Final Project Feedback Opportunity 2 Assignment

(Team Members: louhad, akpavan, gvnsp)

Question: Describe the data set. What is the source of the data set? Who curated or created it? What was the original purpose of the data set?

- The dataset was de-identified and donated to the UCI Machine learning repository in 1995 for public use. It was curated to explore the relationship between malignancy and histo-cytological factors through a process called fine-needle aspiration (FNA). The dataset includes an indication of malignancy or benignity and ten variables related to the tumor's physical qualities.
- There are 569 tumor samples represented in the dataset. There are 212 malignant tumors and 357 benign tumors in the data set.
- Four variables – radius, perimeter, area, and texture – are presented in their standard units as measured by the FNA process.
- Six variables – compactness, smoothness, concavity, symmetry, and fractal dimension – are presented relative to their level of "irregularity." For these seven variables, the more regular the quality is assessed to be, the closer to "0" it will be. The more irregular the quality is assessed to be, the closer to "1" it will be.

What is the exposure/intervention variable you plan to look at? What is the outcome variable? Describe both in detail.

- The exposure variable of interest is texture. It is a measure of the variance of the tumor surface's grayscale intensity as measured by the FNA process.
- The outcome variable is malignancy. It is a categorical variable and very straightforward. If the value is marked as "M," the patient's tumor is cancerous. If the value is marked as "B," the patient's tumor is benign.

Describe your hypothesis. What do you expect to find and why?

Our question: In patients who undergo FNAC imaging for breast tumor screening, how does nuclear texture predict malignancy?

Background: Nuclear atypia can be seen in reactive, pre-neoplastic, and malignancy. Severe nuclear atypia is, in most cases, considered an indicator of malignancy. The texture of the nucleus is the variation of the grayscale intensities in the component pixels recognized by the image processing system.

Wolberg, W.H. et al. (1994) stated that nuclear texture could be used to differentiate between malignant and benign cells. They forecasted the accuracy to be between 95.5 percent and 98.5 percent at a 95% confidence level.

Recognizing the value of nuclear texture as a strong identifier, we intend to study its relation with other nuclear features extracted from the nuclear imaging of FNAC specimens. In doing so, we will study the confounders for nuclear texture and account for it in our machine learning model for predicting malignancy.

We expect to find an intricate relationship between the nuclear texture and some other nuclear features that are expected to affect nuclear morphology in some capacity. We anticipate that our machine learning model's performance will depend on this, and hence, we will conduct an in-depth analysis (unstratified and stratified) to formulate our model recipe.

Our hypothesis(*H*): In patients who undergo Fine Needle Aspiration Cytology (FNAC) of breast tumor mass, nuclear texture can be used to predict malignancy status.

