
**Insightful and vast USA statistics**

Name : Soumyajeet Patra
Course IE6200
Date : 12/4/2019
Course Name : Engineering Probs and Stats
Professor: Shivani Patel

# 1. Abstract

In this project we will explore the Kaggle dataset on Insightful Vast USA Statistics, to determine, explore and provide insight into various factors such as rent demographics. We will explore various statistical factors such as hypothesis testing, employing different tests to solve some questions of interest among many others. Our main aim in this simple statistical analysis is to gain some insight into various factors that affect the decision of visitors to the US who probably want to settle here. Each Question of Interest has been followed up by a comprehensive statistical test which will provide the reader with more insights into each of the variables and their relationships

# 2. Introduction

Our Dataset provides extensive observations on the socio-economic demography of the United States providing us with large datasets on each of the attributes as described below:

- Second Mortgage: Households with second mortgage statistics.
- Home Equity Loan: Households with Home equity Loan statistics.
- Debt: Households with any type of debt statistics.
- Mortgage Costs: Statistics regarding mortgage payments, home equity loans, utilities and property taxes
- Home Owner Costs: Sum of utilities, property taxes statistics
- Gross Rent: Contract rent plus the estimated average monthly cost of utility features
- Gross Rent as Percent of Income Gross rent as the percent of income very interesting
- High school Graduation: High school graduation statistics.

- Population Demographics: Population demographic statistics.
- Age Demographics: Age demographic statistics.
- Household Income: Total income of people residing in the household.
- Family Income: Total income of people related to the householder.

Variables we will explore:

- Average Rent
- Population
- The population proportion of people in different cities
- Type of location(e.g City, Town, Villages, etc.)

Our Dataset can not only be used to provide answers based on geographical location in the United States but can be used as a data source to study factors that can be used to make crucial decisions on the US population and provide a comprehensive insight into factors that affect an individuals decision on relocation, provide an in-depth analysis on Income, Age, Marriage, Mortgage, Home Equity Loan and Demographics.

While we have considered only a few attributes of the 80 odd attributes to answer our questions of Interest. We have focused on questions that can help individuals make decisions, considering variables like State-wise rent mean, population proportions in different cities which will provide insights like which city is preferred by the population in a single state such as Alaska, etc

# 3. Sampling Strategy and Data Collection Method

Our Original Dataset was retrieved from 2012-2016 ACS 5-Year Documentation was provided by the U.S. Census Reports. Retrieved May 2, 2018, from Census estimate, error, and location data and Census location information While the data may be classified as historical data there might be some inconsistencies in the data collection from the source
We have not performed any sampling on the original dataset rather used the dataset entirely for our observations, but some bias may have crept in the original data when the original dataset was created. Wherever sampling was used we have considered in setting a random seed before sampling so that the sample taken is considerably randomized

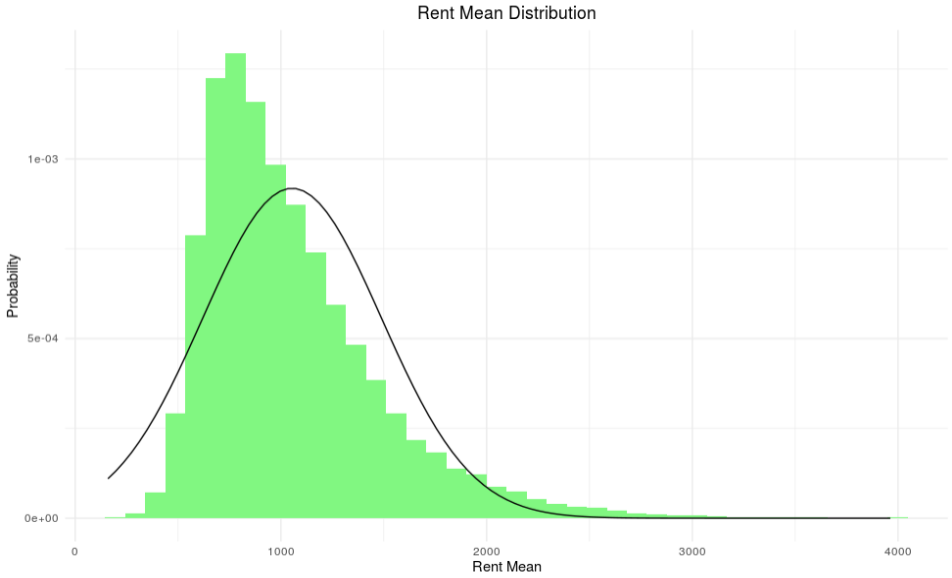
# 4. Tables of Variables

Our variables of concern have been tabulated as below:

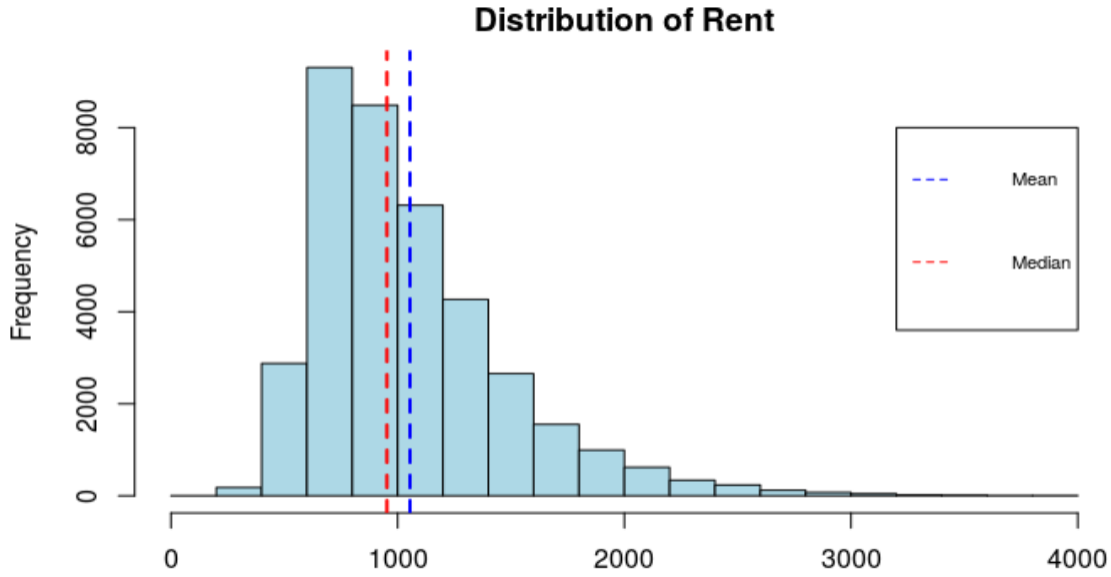
| Variable Name | Type | Scale | Description |
|---|---|---|---|
| pop | Quantitative/Discreet | 0-10000 | The population variable refers to the population number and has been used in our research as a quantitative variable in difference of means |
| female_pop | Quantitative/Discreet | 0-7000 | The population variable refers to the female population demographic in the United States. We can use this variable in our statistical analysis to conduct proportion tests |
| male_pop | Quantitative/Discreet | 0-7000 | The population variable refers to the male population demographic in the United States. We can use this variable in our statistical analysis to conduct proportion tests |
| rent_mean | Quantitative/Continuous | 0-100000 | The rent mean refers to the average mean in a city or a particular area. |
| type | Categorical/Nominal | City,Town,Borough,Village,Urban,CDP | This type refers to the type of region. This is one of the main Categorical variables used in our statistical analysis to provide insight on the over representation or under representation of samples based on location |
| state | Categorical/Nominal | All of the state names in the US | This refers to the name of the state of the US for which we are considering a certain part of the demographics in the dataset. It is also a categorical variable that has been used in our statistical analysis to find population proportion demographics |
| city | Categorical/Nominal | All of the cities in US | This refers to the name of the cities of the US , we have used this as a categorical variable for various statistical tests |

# 5. Exploratory Analysis

In this section, we compare the different effects of confounding variables on each other and their effect on our variables of interest. Let us consider those variables one by one:

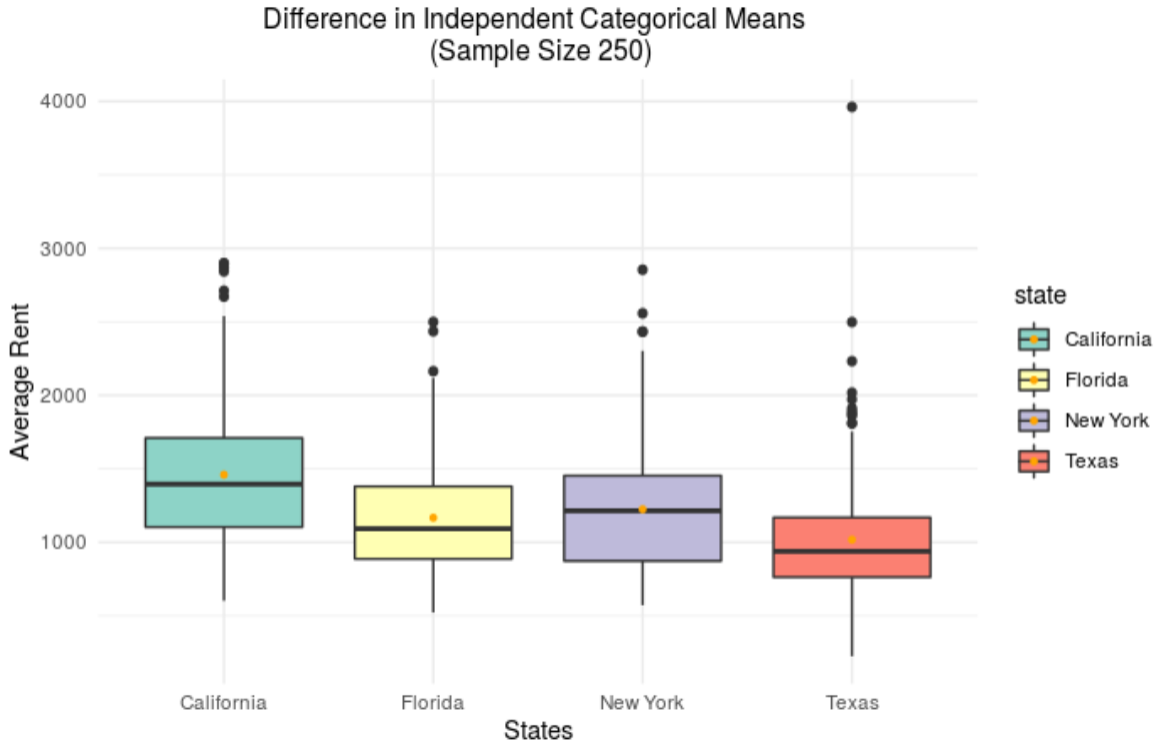



The above plot shows us a distribution of the Mean Rent in all the states of the USA. From the graph, we can derive a notion that the mean rent is lesser than 3000 USD. The mean rent depicts that the max average rent trends towards the lower end of 3000 USD.

