FORECASTING DEMAND

PREDICTING PERISHABLE GOODS EXPENDITURES WITH MULTIPLE REGRESSION

FORECASTING DEMAND

Austin Dowd

- Student at Western Governor's University
- Master of Science, Data Analytics

- Current Role: Marketing Analytics
- IoT Connectivity enabling wireless smart device infrastructure

THE SUPPLY CHAIN

- Perishable goods have a limited shelf life and getting them through production to customers right when they need them is a challenge at every level of the supply chain.
- The ability to foresee demand for perishable goods gives business partners insights to reduce revenue loss from spoilage in cases of over stocking or missed sales through stock outs.
- Additional benefits to demand forecasting include better plan to account for:
  - Labor
  - Materials
  - Equipment
  - Budgeting and Cash Flow

THE DATA AND HYPOTHESIS

	household_code	fruit_paid	vege_paid	ln_fruit_paid	ln_vege_paid	super5_cat	con5_10	regional_10000	dest5_10	mix5_10
0	1	166.260000	148.790000	5.113553	5.002536	1	0.2	1.548161	27.900000	2.966195
1	2	14.030000	39.890000	2.641198	3.686126	1	0.2	1.548161	27.900000	2.966195
2	3	245.490000	425.100000	5.503256	6.052324	1	0.2	1.548161	27.900000	2.966195
3	4	9.720000	142.570000	2.274186	4.					
4	5	330.579999	480.969999	5.800848	6.					
5	6	36.860001	39.220000	3.607127	3.					
6	7	35.340000	93.950000	3.565016	4.					
7	8	79.999999	92.490000	4.382027	4.					
8	9	233.260000	170.889999	5.452154	5.					
9	10	82.060000	213.180001	4.407451	5.					
10	11	309.529995	44.140000	5.735055	3.					
11	12	17.790000	132.450000	2.878637	4.					
12	13	93.120000	177.249999	4.533889	5.					
13	14	428.819999	342.770000	6.061037	5.					
14	15	63.010000	31.630000	4.143293	3.					
15	16	156.240000	286.240001	5.051393	5.					
16	17	46.490000	469.729999	3.839237	6.					
17	18	39.490000	192.059999	3.676048	5.					
18	19	24.810000	76.430000	3.211247	4.					
19	20	161.919998	233.590001	5.087102	5.					

