ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 2.1 | Data Cleaning

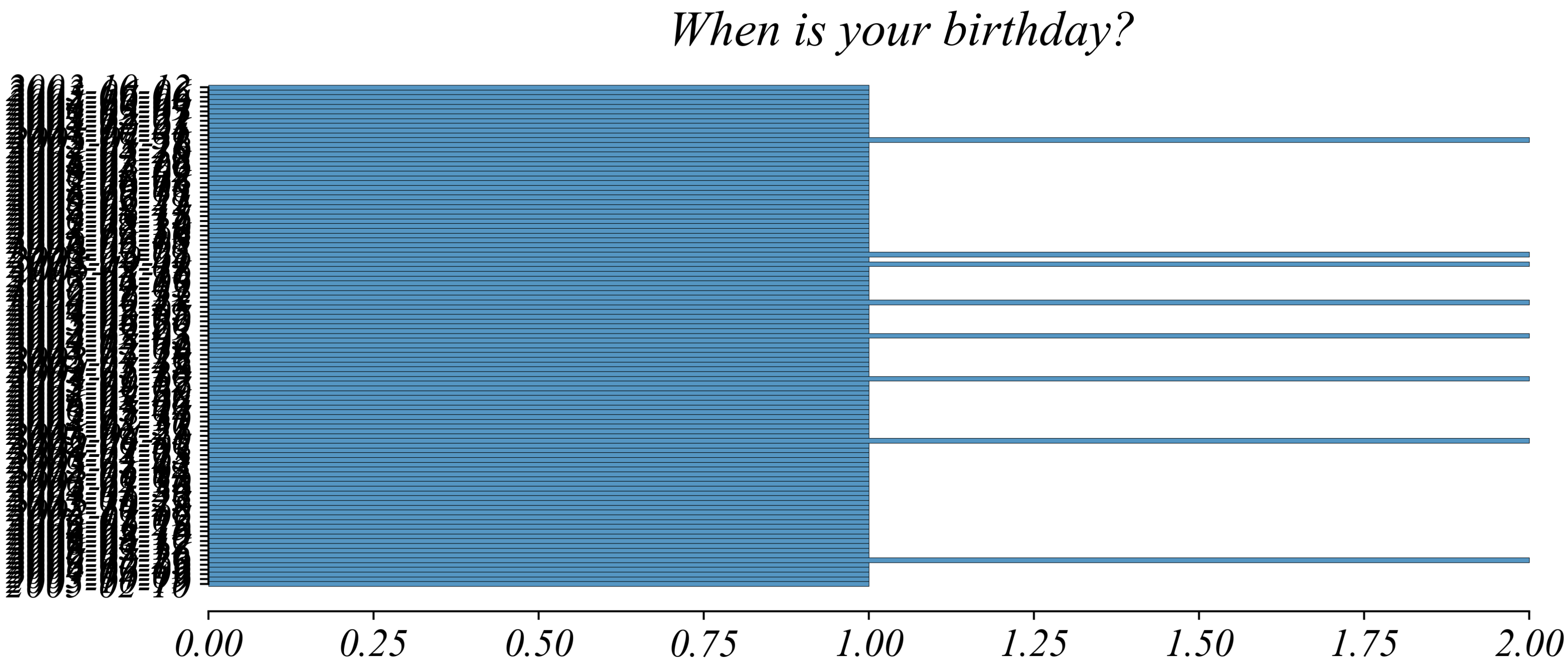
Data Cleaning

Q. Are students who live further away older?

Data Cleaning

Q. Are students who live further away older?