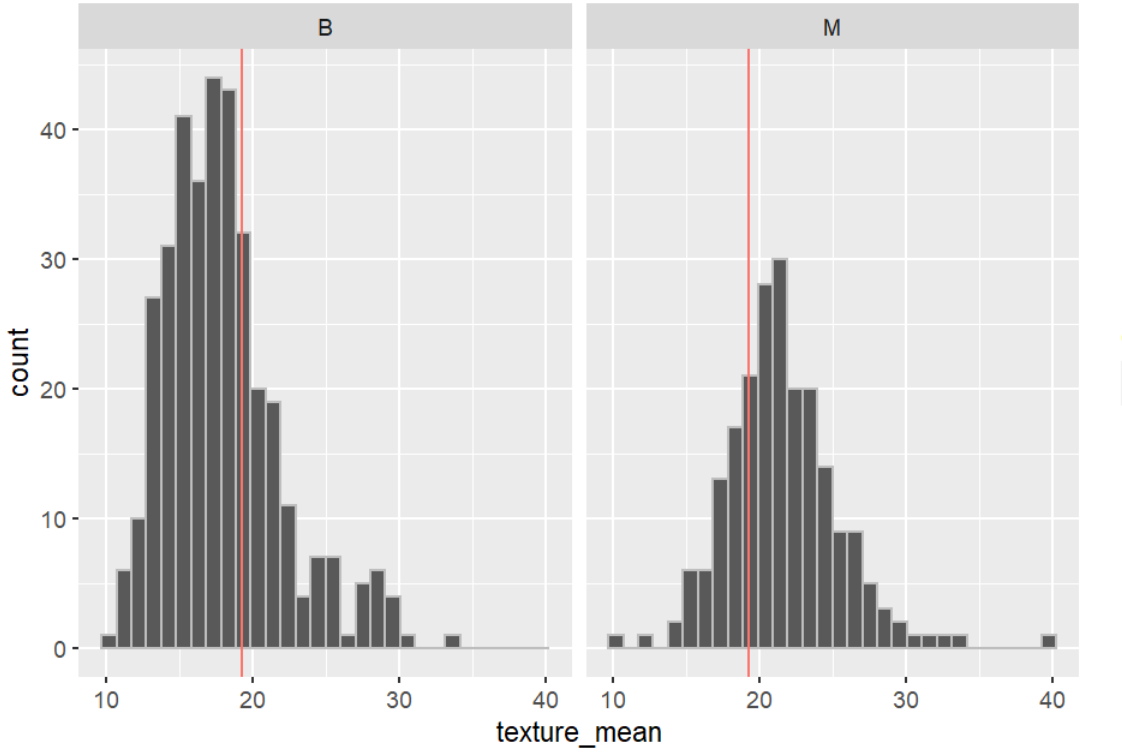
Null Hypothesis (*H0*): In patients who undergo Fine Needle Aspiration Cytology (FNAC) of breast tumor mass, the nuclear texture does not predict malignancy status.

Alternate Hypothesis (*H1*): In patients who undergo Fine Needle Aspiration Cytology (FNAC) of breast tumor mass, nuclear texture can predict malignancy status in some individuals.

Question: Describe two potential confounders that are present in the data set (variables that you think may be associated with both the exposure and the outcome). Why do you think these variables could potentially be confounders?

- Concavity: We have selected this to be our categorical confounding variable. While the texture is an indicator of how uniform the surface is, concavity examines a different external quality of the nucleus. It may be worth conducting a stratified analysis to explore its impact on our primary relationship of interest.
- Area: Nuclear area is measured simply by counting the number of pixels on the interior of the image reference point and adding one-half of the pixels in the perimeter. The nuclear area does directly correspond to the size of the cell, and we are of the opinion that it may impact the nuclear texture. Hence our interest in stratification with it to explore its impact on nuclear texture.

1. Unstratified Analysis



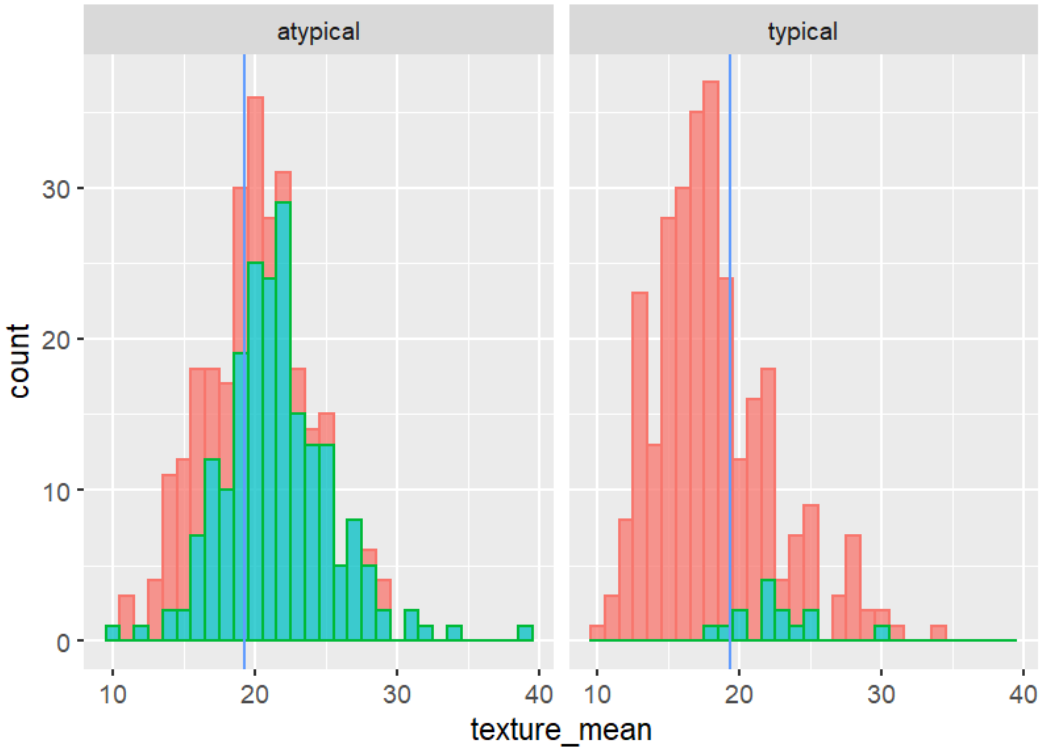
Interpretation: Tumors identified to be malignant (M) were more likely to have the nuclear texture values higher than the calculated mean for all tumors than tumors identified to be benign (B) were.

