STAT 172: Group Project #2

Diana, Maroova, Veronica

Due Tuesday 01/23/2024 at 11pm

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.
- Perform inference for one proportion and two proportions.
- Write accurate inferential conclusions and interpretations for each of the proportions tests.

Useful resources for this homework:

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

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 was created by selecting randomly 1000 observations from the original CCAO dataset whose price is less than $500,000,000 (half a million dollars). This dataset can be loaded using the following command:

```
ccao2_tbl <- read_csv("~/Stats 172 J24/Class/Data/ccao_cat1.csv")
```

```
## 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. 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.

Your Job

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.

Question: Do the majority of houses in Cook County have Central Air or no Central Air?

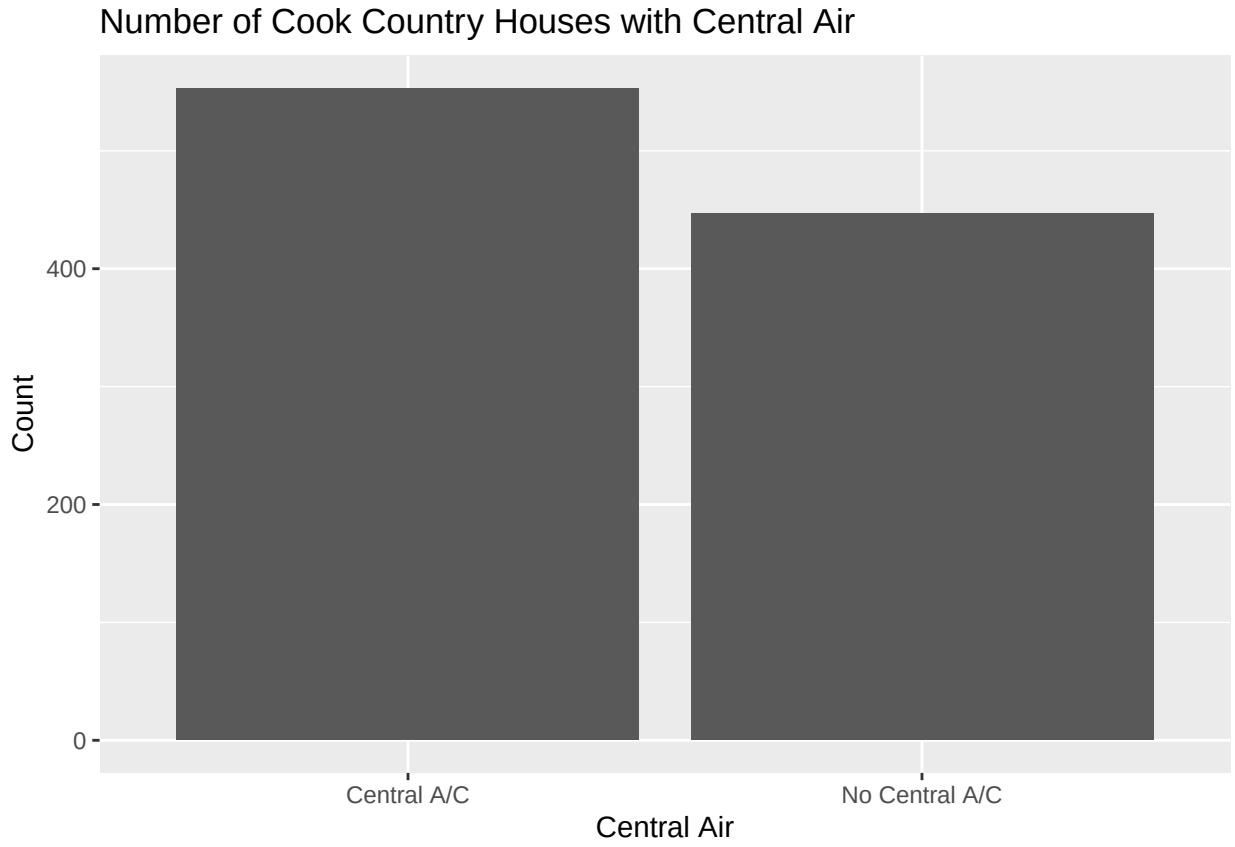
central_air

2. For your variable provide appropriate summary statistic(s) and then describe those statistic(s) in context. Describe in your own words the results of your summaries and visualizations.

