Executive Summary

For our final project, we focused on a statistical analysis on housing sales. Our research proposal is to analyze if the factors *Bedrooms and Floors* and their treatment levels explain significant variations in prices of houses and to see if there is any significant interaction between the Factors. We will be analyzing 5 different models for key results and conclusions. These models are Completely Randomized Design, Randomized Block Design, Completely Randomized Factorial Design, Randomized Block Factorial, and Completely Randomized Analysis of Covariance. We will be utilizing Excel, Python, and SAS to aid with data cleaning and coding the models. All SAS codes can be found in the Appendix.

To begin, we found a dataset on Kaggle that showed house sales in Kings County between 2014-2015. This dataset includes 21 columns and 21, 614 rows, however we decided on 5 columns that will serve as treatment factors, blocking factor, covariate, and response variable. The number of bedrooms and number of floors are our main factors, the blocking factor is the construction grade, the covariate is the square foot of living, and the response variable is the price of the home. We chose 3 levels for each of the independent categorical variables. We decided to have 5 replicates for each of the combinations. Using these combinations, we ran a Python script to convert this into 3\*3\*3\*5 =135 rows of data to feed into the SAS program. From here, we began the analysis on the models.

First, we ran a completely randomized design. We wish to find out if our factor- the number of Bedrooms has a significant effect on price . I.e. if the mean prices significantly differ between the 2 bedroom, 3 bedroom and 4 bedroom houses. For this, our null hypothesis is that there is no significant difference in means of all combinations and our alternative hypothesis is that at least one of our combinations of means vary significantly. For this, we ran Levene’s test for Homogeneity of variance and it gave us a p-value of 0.6755 indicating that our Homogeneity of Variance assumption is tenable. Now, We get a p-value of 0.4292 for the Factor Bedrooms, so there is no significant effect of Bedrooms on Price. Based on the One-Way Anova Results, we conclude that the number of Bedrooms do not significantly affect the House Prices at 5% level of significance. Next, we wish to find out if the number of Floors have significance. I.e. if the mean prices significantly differ between the houses with respect to number of floors. For this, we define null and alternative hypotheses accordingly. Levene’s test for Homogeneity gave us a p-value of 0.1555, so our Homogeneity of Variance assumption is tenable. We get a p-value of 0.0023 for the Factor Bedrooms, so there is significance. So, we reject the Null Hypothesis. We now know that at least one level in Floors is significantly different from another level. We use the REGWQ and Tukey multiple comparisons test to analyse and find out where the source of the significance is. Based on Both REGWQ and Tukey tests, we conclude that Houses with 1.5 floors have a significantly higher mean price compared to the other two types.In the next design RBD, we will be adding a blocking variable to see if our conclusions change accordingly.

Next, we ran a randomized block design. We want to observe if our conclusions change from the ones in CRD if we include a blocking variable. We have chosen Grade as the blocking variable in RBD. Basically, adding a new variable can help us get more accurate insights because this variable will account for some of the variance in Price. We also observe this as our error term becomes smaller, thereby giving us more accurate projection of the influence of Factor Bedroom on our dependent variable, Price. We get a p-value of 0.3309 for the Factor Bedrooms, so there is no significance at 0.05 level of significance. Based on the Two-way Anova Results, we still have the same conclusion as we had in CRD that the number of Bedrooms do not significantly affect the House Prices. But, we have one significant observation that the p-value has decreased from 0.4292 to 0.3309 after adding the Blocking variable concluding the number of bedrooms has a comparatively higher effect than was thought earlier. Next, our factor of interest is Floors, but we add a nuisance variable to see if our conclusions change from the previous design. The Variable Grade will act as a Blocking Variable in this Design. Addition of this variable can help us get more accurate insights because this variable will account for some of the variance in Price. As a result, error term is reduced and it projects the influence of the Factor Bedrooms on Price more accurately. We get a p-value of < 0.001 for the Factor Floors, so there is significance at α = 0.05. Based on the Anova Results, we still conclude that the number of Floors do significantly affect the House Prices. It has to be noted that the p-value has decreased to less than 0.0001 after adding the Blocking variable - the number of floors has a comparatively higher effect than was thought earlier.

