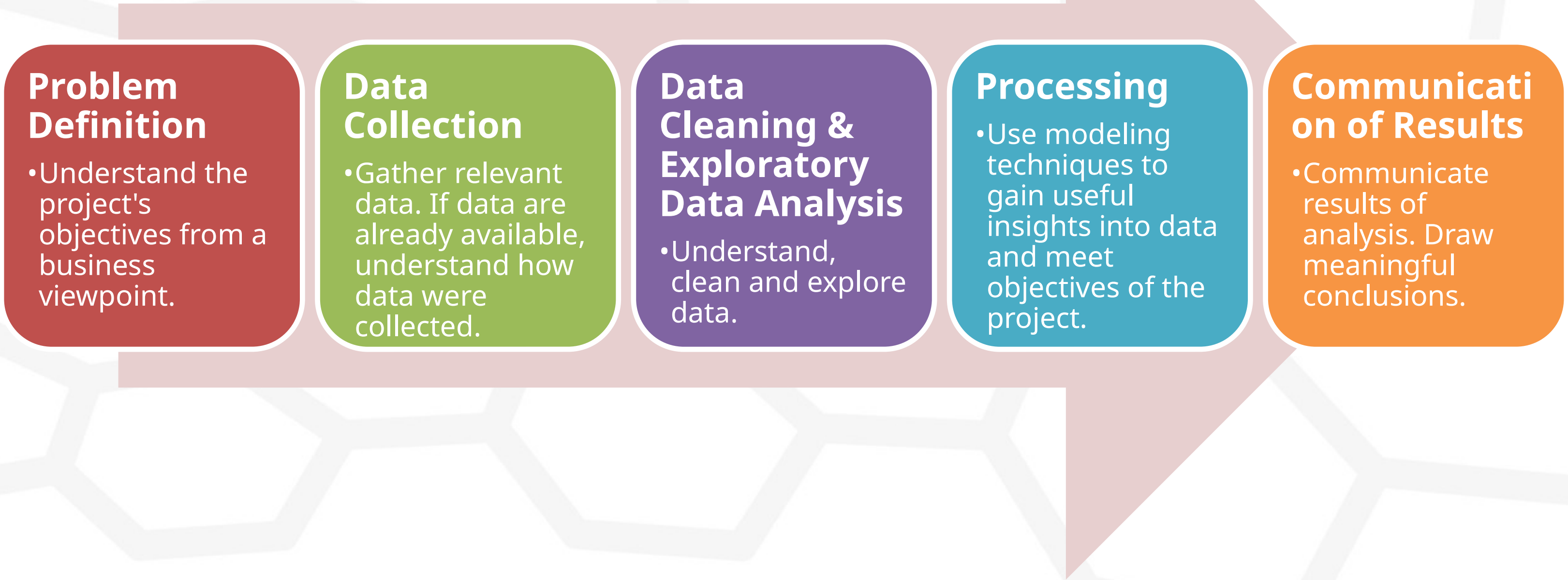
Data Cleaning & Exploratory Data Analysis

CNM Ingenuity, Inc.

Data Science Process



Problem Definition

- Understand the project's objectives from a business viewpoint.

Data Collection

- Gather relevant data. If data are already available, understand how data were collected.

Data Cleaning & Exploratory Data Analysis

- Understand, clean and explore data.

