W203 Lab 2

w203: Lab 2 Pumpkin and weather impact

w203 team 1: Kisha Kim, Patricia Gallagher, Sean Koval, Eric Ellestad

1. An Introduction

Giant Pumpkin competitions have grown in popularity over time and have evolved into a highly competitive sport among farmers with substantial cash prizes and even an organizing body called the Great Pumpkin Commonwealth (GPC) that establishes standards, eligibility requirements, and hosts regional competitions that enable qualification into an annual national championship. While giant pumpkins were the original and are the largest and most popular GPC category, other plant species from the Cucurbitaceae ("Cucurbit") family can be entered which includes field pumpkins, squashes, gourds, and watermelons. A niche but passionate community of growers, researchers, and aficionados have developed novel seed genetics as well as cultivation techniques in order to maximize the harvest weight of this family of gourds, with champion pumpkins frequently weighing in at over 2,000 lbs.

While there are many factors that a grower can control that contribute to the harvest weight of a pumpkin and other Cucurbits such as choice of seed genetics, cultivation method, and the skill of the grower, one of the most important factors in pumpkin plant growth and fruit production is environmental conditions such as temperature and precipitation. The goal of this research study is to investigate the impact of environmental factors on the weight of Cucurbits entered into official GPC competitions in the United States and determine whether certain climate conditions, and therefore geographic location, have an advantage over others when choosing where to grow a competitive giant pumpkin or Cucurbit.

Our research question is:

> **What impact do key environmental factors such as temperature and precipitation have on the weight of giant pumpkins and other Cucurbits grown in the United States that were entered into official GPC competitions?**

In order to answer this question, we evaluated all competitive entries in official GPC competitions from 2013 - 2021 and modeled the official competition weight of the pumpkin or Cucurbit as a function of the weather conditions it was grown under. Given weather is a complex system and pumpkins are grown over many months, we considered several different aggregated weather metrics such as maximum daily temperature, minimum daily temperature, amount of daily precipitation.

According to the Penn State College of Agricultural Sciences

> *Pumpkins are very sensitive to cold temperatures (below 50°F) and plants and fruit will exhibit injury from even a slight frost. The best average temperature range for pumpkin production during the growing season is between 65 and 95°F; temperatures above 95°F or below 50°F slow growth and maturity of the crop. Pumpkins require a constant supply of available moisture during the growing season. Water deficiency or stress, especially during the blossom and fruit set periods, may cause blossoms and fruits to drop, resulting in reduced yields and smaller-sized fruits.*

Given this, we looked at maximum and minimum daily temperatures, their average and variability over the growing season, as well as metrics that track the frequency of extreme temperatures. We will also look at the amount of daily precipitation, its average and variability over the growing season.

2. A description of the Data and Research Design

Causal Model and Research Design

Based on horticultural literature and practical pumpkin growing guides, we identified a number of factors that impact the growth rate, fruit quality, and ultimately the size the of the harvested fruit. Given there was limited data captured regarding the practices used in production for pumpkins in the GPC dataset, our analysis focuses on the environmental factors such as temperature and precipitation. We analyze potential omitted variable bias in section 5.

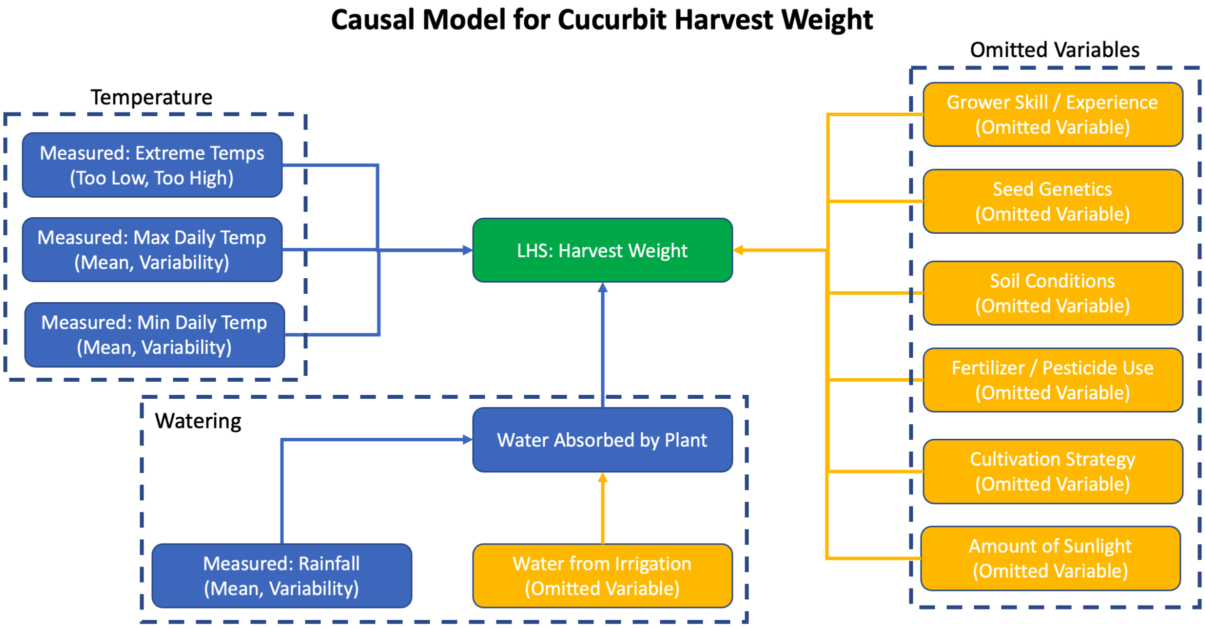

Causal Model for Cucurbit Harvest Weight
Temperature
Measured: Extreme Temps (Too Low, Too High)
Measured: Max Daily Temp (Mean, Variability)
Measured: Min Daily Temp (Mean, Variability)
LHS: Harvest Weight
Watering
Water Absorbed by Plant
Measured: Rainfall (Mean, Variability)
Water from Irrigation (Omitted Variable)
Omitted Variables
Grower Skill / Experience (Omitted Variable)
Seed Genetics (Omitted Variable)
Soil Conditions (Omitted Variable)
Fertilizer / Pesticide Use (Omitted Variable)
Cultivation Strategy (Omitted Variable)
Amount of Sunlight (Omitted Variable)

Figure 1: Cucurbit Harvest Weight Causal Relationships

Pumpkin/Cucurbit Data

We obtained official GPC weights of pumpkins, tomatoes, squashes, and watermelons from the Giant Pumpkins data set which was pulled from www.bigpumpkins.com. This data set included the city and state it was grown in, the year of the competition, and the type of the Cucurbit (Giant Pumpkin, Field Pumpkin, Giant Squash, Giant Watermelon, or Tomato).

City Coordinate Data:

City coordinate data was obtained from the World Cities Database on Kaggle. This data was used to cross reference cities in the official GPC weight dataset and determine their latitude and longitude.

Weather Data:

Weather data was obtained from National Oceanographic and Atmospheric Association Climate Data which provided daily minimum temperature, daily maximum temperature, daily rainfall, and other climate data from weather stations across the United States. The coordinates of each pumpkin/Cucurbit was compared to the coordinates of all weather stations to determine its weather station. The relevant weather metrics were then aggregated over the relevant growing season for the relevant year and associated with that pumpkin/Cucurbit record.

Dataset Preparation and Integration

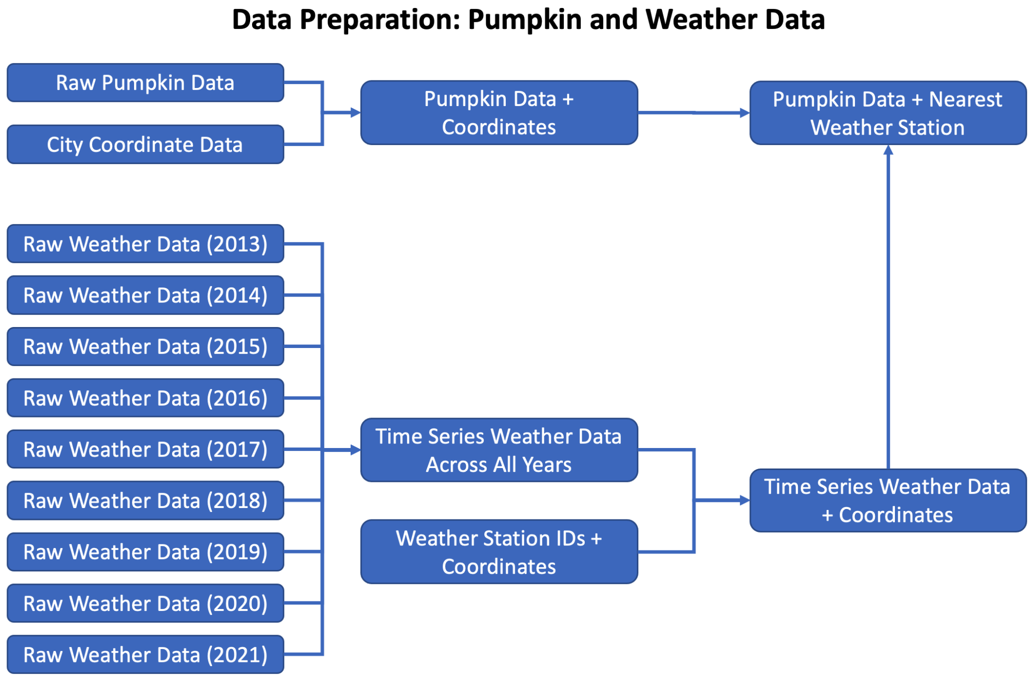
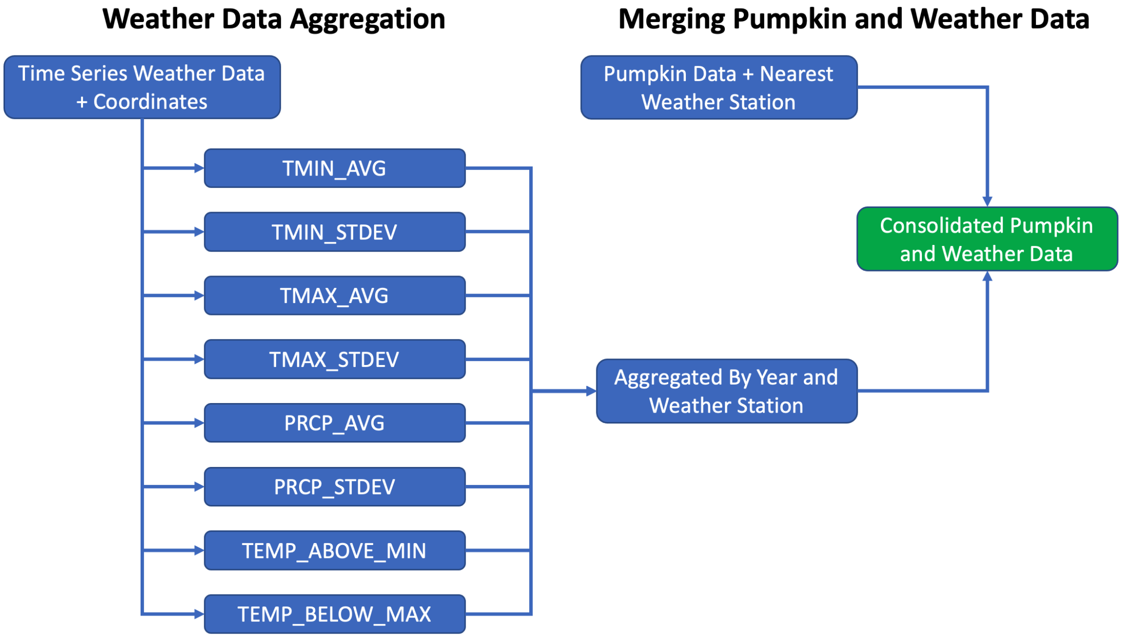


