Phase 1 : initial EDA

Smart Exploration Problem

Business requirements detailed in Phase 0 with 2 questions detailed below

- Question 1. How the company can make better decision based on current marketing campaigns ?
- Question 2. How can we optimize and sell the advertisements ?
  Can we use some methodology to achieve better results with higher confidence ?

> **Info** The aim of this notebook is to describe the approach to the problem and the steps to resolve it

Approach

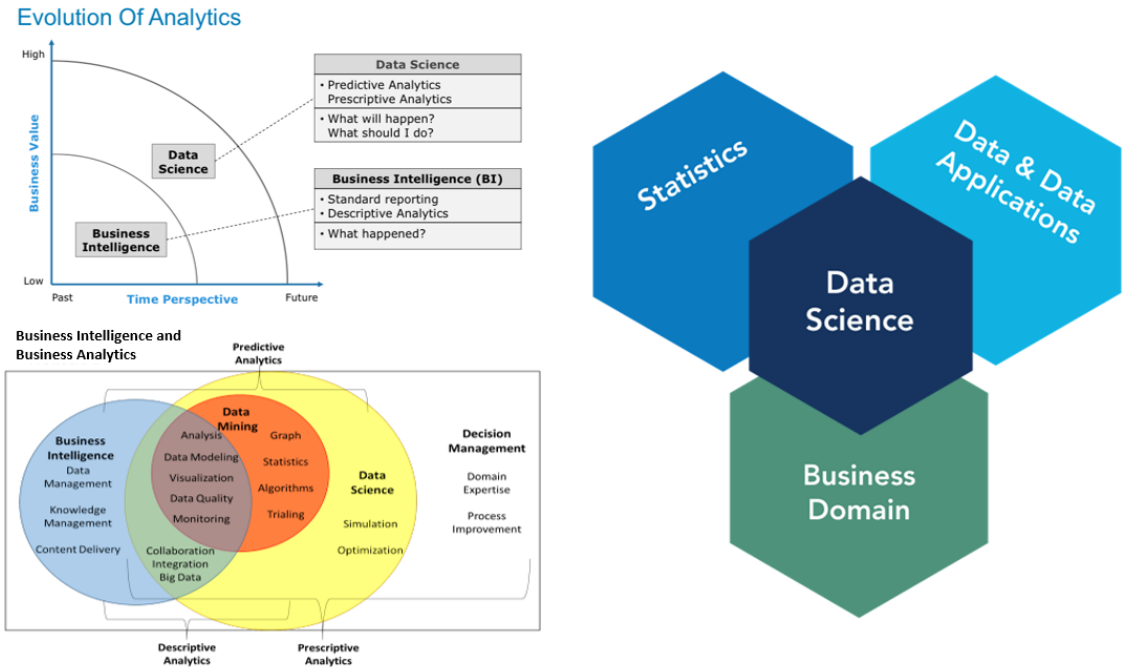
- The approach to this problem will use Business Intelligence concepts and methodologies to answers the first question
- Second question will be resolved using Advanced Analytics/Predictive Analytis to build machine learning models to predict and optimize best campaigns/Ad's in terms of CPE (Cost per Engagement)

The picture bellow show some concepts and methodologies that is going to be used in this Notebook

In [1]:

```
from IPython.display import Image
Image(filename='./data/img_Approach.png')
```

Out[1]:



Demo information provided by IT related to all current marketing campaign

In [2]:

```
import pandas as pd
import pandas_profiling as pp
```

In [3]:

```
df = pd.read_csv('./data/Historical_Data_Smart_Exploration_Demo_Simulation__CPE.csv')
df.head()
```

Out[3]:

	LineItemsID	URL	xyz_campaign_id	channel	channel_ad_id	gender	age	interest	spend
0	1	URL-1	916	1	2	M	45-65+	Interest - 15	1
1	2	URL-1	916	11	2	M	45-65+	Interest - 16	2
2	3	URL-1	916	11	1	M	45-65+	Interest - 20	0
3	4	URL-1	916	11	1	M	45-65+	Interest - 28	1
4	5	URL-1	916	1	1	M	45-65+	Interest - 28	1

Question 1 - How the company can make better decision based on current marketing campaigns ?

