# Capstone Projects

## 1. Haile Hotels and Resorts Reviews Dataset Collection and Analytics

This project is going to be done by three groups and each group shall have **three members**. The first group collects reviews data from **booking.com**, **hoteles.com** and **kayak.com**. The second group will collect reviews from **expedia.com** and **tripadvisor.com**. The third group will collect data from **trip.com** and from Facebook pages (Haile Hotels and Resorts as well as each Haile Hotels and Resorts Facebook pages). The reviews the group collects from these reviews websites and pages is for all Haile Hotels and Resorts listed below.

### 1.1. Data Collection

Each group should collect all reviews with columns as (**rating** (int), **review title** (str), **review comment** (str)) from all sources for the following Haile Hotels and Resorts. Please note that data collected for each hotel and/or resort should be put separately and you may also merge and make overall analysis while you will also do separate analysis of the data about each hotel and resort.

A. Haile Addis Ababa Grand Resort

B. Haile Adama Grand Resort

C. Haile Arba Minch Grand Resort

D. Haile Resort Hawassa

E. Haile Ziway (Batu) Resort or Hotel

F. Haile Resort Gondar

There are some review sites with rating scales from 0-10 and you should convert such scales to 0-5 to make it uniform. Some review sites might not have titles such as Facebook. Facebook might not also have rating values. You can make it missing values for such type non-existent values.

You should employ different techniques to collect your data from review sites. For example, those who collect data from Facebook page reviews can follow this guide to use graph api to collect the public reviews from the main and individual Haile Hotels and Resorts pages. https://developers.facebook.com/docs/graph-api/reference/v2.7/. Trip advisor has paid API and you

should use web scraping techniques from the review pages instead of using API. Other review websites such as booking.com may also have such paid APIs and you should use HTML page scraping instead.

After you finish your data collections separately in each group, then your data should be aggregated to have a single and larger dataset. You should merge the data from the three groups for each hotel and resort separately. For example Haile Resort Arba Minch data collected by the three groups from different sources should be merged to create one larger review dataset about Haile Resort Arba Minch.

**Additional Public Dataset:** In addition to the Haile Hotels and Resorts dataset, you should also collect a public dataset so that your ML model can be trained with larger dataset and can be fine-tuned with Haile Hotels and Resorts dataset. The public review dataset can be downloaded from Kaggle here è https://www.kaggle.com/datasets/datafiniti/hotel-reviews/data.

## 1.2. Sentiment Analysis of Reviews

After collecting your data, you should label each review as positive, negative and neutral. Then you will convert each review to numeric representation such as vectoring the comment (string) data. Then you will train a model to make sentiment classification of each review. Slit the total dataset (from all hotel sites) into 80% training and 20% test split. The data labeling task can be done by both groups in collaboration.

Then an ML model development task should be done separately by each individual group. You should choose classifier algorithms, for example, Naïve Bayes, Decision Trees, Random Forest, Logistic Regression, GradBoost, MLP, CNN, LSTM or SVM. Make sure that you have a very satisfying reason why you choose those algorithms. You may make comparisons with a small number of epochs and choose the best algorithm for example. Finally select three best algorithms and train the models from these three algorithms.

Then train a models instantiated from the selected algorithms, make validations of the models during the training, take performance of the validation results and plot the training plot showing the loss decreasing as the models are trained for more epochs and another plot showing the accuracy of the sentiment classification increasing as the models are trained for more epochs. You can do this routine for each model separately or you can do this using ML Pipeline doing for all the three models altogether.

Finally make tests using your 20% split data to find-out the performance of the classification metrics. You can use different performance metrics for doing the tasks such as classification accuracy, precision, sensitivity, specificity, recall and/or ROC AUC. You should select at least three classification performance metrics but including accuracy anyway. You may also show the confusion matrix for each of the metrics you use.

## 1.3. Areas of Improvement for Haile Hotels and Resorts

Extract the areas of the Haile Hotels and Resorts service segments where reviewers made focus such as reception, rooms, food and catering, shuttle service, bar, sauna and spa, swimming pool, service price, etc. Which part of the area they made a focus on. Then quantify the number of reviews of each topic category

using NLP approaches and prepare plots showing the frequency of the reviews in each category. Then extract areas of improvement where reviewers recommended to be better handled. You may apply best topic identification approaches. Combining the topic identification and the sentiment category of the reviews, you can identify where the opportunity of improvement can be. After making such an analysis and identifying where each review belongs (one review may contain two or more topics), you should label each review with the topic they belong to. For example, if you identified six topics of interest, one review data may belong to “hotel room” and “sauna and spa” topics. It does mean that the review is talking about these two service divisions of the hotels. Then after labeling automatically each review into each topic, you have a dataset which can be used for text classification problems. By the way, you may have an “Other” topic group which doesn’t belong to any of your identified topics.

Then follow similar procedures you went through the sentiment analysis above and train the best three ML models and report your model performance the same way you did in subsection 1.2.

## 1.4. Exploratory Analysis and Dashboard

Prepare a dashboard which shows the word-frequency cloud plots using Seaborn Relplots (https://seaborn.pydata.org/examples/scatterplot_sizes.html), the topics of improvement areas, sentiment analysis plotting the number of negative, positive and neutral reviews, the rating plots for 5, 4, 3, 2 and 1. You should do these plots for each Haile Hotels and Resorts site separately and for all of the hotels aggregated. You may include summary statistical figures included in your dashboard.

## 1.5. Deliverables

**I. Review Dataset:**

A. The review dataset aggregated for each site of the Haile Hotels and Resorts and put into a separate folder and all these folders being put into a single parent folder and being zipped as a ".zip” file.

B. The review sentiment dataset used in subsection 1.2 for all sites aggregated into a single “.csv” file.

C. The topic labeled review dataset used in section 1.3 for all sites aggregated into a single “.csv” file.

D. The scrapping, data collection and data cleaning code used to collect and cleanse data from a folder and being zipped as a ".zip” folder. You should follow standard coding practices to properly commented on part of your code and use proper variables, functions and class naming conventions to make your code readable and can be understood by others.

**II. ML Model Code and Trained Models:**

**A.** The Python Programming code used in subsections 1.2 and 1.3 properly commented and neatly written with each group having been zipped as a ".zip" file.

**B.** Model files (".pt" extension files) developed in subsections 1.2 and 1.3 and put into a single folder with proper names and the folder being zipped as a ".zip" file.

**III. The Dashboard Code:** The dashboard code put in a single file (with proper file structure in it) and being zipped as a ".zip" file if you use multiple files to develop your dashboard.

**IV. Github Public Project Page**: You should create a github public project page, add well explained "README.MD" file, well structured parent folder containing all your codes you developed, any pre-trained models in "/models/" folder, your dataset in "/datasets/" folder both under the parent github folder. In addition to "README" file in the parent folder, you should also have a "requirements.txt" file which specifies the Python libraries one should install before trying to reproduce your solution. The final project report is not included in the github project page.

**V. The Final Project Report:** The final project report should be at most 10 pages excluding the references. You may use small code snippets as part of your documentation to explain how you approached your problem but generally speaking, codes should not be part of your report. The report should have the following sections.

**A. Abstract:** Explain the whole project in under a page and add Keywords at the bottom of your abstract

**B. Introduction:** You should give your reader some introductory briefs about your work.

**C. Statement of the Problem:** You should discuss why review analysis of the hotel and hospitality industry is important and such types of projects are absent from Ethiopian industry. You should emphasize doing the project is relevant for the Ethiopian hotels and hospitality industry. Focus on aspects that you will address (not everything else).

**D. Objectives:** Specify the general objective and the specific objectives you will achieve at the end of your project work.

**E. Related Works:** Make a review of at least four related works done (possibly for other hotel groups in other countries)

**F. Methods:** Describe your methods on how you collect data, how you make data cleaning and transformation, which tools you used, which ML algorithms you employed and how you evaluated your results. Do not forget to add some details about algorithms and methods you used. (Do not skip mentioning it only).

G. **Experimental Description:** Explain how you designed your experiments especially for subsections 1.2 and 1.3 and which ML algorithms used, how you made the selection of those algorithms, how you optimize your model hyper-parameters. You may also describe which performance metrics you used to evaluate (for validation and test) your model.

H. **Results and Discussion:** You should show the results of your work from data collection and cleansing step to the final dashboard. The data collection results are data characterization statistics you found out through exploratory data analysis (EDA). You should also discuss why certain outcomes of your work look like that and show how you interpreted each result.

I. **Conclusion and Recommendations:** Provide brief conclusions and provide actionable recommendations from your entire work.

