AUBURN
UNIVERSITY

RAYMOND J. HARBERT
COLLEGE OF BUSINESS

Business Analytics

Kia Capstone Presentation

Myungsub Cho, Juyan Deng, Carson Goff, Molly Harrop

AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics

Overview

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

Background Information

MANHEIM

- Manheim was established in 1945
- Operates as a “middleman” for auto sales, through traditional methods as well as digitally
- Manheim is an all-encompassing entity, covering solutions from wholesale to retail
- Provides a reliable, safe, and secure market
- A large company, with approx. 18,000 employees and annual revenues of over $2.6 billion

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

Background Information (Cont.)

MANHEIM

- Manheim is also a global company, with operations reaching 11 countries
- Some components of Manheim include dealerships, manufacturers, banks, car rental agencies, and government agencies
- Company promises to help get sellers the highest value for their cars, and to help buyers get the highest quality vehicle possible

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

The Objective

- To develop a pricing model that considers vehicle attributes, seasonality, and regional differences
- Our goal is to understand how to reduce re-driving by looking at the likelihood of car sales at different price levels
- To analyze the data and see how the seller contributes to the resale value of car

AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics

Part 1: General Pricing Models

With this we can compare the developed model with Manheim's existing proprietary model in the real world to determine the accuracy of the model.

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

Step 1: Data cleaning

- Used Microsoft excel functions to reformat the sold date into a version python and excel could interpret.
- Divided sold dates into seasons to be used for the regression model.
- Grouped auction house locations together into regions to be used for the regression model.
- Deleted entries with missing data.

L	M	N	O	P	Q	R
times_run	sold_date	seller	sale_price	sold_date	season	region
1	2018-09-2	5361856	8500	9/27/2018	Fall	Canada
1	2017-05-0	5357510	14000	5/2/2017	Spring	Canada
2	2018-07-1	4920595	22500	7/19/2018	Summer	Canada
2	2018-03-2	4998915	28000	3/20/2018	Spring	Canada
1	2018-05-1	4903345	14000	5/10/2018	Spring	Canada
1	2017-01-3	4903345	13800	1/31/2017	Winter	Canada
1	2017-08-2	4903345	22500	8/24/2017	Summer	Canada
1	2017-12-1	4903345	17000	12/19/2017	Winter	Canada
1	2018-09-2	4909177	23400	9/26/2018	Fall	Canada

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

Step 2: Importing and formatting dataset

- Imported dataset into python notebook.
- Manipulate variables with mixed data types to be suitable for use in a regression model.
- Condition codes were converted to their corresponding numeric values and multiplied by 10 to be compatible with the other condition grades.
- Transmission types combined into either automatic or manual.
- There were only three entries in the entire dataset for manual transmissions so I decided to drop this variable.

```
# Find all the unique condition grades.
print(cars["condition_grade"].unique())

# Replace all the occurences of codes with the corresponding numbers.
cars.loc[cars["condition_grade"] == "AV", "condition_grade"] = '30'
cars.loc[cars["condition_grade"] == "SL", "condition_grade"] = '0'
cars.loc[cars["condition_grade"] == "CL", "condition_grade"] = '40'
cars.loc[cars["condition_grade"] == "EC", "condition_grade"] = '50'
cars.loc[cars["condition_grade"] == "PR", "condition_grade"] = '10'
cars.loc[cars["condition_grade"] == "RG", "condition_grade"] = '20'
```

```
['46' '50' '45' '40' '41' '38' '48' '43' '47' '49' '26' '27' '15' '44'
 '36' '42' '37' '35' '24' '31' '34' '33' 'AV' 'CL' '28' '39' '29' '14'
 '18' '32' '30' '25' '21' '20' '16' '19' '23' '22' '17' '13' 'RG' 'PR'
 'SL' 'EC' '10' '0' '12' '11' '1']
```

```
print(cars["condition_grade"].unique())
```

```
['46' '50' '45' '40' '41' '38' '48' '43' '47' '49' '26' '27' '15' '44'
 '36' '42' '37' '35' '24' '31' '34' '33' '30' '28' '39' '29' '14' '18'
 '32' '25' '21' '20' '16' '19' '23' '22' '17' '13' '10' '0' '12' '11' '1']
```

```
print(cars["transmission"].unique())
cars.loc[cars['transmission'] == 'M', 'transmission'] = 1
cars.loc[cars['transmission'] == 'A', 'transmission'] = 0
cars.loc[cars['transmission'] == '5', 'transmission'] = 0
cars.loc[cars['transmission'] == '6', 'transmission'] = 0
cars.loc[cars['transmission'] == 'O', 'transmission'] = 0
cars.loc[cars['transmission'] == 'P', 'transmission'] = 0
cars.loc[cars['transmission'] == 'Z', 'transmission'] = 0
cars.loc[cars['transmission'] == 'C', 'transmission'] = 0
cars.loc[cars['transmission'] == 'N', 'transmission'] = 0
```

```
['6' 'A' '5' 'O' 'P' 'C' 'M' 'N' 'Z']
```

```
print(cars["transmission"].unique())
```

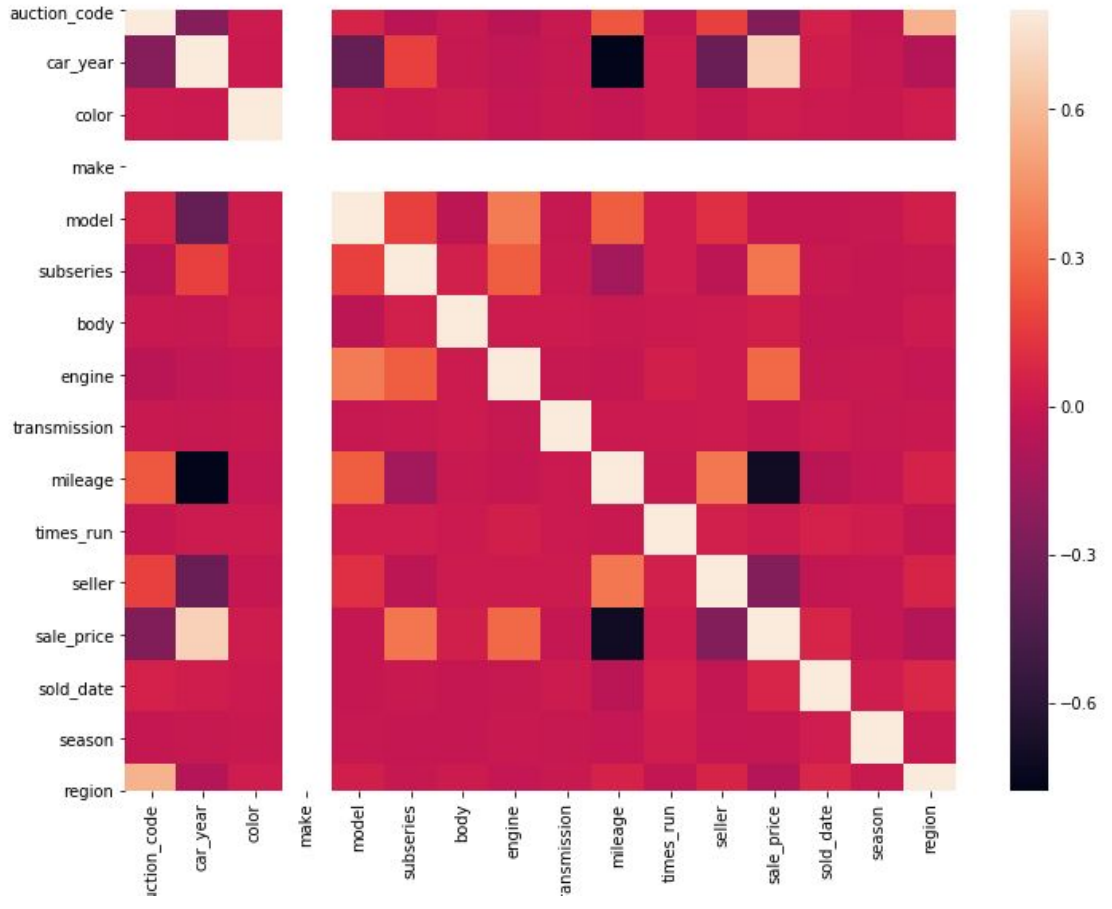
```
[0 1]
```

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

Step 3: Variable Selection

- Used a correlation matrix to determine which of the variables had the strongest influence on the variance in sale price.
- Immediately car year and mileage stand out as having huge sway on sale price
- While immediate it may not appear model is very useful in determining sale price...

```
#correlation matrix
corrmat = cars.corr()
f, ax = plt.subplots(figsize=(12, 9))
sns.heatmap(corrmat, vmax=.8, square=True,);
```

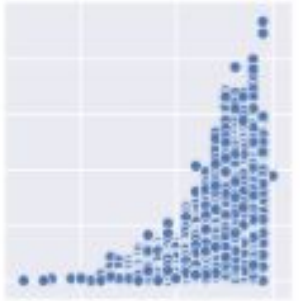


AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

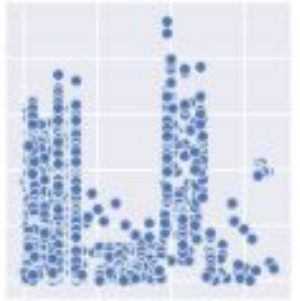
Variable selection cont.

- The models come split into about 50 unique entries including trim level and drivetrain, which my intuition tells me should have significant effect on sale price.
- Split into dummy variables, you can achieve a range for every specific model where the predicted price should fall between
- Again, car year and mileage are both very clearly strongly correlated with sale price.
- Other variables selected include: condition grade, season, and region.

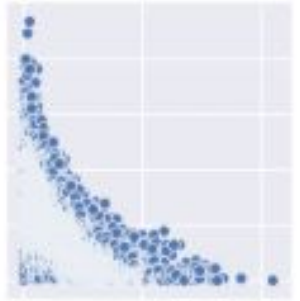
car_year



model



mileage



AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

Step 4: Model Building

- Create X and Y variables from my selection.
- Create dummy variables.
- Split into training and test sets.

```
# Create the x and y variables
X = cars.drop(['auction_code', 'color', 'make', 'subseries', 'body', 'engine', 'transmission',
selectedfeatures = X.columns
print(selectedfeatures)

y = cars['sale_price']
```

```
Index(['car_year', 'model', 'mileage', 'condition_grade', 'season', 'region'], dtype='object')
```

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

Model building: model selection

- I chose to use a random forest regressor model for a multitude of reasons:
  - It provided the highest accuracy score of the different model types I tested at 90%.
  - Offers a predictive function so I can print and compare the values that the model predicts against the actual sale prices.
  - Lowest RMSE of all the model types.

Random forest model

You may also print out the variable importance scores.

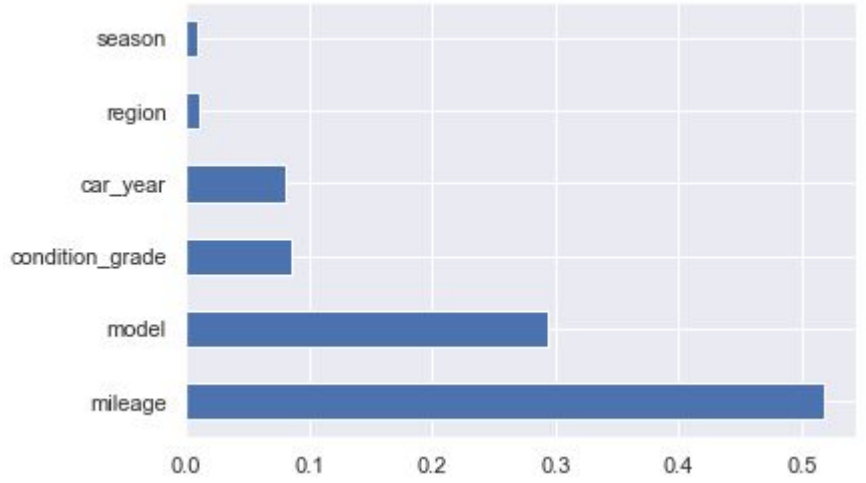
```
from sklearn.ensemble import RandomForestRegressor
forest_reg = RandomForestRegressor(random_state=0)
forest_reg.fit(X_train, y_train)
print("Accuracy Score of Random Forests on train set",forest_reg.score(X_train,y_train))
print("Accuracy Score of Random Forests on test set",forest_reg.score(X_test,y_test))

y_pred_forest = forest_reg.predict(X_test)
forest_mse = mean_squared_error(y_pred_forest, y_test)
forest_rmse = np.sqrt(forest_mse)
print('Random Forest RMSE: %.4f' % forest_rmse)

forest_mae = mean_absolute_error(y_pred_forest, y_test)
print('Random Forest MAE: %.4f' % forest_mae)

# you may also print out variable importance scores.
```

```
Accuracy Score of Random Forests on train set 0.987156663747479
Accuracy Score of Random Forests on test set 0.9104015228603805
Random Forest RMSE: 1609.9417
Random Forest MAE: 1084.3385
```

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

Variable Importance

- After running the random forest regression model, I checked to see which variables had the highest levels of variable importance.
- Season and region both fall way behind the other variables, while mileage is clearly the leading factor in terms of determining the sale price.



Filters

Marks

Automatic

Color | Size | Label

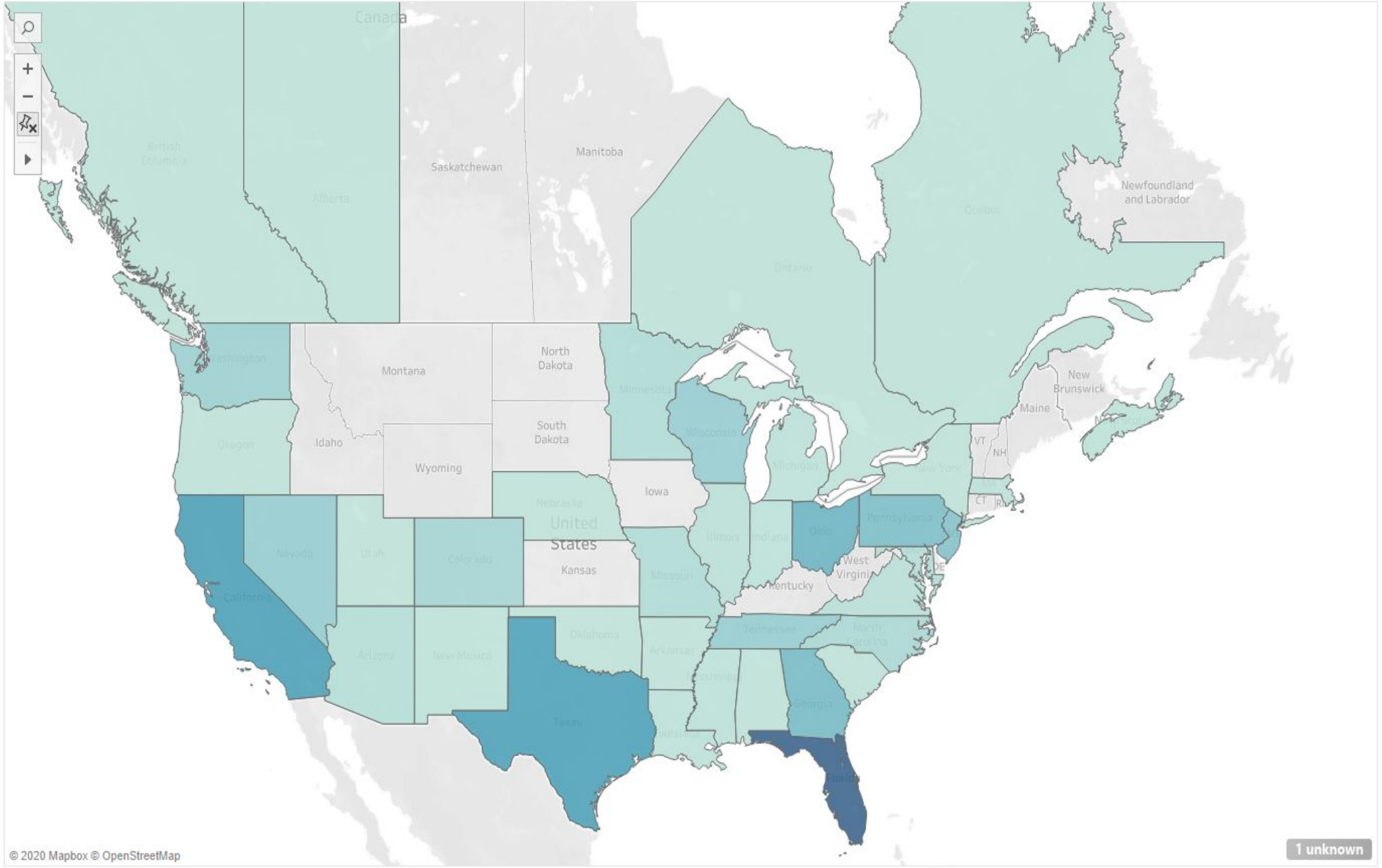
Detail | Tooltip

Country

State

SUM(Sale Pric..

