Janet Anagli and Dr. Eick

COSC 3337 *“Data Science I”* Fall 2025

Last Updated: August 28, 8a

Task1: Exploratory Data Analysis for a 2023 Houston Weather Dataset

Second Draft

Group Task



Will it rain tomorrow in Houston?

Task1 Due: Monday, Sept. 22 end of the day (electronic Submission)

Responsible TA: Janet Anagli

**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using R or other tools)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, density plots,…
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn to distinguish expected from unexpected results in data analysis and data mining—in general, this task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, and also practical experience.

In this task focuses on exploratory data analysis for a 2003 Houston Weather dataset HW2023, which reports maximum temperature, humidity, visibility, cloudiness, wind speed and rain for the 365 days of the year 2023.

Dataset Description:

Houston\_Weather Dataset has the following attributes:

**DATE** / nominal / Each record has a date starting from 01/01/2021 to 12/31/2021

**temp\_max** / continuous / Fahrenheit / maximum temperature of that day

**humidity** / continuous / % / relative humidity at 3pm/ from 0 to

**visibility** / continuous / in meter

**cloudiness** in percent

**wind\_speed** / continuous / mile per hour / wind speed at 3pm/ from 0 to 29

**rain** / continuous / inch / Amount of rainfall of the day; no entry means ‘no rain’

The first 4 entries of the HW2023 dataset are displayed below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| date | temp\_max | humidity | visibility | cloudiness | wind\_speed | rain |
| 1/1/2023 | 81.05 | 83 | 7884.438 | 67.69231 | 11.435 | 1.37 |
| 1/2/2023 | 78.1 | 86 | 6607.971 | 95.29412 | 18.10353 | 14.09 |
| 1/3/2023 | 79.05 | 75.90625 | 8445.219 | 79.53125 | 13.54938 | 5.83 |
| 1/4/2023 | 74.08 | 47.66667 | 10000 | 28.79167 | 7.06125 |  |

Task1 Subtasks:

Apply the following exploratory data analysis techniques **using R** or other tools of your liking to the dataset:

1. Replace the blank entries for the attribute rain by ‘0’ in the HW2023 Dataset. Next,add a column rain\_t to each row which reports the amount of rainfall for the next day and a new column rain\_t? which contains Y if it rains tomorrow and N if it doesn’t.

For example for the 1/1/2023 entry the added attributes would have the values 14.09 and Y and for the 1/3/2023 entry the added attribute would have the values 0 and N. We call this newly created dataset HW2023 from now on. Next, using this data set create a dataset ZHW2023 by z-scoring attributes 2-7 of the dataset and leaving the other attributes (date, rain\_t and rain\_t?) as they are.

1. Compute the covariance matrix for each pair of the following attributes: temp\_max, humidity, visibility, wind\_speed and rain; next, compute the correlations for each of the 10 pairs of the 5 attributes. Interpret the statistical findings! **3 points**
2. Create a scatter plots for the attributes humidity&rain and for visibility& wind-speed. Interpret the two scatter plots**! 4 points**
3. Create histograms for humidity, max\_temp and rain for the examples for which rain\_t? is Y and for the examples for rain\_t is N. Compare the two histograms for each attribute. Assess if the particular attribute is useful to predict if it will rain tomorrow or not. **6 points**
4. Create box plots for the attributes visibility and wind\_speed for the examples for which rain\_t is Y and for the examples for which rain\_t is N. Compare the two boxplots you obtained for each attributes and assess the usefulness of the two attributes for predicting rain.
5. Create supervised scatter plots/supervised density plots for the following 2 pairs of attributes using the attribute rain\_t? as a class variable: humidity&wind\_speed and max\_temp and rain. Use different colors for the class variable. Interpret the obtained plots; in particular, address what can be said about the difficulty in predicting if it rains tomorrow and the distribution of the instances of the two classes in the two supervised scatter plots. **6 points**
6. Create a density plot for the following attribute pair: cloudiness&wind\_speed. Interpret the obtained plot. **3 points**
7. Fit a linear model to the dataset ZHW2023 that predicts the value of the rain\_t attribute using attributes 2-7 in the dataset as the independent variables. Report the R2 of the linear model and the coefficients of each attribute in the obtained regression function. What do the obtained coefficients tell you about the importance of each attribute for predicting if it rains tomorrow? **8 points**
8. Create 3 different decision tree models with 16 or less nodes for the HW2023 dataset using attributes 2-7 and the attribute rain\_t? as the class variable (leaf nodes count; do not submit models with more than 16 nodes!); Report the training accuracy and the testing accuracy of the submitted decision trees. Interpret the learnt decision tree. What does it tell you about the importance of the six attributes for predicting if it will rain tomorrow? **9 points**
9. Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; also report findings you found ‘unexpected’; in particular address the findings obtained related to predicting if it will rain tomorrow using attributes 2 to 7 of the HW2023 dataset. If possible, write about which attributes seem useful for this prediction taskt! Did you learn anything about the weather in Houston you did not know before? **6 points (and up to 4 extra points)**

Remark: About 35-40% of the Task1 points will be allocated to interpreting statistical findings and visualizations!

**Submission Guidelines Task1[[1]](#footnote-1)**: When you submit your task 1 for problem set 1, you should submit a compressed (zipped) folder that contains a word file that displays your graphs and your interpretations. Each interpretation should use complete sentences to describe your findings. Also in the folder, you should include all files used to complete your tasks, such as your R or python files. If you have doubts what to submit send the responsible TA an e-mail.

1. More detailed submission instructions for Task1 will be added to this specification by Sept. 17, 2025 the latest. [↑](#footnote-ref-1)