Next Big Trend
Microvan STP Analysis

Research Objective

BACKGROUND

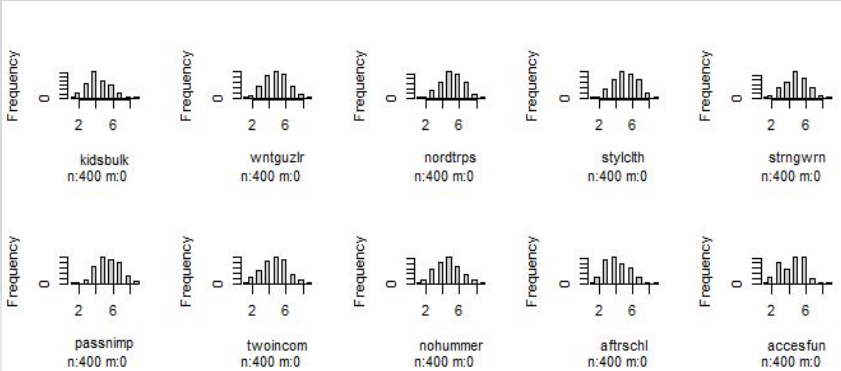
The U.S. Auto Industry is always looking ahead to the next big trend in automotive design. Lead times are lengthy – several years from concept to prototype and several more to the dealer floor – with total costs for a new model line often topping $1B (US). The main method for detecting such trends is primary consumer research, typically starting with targeted focus groups and proceeding to medium- and eventually large-scale surveys.

OBJECTIVE

Make recommendations on which characteristics are most important. To determine whether a customer has a favorable perception of the microvan concept, to **segment** the market, and to determine which segment(s) would be good to **target** based on your analysis of the data.

1 Check Variables

Histograms glance



Overall, we **do not have any concerns** with the distributions of variables. Upon examining the histograms, most variables display a normal distribution, while a few are right-skewed or left skewed (such as secbiggr and next2str). However, these are well expected and even preferred in our dataset, since we want to understand the preference of the studied population.

2 Determine Relationship

As expected, the regression model provides **limited usage for insights**, since there may exist multicollinearity problems among the explanatory variables, which is why we conduct factor analysis to reduce the number of explanatory variables in the next step.

Regression summary

- The regression model containing all 30 variables has an adjusted R-squared of 0.3219, and a large t-statistics test p-value for almost all predictors (only lthrbetr and shdcarpl are significant at the alpha = 0.05 level).
- This indicates that our model has poor fit and explanatory power. Regression at this stage (with raw data and a relatively huge variable set) is therefore useless for insights.

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.023 -1.605 -0.018  1.475  6.508

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.381266   2.958551   0.129   0.8975
kidtrans     0.241114   0.164793   1.463   0.1443
miniboxy     0.177881   0.129258   1.376   0.1696
lthrbetr     0.247630   0.121971   2.030   0.0430 *
secbiggr    -0.104796   0.105833  -0.990   0.3227
safeimpt    -0.018525   0.133630  -0.139   0.8898
buyhghnd     0.112578   0.116162   0.969   0.3331
pricqual     0.105322   0.104796   1.005   0.3155
prmsound     0.010118   0.108312   0.093   0.9256
perfimpt     0.232663   0.128198   1.815   0.0704 .
tkvacatn     0.166171   0.124671   1.333   0.1834
noparkrm     0.178143   0.115804   1.538   0.1248
homlrgst    -0.208684   0.122418  -1.705   0.0891 .
envrminr    -0.033245   0.122777  -0.271   0.7867
needbetw     0.128468   0.102636   1.252   0.2115
suvcmpct     0.215136   0.122643   1.754   0.0802 .
next2str     0.024294   0.106843   0.227   0.8203
carefmny    -0.243143   0.134373  -1.809   0.0712 .
shdcarpl    -0.286783   0.122413  -2.343   0.0197 *
imprtapp     0.059086   0.104214   0.567   0.5711
lk4whldr    -0.064119   0.126739  -0.506   0.6132
kidsbulk    -0.096959   0.122063  -0.794   0.4275
wntguzlr    -0.028943   0.115689  -0.250   0.8026
nordtrps     0.073056   0.127473   0.573   0.5669
stylclth     0.015757   0.113597   0.139   0.8898
strngwrn    -0.196806   0.113448  -1.735   0.0836 .
passnimp     0.161975   0.119056   1.360   0.1745
twoincom     0.170419   0.096469   1.767   0.0781 .
nohummer     0.009052   0.095697   0.095   0.9247
aftrschl    -0.025716   0.116551  -0.221   0.8255
accesfun    -0.003458   0.122112  -0.028   0.9774
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.215 on 369 degrees of freedom
Multiple R-squared:  0.3729,    Adjusted R-squared:  0.3219
F-statistic: 7.314 on 30 and 369 DF,  p-value: < 2.2e-16
```

3 Factor Analysis

Pre-factor diagnostic analysis

Bartlett test of sphericity

X2 = 7884.075

df = 435

p-value < 2.22e-16

Kaiser-Meyer-Olkin

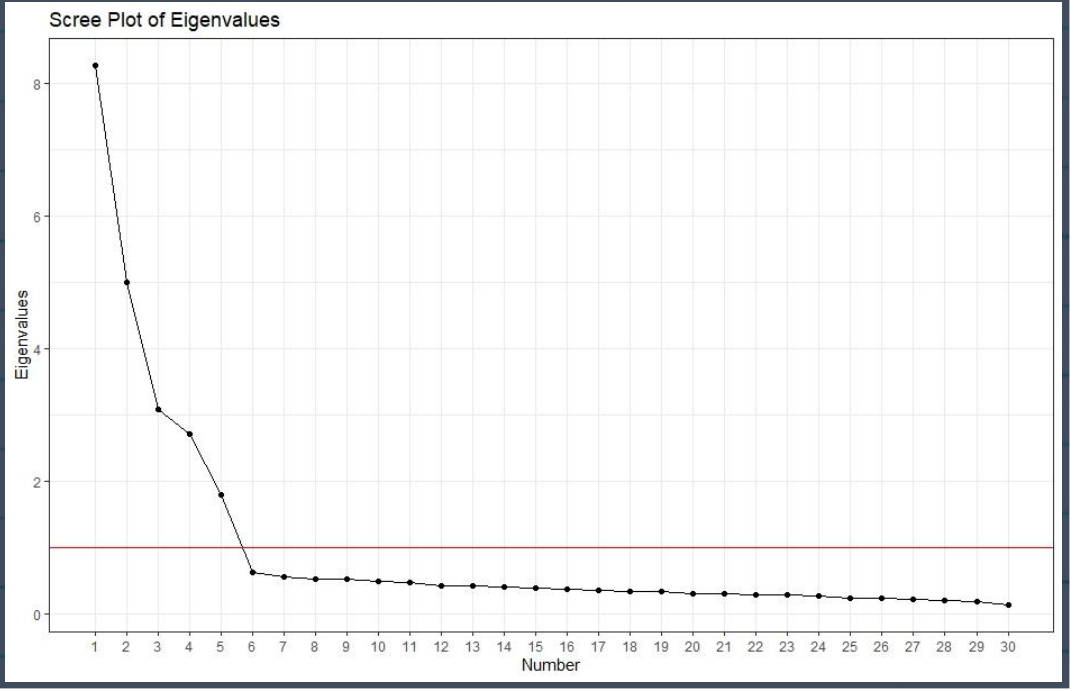
KMO-Criterion: 0.9233418

Pre-factor analysis diagnostics confirmed that factor analysis is appropriate for the data.

Scree Plot

5 factors

are appropriate to conduct the analysis.



Factor loadings

	Factor Loading	Descriptive Names/Labels
RC1	Carefmny:-0.7634; buyhghnd:0.8147; tkvacatn:0.6539; lthrbetr:0.7107; accesfun:0.6784; pricqual:0.7823; passnimp:-0.6475; stylclth:0.6036; twoincom:0.7562; prmsound:0.6819; imprtapp:0.5094	Main: Prestige Secondary: Lifestyle, Quality-seeking, Luxury (High-end)
RC2	Miniboxy:0.8419; suvcmpct:0.8191; homlrgst:-0.6794; noparkrm 0.8066; next2str:-0.7429; secbiggr:0.7593; needbetw:0.7575; nohummer:0.7064;	Main:Space-efficiency Secondary: Size-Alternative, Parking Practicality
RC3	kidtrans 0.9330; nordtrps -0.8665; kidsbulk 0.8248; aftrschl 0.7754;	Main: Family-friendliness Secondary: Family activity / Kids-oriented
RC4	safeimpt 0.9075; perfimpt -0.8836; lk4whldr 0.8556; strngwrn 0.7354;	Main: Safety
RC5	envrminr -0.8666; shdcarpl 0.8665; wntguzlr -0.7628;	Main:Environmental-friendliness

Naming factors

#RC1: Prestige

#RC2: Space-efficiency

#RC3: Family-friendliness

#RC4: Safety

#RC5: Environment-friendliness

4 Regression: Validation

Overall, with better model fit, statistical significance, we can generate insight with the new regression model using **factor scores**, which is definitively better than the regression with 30 variables.

Regression summary

Our second regression using factor scores as explanatory variables has a slight higher Adjusted R-squared of 0.3267 (slightly better model fit).

More importantly, Factors 1,2, and 4 all have high t-test statistics with near-zero p-values (significant at alpha = 0.0001). Factor 1&2 has a positive coefficient, indicating that high preference for luxury/quality and space efficiency have a positive impact on customers' **purchase intent**.

The negative coefficient for factor 4 indicates that high preference for safety negatively impacts the customers' purchase intent.

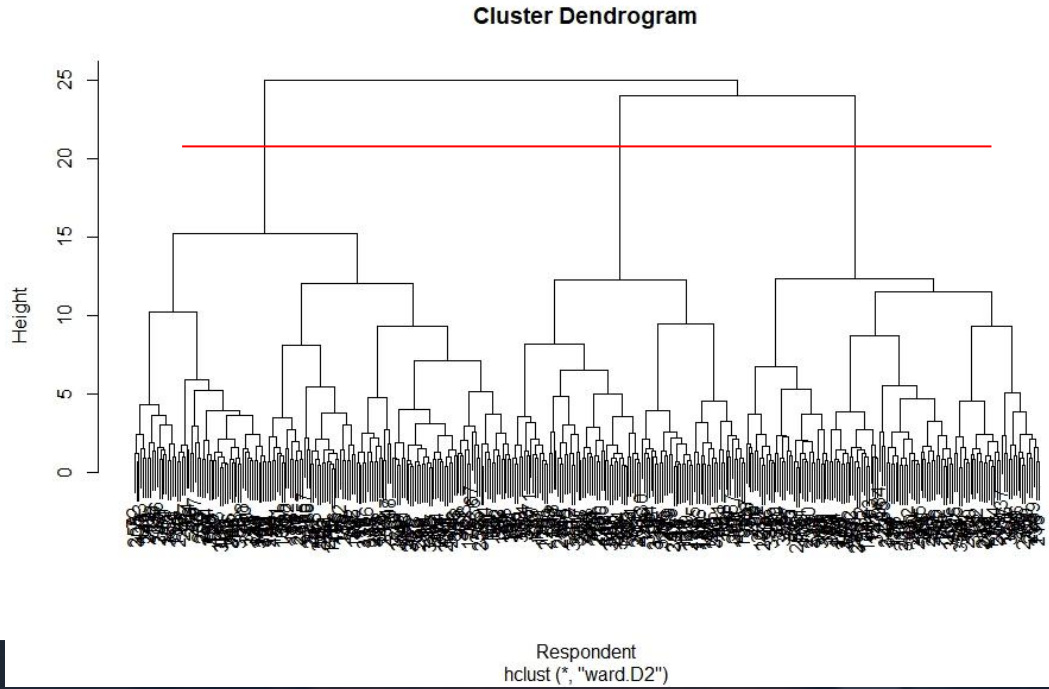
In contrast, factor 3 and 5 are not significant at alpha = 0.05 level, indicating that these factors lack the explanatory power for the concept liking variable (Mvliking) compared to the others.

```
Call:
lm(formula = mydata$mvliking ~ RC1 + RC2 + RC3 + RC4 + RC5, data = q4df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9530 -1.5723 -0.0992  1.6137  6.1489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8425     0.1104  43.878  < 2e-16 ***
RC1           1.0439     0.1105   9.447  < 2e-16 ***
RC2           0.9854     0.1105   8.918  < 2e-16 ***
RC3           0.1838     0.1105   1.664    0.097 .
RC4          -0.5570     0.1105  -5.040  7.1e-07 ***
RC5          -0.1418     0.1105  -1.284    0.200
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.207 on 394 degrees of freedom
Multiple R-squared:  0.3351,    Adjusted R-squared:  0.3267
F-statistic: 39.72 on 5 and 394 DF,  p-value: < 2.2e-16
```

5 Cluster Analysis

A hierarchical clustering dendrogram, K-means analysis and diagnostic plot helped determine that we will form three clusters.

Dendrogram



Three seem to be the optimal number of clusters based on dendrogram.

K-means

```
> k3$size
[1] 157 129 114
> k3$centers
         RC1         RC2        RC3        RC5         RC4
1 -0.9323024  0.04589391 -0.3249301 -0.5114655 -0.09657823
2  0.5949299 -1.09336582  0.1546210  0.2860140  0.01612336
3  0.6107503  1.17402497  0.2725255  0.3807394  0.11476200
```

Cluster 1: RC1+RC5

Cluster 2: RC2+RC1

Cluster 3: RC2+RC1+RC5+RC3

Naming clusters

Segment (Cluster)	RC1	RC2	RC3	RC4	RC5	Factors with high absolute scores	Needs & wants
1	-0.9323024	0.04589391	-0.3249301	-0.09657823	-0.5114655	RC1 (negative) RC5 (negative)	budget
2	0.5949299	-1.09336582	0.1546210	0.01612336	0.2860140	RC2 (negative) RC1 (positive)	spaciousness & prestige
3	0.6107503	1.17402497	0.2725255	0.11476200	0.3807394	RC2 (positive) RC1 (positive)	space-efficiency & prestige, fuel efficiency

6 Clusters vs. Concept Liking

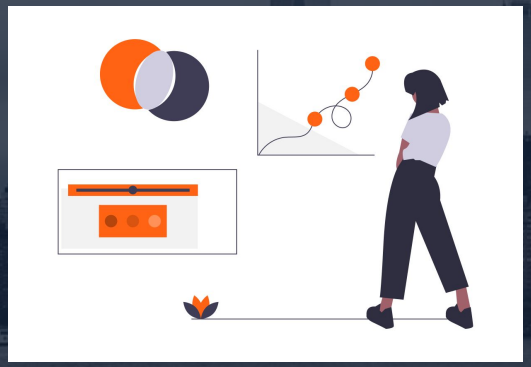
We performed Cross-tab and regression analysis to determine how the clusters vary on the concept liking variable.

Approach 1
Cross-tab

Difference	Purchase Intention								
Cluster	1	2	3	4	5	6	7	8	9
1	2.29	1.50	0.87	1.54	-0.98	-0.37	-0.82	-2.13	-2.28
2	0.58	-0.13	0.94	0.30	1.03	-0.69	0.11	-1.03	-1.36
3	-3.31	-1.62	-2.01	-2.13	0.05	1.16	0.85	3.60	4.12

Approach 2
Regression

The mean of concept linking for **cluster three** is 6.6, ranking the highest among all clusters.



```
Call:
lm(formula = mvliking ~ threecluster, data = mydata)

Residuals:
     Min      1Q  Median      3Q     Max
-5.6053 -1.8981  0.1019  2.1019  5.1019