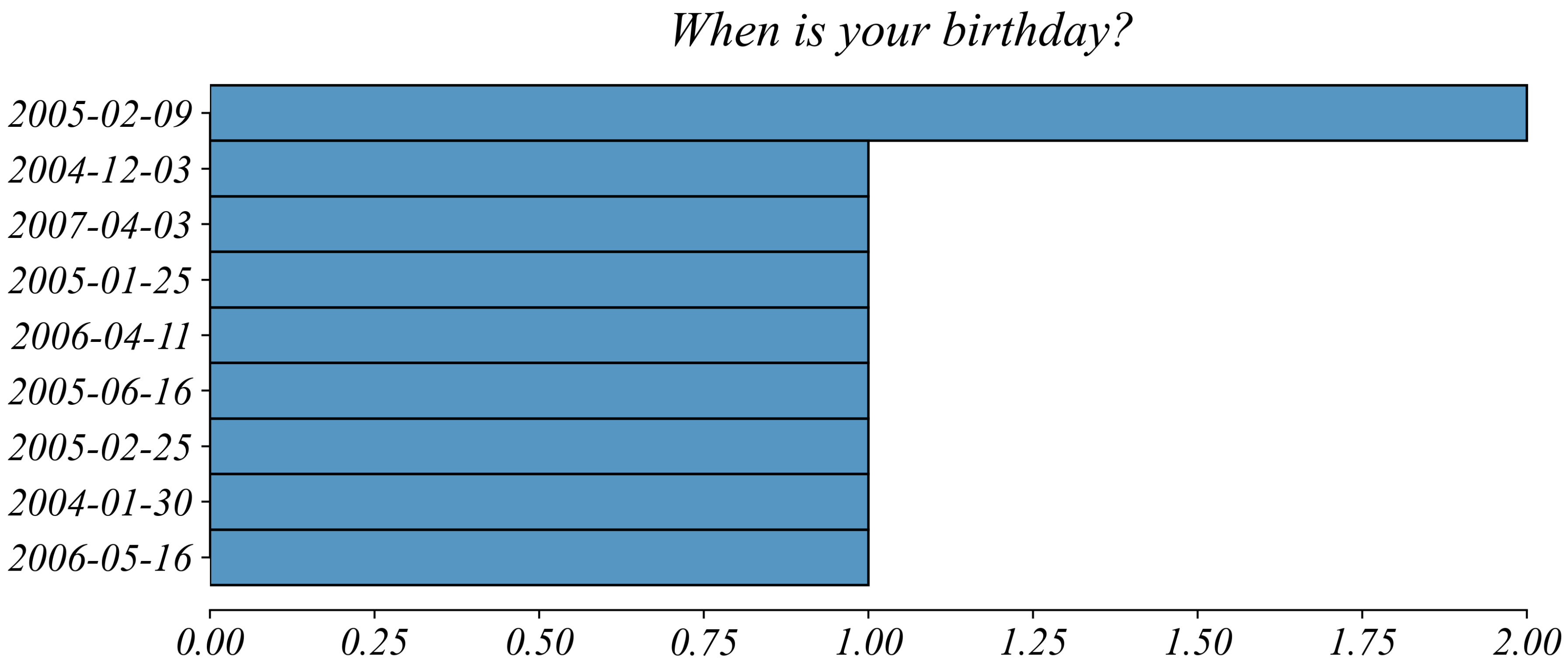
Let’s examine age and distance from Pittsburgh.



Data Cleaning

Q. Are students who live further away older?

Let's examine age and distance from Pittsburgh.



> the birthday data is stored as text: "08/15/2005"