Regional Analyze



SUM(Sale Price)

89,100 165M

Pages

Columns: Longitude (generated)

Rows: Latitude (generated)

This is the virtualization graph between predict sale price model and regional. It shows the same outcome with Manheim's sakes model

Filters

Marks

Automatic

Color | Size | Label

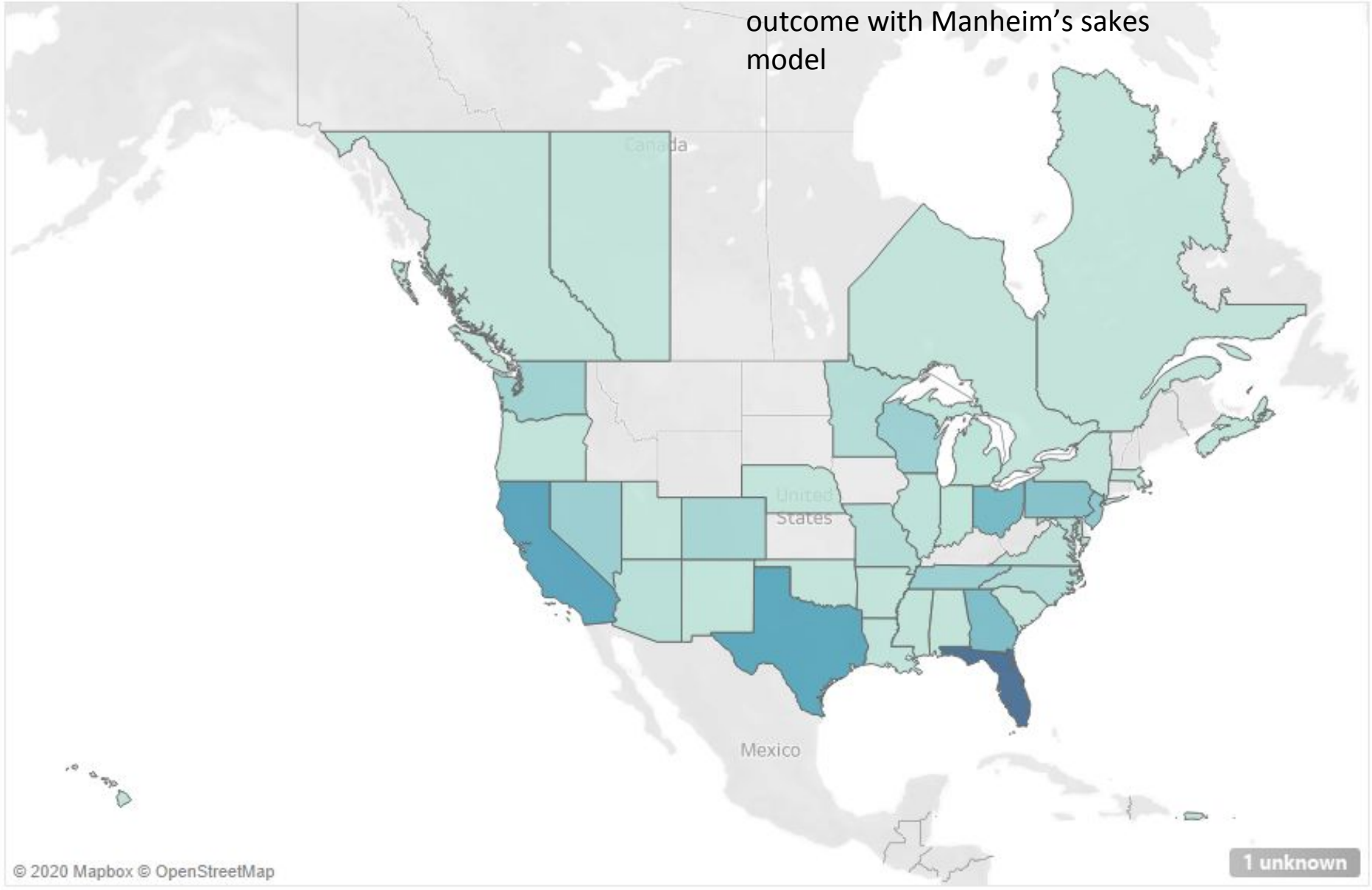
Detail | Tooltip

Country

State

SUM(sale pric..

Sheet 2



SUM(sale price (Kia Ou...

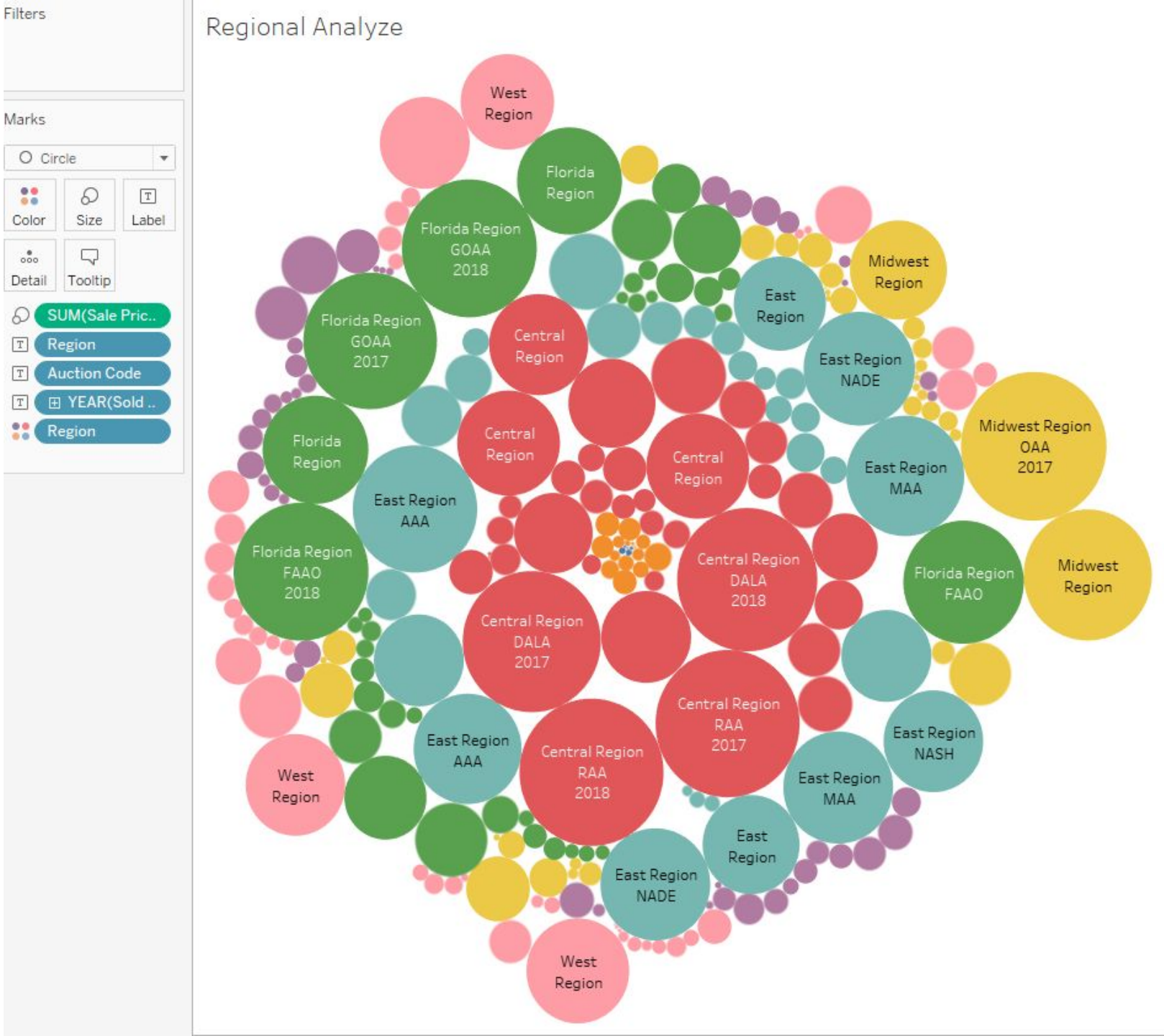
28,350 58,458,421

AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

Although this picture shows that the central region has a higher value than Florida, it contains several states. So we choose Florida, which is more concise and contains a lot of sample data.

