**HW3 (Due date : July.14 PM 19:59, upload in web)**

**Submit the homework named as HW3\_이름.doc, HW3\_이름.R (코드도 첨부)**

**문제**

1. **데이터를 다루는 단계는 데이터 핸들링 -> 데이터 탐색 -> 통계적 모델링(통계모형, 기계학습, 인공지능)의 순서로 이루어진다. 각 단계에서 이루어지는 분석을 각각 2가지씩 예시로 드시오.**

**[데이터 핸들링] – 데이터 결합, 데이터 분할, 데이터 추출**

**[데이터 탐색] – 데이터 기술통계, 데이터 시각화**

**[통계적 모델링] – 두 그룹간 평균비교분석, 짝을 이룬 그룹간 평균비교, 분산분석**

1. **다음 분석들을 하는 경우 각각 필요한 분석 기법은 무엇인가? T-test, paired t-test, ANOVA 중 한 가지를 골라 적으시오.**

**거주 국가에 따른 키 차이의 유의성 검정 ANOVA**

**부부인 아내와 남편의 평균 수명의 차이의 유의성 검정 paired t-test**

**남녀 평균 키 차이의 유의성 검정 T-test**

**손 크기와 평균 키 사이의 유의성 검정 ANOVA**

**실습**

**데이터설명 : NSC2\_M40\_1000.csv (환자의 상병내역을 기록한 표 )**

**STD\_YYYY : 환자의 방문연도**

**RN\_INDI : 환자의 고유번호**

**Number : 환자의 방문횟수**

**데이터설명 : NSC2\_BNC\_1000.csv (건강보험 자격 및 보험료 표)**

**STD\_YYYY : 환자의 방문연도**

**RN\_INDI : 환자의 고유번호**

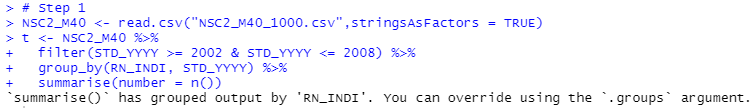
**SEX : 성별**

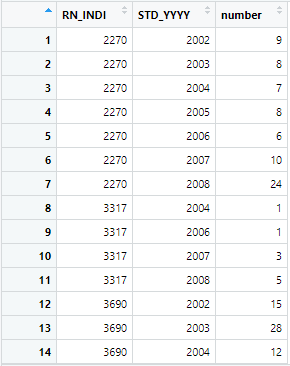
**GAINJA\_TYPE : 가입자 타입**

**기타 : 대상자의 사회경제적 변수**

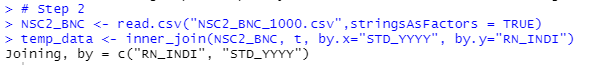
1. **다음 Step을 밟은 다음 문제를 해결하시오**

**Step 1 : NSC2\_M40\_1000.csv 에서 2002년부터 2008년까지의 연도별 환자의 방문횟수를 계산한다. 방문횟수를 number라는 변수로 만들어 기존 Table에 추가한다.**





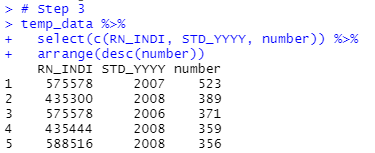
**Step 2 : NSC2\_BNC\_1000.csv와 Step1을 STD\_YYYY, RN\_INDI를 key로 하여 inner join한다.**





**Step 3: Hint를 이용하여 가장 환자의 방문횟수가 높은 환자의 ID와 해당 연도를 적으시오**

**Hint : arrange(데이터이름 ,desc(기준 변수))**



**환자ID : 575578 / 해당 연도 : 2007**

**데이터설명 : U.S SAT scores by state for 2010 (SAT\_2010.csv)**

**A data.frame with 50 rows and 9 variables.**

**state : a factor with levels for each state**

**expenditure : average expenditure per student (in each state)**

**pupil\_teacher\_ratio : pupil to teacher ratio in that state**

**salary : teacher salary (in 2010 US $)**

**read : state average Reading SAT score**

**math state average Math SAT score**

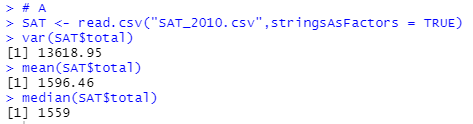
**write state average Writing SAT score**

**total state average Total SAT score**

**sat\_pct percent of students taking SAT in that state**

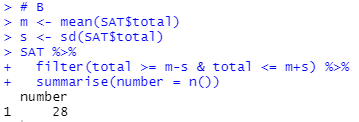
**2. 다음 데이터에 대해 문제를 해결하시오**

1. **Total 변수의 분산과 평균값은? 중앙값은?**



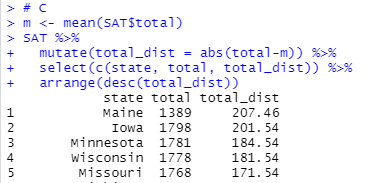
**분산 : 13618.95 / 평균값 : 1596.46 / 중앙값 : 1559**

1. **Total 변수가 (평균 – 표준편차, 평균 + 표준편차) 즉, 1시그마 범위 내에 속하는 state의 수는?**



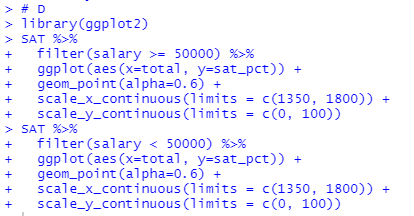
**28개**

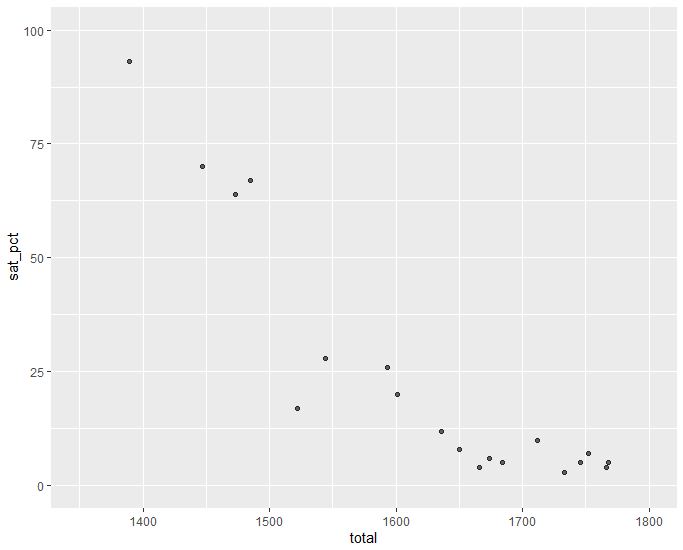
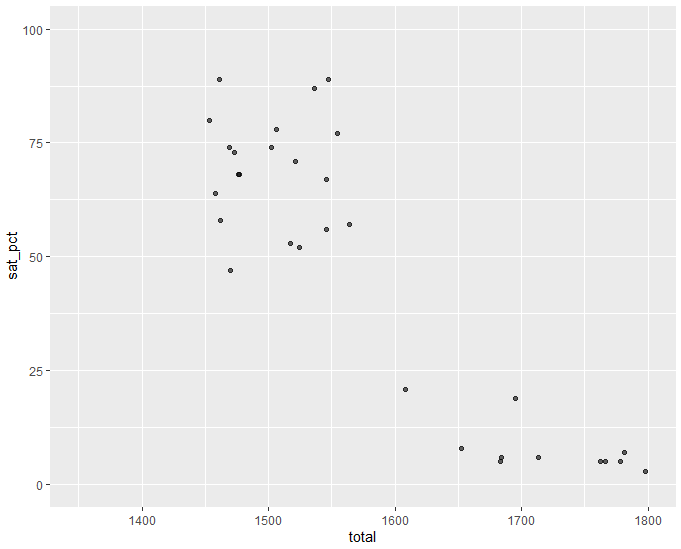
1. **Total의 평균으로부터 가장 멀리 떨어진 Total값을 가진 state의 이름은 무엇인가?**



**Maine**

1. **ggplot을 이용하여 Salary가 50000이상인 데이터와 이하인 데이터에 대해 Sat\_pct와 total 을 이용해 Plot graph를 그리시오. 이 두 그래프를 비교하며 해석하시오. (5점)**



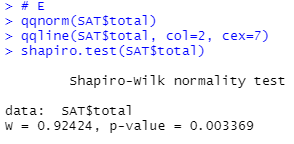


전체적으로 Salary와 무관하게 total값이 높으면 sat\_pct는 낮습니다.

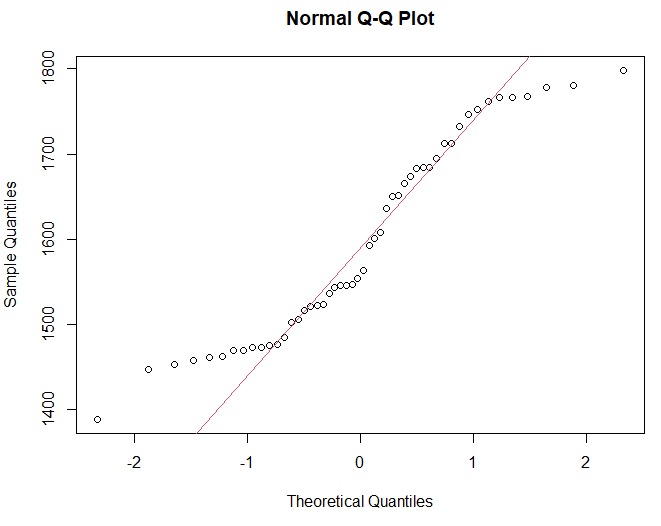
Salary가 50000이상인 그룹에서 sat\_pct(SAT를 치는 비율)이 전체적으로 큰 값을 가지는 경우가 더 많다.

Salary가 50000이상인 그룹에서는 상대적으로 total값이 낮은 주가 더 많다. (연봉은 더 많이 받으면서 실력은 없는 것인가..?)

1. **Total 변수에 대한 정규확률도를 그리고, Shapiro-Wilks test도 하여 비교평가**



p-value가 0.05보다 작으니까 귀무가설을 성립하지 않는다. 즉, 정규분포를 그리지 않는다.

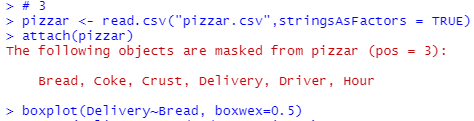


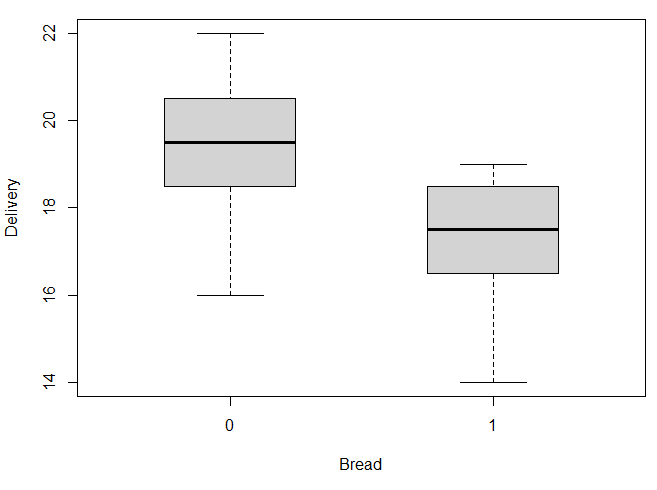
**데이터설명 : pizzar.csv (피자배달 데이터)**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | | |
|  | **Variable** |  | **Description** |
|  |  | | |
|  | **Crust** |  | **Thin=0, Thick=1** |
|  | **Coke** |  | **No=0, Yes=1** |
|  | **Bread** |  | **Garlic bread. No=0, Yes=1** |
|  | **Driver** |  | **Male=M, Female=F** |
|  | **Hour** |  | **Time of order in hours since midnight** |
|  | **Delivery** |  | **Delivery time in minutes** |
|  |  | | |

**3. bread를 주문/비주문에 따라 피자 배달 평균시간에 차이가 있는가?**

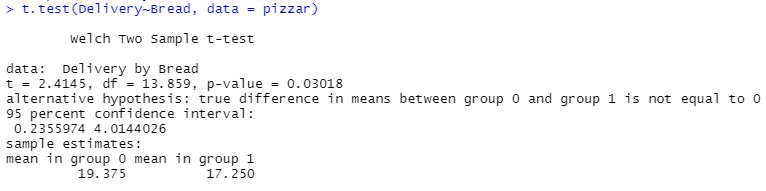
* 1. **상자그림(boxplot)을 그리고 비교설명**





빵을 주문 안한 경우 배달 시간이 전체적으로 더 많이 걸리는 것으로 보인다.

* 1. **ANOVA (혹은 T-Test)를 수행하여 Garlic bread를 주문/비주문에 따라 피자 배달 평균시간에 차이가 유효한지를 설명해라. 그 다음 평균의 차이를 적으시오**



p-value가 약 0.03이므로 유의수준 0.05에서 Garlic Bread 주문 여부에 따라 배달시간은 유의한 차이가 있다고 할 수 있다.

갈릭빵 주문 안했을 때, 평균 배달시간은 19.375

갈릭빵 주문 했을 때, 평균 배달시간은 17.250

입니다.

그래서 평균 차이는 2.125이다.