Answer (question 1)

The first part of the problem is to understand which historical data can be used to make decision on the current campaign's

```
Obs.: Line Item ID is auto increment number and because this could not provide any
useful information and
  was discarded in the analysis
```

All information (except one: Line Item ID) provided in the historical data could provide useful information and also gain insights to make decisions. Using Business Intelligence concepts and tools to fix this problem (descriptive analysis) that based on historical information business users design and define how to improve their business performance.

- One point that could be mentioned is the use of AI and machine learning to find patterns in the historical data that could also drive better investigation and understanting about what can be done on the current campaign's. (Top 3 features at the end)

Decision management and action plan could be triggered on current campaigns, based on answers from historical data

Example 1

- Which campaign (or top 100 campaigns for example) can achieve a great margin on the cost of CPE ?

  Option 1: take the average of CPE by campaign, channel, channel_ad_id, gender, age, interest... and choose the best ones that do not exceed a minimum threshold (for example, the company has the target of maximum CPE of US$ 0.25 - this must be decided by sales and marketing team). With these campaigns that give great margin to the company, so prioritize them

  Option 2: Prioritize ir order all campaigns with higher engagements and lower CPE
  Obs.: It is possible to have lower CPE (lower engagements and lower spent), but the idea is to get higher engagements to drive also other ad's
  from the same campaign. If the actual schema have time period, choose newest historical information if possible

Other business decisions could be done based on questions/answers below

- Which channel is taking the biggest budget and with which margin related to the others?
- Which age distribution provide best margin (in terms of CPE), spend/engagements for specific channel and URL?
- Does the gender influence how many clicks are done by device? and other features?
- How many clicks/spend we must achieve to get a minimal CPE by gender in general?
  Does these clicks converted in a better margin of CPE?
- Which geo location provide the better margin of CPE ?
- Which geo location provide better clicks/spend related to each type of campaign?
- What are the niche users that usually have the most engagement in campaigns? Does location and age influence this decision?
- How is the CPE distribution by interest, device and so on...
- Which campaign can we prioritize based on the best performance (in terms of CPE - historical data)

Technical info

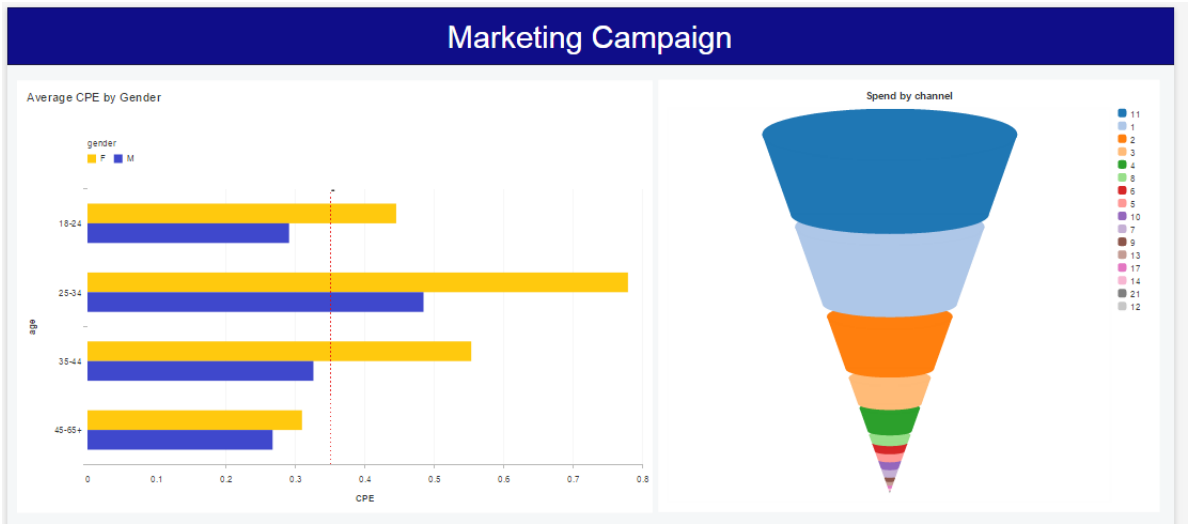
Almost all questions and answers above could be achieved with SQL, relational databases and BI tools.
This analysis can also be complemented using machine learning models to include additional and relevant information that can be consumed by the Line of Business users at the company such as analyst, managers and decision makers

