

# Dino Fun World Analysis Assignment

## Technical Requirements

If you choose to work on your assignment locally, you can use the following versions:

- Python 3.11.3
- Sqlite3
- Pandas == 1.5.3
- Matplotlib == 3.7.2
- Numpy == 1.25.1

## Assignment Description

The administrators of Dino Fun World, a local amusement park, have asked you, one of their data analysts, to perform three data analysis tasks for their park. These tasks will involve understanding, analyzing, and graphing attendance data for three days of the park's operations that the park has provided for you to use. They have provided the data in the form of a database.

**Question 1:** The park's administrators would like your help understanding the different paths visitors take through the park and different rides they visit. In this mission, they have selected five (5) visitors at random whose check-in sequences they would like you to analyze. For now, they would like you to construct a distance matrix for these five visitors. The five visitors have the IDs: 165316, 1835254, 296394, 404385, and 448990.

- The distance matrix should be reported as a dictionary of dictionaries (eg. {1: {2:0, 3:0, 4:0}, 2: {1:0, 3:0, ...}, ...}).

**Question 2:** The park's administrators would like to understand the attendance dynamics at each ride (note that not all attractions are rides). They would like to see the minimum (non-zero) attendance at each ride, the maximum attendance, and the average attendance over the whole day for each ride in a parallel coordinate plot.

- For this question, display a Parallel Coordinate Plot in the notebook and print the data used to create a Parallel Coordinate Plot as a dictionary of dictionaries (eg: { 'Ride1' : {min : 1, max : 3, avg : 2 }, 'Ride2' : { min : 1, max : 3, avg : 2 } ... })

**Question 3:** In addition to a parallel coordinate plot, the administrators would like to see a scatterplot matrix depicting the minimum, maximum, and average attendance for each ride as

above.

- Print the output values of Question 2 in the Jupyter Notebook as the same data will be used for Scatterplot.

## Directions

### Accessing Ed Lessons

You will complete and submit your work through Ed Lessons. Follow the directions to correctly access the provided workspace:

1. Go to the Canvas Assignment, "**Submission: Dino Fun World Analysis Assignment**".
2. Click the "**Load Submission...in new window**" button.
3. Once in Ed Lesson, select the assignment titled "**Dino Fun World Analysis Assignment**".
4. Review the resources provided in all demonstration items.
5. When ready, click on the code challenge and start working in the notebook titled "**Assignment3.ipynb**".

### Assignment Directions

The database provided by the park administration is formatted to be readable by any SQL database library. The course staff recommends the sqlite3 library. The database contains three tables, named 'checkin', 'attractions', and 'sequences'. The database file is named 'dinofunworld.db' and is available in the '**course/data/CSE-578/dinofunworld.db**' path.

**Note:** Please note that the database file is accessible through the learner submission workspace, which requires establishing a connection with the database. For downloading the dataset and potentially working locally, refer to the overview document page.

The information contained in each of these tables is listed below:

#### **checkin:**

- The check-in data for all visitors for the day in the park. The data includes two types of check-ins: inferred and actual checkins.
- Fields: visitorID, timestamp, attraction, duration, type

#### **attraction:**

- The attractions in the park by their corresponding AttractionID, Name, Region, Category, and type. Regions are from the VAST Challenge map such as Coaster Alley, Tundra Land, etc. Categories include Thrill rides, Kiddie Rides, etc. Type is broken into Outdoor Coaster, Other Ride, Carousel, etc.
- Fields: AttractionID, Name, Region, Category, type

#### sequences:

- The check-in sequences of visitors. These sequences list the position of each visitor to the park every five minutes. If the visitor has not entered the park yet, the sequence has a value of 0 for that time interval. If the visitor is in the park, the sequence lists the attraction they have most recently checked in to until they check in to a new one or leave the park.
- Fields: visitorID, sequence

Using the data provided, perform the required analyses and create the distance matrix, parallel coordinate plot, and scatterplot matrix.


## Submission Directions for Assignment Deliverables

This assignment will be auto-graded. You must complete and submit your work through Ed Lesson's code challenge to receive credit for the course:

1. In order for your answers to be correctly registered in the system, you must place the code for your answers in the cell indicated for each question.
  - a. You should submit the assignment with the output of the code in the cell's display area. The display area should contain only your answer to the question with no extraneous information, or else the answer may not be picked up correctly.
  - b. Each cell that is going to be graded has a set of comment lines (ex: `### TEST FUNCTION: test_question1`) at the beginning of the cell. **This line is extremely important and must not be modified or removed.**
2. After completing the notebook, run each code cell individually or click "**Run All**" at the top to print the outputs.

```
[1] ### TEST FUNCTION: test_question1
    # your code here
    print('Hello world')

[1] Hello world
```



3. When you are ready to submit your completed work, click on "**Mark**" at the bottom right of the screen.

4. You will know you have successfully completed the assignment when feedback appears for each test case with a score.
5. If needed: to resubmit the assignment in Ed Lesson
  - a. Edit your work in the notebook
  - b. Run the code cells again
  - c. Click “**Mark**” at the bottom of the screen

Your submission will be reviewed by the course team and then, after the due date has passed, your score will be populated from Ed Lesson into your Canvas grade.

## Evaluation

There are three parts in the grading, and each part has one test case where the total number of points for all parts is 30. If the submission is correct, you will see "The data used for the chart is correct. The plot is a valid chart." with 10 points for each part. If your output data is correct but the graph is not, you will receive a partial score of 5. **The auto-grader first validates your output data, and if it is correct, it proceeds to evaluate the correctness of the graph.** If the submission fails, the grader will return the corresponding error messages.