10 Academy: Artificial Intelligence Mastery

Week 2 Challenge Document  
Date: 18 Dec - 24 Dec 2024

# 

# Table of Contents

# 

[**Table of Contents 2**](#_hqh6balq47w)

[**Situational Overview (Business Need) 3**](#_ukvkad37vo78)

[**Data 3**](#_z4zo7fs8378e)

[Learning Outcomes 3](#_1m4xlttsha20)

[**Competency Mapping 4**](#_d237pvcckupg)

[Team 5](#_xjnt1wnjiu9f)

[**Instructions 5**](#_rp5aibwqpuy5)

[Task 1 - User Overview Analysis 6](#_exrs1izgb6pz)

[Task 2 - User Engagement Analysis 7](#_nsp8t86t43b5)

[Task 3 - Experience Analytics 8](#_kh757nuk8wl2)

[Task 4 - Satisfaction Analysis 9](#_do7sj3z4glm1)

[Task 5 : Dashboard Development 9](#_dt66ak952fmq)

[**Tutorials Schedule 11**](#_53kvbjovdjq8)

[Overview 11](#_t8jssbhnjnj7)

[Wednesday (Dec 18, 2024): Data Pre-processing 11](#_6e3wxnb0mfus)

[Thursday(Dec 19, 2024): EDA and Data Modeling 11](#_hsihdfwwph8e)

[Friday (Dec 20, 2024): Dashboard Development 12](#_3gzjb6nxwcsp)

[Monday (Dec 23, 2024): Containerization 12](#_q1mh9sqh4gp7)

[Tuesday (Dec 24, 2024): Question Answering 12](#_xgnn5t58s17y)

[**Deliverables 13**](#_l6dg85st6ozr)

[Interim Submission (Due Friday 20 Dec .2024 2000 UTC) 13](#_od2mkz5qyazo)

[Feedback 13](#_fo37eyc50kle)

[Final Submission (Due Tuesday 24 Dec 2024 2000 UTC) 13](#_nigr1doxtz44)

[Feedback 14](#_3417wvgygsdf)

[**References 15**](#_dlc7nw4122nf)

# 

# Situational Overview (Business Need)

You are working for a wealthy investor that specialises in purchasing assets that are undervalued. This investor’s due diligence on all purchases includes a detailed analysis of the data that underlies the business, to try to understand the fundamentals of the business and especially to identify opportunities to drive profitability by changing the focus of which products or services are being offered.

Your last role with this investor saw you do a rich analysis of a delivery company and you helped to identify that delivery to university students was the most profitable route to follow, and your analysis helped the investor purchase this delivery company and ramp up profits by 25% within 6 months through focussing on the most profitable aspect of the business. This was driven by university students always being hungry, awake at all hours, willing to purchase from a limited food menu, and tending to live within a small geographical area.

The investor is interested in purchasing TellCo, an existing mobile service provider in the Republic of Pefkakia. TellCo’s current owners have been willing to share their financial information but have never employed anyone to look at the data that is generated automatically by their systems.

Your employer wants you to provide a report to analyse opportunities for growth and make a recommendation on whether TellCo is worth buying or selling. You will do this by analysing a telecommunication dataset that contains useful information about the customers & their activities on the network. You will deliver insights you managed to extract to your employer through an easy-to-use web-based dashboard and a written report.

# Data

* We have sourced the data from a month's aggregation of xDR records [here](https://docs.google.com/spreadsheets/d/1MHOSn4e7YNof0KPrwvrD6cPxxKmJB2Wp/edit?usp=drive_link&ouid=117447142056267428412&rtpof=true&sd=true)
* The description for attributes can be found [here](https://docs.google.com/spreadsheets/d/1NCXlyLDFW0ZoXAZ_uzaxWLiixXNPGorCugqe4xpYabc/edit?usp=drive_link)
* This data should be extracted from a PostgreSQL database, and we've included both the database schema and the corresponding SQL file [here](https://drive.google.com/file/d/1Sl1ei0ToZ7Vr0NsWh0laYuUcWw8j6jg3/view?usp=drive_link).

