# SQL Server와 ODBC와 JDBC를 연동하여 고객 데이터 분석 & 시각화

데이터 분석과 조세진

2021. 04. 30

**CONTACT**
whtpwls777@naver.com

# 목차

# 1. 프로젝트 개요

## 각 직원이 담당하는 고객의 수와 연봉의 상관관계

고객의 정보가 담겨있는 customer 테이블과 직원의 정보가 담겨있는 emp 테이블을 이용하여
직원이 관리하는 고객의 수와 직원의 연봉 간에 상관이 있는지 알아봄

EMP TABLE

| 이름 | 널? | 유형 |
|---|---|---|
| EMPNO | NOT NULL | NUMBER(4) |
| ENAME | | VARCHAR2(10) |
| JOB | | VARCHAR2(9) |
| MGR | | NUMBER(4) |
| HIREDATE | | DATE |
| SAL | | NUMBER(7,2) |
| COMM | | NUMBER(7,2) |
| DEPTNO | | NUMBER(2) |

CUSTOMER TABLE

| 이름 | 널? | 유형 |
|---|---|---|
| ID | NOT NULL | VARCHAR2(20) |
| PWD | NOT NULL | VARCHAR2(20) |
| NAME | NOT NULL | VARCHAR2(20) |
| ZIPCODE | | VARCHAR2(7) |
| ADDRESS1 | | VARCHAR2(100) |
| ADDRESS2 | | VARCHAR2(100) |
| MOBILE_NO | | VARCHAR2(14) |
| PHONE_NO | | VARCHAR2(14) |
| CREDIT_LIMIT | | NUMBER(9,2) |
| EMAIL | | VARCHAR2(20) |
| ACCOUNT_MGR | | NUMBER(4) |
| BIRTH_DT | | DATE |
| ENROLL_DT | | DATE |
| GENDER | | VARCHAR2(1) |

