One Smart Cookie

GIRL SCOUT DIGITAL COOKIE SALES TRENDS

Christy Carter Cathy Potter Stacey Booysen

Cookie Filling

- Introduction
- Database
- Machine Learning
- Summary Graphs
- Interactive Map
- Conclusion

Cookie Mom of 9 Years

- Focus on Covid impacts to Digital Cookie app usage
- Digital Cookie allows credit card transactions
- 2020 was best sales year ever, and before quarantine
- 2021 was full covid impact with higher no-contact sales methods
- 2022 was almost back to normal but with cookie shortage due to supply chainSlide 2

A Batch Made in Heaven

Reason topic was selected

Council Provided:

- Digital Cookie sales detail for 2019-2022
- Service unit zip code assignments

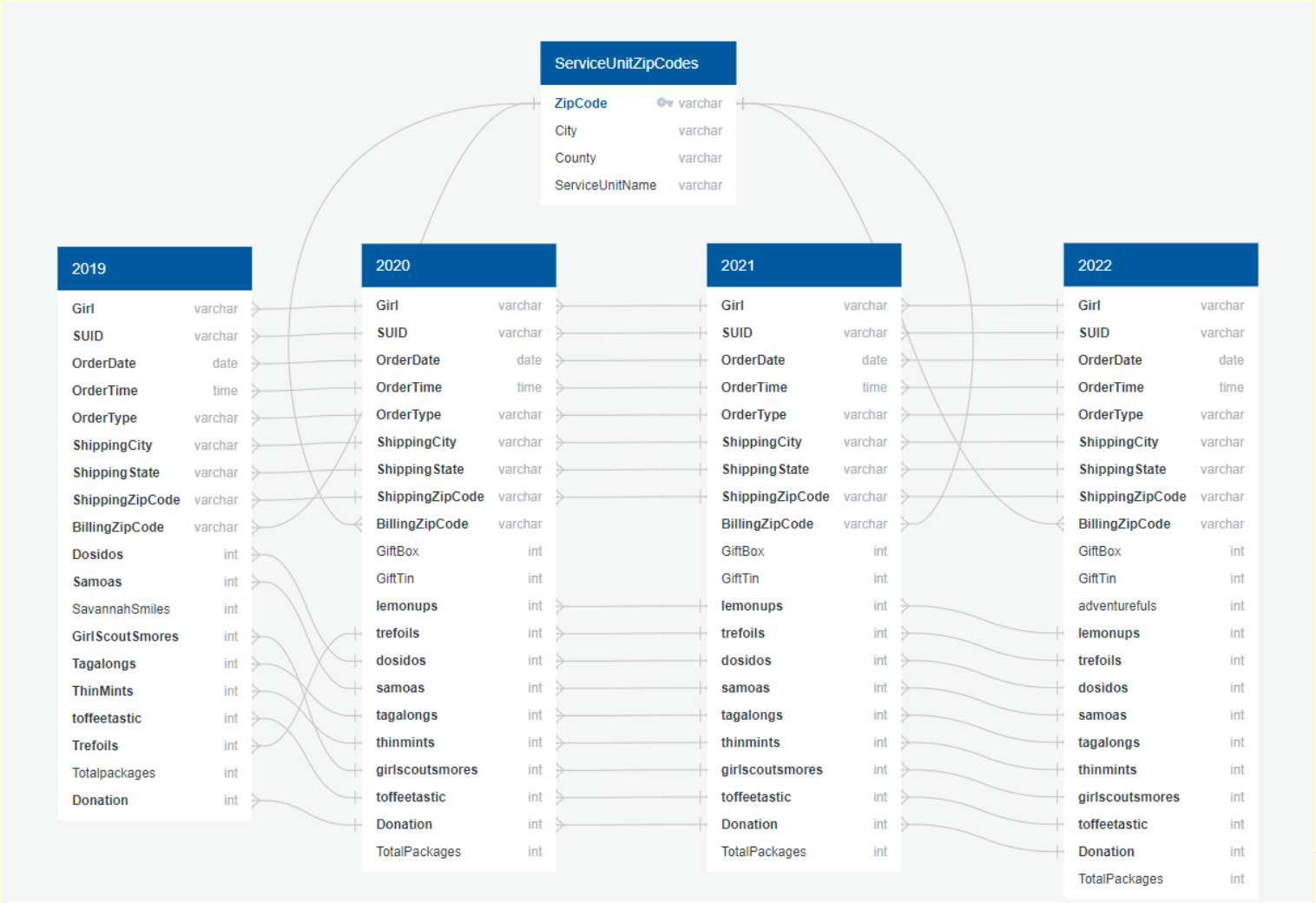
Extracted from the web:

