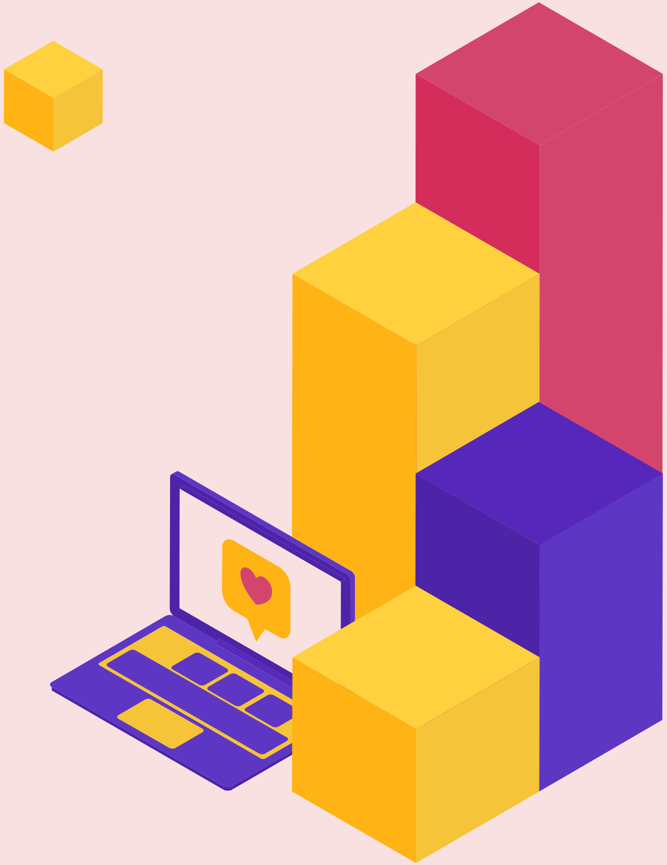


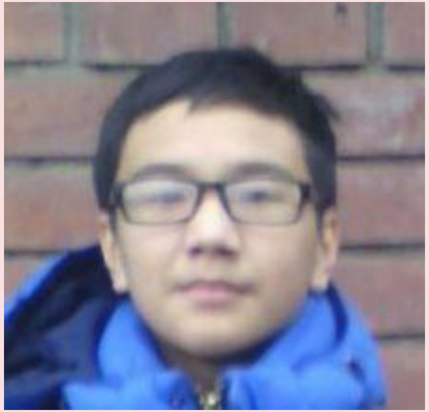
Exam results analysis

Data mining project

Our team

Sultan Kamliyev

Team Lead



Dias Kosmagul

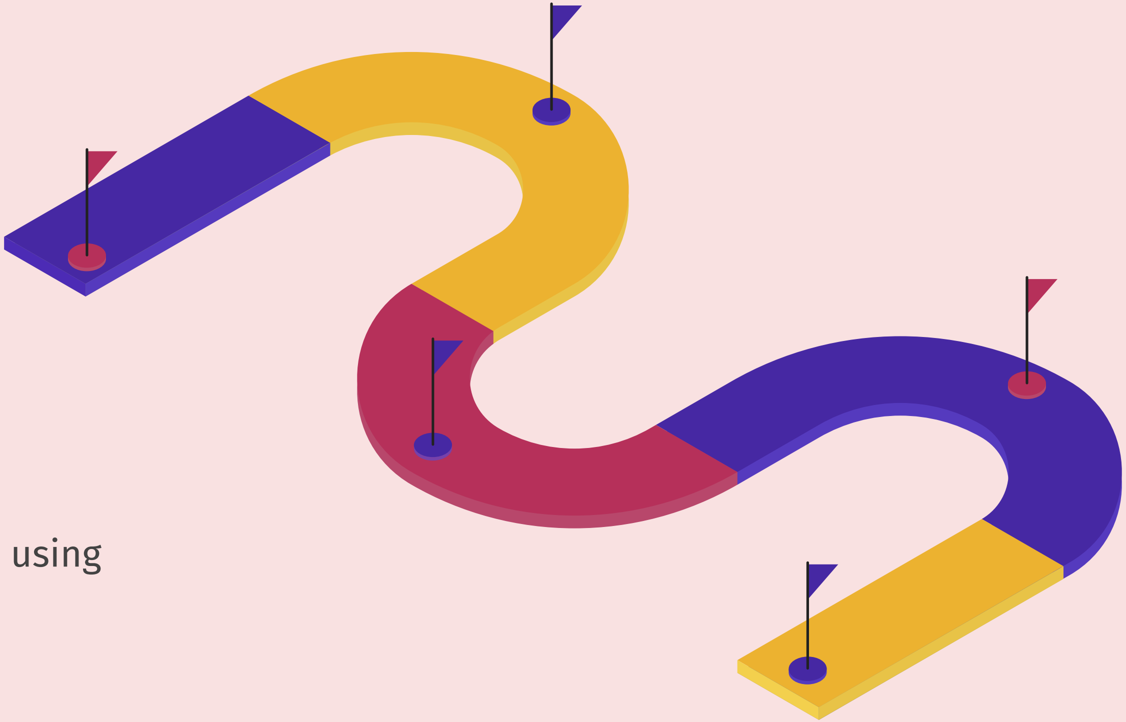
Senior Developer

Daulet Seitzhaparov

Team Spirit Keeper

Mission statement

Classify students into ranks using their results



Using materials

Jupyter

Useful notebook that is really comfortable for work with datasets

Kaggle

Open platform that consists many datasets

Stack overflow

question and answer website for professional and enthusiast programmers

Used materials

Basic	Libraries	
Python	Seaborn	Numpy
Statistics	Matplotlib	Sklearn
Machine learning	Pandas	

Few words about dataset

The dataset includes scores from three exams and a variety of personal, social, and economic factors that have interaction effects upon them.