DEPT TABLE

| 이름 | 널? | 유형 |
|---|---|---|
| DEPTNO | | NUMBER(2) |
| DNAME | | VARCHAR2(14) |
| LOC | | VARCHAR2(13) |

PROJECT.1

# ODBC를 사용하여 SQL SERVER 연동 후 데이터 분석 및 시각화

## - EXCEL

01

ODBC를 사용하여
Oracle DBMS와 Excel을
연동한 후 데이터 분석 및 시각화

# 2. Excel - Oracle

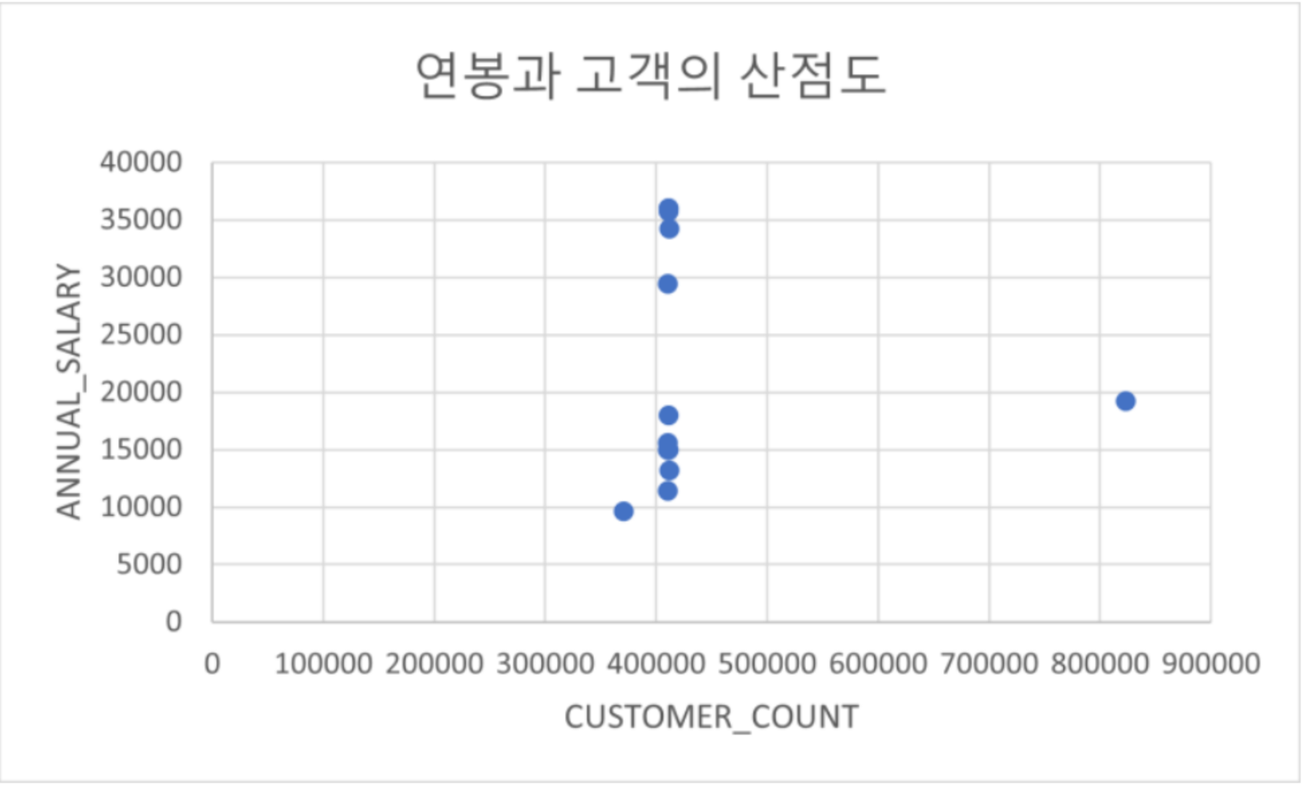
- 데이터 분석 및 시각화

## 사원별 고객의 수와 연봉

- customer table과 emp table을 join한 후

```
select e.ename, count(c.id) as customer_count, avg(sal)*12 as annual_salary
 from CUSTOMER c, emp e
where c.ACCOUNT_MGR = e.EMPNO
group by e.ENAME;
```

- 산점도를 봤을 때 고객의 수와 연봉은
  상관이 없다는 것을 확인할 수 있다.