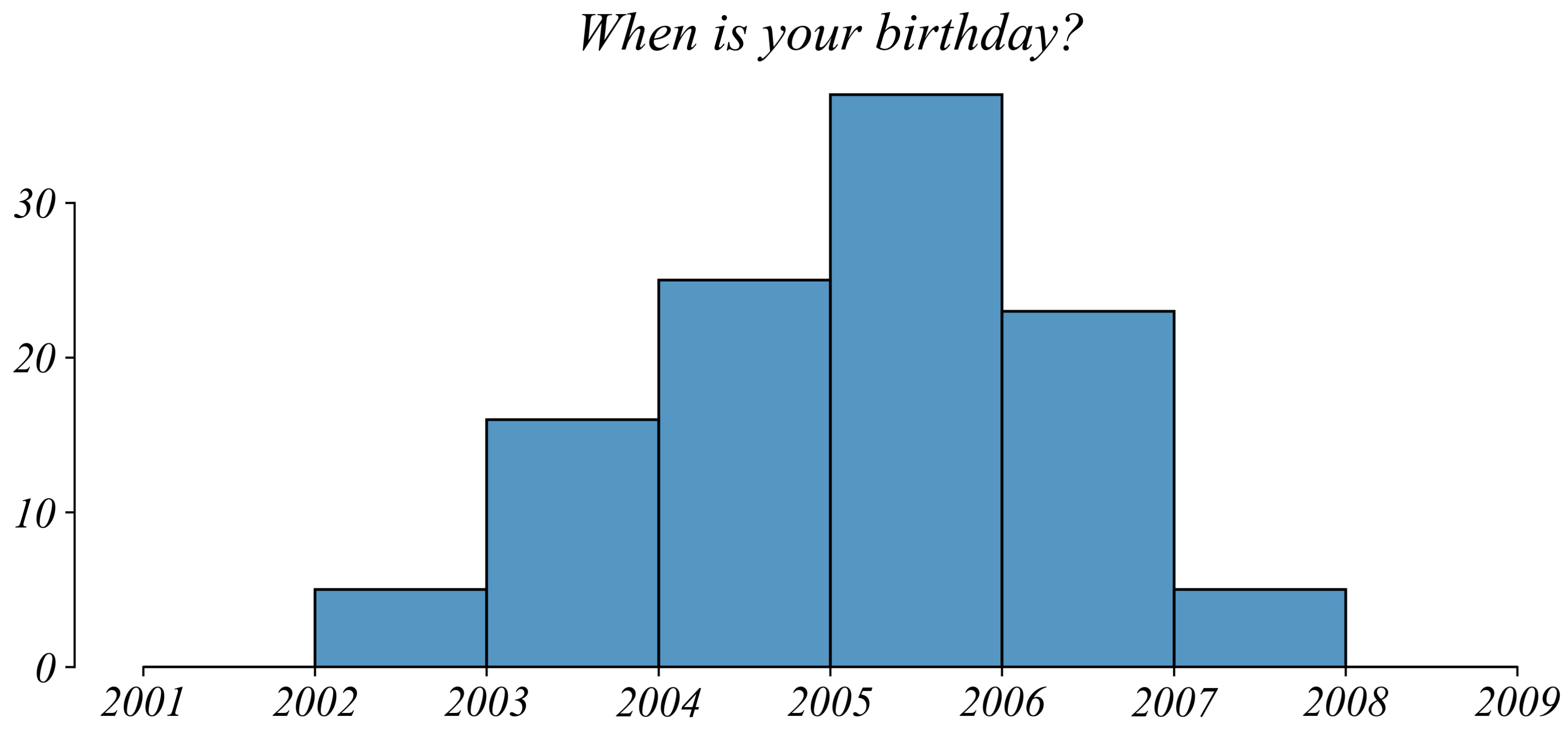
> we need to extract the year to calculate age

String Parsing

Extracting useful information from text

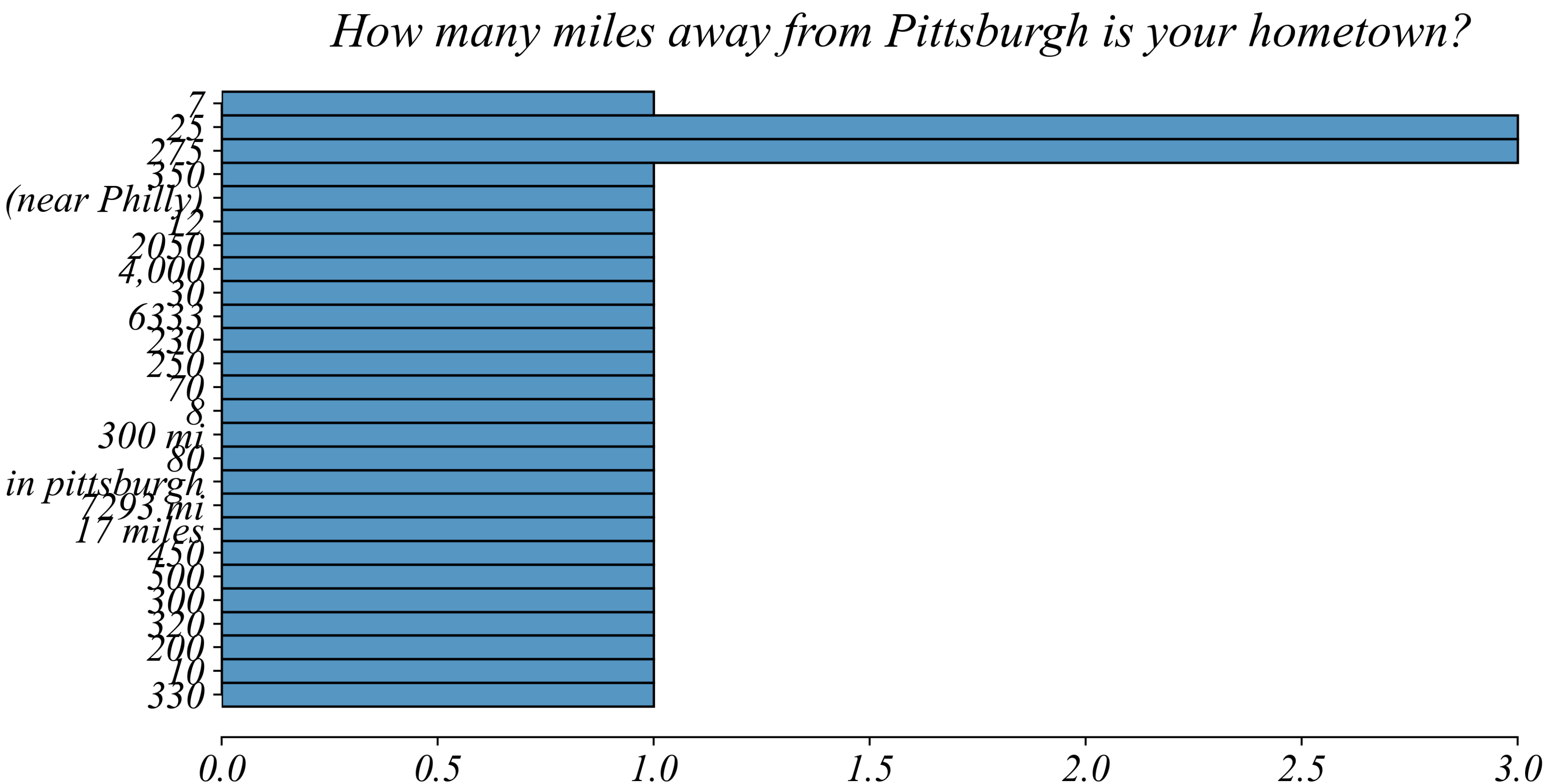
What we have: “08/15/2005”

What we need: 2005



Distance from Pittsburgh

Q. Are students who live further away older?



> lots of different formats!