## Learning Outcomes

* Understanding and reasoning the business context. Thinking about suitable analysis that matches the business need. Thinking about clients and their interests.
* Understanding the data provided and extracting insight. You will have to explore different techniques, algorithms, statistical distributions, sampling, and visualization techniques to gain insight.
* Understand the data structure and algorithms used in EDA and machine learning pipelines
* Building a dashboard to explore data as well as to communicate insight. Advanced use of modules such as plotly, seaborn, matplotlib etc. to build descriptive visualizations. Reading through the modules documentation to expand your skill set.
* Thinking about statistical distributions, sampling, bias, overfitting, and correlations.
* Modular and object-oriented Python code writing. Python package building.

# Competency Mapping

The tasks you will carry out in this week’s challenge will contribute differently to the 11 competencies 10 Academy identified as essential for job preparedness in the field of Data Engineering, and Machine Learning engineering. The mapping below shows the change (lift) one can obtain by delivering the highest performance in these tasks.

| **Competency** | **Potential contributions from this week** |
| --- | --- |
| Professionalism for a global-level job | Articulating business values |
| Collaboration and Communicating | Reporting to stakeholders |
| Software Development Frameworks | Using Github for CI/CD, writing modular codes, and packaging |
| Python Programming | Advanced use of Python modules such as Pandas, Matplotlib, Numpy, Scikit-learn, Prophet, and other relevant Python packages |
| SQL programming | Extensive use of SQL on PostgreSQL |
| Data & Analytics Engineering | data filtering, data transformation, and data warehouse management |
| MLOps & AutoML | Pipeline design, data, and model versioning, |
| Deep Learning and Machine Learning | NLP, topic modeling, sentiment analysis |
| Web & Mobile app programming | HTML, CSS , Flask, Streamlit |

## Team

Tutors:

* Mahlet
* Rediet
* Kerod
* Elias
* Emitinan
* Rehmet

# Instructions

At the end of this week, you are expected to have a complete project that has

* Reusable code for data preparation and cleaning.
* Code connected using scikit pipeline or other form of chaining multiple EDA steps
* Beautiful and insightful Streamlit dashboard that shows your findings.
* SQL database as feature store which can be used to store selected features for dashboard visualization and model training.
* Your project folder structure:
  + Has unit tests with good test coverage
  + Has CI/CD setup - using Github Actions
  + Has Dockerfile to build it as a docker image

├── .vscode/

│ └── settings.json

├── .github/

│ └── workflows

│ ├── unittests.yml

├── .gitignore

├── requirements.txt

├── README.md

├── src/

│ ├── \_\_init\_\_.py

├── notebooks/

│ ├── \_\_init\_\_.py

│ └── README.md

├── tests/

│ ├── \_\_init\_\_.py

└── scripts/

├── \_\_init\_\_.py

└── README.md

The global objective is divided into 4 sub-objectives

* User Overview analysis
* User Engagement analysis
* User Experience analysis
* User Satisfaction Analysis

## Task 1 - User Overview Analysis

The lifeblood of any business is its customers. Businesses are always finding ways to better understand their customers so that they can provide more efficient and tailored solutions to them.

Exploratory Data Analysis is a fundamental step in the data science process. It involves all the processes used to familiarise oneself with the data and explore initial insights that will inform further steps in the data science process.

It is always better to explore each data set using multiple exploratory techniques and compare the results. The goal of this step is to understand the dataset and identify the missing values & outliers if any using visual and quantitative methods to get a sense of the story it tells. It suggests the next logical steps, questions, or areas of research for your project.

For the actual telecom dataset, you‘re expected to conduct a full User Overview analysis & the following sub-tasks are your guidance:

* Start by identifying the top 10 handsets used by the customers.
* Then, identify the top 3 handset manufacturers
* Next, identify the top 5 handsets per top 3 handset manufacturer
* Make a short interpretation and recommendation to marketing teams

In telecommunication, CDR or Call Detail Record is the voice channel and XDR is the data channel equivalent. So here, consider xDR as data sessions Detail Record. In xDR, user behavior can be tracked through the following applications: Social Media, Google, Email, YouTube, Netflix, Gaming, and others.

