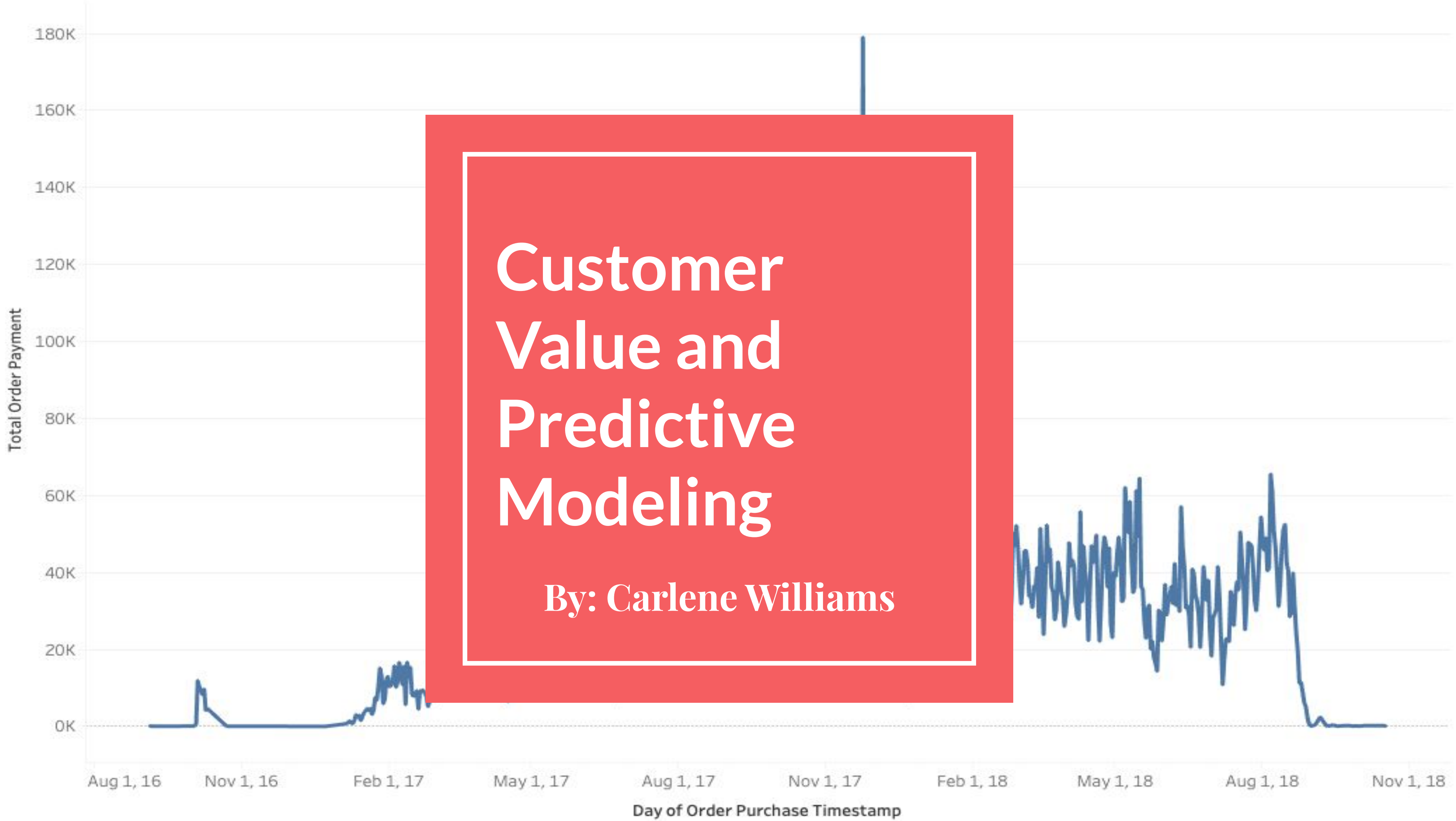


Customer Value and Predictive Modeling

By: Carlene Williams

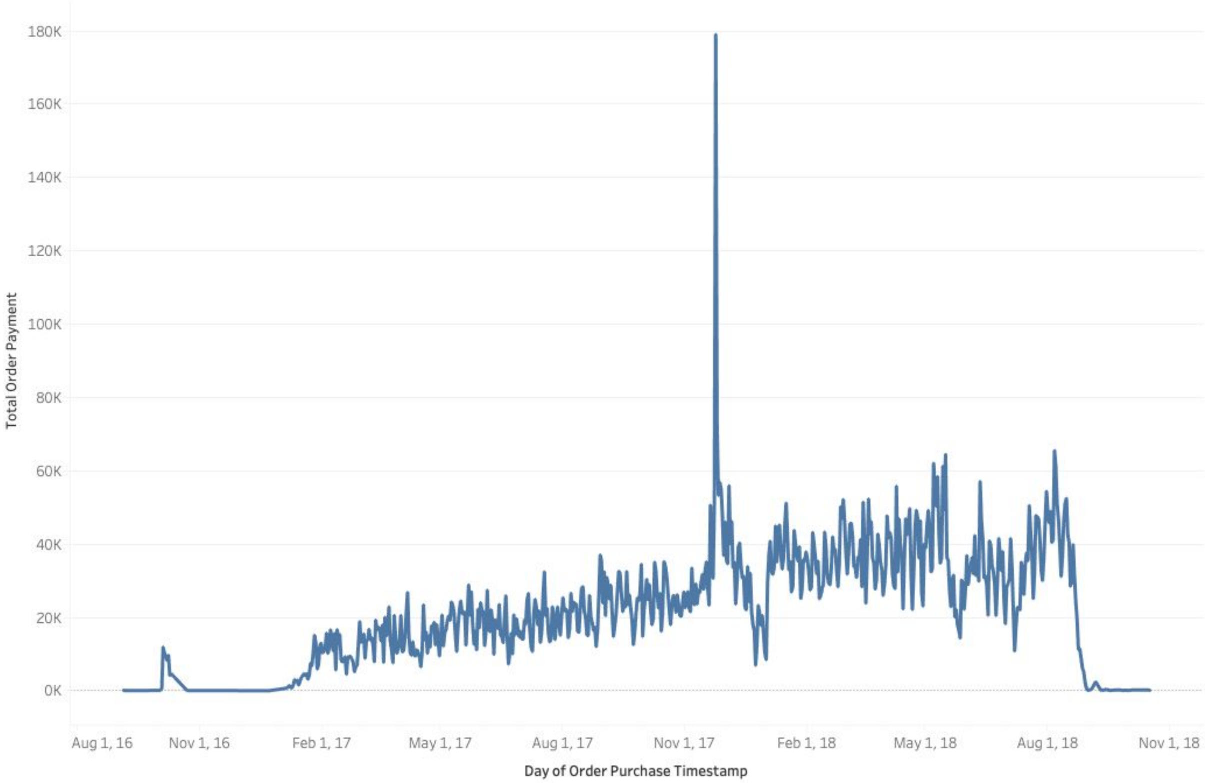
Goal: Predict the Value of a New Customer

- Data about Olist Store from Kaggle.com
  - Brazilian Ecommerce Site

  - Data on around 100k orders, their order dates with customer, product, other information