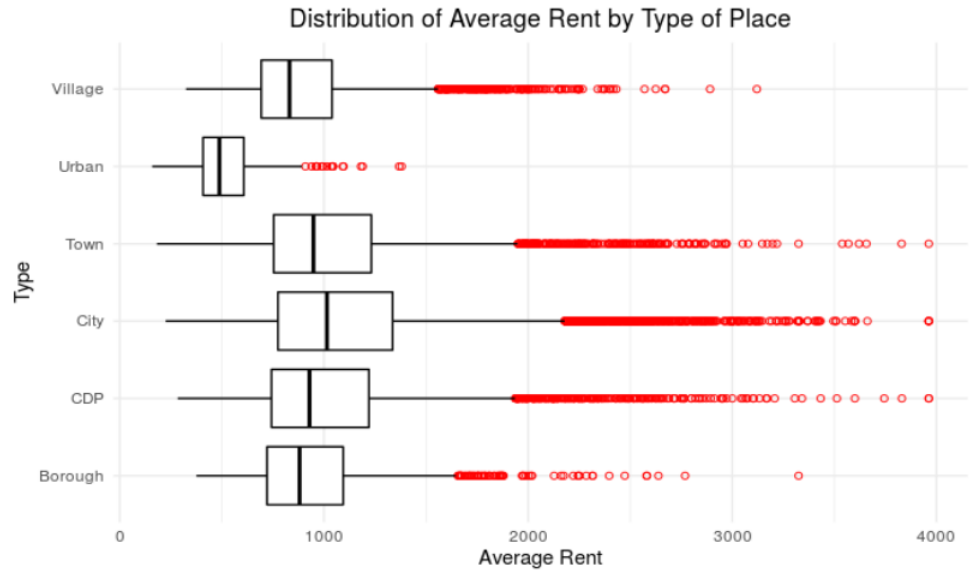



The above plot shows us the Mean and median of the Rent concerning the Rent Distribution Frequency. This Graph provides insight into the data, we can see that the median and mean

are very close together thereby showing that there are very few outliers in the dataset

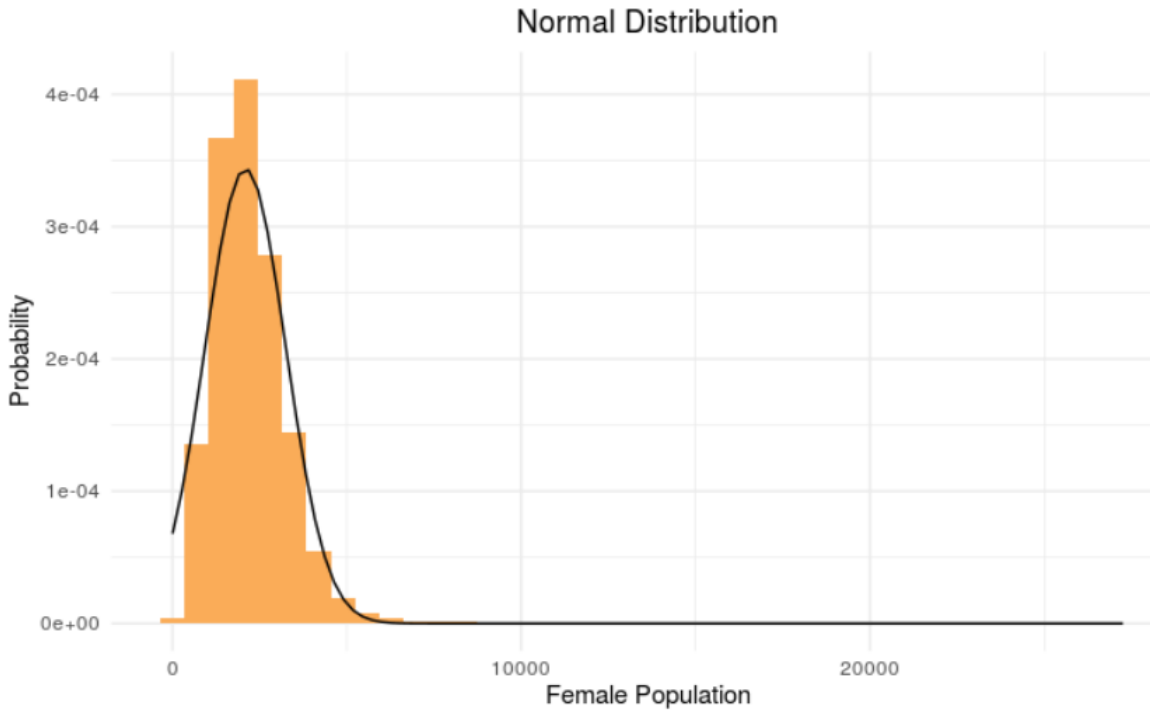



The above plot provides critical insight on the differences in means of rent amount across various cities thereby providing insight into how the rent varies by location We can use this type of analytics to compare means in different cities.

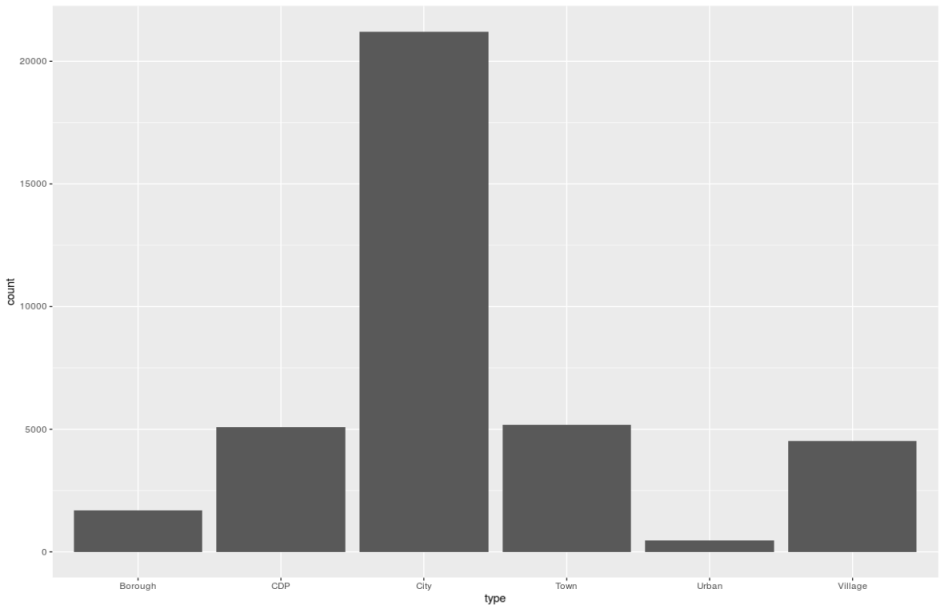



The above plot shows us a distribution of the Mean Rent by location type whether it's a city, Village, Town,city, etc.. From the graph we can derive a notion that the mean rent is

Higher in the city and it's lowest in the Urban Area. We can also infer that the Urban Area has the lowest frequency in our dataset.




The Above plot depicts the distribution of the female population in Alaska. This data is used in statistical analysis to provide us with insight into the difference in the population mean between the proportion of males living in Anchorage to the population of female living in Anchorage in Alaska



The Above plot depicts the frequency of different areas in the dataset, we can use the above data to check on the frequency proportion of each area in the dataset and check if a variable has been overrepresented or underrepresented in our dataset

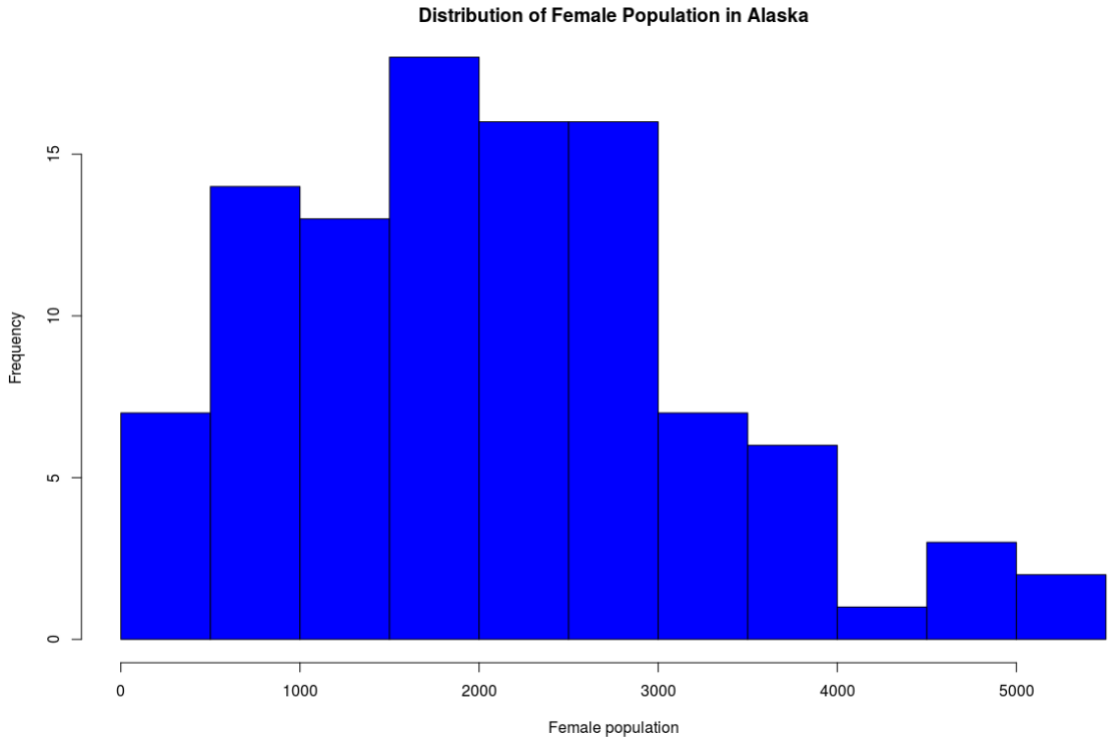



Figura 1: **Graph Depicting the population of women in Alaska**

The Above plot depicts the population of women in Alaska, using the above data we can perform statistical analysis on the preference of the female population in selecting towns or cities to dwell in and this can be further leveraged to gain insights on the city demographics

# 6. Statistical Analysis

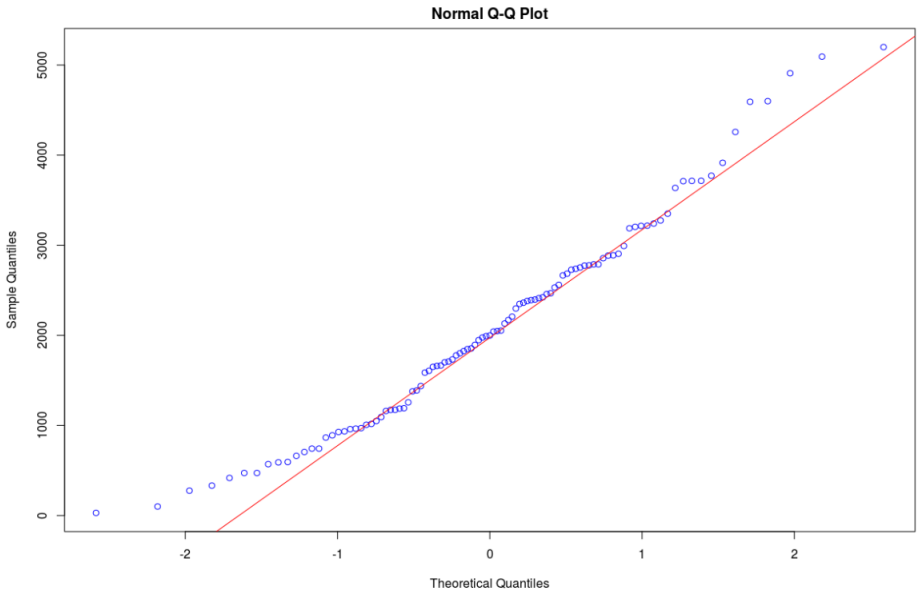
## 6.1. One Sample t-test

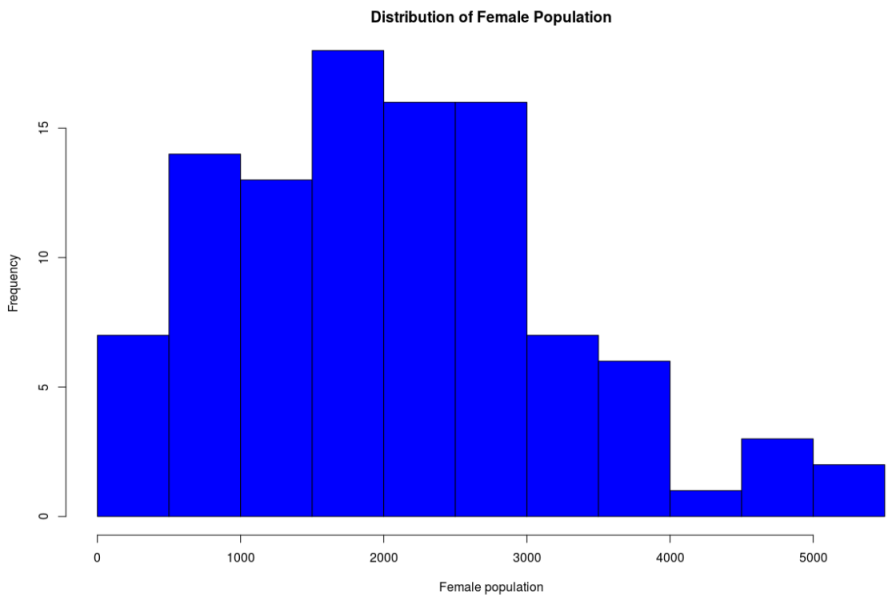
## 6.2. Question of interest

On analyzing the data in our dataset we would like to compare the mean female population in Alaska with the real-time mean population value of Alaska as provided by google i.e. Is the sample mean female population of Alaska $\mu_s = 1800$?

## 6.3. Conditions of One Sample t-test

- The sample is representative of the population - Yes.
- One quantitative variable of interest - yes, the female population is quantitative.
- We want to make inference about the population mean using the sample mean - yes.
- The population variance is unknown so we estimate it using sample data - yes.
- The sample comes from a single population - yes the sample comes from a single population.
- The population data must be normally distributed - need to check.
- To check this condition look at a QQ plot of the sample data.






## 6.4. Population Parameter

The population parameter we want to make an inference to is $\mu$.
Population Variance is unknown

## 6.5. Hypothesis

Two sided:

$$H_o : \mu = 1800$$

The Average female population for New York is 1800

$$H_o : \mu \neq 1800$$

The Average female population for Alaska is different than 1800

## 6.6. Sample Statistic

The sample statistic is the sample mean female population for Alaska $\overline{x}$

## 6.7. Test Statistic

The Test Statistic here can be denoted by :

$$t_{n-1} = \frac{\overline{x} - \mu_o}{\frac{s}{\sqrt{n}}}$$

## 6.8. Distribution of the test Statistic

$$t_{n-1} = \frac{\overline{x} - \mu_o}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

## 6.9. T-test

```
        One Sample t-test

data:  mydata$female_pop[mydata$state == "Alaska"]
t = 2.5684, df = 102, p-value = 0.01167
alternative hypothesis: true mean is not equal to 1800
95 percent confidence interval:
 1866.859 2320.325
sample estimates:
mean of x
 2093.592
```

Our manually calculated P-value is 0,0116656
We can see that the p-value is 0.0116656 almost equal to the computed one, Thereby confirming the computed value.

## 6.10. Confidence Interval

Using the below formula to calculate the confidence Interval in for our two-sample t-distribution

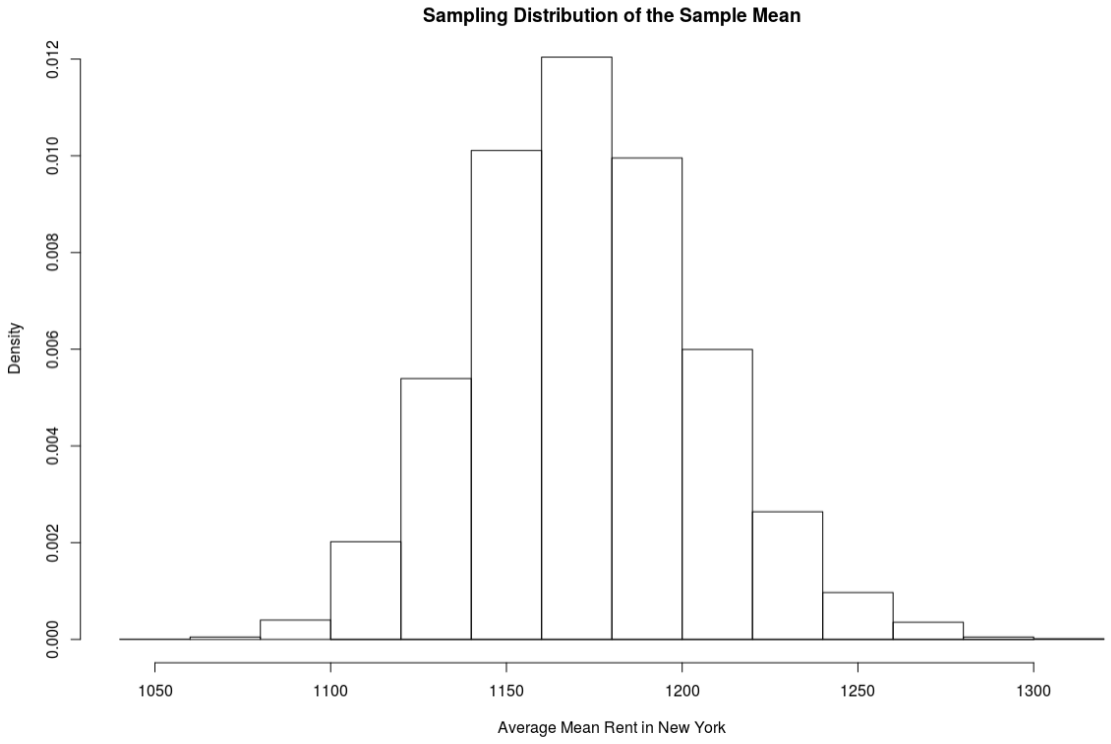
$$P(\bar{x} - t_{(\frac{\alpha}{2},n-1)}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\frac{\alpha}{2},n-1)}\frac{s}{\sqrt{n}}) = (1-\alpha)$$

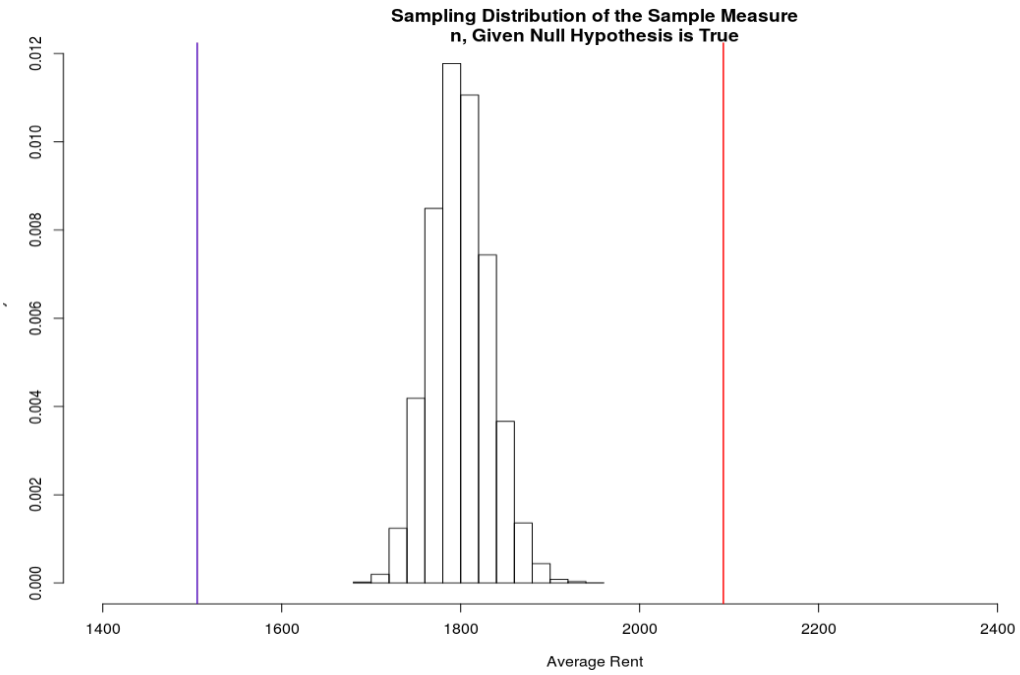
The lower bound of our confidence interval is 1866.859 The upper bound of our confidence interval is 2320.325

## 6.11. Interpretation

There is strong evidence (p-value of nearly 0.0116) to suggest that the true mean female population in Alaska is different than 1800. We reject the null hypothesis that the true mean female population is 1800. With 95 percent confidence, the true mean female population is between 1866.859 and 2320.325 which suggests that the true mean rent is greater than in 1800.

## 6.12. Bootstrapping






Using the Bootstrapping method, we made simulations for 10000 of our samples and we found that bootstrapped p-value and p-value using traditional methods is almost the same (0.0, 0.011) which agrees with a confidence interval from the traditional method and bootstrap method (empirical Method). The reason why the upper bound of a confidence interval for the bootstrap method is different than traditional methods is that theoretically upper bound can be 1 but empirically it has some value at 100

### 6.13. Confidence Interval

Our Confidence Intervals are at 1110.509 and 1241.690

### 6.14. Interpretation

Our bootstrapping p-value agrees with our normal p-value which rejects the null hypothesis thereby we reject our null hypothesis that the mean female population is equal to 1800

## 7. One sample test of proportions

### 7.1. Question of interest

Our question of interest is to determine whether 33.33 or 1/3rd of the population in Alaska live in Anchorage.

## 7.2. Conditions of One Sample proportion test

- Exact Binomial Test
- Categorical variable of interest with 2 categories - Yes! Anchorage or other city in Alaska
- Sample comes from one population - yes.
- Normal approximation - yes.
- $np \geq 10 = 150561$
- $n(1-p) \geq 10 = 301283$

## 7.3. Population Parameter

Our population parameter is the proportion of the population who live in Anchorage.

## 7.4. Hypothesis

Two sided:

$$H_o : p_R = 0{,}3333$$

The True proportion of people who live in Anchorage is .3333

$$H_o : p_R > 0{,}3333$$

The proportion of people who live in Anchorage is greater than .3333

## 7.5. Sample Statistic

Our Sample Statistic is $\frac{150561}{451844} = 0{,}3332146$

## 7.6. Test Statistic

Exact test: there is no test statistic, we can find the probability directly. Normal approximation (using $p_o$ to find the test statistic - Score test statistic)

$$z = \frac{p - p_o}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0{,}3332146 - 0{,}33}{\sqrt{\frac{0{,}33(1-0{,}33)}{451844}}}$$

## 7.7. P Value

Our Manually calculated p value is 0.5484848

```
        Exact binomial test

data:  x and n
number of successes = 150561, number of trials = 451844, p-value = 0.5678
alternative hypothesis: true probability of success is greater than 0.3333333
95 percent confidence interval:
 0.3320608 1.0000000
sample estimates:
probability of success
             0.3332146
```

We can see that the p-value is 0.5678 almost equal to the computed one, Thereby confirming the computed value.

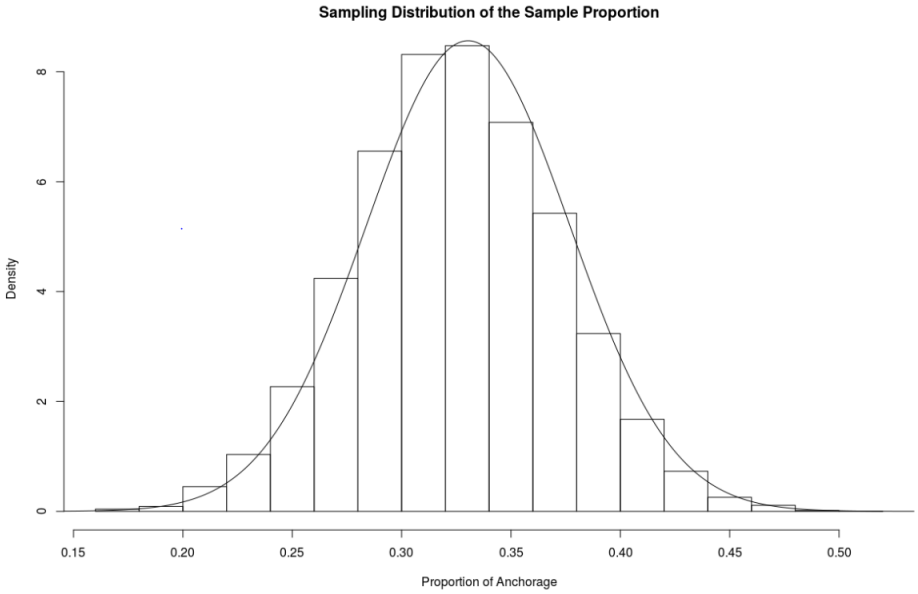
## 7.8. Confidence Interval

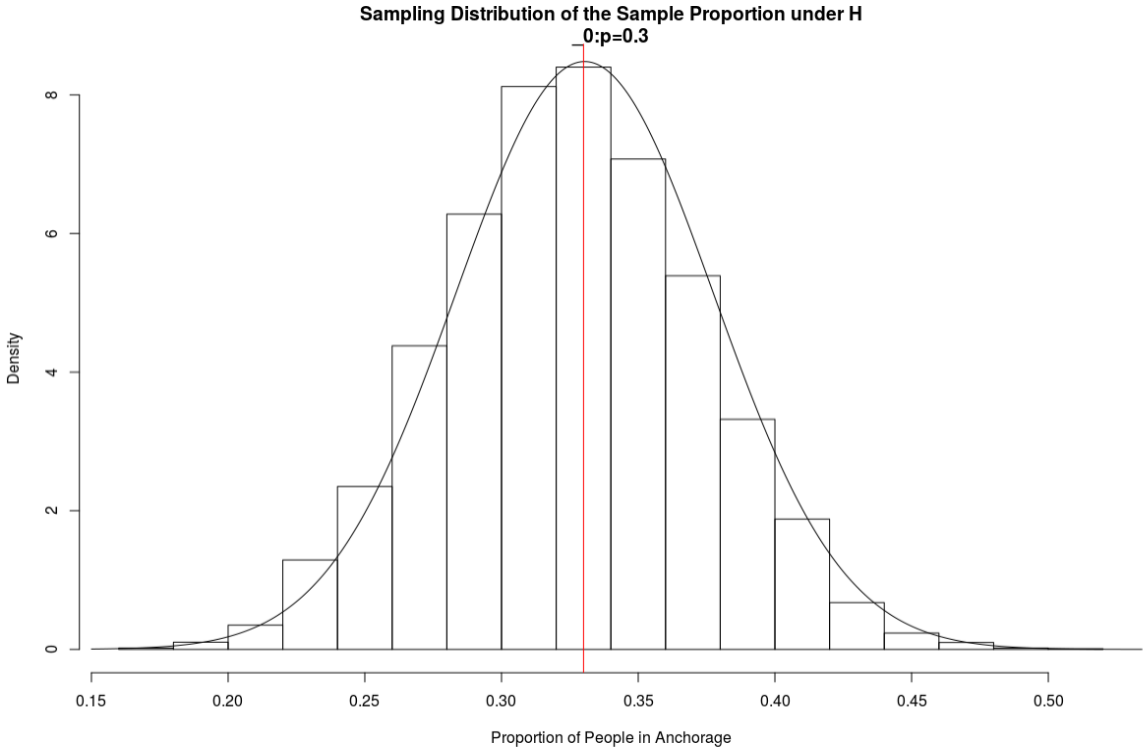
The lower bound of our confidence interval is 0.3320608 The upper bound of our confidence interval is 1.0000000

## 7.9. Interpretation

There is very little evidence (p-value=0.5678) that the true proportion of people who live in anchorage greater than .33333. We can accept the null hypothesis that the true proportion of people who live in anchorage is equal to .3333 We are 95 percent Confident that the true proportion of people who live in anchorage is between 0.3320608 and 1.0000000

## 7.10. Bootstrapping




Using the Bootstrapping method, we made simulations for 10000 of our samples and we found that bootstrapped p-value and p-value using traditional methods is almost the same (0.5678, 0.5655) which agrees with a confidence interval from the traditional method and bootstrap method (empirical Method). The reason why the upper bound of the confidence interval for the bootstrap method is different than traditional methods is that theoretically

upper bound can be 1 but empirically it has some value at 100 percent. Our Confidence Intervals are at 0.2595 and 0.5100

### 7.11. Confidence Interval

Our Confidence Intervals are at 0.2595 and 0.5100

### 7.12. Interpretation

Our bootstrapping p-value agrees with our normal p-value which accepts the null hypothesis thereby we accept our null hypothesis that the proportion of people in Anchorage is 1/3rd of people living in Alaska After using simulations We are 95 percent confident that the true population mean lies between 0.2595 and 0.5100

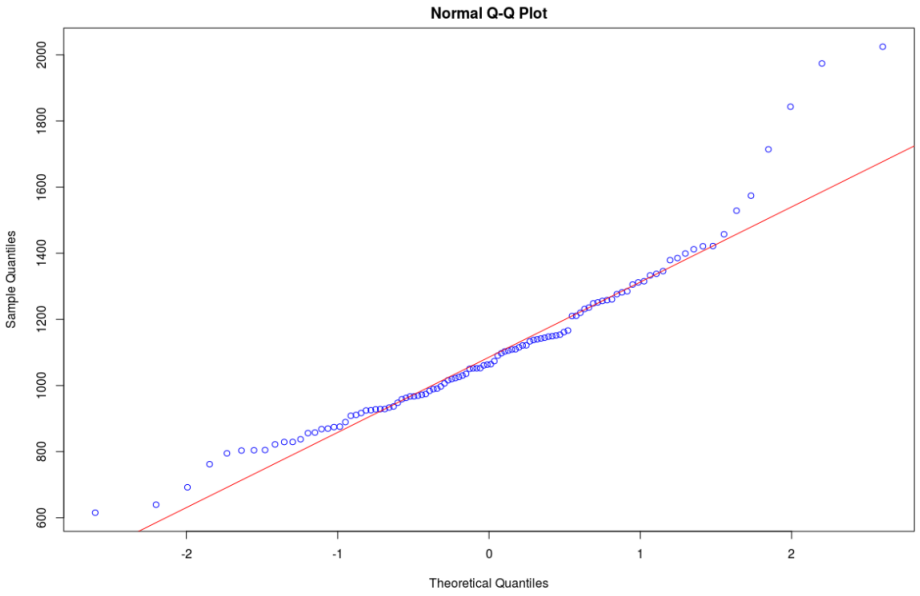
## 8. t-test - Difference of Means

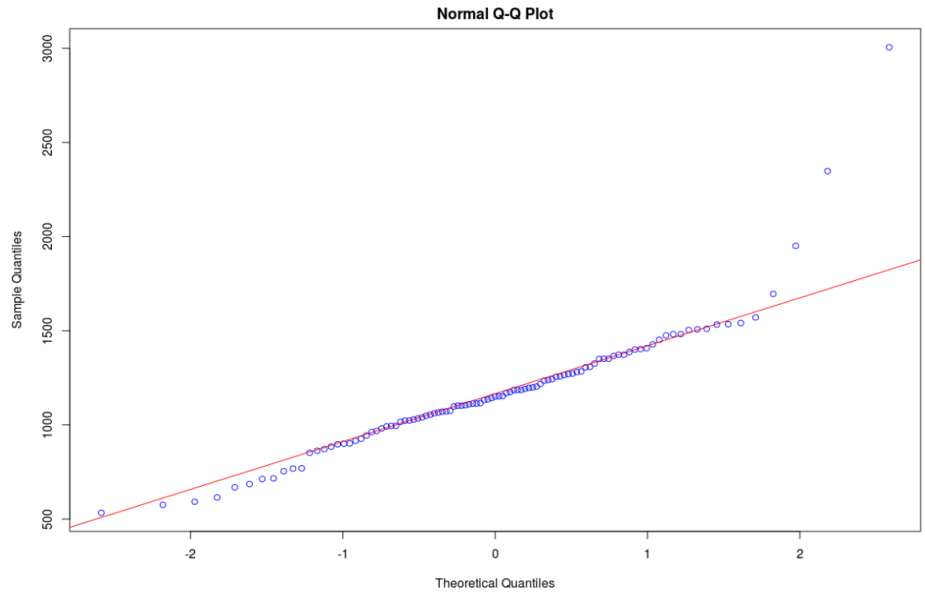
### 8.1. Question of interest

Our Question of Interest is to find out if the mean of average rent in Delaware is higher than Alaska. This will provide us a comprehensive analysis of whether it is more feasible for people to settle in either of the two.

### 8.2. Conditions of two Sample t-test

- The sample is representative of the population -Yes.
- Question of interest has to do with the difference of means between two populations.- Yes the one from Alaska and the other from Delaware
- 2 independent samples from 2 populations - Yes.
- The population data must be normally distributed. Yes




## 8.3. Population Parameter

Our population parameter of interest here is the mean difference of the rent for people living in Delaware as well as in Alaska $\mu_D$ and $\mu_A$

## 8.4. Hypothesis

$$H_o : \mu_D - \mu_A = 0 or H_0 : \mu_D = \mu_A$$

Both being equal statements The true population mean of rent for those in Delaware is equal to the true population mean rent for those who dwell in Alaska.

$$H_o : \mu_D - \mu_A \neq 0 or H_0 : \mu_D \neq \mu_A$$

Both being equal statements The mean rent for people living in Delaware is not equal to the mean rent for people living in Alaska.

## 8.5. Sample Statistic

$\overline{x}_d - \overline{x}_a = 0$ or $\overline{x}_d = \overline{x}_a$

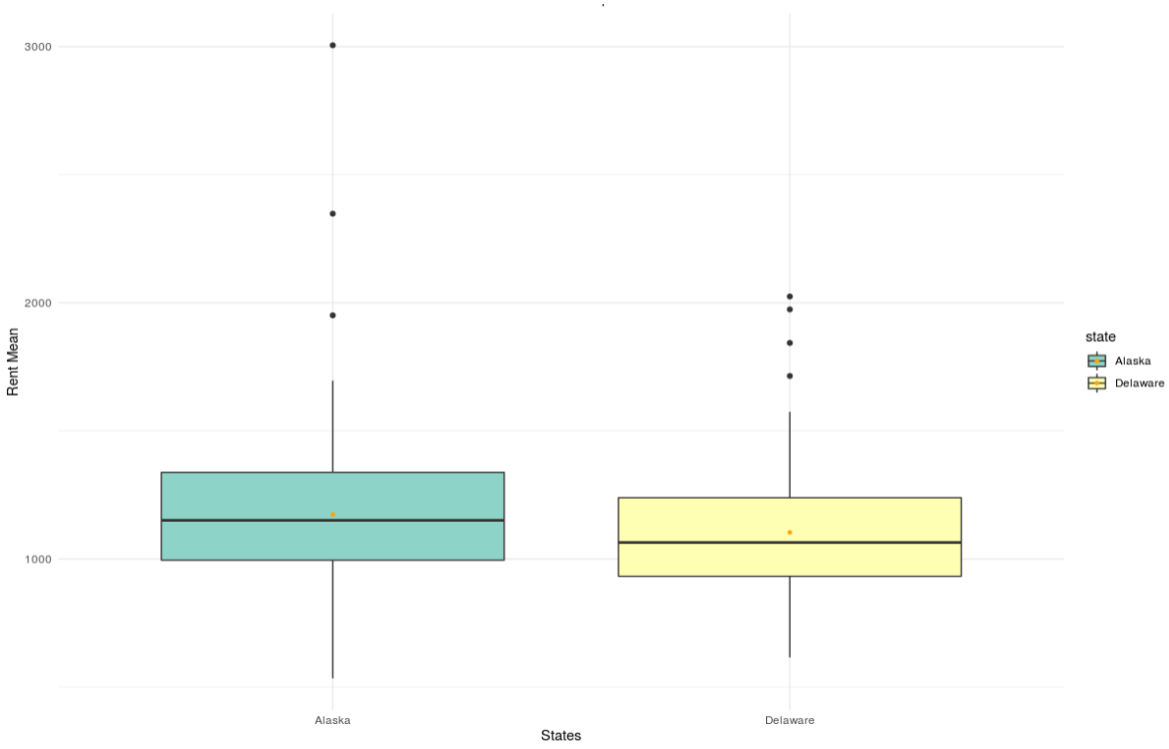
## 8.6. Test Statistic

The Test Statistic here can be denoted by :

$$t_{min(n_d-1,n_a-1)} = \frac{(\overline{x}_d - \overline{x}_a) - (\mu_d - \mu_a)}{\sqrt{\frac{s_d^2}{n_d} + \frac{s_a^2}{n_a}}}$$

## 8.7. Distribution of the test Statistic

$$t_{min(n_d-1,n_a-1)} = \frac{(\overline{x}_d - \overline{x}_a) - (\mu_d - \mu_a)}{\sqrt{\frac{s_d^2}{n_d} + \frac{s_a^2}{n_a}}} \sim t_{min(n_d-1,n_a-1)}$$




## 8.8. T-test

Computed p-value = 0.0905

```
        Welch Two Sample t-test

data:  mydata$rent_mean[mydata$state == "Delaware"] and mydata$rent_mean[mydata$state == "Alaska"]
t = -1.7015, df = 187.02, p-value = 0.0905
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -150.11724   11.07981
sample estimates:
mean of x mean of y
 1103.488  1173.007
```

Our manually calculated P value is 0,0917
We can see that the p value is 0.0917 almost equal to computed one, Thereby confirming the computed value.

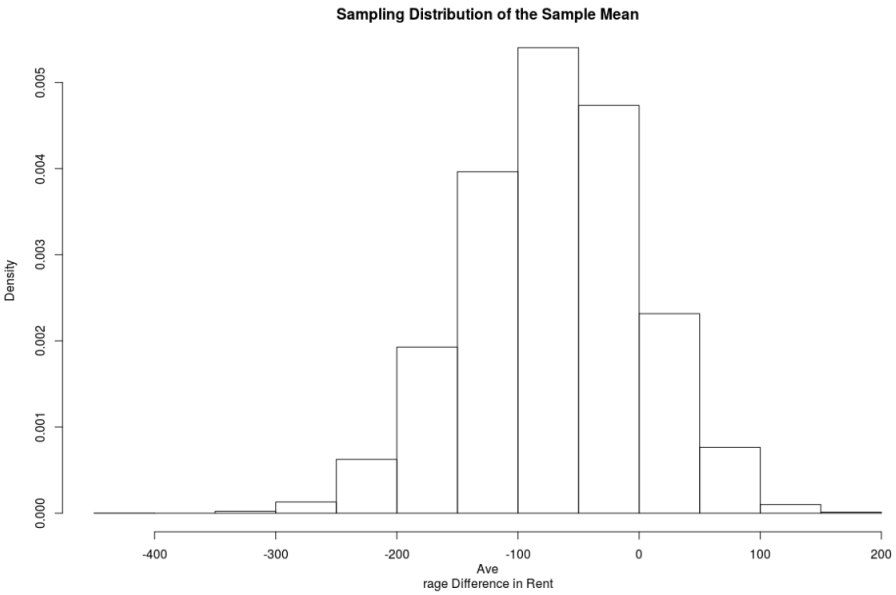
## 8.9. Confidence Interval

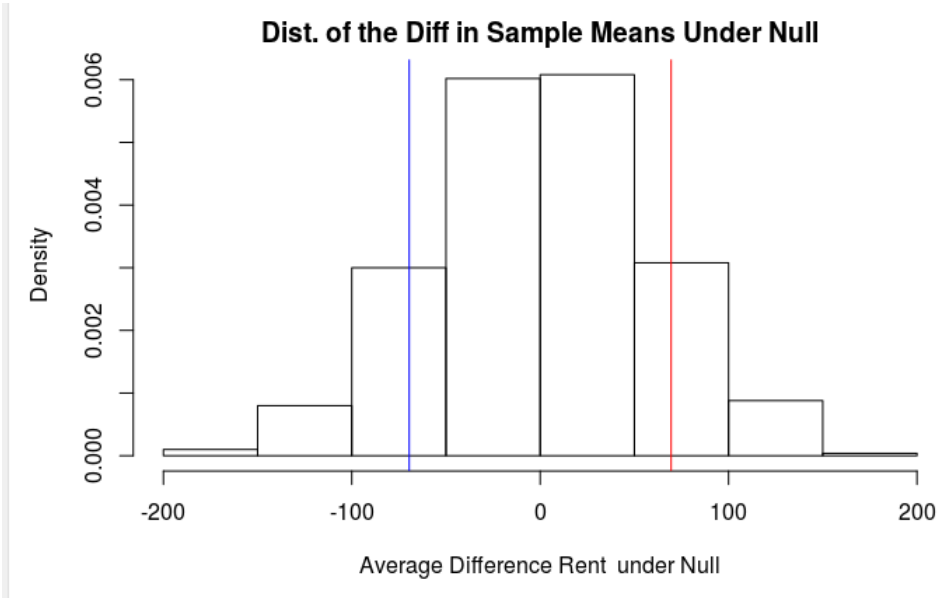
The lower bound of our confidence interval is -150.51164 The upper bound of our confidence interval is 11.47422

## 8.10. Interpretation

By our traditional method of calculating p-value we can see that there is very little evidence (p-value of 0.0917) that there is an actual difference in the mean of rent amount in the two cities, we therefore accept our null hypothesis that the mean rent in Delaware is equal to the mean rent in Alaska and we are 95 percent confident that our difference in mean will lie between -150.51164 and 22.47422

## 8.11. Bootstrapping



Using the Bootstrapping method, we made simulations for 1000 of our samples and we found that bootstrapped p-value and p-value using traditional methods is different (0.249, 0.0917) which agrees with a confidence interval from the traditional method and bootstrap method (empirical Method).

### 8.12. Confidence Interval

Our Confidence Intervals are at -150.51164 and 11.47422

### 8.13. Interpretation

There is very little evidence (p-value=0.0905) to suggest that there is a difference between the mean rent for those in Delaware and those in Alaska. Although our traditional p-value and simulated bootstrap p-value seem to be on the same stand, we know both of them can be different. We fail to reject the null hypothesis that there is no difference between the mean rent in both states. With 95 percent confidence, the true difference in the mean rent between the two states is between -150.11724 and 11.47422

## 9. Two sample test of proportions

### 9.1. Question of interest

We are in interested in the difference between the true proportion of females that live in Anchorage in Alaska and the true population proportion of males living in Anchorage in Alaska $p_f - p_m$

### 9.2. Conditions of two Sample proportion- test

- Sample needs to be representative of the population - Yes The sample is representative of the population
- Categorial response variable with 2 categories - Yes
- 2 independent samples from 2 populations (aka another categorical variable with 2 categories) - Yes.
- $np \geq 10/n(1-p) \geq 10$ for both populations - Yes.

## 9.3. Population Parameter

We are in interested in the difference between the true proportion of females that live in Anchorage in Alaska and the true population proportion of males living in Anchorage in Alaska $p_f - p_m$

## 9.4. Hypothesis

$$H_0 : p_f - p_m = 0$$

There is no difference between the true population proportion of females living in Anchorage and the true population proportion of Males living in Anchorage

$$H_0 : p_f - p_m \neq 0$$

There is a difference between the true population proportion of females living in Anchorage and the true population proportion of Males living in Anchorage

## 9.5. Sample Statistic

$$\hat{p}_f - \hat{p}_m$$

## 9.6. Test Statistic

The Test Statistic here can be denoted by :

$$z = \frac{\hat{p_c} - \hat{p_t}}{\sqrt{\frac{\hat{p_c}(1-\hat{p_c})}{n_c} + \frac{\hat{p_t}(1-\hat{p_t})}{n_t}}}$$

## 9.7. Distribution of the test Statistic

$$z = \frac{\hat{p_c} - \hat{p_t}}{\sqrt{\frac{\hat{p_c}(1-\hat{p_c})}{n_c} + \frac{\hat{p_t}(1-\hat{p_t})}{n_t}}} \sim N(0,1)$$

## 9.8. Two sample proportion test

Our manually calculated P-value is $8{,}616901e-16$
We can see that the p-value is $8{,}616901e-16$ almost equal to 0

## 9.9. Confidence Interval

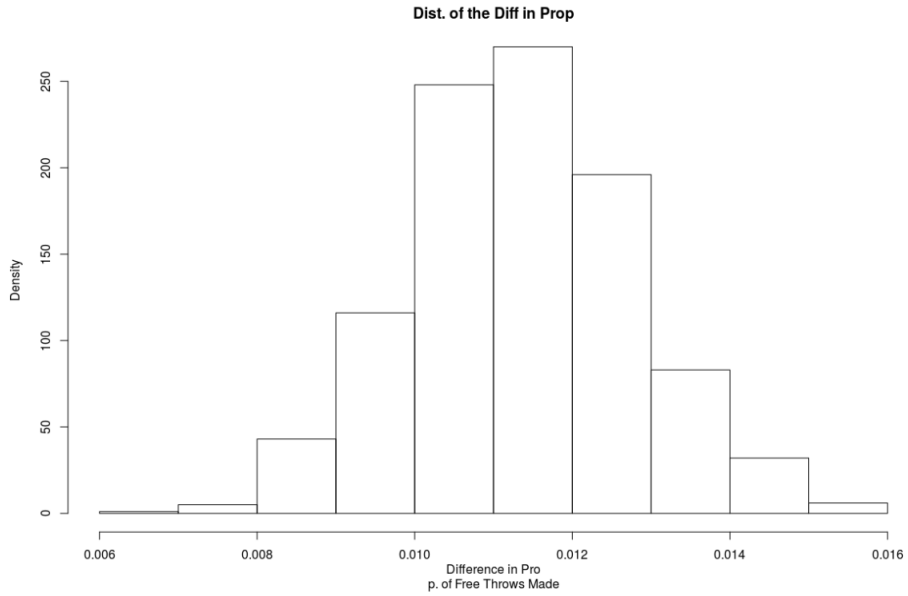
The lower bound of our confidence interval is 0.008545709 The upper bound of our confidence interval is 0.01405068

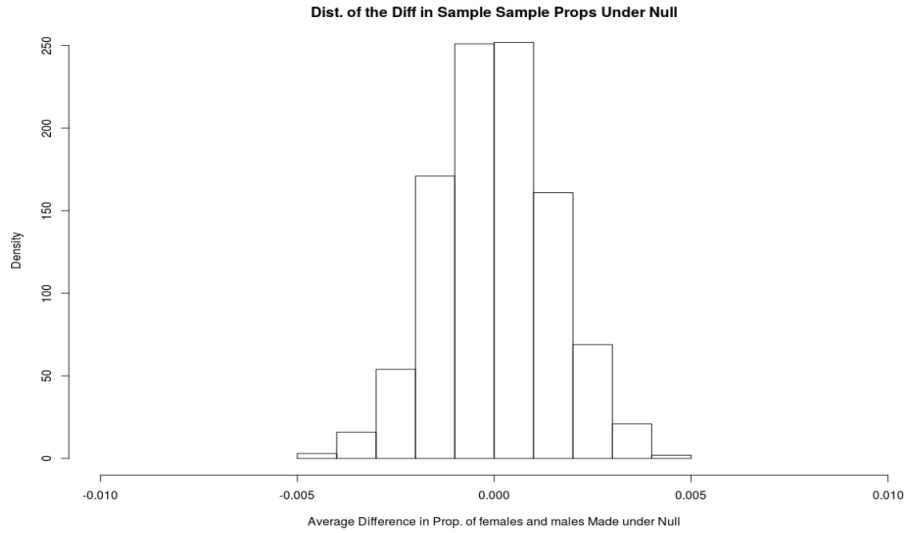
## 9.10. Interpretation

There is strong evidence (p-value of nearly 0.0116) to suggest that the true mean female population in Alaska is different than 1800. We reject the null hypothesis that the true mean female population is 1800. With 95 percent confidence, the true mean female population is between 1866.859 and 2320.325 which suggests that the true mean rent is greater than in 1800.

## 9.11. Bootstrapping

Using Bootstrapping method as seen in Figure 21 and 22, we made simulations for 1000 of our samples and we found that bootstrapped p-value and p-value using traditional methods is almost same (0.0, 8.616901e-16) which agrees with a confidence interval from the traditional method and bootstrap method (empirical Method). Our Confidence Intervals are at 0.008620735 and 0.014193155 with a 95 percent confidence level






.

## 9.12. Confidence Intervals

Our Confidence Intervals are at 0.008620735 and 0.014193155 with 95 percent confidence level

## 9.13. Interpretation

Using randomization methods, there is strong evidence (p-value $\sim 0$) to suggest that there is a difference between the true proportion of females in Anchorage to males in Anchorage. We reject the null hypothesis that the true proportion of females living in Anchorage in Alaska is equal to the true proportion of males in Anchorage. Using confidence intervals created by the bootstrap method, we can say with 95 percent confidence that the true population agrees with our success to reject the null hypothesis. r

# 10. Chi-Squared Test

## 10.1. Question of interest

Our Question of Interest is to find out if any type of area is underrepresented or over-represented in a randomly selected sample of size 500.

## 10.2. Conditions of Chi-Squared test

- Single categorical variable with more than 2 categories - Yes, area type with 6 categories

- The expected count of each count is at least 5 - Yes

## 10.3. Parameters of Interest

We are in interested in the true $p_b, p_c, p_t, p_u, p_v, p_{ci}$ (Proportion of boroughs,cities,towns,urban,villages,cdp)

## 10.4. Hypothesis

$$H_0 : p_b = p_c = p_t = p_u = p_v = p_{ci} = 0$$

The proportion of each type of area is the same and is equal to 0.16

$$H_0 : p_i \neq 0{,}16$$

At least one of the proportions is not equal to 0.16.

