The Data Science Process

The Data Science Process

The Data Science Process

- Ask a question

The Data Science Process

- Ask a question
- Get the data

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning
- Evaluate the trained model

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning
- Evaluate the trained model
- Communicate results/Deploy the model

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning
- Evaluate the trained model
- Communicate results/Deploy the model

Ask a question

Ask a question

Be clear and specific

Examples of poor questions:

Examples of poor questions:

“Which medication is better?”

Examples of poor questions:

“Which medication is better?”

“How can I find more donors?”

JESSICA
CLARE

Who are the best candidates?

Who are the best candidates?

Who are the best candidates given our salary constraint?

Who are the best candidates?

Who are the best candidates given our salary constraint?

Who are the best candidates given our salary constraint and desired tenure?

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning
- Evaluate the trained model
- Communicate results/Deploy the model

Some challenges with data

- Insufficient quantity of data
- Non-representative data
- Poor quality of data
- Irrelevant features

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning using the training dataset
- Evaluate the trained model
- Communicate results/Deploy the model

Data Types

Data Types

Numerical

Discrete: 365, IQ, SAT

Continuous: 3.1415, GPA

Data Types

Numerical

Discrete: 365, IQ, SAT

Continuous: 3.1415, GPA

Categorical

Gender, Race

Data Types

Numerical

Discrete: 365, IQ, SAT

Continuous: 3.1415, GPA

Categorical

Gender, Race

Ordinal

1st, 2nd, freshman, sophomore

SAT Performance

Student ID	Gender	Race	GPA	Lunch	Parent Educ.	SAT
1003987244	female	Hispanic	3.39	Standard	Some H.S.	1014
1008727612	Male	Asian	3.48	Free/ Reduced	H.S.	1060
1005298713	F	Black	4.00	Standard	Bach. Degree	1150
1002145908	male	White	3.91	Standard	Graduate Degree	1075
1006510902	Female	Latino	n/a	Free/ Reduced	Some College	1030
1003908124	Female	White	3.12	Standard		1095
1004187160	Female	Asian	2.80	Standard	Some College	1044
1004481263	Male	White	3.56	Standard	Assoc. Degree	1100
1002990127	M	White	4.00	Standard	Bach. Degree	1270
1005992918	F	Asian	2.94	Free/ Reduced	H.S.	993
1008312208	Male	African-American	2.87	Free/ Reduced	Assoc. Degree	9700

SAT Performance

Student ID	Gender	Race	GPA	Lunch	Parent Educ.	SAT
1003987244	female	Hispanic	3.39	Standard	Some H.S.	1014
1008727612	Male	Asian	3.48	Free/ Reduced	H.S.	1060
1005298713	F	Black	4.00	Standard	Bach. Degree	1150
1002145908	male	White	3.91	Standard	Graduate Degree	1075
1006510902	Female	Latino	n/a	Free/ Reduced	Some College	1030
1003908124	Female	White	3.12	Standard		1095
1004187160	Female	Asian	2.80	Standard	Some College	1044
1004481263	Male	White	3.56	Standard	Assoc. Degree	1100
1002990127	M	White	4.00	Standard	Bach. Degree	1270
1005992918	F	Asian	2.94	Free/ Reduced	H.S.	993
1008312208	Male	African-American	2.87	Free/ Reduced	Assoc. Degree	9700

← Target (y)

SAT Performance

Features (X) →

Student ID	Gender	Race	GPA	Lunch	Parent Educ.	SAT
1003987244	female	Hispanic	3.39	Standard	Some H.S.	1014
1008727612	Male	Asian	3.48	Free/ Reduced	H.S.	1060
1005298713	F	Black	4.00	Standard	Bach. Degree	1150
1002145908	male	White	3.91	Standard	Graduate Degree	1075
1006510902	Female	Latino	n/a	Free/ Reduced	Some College	1030
1003908124	Female	White	3.12	Standard		1095
1004187160	Female	Asian	2.80	Standard	Some College	1044
1004481263	Male	White	3.56	Standard	Assoc. Degree	1100
1002990127	M	White	4.00	Standard	Bach. Degree	1270
1005992918	F	Asian	2.94	Free/ Reduced	H.S.	993
1008312208	Male	African-American	2.87	Free/ Reduced	Assoc. Degree	9700

SAT Performance

Student ID	Gender	Race	GPA	Lunch	Parent Educ.	SAT
1003987244	female	Hispanic	3.39	Standard	Some H.S.	1014
1008727612	Male	Asian	3.48	Free/ Reduced	H.S.	1060
1005298713	F	Black	4.00	Standard	Bach. Degree	1150
1002145908	male	White	3.91	Standard	Graduate Degree	1075
1006510902	Female	Latino	n/a	Free/ Reduced	Some College	1030
1003908124	Female	White	3.12	Standard		1095
1004187160	Female	Asian	2.80	Standard	Some College	1044
1004481263	Male	White	3.56	Standard	Assoc. Degree	1100
1002990127	M	White	4.00	Standard	Bach. Degree	1270
1005992918	F	Asian	2.94	Free/ Reduced	H.S.	993
1008312208	Male	African-American	2.87	Free/ Reduced	Assoc. Degree	9700

Feature Selection

SAT Performance

Student ID	Gender	Race	GPA	Lunch	Parent Educ.	SAT
1003987244	female	Hispanic	3.39	Standard	Some H.S.	1014
1008727612	Male	Asian	3.48	Free/ Reduced	H.S.	1060
1005298713	F	Black	4.00	Standard	Bach. Degree	1150
1002145908	male	White	3.91	Standard	Graduate Degree	1075
1006510902	Female	Latino	n/a	Free/ Reduced	Some College	1030
1003908124	Female	White	3.12	Standard		1095
1004187160	Female	Asian	2.80	Standard	Some College	1044
1004481263	Male	White	3.56	Standard	Assoc. Degree	1100
1002990127	M	White	4.00	Standard	Bach. Degree	1270
1005992918	F	Asian	2.94	Free/ Reduced	H.S.	993
1008312208	Male	African-American	2.87	Free/ Reduced	Assoc. Degree	9700

SAT Performance

Student ID	Gender	Race	GPA	Lunch	Parent Educ.	SAT
1003987244	female	Hispanic	3.39	Standard	Some H.S.	1014
1008727612	Male	Asian	3.48	Free/ Reduced	H.S.	1060
1005298713	F	Black	4.00	Standard	Bach. Degree	1150
1002145908	male	White	3.91	Standard	Graduate Degree	1075
1006510902	Female	Latino	n/a	Free/ Reduced	Some College	1030
1003908124	Female	White	3.12	Standard		1095
1004187160	Female	Asian	2.80	Standard	Some College	1044
1004481263	Male	White	3.56	Standard	Assoc. Degree	1100
1002990127	M	White	4.00	Standard	Bach. Degree	1270
1005992918	F	Asian	2.94	Free/ Reduced	H.S.	993
1008312208	Male	African-American	2.87	Free/ Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
Male	Asian	3.48	Free/Reduced	H.S.	1060
F	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
Female	Latino	n/a	Free/Reduced	Some College	1030
Female	White	3.12	Standard		1095
Female	Asian	2.80	Standard	Some College	1044
Male	White	3.56	Standard	Assoc. Degree	1100
M	White	4.00	Standard	Bach. Degree	1270
F	Asian	2.94	Free/Reduced	H.S.	993
Male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

Clean the Data

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
Male	Asian	3.48	Free/Reduced	H.S.	1060
F	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
Female	Latino	n/a	Free/Reduced	Some College	1030
Female	White	3.12	Standard		1095
Female	Asian	2.80	Standard	Some College	1044
Male	White	3.56	Standard	Assoc. Degree	1100
M	White	4.00	Standard	Bach. Degree	1270
F	Asian	2.94	Free/Reduced	H.S.	993
Male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
Male	Asian	3.48	Free/Reduced	H.S.	1060
F	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
Female	Latino	n/a	Free/Reduced	Some College	1030
Female	White	3.12	Standard		1095
Female	Asian	2.80	Standard	Some College	1044
Male	White	3.56	Standard	Assoc. Degree	1100
M	White	4.00	Standard	Bach. Degree	1270
F	Asian	2.94	Free/Reduced	H.S.	993
Male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
Male	Asian	3.48	Free/Reduced	H.S.	1060
F	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
Female	Latino	n/a	Free/Reduced	Some College	1030
Female	White	3.12	Standard		1095
Female	Asian	2.80	Standard	Some College	1044
Male	White	3.56	Standard	Assoc. Degree	1100
M	White	4.00	Standard	Bach. Degree	1270
F	Asian	2.94	Free/Reduced	H.S.	993
Male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Latino	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Latino	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Latino	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	African-American	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	n/a	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard		1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard	Bach. Degree	1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard	Bach. Degree	1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard	Bach. Degree	1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard	Bach. Degree	1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	9700

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard	Bach. Degree	1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	970

Transform the Data

Transform the Data

Machines want numbers

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard	Bach. Degree	1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	970

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
female	Hispanic	3.39	Standard	Some H.S.	1014
male	Asian	3.48	Free/Reduced	H.S.	1060
female	Black	4.00	Standard	Bach. Degree	1150
male	White	3.91	Standard	Graduate Degree	1075
female	Hispanic	3.25	Free/Reduced	Some College	1030
female	White	3.12	Standard	Bach. Degree	1095
female	Asian	2.80	Standard	Some College	1044
male	White	3.56	Standard	Assoc. Degree	1100
male	White	4.00	Standard	Bach. Degree	1270
female	Asian	2.94	Free/Reduced	H.S.	993
male	Black	2.87	Free/Reduced	Assoc. Degree	970

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
1	Hispanic	3.39	Standard	Some H.S.	1014
0	Asian	3.48	Free/Reduced	H.S.	1060
1	Black	4.00	Standard	Bach. Degree	1150
0	White	3.91	Standard	Graduate Degree	1075
1	Hispanic	3.25	Free/Reduced	Some College	1030
1	White	3.12	Standard	Bach. Degree	1095
1	Asian	2.80	Standard	Some College	1044
0	White	3.56	Standard	Assoc. Degree	1100
0	White	4.00	Standard	Bach. Degree	1270
1	Asian	2.94	Free/Reduced	H.S.	993
0	Black	2.87	Free/Reduced	Assoc. Degree	970

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
1	Hispanic	3.39	Standard	Some H.S.	1014
0	Asian	3.48	Free/Reduced	H.S.	1060
1	Black	4.00	Standard	Bach. Degree	1150
0	White	3.91	Standard	Graduate Degree	1075
1	Hispanic	3.25	Free/Reduced	Some College	1030
1	White	3.12	Standard	Bach. Degree	1095
1	Asian	2.80	Standard	Some College	1044
0	White	3.56	Standard	Assoc. Degree	1100
0	White	4.00	Standard	Bach. Degree	1270
1	Asian	2.94	Free/Reduced	H.S.	993
0	Black	2.87	Free/Reduced	Assoc. Degree	970

