Clinical Data Wrangling

Data Exploration with Burro

Aaron S Coyner, PhD

HIP 523

Acknowledgements

- Slides
  - Adapted from the Clinical Data Wrangling Workshop
  - Nicole Weiskopf, PhD
  - Ted Laderas, PhD
- Data
  - Adapted from the synthetic patient cohort used in BMI 569: Data Analytics

Learning Objectives

- **Understand** the purpose of Exploratory Data Analysis (EDA)
- **Learn** how to perform EDA using `burro`
- **Answer** questions about associations between variables

Overall Goal

- **Predict** 30-day hospital readmissions from our patients
  - **Explore** potential variables in the data to include in our model
    - **Understand** what each variable means
    - **Understand** interactions between variables
  - **Output** a list of potential variables to include in our model

Exploratory Data Analysis

What is it?

- Pioneered by John Turkey
- Detective work on your data
- An **attitude** toward data, not just techniques
- **“Find patterns, reveal structure, and make tentative model assessments.”**
  — John Behrens, *Principles and Procedures of Exploratory Data Analysis* (1997)

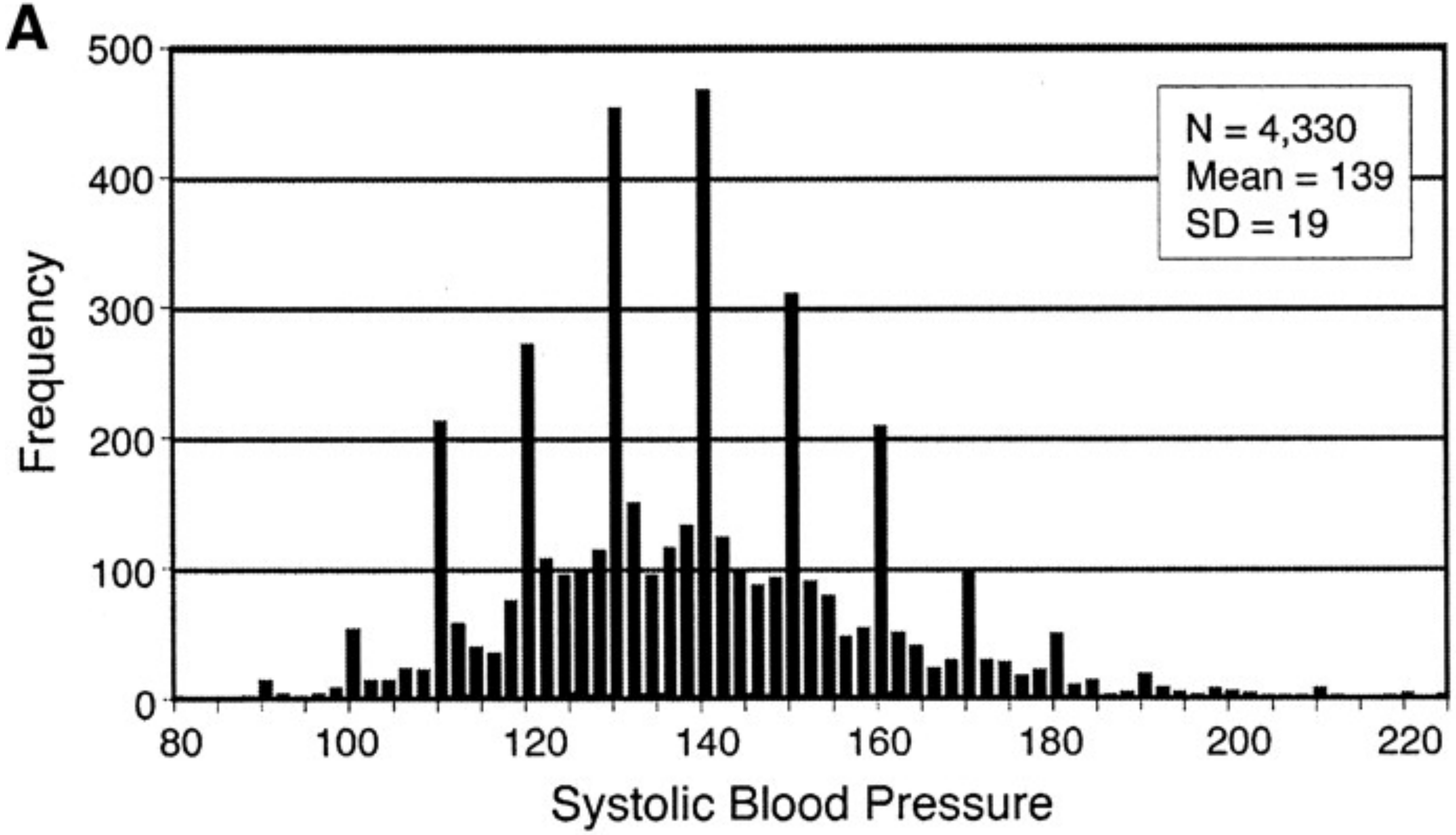
Exploratory Data Analysis

A Quote to Remember

- “Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone.”
  — John Tukey, *Exploratory Data Analysis* (1977)

Exploratory Data Analysis

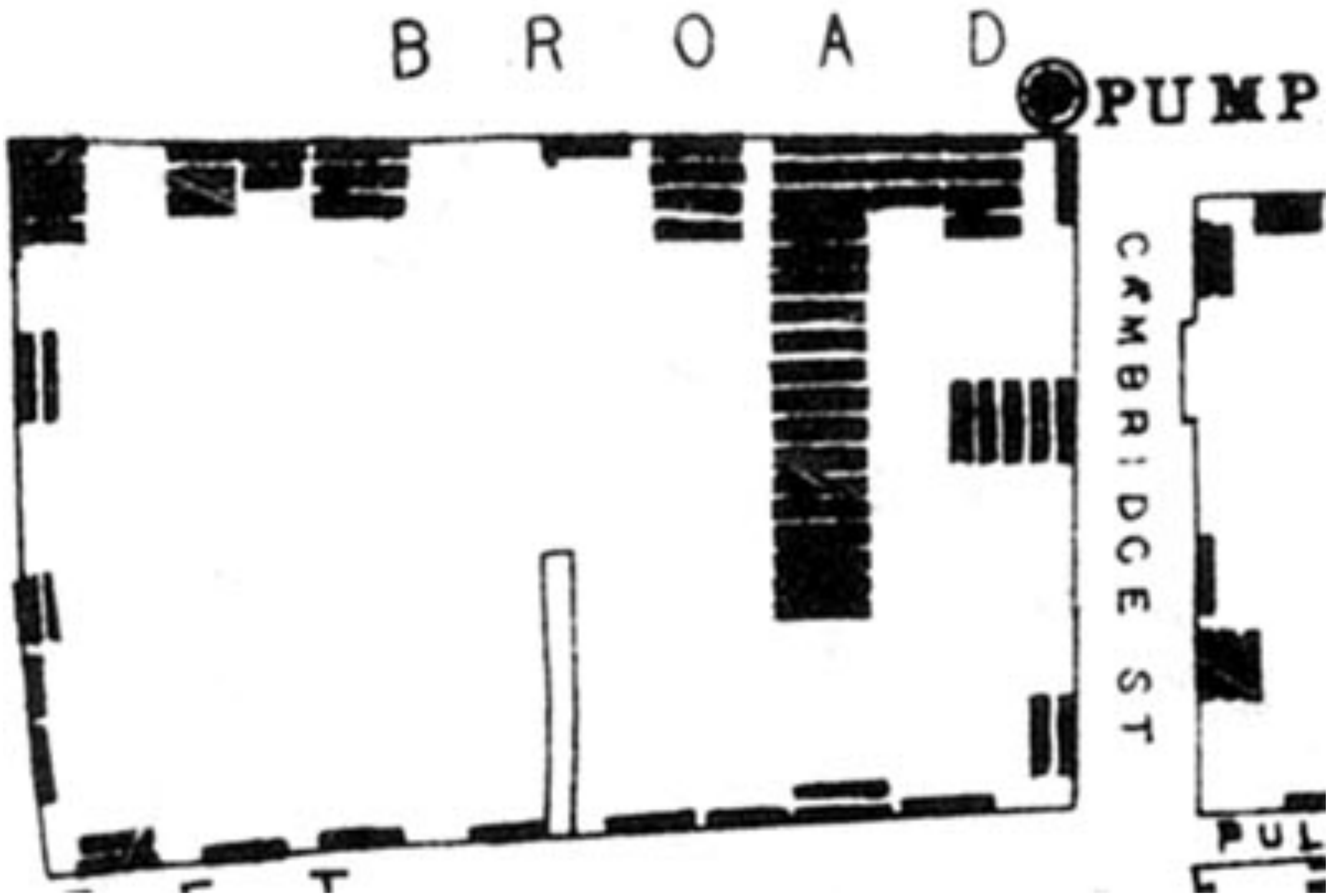
Why should we explore our data?



