

*Toronto*

**Chi-square and ANOVA**

Soni Manana

a *College of Professional Studies****,*** *Master of Professional Studies in Analytics.*

***Subject****: ALY6015* ***NUID****: 002982645*

Under the guidance of

**Dr. Prof. Alex Maizlish**

**Introduction**

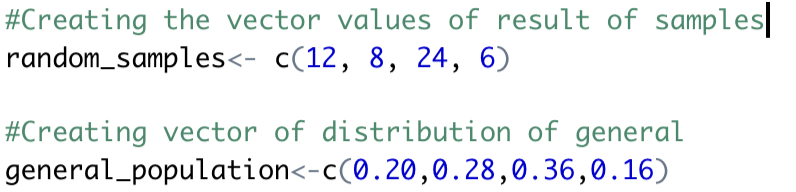
We have given different problems which we will solving using stating hypothesis and using Chi-Square and ANOVA tests.

**What are Chi-Square and ANOVA tests?**

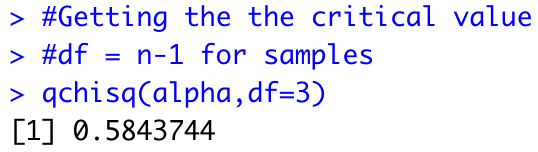
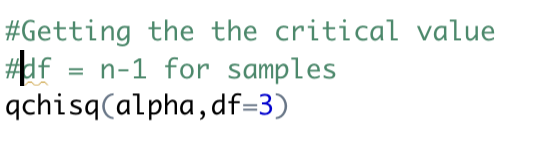
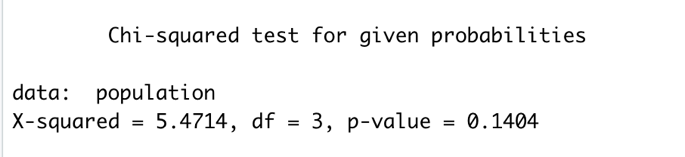
* A Chi-square test is a hypothesis testing. Chi-square tests involve checking if observed frequencies in one or more categories match expected frequencies.
* **ANOVA** means **An**alysis **O**f **Va**riance, It’s same as T-test but calculates for two or more. It compares the effect of independent variable on multi dependent variable. You can understand what is affecting between the variables by ANOVA.

**Problem 11-1.1**

* As we have given distribution of blood groups for the general populations which are type A, 20%; type B, 28%; type O, 36%; and type AB = 16% and Random samples of 50 to test and they found 12 have type A blood, 8 have type B, 24 have type O, and 6 have type AB blood.  
  They want to test if we can conclude that distribution of samples are same as population at 90% level of significance.
* **H0**: Sample Distribution is same as the general population.
* **H1:** Distributions aren’t same.
* We created two vectors as populations and samples

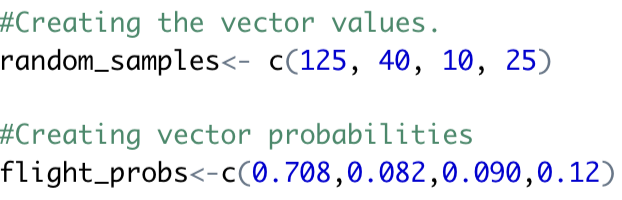
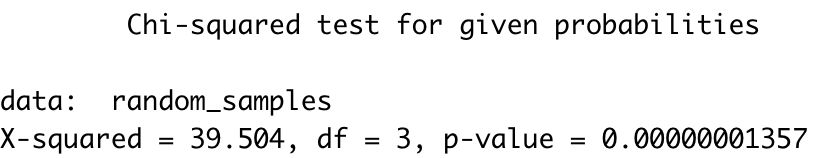


* Got the p-value from Chi-squared test as it’s a traditional hypothesis test just like t-test. Difference is just t-test calculated from means. P-value is 0.14 which is higher than 0.584 (Critical value).
* Hence, we can’t reject **H0.** As we don’t have enough evidence to reject.
* Low value of x-squared means correlations is high. They are inversely proportional to each other.



**Problem 11-1.2**

* We have data of flight of an airline and their regularity provided by government like if a flight lands on airport, there can be chances of status of flight like 70.8% on-time, 8.2% delay because of National Aviation System (NAS) delay, 9% because of late arrival at the airport and 12% due to other reasons like weather or maintenance.
* 200 random samples were taken in which 125 of them were on time, 40 were delayed because of weather, 10 because of NAS delay.
* **H0**: The given probabilities of flights by government and results of random samples are same.
* **H1:** Results and probabilities aren’t same.
* We need to find for 95% level of significance so value of alpha is 0.05.
* Created two vectors with values.

