Car resale Analysis - R Notebook

Prepare

This data has been taken from kaggle and thus it is from a trusted source.The data has thousands of rows and many parameters to analyse so it is a good enough sample size. The data has records from 2002 upto 2023 so it has current data too.Thus the data looks reliable for analysis.But the data is only based on certain cities so that is a limitation in the dataset.

The important packages are loaded. The csv file is loaded into Rstudio using read\_csv function.

library(readr)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ purrr 1.0.1  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(skimr)  
library(ggplot2)  
resale <- read.csv("car\_resale.csv")

Process

First we check for duplicates and also remove the ‘NA’ values

resale <- resale %>% distinct()   
resale <- resale %>% na.omit()

Checking the unique values for all columns to see if there are any problematic values

unique(resale$engine\_capacity)

## [1] "1197 cc" "2179 cc" "1086 cc" "1047 cc" "1196 cc" "1198 cc" "1462 cc"  
## [8] "1493 cc" "1396 cc" "998 cc" "1496 cc" "1364 cc" "1498 cc" "2360 cc"  
## [15] "1248 cc" "796 cc" "1199 cc" "2494 cc" "1186 cc" "1968 cc" "999 cc"   
## [22] "1399 cc" "1956 cc" "814 cc" "1461 cc" "1061 cc" "995 cc" "1499 cc"  
## [29] "799 cc" "1120 cc" "1586 cc" "936 cc" "1353 cc" "1451 cc" "1497 cc"  
## [36] "1582 cc" "1368 cc" "2354 cc" "2184 cc" "1995 cc" "2499 cc" "1395 cc"  
## [43] "1998 cc" "1798 cc" "2143 cc" "2967 cc" "1796 cc" "2894 cc" "1984 cc"  
## [50] "1298 cc" "1799 cc" "1591 cc" "624 cc" "1598 cc" "1172 cc" "2755

Output truncated…

unique(resale$insurance)

## [1] "Third Party insurance" "Comprehensive" "Zero Dep"   
## [4] "Third Party" "Not Available" "2"   
## [7] "1"

unique(resale$insurance)

## [1] "Third Party insurance" "Comprehensive" "Zero Dep"   
## [4] "Third Party" "Not Available" "2"   
## [7] "1"

unique(resale$transmission\_type)

## [1] "Manual" "Automatic"

unique(resale$kms\_driven)

## [1] "40,000 Kms" "70,000 Kms" "80,000 Kms" "1,20,000 Kms"   
## [5] "60,000 Kms" "20,000 Kms" "50,000 Kms" "1,50,000 Kms"   
## [9] "10,000 Kms" "30,000 Kms" "1,00,000 Kms" "90,000 Kms"   
## [13] "1,10,000 Kms" "45,000 Kms" "38,000 Kms" "49,000 Kms"   
## [17] "46,000 Kms" "29,000 Kms" "27,000 Kms" "72,100 Kms"   
## [21] "46,300 Kms" "26,300 Kms" "5,600 Kms" "7,000 Kms"   
## [25] "48,500 Kms" "43,800 Kms" "2,300 Kms" "24,800 Kms"   
## [29] "36,700 Kms" "79,000 Kms" "55,400 Kms" "23,700 Kms"   
## [33] "15,600 Kms" "29,900 Kms" "43,500 Kms" "44,000 Kms"   
## [37] "85,000 Kms" "76,469 Kms" "1,18,976 Kms" "57,689 Kms"   
Output truncated…

unique(resale$owner\_type)

## [1] "First Owner" "Second Owner" "Third Owner" "Fifth Owner" "Fourth Owner"

unique(resale$fuel\_type)

## [1] "Petrol" "Diesel" "CNG" "LPG" "Electric"

unique(resale$max\_power)

## [1] "83.1bhp" "153.86bhp" "83.14bhp"   
## [4] "68.05bhp" "81.86bhp" "69bhp"   
## [7] "73bhp" "86.7bhp" "103.25bhp"   
## [10] "98.6bhp" "89bhp" "67.1bhp"   
## [13] "58.16bhp" "88.7bhp" "87.2bhp"   
## [16] "118.36bhp" "89.84bhp" "170PS"   
## [19] "74bhp" "81.80bhp" "47.3bhp"   
Output truncated…

unique(resale$seats)

## [1] 5 7 8 6 4 9 2 10 14

unique(resale$mileage)

## [1] "21.4 kmpl" "17.6 kmpl" "20.85 kmpl" "19.81 kmpl" "17.19 kmpl"   
## [6] "27.28 kmpl" "15.37 kmpl" "18 kmpl" "20.14 kmpl" "21.56 kmpl"   
## [11] "23.7 kmpl" "23.95 kmpl" "26.6 kmpl" "21.43 kmpl" "18.2 kmpl"   
## [16] "20.65 kmpl" "22.7 kmpl" "11.3 kmpl" "26.59 kmpl" "20.92 kmpl"   
## [21] "21.01 kmpl" "22.74 kmpl" "20.51 kmpl" "18.9 kmpl" "22.9 kmpl"   
## [26] "15.1 kmpl" "11.57 kmpl" "23.84 kmpl" "24.3 kmpl" "25.47 kmpl"   
## [31] "25.24 kmpl" "22.1 kmpl" "22.05 kmpl" "11.4 kmpl" "12.8 kmpl"   
## [36] "18.8 kmpl" "24 kmpl" "21.94 km/kg" "16.55 kmpl" "21 kmpl"   
Output truncated…

unique(resale$body\_type)

## [1] "Hatchback" "MUV" "Sedan" "Minivans"   
## [5] "SUV" "Toyota" "Coupe" "Chevrolet"   
## [9] "Mercedes-Benz" "Audi" "Maruti" "Pickup"   
## [13] "Tata" "Mahindra" "Volvo" "Jaguar"   
## [17] "BMW" "Cars" "Datsun" "Hyundai"   
## [21] "Honda" "Convertibles" "Wagon" "Skoda"   
## [25] "Isuzu"

unique(resale$city)

## [1] "Agra" "Gurgaon" "Lucknow" "Delhi" "Chandigarh"  
## [6] "Bangalore" "Jaipur" "Kolkata" "Ahmedabad" "Chennai"   
## [11] "Pune" "Mumbai" "Hyderabad"

unique(resale$full\_name)

## [1] "2017 Maruti Baleno 1.2 Alpha"   
## [2] "2018 Tata Hexa XTA"   
## [3] "2015 Maruti Swift Dzire VXI"   
## [4] "2009 Hyundai i10 Magna 1.1"   
## [5] "2015 Hyundai i20 Active 1.2"   
## [6] "2017 Tata Tiago 1.05 Revotorq XZ"   
## [7] "2010 Hyundai i10 Magna 1.1"   
## [8] "2016 Maruti Eeco 7 Seater Standard BSIV"   
## [9] "2014 Honda Amaze E i-Vtech"   
## [10] "2018 Hyundai Xcent 1.2 VTVT SX"   
## [11] "2020 Maruti Ciaz Sigma BSIV"   
## [12] "2021 Hyundai Venue SX Opt Executive Diesel"   
Output truncated…

unique(resale$resale\_price)

## [1] "₹ 5.45 Lakh" "₹ 10 Lakh" "₹ 4.50 Lakh" "₹ 1.60 Lakh"   
## [5] "₹ 4.70 Lakh" "₹ 3.60 Lakh" "₹ 2 Lakh" "₹ 3.70 Lakh"   
## [9] "₹ 3.30 Lakh" "₹ 4.75 Lakh" "₹ 6.86 Lakh" "₹ 12.50 Lakh"  
## [13] "₹ 11 Lakh" "₹ 3.95 Lakh" "₹ 3.50 Lakh" "₹ 2.89 Lakh"   
Output truncated…

unique(resale$registered\_year)

## [1] "2017" "2018" "2015" "2009" "2010" "2016" "2014" "2020"   
## [9] "2021" "2019" "2011" "2012" "2013" "2022" "2004" "Dec-18"  
## [17] "Jun-18" "Mar-16" "Feb-18" "Oct-14" "Mar-14" "Feb-13" "Aug-18" "Jun-19"  
## [25] "Dec-21" "Apr-19" "May-18" "Nov-13" "Sep-22" "Jan-21" "Sep-20"

Output truncated…

The insurance column has some rows with ‘not available’,‘1’ and ‘2’ as insurance types which are not valid values and are thus filtered out from the dataset.

resale <- resale %>% filter(!(insurance == "Not Available")) %>% filter(!(insurance == "1")) %>% filter(!(insurance == "2"))   
resale$insurance[resale$insurance == 'Third Party insurance'] <- 'Third Party'  
resale1 <- resale

The body type column has some values that are car makers names rather than car body type and thus are removed from the dataset.

resale1 <- resale1 %>% filter(!(body\_type == "Audi")) %>% filter(!(body\_type == "BMW")) %>% filter(!(body\_type == "Chevrolet")) %>% filter(!(body\_type == "Datsun")) %>% filter(!(body\_type == "Honda")) %>% filter(!(body\_type == "Isuzu")) %>% filter(!(body\_type == "Jaguar")) %>% filter(!(body\_type == "Mahindra")) %>% filter(!(body\_type == "Maruti")) %>% filter(!(body\_type == "Mercedes-Benz")) %>% filter(!(body\_type == "Tata")) %>% filter(!(body\_type == "Toyota")) %>% filter(!(body\_type == "Volvo"))

The dataset has very few rows which do not have power in terms of bhp and thus are removed for simplicity so to have all values in terms of bhp.

resale2 <- resale1 %>% filter(grepl('bhp|Bhp',max\_power))

The kms\_driven column has Kms removed from the values so as to convert it to numeric in order to do analysis on it.

resale3 <- resale2   
resale3$kms\_driven <- gsub("Kms","",resale3$kms\_driven)  
resale3$kms\_driven <- as.numeric(gsub(",","",resale3$kms\_driven))

Only values are extracted from the max\_power column in order to convert the column to numeric

resale3$max\_power <- as.numeric(str\_extract(resale3$max\_power,"[\\d.d]+"))

The ‘cc’ is removed from the engine\_capacity column to also convert it to numeric

resale3$engine\_capacity <- as.numeric(gsub(" cc","",resale3$engine\_capacity))

There are some rows in mileage column which are in terms of km/kg so it has been converted to kmpl by multiplying the value with 1.7 and then the column is converted to numeric.

resale4 <- resale3  
resale4 <- resale4 %>% mutate(mileage = case\_when(str\_detect(mileage,"km/kg")~as.numeric(str\_extract(mileage,"[\\d.d]+"))\*1.7,str\_detect(mileage,"kmpl")~as.numeric(str\_extract(mileage,"[\\d.d]+"))))

The resale\_price has the rupee symbol removed and then the values with Lakh and Crore have the respective numbers multiplied.Finally the column values are all converted to numeric.

resale4$resale\_price <- substring(resale4$resale\_price,3)   
resale4 <- resale4 %>% mutate(resale\_price = case\_when(str\_detect(resale\_price,",")~as.numeric(gsub(",","",resale\_price)),str\_detect(resale\_price,"Lakh")~as.numeric(str\_extract(resale\_price,"[\\d.d]+")) \* 100000,str\_detect(resale\_price,"Crore")~as.numeric(str\_extract(resale\_price,"[\\d.d]+")) \* 10000000))

The registered year column has some rows which have year in a different format than the rest.So the last two digits of the year had to be extracted and formatted to be like the ‘YYYY’ format.Finally this column is also converted to numeric.

resale5 <- resale4  
resale5$registered\_year <- ifelse(str\_length(resale5$registered\_year)==6,substr(resale5$registered\_year,5,6),resale5$registered\_year)  
resale5$registered\_year <- ifelse(str\_length(resale5$registered\_year)==2,paste("20",resale5$registered\_year,sep=""),resale5$registered\_year)  
resale5$registered\_year <- as.numeric(resale5$registered\_year) # converted years to standard 4 digits and type is numeric

Some of the columns are renamed in order to reflect the units as well.

resale5 <- resale5 %>% rename("resale\_price\_rupees" = "resale\_price") %>% rename("mileage\_kmpl" = "mileage") %>% rename("max\_power\_bhp" = "max\_power")

The process stage is completed and we can see the dataset that is properly ready for analysis.

View(resale5)  
options(scipen = 999)

Analysis and Visualization

Lets take a look at the summary of the dataset using the summary and skim\_without\_charts function. We can see that the data is complete with no missing values and also no whitespaces.We can see the various statistical measures such as mean, sd, p0, p25, p50, p75, p100 for the variables.

summary(resale5)

## Id full\_name resale\_price\_rupees registered\_year  
## Min. : 0 Length:15530 Min. : 28000 Min. :2002   
## 1st Qu.: 4596 Class :character 1st Qu.: 400000 1st Qu.:2014   
## Median : 8936 Mode :character Median : 595000 Median :2017   
## Mean : 8887 Mean : 870297 Mean :2017   
## 3rd Qu.:13244 3rd Qu.: 900000 3rd Qu.:2019   
## Max. :17445 Max. :22500000 Max. :2023   
## engine\_capacity insurance transmission\_type kms\_driven   
## Min. : 72 Length:15530 Length:15530 Min. : 300   
## 1st Qu.:1197 Class :character Class :character 1st Qu.: 32000   
## Median :1199 Mode :character Mode :character Median : 53507   
## Mean :1412 Mean : 57568   
## 3rd Qu.:1498 3rd Qu.: 77000   
## Max. :5998 Max. :6275000   
## owner\_type fuel\_type max\_power\_bhp seats   
## Length:15530 Length:15530 Min. : 32.5 Min. : 2.000   
## Class :character Class :character 1st Qu.: 78.9 1st Qu.: 5.000   
## Mode :character Mode :character Median : 88.5 Median : 5.000   
## Mean :102.8 Mean : 5.203   
## 3rd Qu.:117.6 3rd Qu.: 5.000   
## Max. :558.0 Max. :10.000   
## mileage\_kmpl body\_type city   
## Min. : 7.81 Length:15530 Length:15530   
## 1st Qu.: 17.05 Class :character Class :character   
## Median : 19.07 Mode :character Mode :character   
## Mean : 19.71   
## 3rd Qu.: 21.70   
## Max. :140.00

skim\_without\_charts(resale5)

Data summary

|  |  |
| --- | --- |
| Name | resale5 |
| Number of rows | 15530 |
| Number of columns | 15 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 7 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| full\_name | 0 | 1 | 16 | 62 | 0 | 6031 | 0 |
| insurance | 0 | 1 | 8 | 13 | 0 | 3 | 0 |
| transmission\_type | 0 | 1 | 6 | 9 | 0 | 2 | 0 |
| owner\_type | 0 | 1 | 11 | 12 | 0 | 5 | 0 |
| fuel\_type | 0 | 1 | 3 | 8 | 0 | 5 | 0 |
| body\_type | 0 | 1 | 3 | 12 | 0 | 10 | 0 |
| city | 0 | 1 | 4 | 10 | 0 | 13 | 0 |

**Variable type: numeric**

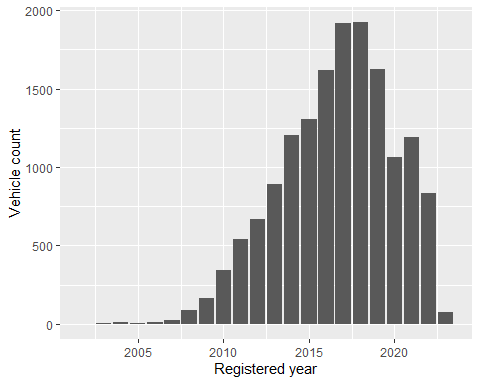
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1 | 8886.74 | 5009.08 | 0.00 | 4596.25 | 8936.50 | 13243.75 | 17445 |
| resale\_price\_rupees | 0 | 1 | 870296.73 | 1013787.55 | 28000.00 | 400000.00 | 595000.00 | 900000.00 | 22500000 |
| registered\_year | 0 | 1 | 2016.65 | 3.30 | 2002.00 | 2014.00 | 2017.00 | 2019.00 | 2023 |
| engine\_capacity | 0 | 1 | 1411.99 | 460.93 | 72.00 | 1197.00 | 1199.00 | 1498.00 | 5998 |
| kms\_driven | 0 | 1 | 57567.81 | 65740.89 | 300.00 | 32000.00 | 53507.00 | 77000.00 | 6275000 |
| max\_power\_bhp | 0 | 1 | 102.80 | 43.57 | 32.50 | 78.90 | 88.50 | 117.60 | 558 |
| seats | 0 | 1 | 5.20 | 0.65 | 2.00 | 5.00 | 5.00 | 5.00 | 10 |
| mileage\_kmpl | 0 | 1 | 19.71 | 4.98 | 7.81 | 17.05 | 19.06 | 21.70 | 140 |

The vehicle count spread across the years shows a consistent increase and a peak around 2017-2018 and then a drop around 2020.

resale5 %>% group\_by(registered\_year) %>% summarise(vehicle\_count = n())

## # A tibble: 22 × 2  
## registered\_year vehicle\_count  
## <dbl> <int>  
## 1 2002 1  
## 2 2003 5  
## 3 2004 12  
## 4 2005 8  
## 5 2006 9  
## 6 2007 25  
## 7 2008 87  
## 8 2009 168  
## 9 2010 345  
## 10 2011 544  
## # ℹ 12 more rows

ggplot(data = resale5) + geom\_bar(mapping = aes(x = registered\_year)) + labs(x = 'Registered year',y = 'Vehicle count')

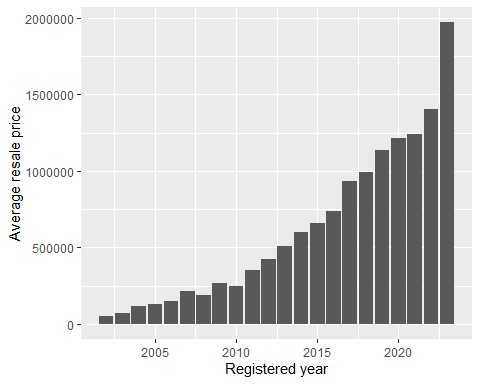


Looking at the average resale price, the average resale price has seen a consistent increase over the years

resale5 %>% group\_by(registered\_year) %>% summarise(avgrupees = mean(resale\_price\_rupees))

## # A tibble: 22 × 2  
## registered\_year avgrupees  
## <dbl> <dbl>  
## 1 2002 50000   
## 2 2003 73000   
## 3 2004 120000   
## 4 2005 131250   
## 5 2006 151667.  
## 6 2007 215920   
## 7 2008 186569.  
## 8 2009 266852.  
## 9 2010 250304.  
## 10 2011 353551.  
## # ℹ 12 more rows

resale5 %>% group\_by(registered\_year) %>% summarise(avgrupees = mean(resale\_price\_rupees)) %>% ggplot(aes(registered\_year,avgrupees)) + geom\_col() + labs(x = 'Registered year',y = 'Average resale price')



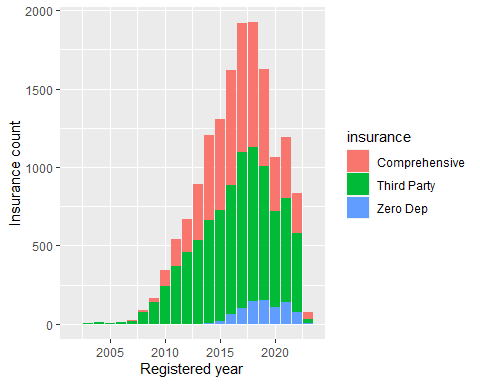
Looking at the spread of insurance types over the years, most cars have had third party insurance and comprehensive insurance has increased over the years. Zero dep only makes up a small portion of the cars involved and it is seen only 2013 onwards.

resale5 %>% group\_by(registered\_year,insurance) %>% summarise(insurecount = n())

## `summarise()` has grouped output by 'registered\_year'. You can override using  
## the `.groups` argument.

## # A tibble: 51 × 3  
## # Groups: registered\_year [22]  
## registered\_year insurance insurecount  
## <dbl> <chr> <int>  
## 1 2002 Third Party 1  
## 2 2003 Third Party 5  
## 3 2004 Third Party 12  
## 4 2005 Comprehensive 2  
## 5 2005 Third Party 6  
## 6 2006 Third Party 9  
## 7 2007 Comprehensive 5  
## 8 2007 Third Party 20  
## 9 2008 Comprehensive 14  
## 10 2008 Third Party 73  
## # ℹ 41 more rows

ggplot(data = resale5) + geom\_bar(mapping = aes(x = registered\_year,fill = insurance)) + labs(x = 'Registered year',y = 'Insurance count')



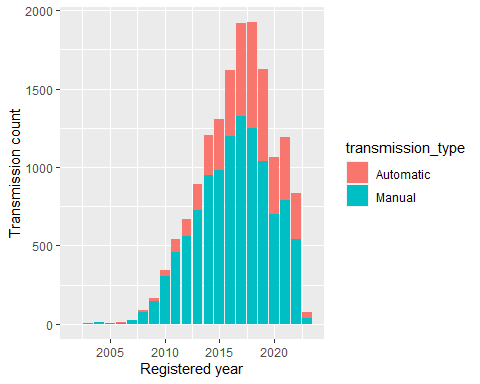
Looking at the vehicles transmission type across the years, manuals were mostly the standard initially, but the automatics have gained a lot of popularity in recent years.

resale5 %>% group\_by(registered\_year,transmission\_type) %>% summarise(transmiscount = n())

## `summarise()` has grouped output by 'registered\_year'. You can override using  
## the `.groups` argument.

## # A tibble: 41 × 3  
## # Groups: registered\_year [22]  
## registered\_year transmission\_type transmiscount  
## <dbl> <chr> <int>  
## 1 2002 Manual 1  
## 2 2003 Manual 5  
## 3 2004 Manual 12  
## 4 2005 Automatic 1  
## 5 2005 Manual 7  
## 6 2006 Automatic 3  
## 7 2006 Manual 6  
## 8 2007 Automatic 2  
## 9 2007 Manual 23  
## 10 2008 Automatic 10  
## # ℹ 31 more rows

ggplot(data = resale5) + geom\_bar(mapping = aes(x = registered\_year,fill = transmission\_type)) + labs(x = 'Registered year',y = 'Transmission count')



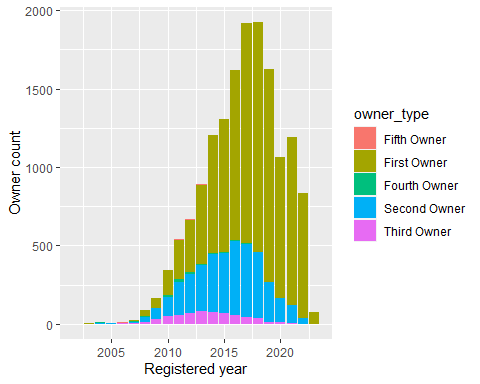
Initially most vehicles had first and second owners but then the first owners have increased sharply. The numbers for third,fourth and fifth owner remain low throughout.

resale5 %>% group\_by(registered\_year,owner\_type) %>% summarise(owncount = n())

## `summarise()` has grouped output by 'registered\_year'. You can override using  
## the `.groups` argument.

## # A tibble: 87 × 3  
## # Groups: registered\_year [22]  
## registered\_year owner\_type owncount  
## <dbl> <chr> <int>  
## 1 2002 Third Owner 1  
## 2 2003 First Owner 3  
## 3 2003 Second Owner 1  
## 4 2003 Third Owner 1  
## 5 2004 First Owner 3  
## 6 2004 Fourth Owner 3  
## 7 2004 Second Owner 4  
## 8 2004 Third Owner 2  
## 9 2005 First Owner 4  
## 10 2005 Fourth Owner 1  
## # ℹ 77 more rows

ggplot(data = resale5) + geom\_bar(mapping = aes(x = registered\_year,fill = owner\_type)) + labs(x = 'Registered year',y = 'Owner count')



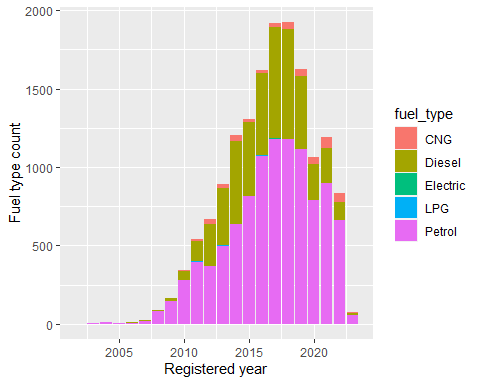
Initially most vehicles were of petrol type but then diesel type vehicles increased by a big margin but still remain behind petrol ones. CNG, LPG and electric fuel types only make up a small portion of the vehicles.

resale5 %>% group\_by(registered\_year,fuel\_type) %>% summarise(fuelcount = n())

## `summarise()` has grouped output by 'registered\_year'. You can override using  
## the `.groups` argument.

## # A tibble: 65 × 3  
## # Groups: registered\_year [22]  
## registered\_year fuel\_type fuelcount  
## <dbl> <chr> <int>  
## 1 2002 Petrol 1  
## 2 2003 Petrol 5  
## 3 2004 Diesel 1  
## 4 2004 Petrol 11  
## 5 2005 Diesel 1  
## 6 2005 Petrol 7  
## 7 2006 Diesel 2  
## 8 2006 Petrol 7  
## 9 2007 CNG 1  
## 10 2007 Diesel 3  
## # ℹ 55 more rows

ggplot(data = resale5) + geom\_bar(mapping = aes(x = registered\_year,fill = fuel\_type)) + labs(x = 'Registered year',y = 'Fuel type count')



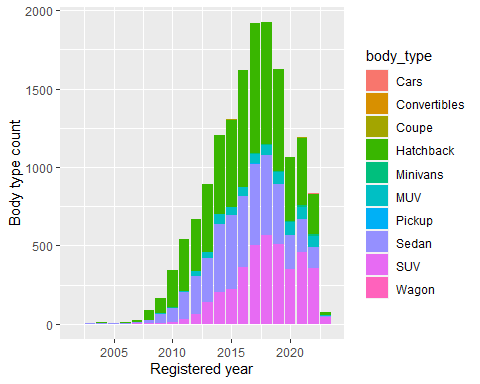
Overall, hatchbacks, SUV and sedans comprise the majority of cars body type and the rest of them are only a small part of the dataset.

resale5 %>% group\_by(registered\_year,body\_type) %>% summarise(btypecount = n())

## `summarise()` has grouped output by 'registered\_year'. You can override using  
## the `.groups` argument.

## # A tibble: 104 × 3  
## # Groups: registered\_year [22]  
## registered\_year body\_type btypecount  
## <dbl> <chr> <int>  
## 1 2002 Hatchback 1  
## 2 2003 Hatchback 2  
## 3 2003 Sedan 3  
## 4 2004 Hatchback 9  
## 5 2004 SUV 2  
## 6 2004 Sedan 1  
## 7 2005 Hatchback 5  
## 8 2005 SUV 2  
## 9 2005 Sedan 1  
## 10 2006 Hatchback 4  
## # ℹ 94 more rows

ggplot(data = resale5) + geom\_bar(mapping = aes(x = registered\_year,fill = body\_type)) + labs(x = 'Registered year',y = 'Body type count')



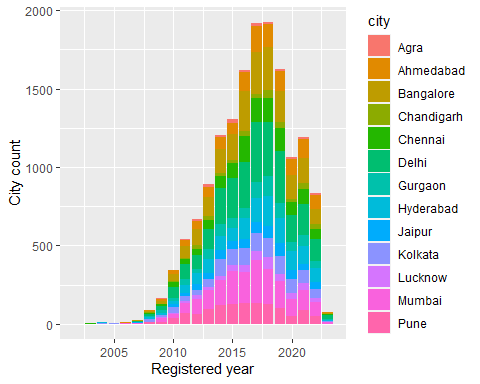
The tier 1 cities like Bangalore, Delhi, Mumbai see the larger part of the vehicles spread compared to other cities.

resale5 %>% group\_by(registered\_year,city) %>% summarise(citycount = n())

## `summarise()` has grouped output by 'registered\_year'. You can override using  
## the `.groups` argument.

## # A tibble: 236 × 3  
## # Groups: registered\_year [22]  
## registered\_year city citycount  
## <dbl> <chr> <int>  
## 1 2002 Ahmedabad 1  
## 2 2003 Bangalore 1  
## 3 2003 Chandigarh 1  
## 4 2003 Chennai 1  
## 5 2003 Jaipur 1  
## 6 2003 Lucknow 1  
## 7 2004 Ahmedabad 2  
## 8 2004 Bangalore 1  
## 9 2004 Hyderabad 4  
## 10 2004 Jaipur 2  
## # ℹ 226 more rows

ggplot(data = resale5) + geom\_bar(mapping = aes(x = registered\_year,fill = city)) + labs(x = 'Registered year',y = 'City count')

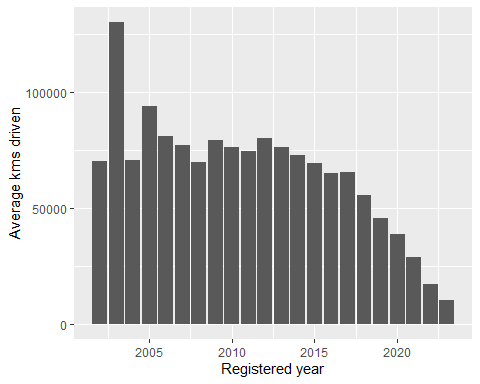


The average kilometers driven starts declining clearly after 2015 after staying consistent till 2015.

resale5 %>% group\_by(registered\_year) %>% summarise(kms\_mean = mean(kms\_driven))

## # A tibble: 22 × 2  
## registered\_year kms\_mean  
## <dbl> <dbl>  
## 1 2002 70000   
## 2 2003 130100   
## 3 2004 70467.  
## 4 2005 94000   
## 5 2006 81111.  
## 6 2007 76984.  
## 7 2008 69743.  
## 8 2009 79389.  
## 9 2010 76041.  
## 10 2011 74511.  
## # ℹ 12 more rows

resale5 %>% group\_by(registered\_year) %>% summarise(kms\_mean = mean(kms\_driven)) %>% ggplot(aes(registered\_year,kms\_mean)) + geom\_col() + labs(x = 'Registered year',y = 'Average kms driven')

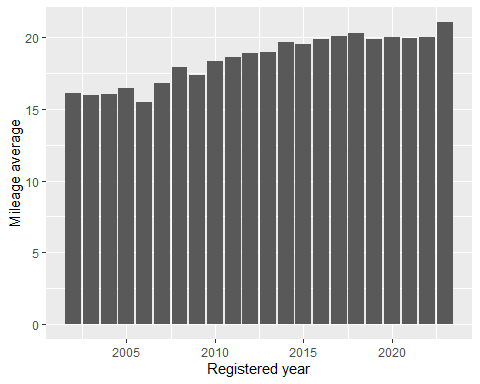


The average mileage has seen a slight increase over the years.

resale5 %>% group\_by(registered\_year) %>% summarise(mileagecount = mean(mileage\_kmpl))

## # A tibble: 22 × 2  
## registered\_year mileagecount  
## <dbl> <dbl>  
## 1 2002 16.1  
## 2 2003 15.9  
## 3 2004 16.0  
## 4 2005 16.4  
## 5 2006 15.5  
## 6 2007 16.8  
## 7 2008 17.9  
## 8 2009 17.4  
## 9 2010 18.3  
## 10 2011 18.6  
## # ℹ 12 more rows

resale5 %>% group\_by(registered\_year) %>% summarise(mileagecount = mean(mileage\_kmpl)) %>% ggplot(aes(registered\_year,mileagecount)) + geom\_col() + labs(x = 'Registered year',y = 'Mileage average')

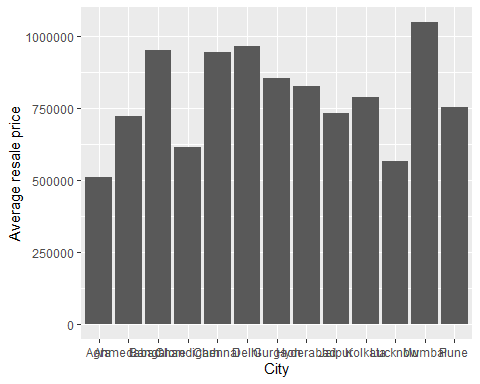


The tier 1 cities of Mumbai, Delhi, Bangalore, Chennai have the higher average resale price compared to other cities.

resale5 %>% group\_by(city) %>% summarise(resalemean = mean(resale\_price\_rupees))

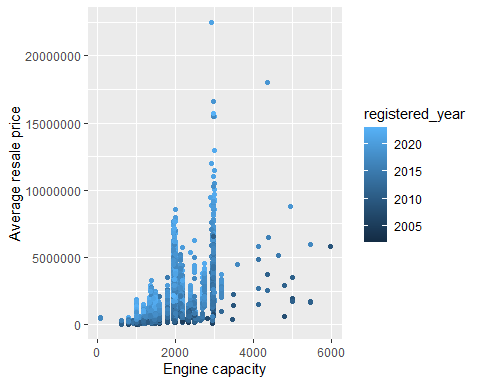
## # A tibble: 13 × 2  
## city resalemean  
## <chr> <dbl>  
## 1 Agra 509191.  
## 2 Ahmedabad 723551.  
## 3 Bangalore 951267.  
## 4 Chandigarh 614251.  
## 5 Chennai 945564.  
## 6 Delhi 967074.  
## 7 Gurgaon 853203.  
## 8 Hyderabad 826383.  
## 9 Jaipur 732418.  
## 10 Kolkata 788961.  
## 11 Lucknow 566427.  
## 12 Mumbai 1049630.  
## 13 Pune 752451.

resale5 %>% group\_by(city) %>% summarise(resalemean = mean(resale\_price\_rupees)) %>% ggplot(aes(city,resalemean)) + geom\_col() + labs(x = 'City',y = 'Average resale price')



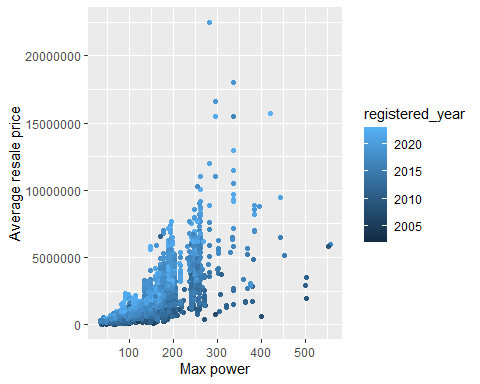
Now lets see how engine capacity affects the price of the vehicles over the years. We can see that the higher engine capacity vehicles have higher average prices but it also applies to the newer vehicles compared to the older models.

ggplot(data = resale5) + geom\_point(mapping = aes(x = engine\_capacity,y = resale\_price\_rupees,color = registered\_year)) + labs(x = 'Engine capacity',y = 'Average resale price')



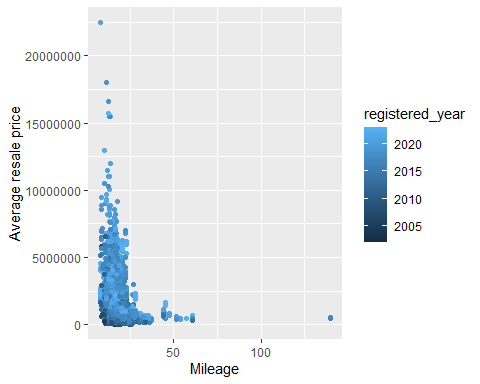
Just like engine capacity, the average resale price increases for higher max power but it again applies to the newer models compared to the older ones.

ggplot(data = resale5) + geom\_point(mapping = aes(x = max\_power\_bhp,y = resale\_price\_rupees,color = registered\_year)) + labs(x = 'Max power',y = 'Average resale price')



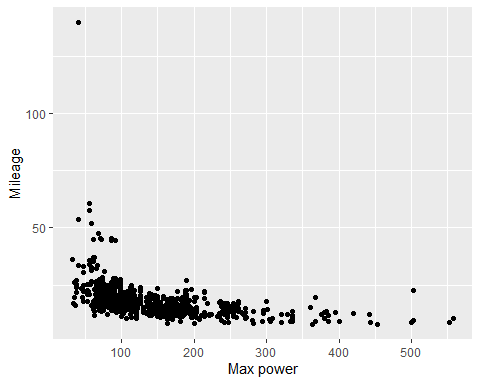
The lesser mileage vehicles have a larger average resale prices compared to the higher mileage ones.

ggplot(data = resale5) + geom\_point(mapping = aes(x = mileage\_kmpl,y = resale\_price\_rupees,color = registered\_year)) + labs(x = 'Mileage',y = 'Average resale price')

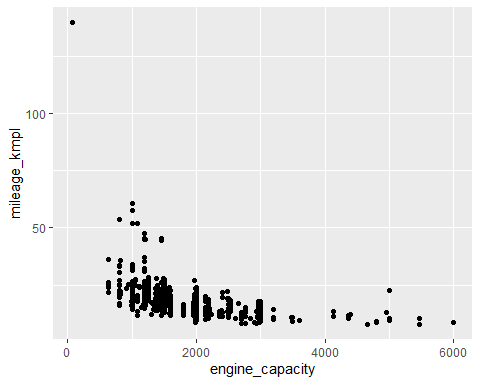


Comparing max power and mileage, the mileage goes on decreasing as the max power increases for vehicles. The same is true when comparing mileage with max power of a vehicle. And finally engine capacity increase corresponds to a max power increase in vehicles.

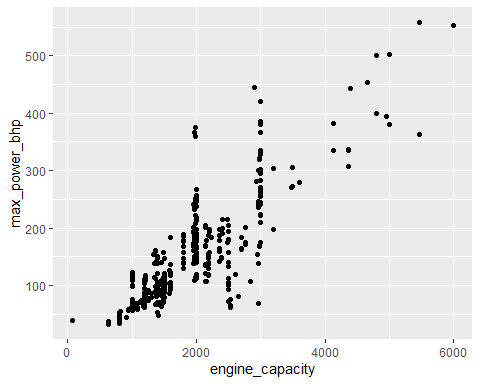
ggplot(data = resale5) + geom\_point(mapping = aes(x = max\_power\_bhp,y = mileage\_kmpl)) + labs(x = 'Max power',y = 'Mileage')



ggplot(data = resale5) + geom\_point(mapping = aes(x = engine\_capacity,y = mileage\_kmpl))

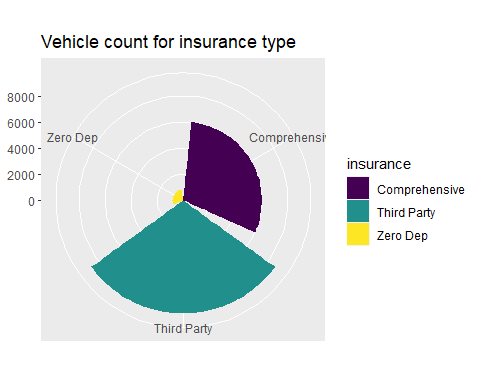


ggplot(data = resale5) + geom\_point(mapping = aes(x = engine\_capacity,y = max\_power\_bhp))



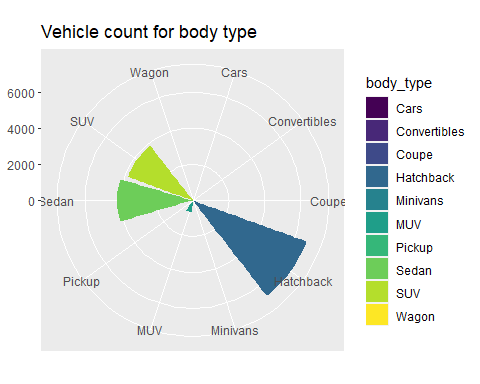
Looking at the insurance types for vehicles, the plot shows that Third party and comprehensive insurance types are the most common for vehicles. Zero dep makes up only a small percentage.

resale5 %>% group\_by(insurance) %>% count(insurance) %>%   
 ggplot(aes(insurance,n,fill = insurance)) +  
 geom\_col()+  
 coord\_polar()+  
 scale\_fill\_ordinal()+  
 labs(x = NULL, y = NULL,title = 'Vehicle count for insurance type')



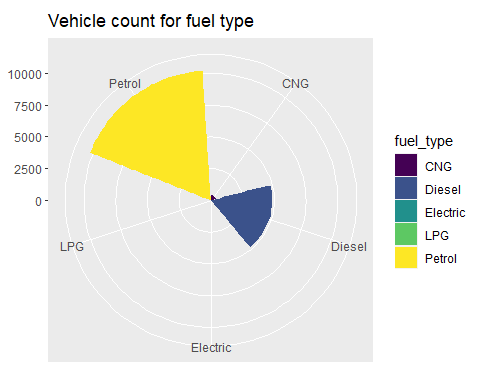
Looking at the body type for vehicles, Sedan, SUV and hatchbacks make up the largest part of vehicles.

resale5 %>% group\_by(body\_type) %>% count(body\_type) %>%   
 ggplot(aes(body\_type,n,fill = body\_type)) +  
 geom\_col()+  
 coord\_polar()+  
 scale\_fill\_ordinal()+  
 labs(x = NULL, y = NULL,title = 'Vehicle count for body type')



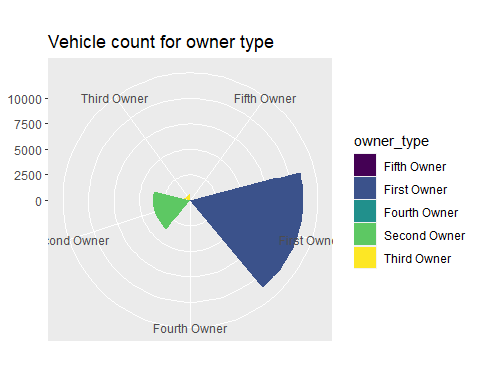
Looking at the fuel type of vehicles, Petrol and Diesel are the most common fuel for vehicles with CNG, LPG and electric making only a small portion of the total.

resale5 %>% group\_by(fuel\_type) %>% count(fuel\_type) %>%   
 ggplot(aes(fuel\_type,n,fill = fuel\_type)) +  
 geom\_col()+  
 coord\_polar()+  
 scale\_fill\_ordinal()+  
 labs(x = NULL, y = NULL, title = 'Vehicle count for fuel type')



Looking at the owner type of vehicles, most vehicles have first or second owners and the remaining vehicles have third,fourth or fifth owners.

resale5 %>% group\_by(owner\_type) %>% count(owner\_type) %>%   
 ggplot(aes(owner\_type,n,fill = owner\_type)) +  
 geom\_col()+  
 coord\_polar()+  
 scale\_fill\_ordinal()+  
 labs(x = NULL, y = NULL,title = "Vehicle count for owner type")



Main observations

1. The vehicle prices have been increasing but many sales have happened in the tier 1 cities and also in recent years the first owners selling have increased which translates to higher prices.
2. The manual transmission vehicles were the standard earlier but people have started using automatics more and more showing that people prefer the comfort of driving offered by automatic vehicles.
3. The average kilometers driven has decreased in recent years and combined with the fact that many sellers are first owners, more people are selling their vehicle quicker than before.
4. Most people sell petrol and diesel vehicles showing that vehicles with other fuel types have a lesser market share.
5. The higher engine capacity corresponds to higher max power leading to lesser mileage in vehicles.
6. Tier 1 cities have a higher resale price showing that it might add an extra premium if the vehicle is sold in a tier 1 city.
7. SUV, sedan and hatchback are the most common vehicles that are sold.
8. The average mileage has slightly increased indicating that the efficiency has become better over the years.
9. The average resale prices are higher when engine capacity and max power are higher but lower when the mileage is higher showing that engines with less power correspond to lesser prices.