Exams are:

- Mathematics
- Reading
- Writing

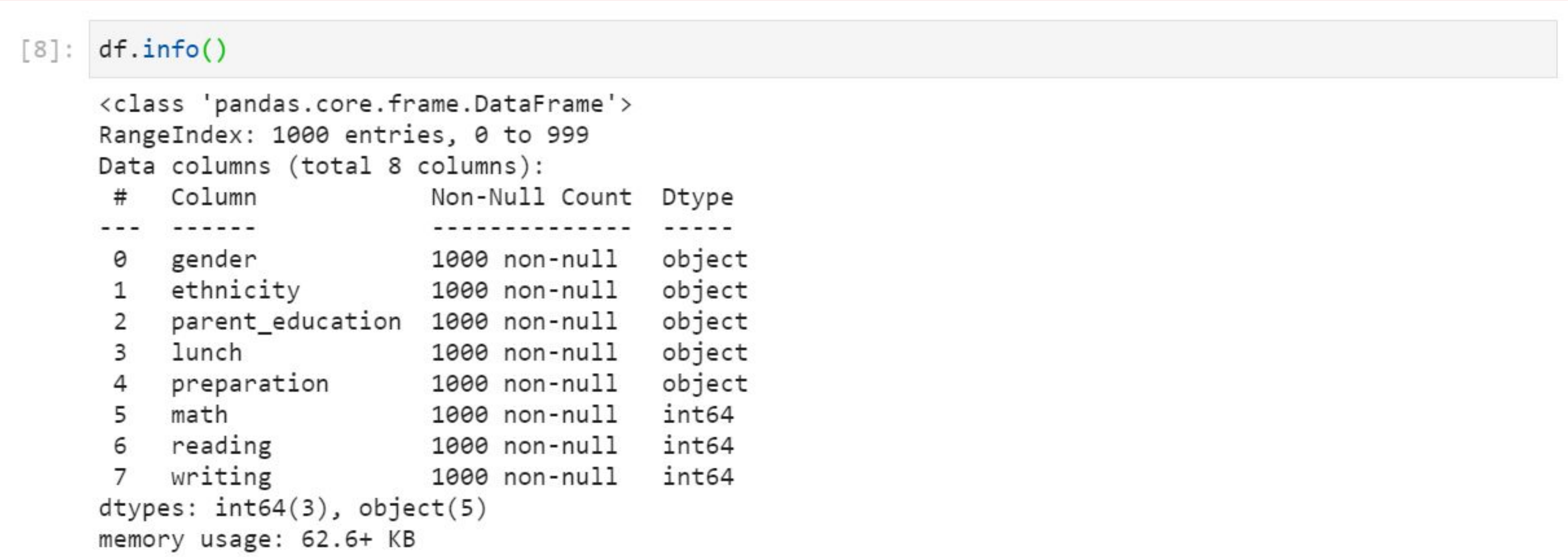
1,000

Number of students

Dataset info

- Gender = Gender
- Ethnicity = Group
- Parent education = Parental degree of education (college, bachelor, master etc.)
- Lunch = Did the student get lunch before exams
- Preparation = Did students complete preparation for the exams
- Math = Result for math exam
- Reading = Results for reading exam
- Writing = Results for writing exam