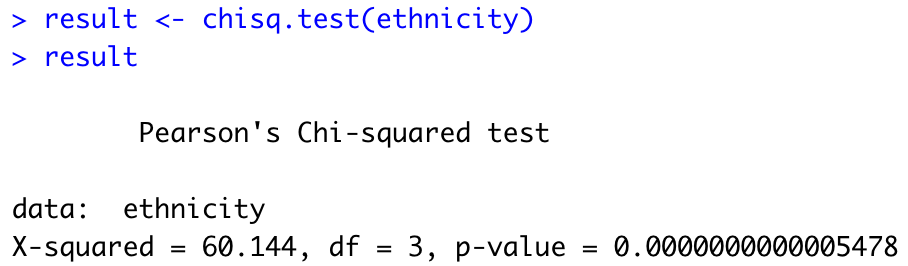
** **

* We can see that value of p is just 0 which is less than alpha. Therefore, we reject H0 the null hypothesis.
* Critical value for this test is 0.35.

****

**Problem 11-2.1**

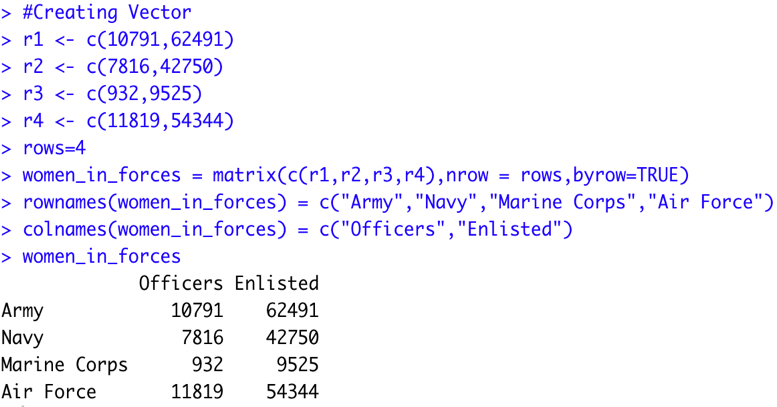
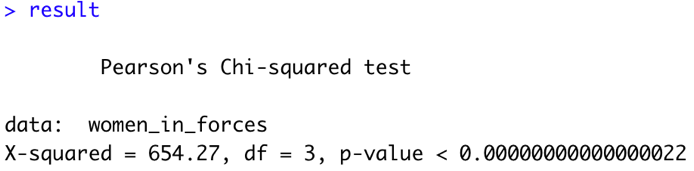
* Table of ethnicity is provided for 2013 and 2014 for movie admissions. Data is in thousands.
* So created a matrix in R named ethnicity.
* From this,
* **H0**: Admissions are not depended upon ethnicity.
* **H1**: They are depending upon ethnicity.

** **

* As we can see from the test p-value is 0 with higher X-squared value and Critical value is 7.814. Therefore, P-value is less than alpha which is 0.05 and X2 value is just 60.144 which shows they hardly have correlation.
* As we can notice number of variables are increasing.

**Problem 11-2.2**

* Table of women in military is provided for different Arm forces like Army and Navy, and ranks are provided who fall in those categories. In nutshell, 4 categories and 2 columns have provided.
* First needs to create a matrix. Then provide row names, column names.
* **H0**: Ranks of women officers are independent from different arm forces.
* **H1:** There is some relationship between rank and branch of forces.
* We need to check that at 95% level of significance or 0.05 (alpha).

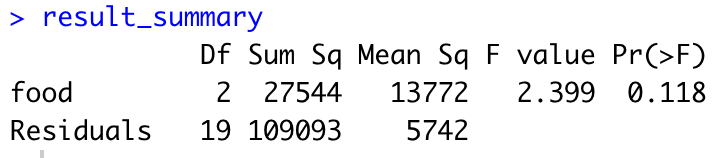
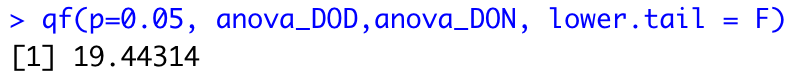
 ****

* As we can notice X-squared value is just way high and p-value is also 0. In that case, we reject the null hypothesis and we found there is a relationship exists.

**Problem 12-1**

| **Condiments** | **Cereals** | **Desserts** |
| --- | --- | --- |
| 270 | 260 | 100 |
| 130 | 220 | 180 |
| 230 | 290 | 250 |
| 180 | 290 | 250 |
| 80 | 200 | 300 |
| 70 | 320 | 360 |
| 200 | 140 | 300 |
|  |  | 160 |

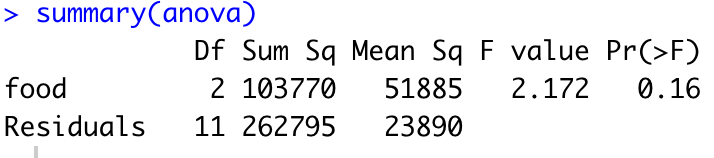
* A table has been provided in which amount of sodium in one serving for 3 different foods like Condiments, Cereals and Desserts are demonstrated.
* **H0**: There is mean sodium difference in three of them.
* **H1:** There isn’t or at least one of them is different.
* We need to find at 0.05 level of significance.
* We will use **ANOVA** for this problem as problem have numerical as well as categorical. There is thumb rule. If every variable is categorical, we can’t use **ANOVA** but **Chi-squared**.
* Actual way is to calculate values like K, N-K, etc. but R provides direct calculation for that.

** **

* DF of food is K which is 2, 19 is N-K, F-value is 2.399 and p-value is 0.118.
* Comparing p-value with level of significance, we failed to found enough evidence to reject the null hypothesis (**H0**).
* The variation between the sodium means is not high enough to the variation of within the samples. F-values is higher.

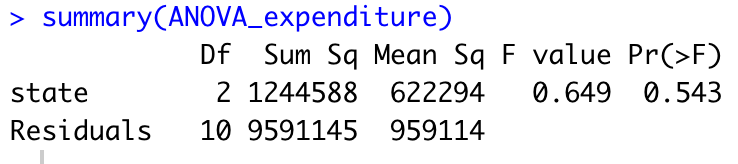
**Problem 12-2.1**

* A table of the sales (in millions of dollars) for a year of a sample of leading companies are provided. At the level of 99% significance to tell if there is difference in means.
* **H0:** Mean difference in sales is same.
* **H1:** Mean difference in sales is not same.
* F-value is 2.172 which is high and P-value is 0.16 which greater than 0.01. In both cases, we cannot reject the null hypothesis.



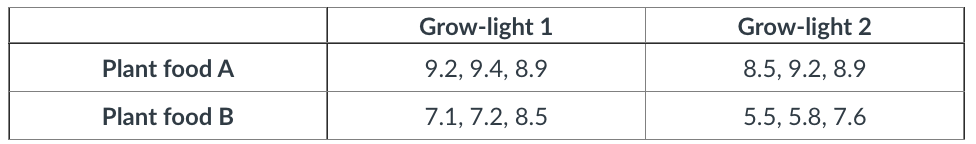
**Problem 12-2.2**

* A table of expenditures (in dollars) per pupils for United States in three sections of country has been provided. A level of significance also provided which is 0.05.
* We need to use ANOVA as there is not only categorical variables. In addition, We need to form null hypothesis in ANOVA is always that there is no difference in means. So
* **H0:** Mean difference in expenditures per pupil is same in all of the sections. (μ1 = μ2 = μ3.)
* **H1:** Mean difference in expenditures per pupil is not same in all of the sections.
* P-value is 0.543 which greater than 0.05. In both cases, we cannot reject the null hypothesis.
* Here F value is good enough where p-value is higher than alpha.

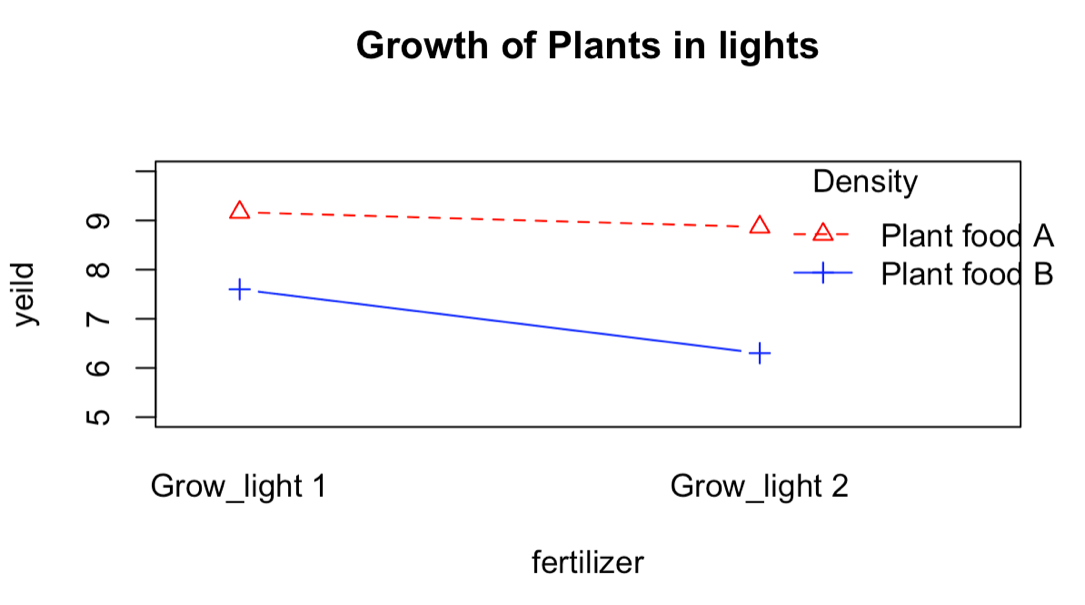


**Problem 12-3**

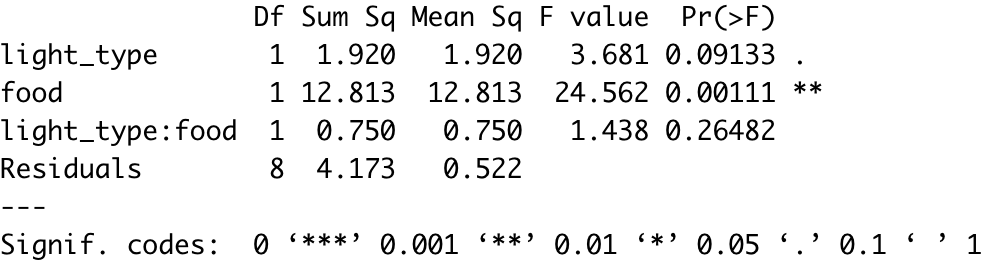
* A table of experiments on plant has given. It conducted that one plant can grow in two different strengths of grow-light and plant food supplement (Fertilizers) by a gardening company.



* From this plot we can notice that in Grow-Light 1, growth of plant has different effects of fertilizers growth can be seen 9 where for plant food B it started at 7.5 around to add to that at Grow-Light 2 the growth numbers decrease more. So according to plot, Grow-Light 1 and Plant food A is good for growth of the plant.



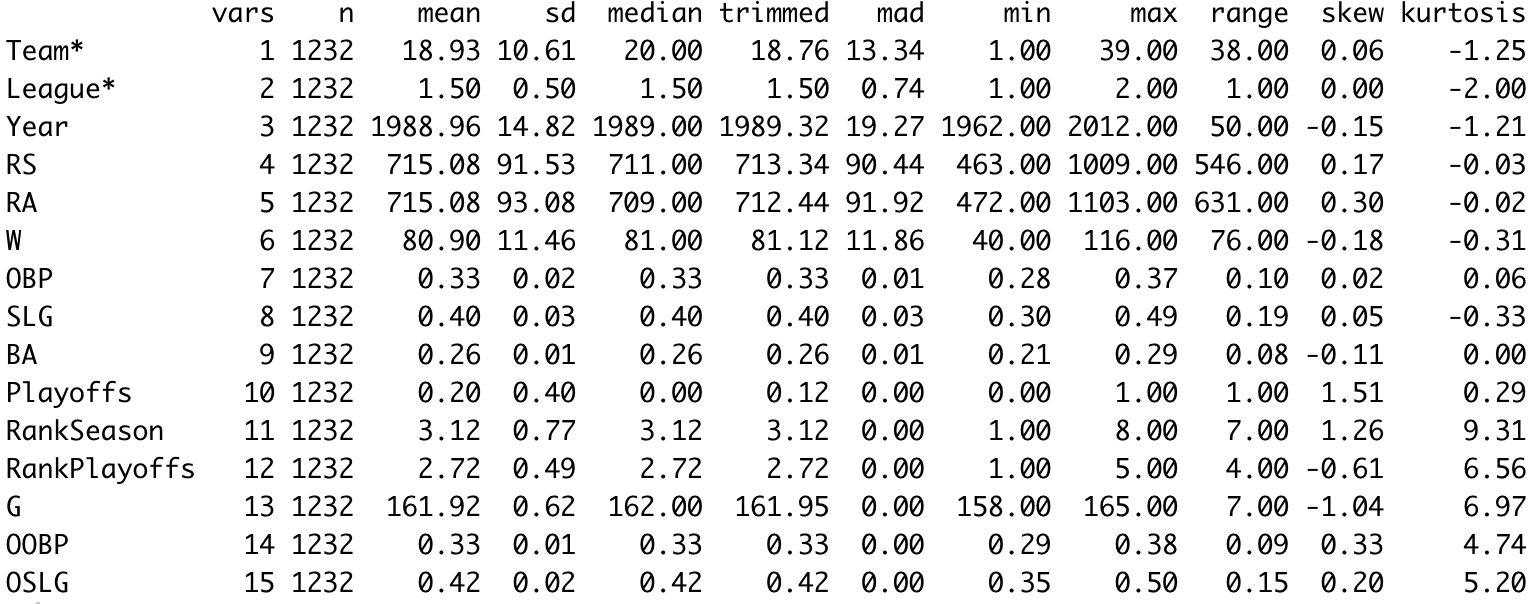
* We can see that there are individual p-values and one interaction of light type and food which is 0.264 and less than 0.05 and hence we failed to reject null hypothesis.
* Lowest value of F also notice for the interaction of light and plant food.

****

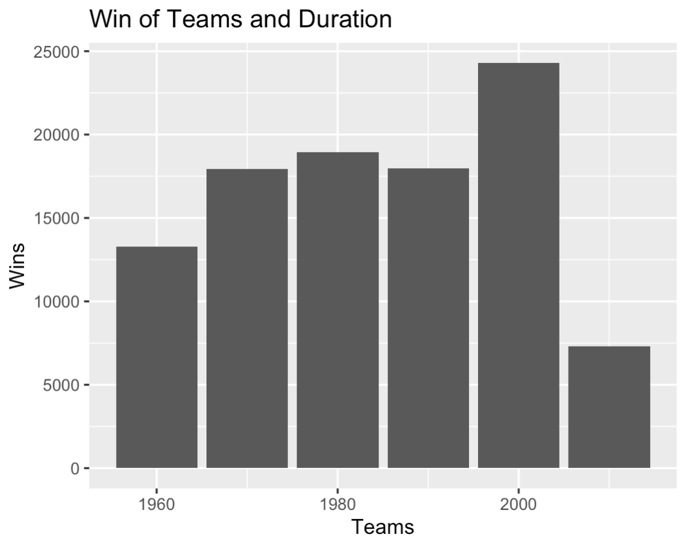
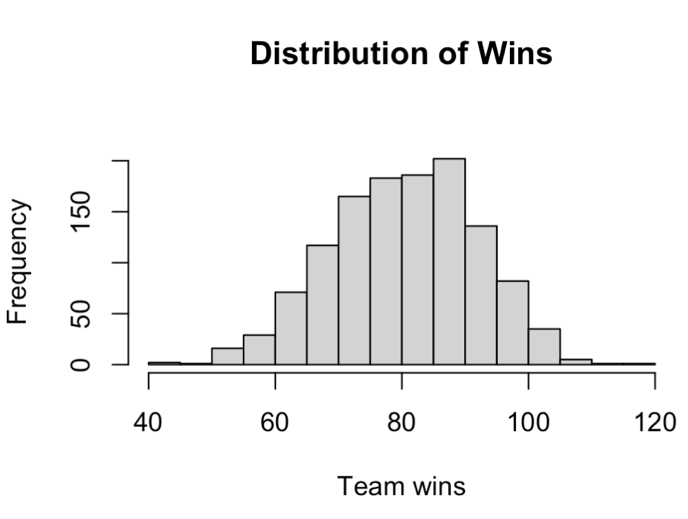
**Baseball Data Analysis**

**Descriptive Analysis**

* Initially, Data was uncleaned, many has 800 or above samples had null values so they were replace by mean of respective variables.
* Dataset has 1232 samples and 15 columns. Two columns such as League name and team name were character columns and others were numerical.
* As we can see from the skewness all of the columns are almost normally distributed except Playoffs, RankSeason, RankPlayoffs.
* Time duration varies from 1962 to 2012.
* As we can see mean of team win is 80 and has standard deviation of 11.46.

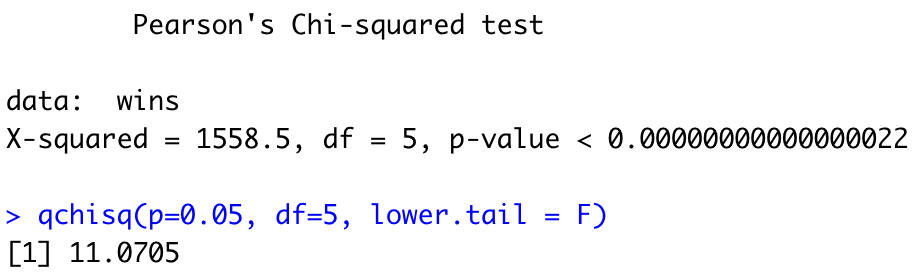
****

* Number of games has to be increased from overall time duration so win is also increased. Low bar is of 2002-2012 because of interleague games were introduced to baseball game which leads to less samples in row thus bar scored less.

****

**Hypothesis of Wins of Team and Decade**

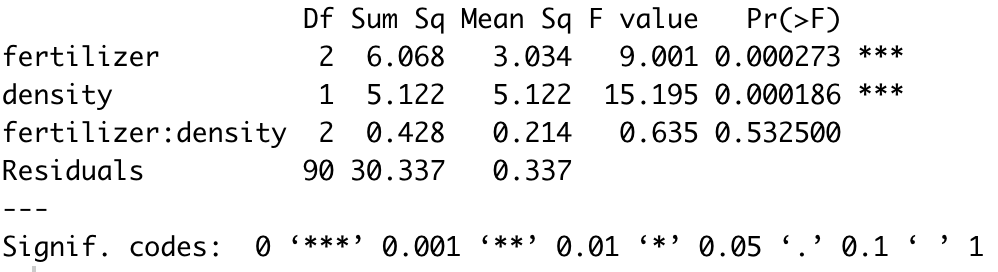
* From numbers of win by decades. Check is required If there is some relation exists between two of them at 95% of level of significance.
* Therefore, **H0**: Numbers of Wins has relation with decades or time.
* **H1**: Number of Wins has no relation with decade or time.

****

* As we can see value of p is 0 which is less than 0.05 and statistics is 1558.5 which is clearly greater than 11.07.
* In both scenario, we reject the null hypothesis.

**Hypothesis of Crop**

* Data of crop is given and it has 4 variables named Block, Fertilizer, Yield and Density of crop.
* We need to find the evidence for if fertilizer and density have an impact on yield
* We need to check for normality, equality of variance and Independence for Two-way ANOVA.
* H0: There is no impact on yield from usage of fertilizer, density and it’s interaction.
* H1: There is impact on yield from usage of fertilizer, density and it’s interaction.
* As from the P values the interaction has value 0.532. So we can’t reject the null hypothesis.
* **So From this, we can say for fertilizer and density we reject the null hypothesis But for the interaction of both we can’t.**

****

**References**

Statology, *2022. The Chi-Square distributions in R Source:* https://www.statology.org/chi-square-distribution-in-r-dchisq-pchisq-qchisq-rchisq/

Wikipedia, *2022. Interleague play Source:* <https://en.wikipedia.org/wiki/Interleague_play>

Sphweb, Lisa Sullivan, PhD, 2022. Hypothesis Testing – Analysis of Variance (ANOVA) *Source:* https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\_hypothesistesting-anova/bs704\_hypothesistesting-anova\_print.html#:~:text=The%20null%20hypothesis%20in%20ANOVA,rather%20than%20in%20mathematical%20symbols.