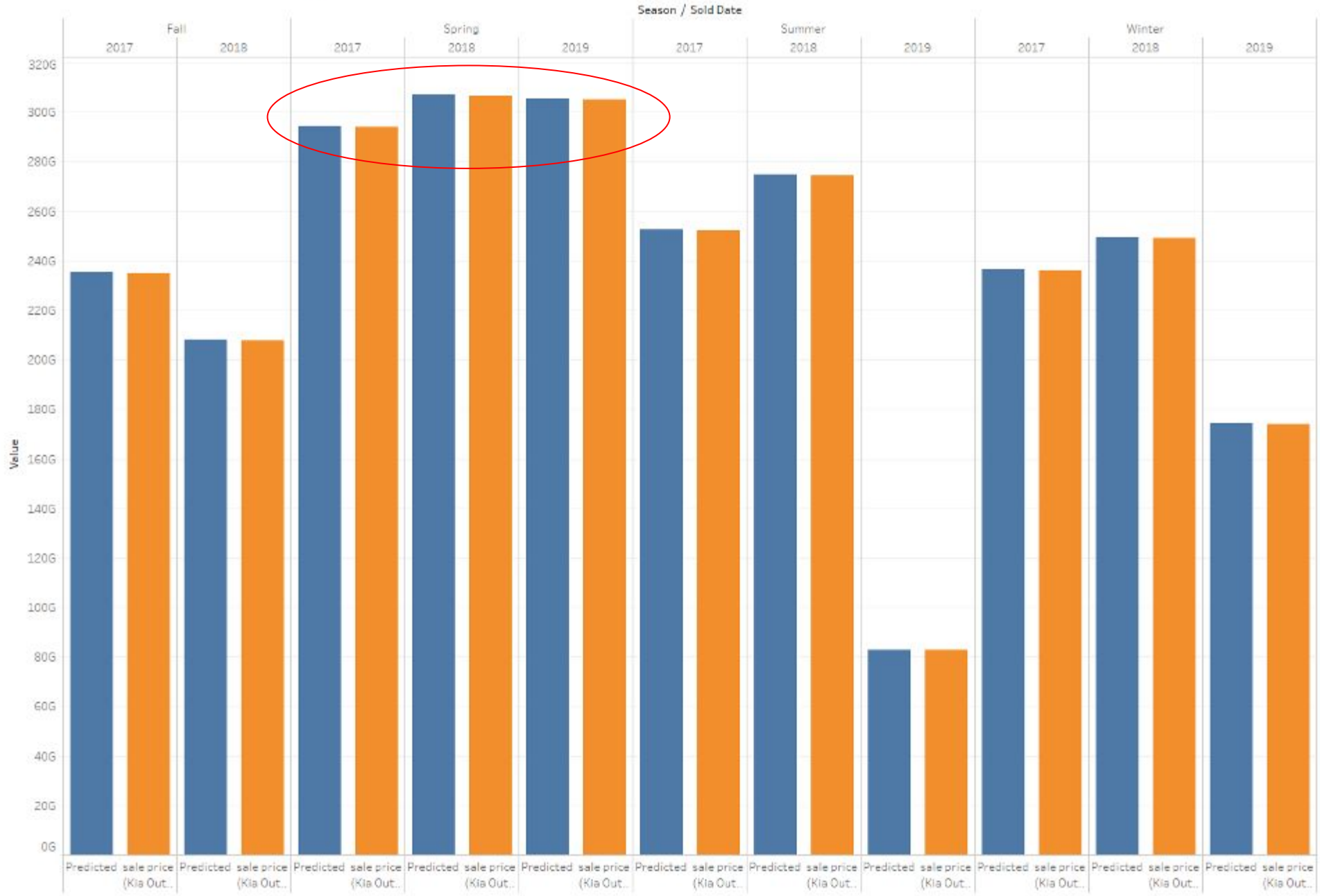


AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

seasonal analysis



AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

Pages

Columns

Rows

Filters

Marks

Circle

Color | Size | Label

Detail | Tooltip

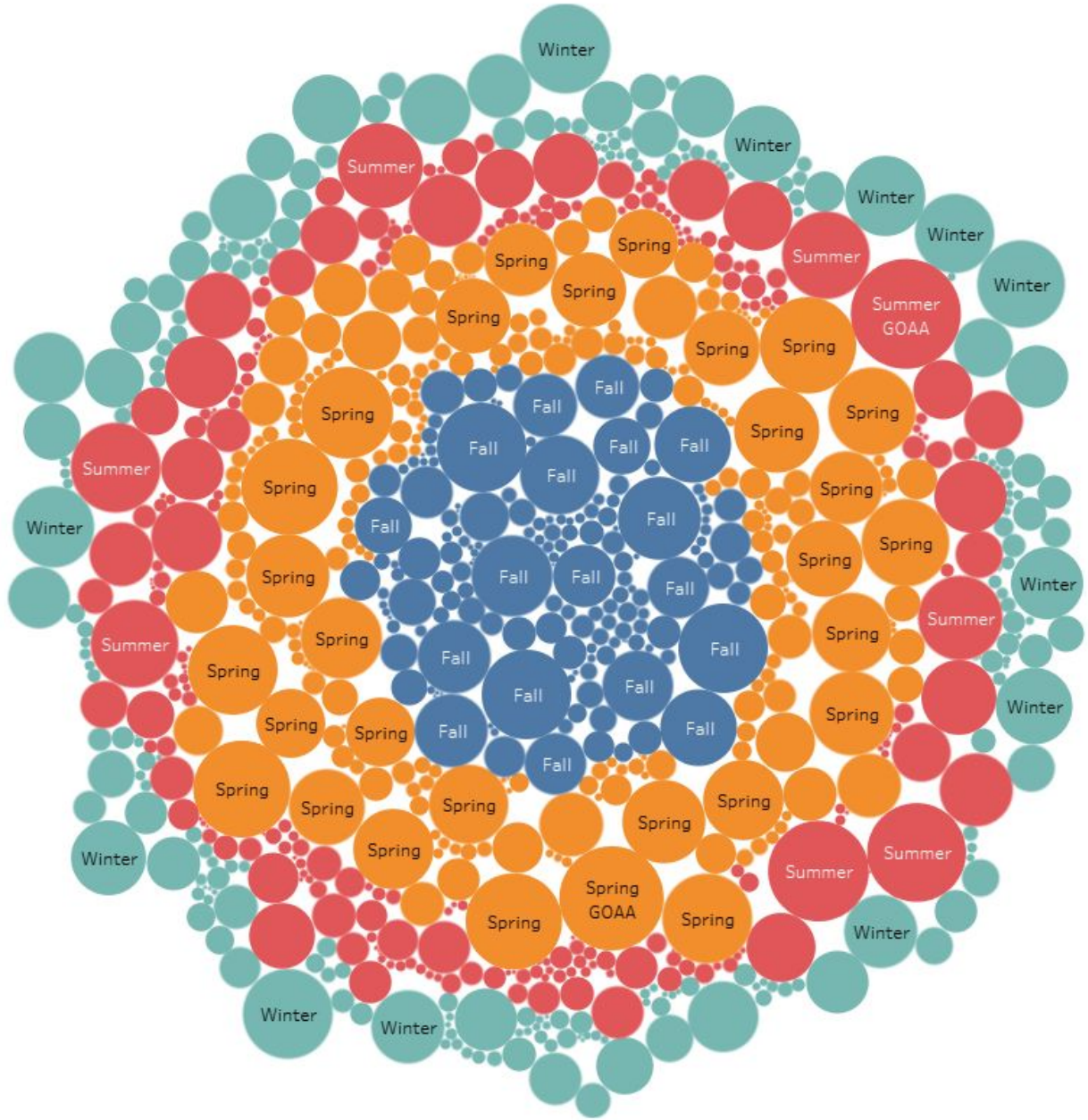
SUM(Sale Pric..

Season

Auction Code

YEAR(Sold ..

