1. **Phương pháp thu thập**

Dữ liệu thời thiết chúng em thu thập từ website [wunderground](https://www.wunderground.com/) là một website chuyên cung cấp dữ liệu về thời tiết trên toán thể giới.

Việc thu thập được sử dụng ngôn ngữ lập trình python với thư viện hỗ trợ là selenium với dữ liệu theo 30 phút, thời gian tứ 01-01-2019 -> 01-01-2022 tại 3 địa điểm Sài Gòn, Hà Nội, Đà Nẵng

Trước tiên, chúng em thu thập từng vùng sau đó là merge các dữ liệu lại thành một dataset lơn với hơn 200 000 quan sát và 12 thuộc tính

1. **Phương pháp tiền xử lý dữ liệu**

Chúng em thực hiện xử lý theo quy trình :

|  |  |
| --- | --- |
| Quy trình xử lý | Nội dung |
| Merge dữ liệu | Bộ dữ liệu sau khi craw là các tệp .csv theo ngày và sau đó chúng em đã merge nó lại thành 1 bộ dataset hoàng chỉnh |
| Thăm dò dữ liệu | Bộ dữ liệu có 204044 quan sát, có 12 thuộc tính và 1 thuộc tính ‘Unnamed: 0’ bị thừa  Ngoài ra trong các quan sát còn có các kí tự: %, và trong tên các thuộc tính còn chứa các khoảng trắng |
| Đổi tên lại các thuộc tính và xóa thuộc tính bị thừa | Xóa thuộc tính ‘Unnamed: 0’  Đổi tên các thuộc tính thành ['Time', 'Temperature', 'Dew\_Point', 'Humidity', 'Wind', 'Wind\_Speed', 'Wind\_Gust', 'Pressure', 'Precip', 'Condition', 'Date', 'Region']  Xóa thuộc tính 'Wind\_Gust' và 'Precip' vì 2 thuộc tính này có giá trị 0.0 trên 200000 giá trị chiếm hơn 99% số lượng mẫu |
| Kiểm tra dữ liệu null và loại bỏ các kí tự đặc biệt trong các quan sát | Chỉ có 13 giá trị null trong thuộc tính Condition => thay thế bằng thuật toán KNN  Với các kí tự đặc biệt ta dùng hàm **replace** để thay thế |
| Ép kiểu dữ liệu | Ép lại kiểu float cho các thuộc tính numeric  Kiểu datetime cho thuộc tính date bằng hàm pd.to\_datetime() |
| Chuyển đổi dữ liệu | Vì các thuộc tính Temperature và Dew\_Point đang đo ở vì thế chúng ta chuyển nó sang theo công thức: |
| Thêm thuộc tính biên độ nhiệt | chúng e nhận thấy với dữ liệu thời tiết thì biên độ nhiệt khá là quan trọng vì thế nên chúng em quyết định tạo thêm 1 thuộc tính biên độ nhiệt (Amplitude) với công thức: |

1. Phân tích khám phá dữ liệu
2. **Thống kê mô tả:**

Table

Description automatically generated

* Bản mô tả dữ liệu cho ta thấy nhìn chung nhiệt độ trung bình và và biên độ nhiệt không quá cao, quy nhiên có một vài lưu ý ta có thể thấy đó là giá trị nhiệt độ (Temperature) và nhiệt độ điểm sương (Dew\_Point) lại là giá trị âm. Điều này là vô lý với thời tiết ở Hà Nội, Đà Nẵng và Sài Gòn

1. **Phân tích các biến numeric:**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Nhìn vào biểu đồ về nhiệt độ (Temperature), nhiệt độ điểm sương (Dew Point), độ ẩm (Humidity) và biên độ nhiệt của 3 địa điểm Hà Nội, Đà Nẵng và Sài Gòn chúng em nhận thấy rằng:

* Giá trị của 3 thuộc tính nhiệt độ, nhiệt độ điểm sương và độ ẩm nhỏ hơn hoặc bằng 0. Đây là những giá trị không thu thập được và sau khi chuyển sang thì giá trị nó chuyển thành âm và cũng tại những điểm đó đã làm biên độ nhiệt tăng quá cao có điểm lên gần tới 60
* Về nhiệt độ thì nhiệt độ của Sài Gòn trong năm thường không có biến động nhiều chỉ tầm 20 - 38, trong khi Hà nội thì con số này biến động rất lớn, có thời điểm lạnh nhất dưới 10 và nóng nhất có ki lên tới 40
* Ngoài ra, ta còn nhận thấy được các thuộc tính này đều có chu kì theo năm và được thể hiện rõ ràng nhất là ở Hà Nội sau đó là Đà Nẵng và cuối cùng là Sài Gòn. Lý giải cho điều triên khi mà Hà Nội có 4 mùa và chịu ảnh hưởng của không khí lạnh vào mỗi dịp mùa đông, trong khi Đà Nẵng ít chịu ảnh hưởng bới không khí lạnh và ở sài gòn thì chỉ có 2 mùa là mùa nắng và mùa mưa.

**Sự tương qan giữa các biến**

Application, table

Description automatically generated

Để nhìn rõ hơn về mối phụ thuộc giữa các biến **numeric** ra có biểu đồ heatmap.

Khá là thú vị khi mà các biến ở mỗi địa điểm lại có một sự tương quan rất là khác nhau.

* Ở Sài gòn, **độ ẩm (Humidity)** và **nhiệt độ điểm sương (Dew Point)** là 2 thuộc tính có mỗi tương quan cao nhất nhưng chỉ ở mức 0.69. Ngược lại **độ ẩm (Humidity)** và **nhiệt độ (temperature)** tuy có tương quan nghịch nhưng sự tương quan khá là cao - 0.77
* ở Đà Nẵng và Hà Nội lại có mối tương thuận quan giữ **nhiệt độ điểm sương(Dew Point)** và **nhiệt độ (Temperature)** với 0.63 và 0.76 theo từng vùng. Ngoài ra, mối tương quan nghich giữa hai biến **độ ẩm (Huminity)** và **nhiệt độ (Temperature)** cũng rất có ý nghĩa khi mà nó có giá trị lên tới – 0.74.

Có thể nói tại mỗi vùng với các điều kiện thời tiết khác nhau thì những thuộc tính này tác động đến thuộc tính kia cũng sẽ khác nhau và để có cái nhìn rõ hơn qua 4 biểu đồ dưới đây

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Với quan sát nhiệt độ cao nhất trong ngày ở cả 3 địa điểm ta nhận thấy sự phân hóa rõ hơn thời tiết của 3 vùng khi mà thời tiết nóng nhất cảu sài dong là vào tháng 3-5 trong khi của 2 địa điểm còn lại là tháng 6 – 8. Ngoài ra, có một số outlier nhưng với bộ dữ liệu trên 200000 thì những outlier này khá là nhỏ để có thể ảnh hưởng nhiều tới mô hình

1. Đối với các biến Categotical

Chart, histogram

Description automatically generated

Đối với dữ liệu thời tiết hướng gió cũng khá là quan trọng khi mà nó cũng ảnh hưởng 1 phần nào đó tới thời tiết và nhiệt độ của vùng. ở đây biến **var** là cao nhất đây là biến cho ta biết thời điểm thay đổi hướng gió trong ngày và giá trị của nó ở Sài gòn cao hơn ở Hà Nội và đà nẵng rất nhiều chứng tỏ thời tiết trong ngày ở Sài gòn biến động nhiều hơn ở hai vùng còn lại. Bên cạnh đó, mật độ gió Đông (E, SE, ESE..) rất nhiều, loại gió này mang hơi ẩm từ biển vào mà cũng là một trong những nguyên nhân khiến lượng nước ta mưa nhiều.

Graphical user interface, chart, bar chart

Description automatically generated

Không có gì quá bất ngờ khi mà với tầm suất gió đông rất là nhiều thì thời tiết mây và mưa cũng rất là nhiều đặc biệt là Sài Gòn, theo sau là Đà Nẵng. Còn lại Hà Nội thì nằm sâu trong đất liền vì thế thời tiết ít mưa và trời đẹp (Fair) với tần suất khá là nhiều