# 2. Ethiopian Airlines Reviews Dataset Collection and Analytics

This project is going to be done by two groups and each group shall have **three members**. The first group collects reviews data from **tripadvisor.com** and **kayak.com**. The second group will collect reviews from **airlinequality.com** and **Ethiopian Airlines Facebook Page**. The reviews the group collects from these reviews websites and pages is for all Haile Hotels and Resorts listed below.

## 2.1. Data Collection

Each group should collect all reviews with columns as (**rating** (int) for each service (entertainment, cabin crew, food, rest-room, sitting comfort etc), **review title** (str), **review comment** (str)) from all sources.

There are some review sites with rating scales from 0-10 and you should convert such scales to 0-5 to make it uniform. Some review sites might not have titles such as Facebook. Facebook might not also have rating values. You can make it missing values for such type non-existent values.

You should employ different techniques to collect your data from review sites. For example, those who collect data from Facebook page reviews can follow this guide to use graph api to collect the public reviews from the main and individual Haile Hotels and Resorts pages. https://developers.facebook.com/docs/graph-api/reference/v2.7/. Trip advisor has paid API and you should use web scraping techniques from the review pages instead of using API. Other review websites such as booking.com may also have such paid APIs and you should use HTML page scraping instead.

After you finish your data collections separately in each group, then your data should be aggregated to have a single and larger dataset. You should merge the data from the two groups.

**Additional Public Dataset:** In addition to the Haile Hotels and Resorts dataset, you should also collect a public dataset so that your ML model can be trained with larger dataset and can be fine-tuned with Ethiopian Airlines Review dataset. The public review dataset can be downloaded from Kaggle here è https://www.kaggle.com/datasets/juhibhojani/airline-reviews.

## 2.2. Sentiment Analysis of Reviews

After collecting your data, you should label each review as positive, negative and neutral. Then you will convert each review to numeric representation such as vectoring the comment (string) data. Then you will train a model to make sentiment classification of each review. Slit the total dataset (from all hotel sites) into 80% training and 20% test split. The data labeling task can be done by both groups in collaboration.

Then an ML model development task should be done separately by each individual group. You should choose classifier algorithms, for example, Naïve Bayes, Decision Trees, Random Forest, Logistic Regression, GradBoost, MLP, CNN, LSTM or SVM. Make sure that you have a very satisfying reason why you choose those algorithms. You may make comparisons with a small number of epochs and choose the best algorithm for example. Finally select three best algorithms and train the models from these three algorithms.

Then train a models instantiated from the selected algorithms, make validations of the models during the training, take performance of the validation results and plot the training plot showing the loss decreasing as the models are trained for more epochs and another plot showing the accuracy of the sentiment classification increasing as the models are trained for more epochs. You can do this routine for each model separately or you can do this using ML Pipeline doing for all the three models altogether.

Finally make tests using your 20% split data to find-out the performance of the classification metrics. You can use different performance metrics for doing the tasks such as classification accuracy, precision, sensitivity, specificity, recall and/or ROC AUC. You should select at least three classification performance metrics but including accuracy anyway. You may also show the confusion matrix for each of the metrics you use.

## 2.3. Areas of Improvement for Ethiopian Airlines

Extract the areas of Ethiopian Airlines service segments where reviewers made focus such as cabin crew, flight delay, luggage handling, food, sit comfort, rest-room quality, airport check, etc. Which part of the area they made a focus on. Then quantify the number of reviews of each topic category using NLP approaches and prepare plots showing the frequency of the reviews in each category. Then extract areas of improvement where reviewers recommended to be better handled. You may apply best topic identification approaches. Combining the topic identification and the sentiment category of the reviews, you can identify where the opportunity of improvement can be. After making such an analysis and identifying where each review belongs (one review may contain two or more topics), you should label each review with the topic they belong to. For example, if you identified six topics of interest, one review data may belong to “cabin crew” and “food service” topics. It does mean that the review is talking about these two

service divisions of the hotels. Then after labeling automatically each review into each topic, you have a dataset which can be used for text classification problems. By the way, you may have an “Other” topic group which doesn’t belong to any of your identified topics.

Then follow similar procedures you went through the sentiment analysis above and train the best three ML models and report your model performance the same way you did in subsection 1.2.

## 2.4. Exploratory Analysis and Dashboard

Prepare a dashboard which shows the word-frequency cloud plots using Seaborn Relplots (https://seaborn.pydata.org/examples/scatterplot_sizes.html), the topics of improvement areas, sentiment analysis plotting the number of negative, positive and neutral reviews, the rating plots for 5, 4, 3, 2 and 1. You may include summary statistical figures included in your dashboard.

