MATHER FINAL PRESENTATION 4/29

Noah Johnson, Katie Foster, & Jackson King

TABLE OF CONTENTS

- Hypotheses
- First Analysis
- Descriptive Analytics
- First Regression Model
- Continued Regression Model
- Scenario Analysis
- Retention Model
- Recommendations

FIRST HYPOTHESES

1. The weekly price and operating margin are associated and positively correlated

2. Frequency of delivery can be affecting delivery cost

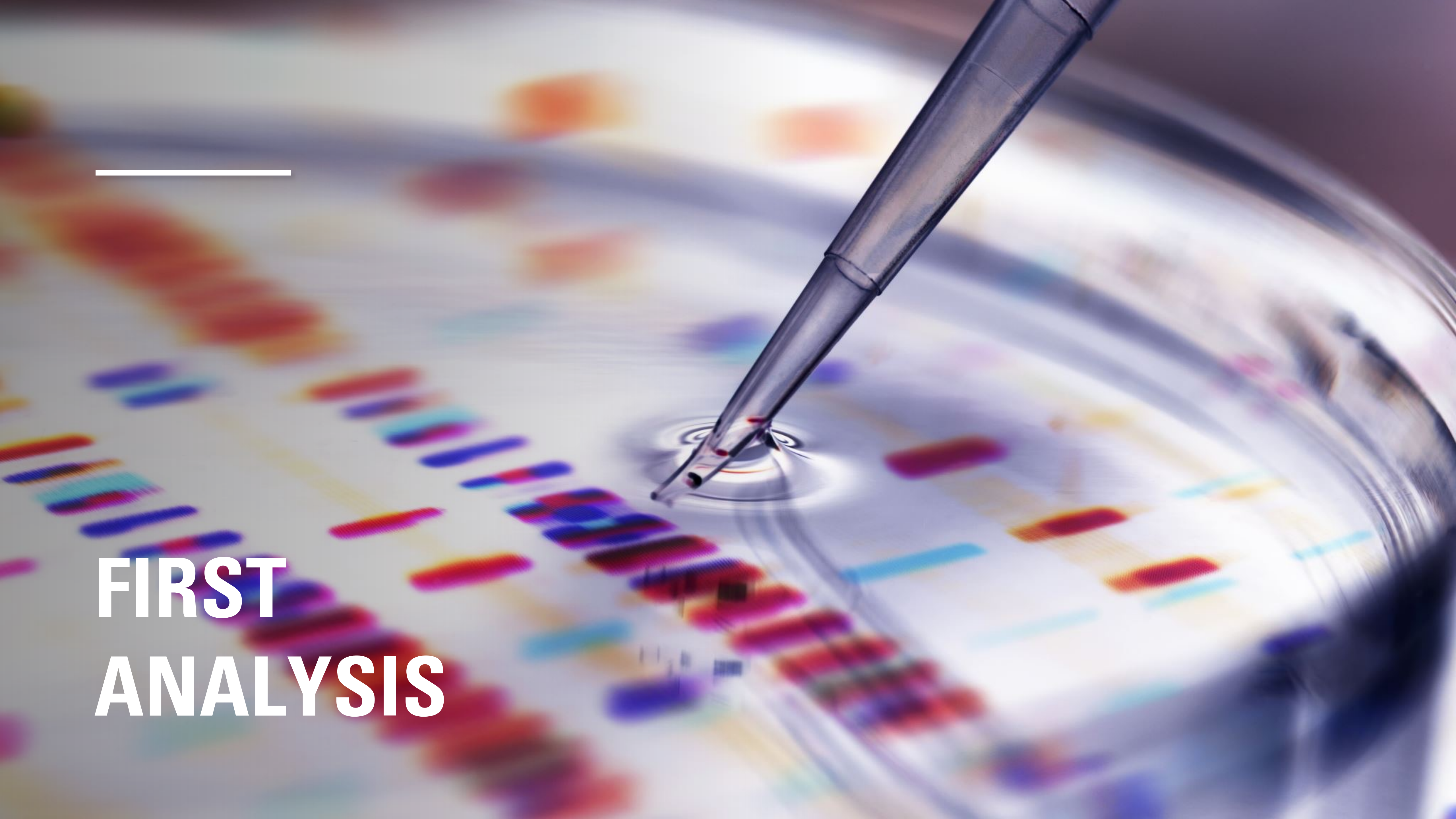
3. Operating margin can be predicted by factors like weekly price, delivery cost, and frequency of delivery. These should be linearly related

CONTINUED HYPOTHESES

1. If we can predict the stay/leave variable with high accuracy, we can predict the 2023 retention

2. If the company pushes 4+ day plans, they can keep active subscribers longer

3. The closer to Bangor the customer is, the more likely they are to stay with the company



FIRST ANALYSIS

DATA AND DATA CHANGES

Original Data:

7 spreadsheets: ‘16, ‘17, ‘18, ‘19, ‘20, ‘21, ‘22

~ 18,000 – 30,000 observations per year

Filters: None

Issues:

- avg_inc in 2 different formats (40 v 40,000)
- Subrate has 382 values

Final Data:

1 spreadsheet: 2016 – 2022

~ 125,000 instances

Filters:

- States: Maine only
- Status: Active only

Added Values:

- Tenure
- Distance (in miles to Bangor from mailcity)
- Stay/Leave binary variable