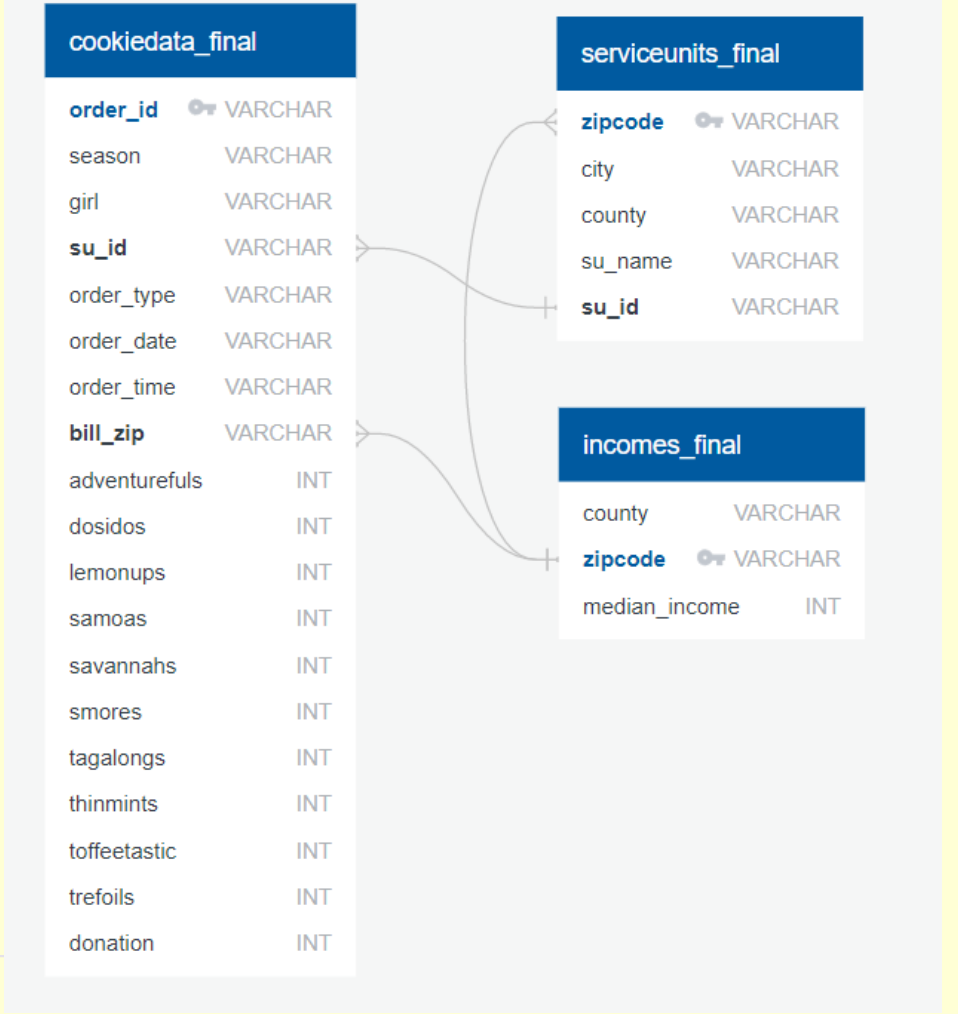
- Median Incomes for specific zip codes from 2020 census data
- geoJSON polygons for specific zip codes

Created:

- geoJSON with zip code center lat/long

PostgreSQL


ServiceUnitZipCodes
ZipCode varchar
City varchar
County varchar
ServiceUnitName varchar
2019
Girl varchar
SUID varchar
OrderDate date
OrderTime time
OrderType varchar
ShippingCity varchar
Shipping State varchar
ShippingZipCode varchar
BillingZipCode varchar
Dosidos int
Samoas int
SavannahSmiles int
GirlScoutSmores int
Tagalongs int
ThinMints int
toffeetastic int
Trefoils int
Totalpackages int
Donation int
2020
Girl varchar
SUID varchar
OrderDate date
OrderTime time
OrderType varchar
ShippingCity varchar
Shipping State varchar
ShippingZipCode varchar
BillingZipCode varchar
GiftBox int
GiftTin int
lemonups int
trefoils int
dosidos int
samoas int
tagalongs int
thinmints int
girlscoutsmores int
toffeetastic int
Donation int
TotalPackages int
2021
Girl varchar
SUID varchar
OrderDate date
OrderTime time
OrderType varchar
ShippingCity varchar
Shipping State varchar
ShippingZipCode varchar
BillingZipCode varchar
GiftBox int
GiftTin int
lemonups int
trefoils int
dosidos int
samoas int
tagalongs int
thinmints int
girlscoutsmores int
toffeetastic int
Donation int
TotalPackages int
2022
Girl varchar
SUID varchar
OrderDate date
OrderTime time
OrderType varchar
ShippingCity varchar
Shipping State varchar
ShippingZipCode varchar
BillingZipCode varchar
GiftBox int
GiftTin int
adventurefuls int
lemonups int
trefoils int
dosidos int
samoas int
tagalongs int
thinmints int
girlscoutsmores int
toffeetastic int
Donation int
TotalPackages int


cookiedata_final
order_id VARCHAR
season VARCHAR
girl VARCHAR
su_id VARCHAR
order_type VARCHAR
order_date VARCHAR
order_time VARCHAR
bill_zip VARCHAR
adventurefuls INT
dosidos INT
lemonups INT
samoas INT
savannahs INT
smores INT
tagalongs INT
thinmints INT
toffeetastic INT
trefoils INT
donation INT
serviceunits_final
zipcode VARCHAR
city VARCHAR
county VARCHAR
su_name VARCHAR
su_id VARCHAR
incomes_final
county VARCHAR
zipcode VARCHAR
median_income INT

Union (rather than Join)

```
-- Adventurefuls were not sold 2019-2021; 0 qty sold added
ALTER TABLE dc2019
ADD adventurefuls INT
DEFAULT '0';

ALTER TABLE dc2020
ADD adventurefuls INT
DEFAULT '0';

ALTER TABLE dc2021
ADD adventurefuls INT
DEFAULT '0';

-- create union
SELECT * INTO cookiedata FROM (
    SELECT order_id, season, girl, su_id, order_type, order_date, order_time, bill_zip, adventurefuls, dosid
    FROM dc2019
    UNION
    SELECT order_id, season, girl, su_id, order_type, order_date, order_time, bill_zip, adventurefuls, dosid
    FROM dc2020
    UNION
    SELECT order_id, season, girl, su_id, order_type, order_date, order_time, bill_zip, adventurefuls, dosid
    FROM dc2021
    UNION
    SELECT order_id, season, girl, su_id, order_type, order_date, order_time, bill_zip, adventurefuls, dosid
    FROM dc2022
) a
```

How the Cookie Crumbles

~Research Questions~

- Does median income predict individual digital cookie sales?

  Linear Regression Model

- Do sales of cookie types based on single digital transactions predict low or high income areas?