- One example will be provided at the end of this notebook using python (basic simulation). The execution of one machine learning algorithm (CPE as target) will provide the most important features/segmentation of the historical data related to the CPE (Cost per Engagement) that correlates business intelligence and feature investigation

In [4]:

```
from IPython.display import Image
Image(filename='./data/img_CPE_Spend_by_Gender_and_Channel.PNG')
```

Out[4]:



BI Analysis - 1: People with age higher than 45 years used to have lower CPE in general
BI Analysis - 2: Channel 11, 1 and 2 have higher spend by the people

Question 2 - How can we optimize and sell the advertisements ?

Can we use some methodology to achieve better results with higher confidence ?

Answer (question 2)

This second part of the problem we are going one step forward with predictive analysis / advanced analytics using the data-driven approach to build and deploy machine learning models. The target is to assign a priority to each lineitem of the current campaign so that we start exploring the ones with higher chances to perform well (in terms of CPE)

The target of the defined problem is to prioritize campaigns with higher chances to perform well (in terms of CPE)

First analysis and considerations

This dataset do not provide some type of information that can shuffle the analysis such as geo location and language. This information context can change drastically how the campaign achieve better or worse results in terms of CPE

For example, people from USA and UK with primary language in English may have some results for specific campaigns and people with Latin culture
and language such as French, Spanish and Portuguese from Europe and Latin America may have other characteristics despite all the globalization
of customs and economy. People with Latin culture that speaks Spanish and leaves in USA may also contribute in different ways related to CPE

One solution to fix this possible issue could be the analyses and predictions with segmentation of this information by country/geo location
and language for example. This approach must be checked if better results could be provided

Methodology and options to deploy machine learning models

- First point is to follow CRISP-DM methodology

- Probably ensemble tree based models will have better accuracy results as all historical data is tabular data. Stacked ensemble could also be an option

- Use of regression models (as CPE is a continuous metric) and to evaluate the discrepancy between the prediction and the actual information could be done with metrics such as RMSE (root mean squared error) or MAE (mean absolute error) and the definition to choose which metric to achieve a better accuracy in the model are correlated with the outliers (if exists) and its distribution in the historical dataset

- CPE probably have skewed distribution and if this scenario confirms it is going to be used CPE in logarithmic scale could provide better prediction

- Create and test different machine learning algorithms to evaluate which provide better results and also try to build models to test the accuracy with partial features, trying to identify the best solution that could explain why partial matches may have better CPE

- Try to build machine learning models with latest information to catch as much as possible all characteristics that is influencing the customer decision. Excellent campaign for company profit could be out of date and could not generate the same results in the future. This could be done only if the time period is presented in the actual schema

- Another approach that may lead to more insights could be the creation of some features / labels. Samples:

  (1). Higher_engagement_feature (Yes of No) based specific threshold: IF ENGAGEMENT HIGHER THAN XXX then Yes otherwise No
  this feature could also be permutated between historical and current data to use for new predictions (must be evaluated)

  (2). To complement the target of CPE, 2 new features could be created as target such as Campaign_Success (Yes or No) and CPE_profitability (High, Medium and Low) according with company rules.

  These 2 new features, Campaign sucess(Yes or No) and CPE profitability(High, Medium or Low) can also

be used to complement the prediction analysis and clustering(identify and compare possible niche of campaign, ads and users).

This new information also permit to create experiments/predictions associated with binomial and multinomial classification machine learning algorithms that could lead to more useful insights.

