1. **Introduction to Statistic**
   1. Library(ggplot2)
   2. Histogram: ggplot(data =ames) + geom\_histogram(mapping = aes(x = sale\_price) + labs(x = “x label”, y = “y label”)
   3. QQplot: ggplot(data = ames, aes(sample = y-var, color = group\_by) + stat\_qq() + stat\_qq\_line()
   4. Box Plot: ggplot(data = ames, aes(y= sale\_price, x= central\_Air, fill= central\_Air)) + geom\_boxplot() + labs(x= “x label”, y = “y label”)
   5. Scatter Plot: ggplot(data = train) + geom\_point(mapping = aes(x=, y=) + labs(x=, y=)
      1. Can also do + coord\_flip()
   6. Type 1 error 🡪 False positive 🡪 rejection of null hypothosis when it is true
   7. Type 2 error 🡪 False negative 🡪 failure to reject a null hypothosis that is false
   8. T-tests (1 and 2 sample T-Tests)
      1. t.test(ames$Sale\_Price, mu = 178000)
         1. the above is testing whether or not the mean sales price of a home is 178000
      2. t.test(ames$Sale\_Price, mu = 178000, alternative = ‘greater’)
         1. the above is testing the null hypothosis that the mean sales price is < 178,000.
      3. To run a 2 sample t-test, you first need to verify the following 3 conditions:
         1. Independent Observations
         2. Normality
            1. Two ways to verify normality

QQ plot

shapiro.test

null hypothesis is that the data is normally distributed

when normality fails, use the Wilcoxon Rank Test

wilcox.test(Sale\_Price ~ Central Air, data =)

* + - 1. Equal variance among groups (requires that both populations are normally distributed)
         1. Null hypothosis: equal variances
         2. var.test(MPG ~ Country, data = cars)

Equal variance -> Pooled variance T-test

Unequal Variance -> Satterthwaite’s t-test

* + - * 1. t.test(MPG ~ Country, data = cars, var.equal = FALSE)

^^use TRUE if the variances are equal

1. **Introduction to ANOVA and Regression**
   1. One way ANOVA: testing if the means between the levels are the same
      1. Ho: all means are the same
      2. Ha: at least one mean is different
      3. Assumptions:
         1. Independent observations
         2. Normality within each category
            1. QQ plot or Histogram
            2. Shapiro-Wilk

Ho: The residuals are normally distributed

Ha: The residuals are not normally distributed

* + - 1. Equal variances
         1. leveneTest🡪 data must be normal

Ho: groups being comapred have equal variances

Ha: groups being compared have unequal variances

* + - * 1. Or plot the fitted vs residuals of the model, and want to see equal spread when looking from left to right
        2. ggplot(model2, aes(x=fitted(model2), y = resid(model2))) + geom\_point()
        3. fligner.test🡪 does not need normality

Ho: groups being comapred have equal variances

Ha: groups being compared have unequal variances

* + 1. Normal and Equal variance 🡪 regular one-way ANOVA
    2. Normal and Unequal variance 🡪 Welch’s ANOVA
       1. Need approximautly normal data, but the variances between groups can be unequal
    3. Not Normal and unequal variance 🡪Kruskal-Wallis
       - 1. Group distributions are identical in shape,variance and symmetric 🡪 difference in means
         2. Group distributions are identical in shape, but not symmetric 🡪 Difference in medians
         3. Else 🡪 Difference in location
  1. One way ANOVA post hoc testing
     1. Two methods for Multiple Comparisons
     2. All pairwise comparisons 🡪 Tukey, Games-Howell, Dunn
     3. Comparisons to a control group 🡪 Dunnet, Wilcoxon(with Bonferroni)
     4. Tukeys 🡪for when you want to do pairwise comparisons for all groups
        1. TukeyHSD(model)
     5. Games-Howell 🡪 used when the model was either Welch’s or Kruskal-Wallis
        1. library(PMCMRplus)
        2. gamesHowellTest(model)
        3. seems that you must make a model for it to run 🡪 as Dr. Simmons about this
     6. Dunn 🡪 used when the model was either Welch’s or Kruskal-Wallis
        1. Dunnett’s test 🡪 used when you want to compare to a control/placebo group
  2. Correlation
     1. Ho: correlation coeficient is equal to 0
     2. Cor.test()
  3. Simple Linear Regression
     1. Normality 🡪 Histogram, QQ-plot or normality test (on residuals)
     2. Equal variances 🡪 want a nice band on the residual plot
     3. Independence 🡪 has to do with data collection
     4. Linearity in the mean 🡪 no pattern in residual plot

1. **More Complex ANOVA and Regression**
   1. Two-way ANOVA
      1. ames\_aov2 <- aov(Sale\_Price ~ Heating\_QC + Central\_Air, data = train)
      2. post-hoc testing 🡪 same as #2
      3. Ho: each variable has no effect on the Y variable
   2. With Interactions
      1. ames\_aov\_int <- aov(Sale\_Price ~ Heating\_QC\*Central\_Air, data = train)
         1. Heating\_QC\*AC\_Air = Heating + AC + Heating:AC
      2. For the above model, an interaction between Central Air and Heating Quaility would imply both of the following:
         1. The impact of Heating Quality on Sale Price differs across levels of Central Air
            1. ex: Difference in price between Excellent and Poor HQ changes if Central Air is or is not present
      3. The impact of Central Air on Sale Price differs across levels of Heating QualityModel Selection
         1. ex: Difference in price between having and not having Central Air changes across levels of HQ
   3. Multiple Linear Regression
      1. Ho: none of the variables are useful in predicting the target variable
      2. Ha: at least one of the variables is useful in predicting the target variable
      3. F–Distribution
         1. (what model is able to do)/(what model is not able to do)
         2. Bounded by 0 (right skewed)
      4. Assumptions
         1. Linear combination of X’s to make Y’s
         2. Normality of data
         3. Equal variance
         4. Errors are independent
      5. R^2 🡪 % change in Y explained by X (which is the model and the variables)
         1. Cannot be used to determine which model is best, as it will always chose the model with more variables
      6. Adjusted-R^2 🡪
         1. penalizes the model for adding extra variables
         2. Can be used to compare models
         3. However, it loses interpretability
2. **Diagnostics and Model Building**
   1. AIC/BIC used to judge models
      1. The lower the better
      2. AIC is for prediction
      3. BIC is for explanation
   2. Forward Selection
      1. Start with empty model, and add variables until adding another variable would be worse than not adding the variable
   3. Backwards Selection
      1. Start with full model, and remove variables one at a time until not removing the next variable is better than removing the next variable
   4. Stepwise Selection
      1. Start with empty model, and then add and subtract variables until the next best move is do to nothing to the model
3. **Diagnostics** 
   1. Violations of linear regression assumptions
      1. **Linearity in the paremeters**:
         1. a misspecified model 🡪therefore the results are not meaningful
         2. look at the residuals vs X values, and a pattern is bad
      2. **Constant variance**:
         1. want constant variance across all levels of X (homoscedasticity)
         2. does not affect the paramter estimates, but the standard errors are compromised
         3. look at residuals vs fitted
         4. Tested using Spearman Rank test between residuals and predicted values
            1. Close to 0 🡪 variance potentially homoscedastic
            2. Positive 🡪 variance increases as the mean increases
            3. Negative 🡪 variance decreases as the mean increases
         5. Variance stabalizing transformations
            1. Log transformation on the y variable

Interpreted as a % change in Y as X increases

* + - * 1. Weighted Least Sqaures

Possible weights include:

1/(fitted values)^2

1/(SD of X)^2 {for when multiple observations at same x values}

1/X^2 {variance}

* + 1. **Normality**:
       1. Does not affect the parameter estimate, but it affects the test results
       2. Use QQ plot or histogram to identify
       3. Tests include either Anderson-Darling or Shapiro-Wilks
       4. Use box-cox transformation
    2. **Independent observations**:
       1. does not affect the parmater estimates, but the standard error is compromised
       2. do not want auto-correlation
  1. Auto-Correlation
     1. Durbin-Watson Test
        1. Ho: no residual correltion
        2. Ha: Residual correlation
        3. Bounded between 0 and 4
        4. D = 2 🡪 fail to reject Ho, and assume we do not think there is autocorrelation
        5. D < 2 🡪 possible positive autocorrelation
        6. D > 2 🡪 possible negative autocorrelation

* 1. Diagnostic Statistics: Infulential points and outliers
     1. Outliers 🡪 Y
        1. Standardized residuals
           1. Not sure used much
        2. Studentized residuals
           1. Obtained by dividing resduals by standard error
           2. Abs(>3) is used as cutoff
     2. Infuential point 🡪 X
        1. Cooks D
           1. Measures the distance in the regression estimate when the ith observation is left out
           2. Cutoff is D >

K = number of vars in model, n = sample size

* + - 1. DFFITS
         1. Measures the impact the ith observation has on the predicted value
         2. One value per datapoint
         3. Cutoff is Abs(DFFITSi) > 2

P = number of parametes

* + - 1. DFBETAS
         1. Measure of change in the ith parameter estimate with the deletion of the ith observation
         2. One value per data point
         3. Cutoff is Abs(DFFITSi) > 2
      2. Hat Values
         1. One value per point
         2. Cutoff is hii
    1. Multicollinearity
       1. VIF (variance inflation factor)
          1. The larger the worse
       2. VIF =
       3. VIF over 10 is considered bad
       4. Center any independent variables in the polynomial regression model
    2. Comparing models
       1. Root MSE 🡪 SD of errors
          1. Not easily interpretable
       2. Mean Absolute Error (MSE) 🡪 on average, how big is your mistake
          1. Not scale invariant (depends on how big the data variation is)

Misisng up is different from missing down

* + - 1. Mean Absolute Percentage Error (MAPE) 🡪 % off from the truth
         1. Not scale invertible due to percents
         2. Prediction MAPE =
  1. Model heiarchy
     1. 🡪if a high order term is in model, all lower order terms must also be in model
     2. 🡪 if an interaction is in the model, both of the regualar variables must also be in model
  2. Transformations done make the model lose interpretability (except log transformations)

