
System Requirements Specification Index

For

MLPython usecase using XGBOOST

Sales forecast analysis L2

1.0

Step to access the work environment

Step 1 use the URL to login provide the username and password

PySpark-Data Analytics-Employee-VIR-Template

180 Mins

1 Sections

1 Skills

1 Questions

5 Total Attempts

70% | -CutOff

100%

System Requirements

- Recommended Browser (Chrome, Safari, Etc)
- Javascript should be enabled in the browser

Link Validity and Cut-off Details

- Link Validity Start Date and Time - 16/9/2024, 7:03 PM
- Cut Off Date and Time - 30/9/2025, 4:05 PM

YAKSHA

Registration Details

First Name *

Last Name *

First Name

Last Name


Email *

Phone (Optional)

Email

Phone

☐ I'm not a robot


reCAPTCHA
Privacy - Terms

Start

Description

PySpark-Data Analytics-Employee-VIR-Template

Step 2 Click on the launch assessment Environment

Speed Test Avg: 4.40Mbps Live: 4.40Mbps Time Remaining: 00:00:00 Final Submit

Question

Data Analytics-Employee-VIR-

Instructions

Introduction

This is a project-based assessment, in which you will be provided with a template code to work. You will be provided with a pre-configured Virtual Machine (VM) to develop the case study.

Launch

You can launch VM by clicking "Launch Assessment Environment". It will take around 5-10 mins to launch the VM.

Configuration

Once launched, if you see any configuration screen, skip it.

Cloning

Once VM is up, it will take another 30 sec-1 min to clone your project template on desktop of your VM. Please wait till then.

Document

The project will be cloned in a folder, named same as your email ID. The folder contains template code and case study document. You are required to open the case study document and thoroughly go through it to understand project and mandatory process.

Development

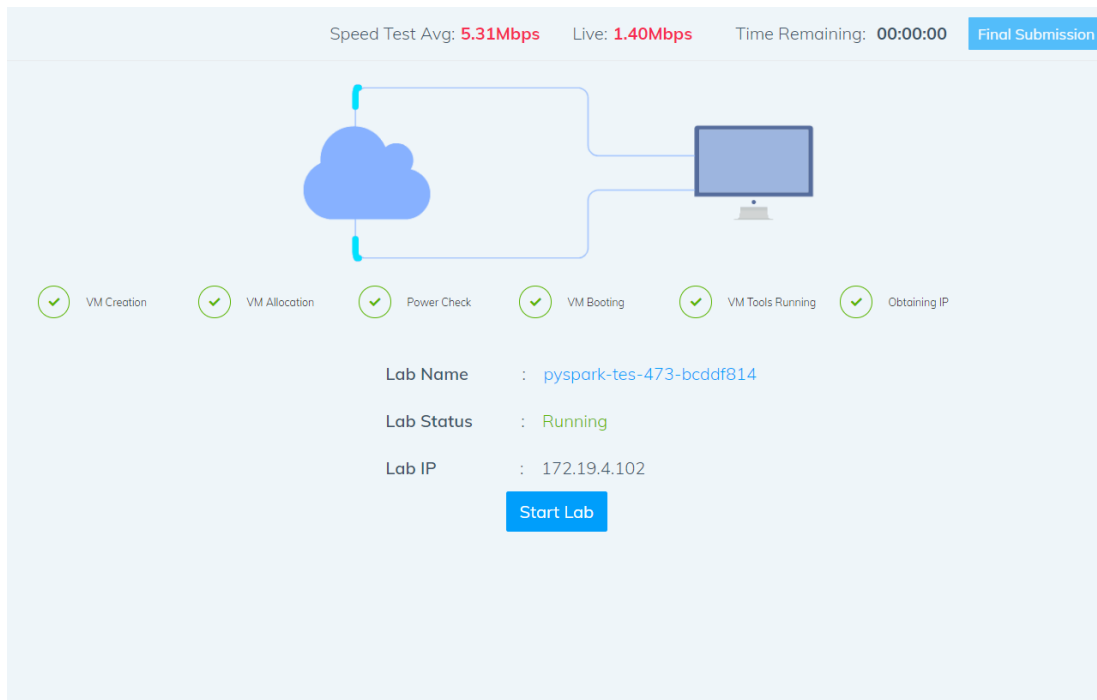
To develop use case all IDEs, Database required are available in VM. There is a README file on desktop containing any credentials you need.

Submission

Before you do final submission of your code there are 2 mandatory things you have to follow, else your evaluation result would be affected

