

Use Case – EDA mit Rapid Miner

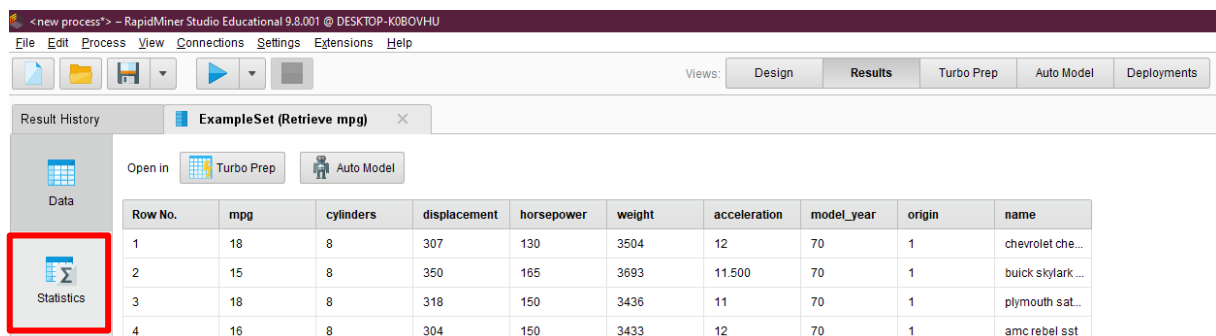
In einem vorherigen Use Case haben Sie die mpg-Daten in Ihr Repository geladen. Dieser Use Case nutzt diese Daten um damit eine kleine Explorative Datenanalyse durchzuführen und einen genaueren Überblick über die Daten zu bekommen.

Vorbereitung

1. Erstellen Sie einen neuen Prozess und ziehen Sie die mpg-Daten in den Prozessbereich.
2. Verbinden Sie die Daten nun mit dem „res“ Knoten.
3. Führen Sie den Prozess aus.

Statistische Kennzahlen

Rapid Miner bietet Ihnen über den „Statistics“-Übersicht des Result Tabs schon eine ungefähre Übersicht über wichtige Kennzahlen, wie beispielsweise Maximum, Minimum, Mittelwert, Abweichung und fehlende Werte.



Row No.	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	name
1	18	8	307	130	3504	12	70	1	chevrolet che...
2	15	8	350	165	3693	11.500	70	1	buick skylark ...
3	18	8	318	150	3436	11	70	1	plymouth sat...
4	16	8	304	150	3433	12	70	1	amc rebel sst

Wechseln Sie zur „Statistics“-Übersicht und schauen Sie sich die verschiedenen statistischen Kennzahlen an, die Sie bereits kennengelernt haben.

Können Sie erkennen bei welcher Variable wohl Datenqualitätsprobleme vorliegen?

Univariate Visualisierungsmethoden

Rapid Miner bietet sowohl die statistischen Kennzahlen, als auch die Möglichkeit einer grafischen Auswertung an. Im folgenden lernen Sie wie die gängigsten grafischen Auswertungen in Rapid Miner umgesetzt werden.

