DATA ANALYSIS PROJECT

Unit Chair: Dr Musa Mammadov
Submission Date: 5:00PM Friday of Week 6

Table of Contents

Section 1: Introduction and Data Description **2**

Section 2: Exploratory Data Analysis and Results **4**

1. Data Cleansing **4**

a. Find and correct errors in categorical variables 4

b. Count the total number of missing values for each column 5

c. Drop columns with more than 10% missing values 6

d. Drop rows with more than 20% missing values 7

e. Calculate z-score for numerical variables and replace outliers 7

f. Replace all missing values by implementing appropriate algorithms 7

2. Main Structure and Key Variables **7**

3. Data Patterns **7**

4. Assumptions **7**

5. Visualization **7**

Section 3: Conclusion **8**

Section 4: References **9**

Dataset Name: Rain in Australia

Group Name: Mon-13 (FANH) **On Campus/Cloud**: On Campus

STUDENT ID	STUDENT FULL NAME	Individual contribution*
218401269	ALEXANDER PAK YU LAI	
218241616	HARRY WILLIAM LODGE	
218459058	VIET NAM NGUYEN	
218271795	JARROD KENG YEN YONG	

* 5 - Contributed significantly, attended all meetings
4 - Partial contribution, attended all meetings
3 - Partial contribution, attended few meetings
2 - No contribution, attended few meetings
1 - No contribution, did not attended any meetings

Section 1: Introduction and Data Description

The Australian weather broadcast or the appropriate agencies comprehends the tomorrow weather, specifically whether it will be raining or not, to deliver updated weather news to citizens or propose proper plans for social activities respectively. Therefore, the dataset of day-to-day weather was collected across various stations over a period of 10 years. The observations extracted from the Australian weather dataset aim to return prediction of tomorrow rain by producing the possibilities of it or two binary labels (Yes and No, 1 and 0 in that order).

The “Rain in Australia” dataset consists of 24 variables in total. Among these variables, there are two variable types, categorical variable and numerical variable.

For the categorical variable, five variables are normal, including “Date”, “Location”, “WindGustDir”, “WindDir9am” and “WindDir3pm” while the remaining two binary variables are “RainToday” and “RainTomorrow”, especially “RainTomorrow” is a target variable. Also, these categorical variables are defined with nominal type. In addition, the “WindDir9am” and “WindGustDir” variables experience the biggest percentages of missing value with around 7 percent (Young 2017). Furthermore, the variable having the highest cardinality, which means that a variable has the largest number of labels, is “Date” with 3436 labels and the second is “Location” with 49.

For the numerical variable, there are seventeen variables, which refer to continuous type, namely “MinTemp”, “MaxTemp”, “Rainfall”, “Evaporation”, “Sunshine”, “WindGustSpeed”, “WindSpeed9am”, “WindSpeed3pm”, “Humidity9am”, “Humidity3pm”, “Pressure9am”, “Pressure3pm”, “Cloud9am”, “Cloud3pm”, “Temp9am”, “Temp3pm” and “RISK_MM”. According the figures in the table indicated the dataset provided by Young (2017), the variables witnessing the most missing values are “Evaporation” and “Sunshine” with roughly 43 and 48 percent respectively, followed by “Cloud9am” and “Cloud3pm” with nearly 38 and 40 percent in that order. Also, the “RISK_MM” variable is in consideration for dropping out of the dataset. According to Young (2017), this variable should be eliminated if it aims to train a classification model instead of regression one or the “RainTomorrow” variable is considered as a target because including this variable which indicates the further information of rain impacts negatively to the predicted values of the trained model or a lower accuracy score.

The observation about the “Rain in Australia” is that the method of training model to return a high accuracy score can be a regression algorithm instead of classification, but the null accuracy, where the accuracy is gained by preferring the most frequent value, should be compared with it to guarantee that the applied model results in the higher accuracy score.

Another observation is that many pairs of variables would have strong or high correlation in the positive trend since they are intuitively correlated in terms of related fields of rain prediction.

Before the dataset is analyzed thoroughly and its comprehensive patterns are extracted, the dataset ensures to be evaluated carefully and cleaned to get rid of missing values or incorrect data by finding the frequency of null values for each variable, calculating z scores or exploring the inner problem of categorical and numerical variables. After finishing the process of data cleansing, the cleaned dataset should be tested, followed by discovering patterns which define relationships between variables due to their correlation coefficient. Eventually, the comprehensive analysis is finalized to sum up as well as conclude findings. Throughout the exploratory data analysis, the appropriate visualizations of data evaluation or results are illustrated.

Section 2: Exploratory Data Analysis and Results

1. Data Cleansing

To achieve a machine learning model with better performance and higher test accuracy score, the “Rain in Australia” dataset should be clean, which removes all missing values, known as null or nan, fixing incorrect data in terms of grammar or format and finding outliers. Therefore, the process of data cleansing is split up into six steps.

a. Find and correct errors in categorical variables

The first stage of cleaning the dataset is to engineer errors in categorical variables, known as nominal columns. There are eight of them in total, “Date”, “Location”, “WindGustDir”, “WindDir9am”, “WinDir3pm”, “RainToday” and “RainTomorrow”, but the important variable scrutinized thoroughly during training and testing process is “RainTomorrow” since it is used as target to produce accuracy scores. Starting finding invalid date format in “Date” is the first step in eliminating errors in the categorical variables and the following function supported by the “datetime” library is applied on this variable to assess every single element.

```
# Define a function to check if the date is valid
def check_valid_format(date):
    date_format = '%Y-%m-%d'
    try:
        datetime.datetime.strptime(date, date_format)
        return True
    except ValueError:
        return False
```

As a result, the “Date” column witnesses no incorrect value in regard to its standard format, e.g. “2008-12-01”.

Focusing on the “Location” variable is the next categorical variable. To evaluate this column, all unique locations should be explored, and one small flaw on the dataset might be fixed. The small flaw is missing proper whitespace in some location names which contains two words. Therefore, the solution is simply that the whitespaces are placed appropriately in two-word locations. The lines of code below illustrates this approach.

```
# There are some locations needed to be fixed
# Insert a space before capitcal letter if the location
# name has 2 words
fixed_location = []
for location in unique_location:
    fixed_name = re.sub(r"(\w)([A-Z])", r"\1 \2", location)
    fixed_location.append(fixed_name)
```

However, this flaw in the “Location” column has insignificant impact on the dataset because the uniqueness of locations are still preserved.

The next three categorical variables whose names are “WindGustDir”, “WindDir9am” and “WinDir3pm” are defined with sixteen compass directions including four caridnal directions, four intercardinal directions and eight secondary intercardinal directions. The fortunate results after processing them through some lines of code shows no incorrect directions in terms of format and the following figure indicates the typical checking.

```
# In order to reduce iteration time, find all unique WindDir9am
unique_WindDir9am = list(set(data[nominal_columns[3]]))
unique_WindDir9am.remove(np.nan)
print("The direction:", ", ".join(unique_WindDir9am))
```

The two remaining variables, “RainToday” and “RainTomorrow”, are represented by two unique elements, “Yes” and “No” which are two binary values and can be considered as 1 and 0 respectively. The positive results gained after these variables experience the validation process are no invalid data.

b. Count the total number of missing values for each column

After the nominal columns finish being corrected, the total number of nan or null values of each column should be found, which becomes a base for the following steps. The below figure illustrates two functions which count the missing values and find their corresponding indices in columns.

```
def count_nan(column, data):
    count = int(data[column].isna().sum())
    return count

def print_index(count, column):
    if count == 0:
        print(f"There are no nan values in {column}.")
    else:
        print(f"The number of nan value in {column}: {count}")
```

The following table gives information on the total number of missing values for each variable.

Variables	Total Missing Values
Date	0
Location	0
MinTemp	637
MaxTemp	322
Rainfall	1406
Evaporation	60843
Sunshine	67816
WindGustDir	9330
WindGustSpeed	9270
WindDir9am	10013
WindDir3pm	3778
WindSpeed9am	1348
WindSpeed3pm	2630
Humidity9am	1774
Humidity3pm	3610
Pressure9am	14014
Pressure3pm	13981
Cloud9am	53657
Cloud3pm	57094
Temp9am	904
Temp3pm	2726
RainToday	1406
RISK_MM	0
RainTomorrow	0

c. Drop columns with more than 10% missing values

Exploiting the result of the total number of missing values in the previous assists in calculating the variables having more than 10 percent nan or null data. Therefore, the percentages of missing values in these variables are easily computed by applying the lines of code in the following figure.

```
# Calculate percent of missing value for each column
percent_nan_dict = dict()
for key, value in missing_value_dict.items():
    percent_nan_dict[key] = (value / data.shape[0]) * 100
```

Consequently, there are four variables holding the percentage of missing values exceeding 10, “Evaporation”, “Sunshine”, “Cloud9am” and “Cloud3pm”, corresponding 44, 48, 48 and 40 percent, which encourages them to be dropped out of dataset by one function of pandas library, “pandas.DataFrame.drop”.

d. Drop rows with more than 20% missing values

Removing the rows whose percentages of missing values are greater than 20 have the same strategy with the previous step. However, one more computation which leads to percentage calculation is counting the total number of missing values for each row by adding the following extra lines of code to figure it out.

```
# Count the total number of missing values for each row
for i in range(0, data.shape[0]):
    total_missing_value = 0
    is_nan_list = np.array(data.iloc[i:i+1,:].isnull())[0]
    total_missing_value = np.count_nonzero(is_nan_list)
    missing_values_dict[i] = total_missing_value
```

The same process of computing the percentage of missing values with the prior step. However, the opposite results are achieved because of no rows with more than 20 percent nan or null data.

e. Calculate z-score for numerical variables and replace outliers

Only numerical variables, known as continuous columns, are able to be applied to the z-score formula to find outliers, and these outliers should be substituted by the mean of appropriate columns. The following figure indicates the computation of z-scores for the entire dataset.

```
def find_outlier_index(col, data):
    z_score = (data[col]-data[col].mean()) / data[col].std(ddof=0)
```

Addressing all outliers assists the process of replacing them with possible proper values which are the average of the corresponding columns.Removing all outliers is the preparation for filling missing values existing in the entire dataset. The function in the below figure gives information on fixing outliers.

```
def replace_outlier(col, data):
    # Fix outlier by replacing it with avarage value of the column
    outlier_indices = find_outlier_index(col, data)

    mean = data[col].mean()
    for index in outlier_indices:
        data.loc[index, col] = mean
```

f. Replace all missing values by implementing appropriate algorithms

To engineer all nan or null data with the possibly highest accuracy, different data types should be applied to replace them by the most proper algorithm. Therefore, filling missing values in numerical variables utilizes the Linear Regression algorithm, “RainToday” uses Naive Bayes”, and “WindGustDir”, “WindDir9am” and “WindDir3pm” simply takes random algorithm provided by “numpy” library.

The reasons why choosing Linear Regression is for numerical variables are that they are continuous and each of these variable arms themselves with independence, which encourages to figure out their relationships by exploiting the linear equation which provides one coefficient and one intercept. Based on the coefficient and intercept, the final purpose is to form a linear equation presenting the relationship between two variables, which are a column containing missing values and the selected stable one which is “RISK_MM”. However, to fit the model successfully, all null or nan should be substituted by 0. The following figure illustrates the process of training the model and producing the linear equation.

```
# # Fit the model
lr = linear_model.LinearRegression()
lr.fit(x,y)

# Gain the model parameters
coef = lr.coef_
intercept = lr.intercept_
```

Filling the missing values in “WindGustDir”, “WindDir9am” and “WindDir3pm” is simplified by the random method of choosing a random direction in sixteen compass directions. The two lines of code in the following figure.

```
random_number = random.randint(0, len(unique_direction)-1)
random_direction = unique_direction[random_number]
```

The “RainToday” variable consists of only two unique values, “Yes” and “No” experiences through Gaussian Classifier which is based on Naive Bayes Theorem to replace all missing values with one of two choices. The first reason why the last variable should be trained by this algorithm is that it is quick, highly accurate and reliable regardless of the size of the dataset, but not simple. Another one is that it emphasizes the independence of each variable and it has high performance on categorically-typed variables. Therefore, after all selected data are encoded to numbers corresponding to labels, they are fitted into Gaussian Classifier to train the model which supports finding the most appropriate values for the missing values. The following lines of code shows the implementation of this classifier technique.

```
#Create a Gaussian Classifier
model = GaussianNB()

# Train the model using the training sets
model.fit(features,encoded_column)
```

2. Main Structure and Key Variables

a. Information used in analysis

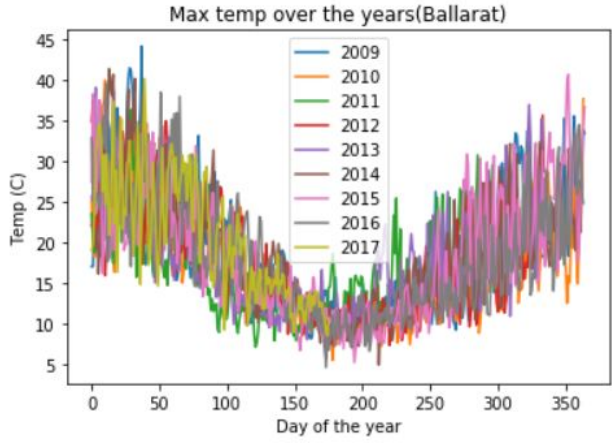
```
AusWeather.columns #looking into the data that has been divided into groups
```

```
Index(['Date', 'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'WindGustDir',
       'WindGustSpeed', 'WindDir9am', 'WindDir3pm', 'WindSpeed9am',
       'WindSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am',
       'Pressure3pm', 'Temp9am', 'Temp3pm', 'RainToday', 'RISK_MM',
       'RainTomorrow'],
      dtype='object')
```

Above is the list of the different values/columns that we could use when analysing the information. It was our job to look at the list of different variables that we could use to help us analyse patterns and trends in the information.

B. Date and location

The main use for these two variables was to help us know the differences between each row of information. Of the 14000+ entries that we had to look into, the combination of the two variables allow us to break all entries into unique rows. The values could also be broken up when looking into each individual location like when we used the location and date to help us know a specific area's max temp as posted throughout the year. By separating the location and then separating by time, we could get area based analysis that we could compare against each year's record as shown in our analysis below.



C. Min temp and Max temp

The max and min temp are used to show us the maximum temperature and minimum temperature that was recorded on that day. We used these temperatures when we wanted to identify if there was any correlation to temperature rising and falling and any of the other variables.

D. Pressure,Humidity and Temp for 9am and 3pm

Perhaps one of the most useful sets of variables that we could use so that we could observe the amount difference between the morning and afternoon. As we originally didn't find much information in the section where there was a rise in temperature during the day. We used pressure and humidity as the key values we looked into for the analysis and helped us reach a usable conclusion with our studies.

E. Rain today and tomorrow

These were the key values that we used to make sure our assumptions were correct. The aim of these variables were to help us identify if there was rain on the following day. By using that information we could predict based on the values given to us if there would be "rain today" on the following day without the need for additional analysis.

G. The other values

The rest of the values, while useful, weren't used too much in our calculations and meant that we didn't have a way to use them during the analysis, other than the generic analysis that we'd used during testing that we had used so that we could make sure all our calculations for both data cleaning and predicting are correct.

3. Data Patterns

Start from here

4. Assumptions

Start from here

5. Visualization

Start from here

Section 3: Conclusion

Section 4: References

Bureau of Meteorology n.d., *Note to accompany Daily Weather Observations*, Australian Government, retrieved 18 April 2020, <http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml>.

Young, J 2017, *Rain in Australia*, Kaggle, retrieved 18 April 2020, <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>.