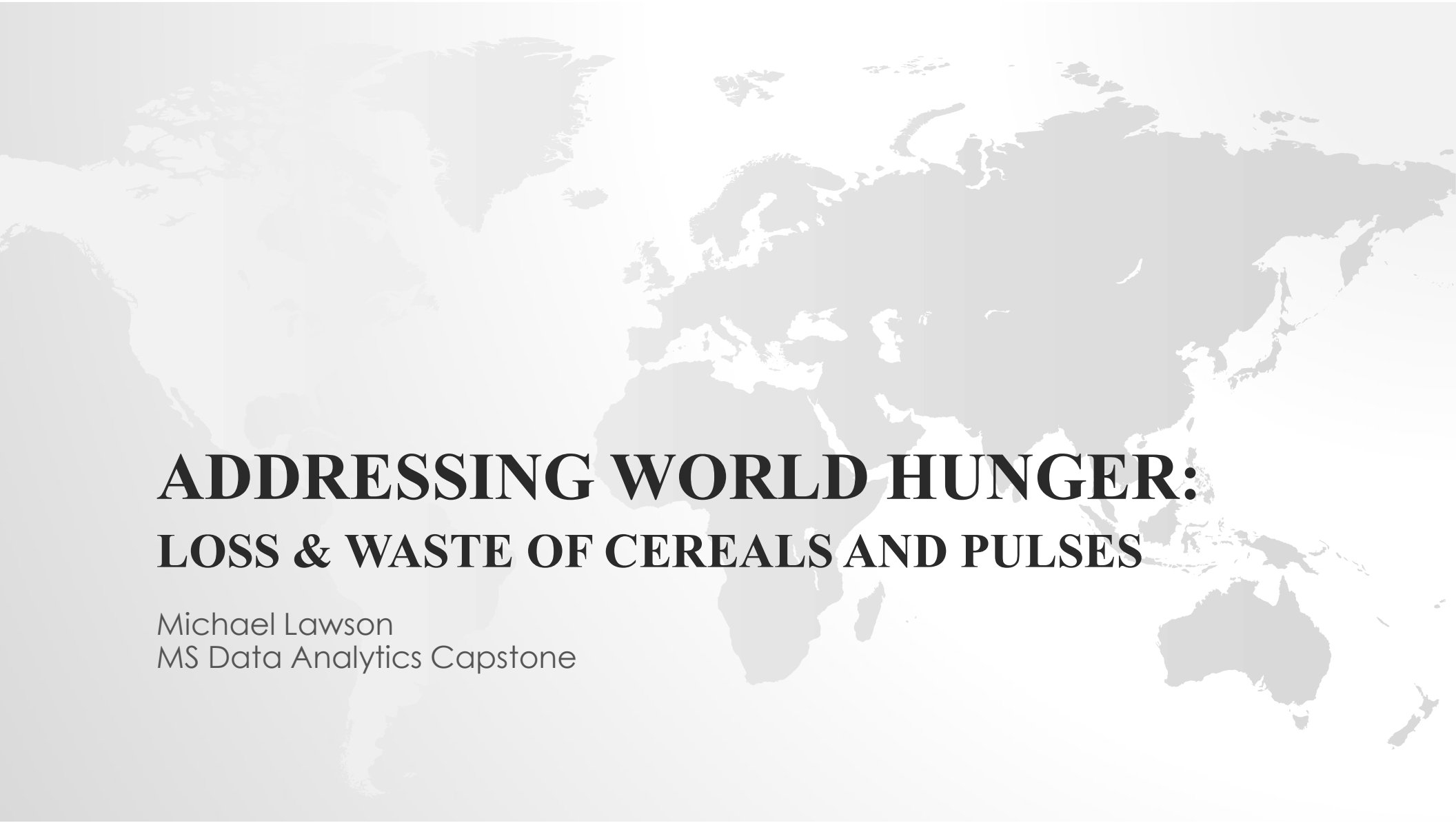


ADDRESSING WORLD HUNGER:
LOSS & WASTE OF CEREALS AND PULSES

Michael Lawson
MS Data Analytics Capstone

STATEMENT OF THE PROBLEM

Question:

- Is there a statistically significant difference in the percentage of food loss between the farm, harvest, and retail food supply stages?

Hypothesis:

- **H_0:** The percentage of food loss does NOT have statistically significant differences between the food supply stages ‘Farm’, ‘Harvest’ and ‘Retail’.

- **H_A:** The percentage of food loss does have statistically significant differences between the food supply stages ‘Farm’, ‘Harvest’ and ‘Retail’.