  Logistic Regression Model

Linear Regression Model

- X = median_income
  y = grand_total
- sklearn train-test-split (80%, 20%)

  Raw data

  Scaled data
- *sklearn LinearRegression()*
- *Calculations*
  - R-squared
  - Model coefficient/intercept
  - Mean squared error

```
# Assign X to 'median_income' and format data
X = lin_reg_df.median_income.values.reshape(-1, 1)
```

```
# Assign y to grand_total column
y = lin_reg_df.grand_total
```

```
# Split data into training (80%) and testing (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

```
# Create linear regression model
model = linear_model.LinearRegression()
```

```
# Fit the data
model.fit(X_train, y_train)
```

```
LinearRegression()
```

```
# Create predictions
y_pred = model.predict(X_test)
```

Linear Regression Results

Raw Data:

```
# View slope and y-intercept
print (f'Model coefficient is {model.coef_}')
print (f'Model intercept is {model.intercept_}')
```

```
Model coefficient is [8.10487992e-06]
Model intercept is 5.613845977492875
```

```
# Score the model--calculate R-squared
print (f'R-squared is equal to {model.score(X_test, y_test)}')
```

```
R-squared is equal to 0.0002452857309155343
```

```
# Calculate mean squared error
print (f'Mean squared error is', metrics.mean_squared_error(y_test, y_pred))
```

```
Mean squared error is 137.5848049593371
```

Scaled Data:

```
# View slope and y-intercept
print (f'Model coefficient is {lin_model.coef_}')
print (f'Model intercept is {lin_model.intercept_}')
```

```
Model coefficient is [0.19980955]
Model intercept is 6.274737605804111
```

```
# Score the model--calculate R-squared
print (f'R-squared is equal to {lin_model.score(X_test_scaled, y_test)}')
```

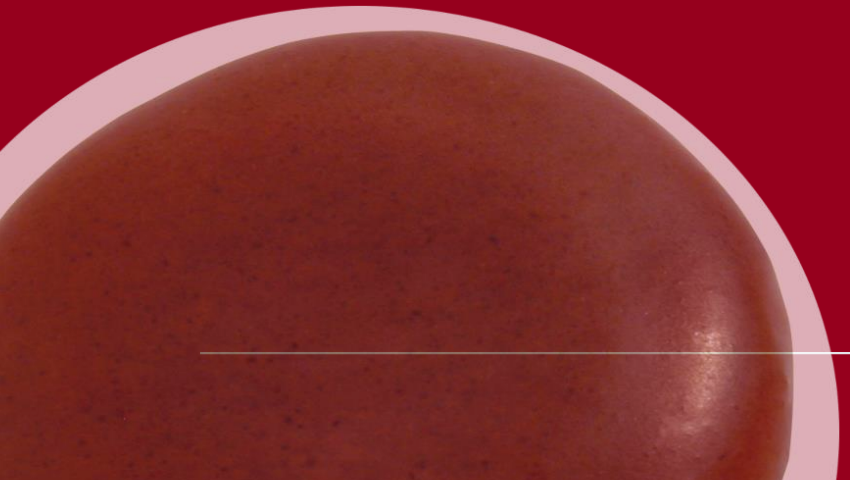
```
R-squared is equal to 0.0002452857309155343
```

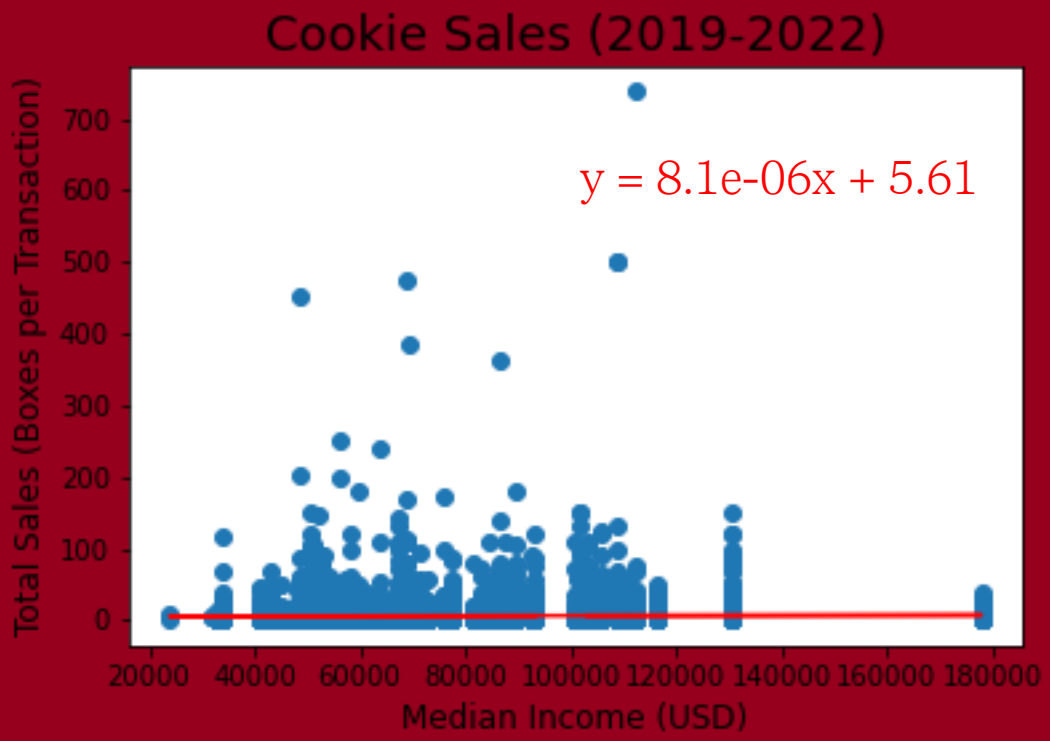
```
# Calculate mean squared error
print (f'Mean squared error is', metrics.mean_squared_error(y_test, y_preds))
```

```
Mean squared error is 137.5848049593371
```

Linear Regression Conclusions

- The model is not a good predictor of boxes sold per digital transaction based on median income.
- Individuals are just as likely to buy the same number of boxes in any income area.
- The high MSE can be attributed to high variation and outliers in sales data.





Logistic Regression Model

- X = all cookie columns in df
  y = high_low_income
- sklearn train-test-split (80%, 20%)
  - *Raw data*
  - *Scaled data*
  - *SMOTEENN*
  - *SMOTE*
- *sklearn LogisticRegression()*
- *Calculations*
  - Accuracy/Balanced accuracy score
  - Confusion matrix
  - Classification report/Imbalanced classification report