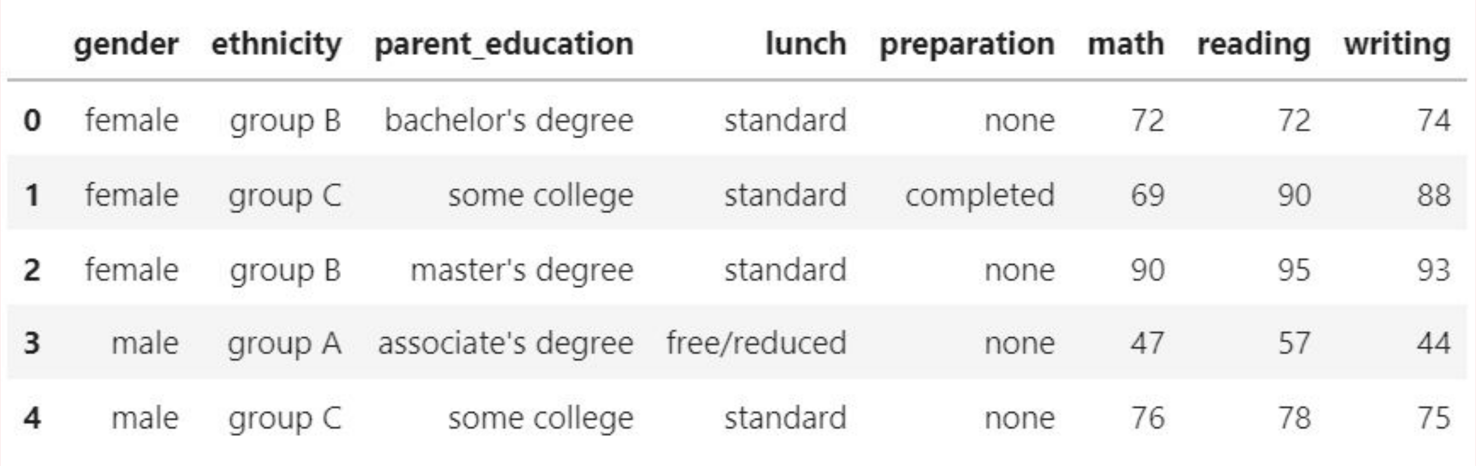
Dataset info



```
[8]: df.info()
```

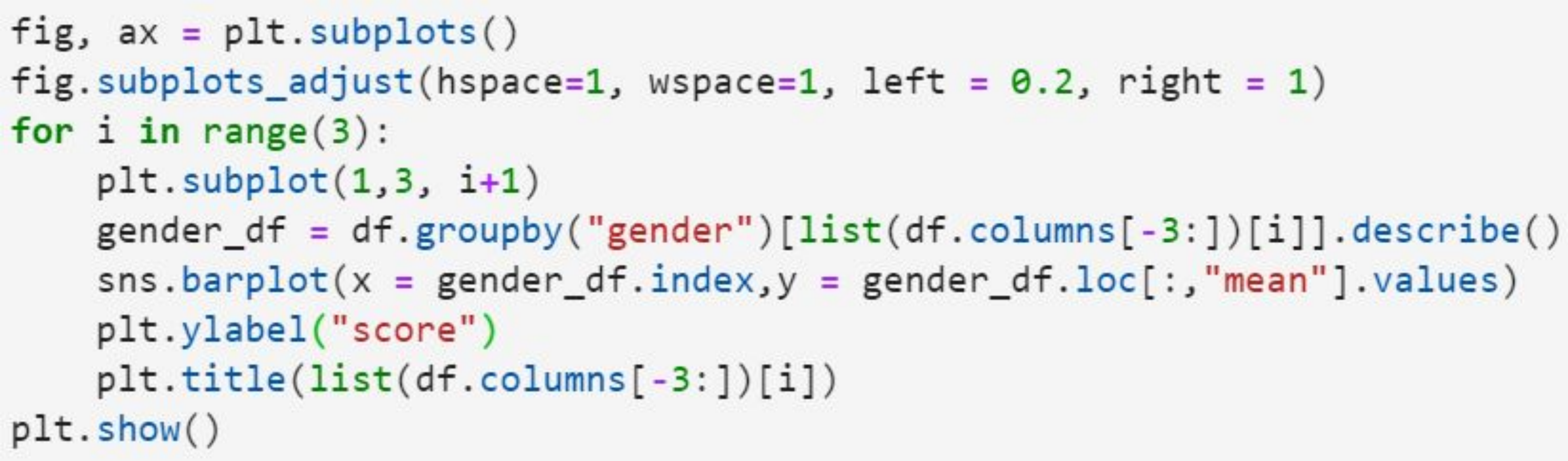
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   gender            1000 non-null   object
 1   ethnicity         1000 non-null   object
 2   parent_education  1000 non-null   object
 3   lunch             1000 non-null   object
 4   preparation       1000 non-null   object
 5   math              1000 non-null   int64 
 6   reading           1000 non-null   int64 
 7   writing           1000 non-null   int64 
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Dataset head



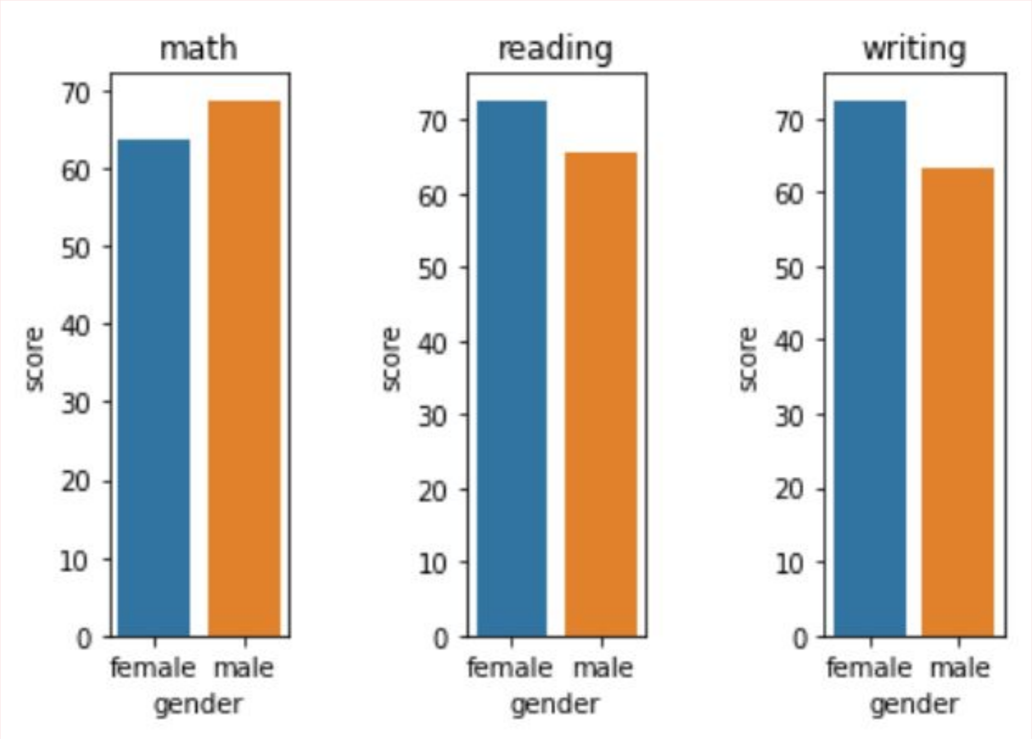
	gender	ethnicity	parent_education	lunch	preparation	math	reading	writing
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Performance for each field by gender