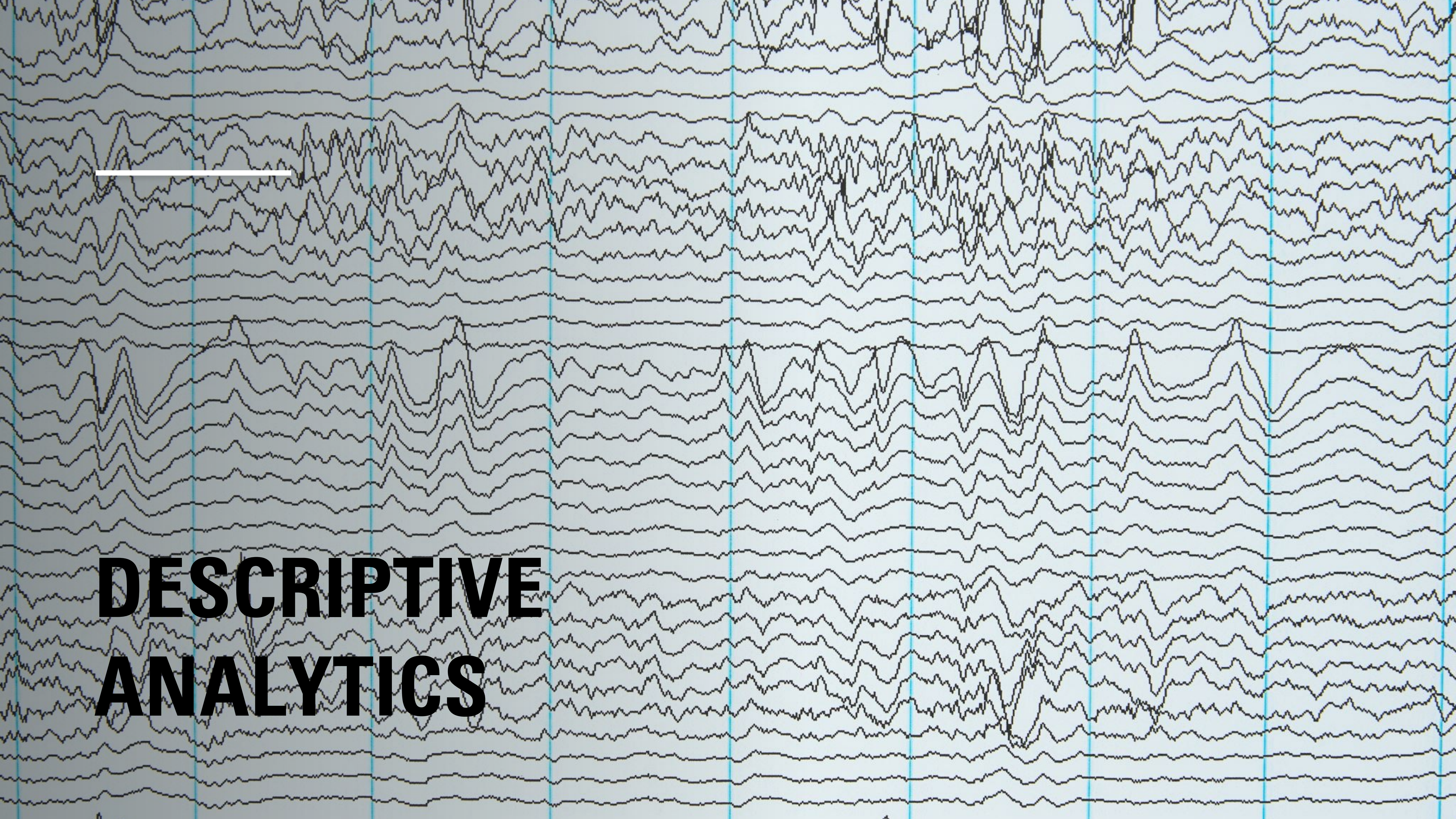
MILES TO BANGOR VARIABLE

- Example:

City	Latitude	Longitude
Bangor (Origin)	44.8012	-68.7778
Abbot	45.1800	-69.4675

Meters= 68671
Miles= 43

1. retrieved generated latitude and longitude data from Tableau
2. Used the "Haversine Equation" to generate the distance in meters between 2 sets of latitude and longitude coordinates
3. converted meters to miles



