# AUFGABENSTELLUNG
# TAKE HOME EXAMINATION
# MULTIVATIATE GEODATENANALYSE
# WS 24/25
# 1. PRÜFUNGSZEITRAUM

Ronny Schomacker

## Vorgaben

Der Bericht sollte die Zeichenzahl von 17.000 Zeichen inklusive Leerzeichen nicht überschreiten!

Der Bericht sollte selbständig und ohne Hilfe anderer von ihnen eigenhändig erzeugt werden! Sollten Bericht-, Codeelemente und Abbildungsgestaltung bei zwei und mehr Abgaben auffällige Ähnlichkeiten besitzen, die kaum durch Zufall oder die Skriptvorlagen erklärt werden können und damit auf eine nicht selbständige Arbeit schließen lassen, werden die betroffenen Berichte als nicht bestanden gewertet!

Layouten sie den Bericht und die Abbildungen und folgen sie der durch die Teilaufgaben vorgegebenen Struktur. Betten sie die Abbildungen (Diagramm) in das PDF-Dokument ein und vergeben sie Abbildungsnummern auf die sie sich im Text beziehen. Für ein ansprechendes Layout und die Gestaltung des Berichtes sowie eine übersichtliche R-Code Strukturierung und Kommentierung gibt es unabhängig von den unten gestellten Aufgaben bis zu 5 Punkte.

## Abgabe

Geben Sie einen vollständig lauffähigen, strukturierten und kommentierten R-Code aller Teilaufgaben in der Form **`Matrikelnummer.R`** (`alternativ auch Matrikelnummer.Rmd`) ab! Fügen sie das erstellten R-Objekte als **`Matrikelnummer.rds`** und Ihre aus INKAR heruntergeladene Originaldatei als **`Matrikelnummer.csv`** hinzu. Geben Sie die Lösung der Aufgaben in Form eines Berichtes ab: **`Matrikelnummer.pdf`** (`alternativ in der Rmd-Datei`)!

Hinweis: Überprüfen sie ihre Abgabe unbedingt auf Vollständigkeit! Lassen sie ihren R-Code auch mal auf einen anderen Rechner laufen! Spätere Abgaben können NICHT berücksichtigt werden.

Geben Sie eine (eingescannte) unterschriebene **Eigenständigkeitserklärung** für den Bericht ab.

Abgabedatum: siehe Moodle!

# Punkte:

- Erreichbar in der Examination: maximal 100 P.
  +25 möglicher Zusatzpunkte, die sich wie folgt zusammensetzen:
  - Zusatzaufgaben in der Examination: +10 P.
  - Zusatzpunkte durch Seminarübungen und Mitarbeit: max. +10 P.
  - Zusatzpunkte für besonders positive Auffälligkeiten in den Lösungen: max. +5 P.

- Bestanden (4,0) >= 50 P. 3,0 >=65 P. 2,0 >= 80 P. 1,0 >= 95 P.

# Aufgabe

Erstellen Sie eine Abfrage bei INKAR (www.inkar.de) mit den von Ihnen separat in Moodle ausgewählten Variablen:

Variable 1: _____

Variable 2: _____

1. Laden Sie die Daten auf Kreisebene als **CSV-Datei (`Matrikelnummer.csv`)** herunter und bearbeiten Sie diese in R so, dass sie als data.frame-Objekt vorliegen. Speichern Sie dieses Objekt als **`Matriknummer.rds`** ab! (R-Code: 5 P.)

2. Erstellen Sie von Ihren vorgegebenen Variablen **Histogramme und Boxplots in R**! Passen Sie die **Anzahl der Klassen (bins) im Histogramm** auf ihre Variablen an. Beschreiben Sie die Verteilung der Variablenn und Begründen Sie Ihre Klassenbildung. (R-Code: 5 P., Diagramme: 5 P., Beschreibung/Begründung: 5 P.)

3. Zusatzaufgabe: Entscheiden und Begründen sie aufgrund der Verteilungssymmetrie/-schiefe der Variablen in den Histogrammen, ob sie eine **Transformation** (Wurzel, Invertierung, Logarithmierung, …) der Daten durchführen, um deren Schiefe anzupassen. Transformieren Sie die Daten gegebenenfalls permanent und speichern sie diese als neue Variable ab! **Arbeiten sie mit den gegebenenfalls (transformierten) Variablen in den nachfolgenden Schritten weiter!!!** (R-Code: 2 P., Begründung: 2 P., weitere Nutzung: 2 P.)

