Ethan McFarlin: HBS Quantitative Exercise

— General Notes:

- Programming language of choice: R (Notebook)
- Directory Contents (/McFarlin Ethan HBS Quant Exercise)
    -

— Responses:

✅ 1. Load the CSV file into the program of your choice.

✅ 2. Conduct the following:

a. Rename AB84002 to acres_corn, and label it “Acres of harvested corn”
b. Rename AB84006 to acres_wheat, and label it “Acres of wheat”
c. Reduce the dataset to only the following variables: GISJOIN, YEAR, STATE, COUNTY, acres_corn, and acres_wheat

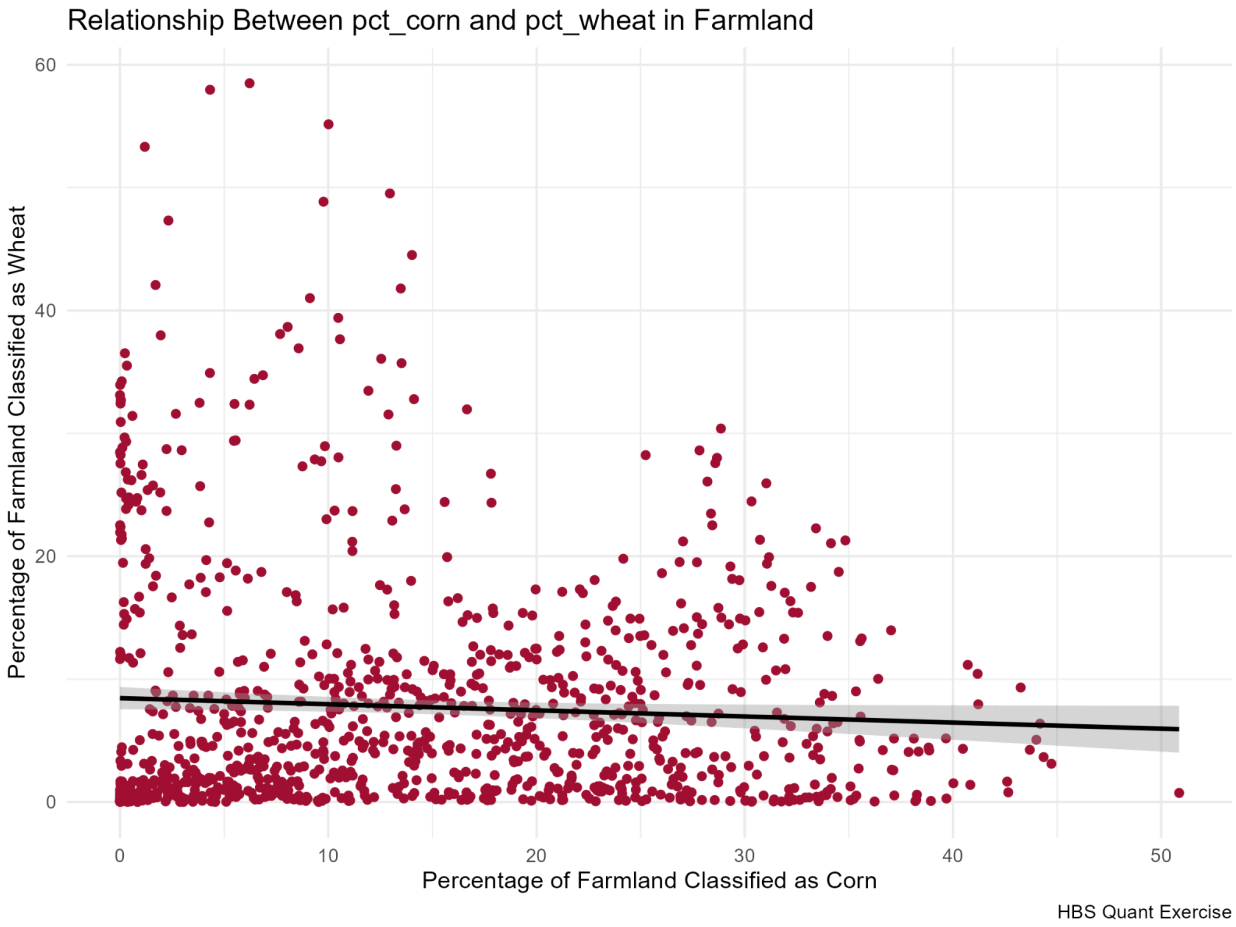
✅ 3. Merge the current dataset with the “test_mergefile.dta” data file, based on the unique identifier GISJOIN, and keep only the observations appearing in both the original and merged-in datasets. This will merge in three new variables: farms, acres_farmland, and tractors.

✅ 4. Create two new variables for the percentage of farmland in corn and the percentage of farmland in wheat, naming them “pct_corn” and “pct_wheat”, and label the variables appropriately.

✅ 5. Find and report the average pct_corn and average pct_wheat in three states: Kansas, Iowa, and Michigan.

State	avg_pct_corn	avg_pct_wheat
Iowa	26.09%	1.42%
Kansas	12.94%	20.27%
Michigan	3.11%	3.06%

✅ 6. Create a scatter plot of pct_corn against pct_wheat, including a line-of-best-fit.



✅ 7. Create a new variable tractors_per_farm equal to the number of tractors per farm, and label the variable appropriately.

✅ 8. Run a regression to examine the relationship between the independent variables pct_corn and pct_wheat and dependent variable tractors_per_farm.

✅ 9. Perform the same regression, but this time take state fixed effects into account.

✅ 10. Interpret the coefficients, significance, and goodness of fit for each of these regressions, and explain what the results indicate. For the second regression, include your interpretation of the state fixed effects.

First Regression: lm(formula = tractors_per_farm ~ pct_corn + pct_wheat, data = df6)	
Type of Analysis	Description
Residuals	The values of tractor_per_farm are mostly symmetric about the median, indicating a relatively close approximation.
Coefficients + Significance	Intercept: When the percentage of wheat and corn are both 0, we expect there to be approximately 0.0866 tractors per farm.
	pct_corn: For every 1% increase in the percentage of corn, we might expect the number of tractors per farm to increase by 0.0017989, indicating a positive relationship. The p-value (which is 4.47e-15) suggests that this is statistically significant.
	pct_wheat: For every 1% increase in the percentage of wheat, we might expect the number of tractors per farm to increase by 0.0047112, indicating a positive relationship. The p-value (which is < 2e-16) suggests that this is statistically significant.
Goodness of Fit	The RSE of 0.082325 can be interpreted as the average deviation between observed and fitted values The multiple R-Squared of 0.2462 tells us that 24.62% of variability in the number of tractors per farm may be accounted for by pct_corn and pct_wheat. The adjusted R-squared of 0.2448 allows us to adapt the R-squared value based on the number of predictors used. The f-statistic of 172.2 with p-value < 2.2e-16 implies

	overall statistical significance of the model.
Overall Interpretation	While it can reasonably assumed from the model that the % of farmland used for corn/wheat are both positively correlated with the number of tractors per farm (with statistical significance), the R-squared value indicates that a large degree of variation in the number of tractors per farm is unexplained by those two singular variables.
Second Regression: lm(formula = tractors_per_farm ~ pct_corn + pct_wheat + factor(STATE), data = df6)	
Residuals	Similar to the first regression, except there is a lower spread and better fit
Coefficients + Significance	Intercept: When the percentage of wheat and corn are both 0, we expect there to be approximately 0.099 tractors per farm.
	pct_corn: For every 1% increase in the percentage of corn, we might expect the number of tractors per farm to increase by 0.0024308, indicating a positive relationship. The p-value (which is < 2e-16) suggests that this is statistically significant.
	pct_wheat: For every 1% increase in the percentage of wheat, we might expect the number of tractors per farm to increase by 0.0045500, indicating a positive relationship. The p-value (which is < 2e-16) suggests that this is statistically significant.
	State-fixed effects: The states with significant coefficients show us which states different significantly from the reference with

	respect to the average number of tractors per farm. - Wisconsin: 0.0374631, significantly higher - South Dakota: 0.0490017, significantly higher - Ohio: -0.0373999: significantly lower - Nebraska: -0.0494244, significantly lower - Missouri: -0.1092375, significantly lower - Michigan: -0.0351221, significantly lower - Indiana: -0.0559878, significantly lower
Goodness of Fit	The RSE of 0.07186 is smaller than the first regression, implying the fit is slightly better. The multiple R-Squared of 0.4442 tells us that 44.42% of variability in the number of tractors per farm may be accounted for by pct_corn and pct_wheat. This appears to be a significant improvement. The adjusted R-squared of 0.4373 allows us to adapt the R-squared value based on the number of predictors used. Higher than the first model. The f-statistic of 64.12 with p-value < 2.2e-16 implies overall statistical significance of the model.
Overall Interpretation	Adding state fixed effects allows us to control for idiosyncratic characteristics of certain states which influence the prevalence of tractors. While pct_corn and pct_wheat are still positive/significant, the state fixed effects show us that several states differ drastically from the reference. Because the R-squared value is improved, the model accounts for twice as much variability, illustrating the importance of state-specific factors.

Overall reflections: The model with state effects offered a better fit and more in-depth insights on state-specific differences. As much as the first model encapsulates some variability in tractors per farm, it's the inclusion of state-level effects that is able to enhance the

explanatory power. In the bigger picture, this goes to show the importance of state-specific considerations when analyzing agricultural practices.