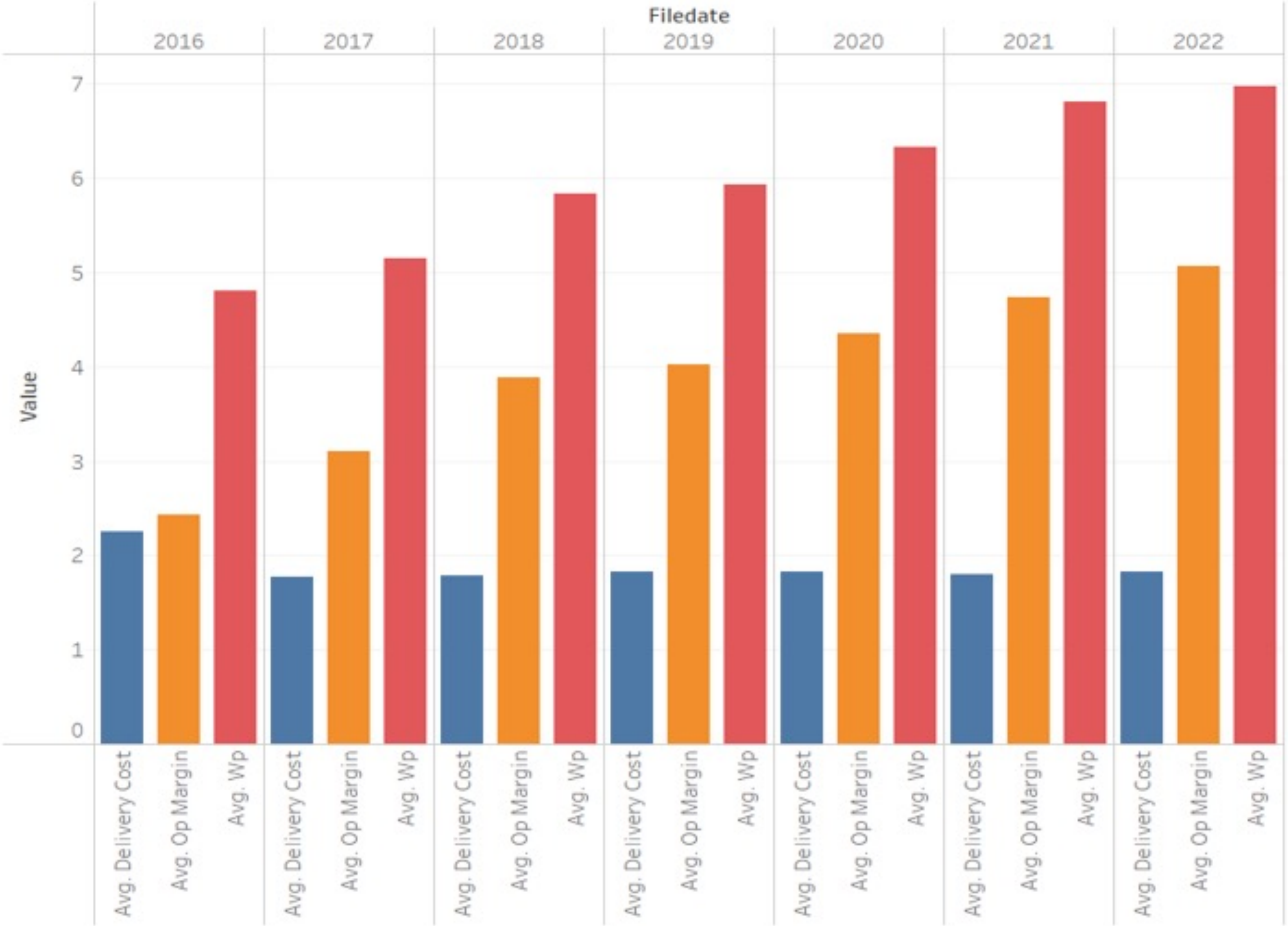
DESCRIPTIVE ANALYTICS

CHANGES OVER TIME

Measure Names

- 🟦 Avg. Delivery Cost
- 🟧 Avg. Op Margin
- 🟥 Avg. Wp

Trends from 2016-2022



INTRO ANALYTICS

99.3% of active subscribers are mailed their paper to Maine

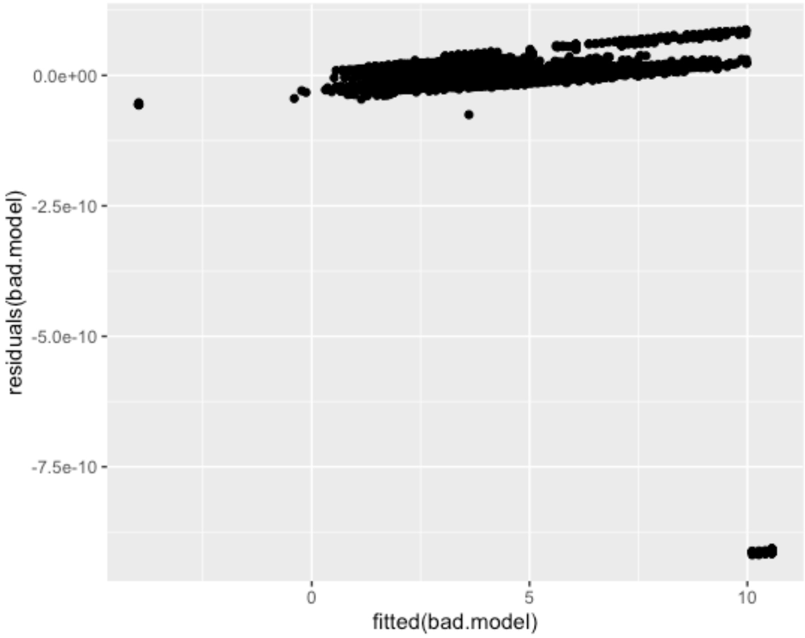
Distinct count of active subscribers in Maine: 32,693

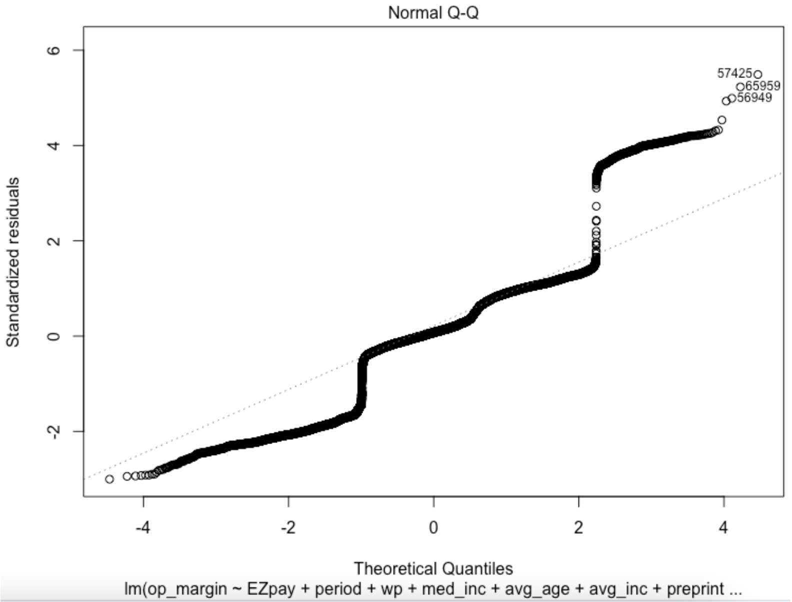
Nearly half of all customers have less than 2 years of tenure

FIRST REGRESSION MODEL

INITIAL REGRESSION MODEL

- Predicting operating margin based on the quantitative variables only
- Adjusted R-squared: .97
- Issues:
    - Multicollinearity
    - Did not meet the assumptions of linear regression
    - Did not include categorical data





