Savoldi Chiara 5014502

Summary

Study design & Code Book: First Dataset

As a row dataset I have chosen "**The Trending YouTube Video Statistics**", which is a daily record statistic for trending YouTube videos. Data includes video title, channel title, publish time, tags, views, likes, and dislikes, description, and comment count. I found this data sources on Kaggle, which is a place that we can use for searching data sets.

My idea is to clean the dataset to develop a future analysis to understand what are the requirements that a video must have to become viral.

I have chosen to analyse the trending YouTube videos for Great Britain. My dataset is composed by 38916 and 16 columns, which are:

- **Link of the video:** represents the usual link that we need on YouTube in order to visualize the video. It's a **character** column like a string in Python "Jw1Y-zhQURU";
- **Trending date:** it contains the date when the video went viral. The column contains **character** data;
- **Title of the video:** every video has a title. The column contains **character** data;
- **Name of the channel:** many people have their own YouTube channel where they can publish videos of different genres. The column contains **character** data;
- **Category of the video:** we can have music videos, funny videos. The column contains **integer** data;
- **Time and date of publishing:** it represents the hour and the date of publication of the video. The column contains **character** data;
- **Tags:** under each video you can post tags to make the video viral when you search for a certain word. The column contains **character** data;
- **Views:** contains the number of views of the video. It is an **integer** column as it contains numbers. For example, the video received 28700 views.
- **Likes:** number of likes of the video. The column contains **integer** data;
- **Dislikes:** number of unlikes. The column contains **integer** data;
- **Number of comments:** number of comments under the video. The column contains **integer** data;
- **Thumbnails:** A thumbnail is a reduced preview image of the original video that is used as a placeholder. The column contains **character** data.
- **Comments disabled:** Comments that have been disabled. The column contains **character** data, but only True or False values: yes the comment was disabled or no, the comment was not disabled;
- **Rating disabled:** the column contains **character** data, but only True or False values;
- **Error video or removed:** video removed from the web. The column contains **character** data, but only True or False values;

- **Video description:** description of the video content. The column contains **character** data.

I started analysing if there were any missing values for each column. Looking at the data I have seen that the missing values were defined as [none]. I have removed all the associated columns.

Outliers in data can distort predictions and affect the accuracy if you do not detect and handle them appropriately. I have found some outliers. I believe the elimination of the outliers depends on the type of analysis I am interested to do. Since I am looking at the most viral videos, it makes sense to eliminate those that have fewer than average views.
In addition to histograms, boxplots are also useful to detect potential outliers. Observations considered as potential outliers are displayed as points in the boxplot: I have built several graphs for visualization.

I considered first the 'Trendind date' column because the dates were indicated in the form "17.14.11". I wanted to adjust the data to have "17-11-14", since the date was in the format YY-DD-MM and we usually consider normal long dates DD-MM-YY or inverted long dates YY-MM-DD. After that, I wanted to manage the 'Publish Time' column to separate the date and time of publication of the various videos. This because for future analysis, it could be interesting to understand what day of the week and at what time, it was preferable to publish a video to increase the number of views. So, I created four new columns:

- **Trending day: character** column that contains the day the video was published ;
- **Trending month: character** column that contains the month the video was published ;
- **Trending year:** character column that contains the year the video was published ;
- **Video time publishing**: Contains hours, minutes, seconds referring to the time of the video's publication. It is a **character** column;
- **Trending wdays:** I have also taken the observations where the year is the 17 and I associate to every day the correct days of weeks, in order to see in which day of the week the videos are most viral. It is a **character** column;
- **Trending Date:** the column becomes in **POSIXc** format. POSIXct stores both a date and time with an associated time zone. The default time zone selected, is the time zone that your computer is set to which is most often your local time zone.

The column ‘Video removed, ‘Rating Disabled’, ‘Column Disabled’ contain only True or False values: I have excluded the values True because if a video has been removed it might not interest me, in particular for my analysis.

Each character column was treated for different reasons:

- Remove the stopwords: in fact, most of the stop words have no particular meaning if isolated from the text and therefore are ignored by the programs;
- Remove punctuation and switch all the character to lower case;
- Removing pronouns, letters with accents ...;
- Removing special characters like Ã¯;
- Removing extra spaces;
- Remove columns left without text;

I kept the Pipe | character typical of tags for the Tags column.

For example, the string “â–º HOW MY RELATIONSHIP STARTED!\nâ–º PB Merch â€¢ ht”, has become “how my relationship started ht”.

In the final part I wanted to better rename the columns of the dataset so that they had a more speaking name.

Study design & Code Book: Second Dataset

The dataset contains files from a bank whose management wants to explore ways to convert their passive customers into personal loan customers (while still keeping them as depositors). Last year, the bank launched a campaign for liability customers that showed a conversion rate of over 9% success. This encouraged the retail marketing department to design campaigns with better-targeted marketing to increase the success rate with a minimal budget.

The Bank.xls includes 5000 rows and 14 columns.

The data includes the customer's demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer's response to the latest personal loan campaign (Personal Loan). In particular:

- **ID client**;
- **Customer's age**: it is an integer column;
- **Number of years of professional experience:** integer column;
- **Annual income** of the customer (dollars): integer column;
- **Home Address ZIP code**: integer column;
- **Family size** of the customer. It can have 4 values: family of type 1,2,3 or 4;
- **Average spending on credit cards** per month (dollars): integer column;
- **Education Level**. It can have 3 values:
  1: Undergrad;
  2: Graduate;
  3: Advanced / Professional;
- Value of house **mortgage** (if any/dollars);
- Did this customer accept the **personal loan** offered in the last campaign? it has only two values: 0 (no) & 1 (yes);
- **Securities_account:** Does the customer have a securities account with the bank? It is a binary variable;
- **CD_Account:** Does the customer have a certificate of deposit (A certificate of deposit (CD) is a savings account that holds a fixed amount of money for a fixed period of time, such as six months, one year, or five years, and in exchange, the issuing bank pays interest.) account with the bank? It is a binary variable;
- **Online:** Does the customer use internet banking facilities? It is a binary variable;
- **CreditCard:** Does the customer use a credit card issued by this Bank? It is a binary variable;

In US, the first digit of a PIN indicates the zone or a region, the second indicates the sub-zone, and the third, combined with the first two, indicates the sorting district within that zone. The final three digits are assigned to individual post offices within the sorting district.

My idea is to consider only the first two digits of the zip code, in order to reduce the possibilities to 7 groups.

I have dropped to ID column since it is not relevant for the analysis.

The 'Experience' column refers to the number of years of work experience. There are 52 rows in the dataset that contain negative values. It is not possible for years to have a value smaller than zero, so I have deleted those columns from my dataset.

I continued the analysis by looking if there were any missing values for each column: there are no missing values in the dataset.

The three variables CCAvg, Mortgage and income present some outliers. To treat the presence of outliers I have used the "interquartile range" (IQR), which represents the width of the box in the boxplot, that is IQR = Q3 – Q1 . The IQR is used as a measure of how spread-out the values are.

The IQR tells how some of the other values are "too far" from the central value. These outliers are outside the range in which we expect them.

If a data point is below Q1 – 1.5×IQR or above Q3 + 1.5×IQR, it is viewed as being too far from the central values to be reasonable.

I have also renamed the 0 to 'No' and the 1 to 'Yes' for easier interpretation of the various graphs.

It is interesting to plot on the x-axis each variable from the dataset, because for example, for 'Education' there are 3 levels, so you can see how many people, in the 3 levels, have said Yes to the campaign or No. It is interesting to understand how much the parameters are related to my variable y.