Season



AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics

Part 2
Reduce Auction Reruns

AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

Filters

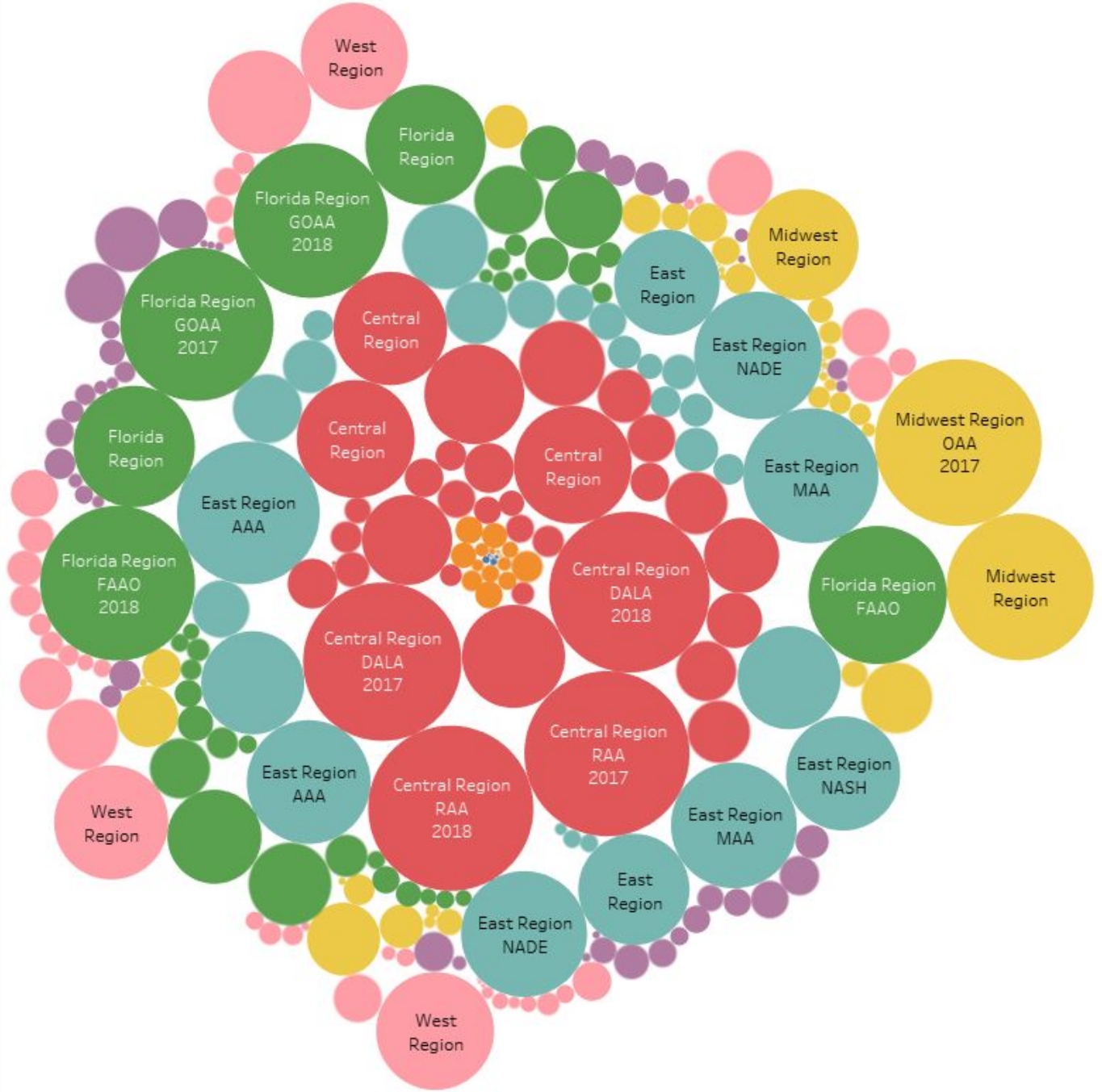
Marks

Circle

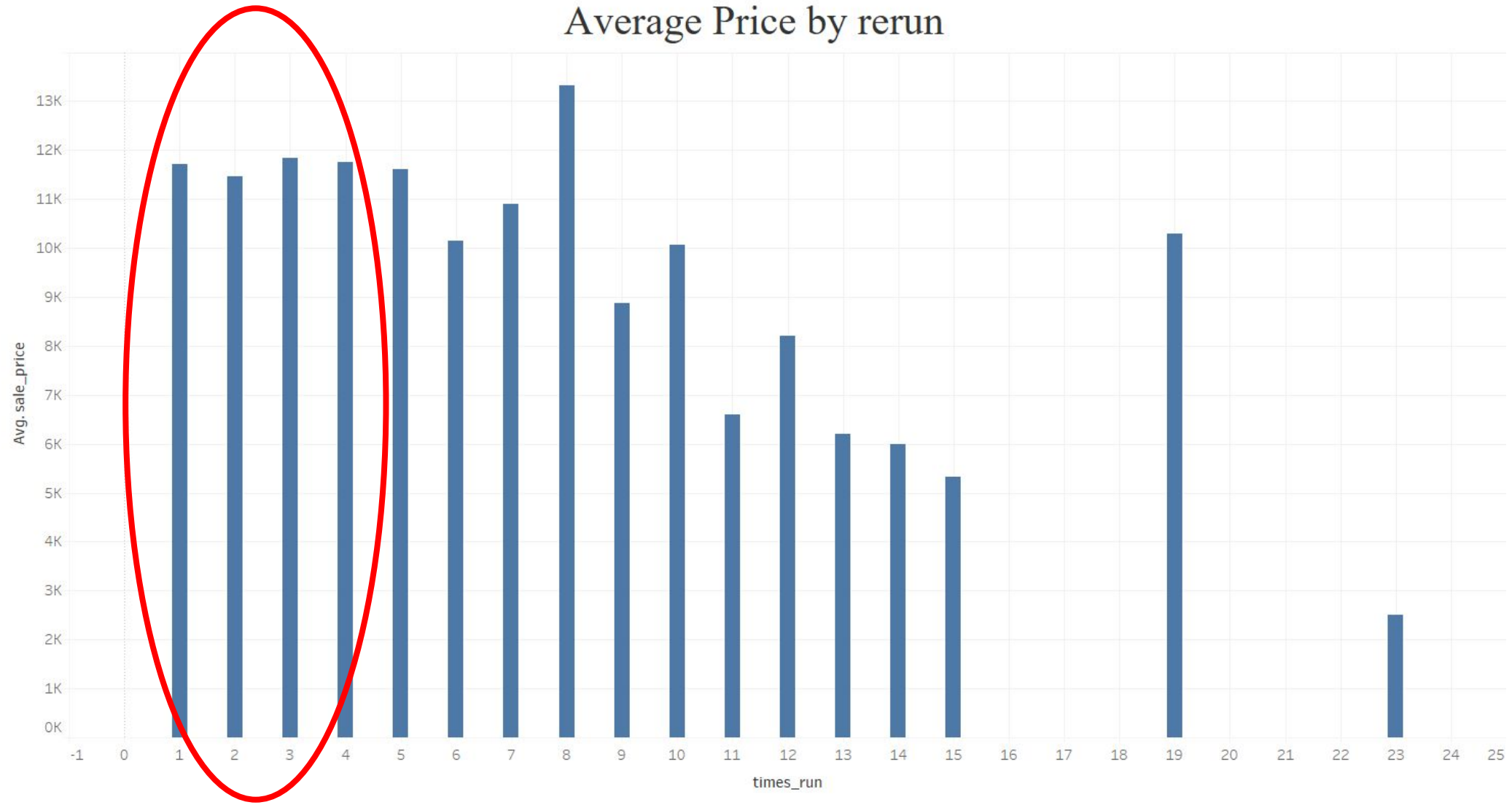
Color | Size | Label | Detail | Tooltip

SUM(Sale Pric..
Region
Auction Code
YEAR(Sold ..
Region

Regional Analyze



AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics



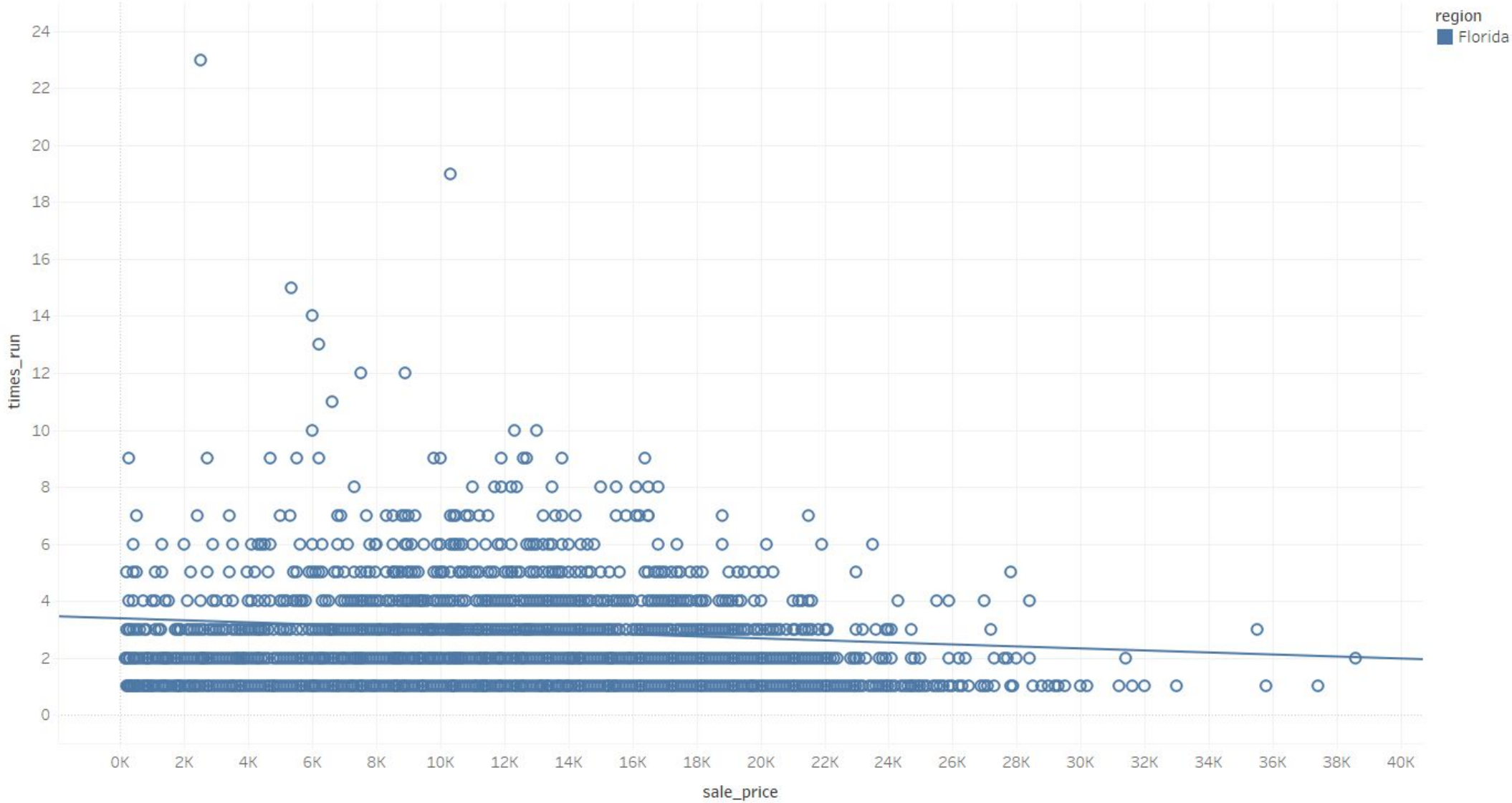
The plot of average of sale_price for times_run.

AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

Price and Times Run in FL



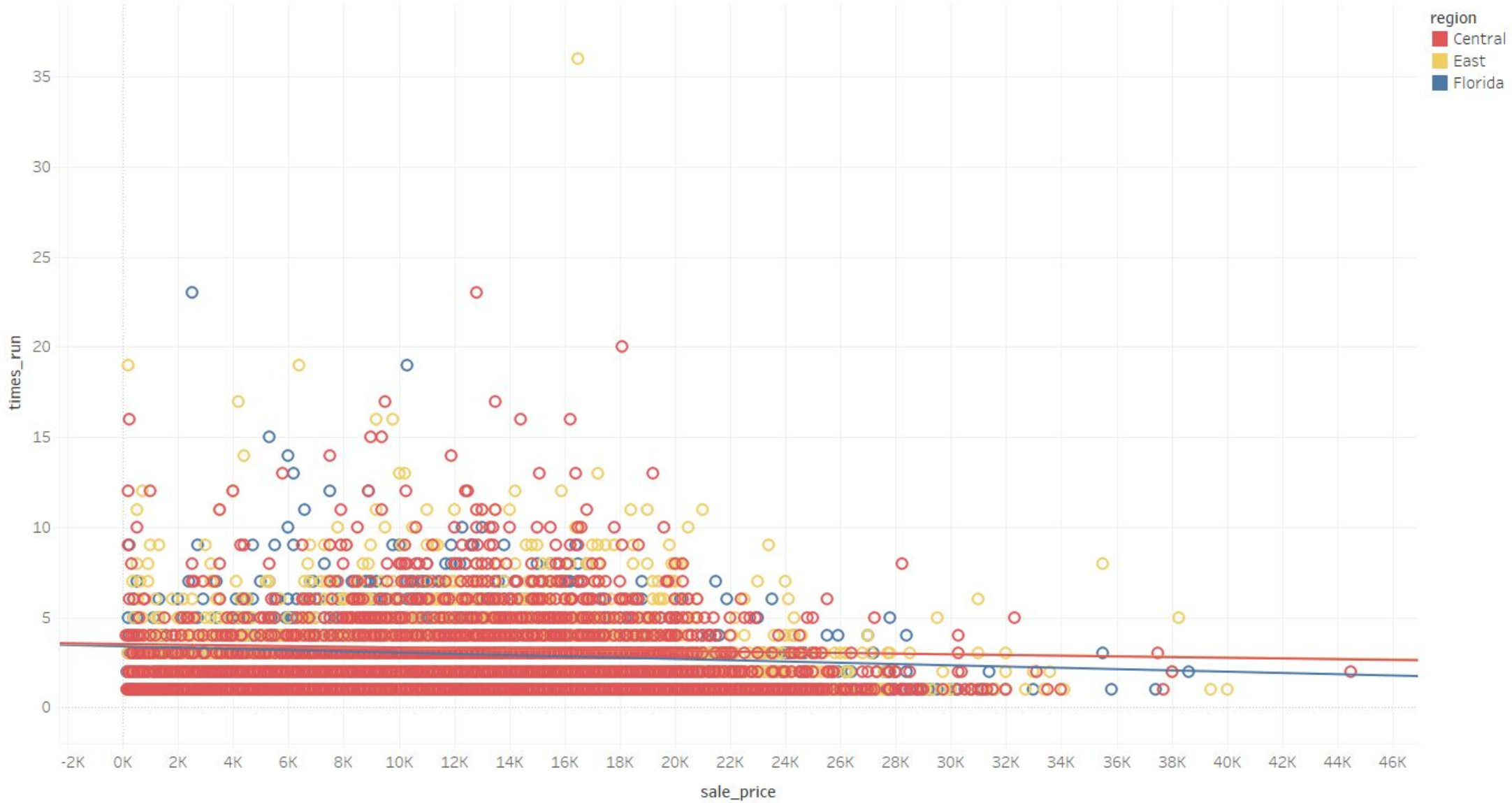
Sale_price vs. times_run. Color shows details about region. The view is filtered on region, which keeps Florida.

AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

Price and Times Run in Top 3 Region



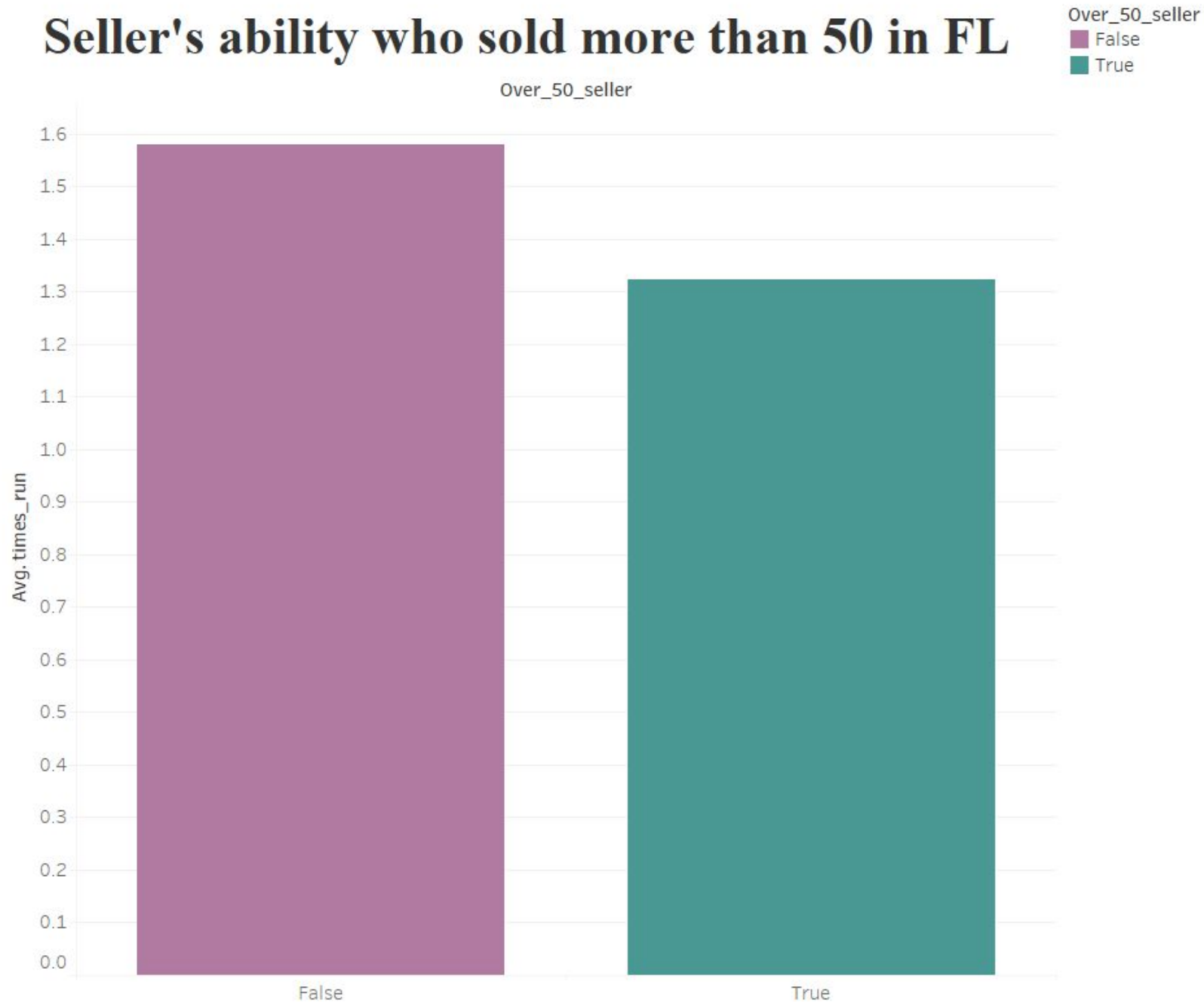
Sale_price vs. times_run. Color shows details about region. The view is filtered on region, which keeps Central, East and Florida.

AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

Seller's ability who sold more than 50 in FL

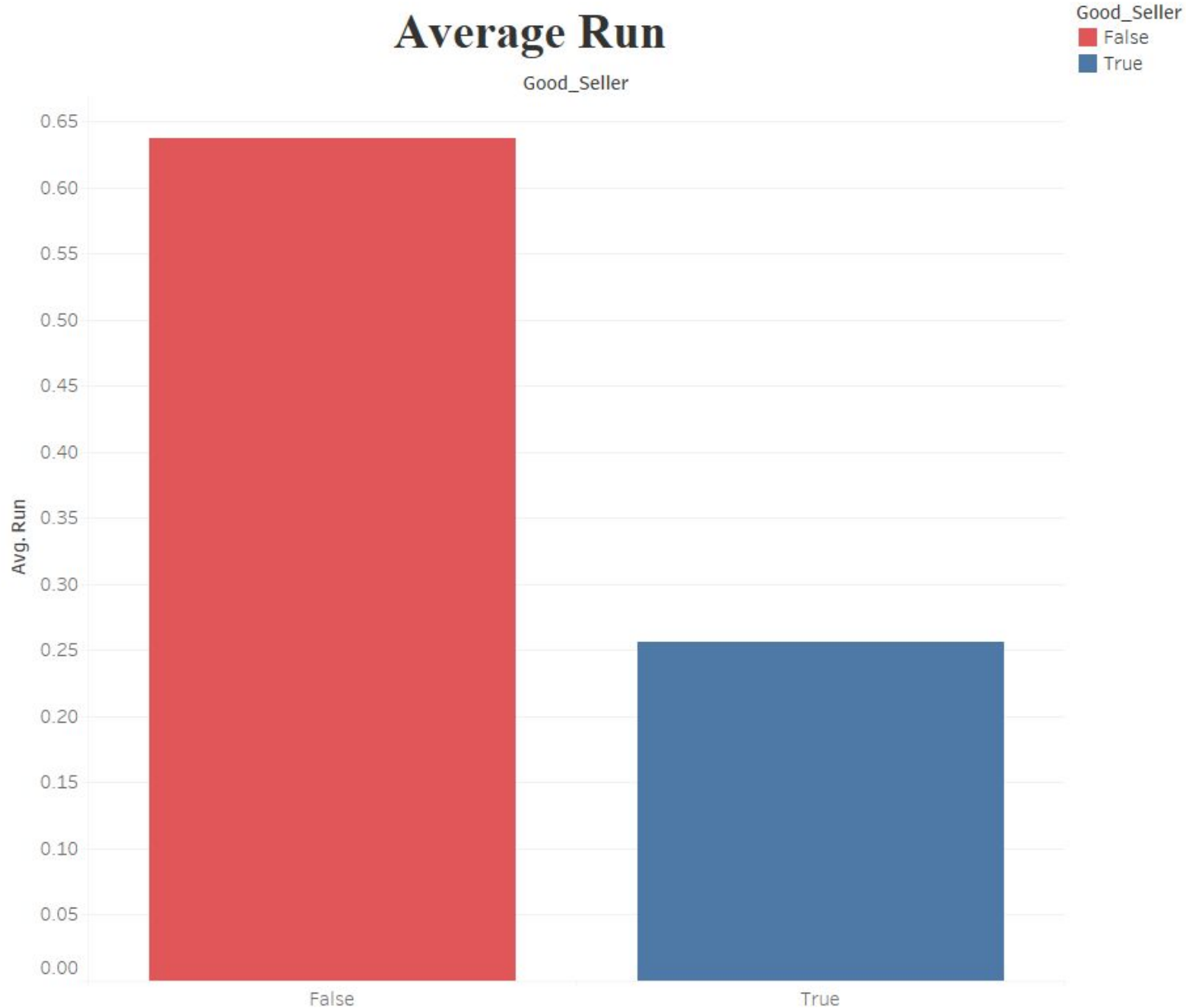


AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

Average Run



AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics

RERUN LOGISTIC REGRESSION MODEL

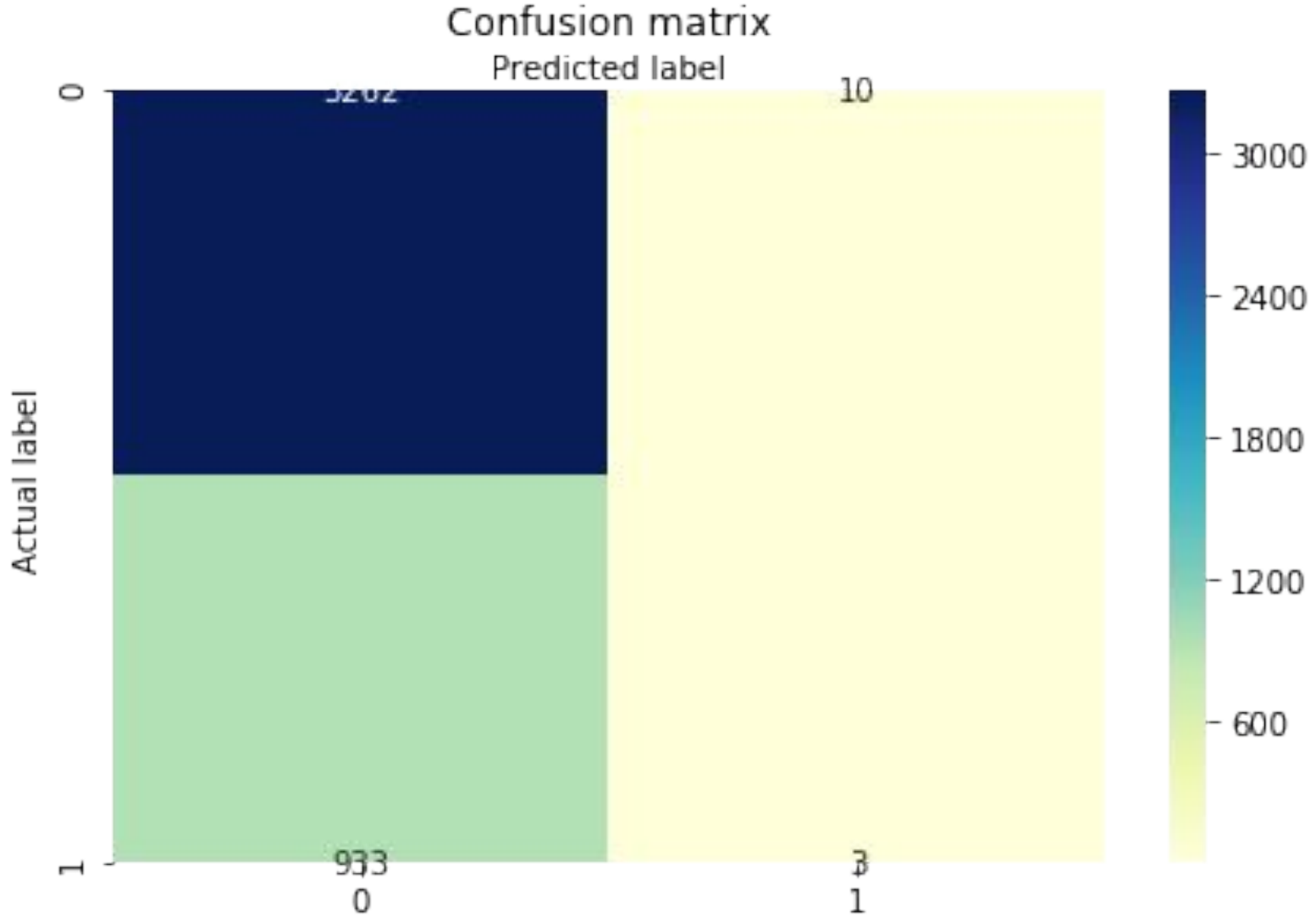
AUBURN UNIVERSITY

RAYMOND J. HARBERT COLLEGE OF BUSINESS

Business Analytics

```
Call:
glm(formula = rerun ~ condition_grade + good_seller + sale_price,
    family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2547  -0.7497  -0.5946  -0.5240   2.0829

Coefficients:
                    Estimate   Std. Error z value             Pr(>|z|)
(Intercept)     -0.473301210  0.101741384  -4.65? 0.0000032872648978?? ***
condition_grade -0.021752416  0.003384397  -6.42? 0.0000000001299196?? ***
good_seller     -1.009809868  0.057314892 -17.61? < 0.0000000000000002 ***
sale_price       0.000054588  0.000006768   8.06? 0.0000000000000007?? ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8992.5  on 8415  degrees of freedom
Residual deviance: 8653.3  on 8412  degrees of freedom
AIC: 8661.3

Number of Fisher Scoring iterations: 4
```

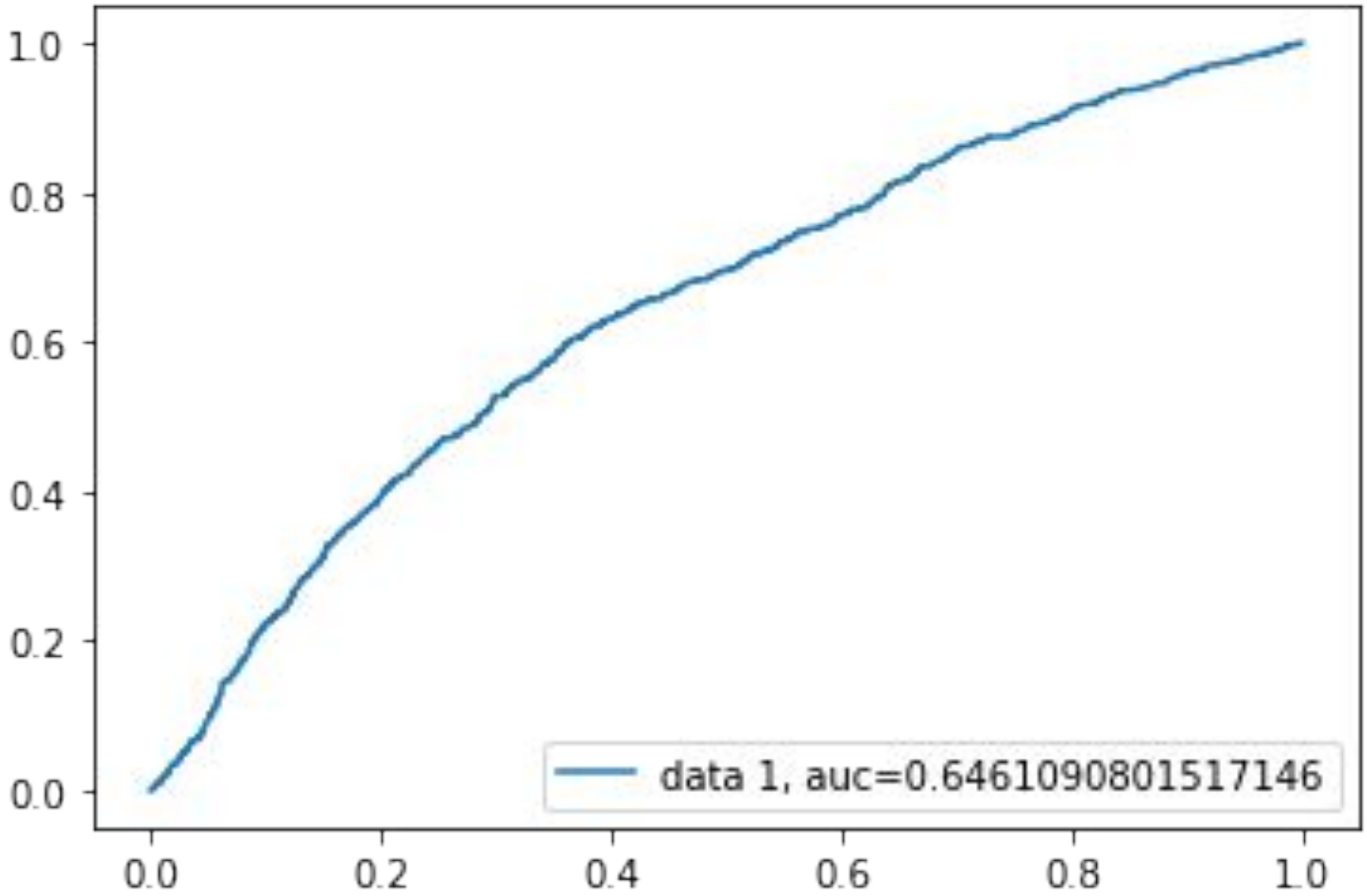
AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics



Accuracy: 0.78

Precision: 0.23

AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics



AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics

```
exp(coef(rerunlogit))
  (Intercept) condition_grade      good_seller       sale_price
    0.6229424       0.9784825        0.3642882        1.0000546
```

When condition grade is increased and other is same as base line (no change), the average probability of rerun is 2% less than base line.

When seller is good seller (over 50 sold with good performance) and other is same as base line (no change), the average probability of rerun is 64% less than base line.

When sale price is increased by $1K ($1,000) and other is same as base line (no change), the average probability of rerun is 5% more than base line.

AUBURN
UNIVERSITY
RAYMOND J. HARBERT
COLLEGE OF BUSINESS
Business Analytics

Conclusion

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

SMALLER MILEAGE → HIGHER PRICE

- Mileage influence price more than any others.
- Regional difference and seasonality do not affect price much

AUBURN UNIVERSITY
RAYMOND J. HARBERT COLLEGE OF BUSINESS
Business Analytics

BETTER SELLER

BETTER PERFORMANCE!!

AUBURN
UNIVERSITY

RAYMOND J. HARBERT
COLLEGE OF BUSINESS

Business Analytics

QUESTION