Distance from Pittsburgh

Answers can be in many creative forms…

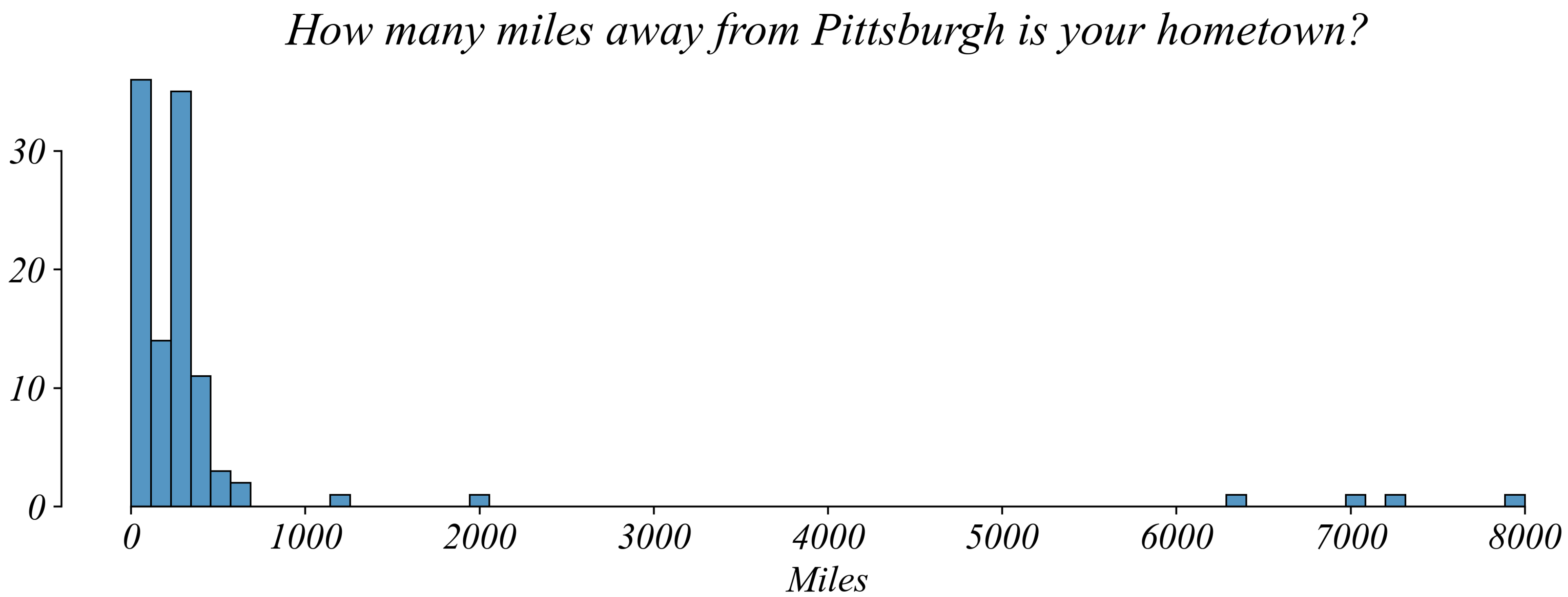
- *"0 miles"*
- *"~500"*
- *"about 1000"*
- *"2.5 hours"*
- *"very far"*

> *computers can't do math with text*

Type Conversion

Converting text to numbers

We can convert text to numbers, forcing errors to become NA.



> entries like "very far" become NA

> entries like "500" become 500.0

Missing Values

What happened to the non-numeric entries?

	new	Approximately how many miles away from Pittsburgh is your hometown?
0	400.0	400
1	16.0	16
2	300.0	300
3	300.0	300
4	400.0	400

Missing Values

What happened to the non-numeric entries?

	new	Approximately how many miles away from Pittsburgh is your hometown?
6	NaN	176 miles away
17	NaN	0 (it’s Pittsburgh)
18	NaN	400-450ish miles
22	NaN	350 miles
23	NaN	240 miles

> they all became NaN (Not Available)

> we need to decide what to do with them

Handling Missing Values

Two main approaches

After replacing problematic values, there are generally two options.

