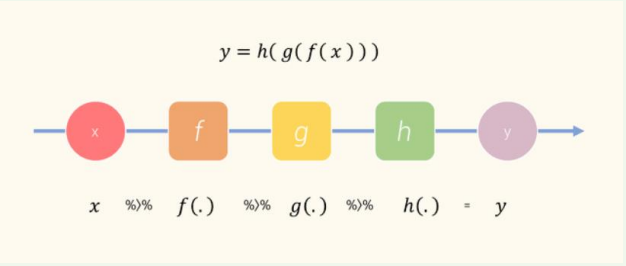
클린업 1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 클린업 1주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다. 그러나 이번 주차의 경우 최대한 R을 사용하는 것을 권장드립니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **21일 목요일 23:59까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실 시도 퇴출이니 유의해주세요.

Chapter 1. EDA, Data Preprocessing

데이터를 분석하면서 가장 많이 사용하는 패키지는 tidyverse입니다. 해당 패키지는 tidyr, dplyr, ggplot 등 데이터를 전처리하고 시각화하는 과정에서 많이 사용되는 패키지들로 구성되어 있습니다. 그 중에서도 특히 데이터 전처리 과정에서 많이 사용되는 tidyr, ggplot2 패키지를 이용하여 데이터를 효율적으로 파악하고, 모델의 가정에 맞추어 전처리하는 과정을 연습해보겠습니다.

데이터를 정제하는 과정에서 pipe 연산자(%>%)를 활용하면 코드를 간결하고 깔끔하게 짤 수 있습니다. 해당 연산자는 ctrl+shift+M을 통해 불러올 수 있으며, 아래의 그림처럼 데이터 흐름을 왼쪽에서 오른쪽으로 흐르도록 하여 직관적으로 파악할 수 있게 하는 장점이 있습니다. 해당 연산자를 최대한 활용하여 익숙해져 봅시다.



[Chapter1 조건 : %>% 연산자를 최대한 활용하여 보기 좋은 코드 작성]

문제0. 메모장의 코드를 실행하여 for_week1.csv를 불러온 뒤 데이터 구조 및 변수들을 파악해보세요.

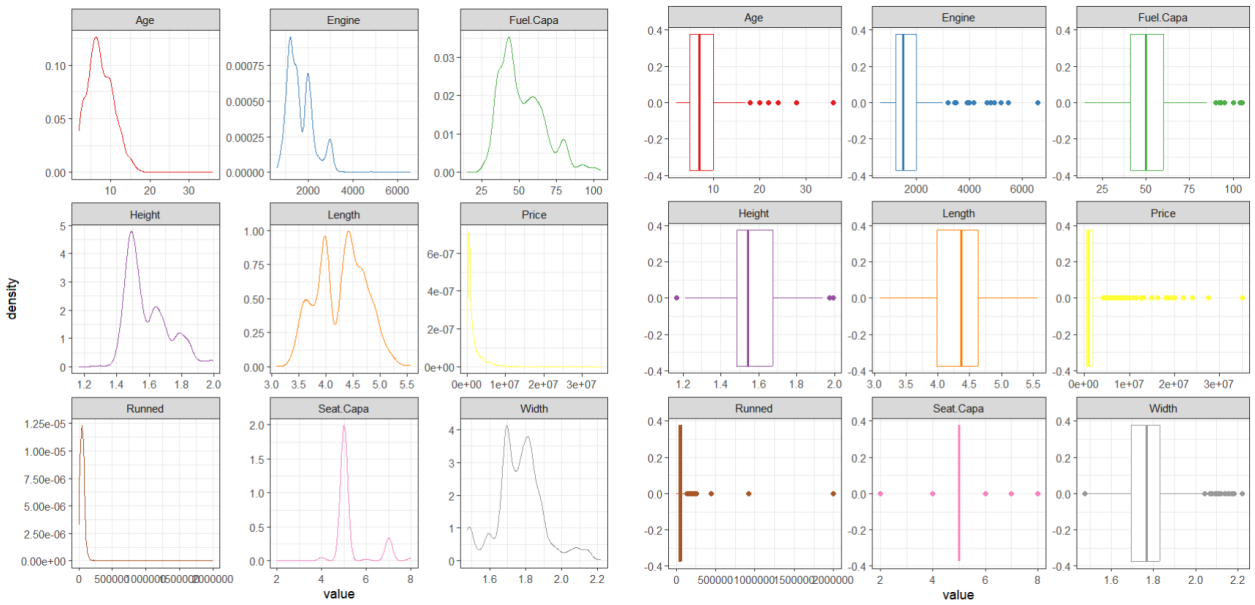
(HINT) head, tail, str, glimpse, summary 등 데이터의 구조를 파악하기 위한 다양한 함수가 있습니다

문제1. 각 변수마다 NA(결측치) 개수와 unique한 값의 개수를 파악해보세요.

(HINT) colSums, is.na, n_distinct 등을 사용하면 편리합니다. 그 외에도 다양한 방법이 있습니다.

문제2. 메모장 아래에 있는 코드를 실행하여, Color 별 Price의 분포를 확인해주세요. 색깔(Color) 별로 가격(Price)에 차이가 존재한다고 말할 수 있을지 생각해본 후, 의견을 작성해주세요. (정답은 없습니다)

문제3. 수치형 변수에 대해 ggplot으로 다음과 같이 density plot과 boxplot을 그려주세요. 그 후 결과에 대해 간단히 해석해보세요. (난이도가 있는 작업입니다.)



(HINT1) 각 변수 별 분포를 ggplot으로 한 눈에 시각화하고 싶을 때, R에서는 보통 gather() 함수를 이용해 데이터를 long type으로 변환한 후 plot을 그리게 되며, facet_wrap()을 이용하여 여러 플롯을 한 번에 그릴 수 있습니다. (scales = "free")

test_table

	country	1999	2000
1	A	0.7K	2K
2	B	37K	80K
3	C	212K	213K

→

	country	year	cases
1	A	1999	0.7K
2	B	1999	37K
3	C	1999	212K
4	A	2000	2K
5	B	2000	80K
6	C	2000	213K

```
gather(data = test_table, key = "year",
       value = "cases", `1999`,`2000`)
```

(HINT2) 범례는 나타나지 않도록 하고, palette = "Set1"을 이용합니다. 그 외 여러 가지 옵션을 변경해보며 최대한 비슷하게 만들어주세요.

문제4. 앞에서 파악한 데이터를 바탕으로 간단한 전처리를 시행해보겠습니다.

문제 4-1. 데이터에서 정수형(integer) 변수를 수치형(numeric) 변수로 바꿔주세요.

문제 4-2. Engine 변수에 대하여, 뒤에 있는 글자 'cc'를 뺀 후 수치형 변수로 바꿔주세요.

문제 4-3. Year 변수에 대하여, `2024 - Year` 의 값으로 Age라는 파생변수를 생성해주세요. 이후 Year 변수는 삭제해주세요.

(HINT) 파생 변수를 만들 때는 mutate() 함수를 이용하면 편리합니다.

문제 4-4. 결측치가 있는 열에 대해, 결측치를 각 열 별 평균값으로 대체해주세요.

(보너스문제) 평균으로 대체하는 것이 괜찮을지, 그렇지 않다면 각 변수 별로 어떻게 보간하는 것이 괜찮을지 문제1, 문제3 등에서 파악한 변수들의 특징을 바탕으로 의견을 간단하게 작성해주세요.

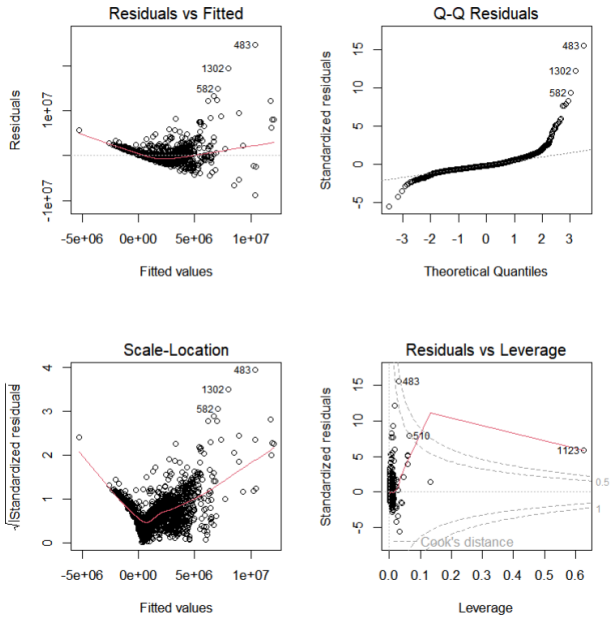
Chapter 2. Statistical Modeling, Issues of Linear Regression

리드오프에서, 간단한 선형회귀분석이라도 그 모델의 가정에 맞추어 적절히 모델링을 하면 더 좋은 결과를 낼 수 있음을 말씀드렸습니다. 따라서 회귀분석 모델링을 통해 전반적인 데이터 분석 과정에 대해 살펴보고, EDA 및 정확한 모델의 이해에 기반한 통계적 분석이 얼마나 중요한지 직접 알아보는 시간을 갖도록 하겠습니다.

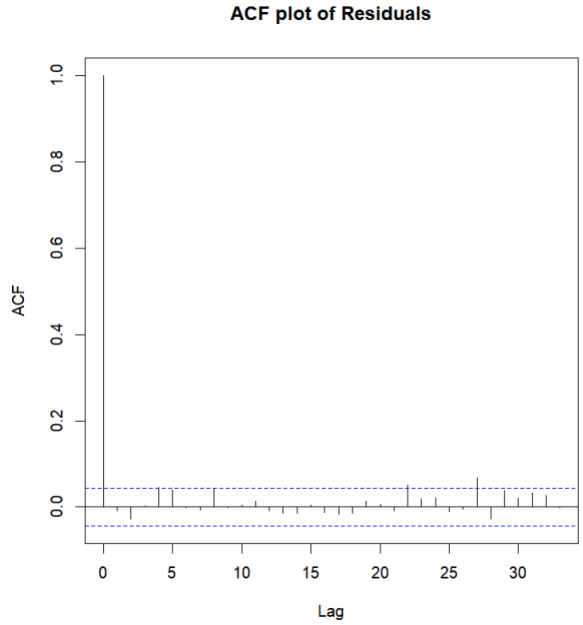
문제1. Chapter 1에서 전처리를 완료한 데이터를 이용할 예정입니다. 그 데이터에서, 수치형 변수만 남겨 num_data를 만들어주세요. (물론 범주형 변수도 모델에 추가하면 더 좋은 성능을 기대할 수 있을 것입니다. 이에 관한 내용은 다른 주차의 패키지과제에서 다뤄보겠습니다.)

문제2. Price를 Y(종속변수)로 하여, 선형회귀모델을 적합해주세요. summary()로 결과를 확인해주세요.

문제3. 아래 그림처럼 residual plot을 출력해주세요. 이후 결과를 간단히 해석해주세요.

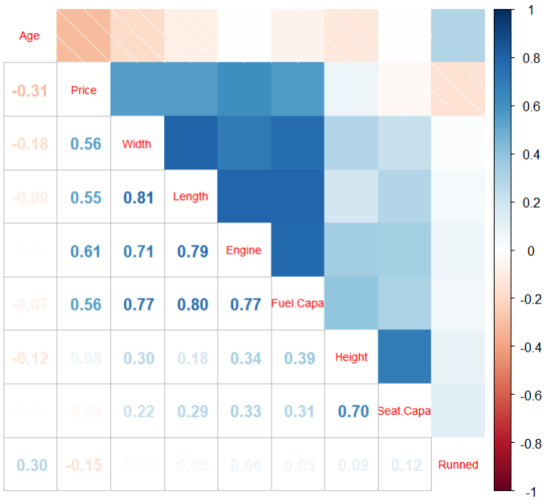


문제4. residual에 대한 ACF plot을 그려주세요. 이후 결과를 간단히 해석해주세요.



(HINT) residuals(model)로 잔차를 추출할 수 있으며, acf()로 ACF plot을 그릴 수 있습니다.

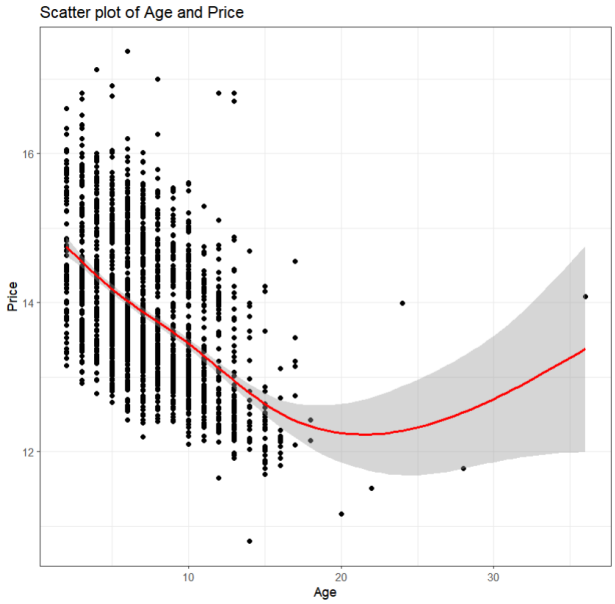
문제5. 수치형 변수에 대해 상관계수 히트맵을 그려주세요.



(HINT) corrplot.mixed 함수를 이용, order="AOE", tl.cex = 0.7 등의 옵션을 추가했습니다. 그 외 여러 가지 옵션을 변경해 가며 최대한 비슷하게 만들어주세요.

문제6. vif() 함수로 각 변수에 대한 VIF를 확인해주세요. 문제5.의 결과와 함께 간단히 해석해주세요.

문제7. Age, Height 변수의 Price에 대한 Scatter plot과 추세선을 그려주세요.



(HINT) ggplot의 geom_smooth를 통해 추세선을 그릴 수 있으며, theme_bw를 사용합니다. method = "gam", color = 'red'를 사용했습니다.

문제8. Age_squared(Age^2) 과 Height_squared(Height^2), Height_cubed(Height^3)을 설명변수로 추가하여 다시 회귀모델을 적합한 후, 결과를 확인해주세요.

(HINT) 파생변수를 만들 때에는 마찬가지로 mutate()를 사용합니다.

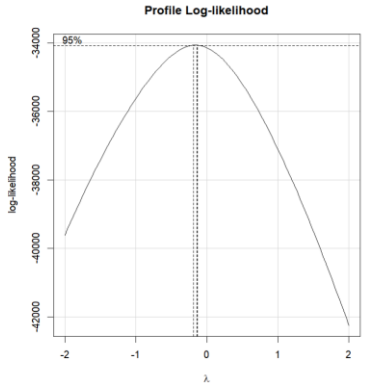
문제8-1. 추가한 파생변수를 이용할 것인지 이용하지 않을 것인지 의견을 1~2줄로 작성해주세요.

문제9. Price 변수에 대해 Box-cox 변환을 실시해보겠습니다.

문제 9-1. Box-cox 변환이 무엇인지 살펴본 후, 간단하게 작성해주세요.

(BONUS) λ에 대한 likelihood가 무엇을 의미하는지에 대해서도 포함하여 작성해주세요.

문제 9-2. λ에 대한 Log-likelihood plot을 그린 후, 최적의 λ를 찾아주세요.

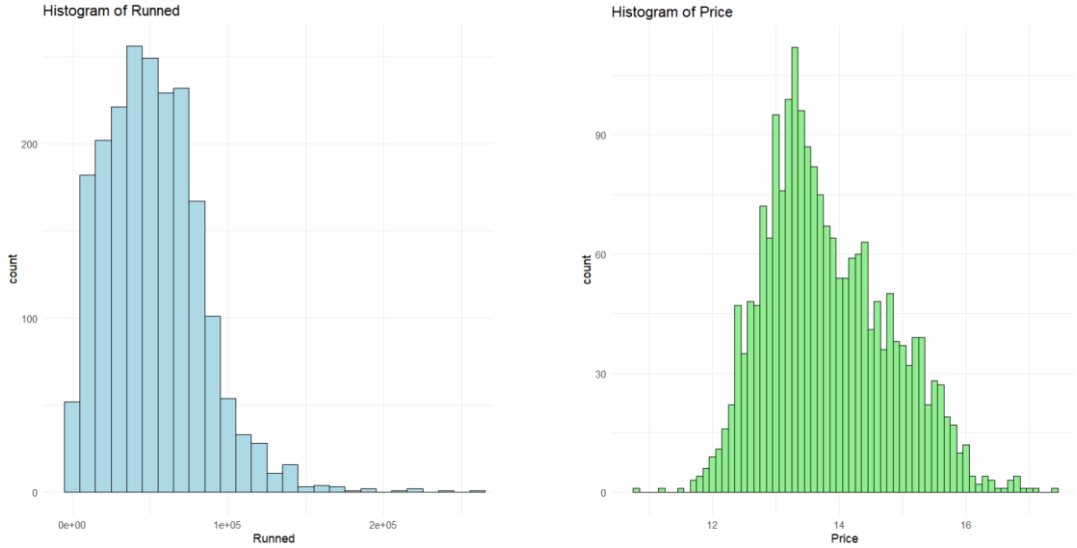


(HINT) model에 대해, boxCox() 함수를 이용해 Log-likelihood plot을 그릴 수 있습니다. 또 powerTransform() 함수를 이용하여 best likelihood를 갖게 하는 λ를 찾을 수 있습니다.

문제 9-3. 최적의 λ가 0에 가까우므로, 로그 변환을 실시하겠습니다. Price 변수를 log변환 해주세요.

문제10. leverage 값이 높은 1123행의 데이터를 확인해주세요. Runned의 Outlier가 회귀 모델링에 영향을 주고 있다고 판단할 수 있으므로, Runned의 값이 300,000을 넘는 행은 삭제해주세요.

문제11. 데이터 수정(변환) 후의 Runned와 Price열의 histogram을 그려주세요.



(HINT) ggplot의 geom_histogram을 이용하며, 각각 (binwidth = 10000, fill = "lightblue", color = "black"), (binwidth = 0.1, fill = "lightgreen", color = "black") 등의 옵션이 추가됩니다.

문제12. 전처리가 완료된 데이터에 대해 다시 문제2, 문제3을 수행해주세요. 처음의 모델 적합 결과와 비교해 주세요.

문제13. Confounding Variable이 무엇인지 살펴보고, 그 개념과 영향에 대해 간단히 설명해주세요.

문제13-1. Price를 Y(종속변수)로, Seat.Capa를 X(설명변수)로 하여 단순 선형회귀모델을 적합시킨 후, summary로 결과를 확인해주세요.

문제13-2. Price를 Y(종속변수)로, Seat.Capa와 Length를 X(설명변수)로 하여 다중 선형회귀모델을 적합시킨 후, summary로 결과를 확인해주세요.

문제13-3. 위 결과로부터 추론할 수 있는 것을 간단하게 설명해주세요.