Figure 2: Pumpkin and Weather Data Preparation

The Pumpkin Dataset included city and state the pumpkin was grown, so we merged this with a City Coordinate Dataset to bring latitude and longitude data into the Pumpkin data. The weather data was downloaded by year and included weather stations all over the world, so we filtered down to just stations in the United States and combined with a weather station coordinate supporting dataset in order to bring latitude and longitude data into the weather data.

Once we had latitude and longitude data in both our Pumpkin and Weather data, we ran a search algorithm that identified for each pumpkin what its nearest weather station was. We then filtered the weather dataset down to just include weather stations marked as nearest a pumpkin record. Then we were ready to aggregate the time series weather data into our target explanatory weather variables.


Weather Data Aggregation
Merging Pumpkin and Weather Data
Time Series Weather Data + Coordinates
TMIN_AVG
TMIN_STDEV
TMAX_AVG
TMAX_STDEV
PRCP_AVG
PRCP_STDEV
TEMP_ABOVE_MIN
TEMP_BELOW_MAX
Pumpkin Data + Nearest Weather Station
Consolidated Pumpkin and Weather Data
Aggregated By Year and Weather Station

#### Aggregating Weather Data into Target Explanatory Variables:

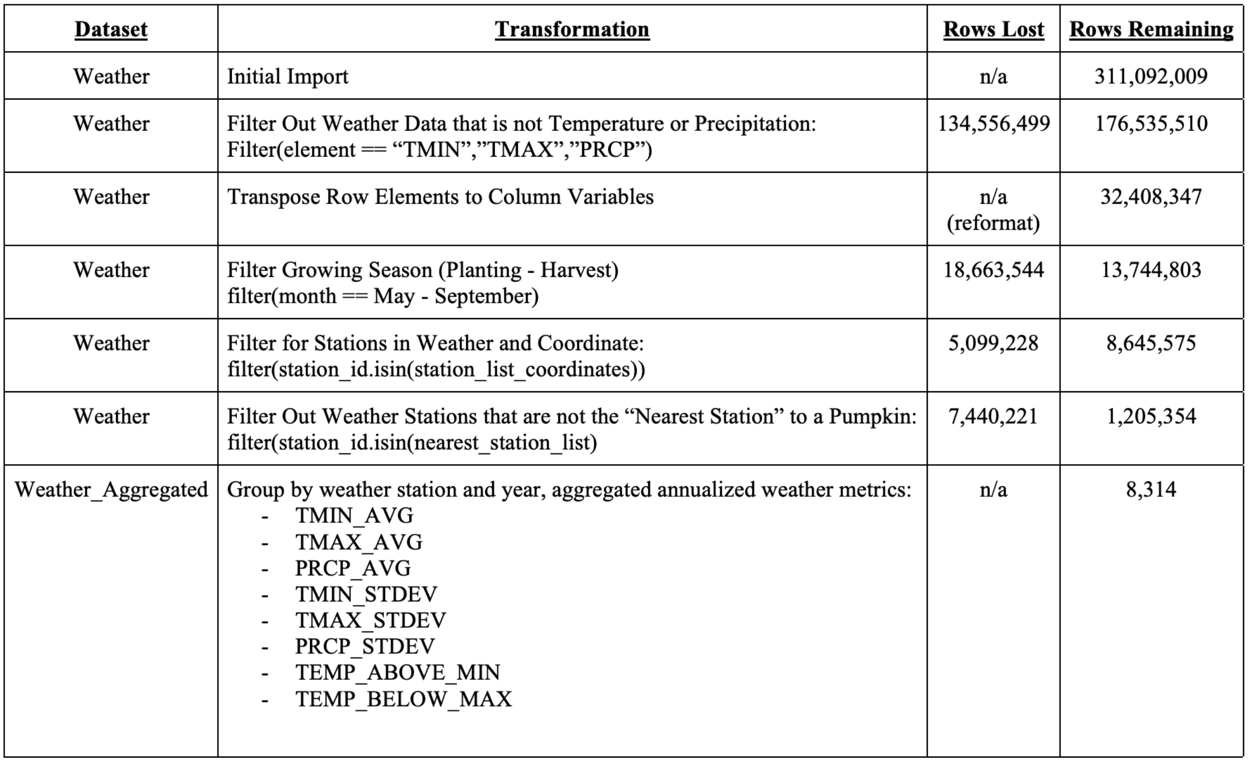
Once we had filtered the weather data down to measurements of relevant from stations near pumpkin records and subsetted based on the growing season, we calculated the summary environmental variables that we used in the regression mnodels:

- `TMIN_AVG`
  - A `TMIN_AVG` value was caculated for each weather station for each year of pumpkin data (2013 - 2021). This variable is the average of the minimum daily temperature over the months of the growing season (May through September) for the given year. The units are in tenths of a degree celsius, so 200 represents 20.0 degrees C.
- `TMAX_AVG`
  - A `TMAX_AVG` value was caculated for each weather station for each year of pumpkin data (2013 - 2021). This variable is the average of the maximum daily temperature over the months of the growing season (May through September) for the given year. The units are in tenths of a degree celsius, so 200 represents 20.0 degrees C.
- `PRCP_AVG`
  - A `PRCP_AVG` value was caculated for each weather station for each year of pumpkin data (2013 - 2021). This variable is the average of the daily precipitation over the months of the growing season (May through September) for the given year. The units are in tenths of a millimeter of precipitation so 100 represents 10mm of rain.
- `TMIN_STDEV`
  - A `TMIN_STDEV` value was caculated for each weather station for each year of pumpkin data (2013 - 2021). This variable is the standard deviation of the minimum daily temperature over the months of the growing season (May through September) for the given year. The units are in tenths of a degree celsius, so 200 represents 20.0 degrees C.
- `TMAX_STDEV`

  - A TMAX_STDEV value was caculated for each weather station for each year of pumpkin data (2013 - 2021). This variable is the standard deviation of the maximum daily temperature over the months of the growing season (May through September) for the given year. The units are in tenths of a degree celsius, so 200 represents 20.0 degrees C.

- PRCP_STDEV
  - A PRCP_STDEV value was caculated for each weather station for each year of pumpkin data (2013 - 2021). This variable is the standard deviation of the daily precipitation over the months of the growing season (May through September) for the given year. The units are in tenths of a millimeter of precipitation so 100 represents 10mm of rain.
- TEMP_BELOW_MAX
  - A TEMP_BELOW_MAX value was caculated for each weather station for each year of pumpkin data (2013 - 2021). Given temperatures for ideal growing conditions based on the literature are below 95F / 35C, this variable is a count of the number of days in a given growing season (May - September) that the maximum daily temperature did NOT exceed 35C (TMAX <= 350).
- TEMP_ABOVE_MINN
  - A TEMP_ABOVE_MINN value was caculated for each weather station for each year of pumpkin data (2013 - 2021). Given temperatures for ideal growing conditions based on the literature are above 65F / 18.3C, this variable is a count of the number of days in a given growing season (May - September) that the maximum daily temperature exceeded 18.3C (TMAX >= 183).

