Toyota Analysis

Sakthivel Sinnasamy

11/02/2020

Introduction

Using the collected data from Toyota dealership in order to understand the relationship between the outcome of interest, dependent variable Price with other sets of factors or independent variables.

Section 1

Linear regression analysis is a suitable analysis method to analyse the given model. The primary objective of this analysis is to provide an inference on the relationship between the outcome of interest, variable price with the other sets of factors or variable in the dataset. Since the outcome of interest, variable price is a numerical variable, the common analysis method used to understand the connection between the dependent variable; price and other independent variable in the collected data. As mentioned by (Worster et al., 2007), to be suitable for linear regression the outcome of interest must be a variable that is not categorical instead must be a continuous numerical range variable, which in this case Price is a continuous numerical range variable. Assuming the data verifies to the linear regression analysis assumption which is that the dependent variable follows normal distribution, has linear relationship with predictors, observation independent of one another, and homoscedastic; where across the regression line the variance of the residuals are the same then the linear regression analysis is a suitable for this model (Uyanık & Güler, 2013). Furthermore, linear regression analysis enables multiple linear regression where there can only be single outcome of interest but 2 or more predictor variables. According to the task objective there is single outcome of interest; Price that need to be explained on its relationship with a set of predictor variable, adding the suitability of using linear regression analysis.

(238 words)

Section 2

Model A

Regression model:- price = age_08_04 + km + fuel_type + met_color + weight

Regression equation:-

$$E(PRICE_i) = INTERCEPT + \beta_1 age0804_{i1} + \beta_2 km_{i2} + \beta_3 diesel_{i3} + \beta_4 petrol_{i4} + \beta_5 metallic_{i5} + \beta_6 weight_{i6} + \epsilon_i$$

In model A, independent variable age_08_04 is selected as it is well known that an older car depreciates in value compare to a newer car. KM independent variable is selected because the more mileage in the car indicates a higher wear and tear resulting price drop (Englmaier et al., 2013). Fuel type is also selected due to the correlation with driving cost that fluctuates with different vehicles and its fuel type (National Research Council, 1992). As per Greim (2017), metallic paint, engine size and weight of a car has positive influence on the residual value of a car. Hence the independent variable met_color, cc and weight have been selected to be the independent variable deeming that these variables have a relationship on the outcome of interest; Price in this case.

Model B

Regression model:- price = age_08_04 + km + fuel_type + cc + automatic

Regression equation:-

$$E(PRICE_i) = INTERCEPT + \beta_1 age0804_{i1} + \beta_2 km_{i2} + \beta_3 diesel_{i3} + \beta_4 petrol_{i4} + \beta_5 cc_{i5} + \beta_6 automatic_{i6} + \epsilon_i$$

In model B, age and km independent variable were selected for similar reasons. In addition to this variables, fuel_type variable were also chosen to identify its relationship with price.This is because fuel economy directly influences the cost of driving and the fluctuations of fuel price correlates to the willingness of customers to pay more or less for fuel economy. Hence fuel type influences car purchase decision, there in the outcome of interest which is the car price (National Research Council, 1992). Independent variables engine size; cc and transmission; automatic have significant and positive influence on the price of the car (Greim, 2017).

Model C

Regression model:- price = age_08_04 + km + hp + abs + airco

Regression equation:-

$$E(PRICE_i) = INTERCEPT + \beta_1 age0804_{i1} + \beta_2 km_{i2} + \beta_3 hp_{i3} + \beta_4 withantilockbrakesystem_{i4} + \beta_5 airconditioning_{i5} + \epsilon_i$$

For model C, independent variable age_08_04 is selected as it is well known that an older car depreciates in value compare to a newer car.KM independent variable is selected because the more mileage in the car indicates a higher wear and tear resulting price drop (Englmaier et al., 2013). According to Greim (2017), horsepower;hp, airconditioning; airco and brake;abs also has significant relationship with the price of a car.

(293 words)

Section 3

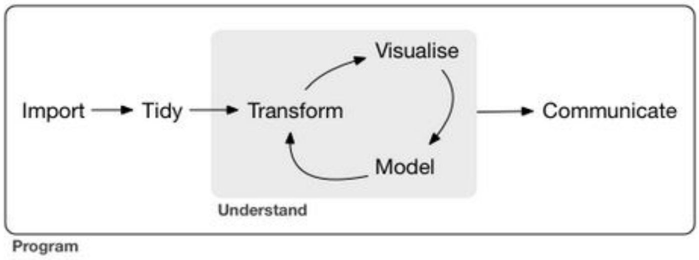


Illustration of workflow for data science

Import

Import is the first workflow stage that is required for the data science project. The objective to achieve in import stage loading data into the R data frame. The loaded will be used to perform data science in R. In this case the Toyota dataset is available in Comma Separated Value .csv file format. Using RStudio, “readr” package will be required to be installed and loaded to library to add the dataset in the RStudio environment. The skills required is to use R code “read_csv” to read the Toyota dataset. Possible challenges are that users may find RStudio may unable to find the package in its library, thereby additional steps need to be taken install “readr” or “tidyverse” package which enables “readr” to be used.

Tidy

The next stage is to tidy up the data’s in the dataset. The objective of this stage is to tidy up data so that each column of the data is a variable and each row of the data is an observation. The package that will be useful for tidy stage is the “janitor” package which must be loaded in RStudio. The skills required is to examine the Toyota dataset for inconsistency and also skills to use functions in “janitor” package as required. Challenges that are faced is the time spent in examining every single observation row for any inclusion of variables especially for large datasets.

Transform

The objective of the transform stage is to narrow down the Toyota dataset by creating a subset of the variables from the main dataset. This subset of variables will be wrangled further to obtain desired observation, statistic or outputs. The packages that will be used in this stage is the “dplyr” package loaded in RStudio. The skills required in this stage to successfully manipulate the variable using “dplyr” functions such as “select()” to generate narrowed down variables dataset, “filter()” to extract data based constrain parameters, “rename()” to create a new mutated variable and others according model output requirement. The challenges faced is that wrong manipulation could result in incorrect observation output and the relationship between the outcome variable and predictors will not be accurate. Another challenge is irrelevant variables are manipulated resulting in desired output unable to be achieved.

Visualisation

Visualisation stage objective is to visualise the data for interpretation. This will assist in identifying outliers, unexpected data variability, and questioning unusual data. The required package for this stage is the “ggplot2” package loaded in RStudio. The skill required for this stage is the ability to spot outliers or interpret the visualisation to identify the unusual data instances. The challenge is the ability to identify data instances that are outliers.

Model

In model stage, the objective is to model applying multiple linear regression on the selected outcome and predictor variable of the Toyota dataset, tidying the result and assessing the model. The required packages for this stages is the “broom” and “kableExtra” packages which enables to use functions like “augment()”, “tidy()”, “glance()” and “kable()’. The required skill is the ability to understand why certain data point is much different with the rest. The challenge is to determine what is the appropriate range for identifying the outliers, as too many data instances could be disregarded if the range parameter not determined appropriately.

Communication

Communication stage where the output results are interpreted and communicated to others. The skill required is too interpret the relationship between predictor variable and outcome variable based on the results output. Challenge is correctly interpreting the estimate and p value.

(585 words)

Analysis

Import

Load libraries to do analysis

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(broom)
library(readxl)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Read data in the folder of project directory

```r
toyota <- read_csv('Toyota_corolla.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   Model = col_character(),
##   Fuel_Type = col_character(),
##   Color = col_character()
## )
## i Use `spec()` for the full column specifications.
```

Tidy

Clean data

```r
toyota <- toyota %>%
  clean_names()
```

View data

```r
summary(toyota)
```

```
##        id            model               price         age_08_04
##  Min.   :   1.0   Length:1436        Min.   : 4350   Min.   : 1.00
##  1st Qu.: 361.8   Class :character   1st Qu.: 8450   1st Qu.:44.00
##  Median : 721.5   Mode  :character   Median : 9900   Median :61.00
##  Mean   : 721.6                      Mean   :10731   Mean   :55.95
##  3rd Qu.:1081.2                      3rd Qu.:11950   3rd Qu.:70.00
##  Max.   :1442.0                      Max.   :32500   Max.   :80.00
##    mfg_month         mfg_year          km           fuel_type
##  Min.   : 1.000   Min.   :1998   Min.   :     1   Length:1436
##  1st Qu.: 3.000   1st Qu.:1998   1st Qu.: 43000   Class :character
##  Median : 5.000   Median :1999   Median : 63390   Mode  :character
##  Mean   : 5.549   Mean   :2000   Mean   : 68533
##  3rd Qu.: 8.000   3rd Qu.:2001   3rd Qu.: 87021
##  Max.   :12.000   Max.   :2004   Max.   :243000
##        hp          met_color         color              automatic
##  Min.   : 69.0   Min.   :0.0000   Length:1436        Min.   :0.00000
##  1st Qu.: 90.0   1st Qu.:0.0000   Class :character   1st Qu.:0.00000
##  Median :110.0   Median :1.0000   Mode  :character   Median :0.00000
##  Mean   :101.5   Mean   :0.6748                      Mean   :0.05571
##  3rd Qu.:110.0   3rd Qu.:1.0000                      3rd Qu.:0.00000
##  Max.   :192.0   Max.   :1.0000                      Max.   :1.00000
##        cc            doors         cylinders     gears       quarterly_tax
##  Min.   : 1300   Min.   :2.000   Min.   :4   Min.   :3.000   Min.   : 19.00
##  1st Qu.: 1400   1st Qu.:3.000   1st Qu.:4   1st Qu.:5.000   1st Qu.: 69.00
##  Median : 1600   Median :4.000   Median :4   Median :5.000   Median : 85.00
##  Mean   : 1577   Mean   :4.033   Mean   :4   Mean   :5.026   Mean   : 87.12
##  3rd Qu.: 1600   3rd Qu.:5.000   3rd Qu.:4   3rd Qu.:5.000   3rd Qu.: 85.00
##  Max.   :16000   Max.   :5.000   Max.   :4   Max.   :6.000   Max.   :283.00
##      weight     mfr_guarantee    bovag_guarantee  guarantee_period
##  Min.   :1000   Min.   :0.0000   Min.   :0.0000   Min.   : 3.000
##  1st Qu.:1040   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.: 3.000
##  Median :1070   Median :0.0000   Median :1.0000   Median : 3.000
##  Mean   :1072   Mean   :0.4095   Mean   :0.8955   Mean   : 3.815
##  3rd Qu.:1085   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 3.000
##  Max.   :1615   Max.   :1.0000   Max.   :1.0000   Max.   :36.000
##       abs             airbag_1         airbag_2          airco
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.8134   Mean   :0.9708   Mean   :0.7228   Mean   :0.5084
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  automatic_airco   boardcomputer      cd_player       central_lock
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.05641   Mean   :0.2946   Mean   :0.2187   Mean   :0.5801
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  powered_windows power_steering       radio          mistlamps
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :1.000   Median :1.0000   Median :0.0000   Median :0.000
##  Mean   :0.562   Mean   :0.9777   Mean   :0.1462   Mean   :0.257
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##   sport_model     backseat_divider  metallic_rim    radio_cassette
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3001   Mean   :0.7702   Mean   :0.2047   Mean   :0.1455
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  parking_assistant     tow_bar
##  Min.   :0.000000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.0000
##  Median :0.000000   Median :0.0000
##  Mean   :0.002786   Mean   :0.2779
##  3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :1.000000   Max.   :1.0000
```