DATA COLLECTION

Food and Agriculture Organization of the United Nations

Year Range
1965 2000 2022
1965 1971 1977 1983 1989 1995 2001 2007 2013 2022

Aggregation
WORLD

Aggregation Options
All

Country
All

Basket Items
All

Commodity (CPC 2.0)
All

Value Chain Stage(s)
All

Method of Data Collection
All

Select to keep only top-10 SDG baskets

Select to show also groups and other broadly defined products

Commodity Basket Aggregation
Production Value- Top 10 by country (Default SDG)

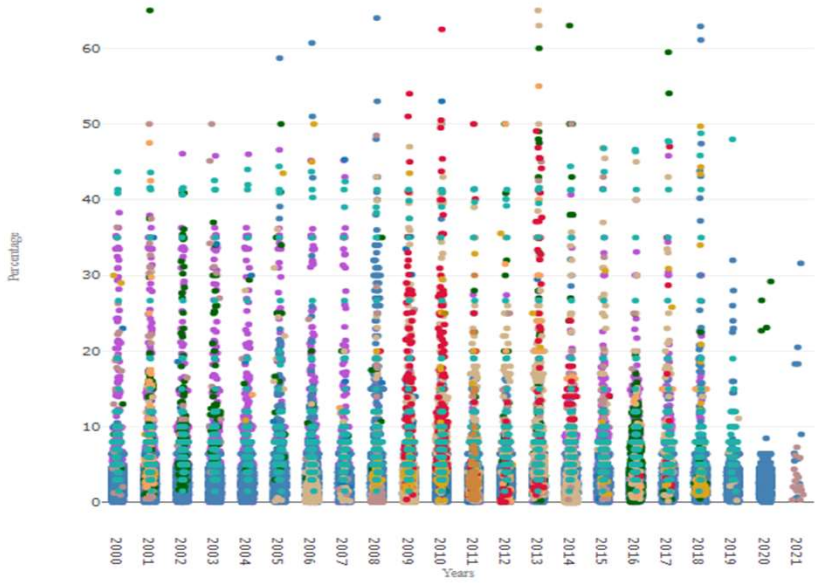
Download Data | Plots

Authenticate

Hide/Show Filters

Plot of loss percentages | Heatmap of available data | Boxplot by stage | Data

Food loss as a percentage of domestic production.



Source: FAO
Date: 2022-10-25 12:14:26
Aggregation: WORLD
Aggregation Option: All
Country : All
Commodity Aggregation: All

DATA ANALYSIS PROCESS

Cleaning & prep:

1. Load Python libraries appropriate for data visualization and regression
2. Load the data using read_csv()
3. Examine the header to see what column names and the values they contain using .head()
4. Examine the shape, dtype, and all column names using .info()
5. Create the reduced data set
6. If null values exist, treat them.
7. Combine categories of 'food_supply_stage' into broader categories: 'Farm', 'Harvest' & 'Retail'
8. Select a sample of the same size from each group
9. Visualize the distribution of the target feature 'loss_percentage'
10. Normalize 'loss_percentage'

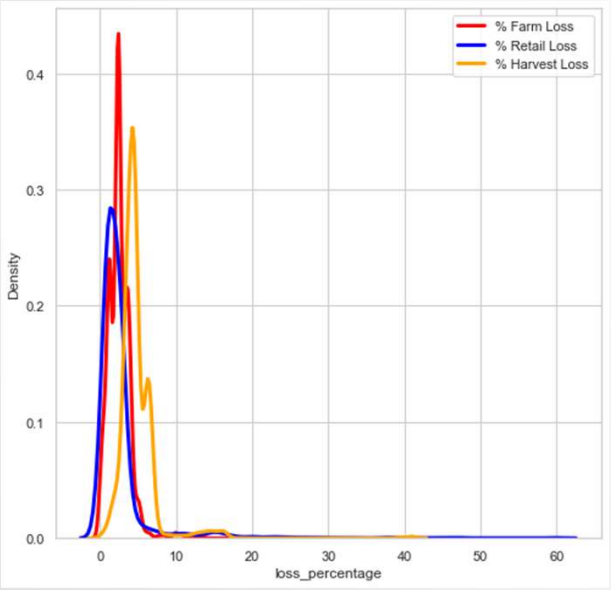
Analysis:

1. Perform 2-Way ANOVA
2. Use Tukey method to determine which pairs are significantly different
3. Create boxplots to compare means and variability of loss percentage
4. Create Q-Q plot to test normality of data (since ANOVA assumes normality)
5. Check for equality of variances of the treatments using the Levene test
6. Export clean data set

PYTHON PACKAGES & LIBRARIES

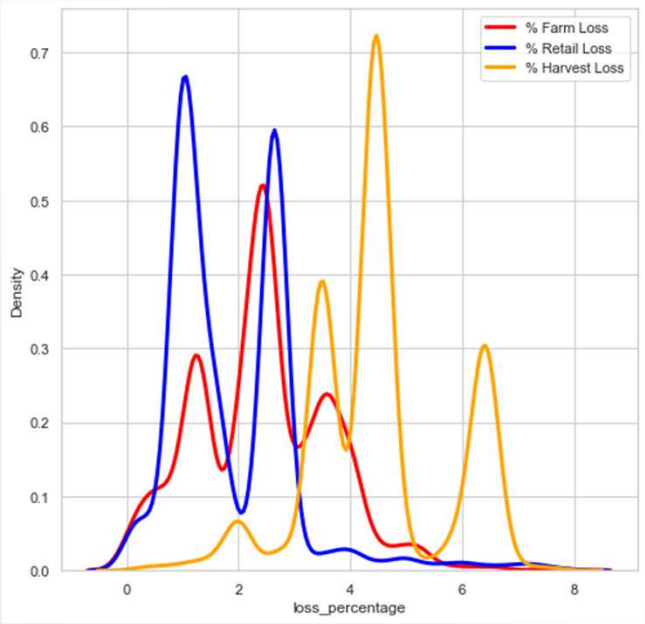
Library	Package	Notes
Pandas		fast and powerful data analysis and manipulation library.
Numpy		wide range of math functions
statsmodels	Formula, stats	Python module that provides classes and functions for the estimation of many different statistical models
Matplotlib	pylab	convenience module that bulk imports matplotlib.pyplot (for plotting) and NumPy (for Mathematics and working with arrays) in a single name space (Tutorialspoint, n.d.)
Matplotlib	pyplot	visualizations of data
seaborn		visualizations of data
scipy		equations and algorithms
statsmodels		provides classes and functions for the estimation of many different statistical models
bioinfokit	analys	easy-to-use functionalities to analyze, visualize, and interpret the biological data
warnings	filterwarnings	loaded to remove filter warnings

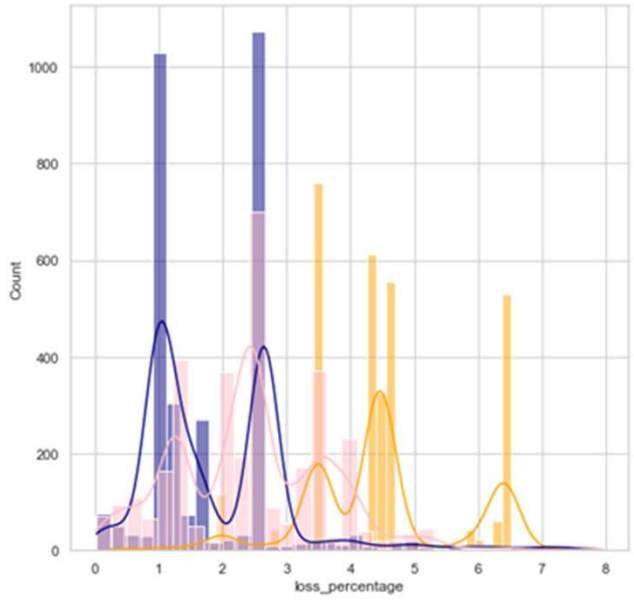
NORMALIZE THE DISTRIBUTION



The distribution is skewed to the right until outliers are removed.

The reduced data set has a normal distribution





ONE-WAY ANOVA

ANOVA MODEL & TUKEY METHOD

```
model = smf.ols('loss_percentage ~ food_supply_stage', data=df).fit()
aov_table = anova_lm(model, typ=2)
print(aov_table)
```

```
                      sum_sq      df          F  PR(>F)
food_supply_stage  12102.187175     2.0  4293.92311     0.0
Residual           13869.569056  9842.0        NaN     NaN
```

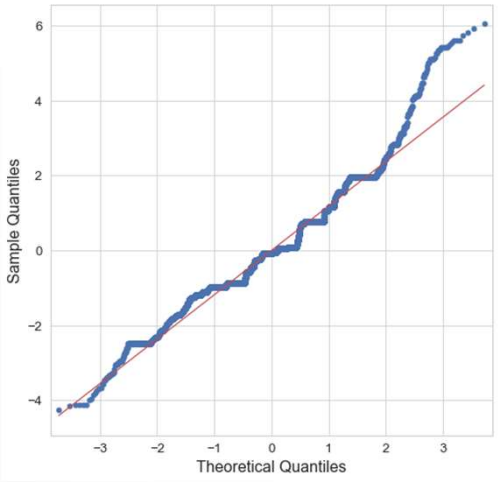
In [22]:

```
# Which stage is significantly different using Tukey method
mcStage = multi.MultiComparison(df['loss_percentage'], df['food_supply_stage'])
results_stage = mcStage.tukeyhsd()
print(results_stage.summary())
```