- Creation of new features (feature engineering) targeting better accuracy for model prediction and keep in mind to not overfit the solution

Demo: basic simulation - Smart Exploration Problem with historical data

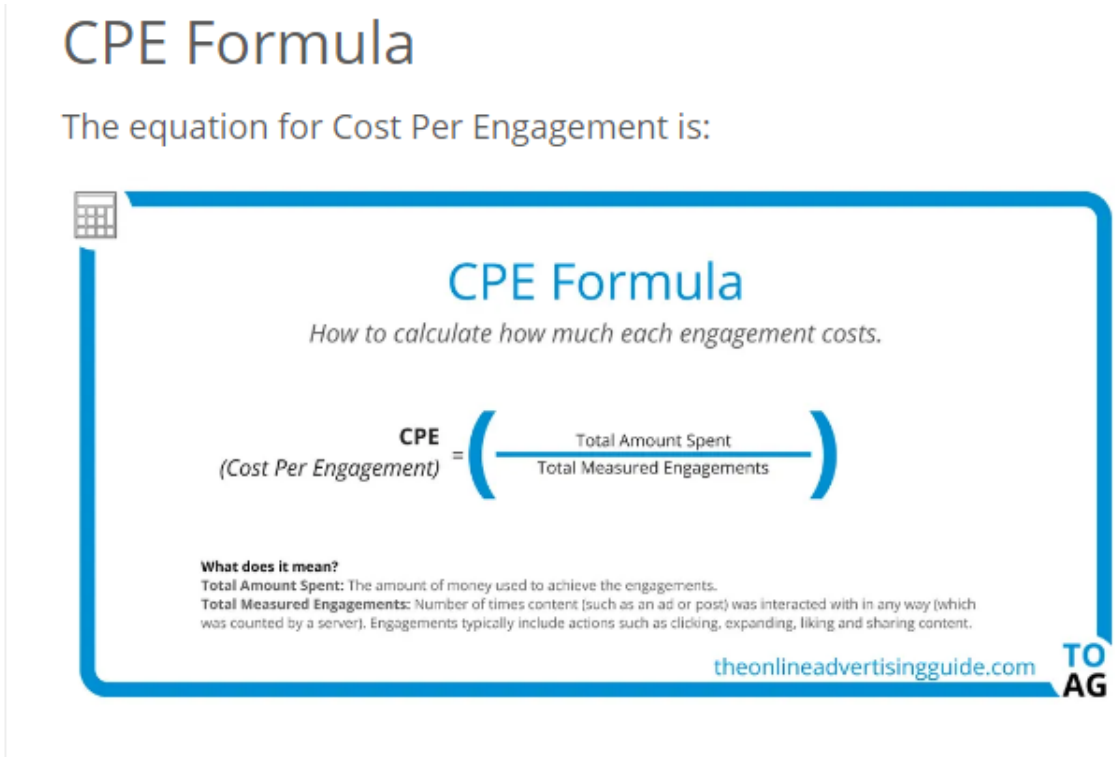
The idea with this simulation below is to present some options on how to start to evaluate the historical information and start to define best ways to assign a priority to campaign with higher chances to perform well (in terms of CPE)

CPE could be defined as

In [5]:

```
from IPython.display import Image
Image(filename='./data/img_CPE_sample_formula.png')
```

Out[5]:



Info: Possible solution: Convert a Regression model (PREDICT CPE) into a Multiclass Classification problem

As presented above - section Methodology (build additional fields), the solution could be a transformation of

Regression problem (PREDICT CPE) into a multiclass classification (PREDICT LEVEL OF CPE) converting the values/range of CPE into 5 levels for predictions. Examples

- Level 1: Best CPE
- Level 2: Good CPE
- Level 3: Standard CPE
- Level 4: Bad CPE
- Level 5: Worse CPE

but let's keep simple and move on with the regression problem and make prediction of CPE and start a quick EDA using python

EDA (Exploratory Data Analysis)

In [6]:

```
df.describe()
```

Out[6]:

	LineItemsID	xyz_campaign_id	channel	channel_ad_id	spend	CPE
count	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000
mean	572.000000	1067.382327	6.323710	2.855643	33.382327	0.410814
std	330.099985	121.629393	4.807888	4.483593	56.896994	0.695316
min	1.000000	916.000000	1.000000	0.000000	0.000000	0.000000
25%	286.500000	936.000000	1.000000	1.000000	1.000000	0.010000
50%	572.000000	1178.000000	9.000000	1.000000	8.000000	0.100000
75%	857.500000	1178.000000	11.000000	3.000000	37.500000	0.480000
max	1143.000000	1178.000000	21.000000	60.000000	421.000000	5.120000

Basic report of the dataset

In [7]:

```
pp.ProfileReport(df, check_correlation = False)
```

Out[7]:

Overview

Dataset info

Number of variables	12
Number of observations	1143
Total Missing (%)	0.0%
Total size in memory	107.2 KiB
Average record size in memory	96.1 B

Variables types

Numeric	6
Categorical	3
Boolean	0
Date	0
Text (Unique)	0
Rejected	3
Unsupported	0

Warnings

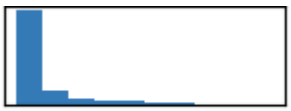
- `CPE` has 216 / 18.9% zeros Zeros
- `URL` has constant value URL-1 Rejected
- `clicks` has constant value yyy Rejected
- `engagement` has constant value xxx Rejected
- `spend` has 216 / 18.9% zeros Zeros

Variables

CPE

Numeric

Distinct count	207
Unique (%)	18.1%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	0.41081
Minimum	0
Maximum	5.12
Zeros (%)	18.9%

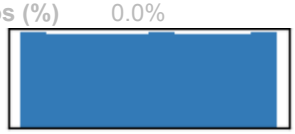


Toggle details

LineItemsID

Numeric

Distinct count	1143
Unique (%)	100.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	572
Minimum	1
Maximum	1143
Zeros (%)	0.0%



Toggle details

~~URL~~

Constant

This variable is constant and should be ignored for analysis

Constant value	URL-1

age

Categorical

Distinct count	4
Unique (%)	0.3%
Missing (%)	0.0%
Missing (n)	0

45-65+	426
25-34	259
18-24	248

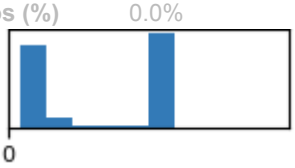
Toggle details

channel

Numeric

Distinct count	16
Unique (%)	1.4%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	6.3237
Minimum	1
Maximum	21
Zeros (%)	0.0%

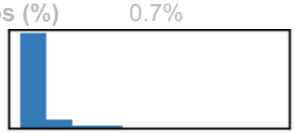


Toggle details

channel_ad_id

Numeric

Distinct count	32
Unique (%)	2.8%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	2.8556
Minimum	0
Maximum	60
Zeros (%)	0.7%



Toggle details

~~clicks~~

Constant

This variable is constant and should be ignored for analysis

Constant value	yyy

~~engagement~~

Constant

This variable is constant and should be ignored for analysis

Constant value	xxx

gender

Categorical

Distinct count	2
Unique (%)	0.2%
Missing (%)	0.0%
Missing (n)	0

M	592
F	551

Toggle details

interest

Categorical

Distinct count	40
Unique (%)	3.5%
Missing (%)	0.0%
Missing (n)	0

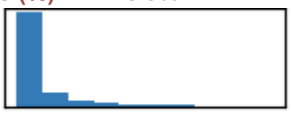
Interest - 16		140
Interest - 10		85
Interest - 29		77
Other values (37)	841	

Toggle details

spend

Numeric

Distinct count	183
Unique (%)	16.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	33.382
Minimum	0
Maximum	421
Zeros (%)	18.9%

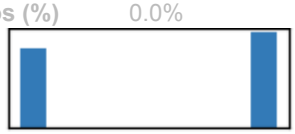


Toggle details

xyz_campaign_id

Numeric

Distinct count	3
Unique (%)	0.3%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	1067.4
Minimum	916
Maximum	1178
Zeros (%)	0.0%



