R Champions

Omar Ahmouda, Leona Bell, Helen Chu, Hoai Do, Delaney Smith

BANA 4080 - Homework 1
EDA Australia Weather Data

Goal

- The purpose of studying this dataset is to explore different variables related to weather in Australia to determine whether it will rain the next day (RainTomorrow)

- Observations from this dataset is provided by the Australian Government's Bureau of Meteorology. For consistency, we named this dataset "weather".

- The first step to analyze this dataset is read it into R with read.csv().

```
weather <- read.csv("weatherAUS.csv", header = T, stringsAsFactors = T)
```

Our Variables

```
> dim(weather)
[1] 87360    23
```

```
 [1] "Date"          "Location"      "MinTemp"
 [4] "MaxTemp"       "Rainfall"      "Evaporation"
 [7] "Sunshine"      "WindGustDir"   "WindGustSpeed"
[10] "WindDir9am"    "WindDir3pm"    "WindSpeed9am"
[13] "WindSpeed3pm"  "Humidity9am"   "Humidity3pm"
[16] "Pressure9am"   "Pressure3pm"   "Cloud9am"
[19] "Cloud3pm"      "Temp9am"       "Temp3pm"
[22] "RainToday"     "RainTomorrow"
```

The function dim() was used to see the number of observations and variables we have. Names() was used to see what specific variables we are working with in this dataset (listed above). This dataset has 23 variables and 87,360 observations.

Variable Descriptions

Daily Australia Weather Dataset

11/1/2007 - 6/25/2017

Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation
1/1/2009	Cobar	17.9	35.2	0	12
1/2/2009	Cobar	18.4	28.9	0	14.8
1/3/2009	Cobar	15.5	34.1	0	12.6
1/4/2009	Cobar	19.4	37.6	0	10.8
1/5/2009	Cobar	21.9	38.4	0	11.4
1/6/2009	Cobar	24.2	41	0	11.2
1/7/2009	Cobar	27.1	36.1	0	13
1/8/2009	Cobar	23.3	34	0	9.8
1/9/2009	Cobar	16.1	34.2	0	14.6

Date

The date of the observation

Location

The common name of the location of the weather stattion

MinTemp

The minimum temperature in degrees Celsius

MaxTemp

The maximum temperature in degrees Celsius

Rainfall

The amount of rainfall recorded for the day in mm

Evaporation

The Class A pan evaporation (mm) in the 24 hours prior to 9AM

Description

Daily Australia Weather Dataset

Sunshine	WindGustDir	WindGustSpeed	WindDir9am
12.3	SSW	48	ENE
13.0	S	37	SSE
13.3	SE	30	Not Available
10.6	NNE	46	NNE
12.2	WNW	31	WNW
8.4	WNW	35	NW
0.0	N	43	N
12.6	SSW	41	S
13.2	SE	37	SE
12.3	ENE	48	ENE

Sunshine

tThe number of hours of bright sunshine in the day

WindGustDir

The direction of the strongest wind gust in the 24 hours prior to midnight

WindGustSpeed

The speed (km/h) of the strongest wind gust in the 24 hours prior to midnight

WindDir9am

The direction of the wind at 9AM

Description

Daily Australia Weather Dataset

WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm
SW	6	20	20	13
SSE	19	19	30	8
N	13	7	68	7
NNW	30	15	42	22
WSW	6	6	37	22
WNW	17	13	19	15
WNW	7	20	26	19
SSE	17	19	33	15
S	15	6	25	9
WSW	30	9	46	28

WindDir3pm

Direction of the wind at 3PM - categorical and based upon the four cardinal directions

WindSpeed9am

Wind speed (km/h) averaged over the 10 minutes prior to 9AM

WindSpeed3pm

Wind speed (km/h) averaged over the 10 minute prior to 3PM

Humidity9am

Humidity (%) at 9AM

Humidity3pm

Humidity (%) at 3PM

Description

Daily Australia Weather Dataset

Pressure9am	Pressure3pm
1006.3	1004.4
1012.9	1012.1
NA	1011.6
1012.3	1009.2
1012.7	1009.1
1010.7	1007.4
1007.7	1007.4
1011.3	1009.9
1013.3	1009.2
1008.3	1004.0

Pressure9am

Atmospheric pressure (hPa) reduced to mean sea level at 9AM

Pressure3pm

atmospheric pressure (hPa) reduced to mean sea level at 3PM

Description

Daily Australia Weather Dataset

	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
1	2	5	26.6	33.4	No	No
2	1	1	20.3	27.0	No	No
3	NA	1	NA	32.7	No	No
4	1	6	28.7	34.9	No	No
5	1	5	29.1	35.6	No	No
6	1	6	33.6	37.6	No	No
7	8	8	30.7	34.3	No	No
8	3	1	25.0	31.5	No	No
9	1	1	20.7	32.8	No	No
10	1	5	23.4	33.3	No	No

Cloud9am

Fraction of sky obscured by clouds at 9AM (This is measured in “oktas”, which are a unit of eighths. It records how many eighths of the sky are obscured by clouds. 0 indicates a complete clear sky, while 8 indicates a completely overcast sky.)

Cloud3pm

Fraction of sky obscured by clouds at 3PM (oktas)

Temp9am

Temperature (degrees C) at 9AM

Temp3pm

Temperature (degrees C) at 3PM

RainToday

1 if precipitation in the 24 hours prior to 9AM exceeds 1mm,
0 otherwise

RainTomorrow

Did it rain the next day? (1 if yes, 0 otherwise)

Conversions

```
> str(weather)
'data.frame':   87360 obs. of  23 variables:
 $ Date         : Factor w/ 3436 levels "1/1/2008","1/1/2009",..: 2 112 222 252 262 272 282 2
92 302 12 ...
 $ Location     : Factor w/ 29 levels "Adelaide","Albany",..: 7 7 7 7 7 7 7 7 7 7 ...
```

When viewing a basic structure of our dataset, we noticed some conversions needed to be made. Str() was used to check the structure of our data. Whereas, all other variables appear to be correctly classified, some changes need to be made for the two variables listed above. Date should be 'date' while location should be a 'character'. We convert with as.Date() and as.character().

```
weather$Date <-  as.Date(weather$Date, format="%m/%d/%Y")
weather$Location <- as.character(weather$Location)
```

Checking the structure again ensures us the variables were properly changed.

```
'data.frame':   87360 obs. of  23 variables:
 $ Date         : Date, format: "2009-01-01" ...
 $ Location     : chr  "Cobar" "Cobar" "Cobar" "Cobar" ...
```

Summary Statistic

```
      Date                Location            MinTemp
 Min.   :2007-11-01   Length:87360       Min.   :-8.00
 1st Qu.:2010-11-19   Class :character   1st Qu.: 8.10
 Median :2013-03-23   Mode  :character   Median :12.50
 Mean   :2013-02-18                      Mean   :12.82
 3rd Qu.:2015-05-09                      3rd Qu.:17.50
 Max.   :2017-06-25                      Max.   :33.90
                                         NA's   :130
    MaxTemp         Rainfall        Evaporation
 Min.   : 4.10   Min.   :  0.000   Min.   :  0.000
 1st Qu.:18.10   1st Qu.:  0.000   1st Qu.:  2.600
 Median :23.00   Median :  0.000   Median :  4.600
 Mean   :23.62   Mean   :  2.342   Mean   :  5.462
 3rd Qu.:28.80   3rd Qu.:  0.800   3rd Qu.:  7.400
 Max.   :48.10   Max.   :371.000   Max.   :145.000
 NA's   :69      NA's   :677       NA's   :10405
```

Here we see some statistical summaries showing the minimum, median, mean, maximum, 1st & 3rd-quartiles and missing values for each numerical variable.

Function used

read.csv(), summary(), indexing

Data type

chr - Character and **dbl** - storing numeric values with decimal points

Missing Data

The "NA's" indicates the amount of missing data of each variable

Summary Statistic

```
    Date               Location           MinTemp          MaxTemp          Rainfall
Length:87360       Length:87360       Min.   :-8.00   Min.   : 4.10   Min.   :  0.000
Class :character   Class :character   1st Qu.: 8.10   1st Qu.:18.10   1st Qu.:  0.000
Mode  :character   Mode  :character   Median :12.50   Median :23.00   Median :  0.000
                                      Mean   :12.82   Mean   :23.62   Mean   :  2.342
                                      3rd Qu.:17.50   3rd Qu.:28.80   3rd Qu.:  0.800
                                      Max.   :33.90   Max.   :48.10   Max.   :371.000
                                      NA's   :130     NA's   :69      NA's   :677

 Evaporation
Min.   :  0.000
1st Qu.:  2.600
Median :  4.600
Mean   :  5.462
3rd Qu.:  7.400
Max.   :145.000
NA's   :10405
```

-->

```
     Date                Location            MinTemp
Min.   :2007-11-01   Length:87360       Min.   :-8.00
1st Qu.:2010-11-19   Class :character   1st Qu.: 8.10
Median :2013-03-23   Mode  :character   Median :12.50
Mean   :2013-02-18                      Mean   :12.82
3rd Qu.:2015-05-09                      3rd Qu.:17.50
Max.   :2017-06-25                      Max.   :33.90
                                        NA's   :130
    MaxTemp          Rainfall        Evaporation
Min.   : 4.10   Min.   :  0.000   Min.   :  0.000
1st Qu.:18.10   1st Qu.:  0.000   1st Qu.:  2.600
Median :23.00   Median :  0.000   Median :  4.600
Mean   :23.62   Mean   :  2.342   Mean   :  5.462
3rd Qu.:28.80   3rd Qu.:  0.800   3rd Qu.:  7.400
Max.   :48.10   Max.   :371.000   Max.   :145.000
NA's   :69      NA's   :677       NA's   :10405
```

Insight

- Missing data: 130 from MinTemp, 69 from MaxTemp. 677 from Rainfall, and 10405 from Evaporation
- Highly skewed variables: Rainfall and Evaporation
- Variables contains outliers: Rainfall and Evaporation
- Variables appears to have erroneous data: None of the first 6 variables
- The average rainfall daily in Australia is 2.342 mm (divided by all days in year, including the days without rain)
- For a rainy day, the average rainfall is 6.58 mm

```
    Rainfall
Min.   :  0.10
1st Qu.:  0.60
Median :  2.20
Mean   :  6.58
3rd Qu.:  6.80
Max.   :371.00
```

Summary Statistic

```
    Sunshine      WindGustSpeed
 Min.   : 0.000   Min.   :  9.00
 1st Qu.: 4.800   1st Qu.: 31.00
 Median : 8.400   Median : 39.00
 Mean   : 7.576   Mean   : 40.41
 3rd Qu.:10.600   3rd Qu.: 48.00
 Max.   :14.500   Max.   :135.00
 NA's   :15740    NA's   :5617
```

Insight

- Missing data: 15740 from Sunshine, 5617 from WindGustSpeed, 697 from WindSpeed9am
- Highly skewed variables: N/A
- Variables contain outliers: WindGustSpeed
- Variables appear to have erroneous data: N/A
- The average sunshine daily in Australia is 7.578 hours of bright sunshine per day.
- For a sunny day, there are 7.824 hours of bright sunshine.

```
    Sunshine
 Min.   : 0.100
 1st Qu.: 5.200
 Median : 8.600
 Mean   : 7.824
 3rd Qu.:10.700
 Max.   :14.500
```

Summary Statistic

WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm
Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00
1st Qu.: 9.00	1st Qu.:13.00	1st Qu.: 56.00	1st Qu.: 36.00
Median :13.00	Median :19.00	Median : 68.00	Median : 51.00
Mean :14.86	Mean :19.37	Mean : 67.05	Mean : 50.22
3rd Qu.:20.00	3rd Qu.:24.00	3rd Qu.: 81.00	3rd Qu.: 64.00
Max. :69.00	Max. :76.00	Max. :100.00	Max. :100.00
NA's :697	NA's :1341	NA's :587	NA's :1172

Insight

- Missing data: 697 from WindSpeed9am, 1341 from WindSpeed3pm, 587 from Humidity9am, and 1172 from Humidity3pm
- Highly skewed variables: WindSpeed9am
- Variables contains outliers: Windspeed9am, WindSpeed3pm
- Variables appears to have erroneous data: N/A

Summary Statistic

```
      WindGustDir     WindDir9am      WindDir3pm
SW             : 6115   N       : 7466   SW      : 6422
E              : 6097   E       : 6199   WSW     : 6374
W              : 6075   W       : 5619   SE      : 6121
N              : 5848   SSE     : 5412   W       : 6086
WSW            : 5776   SE      : 5378   S       : 6066
Not Available: 5641   ENE     : 5370   E       : 5648
(Other)        :51808   (Other):51916   (Other):50643
```

Insight

- Based on this summary, the strongest wind direction flows SW. E is the next strongest, then W, N, and WSW is the least strongest shown. Other wind directions found in 51,808 observations were not as strong.
- WindDir9am - In the morning, wind tends to flow N the most.
- WindDir3pm - Wind tends to flow SW in the evening.
- Missing Data - The strongest wind direction is not available for 5641 observations. This may be because the strongest wind direction is fairly equal between multiple directions. Wind can be unpredictable.
- Overall, whether it is morning or evening, wind gust direction is fairly the same in all directions. Some directions are only slightly stronger. However, there is no skewed-ness.

Summary Statistic

Pressure9am	Pressure3pm	Cloud9am	Cloud3pm
Min. : 980.5	Min. : 977.1	Min. :0.000	Min. :0.00
1st Qu.:1012.9	1st Qu.:1010.4	1st Qu.:1.000	1st Qu.:2.00
Median :1017.5	Median :1015.1	Median :5.000	Median :5.00
Mean :1017.6	Mean :1015.2	Mean :4.354	Mean :4.41
3rd Qu.:1022.3	3rd Qu.:1019.9	3rd Qu.:7.000	3rd Qu.:7.00
Max. :1040.9	Max. :1038.9	Max. :9.000	Max. :9.00
NA's :652	NA's :645	NA's :13755	NA's :15731

Temp9am	Temp3pm
Min. :-1.30	Min. : 3.70
1st Qu.:12.60	1st Qu.:16.90
Median :17.00	Median :21.60
Mean :17.52	Mean :22.16
3rd Qu.:22.30	3rd Qu.:27.10
Max. :39.40	Max. :46.70
NA's :159	NA's :777

Insight

- Missing data: 652 from Pressure9am, 645 from Pressure3pm, 13755 from Cloud9am, 15731 from Cloud3pm,
- 159 from Temp9am, 777 from temp
- Highly skewed variables: Slight positive skew in Temp9am and a more positive skew in Temp3pm
- Variables contains outliers: Pressure9am & Pressure3pm have outliers on both maximum and minimum sides.
  - Temp9am & Temp3pm have outliers lying closer to the maximum.
- Variables appears to have erroneous data: N/A

Summary Statistic

```
         RainToday      RainTomorrow
No              :67260   No :67723
Not Available:   677     Yes:19637
Yes             :19423
```

Insight

- Summaries on these two variables show that majority of the observations do not have rain.
- RainToday - "Yes" means areas had more than 1mm of rain, so the "no" can indicate slightly less than 1mm or none. Therefore, no rain today can be slightly misleading, however it can be used to conclude that less than 1mm of rain or none is insignificant in predicting whether it will rain the next day. Out of 87,360 observations, 677 dates had missing data. This may be attributed to data entry errors or data collection problems. Perhaps sequential dates that had less than 1mm or no rain were not accounted for.
    - Potential erroneous data in this variable.
- RainTomorrow - RainToday appears to be a good predictor of whether it will rain tomorrow. The observation counts for each category (yes vs. no) are similar. 463 more observations did not rain the next day compared to the previous day, while 214 observations had rain the next day.

Handling Missing Data

Use Deploy is.na() and median() function so that missing data is accounted for by replacing it with the median since a decent amount of data is missing within the variables.

```
Aus_wea_hd$MinTemp[is.na(Aus_wea_hd$MinTemp)] <- median(Aus_wea_hd$MinTemp, na.rm = TRUE)
Aus_wea_hd$MaxTemp[is.na(Aus_wea_hd$MaxTemp)] <- median(Aus_wea_hd$MaxTemp, na.rm = TRUE)
Aus_wea_hd$Rainfall[is.na(Aus_wea_hd$Rainfall)] <- median(Aus_wea_hd$Rainfall, na.rm = TRUE)
Aus_wea_hd$Evaporation[is.na(Aus_wea_hd$Evaporation)] <- median(Aus_wea_hd$Evaporation, na.rm = TRUE)
summary(Aus_wea_hd)
```

```
      Date              Location            MinTemp          MaxTemp          Rainfall        Evaporation
 Min.   :2007-11-01   Length:87360       Min.   :-8.00   Min.   : 4.10   Min.   :  0.000   Min.   :  0.000
 1st Qu.:2010-11-19   Class :character   1st Qu.: 8.10   1st Qu.:18.20   1st Qu.:  0.000   1st Qu.:  3.000
 Median :2013-03-23   Mode  :character   Median :12.50   Median :23.00   Median :  0.000   Median :  4.600
 Mean   :2013-02-18                      Mean   :12.82   Mean   :23.62   Mean   :  2.324   Mean   :  5.359
 3rd Qu.:2015-05-09                      3rd Qu.:17.50   3rd Qu.:28.80   3rd Qu.:  0.600   3rd Qu.:  7.000
 Max.   :2017-06-25                      Max.   :33.90   Max.   :48.10   Max.   :371.000   Max.   :145.000
```

However, it is unreasonable to remove it since that may lead to a biased subset of the data. Use median() function on all categorical variables.

Creating Dummy Variables

Deploy fastDummies package, and dummy_cols() function

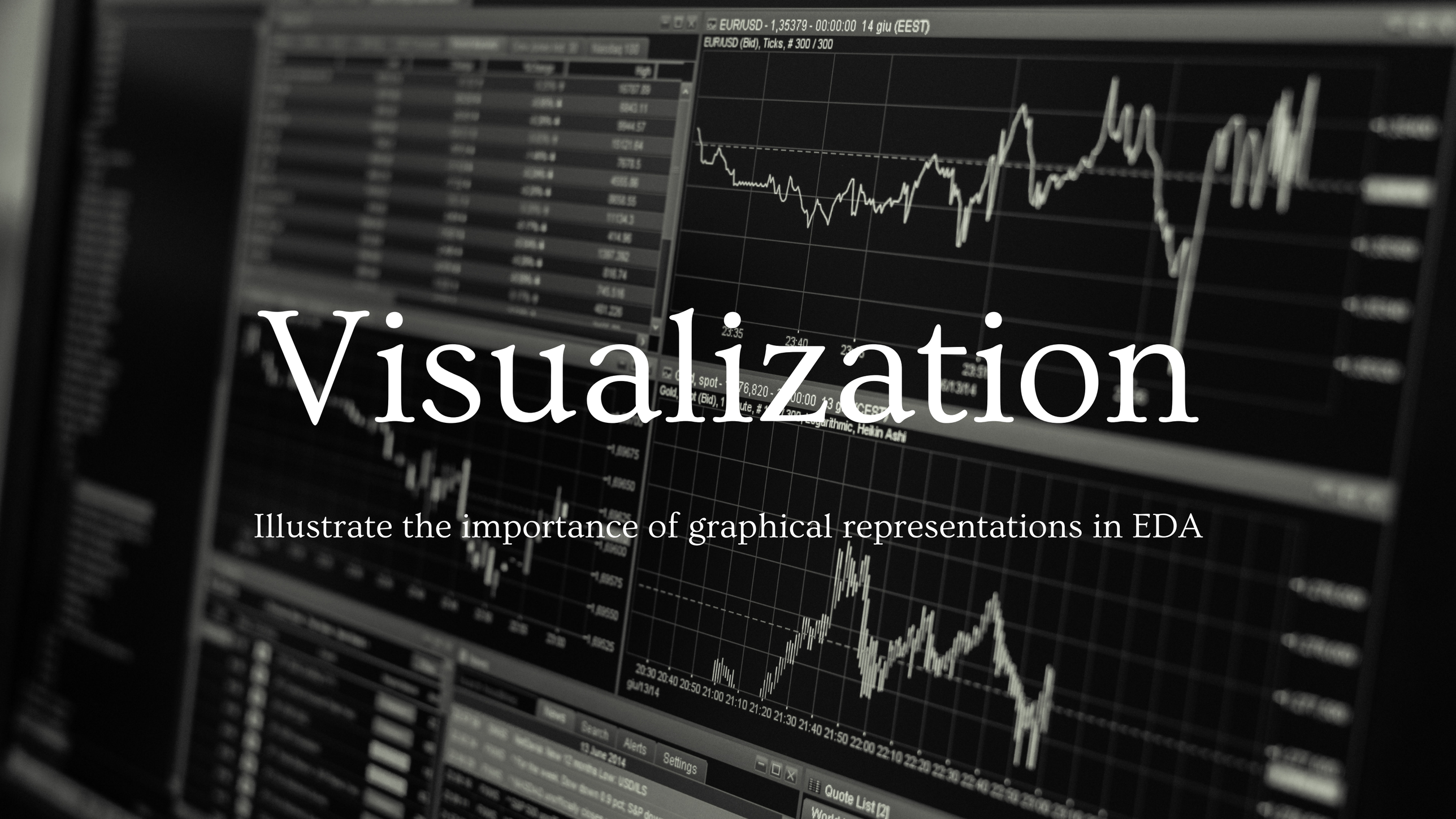
```
library(fastDummies)

#redefine data frame using this data
Aus_wea_hd <- dummy_cols(Aus_wea_hd,
           select_columns = "Location",
           remove_first_dummy = TRUE,
           remove_selected_columns = TRUE)
Aus_wea_hd
```

Create 28 dummy variables for the Location variable including 29 different stations across the Australia.

- Choose the Location as the categorical variable
- Remove the first dummy since we only need n-1 dummy to define n variables.
- Remove the Location column

Location_AliceSprings <int>	Location_Brisbane <int>	Location_Cairns <int>	Location_Canberra <int>	Location_Cobar <int>
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1



Visualization

Illustrate the importance of graphical representations in EDA

Correlation Table & Heatmap

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
MinTemp	1.00	0.74	0.11	0.44	0.06	0.15	0.16	0.15	-0.21	0.04	-0.47	-0.48	0.09	0.03	0.90	0.72
MaxTemp	0.74	1.00	-0.05	0.56	0.42	0.07	0.03	0.04	-0.51	-0.47	-0.34	-0.44	-0.24	-0.23	0.89	0.98
Rainfall	0.11	-0.05	1.00	-0.05	-0.21	0.11	0.07	0.05	0.23	0.25	-0.17	-0.13	0.18	0.16	0.03	-0.06
Evaporation	0.44	0.56	-0.05	1.00	0.31	0.19	0.19	0.12	-0.48	-0.37	-0.25	-0.27	-0.18	-0.17	0.52	0.54
Sunshine	0.06	0.42	-0.21	0.31	1.00	-0.04	0.00	0.05	-0.43	-0.56	0.04	-0.01	-0.58	-0.60	0.25	0.43
WindGustSpeed	0.15	0.07	0.11	0.19	-0.04	1.00	0.58	0.67	-0.20	-0.04	-0.45	-0.40	0.07	0.11	0.13	0.03
WindSpeed9am	0.16	0.03	0.07	0.19	0.00	0.58	1.00	0.49	-0.25	-0.06	-0.23	-0.19	0.02	0.05	0.11	0.02
WindSpeed3pm	0.15	0.04	0.05	0.12	0.05	0.67	0.49	1.00	-0.11	0.02	-0.31	-0.27	0.05	0.02	0.14	0.01
Humidity9am	-0.21	-0.51	0.23	-0.48	-0.43	-0.20	-0.25	-0.11	1.00	0.68	0.13	0.18	0.40	0.31	-0.46	-0.50
Humidity3pm	0.04	-0.47	0.25	-0.37	-0.56	-0.04	-0.06	0.02	0.68	1.00	-0.05	0.03	0.46	0.46	-0.18	-0.52
Pressure9am	-0.47	-0.34	-0.17	-0.25	0.04	-0.45	-0.23	-0.31	0.13	-0.05	1.00	0.96	-0.12	-0.14	-0.44	-0.30
Pressure3pm	-0.48	-0.44	-0.13	-0.27	-0.01	-0.40	-0.19	-0.27	0.18	0.03	0.96	1.00	-0.06	-0.08	-0.49	-0.40
Cloud9am	0.09	-0.24	0.18	-0.18	-0.58	0.07	0.02	0.05	0.40	0.46	-0.12	-0.06	1.00	0.59	-0.11	-0.26
Cloud3pm	0.03	-0.23	0.16	-0.17	-0.60	0.11	0.05	0.02	0.31	0.46	-0.14	-0.08	0.59	1.00	-0.10	-0.27
Temp9am	0.90	0.89	0.03	0.52	0.25	0.13	0.11	0.14	-0.46	-0.18	-0.44	-0.49	-0.11	-0.10	1.00	0.86
Temp3pm	0.72	0.98	-0.06	0.54	0.43	0.03	0.02	0.01	-0.50	-0.52	-0.30	-0.40	-0.26	-0.27	0.86	1.00

Strongly correlated

- Temp9am with MinTemp +0.90
- Temp9am with MaxTemp +0.89
- Temp9am with Temp3pm +0.86
- Temp3pm with MaxTemp +0.98
- Pressure9am with Pressure3pm +0.96

Moderately correlated

- MaxTemp with MinTemp +74
- Temp3pm with MinTemp +72
- WindGustSpeed with WinSpeed 3pm +67
- Sunshine with Cloud3pm -0.60

Multicollinearitty (p > 0.95)

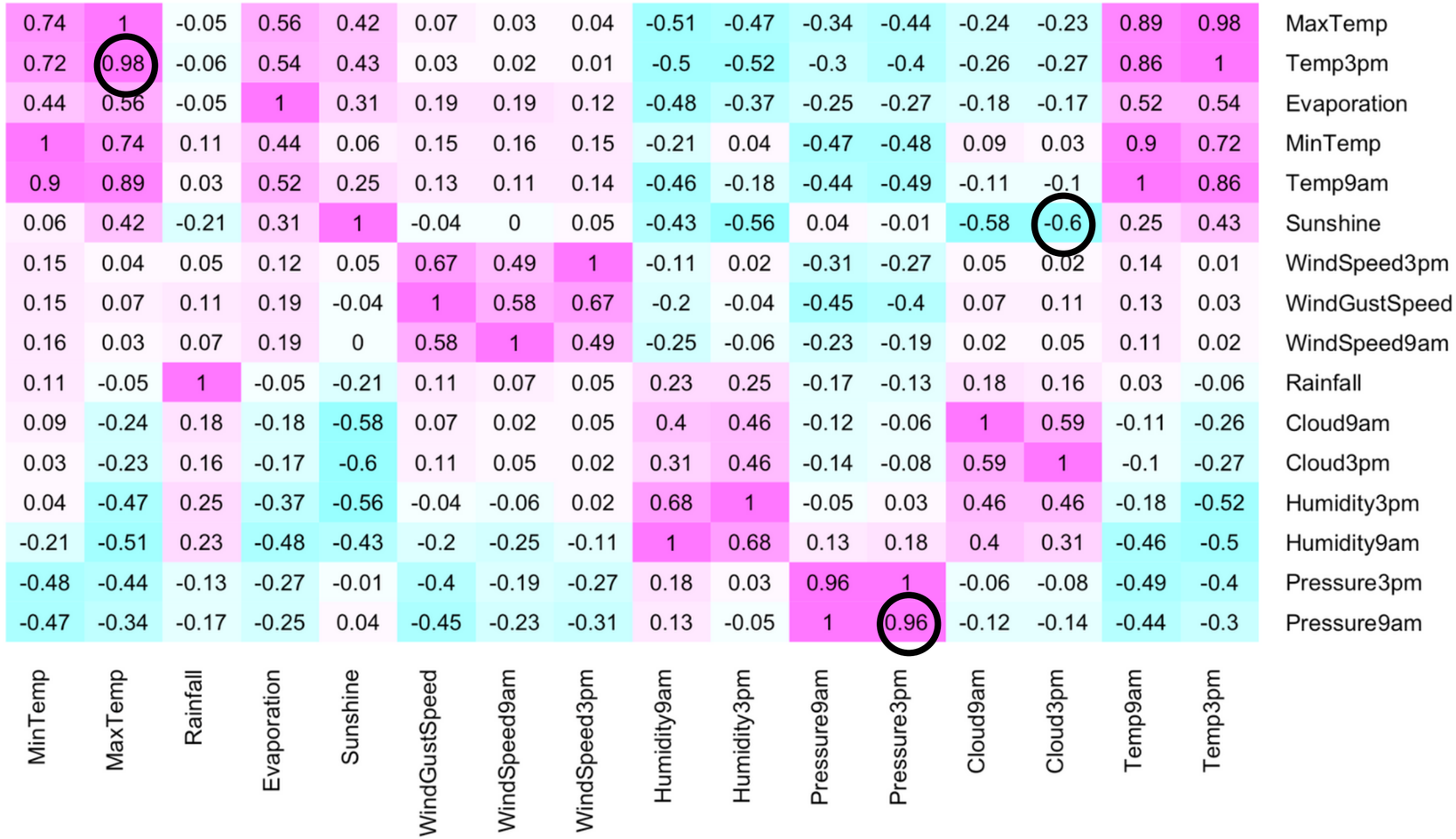
- Temp3pm with MaxTemp +0.98
- Pressure9am with Pressure3pm +0.96

Correlation Summary

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
MinTemp	1.00	0.74	0.11	0.44	0.06	0.15	0.16	0.15	-0.21	0.04	-0.47	-0.48	0.09	0.03	0.90	0.72
MaxTemp	0.74	1.00	-0.05	0.56	0.42	0.07	0.03	0.04	-0.51	-0.47	-0.34	-0.44	-0.24	-0.23	0.89	0.98
Rainfall	0.11	-0.05	1.00	-0.05	-0.21	0.11	0.07	0.05	0.23	0.25	-0.17	-0.13	0.18	0.16	0.03	-0.06
Evaporation	0.44	0.56	-0.05	1.00	0.31	0.19	0.19	0.12	-0.48	-0.37	-0.25	-0.27	-0.18	-0.17	0.52	0.54
Sunshine	0.06	0.42	-0.21	0.31	1.00	-0.04	0.00	0.05	-0.43	-0.56	0.04	-0.01	-0.58	-0.60	0.25	0.43
WindGustSpeed	0.15	0.07	0.11	0.19	-0.04	1.00	0.58	0.67	-0.20	-0.04	-0.45	-0.40	0.07	0.11	0.13	0.03
WindSpeed9am	0.16	0.03	0.07	0.19	0.00	0.58	1.00	0.49	-0.25	-0.06	-0.23	-0.19	0.02	0.05	0.11	0.02
WindSpeed3pm	0.15	0.04	0.05	0.12	0.05	0.67	0.49	1.00	-0.11	0.02	-0.31	-0.27	0.05	0.02	0.14	0.01
Humidity9am	-0.21	-0.51	0.23	-0.48	-0.43	-0.20	-0.25	-0.11	1.00	0.68	0.13	0.18	0.40	0.31	-0.46	-0.50
Humidity3pm	0.04	-0.47	0.25	-0.37	-0.56	-0.04	-0.06	0.02	0.68	1.00	-0.05	0.03	0.46	0.46	-0.18	-0.52
Pressure9am	-0.47	-0.34	-0.17	-0.25	0.04	-0.45	-0.23	-0.31	0.13	-0.05	1.00	0.96	-0.12	-0.14	-0.44	-0.30
Pressure3pm	-0.48	-0.44	-0.13	-0.27	-0.01	-0.40	-0.19	-0.27	0.18	0.03	0.96	1.00	-0.06	-0.08	-0.49	-0.40
Cloud9am	0.09	-0.24	0.18	-0.18	-0.58	0.07	0.02	0.05	0.40	0.46	-0.12	-0.06	1.00	0.59	-0.11	-0.26
Cloud3pm	0.03	-0.23	0.16	-0.17	-0.60	0.11	0.05	0.02	0.31	0.46	-0.14	-0.08	0.59	1.00	-0.10	-0.27
Temp9am	0.90	0.89	0.03	0.52	0.25	0.13	0.11	0.14	-0.46	-0.18	-0.44	-0.49	-0.11	-0.10	1.00	0.86
Temp3pm	0.72	0.98	-0.06	0.54	0.43	0.03	0.02	0.01	-0.50	-0.52	-0.30	-0.40	-0.26	-0.27	0.86	1.00

For the two pairs of Multicollinearity which are Temp3pm - MaxTemp and Pressure9am -Pressure3pm, we could exclude one variables of a pair as the data reduction process to increase the efficiency of data by remove the potential information overlapping.

Heatmap

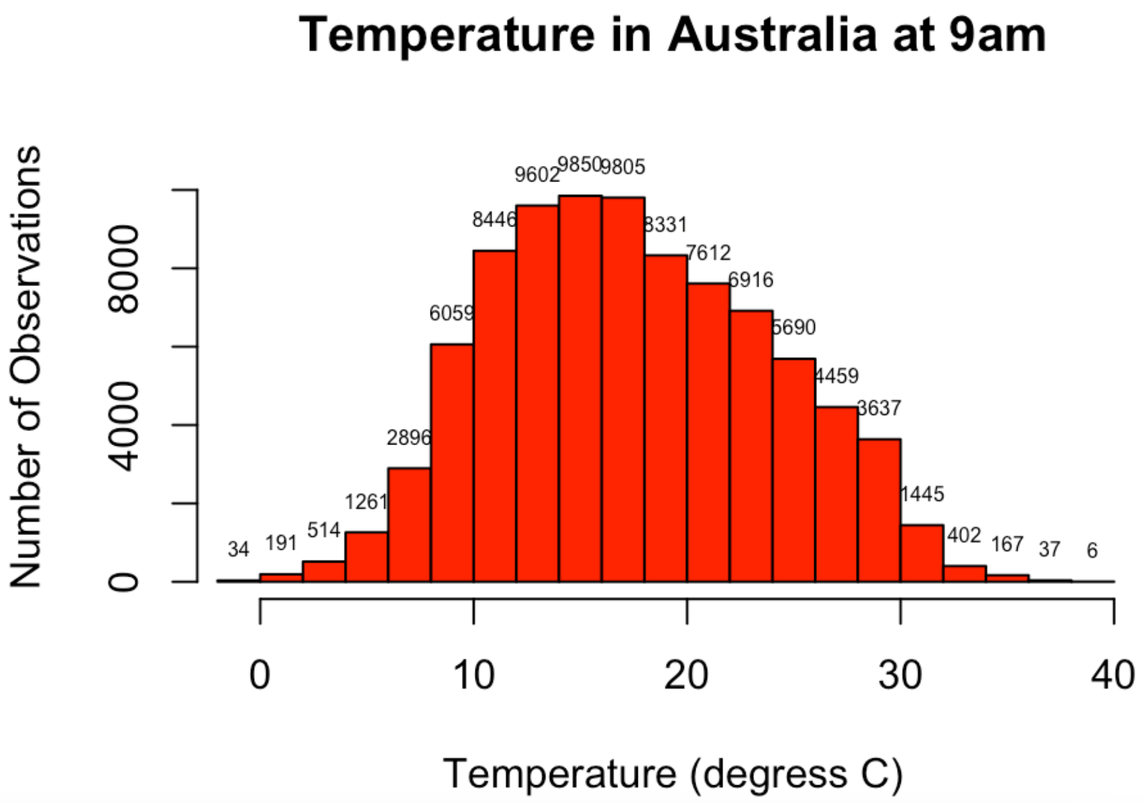


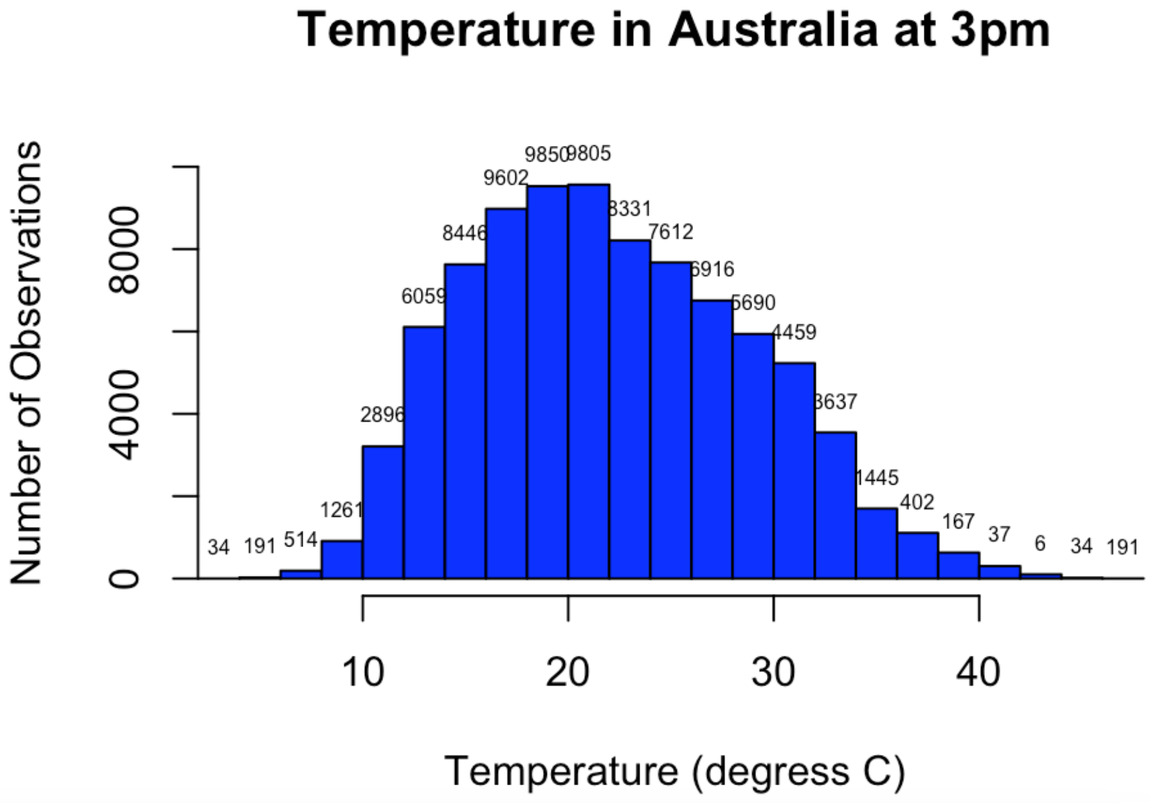
The strongest positively correlated pairs are shown in the darkest pink rectangle. On the flip side, the strongest negatively correlated pairs are shown in the darkest mint blue.

Histogram

Here we look at a comparison of temperature in Australia at 9am versus 3pm to determine a possible relationship to rain.

There is a more positive skew seen in the temperature at 3pm than 9am. This tells us it gets warmer as the day progresses in Australia. At 9am majority of observations are between 15-40 degrees C, while at 3pm, majority of observations are between 18-47 degrees C.





More observations with higher temperatures may indicate less chance of rain.

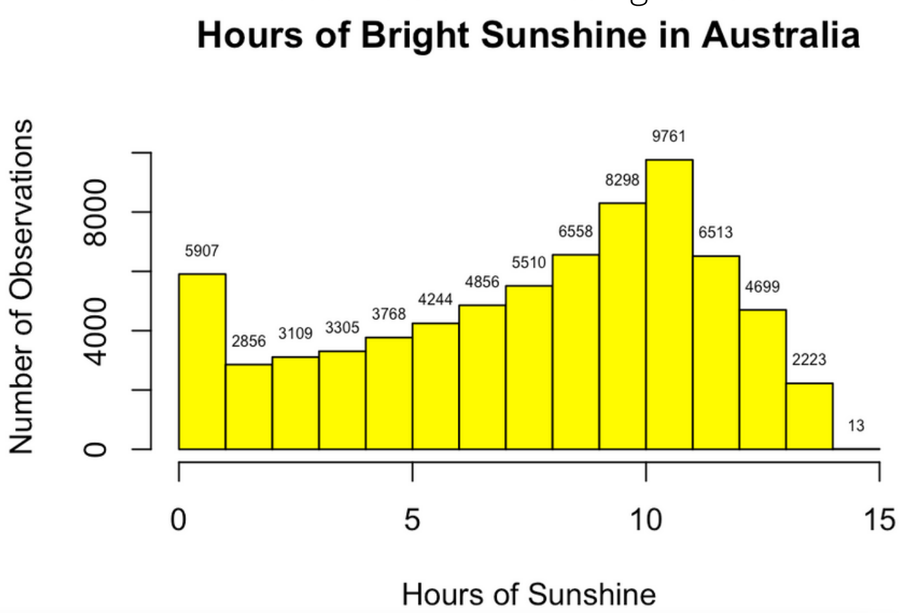
Histogram

Next we look at the hours of bright sunshine in Australia.

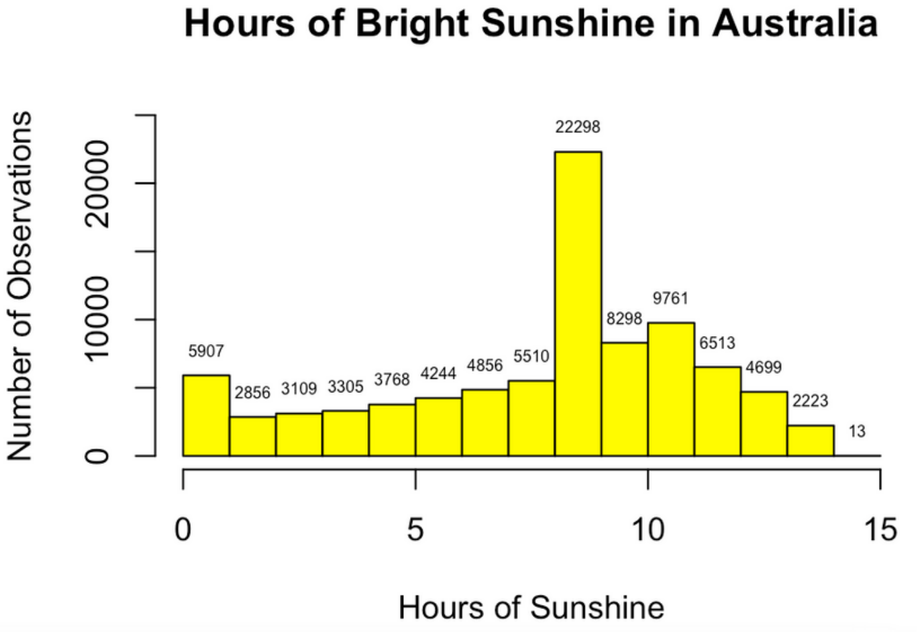
In contrast to the previous histograms, the hours of bright sunshine in Australia is negatively skewed. Majority of observations have between 0-10 hours of bright sunshine. Perhaps more hours of sunshine is needed for rain, since more heat allows the water cycle to work.

When viewing the hours of bright sunshine in Australia, perhaps converting missing values to the median is not a good step to take since it biases the data towards the median dramatically.

No conversion of missing values

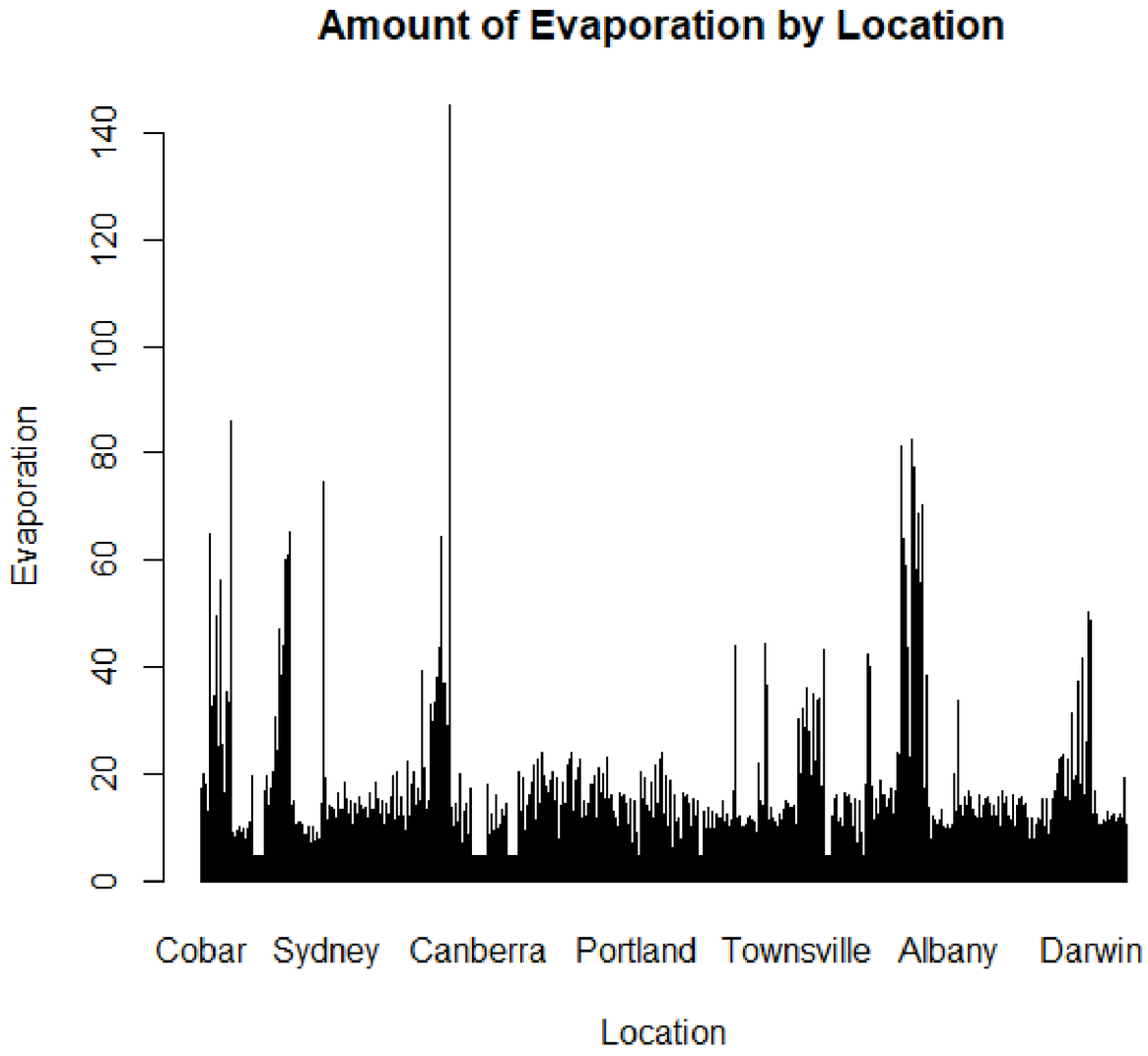


With conversion of missing values to the median.



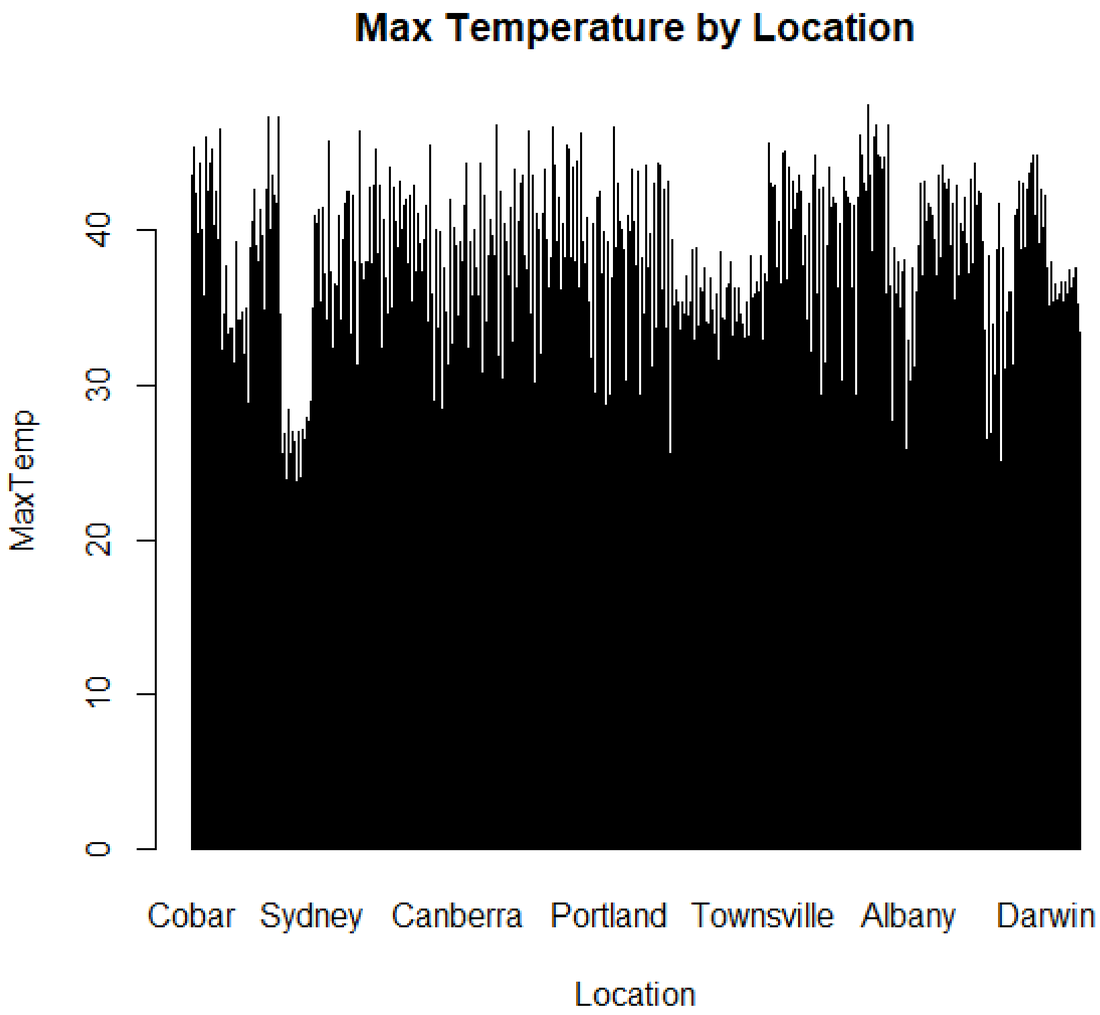
Bar Charts

This chart shows that Canberra and Portland do not frequently have a high amount of evaporation; however, Canberra does have the highest amount of evaporation recorded. Albany and Cobar frequently have high amounts of evaporation.



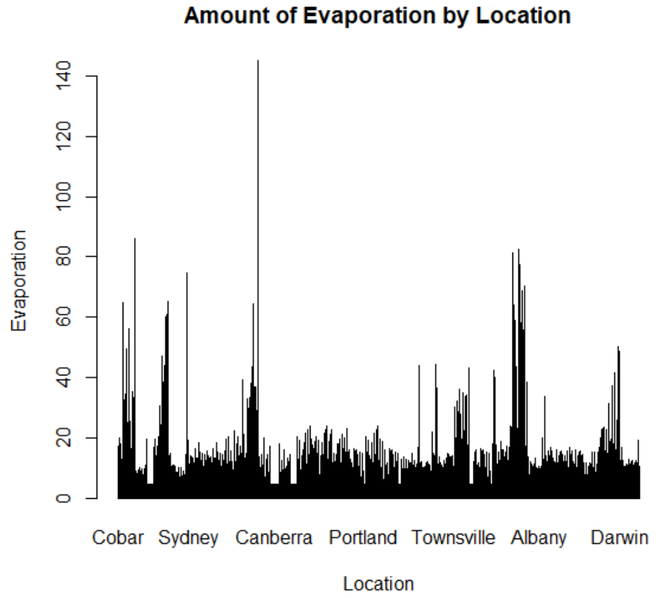
Bar Charts

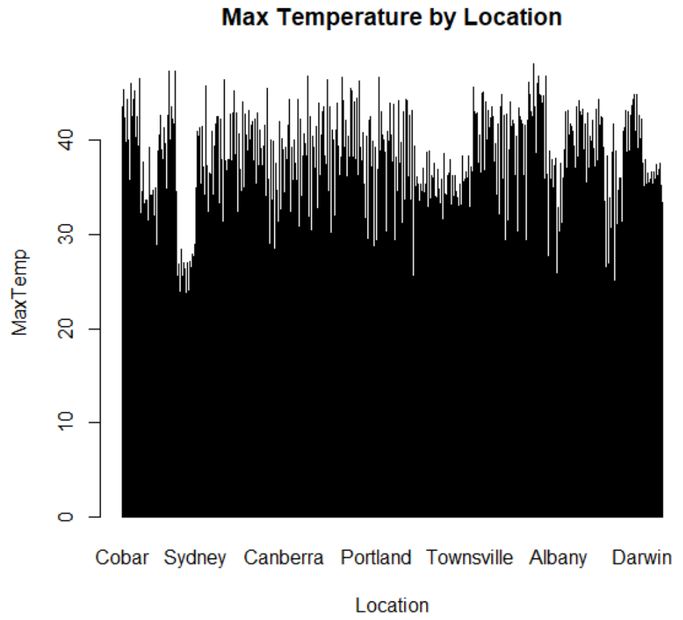
This chart shows that Sydney has the lowest max temperature and Townsville also has a lower max temperature than the other locations. Many of the locations have recorded max temperatures above 40 degrees which shows that all of these locations are in a warmer climate.



Bar Charts

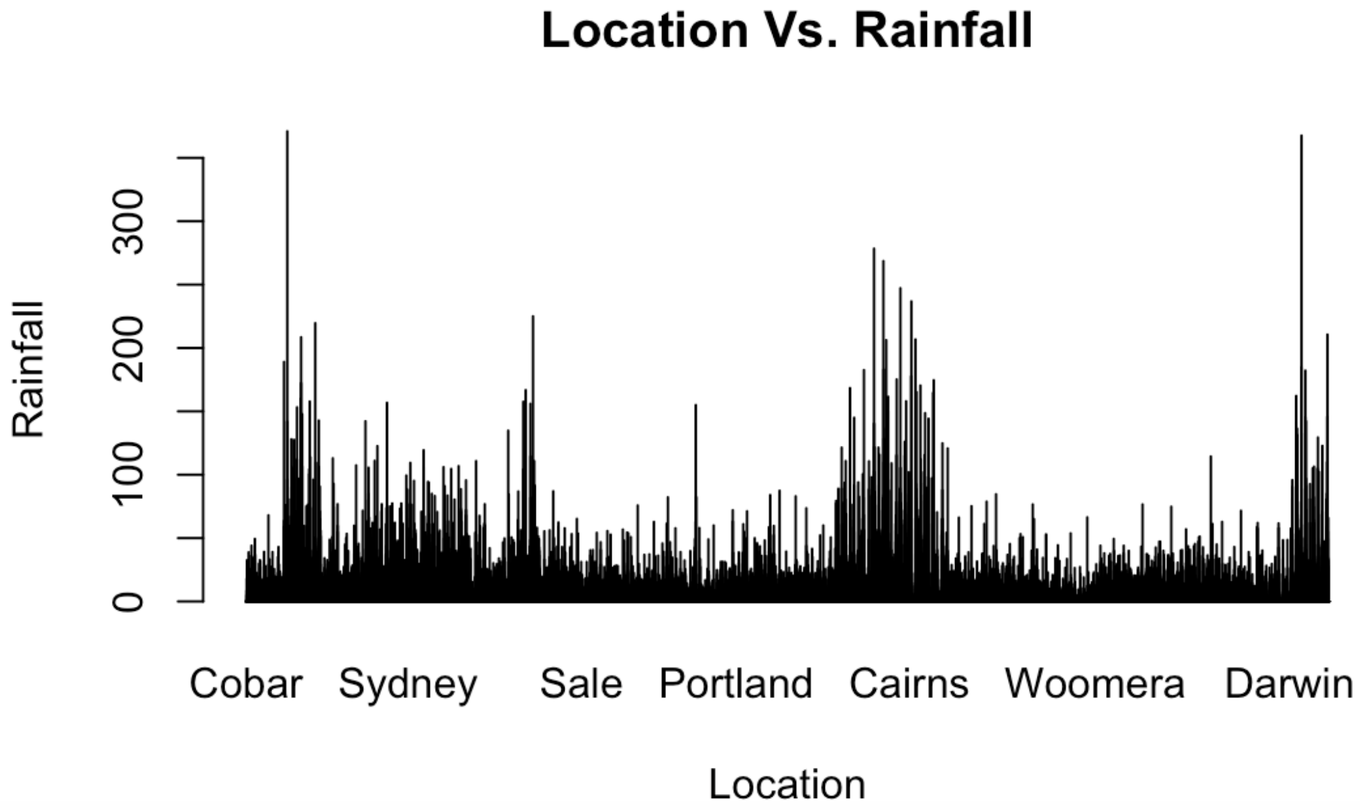
When comparing these charts it is clear there is some relationship between the max temperature and evaporation. Albany frequently has high amounts of evaporation and has one of the highest max temperatures recorded. These bar charts show that evaporation can increase as the max temperature increases.



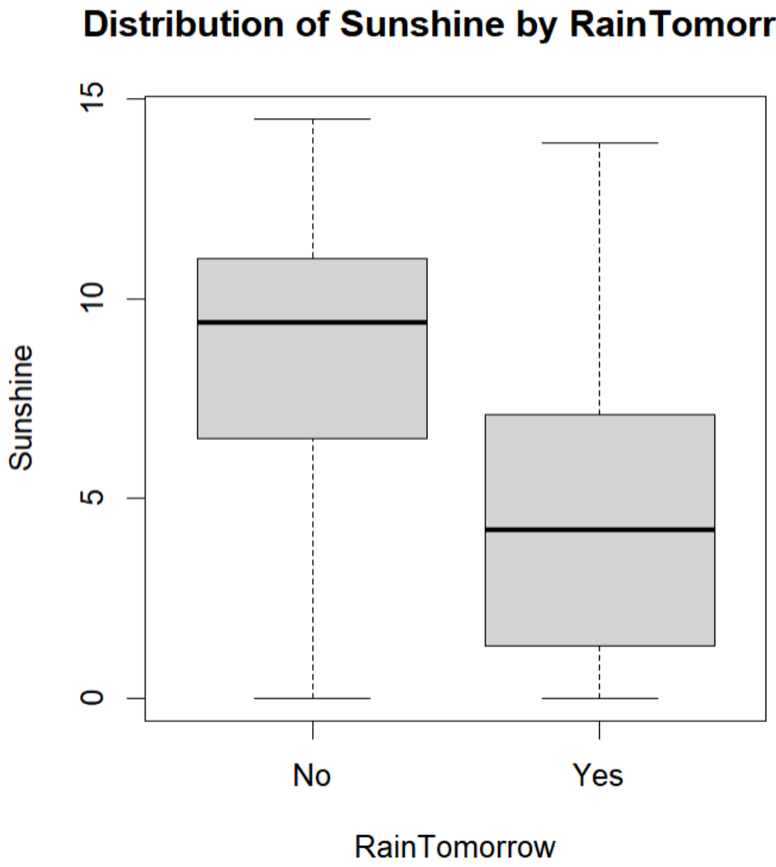


Bar Charts

Looking at this bar chart in comparison to the previous ones we see common location names "Cobar", "Sydney" "Portland", and "Darwin". These are the most significant areas from our dataset. Since they are common in the bar charts produced, we can assume areas with higher max temperature and evaporation are more likely to have the highest rainfall.



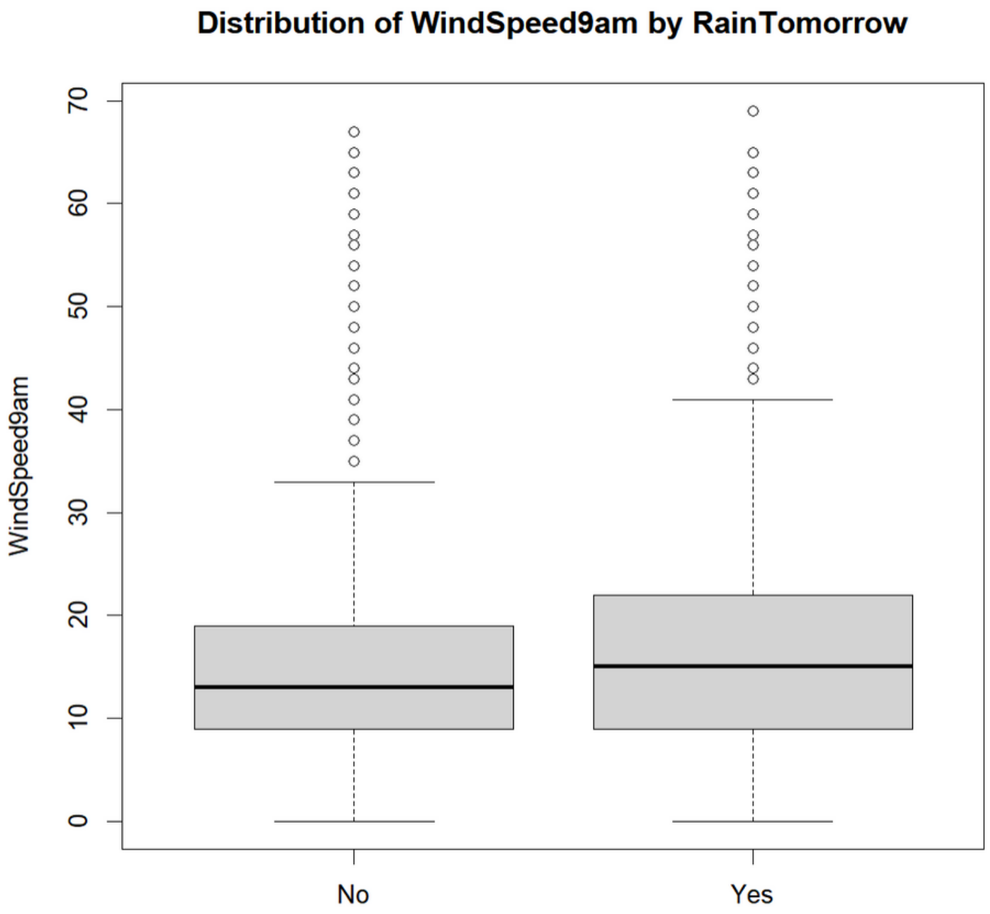
Box Plots



Will it rain tomorrow based on the amount of sunshine?

This box plot shows that generally, the more sunshine there is, the lower the chance of rain tomorrow. There is a strong correlation between low levels of sunshine and rain tomorrow.

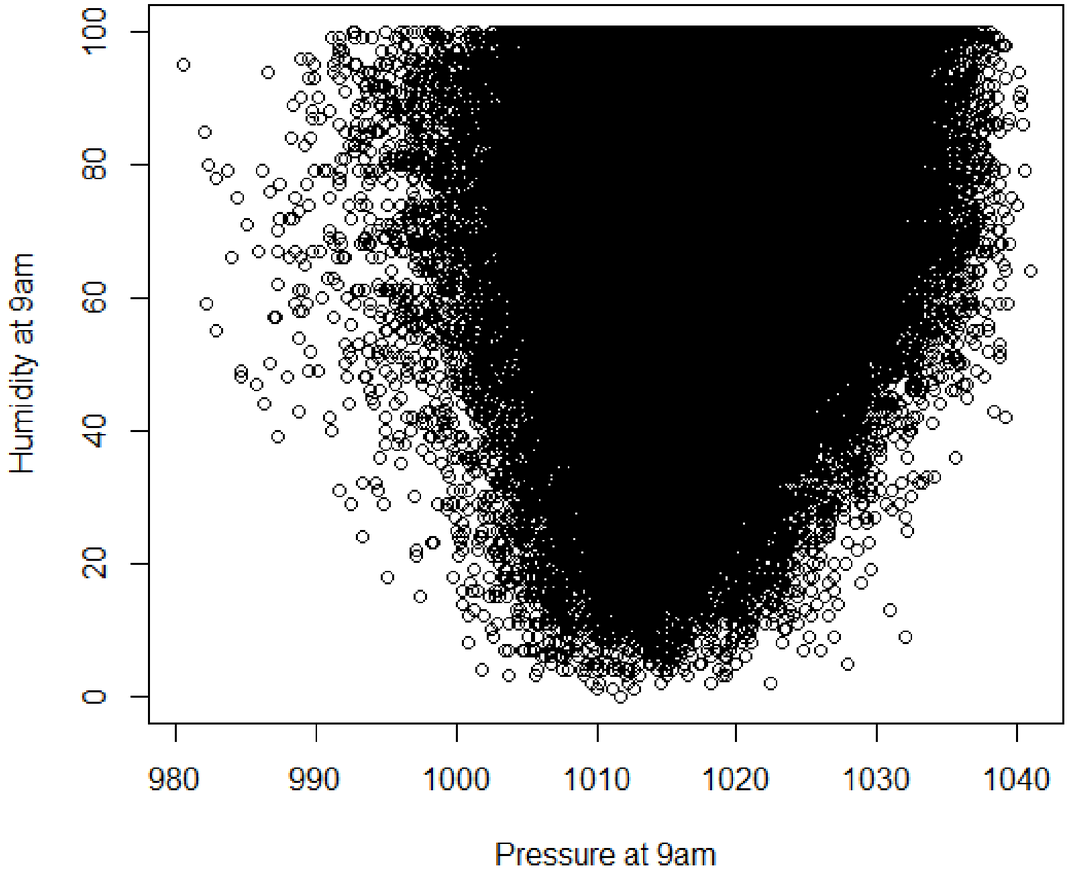
Box Plots



Will it rain tomorrow based on the wind speeds at 9am?

This box plot is harder to read, but still shows that generally, the higher the wind speeds at 9am, the more of a chance there is for there to be rain tomorrow. This correlation would be hard to discern just by viewing this visualization.

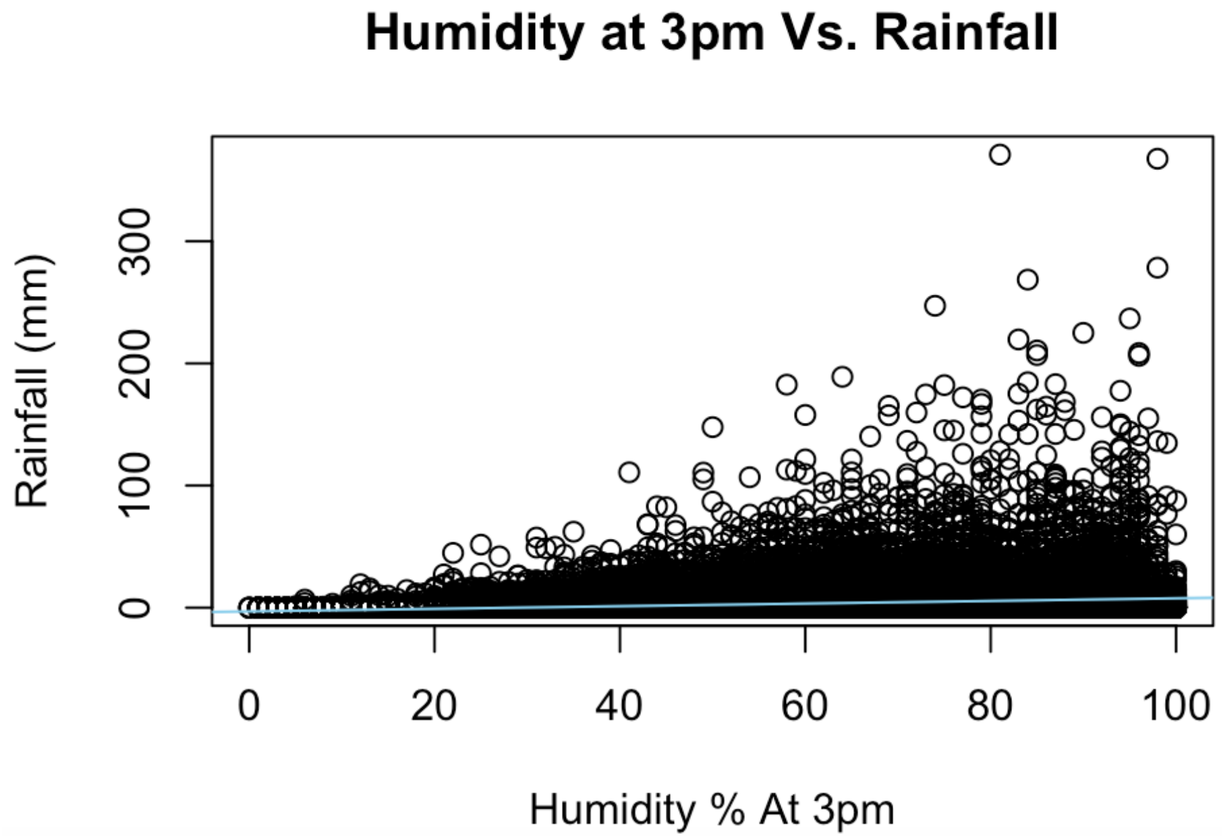
Scatter Plots



This graph explores the relationship between humidity and pressure at 9 am.

No significant correlation can be noted.

Scatter Plots



Based on the correlation plot, humidity at 3pm and rainfall have a slight correlation, so we can visualize this relationship better with a scatter plot.

An abline() function was used to better see the correlation between the two variables.

This small relationship shows that if humidity at 3pm was higher, rainfall tended to be greater as well.

One thing to note is a lot of the observations gather towards 0 mm of rainfall even if the humidity is higher. So humidity may not be the most accurate predictor in determining whether it will rain the next day.

Conclusions

Overall, based on the correlation matrix, and charts/graphs created, no significant relationships can be used to determine whether or not it will rain the next day. The most strongly correlated pairs are between the same variables which is not useful in our predictions. However, we did find that areas with more intense weather conditions like higher humidity, evaporation, and temperature had a better chance at producing rain.

Conclusions

Lots of variables can affect rain, however due to the unpredictability of rain, as well as missing data, it could make predicting whether or not it will rain challenging.

A better way at studying this dataset to better predict whether it will rain the next day is group the variables together of each location and compare the weather conditions by region.

Thank you!

R Champions

Omar Ahmouda, Leona Bell, Helen Chu, Hoai Do, Delaney Smith

BANA 4080 - Homework 1

EDA Australia Weather Data