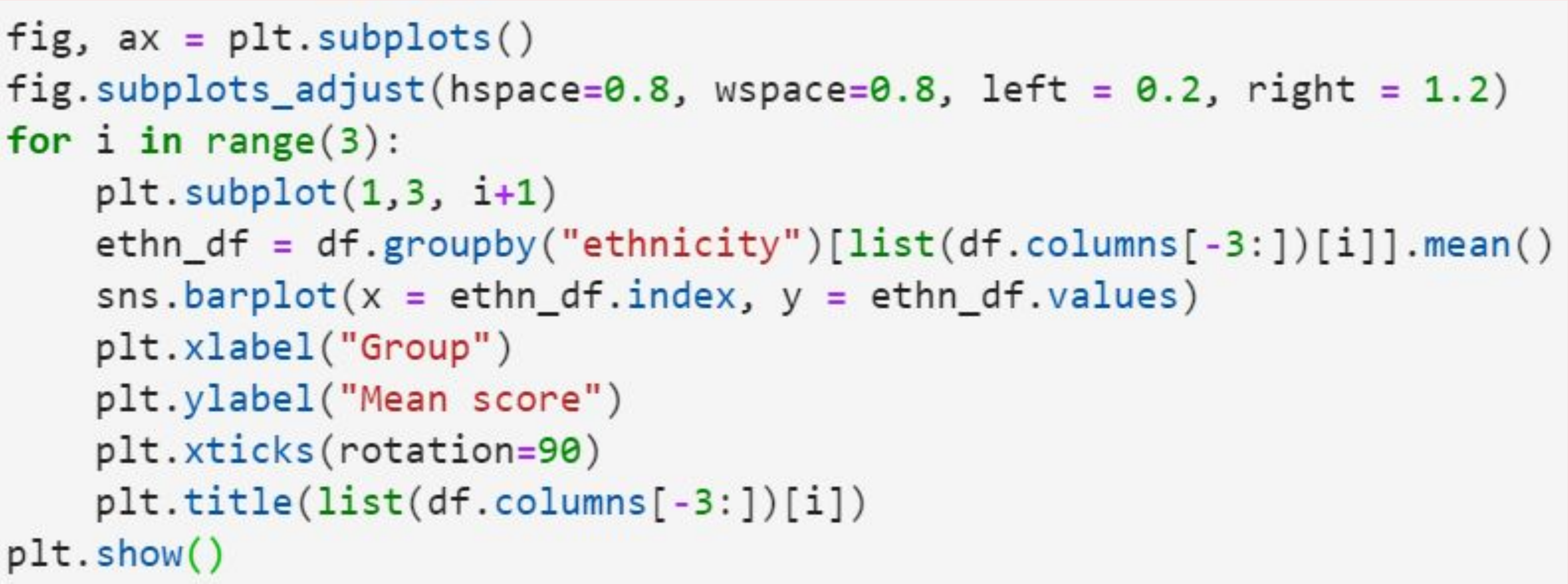


```python
fig, ax = plt.subplots()
fig.subplots_adjust(hspace=1, wspace=1, left = 0.2, right = 1)
for i in range(3):
    plt.subplot(1,3, i+1)
    gender_df = df.groupby("gender")[list(df.columns[-3:])[i]].describe()
    sns.barplot(x = gender_df.index,y = gender_df.loc[:,"mean"].values)
    plt.ylabel("score")
    plt.title(list(df.columns[-3:])[i])
plt.show()
```