1. **Model Building and Scoring for Prediction**
   1. Linear regression is the best linear unbisased estimator (blue)
      1. Unbiased means that the sample means follow a normal distribution for large enough sample of means
      2. Linear regression is unbiased (smallest spread around the truth)
   2. Regularized Regression 🡪 puts constraints on the estimated coefficients in the model and shrinks those estimates to 0 🡪 called the “penalty term”, and is denoted by lambda
      1. for when you have problems with multicollinearity
      2. Coefficients become biased, but potentially improve the variance of the model
   3. Ridge Regression
      1. If lambda = 0, then regular ordinary least squared regression
      2. As lambda approaches infitinty, coefficient shrink to 0
      3. Beta squared
      4. Never drops variables
   4. LASSO Regression
      1. If lambda = 0, then regular ordinary least squared regression
      2. As lambda approaches infitinty, coefficient shrink to 0
      3. Absolute value of beta
      4. Removes variables from equation 🡪 can be used for variable selection
   5. Elastic Net Regression
      1. Combines both Ridge and LASSO models
         1. LASSO does variable selection and drops variables
         2. Ridge keeps all variables
            1. LASSO penalizes numerator, while Ridge penalizes denominator
      2. Can be used for variable selection
   6. Cross-Validation 🡪 used to prevent overfitting of the model
      1. Generally use 10 or 5 “folds”
      2. If 0 is not present on the cv.glmnet graph, then that means a better model than OLS can be made
         1. 1st line on graph is the lambda value for which the model has the lowest cross-validation mean squared error
         2. 2nd line is the minimum value 1 Standard error away from the 1st dashed line
            1. Generally used the 2nd dashed line to select variables
   7. Comparing models
      1. Root MSE 🡪 SD of errors
         1. Not easily interpretable
      2. Mean Absolute Error (MSE) 🡪 on average, how big is your mistake
         1. Not scale invariant (depends on how big the data variation is)
            1. Misisng up is different from missing down
      3. Mean Absolute Percentage Error (MAPE) 🡪 % off from the truth
         1. Generally use this one with MSE
         2. Not scale invertible due to percents
         3. Prediction MAPE =
2. **Categorical Data Analysis**
   1. Categorical Data
      1. Nominal 🡪 categories with no logical ordering
      2. Ordinal 🡪 categories with logical ordering
         1. Can be written forwards and backwards
         2. Binary is always ordinal
   2. An association exists between two categorical variables if there is a change in distribution as another category changes
   3. Tests of Association
      1. Basically the tests of association tell whether or not there is an association, and then the measures of association tell us how much the variables are associated
      2. Chi-sqaured distribution
         1. Ho: no association
         2. Ha: there is an association
         3. Bounded below by 0
         4. Right Skewed
         5. One degree of freedon
      3. Pearson chi-sqaured
         1. Works on any two cateorical variables (can be ordinal and/or nominal
         2. Used of one of the variables is nominal
      4. Likelihood Ratio chi-squared
         1. Works on any two cateorical variables (can be ordinal and/or nominal
      5. Fisher’s Exact Test
         1. Used when 80% or more of the cells do not have an expected count of 5
      6. Mantel-Haenszel chi-squared
         1. Measures association of direction
            1. Tells us there is a notion of relationship direction, but not which way
         2. Used when both variables are either ordinal or binary
            1. R assumes the categories are already in order, and then assumes the order alphabetically
   4. Measures of Association
      1. Odds Ratios
         1. Only for 2x2 tables – binary vs binary
         2. Indicates how much more likely a certain event occurs in one group relative to its occurance in another group
         3. Odds =
         4. For a 2x2 table where inputed by table(row, column), the upper lefthand and lower right hand boxes are the interpetations
      2. Cramers V
         1. Any size table
         2. Bounded between 0 and 1
         3. For nominal variables, can only say yes/no to if they have an association
         4. For comparison purpose only, not for telling if good or bad on its own
      3. Spearmeans Correlation
         1. Ordinal vs ordinal
         2. Tells you direction and strength
         3. Calculated with the Pearson’s correlation on the ranks of the observations instead of the values of the observation
         4. Needs numerical values
   5. Logistic Regression
      1. Used when the target variable is categorical
      2. Has the following properties:
         1. The predicted probability will always be between 0 and 1
         2. The paramter estimates do not enter the model equation linearly
         3. The rate of change of the probability changes as the X’s vary
            1. X’s are not in linear combination with the Y’s
      3. Assumptions:
         1. Independence of observations
         2. Logit in linearity related to variables
      4. Can interpret coefficients of logistic regression through odds ratio
         1. gets odds ratio for that variable
         2. and then🡪 100\*(- 1) gives us interpretability in terms of percent
            1. “increases the expected odds by a%
      5. Continuos X variable Example:
         1. If target is yes/no for house being bonus eligible, and the X var is greater\_living area, then the results can be interpreted as:
            1. Coefficient = 0.0038
            2. = 1.0038
            3. 100\*(1.0038-1) = 0.38%
            4. “Every additional sqaure foot of greater living area increases the expected odds of being bonus eligible by 0.38%”
            5. If we want to do ‘for every 100 additional sqaure feet added’…

= 1.46

100\*(1.46-1) = 46%

“Every additional 100 sqaure feet of greater living area increases the expected odds of being bonus eligible by 46%

* + 1. Categorical X variable example
       1. If target is yes/no for house being bonus eligible, and the X var is yes/no AC then the results can be interpreted as:
          1. Coefficient = 3.56
          2. = 35.16
          3. 100\*(35.16-1) = 3416%
          4. “Homes with AC increases the odds of being bonus eligible by 3416% compared to those without AC
          5. “Homes with AC are 35.16 times as likely to be bonus eligible then compared to homes without central air
    2. Logistic Model Assesment
       1. Compare every pair of 0’s and 1’s of the target variable
          1. Each 0 will be compared to every 1, and each 1 wil be compared to every 0
       2. Concordent pair: the 1 in the model has a high predicted probability in the model than the 0 in the model 🡪 “where the model is right”
       3. Discorden Pair: The 1 in the model has a lower predicted probability in the model than the 0 in the model 🡪 “where the model is wrong”
       4. Tied Pair: The 1 and 0 in the model have the same predicted probability
    3. Logistic Model Variable Selection
       1. All of the variable selection methods from linear regression availible to logistic regression
          1. Forwards, Backwards, StepWise, Ridge, Lasso, Elastic Net