Fruit Pricing Across Regions and Seasons

Amanda Yu
Master of Science, Data Analytics – Data Engineering
Western Governors University
February 2026

TABLE OF CONTENTS

1

INTRODUCTION & BACKGROUND

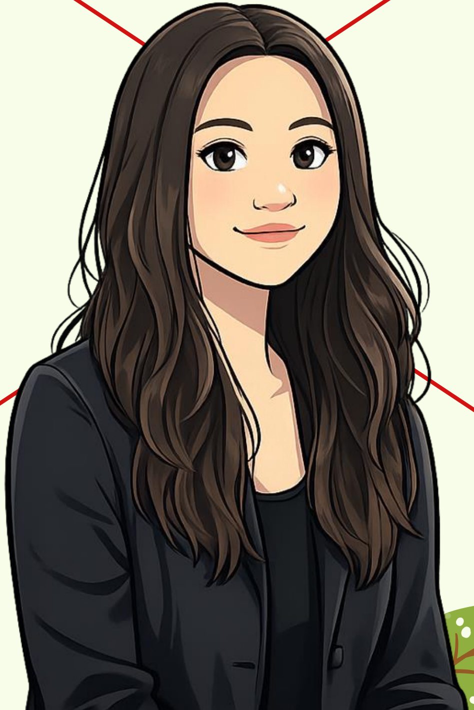
INTRODUCTION & BACKGROUND

About Me

Amanda Yu
Alberta, Canada

Current Role

Data & Reporting Analyst



Education

Bachelor of Commerce – Business Analytics

Experience

Data visualization & reporting

2

RESEARCH QUESTION & HYPOTHESIS

RESEARCH QUESTION & HYPOTHESIS

What factors significantly influence fruit pricing across regions and seasons?

Null Hypothesis

Fruit type, region, ripeness, weight, and season have no statistically significant effect on fruit price
($p \geq 0.05$).

Alternate Hypothesis

At least one of the following variables: fruit type, region, ripeness, weight, or season have a statistically significant effect on fruit price
(p-value < 0.05).

3

SUMMARY OF ANALYSIS & FINDINGS

TOOLS AND TECHNIQUES

Google BigQuery

Cloud data warehouse used to store and analyze large datasets efficiently.

SQL

Query language used to retrieve, filter, and summarize data stored in relational databases.

Multiple Linear Regression

Statistical method to predict the value of one dependent variable based on two or more independent variables, and estimates how much each factor influences the outcome.

DATA ANALYSIS PROCESS

Kaggle
Dataset downloaded from open source

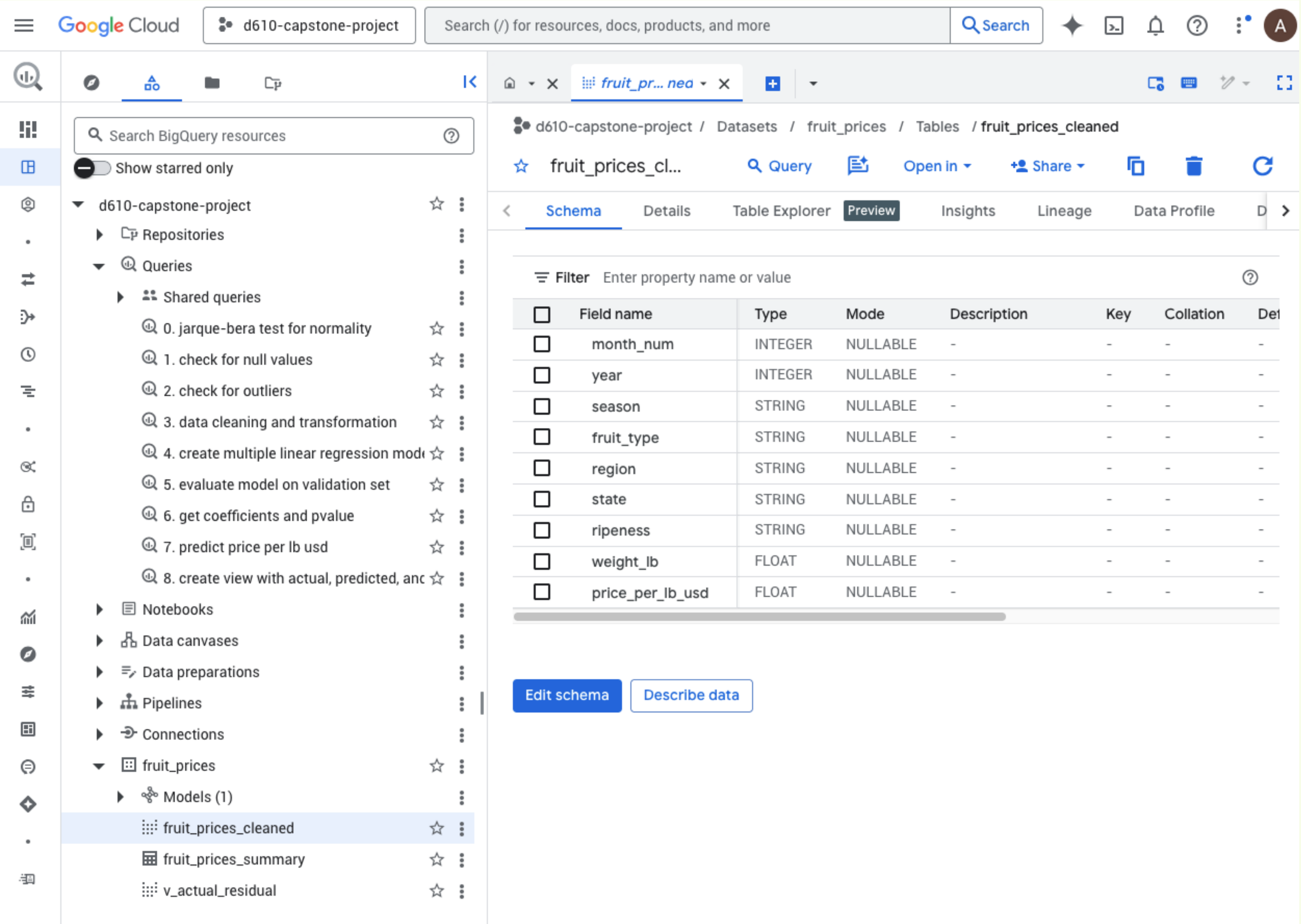
BigQuery
Upload dataset into tables with define schema

SQL
Write and save SQL queries

Data Quality
Check for nulls or outliers

Normality Test
Jarque-Bera test to evaluate distribution

BigQuery ML
Create linear regression model


Google Cloud
d610-capstone-project
Search (/) for resources, docs, products, and more
Search
Search BigQuery resources
Show starred only
d610-capstone-project
Repositories
Queries
Shared queries
0. jarque-bera test for normality
1. check for null values
2. check for outliers
3. data cleaning and transformation
4. create multiple linear regression mode
5. evaluate model on validation set
6. get coefficients and pvalue
7. predict price per lb usd
8. create view with actual, predicted, anc
Notebooks
Data canvases
Data preparations
Pipelines
Connections
fruit_prices
Models (1)
fruit_prices_cleaned
fruit_prices_summary
v_actual_residual
fruit_pr... nea
d610-capstone-project / Datasets / fruit_prices / Tables / fruit_prices_cleaned
fruit_prices_cl...
Query
Open in
Share
Schema
Details
Table Explorer
Preview
Insights
Lineage
Data Profile
Filter
Enter property name or value
Field name
Type
Mode
Description
Key
Collation
month_num INTEGER NULLABLE
year INTEGER NULLABLE
season STRING NULLABLE
fruit_type STRING NULLABLE
region STRING NULLABLE
state STRING NULLABLE
ripeness STRING NULLABLE
weight_lb FLOAT NULLABLE
price_per_lb_usd FLOAT NULLABLE
Edit schema
Describe data

FINDINGS

Significant Predictors

Most variables were statistically significant (19/20 $p < 0.05$, 12 with $p \approx 0$)

Model Performance

$R^2 = 0.95$ (typical error \$0.15–\$0.19/lb)

Conclusion

Reject null hypothesis because fruit type, region, ripeness, and season are significant predictors of fruit price.