Option 1: Drop the missing values

- *Removes entire rows with NA*
- *Reduces sample size*
- *Simple and clean*

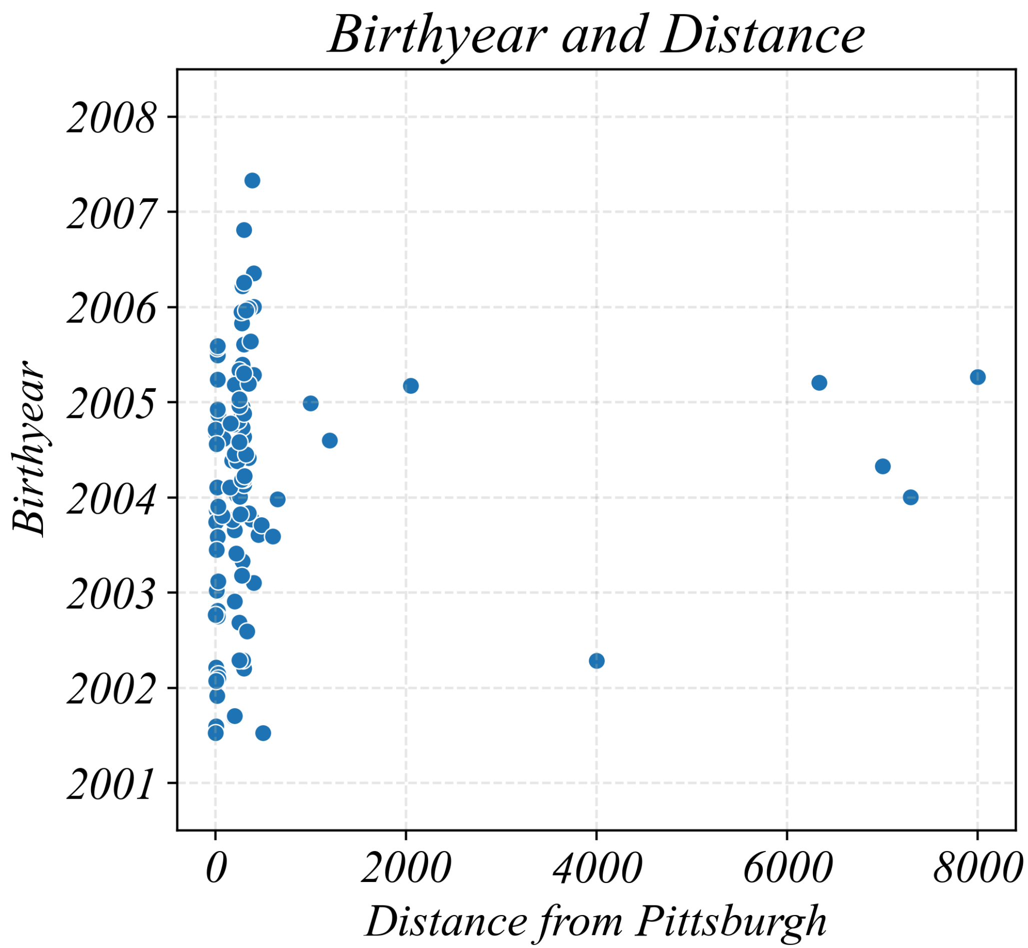
Option 2: Replace with a value

- *Fill with 0, mean, or median*
- *Keeps sample size*
- *May introduce bias*

> *for distance, dropping makes sense - we can’t guess locations*

After Cleaning

Q. Are students who live further away older?



> as expected, there does not seem to be much of a relationship

Summary

Some common data cleaning operations

- ***String Parsing****: Extract information from text*
- ***Type Conversion****: Change text to numbers*
- ***Missing Values****: Drop or replace NAs*

Exercise 2.1 | Data Cleaning

Let’s find the median birthyear and the mean hometown distance from Pittsburgh.

- ***Data***: `Fall_2025_Survey_raw.csv`

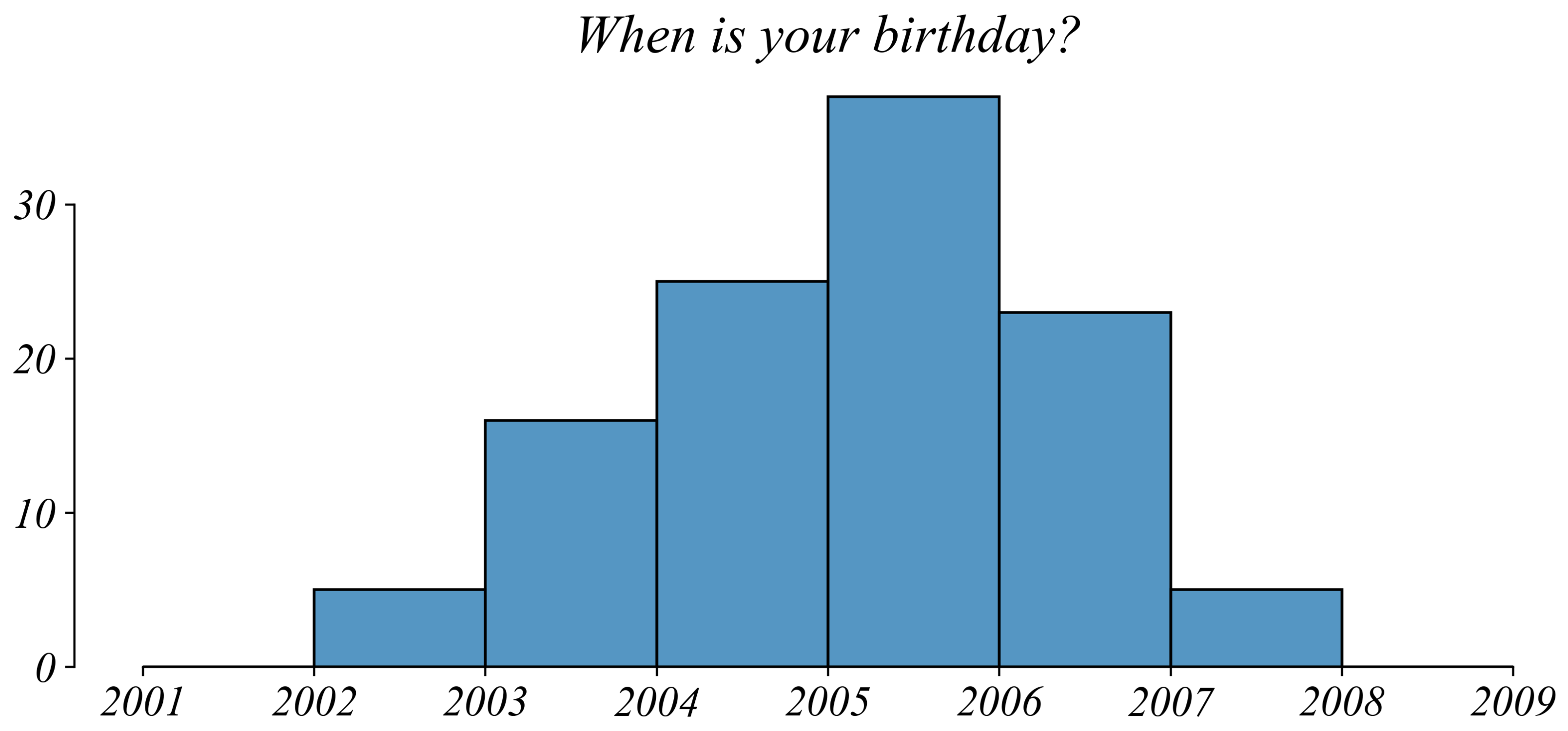
Exercise 2.1 | Birthday to Birthyear

Extract year from birthday text

```
# Convert birthday to datetime
survey['birthday'] = pd.to_datetime(survey['Approximately how many miles away from Pittsburgh
```

```
# Extract year from date
survey['birthyear'] = survey['birthday'].dt.year
```

```
sns.histplot(survey, x='birthyear')
```



Exercise 2.1 | Distance Conversion (Simple)

Convert distance text to numbers

```
# Convert to numeric, errors become NA
survey['distance'] = pd.to_numeric(survey['Approximately how many miles away from Pittsburgh i
```

```
# Check how many became NA
survey['distance'].isna().sum()
```

Exercise 2.1 | Handle Missing Values

Two approaches to NAs

Drop missing values:

```
# Remove rows where distance is NA
survey_dropped = survey.dropna(subset=['distance_clean'])
```

Replace with a value:

```
# Replace NA with median distance
median_dist = survey['distance_clean'].median()
survey['distance_filled'] = survey['distance_clean'].fillna(median_dist)
```

```
# Or replace with 0
survey['distance_zero'] = survey['distance_clean'].fillna(0)
```

Exercise 2.1 | Distance Conversion (Replace)

Convert distance text to numbers

```python
# Replace non-numeric
replacements = {
    '400-450ish miles ': 400,
    'live in pittsburgh': 0,
    '176 miles away': 176,
    '0 (it’s Pittsburgh)': 0,
    '350 miles': 350,
    '240 miles': 240,
    '388 miles': 388,
    '17 miles': 17,
    '300 miles': 300,
    '7293 mi': 7293,
    '4 miles ': 4,
    '27 miles': 27,
    '255 (near Philly)': 255,
    '4,000': 4000,
    '650 miles': 650,
    '250 miles': 250,
    '318 mi': 318,
    '300 mi': 300,
    '1000+': 1000,
    '305 miles': 305
```

```python
# Check how many became NA
survey['distance'].isna().sum()
```

Exercise 2.1 | Scatterplot

Check that it worked

```
# Create scatterplot
sns.scatterplot(data=survey_dropped, x='distance_clean', y='age')
```