```
# Resample the training data with SMOTE
X_resampled, y_resampled = SMOTE(random_state=1, sampling_strategy='auto').fit_resample(
    X_train, y_train
)
Counter(y_resampled)
```

```
Counter({1: 72405, 0: 72405})
```

```
# Train the Logistic Regression model using the resampled data
model = LogisticRegression(solver='lbfgs', random_state=1)
```

```
# Fit the model
model.fit(X_resampled, y_resampled)
```

```
LogisticRegression(random_state=1)
```

```
# Make predictions
y_pred = model.predict(X_test)
results = pd.DataFrame({"Prediction": y_pred, "Actual": y_test}).reset_index(drop = True)
results.head(10)
```

Logistic Regression Results

SMOTEENN:

```
Balanced accuracy score: 0.5023024453714952
```

Confusion Matrix

	Predicted Low Income	Predicted High Income
Actual Low Income	180	7949
Actual High Income	424	23752

Classification report imbalanced

	pre	rec	spe	f1
0	0.30	0.02	0.98	0.04
1	0.75	0.98	0.02	0.85

SMOTE:

```
Balanced accuracy score: 0.5439532736477379
```

Confusion Matrix

	Predicted Low Income	Predicted High Income
Actual Low Income	4884	3245
Actual High Income	12400	11776

Classification report imbalanced

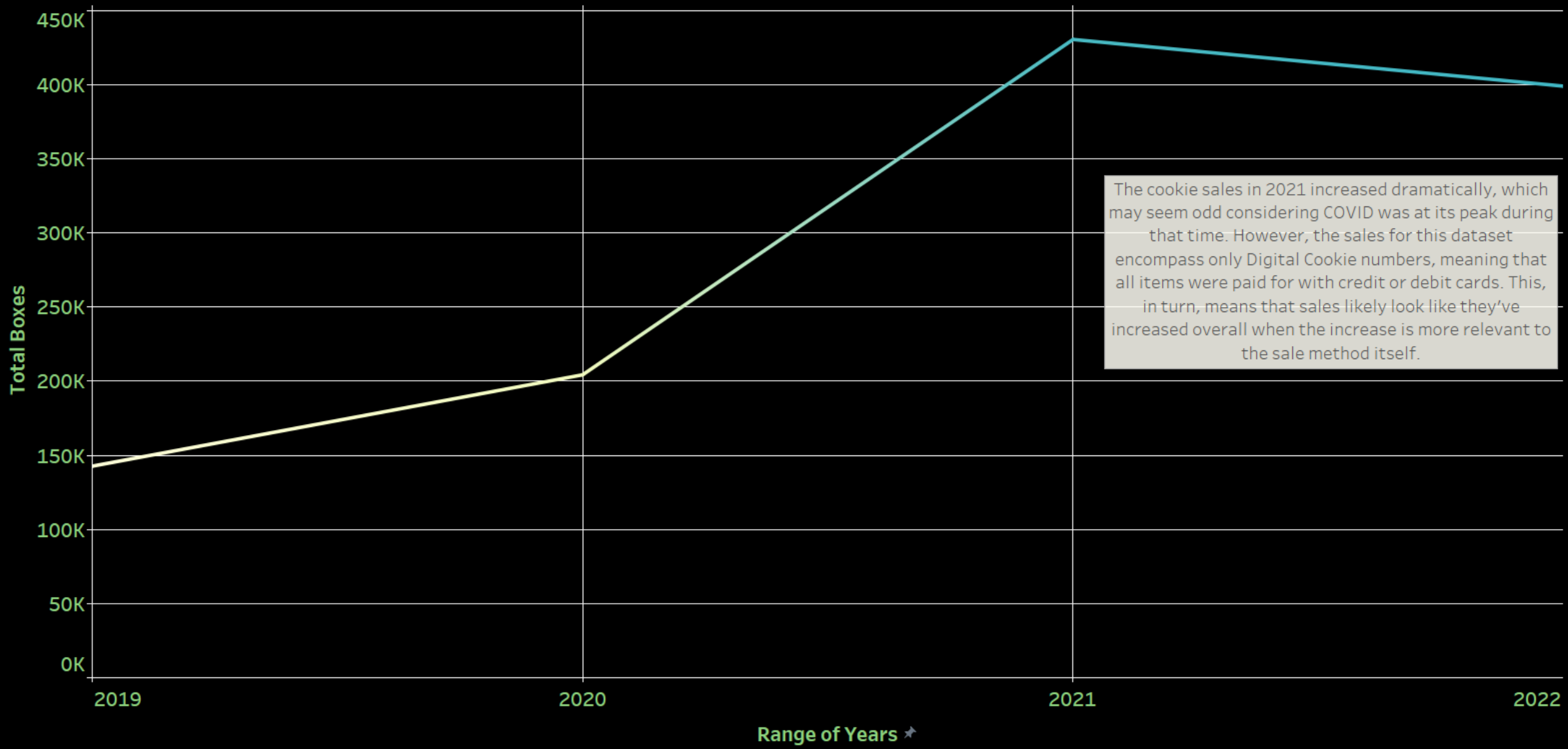
	pre	rec	spe	f1
0	0.28	0.60	0.49	0.38
1	0.78	0.49	0.60	0.60

Logistic Regression Conclusions

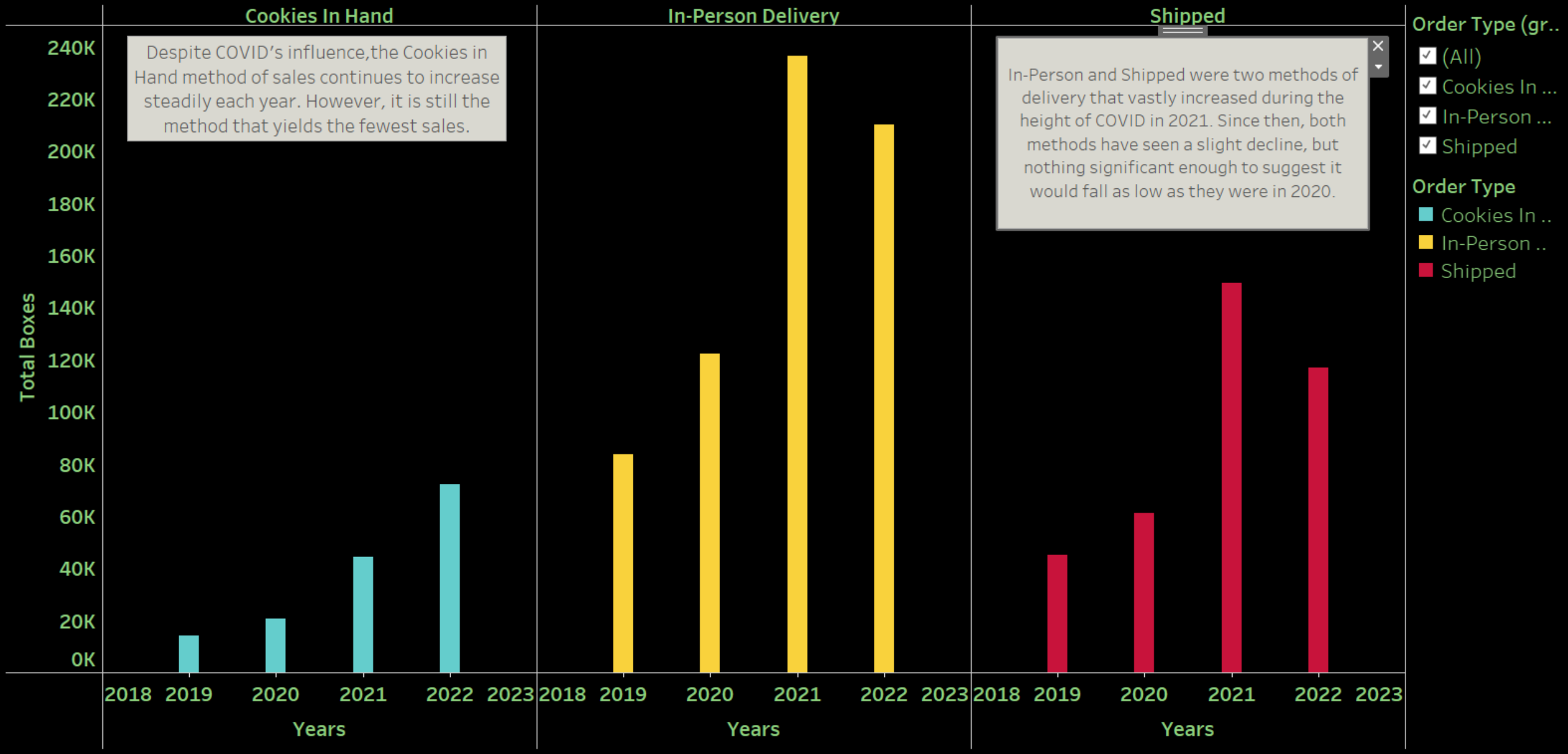
- The model is not a good predictor of high or low income areas when using digital sale transactions.
- Recall improved for predicting low income areas but decreased for high income areas with SMOTE.

Boxes Sold Over the Past Four Years

Story



Delivery Method Story



Best of the Batch Story

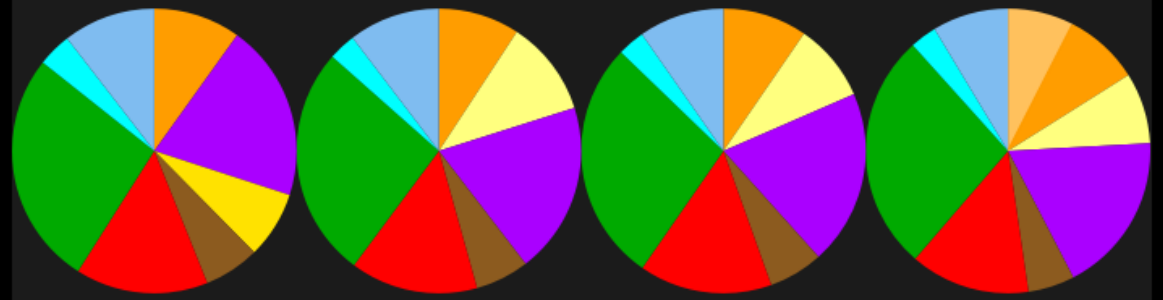
