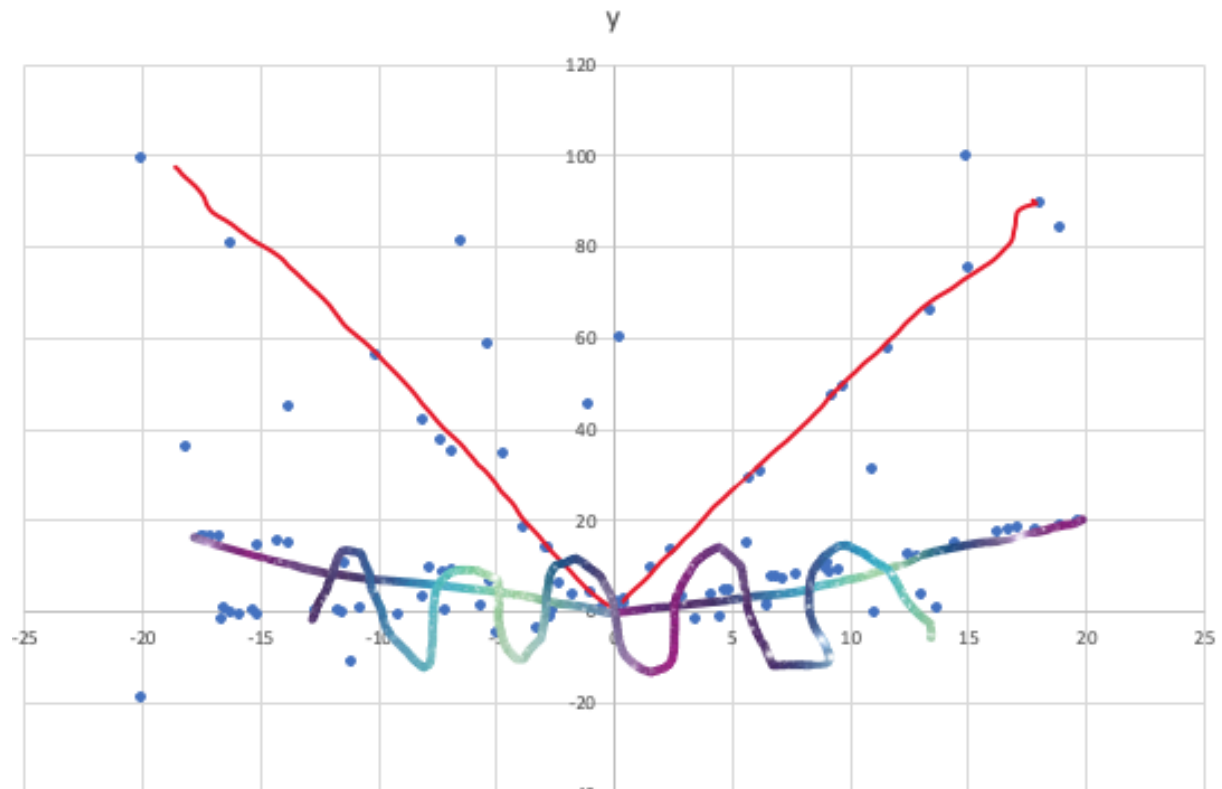
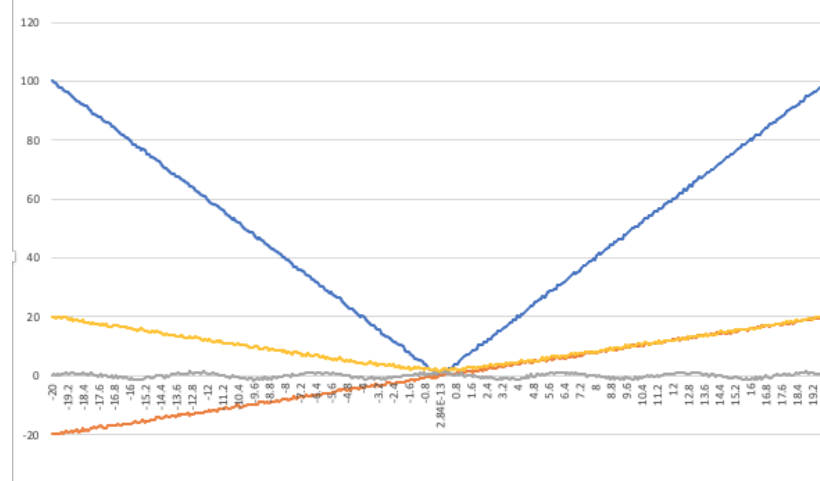


Bevor wir uns mit Python beschäftigen, versuche ich eine Lösungsskizze mit Excel zu erstellen. Das ist manchmal, in der Realität, immer sinnvoll, vor allem wenn es die Datenmenge zulässt. Fangen wir am Ziel an. Wir haben einen Haufen von Testdaten (test.csv), die irgendeine Realität darstellen. Wenn man genau hinschaut, kann man per Hand, wie in der folgenden Abbildung gezeigt, reinmalen, dh, verschiedene Aspekte unserer Realität werden durch verschiedene Funktionen beschrieben.

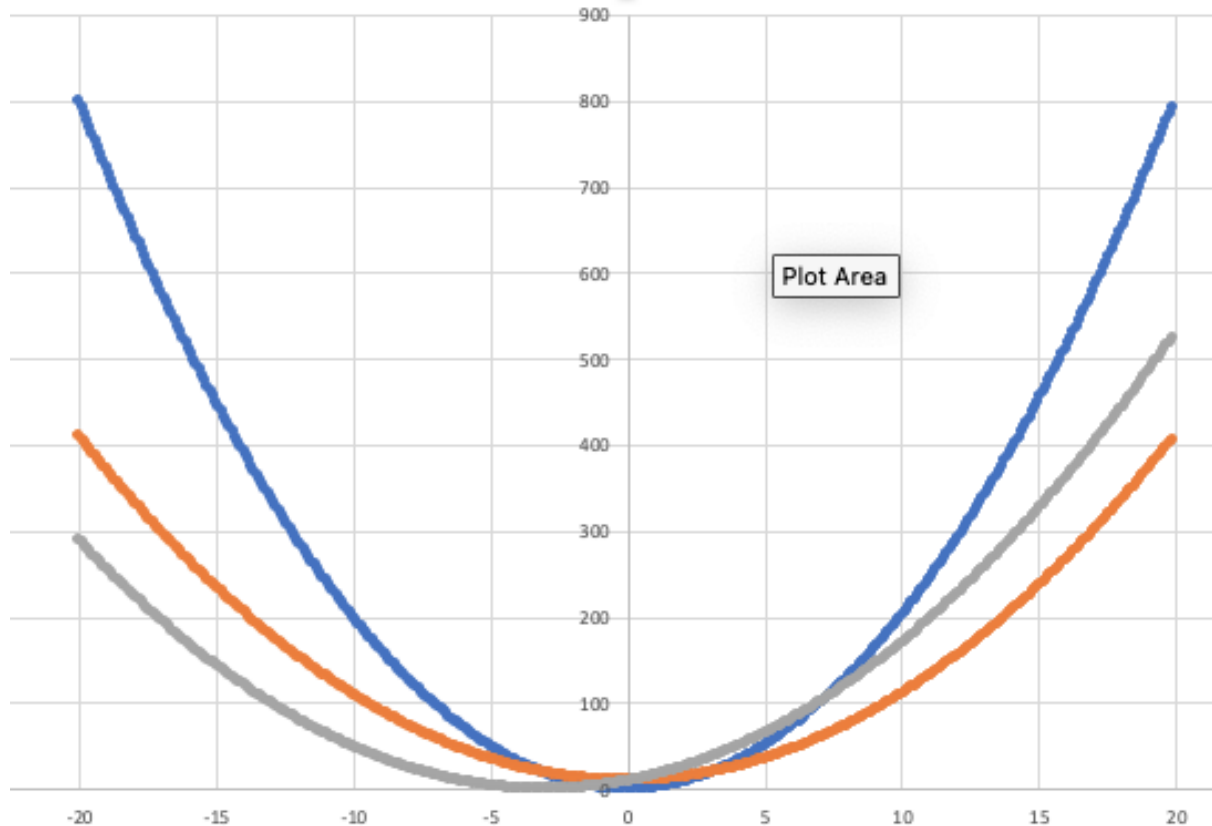


Der Training Dataset enthält diese vier verschiedene Funktionen. Stell Dir diese Funktionen vor als Erkenntnisse, die meine von mir konstruiert Realität abbilden. Wie man in der folgenden Abbildung sieht, sind die vier Darstellungen etwas verrauscht. Stell Dir vor, dass sind individuell gemessene Daten oder auch generierte Daten.



Wenn man algorithmisch an Problem herangeht, hätte man natürlich lieber eine geschlossene Funktion oder zumindest eine gute mathematische Funktion, die die verrauschten Daten relativ gut abbildet.

In dieser Aufgabe stelle ich 50 verschiedenen Versionen von vier Funktionen zur Verfügung, die eine gut Approximation von unseren 4 Funktionen, die wir glauben in dem Testdatensatz zu finden. Wir nennen diese „ideale Funktionen“. Hier drei Beispiele aus dem „ideale Funktionen“ Datensatz.



Jetzt wollen wir natürlich nicht per Hand malen und auch nicht Excel machen, und daher:

- Aufgabe 1: Finde die 4 idealen Funktionen die unserem Modell (die 4 verrauschten Trainings-Funktionen) am nächsten kommen. Least Square und das gibt uns einen Noise-free Datensatz für jede unsere 4 Funktionen. Dh, anstatt meiner noisy Trainingsfunktionen, benutze ich jetzt saubere Daten aus den ideale Funktionen Datensatz.
- Aufgabe 2: Kann ich die 4 idealen Funktionen mit einer Fehlertoleranz an die Realität anpassen. Wie wir in meinem Kunstwerk oben schon gesehen haben, an paar Stellen geht das ganz gut. Andere, wie die sinus Kurve, nicht so sicher. Wie machen wir das und was macht Sinn?

Hier ein Vorschlag, die Aufgabe 2 systematisch anzugehen:

- Gehe in jeden Datenpunkt in der Test/Realität- Datensatz und finde alle Punkte, die „dicht genug“ (zB kleiner als  $\sqrt{2}$ ) an einer der in Aufgabe 1 gefundenen 4 idealen Funktion liegen. Dh ich mache das jetzt vier mal, für jede ideale Funktion.
- Dann kann man sagen: ich habe n Test Daten ( wo  $n < N$  mit  $N$ =Anzahl aller Test Daten), die sich an einer meiner vier idealen Funktionen anfitzen lassen. Natürlich kann es passieren, dass ein Punkt nahe an zwei Funktionen liegt, aber das ist ok. Für jede Funktion speichert man dann die entsprechend gefunden n Test Daten ab.

Dh die vier Funktionen sind Projektionen der Realität. Es gibt aber keine einzige, die die gesamte Realität abbildet. Aufgabe zwei ist die Segmentierung meine „gemessenen“ Test-Daten nach verschiedenen Hypothesen, die durch meine 4 idealen Funktionen dargestellt werden.