Population: Patients who are undergoing FNAC for breast cancer screening.

Exposure variable: Nuclear texture

Outcome variable: Malignancy status

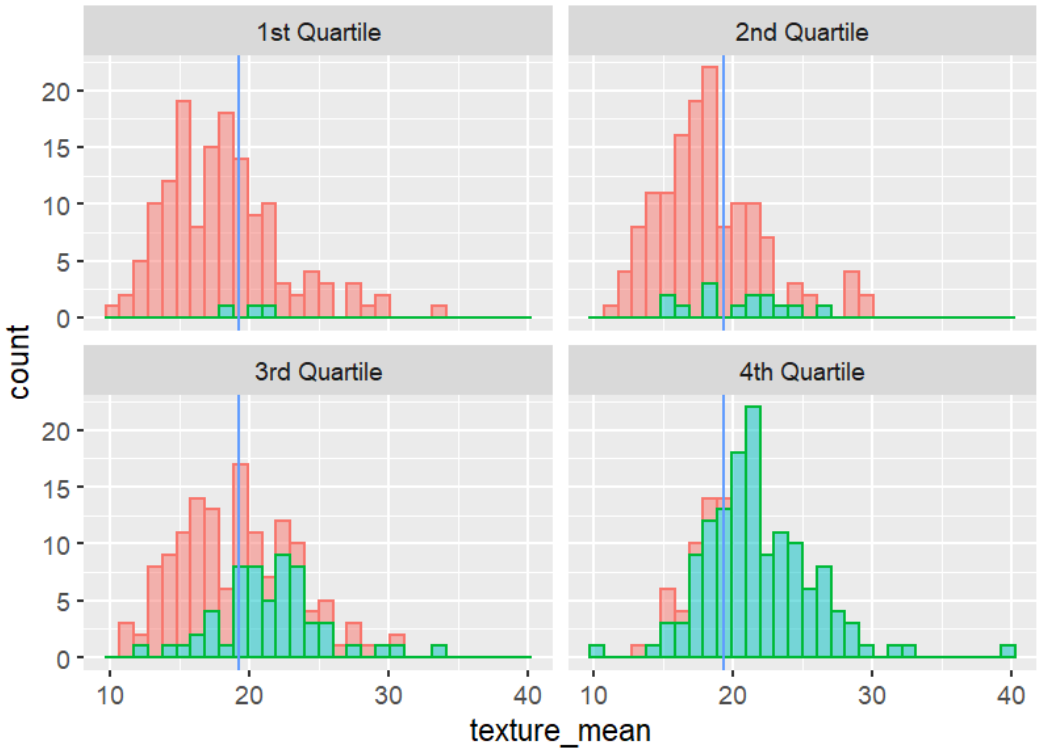
2. **<u>Stratified Analysis (Confounder 1 —Categorical- CONCAVITY)</u>**:



Key: Pink represents benign tumors. Blue represents malignant tumors.

Interpretation: Based on this stratified visualization, it appears that concavity could potentially be an impactful confounder on the relationship between nuclear texture and malignancy. As we can see, it appears that malignancy is significantly more represented among atypical concavity samples, and benignity is significantly more represented among typical concavity samples. Additionally, tumors with typical concavities are more likely to be below the texture mean, and tumors with atypical concavities are more likely to be above the texture mean.

3. **Stratified Analysis (Confounder 2 — Continuous–AREA):**



Key: Pink represents benign tumors. Blue represents malignant tumors.

Interpretation: Stratification enables the control for confounding variables by accounting categories. It will allow us to control confounders by creating two or more categories or subgroups in which the confounding variable either does not vary or does not vary very much.

Here, we studied the Area as a confounder (continuous type). Our analysis indicated that we were better able to group the benign and malignant tumors by stratification with the area. This is because the area was found to impact texture. The mean value of texture means was able to effectively distinguish between the two groups. The shift of values towards higher quartiles indicated a higher likelihood of malignancy.

Conclusions

How would you answer your specific health-related question, taking into account the unstratified and stratified analyses and considering both the Bradford-Hill causal criteria and the strengths and weaknesses of the data set you identified in Part 1 of the Final Project.

Based on the unstratified analysis, there appears to be a correlation between higher nuclear texture values and malignancy status. The variation in nuclear features is consistent on a cellular level for both benign and malignant tumors. Higher mean texture values coincide with malignancy status. Nuclear morphology changes are established to be associated with morphology based on a variety of past studies, so the results of our analysis make sense.

With that being said, after carrying out our continuous and categorical stratified analyses, we hesitate to make any kind of causal claim as it pertains to nuclear texture and malignancy also, we cannot ascertain the temporality of the relationship between nuclear texture and malignancy as our dataset does not contain such information.

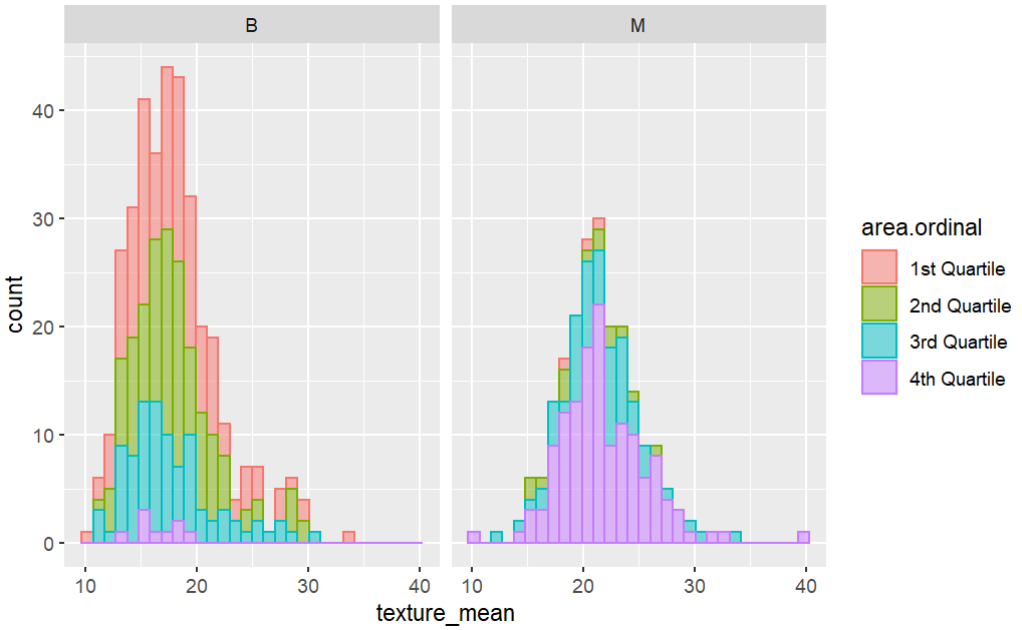
See Appendix B for Hill's Criteria key points.

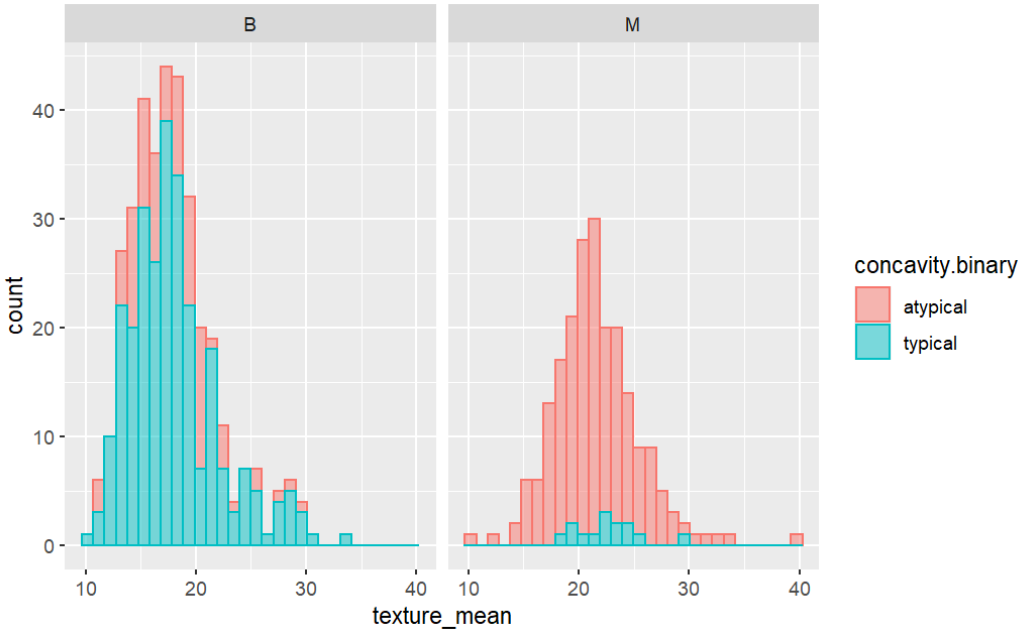
References

1. Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. *Biomedical Image Processing and Biomedical Visualization*, *1905*, 861–870. https://doi.org/10.1117/12.148698
2. Saha, M., Mukherjee, R., & Chakraborty, C. (2016). Computer-aided diagnosis of breast cancer using cytological images: A systematic review. *Tissue and Cell*, *48*(5), 461–474. https://doi.org/10.1016/J.TICE.2016.07.006
3. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine-needle aspirates. *Cancer Letters*, *77*(2–3), 163–171. https://doi.org/10.1016/0304-3835(94)90099-X
4. Wolberg, W. H., Street, W. N., Heisey, D. M., & Mangasarian, O. L. (1995). Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, *26*(7), 792–796. https://doi.org/10.1016/0046-8177(95)90229-5
5. Yu, Y. H., Wei, W., & Liu, J. L. (2012). Diagnostic value of fine-needle aspiration biopsy for breast mass: A systematic review and meta-analysis. *BMC Cancer*, *12*(1), 1–14. https://doi.org/10.1186/1471-2407-12-41/FIGURES/5

Appendix

Appendix A: Additional Visualizations of Stratified Analyses





Appendix B: Bradford Hill Criteria for our analysis.

- Strength

  Based on the unstratified analysis, there appears to be a correlation between higher nuclear texture values and malignancy status. With that being said, after carrying out our continuous and categorical stratified analyses, we hesitate to make any kind of causal claim as it pertains to nuclear texture and malignancy.

- Consistency

  The variation in findings is consistent on a cellular level for benign and malignant tumors.

- Specificity

  Higher the texture_mean values as the exposure coincide with malignancy status as the outcome variable.

- Temporality

  We cannot ascertain the temporality of the relationship between nuclear texture and malignancy as our dataset does not contain such information.

- Biological gradient

  Our unstratified analysis appears that the higher nuclear texture values correspond with a higher likelihood of malignancy status.

- Plausibility

  Nuclear morphology changes are established to be associated with morphology based on a variety of past studies, so the results of our analysis make sense.

- Coherence

  Irrespective of the sample extracted, nuclear morphology changes of breast cancer cells correlate with malignancy.

- Experiment

    If breast cancer is removed as a condition from the patient, their chances of survival will increase.

- Analogy

    The area of study related to finding a correlation between nuclear morphology, specifically texture, in Human breast cancer cells extracted through FNAC are very low