

2021 YDMS 5기 모집과제

정보통계학과 2019251034 김하은

<1. 본인이 생각하는 EDA란 어떠한 것인지 서술하시오.>

우선 EDA가 무엇인지에 대해 자세히 알아보기 앞서, 용어가 대강 어떤 의미를 내포하고 있는지에 대해 생각해보았다. EDA란 Exploratory Data Analysis의 약어로 ‘탐색적 자료 분석’을 뜻한다. 탐색이란, 드러나지 않은 (사물이나) 현상 따위를 찾아내거나 밝히기 위하여 살피어 찾는다는 의미이다. 그러므로 탐색적 자료 분석이란 자료의 숨겨진 정보 및 데이터 분포에 대한 패턴을 찾아내기 위해, 자료를 살피어 분석하는 방식이라고 생각하였다.

EDA를 창안한 통계학자 존 튜키에 따르면, EDA는 가설 검정이 주목적인 기존 통계학과는 달리 데이터 속 정보를 파악하는 것에 주목적을 둔 분석 방식으로, 주어진 자료를 가지고 정보를 찾아내는 방식이다.

다시 말해, EDA는 가설 검정에 쏠려 자료의 본질을 찾기 힘들었던 기존 통계학의 문제점을 보완하는 방법이기도, 주어진 자료를 탐색하여 숨겨진 자료 본연의 정보를 찾는 데에 집중한다. 이때 EDA는 다양한 각도로 데이터를 탐색함으로써 데이터에 대한 더 깊은 이해를 도울 수 있다. 또한 숨겨진 자료 속 문제점을 발견하게 되는 경우, 문제점을 수정하여 재탐색하는 과정이 가능하다는 점은 EDA의 큰 장점이다.

한편 EDA를 통해 자료 본연의 의미를 정확하게 파악하기 위해서는 변수에 대한 이해가 반드시 되어있어야 한다. 따라서 의미있는 정보를 얻기 위해선 변수끼리의 조합을 다양하게 고려해보는 것은 물론 변수끼리의 관계가 잘 맺어진 것인지도 반드시 확인해보아야 한다. 그래야만 EDA를 통해 자료 속 의미있는 패턴과 정보를 찾아낼 수 있기 때문이다. 만약 변수 유형을 무시한 채 진행하게 되면 전혀 다른 내용의 정보가 나올 수 있기 때문에 변수 유형에 주의해야 한다.

마지막으로 EDA의 과정은 다음과 같다. 우선 자료 속 변수에 대한 이해를 바탕으로 시작되어야 한다. 자료의 관측값들을 확인해보는 것도 좋고, 어떠한 형태인지, 변수 유형은 어느 쪽에 속하는지, 변수가 의미하는 것은 무엇인지에 대해 파악하는 것이 여기에 해당된다. 이후 자료를 살피면서 이상치나 결측값과 같은 문제점이 없는지 따져가며 진행해야 하고, 문제점이 있는 경우 자료를 수정한 뒤 탐색하여도 좋다. 이후 자료를 시각화하여 각 변수값들이 어떠한 패턴을 따르는지, 각 변수 간의 관계는 어떻게 되는지에 대해 파악하며 탐색을 진행한다. 이 과정을 통해 데이터 속 숨겨진 패턴이나 정보를 알아낼 수 있으며, 그에 대한 해석하는 과정까지 EDA에 속한다고 할 수 있다.

<2. 기초통계량을 통한 데이터 탐색과 데이터 시각화를 통해 결과를 자신의 언어로 서술하시오. (문제점이 있을 시 개인의 판단 아래 분석 진행)>

우선 T1과 T2의 데이터의 경우, 1000개가 넘는 객체를 가진 자료였기 때문에 우선적으로 두 개의 테이블을 먼저 비교해보았다. 두 자료는 공통변수 ID를 가지고 있었고, 나머지는 다른 변수로 이루어져 있었다. 각각의 자료의 경우 T1은 1438개의 관측치가 있으며, T2의 경우 1437개의 관측치가 있었다. ID를 공통변수로 가지고 있었기 때문에 T1과 T2는 ID로 이어질 것이라 생각하였는데 관측치의 수가 같지 않아 의문이 들었다. 이 때문에 가장 먼저 T1과 T2에서 중복값이나 결측치가 있는지 확인해보았다.

먼저 T1에서 2개의 중복값을 확인하였고, 같은 방식으로 T2에서 2개의 중복값을 확인하였다. T1과 T2 모두에서 중복값이 2개 나온 탓에 관측치 개수는 여전히 맞지 않았지만, T2를 ID 기준으로 오름차순 정리를 해보니 마지막 ID가 T1과 동일함을 확인할 수 있었다. 따라서 자료 중간에 T1에는 있지만 T2에는 없는 ID가 있을 것이라고 생각하였다. 그러므로 T1과 T2는 ID를 기준으로 연결될 수 있겠다고 생각하였다.

이를 바탕으로 T1과 T2의 공통 변수인 ID를 기준으로 데이터를 합하였고, total 이라는 새로운 table 을 만들었다. 가장 먼저 total 의 구조를 확인하기 위해 str() 함수를 이용하여 자료의 구조를 보았다. 결과에 따르면, total은 1442개의 관측치와 15개의 변수로 이루어진 데이터임을 알 수 있고, ID, V1, V2, V3, V5, V6, V7, V8, V9, V11, V12, V13 변수는 int 형태, 나머지 V4, V10, V14의 변수는 factor 형태로 이루어져 있음을 확인하였다.

```
> str(total)
'data.frame': 1442 obs. of 15 variables:
 $ ID : int 1 2 3 4 5 6 7 8 9 10 ...
 $ V1 : int 13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...
 $ V2 : int 23 23 24 26 30 32 27 30 27 23 ...
 $ V3 : int 46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...
 $ V4 : Factor w/ 3 levels "CNG","Diesel",...: 2 2 2 2 2 2 2 2 3 2 ...
 $ V5 : int 90 90 90 90 90 90 90 90 192 69 ...
 $ V6 : int 2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...
 $ V7 : int 3 3 3 3 3 3 3 3 3 3 ...
 $ V8 : int 210 210 210 210 210 210 210 210 100 185 ...
 $ V9 : int 0 0 0 0 0 0 0 0 0 0 ...
 $ V10: Factor w/ 4 levels "0","1","No","Yes": 1 1 2 2 2 1 1 2 1 1 ...
 $ V11: int 3 3 3 3 3 3 3 3 3 3 ...
 $ V12: int 5 5 5 5 5 5 5 5 5 5 ...
 $ V13: int 1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ...
 $ V14: Factor w/ 10 levels "Beige","Black",...: 3 7 3 2 2 9 5 5 6 3 ...
```

그 다음으로 summary() 함수를 이용하여 각 변수들에 대한 간단한 정보들을 확인하였다. 변수별로 요약통계량을 확인할 수 있었으며, V4, V10, V14의 경우 범주형 변수일 것이라는 판단을 하게 되었다.

```
> summary(total)
      ID      V1      V2      V3      V4
Min.   : 1.0   Min.   : 4350   Min.   : 1.00   Min.   : 1   CNG : 17
1st Qu.: 363.2 1st Qu.: 8450   1st Qu.: 44.00 1st Qu.: 43000 Diesel: 135
Median : 724.5 Median : 9900   Median : 61.00 Median : 63390 Petrol: 1270
Mean   : 724.1 Mean   : 10722   Mean   : 56.03 Mean   : 68480
3rd Qu.: 1085.8 3rd Qu.: 11950   3rd Qu.: 70.00 3rd Qu.: 87226
Max.   : 1442.0 Max.   : 32500   Max.   : 80.00 Max.   : 243000

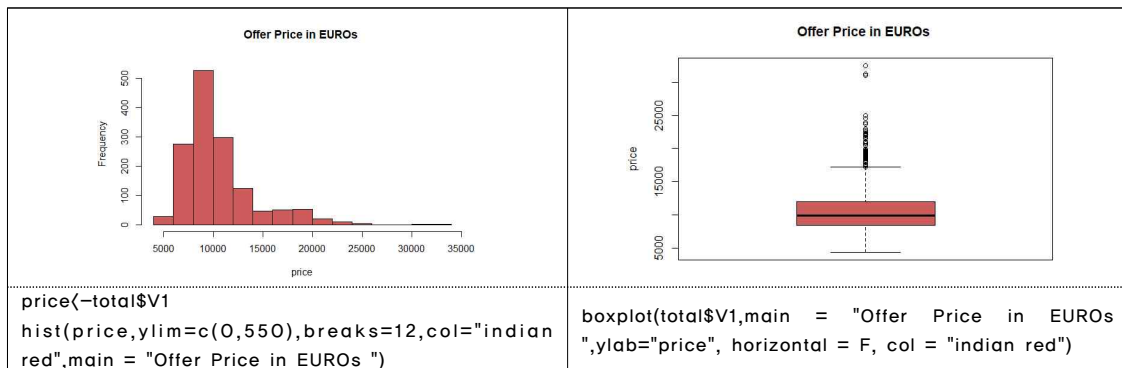
      V5      V6      V7      V8      V9
Min.   : 69.0   Min.   : 1300   Min.   : 2.000   Min.   : 19.00   Min.   : 0.00000
1st Qu.: 87.0   1st Qu.: 1400   1st Qu.: 3.000   1st Qu.: 69.00   1st Qu.: 0.00000
Median : 110.0   Median : 1600   Median : 4.000   Median : 85.00   Median : 0.00000
Mean   : 101.5   Mean   : 1576   Mean   : 4.029   Mean   : 87.05   Mean   : 0.06384
3rd Qu.: 110.0   3rd Qu.: 1600   3rd Qu.: 5.000   3rd Qu.: 85.00   3rd Qu.: 0.00000
Max.   : 192.0   Max.   : 16000   Max.   : 5.000   Max.   : 283.00   Max.   : 2.00000
NA's   : 1

      V10      V11      V12      V13      V14
0 : 786   Min.   : 3.000   Min.   : 3.000   Min.   : 1000   Grey : 300
1 : 549   1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.: 1040   Blue : 289
No : 66   Median : 3.000   Median : 5.000   Median : 1067   Red : 278
Yes : 40   Mean   : 3.816   Mean   : 5.026   Mean   : 1078   Green: 220
NA's : 1   3rd Qu.: 3.000   3rd Qu.: 5.000   3rd Qu.: 1085   Black: 191
Max.   : 36.000   Max.   : 6.000   Max.   : 10000   (Other): 163
NA's : 7       NA's : 1       NA's : 1       NA's : 1
```

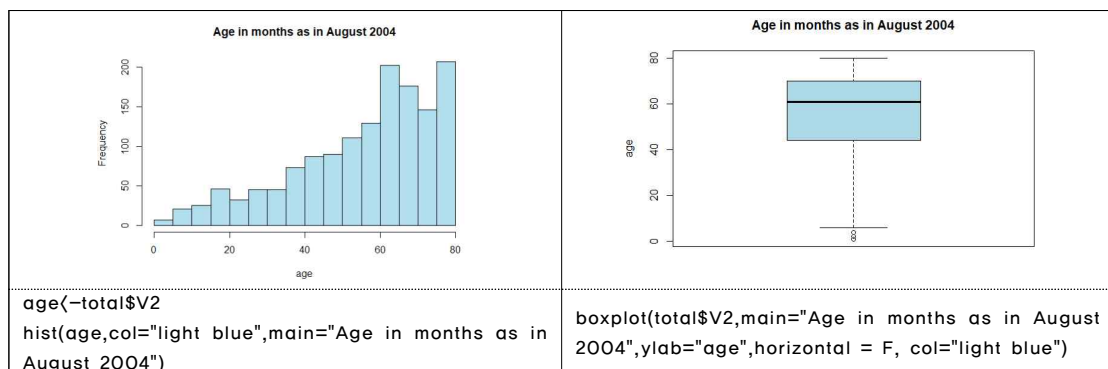
하지만 단순히 구조의 형식과 숫자만으로는 자료를 완전히 판단할 수 없기 때문에 각각의 변수를 단독적으로 시각화하였고, 그래프를 통해 변수를 범주형 변수와 수치형 변수로 나누었다. 이때 변수 설명에 대한 부분도 함께 고려하였다.

우선 ID의 경우에는 직관적으로 생각해보았을 때, 각각의 신원을 구분하는 자료이기 때문에 따로 시각화하여도 정보를 얻어낼 것이 없겠다고 생각하여 변수 판단을 위한 그래프는 그리지 않았다.

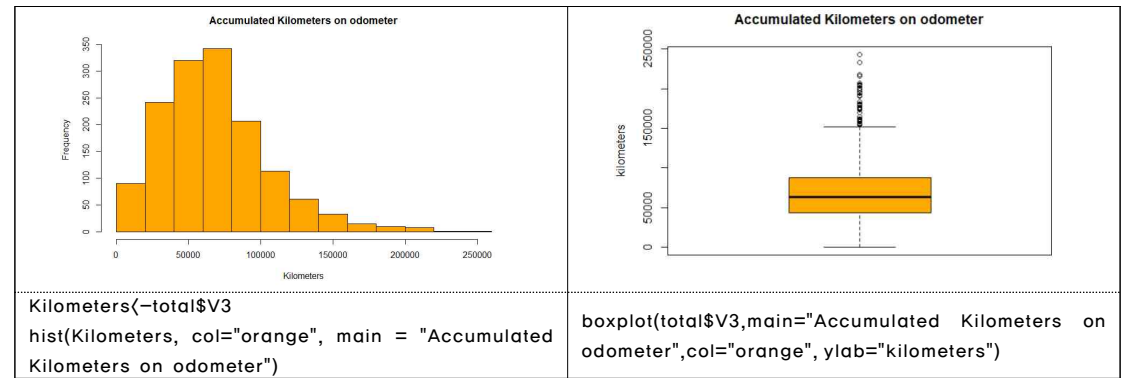
다음으로 V1은 'Offer Price in EUROS'를 나타내는 변수로, 핵심은 'price'라고 생각하였고, 가격과 같은 값은 범주형 자료가 아닌 수치형 자료라고 생각하였다. 이를 바탕으로 수치형 그래프인 히스토그램으로 V1에 대하여 시각화하였다. V1의 그래프를 보면 우향 왜곡 분포로, 정규성을 따르지 않는다는 것을 확인할 수 있다. 따라서 주어진 데이터는 가격이 적은 쪽에 비교적 분포가 많으며, 10000 근방에 가장 많은 객체가 몰려있음을 알 수 있다. 이를 더 자세히 알아보기 위해 상자그림을 그려보았고, 8450(Q1) 과 11950(Q2) 사이에 값들이 주로 분포하고 있음을 확인할 수 있었다. 상자그림 속 자세한 수치는 위에서 보인 summary(total) 그래프의 V1에서 확인할 수 있다.



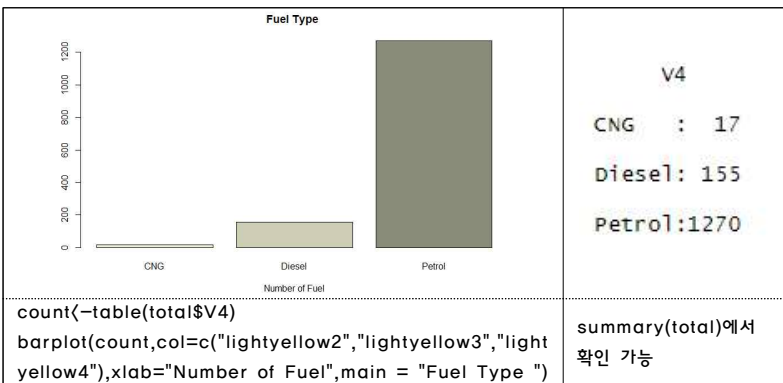
V2는 'Age in months as in August 2004'을 나타내는 변수로 'Age'에 관한 변수이다. age의 경우 시간 개념이 들어간 변수이기 때문에 범주형 변수가 아닌 수치형 변수라고 생각하였다. 이를 바탕으로 수치형 그래프인 히스토그램으로 V2를 시각화하였다. V2의 그래프를 보면 좌향 왜곡 분포로 정규성을 따르지 않음을 알 수 있다. 따라서 주어진 데이터는 age가 큰 쪽으로 갈수록 더 많이 분포하고 있다. 이를 더 직관적으로 판단하기 위해 상자그림을 그려보았고, age가 큰 부분에서 상자그림이 형성된 것을 볼 수 있었다.



V3은 'Accumulated Kilometers on odometer'를 나타내는 변수로 'Kilometers'에 관한 변수이다. 누적 거리의 경우 역시 범주형 변수가 아닌 수치형 변수이므로 히스토그램을 이용하여 시각화하였다. V3의 그래프를 보면 우향 왜곡 분포로 정규성을 따르지 않는다. 따라서 Kilometers가 작은 쪽에 비교적 더 많이 분포하고 있음을 확인하였고 특히 5000에서 10000 사이에 가장 많은 객체가 몰려있음을 확인할 수 있다. 이는 상자그림에서도 마찬가지로 표현이 됨을 확인할 수 있었다.

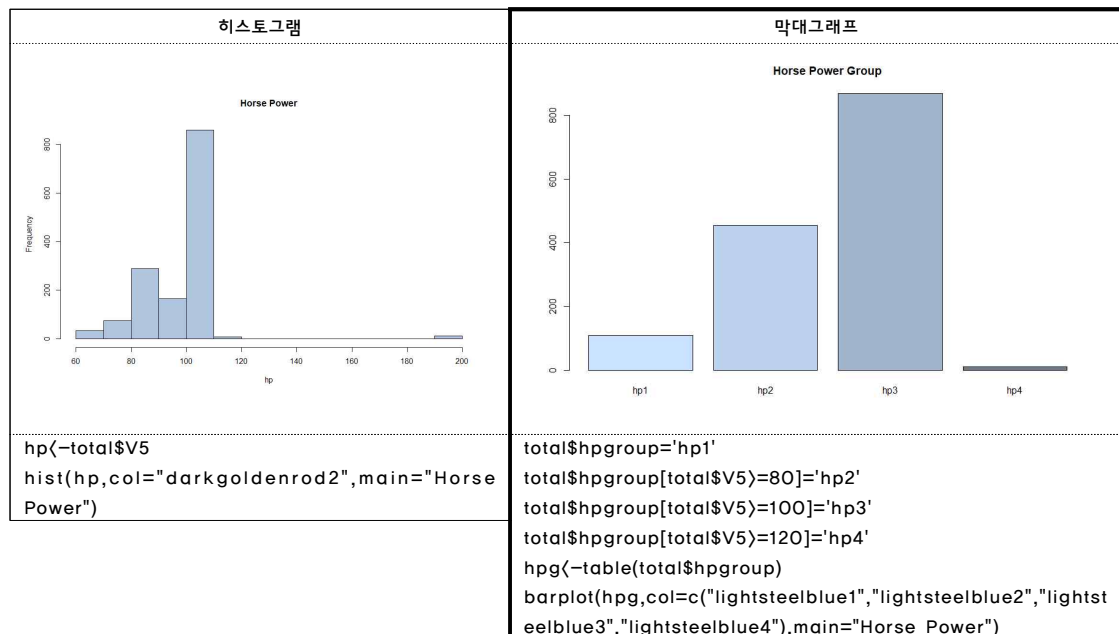


V4는 'Fuel Type'에 관한 변수로 'Type'을 기준으로 데이터가 나누어짐을 알 수 있다. 또한 이 변수를 이루고 있는 객체들은 모두 숫자형 형태가 아닌 문자형 형태였기 때문에 수치형 변수보다는 범주형 변수일 것이라 생각하였다. 이를 바탕으로 범주형 그래프인 막대 그래프로 V4를 시각화하였다. V4의 그래프를 보면, Fuel Type은 'CNG', 'Diesel', 'Petrol'으로 총 3가지이며 'CNG'의 개수가 가장 적고 'Petrol'의 개수가 가장 많음을 알 수 있다. 이때 위에서 보인 summary(total)의 결과창에서 CNG는 17개, Diesel은 155개, Petrol은 1270개임을 알 수 있다.



V5는 'Horse Power'에 관한 데이터로 수치형 변수일 것이라 생각하였고 히스토그램으로 시각화하였다. 그래프를 보면 값들이 이어지지 않아있으며 특정 구간에 몰려있음을 확인할 수 있었다. 그렇기에 더 직관적이고 쉽게 분석할 수 있도록 Horse Power의 값들을 구간을 지어 total에 새로운 변수 hpgroup을 만들어주었다. 마력의 세기를 총 4구간으로 나누었으며, 80미만은 hp1, 80이상 100미만은 hp2, 100이상 120미만은 hp3, 120이상은 hp4로 설정하였다. 각 구간의 도수는 table(total\$hpgroup)를 통해 알 수 있었으며, 순서대로 109, 454, 868,

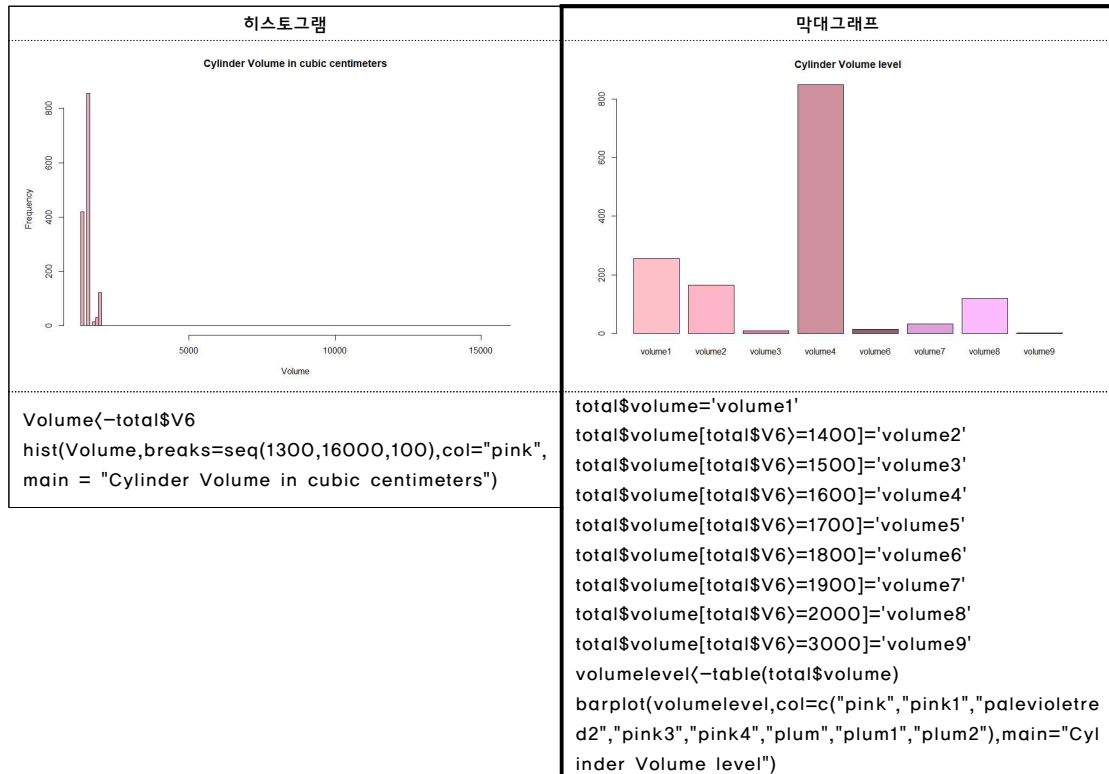
11 임을 확인하였다. 이를 막대그래프로 시각화하여 나타내었고, hp3 구간에 가장 많은 객체가 분포하고 있음을 확인할 수 있다.



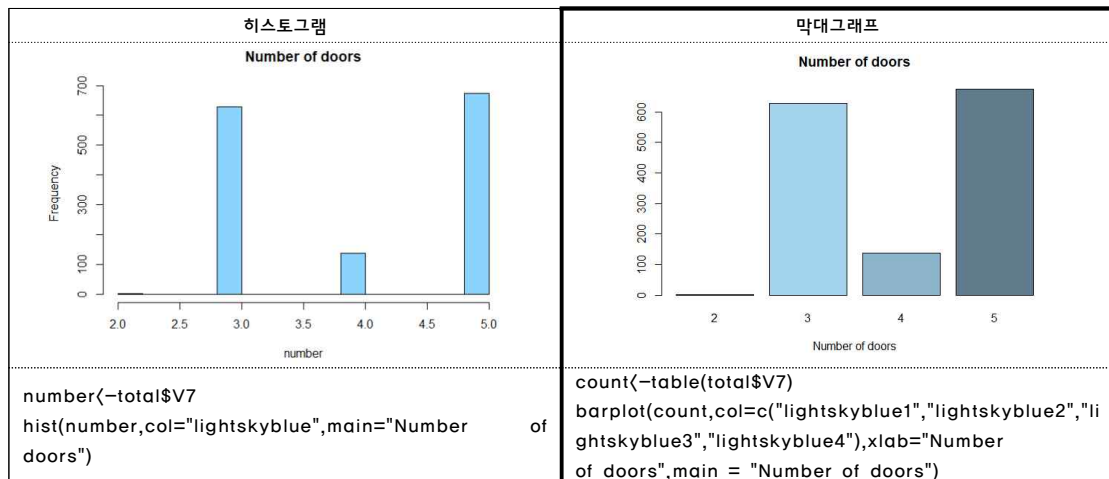
V6은 'Cylinder Volume in cubic centimeters'을 나타내는 변수로 'volume'에 관한 변수라고 판단하여 수치형 변수라고 생각하였다. 따라서 히스토그램으로 시각화하였다. 그래프는 심하게 우향 왜곡 분포를 띠고 있었는데 히스토그램에서 volume의 범위가 길게 잡혀있음을 확인하였다. 어떠한 값이 하나라도 뒤쪽 범위에 존재하기 때문에 그려진 결과라고 생각하여 의문을 가지고 table(total\$V6)을 이용하여 이상치의 존재를 확인하였다. 1개의 값이 16000을 보였고 나머지 값들은 모두 2000이내의 값들로 이루어져 있었다.

수치형 변수이지만 구간을 나누어 새로운 범주형 변수로 구성한다면, 이상치에 대한 영향이 줄어들어 분석하기 더 좋겠다는 판단이 들었다. 따라서 구간별로 나누어 volume이라는 새로운 변수를 만들어주었다. 이를 새로이 volume level로 명명하였고, 이를 기준으로 범주형 그래프인 막대그래프를 그려보았다.

막대그래프는 이전의 히스토그램보다 직관적인 판단이 더 쉬워졌다. 1600이상 1700미만의 값들로 이루어진 volume4 구간에서 가장 높은 분포를 보였으며, volume3 · volume5 · volume9 변수에서는 비교적 작은 분포를 보이는 것을 알 수 있다.

