

به نام خدا

داده‌کاوی

تمرین اول

دکتر شاکری

محمدامین عرب‌خراسانی

۸۱۰۱۰۲۲۰۵

۱۴۰۳
بهار

فهرست

3.....	پاسخ بخش تشریحی
3.....	سوال اول
4.....	سوال دوم
12.....	پاسخ بخش عملی
12.....	پیش‌پردازش
18.....	نمایش دادگان

پاسخ بخش تشریحی

سوال اول

- سن: بسته به نوع دیتا می‌تواند متفاوت باشد. اگر دیتای سن مثل ۱۰، ۳۵، ۹۰ و... سال باشد گسته می‌باشد. در غیر این صورت پیوسته است. (به صورت کلی پیوسته در نظر گرفته می‌شود) با فرض پیوسته بودن، می‌توان از Box plot و Histogram استفاده کرد.

* در صورتی که داده‌های پیوسته به بازه‌های مختلف تقسیم شوند (در این مثال ۰ تا ۱۰، ۱۰ تا ۲۰ و... سال) می‌توان از Pie chart هم استفاده کرد.

- جنسیت: باینری (متقارن) (مرد، زن) می‌توان از Bar chart و Pie chart استفاده کرد.

* میزان درآمد: پیوسته می‌توان از Box plot و Histogram استفاده کرد.
* در صورتی که داده‌های پیوسته به بازه‌های مختلف تقسیم شوند (در این مثال ۰ تا ۱۰، ۱۰ تا ۲۰ و... هزار دلار) می‌توان از Pie chart هم استفاده کرد.

- وضعیت تأهل: اسمی (مجرد، متاهل، مطلقه، بیوه) می‌توان از Bar chart و Pie chart استفاده کرد.

- فرزند دارد: باینری (نامتقارن) (ندارد صفر، دارد یک) می‌توان از Bar chart و Pie chart استفاده کرد.

- شغل: اسمی (مهندس، دکتر، نجار و...) می‌توان از Bar chart و Pie chart استفاده کرد.

- میزان تحصیلات: ترتیبی (دیپلم، کارشناسی، کارشناسی ارشد، دکترا و...)

می‌توان از Bar chart و Pie chart استفاده کرد.

- تعداد اعضای خانواده: گستره
- می‌توان از Histogram و Bar chart، Pie chart استفاده کرد.

سوال دوم

برای حل این سوال از زبان پایتون و کتابخانه‌های مورد نیاز استفاده شده است.

- ۱) در این قسمت توضیحات مربوط به محاسبه‌ی هر کدام از خواسته‌ها آورده شده است و یک مورد از آن‌ها برای نمونه محاسبه خواهد شد.

محاسبه‌ی میانگین:

برای محاسبه‌ی میانگین از رابطه‌ی زیر استفاده شده است.

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$$

به طور مثال میانگین مجموعه داده‌ی A به شکل زیر محاسبه می‌شود.

$$\bar{x} = \frac{55 + 72 + 60 + 54 + 42 + \dots + 38 + 20 + 27}{30}$$

$$\bar{x} = \frac{1522}{30} = 50.7333$$

جدول ۱ میانگین هر کدام از مجموعه داده‌ها را نمایش می‌دهد.

مجموعه داده	میانگین
A	50.7333
B	10.8

جدول ۱ - میانگین مجموعه داده‌ها

محاسبه‌ی میانه:

برای محاسبه‌ی میانگین ابتدا هر مجموعه داده از کوچک به بزرگ مرتب می‌شود. سپس داده‌ی وسط به عنوان میانه در نظر گرفته می‌شود. هر دو مجموعه شامل ۳۰ داده می‌باشند که با توجه به زوج بودن این تعداد، میانگین دو عضو مرکزی به عنوان میانه در نظر گرفته می‌شود.

به طور مثال برای محاسبه‌ی میانه‌ی مجموعه داده‌ی A، داده‌های مرکزی (داده‌ی ۱۵ و ۱۶) ۵۲ و ۵۲ می‌باشند که میانگین آن‌ها برابر ۵۲ خواهد بود.

جدول ۲ میانه‌ی هر دو مجموعه داده را نشان می‌دهد.

مجموعه داده	میانه
A	52
B	11

جدول ۲ - میانه‌ی دو مجموعه داده

محاسبه‌ی چارک اول:

برای محاسبه‌ی چارک اول از مجموعه داده‌ای که برای محاسبه‌ی میانه مرتب شده بود استفاده می‌شود. با توجه به این که میانه از داده‌ی ۱۵ و ۱۶ حاصل شد، برای محاسبه‌ی چارک اول از داده‌ی وسط ۱۵ داده‌ی اول استفاده می‌شود. با توجه به فرد بودن ۱۵، داده‌ی ۸ م معادل چارک اول خواهد بود. به طور مثال برای محاسبه‌ی چارک اول مجموعه داده‌ی A، داده‌ی ۸ م برابر ۳۸ می‌باشد.

جدول ۳ چارک اول هر دو مجموعه داده را نشان می‌دهد.

مجموعه داده	چارک اول
A	38
B	8

جدول ۳ - چارک اول دو مجموعه داده

محاسبه‌ی چارک سوم:

برای محاسبه‌ی چارک سوم از مجموعه داده‌ای که برای محاسبه‌ی میانه مرتب شده بود استفاده می‌شود. با توجه به این که میانه از داده‌ی ۱۵ و ۱۶ حاصل شد، برای محاسبه‌ی چارک سوم از داده‌ی وسط ۱۵ داده‌ی دوم استفاده می‌شود. با توجه به فرد بودن ۱۵، داده‌ی ۲۳ م معادل چارک سوم خواهد بود. به طور مثال برای محاسبه‌ی چارک اول مجموعه داده‌ی A، داده‌ی ۲۳ م برابر ۶۸ می‌باشد.

جدول ۳ چارک اول هر دو مجموعه داده را نشان می‌دهد.

مجموعه داده	چارک سوم
A	68
B	14

جدول 4 - چارک سوم دو مجموعه داده

محاسبه انحراف معیار:

برای محاسبه انحراف معیار از رابطه‌ی زیر استفاده شده است:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

به طور مثال انحراف معیار مجموعه داده‌ی A به صورت زیر محاسبه می‌شود.

$$\sigma = \sqrt{\frac{(55 - 50.7333)^2 + (72 - 50.7333)^2 + \dots + (27 - 50.7333)^2}{30}}$$

$$\sigma = 25.7862$$

جدول 5 انحراف معیار هر کدام از مجموعه داده‌ها را نمایش می‌دهد.

مجموعه داده	انحراف معیار
A	25.7862
B	5.3938

جدول 5 - انحراف معیار دو مجموعه داده

۲) ابتدا برای رسم نمودار جعبه‌ای بررسی می‌شود که آیا داده‌ی پرت در مجموعه داده‌ها وجود دارد یا خیر. برای بررسی این موضوع با توجه به محاسبه‌ی چارک اول و سوم در هر مجموعه داده، IQR هر مجموعه داده از طریق رابطه‌ی زیر محاسبه خواهد شد.

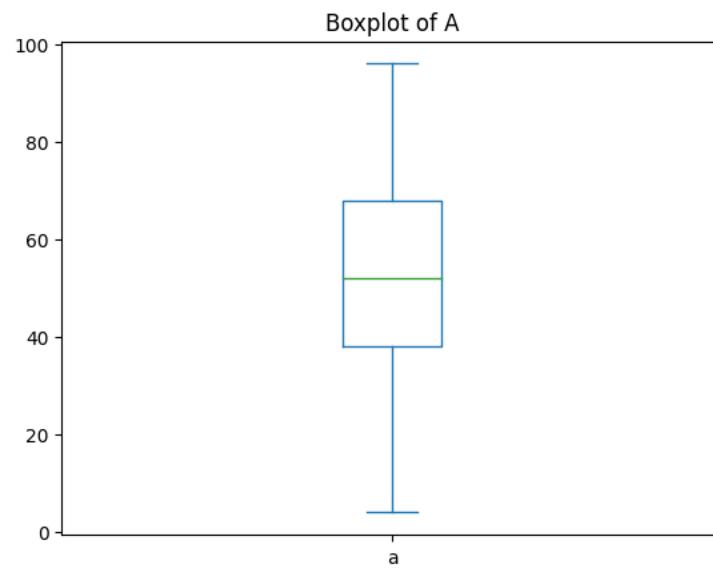
$$IQR = Q_3 - Q_1$$

مقدار IQR برای هر مجموعه داده در جدول 6 آورده شده است.

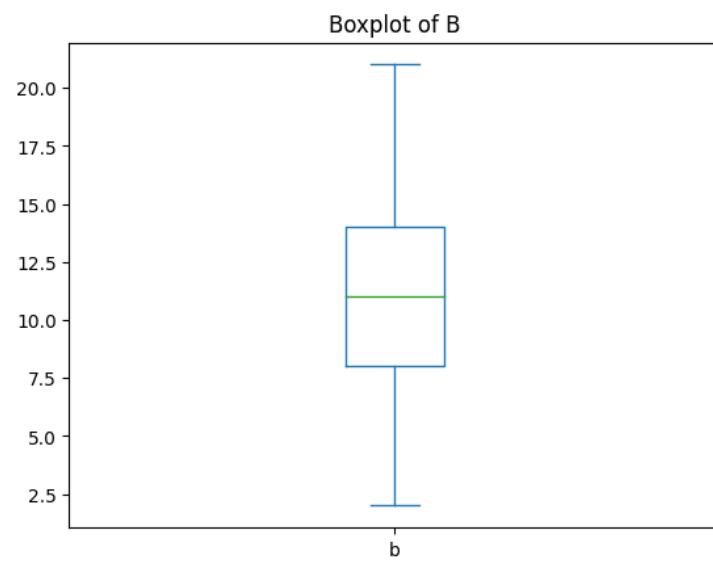
مجموعه داده	IQR
A	30
B	6

جدول 6 - مقدار IQR برای دو مجموعه داده

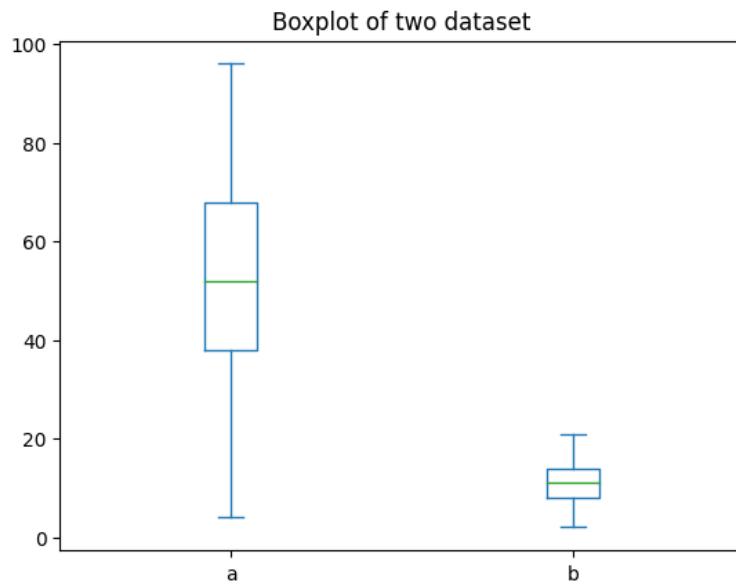
حال برای تشخیص داده‌های پرت می‌بایست برسی شود که آیا داده‌های موجود در مجموعه داده در فاصله‌ی ۱.۵ برابر IQR از چارک بالا و پایین قرار دارد یا خیر. با برسی این موضوع برای دو مجموعه داده مشخص می‌شود هیچ داده‌ای در این دو مجموعه داده، داده‌ی پرت نمی‌باشد. نمودار جعبه‌ای به صورت جداگانه و ترکیبی برای هر دو مجموعه داده به کمک پایتون رسم شده است. شکل ۱ و ۲ به ترتیب نمودار مجموعه داده‌ای A و B را نشان می‌دهند. شکل ۳ نیز این نمودار را برای هر دو مجموعه داده در کنار هم نشان می‌دهد.