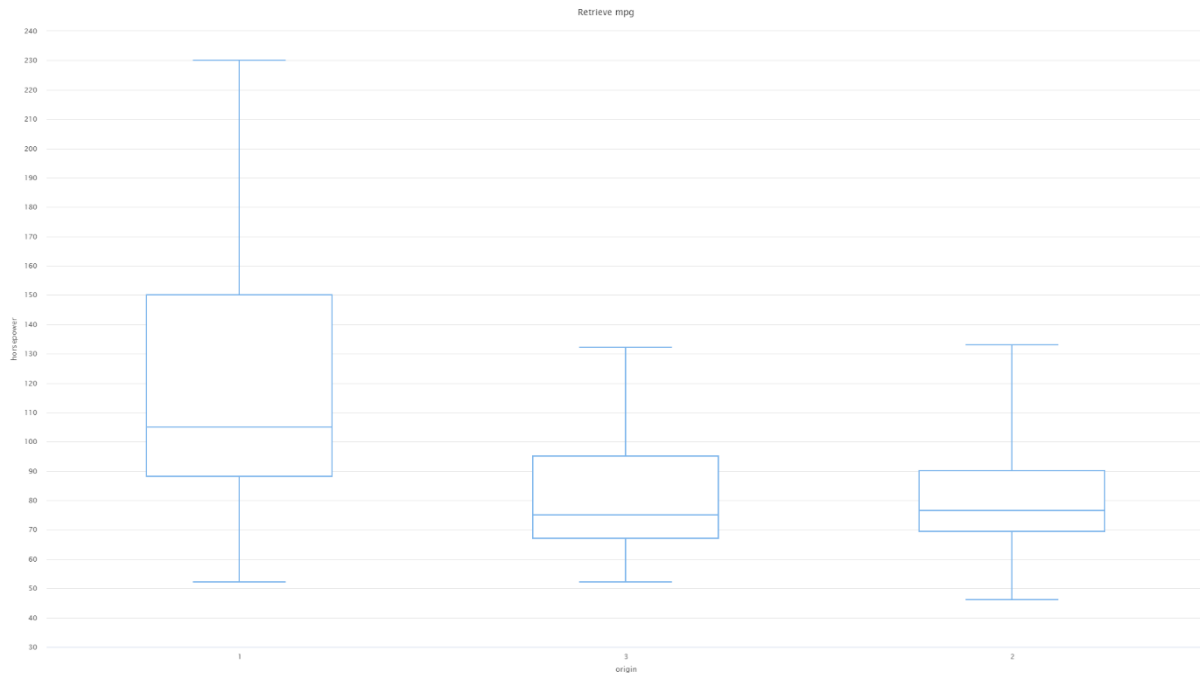
Boxplot:

1. Wechseln Sie auf „Visualizations“.
2. Stellen Sie als „Plot Type“ Boxplot ein.
3. Die „Value Column“ steht für die Variable die Sie näher betrachten möchten. In diesem Fall wählen Sie z.B.: „horsepower“. Es erscheint nun ein Boxplot, welcher die Verteilung der ganzen Variable anzeigt.
4. Um genauere Aussagen bezogen auf die Zielvariable „origin“ treffen zu können, wollen wir jetzt noch bei „Group by“ die Variable „origin“ einstellen.

Hinweis: Falls Sie „origin“ nicht zur Auswahl haben folgen Sie bitte den Anweisungen am Ende des Dokuments !

5. Sie sehen nun 3 Boxplots, aus denen näher ersichtlich wird wie die Werte der Variable horsepower im Bezug auf die Zielvariable verteilt ist.
6. Fahren Sie mit der Maus über eine der „Boxen“. Sie müsste nun die Five-Number-Summary sehen.

Ihre Grafik sollte etwa so aussehen:



Histogramm:

Eine weitere Form der Univariaten Visualisierung ist ein Histogramm. Es zeigt Ihnen die Verteilungen der Daten, gruppiert zu sogenannten „bins“. Umso weniger bins Sie wählen, umso „glatter“ wird Ihre Auswertung. Haben Sie es zum Beispiel mit stark „rauschenden“ Daten zu tun, können Sie dieses Feature benutzen um den Einfluss davon zu minimieren. Im Gegenteil können Sie durch Hinzufügen von „bins“ Verteilungen genauer erkennbar werden.

1. Wählen Sie als „Plot type“ das Histogramm aus.
2. Nehmen Sie das Attribut „mpg“ in die „Value Columns“.
3. Sie sehen nun die Verteilung des Attributs „mpg“ über alle Klassen.
4. Versuchen Sie die Number of bins zu reduzieren. Setzen Sie beispielsweise 10 ein.

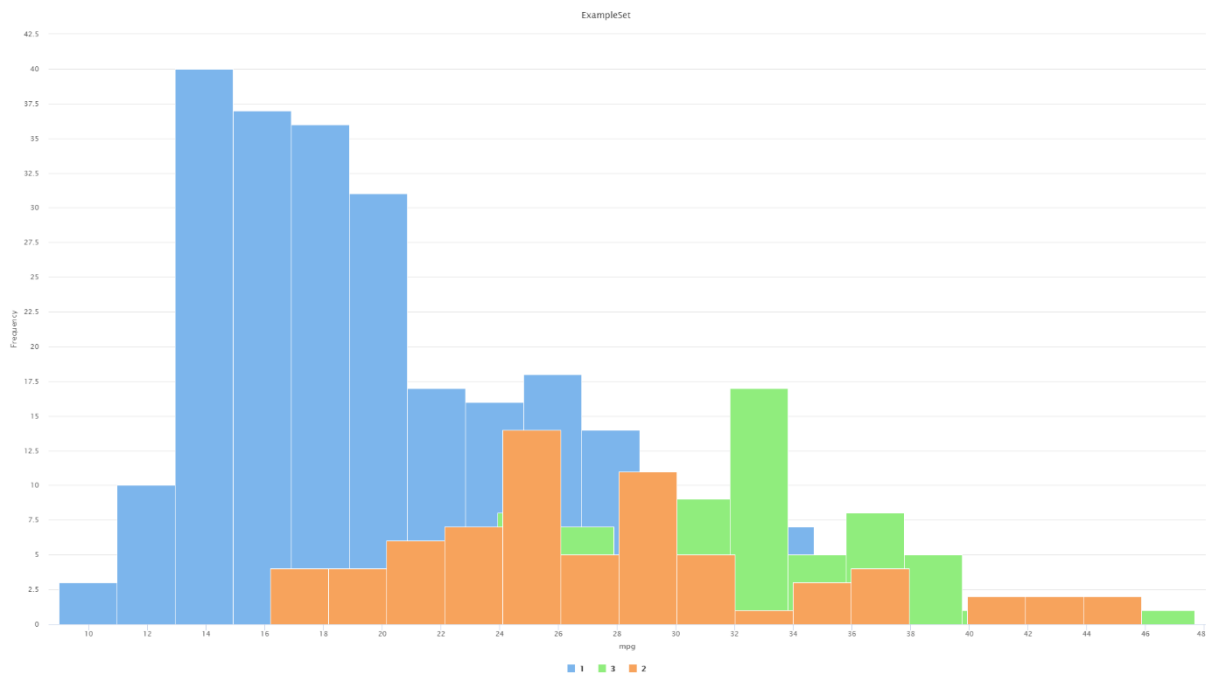
Was stellen Sie fest?

Stellen Sie als Gegenbeispiel die „Number of bins“ auf 30 und analysieren Sie die Grafik erneut.

Es ist ebenso möglich die Verteilungen klassenspezifisch zu betrachten.

5. Stellen Sie das Attribut „Color“ auf „origin“.
6. Wenn Sie nun über die verschiedenen Farben in der Legende hovern, können Sie sehen, dass die Attribute je nach Klasse unterschiedliche Verteilungen aufweisen und eine Gesamtansicht wohl nicht sehr aufschlussreich ist.

Ihre Grafik sollte etwas so aussehen:



Multivariate Visualisierungsmethoden

Wenn Sie Visualisierungen über mindestens 3 oder mehr Dimensionen machen wollen, ist es ratsam eine der multivariaten Visualisierungsmethoden zu benutzen, wie beispielsweise einen Scatterplot.

Hinweis: Durch Features wie „Color“ oder „Group by“ können Univariate Visualisierungsmethoden auch multivariat eingesetzt werden, Sie sollten jedoch aufpassen, wann sich welche Methode besser eignet!

Scatterplot:

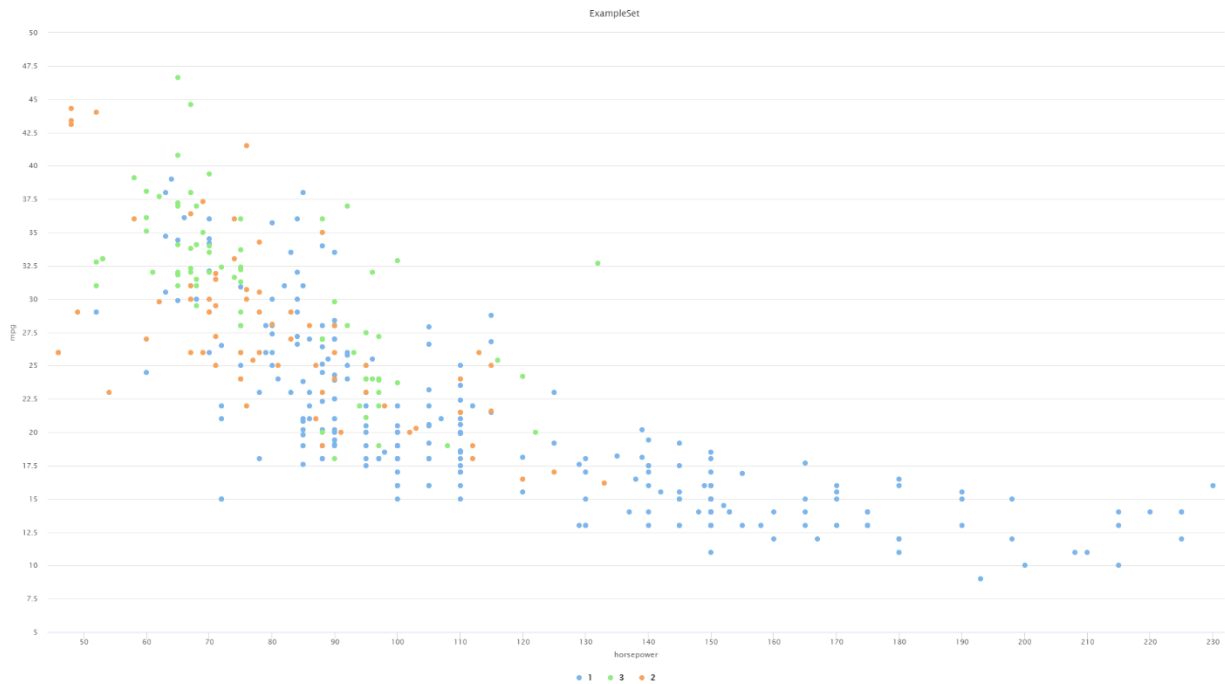
Ein Scatterplot ist eine gute Methode um Zusammenhänge mehrerer Variablen zu untersuchen. Zum einen lässt sich ein erster Eindruck über Korrelation gewinnen, zum anderen können auch durch Features wie „Color“ oder „Size“ eine eventuelle Klassentrennung abgeleitet werden. Falls Sie viele Datenpunkte in Ihrem Dataset haben, können Sie bei einem Scatterplot durch das Feature „Jitter“ ein sogenanntes „Rauschen“ hinzufügen, um Datenpunkte besser voneinander separieren zu können (Hier nicht benötigt).

1. Wählen Sie als Chart Type den „Scatterplot“.
2. Übernehmen Sie „horsepower“ für die X-Achse.
3. Wählen Sie mpg als „Value-Column“ (y-Achse).
4. Selektieren Sie „origin“ für das Feature „Color“.

Was können Sie aus der Visualisierung ableiten?

Was heißt dies für Ihre mögliche Feature Auswahl falls Sie ein ML-Modell erstellen wollten?

