STAT 301: Project – Part I

Sharing your Project Idea

This handout describes the first part of the project, where you will share your project idea.

https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMSDataDict15.txt

Answer the following questions:

1. Who are your group members? Charles Yang, Megan Criley, Jackson Lyons, Steven Verdi

2. Which data set would you like to use? Iowa House

3. What is your intended population? What is your actual population?

Intended: All single family housing units in Iowa

Actual: All single family housing units in Iowa in 2015

4. Describe your sample. How large is it? How were the units selected?

The units were randomly selected from Iowa Housing data of all housing units that were single-family homes.

500 rows

5. Do you have any concerns about the data or questions that could be asked about the data? Note: you should have at least one concern.

Some data is missing. This is a concern.

Variable names are not exactly clear.

6. Choose one quantitative response variable. What is the name of the variable in the dataset (e.g., RNTP)? Describe the variable (e.g., How was it measured? Is it discrete or continuous?).

a. Notes: Your response variable needs to have a range larger than 10 (e.g., The least number of bedrooms in the Air BnB dataset is 0 and most is 8 giving a range of 8. Therefore, number of bedrooms cannot be used as the response variable. You could use it as an explanatory variable.).

b. Make sure the variable is actually quantitative. For example, if Income is listed in groups in the dataset, it is no longer quantitative.

Predicting RNTP (Monthly Rent)

7. Choose three explanatory variables. At least one of the variables should be categorical and at least one should be quantitative. What are the name of the variable in the dataset (e.g., RNTP)? Describe each of the variables (e.g., How was it measured? Is it categorcal or quantitative? Is it discrete, continuous, nominal or ordinal?). For the categorical variable, which category will you use as the reference group (i.e., comparison group or baseline group)?

ACR (Lot Size) Ordinal

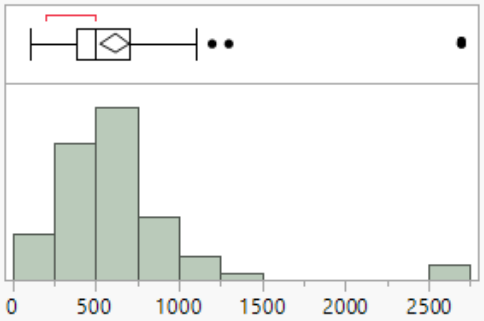
HINCP (Household income (past 12 months)) Continuous

VEH (Vehicles (1 ton or less) available) Ordinal

8. Create histograms or bar graphs for each explanatory and response variable. You do not have to interpret the graphs in this part of the project.

Distributions

RNTP



Quantiles

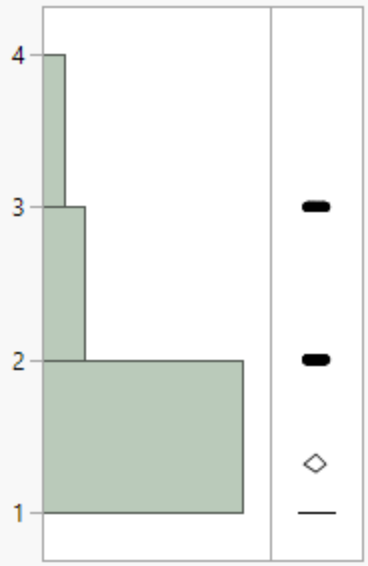
100.0%	maximum	2700
99.5%		2700
97.5%		2700
90.0%		942
75.0%	quartile	700
50.0%	median	500
25.0%	quartile	380
10.0%		250
2.5%		130
0.5%		100
0.0%	minimum	100

Summary Statistics

Mean	614.46602
Std Dev	436.01747
Std Err Mean	42.962078
Upper 95% Mean	699.6811
Lower 95% Mean	529.25094
N	103

Distributions

ACR



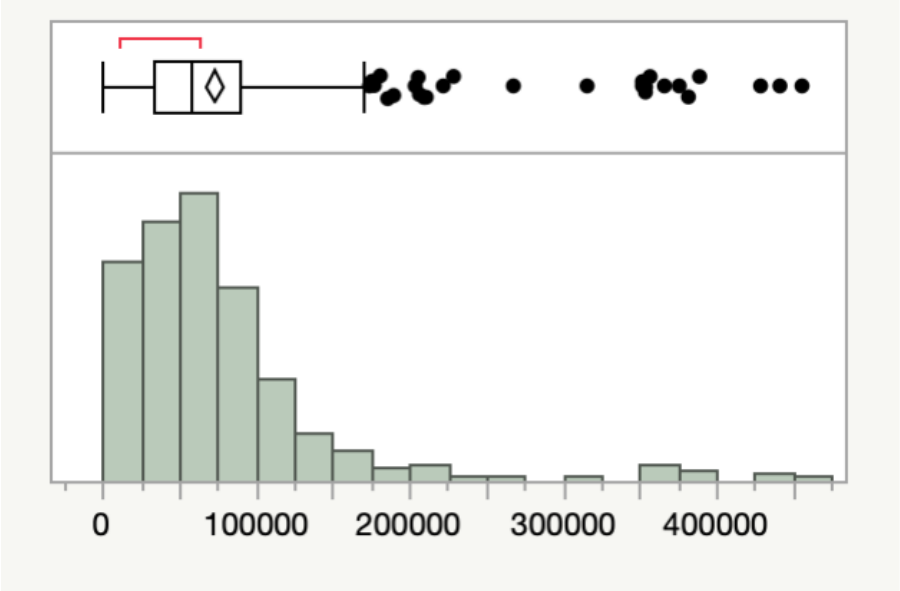
Quantiles

100.0%	maximum	3
99.5%		3
97.5%		3
90.0%		2
75.0%	quartile	1
50.0%	median	1
25.0%	quartile	1
10.0%		1
2.5%		1
0.5%		1
0.0%	minimum	1

Summary Statistics

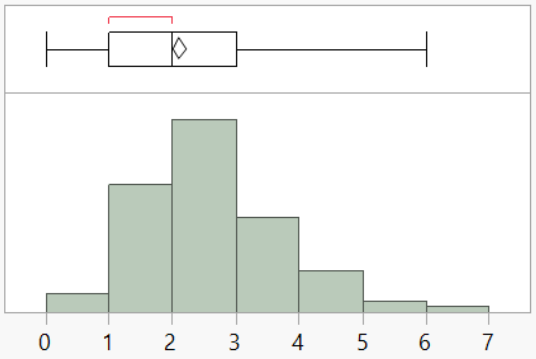
Mean	1.3214286
Std Dev	0.6161769
Std Err Mean	0.0291116
Upper 95% Mean	1.3786412
Lower 95% Mean	1.2642159
N	448

HINCP



Distributions

VEH



Quantiles

100.0%	maximum	6
99.5%		6
97.5%		5
90.0%		4
75.0%	quartile	3
50.0%	median	2
25.0%	quartile	1
10.0%		1
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	2.1136842
Std Dev	1.106045
Std Err Mean	0.0507488
Upper 95% Mean	2.2134047
Lower 95% Mean	2.0139637
N	475

Some categorical variables have many categories. If you plan to use a variable with more than 5 categories, consider collapsing similar categories, especially categories with few observations.

One person in the group upload your group's answers to these questions on Canvas by **12 March** at midnight. Your answers should be uploaded as a Microsoft Word or a pdf file.