Performance for each field by gender

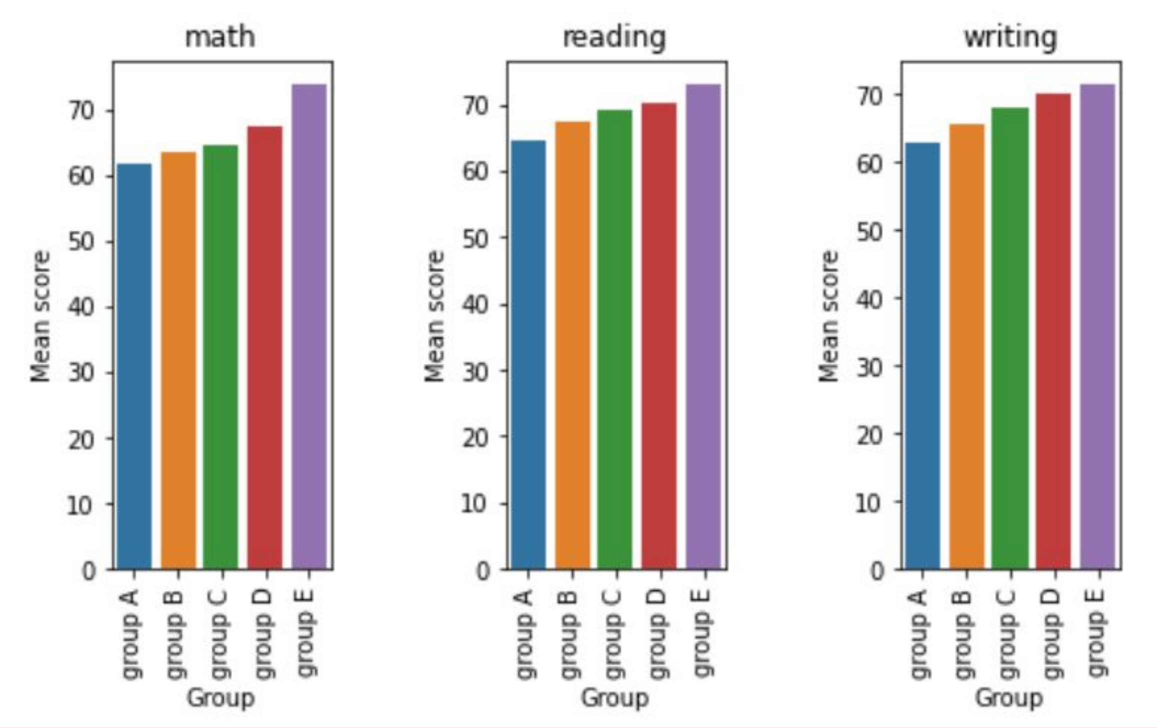


Performance of each group

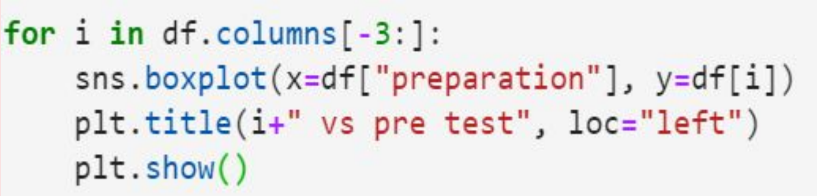


```
fig, ax = plt.subplots()
fig.subplots_adjust(hspace=0.8, wspace=0.8, left = 0.2, right = 1.2)
for i in range(3):
    plt.subplot(1,3, i+1)
    ethn_df = df.groupby("ethnicity")[list(df.columns[-3:])[i]].mean()
    sns.barplot(x = ethn_df.index, y = ethn_df.values)
    plt.xlabel("Group")
    plt.ylabel("Mean score")
    plt.xticks(rotation=90)
    plt.title(list(df.columns[-3:])[i])
plt.show()
```

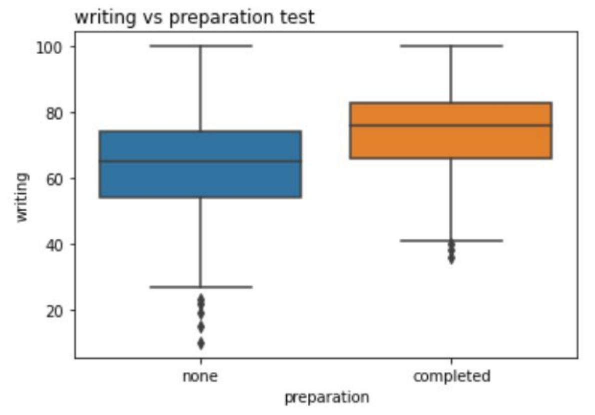
Performance of each group



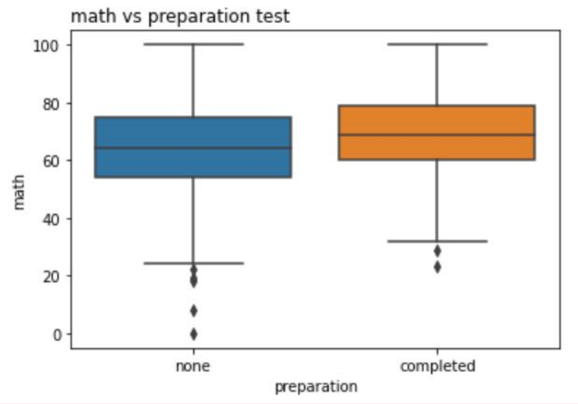
Comparison results and preparation



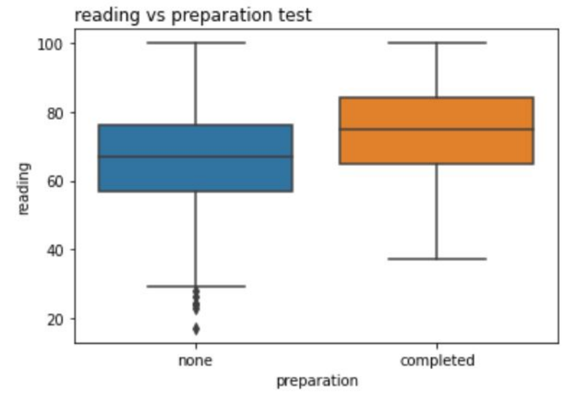
```
for i in df.columns[-3:]:
    sns.boxplot(x=df["preparation"], y=df[i])
    plt.title(i+" vs pre test", loc="left")
    plt.show()
```



