Wir haben insgesamt **4 Datensätze**. Für jeden Datensatz führen wir **den Import und die Transformation** durch.   
- Bei dem ersten Datensatz handelt es sich um **eine Excel-Datei mit den allgemeinen Fakten von 7 Bezirken** wie Bevölkerungsdichte, Arbeitslosenquote usw. Wir teilen diesen Datensatz **in 4 kleinere Datensätze** auf, um die spätere Analyse zu vereinfachen.   
- Der zweite Datensatz ist ebenfalls eine Excel-Datei von statistik-nord.de. Von der Website erhalten wir **die gesamte Bodenfläche der einzelnen Bezirke sowie die Flächen für Siedlung, Verkehr, Vegetation und Gewässer.** Daraus haben wir die **Anteile** berechnet, um die Daten **vergleichbar** zu machen.   
- Der dritte Datensatz ist die **Aufteilung der Siedlungsfläche in wichtige Nutzungszwecke** wie Wohnen, Industrie- und Gewebeflächen, Erholungsflächen und Friedhof. Auch hier haben wir die **Anteile** berechnet.  
- Für den letzten Datensatz mussten wir eine **API-Anfrage** verwenden. Hier erhalten wir die **Anzahl der Händler verschiedener Art in 7 Bezirken**, wie z.B. Drogerien, Supermärkte oder Wochenmärkte usw. Der Datensatz wurde auch **nach Kategorien in 3 Datensätze** unterteilt.   
Das **Hauptproblem** bei der Datenintegration war die **Datenkompatibilität**, d. h. unterschiedliche Maßeinheiten, Namenskonventionen (Synonyme, Groß- und Kleinschreibung usw.) und leider auch unterschiedliche Erfassungszeiten. Technische Unterschiede wie Dateitypen, Schnittstellen.   
**Ein weiteres Problem** ist, dass die **Granularität** begrenzt war. Wir wollten anfangs eigentlich die Analyse **nach Stadtteilen** durchführen, aber für einige Datensätze gibt es **keine Daten** nach Stadtteilen.   
  
Nachdem wir die Daten geladen und transformiert haben, betrachten wir für jeden Datensatz **die statistische Zusammenfassung und die Histogramme der einzelnen Attribute. Hier würde ich nicht jedes einzelne Attribut durchgehen.** Die wichtigsten Erkenntnisse aus dieser explorativen Analyse sind: 1. wir haben keine **Missing Values** und 2. wir haben **Ausreißer** bei einigen Attributen, aber diese Ausreißer sind einfach Extremwerte und keine falsch erfassten Werte.  
  
Wir haben dann einige **Heatmaps** erstellt, um die Korrelationen zwischen den Attributen zu untersuchen. Unser Ziel war es, **die stark korrelierten Attribute zu droppen.**  
  
Dann haben wir **ein schönes Dashboard mit einer sorgfältigen Farbauswahl für die Datenvisualisierung** erstellt, in dem wir **die skalierten Werte für jede Kategorie sowie die absoluten Werte aus jedem Bezirk** anzeigen. Dann sind wir mit der explorativen Analyse fertig. **Das Fazit** aus dieser Phase war, dass unsere Daten zu viele Attribute haben. Es ist schwierig, diese darzustellen, um einen Überblick zu bekommen. Dies ist auch ein **Entscheidungsfaktor für die zu verwendende Analysemethode,** die jetzt von Sophie vorgestellt wird.