شکل ۱ - نمودار جعبه‌ای مجموعه داده‌ی A



شکل 2 - نمودار جعبه‌ای مجموعه داده‌ی B



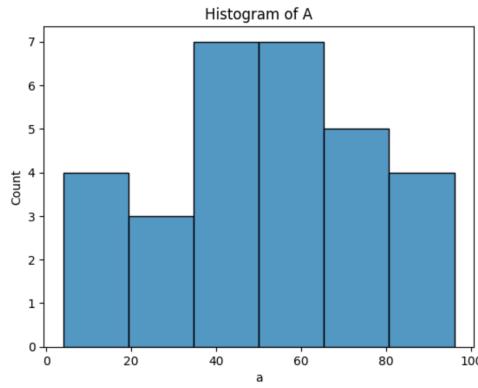
شکل ۳ - نمودار جعبه‌ای دو مجموعه داده A و B

با توجه به شکل ۳ می‌توان نتیجه گرفت با مقایسه بین میزان IQR (کشیدگی جعبه‌ها) هر مجموعه داده و ابتدا و انتهای دیتا در آن‌ها پراکندگی داده‌ها در مجموعه داده A از B بیشتر است.

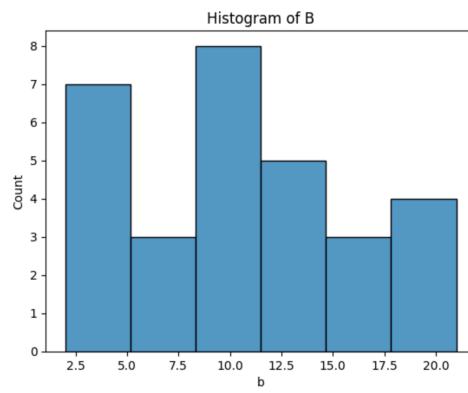
۳) برای رسم نمودار Histogram از کتابخانه seaborn در پایتون استفاده شده است. در رسم Histogram برای داده‌های گستته ابتدا نیاز هست که عرض bin‌ها انتخاب شود. عرض این bin‌ها در کتابخانه ذکر شده به صورت پیش‌فرض در صورتی که مقداری به آن داده نشود از طریق قانون Freedman-Diaconis محاسبه می‌شود. این قانون اشاره می‌کند که به دلیل کمینه کردن انتگرال مریع تفاوت بین Histogram و چگالی توزیع احتمال، بهترین اندازه برای عرض bin را پیشنهاد می‌دهد که از طریق رابطه‌ی زیر محاسبه می‌شود.

$$\text{Bin width} = 2 \frac{IQR(x)}{\sqrt[3]{n}}$$

بنابراین برای هر کدام از مجموعه داده‌ها عرض bin محاسبه می‌شود. شکل ۴ و ۵ به ترتیب نمودار Histogram برای مجموعه داده‌های a و b می‌باشد.



شکل 4 - نمودار Histogram برای مجموعه دادهی A

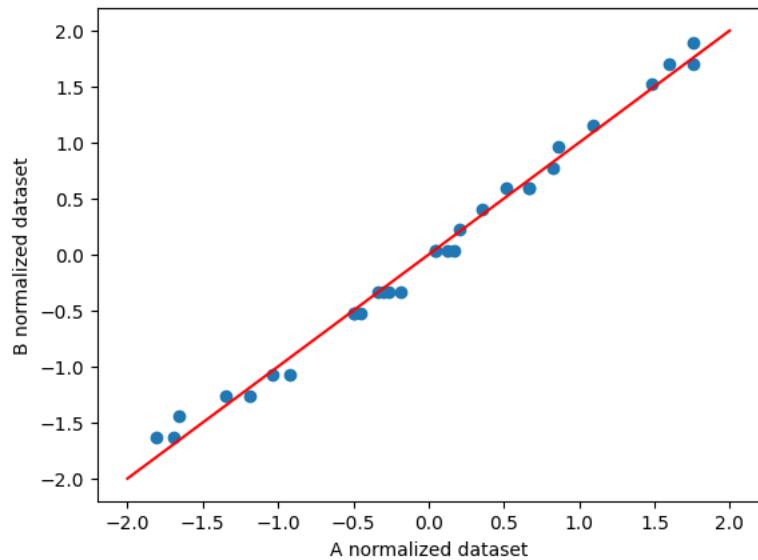


شکل 5 - نمودار Histogram برای مجموعه دادهی B

۴) برای محاسبه‌ی z-score از رابطه‌ی زیر استفاده می‌شود.

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}$$

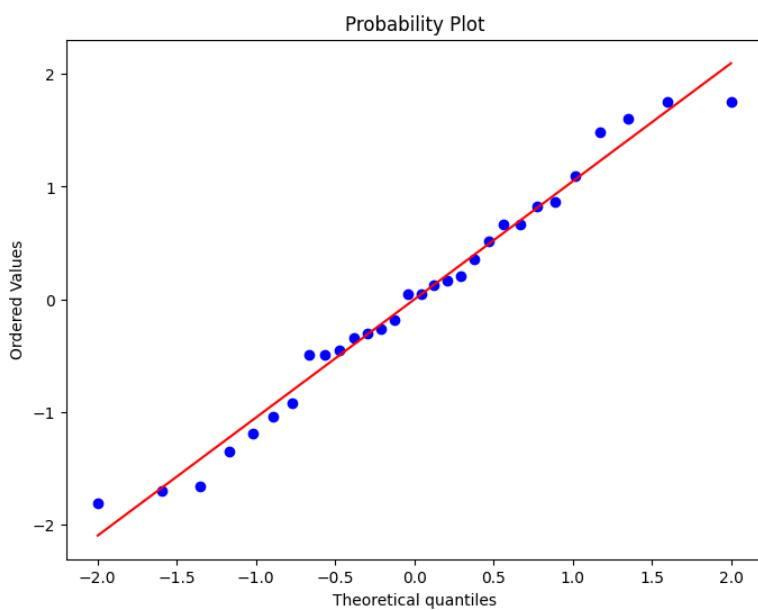
سپس داده‌هایی که نرمال شده‌اند از بزرگ به کوچک مرتب می‌شوند و نمودار Q-Q plot رسم می‌شود. در دو حالت می‌توان توزیع این مجموعه داده‌ها را با یکدیگر مقایسه کرد. حالت اول مقایسه‌ی داده‌های نرمال شده با یکدیگر است که توزیع دو مجموعه دادهی A و B را با یکدیگر مقایسه می‌کند. شکل ۶ این نمودار را برای مجموعه داده‌های نرمال شدهی A و B نشان می‌دهد.



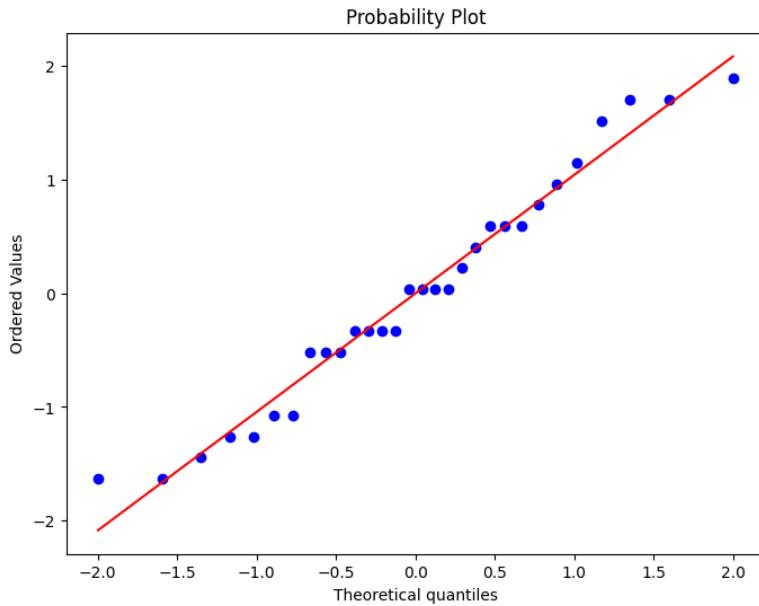
شکل 6 - نمودار Q-Q plot برای دو مجموعه داده نرمال شده

همانطور که از شکل ۶ برمی‌آید، با توجه به نزدیکی این دو داده به خط با زاویه‌ی ۴۵ درجه، توزیع دو مجموعه داده به هم شبیه است و توزیع‌ها یکسان هستند.

حالت دوم به این صورت است که هر مجموعه داده نرمال شده به صورت مجزا با یکتابع احتمال با توزیع نرمال مقایسه می‌شود. شکل‌های ۷ و ۸ به ترتیب نمودار Q-Q plot را برای مجموعه داده‌ی A و B نشان می‌دهند.



شکل 7 - نمودار Q-Q plot برای مجموعه داده‌ی A و تابع احتمال



شکل 8 - نمودار Q-Q plot برای مجموعه داده‌ی B و تابع احتمال

با توجه به شکل‌های ۷ و ۸، نزدیکی این دو مجموعه داده به خط با شیب ۴۵ درجه این نتیجه را می‌دهد که توزیع این دو مجموعه داده نرمال می‌باشد.

۵) برای بررسی همبستگی یا عدم همبستگی این دو مجموعه داده از رابطه‌ی زیر استفاده می‌شود.

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B}$$

مقدار $r_{A,B}$ بین -۱ و ۱ قرار دارد. هر چه این مقدار به ۱ نزدیک‌تر باشد به این معناست که دو مجموعه داده به صورت مثبت و قوی با یکدیگر رابطه همبستگی دارند. هر چه این مقدار به -۱ منفی نزدیک‌تر باشد به این معناست که همبستگی قوی معکوس دارند. هر چه به صفر نزدیک‌تر باشد به این معناست که این همبستگی ضعیفتر است.

این ضریب برای دو مجموعه داده برابر با 0.9956 است. به این معنا که این دو مجموعه داده به صورت مثبت و قوی با یکدیگر همبستگی دارند.

پاسخ بخش عملی

پیش‌پردازش

۱) در مرحله‌ی اول دیتاست بررسی شده است. این دیتاست شامل ۴۹ فایل CSV می‌باشد که هر کدام از این فایل‌ها مربوط به لوکیشن متفاوتی از استرالیا می‌باشد. نام این فایل‌ها از لوکیشن آن‌ها گرفته شده است. همانطور که در صورت سوال ذکر شده است این دیتاست به صورت کلی باید شامل ۲۳ ویژگی باشد که هر کدام از این ویژگی‌ها، پارامتری از وضعیت آب و هوای را مشخص می‌کند. نام این ستون‌ها به همراه توضیحات مربوط به آن در جدول ۷ آورده شده است.

#	Column name	Description
1	Date	The date of observation
2	Location	The common name of the location of the weather station
3	MinTemp	The minimum temperature in degrees Celsius or Fahrenheit
4	MaxTemp	The maximum temperature in degrees Celsius or Fahrenheit
5	Rainfall	The amount of rainfall recorded for the day in mm
6	Evaporation	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
7	Sunshine	The number of hours of bright sunshine in the day.
8	WindGustDir	The direction of the strongest wind gust in the 24 hours to midnight
9	WindGustSpeed	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
10	WindDir9am	Direction of the wind at 9am
11	WindDir3pm	Wind speed (km/hr) averaged over 10 minutes prior to 9am
12	WindSpeed9am	Wind speed (km/hr) averaged over 10 minutes prior to 9am
13	WindSpeed3pm	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
14	Humidity9am	Humidity (percent) at 9am
15	Humidity3pm	Humidity (percent) at 3pm
16	Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9am
17	Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
18	Cloud9am	Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eighths. It records how many eighths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
19	Cloud3pm	Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cloud9am for a description of the values
20	Temp9am	Temperature (degrees C) at 9am
21	Temp3pm	Temperature (degrees C) at 3pm
22	RainToday	Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
23	RainTomorrow	The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".

جدول ۷ - توضیحات مجموعه دادگان

با بررسی دیتاست داده شده مشخص می‌شود هر کدام از این فایل‌ها حداقل ۲۳ ستون دارند و ممکن است بعضی از ۲۳ ستون تعریف شده در فایل‌ها نباشد. همچنین نام برخی از ستون‌ها دقیقاً با نام‌های ذکر شده در جدول ۷ یکسان نیست که می‌بایست نام این ستون‌ها تغییر کند و به نام‌هایی که در جدول ۷ آورده شده است تبدیل شوند.

برای بطرف کردن این مورد، تمامی اسم‌های ستون‌هایی که در دیتاست وجود دارند گرفته می‌شود تا با تمامی اسم‌های متفاوت روپرتو شویم. به تعداد ۱۶۰ ستون با نام‌های متفاوت وجود دارد که می‌بایست به نام درست تغییر کنند. در این اسم‌ها مواردی نظری و . و فاصله باعث تغییر کردن اسم ستون‌ها شده است. بعضی از این نام‌های اشتباہ replace pressure.3.pm rain today، wind_speed_3_pm موارد اضافه حذف می‌شوند.

همچنین مشاهده می‌شود که برخی از نام ستون‌ها به صورت حروف کوچک می‌باشند. برای حل این مشکل ابتدا تمامی حروف تمامی نام ستون‌ها کوچک می‌شوند سپس در انتهای با کمک لیست true_col_name که همان نام‌های موجود در جدول ۷ می‌باشند به نام استاندارد بازگردانی خواهد شد.

همچنین در انتهای برخی از نام ستون‌های مربوط به دما حرف F و C هم به صورت بزرگ هم به صورت کوچک وجود دارد. برای حل این مشکل از مفهوم regex استفاده شده است. C و F بیانگر واحد دما می‌باشند. برای تبدیل فارنهایت به درجهی سلسیوس نیز از رابطه‌ی زیر استفاده شده است.

$$c = \frac{5}{9}(F - 32)$$

در نهایت با for روی فایل‌های csv که آپلود شده‌اند و ایجاد تغییرات اعمال اشاره شده هر دیتافریم به یک دیتافریم اصلی اضافه می‌شود که تمامی عملیات‌ها برای ادامه‌ی سوال روی این دیتافریم خواهد بود. این دیتافریم شامل ۱۴۵۴۶ ردیف و ۲۳ ستون می‌باشد.

همچنین لازم به ذکر است که مقادیر ستون‌های فایل‌هایی که یک ستون مشخص را نداشته باشند در دیتافریم نهایی با NaN به صورت پیش‌فرض ذخیره می‌شوند.

(۲) شکل ۹ نتیجه‌ی اجرای تابع head برای دیتافریم می‌باشد که ۵ ردیف اول این دیتافریم را نشان می‌دهد. همچنین در فایل آپلود شده نیز این شکل رسم شده است.

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
0	2013-03-01	Katherine	24.7	34.6	0.0	4.0	NaN	WNW	43.0	NW	...	74.0	49.0	1008.5	1005.0	7.0	3.0	28.7	34.1	No	No
1	2013-03-02	Katherine	25.2	31.1	0.8	6.4	NaN	WNW	26.0	WNW	...	88.0	77.0	1008.6	1004.1	8.0	8.0	27.0	29.0	No	No
2	2013-03-03	Katherine	25.4	35.9	0.0	3.8	NaN	S	48.0	WSW	...	82.0	55.0	1004.1	999.6	6.0	6.0	29.2	35.1	No	Yes
3	2013-03-04	Katherine	21.6	33.0	81.0	NaN	NaN	W	37.0	NW	...	95.0	64.0	1005.1	1002.0	8.0	5.0	25.1	31.2	Yes	No
4	2013-03-05	Katherine	24.9	NaN	0.0	4.0	NaN	NW	30.0	WNW	...	89.0	61.0	1007.1	1003.8	7.0	3.0	26.9	33.6	No	No

شکل ۹ - نمونه‌ای از ۵ ردیف اول دیتافریم نهایی

۳) برای به دست آوردن نوع دیتاهای موجود در یک دیتافریم از تابع `dtypes` استفاده می‌شود. نتیجه حاصل از اجرای این دستور در زیر آورده شده است.

```

Date          object
Location      object
MinTemp       float64
MaxTemp       float64
Rainfall      float64
Evaporation   float64
Sunshine      float64
WindGustDir   object
WindGustSpeed float64
WindDir9am    object
WindDir3pm    object
WindSpeed9am  float64
WindSpeed3pm  float64
Humidity9am   float64
Humidity3pm   float64
Pressure9am   float64
Pressure3pm   float64
Cloud9am      float64
Cloud3pm      float64
Temp9am       float64
Temp3pm       float64
RainToday     object
RainTomorrow  object
dtype: object

```

۴) برای گرد کردن اعداد مربوط به دما، از تابع `round` استفاده می‌شود. برای اعمال این دستور فقط روی اعداد مربوط به دما بررسی می‌شود که عبارت `Temp` در نام کدام ستون‌ها وجود دارد. برای این منظور از `if` استفاده می‌شود. مقادیر مربوط به آن ستون‌ها گرد می‌شوند.

۵) با استفاده از تابع `info` اطلاعات مربوط به دیتاست حاصل می‌شود. نتیجه‌های اجرای این دستور در زیر آورده شده است.

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 145460 entries, 0 to 3435
Data columns (total 23 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Date        145460 non-null  object 
 1   Location    145460 non-null  object 
 2   MinTemp     143975 non-null  float64
 3   MaxTemp     144199 non-null  float64
 4   Rainfall    142199 non-null  float64
 5   Evaporation 82679 non-null  float64
 6   Sunshine    75625 non-null  float64
 7   WindGustDir 135134 non-null  object 
 8   WindGustSpeed 135197 non-null  float64
 9   WindDir9am  134894 non-null  object 
 10  WindDir3pm  141232 non-null  object 
 11  WindSpeed9am 143693 non-null  float64
 12  WindSpeed3pm 142398 non-null  float64
 13  Humidity9am 142806 non-null  float64
 14  Humidity3pm 140953 non-null  float64
 15  Pressure9am 130395 non-null  float64
 16  Pressure3pm 130432 non-null  float64
 17  Cloud9am    89572 non-null  float64
 18  Cloud3pm    86102 non-null  float64
 19  Temp9am     143693 non-null  float64
 20  Temp3pm     141851 non-null  float64
 21  RainToday   142199 non-null  object 
 22  RainTomorrow 142193 non-null  object 
dtypes: float64(16), object(7)
memory usage: 26.6+ MB

```

همانطور که مشخص است، حافظه‌ای که دیتاست اشغال می‌کند ۲۶.۶ مگابایت می‌باشد.

۶) کتابخانه‌ی pandas به گونه‌ای طراحی شده است که داده‌های category را به یک object بسیار کم حافظه تبدیل می‌کند. وقتی یک ستون به عنوان داده‌های اسمی نمایش داده می‌شود، هر مقدار در ستون به عنوان یک object جداگانه ذخیره می‌شود که باعث مصرف زیادی از حافظه شود، به خصوص زمانی که ستون تعداد زیادی مقدار یونیک دارد یا زمانی که stringها طولانی هستند. اما وقتی یک ستون به عنوان داده‌های category نمایش داده می‌شود، از یک نمایش عددی استفاده می‌شود که هر عدد، مربوط به یک string است. ذخیره کردن داده‌ها به صورت اعداد صحیح نسبت به رشته‌ها حافظه‌ی کمتری اشغال می‌کند.

برای این دیتا است با توجه به قسمت ۳ می‌توان داده‌ای که به صورت object هستند را به داده‌های category تبدیل کرد. داده‌های مربوط به RainToaday، Location، WindDir3pm، WindDir9am و RainTomorrow را به داده‌های category تبدیل کرد.

به کمک تابع astype داده‌های ذکر شده تبدیل به داده‌های category می‌شوند. نتیجه‌ی حاصل به کمک تابع info در زیر آورده شده است.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 145460 entries, 0 to 3435
Data columns (total 23 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Date        145460 non-null  object  
 1   Location    145460 non-null  category
 2   MinTemp    143975 non-null  float64 
 3   MaxTemp    144199 non-null  float64 
 4   Rainfall    142199 non-null  float64 
 5   Evaporation 82670  non-null  float64 
 6   Sunshine    75625  non-null  float64 
 7   WindGustDir 135134 non-null  category
 8   WindGustSpeed 135197 non-null  float64 
 9   WindDir9am  134894 non-null  category
 10  WindDir3pm  141232 non-null  category
 11  WindSpeed9am 143693 non-null  float64 
 12  WindSpeed3pm 142398 non-null  float64 
 13  Humidity9am 142806 non-null  float64 
 14  Humidity3pm  140953 non-null  float64 
 15  Pressure9am 130395 non-null  float64 
 16  Pressure3pm 130432 non-null  float64 
 17  Cloud9am    89572  non-null  float64 
 18  Cloud3pm    86102  non-null  float64 
 19  Temp9am     143693 non-null  float64 
 20  Temp3pm     141851 non-null  float64 
 21  RainToday   142199 non-null  category
 22  RainTomorrow 142193 non-null  category
dtypes: category(6), float64(16), object(1)
memory usage: 20.8+ MB
```

۷) با توجه به استفاده تابع info در قسمت قبل، مقدار حافظه اشغال شده برابر ۲۰.۸ مگابایت خواهد بود. در نتیجه نسبت به حالی که داده‌ها category نباشند ۲۱.۸ درصد حافظه اشغال شده کمتر می‌شود.

۸) به کمک دو تابع sum و isna ابتدا دیتافریم مربوط به مقادیر NaN حاصل می‌شود و در ادامه تعداد آن‌ها برای هر ستون گزارش می‌شود. نتیجه‌ی این گزارش در زیر آورده شده است.

```

Date          0
Location      0
MinTemp       1485
MaxTemp       1261
Rainfall       3261
Evaporation   62790
Sunshine        69835
WindGustDir    10326
WindGustSpeed  10263
WindDir9am     10566
WindDir3pm      4228
WindSpeed9am    1767
WindSpeed3pm    3062
Humidity9am     2654
Humidity3pm     4507
Pressure9am     15065
Pressure3pm     15028
Cloud9am        55888
Cloud3pm        59358
Temp9am         1767
Temp3pm          3609
RainToday        3261
RainTomorrow     3267
dtype: int64

```

۹) به طور کلی ۶ روش رویکرد در برخورد با مقادیر از دست رفته می‌توان اتخاذ کرد.

- نادیده گرفتن مقادیر از دست رفته
- پر کردن مقادیر از دست رفته به صورت دستی
- استفاده از یک global constant با جای مقادیر از دست رفته
- استفاده از مقادیری نظیر میانه و میانگین برای مقادیر از دست رفته
- استفاده از مقادیری نظیر میانه و میانگین برای کل مقادیری که با مقادیر از دست رفته در یک دسته‌بندی هستند
- استفاده از مقادیر احتمالی برای مقادیر از دست رفته نظیر درخت تصمیم‌گیری و رگرسیون

در این دیتابست با توجه به تعداد مقادیر از دست رفته برای هر ویژگی، روش متفاوتی اتخاذ می‌شود.

