#project001_port_stu

2021년 1월 17일 일요일 오후 4:07

☐ 1. Main Question

포루투갈의 2차교육과정에서 학생들의 음주에 영향을 미치는 요소는 무엇인가? 영향을 미친다면 그 정도는 어떠한가?

What are the factors influencing drinking in Portuguese secondary education? If it does, what is its degree?

☐ 2. about DB

#2.1 data source ↓ (kaggle.com)

https://www.kaggle.com/uciml/student-alcohol-consumption

#2.2 data_summary

Student Alcohol Consumption

Social, gender and study data from secondary school students

$ cat student-mat.csv | wc -l --> 395 rows

#2.3 location / posting date / group

Porto, Portugal / April, 2008 / secondary school

#2.4 file location

/home/scott/Database/port_stu/student-mat.csv

#2.5 file processing

delete column name (first line delete)

#2.6 column list used

COLUMN	EXAMPLE	DETAILS
sex,	F,	2. sex - student's sex (binary: 'F' - female or 'M' - male)
age,	18,	3. age - student's age (numeric: from 15 to 22)
Pstatus,	A,	6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
failures,	0,	15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
famrel,	4,	24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
Dalc,	1,	27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc,	1,	28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
G1,	5,	31. G1 - first period grade (numeric: from 0 to 20)
G2,	6,	32. G2 - second period grade (numeric: from 0 to 20)
G3	6,	33. G3 - final grade (numeric: from 0 to 20, output target)

☐ 3. MariaDB

#3.1 create table in MariaDB

==========

```
MariaDB [orcl]>
drop table port_stu;

create table port_stu
(school  varchar(20),
sex  varchar(20),
age  int(10),
address  varchar(20),
famsize  varchar(20),
Pstatus  varchar(20),
Medu  int(10),
Fedu  int(10),
Mjob  varchar(20),
Fjob  varchar(20),
reason  varchar(20),
guardian  varchar(20),
traveltime  int(10),
studytime  int(10),
failures  int(10),
schoolsup  varchar(20),
famsup  varchar(20),
paid  varchar(20),
activities  varchar(20),
nursery  varchar(20),
higher  varchar(20),
internet  varchar(20),
romantic  varchar(20),
famrel  int(10),
freetime  int(10),
goout  int(10),
Dalc  int(10),
Walc  int(10),
health  int(10),
absences  int(10),
G1  int(10),
G2  int(10),
G3  int(10) );
==================================
```

#3.2 load data

```
=========================================================================
MariaDB [orcl]>
LOAD DATA LOCAL INFILE '/home/scott/Database/port_stu/student-mat.csv'
REPLACE
```

```
INTO TABLE  orcl.port_stu
fields TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '₩n';
commit;
```

==========

#3.3 data processing(SQL)

```
<1>
select Pstatus, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu group by Pstatus;
<2>
select sex, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu group by sex;
<3>
select failures, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu group by failures;
<4>
select G1+G2+G3, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu group by G1+G2+G3;
<5>
select age, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu group by age;
<6>
select famrel, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu group by famrel;
```

