Capstone Proposal

David Aragon

January 9, 2018

autos\_csv <- read.csv("~/Downloads/autos (1).csv")  
carPrice <- read.csv("~/Downloads/cnt\_km\_year\_powerPS\_minPrice\_maxPrice\_avgPrice\_sdPrice.csv")

## Capstone Proposal Overview

The used car market lately has been on the rise. With the economy benefiting from lower unemployment rates, the used car market has flourished. The car market has a whole has benefited as well with lease sales increasing. Because lease sales are on the rise, people can either turn in their vehicle or keep it. For those that have kept their vehicle, the used car market gains from consumer turn ins. Consumers that leased their vehicle are much more likely to take care of their vehicles to avert paying high wear and tear charges. As a result, the quality of used cars in a particular car lot is quite good.

## Problem

In my analysis, I want to be able to forecast and display the used car marke based on already used data. The problem I want to solve is eliminating cars that will not sell in a timely manner

## Who is my client and Why do they care

My client will be finance companies looking to forecast, purchase and/or sell vehicles in and out of their respective lots.

## Data

The data used in this analy/ ccvcv fsis will be scraped from Ebay-Kleinanzeigan. These data sets include vehicle information, seller, notRepaired Damage, #ofpictures, lastSeeenOnline, averagePrice, lowPrice, highPrice.

#### Autos.CSV data set

This dataset’s summary is the following:

## dateCrawled name   
## 2016-03-08 15:50:29: 5 Ford\_Fiesta : 336   
## 2016-03-20 16:50:22: 5 Volkswagen\_Golf\_1.4: 335   
## 2016-03-26 10:51:07: 5 BMW\_318i : 334   
## 2016-03-31 17:57:07: 5 Opel\_Corsa : 316   
## 2016-04-02 14:50:21: 5 BMW\_316i : 258   
## 2016-03-05 15:48:41: 4 BMW\_320i : 256   
## (Other) :189320 (Other) :187514   
## seller offerType price abtest   
## gewerblich: 2 Angebot:189341 Min. : 0 control:91131   
## privat :189347 Gesuch : 8 1st Qu.: 1150 test :98218   
## Median : 2950   
## Mean : 10895   
## 3rd Qu.: 7200   
## Max. :99999999   
##   
## vehicleType yearOfRegistration gearbox   
## limousine :48701 Min. :1000 : 10395   
## kleinwagen:40759 1st Qu.:1999 automatik: 39220   
## kombi :34498 Median :2003 manuell :139734   
## :19437 Mean :2005   
## bus :15532 3rd Qu.:2008   
## cabrio :11668 Max. :9999   
## (Other) :18754   
## powerPS model kilometer monthOfRegistration  
## Min. : 0.0 golf : 15286 Min. : 5000 Min. : 0.000   
## 1st Qu.: 70.0 andere : 13453 1st Qu.:125000 1st Qu.: 3.000   
## Median : 105.0 3er : 10528 Median :150000 Median : 6.000   
## Mean : 114.7 : 10398 Mean :125640 Mean : 5.733   
## 3rd Qu.: 150.0 polo : 6714 3rd Qu.:150000 3rd Qu.: 9.000   
## Max. :19208.0 corsa : 6415 Max. :150000 Max. :12.000   
## (Other):126555   
## fuelType brand notRepairedDamage  
## benzin :114106 volkswagen :40687 : 36542   
## diesel : 54968 bmw :20545 ja : 18410   
## : 16936 opel :20427 nein:134397   
## lpg : 2732 mercedes\_benz:17931   
## cng : 308 audi :16676   
## hybrid : 142 ford :13093   
## (Other): 157 (Other) :59990   
## dateCreated nrOfPictures postalCode   
## 2016-04-03 00:00:00: 7392 Min. :0 Min. : 1067   
## 2016-04-04 00:00:00: 7185 1st Qu.:0 1st Qu.:30559   
## 2016-03-20 00:00:00: 6860 Median :0 Median :49661   
## 2016-03-12 00:00:00: 6853 Mean :0 Mean :50892   
## 2016-03-21 00:00:00: 6843 3rd Qu.:0 3rd Qu.:71577   
## 2016-03-14 00:00:00: 6691 Max. :0 Max. :99998   
## (Other) :147525   
## lastSeen   
## 2016-04-07 00:46:04: 14   
## 2016-04-06 04:17:47: 12   
## 2016-04-06 05:44:34: 12   
## 2016-04-06 08:15:55: 12   
## 2016-04-06 09:46:11: 12   
## 2016-04-06 09:46:51: 12   
## (Other) :189275

This dataset’s structure is the following:

str(autos\_csv)

