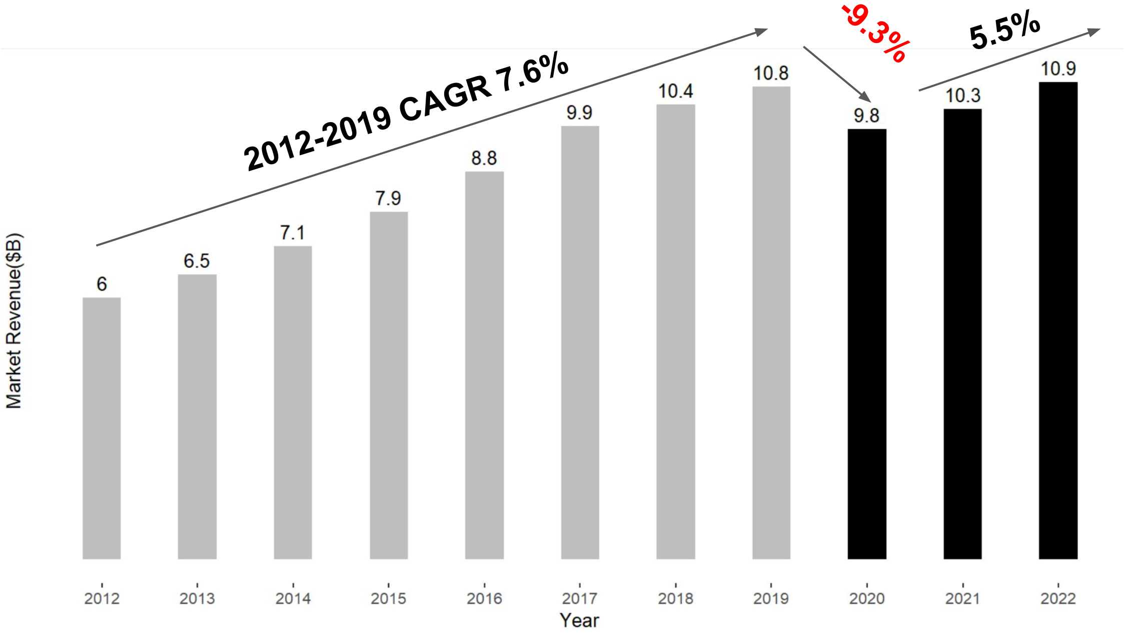
Used Car Sales Price Prediction

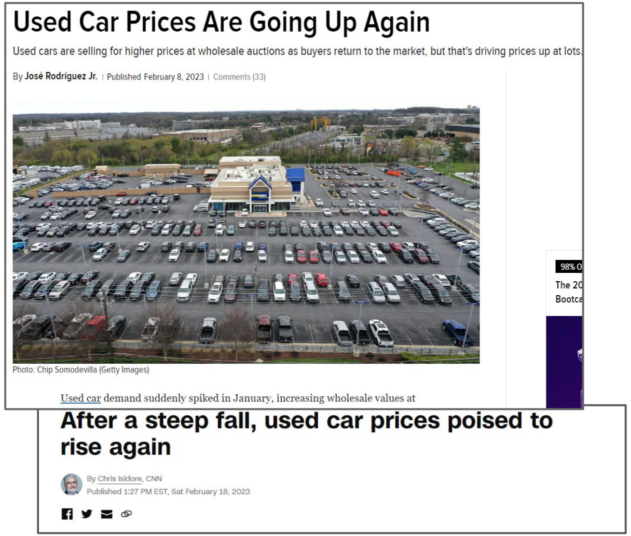
Team 9 Chen Zheng, C K Anand Prakash, Hui Xu, Jaewook Shin, Matthew Kwong

Background

US used car market is still booming.

California used car market revenue
*Statista





⋮

Case Scenario

“ Hello, I am running a used car business in California,
I need 100 more cars to meet the market demand, and I have three options for a supplier.
I want to choose one who can provide me with the best cars with the highest selling price.
Can you help me with this? ”

Who should our client choose?

Supplier A	Supplier B	Supplier C
		
“I’ve statistically chosen the 100 best cars”	“I sourced the best cars through my social network”	“Why do you need other suppliers, why?”

Project Overview

Objective

- **Predict the sales price of cars from each supplier, and select the best one**

Data set used

- **Kaggle, “used-car-auction-prices”**
  - Original: 558,838 rows / 16 variables
  - Cleaned: 20,000 rows / 21 variables*

    Filtered California cars only
    Created dependent variable: sales premium
    Gathered additional variables: awarded

Methodologies applied

- **Linear Regression (Lasso Regression)**
- **Clustering (K-means)**
- **Classification (Support Vector Machine)**

Approach

1. **Explore historical sales data**
2. **Cluster the historical sales data**
3. **Prediction model building**
4. **Predict price premium of cars from each supplier**

“Why do you need this process?”

“First, I will provide you with my sales record.
Can you show me any insights?

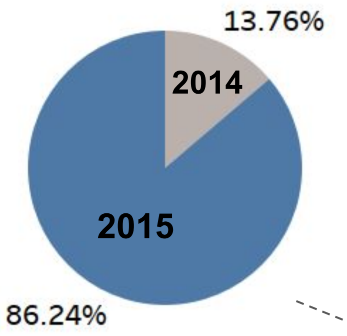
1. **Explore historical sales data**
2. Cluster the historical sales data
3. Prediction model building
4. Predict price premium of cars from each supplier

Exploratory Analysis (1)

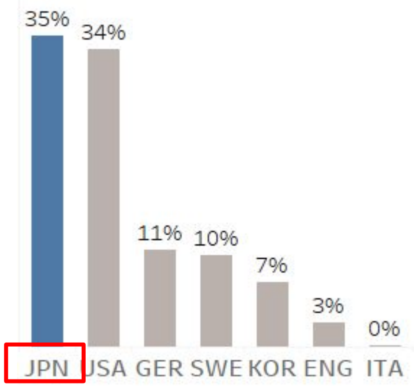
The company has sold 2,906 cars so far. Consumers seem to prefer Japanese make and below 2-year-old cars. As for individual brands, Ford was the most popular one.

Sold 2,906 cars so far…

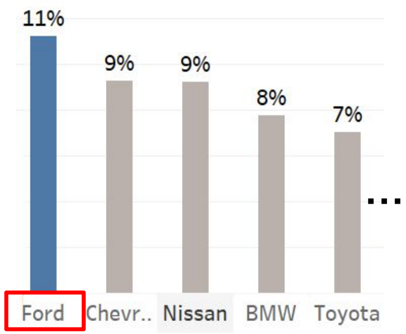
By Fiscal Year



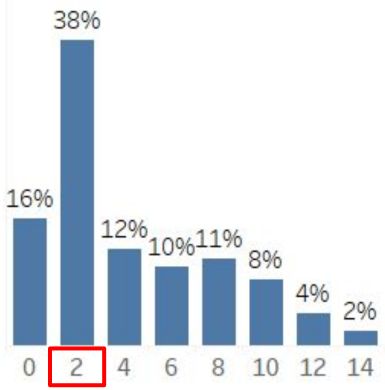
By Nationality



By Make



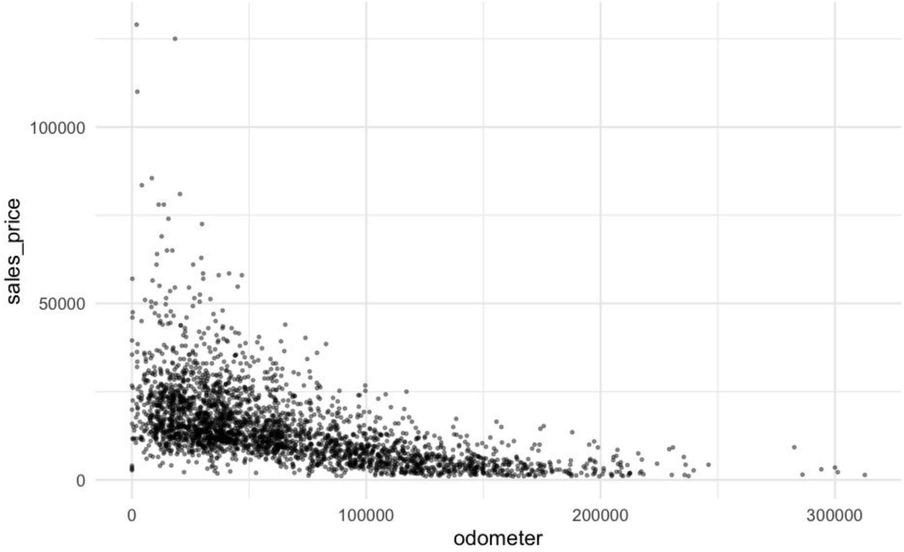
By Car Age

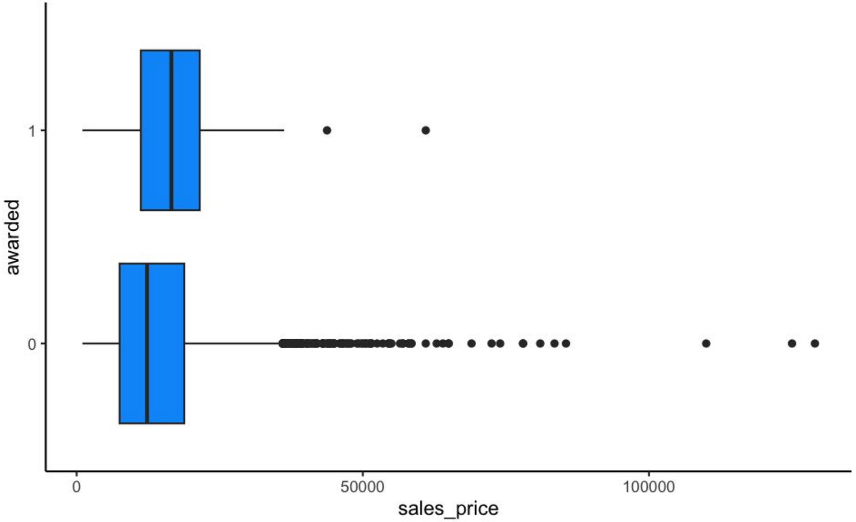


This can be viewed as the consumer preference for used cars in California

Exploratory Analysis (2)

Mileage seems to have negative impacts on the sales prices.
If a certain car is cited as a good one by the media, the car's price tends to increase.





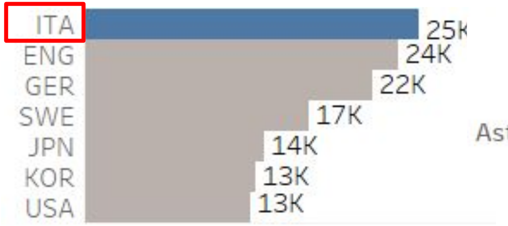
*Award: consumer report, “best cars of the year”

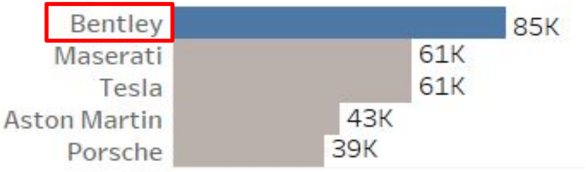
Exploratory Analysis (3)

Apart from the sales price, some cars generated higher price premiums, which means consumers were willing to pay more than the market price for those cars.

By Nationality | **By Make** | **By Body-type**

Avg. Sales Price





Vs.

Avg. Price Premium*

We chose the price premium as our dependent variable

*Price Premium = Sales Price / The Manheim Market Report Price

“I always felt that customers showed different buying patterns for my cars… Can you check if my cars can be segmented?”

1. Explore historical sales data
2. Cluster the historical sales data
3. Prediction model building
4. Predict price premium of cars from each supplier

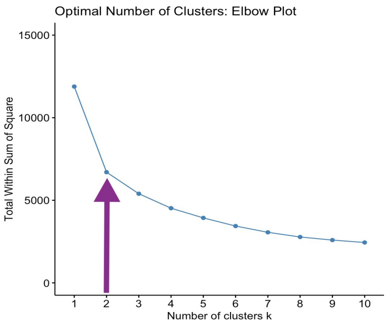
Clustering

Client’s cars can be clustered into two groups.
Given their heterogeneous characteristics, we decided to treat them independently.

Variables considered

```
> str(cardf1)
'data.frame':   2906 obs. of
 $ X           : int  17 18
 $ Premium     : num  0.904
 $ sales_price : int  3300 !
 $ MMR_price   : int  3650 .
 $ make_yr     : int  2008 :
 $ Age         : int  7 10 !
 $ sale_yr     : int  2015 :
 $ sale_mnt    : chr  "Jun"
 $ sale_day    : chr  "Wed"
 $ make        : chr  "Kia"
 $ national    : chr  "KOR"
 $ model       : chr  "Spect
 $ trim        : chr  "EX"  '
 $ awarded     : num  0 0 0
 $ body        : chr  "Sedar
 $ make_yr.1   : int  2008 :
 $ odometer    : int  93632
 $ condition   : num  2.9 2
 $ transmission: int  1 1 1
 $ ext_col     : chr  "blue'
 $ int_col     : chr  "gray'
 $ seller      : chr  "the (
```

Elbow chart



We chose 2 as the optimal size using the knee / elbow method

Segmentation result

```
> clust_data_means
```

	Group.1	Premium	sales_price	MMR_price	Age	awarded	odometer	condition	transmission
1	1	1.01	19063.24	19021.36	2.54	0.07	37733.11	3.72	0.97
2	2	1.04	6431.09	6335.02	9.02	0.05	120572.05	2.73	0.95

Cluster 1

Newer entry-level luxury, good condition: Chevy Camaro, Infiniti G37

Cluster 2

Older well-driven economy, fair condition: Honda Odyssey, Dodge Charger RT

“I see, my cars can be clustered into two groups. Now let’s build a prediction model”

1. Explore historical sales data
2. Cluster the historical sales data
3. Prediction model building
4. Predict price premium of cars from each supplier

Prediction Model Building

We built two different Lasso regression models for each cluster.

Issues

1. **Information discrepancy**
   - Our data has sales related Information, such as sales price and sales data.
     But dealer's data set doesn't

   ☞ **Remove sales-related variables on modeling**

2. **Still too many variables**
   - 103 variables in total, after converting all categorical variables to dummy variables (high risk of multicollinearity)

   ☞ **Take alternative approach**

Alternative Approach

1. **Feature selection with the Client's domain knowledge**
   - "MMR price, Age, Odometer, Condition, Awarded, and Transmission is everything"
   - $R2$: 0.35, RMSE: 0.09738 (for cluster 1)

2. **Lasso regression**
   - $R2$: 0.61, RMSE: 0.0979 (for cluster 1)

   ☞ **Lasso regression showed better model fit and prediction power**

Model building

Model for Cluster 1

```
> coef(lasso.c1)
103 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)      6.283164e-01
MMR_price       -9.798730e-06
Age              7.374449e-03
awarded          .
odometer        -4.037326e-08
condition        1.444955e-01
transmission     .
makeAcura        .
```

Model for Cluster 2

```
> coef(lasso.c2)
103 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)      6.344006e-01
MMR_price       -8.707714e-06
Age              7.032860e-03
awarded          .
odometer         .
condition        1.396396e-01
transmission     .
makeAcura        .
makeAston.Martin .
makeAudi         .
```

Prediction Model Building

"Why do you need
two different models? why?"

		1. Single model for the entire data set	2. Two models optimized for each cluster	
Client's domain knowledge model	R-Squared	0.2226	Cluster1: 0.3545 Cluster2: 0.2179	Avg 0.286
	RMSE	0.1764	Cluster1: 0.0995 Cluster2: 0.2468	Avg 0.173
Lasso Regression	R-Squared	0.6531	Cluster1: 0.6279 Cluster2: 0.756	Avg 0.691
	RMSE	0.1745	Cluster1: 0.0983 Cluster2: 0.2424	Avg 0.170

Two model approach shows better prediction power and model fit

“Now we have prediction models.
Tell me which supplier should I choose?”

1. Explore historical sales data
2. Cluster the historical sales data
3. Prediction model building
4. Predict price premium of cars from each supplier

“This is a tense moment.
Do you need a joke?”

Prediction (1)

First, we classified each supplier's cars and applied the cluster-optimized prediction model. And the winner is…

Support Vector Machine with "radial" kernel

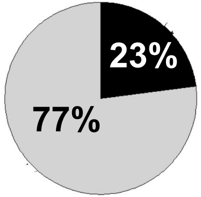
Modeled with the top 10 variables
highly correlated with the cluster variable

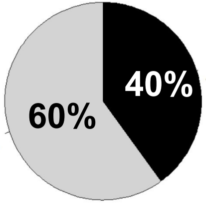
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 907  31
         1   7 498

               Accuracy : 0.9737
                 95% CI : (0.964, 0.9813)
    No Information Rate : 0.6334
    P-Value [Acc > NIR] : < 2.2e-16
```

We chose SVM since the dataset is high-dimensional, and the client favored robustness of a model over interpretability

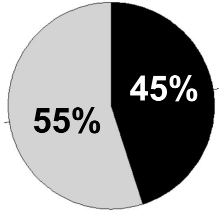
Classify each supplier's cars

Cluster 1
Cluster 2





"My social network said I will get the deal"



Prediction (2) - Conclusion

Supplier A shows the highest and most stable price premium.
Therefore, we highly recommend him for your business.

Forecast of price premium of 100 cars

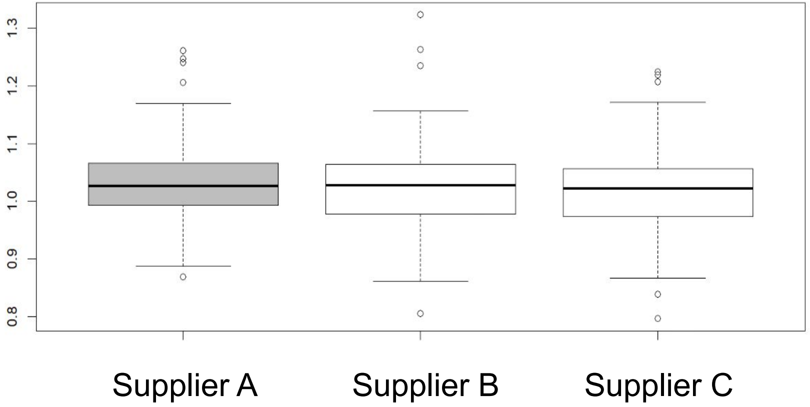
Mean : 1.034
Sd : 0.069

Mean : 1.023
Sd : 0.095

Mean : 1.014
Sd : 0.079

"I will quit and teach marketing instead"

Distribution of price premium



Difference is marginal
since, originally, they are from the same data set

Lessons Learned

Data

Modeling

Storytelling

" Dependent variables are something that can be created "

- Sometimes playing around with variables can provide you with different insights

" Every population has heterogeneous nature "

- It is the era of "micro-segmentation" modeling on segmented groups can be helpful

" It's a message, not a model "

- In real-world business settings, decision makers tend to focus more on messages than models

Appendix

Data Set

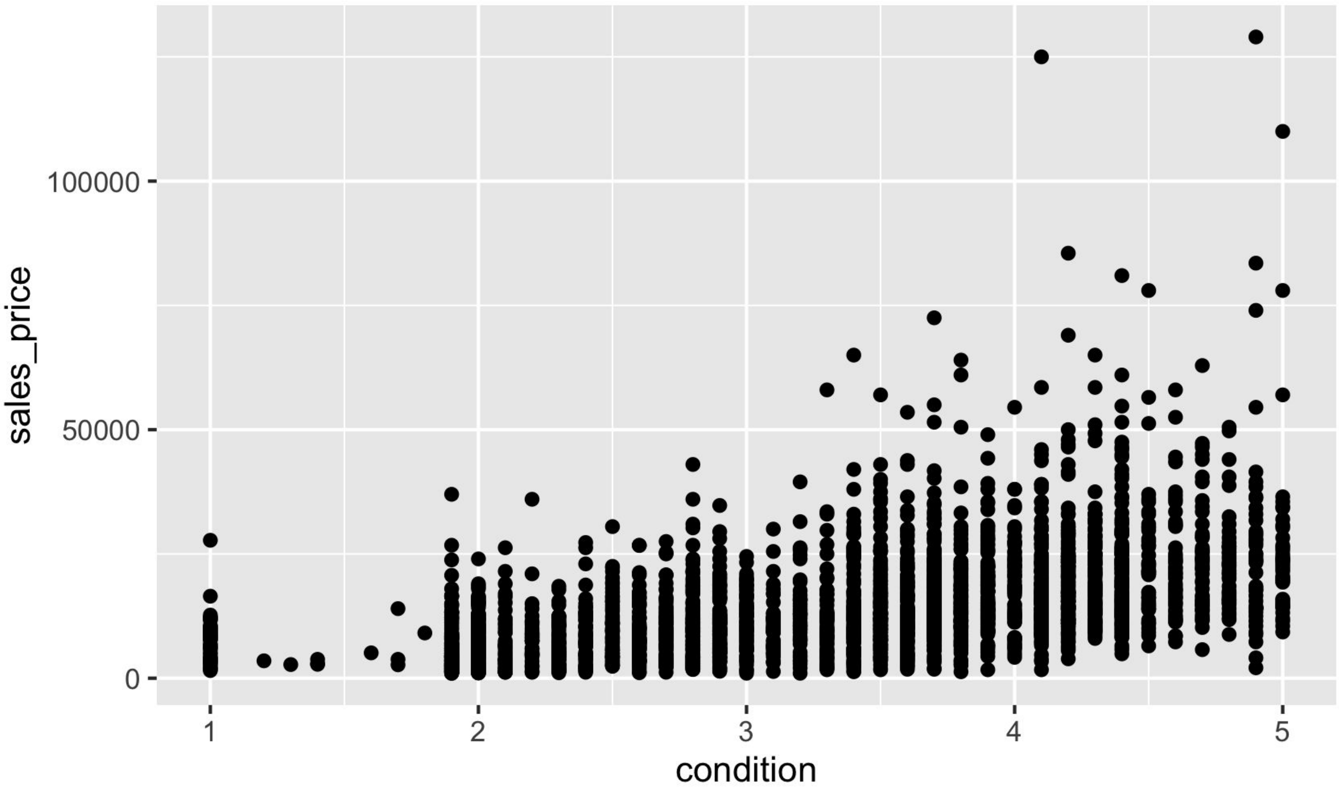
Original Data Set
Kaggle, “used-car-auction-prices” Original: 558,838 rows / 16 variables

year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingprice	saledat
2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg566472	ca	5	16639	white	black	kia motors	20500	21500	Tue De
2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg561319	ca	5	9393	white	beige	kia motors	20800	21500	Tue De
2014	BMW	3 Series	328i SULE	Sedan	automatic	wba3c1c51ek116351	ca	4.5	1331	gray	black	financial se	31900	30000	Thu Jan
2015	Volvo	S60	T5	Sedan	automatic	yv1612tb4f1310987	ca	4.1	14282	white	black	volvo na r	27500	27750	Thu Jan
2014	BMW	6 Series G	650i	Sedan	automatic	wba6b2c57ed129731	ca	4.3	2641	gray	black	financial se	66000	67000	Thu De
2015	Nissan	Altima	2.5 S	Sedan	automatic	1n4al3ap1fn326013	ca	1	5554	gray	black	enterprise	15350	10900	Tue De
2014	BMW	M5	Base	Sedan	automatic	wbsfv9c51ed593089	ca	3.4	14943	black	black	the hertz c	69000	65000	Wed D
2014	Chevrolet	Cruze	1LT	Sedan	automatic	1g1pc5sb2e7128460	ca	2	28617	black	black	enterprise	11900	9800	Tue De
2014	Audi	A4	2.0T Premi	Sedan	automatic	wauffafl3en030343	ca	4.2	9557	white	black	audi missi	32100	32250	Thu De
2014	Chevrolet	Camaro	LT	Convertibl	automatic	2g1fb3d37e9218789	ca	3	4809	red	black	d/m auto	26300	17500	Tue Jan
2014	Audi	A6	3.0T Presti	Sedan	automatic	wauhgafc0en062916	ca	4.8	14414	black	black	desert aut	47300	49750	Tue De
2015	Kia	Optima	LX	Sedan	automatic	5xxgm4a73fg353538	ca	4.8	2034	red	tan	kia motors	15150	17700	Tue De
2015	Ford	Fusion	SE	Sedan	automatic	3fa6p0hdxfr145753	ca	2	5559	white	beige	enterprise	15350	12000	Tue Jan
2015	Kia	Sorento	LX	SUV	automatic	5xyktca66fg561407	ca	5	14634	silver	black	kia motors	20600	21500	Tue De
2014	Chevrolet	Cruze	2LT	Sedan	automatic	1g1pe5sbxe7120097	ca		15686	blue	black	avis rac/sa	13900	10600	Tue De
2015	Nissan	Altima	2.5 S	Sedan	automatic	1n4al3ap5fc124223	ca	2	11398	black	black	enterprise	14750	14100	Tue De
2015	Hyundai	Sonata	SE	Sedan		5npe24af4fh001562	ca		8311	red	??"avis tra'	15200	4200	Tue Dec 16 2014 13:00:	
2014	Audi	Q5	2.0T Premi	SUV	automatic	wa1lfafpxea085074	ca	4.9	7983	white	black	audi north	37100	40000	Thu De
2014	Chevrolet	Camaro	LS	Coupe	automatic	2g1fa1e39e9134494	ca	1.7	13441	black	black	wells fargo	17750	17000	Tue De
2014	BMW	6 Series	650i	Convertibl	automatic	wbayp9c53ed169260	ca	3.4	8819	black	black	the hertz c	68000	67200	Wed D
2015	Chevrolet	Impala	LTZ	Sedan	automatic	2g1165s30f9103921	ca	1.9	14538	silver	black	enterprise	24300	7200	Tue Jul
2014	BMW	5 Series	528i	Sedan	automatic	wba5a5c51ed501631	ca	2.9	25969	black	black	financial se	34200	30000	Tue Feb
2014	Chevrolet	Camaro	LT	Convertibl	automatic	2g1fb3d31e9134662	ca		33450	black	black	avis rac/sa	20100	14700	Tue De

1. **Delete NA values**
2. **Delete invalid values**
   - e.g., sales_price = white
3. **Delete outliers**
   - e.g. price = 30,000,000
4. **Create new variables**
   - Premium (sales_price / MMR_price)
   - Awarded (consumer reports)
   - National (nationality of makes)
5. **Dummy variable conversion**
   - e.g., transmission, state
6. **Sampled 20,000 observations**
7. **Filtered California cars only**

Sales price vs. Condition

"Condition" variable is an important indicator of sales price in the used car market.
Used cars under greater condition is more likely to be sold in higher prices.

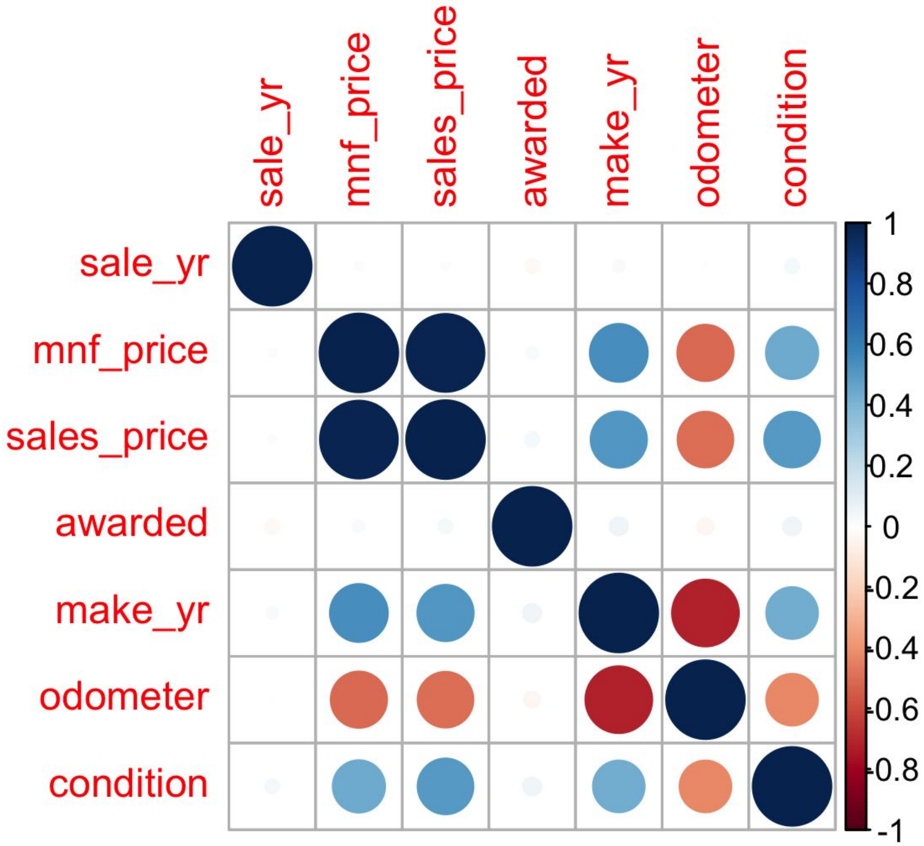


Correlation between features

- “Odometer” is negatively related to sales prices in the used car market.

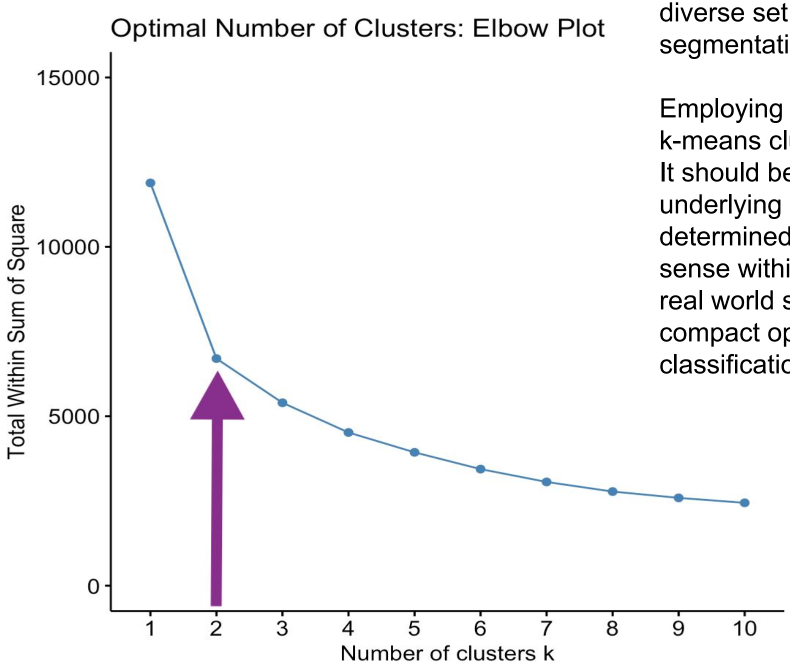
- “Mnf price” is positively related to sales prices.

- “Make_yr” and “condition” variables are also an important indicator of sales prices.



Clustering

```
#Elbow plot
fviz_nbclust(x=clusterdf3.1std, FUNcluster = kmeans, nstart=100, method="wss", k.max = 10) +
  labs(title="Optimal Number of Clusters: Elbow Plot") +
  coord_cartesian(ylim=c(0,15000)) + geom_line(size=2) # ---> looks like K=2 optimal
```



As shown in the exploratory analysis of this project, the cars in this dataset represent a diverse set of vehicular demographics. As such, it was necessary to examine the natural segmentation of the dataset.

Employing the knee/elbow approach to determine the optimal number of clusters k for k-means clustering, we selected 2 for k, as the elbow/knee is distinctly accentuated at k=2. It should be noted that we did also attempt higher values of k and did examine the underlying averages for the segments created for these higher k values, but we ultimately determined that two segments represented a clear delineation across groups that made sense within the scope of this project. Applying the decision making within clustering to our real world setting, both the elbow/knee method as well as the benefit of having more compact operational complexity (from two clusters, as opposed to three or four per classification model) pointed towards using 2 clusters.

```
# Group.1 Premium sales_price MMR_price  Age awarded  odometer condition transmission
# 1       1    1.01    19063.24  19021.36 2.54    0.07  37733.11      3.72         0.97
# 2       2    1.04     6431.09   6335.02 9.02    0.05 120572.05      2.73         0.95

# #  Group.1 Premium sales_price MMR_price  Age awarded  odometer condition transmission
# 1       1    0.98    12625.51  12983.74 3.33    0.05  51798.69      3.06         0.97
# 2       2    1.06     5834.92   5635.82 9.69    0.05 129316.98      2.72         0.95
# 3       3    1.02    24044.02  23667.11 2.25    0.09  29989.77      4.23         0.98
```

Prediction Modeling (1)

Client's Model Summary for train data for cluster 1

```
> summary(fit0)

Call:
lm(formula = Premium ~ MMR_price + Age + awarded + odometer +
    condition + transmission, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67440 -0.04945  0.00075  0.05276  0.57193 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.435e-01  2.232e-02  28.823  < 2e-16 ***
MMR_price   -1.479e-06  2.673e-07  -5.534 3.76e-08 ***
Age          1.634e-02  2.176e-03   7.512 1.07e-13 ***
awarded      9.647e-03  1.079e-02   0.895 0.371206    
odometer     5.580e-07  1.612e-07   3.461 0.000554 ***
condition    8.503e-02  3.542e-03  24.006  < 2e-16 ***
transmission 1.120e-02  1.655e-02   0.677 0.498542    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1008 on 1320 degrees of freedom
Multiple R-squared:  0.3512,	Adjusted R-squared:  0.3483 
F-statistic: 119.1 on 6 and 1320 DF,  p-value: < 2.2e-16
```

Client's Model Summary for entire data for cluster 1

```
> summary(fit.ceo1)

Call:
lm(formula = Premium ~ MMR_price + Age + awarded + odometer +
    condition + transmission, data = a1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67723 -0.04817  0.00085  0.05342  0.57383 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.569e-01  1.858e-02  35.365  < 2e-16 ***
MMR_price   -1.524e-06  2.303e-07  -6.620 4.66e-11 ***
Age          1.745e-02  1.785e-03   9.779  < 2e-16 ***
awarded      1.268e-02  8.924e-03   1.421 0.155493    
odometer     4.614e-07  1.314e-07   3.511 0.000457 ***
condition    8.258e-02  2.903e-03  28.449  < 2e-16 ***
transmission 7.912e-03  1.425e-02   0.555 0.578735    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09977 on 1889 degrees of freedom
Multiple R-squared:  0.3545,	Adjusted R-squared:  0.3524 
F-statistic: 172.9 on 6 and 1889 DF,  p-value: < 2.2e-16
```

Prediction Modeling (2)

Client's Model Summary for train data for cluster 2

```
> summary(fit0)

Call:
lm(formula = Premium ~ MMR_price + Age + awarded + odometer + 
    condition + transmission, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.81445 -0.14524 -0.00665  0.12935  1.78719 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.943e-01  7.627e-02   7.792 2.40e-14 ***
MMR_price    -1.403e-05  3.027e-06  -4.633 4.29e-06 ***
Age           1.190e-02  3.951e-03   3.011   0.0027 ** 
awarded      -7.402e-03  4.453e-02  -0.166   0.8680    
odometer     -2.236e-07  1.601e-07  -1.396   0.1630    
condition     1.665e-01  1.376e-02  12.103  < 2e-16 ***
transmission  4.747e-03  4.306e-02   0.110   0.9122    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2548 on 700 degrees of freedom
Multiple R-squared:  0.2212,	Adjusted R-squared:  0.2145 
F-statistic: 33.13 on 6 and 700 DF,  p-value: < 2.2e-16
```

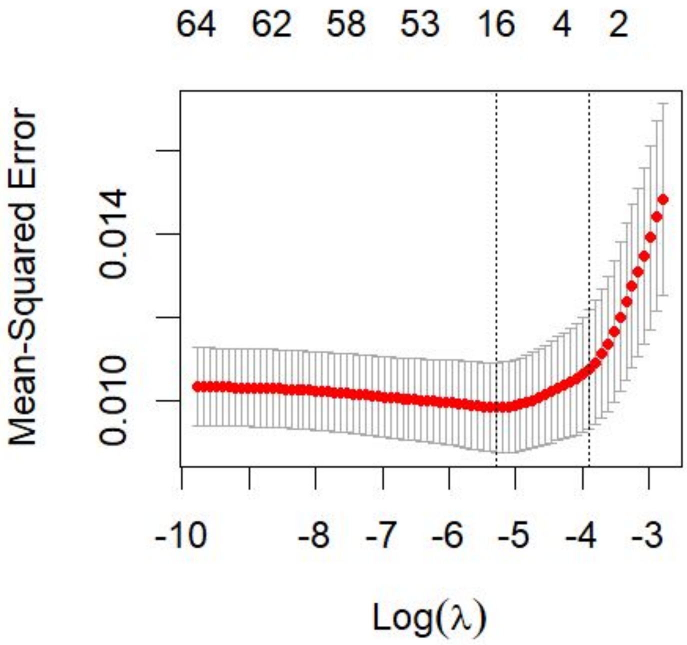
Client's Model Summary for entire data for cluster 2

```
> summary(fit.ceo2)

Call:
lm(formula = Premium ~ MMR_price + Age + awarded + odometer + 
    condition + transmission, data = a2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.81175 -0.15047 -0.00539  0.13223  1.79630 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.998e-01  6.690e-02   8.965  < 2e-16 ***
MMR_price    -1.297e-05  2.591e-06  -5.008 6.49e-07 ***
Age           1.305e-02  3.311e-03   3.942 8.65e-05 ***
awarded       2.593e-02  3.905e-02   0.664   0.5069    
odometer     -2.885e-07  1.499e-07  -1.925   0.0545 .  
condition     1.606e-01  1.152e-02  13.940  < 2e-16 ***
transmission  3.835e-03  3.868e-02   0.099   0.9211    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2578 on 1003 degrees of freedom
Multiple R-squared:  0.2079,	Adjusted R-squared:  0.2031 
F-statistic: 43.87 on 6 and 1003 DF,  p-value: < 2.2e-16
```

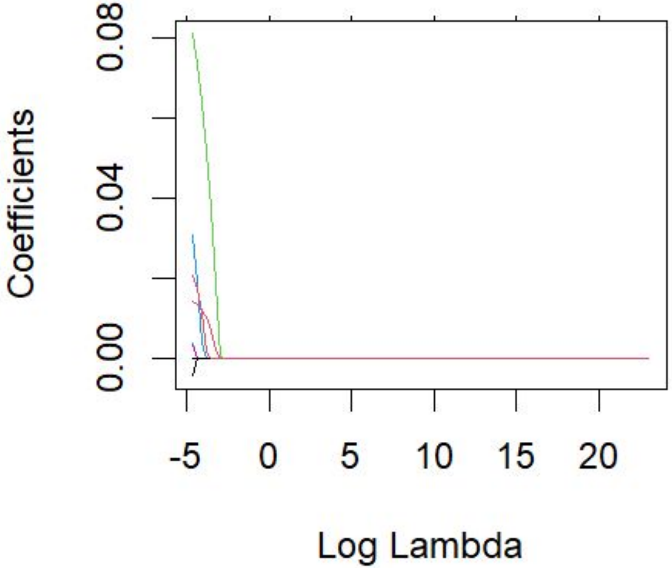
Prediction Modeling (3)

Lasso Regression

To find the best lambda, we first tried a loop method, and then conducted the cross validation



Using the best lambda, we built the lasso regression model, the number of variables included in the model reduced to 11.



Classifications(1)

Random Forest

```
install.packages('randomForest')
library(randomForest)
set.seed(223344)
bag.train.100 <- randomForest(ClusterNo ~ .,
                              data = r2,
                              mtry = 12, ntree = 100,
                              importance = TRUE)
bag.train.100
# Mean of squared residuals: 0.0235618
# % Var explained: 89.61

bag.train.50 <- randomForest(ClusterNo ~ .,
                             data = r2,
                             mtry = 12, ntree = 50,
                             importance = TRUE)
bag.train.50
#  Mean of squared residuals: 0.02322343
#   % Var explained: 89.76
```

The accuracy of random forest is about 89.76%, with 50 number of trees, MSE=0.0232

```
Call:
 randomForest(formula = ClusterNo ~ ., data = r2, mtry = 12, ntree = 50,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 50
No. of variables tried at each split: 12

          Mean of squared residuals: 0.02322343
                    % Var explained: 89.76
```

Logistic Regression Confusion Matrix

```
> yhat.test.class <- ifelse(yhat.test > 0.5, 1, 0)
> table(r2[test, ]$ClusterNo,
+       yhat.test.class,
+       dnn = c("Actual","Predicted"))
      Predicted
Actual   0   1
     0 908  32
     1  31 482
```

This confusion matrix shows the performance of the classification model on the test set. The rows represent the actual class labels, while the columns represent the predicted class labels.

Looking at the table, we can see that the model predicted 482 instances as positive (1) when they were actually positive, and 908 instances as negative (0) when they were actually negative. However, the model incorrectly predicted 31 instances as negative when they were actually positive (false negatives) and 32 instances as positive when they were actually negative (false positives).

accuracy = (908 + 482) / 1453 = 0.9566
The accuracy is quite high, which means that the model is able to correctly classify most instances in the dataset.

Classifications(2)

SVM

```r
library(e1071)

set.seed(100)

dat.train$ClusterNo <- factor(dat.train$ClusterNo)

# Find the best svm model

tune.out <- tune(svm, ClusterNo ~ Age + odometer + MMR_price + condition + int_colblack + int_colgray + nationalKOR + bodySedan + makeInfiniti + bodyCab,
                 data = dat.train, kernel = "radial",
                 ranges = list(cost = c(0.01, 0.1, 1, 10, 100, 1000), gamma = c(0.5, 1, 2, 3, 4)))

# This is the best SVM model

svm.best <- tune.out$best.model

svm.pred <- predict(svm.best, dat.test, type="class")

table(actual=dat.test[[1]], predict=svm.pred)

library(caret)

confusionMatrix(factor(dat.test[[1]]), svm.pred)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 907  31
         1   7 498

               Accuracy : 0.9737
                 95% CI : (0.964, 0.9813)
    No Information Rate : 0.6334
    P-Value [Acc > NIR] : < 2.2e-16
```

- We select the top 10 highest correlated variables, which are:Age, odometer, MMR_price, condition, int_colblack, int_colgray, nationalKOR, bodySedan, makeInfiniti, bodyCab
- SVM model has a high accuracy of 97.37% on the test data, which means that the model was able to correctly predict the target variable for 97.37% of the observations in the test data set
- Therefore, we choose the SVM model because it has the highest accuracy of the three classification models.

Prediction

Data preparation

We randomly assigned 100 observations to each supplier

The original data set

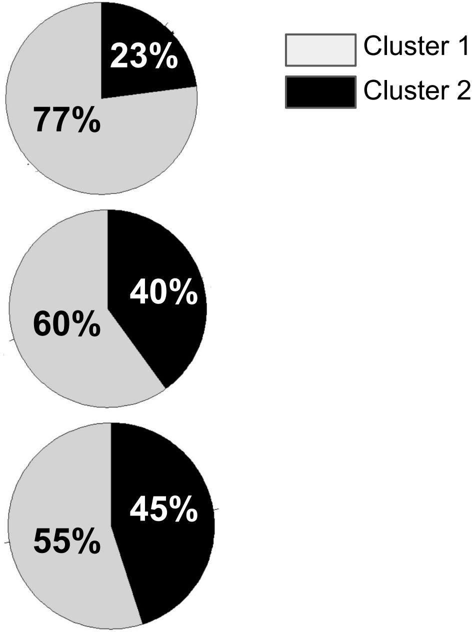
100

100

100

Classification

And then classified each supplier's cars into two groups



Forecasting

Finally, we applied the prediction model to each cluster, and get the results

Prediction for the price premium

	Mean	Standard Deviation
Supplier A	1.034	0.069
Supplier B	1.023	0.095
Supplier C	1.014	0.079