Rain In Australia
Exploratory Analysis

Jacklyn Tellez, jtellez2@bellarmine.edu
Hunter Waldridge, hwaldridge@bellarmine.edu

I. INTRODUCTION

The data set we chose measures the rain fall and other weather data from 2007 to 2017 in 49 cities in Australia. It shows the highest and lowest temperature from each day. It also shows the temperature, wind direction, wind speed, humidity, pressure, and cloud coverage for each day at 9am and 3pm. This data set was chosen because it had data from eleven years and 49 cities. This made it possible to explore the data in endless ways. We cold look at one city, one year, one month, compare cities in a single year, etc.

Rain in Australia | Kaggle

II. DATA SET DESCRIPTION

This data set originally had 23 columns and 145,460 rows. The 4 columns with temperature (MinTemp, MaxTemp, Temp9am, and Temp3pm) were all in Celsius. We added 4 new columns so that we would also have these columns in Fahrenheit. For the 8 temperature columns, we deciphered the difference between Celsius and Fahrenheit by putting _C or _F after each one. After adding our columns, we had 27 columns and 145,460 rows. Of the 27 variables, they were categorized in pandas as the following: six were objects, one we transformed into Datetime, and twenty were float. Five were nominal, one ordinal, nine were interval, and twelve were ratio. Four of the variables were missing over 35% of the data. These four variables are missing so much data that it would not be worth exploring them. Seven variables are missing 7-11% of their data. While a good portion of the data is missing, an analysis could still be made on these variables, but would probably be better looking at other variables first. Sixteen variables are missing 0.5-3.5% of their data. These are the variables that time will be spent most on. Only two variables are not missing any of their data values. These two variables are Location and Date. It is reasonable that these two variables are not missing data because they are the variables that tell us where and when the data was collected from. These two variables are the variables we use to group our data to find the averages. We could find averages by year, day, month, or city.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%) (to the nearest thousand)*
Date	Datetime64[ns]; interval	0%
Location	Object; nominal	0%
MinTemp_C	Float64; interval	1.021%
MinTemp_F	Float64; interval	1.021%
MaxTemp_C	Float64; interval	0.867%
MinTemp_F	Float64; interval	0.867%
Rainfall	Float64; ratio	2.242%
Evaporation	Float64; ratio	43.167%
Sunshine	Float64; ratio	48.010%
WindGustDir	Object; nominal	7.099%
WindGustSpeed	Float64; ratio	7.056%
WindDir9am	Object; nominal	7.264%
WindDir3pm	Object; nominal	2.907%
WindSpeed9am	Float64; ratio	1.215%
WindSpeed3pm	Float64; ratio	2.105%
Humidity9am	Float64; ratio	1.824%
Humidity3pm	Float64; ratio	3.098%
Pressure9am	Float64; ratio	10.357%

Pressure3pm	Float64; ratio	10.331%
Cloud9am	Float64; ratio	38.422%
Cloud3pm	Float64; ratio	40.807%
Temp9am_C	Float64; interval	1.215%
Temp9am_F	Float64; interval	1.215%
Temp3pm_C	Float64; interval	2.481%
Temp3pm_F	Float64; interval	2.481%
RainToday	Object; nominal	2.242%
RainTomorrow	Object; ordinal	2.246%

III. Data Set Summary Statistics

Narrative introduction to the section.