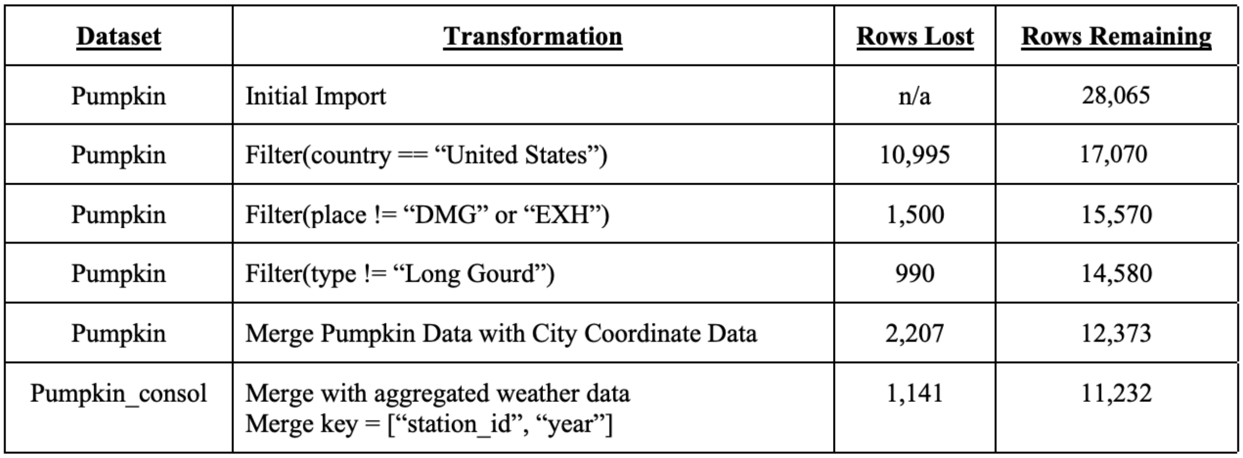
Data Cleaning and Accounting Tables



Dataset	Transformation	Rows Lost	Rows Remaining
Weather	Initial Import	n/a	311,092,009
Weather	Filter Out Weather Data that is not Temperature or Precipitation: Filter(element == "TMIN","TMAX","PRCP")	134,556,499	176,535,510
Weather	Transpose Row Elements to Column Variables	n/a (reformat)	32,408,347
Weather	Filter Growing Season (Planting - Harvest) filter(month == May - September)	18,663,544	13,744,803
Weather	Filter for Stations in Weather and Coordinate: filter(station_id.isin(station_list_coordinates))	5,099,228	8,645,575
Weather	Filter Out Weather Stations that are not the "Nearest Station" to a Pumpkin: filter(station_id.isin(nearest_station_list)	7,440,221	1,205,354
Weather_Aggregated	Group by weather station and year, aggregated annualized weather metrics: - TMIN_AVG - TMAX_AVG - PRCP_AVG - TMIN_STDEV - TMAX_STDEV - PRCP_STDEV - TEMP_ABOVE_MIN - TEMP_BELOW_MAX	n/a	8,314

Figure 3: Weather Cleaning, Filtration, and Aggregation Accounting Table

The weather dataset included many other measurements above and beyond temperature and precipitation, but unfortunately they were only collected at a very small subset of stations which would have reduced the granularity of our localized weather data to meaninglessness. We selected variables that were relevant to the causal model and also had robustness to their completeness and availability.



Dataset	**Transformation**	**Rows Lost**	**Rows Remaining**
Pumpkin	Initial Import	n/a	28,065
Pumpkin	Filter(country == “United States”)	10,995	17,070
Pumpkin	Filter(place != “DMG” or “EXH”)	1,500	15,570
Pumpkin	Filter(type != “Long Gourd”)	990	14,580
Pumpkin	Merge Pumpkin Data with City Coordinate Data	2,207	12,373
Pumpkin_consol	Merge with aggregated weather data Merge key = [“station_id”, “year”]	1,141	11,232

Figure 4: Pumpkin Cleaning, Filtration, and Merging with Weather Data Accounting Table

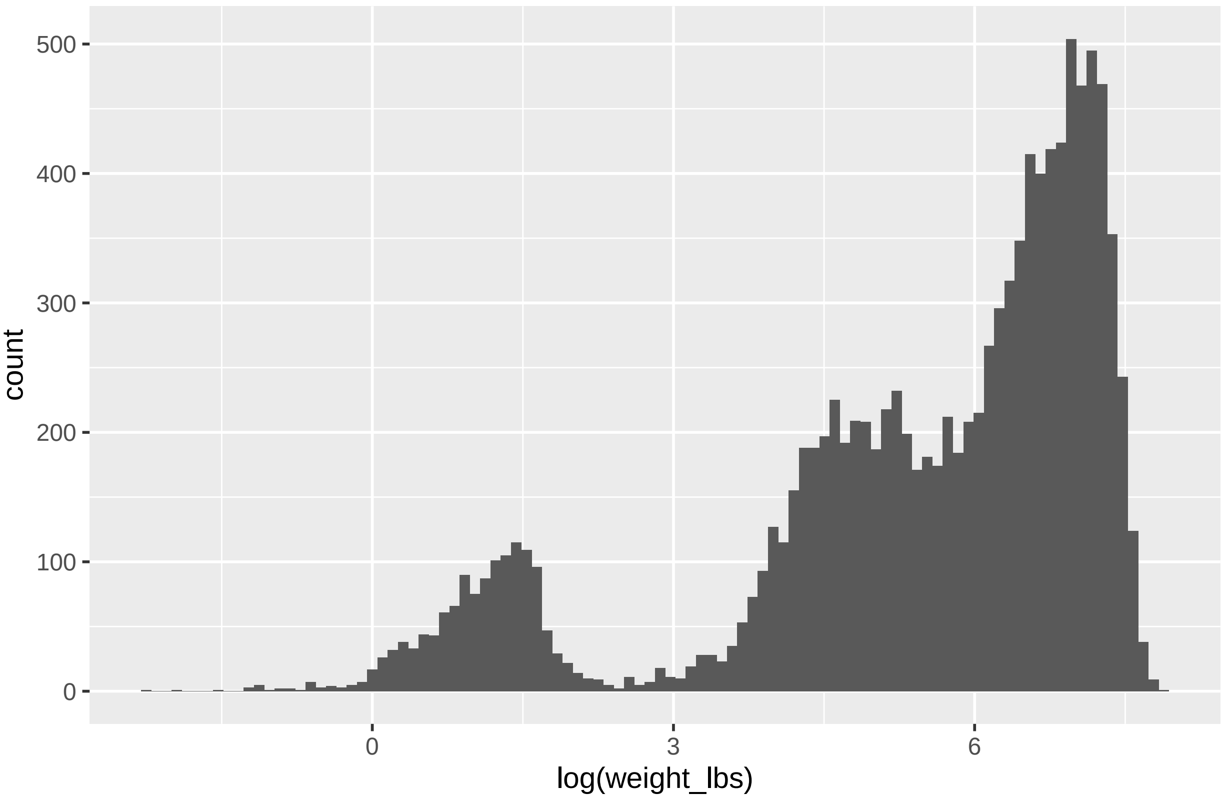
The pumpkin data included additional variables, but were not used due to incomplete and inconsistently recorded data that would have removed the majority of the records from the dataset.

2a. A Model Building Process

Most data filtering is explained in above section 2. See below for EDAs, where key variables used in the model are plotted:

Pumpkin weights are evenly distributed looking at below histogram - Figure 3. There are bumps and these explain the even distribution for each type of pumpkin.

Figure 3: Distribution of Pumpkin Weight



All weather related variables - Average Precipitation, Average Minimum Temperature, Average Maximum Temperature, minimum temperature standard deviation, maximum temperature standard deviation, precipitation standard deviation are evenly distributed. Refer to Figure 4 - 9 below.