This test does not tell us which proportion is not equal to 0.16. All this test tells us is that at least one of the proportions is significantly different than 0.16. There are 500 observations in our sample. If each of the solution choices had the same frequency, then each type of area frequency would have a count of $500 * ,16 = 80$ In other words under the null hypothesis, the expected count $np_i = 80$

## 10.5. Sample-Statistics

In this example we have 6 sample statistics $\hat{p_b}, \hat{p_c}, \hat{p_t}, \hat{p_u}, \hat{p_v}, \hat{p_{ci}}$

## 10.6. Test Statistic and Distribution

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E)^2}{E} \sim X^2_{k=1}$$

## 10.7. P Value

The calculated P-value is 7.928656e-115

## 10.8. Inference

Our P-value is almost 0 and thereby it provides strong evidence that the proportion of any variable is over-represented in the sample thereby we can reject our null hypothesis that each of the types of area is represented in an equal proportion of 0.16

# 11. Summary and Implications

On performing in-depth statistical analysis of our data we were able to gain valuable insights into US Demographics and economic characteristics and which provided us with pieces of evidence that supported our questions of Interest. While we were able to reject some of our Null Hypothesis, we also had to accept a few. We can further gain some comprehensive analysis by going in-depth into the data and work on further features and analyze data using Exploratory Factor Analysis and Incredible Plots, Correlation, heatmaps Linear Regression. Variables we can explore:

- Second Mortgage and Home Equity Debt Statistics
- Household and Family Income
- Population Demographics
- Home Owner and Mortgage Costs
- Gross Rent and Graduation Rates

## 11.1. Limitations and Short-Comings

While our data was collected was retrieved from 2012-2016 ACS 5-Year Documentation was provided by the U.S. Census Reports. Retrieved May 2, 2018, from Census estimate, error, and location data and Census location information While the data may be classified as historical data there might be some inconsistencies in the data collection from the source. Some form of selection bias might have crept in while the collection of the data. While maximum care was taken to consider randomization while sampling our data in the chi-squared test, there might have been already selection or sampling bias where some form of the categorical variable considered might be under or over-represented in the sample thereby causing a bias in our statistical analysis. Furthermore, we need to perform analysis on a realtime data rather than a historical data as that would give us the most up to date results for our statistical analysis

# 12. Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
if (!require("pacman")) install.packages("pacman")
pacman::p_load(tidyverse, skimr, GGally, plotly, viridis, caret, randomForest, e1071, rp
               xgboost, h2o, corrplot, rpart.plot, corrgram, ggplot2, highcharter,
```

```
              ggthemes, psych, scales, treemap, treemapify, repr, cowplot, magrittr, gg
              RColorBrewer, plotrix, ggrepel, tidyverse, gridExtra.)
mydata<-read.csv(file = "c:/real_estate_db.csv")
mydata <- subset(mydata, hc_mortgage_mean != "NaN" & !(is.na(hc_mortgage_mean)),)
mydata <- subset(mydata, rent_mean != "NaN" & !(is.na(rent_mean)),)
mydata <- subset(mydata, state != "NaN" & !(is.na(state)),)
mydata <- subset(mydata, male_pop != "NaN" & !(is.na(male_pop)),)
mydata <- subset(mydata, female_pop != "NaN" & !(is.na(female_pop)),)
mydata <- subset(mydata, hs_degree_male != "NaN" & !(is.na(hs_degree_male)),)
mydata <- subset(mydata, hs_degree_female != "NaN" & !(is.na(hs_degree_female)),)
mydata <- subset(mydata, type != "NaN" & !(is.na(type)),)
testprop<-sample(mydata,size = 300, replace = TRUE)
testofproportions <- subset(mydata, type != "NaN" & !(is.na(type)),)
testofproportions <- subset(mydata, pop != "NaN" & !(is.na(pop)),)
testofproportions <- subset(mydata, city != "NaN" & !(is.na(city)),)
head(mydata)
summary(mydata)
numerics <- select_if(mydata, is.numeric)
colnames(numerics)
# What is the distribution of the family mean

options(repr.plot.width=8, repr.plot.height=7)

# numerics %>% filter(!is.na(family_mean)) %>%
#   summarize(mean=mean(family_mean), sd=sd(family_mean))

subset.rent <- numerics %>% filter(!is.na(rent_mean))

p1 <- ggplot(data=subset.rent, aes(x=rent_mean))+
  geom_histogram(aes(y=..density..), bins = 40, fill="#81F781")+
  stat_function(fun=dnorm, color="black",
                args=list(mean=mean(subset.rent$rent_mean),
                          sd=sd(subset.rent$rent_mean))) + theme_minimal() +
theme(plot.title=element_text(hjust=0.5)) + labs(title="Rent Mean Distribution",
                                                 x="Rent Mean", y="Probability")
p1

subset.female <- numerics %>%
  filter(!is.na(female_pop))


p2 <- ggplot(data=subset.female, aes(x=female_pop))+
```

```
  geom_histogram(aes(y=..density..), bins = 40, fill="#FAAC58")+
  stat_function(fun=dnorm, color="black",
                args=list(mean=mean(subset.female$female_pop[subset.female$STATEID == "2
                          sd=sd(subset.female$female_pop[subset.female$STATEID == "2"]))
theme(plot.title=element_text(hjust=0.5)) + labs(title="Normal Distribution",
                                                 x="Female Population", y="Probability")
p2
ggplot(mydata, aes(mydata$pop, mydata$male_pop)) + geom_point(shape = 7) + geom_smooth(m
subset.debt <- numerics %>%
  filter(!is.na(debt))


p3 <- ggplot(data=subset.debt, aes(x=debt))+
  geom_histogram(aes(y=..density..), bins = 40, fill="#FA5858")+
  stat_function(fun=dnorm, color="black",
                args=list(mean=mean(subset.debt$debt),
                          sd=sd(subset.debt$debt))) + theme_minimal() +
theme(plot.title=element_text(hjust=0.5)) + labs(title="Left Skewed Distribution",
                                                    x="Debt", y="Probability")
plot_grid(p1, p2, p3, align='h', nrow=3)
cols <- numerics %>% select(debt, rent_mean, female_age_mean) %>%
filter(!is.na(debt), !is.na(rent_mean), !is.na(female_age_mean))

do.call(cbind, lapply(cols, summary))
options(repr.plot.width=8, repr.plot.height=4)
# windows(height = 7, width = 3.5)
# Lines: Mean is the blue line and Median the green line

# First Subplot
p4 <- hist(subset.rent$rent_mean, col="LightBlue", xlab="Rent", main="Distribution of Re
abline(v = mean(subset.rent$rent_mean), col = "blue", lwd = 2, lty="dashed")
abline(v = median(subset.rent$rent_mean), col = "Red", lwd = 2, lty="dashed")
legend(x = c(4000, 3200), y = c(8000, 3600), legend=c("Mean", "Median"), col=c("blue","r
      lty="dashed", lwd=1, y.intersp = 3.9, x.intersp=3.8, xjust=-1.8)
four_states <- mydata %>% select(state, rent_mean) %>% filter(!is.na(rent_mean)) %>%
filter(state == "New York" | state == "California"| state == "Florida" | state == "Texas
group_by(state) %>% do(sample_n(., size=250))

ggplot(four_states, aes(x=state, y=rent_mean, fill=state)) + geom_boxplot() +
stat_summary(fun.y=mean, colour="orange", geom="point", size=1) +
theme_minimal() + theme(plot.title=element_text(hjust=0.5, size=12)) +
labs(title="Difference in Independent Categorical Means \n (Sample Size 250)", x="States
```

```
scale_fill_brewer(palette="Set3")
# Third Subplot
p6 <- hist(subset.debt$debt, col="#F78181", xlab="Debt", main="Distribution of Debt")
abline(v = mean(subset.debt$debt), col = "blue", lwd = 2, lty="dashed")
abline(v = median(subset.debt$debt), col = "green", lwd = 2, lty="dashed")
legend(x = c(0.85, 1), y = c(5000, 3500), legend=c("Mean", "Median"), col=c("blue","gree
      lty="dashed", lwd=1, y.intersp = 2, x.intersp=0.7, xjust=0.5)
ggplot(mydata, aes(type, ..count..)) + geom_bar(aes(fill = pop), position = "dodge")
# Use boxplots to explain better the concepts of quartiles

# We will use type of place
t.place <- mydata %>% select(rent_mean, type) %>%
filter(!is.na(rent_mean), !is.na(type)) %>%
ggplot(aes(x=type, y=rent_mean)) + geom_boxplot(fill="white", colour="black",
                                                outlier.colour = "red", outlier.shape =
theme_minimal() + theme(plot.title=element_text(hjust=0.5)) + coord_flip() +
labs(title="Distribution of Average Rent by Type of Place", x="Type", y="Average Rent")

t.place + scale_fill_manual(values=c("#999999", "#E69F00"))
hist(mydata$female_pop[mydata$state == "Alaska"], col="blue", main = 'Distribution of Fe

qqnorm(mydata$female_pop[mydata$state == "Alaska"], col="blue")
qqline(mydata$female_pop[mydata$state == "Alaska"], col="red")

plot(mydata$female_pop[mydata$state == "Alaska"], type="l" , ylab = "Female population")
set.seed(1234)
t.test(mydata$female_pop[mydata$state == "Alaska"], mu = 1800)
x_bar <- mean(mydata$female_pop[mydata$state == "Alaska"])
# null hypothesized population mean
mu_0 <- 1800
# sample st. dev
s <- sd(mydata$female_pop[mydata$state == "Alaska"])
# sample size
n <- length(mydata$female_pop[mydata$state == "Alaska"])
# t-test test statistic
t <- (x_bar - mu_0)/(s/sqrt(n))
# two-sided p-value so multiply by 2
one_sided_t_pval <- pt(q = t, df = n-1, lower.tail = F)*2
one_sided_t_pval
# lower bound
x_bar+(qt(0.025, n-1)*(s/sqrt(n))) # alternately you can use x_bar-(qt(0.975, n-1)*(s/sq
# upper bound
```

```
x_bar+(qt(0.975, n-1)*(s/sqrt(n))) # alternately you can use x_bar-(qt(0.025, n-1)*(s/sq
plot(x = seq(-4, 4, length = 100), dnorm(seq(-4, 4, length = 100)),type = 'l')
abline(v = qt(0.975, n-1),col = "Red")
abline(v = qt(0.025, n-1), col = "Blue")

num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(sample(x = mydata$rent_mean[mydata$state == "Alaska"],
 size = n,
 replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Ave
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(1100, 1300, .1), dnorm(seq(1100, 1300, .1), mean = x_bar, sd(results)))
# Shifting the sample so that the null hypothesis is true
time_given_H0_true <- mydata$rent_mean[mydata$state == "Alaska"] - mean(mydata$rent_mean
# This data is pretty skewed so even though n is large, We are going to perform a lot of
num_sims <- 10000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims)
{
 results_given_H0_true[i] <- mean(sample(x = time_given_H0_true,
 size = n,
 replace = TRUE))
}
# Finally plot the results
hist(results_given_H0_true, freq = FALSE, main='Sampling Distribution of the Sample Meas
n, Given Null Hypothesis is True', xlab = 'Average Rent', ylab = 'Density', xlim = c(140
# adding line to show values more extreme on upper end
abline(v=x_bar , col = "red",)
# adding line to show values more extreme on lower end
low_end_extreme <- mean(results_given_H0_true)+(mean(results_given_H0_true)-x_bar)
abline(v=low_end_extreme, col="red")

low_end_extreme
high_end_extreme <- mean(results_given_H0_true) + x_bar
```

```
high_end_extreme
abline(v=x_bar, col = "red")
abline(v=low_end_extreme, col="blue")
# counts of values more extreme than the test statistic in our original sample, given H0
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= low_end_extreme)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= high_end_extreme)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail
bootstrap_pvalue
# need the standard error which is the standard deviation of the results
bootstrap_SE_X_bar <- sd(results)
# an estimate is to use the formula statistic +/- 2*SE
c(x_bar - 2*bootstrap_SE_X_bar, x_bar + 2*bootstrap_SE_X_bar)