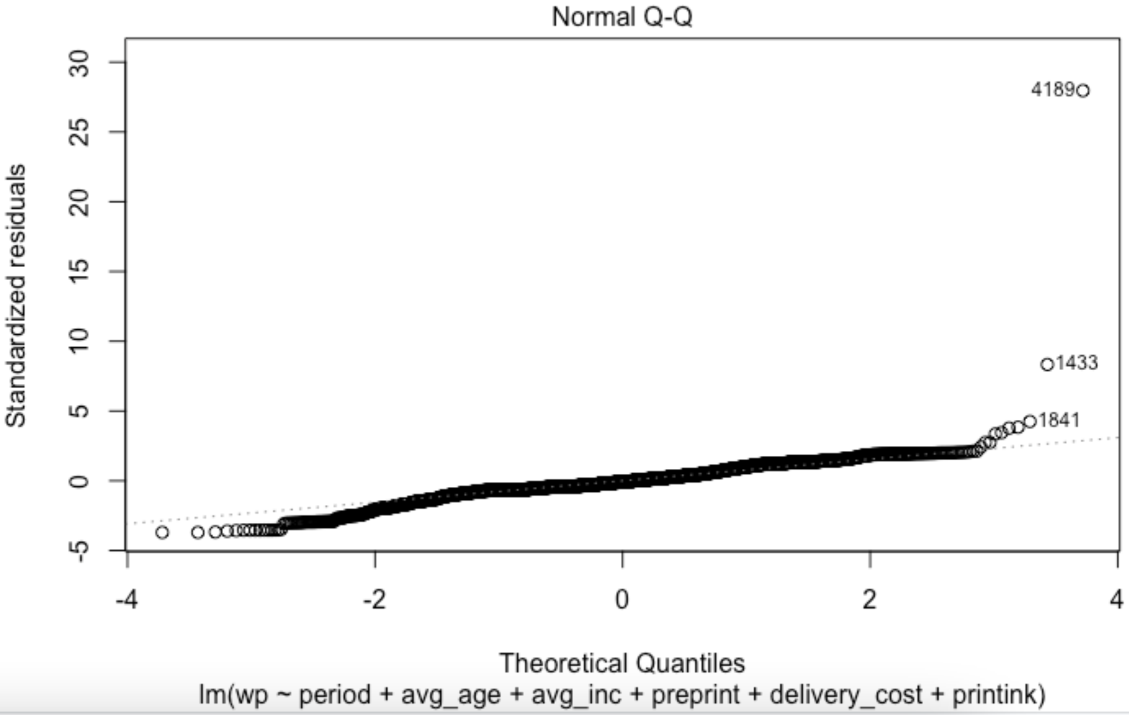
INITIAL WEEKLY PRICE REGRESSION MODEL

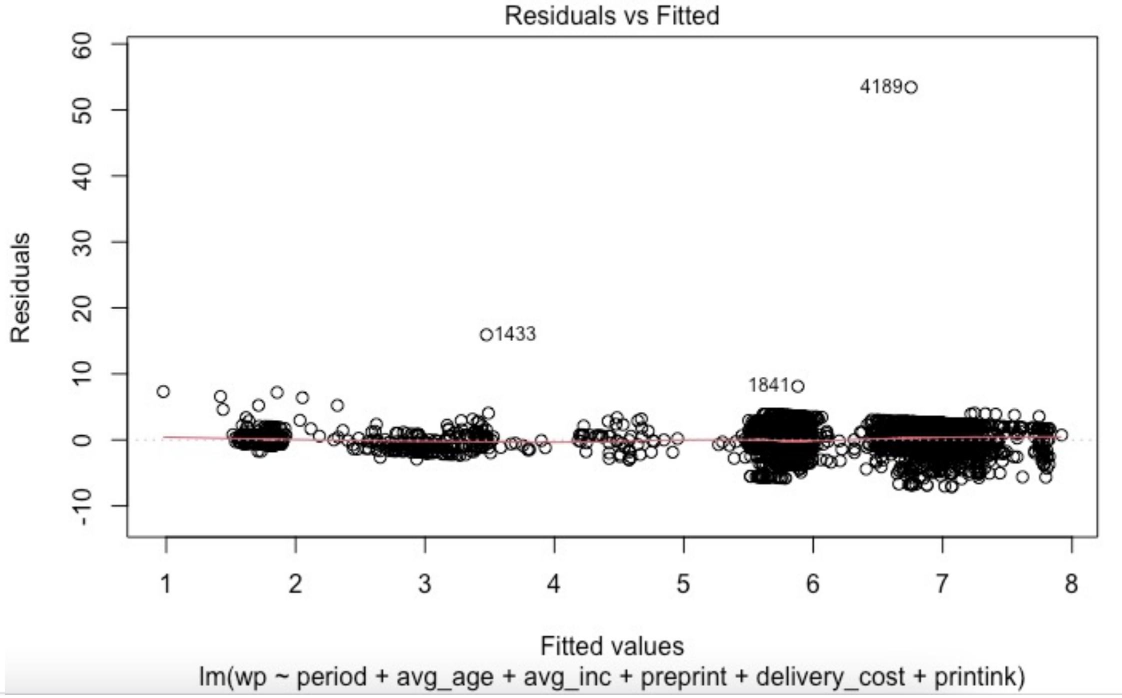
- Period
- Avg_age
- Avg_inc
- Preprint
- Delivery_cost
- Printink

