GitHub Repository

You can find the project repository on GitHub: GitHub Link

a) Title of the Project

"Analyzing Weather Patterns Using Spark and Tableau"

b) Project Idea

- This project aims to analyze weather patterns using a large weather dataset, leveraging Apache Spark for distributed processing, Tableau prep for data cleaning and Tableau for visualization.
- The project will focus on extracting actionable insights related to temperature trends, rainfall patterns, and humidity variations across different locations.
- Insights will be visualized through interactive dashboards, enabling better understanding and forecasting of weather conditions.

c) Tools and Technologies

We will use the following tools and technologies

- **Apache Spark:** Distributed data processing for large-scale datasets.
- **Python:** Writing Spark scripts for analysis.
- **Tableau:** Visualization of results through interactive dashboards.
- **Tableau Prep:** Data cleaning of dataset .
- **SQL:** For querying and managing the weather data, particularly for aggregation and filtering in Apache Spark
- **Jupyter Notebook:** Jupyter Notebooks are ideal for iterative development, allowing for easy integration of Python code with visualizations, making the data analysis process transparent and reproducible.

d) High-Level Architecture

Block Diagram Components

- **Dataset**
- **Data cleaning(Tableau Prep)**

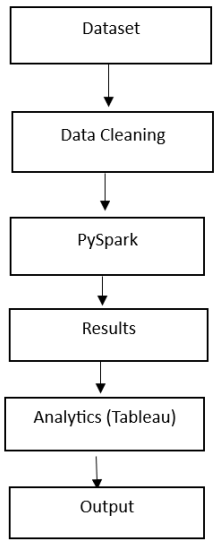

Figure 1: Flowchart

- **PySpark**
- **Results**
- **Analytics(tableau)**
- **Output**

e) Explanation of the Diagram

- **Dataset:** Raw data containing weather attributes like location, temperature, rainfall, and humidity.Acts as the starting point for analysis.
- **Data cleaning(Tableau Prep) :** A tool that used to process raw dataset and remove unnecessary data and produces dataset without any inaccuracy and inconsistency in data.
- **PySpark:** A powerful tool to process the dataset efficiently. Handles tasks like cleaning data, calculating averages, and identifying trends. Prepares the data for visualization by transforming and aggregating it.
- **Results:** Processed and refined data obtained after running PySpark operations. Summarized metrics like average rainfall or temperature variations are ready to be visualized.

- **Analytics(tableau):** A tool to create visualizations like graphs, charts, and dashboards. Makes it easier to understand the data by showing trends and patterns interactively.
- **Output:** Final deliverables, such as Tableau dashboards or reports. Provides clear insights for stakeholders to make informed decisions.

f) Goals to Investigate

- Compute the average MinTemp and MaxTemp for each Location to understand general temperature trends
- Identify Locations where MinTemp or MaxTemp exceeds certain thresholds (e.g., MinTemp less than 0°C or MaxTemp greater than 40°C).
- Find the top 5 locations with the highest average Maximum temperature
- Identify days and locations where there was an unusually large drop in temperature compared to the previous day.
- Analyze the relationship between temperature (MinTemp, MaxTemp), rainfall, and humidity to identify patterns of high humidity conditions. This goal would involve computing: The average humidity (Humidity) under different ranges of MinTemp and MaxTemp. Determine if higher temperatures or rainfall correlate with higher humidity levels.
- Analyze the number of days with rainfall greater than 0 for each location.

g) Implementation Steps

- **Install Tableau Prep:** `https://www.tableau.com/products/prep/download`.
- **Install PySpark:** Ensure you can run JupyterLab from the PowerShell window.
- **Install Tableau Desktop:** `https://www.tableau.com/academic/students`.

Dataset Source

`https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data`

Implementation Steps

1. **Open Tableau Prep:** Click on *Connect*, select **Microsoft Excel**, and load data from Excel. *Note: Ensure the dataset file is in Excel format before importing.*

2. **Add a Clean Step:** Click on the (+) symbol next to the dataset and add a **Clean Step**.

3. **Hide and Rename Columns:**

   - Hide columns that are not relevant to the goals.
   - Rename columns for clarity:
     - **MinTemp** → **MinTemp_in_Celsius**
     - **MaxTemp** → **MaxTemp_in_Celsius**
     - **Rainfall** → **Rainfall_in_millimeters**

4. **Change Row_id to Index:**

   - Use the **Split** function for values like `row1, row2, row3,...`.
   - Click on **Custom Split**, use `'w'` as the separator, and select **Split Off: Last**.
   - Rename the new column **row_id_split** to **Index**.

5. **Export Cleaned Data:** Click on the (+) symbol next to the split step and select **Output**. Export the cleaned dataset as a CSV file and save it locally as `cleaned_weather_data.csv`.

6. **Open JupyterLab:** Launch **JupyterLab** from PowerShell.

7. **Start PySpark Session:** Use the following command to initialize the PySpark session:

   ```
   from pyspark.sql import SparkSession
   spark = SparkSession.builder.appName('WeatherAnalysis').getOrCreate()
   ```

   *Note: The app name is set as **WeatherAnalysis**.*

8. **Create a DataFrame:** Import the cleaned dataset into a PySpark DataFrame:

   ```
   df = spark.read.csv('cleaned_weather_data.csv', inferSchema=True, header=True)
   ```

9. **Goals:**

- **Goal 1:** Analyze average temperatures for each location (MinTemp_in_Celsius and MaxTemp_in_Celsius).

- **Goal 2:** Identify rows where MinTemp or MaxTemp exceeds certain thresholds (e.g., MinTemp ¡ 0°C or MaxTemp ¿ 40°C)

- **Goal 3:** Find the top 5 locations with the highest average Maximum temperature

- **Goal 4:** Identify days and locations where there was an unusually large drop in temperature compared to the previous day.

- **Goal 5:** Analyze the relationship between temperature (MinTemp, MaxTemp), rainfall, and humidity to identify patterns of high humidity conditions. This goal would involve computing: The average humidity (Humidity) under different ranges of MinTemp and MaxTemp. Determine if higher temperatures or rainfall correlate with higher humidity levels.

- **Goal 6:** Analyze the number of days with rainfall greater than 0 for each location.

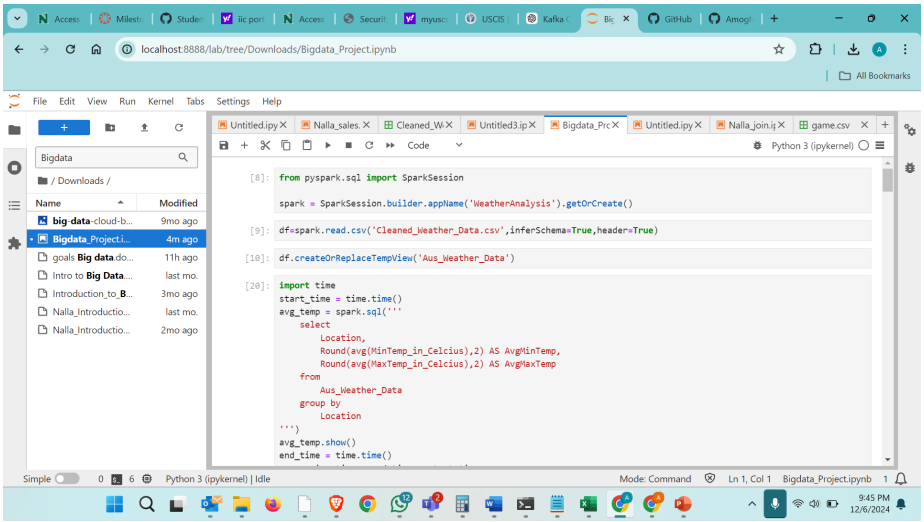

Figure 2:

10. After extracting this excel output files , open tableau desktop and load all these files.Drag and drop the columns in respective rows and columns select appropriate chart to visualize the chart

