MATH 23C FINAL PROJECT

Rudra Barua, Hope Ha, and Matti Tan

P.1 – One-Page Handout
P.2 – Point Breakdown

To conduct our analysis on Boston housing and Boston weather, we used three datasets. Two of these datasets come from the Ames housing dataset, compiled by Dean De Cock. Both datasets include data on Boston housing that range from numerical variables like lot size, to logical variables like whether a house has central air conditioning, and to categorical variables like roof type. The dataset boston.csv, corresponds to test.csv, and the dataset boston2.csv corresponds to train.csv. Our third dataset is a Boston weather dataset, compiled by Jennifer Peng, which covers Boston weather data ranging from temperature, humidity, wind, and precipitation from 2013 to 2018.

How can we use novel statistics to examine the surface area of open porches in Boston houses?
Since not all houses have open porches, we get a very wide spread of values that may lead us astray when we look at central tendency values. The trimmed mean of 26.825 which removes the outliers in the data might be a more accurate measure of central tendency.

Is there a relationship between whether a house has central air conditioning and its overall condition?
Using a permutation test, we found that there is sufficient evidence to reject the null hypothesis that there is no significant difference in the average overall condition score of Boston houses between houses that have and that do not have central air conditioning.

Is there a relationship between whether a house has central air conditioning and a paved driveway?
Using a chi-squared test, we found that there is sufficient evidence to reject the null hypothesis that whether a house has central air conditioning is independent of whether a house has a paved driveway.

Does Lot Area predict Garage Area?
With a line of best fit found through linear regression, we find that there might be a positive relationship/correlation between lot area and garage area.

Is the lot area of a house connected to the type of sale condition of Boston Houses?
Using logistic regression, we find a curve that appears to be a reasonable approximation of how lot area and sale type are related due to the points in the upper portion of the graph being crossed by the curve. Hence, there appears to be a positive relationship between normal type sales and lot area.

Can we check whether a sampling distribution of the above ground living area is standard normal?
In letting the above grade (ground) living area square feet be equal to random variable Y, we found that the random variable $X= (Y- \mu)/\sigma$ has a standard normal distribution, as a histogram of our sampling distribution is well fit by a standard normal curve.

Does the average wind in Boston from 2013-2018 follow a Gamma Distribution?
Although we seemed to fit the average wind in Boston well with a gamma distribution through a histogram and an overlaid density curve, we actually rejected the null hypothesis that our data does follow a gamma distribution, using a chi-square test.

Graphical Representations
We modelled lot area, overall quality, overall condition, and lot configuration using barplots, histograms, and pie-charts. We also made a heatmap of the correlations of the numeric variables, first ignoring ones that are less than 0.05, then isolating ones that have a correlation above 0.5.

Confidence Interval
We used a confidence interval to find what range of house sale prices captures the mean 95% of the time.

Is there a statistically significant difference between the sale price of cul-de-sac and corner houses?
A permutation test (simulation) and a two-sample t-test (classical) reveal that the average sale price of a house on a cul-de-sac in Boston is greater than the average sale price of a house on a corner in Boston.

Math 23c Project Points Breakdown

Required dataset standards

Requirement:	Lines:
1. A dataframe.	4-20
2. At least two categorical or logical columns.	27-35
3. At least two numeric columns.	37-41
4. At least 20 rows, preferably more, but real-world data may be limited.	43-46

Required graphical displays (all graphs must be colored and nicely labeled)

Requirement	Lines
1. A barplot.	84-10, 232-245
2. A histogram.	70-81
3. A probability density graph overlaid on a histogram.	541-543; 618-621
4. A contingency table.	313-369

Required analysis

1. A permutation test.	247-302
2. A p-value or other statistic based on a distribution function.	506-575
3. Analysis of a contingency table.	370-408
4. Comparison of analysis by classical methods (chi-square, CLT) and simulation methods.	506-575

Required submission uploads

1. A .csv file with the dataset	Included in submission, titled: **boston.csv** **boston2.csv** **BostonWeather.csv**
2. A long, well-commented script that loads the dataset, explores it, and does all the analysis.	Included in submission, titled: **FinalMath23cProject.r**
3. A shorter .Rmd with compiled .pdf or .html file that presents highlights in ten minutes.	Included in submission, titled: **FinalMath23cProject.pdf** **FinalMath23cProject.Rmd**
4. A one-page handout that explains the dataset and summarizes the analysis.	Included in submission, titled: **Math23cProjectHandout.pdf**

Additional points for creativity or complexity

1. A data set with lots of columns, allowing comparison of many different variables.	49-53
2. A data set that is so large that it can be used as a population from which samples are taken.	55-60; 579-624
5. A graphical display that is different from those in the textbook or in the class scripts.	111-136
6. Appropriate use of R functions for a probability distribution other than binomial, normal, or chi-square.	615-617 (using qgamma)
8. A convincing demonstration of a relationship that might not have been statistically significant but that turns out to be so.	692-799
11. Nicely labeled graphics using ggplot, with good use of color, line styles, etc., that tell a convincing story.	Histogram (70-81), Barplots (84-101, 232-245), Piecharts (111-136), and Scatter Plot + Linear Regression Line (422-443)
12. An example where permutation tests or other computational techniques clearly work better than classical methods.	506-575
13. Appropriate use of novel statistics (e.g. trimmed mean, maximum or minimum, skewness, ratios).	146-189
14. Use of linear regression.	411-450
15. Calculation and display of a logistic regression curve.	453-502
16. Appropriate use of covariance or correlation.	192-222
17. Use of theoretical knowledge of chi-square, gamma, or beta distributions.	692-699
18. Use of theoretical knowledge of sampling distributions.	579-624
19. A graphical display that is different from those in the class scripts.	104-140
20. Calculation of a confidence interval.	628-688

Subjective impression

1. Immediately disband the search committee and hire them.	**The whole script. :)**