**Task 1.1** - Your employer wants to have an overview of the users’ behavior on those applications.

* Aggregate per user the following information in the column
  + number of xDR sessions
  + Session duration
  + the total download (DL) and upload (UL) data
  + the total data volume (in Bytes) during this session for each application

**Task 1.2** - Conduct an exploratory data analysis on those data & communicate useful insights. Ensure that you identify and treat all missing values and outliers in the dataset by replacing them with the mean or any possible solution of the corresponding column.

You’re expected to report about the following using Python script and slide :

* + Describe all relevant variables and associated data types (slide). findings.
  + Variable transformations – segment the users into the top five decile classes based on the total duration for all sessions and compute the total data (DL+UL) per decile class.
  + Analyze the basic metrics (mean, median, etc) in the Dataset (explain) & their importance for the global objective.
  + Conduct a Non-Graphical Univariate Analysis by computing dispersion parameters for each quantitative variable and provide useful interpretation.
  + Conduct a Graphical Univariate Analysis by identifying the most suitable plotting options for each variable and interpret your findings.
  + Bivariate Analysis – explore the relationship between each application & the total DL+UL data using appropriate methods and interpret your result
  + Correlation Analysis – compute a correlation matrix for the following variables and interpret your findings: Social Media data, Google data, Email data, YouTube data, Netflix data, Gaming data, and Other data
  + Dimensionality Reduction – perform a principal component analysis to reduce the dimensions of your data and provide a useful interpretation of the results (Provide your interpretation in four (4) bullet points maximum).

## Task 2 - User Engagement Analysis

As telecom brands are the data providers of all online activities, meeting user requirements, and creating an engaging user experience is a prerequisite for them. Building & improving the QoS (Quality of Service) to leverage the mobile platforms and to get more users for the business is good but the success of the business would be determined by the user engagement and activity of the customers on available apps.

In telecommunication, tracking the user activities on the database sessions is a good starting point to appreciate the user engagement for the overall applications and per application as well. If we can determine the level of engagement of a random user for any application, then it could help the technical teams of the business to know where to concentrate network resources for different clusters of customers based on the engagement scores.

In the current dataset you’re expected to track the user’s engagement using the following engagement metrics:

* sessions frequency
* the duration of the session
* the session total traffic (download and upload (bytes))

**Task 2.1** - Following the above, perform the following tasks:

* Aggregate the above metrics per customer id (MSISDN) and report the top 10 customers per engagement metric
* Normalize each engagement metric and run a k-means (k=3) to classify customers in three groups of engagement.
* Compute the minimum, maximum, average & total non-normalized metrics for each cluster. Interpret your results visually with accompanying text explaining your findings.
* Aggregate user total traffic per application and derive the top 10 most engaged users per application
* Plot the top 3 most used applications using appropriate charts.
* Using ***k*-means** clustering algorithm, group users in *k* engagement clusters based on the engagement metrics:
  + What is the optimized value of *k (use elbow method for this)*?
  + Interpret your findings.

## Task 3 - Experience Analytics

The Telecommunication industry has experienced a great revolution in the last decade. Mobile devices have become the new fashion trend and play a vital role in everyone's life. The success of the mobile industry is largely dependent on its consumers. Therefore, it is necessary for the vendors to focus on their target audience i.e. what are the needs and requirements of their consumers and how they feel and perceive their products. Tracking & evaluating customers’ experience can help organizations to optimize their products and services so that it meet the evolving user expectations, needs, and acceptance.

In the telecommunication industry, the user experience is related, most of the time, to network parameter performances or the customers’ device characteristics.

In this section, you’re expected to focus on network parameters like [TCP retransmission](https://www.saminiir.com/lets-code-tcp-ip-stack-5-tcp-retransmission/#tcp-retransmission), [Round Trip Time (RTT)](http://www.on-time.com/rtos-32-docs/rtip-32/programming-manual/tcp-ip-networking/tcp/round-trip-time-rtt.htm), [Throughput](https://blog.thousandeyes.com/a-very-simple-model-for-tcp-throughput/#:~:text=A%20Simple%20Model%20for%20TCP%20Throughput&text=Where%20WindowSize%20is%20the%20amount,%2Dto%2Dend%20network%20path.), and the customers’ device characteristics like the handset type to conduct a deep user experience analysis. The network parameters are all columns in the dataset. The following questions are your guidance to complete the task. For this task, you need a Python script that includes all solutions to tasks.

**Task 3. 1** - Aggregate, per customer, the following information (treat missing & outliers by replacing with the mean or the mode of the corresponding variable):

* + Average TCP retransmission
  + Average RTT
  + Handset type
  + Average throughput

**Task 3.2** - Compute & list 10 of the top, bottom, and most frequent:

* 1. TCP values in the dataset.
  2. RTT values in the dataset.
  3. Throughput values in the dataset.

**Task 3.3** - Compute & report:

* 1. The distribution of the average throughput per handset type and provide interpretation for your findings.
  2. The average TCP retransmission view per handset type and provide interpretation for your findings.

**Task 3.4** - Using the experience metrics above, perform a *k-*means clustering (where *k* = 3) to segment users into groups of experiences and provide a brief description of each cluster. (The description must define each group based on your understanding of the data)

## Task 4 - Satisfaction Analysis

Assuming that the satisfaction of a user is dependent on user engagement and experience, you’re expected in this section to analyze customer satisfaction in depth. The following tasks will guide you:

Based on the engagement analysis + the experience analysis you conducted above,

**Task 4.1**  - Write a Python program to assign:

* 1. engagement score to each user. Consider the engagement score as the Euclidean distance between the user data point & the less engaged cluster (use the first clustering for this) ([Euclidean Distance](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.euclidean_distances.html))
  2. experience score for each user. Consider the experience score as the Euclidean distance between the user data point & the worst experience cluster.

**Task 4.2** - Consider the average of both engagement & experience scores as the satisfaction score & report the top 10 satisfied customer

**Task 4.3** - Build a regression model of your choice to predict the satisfaction score of a customer.

**Task 4.4** - Run a k-means (k=2) on the engagement & the experience score.

**Task 4.5**  - Aggregate the average satisfaction & experience score per cluster.

**Task 4.6** - Export your final table containing all user ID + engagement, experience & satisfaction scores in your local MySQL database. Report a screenshot of a select query output on the exported table.

### **Task 5** : Dashboard Development

* Tasks:
  + Designing and developing a dashboard using Visualization tools to visualize data insights.
  + Separate each page by task title for example User overview analysis, User engagement analysis, Experience anysis, Satisfaction analysis. Add a plot for each page.
* KPIs
  + **Dashboard Usability**: Ease of use, with intuitive navigation and clear labels.
  + **Interactive Elements**: Effective use of widgets to enhance user engagement.
  + **Visual Appeal**: Clean and professional design that effectively communicates data insights.
  + **Deployment Success**: Fully functional deployment, accessible via a public URL.

# 

# Tutorials Schedule

## Overview

* Wednesday: Understanding week 2 challenge and Data Pre-processing
* Thursday: Data Exploration and Insights communication
* Thursday: Data modeling
* Friday: Dashboard Development

In the following, the colour **purple** indicates morning sessions, and **blue** indicates afternoon sessions.

## Wednesday (Dec 18, 2024): Data Pre-processing

Here, students will understand this week's challenge and learn how to prepare data for modeling.

* Introduction to week2 challenge(Mahlet)

Key Performance Indicators:

* Understanding advanced code base
* Efficient and modular coding technique
* Understanding data in DB – connecting with DB
* Ability to help others

## Thursday(Dec 19, 2024): EDA and Data Modeling

Here, students will learn how to understand data and communicate insights from the data.

* Data Extraction, Cleaning, Transforming, and formatting using modular Python (Rediet)
* Working with PostgreSQL DB using SQL, Pandas, and SQLAlchemy (Elias)

Key Performance Indicators:

* Proactivity to self-learn - sharing references
* Intermediate to advanced SQL techniques
* Data Understanding and Exploration

## Friday (Dec 20, 2024): Dashboard Development

Here, students will learn how to visually communicate data and insights using dashboards.

* Data visualization, interpretation and communication(Kerod)
* Dashboard development (Rehmet)

Key Performance Indicators:

* Design Thinking
* Dashboard Design
* Proactivity to self-learn - sharing references

## Monday (Dec 23, 2024): Containerization

Here, students will learn how to wrap code files to enable interoperability..

* Containerization and CICD (Emitinan)
* Introduction to Unit Testing: Writing Reliable Code(Rediet)

## Tuesday (Dec 24, 2024): Question Answering

Here, students will ask for clarification on the concepts that are not clear.