## 2.5. Deliverables

I. **Review Dataset:**

A. The review dataset aggregated (from both groups) reviews and rating dataset as “.csv” file..

B. The review sentiment dataset used in subsection 1.2 for all sites aggregated into a single “.csv” file.

C. The topic labeled review dataset used in section 1.3 for all sites aggregated into a single “.csv” file.

D. The scrapping, data collection and data cleaning code used to collect and cleanse data from a folder and being zipped as a ".zip” folder. You should follow standard coding practices to properly commented on part of your code and use proper variables, functions and class naming conventions to make your code readable and can be understood by others.

**II. ML Model Code and Trained Models:**

**A.** The Python Programming code used in subsections 1.2 and 1.3 properly commented and neatly written with each group having been zipped as a ".zip” file.

**B.** Model files (“.pt” extension files) developed in subsections 1.2 and 1.3 and put into a single folder with proper names and the folder being zipped as a ".zip” file.

**III. The Dashboard Code:** The dashboard code is put in a single file (with proper file structure in it) and being zipped as a ".zip” file if you use multiple files to develop your dashboard.

**IV. Github Public Project Page**: You should create a github public project page, add well explained "README.MD" file, well structured parent folder containing all your codes you developed, any pre-trained models in "/models/" folder, your dataset in "/datasets/" folder both under the parent github folder. In addition to "README" file in the parent folder, you should also have a "requirements.txt" file which specifies the Python libraries one should install before trying to reproduce your solution. The final project report is not included in the github project page.

**V. The Final Project Report:** The final project report should be at most 10 pages excluding the references. You may use small code snippets as part of your documentation to explain how you approached your problem but generally speaking, codes should not be part of your report. The report should have the following sections.

**A. Abstract:** Explain the whole project in under a page and add Keywords at the bottom of your abstract

**B. Introduction:** You should give your reader some introductory briefs about your work.

**C. Statement of the Problem:** You should discuss why review analysis of the hotel and hospitality industry is important and such types of projects are absent from Ethiopian industry. You should emphasize doing the project is relevant for Ethiopian hotels and hospitality industry. Focus on aspects that you will address (not everything else).

**D. Objectives:** Specify the general objective and the specific objectives you will achieve at the end of your project work.

**E. Related Works:** Make a review of at least four related works done (possibly for other hotel groups in other countries)

**F. Methods:** Describe your methods on how you collect data, how you make data cleaning and transformation, which tools you used, which ML algorithms you employed and how you evaluated your results. Do not forget to add some details about algorithms and methods you used. (Do not skip mentioning it only).

**G. Experimental Description:** Explain how you designed your experiments especially for subsections 1.2 and 1.3 and which ML algorithms used, how you made the selection of those algorithms, how you optimize your model hyper-parameters. You may also describe which performance metrics you used to evaluate (for validation and test) your model.

**H. Results and Discussion:** You should show the results of your work from data collection and cleansing step to the final dashboard. The data collection results are data characterization statistics you found out through exploratory data analysis

(EDA). You should also discuss why certain outcomes of your work looks like that and show how you interpreted each result.

I. **Conclusion and Recommendations:** Provide brief conclusions and provide actionable recommendations from your entire work.

# 3. Leverage World Bank Data to Identify and Address Critical Challenges in Ethiopia

This project may involve as many as 15 groups with two persons per group as long as your chosen project scopes don't overlap with each other.

## 3.1. The Goal of Each Project

To utilize advanced data analytics, machine learning, and visualization techniques to uncover actionable insights from World Bank data, focusing on specific challenges in Ethiopia.

## 3.2. The Scopes of Each Project

- **Data Acquisition:** Source relevant data from the World Bank database.
- **Data Cleaning and Preparation:** Cleanse, preprocess, and transform the data for analysis.
- **Exploratory Data Analysis (EDA):** Conduct in-depth EDA to uncover patterns, trends, and correlations.
- **Comparative Data Analysis:** Identify similar countries with Ethiopia using economic, demographic, geographic, and/or historical and political similarities and make comparative analysis between Ethiopia and the chosen counties on the problem domain data you collected.
- **Predictive Modeling:** Develop machine learning models to forecast future trends and make informed decisions.
- **Dashboard Development:** Create interactive dashboards using Plotly Dash to visualize key insights and findings.
- **Report Writing:** Prepare a comprehensive report detailing the entire project, including methodology, results, and recommendations.

