Stats141XP: Final Report (Team 6)

Team Members: Amanda Xu, Gloria Won, Jacob Samuels, Jordan Tallman, Priya Jain, Charles Liu
Instructor: Professor Esfandiari
Stats 141XP
3 June 2021
Subject: Neural Networks for Facial Images (Age & AI Project)

Table of Contents:

Abstract

For this project, we wanted to calculate the Accuracies and Reliabilities of 3 AI Models that were used to predict the Actual Age of patient's before ("pre") and after ("post") surgery. To analyze the Accuracies, we would start off by visualizing the differences between the Actual Ages and each AI Model, and we would visualize them using Boxplots and Histograms. Once we have some idea of the AI Models, we would calculate Accuracies for each model by using Mean Square Errors (MSE) and Correlation between the Actual Age and each AI Model.
For Reliabilities, we wanted to find how consistently the AI model is predicting the same age for all photos of a patient. We would group each patient by their "pre", "post", "pre1 & pre2", or "post1 & post2" data. Then, we would find the Correlation, p-value (and Confidence Interval), and finally the raw reliability score to see how reliable each model is.

Clear statement of the problems to be solved

1. How accurately is the AI model predicting the actual age of the patient?
    a. Comparing accuracy across the 3 different models and find if one performs better than the rest
    b. **Defining Accuracy:** how close the predicted age is to the actual age of the patient at the time of the picture
2. How Reliable are the AI Models in predicting the Actual Age?
    a. We will compare the AI Models to the Actual Age to see how close they are to consistently predicting the Patient's actual age
    b. **Defining Reliability:** how close the predicted age is to the actual age of the patient at the time of the picture if the experiment was replicated multiple times. Basically, how consistent we can replicate the experiment for predicting Actual Age from the AI Model.
3. The goal of finding significant variables is to identify which of the given variables provide a statistically significant explanation of the variation in the model.
    a. We were looking to see which variables affected the accuracy and reliability of the model the most.

Schematic showing the relationship between the variables of the study and how they were measured

1. **patient_name:** encoded character variable (format XXX_XXX) that indicates patients' names.
2. **gender:** categorical/binary variable, "F" indicates female and "M" indicates male.

3. **dob:** categorical variable, the date of birth of the patient done in (MM/DD/YYYY) format
4. **surgery_date:** categorical variable, the date the patient was scheduled on to have their surgery done in (MM/DD/YYYY) format
5. **age_on_surg_day:** numerical variable, the age the patient was when going to their surgery, surgery done in (MM/DD/YYYY) format
6. **pre_or_post_operation:** categorical variable, indicates the status/time of the patients' photos, whether it was before surgical operation or after surgical operations. Some patients have three preoperative photos, labeled as "pre", "pre_2" and "pre_3". Some patients also have three postoperative photos, labeled as "post", "post_2" and "pre_3".
7. **age_actual:** numerical/double variable, ranges from 30.26 to 83.82. Represents the actual age of the patients, the time when the photo was taken.
8. **ai_model_1:** numerical variable, integers ranging from 27 to 88. Represents the predicted age of the patients by the AI Model 1.
9. **ai_model_2:** numerical variable, integers ranging from 31 to 88. Represents the predicted age of the patients by the AI Model 2.
10. **ai_model_3:** numerical variable, integers ranging from 31 to 89. Represents the predicted age of the patients by the AI Model 3.
11. **Upper bleph:** categorical variable, indicating the type of surgeries performed on the upper bleph. "0" means no surgery, "1" means surgery performed on the right eye upper bleph, "2" means surgery performed on the left eye upper bleph and "3" means surgery performed on both upper blephs.
12. **Lower bleph:** categorical variable, indicating the type of surgeries performed on the lower bleph. "0" means no surgery, "1" means surgery performed on the right eye lower bleph, "2" means surgery performed on the left eye lower bleph and "3" means surgery performed on both lower blephs.
13. **MRD1 - right:** numerical/double variable measured in millimeter, ranges from 0.395mm to 6.635mm. Measures the distance from patients' pupils to upper eyelids on the right eye.
14. **MRD2 - right:** numerical/double variable, ranges from 3.004mm to 7.758mm. Measures the distance from patients' pupils to lower eyelids on the right eye.
15. **PTB - right:** numerical/double variable, ranges from 0.75mm to 26.29mm. Measures the distance from patients' pupils to the lowest point of the eyebrow on the right eye.
16. **TPS - right:** numerical/double variable, ranges from 0.493mm to 15.364mm. Measures the distance from patients' pupils to the upper eyelid crease on the right eye.
17. **MRD1 - left:** numerical/double variable, ranges from 0.553mm to 6.172mm. Measures the distance from patients' pupils to upper eyelids on the left eye.

18. **MRD2 - left:** numerical/double variable, ranges from 1.824mm to 7.754mm. Measures the distance from patients' pupils to lower eyelids on the left eye.
19. **PTB - left:** numerical/double variable, ranges from 6.194mm to 26.321mm. Measures the distance from patients' pupils to the lowest point of the eyebrow on the left eye.
20. **TPS - left:** numerical/double variable, ranges from 0.671mm to 14.285mm. Measures the distance from patients' pupils to the upper eyelid crease on the left eye.
21. **f-number:** categorical variable, the ratio of the focal length of a camera lens to the diameter of the aperture being used for a particular shot of the patient's photo
22. **exposure time:** categorical variable, the time span in which the film of the camera is actually exposed to the light when it records the patient's photos
23. **focal length:** categorical variable, tells us the angle of view (how much of the scene will be captured) and the magnification (how large individual elements will be)
24. **make:** categorical variable, tells us who was the maker of the camera used to take the photo (i.e. Apple for iPhone 6)
25. **model:** categorical variable, tells us what model was being used for the camera and its type
26. **sensitivity:** categorical variable, the measurement of a camera's ability to capture light of the patient's photo in which it helps reduce blurriness
27. **dimensions:** categorical variable, the dimensions of the patient's photo

- **<u>Schematics:</u>**
    - *Numerical and Categorical Variables*

Numerical		**Categorical**	
age_on_surg_day"	"MRD2-right"	"patient_name"	"f-number"
"age_actual"	"PTB-right"	"gender"	"exposure time"
"ai_model_1"	"TPS-right"	"dob"	"focal length"
"ai_model_2"	"MRD1-left"	"surgery_date"	"make"
"ai_model_3"	"MRD2-left"	"pre_or_post_operation"	"model"
"Lower bleph"	"PTB-left"	"Upper bleph (0 = null, 1 = right, 2 = left, 3 = bilateral)"	"sensitivity"

"MRD1-right"	"TPS-left"	"dimensions"	

- *Variables Used to Investigate Accuracy & Reliability*

Dependent	Independent	Categorical
age_actual	ai_model_1 ai_model_2 ai_model_3	pre_or_post_operation

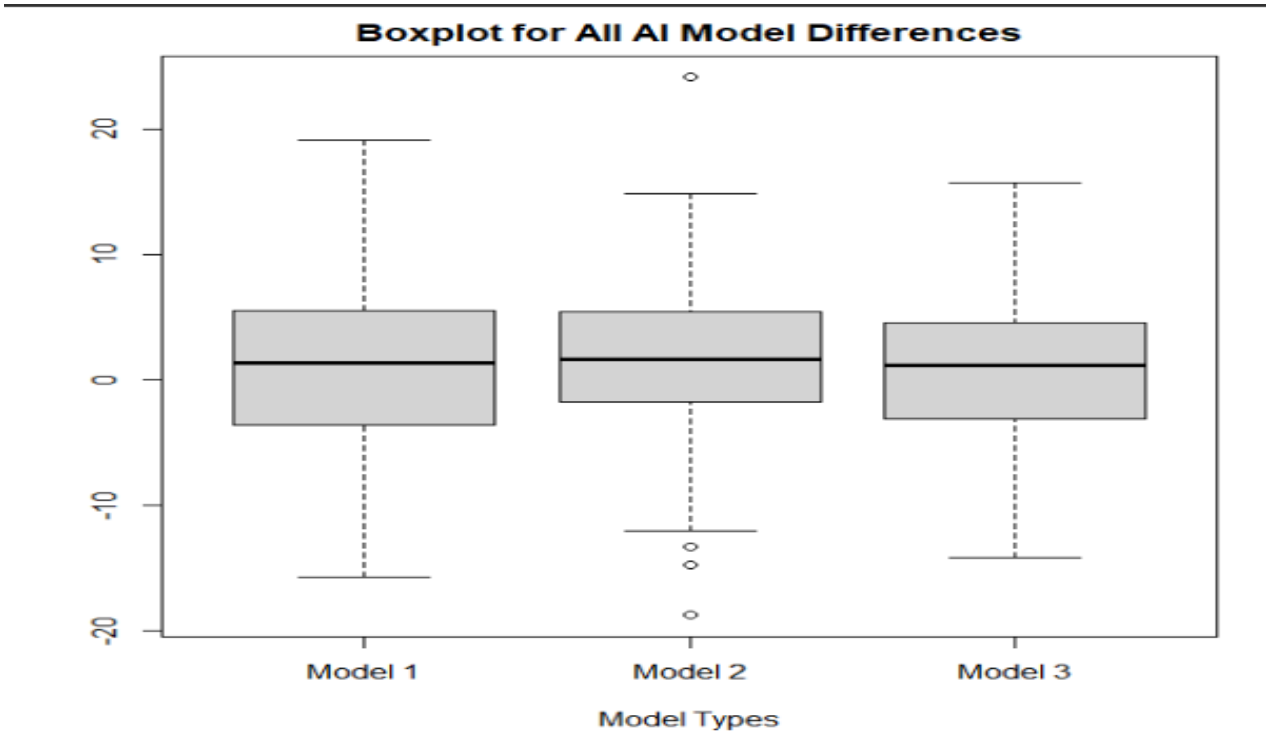
- *pre_or_post_operation*
  - Levels: pre, pre_2, pre_3, post, post_2, post_3
  - Use different combinations of levels during analysis (ex: pre & pre_2)
  - Excluded “pre_3” and “post_3” because not enough observations to help us identity significant findings

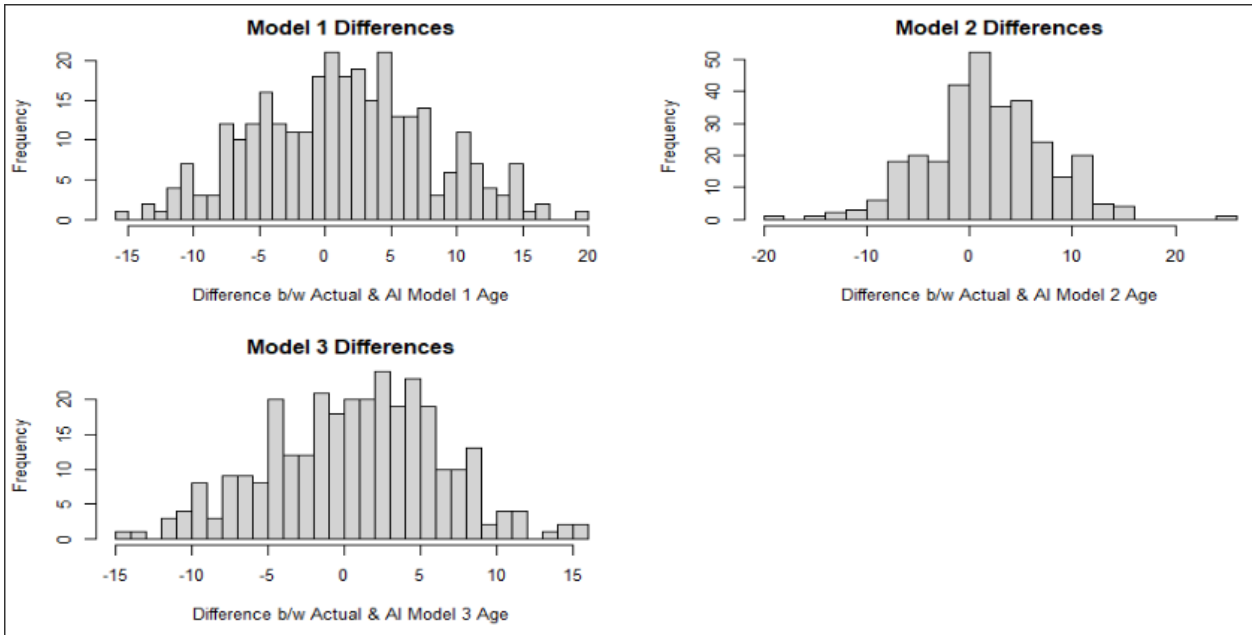
Exploratory data analysis

We are exploring the difference between Actual Age and AI Model Age, and we want to find the model that is closer to zero (most accurate predictions). In the Boxplot and Histogram, they both show that Model 2 and 3 performs the best in terms of being closer to zero and normally distributed around the zero. The summary table is to compare them numerically, showing that Model 2 and 3 still performs the best. Further analysis is needed to see which one performs better than the other.

Graphs/Charts:

(Actual Age - AI Model #)	AI Model 1	AI Model 2	AI Model 3
Mean	1.365	1.759	0.7421
Median	1.329	1.587	1.1389
Min to Max	-15.770 to 19.158	-18.770 to 24.128	-14.2111 to 15.7528





Statistical analysis used to answer the research questions

1. **<u>Accuracy Statistical Methods:</u>**

a. Two methods of comparison
    i. **MSE**
    ii. **Coefficients of Correlation**
b. **Mean Squared Error (MSE)**
    i. What it is: the average of the (actual age - predicted age)^2 for all observations
    ii. Larger the MSE, the lower the accuracy
    iii. Gives us mathematical values we can use to compare the prediction accuracies of the three models
c. **Coefficients of Correlation**
    i. What it is: how correlated the dependent and independent variables are to each other
    ii. Closer it is to 1 or -1, higher the correlation
    iii. Allows us to visually see how close the models are predicting the ages of patients; if the actual age and predicted age are the same, then the correlation should be 1
d. Tests used*:*
    i. **Comparison of correlation test:** to test the hypothesis across all 3 models
        1. Dependent samples because correlations retrieved from same sample
        2. Null hypothesis is that all the models are equal
        3. P-value lower than a significance level of 0.05 gives significant evidence that at least one of the models are different from the others

2. <u>Reliability Statistical Methods:</u>

a. ***Correlation:*** we want to see how consistent the data is with our Actual Age for each AI Model
b. ***p-values:*** we want to check if these AI models are statistically significant in telling us how consistent they are ($p < 0.05$ is statistically significant). If they are statistically significant, we can do further analysis into the reliabilities of these models
c. ***Confidence Intervals:*** This means that if we repeated the experiment many times with datasets drawn from the same population was calculated, the Confidence Intervals would contain the true value at 95% of the time the Confidence Intervals (aka can we reproduce the same model with similar/same results)
d. ***Average "Split Half" Reliability:*** split a test into two halves, administer each half to the same individual, repeat for a large group of individuals,

find the correlation between the scores for both halves. We use it to check how consistent the data is with the Actual Age

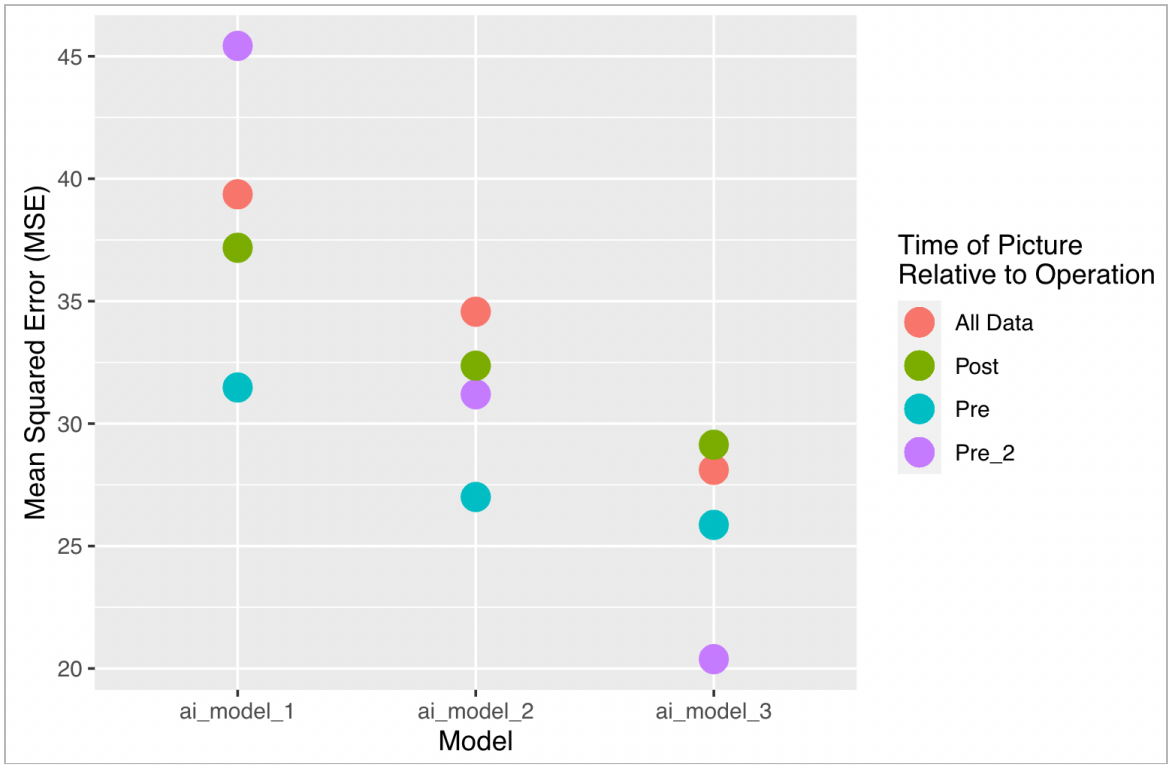
e. ***Raw Score Reliability:*** it basically checks them side-by-side on how correlated they are with each other. We use it to check how consistent the data is with the Actual Age

Summary of results

1. **<u>Accuracy (tables/plots):</u>**

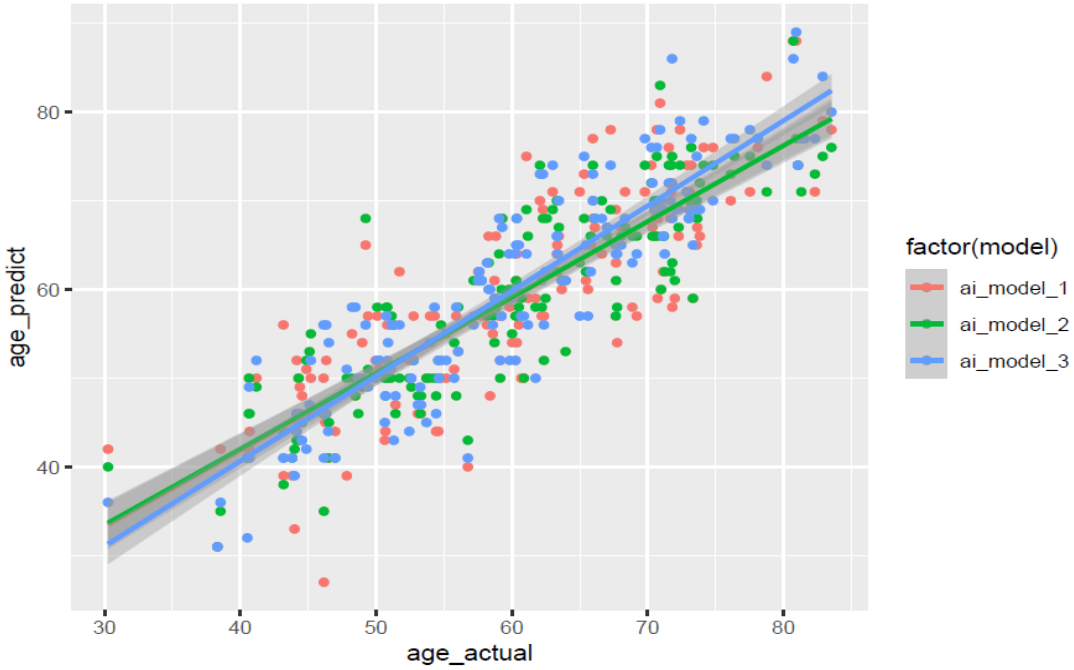
- *MSE (Mean Square Errors)*

Types_for_MSE <chr>	AI_Model_1 <dbl>	AI_Model_2 <dbl>	AI_Model_3 <dbl>
All Pre-Measures	39.19265	28.00477	27.15687
All Post-Measures	43.35426	36.95526	34.69581
Pre1 & Pre2	38.90879	26.98338	27.54190
Post1 & Post2	43.68483	35.05374	33.98027
Both Pre & Post (All Data)	41.14467	32.75085	30.97783

- 

- *Coefficient of Correlation*

- *Model 1:* 0.8402976
- *Model 2:* 0.8768694
- *Model 3:* 0.9011104



- Graph above shows the linear model of age_predict plotted against age_actual for all three models

- AOV performed comparing all pairs of models
  - Model 1 vs model 2 vs model 3
    - P-value = 0.0154
  - Model 1 vs model 2
    - P-value = 0.257
  - Model 1 vs model 3
    - P-value = 0.00225
  - Model 2 vs model 3
    - P-value = 0.0882
- Did the surgery make you look younger? [*Results:* Pre vs Post (Younger/Older)]
  - Average actual age difference between pre and post ~ 1.02 years
  - Found age difference from pre to post for all 3 models
  - Standardized using 1.02 years “natural difference”
  - Post - pre standardized age differences
  - Test performed to identify significance of the average actual age difference
    - Model 1: -0.69 years
      - P-value = 0.283
    - Model 2: -1.77 years

        - P-value = 0.00261
    - Model 3: -1.16 years
        - P-value = 0.0664

2. **Reliability (tables/plots):**

Table 1: All Pre-Measures	Confidence Interval (Lower to Upper)	P-value
AI Model 1	0.7894661 to 0.8796591	2.898158e-46
AI Model 2	0.8366945 to 0.9076984	5.202136e-55
AI Model 3	0.8682801 to 0.9261033	1.666418e-62

Table 2: All Post-Measures	Confidence Interval (Lower to Upper)	P-value
AI Model 1	0.7645162 to 0.8742409	1.393141e-34
AI Model 2	0.7779089 to 0.8818112	3.746188e-36
AI Model 3	0.8174183 to 0.9038364	1.882496e-41

Table 3: Pre1 & Pre2	Confidence Interval (Lower to Upper)	P-value
AI Model 1	0.7937754 to 0.8841353	8.885718e-45
AI Model 2	0.8419960 to 0.9122609	1.765195e-53
AI Model 3	0.8679978 to 0.9271568	2.018188e-59

Table 4: Post1 & Post2	Confidence Interval (Lower to Upper)	P-value
AI Model 1	0.7554866 to 0.8711032	2.698322e-32
AI Model 2	0.7852401 to 0.8876880	1.276184e-35
AI Model 3	0.8165569 to 0.9048618	1.066992e-39

Table 5: All Data	Confidence Interval (Lower to Upper)	P-value
AI Model 1	0.7959473 to 0.8653217	2.010767e-79
AI Model 2	0.8224508 to 0.8833995	6.696195e-88
AI Model 3	0.8572887 to 0.9068893	2.229407e-101

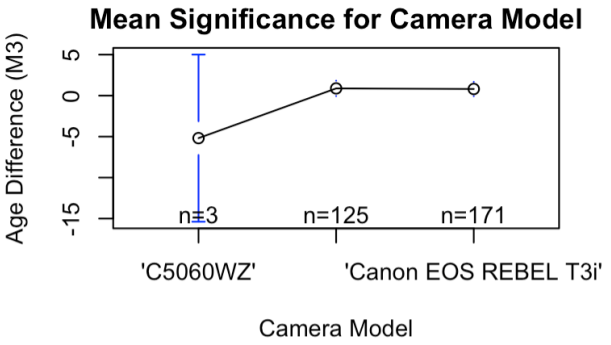
Raw Score Reliability	All Pre	All Post	Pre1 & Pre2	Post1 & Post2	All Data
AI Model 1	0.9536780	0.9473859	0.9544097	0.9440750	0.9508184
AI Model 2	0.9678070	0.9524298	0.9682872	0.9530923	0.9601941
AI Model 3	0.9712402	0.9620903	0.9710032	0.9607307	0.9671768

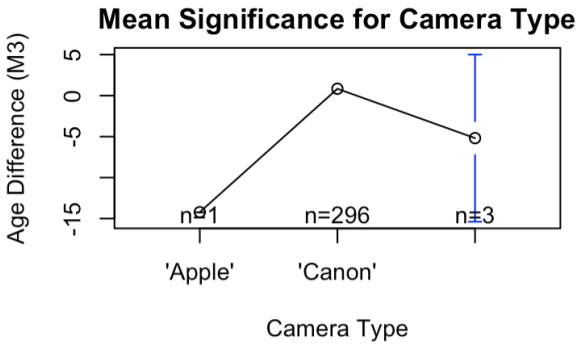
Average Split Half Reliability	All Pre	All Post	Pre1 & Pre2	Post1 & Post2	All Data
Alpha (α) (between all 3 Models & Actual)	0.9708373	0.9625815	0.9713774	0.9619655	0.9668632

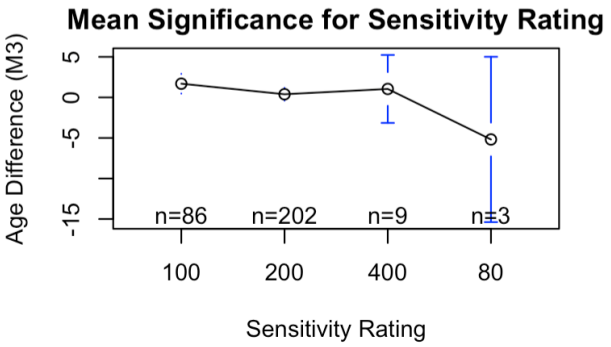
3. **Significant Variables (tables/plots):**

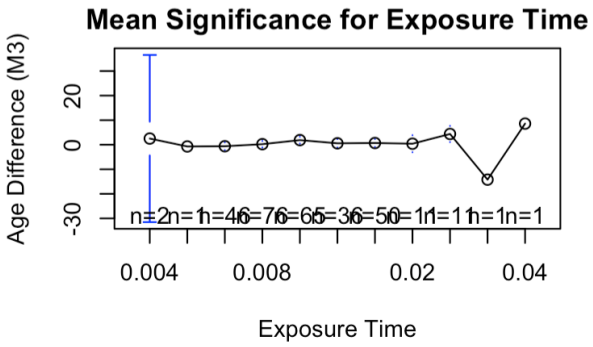
```
                         Df Sum Sq Mean Sq F value Pr(>F)
Age_data$`f-number`       9    380   42.19   1.356  0.209
Age_data$`exposure time`  9    396   43.97   1.413  0.183
Age_data$`focal length`  22    526   23.90   0.768  0.764
Age_data$model            1      6    6.49   0.209  0.648
Age_data$sensitivity      2    109   54.64   1.755  0.175
Age_data$dimensions       6    294   48.98   1.574  0.155
Residuals               252   7844   31.13
```

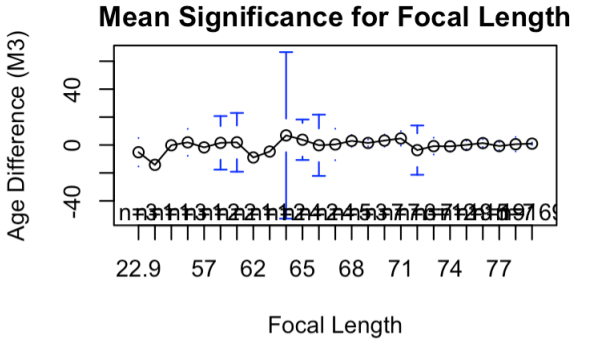
```
                        Df Sum Sq Mean Sq F value Pr(>F)
Age_data$`MRD1-right`    1     95   95.39   3.178 0.0757 .
Age_data$`MRD2-right`    1     20   20.50   0.683 0.4093
Age_data$`PTB-right`     1    201  201.11   6.699 0.0101 *
Age_data$`TPS-right`     1     19   18.77   0.625 0.4297
Age_data$`MRD1-left`     1    103  103.01   3.431 0.0650 .
Age_data$`MRD2-left`     1      0    0.29   0.010 0.9218
Age_data$`PTB-left`      1    120  120.01   3.997 0.0465 *
Age_data$`TPS-left`      1    199  199.09   6.631 0.0105 *
Residuals              293   8796   30.02
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

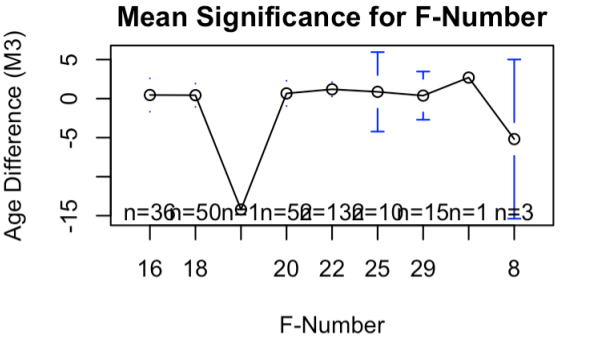


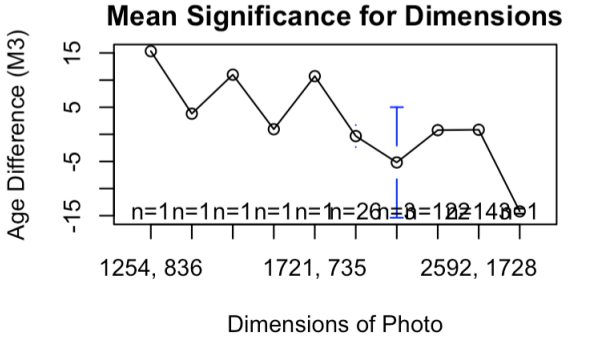


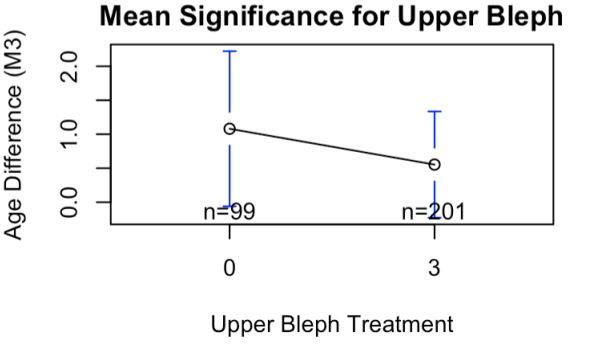












- (Tables of Hard to See Plotmeans' Sample Sizes)

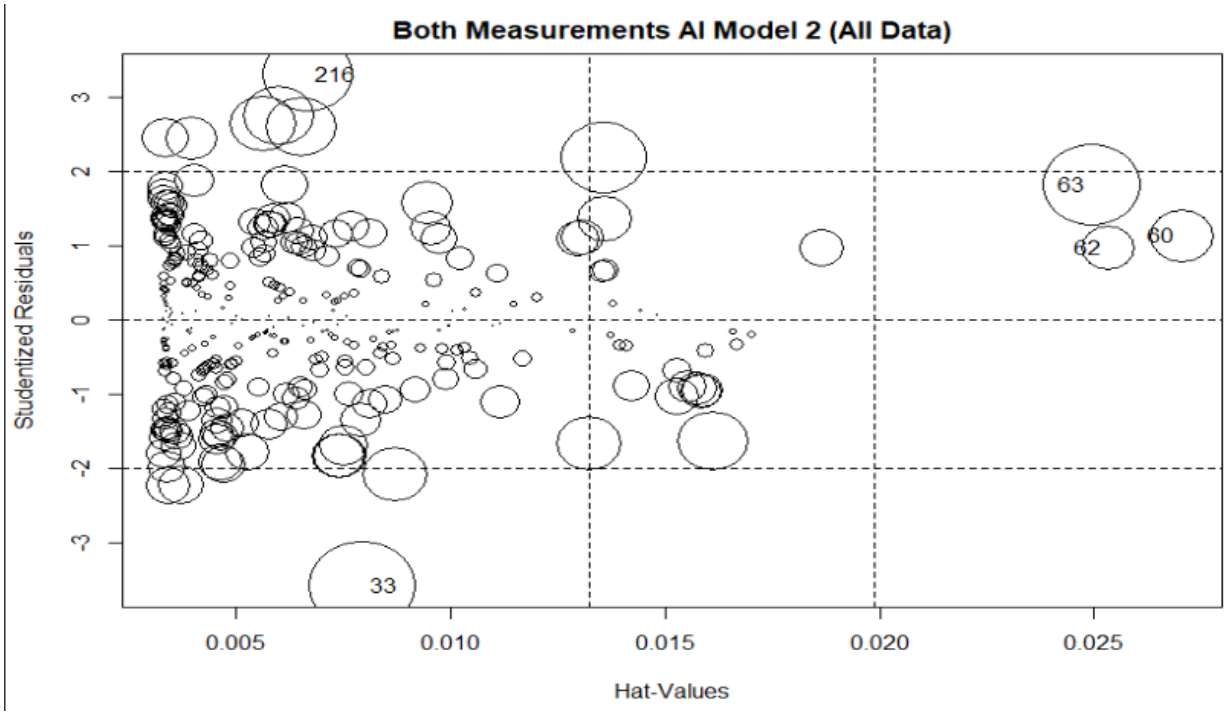
Exposure_Time <chr>	Sample_Sizes <int>
0.004	2
0.005	1
0.00625	46
0.008	76
0.01	65
0.0125	36
0.016666666666667	50
0.02	11
0.025	11
0.033333333333333	1
0.04	1

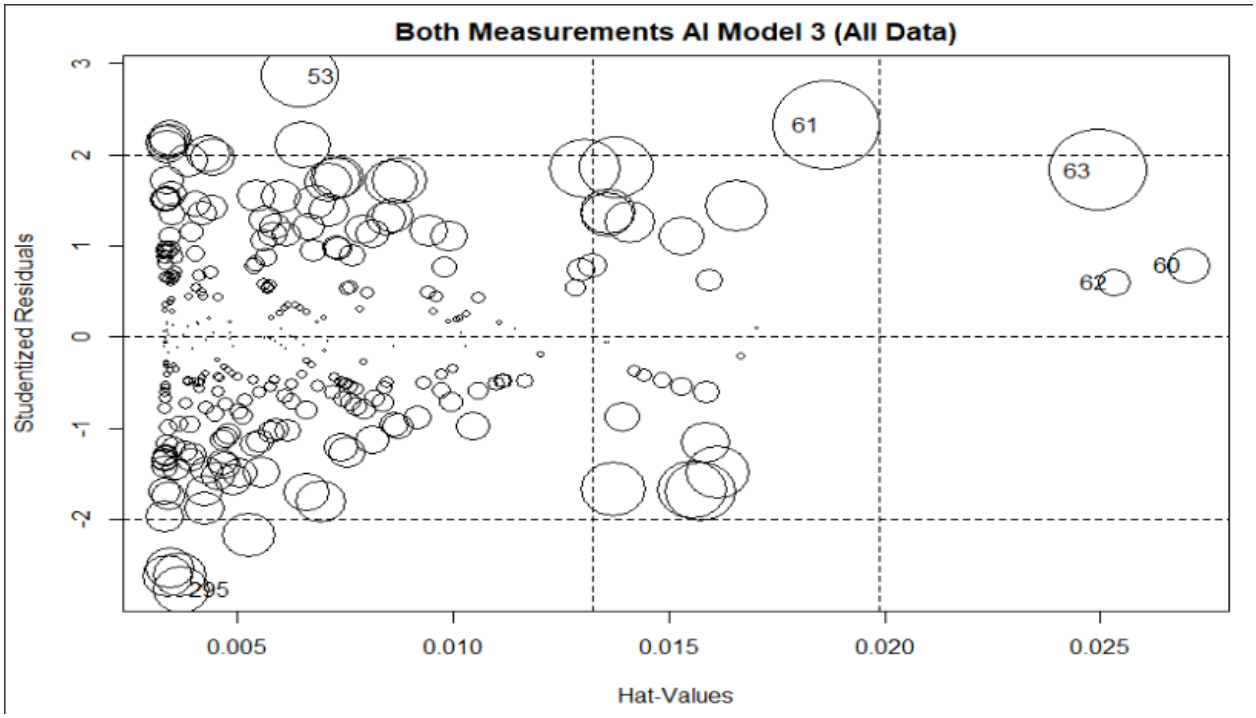
Focal_Length <chr>	Sample_Sizes <int>
22.9	3
4.15	1
53	1
56	3
57	1
60	2
61	2
62	1
63	1
64	2
65	4
66	2
67	4
68	5
69	3
70	7
71	7
72	3
73	7
74	12
75	19
76	15
77	19
78	7
80	169

F_Number <chr>	Sample_Sizes <int>
16	36
18	50
2.2	1
20	52
22	132
25	10
29	15
5.6	1
8	3

Dimensions <chr>	Sample_Sizes <int>
1254, 836	1
1377, 640	1
1431, 678	1
1554, 1087	1
1721, 735	1
1920, 1280	26
2048, 1536	3
2352, 1568	122
2592, 1728	143
3264, 2448	1

4. **Influence Plots (Outliers, Good/Bad/High Leverages):**





Interpretation of results

1. **Accuracy:**

- MSE by AI model and time picture was taken
    - Lower MSE is better
    - Numerically, we can clearly see that Model 3 is the most overall accurate across all 3 AI models. It has the lowest number of Mean Square Errors. The only time Model 3 wasn't the lowest is for "Pre1 & Pre2" measurements only, in which Model 2 was slightly better.
- Coefficients of Correlation by AI model
    - Closer to 1 is better
    - With a p-value of 0.013, based on a significance level of 0.05, there is a significant difference in the coefficients of correlation among the three models, so we reject our null hypothesis. This means that we can see that Model 3 is once again the most overall accurate across all 3 AI Models.

2. **Reliability:**

- From our tables, we can see that each of the AI Models are statistically significant with predicting the Actual Ages of the patients. The Confidence Intervals for AI Model 3 are seemingly tighter, indicating there will be a lower margin of error occurring. We can also see that each of our AI Models are statistically significant. We can see they are all consistent models to the Actual

Ages model due to the p-values being lower than 0.05, however the Confidence Interval determines the magnitude of how consistent the model is.

- As for the Reliability Scores (both raw and Average "Split Half"), we want to make sure they are better than 0.80 in order for it to be considered highly consistent to the Actual Ages model. Using the calculations for each condition, we were able to calculate each of the Model's reliability score (raw). We can see that Model 3 has the best overall raw reliability score, indicating that this is the most consistent model out of all the AI models. We use the Average "Split Half" Reliability to check how reliable all of the AI Models are with the Actual Ages model. On average, they are all greater than 0.80, meaning they are very consistent models overall.
- Overall, we can conclude that AI Model 3 is the most reliable and consistent to the Actual Ages data. The second most consistent is AI Model 2, and AI Model 1 is the worst performing out of all three. However, AI Model 1 is not necessarily bad. It just happens to get outperformed by the other AI Models.

Interpretation of plots

1. **Accuracy:**

- Coefficients of Correlation
    - Plot actual age to the predicted age and analyze coefficient of correlation
    - From the graph visualizing the coefficients of correlation by plotting the points and line of age_actual to age_predict for all three models, we see that model 1 and model 2 are very similar, while we can see that model 3 is visually different from the others.

2. **Significant Variables:**

- the p-values in a small table and just say they are all ($p > 0.05$)
- Plot means for categorical vars. The plot means shows that Camera Model, Camera type, Sensitivity rating and exposure time are relatively sporadic and shows that they are somewhat significant. Focal length seems straight. F-number (sample size of 1 for the downfall) so we may have to remove it. For Dimensions, clear significance. If the line is relatively horizontal, then they are not significant. We can see a majority of the plot means are horizontal. We do have observations that have a small sample size (n=1). It can cause it to seem like the variable is statistically significant, but it really isn't.

3. **Influence Plots (Outliers, Good/Bad/High Leverages):**

- Conditions to be considered Bad Leverage Points (both have to be met):

    - The Hat-value is greater than 4/n (this is a High Leverage point)
    - It is also considered an Outlier ([-2 > StudRes > +2]).
    - **Note:** You can follow more into our analysis of these Bad Leverage Points in the RMD Annotated File, where it'll show you the calculations and numerical values.
- All of the Bad Leverage Points do not necessarily mean they need to be removed from the dataset, but rather they should be studied further into how they might affect the overall analysis. As far as we know, removing these Bad Leverage Points do not significantly affect our analysis.
    - AI Model 2 ("pre_2") [observation 238 from the dataset] is considered a Bad Leverage point. It predicted the age to be approximately 7.28 years more than the actual age.
    - AI Model 3 ("post") [observation 61 from the dataset] is considered a Bad Leverage point. It predicted the age to be approximately 13.79 years more than the actual age.

Overall conclusions

In conclusion, we would recommend using AI Model 3 because it is both the most accurate and reliable for all cases of measurements ("pre" & "post"). It also makes sense that AI Model 3 is the best performing model due to it being trained with 120,000 photos from various angles. Comparatively, Model 2 is only front-facing angles with 50,000 photos, and Model 1 is also only front-facing angles but with only 10,000 photos.

Challenges of the study

1. When dealing with Reliability, a challenge we had to face is that problem with the inconsistency with patient's having the same amounts of before operations (pre) and after operations (post). Not every patient has "pre_2", "pre_3", "post_2", and/or "post_3". This is especially problematic with "pre_3" and "post_3" because of how small the sample sizes are for them. These inconsistencies could explain some errors we were having.
2. We noticed that not every patient has "pre_3" and/or "post_3". This causes difficulties in the calculations and some errors to occur (such as finding the

correlation). The sample sizes for these patients are not big enough, so we happen to not look too deep into these conditions.

3. The sample size seems to pose a problem as it may not be big enough to show statistical significance of the variables in the model.
4. For our Significant Variables, there is a problem being that a lot of the categories in the variables have a small sample size (n=1). This can cause our analysis to seem significant when it really isn't.

Recommendations for the future

1. We would recommend future students/researchers look into how the statistically significant variables might affect the ages of patient's looking younger or older by the AI models.
2. We would also recommend looking into the variables (if able to) of the lifestyles of the patients. For instance, if a patient is a heavy drinker, they might or might not appear older than they actually are before and after the surgery.