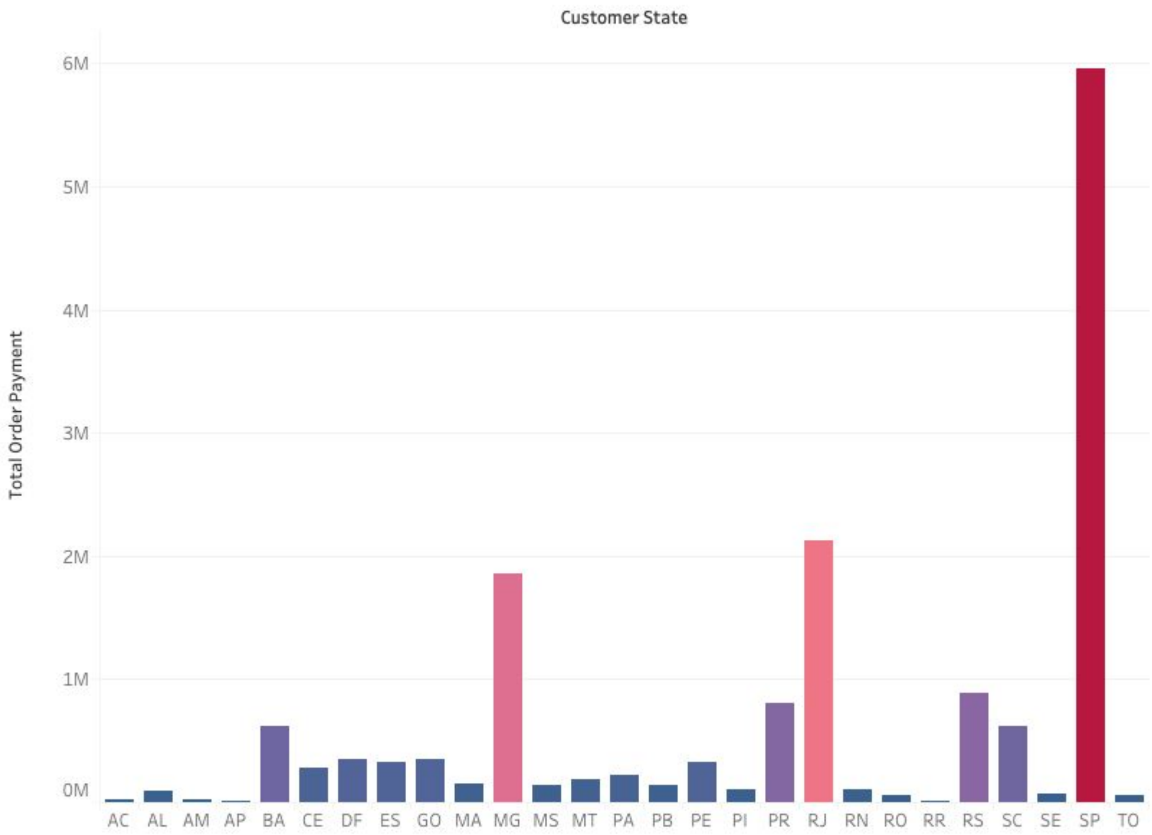
Steps Towards Predicting Customer Value

- Exploratory analysis
- Estimate CV of each historical customer
- Create a linear regression model to test variables (Customer State and Product Type)



