School Data Sample

November 15, 2017

```
In [13]: import pandas as pd
         import numpy as np
```

Event-level data: an attendance log for every student in a school district date | student_id | attendance Dimension-level data: a summary table with demographics for each student in the district student_id | school_id | grade_level | date_of_birth | hometown Using this data, I want to find the following information: What was the overall attendance rate for the school district yesterday? Which grade level currently has the most students in this school district? Which school had the highest attendance rate? The lowest?

```
In [38]: attendance=pd.read_excel("schools.xlsx",sheetname=0)
         attendance.head()
```

```
Out[38]:         date  student_id  attendance
         0 2017-10-28           1           1
         1 2017-10-29           1           1
         2 2017-10-30           1           1
         3 2017-10-31           1           1
         4 2017-11-01           1           1
```

```
In [48]: schools=pd.read_excel("schools.xlsx",sheetname=1)
         schools.head()
```

```
Out[48]:    student_id  school_id  grade_level data_of_birth hometown
         0           1          1           10    1995-04-02   Austin
         1           2          1           10    1995-04-24  Houston
         2           3          2           10    1995-04-24   Austin
         3           4          2           11    1995-04-30  Houston
```

What was the overall attendance rate for the school district yesterday?

```
In [35]: (attendance.loc[attendance["date"]=="2017-10-28","attendance"].mean())
```

```
Out[35]: 0.5
```

Which school had the highest attendance rate? The lowest?

```
In [ ]: Which school had the highest attendance rate? The lowest?
```

```
In [39]: result = (attendance.merge(schools[["school_id","student_id"]])
                   .groupby(["school_id","student_id"])[["attendance"]]
                   .mean()
                   .groupby("school_id")
                   .mean()
                  )
         result
```

```
Out[39]:            attendance
         school_id
         1                 0.9
         2                 0.5
```

Which grade level currently has the most students for school district 1?

```
In [52]: (schools.loc[schools["school_id"] == 1,:]
                   .groupby("grade_level")["student_id"]
                   .count()).to_frame()
```

```
Out[52]:              student_id
         grade_level
         10                    2
```