House Sales in King County, USA

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

**id** : A notation for a house

**date**: Date house was sold

**price**: Price is prediction target

**bedrooms**: Number of bedrooms

**bathrooms**: Number of bathrooms

**sqft\_living**: Square footage of the home

**sqft\_lot**: Square footage of the lot

**floors** :Total floors (levels) in house

**waterfront** :House which has a view to a waterfront

**view**: Has been viewed

**condition** :How good the condition is overall

**grade**: overall grade given to the housing unit, based on King County grading system

**sqft\_above** : Square footage of house apart from basement

**sqft\_basement**: Square footage of the basement

**yr\_built** : Built Year

**yr\_renovated** : Year when house was renovated

**zipcode**: Zip code

**lat**: Latitude coordinate

**long**: Longitude coordinate

**sqft\_living15** : Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area

**sqft\_lot15** : LotSize area in 2015(implies-- some renovations)

Dataset link:

https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DA0101EN/coursera/project/kc\_house\_data\_NaN.csv

# Exploratory Data Analysis Medical Appointments Data

Each patient’s record is characterized by the following features:

* PatientID — a unique identifier of a patient
* AppointmentID — a unique identifier of an appointment
* Gender
* ScheduledDay — a day when an appointment is planned to occur.
* AppointmentDay — a real date of an appointment
* Age — a patient’s age.
* Neighborhood — a neighborhood of each patient
* Scholarship — Does the patient receive a scholarship?
* Hypertension — Does the patient have hypertension?
* Diabetes
* Alcoholism
* Handicap
* SMS\_received — Has the patient received an SMS reminder?
* No\_show — Has the patient decided not to show up?

# Task:

# Perform exploratory data analysis, data cleaning, data preprocessing, data visualization

# Dataset link:

# <https://www.kaggle.com/joniarroba/noshowappointments?select=KaggleV2-May-2016.csv>