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22448 entries, 0 to 22447
Data columns (total 24 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   household_code  22448 non-null  int64
 1   fruit_paid      22448 non-null  float64
 2   vege_paid       22448 non-null  float64
 3   ln_fruit_paid   22310 non-null  float64
 4   ln_vege_paid    22334 non-null  float64
 5   super5_cat      22448 non-null  int64
 6   con5_10         22448 non-null  float64
 7   regional_10000  22448 non-null  float64
 8   dest5_10        22448 non-null  float64
 9   mix5_10         22448 non-null  float64
 10  design5_10      22448 non-null  float64
 11  auto_bg100      22448 non-null  float64
 12  poverty_tr100   22448 non-null  float64
 13  edu             22448 non-null  int64
 14  income          22448 non-null  int64
 15  race            22448 non-null  int64
 16  hh_size         22448 non-null  int64
 17  age             22448 non-null  int64
 18  marital_status  22448 non-null  int64
 19  children        22448 non-null  int64
 20  employed_new    22448 non-null  int64
 21  urban           22448 non-null  int64
 22  msa             22448 non-null  int64
 23  zip             22448 non-null  int64
dtypes: float64(11), int64(13)
memory usage: 4.1 MB
```

Question:

- Can a regression model be built to predict a household's expenditures on perishable goods (fruits and vegetables) based on the available research data?

Hypothesis:

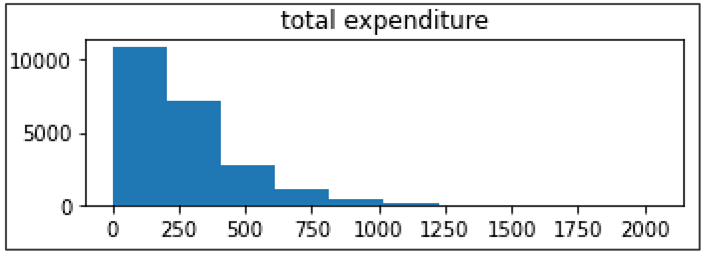
- It is possible to build a regression model to statistically significantly predict household expenditures on perishable goods based on the available variables in the research data.

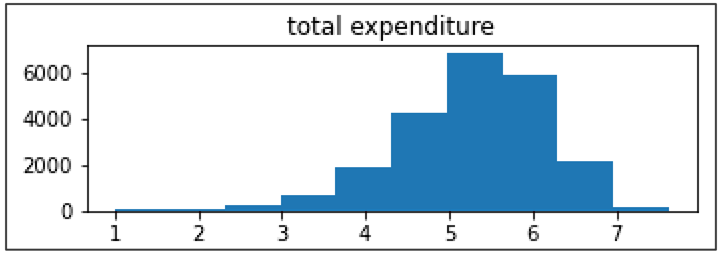
THE ANALYSIS PROCESS - PREPARATION

- Preparation
  - Download and load data into Pandas Dataframe
  - Drop unnecessary and unexplained variables
  - Sum fruit & vegetable expenditures into total expenditures variables
  - Convert Categorical Variables to Dummy Data
  - Examine distributions of continuous numerical variables
  - Logarithmically Transform skewed data

	edu	income	race	marital_status	urban
22089	college or higher	20000-59999	white	divorced	non-urban
1566	no female head	20000-59999	white	seprated/single	urbanized area
20714	college or higher	60000+	asian	married	non-urban
4062	college or higher	60000+	white	married	urbanized area
12615	college or higher	60000+	white	married	urbanized area
5730	college or higher	60000+	white	married	non-urban
1801					
1428					
1195					
1592					

	high school or below	college or higher	no female head	below 20000	20000-59999	60000+	white	black	asian	other	married	widowed	divorced	seprated/single	urbanized area	urban cluster	non-urban
0	0	1	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0
1	0	1	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0
2	0	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0
3	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1
4	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1





THE ANALYSIS PROCESS – INITIAL MODELING

OLS Regression Results

Dep. Variable:	total expenditure	R-squared:	0.121
Model:	OLS	Adj. R-squared:	0.120
Method:	Least Squares	F-statistic:	123.3
Date:	Mon, 05 Sep 2022	Prob (F-statistic):	0.00
Time:	12:53:27	Log-Likelihood:	-27900.
No. Observations:	22448	AIC:	5.585e+04
Df Residuals:	22422	BIC:	5.606e+04
Df Model:	25		
Covariance Type:	nonrobust		

					[0.025	0.975]
					0.008	0.019
					-0.015	0.003
					0.013	0.043
					0.001	0.030
					-0.033	-0.015
					-0.021	0.028
					-0.047	-0.015
					-0.022	-0.009
					0.070	0.114
					0.138	0.220
					-0.009	0.007
					7.19e-07	1.29e-06
					7.96e-07	1.13e-07
college or higher	0.7701	0.012	64.096	0.000	0.747	0.794
high school or below	0.6531	0.014	46.658	0.000	0.626	0.681
no female head	0.4612	0.017	27.340	0.000	0.428	0.494
20000-59999	0.6054	0.012	49.927	0.000	0.582	0.629
60000+	0.8225	0.012	66.963	0.000	0.798	0.847