Figure 4: Distribution of avg min temperature

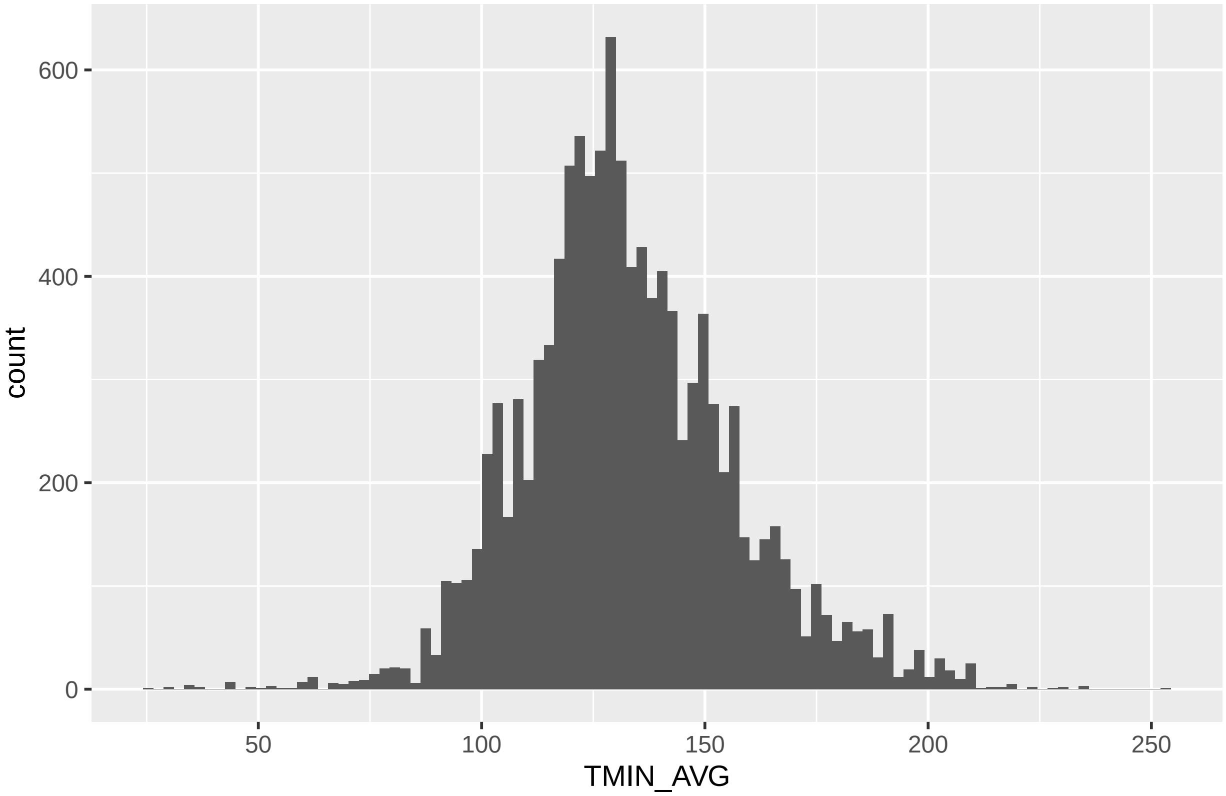


Figure 5: Distribution of stdev of min temp

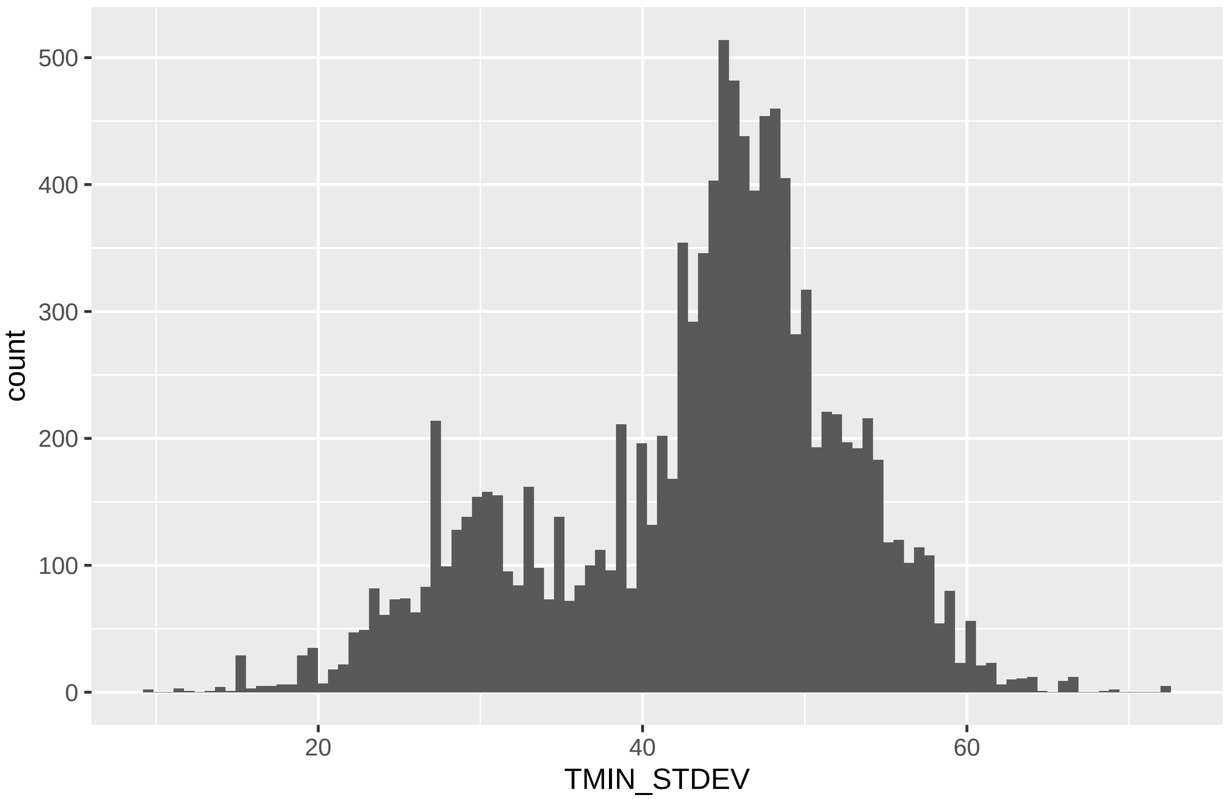


Figure 6: Distribution of avg max temperature

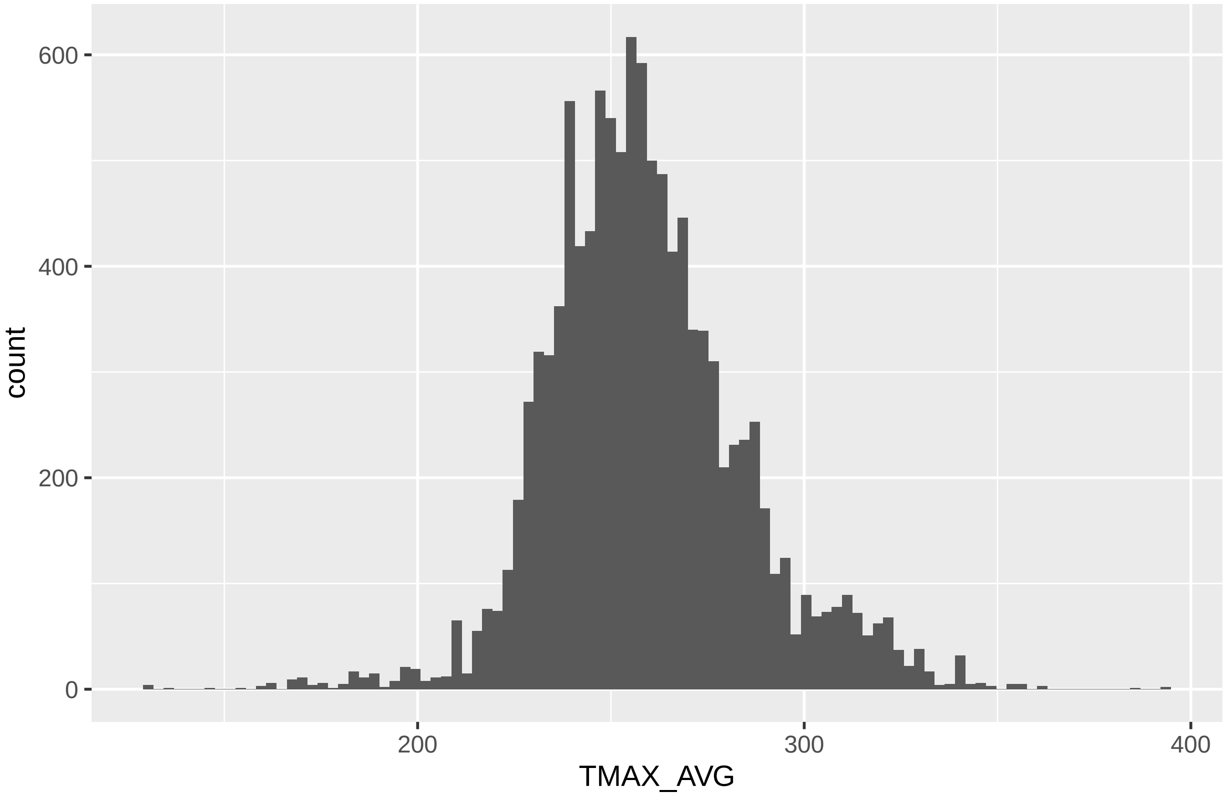


Figure 7: Distribution of stdev max temperature

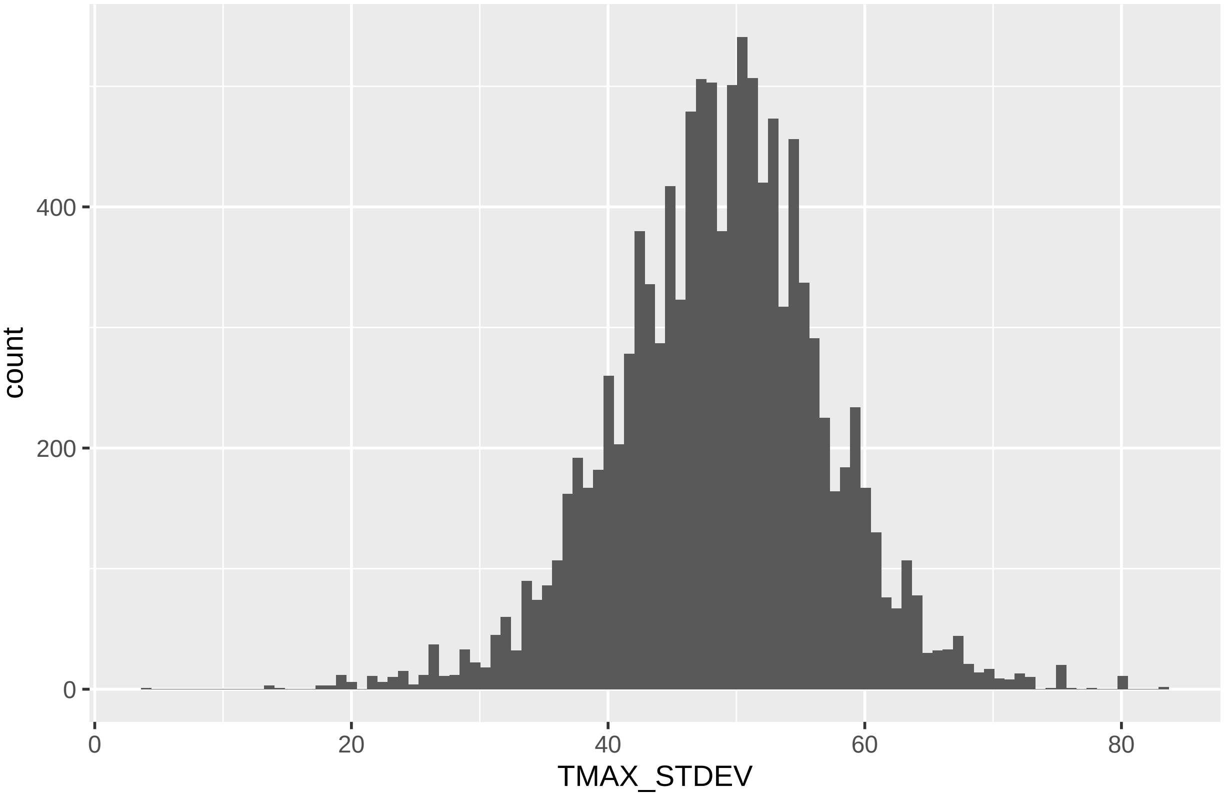


Figure 8: Distribution of count of days per year above min temp

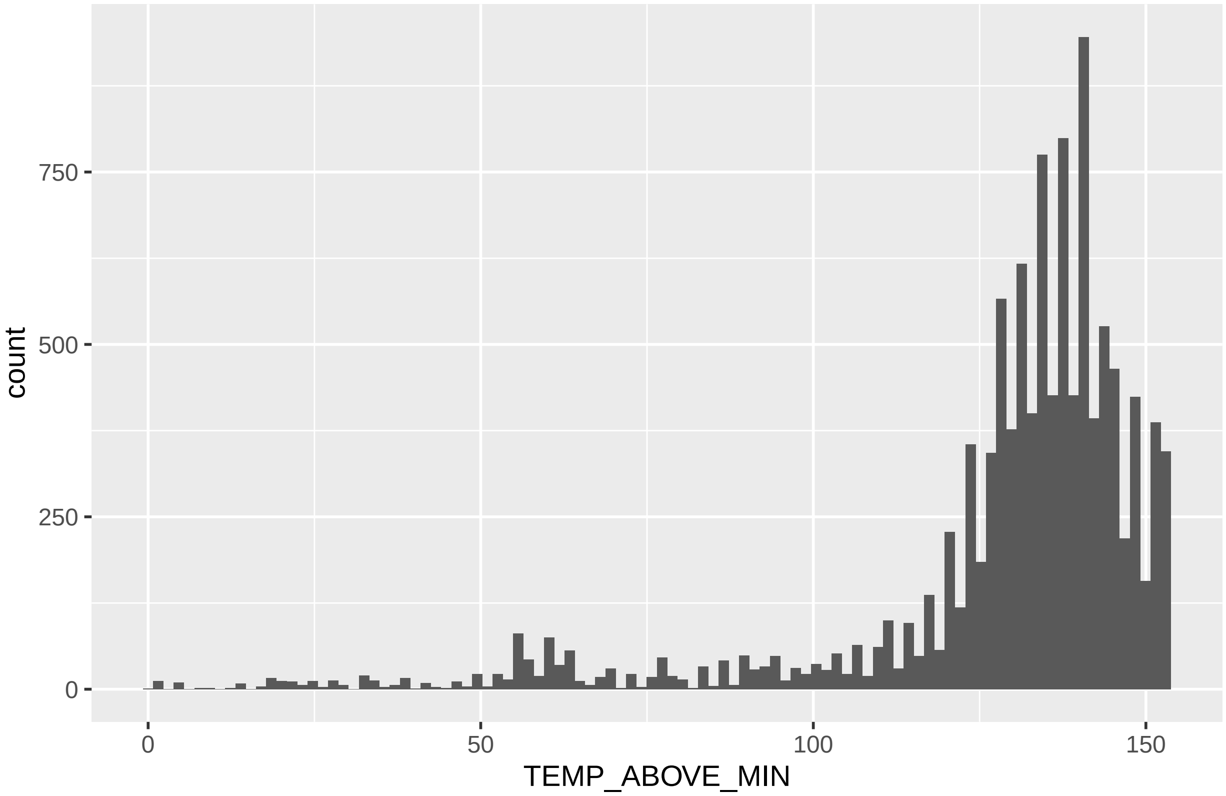
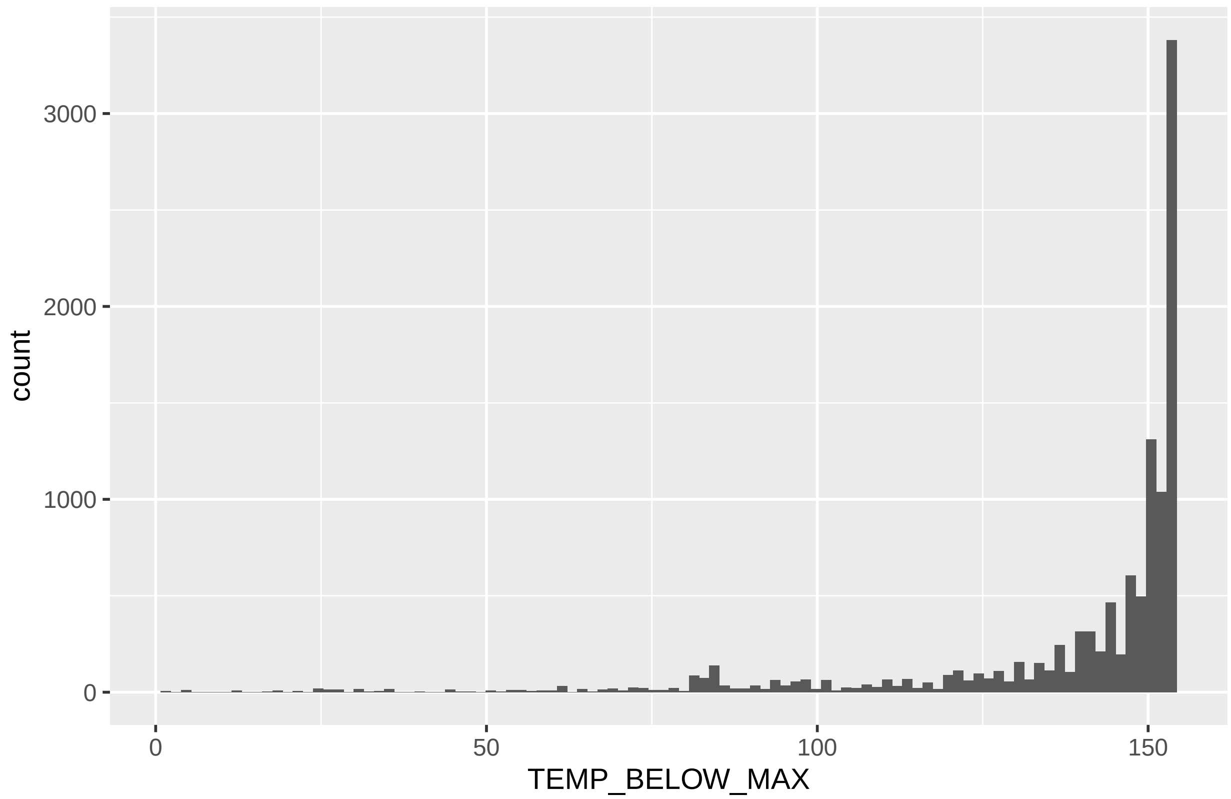


Figure 9: Distribution of count of days per year below max temp



After the initial data analysis, the conducted research attempted to prove the below model:

$$Weight = AveragePrecipitation + AverageMinimumTemperature + AverageMaximumTemperature$$

The output variable the team wanted to measure was the weight of the pumpkin, weight__lbs.

A. We started with a model with only key variables, model1, where minimum temperature, maximum temperature, and precipitation was taken as the predictor. The Rˆ2 value for this initial model was very low, 0.021. **model1: weight__lbs ~ TMIN__AVG+TMAX__AVG+PRCP__AVG**

B. In order to improve the model, in model2, we've added the type as a covariate. The assumption here was that even though all pumpkin/tomatoes measured in this data set are in the same family, the type could impact the weight of the pumpkin. This increased the Rˆ2 value to 0.460, but this was still not optimal. **model2: weight__lbs ~ MIN__AVG+TMAX__AVG+PRCP__AVG+type**

C. In the next model, model3, we took a log transformation of the output variable. Taking the log transformation of the output variable drastically improved the linearity of the model with increased the Rˆ2 value to 0.838. **model3: log transformation of weight__lbs ~ MIN__AVG+TMAX__AVG+PRCP__AVG+type**

D. Adding on to this model, we've created a longer model, model4, by adding in weather-related variables, such as minimum temperature standard deviation, maximum temperature standard deviation, precipitation standard deviation, and year as a covariate. This resulted in Rˆ2 value of 0.840. **model4: log transformation of weight__lbs ~ MIN__AVG+TMAX__AVG+PRCP__AVG+type+year+TMIN__STDEV+TMAX__STDEV+PRCP__ST**

E. In the longest model, model5, we've added in variables like the number of days that exceeded the maximum temperature, and the minimum temperature. This concluded in the R^2 of 0.841. **model5: log transformation of weight__lbs ~ MIN__AVG+TMAX__AVG+PRCP__AVG+type+year+TM + TEMP__ABOVE__MIN**