Need to be aware of issues in the data!

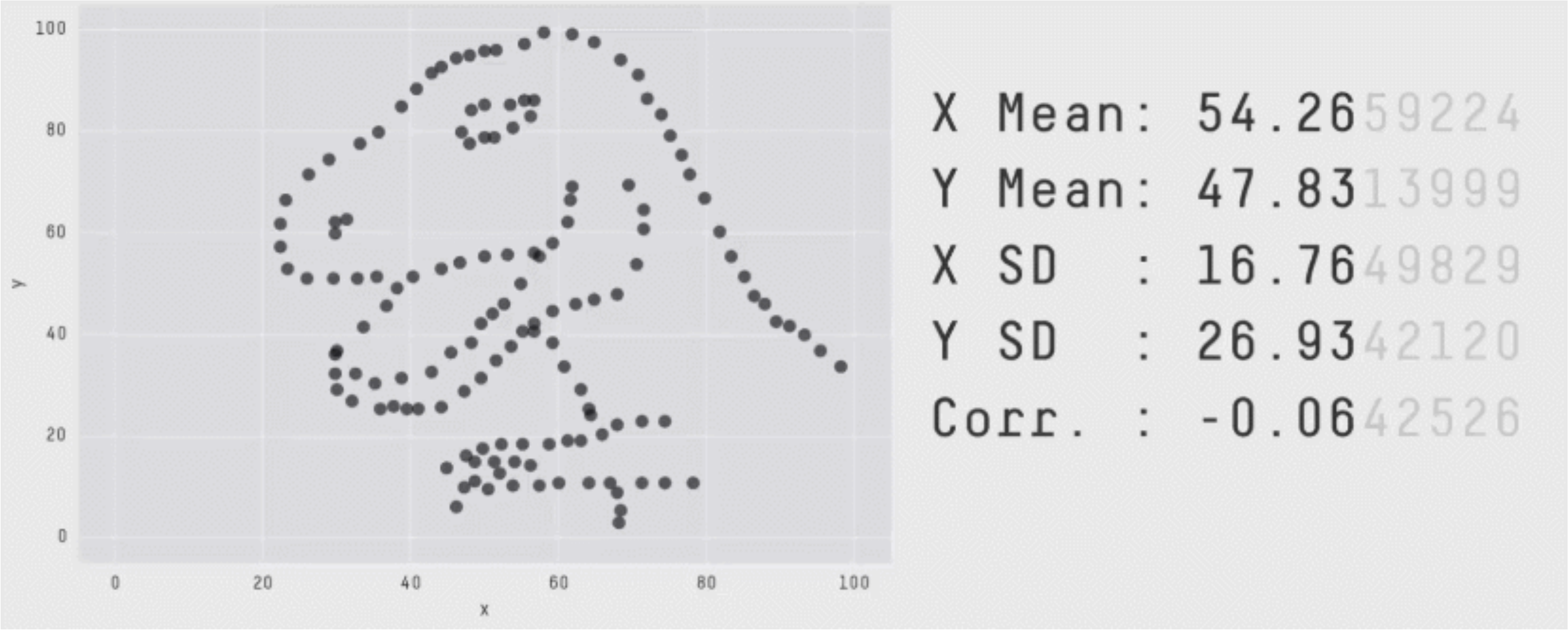
Exploratory Data Analysis

Why should we visualize our data?



Exploratory Data Analysis

BEWARE: Datasaurus Dozen!



12 datasets. Same mean. Same standard deviation. Both dimensions.

Visualization

Look first

- Visualization is a **gateway**
- **Understand** the issues
- Not going to focus on modeling and coding
    - **Build** your foundations and intuitions about your data
    - *Then* we can start getting technical

Burro

A Package for Data Exploration

- Created by Ted Laderas
- Useful for examining issues in your datasets
    - Missing data
    - Associations
    - Correlations
- If you are interested, it is freely-available here: http://laderast.github.io/burro