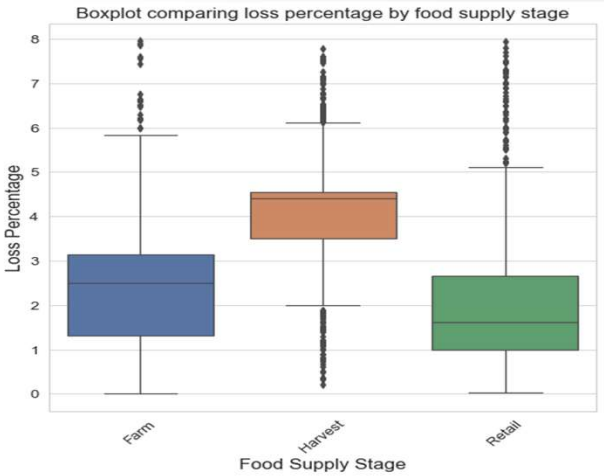
```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
 group1  group2 meandiff p-adj  lower   upper  reject
-------------------------------------------------
   Farm Harvest   2.0351 0.001  1.9667  2.1035   True
   Farm  Retail  -0.5553 0.001 -0.6239 -0.4867   True
Harvest  Retail  -2.5904 0.001 -2.6595 -2.5212   True
-------------------------------------------------
```

REJECT THE NULL HYPOTHESIS

Q-Q Plot



Boxplot



Levene Test

```
# Check for the equality of variances of the treatments using levene test
# The p-value refers to the significance of variation, so a p-value > .05 means variance is equal and ANOVA model is ok
farm3 = df['loss_percentage'][df['food_supply_stage']=='Farm']
harvest3 = df['loss_percentage'][df['food_supply_stage']=='Harvest']
retail3 = df['loss_percentage'][df['food_supply_stage']=='Retail']
(test_statistic, p_value) = stats.levene(farm3,harvest3, retail3)
print("The test statistic is: ", round(test_statistic,5))
print("The P-value is: ", round(p_value,5))
```

```
The test statistic is:  2.88057
The P-value is:  0.05615
```

LIMITATIONS OF ANALYSIS

- ANOVA assumes data has normal distribution
- ANOVA assumes data has equal variance
- Database has limited features
- ANOVA reveals at least one pair of means is significantly different, but not which pair

PROPOSED ACTIONS

- Collect more data
- Dig deeper into what is causing differences in means
- Seek out why more loss happens during harvest

NEXT STEPS

- What is the relationship in the loss during harvest and farming
- Break down the food supply stages for more detailed analysis
- Run new models on other commodities in the database and compare

BENEFITS OF ANALYSIS

- Resources such as water, land, fertilizer, employee wages can be directed at producing more food if the resources aren't wasted with the loss of food.
- Prevention of food loss and waste
- Give stakeholders a direction to seek answers

RESOURCES

- FAO. (n.d.). *Food Loss and Waste Database.*
- https://www.fao.org/platform-food-loss-waste/flw-data/en/
- FAO. (2022, September 29). *Stop Food Loss and waste, for the people, for the planet*.
- https://www.un.org/en/observances/end-food-waste-day
- FAO. (2015). FAO Strategy for Partnerships with Civil Society Organizations.
  - https://sustainabledevelopment.un.org/index.php?page=view&type=400&nr=2213&menu=1515
- Bedre, Renesh. (2022, March 6) ANOVA using Python (with examples).
- https://www.reneshbedre.com/blog/anova.html
- Tutorialspoint. (n.d.) *Matplotlib – Pylab module.*
- https://www.tutorialspoint.com/matplotlib/matplotlib_pylab_module.htm
- Conde, Ximena. (2022, October 19). *They planned a three-day giveaway. But Philadelphians claimed 300,000 free avocados in less than 3 hours.* https://www.inquirer.com/news/philadelphia/avocado-free-giveaway-fdr-park-philadelphia-20221019.html
- Luc Z. (2020, June 26) *One way ANOVA in Python 3.* https://www.youtube.com/watch?v=s6eZ806dqkI