# DataHack 1.0 Problem Statements

## PS1 - AI/ML (Course5i)

### Abstract

In today's world, social media has become a vital part of our lives, and YouTube is one of the most popular platforms for video content creators. With over 2 billion monthly active users, YouTube has become a go-to platform for content creators to showcase their talent and reach a larger audience. However, with such a vast amount of content, it becomes challenging to stand out from the crowd and get the required views and subscribers. To overcome these challenges, YouTubers need to optimize their videos to get maximum engagement, views, and subscribers.

### Problem Statement

The challenge for this Problem Statement is for you to explore various aspects of YouTube optimization, including content quality, relevance, watch time, consistency, SEO, engagement, collaborations, analytics, demographics, and social listening. Use your data analysis skills to analyze the data of 50 Indian YouTubers given to you to determine the reasons behind the increasing growth of certain YouTubers among them and also find the solutions to some analytical questions given to you regarding the given data.

### Dataset

[Youtube\_Channels\_Dataset](https://drive.google.com/file/d/1VifMkZ3Fhh_C1JW9IIa1RPguoJXK9tyG/view?usp=share_link)

[YouTube\_Videos\_Dataset](https://drive.google.com/file/d/1uKs6CKNRLcP3KesAl1MM1_hhTdVwPWYU/view?usp=share_link)

### Schema

**Channel Dataset**

1. Channel ID: Primary Key to Identify the Channel
2. Channel Title: Name of the channel as visible on YouTube
3. Description: Description of the channel as mentioned in the about section of the channel
4. View Count: Total views of the channel
5. Subscriber Count: Total subscribers of the channel
6. Video Count: The number of videos on the channel that are visible to the public
7. Created Time: The time when the channel was created
8. Uploads: ID of a playlist where all the uploads of the channel are present
9. Other Links: Links the other social media handles that the creator might have listed in the about section of the channel

**Video Dataset**

1. Video ID: Primary Key to identify a video
2. Title: The name of the video
3. as visible on YouTube
4. Description: The description of the vides as provided by the creator
5. Category ID: ID of the category the video is from
6. Tags: The tags the creator has used for the video
7. Duration: The length of the video
8. View Count: The number of views the video has received
9. Likes: The number of likes the video has received
10. Comments: The number of comments on the video
11. Upload Time: The time when the video was uploaded to YouTube
12. Channel Title: The name of the channel from where the video is
13. Thumbnail: URL of the thumbnail of the video
14. Topic Categories: A list of wikipedia links showing the category of the video. This is also given by YouTube itself
15. Comments: A list of 100 comments on the video

### Evaluation

1. In what other aspects can a Youtuber drive traffic to grow his channel apart from key tags and descriptive titles, as his/her channel won’t show up in the top videos after a search?
2. Who are the target audience on a video in each category and what should be the ideal upload time for that YouTuber? Create an upload schedule for a YouTuber according to his content category.
3. On the basis of this dataset, mention the top 10 upcoming YouTubers besides the current top 10 YouTubers.
4. What are the most discussed words in YouTube under each category? In what topics should a YouTuber do research in his category to bring about a new trend in his category?
5. What changes can a Youtuber make in his MetaData (title, tags, description, captions) so that his videos show up in the top searches of Youtube and the channel gets more traffic?

**Opportunities :**

**Internship Opportunity At Course5i**

## PS2 - Data Science (Jetson)

### Abstract

The hospitality industry is highly competitive, and companies are always looking for ways to differentiate themselves and provide a better customer experience. One way to achieve this is by leveraging data science to gain insights into customer behavior, preferences, and needs. The challenge for this hackathon is to develop a data-driven solution that can help improve the customer experience in the hospitality industry.

### Problem Statement

**Market Basket Analysis**

In the restaurant industry, market basket analysis can be used to understand which dishes and menu items are commonly ordered together, which can be used to improve menu offerings and optimize sales.

