STATS1: Second Group Project

Andres Lopez, JonJeng Thao, Merrick Reitinger

11/13/2023

By completing this group project, you will be able to:

- Generate a question that can be answered with provided data.
- Perform EDA for two (binary) categorical variables and one quantitative variable.
- Perform inference for one proportion, two proportions, and a single mean
- Write accurate inferential conclusions and interpretations for each of the three procedures

Useful resources for this homework:

- Book chapters 16, 17, and 19.
- In class activities from Unit 4 and Unit 5

Introduction

As in Part 1 of the project we will be using the Cook County Assessor's Office Dataset. This dataset contains a variety of information about the sale price of every property in Cook County (Chicago Area) and can be used to makes predictive models to help determine the assessed value of every property in Cook County.

You will again be working with a subset of the CCAO dataset. In particular we will be using the following variables. Please note that some of the numeric/quantitative variables have been converted to categorical variables.

Variables	Description
`pin`	Unique identifier for each home
`township_code`	Township designation
`num_bedrooms`	Number of bedrooms
`num_fireplaces`	Number of fireplaces
`num_full_baths`	Number of full baths
`num_half_baths`	Number of half baths
`construction quality`	Construction quality, Average, Deluxe, or Poor
`garage_attached`	Garage attached to home? Yes/No
`basement_type`	Type of basement
`basement_finish`	Finish of the basement ("Finished" or "Unfinished)
`porch`	Porch status ("None" or "Finished Porch")

Variables	Description
central_air	Central A/C, Yes/No
central_heating	Central heating, Yes/No
roof_material	Roofing material (Shingle + Asphalt" or "Other")
year_sold	Year home was sold
neighborhood_id	Neighborhood designation
sale_price	Price of sale in dollars
sale_date	Date of sale
year_built2	Year home was built
land_sqft2	Square footage of the land parcel
build_sqft2	Square footage of the home
num_rooms2	Total number of rooms
year_sold_cat	Year home was sold, Pre-2008 or Post 2008
year_build_cat	Year home was build, Pre-1950 or Post 1950
num_full_baths_cat.	Categorical number of full baths (0,1,2, >2)
num_half_baths_cat.	Categorical number of half baths (0,1,>1)
num_fireplaces_cat.	Categorical number of fireplaces (0,1,>1)

The dataset used for each group will be available under the folder `Stats 172 F23/Project/Section...`. Please follow specific instructions from your professor about loading/accessing your own dataset.

```
# Note that the data sets have been filtered so that only houses with
#   sale prices below $500,000 are included.  You should acknowledge this
#   in your writeup.

# If you feel that filtering is appropriate for your analyses, you may do it
#   by creating a new data set like this (making sure you acknowledge your
#   filtering in the writeup):
#newdata <- olddata %>% filter(variable < some value)
```

```
library(tidyverse)
library(mosaic)
library(readr)


ccao_5 <- read_csv("~/Stats 172 F23/Project/Section D (Ziegler-Graham)/JonJeng, Merrick, Andres/ccao_ca
```

```
## Rows: 1000 Columns: 27
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (14): pin, construction_quality, garage_attached, basement_type, centra...
## dbl  (12): township_code, num_bedrooms, num_fireplaces, num_full_baths, num_...
## date  (1): sale_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Overview

You will be exploring two binary categorical variables and one quantitative variable. You are encouraged to select variables that can contribute to a broader narrative related to the Cook County Dataset. You can build on questions your identified in the first project submission. A strong conclusion should reflect connections between the variables you choose for this portion of the project.

Section A: One Proportion

1. Select one of your two binary categorical variables and identify a question that can be answered with it. Clearly state this question for a general audience and explain the variable in context of the data collection.

One of our variables is central air. A question that can be asked is "What percent of houses in Cook County, Illinois have central air?". In the context of the data collection, we want to see if houses are more likely to have central air or not have it, or if there is an equal chance of both.

2. For your variable provide appropriate summary statistic(s) and then describe those statistic(s) in context.

```
table(ccao_5$central_air)
```

```
##
##    Central A/C No Central A/C
##            551            449
```

```
551/1000
```

```
## [1] 0.551
```

These show that 55.1% of houses in the data set have central A/C, and 44.9% of them do not.

3. State the null and alternative hypothesis in symbols and in words for an appropriate statistical test.

Null hypothesis: The proportion of houses with central air will equal 0.5. $H_0 = 0.5$

Alternative hypothesis: The proportion of houses with central air will not equal 0.5 $H_A \neq 0.5$

4. Run your hypothesis test, report your test statistic and p-value, and state your statistical conclusion in context.