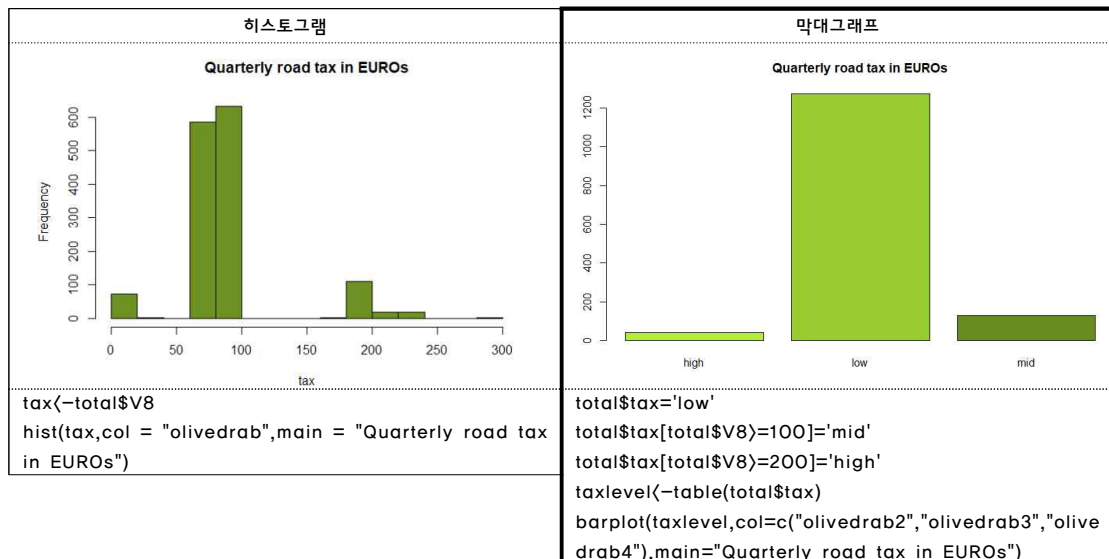


V7은 'Number of doors'를 나타내는 변수로 숫자로 이루어져 있지만 door의 개수에 따라 종류가 나뉜다고 생각하여 수치형 변수보단 범주형 변수라고 생각하였다. 확신을 가지기 위해 히스토그램과 막대그래프를 그려 비교해보았고, 예상대로 히스토그램의 분포를 보고 수치형 변수가 아니라고 생각하였다. 막대그래프를 보면 문의 개수가 2개, 3개, 4개, 5개인 경우로 나뉘어있음을 알 수 있다.

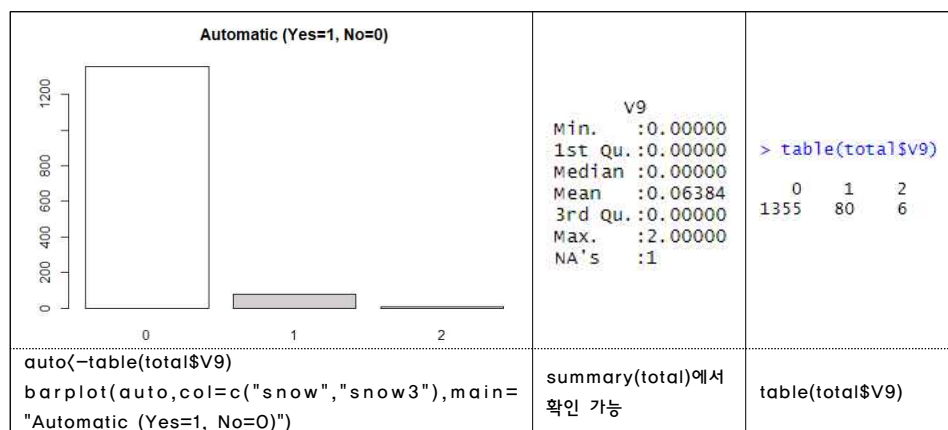


V8은 'Quarterly road tax in EUROS'를 나타내는 변수로 'tax'에 관한 변수이다. tax는 price처럼 가격을 나타내는 값이므로 범주형 변수보단 수치형 변수라고 생각하였다. 따라서 히스토그램으로 시각화하였다. 하지만 히스토그램을 보니 비연속적으로 특정 부분에 몰려 분포하는 것을 알게 되었다. 따라서 구간을 나누어 범주형 변수로 새로 구성하는 것이 더 좋겠

다고 판단하여 tax에 대한 구간을 'low', 'mid', 'high'로 나눈 tax라는 변수를 새로 만들어주었다. 이를 tax level로 지정하여 이에 대한 분포를 막대그래프를 통해 확인하였다. 결론적으로 0이상 100미만의 값들로 이루어진 low구간에 가장 많은 분포를 하고 있음을 알 수 있었다.

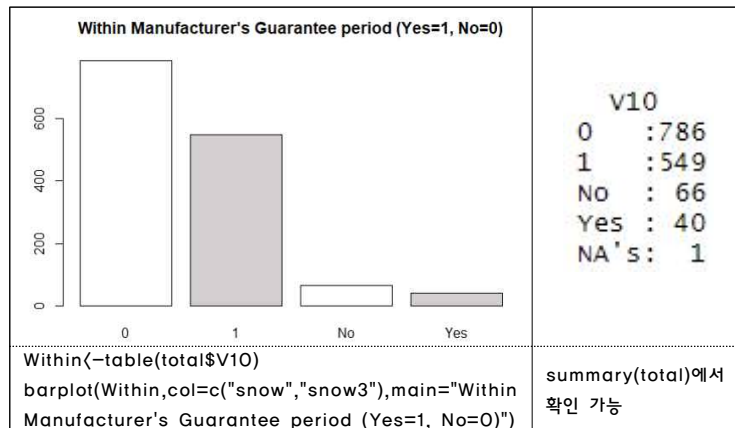


V9는 'Automatic (Yes=1, No=0)'을 나타내는 변수로, 변수 설명에 따라 0 또는 1로만 구성되어야 하는 변수라고 생각하였다. 숫자로만 표시되어있을 뿐, 자료가 Yes나 No를 의미하기 때문이다. 따라서 범주형 그래프인 막대그래프를 이용하여 시각화하였다. 그래프를 본 결과, 대부분의 자료는 0에 존재하기 때문에 Automatic을 갖추고 있지 않음을 알 수 있었다. table 함수를 통해 각 범주별 도수를 파악할 수 있었는데, 0은 1355개, 1은 80개, 2는 6개 (이상치)임을 알 수 있었다. 수치만 보아도 대부분의 분포가 0에 집중되어있음을 알 수 있다. 한편 위에서 보인 summary(total)의 V9부분을 통해 결측값도 1개 존재한다는 것을 알게 되었다.

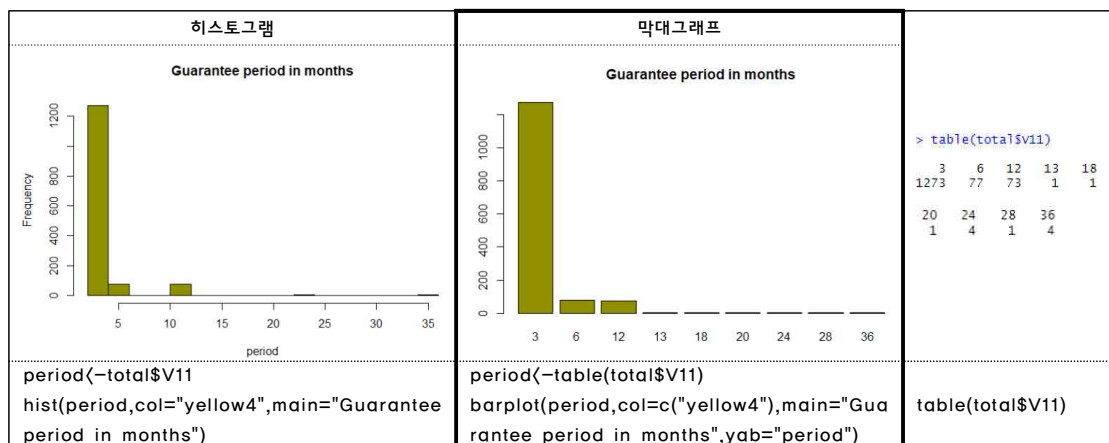


V10은 'Within Manufacturer's Guarantee period (Yes=1, No=0)'을 나타내는 변수로 직전의 V9변수와 동일하게 0과 1로만 이루어져야 하는 변수이다. 이때 0과 1은 곧 No나 Yes를 의미하는 바이므로 범주형 변수라고 생각하였다. 따라서 막대그래프를 이용하여 시각화하였다. 그래프를 보면 대부분의 값들은 0과 1에 분포하였지만 소수의 값들이 Yes나 No로 존재

하고 있음을 알게 되었다. 이는 모두 잘못 입력한 값이라 판단하였다. 한 가지 특이했던 점은 summary(total)에서 V10부분이 V9부분과 다르게 나왔던 것이다. 이는 아마도 V10에는 문자형 객체가 들어갔기 때문에 범주형 자료로 인식된 것 같다고 생각하였다. 한편 summary(total)에서 문자형으로 잘못 입력된 값은 106개이고 결측값 1개가 존재하고 있음을 확인하였다.

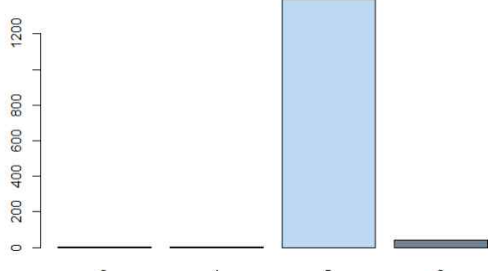


V11은 'Guarantee period in months'를 나타내는 변수로 'period'에 관한 변수이다. 처음에는 기간이 시간의 개념이라 생각하여 연속된 형태의 수치형 변수라고 생각하였고, 수치형 변수의 그래프인 히스토그램으로 시각화하였다. 하지만 그려진 히스토그램은 연속되지 않은 형태로 나오게 되었는데, 이를 통해 V11은 시간의 개념이 아닌 상품의 개념에 가까운 기간에 관한 변수라고 생각하게 되었다. 예를 들자면, '3개월 보증기간 상품', '6개월 보증기간 상품' ... 과 같은 개념이라고 생각한 것이다. 따라서 기간을 범주형 변수로 다시 판단하여 막대그래프를 그려보았고, 그 결과 3개월 보증기간의 범주에 가장 많은 분포를 하고 있음을 알게 되었다. 정확한 수치를 보기 위해 table 함수를 이용하였으며, 1442개의 관측값 중 1273개의 관측값이 3개월 보증기간 범주에 속한다는 것을 알 수 있었다.

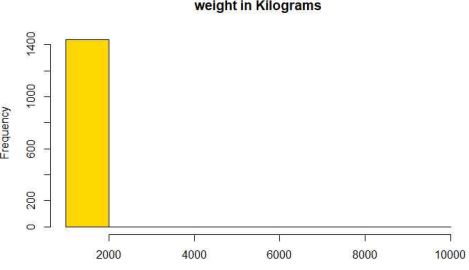
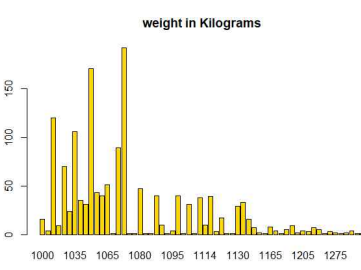


V12는 'Number of gear positions'을 나타내는 변수이고, 'number'가 핵심인 변수라고 생각하였다. 이 경우도 앞서 나온 V7과 같은 이유로 범주형 변수라고 생각하였고 따라서 막대 그래프로 시각화하였다. 막대그래프를 보면 4가지 종류로 나뉘고 있음을 알 수 있었고, 기어의 위치가 5개인 부분에서 가장 많이 분포하고 있었다. 자세한 수치를 알아보기 위해 table함

수를 이용하였으며 3은 2개, 4는 1개, 5는 1395개, 6은 43개 있음을 확인할 수 있었다. 추가로 summary(total)을 통해 결측치가 1개 존재하고 있음을 알게 되었다.

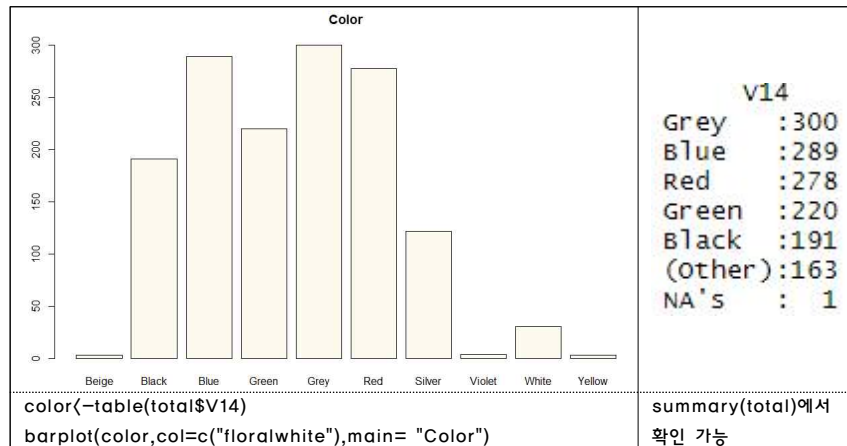
	<pre>V12 Min. :3.000 1st Qu.:5.000 Median :5.000 Mean :5.026 3rd Qu.:5.000 Max. :6.000 NA's :1</pre>	<pre>> table(total\$V12) 3 4 5 6 2 1 1395 43</pre>
<pre>positions<-table(total\$V12) barplot(positions,col=c("slategray1","slategray3", "slategray2","slategray"),main="Number of gear positions")</pre>	<pre>summary(total)에서 확인 가능</pre>	<pre>table(total\$V12)</pre>

V13은 'weight in Kilograms'을 나타내는 변수로 질량에 관한 변수이다. 질량의 경우, 수치형 변수이므로 히스토그램을 이용하여 시각화하였다. 예상과는 달리 히스토그램이 판단하기 쉽지 않게 나왔다. x축 변수의 범위가 지나치게 크게 설정되어있는 것을 보고 이상점이 존재하겠다는 의심을 하게 되었고, 이를 바탕으로 summary(total)의 결과를 보게 되었다. 예상에 맞게 이상치가 존재했으며 대부분의 값들은 2000이내에서 존재하고 있음을 알게 되었다. summary(total)를 참고하면 중앙값은 1067이며 평균은 1085임을 알 수 있다. 결론적으로 V13은 범주형이 아닌 수치형 변수가 맞다고 판단하였다. 그럼에도 불구하고, 확신이 들지 않아 막대그래프를 그려보았다. 막대그래프의 경우 수치형 자료라면 히스토그램에서 볼 수 있는 연속된 형태가 나올 것이라고 생각하였기 때문이다. 또한 이상치로 인한 그래프 왜곡은 히스토그램보다 적을 것이라고 생각하였다. 막대그래프의 형태는 예상했던 대로 연속적인 막대그림이 나왔고, 이상치를 제거했을 때의 히스토그램 모형과 비슷하겠다는 생각을 하게 되었다. 결과적으로 V13은 범주형이 아닌 수치형 변수라고 생각하게 되었다.

		<pre>V13 Min. : 1000 1st Qu.: 1040 Median : 1067 Mean : 1078 3rd Qu.: 1085 Max. :10000 NA's :1</pre>
<pre>weight<-total\$V13 hist(weight,col="gold",main="weight in Kilograms")</pre>	<pre>weight<-table(total\$V13) barplot(weight,col=c("gold"),main="weight in Kilograms")</pre>	<pre>summary(total)에서 확인 가능</pre>

V14는 'Color (Blue, Red, Grey, Silver, Black, etc.)'을 나타내는 변수로 'color'를 기준으로 데이터가 나뉘는 것을 알 수 있다. 따라서 범주형 변수라고 판단하였고 막대그래프로 시각화하였다. 총 10가지 색상으로 구성되어있음을 알 수 있었으며, summary(total)에서 정확한

수치를 확인할 수 있었다. 또한 결측치가 1개 존재한다는 것도 확인하였다.



지금까지 각 변수의 그래프를 그려보면서 변수의 종류를 크게 수치형 변수와 범주형 변수로 나누어 보았다. 결론적으로 다음과 같이 판단되었음을 알 수 있다. 또한 변수 설명을 종합해 본 결과, V4·V5·V6·V8·V12의 설명을 보고 도로 위의 이동수단(자동차·오토바이 등)의 종류일 것이라 추측하였고, V5·V7를 보고 자동차일 것이라 추측하였다. 이때 V3·V10·V11을 보고 중고 제품이라 생각하였으며, Target Variable(목표변수)가 V1이므로 유럽 어느 국가의 중고 거래 데이터일 것이라 생각하였다. 종합해보면, 주어진 데이터는 유럽 어느 국가의 자동차 중고 거래 내역일 것이라 생각하였고, 이를 바탕으로 나머지 분석을 진행해보았다.

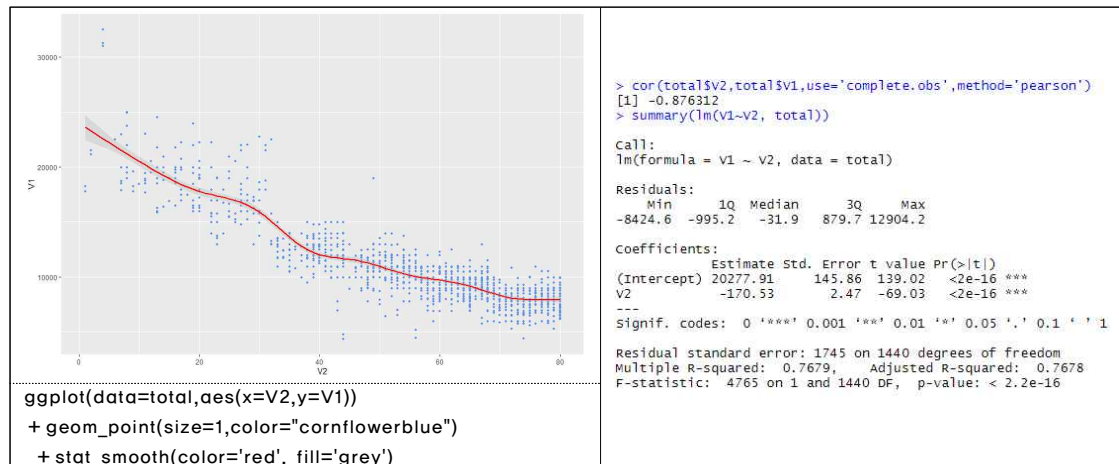
V1 Offer Price in EUROS	수치형 변수	V6→volume level Cylinder Volume	범주형 변수	V11 Guarantee period	범주형 변수
V2 Age	수치형 변수	V7 Number of doors	범주형 변수	V12 Number of gear position	범주형 변수
V3 Kilometers	수치형 변수	V8→tax level tax	범주형 변수	V13 weight in Kilograms	수치형 변수
V4 Fuel Type	범주형 변수	V9 utomatic	범주형 변수	V14 Color	범주형 변수
V5→hp group Horse Power	범주형 변수	V10 period	범주형 변수		

앞에서는 변수 개별마다 시각화하여 분석하였지만, 지금부터는 변수 간의 관계를 살펴며 서로 어떠한 영향을 미치는지를 중점으로 볼 것이다. 이때 주의해야 할 점은 변수의 형태가 어떤 형태인지를 잘 파악해야한다는 것이다.

먼저 **V1**과 **V2**의 관계를 살펴보았다. V1은 'Offer Price in EUROS'에 관한 변수이며, V2는 'Age in months as in August 2004'에 관한 변수이다. V1의 경우 Target Variable(목표변수)이므로 종속변수로 두었고 V2를 독립변수로 두었다. 이를 바탕으로 산점도를 그렸으며, 그 위에 추세선을 그려 시각화하였다. 또한 V1과 V2의 상관계수 및 선형 회귀분석에 대한 분석 결과를 확인해보았다.

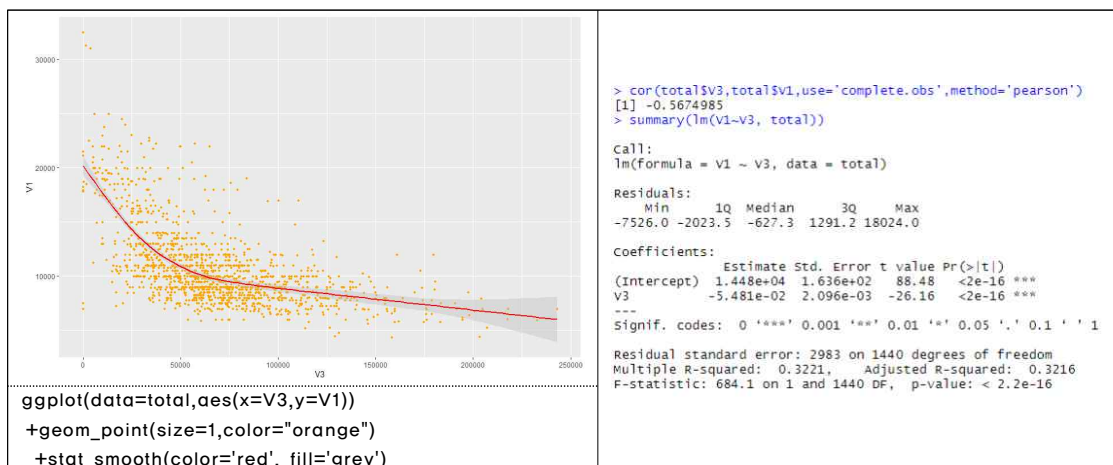
우선 상관계수는 -0.8763이고, 결정계수는 0.7679로 두 변수 사이의 상관관계가 강하다는

것을 알 수 있으며, 그래프의 추세선을 보아도 선형에 가까운 것을 확인할 수 있다. 또한 상관계수는 음수가 나왔는데, 그래프를 보면 상관계수에 맞게 V2가 커질수록 V1의 값이 작아진다는 것을 확인할 수 있다. 다시 말해, age가 커질수록 price는 떨어진다는 것이다. 이에 대해 (중고 자동차라는 가정하에) 자동차의 경우 출시되거나 출고된 지 오래된 차일수록 새로운 제품에 비해 품위가 떨어지기 때문에 가격이 내려갈 수 있겠다고 판단하였다.



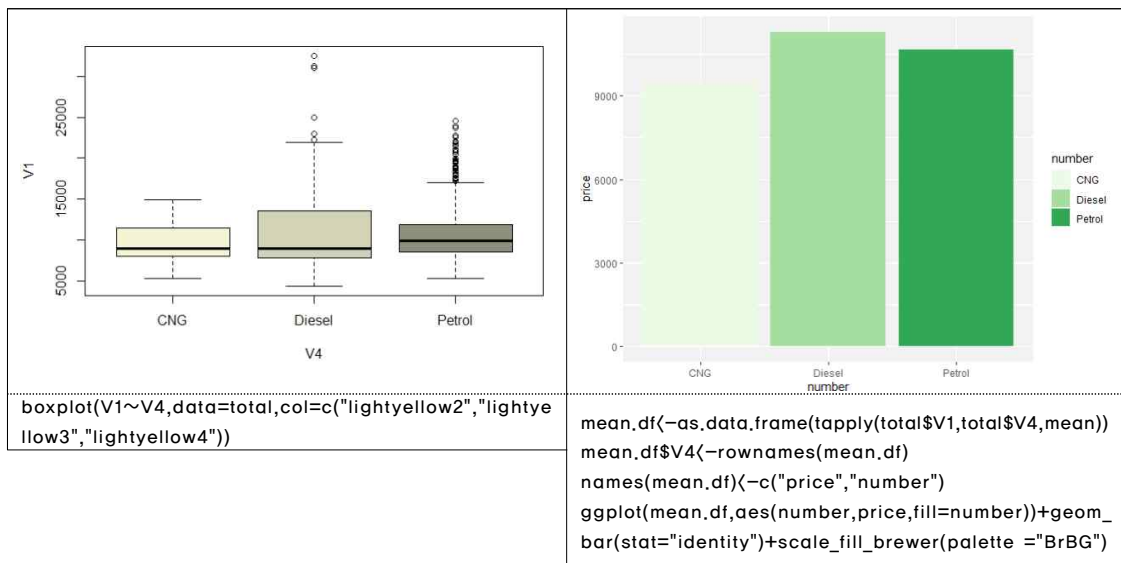
V1과 V3의 관계는 다음과 같다. V1은 'Offer Price in EUROS'에 관한 변수이며, V3는 'Accumulated Kilometers on odometer'에 관한 변수이다. 마찬가지로 V1은 종속변수에 두었으며 V3는 독립변수로 두었다. 이를 바탕으로 산점도를 그렸으며 그 위에 추세선을 그려 시각화하였다.

우선 상관계수는 -0.5674이고, 결정계수는 0.3221로 두 변수 사이의 상관관계가 약하다는 것을 알 수 있으며, 이에 맞게 그래프의 추세선도 비선형에 가까운 것을 확인할 수 있다. 또한 상관계수가 음수인 것을 보아, V3가 증가할수록 V1은 감소한다는 것을 알 수 있었고, 이 또한 그래프에서 확인할 수 있었다. 다시 말해, 주행기록계에 기록된 누적거리가 증가할수록 가격이 감소한다는 것이다. 중고자동차라는 가정 하에, 누적 주행 거리 기록이 높은 차일수록 사용이 많은 제품이므로 비교적 사용이 적은 제품보다 선호도가 낮을 것이라 생각하였다. 그로 인해, 누적 주행 거리 기록이 높은 차일수록 가격이 떨어지는 것이라 생각하였다.



V1과 V4의 관계는 다음과 같다. V1은 'Offer Price in EUROs'에 관한 변수이며 V4은 'Fuel Type'에 관한 변수이다. V1의 경우 수치형 변수이고 V4의 경우 범주형 변수이므로 상자그림을 이용하여 두 변수의 관계를 파악하였다.

상자그림을 보면 V4에 따라 V1의 변화량은 극명한 차이가 있는 것은 아니었으나 Diesel, Petrol, CNG 순으로 최댓값이 지정되었고, 중앙값의 경우 비슷하게 나온 것을 확인하였다. 하지만 Diesel, Petrol의 이상점이 너무 높기 때문에 평균이 중앙값과는 차이가 있을 것이라고 생각하였다. 따라서 각 범주의 평균 price를 계산하여 막대그래프 형식으로 시각화하였다. 상자그림에서는 Diesel, Petrol, CNG의 중앙값이 거의 일치하였는데, 막대그래프에서의 평균은 명확하게 차이남을 확인할 수 있었다. 각 범주별 평균은 Diesel, Petrol, CNG 순으로 높았다. 이상치가 많았기 때문에 평균과 중앙값에 차이가 났음을 알 수 있었다. 정리하자면, Fuel Type에 따라 price의 평균이 달라진다는 것이라 할 수 있다.

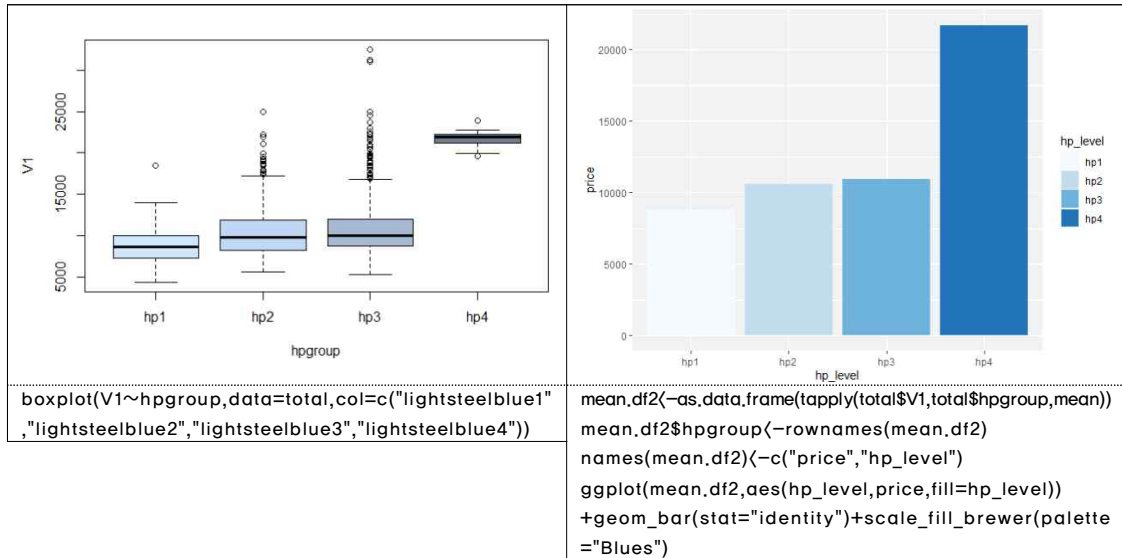


V1과 V5의 관계는 다음과 같다. V1은 'Offer Price in EUROs'에 관한 변수이며 V5는 'Horse Power'에 관한 변수이다. 앞서 V5의 경우, Horse Power의 일정 구간에 따라 나누어 만든 범주형 변수를 다시 만들어 주었기 때문에 V5 대신 hp group의 변수와 V1을 비교하였다. 또한 수치형 변수와 범주형 변수의 비교이기 때문에 상자그림을 이용하였으며, 범주별 평균에 대한 price도 알아보기 위해 범주별 막대그래프를 그려주었다.

먼저 상자그림을 보면 hp4구간이 가장 높은 price에 분포하고 있음을 알 수 있고, 나머지 hp1, hp2, hp3의 구간들은 거의 비슷하다는 것을 알 수 있다. 한편 범주별 막대 그래프를 보면 역시나 hp4가 눈에 띄게 높음을 알 수 있다. 또한 나머지 구간들도 hp1, hp2, hp3 순서대로 점점 평균 가격대가 올라가는 것을 확인할 수 있다. hp구간은 Horse Power의 크기에 따라 범주별로 나눈 것이므로, Horse Power가 클수록 가격대가 높아진 것이라 생각하였다.

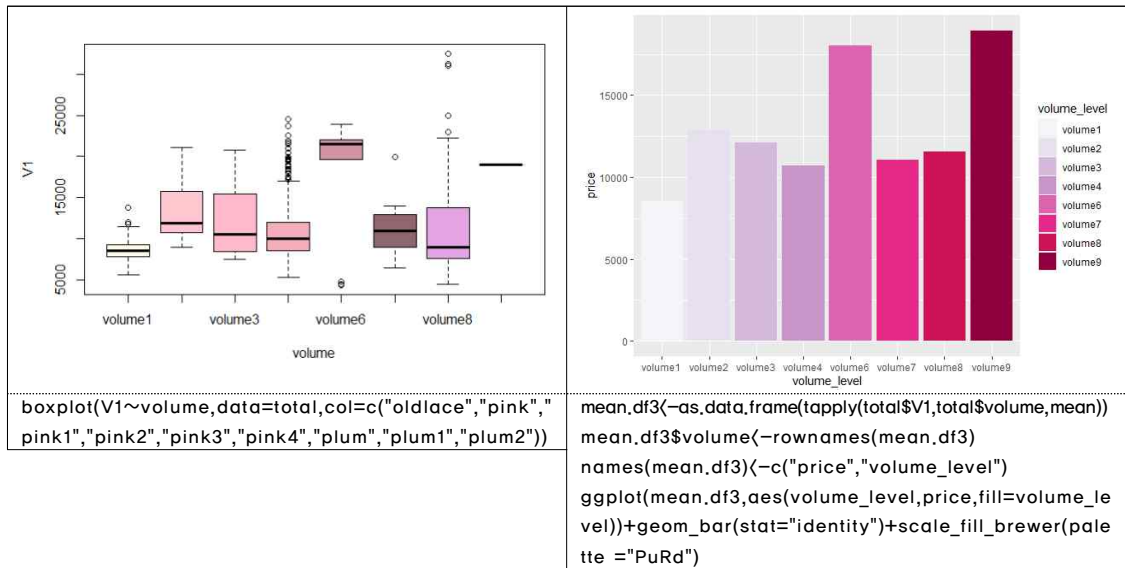
실제로 Horse Power은 마력으로, 자동차의 경우 마력이 높을수록 엔진의 출력이 좋다. 그렇기 때문에, 마력이 높은 차는 엔진의 성능이 더 좋은 것이므로, 다른 차들에 비해 가격이

높게 잡힌 것이라고 생각하였다.



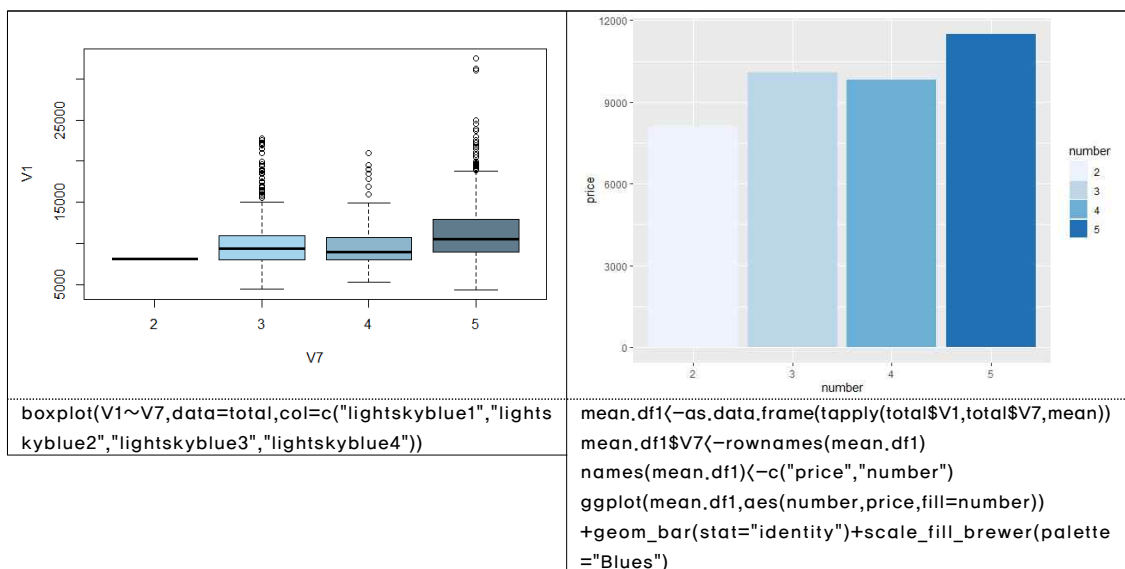
V1과 V6의 관계는 다음과 같다. V1은 ‘Offer Price in EUROS’에 관한 변수이며 V6은 ‘Cylinder Volume in cubic centimeters’에 관한 변수이다. Cylinder Volume in cubic centimeters은 흔히들 말하는 자동차 cc로, 정확히는 자동차 엔진 내부에 위치해있는 실린더 내부의 용적(체적)을 의미한다. 앞서 V6의 경우, Cylinder Volume를 일정 구간에 따라 나누어 범주형 변수인 volume을 새로 만들어 주었기 때문에 V1과 volume을 비교해주었고, 수치형 변수와 범주형 변수의 비교이므로 상자그림을 그려주었다. 또한 범주에 따른 평균값을 비교하기 위해 막대그래프를 그려주었다. 이때 주의해야 할 점은 volume9 구간은 이상치만 존재하는 구간이라는 것이다.

먼저 상자그림을 보면 volume에 따라 값들이 다양하게 분포하고 있었는데, 특히 volume6과 volume9의 상자그림이 높은 price에 위치하고 있음을 볼 수 있었다. 하지만 대부분 구간 속 값들은 3000미만에 분포하고 있는 반면 volume9의 경우 16000이라는 이상치만을 포함하고 있는 구간이었기 때문에 volume9 구간은 논외로 두었다. 이는 막대그래프의 범주별 평균에서도 동일하게 적용하였다. 그렇기 때문에 상자그림과 막대그래프를 종합하면, volume6 구간인 $1800 \leq V6 < 1900$ 에서 가장 높은 가격대를 이루고 있다고 할 수 있다.



V1과 V7의 관계는 다음과 같다. V1은 'Offer Price in EUROS'에 관한 변수이며 V7은 'Number of doors'에 관한 변수이다. V1은 수치형 변수이고 V7은 범주형 변수이므로 상자 그림을 그려주었으며, 범주별 평균 가격을 비교하기 위해 각 범주의 평균을 계산하여 막대 그래프로 시각화하였다.

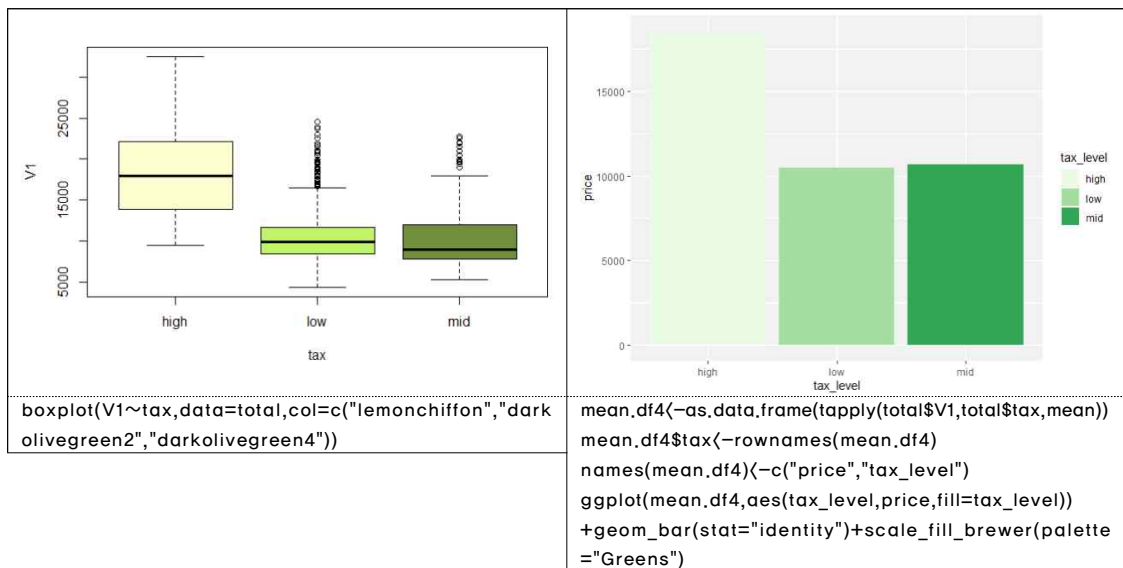
우선 상자그림에서 각 범주별 중앙값은 조금씩 차이남을 볼 수 있는데, 2, 4, 3, 5 순으로 중앙값이 커졌다. 상자그림에서 2를 제외한 나머지 세 구간에 이상점이 많은 것을 보고 각 범주별 평균은 중앙값과 차이날 수 있겠다고 생각하였는데, 범주별 평균을 나타내는 막대 그래프에서는 별반 차이가 없었음을 알 수 있었다. 중앙값과 마찬가지로 2, 4, 3, 5 순으로 평균이 커졌다. 3과 4구간의 평균과 중앙값이 크게 차이 나지 않는 것을 보아, 대체로 자동차의 문의 개수가 많아질수록 가격이 올라가는 양상을 나타낸다 할 수 있겠다고 생각하였다. 이에 대해 자동차의 문의 개수가 많아질수록 자동차의 크기는 커질 것이라고 추측하였고, 크기가 커짐에 따라 자동차의 가격이 오른 것이 아닌가 하는 생각을 하였다.



V1과 V8의 관계는 다음과 같다. V1은 ‘Offer Price in EUROS’에 관한 변수이며 V8은 ‘Quarterly road tax in EUROS’에 관한 변수이다. 앞서 V8의 경우, tax를 세 구간 (high, mid, low)로 나누어 만든 범주형 변수 tax를 만들어주었기 때문에 V1과 tax를 비교하였다. V1의 경우 수치형 변수이고 tax의 경우 범주형 변수이므로 상자그림을 그렸으며 범주별 평균 가격을 비교하기 위해 이를 계산하여 그린 막대그래프도 그렸다.

우선 상자그림을 보면 tax가 높은 high 구간이 가장 높은 가격대에 해당하는 중앙값을 이루고 있었고, mid와 low는 비슷했지만 low의 중앙값이 mid의 중앙값보다 조금 더 컸다. 한편 막대그래프를 보면 범주별 평균 가격은 high구간이 가장 높았고, 그 다음이 mid, low 순이었다.

결론적으로 분기별 자동차의 도로세가 많이 나가는 차 일수록, 특히나 200유로 이상에 해당할수록 자동차의 가격이 높다는 것을 알 수 있다.

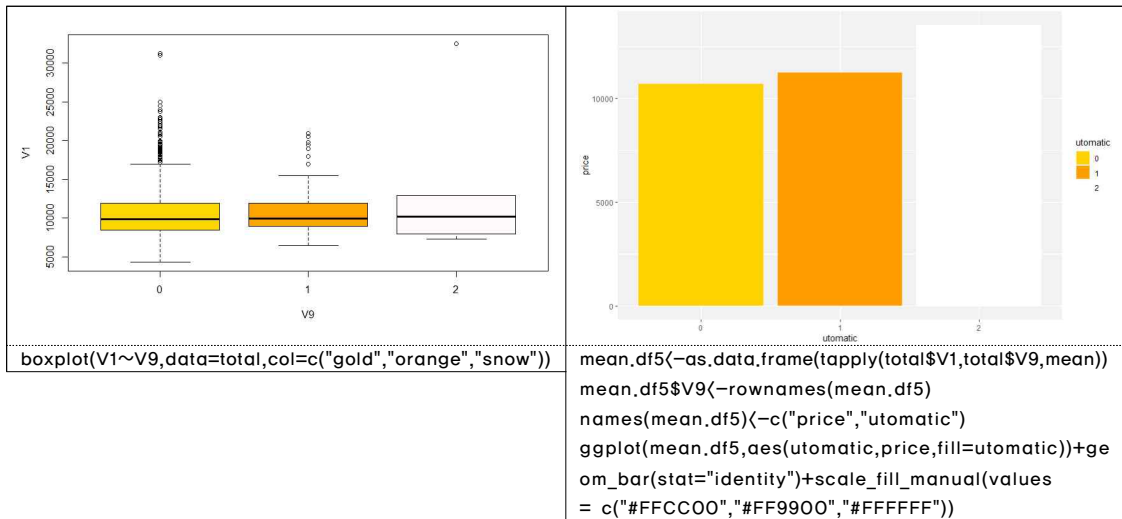


V1과 V9의 관계는 다음과 같다. V1은 ‘Offer Price in EUROS’에 관한 변수이며 V9은 ‘Automatic’에 관한 변수이다. V1의 경우 수치형 변수이고 V9의 경우 범주형 변수이므로 상자그림을 그렸으며 범주별 평균 가격을 비교하기 위해 이를 계산하여 그린 막대그래프도 그렸다. 이때 주의해야할 점은 V9는 0 또는 1로만 구성되어야하는 변수인데, 자료에 이상치인 2가 포함되어있기 때문에 2에 해당하는 값은 고려하지 않아야 한다는 것이다. 이를 그래프를 통해 직관적으로 잘 받아드려질 수 있도록 0 과 1 범주에만 색을 입혔으며 2 범주에는 흰색으로 처리하였다.

우선 상자그림의 경우 0과 1 범주의 중앙값은 비슷했다. 범주별 평균을 계산하여 그린 막대 그래프에서는 미묘한 차이를 보였는데 0보다 1의 가격 평균이 조금 더 높았다.

이를 바탕으로, 자동화 시스템이 구비된 제품은 당연히 품질 면에서 우수하기 때문에 가격이 높은 것이라고 생각하였다. 하지만 가격차이가 크지 않음에도 불구하고 앞서 분석한 V9의 경우 자동화 시스템이 구비되지 않은 0의 범주가 압도적으로 많은 분포를 차지하고 있어서 의외

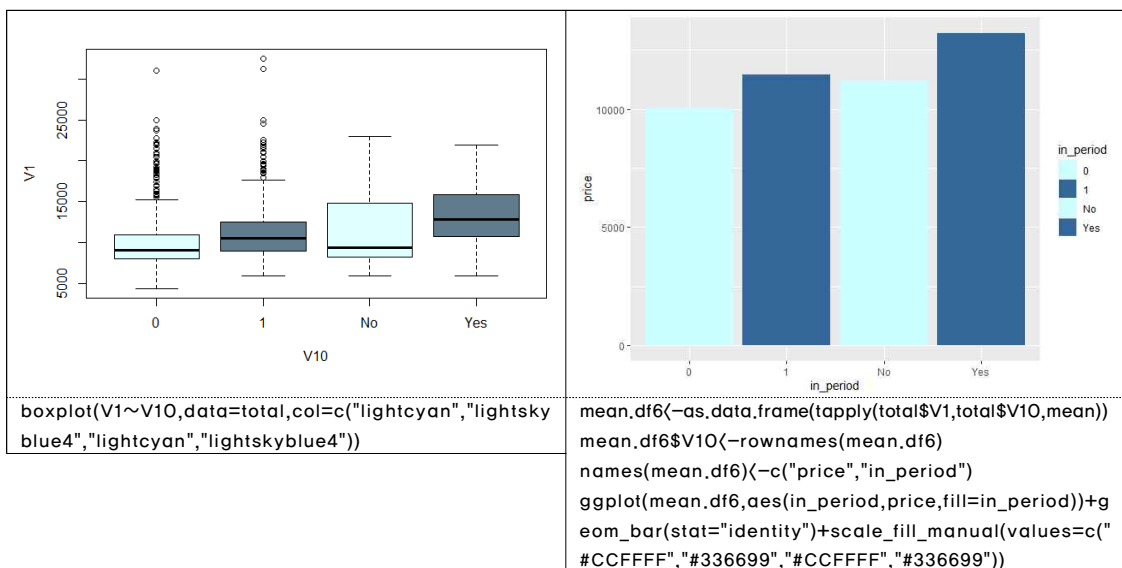
라고 생각하였는데, 만약 주어진 데이터의 배경이 유럽시장이라면 그럴만한 결과라고 생각하였다. 유럽은 대체로 자동변속기 차량보다 수동 변속기 차량이 더 많기 때문이다.



V1과 V10의 관계는 다음과 같다. V1은 'Offer Price in EUROS'에 관한 변수이며 V10은 'Within Manufacturer's Guarantee period'에 관한 변수이다. V1은 수치형 변수이고 V10은 범주형 변수이므로 상자그림을 그려 비교하였으며, 범주별 평균을 비교하기 위해 막대그래프를 그렸다. 한편 V10은 0 또는 1로만 구성되어야 하는 변수인데, No 또는 Yes와 같은 값들이 섞여있기 때문에 주의해주어야 한다. 여기에서는 No는 0과 묶어서, Yes는 1과 묶어서 해석하였고, 직관적인 판단을 돕기 위해 묶이는 범주끼리 같은 색을 지정해주었다.

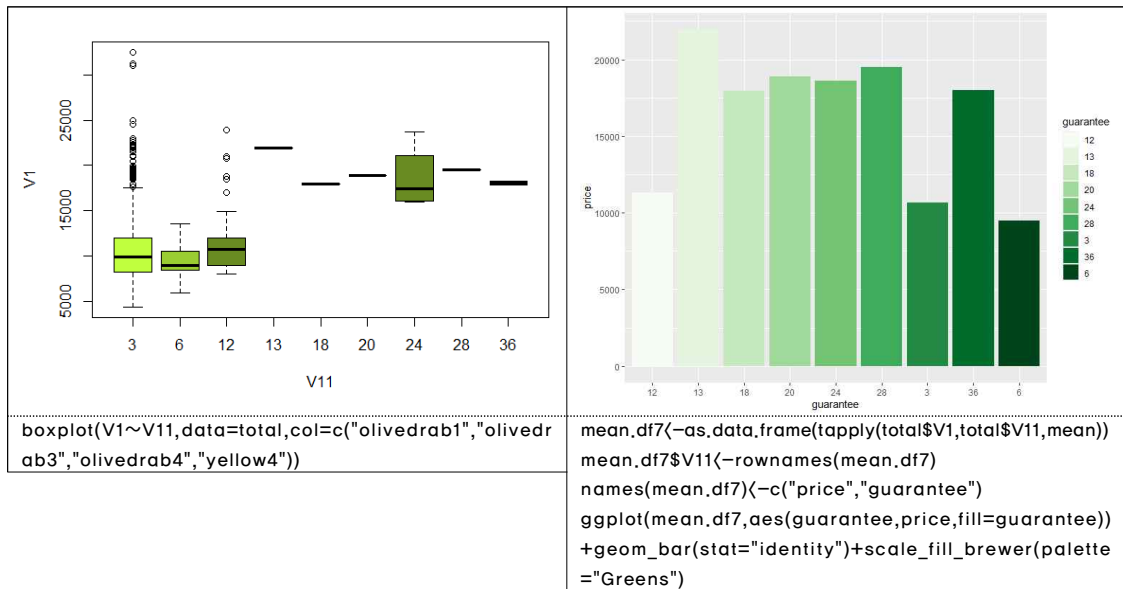
우선 V10의 경우 '제조사 보증기간 내' 인지를 기준으로 범주를 나누었다. 상자그림을 보면 보증기간 내에 해당하는 1과 Yes의 범주가 해당하지 않는 0과 No보다 가격이 더 높은 것을 알 수 있다. 이는 범주별 가격 평균을 비교하는 막대그래프에서도 동일하였다.

결론적으로 '제조사 보증기간 내'에 해당하는 제품은 그렇지 않은 제품들보다 가격이 높다고 할 수 있다.



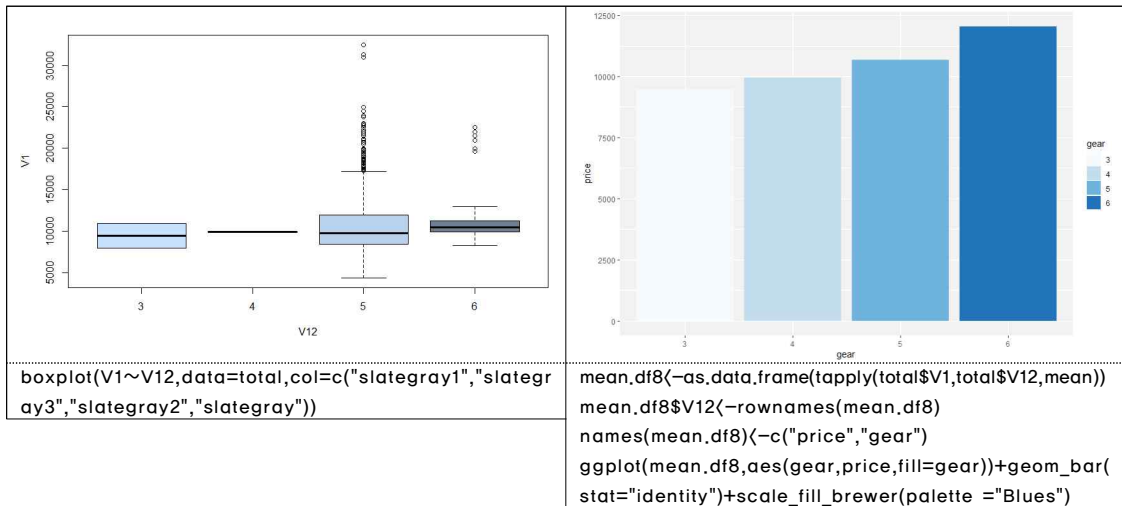
V1과 V11의 관계는 다음과 같다. V1은 'Offer Price in EUROs'에 관한 변수이며 V11은 'Guarantee period in months'에 관한 변수이다. V1은 수치형 변수이고 V11은 범주형 변수이므로 상자그림을 그려 비교하였으며, 범주별 평균을 비교하기 위해 막대그래프를 그렸다.

상자그림과 막대그래프를 보고 직관적인 판단이 힘들다는 생각이 들었다. 두 그래프 모두에서 특정한 패턴을 찾지 못하였는데, 이에 대한 이유를 생각해 보면 V11에 원인이 있었다. 앞서 분석한 V11을 보면 대부분의 값들이 3 범주에 몰려있음을 알 수 있고, 나머지 범주에는 소량의 분포가 있음을 알 수 있다. 3 범주 이외의 범주에 대한 데이터가 충분히 제공되지 않았기 때문에 특정한 패턴을 찾을 수 없었다고 생각했다.



V1과 V12의 관계는 다음과 같다. V1은 'Offer Price in EUROs'에 관한 변수이며 V12는 'Number of gear positions'에 관한 변수이다. V1은 수치형 변수이고 V12는 범주형 변수이므로 상자그림을 그려 비교하였으며, 범주별 평균을 비교하기 위해 막대그래프를 그렸다.

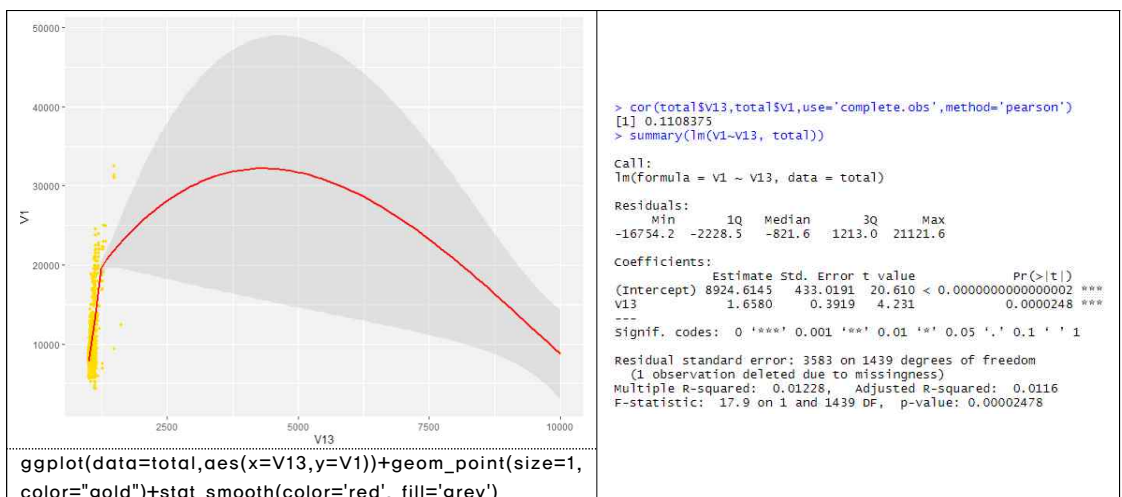
우선 상자그림을 보면 각 범주별 중앙값은 거의 비슷했는데, 앞서 분석한 V12의 자료를 보면 알 수 있듯이 대부분의 분포가 5 범주에 몰려있고 나머지 3, 4, 6에 해당하는 범주에는 소량만 분포하고 있기에, 해당 범주에 대한 충분한 데이터가 쌓이지 않았다고 생각하였고 그렇기에 분석 결과를 일반화할 수 없었다. 그럼에도 불구하고, 자료를 토대로 분석을 해본다면, 막대그래프를 통해 기어의 수가 많아질수록 가격이 올라간다는 결론을 내릴 수 있다.



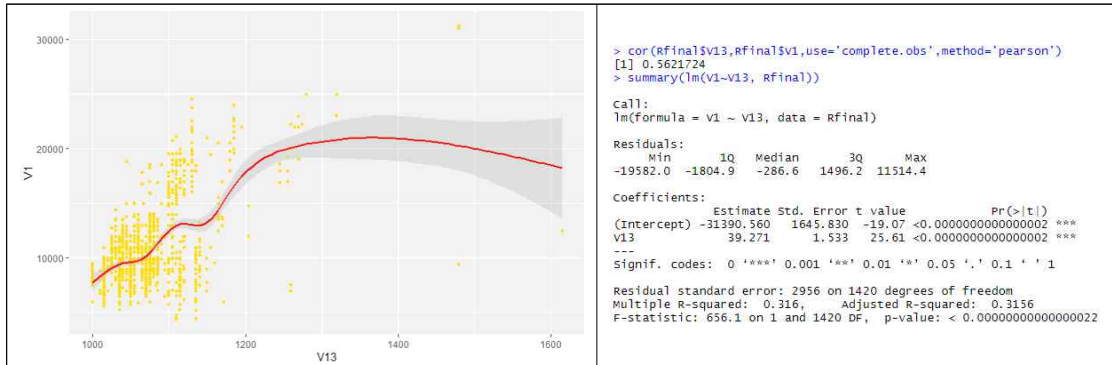
V1과 V13의 관계는 다음과 같다. V1은 'Offer Price in EUROS'에 관한 변수이며 V13은 'weight in Kilograms'에 관한 변수이다. V1과 V13 모두 수치형 자료이기에 산점도를 그렸으며 그 위에 추세선을 그려 시각화하였다.

주어진 산점도와 추세선이 그려진 그래프를 보면 당연히 비선형에 가까운 관계라고 판단할 것이다. 처음 그래프를 보았을 때에는 V13 변수 형태에 대해 의심해보기도 했지만, V13의 경우 명백한 수치형 변수라고 판단하였기에 무엇이 문제인지를 생각해보았다.

가장 먼저 `table(total$V13)`을 확인해보았으며 앞서 분석한 V13도 계속하여 확인해보았다. 그 결과, V13에는 이상치가 있다는 것을 알게 되었으며, 대부분의 값들이 2000미만에 존재하는데 반해 10000이라는 큰 수치의 이상치가 있다는 것을 확인하였다. 따라서, (이상치 10000을 제외한) x축이 2000미만인 산점도 그래프와 추세선을 보았고 이상치를 제외한 값들에서 선형관계를 띠고 있음을 확인하였다. 또한 변수 간 상관관계는 양의 상관관계를 이루고 있으며, 질량이 높을수록 가격이 높다는 결론을 낼 수 있다.



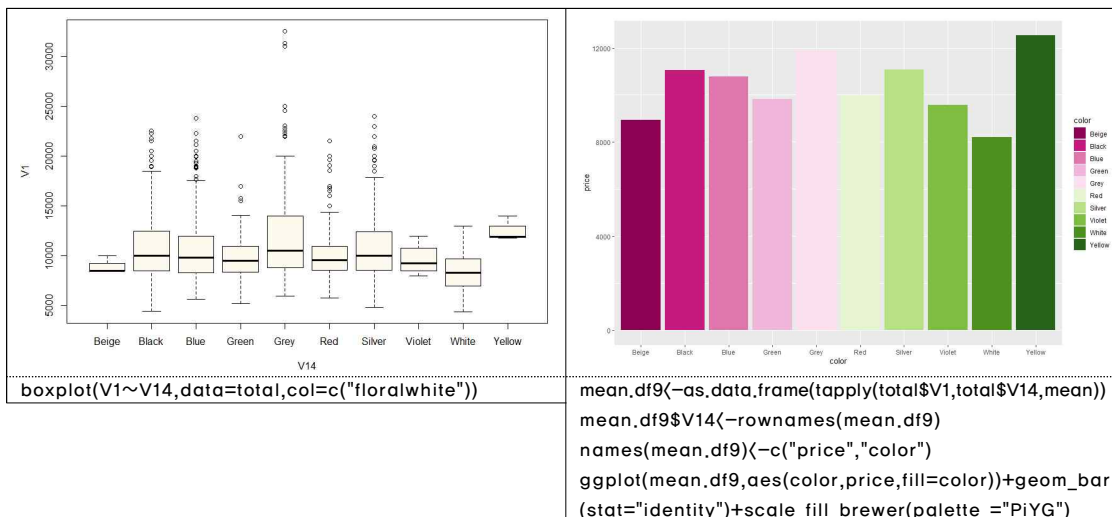
<문제3>에서 후술할 내용을 잠시 가져와 그래프를 보면 다음과 같다. 아래의 그래프는 문제가 됐던 이상치를 제거하고 그린 산점도와 추세선이다. 산점도와 추세선을 보면 앞부분은 실제로 선형에 비슷한 양상으로 증가하고 있음을 볼 수 있다. 또한 상관계수가 0.1108에서 0.5621까지 오름을 볼 수 있으며, 결정계수도 0.0122에서 0.316까지 오른 것을 볼 수 있다. 강한 상관관계라고 할 수 없지만 대체로 양의 상관관계를 이룸을 볼 수 있다.



V1과 V14의 관계는 다음과 같다. V1은 'Offer Price in EUROS'에 관한 변수이며 V14은 'Color'에 관한 변수이다. V1은 수치형 변수이고 V13은 범주형 변수이므로 상자그림을 그려 비교하였으며, 범주별 평균을 비교하기 위해 막대그래프를 그렸다.

그래프를 보기 전, 앞서 분석한 V14에 대해 보면 grey, blue, red, green, black 순으로 많이 분포하고 있음을 알 수 있으며 언급되지 않은 색상의 경우는 분포가 많지 않았다. 그렇기 때문에 언급한 5가지 색상 이외에 대해서는 충분한 데이터가 제공되지 않았기 때문에 선부를 판단을 할 수 없다고 생각하였다.

한편 5가지 색상에서도 눈에 띄는 패턴은 없었고, 그림에도 불구하고 찾는다면 도수가 큰 범주일수록 가격이 높게 측정됨을 알 수 있다. 이에 대해, 색상이 많다는 것은 선호도가 높다는 것이고 선호도가 높을수록 수요가 높기 때문에 가격이 높아질 것이라고 생각하였다.



<3. 주어진 데이터에서 문제점이 있다면, 그 문제점과 해결방안을 서술하시오.>

<2번> 문제에서 언급한 바와 같이, 데이터 탐색 과정에서 가장 먼저 T1과 T2의 데이터를 비교해보았다. 먼저 공통 변수인 ID를 기준으로 두 개의 테이블이 합쳐질 수 있겠다는 생각을 해보았고, 이를 바탕으로 T1의 관측값 개수와 T2의 관측값 개수를 비교해보았다. T1의 경우 1438개의 관측값을 가지고 있으며, T2의 경우 1437개의 관측값을 가지고 있다. 두 테이블의 관측값의 개수가 같지 않았기에, 중복값이나 결측치가 존재할 것이라고 생각하였다. 이를 바탕으로 T1과 T2를 합친 total 테이블에서 중복값과 결측치에 대해 다음과 같이 알아보았다.

먼저 중복값의 존재 유무는 행의 개수를 계산함으로써 확인하였다. total 자체 행의 개수는 1442개가 나온 반면, 중복값을 제거한 행의 개수는 1436개임을 확인하였다. 자세히 알아보기 위해 T1과 T2에서도 확인해보았는데, 각 테이블에서 중복값이 2개씩 나온 것을 확인할 수 있었다.

<pre>> nrow(total) [1] 1442 > nrow(unique(total)) [1] 1436 > total[which(duplicated(total)),]</pre>	<pre>> nrow(t1) [1] 1438 > nrow(unique(t1)) [1] 1436 > nrow(t2) [1] 1437 > nrow(unique(t2)) [1] 1435</pre>
<pre>nrow(total) nrow(unique(total)) total[which(duplicated(total)),]</pre>	<pre>nrow(t1) nrow(unique(t1)) nrow(t2) nrow(unique(t2))</pre>

중복값의 존재를 확인하였기에 이를 제거한 new 라는 새로운 테이블을 만들어주었고, 중복값에 대해 재확인한 결과 new 테이블에는 없다는 결과를 얻었다.

<pre>> new<-unique(total) > view(new) > > nrow(new) [1] 1436 > nrow(unique(new)) [1] 1436</pre>
<pre>new<-unique(total) View(new) nrow(new) nrow(unique(new))</pre>

다음으로 결측치의 유무를 확인하였다. sum(is.na()) 함수를 이용하여 12개의 결측치가 존재함을 확인하였고, colSums(is.na()) 함수를 이용하여 어느 변수에 몇 개의 결측치가 있는지 확인하였다. 결측치는 본래 T2 테이블에 존재했던 변수들에 공통으로 존재하였으며 V10변수에 가장 많은 결측치가 있다는 것을 확인하였다.

<pre>> sum(is.na(new)) [1] 12 > colSums(is.na(new))</pre>	<pre> ID v1 v2 v3 v4 v5 v6 v7 v8 0 0 0 0 0 0 0 0 0 v9 v10 v11 v12 v13 v14 hpgroup volume tax 1 1 7 1 1 1 0 0 0</pre>
<pre>sum(is.na(new)) colSums(is.na(new))</pre>	

결측치가 어느 행에 위치해 있는지 알아보기 위해 `new[!complete.cases(),]` 를 이용하였다. 특이하게도 ID가 114인 V8~V13 변수에 모두 결측치가 확인되었다. 이는 T2에 없던 ID가 T1과 결합하게 되면서 결측치가 발생한 것으로 판단하였다. 따라서 처음 T1과 T2 관측값의 개수가 달랐던 원인은 T2\$ID 에 114라는 값이 존재하지 않았던 것이라고 판단하였다.

한편 V11의 경우, 위에서 언급한 결측치를 제외하고도 6개의 결측치를 가지고 있었다. 결측치에 대한 처리는 크게 [결측치 없애기, 결측치 고치기, 결측치 대체하기] 가 있는데, 이 경우에는 결측치를 제거하기로 하였다.

```
> new[!complete.cases(new),]
```

	ID	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	hpgroup	volume	tax
113	114	24950	8	13253	Diesel	116	2000	5	234	NA	<NA>	NA	NA	NA	<NA>	hp3	volume8	high
724	727	7950	61	88127	Petrol	86	1300	3	69	0	1	NA	5	1015	Red	hp2	volume1	low
778	781	8950	68	77029	Petrol	86	1300	3	69	0	0	NA	5	1015	Black	hp2	volume1	low
931	934	8750	61	55747	Petrol	86	1300	3	69	0	0	NA	5	1020	Blue	hp2	volume1	low
989	993	9995	68	44458	Petrol	86	1300	3	69	0	1	NA	5	1015	Red	hp2	volume1	low
1283	1286	7500	75	74000	Petrol	86	1300	3	69	0	0	NA	5	1015	Black	hp2	volume1	low
1429	1430	7950	80	35821	Petrol	86	1300	3	19	1	0	NA	5	1015	Red	hp2	volume1	low

```
new[!complete.cases(new),]
```

V11의 결측치 데이터에 대해 아예 행 삭제를 하는 이유는, 1400개가 넘는 관측값에서 6개의 결측치는 전체의 0.4 정도의 매우 작은 비율이기 때문에 해당되는 행을 삭제하여도 영향이 크지 않을 것이라 생각하였다.

한편 V11은 범주형 변수로 대부분이 범주 3에 분포하고 있기 때문에, 결측값에 3을 대체해도 괜찮을 것이라 생각하였다. 이 경우도 마찬가지로 전체에 비해 작은 비율이기 때문에 데이터를 대체해도 전체 데이터에 영향이 크지 않을 것이라 생각하였다.

그럼에도 불구하고, 확인되지 않은 불확실한 값을 대체하는 것보다 아예 삭제하는 것이 더 좋을 것 같다는 판단을 하였기 때문에, `na.omit()` 함수를 이용하여 결측치를 제거해주었다. 결측치를 모두 제거한 테이블은 `final` 에 지정해주었으며, `final` 에 대해 결측치가 남아있는지 재확인하였다. 그 결과 다음 결과창과 같이 `final`에는 결측치가 없음을 확인할 수 있다.

```
> final<-na.omit(new)
> sum(is.na(final))
[1] 0
> new[!complete.cases(final),]
```

	ID	V1	V2	V3	V4	V5	V6	V7	V8
	V9	V10	V11	V12	V13	V14	hpgroup	volume	tax

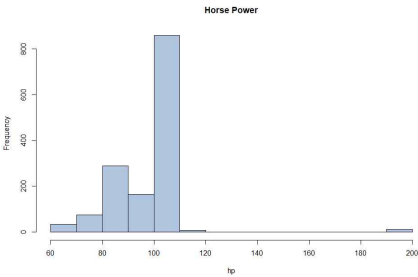
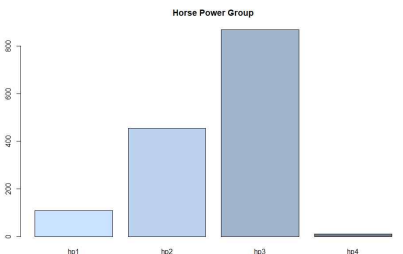
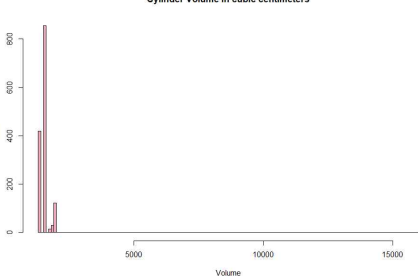
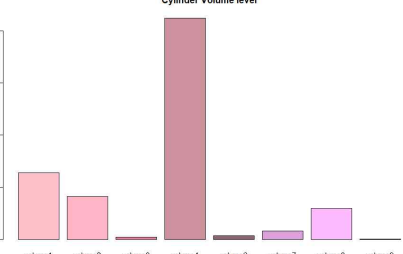
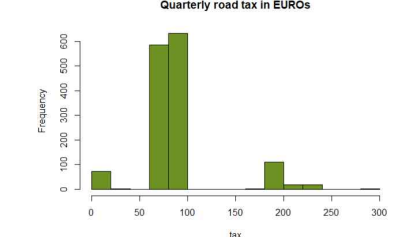
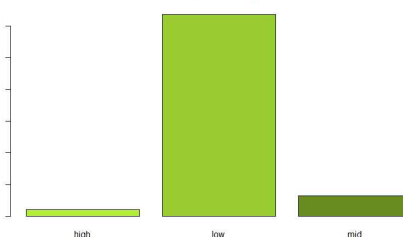
<0 행> <또는 row.names의 길이가 0입니다>

```
final<-na.omit(new)
sum(is.na(final))
new[!complete.cases(final),]
```

이렇게 전체 테이블에 대한 중복값과 결측치에 대한 처리를 해주었다. 전체에 대해 보았으니 지금부터는 앞서 데이터를 분석하면서 찾은 문제점에 대해 처리해 볼 것이다.

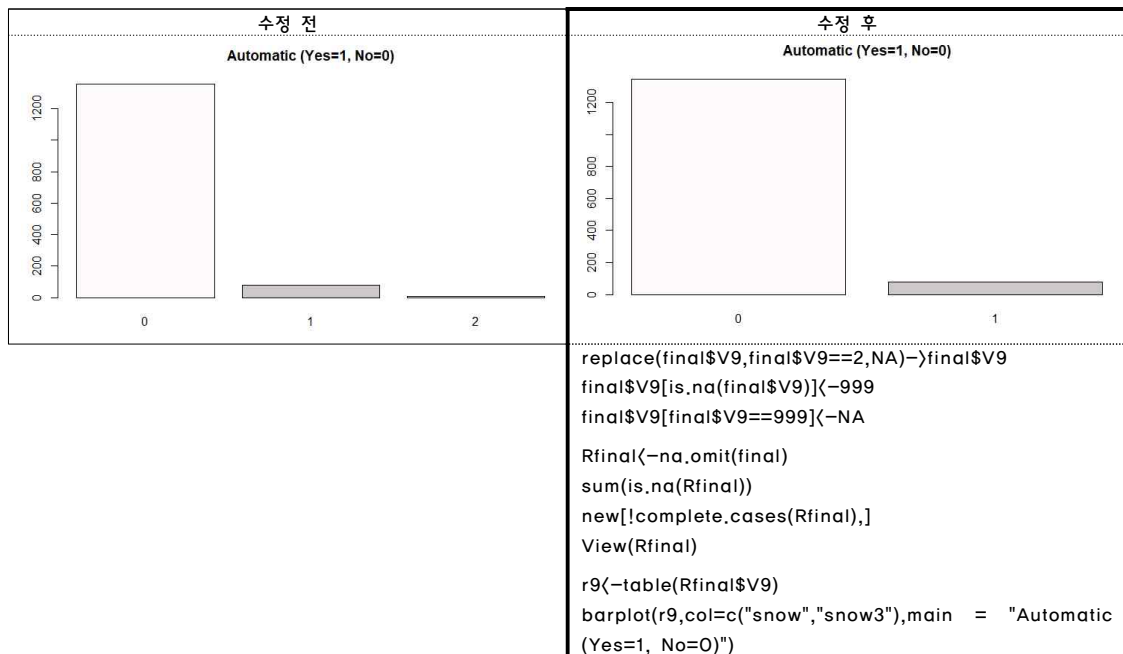
먼저 이번 데이터 분석을 하면서 중점적으로 두었던 변수 유형 설정에 대해 다뤄볼 것이다. 이번 데이터 속 변수의 경우, 수치형 변수로 판단하였지만 그러한 판단 아래에서는 특별한 패턴을 찾을 수 없었거나, 다른 변수와의 관계를 살펴볼 수 없는 경우가 많았다. 이를 해결하기 위해 그러한 변수들에 대해서 일정 구간을 나누어 새롭게 범주형 변수 형태로 만들어 주었다. 이렇게 처리한 변수는 V5, V6, V8 이며, 처리 전과 처리 후의 변수 그래프는 다음과 같다. 구간별로 나누어 범주형 변수로 재설정 해준 결과, 그래프만 보고도 직관적인 판단을 할 수 있었으며 무엇보다도 다른 변수와의 관계를 잘 살펴볼 수 있었다.

※그림의 순서대로 V5, V6, V8

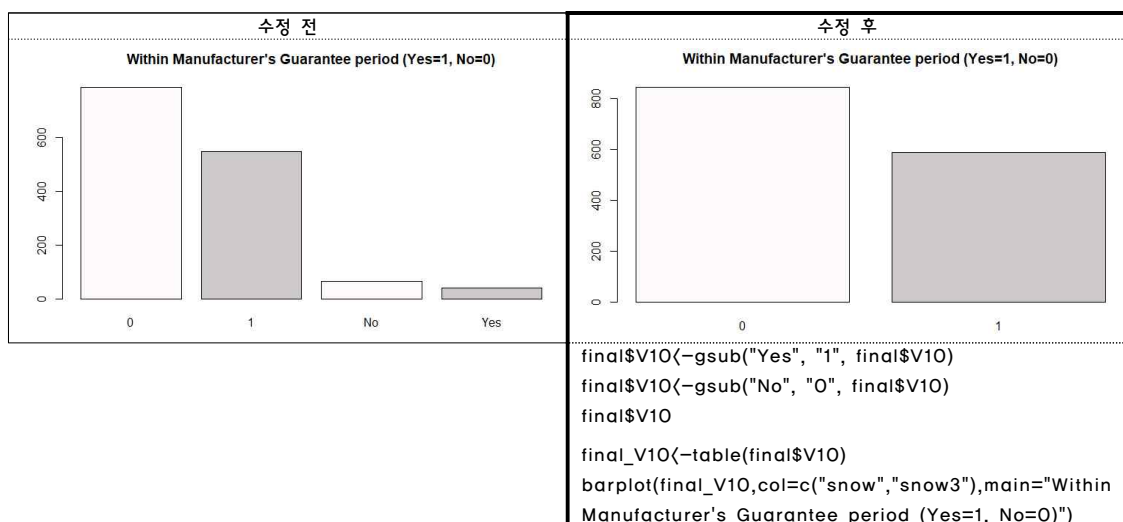
<p>히스토그램</p>  <pre>hp<-total\$V5 hist(hp,col="darkgoldenrod2",main="Horse Power")</pre>	<p>막대그래프</p>  <pre>total\$hpgroup='hp1' total\$hpgroup[total\$V5>=80]='hp2' total\$hpgroup[total\$V5=100]='hp3' total\$hpgroup[total\$V5=120]='hp4' hpg<-table(total\$hpgroup) barplot(hpg,col=c("lightsteelblue1","lightsteelblue2","lightsteelblue3","lightsteelblue4"),main="Horse Power")</pre>
<p>히스토그램</p>  <pre>Volume<-total\$V6 hist(Volume,breaks=seq(1300,16000,100),col="pink",main = "Cylinder Volume in cubic centimeters")</pre>	<p>막대그래프</p>  <pre>total\$volume='volume1' total\$volume[total\$V6=1400]='volume2' total\$volume[total\$V6=1500]='volume3' total\$volume[total\$V6=1600]='volume4' total\$volume[total\$V6=1700]='volume5' total\$volume[total\$V6=1800]='volume6' total\$volume[total\$V6=1900]='volume7' total\$volume[total\$V6=2000]='volume8' total\$volume[total\$V6=3000]='volume9' volumelevel<-table(total\$volume) barplot(volumelevel,col=c("pink","pink1","palevioletred2","pink3","pink4","plum","plum1","plum2"),main="Cylinder Volume level")</pre>
<p>히스토그램</p>  <pre>tax<-total\$V8 hist(tax,col = "olivedrab",main = "Quarterly road tax in EUROS")</pre>	<p>막대그래프</p>  <pre>total\$tax='low' total\$tax[total\$V8=100]='mid' total\$tax[total\$V8=200]='high' taxlevel<-table(total\$tax) barplot(taxlevel,col=c("olivedrab2","olivedrab3","olivedrab4"),main="Quarterly road tax in EUROS")</pre>

다음으로 수정해야하는 변수는 변수 설명에 동떨어진 혹은 조건에 맞지 않는 변수가 들어간 경우이다. 이에 해당하는 변수는 V9와 V10이다.

V9의 경우 Automatic (자동화)의 유무에 대한 변수로, 1과 0으로만 이루어져야 한다. 하지만 변수를 이루고 있는 관측값을 보면, 변수 설명에 맞지 않은 값인 숫자 2가 포함되어 있음을 V9의 막대그래프에서 확인할 수 있다. 2라는 값을 보고 1인지 0인지에 대해 추측을 할 수 없었기 때문에 해당 행들은 결측값으로 만든 뒤 삭제하였다. 이후 잘 수정되었는지 확인하기 위해 다음과 같이 막대그래프를 그려보았으며, 수정 전과 비교해 보았다.



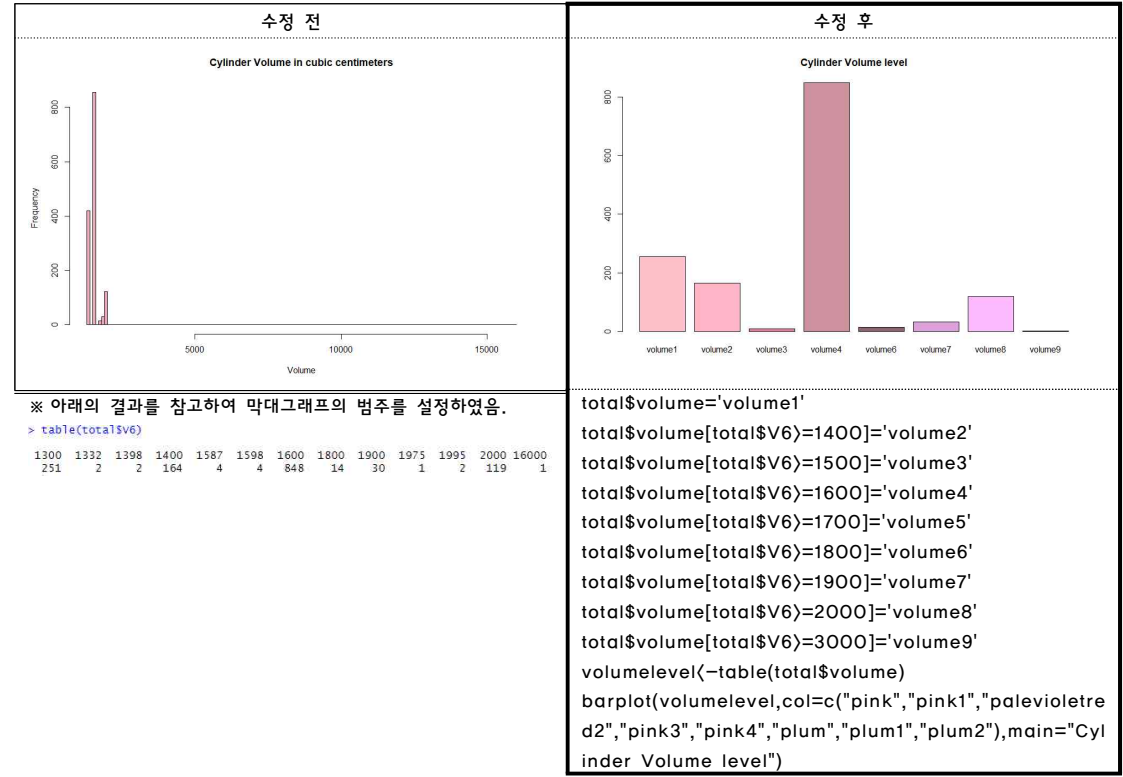
V10의 경우 Within Manufacturer's Guarantee period (제조사 보증 기간 내)에 대한 변수로, 1과 0으로만 이루어져야 한다. 하지만 변수를 이루고 있는 관측값들에 문자형 변수인 Yes나 No가 들어가 있는 것을 V10의 막대그래프에서 확인할 수 있다. 이에 대해 Yes의 경우 1로, No의 경우 0으로 변환하여 고쳐주었다. 이후 잘 수정되었는지 확인하기 위해 막대그래프를 그려보았으며 다음과 같이 나타났다.



마지막으로 수정이 필요했던 변수는 너무 큰 이상치를 가진 경우이다. 이상치가 너무 큰 경우, 데이터 분석에 대한 판단이 오류날 수 있기 때문에 주의해야 한다. 이에 해당되는 변수는 V6과 V13이다. 이상치가 너무 큰 경우에는 수치형 자료를 범주형으로 바꾸는 방법이나 이상치를 제거하는 방법을 이용해 수정하였다.

수치형 자료를 범주형으로 바꾸는 경우는 이상치가 적게 존재할 때에만 사용할 수 있겠다고 생각했다. 그 이유로는 막대그래프는 도수에 영향을 받는 그래프이므로 이상치의 개수가 적어야만 전체 자료에 영향이 덜 갈 거라고 생각했기 때문이다.

이 방법으로 수정한 변수가 바로 V6이다. table()함수의 결과를 참고하여 구간을 적절하게 나누었고, 막대그래프로 시각화하였다. V6의 경우 이상치가 1개만 존재했기 때문에 막대그래프에 큰 영향을 미치지 않았다.



다음으로 이상치가 너무 큰 경우, 아예 이상치를 제거하거나 대체값을 넣는 방법이 있다. 이 방법에 해당하는 변수가 바로 V13이다. V13의 경우, 대부분의 값들이 2000미만에 존재하는데 반해 10000이라는 큰 수치의 이상치가 있는 변수이다. 이로 인해 히스토그램의 경우 구간 범위가 크게 잡혀 일반적인 히스토그램의 모양이 잡히지 않아 분석하는데에 어려움이 있다. 이 뿐만 아니라 다른 (수치형) 변수와의 관계를 볼 때 산점도와 추세선이 심하게 왜곡되어 분석하는데에 어려움이 있다. 이러한 경우 이상치를 다른 값으로 대체하거나 삭제하는 방법이 있다. 예를 들어, V13의 이상치인 10000을 1000으로 고치거나 아예 삭제하는 것이 그 예이다.

결측값 때와 마찬가지로 확인되지 않은 값을 넣는 것보다 1400개 이상의 자료에서 하나의 값을 삭제하는 방향이 비교적 결과에 영향이 덜 미칠 것이라 판단하여, V13의 이상치는 삭제하는 방향으로 갔다. 이후 수정된 그래프를 보기 위해 히스토그램을 그려보았으며 추가로 V1과 V13의 산점도와 추세선 그려보았다. 그래프를 보고 수정 전에 비해 자료가 판단하기 좋아졌다고 생각하였다.