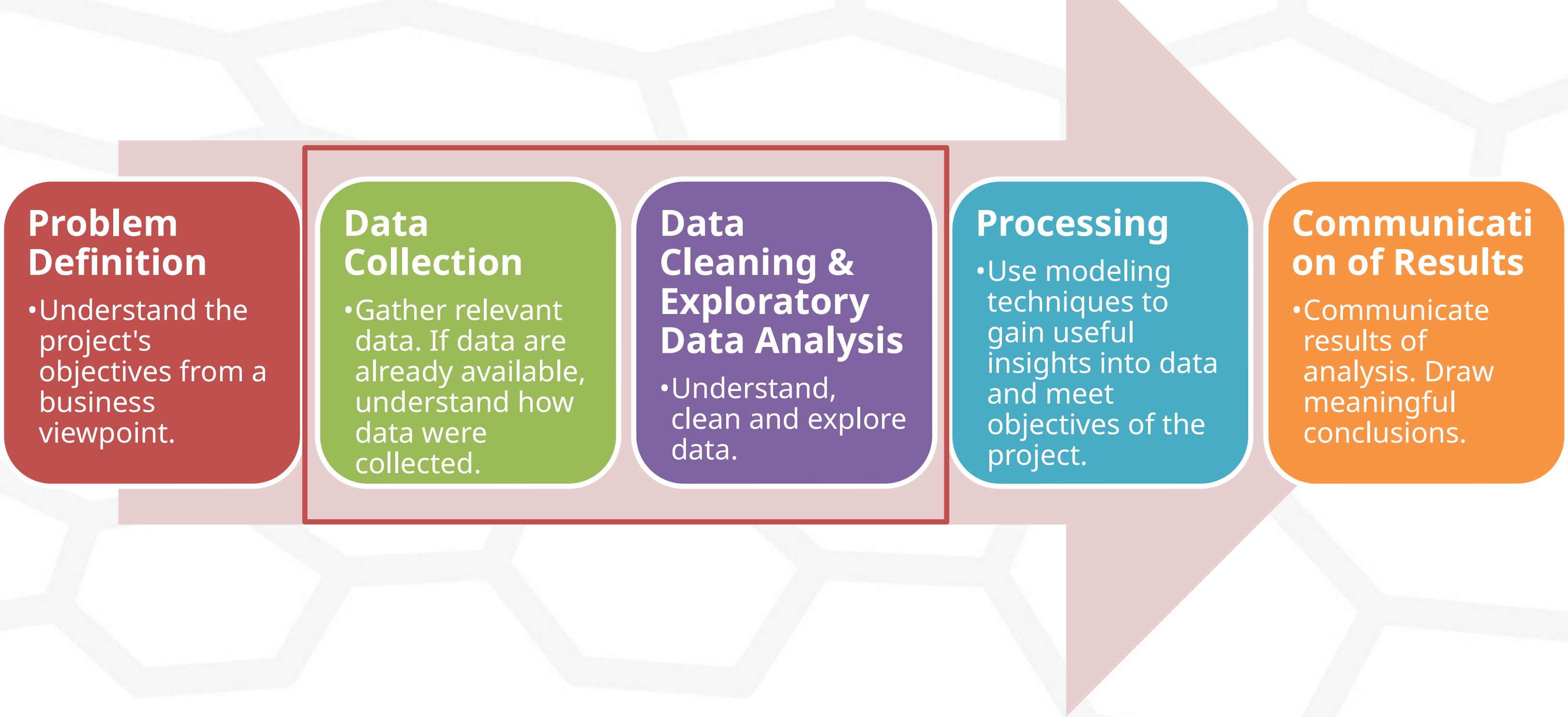
Processing

- Use modeling techniques to gain useful insights into data and meet objectives of the project.

Communication of Results

- Communicate results of analysis. Draw meaningful conclusions.

Data Science Process



Problem Definition

- Understand the project's objectives from a business viewpoint.

Data Collection

- Gather relevant data. If data are already available, understand how data were collected.

Data Cleaning & Exploratory Data Analysis

- Understand, clean and explore data.

Processing

- Use modeling techniques to gain useful insights into data and meet objectives of the project.

Communication of Results

- Communicate results of analysis. Draw meaningful conclusions.

Tidy Data

- It is often said that **80% of data analysis is spent on data cleaning**.
- There is a concept called tidy data that has set a standard for structuring datasets to more easily facilitate analysis.
- What's a relatively easy way to save time and resources? Make sure your team collects data in a tidy format!

> *"Tidy datasets are all alike, but every messy dataset is messy in its own way".*

Source: https://www.jstatsoft.org/article/view/v059i10

Tidy Data Organization

p variables

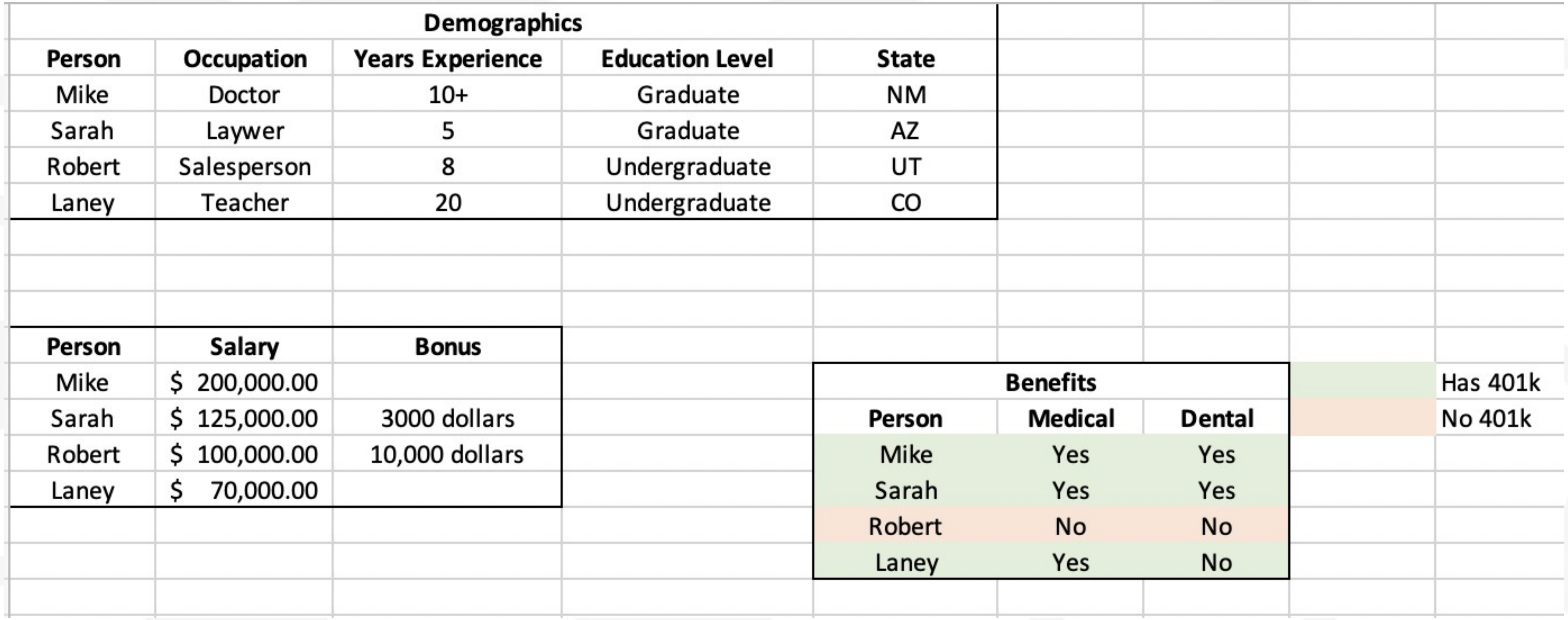
n observations

Patient	Weight	Height	BP	Age	Diabetes
A	150	65	120	35	Yes
B	132	63	140	44	No
C	190	69	130	56	Yes
D	178	70	135	52	No

