National University of Computer and Emerging Sciences



Lab Exercise

Artificial Intelligence Lab

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

Department of Computer Science

FAST-NU, Lahore, Pakistan

# Exercise (30 Marks)

Attempt the tasks given below. Understanding the question is major part of these tasks.

## Create an 1000x700 dp image using NUMPY

## Global Historical Climatology Network (GHCN) Weather Data Cleaning (10 Marks)

Consider the following dataset for this exercise: **GHCN\_daily\_data.csv**

The dataset contains a sample of climate summaries from land surface stations across the globe. You are required to extract the required information only which is as follows:

* Select rows where qflag is null.
* Select rows where station id starts with ‘US’
* Select rows where observation is ‘TMIN’
* Divide the observation value by 10 to record the temperature in °C.
* The output file should only contain columns: station, date, and tmin.

Write the results to a **GHCN\_output.csv** file.

The dataset has no header row so consider the following header row while reading this dataset:

['station', 'date', 'observation', 'value', 'mflag', 'qflag', 'sflag', 'obstime']

## Twitter Dog Rating (10 Marks)

Consider the following dataset for this exercise: **twitter\_dog\_rates\_data.csv**

This dataset contains a sample of dog rating data (rates the cuteness of users' dog pictures).

You are required to create a scatter plot of date vs rating, to see the rating trend in the dataset.

To generate the scatter plot, use the following code snippet:

import matplotlib.pyplot as plt

plt.xticks(rotation=25)

plt.plot(dog\_ratings['timestamp'], dog\_ratings['rating'], 'b.', alpha=0.5)

plt.plot(dog\_ratings['timestamp'], dog\_ratings['prediction'], 'r-', linewidth=3)

plt.show()

An expected scatterplot image has been provided on Google Classroom as well to compare the results. Before we can create this scatter plot, we need to process our dataset.

* We need to create the ‘rating’ column in our dataset using the ‘text’ column. Extract the numeric rating from ‘text’ using regular expression. Check out [this link](https://pandas.pydata.org/docs/reference/api/pandas.Series.str.extract.html#pandas-series-str-extract) for some help.
* Eliminate all null rows for this ‘rating’ column.
* The ‘rating’ contains some outliers (arbitrary large values). Restrict the rating value to less than 25.
* While importing the dataset, the timestamp column is a string. Convert it to proper datetime object using Pandas. Otherwise, we won’t be able to use it properly in our scatterplot.
* We also need to create a ‘prediction’ column. Use the following code to generate this prediction column:

from scipy import stats

fit = stats.linregress(dog\_ratings['timestamp\_seconds'], dog\_ratings['rating'])

dog\_ratings['prediction'] = dog\_ratings['timestamp\_seconds']\*fit.slope + fit.intercept

* Understanding the code snippet above is out of the context of this lab. However, a brief description is as such: We need to predict the rating values of the dogs using Linear Regression which plots a straight line (using line equation) between two attributes to represent the general trend in the dataset. i.e. Increase or decrease in the dog rating over time.
* The notable point here is the ‘timestamp\_seconds’ column. This is generated using a helper function which converts the given ‘timestamp’ into number of seconds so we can feed it to out linregress function in the above code snippter. Explore the Python/Pandas function yourself.
* Visualize your image result using Matplotlib

# Submission Instructions

Always read the submission instructions carefully.

* Rename your Jupyter notebook to your roll number and download the notebook as **.ipynb** extension.
* To download the required file, go to **File->Download .ipynb**
* DO NOT **zip** or **rar** your submission **.ipynb** file.
* Submit the **.ipynb**, **GHCN\_output.csv** files on Google Classroom under the relevant assignment.
* Late submissions will not be accepted