* Q&A(Mahlet & Emitinan)

# 

# Deliverables

**MAKE SURE YOUR SUBMISSION IS IN A PDF FORMAT WHEN ASKED TO UPLOAD A WORD/PPT FILE.**

## Interim Submission **(Due Friday 20 Dec .2024 2000 UTC)**

1. Your employer wants a quick meeting after you’ve done a first quick pass of the data and wants to know whether further investigation is useful. To achieve this, summarize your findings from Task 1 and Task 2 in no more than 15 slides. Ensure that you have used univariate, bivariate, and other creative plots to extract relevant insights, also in each slide comment on what you have observed from the plot and tables you produced.
2. Link to your GitHub repository.

### Feedback

You may not receive detailed comments on your interim submission but will receive a grade.

## Final Submission **(Due Tuesday 24 Dec 2024 2000 UTC)**

1. Summarise your findings from Task 1 to 4 (Customer overview, User Engagement, Experience). Your employer demands no more than 30 slides - 1hr presentation.

* 1. Ensure that you include a title slide at the beginning and a reference slide at the end.
  2. Ensure that you make a recommendation to your employer on the growth potential of the company (positive or negative) based on the data.
  3. Ensure that you share the data and slides justifying your recommendation with data and graphs
  4. Ensure that you outline the limitations of your analysis.
  5. Ensure that you make a recommendation on whether your employer should purchase this company.

1. A Github link to all your codes and a screenshot of your dashboard. Make sure to include as many screenshots as possible so that we are able to see all the functionality of your dashboard.
2. A deployed link or screenshots to the dashboard to show dashboard’s functionality. Make an effort to deploy your dashboard live in Heroku, [Netlify](https://netlify.app/), or any one of the free web app hosting services you can find [here](https://bugfender.com/blog/10-best-app-deployment-platforms/).

### Feedback

You will receive comments/feedback in addition to a grade.

# 

# References

General references

* [Exploratory Data Analysis In Python](https://medium.com/@ramdittakavi_61179/exploratory-data-analysis-a-structured-approach-da4d74646445)
* Non-Graphical [Univariate Analysis 1](https://online.stat.psu.edu/stat504/node/11/)
* Non-Graphical [Univariate Analysis 2](https://www.ocsc.go.th/sites/default/files/attachment/page/3502_exploratory-data-analysis.pdf)
* [Univariate and Bivariate Analysis](https://towardsdatascience.com/how-to-use-python-seaborn-for-exploratory-data-analysis-1a4850f48f14)
* [How to define an outlier](https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm#:~:text=Definition%20of%20outliers,what%20will%20be%20considered%20abnormal.)
* [How to Correlation Analysis](https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable5.html)
* [How to do PCA](https://sebastianraschka.com/Articles/2014_pca_step_by_step.html) ([Video](https://www.youtube.com/watch?v=FgakZw6K1QQ))
* [Define telecoms QoS](https://www.eolss.net/Sample-Chapters/C05/E6-108-14-00.pdf)
* [An Oracle Data Science Case Study in Telecom](https://blogs.oracle.com/datascience/an-oracle-data-science-case-study-in-telecom)
* <https://www.statology.org/deciles-in-python/>
* Use cases and challenges in telecom big data analytics paper [(PDF) Use cases and challenges in telecom big data analytics](https://www.researchgate.net/publication/311628734_Use_cases_and_challenges_in_telecom_big_data_analytics/link/587dcdf508ae9275d4e898fc/download)
* <https://github.com/10-Academy-Self-Learning-Resources/Data-Understanding>
* <https://github.com/10-Academy-Self-Learning-Resources/DataVisualization>
* <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html>
* <https://strategyanalytics.medium.com/pandas-read-excel-removed-support-for-xlsx-files-426e4acfde89>
* <https://www.mlflow.org/docs/latest/tutorials-and-examples/tutorial.html>

Design

* [App Layout & Style Tips | Designing Apps for User (Part II) (streamlit.io)](https://blog.streamlit.io/designing-streamlit-apps-for-the-user-part-ii/)
* [Marvel | Create your free account (marvelapp.com)](https://marvelapp.com/signup)