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
1	1	3.39	Standard	Some H.S.	1014
0	2	3.48	Free/Reduced	H.S.	1060
1	3	4.00	Standard	Bach. Degree	1150
0	4	3.91	Standard	Graduate Degree	1075
1	1	3.25	Free/Reduced	Some College	1030
1	4	3.12	Standard	Bach. Degree	1095
1	2	2.80	Standard	Some College	1044
0	4	3.56	Standard	Assoc. Degree	1100
0	4	4.00	Standard	Bach. Degree	1270
1	2	2.94	Free/Reduced	H.S.	993
0	3	2.87	Free/Reduced	Assoc. Degree	970

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
1	Hispanic	3.39	Standard	Some H.S.	1014
0	Asian	3.48	Free/Reduced	H.S.	1060
1	Black	4.00	Standard	Bach. Degree	1150
0	White	3.91	Standard	Graduate Degree	1075
1	Hispanic	3.25	Free/Reduced	Some College	1030
1	White	3.12	Standard	Bach. Degree	1095
1	Asian	2.80	Standard	Some College	1044
0	White	3.56	Standard	Assoc. Degree	1100
0	White	4.00	Standard	Bach. Degree	1270
1	Asian	2.94	Free/Reduced	H.S.	993
0	Black	2.87	Free/Reduced	Assoc. Degree	970

One-Hot Encoding

SAT Performance

Gender	Race	GPA	Lunch	Parent Educ.	SAT
1	Hispanic	3.39	Standard	Some H.S.	1014
0	Asian	3.48	Free/Reduced	H.S.	1060
1	Black	4.00	Standard	Bach. Degree	1150
0	White	3.91	Standard	Graduate Degree	1075
1	Hispanic	3.25	Free/Reduced	Some College	1030
1	White	3.12	Standard	Bach. Degree	1095
1	Asian	2.80	Standard	Some College	1044
0	White	3.56	Standard	Assoc. Degree	1100
0	White	4.00	Standard	Bach. Degree	1270
1	Asian	2.94	Free/Reduced	H.S.	993
0	Black	2.87	Free/Reduced	Assoc. Degree	970

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	Standard	Some H.S.	1014
0	0	1	0	0	3.48	Free/ Reduced	H.S.	1060
1	0	0	1	0	4.00	Standard	Bach. Degree	1150
0	0	0	0	1	3.91	Standard	Graduate Degree	1075
1	1	0	0	0	3.25	Free/ Reduced	Some College	1030
1	0	0	0	1	3.12	Standard	Bach. Degree	1095
1	0	1	0	0	2.80	Standard	Some College	1044
0	0	0	0	1	3.56	Standard	Assoc. Degree	1100
0	0	0	0	1	4.00	Standard	Bach. Degree	1270
1	0	1	0	0	2.94	Free/ Reduced	H.S.	993
0	0	0	1	0	2.87	Free/ Reduced	Assoc. Degree	970

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	Standard	Some H.S.	1014
0	0	1	0	0	3.48	Free/ Reduced	H.S.	1060
1	0	0	1	0	4.00	Standard	Bach. Degree	1150
0	0	0	0	1	3.91	Standard	Graduate Degree	1075
1	1	0	0	0	3.25	Free/ Reduced	Some College	1030
1	0	0	0	1	3.12	Standard	Bach. Degree	1095
1	0	1	0	0	2.80	Standard	Some College	1044
0	0	0	0	1	3.56	Standard	Assoc. Degree	1100
0	0	0	0	1	4.00	Standard	Bach. Degree	1270
1	0	1	0	0	2.94	Free/ Reduced	H.S.	993
0	0	0	1	0	2.87	Free/ Reduced	Assoc. Degree	970

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	Standard	Some H.S.	1014
0	0	1	0	0	3.48	Free/ Reduced	H.S.	1060
1	0	0	1	0	4.00	Standard	Bach. Degree	1150
0	0	0	0	1	3.91	Standard	Graduate Degree	1075
1	1	0	0	0	3.25	Free/ Reduced	Some College	1030
1	0	0	0	1	3.12	Standard	Bach. Degree	1095
1	0	1	0	0	2.80	Standard	Some College	1044
0	0	0	0	1	3.56	Standard	Assoc. Degree	1100
0	0	0	0	1	4.00	Standard	Bach. Degree	1270
1	0	1	0	0	2.94	Free/ Reduced	H.S.	993
0	0	0	1	0	2.87	Free/ Reduced	Assoc. Degree	970

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	1	Some H.S.	1014
0	0	1	0	0	3.48	0	H.S.	1060
1	0	0	1	0	4.00	1	Bach. Degree	1150
0	0	0	0	1	3.91	1	Graduate Degree	1075
1	1	0	0	0	3.25	0	Some College	1030
1	0	0	0	1	3.12	1	Bach. Degree	1095
1	0	1	0	0	2.80	1	Some College	1044
0	0	0	0	1	3.56	1	Assoc. Degree	1100
0	0	0	0	1	4.00	1	Bach. Degree	1270
1	0	1	0	0	2.94	0	H.S.	993
0	0	0	1	0	2.87	0	Assoc. Degree	970

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	1	Some H.S.	1014
0	0	1	0	0	3.48	0	H.S.	1060
1	0	0	1	0	4.00	1	Bach. Degree	1150
0	0	0	0	1	3.91	1	Graduate Degree	1075
1	1	0	0	0	3.25	0	Some College	1030
1	0	0	0	1	3.12	1	Bach. Degree	1095
1	0	1	0	0	2.80	1	Some College	1044
0	0	0	0	1	3.56	1	Assoc. Degree	1100
0	0	0	0	1	4.00	1	Bach. Degree	1270
1	0	1	0	0	2.94	0	H.S.	993
0	0	0	1	0	2.87	0	Assoc. Degree	970

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	1	0	1014
0	0	1	0	0	3.48	0	1	1060
1	0	0	1	0	4.00	1	4	1150
0	0	0	0	1	3.91	1	5	1075
1	1	0	0	0	3.25	0	2	1030
1	0	0	0	1	3.12	1	4	1095
1	0	1	0	0	2.80	1	2	1044
0	0	0	0	1	3.56	1	3	1100
0	0	0	0	1	4.00	1	4	1270
1	0	1	0	0	2.94	0	1	993
0	0	0	1	0	2.87	0	3	970

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	1	0	1014
0	0	1	0	0	3.48	0	1	1060
1	0	0	1	0	4.00	1	4	1150
0	0	0	0	1	3.91	1	5	1075
1	1	0	0	0	3.25	0	2	1030
1	0	0	0	1	3.12	1	4	1095
1	0	1	0	0	2.80	1	2	1044
0	0	0	0	1	3.56	1	3	1100
0	0	0	0	1	4.00	1	4	1270
1	0	1	0	0	2.94	0	1	993
0	0	0	1	0	2.87	0	3	970

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	1	0	Low
0	0	1	0	0	3.48	0	1	Average
1	0	0	1	0	4.00	1	4	High
0	0	0	0	1	3.91	1	5	Average
1	1	0	0	0	3.25	0	2	Average
1	0	0	0	1	3.12	1	4	Average
1	0	1	0	0	2.80	1	2	Average
0	0	0	0	1	3.56	1	3	High
0	0	0	0	1	4.00	1	4	High
1	0	1	0	0	2.94	0	1	Low
0	0	0	1	0	2.87	0	3	Low

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	1	0	0
0	0	1	0	0	3.48	0	1	1
1	0	0	1	0	4.00	1	4	2
0	0	0	0	1	3.91	1	5	1
1	1	0	0	0	3.25	0	2	1
1	0	0	0	1	3.12	1	4	1
1	0	1	0	0	2.80	1	2	1
0	0	0	0	1	3.56	1	3	2
0	0	0	0	1	4.00	1	4	2
1	0	1	0	0	2.94	0	1	0
0	0	0	1	0	2.87	0	3	0

SAT Performance

Gender	Race_H	Race_A	Race_B	Race_W	GPA	Lunch	Parent Educ.	SAT
1	1	0	0	0	3.39	1	0	0
0	0	1	0	0	3.48	0	1	1
1	0	0	1	0	4.00	1	4	2
0	0	0	0	1	3.91	1	5	1
1	1	0	0	0	3.25	0	2	1
1	0	0	0	1	3.12	1	4	1
1	0	1	0	0	2.80	1	2	1
0	0	0	0	1	3.56	1	3	2
0	0	0	0	1	4.00	1	4	2
1	0	1	0	0	2.94	0	1	0
0	0	0	1	0	2.87	0	3	0

SAT Performance

1	1	0	0	0	3.39	1	0	0
0	0	1	0	0	3.48	0	1	1
1	0	0	1	0	4.00	1	4	2
0	0	0	0	1	3.91	1	5	1
1	1	0	0	0	3.25	0	2	1
1	0	0	0	1	3.12	1	4	1
1	0	1	0	0	2.80	1	2	1
0	0	0	0	1	3.56	1	3	2
0	0	0	0	1	4.00	1	4	2
1	0	1	0	0	2.94	0	1	0
0	0	0	1	0	2.87	0	3	0

SAT Performance

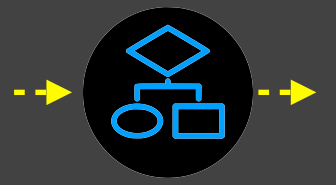
$$x = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 3.39 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 3.48 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 4.00 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 & 3.91 & 1 & 5 \\ 1 & 1 & 0 & 0 & 0 & 3.25 & 0 & 2 \\ 1 & 0 & 0 & 0 & 1 & 3.12 & 1 & 4 \\ 1 & 0 & 1 & 0 & 0 & 2.80 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 & 3.56 & 1 & 3 \\ 0 & 0 & 0 & 0 & 1 & 4.00 & 1 & 4 \\ 1 & 0 & 1 & 0 & 0 & 2.94 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 2.87 & 0 & 3 \end{bmatrix} \quad y = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$

SAT Performance

$X =$

1	1	0	0	0	3.39	1	0
0	0	1	0	0	3.48	0	1
1	0	0	1	0	4.00	1	4
0	0	0	0	1	3.91	1	5
1	1	0	0	0	3.25	0	2
1	0	0	0	1	3.12	1	4
1	0	1	0	0	2.80	1	2
0	0	0	0	1	3.56	1	3
0	0	0	0	1	4.00	1	4
1	0	1	0	0	2.94	0	1
0	0	0	1	0	2.87	0	3

machine learning



$y =$

0
1
2
1
1
1
1
2
2
0
0

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning
- Evaluate the trained model
- Communicate results/Deploy the model

Dataset

Professor Evaluation Scores

Professor Evaluation Scores

Age	Gender	Minority	Native English	Lower Division	Attractive	Evaluation Score	High Evaluation
41	male	no	yes	no	yes	4.1	yes
37	female	no	yes	yes	yes	4.2	yes
34	female	no	yes	no	yes	4.8	yes
44	female	no	no	no	no	4.4	yes
57	male	yes	yes	no	yes	4.7	yes
39	female	yes	yes	yes	yes	4.2	yes
33	female	no	yes	no	yes	4.5	yes
38	male	no	yes	yes	no	3.4	no
50	male	no	yes	yes	no	3.7	yes
43	female	yes	yes	no	no	3.6	yes