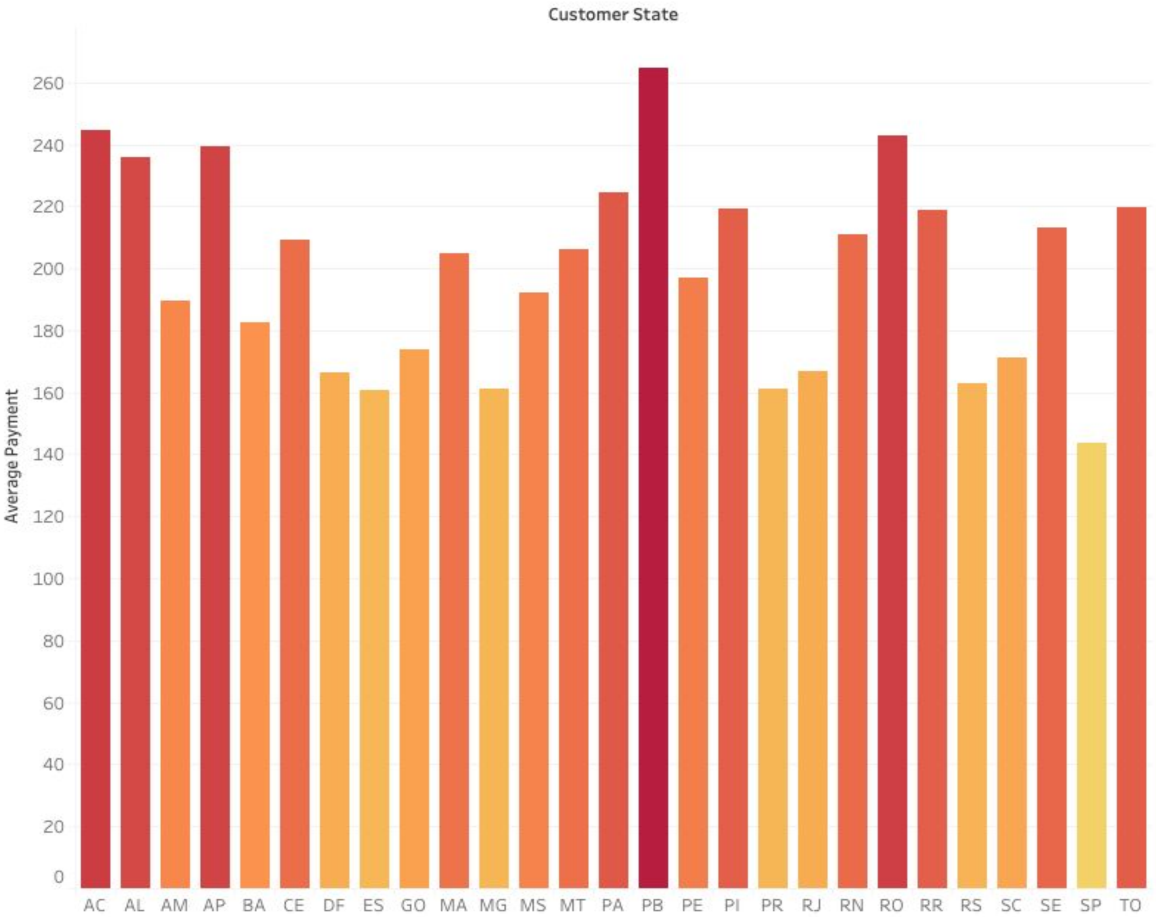
Customers

- 27 different states in total
- Sao Paulo (SP) spends the most money in total



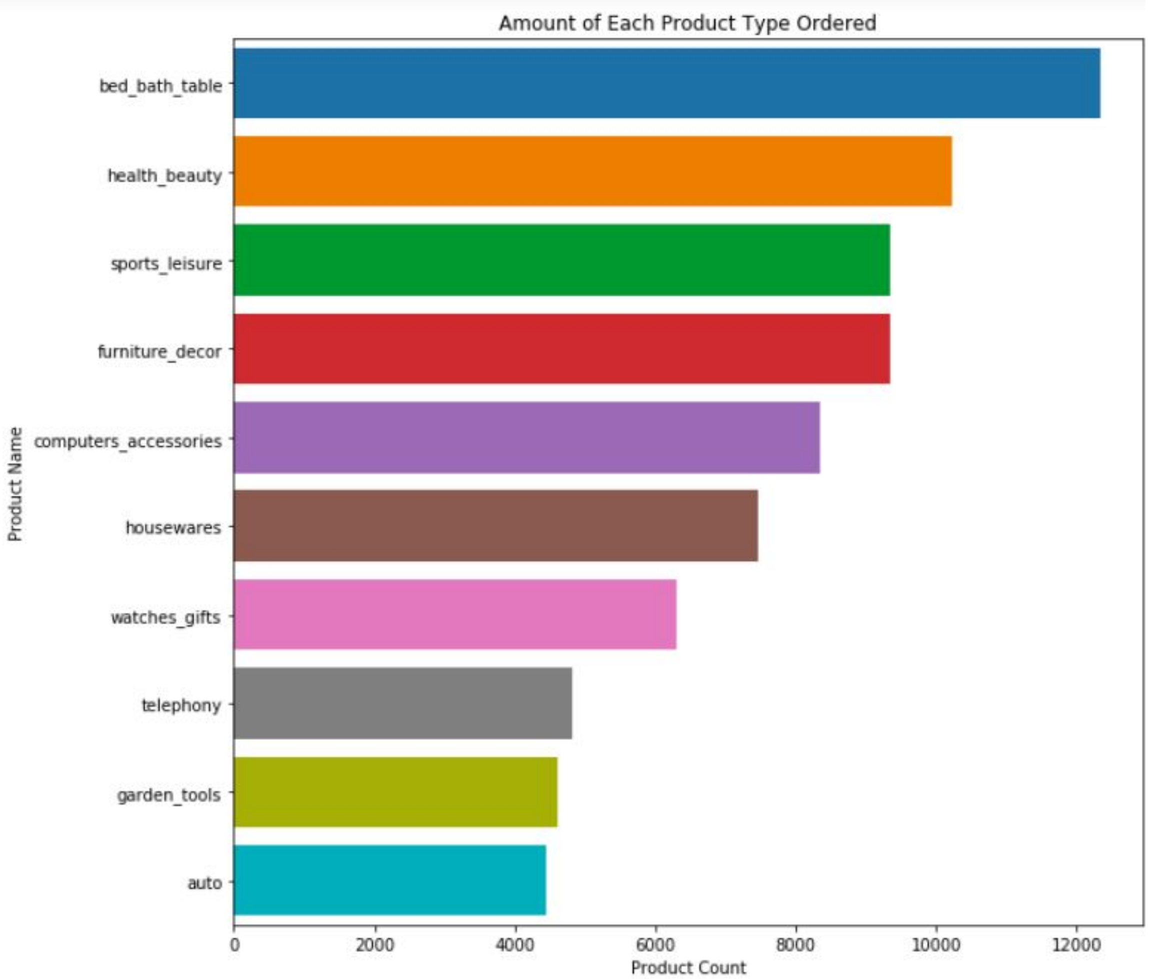
Customers

- 27 different states in total
- Sao Paulo (SP) spends the most money in total, has the most orders
- Paraíba (PB) spends the most per order



