GKL671 Einführung Data Science

Datascience ist ein interdisziplinärer Ansatz zur Extraktion von Wissen aus Daten. Es umfasst eine Vielzahl von Techniken und Methoden, darunter Datenanalyse, Datenvisualisierung, maschinelles Lernen und Statistik.

Machine Learning (ML) ist eine Methode der Künstlichen Intelligenz, bei der Algorithmen und statistische Modelle verwendet werden, um aus Daten zu lernen und Vorhersagen oder Entscheidungen zu treffen. ML-Modelle können in der Lage sein, Muster in Daten zu erkennen und diese Muster zu nutzen, um zukünftige Ereignisse vorherzusagen oder Entscheidungen zu treffen.

Künstliche Intelligenz (KI) bezieht sich auf Technologien und Systeme, die menschenähnliche Intelligenz und kognitive Fähigkeiten demonstrieren. Dazu gehören z.B. Spracherkennung, Bilderkennung, automatisierte Entscheidungsfindung und maschinelles Lernen.

Datenexploration ist der Prozess, bei dem große Mengen von Daten untersucht und analysiert werden, um Muster und Zusammenhänge zu erkennen. Die Daten können auf verschiedene Weise visualisiert werden, z.B. durch Diagramme, Histogramme oder Heatmaps, um Trends und Muster zu identifizieren und zu interpretieren. Die Ergebnisse der Datenexploration können dann als Grundlage für weitere Analysen und Entscheidungen verwendet werden.

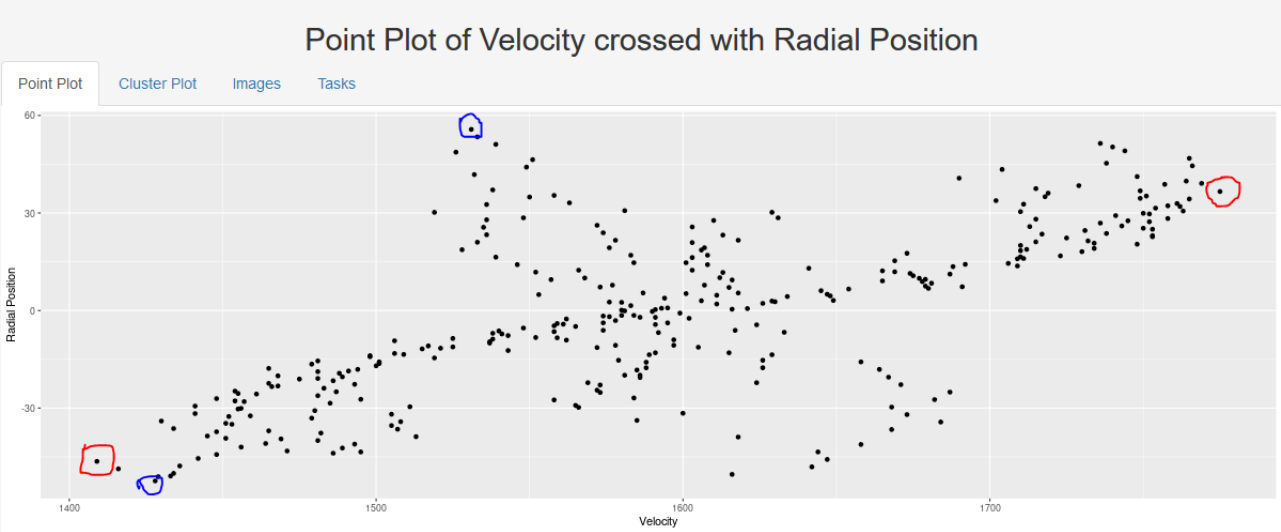
Ein Data Scientist ist eine Person, die sich mit der Analyse von Daten beschäftigt. Die Arbeit eines Data Scientists umfasst das Sammeln, Bereinigen, Speichern und Analysieren von Daten. Sie verwenden verschiedene Tools und Technologien, um Erkenntnisse aus Daten zu gewinnen und diese Erkenntnisse in Entscheidungen umzusetzen.

Zu den Aufgaben eines Data Scientists gehören auch das Entwickeln und Trainieren von maschinellen Lernmodellen, um Vorhersagen und Entscheidungen automatisch zu treffen. Sie müssen auch in der Lage sein, die Ergebnisse ihrer Analysen auf verständliche Weise zu präsentieren und anderen Personen zu erklären, wie sie zu diesen Ergebnissen gelangt sind. Data Scientists arbeiten in der Regel in Unternehmen, Regierungsbehörden oder Forschungseinrichtungen und unterstützen bei der Optimierung von Geschäftsprozessen, Produkten und Dienstleistungen durch den Einsatz von Datenanalyse und maschinellem Lernen.

Clustering ist eine Methode in der Datenanalyse, bei der ähnliche Datenobjekte in Gruppen (Cluster) zusammengefasst werden. Das Ziel von Clustering besteht darin, eine natürliche Struktur in den Daten zu identifizieren, indem Objekte gruppiert werden, die sich in bestimmten Merkmalen ähneln.

Praktische Aufgabestellung:

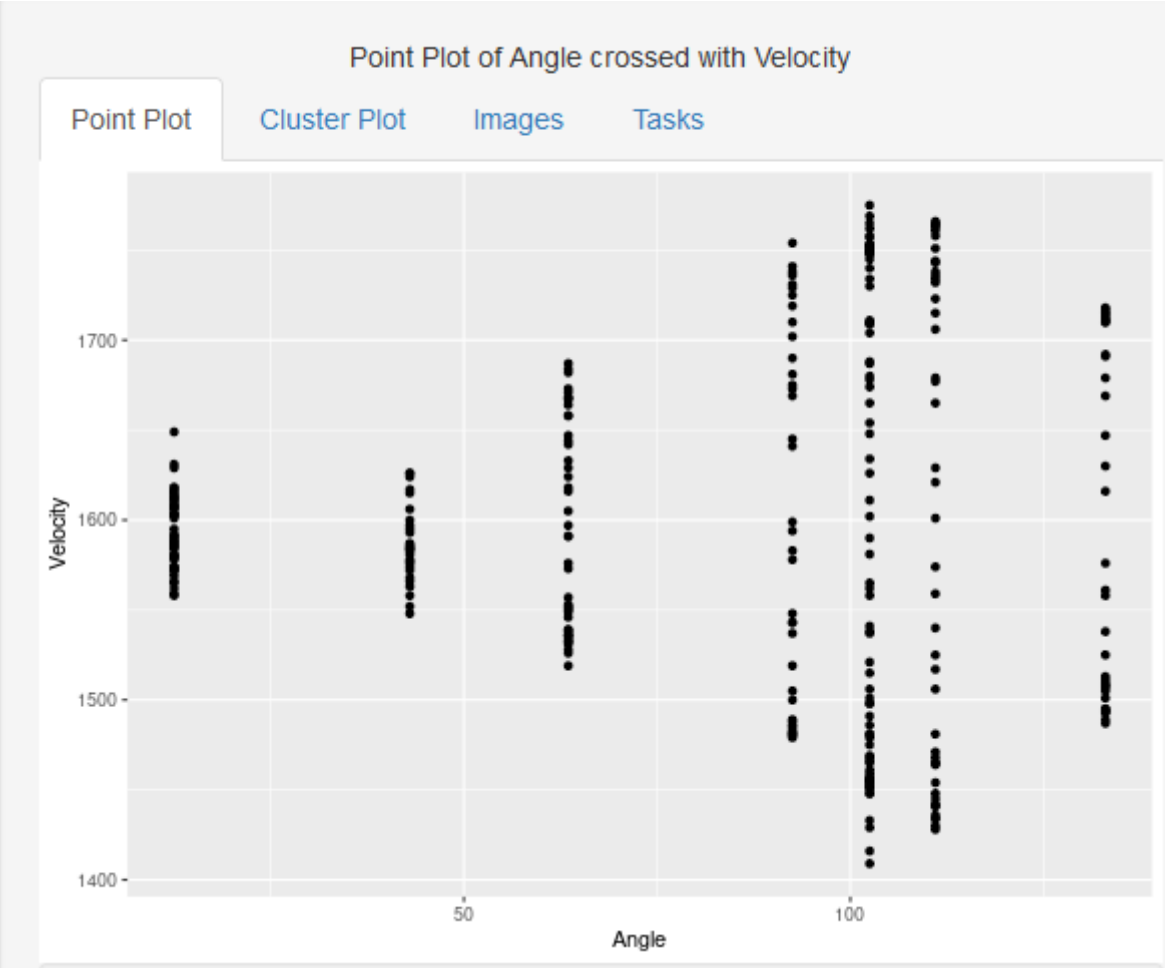
Ermittle die größte und kleinste Geschwindigkeit in der beobachteten Galaxie und ihre minimale und maximale radiale Position.



