Welcome to DigData online Step Up career challenge

This document will help you understand the challenge and give you all the information you need to know.

What resources do you need :

- Dataset (excel file)
- How to guide (this document)
- Presentation (slides we went through in the briefing)
- Prompting material (Python script)

Step Up Career Challenge - Brief

How do ITV promote content on their new streaming platform ITVX effectively to their viewers?

Overview:

ITV have just launched their new streaming platform ITVX and have commissioned new shows by partnering with third parties such as Warner brothers to expand the breadth of ITVX content and the number of hours available to watch (from 3,000 to 15,000!).

Step Up Challenge:

With an average of 30 million registered users, how can we ensure that they discover this new great content, increase the number of hours they stay on the platform and are aware of the new features that ITVX has to offer?

Task :

ITVX is launching a new original show every week, ‘A Spy Among Friends’ is one of the first original shows on ITVX in 2022

- **Task 1 - Data Strategy**
    - **We want to know what the customer base viewing habits look like?**
- **Task 2 - Data Science**
    - **Who should we be promoting this show to?**

- **Based on what you have learnt, what recommendations would you give to product and /or marketing to promote ‘Spy Amongst Friends’.**

Once you’ve come up with your analysis and ideas it will be time to present your results. Imagine your audience is a business leader who may not have the same background as you, so make sure to describe your recommendations in a clear and compelling way! You could do this in many ways, for example by using:

- Infographics and pictures
- Graphs and charts
- Code, text or bullet points

Data Strategy : The DATA

56k Rows, 22 columns

The ‘watched_flag’ column is our target, showing whether a customer watched ‘Spy Amongst Friends’. Watched = 1, Did not watch = 0.

Disclaimer : This is not real ITV data but a representation of the types of data we do use everyday

Data Field	**Data Explanation**
user_id	All data grouped at a distinct user ID level, user_ID unique to a specific customer
stream_id	A unique ID relevant to a customer and viewing session
platform	The platform in which a customer viewed content on, i.e. Desktop, Browser
session_duration_seconds	Length of the session in seconds
is_weekend	Viewing on saturday or sunday = TRUE, weekday viewing = FALSE
session_start_datetime	The time the customer first starts browsing on ITVX
session_end_datetime	The time the customer exists ITVX since starting to browse or view content
stream_type	Viewing is categorised into 2 streams, VOD = Video on demand, Simulcast = Live viewing, i.e. ITV1, ITV2
consumption_seconds	Any amount of viewing on ITVX in seconds

programme_id	Unique ID per programme title
episode_id	Unique ID per episode per programme
programme_title	The title of the programme
series_title	Title of the series, i.e. Love Island is the Programme, Series 1 is the series title.
series_number	Number of the series
episode_title	Title of the episode
episode_number	Episode number
episode_production_year	Year the episode was made
genre	Genre of the programme, i.e. Drama, Comedy
sub_genres	Sub Genre of the programme, i.e. Crime and Thriller
schedule_channel	Channel the programme initially aired on linear TV, i.e. ITV1, ITV2
schedule_programme_slot_duration	The length of the slow the programme was aired on TV

Data Science : The DATA

1763 Rows, 31 columns

The *'any_spy_among_friends_consumption'* column is our target, showing whether a customer watched 'A Spy Among Friends'. Watched = TRUE, Did not watch = FALSE.

Disclaimer : This is not real ITV data but a representation of the types of data we do use everyday

Data Field	**Data Explanation**
user_id	All data grouped at a distinct user ID level, user_ID unique to a specific customer
n_sessions	Total number of distinct sessions
n_programmes_watched	Distinct number of programmes a customer has watched any length of content on

n_episodes_watched	Distinct number of episodes a customer has watched any length of content on
top_3_programmes	Ordered by consumption, top 3 programmes a customer has viewed
top_3_genres	Ordered by consumption, top 3 genres a customer has viewed
total_genre_comedy_consumption_seconds	Total consumption in seconds viewed in the Comedy genre
total_genre_drama_consumption_seconds	Total consumption in seconds viewed in the Drama genre
total_genre_entertainment_consumption_seconds	Total consumption in seconds viewed in the Entertainment genre
total_genre_sport_consumption_seconds	Total consumption in seconds viewed in the Sport genre
total_genre_other_consumption_seconds	Total consumption in seconds viewed in the Other genre
total_genre_factual_consumption_seconds	Total consumption in seconds viewed in the Factual genre
total_channel_ITV_consumption_seconds	Total simulcast consumption in seconds viewed on channel ITV
total_channel_ITV2_consumption_seconds	Total simulcast consumption in seconds viewed on channel ITV2
total_channel_ITV3_consumption_seconds	Total simulcast consumption in seconds viewed on channel ITV3
total_channel_ITVBe_consumption_seconds	Total simulcast consumption in seconds viewed on channel ITVBe
total_channel_ITV4_consumption_seconds	Total simulcast consumption in seconds viewed on channel ITV4
total_watch_morning_consumption_seconds	Total consumption in seconds viewed between 4am and < 11am
total_watch_afternoon_consumption_seconds	Total consumption in seconds viewed between 11am and < 4pm
total_watch_dinner_consumption_seconds	Total consumption in seconds viewed between 4pm and < 9pm
total_watch_night_consumption_seconds	Total consumption in seconds viewed between 9pm and < 4am

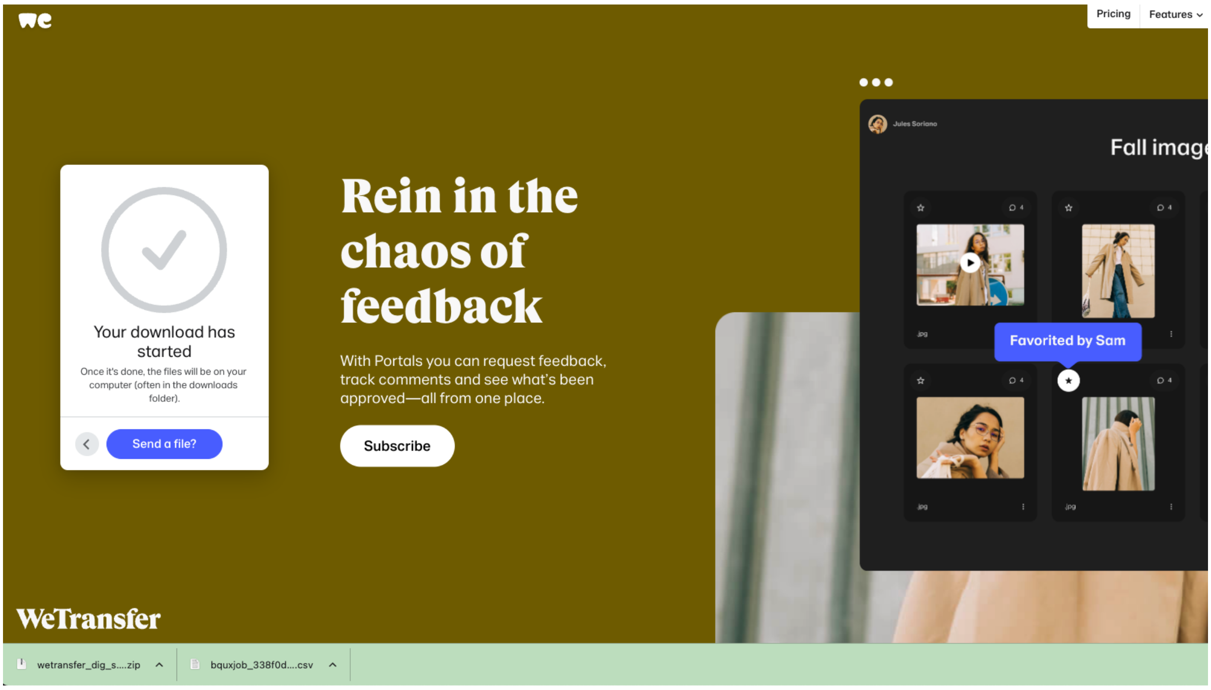
total_platform_connected_tv_consumption_seconds	Total consumption in seconds on connected TV
total_platform_mobile_consumption_seconds	Total consumption in seconds on mobile
total_platform_desktop_consumption_seconds	Total consumption in seconds on desktop
n_devices_watched_on	Number of distinct devices used to view content: this can be different TV's, mobiles, or desktop. The value can be over 3 if a user has e.g. watched on two different mobiles.
total_weekend_consumption_seconds	Total consumption in seconds on saturday and sunday only
total_consumption_seconds	Total consumption in seconds
any_spy_among_friends_consumption	TRUE if that customers has watched any amount of spy among friends, FALSE if they have never watched it
top_3_subgenres	Ordered by consumption, top 3 sub genres a customer has viewed
age	Age of the customer
gender	Gender of the customer

GETTING SET UP :

If you've never worked with Jupyter Notebooks before, here is the quickest way to get started:

1. **Go to https://jupyter.org/try-jupyter/lab/**
2. **Upload the files in called DIG_Step_Up_Data_Science**
3. **Open the python notebook**
4. **Read through some of the hints and tips**
5. **You're all set!**

Your downloaded files screen should look like this (see files at the bottom) :