# you can also use the 5th and 97.5th quantiles to determine the bounds:
c(quantile(results, c(.025, .975)))
# compare to our t-methods
c(x_bar+(qt(0.025, n-1)*(s/sqrt(n))), x_bar+(qt(0.975, n-1)*(s/sqrt(n))))
x <- sum(testofproportions$pop[testofproportions$city == 'Anchorage'])
n <- sum(testofproportions$pop[testofproportions$state  =='Alaska'])
p_cap <- x/n
p_cap
p_0 = 0.3333
z <- (p_cap - p_0)/sqrt((.3333*(1-.3333))/n)
p_val <- pnorm(z, lower.tail = F)
p_val
binom.test(x=x, n = n, p=(1/3), alternative="greater")
binom.test(x=x, n = n, p=(1/3), alternative="greater")$conf.int
binom.test(x=x, n = n, p=(1/3), alternative="two.sided")

Anchorage <- rep(c(1,0), c(33, 100-33))
Anchorage

table(Anchorage)
# This data is pretty skewed so even though n is large, We are going to do a lot of simu
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(sample(x = Anchorage,
 size = 100,
```

```
 replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Proportion', xlab
 = 'Proportion of Anchorage', ylab = 'Density')
# estimating a normal curve over it - this looks pretty good!
lines(x = seq(.1, .75, .001), dnorm(seq(.1, .75, .001), mean = mean(results), sd = sd(re
c(quantile(results, c(.05, 1)))
cat("exact binomial test")
binom.test(x=33, n = 100, p=(1/3), alternative="greater")$conf.int
# Under the assumption that the null hypothesis is true, we have 33% population in Ancho
Anchorage<- rep(c(1, 0), c(33, 100-33))
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(sample(x = Anchorage,
 size = 100,
 replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Proportion under H
_0:p=0.3', xlab = 'Proportion of People in Anchorage', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(.15, .75, .001), dnorm(seq(.15, .75, .001), mean = mean(results), sd = sd
(results)))
abline(v=.33, col="red")
count_of_more_extreme_upper_tail <- sum(results >= .33)
bootstrap_pvalue <- count_of_more_extreme_upper_tail/num_sims
cat("Bootstrap p-value")
bootstrap_pvalue
binom.test(x=33, n = 100, p=(1/3), alternative="greater")$p.value
c(quantile(results, c(.05, 1)))
set.seed(0)
options(repr.plot.width=8, repr.plot.height=5)


south_states <- mydata %>% select(state, rent_mean) %>% filter(state == "Delaware" | sta
ggplot(aes(x=state, y=rent_mean, fill=state)) + geom_boxplot() +
stat_summary(fun.y=mean, colour="orange", geom="point", size=1) +
theme_minimal() +
```

```
theme(plot.title=element_text(hjust=0.5, size=10)) +
labs(title="Difference Between Two Independent Means", y="Rent Mean", x="States") +
scale_fill_brewer(palette="Set3")

south_states
qqnorm(mydata$rent_mean[mydata$state == "Delaware"],col = 'blue')
qqline(mydata$rent_mean[mydata$state == "Delaware"] ,col = 'red')
qqnorm(mydata$rent_mean[mydata$state == "Alaska"],col = 'blue')
qqline(mydata$rent_mean[mydata$state == "Alaska"],col = 'red')
t.test(mydata$rent_mean[mydata$state == "Delaware"], mydata$rent_mean[mydata$state == "A
x1_bar <- mean(mydata$rent_mean[mydata$state == "Delaware"])
x2_bar <- mean(mydata$rent_mean[mydata$state == "Alaska"])
n<-min(length(mydata$rent_mean[mydata$state == "Delaware"]),mydata$rent_mean[mydata$stat
se <- sd(mydata$rent_mean[mydata$state == "Delaware"])**2/length(mydata$rent_mean[mydata

t <- (x1_bar - x2_bar) / sqrt(se)

two_sided_t_pval <- pt(t, df=n-1, lower.tail = T)*2
two_sided_t_pval
c((x1_bar - x2_bar)+qt(0.025,df=n-1)*sqrt(se),(x1_bar - x2_bar)+qt(0.975,df=n-1)*sqrt(se
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 mean_dela <- mean(sample(x = mydata$rent_mean[mydata$state == "Delaware"],
 size = 34,
 replace = TRUE))
 mean_alas <- mean(sample(x = mydata$rent_mean[mydata$state == "Alaska"],
 size = 34,
 replace = TRUE))
 results[i] <-   mean_dela - mean_alas
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Ave
rage Difference in Rent', ylab = 'Density')
c(quantile(results, c(.025, .975)))
t.test(mydata$rent_mean[mydata$state == "Delaware"], mydata$rent_mean[mydata$state == "A
transform(mydata, state=sample(state))
num_sims <- 1000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
```

```
# A loop for completing the simulation
for(i in 1:num_sims){
 # idea here is if there is no relationshipm we should be able to shuffle the groups
 shuffled_groups <- transform(mydata, state=sample(state))
 mean_delaware <- mean(shuffled_groups$rent_mean[shuffled_groups$state=="Delaware"])
  mean_alaska <- mean(shuffled_groups$rent_mean[shuffled_groups$state=="Alaska"])
 results_given_H0_true[i] <-mean_delaware - mean_alaska
}

# Finally plot the results
hist(results_given_H0_true, freq = FALSE,
 main='Dist. of the Diff in Sample Means Under Null',
 xlab = 'Average Difference Rent  under Null',
 ylab = 'Density')

diff_in_sample_means <- mean(mydata$rent_mean[mydata$state == "Delaware"]) - mean(mydata
diff_in_sample_means
lower_extreme <-
abline(v=diff_in_sample_means, col = "blue")
abline(v=abs(diff_in_sample_means), col = "red")

count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= diff_in_sample_means)
count_of_more_extreme_lower_tail
count_of_more_extreme_upper_tail <- sum(results_given_H0_true > abs(diff_in_sample_means
count_of_more_extreme_upper_tail
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail
cat("Bootstrap p-value")

bootstrap_pvalue
t.test(mydata$rent_mean[mydata$state == "Delaware"], mydata$rent_mean[mydata$state == "A
c((x1_bar - x2_bar)+qt(0.025,df=n-1)*sqrt(se),(x1_bar - x2_bar)+qt(0.975,df=n-1)*sqrt(se
# the parts of the test statistic
# sample props
p_hat_c <- sum(mydata$female_pop[mydata$city == 'Anchorage'])/sum(mydata$female_pop[myda
p_hat_t <- sum(mydata$male_pop[mydata$city =='Anchorage'])/sum(mydata$male_pop[mydata$st
p_hat_c
p_hat_t
# null hypothesized population prop difference between the two groups
p_0 <- 0
# sample size
n_c <- sum(mydata$female_pop[mydata$state == 'Alaska'])
n_t <- sum(mydata$male_pop[mydata$state == 'Alaska'])
```

```
# sample variances
den_p_c <- (p_hat_c*(1-p_hat_c))/n_c
den_p_t <- (p_hat_t*(1-p_hat_t))/n_t
# z-test test statistic
z <- (p_hat_c - p_hat_t - p_0)/sqrt(den_p_c + den_p_t)
z
# two sided p-value
two_sided_diff_prop_pval <- pnorm(q = z, lower.tail = FALSE)*2
two_sided_diff_prop_pval
# lower bound
(p_hat_c - p_hat_t)+(qnorm(0.025)*sqrt(den_p_c + den_p_t))
# upper bound
(p_hat_c - p_hat_t)+(qnorm(0.975)*sqrt(den_p_c + den_p_t))
# Make the data
females <- rep(c(1,0), c(sum(mydata$female_pop[mydata$city == 'Anchorage']), n_c - sum(m
males <- rep(c(1,0), c(sum(mydata$male_pop[mydata$city == 'Anchorage']),
                       n_t- sum(mydata$male_pop[mydata$city == 'Anchorage'])))

num_sims <- 1000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 prop_female <- mean(sample(females,
 size = n_c,
 replace = TRUE))
 prop_male <- mean(sample(x = males,
 size = n_t,
 replace = TRUE))
 results[i] <- prop_female - prop_male
}
# Finally plot the results
hist(results, freq = FALSE, main='Dist. of the Diff in Prop', xlab = 'Difference in Pro
p. of Free Throws Made', ylab = 'Density')

c(quantile(results, c(.025, .975)))
c((p_hat_c - p_hat_t)+(qnorm(0.025)*sqrt(den_p_c + den_p_t)), (p_hat_c- p_hat_t)+(qnorm
(0.975)*sqrt(den_p_c + den_p_t)))
# Make the data
df_combined <- data.frame("population_proportion" = c(females, males),
 "Anchorage" = rep(c("females", "males"), c(n_c, n_t)))
num_sims <- 1000
```

```
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 # idea here is if there is no relationshipm we should be able to shuffle the groups
 shuffled_groups <- transform(df_combined, Anchorage=sample(Anchorage))
 prop_females <- mean(shuffled_groups$population_proportion[shuffled_groups$Anchorage=="
 prop_males <- mean(shuffled_groups$population_proportion[shuffled_groups$Anchorage=="ma
 results_given_H0_true[i] <- prop_females - prop_males
}
results_given_H0_true
# Finally plot the results
hist(results_given_H0_true, freq = FALSE,
 main='Dist. of the Diff in Sample Sample Props Under Null',
 xlab = 'Average Difference in Prop. of females and males Made under Null',
 ylab = 'Density',xlim = c(-0.010,0.010) )
diff_in_sample_props <- p_hat_c - p_hat_t
abline(v=diff_in_sample_props, col = "blue")
abline(v=-diff_in_sample_props, col = "red")
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= -diff_in_sample_props)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true > diff_in_sample_props)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail
cat("Bootstrap p-value")
bootstrap_pvalue
two_sided_diff_prop_pval
 c(quantile(results,c(0.025, 0.975)))
set.seed(315)
newdata<-sample(mydata$type, size = 500)
newdata
sum(((table(newdata) - 80)^2)/80)
pchisq(541.7, df = 6-1, lower.tail = FALSE
```