vehicle-colour

Michael O'Rourke

April 27, 2018

## Does colour effect the value of a used car in UK auctions?

In simpler times, there were 7 colours. 7 colours that could be remembered using a simple mnemonic. These days there are nearly 2,000 pantone colours and laboratory tests estimate that humans can see approximately 10,000,000 different colours.

Colour can influence our mood, the taste of our food, even our perception of attractiveness. And if it can influence our most primitive emotions, it should be no surprise that it can influence our buying decisions too. Marketing agencies realised this a long time ago, which is why colour psychology plays such an important part in their campaigns.

The automotive industry is not short of marketeers (or designers). Their influence can be seen in the vast array of colours available to consumers. There are literally thousands of colours available for the exterior of a car. Many with names that could be given to a pop stars first born ('Lunar sky brown' and 'Pepper dust' are a couple of my current favourites).

Clearly, someone within the industry thinks that colour is important. But with so much variety, and so many contrasting tastes, does colour make a difference to the value of a vehicle?

Internally, our valuation experts often tell us about the importance of colour; "white cars will almost always be worth more" or "avoid doom blue coloured cars". There will always be cases where colour does have an influence, just as there will always be cases where it doesn't. What we are interested in, is whether colour has a systematic, market level effect on the value of a vehicle. All else being equal, does the colour of the vehicle influence the selling price?

### Main drivers of price

There is one thing we can be sure about, colour is not the main driver of a vehicles price. The most important factor is the type of vehicle. No one would expect the difference in price between a BMW i5 and a Nissan Micra to be the colour. Even if the BMW was in opaque couché. \_<https://www.digitaltrends.com/photography/pantone-448-c-ugliest-color/_> This is followed by the condition of the vehicle. A brand new, low mileage vehicle will almost always be worth more than an old, high mileage vehicle.

Then come the market influences. Supply and demand for a particular vehicle can change over time. The hot, new hatchback can soon become garden variety through no fault of its own (here's looking at you Fiat 500). However, it is not only fashions and tastes that drive demand, market seasonality also plays its part. There is an expectation within the industry that certain vehicle types are worth more at different times of the year. For example, convertibles are thought to sell better before summer than before winter so are worth more. Likewise, 4X4s are anticipated to sell better as weather conditions worsen in late autumn.

To able to attempt to answer our question, these factors will need to be mitigated and controlled.

### Our data

The data used for this analysis is wholesale sales data from Manheim Auctions between the beginning of 2016 and the end of 2017. While this dataset is smaller than website listing data over the same period, it has a couple of advantages. The first is that it is actual transactional data. We do not need to infer a sale or be forced to assume that the selling price is the same the listing price.

The second advantage is that there is no fixed price for a vehicle before the bidding process. Buyers base their bids based on their perception of the worth of the vehicle. This gives us a dataset that, once vehicle and market conditions have been controlled for, gives us artefacts about a vehicles relative worth. If colour is an influencer of vehicle price, we should be able to see evidence to support it.

### Controlling for main price drivers

#### Difference in vehicle specification

The industry standard vehicle identifier in the market is the CAP ID/CAP Code. This identifies a vehicle to a high degree of granularity. Vehicles with the same CAP ID, have the same manufacturer, range, model, derivative, engine size, fuel type, drive train, transmission, no of doors, and many more technical specifications. They also have relatively short production life-cycles, usually between 3 to 5 years. This makes it ideal for controlling for differences in vehicle specification. There are limitations to it. Although it categorises vehicles with standard equipment (such as leather seats or navigation systems), it does not control for any optional extras a consumer might have added. These extras aren't something we can control for within our dataset.

*Only vehicles with the same CAP ID will be compared.*

### Vehicle conditions

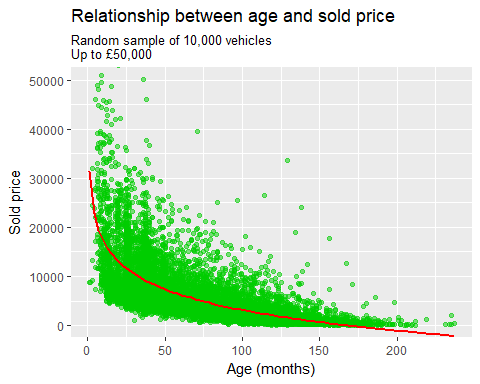
#### Age of a vehicle

As a vehicle gets older, it's value decreases. The rate of depreciation (influence of age on price) also changes as a vehicle gets older.

The graph below shows the relationship between age and wholesale price. We've used a random sample of the data to make it easier to see the relationship. Age, measured in months between the registration date and sold date, is along the x-axis (young vehicles towards the left, older vehicles to the right). Sold price, measured in £, is along the y-axis, lower values at the bottom, higher prices at the top. Each dot is a wholesale vehicle sale.