Toggle details

Correlations

Sample

	LineItemsID	URL	xyz_campaign_id	channel	channel_ad_id	gender	age	
0	1	URL-1	916	1	2	M	45-65+	Int
1	2	URL-1	916	11	2	M	45-65+	Int
2	3	URL-1	916	11	1	M	45-65+	Int
3	4	URL-1	916	11	1	M	45-65+	Int
4	5	URL-1	916	1	1	M	45-65+	Int

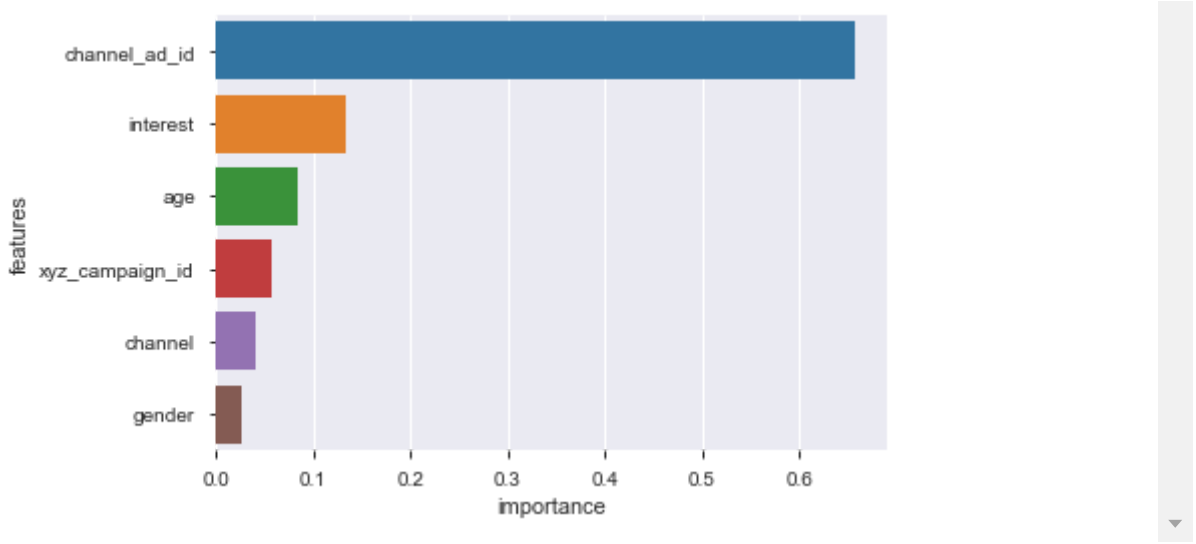
Machine Learning model - Random Forest Regressor

- Plot the feature importance related to the target variable (CPE)

In [8]:

```
## sklearn import
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestRegressor

import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline


## drop items not used
drop_items = ['LineItemsID','URL', 'spend', 'engagement', 'clicks' ]

df.drop(drop_items, axis = 1, inplace = True)

# LabelEncoder
le = LabelEncoder()

# apply "le.fit_transform"
df = df.apply(le.fit_transform)

target = 'CPE'
features = df.columns.tolist()
features.remove(target)

X_train = df[features]
y_train = df[target]

## This line instantiates the model.
fit_rf = RandomForestRegressor()

## Fit the model on your training data.
fit_rf.fit(X_train, y_train)

# Plot feature importance
feature_importances = pd.DataFrame({'features': X_train.columns,
                            'importance': fit_rf.feature_importances_}).sort_values('import

ax = sns.barplot(x="importance", y="features", data=feature_importances)
```

```
D:\Z_PRJ_FILES_DS2\Miniconda3\lib\site-packages\sklearn\ensemble\forest.py:2
45: FutureWarning: The default value of n_estimators will change from 10 in
version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```



Summary - initial analysis

- This Notebook provide one option to approach the problem and show an overview of the data and features / segmentation that impact CPE

The 3 most important features that impact CPE in this historical data are:

- channel_ad_id
- interest and
- age

More EDA (exploratory data analysis) will also be done in following notebook

In [ ]: