# TDP Data Analyst - Assignment

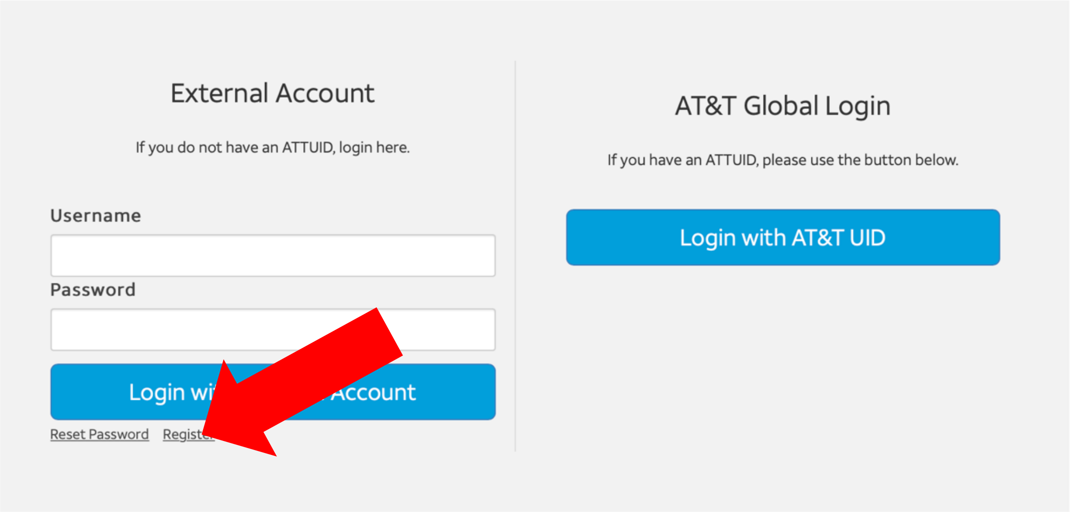
## Overview:

This assignment will assess your competencies as a data analyst including statistical analysis, data engineering, data cleanup, data science, data visualization, and presentation skills. The assignment will be divided into two parts: Part A and Part B.

In this assignment, it is your task to derive insights from the data. Part A will give you a set list of problems that you must solve. Part B will be free form experimentation. This is where you have the opportunity to show your creativity. You'll be asked to experiment with a hypothesis of your choice and provide the insights that you've derived.

## The Data:

The data will be provided to you via an emailed link that you will need to download. After clicking the link in the email, make sure to create an account and follow the instructions.



External Account

If you do not have an ATTUID, login here.

AT&T Global Login

If you have an ATTUID, please use the button below.

The data provided to you is composed of 5 different datasets, listed below:

| Dataset | Description |
|---|---|
| BayerAG Historical | Historical stock data for BayerAG |
| Dupont Historical | Historical stock data for Dupont |
| US Bee Colony data | Bee colony data for states in the United States |
| US Blueberries data | Blueberry agriculture data for the United States |
| Almond data | Almond agriculture data for the United States |

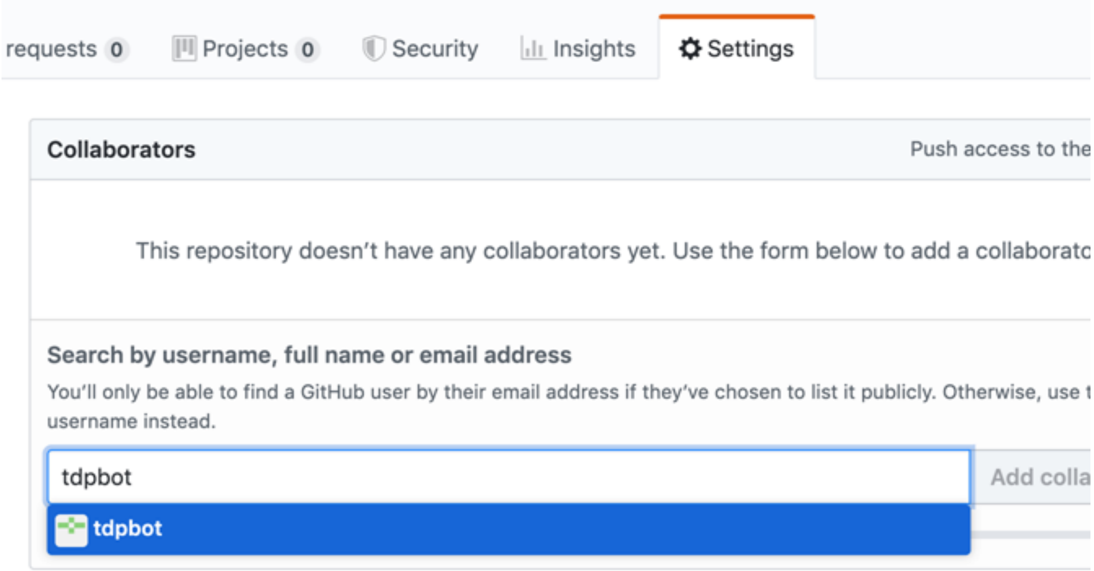
These datasets come from several different sources, so they don't all follow the same conventions. In order to use all datasets for your insights, you'll have to make decisions on how to engineer them together.

## Data Cleaning:

There are data discrepancies littered across the different datasets. There are issues such as wrong dates, missing values, and typos. It is up to you to decide how you want to handle these issues. There is no right answer for how to handle any issue, as long as you can document and explain your reasoning.

## Submission:

1. Create a **private** GitHub repository
2. Add us as a contributor:



**Collaborators**

Push access to the

This repository doesn't have any collaborators yet. Use the form below to add a collaborato

**Search by username, full name or email address**

You'll only be able to find a GitHub user by their email address if they've chosen to list it publicly. Otherwise, use t username instead.

tdpbot

tdpbot

3. Please make sure to include your **full name** and **email** in the README.

Be sure to push all your work before the deadline: **Sunday November 17th, 11:59 PM Central Time (CST)**. No submissions will be accepted after the deadline.

## Minimum Criteria:

In order to move onto the technical interview, the following minimum criteria must be met:

1. **Part A**: Complete 4 out of the 6 questions correctly
2. **Part B**: Complete Part B

# Part A

This part we will be assessing multiple data analyst skills. Try to answer as many questions as you can. Also, be ready to present your work during the interview.

**Question 1:** Calculate the correlation between adjusted closing prices for the Dupont and BayerAG stocks.

**Question 2:** In the blueberry dataset, which state had the highest increase of yield per acre for a given year.

**Question 3:** Which year had the highest standard deviation of colonies

**Question 4:** For the year 2014, visualize the histograms of Blueberry Yield Per Acre for states that had a Total Colony Loss less than 35% and states that had a Total Colony Loss more than 35%

**Question 5:** Calculate and visualize the correlation matrix for the variables below. Which two variables had the highest correlation

- Dupont Adjusted Closing Stock price
- BayerAG Adjusted Closing Stock price
- Average Colony Loss
- Average Blueberry Yield per Acre
- Almond Yield per Acre

**Question 6:** Using Dupont and BayerAG Adjusted Closing Stock price as input, forecast the number of Colonies for the states of California, North Dakota, and Texas if Dupont's stock price closed at **$90** and BayerAG at **$80**

# Part B

## Data Insights Experiment

Part B of this assignment will allow you to showcase your creativity, problem solving skills, and your competencies in data analytics. Here, we're asking you to define a hypothesis based on the provided datasets and develop a solution for your experiment. Don't worry if your hypothesis proves to be incorrect, as long as you're able to demonstrate your conclusion in an effective manner.

Here, we provide an example using a different dataset to develop a hypothesis and solution:

**Datasets:**

Dallas Murders  
Historical temperature in Dallas

**Hypothesis**:

Murder rate is correlated to temperature in Dallas, Texas.

**Conclusion**:

In the past 4 years, murders are 34% correlated to the monthly average temperature.

**Solution**:

See the Jupyter notebook attached.