Tutorial: Exploratory Data Analysis + Web Scraping

Daniel Christian Mandolang – 2106630006

1. Preprocessing Data
   Pada tahapan preprocessing dari kode yang diberikan dapat dilihat bahwa data telah dapat diproses. Namun, jika dilihat kita dapat melakukan beberapa hal untuk menyempurnakan tahap preprocessing data ini.

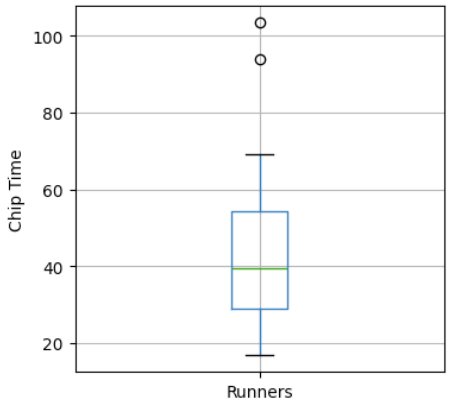
   a. Strip ‘Name’, ‘Gender Place’, and ‘Age Group Place’ column
      ‘Name’, ‘Gender Place’, and ‘Age Group Place’ column in the dataframe still contains leading and trailing whitespaces that make them hard to be read.
      ```
      df['Name'] = df['Name'].str.strip()
      df['Gender Place'] = df['Gender Place'].str.strip()
      df['Age Group Place'] = df['Age Group Place'].str.strip()
      ```

   b. Normalize ‘Gender Place’ and ‘Age Group Place’ Column
      ‘Gender Place’ and ‘Age Group Place’ is still in form of ‘4 of 78’ which is difficult to be processed by computer. It is better to be normalized, for example ‘4 of 78’ is normalized into 4/78 = 0.0512820513.
      ```
      def normalize_place(s: str) -> float:
          place = s.split(' of ')
          return int(place[0]) / int(place[1])

      df['Gender Place'] = df['Gender Place'].apply(normalize_place)
      df['Age Group Place'] = df['Age Group Place'].apply(normalize_place)
      ```

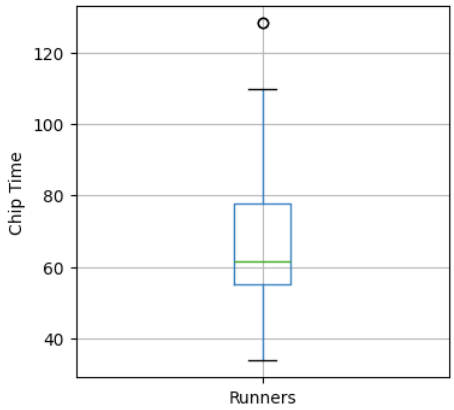
2. Analisis Perbandingan Outlier
   Dapat dilihat pada ketiga box plot di bawah, terdapat perbedaan mengenai outlier dari ketiga dataset. Berikut ini analisis dan perbandingan hasil ketiga dataset.

a. 2018MLK5K



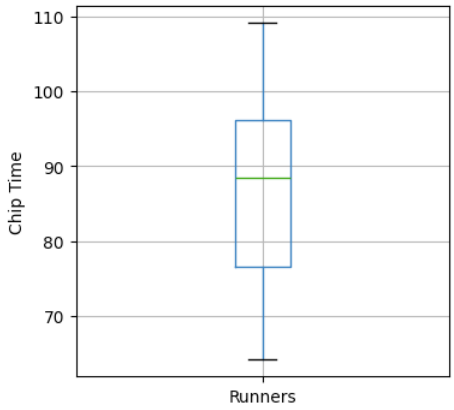
- Median berada di sekitar nilai 50, yang menunjukkan bahwa setengah dari pelari menyelesaikan lomba di bawah waktu ini.
- IQR, yang ditunjukkan oleh panjang kotak, adalah sekitar 30 hingga 70 menit, yang relatif lebar menunjukkan variasi yang cukup signifikan dalam waktu penyelesaian pelari.
- Ada beberapa outlier di atas batas atas whisker (sekitar nilai 100), menandakan bahwa ada beberapa pelari dengan waktu yang jauh lebih lambat dibandingkan dengan sisanya.

b. 2018MLK10K



- Median terlihat lebih tinggi dibandingkan dengan 5K, berada di sekitar angka 60 menit.
- IQR yang lebih sempit dari 5K, kira-kira antara 50 hingga 80 menit, menunjukkan bahwa pelari 10K memiliki variasi waktu yang lebih sedikit.
- Seperti 5K, ada outlier di atas batas atas whisker, namun tampaknya jumlahnya lebih sedikit.

c. 2018MLK15K



- Median sekali lagi naik, sekarang berada di sekitar 95 menit, yang masuk akal mengingat jarak yang ditempuh lebih jauh.
- IQR adalah yang terkecil di antara ketiganya, menunjukkan konsistensi waktu penyelesaian yang lebih tinggi di antara pelari 15K.
- Tidak ada outlier yang terlihat, yang bisa menunjukkan bahwa pelari 15K secara keseluruhan adalah grup yang lebih konsisten.

Outlier diidentifikasi sebagai titik-titik di luar 'whisker' atau garis horizontal terjauh dari kotak.

- 5K memiliki jumlah outlier terbanyak, yang dapat menunjukkan bahwa ada lebih banyak pelari kasual atau pemula yang waktu penyelesaiannya bervariasi secara signifikan.
- 10K memiliki beberapa outlier, tetapi lebih sedikit dari 5K.
- 15K tampaknya tidak memiliki outlier, yang mungkin menunjukkan bahwa pelari ini lebih seragam dalam kemampuan dan pelatihan mereka.