6. get coefficients and pvalue

```sql
SELECT
  *
FROM ML.ADVANCED_WEIGHTS(
  MODEL `fruit_prices.fruit_prices.fruit_price_linear_regression`
)
ORDER BY p_value;
```

Query completed

Using on-demand processing quota

Query results

Row	processed_input	category	weight	standard_error	p_value
1	__INTERCEPT__	null	2.301403552189...	null	null
2	fruit_type	Peach	0.0	0.0	NaN
3	region	South	0.0	0.0	NaN
4	ripeness	Slightly Unripe	0.0	0.0	NaN
5	season	Winter	0.0	0.0	NaN
6	fruit_type	Pineapple	0.650794880679...	0.008500209317...	0.0
7	fruit_type	Blueberry	2.493071186793...	0.008543439712...	0.0
8	fruit_type	Apple	-0.34238919215...	0.008549803269...	0.0
9	fruit_type	Avocado	-0.45646441618...	0.008555249301...	0.0
10	fruit_type	Strawberry	1.103273700129...	0.008556821712...	0.0
11	fruit_type	Mango	-0.73384918342...	0.008557834515...	0.0
12	fruit_type	Orange	-0.92633471095...	0.008577855681...	0.0
13	fruit_type	Banana	-1.48249184589...	0.008585782234...	0.0
14	region	West	0.207488497453...	0.005096619589...	0.0
15	region	Northeast	0.170960692000...	0.005627243142...	0.0
16	ripeness	Overripe	-0.53771043502...	0.006088357684...	0.0
17	season	Summer	-0.33481592251...	0.005447107875...	0.0
18	ripeness	Very Ripe	-0.13836849659...	0.006068530923...	1.332267629550...
19	fruit_type	Grape	0.190923115927...	0.008585792353...	1.776356839400...
20	season	Spring	-0.12093368147...	0.005435842329...	1.776356839400...
21	season	Fall	-0.09065865149...	0.005431815912...	1.718625242119...
22	ripeness	Unripe	-0.06541444045...	0.006067028427...	1.607620703225...
23	region	Midwest	0.044777373685...	0.005199561177...	5.010600157007...
24	ripeness	Ripe	0.050940803174...	0.006084887456...	7.675138080642...
25	weight_lb	null	-0.00139633364...	0.001406792514...	0.321357964099...

4

LIMITATIONS OF TECHNIQUES USED

LIMITATIONS

DATASET

Dataset is synthetic and restricted to U.S. states over only two years, which limits real-world generalizability and may not reflect true market dynamics.

MULTIPLE LINEAR REGRESSION

Assumes straight-line relationships and may miss more complex patterns or interactions

BIG QUERY

Heavy feature engineering (like one-hot encoding) can make SQL pipelines lengthy and may increase cost if repeated queries re-scan large tables

5

SUMMARY OF PROPOSED ACTIONS

PROPOSED ACTIONS

PRICE BENCHMARKING

Use the model to estimate expected price/lb and compare pricing by fruit, region, season, ripeness, and weight.

FOCUS ON KEY DRIVERS

Track the largest and most significant coefficients to explain and anticipate price changes.

OPERATIONALIZE

In BigQuery, refresh data, retrain the model, and publish predictions to dashboards/reports.

FUTURE ENHANCEMENTS

Validate with real-world data, expand time/geography, and test interactions/nonlinear models.

6

BENEFITS OF ANALYSIS

BENEFITS OF ANALYSIS

ACTIONABLE INSIGHTS

Identifies the biggest price drivers from the interpretable coefficients

ACCURATE FORECASTS

R^2 ~0.95
Typical error ~$0.15–$0.19/lb

SCALABLE WORKFLOW

Repeatable BigQuery ML pipeline for consistent refreshes

REFERENCES

- Google Cloud. (n.d.). BigQuery [Documentation]. Google Cloud. Retrieved January 25, 2026, from https://cloud.google.com/bigquery
- Google Cloud. (n.d.). BigQuery ML introduction [Documentation]. Google Cloud. Retrieved January 25, 2026, from https://docs.cloud.google.com/bigquery/docs/bqml-introduction
- Sewell, W. (n.d.). Fruit Pricing Dataset [Dataset]. Kaggle. Retrieved December 27, 2025, from https://www.kaggle.com/datasets/williamsewell/fruit-pricing-dataset/data
- Slidesgo. (n.d.). Strawberry and Fruits [Presentation template]. Slidesgo. Retrieved January 25, 2026, from https://slidesgo.com/theme/strawberry-and-fruits#search-fruit&position-12&results-215&rs=search

THANK YOU!