In the model, both type and the year as covariates greatly helped achieve our modeling goals, proven in the increased R^2 values. This explains that the same weather conditions could impact the size of the pumpkin differently depending on the year, and the variety of the pumpkin.

```
##
## ===================================================================================
##                                               Dependent variable:
##                              ------------------------------------------------------
##                                                    weight_lbs
##                                         (1)                        (2)
## -----------------------------------------------------------------------------------
## TMIN_AVG                             -2.566***                  -1.542***
##                                       (0.329)                    (0.247)
##
## TMAX_AVG                              -0.461                      -0.350
##                                       (0.296)                    (0.220)
##
## PRCP_AVG                             -1.321***                   1.745***
##                                       (0.397)                    (0.298)
##
## factor(type)Giant Pumpkin                                       786.471***
##                                                                  (12.979)
##
## factor(type)Giant Squash                                        500.272***
##                                                                  (20.197)
##
## factor(type)Giant Watermelon                                    94.928***
##                                                                  (16.398)
##
## factor(type)Tomato                                              -67.218***
##                                                                  (16.365)
##
## Constant                           1,077.197***                 312.506***
##                                      (56.732)                    (43.918)
##
## -----------------------------------------------------------------------------------
## Observations                          11,311                      11,311
## R2                                     0.021                      0.460
## Adjusted R2                            0.020                      0.460
## Residual Std. Error            533.054 (df = 11307)        395.791 (df = 11303)
## F Statistic                  79.212*** (df = 3; 11307) 1,376.810*** (df = 7; 11303)
## ===================================================================================
## Note:                                                   *p<0.1; **p<0.05; ***p<0.01
```

```
##
## ====================================================================================================
##                                                                Dependent variable:
##                              -----------------------------------------------------------------------
##                                                                    log(weight_lbs)
##                                           (1)                           (2)
## ----------------------------------------------------------------------------------------------------
## TMIN_AVG                               -0.002***                     -0.003***                    -0
##                                         (0.0005)                      (0.001)                      (
##
## TMAX_AVG                               -0.001***                     -0.001***                    -0
##                                         (0.0004)                      (0.0004)                     (
##
## PRCP_AVG                                0.004***                       0.001
##                                         (0.001)                       (0.001)                      (
##
## factor(type)Giant Pumpkin               2.241***                      2.242***                     2.
##                                         (0.025)                       (0.025)                      (
##
## factor(type)Giant Squash                1.661***                      1.662***                     1.
##                                         (0.039)                       (0.039)                      (
##
## factor(type)Giant Watermelon            0.516***                      0.510***                     0.
##                                         (0.032)                       (0.032)                      (
##
## factor(type)Tomato                     -3.217***                     -3.225***                    -3
##                                         (0.032)                       (0.032)                      (
##
## factor(year)2014                                                      0.213***                     0.
##                                                                       (0.029)                      (
##
## factor(year)2015                                                      0.225***                     0.
##                                                                       (0.029)                      (
##
## factor(year)2016                                                      0.296***                     0.
##                                                                       (0.030)                      (
##
## factor(year)2017                                                      0.274***                     0.
##                                                                       (0.031)                      (
##
## factor(year)2018                                                      0.225***                     0.
##                                                                       (0.031)                      (
##
## factor(year)2019                                                      0.158***                     0.
##                                                                       (0.031)                      (
##
## factor(year)2020                                                      0.269***                     0.
##                                                                       (0.033)                      (
##
## factor(year)2021                                                      0.254***                     0.
##                                                                       (0.032)                      (
##
## TMIN_STDEV                                                            0.007***                     0.
##                                                                       (0.001)                      (
##
## TMAX_STDEV                                                           -0.007***                    -0
##                                                                       (0.001)                      (
##
```

```
## PRCP_STDEV                                                           0.0004                       -0
##                                                                      (0.001)                       (0
##
## TEMP_BELOW_MAX                                                                                   -0
##                                                                                                    (0
##
## TEMP_ABOVE_MIN                                                                                   0.0
##                                                                                                    (0
##
## Constant                               4.790***                     4.746***                      5.0
##                                        (0.086)                       (0.105)                       (0
##
## -----------------------------------------------------------------------------------------------------
## Observations                            11,311                       11,311                        11
## R2                                      0.838                         0.840                         0
## Adjusted R2                             0.838                         0.840                         0
## Residual Std. Error                0.771 (df = 11303)          0.765 (df = 11292)            0.764 (d
## F Statistic                  8,330.414*** (df = 7; 11303) 3,299.626*** (df = 18; 11292) 2,985.490***
## =====================================================================================================
## Note:                                                                                     *p<0.1; **p
```

```
##
## =====================================================================================================
##                                                              Dependent variable:
##                              ------------------------------------------------------------------------
##                                     weight_lbs                              log(weight_lbs)
##                                         (1)                        (2)                           (3)
## -----------------------------------------------------------------------------------------------------
## TMIN_AVG                             -2.566***                  -0.002***                     -0.003*
##                                       (0.329)                    (0.0005)                      (0.00
##
## TMAX_AVG                              -0.461                    -0.001***                     -0.003*
##                                       (0.296)                    (0.0004)                      (0.00
##
## PRCP_AVG                             -1.321***                   0.004***                       0.00
##                                       (0.397)                    (0.001)                       (0.00
##
## factor(type)Giant Pumpkin                                        2.241***                     2.248**
##                                                                  (0.025)                       (0.025
##
## factor(type)Giant Squash                                         1.661***                     1.672**
##                                                                  (0.039)                       (0.039
##
## factor(type)Giant Watermelon                                     0.516***                     0.521**
##                                                                  (0.032)                       (0.032
##
## factor(type)Tomato                                              -3.217***                     -3.217*
##                                                                  (0.032)                       (0.032
##
## factor(year)2014                                                                              0.203**
##                                                                                                (0.029
##
## factor(year)2015                                                                              0.211**
##                                                                                                (0.029
##
## factor(year)2016                                                                              0.296**
##                                                                                                (0.03
##
## factor(year)2017                                                                              0.277**
##                                                                                                (0.03
##
## factor(year)2018                                                                              0.215**
##                                                                                                (0.03
##
## factor(year)2019                                                                              0.150**
##                                                                                                (0.03
##
## factor(year)2020                                                                              0.280**
##                                                                                                (0.03
##
## factor(year)2021                                                                              0.259**
##                                                                                                (0.03
##
## TMIN_STDEV                                                                                    0.007**
##                                                                                                (0.00
##
## TMAX_STDEV                                                                                    -0.007*
##                                                                                                (0.00
##
## PRCP_STDEV                                                                                     -0.00
##                                                                                                (0.00
##
## TEMP_BELOW_MAX                                                                                -0.002*
##                                                                                                (0.00
##
## TEMP_ABOVE_MIN                                                                                0.004**
##                                                                                                (0.00
##
## Constant                           1,077.197***                 4.790***                     5.052**
##                                      (56.732)                    (0.086)                       (0.18
##
## ----------------------------------------------------------------------------------------------------
## Observations                          11,311                      11,311                       11,31
## R2                                     0.021                       0.838                         0.84
## Adjusted R2                            0.020                       0.838                         0.84
## Residual Std. Error            533.054 (df = 11307)        0.771 (df = 11303)           0.764 (df =
## F Statistic                  79.212*** (df = 3; 11307) 8,330.414*** (df = 7; 11303) 2,985.490*** (df
## ====================================================================================================
## Note:                                                                                    *p<0.1; **p<0.0
```

Once the models are built, F-test was conducted to compare few models:

- Model 1 vs Model2 : p-value was lower than 0.05, which rejects the null hypothesis. This proves that the full model, model2 is a better model.

```
## Analysis of Variance Table
##
```

```
## Model 1: weight_lbs ~ TMIN_AVG + TMAX_AVG + PRCP_AVG
## Model 2: weight_lbs ~ TMIN_AVG + TMAX_AVG + PRCP_AVG + factor(type)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  11307 3212840439
## 2  11303 1770619549  4 1442220890 2301.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- Model 3 vs Model 4: p-value was lower than 0.05, which rejects the null hypothesis. This proves that the full model, model4 is a better model.

```
## Analysis of Variance Table
##
## Model 1: log(weight_lbs) ~ TMIN_AVG + TMAX_AVG + PRCP_AVG + factor(type)
## Model 2: log(weight_lbs) ~ TMIN_AVG + TMAX_AVG + PRCP_AVG + factor(type) +
##     factor(year) + TMIN_STDEV + TMAX_STDEV + PRCP_STDEV
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  11303 6720.0
## 2  11292 6611.9 11     108.1 16.784 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- Model 4 vs Model 5: p-value was lower than 0.05, which rejects the null hypothesis. This proves that the full model, model5 is a better model.

```
## Analysis of Variance Table
##
## Model 1: log(weight_lbs) ~ TMIN_AVG + TMAX_AVG + PRCP_AVG + factor(type) +
##     factor(year) + TMIN_STDEV + TMAX_STDEV + PRCP_STDEV
## Model 2: log(weight_lbs) ~ TMIN_AVG + TMAX_AVG + PRCP_AVG + factor(type) +
##     factor(year) + TMIN_STDEV + TMAX_STDEV + PRCP_STDEV + TEMP_BELOW_MAX +
##     TEMP_ABOVE_MIN
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  11292 6611.9
## 2  11290 6581.4  2    30.457 26.124 4.795e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

This aligns with the Rˆ2 value from above as well. Per the F-test analysis and the Rˆ2, we have decided to proceed with model 5 to validate assumptions and run additional tests to observe the regression fit.

3. Model Assumptions

Due to the size of our data set, we utilized an OLS Large Sample model. Below is a discussion of the model assumptions.

Independently and Identically Distributed Our data includes a certain amount of dependency due to a) standard practices for cucurbit growing and harvesting and b) geographic similarities. However, our data is relatively granular since the fruit attributes are observed at the individual level and the weather data is observed by stations rather than at a more highly aggregated level. The size of the dataset should also help to offset loss of power caused by any dependency.

Furthermore, our model utilizes data for several types of cucurbits. This can potentially be seen as a violation of the identically distributed assumption. However, since all types of fruit included in our dataset belong to the cucurbit family and are affected by weather patterns in similar ways, we concluded that this violation was negligible. Additionally, within each type of cucurbit, the data should be identically distributed.

No Perfect Collinearity and Unique BLP Perfect collinearity exists when multiple variables from a model contribute the same information. In other words, the perfectly collinear variables are related to the point that including all of them adds no additional value to the data. The model can instead be simplified to use the linear transformation of just one of the variables.

Within R, variables are dropped from models if they exhibit perfect collinearity with another variable. When running coefficient tests on the models included in this report, since none of the variables were dropped, we can assume that no perfect collinearity exists in our models. (In the table below, we can observe that all of the variables included in model5 were also included in the output of our coefficient test.) Furthermore, since there is no perfect collinearity, we can also conclude that a Best Linear Predictor exists for our models.

```
coeftest(model5, vcov=vcovHC(model5))
```

```
##
## t test of coefficients:
##
##                                 Estimate  Std. Error   t value  Pr(>|t|)
## (Intercept)                   5.05206192  0.18801218   26.8709 < 2.2e-16 ***
## TMIN_AVG                     -0.00324979  0.00055887   -5.8149 6.232e-09 ***
## TMAX_AVG                     -0.00331572  0.00078140   -4.2433 2.220e-05 ***
## PRCP_AVG                      0.00211509  0.00135433    1.5617   0.11838
## factor(type)Giant Pumpkin     2.24833282  0.01750822  128.4159 < 2.2e-16 ***
## factor(type)Giant Squash      1.67237022  0.04809389   34.7730 < 2.2e-16 ***
## factor(type)Giant Watermelon  0.52099020  0.02685510   19.4000 < 2.2e-16 ***
## factor(type)Tomato           -3.21683714  0.02155732 -149.2225 < 2.2e-16 ***
## factor(year)2014              0.20288630  0.02957377    6.8603 7.229e-12 ***
## factor(year)2015              0.21108083  0.03052700    6.9146 4.946e-12 ***
## factor(year)2016              0.29604572  0.03097291    9.5582 < 2.2e-16 ***
## factor(year)2017              0.27746386  0.03208289    8.6483 < 2.2e-16 ***
## factor(year)2018              0.21471975  0.03309752    6.4875 9.093e-11 ***
## factor(year)2019              0.15040858  0.03381814    4.4476 8.768e-06 ***
## factor(year)2020              0.28031584  0.03477068    8.0618 8.273e-16 ***
## factor(year)2021              0.25859174  0.03388215    7.6321 2.496e-14 ***
## TMIN_STDEV                    0.00662045  0.00109080    6.0693 1.326e-09 ***
## TMAX_STDEV                   -0.00691112  0.00129874   -5.3214 1.049e-07 ***
## PRCP_STDEV                   -0.00009131  0.00057907   -0.1577   0.87471
## TEMP_BELOW_MAX               -0.00194826  0.00100602   -1.9366   0.05282 .
## TEMP_ABOVE_MIN                0.00437206  0.00111411    3.9243 8.751e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

For completeness, we also tested our model against the Classical Linear model assumptions that were not discussed above.

