

Introduction to Data Science

Numerical vs. Categorical Data

Joanna Bieri DATA101

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 [Click Here for Joanna's Schedule](#)

Announcements

Please come to office hours to get help!

Remember to come to lab! Friday 10:50-12pm Duke 200.

Data Types

Data Set Types:

- Structured Data
- Unstructured Data
- Semi-structured Data

Number of Variables:

- Univariate - single variable
- Bivariate - two variables
- Multivariate - many variables

Types of Variables:

- **Numerical variables** - values that are numbers
- **Categorical** - values that are not numbers

Types of Variables:

Numerical variables can be classified as

- **continuous** - infinite number of possible values, subdivisions (real numbers)
- **discrete** - only specific values (integers)

Categorical variables can be classified as

- **ordinal** - has some natural ordering.
- **nominal** - lacks natural ordering

Your Examples?

Loan Data Set

Loan data from Lending Club

Loan Data Variable Types

variable	type
loan_amount	numerical, continuous
interest_rate	numerical, continuous
term	numerical, discrete
grade	categorical, ordinal
state	categorical, nominal
annual_income	numerical, continuous
homeownership	categorical, nominal
debt_to_income	numerical, continuous

Visualizing Numerical Data

Numerical Data - Summary Statistics

With numerical data it is possible to quickly do summary statistics. We might want to know things like:

- **Measures of Center** - mean, median
- **Measures of Spread** - range (max-min), standard deviation, inter-quartile range ($IQR = Q3 - Q1$)
- **The Shape of the Data** - right-skewed, left-skewed, symmetric, unimodal, bimodal, etc.
- **Observations of Outliers** - use IQR or or visually

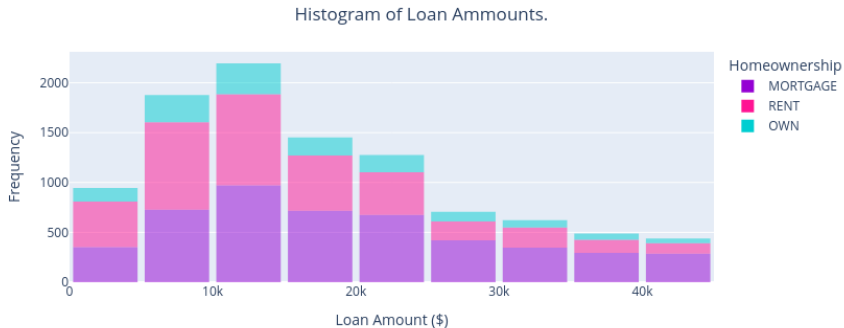
Numerical Data - Summary Statistics

	loan_amount	interest_rate	term	annual_income	debt_ratio
count	10000.000000	10000.000000	10000.000000	1.000000e+04	99
mean	16361.922500	12.427524	43.272000	7.922215e+04	19
std	10301.956759	5.001105	11.029877	6.473429e+04	15
min	1000.000000	5.310000	36.000000	0.000000e+00	0.0
25%	8000.000000	9.430000	36.000000	4.500000e+04	11
50%	14500.000000	11.980000	36.000000	6.500000e+04	17
75%	24000.000000	15.050000	60.000000	9.500000e+04	25
max	40000.000000	30.940000	60.000000	2.300000e+06	46

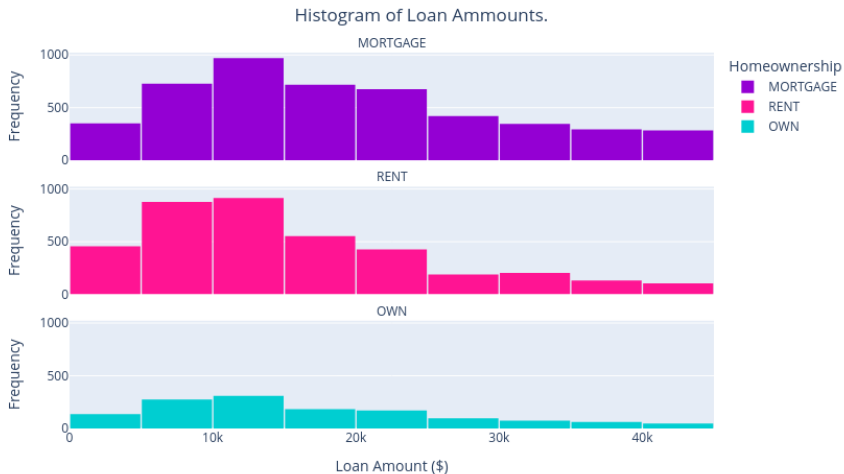
Numerical Data - Visualizations

- **Histograms**
 - Histograms with Facets
 - Histograms with Distribution Information
- **Distributions**
 - Violin Plots - hybrid of a box plot and a kernel density plot
 - Box Plots

