# Opschonen van de data

Hoge waarden weggooien:

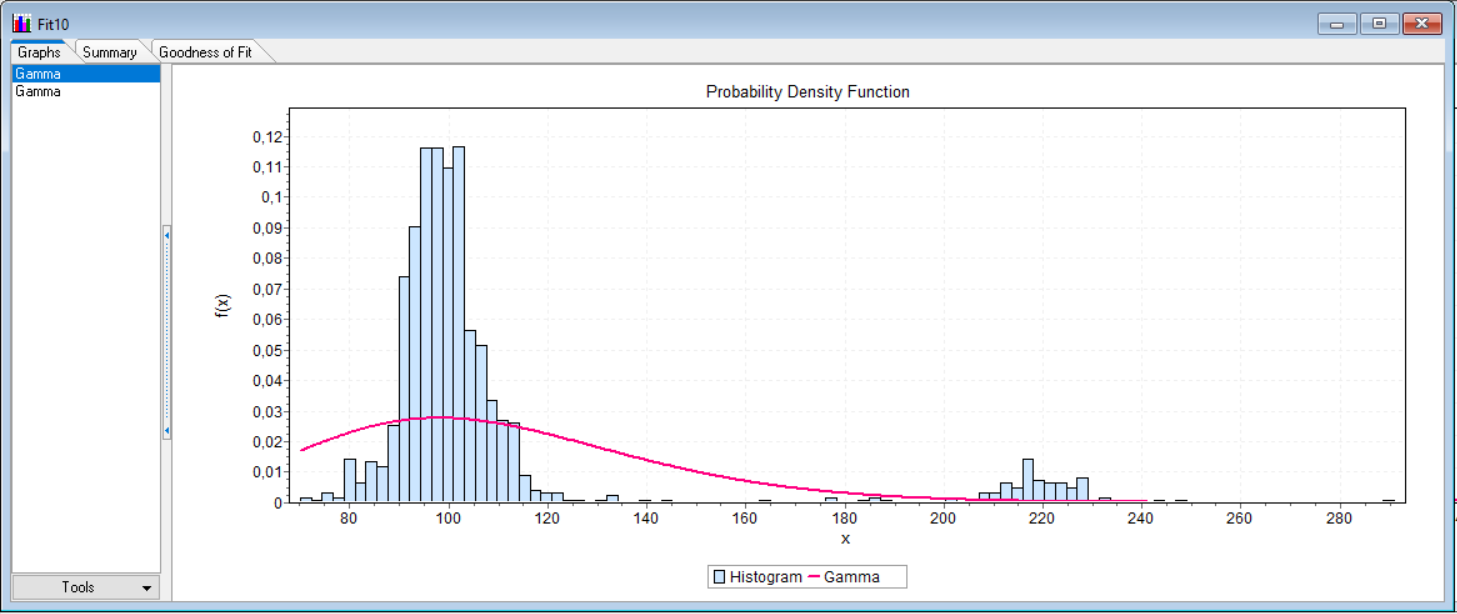
* Veroorzaakt door calamiteiten die buiten de scope van onze simulation study liggen

Lage waarden weggooien:

* Waarschijnlijk door fout in de metingen. Extreem lage waarden zijn praktisch gezien ook onmogelijk aangezien de tram een maximum snelheid heeft en een gegeven afstand moet afleggen.

Clusters weggooien:

* Waarschijnlijk veroorzaakt door een tijdelijke halte of omleiding.
* Buiten de scope van onze simulation study: we nemen aan dat alle obstructies op het traject wijken voor de tram (geen stoplichten) en dat de verschillen in reistijd slechts worden veroorzaakt door weersomstandigheden en variërende machinisten.
* Veroorzaakt extreem slechte fit van een functie op de data. Een voorbeeld hiervan zien we hieronder: hier is een poging gedaan een gamma distributie te vinden die de ruwe data het best beschrijft. De fit is verre van goed.



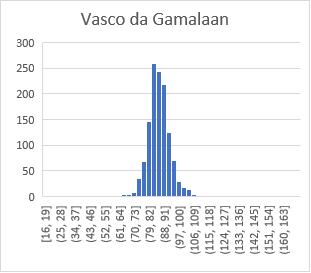
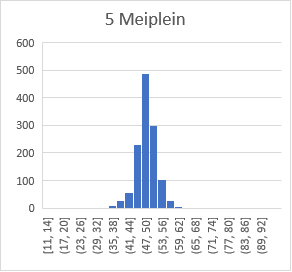
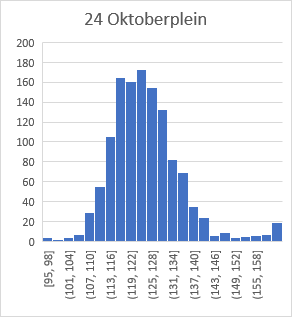
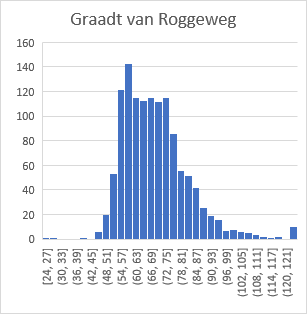
# Distributie kiezen

Aan de hand van alle scatterplots van de data per rit.

Hieronder, als voorbeeld, 4 gegeven.

Conclusie: we kiezen een **gamma distributie.**

* De vorm van de grafieken komt het meest overeen met de dichtheidsfunctie van de gamma distributie.
* We kiezen de voor een distributie met slechts 2 parameters om de simulatie niet te complex maken.



# Een geschikte fit zoeken per traject.

Gamma distributie wordt gedefinieerd door 2 parameters:

* Shape parameter k
* Scale parameter phi

Deze twee variabelen verhouden zich direct tot het gemiddelde en de standaard deviatie, immers:

* Gemiddelde = k \* phi
* Std.Deviatie = wortel(variance) = wortel(k \* phi ^2)

Analyse van de gemiddelden en de standaard deviaties van de data:

* Plot gemaakt van:
  + Alle gemiddeldes op de x as, alle standaard deviaties op de y as
* Vermoeden: er is een lineare relatie tussen gemiddelde en standaard deviatie
  + *Dit vermoeden wordt ondersteund door onze aanname over de variatie in reistijd: indien deze volledig wordt veroorzaakt door het weer en de rijstijl van de machinist, is het effect hiervan merkbaar gedurende het gehele traject. De schommeling zal dus groter zijn wanneer het traject langer is, en kleiner zijn wanneer het traject korter is.*

**Graadt v. Roggeweg (beginhalte) nemen we niet mee in de samenstelling van de lineare vergelijking**

* In de grafiek is duidelijk te zien: dit is een outlier
* Resulteert in een enorm lage waarde van R^2 (correlatie coefficient)
  + R is de waarde van de correlatiecoëfficiënt van Pearson, met als maximale waarde 1
* Argumentatie: dit is de eerste halte van het traject, met als vertrekpunt het Centraal Station. Een hoge variatie zou kunnen worden veroorzaakt door het lang moeten wachten op arriverende passagiers en binnenrijdende trams bij het station.

**Resultaat:** een nette vergelijking voor de relatie tussen het gemiddelde en de standaard deviatie (in het groene vlak: y = 0,0682x + 0,759

* Met bovendien een correlatie coefficient van 0,86 = goed.

**In de simulation study is de verwachtte gemiddelde reistijd gegeven**

* Hiermee kunnen we volgens bovenstaande vergelijking de verwachtte standaard deviatie berekenen
* Hiermee wordt een gamma distributie volledig vastgelegd van waaruit gesampled kan worden voor de reistijden per traject.