Dr. Eick

COSC 6335 *“Data Mining”* Fall 2023

Problem Set1

Task1: Exploratory Data Analysis for a Basel Weather Dataset



Task1 Due: Friday, September 23, 11:59p (electronic Submission)

**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.

The goal of this project is to perform exploratory data analysis for the Basel\_Weather Dataset (Available at: **Teams/ H\_20233\_COSC\_6335\_18523/General/Files/Data** ) which is a modification of the Weather Prediction Dataset (<https://github.com/florian-huber/weather_prediction_dataset> ). The original Weather Prediction Dataset contains weather data for 18 cities, while given dataset contains data only for the city of Basel with an additional categorical humidity class added. Basel\_Weather Dataset has the the following attributes:

DATE / nominal / Each record has a date starting from 01/01/2000 to 01/01/2010

MONTH / nominal / -- / 1 to 12

cloud\_cover / nominal / %/ 0 to 10, The fraction or percentage of the sky that is covered by clouds at a specific location and time

global\_rediation / continuous / W/m2/ The solar radiation that reaches the Earth's surface

precipitation / continuous / mm / Amount of rainfall

subshine / continuous / hours / The amount of time during which the sun is visible and unobscured by clouds or other atmospheric obstructions.

temp\_mean / continuous / celcicus / Average temperture of the day

temp\_min / continuous / celcicus / Minimum temperture of the day

temp\_max / continuous / celcicus / Maximum temperture of the day

humidity / continuous / % /

humidity\_class / ordinal/ ---/ Low, Mid, and High (derived from Humidity attribute; see below)

3 Examples in the Weather Prediction Dataset:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20000101 | 1 | 7 | 0.82 | 1.34 | 0 | -15.2 | -17 | -13.4 | 0.89 | Mid |
| 20000102 | 1 | 5 | 0.6 | 0.39 | 2.8 | -13.7 | -15 | -12.3 | 0.86 | Low |
| 20000110 | 1 | 8 | 0.25 | 0.38 | 0 | -13.3 | -15 | -11.6 | 0.97 | High |

The values of the humidity\_class attribute have been computed from the humidity attribute as follows: values till 33 percentile (0.1 to 0.86) is “Low”, till 66 percentile (0.86 to 0.96) is “Mid” and remaining (0.96 to 1) are treated as “High”. In general, we are interested to predict Attributes humidity and humidity\_class using the other attributes; that is, we like to predict the humidity\_class of the weather based on other weather properties described by attributes DATE, MONTH, cloud\_cover, global\_rediation, precipitation, subshine, temp\_mean, temp\_min, temp\_max . Other things we are interested in are finding relationships between the continous attributes in the dataset, and to understand relationship between humidity and temperature variables.

Assignment1 Tasks:

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

1. Here we will try to understand relationship among values of a variable and the categories of the target class. Compute the mean value and standard deviation of the 7 numerical attributes cloud\_cover, global\_rediation, precipitation, subshine, temp\_mean, temp\_min, temp\_max **[[1]](#footnote-1)**. Next calculate mean and standard deviation for each humidity\_class (for example mean and standard deviation for temp\_mean when humidity\_class is Low/Mid/High, repeat for all other variables). Interpret class wise variations in attribute mean values and standard deviations**. 4 points**
2. Next, we will analyze the overall relationship among the attributes. Compute the covariance matrix for each pair of the following attributes: global\_rediation, subshine, temp\_mean, and humidity (treat this attribute as a continuous attribute); next, compute the correlations for each of the 6 pairs of the 4 attributes. Interpret the obtained statistical findings! **4 points**
3. Next, we will visually understand the relationship between attributes in 2D. Create a scatter plot for the cloud\_cover and humidity attributes. Create two more scatter plots cloud\_cover and temp\_mean ; temp\_mean and humidity. Interpret the scatter plots**! 6 points**
4. Next, we will create summaries of the attribute value distribution for the three classes. Create histograms for temp\_mean, subshine, and global\_rediation attributes for Low, Mid and High humidity; interpret the obtained 9 histograms.**6 points**
5. Create box plots for the precipitation attribute for the instances of each humidity\_class —one for Low, Mid and High— and a fourth box plot for all instances in the dataset. Interpret and compare the 4 box plots **4 points**
6. Create supervised scatter plots/supervised density plots for the following 3 pairs of attributes using the humidity\_class attribute as a class variable: cloud\_cover & subshine, global\_rediation & temp\_mean and precipitation & temp\_mean. Use different colors for the class variable. Interpret the obtained plots; in particular address what can be said about the difficulty in predicting the humidity\_class using the analyzed weather attributes. **6 points**
7. Create a new dataset ZBasel\_Weather from the Basel\_Weather dataset by transforming the 7 attributes cloud\_cover, global\_rediation, precipitation, subshine, temp\_mean, temp\_min, temp\_max into z-scores. Fit a linear model that predicts the Humidity attribute using the 7 z-scored, continuous attributes 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 the humidity level? **8 points**
8. Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; in particular address the findings obtained related to predicting humidity using the other continuous weather attributes). **5 points (and up to 3 extra points)**

Remark: About 30-40% of the Task1 points will be allocated to interpreting statistical findings and visualizations!

**Submission Guidelines Task1**: Task 1 will be a Kritik submission. For each task you need to create:

* A separate code file with proper commenting
* A report showcasing your results and discussions

**Your name should not appear in these documents.** Rubrics for each task is given below:

**Rubrics**:

**Q1** Here we will try to understand relationship among values of a variable and the categories of the target class. Compute the mean value and standard deviation of the 7 numerical attributes cloud\_cover, global\_rediation, precipitation, subshine, temp\_mean, temp\_min, temp\_max **[[2]](#footnote-2)**. Next calculate mean and standard deviation for each humidity\_class (for example mean and standard deviation for temp\_mean when humidity\_class is Low/Mid/High, repeat for all other variables). Interpret class wise variations in attribute mean values and standard deviations**. 4 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Presentation of the Results | No mean values and standard deviations from seven attributes are presented. | All mean values and standard deviations from seven attributes are presented in a random manner with no clear story about data | All mean values and standard deviations from seven attributes are presented in a structured manner, but do not tell any clear story about data | All mean values and standard deviations from seven attributes are presented in a manner to tell a clear story about data (How values influence classes from different attributes) | 0.5 |
| Correctness of the Results | Results are completely incorrect. | Only 33% results are correct. | Only 66% results are correct. | All results are completely correct. | 1.5 |
| Discussion Writing Quality | No discussion is presented | The discussion is poorly written with lots of mistakes and contains many redundant comments | The discussion quality is moderate with some mistakes and contains a few redundant comments | The discussion is very well written with no redundancy | 0.5 |
| Discussion  Quality | No discussion is presented | The discussion misses a lot of the important information and /or contains many wrong information | The discussion misses a some of the important information and/or contains some wrong information | The discussion is fully logical and contains all important information | 1.5 |

**Q2**: Next, we will analyze the overall relationship among the attributes. Compute the covariance matrix for each pair of the following attributes: global\_rediation, subshine, temp\_mean, and humidity (treat this attribute as a continuous attribute); next, compute the correlations for each of the 6 pairs of the 4 attributes. Interpret the obtained statistical findings! **4 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Covariance  Matrix Presentation | All covariance matrices have flaws or all are missing | Covariance matrices has a lot of flaws such as most data are missing or wrong and presented in a manner that does not show clear relationship among the variables | Covariance matrices has some flaws such as a few missing or wrong data and presented in a manner that does not show clear relationship among the variables | All covariance matrices are correct and presented in a manner that shows clear relationship among the variables | 0.25 |
| Correctness of the Covariance  Matrix | Results are completely incorrect. | Only 33% results are correct. | Only 66% results are correct. | All results are completely correct. | 0.75 |
| Covariance  Matrix Interpretation  Quality | No Interpretation is presented | The Interpretations are poorly written with lots of mistakes and contains many redundant comments | The Interpretations are moderate with some mistakes and contains a few redundant comments | The Interpretations are very well written with no redundancy | 0.25 |
| Covariance  Matrix Interpretation  Correctness | All covariance matrices have missing information or cannot represent relationships among the variables | Most of the covariance (>50%) matrices have missing information or cannot represent relationships among the variables | Some of the covariance (<=25%) matrices have serious missing information or cannot represent relationships among the variables | All covariance matrices are correct, convenient and can capture the relationships among the variables well | 0.75 |
| Correlation  Matrix  Presentation | All correlation matrices have flaws or all are missing | Correlation matrices has a lot of flaws such as most data are missing or wrong and presented in a manner that does not show clear relationship among the variables | Correlation matrices has some flaws such as a few missing or wrong data and presented in a manner that does not show clear relationship among the variables | All correlation matrices are correct and presented in a manner that shows clear relationship among the variables | 0.25 |
| Correctness of the Correlation  Matrix | Results are completely incorrect. | Only 33% results are correct. | Only 66% results are correct. | All results are completely correct. | 0.75 |
| Correlation  Matrix Interpretation Quality | No Interpretation is presented | The Interpretations are poorly written with lots of mistakes and contains many redundant comments | The Interpretations are moderate with some mistakes and contains a few redundant comments | The Interpretations are very well written with no redundancy | 0.25 |
| Correlation Matrix Interpretation  Correctness | All Correlation matrices have missing/wrong information or cannot represent relationships among the variables | Most of the Correlation (>50%) matrices have missing/wrong information or cannot represent relationships among the variables | Some of the Correlation (<=25%) matrices have serious missing/wrong information or cannot represent relationships among the variables | All Correlation matrices are correct, convenient and can capture the relationships among the variables well | 0.75 |

**Q3**: Next, we will visually understand the relationship between attributes in 2D. Create a scatter plot for the cloud\_cover and humidity attributes. Create two more scatter plots cloud\_cover and temp\_mean ; temp\_mean and humidity. Interpret the scatter plots**! 6 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Scatter Plot Quality | No or wrong scatter plots are given | Quality of the scatter plots are very poor and lots of missing details within the visualization | Quality of the scatter plots are modest and number of missing details within the visualization | Quality of the scatter plots are very high and the scatter plot perfectly visualize the relationship among the variables | 2 |
| Scatter Plot Presentation | No scatter plots are presented. | All scatter plots are presented in a random manner with no clear story about data | All scatter plots are presented in a structured manner, but do not tell any clear story about data | All scatter plots are presented in a manner to tell a clear story about data (e.g. represents inter-relationship) | 1 |
| Scatter Plot Interpretation  Quality | No Interpretation is presented | The Interpretations are poorly written with lots of mistakes and contains many redundant comments | The Interpretations are moderate with some mistakes and contains a few redundant comments | The Interpretations are very well written with no redundancy | 1 |
| Scatter Plot Interpretation  Correctness | No discussion is presented | The discussion contains many redundant/wrong comments and/or also misses a lot of important information | The discussion contains some redundant/wrong comments and/or also misses some important information | The results are discussed in a **concise, clear** manner to show relationship among attributes | 2 |

**Q4:** Next, we will create summaries of the attribute value distribution for the three classes. Create histograms for temp\_mean, subshine, and global\_rediation attributes for Low, Mid and High humidity; interpret the obtained 9 histograms.**6 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Histogram Quality | No or wrong histograms are given | Quality of the histograms are very poor and lots of missing details within the visualization | Quality of the histograms are modest and number of missing details within the visualization | Quality of the histograms are very high and the scatter plot perfectly visualize the relationship among the variables | 2 |
| Histogram Presentation | No histograms are presented. | All histograms are presented in a random manner with no clear story about data | All histograms are presented in a structured manner, but do not tell any clear story about data | All histograms are presented in a manner to tell a clear story about data (e.g. represents inter-relationship) | 1 |
| Histogram Interpretation  Quality | No Interpretation is presented | The Interpretations are poorly written with lots of mistakes and contains many redundant comments | The Interpretations are moderate with some mistakes and contains a few redundant comments | The Interpretations are very well written with no redundancy | 1 |
| Histogram Interpretation  Correctness | No discussion is presented | The discussion contains many redundant/wrong comments and/or also misses a lot of important information | The discussion contains some redundant/wrong comments and/or also misses some important information | The results are discussed in a **concise, clear** manner to show relationship among attributes (How values influence classes from different attributes) | 2 |

**Q5:** Create box plots for the precipitation attribute for the instances of each humidity\_class —one for Low, Mid and High— and a fourth box plot for all instances in the dataset. Interpret and compare the 4 box plots **4 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Box Plots Quality | No or wrong boxplots are given | Quality of the boxplots are very poor and lots of missing details within the visualization | Quality of the boxplots are modest and number of missing details within the visualization | Quality of the boxplots are very high and the scatter plot perfectly visualize the relationship among the variables | 1.5 |
| Box Plots Presentation | No boxplots are presented. | All boxplots are presented in a random manner with no clear story about data | All boxplots are presented in a structured manner, but do not tell any clear story about data | All boxplots are presented in a manner to tell a clear story about data (e.g. represents inter-relationship) | 0.5 |
| Box Plots Interpretation  Quality | No Interpretation is presented | The Interpretations are poorly written with lots of mistakes and contains many redundant comments | The Interpretations are moderate with some mistakes and contains a few redundant comments | The Interpretations are very well written with no redundancy | 0.5 |
| Box Plots Interpretation  Correctness | No discussion is presented | The discussion contains many redundant/wrong comments and/or also misses a lot of important information | The discussion contains some redundant/wrong comments and/or also misses some important information | The results are discussed in a **concise, clear** manner to show relationship among attributes | 1.5 |

**Q6:** Create supervised scatter plots/supervised density plots for the following 3 pairs of attributes using the humidity\_class attribute as a class variable: cloud\_cover & subshine, global\_rediation & temp\_mean and precipitation & temp\_mean. Use different colors for the class variable. Interpret the obtained plots; in particular address what can be said about the difficulty in predicting the humidity\_class using the analyzed weather attributes. **6 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Scatter Plot Quality | No or wrong scatter plots are given | Quality of the scatter plots are very poor and lots of missing details within the visualization | Quality of the scatter plots are modest and number of missing details within the visualization | Quality of the scatter plots are very high and the scatter plot perfectly visualize the relationship among the variables | 2 |
| Scatter Plot Presentation | No scatter plots are presented. | All scatter plots are presented in a random manner with no clear story about data | All scatter plots are presented in a structured manner, but do not tell any clear story about data | All scatter plots are presented in a manner to tell a clear story about data (e.g. represents inter-relationship) | 1 |
| Scatter Plot Interpretation  Quality | No Interpretation is presented | The Interpretations are poorly written with lots of mistakes and contains many redundant comments | The Interpretations are moderate with some mistakes and contains a few redundant comments | The Interpretations are very well written with no redundancy | 1 |
| Scatter Plot Interpretation  Correctness | No discussion is presented | The discussion contains many redundant/wrong comments and/or also misses a lot of important information | The discussion contains some redundant/wrong comments and/or also misses some important information | The results are discussed in a **concise, clear** manner to show relationship among attributes | 2 |

**Q7:** Create a new dataset ZBasel\_Weather from the Basel\_Weather dataset by transforming the 7 attributes cloud\_cover, global\_rediation, precipitation, subshine, temp\_mean, temp\_min, temp\_max into z-scores. Fit a linear model that predicts the Humidity attribute using the 7 z-scored, continuous attributes 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 the humidity level? **8 points**

Linear model rubric

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| LM-correctness  (Code+Report) | The model is completely wrong | There are some major mistakes during the model creation such as failure of z-score conversion or missing attributes | There are some minor mistakes during the model creation such as a few missing attributes | Model creation is completely right | 3 |
| LM-Presentation | All coefficients are missing or reported coefficients and R\*\*2 are completely wrong | Most coefficients are missing or reported coefficients and R\*\*2 contain a lot of error | Some coefficients are missing or reported R\*\*2 contain some minor errors or R\*\*2 is incorrect or missing | All coefficients are present and reported coefficients and R\*\*2 are correct | 1.5 |
| LM-Interpretation  Quality | No Interpretation is presented | The Interpretations are poorly written with lots of mistakes and contains many redundant comments | The Interpretations are moderate with some mistakes and contains a few redundant comments | The Interpretations are very well written with no redundancy | 1.5 |
| LM-Interpretation  Correctness | The interpretation of the obtained coefficients to assesses attribute usefulness to estimate humidity attribute is totally wrong or missing | The interpretation of the obtained coefficients to assesses attribute usefulness to estimate humidity attribute has some major flaws or is incomplete | The interpretation of the obtained coefficients to assesses attribute usefulness to estimate humidity attribute is mostly convincing | The interpretation of the obtained coefficients to assesses attribute usefulness to estimate humidity attribute is convincing | 2 |

**Q8:** Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; in particular address the findings obtained related to predicting humidity using the other continuous weather attributes). **5 points (and up to 3 extra points)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Conclusion Writing  Quality | No conclusion | The conclusion is very poorly written and very hard to understand clearly what the author tried to explain | The conclusion is moderately written and there are some minor details that are hard to understand | Every sentence written in the conclusion is very clear and has a strong purpose | 2 |
| Conclusion  Information | No conclusion | The conclusion misses most of the important points | The conclusion misses some of the important points | The conclusion contains all the important points | 3 |
| Important Points(extra) | There are no additional insights that makes the conclusion very different than other conclusions | There are some additional insights | There are many additional insights that makes the conclusion very different than other conclusions | The conclusion provides very different and extra-ordinary insights about the data, and missed by all other conclusions | 3 |

1. This is more a verification of that you have the correct dataset! [↑](#footnote-ref-1)
2. This is more a verification of that you have the correct dataset! [↑](#footnote-ref-2)