Three characteristics of tidy data:

1. Every column is a variable.
2. Every row is an observation.
3. Every cell is a single value.

Not So Good Data Structure



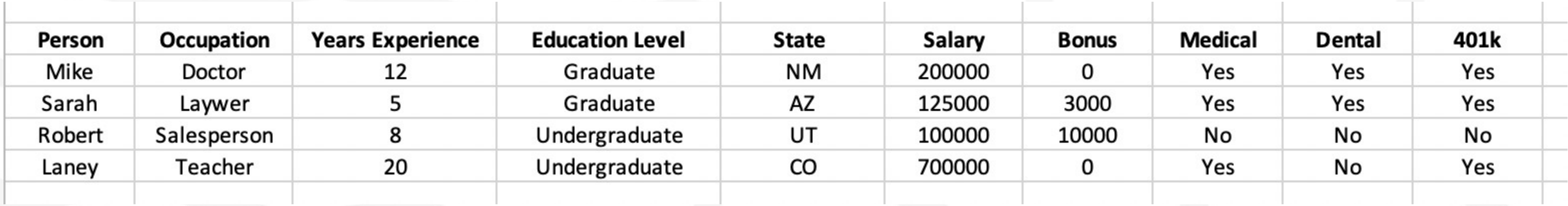
Demographics				
Person	Occupation	Years Experience	Education Level	State
Mike	Doctor	10+	Graduate	NM
Sarah	Laywer	5	Graduate	AZ
Robert	Salesperson	8	Undergraduate	UT
Laney	Teacher	20	Undergraduate	CO

Person	Salary	Bonus
Mike	$ 200,000.00	
Sarah	$ 125,000.00	3000 dollars
Robert	$ 100,000.00	10,000 dollars
Laney	$ 70,000.00	

Benefits		
Person	Medical	Dental
Mike	Yes	Yes
Sarah	Yes	Yes
Robert	No	No
Laney	Yes	No

🟩	Has 401k
🟥	No 401k

Good Data Structure

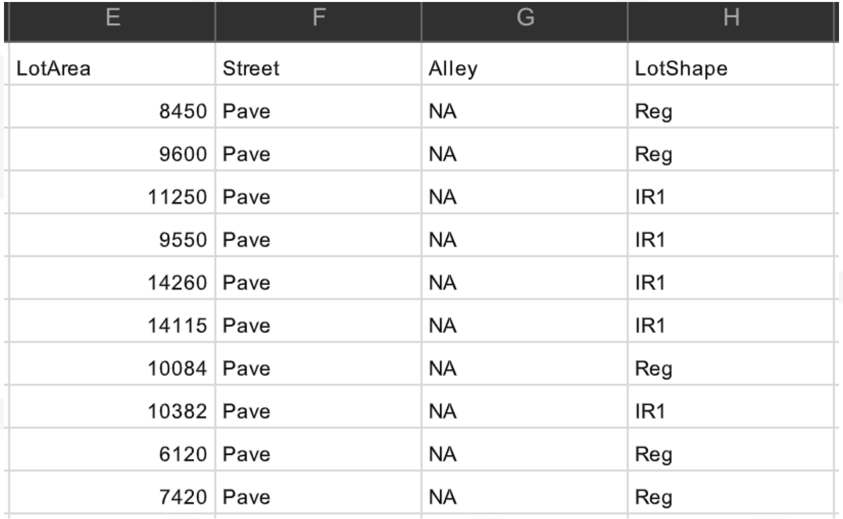


Person	Occupation	Years Experience	Education Level	State	Salary	Bonus	Medical	Dental	401k
Mike	Doctor	12	Graduate	NM	200000	0	Yes	Yes	Yes
Sarah	Laywer	5	Graduate	AZ	125000	3000	Yes	Yes	Yes
Robert	Salesperson	8	Undergraduate	UT	100000	10000	No	No	No
Laney	Teacher	20	Undergraduate	CO	700000	0	Yes	No	Yes

Data Cleaning

Below are some ways data may need to be cleaned

- Remove redundant variables
- Address null/missing values
- Transform data
- Convert data types
- Restructure data



E	F	G	H
LotArea	Street	Alley	LotShape
8450	Pave	NA	Reg
9600	Pave	NA	Reg
11250	Pave	NA	IR1
9550	Pave	NA	IR1
14260	Pave	NA	IR1
14115	Pave	NA	IR1
10084	Pave	NA	Reg
10382	Pave	NA	IR1
6120	Pave	NA	Reg
7420	Pave	NA	Reg

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is used to do the following:

- Understand your data
- Generate summary statistics
- Identify anomalies/outliers
- Understand variables & their relationships
- Help determine next steps in the analysis