```
table(ccao2_tbl$central_air)
```

```
##
##    Central A/C No Central A/C
##            553            447
```

```
ggplot(ccao2_tbl, aes(x = central_air)) +
  geom_bar() +
  xlab("Central Air") +
  ylab("Count") +
  ggtitle("Number of Cook Country Houses with Central Air")
```




```
table(ccao2_tbl$central_air) %>% proportions() %>% round(2)
```

```
##
##    Central A/C No Central A/C
##           0.55           0.45
```

We can see that there are more houses with A/C in Cook County houses rather than without. This categorical variable is non-ordinal (2-levels) as well as there is no natural order to the levels (only yes or no)

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

$$H_0 : p = 0.5$$

The proportion of houses with central air in Chicago is equal to 0.50

$$H_a : p \neq 0.5$$

The proportion of houses with central air in Chicago is not equal to 0.50

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

Houses have A/C (Success) : 553 Total of houses (n) : 1000

```r
prop.test(x = 553, n = 1000, p = 0.50, alternative = "two.sided")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  553 out of 1000
## X-squared = 11.025, df = 1, p-value = 0.0008989
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5215390 0.5840488
## sample estimates:
##     p
## 0.553
```

P-value: 0.0008989 Z-Score: 3.320392 (sqrt of 11.025)

CONCLUSION: Our z-score is greater than 2 (3.32 > 2), which tells us that our p-value will be less than 5%, so we will reject the null hypothesis. We are given our p-value which is 0.0008989, this is smaller than 5% so we reject the null hypothesis that the majority of houses in Chicago have A/C.

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

From the hypothesis test we are given our 95% confidence interval of [0.521, 0.584] We are 95% confident that the true proportion of houses that have A/C in Chicago is between 52% and 58%. Based on the 95% confidence interval [0.521, 0.584] out point estimate of 50% does not fall within the interval, so we can conclude that there is sufficient evidence to reject the null hypothesis.

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

95% confidence interval assumes the conditions of the central limit theorem. 1. The conditions for independence are met as one house having A/C does not affect another house having A/C 2. The conditions for sample size are met as we have more than 10 success (553 > 10) and 10 failures (447>10).

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.

Question: Does the year that the house was built affect whether it has central air or not?

Explanatory = year_build_cat

-Year_build-_cat divides houses into pre-1950 built and post-1950 built (binary).

Response = central_air

-Central_Air determines whether a house has A/C or not and is binary Y/N in our dataset.

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

```
table(ccao2_tbl$year_built_cat, ccao2_tbl$central_air)
```

```
##
##             Central A/C No Central A/C
##   Post-1950         519            359
##   Pre-1950           34             88
```

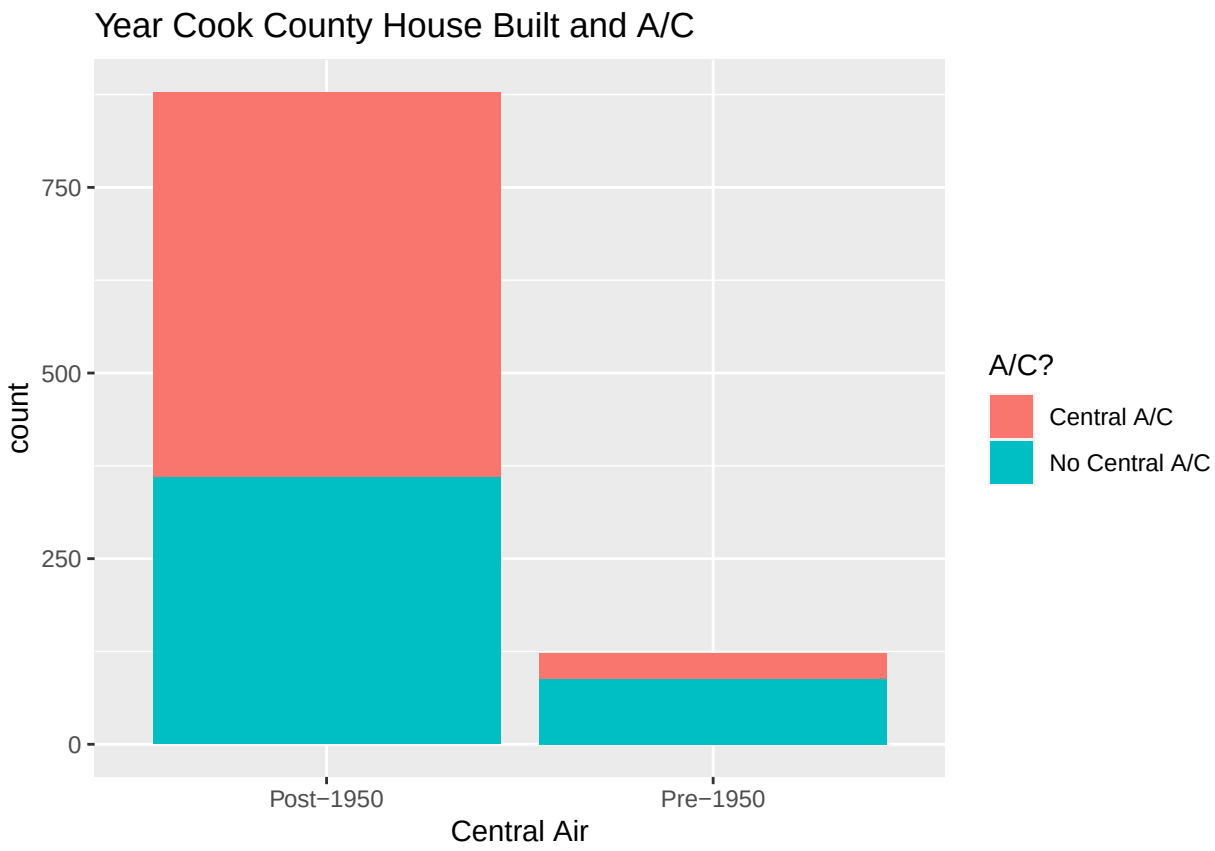
```
table(ccao2_tbl$year_built_cat, ccao2_tbl$central_air) %>% addmargins()
```

```
##
##             Central A/C No Central A/C  Sum
##   Post-1950         519            359  878
##   Pre-1950           34             88  122
##   Sum               553            447 1000
```

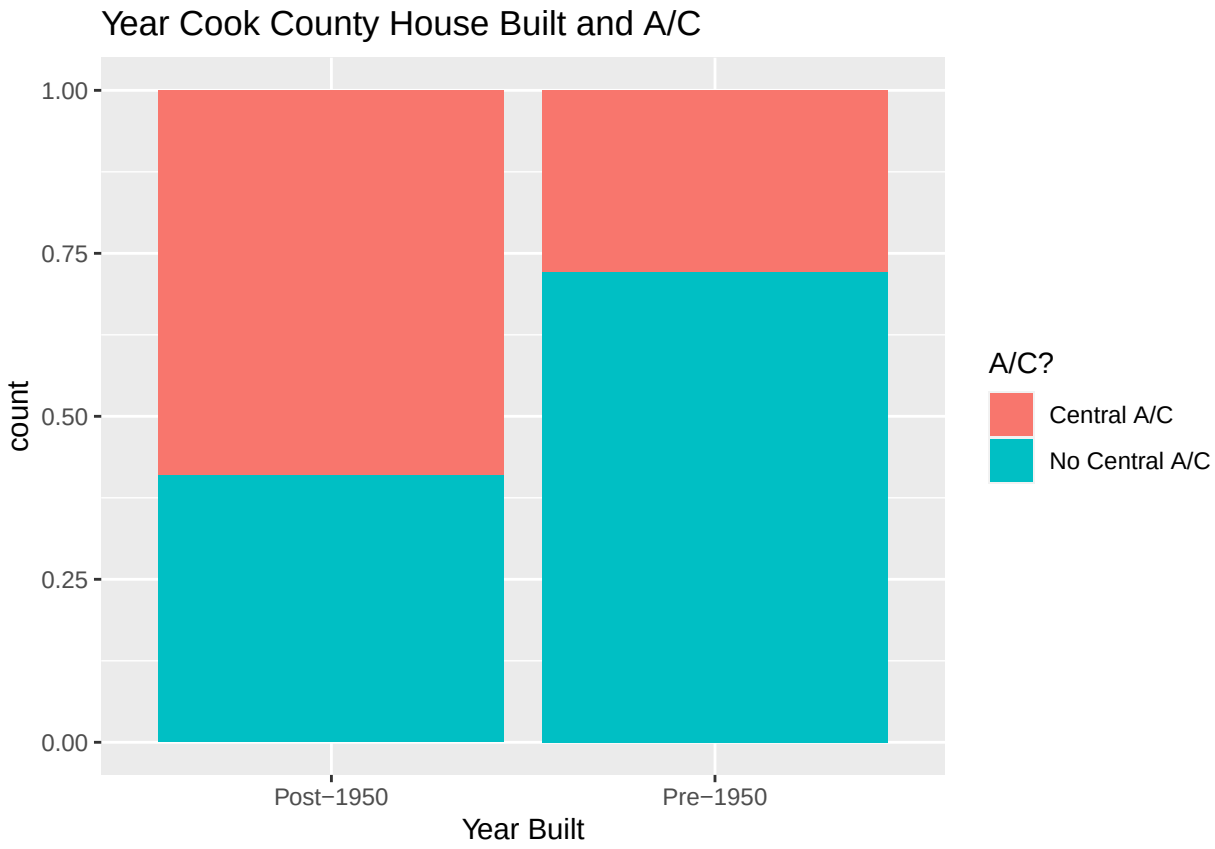
```
table(ccao2_tbl$year_built_cat, ccao2_tbl$central_air) %>% proportions(margin=1) %>% round(3)
```