The red line highlights the overall relationship. There is a steep fall in price due to age in the first few months, which softens and stabilises, before gradually levelling off. This means that vehicles lose most of their value in the first couple of years, followed by a period of steady reduction in price (2-12 years), until there isn't much value left in the vehicle and the price levels off close to 0. It is easy to see why controlling for the age of a vehicle is important.

# Create a scatterplot showing age and wholesale price relationship  
# Use a sample of the data so that the curve is easier to see  
# Limited to £50k for visibilities sake  
  
ggplot(sample\_n(data\_subset, 10000), aes(age\_months, sold\_price)) + geom\_point(colour = 43, alpha = 0.5) + stat\_smooth(method = "lm", formula = y ~ log(x), se = FALSE, colour = 'red') + labs(x = "Age (months)", y = "Sold price") + coord\_cartesian(ylim = c(0,50000)) + ggtitle("Relationship between age and sold price", "Random sample of 10,000 vehicles\nUp to £50,000")



We can control for age in two ways, compare vehicles with exactly the same age or create small cohorts of vehicles with similar ages (e.g. group vehicles with ages between 3 and 6 months together). There are pros and cons to both. Using the exact age of a vehicle will allow us to completely control for age but will leave us with few vehicles to compare, especially when we take into consideration other influencers on price (mileage, vehicle condition, etc). Creating small groups would only allow us to control for some of the variance in price due to age but create groups with more vehicles to compare.

For our purposes, small age cohorts will be created. The size of the groups is dependent on the age of the vehicle. Depreciation is at its most dramatic during the first year of a vehicles life-cycle, so these age cohorts will be split into 3-month sizes. Between ages 1 to 4 years, age cohorts will be split into 6-month sizes, 4 to 15 will be 1-year sized groups and 2-year sized groups for anything older.

*Only vehicles with the same age band will be compared.*

# Create the age cohorts  
# Create a sequence of cutoffs. Use unique so that the beginning/end of each sequence is not duplicated  
age\_cutoffs <- unique(c(seq(3, 12, by = 3), seq(12, 48, by =6),   
 seq(48, 144, by = 12), seq(144, 264, by = 24)))  
  
# Create labels for these cutoffs  
age\_labels <- as.character(age\_cutoffs[1:22]) # leave of the last cutoff as it is out of range of the data  
  
# Use the cut function to create a new column for the age cohorts  
data\_subset\_with\_age\_cohorts <- data\_subset %>%  
 mutate(age\_cohort = cut(age\_months, breaks = age\_cutoffs, labels = age\_labels))

#### Mileage of a vehicle

The relationship between mileage and sold price is similar to the relationship between age and sold price. As mileage increases, the price decreases. The rate of depreciation also shares the same shape as age; a steep initial drop in price for low mileage vehicles, then a steady, consistent rate of depreciation, before a levelling off towards the higher mileages.

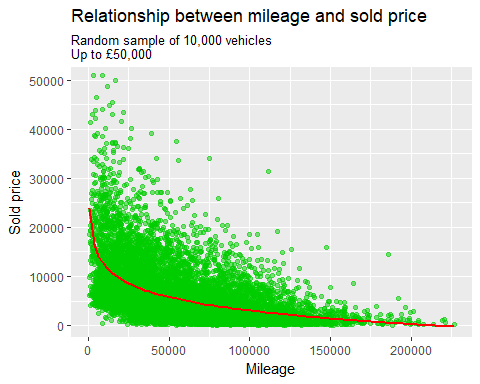
The graph below illustrates this relationship. Mileage is measured on the x-axis, low mileages on the left, high mileages to the right. Price is measured on the y-axis, low prices at the bottom, high prices at the top.

As with age, completely controlling for the mileage by exactly matching vehicles on mileage will result in too few vehicles to compare. Instead we'll group vehicles together based on their relative similarity in mileage. The depreciation rate is most abrupt for lower mileages, so the mileage cohorts will be split into 1,000-mile segments. 10,000 to 30,000 will be split by 5,000-mile segments, 30,000 to 100,000 by 10,000-mile segments and anything above 100,000 miles into cohorts of 20,000-mile segments.

This should allow us to balance specificity with sample size.

*Only vehicles with the same mileage band will be compared.*

# Create a scatterplot showing age and wholesale price relationship  
# Use a sample of the data so that the curve is easier to see  
# Limited to £50k for visibilities sake  
  
ggplot(sample\_n(data\_subset, 10000), aes(mileage, sold\_price)) + geom\_point(colour = 43, alpha = 0.5) + stat\_smooth(method = "lm", formula = y ~ log(x), se = FALSE, colour = 'red') + labs(x = "Mileage", y = "Sold price") + coord\_cartesian(ylim = c(0,50000)) + ggtitle("Relationship between mileage and sold price", "Random sample of 10,000 vehicles\nUp to £50,000")



mileage\_cutoffs <- unique(c(seq(1000, 10000, by = 1000), seq(10000, 30000, by = 5000),  
 seq(30000, 100000, by = 10000), seq(100000, 260000, by = 20000)))  
  
mileage\_labels <- as.character(mileage\_cutoffs[1:28])  
  
  
# Create a column which month each sale occured  
# This will be used to mitigate the effect of seasonality  
  
# Create function to convert month numbers to equivalent string  
  
convert\_months <- function(x){  
 ifelse(x == 1, "Jan",  
 ifelse(x == 2, "Feb",  
 ifelse(x == 3, "Mar",  
 ifelse(x == 4, "Apr",  
 ifelse(x == 5, "May",  
 ifelse(x == 6, "Jun",  
 ifelse(x == 7, "Jul",  
 ifelse(x == 8, "Aug",  
 ifelse(x == 9, "Sep",  
 ifelse(x == 10, "Oct",  
 ifelse(x == 11, "Nov",  
 ifelse(x == 12, "Dec", NA))))))))))))  
}  
  
data\_subset\_cohorts <- data\_subset\_with\_age\_cohorts %>%  
 mutate(mileage\_cohort = cut(mileage, breaks = mileage\_cutoffs, labels = mileage\_labels),  
 month = convert\_months(month(date\_sold)))

#### Condition grade

Age and mileage can be good measures of the usage of the vehicle, but they don't directly describe the physical condition of a vehicle. A well looked after, 2-year old, 20,000-mile company car is likely to be in better condition than a 2-year old, 20,000-mile family car which has been subjected to three toddlers in the back.

Within the wholesale industry, the physical condition of a vehicle is given a grade. Within Manheim auctions, these scored range from one to five. Well-conditioned vehicles are given a grade of one, poorly conditioned vehicles are given a grade of five. Vehicles that do not meet the criteria to reach grade five are 'unclassified'. The price penalty between grade one and five can be significant.

*Only vehicles with the same condition grade will be compared.*

In part 2: *How do we control for market levels effects such as inflation and seasonality?* What effect does all of this have on the sample sizes of our comparable vehicles? \*How do we translate the near 2,000 colour types into something more usable?

### Market conditions

#### Adjusting for inflation

Our data spans 24 months, over this time inflation has increased

The UK Office of National Statistics (ONS) publishes monthly Consumer Price Index (CPI) figures for each major consumer sector. The closest CPI sector available for wholesale vehicles is the used car retail CPI figures. As the retail market drives the wholesale market, assuming that the wholesale CPI figures are similar to the retail CPI figures does not seem unreasonable. All sold prices will be adjusted to the December 2017 levels (the last month in the dataset). The ONS data can be downloaded here: *link to ONS CPI scores*

#CPI data  
#Loaded with 1,391 columns. All blank after column 9. Keep only columns I'm interested in (used car CPI)  
used\_car\_cpi\_scores <- used\_car\_cpi\_scores %>% select(1,3)  
#Rename columns  
colnames(used\_car\_cpi\_scores) <- c("title", "used\_car\_cpi\_score")  
  
#CPI dataset contains CPI scores at a annual, quarterly and monthly level. Use monthly  
#Example format: 2016 JUL  
  
# Create a column with the CPI data format  
data\_subset\_cohorts\_temp <- data\_subset\_cohorts %>%  
 mutate(cpi\_date\_ref = convert\_date\_to\_cpi\_date(date\_sold))  
  
# Stop if any of the references are null values  
stopifnot(anyNA(data\_subset\_cohorts\_temp$date\_sold)== FALSE)  
  
# Join the CPI score data to the vehicle data using the CPI date reference  
joined\_data\_subset <- data\_subset\_cohorts\_temp %>%  
 left\_join(used\_car\_cpi\_scores, by = c("cpi\_date\_ref" = "title")) %>%   
 mutate(used\_car\_cpi\_score = as.numeric(used\_car\_cpi\_score)) # convert CPI values to numeric  
  
# Set December 2017 as the reference point  
# All prices will be adjusted to December 2017  
inflation\_reference\_value <- as.numeric(used\_car\_cpi\_scores %>% filter(title == '2016 DEC') %>% select(used\_car\_cpi\_score))  
  
# Adjust prices  
adjusted\_price\_dataset <- joined\_data\_subset %>%  
 mutate(adjusted\_sold\_price = round((sold\_price/used\_car\_cpi\_score) \* inflation\_reference\_value))  
  
# Histogram showing the effect of the adjustment  
# ggplot(adjusted\_price\_dataset %>%   
# mutate(adjustment\_percentage = ((adjusted\_sold\_price - sold\_price)/ sold\_price)), # create inflation % change  
# aes(adjustment\_percentage)) + geom\_histogram(fill = 43, colour = 'grey', alpha = 0.5, binwidth = 0.01) +  
# ggtitle("Distribution of inflation adjustments")

### Seasonality

Within the automotive industry, there is an assumption that certain types of vehicles have a higher demand (and therefore higher price) at different times of the year. For example, convertible vehicles are expected to be in higher demand before the summer than after. Likewise, 4x4 vehicles are expected to be in higher demand in the run up to winter.

*Something about plate change?*

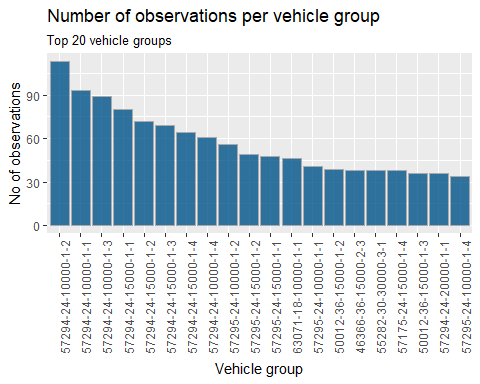
To take this into effect, sales will be split into cohorts based on the time of the year of their sale. If our dataset was larger, the cohorts could be as granular as months. Unfortunately, when you take into consideration the vehicle type, age, mileage and condition cohorts the number of vehicles available for comparison is likely to be small. For that reason, the cohorts will be at quarter level.

seasonality\_dataset <- adjusted\_price\_dataset %>%  
 mutate(seasonality\_cohort = quarter(date\_sold))

To summarise, only vehicles with the same cap ID, age cohort, mileage cohort, condition grade and seasonality cohort will be compared. In addition, all prices have been adjusted for inflation to December 2017 prices.

That's a lot of restrictions, will there be any groups of vehicles with large enough sample sizes to compare against each other?

# Create groups with the same age, mileage, condition, cap\_id, and seasonality  
vehicle\_cohort\_dataset <- seasonality\_dataset %>%  
 mutate(vehicle\_group = paste(cap\_id, age\_cohort, mileage\_cohort, condition\_grade, seasonality\_cohort, sep = '-'))  
  
# count how many observations per vehicle group  
observations\_per\_vehicle\_group <- vehicle\_cohort\_dataset %>%  
 group\_by(vehicle\_group) %>%  
 summarise(no\_of\_observations = n()) %>%  
 arrange(-no\_of\_observations)   
  
ggplot(observations\_per\_vehicle\_group %>% top\_n(20, no\_of\_observations),   
 aes(reorder(vehicle\_group, -no\_of\_observations), no\_of\_observations)) +  
 geom\_col(fill = "#00528A", colour = "#a8aaac", alpha = 0.8) + labs(x = "Vehicle group", y = "No of observations") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +   
 ggtitle("Number of observations per vehicle group", "Top 20 vehicle groups")



As expected, the sample size of each group is quite small. The largest group has only 113 observations, the 10th has 49 and the 20th has 34. As much as we would like these groups to be larger, the importance of specificity far outweighs the need for large sample sizes.

#### What should be our minimum sample size?

If our cut off is too high, then there are two problems. The most obvious is that it gives us too few groups to compare colours in. The second, is to do with the variety of vehicles. The eagle eyed amongst us would have noticed that the CAP IDs of the top 20 groups are very similar. In fact, 14 of the top 20 groups are Fiesta's, and 17 are Fords. If we set the cut off too high, we will only be able to compare how colour affects young, low mileage, Ford Fiestas.

If our cut off it too low, then we end up with thousands of groups with a couple of vehicles to compare which is unlikely to tell us anything more than random chance would. Approximately, 99% of all the vehicle groups have 5 or less observations.

Ten observations feels too low and twenty observations reduces the breadth of vehicle models we can compare. A reasonable compromise seems *why is this reasonable* to be 15 observations per vehicle group. This gives us 232 vehicle groups to compare.

### Colours

The simple days of only having to remember the seven colours of the rainbow have long done. In a world where there are nearly 2,000 pantone colours available and even black has different shades. It should be no surprise that there are 1,614 'different' colours of vehicles within our dataset. When designers and marketing teams combine the possibilities are endless. A couple favourites are 'Lunar sky brown' and 'Pepper dust'.

For this analysis, we'll be going back to basics with our colour descriptions. 'Moroccan red' and 'Tango red metallic' will be both classified as 'red', 'Sea grey' and 'Storm grey' are both 'grey'. This will result in some dropped observations, what colour of the rainbow is 'fashionista' or 'vision'.