Understand Your Data

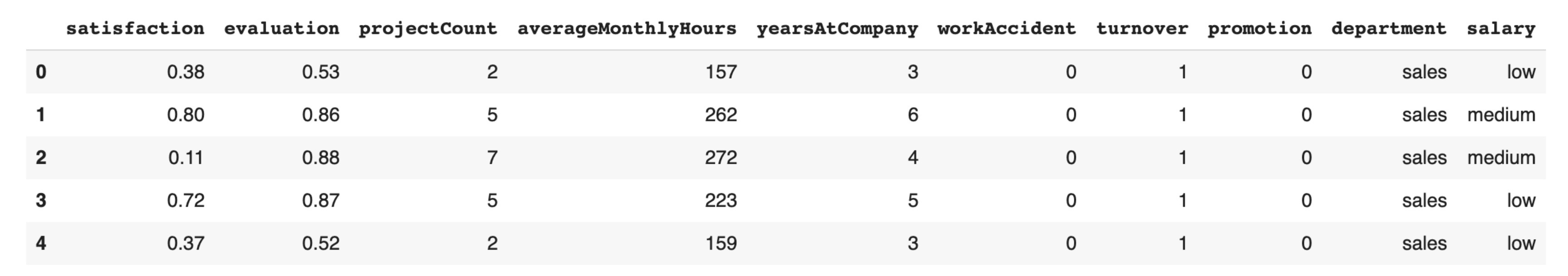
Before performing any analysis, it’s important that you take the time to understand your data and the types of variables you will be dealing with. Some useful pandas methods are given below:

- **.shape** - gives the dimensions of a dataset
- **.head()** - gives the first *n* rows of a dataset. Default is *n = 5*.
- **.describe()** - provides summary statistics of a dataset
- **.info()** - gives information about the columns in your dataset including the datatype and the number of non-null values
- **.columns** - returns a list of the column names
- **.unique()** - returns the unique values of a column

Things to Look Out for In Visualizations

- Trends
- Patterns
- Gaps
- Relationships
- Outliers
- Message

An Example Analysis – HR Retention



	satisfaction	evaluation	projectCount	averageMonthlyHours	yearsAtCompany	workAccident	turnover	promotion	department	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low

What is the business problem? What is the data science problem?

https://www.kaggle.com/randylaosat/predicting-employee-ker

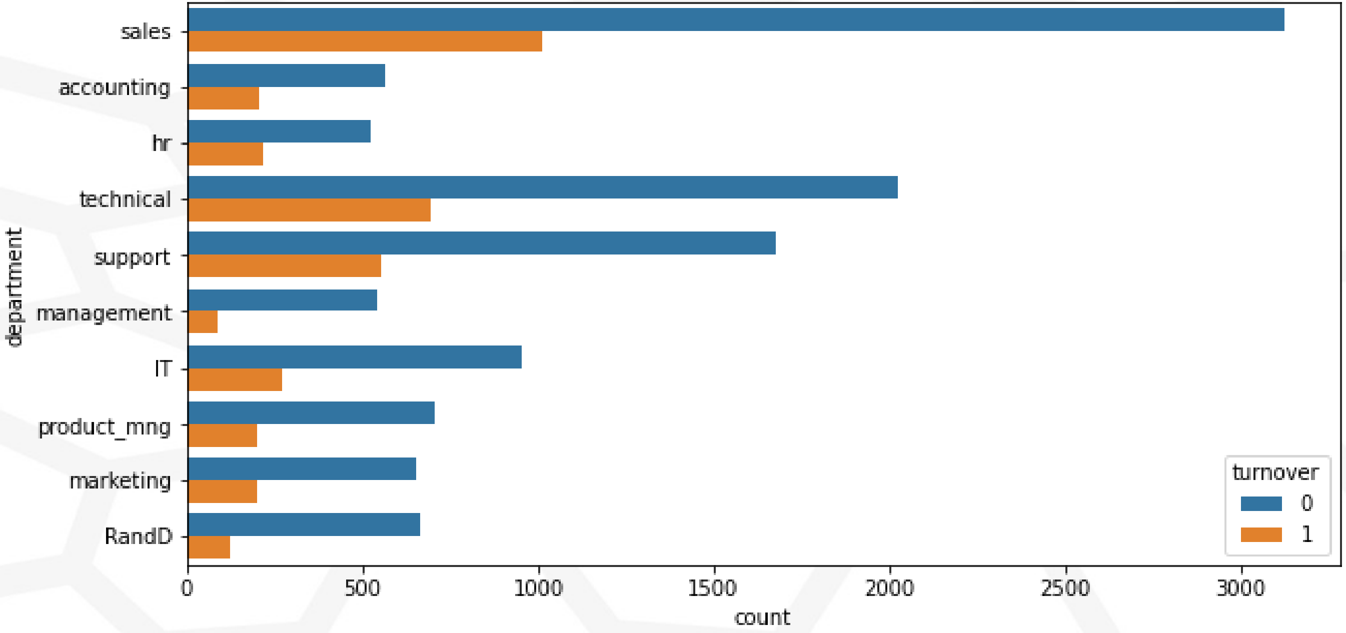
HR Retention Problem Definition

- Our company wants to understand what factors contribute most to employee turnover. We also want to be able to predict turnover for certain employees so we can focus our retention strategies on employees likely to turnover.

- A quarterly HR report will be generated that lists employees that are predicted to turnover.

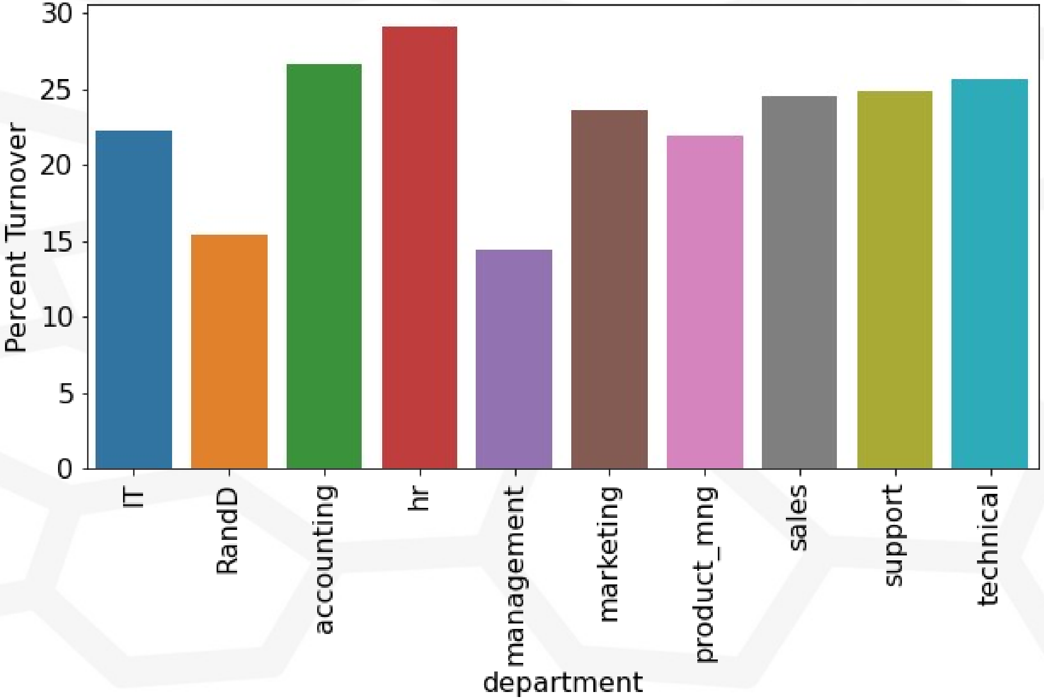
- This is a supervised classification problem.

Department vs Turnover



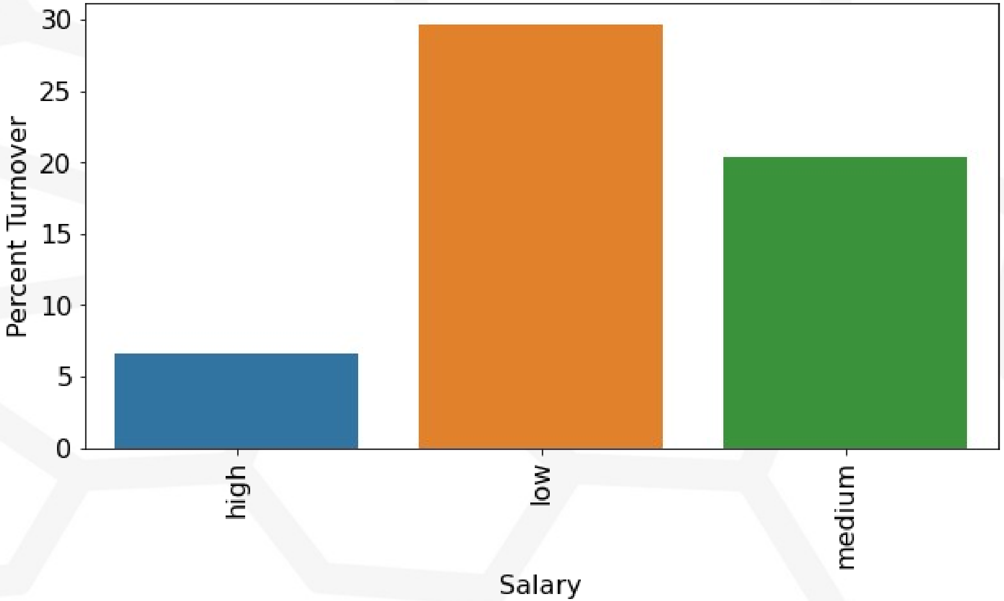
Why might this not be a good way to represent this data?

Department vs Percent Turnover

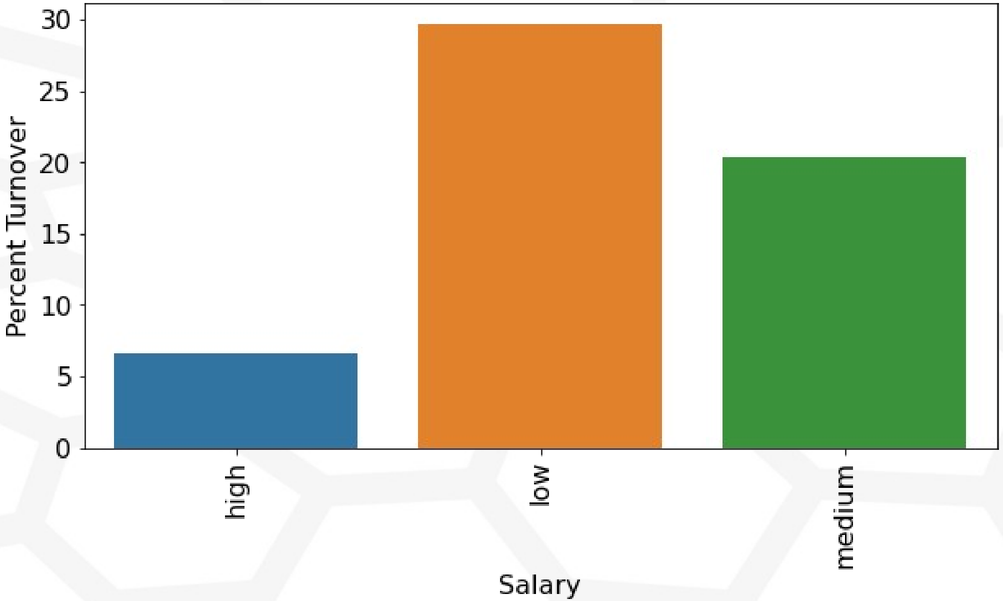


We can see that HR and accounting have the highest percent turnover. Management has the lowest turnover. Why might this be happening?