(Most popular cookies overall) 2019-2022

Over the years, it looks as though certain cookies have retained their popularity, the highest three being Thinmints, Samoas, and Tagalongs.

(Most popular cookies per year) 2019-2022

Season

2019 2020 2021 2022



However, we notice a dip in the sales for these cookies in 2022. This can be attributed to both COVID's impact, and the cookie shortage going on at the time. If not for these factors, the data suggests that sales would have continued to increase.

For Reference, the Thinmint sales ranged per the below:

34,016 in 2019
48,558 in 2020
107,067 in 2021
98,667 in 2022

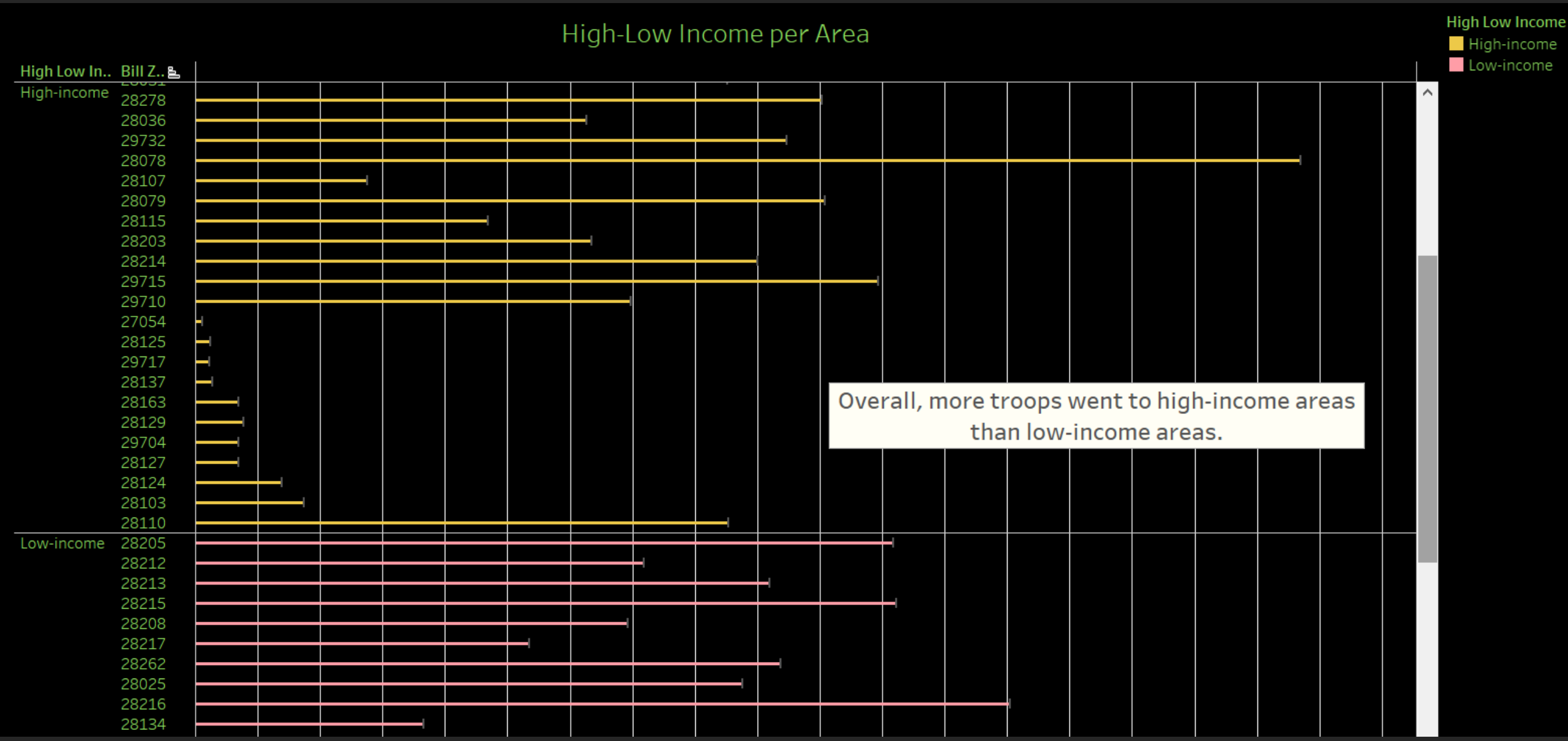
- Dosidos
- Lemonups
- Samoas
- Savannah
- Smores
- Tagalongs
- Thinmints
- Toffeetas
- Trefoils

Season

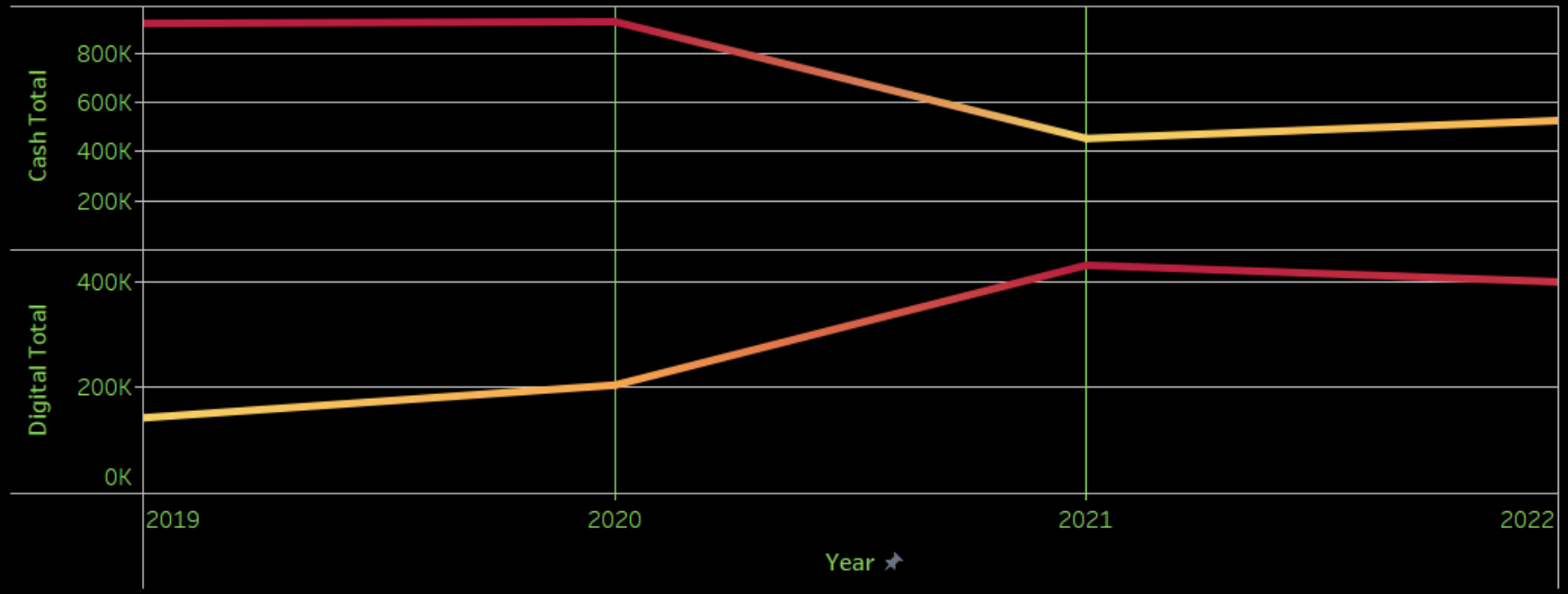
- [x] (All)
- [x] 2019