```
##
##             Central A/C No Central A/C
##   Post-1950       0.591          0.409
##   Pre-1950        0.279          0.721
```

```
ggplot(data = ccao2_tbl, aes(x = year_built_cat, fill = central_air)) +
  geom_bar() +
  labs(x = "Central Air",
       title = "Year Cook County House Built and A/C",
       fill = "A/C?")
```

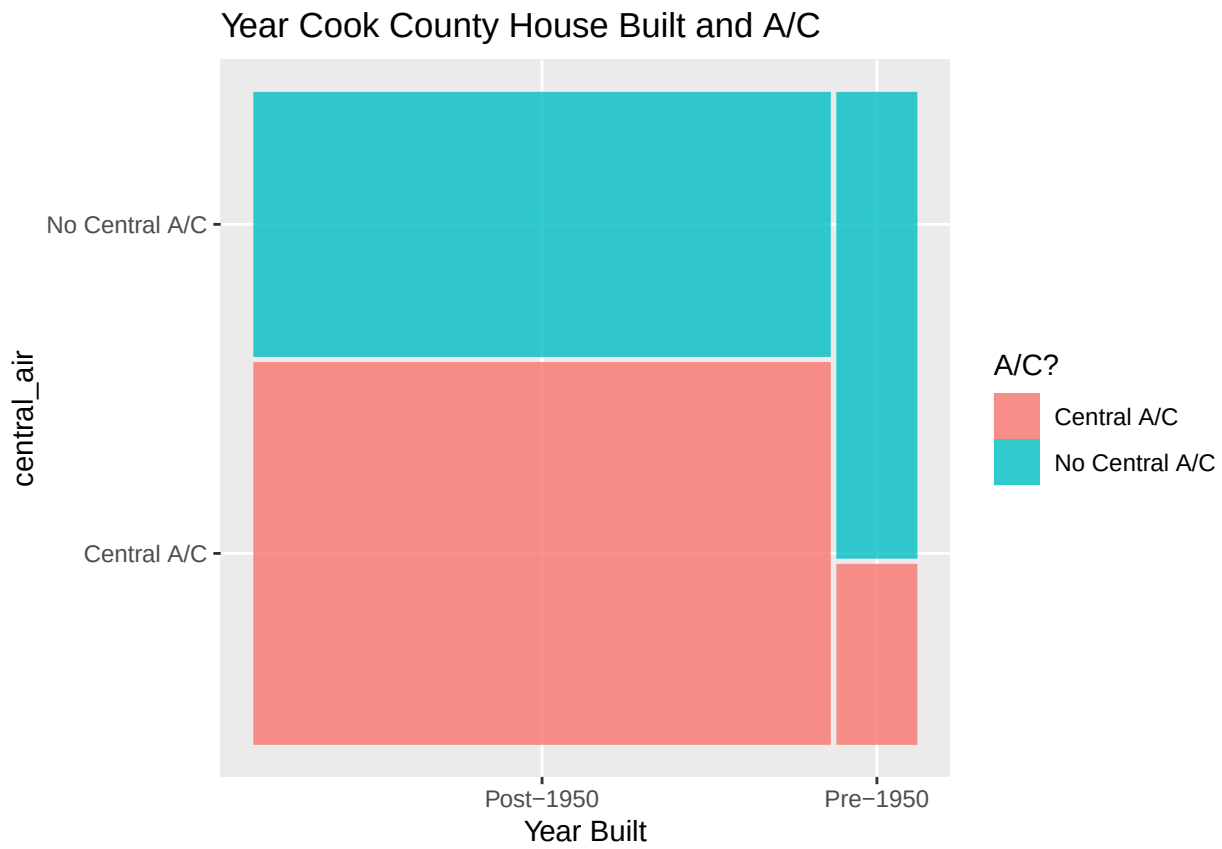



```
ggplot(data = ccao2_tbl, aes(x = year_built_cat, fill = central_air)) +
  geom_bar(position="fill") +
  labs(x = "Year Built",
       title = "Year Cook County House Built and A/C",
      fill = "A/C?")
```




```
library(ggmosaic)
ggplot(data = ccao2_tbl) +
  geom_mosaic(aes(x = product(year_built_cat), fill = central_air))+
  labs(x = "Year Built",
  title = "Year Cook County House Built and A/C",
       fill = "A/C?")
```

```
## Warning: `unite_()` was deprecated in tidyr 1.2.0.
## i Please use `unite()` instead.
## i The deprecated feature was likely used in the ggmosaic package.
##   Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



878 houses were built post-1950, and 122 were built pre-1950. 553 houses have A/C, and 447 do not. Majority of post-1950 houses (519/878, 59.1%) have A/C. Minority of pre-1950 houses have A/C (34/122, 27.9%). Pre-1950 are more likely to not have A/C (88/122, 72.1%). Most houses that have A/C (519/553, 93.9%) were built post-1950. Most houses without A/C, (359/447, 80.3%) were built post-1950.

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

$$H_0 : \beta_1 = 0$$

The is no difference between what year the house was built and whether it has central air.

$$H_a : \beta_1 \neq 0$$

The is a difference between what year the house was built and whether it has central air.

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

Houses have A/C (Success) : 553 Total of houses (n) : 1000 Houses built post 1950 (success) : 878 Total of houses (n) : 1000

122 pre 1950 878 post 1950

```r
mosaic::prop.test(central_air ~ year_built_cat, data = ccao2_tbl, conf.level = 0.95, alternative = "two
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(central_air ~ year_built_cat)
## X-squared = 41.044, df = 1, p-value = 1.489e-10
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2218115 0.4030438
## sample estimates:
##    prop 1    prop 2
## 0.5911162 0.2786885
```

P-value: 1.489e-10 Z-Score: 6.406559 (sqrt of 41.044)

CONCLUSION: Our z-score is greater than 2 (6.406559 > 2), which tells us that our p-value will be less than 5%, so we will reject the null hypothesis. We are given our p-value which is 1.489e-10, this is smaller than 5% so we reject the null hypothesis that there is no difference between what year the house was built and whether it has central air or not.

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

From the hypothesis test we are given our 95% confidence interval of [0.221, 0.403] We are 95% confident that the true proportion of houses that have A/C in Chicago is between 22% and 40%. Based on the 95% confidence interval [0.221, 0.403] our point estimate of 0 does not fall within the interval, so we can conclude that there is sufficient evidence to reject the null hypothesis

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

95% confidence interval assumes the conditions of the central limit theorem. 1. The conditions for independence are met as one house having A/C does not affect another house having A/C 2. The conditions for sample size for post 1950 are met as we have more than 10 success (Post 1950 519 > 10) and 10 failures (Post 1950 359 > 10). The conditions for sample size for pre 1950 are met as we have more than 10 success (Pre 1950 34 > 10) and 10 failures (pre 1950 88 > 10)

Section C: Conclusion

Write one to two paragraphs summarizing your findings from your two analyses to a general audience.

Based on the first analysis of Cook County Houses the proportion of houses containing A/C is different from 50%. We found that the p-value was 0.0008989, which is smaller than 5%, and our point estimate was not included in the confidence interval [0.521, 0.584], which gives us enough evidence to claim that more than 50% of houses in Cook County have A/C.

Based on the second analysis of Cook County Houses we investigated if there is a difference in proportion between houses pre 1950 and post 1950 and whether they contain A/C or not (p1 = p2?). After conducting a prop.test we observed that the p-value was 1.489e-10 which is smaller than 5%. Our confidence interval was [0.221, 0.403] which did not include our null value of 0, allowing us to reject the null hypothesis. Our findings suggest there is a difference in proportions between what year the house was built and whether it has central air ($p1 \neq p2$).

Knit this file. Submit your pdf to Moodle. Make sure to submit only one PDF per group