Transform

Select variable

Model A

```
toyotaA <- toyota %>%
  select(id, price, age_08_04, km, fuel_type, met_color, weight)
glimpse(toyotaA)
```

```
## Rows: 1,436
## Columns: 7
## $ id        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ price     <dbl> 13500, 13750, 13950, 14950, 13750, 12950, 16900, 18600, 2150~
## $ age_08_04 <dbl> 23, 23, 24, 26, 30, 32, 27, 30, 27, 23, 25, 22, 25, 31, 32, ~
## $ km        <dbl> 46986, 72937, 41711, 48000, 38500, 61000, 94612, 75889, 1970~
## $ fuel_type <chr> "Diesel", "Diesel", "Diesel", "Diesel", "Diesel", "Diesel", ~
## $ met_color <dbl> 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, ~
## $ weight    <dbl> 1165, 1165, 1165, 1165, 1170, 1170, 1245, 1245, 1185, 1105, ~
```

Model B

```
toyotaB <- toyota %>%
  select(id, price, age_08_04, km, fuel_type, cc, automatic)
glimpse(toyotaB)
```

```
## Rows: 1,436
## Columns: 7
## $ id        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ price     <dbl> 13500, 13750, 13950, 14950, 13750, 12950, 16900, 18600, 2150~
## $ age_08_04 <dbl> 23, 23, 24, 26, 30, 32, 27, 30, 27, 23, 25, 22, 25, 31, 32, ~
## $ km        <dbl> 46986, 72937, 41711, 48000, 38500, 61000, 94612, 75889, 1970~
## $ fuel_type <chr> "Diesel", "Diesel", "Diesel", "Diesel", "Diesel", "Diesel", ~
## $ cc        <dbl> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 1800, 1900, ~
## $ automatic <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

Model C

```
toyotaC <- toyota %>%
  select(id, price, age_08_04, km, hp, abs, airco)
glimpse(toyotaC)
```

```
## Rows: 1,436
## Columns: 7
## $ id        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ price     <dbl> 13500, 13750, 13950, 14950, 13750, 12950, 16900, 18600, 2150~
## $ age_08_04 <dbl> 23, 23, 24, 26, 30, 32, 27, 30, 27, 23, 25, 22, 25, 31, 32, ~
## $ km        <dbl> 46986, 72937, 41711, 48000, 38500, 61000, 94612, 75889, 1970~
## $ hp        <dbl> 90, 90, 90, 90, 90, 90, 90, 90, 192, 69, 192, 192, 192, 192,~
## $ abs       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ airco     <dbl> 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

Mutate variables

Model A

```
toyotaA <- toyotaA %>%
  mutate(met_color = factor(met_color,
                            label = c('No', 'Yes')))
```

Model B

```
toyotaB <- toyotaB %>%
  mutate(automatic = factor(automatic,
                            label = c('No', 'Yes')))
```

Model C

```
toyotaC <- toyotaC %>%
  mutate(abs = factor(abs,
                      label = c('No', 'Yes')))
```

```
toyotaC <- toyotaC %>%
  mutate(airco = factor(airco,
                      label = c('No', 'Yes')))
```

Section 4

Categorize a numerical variable and regroup a categorical variable

Model A

```
toyotaA <- toyotaA %>%
  mutate(catA = cut(weight,
                    breaks = c(0, 1134, 1360, 1587, 1814),
                    labels = c('Passenger cars light', 'Passenger cars compact', 'Passenger cars medium', 'Pass
enger cars heavy')))
```

The cut points for the weight variable in model A is chosen due to the cut points being the vehicle size classes of the National Highway Traffic Safety Administrations (NHTSA). This weight classes are used as the standard for New Car Assessment Program (NCAP) testing in United States (U.S.). The cuts point for passenger cars according to the standard; light, above 907kg to below 1134kg; compact, up to 1360kg, medium, up to 1587kg; and heavy, up to 1814kg. The changes to the model parameters are the weight variable was initially a numerical variable. Now the mutated categorical variable catA will result in the multiple linear regression model to have the cut points category to be referenced to the dummy category of “Passenger cars light”.

(124 words)

Model B

```
toyotaB <- toyotaB %>%
  mutate(catB = recode(fuel_type,
                       'Diesel' = 'liquid fuel',
                       'Petrol' = 'liquid fuel',
                       'CNG' = 'fuel gas'))
```

The new categories are chosen so that the fuel types can be segregated by liquid fuel and fuel gas. These two particular categories were chosen to investigate the impact of having gas cylinders on vehicle price. The fuel gas requires gas tanks to be installed in the boot of sedans such as Toyota corolla in our case, which will limit the boot space. This could potentially have a negative impact on the price of the vehicle. The changes to the model parameters is that now Model B’s multiple linear regression model predictor variable catB which replaces fuel_type, will only have ‘liquid fuel’ as a reference category to dummy category ‘fuel gas’.

(111 words)

View data

Model A

```
summary(toyotaA)
```

```
##        id             price         age_08_04           km
##  Min.   :   1.0   Min.   : 4350   Min.   : 1.00   Min.   :      1
##  1st Qu.: 361.8   1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 43000
##  Median : 721.5   Median : 9900   Median :61.00   Median : 63390
##  Mean   : 721.6   Mean   :10731   Mean   :55.95   Mean   : 68533
##  3rd Qu.:1081.2   3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 87021
##  Max.   :1442.0   Max.   :32500   Max.   :80.00   Max.   :243000
##   fuel_type         met_color     weight                         catA
##  Length:1436        No :467   Min.   :1000   Passenger cars light  :1314
##  Class :character   Yes:969   1st Qu.:1040   Passenger cars compact: 117
##  Mode  :character             Median :1070   Passenger cars medium :   4
##                               Mean   :1072   Passenger cars heavy  :   1
##                               3rd Qu.:1085
##                               Max.   :1615
```

Model B

```
summary(toyotaB)
```

```
##        id             price         age_08_04           km
##  Min.   :   1.0   Min.   : 4350   Min.   : 1.00   Min.   :     1
##  1st Qu.: 361.8   1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 43000
##  Median : 721.5   Median : 9900   Median :61.00   Median : 63390
##  Mean   : 721.6   Mean   :10731   Mean   :55.95   Mean   : 68533
##  3rd Qu.:1081.2   3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 87021
##  Max.   :1442.0   Max.   :32500   Max.   :80.00   Max.   :243000
##   fuel_type               cc        automatic      catB
##  Length:1436        Min.   : 1300   No :1356   Length:1436
##  Class :character   1st Qu.: 1400   Yes:  80   Class :character
##  Mode  :character   Median : 1600              Mode  :character
##                     Mean   : 1577
##                     3rd Qu.: 1600
##                     Max.   :16000
```

Model C

```
summary(toyotaC)
```

```
##        id             price         age_08_04           km
##  Min.   :   1.0   Min.   : 4350   Min.   : 1.00   Min.   :     1
##  1st Qu.: 361.8   1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 43000
##  Median : 721.5   Median : 9900   Median :61.00   Median : 63390
##  Mean   : 721.6   Mean   :10731   Mean   :55.95   Mean   : 68533
##  3rd Qu.:1081.2   3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 87021
##  Max.   :1442.0   Max.   :32500   Max.   :80.00   Max.   :243000
##        hp          abs        airco
##  Min.   : 69.0   No : 268   No :706
##  1st Qu.: 90.0   Yes:1168   Yes:730
##  Median :110.0
##  Mean   :101.5
##  3rd Qu.:110.0
##  Max.   :192.0
```

Exploratory data analysis (EDA)

Summarize data

Model A

```
options(scipen = 999)
toyotaA %>% summarize(mean_age = mean(age_08_04, na.rm = T),
                      mean_KM = mean(km, na.rm = T),
                      mean_Weight = mean(weight, na.rm = T),
                      sd_age = sd(age_08_04, na.rm = T),
                      sd_KM = sd(km, na.rm = T),
                      sd_Weight = sd(weight, na.rm = T)) %>%
  gather()
```

```
## # A tibble: 6 x 2
##   key           value
##   <chr>         <dbl>
## 1 mean_age       55.9
## 2 mean_KM     68533.
## 3 mean_Weight  1072.
## 4 sd_age         18.6
## 5 sd_KM       37506.
## 6 sd_Weight      52.6
```

Model B

```
options(scipen = 999)
toyotaB %>% summarize(mean_age = mean(age_08_04, na.rm = T),
                      mean_KM = mean(km, na.rm = T),
                      mean_cc = mean(cc, na.rm = T),
                      sd_age = sd(age_08_04, na.rm = T),
                      sd_KM = sd(km, na.rm = T),
                      sd_cc = sd(cc, na.rm = T)) %>%
  gather()
```

```
## # A tibble: 6 x 2
##   key        value
##   <chr>      <dbl>
## 1 mean_age    55.9
## 2 mean_KM  68533. 
## 3 mean_cc   1577. 
## 4 sd_age      18.6
## 5 sd_KM    37506. 
## 6 sd_cc      424. 
```

Model C

```
options(scipen = 999)
toyotaC %>% summarize(mean_age = mean(age_08_04, na.rm = T),
                      mean_KM = mean(km, na.rm = T),
                      mean_hp = mean(hp, na.rm = T),
                      sd_age = sd(age_08_04, na.rm = T),
                      sd_KM = sd(km, na.rm = T),
                      sd_hp = sd(hp, na.rm = T)) %>%
  gather()
```

```
## # A tibble: 6 x 2
##   key        value
##   <chr>      <dbl>
## 1 mean_age    55.9
## 2 mean_KM  68533. 
## 3 mean_hp    102. 
## 4 sd_age      18.6
## 5 sd_KM    37506. 
## 6 sd_hp       15.0
```

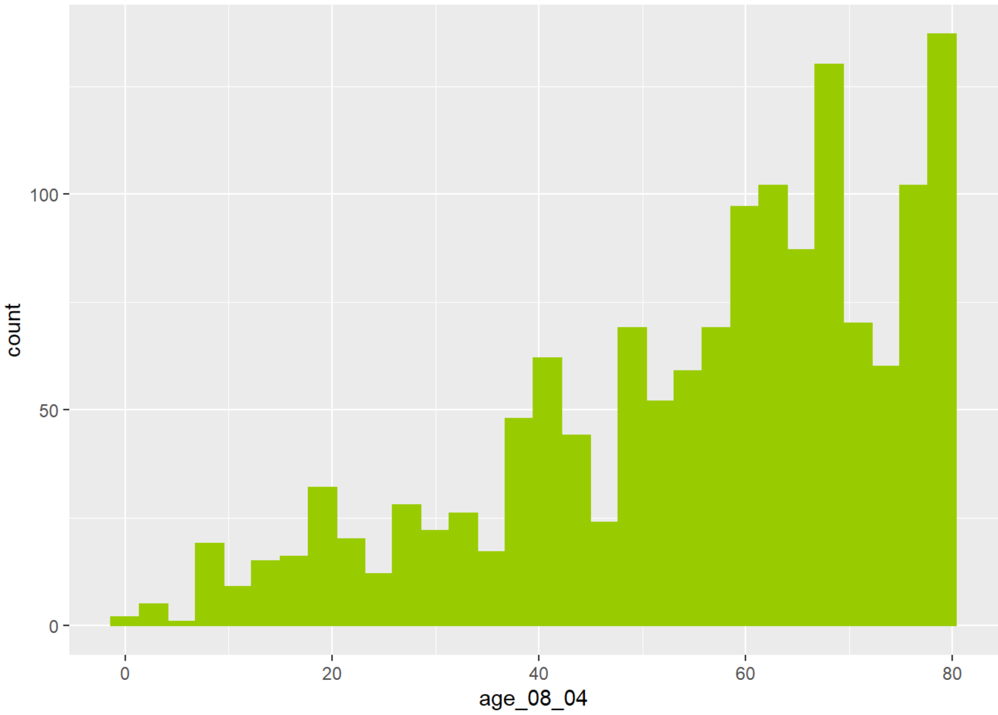
Visualise

Model A

Age of cars

```
ggplot(toyotaA, aes(age_08_04)) + geom_histogram(color="#99CC00", fill="#99CC00")
```