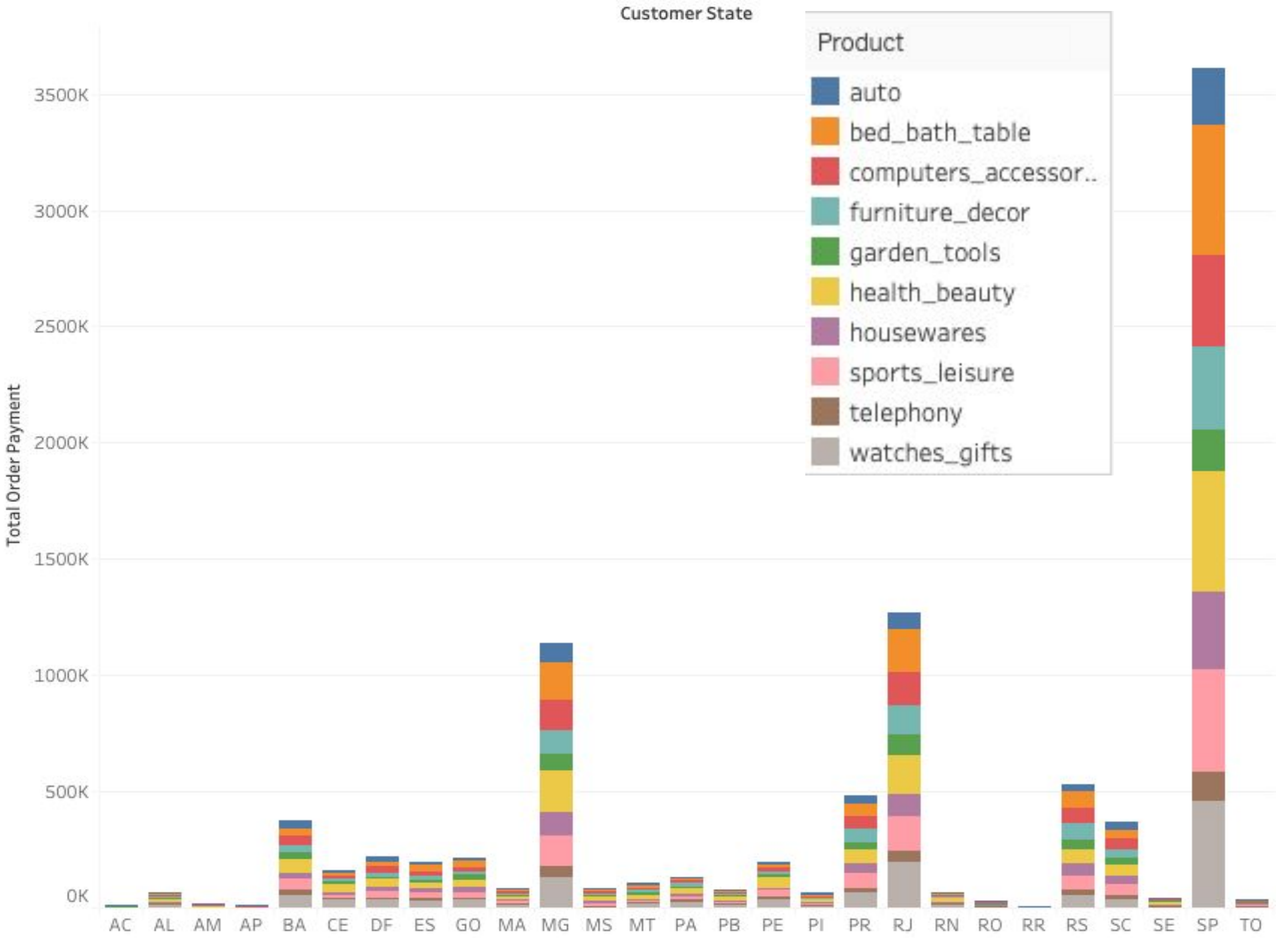
Products

- 71 different product types in total
- Bed/Bath/Table is ordered the most



Products

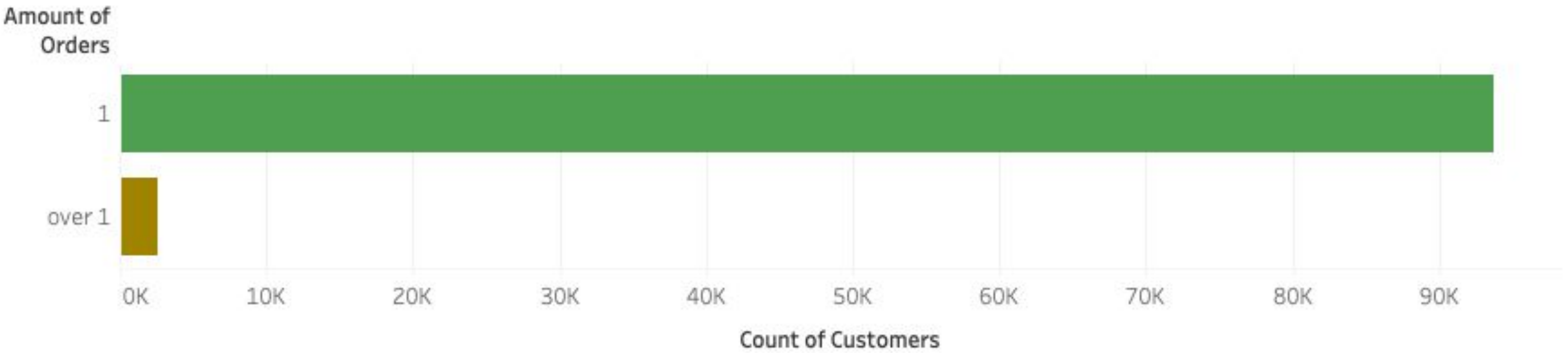
- 71 different product types in total
- Health/Beauty, Watches/Gifts, Bed/Bath/Table have the most money spent