Salary vs Percent Turnover

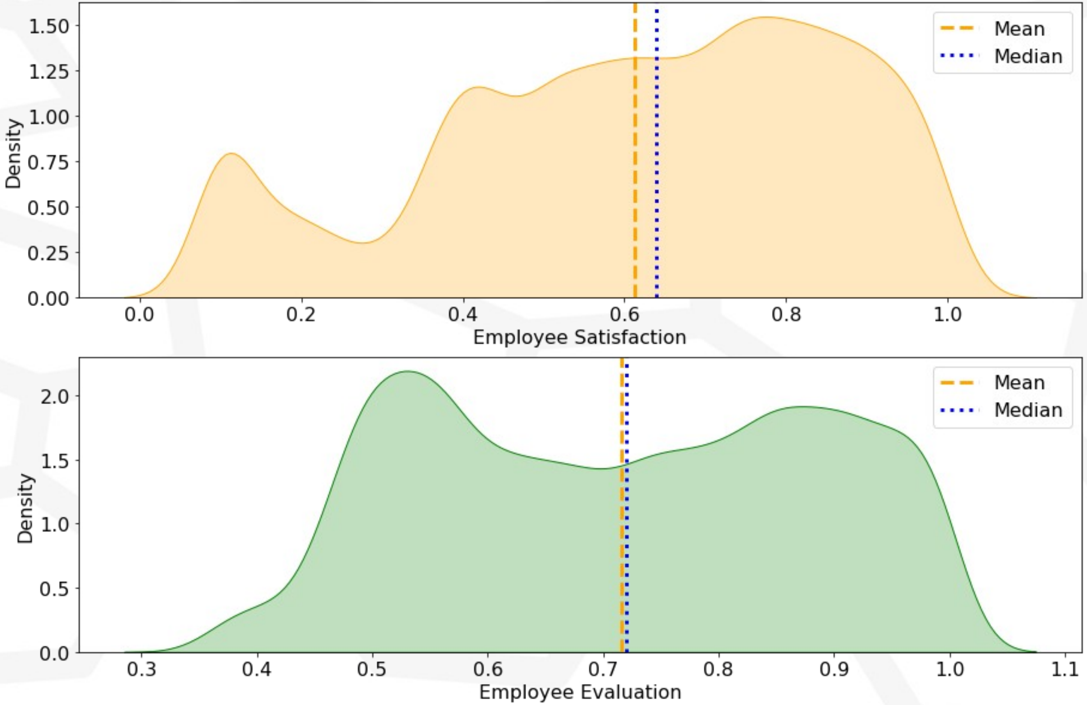


Salary vs Percent Turnover

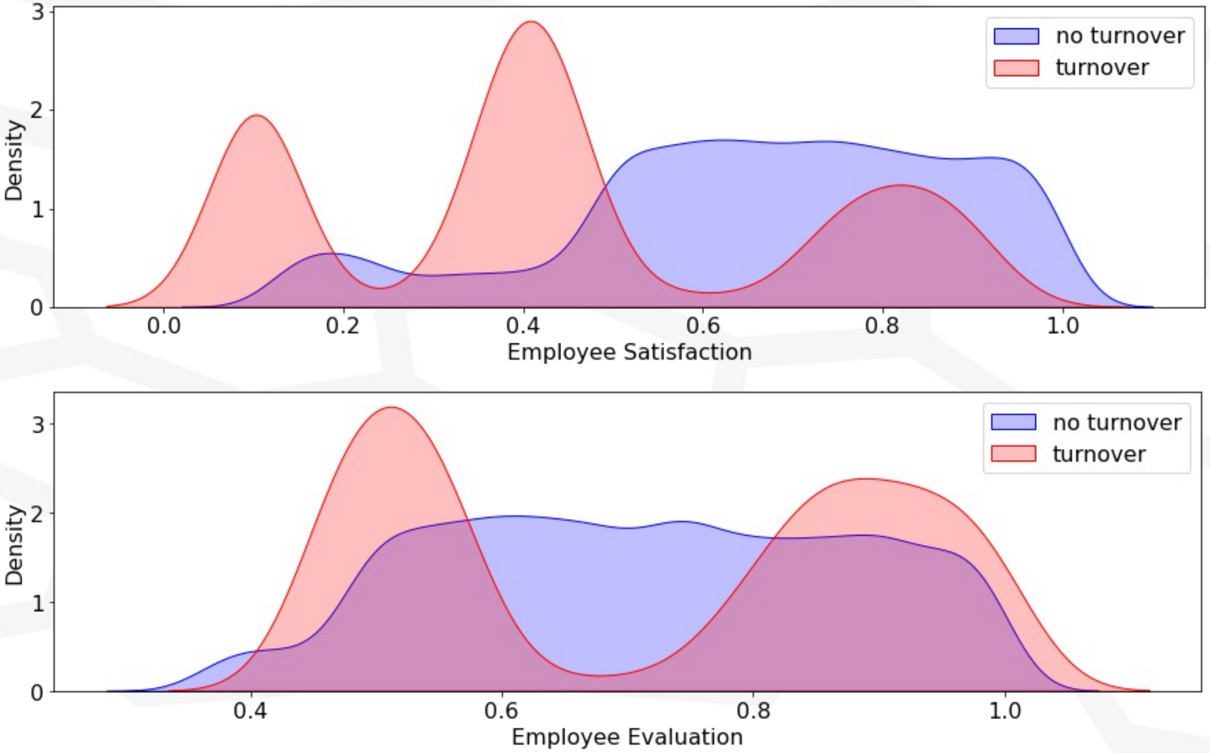


Unsurprisingly, most people who left had low or medium salaries.