1. RUN TEST CASE (AS MENTIONED IN DOCUMENT)
2. PUSH CODE TO GIT (AS MENTIONED IN DOCUMENT)

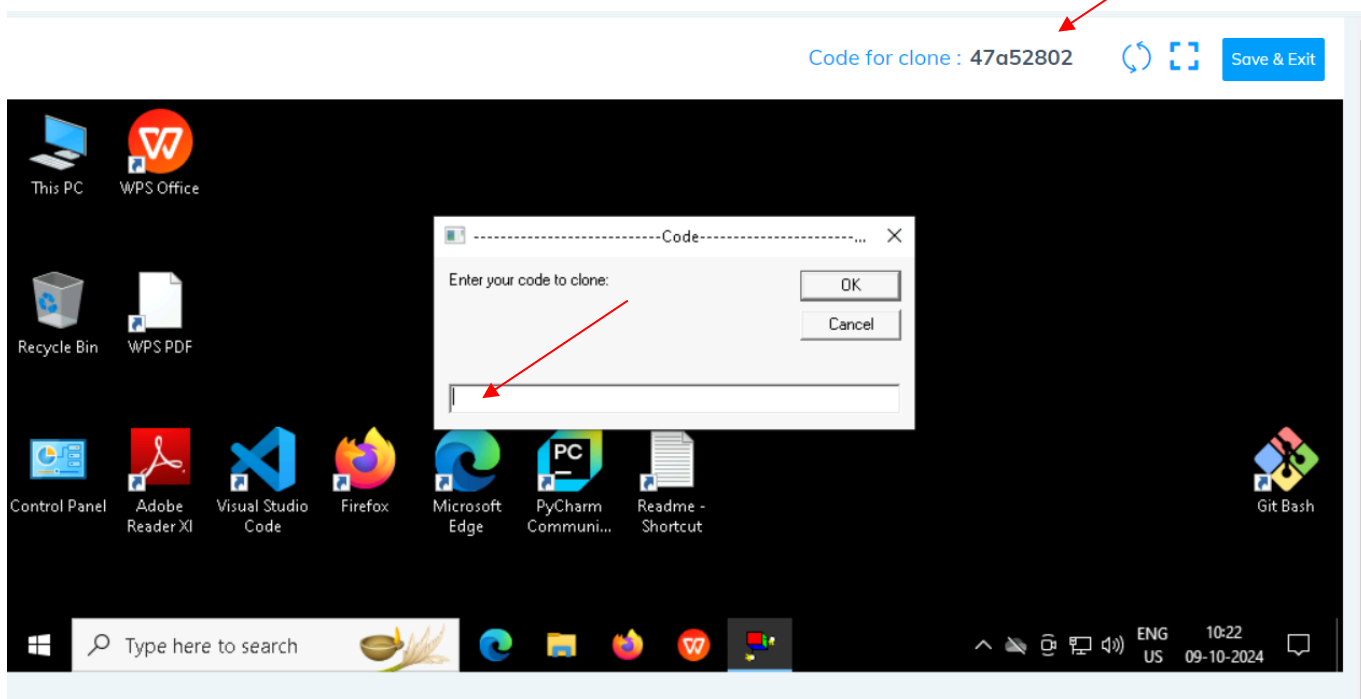
Launch Assessment Environment



Step 3 Click on the start lab button

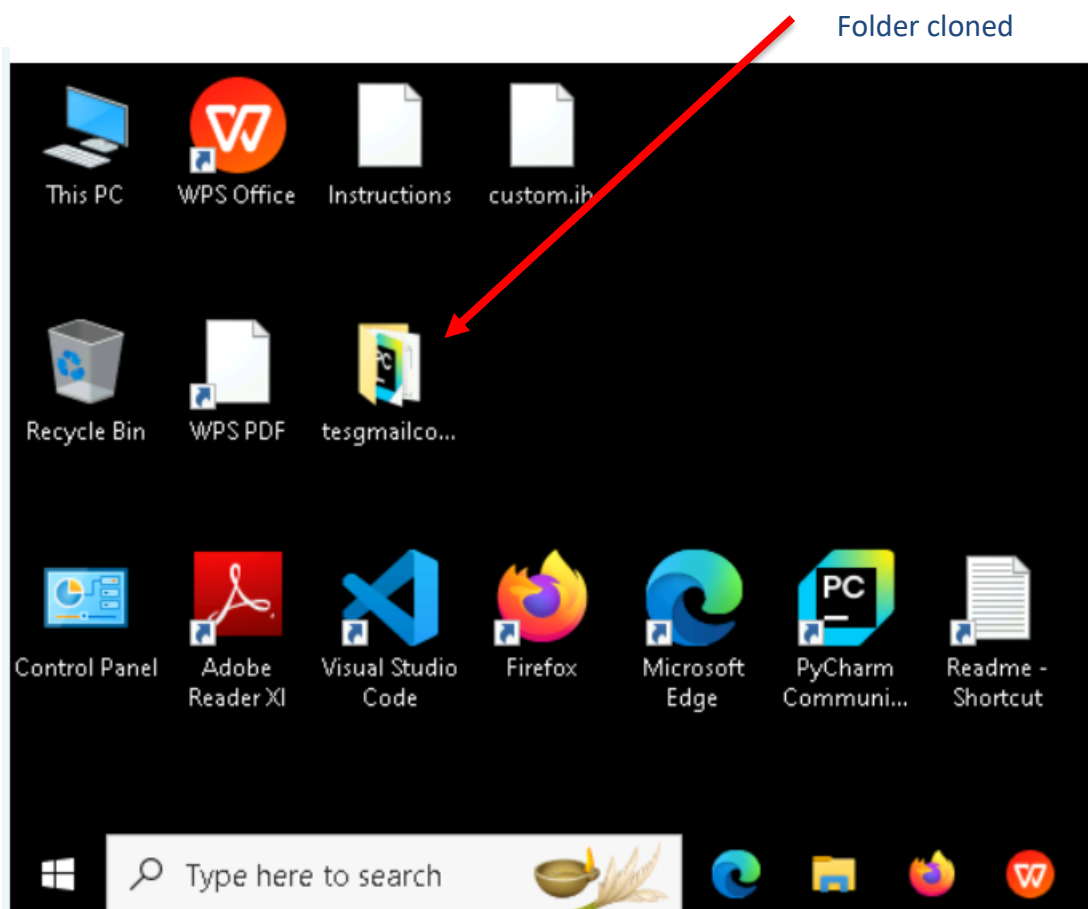
Step 4 you will get a window you need to type the code from that top corner

- You need to type the code in the window . It will take few minutes to start the window