R-Squared = 0.323

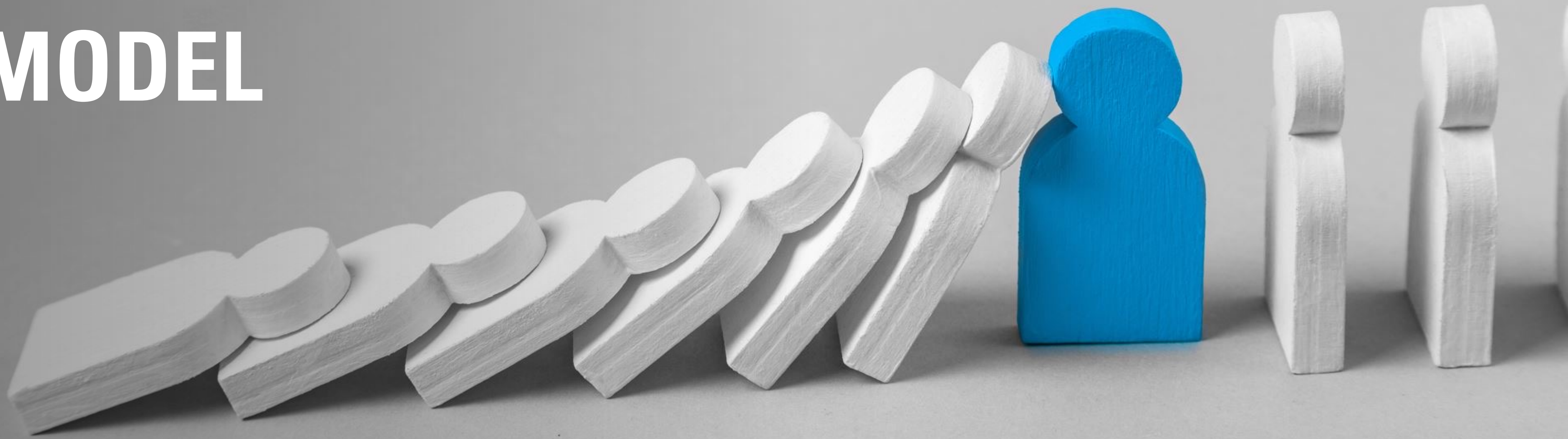
- Assumptions are better for weekly price than operating margin
- Model needs improvement by adding categorical variables and removing multicollinearity

WP REGRESSION ASSUMPTIONS





CONTINUED REGRESSION MODEL



MODEL 1: LINEAR REGRESSION FOR WEEKLY PRICE

Quantitative:

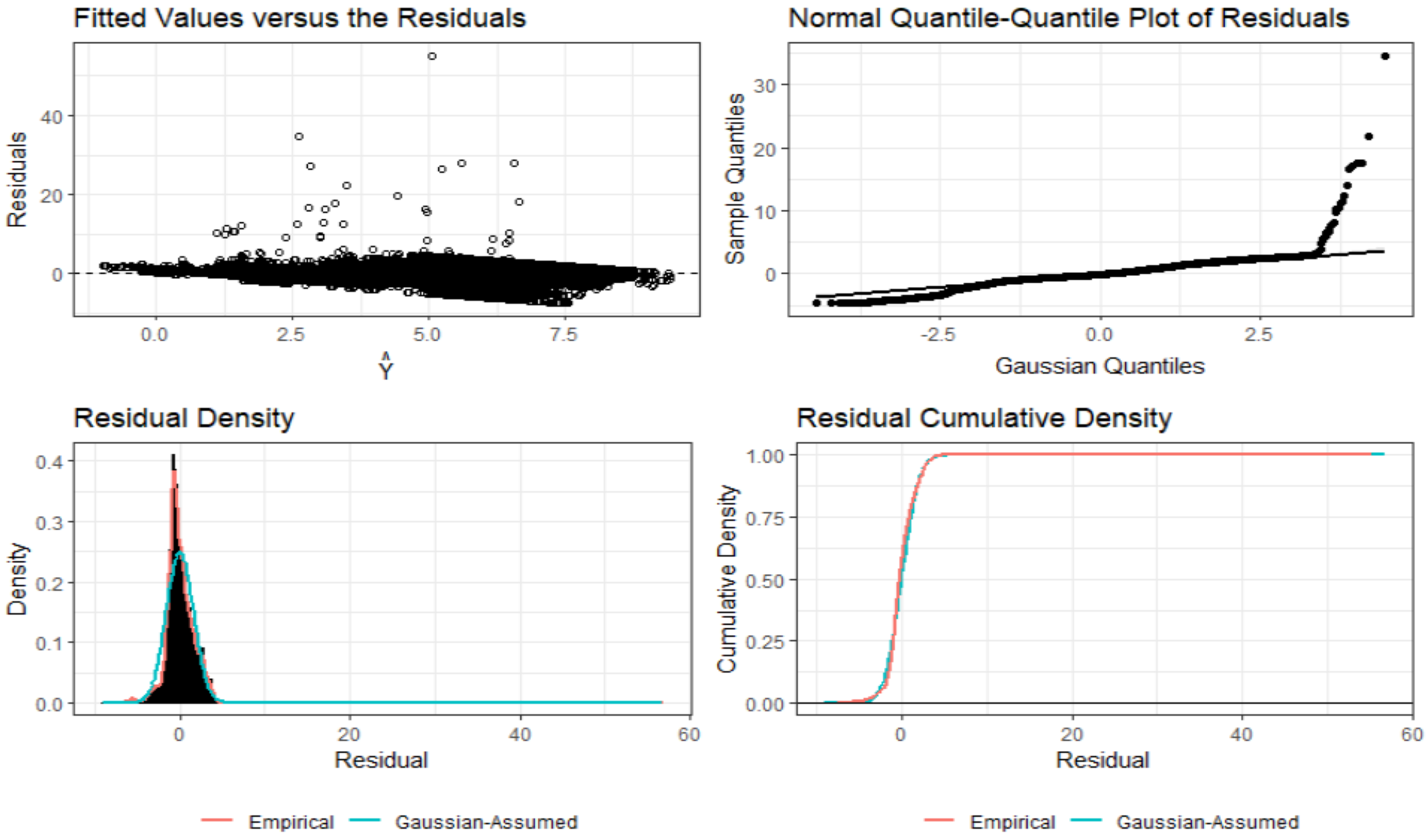
- Tenure
- Avg_age
- miles

Categorical:

- Fod
- EZpay
- Carrier_flag
- Isgift
- Period
- Income

- Adjusted R-Squared: .4255
- Many significant p-values
- All significant except for fod_TWTF & period7
- Most significant coefficients:
  1. fod_MTWTFS
  2. tenure
  3. Income 55k-68k

MODEL 1 DIAGNOSTIC PLOTS



Does not meet all the assumptions for linear regression:

- The assumption of homoscedasticity of the residuals is not satisfied

MODEL 2: LINEAR REGRESSION FOR DELIVERY COST

Quantitative:

