A4LR1 „Absatzprognose“

Hintergrund:

Da es für jedes Unternehmen sinnvoll ist, eine Einschätzung zu haben, welche Mengen in Zukunft verkauft werden, möchte auch die Lichtgestalten GmbH eine Prognose für ihren Absatz vornehmen. Dabei sollen die abgesetzten Mengen an Leuchtmitteln und Ausgaben für Werbung für die Jahre 2020, 2021 und 2022 herangezogen werden. Ihre Aufgabe ist es nun, zunächst zu ermitteln, welche Variable am besten als unabhängige Variable geeignet ist (Zeit vs. Werbeausgaben). Anhand davon wird entschieden, ob mit einem Kausalmodell oder einer Zeitreihenanalyse weitergearbeitet wird. (Kleiner Spoiler vorab zur Selbstkontrolle: wir arbeiten mit einem Kausalmodell weiter). Darauf aufbauend sollen Sie einen dafür geeigneten Workflow erstellen und ein Lineares Regressionsmodell trainieren, mit dem eine Prognose für die verkauften Leuchtmittel vorgenommen werden kann.

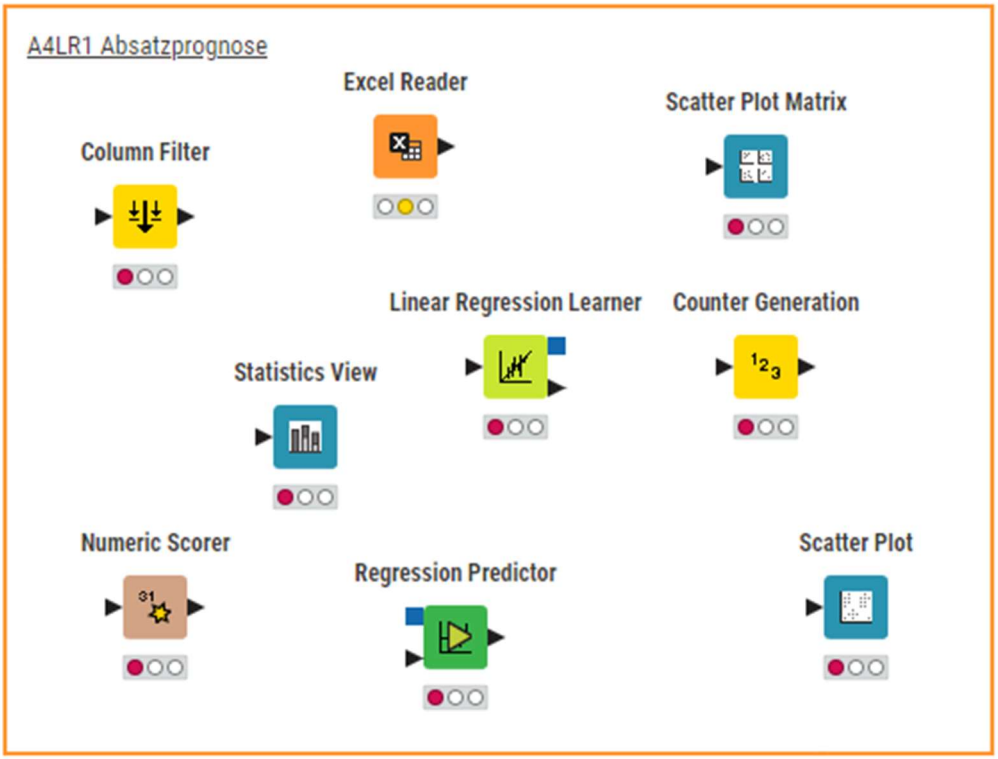
An dieser Stelle sei betont, dass diese Aufgabe stark vereinfacht ist. Die Realität ist selbstverständlich vielschichtiger, da werden für Absatzprognosen oft Modelle, in denen die Zeit die unabhängige Variable ist, verwendet. Diese gehören in das Gebiet der Zeitreihenanalyse. Darauf wird hier allerdings zu Gunsten eines leicht verständlichen Beispiels verzichtet, welches die Tür öffnen soll, die Fähigkeit zu erwerben, Regressionsmodelle mit KNIME trainieren und bewerten zu können.