Third, we ran a completely factorial design. We used Completely Randomized Factorial Design to see if the combinations of number of floors and number of bedrooms will affect the prices. From the model Price = Floors Bedrooms Floors \* Bedrooms, we found that the p-value was 0.0398. We can reject the null hypothesis, which means that the number of floors or the number of bedrooms or the interaction has an affect on the price. We also found that the interaction of floors and bedrooms p-value is 0.6519, which is higher than the significance level.

Next, we remove the interaction from the model and rerun it because the interaction is insignificant. We found out that the *Floors* alone had significance with a p-value of 0.0023. Post Hoc results showed that houses with 1.5 Floors have a significantly higher mean price than the other groups. And, the other two groups were not significantly different in terms of mean Price.

Next, we ran a randomized block factorial design. We conduct the analysis to check if there is significance due to levels in any of the two factors accounting for variance created by the blocking variable. We still found that the interaction was insignificant. We reran the model again without the interaction. We found that the p-value decreased from the CRF model from 0.0023 to 0.0003 and the factor Bedrooms does not have statistical significance(although there is a decrease in p-value from 0.41 to 0.29). The REGWQ post hoc test results are also consistent with the CRF design’s REGWQ post hoc results. I.e. Houses with 1.5 Floors have a significantly higher mean price than the other groups. And, the other two groups were not significantly different in terms of mean Price.

Last, we ran a completely randomized analysis of covariance. We used ANCOVA as an extension of ANOVA to see if our Factors show significance after accounting for the effects of our covariate: square foot of living. In the first model, we used Floors as the main independent factor. When checking for the Homogeneity of the regression slopes assumption, we found that the interaction SQFT\_living\*Floors had a p-value lower than 0.05, which corresponds to significant interaction and hence The homogeneity of regression slopes assumption is not tenable in this case. Then, we considered an ANCOVA with Bedrooms as the main Factor, in this case, the regression slopes were homogenous. We found that the levels in bedrooms were again not significant with a p-value of 0.3529. Therefore, we cannot reject the null hypothesis that there is no significance between the levels. We infer that the levels in bedrooms do not explain a significant amount of variation in price. Consistent with this, the pairwise comparison of least square means, adjusted for the covariate did not show significance.

Ultimately, our final conclusions for this project were that the factor *Floor* has the most significant effect on the price of the house. The number of bedrooms do not significantly affect the price. We also found that houses with 1.5 floors have a significantly higher mean Price than the other homes.

**Appendix:**

**Completely Randomized Design:**

A: Independent Variable : Bedrooms

\*;

\* ODS LISTING;

\*;

ODS Graphics On;

\*;

Title 'SAS Program for CRAC Design-HOUSE\_PRICES'; Options Linesize = 80;

\*;

proc import datafile="/folders/myfolders/BIA654\_Clean.xlsx" out=HousePrices dbms=xlsx replace;

run;

\*Perform CRF ANOVA for the miles per gallon with class and type.;

Proc Print Data=HousePrices;

\*;

Proc GLM Data=HousePrices;

Class BEDROOMS;

Model PRICE = BEDROOMS;

Means BEDROOMS/regwq tukey bon alpha=0.05;

Means BEDROOMS/hovtest= levene hovtest=bartlett;

Means BEDROOMS;

\*;

Proc Sort Data=HousePrices;

By BEDROOMS;

\*;

Proc Univariate Data=HousePrices;

Var PRICE;

By BEDROOMS;

ID BEDROOMS;

\*;

Run;

\*;

**Completely Randomized Design**

B: Independent Variable : Floors

\*;

\* ODS LISTING;

\*;

ODS Graphics On;

\*;

Title 'SAS Program for CRAC Design-HOUSE\_PRICES'; Options Linesize = 80;

\*;

proc import datafile="/folders/myfolders/BIA654\_Clean.xlsx" out=HousePrices dbms=xlsx replace;

run;

\*Perform CRF ANOVA for the miles per gallon with class and type.;

Proc Print Data=HousePrices;

\*;

Proc GLM Data=HousePrices;

Class FLOORS;

Model PRICE = FLOORS;

Means FLOORS/regwq tukey bon alpha=0.05;

Means FLOORS/hovtest= levene hovtest=bartlett;

Means FLOORS;

\*;

Proc Sort Data=HousePrices;

By FLOORS;

\*;

Proc Univariate Data=HousePrices;

Var PRICE;

By FLOORS;

ID FLOORS;

\*;

Run;

\*;

ODSGraphicsOff;

\*;

\*;

Quit;

**Randomized Block Design:**

A: Factor: Bedrooms, Blocking Factor: Grade

\*;

\* ODS LISTING;

\*;

ODS Graphics On;

\*;

Title 'SAS Program for CRAC Design-HOUSE\_PRICES'; Options Linesize = 80;

\*;

Proc import datafile="/folders/myfolders/BIA654\_Clean.xlsx" out=HousePrices dbms=xlsx replace;

run;

Proc Print Data=HousePrices;

PROC SORT Data = HousePrices;

BY BEDROOMS;

Proc GLM Data=HousePrices;

Class GRADE BEDROOMS;

Model PRICE = GRADE BEDROOMS;

Means BEDROOMS/regwq tukey bon alpha= 0.05;

Means BEDROOMS/hovtest= levene hovtest= bartlett;

\*;

Proc Univariate Data=HousePrices;

Var PRICE;

By BEDROOMS GRADE;

ID BEDROOMS GRADE;

\*;

\*;

ODS Graphics Off;

\*;

Quit;

**Randomized Block Design:**

B: Factor: Floors Blocking Factor: Grade

\*;

\* ODS LISTING;

\*;

ODS Graphics On;

\*;

Title 'SAS Program for CRAC Design-HOUSE\_PRICES'; Options Linesize = 80;

\*;

Proc import datafile="/folders/myfolders/BIA654\_Clean.xlsx" out=HousePrices dbms=xlsx replace;

run;

Proc Print Data=HousePrices;

PROC SORT Data = HousePrices;

BY FLOORS;

Proc GLM Data=HousePrices;

Class GRADE FLOORS;

Model PRICE = GRADE FLOORS;

Means FLOORS/regwq tukey bon alpha= 0.05;

Means FLOORS/hovtest= levene hovtest= bartlett;

\*;

Proc Univariate Data=HousePrices;

Var PRICE;

By FLOORS GRADE;

ID FLOORS GRADE;

\*;

\*;

ODS Graphics Off;

\*;

Quit;

3. SAS Code for CRF Design:

\*;

\* ODS LISTING;

\*;

ODS Graphics On;

\*;

Title 'SAS Program for CRF-33 Design'; Options Linesize = 80;

\*;

proc import datafile="/folders/myfolders/BIA654\_Clean.xlsx" out=HousePrices dbms=xlsx replace;

run;

\*Perform CRF ANOVA for the miles per gallon with class and type.;

Proc Print Data=HousePrices;

\*;

Proc GLM Data=HousePrices;

Class FLOORS BEDROOMS;

Model PRICE = FLOORS BEDROOMS FLOORS\*BEDROOMS;

Proc GLM Data=HousePrices;

Class FLOORS BEDROOMS;

Model PRICE = FLOORS BEDROOMS;

Means FLOORS BEDROOMS/regwq tukey bon alpha=0.05;

Means FLOORS BEDROOMS/hovtest= levene hovtest=bartlett;

Means FLOORS BEDROOMS;

\*;

Proc Sort Data=HousePrices;

By FLOORS BEDROOMS;

\*;

Proc Univariate Data=HousePrices normal;

Var PRICE;

By FLOORS BEDROOMS;

ID FLOORS BEDROOMS;

\*;

Run;

\*;

\*;

ODS Graphics Off;

\*;

Quit;

4. Randomized Blocking Factorial\_Design:

\*;

\* ODS LISTING;

\*;

ODS Graphics On;

\*;

Title 'SAS Program for CRF Design-HOUSE\_PRICES'; Options Linesize = 80;

\*;

proc import datafile="/folders/myfolders/BIA654\_Clean.xlsx" out=HousePrices dbms=xlsx replace;

run;

\*Perform CRF ANOVA for the miles per gallon with class and type.;

Proc Print Data=HousePrices;

\*;

Proc GLM Data=HousePrices;

Class GRADE FLOORS BEDROOMS;

Model PRICE = GRADE FLOORS BEDROOMS FLOORS\*BEDROOMS;