- tenure
- avg_age
- miles

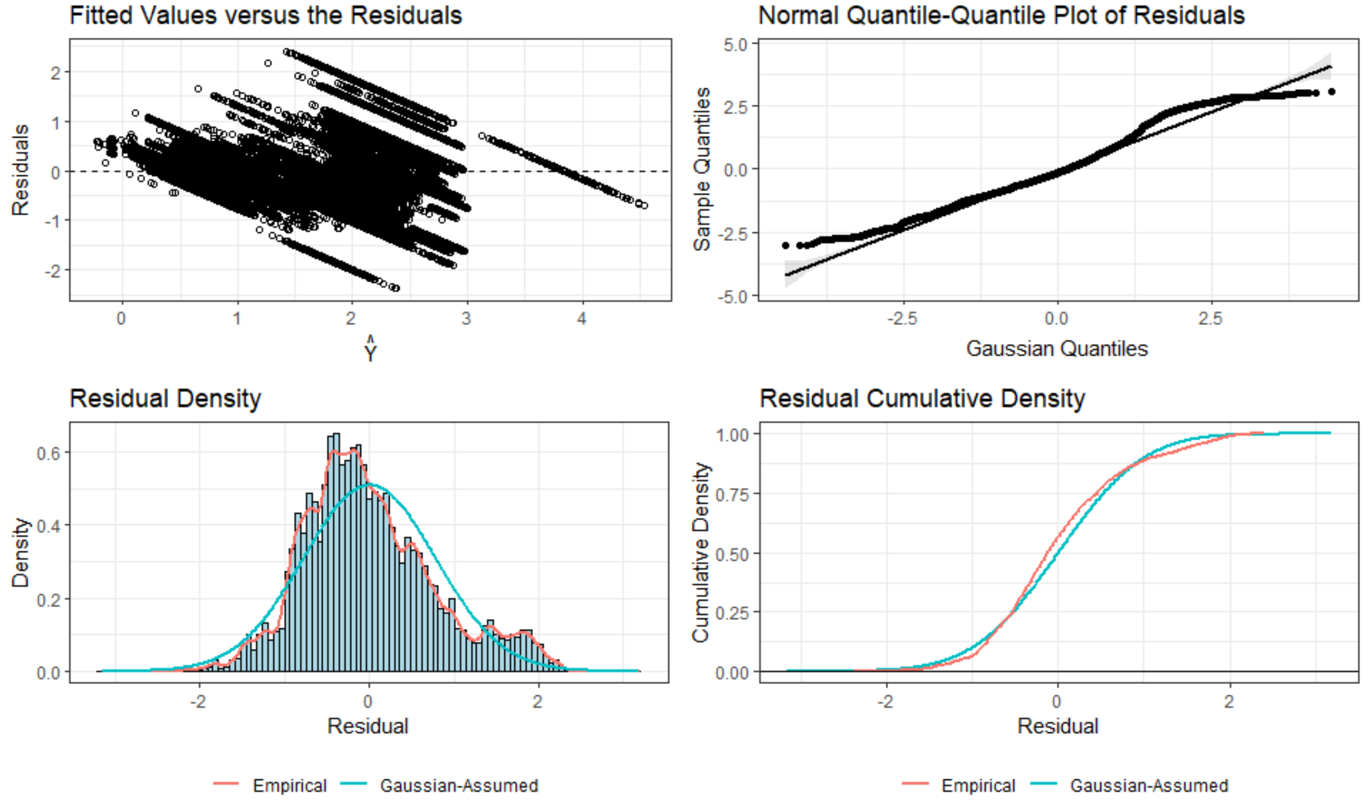
Categorical:

- fod
- EZpay
- carrier_flag
- isgift
- period
- income

Summary:

- Adjusted R-squared: .3192
- Most significant coefficients:
  1. fod_MTWTFS
  2. miles
  3. avg_age

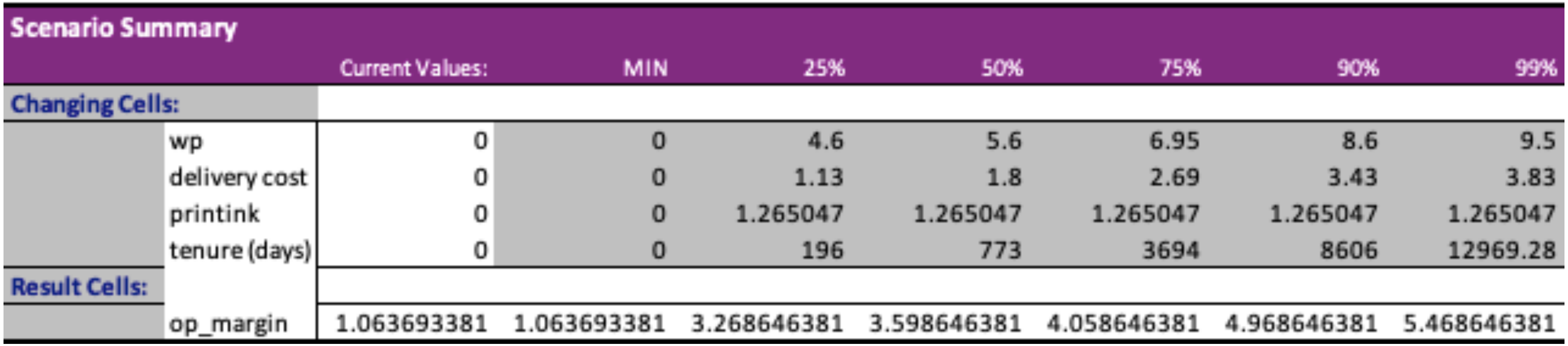
MODEL 2 DIAGNOSTIC PLOTS



Assumptions are better suited for linear regression

SCENARIO ANALYSIS

SCENARIO – QUARTILES (MIN, 25%, 50%, 75%, 90%, 99%)



Scenario Summary

		Current Values:	MIN	25%	50%	75%	90%	99%
Changing Cells:								
	wp	0	0	4.6	5.6	6.95	8.6	9.5
	delivery cost	0	0	1.13	1.8	2.69	3.43	3.83
	printink	0	0	1.265047	1.265047	1.265047	1.265047	1.265047
	tenure (days)	0	0	196	773	3694	8606	12969.28
Result Cells:								
	op_margin	1.063693381	1.063693381	3.268646381	3.598646381	4.058646381	4.968646381	5.468646381