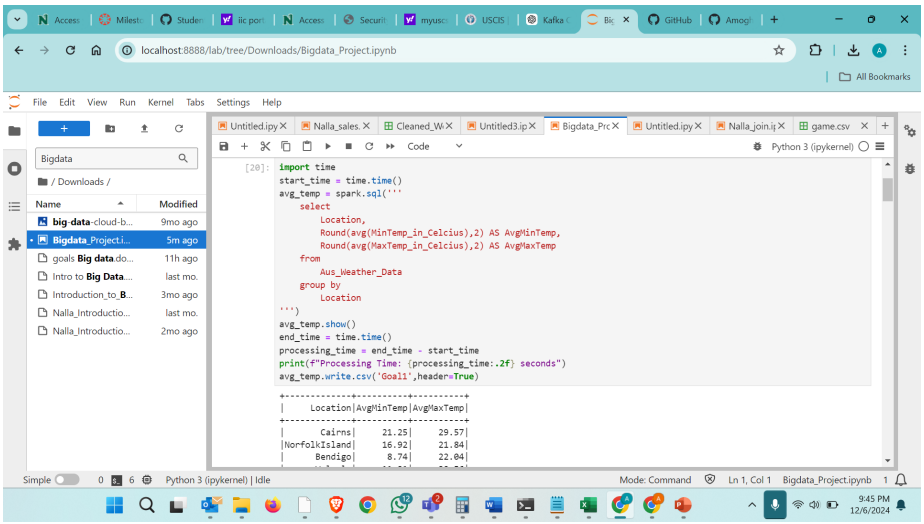

Figure 3: Goal 1

Goal 1 : Analyze average temperatures for each location (MinTemp_in_Celsius and MaxTemp_in_Celsius).

Story:- The Above bar chart and column chart shows to Compute the average MinTemp and MaxTemp for each Location to understand general temperature trends.First, Darwin city has highest average Minimum temperature with 23.31 celcius, Second location is Cairns with 21.25 celcius followed by katherine with 20.64 celcius at third and other cities has less averages .The Katheine has highest Average maximum temperature with 34.97 celcius at first, Darwin city has 32.62 celcius at second and followed by Ulluru and Carins cities.This chart summarized that cities which more average minimum temperature also has more average maximum temperature.

Goal 2: s where MinTemp or MaxTemp exceeds certain thresholds (e.g., MinTemp ¡ 0°C or MaxTemp ¿ 40°C).

story: This above 2 Filled Map shows the cities with lowest temperature and highest temperature among the data. The Tuggerano city has the lowest temperature with -8.2 celcius and Richmond has highest temperature with 47 celcius.The cities with moderate temperature has highest temperature and cities with low temperature has no increase in temperatures.

Goal 3: Find the top 5 locations with the highest average maxTemp

story: The pie chart shows the top 5 cities with highest temperature in the given dataset. The Katherine has 34.97 celcius, Darwin has 32.61 celcius, Ulluru has 30.65 celcius, followed by Cairns and Townsville with 29.57 and 29.31 Celcius.

Goal 4: Identify days and locations where there was an unusually large drop in temperature compared to the previous day. Story: The clustured bar chart shows the cities with min temp drop and max temp drop for each city. The

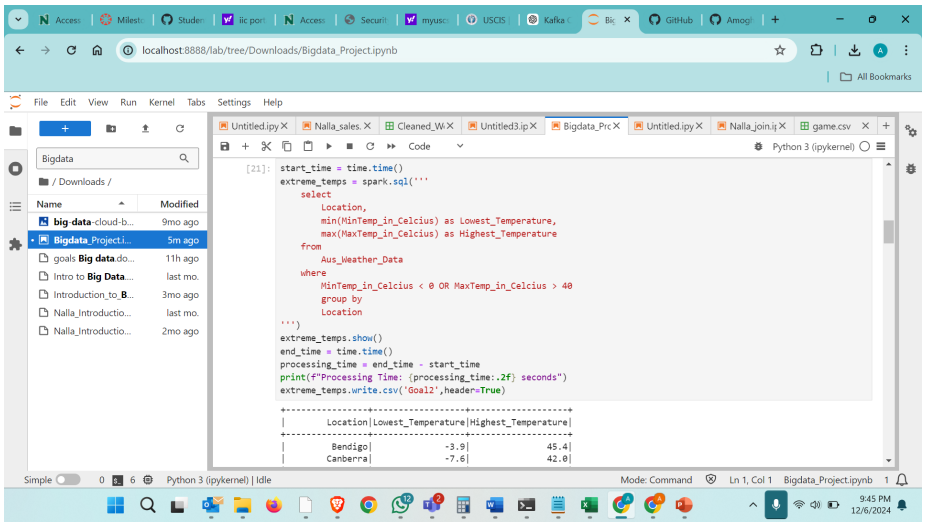

Figure 4: Goal 2

Bendigo city has minimum temperature drop with -10.8 Celcius and Cobar has maximum temperature with 25.1 Celcius. The cities with low temperature has high prediction of temperature drop compared to cities with low temperature cities

Goal 5: Analyze the relationship between temperature (MinTemp, MaxTemp), rainfall, and humidity to identify patterns of high humidity conditions. This goal would involve computing: The average humidity (Humidity) under different ranges of MinTemp and MaxTemp. Determine if higher temperatures or rainfall correlate with higher humidity levels

story:

Goal 6: Analyze the number of days with rainfall greater than 0 for each location.

story:

@

misckaggle$_{w}eather_{d}ata, title = AustraliaWeatherData, author = ArunavaKChakraborty, year = 2021, note = Availableat$

@manualapache$_{s}park, title = ApacheSpark : Lightning - FastUnifiedAnalyticsEngine, author = TheApacheSoftwareFoundation, year = 2023, note = Availableat$

@manualtableau, title = Tableau Desktop and Tableau Prep, author = Tableau Software, year = 2023, note = Available at `https://www.tableau.com/`

@articlehumidity$_{r}ainfall_{c}orrelation, author = R.SmithandJ.Doe, title = AnalyzingtheCorrelationBetweenRainfallandHumidityinAustralia, journal =$

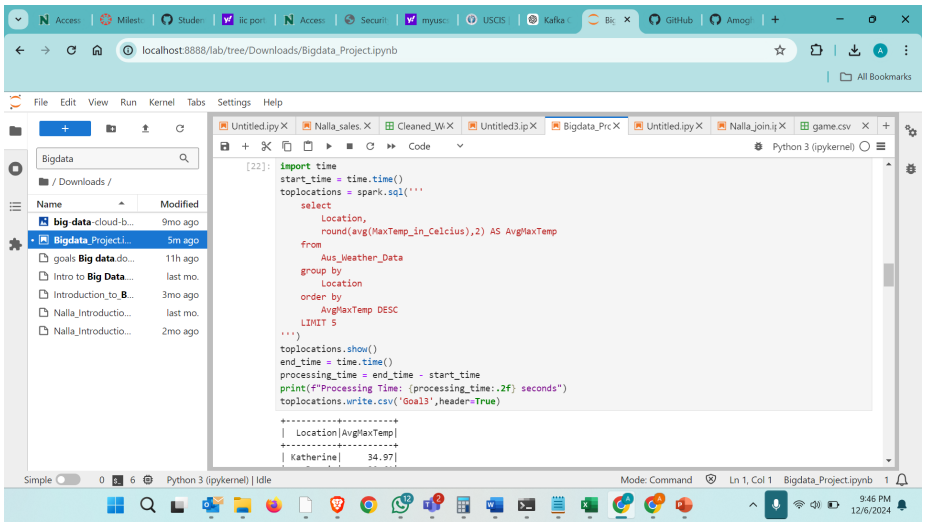

Figure 5: Goal 3

$Journal of Meteorological Studies, volume = 45, number = 3, pages = 234 - 245, year = 2020, publisher = Springer$

@book$big_data_analysis, author = A. Johnson and B. Lee, title = Big Data Analytics in Weather Forecasting, 2018, publisher = Elsevier, address = London, UK$

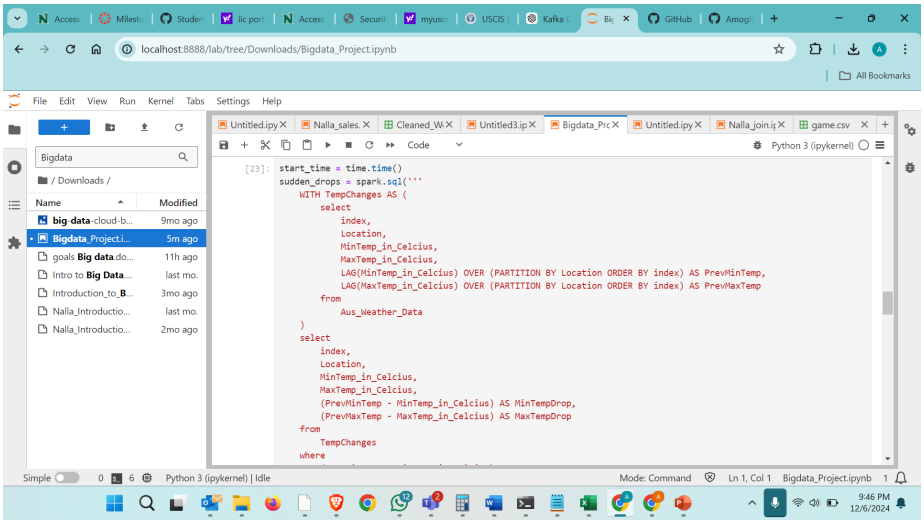

Figure 6: Goal 4

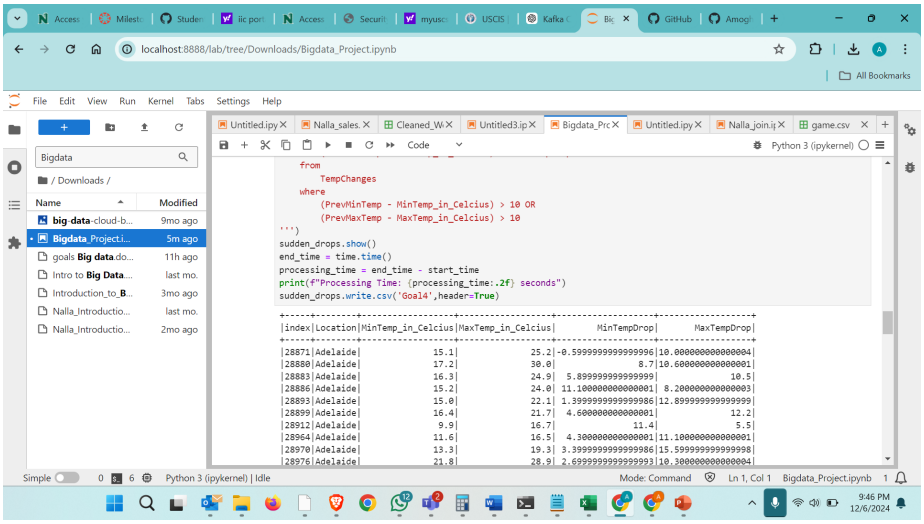

Figure 7: Goal 4 - 2

