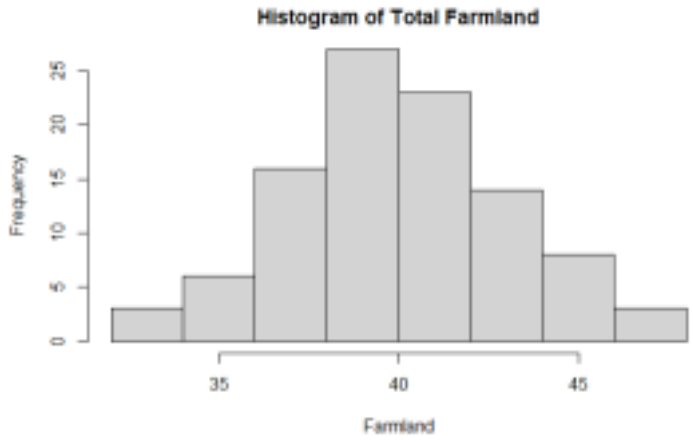
Causal Inference

We are interested in predicting the total crop yield for various farms across North California. We have data on 100 farms, and have recorded the following:

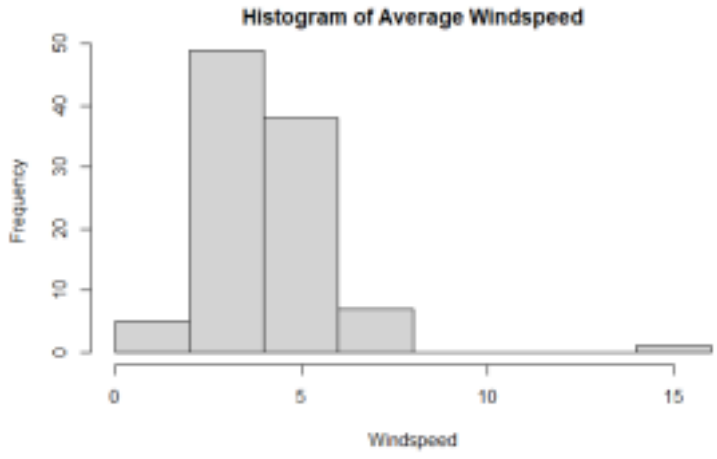
- Total Farmland (Acres)
- Average Wind Speed (MPH)
- Average Rainfall (In)
- Total Crop Yield (Bushels)
- County

We would like to use our data to predict the total crop yield. For now, we are going to ignore the county data and only focus on the numerical values. Before we do any modeling, we first want to look at the distribution of our variables:

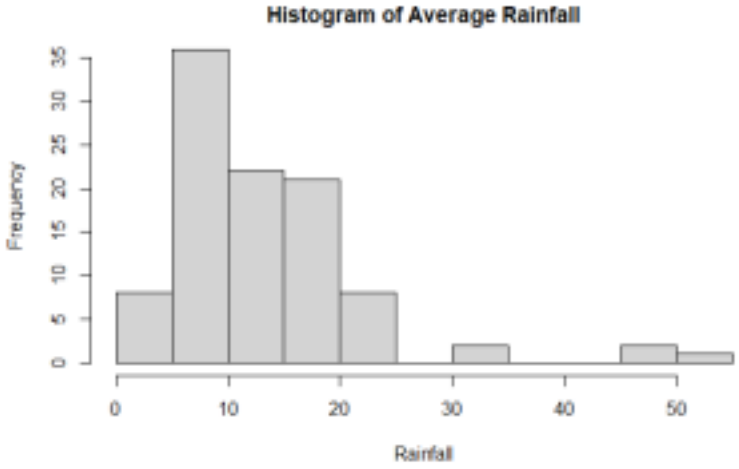
Farmland:



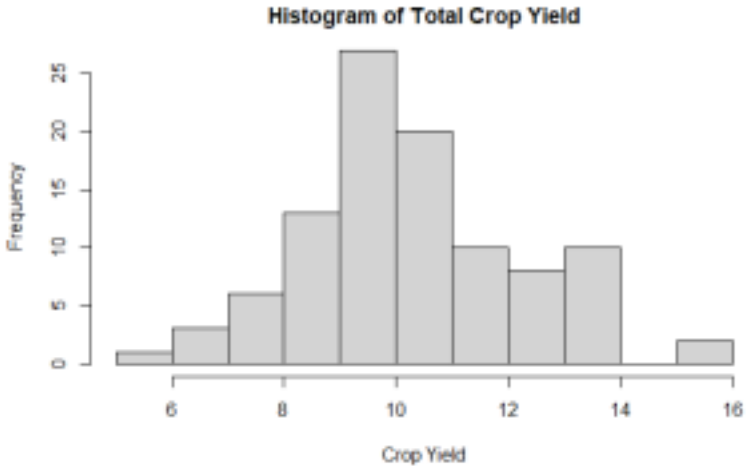
Windspeed:



Rainfall:



Crop Yield:



Question 1: Given the above distributions, what data cleaning and/or transformations would you recommend we do? Why would you make these changes and what would you expect the end result to look like?

I will do a logarithmic transformation for **Rainfall** and for **Wind Speed** because this will pull in the tail on the right and make the distribution more symmetric and standarization for all the data for the following reasons:
The distribution should be closer to normal, which is often an assumption or beneficial for many Ml algorithms. Also, all the data would be on a similar scale, which is crucial for Ml models that rely on distance; like k-NN or linear regression.

Assume we have implemented the changes you have suggested. We now run a linear regression on the new data and get the following results.

MODEL INFO:
Observations: 100
Dependent Variable: crop_yield
Type: OLS linear regression

MODEL FIT:
$F(3,96) = 131.87$, $p = 0.00$
$R^2 = 0.80$
Adj. $R^2 = 0.80$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	0.32	2.29	0.14	0.89
farmland	0.21	0.06	3.77	0.00
wind	0.05	0.09	0.54	0.59
rain	0.37	0.02	18.71	0.00

Question 2: How would you explain, in plain English, the meaning behind the coefficient on farmland.

For every one unit increase in the amount of farmland, we can expect the crop yield to increase by 0.21 units, **keeping all other factors constant**.

Question 3: Say we are going to use this model to predict crop yield for other farms in the same region. If we are worried about overfitting, what would be some potential improvements we could make? How could we test if these changes help prevent overfitting?

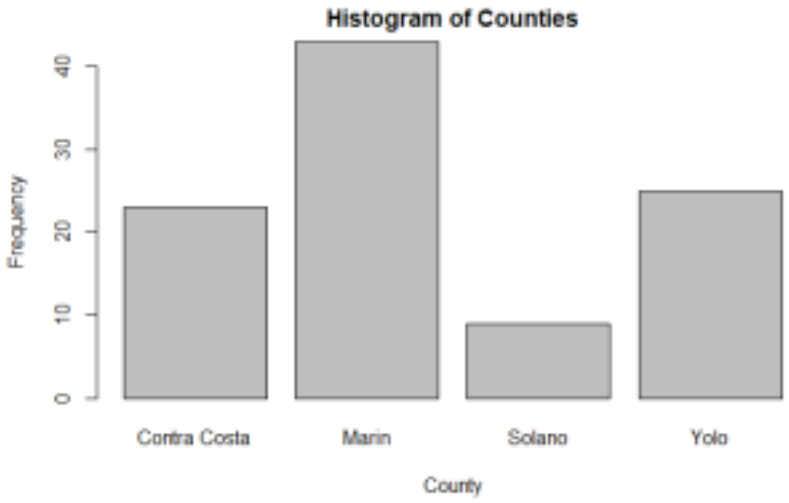
Potential improvements for prevent overfitting:

- Feature Selection/Engineering: Some features might be irrelevant or less relevant. By reducing the numbers of predictors or transforming them we can reduce the complexity of the model.
- Gather more data: More data can help in providing a better perspective and can make the model more robust.
- Use regularization techniques.

Test changes to prevent overfitting:

- Comparison with a Baseline: We can compare the performance of the model with a simpler model or a known baseline.
- Plot the model's learning curve performance on both the training and test sets. If the model is overfitting, the performance will be much better on the training set than the test set.
- Cross-validation: This will give us a better understanding of how the model performs across different subsamples of the data.

Now we are going to include the county data. First let's look at the distribution of the counties.



We will create a dummy variable is_yolo, which is 1 if the county is yolo, and 0 otherwise. We will now rerun the same regression as earlier but now with the is_yolo variable.

MODEL INFO:
Observations: 100
Dependent Variable: crop_yield
Type: OLS linear regression

MODEL FIT:
$F(4,95) = 288.34$, $p = 0.00$
$R^2 = 0.92$
Adj. $R^2 = 0.92$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	2.01	1.45	1.39	0.17
farmland	0.15	0.04	4.26	0.00
wind	-0.03	0.06	-0.58	0.57
rain	0.39	0.01	31.36	0.00
is_yolo	3.05	0.25	12.20	0.00

Question 4: How would you explain, in English, the meaning behind the coefficient on is_yolo. If we were interested in further exploring the effect of the county on crop yield, what would you recommend trying?

If two farms are similar in every aspect such as the size of the farmland, the amount of rain, and wind,

but one is located in Yolo county and the other is not, we would expect the farm in Yolo county to have a crop yield that is 3.05 units higher.
Instead of only creating a dummy variable for Yolo, we could create dummy variables for each of the other counties (e.g., is_marin, is_solano, is_contra_costa). This would allow us to compare the effect of each county on crop yield.

Question 5: Currently we have a fairly simple approach to predicting crop yield. If we wanted to make our model better, what would be some potential improvements? What else in the data would you want to look at in order to better the model? How would we be able to measure if our changes were improving the model?

To improve our model capacity we could think in adding more features like: Soil quality, temperature data, pest/disease information can all provide valuable information.
We could check for potential biases; outliers or influential data points which can disproportionately impact the fit of a linear regression.
For better our model we could do a correlation analysis and keep the features that have a strong and positive correlation with our target variable crop_yield.
And to measure if our changes improve the model we could check the model evaluation metrics like R^2, Adjusted R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). An increase in R^2 or a decrease in RMSE on the test set indicates improvement.

SQL

Imagine a table of all employees (called 'Employees'), with the following fields:

- id (num)
- company_id (num)
- name (char)
- salary (num)
- department (char)
- hiring_date (datetime)

And another table of firms (called 'Firms')

- id (num)
- name (char)
- industry (char)
- date (datetime)
- earnings (num)

I need you to do a few data pulls for me:

1. The names of all of the different firms

```
SELECT DISTINCT name FROM Firms;
```

2. All employees whose name starts with the letter ‘A’ hired before March 14th of this year

```
SELECT * FROM Employees
WHERE name LIKE 'A%' AND hiring_date < '2023-03-14';
```

3. Total earnings by industry after February 21st, 2020

```
SELECT industry, SUM(earnings) as total_earnings
FROM Firms
WHERE date > '2020-02-21'
GROUP BY industry;
```

4. The names of the top 40 highest-salaried employees working for the firm ‘ToysRUs’

```
SELECT E.name
FROM Employees E
INNER JOIN Firms F ON E.company_id = F.id
WHERE F.name = 'ToysRUs'
ORDER BY E.salary DESC
LIMIT 40;
```

5. The name of the third-highest-salaried employee for each firm in the ‘FMCG’ industry

```
WITH RankedEmployees AS (
SELECT E.name, E.company_id, RANK() OVER (PARTITION BY E.company_id
ORDER BY E.salary DESC) as rank
FROM Employees E
INNER JOIN Firms F ON E.company_id = F.id
WHERE F.industry = 'FMCG'
)
SELECT name
FROM RankedEmployees
```

```
WHERE rank = 3;
```