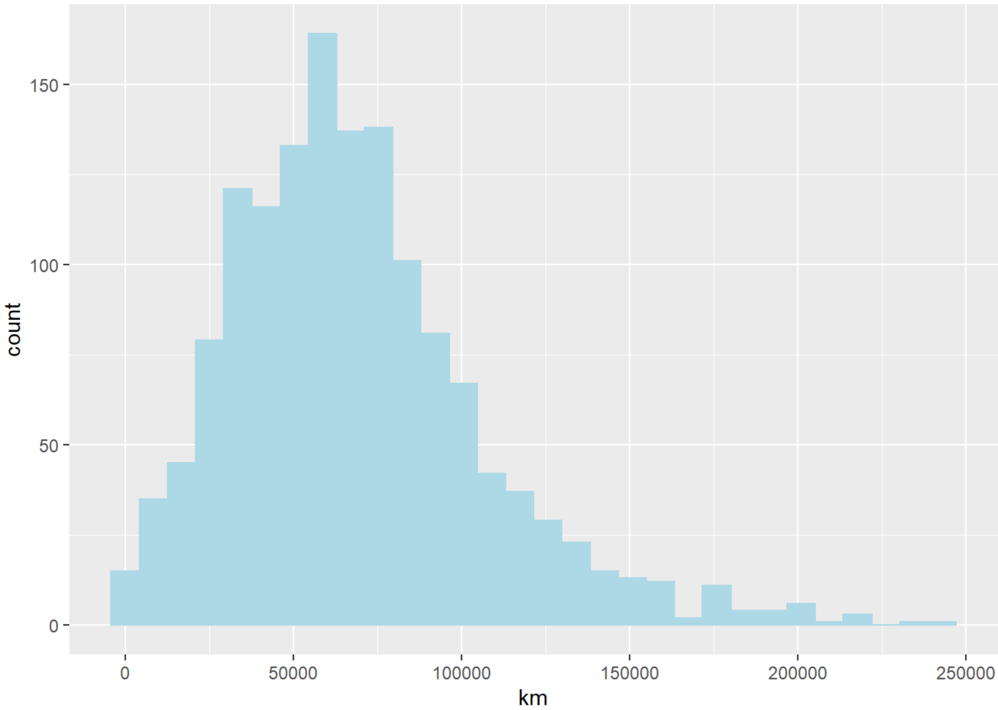
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Mileage

```
ggplot(toyotaA, aes(km)) + geom_histogram(color="lightblue", fill="lightblue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



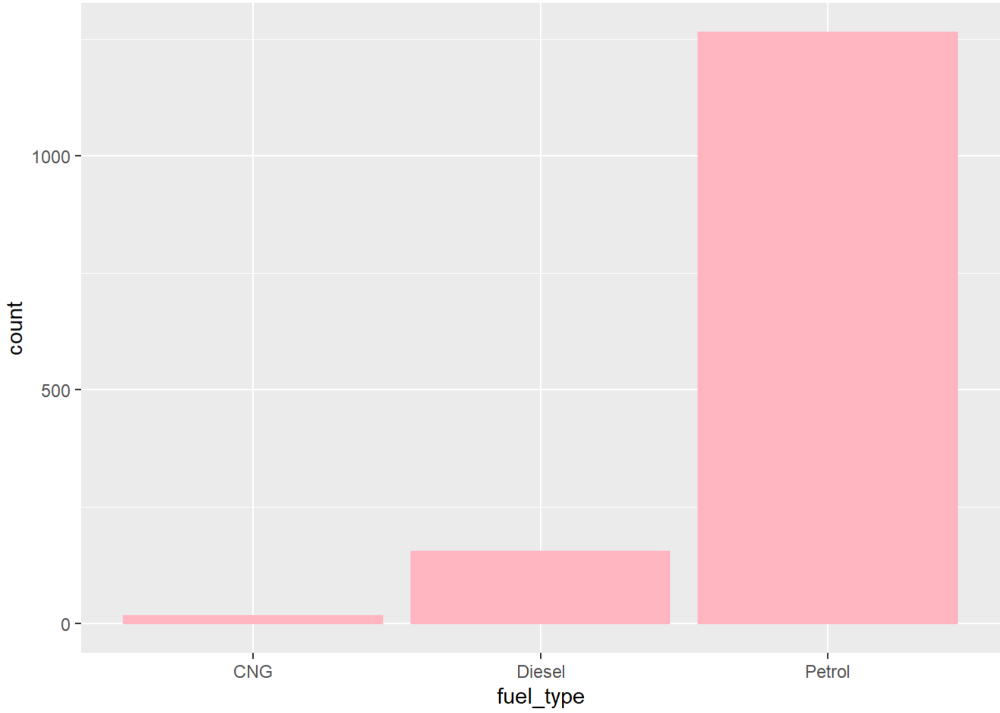
Metallic colour, Fuel type, catA (Count)

```
toyotaA %>%
  count(met_color, fuel_type, catA)
```

```
## # A tibble: 14 x 4
##    met_color fuel_type catA                     n
##    <fct>     <chr>     <fct>                <int>
##  1 No        CNG       Passenger cars light     4
##  2 No        Diesel    Passenger cars light    26
##  3 No        Diesel    Passenger cars compact  26
##  4 No        Diesel    Passenger cars medium    1
##  5 No        Petrol    Passenger cars light   399
##  6 No        Petrol    Passenger cars compact  10
##  7 No        Petrol    Passenger cars medium    1
##  8 Yes       CNG       Passenger cars light    13
##  9 Yes       Diesel    Passenger cars light    34
## 10 Yes       Diesel    Passenger cars compact  66
## 11 Yes       Diesel    Passenger cars medium    2
## 12 Yes       Petrol    Passenger cars light   838
## 13 Yes       Petrol    Passenger cars compact  15
## 14 Yes       Petrol    Passenger cars heavy     1
```

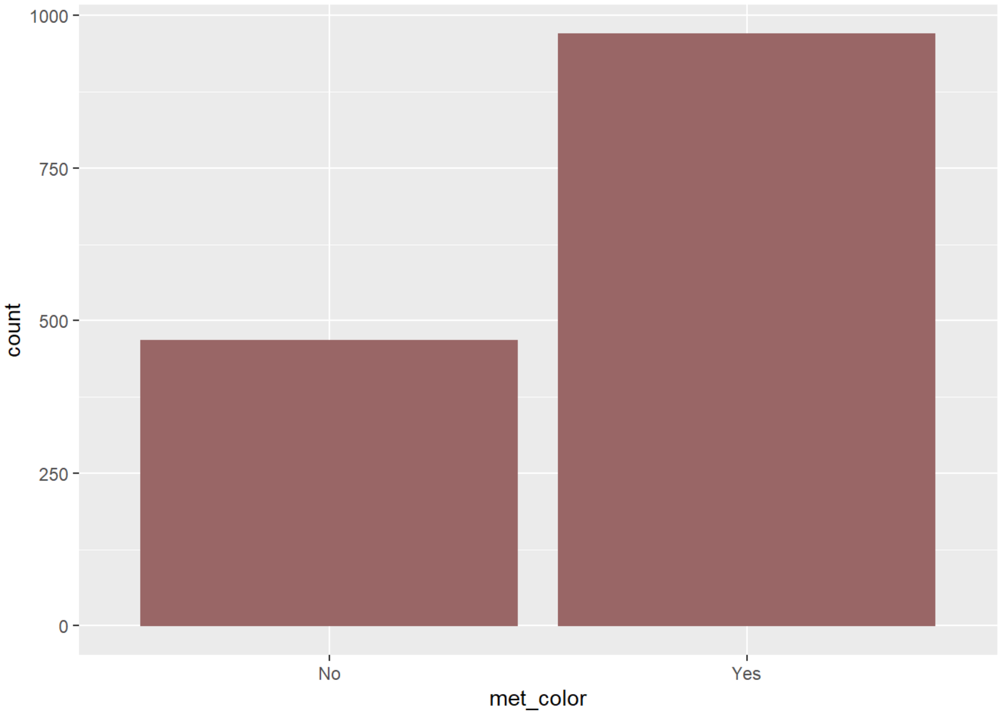
Fuel type (count) plot

```
ggplot(toyotaA, aes(fuel_type)) + geom_bar(color="lightpink", fill="lightpink")
```



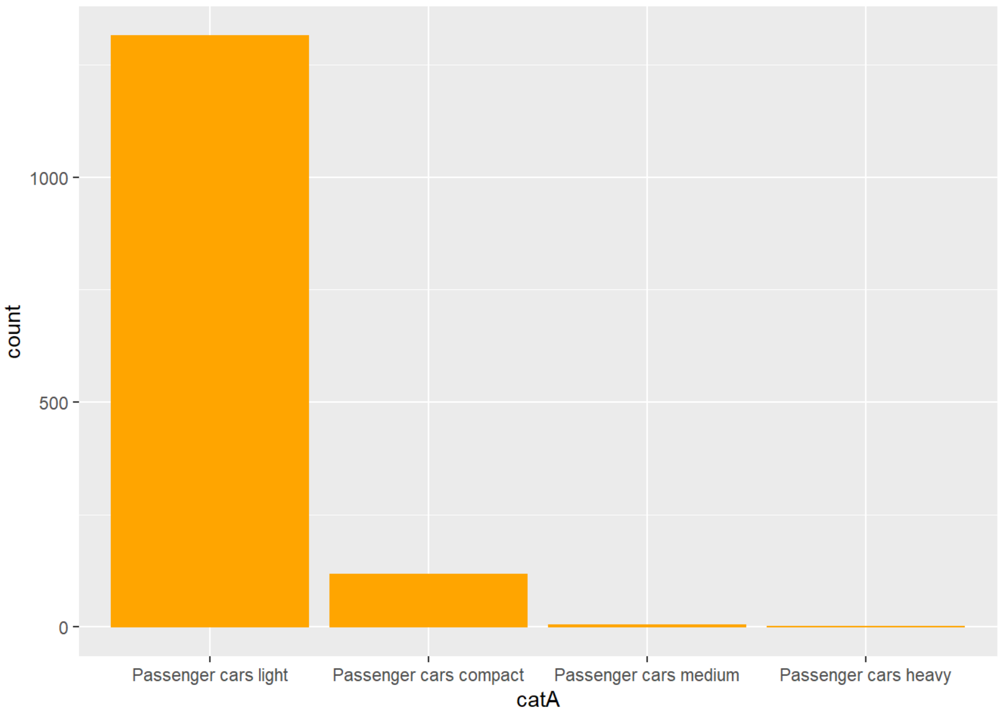
Metallic color (count)plot

```
ggplot(toyotaA, aes(met_color)) + geom_bar(color="#996666", fill="#996666")
```



Metallic color (count)plot

```
ggplot(toyotaA, aes(catA)) + geom_bar(color="orange", fill="orange")
```

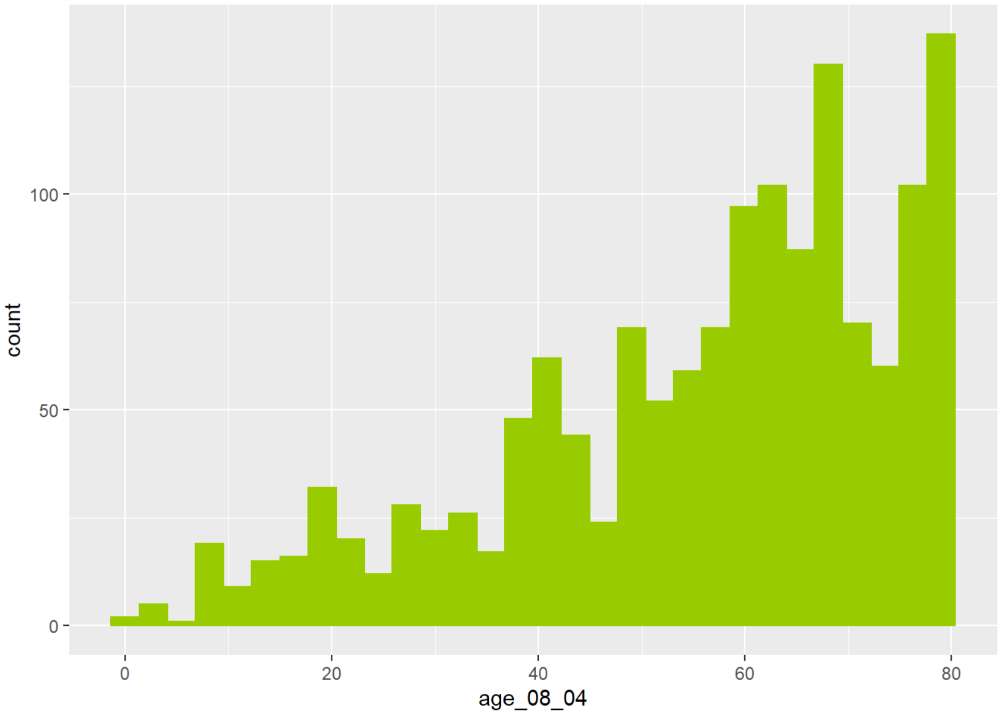


Model B

Age of cars

```
ggplot(toyotaB, aes(age_08_04)) + geom_histogram(color="#99CC00", fill="#99CC00")
```

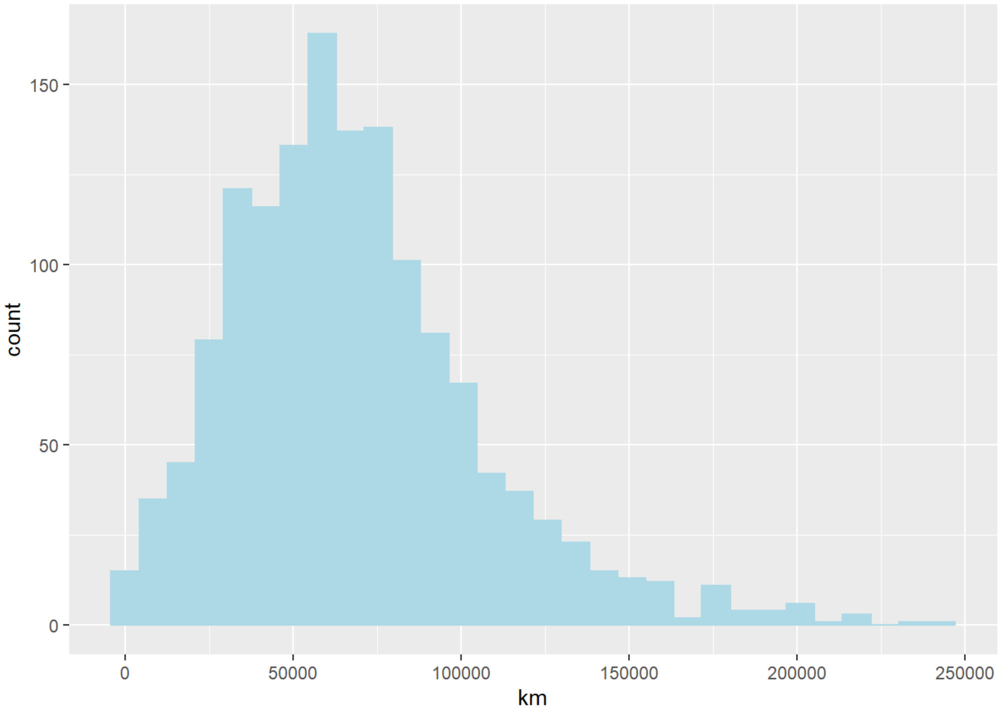
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Mileage

```
ggplot(toyotaB, aes(km)) + geom_histogram(color="lightblue", fill="lightblue")
```

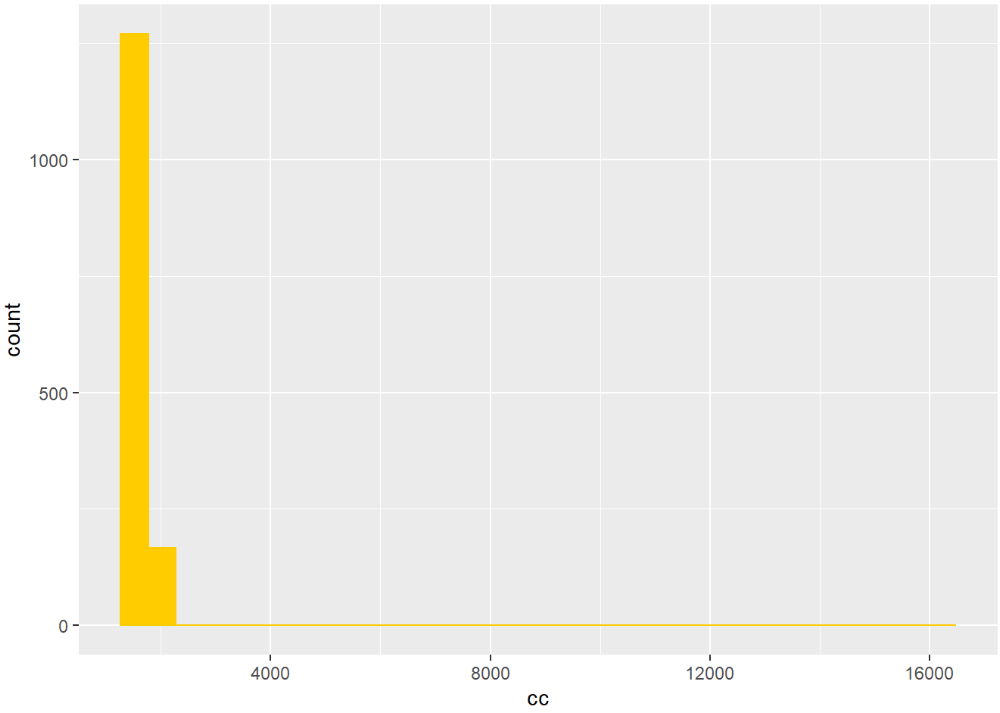
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Engine size, cc

```
ggplot(toyotaB, aes(cc)) + geom_histogram(color="#FFCC00", fill="#FFCC00")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



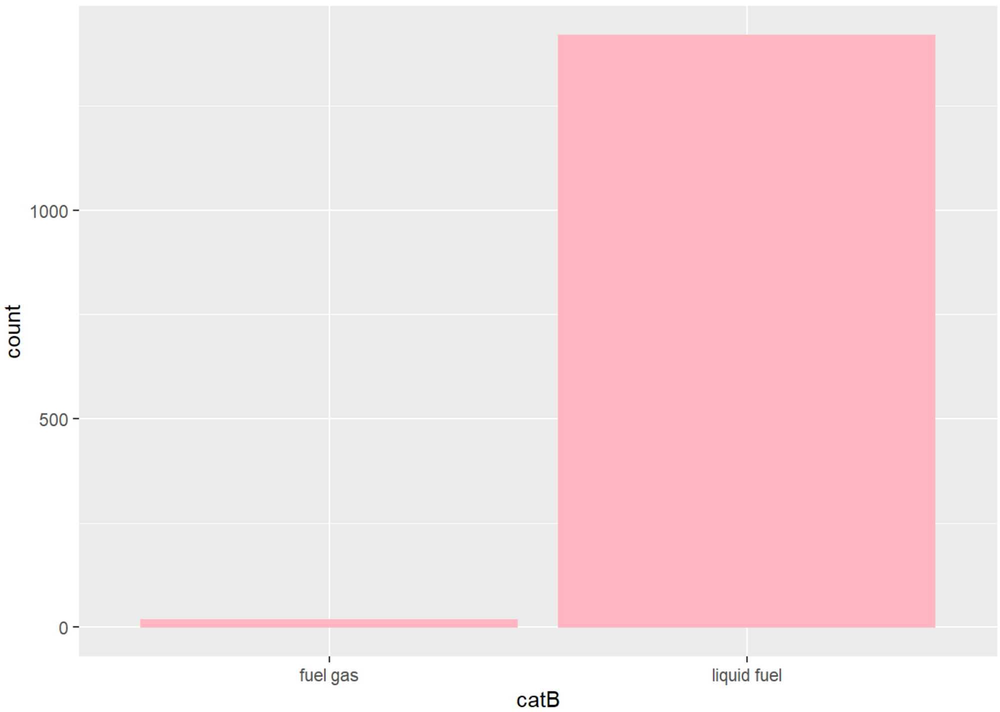
Automatic, CatB (Count)

```
toyotaB %>%
  count(automatic, catB)
```

```
## # A tibble: 4 x 3
##   automatic catB            n
##   <fct>     <chr>       <int>
## 1 No        fuel gas       16
## 2 No        liquid fuel  1340
## 3 Yes       fuel gas        1
## 4 Yes       liquid fuel    79
```

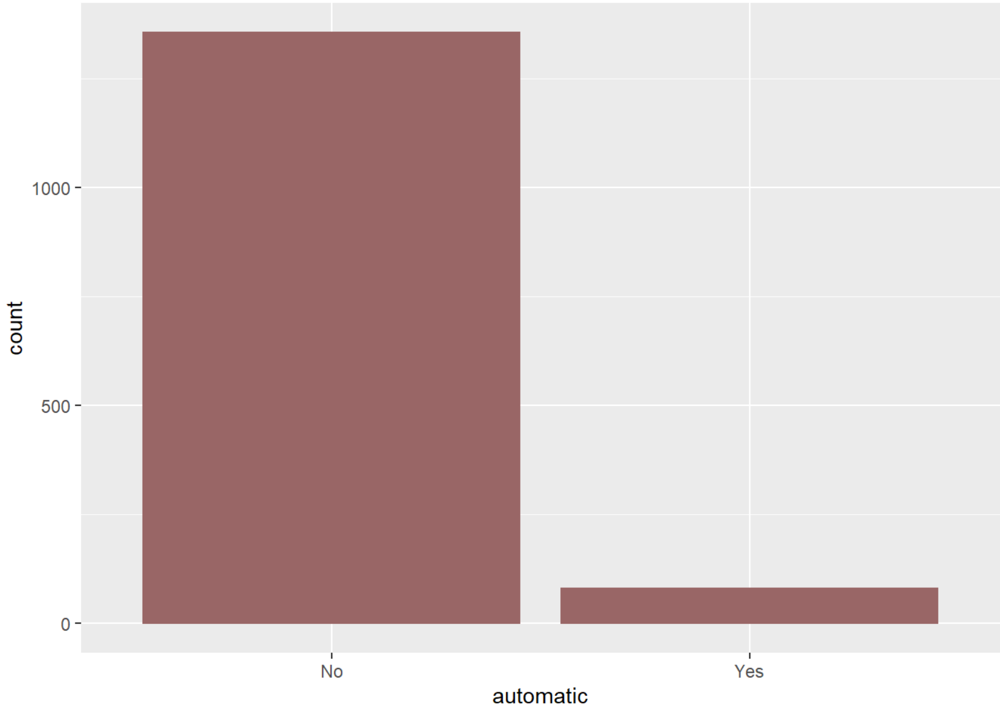
catB (count) plot

```
ggplot(toyotaB, aes(catB)) + geom_bar(color="lightpink", fill="lightpink")
```



Automatic (count) plot

```
ggplot(toyotaB, aes(automatic)) + geom_bar(color="#996666", fill="#996666")
```

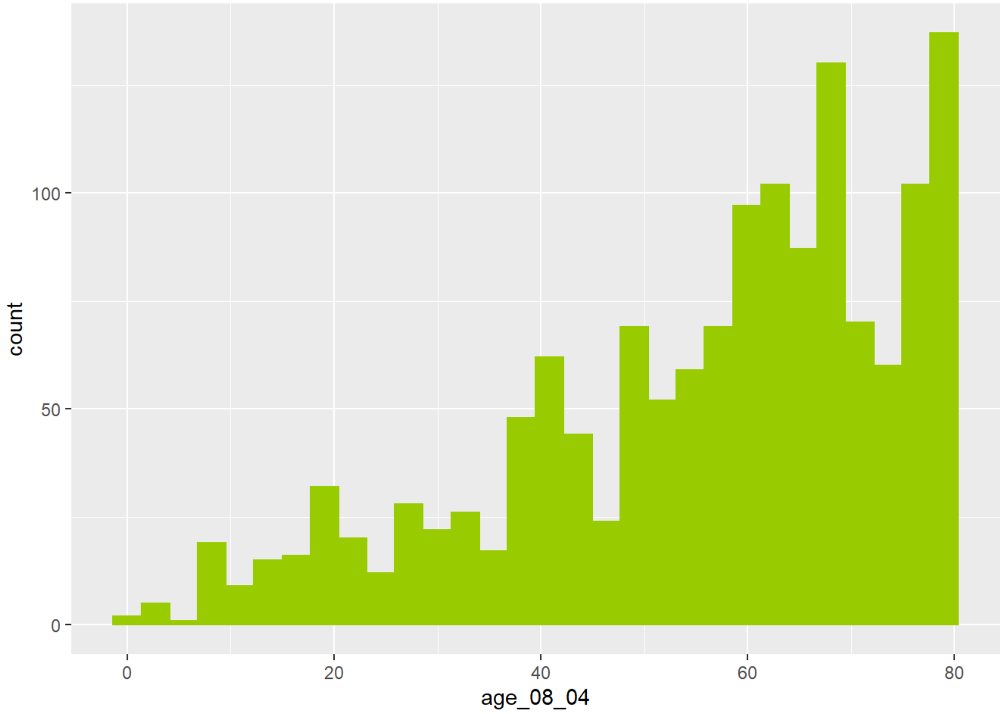


Model C

Age of cars

```
ggplot(toyotaC, aes(age_08_04)) + geom_histogram(color="#99CC00", fill="#99CC00")
```

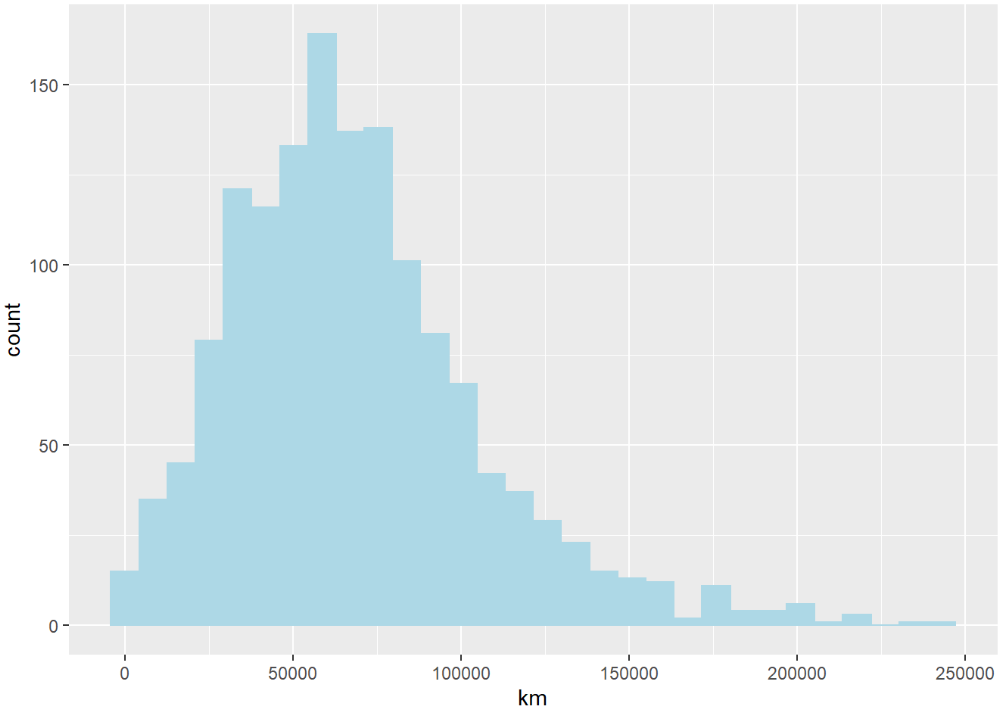
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Mileage

```
ggplot(toyotaC, aes(km)) + geom_histogram(color="lightblue", fill="lightblue")
```

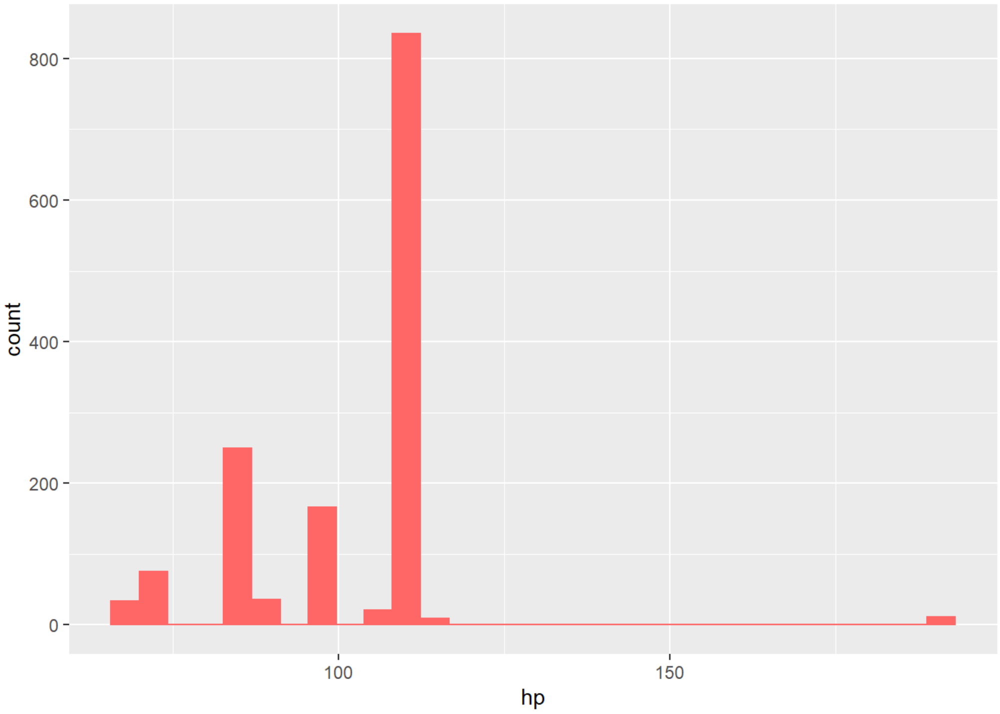
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Horsepower

```
ggplot(toyotaC, aes(hp)) + geom_histogram(color="#FF6666", fill="#FF6666")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



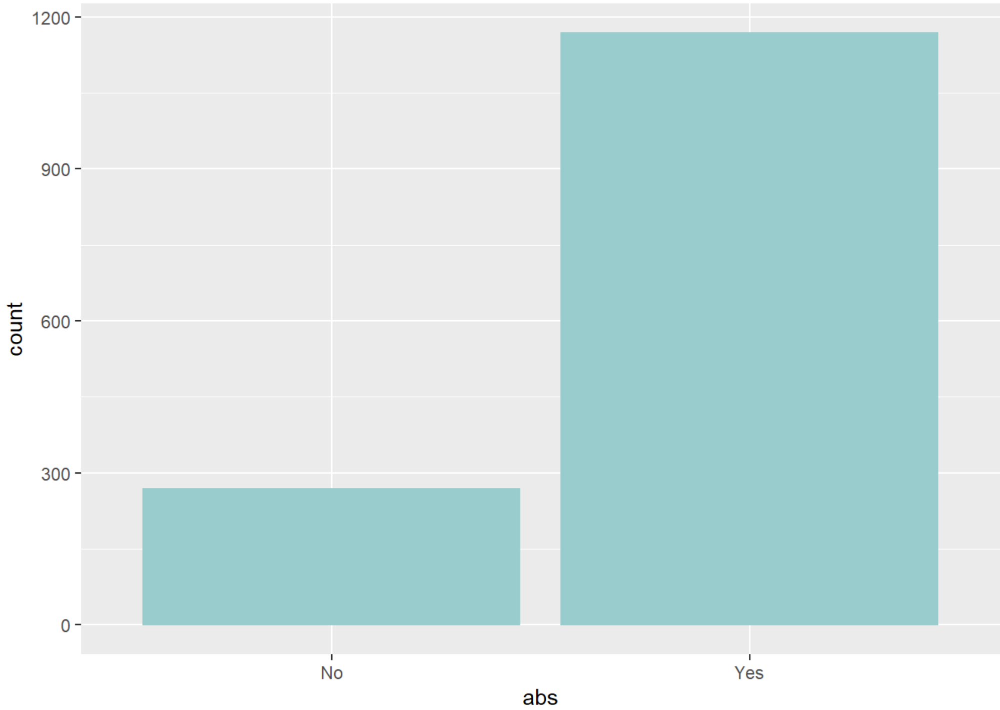
Anti-lock braking system, air-conditioning (Count)

```
toyotaC %>%
  count(abs, airco)
```

```
## # A tibble: 4 x 3
##   abs   airco     n
##   <fct> <fct> <int>
## 1 No    No      195
## 2 No    Yes      73
## 3 Yes   No      511
## 4 Yes   Yes     657
```

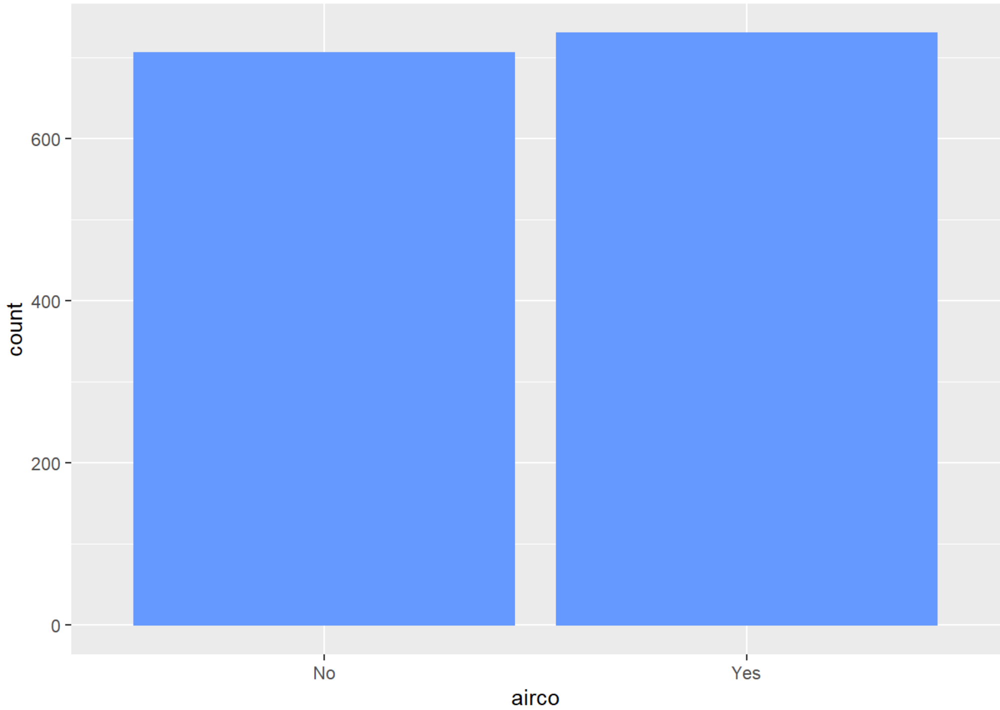
Anti-lock braking system (count) plot

```
ggplot(toyotaC, aes(abs)) + geom_bar(color="#99CCCC", fill="#99CCCC")
```



Air-conditioning (count)plot

```
ggplot(toyotaC, aes(airco)) + geom_bar(color="#6699FF", fill="#6699FF")
```



Model

Multiple Linear Regression

Model A

```
modA_mlr <- lm(price ~ age_08_04 + km + fuel_type +
               met_color + catA, data = toyotaA)
```

Model B

```
modB_mlr <- lm(price ~ age_08_04 + km + catB +
               automatic + cc, data = toyotaB)
```

Model C

```
modC_mlr <- lm(price ~ age_08_04 + km + hp +
                 abs + airco, data = toyotaC)
```

View data

Model A

```
summary(modA_mlr)
```

```
##
## Call:
## lm(formula = price ~ age_08_04 + km + fuel_type + met_color +
##     catA, data = toyotaA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10036.6   -876.1    -21.6    832.2   6979.6
##
## Coefficients:
##                              Estimate   Std. Error t value
## (Intercept)               19446.231741   404.915161  48.025
## age_08_04                  -143.812706     2.748241 -52.329
## km                           -0.017041     0.001498 -11.375
## fuel_typeDiesel            -706.963259   403.959699  -1.750
## fuel_typePetrol             300.913379   378.225017   0.796
## met_colorYes                114.711288    85.732438   1.338
## catAPassenger cars compact  2507.393703   198.662622  12.621
## catAPassenger cars medium  10036.881510   776.623944  12.924
## catAPassenger cars heavy     179.842045  1512.305834   0.119
##                                     Pr(>|t|)
## (Intercept)               <0.0000000000000002 ***
## age_08_04                 <0.0000000000000002 ***
## km                        <0.0000000000000002 ***
## fuel_typeDiesel                       0.0803 .
## fuel_typePetrol                       0.4264
## met_colorYes                          0.1811
## catAPassenger cars compact <0.0000000000000002 ***
## catAPassenger cars medium  <0.0000000000000002 ***
## catAPassenger cars heavy              0.9054
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1511 on 1427 degrees of freedom
## Multiple R-squared:  0.8275, Adjusted R-squared:  0.8265
## F-statistic: 855.5 on 8 and 1427 DF,  p-value: < 0.00000000000000022
```

Model B

```
summary(modB_mlr)
```

```
##
## Call:
## lm(formula = price ~ age_08_04 + km + catB + automatic + cc,
##     data = toyotaB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6831.0  -933.1   -45.5   821.3 12471.1
##
## Coefficients:
##                     Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)     19097.064819   457.604243  41.733 < 0.0000000000000002 ***
## age_08_04        -152.573627     2.764891 -55.183 < 0.0000000000000002 ***
## km                 -0.017015     0.001391 -12.233 < 0.0000000000000002 ***
## catBliquid fuel   428.160802   406.011221   1.055              0.29181
## automaticYes      619.096187   190.905670   3.243              0.00121 **
## cc                  0.557002     0.104656   5.322          0.000000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1640 on 1430 degrees of freedom
## Multiple R-squared:  0.7962, Adjusted R-squared:  0.7955
## F-statistic:  1117 on 5 and 1430 DF,  p-value: < 0.00000000000000022
```

Model C

```
summary(modC_mlr)
```

```
##
## Call:
## lm(formula = price ~ age_08_04 + km + hp + abs + airco, data = toyotaC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6136.9  -941.4   -93.6   796.3 12315.0
##
## Coefficients:
##                 Estimate   Std. Error t value             Pr(>|t|)
## (Intercept) 17097.220539   375.171981  45.572 < 0.0000000000000002 ***
## age_08_04    -153.411359     2.954263 -51.929 < 0.0000000000000002 ***
## km             -0.012474     0.001342  -9.294 < 0.0000000000000002 ***
## hp             32.430025     2.980494  10.881 < 0.0000000000000002 ***
## absYes       -621.560341   115.599135  -5.377        0.00000008842 ***
## aircoYes      561.132430    92.574712   6.061        0.00000000172 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1549 on 1430 degrees of freedom
## Multiple R-squared:  0.8181, Adjusted R-squared:  0.8175
## F-statistic:  1286 on 5 and 1430 DF,  p-value: < 0.00000000000000022
```

View tidied result

Model A

```
tidy(modA_mlr)
```

```
## # A tibble: 9 x 5
##   term                         estimate  std.error statistic   p.value
##   <chr>                           <dbl>      <dbl>     <dbl>     <dbl>
## 1 (Intercept)                19446.      405.          48.0   2.57e-300
## 2 age_08_04                   -144.        2.75       -52.3   0
## 3 km                            -0.0170    0.00150    -11.4   9.22e- 29
## 4 fuel_typeDiesel             -707.      404.          -1.75  8.03e-  2
## 5 fuel_typePetrol              301.      378.           0.796 4.26e-  1
## 6 met_colorYes                 115.       85.7          1.34  1.81e-  1
## 7 catAPassenger cars compact  2507.      199.          12.6   1.07e- 34
## 8 catAPassenger cars medium  10037.      777.          12.9   3.26e- 36
## 9 catAPassenger cars heavy     180.     1512.           0.119 9.05e-  1
```

Model B

```
tidy(modB_mlr)
```

```
## # A tibble: 6 x 5
##   term             estimate std.error statistic   p.value
##   <chr>               <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     19097.     458.          41.7  1.27e-249
## 2 age_08_04        -153.       2.76       -55.2  0
## 3 km                 -0.0170   0.00139    -12.2  8.51e- 33
## 4 catBliquid fuel   428.     406.           1.05 2.92e-  1
## 5 automaticYes      619.     191.           3.24 1.21e-  3
## 6 cc                  0.557    0.105        5.32 1.19e-  7
```

Model C

```
tidy(modC_mlr)
```

```
## # A tibble: 6 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 17097.     375.          45.6  7.83e-281
## 2 age_08_04    -153.       2.95       -51.9  0
## 3 km             -0.0125   0.00134     -9.29 5.34e- 20
## 4 hp             32.4      2.98        10.9  1.51e- 26
## 5 absYes       -622.     116.          -5.38 8.84e-  8
## 6 aircoYes      561.      92.6          6.06 1.72e-  9
```

Generate predicted and residuals value

Model A

```
modA_mlr_fit <- augment(modA_mlr)
modA_mlr_fit
```

```
## # A tibble: 1,436 x 12
##    price age_08_04    km fuel_type met_color catA  .fitted .resid    .hat .sigma
##    <dbl>     <dbl> <dbl> <chr>     <fct>     <fct>   <dbl>  <dbl>   <dbl>  <dbl>
##  1 13500        23 46986 Diesel    Yes       Pass~  17253. -3753. 0.0113   1508.
##  2 13750        23 72937 Diesel    Yes       Pass~  16811. -3061. 0.0103   1509.
##  3 13950        24 41711 Diesel    Yes       Pass~  17199. -3249. 0.0116   1509.
##  4 14950        26 48000 Diesel    No        Pass~  16690. -1740. 0.0128   1511.
##  5 13750        30 38500 Diesel    No        Pass~  16276. -2526. 0.0136   1510.
##  6 12950        32 61000 Diesel    No        Pass~  15605. -2655. 0.0119   1510.
##  7 16900        27 94612 Diesel    Yes       Pass~  15866.  1034. 0.0101   1511.
##  8 18600        30 75889 Diesel    Yes       Pass~  15754.  2846. 0.00989  1509.
##  9 21500        27 19700 Petrol    No        Pass~  18036.  3464. 0.0186   1508.
## 10 12950        23 71138 Diesel    No        Pass~  14219. -1269. 0.0181   1511.
## # ... with 1,426 more rows, and 2 more variables: .cooksd <dbl>,
## #   .std.resid <dbl>
```

Model B

```
modB_mlr_fit <- augment(modB_mlr)
modB_mlr_fit
```

```
## # A tibble: 1,436 x 12
##     price age_08_04    km catB      automatic    cc .fitted .resid    .hat .sigma
##     <dbl>     <dbl> <dbl> <chr>     <fct>     <dbl>   <dbl>  <dbl>   <dbl>  <dbl>
##   1 13500        23 46986 liquid f~ No         2000  16331. -2831. 0.00344  1639.
##   2 13750        23 72937 liquid f~ No         2000  15889. -2139. 0.00419  1640.
##   3 13950        24 41711 liquid f~ No         2000  16268. -2318. 0.00329  1640.
##   4 14950        26 48000 liquid f~ No         2000  15856.  -906. 0.00307  1641.
##   5 13750        30 38500 liquid f~ No         2000  15407. -1657. 0.00269  1640.
##   6 12950        32 61000 liquid f~ No         2000  14719. -1769. 0.00254  1640.
##   7 16900        27 94612 liquid f~ No         2000  14910.  1990. 0.00476  1640.
##   8 18600        30 75889 liquid f~ No         2000  14771.  3829. 0.00321  1638.
##   9 21500        27 19700 liquid f~ No         1800  16073.  5427. 0.00287  1634.
## 10 12950        23 71138 liquid f~ No         1900  15864. -2914. 0.00393  1639.
## # ... with 1,426 more rows, and 2 more variables: .cooksd <dbl>,
## #   .std.resid <dbl>
```

Model C

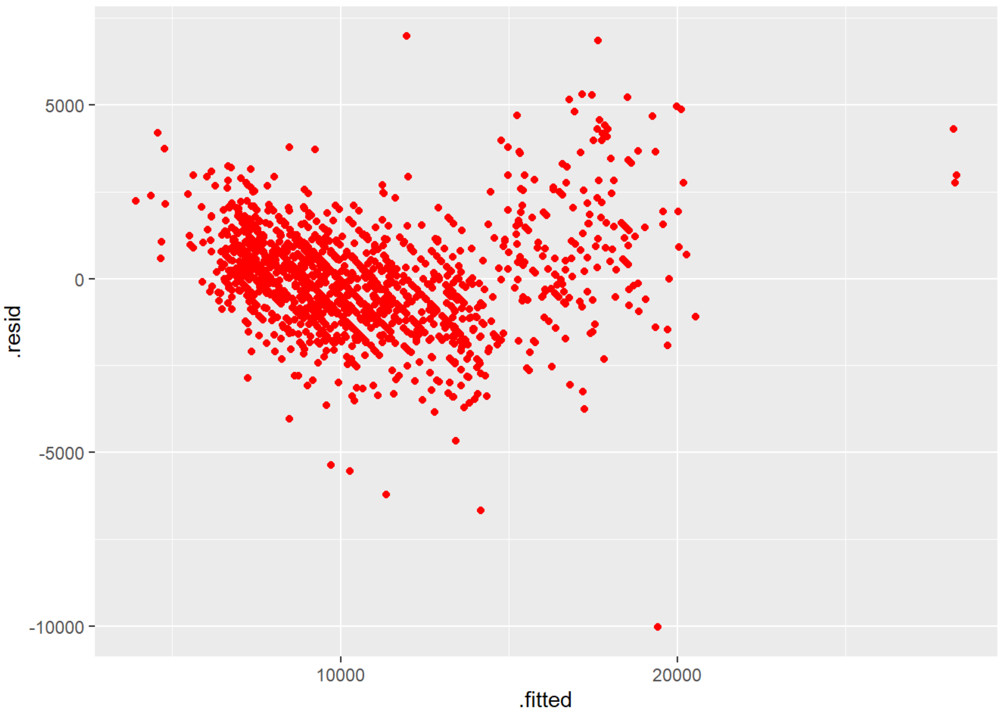
```
modC_mlr_fit <- augment(modC_mlr)
modC_mlr_fit
```

```
## # A tibble: 1,436 x 12
##     price age_08_04    km    hp abs   airco .fitted .resid    .hat .sigma .cooksd
##     <dbl>     <dbl> <dbl> <dbl> <fct> <fct>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>
##   1 13500        23 46986    90 Yes   No     15280. -1780. 0.00582  1549. 1.29e-3
##   2 13750        23 72937    90 Yes   Yes    15517. -1767. 0.00441  1549. 9.64e-4
##   3 13950        24 41711    90 Yes   No     15192. -1242. 0.00555  1550. 6.01e-4
##   4 14950        26 48000    90 Yes   No     14807.   143. 0.00520  1550. 7.47e-6
##   5 13750        30 38500    90 Yes   Yes    14873. -1123. 0.00328  1550. 2.89e-4
##   6 12950        32 61000    90 Yes   Yes    14285. -1335. 0.00290  1550. 3.61e-4
##   7 16900        27 94612    90 Yes   Yes    14633.  2267. 0.00479  1549. 1.72e-3
##   8 18600        30 75889    90 Yes   Yes    14407.  4193. 0.00341  1546. 4.19e-3
##   9 21500        27 19700   192 Yes   Yes    18876.  2624. 0.0281   1548. 1.42e-2
## 10 12950        23 71138    69 Yes   Yes    14859. -1909. 0.00775  1549. 1.99e-3
## # ... with 1,426 more rows, and 1 more variable: .std.resid <dbl>
```

Assess model

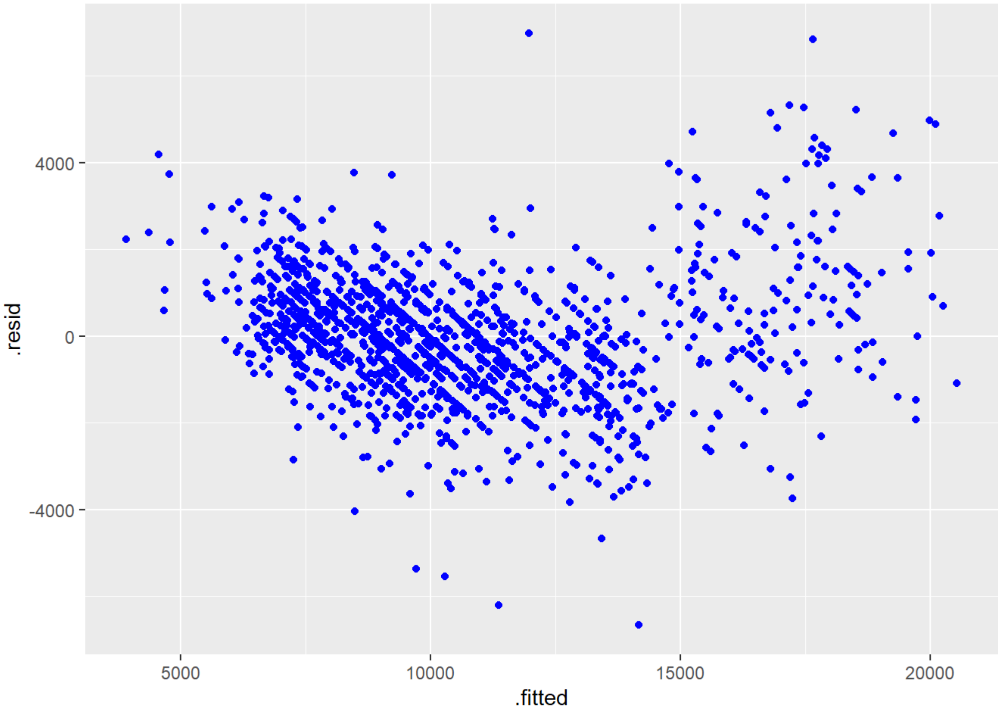
Model A

```
ggplot(modA_mlr_fit, aes(x = .fitted, y = .resid)) +
  geom_point(color="red")
```



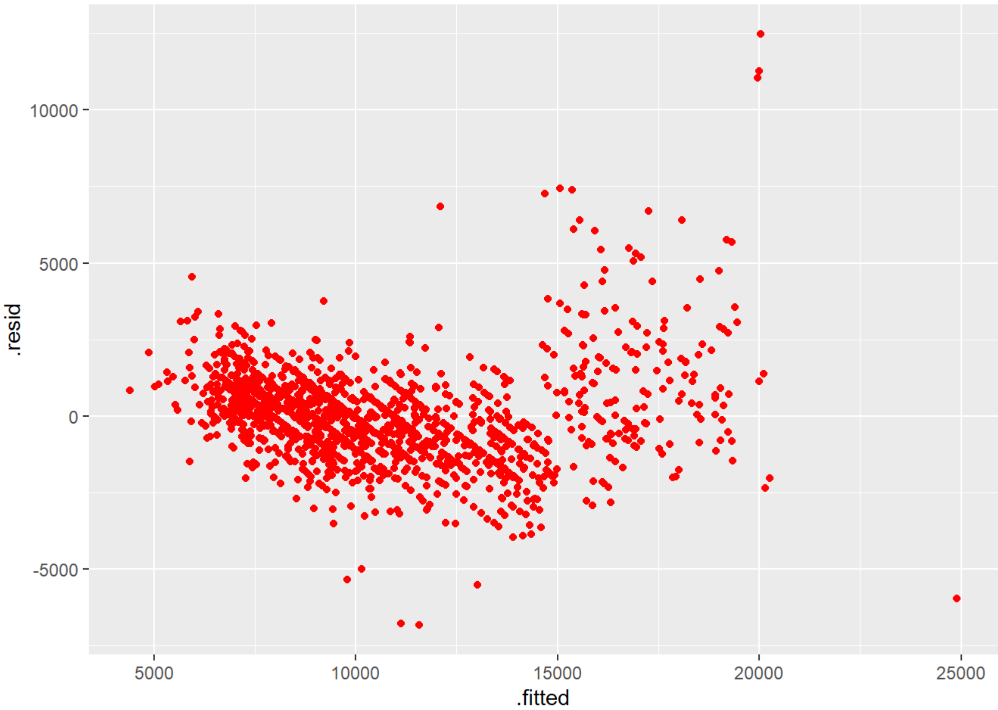
```
modA_mlr_fit2 <- modA_mlr_fit %>%
  filter(.resid > -7500 ,
         .fitted < 25000)

ggplot(modA_mlr_fit2, aes(x = .fitted, y = .resid)) +
  geom_point(color="#0000FF")
```



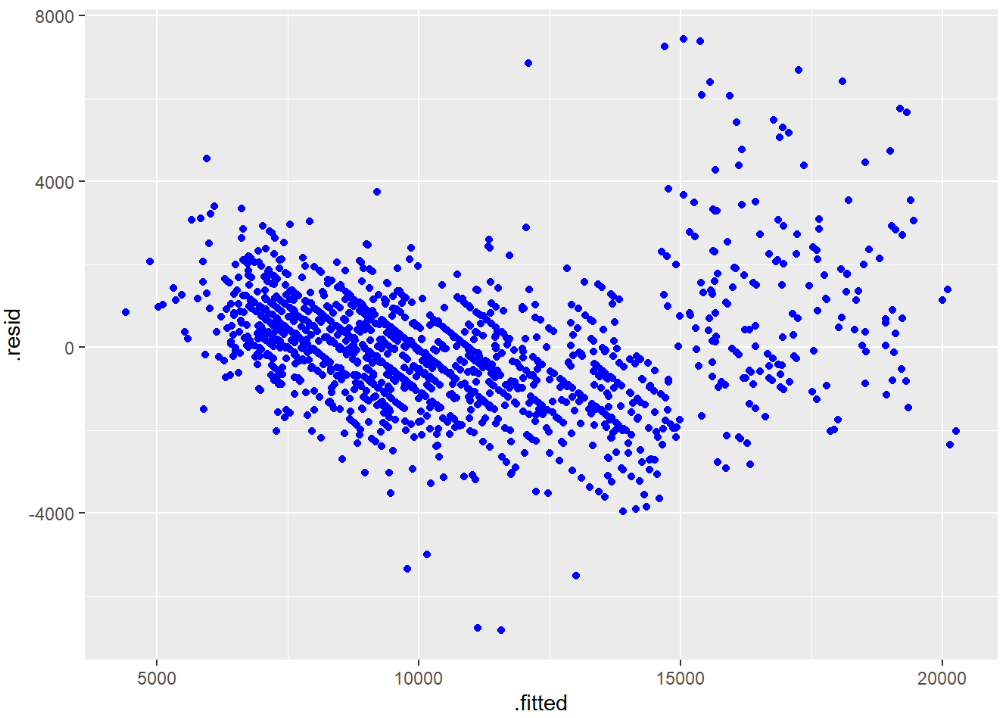
Model B

```
ggplot(modB_mlr_fit, aes(x = .fitted, y = .resid)) +
  geom_point(color="red")
```




```
modB_mlr_fit2 <- modB_mlr_fit %>%
  filter(.resid < 10000 ,
         .fitted < 22500)

ggplot(modB_mlr_fit2, aes(x = .fitted, y = .resid)) +
  geom_point(color="#0000FF")
```



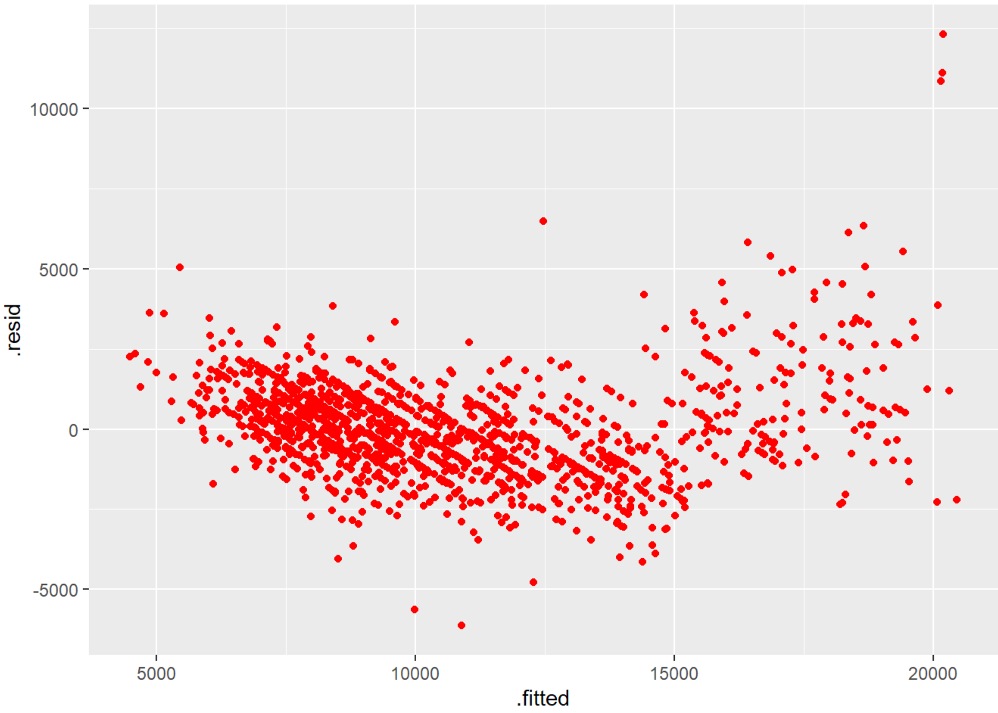
Re- estimate multiple linear regression with fited data (Outliers are removed due to erroneous data i.e. cc=16000)

```
modB_mlr2 <- lm(price ~ age_08_04 + km + catB +
               automatic + cc, data = modB_mlr_fit2)
tidy(modB_mlr2)
```

```
## # A tibble: 6 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      16128.      538.         30.0   1.14e-153
## 2 age_08_04         -143.        2.68      -53.2   0
## 3 km                  -0.0213   0.00140    -15.2   1.95e- 48
## 4 catBliquid fuel    261.      375.          0.696 4.86e-  1
## 5 automaticYes       739.      177.          4.19  3.02e-  5
## 6 cc                   2.38      0.241       9.87  2.92e- 22
```

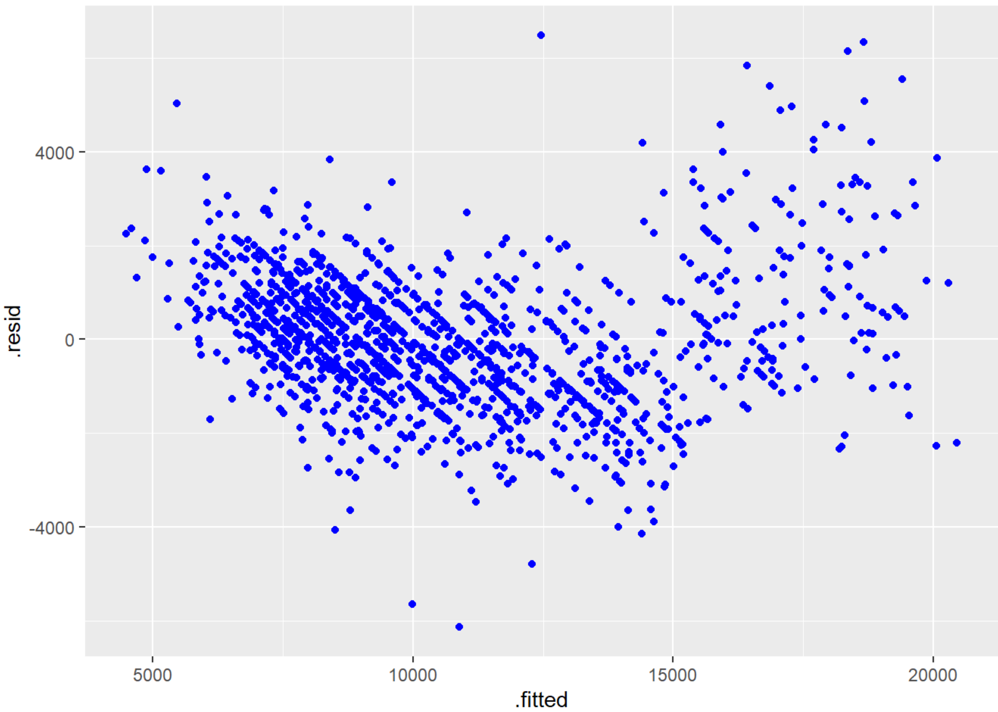
Model C

```
ggplot(modC_mlr_fit, aes(x = .fitted, y = .resid)) +
  geom_point(color="red")
```




```
modC_mlr_fit2 <- modC_mlr_fit %>%
  filter(.resid < 10000,
         .fitted < 25000)

ggplot(modC_mlr_fit2, aes(x = .fitted, y = .resid)) +
  geom_point(color="#0000FF")
```



Re- estimate multiple linear regression with fited data (Outliers are removed)

```
modC_mlr2 <- lm(price ~ age_08_04 + km + hp +
                abs + airco, data = modC_mlr_fit2)
tidy(modC_mlr2)
```

```
## # A tibble: 6 x 5
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 16881.     353.          47.8  2.31e-298
## 2 age_08_04    -150.       2.79       -53.5  0
## 3 km             -0.0123   0.00126     -9.74 9.48e- 22
## 4 hp             31.6      2.80        11.3  2.76e- 28
## 5 absYes       -572.     109.          -5.26 1.65e-  7
## 6 aircoYes      571.      87.1          6.56 7.73e- 11
```

Present linear regression result

Model A

```
options(scipen = 999)
modA_model <- tidy(modA_mlr, conf.int = TRUE)
kable(modA_model) %>%
  kable_styling()
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	19446.2317413	404.915161	48.0254473	0.0000000	18651.9389089	20240.5245736
age_08_04	-143.8127062	2.748241	-52.3289973	0.0000000	-149.2037325	-138.4216799
km	-0.0170407	0.001498	-11.3753599	0.0000000	-0.0199792	-0.0141021
fuel_typeDiesel	-706.9632586	403.959699	-1.7500836	0.0803188	-1499.3818310	85.4553139
fuel_typePetrol	300.9133791	378.225017	0.7955935	0.4264006	-441.0233253	1042.8500834
met_colorYes	114.7112879	85.732438	1.3380150	0.1811047	-53.4638441	282.8864198
catAPassenger cars compact	2507.3937026	198.662622	12.6213662	0.0000000	2117.6915828	2897.0958224
catAPassenger cars medium	10036.8815097	776.623944	12.9237343	0.0000000	8513.4343986	11560.3286209
catAPassenger cars heavy	179.8420449	1512.305834	0.1189191	0.9053562	-2786.7391052	3146.4231951

Model B

```
options(scipen = 999)
modB_model <- tidy(modB_mlr2, conf.int = TRUE)
kable(modB_model) %>%
  kable_styling()
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	16127.6569072	537.6583198	29.996108	0.0000000	15072.9707777	17182.3430367
age_08_04	-142.7366303	2.6829590	-53.201196	0.0000000	-147.9996004	-137.4736601
km	-0.0212813	0.0014007	-15.193761	0.0000000	-0.0240289	-0.0185337
catBliquid fuel	261.2973994	375.3302526	0.696180	0.4864296	-474.9612927	997.5560916
automaticYes	739.0471959	176.5709165	4.185555	0.0000302	392.6805731	1085.4138187
cc	2.3764299	0.2408329	9.867545	0.0000000	1.9040050	2.8488548

Model C

```
options(scipen = 999)
modC_model <- tidy(modC_mlr2, conf.int = TRUE)
kable(modC_model) %>%
  kable_styling()
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	16880.5120124	353.2920522	47.780616	0.0000000	16187.4845047	17573.5395200
age_08_04	-149.5008490	2.7937862	-53.511915	0.0000000	-154.9812176	-144.0204804
km	-0.0122986	0.0012626	-9.740461	0.0000000	-0.0147754	-0.0098218
hp	31.6087746	2.8045119	11.270687	0.0000000	26.1073661	37.1101831
absYes	-572.4955824	108.8075098	-5.261545	0.0000002	-785.9354174	-359.0557473
aircoYes	570.9249728	87.0916729	6.555448	0.0000000	400.0835269	741.7664188

Discussion

Communicate

Section 5

Process of model assessment

The process to assess the multiple linear regression is to use the ggplot2 package to plot a residual vs fitted plot to identify the outliers. For this, we use the estimation linear regression model which will be fitted in scatter plot. There is a linear relationship between the residual (observed price of vehicle) and the fitted (predicted price of the vehicle). Any outlier that do not confirm with the linearity will be further analysed whether they actually are outliers. After deciding, the outliers is either retained or removed by range, plotting another fitted vs residual value that now conforms on linearity between the observed price and predicted price.

Comment of the findings

In model A, due to a p-value of more than the common alpha level of 0.05, it is indicative that predictor variables 'fuel_typeDiesel', 'fuel_typePetrol', 'met_colorYes', 'catAPassenger cars heavy' are not statistically significant, therefore the change in these predictor variable has no effect on outcome variable; Price of vehicle. Predictor variable vehicle age in month and mileage has negative affect on price, where for every 1 month increase in age, the price reduces by €144 and similarly for every increase of 1km mileage, the price is reduced by €0.02. Compact and medium passenger car is €2507 and €10037 more expensive respectively compared to light passenger car. In model B, similar results were obtained for predictor variables vehicle age in month from august 2004 and mileage. P-value for liquid fuel vehicle type is higher than common alpha value 0.05 hence deemed to have no effect on price. Automatic transmission vehicle price is €739 higher than manual transmission vehicle price. For ever 1 cubic capacity increase in engine size, €2.40 increase in vehicle price is detected. For model C, besides age in month and mileage predictor variables which have almost similar effect as in above models, ever 1hp increase in horsepower, the vehicle price goes up by €32 and vehicles with air-conditioning is €571 more expensive than vehicle without air-conditioning. Vehicles with anti-lock braking system however is €572 cheaper than vehicle with ABS, which is a questionable output, that could be due to metadata error.

Suggested measures to improve

By adding interaction plots between all 5 of the predictor variables to the outcome variable to understand possible interactions among variables (Burrill, 1997). Also checking for multicollinearity will help in reducing unreliability of the linear regression model.

(385 words)

Reference

Burrill, D. F. (1997). Modeling and interpreting interactions in multiple regression. The Ontario Institute for Studies in Education Toronto, Ontario Canada.

Englmaier, F., Schmöller, A., & Stowasser, T. (2013). Price discontinuities in an online used car market.

Greim, L. (2017). Impact of the model life cycle on the residual car value in the leasing industry (Master's thesis, University of Twente).

National Research Council. (1992). Automotive fuel economy: How far can we go?. National Academies Press.

Schnitman, R. A Short Introduction to Applied Statistical Programming in R.

Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106, 234-240.

Worster, A., Fan, J., & Ismaila, A. (2007). Understanding linear and logistic regression analyses. Canadian Journal of Emergency Medicine, 9(2), 111-113.