Ihre Grafik sollte etwas so aussehen:

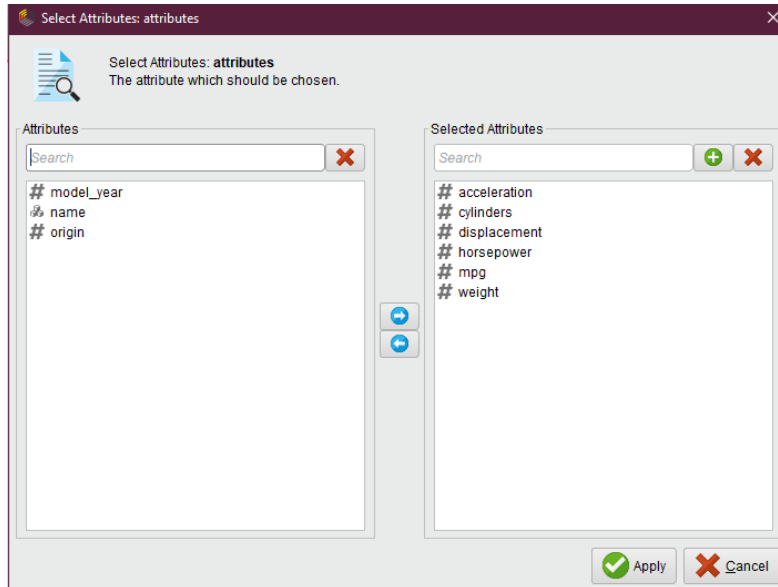


Sie haben nun einen kleinen Überblick über mögliche Kennzahlen & Visualisierungsmethoden um einen ersten Eindruck über Ihre Daten zu bekommen. Natürlich gibt es noch viele weitere Analysen, die jedoch an dieser Stelle nicht vertieft werden.

Correlation Matrix

Eine sehr wichtige Analysemethode ist das Erforschen von Beziehungen zwischen verschiedenen Variablen und Einflussfaktoren. Hierbei kann eine Correlation Matrix hilfreich sein, gerade wenn es sich um den Vergleich mehrerer Dimensionen handelt.

1. Wechseln Sie in das „Design“-Tab.
2. Fügen Sie den „Correlation Matrix“- Operator in den Prozessbereich ein.
3. Wählen Sie den Operator per Linksklick aus → attribute filter type : subset → Select Attributes
4. Wählen Sie alle numerischen Variablen aus → Apply.



5. Verbinden Sie den „mat“ Knoten mit einem „res“ Knoten.
6. Verbinden Sie den „exa“ Knoten mit einem „res“ Knoten.

Ihr Prozess sollte so aussehen:



Sie sollten folgende Matrix sehen:

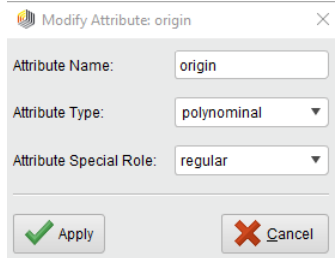
Attribut...	mpg	cylinders	displac...	horsepo...	weight	acceler...
mpg	1	-0.775	-0.803	-0.778	-0.831	0.418
cylinders	-0.775	1	0.951	0.843	0.896	-0.505
displace...	-0.803	0.951	1	0.897	0.933	-0.544
horsepo...	-0.778	0.843	0.897	1	0.865	-0.689
weight	-0.831	0.896	0.933	0.865	1	-0.417
accelerat...	0.418	-0.505	-0.544	-0.689	-0.417	1

Wie sieht die Beziehung zwischen „mpg“ und „horsepower“ aus? Ist dies ein Kausalzusammenhang?

Workaround falls „origin“ nicht als „Group By“ angezeigt wird:

In diesem Fall wurde „origin“ beim Einlesen als numerischer Wert erkannt, obwohl es sich hierbei eindeutig um eine Klasse handelt. Diese repräsentieren meist kategorische Features. Dies muss nachträglich noch geändert werden.

1. Navigieren Sie in Ihrem Repository zum mpg-Dataset
2. Rechtsklick → Edit
3. Sie sehen unter dem Prozessbereich, dass Sie eine Vorschau des Datasets angezeigt bekommen.
4. Rechtsklick auf Origin → Modify Attribute → Attribute Type = polynomial



5. Apply → Speichern des Datasets



Der Fehler sollte nun behoben sein.