Rot: Für die Geschwindigkeit

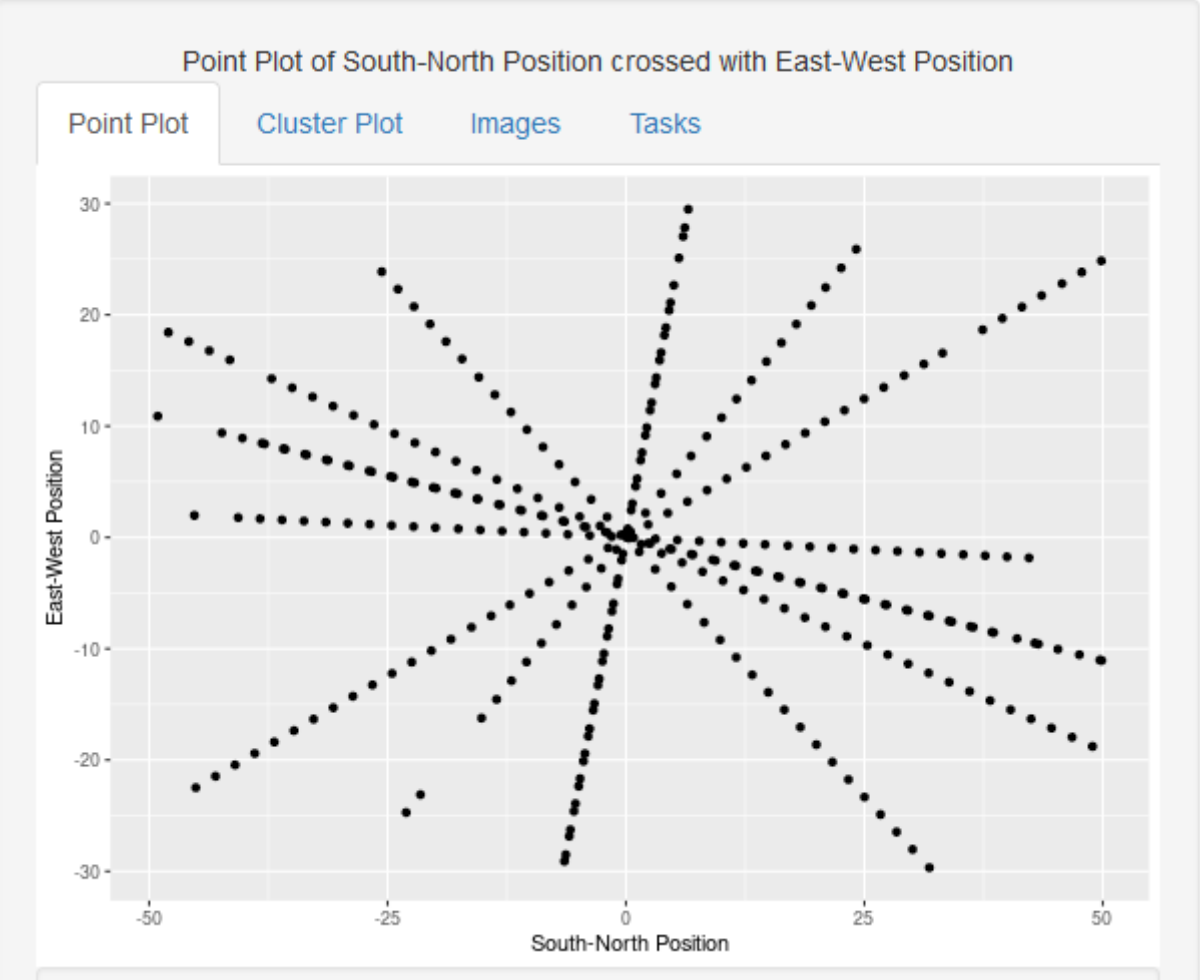
Blau: Für die Radiale Position

Ermittle, bei welchem Winkel die extremsten Geschwindigkeiten angenommen werden.



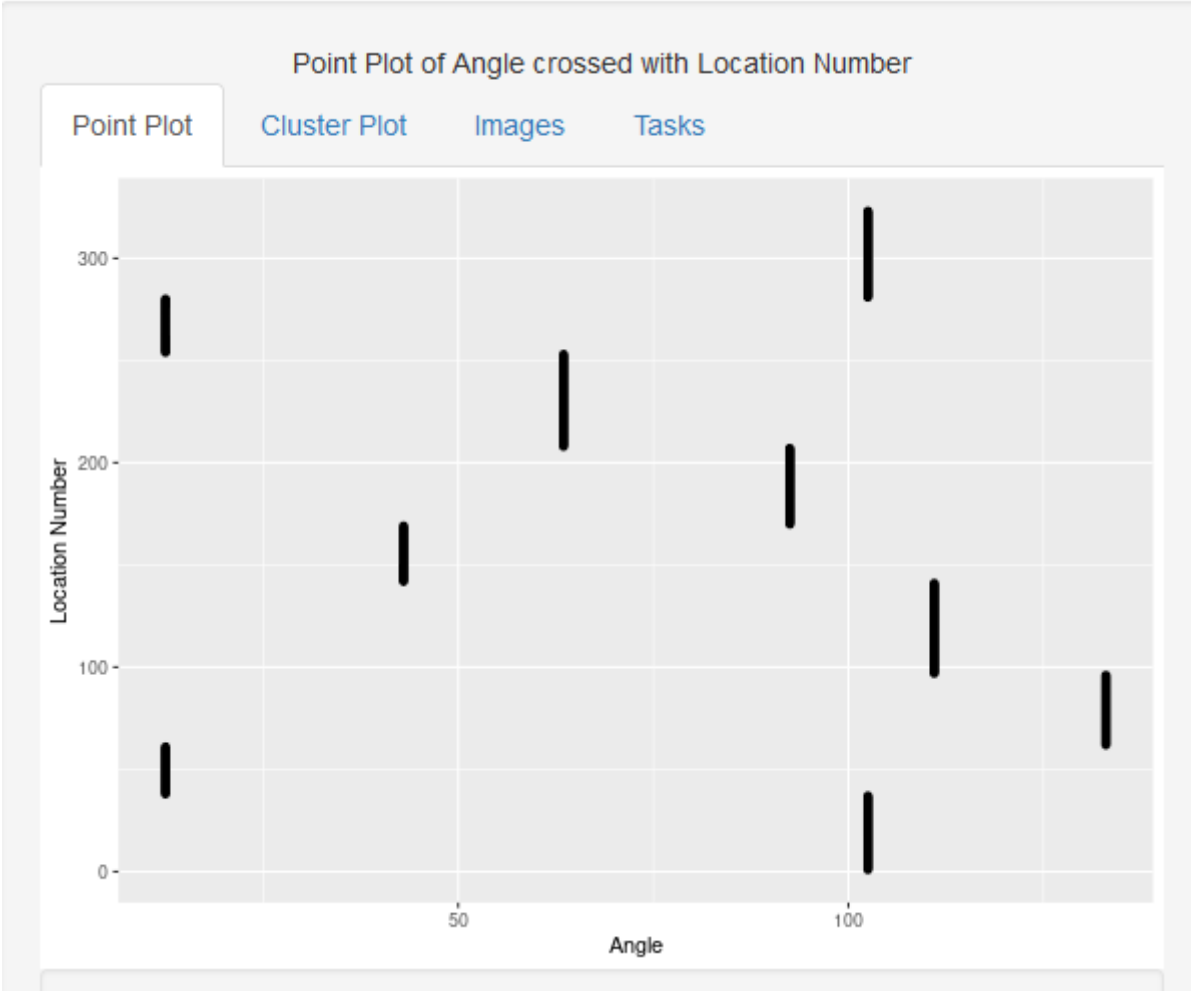
Die Maximalgeschwindigkeit ist ungefähr bei 107°

Beschreibe, welche Information sich in einer Punktwolke darstellen lässt, wenn die Nord-Süd-Richtung gegen die Ost-West-Richtung aufgetragen wird.



Wenn man die Beiden Skalierungen kreuzt und das Diagramm 1:1 ist, ist es möglich Winkel der Messungen zu Visualisieren

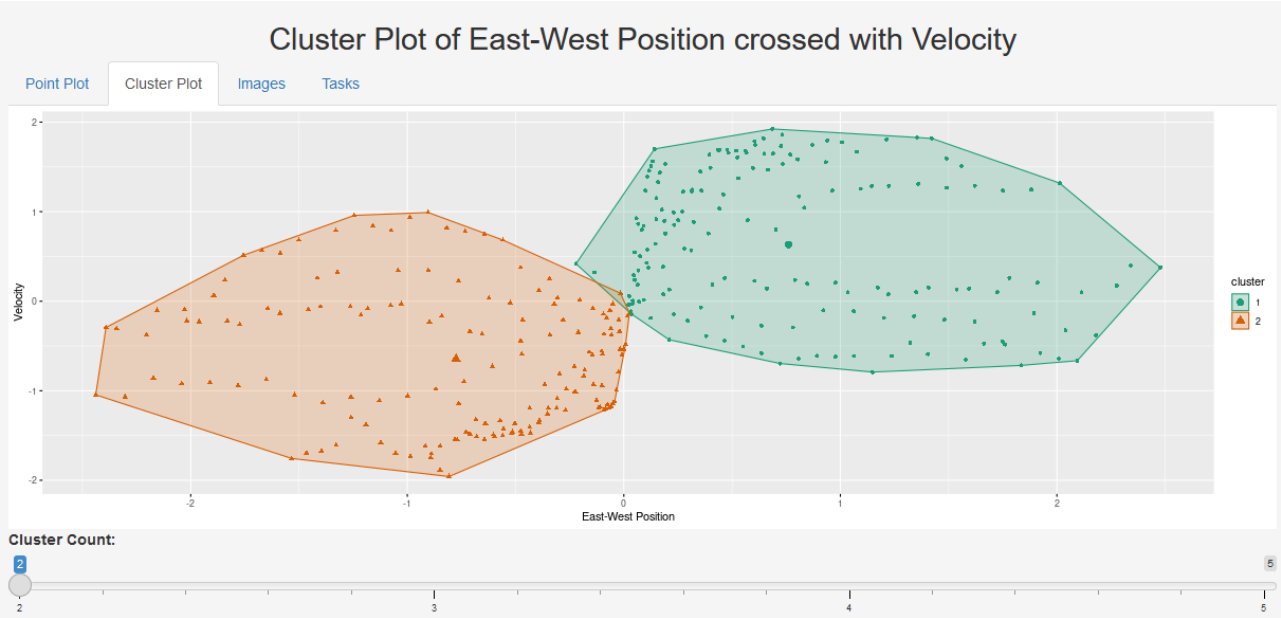
Überlege, welche Information sichtbar wird, wenn der Winkel gegen die Location Number aufgetragen
wird.



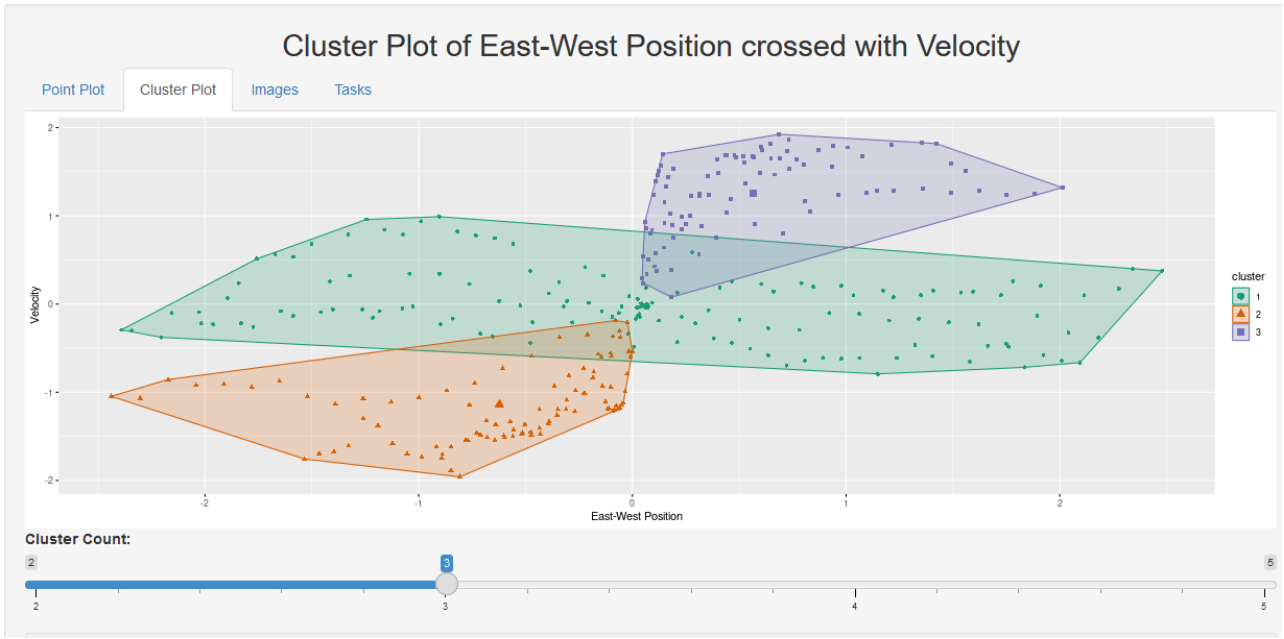
Durch die Kreuzung der Positionsnummer und des Winkels kann man herausfinden, wie oft das Teleskop die Galaxie durchquert und Messungen durchgeführt hat. Zusätzlich kann man die Reihenfolge der Messungen einsehen.

Vergleiche, ob die Anpassung von 2, 3, 4 oder 5 Clustern am geeignetsten ist und woran man das
erkennen kann. Betrachte dabei auch die Darstellung der Cluster als Überlagerung der
zugrundliegenden Aufnahme des Teleskops.

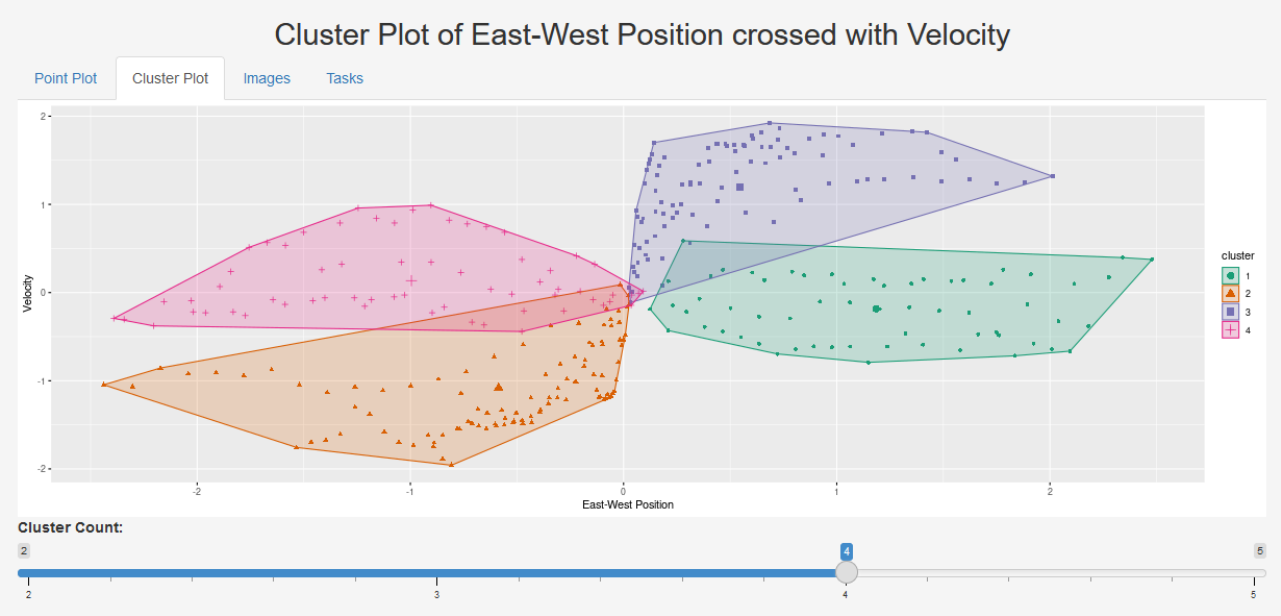
Plot mit 2 Cluster:



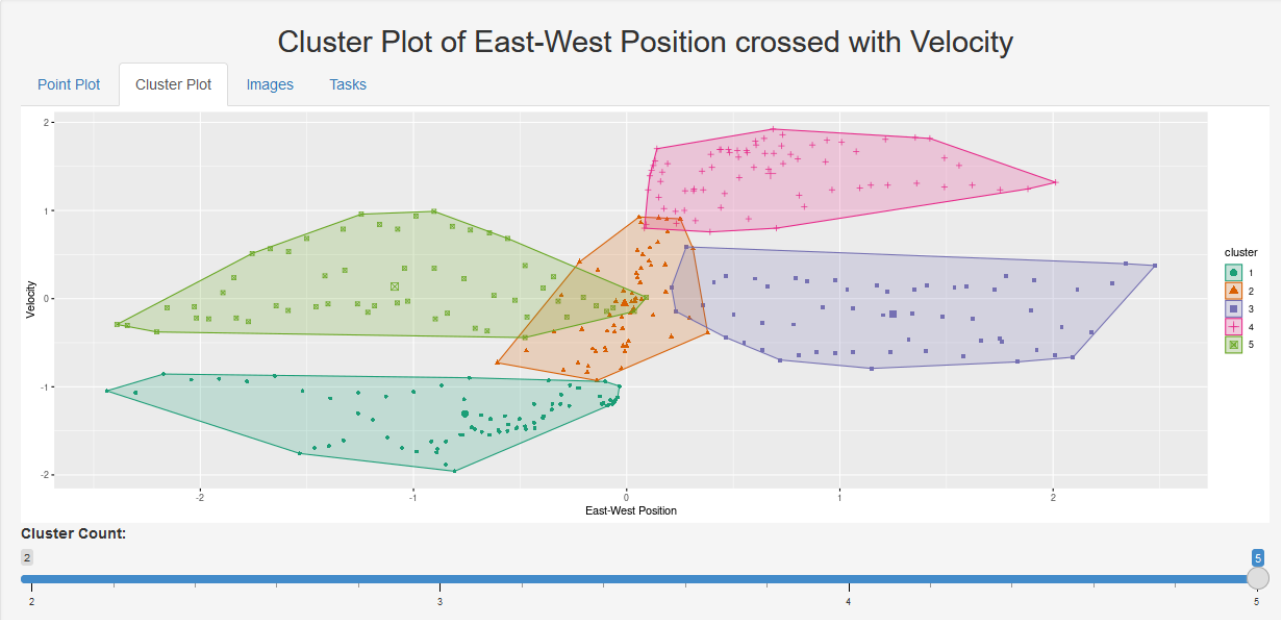
Plot mit 3 Cluster:



Plot mit 4 Clustern:



Plot mit 5 Clustern:



Durch die Anpassung der Cluster können Gruppierungen von Daten visualisiert werden, die bestimmte Ähnlichkeiten aufweisen. Basierend auf diesem Kriterium kann entschieden werden, welche Cluster angewendet werden sollen. In diesem speziellen Beispiel wurden 5 Cluster ausgewählt, um die langsamsten, schnellsten, höchsten und niedrigsten Punkte einfach bestimmen zu können.