# BERKELEY
# UNIVERSITY OF CALIFORNIA


## Capstone Project 2023

**HANDE GABRALI-KNOBLOCH**

**OUTLINE**

## 1.Executive Summary

### Summary of Methodologies

1. Data Collection Methodology:
2. Data Wrangling:

- Data Cleaning
- Data Standardization, Feature Scaling

3. Exploratory Data Analysis (EDA) using Visualization

4. Predictive Analysis using Supervised & Unsupervised Learnings

5. Modul Evaluation

## 2. Introduction

### 2.1 Project Background and Context

With the effect of the Covid epidemic, consumers' staying at home for a long time has also changed their media usage habits.

While the use of digital media in developed countries is progressing very rapidly, it is not possible to talk about the same speed in developing countries.

The Coca-Cola Company is 137 years old in the world trade market with 200 brands worldwide. The innovations and changes it has brought to the market, it operates in approximately 200 countries globally.

The Coca-Cola shapes its investments in the countries where it is active, as in every brand today, where the use of digital media surpasses traditional media.

The social media and global video streaming services can potentially reshape consumers purchasing behaviors and media usage habits. Therefore, the media has an unavoidable function in conveying information and messages to the community.

There are two leading media: traditional media, printed newspapers, magazines, billboards, books, brochures, television, radio, and others. Second is social or cyber media such as newspapers, blogs, Twitter, Instagram, and others. Many factors influence consumer intention toward brands and their purchase intentions (Noor, at., 2023, p.15).

This capstone project also starts from the example of Turkey, which will be tried to reveal how the global brand Coca-Cola, particularly traditional media investments, are shaped in developing countries and their markets.

### 2.1.2 Problem Statement

In Turkey;

- TV advertising investments are with 78% increase in 2022 (fastest annual growth rate reached).

- Media investments in 2022, 24.03% in newspaper while the increase on the magazine side was 31.03%. took place.

- Outdoors closed 2022 with record growth: 124.63%.

- In 2022, radio investments exceeded 818 million TL. It has grown by nearly 100%.

- Cinema investments took place 110.7 million TL in 2022.

Considering the growth in the traditional media industry in Turkey, The Coca-Cola company decided to conduct research on the setups of the channels to determine which variables in the campaign had an impact on traditional media and greater impact on the performance and outcome of the campaign.

Thus, starting from the total media investments in Turkey in 2020, it is desired to evaluate the general investment tables of 2022 and get an idea about 2023 and beyond. As a result, the consistency of the analyzes made regularly by the brand will also be reviewed.

### 2.1.3. Common Problems that are needed to be solved:

- What affects the traditional media investments and communications of brands?
- Why should the Coca-Cola company continue to invest in mainstream media?
- Does the usage of TV's GRP positively affects the TV investment of brands?
- What do the traditional media investments costs depend on?
- What and how is it determined (whether or not) the first stage will be used again?
- Under what conditions can the traditional media investments rates be determined?

## 3. Methodology

### 3.1 Data Collection Methodology

Sources of dataset used within the scope of this capstone project are Kantar Media, RTÜK (Radio and Television Supreme Council), RD (Advertisers Association) Members, cinema companies operating in the industry, ARVAK members and in the Outdoor sector companies, and Press Organizations operating in Turkey. This data set, which is provided by the media agencies where The Coca-Cola company operates in the Turkish market, includes the TV, radio, OHH, cinema and press usages of the company and its competitors for 2020.

The main purpose of the data is designed to better understand the performance of each brand's campaigns in the market in 2020 relative to traditional media usages and budgets. Also, this also allows for a more accurate reflection of current mainstream media usage trends and brands investment behaviors.

Since the variables in the original dataset belong to all brands operating in the Turkish market, their number was higher than expected. Therefore, data cleaning was applied to ensure that only the appropriate variables in the NARD category ( Non-Alcoholic Ready to Drink) were used to answer the research questions.

Since the data were obtained from various platforms, special formats were applied to the datasets when they were first obtained. Therefore, data cleaning was necessary to reform errors and inconsistencies in the data, as well as to observe and eliminate all missing values and outliers in the data sets.

**Dataset:**

Based on the problem statement and questions, the following dataset may be used in this research:

**Traditional/ Mainstream advertising data from 2020:** This dataset would provide information on the performance of mainstream campaigns during the Corona pandemic. It would include metrics such as GRP rates (Gross Rating Point), day parts, frequencies, Column x cm (Press), item count, and costs.