Table 2: Summary Statistics for Rain in Australia

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
MinTemp_C	143,975.00	12.194	6.398	-8.500	7.600	12.000	16.900	33.900
MinTemp_F	143,975.00	53.949	11.517	16.700	45.680	53.600	62.420	93.020
MaxTemp_C	144,199.00	23.221	7.119	-4.800	17.900	22.600	28.200	48.100
MinTemp_F	144,199.00	73.798	12.814	23.360	64.220	72.680	82.760	118.580
Rainfall	142,199.00	2.361	8.478	0.000	0.000	0.00	0.800	371.000
Evaporation	82,670.00	5.468	4.194	0.000	2.600	4.800	7.400	145.000
Sunshine	75,625.00	7.611	3.785	0.000	4.800	8.400	10.600	14.500
WindGustSpeed	135,197.00	40.035	13.607	6.000	31.000	39.000	48.000	135.000
WindSpeed9am	143,693.00	14.043	8.915	0.000	7.000	13.000	19.000	130.000
WindSpeed3pm	142,398.00	18.663	8.810	0.000	13.000	19.000	24.000	87.000
Humidity9am	142,806.00	68.881	19.029	0.000	57.000	70.000	83.000	100.000
Humidity3pm	140,953.00	51.539	20.796	0.000	37.000	52.000	66.000	100.000
Pressure9am	130,395.00	1017.650	7.107	980.500	1012.900	1017.600	1022.400	1041.000
Pressure3pm	130,432.00	1015.256	7.037	977.100	1010.400	1015.200	1020.000	1039.600
Cloud9am	89,572.00	4.447	2.887	0.000	1.000	5.000	7.000	9.000
Cloud3pm	86,102.00	4.510	2.720	0.000	2.000	5.000	7.000	9.000
Temp9am_C	143,693.00	16.991	6.489	-7.200	12.300	16.700	21.600	40.200
Temp9am_F	143,693.00	62.583	11.680	19.040	54.140	62.060	70.880	104.360
Temp3pm_C	141,851.00	21.683	6.937	-5.400	16.600	21.100	26.400	46.700
Temp3pm_F	141,851.00	71.030	12.486	22.280	61.880	69.980	79.520	116.060

There should be a table for **EACH** categorical variable.

Table 3: Proportions for XXX (n=yyy)

Category	*Frequency*	*Proportion (%)*
Location: Canberra	3436	*2.362%*
Location: Sydney	3344	*2.299%*
Location: Darwin	3193	*2.195%*
Location: Hobart	3193	*2.195%*
Location: Brisbane	3193	*2.195%*
Location: Melbourne	3193	*2.195%*
Location: Perth	3193	*2.195%*
Location: Adelaide	3193	*2.195%*
Location: Wollongong	3040	*2.090%*
Location: AliceSprings	3040	*2.090%*
Location: Albany	3040	*2.090%*
Location: Launceston	3040	*2.090%*
Location: Cairns	3040	*2.090%*

Location: Bendigo	3040	*2.090%*
Location: MountGambier	3040	*2.090%*
Location: Ballarat	3040	*2.090%*
Location: GoldCoast	3040	*2.090%*
Location: Albury	3040	*2.090%*
Location: MountGinini	3040	*2.090%*
Location: Townsville	3040	*2.090%*
Location: Tuggeranong	3039	*2.089%*
Location: Newcastle	3039	*2.089%*
Location: Penrith	3039	*2.089%*
Location: Portland	3009	*2.069%*
Location: Witchcliffe	3009	*2.069%*
Location: Woomera	3009	*2.069%*
Location: Cobar	3009	*2.069%*
Location: Williamtown	3009	*2.069%*
Location: sale	3009	*2.069%*
Location: Waggawagga	3009	*2.069%*
Location: Mildura	3009	*2.069%*
Location: Moree	3009	*2.069%*
Location: Sydney airport	3009	*2.069%*
Location: Perth airport	3009	*2.069%*
Location: norfolks island	3009	*2.069%*
Location: Badgerys Creek	3009	*2.069%*
Location: Melbourne airport	3009	*2.069%*
Location: Dartmoor	3009	*2.069%*
Location: Watsonia	3009	*2.069%*
Location: Richmond	3009	*2.069%*
Location: Nuriootpa	3009	*2.069%*
Location: coffs harbour	3009	*2.069%*
Location: Pearce RAAF	3009	*2.069%*
Location: Walpole	3006	*2.067%*
Location: Norah Head	3004	*2.065%*
Location: SalomGums	3001	*2.063%*
Location: Nhil	1578	*1.085%*
Location: Katherine	1578	*1.085%*
Location: Uluru	1578	*1.085%*
WindGustDir: W	9915	*7.337%*
WindGustDir: SE	9418	*6.969%*
WindGustDir: N	9313	*6.892%*
WindGustDir: SSE	9216	*6.820%*
WindGustDir: E	9181	*6.794%*
WindGustDir: S	9168	*6.784%*
WindGustDir: WSW	9069	*6.711%*
WindGustDir: SW	8967	*6.636%*
WindGustDir: SSW	8736	*6.465%*
WindGustDir: WNW	8252	*6.107%*
WindGustDir: NW	8122	*6.010%*
WindGustDir: ENE	8104	*5.997%*
WindGustDir: ESE	7372	*5.455%*
WindGustDir: NE	7133	*5.278%*
WindGustDir: NWN	6620	*4.899%*
WindGustDir: NNE	6548	*4.846%*
WindDir9am: N	11,758	*8.716%*

WindDir9am: SE	9287	*6.885%*
WindDir9am: E	9176	*6.802%*
WindDir9am: SSE	9112	*6.755%*
WindDir9am: NW	8749	*6.486%*
WindDir9am: S	8659	*6.419%*
WindDir9am: W	8459	*6.271%*
WindDir9am: SW	8423	*6.244%*
WindDir9am: NNE	8129	*6.026%*
WindDir9am: NNW	7980	*5.916%*
WindDir9am: ENE	7836	*5.809%*
WindDir9am: NE	7671	*5.687%*
WindDir9am: ESE	7630	*5.656%*
WindDir9am: SSW	7587	*5.624%*
WindDir9am: WNW	7414	*5.496%*
WindDir9am: WSW	7024	*5.207%*
WindDir3pm: SE	10,838	*7.674%*
WindDir3pm: W	10,110	*7.158%*
WindDir3pm: S	9926	*7.028%*
WindDir3pm: WSW	9518	*6.739%*
WindDir3pm: SSE	9399	*6.655%*
WindDir3pm: SW	9354	*6.623%*
WindDir3pm: N	8890	*6.295%*
WindDir3pm: WNW	8874	*6.283%*
WindDir3pm: NW	8610	*6.096%*
WindDir3pm: ESE	8505	*6.022%*
WindDir3pm: E	8472	*5.999%*
WindDir3pm: NE	8263	*5.851%*
WindDir3pm: SSW	8156	*5.775%*
WindDir3pm: NNW	7870	*5.572%*
WindDir3pm: ENE	7857	*5.563%*
WindDir3pm: NNE	6590	*4.666%*
Rain Today: No	110319	*77.581%*
Rain Today: Yes	31880	*22.419%*
Rain Tomorrow: No	110316	*77.582%*
Rain Tomorrow: Yes	31877	*22.418%*

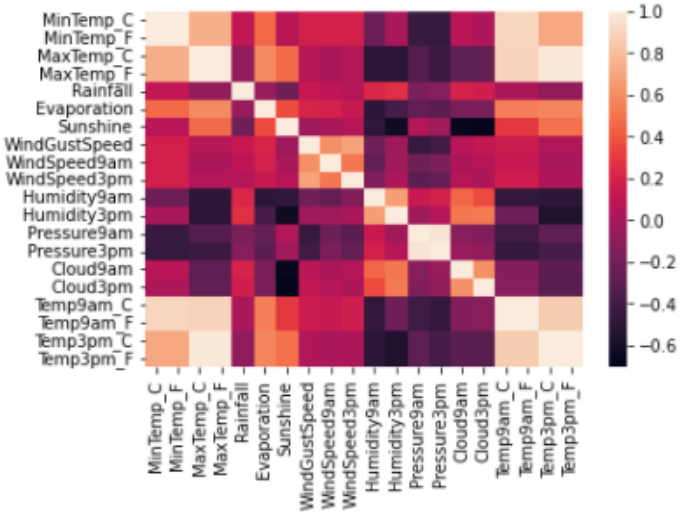
After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn't make sense)

Table 4: Correlation Table/Tables

[7]:

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
MinTemp	1.000000	0.736555	0.103938	0.466993	0.072586	0.177415	0.175064	0.175173	-0.232899	0.006089	-0.450970	-0.461292	0.078754	0.021605	0.901821	0.708906
MaxTemp	0.736555	1.000000	-0.074992	0.587932	0.470156	0.067615	0.014450	0.050300	-0.504110	-0.508855	-0.332061	-0.427167	-0.289370	-0.277921	0.887210	0.984503
Rainfall	0.103938	-0.074992	1.000000	-0.064351	-0.227549	0.133659	0.087338	0.057887	0.224405	0.255755	-0.168154	-0.126534	0.198528	0.172403	0.011192	-0.079657
Evaporation	0.466993	0.587932	-0.064351	1.000000	0.365602	0.203021	0.193084	0.129400	-0.504092	-0.390243	-0.270362	-0.293581	-0.183793	-0.182618	0.545115	0.572893
Sunshine	0.072586	0.470156	-0.227549	0.365602	1.000000	-0.034750	0.005499	0.053834	-0.490819	-0.629130	0.041970	-0.019719	-0.675323	-0.703930	0.291188	0.490501
WindGustSpeed	0.177415	0.067615	0.133659	0.203021	-0.034750	1.000000	0.605303	0.686307	-0.215070	-0.026327	-0.458744	-0.413749	0.071736	0.109168	0.150150	0.032748
WindSpeed9am	0.175064	0.014450	0.087338	0.193084	0.005499	0.605303	1.000000	0.519547	-0.270858	-0.031614	-0.228743	-0.175817	0.025112	0.054639	0.128545	0.004569
WindSpeed3pm	0.175173	0.050300	0.057887	0.129400	0.053834	0.686307	0.519547	1.000000	-0.145525	0.016432	-0.296351	-0.255439	0.053337	0.025396	0.163030	0.027778
Humidity9am	-0.232899	-0.504110	0.224405	-0.504092	-0.490819	-0.215070	-0.270858	-0.145525	1.000000	0.666949	0.139442	0.186858	0.452297	0.357326	-0.471354	-0.498399
Humidity3pm	0.006089	-0.508855	0.255755	-0.390243	-0.629130	-0.026327	-0.031614	0.016432	0.666949	1.000000	-0.027544	0.051997	0.517120	0.523120	-0.221019	-0.557841
Pressure9am	-0.450970	-0.332061	-0.168154	-0.270362	0.041970	-0.458744	-0.228743	-0.296351	0.139442	-0.027544	1.000000	0.961326	-0.129796	-0.147861	-0.422556	-0.286770
Pressure3pm	-0.461292	-0.427167	-0.126534	-0.293581	-0.019719	-0.413749	-0.175817	-0.255439	0.186858	0.051997	0.961326	1.000000	-0.060772	-0.084778	-0.470187	-0.389548
Cloud9am	0.078754	-0.289370	0.198528	-0.183793	-0.675323	0.071736	0.025112	0.053337	0.452297	0.517120	-0.129796	-0.060772	1.000000	0.603564	-0.136959	-0.302060
Cloud3pm	0.021605	-0.277921	0.172403	-0.182618	-0.703930	0.109168	0.054639	0.025396	0.357326	0.523120	-0.147861	-0.084778	0.603564	1.000000	-0.126659	-0.317420
Temp9am	0.901821	0.887210	0.011192	0.545115	0.291188	0.150150	0.128545	0.163030	-0.471354	-0.221019	-0.422556	-0.470187	-0.136959	-0.126659	1.000000	0.860591
Temp3pm	0.708906	0.984503	-0.079657	0.572893	0.490501	0.032748	0.004569	0.027778	-0.498399	-0.557841	-0.286770	-0.389548	-0.302060	-0.317420	0.860591	1.000000

Figure 1: Heatmap of the correlation matrix



IV. DATA SET GRAPHICAL EXPLORATION

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

A. Distributions

From Figure 2, we can see that in the year 2102 had the lowest average temperature. From 2010 to 2011, the average temperature stays at a steady level. The year 2014 had the highest average temperature.

From Figure 3, 2014 had the lowest average rainfall. The higher temperatures in 2014 could have resulted in less rain fall. The same does not hold true for the opposite. The lowest yearly average temperature (2012) is not the highest rainfall year but rather 2011 has the highest rainfall.

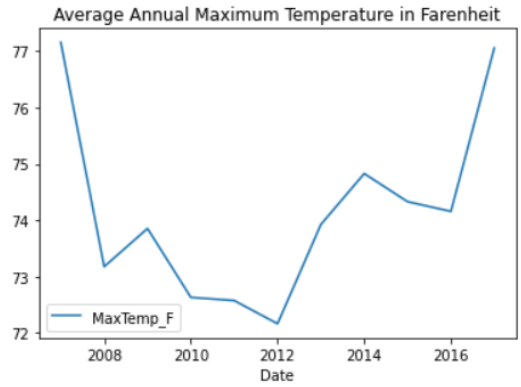


Figure 2: The Average Maximum Temperature for each year

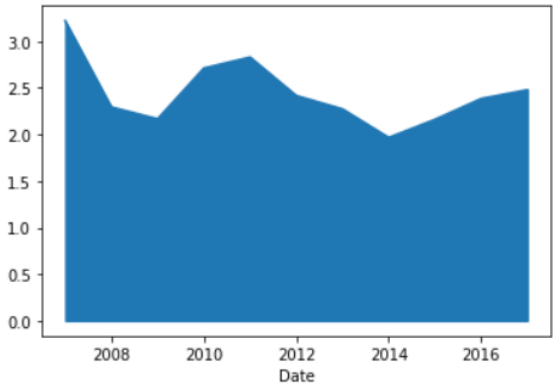


Figure 3: An area graph of the yearly rainfall

B. ScatterPlots / Pairwise Plots (continuous variables)

In Figure 4, we see comparisons for the cities Melbourne and Adelaide. These graphs compare rainfall, evaporation, and sunrise. Adelaide has a higher rain fall, while Melbourne has more evaporation and sunshine. It is scientific fact that a higher sunshine leads to a higher evaporation level. In the sunshine to rain fall graphs (top right corner and bottom left corner), it shows that the Adelaide plots are closer to sunshine axis with only two outliers while the Melbourne plots extends furthers then the Adelaide plots hand has more outliers. In the evaporation/sunshine (bottom middle and right column middle), the Melbourne plots mostly cover up by the majority of the Adelaide plots. It's a thick rectangle along the sunshine axis. Only Adelaide plots have outliers. The outliers have high levels of sunshine with high levels of evaporation. In the rainfall/evaporation graphs (top middle and left column middle), most of the plots create a triangle with 20 evaporation zero rainfall and zero evaporation and 40 rainfall for both cities. Melbourne's outliers have low evaporation and high rain fall. Adelaide's outliers have high evaporation and low rainfall.

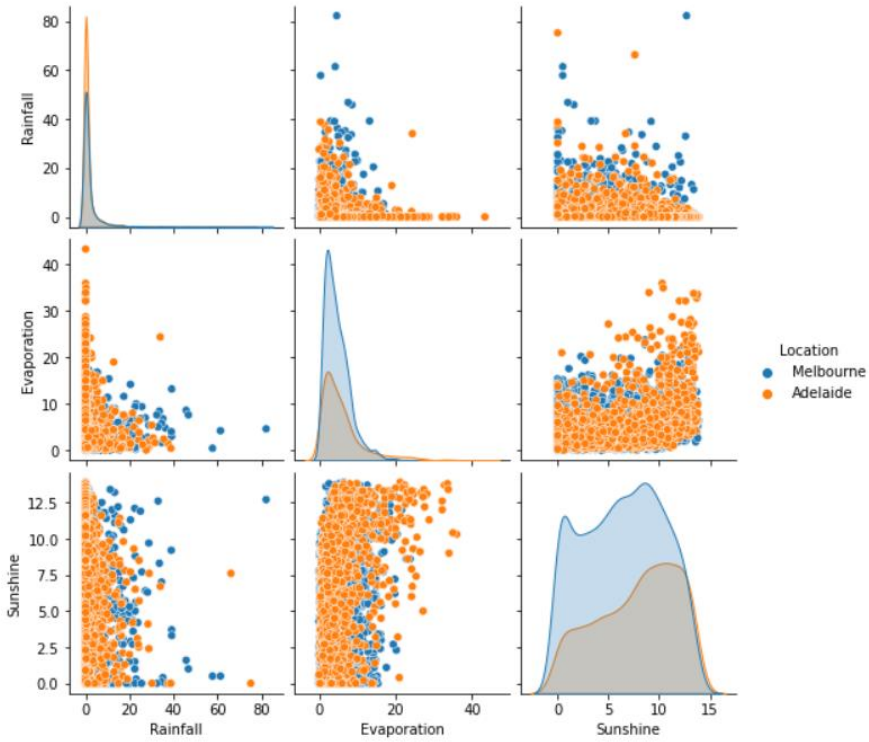


Figure 4: These 9 graphs compare the sunshine, evaporation, and rainfall in the cities of Melbourne and Adelaide

C. Barcharts and Histograms

In Figure 5, we see the average rain fall for each year. We can easily see that 2007 had the highest average and 2014 had the lowest average. We can better compare the averages in figure 5 than in figure 3. The majority fall between 2.0 and 2.5. The other three fall between 2.5 and 3.5.

It is hard to depict data from figure 6, because it is comprised of data from all of the cities. It is easier to read the graphs with not as much data. Since the data set contains 49 cities, we looked at just one city, Sydney. From figure 7, we see that the majority of Sydney's evaporation levels fall between 2.5 and 7.5. the histogram is skewed to the left with outliers beyond 17.5.

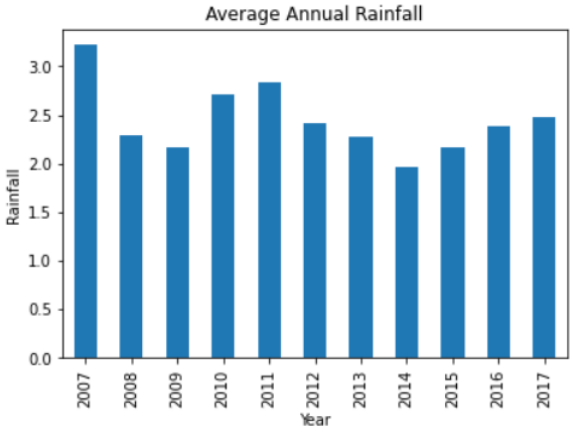


Figure 5: Bar chart of the average annual rainfall

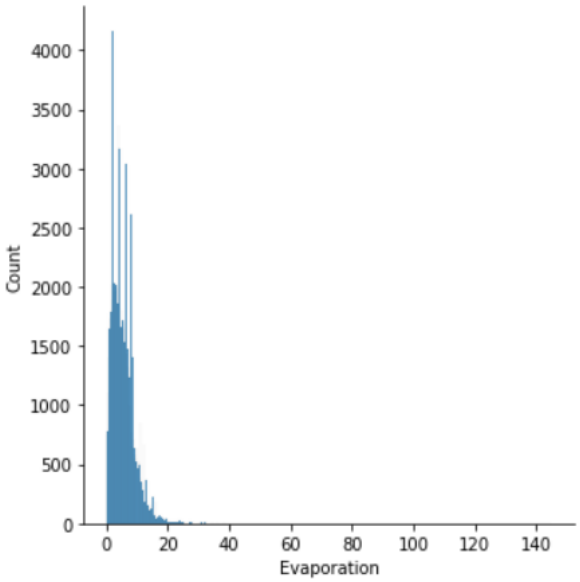


Figure 6: Histogram Shows the evaporation level for all of the Cities in the data set

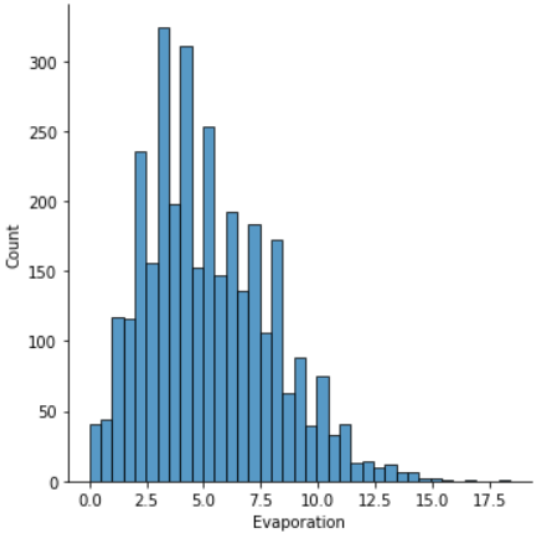


Figure 7: Histogram shows the evaporation levels of Sydney

D. Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.

This data set gave five variables that were measured at 9am and 3pm. In Figures 8, 9, 10, and 11, show these comparisons in box plots. Wind Direction was left out because it was a categorical variable and Cloud coverage was left out because 38% at 9 am and 40% at 3pm of the data was missing. In figure 8, MinTemp has no outliers and ranges from 40 degrees to just over 80 degrees with a mean of 58/59 degrees. MaxTemp has a few outliers the highest being over 100 degrees and all of them whiskers plot. MaxTemp has a mean of 73/74 degrees. This is over 10 degrees hotter than the MinTemp. It is reasonable that the MaxTemp mean is higher than the MinTemp mean. There are data points for both that fall between 55 and 81which I reasonable because the lowest temperature on the hottest day can be close to the highest temp on the coldest day. Figure 9, compares the temperatures in Sydney at 9am and 3pm. Both variables have outliers, but at 3pm there are a lot more outliers. All outliers fall above the box plots. Without outliers, 9am ranges from 43 to 91 degrees and 3pm ranges from 54 to 95 degrees. The range at 9 am is a lot bigger than at 3pm. Nine am has a mean of 65 and 3pm has a mean of 70. Even though their ranges are different, their means are relatively clothes to each other. In Figure 10, the humidity in Sydney is compared at 9am

and 3pm. The 25 to 75 percentile box is higher a up the graph at 9 am than at 3pm. It is known that mornings typically have a higher humidity. The outliers at 9am fall bellow the box plot and the outliers at 3pm are above and below the box lot. The mean at 9am is 70 and the mean at 3pm is 55. In figure 11, wind speed is compared at 9am and 3pm in Sydney. This is the first comparison where the ranges at the two times are similar. (Note: not the 25th and 75th percentile boxes). Both have a handful of outliers above the boxes and at 3pm there is one below the box. The mean at 9am is 15 and the mean at 3pm is 19. This comparison shows that wind speed is not effect much by the time of day like is shown in the other three graphs.

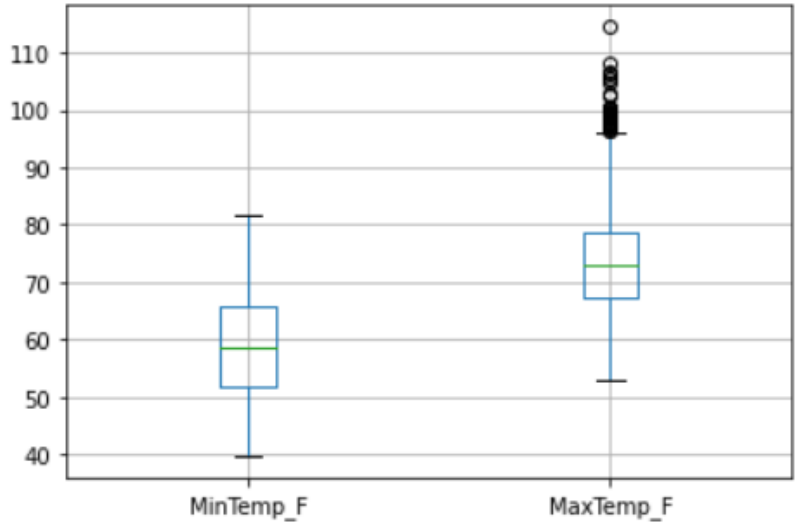


Figure 8: Box plot shows the difference in the minimum and maximum temperatures in Sydney

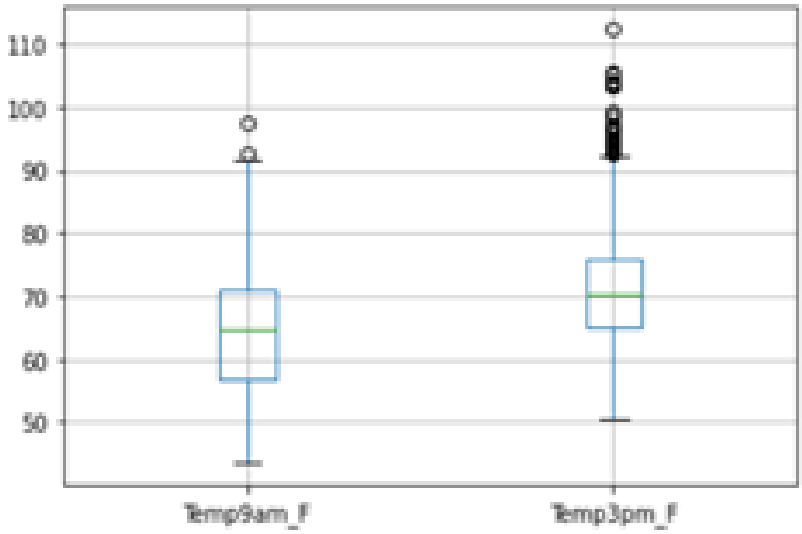


Figure 9: box plot compares the temperatures at 9am and 3pm in Sydney

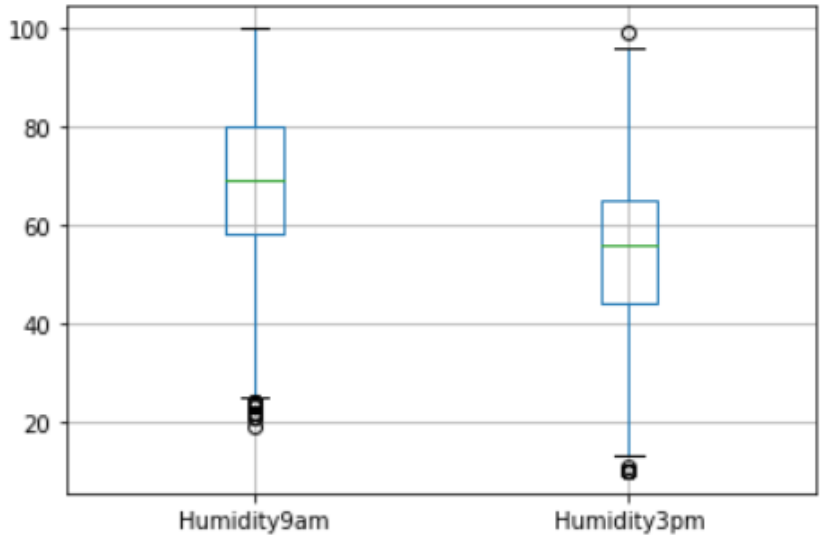


Figure 10: Box plot compares the Humidity in Sydney at 9am and 3pm

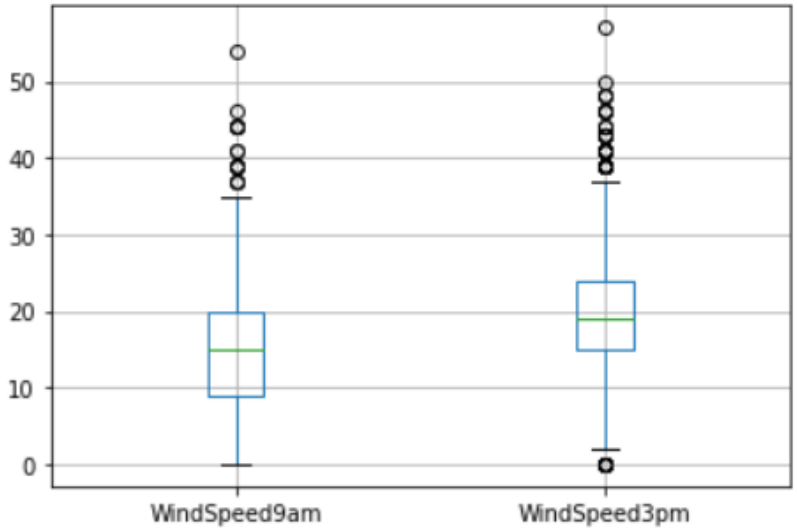


Figure 11: Box plot compares Wind speed at 9am and 3pm in Sydney

V. SUMMARY OF FINDINGS

After the exploratory analysis of the data set rainfall in Australia, multiple things can be concluded. The first being that this data set can be looked at in multiple ways. Cities Can be compared to each other with the continuous variables or we could look at just one city compare the different years of that city. the data can also be looked at by finding averages of each year or comparing years to each other. the same can be done with both months and days. this data set also had variables that were able to be compared to each other. variables were taken to be compared at 9:00 AM and 3:00 PM. By having these variables, comparisons were easier to be thought out.

From the exploratory analysis, conclusions were made about certain variables. A higher sunshine showed a higher evaporation level, which is supported with scientific fact. From 9:00 AM to 3:00 PM wind speed did not change throughout the day as much as humidity and temperature did. in the morning the humidity was higher and in the afternoon the temperature was higher, typically. While these conclusions came from just looking at Sydney, Australia similar conclusions would most likely occur in other cities around Australia.