Aufgabenstellung:

1. Beginnen Sie damit die Eingangsdaten in den Workflow zu laden: „Absatz_und_Werbeausgaben_2012_bis_2022.xlsx“. Nutzen Sie hierfür den „Excel Reader“-Knoten.
2. Zunächst soll eine Analyse der Eingangsdaten stattfinden. Dies soll mit Hilfe einer Scatter Plot Matrix geschehen. Für eine sauberere Darstellung soll in diesem Schritt ein Counter hinzugefügt werden. Verwenden Sie hierfür den „Counter Generation“-Knoten.
3. Der Knoten „Scatter Plot Matrix“ kommt nun zum Einsatz, um Scatter Plots zu erstellen, mit denen Sie die Korrelation der Bestandteile der Eingangsdaten betrachten können. Dabei ersetzt der Counter die Spalte Monat zur besseren Darstellung. Nun sollten Sie erkennen, dass die Werbeausgaben einen stärkeren Einfluss auf den Absatz haben als die Zeit. Daher werden wir für eine die folgende, einfache lineare Regression die Werbeausgaben als unabhängige Variable auswählen.
4. Filtern Sie nun die Eingangsdaten und entfernen Sie dafür die Spalte mit den Monaten. Das bewerkstelligen Sie mit dem „Column Filter“-Knoten.
5. Trainieren Sie mit den nun geladenen Daten ein Lineares Regressionsmodell. Ihr Ziel ist es, in Zukunft die Werbeausgaben zielgenau auf den gewünschten Absatz einzustellen. Hierbei hilft Ihnen der „Linear Regression Learner“-Knoten.
6. Das nun trainierte Modell gilt es nun anzuwenden. Verwenden Sie dafür den „Regression Predictor“-Knoten. Mit seiner Hilfe entwickeln Sie die Prognose des Absatzes in Abhängigkeit von den Werbeausgaben. Diese wird in den nächsten Schritten visualisiert und bewertet.
7. Mit einem Scatter Plot-Knoten können Sie nun die Punkte auf der Regressionsgeraden visualisieren, indem Sie die unabhängige Variable auf die X-Achse und die berechnete abhängige Variable auf die Y-Achse legen. (Tipp: Setzen Sie einen Haken bei „Domain bounds“).
8. Mit dem „Numeric Scorer“-Knoten ermitteln sie die Bestimmungsmaße des Modells. Richten Sie dabei ein besonderes Augenmerk auf den Wert „R^2“. Denken Sie darüber nach, was er Ihnen sagt. Vergessen Sie aber nicht auch die anderen Kennzahlen zu betrachten.

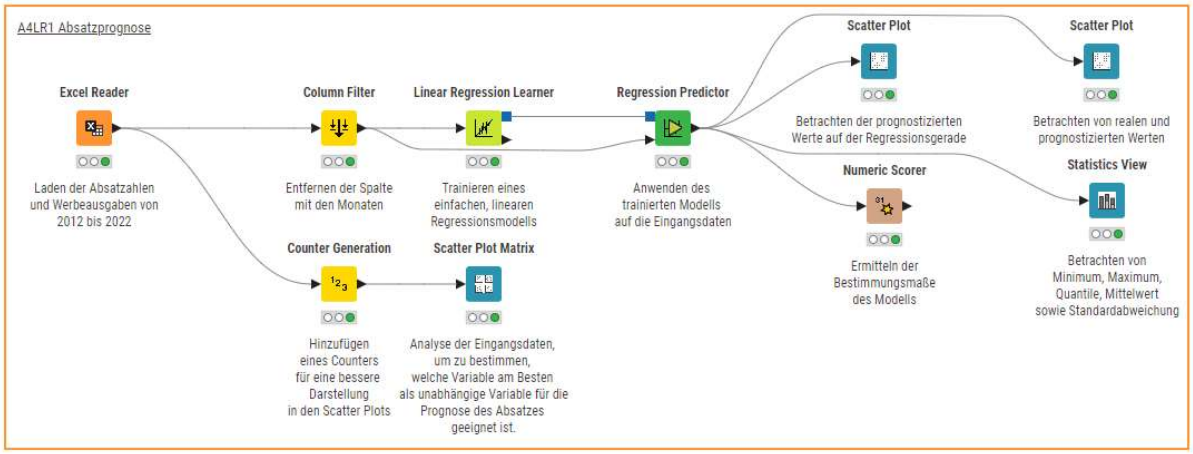
9. Ein weiterer „Scatter Plot“-Knoten wird nun dafür eingesetzt, Prognose und reale Werte zu vergleichen. Tragen Sie dabei den (realen) Absatz auf die x-Achse auf und die Prognose auf die y-Achse. (Tipp: Setzen Sie auch hier einen Haken bei „Domain bounds“ für die Bestimmung der unteren und oberen Grenzen der Achsen).
10. Zuletzt vergleichen wir noch diverse Werte wie Minimum, Maximum, Quantile, etc. für den realen Absatz und die Prognose. Dafür verwenden Sie den „Statistics View“-Knoten.