☐ 4. matplotlib in Linux_python

==========

```
# SQL script
# import module
import mysql.connector
import pandas as  pd
import matplotlib.pyplot as plt

# MariaDB ---> Linux_Python
config = {
    "user": "root",
    "password": "1234",
    "host": "192.168.56.101", # local
    "database": "orcl", # Database name
    "port": "3456" # default port:3306
```

```
}
conn = mysql.connector.connect(**config)    # connect code
cursor = conn.cursor()    # db select, insert, update, delete OBJECT

# SQL script
sql1 = """
select Pstatus, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu
  group by Pstatus;
""" # cast(A as double) ---> **transform datatype decimal to double

sql2 = """
select sex, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu
  group by sex;
"""

sql3 = """
select failures, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu
  group by failures;
"""

sql4 = """
select G1+G2+G3, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu
  group by G1+G2+G3;
"""

sql5="""
select age, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu
  group by age;
"""

sql6="""
select famrel, cast( truncate(avg(Dalc+Walc)/2*100, 0) as double) as res
  from port_stu group by famrel;
"""

# cursor select / processing
cursor.execute(sql1)    # exe sql script into cursor(memory)
resultList1 = cursor.fetchall()    # tuple list
cursor.execute(sql2)
resultList2 = cursor.fetchall()
```

```
cursor.execute(sql3)
resultList3 = cursor.fetchall()
cursor.execute(sql4)
resultList4 = cursor.fetchall()
cursor.execute(sql5)
resultList5 = cursor.fetchall()
cursor.execute(sql6)
resultList6 = cursor.fetchall()


# Pandas
df1 = pd.DataFrame(resultList1)    # transform resultList into Pandas
df2 = pd.DataFrame(resultList2)
df3 = pd.DataFrame(resultList3)    # df=df.astype(float) ↓ ↓ ↓
df4 = pd.DataFrame(resultList4)    # transform datatype to float(err_sol)
df5 = pd.DataFrame(resultList5)
df6 = pd.DataFrame(resultList6)

# matplotlib
plt.subplot(231)    # nrows,ncols,index(Z_shape)
plt.bar(df1[0],df1[1],color='red',width=0.6)
plt.xlabel('famrel₩n')
plt.ylabel('Alcohol')
plt.yticks(range(0,301,100))
x=list(df1[0])
y=list(df1[1])
for i, v in enumerate(x):
    plt.text(v, y[i], y[i],     # 좌표 (x축 = v, y축 = y[0]..y[1], 표시 = y[0]..y[1])
             fontsize = 10,
             color='black',
             horizontalalignment='center',    # horizontalalignment (left, center, right)
             verticalalignment='bottom')    # verticalalignment (top, center, bottom)

plt.subplot(232)
plt.bar(df2[0],df2[1],color='blue')
plt.title('#Variances influencing to Alcohol ₩n')
plt.xlabel('Sex₩n')
plt.ylabel('Alcohol')
plt.yticks(range(0,301,100))
x=list(df2[0])
y=list(df2[1])
for i, v in enumerate(x):
    plt.text(v, y[i], y[i],
             fontsize = 10,
             color='black',
```

```
            horizontalalignment='center',
            verticalalignment='bottom')

plt.subplot(233)
plt.barh(df3[0],df3[1],color='slategray')
plt.ylabel('Failures₩n')
plt.xlabel('Alcohol')
plt.yticks(range(0,4,1))
plt.xticks(range(0,301,100))

plt.subplot(234)
plt.scatter(df4[0],df4[1],color='teal',s=10)
plt.xlabel('Grades per year₩n')
plt.ylabel('Alcohol')
plt.xticks(range(0,61,10))
plt.yticks(range(0,501,100))

plt.subplot(235)
plt.plot(df5[0],df5[1],color='pink')
plt.grid()
plt.xlabel('age₩n')
plt.ylabel('Alcohol')
plt.xticks(range(15,23,2))
plt.yticks(range(0,501,100))


plt.subplot(236)
plt.bar(df6[0],df6[1],color='#FF8200')
plt.xlabel('Family relationship₩n')
plt.ylabel('Alcohol')
plt.xticks(range(1,6,1))
plt.yticks(range(0,301,100))
x=list(df6[0])
y=list(df6[1])
for i, v in enumerate(x):
    plt.text(v, y[i], y[i],
            fontsize = 6,
            color='black',
            horizontalalignment='center',
            verticalalignment='bottom')


plt.tight_layout()
plt.show()
```