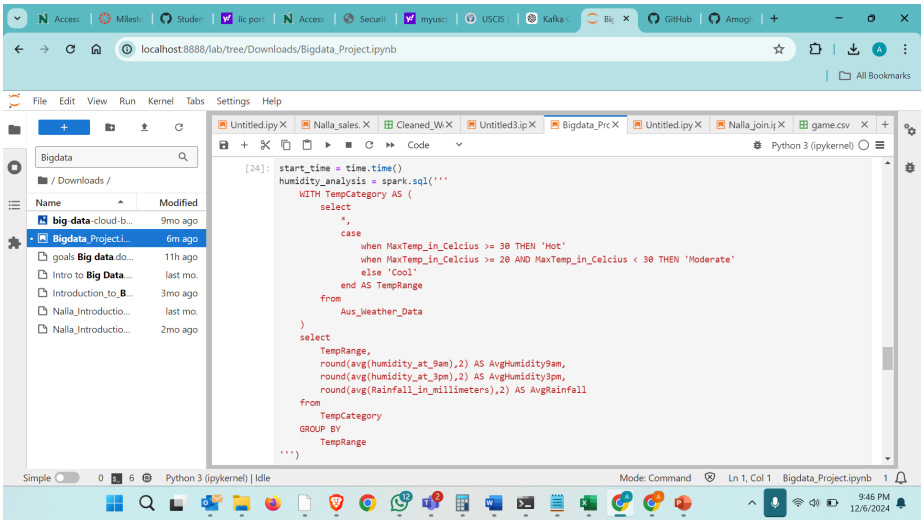

Figure 8: Goal 5

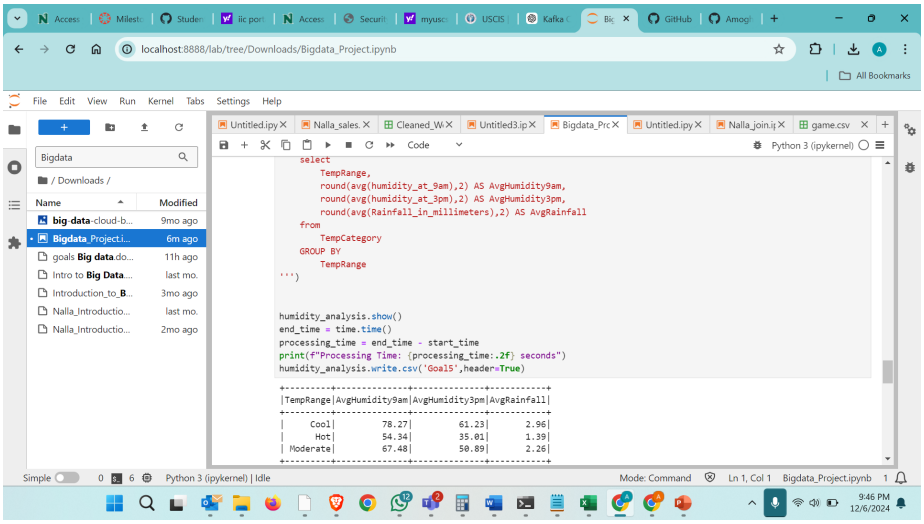

Figure 9: Goal 5-2

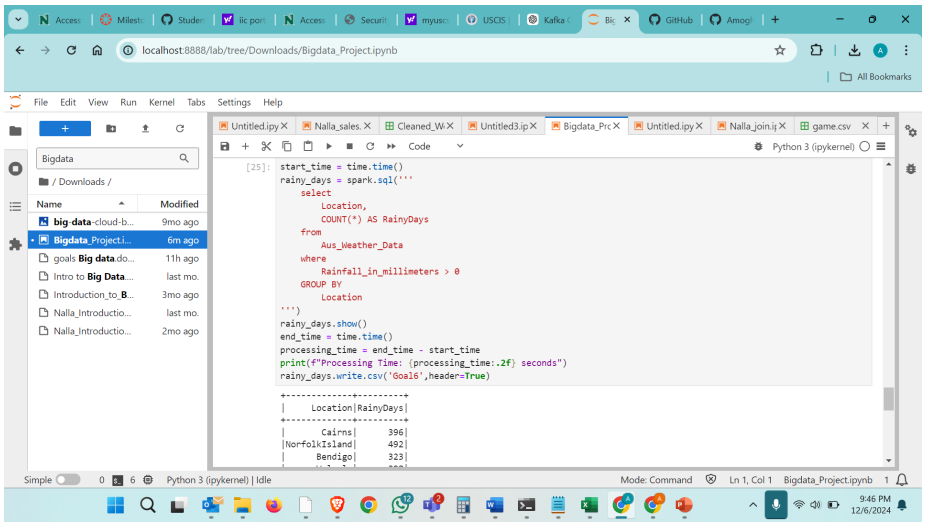