برای ویژگی‌هایی که بیشتر از ۱۰۰۰۰ داده‌ی از دست رفته دارند مقادیر از دست رفته با میانگین جایگزین می‌شوند و برای ویژگی‌هایی که کمتر از این مقدار داده‌ی از دست رفته دارند، از آن‌ها صرف نظر می‌شود و از دیتابست حذف می‌شوند. با توجه به این که ستون‌هایی که تعداد داده‌های از دست رفته‌ی زیادی دارند زیاد است، حذف کردن ویژگی‌های دیگر همراه با این مقدار داده‌ی زیاد باعث از بین رفتن دیتابست می‌شود. اما با حذف ویژگی‌هایی که تعداد داده‌ی از دست رفته‌ی کمی دارند مشکلی برای دیتابست ایجاد نمی‌شود. گزارشی از تعداد NaN‌ها در هر ستون در زیر آورده شده است.

```

Date          0
Location      0
MinTemp       1485
MaxTemp       1261
Rainfall       3261
Evaporation   62790
Sunshine        69835
WindGustDir    10326
WindGustSpeed  10263
WindDir9am     10566
WindDir3pm      4228
WindSpeed9am    1767
WindSpeed3pm    3062
Humidity9am     2654
Humidity3pm     4507
Pressure9am     15065
Pressure3pm     15028
Cloud9am        55888
Cloud3pm        59358
Temp9am         1767
Temp3pm          3609
RainToday        3261
RainTomorrow     3267
dtype: int64

```

همانطور که از گزارش مشخص است WindDir9am و WindGustDir تعداد داده‌ی categorical به عنوان یک داده‌ی WindDir9am است. از آنجایی که برای داده‌های categorical میانگین تعريف نمی‌شود که جایگزین مقادیر از رفته بیشتری ۱۰۰۰۰ تا دارند. از آنجایی که برای داده‌های categorical میانگین تعريف نمی‌شود که جایگزین مقادیر از دست رفته شود این داده‌ها یا باید حذف شوند یا با مد جایگزین شوند. از آنجایی که ۱۰۰۰۰ داده برای ۲۲ ویژگی دیگر با حذف این مقادیر حذف می‌شوند، برای حفظ ارزش دیتابست این مقادیر از دست رفته با مد جایگزین می‌شوند. به علاوه‌ی این که چون در این سوال پیش‌بینی مورد خاصی بر اساس این ویژگی‌ها مطرح نیست و در ادامه‌ی سوال به صورت مستقیم با از ویژگی WindDir9am و WindGustDir استفاده نمی‌شود، جایگزینی مقادیر از دست رفته با مد، تصمیم خوبی می‌باشد.

۱۰) با توجه به توضیحات بخش قبل، کد مربوط به این مرحله با حلقه‌ی for و جملات شرطی if و توابع fillna و dropna پیاده‌سازی می‌شود. بعد از حذف شدن یا جایگزینی مقادیر از دست رفته، تعداد NaN‌ها باید صفر باشد. گزارش زیر تعداد NaN‌ها بعد از این عملیات‌ها می‌باشد. بعد از این مرحله، تعداد داده‌ها از ۱۴۵۴۶ به ۱۳۴۵۹ می‌رسد.

```

Date          0
Location      0
MinTemp       0
MaxTemp       0
Rainfall       0
Evaporation   0
Sunshine       0
WindGustDir    0
WindGustSpeed  0
WindDir9am     0
WindDir3pm     0
WindSpeed9am   0
WindSpeed3pm   0
Humidity9am    0
Humidity3pm    0
Pressure9am    0
Pressure3pm    0
Cloud9am       0
Cloud3pm       0
Temp9am        0
Temp3pm        0
RainToday      0
RainTomorrow   0
dtype: int64

```

۱۱) این قسمت در بخش اول آورده شده است.

۱۲) برای شناسایی داده‌های پرت از IQR استفاده می‌شود. به این صورت که ابتدا Q1 و Q3 برای هر ستون در یک حلقه‌ی for روی ستون‌ها با شرط عددی بودن محاسبه می‌شود. سپس IQR محاسبه می‌شود. با توجه به نوع دیتابست استفاده از ضریب ۱.۵ برای IQR برای شناسایی داده‌های پرت، بخش زیادی از دیتا را حذف می‌کند. بنابرین این ضریب ۵ انتخاب می‌شود. در نهایت با شناسایی داده‌های پرت بر اساس فاصله‌ی ۳ برابر IQR از Q1 و Q3، از دیتا فریم حذف می‌شوند و در نهایت دیتا فریم نهایی حاصل می‌شود.

تعداد داده‌ها در این مرحله از ۱۳۴۵۹ به ۱۱۳۸۳۴ می‌رسد و این بدان معناست که ۱۵ درصد داده‌ها نسبت به حالت قبل حذف شده‌اند. گزارش زیر حاصل اجرایتابع info برای دیتا فریم نهایی می‌باشد.

```

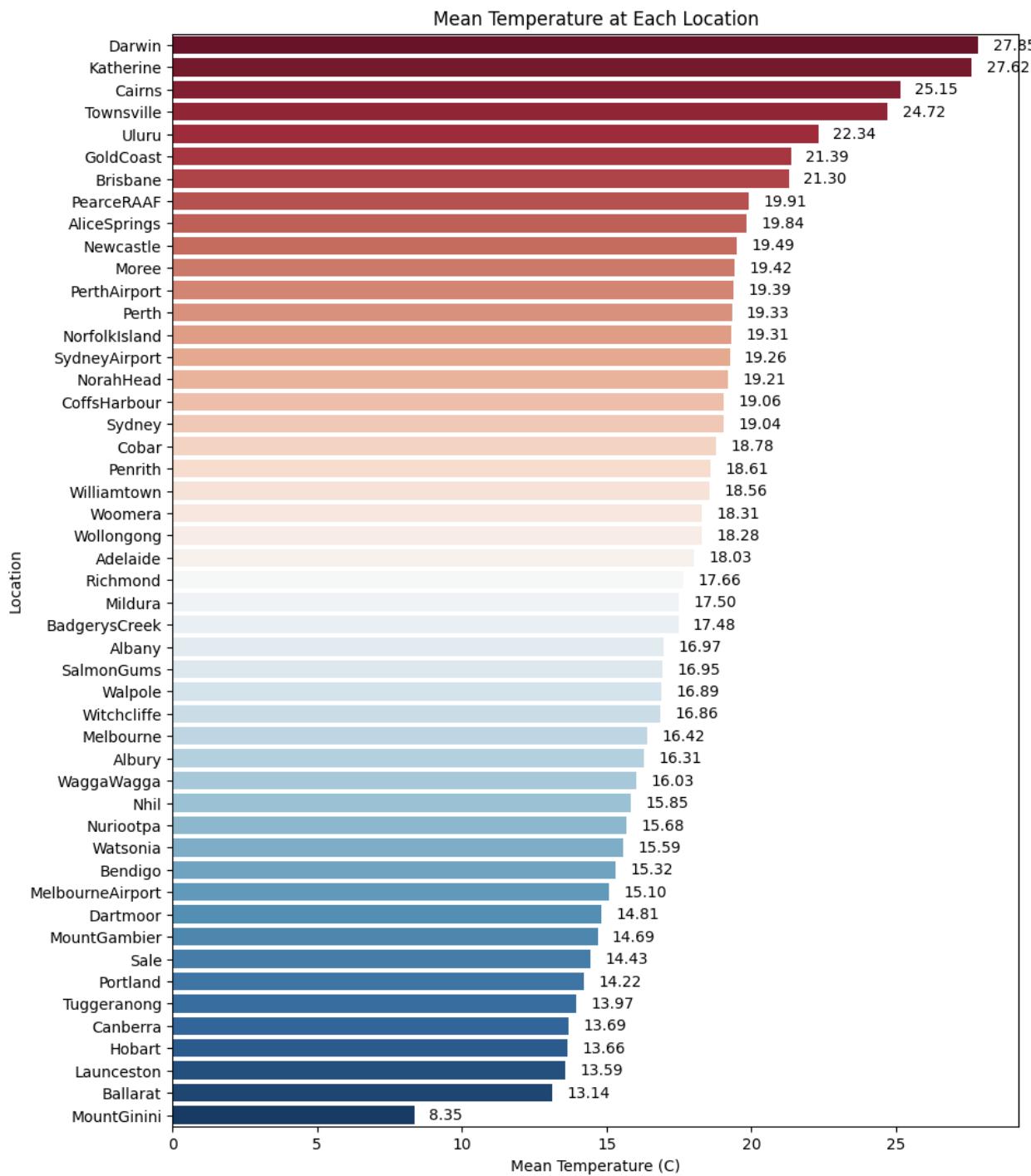
<class 'pandas.core.frame.DataFrame'>
Int64Index: 113834 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   Date              113834 non-null   object  
 1   Location          113834 non-null   category
 2   MinTemp           113834 non-null   float64 
 3   MaxTemp           113834 non-null   float64 
 4   Rainfall          113834 non-null   float64 
 5   Evaporation       113834 non-null   float64 
 6   Sunshine          113834 non-null   float64 
 7   WindGustDir       113834 non-null   category
 8   WindGustSpeed    113834 non-null   float64 
 9   WindDir9am        113834 non-null   category
 10  WindDir3pm        113834 non-null   category
 11  WindSpeed9am     113834 non-null   float64 
 12  WindSpeed3pm     113834 non-null   float64 
 13  Humidity9am      113834 non-null   float64 
 14  Humidity3pm      113834 non-null   float64 
 15  Pressure9am      113834 non-null   float64 
 16  Pressure3pm      113834 non-null   float64 
 17  Cloud9am          113834 non-null   float64 
 18  Cloud3pm          113834 non-null   float64 
 19  Temp9am           113834 non-null   float64 
 20  Temp3pm           113834 non-null   float64 
 21  RainToday         113834 non-null   category
 22  RainTomorrow      113834 non-null   category
dtypes: category(6), float64(16), object(1)
memory usage: 16.3+ MB

```

نمایش دادگان

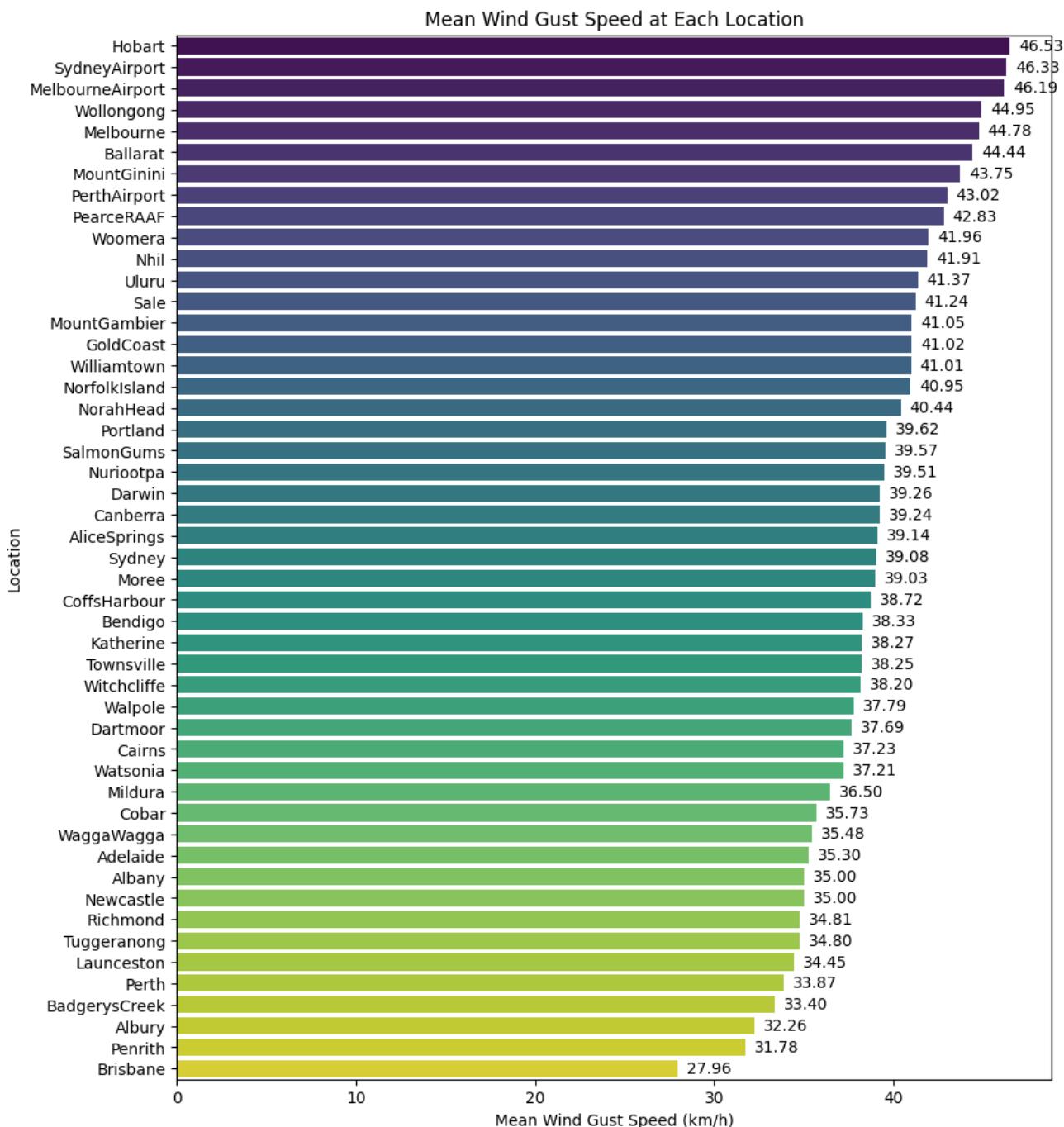
۲) در این بخش از سوال از کتابخانه seaborn استفاده می‌شود. با توجه به آن که میانگین دما در دیتاست وجود ندارد ابتدا برای هر شهر میانگین MinTemp و MaxTemp با کمک groupby() محاسبه می‌شود. (فرض شده است برای محاسبه میانگین دمای روزانه کاری به ستون‌های Temp9am و Temp3pm نداریم) در ادامه، از این دو مقدار میانگین گرفته شده و به ازای هر شهر در روزهای مختلف، دمای میانگین در نمودار میله‌ای به صورت افقی و نزولی رسم می‌شود. همچنین مقادیر مربوط به هر میله در جلوی آن آورده شده است.

نمودار ۱ میانگین دمای هر شهر را به صورت میله‌ای نشان می‌دهد.



نمودار ۱ - نمودار میله‌ای میانگین دما برای هر مکان

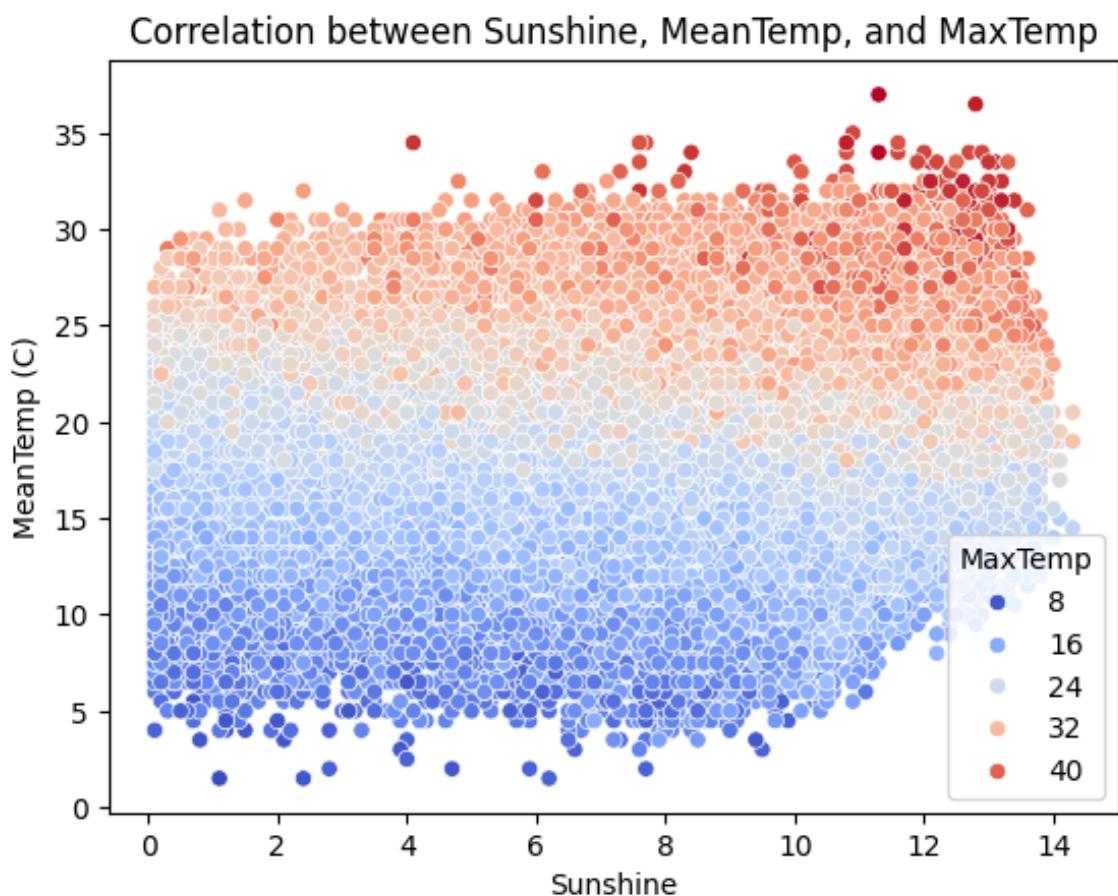
۳) برای این قسمت از نمودار میله‌ای استفاده شده است. اگر هدف از طرح این سوال آن باشد که مناطقی که باد در آنها سرعت بیشتری دارد شناسایی شود، بهترین رویکرد آن است که همانند قسمت قبل میانگین سرعت باد ثبت شده در هر ایستگاه در روزهای مختلف محاسبه شود و نتایج در یک نمودار میله‌ای با سایر ایستگاه‌ها مقایسه شود. به این منظور همانند بخش قبل عمل می‌شود. نمودار ۲ نتیجه را به نمایش در می‌آورد.



نمودار 2 - نمودار میله‌ای برای میانگین بیشترین سرعت باد هر منطقه

۴) برای این قسمت از نمودار scatter و hue استفاده می‌شود. برای آن که بتوان مقایسه‌ی دقیق‌تری داشت و سه پارامتر در نمودار مقایسه شود از hue scatter استفاده شده است. در این بخش تابش خورشید محور x میانگین دمای روزانه محور y و بیشترین دمای روزانه با گرادیان رنگ نمایش داده شده است. به این ترتیب که رنگ قرمز نمایانگر دمای بیشتر و

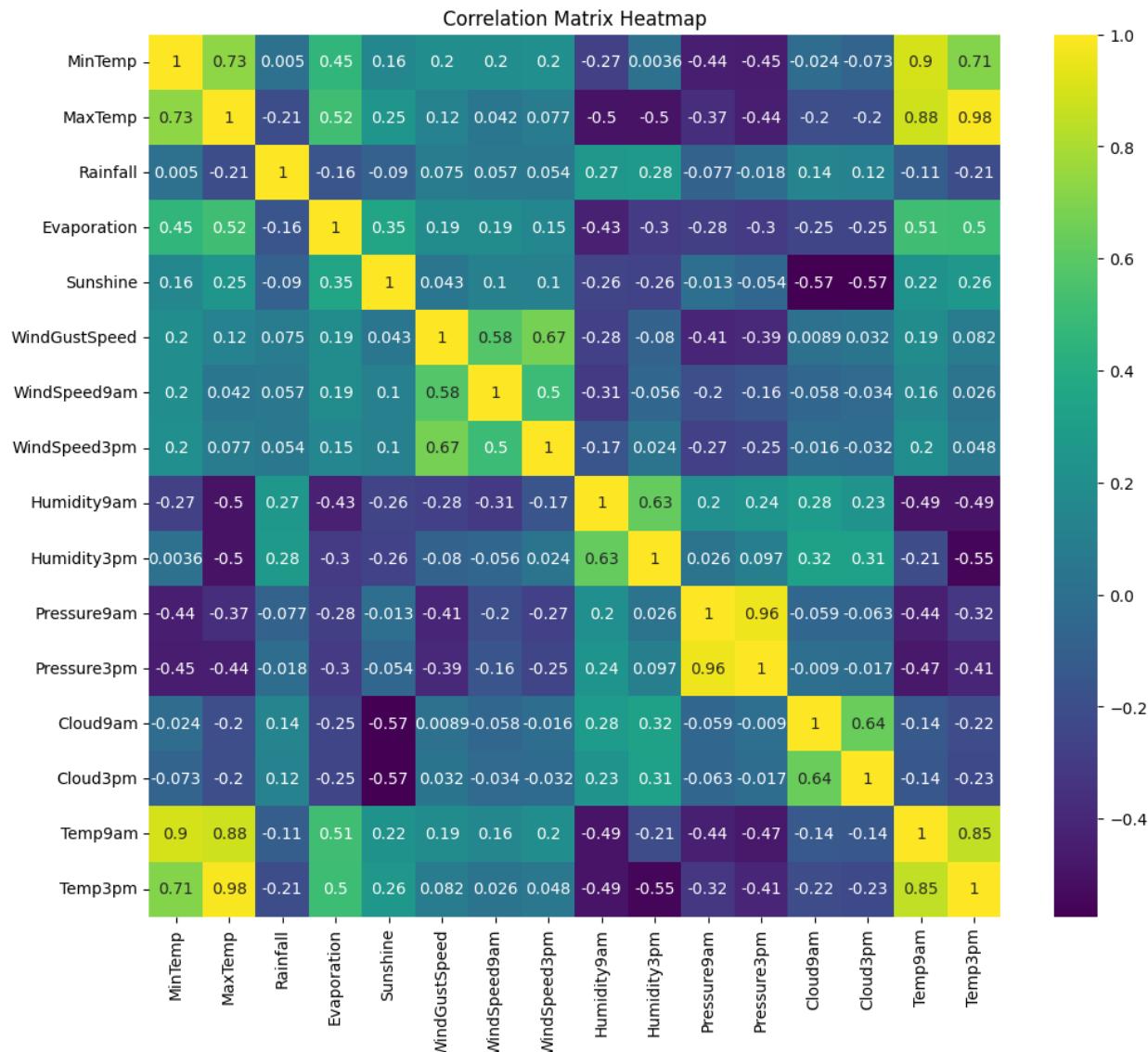
رنگ آبی نمایانگر دمای کمتر می‌باشد. نقطه‌های رسم شده همه‌ی دیتا در همه‌ی روزها و مکان‌ها است. لازم به ذکر است اطلاعات مربوط به تابش خورشید که صفر بودند از دیتاست حذف شدند. نمودار ۳ این مقایسه را نشان می‌دهد.



نمودار ۳ - نمودار scatter همبستگی تابش خورشید، میانگین و بیشترین دمای روزانه

همانطور که از نمودار ۳ مشخص است، با توجه به شیب مثبت روند داده‌ها (همبستگی مستقیم بین میانگین دما و تابش خورشید) و قرمزتر شدن نقطه‌ها در تابش‌های بالا (همبستگی مستقیم بین بیشترین دما و تابش خورشید) بین تابش خورشید و دمای روزانه همبستگی مستقیم وجود دارد.

۵) برای این منظور ابتدا ماتریس همبستگی با تابع `corr()` تشكیل می‌شود. سپس با کمک `seaborn` و دستور `() heatmap` نمودار خواسته شده رسم می‌شود. در این نمودار میزان همبستگی بین هر ویژگی عددی نمایش داده می‌شود که برگرفته شده از ماتریس همبستگی می‌باشد. نمودار ۴ نشان دهنده‌ی این نمودار است.



نمودار ۴ - ماتریس همبستگی به شکل heatmap

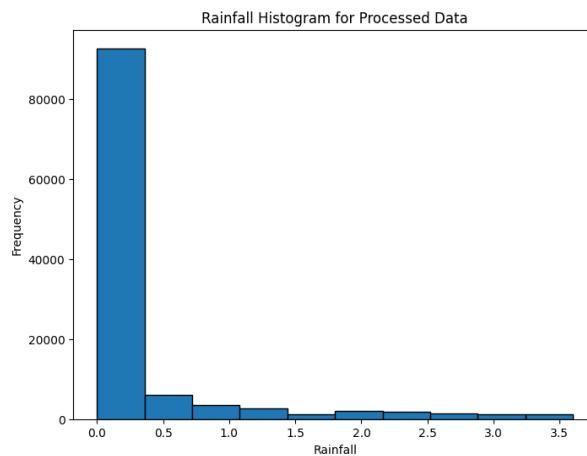
از نمودار ۴ می‌توان متوجه شد که Pressure9am، Temp9am با MinTemp، MaxTemp، Temp3pm با Temp9am، Temp3pm با Temp9am و Temp3pm با MaxTemp، Pressure3pm دارند. هر چه این عدد به ۱ نزدیکتر باشد همبستگی بیشتر است.

۶) در این قسمت برای داده‌های مختلف این نمودارها رسم شده‌اند که در ادامه توضیح داده شده‌اند.

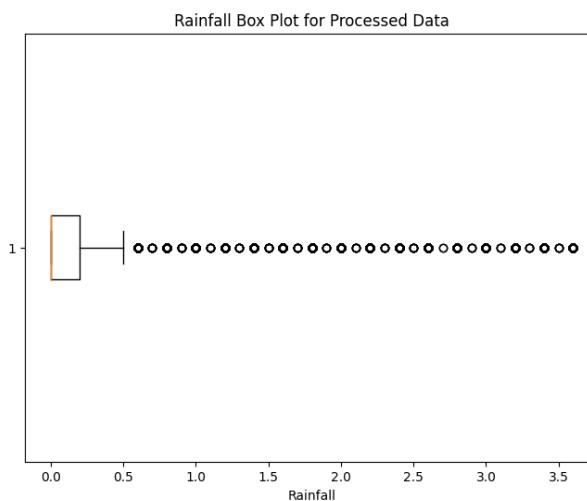
- برای داده‌ی پردازش شده
- برای داده‌ی پردازش شده که ۵ درصد اول و آخر آن حذف شده
- برای داده‌ی پردازش شده که روزهای بدون باران از آن حذف شده
- برای داده‌ی پردازش نشده‌ی بدون حذف داده‌های پرت

برای داده‌ی پردازش نشده که ۵ درصد اول و آخر آن حذف شده •

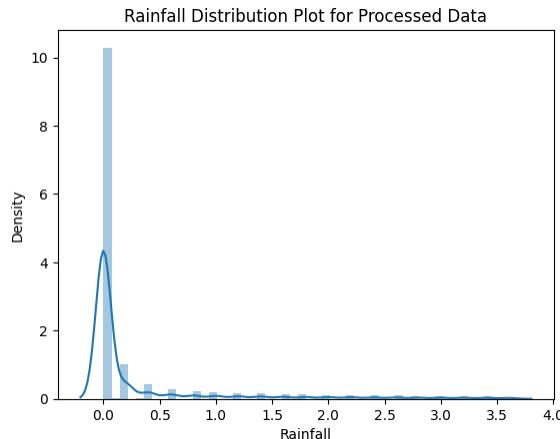
برای قسمت اول نمودارهای ۵ تا ۷ نتایج را نشان می‌دهند.



نمودار 5 - نمودار Histogram تعداد بارندگی بر اساس مقدار



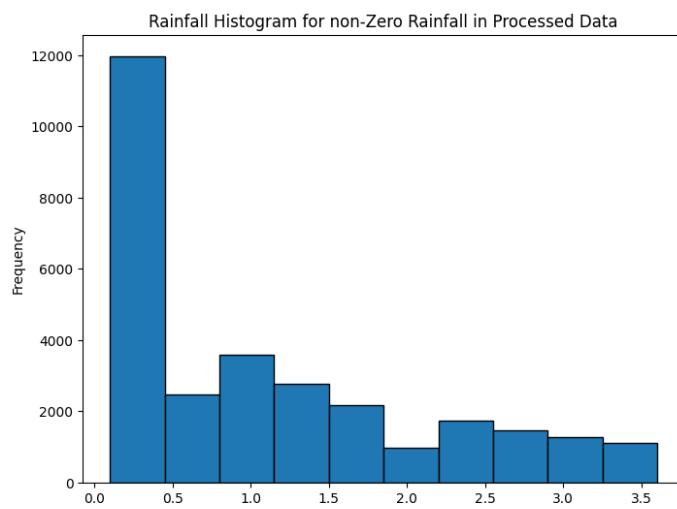
نمودار 6 - نمودار جعبه‌ای تعداد بارندگی بر اساس مقدار



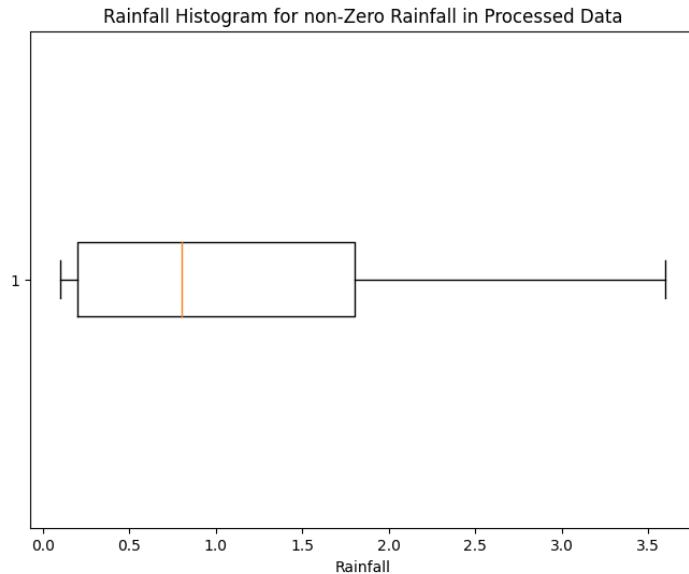
نمودار 7 - نمودار توزیع بارندگی بر اساس مقدار

برای قسمت دوم نمودارها رسم شدند اما نمایش خوبی ندارند و مفاهیمی از آن استنتاج نمی‌شود.

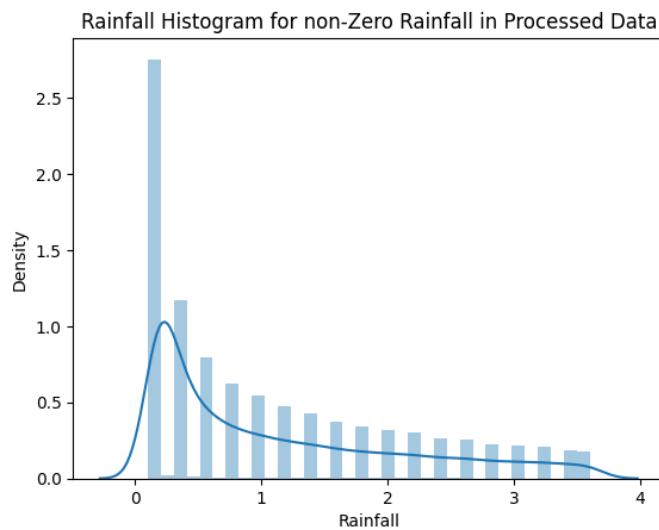
برای قسمت سوم همانطور که از نمودارهای ۵ تا ۷ مشخص است، بیشتر روزها بارندگی وجود ندارد یا مقدار آن بسیار کم است. به همین منظور اگر هدف بررسی میزان بارندگی در روزهای بارانی باشد فقط برای روزهای بارانی نمودارهای ۸ تا ۱۰ رسم می‌شوند. همچنین نمایش بهتری از تعداد بارندگی‌ها بر اساس مقدار ارائه می‌دهد.



نمودار 8 - نمودار Histogram تعداد بارندگی بر اساس مقدار



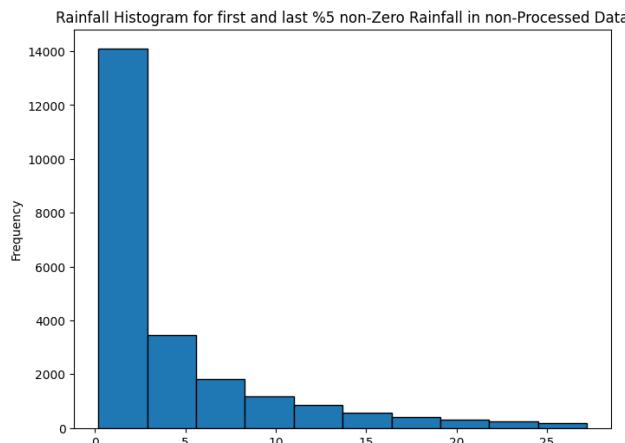
نمودار 9 - نمودار جعبه‌ای تعداد بارندگی بر اساس مقدار



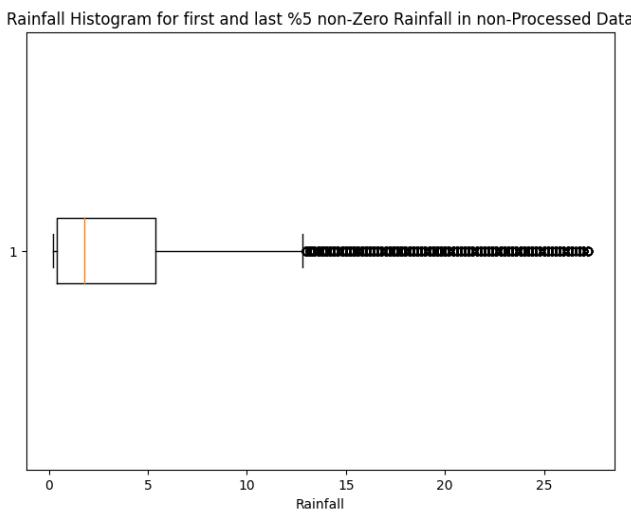
نمودار 10 - نمودار توزیعی بارندگی بر اساس مقدار

با توجه به آن که برای بررسی میزان بارندگی داده‌های مربوط به روزهای بدون باران حذف شد، برای بررسی بهتر این رویکرد، وجود تعداد زیادی روز بدون بارون باعث شده است در حذف داده‌های پرت مقداری از داده‌های مفید میزان بارندگی از بین برود.

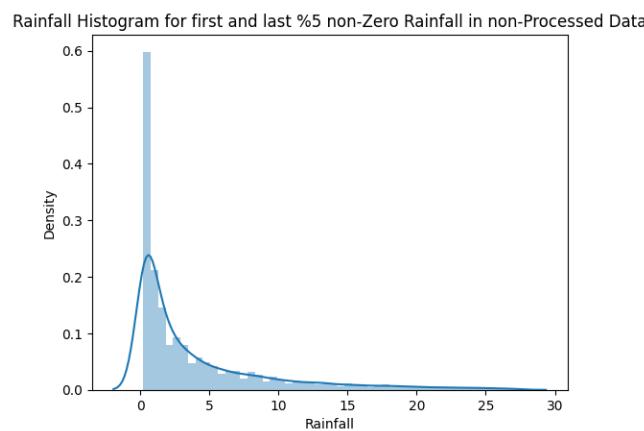
به این منظور دیتاست اولیه‌ی قبل از پیش پردازش گرفته می‌شود. ابتدا همه‌ی این نمودارها برای آن رسم می‌شود. سپس ۵ درصد اول و آخر آن حذف می‌شود و نتایج جدید در نمودارهای ۱۱ تا ۱۳ به نمایش در می‌آیند.



نمودار 11 - نمودار Histogram تعداد بارندگی بر اساس مقدار



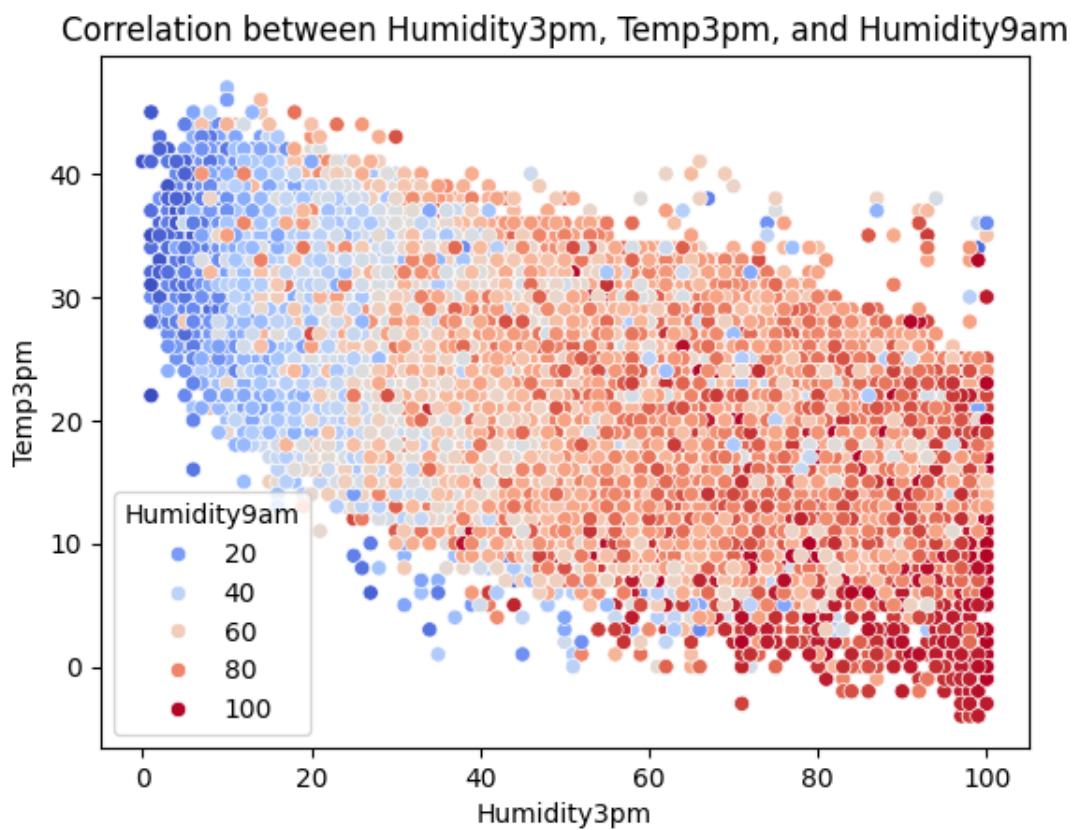
نمودار 12 - نمودار جعبه‌ای تعداد بارندگی بر اساس مقدار



نمودار 13 - نمودار توزیعی بارندگی بر اساس مقدار

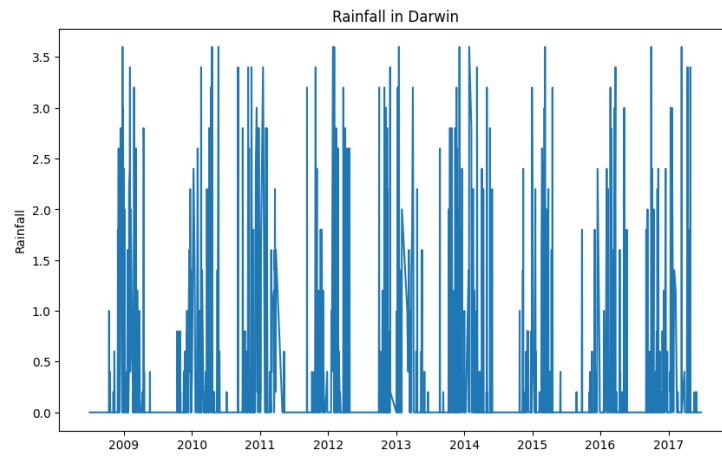
همانطور که از نمودارهای توزیعی و فراوانی بر می‌آید، توزیع تعداد بارندگی بر اساس میزان بارندگی به خوبی نمایش داده شده است. همچنین ين نمودارها نشان می‌دهند که positively skewed هستند. همچنین میانه‌ی این اطلاعات از نمودار جعبه‌ای قابل نمایش است.

۷) از این روش در بررسی تاثیر تابش خورشید بر دما در بخش ۴ استفاده شده است. برای بررسی همبستگی بین رطوبت ۳ بعد از ظهر با رطوبت ساعت ۹ صبح و دمای ۳ بعد از ظهر نمودار scatter hue رسم می‌شود. نمودار ۱۴ این همبستگی را نشان می‌دهد.

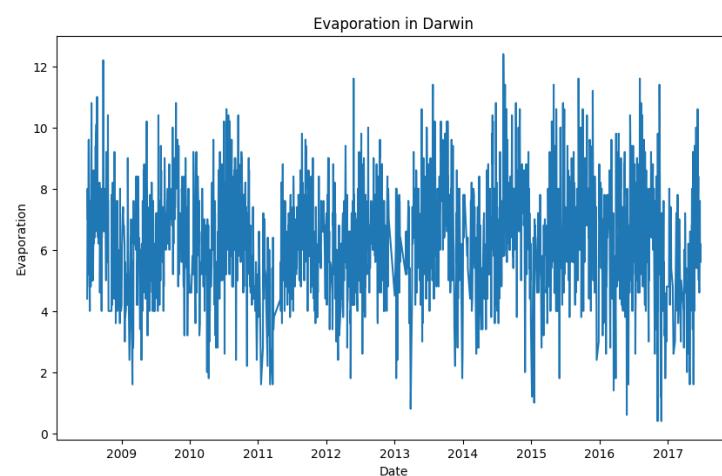


نمودار ۱۴ - نمودار scatter به منظور نمایش همبستگی

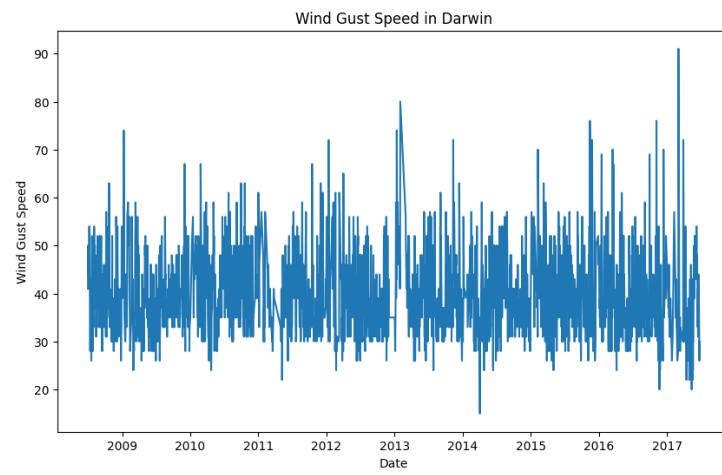
۸) موارد خواسته شده به تفکیک روز برای Darwin رسم شده‌اند. نمودارهای ۱۵ تا ۱۷ این اطلاعات را به خوبی نشان می‌دهند.



نمودار 15 - میزان بارش به تفکیک روزها

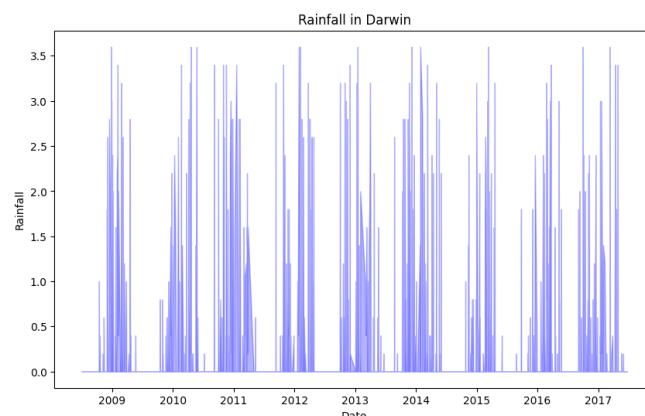


نمودار 16 - میزان تبخیر به تفکیک روزها

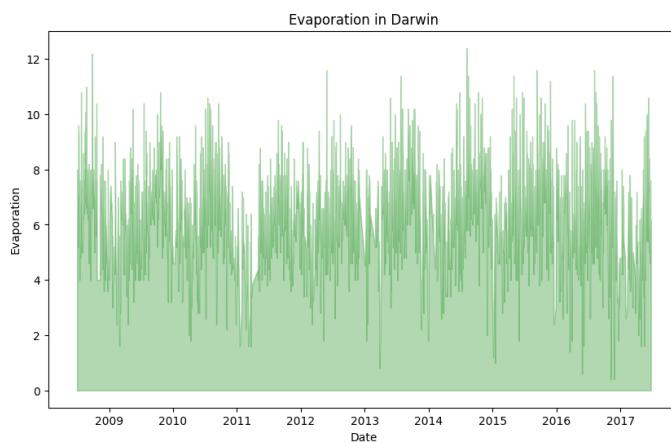


نمودار 17 - میزان بیشترین سرعت باد به تفکیک روزها

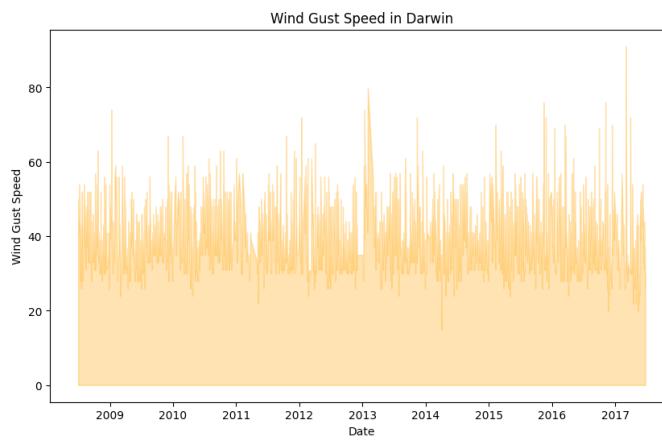
همچنین برای نمایش بهتر از نمودارهای سطحی استفاده شده است. نمودارهای ۱۸ تا ۲۰، نتیجه را نشان می‌دهد.



نمودار ۱۸ - میزان بارش به تفکیک روزها

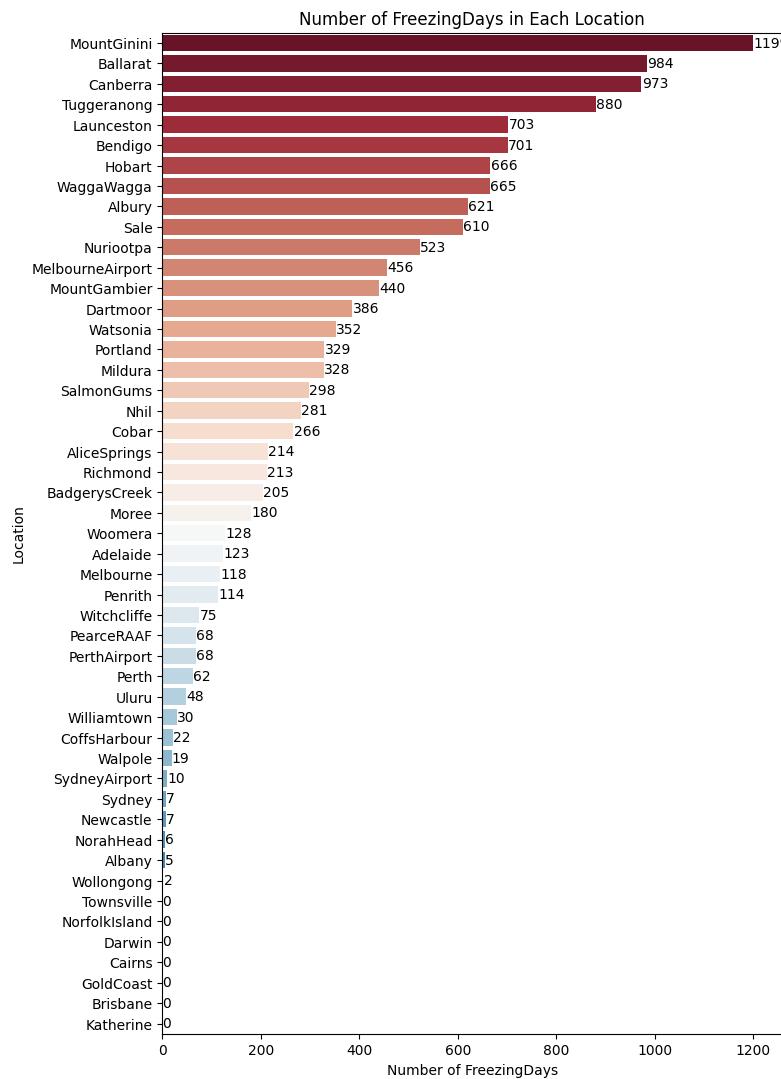


نمودار ۱۹ - میزان تبخیر به تفکیک روزها

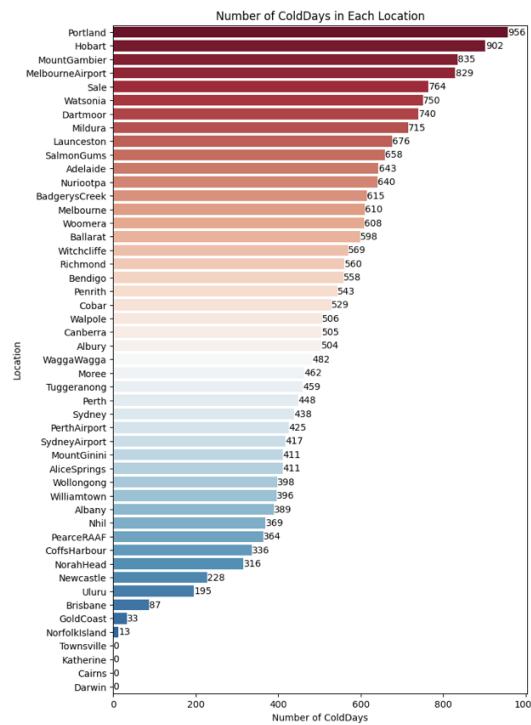


نمودار ۲۰ - میزان بیشترین سرعت باد به تفکیک روزها

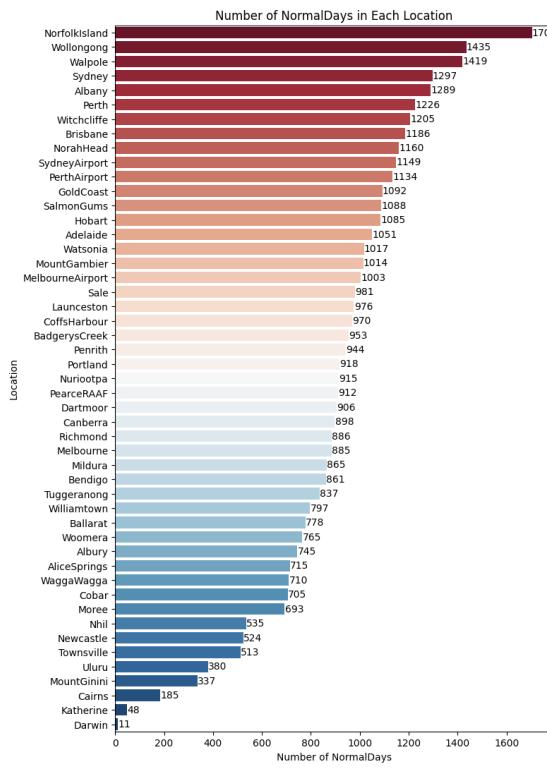
۹) در این قسمت ابتدا یک دیتا فریم جدید بر اساس ویژگی‌های عنوان شده در صورت سوال ایجاد می‌شود. سپس ۵ دسته‌بندی عنوان شده تشکیل می‌شوند. بر اساس این دسته‌بندی ایجاد شده اطلاعات مربوط به دماها بین این ۵ دسته تقسیم شده و تعداد آن‌ها گزارش می‌شود.
نمودارهای ۲۱ تا ۲۵ نشان‌دهنده تعداد روزهای هر ۵ دسته‌بندی در هر مکان می‌باشد.



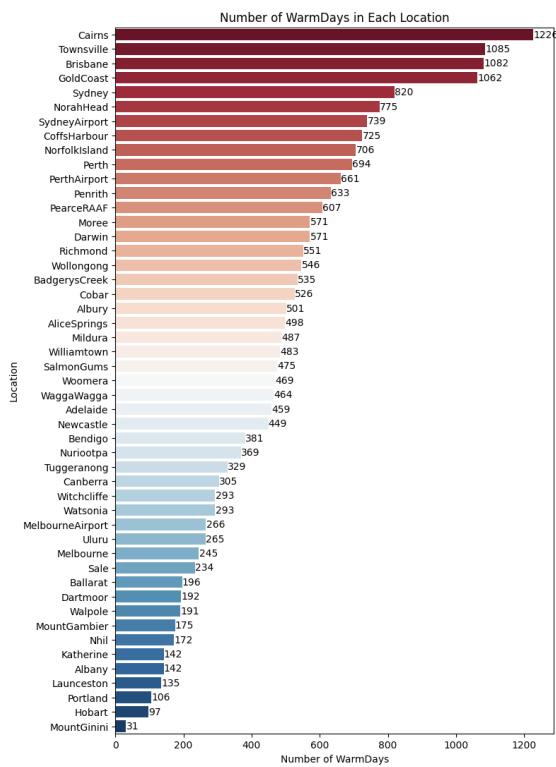
نمودار ۲۱ - نمودار میله‌ای تعداد روزهای خیلی سرد در هر استگاه



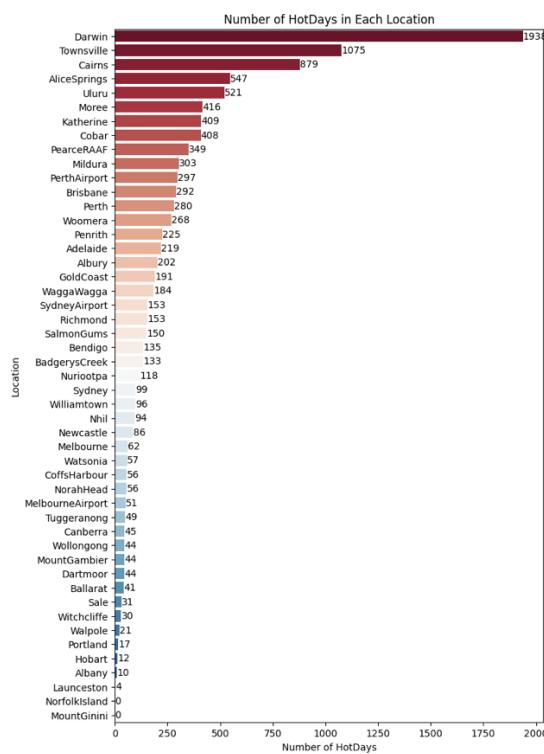
نمودار 22 - نمودار میله‌ای تعداد روزهای سرد در هر ایستگاه



نمودار 23 - نمودار میله‌ای تعداد روزهای عادی در هر ایستگاه



نمودار 24 - نمودار میله‌ای تعداد روزهای گرم در هر ایستگاه



نمودار 25 - نمودار میله‌ای تعداد روزهای خیلی گرم در هر ایستگاه

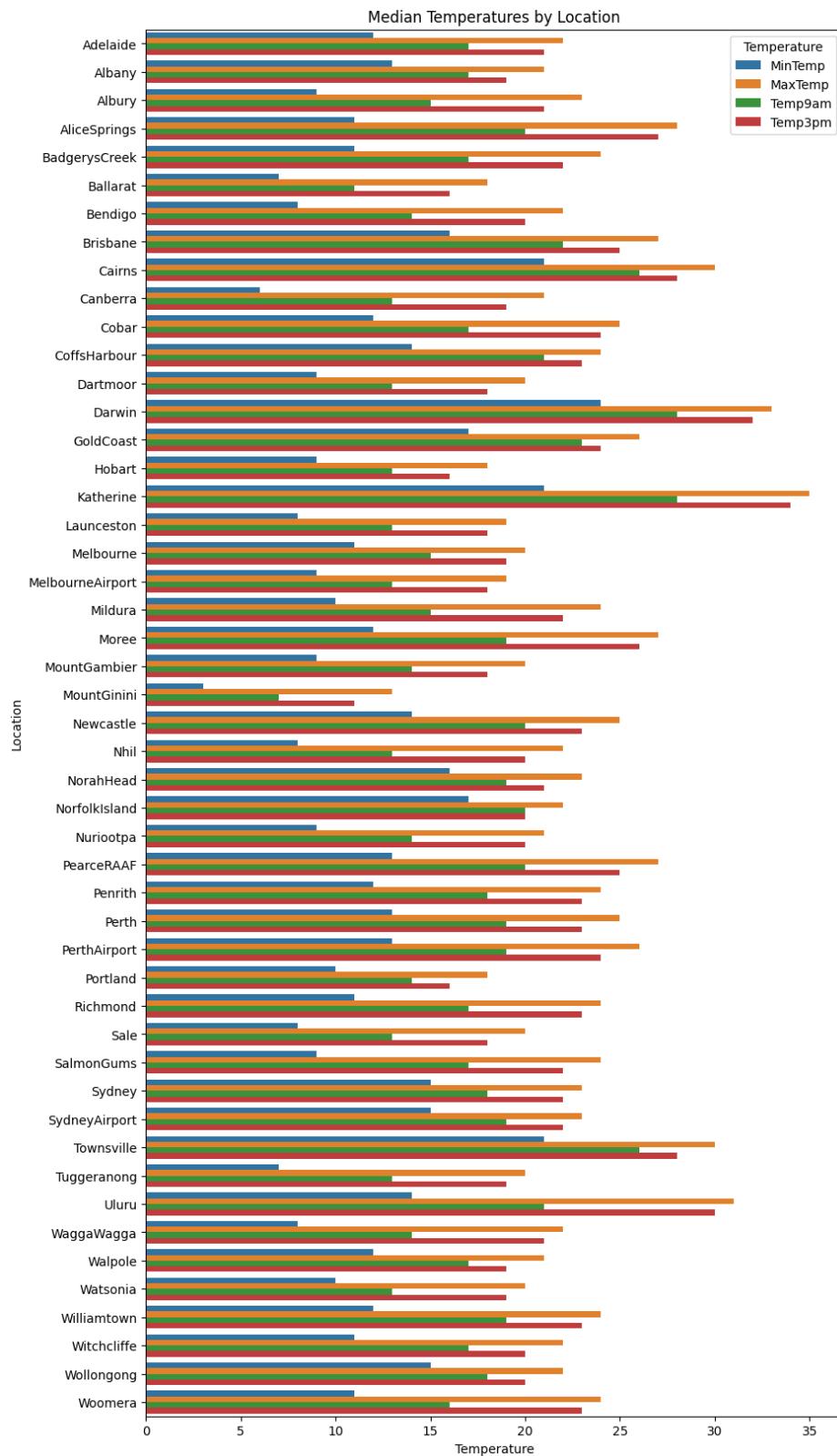
با توجه به نمودارهای رسم شده، نتایج زیر حاصل می‌شود.

- بیشترین تعداد روز خیلی سرد: MountGinini
- بیشترین تعداد روز سرد: Portland
- بیشترین تعداد روز عادی: NorfolkIsland
- بیشترین تعداد روز گرم: Cairns
- بیشترین تعداد روز خیلی گرم: Darwin

۱۰) با استفاده از `melt()` شکل دیتا فریم را می‌توان به منظور استفاده‌های متفاوت تغییر داد و زمانی کاربرد دارد که یکی از ستون‌ها ستون شناسایی هستند و بقیه ستون‌ها مقدار عددی برای آن می‌باشند. داکیومنت آن در [این لینک](#) وجود دارد.

تابع مخالف آن (`pivot()`) می‌باشد که دقیقاً برعکس `melt()` می‌باشد. در مواردی که بخواهیم تعداد سطرهای یک دیتا فریم را تغییر دهیم و کاهش دهیم از آن استفاده می‌شود و اطلاعات دیتا فریم حفظ می‌شود.

بتدا متغیرهای عنوان شده از دیتا فریم اصلی جدا شده و در یک دیتا فریم قرار می‌گیرد. سپس اطلاعات جدا شده برای هر مکان خاص انتخاب می‌شوند و همزمان میانه‌های آن‌ها برای هر ویژگی محاسبه می‌شود. در نهایت از `melt()` استفاده می‌شود که دیتا فریم را طولانی‌تر و باریک‌تر می‌کند. تمامی ستون‌ها و ویژگی‌ها در یک ستون قرار می‌گیرند. در نهایت نمودار آن‌ها به صورت میله‌ای رسم می‌شود که در نمودار ۲۶ آورده شده است.



نمودار 26 - نمودار میله‌ای برای نمایش ویژگی‌های دما برای هر مکان