below 20000	0.4565	0.017	27.260	0.000	0.424	0.489
asian	0.6382	0.027	23.257	0.000	0.584	0.692
black	0.3106	0.018	16.895	0.000	0.275	0.347
other	0.4804	0.023	20.781	0.000	0.435	0.526
white	0.4552	0.013	35.722	0.000	0.430	0.480
divorced	0.3933	0.014	28.728	0.000	0.366	0.420
married	0.6507	0.016	41.217	0.000	0.620	0.682
seprated/single	0.3185	0.015	21.954	0.000	0.290	0.347
widowed	0.5219	0.020	25.544	0.000	0.482	0.562
non-urban	0.6366	0.013	48.540	0.000	0.611	0.662
urban cluster	0.6157	0.022	28.148	0.000	0.573	0.659
urbanized area	0.6322	0.014	46.511	0.000	0.606	0.659
intercept	1.8845	0.024	79.750	0.000	1.838	1.931

Omnibus:	1901.420	Durbin-Watson:	1.951
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2730.123
Skew:	-0.686	Prob(JB):	0.00
Kurtosis:	4.019	Cond. No.	1.33e+16

- R-squared: 0.121
- Condition Number: 1.33e+16
- Reduce number of variables:
  - P-values below 0.05
  - T-statistic above 0

THE ANALYSIS FINDINGS – FINAL MODEL

Dep. Variable:	total expenditure	R-squared:	0.106
Model:	OLS	Adj. R-squared:	0.105
Method:	Least Squares	F-statistic:	235.8
Date:	Mon, 05 Sep 2022	Prob (F-statistic):	0.00
Time:	12:53:27	Log-Likelihood:	-22499.
No. Observations:	17958	AIC:	4.502e+04
Df Residuals:	17948	BIC:	4.510e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	4.3484	0.046	94.906	0.000	4.259	4.438
60000+	0.4024	0.025	16.012	0.000	0.353	0.452
college or higher	0.3561	0.025	14.372	0.000	0.308	0.405
20000-59999	0.1560	0.025	6.207	0.000	0.107	0.205
non-urban	0.0697	0.033	2.133	0.033	0.006	0.134
high school or below	0.2414	0.028	8.769	0.000	0.187	0.295
urbanized area	0.1074	0.032	3.338	0.001	0.044	0.170
married	0.3219	0.018	17.469	0.000	0.286	0.358
white	0.0401	0.017	2.338	0.019	0.006	0.074
divorced	0.0082	0.022	0.381	0.703	-0.034	0.050

Omnibus:	1430.606	Durbin-Watson:	1.963
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2013.331
Skew:	-0.662	Prob(JB):	0.00
Kurtosis:	3.967	Cond. No.	18.4

- Test and Training Data
- R-squared: 0.106
- Condition Number: 18.4

Not statistically significant

Can not reject the null hypothesis

TOOLS & TECHNIQUES

Python
programming language

- Widely used in data science
- Large selection of libraries for data management, statistics and machine learning

Pandas
Data management library

- Data frame management
- Functions for exploring and manipulating data frames

Matplotlib
Data visualization library

- Ideal for visualizing data and analysis results

Statsmodels
Statistical modeling library

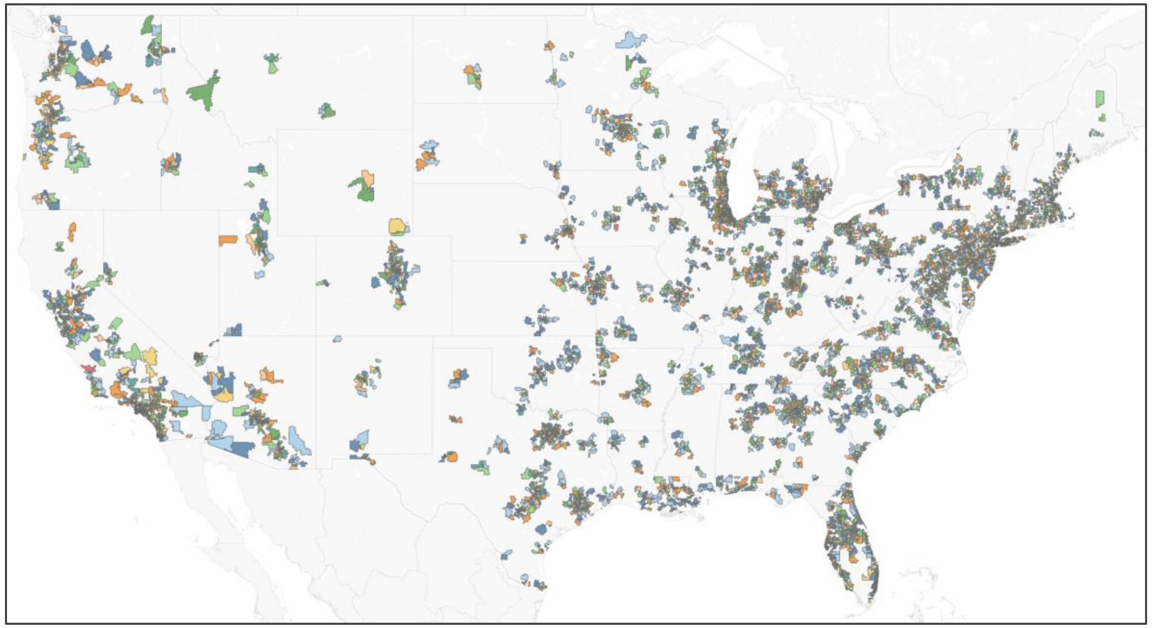
- Suited for statistical modeling
- Multiple functions for predictive modeling

Ordinary Least Squares
Regression Model

- Capable of multivariate regression
- Works with numerical and categorical data

PROPOSED ACTION

- 1. Seek more data
  - The results of this regression model indicate the data available was not well suited for regression modeling. Additional variables for these households can provide more options for modeling and discovering better predictors.
- 2. Study further with a different model
  - Regression modeling did not result in accurate predictions. Developing new research questions and examining the capabilities of other modeling techniques like classification may result in a statistically significant model that can provide actionable insights
- 3. Examine predictive modeling by groups
  - Regression modeling may be more successful at examining expenditures by groups within the data. Exploring predictive options with regression modeling for zip codes or metropolitan statistical areas. Having area wide forecasting for demand still meets the needs of most companies within a supply chain.



EXPECTED BENEFITS

1. Production planning for future demand
2. Labor management at all levels of the supply chain
3. Reduced costs by minimizing storage needs during shipping and distribution
4. Retail inventory management
5. Planned downtime for equipment maintenance during off peak times

THANK YOU