DATA ANALYTICS CAPSTONE PROJECT

Boolean Course in Data Analytics #02
November 2023

You have been divided into the following **three groups**:

1. Anuoluwapo Ajani Adeniran, Jane Yu, Tiago Santos, Marketa Chalupova, Zeinab Toghani
2. Gonçalo Bernardo, Mahshid Pashootan, Matthew Ray, Valéria Onofrei
3. Elisabete Domingues, Enrico Freddo, Mohammed Abul Umayer, Morteza Abbaszadeh, Niyaz Rowshan

Each group should **choose between one of the following three tracks** (different groups may choose the same track) and, together, you will:

- study the problem at hand,
- devise a strategy to tackle it,
- Query/load, clean, analyse and visualise your data
- present your results.

On the **2nd of December 2023 between 9:30 and 11:30**, each group will present their work making sure to:

- properly introduce the problem setting,
- explain which tool / software were used (eg: SQL for data extraction, Python for analysis, Tableau for storytelling / visualisation, etc)
- show the data and insights you found along the way that support your case,
- clearly state your conclusions and actions that need to be taken.

Each group will have between 30 - 40 minutes to present their work, that is about 7 minutes per person.

Each person in the group needs to be involved in the background work as well as in the presentation of the results (that is, everyone should present a part of the project).

Summary of the proposed tracks (see the next pages for a detailed description):

1. E-commerce business / marketing setting
2. Product feature AB testing
3. Freelance open track

Track 1: e-commerce business / marketing setting

You're a Data Analyst at an e-commerce company and the business requires a comprehensive analysis about the progress of their activity. In particular, they require a study about the progress of sales, inventory and products.

The company data is stored in the following BigQuery database: **bigquery-public-data.thelook_ecommerce**. Your task is to query and extract the data, perform any cleaning if necessary, manipulate it, analyse it and create a compelling presentation of your results in Tableau, that you'll deliver to the business.

Your final output should provide insights about the following topics:

- **Website activity**: the **events** table contains session-level data about the user's online behaviour.
  Some examples: which parts of the website are the most visited? where is the traffic coming from? …
- **Demographic composition**: the **users** table includes customer information.
  Some examples: age / gender of your customers, popular countries / cities, customers that are particularly loyal (multiple purchases), …
- **Product performance**: the **product** table describes the product catalogue as well as costs and retail prices of each element, but it's in the **order-item** table that you will be able to see how many times a certain product was sold.
  Some examples: best/worst sellers, which products bring the highest/lowest revenue, gender/age preferences for certain products / categories, …
- **Inventory status**: the **inventory_items** sheet shows stock availability and the **distribution_centres** table shows where each distribution centre is located.
  Some examples: are there availability issues for any product / category? is there a distribution centre that is problematic? …
- **Other insights**: don't forget about the **orders** table, which provides order-level information on your customer's transactions.

Finally, as explained above, you will use the insights gained from your analysis above to prepare a presentation in the form of a dashboard, story or PowerPoint/Google Slides in which you'll explain the context, the insights you found, their interpretation and any recommendations / next steps, explaining how they can be leveraged to improve the future of the business.

As an added bonus, implement one of the following predictive analysis and try to fit them into your narrative:

- What-if analysis (eg: profit margin)
- Forecast (eg: sales)
- Linear regression (eg: predict age or gender given items purchased)

Note: for this project, you should use all three of: SQL (grouping, joining data, etc), Python (cleaning, exploring, analysing) and Tableau (visualisation and presentation).

Track 2: product feature AB testing

You're a Product Data Analyst, working for an online store. You just came out from a meeting where it emerged that the overall conversion rate is 2.8%, which is not bad, but it could be improved by reducing attrition (that is, reducing drop-offs) in the checkout funnel.

Your **first objective** is to identify a problematic step (in terms of customers' drop-off) in the conversion funnel; you'll analyse the data via Google Analytics using the Google Merchandise Store account. The product manager and the design team will then propose a redesign of a feature in that web-page (some examples are: adding the shipping costs up-front or redesigning the billing details prompt or adding a new product bundle feature, etc; come up with a plausible option).

Once the redesign is completed and the new feature has been implemented by the web development team and is ready to go live into production, your **second task** will be to design an experiment (an AB test) that will allow you to measure statistically (via hypothesis testing) whether the new feature is actually improving the funnel conversion rate or not. For this task, you will have to:

1. *describe the experiment*;
2. *decide the minimum improvement you hope to achieve* with the new feature implementation (that is: the conversion rate should increase to at least… x%);
3. *calculate the minimum sample size*: number of days / observations required to collect the data for the sample to be significant / representative;
4. *generate the data*: in a real scenario you'd be collecting the data, but in this exercise, you will have to simulate two series (control and test) of "conversion data" (that is: series of 1s=converted and 0s=did not convert);
5. use a test of hypothesis to figure out whether the new feature has significantly increased the checkout conversion rate or not.

The two series that you will simulate at step 4. will have a binomial distribution with an appropriate size parameter “n” (that you’ll have calculated at step 3.) and probability parameter “p” of:

- control series: p=0.028
- test series: p=0.035

Finally, your **third task** will be to present your results to your boss (the Head of Product), offering evidence for your findings and recommendations on how to proceed with next steps and actions (whether your test results were positive or not). You need to convince him to take action and follow your recommendations.

Note: for the first objective you should use Google Analytics and (optionally) Google Sheets, for the second objective you should be using Python and for the third objective you should be using Tableau or Google Data Studio.

Track 3: freelance open track

You’re a freelance Data Analyst / Data Journalist and you’ve been commissioned for a project by a company / newspaper.

This is an **open track**, which has the advantage of not having instructions to follow, but it also bears the challenge of having to structure the analysis on your own.

Below is an ideal skeleton of what your project outline may look like:

1. **Set the stage**: pick an area of interest and find an appropriate business problem / news topic that a company / newspaper is likely to commission to a freelance analyst.

2. **Find the data**: search for one or more datasets that can be used to initialise your analysis and solve your problem (as you may have realised, you have complete freedom in the choice of topic and data). Here are some examples of how you may retrieve data:
    a. Excel / csv / json files downloaded from the internet
    b. Data queried from a database
    c. Data retrieved via API or scraped from the web
    d. Synthetic data (extracted/simulated)

3. **The core of the project**: begin your analysis and problem solving. Again, you have total freedom to implement the tools and techniques that are most relevant to your project, however, be sure to include at least two of the following topics that we have covered in class:
    a. use of APIs or web scraping

b. data cleaning / data manipulation
c. joining / unioning multiple datasets
d. exploratory data analysis
e. statistics / predictions / optimisation (solver)

Also, be sure to use Python for your analysis as well as any of the following tools that we have studied in class if necessary:

a. Google Sheets
b. Google Analytics
c. Google Big Query / SQL

4. **Tell your story**: finally, devise a communication strategy to be sure that your message / analysis / main findings are properly received by your client / audience (which you'll have to define). You can decide the method of delivery, however any chart and dashboard should be created in either:

a. Google Data Studio
b. Tableau
c. Python