For example, if market basket analysis reveals that customers who order pizza are likely to also order garlic bread, a restaurant could offer a discounted bundle of pizza and garlic bread to encourage customers to order both items together. Additionally, the restaurant could consider adding new items to the menu that pair well with popular items to increase sales.

Overall, market basket analysis can be a powerful tool for restaurants looking to optimize their menu offerings and improve sales by identifying which items are commonly ordered together.

**ChatGPT based customer engagement**

how to leverage technology to improve customer engagement, foster brand loyalty, and generate repeat business in an increasingly competitive and digital landscape.

We can create personalized email campaigns by analyzing the past order history of customers. For instance, if a customer has not visited the restaurant in the last 30 days, we can send them an email with a 15% discount offer to incentivize their return. Similarly, we can inform all our vegan customers about the newly added vegan dish to our menu to cater to their specific preferences.

Students are free to use any other data science technique. We are looking for solutions which are not straight forward data analysis showing the reports/dashboard on existing facts.

### Dataset

[Sample Data](https://jetson-datathon-sample-data.s3.us-east-2.amazonaws.com/jetson-sample-data.csv)

### Schema

| Field Name | Data Type | Description |
| --- | --- | --- |
| client\_id | String | Unique id to identify client. |
| order\_id | String | Unique id for Order. One Order can have multiple ordered items, in that case the same Order\_id is used for multiple different item\_name. |
| item\_name | String | Name of the ordered item |
| price | Double | Cost of the item. |
| date | Date | Order Date |

### Evaluation

* Our evaluation process will consider both the data science solution and the presentation.
* We will place particular emphasis on the approach used by students in developing the solution.
* As part of their data science solution, students are encouraged to utilize Kaggle to analyze data and share their Kaggle notebook. Kaggle provides an excellent platform for students to gain experience in data analysis and explore real-world datasets.
* For ChatGPT based solutions students are free to use whatever technology they want to use. Then can host the solution as a web application.
* For the final outcome, students should create a PowerPoint presentation that provides a clear and concise overview of their data science solution. This presentation should include details about the approach used, the data sources and analysis methods, and the key findings and insights.
* We are looking for presentations that effectively communicate the results of the data analysis and provide actionable insights for the hospitality industry. The final presentation should be well-structured, visually appealing, and engaging to the audience.

**Opportunities :**

**Internship Opportunities at Jetson**

## PS3 - Data Analytics (TechGig , LilacBud)

### Abstract

The used car market is a significant and growing industry, with millions of used cars being sold every year. One of the biggest challenges for buyers and sellers is determining the fair market value of a used car, which can vary widely depending on a variety of factors, including make, model, year, mileage, and condition.

### Problem Statement

This datathon aims to leverage machine learning techniques to develop a predictive model that can accurately predict the price of used cars. Participants are provided with a large dataset that includes information about thousands of used cars, including make, model, year, mileage, condition, and price. The goal of the datathon is to develop a predictive model that can accurately predict the price of a used car based on these and other factors.

### Dataset

[train\_data.csv](https://drive.google.com/file/d/1MXEcWQ986cg6elbRNKcoiBmDoq8hxNR_/view?usp=sharing)

[sample.csv](https://drive.google.com/file/d/1d1pH95KqYNE7M9rgCZerK892o3O745td/view?usp=sharing)

[test\_data.csv](https://drive.google.com/file/d/1IY_Pky06N1a8TeVkfHFsgFPjS7D_B3my/view?usp=sharing)

[Image Dataset](https://data.mendeley.com/datasets/hj3vvx5946/1)

### Schema

Name - String

Location - String

Year - Numeric

Kilometers\_Driven - Numeric

Fuel\_Type - String

Transmission - String

Owner\_Type - String

Mileage - Numeric with unit

Engine - Numeric with unit

Power - Numeric with unit

Seats - Numeric

New\_Price - Numeric with unit

Price - Numeric

### Evaluation

Based on the given problem statement, the following are the tasks involved:

* Develop a model using the features in the given dataset to predict the price of a car.
* Use the image dataset to extract metrics that specify various features of the car such as scratches, dents, number of seats, Type of Car (Sedan, SUV, etc, and color.
* Use the extracted metrics from car images as additional features to support the final predicted price.
* Use web scraping to scrape data from various online sources such as car marketplaces, auctions, and classified sites to enhance the fair estimation of the market sale value of a car, specifically based on its model, color, and other relevant features.
* Analyze the importance of individual features that contribute to the final predicted price and provide appropriate visualizations to support the findings
* What changes would you make to these datasets so as to make an analyst’s life easier?
* Additionally, the model could be used to develop pricing strategies for used car dealerships or inform policy decisions related to the used car market.

**Opportunities :**

**Student Ambassador at TechGig**

**Internship Opportunities at LilacBud**

## PS4 - DevOps (Streamlit , Snowflake)

### Abstract

As a participant in this open innovation datathon, your task is to perform an analytical case study and create a data analysis project using Streamlit and connect it to a Snowflake database. Your project should be based on a dataset from the finance domain, and you should explore and analyze the data to gain insights and make recommendations.

### Problem Statements

Some suggested topics include analyzing stock market trends, exploring financial statements of a company, or examining patterns in credit card transactions. You may also choose to focus on one of the following topics:

1. What led to the fall of Silicon Valley Bank, and how can we use data to identify and prevent such collapses in the future? Explore the financial statements of the bank and identify key factors that contributed to its downfall. Use data analysis to identify trends and patterns that can help predict potential bank failures in the future. Predict any other financial institutions following a similar trend, on the verge of collapse. Your analysis can reflect the correlation between current tech layoffs and Tech Banks profitability/trajectory.

2. How was the economy of India maintained during the COVID lockdown? Which government decisions led to the downfall, and how was it revived two years later? Analyze economic indicators such as GDP, inflation, and employment to gain insights into how the Indian economy responded to the pandemic. Identify key policy decisions that helped revive the economy and make recommendations for future crises. One possible idea is to analyze news headlines about prominent economic policies and monetary decisions that led to financial trajectories.

### Dataset

You may choose any dataset within the finance domain that interests you. You’re welcome to use any data source. There are a ton of free data sources available in the Snowflake Data Marketplace.

### Evaluation

Your final deliverable should be a Streamlit website that showcases your data analysis project, including visualizations, insights, and recommendations. Your website should be connected to a Snowflake database and should be easy to navigate and understand. Good luck!

* Build a Streamlit app that connects to data stored in Snowflake and displays this data using one of Streamlit’s newest features, st.experimental\_data\_editor.
* You can start an extended free trial of Snowflake here. There are a ton of free data sources available in the Snowflake Data Marketplace.
* Connecting your Streamlit app to Snowflake can be done in two ways: through using Snowflake’s Python connector or through using Snowpark for Python.
* In addition to displaying your data with st.experimental\_data\_editor, feel free to create data visualizations using any of Streamlit’s powerful charting widgets.
* Does the app successfully connect to data stored in Snowflake?
* Does the app display data with st.experimental\_data\_editor?
* Does the app use other Streamlit widgets to create meaningful data visualizations?

**Bonus points**

* If the data visualizations incorporate input widgets to enable users to adjust specific variables.
* If the app uses Streamlit’s new caching methods: st.cache\_resource and st.cache\_data
* In addition to the requirements outlined above, please note that there will also be points awarded for the complexity of the case study chosen. This means that if you choose a dataset and topic that is more challenging and requires a higher level of analysis, you may receive more points than if you choose a simpler dataset.
* The complexity of your case study will be evaluated based on the depth and breadth of the analysis you conduct, the level of sophistication of your visualizations and modeling techniques, and the overall level of creativity and innovation in your approach.

**Opportunities :**

**Referral for any openings at Streamlit and Snowflake**

**Mentorship session with Data Science team at Snowflake**

**Shoutout from Streamlit on their official social media pages and a blog on their website**