Figure 10: Goal 6

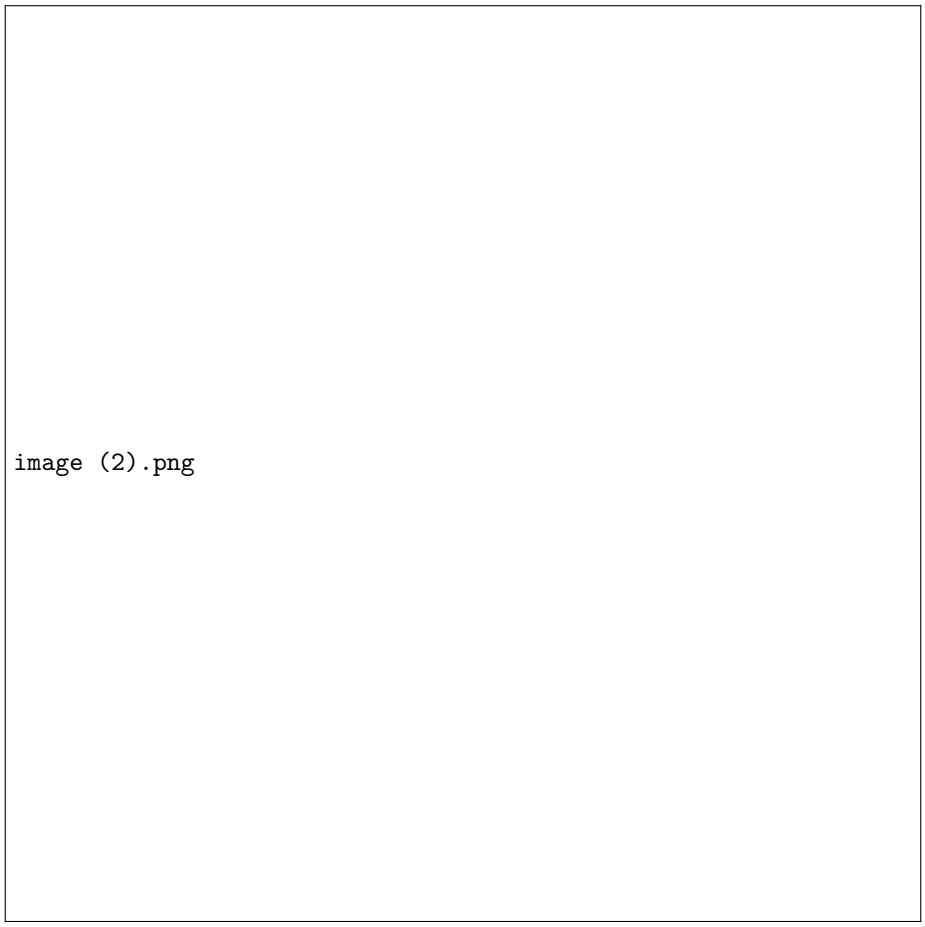

Figure 11:

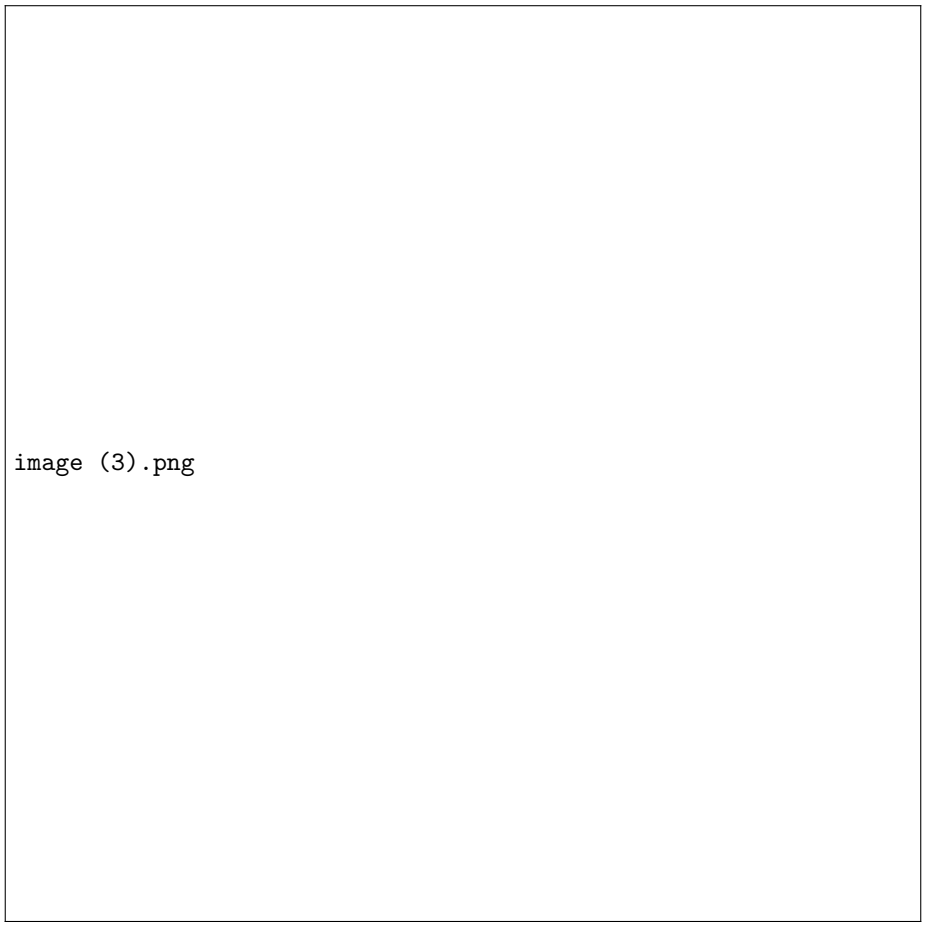

Figure 12: