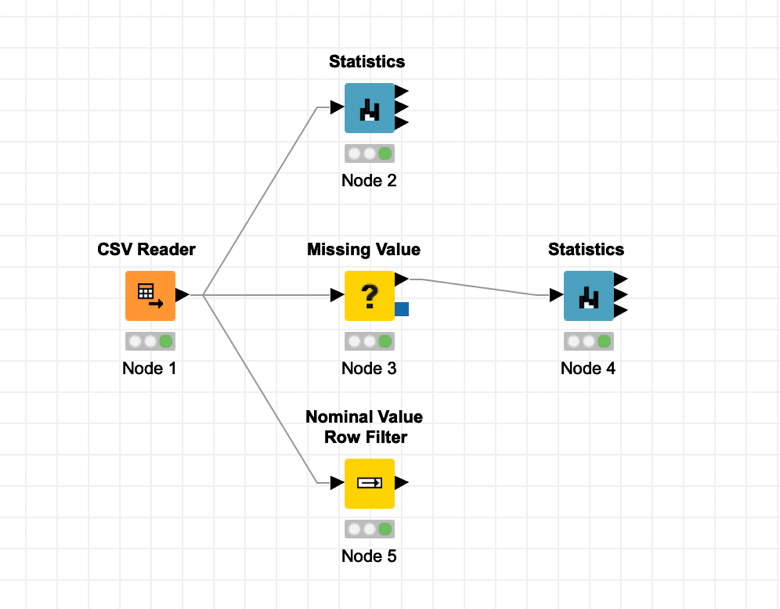
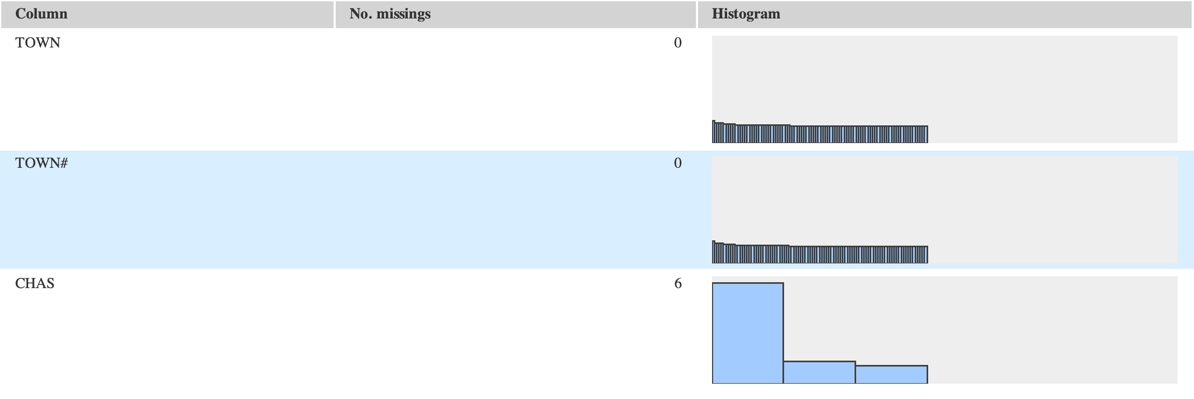
2019-DM

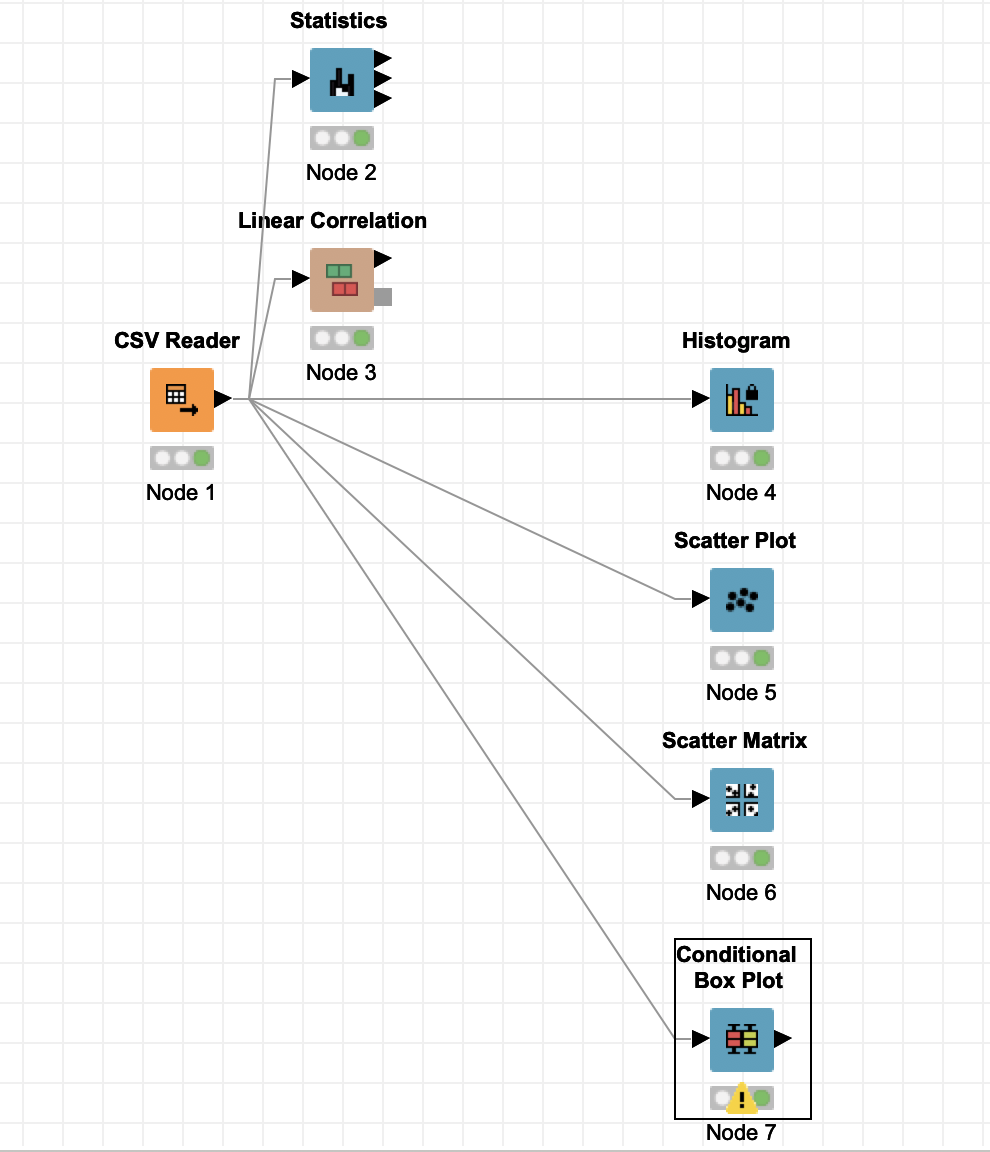
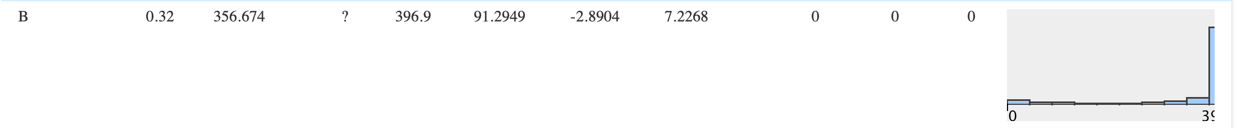
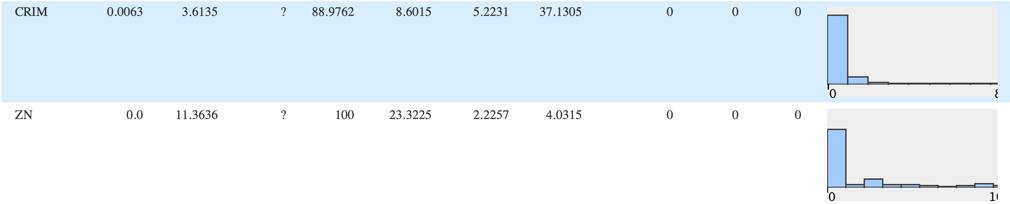
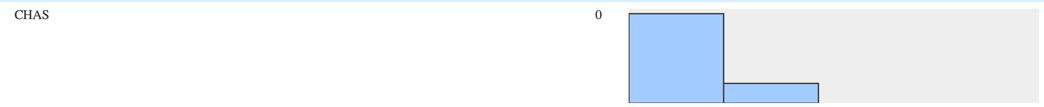
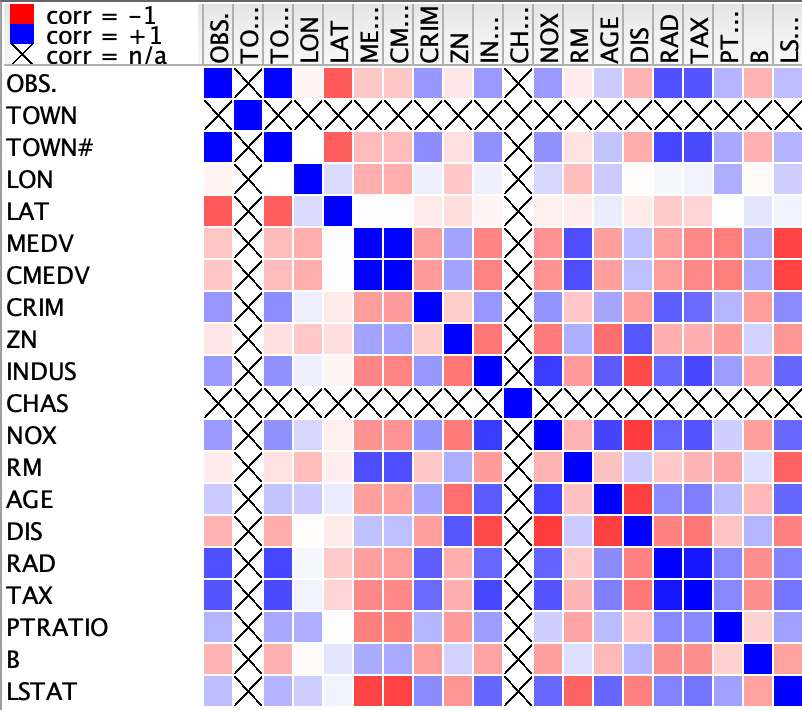
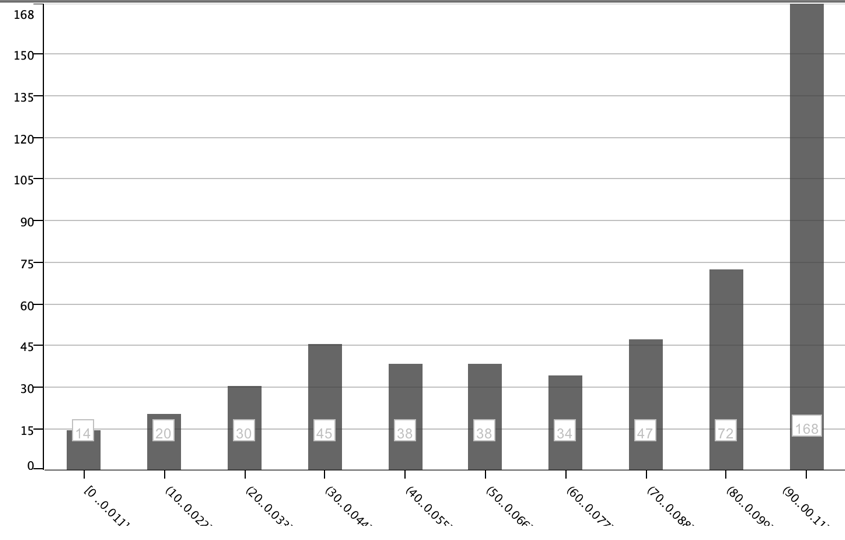
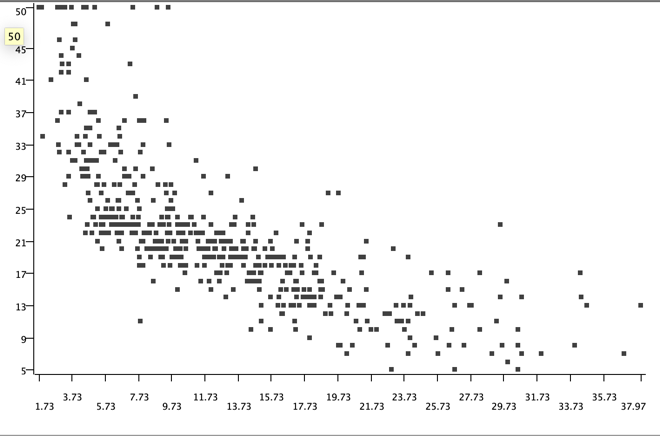
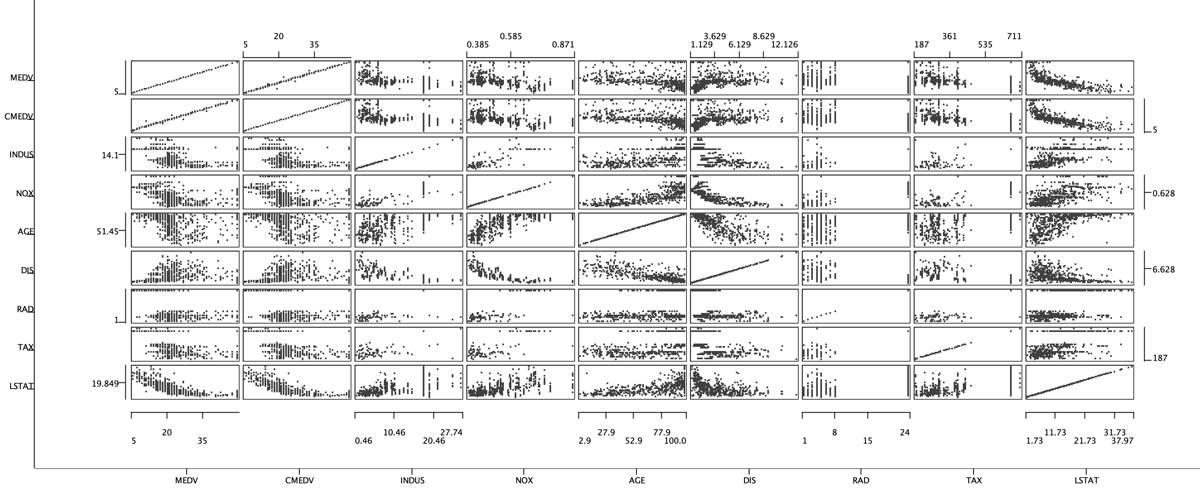
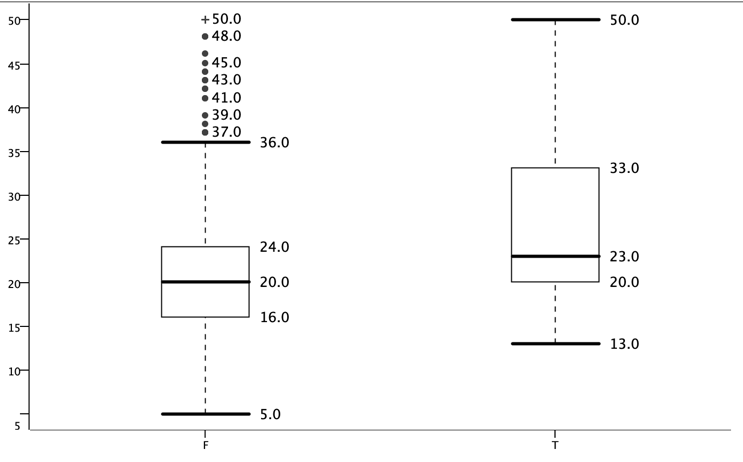
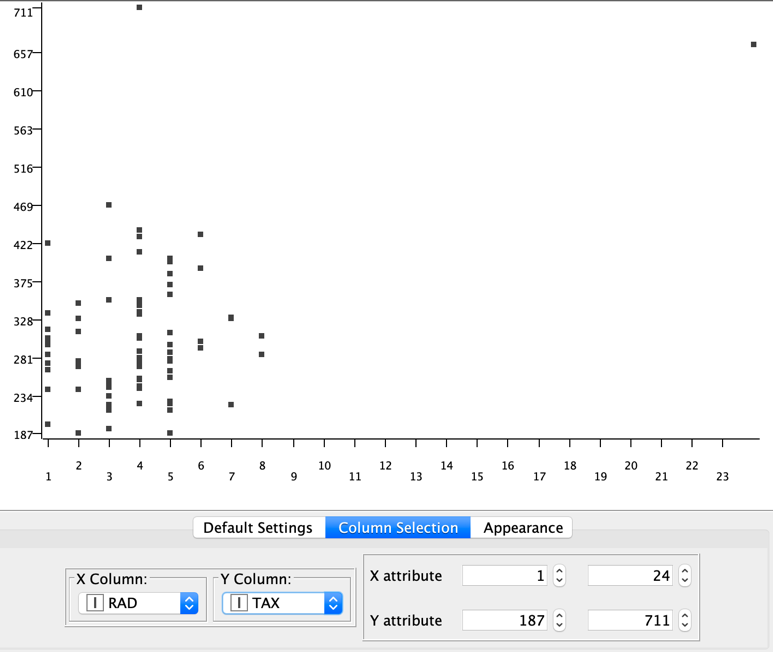
Week 2 실습 리포트

한두솔 (비즈니스 빅데이터 분석, 2018522113)

**실습 2-2**

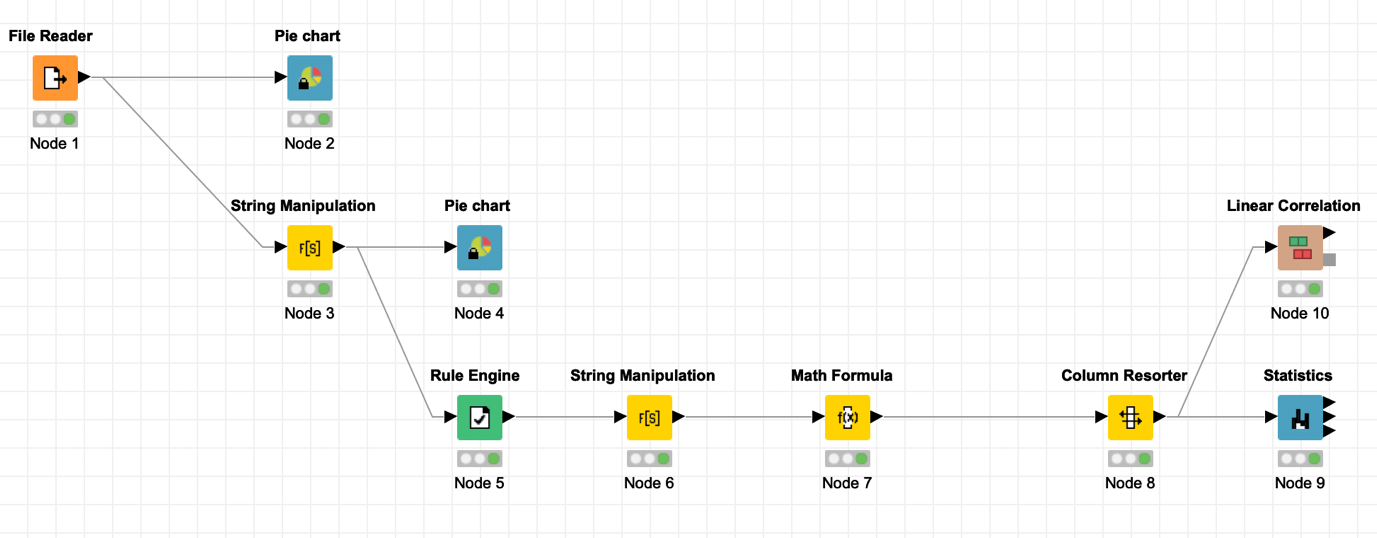