Histogram with Categorical Coloring



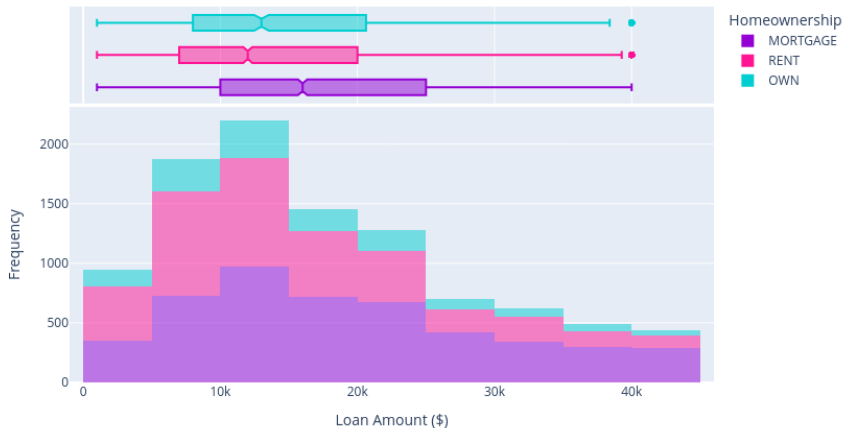
Histogram with Facets



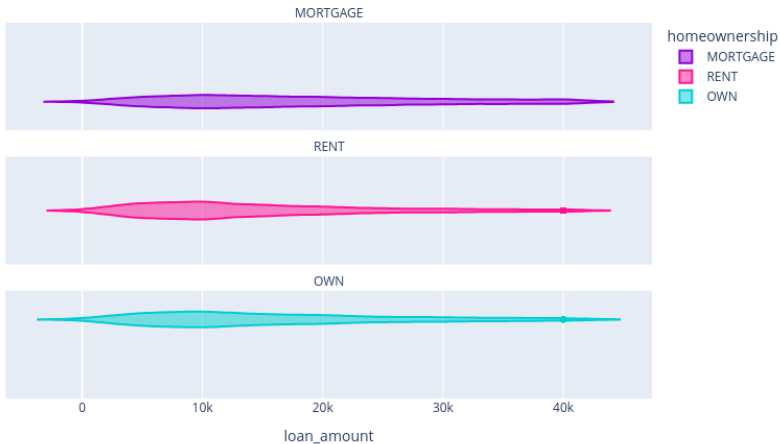
Histogram with Box Plot

Unable to display output for mime type(s): text/html

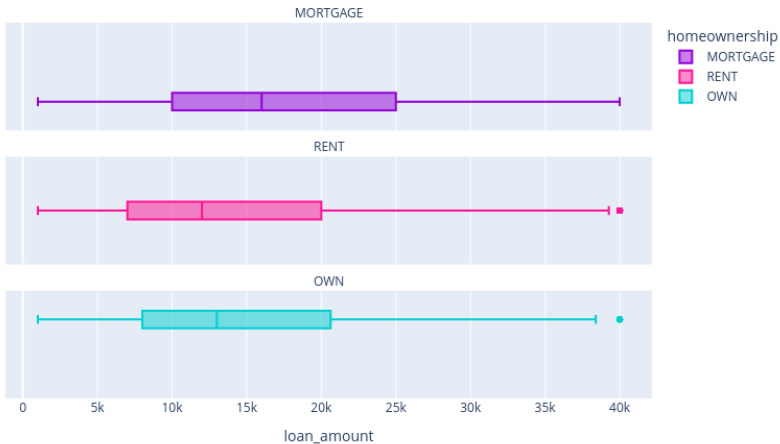
Histogram of Loan Ammounts.



Violin Plot



Box Plot (Box and Whiskers)



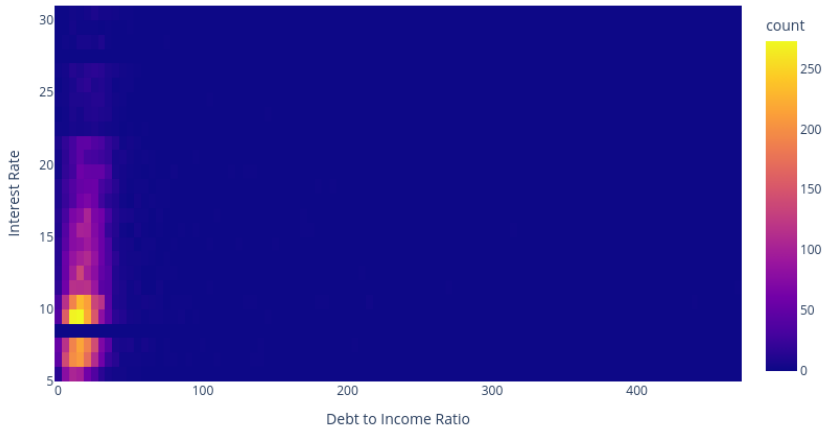
Scatter Plots

Debt to Income Ratio vs. Interest Rate



Heat Maps

Debt to Income Ratio vs. Interest Rate



Visualizing Categorical Data

Visualizing Categorical Data

- Can't do simple summary statistics.
- Often more interested in counts (`value_counts`)

```
homeownership
```