## 'data.frame': 189349 obs. of 20 variables:  
## $ dateCrawled : Factor w/ 164590 levels "2016-03-05 14:06:22",..: 96234 96080 44613 62122 135465 158918 142369 82743 161561 60299 ...  
## $ name : Factor w/ 128113 levels "\_\_\_\_\_AUDI\_A4\_S\_LINE\_\_\_\_\_\_VOLLAUSSTATUNG\_\_\_\_\_\_",..: 43627 2401 50197 44557 95036 15652 80956 118428 34979 119710 ...  
## $ seller : Factor w/ 2 levels "gewerblich","privat": 2 2 2 2 2 2 2 2 2 2 ...  
## $ offerType : Factor w/ 2 levels "Angebot","Gesuch": 1 1 1 1 1 1 1 1 1 1 ...  
## $ price : int 480 18300 9800 1500 3600 650 2200 0 14500 999 ...  
## $ abtest : Factor w/ 2 levels "control","test": 2 2 2 2 2 2 2 2 1 2 ...  
## $ vehicleType : Factor w/ 9 levels "","andere","bus",..: 1 5 9 6 6 8 4 8 3 6 ...  
## $ yearOfRegistration : int 1993 2011 2004 2001 2008 1995 2004 1980 2014 1998 ...  
## $ gearbox : Factor w/ 3 levels "","automatik",..: 3 3 2 3 3 3 3 3 3 3 ...  
## $ powerPS : int 0 190 163 75 69 102 109 50 125 101 ...  
## $ model : Factor w/ 251 levels "","1\_reihe","100",..: 119 1 120 119 104 13 9 42 58 119 ...  
## $ kilometer : int 150000 125000 125000 150000 90000 150000 150000 40000 30000 150000 ...  
## $ monthOfRegistration: int 0 5 8 6 7 10 8 7 8 0 ...  
## $ fuelType : Factor w/ 8 levels "","andere","benzin",..: 3 5 5 3 5 3 3 3 3 1 ...  
## $ brand : Factor w/ 40 levels "alfa\_romeo","audi",..: 39 2 15 39 32 3 26 39 11 39 ...  
## $ notRepairedDamage : Factor w/ 3 levels "","ja","nein": 1 2 1 3 3 2 3 3 1 1 ...  
## $ dateCreated : Factor w/ 97 levels "2014-03-10 00:00:00",..: 83 83 73 76 90 94 91 80 94 76 ...  
## $ nrOfPictures : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ postalCode : int 70435 66954 90480 91074 60437 33775 67112 19348 94505 27472 ...  
## $ lastSeen : Factor w/ 111190 levels "2016-03-05 14:15:08",..: 107390 106918 92424 25353 100813 104198 94797 50176 88971 71117 ...

In review of this dataset, some of the fields are in german. I will have to translate those field into English for my clients. The rest of the data looks to be in fair order and will be an easy import for further analysis

#### Car Prices data sets

This dataset’s summary is the following:

## count km year powerPS   
## Min. : 11.0 Min. : 5000 Min. :1991 Min. : 40   
## 1st Qu.: 20.0 1st Qu.: 60000 1st Qu.:2001 1st Qu.: 80   
## Median : 47.0 Median : 90000 Median :2007 Median :140   
## Mean : 161.0 Mean : 88653 Mean :2005 Mean :152   
## 3rd Qu.: 123.8 3rd Qu.:125000 3rd Qu.:2011 3rd Qu.:200   
## Max. :3623.0 Max. :150000 Max. :2015 Max. :520   
## minPrice maxPrice avgPrice sdPrice   
## Min. : 101 Min. : 950 Min. : 412.2 Min. : 193.4   
## 1st Qu.: 350 1st Qu.: 8900 1st Qu.: 3244.9 1st Qu.: 1474.9   
## Median : 1692 Median :17000 Median : 7698.0 Median : 2682.6   
## Mean : 4162 Mean :21440 Mean :10464.5 Mean : 3432.7   
## 3rd Qu.: 6000 3rd Qu.:29800 3rd Qu.:15410.0 3rd Qu.: 4468.5   
## Max. :36675 Max. :99999 Max. :66936.0 Max. :22338.8

This dataset’s structure is the following:

## 'data.frame': 1770 obs. of 8 variables:  
## $ count : int 19 69 11 26 42 233 13 15 19 234 ...  
## $ km : int 125000 150000 90000 100000 125000 150000 90000 100000 125000 150000 ...  
## $ year : int 1991 1991 1991 1991 1991 1991 1991 1991 1991 1991 ...  
## $ powerPS : int 40 40 60 60 60 60 80 80 80 80 ...  
## $ minPrice: int 300 110 330 200 150 110 150 150 150 120 ...  
## $ maxPrice: int 1499 1600 10000 8750 4600 9000 4699 3990 3900 12000 ...  
## $ avgPrice: num 648 516 2342 1392 1141 ...  
## $ sdPrice : num 352 347 3017 1599 1010 ...

In this dataset, it has 8 variables : count, year, km, powerPS, minPrice, maxPrice, avgPrice, and sdPrice. 6 of these variable are integer types and 2 are numeric types. This dataset is fairly clean so not much clean up will be involved in data analysis.

## Approach

My approach to analyzing this data will be to clean up the data significantly. The data is in german so I will need to translate as my client will be mostly english speakers. Once data is cleaned, I will join the datasets are prices and vehicle info are two different datasets. Using R language and RStudio, I will display which vehicle are the best selling and also display a table of value loss of an average used car per year.

## Deliverables

I plan on using R language to write the necessary code to display line graphs to show the progression or regression of certain vehicle in a particular lot. I plan on converting this visualization as a PDF file to share with client