writing vs preparation test

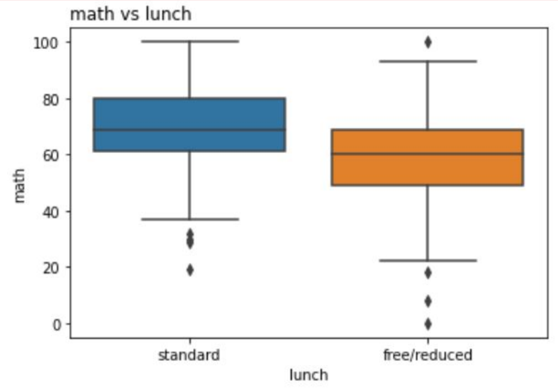


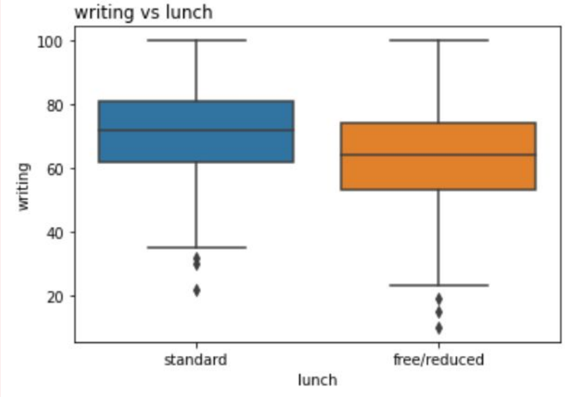
math vs preparation test

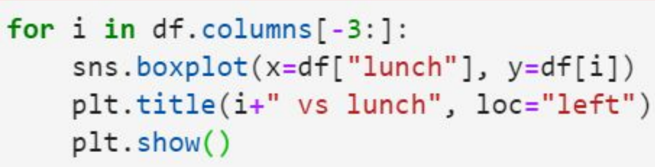


reading vs preparation test

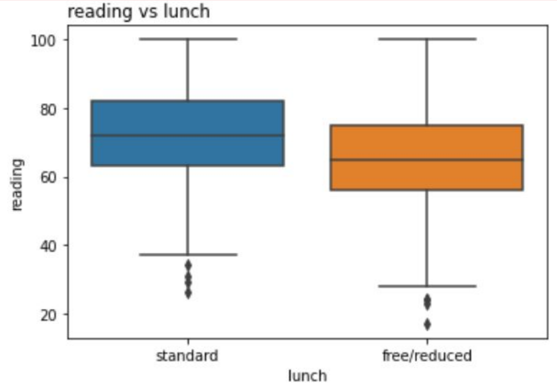
Comparison results and lunch

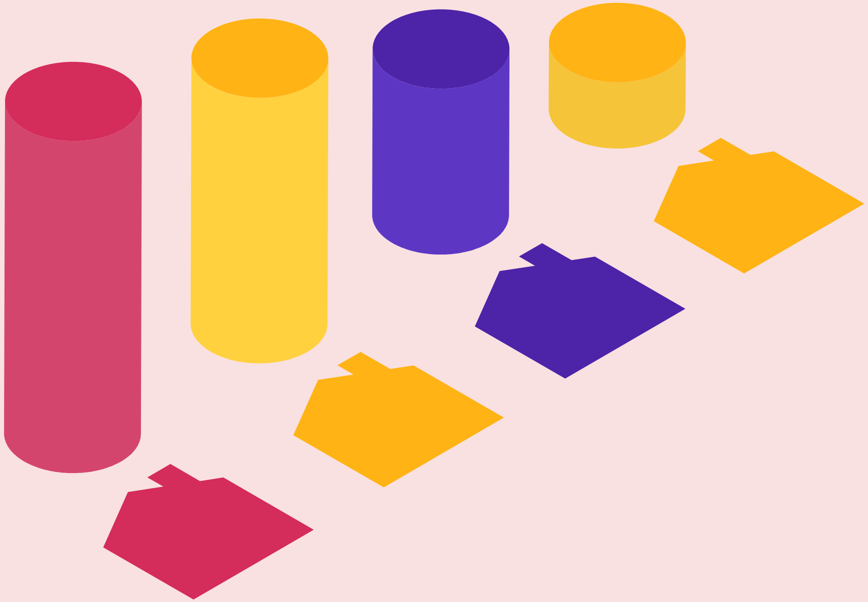






```python
for i in df.columns[-3:]:
    sns.boxplot(x=df["lunch"], y=df[i])
    plt.title(i+" vs lunch", loc="left")
    plt.show()
```





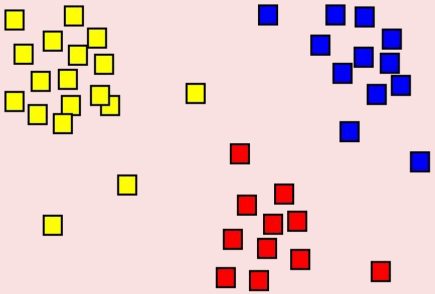
02

K-Means Clustering

What is Clustering?

Clustering loosely defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters. Clustering **helps identify two qualities** of data:

1. **Meaningfulness**
2. **Usefulness**



Examples:

1. **Partitional clustering**
2. **HIerarchical clustering**
3. **Density-based clustering**

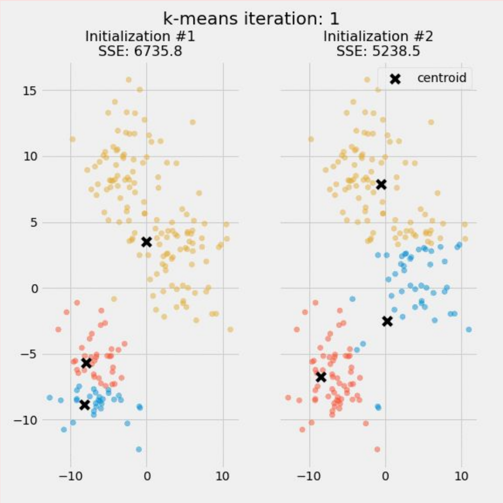
K-Means Clustering

K-Means Clustering method is an unsupervised machine learning technique used to identify clusters of data objects in dataset. K-Means is one the oldest and most approachable. So these traits male implementing k-mean clustering in Python **reasonably straightforward**.