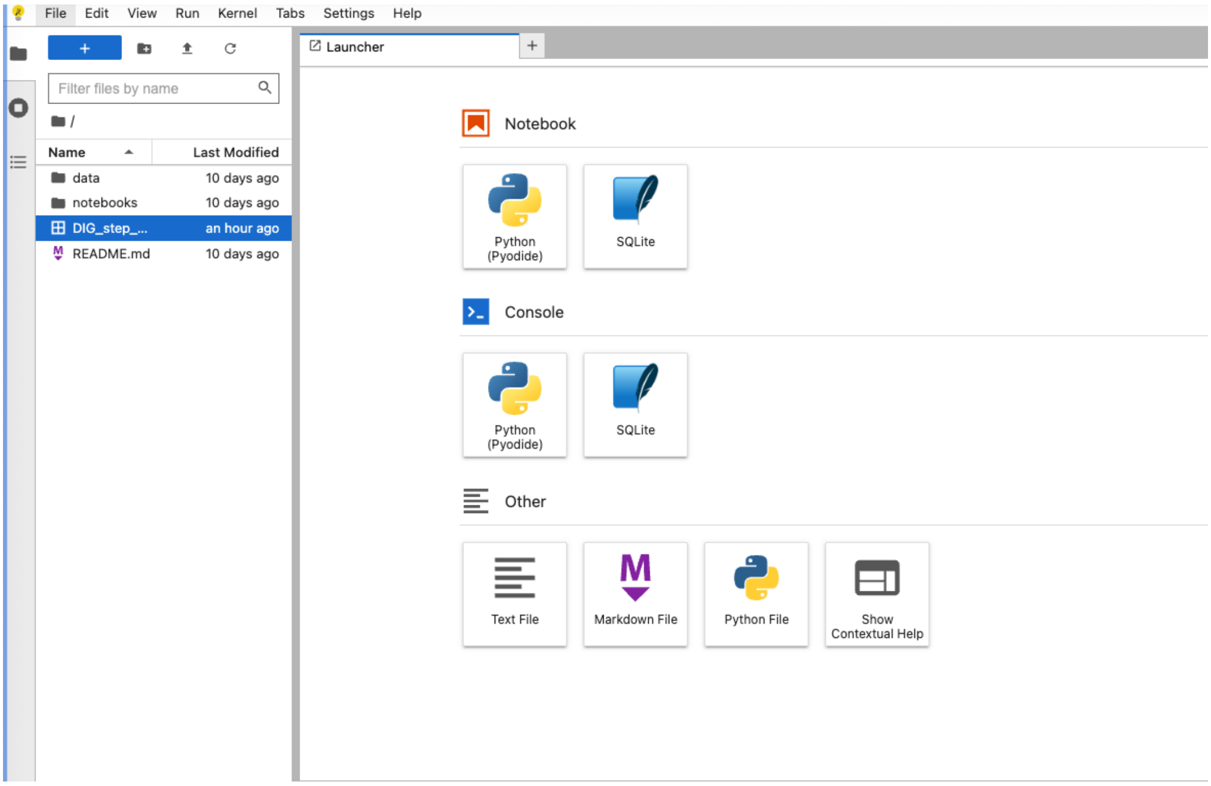
Your download has started

Once it's done, the files will be on your computer (often in the downloads folder).

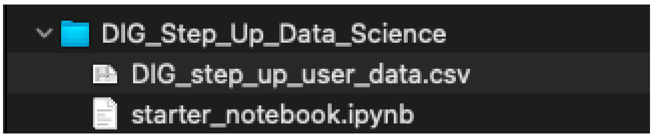
Rein in the chaos of feedback

With Portals you can request feedback, track comments and see what's been approved—all from one place.

This is the jupyter interface, click on the small arrow button highlighted to upload the python file and excel.

