Find a dataset and Form the big idea:

Find a dataset

The datasets chosen are used car sales datasets from craigslist USA and eBay Germany. These datasets are both north of a third of a million rows and as such will be cut down. The cut down process was done using a python library called dask as opening the csv in R is unresponsive. Prior experience with dask lead me to believe this would be the correct tool to minimise time spent on this problem. Since dask is using panda's data frames and this aims to implement R functionality, I'm basically using R for this step and its outside the scope of the find a dataset instruction. Both datasets were shortened to 500 rows for the purposes of this assignment. Regarding the US dataset the URL data and description columns were dropped as these rows text data often went north of 15 characters.

The final datasets used will be these modified datasets and are included in the upload.

German:

https://www.kaggle.com/orgesleka/used-cars-database

As this data was taken from a German site many words are in German. In order to merge this data correctly this must be remedied. The location data is in postal code format this should be converted to human readable location format. In many cases, the headers are synonym of each other for example gearbox and transmission.

DateCrawled not in the other dataset - drop
Name not in the other dataset - drop
seller one value category – drop
Offertype one value category – drop
price - good
abtest – not in other dataset - drop
vehicleType – convert language and rename
yearOfRegistration rename to year
gearbox – rename
powerPS – drop as not in other dataset
Model – good
Kilometer - rename to odometer
monthOfRegistration drop as not in other dataset
fuelType – rename to type
brand – rename to manufacturer
notRepairedDamage like title status
dateCreated not in the other dataset - drop
nrOfPictures not in the other dataset - drop
postalCode – used to join zip code dataset then drop
LastSeen not in the other dataset - drop

Zip codes:

http://download.geonames.org/export/zip/DE.zip

country code: iso country code, 2 characters

postal code: varchar (20)

place name: varchar (180)

admin name1: 1. order subdivision (state) varchar (100)

admin code1: 1. order subdivision (state) varchar (20)

admin name2: 2. order subdivision (county/province) varchar (100)

admin code2: 2. order subdivision (county/province) varchar (20)

admin name3: 3. order subdivision (community) varchar (100)

admin code3: 3. order subdivision (community) varchar (20)

latitude: estimated latitude (wgs84)

longitude: estimated longitude (wgs84)

accuracy: accuracy of lat/lng from 1=estimated, 4=geonameid, 6=centroid of addresses or shape

The latitude, longitude and admin name will be kept from this dataset to merge on the postal code value. This will match with state, lat and long in the us dataset allowing the use of Location data if required.

USA:

https://www.kaggle.com/austinreese/craigslist-carstrucks-data

region fine
price fine
year fine
manufacturer – fine
model – fine
condition – change to binary value
cylinders not in the other dataset - drop
fuel rename to fuelType
odometer - good
title_status like notRepairedDamage
Transmission renames to gearbox
Vin not in the other dataset - drop
drive drop
size mostly empty – drop
type fine
paint_color drop
county drop as wrong data
state
lat
long

Form your Big Idea

The dataset has two distinct regions, Germany and America. Comparing the differences between these two may be interesting. Exploring the data lead to some interesting graphs but as both regions are together not much can be made in the way of conclusions.

Building on this idea, the year, fuel type, distance travelled, and price can be considered to prove what prompts a regional driver to sell their vehicle online.

A discovery into what makes US and German car drivers want to sell their beloved vehicles using online marketplace data!