## 3.3. The Approaches for Each Project

1. **Problem Identification:**
    - **Identify a pressing issue:** Select a specific challenge in Ethiopia that aligns with their interests and can benefit from data-driven solutions.
    - **Define clear objectives:** Clearly articulate the specific goals of the project.

- **Formulate research questions:** Develop focused questions to guide the analysis.

2. **Data Acquisition and Preparation:**
   - **Access World Bank data:** Utilize the World Bank's Open Data platform to obtain relevant datasets from https://databank.worldbank.org/databases.
   - **Data cleaning:** Handle missing values, outliers, and inconsistencies.
   - **Data transformation:** Normalize, standardize, and engineer features as needed.

3. **Exploratory Data Analysis (EDA):**
   - **Univariate analysis:** Explore individual variables using descriptive statistics and visualizations.
   - **Bivariate analysis:** Investigate relationships between pairs of variables.
   - **Multivariate analysis:** Analyze the relationships among multiple variables.

4. **Comparative Data Analysis:**
   - **Choose Similar Countries:** Identify similar countries with Ethiopia using economic, demographic, geographic, and/or historical and political similarities.
   - **Select variables:** Select variables on which the chosen countries and Ethiopia can be compared.
   - **Dashboard:** Develop at least one dashboard page (as part of the dashboard) and show the comparative charts and plots.

5. **Predictive Modeling:**
   - **Feature engineering:** Create new features to improve model performance.
   - **Model selection:** Choose appropriate machine learning algorithms (e.g., regression, classification, time series forecasting). As the data samples for a given country are yearly and a maximum of 60 years of data is available, you may use models such as ARIMA, Exponential Smoothing, or similar models from statsmodels for your predictive model development.
   - **Model training and evaluation:** Train and evaluate models using relevant metrics (e.g., accuracy, precision, recall, F1-score, RMSE, MAE).

6. **Dashboard Development:**
   - **Interactive visualization:** Create interactive visualizations using Plotly Dash.
   - **User-friendly interface:** Design an intuitive dashboard for easy understanding.
   - **Key insights:** Highlight important findings and trends.

7. **Report Writing:**
   - **Executive summary:** Provide a concise overview of the project.
   - **Introduction:** Introduce the problem, objectives, and methodology.
   - **Data and Methods:** Describe data sources, cleaning, preprocessing, and modeling techniques.
   - **Results and Discussion:** Present key findings, visualizations, and model performance.
   - **Conclusion:** Summarize the main conclusions and implications.

- **Recommendations:** Offer actionable recommendations for stakeholders.

## 3.4. Problem Selection

The problem you will work on should have better data with less missing data especially for recent years. There is a lot of data categories such as:

1. **World Development Indicators**
2. **Statistical Capacity Indicators**
3. **Education Statistics - All Indicators**
4. **Health Nutrition and Population Statistics**
5. **Millennium Development Goals**

Each of these categories do have a lot of variables. You may have a way to navigate through multiple variables to make an overall analysis or you may focus on a few variables and bring deep analytics insights.

## 3.5. Deliverables

I. **The Report:** Report of your entire work containing the following elements:
   A. **Abstract:** Explain the whole project in under a page and add Keywords at the bottom of your abstract
   B. **Introduction:** You should give your reader some introductory briefs about your work.
   C. **Statement of the Problem:** You should discuss why review analysis of the world data variables from an Ethiopian perspective. You should emphasize doing the project is relevant for Ethiopian national goals and strategies as well as policing the field you are analyzing. Focus on aspects that you will address (not everything else).
   D. **Objectives:** Specify the general objective and the specific objectives you will achieve at the end of your project work.
   E. **Related Works:** Make a review of at least four related works done (possibly for other countries or similar work)
   F. **Methods:** Describe your methods on how you choose which category and which variables and why, how you make data cleaning and transformation, which tools you used, which ML algorithms you employed (if any used) and how you evaluated your results. Do not forget to add some details about algorithms and methods you used. (Do not skip mentioning it only).
   G. **Experimental Description:** Explain how you designed your experiments which algorithms used, how you made the selection of those algorithms, etc. You may also describe which performance metrics you used to evaluate the goodness or relevance of your work.
   H. **Results and Discussion:** You should show the results of your work from data collection and cleansing step to the final dashboard. The data collection results are data characterization statistics you found out through exploratory data analysis (EDA). You should also discuss why certain outcomes of your work look like that and show how you

interpreted each result. Use plots, charts and diagrams you developed in your visualization part in your documentation.

l. **Conclusion and Recommendations:** Provide brief conclusions and provide actionable recommendations from your entire work.

II. **The Source Codes:** The zipped folder containing each step of your analytics and visualization works including the dashboard source codes.

III. **The Dataset(s):** The cleansed and possibly transferred datasets you used along with the "readme.md" file where users can read how they can use these datasets if they want to reproduce and enhance what you have done. If you have multiple datasets, you should zip a ".zip" file in a single folder and submit it as a single zipped file.

IV. **Github Public Project Page**: You should create a github public project page, add well explained "README.MD" file, well structured parent folder containing all your codes you developed, any pre-trained models in "/models/" folder, your dataset in "/datasets/" folder both under the parent github folder. In addition to "README" file in the parent folder, you should also have a "requirements.txt" file which specifies the Python libraries one should install before trying to reproduce your solution. The final project report is not included in the github project page.

# 4. Crypto Market Data Analysis, Visualization and Best Predictive Models Development

This project is intended to be done by at most five groups of two persons in each group. The instruction in selection of the cryptocurrencies each group will work on is as follows.

1. Two groups doing four crypto-currencies may overlap with not more than two crypto-currencies. This means that each group should have two crypto-currencies unique to the group.
2. The methods and approaches are expected to be different and, therefore, similar programs, source codes and reports are not acceptable.
3. You should consider data collection for each of your crypto currencies using the following three groups of parameters.
    a. **Short Range:** data for a 1 month period with time-interval of 2 minutes.
    b. **Medium Range:** data for 2 years period with a time-interval of 1 hour.
    c. **Long Range:** data for a period of "max" with a time-interval of 1 day.

## 4.1. Project Goal

Develop a comprehensive analysis of at least four cryptocurrencies, including:

- Data collection and cleaning
- Exploratory Data Analysis (EDA)
- Predictive modeling for price forecasting
- Interactive dashboard development

## 4.2. How You Should Do Your Project

1. **Cryptocurrency Selection:**
    - Choose at least **four cryptocurrencies** for analysis. Consider a mix of established and emerging cryptocurrencies.
    - Justify your selection based on market capitalization, trading volume, and potential for future growth.

2. **Data Collection:**
    - **Source:** Use the Yahoo Finance API (or other reliable sources like CoinGecko, Binance API or Google Finance) to collect historical price data (e.g., daily closing prices, trading volumes) for the selected cryptocurrencies.
    - **Data Cleaning:**
        - Handle missing values (e.g., forward fill, backward fill, interpolation).
        - Address outliers (e.g., Winsorization, removal).
        - Clean and preprocess data for further analysis.
3. **Exploratory Data Analysis (EDA):**
    - **Univariate Analysis:**
        - Calculate summary statistics (mean, median, standard deviation, etc.) for each cryptocurrency in a given period for example for the last one year, for the last one month, for the last one week, five days etc.
        - Visualize price trends using line charts and box plots.
    - **Bivariate Analysis:**
        - Calculate and visualize correlations between the prices of different cryptocurrencies.
        - Analyze the relationship between price and trading volume.

4. **Predictive Modeling:**
    - **Feature Engineering:**
        - Create new features (e.g., moving averages, price momentum, volatility indicators) to improve model performance.
    - **Model Selection:**
        - Experiment with different time series forecasting models:
            - **ARIMA:** Autoregressive Integrated Moving Average
            - **LSTM:** Long Short-Term Memory (for deep learning)
            - **Prophet:** Facebook's time series forecasting library
        - Evaluate model performance using appropriate metrics (e.g., RMSE, MAE, MAPE).
    - **Model Tuning:** Fine-tune model hyperparameters to optimize performance.

5. **Interactive Dashboard Development:**
    - **Choose a Framework:** Use a suitable framework like Plotly Dash, Streamlit, or Flask to create an interactive dashboard.
    - **Dashboard Components:**

        - **Price Charts:** Interactive line charts displaying historical and forecasted prices.
        - **Key Metrics:** Display key performance indicators (e.g., daily returns, volatility, trading volume).
        - **Model Comparisons:** Visualize the performance of different predictive models.
        - **Interactive Controls:** Allow users to:
            - Select cryptocurrencies.
            - Adjust forecasting horizons.
            - Explore different model parameters.