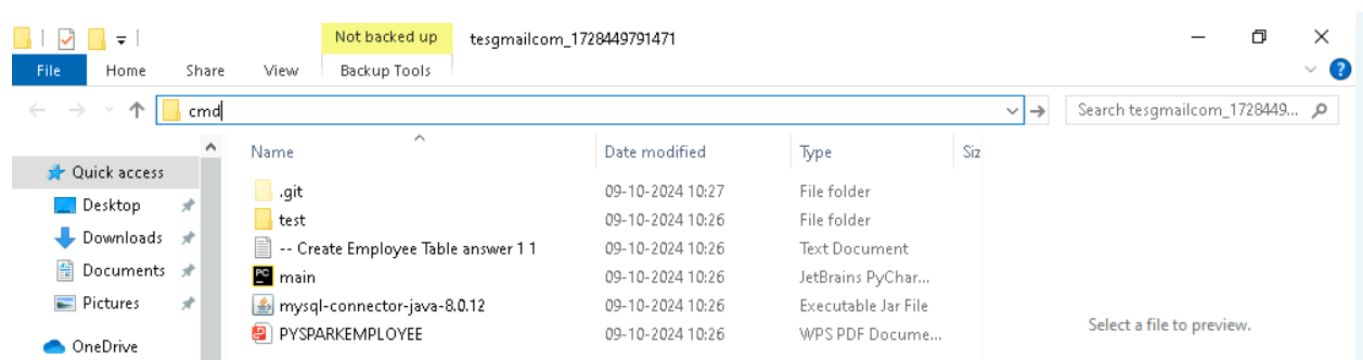


Click on ok

Step 5 after few seconds we can see that the your folder is cloned in the desktop .

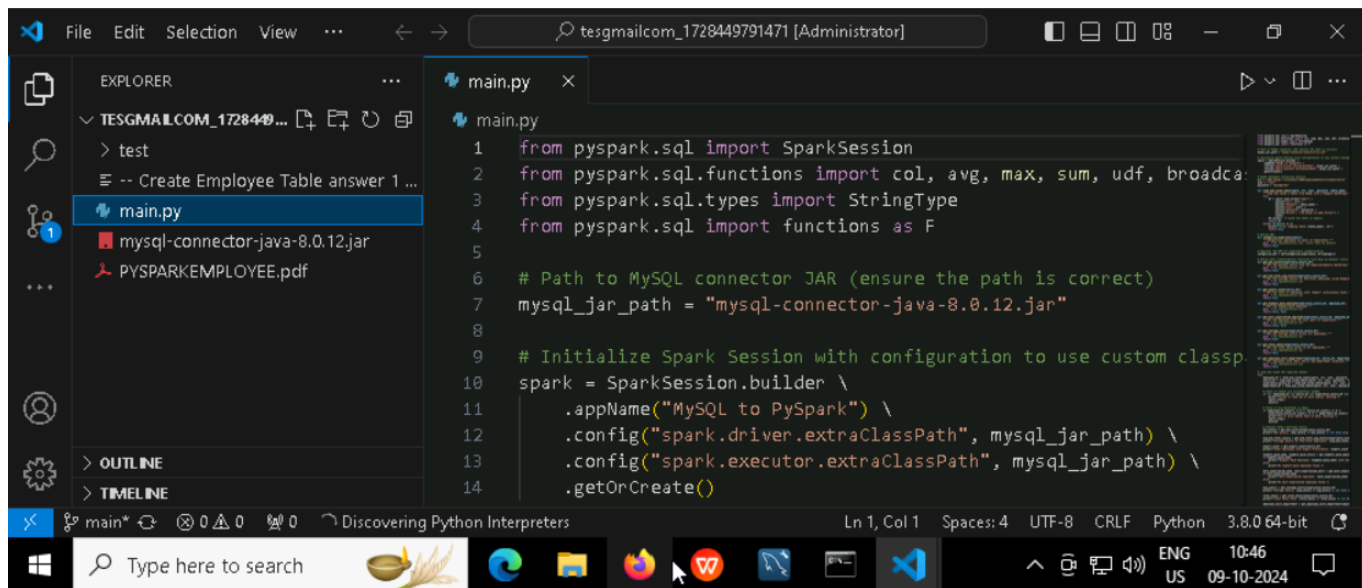


Step 6 go inside the folder type cmd in the top of the file explorer



- Type **code**. And hit enter you can see that workspace is opened in the visual code



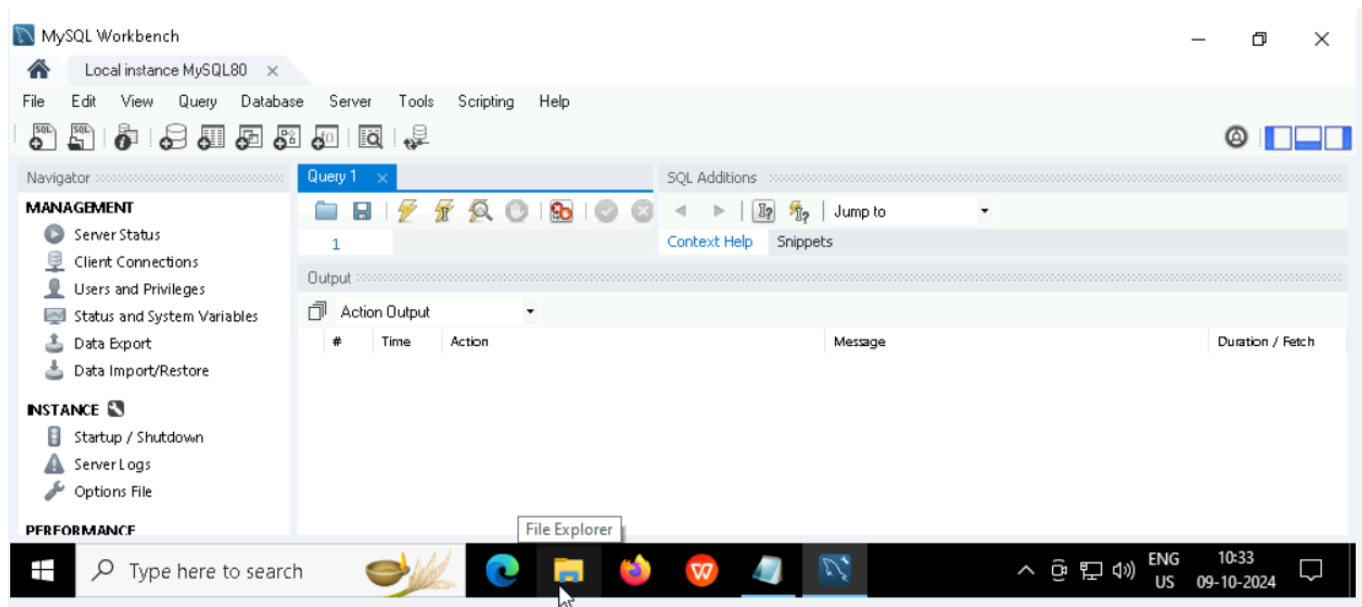


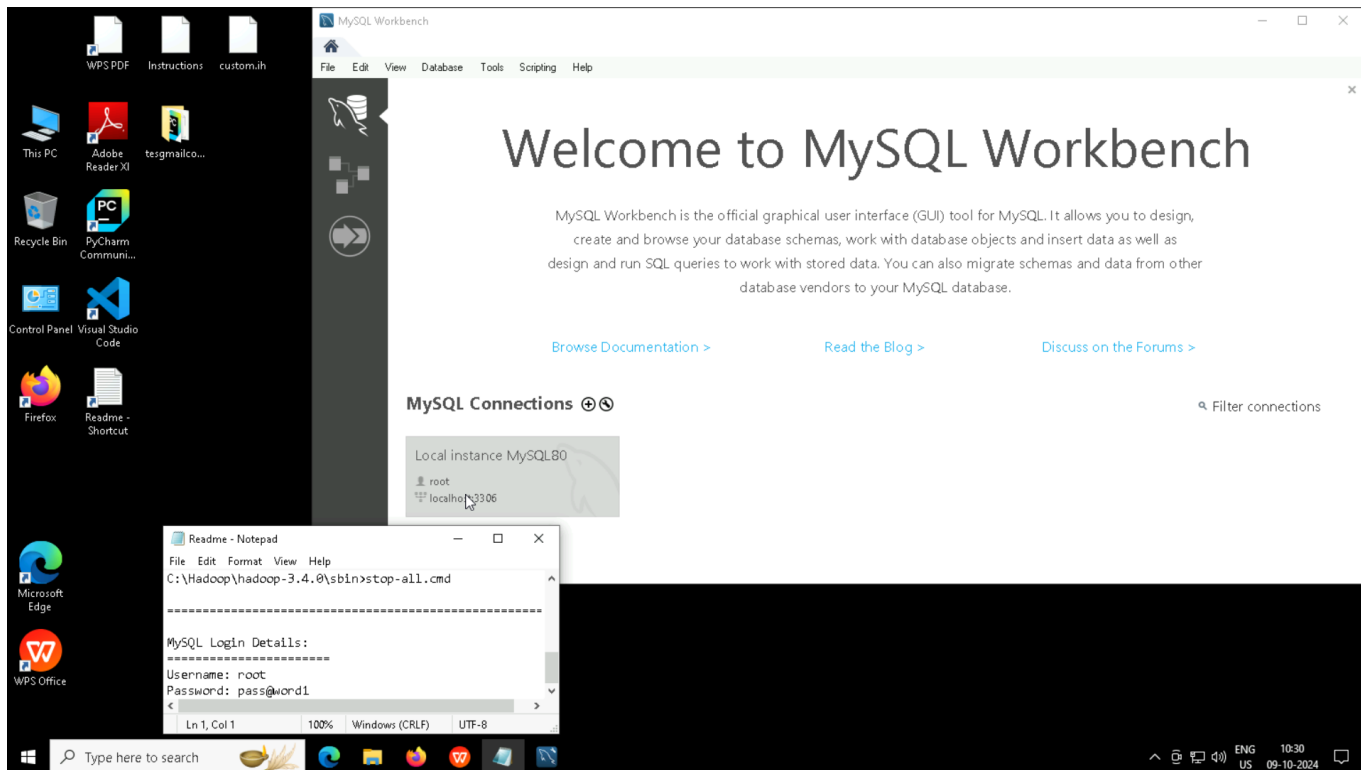
- You can see that workspace is ready to code

Note Please only work with visual code not with any other IDE

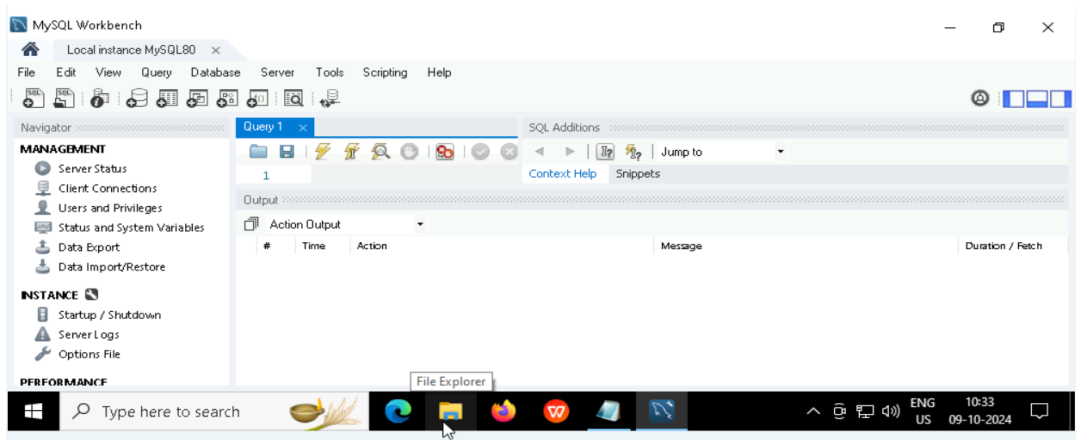
- In the folder cloned you will have all the project files needed .
- In the search bar open type workbench .
- The username & password for the mysql database is given the readme shortcut file .

Create the database with the same name **employeeetails**



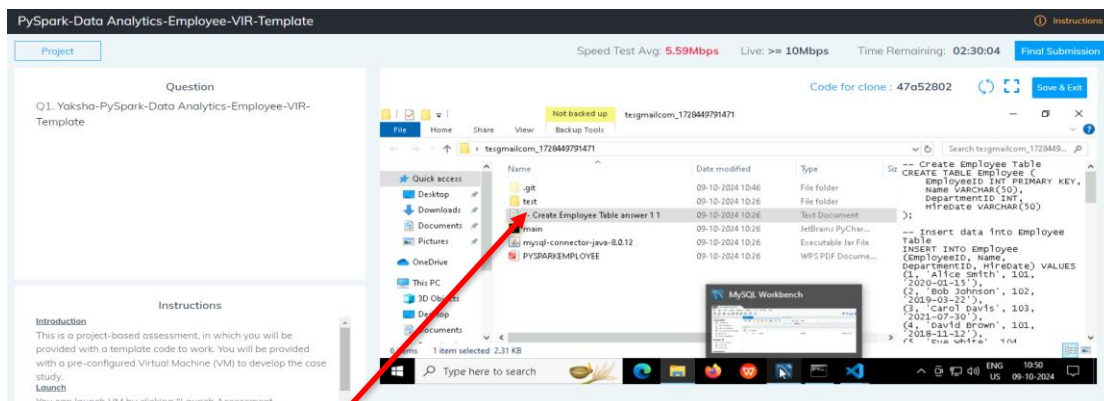


You are requested to input the password for the MySQL workbench .



Create the required database with the with the given instruction below

You can create the table using the following script which is given folder



You can copy and paste in the workbench to create the required tables

Problem Statement : **Sales forecast analysis**
Description : Use relevant methods operations to perform specified activities which are given in the instructions.

The retail company, FutureMart, has been growing rapidly, expanding its reach across multiple regions and adding a diverse range of products across various categories. With a wide network and a large customer base, FutureMart now aims to improve its sales forecasting capabilities and optimize operations by leveraging machine learning models and data analytics. The goal is to make informed decisions on inventory management, regional sales strategies, and targeted marketing based on predicted and historical sales trends.

FutureMart's data includes key details about each transaction, such as:

- Order Date and Ship Date: Dates when the order was placed and shipped.
 - Customer Segment: Classification of customers into segments like Corporate, Home Office, or Consumer.
 - Region, Country, State, City: Location data to analyze sales distribution geographically.
 - Product Category and Sub-Category: Classification of items to study category-specific performance.
 - Sales: Actual sales amount, which serves as the target variable in the model for prediction.
- Data Challenges and Goals:
- Data Cleaning and Preprocessing: Some fields may contain missing values or need transformation, especially date fields and categorical values.
 - Predictive Modeling: FutureMart is using XGBoost, a powerful gradient boosting algorithm, to build a regression model that predicts sales based on selected features.
 - Sales Insights and Comparison: FutureMart needs to analyze trends in specific regions and product categories to tailor business strategies.
 - Forecasting and Scenario Planning: By predicting sales trends and identifying high-performing regions and products, FutureMart can plan for future demand and optimize stock levels.

Analytical and Predictive Objectives:

1. Sales Forecasting: Generate sales predictions for specific input features to estimate revenue for future orders.
2. Regional and Category Analysis: Determine high-performing regions, cities, and categories to allocate resources efficiently.
3. Sales Trend Analysis: Understand historical sales trends across categories to prepare for seasonal demand and optimize inventory.
4. Product-Level Insights: Identify top-selling products and areas with untapped potential to boost sales through targeted marketing.
5. Sales Comparison: Analyze differences between regions and categories to uncover patterns and inform business expansion decisions.

Hints and Steps to Achieve This Use Case:

1. Data Preparation:
 - Ensure that date fields, like "Order Date" and "Ship Date," are properly parsed and transformed into datetime objects.
 - Use Label Encoding for categorical variables (e.g., 'Ship Mode', 'Region') to make them compatible with machine learning models.

- Handle missing values by filling or imputing them, especially in fields that may impact predictions (e.g., postal codes).
- 2. Feature Engineering:
 - Select Relevant Features: Focus on features likely to influence sales, such as customer segment, region, and category.
 - Add New Features: Derived features like "Days to Ship" (difference between order and ship dates) could provide valuable insights.
- 3. Model Training:
 - Train an XGBoost Regressor model on the training data, tuning parameters like the number of estimators and learning rate.
 - Evaluate Model Performance using validation data to ensure it accurately predicts sales without overfitting.
- 4. Prediction Function:
 - Develop a function (`predicted_sales_for_input`) that takes in specific input features to predict future sales. Ensure the function standardizes and preprocesses inputs to match the training data format.
- 5. Analytical Functions for Insights:
 - Implement functions like `average_sales_for_region` and `total_sales_for_category` to summarize historical data.
 - These functions provide business insights, such as which region has the highest average sales or which category generates the most revenue, helping to make data-driven decisions.
- 6. Trend and Distribution Analysis:
 - Use functions like `trend_for_category` and `region_sales_distribution` to visualize sales patterns and distributions.
 - Hint: Look at seasonality and year-over-year trends to understand how sales behave over time and adjust inventory accordingly.
- 7. Evaluation and Model Loading:
 - Save the model in a pickle file and ensure it can be reloaded for future use.
 - Verify that the model and label encodings are compatible to make predictions on new, unseen data.

Solve these Questions

1. **What is the predicted sales for a specific input?**
 - Given a set of input features like Ship Mode, Segment, Region, Category, Sub-Category, and Postal Code, predict the sales value using the XGBoost model.
 - *Why it's useful:* This function helps the company forecast sales for potential future orders, allowing them to anticipate revenue.
2. **What is the average sales for Region X?**
 - Calculate the average sales for a specific region based on historical data.
 - *Why it's useful:* Identifying average sales per region helps in understanding geographic performance and planning region-specific strategies.
3. **What is the total sales for Category X?**
 - Sum up all sales for a specified product category.
 - *Why it's useful:* It reveals which categories generate the most revenue, guiding inventory and marketing efforts.
4. **Which region has the highest average sales?**
 - Determine the region with the highest average sales value.
 - *Why it's useful:* Pinpointing high-performing regions enables the company to optimize resource allocation and focus efforts on expanding market share.
5. **What is the trend of sales for Category X?**

- Analyze the trend in sales over time for a specific product category.
- *Why it's useful:* Understanding sales trends aids in forecasting demand and managing seasonal inventory.
- 6. What is the sales distribution by region?**
 - Display the distribution of sales across different regions.
 - *Why it's useful:* This breakdown highlights regional contributions to total sales, supporting strategic planning and growth initiatives.
- 7. What is the total and average sales for Product X?**
 - Calculate both total and average sales for a specific product by product ID.
 - *Why it's useful:* Assessing individual product performance assists in product line decisions and marketing focus.
- 8. What is the predicted average sales for City X?**
 - Estimate the average sales value for a given city using historical data.
 - *Why it's useful:* Predicting sales at the city level helps in fine-tuning regional strategies and aligning local marketing campaigns.
- 9. What is the total sales distribution by category?**
 - Summarize sales distribution across different product categories.
 - *Why it's useful:* It helps the company focus on high-revenue categories and make inventory and product strategy decisions.
- 10. What is the difference in sales between Region X and Region Y?**
 - Calculate the sales difference between two specific regions.
 - *Why it's useful:* This insight allows a comparison of regions, highlighting areas for improvement or growth.
- 11. What is the product number with the highest sales?**
 - Identify the product with the highest sales value in the dataset.
 - *Why it's useful:* Knowing top-selling products allows the company to focus marketing and inventory efforts on their best performers

Execution Steps to Follow:

1. All actions like build, compile, running application, running test cases will be through Command Terminal.
2. To open the command terminal the test takers, need to go to Application menu (Three horizontal lines at left top) -> Terminal -> New Terminal
3. This editor Auto Saves the code
4. If you want to exit (logout) and continue the coding later anytime (using Save & Exit option on Assessment Landing Page) then you need to use CTRL+Shift+B-command compulsorily on code IDE. This will push or save the updated contents in the internal git/repository. Else the code will not be available in the next login.
5. These are time bound assessments the timer would stop if you logout and while logging in back using the same credentials the timer would

resume from the sametime it was stopped from the previous logout.

6. To setup environment:

You can run the application without importing any packages

7. To launch application:

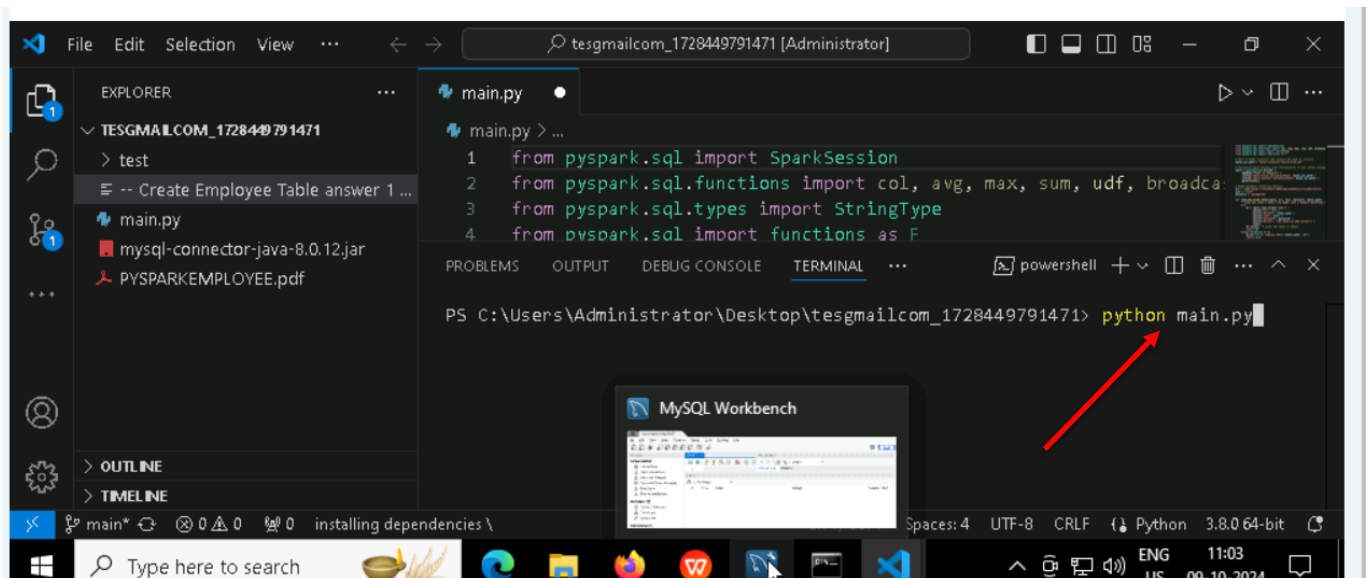
Python sales.py

8. To run Test cases:

python -m unittest

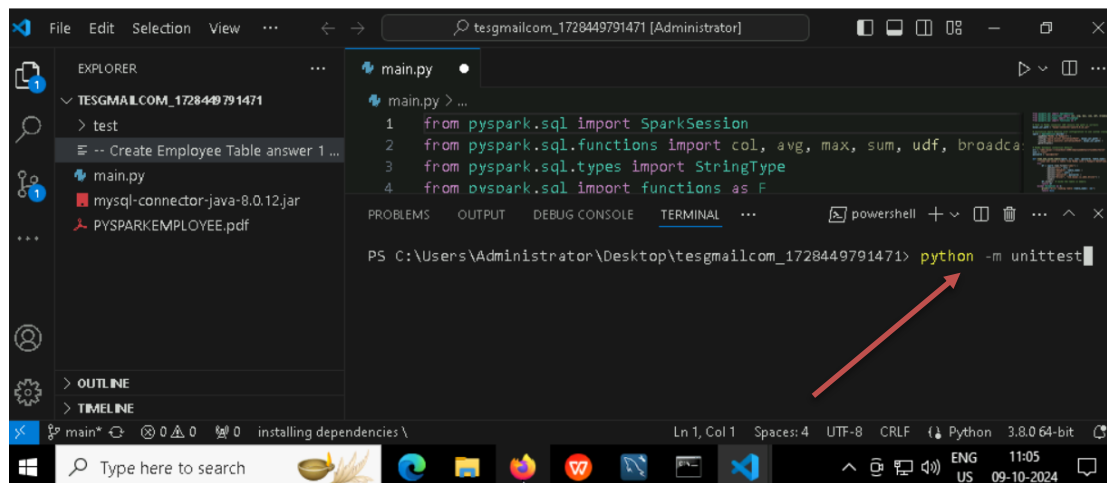
Before Final Submission also, you need to use CTRL+Shift+B-
command compulsorily on code IDE. This will push or save the
updated contents in the internalgit/repository for code