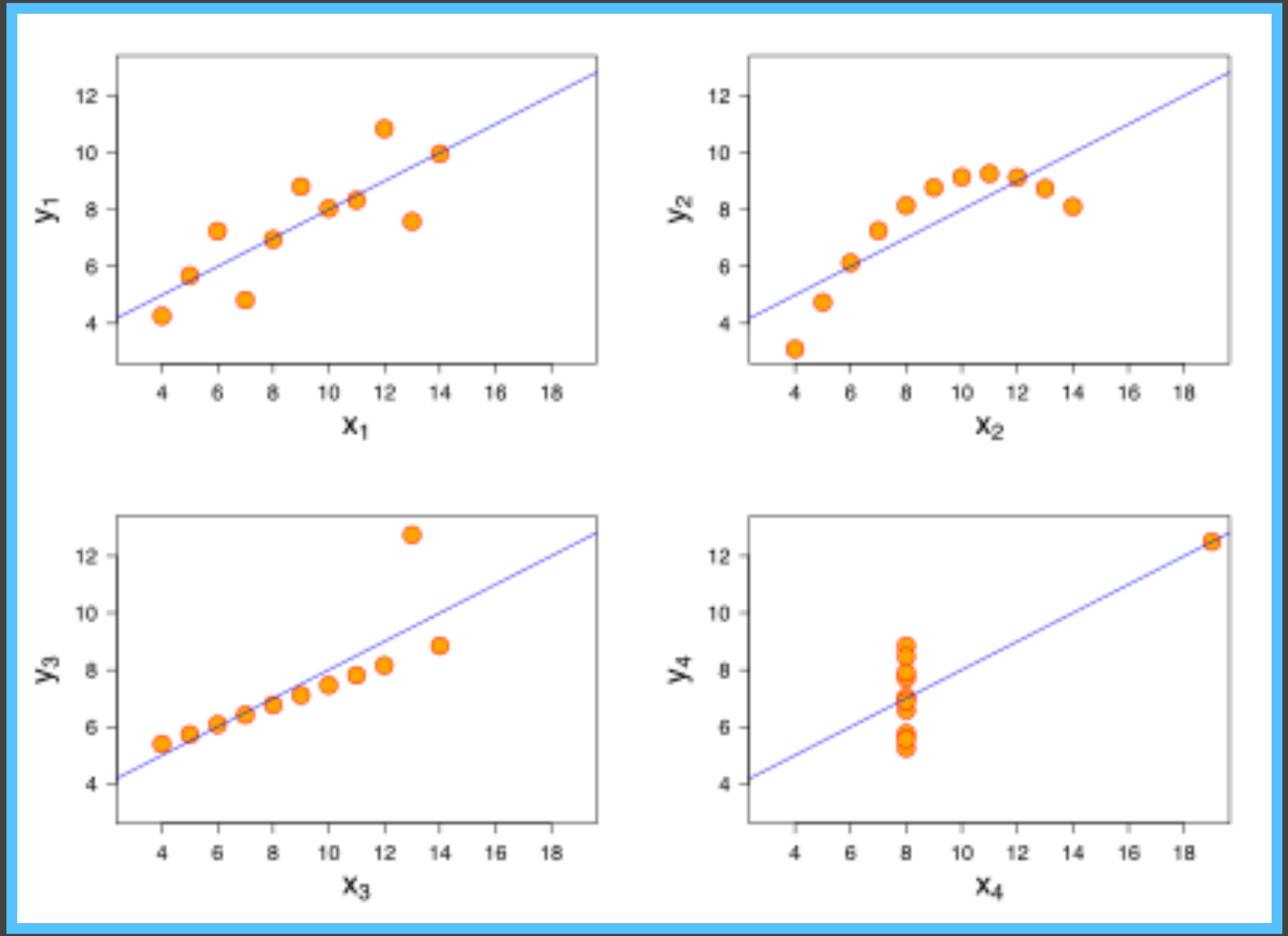
Summarize the data

Data Summary

	Age	Gender	Minority	Native English	Lower Division	Attractive	Evaluation Score
count	463	463	463	463	463	463	463
mean	48.365	0.421	0.138	0.939	0.339	0.403	3.998
std	9.802	0.494	0.345	0.238	0.473	0.491	0.554
min	29	0	0	0	0	0	2.1
25%	42	0	0	1	0	0	3.6
50%	48	1	1	1	1	1	4.0
75%	57	1	0	1	1	1	4.4
max	73	1	1	1	1	1	5.0

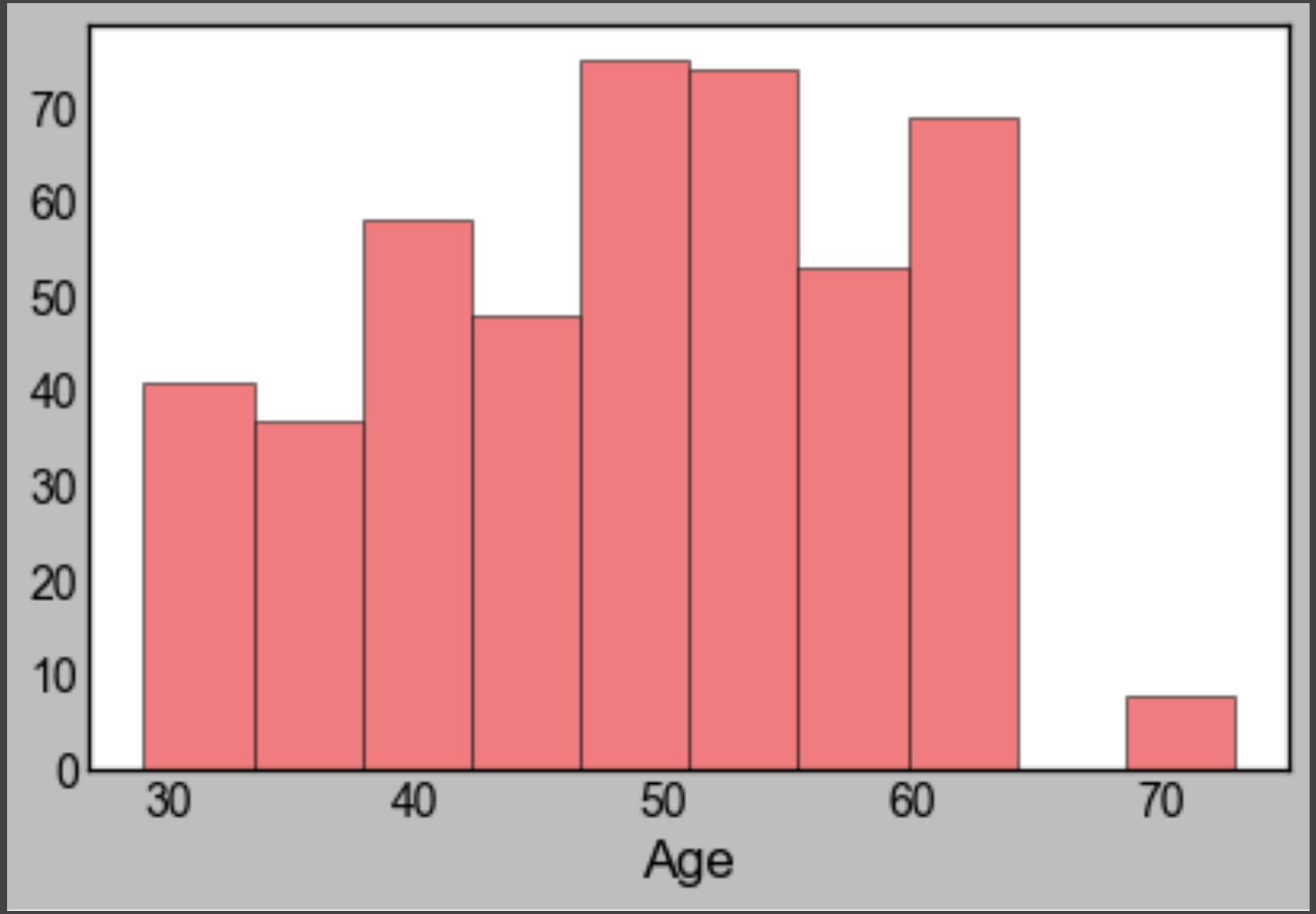
Visually explore the data

Anscombe's quartet

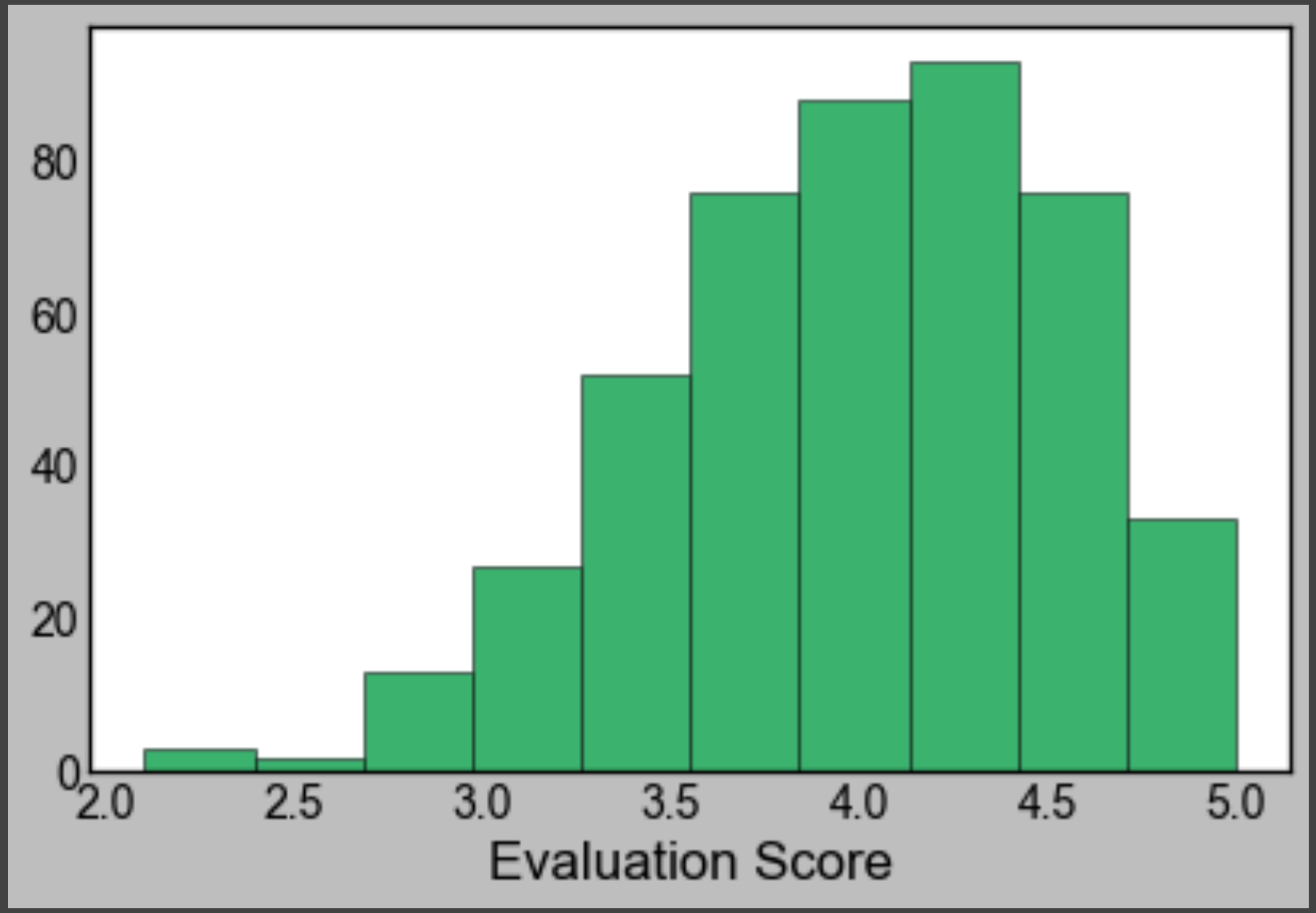


Single Features
(univariate distributions)