6. **Report Writing:**
    - **Executive Summary or abstract:** Concisely summarize the project goals, methodology, and key findings.
    - **Introduction:** Introduce the project, its objectives, and the selected cryptocurrencies.
    - **Data Collection and Preparation:** Describe the data sources, cleaning, and preprocessing steps.
    - **Exploratory Data Analysis:** Present key findings from the EDA, including visualizations and statistical analyses.
    - **Predictive Modeling:** Detail the model selection, training, evaluation, and tuning process.
    - **Dashboard Development:** Describe the dashboard design, functionality, and key features.
    - **Results and Discussion:** Discuss your findings based on the results you have gotten.
    - **Conclusions and Recommendations:** Summarize the main conclusions, discuss limitations, and suggest potential future research directions.

7. **Github Repository:**

The github repository should be public and should contain the following elements.

- **README.md:**
    - Project title and description.
    - Instructions on how to run the code and use the dashboard.
    - List of dependencies (libraries used in the project).
    - Acknowledgements and contributions.
- **requirements.txt:** List all the necessary Python libraries for the project.
- **Source code:** Your project's source code you used throughout the project.
- **Trained Models:** Your trained models in “/models/” folder of the parent folder.
- **Datasets:** Your cleansed and (transformed) datasets in “.csv” files in under “/datasets/” folder.

8. **Evaluation Criteria:**

These apply to all projects.

- **Data Collection and Cleaning:** Accuracy, completeness, and handling of missing values.
- **EDA:** Depth of analysis, quality of visualizations, and insights gained.
- **Predictive Modeling:** Accuracy of forecasts, model selection, and evaluation.
- **Dashboard Development:** Interactivity, user-friendliness, and effectiveness of visualizations.
- **Report Writing:** Clarity, conciseness, and presentation of findings.
- **Code Quality:** Organization, readability, and use of best practices.
- **GitHub Repository:** Completeness, organization, and adherence to best practices.

## 4.3. Project Deliverables

I. **Report:** A well-organized report detailing the entire project, including methodology, results, and findings.
II. **Source Code:** Well-commented and organized source code in a `.zip` file.
III. **Public GitHub repository:** a public github project page with the source code, datasets, trained models, a `README.md` file (with project description, instructions, and dependencies), and a `requirements.txt` file.
IV. **Datasets:** The final cleansed and transformed clean data you collected and cleansed in during the project with all dataset files put into ".csv" files in a single folder and the folder being zipped as a ".zip" file.

## 4.4. Possible Cryptocurrencies for the Projects

You may check cryptocurrencies with longtime market data from https://finance.yahoo.com/markets/crypto/all/ and select those cryptocurrencies with sufficient data. You should avoid fiat equivalent crypto-currencies such USDT because their market price is set to be equivalent to fiat currency USD.

# 5. Stock Market Data Analysis, Visualization and Best Predictive Models Development

This project is intended to be done by at most ten groups of two persons in each group. The instruction in selection of the stocks each group will work on is as follows.

1. No overlap is allowed and each group should work on at least four stocks.
2. The methods and approaches are expected to be different and, therefore, similar programs, source codes, results and reports are not acceptable.
3. You should consider data collection for each of your stock symbols using the following three groups of parameters.
    a. **Short Range:** data for a **1 month period** with time-interval of **2 minutes**.

b. **Medium Range:** data for **2 years period** with a time-interval of **1 hour**.
c. **Long Range:** data for a period of **"max"** with a time-interval of **1 day**.

## 5.1. Project Goal

Develop a comprehensive analysis of at least four stocks, including:

- Data collection and cleaning
- Exploratory Data Analysis (EDA)
- Predictive modeling for price forecasting
- Interactive dashboard development

## 5.2. How You Should Do Your Project

9. **Stock Selection:**
    - Choose at least **four stocks** for analysis. Consider (not a strict requirement) a mix of established and emerging stocks.
    - Justify your selection based on market capitalization, trading volume, and potential for future growth.