Workflow

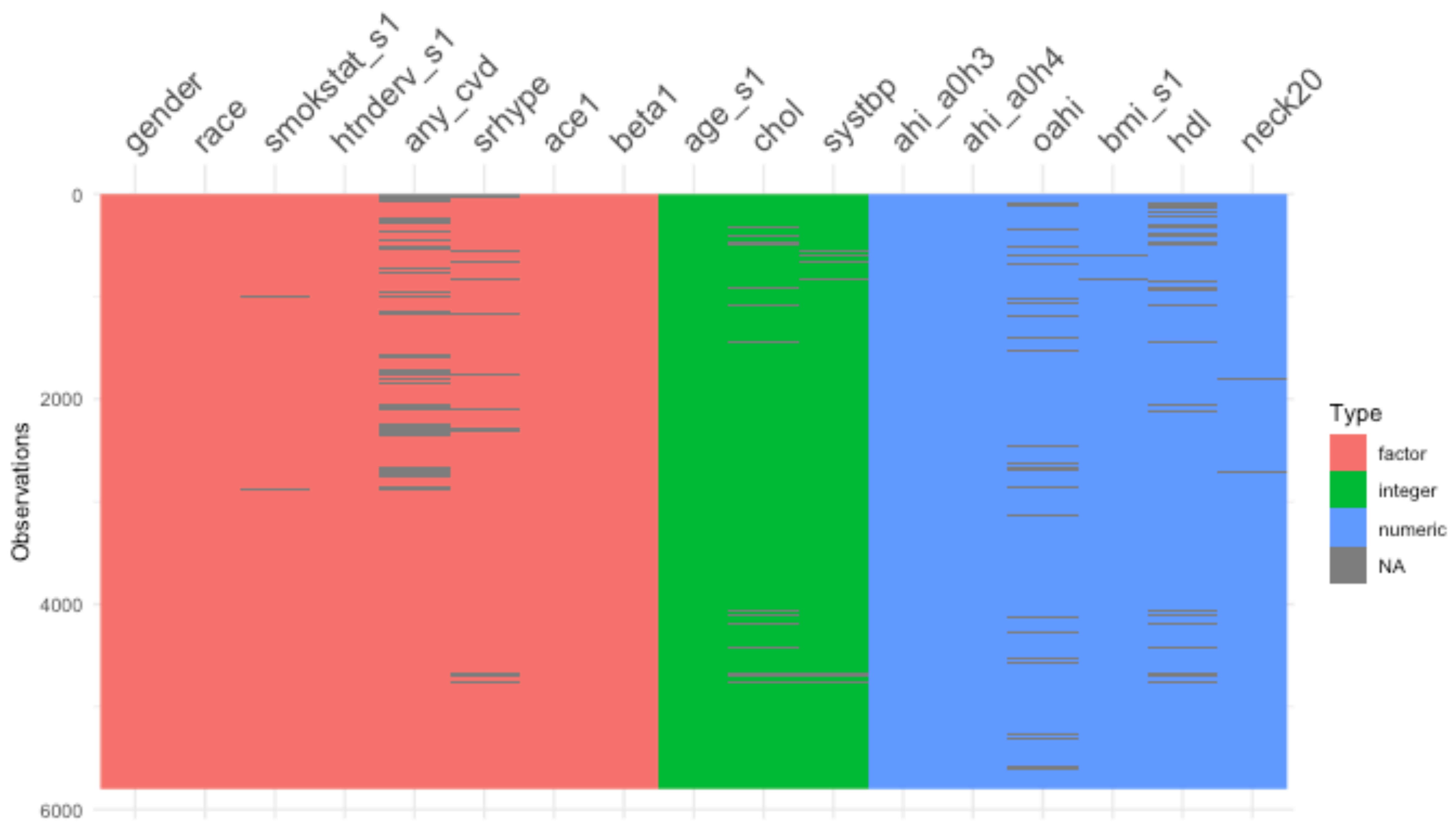
Selecting Variables for Modeling

- Ultimately, need to make decisions about which variables we think may be useful for predicted 30-day readmissions
- **Missingness**
    - Are there too many missing cases in our variable?
- **Usefulness**
    - Is there a correlation between the variable and our outcome?
- **Association**
    - How associated is our variable with other variables in the model? Should we choose on or the other?
- **Clinical**/domain-specific considerations
    - How were the data collected and does that affect our measurement?

Burro

https://bit.ly/hip_dw

The Overview Panel



HIP523 Data | Data Overview | Categorical Data | Continuous Data

Visual Summary | Tabular Summary | Data Dictionary

Data summary

Name	my_data_table
Number of rows	5603
Number of columns	8
_______________________	
Column type frequency:	
character	1
factor	5
numeric	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
patientid	0	1	1	5	0	5603	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
readmit30	1635	0.71	FALSE	2	0: 3411, 1: 557
diabetes_complications	755	0.87	FALSE	2	0: 4767, 1: 81
myocardial_complications	951	0.83	FALSE	2	0: 3474, 1: 1178
outcome	0	1.00	FALSE	3	Dis: 2787, SNF: 2238, Reh: 578
admit_source	0	1.00	FALSE	4	Eme: 2652, Cli: 1463, Tra: 1177, SNF: 311

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	755	0.87	35.18	6.00	21.29	30.09	35.13	40.04	53.9	
length_of_stay	0	1.00	7.97	10.41	2.00	3.00	5.00	9.00	298.0	

HIP523 Data | Data Overview | Categorical Data | Continuous Data

Visual Summary | Tabular Summary | Data Dictionary

Show 50 entries

Search:

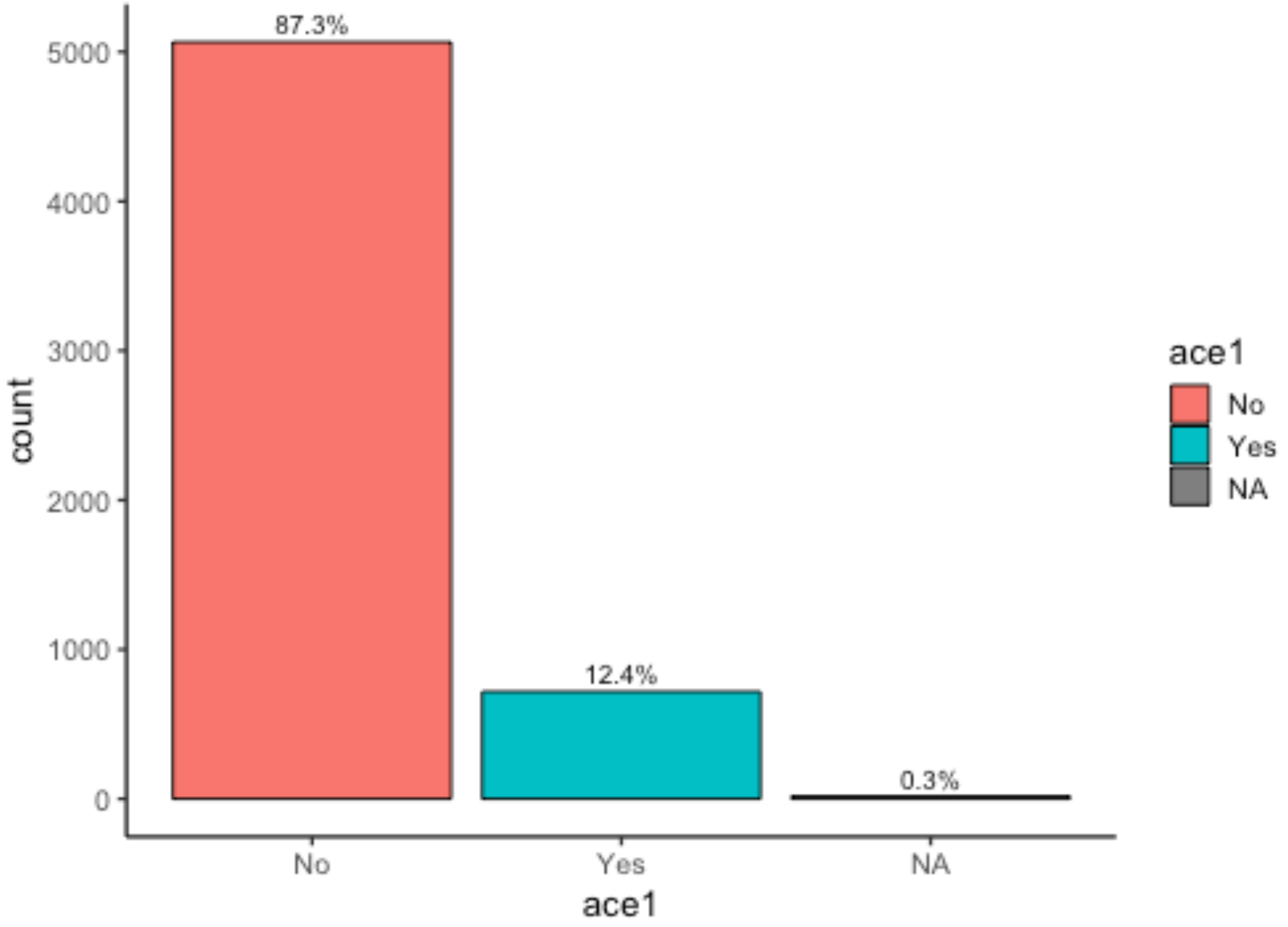
	ï..variable.name	description
1	age	Age of patient in years
2	patientid	Numeric ID
3	length_of_stay	length of stay in the hospital for previous admission in days
4	readmit30	1/0 value of whether the patient was readmitted to the hospital within 30 days
5	diabetes_complications	Whether the patient has complications related to diabetes. Calculated by using the Charleson Comorbidity Index on ICD9 codes
6	myocardial_complications	Whether the patient has complications related to myocardial infarctions. Calculated by using the Charleson Comorbidity Index on ICD9 codes
7	outcome	Where the patient ended up after admission
8	admit_source	Department where the patient was admitted

Questions

From the Overview Panel

- How big is the dataset?
- How many categorical variables (factors) are there?
- How many missing `readmit30` cases (coded as `NA`) are there?
- What is the mean age of the dataset?
  - Is it what you would expect?
- Link to `Burro`: https://bit.ly/hip_dw
- Link to the data: https://bit.ly/hip_sheet

The
Category Panel



HIP523 Data | Data Overview | Categorical Data | Continuous Data

Single Variable | Outcome View | Tabular

Select Categorical Variable

admit_source


```
readmit_data %>%
  mutate(gr = 1) %>%
  ggplot(aes_string(x = admit_source, fill = admit_source)) +
  geom_bar(aes(y = ..count..), color = "black") +
  viridis::scale_fill_viridis(discrete = TRUE, option = "magma") +
```

HIP523 Data | Data Overview | Categorical Data | Continuous Data

Single Variable | Outcome View | Tabular

Select Categorical Variable

diabetes_complications


```
readmit_data %>%
  mutate(gr = 1) %>%
  ggplot(aes_string(x = diabetes_complications, fill = diabetes_complications)) +
  geom_bar(aes(y = ..count..), color = "black") +
  viridis::scale_fill_viridis(discrete = TRUE, option = "magma") +
```

HIP523 Data | Data Overview | Categorical Data | Continuous Data

Single Variable | Outcome View | Tabular

Select Variable

diabetes_complications

Select Outcome Variable

readmit30

HIP523 Data | Data Overview | Categorical Data | Continuous Data

Single Variable | Outcome View | Tabular

Select Variable

readmit30

Select Outcome Variable

diabetes_complications

HIP523 Data | Data Overview | Categorical Data | Continuous Data

Single Variable | Outcome View | Tabular

Select Crosstab Variable (x)

readmit30

Select Crosstab Variable (y)

diabetes_complications

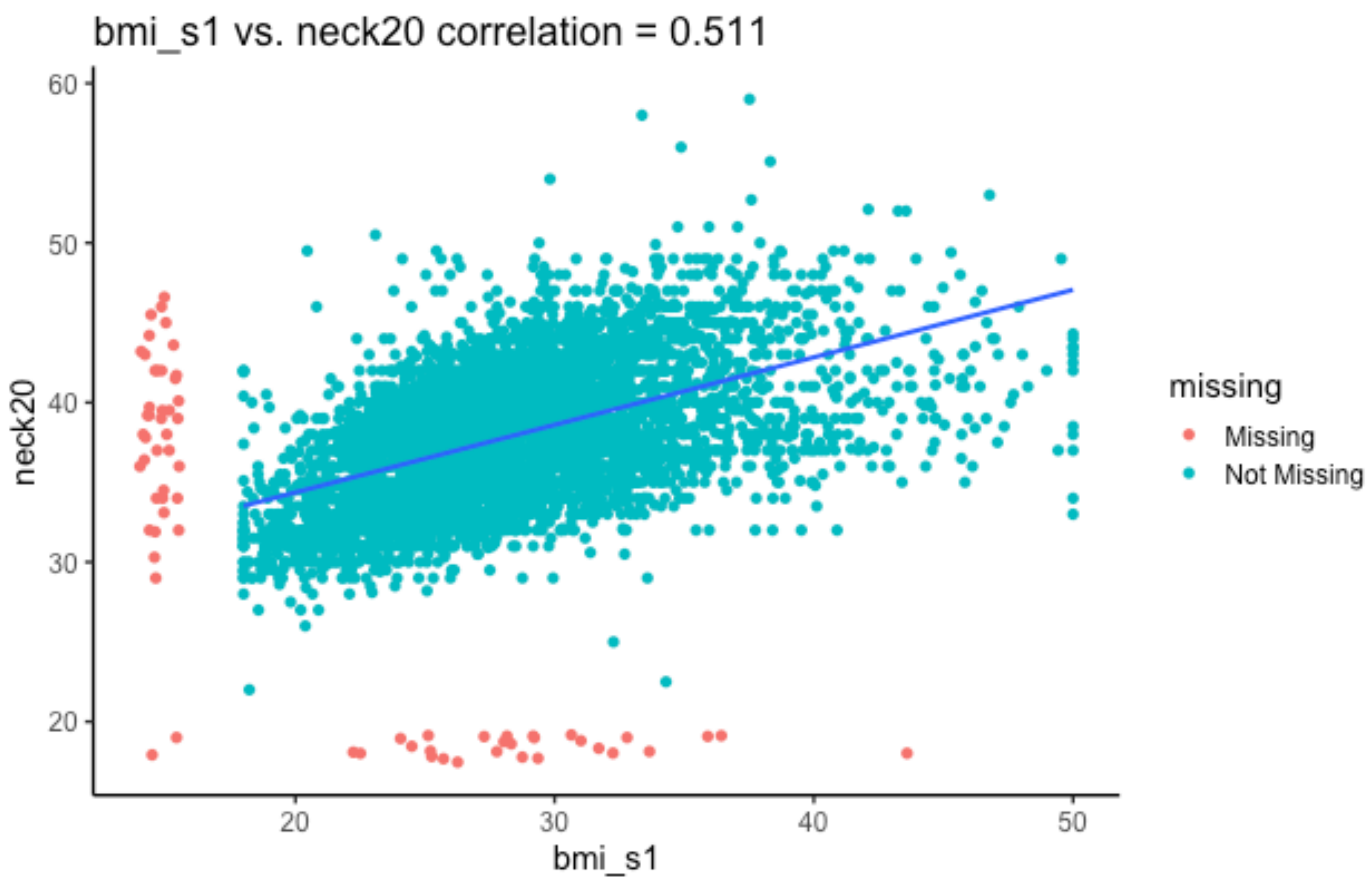

```
readmit_data %>%
  data.frame() %>%
  ggplot(aes_string(y = readmit30, x = diabetes_complications)) +
  geom_count() +
  theme(axis.text.x = element_text(angle = 90))
```