Proc GLM Data=HousePrices;

Class GRADE FLOORS BEDROOMS;

Model PRICE = GRADE FLOORS BEDROOMS;

Means FLOORS BEDROOMS/regwq tukey bon alpha=0.05;

Means FLOORS BEDROOMS/hovtest= levene hovtest=bartlett;

Means FLOORS BEDROOMS;

\*;

Proc Sort Data=HousePrices;

By FLOORS BEDROOMS;

\*;

Proc Univariate Data=HousePrices Normal;

Var PRICE;

By FLOORS BEDROOMS;

ID FLOORS BEDROOMS;

\*;

Run;

\*;

\*;

ODS Graphics Off;

\*;

Quit;

5. **Completely Randomized Analysis of Covariance**

A.Main Factor: Bedrooms| Covariate: SQFT\_living

\*;

\* ODS LISTING;

\*;

ODS Graphics On;

\*;

Title 'SAS Program for CRAC Design-HOUSE\_PRICES'; Options Linesize = 80;

\*;

proc import datafile="/folders/myfolders/BIA654\_Clean.xlsx" out=HousePrices dbms=xlsx replace;

run;

Proc Print Data=HousePrices;

\*;

PROC GLM DATA = HousePrices;

Model PRICE = SQFT\_LIVING;

\*SIMPLE LINEAR REGRESSION WITH EACH LEVEL OF BEDROOMS;

Proc sort Data = HousePrices;

BY BEDROOMS;

Proc GLM Data=HousePrices;

Model PRICE = SQFT\_LIVING;

By BEDROOMS;

\*HOMOGENEITY OF REGRESSION SLOPES ;

PROC GLM DATA = HousePrices;

Class BEDROOMS;

Model PRICE = BEDROOMS SQFT\_LIVING BEDROOMS\*SQFT\_LIVING;

\*ANCOVA WITH SQFT\_LIVING AS THE COVARIATE- FLOOR REMOVED;

PROC GLM DATA = HousePrices;

CLASS BEDROOMS;

Model PRICE = BEDROOMS SQFT\_LIVING;

Means BEDROOMS / REGWQ Tukey;

LSMEANS BEDROOMS/ STDERR PDIFF ADJUST = tukey;

Proc Sort Data = HousePrices;

By BEDROOMS;

Proc Univariate Data=HousePrices normal;

Var PRICE;

By BEDROOMS;

ID BEDROOMS;

\*;

Run;

\*;

\*;

ODS Graphics Off;

\*;

Quit;

CRAC:

B. Main Factor: Floors | Covariate: SQFT\_living

\*;

\* ODS LISTING;

\*;

ODS Graphics On;

\*;

Title 'SAS Program for CRAC Design-HOUSE\_PRICES'; Options Linesize = 80;

\*;

proc import datafile="/folders/myfolders/BIA654\_Clean.xlsx" out=HousePrices dbms=xlsx replace;

run;

Proc Print Data=HousePrices;

\*;

PROC GLM DATA = HousePrices;

Model PRICE = SQFT\_LIVING;

\*SIMPLE LINEAR REGRESSION WITH EACH LEVEL OF FLOORS;

Proc sort Data = HousePrices;

BY FLOORS;

Proc GLM Data=HousePrices;

Model PRICE = SQFT\_LIVING;

By FLOORS;

\*HOMOGENEITY OF REGRESSION SLOPES ;

PROC GLM DATA = HousePrices;

Class FLOORS;

Model PRICE = FLOORS SQFT\_LIVING FLOORS\*SQFT\_LIVING;

\*ANCOVA WITH SQFT\_LIVING AS THE COVARIATE- FLOOR REMOVED;

PROC GLM DATA = HousePrices;

CLASS FLOORS;

Model PRICE = FLOORS SQFT\_LIVING;

Means FLOORS / REGWQ Tukey;

LSMEANS FLOORS/ STDERR PDIFF ADJUST = tukey;

Proc Sort Data = HousePrices;

By FLOORS;

Proc Univariate Data=HousePrices normal;

Var PRICE;

By FLOORS;

ID FLOORS;

\*;

Run;

\*;

\*;

ODS Graphics Off;

\*;

Quit;