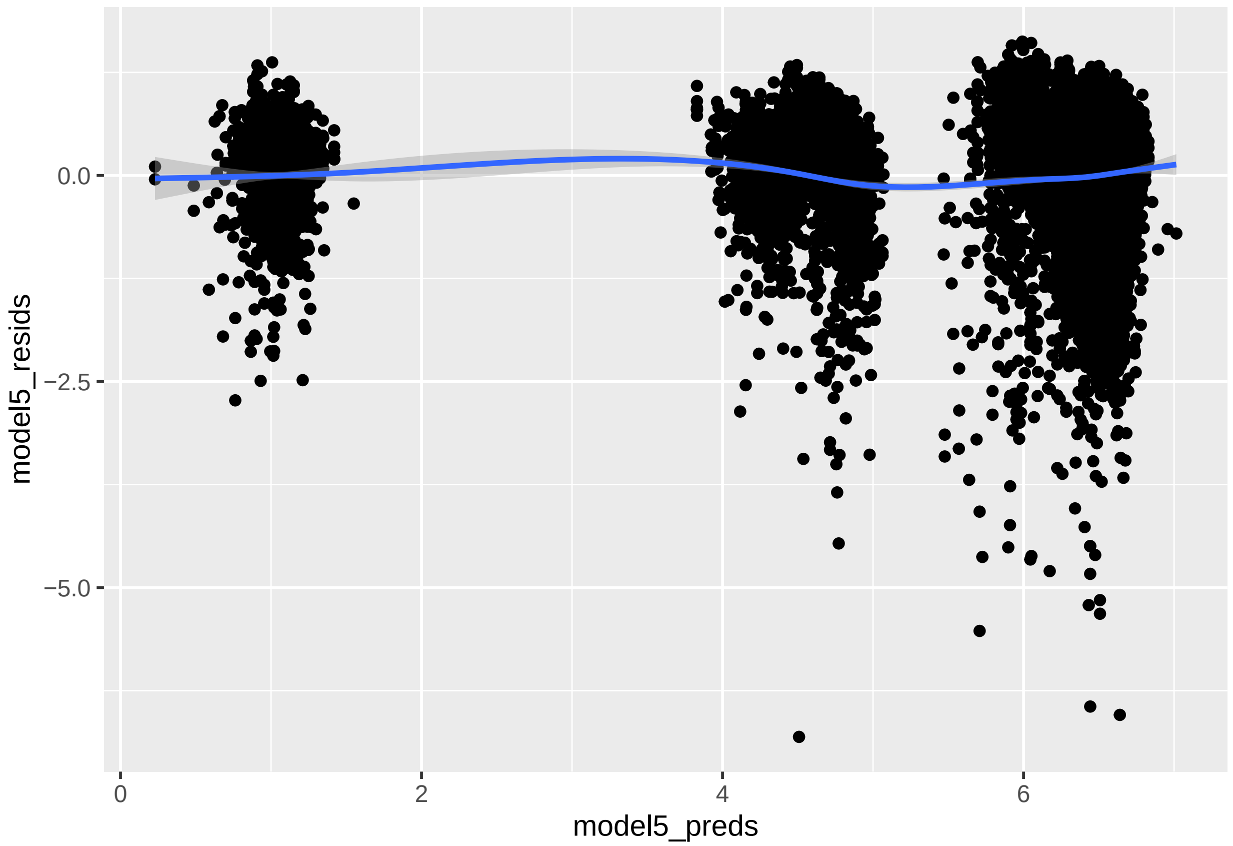
Homoscedasticity An additional assumption of the classical linear model is that the variance of the residuals is constant, or homoscedastic. To test this assumption, we plotted the residuals against the predicted values, as can be seen in the below chart. A plot for a model with homoscedastic residuals would show points evenly distributed. For our main model (model5), there were large amounts of clustering. When taking into account the different types of Cucurbits included in this model, the clustering is intuitive. Within each cluster, the data points were relatively evenly distributed. Therefore, we concluded that our model met the homoscedasticity requirement reasonably well.

```r
# model 5
model5_preds <- predict(model5)
model5_resids <- resid(model5)



plot5 <- ggplot(data=d_pumpkin, aes(model5_preds, model5_resids)) + geom_point() +
stat_smooth()
print(plot5)
```

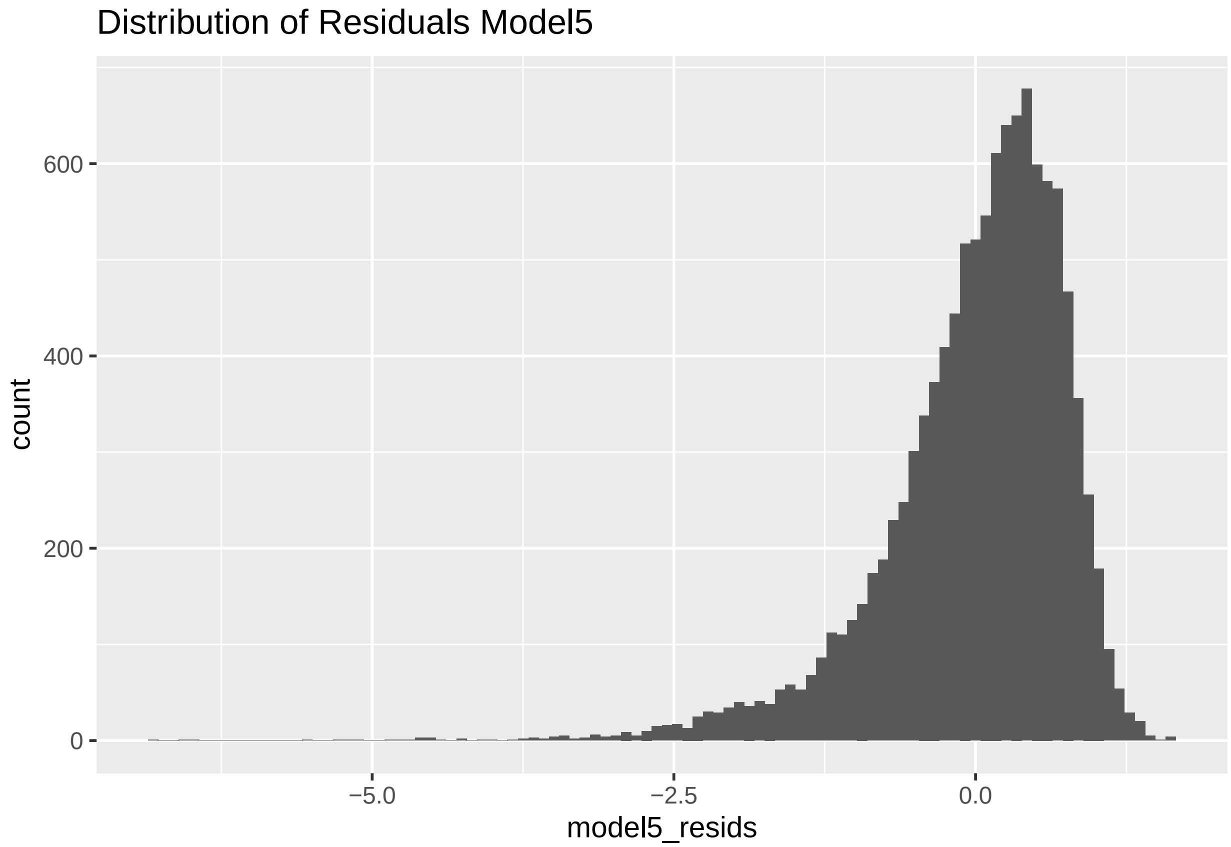
```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



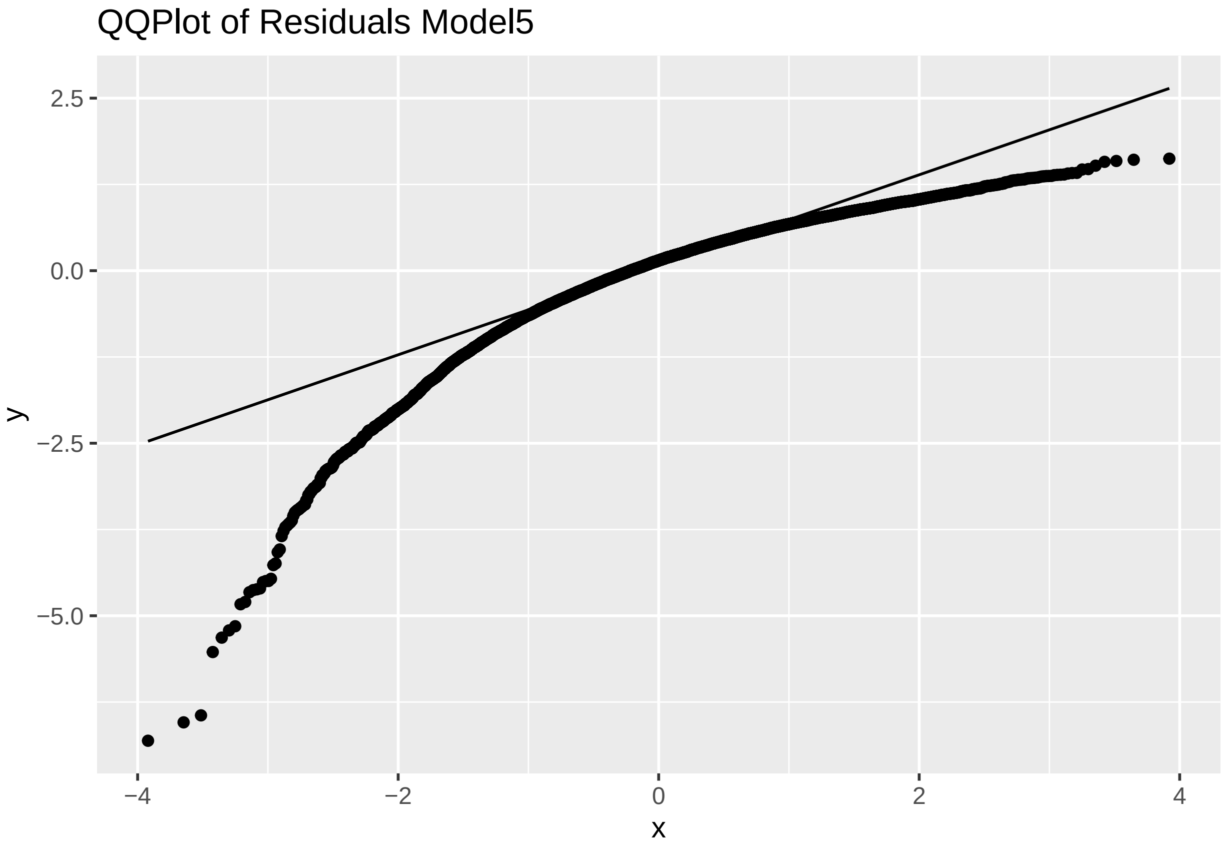
Linear Conditional Expectations Furthermore, the classical linear model assumes linear conditional expectations. To test this assumption, we again observed the plot of our model residuals against the predicted values. In these plots, we focused our analysis on the smoothing curve to better distinguish the linearity of the data points. While we observed nonlinear smoothing curves for most of our models, our main model (model5) plot showed reasonably (although not perfectly) linear data. Therefore, we concluded that model5 met the linear conditional expectations assumption.

Normally Distributed Errors Finally, the classical linear model assumes normally distributed residuals. To test the normality of the residuals, we created a histogram and a qq plot of the residuals for each of the models. As seen in the histogram chart below, the residuals of model5 are left skewed. Furthermore, we can also see that our residuals do not fit well to a linear line. Ultimately, our models failed this assumption. However, since we utilized over eleven thousand data points for our model, we can apply the central limit theorem to our residuals and assume that the distribution of our errors does not compromise the validity of our t-tests.