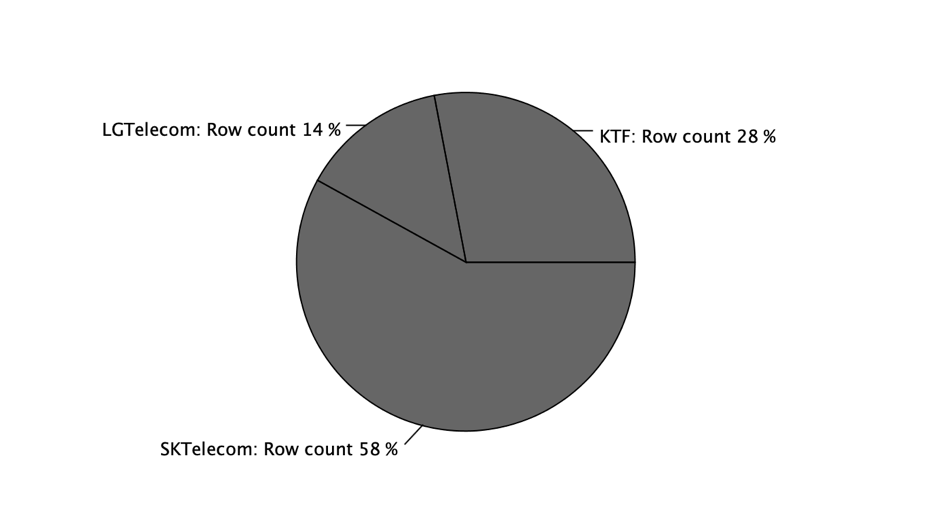
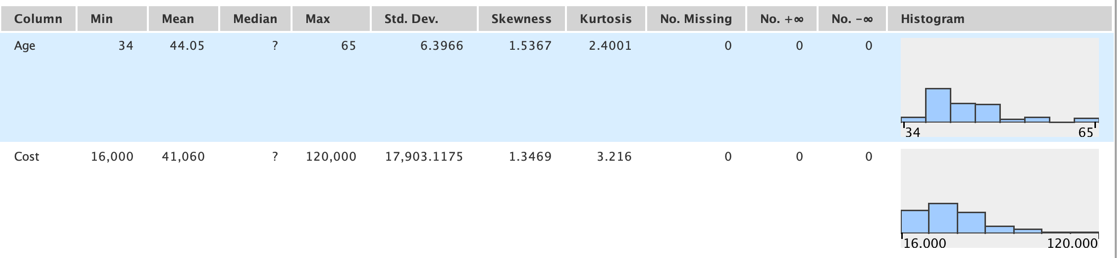
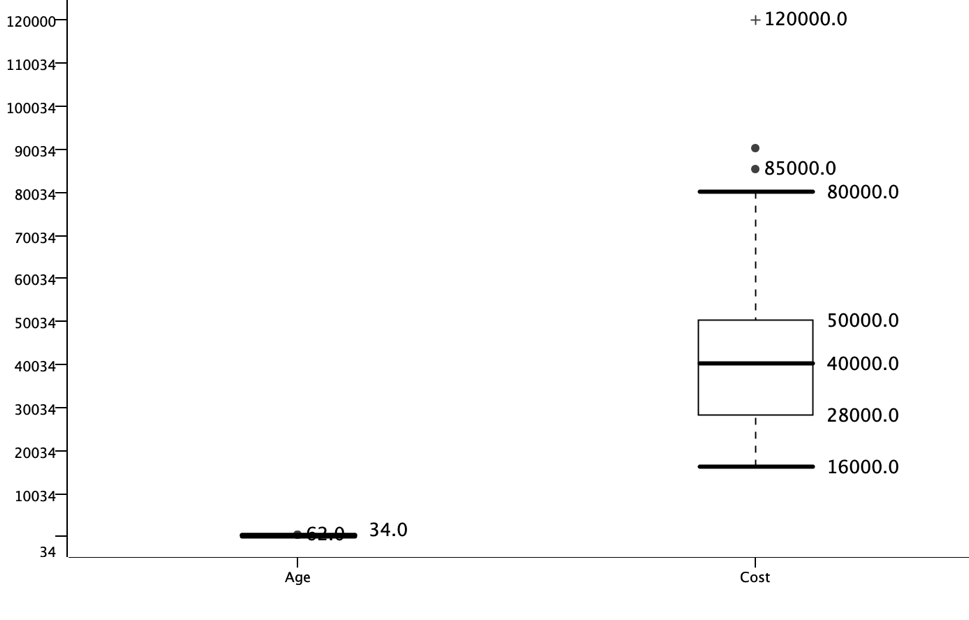
1. 실습 개요
   1. 실습 목적 : 테이블 데이터에서 결측값을 확인하고 대체 또는 제거하는 과정 이해
   2. 실습 데이터 : boston\_missing.csv (506 observations)
      1. Comma separated value 형태의 데이터
      2. 보스턴 주택 데이터로, 주택 가격에 영향을 미치는 변수들이 존재
   3. 실습 과정
      1. Import data
      2. Import 된 데이터를 테이블로 확인 및 이해
      3. Summary Statistics 확인
      4. 데이터 검토 및 결측값 대체
      5. 결측값 데이터 제거
2. Workflow
3. 분석 데이터 특징 사항
   1. 506개의 row가 있고, 20개의 column이 존재하며 그중 3가지의 nominal value column이 있다. CHAS Column에서 6개의 observation이 결측치를 포함하고 있다.
   2. Missing Value 처리는 두가지 방법을 시행 하였다.
      1. 해당 결측치들을 most frequent value로 replace 하였다.
      2. 해당 결측치들을 nominal value filter로 제거하였다.
4. 본 실습 내용의 활용/응용 방안
   1. 데이터를 불러오는 과정에서 사용자의 분석 목적에 따라 Nominal 과 Continuous 값들을 잘 분류할 것.
   2. 결측치 처리를 할때 해당 변수와 dependent variable과의 관계를 잘 파악하고 replace할지 remove할지 결정할 것.

**실습 2-3**

1. 실습 개요
   1. 실습 목적
      1. 테이블 데이터를 시각화하여 데이터 특성 이해
   2. 실습 데이터
      1. Boston.csv
      2. 보스턴 주택 데이터로, 주택 가격에 영향을 미치는 변수들이 존재
   3. 실습 과정
      1. 데이터 불러오기
      2. 불러온 데이터 테이블로 확인하기
      3. 데이터 검토 및 이상치 확인
      4. 변수 별 기초통계 및 변수간 상관관계
      5. 히스토그램
      6. 산포도
      7. 산포도 매트릭스
      8. 박스 플롯
2. 직접 작성한 전체 스트림
3. 데이터 전처리를 통해 파악한 분석 데이터의 특징 사항
   1. Sample Statistics
      1. CRIM (min:0.0063, max:88.9762, mean: 3.6135), ZN (min:0.0, max:100, mean: 11.3636), B (min:0.32, max:396.9, mean:356.674) 칼럼들의 skewness가 높다. 특히 CRIM은positively skewed.