10. **Data Collection:**
    - **Source:** Use the Yahoo Finance API (or other reliable sources like CoinGecko, Binance API or Google Finance) to collect historical price data (e.g., daily closing prices, trading volumes) for the selected stocks.
    - **Data Cleaning:**
        - Handle missing values (e.g., forward fill, backward fill, interpolation).
        - Address outliers (e.g., Winsorization, removal).
        - Clean and preprocess data for further analysis.
11. **Exploratory Data Analysis (EDA):**
    - **Univariate Analysis:**
        - Calculate summary statistics (mean, median, standard deviation, etc.) for each stock in a given period for example for the last one year, for the last one month, for the last one week, five days etc.
        - Visualize price trends using line charts and box plots.
    - **Bivariate Analysis:**
        - Calculate and visualize correlations between the prices of different stock.
        - Analyze the relationship between price and trading volume.

12. **Predictive Modeling:**
    - **Feature Engineering:**
        - Create new features (e.g., moving averages, price momentum, volatility indicators) to improve model performance.
    - **Model Selection:**
        - Experiment with different time series forecasting models:
            - **ARIMA:** Autoregressive Integrated Moving Average

        - **LSTM:** Long Short-Term Memory (for deep learning)
        - **Prophet:** Facebook's time series forecasting library
    - Evaluate model performance using appropriate metrics (e.g., RMSE, MAE, MAPE).
- **Model Tuning:** Fine-tune model hyperparameters to optimize performance.

13. **Interactive Dashboard Development:**
    - **Choose a Framework:** Use a suitable framework like Plotly Dash, Streamlit, or Flask to create an interactive dashboard.
    - **Dashboard Components:**
        - **Price Charts:** Interactive line charts displaying historical and forecasted prices.
        - **Key Metrics:** Display key performance indicators (e.g., daily returns, volatility, trading volume).
        - **Model Comparisons:** Visualize the performance of different predictive models.
        - **Interactive Controls:** Allow users to:
            - Select a stock.
            - Adjust forecasting horizons.
            - Explore different model parameters.

14. **Report Writing:**
    - **Executive Summary or abstract:** Concisely summarize the project goals, methodology, and key findings.
    - **Introduction:** Introduce the project, its objectives, and the selected stocks.
    - **Data Collection and Preparation:** Describe the data sources, cleaning, and preprocessing steps.
    - **Exploratory Data Analysis:** Present key findings from the EDA, including visualizations and statistical analyses.
    - **Predictive Modeling:** Detail the model selection, training, evaluation, and tuning process.
    - **Dashboard Development:** Describe the dashboard design, functionality, and key features.
    - **Results and Discussion:** Discuss your findings based on the results you have gotten.
    - **Conclusions and Recommendations:** Summarize the main conclusions, discuss limitations, and suggest potential future research directions.

**15. Github Repository:**

The github repository should be public and should contain the following elements.

- **README.md:**
    - Project title and description.
    - Instructions on how to run the code and use the dashboard.
    - List of dependencies (libraries used in the project).

    - Acknowledgements and contributions.
- **requirements.txt:** List all the necessary Python libraries for the project.
- **Source code:** Your project's source code you used throughout the project.
- **Trained Models:** Your trained models in “/models/” folder of the parent folder.
- **Datasets:** Your cleansed and (transformed) datasets in “.csv” files in under “/datasets/” folder.

**16. Evaluation Criteria:**

These apply to all projects.

- **Data Collection and Cleaning:** Accuracy, completeness, and handling of missing values.
- **EDA:** Depth of analysis, quality of visualizations, and insights gained.
- **Predictive Modeling:** Accuracy of forecasts, model selection, and evaluation.
- **Dashboard Development:** Interactivity, user-friendliness, and effectiveness of visualizations.
- **Report Writing:** Clarity, conciseness, and presentation of findings.
- **Code Quality:** Organization, readability, and use of best practices.
- **GitHub Repository:** Completeness, organization, and adherence to best practices.

## 5.3. Project Deliverables

V. **Report:** A well-organized report detailing the entire project, including methodology, results, and findings.

**VI.** **Source Code:** Well-commented and organized source code in a `.zip` file.

VII. **Public GitHub repository:** a public github project page with the source code, datasets, trained models, a `README.md` file (with project description, instructions, and dependencies), and a `requirements.txt` file.

VIII. **Datasets:** The final cleansed and transformed clean data you collected and cleansed in during the project with all dataset files put into “.csv” files in a single folder and the folder being zipped as a ".zip” file.

## 5.4. Stocks for the Projects

You may check stocks with longtime market data from https://finance.yahoo.com/markets/stocks/most-active/ and select those stocks with sufficient data.

# 6. Your Own Projects

You can bring your own project ideas and discuss with us to see if it can fit a capstone project and lets you learn and gain valuable experience.