Screen shot to run the program



To run the application

- Python main.py



To run the testcase

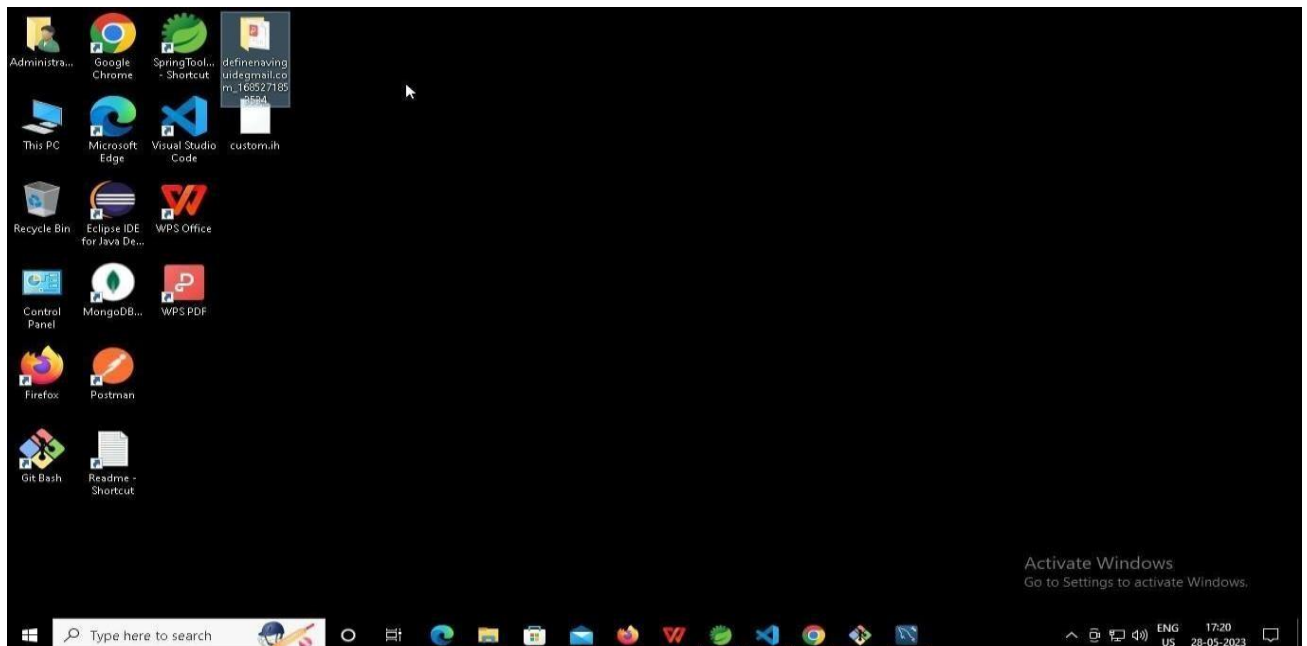
- **Python -m unittest**

Screenshot to push the application to github

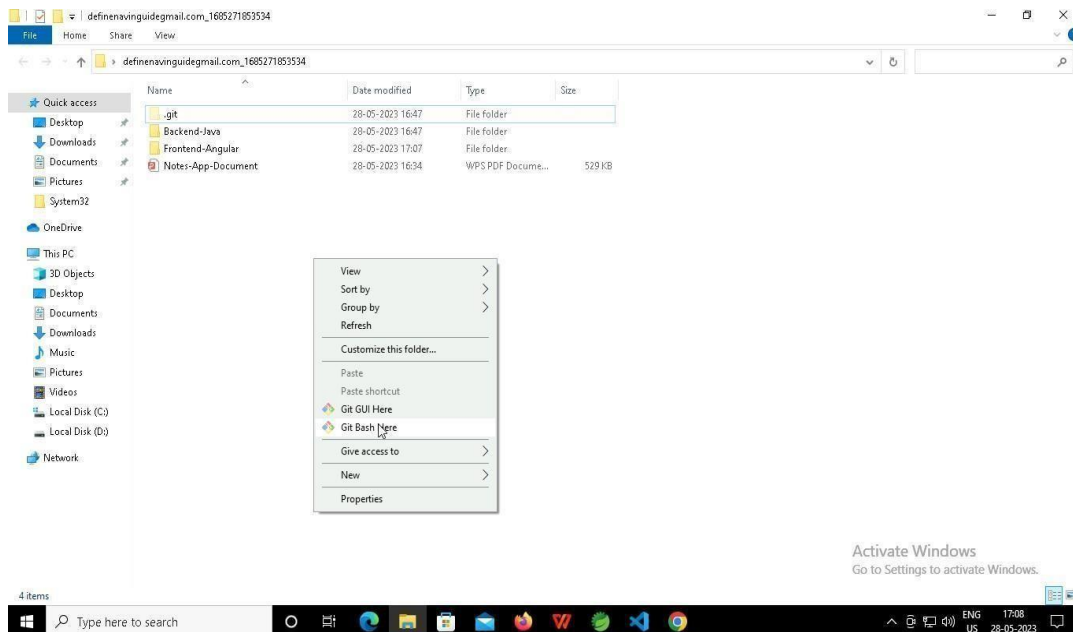
-----X-----

You can run test cases as many numbers of times and at any stage of Development, to check howmany test cases are passed/failed and accordingly refactor your code.

1. **Make sure before final submission you commit all changes to git.** For that open theproject folder available on desktop



a. **Right click in folder and open Git Bash**



b. In Git bash terminal, run following commands

c. `git status`

```
Administrator@2a5ee7ad258f58c MINGW64 ~/Desktop/tesgmailcom_1728449791471 (main)
$ git status
On branch main
Your branch is up to date with 'origin/main'.

Changes not staged for commit:
  (use "git add/rm <file>..." to update what will be committed)
  (use "git restore <file>..." to discard changes in working directory)
        deleted:    templatespark.py

no changes added to commit (use "git add" and/or "git commit -a")

Administrator@2a5ee7ad258f58c MINGW64 ~/Desktop/tesgmailcom_1728449791471 (main)
$
```

d. `git add .`

```
Administrator@2a5ee7ad258f58c MINGW64 ~/Desktop/tesgmailcom_1728449791471 (main)
$ git add .
```

e. `git commit -m "First commit"`

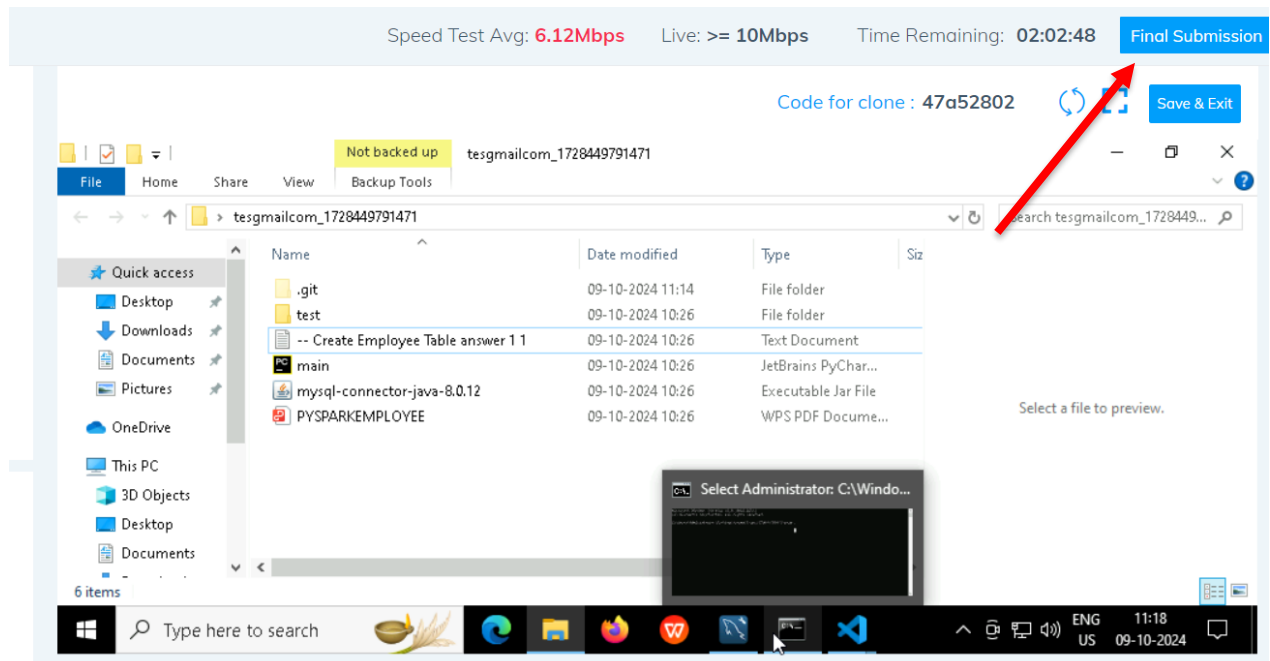
(You can provide any message every time you commit)

```
Administrator@2a5ee7ad258f58c MINGW64 ~/Desktop/tesgmailcom_1728449791471 (main)
$ git commit -m "first commit"
[main f97ce24] first commit
1 file changed, 91 deletions(-)
delete mode 100644 templateespark.py
```

f. git push

```
Administrator@2a5ee7ad258f58c MINGW64 ~/Desktop/tesgmailcom_1728449791471 (main)
$ git push
Enumerating objects: 3, done.
Counting objects: 100% (3/3), done.
Delta compression using up to 4 threads
Compressing objects: 100% (2/2), done.
Writing objects: 100% (2/2), 212 bytes | 212.00 KiB/s, done.
Total 2 (delta 1), reused 0 (delta 0), pack-reused 0 (from 0)
remote: Resolving deltas: 100% (1/1), completed with 1 local object.
To https://github.com/IIHTDevelopers/tesgmailcom_1728449791471.git
a1c1905..f97ce24 main -> main
```

After you have pushed your code Finally click on the final submission button



You should see a screen like this you will have to wait for the results . after getting this page you can leave the system