```
prop.test(ccao_5$central_air)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  $  [with success = Central A/C]ccao_5  [with success = Central A/C]central_air  [with success
## X-squared = 10.201, df = 1, p-value = 0.001404
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
```

```
##  0.5195339 0.5820694
## sample estimates:
##     p
## 0.551
```

The test statistic was 10.201, and the p-value was 0.001404. Because the p-value is below 0.05, we can reject the null hypothesis.

5. Interpret an appropriate 95% confidence interval in context.

We can be 95% confident that the true percentage of houses with central air is between 51.9% and 58.2%.

6. Comment on the role of assumptions/conditions for the test.

The observations are independent because one house having central air should not affect another house having central air, and there are 1,000 observations in the data set, so we can assume nearly normal distribution.

Section B: Two Proportions

1. Identify a question that can be answered with two binary categorical variables in the dataset. Clearly state this question for a general audience, identify the explanatory and response variable, and explain the two variables in context of the data collection.

Does a house being built pre of post-1950 have an impact on whether it has central air or not? The explanatory variable is year built, and the response variable is whether the house has central air. In this context the variables refer to the when the houses in the dataset were built and whether they have central air.

2. Explore and describe the relationship between the two variables with appropriate summary statistics and visualizations. Make sure to describe the results of your summary and visualization.

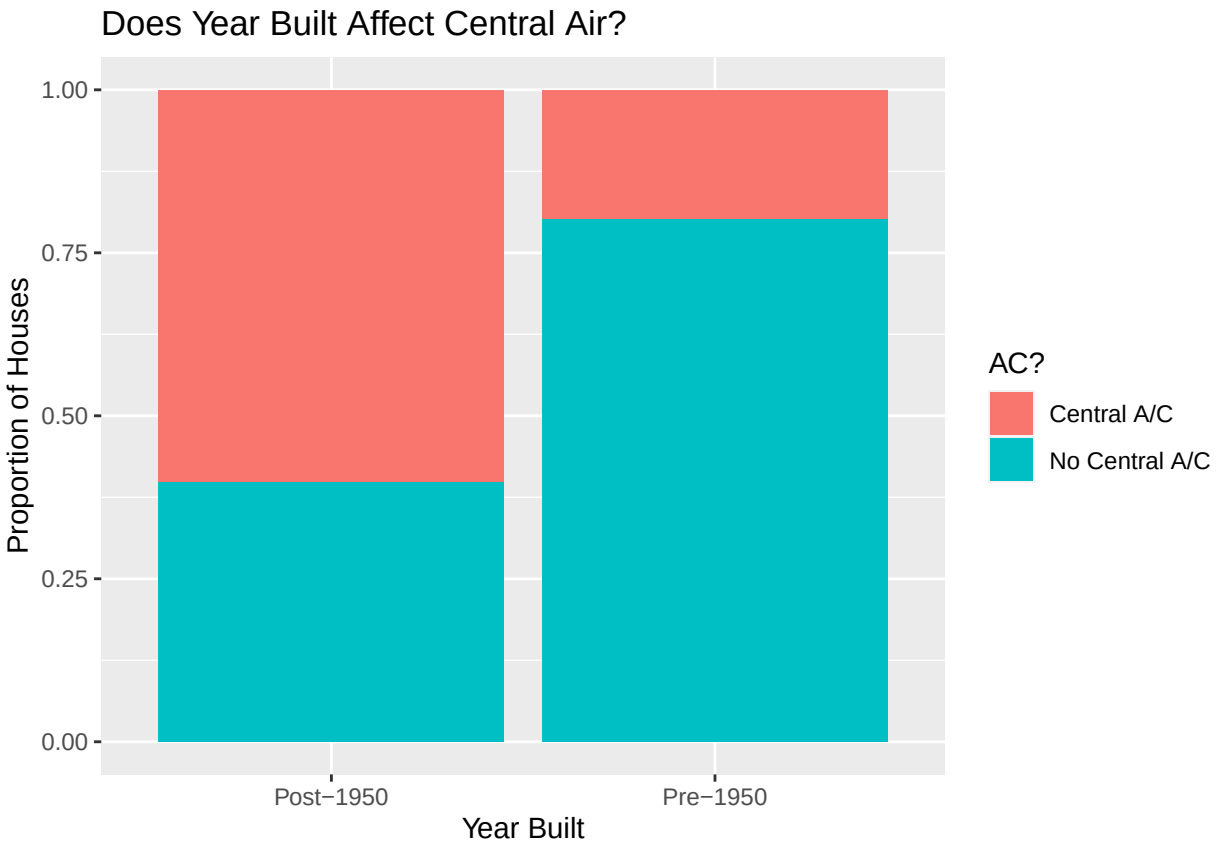
```
table(ccao_5$year_built_cat, ccao_5$central_air)
```

```
##
##             Central A/C No Central A/C
##   Post-1950         526            348
##   Pre-1950           25            101
```

```
table(ccao_5$year_built_cat, ccao_5$central_air) %>%
  proportions(margin = 1) %>%
  round(3)
```

```
##
##             Central A/C No Central A/C
##   Post-1950       0.602          0.398
##   Pre-1950        0.198          0.802
```

```
ggplot(ccao_5, aes(x = year_built_cat, fill = central_air)) +
  geom_bar(position = "fill") +
  labs(title = "Does Year Built Affect Central Air?",
       x = "Year Built",
       y = "Proportion of Houses",
       fill = "AC?")
```



These statistics and this visualization show that houses built post 1950 are much more likely to have central air than houses built pre 1950.

3. State the null and alternative hypothesis in symbols and in words for an appropriate statistical test.

Null hypothesis: The proportion of houses with central air will be the same for houses built before and after 1950.

Alternative hypothesis: The proportion of houses with central air will be different for houses built before and after 1950. $H_A \neq 0.5$

4. Run your hypothesis test, report your test statistic and p-value, and state your statistical conclusion in context.

```
prop.test(ccao_5$year_built_cat, ccao_5$central_air)
```

```
##
```

```
##  1-sample proportions test with continuity correction
##
## data:  $  [with success = Post-1950]ccao_5  [with success = Post-1950]year_built_cat  [with success =
## X-squared = 558.01, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.8514578 0.8936105
## sample estimates:
##     p
## 0.874
```

Test Statistic = 558.01 p-value < 2.2e-16

Since our P value is almost 0, well below 0.05, we can reject the null hypothesis. Our test statistic resulted in this low P value that provides strong evidence towards the association of central air conditioning and the year a house was built.

5. Interpret an appropriate 95% confidence interval in context.

We can be 95% confident that the true proportion of homes with central air conditioning built after 1950 is to be between 0.8514578 and 0.8936105. Since the interval does not include 0.5, which is the null proportion, we can reject the null hypothesis.

6. Comment on the role of assumptions/conditions for the test.

Our test assumes that the values of our data are independent and have been randomly sampled. Knowing the category of one home in terms of whether it has central air or the year it was built does not provide information about the categories of another home. We can

Section C: One Mean

1. Identify a question of a single numeric variable you are interested in estimating from the dataset. Clearly state this question for a general audience and explain the variable in context of the data collection.

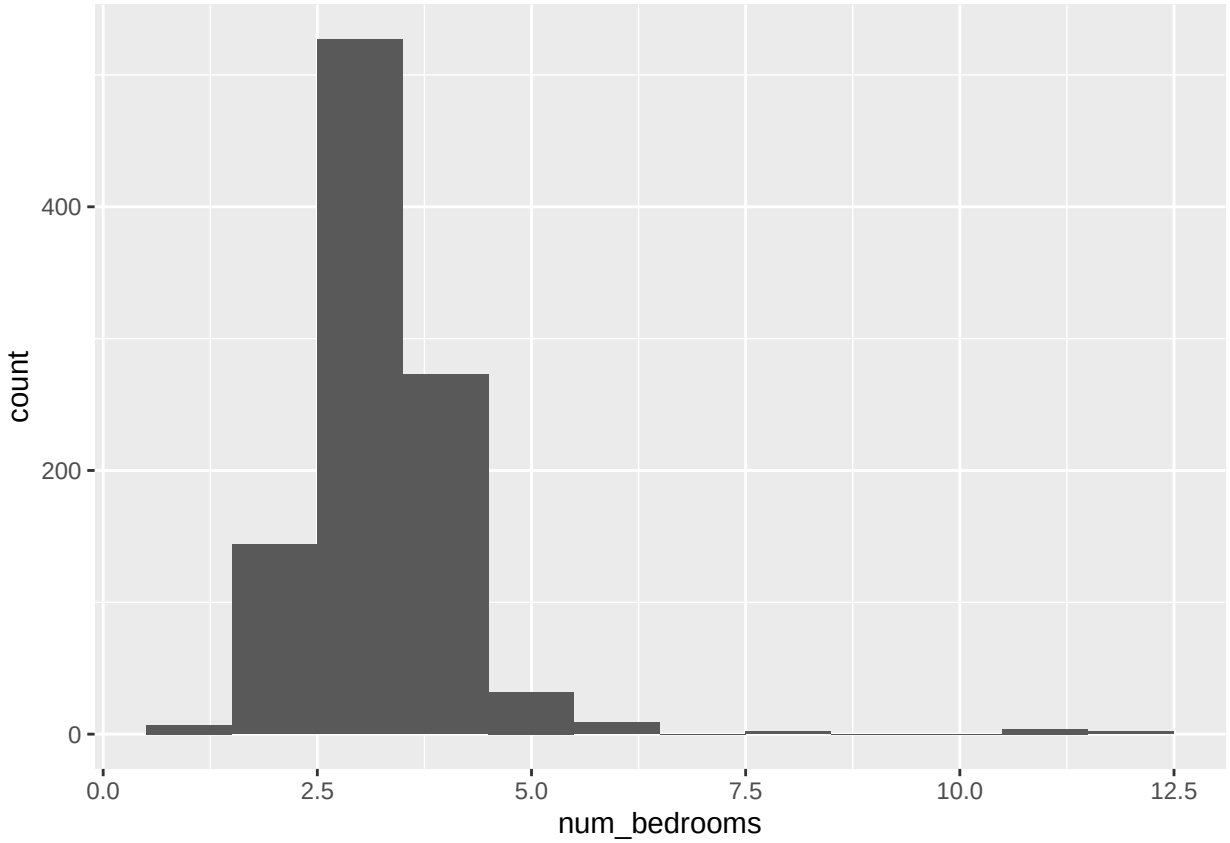
What is the mean number of bedrooms for houses in Cook County? In this context our variable is the number of bedrooms in a house, which in the data set is the variable num_bedrooms.

2. For your variable perform a single variable exploratory analysis, including summary statistics and visualizations. Describe in your own words the results of your summaries and visualizations.

```
favstats(ccao_5$num_bedrooms)
```

```
##  min Q1 median Q3 max  mean       sd    n missing
##    1  3      3  4  12 3.266 1.025815 1000       0
```

```
ggplot(ccao_5, aes(x = num_bedrooms)) +
  geom_histogram(binwidth = 1)
```



These statistics and visualizatiopns show us that the mean house in this dataset has 3 bedrooms, with most houses being close to that number, and a few outliers with much more.

3. State the null and alternative hypothesis in symbols and in words for an appropriate statistical test.

Null hypothesis: The true mean number of bedrooms for houses in Cook County is equal to the observed mean of 3.

$H_0 : \mu = 3$

Alternative hypothesis: The true mean number of bedrooms for houses in Cook County is not equal to the observed mean of 3.

$H_A : \mu \neq 98.6$

4. Run your hypothesis test, report your test statistic and p-value, and state your statistical conclusion in context.

```
t.test(~ num_bedrooms, data = ccao_5)
```

```
##
##  One Sample t-test
##
## data:  num_bedrooms
## t = 100.68, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  3.202343 3.329657
## sample estimates:
## mean of x
##     3.266
```

Test Statistic: 100.68

P value = 2.2e-16

5. Interpret an appropriate 95% confidence interval in context.

We are 95% confident that the true mean number of bedrooms for houses in Cook County falls within the interval (3.202343, 3.329657). Because of the context of the situation ,however, we round these values to 3.

6. Comment on the role of assumptions/conditions for your confidence interval.

The conditions for the confidence interval are met. The data are independent because the number of bedrooms in one house would not have an impact on the number of bedrooms in a different house, and because we have a large sample size of 1,000 we can assume that there would be roughly normal distribution.

Section D: Conclusion

Write a paragraph summarizing your findings from your 3 analyses to a general audience.

We found that a majority of the houses in Cook County have central air, with the number being between 51.9% and 58.2%. This is impacted by when a house was built, while 60.2 percent of post 1950 houses have central air, only 19.8 percent of pre-1950 houses have central air. We found that the average house in Cook County has 3 bedrooms. These make conclusions make sense, because many houses built pre-1950 would not have originally had central air, so it would have to be added on later. The mean number of bedrooms being 3 also makes sense.

Knit this file. Submit your pdf to Moodle.