**Appendix**

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

#1

mData = read.csv('./AmesHousing.csv')

str(mData)

# 2

# getting rows and cols count

length(mData) # columns

nrow(mData) # rows

library(dplyr)

nData = mData %>% select(where(is.numeric))

for(i in 1:ncol(nData)){

nData[is.na(nData[,i]), i] <- mean(nData[,i], na.rm = TRUE)

}

psych::describe(nData[c('SalePrice','X1st.Flr.SF','Garage.Cars','Garage.Area','Gr.Liv.Area','Overall.Qual','Overall.Cond')])

options(scipen=999)

hist(nData$SalePrice,xlab='House Sale Price',main='Ames Iowa House Prices',col='red',border='white')

hist(nData$Overall.Qual,xlab='Overall Quality',main='Quality of Houses of Ames',col='cyan4',border='white')

cor(nData['SalePrice'],nData[1:38])

library(ggcorrplot)

ggcorrplot(cor(nData), insig = "blank", lab\_size=2,type='lower',tl.cex = 8)

ggcorrplot(cor(nData['SalePrice'],nData[1:39]), lab=TRUE,lab\_size=1.5, insig = "blank", tl.cex=8)

library(ggplot2)

options(scipen=999)

# highest

ggplot(nData, aes(x=Gr.Liv.Area, y=SalePrice)) +

geom\_point(size=2, shape=1)

# lowest

ggplot(nData, aes(x=Yr.Sold, y=SalePrice)) +

geom\_point(size=2, shape=1)

# 0.5

ggplot(nData, aes(x=Year.Built, y=SalePrice)) +

geom\_point(size=2, shape=1)

# 7

par(mfrow=c(1,2))

# a

SL\_lmodel = lm(nData$SalePrice ~ nData$Gr.Liv.Area)

plot(nData$SalePrice ~ nData$Gr.Liv.Area,xlab="Above Ground Level (Sq.Ft)",ylab="Sales Price",main='House Price ~ Sq.ft Above Ground Level')

abline(a= SL\_lmodel$coefficients[1],b=SL\_lmodel$coefficients[2])

summary(SL\_lmodel)

# b

ML\_lmodel = lm(nData$SalePrice ~ nData$Gr.Liv.Area + nData$X1st.Flr.SF + nData$Total.Bsmt.SF)

plot(nData$SalePrice ~ nData$Gr.Liv.Area,xlab="Above Ground Level (Sq.Ft)",ylab="Sales Price",main='Price ~ Sq.ft Above Ground + Basement + 1st Floor')

abline(a= ML\_lmodel$coefficients[1], b=ML\_lmodel$coefficients[2]+ML\_lmodel$coefficients[3]+ML\_lmodel$coefficients[4])

summary(ML\_lmodel)

par(mfrow=c(1,1))

plot(SL\_lmodel)

plot(ML\_lmodel)

library(car)

vif(ML\_lmodel)

# 11

outlierTest(ML\_lmodel)

par(mfrow=c(1,1))

hat.plot <- function(ML\_lmodel) {

p <- length(coefficients(ML\_lmodel))

n <- length(fitted(ML\_lmodel))

plot(hatvalues(ML\_lmodel), main="Index Plot of Hat Values")

abline(h=c(2,3)\*p/n, col="red", lty=2)

identify(1:n, hatvalues(ML\_lmodel), names(hatvalues(ML\_lmodel)))

}

hat.plot(ML\_lmodel)

# 12

cooksValue = cooks.distance(ML\_lmodel)

influential\_points = cooksValue[(cooksValue>3\*mean(cooksValue,na.rm=T))]

names = names(influential\_points)

outliers = nData[names,]

cleanedData = nData%>%anti\_join(outliers)

#qqnorm()

par(mfrow=c(1,1))

WO\_ML\_lmodel<- lm(cleanedData$SalePrice~cleanedData$Gr.Liv.Area+cleanedData$X1st.Flr.SF +cleanedData$Total.Bsmt.SF)

plot(cleanedData$SalePrice ~ cleanedData$Gr.Liv.Area)

abline(a=WO\_ML\_lmodel$coefficients[1],b=WO\_ML\_lmodel$coefficients[1]+WO\_ML\_lmodel$coefficients[2])

summary(WO\_ML\_lmodel)

plot(WO\_ML\_lmodel)

#13

library(MASS)

stepAIC(WO\_ML\_lmodel,direction = "both")

par(mfrow=c(1,1))

library(leaps)

leaps <- regsubsets(SalePrice~Gr.Liv.Area+X1st.Flr.SF +Total.Bsmt.SF + Garage.Area + Garage.Cars,

data=withoutOutliers,nbest=4)

plot(leaps,scale="adjr2",ylab='Adjusted R-Square Value')