Verwendete Knoten:

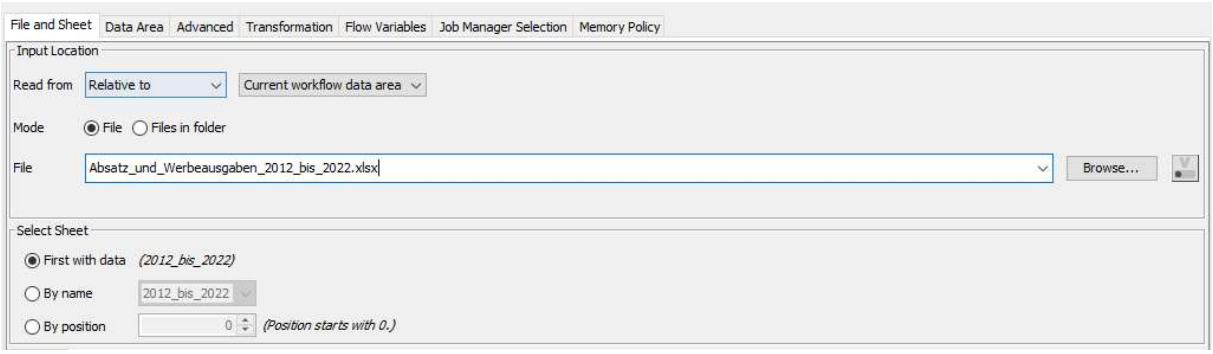


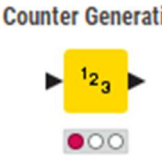
Einführung in die Aufgabe

Ziel der Aufgabe ist es, folgenden Workflow zu erschaffen:

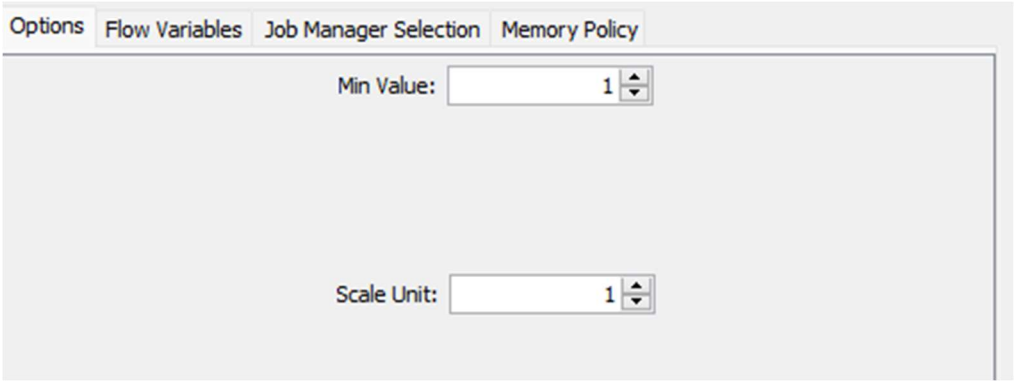


Beginnen wir zunächst mit dem „Excel Reader“-Knoten. Wie aus früheren Aufgaben hoffentlich bekannt, laden Sie hiermit die Ausgangsdaten „Absatz_und_Werbeausgaben_2012_bis_2022.xlsx“. Sollten Sie nicht mehr wissen, wie das funktioniert, so finden Sie eine detaillierte beschreibung zur Aufgabe A1DV1.

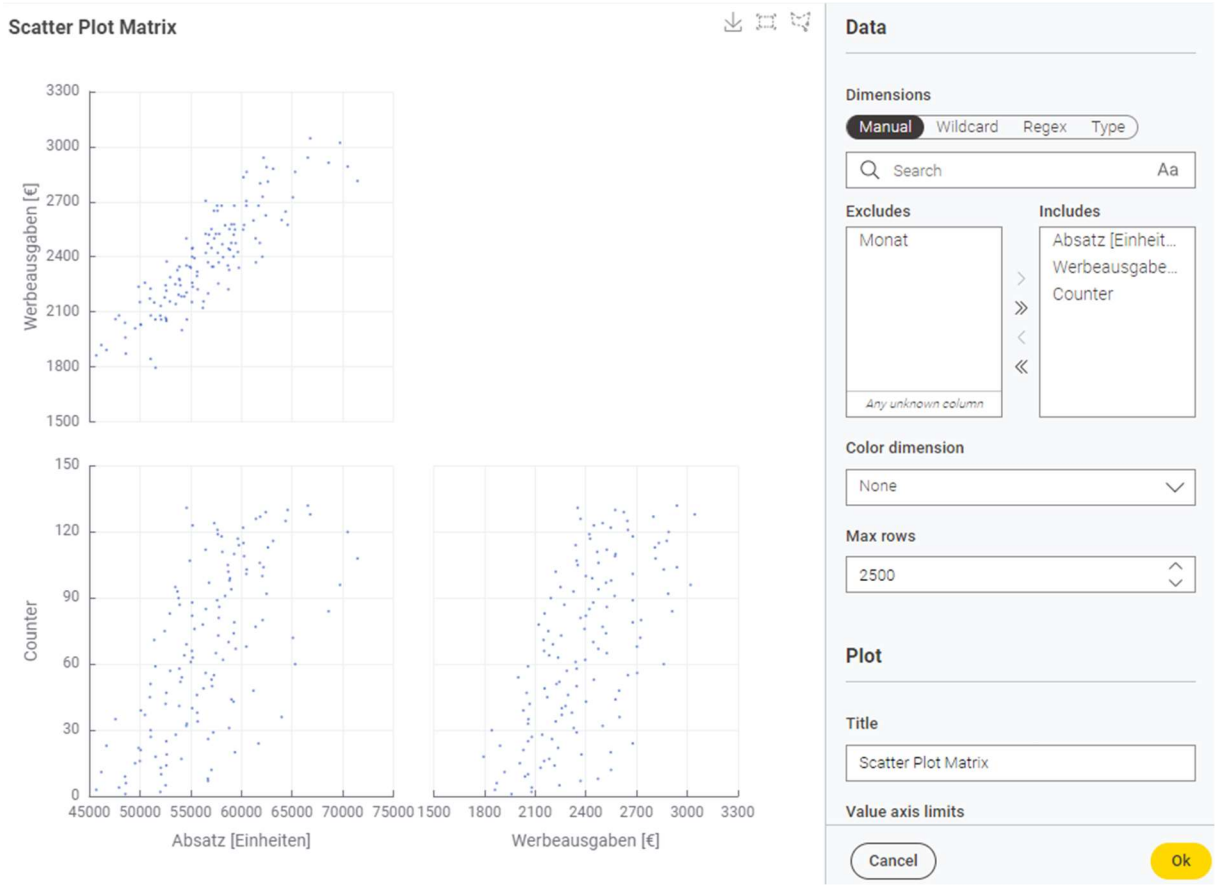




Zur Datenanalyse verwenden wir mehrere Scatter Plots. Bevor das machen, wollen wir allerdings noch die Daten für eine bessere Darstellung vorbereiten. Dafür verwenden wir den „Counter Generation“-Knoten. Stellen Sie ihn so ein, dass er bei eins startet und in 1er Schritten hochzählt.

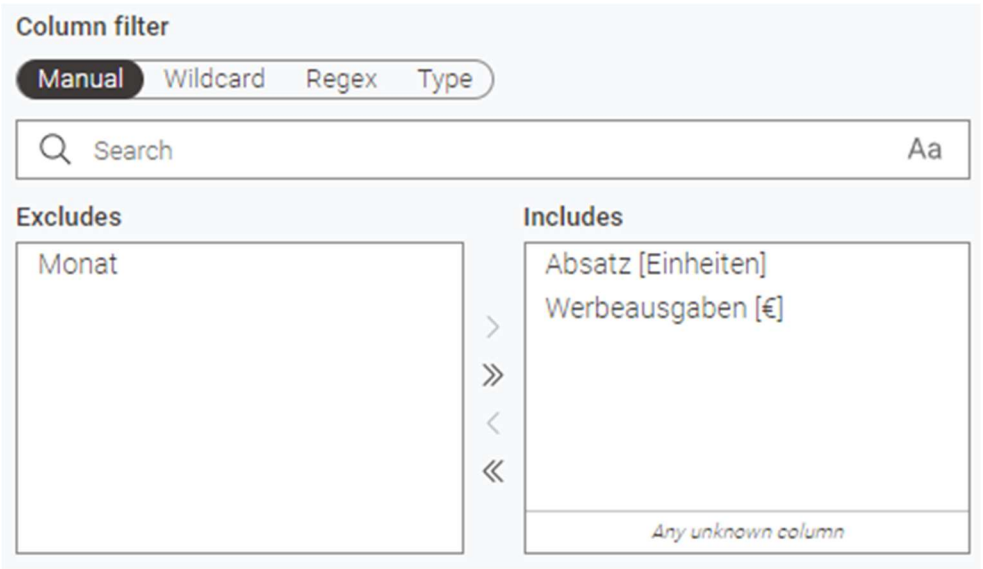


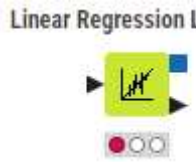
Damit wir die Korrelationen optisch betrachten können, wollen wir sie in mehreren Scatter Plots beieinander darstellen. Konfigurieren Sie den Knoten dafür folgendermaßen:



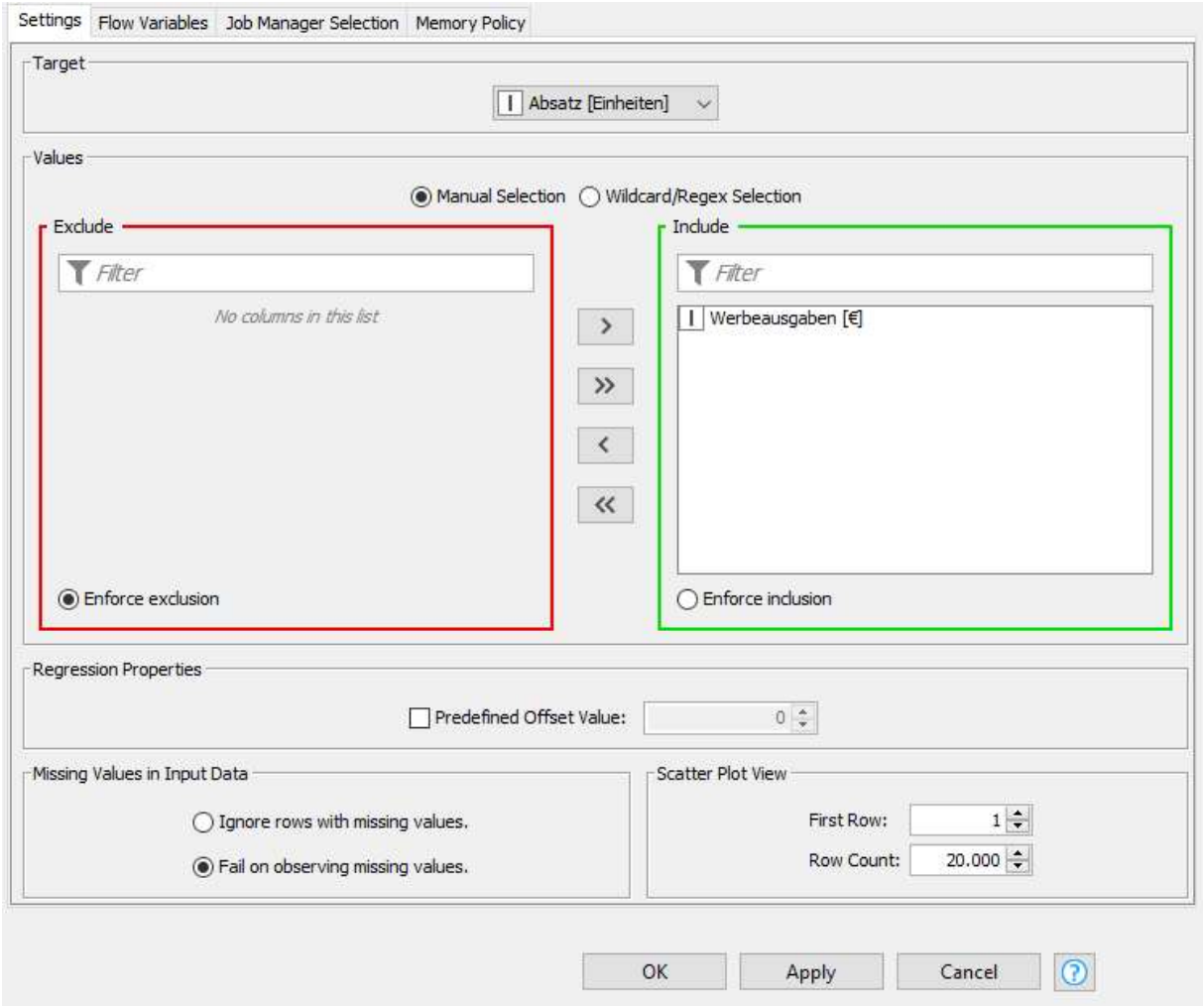
Nun sollten Sie erkannt haben, dass Werbeausgaben und Absatz stärker korrelieren als Zeit und Absatz. Aufgrund dieses Wissens wollen wir in den folgenden Schritten ein einfaches, lineares Regressionsmodell mit den Werbeausgaben als unabhängige Variable erstellen.

Zunächst filtern wir mit dem „Column Filter“-Knoten die Monatsspalte, die nicht benötigt wird.



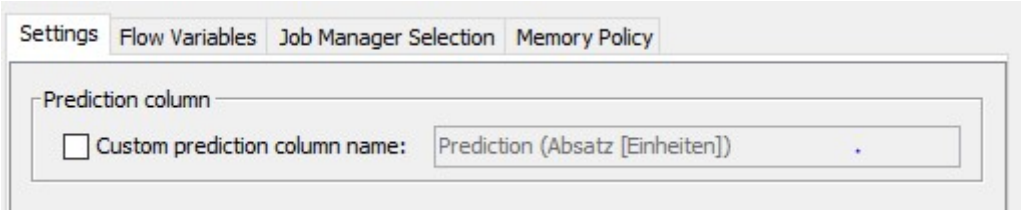


Nun fahren wir fort, das Modell zu trainieren. Verbinden Sie dafür die Eingangsdaten mit dem „Linear Regression Learner“-Knoten. Als Ziel stellen wir dort den Absatz [Einheiten] ein. Unter „Values“ inkludieren wir die Werbeausgaben [€].



Kontrollfrage: Wie lautet die Gleichung der Regressionsgeraden, die von dem Knoten gelernt wird?

Weiter geht es mit dem „Regression Predictor“-Knoten. In diesen laden wir das trainierte Modell und die Eingangsdaten. Für ebendiese soll eine Prognose erstellt werden. In diesem Knoten müssen wir ausnahmsweise nichts einstellen. Es besteht die Möglichkeit, die „Prognose-Spalte“ mit einem eigenen Namen zu versehen.

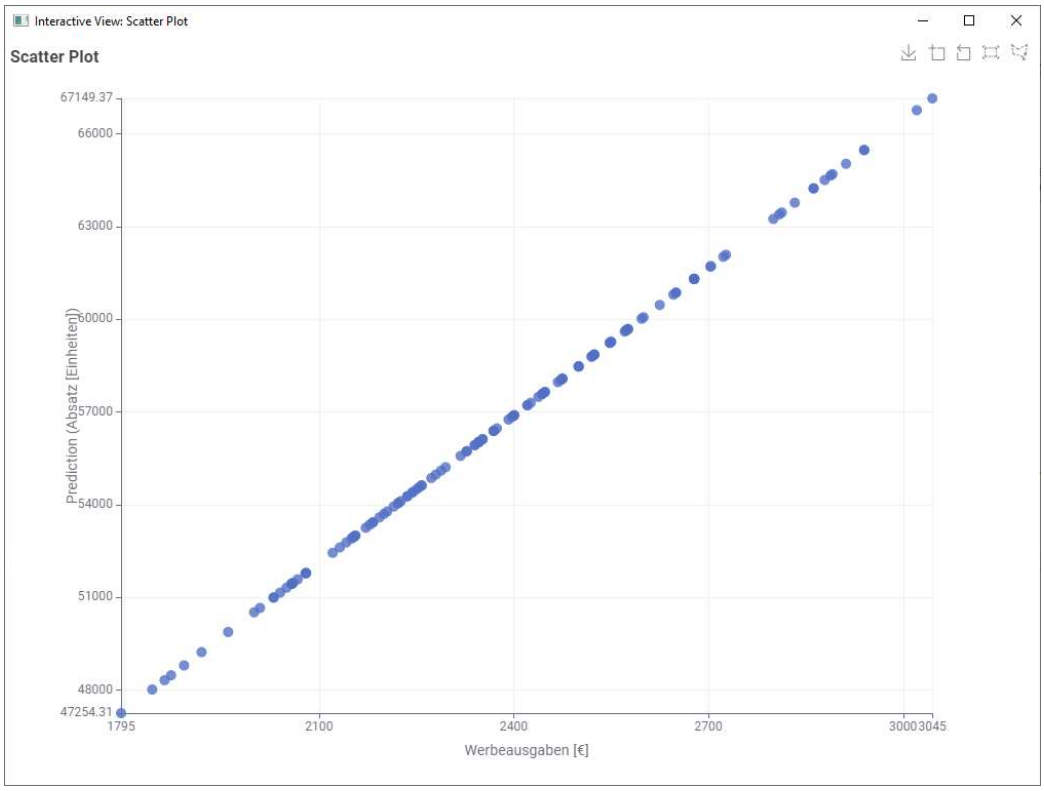


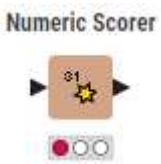
Folgendes Ergebnis sollten Sie nun erhalten:

Rows: 132 | Columns: 4 — Table | Statistics

#	RowID	Monat (Local Date)	Absatz [Einheiten] (Number (integer))	Werbeausgaben [€] (Number (integer))	Prediction (Absatz [Einheiten]) (Number (double))
1	Row0	2012-01-01	48548	1960	49,880.461
2	Row1	2012-02-01	52006	2079	51,774.471
3	Row2	2012-03-01	45671	1862	48,320.688
4	Row3	2012-04-01	47925	2080	51,790.387
5	Row4	2012-05-01	52523	2244	54,400.619