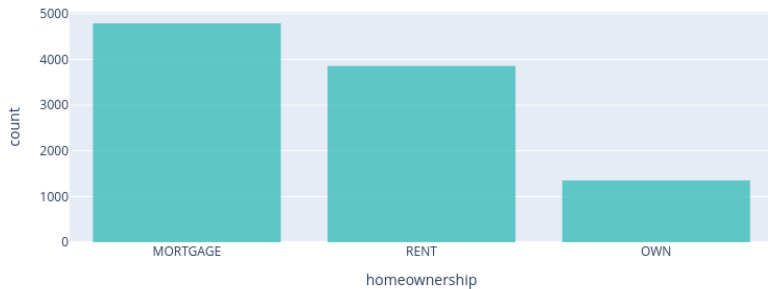
```
MORTGAGE      4789
```

```
RENT          3858
```

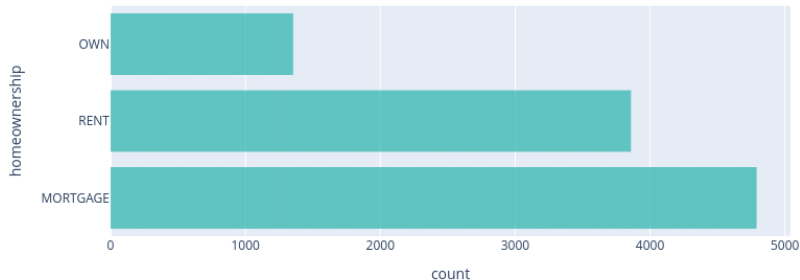
```
OWN           1353
```

```
Name: count, dtype: int64
```

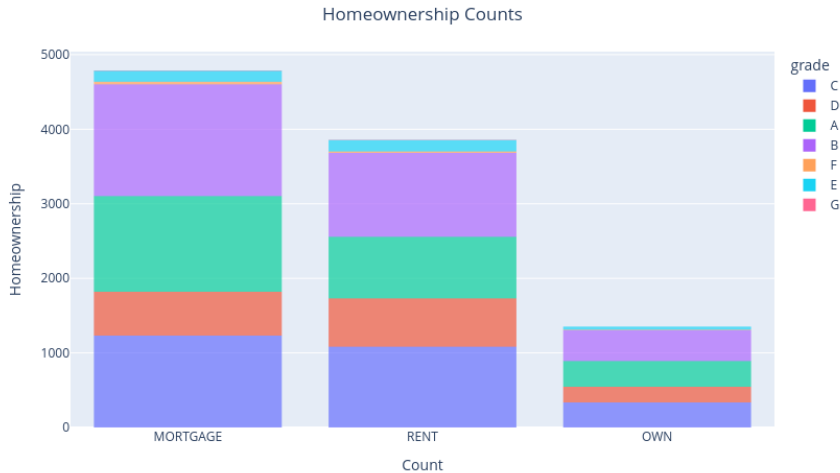
Bar Plot



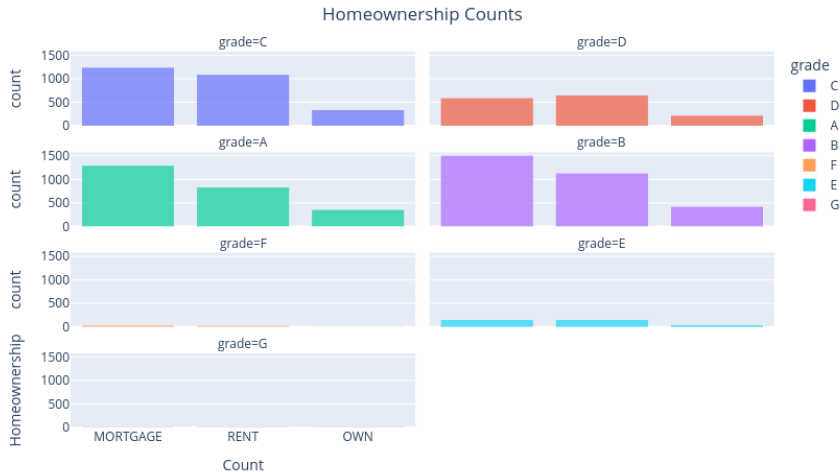
Bar Plot



Bar Plot with categorical fill



Bar Plot with Facets



Numerical and Categorical Data - Visualization

Is anyone willing to show off their work
Questions?