Age Distribution

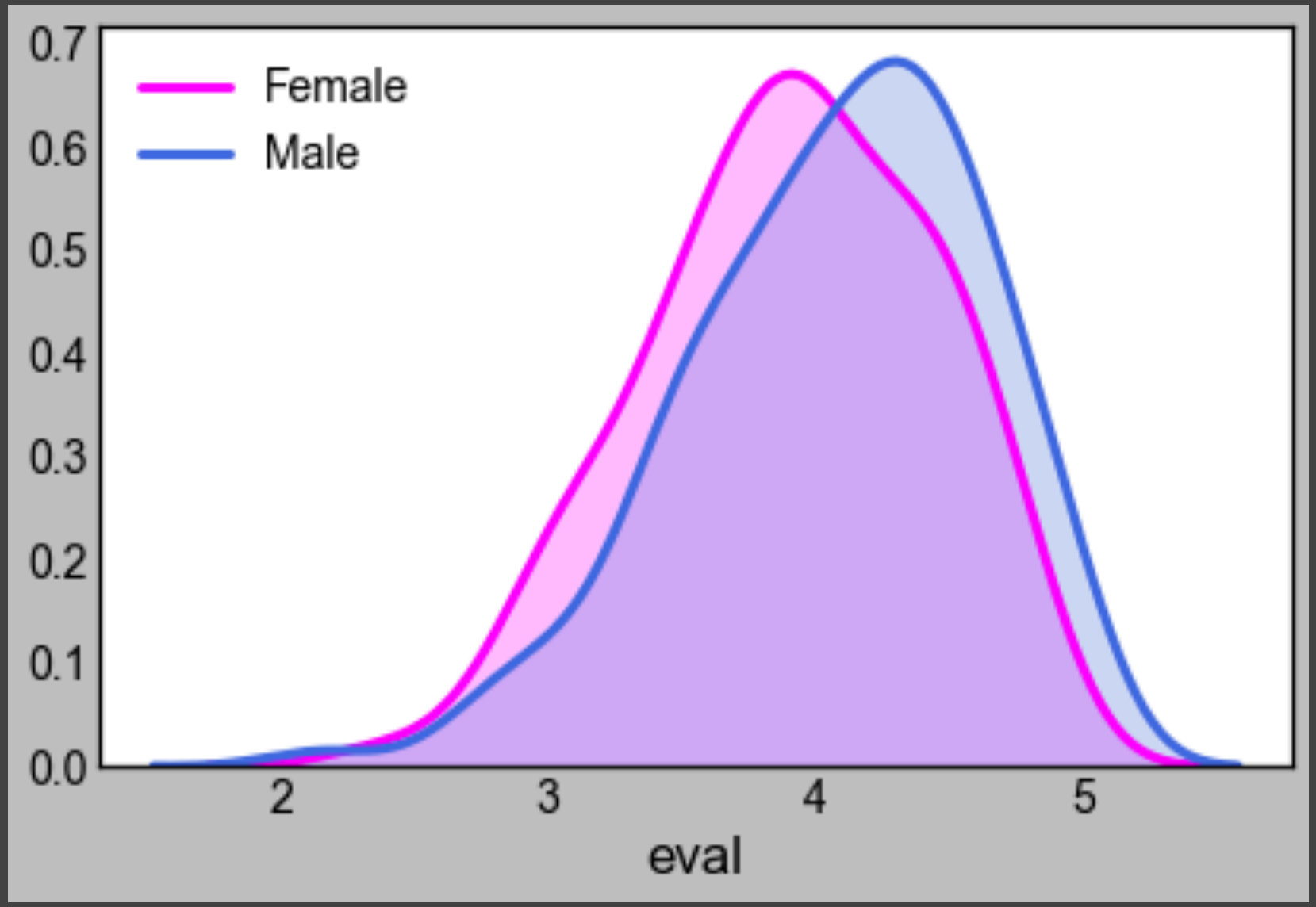


Evaluation Score Distribution



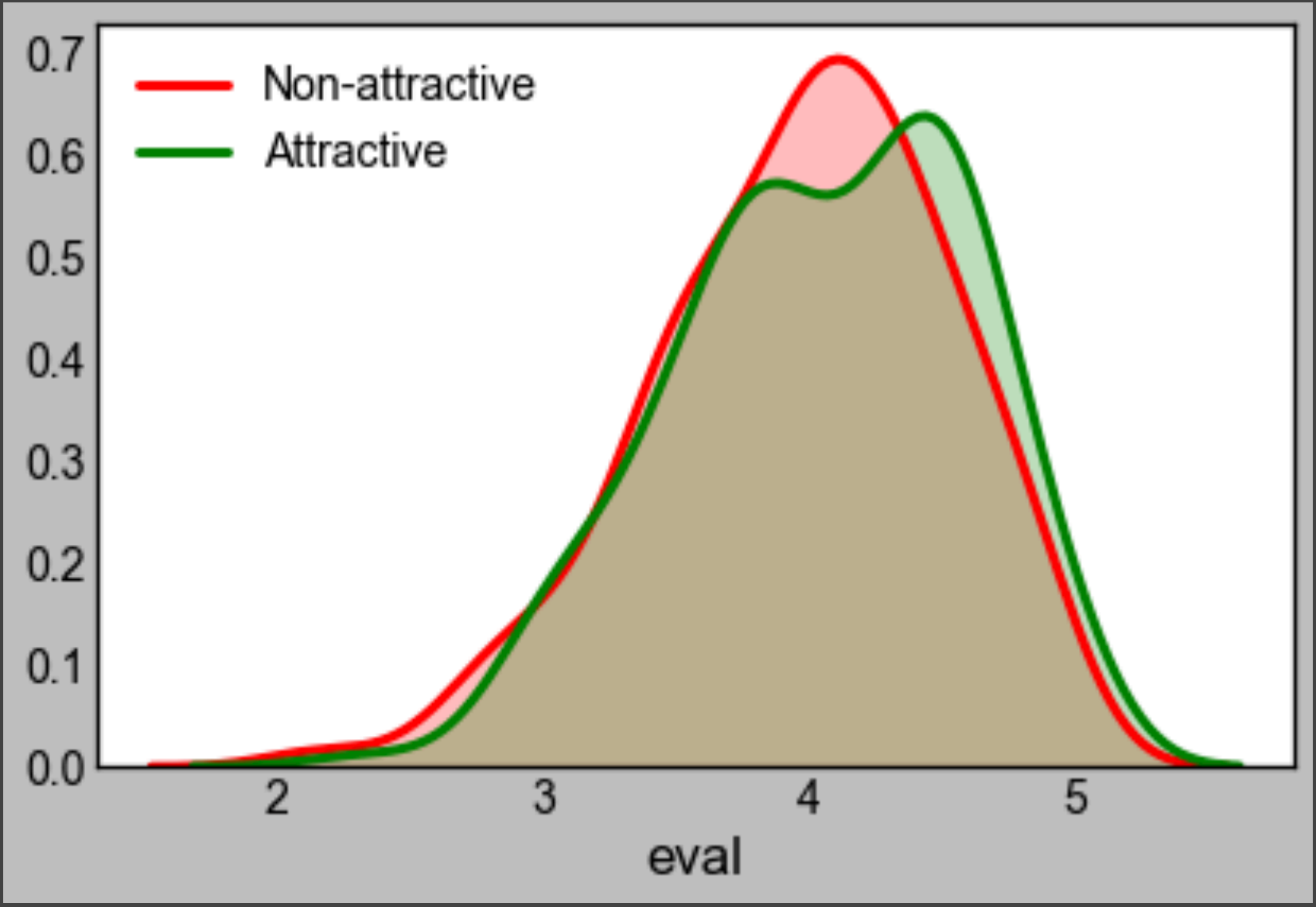
Two Features
(Bivariate distributions)

Female Male



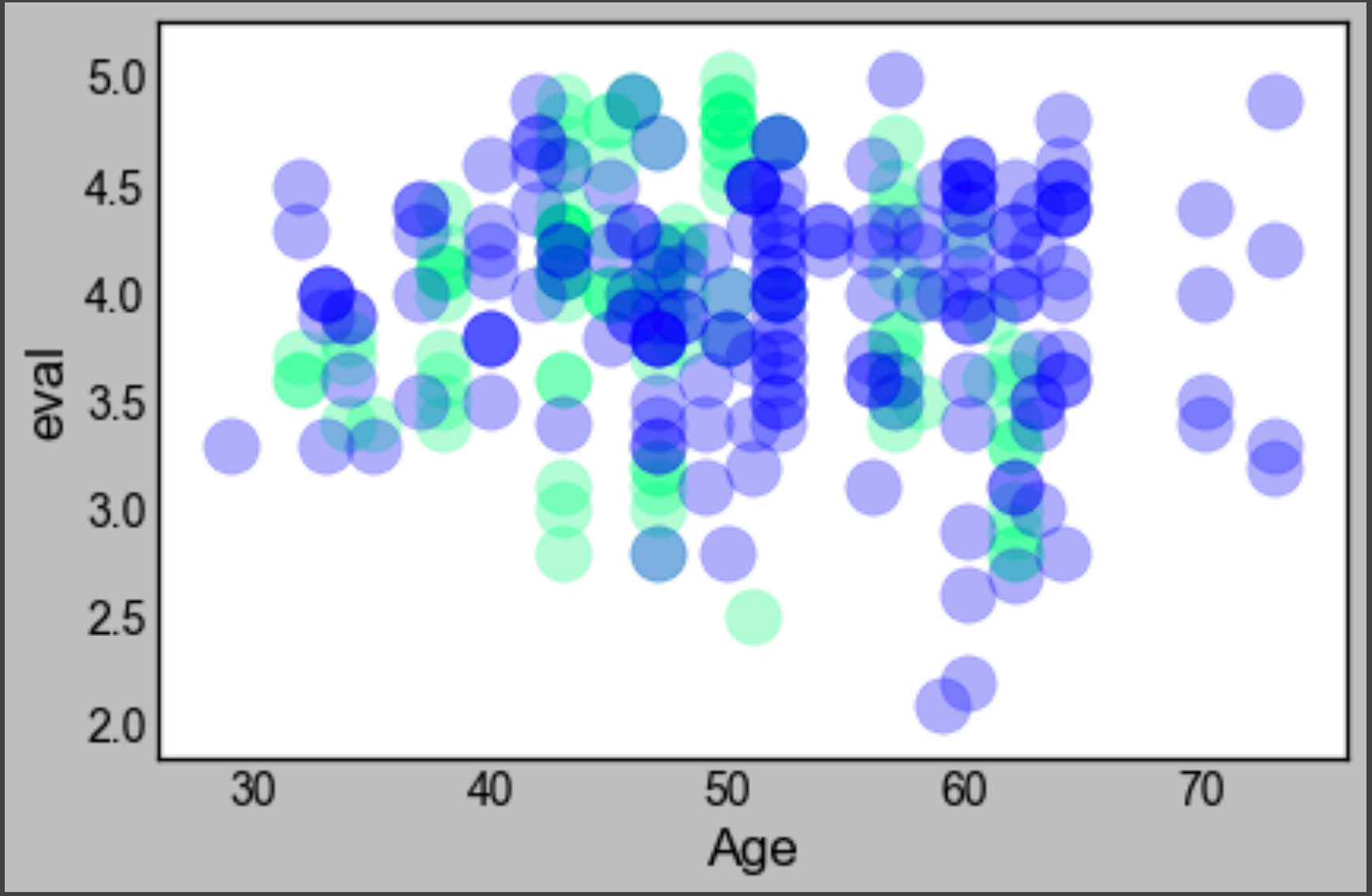
non-attractive

attractive



Non-Attractive Professors

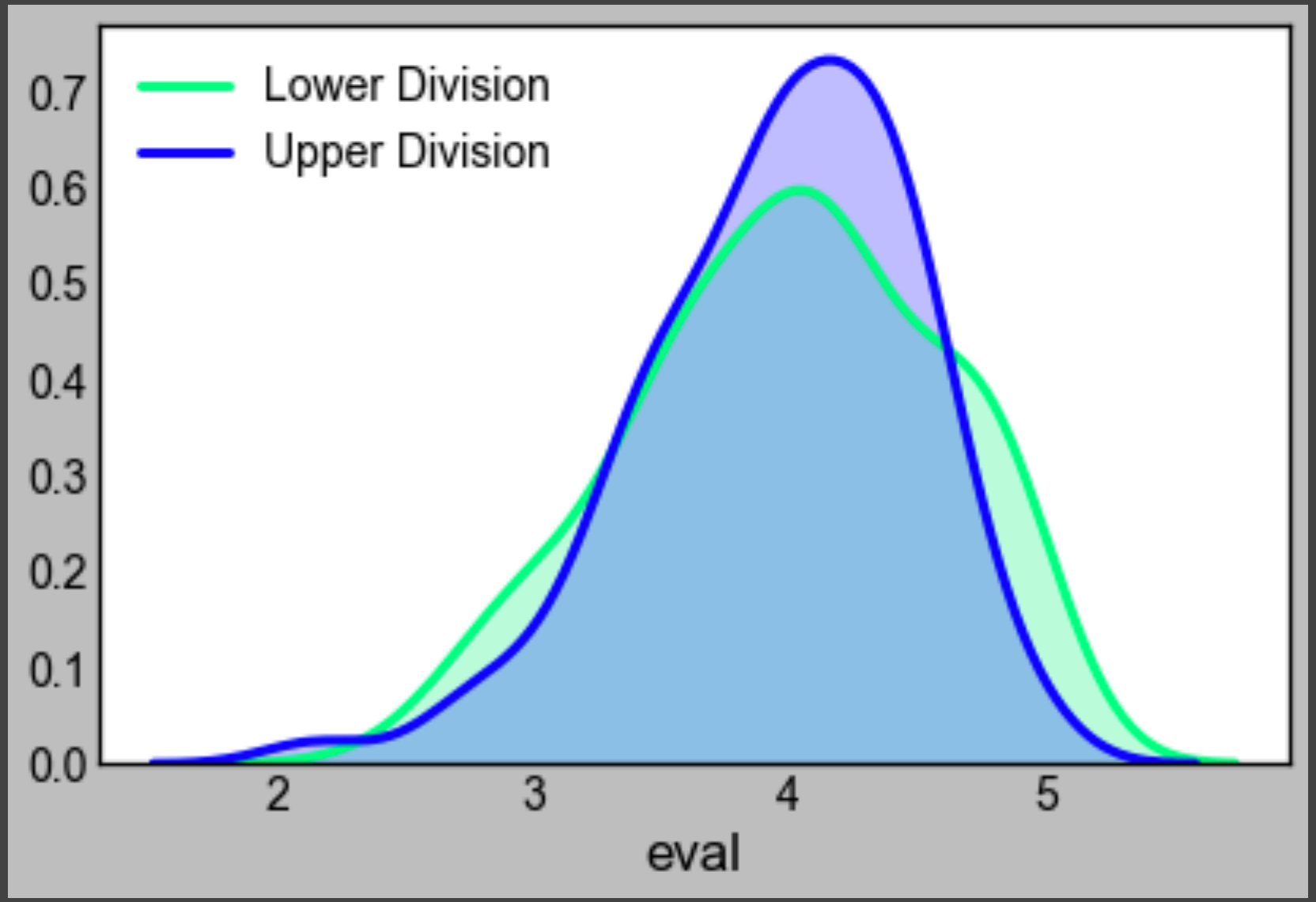
lower division upper division



Non-Attractive Professors

lower division

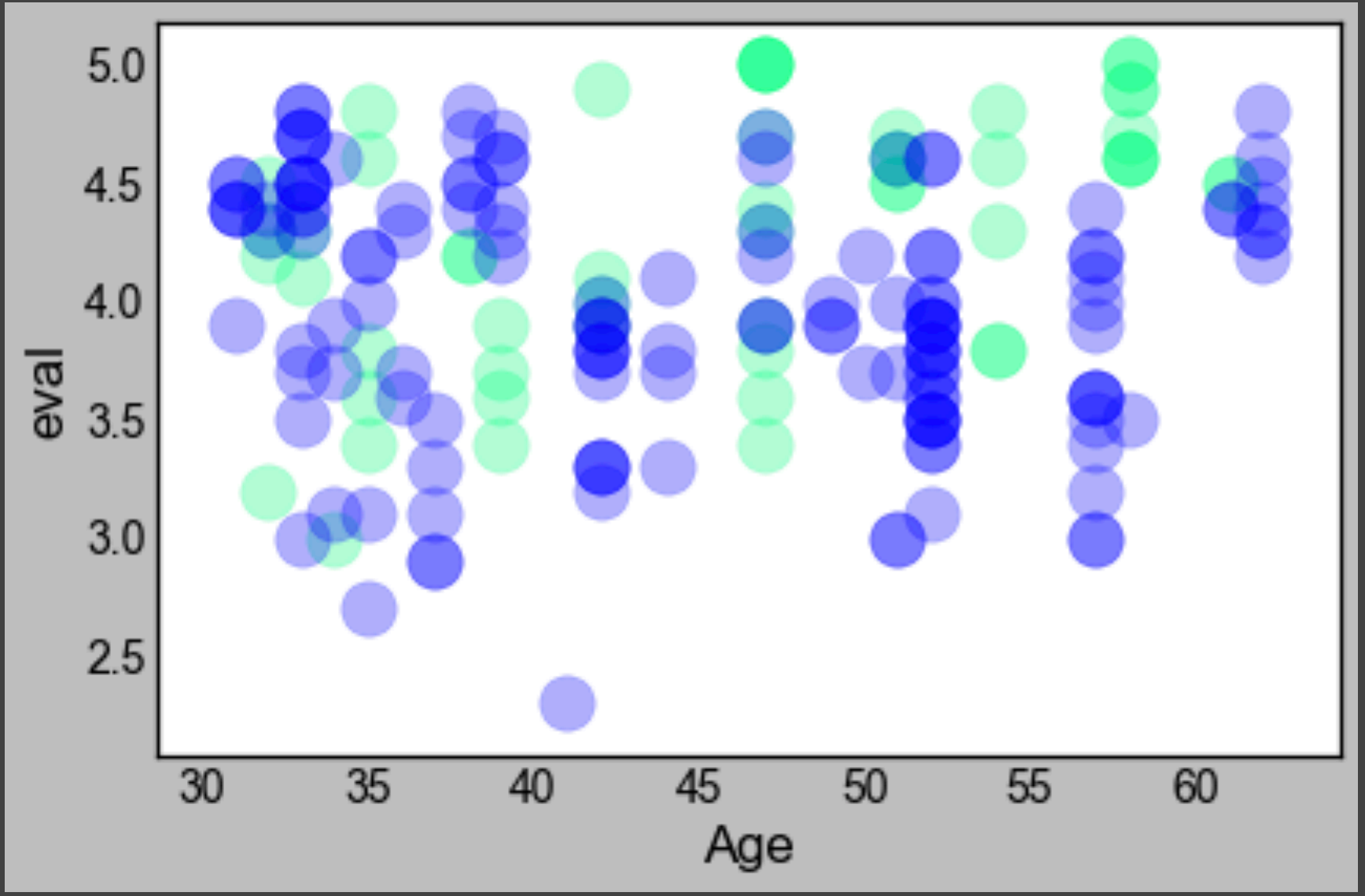
upper division



Attractive Professors

lower division

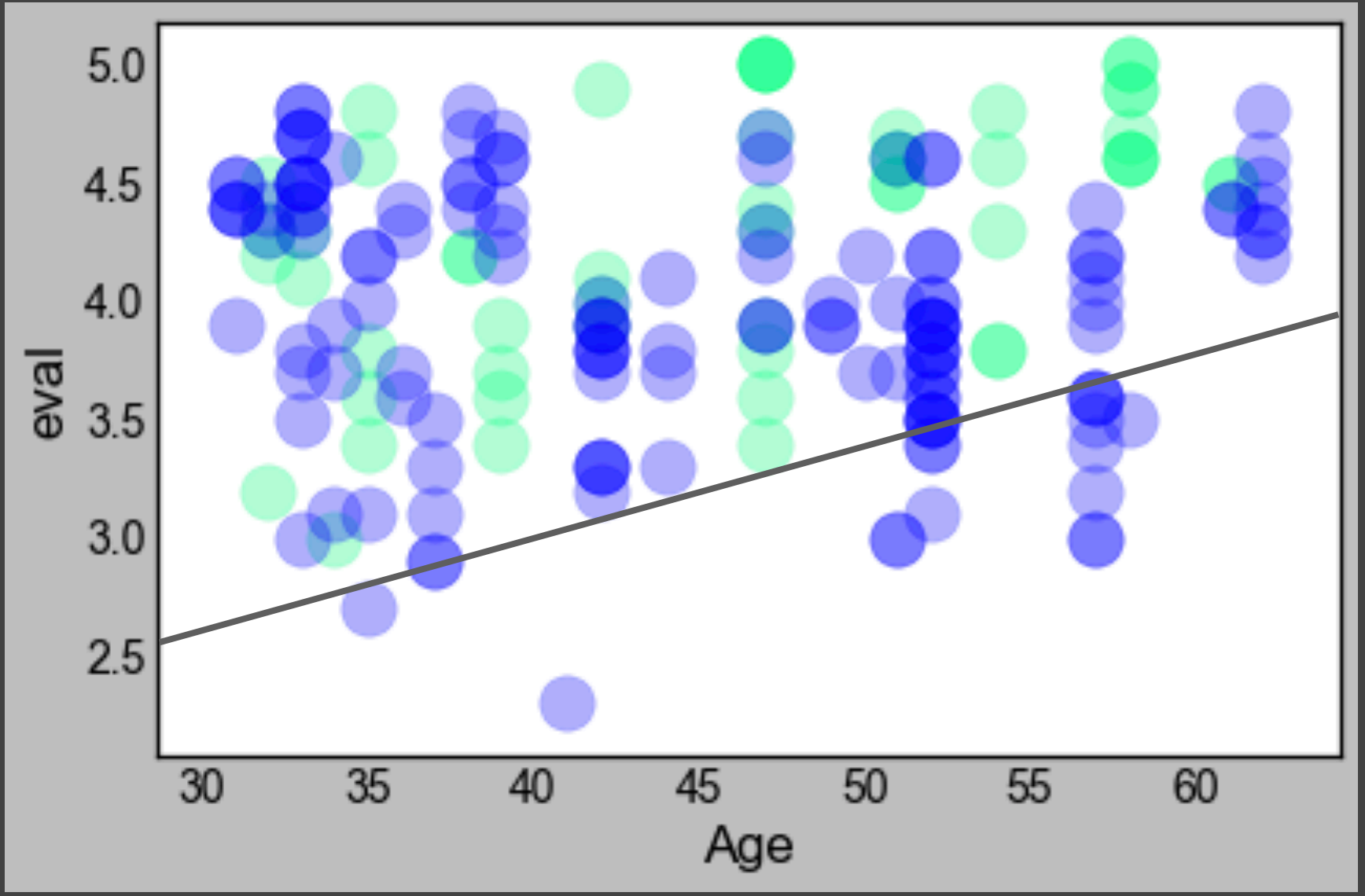
upper division



Attractive Professors

lower division

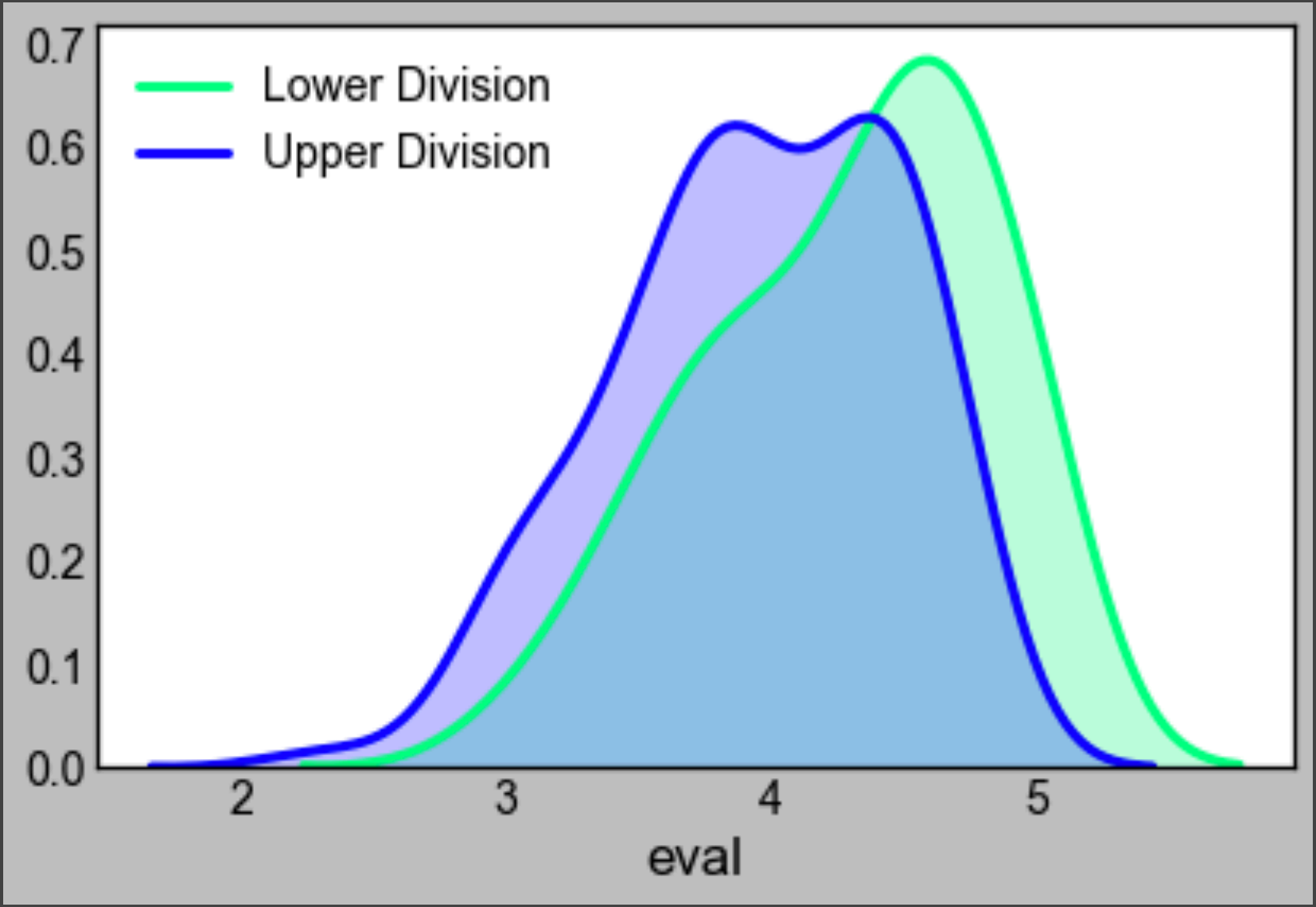
upper division



Attractive Professors

lower division

upper division



Professor Evaluation Scores

Professor Evaluation Scores

Upper Division Students

High Evaluation

Professor Evaluation Scores

Upper Division Students

High Evaluation

Attractive Professor: .82

Professor Evaluation Scores

Upper Division Students

High Evaluation

Attractive Professor: .82

Non-attractive Professor: .83

Professor Evaluation Scores

Lower Division Students

High Evaluation

Professor Evaluation Scores

Lower Division Students

High Evaluation

Attractive Professor: .91

Professor Evaluation Scores

Lower Division Students

High Evaluation

Attractive Professor: .91

Non-attractive Professor: .81

The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning using the training dataset
- Evaluate the trained model
- Communicate results/Deploy the model

Professor Evaluation Scores

Age	Gender	Minority	Native English	Lower Division	Attractive	Evaluation Score	High Evaluation
41	0	0	1	0	1	4.1	1
37	1	0	1	1	1	4.2	1
34	1	0	1	0	1	4.8	1
44	1	0	0	0	0	4.4	1
57	0	1	1	0	1	4.7	1
39	1	1	1	1	1	4.2	1
33	1	0	1	0	1	4.5	1
38	0	0	1	1	0	3.4	0
50	0	0	1	1	0	3.7	1
43	1	1	1	0	0	3.6	1

Professor Evaluation Scores

Age	Gender	Minority	Native English	Lower Division	Attractive	Evaluation Score	High Evaluation
41	0	0	1	0	1	4.1	1
37	1	0	1	1	1	4.2	1
34	1	0	1	0	1	4.8	1
44	1	0	0	0	0	4.4	1
57	0	1	1	0	1	4.7	1
39	1	1	1	1	1	4.2	1
33	1	0	1	0	1	4.5	1
38	0	0	1	1	0	3.4	0
50	0	0	1	1	0	3.7	1
43	1	1	1	0	0	3.6	1

Professor Evaluation Scores

Age	Gender	Minority	Native English	Lower Division	Attractive	Evaluation Score	High Evaluation
41	0	0	1	0	1	4.1	1
37	1	0	1	1	1	4.2	1
34	1	0	1	0	1	4.8	1
44	1	0	0	0	0	4.4	1
57	0	1	1	0	1	4.7	1
39	1	1	1	1	1	4.2	1
33	1	0	1	0	1	4.5	1
38	0	0	1	1	0	3.4	0
50	0	0	1	1	0	3.7	1
43	1	1	1	0	0	3.6	1

Lower Division	Attractive
0	1
1	1
0	1
0	0
0	1
1	1
0	1
1	0
1	0
0	0

Lower Division	Attractive	Beauty Bias
0	1	
1	1	
0	1	
0	0	
0	1	
1	1	
0	1	
1	0	
1	0	
0	0	

Lower Division x Attractive = Beauty Bias

Lower Division x Attractive = Beauty Bias

Lower Division	Attractive	Beauty Bias
0	1	
1	1	
0	1	
0	0	
0	1	
1	1	
0	1	
1	0	
1	0	
0	0	

Lower Division x Attractive = Beauty Bias

Lower Division	Attractive	Beauty Bias
0	1	
1	1	
0	1	
0	0	
0	1	
1	1	
0	1	
1	0	
1	0	
0	0	

Lower Division x Attractive = Beauty Bias

Lower Division	Attractive	Beauty Bias
0	1	
1	1	1
0	1	
0	0	
0	1	
1	1	1
0	1	
1	0	
1	0	
0	0	

Lower Division x Attractive = Beauty Bias

Lower Division	Attractive	Beauty Bias
0	1	0
1	1	1
0	1	0
0	0	0
0	1	0
1	1	1
0	1	0
1	0	0
1	0	0
0	0	0

Professor Evaluation Scores

Age	Gender	Minority	Native English	Lower Division	Attractive	Beauty Bias	Evaluation Score	High Evaluation
41	0	0	1	0	1	0	4.1	1
37	1	0	1	1	1	1	4.2	1
34	1	0	1	0	1	0	4.8	1
44	1	0	0	0	0	0	4.4	1
57	0	1	1	0	1	0	4.7	1
39	1	1	1	1	1	1	4.2	1
33	1	0	1	0	1	0	4.5	1
38	0	0	1	1	0	0	3.4	0
50	0	0	1	1	0	0	3.7	1
43	1	1	1	0	0	0	3.6	1

Professor Evaluation Scores

Age	Gender	Minority	Native English	Lower Division	Attractive	Beauty Bias	Evaluation Score	High Evaluation
41	0	0	1	0	1	0	4.1	1
37	1	0	1	1	1	1	4.2	1
34	1	0	1	0	1	0	4.8	1
44	1	0	0	0	0	0	4.4	1
57	0	1	1	0	1	0	4.7	1
39	1	1	1	1	1	1	4.2	1
33	1	0	1	0	1	0	4.5	1
38	0	0	1	1	0	0	3.4	0
50	0	0	1	1	0	0	3.7	1
43	1	1	1	0	0	0	3.6	1

Professor Evaluation Scores

41	0	0	1	0	1	0	4.1	1
37	1	0	1	1	1	1	4.2	1
34	1	0	1	0	1	0	4.8	1
44	1	0	0	0	0	0	4.4	1
57	0	1	1	0	1	0	4.7	1
39	1	1	1	1	1	1	4.2	1
33	1	0	1	0	1	0	4.5	1
38	0	0	1	1	0	0	3.4	0
50	0	0	1	1	0	0	3.7	1
43	1	1	1	0	0	0	3.6	1

Professor Evaluation Scores

X =
features

41	0	0	1	0	1	0	4.1
37	1	0	1	1	1	1	4.2
34	1	0	1	0	1	0	4.8
44	1	0	0	0	0	0	4.4
57	0	1	1	0	1	0	4.7
39	1	1	1	1	1	1	4.2
33	1	0	1	0	1	0	4.5
38	0	0	1	1	0	0	3.4
50	0	0	1	1	0	0	3.7
43	1	1	1	0	0	0	3.6

y =
target

1
1
1
1
1
1
1
0
1
1

Feature Engineering

Machine Learning

GRE	GPA	University Rating	TOEFL	SOP	LOR	Research	Admitted
337	3.89	4	118	4.5	4.5	1	Yes
324	3.58	4	107	4.0	4.5	1	Yes
316	3.23	3	104	3.0	3.5	1	No
322	3.50	3	110	3.5	2.5	1	Yes
314	3.31	2	103	2.0	3.0	0	No
330	3.77	5	115	4.5	3.0	1	Yes
321	3.31	3	109	3.0	4.0	1	Yes

Machine Learning

GRE	GPA	University Rating	TOEFL	SOP	LOR	Research	Admitted
337	3.89	4	118	4.5	4.5	1	Yes
324	3.58	4	107	4.0	4.5	1	Yes
316	3.23	3	104	3.0	3.5	1	No
322	3.50	3	110	3.5	2.5	1	Yes
314	3.31	2	103	2.0	3.0	0	No
330	3.77	5	115	4.5	3.0	1	Yes
321	3.31	3	109	3.0	4.0	1	Yes

Machine Learning

GRE	GPA	University Rating	TOEFL	SOP	LOR	Research	Admitted
337	3.89	4	118	4.5	4.5	1	Yes
324	3.58	4	107	4.0	4.5	1	Yes
316	3.23	3	104	3.0	3.5	1	No
322	3.50	3	110	3.5	2.5	1	Yes
314	3.31	2	103	2.0	3.0	0	No
330	3.77	5	115	4.5	3.0	1	Yes
321	3.31	3	109	3.0	4.0	1	Yes

Target

Machine Learning

Features

GRE	GPA	University Rating	TOEFL	SOP	LOR	Research	Admitted
337	3.89	4	118	4.5	4.5	1	Yes
324	3.58	4	107	4.0	4.5	1	Yes
316	3.23	3	104	3.0	3.5	1	No
322	3.50	3	110	3.5	2.5	1	Yes
314	3.31	2	103	2.0	3.0	0	No
330	3.77	5	115	4.5	3.0	1	Yes
321	3.31	3	109	3.0	4.0	1	Yes

Data Analysis

Correlation

	GRE Score	TOEFL Score	University Rating	SOP	LOR	GPA	Research
GRE Score	1.000000	0.835977	0.668976	0.612831	0.557555	0.832669	0.580391
TOEFL Score	0.835977	1.000000	0.695590	0.657981	0.567721	0.828019	0.489858
University Rating	0.668976	0.695590	1.000000	0.734523	0.660123	0.745734	0.447783
SOP	0.612831	0.657981	0.734523	1.000000	0.729593	0.718403	0.444029
LOR	0.557555	0.567721	0.660123	0.729593	1.000000	0.669626	0.396859
GPA	0.832669	0.828019	0.745734	0.718403	0.669626	1.000000	0.521163
Research	0.580391	0.489858	0.447783	0.444029	0.396859	0.521163	1.000000

Data Analysis

Correlation

	GRE Score	TOEFL Score	University Rating	SOP
GRE Score	1.000000	0.835977	0.668976	0.612831
TOEFL Score	0.835977	1.000000	0.695590	0.657981
University Rating	0.668976	0.695590	1.000000	0.734523
SOP	0.612831	0.657981	0.734523	1.000000
LOR	0.557555	0.567721	0.660123	0.729593
GPA	0.832669	0.828019	0.745734	0.718403
Research	0.580391	0.489858	0.447783	0.444029

Data Analysis

GRE	GPA	University Rating	TOEFL	SOP	LOR	Research	Admitted
337	3.89	4	118	4.5	4.5	1	1
324	3.58	4	107	4.0	4.5	1	1
316	3.23	3	104	3.0	3.5	1	0
322	3.50	3	110	3.5	2.5	1	1
314	3.31	2	103	2.0	3.0	0	0
330	3.77	5	115	4.5	3.0	1	1
321	3.31	3	109	3.0	4.0	1	1

Feature Engineering

GRE	GPA
337	3.89
324	3.58
316	3.23
322	3.50
314	3.31
330	3.77
321	3.31

Feature Engineering

GRE	GPA	Performance Score
337	3.89	
324	3.58	
316	3.23	
322	3.50	
314	3.31	
330	3.77	
321	3.31	

Feature Engineering

GRE	GPA	Performance Score
337	**389**	
324	**358**	
316	**323**	
322	**350**	
314	**331**	
330	**377**	
321	**331**	

Feature Engineering

GRE	GPA	Performance Score
337	389	**363**
324	358	**341**
316	323	**319.5**
322	350	**336**
314	331	**322.5**
330	377	**353.5**
321	331	**326**

Data Analysis

Feature Engineering

GRE	GPA	Performance Score	University Rating	TOEFL	SOP	LOR	Research	Admitted
337	3.89	363	4	118	4.5	4.5	1	1
324	3.58	341	4	107	4.0	4.5	1	1
316	3.23	319.5	3	104	3.0	3.5	1	0
322	3.50	336	3	110	3.5	2.5	1	1
314	3.31	322.5	2	103	2.0	3.0	0	0
330	3.77	353.5	5	115	4.5	3.0	1	1
321	3.31	326	3	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

GPA	Performance Score	University Rating	TOEFL	SOP	LOR	Research	Admitted
3.89	363	4	118	4.5	4.5	1	1
3.58	341	4	107	4.0	4.5	1	1
3.23	319.5	3	104	3.0	3.5	1	0
3.50	336	3	110	3.5	2.5	1	1
3.31	322.5	2	103	2.0	3.0	0	0
3.77	353.5	5	115	4.5	3.0	1	1
3.31	326	3	109	3.0	4.0	1	1

Feature Engineering

GPA	University Rating
3.89	4
3.58	4
3.23	3
3.50	3
3.31	2
3.77	5
3.31	3

Feature Engineering

GPA	University Rating
3.89	4
3.58	4
3.23	3
3.50	3
3.31	2
3.77	5
3.31	3

Feature Engineering

GPA	University Rating
3.89	4
3.58	4
3.23	3
3.50	3
3.31	2
3.77	5
3.31	3

Feature Engineering

GPA	University Rating	GPA Power
3.89	4	
3.58	4	
3.23	3	
3.50	3	
3.31	2	
3.77	5	
3.31	3	

Feature Engineering

GPA	University Rating	GPA Power
3.89	4	**15.56**
3.58	4	**14.32**
3.23	3	**9.69**
3.50	3	**10.5**
3.31	2	**6.62**
3.77	5	**18.85**
3.31	3	**9.93**

Feature Engineering

GPA	University Rating	GPA Power
3.89	4	**15.56**
3.58	4	**14.32**
3.23	3	**9.69**
3.50	3	**10.5**
3.31	2	**6.62**
3.77	5	**18.85**
3.31	3	**9.93**

Feature Engineering

GPA	University Rating	GPA Power
3.89	4	**15.56**
3.58	4	**14.32**
3.23	3	**9.69**
3.50	3	**10.5**
3.31	2	**6.62**
3.77	5	**18.85**
3.31	3	**9.93**

Feature Engineering

GPA	University Rating	GPA Power
3.89	4	**15.56**
3.58	4	**14.32**
3.23	3	**9.69**
3.50	3	**10.5**
3.31	2	**6.62**
3.77	5	**18.85**
3.31	3	**9.93**

Data Analysis

Feature Engineering

GPA	Performance Score	University Rating	GPA Power	TOEFL	SOP	LOR	Research	Admitted
3.89	363	4	15.56	118	4.5	4.5	1	1
3.58	341	4	14.32	107	4.0	4.5	1	1
3.23	319.5	3	9.69	104	3.0	3.5	1	0
3.50	336	3	10.5	110	3.5	2.5	1	1
3.31	322.5	2	6.62	103	2.0	3.0	0	0
3.77	353.5	5	18.85	115	4.5	3.0	1	1
3.31	326	3	9.93	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

Performance Score	University Rating	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	15.56	118	4.5	4.5	1	1
341	4	14.32	107	4.0	4.5	1	1
319.5	3	9.69	104	3.0	3.5	1	0
336	3	10.5	110	3.5	2.5	1	1
322.5	2	6.62	103	2.0	3.0	0	0
353.5	5	18.85	115	4.5	3.0	1	1
326	3	9.93	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

Performance Score	University Rating	SES	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	low	15.56	118	4.5	4.5	1	1
341	4	middle	14.32	107	4.0	4.5	1	1
319.5	3	middle	9.69	104	3.0	3.5	1	0
336	3	low	10.5	110	3.5	2.5	1	1
322.5	2	high	6.62	103	2.0	3.0	0	0
353.5	5	middle	18.85	115	4.5	3.0	1	1
326	3	high	9.93	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

Performance Score	University Rating	SES	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	1	15.56	118	4.5	4.5	1	1
341	4	2	14.32	107	4.0	4.5	1	1
319.5	3	2	9.69	104	3.0	3.5	1	0
336	3	1	10.5	110	3.5	2.5	1	1
322.5	2	3	6.62	103	2.0	3.0	0	0
353.5	5	2	18.85	115	4.5	3.0	1	1
326	3	3	9.93	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

Performance Score	University Rating	SES	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	1	15.56	118	4.5	4.5	1	1
341	4	2	14.32	107	4.0	4.5	1	1
319.5	3	2	9.69	104	3.0	3.5	1	0
336	3	1	10.5	110	3.5	2.5	1	1
322.5	2	3	6.62	103	2.0	3.0	0	0
353.5	5	2	18.85	115	4.5	3.0	1	1
326	3	3	9.93	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

Performance Score	University Rating	SES	Grit	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	1		15.56	118	4.5	4.5	1	1
341	4	2		14.32	107	4.0	4.5	1	1
319.5	3	2		9.69	104	3.0	3.5	1	0
336	3	1		10.5	110	3.5	2.5	1	1
322.5	2	3		6.62	103	2.0	3.0	0	0
353.5	5	2		18.85	115	4.5	3.0	1	1
326	3	3		9.93	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

Performance Score	University Rating	SES	Grit	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	1		15.56	118	4.5	4.5	1	1
341	4	2		14.32	107	4.0	4.5	1	1
319.5	3	2		9.69	104	3.0	3.5	1	0
336	3	1		10.5	110	3.5	2.5	1	1
322.5	2	3		6.62	103	2.0	3.0	0	0
353.5	5	2		18.85	115	4.5	3.0	1	1
326	3	3		9.93	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

Performance Score	University Rating	SES	Grit	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	1	4.0	15.56	118	4.5	4.5	1	1
341	4	2	2.0	14.32	107	4.0	4.5	1	1
319.5	3	2	1.5	9.69	104	3.0	3.5	1	0
336	3	1	3	10.5	110	3.5	2.5	1	1
322.5	2	3	0.67	6.62	103	2.0	3.0	0	0
353.5	5	2	2.5	18.85	115	4.5	3.0	1	1
326	3	3	1	9.93	109	3.0	4.0	1	1

Data Analysis

Feature Engineering

Performance Score	University Rating	Grit	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	4.0	15.56	118	4.5	4.5	1	1
341	4	2.0	14.32	107	4.0	4.5	1	1
319.5	3	1.5	9.69	104	3.0	3.5	1	0
336	3	3	10.5	110	3.5	2.5	1	1
322.5	2	0.67	6.62	103	2.0	3.0	0	0
353.5	5	2.5	18.85	115	4.5	3.0	1	1
326	3	1	9.93	109	3.0	4.0	1	1

Data set

Performance Score	University Rating	Grit	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	4.0	15.56	118	4.5	4.5	1	1
341	4	2.0	14.32	107	4.0	4.5	1	1
319.5	3	1.5	9.69	104	3.0	3.5	1	0
336	3	3	10.5	110	3.5	2.5	1	1
322.5	2	0.67	6.62	103	2.0	3.0	0	0
353.5	5	2.5	18.85	115	4.5	3.0	1	1
326	3	1	9.93	109	3.0	4.0	1	1

Data set

Performance Score	University Rating	Grit	GPA Power	TOEFL	SOP	LOR	Research	Admitted
363	4	4.0	15.56	118	4.5	4.5	1	1
341	4	2.0	14.32	107	4.0	4.5	1	1
319.5	3	1.5	9.69	104	3.0	3.5	1	0
336	3	3	10.5	110	3.5	2.5	1	1
322.5	2	0.67	6.62	103	2.0	3.0	0	0
353.5	5	2.5	18.85	115	4.5	3.0	1	1
326	3	1	9.93	109	3.0	4.0	1	1

Data set

features

target

363	4	4.0	15.56	118	4.5	4.5	1
341	4	2.0	14.32	107	4.0	4.5	1
319.5	3	1.5	9.69	104	3.0	3.5	1
336	3	3	10.5	110	3.5	2.5	1
322.5	2	0.67	6.62	103	2.0	3.0	0
353.5	5	2.5	18.85	115	4.5	3.0	1
326	3	1	9.93	109	3.0	4.0	1

1
1
0
1
0
1
1

Data set

X = features

y = target

363	4	4.0	15.56	118	4.5	4.5	1
341	4	2.0	14.32	107	4.0	4.5	1
319.5	3	1.5	9.69	104	3.0	3.5	1
336	3	3	10.5	110	3.5	2.5	1
322.5	2	0.67	6.62	103	2.0	3.0	0
353.5	5	2.5	18.85	115	4.5	3.0	1
326	3	1	9.93	109	3.0	4.0	1

1
1
0
1
0
1
1

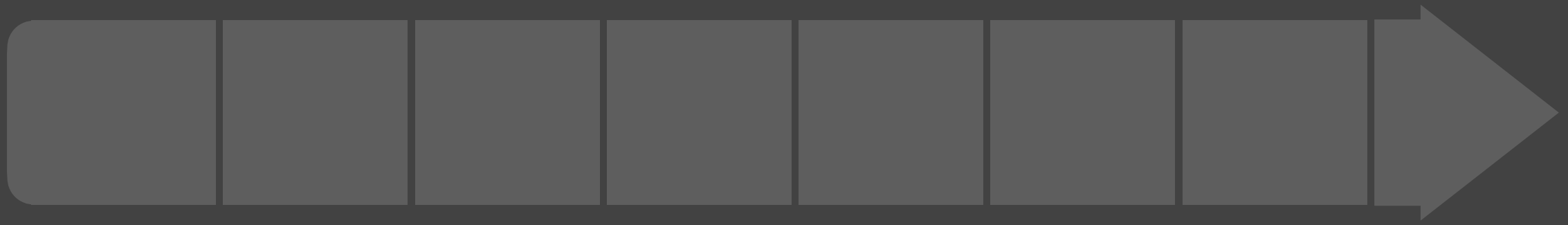
The Data Science Process

- Ask a question
- Get the data
- Clean and transform the data
- Exploratory data analysis
- Feature engineering and selection
- Machine learning
- Evaluate the trained model
- Communicate results/Deploy the model

Most data scientists spend:

- 60% of their time cleaning and preparing data
- 19% of their time collecting datasets
- 9% of their time analyzing the data for patterns
- 3% of their time using ML to train their datasets
- 4% of their time refining their ML algorithms
- 5% of their time on other tasks

The Data Science Process

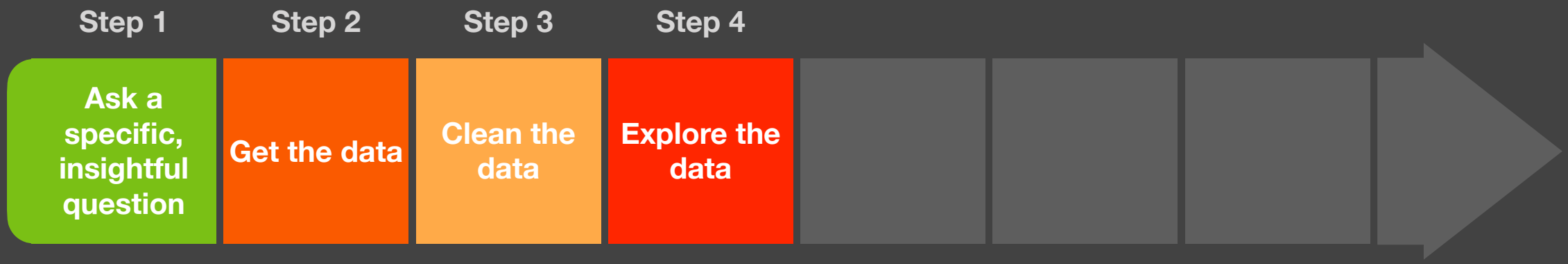


The Data Science Process

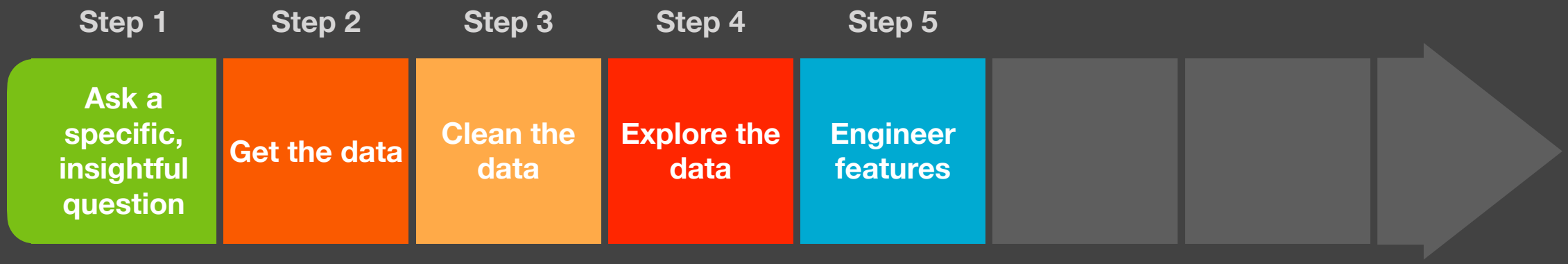
The Data Science Process

The Data Science Process

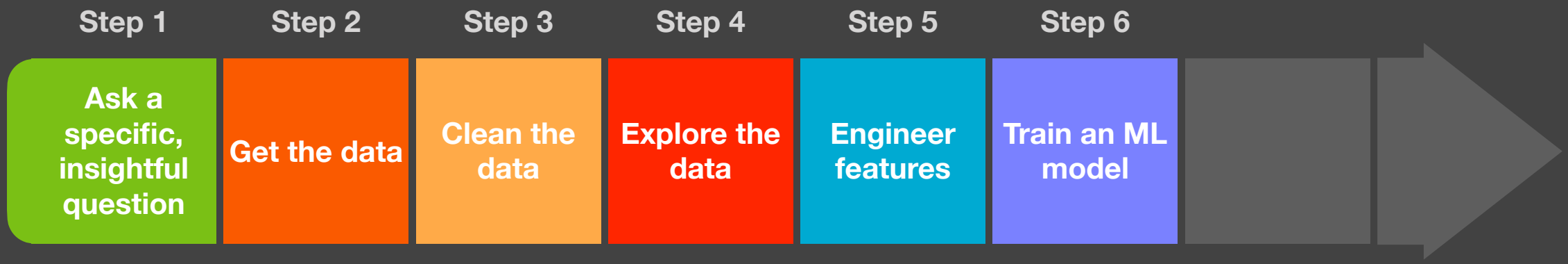
The Data Science Process



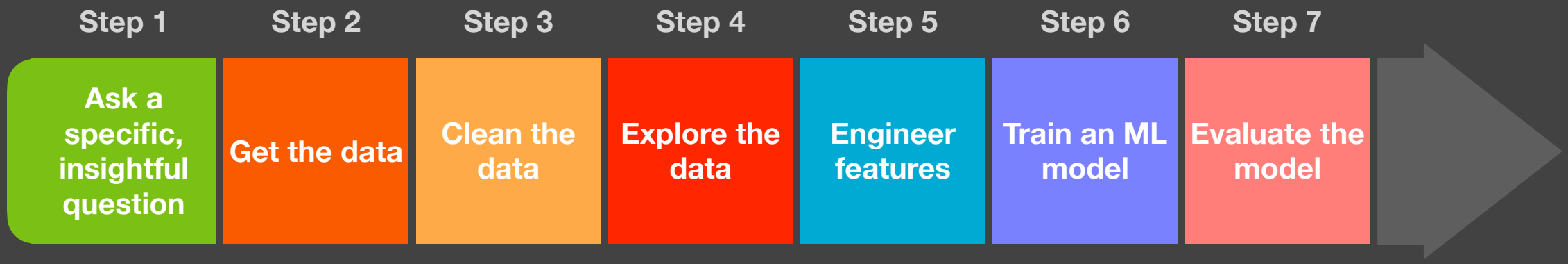
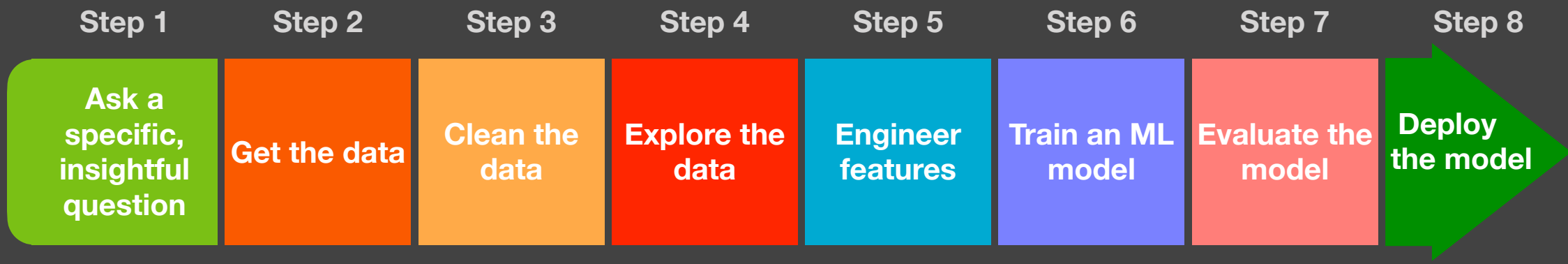
The Data Science Process



The Data Science Process



The Data Science Process



The Data Science Process