```
plot_errors_5 <- d_pumpkin %>%
  ggplot(aes(x=model5_resids)) +
  geom_histogram(bins=100) +
  ggtitle("Distribution of Residuals Model5")
print(plot_errors_5)
```



```
qqplot_5 <- d_pumpkin %>%
  ggplot(aes(sample=model5_resids)) +
  stat_qq() + stat_qq_line() +
  ggtitle("QQPlot of Residuals Model5")
print(qqplot_5)
```



4. A Results Section

Below are the results of the five regression models discussed in the previous section. The dependent variable for the first two models was cucurbit weight in pounds. The dependent variable for the last three models was the log of the cucurbit weight in pounds.

In our main model (model5), most of the independent variables were shown to be statistically significant with the exception of the precipitation variables. (We hypothesized that the insignificance of precipitation might be due to well controlled irrigation systems developed by growers.)

Cucurbit type had the most effect on our model. For our model, Field Pumpkins were used as our base type. The effect of type is very intuitive; the type of cucurbit will have a very large effect on the expected size of the fruit. Our interpretations of the type coefficients are as follows: -Giant Pumpkin: All else being equal, a giant pumpkin will weigh 2.248 pounds more than a field pumpkin. -Giant Squash: All else being equal, a giant squash will weigh 1.672 pounds more than a field pumpkin. -Giant Watermelon: All else being equal, a giant watermelon will weigh 0.521 pounds more than a field pumpkin. -Tomato: All else being equal, a tomato will weigh 3.217 pounds less than a field pumpkin.

Following type, year also had a very large effect on our model. 2013 was used as our base year. Each year had a relatively high coefficient compared to our other variables and was statistically significant, suggesting that there is a time-based component affecting the weight of the pumpkins. (Potential rationale for this could be changes in growing practices in the competitive cucurbit community or improving/changing genetics in the cucurbits.) All of our year coefficients were positive suggesting that the size of cucurbits had generally increased since 2013. For example, our interpretation of the 2014 variable coefficient would be as follows: All else being equal, a cucurbit grown in 2014 would weigh 0.203 pounds more than a cucurbit grown in 2013.

Our last group of variables was our weather variables. While most of the weather variables were statistically significant, they all had very small coefficients suggesting that their practical significance was rather limited. Furthermore, for some of the weather variables, the signs of the coefficients behave in a manner contradictory to what has been asserted in previous studies, which produces some concern about their validity.

One of our more interesting findings for the weather variables was in regards to our TEMP_ABOVE_MIN and TEMP_BELOW_MAX variables. As previously discussed, since literature on cucurbit growing suggests that the fruit growing is best done in temperatures between 65 F and 95 F, these two variables represent the number of days the temperature actually fell within this range. Our expectation was that the coefficient for both variables would be positive. TEMP_ABOVE_MIN followed our expectation with a positive coefficient significant at the one percent significance level. Our interpretation of the TEMP_ABOVE_MIN coefficient is as follows: One additional day of max daily temperature being below the max optimal temperature for cucurbit growth increases the cucurbit weight by 0.004 pounds.

The TEMP_BELOW_MAX variable had a negative coefficient. However, it was only significant at the ten percent significance level while our cutoff was the five percent significance level. Therefore, we disregarded this result as insignificant.

An additional finding for the weather variable was in regards to our TMIN_STDEV and TMAX_STDEV variables. Since cucurbits are supposed to thrive in consistent climates, we had hypothesized that the coefficients would be negative for both of these variables. TMAX_STDEV followed our expectation and is significant at the one percent significance level. Our interpretation of the TMAX_STDEV coefficient is as follows: An increase in the standard deviation of the daily maximum temperature by one decreases the weight of the cucurbit by 0.007 pounds.

While the TMIN_STDEV variable deviated from our expectations with a positive coefficient, the variable coefficient was also statistically significant. Our interpretation of the TMIN_STDEV coefficient is as follows: An increase in the standard deviation of the daily minimum temperature by one increases the weight of the cucurbit by 0.007 pounds.

Below is additional interpretations of the significant variable coefficients not discussed previously in this section: -An increase in the TMIN_AVG temperature by one degree will decrease the weight of the cucurbit by 0.03 pounds. -An increase in the TMAX_AVG temperature by one degree will decrease the weight of the cucurbit by 0.03 pounds.

5. Structural limitations of your model

There were many structural limitations to building a regression model with the combined data set we chose. The first limitation was our lack of access to data around irrigation systems and amount of sunlight exposed to the Cucurbits. This is important as the net water intake during the growing season is a combination of precipitation and supplemental watering. The omitted variable regarding irrigation is positively correlated with the weight_lbs variable and negatively correlated with PRCP suggesting that the omitted variable bias is negative. Alongside supplemental watering, another opportunity for omitted variable bias comes from omitting the amount of sunlight the Cucurbits received. The variable has a causal relationship with our outcome variable and is negatively correlated with the precipitation variable that we used in our regression. The sunlight variable was omitted due to lack of data and likely has a negative bias effect on the model further supporting the notion that the precipitation variable is overestimated in the model.

The negative omitted variable bias of both irrigation and sunshine variables moves our positive coefficient for PRCP towards zero reducing the explanatory power of the PRCP variable in our model. Because of this, we are comfortable with the results of our model. Since we can only measure precipitation in a region and not any sort of irrigation system setup by the grower, we are assuming that the precipitation is the only form of water the Cucurbits are receiving. This is likely a false assumption, but since we can't measure how much water irrigation systems were supplying to various Cucurbits, the PRCP variable is a reasonable inclusion into the model. To resolve the missing data surrounding the omitted variable bias we could take two different approaches. We can collect the data from the growers about any supplemental water added and sunshine to Cucurbits on a time series basis. This would allow us to have a better understanding of the net water consumed by the Cucurbits and the amount of light they were exposed to. The second option would be to create proxy variables for the omitted variables which would call into question the reproducibility of the research.

The second limitation was missing data within our pumpkin data set. We omitted the seed_mother and pollinator_father variables from our model - though we believed they would be a deterement of the outcome variable - due to missing data. While research suggests seed_mother and pollinator_father have explanatory power of our outcome variable, omitting the variables does not create any bias in the model since the omitted variables do not have any correlation with the chosen independent variables. Other than the aforementioned variables, we omitted: grower_name, city, state_prov, country, gpc_site, ott (over the top inches), est_weight, pct_chart (weight percentile), and variety. These variables were omitted due to their lack of explanatory power of the outcome variable. We concluded that the omission of these variables did not create any omitted variable bias due to the lack of correlation between the listed variables and the independent variables chosen in the model.

7. Conclusion

8. References

1. The Great Pumpkin Commonwealth. 27 Oct. 2014, gpc1.org/.

2. “GPC Weighoff Results and Information.” Www.bigpumpkins.com, www.bigpumpkins.com/ViewArticle.asp?id=13

3. GPC Rules and Handbook. Great Pumpkin Commonwealth, 1 Mar. 2021.

4. Daily Temperature and Precipitation Reports - Data Tables. National Oceanographic and Atmospheric Administration, www.climate.gov/maps-data/dataset/daily-temperature-and-precipitation-reports-data-tables.

5. “World Cities Database.” Kaggle.com, www.kaggle.com/max-mind/world-cities-database.

6. Penn State Extension. “Pumpkin Production.” Penn State Extension, 20 June 2005, extension.psu.edu/pumpkin-production.

7. KGO. “2,191-Pound Mega Gourd Wins World Championship Pumpkin Weigh-off in Half Moon Bay.” ABC7 San Francisco, 11 Oct. 2021, abc7news.com/half-moon-bay-pumpkin-weigh-off-winner-biggest-2021-peninsula-patch/11112151/. Accessed 9 Dec. 2021.

8. “How to Grow a Giant Pumpkin.” Www.pumpkinfest.org, www.pumpkinfest.org/giant-vegetables/how-to-grow-a-giant-pumpkin/.

9. Group, EMILY FABER, Sinclair Broadcast. “Growing These Massive One-Ton Pumpkins Takes a Green Thumb and the Perfect Seed.” WJLA, 29 Oct. 2020, wjla.com/news/offbeat/growing-these-massive-one-ton-pumpkins-takes-a-green-thumb-and-the-perfect-seed.

10. “Beginner’s Seminar Presentation.” Great Pumpkin Commonwealth, Great Pumpkin Commonwealth, gpc1.org/wp-content/uploads/2019/02/beginners-power-point.ppt.