Employee Satisfaction & Evaluation

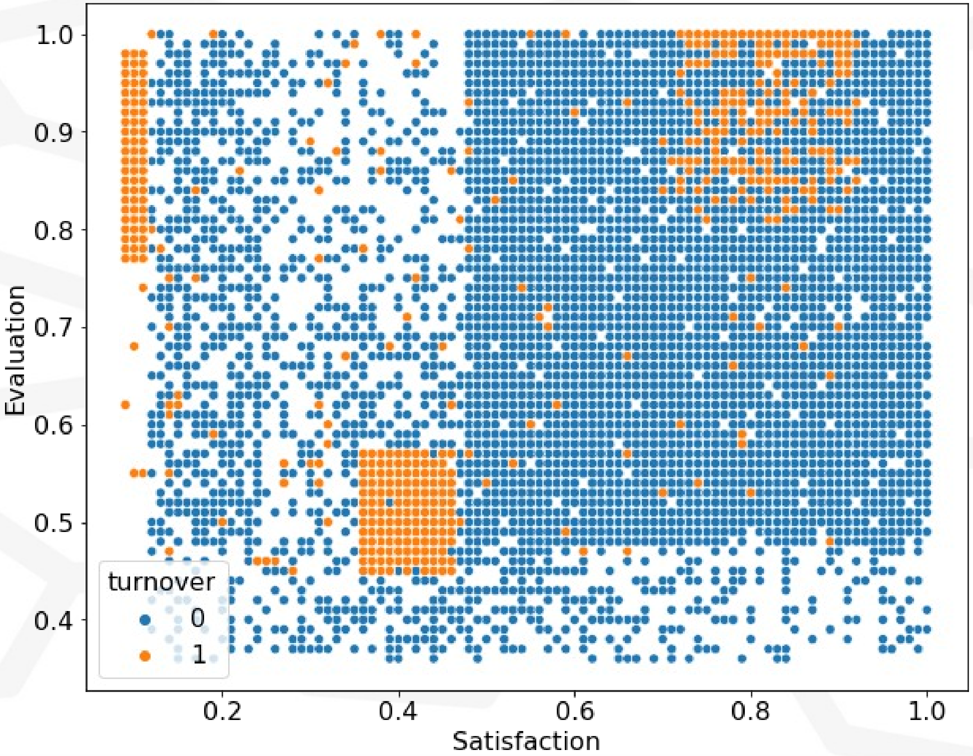


Satisfaction & Evaluation vs Turnover

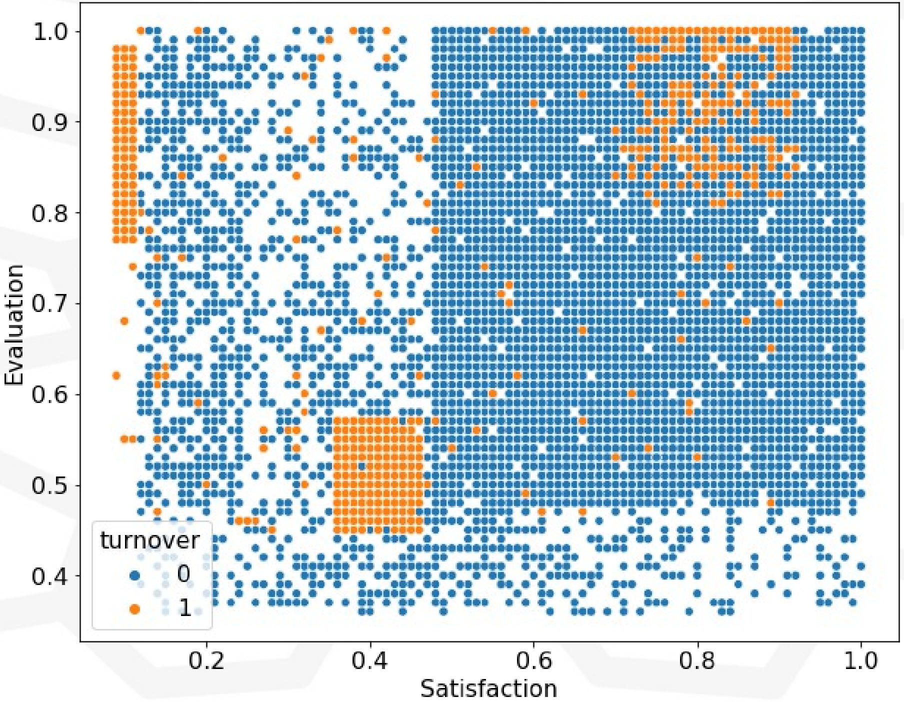


The distributions for turnover vs no turnover are very different. We also see that the turnover group has multiple modes.

Satisfaction & Evaluation vs Turnover Pt 2



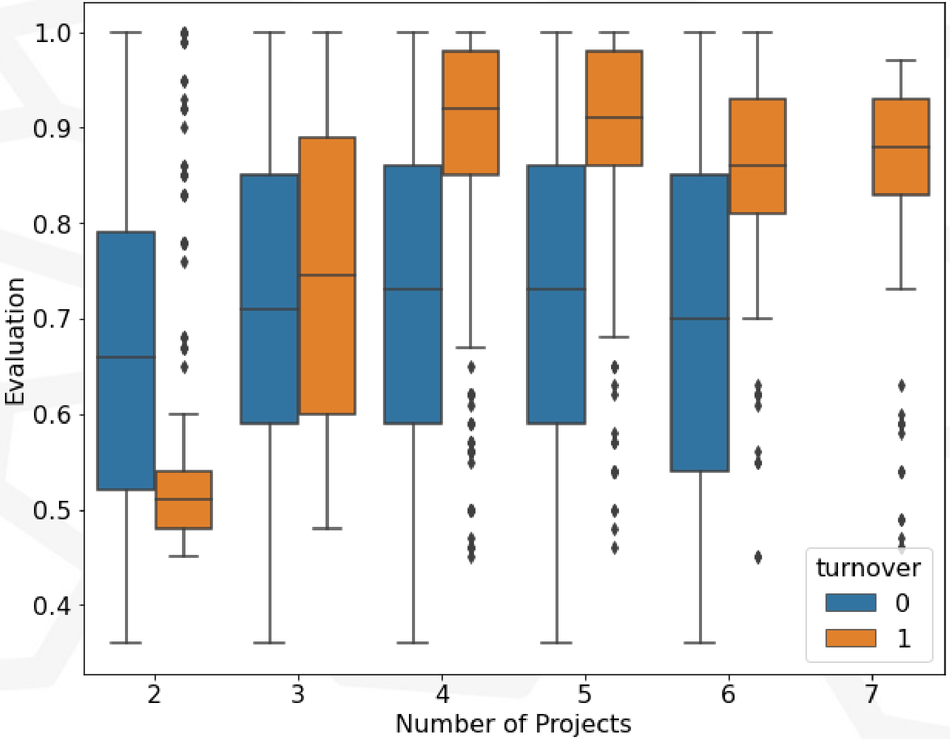
Satisfaction & Evaluation vs Turnover Pt 2



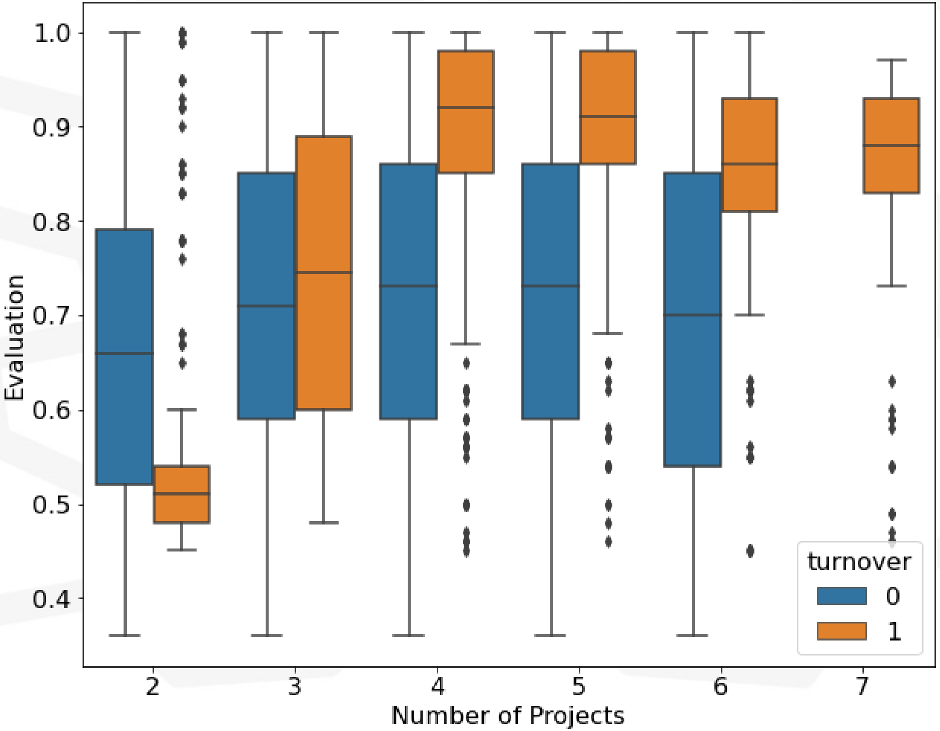
There appear to be three groups of employees who turnover:

1. Low satisfaction, high evaluation
2. Low satisfaction, low, evaluation
3. High satisfaction, high evaluation

Project Count & Evaluation vs Turnover



Project Count & Evaluation vs Turnover



For > 3 projects, employees that have a higher evaluation tend to leave.

What Next?

- Based on our EDA, we can see that there are likely complex relationships between our variables and turnover.
- It's easy to understand why some employees are leaving (e.g., those with low satisfaction, low performance, low salary), but it is more challenging to understand why others are leaving (e.g., those with high satisfaction and high performance).
- Building a model can help us understand these more complex relationships.

Recap

- Think about the population of interest when collecting data.
- Up to 80% of the analysis involves data cleaning - encourage your team to store their data in a tidy format when possible.
- Exploratory data analysis (EDA) is a way to get an understanding of your data and potential relationships between variables.