Besides, this dataset would provide information on the performance of NARDT market campaigns during different months throughout the year.

**Demographic data:** The target audiences used by the Coca-Cola Company in its campaigns and purchases are not included in the dataset due to the confidentiality of purchasing.

Before all applications, raw dataset has 28.869 rows x 30 columns.

### 3.2. Data Wrangling

#### 3.2.1. Data Cleaning

**Removing unnecessary columns:**

At this stage, the columns that are not necessary for the study were deleted and the data was used more efficiently.

**Check for and remove duplicates:**

Duplicate data can skew the results of data analysis, and it is important to remove them before analysis. This step involves identifying and removing any duplicate data in the dataset. Duplicates can occur due to data entry errors or technical issues and need to be removed from the dataset to guarantee of our data accuracy.

**Handle missing values:**

Missing values can cause analysis problems when working on the data and therefore need to be handled adequately. This step regards identifying missing values in the dataset and then determining how best to handle them. This may include loading values, removing rows or columns with missing values, or leaving missing values intact.

**Check for and remove outliers:**

Outliers can have a significant impact on data analysis, and therefore it is necessary to identify and remove them before analysis. This step regards distinguishing outliers and determining the primary way to manage them, whether by removing them or adjusting them.

**Data Standardization, Feature Scaling:**

This section aims to standardize all data types in the dataset. This will facilitate the analysis of data. For example, ensuring that all medium types or dayparts are in the same format and numeric data are in the same units.

**Constraints**

Some of the limitations of this research are that certain data, such as conversion rate in a digital campaign, are not included in the data. For this reason, the effects of digital advertisements cannot be compared with traditional media.

There are many external factors that can affect the success of a mainstream media campaign, including changes in the competitive landscape, consumer behavior and broader economic trends. This may complicate the research and cause limitations and affect the accuracy of the findings.

Different research patterns can be created with the data obtained as a result of integrating and using together the uses in mainstream media and digital media tools.

## 4. Exploratory Data Analysis (EDA) using Visualization

Exploratory analysis is a data analysis approach used to provide first-hand understanding of data and identify potential patterns or trends. In the context of mainstream advertising campaigns, exploratory analysis can be used to identify variables that are likely to have an impact on campaign performance, such as channel type, vehicle type, total GRP, Cost or seconds. In the analysis here, bar, scatter, categorical LM plots and heat map are used. Exploratory analysis may involve various statistical techniques such as regression analysis or correlation analysis to determine the relationships between variables.

### Packages

In the analysis, Python programming language and Jupyter notebook and Pandas, Numpy, Seaborn, Matplotlib, Sklearn, Plotly, Mglearn are mainly used.

## 5. Predictive Analysis using Supervised & Unsupervised Learnings

### 5.1.Principal Component Analysis (PCA)

These methods give us some methods of roughly how we can operate between independent variables when we don't have a variable at hand. We have 22 components with main explained variance is 0.045.

Factor loading for the first component explains 0.19 of the variances.

After a sharp decline of explained variance from the second component, principal components stay at the average level until the nineth component. The correlation pattern between them may indicate a latent factor and dependency among multiple traits.

9 out of 22 components suffice to capture at least 90% of the original variance. While some variables do not contribute much to the variance in the data, more space is opened up to reduce dimensionality.

The top-3 highest and lowest factor loadings for the first principal component. Also, the lowest coefficients represent zero connection to the keywords Source Company, Report Brand and Report Product.

30” GRP_Total, Cost_TL and GRP_Total are dimensions that seem partially strong correlates (partially) according to other variables.

### 5.2.K-means Clustering

In K-Means clustering method, our aim is to try to cluster observations or variables by using similarity matrices and building various distance calculations.

It is wanted that the clusters formed are homogeneous within themselves but heterogeneous with respect to each other. Also, K-Means is a non-hierarchical clustering method.

Again, our goal here is to achieve high similarity within clusters and low similarity outside clusters. When we want to compartmentalize, K-means the most units of observation we have.

The catch that will be encountered in K-Means is; It is a good understanding of its iterative nature. After calculating the distances of the initially created centers, the first primitive clustering structure occurs. We then recalculate the distances to a center.

Clustering structure and distance measurement are done again. In each iteration, the sets that the elements enter may change. Another problem is; It's a question of choosing a starting point. Determining how many digits K will be is a separate problem. Because different error may occur when K=4 or K=6.