      2. Unusualness를 판별하기 위해 categorical data의 frequency count를 보았더니, CHAS칼럼의 값들이 T와 F중 F에 더 많이 해당이 되는 것을 발견.
   2. Correlation
      1. 변수간 상관관계가 높은 조합
         1. LSTAT – MEDV (-0.7363)
         2. DIS – INDUS (-0.708)
         3. NOX – DIS (-0.7692)
         4. AGE – DIS (-0.7479)
         5. RAD – TAX (0.9102)
         6. CMEDV – MEDV (0.9969)
   3. 데이터 시각화 결과
      1. Histogram(Age)
      2. Scatter Plot(LSTAT – MEDV)
      3. Scatter Matrix (LSTAT,TAX,RAD,DIS,AGE,NOX,INDUS,CMEDV,MEDV)
      4. Box Plot(CHAS-MEDV)
4. 데이터 전처리와 시각화를 통해 파악한 보스턴 주택 데이터의 특징 사항
   1. 변수간 다중 공선성 존재 (rule of thumb: r > 0.9)
      1. MEDV-CMEDV (0.9969)
      2. RAD-TAX (0.9102)
   2. Feature Scaling(Normalization)을 하지 않아서 몇몇 highly correlated 변수들이 수치만큼의 관계성을 산포도로 파악하기 어려움.(RAD-TAX) 
   3. Box-plot에 의하면, 찰스강에서 가깝지 않은 집들의 MEDV 수치가outlier가 많고, 그중 extreme한 outlier도 존재하였다.
5. 본 실습 내용의 활용/응용 방안
   1. 데이터를 수치적으로 파악하는 것 보다 시각화 하여 파악하는게 효율적 이라는 것을 배웠기 때문에 각각의 시각화 방법들을 적재적소에 잘 활용 할 것이다.
   2. 모델 설계 단계 이전에 변수 간 correlation 을 시각화 하기 위해선 normalization이 필요해 보인다.
   3. 모델 목적에 따라 다르겠지만, 일반적인 경우 box plot을 이용해 outlier들을 제거할 것이다.

**실습 2-4**

(OSX 운영체제의 한글 인코딩 문제로 데이터를 영어로 변환시키고 진행했습니다.)

1. 본 실습의 개요
   1. 목적: 데이터 전처리 방식 이해 (분포확인, 공백제거, 변수제거, 새로운변수 생성, 히스토그램)
   2. 실습 데이터
      1. 통신.txt
      2. 통신 서비스 가입자 100명의 id, 주민등록번호, 직종, 가입 통신사, 통화료 데이터
   3. 실습 내용
      1. 데이터 불러오기
      2. 불러온 데이터 확인하기
      3. 불러온 데이터 분포도 확인하기
      4. 데이터 전처리: 공백 제거
      5. 공백 제거한 데이터 분포 확인하기
      6. 주민번호 데이터에서 나이를 추출해 새로운 변수 생성
      7. 주민번호 데이터에서 성별을 추출해 새로운 변수 생성
      8. 전체 데이터의 필드 순서 변경하기
      9. 데이터의 통계적 특성 및 상관관계 파악하기
2. 직접 작성한 전체 스트림
3. 데이터 전처리와 시각화를 통해 파악한 통신서비스가입자 데이터의 특징 사항
   1. Pie Chart를 통해 SKT가입자가 KT+LG보다 더 많은것을 발견
   2. Summary Statistics를 통해 Cost변수(min:16k, max:120k, mean: 41k)에 outlier가 있을 것으로 예상
      1. 이에 따른boxplot시각화를 통해 검증
4. 본 실습 내용의 활용/응용 방안
   1. 텍스트 공백 제거 방법으로 전처리 과정에 지장을 초래하지 않게끔 활용할 계획.
   2. 숫자로 된 String을 numeric으로 변환시켜 산술함수 구현하는 것으로 향후 직관적이지 않을 수도 있는 데이터를 더 직관적인 표현으로 변형시킬 계획.
   3. 칼럼의 순서를 변경하는 방법을 마찬가지로 가독성 향상에 도움을 줄 수 있는 방향으로 활용할 계획.