4. Entfernen sie die **NA-Werte** und **Ausreißer** ihres Datensatzes, wie sie in R bei einem Boxplot (Regel nach Interquartilsabstand) gekennzeichnet werden und bilden sie ein **Subset** mit dem sie in den folgenden Schritten weiterarbeiten! (R-Code: 5 P., Subset: 5 P.)

5. Entscheiden sie sich durch Formulierung einer Hypothese für die Zuordnung der abhängigen und unabhängigen Variable. Erzeugen sie ein **Punktstreuungsdiagramm** der beiden Variablen des Subsets in R! Beschreiben sie die Beziehung der Variablen anhand der Muster (Anordnung, Cluster) im Diagramm (R-Code: 5 P., Beschreibung: 5 P.)

6. Diskutieren sie anhand des Erscheinungsbildes der Punktwolke im **Punktstreuungsdiagramm**, inwiefern eine lineare Regression sinnvoll bzw. kritisch sein könnte. Gehen Sie dabei auf die Bedingungen der Regression ein (Normalität, Homoskedastizität, Linearität und Unabhängigkeit). Sie können dazu auch den im Skript vorgestellten Boxplot-Ansatz umsetzen. Welche Konsequenzen werden die sichtbaren Verletzungen der Kriterien für die Über- und Unterschätzung der abhängigen Variable im Wertebereich der unabhängigen Variable haben? (Beschreibung: 10 P.)

7. **Führen sie die Regression unabhängig von ihrer (eventuell negativen) Bewertung in R durch** und interpretieren Sie **Güte und Aussage ihres Regressionsmodells** möglichst quantitativ mit Hilfe von Güteparametern des Regressionsmodells. In welcher Relation stehen die beiden Variablen in den Landkreisen von Deutschland generell zueinander? Geben Sie dafür auch die sich aus den ermittelten Parametern ergebende mathematische Regressionsgleichung an (ggf. mit Transformation). Wieviel Varianz von der abhängigen Variable kann die Regression erklären? (R-Code: 5 P., Beschreibung: 5 P.)

8. Zusatzaufgabe: Inwiefern kann die Relation der beiden Variablen durch eine **Transformation oder Polynombildung** optimiert werden? Begründen sie ihre Versuche anhand der Punktstreungswolke der beiden Variablen. Entscheiden sie mithilfe des Bestimmtheitsmaßes $R^2$, ob die Transformation zu einer Verbesserung beiträgt. Sie können dafür in R die Regressionsformel der Form `formula= I(transform(y))~I(transform(x))` nutzen. Geben sie die mathematische Gleichung der durch Transformation optimierten Regression an. (R-Code: 3 P., Erläuterung: 1 P.)

9. Verknüpfen Sie die Subset-Daten mit Geometriedaten der Kreise mithilfe von beispielsweise `merge()` in R. Bereiten sie die Daten gegebenenfalls so vor, dass dieser Schritt möglich ist. (R-Code: 5 P.)

10. Erstellen Sie eine **ansprechende Karte** für die beiden ursprünglichen Variablen in R, beispielsweise mit **`tmap`**! Achten Sie dabei insbesondere auf die Legende (Beschriftung, Klassen (Klassengrenzenrücktransformation zu ursprünglichen Einheiten)! Beschreiben Sie beobachtbare räumliche Muster (Nord/Süd, Ost/West, städtisch/ländlich, …) der einzelnen Variablen. (R-Code: 5 P., Beschreibung: 5 P., Kartenlayout: 5 P.).

11. Erstellen Sie eine **ansprechende Karte** für die **Residuen der Linearen Regression** in R! Was sagt ihnen das in Erscheinung tretende räumliche Muster der Residuen auf der Karte? Gibt es Beziehungen zu den Karten in Aufgabe 9? Welche Konsequenzen ergeben sich für die räumliche Güte des Linearen Modells? (R-Code: 5 P., Karte: 5 P., Beschreibung: 5 P.)

Viel Erfolg und gutes Gelingen!

Ronny Schomacker