Questions

From the Category Panel

- How many categories are there for `outcome`?
- Are the proportions of `readmit30` balanced across `admit_source`?
- Are older people more likely to have `myocardial_complications`?
- Are the proportions of missing data for `readmit30` balanced across `outcome` categories?
- Link to `Burro`: https://bit.ly/hip_dw
- Link to the data: https://bit.ly/hip_sheet

The Continuous Panel



HIP523 Data | Data Overview | Categorical Data | Continuous Data

Histogram Explorer | Boxplot Explorer | Correlation Explorer

Select Numeric Variable

length_of_stay

Number of bins:

1 — 30 — 50

HIP523 Data | Data Overview | Categorical Data | Continuous Data

Histogram Explorer | Boxplot Explorer | Correlation Explorer

Select Numeric Variable

age

Select Category Variable

myocardial_complications

HIP523 Data | Data Overview | Categorical Data | Continuous Data

Histogram Explorer | Boxplot Explorer | Correlation Explorer

Select Y Variable

age

Select Y Variable

length_of_stay

Questions

From the Continuous Panel

- What is the distribution of `age` in our patients?
- Is `age` evenly distributed across `readmit30`? If not, how is it distributed?
- Are `age` and `length_of_stay` correlated? Are you surprised?
- Should we include both `age` and `myocardial_complications` in our model?
- Link to `Burro`: https://bit.ly/hip_dw
- Link to the data: https://bit.ly/hip_sheet

Selecting Predictors

For Next Time

- **Missingness**
  - Which variables have missing data?
  - Is the missingness correlated for any two variables?
  - How could we deal with this?
- **Associations and Correlations**
  - Including interacting variables as predictors can affect their predictive power
  - For example, `age` and `myocardial_complications`
- **Select your predictor covariates of** `readmit30`
  - **We're going build predictive models of the dataset using these predictors**

Wrap Up

- Data exploration can be fun “detective” work
- Be curious! Start with a question.
- Assess the impact of adding a covariate to a model:
  - Does the distribution look like other populations?
  - Is it associated with your outcome?
  - Is it associated with other variables?
  - Is the data missing in a suspicious way?