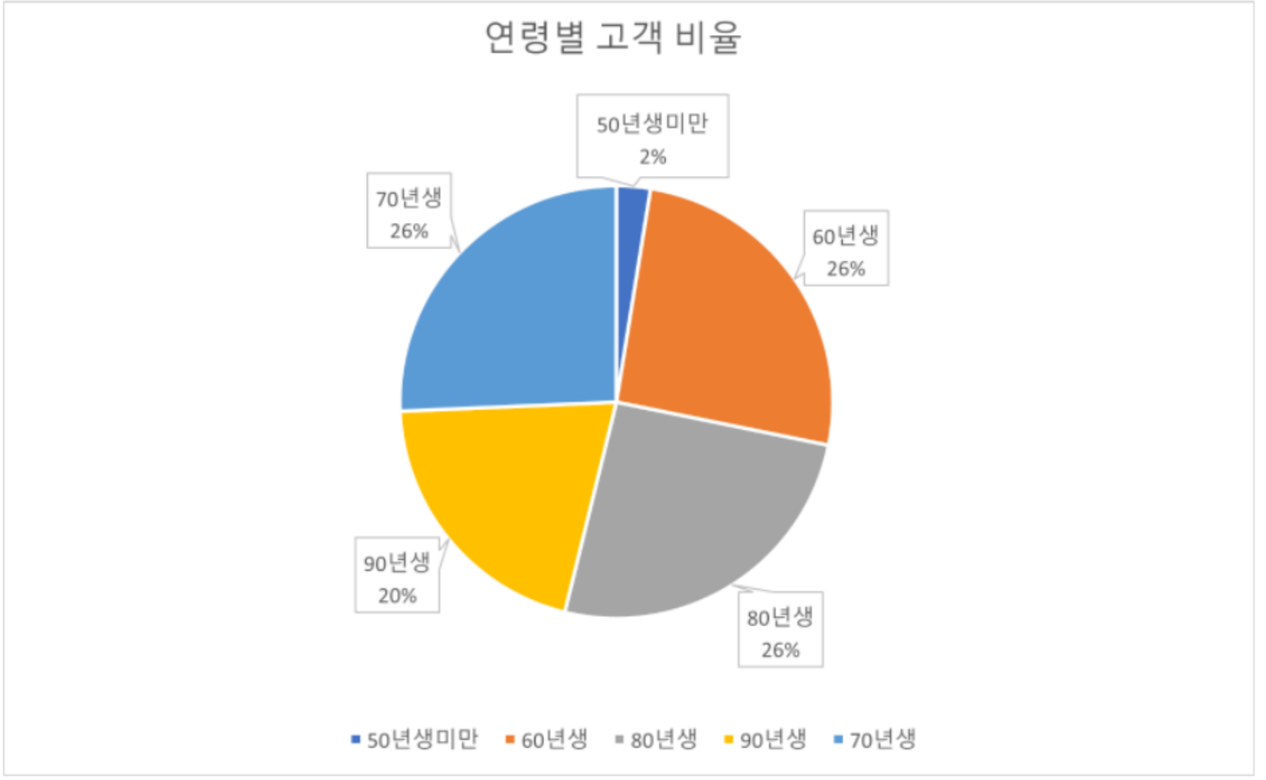


# 2. Excel - Oracle

- 데이터 분석 및 시각화

## 연령대별 고객 비율(주제 외에 고객 테이블 분석)

```
select case
    when birth_dt between '50/01/01' and '59/12/31' then 50 --50년생
    when birth_dt between '60/01/01' and '69/12/31' then 60 --60년생
    when birth_dt between '70/01/01' and '79/12/31' then 70 --70년생
    when birth_dt between '80/01/01' and '89/12/31' then 80 --80년생
    when birth_dt between '90/01/01' and '99/12/31' then 90 --90년생
end as age_group, count(*) as 고객수 from customer group by case
    when birth_dt between '50/01/01' and '59/12/31' then 50
    when birth_dt between '60/01/01' and '69/12/31' then 60
    when birth_dt between '70/01/01' and '79/12/31' then 70
    when birth_dt between '80/01/01' and '89/12/31' then 80
    when birth_dt between '90/01/01' and '99/12/31' then 90
end order by age_group ;
```




- 60-80년생이 가장 많은 비율을 차지하고 있다.

# 2. Excel - Oracle

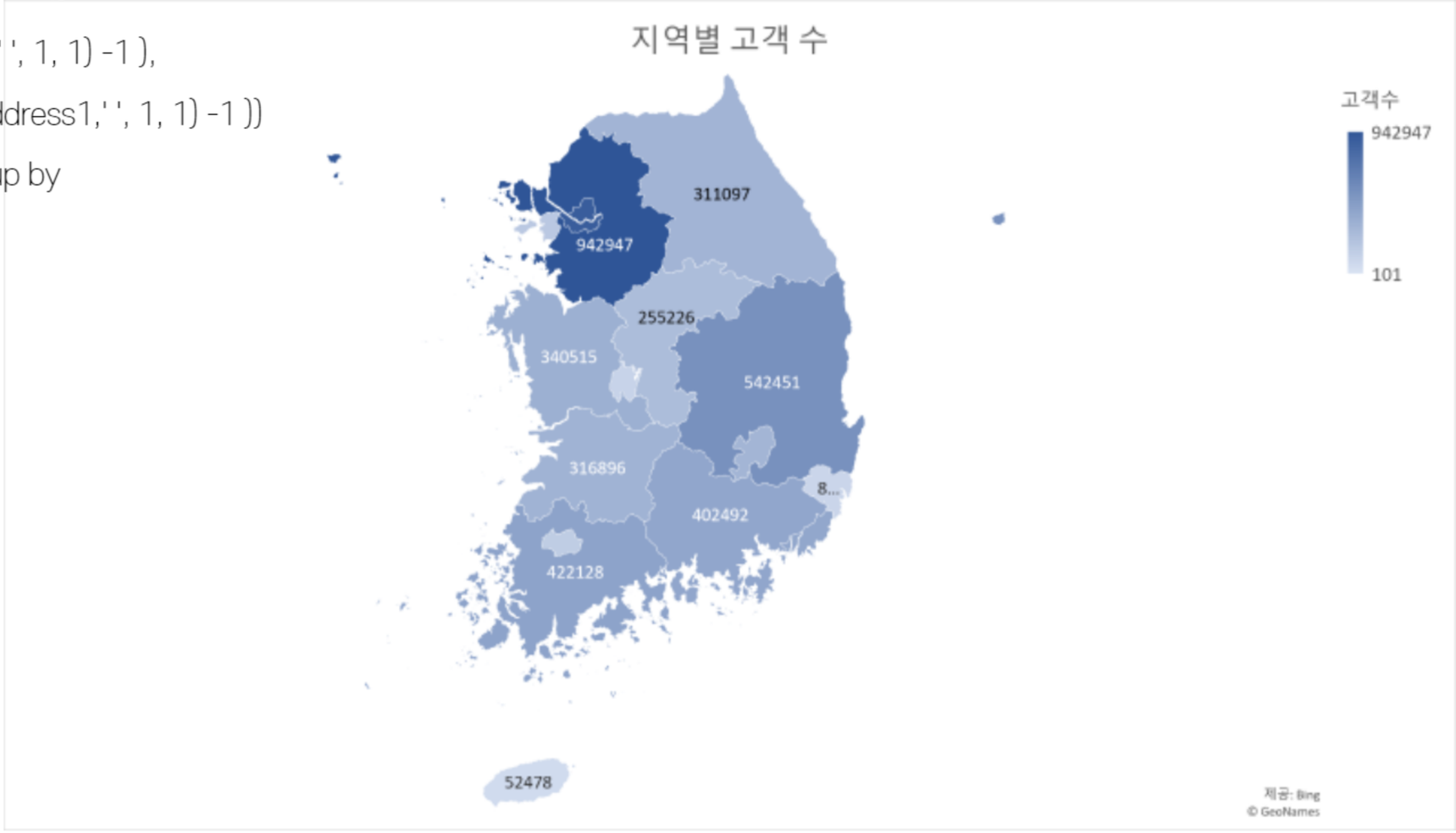
- 데이터 분석 및 시각화

## 지역별 고객 비율(주제 외에 고객 테이블 분석)

```
select decode(substr(address1,1,instr(address1,' ', 1, 1) -1 ),
'uC778천','인천광역시',substr(address1,1,instr(address1,' ', 1, 1) -1 ))
as local, count(*) as "고객수" from customer group by
substr(address1,1,instr(address1,' ', 1, 1) -1 );
```

- 지역은 ADDRESS1 컬럼을 이용하여 데이터 추출
- 인천광역시의 데이터는 'uC778천'로 나와있어서 데이터 처리 후 데이터 출력

- 경기와 서울에 많은 인원이 분포되어 있는 것을 확인할 수 있다.



PROJECT.1

# JDBC를 사용하여 SQL SERVER 연동 후 데이터 분석 및 시각화

- R

02

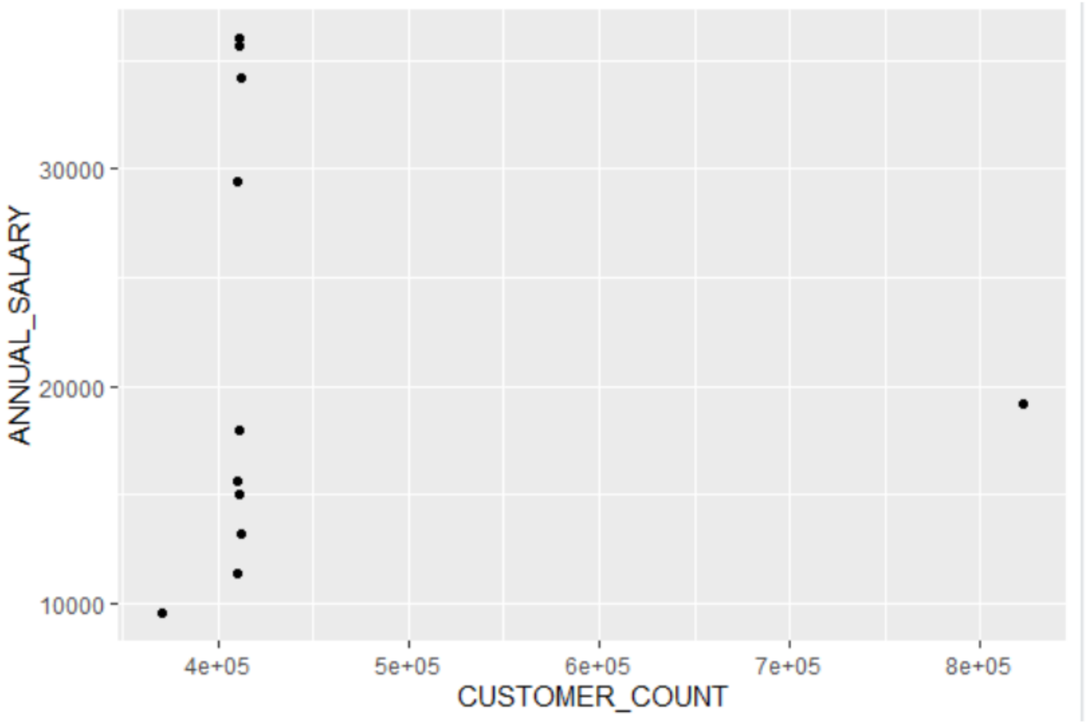
JDBC를 사용하여
R과 Oracle DBMS를
연동한 후 데이터 분석 및 시각화

# 3. R - Oracle

- 데이터 분석 및 시각화

## 사원별 고객의 수와 연봉

```
query <- "select e.ename, count(c.id) as customer_count, avg(sal)*12
as annual_salary from CUSTOMER c, emp e
where c.ACCOUNT_MGR = e.EMPNO
group by e.ENAME"
a <- dbGetQuery(conn, query)
library(ggplot2)
ggplot(data=a,aes(x=CUSTOMER_COUNT, y=ANNUAL_SALARY))+geom_point()
```



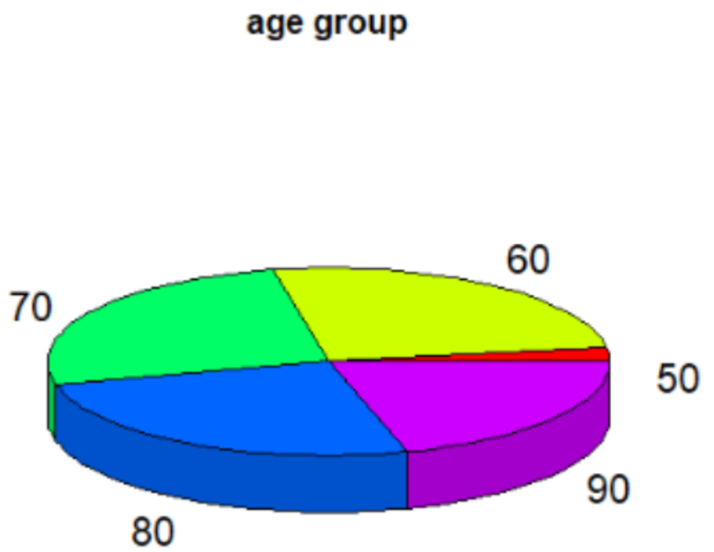

- 산점도를 봤을 때 고객의 수와 연봉은 상관이 없다는 것을 확인할 수 있다.

# 3. R - Oracle

- 데이터 분석 및 시각화

## 연령대별 고객 비율(주제 외에 고객 테이블 분석)

```
query <- "select case when birth_dt between '50/01/01' and '59/12/31' t
hen 50 when birth_dt between '60/01/01' and '69/12/31' then 60 when b
irth_dt between '70/01/01' and '79/12/31' then 70 when birth_dt betwee
n '80/01/01' and '89/12/31' then 80 when birth_dt between '90/01/01' a
nd '99/12/31' then 90 end as age_group, count(*) as customer_count fro
m customer group by case when birth_dt between '50/01/01' and '59/1
2/31' then 50 when birth_dt between '60/01/01' and '69/12/31' then 60
when birth_dt between '70/01/01' and '79/12/31' then 70 when birth_dt
between '80/01/01' and '89/12/31' then 80 when birth_dt between '90/0
1/01' and '99/12/31' then 90 end order by age_group"
library(plotrix)
pie3D(a$CUSTOMER_COUNT, labels=a$AGE_GROUP, main = "age group")
```




- 60-80년생이 가장 많은 비율을 차지하고 있다.

PROJECT.1

# ODBC를 사용하여 SQL SERVER 연동 후 데이터 분석 및 시각화

## - PYTHON

03

ODBC를 사용하여
ORACLE DBMS와 EXCLE을 연동한 후
ORACLE DBMS에 존재하는
고객 데이터 분석 및 시각화

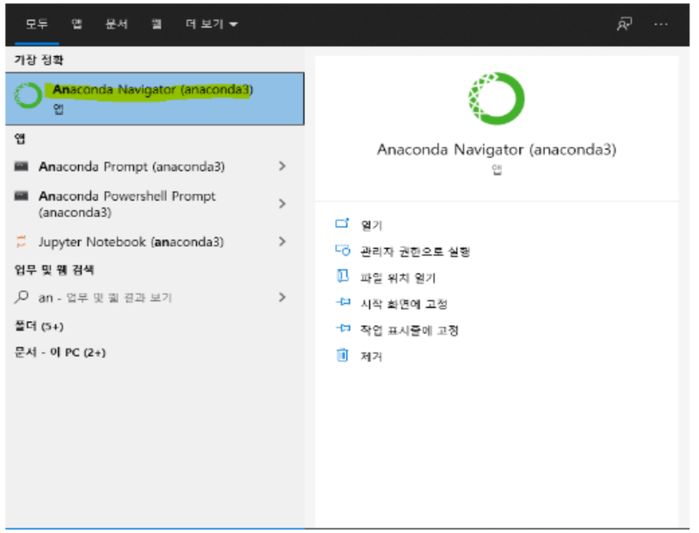
# 4. Python - Oracle

- 데이터 분석 및 시각화

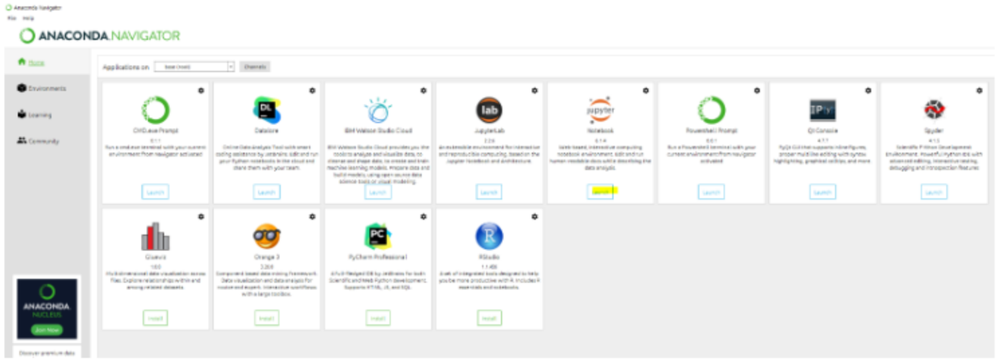
https://www.anaconda.com/products/individual#download-section