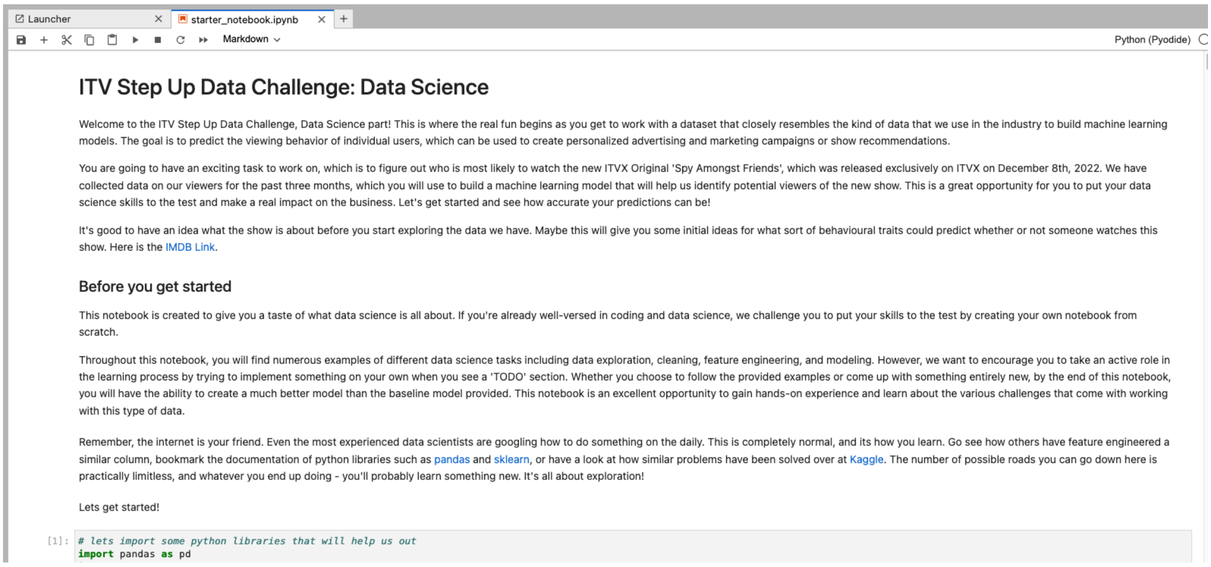


You will need to upload the 2 files in the folder below:



DIG_Step_Up_Data_Science
DIG_step_up_user_data.csv
starter_notebook.ipynb

Your Jupyter notebook should look like the screenshot below - this can be your starting point for the data science task.



ITV Step Up Data Challenge: Data Science

Welcome to the ITV Step Up Data Challenge, Data Science part! This is where the real fun begins as you get to work with a dataset that closely resembles the kind of data that we use in the industry to build machine learning models. The goal is to predict the viewing behavior of individual users, which can be used to create personalized advertising and marketing campaigns or show recommendations.

You are going to have an exciting task to work on, which is to figure out who is most likely to watch the new ITVX Original 'Spy Amongst Friends', which was released exclusively on ITVX on December 8th, 2022. We have collected data on our viewers for the past three months, which you will use to build a machine learning model that will help us identify potential viewers of the new show. This is a great opportunity for you to put your data science skills to the test and make a real impact on the business. Let's get started and see how accurate your predictions can be!

It's good to have an idea what the show is about before you start exploring the data we have. Maybe this will give you some initial ideas for what sort of behavioural traits could predict whether or not someone watches this show. Here is the IMDB Link.

Before you get started

This notebook is created to give you a taste of what data science is all about. If you're already well-versed in coding and data science, we challenge you to put your skills to the test by creating your own notebook from scratch.

Throughout this notebook, you will find numerous examples of different data science tasks including data exploration, cleaning, feature engineering, and modeling. However, we want to encourage you to take an active role in the learning process by trying to implement something on your own when you see a 'TODO' section. Whether you choose to follow the provided examples or come up with something entirely new, by the end of this notebook, you will have the ability to create a much better model than the baseline model provided. This notebook is an excellent opportunity to gain hands-on experience and learn about the various challenges that come with working with this type of data.

Remember, the internet is your friend. Even the most experienced data scientists are googling how to do something on the daily. This is completely normal, and its how you learn. Go see how others have feature engineered a similar column, bookmark the documentation of python libraries such as pandas and sklearn, or have a look at how similar problems have been solved over at Kaggle. The number of possible roads you can go down here is practically limitless, and whatever you end up doing - you'll probably learn something new. It's all about exploration!

Lets get started!

```
[1]: # lets import some python libraries that will help us out
     import pandas as pd
```

Our data science team suggest running 3 models to predict the likelihood of our remaining customer base that might watch the new original show ‘Spy Among Friends’

Now it's time to take it to the next level and showcase your expertise. The business is going to have some follow-up questions for you, so be prepared to provide insightful answers. Here are some key questions to think about:

- Out of the three models, which one performed the best and why?
- Which features played the most significant role in determining the outcome?
- Are there any additional data points that could be collected to enhance the model's performance?
- Can you explain, in your own words, how each of these models work and their underlying mechanisms? Remember to do some research first to deepen your understanding.

Write your answer either in a separate document or within the notebook. Be creative and use whatever format you'd like - for example, it might be easier for you to stick to making graphs in the notebook and then exporting them to a powerpoint to add comments.

Hint: For a good stakeholder presentation - it is important to include visualisations and key metrics.