### 5.3. Multiple Linear Regression, Partial least squares regression (Partial-Least-Squares-Regression), Ridge and LASSO

#### 5.3.1. Multiple Linear Regression

Main Purpose: To find the linear function that expresses the relationship between dependent and independent variables.

We do this by trying to find the coefficient estimates that will minimize the sum of squares of error.

There are two Purposes:

1) Estimating the values of the dependent variable by means of the variables determined to affect the dependent variable.

(2) Which of the independent variables thought to affect the dependent variable, in which direction? to detect its impact. Trying to define the relationship between them.

Among its assumptions, especially “Multiple Linear Connection Problem” and “Autocorrelation Problem” are unpopular assumptions. When there is a Multicollinearity Problem, it means that the independent variables are very highly correlated with each other. This causes some problems. Methods such as PCR and LASSO have been suggested to overcome these problems.

In the multilinear regression model, there is no external parameter other than ß0. It is not actually an external parameter. Therefore, the Model Tuning process will be considered here as model validation.

The purpose of model validation; It was an effort to obtain more accurate errors with the obtained errors.

We have a problem that if you enter different values into random_state= -"We will select a particular part of the model, but which particular part you will select".

In order to eliminate this problem: Cross-Validation method is used.

### 5.3.2. **Partial least squares regression (Partial-Least-Squares-Regression)**

It is based on the idea of establishing a regression model by reducing the variables to a smaller number of components that do not have multicollinearity problems between them (Herman Wold, 1966-1982).

In the size reduction approach; It is being tried to express the information carried by P variables with less than ten variables. Here, too, it can be thought of as trying to move P variables with C components, which are less than ten.

We should keep in mind that there is a difference between PCR and PLS: PCR breaks the links with component reduction, but because the independent variables are processed entirely within itself, a reduction process takes place with no ability to explain the dependent variable. In PLS, on the other hand, it does dependent variable-oriented component rendering.

### 5.3.3. Ridge Regression

Its purpose is to find the coefficients that minimize the sum of squares error by applying a penalty to these coefficients (Hoerl & Kennard, 1970).

When lambda values change (each color represents a separate parameter). The trick of Ridge regression is to never set the lambda to zero. But keeping all the based numbers in the model and adjusting the effects of these based numbers on the model. For example, approaching zero. This is where Ridge and Lasso regression differentiated.

Sometimes these coefficients lose their importance according to the lambda value. We are able to adjust the relative effects of the variables -the effects on the estimation function- according to the lambda to be selected, but we still say let it stay in the model. We are saying that we should bring it closer to zero but not remove it from the complete model.

### 5.3.4. Least Absolute Shrinkage and Selection Operator (LASSO) Regression

Lasso regression achieves L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can perform in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g., Ridge regression) doesn't result in elimination of coefficients or sparse models. This builds the Lasso far easier to interpret than the Ridge**.**

## 6. Results

| Which model gives the best result for our dataset in general? | | | |
|---|---|---|---|
| | | | |
| **Model Names** | **Results** | **Cross_Val_Score** | |
| | | **Test** | **Train** |
| **Multiple Linear Regression** | | 7549.517791 | 9288.241897 |
| **PLS** | 10974.56 | | |
| **Ridge** | 9705.984 | | |
| **Lasso** | 9709.243 | | |

To determine how good is a model, let us understand the impact of wrong predictions that are “mean_squared_error(y_test, y_pred)”. MLR, Ridge and Lasso are the best predicted models and the wors one is the PLS.

## 7. Conclusion

In conclusion, data on TV, radio, press, cinema and OHH for 2020 were analyzed. The data was presented through various plots and charts to help understand the trends in frequencies, period of use, number of pages used for each medium types.

While New_Cost TL is expressing the mainstream media investments made by the Coca-cola company in 2020, Multi linear regression analysis was used to determine the relationship between continuous variables in the dataset.

X = df.drop('NEW Cost_TL', axis = 1)

y = df['NEW Cost_TL']

To identify the strength of the effect that the independent variables have on a dependent variable. So, determining the strength of relationship between dose and effects of New_Cost TL and other variables. Also, we tried to forecast effects or impacts of changes.

So cross_val_score.mean estimates the expected 76% accuracy of our model on out-of-training data. Based on the results, New_Cost TL in 2020 has statistically significant coefficients, indicating that ad cost has a positive effect on performance impressions for each medium types.