- [x] 2020
- [x] 2021
- [x] 2022

Troops per Zip Code

Story



Cash vs Digital



Digital Total

142,631 430,3

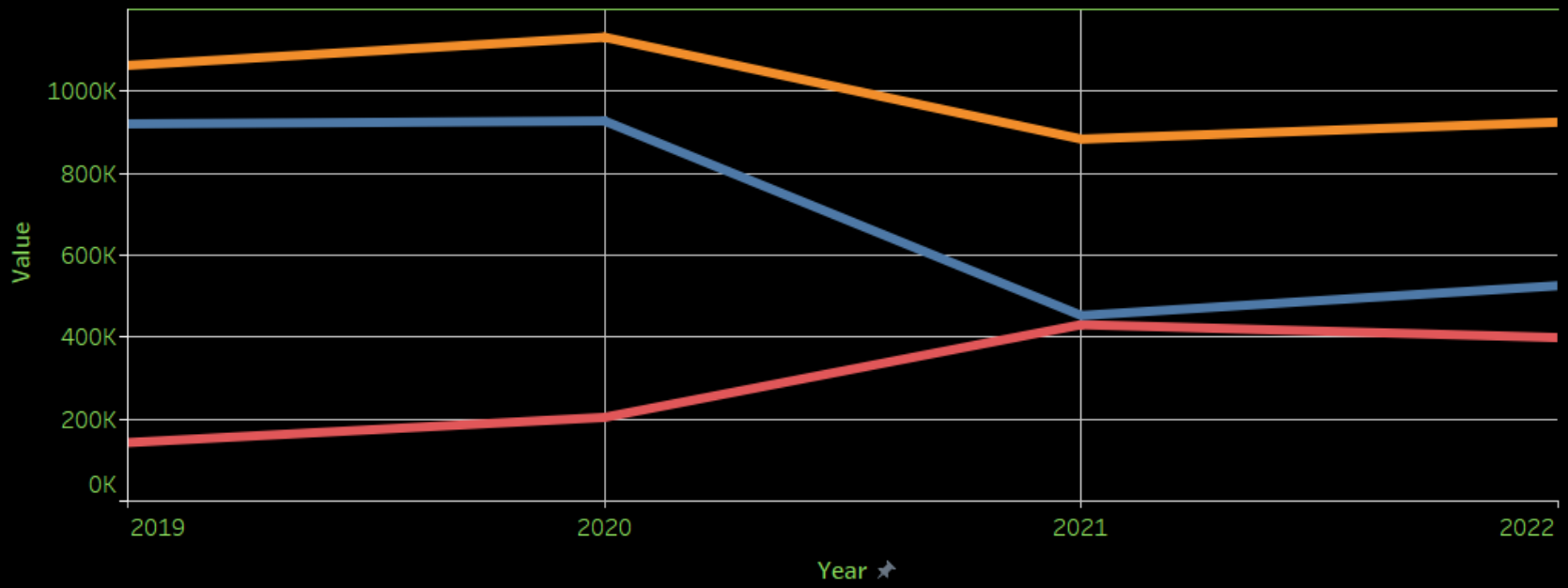
Cash Total

452,541 927,0

As expected, likely due to the impact of COVID, the digital sales went up while the cash sales went down.
In 2022 there looks to be a slight rise in cash purchases now that people are venturing outside again. This implies that cash sales will continue going up, just at a slower pace.

Story

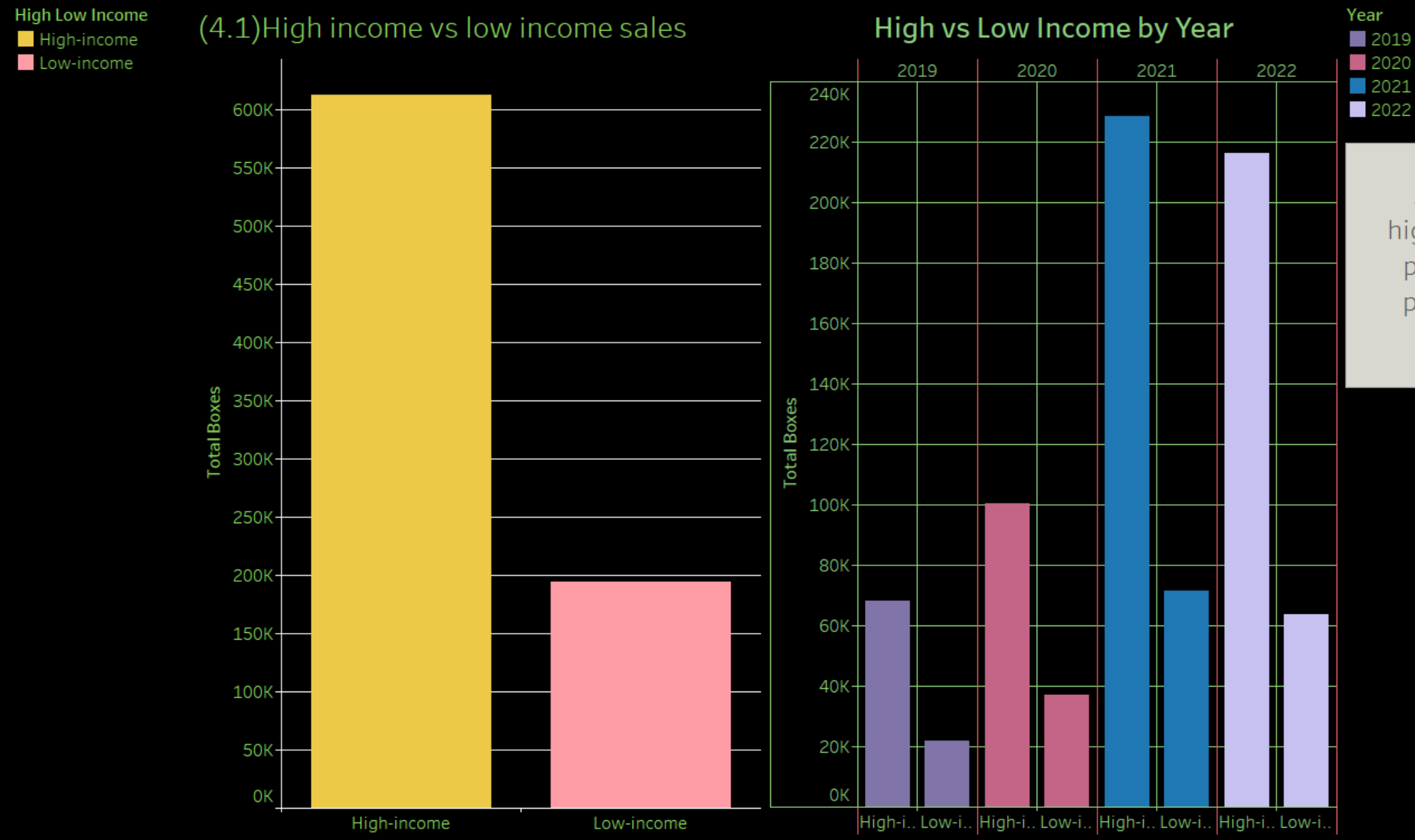
COVID Impacts on Sales Type



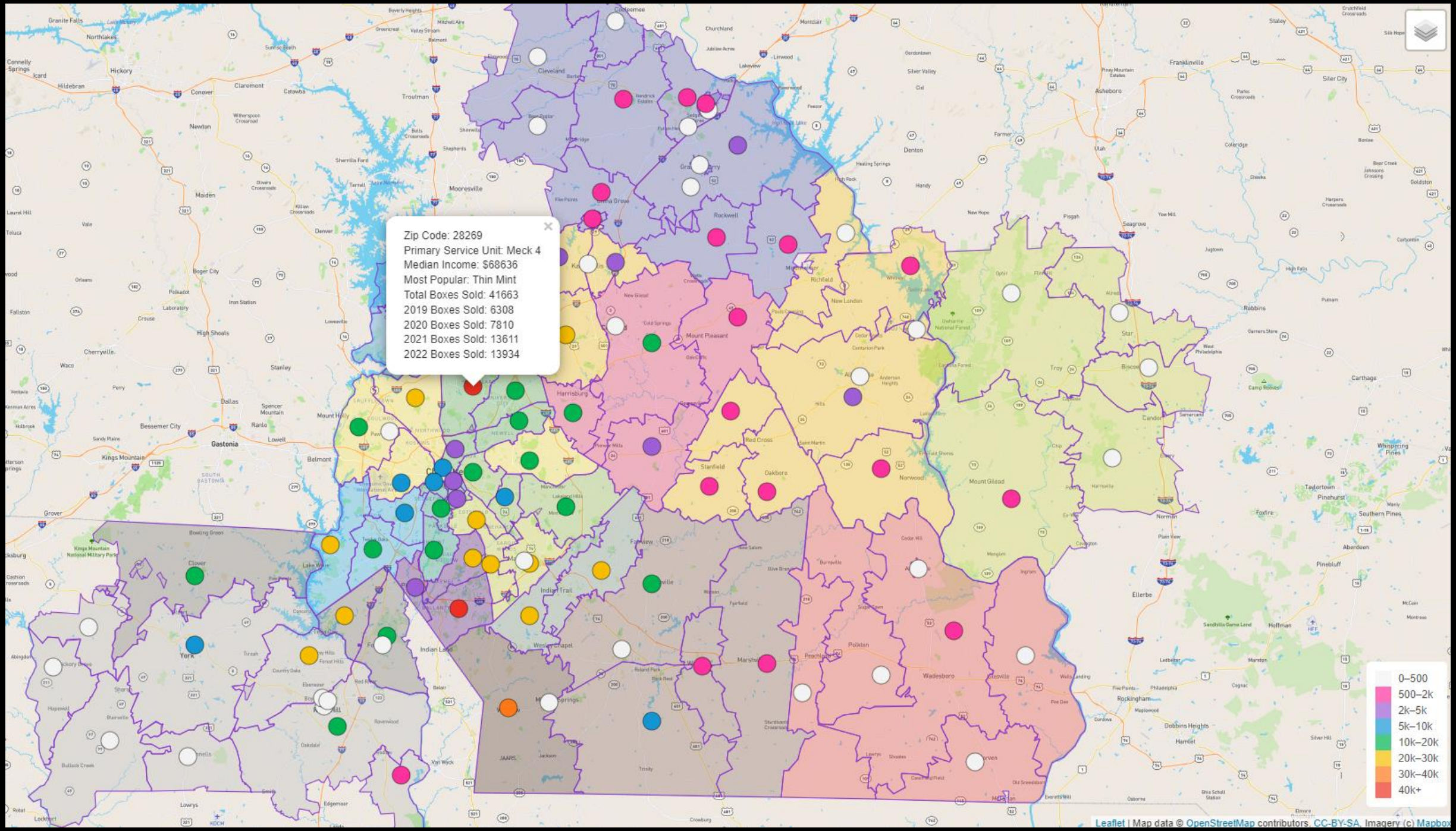
Measure Names
- Cash Total
- Digital Total
- Total Overall

Boxes per High and Low Income Areas

Story



At a glance, we see that higher-income areas tend to purchase the most boxes, providing troops with the majority of their sales.


Zip Code: 28269
Primary Service Unit: Meck 4
Median Income: $68636
Most Popular: Thin Mint
Total Boxes Sold: 41663
2019 Boxes Sold: 6308
2020 Boxes Sold: 7810
2021 Boxes Sold: 13611
2022 Boxes Sold: 13934
0–500
500–2k
2k–5k
5k–10k
10k–20k
20k–30k
30k–40k
40k+
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery (c) Mapbox

You Gotta Risk it for the Biscuit

Closing Remarks

- Machine Learning Models were not useful for this type of dataset.
- As expected COVID did impact sales but did increase app usage.
- The data showed that more boxes were purchased in high-income areas, however, it was found that more troops sold in those areas to begin with.
- Thin Mints are the undisputed champion.
- The map showed the areas with the highest digital sales, with zip code 28269 coming out as the top area for number of boxes sold.