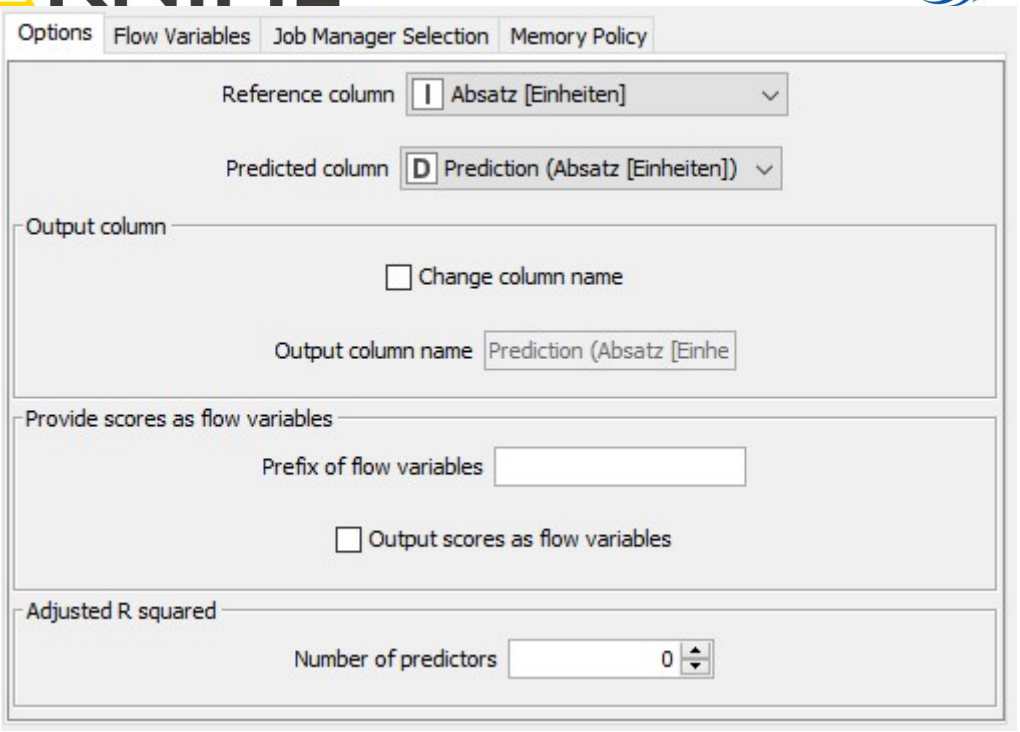
Die ermittelten Punkte sollten alle auf der Regressionsgeraden liegen. Um dies zu überprüfen, setzen wir einen Scatter Plot-Knoten in das Modell ein und legen die Werbeausgaben auf die X-Achse und die prognostizierten Absatzwerte auf die Y-Achse. Die Wertebereiche der Achsen können wir automatisch setzen lassen, indem wir für die Einstellung „Axis limits“ den Wert „Domain bounds“ wählen. Der Scatter Plot sollte folgendes Ergebnis zeigen:





Nun müssen wir allerdings auch eine Bewertung der Prognose vornehmen. Dafür werden wir im Folgenden drei weitere Knoten einsetzen.

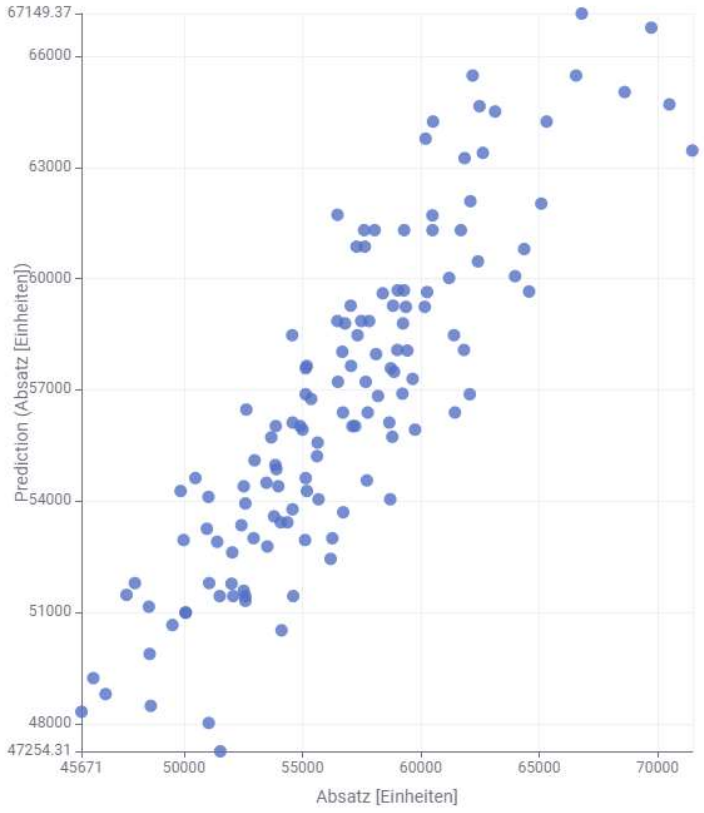
Beginnen wir mit dem „Numeric Scorer“. Er zeigt uns die Bestimmungsmaße. Verbinden Sie ihn dafür mit den Daten aus dem Regression Predictor“-Knoten. Stellen Sie dann die Referenzspalte und die prognostizierte Spalte ein.

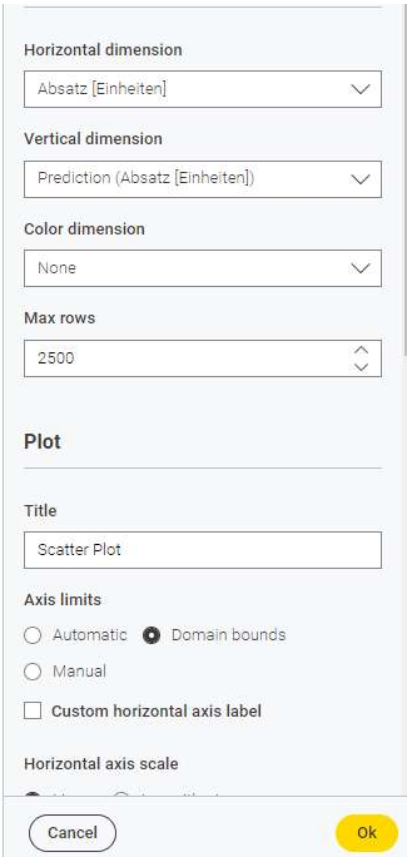


Die Bestimmungsmaße können Sie mit einem Blick in die Tabelle oder durch öffnen mit der Lupe einsehen. Sie sollten nun so aussehen:

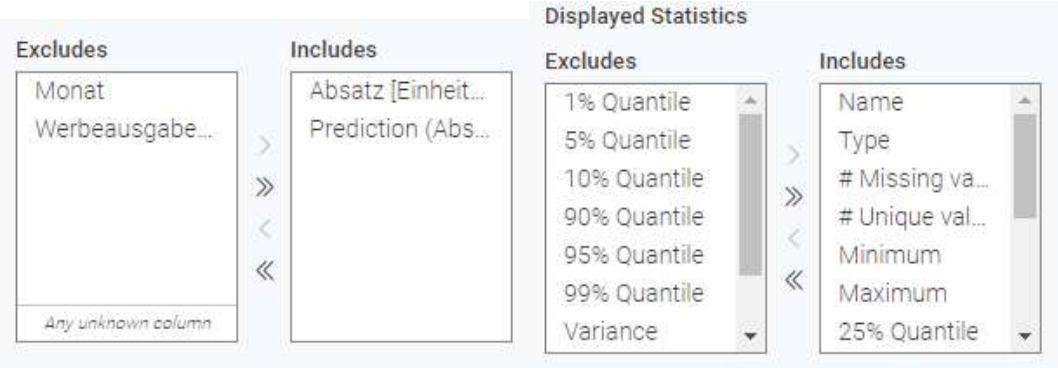
R^2:	0,758
Mean absolute error:	1.982,563
Mean squared error:	6.082.244,427
Root mean squared error:	2.466,221
Mean signed difference:	-0
Mean absolute percentage error:	0,035
Adjusted R^2:	0,758

Weiter geht es nun mit dem „Scatter Plot“-Knoten. Konfigurieren Sie ihn so, dass auf der horizontalen Achse der „reale“ Absatz aufgetragen wird und auf der vertikalen Achse die Prognose. Hilfreich ist es noch, wenn sie die Grenzen der Achsen auf die Grenzen der Werte legen. Daraus sollten Sie folgende Darstellung erhalten.





Zu guter Letzt wollen wir noch statistische Kennwerte der realen Werte und der Prognose betrachten können. Dabei kommt der „Statistics View“-Knoten zur Anwendung. Achten Sie darauf, dass Sie den „realen“ Absatz und die Prognose inkludieren. Alles andere wird exkludiert. Unter „Displayed Statistics“ können Sie sich konfigurieren, was Sie betrachten möchten.



Kontrollfrage: Geben Sie die 25-, 50- und 75%-Quantile sowie den Mittelwert dieser Statistik an!