Customer Value of Current Customers

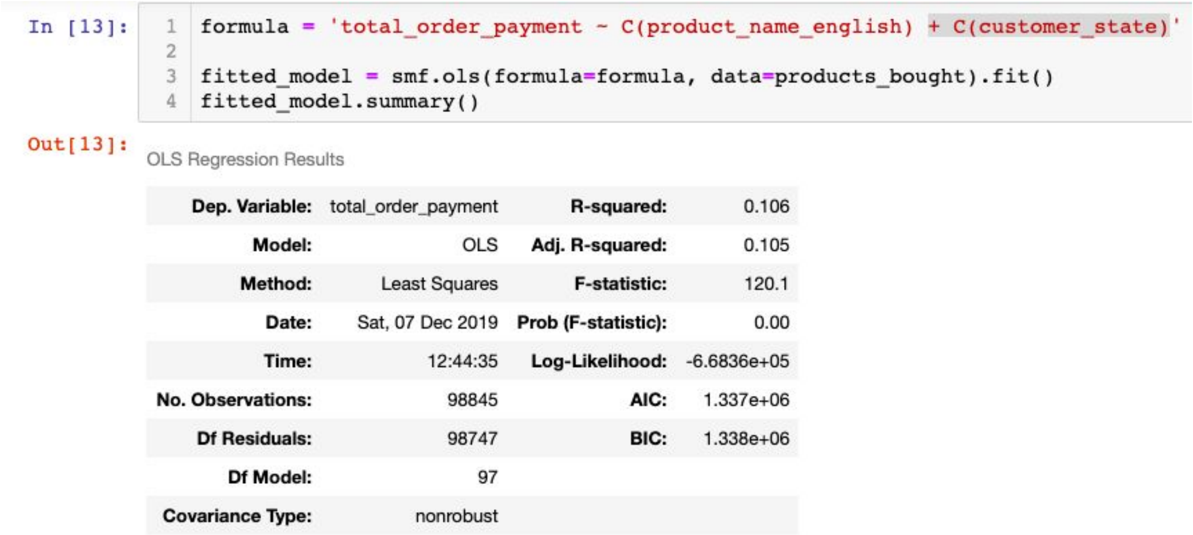
- First six months of purchases per customer
- Sum up payments over that time period

96,095 customers



The Model

- Linear Regression Model
- Predictors: Customer State and Product Name



In [13]:

```
formula = 'total_order_payment ~ C(product_name_english) + C(customer_state)'

fitted_model = smf.ols(formula=formula, data=products_bought).fit()
fitted_model.summary()
```

Out[13]:

OLS Regression Results

Dep. Variable:	total_order_payment	**R-squared:**	0.106
Model:	OLS	**Adj. R-squared:**	0.105
Method:	Least Squares	**F-statistic:**	120.1
Date:	Sat, 07 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:44:35	**Log-Likelihood:**	-6.6836e+05
No. Observations:	98845	**AIC:**	1.337e+06
Df Residuals:	98747	**BIC:**	1.338e+06
Df Model:	97		
Covariance Type:	nonrobust		

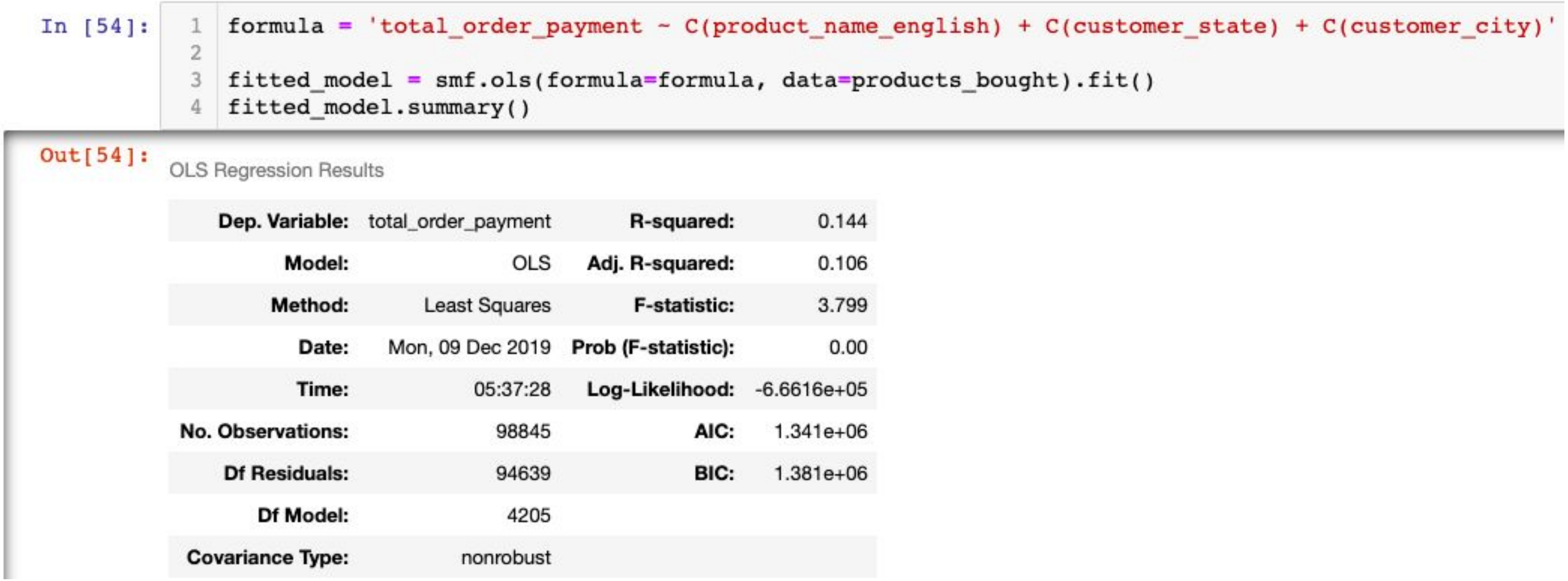
Conclusions

- Only 10% of the behavior of Customer Value in relation to Customer State and the Product Type is explained by this model

- These two variables alone aren’t useful enough to predict Customer Value

Next Step

- Accuracy:
  - Test other ways of calculating CV for this dataset full of customers that only order once



In [54]:

```
formula = 'total_order_payment ~ C(product_name_english) + C(customer_state) + C(customer_city)'

fitted_model = smf.ols(formula=formula, data=products_bought).fit()
fitted_model.summary()
```

Out[54]:

OLS Regression Results

Dep. Variable:	total_order_payment	**R-squared:**	0.144
Model:	OLS	**Adj. R-squared:**	0.106
Method:	Least Squares	**F-statistic:**	3.799
Date:	Mon, 09 Dec 2019	**Prob (F-statistic):**	0.00
Time:	05:37:28	**Log-Likelihood:**	-6.6616e+05
No. Observations:	98845	**AIC:**	1.341e+06
Df Residuals:	94639	**BIC:**	1.381e+06
Df Model:	4205		
Covariance Type:	nonrobust		