Alpha (α) is the penalty term that denotes the amount of shrinkage (or constraint) that will be implemented in the equation So, Ridge Regression alpha= 0.005, Lasso Regression alpha equals to 0.11079288755091725.

K-means tries to discover the least-squares partition of the data. PCA determines the least-squares cluster membership vector.

Elbow curve at k=6 and score= 1880837054500.593 so that, our optimal number of clusters for K-means clustering is 6.

## 8. Discussion

Brand awareness, brand image, brand attitude, prosumers, lead users and purchase intention are the key role in the brand communication. Also, customers need information about the brand to develop their awareness. The buying decisions of consumers always influenced by advertising. Regardless, traditional media applications are seen as more reliable than social media for consumers today. Especially considering the information clustering in the digital environment.

In order to increase brand awareness and awareness among consumers, brands should increase their frequency and increase their visibility with the periodic investments and campaigns they have made on the basis of media, as seen in their traditional media investments and communications.

For these reasons, especially in developing economies and their markets, by continuing to invest in mainstream media, the Coca-Cola company can build brand awareness and consumer information in a more reliable legacy media landscape.

It is seen that the use of TV's GRP has positive effects on TV investments of brands. It is noteworthy that The Coca-Cola company invests heavily in the channels measured in Turkey.

The cost of traditional media investments may depend on the time and unit price of the sport used in TV, at the beginning the cost of the line x column, and on the purchase and rental of boards in OHH through agreements. Here again, the importance of media buying and strategies comes to the fore.

In this study, it is not known to what extent it contributes to traditional media investments, since gender cannot be examined in the context of the target audience.

**References:**

**Articles:**

Noor., N, Puteh.,K, Nordin.,N, Amir., M, Amir., H, Sazali., F, Kamaluddin., M. ” Effectiveness of Traditional Media and Social Media Sustainability Communication in Influencing Green Consumption Intention”, Journal of Media and Information Warfare, Vol. 16 (1), 14-27, April 2023. Pp. 15-18.

Shao., C. (2023), “Changing Mass Media Consumption Patterns Before/After Relocation: East

Asian International Students’ Mass Media Use and Acculturation Strategies”, International Journal of Communication 17(2023), 1592–1612.

**Websites:**

ARVAK: **https://www.arvak.com.tr/Uyeler**

The Coca-Cola Company: Access Date: 07.05.23, https://www.coca-colacompany.com/

Turkish Government: Access Date: 07.05.23, https://www5.tbmm.gov.tr/kanunlar/k6487.html#:~:text=MADDE%202%2D%208%2F6%2F,ve%20t%C3%BCketicilere%20y%C3%B6nelik%20tan%C4%B1t%C4%B1m%C4%B1%20yap%C4%B1lamaz.

Word Investment Report 2020: Access Date: 07.05.23,

https://unctad.org/system/files/official-document/wir2020_overview_en.pdf

Turkish Advertising Association Reports: Access Date: 07.05.23,

https://rd.org.tr/assets/uploads/medya_yatirimlari_2019_.pdf

https://rd.org.tr/Assets/uploads/7587437b-563d-4917-b767-676021317bb1.pdf

Dentsu ad spend report: Access Date: 07.05.23,

https://www.dentsu.com/news-releases/dentsu-ad-spend-forecast-july-2022-release

Word Economic Forum: Access Date: 07.05.23,

https://www.weforum.org/agenda/2022/05/a-digital-silver-bullet-for-the-world/

https://www.insiderintelligence.com/forecasts/5d13a07a64fe7d034c2cc15a/5d139fb0b88aeb0b7c481d6c/

Statista: Digital Newspapers & Magazines – Worldwide: Access Date: 07.05.23,

https://www.statista.com/outlook/amo/media/newspapers-magazines/digital-newspapers-magazines/worldwide

Kantar: Access Date: 08.05.23

https://www.kantar.com/locations/turkey#_=

RTÜK: https://en.wikipedia.org/wiki/Radio_and_Television_Supreme_Council

Reklamcilar Dernegi (Advertisers Association): https://rd.org.tr/

https://rpubs.com/swal/1035300

https://www.linkedin.com/pulse/capstone-projects-data-science-machine-learning-full-alaa/

Data cleaning:

https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/

https://careerfoundry.com/en/blog/data-analytics/what-is-data-cleaning/

LASSO: Access Date: 15.05.23

https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#:~:text=Lasso%20regression%20is%20a%20regularization,i.e.%20models%20with%20fewer%20parameters).

https://www.statisticshowto.com/lasso-regression/