아나콘다 설치


아나콘다 실행




Jupyter를 이용하여 Oracle과 연동



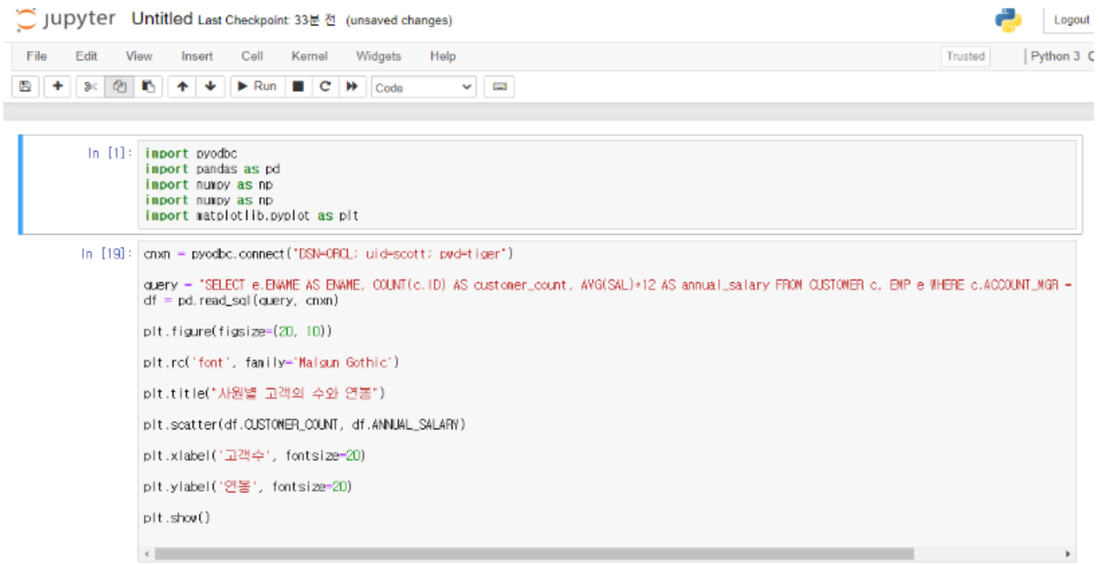

Unitied.ipynb 선택

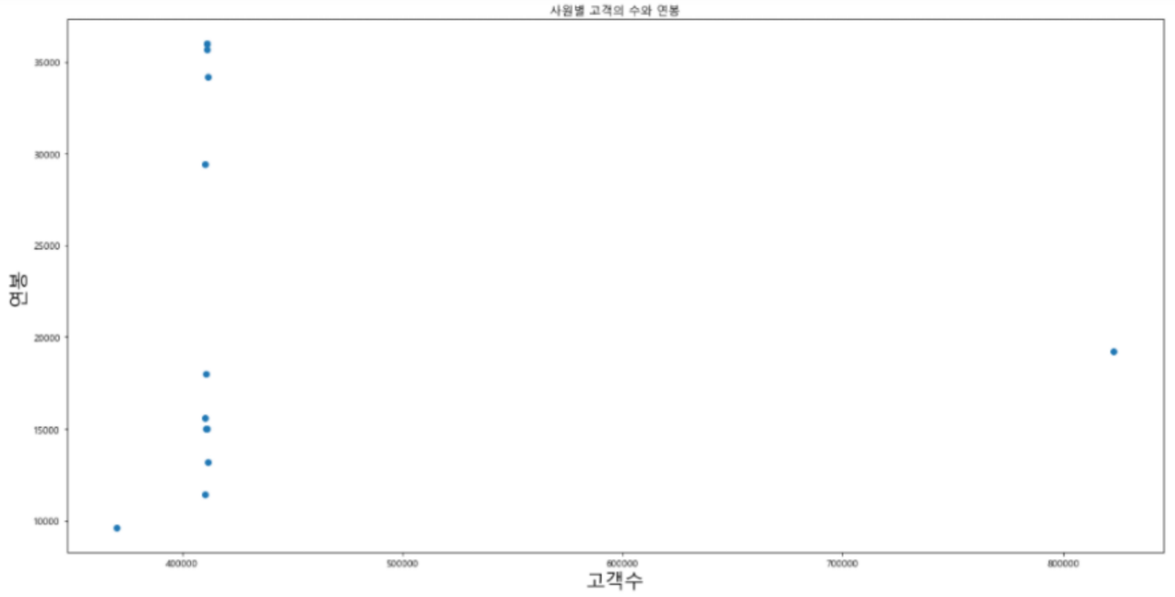
# 4. Python - Oracle

- 데이터 분석 및 시각화

## 사원별 고객의 수와 연봉

```
import pyodbc
import pandas as pd
import numpy as np
import numpy as np
import matplotlib.pyplot as plt
cnxn = pyodbc.connect("DSN=ORCL; uid=scott; pwd=tiger")
query = "SELECT e.ENAME AS ENAME, COUNT(c.ID) AS customer_count, AVG(SAL)*12 AS annual_salary FROM CUSTOMER c, EMP e WHERE c.ACCOUN
T_MGR = e.EMPNO GROUP BY e.ENAME;"
df = pd.read_sql(query, cnxn)
plt.figure(figsize=(20, 10))
plt.rc('font', family='Malgun Gothic')
plt.title("사원별 고객의 수와 연봉")
plt.scatter(df.CUSTOMER_COUNT, df.ANNUAL_SALARY)
plt.xlabel('고객수', fontsize=20)
plt.ylabel('연봉', fontsize=20)
plt.show()
```



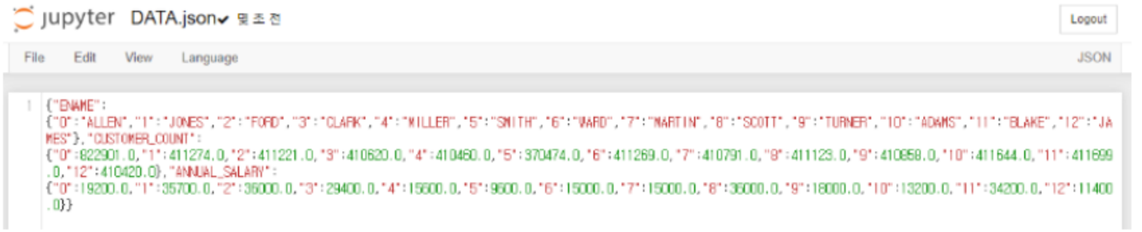
# 4. Python - Oracle

- 데이터 분석 및 시각화

## 분석된 Source 데이터를 JSON로 생성

```
js = df.to_json(orient = 'table')
with open('DATA.json', 'w', encoding='utf-8') as file:
    df.to_json(file, force_ascii=False)
```

실행하면 DATA.json 파일이 생성

# 감사합니다!
# 잘 부탁드립니다!

2021.04. 30

**CONTACT**
whtpwls777@naver.com