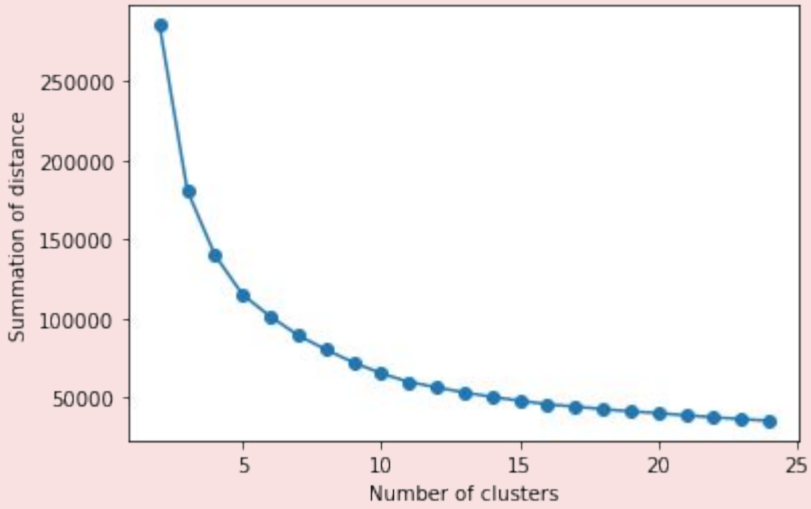
The **first step** is randomly select **k centroids**, where k is equal to the number of clusters of your choose. **Centroids** are data points representing the center of a cluster. Initialization of the centroids is an **important step**.

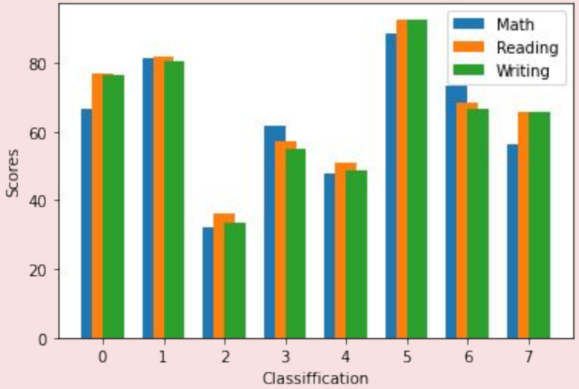
Expectation - Maximization

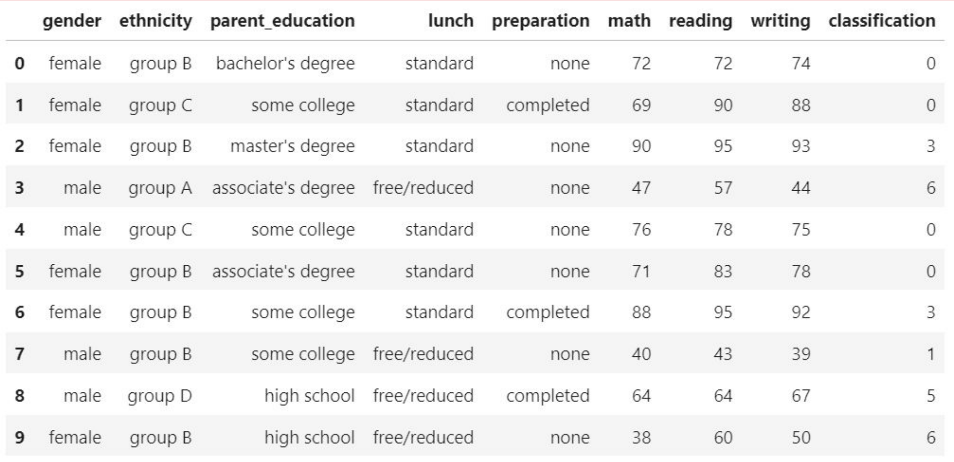
SSE an measure of clustering performance