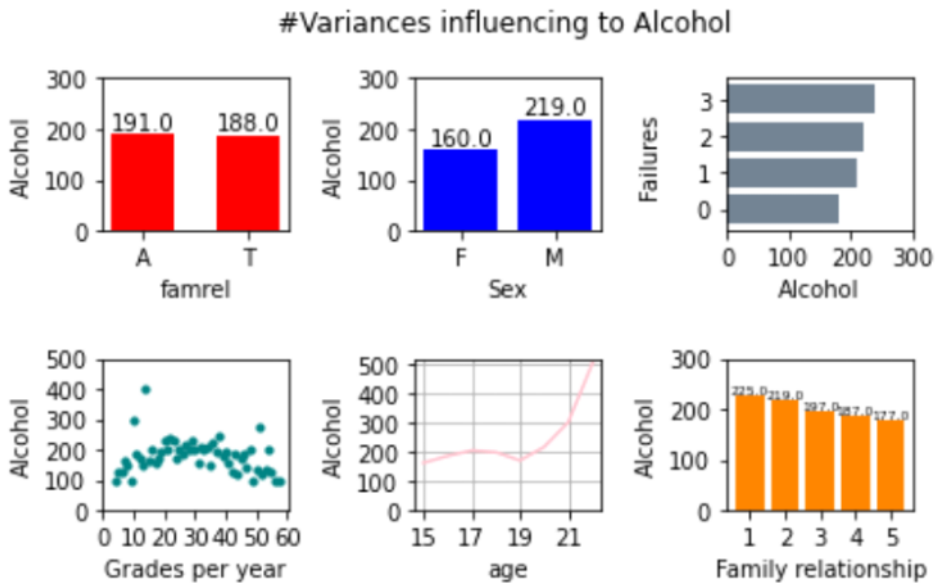
#[graph]
#https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.plot.html
#https://matplotlib.org/3.1.1/tutorials/introductory/sample_plots.html
==========

☐ 5. explain result : conculusion

학생들의 음주에 미치는 영향을 조사하기 위해 6가지의 변수 후보들을 선정하였다.
독립변수 : famrel(부모거주여부), sex(성별), failures(수업낙제횟수), grade(G1+G2+G3 : 연간성적), age(15~22), famrel(가족관계)
종속변수 : Alcohol((Dalc+Walc)/2*100 : 1주간 알코올 섭취정도)
이 변수들 중 grade 는 0~60 (60점이 고득점), Alcohol 은 0~500 (100:매우낮음, 500:매우높음)으로 가공하여 사용하였다.
나머지 변수들은 위의 2.6 을 참고하여 그래프를 해석할 수 있다.

fermer 변수를 보면 A:apart 는 191, T:together 는 188 로 매우 근소한 차이를 보이고 있다.
즉, 부모의 거주결합상태는 학생들의 음주에 큰 영향을 미치지 않는 것으로 판단할 수 있다.
sex 변수는 여자가 160, 남자가 219의 수치로 남학생이 평균적으로 더 많은 양의 음주를 하는 것으로 파악할 수 있다.
failures 변수는 낙제한 과목이 많을수록 (0:181, 1:209, 2:220, 3:237)로 수치가 높아지는 것을 볼 수 있다.
이는 낙제한 과목이 많을수록 더 음주를 하는 경향을 보여준다.
grade(그래프에서는 Grades per year) 변수는 0점(최하점)부터 60점까지의 Alcohol 산포도를 나타내고 있다.
흥미로운 점은 0~20 점대는 점수가 증가함에 따라 음주량이 증가하며 20~40점대는 음주량의 변화가 미세하고,
40~60점대는 점수증가에 따라 음주량이 감소하는 것을 볼 수 있다. 이는 일정 수준 이하로 음주를 하는 집단은 음주를 하는 집단에 비해 성적이 높거나, 낮은 부분에 속해 있다. 즉, 음주량이 보다 적은 집단은 어떤 다른 요인에 따라 성적이 상위권이거나, 하위권에 드는 특성을 보인다고 추측할 수 있다.
age 변수는 Alcohol 변수와 양의 관계를 보여준다. 18,19세의 경우를 제외하면 나이가 증가함에 따라 알코올 섭취량이 증가한다. 다른 변수들에 비해 Alcohol 변수의 변화폭도 크기 때문에 가장 강력한 변수로 선정할 수 있다.
famrel 변수는(그래프에서 Family relationship)가족간의 관계가 좋을수록 학생들의 음주경향이 감소하는 것을 보여준다.

결론적으로, 포루투갈 학생들의 음주는 나이의 증가와 강한 연관성을 가지고 있으며, 성별이 남자이고, 낙제과목이 많아질수록, 가족간 관계가 안 좋을수록 음주량이 증가한다. 부모의 결합상태는 큰 영향이 없으며, 40점 이상의 연간성적 범주에서는 성적과 음주량이 반비례, 20점 이하에서는 성적과 음주량이 비례, 20점과 40점 사이에서는 변화가 없다.


#Variances influencing to Alcohol
Alcohol
191.0
188.0
A
T
famrel
160.0
219.0
F
M
Sex
Failures
Alcohol
Grades per year
age
Family relationship

===end line===