RETENTION MODELS

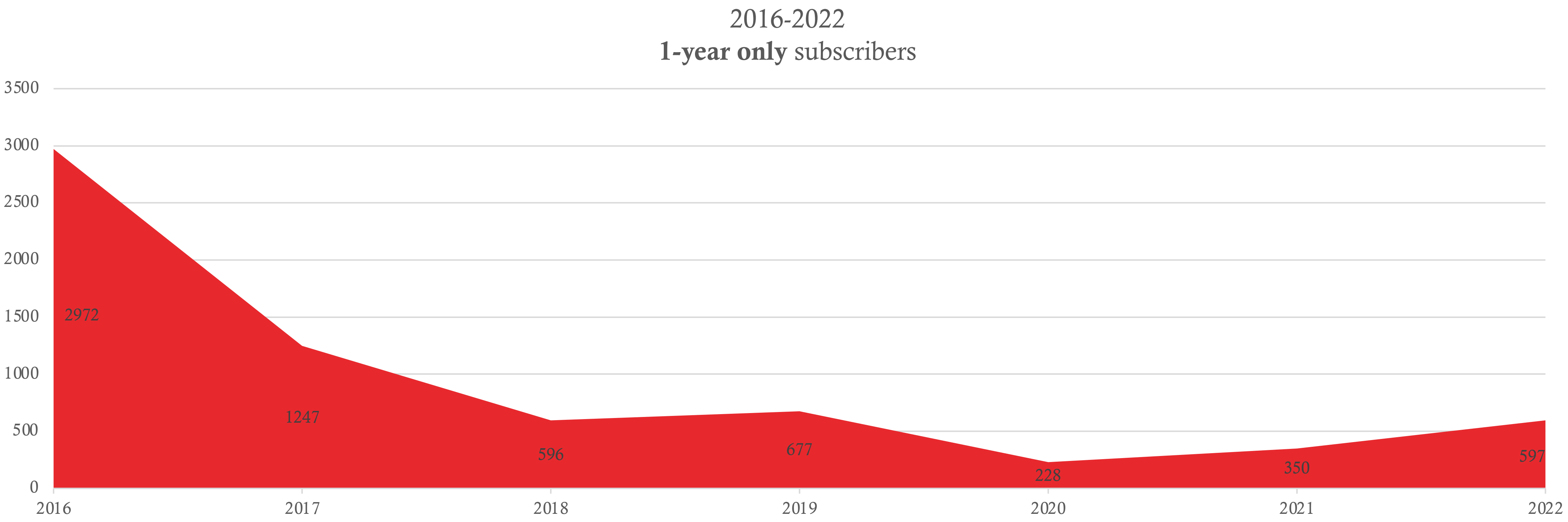
IDEAS BEHIND RETENTION

Want to see which factors go into whether a current customer will stay next year or leave

Created a new variable "ny_status" that encapsulates this, has values for 2016-21

Once model is created, use 2022 as a testing set to predict 2023 customer behavior

1-YEAR RETENTION STATS



LOGISTIC REGRESSION

- Since the decision to stay or leave is binary, started with binary logistic regression
- Model provides a **probability** of whether a customer will stay or go
  - **If higher** than sample proportion of those that stayed, **predict success**, or stay
  - **Else, predict failure**, or leave
- Only requires independence of observations and the absence of multicollinearity

DISTRIBUTION OF NEXT-YEAR STATUS

Next-Year Status	**Percentage of Total 2016-21**	**Percentage of Total 2022**
Leave	**20.33%**	**23.65%**
Stay	**79.67**	**76.35%**

WEEKLY PRICE

Next-Year Status	Percentage of Total 2016-21	Percentage of Total 2022
Leave	**20.33%**	**23.65%**
Weekly Price Quartile 1	33.64%	79.54%
Weekly Price Quartile 2	25.40%	16.32%
Weekly Price Quartile 3	22.05%	3.96%
Weekly Price Quartile 4	18.91%	0.18%
Stay	**79.67%**	**76.35%**
Weekly Price Quartile 1	15.19%	6.25%
Weekly Price Quartile 2	31.01%	27.04%
Weekly Price Quartile 3	27.16%	33.99%
Weekly Price Quartile 4	26.64%	32.72%

FREQUENCY OF DELIVERY

Next-Year Status	Percentage of Total 2016-21	Percentage of Total 2022
Leave	**20.33%**	**23.65%**
Saturday	17.93%	43.30%
Friday, Saturday	0.11%	0.18%
Thursday, Friday, Saturday	2.72%	10.37%
5-Day Plan*	0.73%	10.19%
Mon, Tues, Wed, Thurs, Fri, Sat	78.51%	35.96%
Stay	**79.67%**	**76.35%**
Saturday	7.06%	0%
Friday, Saturday	0.05%	0%
Thursday, Friday, Saturday	1.38%	0%
Tues, Wed, Thurs, Fri	0%	0.02%
5-Day Plan*	0.46%	0.35%
Mon, Tues, Wed, Thurs, Fri, Sat	91.05%	99.63%

AVERAGE AGE

Next-Year Status	Percentage of Total 2016-21	Percentage of Total 2022
Leave	**20.33%**	**23.65%**
Average Age Quartile 1	26.30%	29.08%
Average Age Quartile 2	25.29%	27.62%
Average Age Quartile 3	24.14%	25.09%
Average Age Quartile 4	24.27%	18.21%
Stay	**79.67%**	**76.35%**
Average Age Quartile 1	24.27%	23.65%
Average Age Quartile 2	24.91%	21.22%
Average Age Quartile 3	25.62%	26.60%
Average Age Quartile 4	25.19%	28.53%