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.8981     0.1951  19.985   <2e-16 ***
threecluster2  0.5360     0.2904   1.846   0.0657 .
threecluster3  2.7072     0.3007   9.002   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.444 on 397 degrees of freedom
Multiple R-squared:  0.1787,    Adjusted R-squared:  0.1745
F-statistic: 43.18 on 2 and 397 DF,  p-value: < 2.2e-16
```

7

Reality Check

As a final step, examine the demographic profile of each of the three clusters.

Cluster Demographic glance

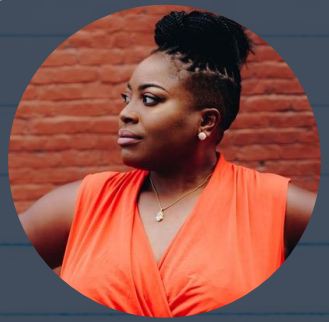
Segment (Cluster)	Segment size (# respondents)	Purchase intent (mean mvliking)	Needs & wants	Demographics						
				Age	Annual household income	Annual miles driven by household	# of kids living in household	Education	Gender	Recycling activity
1	157	3.9	budget	Below 40	Below $50,000	Below 20,000 miles	0-1 kid	Undergrad Degree and below	(undifferentiated across segments)	(undifferentiated across segments)
2	129	4.4	spaciousness & prestige	40-49	$50,000 and above	15,000 miles and above	At least 1 kid	Undergrad Degree and above		
3	114	6.6	space-efficiency & prestige	40-49	Above $50,000	20,000 miles and above	At least 1 kid	Undergrad Degree and above		

Persona

Bill Jackson

(Cluster 1: size 157)

Bill is a 36-year old Bank Teller without any children. Bill has an undergraduate degree and an annual income of less than $50,000.



Raina Galán

(Cluster 2: size 129)

Raina is a Hollywood producer & a mother of one. She makes 150k$ a year and loves to drive around to travel with her family. She drives over 15,000 miles every year.

Daniel Årud

(Cluster 3: size 114)

Daniel is a board member at balarah living in Beverly Hills, who works remotely and loves to enjoy family activities with his kids. He makes 130k$ and drives his car for more than 20,000 miles every year.

Notes

- Demographics for each segment are aligned with their needs and wants, e.g. segments 2 and 3 who have high annual household income want luxury.
- Cluster 2 and 3 cover multiple age groups (30 to 60). However, since targeting a wide range of age groups is an inefficient targeting strategy, we elect to describe/label both clusters with the predominant age group (40 to 49).

8 Who are going to drive Microvan?

Prestige MicroBus

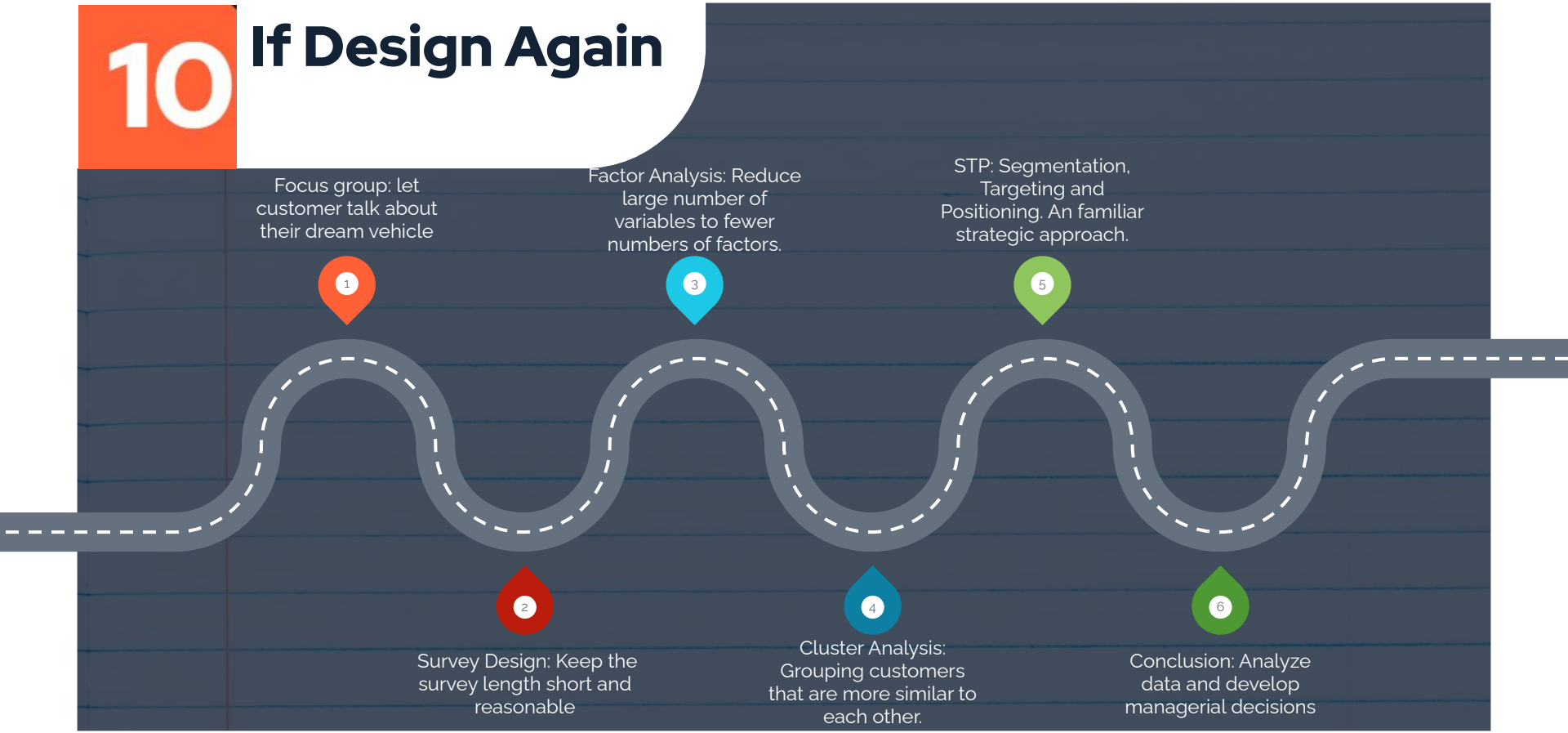
For middle to upper-middle class generation X,

our product is a symbol of prestige among all **space-efficient** microvans because it fills the market vacancy for high-end, quality yet spacious vehicles.

9 Research Limitations

- Possibly erroneous demographic data
- Inclusion of currently irrelevant variables which could be useful for further applications
- The survey focus on the customers and the manufacturer without consideration of the overall market and the competitors

10 If Design Again



Thank you

- Group 15:
- Dongzhou Wei, Bill Ye, Zhao Zhang, Raina Lang, Li Fang