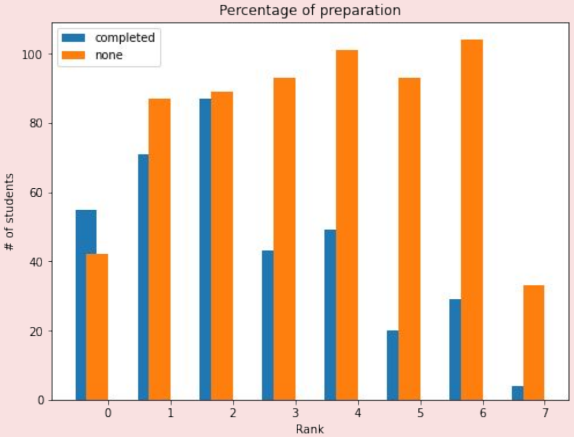
Analysis and Graphs

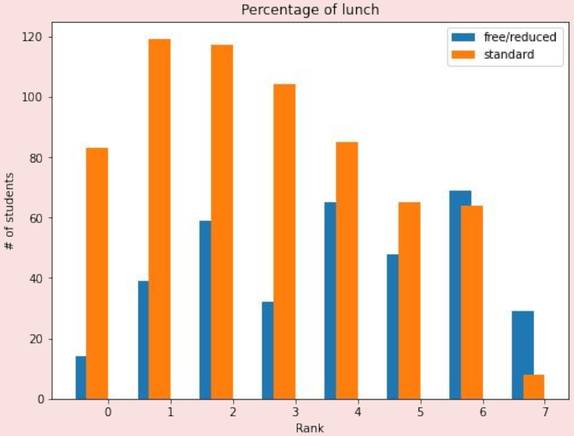


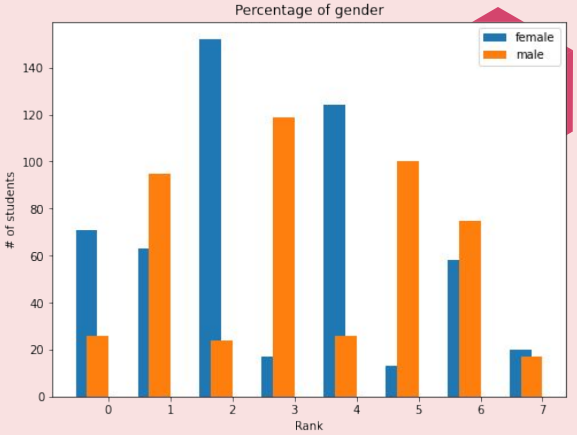




	gender	ethnicity	parent_education	lunch	preparation	math	reading	writing	classification
0	female	group B	bachelor's degree	standard	none	72	72	74	0
1	female	group C	some college	standard	completed	69	90	88	0
2	female	group B	master's degree	standard	none	90	95	93	3
3	male	group A	associate's degree	free/reduced	none	47	57	44	6
4	male	group C	some college	standard	none	76	78	75	0
5	female	group B	associate's degree	standard	none	71	83	78	0
6	female	group B	some college	standard	completed	88	95	92	3
7	male	group B	some college	free/reduced	none	40	43	39	1
8	male	group D	high school	free/reduced	completed	64	64	67	5
9	female	group B	high school	free/reduced	none	38	60	50	6


Percentage of preparation
completed
none
# of students
0
20
40
60
80
100
Rank
0
1
2
3
4
5
6
7


Percentage of lunch
free/reduced
standard
# of students
0
20
40
60
80
100
120
Rank
0
1
2
3
4
5
6
7


Percentage of gender
female
male
# of students
0
20
40
60
80
100
120
140
Rank
0
1
2
3
4
5
6
7

Conclusion

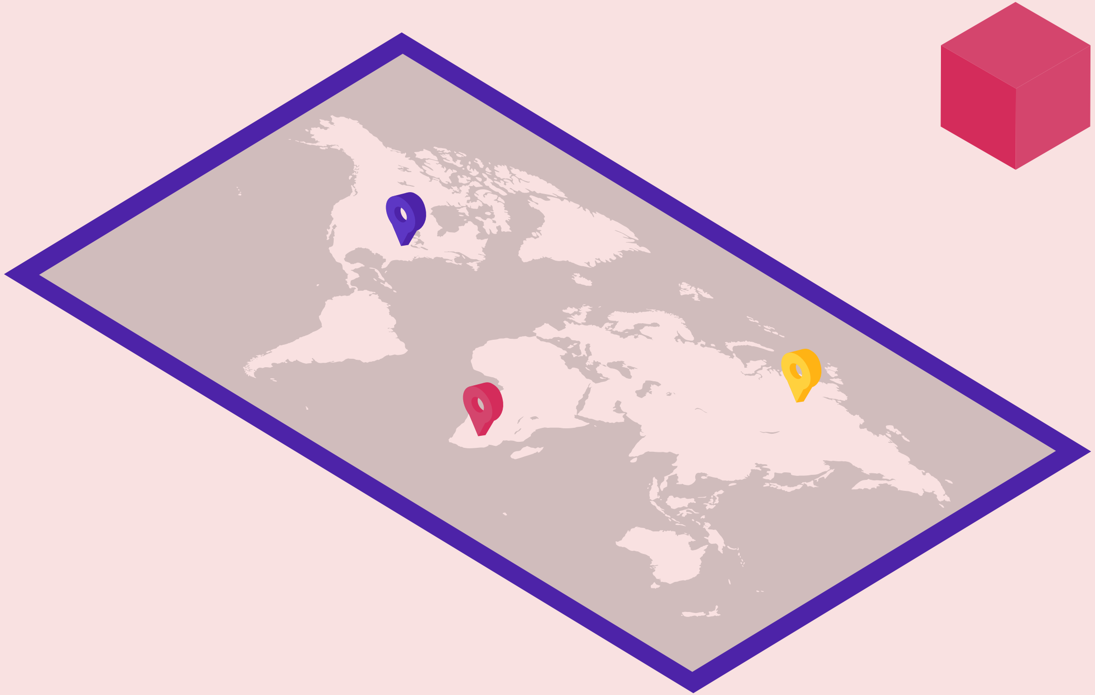
Education from parents can be helpful, but we can't count it as the most important

Completing the entire course is very important

One of the most significant things to the students is the lunch

it doesn't matter who you are, a girl or a boy. the effect on correlation is zero

in the end, we want to note that in order to get a good grade for a student, it is important to eat well and study carefully



Questions

Resources

- **Dataset:** http://roycekimmons.com/tools/generated_data/exams/
- **Pandas:** https://pandas.pydata.org/
- **Numpy:** https://numpy.org/
- **Sklearn:** https://scikit-learn.org/stable/
- **Seaborn:** https://seaborn.pydata.org/
- **Matplotlib:** https://matplotlib.org/

Thanks!

CREDITS: This presentation template was created by **Slidesgo**,
including icons by **Flaticon** and infographics & images by **Freepik**