Deskriptive Analyse

# Merkmale des Datensatzes

## Der originale Datensatz besteht aus:

* 81.504 Instanzen/Tupeln/Beobachtungen, davon sind
  + **43.319 in scope**
  + 38.185 out of scope
* Jede Instanz wird durch **10 Attribute** beschrieben:
* ORG (Origin/Startflughafen)
* DST (Destination/Zielflughafen)
* SECTOR\_ID (Sektor/Verbindung)
* SDD (Shipment Departure Date?)
* TOTAL\_CORE (Core Weight/Frachtgewicht - Masseneinheit unbekannt)
* ORG\_CTRY (Country of Origin/Herkunftsland)
* DST\_CTRY (Country of Destination/Zielland)
* ORG\_REGION (Region of Origin/Herkunftsregion/-kontinent)
* DST\_REGION (Region of Destination/Zielregion)
* IN\_SCOPE (Binärer Wert)
* Es gibt keinen Index / kein Attribut, dass eine Instanz eindeutig benennt

## Der Arbeitsdatensatz...

* ...wurde anonymisiert, indem wir alle Attributswerte mit Geoinformationen in eindeutige, 3-stellige Ziffernfolgen umwandelten. Die betroffenen Attribute sind:
* ORG
* DST
* SECTOR\_ID
* ORG\_CTRY
* DST\_CTRY
* ORG\_REGION
* DST\_REGION
* ...wurde auf die Instanzen mit Attributswert IN\_SCOPE = 1 eingeschränkt.

# Erste Annäherung an den Datensatz

## Darstellung des Transportgewichts, summiert über alle Sektoren

Im Folgenden werden die Transportgewichte über alle Sektoren tageweise in Jahresdiagrammen gezeigt. Die Jahre 2010 bis 2012 zeigen folgende Charakteristiken:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mittelwert | Max | Min | Stabw | Regressionsgerade | Bestimmtheitgrad |
| 2010 | 400126.663 | 726104 | 22785 | 154955.489 | y = 528.06x - 2E+07 | R² = 0.1291 |
| 2011 | 487853.088 | 771505.5 | 117446.5 | 141313.522 | y = 100.15x - 4E+06 | R² = 0.0056 |
| 2012 | 537016.347 | 917775 | 51696 | 179145.773 | y = 362.46x - 1E+07 | R² = 0.0457 |

Der Graph bildet die Entwicklung des akkumulierten Core Weight über alle „in scope“-Sektoren ab.

Das Jahr 2010 umfasst 364 Instanzen (Kalendertage). Die Range der Werte des Core Weight erstreckt sich dabei von 22785 bis 726104 und der Mittelwert beträgt 400126,66. Die maximalen Abweichungen vom Mittelwert belaufen sich auf -377341,66 downwards, sowie 325977,34 (nahezu symmetrisch, Differenz Max-Min: 51364,32). Im Jahresverlauf ist ein Aufwärtstrend zu erkennen (siehe Regressionsgerade, grün eingezeichnet). Der Bestimmtheitsgrad ist gering.

Die Daten verlaufen (abgesehen vom Wachstumstrend) in einem wiederkehrenden Muster und lassen sich in mehrere Teilabschnitte unterteilen. Diese Abschnitte beginnen mit einem Tief, nahe 0 und enden mit einem Peak, auf welchen ein direkter Absturz erfolgt.

Hypothesen:

* Teilabschnitte bilden Wochenverläufe ab
* Core Weight einer Instanz ist abhängig von der Anzahl der geflogenen Sektoren
* Der geringe Bestimmtheitsgrad spiegelt die extremen Schwankungen der Einzelwerte ab, die Verlässlichkeit einer naiven Prognose für die folgende Instanz ist gering
* Anstieg am Ende des Betrachtungszeitraumes ist bedingt durch das Weihnachtsgeschäft
* Die geringeren Werte (im Vergleich zu den Folgejahren) sind bedingt durch eine wachsende Anzahl an Sektoren „in scope“ oder aber eine bessere Datensammlung

Kritische Bereiche:

* Relevante Bank Holidays als spezieller Untersuchungsgegenstand (könnten durch farbliche Hervorhebung erkenntlich gemacht werden)
* Relevante Bank Holidays bzw. deren Einfluss auf das Core Weight sollten gesondert untersucht werden
* Der Bestimmtheitsgrad lässt keine zuverlässige naive Prognose für Instanz n+1 zu, allerdings handelt es sich um einen rollierenden Forecast für den selben Wochentag der Folgewoche, dies muss genauer untersucht werden (Abhängigkeit der Instanzen muss analysiert werden)

Ungewissheiten/Irrationalitäten:

Der Graph bildet die Entwicklung des akkumulierten Core Weight über alle „in scope“-Sektoren ab.

Das Jahr 2011 umfasst 365 Instanzen (Kalendertage). Die Range der Werte des Core Weight erstreckt sich dabei von 117446,5 bis 771505,5 und der Mittelwert beträgt 487853,088. Die maximalen Abweichungen vom Mittelwert belaufen sich auf -370406,588 downwards, sowie 283652,412 (weniger symmetrisch als in 2010, Differenz Max-Min: 86754,176). Im Jahresverlauf ist ein Aufwärtstrend zu erkennen (siehe Regressionsgerade, grün eingezeichnet). Dieser ist wesentlich flacher als in 2010. Der Bestimmtheitsgrad ist noch geringer als in 2010 (Differenz: -0,1235).

Die Daten verlaufen (abgesehen vom Wachstumstrend) in einem wiederkehrenden Muster und lassen sich in mehrere Teilabschnitte unterteilen. Diese Abschnitte beginnen mit einem Tief, nahe 0 und enden mit einem Peak, auf welchen ein direkter Absturz erfolgt.

Hypothesen:

* Teilabschnitte bilden Wochenverläufe ab
* Core Weight einer Instanz ist abhängig von der Anzahl der geflogenen Sektoren
* Der geringe Bestimmtheitsgrad spiegelt die extremen Schwankungen der Einzelwerte ab, die Verlässlichkeit einer naiven Prognose für die folgende Instanz ist gering(er als in 2010)
* Anstieg am Ende des Betrachtungszeitraumes ist bedingt durch das Weihnachtsgeschäft

Kritische Bereiche:

* Relevante Bank Holidays als spezieller Untersuchungsgegenstand (könnten durch farbliche Hervorhebung erkenntlich gemacht werden)
* Relevante Bank Holidays bzw. deren Einfluss auf das Core Weight sollten gesondert untersucht werden
* Der Bestimmtheitsgrad lässt keine zuverlässige naive Prognose für Instanz n+1 zu, allerdings handelt es sich um einen rollierenden Forecast für den selben Wochentag der Folgewoche, dies muss genauer untersucht werden (Abhängigkeit der Instanzen muss analysiert werden)

Ungewissheiten/Irrationalitäten:

Der Graph bildet die Entwicklung des akkumulierten Core Weight über alle „in scope“-Sektoren ab.

Das Jahr 2011 umfasst 366 Instanzen (Kalendertage). Die Range der Werte des Core Weight erstreckt sich dabei von 51696 bis 917775 und der Mittelwert beträgt 537016,347. Die maximalen Abweichungen vom Mittelwert belaufen sich auf -485320,347 downwards, sowie 380758,653 (weniger symmetrisch als in 2010/2011, Differenz Max-Min: 104561,694). Im Jahresverlauf ist ein Aufwärtstrend zu erkennen (siehe Regressionsgerade, grün eingezeichnet). Dieser ist flacher als in 2010, aber steiler als in 2011. Der Bestimmtheitsgrad ist geringer als in 2010, aber höher 2011 (Differenz zu 2010: -0,0834; Differenz zu 2011: 0,0401).

Die Daten verlaufen (abgesehen vom Wachstumstrend) in einem wiederkehrenden Muster und lassen sich in mehrere Teilabschnitte unterteilen. Diese Abschnitte beginnen mit einem Tief, nahe 0 und enden mit einem Peak, auf welchen ein direkter Absturz erfolgt.

Hypothesen:

* Teilabschnitte bilden Wochenverläufe ab
* Core Weight einer Instanz ist abhängig von der Anzahl der geflogenen Sektoren
* Der geringe Bestimmtheitsgrad spiegelt die extremen Schwankungen der Einzelwerte ab, die Verlässlichkeit einer naiven Prognose für die folgende Instanz ist gering(er als in 2010)
* Anstieg am Ende des Betrachtungszeitraumes ist bedingt durch das Weihnachtsgeschäft
* Die unterschiedliche Anzahl an Instanzen pro Jahr ist bedingt durch ein Schaltjahr, eine Prognose auf Basis des Datumswertes (indizierte Instanz des Vorjahres) ist nicht möglich

Kritische Bereiche:

* Relevante Bank Holidays als spezieller Untersuchungsgegenstand (könnten durch farbliche Hervorhebung erkenntlich gemacht werden)
* Relevante Bank Holidays bzw. deren Einfluss auf das Core Weight sollten gesondert untersucht werden
* Der Bestimmtheitsgrad lässt keine zuverlässige naive Prognose für Instanz n+1 zu, allerdings handelt es sich um einen rollierenden Forecast für den selben Wochentag der Folgewoche, dies muss genauer untersucht werden

Ungewissheiten/Irrationalitäten: