Exploratory data analysis, EDA

Nazgul Rakhimzhanova

COURSE SCHEDULE

week	Mid Term (weeks 01-07)	End Term (weeks 08-14)	week
01	Intro: Data Science Area and open source tools for Data Science	Statistics: Distribution – Lognormal, Exponential	08
02	NumPy package for data science	Sampling and Estimation	09
03	Pandas package for data science	Correlation and Covariance	10
04	Visualization with matplotlib	Hypothesis testing	11
05	Statistics: Distribution – Normal	Decision Tree	12
06	Exploratory Data Analysis (EDA)	Linear Regression	13
07	**Summary for 6 weeks QA session**	**Summary for 6 weeks QA session**	**14**
15	**Course summary**		

OUTLINE

- Previously
- EDA
- Readings

PREVIUOSLY

- Discussed about **distributions**
- Did overview of the **EDA**
- Practiced **with histograms to** identify he distribution

PREVIUOSLY

Have you thought about additional DS application for real life tasks?

What We Do

- Data Collection
- EDA
- Feature Engineering
- Model Training
- Model Improvement

EDA common steps

- **Data specification** (understanding)
  - First look at data
  - How many data types?
  - How many missing values?
- **Handling missing values**
  - Drop / restore
- **Data editing** (correction)
  - Inconvenient formats
  - Messy data – noise, outliers.
  - Categorical data handling
- **Relationships**
  - Patterns, pre-summaries
- Normalization
- Feature Extraction

DATA SPECIFICATION

- Understanding the data
  - **pandas.dataframe.head()**
  - **pandas.dataframe.tail()**
  - **pandas.dataframe.info()**
  - **pandas.dataframe.describe()**
  - **pandas.dataframe[‘column’].value_counts()**

DATA UNDERSTANDING

- Understanding -> Visualization
  - **pandas.dataframe.hist()**
  - **pandas.boxplot()**
  - **matplotlib etc.**

DATA EDITING

- Formats -> Date to DateTime
- Dropping inconsistent data -> Dropping IDs and similar data with no knowledge.
- Dropping NaNs: rule – **>5% -- 30%<**
  - **df.drop('column', axis=1, inplace = True)**
  - **df.drop(np.arrange(10), axis=0, inplace = True)**
  - **df.drop(df[(df['age'] < 18) | (df['age'] > 50)].index, axis = 0, inplace = True)**

HANDLING MISSING VALUES

- Missing values of column
- Missing values of rows

- **How to restore?**

DATA EDITING

- Restoring NaN values:
  - Restore by mean
  - Restore by category mean
  - Restore by median
  - Restore by mode
  - Restore by sliding windows’ mean/median/mode
  - Restore by backward/forward replication
  - Restore by interpolation
  - Restore by filling with 0
    - **NaN != 0**

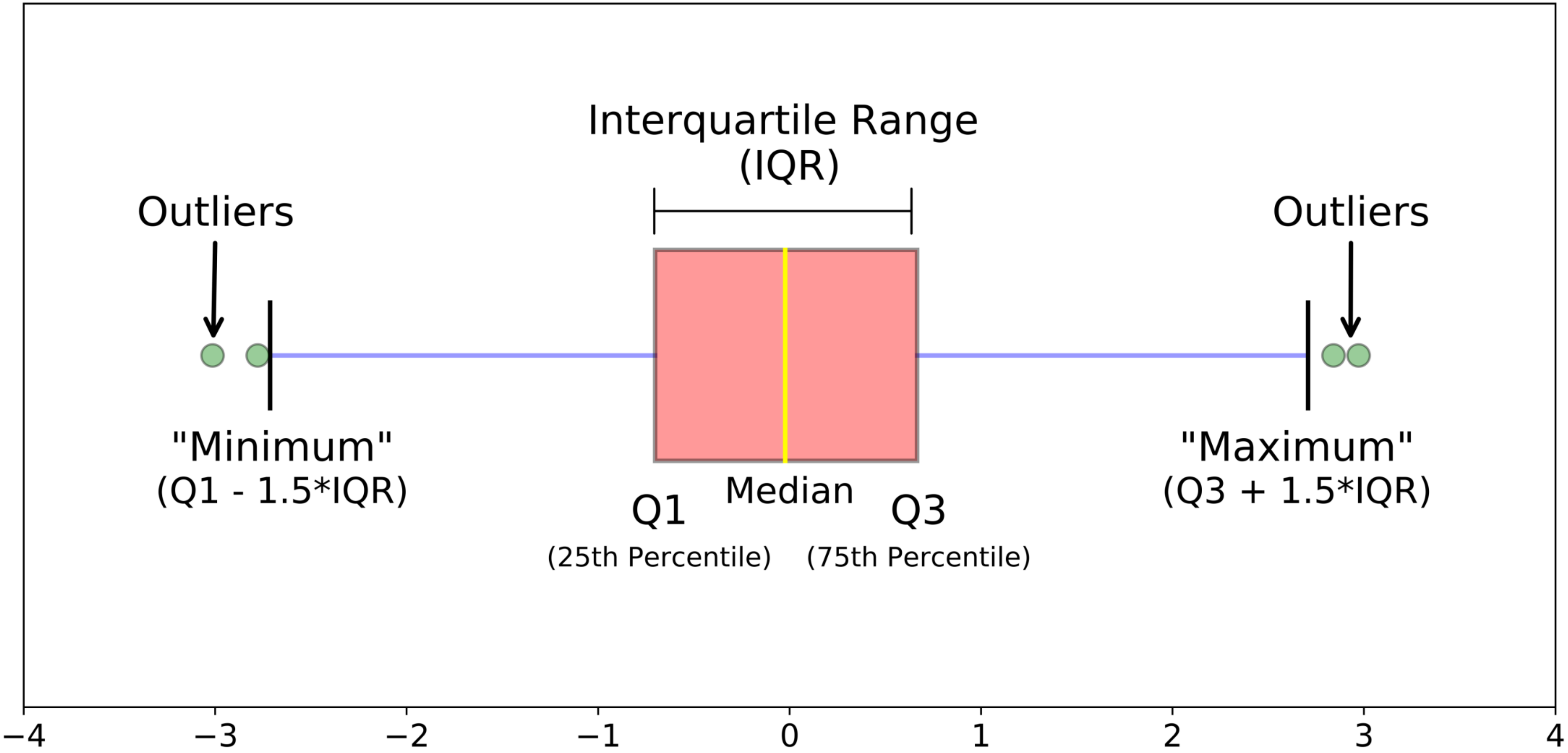
DATA EDITING

DATA EDITING

OUTLIERS

- Boxplots
- Histograms
- IQR

OUTLIERS



Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

CATEGORICAL DATA ENCODING

- **Label** Encoding
- **One-Hot** Encoding
- **Hashing** (HASH function)

DATA ENCODING

- **Label Encoding**

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	ca
0	26	student	single	high.school	no	no	no	telephone	jun	mon	901	1
1	46	admin.	married	university.degree	no	yes	no	cellular	aug	tue	208	2
2	49	blue-collar	married	basic.4y	unknown	yes	yes	telephone	jun	tue	131	5
3	31	technician	married	university.degree	no	no	no	cellular	jul	tue	404	1
4	42	housemaid	married	university.degree	no	yes	no	telephone	nov	mon	85	1

education
university – 6
professional courses – 5
college - 4
high school – 3
basic 12y – 2
basic 9y - 1
basic 4y - 0

martial
Single – 2
Married – 1

loan
no – 0
yes – 2

housing
no – 0
yes - 2

contact
telephone – 1
cellular – 0

DATA ENCODING

- Label Encoding

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	ca
0	26	student	single	high.school	no	no	no	telephone	jun	mon	901	1
1	46	admin.	married	university.degree	no	yes	no	cellular	aug	tue	208	2
2	49	blue-collar	married	basic.4y	unknown	yes	yes	telephone	jun	tue	131	5
3	31	technician	married	university.degree	no	no	no	cellular	jul	tue	404	1
4	42	housemaid	married	university.degree	no	yes	no	telephone	nov	mon	85	1

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign
0	26	8	2	3	0	0	0	1	4	1	901	1
1	46	0	1	6	0	2	0	0	1	3	208	2
2	49	1	1	0	1	2	2	1	4	3	131	5
3	31	9	1	6	0	0	0	0	3	3	404	1
4	42	3	1	6	0	2	0	1	7	1	85	1

DATA ENCODING

- **One-Hot Encoding**

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	ca
0	26	student	single	high.school	no	no	no	telephone	jun	mon	901	1
1	46	admin.	married	university.degree	no	yes	no	cellular	aug	tue	208	2
2	49	blue-collar	married	basic.4y	unknown	yes	yes	telephone	jun	tue	131	5
3	31	technician	married	university.degree	no	no	no	cellular	jul	tue	404	1
4	42	housemaid	married	university.degree	no	yes	no	telephone	nov	mon	85	1

Single	Married
1	0
0	1
0	1
0	1
0	1
0	1
1	0

DATA ENCODING

- **One-Hot Encoding**

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	ca
0	26	student	single	high.school	no	no	no	telephone	jun	mon	901	1
1	46	admin.	married	university.degree	no	yes	no	cellular	aug	tue	208	2
2	49	blue-collar	married	basic.4y	unknown	yes	yes	telephone	jun	tue	131	5
3	31	technician	married	university.degree	no	no	no	cellular	jul	tue	404	1
4	42	housemaid	married	university.degree	no	yes	no	telephone	nov	mon	85	1

	0	1	2	3	4	5	6	7	8	9	...	43	44	45	46	47	48	49	50	51	52
0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
4	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0

RELATIONSHIPS

- Patters
  - **Barplots –** to compare the same parameter in two or more groups
  - **GroupBy –** to see some specific groups or parameters
- Correlations
  - **corr() –** to check if our data happen to have linear dependency

RELATIONSHIPS

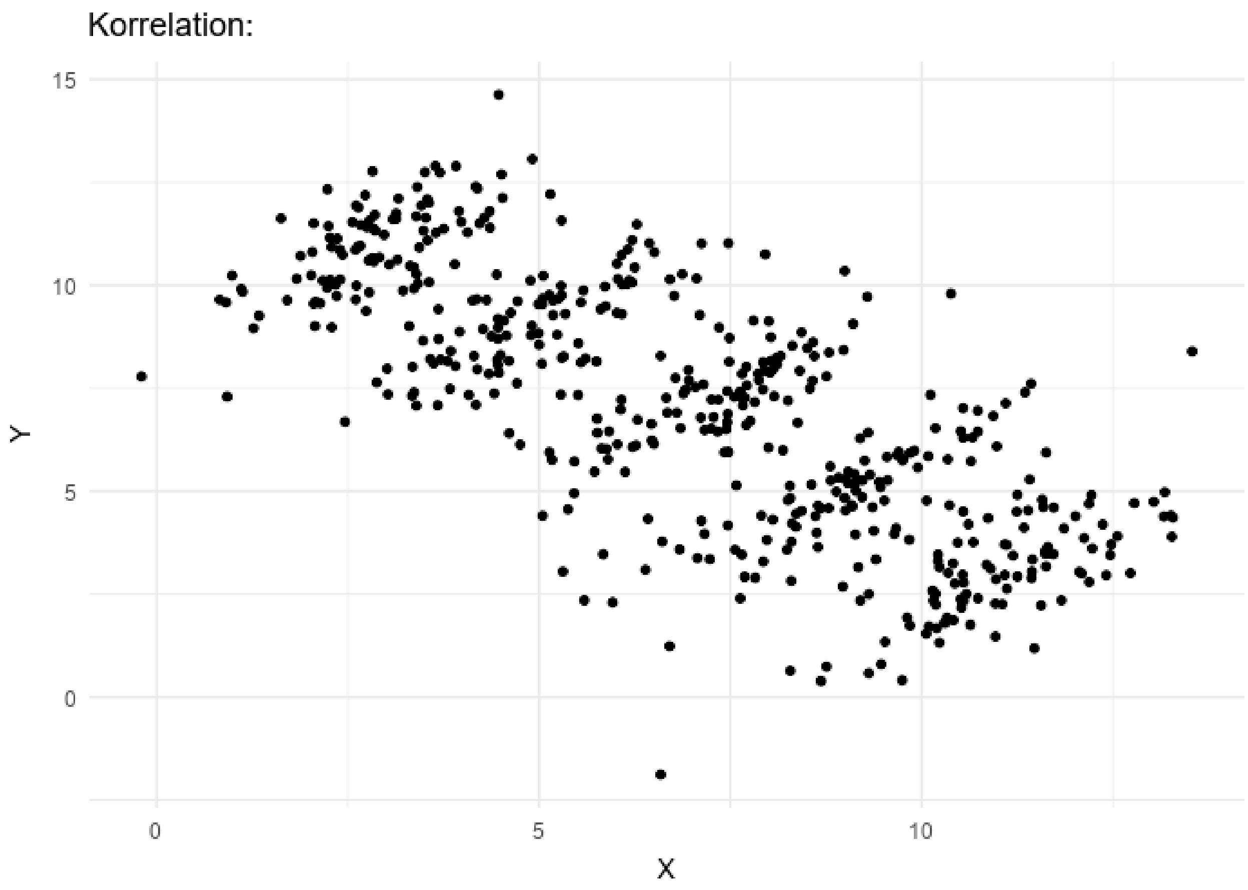
"Never trust a statistic if you haven't falsified it yourself".
Winston Churchill

The Simpson-Paradox

Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

https://en.wikipedia.org/wiki/Simpson%27s_paradox

RELATIONSHIPS



https://en.wikipedia.org/wiki/Simpson%27s_paradox

RELATIONSHIPS

Fact 01 :

Gender	Applicants	Accepted	Rejected	% Accepted
Male	2175	1025	1150	0.471
Female	849	261	588	0.307

University of California, Berkeley, Admission rate 1973 y.

RELATIONSHIPS

Fact 02 :

Gender	Applicants	Accepted	Rejected	% Accepted
Male	825	512	313	0.621
Female	108	89	19	0.824

Data for Department 1.

Gender	Applicants	Accepted	Rejected	% Accepted
Male	417	138	279	0.331
Female	375	131	244	0.349

Data for Department 3.

Gender	Applicants	Accepted	Rejected	% Accepted
Male	560	353	207	0.63
Female	25	17	8	0.68

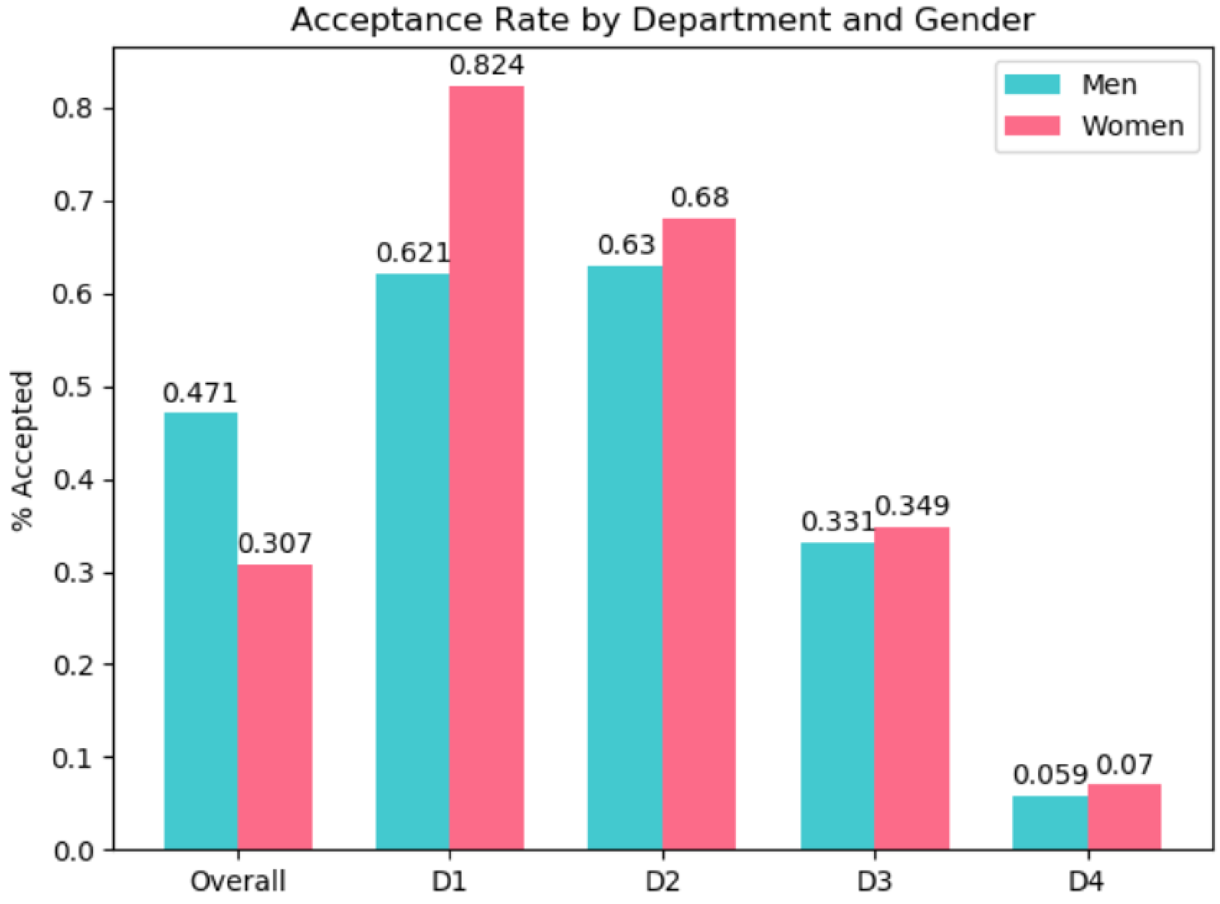
Data for Department 2.

Gender	Applicants	Accepted	Rejected	% Accepted
Male	373	22	351	0.059
Female	341	24	317	0.070

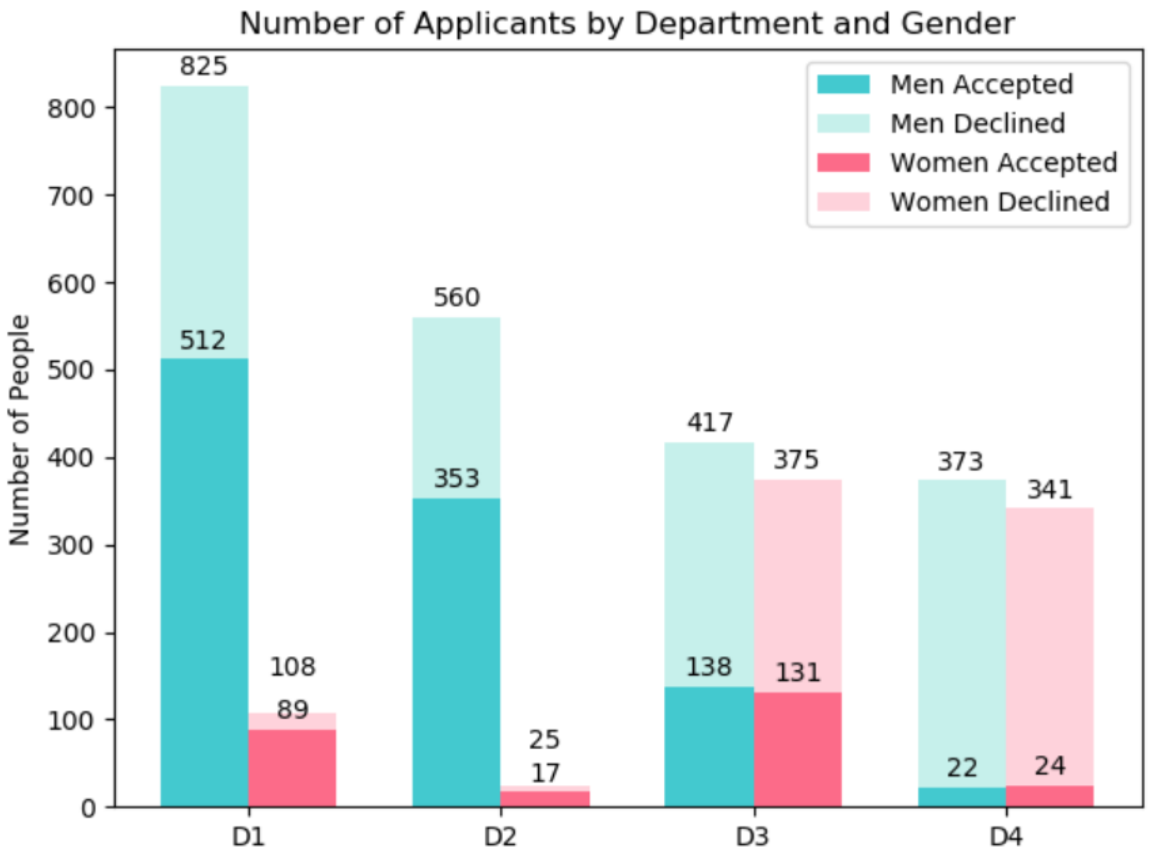
Data for Department 4.

The overall acceptance rate of women is lower than the overall acceptance rate of men. Yet, in each department, the acceptance rate for women is higher than the acceptance rate for men.

RELATIONSHIPS



RELATIONSHIPS



- Department 1 seemed to accept both a high number of people and a high percentage of applicants, yet very **few women applied.**
- The same goes for Department 2.
- In Departments 3 and 4, the number of women who applied was almost the same as the number of men who applied — but the overall acceptance rate was quite low compared to the other departments.

This is also the explanation of why the overall acceptance rate for women is lower than the rate for men. It's not that women were discriminated against by any departments (at least as far as we know!), it's that women — in comparison to men — seemed to apply more to very competitive departments where it was hard to get in.

Readings

- **https://towardsdatascience.com/exploratory-data-analysis-of-kaggle-datasets-9a293886f644**
- https://towardsdatascience.com/gender-bias-in-admission-statistics-the-simpson-paradox-cd381d994b16