MILES FROM BASE

Next-Year Status	Percentage of Total 2016-21	Percentage of Total 2022
Leave	**20.33%**	**23.65%**
Miles from Base Quartile 1	22.55%	22.45%
Miles from Base Quartile 2	23.93%	25.27%
Miles from Base Quartile 3	25.29%	19.28%
Miles from Base Quartile 4	28.23%	33.00%
Stay	**79.67%**	**76.35%**
Miles from Base Quartile 1	24.23%	25.13%
Miles from Base Quartile 2	26.19%	23.54%
Miles from Base Quartile 3	25.30%	28.79%
Miles from Base Quartile 4	24.27%	22.54%

LOGISTIC REGRESSION MODEL

- Used four variables to model retention: weekly price, average age, miles from base, and FOD
- If probability of staying exceeds probability that someone in the training set stayed (**79.65%**), we predict they stay; otherwise, we predict they leave
- Testing on split dataset gave **~68% accuracy** on average
- Mapped same model trained on entire 2016-21 dataset to predict 2022 model

ANALYSIS OF PREDICTED 2022 RETENTION: STAY

- **First 7854** observations of predicted 2022 data sorted by highest probability of staying all had a 6-day subscription plan (_MTWTFS)
- **14 of first 15** observations were in the fourth quartile of average age (that is, older than average)
- **First 70** observations and **452 of first 453** were in the fourth quartile of weekly price
- **First 11** observations were in either the second or third quartiles of miles from the newspaper's base (intermediate distances from Bangor)

ANALYSIS OF PREDICTED 2022 RETENTION: LEAVE

- **First 1337** instances of 2022 data sorted by lowest probability of staying have a weekly price in the first quartile (**highest 11.26%** of probabilities)
  - **Highest 30** probabilities have the Saturday-only delivery plan
- Each of **first 9** observations and **18 of the first 20** observations are in the fourth quartile of distance from the newspaper's base
- Each of the **first 14** observations and **26 of the first 28** are in either the first or second quartile of average age (that is, younger than average)

DISCUSSION OF LOGISTIC REGRESSION

- **Pros**
  - **Simple** and interpretable
  - Values map well to 2022 data points
- Cons
  - ~68% accuracy on average does not inspire confidence
  - Map of actual percentages to model percentages **encapsulates trends**, not correct values
- New model…

RANDOM FOREST FOR RETENTION

- Random forest – constructing multiple decision trees from random subsets of explanatory variables, and predicting the most common result of each
- Ensemble ML algorithm which introduces complexity and large sample size into previous model
- Use **SMOTE** to over-sample "leave"
  - With an under-sampling method tied on, results in higher accuracy on average on binary decision
- **85% accuracy** on average
- **6** variables: FOD, weekly price, delivery cost, tenure, average age, and miles from base

WEEKLY PRICE

Next-Year Status	Percentage of Total 2016-21	Percentage of Total 2022
Leave	**20.33%**	**23.65%**
Weekly Price Quartile 1	33.64%	39.01%
Weekly Price Quartile 2	25.40%	26.24%
Weekly Price Quartile 3	22.05%	18.16%
Weekly Price Quartile 4	18.91%	16.59%
Stay	**79.67%**	**76.35%**
Weekly Price Quartile 1	15.19%	21.18%
Weekly Price Quartile 2	31.01%	24.24%
Weekly Price Quartile 3	27.16%	28.25%
Weekly Price Quartile 4	26.64%	26.34%

FREQUENCY OF DELIVERY

Next-Year Status	Percentage of Total 2016-21	Percentage of Total 2022
Leave	**20.33%**	**13.46%**
Saturday	17.93%	15.40%
Friday, Saturday	0.11%	0.25%
Thursday, Friday, Saturday	2.72%	7.83%
Tues, Wed, Thurs, Fri	0%	0.13%
5-Day Plan*	0.73%	19.54%
Mon, Tues, Wed, Thurs, Fri, Sat	78.51%	56.86%
Stay	**79.67%**	**86.54%**
Saturday	7.06%	9.44%
Friday, Saturday	0.05%	0.01%
Thursday, Friday, Saturday	1.38%	1.62%
5-Day Plan*	0.46%	0.06%
Mon, Tues, Wed, Thurs, Fri, Sat	91.05%	88.88%

TENURE

Next-Year Status	Percentage of Total 2016-21	Percentage of Total 2022
Leave	**20.33%**	**13.46%**
Tenure Quartile 1	36.99%	39.01%
Tenure Quartile 2	27.72%	28.43%
Tenure Quartile 3	20.30%	18.16%
Tenure Quartile 4	14.99%	14.40%
Stay	**79.67%**	**86.54%**
Tenure Quartile 1	21.87%	22.59%
Tenure Quartile 2	24.32%	24.64%
Tenure Quartile 3	26.24%	26.11%
Tenure Quartile 4	27.56%	26.66%

MILES FROM BASE

Next-Year Status	Percentage of Total 2016-21	Percentage of Total 2022
Leave	**20.33%**	**13.46%**
Miles from Base Quartile 1	22.55%	22.48%
Miles from Base Quartile 2	23.93%	16.41%
Miles from Base Quartile 3	25.29%	27.74%
Miles from Base Quartile 4	28.23%	33.38%
Stay	**79.67%**	**86.54%**
Miles from Base Quartile 1	24.23%	24.81%
Miles from Base Quartile 2	26.19%	25.12%
Miles from Base Quartile 3	25.30%	26.36%
Miles from Base Quartile 4	24.27%	23.71%

RECOMMENDATIONS

- Prioritize selling 6-day plans
- Promote higher weekly price sales – majority of people stay after the 2nd Quartile of weekly price
- Use the logistic regression model to predict if customer stays/leaves, and target those who stay with better subscription packages

FUTURE WORK

Continue to correct linear models of operating margin and weekly price to fit assumptions

Experiment with other machine-learning algorithms to predict retention

If the company is in a position to expand to other states, collecting out-of-state data suitable for analysis would help the company diversify

ACKNOWLEDGEMENTS

We would like to thank the Mather Economics team for their support and help all semester.

Thank you to Dr. Grannan for all your help, and we wish you luck in Charlotte!

Thanks to Furman University for affording us this meaningful opportunity.